COGSCI'19

Creativity+Cognition+Computation

24 - 27 JULY 2019 MONTREAL, CANADA

Invited Speakers

Elizabeth Churchill | Mary Lou Maher | Takeshi Okada

Co-Chairs

Ashok Goel | Colleen Seifert | Christian Freksa

Introduction

Dear Cognitive Science Colleagues,

Welcome to the 41st Annual Conference of the Cognitive Science Society in Montreal, Canada! Our meeting brings together some of the most innovative and exciting research in Cognitive Science today, and highlights the conference theme of *Creativity + Cognition + Computation*.

In addition to the Rumelhart Prize presentation by Michelene Chi and the Carvalho-Heineken Prize presentation by Nancy Kanwisher, the program features three plenary speakers: Elizabeth Churchill (Google Research), Mary Lou Maher (University of North Carolina), and Takeshi Okada (University of Tokyo). Further, the program includes the Jacobs Foundation Symposium, *How Curious? The Need for Exploration and Discovery*, as well as an invited symposium on *Creativity in the Arts* in addition to the Rumelhart Symposium on *Translation Research in STEM Learning* and the Glushko Ph.D. Dissertation Awards Symposium.These invited symposia and talks showcase the conference theme.

The program committee for CogSci 2019 received 1110 submissions, including 810 full papers, 256 member abstracts, 13 publication-based short papers, as well as 14 proposals for symposia, 10 for workshops, and 8 for tutorials. After a rigorous review process, the committee selected 202 papers for oral presentation and inclusion in the conference proceedings (25%), 306 papers for poster presentation and inclusion in the proceedings (38%), and 163 papers for poster presentation with inclusion of abstracts in the proceedings (20%). We also selected 204 submitted member abstracts and accepted another 19 abstracts from full paper submissions as invited member abstracts. In addition, we accepted 12 publication-based talks, 10 symposia, 7 workshops, and 4 tutorials to make for a very rich and inclusive program.

We hope that you enjoy the program this year as well as the beautiful city of Montreal!

Your Program Co-Chairs,
Ashok Goel (Georgia Institute of Technology, USA)
Colleen Seifert (University of Michigan, USA)
Christian Freksa (University of Bremen, Germany)

Acknowledgements

We are very grateful to everyone who contributed to the planning and organization of this year's Cognitive Science meeting, to all authors who submitted their contributions, and to all reviewers who generously donated their expertise and time to evaluate the submissions. We thank the members of the Program Committee who coordinated the reviews and made the tough decisions about submissions, and the members of the conference organizing subcommittees who showed initiative in completing their demanding tasks. These Organizing and Program Committee members are listed below.

We are especially grateful for the assistance of a number of individuals and groups critical to handling the many organizational aspects of the meeting. We thank Michael Frank, the Chair of the Cognitive Science Society, Anna Drummey, the Executive Officer of the Society, and the entire Governing Board of the Society, for their advice and support throughout the process. Lily Chang at *International Conference Services*, Jude Ross at *Podium Conferences*, and James Stewart at *Precision Conference Solutions* have been helpful, effective, and constant partners during the long process. Chuck Kailish and Timothy Rogers, two of the Co-Chairs of last year's conference, offered help whenever we needed them. Additional help included key contributions from Thomas Barkowsky for the reviewing process, Andrea Patalano for the awards organization, and Sungeun An for creating the conference poster.

Finally, we are grateful to the Cognitive Science Society and to the sponsors of this conference, including the Robert J. Glushko and Pamela Samuelson Foundation, the Jacobs Foundation, Facebook AI, DeepMind Technologies and the Weinberg Institute of Cognitive Science for their support.

Enjoy!

Ashok Goel, Colleen Seifert, and Christian Freksa
Co-Chairs, Cognitive Science 2019

Sponsors

The Robert J. Glushko and Pamela Samuelson Foundation
The Jacobs Foundation
Facebook AI
DeepMind Technologies
The Weinberg Institute for Cognitive Science
The Cognitive Science Society

Thank you again for your support!

How to Cite Your Paper

APA formatted citation for a paper:

Author, A. & Author, B. (2019). This is the title of the paper. In A.K. Goel, C.M. Seifert, & C. Freksa (Eds.), *Proceedings of the 41st Annual Conference of the Cognitive Science Society* (pp. PAGES). Montreal, QB: Cognitive Science Society.

APA formatted citation for a published abstract:

Author, A. & Author, B. (2019). This is the title of the abstract. In A.K. Goel, C.M. Seifert, & C. Freksa (Eds.) *Proceedings of the 41st Annual Conference of the Cognitive Science Society* (p. NUMBER). Montreal, QB: Cognitive Science Society.

APA formatted citation for a talk/poster presentation:

Author. A. & Author, B. (2019, July). This is the title of the talk or poster. Paper (or Poster) presented at the 41st Annual Conference of the Cognitive Science Society. Montreal, Quebec, Canada.

Organizing Committee

Program

Ashok Goel (Georgia Institute of Technology, USA)
Colleen Seifert (University of Michigan, Ann Arbor, USA)
Christian Freksa (University of Bremen, Germany)

Awards

Andrea Patalano (Wesleyan University, USA)
Richard Lewis (University of Michigan, USA)
Michael Mozer (University of Colorado, USA)
Wendy Newstetter (Georgia Institute of Technology, USA)
Steve Sloman (Brown University, USA)

Communications

Rick Dale (University of California, USA)
Rui Meng (University of Wisconsin, USA)

Student Volunteers

Matthew Setzler (Indiana University, USA)
Peter Felsman (University of Michigan, USA)

Local CogSci Community

Isabelle Soulieres (Université du Québec à Montréal, Canada)
Thomas Shultz (McGill University, Canada)

Member Abstracts

Maithilee Kunda (Vanderbilt University, USA)
Sebastien Hélie (Purdue University, USA)
Ana-Maria Olteteanu (Freie Universität Berlin, Germany)
Michael Shafto (Cognitive Science Associates)

Member Publication-Based Talks Committee

Jim Davies (Carleton University, Canada)
Evangelia Chrysikou (Drexel University, USA)

Symposia Committee

Nora Newcombe (Temple University, USA)
Garrison Cottrell (University of California, USA)

Tutorials and Workshops Committee

Thomas Barkowsky (University of Bremen, Germany)
Travis Seymour (University of California, USA)
Elizabeth Bonawitz (Rutgers University, USA)
Swaroop Vattam (MIT Lincoln Laboratory, USA

Sponsorship

Keith McGreggor (Georgia Institute of Technology, USA)

Program Committee

Dor Abrahamson
Erik Altmann
Elena Andonova
Elisabeth André
Blair Armstrong
Mike Barley
Daniel Bartels
Paul Bello
Andrea Bender
Sven Bertel
Julie Boland
Bert Bredeweg
Tad Brunye
Nicholas Bryan-Kinns
Daphna Buchsbaum
Bruce Burns
Martin Butz
Amilcar Cardoso
Peter Carruthers
Peter Cheng
Evangelia Chrysikou
Tim Clausner
Garrison Cottrel
Scotty Craig
Jennifer Culbertson
David Danks
Jim Davies
Virginia De Sa
Gedeon Deak
Morteza Dehghani
M. Belen Diaz-Agudo
Lauren Dilley
Steven Dow
Catharine Echols
Susan Epstein
Caitlin Fausey
Anna Fisher
Kenneth Forbus
Wai-Tat Fu
Susan Gelman
Dedre Gentner
John Gero
Samuel Gershman
Tobias Gerstenberg
Kevin Gluck
Adele Goldberg
Laura Gonnerman
Cleotilde Gonzalez
Tom Griffiths
Maurice Grinberg
Elizabeth Gunderson
Glenn Gunzelmann
Todd Gureckis
Mary Hayhoe
Mary Hegarty
Sebastien Hélie
Keith Holyoak
Janet Hsiao
Edward Hubbard
Mutsumi Imai
Bipin Indurkhya
Ion Juvina
Irene Kan
Mark Keane
Christopher Kello
Celeste Kidd
Pia Knoeferle
Kenneth Koedinger
Stefan Kopp
Gustav Kuhn
Maithilee Kunda
David Leake
Michael Lee
Roger Levy
Rick Lewis
Peggy Li
Antonio Lieto
Tania Lombrozo
Andrew Lovett
Christian Luhmann
Maryellen MacDonald
Michael Mack
Brian Magerko
Lorenzo Magnani
Asifa Majid
Art Markman
Amy Masnick
Jay McClelland
Catherine Mello
Janet Metcalfe
Laura Michaelis
Bradley Morris
Stefan Münzer
J. William Murdock
Nora Newcombe
Sergei Nirenburg
David Noelle
John Pani
Anna Papafragou
Andrea Patalano
Sturt Patrick
David Peebles
Rafael Perez y Perez
Amy Perfors
Hugh Rabagliati
Marco Ragni
Gisela Redeker
Serge Robert
Robert Saunders
Ute Schmid
Gregor Schöner
Chris Schunn
Travis Seymour
Meredith Shafto
Michael Shafto
Priti Shah
Chris Sims
Steven Sloman
Jennifer Spenader
Michael Spranger
Mahesh Srinivasan
Keith Stenning
Stephen Smith
Ron Sun
Heike Tappe
Thora Tenbrink

Joshua Tenenbaum
Hannu Toivonen
Greg Trafton
Barbara Tversky
Dan Ventura
Gabrialla Vigilocco
Alan Wagner
Christoph Weidemann
Katherine White
Geraint Wiggins
Fei Xu
Yang Xu
Jerry Zhu

Student Volunteers

Nadia Blostein, McGill University
Zixian Chai, Stanford University
Dania Chatila, McGill University
Gabriela Iwama, Max Planck Institute for Intelligent Systems
Dimosthenis Kontogiorgos, KTH Royal Institute of Technology
Ezgi Mamus, Radboud University
Sebastian Musslick, Princeton University
Pauline Palma, McGill University
Armand Rotaru, University College London
Matt Rounds, University of Edinburgh
Rose Schneider, UC San Diego
Tanmay Sinha, ETH Zurich
Oana Stanciu, Central European University; BPF
Robert Thorstad, Emory University
Mehrgol Tiv, McGill University
Naomi Vingron, McGill University

Awards

Robert J. Glushko Dissertation Prizes

The Cognitive Science Society and the Glushko-Samuelson Foundation award up to five outstanding dissertation prizes in cognitive science each year. The goals of these prizes are to increase the prominence of cognitive science and encourage students to engage in interdisciplinary efforts to understand minds and intelligent systems. The hope is that the prizes will recognize and honor young researchers conducting ground-breaking research in cognitive science. The eventual goal is to aid in efforts to bridge between the areas of cognitive science and create theories of general interest to the multiple fields concerned with scientifically understanding the nature of minds and intelligent systems. Promoting a unified cognitive science is consistent with the belief that understanding how minds work will require the synthesis of many different empirical methods, formal tools, and analytic theories. 2011 was the inaugural year of this prize, and a new competition is held annually. The 2019 recipients of the Robert J. Glushko Prizes for Outstanding Doctoral Dissertations / Theses in Cognitive Science are:

Kirsten Adam – University of Chicago, 2018
Characterizing the Limits of Visual Working Memory

Max Kleiman-Weiner – **Massachusetts Institute of Technology, 2018**
Computational Foundations of Human Social Intelligence

Martin Maier – Humboldt University, 2018
Language, Meaning, and Visual Perception: Event-Related Potentials Reveal Top-Down Influences on Early Visual Processing

Jean-Paul Noel – Vanderbilt University, 2018
Leveraging Multisensory Neurons, Circuits, Brains, and Bodies to Study Consciousness: From the Outside-In and the Inside-Out

Katharine Tillman – University of California, 2017
Constructing the Concept of Time: Roles of Language, Perception, and Culture

Marr Prize

The Marr Prize, named in honor of the late David Marr, is awarded to the best student paper at the conference. All student first authors were eligible for the Marr Prize for the best student paper. The Marr Prize includes an honorarium of $1000 and is sponsored by The Cognitive Science Society. The winners of the 2019 Marr Prize for the Best Student Paper is:

Jose M. Ceballos, University of Washington, *The Role of Basal Ganglia Reinforcement Learning in Lexical Priming and Automatic Semantic Ambiguity Resolution*

Nicolas Oliver Riesterer, Universität Freiburg, *Modeling Human Syllogistic Reasoning: The Role of "No Valid Conclusion"*

Computational Modeling Prizes

Four prizes worth $1000 each are awarded for the best full paper submissions to CogSci 2019 that involve computational cognitive modeling. The four prizes represent the best modeling work in the areas of perception/action, language, higher-level cognition, and applied cognition. These prizes are sponsored by The Cognitive Science Society. The winners of the 2019 Computational Modeling Prizes are:

Applied Cognition:

Douglas Guilbeault, University of Pennsylvania, *The Social Network Dynamics of Category Formation*

Higher-Level Cognition:

Ardavan S. Nobandegani, McGill University, *A Resource-Rational Process-Level Account of the ST. Petersburg Paradox*

Perception & Action:

Yunyan Duan, Northwestern University, *A Rational Model of Word Skipping in Reading: Ideal Integration of Visual and Linguistic Information*

Language:

Benjamin Peloquin, Stanford University, *The Interactions of Rational, Pragmatic Agents Lead to Efficient Language Structure and Use*

Sayan Gul Award

Sayan Gul was an undergraduate at UC Berkeley studying cognitive science and computer science, and had great potential as a cognitive scientist. He died tragically while traveling to the Annual Conference of the Cognitive Science Society for the presentation of his research. This award is intended to support similarly outstanding undergraduates conducting research in cognitive science. In honor of Sayan Gul, the Sayan Gul Award supports undergraduate students with travel related costs who are presenting authors at the conference. The Sayan Gul Award includes a cash award of $500. This year's winner of the award is:

Megumi Sano, Stanford University, *Graphical Convention Formation During Visual Communication*

Diversity & Inclusion Travel Awards

Five prizes will be award to support travel to the conference for graduate students who bring diversity to the society, in particular under-represented racial/ethnic groups and citizens of under-represented countries (Zone B Society members) who are presenting at the conference. Each travel award includes a cash award of $1,000. This year's travel awards recipients are:

Jose M. Ceballos, University of Wisconsin, *The Role of Basal Ganglia Reinforcement Learning in Lexical Priming and Automatic Semantic Ambiguity Resolution*

Tania Delgado, University of California San Diego, *Differences in Learnability of Pantomime Versus Artificial Sign: Iconicity, Cultural Evolution, and Linguistic Structure*

Nianyu Li, Peking University, *A Conceptual Model of Self-Adaptive Systems Based on Attribution Theory*

Che Lucero, Cornell University, *Unconscious Number Discrimination in the Human Visual System*

Mukesh B. Makwana, Centre of Behavioural and Cognitive Sciences, Mumbai, *Hands in Mind: Learning to Write with Both Hands Improves Inhibitory Control, but Not Attention*

Guilherme Sanches de Oliveira, University of Cincinnati, *Bee-ing In the World: Phenomenology, Cognitive Science, and Interactivity in a Novel Insect-Tracking Task*

Staci Meredith Weiss, Temple University, *Individual Differences in Bodily Attention: Variability in Anticipatory Mu Rhythm Power Is Associated with Executive Function Abilities and Processing Speed*

Student Travel Awards

The Robert J. Glushko and Pamela Samuelson Foundation generously sponsored $10,000 for student travel awards. Travel awards have been provided to students whose submissions were accepted as full papers, received high rankings, and who indicated a need for travel funding. This year's travel awards went to:

Nicolas Collignon, University of Edinburgh
Douglas Guilbeault, University of Pennsylvania
Ethan Hurwitz, University of California, San Diego
Akila Kadambi, University of California, Los Angeles
Kei Kashiwadate, Deniki University
Lara Kirfel, University College London
Sang Ho Lee, Ohio State University
Ashley Leung, University of Chicago
Mahi Luthra, Indiana University
Olivia Miske, Arizona State University
Sebastian Musslick, Princeton University
Benjamin Peloquin, Stanford University
Nicolas Riesterer, University of Freiburg
Harrison Ritz, Brown University
Jennifer Sloane, University of New South Wales
Leila Straub, ETH Zurich
Karina Tachihara, Princeton University
Charley Wu, Max Plack Institute for Human Development
Yueyuan Zheng, University of Hong Kong

Rumelhart Prize Presentation

Michelene Chi, Arizona State University

Translating the ICAP Theory of Cognitive Engagement Into Practice

Carvalho-Heineken Prize Presentation

Nancy Kanwisher, MIT

Functional Imaging of the Human Brain: A Window in the Architecture of the Human Mind

Keynote Talks

Elizabeth Churchill, Google Research

Cognition, Collaboration, and Creativity: Google's Material Design as a Case Study

Mary Lou Maher, University of North Carolina

Computational Models of Creativity: Curiosity, Novelty, and Surprise.

Takeshi Okada, University of Tokyo

Inspiration and Artistic Creation

Rumelhart Symposium

Translation Research in STEM Learning

Jim Slotta, University of Toronto, Moderator
Kristy Boyer, University of Florida
Kirsten R Butcher, University of Utah
Percival G Matthews, University of Wisconsin
Jodi Davenport, WestEd

Jacobs Foundation Symposium

How Curious? The Cognitive Need for Exploration and Discovery

Elizabeth Bonawitz, University of New Jersey, Rutgers
Tobias Hauser, University College London
Allyson Mackey, University of Pennsylvania
Celeste Kidd, University of California, Berkeley

Invited Symposium

Creativity in the Arts

David Kirsh, University of California San Diego, Moderator
Gil Weinberg, Georgia Tech
Brian Magerko, Georgia Tech
Valentina Nisi, University of Madeira

Glushko Awards Symposium

Kirsten C. S. Adam, University of California San Diego
Martin Maier, Humboldt-University Berlin
Jean-Paul Noel, Vanderbilt University and New York University
Katharine A. Tillman, University of Texas
Max Kleiman-Weiner, Harvard University

Table of Contents

Papers with Oral Presentations

Papers with Poster Presentations

Heuristics, hacks, and habits: Boundedly optimal approaches to learning, reasoning and decision making

Ishita Dasgupta[1], **Eric Schulz**[1], **Jessica B. Hamrick**[2] & **Joshua B. Tenenbaum**[3]
[1]Department of Psychology, Harvard University
[2]DeepMind, London, UK
[3]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology

Humans regularly perform tasks that require combining information across several sources of information to learn, reason, and make decisions. Bayesian models provide a computational framework, and a normative account, for how humans carry out these tasks. However, exact inference is intractable in most real-world situations, and extensive empirical work shows that human behavior often deviates significantly from the Bayesian optimum. A promising possibility is that people instead approximate rational solutions using bounded available resources. In this workshop, we bring together leading researchers from cognitive science, neuroscience and machine learning to build a better understanding of bounded optimality in how humans learn, reason and make decisions.

Keywords: Heuristics; Resource rationality; Reasoning; Decision making; Reinforcement learning; Machine learning

Introduction

This workshop will cover work that casts human and machine learning, decision making and reasoning as *boundedly optimal*. In particularly, we will focus on meta-reasoning, reinforcement learning, active information acquisition, and probabilistic reasoning.

The notion that the mind approximates rational (Bayesian) inference has had a strong influence on thinking in psychology since the 1950s. However, people deviate from Bayesian ideals in several well-documented instances (6), giving rise to the idea that they rely on heuristic rules instead (5). Nonetheless, people can behave in ways that approximate Bayesian inference in complex domains such as (active) learning (2), reasoning (1) and decision making (14). How can these apparently contradictory findings be explained?

One idea is that people approximate rational solutions using limited available resources, a proposal often discussed under the terms of resource or computational rationality (4; 7). In light of limited resources, boundedly optimal solutions to complex problems can take the form of sampling-based approximations (3), simplified decision rules (13), pruning of low-value options (9), or through an adaptation of information acquisition to the structure of the task (12). However, how exactly the different approaches should be combined to produce a fully-developed theory of bounded optimality that transfer across domains and tasks is still an open question, with some researchers proposing that intelligent agents can meta-reason about which strategies to apply (10), and others stressing the connections between heuristic and Bayesian inference (11) and the role of inductive biases (8).

Goal and scope

The aim of this workshop is to bring together scientists who have a joint interest in how resource-constrained agents solve realistic problems, such as making decisions, finding rewards, acquiring information or reasoning and learning about the world. We have invited leading researchers from cognitive science and machine learning interested in the computational foundations of bounded optimality. In particular, our goal is to facilitate discussion and help build a more unified notion of rationality that takes resource and computational limitations into consideration. Key questions of discussion will include:

- How can we formalize theories of bounded optimality?
- What is a good framework and what are good domains in which to benchmark progress in developing such theories?
- What can we learn from past debates on and formalizations of rationality?
- Do agents learn different context-specific boundedly optimal strategies? How might they recognize when to apply which strategy?
- What does a bounded agent optimize, if at all? How can bounded optimality cope with the curse of dimensionality?

Target audience

This workshop fits well with this year's focus on "Creativity + Cognition + Computation". These key elements of cognition are precisely those that drive modern accounts of bounded optimality and are features of human intelligence that modern theories of rationality seek to explain. Our target audience is interdisciplinary and almost as broad as the conference as a whole — we expect this workshop to be of interest to cognitive psychologists, linguists, developmental psychologists, neuroscientists, philosophers and machine learning researchers alike. The workshop's webpage can be found at: https://hacksandhabits.github.io

Organizers and presenters

Ishita Dasgupta (Organizer) is a PhD-student at Harvard University working in Samuel Gershman's Computational Cognitve Science lab. Ishita's work explores how people and machines make resource rational approximations to difficult problems, in particular in the domains of probability estimation, hypothesis generation, and intuitive physics.

Eric Schulz (Organizer) is a Data Science Postdoctoral Fellow at Harvard University. Eric studies generalization as function learning with a particular focus on compositionality and reinforcement learning.

Jessica B. Hamrick (Organizer) is a Research Scientist at DeepMind. Her research focuses on cognitive science-inspired theories of machine learning. In particular, she focuses on the role of mental simulation and resource rational approximations.

Joshua B. Tenenbaum (Organizer) is Professor of cognitive science at MIT. Josh's lab sits at the intersection of cognitive science and machine learning, with a focus on hallmarks of human intelligence; in particular, the ability to learn efficiently and flexibly from limited data.

Paula Parpart is a postdoc at the University of Warwick working with Prof. Neil Stewart. Her research has focused on reconciling heuristic and Bayesian views of rationality in decision making.

Falk Lieder leads the Rationality Enhancement Group at the Max Planck Institute for Intelligent Systems in Tübingen. His mission is to build a scientific foundation and practical tools for helping people become more effective by supporting cognitive growth, goal setting, and goal achievement.

Tom Griffiths is a Professor of Psychology and Computer Science at Princeton University. Tom develops mathematical models of higher level cognition to understand the formal principles that underlie people's ability to solve everyday computational problems.

Özgür Şimşek is a Senior Lecturer in Machine Learning at the University of Bath. Her research is on algorithms that can learn from limited experience in complex, real-word environments, with a focus on reinforcement learning.

Neil Bramley is a Lecturer of Cognitive Psychology at the University of Edinburgh. His work focuses on how people actively construct and use causal models to guide their interactions with the natural world.

Azzurra Ruggeri is a Max Planck Research Group Leader at the MPI for Human Development in Berlin. Her research focuses on how children and adults actively search for information when making decisions, drawing causal inferences and solving categorization tasks.

Kelsey Allen is a graduate student advised by Josh Tenenbaum at MIT. She uses computational models and behavioral experiments to study the development of intuitive theories, in particular intuitive physics in planning and reinforcement learning contexts.

Peter Dayan is a director at the Max Planck Institute for Biological Cybernetics in Tübingen. His research focuses on the computational neuroscience of learning and decision making, with a focus on neuromodulation, meta-control and computational psychiatry.

Workshop structure

We propose a full-day workshop consisting of three parts. The first two parts will be a series of 20 minute talks. The final part will be a panel discussion about the limits and future of bounded optimality in cognitive science.

The morning session will consist of the following talks:

Presenter	Topic
Eric Schulz	Optimizing with confidence
Paula Parpart	Heuristics as Bayesian inference
Falk Lieder	Learning how to decide
Ishita Dasgupta	Learning to infer
Josh Tenenbaum	Computational rationality

The afternoon session will consist of the following talks:

Presenter	Topic
Jessica Hamrick	Resource-rational mental simulation
Tom Griffiths	Bridging Marr's levels
Özgür Şimşek	Exploiting the statistical properties of decision environments
Neil Bramley	Neurath's ship: Incremental active theory-building
Azzurra Ruggeri	Ecological active learning
Kelsey Allen	Hacks in intuitive theories
Peter Dayan	Slothful serial; perilous parallel processing

The final 45 minutes will be a *panel discussion*.

References

P. W. Battaglia, J. B. Hamrick, and J. B. Tenenbaum. Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, page 201306572, 2013.

N. R. Bramley, P. Dayan, T. L. Griffiths, and D. A. Lagnado. Formalizing neurath's ship: Approximate algorithms for online causal learning. *Psychological review*, 124(3):301, 2017.

I. Dasgupta, E. Schulz, and S. J. Gershman. Where do hypotheses come from? *Cognitive psychology*, 96:1–25, 2017.

S. J. Gershman, E. J. Horvitz, and J. B. Tenenbaum. Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, 349(6245):273–278, 2015.

G. Gigerenzer and H. Brighton. Homo heuristicus: Why biased minds make better inferences. *Topics in cognitive science*, 1(1):107–143, 2009.

T. Gilovich, D. Griffin, and D. Kahneman. *Heuristics and biases: The psychology of intuitive judgment*. Cambridge university press, 2002.

T. L. Griffiths, F. Lieder, and N. D. Goodman. Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in cognitive science*, 7(2):217–229, 2015.

J. B. Hamrick, K. R. Allen, V. Bapst, T. Zhu, K. R. McKee, J. B. Tenenbaum, and P. W. Battaglia. Relational inductive bias for physical construction in humans and machines. *arXiv preprint arXiv:1806.01203*, 2018.

Q. J. Huys, N. Eshel, E. O'Nions, L. Sheridan, P. Dayan, and J. P. Roiser. Bonsai trees in your head: how the pavlovian system sculpts goal-directed choices by pruning decision trees. *PLoS computational biology*, 8(3):e1002410, 2012.

F. Lieder, D. Plunkett, J. B. Hamrick, S. J. Russell, N. Hay, and T. Griffiths. Algorithm selection by rational metareasoning as a model of human strategy selection. In *Advances in neural information processing systems*, pages 2870–2878, 2014.

P. Parpart, M. Jones, and B. C. Love. Heuristics as bayesian inference under extreme priors. *Cognitive psychology*, 102:127–144, 2018.

A. Ruggeri and T. Lombrozo. Children adapt their questions to achieve efficient search. *Cognition*, 143:203–216, 2015.

Ö. Şimşek. Linear decision rule as aspiration for simple decision heuristics. In *Advances in Neural Information Processing Systems*, pages 2904–2912, 2013.

C. M. Wu, E. Schulz, M. Speekenbrink, J. D. Nelson, and B. Meder. Generalization guides human exploration in vast decision spaces. *Nature Human Behaviour*, 2(12):915, 2018.

Guided Playful Learning: Developmental, Computational, and Educational Perspectives

Emily N. Daubert (emily.daubert@rutgers.edu)
Department of Psychology, 101 Warren Street
Newark, NJ 07102 USA

Patrick Shafto (patrick.shafto@gmail.com)
Department of Mathematics and Computer Science, 110 Warren Street
Newark, NJ 07102 USA

Keywords: guided play; computational modeling; science of learning; playful learning

Workshop Summary

Learning is a continuous process that is contingent on temporal, developmental, and social factors. Well-timed guidance is critical for successful learning. Further, the needs of learners vary with their stage of development, and the social context in which learning takes place has important implications for learning. Researchers from a variety of backgrounds, including cognitive development, computational modeling, and education have explored these various components in isolation, however, understanding learning requires the examination of the interactions between the temporal, developmental, and social factors involved.

Interactions among these factors are of critical importance in fields such as cognitive development, computational modeling, and education. Take, for example, educational settings, where didactic approaches such as direct instruction have been favored over more free-play based approaches (Stockard & Engelmann, 2008). In direct instruction, learning is not just social, it is adult-initiated and adult-led, and by its nature less responsive to temporal factors that may affect a learners performance. Free play, in contrast, allows the learner to lead, which allows greater responsivity to temporal changes. Aside from the developmental merits, the debate between direct instruction and free play is emblematic of the need for a better understanding of how the social and temporal components interact to foster learning (Yu et al., 2018). Similar issues arise in the developmental and modeling literatures.

Recently, guided playful learning has been put forth as an integrative child-led, adult-assisted approach for promoting learning. However, many unanswered questions remain regarding the interplay of factors involved in guided playful learning. The goal of this workshop is to bring together an interdisciplinary group of researchers, with expertise in cognitive development, computation, and education in an effort to merge these separate literatures, draw general conclusions, and develop directions for future research.

Research in cognitive development on the effectiveness of guided playful learning is mixed. There is some evidence that guided learning is more effective than adult-led discovery (i.e. direct instruction) and unassisted discovery (i.e. free play) for promoting learning in children (Honomichl & Chen, 2012). However, some research indicates that direct instruction is equally, if not more effective, in achieving explicit learning goals (Becker & Gersten, 1982). Others still find that there is no substitute for the wide-ranging benefits of child-initiated free play, which is intrinsically motivated (Rubin, Fein, & Vandenberg, 1983). One possible reason for the differing conclusions is different definitions of guidance, which have included questioning, modeling, enhanced materials, and feedback. Thus, it remains unclear what kinds of guidance are most effective for promoting learning. Understanding the nature of effective guidance will also help to clarify the underlying cognitive mechanisms that lead to changes in children's knowledge.

In computational modeling, there has not yet been significant progress toward an understanding of guided playful learning. Research has investigated free exploration. Two versions of this that are prominent in the literature are active learning, which is commonly formalized as maximizing Expected Information Gain of the next observation (Russo & Van Roy, 2014), and reinforcement learning which maximizes expected reward over time (Niv et al., 2015). Research has also investigated instruction. For example, models have formalized selection of data by a knowledgeable and helpful teacher as well as formalized learning from such data, where the learner reasons both about the data and the teacher's intent (Shafto, Goodman, & Griffths, 2014). Guided playful learning lies at the nexus of these three themes, where guidance aims to foster learning over time through self directed exploration. Moreover, guided playful learning requires modeling of when to provide guidance, which adds layer of complexity not considered in this previous work.

In education, researchers have asked if guided playful learning is effective in various domains of learning. Specifically, guided playful learning may be more effective in domains in which learning is promoted through child-led exploration, as with causal learning (Bonawitz et al., 2011). Similarly, guided playful learning may promote learning in domains in which child engagement is crucial, such as literacy (Lillard & Else-Quest, 2006). But it remains unclear if guided learning is effective in domains that are

traditionally associated with rote memorization, such as mathematics. In addition, educational researchers have focused on the role of individual differences in guided playful learning. The effectiveness of guidance content can be influenced by individual differences, such as children's cognitive style, background knowledge, socioeconomic status, and language learner status.

Developing a unified theoretical and empirical understanding of guided playful learning will allow for the discovery of the complex interplay of temporal, developmental, and social factors in children's learning. By bringing together researchers from traditionally distinct communities, we hope to begin to answer this foundational set of questions about the nature of cognition.

Workshop Structure

The workshop will feature well-known experts from different fields. The workshop will also invite poster submissions from the broader cognitive science community, with "poster teasers" flash talks related to guided playful learning. Additionally, the schedule has built in ample time for questions for mini-panels of each sub-area of guided playful learning, ensuring maximum opportunity for audience engagement.

Proposed Schedule

9:00-9:15: Opening Remarks (Elizabeth Bonawitz)

9:15-10:45: The Role of Play in the Development of Knowledge

Roberta M. Golinkoff *"A helping hand: Adult-infant play and infant category learning"*
Yuan Meng *"Leveraging self-explanation to scaffold causal learning in children"*
Pierre-Yves Oudeyer *"Computational models of intrinsically motivated learning, autonomous goal setting, and how it can self-organize long-term developmental structures"*

10:45 – 11:00: Coffee/Tea Break

11:00–12:30: Intuitive Pedagogy in Playful Learning

Kathleen H. Corriveau *"Variability in parent-child guidance during dyadic STEM learning"*
Todd Gureckis *"Modeling intuitive teaching using POMDPs"*
Maureen Callanan *"Children learning about science through family conversations"*

12:30-1:40: Lunch

1:40-2:00: Poster Teasers

2:00-3:30: Inferential Consequences in Guided Play

Ilona Bass *"A computational account for the exploratory benefits of guided play"*
Emily N. Daubert *"Promoting psychosomatic understanding using pedagogical questions during storybook reading"*
Patrick Shafto *"A unified computational framework of learning for oneself and from others"*

3:30-3:45: Coffee/Tea Break

3:45-4:15: Poster Viewing

4:15-5:15: Bringing Guided Play to the Classroom and Beyond

Jamie Jirout *"Exploring to learn: Methods of encouraging curiosity in the lab and in the classroom"*
Kathy Hirsh-Pasek *"Playful learning landscapes: Where guided play meets architectural design"*

5:15-5:30: Closing Remarks (Elizabeth Bonawitz)

References

Bonawitz, E., Shafto, P., Gweon, H., Goodman, N.D., Spelke, E., & Schulz, L. (2011). The double-edged sword of pedagogy: Instruction limits spontaneous exploration and discovery. *Cognition, 120*(3), 322-330.

Becker, W., & Gersten, R. (1982). A follow-up of Follow Through: The later effects of the direct instruction model on children in fifth and sixth grades. *Am Ed Res J, 19,* 75-92.

Honomichl, R. D., and Chen, Z. (2012). The role of guidance in children's discovery learning. *Review Cog Sci 3*, 615–622.

Lillard, A. & Else-Quest, N. (2006). The early years: Evaluating Montessori education, *Science, 313,* 1893.

Niv, Y., Daniel, R., Geana, A., Gershman, S. J., Leong, Y. C., Radulescu, A., & Wilson, R. C. (2015). Reinforcement learning in multidimensional environments relies on attention mechanisms. *The J of Neuro, 35*(21), 8145-57.

Rubin, K.H., Fein, G., & Vandenberg, B. (1983). *Play*. In E.M. Hetherington (Ed.), Handbook of child psychology: Socialization, personality, and social development. Wiley.

Russo, D. & Van Roy, B. (2017). Learning to optimize via information-directed sampling. *Ops Res, 66*(1), 230-252.

Shafto, P., Goodman, N. D., and Griffths, T. L. (2014). A rational account of pedagogical reasoning: teaching by, and learning from, examples. *Cog Psych, 71*, 55–89.

Stockard, J. & Engelmann, K. (2008). *Academic kindergarten and later academic success: The impact of direct instruction*. Eugene, OR: National Institute for Direct Instruction.

Yu, Y., Shafto, P., Bonawitz, E., Yang, S., Golinkoff, R.M., Corriveau, K.H., Hirsh-Pasek, K., & Xu, F. (2018). The theoretical and methodological opportunities afforded by guided play with young children. *Front Psych, 9*, 1152.

Using replication studies to teach research methods in cognitive science

Joshua R. de Leeuw (jdeleeuw@vassar.edu)
Jan Andrews (andrewsj@vassar.edu)
Ken Livingston (livingst@vassar.edu)
Department of Cognitive Science, Vassar College

Michael Franke (mchfranke@gmail.com)
Institute for Cognitive Science, University of Osnabrück

Josh Hartshorne (joshua.hartshorne@bc.edu)
Department of Psychology, Boston College

Robert Hawkins (rxdh@stanford.edu)
Department of Psychology, Stanford University

Jordan Wagge (jordan.wagge@avila.edu)
Department of Psychology, Avila University

Keywords: pedagogy; replication; research methods; education

Overview

Some instructors of research methods classes are conducting authentic (i.e., publishable) replication studies with their classes (de Leeuw et al., 2018; Hartshorne et al., 2019; Hawkins et al., 2018; Leighton, Legate, LePine, Anderson, & Grahe, 2018; Wagge et al., 2019). This practice has, potentially, both pedagogical benefits for students and broader benefits for the scientific community (Frank & Saxe, 2012; Standing et al., 2014). Students experience an authentic research process from design through publication, providing opportunities for instruction on many different aspects of the research pipeline. When done with care, replications from the classroom become a valuable part of the scientific literature, and students fulfill an underserved role in science: performing direct replications (Everett & Earp, 2015).

Adding authentic replication work to a research methods class naturally raises many questions about pedagogy and implementation. What studies should be replicated? How can an appropriate sample for the replication be obtained, especially at small institutions? What can instructors do to ensure that students, who may be conducting research for the first time, are able to produce quality work that meets the standards of publication? What aspects of the research process should students contribute the most to, and what aspects should be controlled by the instructor?

There are many reasonable answers to these questions. With the growing adoption of replication studies in courses, a diverse set of classroom-tested approaches now exists. This creates the possibility for sharing, synthesizing, and improving teaching strategies, which is the goal of this workshop. This workshop brings together instructors who have conducted replication work with their research methods classes to discuss their successes and failures. These instructors have taught classes at the undergraduate and graduate level. Students in the classes have conducted behavioral studies (both in-lab and online) and EEG studies. The classes vary in structure (students may work as an entire class, in small groups, individually, or as part of a larger collaborative endeavor across many classes) and points of emphasis in the research process.

Workshop Structure

This is a half-day workshop. The first portion of the session will feature presentations from instructors (listed below) describing how replication studies have been utilized in their classes and how replication studies fit into the broader pedagogical goals of the class. In the second portion of the workshop, the presenters will discuss questions from the audience and a moderator in a panel format. Audience contributions to the discussion will be welcome.

Target Audience

The workshop welcomes anyone with an interest in teaching research methods, including both current instructors and students and postdocs who plan to teach research methods in the future. We hope that workshop attendees will leave with concrete ideas for how to incorporate replication work into their own research methods classes.

Presenters

Josh de Leeuw, Jan Andrews, & Ken Livingston (Vassar College) have co-taught undergraduate Research Methods in Cognitive Science. In their course, students begin the semester by conducting a replication study and then develop one or more novel follow-up experiments. They will discuss how conducting a replication prepares students to design and execute their own original research, and how

working with undergraduate students on drafting a manuscript for submission to a journal provides a different kind of opportunity for teaching scholarly writing.

Jordan Wagge (Avila University) is the Associate Director of the Collaborative Replications and Education Project (CREP), a project that promotes and scaffolds crowdsourced replication work through student research. She will discuss how CREP can support replication work in methods courses, including sample assignment guidelines for instructors who seek to incorporate CREP work into their courses.

Joshua Hartshorne (Boston College) has taught three iterations of his course Language Acquisition & Development. Although not a methods course, it contains a substantial lab component. Through a series of group projects, each class of approximately 10 students completes 5-6 replications. The presentation will discuss how to incorporate a lab component into a content class. It will also discuss how to use replications as a vehicle for teaching programming, statistics, and best practices.

Robert Hawkins (Stanford University) recently led a classroom replication effort as part of the graduate course "Lab in Experimental Methods". In this course, each student chooses a paper to replicate based on their own research interests and proceeds independently through a structured series of milestones with supervision from instructors. The presentation will discuss this pedagogical workflow, how the replication model can be adapted for students of different levels, and the challenges that arise in managing a wide diversity of projects.

Michael Franke (University of Osnabrück) has taught two classes that combine undergraduate/graduate levels with a dual focus: one on theoretical issues concerning reproducibility and open science, and another on conveying first practical experiences with behavioral experiments by means of a replication project. The courses required students to preregister their replication and make all analysis scripts available prior to data collection. The presentation will discuss the challenges and opportunities of making especially undergraduates appreciate solutions (e.g., preregistration & large-scale replications) to problems (e.g., abundant researcher degrees of freedom) they have not experienced first-hand yet.

References

de Leeuw, J. R., Andrews, J., Altman, Z., Andrews, R., Appleby, R., Bonanno, J., … Shriver, A. (2018). Similar event-related potentials to structural violations in music and language: A replication of Patel, Gibson, Ratner, Besson, & Holcomb (1998). *PsyArXiv.* doi:10.31234/osf.io/e9w3v

Everett, J. A. C., & Earp, B. D. (2015). A tragedy of the (academic) commons: Interpreting the replication crisis in psychology as a social dilemma for early-career researchers. *Frontiers in Psychology, 6*, Article 1152. doi:10.3389/ fpsyg.2015.01152

Frank, M. C., & Saxe, R. (2012). Teaching replication. *Perspectives on Psychological Science*, 7(6), 600–604.

Hartshorne, J. K., Skorb, L., Dietz, S. L., Garcia, C. R., Iozzo, G. L., Lamirato, K. E., … Others. (2019). The Meta-Science of Adult Statistical Word Segmentation: Part 1. *Collabra: Psychology, 5*(1). doi:10.1525/collabra.181

Hawkins, R. X. D., Smith, E. N., Au, C., Arias, J. M., Catapano, R., Hermann, E., … Frank, M. C. (2018). Improving the replicability of psychological science through pedagogy. *Advances in Methods and Practices in Psychological Science*, 1(1), 7–18.

Leighton, D. C., Legate, N., LePine, S., Anderson, S. F., & Grahe, J. (2018). Self-Esteem, self-disclosure, self-expression, and connection on Facebook: A collaborative replication meta-analysis. *Psi Chi Journal of Psychological Research*, 23(2), 98–109.

Standing, L. G., Grenier, M., Lane, E. A., Roberts, M. S., & Sykes, S. J. (2014). Using replication projects in teaching research methods. *Psychology Teaching Review, 20*(1), 96-104.

Wagge, J. R., Baciu, C., Banas, K., Nadler, J. T., Schwarz, S., Weisberg, Y., … Grahe, J. E. (2019). A demonstration of the Collaborative Replications and Education Project: Replication attempts of the red-romance effect, *Collabra: Psychology*, 5(1), 5. doi: 10.1525/collabra.177

Measuring Creativity - Workshop Proposal

Ana-Maria Olteţeanu (ana-maria.olteteanu@fu-berlin.de)
Human-Centered Computing, Freie Universität Berlin,
Germany

Keywords: creative cognition; creativity psychometrics; creative problem solving; computational modelling; computational creativity; creativity metrics and methods; intelligent systems; creativity assistive systems

Workshop Proposal

Various methods exist for measuring creativity, most of them in the form of creativity tests, like the Remote Associates Test [Mednick, 1962], the Alternative Uses Test [Guilford, 1956], TTCT [Kim, 2006], the Wallach-Kogan tests [Wallach and Kogan, 1965], insight problems [Maier, 1931, Duncker, 1945, Cunningham et al., 2009], etc.

However, the feasibility and dependability of various types of psychometric assessment and administration of measures, as pertaining to various creativity tasks, have recently been questioned and enriched [Beisemann et al., 2018, Hass, 2015, Hass et al., 2018, Hass and Beaty, 2018, Wilken et al., 2018]. The thought and work on the measurement of creativity are witnessing a new revival.

Recently, new methods of computationally creating stimuli for greater measurement accuracy have been developed [Olteţeanu et al., 2017, Olteţeanu, 2016, Olteteanu and Yoopoo, 2017], inspired by artificial cognitive systems that solve creativity tests [Olteţeanu et al., 2018]. Such computational psychometrics methods have already shown to provide designs with greater control [Olteţeanu and Schultheis, 2017] and the computational resurrection of tests which were initially proposed theoretically [Olteţeanu et al., 2018].

This workshop will focus on building a red thread of discussion on the current state of creativity psychometrics, integrating topics on existing classic and novel, manual and computational methods of testing and measuring creativity. The following questions will be addressed:

(i) What creativity measuring methods exist and what are their strengths and weaknesses?
(ii) Which creativity factors are measured by the existing creativity methods? Is there an overlap of measuring methods for different factors? Are they factors for which no methods exist or current methods are not yet up to the task?
(iii) What is the suitability of existing current methods for empirical testing versus computational modelling?
(iv) How can comparability be ensured across creativity test item sets?
(v) What creativity metrics and methods can be used in evaluating the computational modeling of creativity?
(vi) What is the impact of artificial cognitive systems and their evalution on creativity metrics? Of computational creativity systems and their evaluation?
(vii) What are the new computational and automatized measures of creativity, and what is their role in the ecosystem of measures?
(viii) Subjective and objective measures in creativity.

Workshop Duration and Organization

We propose a half a day workshop for the presentation, discussion and elaboration of creativity measuring methods. The workshop will involve three elements:

(i) Three invited speakers from different backgrounds (Cognitive Psychology, Cognitive Neuroscience, Cognitive Systems - Computer Science) will present existing creativity measuring methods (details below).
(ii) Short presentations of papers and posters will be accepted on the topic.
(iii) The workshop will end with a panel discussion, focused on establishing future directions for methods and systems aimed at supporting creativity and problem solving.

Publication: The papers submitted for this workshop will be published as a CEUR-WS volume. If enough high quality papers are received, a Special Issue will be proposed by the organizer to the *Cognitive Systems Research* journal, or a topic proposal will be made to TopiCS in Cognitive Science.

Topics for this workshop will be centered around, but not limited to:

- Creative cognition
- Creativity measures and Tests
- Psychometrics for Creative Cognition
- Computational methods for measuring creative cognition
- Computational modelling
- Artificial creative cognitive systems
- Creative problem solving
- Computational Creativity
- Evaluation of natural and computational cognitive systems
- Associativity and Conceptual Spaces
- Semantic networks and semantic graphs
- Ill structured problem solving and Structured representations
- Knowledge discovery
- Creativity modeling approaches and their relation to evaluation, including Case based reasoning, Neural networks, Evolutionary algorithms
- Analogy and Metaphor
- Creative assistive systems

Speakers

- **Richard Hass** – Thomas Jefferson University, US. Talk topic: *Improving Measures on Creative Object Uses*. Background: Cognitive Psychology.
- **Evangelia Chrysikou** – Drexel University, US. Talk topic: *A standardized test for creativity based on the Alternative Uses Task*. Background: Cognitive Neuroscience.

- **Ana-Maria Olteţeanu** – Head of Cognitive Systems, Freie Universitat Berlin– Talk topic: *Computational Measures of Creativity*. Background: Cognitive Systems – Computer Science.

Organizer - Short biography

Ana-Maria Olteţeanu is the Principal Investigator of the „Creative problem solving in cognitive systems" (CreaCogs) project funded by the German Research Foundation (DFG) at the Freie Universitat Berlin, Germany.

Ana-Maria has a cross-disciplinary background: she holds a PhD in Musicology (2011) and a *summa cum laude* Doctorate in Cognitive Systems and Artificial Intelligence (2016). Her thesis got nominated for the EurAI Dissertation Prize, and won the OLB 1st Prize for the best Doctoral Dissertation in Science in NW Germany in the last two years (2017).

Ana-Maria authored more than 30 papers on the topic of creative problem solving, of which five journal articles focus on developing artificial cognitive systems and computational measures for creativity psychometrics. Her book *Cogs in the Creative Machine* will be published by Springer in June 2019. Ana-Maria has reviewed more than 40 papers for over 20 international conferences and journals, and gave over 20 conference and invited talks on creative cognitive systems. Dr. Dr. Olteţeanu has been a program committee member of 15 workshops and conferences in the field. She organized and chaired 4 Symposia/Workshops/conference tracks, and is the editor of four volumes and special issues on creativity related topics. Together with Sebastien Helie, Ana-Maria will write the chapter on *Computational Models of Creativity* in the upcoming edition of The Cambridge Handbook of Computational Cognitive Sciences. Ana-Maria's interests are related to natural and artificial cognitive systems, creative problem solving, cognitive modeling, computational psychometrics, knowledge discovery and spatial reasoning.

Recent Organizing and Editorial Experience

2018 - 2021 – Editorial Board member, Cognitive Systems Research Journal.

2018 – Organizer and Chair of the workshop *Computational Methods and Systems for the Cognitive Modelling and Support of Creativity and Creative Problem Solving*, at the Cognitive Science Conference in Madison, Wisconsin, 2018 (over 50 participants)

2018-2019 – Topic Editor for *Frontiers in Psychology-Cognitive Science* and *Frontiers in Artificial Intelligence and Robotics*, for the Topic *Creativity from Multiple Cognitive Science Perspectives* (with Bipin Indurkhya).

2018-2019 – Guest Associate Editor for *Frontiers in Psychology-Cognitive Science* and *Frontiers in Artificial Intelligence and Robotics*, for the Topic *Creativity from Multiple Cognitive Science Perspectives* (with Bipin Indurkhya).

2017-2018 – Guest editor of the Cognitive Systems Research journal, for the special issue on *Problem-solving, Creativity and Spatial Reasoning in Cognitive Systems* (with Zoe Falomir).

2017 – Editor of the *Proceedings of the 2nd Symposium on Problem-solving, Creativity and Spatial Reasoning in Cognitive Systems*, CEUR-Ws vol. 1869 (with Zoe Falomir).

2017 – Co-organized the *ProSocrates - Problem solving, creativity and spatial reasoning in cognitive systems* Symposium, at the Hanse Wissenschafts-Kolleg, Delmenhorst, Germany.

2016 – Co-organized the *ProSocrates - Problem solving, creativity and spatial reasoning in cognitive systems* Symposium, at the German Cognitive Science Society conference - Space for Cognition, Bremen (Germany).

References

Beisemann, M., Forthmann, B., Christian Bürkner, P., and Holling, H. (2018). Psychometric evaluation of an alternate scoring for the remote associates test. *The Journal of creative behavior*.

Cunningham, J. B., MacGregor, J. N., Gibb, J., and Haar, J. (2009). Categories of insight and their correlates: An exploration of relationships among classic-type insight problems, rebus puzzles, remote associates and esoteric analogies. *The Journal of Creative Behavior*, 43(4):262–280.

Duncker, K. (1945). On problem solving. *Psychological Monographs*, 58(5, Whole No.270).

Guilford, J. P. (1956). The structure of intellect. *Psychological bulletin*, 53(4):267.

Hass, R. (2015). Feasibility of online divergent thinking assessment. *Computers in Human Behavior*, 46.

Hass, R. and Beaty, R. (2018). Use or consequences: Probing the cognitive difference between two measures of divergent thinking. *Frontiers in Psychology*, 9:2327.

Hass, R., Rivera, M., and Silvia, P. (2018). On the dependability and feasibility of layperson ratings of divergent thinking. *Frontiers in Psychology*, 9.

Kim, K. H. (2006). Can we trust creativity tests? A review of the Torrance Tests of Creative Thinking (TTCT). *Creativity Research Journal*, 18(1):3–14.

Maier, N. R. (1931). Reasoning in humans. II. The solution of a problem and its appearance in consciousness. *Journal of Comparative Psychology*, 12(2):181.

Mednick, S. (1962). The associative basis of the creative process. *Psychological review*, 69(3):220.

Olteţeanu, A.-M. (2016). In *Proceedings of the Workshop on Computational Creativity, Concept Invention, and General Intelligence (C3GI2016)*, volume 1767, Osnabrück. CEUR-Ws.

Olteţeanu, A.-M., Schottner, M., and Schuberth, S. (2018). Computationally resurrecting the functional remote associates test using cognitive word associates and principles from a computational solver. *Knowledge-Based Systems*.

Olteţeanu, A.-M. and Schultheis, H. (2017). What determines creative association? revealing two factors which separately influence the creative process when solving the remote associates test. *The Journal of Creative Behaviour*.

Olteţeanu, A.-M., Schultheis, H., and Dyer, J. B. (2017). Computationally constructing a repository of compound Remote Associates Test items in American English with comRAT-G. *Behavior Research Methods, Instruments, & Computers*.

Olteţeanu, A.-M., Falomir, Z., and Freksa, C. (2018). Artificial cognitive systems that can answer human creativity tests: An approach and two case studies. *IEEE Transactions on Cognitive and Developmental Systems*, 10(2):469–475.

Olteteanu, A.-M. and Yoopoo, K. (2017). Towards computationally creating multi-answer queries for the remote associates test. In *Proceedings of the 5th International Workshop on Artificial Intelligence and Cognition*, volume 2090, pages 34–40. CEUR-Ws.

Wallach, M. A. and Kogan, N. (1965). *Modes of thinking in young children: A study of the creativity-intelligence distinction.* Holt, Rinehart & Winston.

Wilken, A., Forthmann, B., and Holling, H. (2018). Instructions moderate the relationship between creative performance in figural divergent thinking and reasoning capacity. *The Journal of creative behavior*.

Acknowledgements

The author acknowledges the support of the German Science Foundation (DFG) for the organization of this workshop, via the grant Creative Cognitive Systems (CreaCogs – OL 518/1-1).

Predicting Individual Human Reasoning: The PRECORE-Challenge

Marco Ragni (ragni@cs.uni-freiburg.de) and Nicolas Riesterer (riestern@cs.uni-freiburg.de)
Cognitive Computation Lab, Georges-Khler-Allee 52
Freiburg, 79110

Sangeet Khemlani (sunny.khemlani@nrl.navy.mil)
Navy Center for Applied Research in Artificial Intelligence, US Naval Research Laboratory
Washington, DC, USA 20002

Keywords: Cognitive Modeling; Syllogistic Reasoning; Individual Modeling; Challenge

Short Summary

Most computational models of cognition are based on aggregate data. In recent years, skepticism about group-to-individual generalizability has begun to emerge (Fisher, Medaglia, & Jeronimus, 2018). Simultaneously, results have shown that the current state in modeling reasoning is approaching a ceiling caused by the focus on aggregation (Riesterer, Brand, & Ragni, 2018). The time is ripe to adopt a new perspective on the challenge of cognitive modeling: how to model the individual reasoner. In addition to explaining aggregate data from training datasets, computational cognitive models can adapt to an individual by integrating knowledge about past responses into the prediction mechanism. This workshop will tackle conceptual, computational, theoretical, and methodological challenges in modeling individual reasoning behavior. A recent methodological advancement in assessing both aggregate and individual reasoning behavior, the Cognitive Computation for Reasoning Analysis (CCOBRA) framework, will be used to propose a new competition for theory-driven computational models of individual reasoning behavior. This workshop, and its underlying theoretical challenge, invites participants from cognitive science, AI, and all related fields to learn to build computational models of individual reasoners.

Core challenge: Modeling individuals

How can cognitive scientists build robust simulations of individual reasoners? This workshop will address the theoretical and methodological challenges in developing PREdictive, individualized COgnitive models of REasoning – the PRECORE Challenge. An orthodox methodology for fitting cognitive models to a dataset concerns a two-fold procedure: a given cognitive model's parameters are set by learning to predict the outcomes from a training dataset, and then it is applied to a novel dataset that the model never encountered before. The methodology is often used to build models of aggregated behavior form multiple individuals, but in principle, it can be applied to assessing individual reasoning behavior as well. The Cognitive Computation for Behavioral Reasoning Analysis (CCOBRA) framework is a benchmarking tool implemented in Python that actively integrates the individual human into the prediction loop. At its core lies a close connection to psychological experiments. Models are expected to simulate the experimental procedure for individual participants. They are presented with the same task in the same sequence with the same response options. By providing precise responses to individual tasks, models are evaluated based on their predictive accuracies. In the CCOBRA framework, computational models are supplied with the true response, both in the training phase, as well as in the evaluation phase; in this way, models can learn a default set of parameter settings in training and then be used to detect individual strategies in reasoning in the evaluation phase to refine their predictions further. Models are allowed to train on a dataset consisting of tasks and the actual human responses of individuals not present in the evaluation data. Additionally, after predicting the response to a task, they are presented with the true response and thus allowed to adapt to an individual participant. Hence, CCOBRA extends the traditional cognitive modeling problem by moving beyond the level of aggregates. As a result, the challenge for computational cognitive models is more difficult, but the payoffs are greater, i.e., they can lean to the development of robust computational models of individual reasoning strategies and adapt to the constraints of individual reasoners.

Models are ultimately compared via their predictive accuracy on unseen data. If a model manages to hit the true response more often than another model, the CCOBRA framework assigns it a higher score. The framework operates in a domain-agnostic fashion, i.e., it is compatible with computational cognitive models based on symbolic, probabilistic, connectionist, or hybrid approaches. Hence, computational cognitive models in the CCOBRA framework are assessed and compared on a fair and neutral ground. The only requirements imposed by CCOBRA is an implementation based on Python and the capability of generating a precise prediction for a given task. The problem of overfitting will be tackled by computing the final evaluation scores on previously unreported data. Higher predictive scores in the CCOBRA framework correspond directly to a better grasp of the processes underlying an individual human reasoner's cognitive system. The project is entirely open-source and accessible via Github[1].

[1] https://github.com/CognitiveComputationLab/ccobra

Benchmarking data and example model implementations can be found in the repository. A companion website[2] exists which allows to quickly upload and evaluate model implementations without the need to install the framework.

A domain-general challenge

Cognitive scientists have built computational models that simulate a wide variety of reasoning behavior, e.g., reasoning about syllogisms, reasoning about relations, reasoning about sentences and propositions, and reasoning about causation. Theorists have built computational models of reasoning in only some of these domains – and they've constructed models of individual reasoners in only one of them. Hence, the challenge of analyzing individual reasoning behavior is acute. This workshop, and its underlying benchmarking methodology, seeks to develop domain-general solutions for developing models of individuals. Consider the domain of syllogistic reasoning, for instance. Syllogisms are problems built from categorical assertions of the form "All of the As are Bs" and "All of the Bs are Cs". Reasoners deduce conclusions from syllogisms by comprehending two premises responding to the prompt: "What, if anything, follows?" Most reasoners generate spontaneously generate a conclusion of the form "All of the As are Cs" to the two premises above. As a recent meta-analysis shows, some syllogisms are easy, and some are difficult (Khemlani & Johnson-Laird, 2012). The same meta-analysis showed that twelve theories syllogistic reasoning had difficulty explaining the variation reasoners exhibit. The problem is endemic to computational models of reasoning: many of them perform well on aggregated data, but they they are unable to account for the individual differences that become relevant when attempting to predict how individual reasoners respond to various problems (Riesterer et al., 2018). Models in all reasoning domains are presently have an upper bound by the most frequent response.

Goals and Scope

The central goal of the workshop is to encourage and enhance cognitive modeling of syllogistic reasoning on an individual level and discussions by researchers of such diverse fields of cognitive science as psychology, AI, linguistics, and philosophy. Participation is possible by any of the following: Presenting a 15 minutes talk about cognitive modeling (please send us an email by July, 1), submitting a model for the modeling task in CCOBRA, discussing statistical analysis of aggregated vs. individual reasoning, or providing any insights in the discussion for advancing the current state of modeling beyond the level of aggregate syllogistic data.

Workshop Organization

Marco Ragni is a DFG-Heisenberg fellow and associate professor at the technical faculty of the Albert-Ludwigs-University Freiburg and leads the Cognitive Computation Lab. His research interests include qualitative spatio-temporal reasoning, knowledge representation and reasoning, cognitive modeling, and complex cognition with a special focus on analyzing why and how human reasoning often deviates from classical logical approaches.

Homepage: www.cc.uni-freiburg.de
Email: ragni@cs.uni-freiburg.de
Pub: dblp.uni-trier.de/pers/hd/r/Ragni:Marco

Nicolas Riesterer is a PhD student at the Cognitive Computational Lab, associated with the Department of Computer Science of the Albert-Ludwigs-University Freiburg. His research interests are centered around developing predictive models for human reasoning based on approaches from both cognitive science and AI.

Homepage: www.cc.uni-freiburg.de
Email: riestern@cs.uni-freiburg.de
Pub: dblp.uni-trier.de/pers/hd/r/Riesterer:Nicolas

Sangeet Khemlani is a computational cognitive scientist in the Navy Center for Applied Research in Artificial Intelligence at the US Naval Research Laboratory. His work focuses on building computational cognitive models of deductive, inductive, and abductive reasoning, and testing those models against a wide variety of behavioral data.

Homepage: www.khemlani.net
Email: sunny.khemlani@nrl.navy.mil
Pub: www.khemlani.net/publications/

Committee

- Ruth Byrne, University of Dublin
- Christoph Beierle, Fernuniversität Hagen, Germany
- Ulrich Furbach, University of Koblenz, Germany
- Steffen Hölldobler, University of Dresden, Germany
- Markus Knauff, Universitt Gießen, Germany
- Gabriele Kern-Isberner, TU Dortmund, Germany
- Frieder Stolzenburg, Harz University of Applied Sciences, Germany

References

Fisher, A. J., Medaglia, J. D., & Jeronimus, B. F. (2018, jun). Lack of group-to-individual generalizability is a threat to human subjects research. *Proceedings of the National Academy of Sciences*, *115*(27), E6106–E6115.

Khemlani, S., & Johnson-Laird, P. N. (2012). Theories of the syllogism: A meta-analysis. *Psychological Bulletin*, *138*(3), 427–457.

Riesterer, N., Brand, D., & Ragni, M. (2018). The predictive power of heuristic portfolios in human syllogistic reasoning. In *Lecture notes in computer science* (pp. 415–421). Springer International Publishing.

[2] http://orca.informatik.uni-freiburg.de/ccobra/

Everyday Activities

Holger Schultheis (schulth@informatik.uni-bremen.de)
Institute for Artificial Intelligence, University of Bremen
Am Fallturm 1, 28359 Bremen, Germany

Richard P. Cooper (r.cooper@bbk.ac.uk)
Department of Psychological Sciences, Birkbeck, University of London
Malet Street, London, London WC1E 7HX UK

Keywords: everyday activities; complex tasks; control of action sequences; action planning; demographic change

Introduction

Humans perform a wide range of everyday activities (e.g., preparing a meal, setting the table) frequently, and often without conscious thought. Despite the experienced ease with which we perform such activities, their successful completion involves a complex set of abilities and mechanisms. This becomes apparent when considering that even healthy adults exhibit occasional errors (Norman, 1981, e.g., failing to spoon coffee grinds into the filter before switching on the coffee machine) in performing the necessary actions, while mild cognitive impairment may interfere with successful performance of highly familiar everyday activities (Gold, Park, Troyer, & Murphy, 2015).

Successful performance of everyday activities taxes at least the following abilities:

- *Perception*: The environment in which the actions are performed has to be adequately perceived to properly act in it. Among others this comprises the ability to recognize largely occluded objects in cluttered environments (e.g., plates in a stack of plates or objects in a dishwasher).
- *Action Planning*: Everyday activities consist of several actions and the effectiveness and efficiency of performing activities will often depend on the order in which the actions are executed (see coffee making example above). Accordingly, planning one's actions is an important aspect of everyday activity performance.
- *Spatial Reasoning*: Spatial relations of objects to each other and to one's body are crucial for everyday activity. Without knowledge about these relations, locomotion in the environment as well as collecting and properly arranging objects would not be possible.
- *Movement Planning*: Individual (motor) actions require planning to, for example, avoid obstacles, remain in the operational range of one's effectors, and to reduce the chance for mishaps (reaching with a full cup over – instead of around – your laptop is not a good idea)
- *Controlling Action Sequences*: Action sequences not only have to be planned, but also controlled during execution to ensure that no actions are left out, actions are not executed in the wrong order, or that inappropriate (i.e., not part of the plan) actions that are habitual or appropriate given the current state of the environment are avoided.
- *Monitoring and Error Correction*: Given that slips and lapses in action execution occur, monitoring of progress towards the goal and error correction mechanisms are also need to ensure successful action completion.

Considering that the listed abilities constitute research areas in their own right, it seems clear that gaining a (more) comprehensive understanding of everyday activities is an ambitious endeavor. At the same time, everyday activities provide an opportunity to jointly research several cognitive abilities in what Newell (1973) has called *complex tasks*. Everyday activities such as "setting the table" are circumscribed enough to study them in the lab, while being complex enough to require the combination of several cognitive abilities. As such, investigation of everyday activities has the potential to not only foster our understanding of the cognitive processes involved, but also of their interaction and integration.

Gaining a deeper understanding is also of applied relevance. Given the demographic change and an aging society, the number of people unable to perform independently all necessary everyday activities is increasing (e.g., Nicholas & Smith, 2006). A deeper understanding of what drives successful everyday activities, how the underlying mechanisms develop, and how and what in the process may break down with age and cognitive impairment (dementia) can help support those who have trouble with everyday activities in two ways. First, with knowledge about which abilities may decline with age and impairment, specific training regimes can be developed to counter the decline in ability (e.g., Bettcher et al., 2011). Second, support could be given by artificial cognitive agents (e.g., robots) performing or prompting those activities that people are less able to do themselves. Currently available (household) robots are missing the flexibility and versatility to stand in for a human housekeeper (Ersen, Oztop, & Sariel, 2017), and a deeper understanding of the mechanisms that underlie learning and mastery of everyday activities may therefore inform the design of improved artificial agents.

This workshop will assemble six speakers with multidisciplinary backgrounds to discuss (a) the cognitive abilities un-

derlying everyday activities, (b) how these abilities develop ontogenetically, (c) how abilities may break down with cognitive impairment, (d) possible integration of different abilities in the scope of everyday activities, and (e) how insights from (a)-(d) could inform building artificial cognitive agents mastering everyday activities.

Speakers

Speakers have been selected to cover important areas that are relevant to the issues raised in the preceding section. Our speakers combine expertise in abilities involved in everyday activities, how they develop (Kaichi Yanaoka, Satoru Saito), how they may decline with cognitive impairment (Tania Giovannetti), how they may be formalized and integrated in computational models (Falk Lieder, Gregor Schöner, John Laird), and how cognitive principles may be transferred to artificial cognitive agents (John Laird, Gregor Schöner). Talks will address the following topics:

Falk Lieder, MPI Tübingen will present work on *discovering rational planning strategies*. To succeed in everyday life people have to quickly solve complex sequential decision problems with bounded cognitive resources. Lieder and colleagues' resource-rational analysis suggested that people's planning strategies are jointly shaped by these adaptive pressures and the structure of the environment. Lieder will present an automatic method that leverages this principle to predict which planning strategy people are going to use in a given environment and test it in a series of experiments.

Gregor Schöner, Ruhr-Universität Bochum will present *how neural dynamic architectures generate physical and mental acts*. Acting in the real world involves the coordination of perception, cognitive processes, and movement generation. Schöner will discuss how the balance between stability and flexibility that is necessary for successful coordination can be achieved in a framework of neural dynamics.

Kaichi Yanaoka & Satoru Saito, Kyoto University will present work on *the role of executive functions in routine sequential actions in young children*. They will provide an overview of research on executive functions and action control from a developmental perspective before presenting new data on learning and control of routine sequential actions in young children.

John Laird, University of Michigan will present *a cognitive architecture approach to everyday activities*. Laird will explore how the myriad of cognitive capabilities required to perform everyday activities can be supported by an integrated cognitive architecture, drawing examples from research with the Soar architecture. One capability Laird will focus on is Interactive Task Learning — how the cognitive architecture approach can support learning new tasks from natural instruction.

Tania Giovannetti, Temple University will present work on *everyday action in cognitive aging, mild cognitive impairment, and dementia*. Giovannetti will provide an overview of how deterioration of older adults' performance of everyday tasks is related to level and type of cognitive impairment. In doing so, she will also highlight the implications observed difficulties have for understanding the cognitive mechanisms that are required for accurate performance of everyday activities in healthy populations.

Schedule

The workshop is planned as a half-day event. Speakers will be allotted 25 minutes each for their presentations (20 minutes talk + 5 minutes discussion). The workshop will begin with a brief introduction by the organizers followed by the first three talks (Lieder, Schöner, Yanaoka & Saito). After the break, the two remaining talks (Laird, Giovannetti) will be delivered. The organizers will then lead a discussion of all presentations. The workshop will be concluded with a 30 min. poster session. Posters will be solicited by a Call for Posters with rolling acceptance. Poster presenters will be asked to put up their posters before the workshop to allow attendees to begin discussing them during the break.

Acknowledgments

We gratefully acknowledge support by the German Research Foundation (DFG) through the project P3 "Spatial Reasoning in Everyday Activity" as part of the Collaborative Research Center (Sonderforschungsbereich) 1320 "EASE - Everyday Activity Science and Engineering", University of Bremen (http://www.ease-crc.org/).

References

Bettcher, B. M., Giovannetti, T., Libon, D. J., Eppig, J., Wambach, D., & Klobusicky, E. (2011). Improving everyday error detection, one picture at a time: A performance-based study of everyday task training. *Neuropsychology*, *25*(6), 771–783. doi: 10.1037/a0024107

Ersen, M., Oztop, E., & Sariel, S. (2017). Cognition-enabled robot manipulation in human environments: Requirements, recent work, and open problems. *IEEE Robotics Automation Magazine*, *24*(3), 108-122. doi: 10.1109/MRA.2016.2616538

Gold, D. A., Park, N. W., Troyer, A. K., & Murphy, K. J. (2015). Compromised naturalistic action performance in amnestic mild cognitive impairment. *Neuropsychology*, *29*(2), 320–333. doi: 10.1037/neu0000132

Newell, A. (1973). You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. In W. G. Chase (Ed.), *Visual information processing.* New York: Academic Press.

Nicholas, P. K., & Smith, M. F. (2006). Demographic challenges and health in Germany. *Population Research and Policy Review*, *25*(5/6), 479–487.

Norman, D. A. (1981). Categorization of action slips. *Psychological Review*, *88*(1), 1.

Beyond the Ivory Tower: Non-Academic Career Paths for Cognitive Scientists

Vanessa R. Simmering (vanessa.simmering@act.org)
ACTNext by ACT, Inc., 500 ACT Drive
Iowa City, IA 52243 USA

Carissa L. Shafto (carissa.shafto@gmail.com)
Brightfield Strategies, LLC, 575 8th Ave
New York, NY 10018 USA

Keywords: professional development; non-academic; careers; industry; non-profit; government

Objectives and Scope

Cognitive science research has far-reaching implications, but many graduate students are trained with only an academic career in mind. Academic training develops a wide range of skills in service of behavioral research, literature reviewing, data analysis, scientific publishing, grant writing, teaching, and student mentorship. These skills also have direct application in non-academic positions, but training within academia typically neglects to address how these skills translate to other work environments and career paths. As growth in the number of doctoral trainees continues to outpace permanent academic positions (Kolata, 2016; Larson, Ghaffarzadegan, & Xue, 2013; Lederman, 2016), more doctoral recipients have been seeking non-academic employment (National Science Board, 2018). Doctoral students and recipients who are interested in exploring non-academic employment options may not know where to turn for guidance. Our goal in this professional development workshop is to offer such guidance and an opportunity to network with scholars in similar situations.

The session will be led by two scholars with doctoral degrees in psychology who worked in academic positions previous to their industry careers: Carissa Shafto is a senior data scientist and data governance specialist for Brightfield Strategies; Vanessa Simmering is a senior research scientist for ACTNext by ACT, Inc. They will draw on their individual experiences navigating from academic to non-academic positions to guide the activities and discussion. Additionally, they will solicit contributions and participation from other scholars with a diverse range of backgrounds and positions to increase the breadth of experiences participants consider.

Workshop Schedule

The time for the half-day session will be divided approximately in thirds, beginning with a presentation by the leaders, followed by a set of interactive activities among participants, and closing with group discussion of the activities and questions they raised. Because the workshop will occur before the conference, we hope that participants can use this opportunity to connect with each other and continue the conversations and networking beyond the end of the workshop.

Part 1: Introduction of Contributing Scholars and Different Career Paths

The leaders will begin with an overview of the goals of the session, followed by a series of narrated slides in which scholars (the leaders plus additional contributors) describe their backgrounds and employment. Specifically, we will ask all contributors to list the discipline of their degree and the general area of their research training, followed by (when relevant) any academic and non-academic positions they held before their current position, then a description of their current job, ending with a comment on what motivated them to seek out a non-academic career. Each contributor's description will be brief (3 minutes or less) and compiled into a single presentation in advance to maximize the number of examples we can present to participants. We have agreements to contribute narrated slides from thirteen participants thus far, listed in Table 1, and will invite more contributors if needed to ensure diverse representation of participants' backgrounds, interests, and employment types. Contributors will be encouraged to attend if possible, but attendance will not be required as this may limit which types of people and careers that can be represented, since many non-academic careers do not require or fund conference travel.

Part 2: Developing Your Pitch

Participants will be given time to work individually and then in small groups on two related activities developing "elevator pitches", which are brief but persuasive speeches designed to spark the listener's interest to learn more. The first pitch will be focused on what the participant is looking for in a career. The second will focus on what the participant has to offer to an employer. The leaders will scaffold this activity by highlighting successful strategies (e.g., focusing on skills over content, considering opportunities rather than obstacles) and potential individual considerations (e.g., whether one is leaving a temporary versus permanent position, whether relocation is possible). As relevant, these activities may include brainstorming a wide range of potential employment opportunities, or focusing on a specific position the participant already has in mind. During this portion, the leaders and any contributors in attendance will circulate through the room to talk to participants and answer questions that arise.

Table 1: PhD Scholars Contributing Narrated Slides on their Backgrounds and Careers

Name	Position	Institution / Company / Agency
Dan Acheson	Data Science Manager	Uptake
Keith Apfelbaum	Research Director	Foundations in Learning, Inc.
Aimee Arnoldussen	Medical Technology Assessment	University of Wisconsin Hospital & Clinics
Megan C. Brown	Lead Decision Scientist	Consumer and Partner Insights, Starbucks
John Lipinski	Director of Client Management	Certilytics
April Murphy	Data Scientist	Tulco Labs
Maggie Renno	Research Analyst	Wisconsin Department of Children and Families
Alexa Romberg	Research Manager	Schroeder Institute at Truth Initiative
Sarah Sahni	Associate in Social & Economic Policy	Abt Associates
Matthew Schlesinger	Senior Data Scientist	ReThink Medical
Sean Taylor	Research Scientist Manager	Core Statistics Team, Facebook
Dan Vatterott	Data Scientist	Showtime
Tim Wifall	Senior User Experience Researcher	Samsung Research America

Part 3: Questions, Feedback, Discussion, Networking, and Resources

Following the activity, participants will have an opportunity to ask questions, seek feedback, and discuss concerns within the larger group. The leaders will structure the time of the final third of the session based on interest from participants, including references to resources participants may want to use as they pursue non-academic careers. For example, a number of consulting services can be found online (The Professor Is In, Cheeky Scientist, Beyond the Professoriate, Next Scientist) but each varies slightly in their scope (i.e., some cater more to "hard" sciences, others to social sciences and humanities) and therefore their potential utility for participants with different backgrounds. They also vary in the amount of information offered free of charge and services provided at a cost. Social media sites (e.g., Post-Academic Athenas, Facebook groups, and LinkedIn) also offer more informal support, through discussion and peer mentoring, and can help participants expand their networks. The leaders and contributors will be able to provide some specific experiences to help participants evaluate what approaches could be of most use to them.

At the conclusion of the event, Dr. Simmering will survey interest from participants in potentially forming a group on LinkedIn or another platform to stay connected and follow up on conversations started during the session. Participants will also be provided with contact information from any contributors who agree to offer this opportunity to connect. We hope the session will give participants an entry point to exploring a wide range of career options and the necessary resources to pursue non-academic career paths.

Acknowledgments

Thanks are owed to Alexa Romberg for help in organizing a variant of this workshop submitted to another conference, and to Shevaun Lewis for permission to adapt and use materials she developed for a professional development workshop at the University of Maryland.

References and Resources

Kolata, G. (2016, July 14). So many research scientists, so few openings as professors. *The New York Times*. Retrieved from https://www.nytimes.com/2016/07/14/upshot/so-many-research-scientists-so-few-openings-as-professors.html

Larson, R. C., Ghaffarzadegan, N., & Xue, Y. (2014). Too many PhD graduates or too few academic job openings: the basic reproductive number R_0 in academia. *Systems Research and Behavioral Science*, 31(6), 745-750.

Lederman, D. (2016, December 9). The new Ph.D.s. *Inside Higher Education.* Retrieved from https://www.insidehighered.com/news/2016/12/09/phd-recipients-increase-number-job-prospects-vary-new-us-data-show

National Science Board. (2018). Academic Research and Development. In *Science & Engineering Indicators 2018*. National Science Foundation: Arlington, VA. Retrieved from https://nsf.gov/statistics/2018/nsb20181/report/sections/academic-research-and-development/doctoral-scientists-and-engineers-in-academia

Beyond the Professoriate Career Advice. Retrieved from https://beyondprof.com/

Cheeky Scientist Association & Training. Retrieved from https://cheekyscientist.com/

Next Scientist. Retrieved from http://www.nextscientist.com/

Post-Academic Athenas. Retrieved from https://www.postacathenas.com/

The Professor Is In: Real-Ac Transition Services. Retrieved from https://theprofessorisin.com/post-ac-services/

Daylong data: Raw audio to transcript via automated & manual open-science tools

John Bunce (john.bunce@umanitoba.ca)
Department of Psychology, University of Manitoba
190 Dysart Road, Winnipeg, MB R3T 2N2 Canada

Elika Bergelson (elika.bergelson@duke.edu)
Department of Psychology and Neuroscience, Duke University
417 Chapel Drive, Campus Box 90086, Durham, NC 27708-0086

Anne Warlaumont (warlaumont@ucla.edu)
Department of Communication, University of California, Los Angeles
2225 Rolfe Hall, UCLA, Box 951538, Los Angeles, CA 90095

Marisa Casillas (marisa.casillas@mpi.nl)
Language Development Department, MPI for Psycholinguistics
Wundtlaan 1, 6525 XD, Nijmegen, The Netherlands

Abstract

Several of the central questions in language, social cognition, and developmental research focus on the roles of input, output, and interaction on learning and communication. While it has become easy to collect long-form recordings, getting useful data out of them is a more daunting task. Across four mini-sessions, this tutorial aims to address pre- and post-data collection concerns, and provide a hands-on introduction to manual and automated annotation techniques. Attendees will leave this tutorial with resources and concrete experience for collecting, annotating, and sharing/archiving naturalistic recordings, including specific open-science practices relevant for these data.
Keywords: daylong recordings; natural language; speech technology; automated annotation; open science

Introduction

The ability to record and efficiently analyze everyday talk from a variety of different populations is crucial for many topics in language science, including: variation in children's linguistic input, distributional patterns of language in adult speech, atypical speech patterns for medical diagnosis, and more (e.g., see Casillas and Cristia (under review) for a review). However, even the most basic facts about everyday speech experience have remained elusive given the technological constraints of capturing and analyzing daylong speech for large samples of participants. In the last two decades, the LENA™ system has emerged as a potential solution to this methodological gap (see Ganek and Eriks-Brophy (2018) for a review). However, due to its costs and proprietary, aging technology, LENA™'s usefulness is increasingly limited.

In this half-day tutorial we will describe a new approach for getting the most from daylong recordings; one that uses community-based norms to support researchers at every step, from ethics review and initial data collection to automated analysis, manual annotation, and data archival. The tools and databases we include are all open-source and oriented toward usability on new populations and new technical challenges—an ideal next step to enable researchers to tackle new scientific questions about everyday language use. These tools has developed out of the ACLEW project (`http:\sites.google.com/view/aclewdid/home`).

Tutorial aims

This tutorial is focused on facilitating current research using daylong recordings while also boosting the future development of *even better* tools for the collection, annotation, and analysis of daylong recordings.

Our first aim is to **lower the barrier to using daylong recordings for language research**. Many researchers who are interested in this method are held back from doing so because there is no clear cost- and time-efficient way to annotate the data. We hope to allay some of these concerns by introducing a set of tools and techniques participants can use to extract usable data from their recordings. We will provide a hands-on training session demonstrating how to use our ACLEW audio-processing pipeline (automated tools for exploring voice activity, utterance segmentation, speaker diarization, and speech rate estimation) and manual annotation framework suitable for cross-corpus comparison. All software is free, open-source, and multi-platform.

Our second aim is to **promote an open-science framework for natural language data**, with an eye toward improving access to shared data and comparative analysis. The daylong recording community is just getting off the ground (HomeBank; VanDam et al., 2016), and there is vast potential for scientific advancement if more researchers were to participate. To demonstrate the benefits of data sharing and re-use for daylong recordings, we will show how the use of unified tools and annotation templates can lead to new breakthroughs in comparing natural language environments across cultures. Our motivation is that the long-term non-commercial success of our toolkit depends on an active community of users. Active users contribute new training data, give feedback on quality, and make requests for new functionality. We therefore hope to convince researchers that these tools can meet their immediate analytic needs while also persuading them to invest in the community so that we can establish the mega corpora necessary for continued tool improvement.

Participants

This tutorial is intended for researchers at all levels of experience who are interested in the collection, analysis, curation, and computational modeling of natural language data. While the tutorial will be accessible to a general CogSci audience, we also hope to attract participants who are interested in daylong recordings but daunted by the prospect of collecting or processing them. We also encourage participation by researchers who have already invested in daylong recordings and are looking for new ways to utilize them. Indeed, as part of DARCLE we have a commitment and track record of supporting new investigators (`http://darcle.org/newInvestigators.html`).

Learning outcomes

After this tutorial, participants will be able to (1) assess the pros and cons of using naturalistic recordings for their research questions, (2) locate, use, and adapt our online, self-guided tutorials and templates for creating machine-friendly annotations, (3) download, install, run, and interpret the output provided by the (open source) audio-processing software, and (4) understand how to gain access to and use HomeBank, a repository for daylong audio recordings.

Tutorial structure

This half-day hands-on tutorial will introduce: issues surrounding daylong recording collection, a standardized manual annotation process, the use of automated annotation tools, and best practices for data archiving. This will be organized into four sessions (separated by 5-min breaks). Participants will work with sample media file to get hands-on experience in each session.

Session 1. Pre-data collection concerns (25 min) A brief introduction to the method, its costs and benefits, and what to consider before collecting data. Topics include: how to decide whether daylong recordings are suitable for the research question, considerations when applying for ethical approval, and off-the-shelf hardware and software options. We will relate these topics to individual research interests.

Session 2. Manual annotation (55 min) A 3-part interactive training session introducing participants to manual annotation in the machine-friendly template we have developed for ELAN (Casillas et al., 2017). Part 1 focuses on the basic setup of the annotation scheme. Part 2 focuses on the use and adaptability of the annotation conventions. Part 3 focuses on the annotator training standards and reliability estimation using the automated tools provided by ACLEW.

Session 3. Automated annotation (55 min) An interactive tour of the ACLEW automated tools package. Each tool will be introduced and demonstrated with example media files. We will also take this opportunity to demonstrate the value of adding new training and testing data and will open the floor to discussion about future tool development.

Session 4. Archiving and community (25 min) A brief discussion focused on the issues surrounding the long-term storage of daylong recordings. We will also discuss efficient and accessible ways to share data, annotations, and analysis, and review the benefits of open-science practices.

Learning materials

Participants will need an Internet-connected laptop and a pair of headphones. The organizers will create an OSF page with links to all training materials and instructions for future use. Although sample data will be provided, participants are encouraged to bring their own data to demonstrate the challenges of different research questions using daylong audio.

Tutor credentials

The materials and instruction for this tutorial will come from the cognitive scientists and software developers who created the tools being covered. Collectively, they have expertise in training dozens of researchers (undergraduate to PhD) on the steps covered in sessions 1–4. That said, this tutorial will be the very first to cover the end-to-end use of this pipeline for researchers working on daylong audio recordings.

Summary of significance

The study of everyday talk is fundamental for understanding the relationship between cognition, culture, and language. Recent technological advancements afford researchers the ability to study everyday language on a much larger scale than before, but these technologies are challenging and therefore remain somewhat underutilized. We aim to further the use and usefulness of this technology by spreading knowledge of how to effectively employ it and by facilitating the continued improvement of the associated tools for language science.

Acknowledgments

This work was supported by a TransAtlantic Platform "Digging into Data" collaboration grant (ACLEW: Analyzing Child Language Experiences Around the World) and an NWO Veni Innovational Research Scheme (275-89-033) to MC.

References

Casillas, M., Bunce, J., Soderstrom, M., Rosemberg, C., Migdalek, M., Alam, F., ... Garrison, H. (2017). *Introduction: The ACLEW DAS template [training materials].* Retrieved from `https://osf.io/aknjv/`

Casillas, M., & Cristia, A. (under review). A step-by-step guide to collecting and analyzing long-format speech environment (LFSE) recordings. *XX*, *XX*, XX–XX.

Ganek, H., & Eriks-Brophy, A. (2018). Language ENvironment analysis (LENA) system investigation of day long recordings in children: A literature review. *Journal of Communication Disorders*, *72*, 77–85.

VanDam, M., Warlaumont, A. S., Bergelson, E., Cristia, A., De Palma, P., & MacWhinney, B. (2016). *HomeBank: An online repository of daylong child-centered audio recordings.*

EMHMM: Eye Movement Analysis with Hidden Markov Models and Its Applications in Cognitive Research

Janet H. Hsiao (jhsiao@hku.hk)
Department of Psychology, University of Hong Kong
Pokfulam Road, Hong Kong, Hong Kong

Antoni B. Chan (abchan@cityu.edu.hk)
Department of Computer Science, City University of Hong Kong
Tat Chee Avenue, Kowloon, Hong Kong

Keywords: EMHMM; eye movement; hidden Markov model

Significance of the Method

In many daily life activities, eye movements provide strong clues about underlying cognitive processes. For example, patients with cognitive deficits have atypical eye movement patterns. Users with different experiences show different eye movement behavior in viewing websites. Thus, eye movement has become an important measure in the broad research fields in cognitive science.

Recent research has reported substantial individual differences in eye movements during cognitive tasks. Nevertheless, most of the current analysis methods do not adequately reflect these individual differences. Also, they focus on spatial information (fixation locations), whereas temporal information (transitions among fixation locations) is typically overlooked. The most common method has been the use of predefined regions of interests (ROIs) on the stimuli. However, predefined ROIs are often subject to experimenter bias and inconsistency across studies. To address these problems, Caldara and Miellet (2011) proposed to directly perform by-pixel statistical tests on fixation heat maps (where fixations are smoothed with a Gaussian function) to determine the regions with significant difference between conditions. Nevertheless, these regions are often irregularly shaped and difficult to interpret. Also, fixation maps at different times only show the transition of overall fixation distribution and do not provide information about transitions between regions. Another method (Jack et al., 2009) is to define ROIs as regions formed by running the k-means clustering algorithm on significantly fixated regions of a fixation map. However, this approach assumes that all ROIs are circular and the same size, and the number of ROIs must be preset by the experimenter.

Thus, we have developed a novel eye movement data analysis method, Eye Movement analysis with Hidden Markov Models (EMHMM; Chuk, Chan, & Hsiao, 2014), which summarizes each individual's eye movement pattern using a hidden Markov model (HMM; a type of machine learning model for time series data), including person-specific ROIs and transition probabilities among the ROIs. Individual HMMs can be clustered according to similarities to discover common patterns (Fig. 1a), and the similarity between an individual pattern and a common pattern can be quantitatively assessed through estimating the likelihood of the individual's data being generated by the common pattern HMM. This similarity measure then can be used to examine associations between eye movement patterns and other cognitive measures (Fig. 1b & 1c). We have applied this method to face recognition research and made discoveries thus far not revealed by other methods, including how eye movements are associated with recognition performance, cognitive abilities (Chan, Chan, Lee, & Hsiao, 2018), cultural differences (Chuk, Crookes, et al., 2017), memory encoding/retrieval (Chuk, Chan, & Hsiao, 2017), sleep loss (Zhang, Chan, Lau, & Hsiao, 2019), and activations in brain regions important for top-down attention control (Chan et al., 2016). We have also recently developed new methodologies for more complex cognitive tasks, including using switching HMMs for tasks involving cognitive state changes (Chuk, Chan, Shimojo, & Hsiao, 2016), and using the machine learning algorithm co-clustering for tasks involving stimuli with different feature layouts (Hsiao, Chan, Du, & Chan, 2019).

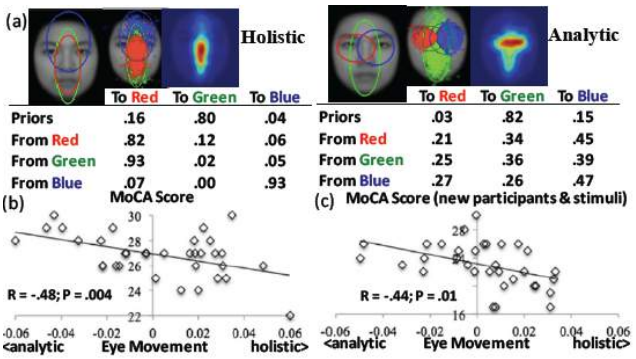


Fig. 1: (a) Analytic and holistic patterns in face recognition (Chan et al., 2018). Ellipses show ROIs as 2-D Gaussian emissions. The table shows transition probabilities among the ROIs. Priors show the probabilities that a fixation sequence starts from the ellipse. (b) In older adults, the more holistic the pattern, the lower the cognitive status (by MoCA), and (c) this correlation was replicated with new participants viewing new face images using the representative HMMs in (b).

In short, the EMHMM methodology will allow us to summarize, quantitatively assess, and compare individual eye movement patterns across stimuli and tasks, and examine how they are associated with other cognitive measures. It will lead to innovative findings not revealed by any existing methods with a lasting impact on how eye tracking is used for understanding cognition across disciplines. The Matlab Toolbox for EMHMM is available at http://visal.cs.cityu.edu.hk/research/emhmm/.

Structure and Activities

This half-day tutorial consists of 2 sessions:

1. Introduction to EMHMM and Its Applications: We will first introduce current methods in eye movement data analysis to illustrate the advantages of the EMHMM method. We will then introduce how we can apply EMHMM to research on face recognition, reading, cultural difference, ageing, sleep, information systems, decision making, scene perception, and video viewing. In the end we will provide a short demo in which attendees can come to perform a face recognition task with eye tracking, and get a personalized EMHMM report on site.

2. Tutorial and Hands-on Experience: We will first present an EMHMM simulation study (Chan & Hsiao, 2018) and provide recommendations for using EMHMM in cognitive research. We will then provide an EMHMM Matlab Toolbox tutorial with sample data for attendees to practice using the toolbox on their own laptops. We will have at least one laptop available onsite for attendees who do not have access to Matlab. Attendees may also bring their own data and ask questions on site.

Credentials of the Tutorial Organizers

The tutorial organizers have been developing the EMHMM method for 7 years. Since the first paper/talk presented at the Annual Meeting of the Cognitive Science Society (Chuk, Chan, & Hsiao, 2013), they have published 6 journal papers (including *Cognition* and *Sleep*) and 23 conference/invited presentations (including VSS and ICIS) using this method with collaborators from the UK, the US, Germany, and Australia, etc. on various topics.

Janet Hsiao is a world-leading expert in using eye tracking and computational modeling methods to understand human cognition. She has published in several high-profile cognitive science journals including *Psychological Science* and *Cognition*. She is currently an Associate Editor for *Cognitive Science*, and has been served on the Program Committee for the annual meetings of the Cognitive Science Society since 2016.

Antoni Chan is a world-leading expert in probabilistic models for time series data analysis and pattern recognition. He has published in several high-profile machine learning and computer vision journals, including *IEEE Trans. on Pattern Analysis and Machine Intelligence* and the *Journal of Machine Learning Research*. He is currently a Senior Area Editor for *IEEE Signal Processing Letters*, and served as an Area Chair for ICCV'15, '17, and '19.

References

Caldara, R. & Miellet, S. (2011). iMap: a novel method for statistical fixation mapping of eye movement data. *Behav. Res. Methods*, *43*, 864-878.

Chan, A. B., & Hsiao, J. H. (2018). EMHMM Simulation Study. http://arxiv.org/abs/1810.07435

Chan, C. Y. H., Chan, A. B., Lee, T. M. C., & Hsiao, J. H. (2018). Eye movement patterns in face recognition are associated with cognitive decline in older adults. *Psychonomic Bulletin & Review*, *25*(6), 2200-2207.

Chan, C. Y. H., Wong, J. J., Chan, A. B., Lee, T. M. C., & Hsiao, J. H. (2016). Analytic eye movement patterns in face recognition are associated with better performance and more top-down control of visual attention: an fMRI study. *Proceeding of the 38th Annual Conference of the Cognitive Science Society* (pp. 854-859).

Chuk, T., Chan, A. B., & Hsiao, J. H. (2014). Understanding eye movements in face recognition using hidden Markov models. *J. Vis.*, *14*(11):8, 1-14.

Chuk, T., Chan, A. B., & Hsiao, J. H. (2017). Is having similar eye movement patterns during face learning and recognition beneficial for recognition performance? Evidence from hidden Markov modeling. *Vision Research*, *141*, 204-216

Chuk, T., Chan, A. B., Shimojo, S., & Hsiao, J. H. (2016). Mind reading: Discovering individual preferences from eye movements using switch hidden Markov models. *Proceeding of the 38th Annual Conference of the Cognitive Science Society* (pp. 182-187).

Chuk, T., Crookes, K., Hayward, W. G., Chan, A. B., & Hsiao, J. H. (2017). Hidden Markov model analysis reveals the advantage of analytic eye movement patterns in face recognition across cultures. *Cognition*, *169*, 120-117.

Coutrot, A., Hsiao, J. H., & Chan, A. B. (2018). Scanpath modeling and classification with Hidden Markov Models. *Behavior Research Methods*, *50*(1), 362-379

Hsiao, J. H., Chan, K. Y., Du, Y. & Chan, A. B. (2019). Understanding individual differences in eye movement pattern during scene perception through hidden Markov modeling. *Proceeding of the 41th Annual Conference of the Cognitive Science Society*

Jack, R. E., Blais, C., Scheepers, C., Schyns, P. G., & Caldara, R. (2009). Cultural confusions show that facial expressions are not universal. *Curr. Biol.*, *19*, 1-6.

Zhang, J., Chan, A. B., Lau, E. Y. Y., & Hsiao, J. H. (2019). Individuals with insomnia misrecognize angry faces as fearful faces while missing the eyes: An eye-tracking study. *Sleep*, *42*(2), zsy220.

Optimizing the Design of an Experiment using the ADOpy Package: An Introduction and Tutorial

Jay I. Myung (Myung.1@osu.edu)
Mark A. Pitt (Pitt.2@osu.edu)
Department of Psychology, Ohio State University
Columbus, OH 43210 USA

Jaeyeong Yang (urisa12@snu.ac.kr)
Woo-Young Ahn (wahn55@snu.ac.kr)
Department of Psychology, Seoul National University
Seoul, 08826 KOREA

Keywords: computational cognition; Bayesian active learning; autonomous experimentation; adaptive design optimization; Python software package

Introduction

Experimentation is one of the cores of cognitive science, whether one is interested in understanding the mechanisms underlying cognitive control or the neural basis of decision-making. Through accurate measurement in a well-thought-out experimental design, the goal is to obtain sufficiently noise-free data to make inferences about processing. The design of an experiment can be especially tricky, requiring consideration of many factors (e.g., what levels and how many levels of a variable should be presented, how many stimuli per level, etc.). The final design can sometimes result in only a subset of the design space (i.e., conditions) yielding interesting results, with the remaining data being minimally informative.

Advances in Bayesian statistics and machine learning offer algorithm-based ways to generate optimal and efficient experimental designs so as to minimize uninformative and wasted experimental trials (e.g., Cavagnaro, Myung, Pitt, & Kujala, 2010; Lesmes, Lu, Baek, & Dosher, 2010). In an optimized experiment, stimuli are selected adaptively and optimally (i.e., in an information theoretic sense; Lindley, 1956) on each trial by real-time data analysis of observed responses from earlier trials. What is being optimized is the values of the design variables that can be manipulated experimentally, such as the intensity of a stimulus in a psychophysics experiment or the monetary rewards and probability of occurrence in a preferential choice experiment. This is unlike a traditional experiment in which the design is fixed for all participants and stimulus presentation is either random or follows a predetermined schedule.

One such approach is referred to as Adaptive Design Optimization (ADO; Cavagnaro et al., 2010). ADO derives from optimal experimental design in statistics (Atkinson & Donev, 1992; Chaloner & Verdinelli, 1995) and active learning in machine learning (Cohn, Atlas, & Ladner, 1994; Settles, 2012). ADO is a general-purpose, algorithm-based method for autonomously conducting adaptive experiments that lead to rapid accumulation of information about the phenomenon of interest with the fewest number of trials. ADO can improve significantly the informativeness and efficiency of data collection (e.g., Cavagnaro et al., 2011 & 2016).

ADOpy

Expertise in statistics and computational modeling is required to use these machine-learning methods. To improve their accessibility to a wide range of researchers, we have developed an open-source Python package. The package, dubbed ADOpy, implements ADO for optimizing experimental designs. ADOpy is currently available on GitHub (https://github.com/adopy), with three pre-installed adaptive experimental tasks as of January 2019: (a) the slope and threshold estimation of the psychometric function (Kontsevich & Tyler, 1999); (b) the delay discounting experiment (Cavagnaro et al., 2016); and (c) the choice under risk and ambiguity experiment (Levy et al., 2010).

ADOpy is written using high-level semantic-based commands in a such way that the whole ADO procedure is broken into a set of meaningful function calls that can be easily edited and modified by users. Further and importantly, the package is user-friendly in that users can use the package without having to understand the computational details of the ADO algorithm. Additionally, the package is modular so that new models and/or experimental tasks can be easily added. Thus, only a modest amount of programming and modeling experience is required to use ADOpy.

The purpose of the proposed tutorial is to introduce ADOpy to cognitive scientists in a hands-on training environment, first providing a conceptual introduction to optimal experimental design and then walking through examples that demonstrate how to use methodology. The tutorial will be based on a manuscript (in preparation) to be submitted for publication in the near future.

Tutorial Format

This half-day tutorial will be organized into two 1.5-hour sessions with a 30-min coffee break between them. The first part, given by the first two authors, will consist of a general

overview of the conceptual and statistical foundations of ADO (1 hour) and then 30 minutes to answer questions and set up for the tutorial session. After the break, the second 1.5 hours will be a tutorial on the ADOpy package, with hands-on training using concrete, work-through examples, run jointly by the third and fourth authors.

There will be a website with a program, a web link to the GitHub site, the abstracts and slides of all presentations, supplementary Python code to be used in the hands-on session and recommended readings.

Target Audience

Graduate students, postdoctoral researchers, and scientists, who are new to ADO and have workable knowledge of Bayesian statistics on a graduate level and also of basic Python programming.

Organizers/Presenters

Jay I. Myung is Professor of Psychology at the Ohio State University. He received a PhD in 1990 in psychology at Purdue University. His research interests in the fields of cognitive and mathematical psychology include computational cognition, optimal experimental design, Bayesian modeling, and model comparison. Homepage: https://faculty.psy.ohio-state.edu/myung/personal/

Mark A. Pitt is Professor of Psychology at the Ohio State University. He received his PhD in 1989 in psychology at Yale University. In addition to researching computational approaches to improving inference in experimentation, he researches questions in psycholinguistics, such as how listeners recognize spoken words. Homepage: http://lpl.psy.ohio-state.edu/.

Jaeyeong Yang is a second-year graduate student of psychology in the Department of Psychology at Seoul National University. He received a double major B.S. in psychology and computer science, and he wrote the *ADOpy* package in Python.

Woo-Young Ahn is Assistant Professor of Psychology at Seoul National University. He received a PhD in 2012 in clinical psychology at Indiana University and has published over 20 papers in journals such as *Cognitive Science, Proceedings of the National Academy of Sciences, Current Opinion in Behavioral Sciences, Journal of Mathematical Psychology,* and *Computational Psychiatry*. His research interests include decision neuroscience and computational psychiatry, and he developed the Bayesian modeling package *hBayesDM* (https://github.com/CCS-Lab/hBayesDM).

Acknowledgments

This research is supported in part by National Institute of Health Grant R01-MH093838 to JIM and MAP, and also by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT, & Future Planning (2018R1C1B3007313) to WYA.

References

Atkinson, A., & Donev, A. (1992). *Optimum Experimental Designs*. Oxford University Press.

Cavagnaro, D. R., Aranovich, G. J., McClure, S. M., Pitt, M. A., & Myung, J. I. (2016). On the functional form of temporal discounting: An optimized adaptive test. *Journal of Risk & Uncertainty, 52,* 233-254.

Cavagnaro, D. R., Myung, J. I., Pitt, M. A., & Kujala, J. (2010). Adaptive design optimization: A mutual information based approach to model discrimination in cognitive science. *Neural Computation, 22*, 887-905.

Cavagnaro, D. R., Pitt, M. A., & Myung, J. I. (2011). Model discrimination through adaptive experimentation. *Psychonomic Bulletin & Review, 18(1),* 204–210.

Chaloner, K., & Verdinelli, I. (1995). Bayesian experimental design: A review. *Statistical Science, 10(3)*, 273–304.

Cohn, D., Atlas, L., & Ladner, R. (1994). Improving generalization with active learning. *Machine Learning, 15(2*), 201–221.

Kontsevich, L. L., & Tyler, C. W. (1999). Bayesian adaptive estimation of psychometric slope and threshold. *Vision Research, 39,* 2729-2737.

Lesmes, L. A., Lu, Z.-L., Baek, J., & Dosher, B. A. (2010). Bayesian adaptive estimation of contrast sensitivity function: the quick CSF method. *Journal of Vision, 20*, 1–21.

Levy, I., Snell, J., Nelson, A. J., Rustichini, A., & Glimcher, P. W. (2010). Neural representation of subjective value under risk and ambiguity. *Journal of Neurophysiology, 103*, 1036-2047.

Lindley, D. V. (1956). On a measure of the information provided by an experiment. *Annals of Mathematical Statistics, 27(4)*, 986–1005.

Settles, B. (2012). Active learning. Synthesis Lectures on *Artificial Intelligence and Machine Learning, 6(1),* 1–114.

Full Day Tutorial on Quantum Theory in Cognitive Modeling

Emmanuel M. Pothos (Emmanuel.pothos.1@city.ac.uk) and James M. Yearsley (James.Yearsley@city.ac.uk)
Department of Psychology, City, University of London, London, EC1V 0HB, UK

Zheng (Joyce) Wang (wang.1243@osu.edu)
School of Communications, Center for Cognitive and Brain Sciences, Ohio State University, Columbus, OH 43210 USA

Peter D. Kvam (kvam.peter@gmail.com) and Jerome R. Busemeyer (jbusemey@indiana.edu)
Department of Psychological and Brain Sciences, Indiana University, Bloomington, IN 47405, USA

Keywords: quantum probability theory; classical/ Bayesian probability theory; Markov processes; contextuality; decision making; memory; similarity.

Introduction

Even though the generally acknowledged normative and descriptive standard for modeling human inference is classical/ Bayesian probability theory (CPT), there have also been several reports which challenge CPT's universal applicability. Some of the most influential empirical demonstrations of such so-called fallacies have been reported by Kahneman, Tversky and their collaborators. For example, consider the evocative conjunction fallacy. In the Tentori et al. (2004) demonstration of the conjunction fallacy, participants are quite happy to consider it more probable to randomly select a Scandinavian person with both blue eyes and blond hair, than just blond hair. Even though we can imagine a line-up of Scandinavian individuals (making the set theoretic structure of CPT explicit and so the impossibility of a conjunction fallacy), there just seems a persistent feeling that somehow the conjunction is more likely than the marginal (cf. Gilboa, 2000). How can our intuition be so much at odds with CPT prescription?

We call quantum probability theory (QPT) the rules for how to assign probabilities to events from quantum mechanics, without any of the physics. QPT is in principle applicable in any situation where there is a need to formalize uncertainty. In psychology, one way to motivate QPT is as a bounded rationality approach to CPT: whereas in CPT we require conjunctions/ disjunctions across all possible questions (and the underlying logical structure is a Boolean algebra), in QPT (classical) conjunctions/ disjunctions are possible only for so-called compatible questions, while for incompatible ones they are undefined (they have to be computed with sequential operations; the underlying logical structure is a partial Boolean algebra).

Where incompatible questions are concerned, QPT provides a radically different perspective on probabilistic inference, compared to CPT, characterized by, for example, interference effects, violations of the law of total probability, supercorrelations, and constructive influences from judgments. These characteristics have provided a rich modeling framework for accommodating behavioral results superficially at odds with classical structure, across several areas including decision making, memory, similarity, perception, and logical reasoning, to mention but a few (overviews in Bruza et al., 2015; Busemeyer & Bruza, 2012; Haven & Khrennikov, 2013; Pothos & Busemeyer, 2013).

The purpose of the tutorial is to provide a comprehensive introduction to the QPT techniques commonly employed in cognitive modeling and illustrate the breadth of cognitive findings for which successful QPT models have been proposed.

Presenters

Emmanuel Pothos is a Professor of Psychology at City, University of London. He has been involved with the quantum cognition research programme since its inception, more than 10 years go. James Yearsley is a mathematical psychologist, originally trained in quantum theory. He has provided one of the most compelling a priori behavioral predictions of QPT (Yearsley & Pothos, 2016). Zheng (Joyce) Wang is a Professor at The Ohio State University. She was Co-Editor for a special issue on quantum cognition that appeared in *Topics in Cognitive Science,* 2013, Vol. 5). Peter Kvam is a postdoctoral researcher at Indiana University, who has published many articles on quantum cognition including in top journals such as PNAS. Finally, Jerome Busemeyer is Distinguished Professor of Cognitive Science at Indiana University and fellow of the Cognitive Science Society. He is one of the instigators of the quantum cognition research programme.

Previous Tutorials and Symposia

The tutorial has been presented at the Cognitive Science meetings in Nashville (2007), Washington DC (2008), Amsterdam (2009), Sopporo (2012), Berlin (2013), Quebec City (2014), Pasadena (2015), Philadelphia (2016), and Madison (2018), with about 30 to 50 participants each time. The ratings from participants after the tutorial were all very positive. In 2017, we held a workshop on quantum cognition supported by the Estes Foundation to 60 participants at a joint meeting of

the Society for Mathematical Psychology and the International Conference on Cognitive Modeling at the University of Warwick, UK. Also, this tutorial follows a symposium on quantum cognition at the Cognitive Science meeting 2011, whose papers appeared as a special issue in *Topics in Cognitive Science* (2013).

Assumptions about Participants Background

Most of the techniques we will cover involve elementary linear algebra and should be accessible to participants with minimal mathematical background. Note, no knowledge of physics is required and, with the exception of providing some historical context, no references to physics will be made.

Material to be Covered

We intend to organize the tutorial in three sessions, but with multiple speakers per session and short breaks, to make presentations more engaging for the audience. We note below how each session will be broken up into parts, with an approximate indication of time per part.

Introduction and background (2 hours)

Why employ QPT in cognitive modeling? Busemeyer will provide a brief introduction to the tutorial (0.25 hours). We will then consider a simple QPT model for the conjunction fallacy, explaining how the representations can be set up, how are probabilities computed, and how the interference term necessary to accommodate the conjunction fallacy emerges. We will also discuss the way the QPT prediction of a CF can be interpreted in rational terms (Pothos, 1 hour). We will then provide an overview of empirical findings which have been modeled with QPT, with a focus on other decision findings (e.g., disjunction effect; disjunction fallacy), questionnaire response biases (e.g., order effects), memory (e.g., the overdistribution effect), similarity, and perception (e.g., violations of the law of total probability; Wang, 0.75 hours).

Dynamical models; advanced techniques (2 hours)

We will discuss how dynamical cognitive processes can be modeled with QPT and introduce related technical concepts, e.g., unitary operators and Hamiltonians, side by side with classical counterparts, in the context of well-known empirical results from decision making (Busemeyer, 0.75 hours). We will then introduce some more advanced QPT methods. Notably QPT includes a sophisticated formalism for noise in probabilistic inference (with the formalism of POVMs), that is relevant in psychological processes where noise is assumed to play a substantial role. Additionally, the standard dynamical formalism in QPT can be extended to situations where there is an interaction (information exchange) with the environment (cf. open system dynamics; Yearsley, 0.75 hours). Finally, we will consider Bayesian model comparisons between QPT and matched CPT models and discuss their relative complexity in general terms and in relation to specific examples (Yearsley & Kvam, 0.5 hours).

Generative value (2 hours)

We will consider the generative value of the quantum cognition research programme, with emphasis on explaining the techniques and allowing insight into the thought process leading to model creation. Kvam (1 hour) will present a research programme on modeling heuristics within QPT. In particular, he will demonstrate how several fast and frugal heuristics can be reconstructed by integrating them with a quantum logic structure, introducing qubits, U-gates, and quantum information theory more generally. He will consider several applications including regarding expertise, game theory, and the hindsight bias. Wang (0.75 hours hours) will present one of the most surprising and robust predictions from QPT, the so-called QQ equality, which is a parameter free constraint on how order effects in question pairs ought to add up to zero (Wang et al., 2014). Yearsley (0.25 hours) will discuss the prediction of the Quantum Zeno effect, that the density of intermediate judgments slows down opinion change; this prediction relates to one of the most distinctive properties of QPT, the collapse postulate, which entails state changes from measurements. Pothos (0.25 hours) will illustrate this in a simpler paradigm, leading to a prediction of a novel decision bias. And finally, Busemeyer (0.5 hours) will outline the future directions of the quantum cognition research programme.

Acknowledgments

EMP was supported by ONRG grant N62909-19-1-2000.

References

Bruza, P., Wang, Z., & Busemeyer, J. R. (2015). Quantum cognition: a new theoretical approach to psychology. *Trends in Cognitive Sciences*, 19, 383-393.

Busemeyer, J. R., & Bruza, P. D. (2012). *Quantum models of cognition and decision*. Cambridge, UK: Cambridge University Press.

Gilboa, I. (2000). *Theory of decision under uncertainty*. Cambridge University Press: Cambridge, UK.

Haven, E. and Khrennikov, A. (2013). *Quantum Social Science*. Cambridge University Press: Cambridge, UK.

Pothos, E. M., & Busemeyer, J. R. (2013). Can quantum probability provide a new direction for cognitive modeling? *Behavioral & Brain Sciences*, 36, 255-274.

Tentori, K., Bonini, N., & Osherson, D. (2004). The conjunction fallacy: a misunderstanding about conjunction? *Cognitive Science*, 28, 467-477.

Wang, Z., Solloway, T., Shiffrin, R. M., & Busemeyer, J. R. (2014). Context effects produced by question orders reveal quantum nature of human judgments. *PNAS*, 111, 9431-9436.

Yearsley, J. M. & Pothos, E. M. (2016). Zeno's paradox in decision making. *Proceedings of the Royal Society B*, 283, 20160291.

Individual Differences in Spatial Representations and Wayfinding

Navigation is a well-specified computational problem, and solving it is vital for survival. Given these constraints, we might expect that humans differ minimally in their wayfinding capabilities. Indeed, a lack of variation is often implicitly assumed when cognitive scientists debate the existence of cognitive maps or when cognitive neuroscientists search for the neural substrates of navigation. However, in everyday life, we frequently discuss how some people get lost with some frequency, or how women ask for directions while men use maps. Indeed, it is increasingly apparent in the scientific data on navigation (and other cognitive domains) that the study of normative functioning needs to be integrated with the study of human variation, with its attendant challenges regarding experimental design and use of psychometrics. The four papers in this symposium gather together current work in cognitive science and neuroscience that aim to integrate the study of variation into the more common normative approach.

Mechanisms of Differences in Cognitive Mapping and Navigational Ability: Explorations Using Virtual Reality Manipulations

Thackery I. Brown[1], Qiliang He[1], Timothy P. McNamara[2], Jon Starnes[1], Sarah Goodroe[1]
[1]Georgia Institute of Technology [2]Vanderbilt University

Daily function depends on an ability to mentally map our environment. Environmental visibility and complexity can increase this challenge. Importantly, people vary dramatically in their ability to navigate flexibly and overcome such environmental challenges. In this paper, we will present experimental work targeting the mechanisms that underlie different navigational abilities, and how objective and introspective measures of ability interact to influence navigational strategy use. Using virtual reality, we manipulated environmental visibility and complexity. Participants then performed wayfinding, pointing, and route following tasks to probe cognitive map memory and navigational flexibility. Our findings reveal that individual differences in metacognition - such as perceived sense of direction - and in navigational strategy preference powerfully impact how environmental features affect spatial memory. We also gathered data on the neurocognitive foundations of these differences. Importantly, our methods highlight individualized interventions that can improve spatial learning and specify the mechanisms through which they operate.

A Meta-analysis of Sex Differences in Human Navigation Skills

Alina Nazareth[1], Lucy Huang[2], Nora S. Newcombe[1], Daniel Voyer[2]
[1] Temple University [2]University of New Brunswick

Popular sources often assume the existence of a male advantage in navigation, but the scientific data are inconsistent. This meta-analysis evaluates the literature on behavioral sex differences in human navigation. We quantify the overall magnitude of sex differences in a variety of paradigms and populations and examine potential moderators in large-scale navigation skills, using 694 effect sizes from 266 studies and a multilevel linear modeling approach. Overall, we found that male participants outperform female participants, with a small to medium effect size (d= 0.34 to 0.38). The type of task, the type of dependent variable and the testing environment significantly contribute to variability in effect sizes. Pointing and recall tasks show larger sex differences than distance estimation tasks or learning to criterion; among the dependent

variables, the deviation scores associated with pointing tasks show larger effect sizes. The largest estimate was d = .55 for tasks than required coordinating indoor and outdoor views. Interestingly, studies with children younger than 13 years showed very small effect sizes (d = .15) as compared to older age groups. We discuss the implications of these findings for the study of sex differences and identify avenues for future navigation research.

Measuring Spatial Perspective Taking:
Analysis of Four Measures using Item Response Theory

Maria Brucato[1], Andrea Frick[2], Alina Nazareth[1], Nora S. Newcombe[1]
[1] Temple University [2] University of Fribourg

Research on spatial thinking needs reliable and valid measures of individual differences in skills. Visuospatial Perspective Taking (PT)—the ability to mentally maintain and transform spatial relationships between objects within an environment—is one kind of spatial skill that is especially relevant to navigation and building cognitive maps. However, the psychometric properties of various PT tasks have yet to be examined. The present study examines three main psychometric properties of PT tasks: 1) the reliability of two tasks developed for children but adapted in difficulty level for use in adult populations, 2) item difficulty and discriminability within and between four tasks using item response theory, and 3) relation of scores with general intelligence, working memory, and mental rotation. Results showed that two of the four PT tasks have promising psychometric properties for measuring a wide range of PT ability based on item difficulty, discriminability, and efficiency of a test information function.

Genetics and Experience Modulate Individual Differences in Navigation

Veronique Bohbot
McGill University

Different memory systems, dependent on separate parts of the brain, can sustain successful navigation. The hippocampus is implicated in spatial memory strategies used when finding one's way in the environment, i.e. it is allocentric and involves remembering the relationship between landmarks. On the other hand, another strategy dependent on the caudate nucleus can also be used, i.e. the response strategy, which relies on making a series of stimulus-response associations (e.g. right and left turns from given positions). Participants who use the response strategy are faster at learning navigation tasks lending themselves to using a single specified route. Young adult response learners have increased fMRI activity and grey matter in the caudate nucleus, but decreased fMRI activity and grey matter in the hippocampus. Research in my laboratory has shown that specific navigation strategies are associated with several genes, such as BDNF and ApoE, as well as hormones, such as cortisol and progesterone, but not estrogen and progesterone. Experiences dependent modulators such as age, habit, stress and rewards also modulate strategies dependent on the hippocampus and caudate nucleus. These results have important translational implications because a larger hippocampus has been associated with healthy cognition in normal aging and with a reduced risk of numerous neurological and psychiatric disorders such as Alzheimer's disease, Schizophrenia, Post-Traumatic Stress disorder and Depression.

What makes a good explanation? Cognitive dimensions of explaining intelligent machines

Roberto Confalonieri, Tarek R. Besold
(name.surname@telefonica.com)
Alpha Health AI Lab
Telefónica Innovación Alpha

Tillman Weyde
(t.e.weyde@city.ac.uk)
Dept. of Computer Science
City, University of London

Kathleen Creel
(k.creel@pitt.edu)
Dept. of History and Philosophy of Science
University of Pittsburgh

Tania Lombrozo
(lombrozo@princeton.edu)
Dept. of Psychology
Princeton University

Shane Mueller
(shanem@mtu.edu)
Cognitive and Learning Sciences
Michigan Technological University

Patrick Shafto
(patrick.shafto@gmail.com)
Dept. of Math. and Computer Science
Rutgers University

Keywords: Explainability; Artificial Intelligence; Philosophy of Artificial Intelligence; Psychology; Cognitive Science

Explainability is assumed to be a key factor for the adoption of Artificial Intelligence systems in a wide range of contexts (Hoffman, Mueller, & Klein, 2017; Hoffman, Mueller, Klein, & Litman, 2018; Doran, Schulz, & Besold, 2017; Lipton, 2018; Miller, 2017; Lombrozo, 2016). The use of AI components in self-driving cars, medical diagnosis, or insurance and financial services has shown that when decisions are taken or suggested by automated systems it is essential for practical, social, and increasingly legal reasons that an explanation can be provided to users, developers or regulators.[1]Moreover, the reasons for equipping intelligent systems with explanation capabilities are not limited to user rights and acceptance. Explainability is also needed for designers and developers to enhance system robustness and enable diagnostics to prevent bias, unfairness and discrimination, as well as to increase trust by all users in *why* and *how* decisions are made. Against that background, increased efforts are directed towards studying and provisioning explainable intelligent systems, both in industry and academia, sparked by initiatives like the DARPA Explainable Artificial Intelligence Program (DARPA, 2016). In parallel, scientific conferences and workshops dedicated to explainability are now regularly organised, such as the 'ACM Conference on Fairness, Accountability, and Transparency (ACM FAT)' (Friedler & Wilson, n.d.) or the 'Workshop on Explainability in AI' at the 2017 and 2018 editions of the International Joint Conference on Artificial Intelligence.

However, one important question remains hitherto unanswered: *What are the criteria for a good explanation?*

Explainable Artificial Intelligence

While Explainable Artificial Intelligence (XAI) has recently received significant attention, its origins stem from several decades ago when AI systems were mainly developed as knowledge-based or expert systems, such as in MYCIN (Buchanan & Shortliffe, 1984) and NEOMYCIN (Hasling, Clancey, & Rennels, 1984). In these systems, explanations were conceived mainly as reasoning traces of the system — at first resulting in a very technical notion of what an explanation is, with only limited regard to cognitive aspects on the user's side. Still, in the context of REX (Wick & Thompson, 1992), there was already a discussion of how to adapt explanations to different user groups and the trade-offs involved. While interest in XAI subsided after the mid-1990s, recent successes in machine learning technology have brought explainability back into the focus. This has led to a plethora of new approaches for both autonomous and humans-in-the-loop systems, aiming to achieve explainability, as defined by respective system creators, without sacrificing system performance.

Many systems focus on interpretable *post-hoc* approximations of black-box models (Guidotti et al., 2018), using symbolic representations such as decision trees (Craven, 1996; Sarkar et al., 2016) or decision rules (Ribeiro, Singh, & Guestrin, 2018), feature importance (Lou, Caruana, & Gehrke, 2012), saliency maps (Selvaraju et al., 2017), or local regression models (Ribeiro, Singh, & Guestrin, 2016). On the other hand, there are efforts to design intelligent systems to be interpretable by design, e.g., in recommender systems (Zhang & Chen, 2018), or in a recently started project developing the concept of *perspicuous computing*.[2]

In these heterogeneous origins and developments of XAI, a discussion is still to be had on what precisely the roles of explanations are and, in particular, what makes an explanation a good explanation. To this end, we will bring together several experts of different aspects of the phenomenon "explanation" in this symposium, to analyze the notion of explanation in the context of artificial intelligence from different cognition-related perspectives.

What Makes a Good Explanation?

Starting out from the cognition of explanations, this symposium will foster scientific discourse about what

[1]As a case in point, the European Union's General Data Protection Regulation (GDPR) stipulates a right to "*meaningful information about the logic involved*"— commonly interpreted as a 'right to an explanation'— for consumers affected by an automatic decision (Parliament and Council of the European Union, 2016).

[2]https://www.perspicuous-computing.science

functions an explanation needs to fulfill and the criteria that define its quality. Some of the aspects to be addressed are:

- Objective and subjective value of explanations
- Dimensions of explanations: complete vs compact, abstract vs concrete, reduced vs simplified, ...
- Anchoring to known concepts
- Counter-factual explanations and actionability
- Personalisation
- Legal requirements
- Grounding in personal and social experience and intuition

A panel of recognised scholars and researchers will bring insights and expertise from different points of view, including psychology, cognitive science, computer science, and philosophy, and will foster knowledge exchange and discussion of the multiple facets of explanation:

- Kathleen Creel will talk about 'Understanding Machine Science: XAI and Scientific Explanations', drawing on the literature on scientific explanation in philosophy and cognitive science, and arguing that for scientific researchers, good explanations require more access to the functional structure of the intelligent system than is needed by other human users.
- Tania Lombrozo will talk about 'Explanatory Virtue & Vices', considering the multiple functions and malfunctions of human explanatory cognition with implications for XAI. In particular, she will suggest that we need to differentiate between different possible goals for explainability, and that doing so it highlights why human explanatory cognition should be a crucial constraint on design.
- Shane Mueller will talk about 'Ten fallacies of Explainable Artificial Intelligence', reviewing some of the assumptions made until now about what properties lead to good explanations, and describing how each constitutes a fallacy that might backfire if used for developing XAI systems. He will then describe a framework developed for the DARPA XAI Program for measuring the impact of explanations that incorporates cognitive science theory related to mental models, sensemaking, context, trust, and self-explanation that can provide a principled approach for developing explainable systems.
- Patrick Shafto will talk about 'XAI via Bayesian Teaching', raising questions about the use of modern machine learning algorithms in societally important processes, and theoretical questions about whether and how the opaqueness of these algorithms can be ameliorated, in the framework of Bayesian teaching.
- Roberto Confalonieri and Tillman Weyde will talk about 'An Ontology-based Approach to Explaining Artificial Neural Networks', addressing the challenges of extracting symbolic representations from neural networks, exploiting domain knowledge, and measuring understandability of decision trees with users both objectively and subjectively.

References

Buchanan, B. G., & Shortliffe, E. H. (1984). *The Mycin Experiments of the Stanford Heuristic Programming Project*. Addison-Wesley Longman Publishing Co., Inc.

Craven, M. W. (1996). *Extracting comprehensible models from trained neural networks*. (Ph.D. Thesis)

DARPA. (2016). *Explainable AI - program.*

Doran, D., Schulz, S., & Besold, T. R. (2017). What does explainable AI really mean? A new conceptualization of perspectives. In *CEUR Workshop Proc.* (Vol. 2071).

Friedler, S. A., & Wilson, C. (Eds.). (n.d.). *Proceedings of machine learning research* (Vol. 81). PMLR.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *Comp. Surv.*, *51*(5), 1–42.

Hasling, D. W., Clancey, W. J., & Rennels, G. (1984). Strategic explanations for a diagnostic consultation system. *Int. Journal of Man-Machine Studies*, *20*(1), 3 - 19.

Hoffman, R. R., Mueller, S. T., & Klein, G. (2017). Explaining explanation, part 2: Empirical foundations. *IEEE Intelligent Systems*, *32*(4), 78-86.

Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). Metrics for explainable AI: challenges and prospects. *CoRR*, *abs/1812.04608*.

Lipton, Z. C. (2018, June). The mythos of model interpretability. *Queue*, *16*(3), 30:31–30:57.

Lombrozo, T. (2016). Explanatory preferences shape learning and inference. *Trends in CogSci*, *20*(10), 748-759.

Lou, Y., Caruana, R., & Gehrke, J. (2012). Intelligible models for classification and regression. In *Proc. of the 18th ACM KDD* (pp. 150–158). ACM.

Miller, T. (2017). Explanation in artificial intelligence: Insights from the social sciences. *CoRR*, *abs/1706.07269*.

Parliament and Council of the European Union. (2016). *General Data Protection Regulation.*

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proc. of the 22nd Int. Conf. on Knowledge Discovery and Data Mining* (pp. 1135–1144). ACM.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. In *AAAI* (pp. 1527–1535). AAAI Press.

Sarkar, S., Weyde, T., Garcez, A., Slabaugh, G. G., Dragicevic, S., & Percy, C. (2016). Accuracy and interpretability trade-offs in machine learning applied to safer gambling. In *CEUR Workshop Proc.* (Vol. 1773).

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV* (pp. 618–626).

Wick, M. R., & Thompson, W. B. (1992, March). Reconstructive expert system explanation. *Artificial Intelligence*, *54*(1-2), 33–70.

Zhang, Y., & Chen, X. (2018). Explainable recommendation: A survey and new perspectives. *CoRR*, *abs/1804.11192*.

How Does Current AI Stack Up Against Human Intelligence?

Ken Forbus (forbus@northwestern.edu) and Dedre Gentner (gentner@northwestern.edu)
Department of Electrical Engineering and Computer Science, & Department of Psychology, Northwestern University
Evanston, IL. 60208 USA

John E. Laird (laird@umich.edu)
Department of Electrical Engineering and Computer Science, University of Michigan
Ann Arbor, MI. 48109 USA

Thomas Shultz (thomas.shultz@mcgill.ca) and Ardavan Salehi Nobandegani (ardavan.salehinobandegani@mail.mcgill.ca)
Department of Psychology & School of Computer Science, McGill University
Montreal, QC. H3A 1G1 Canada

Paul Thagard (pthagard@uwaterloo.ca)
Department of Philosophy, University of Waterloo
Waterloo, ON. N2L 3G1 Canada

Keywords: artificial intelligence, human intelligence, problem solving, learning, cognitive architecture

Introduction

The past decade has seen remarkable progress in artificial intelligence, with such advances as self-driving cars, IBM Watson, AlphaGo, Google Translate, face recognition, speech recognition, virtual assistants, and recommender systems. Ray Kurzweil and others think that it is only a matter of decades before AI surpasses human intelligence. This symposium will evaluate the extent to which AI currently approximates the full range of human intellectual abilities, and critically discuss the prospects for closing the gap between artificial and human intelligence. Participants will combine the perspectives of computer science, psychology, and philosophy.

The Comparative Cognition of Humans and Machines

Ken Forbus and Dedre Gentner

While there has been great progress in both cognitive science and artificial intelligence, both would benefit from better communication between them. The comparative study of cognition in humans and intelligent machines can shed light on both kinds of systems. In the last decade, the confluence of massive computational resources, massive data sets, and several decades of incremental advances has led to a substantial increase in the ability to build applications with neural networks. Deep learning systems have shown impressive performance in image classification and game learning. However, they still fall far short of capturing human abilities such as explanation and inference, and they require orders of magnitude more data than humans do. We argue that a fundamental lack in these systems is their lack of explicit relational representations. The ability to represent and reason about relational patterns is central to our human ability to explain and predict, and to learn rapidly via analogies with prior knowledge. Fortunately, many of the same factors that have led to gains in deep learning systems are also acting to increase our ability to build large-scale systems with relational representations, which reason and learn in human-like ways. We discuss examples from recent experiments in which analogical learning over relational representations leads to far more humanlike and data-efficient learning than deep learning.

AI and Cognitive Architecture

John E. Laird

There is more talk than ever about general AI, but all the emphasis appears to be on recognition, classification, or reactive decision making with very little on cognition. The emphasis seems to be on only slices of System 1. Within those slices, we see human-level or even super-human performance, but these are very thin slices. Each system is focused on one phenomenon, and given the emphasis on learning from large data sets; it leads to overfitting, not necessarily to specific data, but to the specific problem to the exclusion of developing anything that can work on another problem, or even interact with another cognitive capabilities. In contrast, humans are defined by their flexibility – they can work on many different problems, switching effortlessly from one task to another. They also can learn from many sources of knowledge, on line and in real time, and using a variety of learning techniques. Moreover, they can learn new tasks from scratch in real-time from natural language instruction. A growing field

called Interactive Task Learning has developed an AI system that is embodied in a variety of robotic platforms and that can learn over 50 games and puzzles as well as navigation tasks. It integrates natural language processing, planning, perception, motor control, and learning within a cognitive architecture. Christian Lebiere, Paul Rosenbloom and I have proposed the Common Model of Cognition (CMC) to unify the theoretical underpinnings of many cognitive architectures, starting with Soar, ACT-R, and Sigma. CMC has a vastly different structure than current AI approaches, including procedural and declarative memories, working memory, multiple learning mechanisms. Although these components are common in cognitive science, they are the exception in current AI systems, in large part because of the emphasis on System 1, and off-line batch learning. Until AI takes cognitive architecture, as exemplified by the CMC, seriously, it will not achieve the flexibility, breadth, and adaptability we associate with human intelligence.

Close the Gap and Cooperate

Thomas Shultz and Ardavan Salehi Nobandegani

We will argue that attempts towards achieving artificial general intelligence (AGI) should pay more attention to human intelligence and its neural underpinnings. Having to interact with humans, AGI will need an adequate grasp of human judgment and decision-making and moral principles. Human intelligence not only surpasses current AGI systems, but, importantly, it does so in a resource-efficient way, setting a gold standard for future AI systems. Many of the important AI algorithms originated in psychology, and that strategy is still worth pursuing. A current shortcoming of many AI systems is their limited capacity for generalization – the ability to transfer knowledge from a newly or previously learned task to other relevant tasks. AI could also benefit tremendously from cognitive and developmental psychology to better understand the developmental stages that human infants go through on their way toward adult-level intelligence. To illustrate, we'll focus a bit on the significance of autonomous learning (aka active learning) for bridging the current gap with humans. Even infants take an active role in their own learning by selecting what to work on, what to abandon, and perhaps which examples would be most useful. There is a key role here for learning cessation, the ability to give up on impossible learning tasks, identifiable by lack of continued progress. This paves the way for focusing on tasks in which progress and mastery are more likely. We can suggest ways of implementing these important human capacities in future AI systems. Finally, we want to stress the importance of a cooperative relationship between humans and machines. The notion of gap between us and them that can be closed or even surpassed suggests a more competitive relationship than there perhaps needs to be. The results of mutual cooperation between humans and machines could be much more interesting and desirable to achieve.

How AI Can Understand Causality

Paul Thagard

Causality is important for operating in the world and explaining how it works. Yoshua Bengio and others have pointed out that deep learning and other AI systems lack a human-level understanding of causality. Thagard (2019) argues that human understanding of causality originates with sensory-motor-sensory schemas found in infants as young as 2.5 months. For example, a baby can see a rattle, hit it with hands, and see the rattle move and make a noise. Learning robots could potentially form such schemas, but would have to go beyond current AI systems in several ways. First, they would need *modal retention*, the capacity to save and work with sensory and motor representations. This capacity is found in the Semantic Pointer Architecture of Chris Eliasmith (2013), but not in other cognitive architectures or AI systems. Second, they would need the capacity to learn dynamic patterns that capture changes in series of events. Third, they would need to be able to expand the rudimentary sensory-motor appreciation of causality to cover advanced elements that included regularities, probabilities, and manipulations.

References

Eliasmith, C. (2013). *How to build a brain: A neural architecture for biological cognition*. Oxford: Oxford University Press.

Forbus, K. D., & Gentner, D. (2017). Evidence from machines that learn and think like people. *Behavioral and Brain Sciences*, *40*.

Forbus, K., Liang, C. and Rabkina, I. (2017). Representation and computation in cognitive models. *Topics in Cognitive Science*, doi:10.1111/tops.12277.

Laird, J. E., Gluck, K., Anderson, J., Forbus, K., Jenkins, O., Lebiere, C., Salvucci, D., Scheutz, M., Thomaz, A., Trafton, G., Wray, R. E., Mohan, S., Kirk, J. R. (2017). Interactive task learning, *IEEE Intelligent Systems,* 32(4), 6-2.

Laird, J. E., Lebiere, C., & Rosenbloom, P. S. (2017). A standard model of the mind: Toward a common computational framework across artificial intelligence, cognitive science, neuroscience, and robotics. *AI Magazine*, 38(4), 13-26.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40(e253), 1–72.

Shultz, T. R., & Doty, E. (2014). Knowing when to quit on unlearnable problems: another step towards autonomous learning. In J. Mayor & P. Gomez (Ed.), *Computational Models of Cognitive Processes* (pp. 211–221). London: World Scientific.

Thagard, P. (2019). *Brain-mind: From neurons to consciousness and creativity*. New York: Oxford University Press.

In Vivo Studies of Solo and Team Performance

Ray Perez (Co-Organizer)
Office of Naval Research (ray.perez@navy.mil)

Wayne D. Gray (Co-Organizer)
Rensselaer Polytechnic Institute (grayw@rpi.edu)

Jerad H. Moxley
Weill Cornell Medicine (jhm2006@med.cornell.edu)

David Mendonça
Rensselaer Polytechnic Institute (mendod@rpi.edu)

Jamie C. Gorman
Georgia Institute of Technology (jamie.gorman@psych.gatech.edu)

We bring together four researchers who study expertise in team or in solo (i.e., individual) performance. Team research tends to either collect a lot of questionnaire data after performance or a little data, in real-time, by human observers. Studies of solo performers are often restricted to convenience samples of task novices, who often spend less than an hour learning and performing the task. In contrast, the research of all four of our panelists is notable for using tasks which require days-to-years of practice and for the quantities of data collected. Discussions will emphasize the contributions these approaches are making to theoretical cognitive science.

Jamie C. Gorman – *Theory of Interactive Team Cognition*

By recreating environments for Drone pilots or Submariners, Jamie Gorman and colleagues collect communications and responses among team members in real-time longitudinal studies. These data allowed researchers to apply the power of nonlinear dynamical systems theory to further develop the theory of Interactive Team Cognition (ITC, Cooke, Gorman, Myers, & Duran, 2013). The approach has been extended to teams composed of humans and machines.

ITC proposes that team cognition: (1) is an activity, not a property or product; (2) must be measured and studied at the team level; and (3) is inextricably tied to context.

ITC Prop 1 maintains that team cognition is dynamic and context dependent. ITC Prop 2 leads to a systems perspective in which models and metrics are focused at the team level, with individual cognition and behavior viewed as emergent team dynamics. Team member behavior and cognition are dynamically reorganized (or rearranged) in real time (ITC Prop 1) to maintain functionality as the team adapts to changing task environments to achieve its goal. Hence, teams with high cognitive skill achieve their goal even if environmental context varies and roadblocks to team effectiveness are encountered (ITC Prop 3).

Unlike individual cognition, there are no standard tests to measure the general cognitive skill or ability of a team. One theoretical and methodological development has been to determine a generalizable way to identify and measure team cognitive skill through a team's "general adaptive response".

Our research on team cognition has shown that teams that achieve their goals have (a) a faster general adaptive response, (b) adapt their responses to the variability in obstacles they encounter, and (c) generate responses appropriate to the particular roadblocks they encounter. For examples, I will draw on research with medical teams, submarine crews, UAV teams, as well as in vitro, laboratory, team coordination tasks. This variety of teams illustrates the concept of the *general adaptive response* as an ITC-based measure of team cognitive skill. These teams also illustrate the real-time dynamical system modeling techniques that we use to track team cognition in dynamic environments.

David Mendonça – *Adaptation in Adversarial Games*

David Mendonça's prior research has focused on intensive studies of teams in high-stakes, time-constrained environments. His most recent work is an extremely retrospective analysis of "An historical perspective on community resilience: The case of the 1755 Lisbon Earthquake" (Mendonça, Amorim, & Kagohara, 2018).

In addition to being the most played game in the world (with approximately 10M active users), *League of Legends* (LoL) is an adversarial game (similar to "capture the flag") in which teams must adapt to (and even precipitate) unplanned-for contingencies. Elite players (such as those we study), have played thousands of such matches, with the average match consisting of two teams, each of 5 players, battling for 30 min.

Our work explores the relationship between (i) prematch composition of a team, (ii) decision processes within the match, and (iii) match outcomes in LoL. Respective methodological challenges include (i) characterizing team capabilities, (ii) quantifying adaptation, and (ii) validating measures of performance.

In contrast to traditional work on teams, we utilize no psychometric instruments, instead deriving measures that are validated against salient theoretical constructs and instantiated with gameplay data. And while these data are freely available, their allure is offset by some hard realities: researchers have no influence over either the data stream or the game architecture, and the formulas used to benchmark individual and team expertise are

held as trade secrets. Matches are scheduled by the developers on a rolling basis and–unlike in "regular" sports–are designed so that opponents are closely matched.

After briefly summarizing results to date, we explore within-match performance of teams whose members have weaker or stronger histories of working together, focusing specifically on behavioral responses to the temporary loss of one or more team members. We present data on how the experience of "playing shorthanded" translates (or fails to translate) into longer-term behavioral adaptations.

The talk concludes with issues and implications for the design and/or modification of open-source, team-based games and the data associated with them. ïż£

Jerad Moxley – *Chess: The Once & Future Paradigm*

The distinction of having studied more types of gameplay by solos or teams, than any other researcher on this panel may go to Jerad Moxley. His studies have spanned crossword puzzles, chess, basketball, elderly game players, videogames, as well as gender differences among SCRABBLE players.

For researchers interested in skilled performance, an important feature of chess is the reliability of the chess rating system and the fact that one experimental task (the choose the *best move task*), can measure skill and age effects about as well as tournament play, thereby making Chess ideal for studying domain-specific performance. Complimentary, another common task, the *recall task*, diverges from tournaments in ways that make it useful for studying a mixture of domain specific and domain-general abilities.

Applying the *best move task* and *the recall task* across the lifespan of chess players has increased our understanding of how domain-specific processes and domain-general abilities develop. Research on older adults and children now converges to show strong aging effects of chess tasks that tap into both specific and general abilities. In contrast, the best move tasks captures relatively small aging effects consistent with tournament performance.

As noted, performance on the best move task shows developmental trends in both youth and older adults that mirror tournament performance. Importantly, however, process tracing shows clear differences between the growth of skill in youth and the decline of skill with aging. Although skill development is broadly consistent with what we expect based on tournament performance, the age-related decline of prior skill levels shows process differences that dissociate from skill. In particularly, the age-related declines are not uniform. On easy problems, better players immediately gain an advantage over weaker players.

In contrast, on difficult problems, process tracing has shown that better players initially resemble weaker players but as problem solving continues, better players massively improve their move selection. In contrast, more time does not improve the performance of the weaker players. Methodologically, these conclusions follow from the combination of verbal protocol analysis and the traditional behavioral measures.

We view chess not as a standalone domain, divorced from the rest of human cognition but, rather, as a viable paradigm for studying the big questions in cognitive science. Indeed, the tasks and domains discussed here can easily be used by researchers who have no interest in chess itself to answer their questions of interest.

Wayne D. Gray – *Plateaus, Dips, & Leaps to Expertise*

After several years of working in applied labs, Wayne Gray became concerned that basic researchers were not working on the types of theory he needed to do his job. That concern led him to shift to academe where he has since attempted to pursue theories and research applicable to problems of interactive behavior.

Learning a new task can be hard but, apparently, learning and using a new procedure for an old task can be even harder. That is the message from work on *stable suboptimal performance* from the early 2000s. Wai-tat Fu and I demonstrated time and again that people who knew the optimal procedures would fail to apply them, falling back on older ways of doing things.

Although that battle is still being fought (e.g., Lafreniere, Gutwin, & Cockburn, 2017), the focus in my lab has shifted. After a few years of looking at learning curves for individuals, we realized that none of our curves were close to being picture-perfect power law curves. All of our curves showed plateaus, dips, and leaps. Indeed, what we had thought of as noise was, in fact, the message; namely, that learning a real-time, complex, dynamic task entails a series of explorations and discoveries, trials and errors, in search of methods or strategies that will move performance forward.

We now refer to complete mastery of a task as *asymptotic performance* and to stable suboptimal behavior as performance *plateaus*. However, the most interesting parts of the curve are those periods in which performance dips and, sometimes, leaps. The talk will provide several examples of the use of dips and leaps to identify periods of method discovery or invention.

Ray Perez – Basic Research for Complex Problems

For the last 3 decades, Ray Perez has been pursuing applied problems by finding or encouraging others to find theory-based solutions. Most recently, Ray has been the Program Officer of the Office of Naval Research's Cognitive Science of Learning program.

Ray Perez is co-organizer of this symposium as well as its moderator and discussant. In each of these three roles, Ray is focused on how complex tasks, sometime performed by a single person and other times performed by teams, are learned and executed.

Cognitive Network Science: Quantitatively Investigating the Complexity of Cognition

Yoed N. Kenett* (yoedk@sas.upenn.edu)
Department of Psychology, University of Pennsylvania
Philadelphia, PA 19104 USA
Nichol Castro (castro070503@gmail.com)
Department of Speech and Hearing Sciences, University of Washington
Seattle, WA 98195 USA
Elisabeth Karuza (ekaruza@psu.edu)
Department of Psychology, Pennsylvania State University
University Park, PA 16801 USA
Michael S. Vitevitch (mvitevit@ku.edu)
Department of Psychology, University of Kansas
Lawrence, KS 66045 USA
*** - Organizer**

Keywords: cognitive networks; aging; learning; network science; multiplex networks; complexity

Cognition is complex. This complexity is related to multiple, distributed neurocognitive processes dynamically operating across parallel scales, resulting in cognitive processing. A major challenge in studying this complexity, relates to the abstractness of theoretical cognitive constructs, such as language, memory, or thinking in general. Such abstractness is operationalized, indirectly, via behavioral, measures or in neural activity. In the past two decades, an increasing number of studies have been applying network science methodologies across diverse scientific fields to study complex systems.

Network science is based on mathematical graph theory, providing quantitative methods to investigate complex systems as networks (Baronchelli, Ferrer-i-Cancho, Pastor-Satorras, Chater, & Christiansen, 2013; Siew, Wulff, Beckage, & Kenett, 2018). A network is comprised from nodes, that represent the basic unit of the system (e.g., concepts in semantic memory) and links, or edges, that signify the relations between them (e.g. semantic similarity). While the application of network science methodologies has become an extremely popular approach to study brain structure and function, it has been used to study cognitive phenomena to a much lesser extent. This, despite classic cognitive theory in language and memory being highly related to a network perspective (Collins & Loftus, 1975; Siew et al., 2018). Already, network science in cognitive science has enabled the direct examination of the theory that high creative individuals have a more flexible semantic memory structure, identified mechanisms of language development through preferential attachment, shed novel light on statistical learning, shown how specific semantic memory network parameters influence memory retrieval, and provided new insight on the structure of semantic network of second language in bilinguals (Siew et al., 2018).

The aim of the current symposia is to demonstrate the potential and strength of applying network science methodologies to study cognition. This will be achieved by bringing together leading researchers that apply such methods to study various aspects of cognition, including language, learning, aging, and creativity. The presentations will describe state-of-the-art progress and perspectives that are achieved in applying these methods to study cognition. Importantly, these talks aim at stimulating discussion of the fruitfulness of such an approach and how such an approach can powerfully and quantitatively study the complexity of cognitive phenomena. Finally, this symposium aims to demonstrate how network science in cognitive science can be used to quantitatively bridge across different levels of analysis, spanning the computational, behavioral, neural, and social.

Yoed Kenett: Introducing cognitive network science

In recent years, network science has become a popular tool in the study of structure and dynamics at the neural level of the brain. Despite its rich potential, this has been the case to a lesser extent to the study of cognitive phenomena. This, despite classic cognitive theory in domains such as memory and language being heavily based on a network perspective (Collins & Loftus, 1975; Siew et al., 2018). In this talk, I will argue for the potential of applying the quantitative language of networks to study cognition. I will first describe methodological approaches to estimate cognitive networks and relevant network science measures. I will then briefly describe how cognitive network research can be applied to study the structure, processes, and dynamics of cognitive domains. These examples will focus on semantic memory and relate to aspects of creativity, spreading activation, and semantic memory restructuring. Finally, I will argue that cognitive network science can be used to quantitatively bridge across multiple domains of analysis, spanning the neural, cognitive, and social.

Nichol Castro: Capturing the aging lexicon using network science techniques

Word findings problems increase with age, even in the absence of disease or impairment. Although some accounts attribute word finding problems to changes in domain general cognitive processes, the prominent explanation is a deficit in accessing phonology due to weakened connections between lexical items and their phonological constituents between lexical items and their phonological constituents (Burke, MacKay, Worthley, & Wade, 1991). In other words, there is a change in the structure of the mental lexicon that occurs with age. However, quantifying structural change in the mental lexicon has remained understudied. This talk will show how the tools of network science can be used to identify key structural changes in phonological and semantic networks occurring across adulthood (e.g., Dubossarsky, De Deyne, & Hills, 2017). A discussion of how structural change could impact language processing will ensue, followed by a brief foray into the implications of aging lexicon networks in clinical populations. In particular, it's important that we consider how aging impacts the lexicon of not just "healthy" adults, but also those who have suffered brain insult (e.g., in the case of stroke-induced aphasia).

Elisabeth Karuza: Probing the level at which learners track co-occurrence patterns

Humans are highly attuned to the clustering of elements in their surroundings. For example, when learners are confronted with novel sequential input, their element-by-element processing times have been shown to reflect the community structure (i.e., multi-element patterns of co-occurrence) underpinning those sequences. In this talk, I will detail recent developments in a framework for examining learners' sensitivity to the network structure of their environment. Prior applications of this framework have generally involved assigning a handful of unnatural stimuli (e.g., fractal images) to nodes in a small network and generating sequences by walking along its edges (Karuza, Kahn, Thompson-Schill, & Bassett, 2017; Schapiro, Rogers, Cordova, Turk-Browne, & Botvinick, 2013). Here, I will describe the expansion of this approach to encompass the study of larger temporal networks comprised of more naturalistic stimuli (e.g., manipulable objects and phonotactically legal pseudowords). Finally, I will examine how the collection of off-line measures might serve to clarify the previously observed relationship between on-line processing times and network architecture.

Michael Vitevitch: Connecting the MIND and the BRAIN with multiplex networks

Poeppel and Embick (2017) describe two problems researchers face when trying to bring together the mind and the brain: (1) granularity mismatch problem, and (2) ontological incommensurability problem. In the granularity mismatch problem, the elemental concepts and operations of Cognitive Science doesn't match the elemental concepts and operations of Neuroscience. In the ontological incommensurability problem, the fundamental elements of Cognitive Science cannot be reduced to or matched up with the fundamental biological units of neuroscience. Poeppel and Embick (2017) suggest that computational models may overcome these problems and provide the desired bridge between mind and brain. As an alternative to bridging the mind and brain, I discuss the possibility (and potential problems) of using multiplex networks to bridge mind to brain, and to bridge the individual mind-brain to the mind-brains of others.

References

Baronchelli, A., Ferrer-i-Cancho, R., Pastor-Satorras, R., Chater, N., & Christiansen, M. H. (2013). Networks in cognitive science. *Trends in Cognitive Sciences, 17*(7), 348-360.

Burke, D. M., MacKay, D. G., Worthley, J. S., & Wade, E. (1991). On the tip of the tongue: What causes word finding failures in young and older adults? *Journal of Memory and Language, 30*(5), 542-579.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review, 82*, 407-428.

Dubossarsky, H., De Deyne, S., & Hills, T. T. (2017). Quantifying the structure of free association networks across the life span. *Developmental Psychology*.

Karuza, E. A., Kahn, A. E., Thompson-Schill, S. L., & Bassett, D. S. (2017). Process reveals structure: How a network is traversed mediates expectations about its architecture. *Scientific Reports, 7*(1), 12733.

Poeppel, D., & Embick, D. (2017). Defining the relation between linguistics and neuroscience, *Twenty-first century psycholinguistics* (pp. 103-118): Routledge.

Schapiro, A. C., Rogers, T. T., Cordova, N. I., Turk-Browne, N. B., & Botvinick, M. M. (2013). Neural representations of events arise from temporal community structure. *Nature Neuroscience, 16*(4), 486-492.

Siew, C. S. Q., Wulff, D. U., Beckage, N. M., & Kenett, Y. N. (2018). Cognitive network science: A review of research on cognition through the lens of network representations, processes, and dynamics. *arXiv*.

Symposium in Memory of Jeff Elman: Language Learning, Prediction, and Temporal Dynamics

James L. McClelland (jlmcc@stanford.edu)
Department of Psychology, Stanford University
Stanford, CA 94305 USA

Ken McRae (kenm@uwo.ca)
Department of Psychology, University of Western Ontario, London, ON N6A 5C2 Canada

Arielle Borovsky (aborovsky@purdue.edu)
Speech, Language, & Hearing Sciences
Purdue University
West Lafayette, IN 47907 USA

Gina R. Kuperberg (gkuperberg@mgh.harvard.edu)
Department of Psychology, Tufts University, Medford, MA, 02155 USA

James S. Magnuson (james.magnuson@uconn.edu)
Department of Psychological Sciences, University of Connecticut, Storrs, CT 06269-1020, USA

Felix Hill (felixhill@google.com)
DeepMind, 6 Pancras Square
London, N1C 4AG UK

Keywords: Jeff Elman, simple recurrent networks, prediction, TRACE, language development, event knowledge

Introduction

Jeffrey Locke Elman (1948-2018) devoted his career to studying human language. He investigated how people use language flexibly and productively, and how these abilities are learned from linguistic and other input. Jeff was a faculty member at the University of California, San Diego from 1977 until he passed away in 2018. His early research focused on phonetics and phonology. This work began his theoretical journey that resulted in the ideas for which he is best known: seemingly discrete combinatorial units of language, such as phonemes, may best be understood as emergent properties of underlying continuous multidimensional representations, such as phonetic input.

In the early 1980's, Jeff was a key part of transformative developments at UCSD in connectionist modeling, working with Jay McClelland, Dave Rumelhart, Geoff Hinton, and Liz Bates. With McClelland, he developed the TRACE model of speech perception. In TRACE, speech perception is seen as a constraint satisfaction process in which prior and subsequent context combine with incoming sensory evidence to determine how humans perceive speech.

Jeff then turned his attention to a central but often neglected aspect of cognition: time. Jeff's work on Simple Recurrent Networks, beginning with his classic 1990 article *Finding Structure in Time*, proposed that time-evolving continuous hidden-state representations are fundamental to language processing, and enable prediction-based learning of language. This work remains among the most influential in the history of Cognitive Science. Jeff's subsequent work explored the initial conditions under which a simple recurrent network would recover grammatical structure. He then led a collaborative project to rethink the nature of what must be built in as a foundation for language, and more generally for cognition (Elman et al., 1996). In later work, he focused on the relationship between language and event knowledge. He argued that words do not have meanings, but instead provide clues that a listener uses to understand language. He also focused on event knowledge as a basis for prediction during language comprehension (Elman, 2009; Metusalem et al., 2012). Jeff's final major contribution was a model of how event knowledge is learned. He argued that knowledge of the components and temporal structure of events emerges as a consequence of prediction-based learning (Elman & McRae, 2019).

Jeff also played a major role in advancing Cognitive Science as a field. At UCSD, he and colleagues co-founded the interdisciplinary Center for Research in Language in 1985. In 1986, Jeff was a major part of the first Cognitive Science department, which he chaired from 1995 to 1998. Jeff also served as Dean of Social Sciences, and a founder of both the Kavli Institute for Mind and Brain and the Halicioğlu Data Sciences Institute. Finally, Jeff provided guidance for the field by serving as President of the Cognitive Science Society, and a highly respected Chair of the NIH Language and Communication study section.

This symposium honors Jeff's memory. The introduction and discussion will be led by the organizers (McClelland & McRae). In between, four speakers whose work reflects the legacy of Jeff's contributions will present research from the perspectives of cognitive neuroscience, cognition and perception, language development, computational modeling, and deep learning in simulated embodied agents.

Talks

Gina R. Kuperberg
Language prediction over time and space: Evidence from multimodal neuroimaging studies

In his seminal paper, *Finding structure in Time*, Elman argued that predictions are based not just on input from the world, but on the ever-changing state of the cognitive system. He emphasized the idea that these predictions are

non-deterministic, implicit, and inevitable. He also pointed out that prediction error not only provides feedback to the system (to learn and maximize its performance), but that it also provides valuable clues for the scientist: it can tell us about the structure of the input and the nature of cognition. These ideas have far-reaching implications for thinking about what neural measures can tell us about the architecture of language comprehension. I will discuss evidence from multimodal neuroimaging studies (ERP, fMRI and MEG) that, during comprehension, spatiotemporally distinct neural signatures reflect neural prediction error and updating at multiple time scales. I will argue that they point to a language comprehension system in which probabilistic predictions are generated and incrementally updated over time, at multiple levels and grains of representation, with the ultimate goal of inferring the latent cause that best explains the full set of inputs encountered — the message that the communicator intended to convey. Consistent with Elman's ideas, I also will argue that the neural responses evoked by prediction violations play a crucial role in triggering us to rapidly adapt to the statistical structure of our ever-changing communicative environments so that we can predict more efficiently in the future.

Arielle Borovsky
Prediction in a changing world

This talk connects with several of Elman's contributions, including his perspective on prediction, learning over time, and event knowledge in language learning and processing. Numerous language processing models emphasize the importance of listeners' ability to predict upcoming information for efficient language comprehension and learning. Much of the evidence for these models is derived from studies of comprehension in well-known or familiar (i.e. predictable) contexts. However, speakers are pressed to prioritize novel information, suggesting that everyday conversation does not typically rehash redundant events. In developmental and learning contexts, this problem may be compounded by the fact that listeners may still be learning about the language and the world. Therefore, they may not have sufficient knowledge to generate predictions. In all of these circumstances, prediction might be counter-productive for comprehension. I will discuss recent studies of how adults and children engage in prediction while learning about new events. The findings illustrate that while adult listeners can rapidly modify their predictions in the face of change, children develop this flexibility gradually over a protracted period. By incorporating developmental insights and learning paradigms into studies of linguistic prediction, we can develop richer models of how predictive mechanisms support everyday communication and learning.

James S. Magnuson
Elman's agenda for the cognitive science of language processing

I will review the remarkable breadth and depth of one of Elman's major contributions: the TRACE model of speech perception and spoken word recognition (McClelland & Elman, 1986). I then will apply one of his other major contributions – the simple recurrent network (SRN; Elman, 1990) – to the same domain. Remarkably, SRNs have not been applied deeply to problems in spoken word recognition. Even more remarkably, despite seemingly large differences in architecture, TRACE and SRNs make extremely similar predictions, including item-specific predictions for large sets of items. I will conclude by considering how deeply Elman's ideas and work have shaped the cognitive science of language processing.

Felix Hill
Embodied neural network agents that learn language in a simulated world

I will describe a neural network 'agent' that is situated in a fully-navigable simulated 3D world as a model of early child word learning. The agent perceives its world via first-person continuous raw visual input and must learn to respond, with appropriate sequences of fine-grained motor actions, to symbolic language-like stimuli that describe simple goals. Recurrent components inspired by Jeff Elman's work play an important part in this architecture both for processing language word-by-word and for making sense of experience timestep-by-timestep. I explain how, under certain training conditions, the agent learns to reflect some known aspects of human word learning, including the emergence of semantic classes, vocabulary spurts, curriculum effects and word-learning biases. I further demonstrate how word learning can be sped up by incorporating an offline experience replay mechanism. Finally, I discuss the strengths and weaknesses of modelling early word learning with deep reinforcement learning agents in this way.

References

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*, 179-211.

Elman, J. L. (2009). On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive Science*, *33*, 547-582.

Elman, J. L., Bates, E., Johnson, M. H., Karmiloff-Smith, A., Parisi, D. & Plunkett, K. (1996). *Rethinking Innateness.* Cambridge, MA: MIT Press.

Elman, J. L., & McRae, K. (2019). A model of event knowledge. *Psychological Review*, *126*, 252-291.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*, 1-86.

Metusalem, R., Kutas, M., Urbach, T. P., Hare, M., McRae, K., & Elman, J. L. (2012). Generalized event knowledge activation during on-line language comprehension. *Journal of Memory & Language*, *66*, 545-567.

Understanding interactions amongst cognitive control, learning and representation

Sebastian Musslick, Abigail Hoskin Novick, Taylor Webb, Steven Frankland & Jonathan Cohen
Department of Psychology & Neuroscience Institute,
Princeton University, Princeton, NJ

Rebecca J. Jackson & Matthew A. Lambon Ralph
MRC Cognition and Brain Science Unit
Cambridge University, Cambridge, UK

Lang Chen
Department of Psychiatry and Behavioral Sciences
Stanford University, Palo Alto, CA

Timothy T. Rogers
Department of Psychology
UW-Madison, Madison, WI

Introduction

Research in cognitive control investigates how cognition and behavior get tailored to suit behavioral goals in particular task contexts. The work often focuses on mechanisms that adjudicate competition amongst simultaneously active but mutually incompatible representations. The objects of control—the competing representations—are often cast as fixed entities: control influences interactions among these but does not shape the representations themselves. Conversely, research into the origins of mental representations (perceptual, linguistic, semantic, etc.) often neglects questions central to theories of control: whether and how the acquired representations support flexible task-dependent behaviors, the degree to which learning produces representations that compete or cooperate within and across tasks, or the extent to which learned representations require task-dependent potentiation to operate effectively.

Recent work within each tradition suggests, however, that control, learning, and representation are tightly interconnected. First, the degree to which control is required for any given task and stimulus domain depends critically on the nature, structure, strength, and compatibility of the underlying representations, which in turn arise from learning and experience. Second, when the same items are engaged in a variety of different tasks, it may be useful to exploit a common representation across tasks, or to learn different representations for each, or to find some middle ground—thus learning must produce a flexible set of representations suited to control demands and capturing shared structure within and across task contexts. Third, since control shapes the flow of activation within sensory, motor, and associative systems, it must also constrain activation-dependent learning within and between these systems—that is, the representations acquired must depend to some degree upon control.

This symposium brings together four perspectives on the mutual constraints existing among systems of control, learning, and representation. In each case, consideration of these mutual influences leads to new and often surprising resolutions to long-standing questions across seemingly disparate domains of cognitive neuroscience.

The rational boundedness of cognitive control: Shared versus separated representations

Sebastian Musslick, Abigail Hoskin Novick &
Jonathan D. Cohen, Princeton University

A fundamental and striking limitation of human cognition is the constraint on the number of control-dependent processes that can be executed simultaneously, which forms one of the most basic and influential tenets of cognitive psychology: controlled processing relies on a central, limited capacity processing mechanism that imposes seriality on control-dependent processes. We present a challenge to this view that distinguishes control-dependent and automatic processing by their reliance on shared vs. separated representations. Specifically, we propose that: task performance relies on sets of representations that may be shared with others; the inability to perform more than one task at a time may reflect conflict that arises when the tasks involved make use of the same set of representations for different purposes; and the purpose of control is to prevent such conflict by restricting use of such shared sets of representations to just one task at a time. That is, constraints associated with control-dependent processing reflect a rational response to sharing of representations, rather than limitations in the control mechanism itself. We use graph-theoretic methods to formalize this theory, and analyze the multitasking capability of two-layer neural networks when representations are shared/not shared across tasks. The multitasking capability of a network drops precipitously with an increase in shared representations, and is virtually invariant to network size.

Why then should a network use shared representations at all? In computational simulations and behavioral experiments we demonstrate a tradeoff between learning efficiency, promoted by shared representations, and multitasking, best achieved via separated representations. The commonly-observed trajectory from controlled to automatic processing may therefore reflect an optimization of this tradeoff: shared representations initially afford a bias toward efficient learning in novel task environments at the expense of seriality and control-dependence; but experience in environments where multitasking affords sufficient advantage ultimately promotes acquisition of separated, task-dedicated representations.

Canonical representations for generalization in relational reasoning

Taylor Webb, Steven Frankland, Alexander Petrov[1], Randall C. O'Reilly[2] & Jonathan D. Cohen, Princeton University, Ohio State[1], and U. Colorado-Boulder[2]

The preceding talk suggests that capacity limits on control-dependent tasks fundamentally arise from the use of shared representations across tasks. Why then should cognitive systems employ shared representations? The answer may lie in the remarkable human capacity to generalize far beyond the scope of experience. By contrast, state-of-the-art neural network algorithms tend to do well at interpolating between data points in their training corpora, but generally fail to extrapolate beyond the scope of those data points.

We propose that one way to enable human-like generalization in neural networks is by giving them access to a basis set of canonical, general-purpose representations that capture the abstract relations inherent in common structural motifs (e.g. lines, rings, or trees). We present a method for transforming domain-specific representations into a canonical form, and show that these transformed representations enable robust extrapolation to data points far from the training domain — that is, out of domain generalization. Such broad generalization requires, however, that processes within and across task and item domains share use of the canonical representations, thus making them dependent on control. Understanding the conditions under which canonical representations arise thus provides insight into both the human capacity for generalization and the relationship of this ability to cognitive control.

Toward a neural architecture for controlled semantic cognition

Rebecca J. Jackson, Timothy T. Rogers & Matthew A Lambon Ralph, Cambridge University

We consider how opposing demands of task-specific control versus broad generalization might constrain the architecture of the networks that support semantic cognition—the remarkable human ability to flexibly deploy conceptual knowledge across a variety of behavioral contexts. The semantic system must acquire context-invariant representations that express conceptual structure by abstracting over episodes, time, and modality (sensory, motor, linguistic, and affective), while also dynamically tailoring representations to produce context-appropriate similarity structures and behaviors. How should a semantic system be structured to promote both functions?

We report simulations with models varying in five architectural features, representing different hypotheses about the influence of control on semantic processing and the structure of the semantic network itself. We compared model variants in their acquisition of both context-invariant conceptual structure and context-dependent tailoring of representations and outputs. The system's functioning was best served by an architecture employing a single, deep multimodal hub containing sparse long-range connections from modality-specific inputs, and with control systems operating on peripheral modality-specific representations without affecting the hub. This architecture creates regions of relative specialization for control and representation, explaining distinct patterns of semantic dysfunction arising from temporal versus fronto-parietal pathology. The simulations thus suggest that the cortical anatomy of semantic cognition can be understood as balancing demands of representation and control.

Learning, control, and modularity in lexical semantics.

Lang Chen, Stanford University
Timothy T. Rogers, University of Wisconsin-Madison

A central goal for cognitive approaches to language has been to understand whether various sub-processes operate independently or are mutually interdependent. In accordance with the preceding talks, we suggest the tension between views can be resolved by considering how cognitive control and task-specific experience jointly impact learning in lexical semantic systems, taking visual word recognition as a well-studied example of the controversy. On one hand, patients with acquired semantic impairments typically show difficulty recognizing low-frequency words with unusual orthography, suggesting an interdependence between lexical and semantic representations. On the other, a handful of cases show serious semantic impairment with normal word recognition, suggesting that recognition and semantic processes are independent. Similar patterns in other aspects of language have produced fundamentally different perspectives: one in which all varieties of linguistic representation mutually constrain one another, and another in which different representations are modular and independent.

We show that lexical and semantic representations in a recurrent neural network can become modular when (1) words appear in task-contexts requiring independent activation of each representation and (2) a context-dependent control signal strongly constrains activation in the network. This model suggests that individuals with strong executive control and unusually frequent experience with orthography may develop relatively independent lexical and semantic representations. We tested this hypothesis using dual-task studies to assess semantic interference on word recognition. Most participants showed degraded recognition with concurrent semantic processing but a small percentage showed no such effect. These exceptions uniformly showed exceptional orthographic knowledge and no interference in a Stroop task—suggesting that strong control and practiced orthography jointly promote independent lexical and semantic processing. The results offer a middle ground between fully modular and fully interactive perspectives, and suggest that control and learning play critical roles in shaping the degree to which various linguistic representations interact.

Beyond number: Towards a unified view of dimensional reasoning in perception, cognition, and language

Speakers:
Stella Lourenco & Lauren Aulet
(stella.lourenco@emory.edu; lauren.s.aulet@emory.edu)
Dept of Psychology, Emory University, Atlanta, GA

Jessica Cantlon (jcantlon@andrew.cmu.edu)
Department of Psychology, Carnegie Mellon University, Pittsburgh, PA

Anna Papafragou (papafragou@psych.udel.edu)
Department of Psychological and Brain Sciences, University of Delaware, Newark, DE

Pooja Paul (*organizer*) (ppaul18@stanford.edu)
Center for the Study of Language and Information, Stanford University, Stanford, CA

Keywords: quantitative scales; analog magnitude systems; transitive inference; ordinality; comparative reasoning; animal behavior; conceptual development; language; psychophysics;

Introduction

Natural number concepts play a fundamental role in abstract human thought, being central to mathematics, science and measurement, as well as pervasive in our everyday reasoning. A major turning point in our understanding of the psychological bases of number came with the discovery of an approximate, analog system for representing numerical magnitude found to underlie our numerical intuitions in tasks ranging from relative numerosity judgments, to addition, subtraction, and ordering, among others. Crucially, such analog magnitude representations are involved not just for number, but for all other kinds of dimensions as well, from *physical size*, *loudness, brightness*, and *duration* (e.g. Fias et al. 2003; Cordes & Brannon, 2008), to more evaluative dimensions like *likelihood* (Wellman, Kushnir, Xu & Brink, 2016). Based on extensive studies on human adults, children, pre-verbal infants and non-human species, we now understand the systems underlying the mental representation of scalar dimensions to be best characterized as approximate, analog representations with signature ratio limits (obeying Weber's Law), operational in humans from birth and throughout the lifespan, and shared with a wide range of other animal species.

The primary goal of this symposium is to bring recent developments from infant and comparative psychological research pertaining to our understanding of analog magnitude systems to a broader audience of cognitive scientists, to discuss their implications for human cognition. With a more complete picture of the kinds of inferential capacities afforded by analog magnitude and other systems in non-human animals and preverbal infants, we are in a better position to understand the interplay between language and non-linguistic systems in the human mind.

Recent developments

A wealth of research in developmental and comparative cognition in recent years has revealed previously unexpected inferential capacities in infants and non-human animals that are evidently supported by analog magnitude representations.

Cross-dimensional mapping in infancy

It is well-known that adults and children readily map analog magnitude representations to one another (e.g. Stevens & Marks, 1965), but it is a more recent discovery that this tendency in fact begins in infancy. For example, given evidence for a correspondence between numerosity and line length in a visual habituation task, human newborns expect shorter lines to correspond to smaller numerosities, and longer lines to correspond to larger numbers (de Hevia & Spelke, 2010). That newborn infants spontaneously map between number and space, as well as duration (de Hevia et al., 2014), suggests that at least some kinds of scalar mappings may precede experience. Importantly, older infants have been shown to learn more arbitrary mappings in a context-specific manner as well (Lourenco & Longo, 2010), raising the possibility that tracking correspondences between environmentally co-occurring variables may be one way in which infants learn about their physical (and social) worlds in infancy, before access to language.

Transitive inference in animals

Yet another reasoning strategy implicated to be subserved by the analog magnitude systems is transitive inference (TI), the ability to infer from $A > B$ and $B > C$ that $A > C$. Extensive and well-controlled studies of non-human animals in recent decades have revealed a pervasive capacity for transitive inference in species ranging from fellow primates and mammals, to birds, amphibians, and fish. The capacity to represent ordinal relationships is a prerequisite for transitive inference, and as such, TI can be considered a kind of order-based reasoning. Cantlon and Brannon (2006) find behavioural evidence for shared systems for ordering numerical magnitudes in humans and monkeys, and moreover that both groups exhibit semantic congruity effects, signalling a common mental comparison process (Cantlon & Brannon, 2005). The preponderance of evidence for successful non-symbolic TI and order-sensitivity in the animal literature has important implications for human reasoning that are yet to be fully explored by the cognitive scientific community. Such evidence should be of particular interest to those investigating the conceptual foundations of symbolic thought, given the implication that the binary *more than* relation ('<') in language and mathematics may have its basis in analog magnitude systems.

Scalar phenomena in language

In linguistics, conceptual and pragmatic scales are invoked in explanations of linguistic phenomena ranging from gradable adjectives ('*tall*', '*fast*', '*large*', '*ambitious*') and comparative and superlative constructions ('*Ben is taller*

than Dan'; '*Ben is the tallest*'), to scalar implicature (Horn 1972; Hirschberg 1985), to name a few. That classic behavioral signatures of analog magnitude systems --- the symbolic distance effect (e.g. Moyer & Landauer, 1967) and semantic congruity effects (e.g. Banks, Clark & Lucy, 1975) --- arise in tasks involving gradable adjectives, provides some support for links between these linguistic labels and underlying analog format representations. But most well-studied in this regard are the bidirectional linkages between natural numbers (<'*one*', '*two*', '*three*', …>), and corresponding analog magnitude representations in the numerate human mind (Odic, Le Corre & Halberda, 2015). Given that both the number scale as well as scales comprised of gradable adjectives give rise to scalar implicatures, it is worthwhile to consider whether similar mechanisms to those supporting dimensional inference in infants and animals, may also be involved in scale-based reasoning in humans. As it happens, there is recent evidence for the use of parallel strategies for scalar inference by children and adults in non-linguistic tasks (Kampa & Papafragou, 2019; Gweon & Asaba, 2018) lending credence to this possibility.

The pervasiveness of scalar phenomena cross-linguistically, in light of the developments highlighted above, raises the following questions: First, taking for granted that conceptual scales are indeed psychologically 'real', how should they be characterized in psychological terms? What is the precise nature of the relationship between conceptual and/or pragmatic scales, and associated analog magnitude representations? Finally, are there deeper connections between the inferential capacities afforded by analog magnitude systems in preverbal infants and nonverbal animals, and the widely-studied phenomena of scalar and quantity-based inference in linguistically-savvy humans? More specifically, might there be shared neural and cognitive mechanisms for the computation of dimensional inferences in the linguistic, cognitive, and perceptual domains?

Linguists in the 1980's and 90's theorized the existence of 'scalar models' that map between two or more correlated dimensions to support implicit inferences arising with scalar language (e.g. Fauconnier, 1975; Kay, 1990; Israel 1996). Although such cognitive accounts subsequently fell out of favor within mainstream linguistic theory, the empirical clarity provided by psycholinguistic findings in recent years has convinced some that a better understanding of the conceptual structures that language links up to "under the hood" may be essential to account for the distribution of various classes of linguistic inference (Paul, 2018). The superficial similarity of the early theoretical models of scalar linguistic reasoning to the recent empirical results from the infant literature (i.e., bidirectional mappings between statistically correlated properties), suggests the former may be ripe for revisiting. The different disciplines studying phenomena involving the representation of dimensional attributes stand to gain from sharing insights across disciplinary boundaries, something we hope to foster with this symposium. Moreover, this symposium has the potential to inspire renewed efforts towards a more psychologically informed model of scalar reasoning in language, and possibly even a unified model of dimensional reasoning in human and animal cognition.

Speakers:

Stella Lourenco will represent the perspective from infant cognition, specifically her research on cross-dimensional mappings in infancy, as well as some brand new cognitive neuroscientific results from her lab supporting a generalized system of magnitude representation.

Jessica Cantlon will discuss the comparative cognitive perspective, including findings of parallel behavioral patterns in human adults and monkeys in numerical ordering and other tasks, and what this reveals about our shared mental processes for magnitude comparison.

Anna Papafragou will focus on the development of scalar implicature, and present new work showing that adults and children's behavioral patterns in non-linguistic and linguistic versions of a task eliciting scalar implicature are guided by a common principle.

Pooja Paul will employ her background in linguistics and developmental psychology to disentangle the contributions of extra-linguistic domains from that of language in scalar reasoning. Her presentation will synthesize the different strands of research presented during the symposium, and paint a picture of what a unified theory of dimensional reasoning might look like.

References

Banks, Clark & Lucy (1975). The locus of semantic congruity effects in comparative judgments. *J. of Exp. Psych.*

Cantlon & Brannon (2005). Semantic congruity affects numerical judgments similarly in monkeys and humans.

Cantlon & Brannon (2006). Shared system for ordering small and large numbers in monkeys and humans. *Psych. Science.*

De Hevia, Izard, Coubart, Spelke & Streri (2014). Representations of space, time, and number in neonates.

Fauconnier (1975). Pragmatic Scales and Logical Structure.

Gweon, H. & M. Asaba (2018). Order matters: Children's evaluation of underinformative teachers depends on context.

Holmes & Lourenco (2009). Spatial organization of magnitude in the representation of number and emotion.

Kampa & Papafragou (2019). Do children extend pragmatic principles to non-linguistic symbols? *Symposium talk at SRCD*, Baltimore, March 21-23.

Lourenco & Longo (2010). General magnitude representation in human infants. *Psychol Sci.* 21(6).

Moyer & Landauer (1967). Time required for judgments of numerical equality. *Nature.*

Odic, Le Corre & Halberda (2015). Children's mappings between number words and the approximate number system.

Paul (2018). The linguistic and conceptual representation of scalar alternatives: Number and *only* as case studies. *Ph.D Thesis*. Harvard University.

Stevens & Marks (1965). Cross-modality matching of brightness and loudness. *Proc. Natl. Acad. of Sciences 54*(2).

Walsh (2003). A theory of magnitude: common cortical metrics of time, space and quantity.

Extending Rationality

Emmanuel M. Pothos[1] (Emmanuel.pothos.1@city.ac.uk), Jerome R. Busemeyer[2] (jbusemey@indiana.edu), Tim Pleskac[3] (pleskac@ku.edu), James M. Yearsley[1] (James.Yearsley@city.ac.uk), Joshua B. Tenenbaum[4] (jbt@mit.edu), Noah D. Goodman[5] (ngoodman@stanford.edu), Michael Henry Tessler[4] (tessler@mit.edu), Thomas L. Griffiths[6] (tomg@princeton.edu) , Falk Lieder[7] (falk.lieder@tuebingen.mpg.de), Ralph Hertwig[8] (hertwig@mpib-berlin.mpg.de), Thorsten Pachur[8] (pachur@mpib-berlin.mpg.de), Christina Leuker[8] (leuker@mpib-berlin.mpg.de) & Richard M. Shiffrin[2] (shiffrin@indiana.edu)

[1]Department of Psychology, City, University of London, Northampton Square London, EC1V 0HB, UK
[2]Department of Psychological and Brain Sciences, Indiana University, Bloomington, IN 47405, USA
[3]Tim Pleskac, Department of Psychology, University of Kansas, Lawrence, KS 66045, USA
[4]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[5]Noah D. Goodman, Department of Psychology, Stanford University, Stanford, CA 94305, USA
[6] Department of Psychology, Princeton University, Princeton, NJ 08540, USA
[7]Max Planck Institute for Intelligent Systems, Tübingen 72076, Germany
[8]Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin 14195, Germany

Keywords: rationality, bounded rationality, fallacies, heuristics, resource-rational, probabilistic programming language, classical and quantum probability theory

Fallacies?

Since antiquity, we have wondered about the foundations of our (apparent) intellectual superiority. A way to approach this issue is to seek rational standards in decision making and examine convergence between such standards and behavior. However, establishing a rational framework is not straightforward. One of the most unique contributions of cognitive science is the varied perspectives it has provided for rationality. With many recent advances in decision theory (including novel techniques in probabilistic inference and sophisticated frameworks for heuristics-driven reasoning), it is particularly timely to reevaluate rational standards and our assumptions regarding rational behavior. This is the purpose of this interdisciplinary symposium, bringing together expertise in psychology, computer science, mathematics, physics, and philosophy of mind.

Cognitive science research has already instigated major shifts in our perception of rationality and optimality. For most of our history, it has been considered that classical logic is the source of human rationality and the appropriate normative standard against which to assess decisions. Wason sought a general test of whether natural reasoning is consistent with classical logic, by asking participants to select which evidence was best suited to test a given rule. Logic prescribes selections which can definitely falsify the rule (a falsificationist mentality which has had a pervasive influence in scientific reasoning, including in frequentist statistics), but instead participants selected evidence with potential to confirm the rule. Oaksford and Chater (1994) proposed that participants prefer the cards which minimize the information-theoretic uncertainty regarding the validity of the rule, employing Anderson's (1990) idea of optimal adaptation. Classical probability theory (CPT) thus revealed an alternative perspective for the 'correct' selections in Wason's task.

CPT is currently recognized as the right starting point for understanding rational decision making, benefiting from powerful formal justifications and excellent descriptive coverage. Equally, it has been increasingly appreciated that a *baseline* CPT framework is unlikely to provide either a complete descriptive framework for cognition or indeed an appropriate normative framework, without suitable extensions (e.g., Tenenbaum et al., 2011). One influential source of indication that this is the case concerns reports of persistent apparent violations of CPT principles, usually called fallacies. Tversky, Kahneman and their colleagues have produced some of the most evocative examples, for example, the conjunction fallacy, according to which naïve observers are quite happy to accept that *Prob*(A&B)>*Prob*(A) (Tversky & Kahneman, 1983). The most telling instantiation of this result involves the probability of a Scandinavian person having blue eyes and blond hair vs. just having blond eyes (Tentori et al., 2004). Imagining a line-up of Scandinavian individuals makes it immediately obvious why the conjunction fallacy is, well, a fallacy, and yet the conjunctive statement still feels natural – it is this persistence that makes fallacies so puzzling. There are several similar results. For example, a famous Gallup poll study showed a *Prob*(Clinton is honest) of 50% when this question was first but 57% after a similar question for Gore (Moore, 2002); in another famous study, a mixture of weak and strong evidence had less impact than just the strong evidence (the dilution effect; Nisbett et al., 1981).

Such findings appear to challenge our expectation of rationality. But do they have to? Over the last decades, new, sophisticated techniques and ideas have emerged, which require drastic revision to our perception of applicability of *baseline* CPT frameworks in thought. In this symposium we explore four approaches, some of which directly extend baseline CPT ideas while others are motivated from baseline CPT ideas to develop in more alternative directions, with sometimes surprising implications for empirical coverage and normative evaluation.

Resource-rational analysis: Griffiths, Lieder

Baseline CPT inference is expensive, and practical models often involve some kind of sampling-based approximation to posterior probabilities. In the tradition of bounded

rationality, the resource-rational analysis is about finding the optimal balance between the accuracy of probabilistic approximations and resource allocation, with the latter formulated in terms of computational cost (Griffiths et al., 2015). This approach can recover previously-identified heuristics and discover new ones, as well as shed light in the way resource limitations can lead to apparent deviations from CPT prescription.

Quantum: Busemeyer, Pleskac, Yearsley, Pothos

Another way in which CPT probabilistic inference can be made more tractable is by limiting the size of the probabilistic space. The logical structure of CPT is a Boolean algebra, but for Quantum probability theory (QPT) it is a partial Boolean algebra, which means a collection of smaller (simpler) parts, which are classical individually, but inconsistencies/ contextuality/ apparent fallacies arise when reasoning between parts. We think that QPT representations are more likely when e.g. participants are unfamiliar with a problem or unwilling to engage thoughtfully. We show how QPT can reveal rational perspectives to established fallacies (Pothos et al., 2017) and further consider whether QPT can shed light on rational status of behavior in strategic games, in situation when decisions appear inconsistent with the Nash equilibrium or sub game perfect equilibrium.

Heuristics: Hertwig, Pachur, Leuker

Rather than simplify or approximate CPT inference through e.g. more efficient sampling procedures, an alternative, influential possibility is that the mind adopts heuristics. Heuristics can be as accurate and sometimes even more accurate than strategies that employ the greatest possible amount of information and computation. Can such advantages generalize to situations involving interactions with other intelligent, competitive actors? We will explore the effectiveness of heuristics in stationary games against nature and in strategic games and show that heuristics are particularly competitive when the level of epistemic uncertainty is high. We will also consider in general the ecological structures that heuristics can harness, and how theories of heuristics can be integrated with other frameworks of human choice.

Probabilistic language of thought: Tenenbaum, Goodman, Tessler

An important extension to baseline CPT frameworks concerns incorporating language-like properties (such as compositionality), representations, and pragmatic reasoning in probabilistic inference. The probabilistic programming language (PPL) / probabilistic language of thought (PLoT) can more naturally apply to richer forms of reasoning, including everyday reasoning under uncertainty (e.g., Goodman et al., 2015). Furthermore, enriching these models with an understanding of natural language pragmatics can explain apparent fallacies in classical reasoning tasks (e.g., Tessler & Goodman, 2014). Assuming a communicative context to a task involving language allows a reasoner in a PPL/PLoT model to incorporate the goals of a speaker (e.g., assuming the speaker intends to be informative), so providing a rational perspective on reasoning fallacies. We will also consider the way resource limitations guide practical models in PPL.

Discussion: Shiffrin

The discussion part of the symposium will address these varying perspectives on rationality and bring together the themes raised in the presentations. The overarching questions concern what is rationality, and whether 'bounded rational' approaches capture enough of what humans mean by this concept. The discussion will be open to all presenters and the audience.

Acknowledgments

EMP was supported by ONRG grant N62909-19-1-2000.

References

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.

Goodman, N. D., Tenenbaum, J. B., & Gerstenberg, T. (2015). Concepts in probabilistic language of thought. In Eds. E .Margolis & S. Laurence, *New Directions in the Study of Concepts*, pp.623-653. MIT Press: Cambridge.

Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). Rational use of cognitive resources: levels of analysis between the computational and the algorithmic. *Topics in Cognitive Science*, 7, 217-229.

Moore, D. W. (2002). Measuring new types of question-order effects. *Public Opinion Quarterly*, 66(1), 80-91.

Nisbett, R. E., Zukier, H., & Lemley, R. E. (1981). The dilution effect: nondiagnostic information weakens the implications of diagnostic information. *Cognitive Psychology*, 13, 248-277.

Oaksford, M. & Chater, N. (1994). A Rational Analysis of the Selection Task as Optimal Data Selection. *Psychological Review*, 101, 608-631.

Pothos, E. M., Busemeyer, J. R., Shiffrin, R. M., & Yearsley, J. M. (2017). The rational status of quantum cognition. *Journal of Experimental Psychology: General*, 146, 968-987.

Tenenbaum, J. B, Kemp, C., Griffiths, T. L., & Goodman, N. (2011). How to grow a mind: statistics, structure, and abstraction. *Science*, 331, 1279-1285.

Tentori, K., Bonini, N., & Osherson, D. (2004). The conjunction fallacy: a misunderstanding about conjunction? *Cognitive Science*, 28, 467-477.

Tessler, M.H., & Goodman, N. (2014). Some arguments are probably valid: Syllogistic reasoning as communication. In *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjuctive fallacy in probability judgment. *Psychological Review*, 90, 293-315.

Insight and the Genesis of New Ideas

Frédéric Vallée-Tourangeau
(f.vallee-tourangeau@kingston.ac.uk)
Department of Psychology, Kingston University
Kingston upon Thames, UNITED KINGDOM, KT1 2EE

Linden J. Ball
(lball@uclan.ac.uk)
School of Psychology, University of Central Lancashire
Preston, Lancashire, UNITED KINGDOM, PR1 2HE

Keywords: insight, creativity, working memory, phenomenology, first-order problem solving

Research on insight problem solving focuses on the genesis of new ideas and aims to identify underpinning processes that turn an initially unproductive problem representation into one within which the solution offers itself in the agent's mental look-ahead horizon. To address this aim, researchers typically create laboratory-based tasks designed to encourage an incorrect representation of an ostensibly simple problem or riddle such as "how do you throw a ping pong ball in such a way that it travels a certain distance, comes to a dead stop and then reverses direction" (Ansburg & Dominowski, 2000). Such riddles are created to encourage an incorrect interpretation and engender an *impasse*. Researchers can then observe how this impasse is overcome by: (i) examining the phenomenology of insight; (ii) analysing strategic processing (e.g., via protocol analysis); and (iii) exploring brain areas that are active when insight arises (e.g., using neuroimaging).

The current debate in insight research (e.g., Gilhooly, Ball, & Macchi, 2015; Gilhooly & Webb, 2018) pitches the *business-as-usual* view against the *special-processes* view. The latter has roots in Gestalt ideas: insight is the result of a swift change in the way a problem is represented in the mind. The sudden awareness of the solution suggests that insight is not the product of a conscious, incremental, deliberate analysis of the problem helping the agent formulate a solution gradually over time. The 'special' in special processes underscores insight as the product of non-routine cognition largely operating non-consciously (Ohlsson, 2018). If routine cognition, in turn, is in the business of helping an agent plan and solve problems, then the business-as-usual view holds that insight is the product of conscious, deliberate, and incremental effort to solve a problem. From this perspective, a breakthrough may yield a *eureka* moment, but this distinct phenomenological signature does not imply that something other than routine cognition is involved in insight.

Insight research has laboured a fertile ground of methodological and theoretical development in the past 20 years. When the important edited volume by Sternberg and Davidson (1995) was published, research was predicated on a dichotomy whereby problems were deemed to be either *analytic* (e.g., the Tower of Hanoi) or *insight* problems (e.g., the 9-dot problem). This missed the critical point that insight and analysis are underpinning *processes* rather than solution outcomes. Developments in theory (e.g., Weisberg's, 2018, integrated framework) have underscored this point, as has the introduction of new problem types that can be solved either by insight or analysis, as reflected in self-reports (Bowden, Jung-Beeman, Fleck, & Kounios, 1995; Salvi, Costantini, Bricolo, Perugini, & Beeman, 2015; Threadgold, Marsh, & Ball, 2018). Such problems have facilitated investigations of the neural correlates of insight (Abraham, 2018; Kounios & Beeman, 2014) and associated biomarkers (e.g., eye blinks; Salvi, Bricolo, Franconeri, Kounios, & Beeman 2015). Individual differences approaches have also revealed the role of working memory capacity in insight (Chuderski & Jastrzębski, 2018). This symposium will showcase important aspects of current insight research, with presentations by Anna Abraham, Carola Salvi, Ut Na Sio, Margaret Webb, Frédéric Vallée-Tourangeau and Linden Ball (discussant).

Abraham will explore how the study of the brain informs the workings of the human mind as it arrives at insights. Functional neuroimaging studies have revealed key brain regions and networks of relevance, also highlighting the intimate roles played in insight by creative processes such as analogical reasoning and conceptual expansion. EEG studies using event-related potentials indicate a unique neural activity pattern when processing creative associations that are personally experienced as being novel and fitting, as distinct from processing associations that are merely novel or merely fitting. In addition, neuropsychological studies indicate that disruptions at the level of brain structure can both aid and impede creative thinking. The former occurs in contexts where distractibility facilitates creative ideation, a finding indirectly supported by personality-based studies of schizotypy and creativity. These results highlight the value of the neuroscientific approach in advancing an understanding of how creative insights come to pass.

Salvi will present her findings on the "accuracy effect" (i.e., insight solutions have a higher probability of being correct than analytic solutions when tested using convergent thinking problems) and will discuss the model behind this result. The effect is explained by the fact that insight processing yields no partial solution information because of subthreshold processing prior to the suddenly available solution. In contrast, analytic processing can yield better-than-chance guessing based on processing of suprathreshold activation candidates. Further**,** Salvi will present her latest results on the neural correlates and biomarkers associated with insight solutions and the underlying cognitive processes.

Sio will focus on the circumstances that promote creativity. Despite the common belief that distraction will cause productivity loss and that individuals should focus on a single task to achieve optimal performance, recent studies have demonstrated that distraction (e.g., incubation and multitasking) can enhance performance for problems requiring creative thinking. Different potential mechanisms for this distraction effect will be discussed. Sio will also

present findings of recent studies aimed at identifying moderators of the effect to help explain why the positive effect of distraction might emerge and to identify the conditions under which distraction becomes facilitating.

Webb will present research on individual differences associated with insight phenomenology. Investigating individual differences in possible biases in reporting insight is constrained by the "problem of problems", that is, problem-solving skills (e.g., working memory) are required for insight problem solving itself. These individual differences may not be the same as those associated with a bias towards insight experiences. In her recent work, Webb has explored divergent thinking tasks, in which subjective accuracy is high. Participants completed a form of the alternative uses task, reporting on their insight phenomenology ("aha!" experiences) in a trial-wise manner. They were then presented with various solutions to the previous task and also completed a measure of schizotypy (the O-LIFE) to assess whether positive schizotypy (associated with the tendency to perceive meaning in noise) predicted a tendency to report feelings of insight. Findings indicate that generating a use is significantly more likely to result in an "aha!" experience than being presented with a use; positive schizotypy is also a positive predictor of feelings of insight.

Vallée-Tourangeau will outline an ecological perspective on insight, critically reflecting on how insight research often proceeds in the laboratory and how the psychometric methodology validates and reinforces a model of problem solving in which working memory plays a central role. His reflections draw on a distinction between *first-order* versus *second-order* problem solving (Vallée-Tourangeau & March, forthcoming). Research typically assumes the world is represented *inside* a person's head, with mental representations being transformed by rules and operators. It is, therefore, not surprising that individual differences in working memory capacity explain a substantial proportion of the variance in problem solving performance, as working memory underpins a person's ability to construct, maintain and transform mental representations. Crucially, the standard methodology requires participants to think about short vignettes (a few words or sentences) that describe (ambiguously) some state of the world. In other words, participants are not embedded in the *physical* world to solve a problem (first-order problem solving) but are instead processing representations of the world based on abstractions of varying complexity (second-order problem solving). First-order problem solving is impossible as participants cannot interact with a *physical* problem presentation. Second-order problem solving carries a representational toll and, as a result, individual differences in the ability to maintain and transform mental representations—gauged in terms of working memory capacity—correlate with problem solving performance.

References

Abraham, A. (2018). *The neuroscience of creativity*. Cambridge, UK: Cambridge University Press.

Ansburg, P. I., & Dominowski, R. L. (2000). Promoting insightful problem solving. *Journal of Creative Behavior, 34*, 30-60.

Bowden, E., Jung-Beeman, M., Fleck, J., & Kounios, J. (2005). New approaches to demystifying insight. *Trends in Cognitive Sciences*, *9*, 322-328.

Chuderski, A., & Jastrzębski, J. (2018). Much ado about aha!: Insight problem solving is strongly related to working memory capacity and reasoning ability. *Journal of Experimental Psychology: General, 147*, 257-281.

Fleck, J. I., & Weisberg, R. W. (2013). Insight versus analysis: Evidence for diverse methods in problem solving. *Journal of Cognitive Psychology 25,* 436-463.

Gilhooly, K. J., Ball, L. J., & Macchi, L. (2015). Insight and creative thinking processes: Routine and special. *Thinking & Reasoning, 21*, 1-4.

Gilhooly, K., & Webb, M. E. (2018). Working memory and insight problem solving. In F. Vallée-Tourangeau (Ed.), *Insight: On the origins of new ideas* (pp. 105-119). London: Routledge.

Kounios, J., & Beeman, M. (2014). The cognitive neuroscience of insight. *Annual Review of Psychology*, *65*, 71-93.

Ohlsson, S. (2018). The dialectic between routine and creative cognition. In F. Vallée-Tourangeau (Ed.), *Insight: On the origins of new ideas* (pp. 8-27). London: Routledge.

Salvi, C., Costantini, G., Bricolo, E., Perugini, M., & Beeman, M. (2016). Validation of Italian rebus puzzles and compound remote associate problems. *Behavior Research Methods, 48*, 664-685.

Salvi, C., Bricolo, E., Franconeri, S. L., Kounios, J., & Beeman, M. (2015). Sudden insight is associated with shutting out visual inputs. *Psychonomic Bulletin & Review*, *22*, 1814-1819.

Sternberg, R. J., & Davidson, J. E. (Eds) (1995). *The nature of insight*. Cambridge, MA: The MIT Press.

Threadgold, E., Marsh, J. E., & Ball, L. J. (2018). Normative data for 84 UK English rebus puzzles. *Frontiers in Psychology, 9,* 2513, 1-15.

Vallée-Tourangeau, F., & March, P. L. (forthcoming). Insight out: Making creativity visible. *Journal of Creative Behavior.*

Weisberg, R. W. (2018). Insight, problem solving, and creativity: An integration of findings. In F. Vallée-Tourangeau (Ed.), *Insight: On the origins of new ideas* (pp. 191-215). London: Routledge.

Logicist Computational Cognitive Modeling of *Infinitary* False Belief Tasks

Selmer Bringsjord (selmerbringsjord@gmail.com)
Rensselaer AI & Reasoning Lab, Dept of Cognitive Science, RPI; Troy, NY 12180 USA

Naveen Sundar Govindarajulu (naveensundarg@gmail.com)
Rensselaer AI & Reasoning Lab, Dept of Cognitive Science, RPI; Troy, NY 12180 USA

Abstract

We synoptically describe having achieved the unprecedented logicist cognitive computational simulation of quantified versions of any n-level (FBT_n, $\forall n \in \mathbb{N}$) false-belief task, and hence of what we call the *infinitary* false-belief task (FBT_ω); the achievement is enabled by the automated reasoner ShadowProver. Logicist cognitive computational simulation of the level-one (or, as it's currently known, "first-order") false-belief task (FBT_1) was achieved circa 2007 by Bringsjord et al. But subsequently cognitive science has seen the arrival such modeling and simulation successfully applied to the *second-order* false-belief task (FBT_2); see e.g. (Blackburn & Polyanskaya, forthcoming). (This is the level-two FBT in our hierarchy of tasks.) But now, courtesy of what we report, logicist cognitive computational simulation of any FBT_n is accomplished for the first time, and hence the infinitary false-belief task (FBT_ω) is reached as well.

Keywords: logic; cognitive modeling; false-belief task; sally-anne task; infinitary reasoning

The Level-1 and Level-2 False-Belief Tasks Many readers will be familiar with the standard false-belief task (FBT_1; a.k.a. the Sally-Anne task), first introduced by Wimmer and Perner (1983). But to ensure self-containedness we recapitulate: A subject (in an experiment carried out by e), agent a, perceives two agents a_1 and a_2 in front of two boxes b_1 and b_2. Agent a_1 puts an object o into b_1 in plain view of a_2. Agent a_2 then leaves, and in the absence of a_2, a_1 moves o from b_1 into b_2; this movement isn't perceived by a_2. Agent a_2 now returns, and a is asked by the experimenter e: "If a_2 desires to retrieve o, which box will a_2 look in?" If younger than four or five, a will reply "In b_2" (which of course fails the task); after this age subjects respond with the correct "In b_1." While some refer to this task as the "first-order" version of the false-belief task, we refer to it as the "level-one" version of the task.[1]

Table 1 lists some of the key epistemic propositions that hold of FBT_1^P after the switch happens, paired with their obvious symbolizations in our multi-operator quantified cognitive calculus used for handling false-belief tasks. We use the superscript 'P' to indicate that the task in question is passed; we reserve superscript 'F' to indicate that the task is failed.

[1]Use of the locution "n-order" is quite infelicitous, because this locution is long established in formal logic as a way to pick out the expressive power of extensional logics within a hierarchy of them. For instance, there is first-order logic, second-order logic, and so on. Since which of these logics is used to model and simulate a given false-belief task is a key parameter in the logicist modeling in question, we judge it to be wise to refer to such tasks at a given *level*, not an order, so as to avoid confusion that will otherwise obtain.

Table 1: Table for Level-1 (L1) FBT = FBT_1

Label	English Declarative Content	Formula
L1.1	a_1 believes a_2 believes o is in b_1.	$\mathbf{B}_{a_1}\mathbf{B}_{a_2}I(b_1)$
L1.2	a_1 believes o is in b_1.	$\mathbf{B}_{a_1}I(b_1)$
L1.3	a believes a_2 believes o is in b_1.	$\mathbf{B}_{a}\mathbf{B}_{a_2}I(b_1)$
L1.4	a bel. a_1 bel. a_2 bel. o is in b_1.	$\mathbf{B}_{a}\mathbf{B}_{a_1}\mathbf{B}_{a_2}I(b_1)$

The level-two (or "second-order") FBT is easily captured, as follows.[2] First, when agent a_2 leaves, he/she secretly perceives a_1 move o to box b_2. Formally, the key adjustment is an *addition* to (adjustments of) the lines seen in Table 1: e.g.

L2 a_2 believes a_1 believes a_2 believes o is in b_1.

Prior Relevant Achievements Circa 2007, cognition associated with the *false-belief task* (FBT_1, including both FBT_1^P and FBT_1^F) was modeled in formal logic expressive enough to handle quantification, and computationally simulated (Arkoudas & Bringsjord, 2008, 2009).[3] This type of research falls under what Bringsjord (2008) calls *logicist computational cognitive modeling* (LCCM). As far as we are aware, this work in 2007 marks the first robust logicist modeling and simulation of both passing and failing cognition in FBT.[4] Here is the crucial takeaway from study of prior work: No one, before now, has achieved logicist computational cognitive modeling of quantified false-belief tasks at level 3, 4, ..., even in the non-quantificational case; and no one has reached the infinitary case.

Level-k ($k \geq 3$) False-Belief Tasks In the level-three false-belief task, agent a_1 secretly views a_2's secretly viewing into the room from outside it. (All of this is easily visualized with help from iterated, hidden cameras that feed information to the agents. Because of space limitations we forego visual depictions.) For FBT_3, the characteristic formula is:

$$\mathbf{B}_{a_1}\mathbf{B}_{a_2}\mathbf{B}_{a_1}\mathbf{B}_{a_2}I(b_1) \tag{1}$$

[2]A nice place to start reviewing the literature on FBT_2 is (Baron-Cohen, O'Riordan, Stone, Jones, & Plaisted, 1999), which has complete references to the earliest introduction of FBT_1 and FBT_2 in the (empirical) literature. (In this regard, we certainly recommend that interested readers review (Perner & Wimmer, 1985).) There is no discussion in this literature of level-3-and-above FBTs, let alone of infinitary FBTs such as FBT_ω; and we haven't found any formal/mathematical literature on these more demanding FBTs either.

[3]While formal but certainly declarative, very impressive computational cognitive modeling of FBT_1 was achieved earlier by Wahl and Spada (2000). Stenning and van Lambalgen (2008) provide informal declarative notation for modeling false belief, but have no implementation/simulation.

[4]Bello, Bignoli, and Cassimatis (2007), as in the aforecited (Wahl & Spada, 2000), achieve computational cognitive cognitive modeling of FBT_1 that makes use of declarative representations, but not of any logics.

Quantified False-Belief Tasks The sub-formulae $I(b_n)(n \in \{1,2\}$ is expressible within the formal language of only the propositional calculus. If instead of a single object o being used, a given FBT involves a *group G* of, say, n objects, then the correlate to this sub-formula will require the machinery of at least the quantificational machinery of first-order logic. We are able to model and computationally simulate in this more demanding case, so that even if subjects have beliefs about a quantity m from G ($m \leq n$) being placed in the box, their cognition can't be captured.

FBT_ω: An Infinitary Quantified False-Belief Task Our inference system leverages a computable version of an infinitary inference rule to prove FBT_ω given that we can prove FBT_n $\forall n$ (N. Govindarajulu, Licato, & Bringsjord, 2013).[5]

Automation We use an automated reasoning system, ShadowProver, to model FBT_n and FBT_ω. ShadowProver is a quantified modal logic theorem prover that has been used to model, in LCCM fashion, intricate reasoning tasks, e.g. ethical reasoning in (N. Govindarajulu & Bringsjord, 2017; N. S. Govindarajulu, Bringsjord, Ghosh, & Peveler, Forthcoming in 2019) and self-consciousness in (Bringsjord, Licato, Govindarajulu, Ghosh, & Sen, 2015). Since characteristic statements for FBT_n and FBT_ω are structurally similar to common knowledge, we leverage ShadowProver's ability to use the operator ($\mathbf{C}$) for such knowledge.[6]

Objections We mention here only that while it might be objected that humans have trouble with even third-order belief, many of our college-level subjects on the contrary have little trouble proving correct answers for any FBT_n.

Acknowledgments The "late-breaking" achievements described herein have been enabled by generous support from ONR and AFOSR, for which we are deeply grateful.

References

Arkoudas, K., & Bringsjord, S. (2008). Toward Formalizing Common-Sense Psychology: An Analysis of the False-Belief Task. In T.-B. Ho & Z.-H. Zhou (Eds.), *Proceedings of the Tenth Pacific Rim International Conference on Artificial Intelligence (PRICAI 2008)* (pp. 17–29). Springer-Verlag. Retrieved from http://kryten.mm.rpi.edu/KA_SB_PRICAI08_AI_off.pdf

Arkoudas, K., & Bringsjord, S. (2009). Propositional Attitudes and Causation. *International Journal of Software and Informatics*, *3*(1), 47–65. Retrieved from [7]

Baron-Cohen, S., O'Riordan, M., Stone, V., Jones, R., & Plaisted, K. (1999). Recognition of faux pas by Normally Developing Children and Children with Asperger Syndrome or High-Functioning Autism. *Journal of Autism and Developmental Disorders*, *29*, 407–418.

Bello, P., Bignoli, P., & Cassimatis, N. (2007). Attention and Association Explain the Emergence of Reasoning About False Belief in Young Children. In *Proceedings of the 8th international conference on cognitive modeling* (pp. 169–174). Ann Arbor, MI: University of Michigan.

Blackburn, T. B. P., & Polyanskaya, I. (forthcoming). Being Deceived: Information Asymmetry in Second-Order False Belief Tasks. *topiCS*.

Bringsjord, S. (2008). Declarative/Logic-Based Cognitive Modeling. In R. Sun (Ed.), *The Handbook of Computational Psychology* (pp. 127–169). Cambridge, UK: Cambridge University Press. Retrieved from [8]

Bringsjord, S., Licato, J., Govindarajulu, N., Ghosh, R., & Sen, A. (2015). Real Robots that Pass Tests of Self-Consciousness. In *Proccedings of the 24th ieee international symposium on robot and human interactive communication (ro-man 2015)* (p. 498-504). New York, NY: IEEE.

Govindarajulu, N., & Bringsjord, S. (2017). On Automating the Doctrine of Double Effect. In C. Sierra (Ed.), *Proceedings of the twenty-sixth international joint conference on artificial intelligence (IJCAI-17)* (pp. 4722–4730). International Joint Conferences on Artificial Intelligence. Retrieved from https://doi.org/10.24963/ijcai.2017/658

Govindarajulu, N., Licato, J., & Bringsjord, S. (2013). Small Steps Toward Hypercomputation via Infinitary Machine Proof Verification and Proof Generation. In M. Giancarlo, A. Dennuzio, L. Manzoni, & A. Porreca (Eds.), *Unconventional computation and natural computation; lecture notes in computer science; volume 7956* (pp. 102–112). Berlin, Germany: Springer-Verlag.

Govindarajulu, N. S., Bringsjord, S., Ghosh, R., & Peveler, M. (Forthcoming in 2019). *Beyond The Doctrine Of Double Effect: A Formal Model of True Self-Sacrifice.* (Robots and Well Being: Proceedings of the International Conference on Robot Ethics and Safety Standards 2017)

Perner, J., & Wimmer, H. (1985). "John thinks that Mary thinks that ...": Attribution of Second-Order Beliefs by 5–10-Year-Old Children. *Journal of Experimental Child Psychology*, *39*, 437–471.

Stenning, K., & van Lambalgen, M. (2008). *Human Reasoning and Cognitive Science*. Cambridge, MA: MIT Press.

Wahl, S., & Spada, H. (2000). Children's Reasoning About Intentions, Beliefs and Behaviour. *Cognitive Science Quarterly*, *1*(1), 3–32.

Wimmer, H., & Perner, J. (1983). Beliefs about Beliefs: Representation and Constraining Function of Wrong Beliefs in Young Children's Understanding of Deception. *Cognition*, *13*, 103–128.

[5] Some infinitary logics allow various uncountably infinite elements, and certainly from a purely logico-mathematical perspective there's no reason why false-belief tasks involving such elements can't be specified, but so far we have only explored, and met with success on, *countably* infinite false-belief tasks.

[6] $\mathbf{C}\phi \equiv$ 'All agents $\mathbf{K}\phi$, all agents $\mathbf{K}$ that all agents $\mathbf{K}\phi$, ...

[7] http://kryten.mm.rpi.edu/PRICAI_w_sequentcalc_041709.pdf

[8] http://kryten.mm.rpi.edu/sb_lccm_ab-toc_031607.pdf

PUBLICATION-BASED PRESENTATION: Modeling Human Creative Cognition using AI Techniques

Steve DiPaola
School of Interactive Arts and Technology, Simon Fraser University
250-13450 102nd Avenue, Surrey, B.C., Canada V3T 0A3 sdipaola@sfu.ca

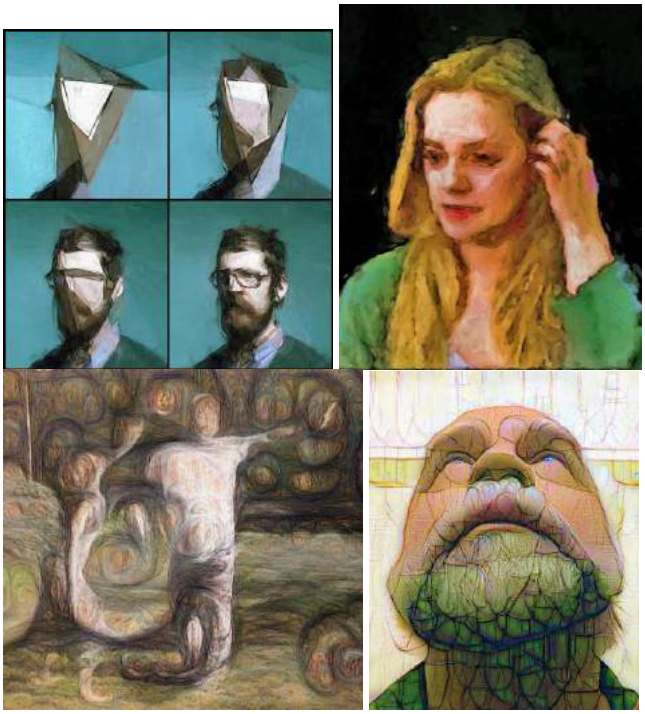
Keywords: computational creativity; fine art painting; creativity; empathy; artificial intelligence; deep learning; evolutionary programming

DiPaola's research endeavors to build top down Artificial Intelligence (AI) models of human creativity, empathy and expression for both use in new forms of computation systems as well as analysis of how the creative mind works. In doing so he has interviewed hundreds of artists, writers and musicians on how they perceive their creative talent and its originals. Combined with research from neuro-aesthetics and computer modelling, DiPaola notes that while many creative individuals report that they believe new insights as coming into them from an external source during creative flow, that evidence point to these new creative ideas and interpretations often more likely have internal roots from the individual's, mid and long term past experiences and processes. DiPaola attempts to model this and other human creativity processes in computational form often as AI systems such as deep learning, reinforcement learning and evolution programming. Two efforts underway in DiPaola's research lab are mapping out the creative process of a fine art portrait painter using 5 hierarchical AI systems, as well as modelling an empathetic embodied character agent who can understand emotions from those she talks with and construct creative narrative or quote like responses.

The common view that our creativity is what makes us uniquely human suggests that incorporating research on human creativity into Artificial Intelligence (AI) based generative deep learning techniques might be a fruitful avenue for making their outputs more compelling and human-like, especially in arts such as the creative arts. Using our labs original AI systems such as our deep learning convolutional neural networks and cognitive based computational art rendering systems, we attempt to show how human creativity can be implemented/modelled computationally, and demonstrate their impact on the resulting digital generative art. Conversely, he will discuss how explorations in creativity AI can inform our understanding of human creativity and its foundations.

DiPaola will discuss and demonstrate his lab's approach (ivizlab.sfu.ca) to cognitively modeling a fine art painter process by integrating Deep Learning AI with novel computational novel NPR approaches. This interdisciplinary (cognitive science / arts / AI) work brings cognitive creative fields together with Deep Learning neural networks. DiPaola will demonstrate and discuss the lab new work as well as the applications spaces in interactive arts, health and a recent Google / Knight Foundation granted project using creative painterly emulation as a new approach to anonymize interviewees in documentary videos where the study data shows improvement to overall empathy and engagement compared to current techniques.

Steve DiPaola, past Director of the Cognitive Science Program at Simon Fraser University (SFU), is currently is a Professor and lab director of the iVizLab, a PhD based lab on Artificial Intelligence using human cognition theories of creativity, empathy and expression. He came to SFU from Stanford University where some of his creative AI systems were used in generative game creation including the best-selling game of that year, "The Sims". DiPaola has over 100+ peer reviewed papers in AI/cognition and $2 million in past/current funding in AI related areas of cognitive creativity and expression. As both a scientist and artist, DiPaola has written code for his AI "creative on its own" artworks that has been shown in major galleries and museums including The Whitney, The Smithsonian, Tate, and gallery's in NYC, London and LA.



DiPaola Papers

Journals:

- DiPaola S, McCaig G, Gabora L, (2018). Informing Artificial Intelligence Generative Techniques using Cognitive Theories of Human Creativity. Procedia Computer Science. Special Issue: Bio Inspired Cognitive Arch. Vol 145 pp 158-168.
- DiPaola S, (2017) Exploring the Cognitive Correlates of Artistic Practice Using a Parameterized Non-Photorealistic Toolki", Leonardo, Winner of 2016 LABS Leonardo Award. Vol. 50, pp 531-452.
- Shakeri H, Nixon M, & DiPaola S, (2017) Saliency-Based Artistic Abstraction with Deep Learning and Regression Trees, Journal of Imaging Science and Technology, Vol 61, No 6, pp. 60402-1-60402-9(9), 2017.
- DiPaola S, (2014)Using a Contextual Focus Model for an Automatic Creativity Algorithm to Generate Art Work, Procedia Computer Science. Special Issue: Bio Inspired Cognitive Architectures, Vol 41, pp. 212-219..
- DiPaola S, Riebe C, Enns J T, (2013) Following the masters: Portrait viewing and appreciation is guided by selective detail, Perception, Vol 42, No 6, pp 608–630.
- DiPaola S, Riebe C, Enns J, (2010). Rembrandt's Textural Agency: A Shared Perspective in Visual Art and Science", Leonardo, Vol 43, No 3, pp 145-151.
- Riebe C, DiPaola S., & Enns J, (2009). Following The Masters: Viewer Gaze is Directed by Relative Detail in Painted Portraits, Journal of Vision, Vol 9, No 8, pp 368-368.
- DiPaola S, Gabora L, (2009). Incorporating Characteristics of Human Creativity into an Evolutionary Art Algorithm, Genetic Programming and Evolvable Machines Journal, Vol 10, No 2, pp 97-110.

Conference Papers

- Feldman S, Yalcin ON, DiPaola S, (2017). Engagement with artificial intelligence through natural interaction models, Proc: Electronic Visualisation and the Arts, British Computer Society, 296-303.
- McCaig R, DiPaola S, Gabora L, (2016). Deep Convolutional Networks as Models of Generalization and Blending Within Visual Creativity, Proceedings of International Conference on Computational Creativity, 8 pages.
- DiPaola S, McCaig R, (2016). Using Artificial Intelligence Techniques to Emulate the Creativity of a Portrait Painter, Proceedings of Electronic Visualisation and the Arts, British Computer Society, 8 pages, London.
- Choi S K, DiPaola S, (2015). Touch of the Eye: Does Observation Reflect Haptic Metaphors in Art Drawing?, Proceedings of ACM Conf on Human Factors in Computing Systems (CHI '15), pp 579-588.
- Salevati S, DiPaola S, (2015). A Creative Artificial Intelligence System to Investigate User Experience, Affect, Emotion and Creativity, Proceedings of Electronic Visualisation and the Arts, British Computer Society, 8 pages, London.
- DiPaola S, (2014). Computer Modelling Fine Art Painting using a Cognitive Correlative Heuristics Approach, Proceedings of Biologically Inspired Cognitive Architectures. 5 pages. MIT, MA.
- Salevati M, DiPaola S, (2014). Using a Creative Evolutionary System for Experiencing the Art of Futurism, Proceedings of Electronic Visualisation and the Arts, Florence, Italy, 8 pages.
- Choi S K, DiPaola S, (2013). How a Painter Paints: An Interdisciplinary Understanding of Embodied Creativity, Proceedings of Electronic Visualisation and the Arts, pp. 127-134. British Computer Society, London.
- DiPaola S, Smith A, (2013). Interactively Exploring Picasso's Multi-dimensional Creative Process in Producing Guernica, Proceedings of Electronic Visualisation and the Arts, pp. 25-31. British Computer Society, London.
- Gabora L, DiPaola S, (2012). How Did Humans Become So Creative? A Computational Approach, Proceedings of International Conference on Computational Creativity, pp 203-211.
- DiPaola S, Smith A, (2012). Formalizing An Interconnected Syntax For Picasso's Creative Process In Producing Guernica", Proceedings of Conceptual Structure, Discourse and Language, 6 pages.
- Choi S K, DiPaola, S, Schiphorst T, 2012. The Tacit And The Trace: Towards Syntax Of The Creative Act, Proceedings of Conceptual Structure, Discourse and Language, 6 pages.
- DiPaola S, (2009). Quantifying artist's use of human vision constructs to influence viewer eye gaze,", In Proc: SPIE Human Vision and Imaging, Int. Society for Optical Engineering, 6 pages.
- DiPaola S, (2008). "The Trace and the Gaze: Textural Agency in Rembrandt's Late Portraiture from a Vision Science Perspective", Proceedings of Electronic Visualisation and the Arts, 8 pages, London.
- DiPaola S, Gabora L, (2007). Incorporating Characteristics of Human Creativity into an Evolutionary Art Algorithm", In Proceedings of the 2007 GECCO Conference Companion on Genetic and Evolutionary Computation (London,, July 07 - 11, 2007). GECCO '07, pp 2450-2456., ACM, New York, NY.
- DiPaola S, (2007). A Knowledge Based Approach to Modeling Portrait Painting Methodology, Proceedings of Electronic Visualisation and the Arts, 10 pages, London.
- DiPaola S, (2007). Painterly Rendered Portraits from Photographs using a Knowledge-Based Approach", In Proc: SPIE Human Vision and Imaging, Int. Society for Optical Engineering, Keynote paper. pp 33-43.
- DiPaola S, (2006). Evolving Portrait Painter Programs using Genetic Programming to Explore Computer Creativity", Proceedings of iDMAa Conference (International Digital Media and Arts Association), 7 pages.
- DiPaola S, (2005). Evolving Creative Portrait Painter Programs Using Darwinian Techniques with an Automatic Fitness Function", Proceedings of Electronic Visualisation and the Arts, 10 pages, London. July.

A Cultural Evolution Framework for Human Creativity

Liane Gabora (liane.gabora@ubc.ca)
Department of Psychology, University of British Columbia, Kelowna BC, V1V 1V7, CANADA

Keywords: concepts; convergent thinking; contextual focus; creativity; cultural evolution; divergent thinking; representational redescription; self-organized criticality

Introduction: Honing Theory of Creativity

Other species perceive, make decisions, take action, and even create. However, our species is exceptional with respect to our predilection to adapt ideas to our own needs, tastes, and perspectives, and express ourselves through language, technology, art, and other means. I will present ongoing theoretical and empirical research on how the creative process works and how human creativity evolved. What makes this research program unique is that it examines creativity from the perspective of its role in fueling the evolution of culture, and includes both studies with human participants and computational models.

Creativity research has emphasized the generation of multiple ideas over *honing*—recursively reflecting on a question or idea by viewing it from different perspectives (Gabora & Kauffman, 2016; Gabora, 2017). Just as a single object may cast separate shadows when lit from different directions, the mental representation of a creative work-in-progress may be a single entity with the potentiality to be articulated as different prototypes, sketches, or story ideas.

Honing does not encompass additions or modifications to an idea that are tacked on willy-nilly; it refers specifically to modifications that arise in response to an overarching conceptual framework that is shepherding[1] the creative process. The structure of this overarching framework reflects the individual's *worldview:* their self-organizing web of understandings about their world and their place in that world (in other words, the creator's mind as experienced 'from the inside').

The term *psychological entropy* has been used to refer to arousal-provoking uncertainty, which can be experienced not just negatively as anxiety but also positively as a wellspring for creativity (or both) (Gabora, 2017). It is proposed that psychological entropy—a macro-level variable acting at the level of the worldview as a whole—generates emotions that play a role in guiding and monitoring creative tasks. Thus, honing continues until psychological entropy decreases to an acceptable level. In Piagetian terms, during honing the individual assimilates each new understanding of the idea, and the individual's worldview changes to accommodate this new understanding. Insight is then explained in terms of *self-organized criticality* (SOC) (Gabora, 2017), a phenomenon wherein, through simple local interactions, complex systems tend to find a critical state poised at the cusp of a transition between order and chaos, from which a single small perturbation occasionally exerts a disproportionately large effect. Thus, while most thoughts have little effect on one's worldview, an idea we call *insightful* is one for which one thought triggers another, which triggers another, and so forth in a avalanche of conceptual change.

Convergent thought has been defined and measured in terms of the ability to perform on tasks where there is a single correct solution, and *divergent thought* in terms of the ability to generate multiple different solutions. I will explain why *honing theory* (HT) leads us to redefine convergent thought as thought in which the relevant concepts are considered from *conventional contexts,* and divergent thought as thought in which they are considered from *unconventional contexts* (Gabora, 2018).

Implications for Cultural Evolution Theory

I propose that creativity fuels worldview transformation, and that worldviews are what evolve through culture, in a piecemeal fashion, through a process of *Self-Other Reorganization* (SOR) involving (internal) self-organization and (external) interaction with other worldviews (Gabora, 1999, 2013, 2019). SOR solves dilemmas associated with the high degree of human cooperation (Voorhees, Read, & Gabora, in press), which enables the cumulative building of ideas on one another. I will present a set of agent-based model experiments which show, in different ways, that the effectiveness of this cumulative building depends on the balance between continuity (via imitation) and novelty (via creativity) (Gabora & Tseng, 2017).

I propose that creative outputs merely provide evidence concerning the evolutionary states of worldviews (just as shadows provide evidence concerning the shape casting the shadow). This stands in contrast to the traditional view that behaviors, artifacts, or memes, are the objects of cultural evolution, i.e., they are *what* evolves through culture.

Cross-Domain Influence

The view that it is worldviews that evolve through culture follows naturally from studies of *cross-domain influence,* wherein a creative output in one domain (e.g., art) is influenced by another domain (e.g., music). I will report on a set of studies in which creative individuals in multiple disciplines were asked to list as many influences on their creative work as they could. Results indicate that cross-domain influences are surprisingly ubiquitous, particularly in the arts, where they appear to be even more widespread than within-domain influences (Scotney, Weissmeyer, & Gabora, 2018). The discontinuities in cultural lineages that result from cross-domain influence (e.g., Led Zeppelin's use of Tolkien's *Lord of the Rings* as inspiration for the song

[1] This word is chosen deliberately because it implies that the process is neither entirely top-down nor entirely bottom-up.

"Battle of Evermore") are difficult to account for without resorting to the view that it is not the outputs themselves but the worldviews generating them that evolve through culture.

The Origins of Creative, Cultural Evolution

Like the origin of life, the origin of the kind of integrated worldview needed for cultural evolution has been modeled using an *autocatalytic framework* (Gabora & Steel, 2017). In an autocatalytic network, for each component there exists a means to catalyze the reaction that generates it. Although no component can catalyze its own formation, the network of components as a whole is collectively autocatalytic. In culture, the role of catalysis is played by association and reminding events, and the 'reactions' are between, not catalytic molecules, but concepts and ideas. As parents and others share knowledge with children, an integrated understanding of the world takes shape in their minds, such that they become able to reframe new information in terms of existing mental structure, and become themselves creative contributors to cultural evolution.

I propose that two key steps toward cognitive modernity were (1) onset of *representational redescription* (RR) in *Homo erectus* 2 MYA, and (2) onset in the Middle/Upper Paleolithic of *contextual focus* (CF): the ability to shift between convergent and divergent modes of thought (Gabora & Smith, 2018). In terms of the autocatalytic model, representational redescription entails an interaction or 'catalysis event' between different representations or perspectives, and CF entails the capacity to vary the 'reactivity' of the network. CF may have originated with mutation of the FOXP2 gene, which is known to have undergone human-specific mutations in the Paleolithic (Gabora & Smith, 2019). FOXP2, once thought to be the "language gene", is not uniquely associated with language. In its modern form, FOXP2 may have enabled fine-tuning of the neurological mechanisms underlying the capacity to shift between convergent and divergent processing modes by varying the size of the activated region of memory.

Computer-generated Art and Music

Finally, I will discuss ongoing applications of HT to the development of computer-generated art and music (Bell & Gabora, 2016; DiPaola, & Gabora, & McCaig, 2018; McCaig, DiPaola, & Gabora, 2016). I will show how such efforts are useful for bringing to light the strengths and limitations of our understanding of the creative process.

Acknowledgments

This work was supported by grant 62R06523 from the Natural Sciences and Engineering Research Council of Canada.

References

Aerts, D., Gabora, L., & Sozzo, S. (2013). Concepts and their dynamics: A quantum theoretical model. *Topics in Cognitive Science*, *5*, 737–772.

Aerts, D., Broekaert, J., Gabora, L., & Sozzo, S. (2016). Generalizing prototype theory: A formal quantum framework. *Frontiers in Psychology* (Cognition), *7*(418).

Bell, S. & Gabora, L. (2016). A music-generating system based on network theory. In *Proceedings of the seventh international conference on computational creativity*. Palo Alto: AAAI Press.

DiPaola, S., & Gabora, L. & McCaig, G. (2018). Informing artificial intelligence generative techniques using cognitive theories of human creativity. *Procedia Computer Science, 145,* 158–168.

Gabora, L. (1999). Weaving, bending, patching, mending the fabric of reality: A cognitive science perspective on worldview inconsistency. *Foundations of Science, 3*, 395–428.

Gabora, L. (2013). An evolutionary framework for culture: Selectionism versus communal exchange. *Physics of Life Reviews, 10*, 117–145.

Gabora, L. (2017). Honing theory: A complex systems framework for creativity. *Nonlinear Dynamics, Psychology, and Life Sciences, 21*, 35–88.

Gabora, L. (2018). The neural basis and evolution of divergent and convergent thought. In O. Vartanian & R. Jung (Eds.) *The Cambridge handbook of the neuroscience of creativity*. Cambridge: Cambridge University Press.

Gabora, L. (2019). Creativity: Linchpin in the quest for a viable theory of cultural evolution. *Current Opinion in Behavioral Sciences, 27,* 77–83.

Gabora, L. & Kauffman, S. (2016). Toward an evolutionary-predictive foundation for creativity. *Psychonomic Bulletin & Review, 23*, 632–639.

Gabora, L., & Tseng, S. (2017). The social benefits of balancing creativity and imitation: Evidence from an agent-based model. *Psychology of Aesthetics, Creativity, and the Arts*, *11*, 457–473.

Gabora, L. & Smith, C. (2018). Two cognitive transitions underlying the capacity for cultural evolution. *Journal of Anthropological Sciences, 96,* 1–26.

Gabora, L. & Smith, C. (2019). Exploring the psychological basis for transitions in the archaeological record. In: T. Henley, E. Kardas, & M. Rossano (Eds.) *Handbook of cognitive archaeology*. Routledge / Taylor & Francis.

Gabora, L., & Steel, M. (2017). Autocatalytic networks in cognition and the origin of culture. *Journal of Theoretical Biology, 431,* 87–95.

McCaig, G, DiPaola, S., & Gabora, L. (2016). Deep convolutional networks as models of generalization and blending within visual creativity. In *Proceedings of the seventh international conference on computational creativity* (pp. 156–163). Palo Alto: AAAI Press.

Scotney, V., Weissmeyer, S., & Gabora, L. (2018). Cross-domain influences on creative processes and products. *Proceedings of the 40th meeting of the cognitive science society* (pp. 2452-2457). Austin TX: Cog Science Society.

Voorhees, B., Read, D., & Gabora, L. (in press). Identity, kinship, and the evolution of cooperation. *Current Anthropology*.

From Design Cognition to Design Neurocognition

John S Gero (john@johngero.com)
Department of Computer Science and School of Architecture, University of North Carolina at Charlotte
Charlotte, NC 28223 USA

Keywords: design cognition, design neurocognition, protocol analysis, EEG, fNIRS

Introduction

Design is one of the most profound acts of humans and is the way in which we intentionally change both the physical and virtual worlds around us. Design is mentioned in the earliest extant writings of humans. It appears in *The Epic of Gilgamesh,* which dates back over 4,000 years. The first mention of design appears around the same time as the earliest writings about mathematics, philosophy and science. Design is one of the ways a society increases its economic and social wealth. Given its longevity it is surprising that the formal study of design dates back only to the twentieth century. The scientific study of design, design science, commenced only about 60 years ago.

In English the word "design" is used both as a noun and a verb and its use is disambiguated by its context. We will, in general, use the word "design" to mean the outcome and "designing" to mean the process of producing a design.

There are many designers and teachers of designing who claim that designing cannot be studied scientifically since its results are not reproducible. Whilst designs can be studied what we are interested in when studying designing are the processes that go to make up the acts of designing. It is assumed that there is some regularity exhibited by those processes and it is those processes and that regularity that is being studied. The scientific study of designing borrows it methods directly from the scientific method. It carries out controlled experiments in laboratories and in-situ studies in the field.

Designing was initially studied within the framework of information processing before moving to an artificial intelligence frame. However, when designing was treated as cognitive processes, it used the frame of cognitive science and the field of research became known as "design cognition".

The talk will present recent advances in the study of design cognition and the extension of those studies into the study of brain behavior while designing – "design neurocognition" in the Gero lab. The Gero lab is a disaggregated lab with projects in locations in multiple countries including Australia, Croatia, France, Italy, Sweden, Switzerland and the USA.

Design Cognition Through Protocol Analysis

Protocol analysis (Ericsson & Simon, 1993) has become the preferred research method for the elicitation of design cognition. Around it a range of analysis methods have been developed (Kan & Gero, 2017) that form the basis of new results. The results presented in the talk are derived from a newly developed model of co-design in teams by Gero & Milovanovic (unpublished) based on the situated version of the FBS ontology, sFBS, (Gero & Kannengiesser, 2014). The model provides for fine grained behavior of individuals in teams.

Results from a protocol study of cohorts of two-person homogeneous and heterogeneous teams, where the heterogeneity is due to gender, are presented in Figure 1 (Milovanovic & Gero, submitted). The cohorts were undergraduate mechanical engineering students at a state university in Utah and were given the same design task. In Figure 1, each ellipse contains the sFBS behavior of team members, where the top ellipses represent team member A activations and the bottom ellipses represent team member B activations. For a detailed development of the situated Function-Behavior-Structure ontology consult Gero & Kannengiesser (2004). The links between the activation variables are a measure of the cumulative occurrences of cognitive design processes. The variables outside the team members' individual spaces are externalizations in the forms of verbalizations, sketches or gestures. The externalization of thought through verbalization, gestures and sketching provides the basis for co-designing. The Gero & Milovanovic (submitted) model of co-design uses the notion that co-design occurs when designers cross the externalization boundary.

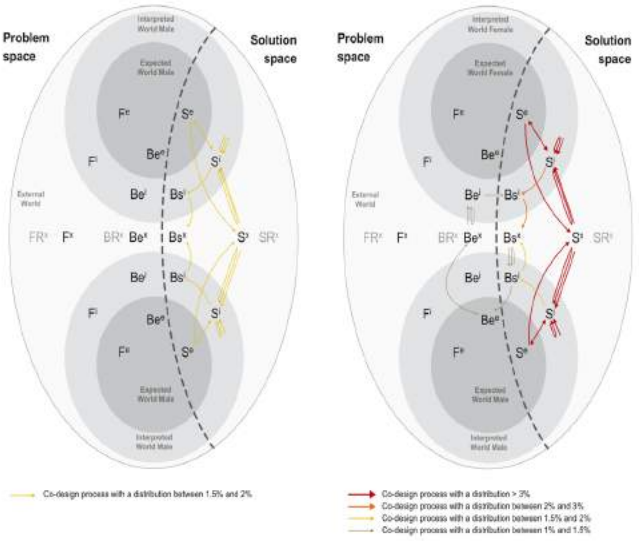
The results in Figure 1 show that heterogeneous teams containing one female and one male member exhibit more co-design processes than do homogeneous all-male teams. Further, such mixed-gender teams distribute more of their cognitive effort between the problem and the solution than do all-male teams, who expend more of their cognitive effort on the solution.

The presentation will show results of studying the design cognition of students and tutors in a studio pedagogy setting. It will present the change in student-student design cognition interaction over multiple studio sessions.

From Design Cognition to Design Neurocognition

The drop in the cost of non-invasive brain measurement has opened avenues of research into design neurocognition. In particular EEG and fNIRS, which collect temporal data, are both well suited for design neurocognition studies since design is a temporal activity. fMRI is less suited to study the

temporal behavior of designing. It is well suited where high spatial resolution is required.



(a) (b)

Figure 1: (a) Dominant sFBS co-design processes for homogeneous, all-male, teams; (b) dominant sFBS co-design processes for heterogeneous, mixed-gender, teams (Milovanovic & Gero, to appear).

The presentation will report on using a 14 channel EEG block experiment to measure the effect of design task on brain behavior. The tasks range from highly constrained to unconstrained. The total task related power of measured signals is presented in Figure 2 for the pre-task and the four design tasks for 58 participants covering multiple domains. Results for individual domains indicate significant differences due to domain and task.

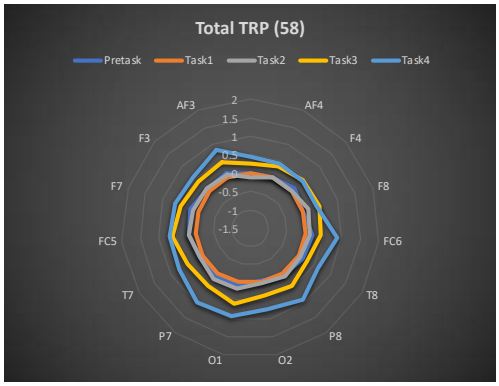


Figure 2: Total TRP for each of 14 channels across all participants for Pre-task, Task 1, Task 2, Task 3 and Task 4 (Vieira, Gero, et al, unpublished data).

While EEG measures electrical signals at the surface of the brain with high temporal resolution, functional near infrared spectroscopy (fNIRS) measures BOLD demand with medium temporal resolution. The presentation will report on an fNIRS experiment that repeats a previous protocol study for which we have cognitive results. The results of dominant hemisphere activation over time are presented in Figure 3 showing an unexpected pattern of behavior. Additional results cover other concept generation techniques.

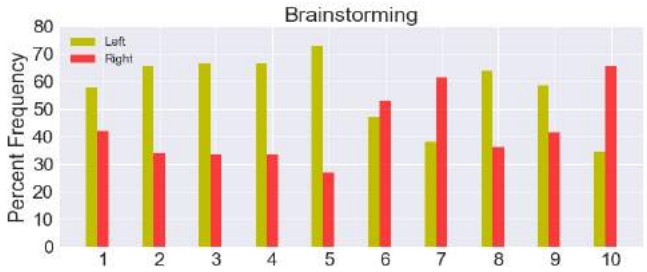


Figure 3: Percent frequency across time deciles of dominant hemisphere during brainstorming (Shealy & Gero, submitted).

These three exemplary results from there different measuring approaches, protocol analysis, EEG and fNIRS, demonstrate the expanding capacity to measure design cognition through measurement of the mind and indirectly through measurement of the brain. Until recently, only measurement of the cognition through the behavior of the mind was reliably available. The development of relatively inexpensive tools for non-invasive brain measurement has opened novel approaches to the measurement of design neurocognition. Bringing cognitive studies of the mind and neurocognitive studies of the brain together offers opportunities to both increase our understanding of designing and to provide the foundation for the development of tools to aid designing and the development of curricula to improve design education.

Acknowledgments

The author wishes to acknowledge the researchers who have collaborated on the projects alluded to in this abstract. These include, in particular, Julie Milovanovic (UMR AAU-CRENAU, Graduate School of Architecture and Ecole Centrale Nantes, France), Sonia Vieira (INEGI, Institute of Science and Innovation in Mechanical and Industrial Engineering, Porto, Portugal) and Tripp Shealy (The Charles E. Via, Jr. Department of Civil & Environmental Engineering, Virginia Tech, USA). The author wishes to acknowledge the National Science Foundation Grant Nos. CMMI-1161715, EEC-1160345, CMMI-1400466, EEC-1463873 and CMMI-1762415.

References

Ericsson, K. A. & Simon, H. A. (1993). *Protocol Analysis; Verbal Reports as Data*. Cambridge: MIT Press.

Gero, J. S. & Kannengiesser, U. (2004). The situated Function-Behaviour-Structure framework, *Design Studies, 25(4),* 373-391.

Gero, J. S. & Kannengiesser, U. (2014). The Function-Behaviour-Structure ontology of design. In A. Chakrabarti

& L. Blessing (Eds), *An Anthology of Theories and Models of Design*. London: Springer.

Gero, J. S. & Milovanovic, J. (submitted). The situated Function-Behavior-Structure co-design model.

Kan, W. T. & Gero, J. S. (2017). *Quantitative Methods for Studying Design Protocols*. Dordrecht: Springer.

Milovanovic, J. & Gero, J. S. (to appear). Exploration of gender diversity effects on design team dynamics, *Human Behavior in Design Conference,* Tutzing, Germany, 23-24 April 2019.

Shealy, T. & Gero, J. S. (submitted). The neurocognition of three engineering concept generation techniques.

Vieira S., Gero J. S., Delmoral J., Gattol V., Fernandes, C., Parente, M. & Fernandes, A. (to appear). Insights from an EEG study of mechanical engineers problem solving and designing, *Human Behavior in Design Conference,* Tutzing, Germany, 23-24 April 2019.

Relevant Previous Publications

Gero, J. S., Jiang, H. & Williams, C. (2013). Design cognition differences when using unstructured, partially structured and structured concept generation creativity techniques. *International Journal of Design Creativity and Innovation, 1(4),* 196-214.

Hu, M., Shealy, T. & Gero, J. S. (2018). Neuro-cognitive differences among engineering students when using unstructured, partially structured and structured concept generation techniques. *Proceedings of the ASEE annual conference,* Salt Lake City, Utah: ASEE2018.

Kan, J. W. T. & Gero, J. S. (2017). Characterizing innovative processes in design spaces through measuring the information entropy of empirical data from protocol studies. *AIEDAM, 32(1),* 32-43.

Kannengiesser, U. & Gero, J. S. (2015). Is designing independent of domain? Comparing models of engineering, software and service design. *Research in Engineering Design, 26(3),* 253-275.

Milovanovic, J. & Gero, J. S. (2018). Exploration of cognitive design behavior during design critiques. In D. Marjanovic, P. J. Clarkson, U. Lindemann, T. McAloone & C. Weber (Eds), *Human Behavior in Design Vol. 5*, pp. 2099-2110. doi.org/10.21278/idc.2018.0547

Shealy, T., Hu, M. & Gero, J. S. (2018). Neuro-cognitive differences between brainstorming, morphological analysis and TRIZ. *Proceedings of the ASME IDETC*, paper DETC2018-86272.

Vieira, S., Gero, J. S., Delmoral, J., Fernandes, C., Gattol, V. & Fernandes, A. (2018). Workshop Paper: Studying the neurophysiology of designing through an EEG study of layout design: Preliminary results, *DCC'18 Workshop on Neurophysiological Measures and Biometric Analyses in Design Research*, Lecco, Italy, July 2018.

Yu, R., Gu, N., Ostwald, M. & Gero, J. S. (2015). Empirical support for problem-solution co-evolution in a parametric design environment. *AIEDAM 25(1),* 33-44.

Towards emotion based music generation: A tonal tension model based on the spiral array

Dorien Herremans (dorien_herremans@sutd.edu.sg)
Information Systems, Technology and Design
Singapore University of Technology and Design
8 Somapah Road, 487273 Singapore

Elaine Chew (elaine.chew@qmul.ac.uk)
School of Electronic Engineering and Computer Science
Queen Mary University of London
Mile End Road, E1 4NS London, UK

Keywords: Tension; Tonal tension, Music, Computational modelling; Music and emotion; Music structure

Introduction

Tension is an integral part of the music listening experience, closely connected to the sensing of emotions. We have explored how a particular aspect of tension, tonal tension, can be modelled and used to guide for automatic music generation.

A model was developed for extracting three aspects of tonal tension (Herremans & Chew, 2016b) from a musical score. The model is based on the spiral array, a three-dimensional model for tonality developed by Chew (2014). This was then integrated in an online interactive system, for easy visualisation, in sync with audio and score. Finally, the tension model was included in a state-of-the-art music generation system called MorpheuS, whereby we use tension to guide the underlying tension fabric of the generated music.

Spiral Array Model

In order to model tonal tension, we first need to be able to model pitches in a meaningful way. This was achieved through the three-dimensional model of tonality called the spiral array (Chew, 2014). The spiral array consists of three sets of helices: one that represents pitch classes, a pair for major and minor triads, and a pair for major and minor keys. The pitch spiral is the one we use for modelling tension. The triads are generated as convex combinations of their member pitches, and keys are represented as convex combinations of their defining chords.

Three new indicators of tension

Tension is a composite characteristic. There are many factors that contribute to the listener's feeling of tension, including loudness, timbre, dissonance, and harmonies. We chose to focus on tonal tension, and propose three characteristics, that are calculated for each time window, or cloud, of notes:

Cloud diameter Calculated as the largest distance between different notes in a cloud, thus capturing the dissonance of a note cluster.

Cloud momentum Calculated as the position change or movement between two adjacent clouds of notes, capturing the amount of harmonic change from one time slice to the next.

Tensile strain Computed as the distance between the centroid of the current slice and that of all pitches, representing the global key.

Figure 1 shows an example of the Tristan chord in the spiral array, a famous tense chord from Wagner's opera Tristan and Isolde. One can immediately see that it spans a large region in the pitch helix, which results in a high cloud diameter.

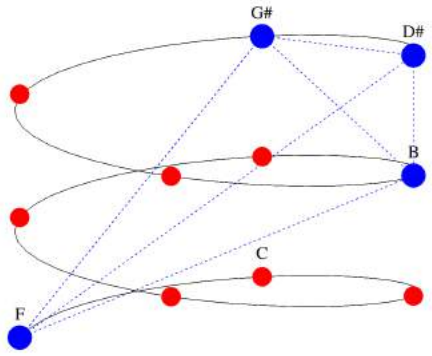


Figure 1: The Tristan chord in the Spiral Array pitch helix.

For a more complete overview of the proposed novel tonal tension model, the reader is referred to (Herremans & Chew, 2016b).

The model can be used for musicological or cognitive science purposes, as we have created an interactive online platform that visualises both tension (Herremans & Chuan, 2017) and arousal valence data (Herremans, Yang, et al., 2017) in sync with musical scores and audio.

Scaffolding music generation

In recent years, automatic music generation systems have become ever more popular due to advances in deep learning. There is a wide range of music generation systems available, e.g. for generating music that matches computer games, harmonizing a melody, etc. For a complete overview, the reader

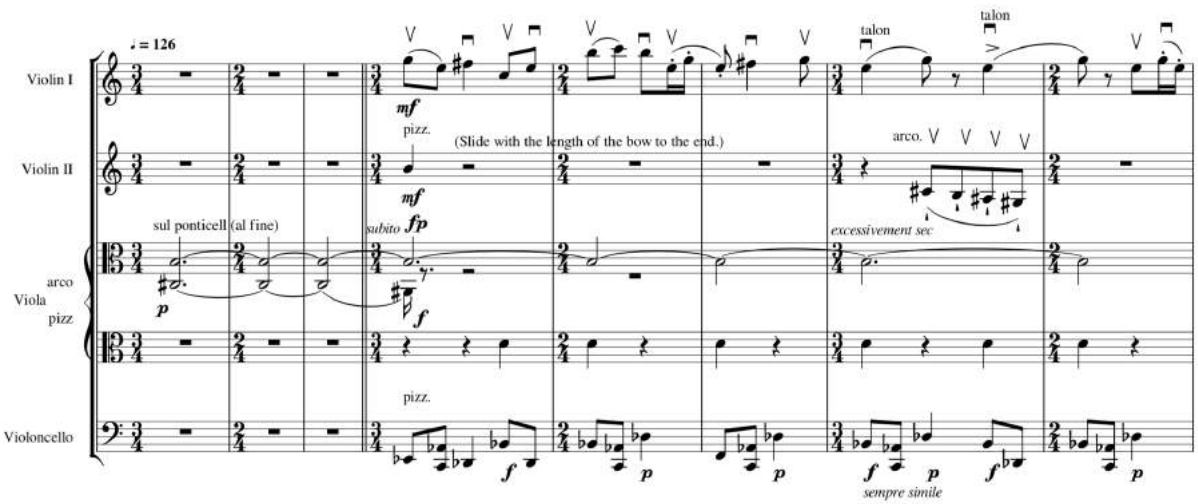


Figure 2: Excerpt (bars 1-8) of MorpheuS' piece based on the first of Stravinsky's *Three Pieces for String Quartet*

is referred to the survey paper by Herremans, Chuan, & Chew (2017). In this paper, the current challenges for music generation are identified as generating music with long term structure, and music that communicates certain emotions.

We developed a music generation system, called MorpheuS, which uses combinatorial optimization techniques to generate music with specific tension values over time (i.e., a given tension profile) and recurring pattern structure (Herremans & Chew, 2016a, 2017). MorhpeuS takes as input an existing musical score in MusicXML format. From this piece, the tension is calculated using the model described above. Secondly, recurring note patterns are extracted using the SIA algorithm by Meredith et al. (2002). The user can then use this original tension profile and the detected recurrent pattern structure, or create a new version of these, to scaffold the music generation process.

In the first step of the music generation process, all pitches of the original template piece are erased, but the rhythm is kept intact. A variable neighborhood search algorithm then populates the rhythm template with random pitches, while preserving the repeated pattern structure. The pitches are then optimized to maximize the fit between the current tension profile and the desired tension. For a more in depth explanation, the user is referred to Herremans & Chew (2017).

Figure 2 shows an example of one of the generated pieces by MorpheuS, based on Stravinsky's *Three Pieces for String Quartet*, composed for performance by members of the Singapore Symphony Orchestra on Channel News Asia.

Conclusions

The MorpheuS music generation system tackles one of the biggest remaining challenges in automatic music generation: generating music with structure and with the goal of communicating particular emotions over time. MorpheuS pieces have been performed internationally; recordings of selected pieces can be found online[1].

[1] dorienherremans.com/morpheus

Acknowledgments

This research started as part of a Marie Skłodowska-Curie Action Fellowship (EU MSCA grant No 658914), and has been supported by MOE2018-T2-2-161 and SMART-MIT Grant ING1611118-ICT.

References

Chew, E. (2014). *Mathematical and computational modeling of tonality: Theory and applications*. New York: Springer.

Herremans, D., & Chew, E. (2016a). Morpheus: Automatic music generation with recurrent pattern constraints and tension profiles. In *IEEE tencon.* Singapore: IEEE.

Herremans, D., & Chew, E. (2016b). Tension ribbons: Quantifying and visualising tonal tension. In *Second international conference on technologies for music notation and representation (tenor)* (Vol. 2, p. 8-18). Cambridge, UK.

Herremans, D., & Chew, E. (2017). Morpheus: generating structured music with constrained patterns and tension. *IEEE Transactions on Affective Computing*, *PP*. doi: 10.1109/TAFFC.2017.2737984

Herremans, D., & Chuan, C.-H. (2017). A multi-modal platform for semantic music analysis: visualizing audio- and score-based tension. In *11th international conference on semantic computing IEEE ICSC 2017.* San Diego.

Herremans, D., Chuan, C.-H., & Chew, E. (2017). A functional taxonomy of music generation systems. *ACM Computing Surveys (CSUR)*, *50*(5), 69.

Herremans, D., Yang, S., Chuan, C.-H., Barthet, M., & Chew, E. (2017). Imma-emo: A multimodal interface for visualising score-and audio-synchronised emotion annotations. In *Proc. of the 12th int. audio mostly conf. on augmented and participatory sound and music experiences.*

Meredith, D., Lemström, K., & Wiggins, G. A. (2002). Algorithms for discovering repeated patterns in multidimensional representations of polyphonic music. *Journal of New Music Research*, *31*(4), 321–345.

Cognitive Chrono-Ethnography (CCE): A Behavioral Study Methodology Underpinned by the Cognitive Architecture, MHP/RT

Muneo Kitajima (mkitajima@kjs.nagaokaut.ac.jp)
Department of Management and Information Systems Engineering, Nagaoka University of Technology
1603-1 Kamitomioka Nagaoka Niigata, 940-2188 JAPAN

Keywords: cognitive architecture; action selection; multidimensional memory; ethnography

Introduction: MHP/RT and CCE

At the 0-th order approximation, a person interacts with his or her environment by running an endless cycle of perceiving the external and internal environment through five senses via sensory neurons as parallel processing, and acting to the external environment through body parts via motor neurons as serial processing. As s/he perceives the results of movement of his/her body parts as well as the changes of the external environment as time goes by, the next cycle of Perceptual–Motor should occur. Interneurons in-between the sensory neurons and motor neurons convert the input patterns to the output patterns – these constitute a Perceptual-Cognitive-Motor process (PCM process). Starting from this basic cycle, we (M. Toyota and the author) constructed a comprehensive theory of action selection and memory, Model Human Processor with Realtime Constraints (MHP/RT), that should provide a basis for constructing any models for users interacting with ever-changing environments, and an accompanying behavioral study methodology, Cognitive Chrono-Ethnography (CCE) (Kitajima, 2016; Kitajima & Toyota, 2013) to be used to utilize, validate, and/or refine MHP/RT. MHP/RT and CCE are two wheels for conducting cognitive behavioral sciences, that complement each other from theoretical and experimental perspectives, respectively. Visit `http://oberon.nagaokaut.ac.jp/ktjm/organic-self-consistent-field-theory/index.html` for more information for the entire project.

Model Human Processor with Realtime Constraints (MHP/RT)

MHP/RT is an extension of Model Human Processor developed by Card, Moran, and Newell (1983). The purpose of MHP/RT is to implement at a higher level the facts that the fundamental processing mechanism of brain is Parallel Distributed Processing (PDP) (McClelland & Rumelhart, 1986), that human behavior emerges as the results of competition of the dual processes of System 2, slow *conscious* processes for deliberate reasoning with feedback control, and System 1, fast *unconscious* processes for intuitive reaction with feedforward control for connecting perception and motor movements, called Two Minds (Kahneman, 2003), and that human behavior is organized under happiness goals (Morris, 2006), on the assumption that the processing involved in action selection is truly dynamic interaction that evolves in the irreversible time dimension. The extension is done by considering that the endless PCM cycle continues from his or her birth to death in the ecological system that consists of the person and the environments, and it is a periodic circulation system.

MHP/RT consists of two parts. The first part is the cyclic PCM processes, in which PDP for those processes is implemented in hierarchically organized bands having their respective characteristic times for operations, i.e., biological, cognitive, rational, and social bands (Newell, 1990) by associating relative times (not absolute) to the PCM processes that carry out a series of events. The second part is memory, which supports the PCM processes. It is implemented as a distributed memory system and at the same time it serves as a mechanism to establish synchronization among multiple PCM processes.

Cognitive Chrono-Ethnography: CCE

Equipped with the cognitive architecture, MHP/RT, how can we study people's behaviors, characterized by Two Minds working dynamically along the time dimension? We came up with a solution in the form of a study methodology, called CCE. Cognitive Chrono-Ethnography combines three concepts. "Cognitive" declares that CCE deals with interactions between consciousness and unconsciousness in the PCM cycles. "Chrono(-logy)" is about time ranging from $\sim$100 msec to days, months, and years, and CCE focuses on such time ranges. "Ethnography" indicates that CCE takes ethnographical observations as the concrete study method because in daily life people's Two Minds tends to re-use experientially effective behavioral patterns, which is called "bias". Ethnographical field observations are essential for understanding each person's bias in his/her daily life.

CCE Procedure

Figure 1 shows the seven steps to conduct a CCE study:
1) Ethnographical Field Observation: Use the basic ethnographical investigation method to clarify the outline of the structure of social ecology that underlies the subject to study.
2) Mapping the Observed Phenomena on Cognitive Architecture: With reference to the behavioral characteristics of people which have been made clear so far and MHP/RT, consider what kind of characteristic elements of human behavior are involved in the investigation result in (1).
3) Identifying Study Parameters through Model-Based Simulation: Based on the consideration of (1) and (2), construct an initial simple model with the constituent elements of activated memories, i.e., meme, and the characteristic PCM processing to represent the nature of the ecology of the study space.
4) Design a CCE Study: Based on the simple ecological

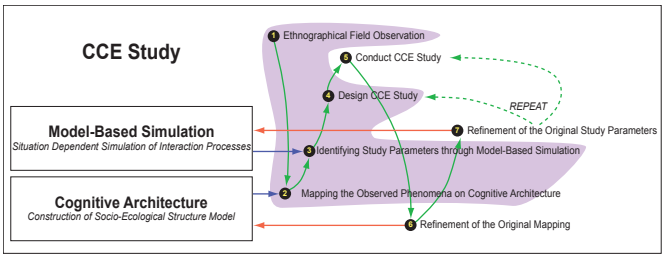


Figure 1: The CCE procedure (Kitajima, 2016, Figure 5.1).

model, identify a set of typical behavioral characteristics from a variety of people making up the group to be studied. Then formulate screening criteria of elite monitors who represent a certain combination of the behavioral characteristics, and define ecological survey methods for them.

5) Conduct CCE Study: Select elite monitors and conduct an ethnographical field observation.

6) Refinement of the Original Mapping: Check the results of (5) against the results of (2) for appropriateness of the mapping. If inappropriate, back to (2) and redo from there.

7) Refinement of the Original Study Parameters: If the result of (5) is unsatisfactory, go back to (4) and re-design and conduct a revised CCE study, otherwise go back to (3) to redo the model-based simulation with a set of refined parameters.

Completed CCE Studies: A Few Examples

Navigation in a train station by following signs: With the focus of action selection processes involved in slow navigation, Kitajima and Toyota (2012) reported a CCE study to investigate how elderly people use guide signs at train stations when they have to transfer lines, in addition to use some facilities such as restrooms, lockers, and so on. The results showed: 1) persons with inferior planning function with normal attention function did not use guide signs when they had mental models, whereas they did not gather task-relevant information but irrelevant one when they had no mental model, and 2) persons with inferior planning function and inferior attention consistently had problems in gathering task-relevant information by using guide signs because of vague description of behavioral goals. The interactions between planning and attention functions and the existence of mental models are consistent with MHP/RT's simulation results.

Sightseeing in a hot spring resort: Hot spring resorts are popular tourist attractions in Japan. However, little is known about why they are popular destinations. To answer this question, Kitajima, Tahira, Takahashi, and Midorikawa (2012) conducted a CCE study. Forty-three groups participated in the study as elite monitors. Each group arrived at Kinosaki-Onsen and were asked to tour the place. They were instructed to carry a GPS and a digital camera for recording their activities. By analyzing the results of the interviews, we identified six types of tourist activities including: bathing, staying, eating, exploring, touring, and shopping, each of which corresponds to a different set of happiness goals.

On-Going CCE Studies

Designing Memorable Events: People live in the environment filled with artifacts, part of which is real and the rest is virtual. Kitajima, Shimizu, and Nakahira (2017) conducted initial steps of CCE to understand how the PCM processes along with the memory process result in memorable experiences. Preliminary experiments were conducted to see how omnidirectional movies in virtual reality augmented with audio-guide made the experience memorable by timely provision of multi-modal information as designed by MHP/RT.

Designing Immersive Events: Immersive virtual environments are distinct from other types of multimedia learning environments. Dinet and Kitajima (2018) reported initial steps of CCE that focused on the conditions necessary to produce "immersive experience" for the user. The CCE study will continue in the context of developing a multimodal interface to help young pedestrians acquire necessary skills for safe navigation in dangerous traffic environments.

Six Publications Relevant to this Abstract

Dinet, J., & Kitajima, M. (2018). Immersive interfaces for engagement and learning: Cognitive implications. In *Proceedings of the 2015 virtual reality international conference* (pp. 18/04:1–18/04:8). New York, NY, USA: ACM.

Kitajima, M., Shimizu, S., & Nakahira, K. T. (2017). Creating memorable experiences in virtual reality. In *3rd IEEE International Conference on Cybernetics* (p. 1-8).

Kitajima, M. (2016). *Memory and Action Selection in Human-Machine Interaction*. Wiley-ISTE.

Kitajima, M., & Toyota, M. (2013). Decision-making and action selection in Two Minds: An analysis based on Model Human Processor with Realtime Constraints (MHP/RT). *Biologically Inspired Cognitive Architectures*, *5*, 82–93.

Kitajima, M., Tahira, H., Takahashi, S., & Midorikawa, T. (2012). Understanding tourist's *in situ* behavior: a Cognitive Chrono-Ethnography study of visitors to a hot spring resort. *Journal of Quality Assurance in Hospitality and Tourism*, *12*, 247–270.

Kitajima, M., & Toyota, M. (2012). Simulating navigation behaviour based on the architecture model Model Human Processor with Real-Time Constraints (MHP/RT). *Behaviour & Information Technology*, *31*(1), 41–58.

References

Card, S. K., Moran, T. P., & Newell, A. (1983). *The Psychology of Human-Computer Interaction*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Kahneman, D. (2003). A perspective on judgment and choice. *American Psychologist*, *58*(9), 697–720.

McClelland, J. L., & Rumelhart, D. E. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition*. The MIT Press.

Morris, D. (2006). *The nature of happiness*. London: Little Books Ltd.

Newell, A. (1990). *Unified Theories of Cognition*. Cambridge, MA: Harvard University Press. (p. 122, Fig. 3-3)

Warning: The Exemplars in Your Category Representation May Not Be the Ones Experienced During Learning

Kenneth J. Kurtz (kkurtz@binghamton.edu)
Department of Psychology, 4400 Vestal Parkway East
Binghamton, NY 13902 USA

Daniel C. Silliman (dsillim1@binghamton.edu)
Department of Psychology, 4400 Vestal Parkway East
Binghamton, NY 13902 USA

Keywords: Connectionist; Exemplar; Category learning; Computational modeling

The WARP Model of Category Learning

Research on categorization and classification learning has greatly benefitted from the use of computational modeling which requires making all theoretical assumptions explicit and provides a direct means of theory evaluation by *fitting* behavioral data. The field has advanced notably through model comparison relative to benchmark data on human category learning performance. Exemplar theory has become a leading psychological explanation largely due to the success of its formal models in fitting human data across a number of tasks (Kruschke, 1992; Nosofsky & Palmeri, 1997).

The exemplar view casts categorization as based on an explicit calculation of similarity between the to-be-categorized stimulus and instances stored in long-term memory (exemplars) associated with each category. The similarity is computed as an inverse exponential function of distance between psychological representations in a multidimensional space. This representational space can be transformed by stretching or shrinking dimensions using selective attention. The category with exemplars of greater similarity (less distance) to the stimulus is activated. This account has been extended in the ALCOVE model (Kruschke, 1992) which implements adaptive learning of attentional weights on the stimulus dimensions and association weights between each exemplar and category.

While exemplar models have shown a high degree of success in fitting behavioral data, they do not provide an account of representation learning. These models generally assume that each item in the input domain has a unique psychological representation (estimated via multidimensional scaling) that remains fixed throughout the category learning process. Further, a strict correspondence holds between the category representation and the stimulus items known to be members of that category (note: reference point models can also use centroids of clusters of exemplars).

This is in strong contrast to feedforward artificial neural networks that gradually learn representations to optimize task performance (Rumelhart, Hinton, & Williams, 1986). In standard connectionist models, each stimulus gets recoded at a "hidden" layer based on a set of optimized synapse-like weights that yield a distributed representation across the hidden nodes—which can be seen as a point in a constructed multidimensional space. A second set of weights connects these hidden nodes to an output layer of class nodes. The internal representations are incrementally repositioned in weight space via gradient descent to optimize accurate prediction at the output layer.

The Weights-as-Adaptive-Reference-Points (WARP) model is designed to bridge the reference point similarity-based approach of exemplar models with the flexibility and psychological plausibility of learned representations in neural networks. This merger of design principles is achieved by replacing the localist exemplar node representations (as in ALCOVE) with a layer that follows the foundational connectionist design principles of: 1) a forward pass that computes activation based on a function of the 'net input,' i.e., the input activations multiplied by their weights; and 2) a backward pass that modifies the weights to minimize task error and estimate the function to be approximated.

On the connectionist view, the hidden nodes are constructed dimensions that usefully transform the values of a stimulus in input space to a set of values in another representational space. On the exemplar view, each hidden node is a reference point to the location of a training item in input space and its activity indexes the proximity of that point to a stimulus. We propose a new formulation that allows the hidden nodes to function according to connectionist mechanics and yet act as reference points. The result is that the model discovers its own reference points using task-driven error minimization as opposed to making a commitment to the inputs themselves as the basis for the reference points.

The WARP model functions by taking the encoding weights to each hidden node as its "address" or reference point location in input space. As the weights change via learning by backpropagation, each node follows a trajectory in weight space from its initial random location toward a place where its task is functionality optimized. The 'net input' is the vector multiplication between the input activations and the incoming weights to a node. This is a dot product or linear algebraic measure of similarity (i.e., the angle between the vectors) as opposed to a spatial distance metric. The critical similarity computation between stimulus and reference point occurs implicitly in the forward pass. To

make this work as intended, a simple, novel activation function at the hidden layer is used which takes the form of Equation 1:

$$exp[(a \cdot b) - k] \quad (1)$$

where a is the vector of input activations, b is the vector of incoming input->hidden node weights, and k is a constant value set to the number of dimensions in the category structure. The key property of this function is this: the more closely the incoming weight vector for a hidden node approximates the values of an input vector, the greater the activation of the hidden node. Over the course of training, different hidden nodes will be repositioned to parts of weight space that allow them to respond to particular regions in input space: to get better at classifying is to move the adaptive reference points to useful positions. A standard association layer connects the hidden nodes to class nodes and a softmax output layer is used to determine the class probabilities

WARP utilizes a set of connectionist-style free parameters: learning rate, number of hidden nodes (i.e., density of the implicit covering map), and range of random initialization for incoming weights; and can also incorporate a set of reference point model-style free parameters: degree of sensitivity of reference points and a response mapping constant for determining class activations.

Preliminary testing has shown promising fits to the classic behavioral benchmark of the Shepard, Hovland, and Jenkins (1961) six types of elemental category structures (dataset from Nosofsky et al., 1994). This investigation also revealed that the WARP model discovers more parsimonious reference points when available: instead of always dedicating each hidden node to a single input, WARP can develop reference points that respond strongly to particular feature correlations or unidimensional rules. In conjunction with classic exemplar-style nodes, these feature detector-style nodes allow the model to efficiently handle various and complex category structures. The use of this multi-strategy toolkit mirrors the diversity and flexibility of human category learning (Ashby, Alfonso-Reese, & Waldron, 1998).

In addition to modeling human behavior, WARP has also been initially tested for potential application as a classifier in the domain of machine learning. Different parameterizations of the model, while inappropriate for capturing the pace and nuance of human learning, show highly rapid and efficient performance on the iris flower benchmark dataset. Interestingly, the model solves the classification problem using *discriminative prototypes* that maximize distance to competing classes while minimizing distance to the target class. Continued investigations of the model are underway to better reveal the nature and diversity of the solutions WARP discovers for different types of classification problems; and to determine the power of the model in addressing the goals of psychological explanation and advancing AI.

Relevant Publications

Corral, D., Kurtz, K. J., & Jones, M. (2018). Learning relational concepts from within-versus between-category comparisons. *Journal of Experimental Psychology: General*, *147*(11), 1571.

Conaway, N., & Kurtz, K. J. (2017a). Similar to the category, but not the exemplars: A study of generalization. *Psychonomic bulletin & review*, *24*(4), 1312-1323.

Conaway, N., & Kurtz, K. J. (2017b). Solving nonlinearly separable classifications in a single-layer neural network. *Neural computation*, *29*(3), 861-866.

Honke, G. & Kurtz K.J. (in press). Similarity is as similarity does? A critical inquiry into the effect of thematic association on similarity. *Cognition*.

Kurtz, K. J., Boukrina, O., & Gentner, D. (2013). Comparison promotes learning and transfer of relational categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(4), 1303.

Kurtz, K. J., Levering, K. R., Stanton, R. D., Romero, J., & Morris, S. N. (2013). Human learning of elemental category structures: Revising the classic result of Shepard, Hovland, and Jenkins (1961). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(2), 552.

Levering, K. R., & Kurtz, K. J. (2015). Observation versus classification in supervised category learning. *Memory & cognition*, *43*(2), 266-282.

Pape, A. D., Kurtz, K. J., & Sayama, H. (2015). Complexity measures and concept learning. *Journal of Mathematical Psychology*, *64*, 66-75.

References

Ashby, F. G., Alfonso-Reese, L. A., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological review*, *105*(3), 442.

Kruschke, J. K. (1992). ALCOVE: an exemplar-based connectionist model of category learning. *Psychological review*, *99*(1), 22.

Nosofsky, R. M., Gluck, M. A., Palmeri, T. J., McKinley, S. C., & Glauthier, P. (1994). Comparing modes of rule-based classification learning: A replication and extension of Shepard, Hovland, and Jenkins (1961). *Memory & cognition*, *22*(3), 352-369.

Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological review*, *104*(2), 266.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, *323*(6088), 533.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological monographs: General and applied*, *75*(13), 1.

Concept Learning with Energy-Based Models

Igor Mordatch (mordatch@openai.com)
OpenAI, San Francisco, CA

Keywords: concepts;meta-learning;imitation learning

Introduction

Many hallmarks of human intelligence, such as generalizing from limited experience, abstract reasoning and planning, analogical reasoning, creative problem solving, and capacity for language require the ability to consolidate experience into *concepts*, which act as basic building blocks of understanding and reasoning.

Examples of concepts include visual (*"red"* or *"square"*), spatial (*"inside"*, *"on top of"*), temporal (*"slow"*, *"after"*), social (*"aggressive"*, *"helpful"*) among many others (Rosch, Mervis, Gray, Johnson, & Boyes-braem, 1976; Lakoff & Johnson, 1980). These concepts can be either identified or generated - one can not only find a square in the scene, but also create a square, either physical or imaginary. Importantly, humans also have a largely unique ability to combine concepts compositionally (*"red square"*) and recursively (*"move inside moving square"*) - abilities reflected in the human language. This allows expressing an exponentially large number of concepts, and acquisition of new concepts in terms of others. We believe the operations of identification, generation, composition over concepts are the tools with which intelligent agents can understand and communicate existing experiences and reason about new ones.

Crucially, these operations must be performed on the fly throughout the agent's execution, rather than merely being a static product of an offline training process. Execution-time optimization, as in recent work on meta-learning (Finn, Abbeel, & Levine, 2017) plays a key role in this. We pose the problem of parsing experiences into an arrangement of concepts as well as the problems of identifying and generating concepts as optimizations performed during execution lifetime of the agent. The meta-level training is performed by taking into account such processes in the inner level.

Specifically, a concept in our work is defined by an energy function taking as input an event configuration (represented as trajectories of entities in the current work), as well as an attention mask over entities in the event. Zero-energy event and attention configurations imply that event entities selected by the attention mask satisfy the concept. Compositions of concepts can then be created by simply summing energies of constituent concepts. Given a particular event, optimization can be used to identify entities belonging to a concept by solving for attention mask that leads to zero-energy configuration. Similarly, an example of a concept can be generated by optimizing for a zero-energy event configuration. See Figure 1 for examples of these two processes.

The energy function defines a family of concepts, from which a particular concept is selected with a specific concept code. Encoding of event and attention configurations can be achieved by execution-time optimization over concept codes. Once an event is encoded, the resulting concept code structure can be used to re-enact the event under different initial configurations (task of imitation learning), recognize similar events, or concisely communicate the nature of the event. We believe there is a strong link between concept codes and language, but leave it unexplored in this work.

Description of events we consider and video results of our model learning on these events are available at: `sites.google.com/site/energyconceptmodels`

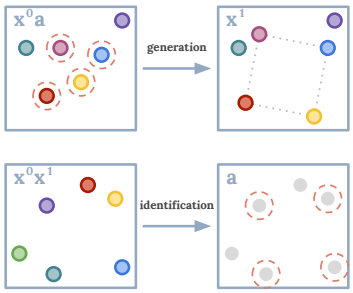


Figure 1: Examples of generation and identification processes for a *"square"* concept. a) Given initial state $\mathbf{x}^0$ and attention mask $\mathbf{a}$, square consisting of entities in $\mathbf{a}$ is formed via optimization over $\mathbf{x}^1$. b) Given states $\mathbf{x}$, entities comprising a square are found by optimization over attention mask $\mathbf{a}$.

Method

Existence of a particular concept is given by energy function $E(\mathbf{x},\mathbf{a},\mathbf{w}) \in \mathbb{R}^+$, where parameter vector $\mathbf{w}$ specifies a particular concept from a family. $E(\mathbf{x},\mathbf{a},\mathbf{w}) = 0$ when state trajectory $\mathbf{x}$ under attention mask $\mathbf{a}$ over entities satisfies the concept $\mathbf{w}$. Otherwise, $E(\mathbf{x},\mathbf{a},\mathbf{w}) > 0$. The conditional

probabilities of a particular event configuration belonging to a concept and a particular attention mask identifying a concept are given by the Boltzmann distributions:

$$p(\mathbf{x}|\mathbf{a},\mathbf{w}) \propto \exp\{-E(\mathbf{x},\mathbf{a},\mathbf{w})\} \quad (1)$$

$$p(\mathbf{a}|\mathbf{x},\mathbf{w}) \propto \exp\{-E(\mathbf{x},\mathbf{a},\mathbf{w})\} \quad (2)$$

Given concept code $\mathbf{w}$, the energy function can be used for both generation and identification of a concept implicitly via optimization (see Figure 1):

$$\mathbf{x}(\mathbf{a}) = \underset{\mathbf{x}}{\operatorname{argmin}} E(\mathbf{x},\mathbf{a},\mathbf{w}) \quad \mathbf{a}(\mathbf{x}) = \underset{\mathbf{a}}{\operatorname{argmin}} E(\mathbf{x},\mathbf{a},\mathbf{w}) \quad (3)$$

To learn concepts from experience grounded in events, we pose a few-shot prediction task. Given a few demonstration example events and initial state for a new event, the task is to predict attention $\mathbf{a}$ and the future state $\mathbf{x}$ of the new event. We

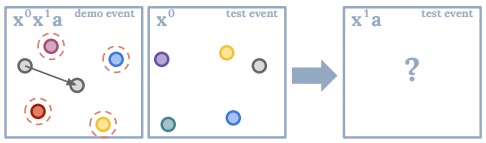


Figure 2: Example of a few-shot prediction task.

follow the maximum entropy inverse reinforcement learning formulation (Ziebart, Maas, Bagnell, & Dey, 2008) and assume demonstrations are samples from the distributions given by the energy function E and find energy function parameters θ via maximum likelihood estimation over future state and attention given initial state. The resulting loss for a particular dataset X is

$$\mathcal{L}_p^{\mathrm{ML}}(X,\mathbf{w}) = \mathbb{E}_{(\mathbf{x},\mathbf{a})\sim X}\left[-\log p\left(\mathbf{x}^1,\mathbf{a} \mid \mathbf{x}^0,\mathbf{w}\right)\right]$$

Where the joint probability can be decomposed in terms of probabilities in (1) and (2) as

$$\log p\left(\mathbf{x}^1,\mathbf{a} \mid \mathbf{x}^0,\mathbf{w}\right) = \log p\left(\mathbf{x}^1 \mid \mathbf{a},\mathbf{w}_x\right) + \log p\left(\mathbf{a} \mid \mathbf{x}^0,\mathbf{w}_a\right)$$

We use two concept codes, $\mathbf{w}_x$ and $\mathbf{w}_a$ to specify the joint probability. The interpretation is that $\mathbf{w}_x$ specifies the concept of the action that happens in the event (i.e. *"be in center of"*) while $\mathbf{w}_a$ specifies the argument the action happens over (i.e. *"square"*). This is a concept structure or syntax that describes the event. The concept codes are interchangeable and same concept code can be used either as action or as an argument because the energy function defining the concept can either be used for generation or identification. This importantly allows concepts to be understood from their usage under multiple contexts.

Experimental Results

We introduce a simulated environment and tasks for a two-dimensional scene consisting of a varying collection of entities, each processing position, color, and shape. We observe the following properties:

Concept inference in multiple contexts: An important property of our model is ability to learn from and apply it in both generation and identification contexts. We qualitatively observe that the model performs sensible behavior in both contexts. For example, we considered events with proximity relations *"closest"* and *"farthest"* and found model able to both attend to entities that are closest or furthest to another entity, and to move an entity to be closest or furthest to another entity.

Transfer between contexts: When our model trained on both contexts it shares experience between contexts. Knowing how to act out a concept should help in identifying it and vice versa. We perform an experiment where we train the energy model only in identification context and test the model's performance in generation context (and conversely). We observe that even without explicitly being trained on the appropriate context, the networks perform much better than baseline of two independently-trained networks.

Sharing codes across contexts: Another property of our model is that codes $\mathbf{w}_x$ and $\mathbf{w}_a$ for identifying concepts are interchangeable and can be shared between generation and identification contexts. For example, either turning an entity red would or identifying all red entities in the scene would ideally use the same concept of *"red"*. We indeed observe that events which involve recognizing entities of a particular color, the codes $\mathbf{w}_a$ match the codes $\mathbf{w}_x$ for setting entities to that color and find similar correlation in the other events as well.

Conclusion

We believe that execution-time optimization plays a crucial role in acquisition and generalization of knowledge, planning and abstract reasoning, and communication. In the current work we used a simple concept structure, but more complex structure with multiple arguments or recursion would be interesting to investigate in the future. It would also be interesting to test compositionality of concepts, which is very suited to our model as compositions corresponds to the summation of the constituent energy functions.

References

Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. *arXiv preprint arXiv:1703.03400*.

Lakoff, G., & Johnson, M. (1980). The metaphorical structure of the human conceptual system. *Cognitive science*, *4*(2), 195–208.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-braem, P. (1976). Basic objects in natural categories. *COGNITIVE PSYCHOLOGY*.

Ziebart, B. D., Maas, A. L., Bagnell, J. A., & Dey, A. K. (2008). Maximum entropy inverse reinforcement learning. In *Aaai* (Vol. 8, pp. 1433–1438).

On the nature of creative processes: performativity as a missing algorithm

Antonino Pennisi (apennisi@unime.it)
Department of Cognitive Sciences, Psychology, Education and Cultural Studies (COSPECS), 6/8 Concezione Street
Messina, ME 98121 Italy

Gessica Fruciano (frucianogessica@gmail.com)
CRISCAT (International Research Center for Theoretical and Applied Cognitive Sciences) University of Messina and
Universitary Consortium of Eastern Mediterranean, Noto (CUMO) 78 A. Sofia Street
Noto, SR 96017 Italy

Giovanni Pennisi (gpennisi@unime.it)
Department of Cognitive Sciences, Psychology, Education and Cultural Studies (COSPECS), 6/8 Concezione Street
Messina, ME 98121 Italy

Keywords: performativity; creativity; embodied cognition; biolinguistics; schizophrenia

The creative role of performativity

In our project the performance is a product of performativity. Performativity is the cognitive ability to produce physical or mental actions. Studying performance and studying performativity sets different scientific activities. Studying how to enhance performance belongs to the behavioral science. On the contrary, studying performativity belongs to a general cognitive procedure that must not be confused with the description of behaviors, requiring instead a specific theorization in the cognitive sciences. The aim of this research project is to focus on the hypothesis that performativity is not a property confined to certain specific human skills, or to certain specific acts of language. Instead, the executive and motor component of cognitive behavior should be considered an intrinsic part of the physiological functioning of the mind and as endowed with self-generative power (Pennisi A., 2019; Pennisi A.-Falzone, 2016).

We believe that performativity has evolved alongside with those natural selection processes which have led the human species to develop articulated language and the embodied simulation (Pennisi A.-Falzone, 2016; Falzone 2018). In such framework, cognition is a form of mediated action rather than the link between inner thought and overt behavior. According to our model, thus, action is not the mere externalization of a mental process, but is the process itself (Pennisi A., 2018 and 2019; Pennisi A.-Falzone, 2019; Gallese, 2019). Since such process is carried out through the body, we think that the species-specificity of the bodies occurring in nature paves the way for every individual's knowledge of reality.

Performativity as a physiological tool of cognitive creativity has precise neural correlates and procedural properties.

From the point of view of procedures, performativity is a cognitive property that arises from the absence of an algorithm designed to carry out a given performance. Acting in a non-planned way, learning by trial and error, applying familiar behavioral patterns to new situations: these are just a few examples of what is performativity and of how it works.

Thus, performativity is intrinsically creative because its nature is to face situations that cannot be solved by the application of already known algorithms. In a nutshell, performative creativity is a procedural system that is somewhere between what Chomsky called "rule governed creativity" and "rule-changing creativity". Performativity however bears a peculiar kind of creativity, which is different from the one generated by the competence but still shares some features with the latter: in fact, it is a fully embodied and free-from-rules process that is carried out through trial and error, that is to say it depends on the bodily practice (locomotion, language, perception, etc.) made in everyday experience (Pennisi A., 2019; Pennisi A.-Falzone, 2016; Gallese 2018; Matteucci, 2018; Montani 2018). In functional terms, hence, the brain is a powerful biological instrument which permits continuous reorganization of the activity of organisms. An incessant activity of biological agents that move and act, that perceive and explore the world around them through a network of sensors and nerves, whose complexity of articulation is directly dependent on the species-specific structure. This activity relentlessly stimulates the rewiring of sensorimotor networks and remodeling of cognitive interactions. Our mind is the result of this close cooperation between the performative competence triggered by sensory-motor systems and the readjustment of the computational procedures of our deep brain to allow the survival and growth in the fitness of individuals and the entire species within environmental variation.

Insights from neurolinguistics

A large amount of literature has been devoted to the aforementioned mapping process, carried out through both brain imaging (Monchi et al. 2001, 2006; Nagano-Saito et al. 2008) and the study of the biochemical reactions involved in the plasticity of synaptic processes (Thivierge et al. 2007; Ko et al. 2013; Tamburrini-Prevete, 2018). Such researches have demonstrated "that the caudate nucleus and the putamen are particularly important, respectively, in the planning and the

execution of a self-generated novel action, whereas the subthalamic nucleus may be required when a new motor program is solicited independently of the choice of strategy" (Monchi et al. 2006, 257). Examining the biolinguistic aspects of these discoveries in depth, Lieberman and his team have shown that the neural circuits connecting different brain parts during human speech exploit the putamen for neuromotor control, changing "on the run" - that is, during verbal action performance - "the direction of our thought processes based on new stimuli such as the understanding of meaning conveyed by the syntax of language" (Lieberman & McCarthy 2007, 16).

Furthermore, a similar activation of brain motor components is registered when language data are processed in the absence of grammatically well-tested algorithms, such as when a second language is learned (Klein et al. 1994), or when a subject switches from listening to informal speech to a more formal one (Abutalebi et al. 2007).

In short, the management of neurocerebral performative strategies seems to be responsible for the most dynamic processes of linguistic behavior. This kind of behavior needs an attempt, or an active effort, that cannot be accomplished only through the mechanical application of already known and stabilized rules because it requires "the execution of a self-generated action among competitive alternatives" (Lieberman 2013, 80): an activity that is prolonged virtually forever, after the first acquisition step of ontogenetic speech, moving from mechanical physiology to the physiology of thought.

This overall framework also explains why the paths of speech often follow the hesitational phenomena of breaking up, recomposition, reunion, syncretism, propositional chiselling, semantic and lexical refinement: that is, all that is stigmatized by Chomsky's idea of performance as the deposit of cognitive junk produced by externalization devices (to repeat his words: "numerous false starts, deviations from rules, changes of plan in mid course, and so on", 1960, 530). On the contrary, the most advanced neurolinguistic research reveals the close interconnection between motor performativity and the continuous reorganization of propositional and abstract thinking: "the cortico-striatal regions that regulate language comprehension also regulate many aspects of behavior such as motor control and abstract reasoning" (Simard, Monchi et al. 2010, 1092). Evolutionarily, in fact, the performative motricity of thought could have been decisive for understanding the subsequent development of human language, "because it indicates that our modem brains may actually have been shaped by an enhanced capacity for speech motor control that evolved in our ancestors" (Lieberman & McCarthy 2007, 16).

Schizophrenia as the realm of anti-performativity

Another field of research which supports our idea of performativity is phenomenological psychopathology. Authors like Sass (1992), Stanghellini (2004) and Fuchs (2005), in fact, claim that one of the core symptoms of schizophrenia is a sort of "disembodiment", the onset of a problematic relationship between the patient and his own body in which the parts of the latter become heavy, distorted and even "stranger". This peculiar kind of corporeity is reflected in a total lack of fluidity in any patient's performance: "patients frequently experience a disintegration of habits or automatic performances, a «disautomation». Instead of simply dressing, driving, walking, etc., they have to prepare and produce each single action deliberately, in a way that could be called a «Cartesian» action of the mind on the body" (Fuchs & Röhricht 2017).

Such schizophrenic tendencies might be described as the attempt to apply procedural rules - algorithms - to the everyday and well-mastered situations that make up our "being in the world", as the following words by a schizophrenic patient show: "If I do something like going for a drink of water, I've to go over each detail – find cup, walk over, turn tap, fill cup, turn tap off, drink it" (Chapman 1966, 239). As we have already claimed (Pennisi G. 2018), schizophrenia might be read as the disruption of the mechanisms that make a performance efficient, namely the selective target control, the softly conscious monitoring of one's bodily configurations and the implicit sense of body-as-subject (Gallagher 2018).

Instead of having this tacit, self-transparent and immediate relationship with their own bodies, patients often exercise a thematic control on the latter that goes from repetitively touching their own body parts – as if they try to verify if their body still «belongs» to them – to the fragmentation of every goal-related movement in many sub-movements, like in the previous example. Schizophrenics' inability to get in the flow of the action is what makes such illness "the realm of anti-performativity" (Pennisi G. 2018): this is why we think that the study of the role of performativity on human cognition cannot be separated from the phenomenological analysis of psychopathologies.

Conclusion

In the light of the above, we will define performativity as a constituent component of the cognitive processes. The actions that we perform in the environment, in fact, allow us to know both the surrounding world and our physical possibilities. In such model, the body is not only the means by which the individual explores and acts on the environment, but the precondition for the development of any cognitive ability.

Our intention is to validate our ideas on the role of the body and on performativity by applying the interdisciplinary methods of Cognitive Science. The issues we have raised, in fact, not only are the subject of a debate between the embodied/extended mind models and the mentalist hypotheses carried out by cognitive psychology and computationalism, but can only be clarified by providing an overview of the scientific literature on psychopathology and on cognitive neuropsychology.

References

Abutalebi, J., Brambati, S. M., Annoni, J. M., Moro, A., Cappa, S. F., & Perani, D. (2007). The neural cost of the auditory perception of language switches: An event-related functional magnetic resonance imaging study in bilinguals. *The Journal of Neuroscience, 27(50)*, 13762-13769.

Chapman J. (1966). The Early Symptoms of Schizophrenia. *The British Journal of Psychiatry, 112(484)*, 225-251.

Chomsky, N. (1960). Explanatory models in linguistics. *Studies in Logic and the Foundations of Mathematics, 44*, 528-550.

Falzone, A. (2018). Performativity and evolution. *Reti, saperi, linguaggi. Italian Journal of Cognitive Sciences*, 1/2018, 149-60.

Fuchs, T. (2005). Corporealized and disembodied minds: a phenomenological view of the body in melancholia and schizophrenia. *Philosophy, Psychiatry, & Psychology, 12(2)*, 95-107.

Fuchs, T., & Röhricht, F. (2017). Schizophrenia and intersubjectivity: An embodied and enactive approach to psychopathology and psychotherapy. *Philosophy, Psychiatry, & Psychology, 24(2)*, 127-142.

Gallagher S. (2018). Mindfulness and mindlessness in performance. *Reti, saperi, linguaggi. The Italian journal of Cognitive Sciences*, 1/2018, 5-18.

Gallese, V. (2018). Embodied simulation and its role in cognition. *Reti, saperi, linguaggi. Italian Journal of Cognitive Sciences*, 1/2018, 31-46.

Klein, D., Zatorre, R. J., Milner, B., Meyer, E., & Evans, A. C. (1994). Left putaminal activation when speaking a second language: Evidence from PET. *Neuroreport, 5(17)*, 2295-2297.

Ko, J. H., Antonelli, F., Monchi, O., Ray, N., Rusjan, P., Houle, S., et al. (2013). Prefrontal dopaminergic receptor abnormalities and executive functions in Parkinson's disease. *Human brain mapping, 34(1)*, 1591-1604.

Lieberman, P., & McCarthy, R. (2007). Tracking the evolution of language and speech: Comparing vocal tracts to identify speech capabilities. *Expedition: The magazine of the University of Pennsylvania, 49(2)*, 15-20.

Lieberman, P. (2013). *The unpredictable species. What makes humans unique.* Princeton: Princeton University Press.

Matteucci, G. (2018). Creativity as extended mind's aesthetic performativity. *Reti, saperi, linguaggi. Italian Journal of Cognitive Sciences*, 1/2018, 69-80.

Monchi, O., Petrides, M., Petre, V., Worsley, K., & Dagher, A. (2001). Wisconsin card sorting revisited: Distinct neural circuits participating in different stages of the task identified by event-related functional magnetic resonance imaging. *The Journal of Neuroscience, 27(19)*, 7733-7741.

Monchi, O., Petrides, M., Strafella, A. P., Worsley, K. J., & Doyon, J. (2006). Functional role of the basal ganglia in the planning and execution of actions. *Annals of neurology, 59(2)*, 257-264.

Montani, G. (2018). Once again on «Narrative imagination». «Schematizing without concept» in language, image and dreaming brain. *Reti, saperi, linguaggi. Italian Journal of Cognitive Sciences*, 1/2018, 127-134.

Nagano-Saito, A., Leyton, M., Monchi, O., Goldberg, Y. K., He, Y., & Dagher, A. (2008). Dopamine depletion impairs front os tri atal functional connectivity during a set- shifting task. *The journal of neuroscience, 28(14)*, 3697- 3706.

Pennisi, A. (2018). Performative Dimensions in cognitive sciences. *Reti, saperi, linguaggi. Italian Journal of Cognitive Sciences*, 1/2018, pp. 25-30.

Pennisi, A. (2019). Dimensions of the bodily creativity. For an extended theory of performativity. In A. Pennisi-A.Falzone (eds), *The Extended Theory of Cognitive Creativity. Interdisciplinary Approaches to Performativity*, Berlin: Springer, 11-43.

Pennisi A., Falzone A. (2016). *Darwinian biolinguistics. Theory and history of naturalistic philosophy on language.* Berlin-Heidelberg-New York-Cham: Springer.

Pennisi A., Falzone A. (2019). *The Extended Theory of Cognitive Creativity. Interdisciplinary Approaches to Performativity,* Berlin: Springer.

Pennisi, G. (2018). Towards a deeply embodied Enactivism. *Reti, saperi, linguaggi. Italian Journal of Cognitive Sciences*, 2/2018, 271-280.

Sass, L. A. (1992). *Madness and modernism: Insanity in the light of modern art, literature, and thought.* New York: Basic Books.

Simard, F., Joanette, Y., Petrides, M., Jubault, T., Madjar, C., & Monchi, O. (2010). Fronto-striatal contribution to lexical set-shifting. *Cerebral Cortex, 21*, 1084-1093.

Stanghellini, G. 2004. *Disembodied spirits and deanimated bodies: The psychopathology of common sense.* Oxford: Oxford University Press.

Thivierge, J. P., Rivest, F., & Monchi, O. (2007). Spiking neurons, dopamine, and plasticity: Timing is everything, but concentration also matters. *Synapse-New York, 61(6)*, 375-390.

Tamburrini, G., Prevete, R. (2018). Neuromodulation and neural circuit performativity: towards a computational model. *Reti, saperi, linguaggi. Italian Journal of Cognitive Sciences*, 1/2018, 111-126.

Why sociality affects creativity: lessons from autism

Pennisi Paola (ppennisi@unime.it)
Linguistic Centre of Messina University (CLAM), 54 Via Luciano Manara
Messina, ME 98123 Italy

Giallongo Laura (lgiallongo@unime.it)
Department of Cognitive Sciences, 6/8 via Concezione
Messina, ME 98121 Italy

Keywords: creativity; autism; imagination; social cognition; divergent thinking; insight

Introduction

As human beings we are social. All of us had to be included in a group to survive; most of us highly desire to live and collaborate with others on a daily basis. In this paper we will try to show how our sociality (considered as the inclination to live and collaborate with other co-specifics) affects our creativity.

How sociality affects creativity

Creativity, in fact, means being yourself, seeing the world in a way that is different from that of others. Each time that we perceive the world, we collect or ignore some data, we focus on something and neglect something else. Each perception is a creative act and this is showed not only by the Kanizsa's triangle or other similar optic illusions, but even by our spontaneous impulse to build our reality. When we are in love, for example, we are more inclined to interpret the gestures of our object of love in the direction that we would like to be the real one. In this condition we could easily mistake a wink aimed at the expulsion of a hair from the other's eye with a wink towards us. The thirst makes us see the water even where it is not there. What we call reality is an interspecific bargaining of the meaning of a perception.

Our sociality can push us to creativity in many ways: inviting us to solve problems, providing new information, criticizing one of our acts of creation or even inviting us to brainstorm. Societies also often reward creativity. But the *eureka,* the act of creating a different way of thinking something will take place only if we are able to go beyond the conformity of our perceptions with those of others.

Working definition for "creativity"

Creativity is a very heterogeneous concept. Here we will consider "creativity" as the ability to generate multiple solutions to a problem.

This definition encompasses in the same category the divergent thinking, insights and artistic creativity[1].

[1] This last can be seen as the essay of the artist to resolve the problem of representing his subjects.

Creativity in autism

Autism is a neurodevelopmental disorder characterized by persistent deficits in social communication and interaction and restricted and repetitive patterns of behaviour, interests or activities (APA 2013). Among the numerous consequences of the disorder, there are the lack of spontaneous symbolic play (Jarrold et al. 1993); anomalies in imagination (Low et al. 2009); difficulties to understand metaphors (Hobson 2012; Rumblad & Annaz 2010); very poor dreamlike activity (Daoust et al. 2007). For these reasons, subjects with autism are frequently considered less creative than subjects without autism. I.e. Craig and Baron Cohen (1999) described autistic creativity as a reality-based creativity and opposed it to the imaginative creativity of people without autism (Craig & Baron Cohen 1999).

The artistic productions of some savants with autism are famous for their proximity to reality – i.e. Stephen Wiltshire's productions, or Nadia's drawings (Selfe 2011). However a lot of other productions of autistic subjects show that the disorder doesn't affect the imaginative creativity: see i.e. Tammet (2008) or fig.1, which is a drawing made a 7 years autistic child.

Figure 1

Moreover, also among those who show the reality-based style of creativity described Craig & Baron Cohen (1999), subjects frequently solve problems in non-conformist ways. I.e., Temple Grandin managed to solve a major technical problem in the slaughtering of cows thanks to her style of thought which is indeed based on a reality-based form of creativity that is impossible to artlessly catch for people without autism (Grandin 1995).

As we will try to show in the full paper, the lacking of social affordances in subjects with autism greatly enhance their creativity, making their professional or artistic contribution very original for many fields of studies.

Subjects with autism, in fact, can think and imagine things in different ways than that of the most part of the population because they are less subject to perceptive and psychological biases linked to human sociality. I.e. their ability to make physical causation inference is superior than that of the most part of the population; on the contrary

emotional and intentional inferences are more difficult for subjects with autism than for the rest of the population (Pennisi 2016).

Why not all subjects with autism are creative?

Unfortunately, neurodevelopmental disorders are frequently associated with a low QI. Below a certain IQ, it is rarely possible to express one's creativity in a way that is comprehensible to others. Some talents sometimes manage to emerge, such as in the case of Nadia (Selfe 2011), but normally too low intellectual quotients do not allow the expression of the creativity of one's own creativity.

For all those subjects with autism who have an average or above average IQ, creativity is probably hidden where we are not used to looking for it. The absence of social motivation (Chevallier et al. 2012) turns into the habit of not asking others to help solve their problems and not to receive requests for help in solving problems. But in a world where the rules of sociability are a far-off buzz, the need to solve everyday problems requires the use of creativity. I.e., a child with autism who wants to open a door handle too high for him could easily take the adult's hand next to him and use it as a tool to open the door, rather than explicitly asking for help. Certainly this is a not very conventional way of "using" the adult's arm. Italian journalist Gianluca Nicoletti, father of a boy with autism (Tommaso), tells how his son, interested in not losing his favorite cassette during a move, was able to find a way to identify the right tape in a mountain of identical boxes (Nicoletti 2015). Nobody knows exactly what strategy the boy used, but certainly it hides an attitude to think and perceive the mountain of boxes in a totally different way from the rest of the family. An ordinary child would have simply asked the mother to remember for him and she would have drawn something on the outside of the box.

The point is that creativity is always linked to something pre-existing. It is likely that, in the eyes of people without autism, many tactics used by individuals with autism are creative, whereas for Tommaso, the ability to locate the cassette in the box was not an act of creativity, but just the result of having followed his normal flow of thought, which simply has characteristics different from that of most of the population.

We all have a creative mind, but the pressure of sociability pushes us to inhibit part of our potential in order to better understand others and be better integrated into social groups.

İn the full paper we will try to prove our hypothesis by providing a wider analysis of numerous case studies.

Conclusions

The study of autistic cognition is a precious source of information on the usual functioning of human cognition. In fact, it shows the link between attitude to sociality and all the rest of cognitive processes. Autistic cognition teaches us that creativity is not an empyrean concept and that we are always creative with respect to something else.

Creativity with respect to the usual ways of thinking is the active effort to alter our usual flow of thought in order to solve a problem that we are not able to solve with previously used methods. Creativity with respect to society, on the other hand, is a style of thought that deviates from the one that is accepted by the rest of the group. In most people the two things often coincide; but at any moment we have the possibility of exerting an active effort to get rid of a habit of thought and to create one that has not yet been explored yet.

References

American Psychiatric Association., & American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders: DSM-5*. Washington, D.C: American Psychiatric Association.

Chevallier, C., Kohls, G., Troiani, V., Brodkin, E. S., & Schultz, R. T. (2012). The social motivation theory of autism. *Trends in cognitive sciences*, *16*(4), 231-239.

Craig, J., & Baron-Cohen, S. (1999). Creativity and imagination in autism and Asperger syndrome. *Journal of Autism and Developmental Disorders*, *29*(4), 319-326.

Daoust, A. M., Lusignan, F. A., Braun, C. M., Mottron, L., & Godbout, R. (2008). Dream content analysis in persons with an autism spectrum disorder. *Journal of autism and developmental disorders*, *38*(4), 634-643.

Grandin, T. (1995). *Thinking in pictures: And other reports from my life with autism*. New York: Vintage Books.

Hobson, R. P. (2012). Autism, literal language and concrete thinking: Some developmental considerations. *Metaphor and Symbol*, *27*(1), 4-21.

Jarrold, C., Boucher, J., & Smith, P. (1993). Symbolic play in autism: A review. *Journal of autism and developmental disorders*, *23*(2), 281-307.

Low, J., Goddard, E., & Melser, J. (2009). Generativity and imagination in autism spectrum disorder: Evidence from individual differences in children's impossible entity drawings. *British Journal of Developmental Psychology*, *27*(2), 425-444.

Nicoletti, G., (2015). *Una notte ho sognato che parlavi: Così ho imparato a fare il padre di mio figlio autistico*. Milano: Mondadori.

Pennisi, P. (2016). Inferential abilities and pragmatic deficits in subjects with Autism Spectrum Disorders. In *Pragmemes and Theories of Language Use* (pp. 749-768). Springer, Cham.

Rundblad, G., & Annaz, D. (2010). The atypical development of metaphor and metonymy comprehension in children with autism. *Autism*, *14*(1), 29-46.

Selfe, L. (2011). *Nadia Revisited: A Longitudinal Study of an Autistic Savant*. Hoboken: Taylor & Francis.

Tammet, D. (2008). *Born on a blue day: A memoir of Asperger's and an extraordinary mind*. Anstey, Leicester: F.A. Thorpe.

Author's relevant publications

Capone, A., Falzone, A., Pennisi, P. (2018), *Pronominals and presuppositions in that-clauses of indirect reports*, in Capone, A., García-Carpintero, M., Falzone, A. (editors) (2018), *Indirect Reports and Pragmatics in the World Languages*, Cham: Springer, pp. 227-242

Cazzato D., Adamo F., Palestra G. C., Crifaci G., Ruta L., Pioggia G., Pennisi, P., Leo M., Distante C., (2015, November). *Non-intrusive and calibration free visual exploration analysis in children with Autism Spectrum Disorder*. In *Computational Vision and Medical Image Processing V: Proceedings of the 5th Eccomas Thematic Conference on Computational Vision and Medical Image Processing* (VipIMAGE 2015, Tenerife, Spain, October 19-21, 2015) (p. 201). CRC Press. (ISBN 9781315642796)

Falzone, A., Gangemi, A., Pennisi, P., Fabio, R. A. (2015) *Correlations Between Linguistic Phenotype and GeneticAlterations in Rett Syndrome*. «CEUR Workshop Proceedings». O8/2015, Vol 1419, pp. 605 – 610

Pennisi, P. (2016), *What the autistic style of drawing says about the development of language?*, in «Reti, saperi e linguaggi», n. 10, anno 5, 2/2016 (ISSN 2279-7777)

Pennisi, P. (2018), *Mente incarnata e linguaggio: la dimensione aspettuale nella cognizione autistica*, «Lexia», gennaio 2018, nn. 27-28, pp. 465-492. ISSN: 1720-5298.

Pennisi, P. (2018), *Our mind is still inside our skin*, in «RSL, Italian Journal of Cognitive Sciences», 1/2018 a. 7 (13), pp. 19-24, ISSN 1826-8889

Pennisi, P. (2019), "Personal reference in subjects with autism", in Capone, A., Carapezza, M., Lo Piparo, F. (eds.). *Further Advances in Pragmatics and Philosophy Part 2 Theories and Applications*, Cham: Springer

Pennisi, P. (2019), "Research in Clinical Pragmatics: The essence of a new philosophy, the state of the art and future research", in Capone, A., Carapezza, M., Lo Piparo, F. (eds.). *Further Advances in Pragmatics and Philosophy Part 2 Theories and Applications*, Cham: Springer

Pennisi, P., (2016) *Inferential abilities and pragmatic deficits in subjects with Autism Spectrum Disorders*, in Allan, K., Capone, A., Keckes, I., «Pragmemes and Theories of Language Use». Springer: Cham, pp. 749 – 768

Pennisi, P., (2016). *Il linguaggio dell'autismo: studi sulla comunicazione silenziosa e la pragmatica delle parole*. Il Mulino, Bologna, 2016 (ISBN: 978-88-15-26595-

Pennisi, P., (*in press*), "Happiness and unhappiness of performative acts: second language acquisition and psychopathological behaviors" in Pennisi, P.. Falzone A. (eds.) *The Extended Theory of Cognitive Creativity. Interdisciplinary Approaches to Performativity*, Cham: Springer

Pennisi, P., (*in press*), "The contextual, enabling, and constitutive role of physical experience in narratives", in Sinding, M. (ed.) *Narrative, Cognition & Science*

Pennisi, P., Tonacci, A., Tartarisco, G., Billeci, L., Ruta, L., Gangemi, S., Pioggia, G. (2015) *Autism and Social Robotics: a systematic review*, «Autism Research», 2015. Doi 10.1002/aur.1527

Language and event recall in memory for time

Yaqi Wang (yw1308@york.ac.uk)
Department of Psychology, University of York
York, YO10 5DD UK

Silvia P. Gennari (silvia.gennari@york.ac.uk)
Department of Psychology, University of York
York, YO10 5DD UK

Introduction

We remember and talk about events as unfolding over time. When recalling a recent past event, we are often able to recollect its duration and mentally reproduce its component actions, as if replaying the event in our mind. However, key aspects of duration memory are poorly understood, in particular, how event memories map onto clock time and how this mapping is modulated by language. Indeed, we do not perceive and remember objective clock time, unless we pay attention to clocks. Instead, we build event representations that are not replicas of our experiences but are rather temporally compressed, and thus do not often coincide with the real time it took these experiences to unfold.

Here, we investigate the relationship between time, memory and language by examining how people recall and mentally reproduce (replay) events that were conceptualized through language. We specifically ask two main questions. First, what determines the duration and clock accuracy of event reproductions from memory. Second, how these reproductions are modulated by linguistic descriptions, thus potentially leading to distorted reproductions.

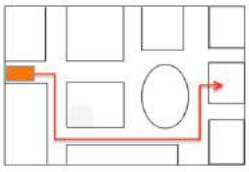


Figure 1: Stimulus example. The arrow indicates path.

In four studies, participants were first asked to study cartoon-like animations accompanied by descriptive phrases for a later memory test. The animations varied in duration from 3-9s and showed geometric figures moving in a familiar setting. Each animation was paired with one of two possible descriptions implying either fast or slow-motion speed, for example, *grandma taking the bus to the hospital* vs *an ambulance taking someone to the hospital* (cf. Fig. 1). Thus, the two phrases implied a shorter or longer event duration. The descriptions provided critical information to understand the animation, which would otherwise be unspecific as to the nature of the moving object. After learning, participants were asked to replay the animations in their minds *exactly as they occurred in their original time course* when prompted with either an animation frame or the corresponding description. Participants clicked the mouse at the start and finish of their mental replays. The reproduced or replayed duration was measured. Finally, they were asked to verbally recall as many details as they could about each animation when prompted.

We hypothesized following Ornstein (1969) recall-based account that the amount of information recalled would explain duration memory over and above the influence of stimulus duration and number of stimulus segments (Faber & Gennari, 2017). Importantly, when the amount of exposure to the same stimuli increases (as in Exps. 2 & 4), we expected that more event information should be learned, and thus, the reproduced duration should lengthen particularly for the stimuli where reproduction accuracy could be improved. Under this hypothesis, memory biases previously reported (Roy & Christenfeld, 2008), such as the tendency to shorten long events and lengthened short ones should emerge from the information recalled: people remember proportionally more information per time unit for short events, leading to lengthened duration reproductions, whereas they remember proportionally less information per time unit for long events, leading to shortened reproductions.

Additionally, we hypothesized following interactive-encoding accounts (Lupyan, 2008; Feist & Gentner, 2007) that if verbal encoding distorts the initially encoded memory, events described with slow-phrases should be mentally reproduced as longer than those described with fast-phrases, even when a visual-cue prompts event replay. Alternatively, retrieval-based accounts (Alba & Hasher,1983) argue that language should only play a role at retrieval, i.e., the effect of descriptions should be observed only when descriptions prompt replays.

Experiments 1 & 2: visually-cued event replays after one and three stimulus viewings

Participants learned 21 animations alongside descriptions either once in random order (exp. 1) or three times (3 runs through the stimuli in random order, exp. 2). Mental replay and verbal recall were cued by a visual frame. Separate pre-tests determined stimulus characteristics (implied motion speed, familiarity, number of segments, etc.).

Results indicated no influence of language on duration reproductions. However, the event information recalled (as measured by the number of words used in verbal recall) modulated reproduced duration. As the number of words recalled increased, so did the reproduced duration, over and above the influence of stimulus duration and the number of stimulus segments. Moreover, the extent to which reproduced duration deviated from the clock stimulus duration (deviation index = reproduced duration/stimulus duration) were

lengthened for shorter animations but were shortened for longer animations, as previously reported (Fig. 2). A deviation index of 1 indicates accurate reproductions, with smaller or larger indices indicating under- or over-reproductions. Moreover, more accurate reproductions (closer to 1) were obtained in Exp. 2 compared to Exp. 1, particularly for longer animations, where more details could be learned with more exposure. This suggests that the number of details recalled underpins duration reproductions. Critically, the density of the details recalled (the number of words recalled per seconds in an animation) explained deviation indices. Thus, shorter animations were reproduced as longer because more details were proportionally recalled for them compared to longer animations, thus providing a possible explanation for the temporal bias observed.

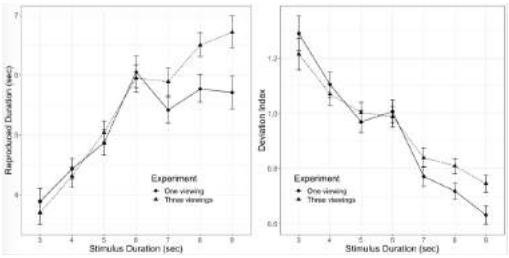


Figure 2: Reproduced duration and deviation index as a function of the stimulus duration

Experiment 3 & 4: verbally cued event replays after one or three stimulus viewings

These studies used the linguistic descriptions to prompt event replays and verbal recall, instead of a visual cue. Note that language may influence reproductions because participants are unsure of what they saw after a single viewing. Thus, testing deeper learning may reveal whether weak memory traces play a role in language effects.

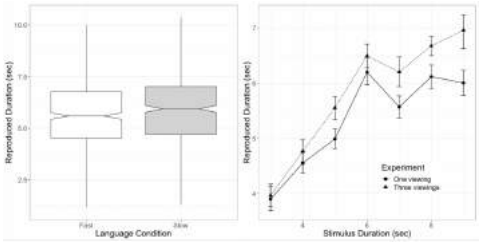


Figure 3: Reproduced duration as a function of language condition and stimulus duration in Exp. 4

Results indicated that for one and three stimulus viewings, there was a language effect. Slow-phrases led to longer reproductions whereas fast phrases led to shorter ones across all stimulus durations. This suggests that the memory representation retrieved is combined with top-down conceptual information present at retrieval, leading to a biased reproduction. In addition, the density of the information recalled predicted deviation indices (temporal bias), replicating exps. 1 & 2. Thus, both language and event recall influenced replays, leading to linguistically and temporally distorted retrieval.

Discussion

We investigated how event reproductions from memory were modulated by event descriptions and the event information recalled. Visually cued event reproductions did not vary as a function of language, suggesting that language did not modulate the way the animations were encoded or subsequently retrieved. Instead, event memory was the main source of information guiding duration reproductions, as evidenced by the predictive role of the number of words used in recall, over and above stimulus duration and segments. Critically, irrespective of cue type, better learning led to longer event reproductions for animations where accuracy could be improved, consistent with the recall-based view.

Verbally-cued reproductions led to shorter or longer reproductions according to the phrases, even after extensive learning. The concurrent influence of recalled information and language, therefore, suggests that the retrieved episodic event representations were combined with linguistic concepts, leading to hybrid event reproductions modulated by both event memory and language.

In all experiments, shorter stimuli were lengthened, and longer stimuli shortened, despite modulations by learning and language. The deviation index in all studies was explained by the information density recalled (the number of words recalled per second). We argue that information density and temporal biases stems from event perception and encoding mechanisms: Information at event boundaries is recalled better than within-event information (Zacks et al., 2007). In longer events, which tend to have longer segments, more within-segment information is forgotten, whereas for short events, which have relatively short segments, more information is proportionally retrieved.

Taken together, these results are consistent with both a recall-based view of memory for duration and a retrieval account of the role of language in memory.

References

Alba, J. W., & Hasher, L. (1983). Is memory schematic? *Psychological Bulletin*, 93(2), 203–231.

Faber, M., & Gennari, S. P. (2015). In search of lost time: Reconstructing the unfolding of events from memory. *Cognition*, 143, 193–202.

Feist, M. I., & Gentner, D. (2007). Spatial Language influences memory for spatial scenes. *Memory and Cognition*, 35(2), 283–296.

Lupyan, G. (2008). From Chair to "Chair": A Representational Shift Account of Object Labeling Effects on Memory. *Journal of Experimental Psychology: General*, 137, 348-369.

Ornstein, R. E. (1969). *On the experience of time*. Harmondsworth, England: Penguin.

Zacks, J., Speer, N., Swallow, K., Braver, T., & Reynolds, J. (2007). Event perception: a mind-brain perspective. *Psychological Bulletin*, 133(2), 273–293.

Evolution and efficiency in color naming: The case of Nafaanra

Noga Zaslavsky[*,1,2] **(noga.zaslavsky@mail.huji.ac.il)**
Karee Garvin[*,2] **(karee_garvin@berkeley.edu)**
Charles Kemp[3] **(c.kemp@unimelb.edu.au)**
Naftali Tishby[1,4] **(tishby@cs.huji.ac.il)**
Terry Regier[2,5] **(terry.regier@berkeley.edu)**

[1]Edmond and Lily Safra Centre for Brain Sciences, Hebrew University, Jerusalem 9190401, Israel
[2]Department of Linguistics, University of California, Berkeley, CA 94720 USA
[3]School of Psychological Sciences, University of Melbourne, Parkville, Victoria 3010, Australia
[4]Benin School of Computer Science and Engineering, Hebrew University, Jerusalem 9190401, Israel
[5]Cognitive Science Program, University of California, Berkeley, CA 94720 USA
[*]Contributed equally

Keywords: language evolution; color naming; efficient communication; information theory

Many theories hold that languages acquire new color terms with time, resulting in finer-grained color naming systems (e.g. Berlin & Kay, 1969; MacLaury, 1997; Levinson, 2000). More recently, it has also been claimed (e.g. Lindsey et al., 2015; Regier et al., 2015; Gibson et al., 2017) that this historical evolutionary process, and color naming more generally, are shaped by the need for efficient communication — that is, the need to communicate accurately, with a simple lexicon. Zaslavsky et al. (2018) [henceforth ZKRT] showed that an independent information-theoretic principle of efficiency, the Information Bottleneck (IB) principle (Tishby et al., 1999), explains much cross-language variation in color naming, and they hypothesized that color naming systems evolve under pressure to remain near the theoretical limit of efficiency. However, most research concerning the evolution of color naming, including ZKRT, has been based on synchronic cross-language comparisons, rather than on diachronic data.

Here, we examine color naming evolution using diachronic data for a single language: Nafaanra, a Senufo language spoken in Western Ghana. Color naming data for Nafaanra were first collected in 1978 in the village Banda Ahenkro, as part of the World Color Survey (WCS, Kay et al., 2009). The data revealed a 3-term system with terms for light/white, dark/black and red. ZKRT found that 93% of the WCS systems, including this one, are near-optimally efficient in the IB sense. Nafaanra data were collected again in Banda Ahenkro by one of us (K.G.) in summer 2017, and revealed a 7-term system. The three terms from 1978 are still used but they now name smaller categories, and there are also new terms for (roughly) yellow, green, blue and purple. These findings are consistent with the claim that languages add new color terms with time. To investigate whether Nafaanra had changed under pressure to remain efficient, we analyzed the 2017 system in the same way ZKRT had analyzed the 1978 system. We found that the 2017 Nafaanra system, like the 1978 system, lies near the theoretical limit of efficiency, and that this outcome would be unlikely without pressure for efficiency. To our knowledge, this is the first evidence that directly supports the proposal that color naming evolves under pressure for efficient communication. How broadly this finding generalizes across languages and domains (Regier et al., 2015), and how efficiency interacts with other factors such as language contact, are questions for future research.

References

Berlin, B., & Kay, P. (1969). *Basic color terms: Their universality and evolution*. Berkeley and Los Angeles: University of California Press.

Gibson, E., Futrell, R., Jara-Ettinger, J., Mahowald, K., Bergen, L., Ratnasingam, S., ... Conway, B. R. (2017). Color naming across languages reflects color use. *PNAS*, *114*(40), 10785-10790.

Kay, P., Berlin, B., Maffi, L., Merrifield, W. R., & Cook, R. (2009). *The World Color Survey*. Stanford: Center for the Study of Language and Information.

Levinson, S. C. (2000). Yélî Dnye and the theory of basic color terms. *Journal of Linguistic Anthropology*, *10*(1), 3-55.

Lindsey, D. T., Brown, A. M., Brainard, D. H., & Apicella, C. L. (2015). Hunter-gatherer color naming provides new insight into the evolution of color terms. *Current Biology*, *25*(18), 2441–2446.

MacLaury, R. E. (1997). *Color and cognition in Mesoamerica: Constructing categories as vantages*. University of Texas Press.

Regier, T., Kemp, C., & Kay, P. (2015). Word meanings across languages support efficient communication. In B. MacWhinney & W. O'Grady (Eds.), *The handbook of language emergence* (pp. 237–263). Hoboken, NJ: Wiley-Blackwell.

Tishby, N., Pereira, F. C., & Bialek, W. (1999). The Information Bottleneck method. In *Proceedings of the 37th annual Allerton conference on communication, control and computing*.

Zaslavsky, N., Kemp, C., Regier, T., & Tishby, N. (2018). Efficient compression in color naming and its evolution. *PNAS*, *115*(31), 7937–7942.

Evaluating Theories of Collaborative Cognition Using the Hawkes Process and a Large Naturalistic Data Set

Mohsen Afrasiabi (afrasiabi@wisc.edu)
University of Wisconsin-Madison
Department of Psychology, 1202 W. Johnson Street
Madison, WI 53706 USA

Mark G. Orr (mo6xj@Virginia.edu)
University of Virginia
Biocomplexity Institute and Initiative
charlotsville, VA 22911

Joseph L. Austerweil (austerweil@wisc.edu)
University of Wisconsin-Madison
Department of Psychology, 1202 W. Johnson Street
Madison, WI 53706 USA

Abstract

People spontaneously collaborate to solve a common goal. What factors affect whether teams are successful? Due to lack of large-scale naturalistic data and methods for investigating scientific questions on such data, previous work has either focused on very concrete cases, such as surveys of business teams, or abstract cases, such as GridWorld games, where agents coordinate their movement so that each agent can get to their own goal without obstructing other agents. We propose a computational framework based on the multivariate Hawkes process and a novel algorithm for parameter estimation on large data sets. We demonstrate the potential of this method by applying it to a large database of programming teams, public GitHub repositories. We analyze factors known to influence team performance, such as leader organization style and team cognitive diversity, as well as other factors, such as the burstiness of effort, that are difficult to test using existing methods.
Keywords: Collaborative cognition; Hawkes process; Organizational psychology; Bayesian nonparametrics

Introduction

People naturally form groups to collaborate towards a common goal. We coordinate to navigate the world (Ho et al., 2016), to protest inequalities (Korkmaz et al., 2018), to increase efficiency and well-being (Simon, 1991), to solve problems (Miller, 1951) and crises (Militello et al., 2007), to conduct science (Wuchty et al., 2007) and for many other goals. Previous work on *collaborative cognition* tends to either focus on case studies, such as using surveys of company employees (Kozlowski & Ilgen, 2006), or abstract situations, such as game-theoretic analyses of whether to cooperate or defect in the Prisoner's Dilemma (Rand & Nowak, 2013). Although these methods have been drastically increased our understanding of collaboration and competition (e.g., what mechanisms promote cooperation in competitive scenarios; Kleiman-Weiner et al. 2016; Rand & Nowak 2013), there is a need to bridge this gap. In this paper, we propose a large-scale natural data set and computational framework for analyzing human collaboration.

Technical approaches for theoretical development, conceptualization, and modeling of collaborative cognition come in many forms, each with specific strengths and weakness. For example, agent-based simulation can represent individuals interacting in dynamic network structures, but suffer from issues, such as computational difficulties in scaling the number of agents to realistic numbers, the number of free parameters (whether in model choice or explicit parameters), what the right level of abstraction should be, and how to evaluate them with respect to empirical data. This methodology has been extremely powerful, for example, it is unclear we would have discovered without these models that cooperation in the Prisoner's Dilemma can emerge from natural selection when the agents play according to how they are networked (Ohtsuki et al., 2006). But due to the simplifications, it is unclear whether this approach can be applied to any phenomena of interest (Louie & Carley, 2008).

In this paper, we focus on one aspect of collaborative cognition: how teams act as if they are a single mind when solving a common task (Searle, 1995; Bacharach et al., 2006). There are two major challenges facing collective cognition research on this perspective: (1) a lack of naturalistic data of real-world problems in the process of being solved and (2) a lack of formal methods for evaluating such data, which are richly-structured discrete data over continuous time (Kozlowski et al., 2016). For example, recent work has explored how pairs of agents can learn to coordinate and generalize their coordination in "Grid Worlds" – an environment consisting of a grid, two circle avatars in the grid, and two goals that the avatars try to get to without impeding each other (Austerweil et al., 2016; Ho et al., 2016). To address the first problem, we propose analyzing projects (called repositories) on GitHub, an online social coding platform, as a source of large-scale, naturalistic data of humans self-organizing towards solving a common goal. To address the second problem, we propose using the multivariate Hawkes process (Hawkes, 1971), a Bayesian nonparametric process, that, unlike Poisson processes, can capture the bursty nature of work on GitHub. To do so, we derive a novel approximation technique that can estimate parameters for a set of richly structured discrete data.

Introduction to GitHub

GitHub is an online social coding platform. Users can create projects, called repositories, which are publicly accessible. It is built on the decentralized software version control platform `git`. Each `git` user of a repository has a full-fledged version of the project and full control of their local version. They then can share their changes to others working on the project who can decide whether to merge them into their own repository.

Given how decentralized projects managed by `git` are and the importance of clear leadership for project success in some tasks from empirical research in Industrial and Organizational Psychology (Kozlowski & Ilgen, 2006; D. Wang et al., 2014), one may be surprised that GitHub is one of the most popular platforms for collaborative programming projects. This is because GitHub affords coordination with other team members in a few ways. (1) Only some members are "owners" of the repository, who are allowed to accept proposed changes to the project (any owner can make another member an owner – the original creator is the first owner of the repository), (2) a set of `Events` that keep track of actions taken by each member to global repositories, and (3) conversations through different media, such as e-mail lists or Reddit. Although the third method of coordinating is important, we leave it for future research. We will focus on repository ownership and events to analyze collaborative cognition on GitHub.

There are six main types of `Events` that we focus on: `CreateEvent`, `ForkEvent`, `DeleteEvent`, `PullRequestEvent`, `PushEvent`, `IssueEvent`, and `WatchEvent`. Every event is stored with the time when it occured. Some event types have subtypes that enable team members to discuss the event. A `CreateEvent` occurs when someone creates a new repository or (more commonly) creates a new "branch", which is a copy of the project attached to the main one. Branches are often used to prototype new features. Sometimes the prototype works and a team member proposes incorporating it back into the main project, which is a `PullRequestEvent` (an owner then either accepts or rejects the merger, sometimes after comments from different members). Sometimes the prototype does not work, in which case it gets deleted, which is catalogued by a `DeleteEvent`. A `PushEvent` occurs when someone updates a file in the main public repository. Team members that discover problems or want to raise other issues can do so with an `IssueEvent`. Finally, anyone interested in a project can get regular updates to any changes by "watching" the repository. Whenever a new person watches the repository, a `WatchEvent` occurs. Although these events do not catalogue all work by a team, they provide a lot of information about how team members collaborate and develop a project. We will analyze them to test theories of collaboration, but first we present our computational framework.

A Computational Framework for Teamwork

We formulate our model as a Bayesian nonparametric Point Process. It is a multivariate Hawkes process, where the dimensions correspond to the different types of Events and marks correspond to the properties of the Event. For example, an IssueEvent will be one dimension in the multivariate Hawkes process, and values of the IssueEvent (such as the user, the repository, etc) are all part of the mark.

In this section, we first define Stochastic Marked Non-Homogeneous Poisson Point Processes. Next, we define the univariate Hawkes process with a simple mark. Then, we formulate a multivariate Hawkes process. Throughout, we will introduce notation that will become increasingly catered to the special case of modeling GitHub.

Stochastic Marked Non-Homogeneous Poisson Point Processes

A *Marked Point Process* is a sequence of marked random points, where each point $H_i = (t_i, e_i)_{i=1,\dots}$ is composed of a continuous-valued time value ($t_i \in \mathcal{R}_+$, positive real numbers) and a *mark* ($e_i \in \mathcal{E}$, an arbitrary event space $\mathcal{E}$). For the specific case of modeling GitHub, marks are multivariate points taking values in the space, $\{1,2,\dots,E\} \times \{1,2,\dots,U\} \times \{1,2,\dots,R\}$, where E is the number of Event Types, U is the number of agents, and R is the number of repositories.[1] The framework allows for observed mark types to influence the rates of EventTypes, which will be important for capturing dependencies betweeen EventTypes. For example, a PushEvent is more likely after a CreateEvent then a WatchEvent.

A *Non-homogeneous Marked Poisson Point Process* is a special case of a Marked Point Process, where the number of points in a period of time $[a,b]$ is Poisson distributed with parameter $\int_a^b \lambda(t)dt$. $\lambda(t)$ is an intensity function or the instantaneous rate for points to arrive at time t. To capture relations between EventTypes, agents, and repositories, $\lambda_\theta(t)$ will be dependent on $\theta = (e,u,r)$, which corresponds to the rate of users u producing events of type e in repository r. The interactions between the stream of events for users in different repositories can be distributions other than *Possion*. They are defined as appropriate for the domain, which is how we will include psychologically-based representations in future work. For this article, we assume each repository, event types, and users are marked processes with empirical distributions extracted from real repository data.

Multivariate Hawkes Process with Agent Types, Repositories, and Communities

In the models discussed above, all events arrive independently, either at a constant rate (for Poisson process) or governed by an intensity function (for the non-homogeneous Poisson processes). In both cases, they are independent of events that previously occurred. However, in social environments, the arrival of an event increases the likelihood of observing events in the future. To model this phenomena we use a *Hawkes Point Process* with a *self-exciting* kernel in which an event arrival explicitly depend on past events (Hawkes, 1971). A *Point Process* is a *Hawkes Process* if the conditional intensity function $\lambda_r(t|H_i = (t_i, e_i)_{i=1,\dots})$ is:

$$\lambda_r^*(t) = \lambda_r(t|H_1,\dots,H_n) = \lambda_{r,0}(t) + \sum_{i:t>t_i} \phi(t-t_i;\beta) \qquad (1)$$

[1] Technically, the number of users and repositories are random variables themselves. Then the second and third dimension of the mark would each be counting processes. $U(t)$ could encode the number of users at time t and the probability of a point having a value on the second-dimension beyond $U(t)$ is null. The same can be done for repositories.

where $\lambda_{r,0}(t)$ is the repository intensity based on prior or exogenous information. The events generated from $\lambda_{r,0}(t)$ are called *immigrant* events. Note that when $\phi = 0$, we recover a *Poisson Process*. $\phi(t;\beta)$ is a kernel function and typically decays with increasing t and β are its parameters. The most common decay function is the scaled exponential taking the following form: $\phi(t;\alpha,\omega) = \alpha\omega\exp\{-\omega t\}$, where $\beta = (\alpha,\omega)$, $\alpha \geq 0$ and $\omega > 0$ and $\alpha < \omega$. Another widely used kernel for modeling social behavior is the power-law function: $\phi(t;\alpha,\eta,\gamma) = \alpha(t+\gamma)^{-(\eta+1)}$, where $\alpha \geq 0$, $\gamma > 0$, $\eta > 0$ and $\alpha < \eta\gamma^{\eta}$.

After observing an event, the intensity is large for some time and then decays to zero. Thus, more recent events influence the current event's intensity more than older events. This results in a *self-excitatory* process, where bursts of points in a small time period lead to a large increase in intensity in that region. By defining $\phi(t)$ differently, it is also possible to capture *self-inhibiting* processes (Yang et al., 2015), which will be important in capturing an user waiting for other users (e.g., respond to an `IssueEvent`). Both properties violate the *memoryless* property, and thus, Hawkes processes capture a broader set of Point Processes then standard nonhomogeneous Possion Processes.

As our model is multi-user, multi-event and multi-repository we will use the *multivariate* formulation of the Hawkes process. The basic assumption behind the multivariate Hawkes process is that the arrival of an event in one dimension can affect the arrival rates of events in other dimensions according to some generative process. The specification of the generative process can be as richly structured as appropriate for the domain. This enables analysis of structured discrete data over continuous events. We model this dependence in the following manner: each repository is a Hawkes process, the Hawkes processes for repositories are interdependent, and the event types and users as marks. In this paper, we use pairwise correlations to capture repository interdependence and the joint probability of pairs of Event Types is estimated from our data set.

Using an exponential kernel function, the conditional intensity $\lambda_r^*(t)$ is:

$$\lambda_r^*(t) = \lambda_{r,0}(t) + \sum_{i:t>t_i} \alpha_{r_i,r}\omega_{r_i,r}\exp(\omega_{r_i,r}(t-t_i)), \qquad (2)$$

where $\alpha_{r_i,r}$ is an interactivity matrix defining how the r_i dimension influences the r dimension given the values of features across the different dimensions at time t. We approximate this matrix via maximum likelihood estimation. The likelihood of repository r with parameter set $\beta = (\alpha,\omega)$ and λ_0 is (Ozaki, 1979):

$$l_r = \exp\left\{-\int_0^T \lambda_r(t|\{t_j\}_{j=1}^N)dt\right\}\prod_{i=1}^N \lambda_r(t_i|\{t_j\}_{j=1}^{i-1}) \qquad (3)$$

and the log-likelihood, with some simplification, is:

$$\log l_r(\{t_i\}|\eta_r) = -\lambda_{r,0}T + \sum_{i=1}^N \alpha_r(\exp(-\omega_r(T-t_i))-1) + \sum_{i=1}^N \log(\lambda_{r,0} + \alpha_r\omega_r\Omega_r(i))$$

where $\Omega_r(i) = \sum_{t_j<t_i}\exp(-\omega_r(t_j-t_i))$, $\forall i \geq 2$ and $\Omega_r(1) = 0$.

Unfortunately we cannot optimize the log-likelihood directly, because the curvature vanishes. So, we estimate the parameters by extending a version of *Maximum a Posteriori Expectation Maximization* (Zipkin et al., 2016). Let $\tau = (t_i)$ be the sequence of actions performed on a repository and $M = M_{ij}$ be a branching matrix of an immigrant event, where $M_{ij} = 1$ if event i is an offspring of event j. M is the causal cascade structure of sequence of actions performed in a repository. Let $p(\Upsilon;F)$ be a prior on $\Upsilon = (\eta,\lambda_0)$ with hyperparameter F. We perform *MAP* estimation using the EM algorithm to maximize the event stream posterior, $p(\Upsilon|\tau,M) \propto p(\tau,M|\Upsilon)p(\Upsilon|F)$. Let $\log P(\tau,M|\Upsilon,F) = \log p(\tau,M|\Upsilon) + \log p(\Upsilon|F)$ be the event stream probability. We decompose the first term in the following manner: $\log p(\tau,M|\Upsilon) = \mathfrak{L}_1(\lambda_0,\tau) + \mathfrak{L}_2(\eta,\tau) + \mathfrak{L}_3(\eta,\tau)$ where

$$\begin{aligned}
\mathfrak{L}_1(\lambda_0,\tau) &= -\lambda_0 T + b(\log\lambda_0 + \log T) - \log m! \\
\mathfrak{L}_2(\eta,\tau) &= -n\Phi(\eta) + \sum_i d_i\Phi(\eta) - \log m_i! \\
\mathfrak{L}_3(\eta,\tau) &= \sum_{ij} M_{ij}[\log\phi(t_i-t_j;\theta) - \log\Phi(\theta)]
\end{aligned}$$

where $m = \sum_i M_{ii}$, $m_i = \sum_j M_{ij}$, and $\Phi(\eta) = \int_0^\infty \phi(t;\eta)dt$.

$$\log p(\tau,M;\Upsilon) = -\lambda_0 T + m\log\lambda_0 + b\log T - \log(m!) + \sum_i[-\Phi(\eta) + m_i\log\Phi(\eta)\log(m_i!)] + \sum_{ij} M_{ij}\log\phi(t_i-t_j;\eta) - \log\Phi(\eta)$$

In the E-step of the MAP EM algorithm, we compute the current distribution over M. As M is a matrix of branching variables, each is Bernoulli and so M can be expressed as the expected branching matrix $P = [p_{ij}]$ based on the data τ and our current parameter estimate Υ^k. The expected branching matrix at each iteration is $P^{k+1} = \mathbb{E}[M|\tau,\Upsilon^k]$. In the M-step, we update our parameter estimate to maximize the expectation of the event stream posterior log-likelihood:

$$\begin{aligned}
\Upsilon^{k+1} &= \arg\max_\Upsilon \mathbb{E}[\mathfrak{L}(\tau,M;\Upsilon,F)|M = P^{k+1}] \\
&= \arg\max_\Upsilon (\mathbb{E}[\log p(\tau,M;\Upsilon)|M = P^{k+1}] + (\mathbb{E}[\log p(\Upsilon,F)])
\end{aligned}$$

We use a Gamma prior on α and ω, with parameters (s,t) and (u,v), respectively. Extending the method in Zipkin et al. (2016), the EM update steps can be derived using the immigrant/offspring interpretation. The ith event is either an immigrant or an offspring of one of the previous events. The probability that the ith event is an immigrant event is proportional

to λ_0^k, while the probability that it is an offspring of event j for $j < i$ is proportional to the kernel function $\phi(t_i - t_j; \alpha^k, \omega^k)$. The E-step update then is

$$P_{ij}^{k+1} = \begin{cases} \frac{1}{\Lambda^k(i)} & \text{for } i = j \\ \frac{1}{\Lambda^k(i)} \phi(t_i - t_j; \alpha^k, \omega^k) & \text{for } j < i \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

where the normalization factor is $\Lambda^k(i) = \lambda_0^k + \sum_{j<i} \phi(t_i t_j; \alpha_k, \omega_k)$. Finally the M-step is

$$\mu^{k+1} = \frac{1}{T} \sum_i P_{ii}^{k+1} \quad \alpha^{k+1} = \frac{1}{n+l} [\sum_{j<i} P_{ij}^{k+1} + s - 1] \quad (5)$$

$$\omega^{k+1} = \frac{\sum_{j<i} P_{ij}^{k+1} + s - 1}{\sum_{j<i} P_{ij}^{k+1} (t_i - t_j) + v} \quad (6)$$

Analyzing Teamwork on GitHub

We now present how GitHub can be used as a naturalistic, large-scale data set and the Hawkes process to analyze the dynamics of collaborative cognition. We used a data set of events from public repositories on GitHub at the start of midnight on March 1st 2017 to 11:59pm on August 31st 2017. We retrieved 456,195 events across 8,083 repositories.

One issue is that not all repositories are collaborative projects. For example, many repositories are used for web pages, software tutorials (e.g., learning how to fork repositories), and other personal usage. Further, many projects become inactive and abandoned without being deleted. We follow best practices for studying GitHub repositories from previous work in computer science (Kalliamvakou et al., 2016) by filtering repositories according to the following criteria: (1) there are at least 10 Events (not counting `WatchEvent`) in the data set, and (2) at least three unique "active" users. We define an active user of a repository to be someone who had at least one `CreateEvent` or `PushEvent` with it. Using these criteria, our filtered data set was comprised of 390,277 events across 1,235 repositories. This leaves us with 86% and 15% of the total events and repositories, respectively.

Are Hawkes Processes Really Necessary?

Before testing collaborative cognition hypotheses, we provide some justification for using a more complex process, a Hawkes process, rather than a standard Poisson process. From a qualitative perspective, Figure 2 shows the stream of events over time from a representative project and the best fits from a Poisson process and a Hawkes process using an exponential and power-law kernel. Due to its memorylessness property, the Poisson process is simply unable to recreate the bursty dynamics of the event stream. For our data, the Hawkes process with an exponential kernel provides the best qualitative and quantitative fit. Thus, for the remainder of the paper, we only consider the Hawkes process with an exponential kernel. A quantitative comparison of the model fits is computationally challenging due to the large number of repositories. Thus, we approximated by calculating the root

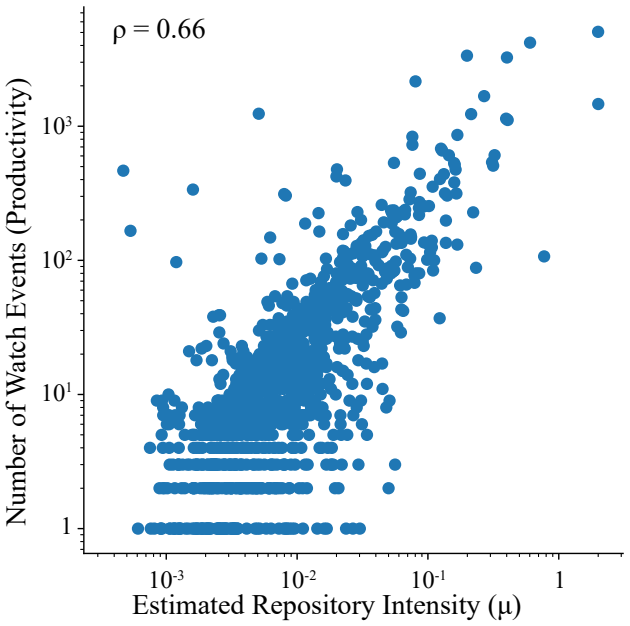


Figure 1: The repository intensity (μ) of the Hawkes Process as estimated from the GitHub data. It corresponds closely to the productivity of the repository.

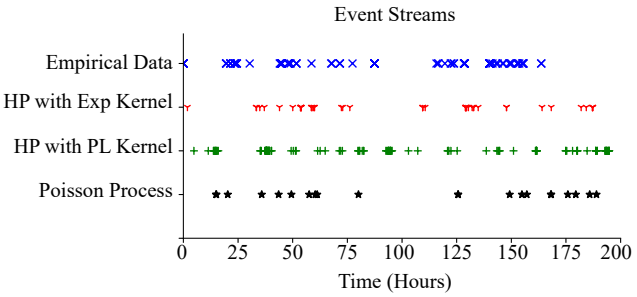


Figure 2: A representative GitHub event stream and samples from a best-fit Poisson Process and Hawkes Processes with an exponential and a power-law kernel.

mean squared error (RMSE) of 200 randomly sampled repositories and then 200 randomly sampled events within each of those repositories. The approximate RMSE for the Hawkes and Poisson processes were 7.27 and 11.81. Further, Figure 1 the number of watch events is closely related to the estimated repository intensity ($\rho = 0.66, p < 0.001$), validating our novel estimation procedure.

.

Testing collaborative cognition

We now turn to testing three different phenomena in collaborative cognition and assess how they affect performance: leadership organization style, diversity, and event dynamics. There is no clear definition of what makes a repository successful on GitHub (especially one that can be automatically applied to all repositories). We use the number of

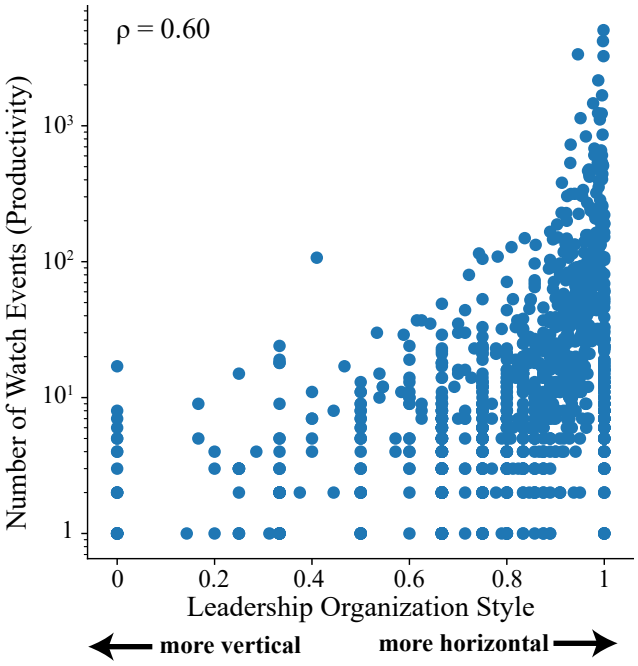


Figure 3: Shared leadership is more successful.

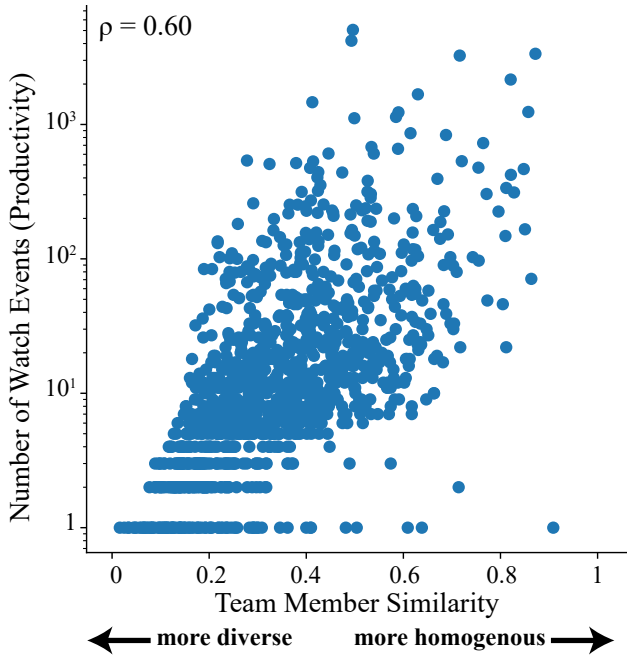


Figure 4: Teams with less cognitive diversity are more productive. The cognitive role of a team member was quantified as the distribution of event types that they produced.

`WatchEvents` for a repository in the six month period as a measure of project success. When a person chooses to *watch* a repository, it means they receive regular updates on any changes to the repository. These are people who are interested in the progress of a project, but do not necessarily contribute to it. In fact, they probably do not, as previous work found that only about 5% of people who watch a repository end up contributing to it (Sheoran et al., 2014).

Leadership organization style. Previous survey studies and meta-analyses of them have found that shared leadership (what we call "horizontal") is positively associated with group performance (D. Wang et al., 2014). We test whether this relationship holds in our large-scale, naturalistic collaboration data set. Team members in a repository are split into two groups: *owners* and *users*. Users can create their own version of a project and build on it on their own. However, they can only propose changes to the global repository (or the team's project). We define leadership style as the percentage of active users who are not owners that work on the project. Lower scores imply a vertical leadership style, where only a few team members are leaders. Larger scores imply a horizontal leadership style, where most team members are leaders. As shown in Figure 3, most teams are horizontally organized and there is a strong positive relation between horizontal organization and performance ($\rho = 0.60, p < 0.001$).

Cognitive Diversity. How does the diversity of roles within a team affect performance? Recent work found that diversity of roles (cognitive diversity) is positively related to team creativity when there are leaders that serve as role models for other team members, but negatively associated otherwise (X.-H. Wang et al., 2016). Given that we found higher performance in programming projects when the leadership style was more distributed, we expect that cognitive diversity may hurt productivity on GitHub, rather than enhance it.

To assess the role of cognitive diversity in team performance on GitHub, we quantified the similarity between two users as the inner product of the distributions of events produced by each user across all repositories. The diversity score of a repository was defined to be the average pairwise similarity of active repository users. Due to computational constraints, for repositories with many users, we approximated the quantity by averaging 10,000 randomly selected pairs of users. Figure 4 shows that teams with less diverse roles performed better ($\rho \approx 0.60, p < 0.001$).

Bursts. Are particular leadership organizations related with differences in how bursty the team's progress is on the project? Is burstiness related to performance? Thanks to the Hawkes process formalism, we can address this question by examining the relation between leadership style and the fit α parameter associated with the repository. Interestingly, Figure 5 shows that more centralized leadership organization is associated with burstier progress ($\rho = 0.39, p < 0.001$). However, burstiness has only a very weak effect on performance ($\rho = -0.13, p < 0.001$). Note that this analysis was only possible to conduct due to the computational formalism for analyzing teamwork presented in this paper.

Discussion, Limitations, and Conclusions

In this article, we proposed, validated, and used a novel computational framework for analyzing large-scale real-world

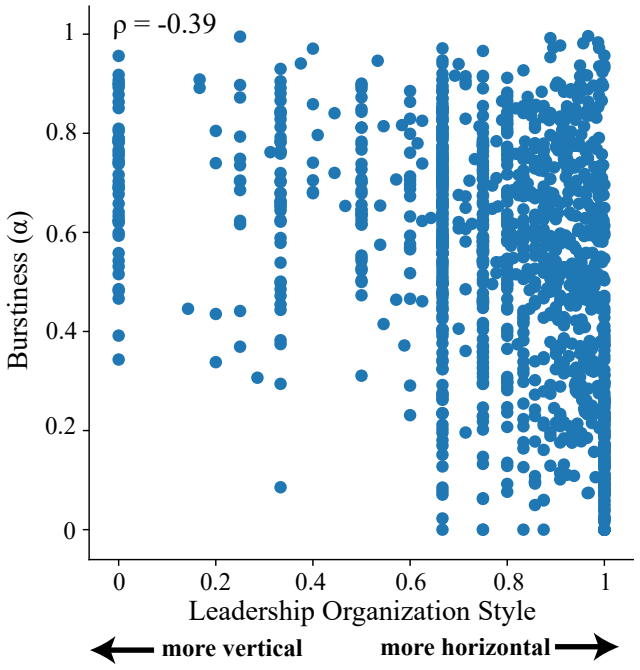


Figure 5: More vertically organized leadership styles are burstier.

collaboration data: The Multivariate Hawkes Process. We demonstrated how it can be used to test constructs in collaborative cognition. For example, we found that horizontal leadership structures were more successful. This may be specific to programming projects that naturally break into different pieces that can be worked on individually and integrated later. Future work will need to follow up on this and the other findings

As a proof of concept, we made a number of assumptions and simplifications. We assumed the only relation between events and teams are pairwise correlations. Further, we ignored an event's content, focusing on statistical patterns. In future work we plan to extend our work to address these limitations and incorporate social and cognitive principles (e.g., scripts for how events usually occur on GitHub; Schank & Abelson 1977), and examine whether the framework generalizes to analyzing other social domains (e.g., Reddit). Recent work suggests cognitive structures, such as shared memory, are essential for understanding team performance (DeChurch & Mesmer-Magnus, 2010). Additionally, we assumed that our results generalize to all task types solved by teams. However, psychologists have organized task types into ontologies (Wildman et al., 2012), and we plan to examine whether our results generalize across tasks. Shared programming projects may lend themselves more naturally to distributed, horizontal leadership structure, whereas a clear leader or established organizational identity may be needed to solve other tasks, such as putting out a fire (Mesmer-Magnus et al., 2018).

Our computational framework is built using probabilistic modeling. This enables us to conduct principled analyses that would otherwise be difficult or impossible in other frameworks. Recent work has analyzed determining automated interventions on social media using a similar probabilistic modeling framework (Farajtabar et al., 2017). For example, using point processes and Markov decision processes, Farajtabar et al. (2017) created a method for mitigating the spread of Fake News through online social networks. We are excited to adapt these techniques into our framework, which would enable us to see how intervening on GitHub repositories (e.g., stopping support for TensorFlow) or counterfactual questions (e.g., how would machine learning applications be affected if TensorFlow were never made public).

Acknowledgments

The research is based upon work supported by the Defense Advanced Research Projects Agency (DARPA), via the Air Force Research Laboratory (AFRL; contract #: FA8650-18-C-7826). The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA, the AFRL or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

References

Austerweil, J. L., Brawner, S., Greenwald, A., Hilliard, E., Ho, M. K., Littman, M. L., ... Trimbach, C. (2016). The impact of outcome preferences in a collection of non-zero-sum grid games. In *AAAI Spring Symposium 2016 on Challenges and Opportunities in Multiagent Learning for the Real World.*

Bacharach, M., Gold, N., & Sugden, R. (2006). *Beyond individual choice: Teams and frames in game theory*.

DeChurch, L. A., & Mesmer-Magnus, J. R. (2010). The cognitive underpinnings of effective teamwork: A meta-Analysis. *Journal of Applied Psychology*, *95*(1), 32–53.

Farajtabar, M., Yang, J., Ye, X., Xu, H., Trivedi, R., Khalil, E., ... Zha, H. (2017). Fake news mitigation via point process based intervention. In *ICML*.

Hawkes, A. (1971). Spectra of some self-exciting and mutually exciting point processes. In *Biometrkia* (pp. 89–90).

Ho, M. K., MacGlashan, J., Hilliard, E., Trimbach, C., Brawner, S., Gopalan, N., ... Austerweil, J. L. (2016). Feature-based joint planning and norm learning in collaborative games. In *Proceedings of the 38th annual meeting of the cognitive science society.*

Kalliamvakou, E., Gousios, G., Blincoe, K., Singer, L., German, D. M., & Damian, D. (2016). An in-depth study of the promises and perils of mining GitHub. *Empirical Software Engineering*, *21*, 2035–2071.

Kleiman-Weiner, M., Ho, M. K., Austerweil, J. L., Littman, M. L., & Tenenbaum, J. B. (2016). *Coordinate to coop-*

erate or compete: Abstract goals and joint intentions in social interaction.

Korkmaz, G., Monica, C., Kraig, A., Lakkaraju, K., Kuhlman, C. J., & Vega-Redondo, F. (2018). Coordination and common knowledge on communication networks. In *Proceedings of the 17th international conference on autonomous agents and multiagent systems* (pp. 1062–1070).

Kozlowski, S. W. J., Chao, G. T., Grand, J. A., Braum, M. T., & Kuljanin, G. (2016). Capture the multilevel dynamics of emergence: Computational modeling, simulation, and virtual experimentation. *Organizational Psychological Review*, *6*(1).

Kozlowski, S. W. J., & Ilgen, D. R. (2006). Enhancing the effectiveness of work groups and teams. *Psychological Science in the Public Interest*, *7*(3), 77-124.

Louie, M. A., & Carley, K. M. (2008). Balancing the criticisms: Validating multi-agent models of social systems. *Simulation Modelling Practice and Theory*, *16*, 242–256.

Mesmer-Magnus, J. R., Asencio, R., Seely, P. W., & DeChurch, L. A. (2018). How organizational identity affects team functioning: The identity instrumentality hypothesis. *Journal of Management*, *44*(4), 1530–1550.

Militello, L. G., Patterson, E. S., Bowman, L., & Wears, R. (2007). Information flow during crisis management: challenges to coordination in the emergency operations center. *Cognition, Technology & Work*, *9*(1), 25–31.

Miller, G. A. (1951). *Language and communication*.

Ohtsuki, H., Hauert, C., Lieberman, E., & Nowak, M. A. (2006). A simple rule for the evolution of cooperation on graphs and social networks. *Nature*, *441*, 502–505.

Ozaki, T. (1979). Maximum likelihood estimation of hawkes' seld-exciting point processes. In *Annual institure of statistical mathematics 31* (pp. 145–155).

Rand, D. G., & Nowak, M. A. (2013). Human cooperation. *Trends in Cognitive Sciences*, *17*(8), 413-425.

Schank, R., & Abelson, R. (1977). *Scripts, plans, goals, and understanding*. LEA.

Searle, J. R. (1995). *The construction of social reality*.

Sheoran, J., Blincoe, K., Kalliamvakou, E., Damian, D., & Ell, J. (2014). Understanding "Watchers" on GitHub. In *Proceedings of the 11th working conference on mining software repositories* (p. 336-339). NY, NY: ACM.

Simon, H. A. (1991). Organization and markets. *Journal of economic perspectives*, *5*(2), 25-44.

Wang, D., Waldman, D. A., & Zhang, Z. (2014). A meta-analysis of shared leadership and team effectiveness. *Journal of Applied Psychology*, *99*(2), 181–198.

Wang, X.-H., Kim, T.-Y., & Lee, D.-R. (2016). Cognitive diversity and team creativity: Effects of team intrinsic motivation and transformational leadership. *Journal of Business Research*, *69*, 3231–3239.

Wildman, J. L., Thayer, A. L., Rosen, M. A., Salas, E., Mathieu, J. E., & Rayne, S. R. (2012). Task types and team-level attributes: Synthesis of team classification literature. *Human Resource Development Review*, *11*(1), 97–129.

Wuchty, S., Jones, B. F., & Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science*, *316*, 1036–1039.

Yang, Q., Wooldridge, M. J., & Zha, H. (2015). Trailer generation via a point process-based visual attractiveness model. In *ICAI* (pp. 2198–2204).

Zipkin, J. R., Schoenberg, F. P., Coronges, K., & Bertozzi, A. L. (2016). Point-process models of social network interactions: parameter estimation and missing data recovery. In *European journal of applied mathematics* (p. 27(03):502529).

Measuring Programming Competence by Assessing Chunk Structures in a Code Transcription Task

Noorah Albehaijan[1,2] (N.albehaijan@sussex.ac.uk, naalbehaijan@iau.edu.sa)
Peter C-H. Cheng[1] (p.c.h.cheng@sussex.ac.uk)
[1]Department of Informatics, University of Sussex Brighton, BN1 9QJ, UK
[1,2]Department of Computer Science, Imam Abdulrahman Bin Faisal University Jubail, P.O. Box 12020, Saudi Arabia

Abstract

In a simple transcription task in which sections of Java program code are copied by freehand writing, it is demonstrated that chunk related temporal signals are sufficiently robust to permit the measurement of programming competence. An experiment with 24 participants revealed that the number of views of the stimulus per trial and the duration of writing per stimulus view are both strongly correlated with independent measures of Java competence.

Keywords: Chunking, program comprehension; competence measurement; transcription.

Introduction

Chunking (Miller, 1956; Cowan, 2001) underpins cognition in tasks that involve information of any complexity. Many phenomena are explained by the notion. For instance, at a long timescale, chunk acquisition explains many of the elevated abilities of experts over novices (e.g., Chase & Simon, 1973; Egan & Schwartz, 1979). At medium timescales, learning relies on the acquisition of chunks (Gobet et al., 2001). The organization of chunks changes during learning with the accretion of new chunks and the restructuring of networks of chunks. At short timescales, the structure of chunks in memory is one substantial factor in the control of routine sequential behaviour, such as the writing of memorised sentences (Cheng & van Genuchten, 2018) or the drawing of geometric diagrams (Obaidellah & Cheng, 2015).

All this suggests that it should be feasible to assess a learner's understanding or competence in a particular knowledge domain by evaluating behavioral measures that are dependent on the underlying structure of that learner's chunk network. And that such assessments can be done using simple production tasks, such as the written transcription of text or formulas, or the copying of diagrams.

Various studies have shown that certain measures of the distribution of the durations of inter-stroke pauses provide feasible measures of competence (Cheng, 2014, 2015; Cheng & Rojas-Anaya, 2007; van Genuchten et al., 2009; Zulkifli, 2013). An *inter-stroke pause* is the time that the pen is off the paper between written strokes, which provides measures at times scales in the range of 100 ms to 1 second . These studies typically used simple transcription tasks, in which the participants copied simple stimuli in each trial, such as a mathematical equations or one English sentence. Strong correlations with independent measures of domain comprehension were found. Further, the relative difficult of stimuli were clearly related to the magnitude of the pause measures. These findings were obtained across diverse domains (algebraic formulas and natural language), classes of users (children and adults) and interface media (pen on paper and on screen mouse driven symbol selection).

Pause measures in typewriting, keystroke logging, have been used extensively to study writing behaviour and performance (e.g., Spelman Miller & Sullivan, 2006), but this requires the aggregation of relatively large amounts of data in order to find effects. Also, our pilot experiments have shown that individual differences, such as variations in typing strategy and skill, tend to obscure the temporal chunk signals. So, inter-keypress pause measures do not appear to be reliable.

What other behaviors might provide strong and robust temporal chunk signals that can serve as a measure of comprehension? Can the scope of chunk-based measures of comprehension be extended to other domains beyond mathematics and natural language? The present experiment addresses these questions.

As chunking is important in the doing and learning of programming (e.g., Shneiderman, 1976; McKeithen, et al., 1981; Pennington, 1987), here we will focus on the assessment of learners' comprehension of programming code, specifically Java. Some studies have used response times to study programming comprehension in whole tasks, such as sets of multiple choice questions, lasting minutes (e.g., Adelson, 1981, 1984; Ye & Salvendy, 1996). Here, the focus is on the time required for component activities within a task, rather than overall task time, and the examination of process durations that may directly depend upon the chunks possessed by participants.

Again we will use a transcription task, as in the experiments cited above. In those experiments, typically, the stimulus was presented on a card or computer-screen placed near a writing tablet, so that the participants could switch their gaze between the stimulus and the tablet. In this experiment we will record when the participant switches between the stimuli and tablet using a participant-driven "hide-show" interaction method. The stimulus appears on the computer screen when the participant holds down a special button. To write the participant must release the button and the stimulus is masked. This extends the repertoire of techniques that may be used to assess chunk structures with a method that targets the processing of several chunks, at a 10 s timescale, which contrasts to the previous methods that analyse elements within a single chunk.

This method makes available various measures: (a) *view-numbers* – the total number of views of the stimulus in a trial;

(b) *writing-times* – the time spent writing between two successive views; (c) *view-times* – the duration of each look at the stimulus.

Various predictions can be derived for these measures. Experts perceive the stimuli using larger chunks than novices. Assume that working memory capacity for chunks does not vary substantially with expertise, which is plausible given that transcription is a relatively complex task (Cowan, 2001) rather than a simple decision making or capacity test (Miller, 1956). So, as the size of a stimulus is fixed, we predict:

H1) View-numbers*: the number of views of the stimulus in a trial will be less for more competent participants.*

As more competent participants' chunks contain more content, we predict:

H2) Writing-times*: the duration of written responses after each stimulus view will be longer for more competent participants.*

This assumes that writing speed is independent of expertise in the target domain, which is plausible for adult participants. Now, as the time to perceive a chunk is approximately constant (Chase & Simon, 1973), and if the number retained per view is independent of competence, then we predict:

H3) View-times*: the time spent on each separate view of the stimuli will not be directly related to competence.*

Frequently used components of Java are introduced earlier during instruction, so we predict:

H4) The performance on basic stimuli will be superior to advanced stimuli, with fewer view-numbers and longer writing-times, but no impact on view-time.

Note that H3 is framed negatively, so care is required to interpret data that might support it. In particular, the magnitude of other effects must be strong so that the likelihood of the absence of an overall view-time effect is not merely due to lack of statistical power. The underlying pattern of view-time data can also be examined for supporting evidence.

Clearly, the predictions depend on some strong assumptions, so unless the effects of chunking produce substantial temporal signals, no effective measures of competence will be obtained.

Method

The experiment was conducted at the University of Sussex with approval from the Science School's ethics committee.

Design

The experiment is a within participant design with each person transcribing basic and advanced sections of Java program code. The order of these trials was counter-balanced. The trials were preceded with two practice stimuli.

(Originally, the experiment was a counter-balanced 2X2 design with a fixed stimuli factor to provide pause distribution measures for comparison. Unfortunately, an obscure software-hardware interaction on the experimental computer was found during analysis. As the original counterbalancing does not appear to have affected the reported conditions, for clarity, the experiment is presented just as single factor.)

Participants

The participants were 24 adults from the School of Engineering and Informatics. Recruitment spanned first year undergraduate students through to members of faculty, to obtain good range of programming expertise. Age ranged from 19 to 59 years (*mean*=25, *SD*=8.51), and 15 were male and 9 females. They received £8 for participating.

```
#
public class Person{
    public String name;
    public int age;}

public void Balance(){
    System.out.println("#");
    Total += balance;}

int h=0;
while(h<hCount.length){
  System.out.println(h+hCount[h]);h++;}
```

Figure 1: Stimulus sample (basic).

Materials

The two practice stimuli consisted of series of simple statements, such as 'Computer Science', 'Programming Course', 'JAVA Programming Language'. Each of the four Java program code stimuli consisted of nine lines of code divided into three separate blocks. Each stimulus had an equal number of lines and the total number of characters differed by less than 5%. Figure 1 shows an example of one stimulus. Two *basic* and two *advanced* versions of the stimuli were created by consulting the course content of the student participants. The expressions in the basic stimuli were a core part of their JAVA instruction in their first year. The expressions in the advanced stimuli are more specialist items that would only have been seen by the better performing students.

The experiment was conducted using a standard graphics tablet (Wacom – Intuous3) connected to a PC running a logging program specially written in our lab. Participants wrote with an inking pen on a response sheet. The response sheet was printed with a grid of 17 lines; each consisting of 42 spaces for the writing of separate characters. The sheet was designed for non-cursive writing in order to provide rich inter-stroke pause data (see parenthetical note in the Design section). Participants adjust to this style of writing quickly and it does not appear to adversely affected other aspects of their performance (Cheng, 2014; Zulkifli, 2013).

Following the trials, the participants completed a questionnaire with four parts (on an internet survey platform). Part 1 included biographic questions relating to educational level. Part 2 assessed programming experience in general with five graduated rating items, such as 'I can develop programs using more than one object-oriented programming language". Part 3 assessed Java programming expertise level using eight graduated items, such as 'I am familiar with both objects and classes in Java'. Part 4 measured the participants' familiarity with the four specific Java stimuli that they were presented with during the trial. Participants were asked to judge what

their degree of familiarity would have been for each item *prior* to the experiment, on a 5 point Likert Scale.

Procedure

Participants were asked to hold the pen in their preferred hand and trained to: start writing at the beginning of each line, even for indented code; start writing as soon as the stimulus is revealed; copy the code as quickly and as accurately as they can; continue writing without correcting if they made a mistake; draw an upside down triangle symbol (inverted capital delta) in place of spaces; to start each trial with a hash (#); to hold down the special key to reveal with stimulus, with their preferred hand, which ensures that they write only when the stimulus key is released. The participants easily complied with these requirements and quickly became fluent in the practice trials. (Several of these conditions were needed for the pause measurements.) Similar trial requirements were successfully used in our previous experiments, so it is clear that they do not, on their own, undermine the reliability of the results.

For each trial, the response sheet was taped to the tablet. The participants finished the experiment within an hour.

Table 1: Correlation between competence measures. (N=24, Pearson correlation, 1 tail, critical value is 0.472 at $p<.01$)

	Education level	General programming	Java	Familiarity
Education level	–	0.366	0.183	0.181
General programming		–	0.759	0.734
Java			–	0.849
Familiarity				–

Results

Independent measure of competence

Questionnaire responses were coded to obtain independent competence measures against which to compare the chunk-based measures. Education level was scored on a scale from one to six (1=1st year undergraduate student, 6=faculty member). General programming and Java experience were scored by giving one point for each positive answer related to the measure, so had scales from zero to five and zero to eight, respectively. Ratings of the familiarity were scored from 0 (low) to 4 (high), so with the four stimuli, the overall scale runs from zero to twelve. Table 1 presented correlations between all combination of the measures, and is unsurprising. Education level is only weakly (and not significantly) correlated to the other measures. General programming experience has a strong positive relation to both Java experience and familiarity. The correlation between Java experience and familiarity are particularly strong. All this suggests that both Java experience or familiarity are specific to Java, rather than wider programming competence, and that either is suitable to serve as an independent measure. As the actual pattern of results is equivalent with either measure, just the analyses using familiarity are reported here.

Behavioural measures

The dependent behaviour measures were computed from the logs of each participant. The median writing-times and view-times were calculated for each trial. View-numbers is a count of interface switches to the stimuli (button presses). (We also computed a view related measure that discounted views of a stimulus without any accompanying writing before the next view, as some participants occasionally made such repeated views. The pattern of results using this measure is essentially the same as that with view-numbers.)

Figures 2, 3 and 4 show the total view-numbers, median writing-times and median view-times for participants rank ordered by their familiarity scores. Figures 5, 6 and 7 aggregate the data across low and high competent participants by showing the mean of the total view-numbers, the mean of the median writing-times and the mean of the median view-times. A binary split of participants' familiarity scores conveniently creates two equal size groups, with low scores exclusively below 6 or and high score exclusively above 8.

The first thing to note is that the total view-numbers, Figure 5, for the practice items is considerably lower than for the Java stimuli, but that the value is essentially equal at low and high competency (6.6 and 5.7, respectively). Similarly, the mean of the median writing-times, Figure 6, for the practice items is substantially longer than the Java stimuli, and although the value is greater for higher than lower competence (means of 14.2 and 12.1 s), it is not significantly so (by a *t* test; t=1.09, df=22, 1 tail, p=.24). These results reassuringly suggest that an effect of transcribing the Java stimuli exists beyond the act of merely transcribing any stimuli.

Consistent with prediction H1, Figure 5 shows that the high competence participants required fewer views than those with low competence, which is significant at both levels of stimuli (basic: 16.3 vs. 25.2, t=4.40, p=.0002; advanced, 20.0 vs. 28.5; t=4.05, p=.0005; both df=22, 1 tail).

Consistent with prediction H4, the basic stimuli demand fewer views than the advance stimuli across all participants (20.8. vs. 24.3; t=4.05, p=.0003; df=22, 1 tail). Further, for high competence participants the view-numbers is still significant despite the small group size (19.2 vs. 22.2; t=2.88, p=.016; df=10, 1 tail).

Consistent with prediction H2, Figure 3 and 6 show that the high competence participants had longer writing-times than those with low competence, which is significant at both levels of stimuli (basic: 10.7 vs. 6.5 s, t=3.86, p=.0008; advanced, 8.0 vs. 5.7; t=3.14, p=.005; both df=22, 1 tail).

Consistent with prediction H4, the advanced stimuli had shorter writing-times than the basic stimuli across all participants (6.9. vs. 8.6 s; t=3.29, p=.002; df=22, 1 tail). Further, for high competence participants the writing-time is still significant despite the small group size (8.0 vs. 10.7 s; t=3.7, p=.003; df=10, 1 tail), but not for the low competence participants (5.7 vs. 6.5 s, t=1.8, p=.1, df=10, 1 tail).

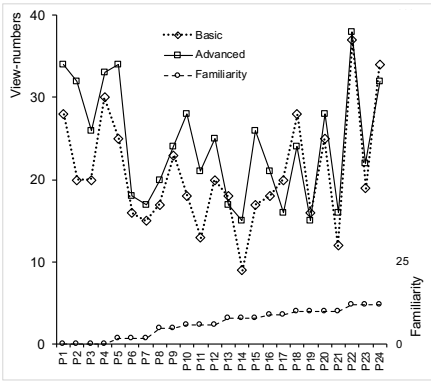


Figure 2. Total view-numbers for participants across basic and advance stimuli.

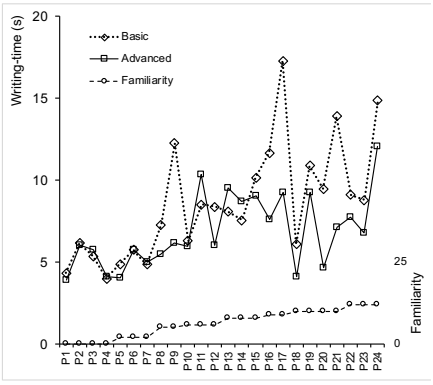


Figure 3. Median writing-times for participants across basic and advance stimuli.

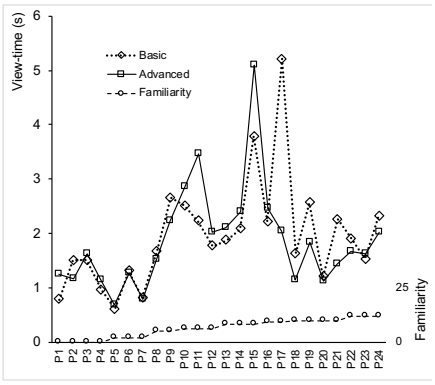


Figure 4. Median view-times for participants across basic and advance stimuli.

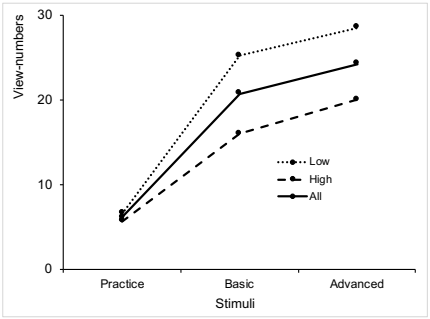


Figure 5: Mean view-numbers across stimuli type and level of competence.

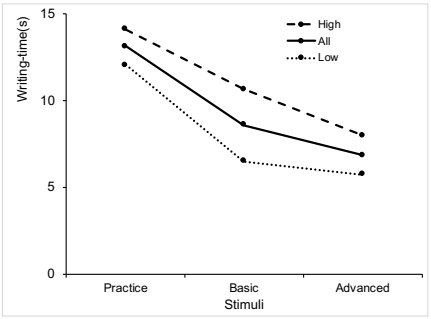


Figure 6. Mean of median writing-times across stimuli type and level of competence.

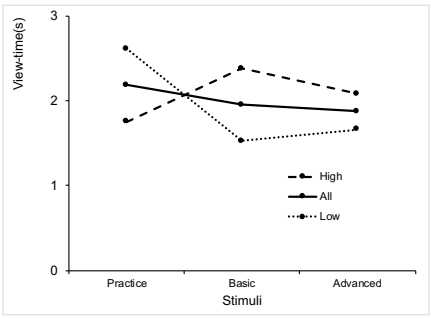


Figure 7. Mean of median view-times across stimuli type and level of competence.

Turning to H3, which concerns the absence of an overall effect of view-times, Figure 4 does not show a clear overall upward or downward trend in view-times, for both levels of stimuli difficulty. If anything, the overall pattern is an inverted 'u', in contrast to the trends in Figure 2 and 3. Figure 7 reveals that high competence participants have longer view-times than those with low competence, but this is not significant for the advanced stimuli (2.1 vs. 1.7 s; t=1.50, p=.15, df=22, 1 tail), but is marginally significant for the basic stimuli (2.4 vs. 1.5 s; t=2.62, p=.02, df=22, 1 tail). Further, comparing the view-times on the practice stimuli with the Java stimuli view-times we see they are similar, whereas for view-numbers and for writing-times the practice values are quite different to the Java stimuli values, as noted above.

Consistent with prediction H4, Figure 4 shows that nearly equal numbers of participants had longer view-times for basic stimuli or for advanced stimuli. In terms of the means across all participants, Figure 7, no significant differences occur for the basic stimuli (1.5 vs. 1.7, t=1.03, p=.3, df=22, 1 tail) nor the advanced stimuli (2.4 vs. 2.3; t=1.21, p=.25; df=22, 1 tail).

In summary, with respect to total view-numbers, means writing-times and view-times, all the predictions are supported.

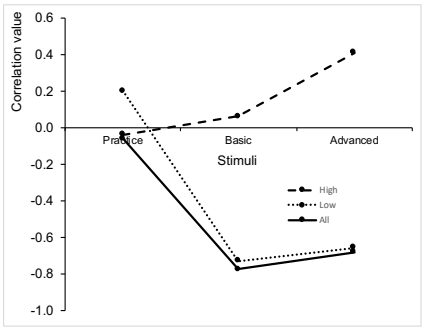


Figure 8: Correlation of view-numbers with familiarity across stimuli and competence.

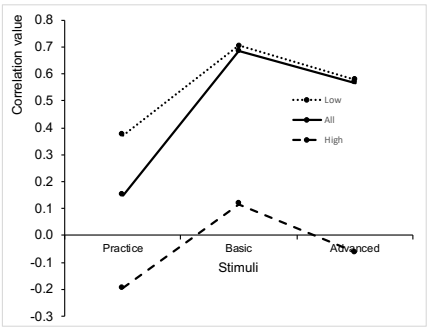


Figure 9: Correlation of writing-times with familiarity across stimuli and competence.

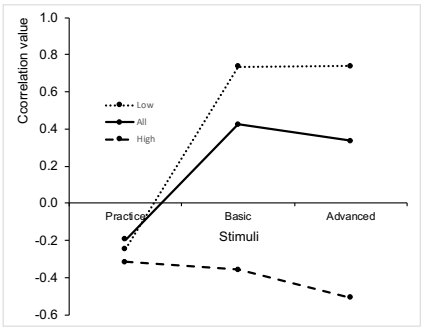


Figure 10: Correlation of view-times with familiarity across stimuli and competence.

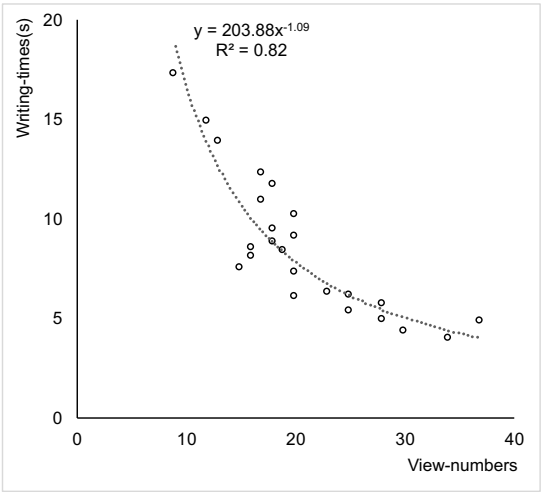


Figure 11: Relation of writing-times to view-numbers (basic stimulus)

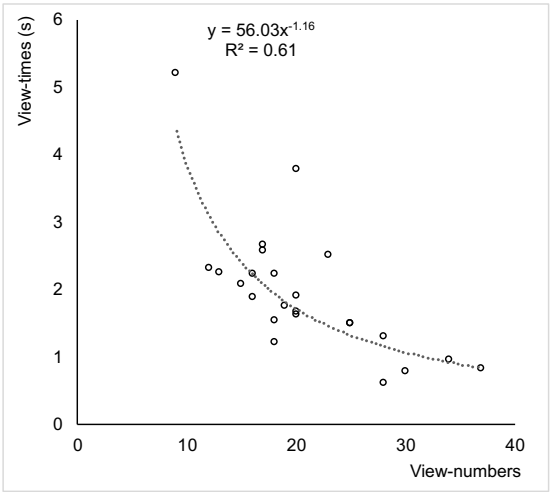


Figure 12: Relation of view-times to view-numbers (basic stimulus)

Correlation values for various measures

The correlations between view-numbers, writing-times and view-times versus familiarity were computed in order to further examine our four predictions. Figures 8, 9 and 10 show the Pearson correlations of familiarity score with, respectively, view-numbers, writing-times and view-times. The scale ranges are not the same. For correlations over all participants (solid line in Figs. 8-10) the critical value is 0.344 for significant correlations at $p<.05$, and 0.472 at $p<.01$ (1 tail, df=22). For correlations with just high competence or low competence participants (dashed or dotted lines) the critical value is 0.497 at $p<.05$ and 0.658 at $p<.01$ (1 tail, df=10).

As expected, none of the correlations for the practice items are significant. With view-numbers, Figure 8, across all participants the negative correlations are strong and significant: numbers of views decrease with competence (H1). The result is similar when just the low competence group is considered, but correlation for the high competence participants is positive but not significant. For writing-times the pattern of results is similar but the direction of the correlations are reversed, Figure 9: writing-time increases with competence (H2). For the whole group and the low competence subgroup the correlation for advanced stimuli is less than for the basic stimuli.

The view-times correlation, Figure 10, for the whole group and the high competent sub-group are not significant, but the correlations of the low competence participants are strong for both Java stimuli.

In summary, correlations for the view-numbers, writing-times and view-times are consistent with our four predictions, overall, but with some divergence in detail. In particular, view-numbers and writing-times did not differentiate high competence participants. Also, view-times did unexpectedly differentiate low competence participants, who needed more view time with increasing competence.

View-numbers vs. writing-times and view-times

The relation between our three main behavioural measures are examined because a systematic relation between them could provide further support for the hypotheses and more precise chunk-based explanations of the results. View number and writing-time are both predicted to be dependent upon chunking processes, so there should be some consistent and systematic relation between them. View-time is not expected to be chunk dependent, so no regular relation between it and view-numbers (or writing-duration) is anticipated. Scatter plots of these variables were drawn for all the participants in all the conditions of the experiment. Figure 11 plots writing-times versus view-numbers for the basic stimuli and Figure 12 is similar but for view-times. The pattern of data in Figure

Table 2. Parameter of best-fit power relation for writing-times and view-times to view-numbers

	Writing-times vs. view-numbers			View-times vs. view-numbers		
	Practice	Basic	Advanced	Practice	Basic	Advanced
Index, *i*	-0.95	-1.09	-0.97	-1.05	-1.16	-1.24
Constant, *C*	57.9	203.9	136.5	8.9	56	83.7
R-squared	0.459	0.818	0.747	0.603	0.615	0.623

11 has a particularly distinctive form, which is also apparent in the graph for the advanced stimuli. Thus, a power law curve for an inverse proportional relation was fitted to the data: the parameters of the best fit equations are given in Table 2, along with the R^2 values. The quality of fit for other equation forms (e.g., linear) were worse than a power law.

The power law for writing-time versus view-numbers is noteworthy, across both stimuli: the index is close to minus one and the R^2 values are large. This implies that the data is governed by a direct inverse proportional law. The relation between the view-times and view-numbers is less clear, with an absolute value of the index further from unity and lower R^2 values.

In other experiments, as yet unpublished, we have found similar patterns in view-numbers and writing-times data that closely fit an inverse proportional power law, so we are confident that the pattern is not accidental.

In summary, there appears to be a simple relation between the view-numbers and writing-times: a participant who takes twice the view-numbers of another will use half the time each time they write. But this simple relation does not hold for view-times.

Discussion

Previous studies have shown that measures of the distribution of inter-stroke pauses, captured in a simple transcription task, appear to reflect the different chunk structures of learners and hence may be used to assess the competence of the learners (Cheng, 2014; 2015, van Genuchten & Cheng, 2010; Zulkifli, 2013). This experiment extends those findings, in three ways.

First, allowing the user to reveal the stimulus at will, and hiding it during writing, allows two alternative temporal chunk measures to be captured: view-numbers — the total number of views of the stimulus in a trial; writing–times – the median duration of writing time between views. Predictions H1, H2, and H4 associated with these measures are well supported by converging evidence. The measures strongly correlated with our independent measures of competence. Further, no support for view-times as a suitable measure of competence was obtained, as predicted in H3, despite the relative strength of the effects for the other two measures.

Second, the experiment has shown that measures based on temporal chunk signals are applicable beyond mathematics (algebraic formula) and natural language, in a domain that happens to share some characteristics of both those domains.

Third, in contrast to the single line stimuli used in the previous experiments mentioned above, the present stimuli were larger (nine lines). The greater amount of data per trial means that single trials can yield strong usable correlations with competence, without the theoretical problems of deciding how to aggregate data from multiple trials or the practical problems associated with switching between multiple trials.

The overall correlations of view-numbers and writing-times with competence are strong, and this also holds for the low competence group. However, we must consider two qualifications. First, it is clear from Figures 2 and 3, that there is considerable variability between participants, such that some of the best low competence participant have better scores than many of the high competence participants, and vice versa. Clearly the development of a real educational test of programming competence must address the accuracy and sensitivity of the measures, perhaps by combining measures. Second, the curves in Figure 2 and 3 suggest that the view-numbers and writing-times may have plateaued for the high competent participants; in other words the difficulty of the advance stimuli may be insufficient to differentiate those within that group. This seems plausible, in hindsight, as the range of difficulty captured in the stimuli design was based on the undergraduate Java programming curriculum, but a proportion of the participants were drawn from more senior groups. This plateauing was also seen in previous studies (Cheng, 2014, 2015). One implication of this is the importance of designing stimuli with a sufficient range for the target test group.

The clear inverse proportional relation between writing-times versus view-numbers (Figure 11, Table 2) supports the chunk based explanations underpinning the predictions H1 and H2, and the poor fit of such a power law for view-times versus view-numbers is consistent with prediction H3. In particular, this implies that assumptions made for the predictions concerning the variability in participants working memory capacity and speed of writing are relatively small effects in comparison to chunk size variability with competence. In other words, the primary process in the transcription tasks appears to be the selection of chunks from the stimulus, with more competent participants retaining more characters – because they possess larger chunks – and this determines that time required for writing is in a direct proportion to the number of characters. Nevertheless, Figure 2 and 3 show much individual variability, so a useful line for future work is to investigate the possibility of separately measuring working memory capacity and writing speeds of participants in order to consider whether there is a need to devised methods to normalize for them.

Two observed effects might be spurious results, but they are sufficiently striking to deserve fuller investigation in further work. The first is the positive correlations of view-times with competence, specifically for low competence participants, is counter to prediction H3, Figure 10. The second is the increase in view-times with decreasing view-numbers, Figure 12: theoretically, there ought to be little relation between the two. One approach to study these effects is to probe the contents of participants' individual sets of chunks, which we are currently doing by extracting the locations of onset of views from the written logs in order to identify the precise content of participants' chunks. Our current hypothesis is that view-time variations may be due to differences in stimuli encoding strategies that fluctuate with content type.

This paper contributes a method for evaluating competence in a programming using a transcription task and measures with timescale of 10 s. This extends the range of techniques beyond the pause distribution measures of previous work (e.g., Cheng, 2014, 2015; Cheng & Rojas-Anaya, 2007).

Uses of the technique in education are readily imagined that exploit the relative simplicity, short trial times and the potential for fully automated scoring. Simply, such transcription tasks might be administered as a component of summative end-of-course evaluations or as standalone screening tests at the outset of a course. More interestingly with appropriately designed test items, the approach might be used as a form of formative assessment to provide tutors with information about individuals' growing understanding of targeted programming concepts. We are planning work on the development of the approach as a tool for use in computer-based tutoring systems.

Acknowledgments

Noorah's PhD study is supported by Imam Abdulrahman Bin Faisal University.

References

Adelson, B. (1981). Problem solving and the development of abstract categories in programming languages. *Memory & Cognition*, *9*(4), 422–433.

Adelson, B. (1984). When novices surpass experts: The difficulty of a task may increase with expertise. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(3), 483–495.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology*, *4*(1), 55–81.

Cheng, P. C-H. (2014). Copying equations to assess mathematical competence: An evaluation of pause measures using graphical protocol analysis. In *Proc. of the 36th Annual Conf. of the Cognitive Science Society*, 319–324.

Cheng, P. C-H. (2015). Analyzing chunk pauses to measure mathematical competence : Copying equations using 'centre-click' interaction . In *Proc. of the 37th Annual Conference of the Cognitive Science Society. Cognitive Science Society., Austin, TX*, 345–350.

Cheng, P. C-H., & Rojas-Anaya, H. (2007). Measuring Mathematical Formula Writing Competence: An Application of Graphical Protocol Analysis. In *Proc. of the Twenty Ninth Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society, 869-874

Cheng, P. C-H., & van Genuchten, E. (2018). Combinations of simple mechanisms explain diverse strategies in the free-hand writing of memorised sentences. *Cognitive Science, 42*, 1070–1109.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Science, 24*(1), 87-114.

Egan, D. E., & Schwartz, B. J. (1979). Chunking in recall of symbolic drawings. *Memory & Cognition*, *7*(2), 149–58.

Gobet, F., Lane, P. C. R., Croker, S., Cheng, P. C.-H., Jones, G., Oliver, I., & Pine, J. M. (2001). Chunking mechanisms in human learning. *Trends in Cognitive Science*, 5(6), 1236-1243.

McKeithen, K. B., Reitman, J. S., Rueter, H. H., & Hirtle, S. C. (1981). Knowledge organization and skill differences in computer programmers. *Cognitive Psychology*, *13*(3), 307–325.

Miller, G. A. (1956). The magical number seven plus or minus two: Some limits on our capacity for information processing. *Psychological Review, 63*, 81-97.

Obaidellah, U. H., & Cheng, P. C.-H. (2015). The role of chunking in drawing Rey complex figure. Perception and Motor Skills, 120(2), 535-555.

Pennington, N. (1987). Stimulus structures and mental representations in expert comprehension of computer programs. *Cognitive Psychology*, *19*(3), 295–341.

Shneiderman, B. (1976). Exploratory experiments in programmer behavior. *International Journal of Computer & Information Sciences*, *5*(2), 123–143.

Spelman Miller, K. & Sullivan, K. P. H. (2006). Keystroke logging: an Introduction. In G. Rijlaarsdam (Series Ed.) and K.P.H. Sullivan, & E. Lindgren. (Vol. Eds.), *Studies in Writing, Vol.18, Computer Keystroke Logging: Methods and Applications*. Oxford; Elsevier.

van Genuchten, E., & Cheng, P. C.-H. (2010). Temporal Chunk Signal Reflecting Five Hierarchical Levels in Writing Sentences. In *Proc. of the 32nd Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society,* 1922-1927.

van Genuchten, E., Cheng, P. C.-H., Leseman, P. P. M., & Messer, M. H. (2009). Missing working memory deficit in dyslexia: Children writing from memory. In *Proc. of the 31st Annual Conference of the Cognitive Science Society (Pp. 1674-1679). Austin, TX: Cognitive Science Society*.

Ye, N., & Salvendy, G. (1996). Expert-novice knowledge of computer programming at different levels of abstraction. *Ergonomics*, *39*(3), 461–481.

Zulkifli, M. (2013). *Applying Pause Analysis to Explore Cognitive Processes in the Copying of Sentences by Second Language Users*. University of Sussex (Unpublished PhD Thesis).

The Role of Information in Visual Word Recognition: A Perceptually-Constrained Connectionist Account

Raquel G. Alhama (rgalhama@mpi.nl)
Max Planck Institute for Psycholinguistics, Wundtlaan 1, 6525 XD Nijmegen, The Netherlands

Noam Siegelman (noam.siegelman@yale.edu)
Haskins Laboratories, 300 George street, New Haven, CT, USA

Ram Frost (ram.frost@mail.huji.ac.il)
Department of Psychology, The Hebrew University, Jerusalem, Israel

Blair C. Armstrong (blair.armstrong@utoronto.ca)
Department of Psychology, University of Toronto, 1265 Military Trail, Toronto, ON, Canada

Abstract

Proficient readers typically fixate near the center of a word, with a slight bias towards word onset. We explore a novel account of this phenomenon based on combining information-theory with perceptual constraints in a connectionist model of visual word recognition. This account posits that the amount of information-content available for word identification varies across fixation locations and across languages. These differences contribute to the overall fixation location bias in different languages, make the novel prediction that certain words are more readily identified when fixating at an atypical fixation location, and predict specific cross-linguistic differences. We tested these predictions across several simulations in English and Hebrew, and in a behavioral experiment. The results confirmed that the bias to fixate closer to word onset aligns with reducing uncertainty in the visual signal, that some words are more readily identified at atypical fixation locations, and that these effects vary across languages.

Keywords: visual word recognition; computational modelling; connectionism; information theory; fixation location

Introduction

The fundamental aim of visual word recognition is to identify a word based on its constituent letters. Considerable computational and behavioral evidence from studying isolated visual word recognition, which typically involves seeing a single word presented at the center of visual fixation, suggests that a graded constraint satisfaction process selects a candidate that fits with the lower-level (visual/orthographic) and higher level representations (e.g., lexical information, McClelland & Rumelhart, 1981). In contrast to this methodology, in more naturalistic studies of reading via eye-tracking, considerable evidence suggests that readers tend to fixate more frequently near the middle of words, typically with a bias towards beginning of a word, with some variation across languages (see Figure 1, for example distributions from initial fixations during natural reading in English and Hebrew Siegelman et al., 2019).

A key question from considering this body of work, then, is how and why the visual system of a proficient reader tends to fixate at particular positions in a word, and on a related front, why these fixation distributions vary as a function of the language. Classic accounts focused on the low-level operations of the occulomotor system do not appear to offer a ready explanation of these effects, particularly in terms of cross-linguistic differences (see McConkie, Kerr, Reddix, Zola, & Jacobs, 1989 for a review of occulomotor theories; also Reichle, Rayner, & Pollatsek, 1999; Engbert, Nuthmann, Richter, & Kliegl, 2005). Accounts that hold more promise in this regard consider higher-level factors (e.g., morphology; Deutsch & Rayner, 1999).

Here, we explore an alternative more general account based on information theory in the visual signal and how it maps onto lexical representations. This work shares some conceptual similarity with prior work by Brysbaert and Nazir (2005), although the latter did not quantify information in the formal terms that we do, which may, as outlined in the discussion, explain some discrepancies between their results and ours. In our first study we examined the differences in information distributions as a function of fixation location in Hebrew and English, and found that these distributions shared key characteristics of the human fixation location distributions. In our second study, we instantiated a feed-forward connectionist model with a psychophysically-derived constraint on letter identification as a function of distance (eccentricity) from the target fixation. This model allowed us to examine how different amounts of information content can be extracted at different fixation locations in different languages during word recognition. If it succeeded in doing so, it could explain why there is a preferred fixation location in different languages due simply to how low-level constraints interact when identifying a word, in the absence of higher-level constraints (e.g., morphology, semantics). This model also served as a test-bed for probing whether words exist in different languages that, due strictly to the information content available at different fixation locations, are, perhaps counter-intuitively, more efficiently recognized by looking at fixation locations other than the overall preferred location in the language. These predictions were corroborated in a pilot behavioral experiment. Taken together, this research highlights how maximizing information in the visual signal could be a major driver of many behaviors observed within and between languages. It also offers specific predictions for broadening this account in future work, for instance, in maximizing information across words

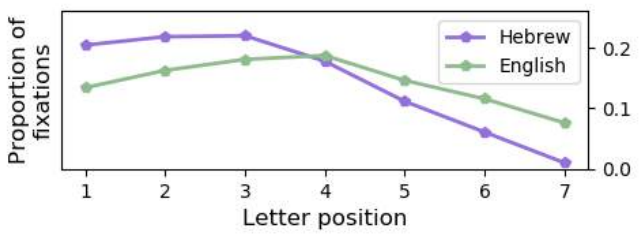


Figure 1: Distribution of fixation locations for 7-letter words in English and Hebrew. 1=start of word (left in English, right in Hebrew)

rather than within the processing of single words.[1]

Study 1: Information-content in the early visual-orthographic representation.

In our first study, we explored how much information-content was available for word recognition in the early visual-orthographic signal when fixating at different locations in the word. This was achieved computationally by passing the visual representation of a word at a particular fixation location through a visual filter that reflects how more visual information is extracted from the fixated location in a word and less information is extracted as a function of eccentricity (distance) from this location. We applied this procedure to samples of words from English and Hebrew, which belong to different language families, to gain insight into the language-specific versus language-general nature of the results.

Data We analyzed the 50,000 highest frequency words from the OpenSubtitles translated movie subtitle database (Tiedemann, 2012)[2]. We removed all words that contained foreign alphabet characters. For simplicity, we selected for our study only 7-letter words, because we predicted that strong effects of fixation location and information content would be more readily detected in longer words that could nevertheless be perceived with a single fixation. The resulting lists contains 5565 words in Hebrew, and 8145 in English.[3]

Architecture To simulate the constraints on visual perception imposed by the early visual perception system, we passed the representation of each word in each language through perceptual filters adapted from McConkie et al. (1989). In the original formulation of this model of perceptual filtering, the fixated letter was perceived with 100% accuracy, and the likelihood of successful perception fell off linearly as a function of eccentricity (see Figure 2, for examples from fixating letter 2 or letter 6 in a 7-letter word). The exact slope of this function, as exemplified by the $drop = 0.1$ and $drop = 0.25$ lines in the figure, leads to an initial linear change in the amount of extracted information, which eventually reaches floor.

In the original paper, the authors noted that the optimal value of the *drop* parameter remained to be determined. Thus, for this initial work, we opted to use a *drop* parameter of 0.25. This value was selected so as to capture most but not all the letters in a word when perceived from the start or end of the word, which we predicted would lead to relatively high, but below ceiling, recognition rates (confirmed and described in a later section) and substantial differences in information as a function of fixation location.

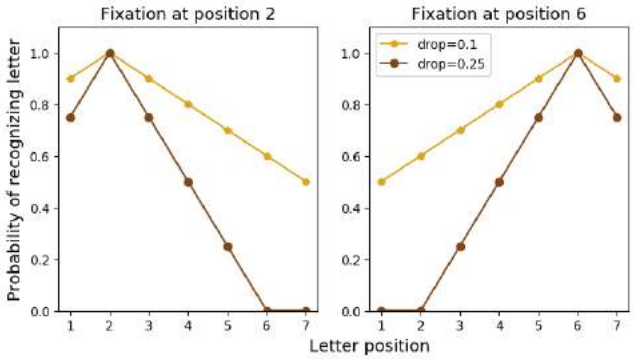


Figure 2: Probability of recognizing the constituent letters in a word when fixating letter position 2 (left) and letter position 6 (right) in a word according to the McConkie model.

Procedure We tested for how the fixation location could impact the information content extracted from the perceived word. For simplicity, and to test for strong modulations in word recognition due to the perceived information, here we focused on the information content extracted from a fixation near the beginning of a word (at the second letter position) and near the end of a word (at the sixth letter position). After passing each word through the McConkie filter, for each word, we calculated the remaining amount of uncertainty on the identity of the word *after* fixating at each of these fixation locations (a proxy of the information content in each location[4]). The measure of uncertainty we used was *entropy*, as proposed in Shannon (1948). Concretely, given the letters retrieved after fixating on a word, we computed the remaining entropy as $H = -\sum_{w=1}^{m} p_w log_2(p_w)$, where the words w belonged to the set of words m that have a perfect match with the identified letters, both in letter identity and letter position (e.g. the word 'therapy' would be in the set of matching words for the recognized letters 'ther - - -'). The probability of a matching word p_w was estimated as its relative frequency

[1]The code for our models and analyses is released at https://github.com/rgalhama/nnfixrec_cogsci2019.

[2]From https://github.com/hermitdave.

[3]We ruled out the possibility that vocabulary size drove any of our simulated behavioral effects by down-sampling the English corpus to be the same size as the Hebrew corpus in our simulations.

[4]Note that, throughout our paper, we use the term "information content" of a fixation location to quantify the contribution of the observed letters in minimizing the uncertainty *on the identity of the word*. This should not be confused with the *surprisal* conveyed by the letters in a fixation location. The former concerns a probabilistic models for words (based on word frequency and component letters), while the latter would be based on a probabilistic model of letter strings.

in the corpora.[5] To test whether information was distributed evenly across the beginning and end of all words in each language, we subtracted the entropy for each word at fixation location 2 from that at fixation location 6. If entropy were evenly distributed, these values should cluster around 0.

Additionally, from the distribution of entropy difference scores, we identified 100 words with the most extreme positive (50 words) and negative (50 words) values.[6] We refer to the words with more information content when fixating at position 2 (i.e. negative entropy differences) as the $maxIC(2)$ words, and the words with more information content at position 6 (i.e,. positive entropy differences) as the $maxIC(6)$ words. These words served as test items in Study 2.

Results

The difference in entropy values when a word was perceived at fixation location 2 versus fixation location 6 are plotted in Figure 3. It clearly shows that entropy information is not uniformly distributed across words, as in both English and Hebrew there are more negative scores. It also shows a relatively wide range of entropy values, with both languages containing words with entropy difference scores ranging from approximately -5 to 3. Further, the Hebrew scores tend to be more negative than the English scores.

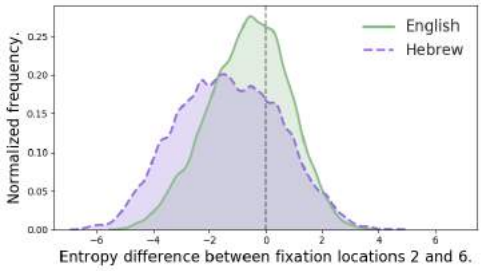


Figure 3: Distribution of entropy differences for 7-letter words.

To give a more concrete intuition into the magnitude of the difference scores and their relationship to successful visual word recognition, consider the case of the word "zooming", which has an entropy difference score of -5.1. If the fixated letter and the two letters on either side of this letter are preceived correctly, there is a 100% likelihood of successful recognition of this word when fixated at position 2 (i.e. when perceiving 'zoo-'). However, there is only a 3% success rate when fixated at position 6 (i.e. when perceiving '-ing').

Next, we selected 100 words per language with "extreme" entropy difference scores for use in Study 2. These items had mean difference scores, for $maxIC(2)$, of -2.68 in English and -3.32 in Hebrew, and for $maxIC(6)$, it was 2.19 in English and 2.58 in Hebrew.

In additional simulations, not reported in detail due to space constraints, we also confirmed that varying the exact shape of the McConkie function and the value of the *drop* parameter did not qualitatively alter these trends unless only the nearest items to the fixation location, or nearly all the words in the word, were perceived with 100% accuracy.

Discussion

The first simulation substantiated our predictions that different amounts of information content can be extracted by fixating at different locations in a word. Overall, there appears to be more information content present at the start of words in both languages, providing initial evidence for a language-general trend. Thus, the fixation distributions in different languages may at least be partially attributed to a system that attempts to minimize entropy in the visual signal in service of word recognition. This claim is further bolstered by the fact that the Hebrew distribution was even more shifted to contain more information when fixating at the beginning of a word, consistent with the stronger preference to fixate earlier in Hebrew words in behavioral data (see Figure 1). The broad distribution of values in each language also enabled us to select items with "extreme" entropy difference scores across fixation locations. This enabled us to test whether some words are more readily identified by fixating at a location other than the overall preferred fixation of the language (which is off-center, nearer to the beginning of the word).

Having thus established that the perceptual input to the word recognition system contains major differences in entropy based on fixation location, we next explored how these inputs could shape processing in a connectionist model of word recognition.

Study 2: A perceptually-constrained connectionist model of visual word recognition

The previous study focused on the distribution of information contained in the languages. In this study, we employed a connectionist model and a coordinated pilot behavioral experiment to investigate whether a learning model of word recognition is sensitive to these information patterns. This allowed us to develop new predictions about how performance in different fixation locations evolves in relation to reading proficiency: although it is beyond our goals to align model training (in epochs) with human reading experience —which is a non-trivial question–, our learning model provided us with insights into novel emergent processing dynamics that are not visible from an information-theoretic approach. In particular, we focused on whether the model and the human participants displayed an interaction between fixation location and the location of maximum information content in our "extreme" items selected in Study 1. Because some of the implementational decisions for the model were made to increase

[5] An alternative approach is to compute these values over word types rather than word tokens. Control simulations showed that both of these approaches were highly correlated in both languages, $r > .72$, and that the correlations between entropy differences over types or tokens and word frequency were extremely small, $|r| < .04$

[6] We filtered some items to avoid the over-representation of letter combinations like "-ing" and to eliminate extremely low and high frequency items.

the similarity between the simulated task and the pilot behavioral task, we provide a brief overview of the behavioral task and findings before turning to the details of the simulation (for the complete report of this experiment, see Siegelman et al., 2019).

Overview of the behavioral task. A total of 23 native speakers of Hebrew (14 females, age range: 22-30, mean: 24.9) were presented with the a set of words including the 100 words with extreme information differences described previously, and an additional selection of 100 words (also of length 7)[7] with intermediate entropy differences, which we treat as fillers for the purpose of this paper.[8] In the task, participants first focused on a fixation cross for 1000 ms, and then were presented with a word for 100 ms. This brief presentation prevented multiple fixations and was expected to lower performance below ceiling. Critically, the word was presented in different locations, including cases wherein the letter at position 2 or position 6 appeared at the same location as the fixation cross (i.e., the word was left- or right-shifted from screen center where participants were fixating). Participants were then instructed to say or guess the word that had been presented. Responses were coded either as correct or incorrect. All words were presented in both fixation locations simulated in the model (i.e., fixation at position 2 vs. position 6). Words were presented in a random order.

The results are presented in Figure 4. The pattern of results indicated that accuracy was highest when a word was fixated at the location which contained the most information content, and lower at the location which contained less information content. Critically, this was true not only for words that had more information content early in the word, but also for words that contained more information content near the end of the word: a logistic mixed-effect model (with condition, fixation location, and their interaction as fixed effects, trial number and log-transformed frequency as control variables, by-subject and by-item random intercepts, and by-subject random slope for condition) revealed a significant interaction (B = 0.59, SE = 0.05, $p < 0.001$ and a significant main effect of fixation location (B = -0.23, SE = 0.05, $p < 0.001$). Thus, these findings do not simply reflect a preference to process words in the more frequently fixated location in a given language, which our prior study showed contains, on average, more information content. The presence of numerically larger differences across fixation locations for words

[7] All the words we use are multisyllabic. This could create confounds in the modelling work if we mapped the input with phonological representations, but we intentionally focused only on orthographic factors, since our goal is to find out what structure exists in the orthographic signal alone in the absence of phonological considerations. Future work may investigate how these visual representations interact with phonological representations, similar to a classic "triangle" model (Seidenberg & McClelland, 1989).

[8] Although we did not test these items in the model, the behavioral results indicated that items with intermediate entropy difference scores were relatively unaffected by whether they were fixated near onset or offset, as predicted by the account.

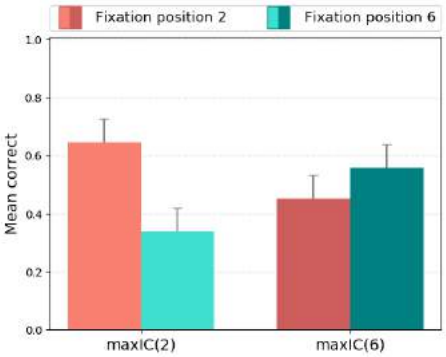


Figure 4: Mean correct responses of participants in the four conditions. Error bars = SEM.

with more information content near the start of a word relative to near the end of a word may suggest a more subtle interaction between information content and frequency of exposure to different locations, however. Additionally, averaging across the four experimental conditions, overall accuracy was significantly below ceiling.

Simulating the behavioral experiment

Model Architecture We implemented a feed-forward connectionist model that mapped perceptually-constrained distributed input of a word's constituent letters onto a localist representation of each word in the training vocabulary, as illustrated in Figure 5. There were 7 letter input slots, one for each position in a 7-letter word. Each of these slots had one unit for every letter in the alphabet and coded for the presence (1) or absence (0) of a given letter in that position (i.e., a binary one-hot coding). The distributed representation of the visual word was then input to a McConkie filter set to perceive the word at a particular fixation location (the procedure for specifying fixation locations is described later). Thus, the one-hot vectors would be down-scaled (using a drop parameter, d, of 0.25) to reduce the activity of having perceived a given letter as a function of eccentricity from the fixation location, as quantified in Equation 1:

$$x(i) = x(i) * max(0, 1 - eccentricity(i) * d) \quad (1)$$

Thus, the activity of the fixated letter remained unchanged, the activity of letters more than four slots distant from the fixated letter was set to 0, and activity in each letter-slot would decrease linearly between these two bounds.

To simulate the noisy nature of perceptual inputs in the human visual system, we next injected normally-distributed random noise ($\mu = 0.2$, $\sigma^2 = 0.05$) into the unit activations (clipping activations to $[0, 1]$). We assumed that the activity after these processing steps was analogous to what would be available in an early visual-orthographic representation ("perceived input" in the Figure).

Next, we mapped the perceived input onto a one-hot log-softmax target output representation for each word in the vocabulary through a pool of 125 hidden units. The output of

the hidden units was determined first by computing the sigmoidal function of their net input, followed by the injection of uniform random (output) noise (mean = 0, range = 0.05). All the weights in the network were randomly initialized from a uniform distribution in the range $[\frac{-1}{\sqrt{fan\_out}}, \frac{1}{\sqrt{fan\_out}}]$, where fan_out was the number of units in the subsequent layer of the model.

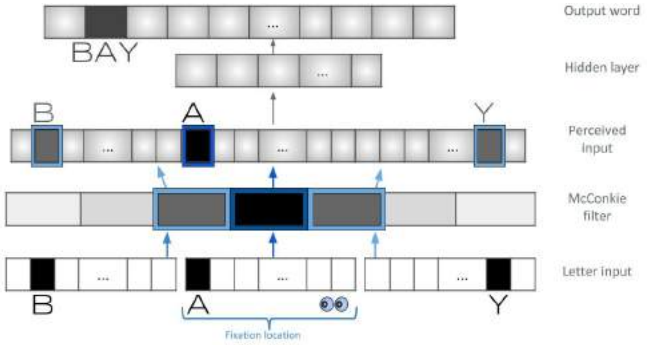


Figure 5: Model architecture. The implemented model perceived 7-letter words; here, we illustrate the model processing a 3-letter word (BAY) for simplicity. Full (1) and no (0) activity are shaded in black and white; intermediate values are shaded in grey.

Determining fixated locations in a word We evaluated several methods for selecting how often a word was perceived from a different fixation location, which we term "fixation location distribution schemes". One model sampled each fixation location equally (hereafter, the *uniform* fixation model). This provided an estimate of the impact of the entropy at different fixation locations that was unconfounded with how often humans typically fixate at each location, and how those distributions varied across the two languages under study. Another model employed the language-specific behaviorally-derived fixation distributions illustrated in Figure 1 (hereafter, the *behavioral* fixation model). A third model averaged these two fixation distribution schemes (hereafter, the *50/50* model). This "blended" model allowed us to interpolate between these two previously described schemes and simulated a case where a model was sensitive to frequency of exposure, but not necessarily to the raw values. The logic here was that a good model might standardize frequency information to ensure low-frequency information is also learned.

Training The model was trained by presenting a 7-letter word at a particular fixation location and computing the cross-entropy error between the output and the target representation. Error was accumulated in batches in which every seven-letter word in the target language was presented 20 times, with the likelihood of fixating at a particular location determined by the fixation distribution sampling scheme. Error was then backpropagated to adjust the weights between the perceived input and the output layer (learning rate = .005; weight decay = .0001) using stochastic gradient descent for the first 10 epochs, and the Adam algorithm thereafter (Kingma & Ba, 2014). The model was trained for 200 epochs (runs through each batch).

All models reached a stable high level of overall word recognition accuracy (near 80%) for approximately the last 50 epochs of training. The vast majority of the incorrect responses originate from words perceived at a suboptimal—and where applicable, less frequent—fixation location. Figure 6 provides representative data for the Hebrew words with extreme entropy values using the *uniform* model. (space constraints prevented the inclusion of plots from the other models, which were broadly similar). The presence of different effects during early training than at the end of training also makes novel predictions for future developmental studies.

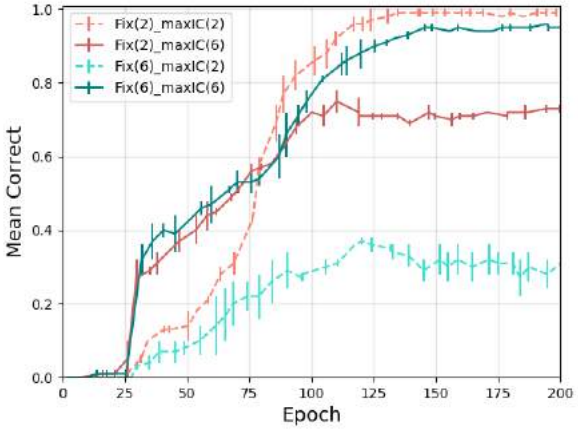


Figure 6: Accuracy for the *uniform* model trained on Hebrew for the words with extreme entropy difference scores. Error bars = SEM.

Testing We froze the weights on the trained models before testing them in a manner analogous to the behavioral experiment. In the test, we presented all the *maxIC*(2) and *maxIC*(6) words at both fixation location 2 and fixation location 6. We also tested several methods of bringing performance in the task below ceiling as in the behavioral experiment, including dimming model inputs (multiplying all input letter activations by a value less than 1), and increasing variance of the noise applied to the perceptual input (cf. Lambon Ralph, Lowe, & Rogers, 2007). These methods yielded similar overall results, so here we report only the results of dimming (dimming parameter = .35). We ran this simulation twice on models initialized with different random weights and report the average results.

Results

The results for the *uniform*, 50/50, and *behavioral* fixation models of English and Hebrew are presented in Figure 7. First, in contrast to the non-dimmed model at the end of training (see Figure 6), our testing procedure clearly succeeded in

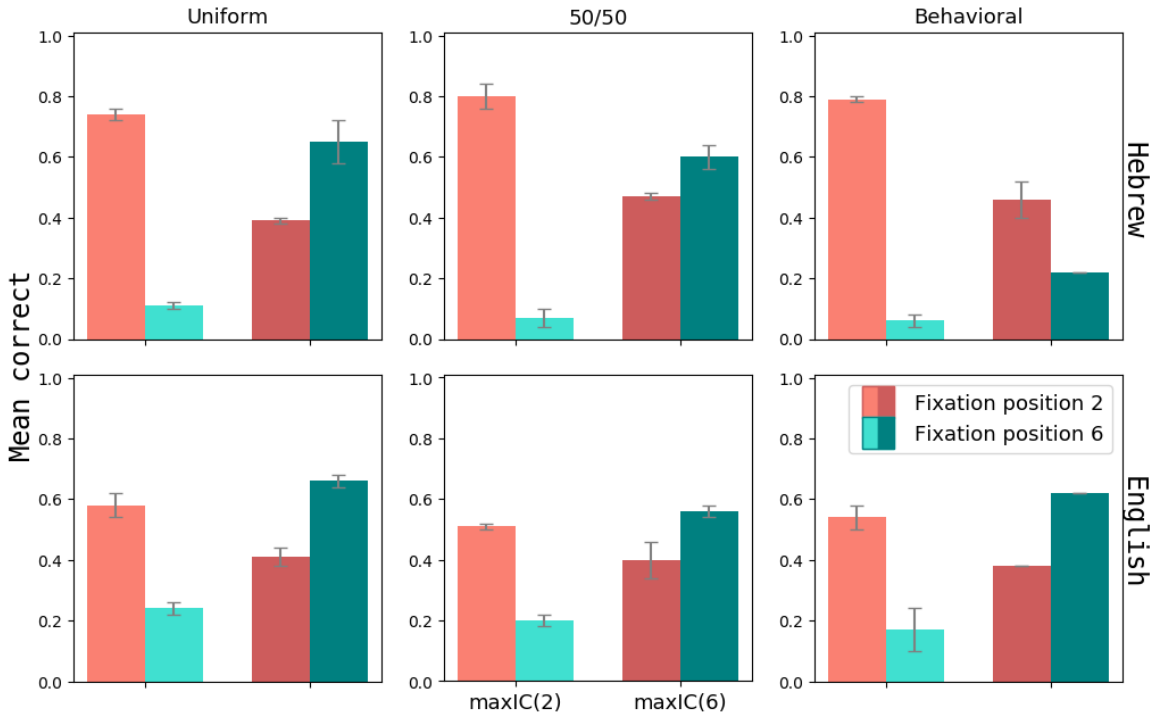


Figure 7: Results of testing the different fixation location distribution models in Hebrew [top] and English [bottom]. Error bars = SEM.

lowering overall accuracy to a level similar to that in the behavioral experiment.

The most critical finding was that, with the exception of the Hebrew *behavioral* fixation model, every one of these models produced a qualitatively similar interaction between fixation location and whether more information content was located at the beginning or end of a word, as in the behavioral results. The Hebrew simulations also show more pronounced effects for the $maxIC(2)$ items overall, as in the behavioral data. The reduced effects in English make for novel predictions for an experiment in that language.

Moving from the $uniform$ model through the $50/50$ model to the *behavioral* model, the effects of the behavioral fixation sampling scheme, which fixates words at position 2 more of then than position 6, is more apparent in Hebrew than in English. This is reflected by the fact that $max(IC2)$ words are perceived more accurately when fixating at position 2 in Hebrew and less accurately when fixating at position 6 when moving toward the behavioral fixation scheme.

The exceptional Hebrew *behavioral* fixation model appears to be an exaggerated extension of the effects of fixation location frequency outlined above. In the case of this model, even the $maxIC(6)$ words were responded to more accurately when fixating earlier in the word. The presence of this pattern only in Hebrew is at least partially explained by the more extreme differences in fixation location sampling distributions in Hebrew than in English. These results also suggest that the human visual recognition system may at least partially normalize the effects of fixation location frequency, given that the $50/50$ and $uniform$ fixation models produced qualitative results more similar to those in the behavioral experiment.

Discussion

The results of the second set of simulations largely paralleled those of the behavioral experiment, with both exhibiting an interaction between fixation location and the location with most information content. The simulations also showed the influence of the behavioral fixation location distributions in enhancing the perception of words at the most frequent fixation location, and suggest that the word recognition system normalizes the fixation location distribution to some degree. Further, although the qualitative findings were similar across languages, suggesting that a general principle is at play, at a quantitative level there were some differences between the two target languages. These differences align with the relatively higher information content at the beginning of Hebrew words and the greater likelihood of fixating at the beginning of words. Collectively, this work therefore indicates that the word recognition system is sensitive to the information content in different locations in a word, as constrained by the perceptual system.

These results are only in partial agreement with past work (Brysbaert & Nazir, 2005). In that work, participants were presented with partial word information for 5-letter words and asked to "guess" the word. The distribution of "guesses" relative to the correct response was then taken as their measure of uncertainty. Their results showed similar effects as in our study at word onset, but no effects at word offset. These

findings were interpreted as suggesting that the effects of information content were only present at the preferred fixation location. A combination of factors likely explain the discrepancy between their claims and ours, including a more adequate formal quantification of information content and the use of longer words that may be more sensitive to perceptually-constrained information content effects.

The success of this work at the individual word level also points to important directions for future work. One major question raised by this work is how these principles could generalize to the multi-word level. Can a preceding word provide top-down context and reduce the uncertainty (i.e. the average information content) on the set of upcoming words so as to not only facilitate processing, but also alter the location of an upcoming fixation? If so, this result could help explain the relatively broad fixation distributions obtained in different languages, because the optimal location to fixate in a word may deviate from the average location from the language as a function of context. The somewhat broad overall fixation location distributions may therefore in actuality reflect the averages of narrower fixation location distributions that are conditioned by the preceding word.

Our work shows that the observed behavioral effects in word recognition can be explained based on low-level information structures in the visual signal, without the need to resort to higher-level morphological structures. Higher-level structures can enter the visual-orthographic system in two ways: first, in shaping the word forms of a language, and second, as representations that mediate word recognition. The former is subsumed in our information-theoretic approach, which encompasses all the constraints that provided word forms with their actual shape, providing us with a quantitative comparative framework for a crosslinguistic perspective. The latter cannot be completely ruled out: although our model does not require morphological representations to succeed, the contribution of these representations should be assessed with more targeted experiments that aim to tease apart the visual/orthographic from morphological (e.g. looking at performance for regular and irregular morphemes such as *brothel/broth, corner/corn, farmer/farm*, Rastle, Davis, & New, 2004).

To sum up, this work offers a language-general and parsimonious account of how a specific type of statistical information drives performance in the perceptually-constrained word recognition system, complementing accounts based on the operation of the occulomotor system, as well as complementing or or subsuming accounts based on higher-level information. In so doing, this work reinforces the importance of studying how the structure of language itself interacts with the perceptual constraints of the visual/orthographic system (Lerner, Armstrong, & Frost, 2014) in shaping reading behaviors, and opens new avenues for combining isolated word and naturalistic reading research.

Acknowledgments

This project was funded by the ERC (ERC-2015-AdG-692502 to RF) and NSERC (DG-502584 to BCA).

References

Brysbaert, M., & Nazir, T. (2005). Visual constraints in written word recognition: evidence from the optimal viewing-position effect. *Journal of Research in Reading*, *28*(3), 216–228.

Deutsch, A., & Rayner, K. (1999). Initial fixation location effects in reading hebrew words. *Language and Cognitive Processes*, *14*(4), 393–421.

Engbert, R., Nuthmann, A., Richter, E. M., & Kliegl, R. (2005). SWIFT: a dynamical model of saccade generation during reading. *Psychological review*, *112*(4), 777.

Kingma, D. P., & Ba, J. L. (2014). Adam: A method for stochastic optimization. In *Proc. 3rd Int. Conf. Learning Representations.*

Lambon Ralph, M. A., Lowe, C., & Rogers, T. T. (2007, 04). Neural basis of category-specific semantic deficits for living things: evidence from semantic dementia, HSVE and a neural network model. *Brain*, *130*(4), 1127-1137.

Lerner, I., Armstrong, B. C., & Frost, R. (2014). What can we learn from learning models about sensitivity to letter-order in visual word recognition? *Journal of Memory and Language*, *77*, 40–58.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. an account of basic findings. *Psychological Review*, *88*(5), 375.

McConkie, G. W., Kerr, P. W., Reddix, M. D., Zola, D., & Jacobs, A. M. (1989). Eye movement control during reading: Ii. frequency of refixating a word. *Perception & Psychophysics*, *46*(3), 245–253.

Rastle, K., Davis, M. H., & New, B. (2004). The broth in my brother's brothel: Morpho-orthographic segmentation in visual word recognition. *Psychonomic Bulletin & Review*, *11*(6), 1090–1098.

Reichle, E. D., Rayner, K., & Pollatsek, A. (1999). Eye movement control in reading: Accounting for initial fixation locations and refixations within the EZ reader model. *Vision research*, *39*(26), 4403–4411.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological review*, *96*(4), 523.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379–423.

Siegelman, N., Alhama, R. G., Bogaerts, L., Armstrong, B., Kuperman, V., & Frost, R. (2019). *Reading across writing systems: an information-theoretic perspective, in prep.*

Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In N. Calzolari et al. (Eds.), *Proc. 8th International Conf. on Language Resources and Evaluation.* Istanbul, Turkey: European Language Resources Association (ELRA).

Rapid Trial-and-Error Learning in Physical Problem Solving

Kelsey Allen
Massachusetts Institute of Technology, Cambridge, Massachusetts, United States

Kevin Smith
Massachusetts Institute of Technology, Cambridge, Massachusetts, United States

Josh Tenenbaum
MIT, Cambridge, Massachusetts, United States

Abstract

We introduce a new problem solving paradigm: solving physical puzzles by placing tool-like objects in a scene. The puzzles are designed to explicitly evoke different physical concepts such as support, blocking, tipping, and launching, and are typically solved in a handful of trials. We study human participants' problem solving strategies, including what they try first, how they update their actions based on failed attempts, and how many attempts they eventually take to solve the puzzles. We introduce the 'Sample, Simulate, Remember' model that incorporates object-based priors to generate hypotheses, mental simulation to test hypotheses, and a memory and generalization system to update across simulations and real-world trials, and show that all three components are needed to explain human performance. Further results can be found at https://k-r-allen.github.io/tool-games/

Self-Organized Division of Cognitive Labor

Edgar Andrade-Lotero (edgar.andrade@urosario.edu.co)
Department of Applied Mathematics and Computer Science, Carrera 6 No. 12C-06, Of. 506
Bogotá, COLOMBIA

Robert L. Goldstone (rgoldsto@indiana.edu)
Department of Psychological and Brain Sciences, Psychology Building 338
Bloomington, IN 53706 USA

Abstract

The division of labor phenomenon has been observed with respect to both manual and cognitive labor, but there is no clear understanding of the intra- and inter-individual mechanisms that allow for its emergence, especially when there are multiple divisions possible and communication is limited. Situations fitting this description include individuals in a group splitting a geographical region for resource harvesting without explicit negotiation, or a couple tacitly negotiating the hour of the day for each to shower so that there is sufficient hot water. We studied this phenomenon by means of an iterative two-person game where multiple divisions are possible, but no explicit communication is allowed. Our results suggest that there are a limited number of biases toward divisions of labor, which serve as attractors in the dynamics of dyadic coordination. However, unlike Schelling's focal points, these biases do not attract players' attention at the onset of the interaction, but are only revealed and consolidated by the in-game dynamics of dyadic interaction.

Keywords: Group cognition; Divergent behavioral norms; Focal points; Cooperation.

Introduction

An individual can often benefit from participating in a group when (s)he can perform just one component of the group's task while other individuals take care of other parts. When the other individuals also benefit from this arrangement, we speak of an efficient division of labor. For example, two roommates can choose between (a) preparing their lunch for themselves every day, and (b) dividing the days of the week on which one prepares lunch for two. In the latter case, both roommates benefit from not having to cook every day.

The benefits of division of labor have been studied not only with respect to manual labor (Smith, 2008), but also with respect to cognitive labor (Sloman & Fernbach, 2017; Kennedy, Eberhart, & Shi, 2001). For instance, one study showed that the puzzle of assigning categories to the nodes of a network such that no adjacent nodes have the same category could be efficiently solved as a self-organized, collective task if each individual is assigned to a single node and is only concerned about the acceptability of their local sub-network (Kearns, Suri, & Montfort, 2006).

In some collective groups, such as ant colonies or beehives, the division of labor occurs as a genetically designed organization (Weitekamp, Libbrecht, & Keller, 2017; Robinson, 1992). However, it can also emerge as a self-organized process, without leaders or explicit negotiations (Heylighen, 2013). For example, when a group of individuals has to collectively guess a target number, where the collective guess is the sum of their individual guesses, and the only feedback they receive is for how much their collective guess is greater or lesser than the target, individuals spontaneously differentiate their behaviors to either react or not react to the feedback, and the extent to which role differentiation occurs is predictive of group performance (Roberts & Goldstone, 2011).

What are the cognitive mechanisms that facilitate the self-organized division of labor? One possibility is that it arises from the principle of maximization of expected utility. In our previous example of the two roommates, successful division of the days of the week might be said to arise because it constitutes a Nash equilibrium, that is, a combination of choices in which no roommate can obtain a higher payoff by changing only their choice—fixing the other roommate's choice (Ross, 2018). However, as it turns out, maximization of expected utility is not sufficient to explain why roommates act in accord with a particular Nash equilibrium instead of another (Arthur, 1994; Colman, 2003). Some scholars have suggested that games with multiple Nash equilibria are not solved on the basis of maximization of expected utility, but rather by means of rough-and-ready rules of thumb based on limited knowledge and time. This approach is known as 'bounded rationality' to emphasize that people frequently have memory, attentional, and calculation limitations that prevent them from employing perfectly rational strategies (Holbrook, 2002; Simon, 1957). It could be claimed, returning to our roommates example, that the division of labor according to which Roommate A prepares lunch only on weekdays and Roommate B only prepares lunch on the weekend is achieved because they cannot think of a different division, or because this division is the most natural for both of them, even though there are many other possible divisions. This is an example of the focal point approach, according to which the set of all possible Nash equilibria is reduced to just a single point that is psychologically salient for all players (Mehta, Starmer, & Sugden, 1994; Schelling, 1960). Another possible proposal is that individuals possess a small set of simple strategies that they can apply in their search for a division. For example, they may stick to one strategy for as long as it provides acceptable results, and when it fails, they would swap it for another in their strategy set (a win-stay, lose-shift heuristic). There are cooperative scenarios, such as the famous El Farol problem

(Arthur, 1994), in which this heuristic works well. Another strategy could be to adapt one's own reactivity to the task based on how much the whole group is contributing. Indeed, as pointed out by Roberts and Goldstone (2011), there may be situations in which an individual helps the collective effort by refraining from acting or reducing their activity, allowing the other players to dominate the task.

We studied this phenomenon by means of an iterative two-person game where multiple divisions are possible, but no explicit communication is allowed. Our results suggest that there are a limited number of biases toward divisions of labor and that they work as attractors in the negotiation dynamics. Unlike Schelling's focal points, these biases do not attract players' attention at the onset of the interaction, but are only revealed and consolidated by the in-game dynamics of the dyadic interaction. In other words, these biases do not determine players' *a priori* actions, via some sort of iterative reasoning, for dividing up their task. Rather, the attractors only become salient as a result of the interaction.

Materials and methods

Participants and procedure

Participants were 90 undergraduate students at Indiana University in Bloomington who received course credits for approximately 1 hour of participation. Participants were run in 10 experimental sessions, each one requiring an even number of participants to be grouped into dyads. If an odd number of participants turned up to the session, one of them was randomly chosen and sent home. The number of dyads in each session were as follows: 4, 5, 3, 6, 4, 2, 6, 3, 8, and 4. Participants sat in a university computer lab, each at a sound- and sight-isolated personal computer running a version of the game implemented in the nodeGame platform (Balietti, 2017). The computer randomly paired participants into dyads and each dyad participated in 60 rounds of the game. Participants were instructed not to talk to each other and were not informed about who was paired with whom.

The task

The task is a two-player game, which we dub "Seeking the unicorn," in which players interact with 64 tiles arranged in an 8×8 grid (see the top panel in Figure 1). The grid can either hide a unicorn beneath one of the tiles or else the unicorn can be absent from the grid, either event can occur with equal probability. At the beginning of each round, the computer chooses with equal probability whether or not there is a unicorn, and if there is one, it randomly chooses a tile in which to hide it, each tile having an equal probability of being chosen. Then, players seek for the unicorn by uncovering tiles one at a time, with both players uncovering tiles simultaneously, in order to see what lies beneath them. What tiles have been uncovered and whether there is or not a unicorn is only known to the player that uncovers these tiles. Tiles uncovered by both players instantly change their color and both players can see this. At any time during the round, each player can guess whether the unicorn is present or absent. The other player will know this player's decision and can use it to inform their own guess. The round ends when both players announce that their guess is a final decision, and then they are shown their scores (see the bottom panel in Figure 1). The score depends on whether the player's guess is correct (32 points) or incorrect (-64 points), subtracting the number of tiles that were uncovered by both players.

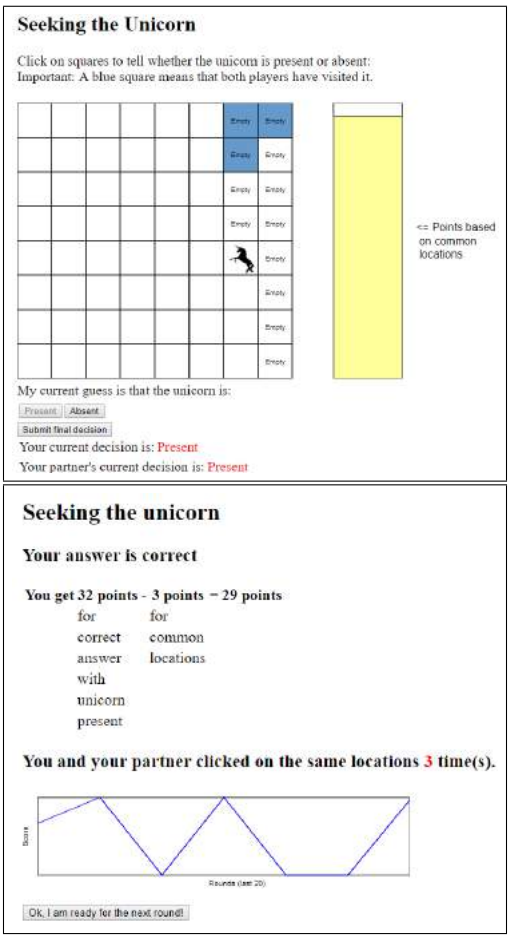


Figure 1: The experimental task. The top panel shows the grid as displayed to each player. By uncovering a tile, they know whether it is empty or contains the unicorn. Such information is private for the player. Tiles uncovered by both players have a blue background and both can see this coloration. They also have access to each other's guesses. The yellow column on the right decreases as the number of overlapping tiles increases. The round ends when both players submit their final decision. In the bottom panel we show the screen displaying the score and the score history over the last 20 rounds.

Measures

The following measure, which we call the Division of Labor Index (DLINDEX), determines the extent to which

players split the grid into complementary regions:

$$\text{DLINDEX} = \frac{\text{Tiles uncovered by one or both of the players} - \text{Overlapping tiles}}{\text{Tiles in the grid}}$$

This measure instantiates the intuition that it is beneficial if a dyad collectively uncovers all of the tiles (first term) and does not overlap in any tiles uncovered (second term). Observe that it ranges from 0 to 1 with 1 being ideal division of labor and 0 being least efficient. There is only one way of being ideal, namely, when both players uncover the entire grid and do not overlap at all. Additionally, we measure how consistently a player uncovers tiles from one round to the next:

$$\text{Consistency}_n = \frac{\text{Overlapping uncovered tiles from Round } n-1 \text{ to Round } n}{\text{Tiles uncovered in either of the two rounds}}$$

This number ranges from 0 to 1 with 1 meaning that the player uncovers the same tiles on both rounds, and 0 meaning that the player uncovers a completely different set of tiles from one round to the next. We also define the distance and the similarity between two regions a and b in the grid in the following way:

$$\text{dist}(a,b) = \sqrt{\sum_{t \in \text{Tiles}} (a_t - b_t)^2}, \quad \text{sim}(a,b,\varepsilon) = e^{-\varepsilon * \text{dist}(a,b)}$$

Here, t represents the t-th tile in the 8×8 grid (represented as the list $[1,\ldots,64]$), and a_t and b_t can be either 1 or 0, representing whether or not tile t belongs to a and b, respectively. The parameter ε in the definition of sim determines the extent to which the distance between two regions determines the similarity between them and, unless explicitly stated otherwise, we assume that ε=1.

Results

We should note at the outset that we have not used the entire dataset in our analysis. The reason is that rounds on which the unicorn is present provide us only with partial information as to how players split the grid, because on those rounds players do not have to uncover every tile. Once they find the unicorn, they will say that the unicorn is present and finish the round. Therefore, unless explicitly stated otherwise, we are only reporting results for trials on which the unicorn is absent.

For each dyad we created a figure displaying two grids, one for each player. In this figure we magnitude-coded each tile according to the number of times the player selected it through 60 rounds of the experiment in such a way that the darker the tile, the more times it was selected. Figure 2 shows the types of regions that were obtained. There were only four stable, successful pairs of complementary regions in the grid: the Left-Right, Top-Bottom, All-Nothing, and Inside-Outside splits. We call them the focal splits. Only dyads in the focal splits obtained an above-average DLINDEX, except for one dyad with no discernible stable region that nevertheless has

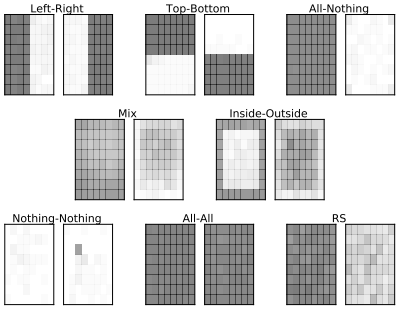


Figure 2: The seven types of splits of the grid that could be observed from our data. Each panel shows two grids, one for each player, with the regions uncovered through 60 rounds. The darker the tile, the more rounds the player uncovered it.

an average DLINDEX of 0.82 (this is over 0.45 standard deviations above the average of the 45 dyads; this dyad determined the Mix type of split in Figure 2). We conclude that 26 out of 45 dyads successfully split the grid. This represents over 57% success in self-organizing division of labor.

If our paradigm were a one-shot task in which players had to converge on a split of the grid on only one round, our data show that the average DLINDEX would be close to 0.43 (s.d.$\approx$ 0.32). By comparison, in our iterated task, the average DLINDEX rose up to almost 0.68 after 60 rounds (s.d.$\approx$ 0.32). The difference between these averages is statistically significant ($p< 0.001$). This shows that an efficient division of labor does not emerge on the first round, and that the iterated nature of our task facilitates its emergence.

But how did the division of labor emerge? We observed that, in general, dyads moved from lower to higher levels of DLINDEX, and that players in a low-level dyad tended to more frequently change their tile selection strategy from one round to the next with respect to players in a high-level dyad. Moreover, we found a positive correlation between a player's consistency on Round n and their score on Round $n-1$ ($\beta \approx 0.51$; $p < 0.001$). This supports the hypothesis that players used, at least to some extent, a win-stay, lose-shift heuristic (WSLS). That is to say, if their score is relatively high, which often occurs when the dyad splits the grid into complementary regions, each player tends to re-select their previously selected tiles; but if their score is low, they will be more likely to shift to different tiles.

However, WSLS does not seem to account for all the characteristics of the dyadic interaction. When we predict DLINDEX as a function of consistency, we see that, perhaps not surprisingly, dyads consisting of individuals who are relatively consistent in their tile selection strategies tend to divide labor better ($\beta \approx 0.36$; $p < 0.001$). However, we also observe an interaction such that dyads with players that differ in their consistencies tend to divide labor better than predicted when players have a large amount of overlap in their selected tiles. That is, if both players overlap considerably, it is best if one player is consistent and the other player is not. The evidence

for this claim comes from comparing the linear regression model above with a model that includes the interaction between, on the one hand, the absolute difference in consistency between players on a given round and, on the other hand, the number of overlapping tiles on the previous round:

$$\text{DLINDEX}(n) \sim \alpha + \beta_1 * \text{Consistency}(n) + \beta_2 * \text{difConsist}(n) + \beta_3 * \text{Overlap}(n-1) + \beta_4 * \text{difConsist}(n) * \text{Overlap}(n-1)$$

Our data show that this interaction is positive ($\beta_4 \approx 0.01$; $p < 0.001$). Moreover, an analysis of variance test ($p < 0.001$) confirms that this interaction effect accounts for significantly more variance in performance relative to the main effects. These results indicate that dyads eventually tend to most effectively divide labor despite initially overlapping in their tiles when one player is consistent/stubborn and the other player is inconsistent/flexible, giving rise to complementary degrees of reactivity to occasions of overlap (Roberts & Goldstone, 2011). But why did one of the players become more stubborn? We found that if a player tends to select tiles consistent with a focal region (that is, one half of a focal split), they tend to be more consistent. In other words, the closer a player's tile selection strategy is to a focal region, the more stubborn they become, presumably because they believe that they are forming one half of a viable division of labor. The regression model of consistency with respect to distance to closest focal region confirms this effect ($\beta \approx -0.12$; $p < 0.001$). The interesting question now is how the other player figured out that they have to select tiles in the appropriate complementary region, given that a player only had access to their own uncovered tiles and not the other player's uncovered tiles. The key seems to lie in the fact that players do have access to overlapping tiles, from which the other player's selected tiles can be inferred with reasonably high validity.

One mechanism that can account for many players' shifts in selected tiles is based on a measure of the similarity between a focal region and the overlapping tiles. If one player's selected tiles are sufficiently close to a focal region, then this can be used as a signal for the other player to select the corresponding, complementary region. In Figure 3 we take a closer look at an actual game play from a dyad in which this mechanism is prominent, as exhibited by Player B's transition. On Round 23 the overlapping tiles are similar to the focal region RIGHT, which inclines B to select every single tile in the complementary LEFT region. Observe that B not only re-selected the left region's tiles from the previous round, but uncovered the whole LEFT region. More generally, the player's attention is attracted toward a focal region k when the region that is complementary to k is sufficiently similar to the overlapping tiles. To be sure, even though the process seems to be gradual and there are other factors at play, these complementary focal regions have attraction power. Last but not least, observe that the overlapping tiles are the same for both players, so Player A's attention is also attracted by LEFT. Nevertheless, given that A has uncovered the focal region RIGHT, they tend to become "stubborn" in the sense of resisting substantial change to their uncovered tiles. The combination of this retention and the attraction powers of a focal region informs a decision process that we call the Focal Regions as Attractors heuristic (FRA).

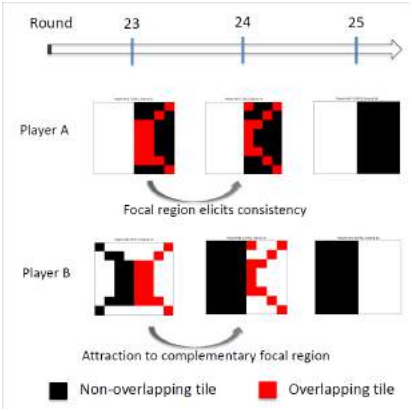


Figure 3: Evidence in favor of Focal Regions as Attractors (FRA). We see the transition from one round to the next, taken from an actual game play. In each grid, black tiles represent uncovered tiles and red tiles were uncovered by both players. Player A's transition illustrates 'stubbornness' and Player B's illustrates the attraction exerted by the complement of A's focal region, which is also a focal region. See details in the text.

Computational models

We put our previous explanations to the test by providing a computational model for each one of these two heuristics. The first model is an implementation of WSLS. To motivate it, suppose that on round n the player uncovered tiles determining BOTTOM. We want to determine the probability of choosing each region k in $\mathcal{K}$ on round $n+1$, where k in $\mathcal{K}$={RS, ALL, NOTHING, BOTTOM, TOP, LEFT, RIGHT, INSIDE, OUTSIDE}. $\mathcal{K}$ contains the focal regions, plus the type of region we call RS, which represents all remaining regions in the grid. Now, if the player is in a win situation, we should increase the probability of choosing again BOTTOM. This effect can be obtained by means of a threshold function (see Figure 4). More formally, the model defines a probability function, determined by the following formula:

$$P(k) = \frac{\text{attract}(k)}{\sum_{r \in \mathcal{K}} \text{attract}(r)} \quad (1)$$

The attract(k) function represents the extent to which a player is inclined to choose region k, given the current state of the game. For the WSLS model, we assume that this state is represented by the vector (i,s), where i is the region explored on the previous round and s the obtained score. The attract function for the WSLS model is defined in the following way:

$$\text{attract}(k,i,s) = \text{bias}_k + \alpha * \text{thresh}(s_n, \beta, \gamma) * \text{I}(k,i) \quad (2)$$

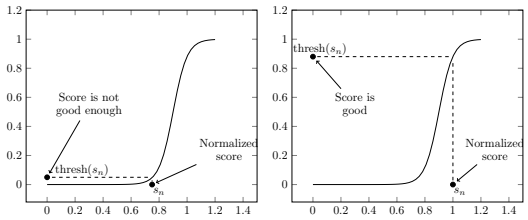


Figure 4: The thresh(s_n) function representing how good the score was. The left panel illustrates a situation where the score is not good enough, which is captured by the low value of thresh(s_n). The right panel illustrates a good situation, captured by the high value of thresh(s_n).

Here, the term bias$_k$ represents how inclined the player feels toward k, all other things being equal, and is expected to be higher for more pre-experimentally salient regions. The second term contains the functions thresh and I, which are defined in the following way:

$$\text{thresh}(s_n,\beta,\gamma)=\frac{1}{1+e^{-\beta(s_n-\gamma)}},\quad \text{I}(k,i)=\begin{cases}1, & \text{if } i=k\neq \text{RS}\\ 0, & \text{otherwise}\end{cases}$$

Here, s_n is the normalized score, which takes values between 0 and 1. The function thresh(s_n,β,γ) has an S shape and takes values in the open interval $(0,1)$. It goes from values near 0 to values near 1 when s_n is near γ, and the steepness of this transition is determined by β (see Figure 4). The second term in Equation 2 contains the parameter α, which determines the extent to which the score increases the player's tendency to choose k, when the normalized score is greater than γ. The effect of I(k,i) in this expression is that the only region that has its bias modified is region i (i.e., the region explored on the previous round) and only if this region is a focal region. The value of attract(k) for the remaining regions is equal to bias$_k$.

The model defined by FRA extends the previous model. To motivate it, suppose that on round n the player uncovered tiles in the i region as defined in Figure 5. Now, we should consider the overlapping region, j, and consider its similarity to each focal regions (see right panel of Figure 5). The more similar to k, the more attractive *the complement* of k becomes. In our example, the overlapping region is more similar to UP, so the probability of choosing BOTTOM on round $n+1$ is increased. More formally, we assume that the current state of the game is represented by the vector (i,s,j), where i is the region explored on the previous round and s the obtained score, and j the area formed by the overlapping tiles. The attractiveness of k is defined in the following way:

$$\begin{aligned}\text{attract}(k,i,j,s)=&\ \text{bias}_k+\alpha*\text{thresh}(s_n,\beta,\gamma)*\text{I}(k,i)\\&+\delta*\text{sim}(j,\bar{k},\varepsilon)*\text{Focal}(k)+\zeta*\text{I}(k,i)\end{aligned}\qquad(3)$$

Observe that the first two terms in Equation 3 are the same as in Equation 2. The third and fourth terms are new. In the third term, the function sim$(j,\bar{k},\varepsilon)$ determines the similarity

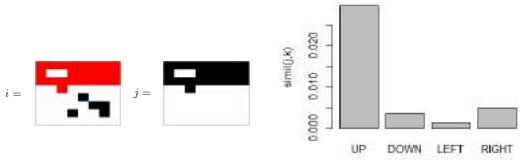


Figure 5: An example of a region visited, i, the overlapping tiles, j, and the similarity between j and other regions.

Model	θ_1	θ_2	θ_3	θ_4	α	β
WSLS	0.14	0.0674	0.0123	0.0009	39	405
FRA	0.077	0.048	≈ 0	≈ 0	48	402

Model	γ	δ	ε	ζ	Dev.	AIC
WSLS	0.933	0	0	0	3060	3074
FRA	0.99	1.57	0.94	3	2709	2709

Table 1: Best parameters and deviance for each model. The first four parameters correspond to the biases in the model: θ_1=bias$_{\text{ALL}}$, θ_2=bias$_{\text{NOTHING}}$, θ_3=bias$_{\text{BOTTOM}}$=bias$_{\text{TOP}}$=bias$_{\text{LEFT}}$=bias$_{\text{RIGHT}}$, and θ_4=bias$_{\text{IN}}$=bias$_{\text{OUT}}$. Moreover, bias$_{\text{RS}}$ is defined as 1 minus the sum of the other biases, and we require that the sum of all biases adds to 1.

between j and the complement of k, denoted as $\bar{k}$. The function Focal(k) is defined in the following way:

$$\text{Focal}(k)=\begin{cases}1, & \text{if } k\notin\{\text{RS, ALL}\}\\ 0, & \text{otherwise}\end{cases}$$

The parameter δ in Equation 3 determines the extent to which the similarity between j and $\bar{k}$ modifies attract(k), but this only occurs when k is a focal region and is different from ALL. This effect is obtained by multiplying δ by Focal(k). Finally, the parameter ζ determines the extent of the player's stubbornness when i is a focal region.

Note that the extra parameters from FRA with respect to WSLS are δ, ε, and ζ, and that Equation 2 for WSLS can be obtained from Equation 3 when $\delta=\zeta=0$. That is, WSLS is a nested model within FRA.

Using maximization of log likelihood of the multinomial distribution of the observed transition frequencies and the respective predicted probabilities given by the model, we found the optimal parameters and the *deviance* of the two models, summarized in Table 1. Both the Likelihood Ratio Test ($\chi^2=351$; 3 d.o.f.; p<0.001) and the ΔAIC=365 provide quantitative evidence that the additional parameters contributed by FRA provide a better account of the underlying choice process and that this model's better fit to the data is not due to overfitting.

We simulated our game in the same conditions as the experimental task. For each model, we ran 100 simulations of 60 rounds of the game, obtaining two collections of simulated data. In the two top panels of Figure 6 we can observe

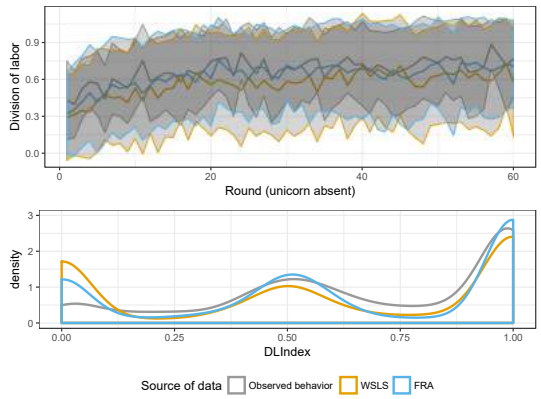


Figure 6: Comparison between observed and simulated behavior. The top panel shows the behavior of DLINDEX through all rounds. The vertical axis represents, for each round, the average DLINDEX. Shadow regions represent an error margin of one standard deviation. The bottom panel shows the kernel density estimate of DLINDEX for observed and simulated data, where each observation is the DLINDEX of a dyad at the end of a round.

the behavior of DLINDEX through all rounds. The vertical axis represents, for each round, the average DLINDEX with respect to all dyads in the respective set (45 for humans; 100 for each model). In the three cases we observe that the indexes are within the error margins of the others, and that the three sets of data show the same positive trend through rounds. However, in the case of WSLS, the t-test of mean difference ($p \approx 0.01$) does not provide conclusive evidence to assert that the means are the same, whereas the t-test of mean difference of DLINDEX ($p \approx 0.3$) determines that there is no statistical evidence to claim that the means are different, which means that FRA is better at capturing the tendency of DLINDEX in human subjects.

In the bottom panel of Fig. 6 we can see the kernel density estimate of DLINDEX for observed and simulated data. When the density curve is high (y-axis) for a given value of DLINDEX (x-axis), it means that there were many rounds for which a dyad obtained a DLINDEX close to x. Observe that, for humans, high values of DLINDEX are more frequent than medium and low values—representing the fact that many dyads split the grid satisfactorily. However, in WSLS there is a considerable tail on the left, indicating many more trials on which dyads did not split the grid into complementary regions, as compared to humans. Moreover, in WSLS the frequency of low values is higher than that of medium values, which is not in accordance with the observed data. For FRA, the frequency of low values is not greater than that of medium values, which is closer to what is observed in human data. To sum up, it seems that WSLS predicts a less efficient division of labor than exhibited by people, whereas FRA and people show a comparable degree of division of labor.

Discussion

57% of human dyads finished 60 rounds of game play with an efficient division of labor. The results from our experiment and our computational models allow us to explain how most dyads managed to split the grid without being able to engage in explicit negotiations. First of all, even though there are 2^{64} ways to split the grid, dyads split it in only four different ways. In some sense, these splits are focal points because they have a certain psychological salience (Schelling, 1960). One might have thought that these individual cognitive biases (focal points) would exert an early (in terms of rounds) influence on choices exactly because they are *a priori*, so that agents would have started on Round 1 with strategies of selecting all tiles on the left, top, bottom, or right. If agents understand that these are natural attractors, then through engaging in many levels of iterated thinking based on common knowledge (Lewis, 1969), these would be logical starting points. However, players do not generally start with strategies that resemble focal points. Humans are far more idiosyncratic and exploratory in their initial selections of tiles. It is only through repeated interactions that players manifest their *a priori* predispositions/biases toward certain focal points. In other words, a *priori biases* do not entail that the biases are manifest at the onset of play. It is only through dyadic interaction that these biases are revealed (Kaush ML, Griffiths TL, & Lewandowsky S, 2007). Returning to our two roommates example, there are 128 different ways to divide the days of the week in order to alternate one roommate cooking for two. We suppose that not every possible division is equally salient for them, and that only a handful of divisions will actually attract and retain their attention, such as the division between weekdays or weekend, or a division based on the idea of cooking every other day. If the roommates cannot explicitly negotiate a division but are given the daily chore of preparing lunch(es), one roommate will eventually follow one of these psychologically salient divisions and will tend to persist in the strategy because it is a focal point. To the extent that the other roommate wants to avoid overlapping days, soon they will be attracted to the psychologically salient strategy of choosing complementary days of the week. Interacting individuals, both human and algorithmic, can often arrive at efficient coordinating solutions in a paradigm that incorporates two challenging conditions – individuals cannot explicitly communicate, and there are multiple coordinating solutions that are initially equally salient. The human and computational results indicate that agents solve this coordination task by beginning with a set of possibly incompatible focal points. Then, via iterated interactions they adjust their behaviors to move toward focal points when they are not at a focal point, stay in a focal point once reached, and shift to a complementary focal point relative to the other player. In this way, the coordination that a group forms results from the interplay over time between their *a priori* cognitive biases and the dynamics of their interpersonal interaction (Hawkins, Goodman, & Goldstone, in press).

Acknowledgements

EAL was partially supported by the following research awards from Universidad del Rosario: "Becas para Estancias de Docencia e Investigación, 2018 and "Fondo de Capital Semilla, 2018". EAL is also grateful with Luis Andrade for useful discussions. Both authors are grateful with Evan Nix for running the experimental sessions, and with the anonymous referees at CogSci2019 for useful comments.

References

Arthur, W. B. (1994). Inductive Reasoning and Bounded Rationality. *The American Economic Review*, *84*(2), 406–411.

Balietti, S. (2017). nodeGame: Real-time, synchronous, online experiments in the browser. *Behavior Research Methods*, *49*(5), 1696–1715.

Colman, A. M. (2003). Cooperation, psychological game theory, and limitations of rationality in social interaction. *Behavioral and Brain Sciences*, *26*(2), 139–153.

Hawkins, R., Goodman, N., & Goldstone, R. (in press). The emergence of social norms and conventions. *Trends in Cognitve Science*.

Heylighen, F. (2013). Self-organization in Communicating Groups: The Emergence of Coordination, Shared References and Collective Intelligence. In *Complexity Perspectives on Language, Communication and Society* (pp. 117–149). Berlin, Heidelberg: Springer.

Holbrook, M. B. (2002). Bounded rationality: The adaptive toolbox. *Psychology & Marketing*, *20*(1), 87–92.

Kaush ML, Griffiths TL, & Lewandowsky S. (2007). Iterated learning: intergenerational knowledge transmission reveals inductive biases. *Psychonomic Bulletin & Review*, *14*(2), 288–288–94.

Kearns, M., Suri, S., & Montfort, N. (2006). An Experimental Study of the Coloring Problem on Human Subject Networks. *Science*, *313*(5788), 824–824–827.

Kennedy, J. F., Eberhart, R. C., & Shi, Y. (2001). *Swarm Intelligence*. San Francisco: Morgan Kaufmann.

Lewis, D. (1969). *Convention: A philosophical study*. Wiley-Blackwell.

Mehta, J., Starmer, C., & Sugden, R. (1994). The Nature of Salience: An Experimental Investigation of Pure Coordination Games. *The American Economic Review*, *84*(3), 658–673.

Roberts, M. E., & Goldstone, R. L. (2011). Adaptive Group Coordination and Role Differentiation. *PLOS ONE*, *6*(7), e22377.

Robinson, G. E. (1992). Regulation of Division of Labor in Insect Societies. *Annual Review of Entomology*, *37*(1), 637–665.

Ross, D. (2018). Game Theory. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2018 ed.). Metaphysics Research Lab, Stanford University.

Schelling, T. (1960). *The Strategy of Conflict* (1st ed.). Cambridge: Harvard University Press.

Simon, H. A. (1957). *Models of man: social and rational; mathematical essays on rational human behavior in society setting*. New York: Wiley.

Sloman, S., & Fernbach, P. (2017). *The Knowledge Illusion: Why we never think alone*. Riverhead.

Smith, A. (2008). *An inquiry into the nature and causes of the wealth of nations Adam Smith*.

Weitekamp, C., Libbrecht, R., & Keller, L. (2017). Genetics and Evolution of Social Behavior in Insects. *Annual Review Of Genetics*, *51*, 219–219–239.

A friend or a toy? Four-year-olds strategically demonstrate their competence to a puppet but only when others treat it as an agent

Mika Asaba[1*] (masaba@stanford.edu), Xiaoqian Li[2*] (xiaoqian_li@mymail.sutd.edu.sg), W. Quin Yow[2] (quin@sutd.edu.sg), and Hyowon Gweon[1] (hyo@stanford.edu)

[1]Department of Psychology, Stanford University, California, USA
[2]Humanities, Arts, and Social Sciences, Singapore University of Technology and Design, Singapore

Abstract

Others' beliefs about the self can powerfully influence our everyday interactions with others. Recent work suggests that even preschool-aged children are sensitive to what others think of the self and actively attempt to manage these beliefs (Asaba & Gweon, 2018). What cognitive capacities underlie these early self-presentational behaviors, and in what contexts do these behaviors emerge? Here we show that preschoolers strategically demonstrate their competence to even a *puppet*, but only when an adult treats the puppet as an agent and specifically asks which toy the child wants to "show" to the puppet (Exp.1). However, they do not show such strategic demonstration of their competence when the same puppet is treated as an object (Exp.2). These results suggest that self-presentational behaviors can emerge even in the absence of any immediate prospect of social evaluation insofar as children consider the target entity as capable of holding beliefs. Furthermore, whether or not children ascribe a belief about the self to the target is heavily modulated by how an entity is treated by others. We discuss the relevance of these findings to early reputation management behaviors, and more broadly, the use of make-believe agents in developmental research.

Keywords: cognitive development; social cognition; Theory of Mind; reputation management; agency

Introduction

What others think of us – our competence, kindness, fairness – is central in our minds. Others' beliefs about the self have the power to influence our social interactions, well-being, and even long-term life outcomes. Fortunately, we have some control over how others think of us: We can change our behaviors in the presence of others (e.g., act more generously; Novak & Sigmund, 2005) or actively disclose information about the self (Hicks, Liu, & Heyman, 2015). Knowing how to manage others' beliefs about us, or our reputation more broadly, can help us better navigate the complex social world and build healthy relationships with others. Despite the ubiquity of self-presentational behaviors, however, the ontogenetic origins of the ability to represent and modulate others' beliefs about the self remain poorly understood. What cognitive capacities underlie self-presentational behaviors, and in what contexts do these behaviors manifest?

Recent developmental work has provided some initial insights into young children's sensitivity to others' evaluations of them. Young children attempt to promote a positive reputation by sharing more and cheating less in the presence of others (e.g., Engelmann, Hermann, & Tomasello, 2012; Leimgruber, Shaw, Santos, & Olson, 2012) and try to maintain a positive reputation of being "smart" or "nice" (e.g., Fu, Heyman, Qian, Guo, & Lee, 2014). Their behaviors are further modulated by the potential social consequences; they share more when the observer could reciprocate their good deeds in the future than in one-time interactions (Engelmann, Over, Hermann, & Tomasello, 2013). These findings suggest that children care about others' evaluations and engage in behaviors to manage their reputations.

The ability to represent and reason about others' beliefs – Theory of Mind (ToM) – may be particularly important for effective self-presentational behaviors (Asaba & Gweon, 2018; Engelmann & Rapp, 2018; Silver & Shaw, 2018). By using an intuitive theory of others' minds, children can not only infer what others think of them, but also figure out what evidence could improve or maintain these beliefs. Surprisingly, however, there is little empirical support for ToM as a potential mechanism underlying self-presentational behaviors. Prior work in early reputation management behaviors has primarily manipulated whether or not children were being observed by another person. Thus, the role of ToM in self-presentational behaviors remains unclear; some self-presentational behaviors may only require the mere presence of others while more complex interactions may involve sophisticated inferences about the observer's beliefs.

A recent study provides suggestive evidence that preschoolers' self-presentational communicative behaviors depend on the content of others' beliefs about the self, rather than the mere experience of being observed by others (Asaba & Gweon (2018). Findings from this study suggested that 3- and 4-year-old children strategically presented their own competence depending on the observer's prior observations of their failures and successes, even when doing so meant foregoing an opportunity to teach new information to the observer. When the observer had seen the child's failures as well as their final success on a toy (i.e., believing the child can make the toy go), given a chance to demonstrate either the same toy or a toy she had never seen, children strongly preferred to demonstrate the novel toy. However, when the observer left before the child's final success (i.e., believing the child cannot make the toy go), children were more likely

[*] These authors contributed equally to this work.

to demonstrate their success on the same toy rather than the novel toy. Such selective demonstration of one's competence might require the ability to understand how others' observations of the self can generate certain beliefs in others' minds (i.e., "She thinks I cannot operate the toy") and the capacity to infer how additional evidence can change these beliefs (i.e., "demonstrating my success on this toy will make her think I can operate the toy").

These results suggest that children are sensitive to more than the mere presence of others; rather, their self-presentational actions can be *modulated* by representations of others' beliefs about the self. Following prior work on the early development of reputation management, Asaba & Gweon (2018) used a human confederate as the agent who observed children's failures and successes. However, if the process of ascribing beliefs to an observer can elicit reputation management behaviors, the target of such actions should not be limited to other human beings; these behaviors may also manifest in children's interactions with non-human entities, even if there are no real-world consequences to protecting or promoting one's reputation in front of them. If so, even the presence of *a puppet* in the room as children repeatedly fail to activate a toy would lead children to demonstrate their success (i.e., to "change the puppet's belief"), but only in contexts in which children would readily attribute beliefs to the puppet. In other words, the results from Asaba & Gweon (2018) should replicate even when the human confederate is replaced with a hand puppet, specifically when children consider the puppet as a social entity capable of holding a belief.

Decades of work on ToM provide reasonable support for this hypothesis. A large meta-analysis (Wellman, Cross, & Watson, 2001) has shown that children's responses in classic false-belief tasks do not systematically vary depending on the nature of the protagonist (i.e., a drawing, a hand puppet, or a real person); children are willing to attribute perceptual, epistemic, and emotional states to non-human, make-believe entities insofar as they are described and treated as sentient agents that think, feel, and act like humans. These classic ToM tasks usually require children to predict someone's action (e.g., Sally will go to where she thinks the ball is), but would children be motivated enough to share information about their own competence over novel information about a toy in such settings? Such results might attest to the power of mental-state reasoning that encourage children to engage in rich social interactions even with make-believe entities.

What factors may influence children's willingness to ascribe agency to various non-human entities? Prior theoretical work has proposed that children may evaluate an entity's *cognitive property* (Leslie, 1994) – that agents hold certain attitudes (e.g., desires, beliefs) to the truth of propositions. Interestingly, empirical work on children's understanding of agency suggests that children's agency attribution not only relies on the observable features of an entity (e.g., whether or not it has eyes; Johnson, Slaughter, & Carey, 1998) but also can be informed by how others communicate about it (e.g., how often parents talk about psychological property of nonliving kinds; Jipson, Labotka, Callanan, & Gelman, 2018). Critically, adults often *depict* make-believe or imaginary scenes, objects, or agents to children as if they were real, and children readily understand such communicative intent and "play along" with them (Clark, 2016). Thus, children may ascribe agency to a puppet to the extent that other adults treat or depict it as an agent, especially one holding certain beliefs.

The main objective of the current work is to bridge prior work in Theory of Mind and reputation management by clarifying the role of belief-attribution in young children's self-presentational behaviors. To this end, our primary goal was to replicate the findings of Asaba & Gweon (2018) using a puppet that was introduced as the experimenter's "friend" and treated as such (Experiment 1). We predicted that children would go as far as demonstrating their competence to a puppet to change its "belief" about their competence when the adult experimenter treats the puppet as if it were an agent with mental capacities (i.e., with the *cognitive property*). We then provide additional evidence that such behavior is selective to contexts where children have reasons to consider the puppet as a social being capable of holding beliefs (Experiment 2).

Experiment 1: Puppet as Agent

In Experiment 1, we replicated Asaba & Gweon (2018) with 4-year-olds using the same design except that children were "observed" by a puppet rather than a human confederate. Importantly, the experimenter treated this puppet as an agent, calling the puppet her "friend" and referring to the puppet's mental states (i.e., ignorance) about the toys, similarly to how the confederate was treated in Asaba & Gweon (2018).

Methods

Participants 50 4-year-olds (M_{Age}(SD)= 4.49(.29), range = 4.01–4.99; 30 females) were recruited from a university preschool and randomly assigned to the Present (N=25) or Absent (N=25) condition. An additional 14 children were recruited but excluded due to failure on a memory check question (N=13) or technical error (N=1).

Materials We designed two distinct novel toys with different causal mechanisms that each lit up when activated (see Figure 1). The blue toy had two green buttons on the top; pressing the two buttons at the same time would make a rubber frog on the top of the toy light up. The yellow toy had two gray knobs on the left and right sides; turning the two knobs at the same time would make a rubber owl on the top of the toy light up. In reality, the toys were not actually functional but were activated by the experimenter with a remote control switch

Figure 1: Schematic of the toys used in Experiments 1 and 2.

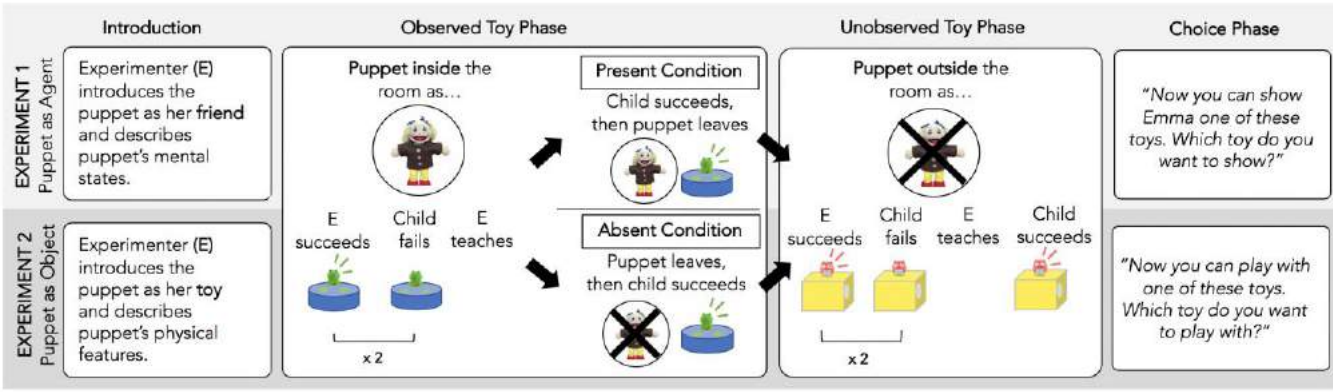


Figure 2: Procedures for Experiments 1 and 2.

underneath the table hidden from the participants' view. A girl hand puppet and a 3" x 4" picture of the puppet was used.

Procedure Children were tested individually in a quiet room at their preschool. The child sat across from an experimenter at a rectangular table. The experiment consisted of the Introduction, Observed Toy, Unobserved Toy, and Choice Phases; only the Observed Toy Phase differed between conditions (see Figure 2).

Introduction Phase: The experimenter showed the child the two toys and said that her *friend* "Emma" (a hand puppet) would watch them play. The experimenter put the puppet on the table facing the child and asked the child to say hello to Emma. Critically, the experimenter described the puppet with respect to its mental states: "Emma has never seen these toys before, and she doesn't know anything about them."

Observed Toy Phase: The puppet "watched" as the child and the experimenter played with one of the two novel toys (i.e., the Observed Toy; blue and yellow toy counterbalanced across participants). The experimenter successfully activated the toy by pressing the two buttons simultaneously (blue toy) or turning the two knobs simultaneously (yellow toy). The child then attempted to operate the toy (i.e., the child pressed the buttons of the blue toy or turned the knobs of the yellow toy) but failed, and the experimenter acknowledged the failure by saying "Hm." The experimenter then succeeded on the toy again and the child failed again. Then, the experimenter instructed the child how to activate the toy: "You have to push this button and this button at the exact same time" (blue toy) or "…turn this and this at the exact same time" (yellow toy). Then, the child was given another chance and succeeded. The experimenter acknowledged the success by saying "Now you know how to play with this toy!"

The critical manipulation between conditions was when the puppet was in the room. In the Present condition, the puppet "watched" the child's initial two failures and final success, then the experimenter brought the puppet outside the room. In the Absent condition, the puppet "watched" the child's initial two failures but was then brought outside the room after the experimenter's instruction on the toy; next, the child succeeded. In both conditions, the experimenter said that "Emma has to go now," before bringing the puppet outside.

Unobserved Toy Phase: The child and experimenter played with the other toy (i.e., the Unobserved Toy) while the puppet was out of the room. The sequence of failures and successes and the experimenter's instruction were identical as in the Observed Toy Phase. The child first failed to activate the toy twice, the experimenter taught the causal mechanism, then the child succeeded. Then, the child successfully activated both the Observed Toy and Unobserved Toy twice more, ensuring that the child was confident in operating both toys.

Choice Phase: With the puppet still outside the room, the experimenter positioned the two toys equidistant from the child. The experimenter placed a photo of the puppet in front of the child and asked, "Now you can show Emma one of these toys. Which toy do you want to show?" Children responded by touching or pointing to one of the toys. Then, children were asked a memory check question, "Did Emma watch when you were playing with this toy or this toy?" Only children who correctly responded to this question (i.e., selecting the Observed Toy) were included in the final sample. At the end, the puppet was brought back into the room, and children demonstrated the chosen toy.

Results and Discussion

In the Absent Condition, the puppet only observed the child's failures, whereas in the Present Condition, the puppet observed the child's failures *and* success. Thus in the Absent Condition, the puppet had an *incorrect* belief about the child's ability on the Observed Toy. We predicted that children would choose the Observed Toy more often in the Absent Condition than in the Present Condition if they were able to track the puppet's beliefs about their abilities and wanted to improve these beliefs.

We ran a generalized linear model (family = binomial) with Condition (dummy coded; Present = 0, Absent = 1), Observed Toy Type (dummy coded; Blue Toy = 0, Yellow Toy = 1), and Age (continuous) as predictors: Toy Choice ~ Condition + Age + Observed Toy Type. We found that Condition significantly predicted children's choice of toys (β = 1.839, z = 2.762, p = .006), but Age (β = .618, z = .519, p = .604) and Toy Type (β = -.268, z = -.400, p = .690) did not. Follow-up analyses confirmed that participants chose the

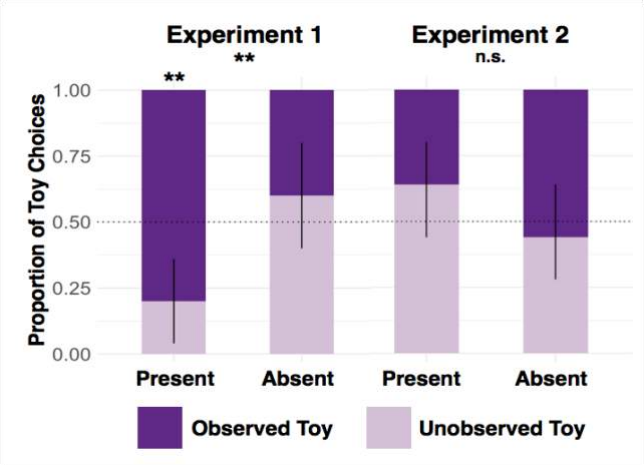


Figure 3: Results from Experiments 1 and 2. Error bars represent 95% confidence intervals. $^{**}p < .01$.

Observed Toy more often in the Absent Condition than in the Present Condition (% choosing the Observed Toy: 60% (Absent) vs. 20% (Present), $p = .009$, Fisher's Exact Test; see Fig.3). In the Present Condition, children strongly preferred to show the Unobserved Toy (80%, $p = .004$, Binomial Test), suggesting that they wanted to show the puppet a novel toy. In the Absent condition, however, children did not show a preference for either toy ($p = .424$, Binomial Test).

These results are in line with our main hypothesis that young children may go so far as to demonstrate their competence to a puppet when it is treated as an agent who is capable of holding beliefs. Specifically, when the puppet had only "observed" (i.e., was on the table facing the child) their failures but not their final success at operating a toy, children were motivated to demonstrate their competence by choosing the same toy to show the puppet, foregoing the opportunity to show a novel toy. These results are consistent with the literature on false-belief reasoning in preschoolers using puppets instead of human experimenters. However, it is nevertheless striking that children went as far as showing off their competence to a *puppet* that is incapable of any real-world social evaluation.

One critical prerequisite for such behaviors is the attribution of a belief state to the puppet. In Experiment 1, we provided a number of contextual cues to encourage children to consider the puppet as an agent, such as introducing the puppet as a "friend", verbally describing its mental states, and asking children to choose a toy to "show" the puppet. However, such behaviors would fail to emerge in the absence of any contextual support for ascribing agency to puppet, such that the puppet is no longer considered as an agent that is capable perceiving the environment or holding a mental state. In Experiment 2, we test this idea by asking whether the pattern of results in Experiment 1 would go away in a context where the puppet is treated as an object (i.e., the experimenter's toy) and children are asked to simply choose a toy to play with (rather than "showing" a toy to the puppet). Note that this is a control experiment where we expect a *failure to replicate* Asaba & Gweon (2018); critically, we also predicted an interaction of condition (Present vs. Absent) and experiment (Exp.1: Agent vs. Exp.2: Object); this would provide additional support that children's representation of the puppet (as an agent vs. an object) modulates the pattern of results.

Experiment 2: Puppet as Object

Methods

Participants 50 4-year-olds ($M_{Age}(SD) = 4.59(.31)$, range = 4.01–4.98; 27 females) were recruited from a university preschool and randomly assigned to the Present (N=25) or Absent (N=25) condition. An additional 28 children were recruited but excluded due to failure on the memory check.

Materials The same materials from Exp.1 were used.

Procedure The procedure was nearly identical to Exp.1, except for three critical modifications as described below, aiming to minimize the perceived agency of the puppet.

Introduction Phase: After introducing the two novel toys, the experimenter told the child that she had another *toy* (the puppet, which was introduced as a "friend" in Exp.1) and that she would put the puppet on the table as they were playing. Rather than referencing the puppet's ignorance about the toys, here the experimenter only described the puppet's physical features: "My puppet has blond hair and brown eyes. I also got the blue ribbons to tie my puppet's hair."

Observed Toy Phase: Children failed twice and succeeded once on the Observed Toy, and the Present and Absent conditions varied by whether the puppet was present for children's final success. In contrast to Experiment 1, the experimenter stated that someone else needed the puppet, rather than that the puppet needed to go. Additionally, the puppet's presence was emphasized at the beginning ("Now the puppet is on the table") and children helped bring the puppet outside the room to ensure that children were paying attention to the puppet. These changes were included to help children remember when the puppet was in the room.

Unobserved Toy Phase: Same as in Exp.1.

Choice Phase: The experimenter brought the puppet back onto the table. Importantly, the test question used in Experiment 1 (and in Asaba & Gweon, 2018; "Which toy do you want to show my friend?") implies that the puppet should be treated as an agent; using the same question would provide a strong signal to the child that the experimenter wants the child to "communicate" to the puppet. Thus, in Experiment 2, the experimenter asked instead: "Now you can play with one of these toys. Which toy do you want to play with?" while the puppet (instead of a photo of the puppet) was placed on the table, facing the child. We come back to the role of the final question in the General Discussion. However, we did use a similar memory check as Exp.1 by asking: "Was the puppet here on this table when you were playing with this toy or this toy?" Only children who correctly responded to this question were included in the final sample.

Results and Discussion

Here, the experimenter treated the puppet as her toy, and we predicted that if children's strategic self-promotion was sensitive to the nature of their "observer", then children would not strategically communicate to a puppet depicted as an object. Specifically, we predicted that there would be no difference in children's choices across conditions.

We ran the same generalized linear model as in Exp. 1 and found that Condition (β = -.747, z = -1.259, p = .208), Age (β =1.28, z = 1.232, p = .214), and Observed Toy Type (β = .603, z = .997, p = .319) did not predict children's choice of toys. Indeed, children chose the Observed Toy at similar rates in the Absent Condition and the Present Condition (% choosing the Observed Toy: 44% (Absent) vs. 64% (Present), p = .256, Fisher's Exact Test). Further, children did not selectively choose a toy in either condition (Present: p = .23, Absent: p = .69, Binomial Tests). Given the high rate of exclusion, we ran analyses including participants who failed the memory check question and found the same pattern of results: 53.4% (Present) vs. 38.9% (Absent) of participants chose the Observed Toy (p = .259, Fisher's Exact Test). As predicted, in this study, the results did not show a clear pattern for children's choice of toys as in Asaba & Gweon (2018).

The critical difference between experiments was whether the social context encouraged children to consider the puppet as an agent (capable of holding a belief) or an object. This allowed us to test the additional hypothesis that children would strategically choose which toy to show the puppet in a context where children had reason to attribute beliefs to the puppet (Exp.1), but not when it was treated as an object (Exp.2). To compare across experiments, we ran a generalized linear model (family = binomial) with Condition, Experiment, and Age (continuous) as predictors: Toy Choice ~ Condition * Experiment + Age. As expected, we found a significant Condition x Experiment interaction (β = -2.709, z = -2.709, p = .007), as well as significant main effects for Condition (β = 1.937, z = 2.608, p = .009) and Experiment (β = 2.013, z = 2.702, p = .007), but not Age (β = .816, z = 1.06, p = .289). The significant interaction between condition and experiment provides additional support for the idea that children's self-promotional behaviors are driven by the belief that children ascribe to the observer rather than the mere presence of an observer.

General Discussion

Across two experiments, we found that young children readily demonstrated their competence to a puppet, and that their self-promotional communication was modulated by the social context in which children interacted with the puppet.

As our primary goal, Exp. 1 provided a conceptual replication of Asaba & Gweon (2018). Remarkably, when a puppet had only observed their failures on a toy, four-year-olds demonstrated their success on the same toy to the puppet (rather than demonstrating a novel toy they played with in the absence of the puppet). This suggests that they attributed beliefs to the puppet about their own competence when the puppet was present for their failures, and they wanted to demonstrate their success to revise the puppet's (arguably false) beliefs. However, this pattern of results was found only in Experiment 1, when the experimenter treated the puppet as an agent and asked the child to demonstrate a toy to it; we did not find this pattern in Experiment 2 when the experimenter treated the puppet as an object and asked the child to choose a toy to play with.

Collectively, these results suggest that children are willing to engage in self-presentation behaviors to a non-human agent. Even though "losing face" in front of a puppet could not bear any foreseeable, real-world social consequences, children nonetheless tried to present positive information about the self (i.e., their success on the Observed Toy) instead of information about a novel toy (Unobserved Toy). Note that the puppet's belief was never explicitly mentioned; children inferred the belief from its "observations" of their own failures and successes and selectively provided evidence that might improve the puppet's beliefs. As irrational as these behaviors might seem, children were not indiscriminately showing off their competence to any entity; in the absence of any contextual cues to ascribe mental capacities to the puppet, children did not show these behaviors.

What makes children *want* to have a positive impression, even to a puppet? A large literature documents strong human desires to make positive impressions in the minds of others, regardless of age, gender, or culture. One perspective suggests that these attempts reflect a desire to build a shared reality with others (Harris, 2017); when children perceive gaps in knowledge or understanding between themselves and other people, they are motivated to remedy them by providing additional information. Our results provide additional support for this idea, and further show that the process of belief attribution (about the self) may be a key modulator of the motivation to preserve (good) or improve (bad) images of ourselves. Another theoretical perspective is that this motivation comes from the desire to be selected by others as social partners (Engelmann & Rapp, 2018). From this view, without the social pressure to be seen as desirable social partners, reputation management behaviors would not emerge. Our results do not necessarily contradict this view. Though there are no clear consequences to showing off to a puppet, children may still consider the pragmatic demands communicated by the experimenter or they may be motivated to practice self-promoting. Further, the desire to be accepted by others may be a more basic instinct even when people are not explicitly aware of them (Dweck, 2017).

More specifically, our task might have encouraged such behavior by asking children to "show" one of the toys to the puppet (although this was an ambiguous request either to show off or to teach novel information). By contrast, Experiment 2 provided little contextual support for these motivations to manifest. Although the two experiments were well matched in children's experiences with the toy and the time children spent in the presence of the puppet, Experiment 2 differed from Experiment 1 in two ways: the experimenter did not provide agency cues about the puppet and also asked the child to choose a toy to play with (rather than asking to

choose what to "show" to the puppet). Although this was an important design decision to prevent children from retrospectively attributing agency to the puppet ("showing" implies the ability to perceive), these results do not allow us to tease apart the relative importance of others' treatment of the puppet versus the nature of the final question.

An intriguing possibility is that even in contexts where adults initially treat the puppet as a toy, children might retrospectively ascribe a belief to the agent (see Király, Oláh, Csibra, & Kovács, 2018) when the experimenter asks which toy children want to "show" to the puppet. Such results might suggest that children are picking up subtle cues that reflect the ways adults communicate about the sentience of nonhuman entities (Weisman et al., 2017). Conversely, prior work in reputation management (e.g., Engelmann et al., 2012) suggests that children exhibit self-presentational behaviors even in the absence of explicit requests to communicate with their observer; thus, given clear evidence that adults treat the puppet as an agent (as in Exp.1), children might have still show similar self-presentational behaviors even when they are simply asked to choose a toy to play with. Future work might test the idea that adults' explicit treatment of the puppet and the nature of the final question might independently contribute to these behaviors.

Interestingly, the exclusion rate was noticeably high in Exp.2. It is possible that children may have not paid much attention to the puppet and subsequently had difficulty answering the memory question because of the social context. Understanding how children's memory might depend on the social context of their interactions is an area for future work.

Note that children were split between the Observed and Unobserved Toys in the Absent Condition in Exp. 1, as in Asaba & Gweon (2018). This might reflect genuine conflict between the desire to provide new information with the Unobserved Toy versus demonstrate their abilities on the Observed Toy; however, one might wonder if children were not considering the puppet's *beliefs* about their abilities, but simply wanted to show a success on either toy. While this still requires attributions of ignorance, ongoing work shows that when the confederate is fully knowledgeable about the toys in the Absent Condition, children selectively choose the Observed Toy, suggesting that they want to specifically revise the confederate's beliefs about their ability on that toy.

Broadly, these findings are consistent with the hypothesis that belief-reasoning capacities play a role in children's reputation management behaviors. Although work in Theory of Mind has traditionally focused on reasoning about others' beliefs about observable, objective physical states of the world (e.g., Wimmer & Perner, 1983), our work suggests that young children can also reason about beliefs concerning unobservable, subjective qualities of the self. Just as young children understand that others' observations (e.g., Anne *sees* the ball in the box) lead to others' beliefs (e.g., Anne *thinks* the ball is in the box, Wellman et al., 2001), they also understand that others' observations of their failures and successes informs others' beliefs about the self. Further, just as young children provide information to improve others' beliefs about the world (e.g., Gweon, Shafto, & Schulz, 2018), children in this study actively provided information about the self given others' beliefs about the self.

However, although belief attribution was critical for children's strategic communication in our task, not all reputation management behaviors may require rich psychological reasoning abilities. Rather, some behaviors may be a response to the mere presence of others, and mental-state reasoning may be involved only in certain contexts where belief attribution is necessary to motivate the behavior (e.g., when children are attempting to *revise* others' beliefs about them). If this is the case, even among children who clearly employ some reputation management behaviors (e.g., cheating less when others are present), the individual differences in their ability to select appropriate information or action to change others' beliefs about their competence might positively correlate with their performance on standard measures of Theory of Mind.

Here, we took advantage of prior work suggesting children's willingness to attribute mental states to puppets (Wellman et al., 2001). Critically, whereas prior work has manipulated the physical features of an entity (e.g., Johnson et al., 1998), we manipulated how the experimenter *treated* it. Our findings suggest that children differentially perceived the puppet depending on the experimenter's interactions with it. Along with recent work suggesting that children attend to agency cues and interact with non-human entities accordingly (Breazeal et al., 2018), the ways in which adults *treat* entities (e.g., other humans, toys, pets, deities) might have deep consequences for how young children subsequently treat them. Further, it is possible that young children might use graded levels of agency (see Weisman et al., 2017) to determine what to communicate to others. Whether children might prioritize their reputations for those with greater mental capacities (e.g., robots over puppets, or adults over babies) is an open area for future work. Further work could also consider directly asking children to evaluate the perceived agency of the interlocutor that has varying degrees of cues to agency, and investigating how this perceived agency influences children's behaviors.

These findings may also be useful to researchers in cognitive development whose work utilizes puppets in their methodology. For many studies, puppets are more than just logistically convenient stand-ins for human experimenters; they are often a necessary piece of the methodology, especially for studies that must present properties of agents that are implausible in human adults (e.g., someone who does not know labels of simple household objects) or tricky to convey with human actors in experimental contexts (e.g., someone who attempts to climb a hill). A useful takeaway from the current results is that just as children attribute beliefs about the external world to puppets (e.g., location of Sally's ball, see Wellman et al. 2001), they also attribute rich beliefs about abstract qualities of the self such as competence or abilities; Four-year-olds readily attempted to change a puppet's beliefs as if it was human. Importantly, their tendency to treat a puppet as an agent may be critically

modulated by the ways in which the experimenter had treated it. One open question is whether children genuinely believe that the puppet is an agent, or whether they are perceiving the experimenter's communicative intent (i.e., the experimenter wants to communicate to the child that the puppet is a friend of hers) and therefore following along by engaging in a pretend play with the experimenter. Although the current work does not directly address this question, it is possible that children's reasons for attributing beliefs to a non-human entity depends on age (e.g., younger children may treat it as an actual agent, whereas older children are aware that they are make-believe but still engage in pretense).

What others think of us is deeply important for our everyday interactions with others, and the ability to reason about others' minds might allow us to reason about others' beliefs about us in savvy, sophisticated ways. Our findings suggest that children's strategic self-presentational behaviors are specific to the social context. Children do not promiscuously show off to anyone or anything; rather, they are sensitive to cues about the object's agency and specifically communicate about the self to other agents.

All data and analyses are available here:
https://osf.io/3zsb7/

Acknowledgments

We thank Habin Shin and Hannah French Levy for help with data collection. We also thank the parents and families of Bing Nursery School. This work was supported by an NSFGRFP to MA, an SUTD PGF to XL, and a McDonnell Scholars Award to HG.

References

Asaba, M., & Gweon, H. (2018). Look, I can do it! Young children forego opportunities to teach others to demonstrate their own competence. In T.T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 106-111). Austin, TX: Cognitive Science Society.

Breazeal, C., Harris, P. L., DeSteno, D., Kory Westlund, J. M., Dickens, L., & Jeong, S. (2016). Young children treat robots as informants. *Topics in Cognitive Science*, *8*(2), 481-491.

Clark, H. H. (2016). Depicting as a method of communication. *Psychological Review*, *123*(3), 324.

Dweck, C. S. (2017). From needs to goals and representations: Foundations for a unified theory of motivation, personality, and development. *Psychological Review*, *124*(6), 689-719.

Engelmann, J. M., Herrmann, E., & Tomasello, M. (2012). Five-year olds, but not chimpanzees, attempt to manage their reputations. *PLoS One*, *7*(10), e48433.

Engelmann, J. M., Over, H., Herrmann, E., & Tomasello, M. (2013). Young children care more about their reputation with ingroup members and potential reciprocators. *Developmental Science*, *16*(6), 952-958.

Engelmann, J. M., & Rapp, D. J. (2018). The influence of reputational concerns on children's prosociality. *Current Opinion in Psychology*, *20*, 92-95.

Fu, G., Heyman, G. D., Qian, M., Guo, T., & Lee, K. (2016). Young children with a positive reputation to maintain are less likely to cheat. *Developmental Science*, *19*(2), 275-283.

Gweon, H., Shafto, P. & Schulz, L.E. (2018). Development of children's sensitivity to over-informativeness in learning and teaching. *Developmental Psychology, 54*(11), 2113-2125.

Harris, P. (2017). Tell, ask, repair: Early responding to discordant reality. *Motivation Science, 3*(3), 275-286.

Hicks, C. M., Liu, D., & Heyman, G. D. (2015). Young children's beliefs about self-disclosure of performance failure and success. *British Journal of Developmental Psychology*, *33*(1), 123-135.

Jipson, J. L., Labotka, D., Callahan, M. A., & Gelman, S. A. (2018). How conversations with parents may help children learn to separate the sheep from the goats (and the robots). In Saylor, M. M. & Ganea, P. (Eds), *Active learning from infancy to childhood* (pp. 189-212). New York: Springer.

Johnson, S., Slaughter, V., & Carey, S. (1998). Whose gaze will infants follow? The elicitation of gaze-following in 12-month-olds. *Developmental Science, 1*, 233-238.

Király, I., Oláh, K., Csibra, G., & Kovács, Á. M. (2018). Retrospective attribution of false beliefs in 3-year-old children. *Proceedings of the National Academy of Sciences*, *115*(45), 11477-11482.

Leimgruber, K. L., Shaw, A., Santos, L. R., & Olson, K. R. (2012). Young children are more generous when others are aware of their actions. *PloS One*, *7*(10), e48292.

Leslie, A. (1994). ToMM, ToBY, and Agency: Core architecture and domain specificity. In L. Hirschfeld & S. Gelman (Eds.), *Mapping the Mind: Domain Specificity in Cognition and Culture* (pp. 119-148). Cambridge: Cambridge University Press.

Nowak, M. A., & Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature*, *437*(7063), 1291.

Silver, I. M., & Shaw, A. (2018). Pint-Sized Public Relations: The Development of Reputation Management. *Trends in Cognitive Sciences*, *22*(4), 277-279.

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: the truth about false belief. *Child Development*, *72*(3), 655–684.

Weisman, K., Dweck, C. S., & Markman, E. M. (2017). Rethinking people's conceptions of mental life. *Proceedings of the National Academy of Sciences*, *114*(43), 11374-11379.

Modifying social dimensions of human faces with ModifAE

Chad Atalla[1]
Computer Science and Engineering
University of California, San Diego
chada@ucsd.edu

Amanda Song[1]
Cognitive Science
University of California, San Diego
mas065@ucsd.edu

Bartholomew Tam
Electrical and Computer Engineering
University of California, San Diego
b4tam@ucsd.edu

Asmitha Rathis
Computer Science and Engineering
University of California, San Diego
arathis@eng.ucsd.edu

Gary Cottrell
Computer Science and Engineering
University of California, San Diego
gary@eng.ucsd.edu

Abstract

At first glance, humans extract social judgments from faces, including how trustworthy, attractive, and aggressive they look. These impressions have profound social, economic, and political consequences, as they subconsciously influence decisions like voting and criminal sentencing. Therefore, understanding human perception of these judgments is important for the social sciences. In this work, we present a modifying autoencoder (ModifAE, pronounced "modify") that can model and alter these facial impressions. We assemble a face impression dataset large enough for training a generative model by applying a state-of-the-art (SOTA) impression predictor to faces from CelebA. Then, we apply ModifAE to learn generalizable modifications of these continuous-valued traits in faces (e.g., make a face look slightly more intelligent or much less aggressive). ModifAE can modify face images to create controlled social science experimental datasets, and it can reveal dataset biases by creating direct visualizations of what makes a face salient in social dimensions. The ModifAE architecture is also smaller and faster than SOTA image-to-image translation models, while outperforming SOTA in quantitative evaluations.

Keywords: neural networks; generative models; face recognition; social perception; image modification

Introduction and Related Work

Humans quickly form subjective impressions of faces, judging traits like facial attractiveness, trustworthiness, and aggressiveness (Todorov, Olivola, Dotsch, & Mende-Siedlecki, 2015). Despite the continuous scale and subjective nature of these social judgments, there is often a consensus among humans in how traits are perceived (e.g., human raters agree that certain faces appear relatively more trustworthy) (Falvello, Vinson, Ferrari, & Todorov, 2015; Eisenthal, Dror, & Ruppin, 2006). Social judgments of faces have a significant impact on social outcomes, ranging from electoral success to sentencing decisions (Dumas & Testé, 2006; Oosterhof & Todorov, 2008). Modeling is one way to understand these critical split-second impressions. Another way is through explicit human-judged experiments, which require carefully controlled datasets (e.g., building a dataset of faces which vary in "trustworthiness" while remaining consistent across age, gender, and "attractiveness"). In this work, we develop a system to model these impressions, visualize human perceptual biases, and create isolated image modifications for experimental datasets.

Choosing a subset of social impressions for modeling, we look to the 10k US Adult Faces Database (Bainbridge, Isola, & Oliva, 2013a). Bainbridge et al. (2013a) investigated what social attributes influence the memorability of a face. They compiled a list of 20 spontaneous social judgments and the corresponding opposite traits. Then, they assembled a human-judged dataset of trait ratings on 2,222 faces from the 10k US Adult Faces Database. Among the 40 traits, "aggressive," "attractive," "intelligent," "emotional," and "trustworthy" were frequently used in human-written face descriptions, played a significant role in face memorability, and had high rating agreement levels between human judges. Therefore, we choose them as the subset of social impressions for modeling in this paper.

To create controlled face datasets and visualize perceptual biases, a generative model is needed. Recent generative image models have been successful in creating high-resolution, high fidelity, and diverse images (Brock, Donahue, & Simonyan, 2018; Karras, Aila, Laine, & Lehtinen, 2017; Choi et al., 2017). However, in the face space, most generative models have focused on editing or modifying categorical and objective attributes, such as expression, gender, hair color, and identity (Choi et al., 2017). These categorical changes are referred to as "image to image translation." Here, we focus on modifying continuous attributes of an image, which we refer to as "continuous image modification" (Isola, Zhu, Zhou, & Efros, 2016). Regarding continuous image modification, there has been work on modifying the memorability (Khosla, Bainbridge, Torralba, & Oliva, 2013), and attractiveness of a face (Leyvand, Cohen-Or, Dror, & Lischinski, 2008), but these models do not generalize to wider sets of social impressions. Also, some researchers have generated fake faces with particular social impressions, but these models cannot modify real face images (Vernon, Sutherland, Young, & Hartley, 2014; Oosterhof & Todorov, 2008). So, no prior work has attempted to automatically modify general continuous social impressions of real face photographs.

Conditional generative adversarial networks (GANs) (Goodfellow et al., 2014) have become the most popular tool for the image to image translation task, so we compare against a recent GAN as a state-of-the-art (SOTA) reference point (Isola et al., 2016; Mirza & Osindero, 2014; Lee & Seok, 2017). StarGAN (Choi et al., 2017) is a SOTA conditional GAN that can modify multiple binary categorical traits at once, maintaining identifying traits of the original image using "cycle consistency" (Zhu, Park, Isola, & Efros, 2017). StarGAN consists of two networks: a generator and discrim-

inator. The generator takes an image and a set of desired categorical traits, producing a modified image. The discriminator takes an image and makes a prediction about its realism and categorical traits. By comparing the fake images to genuine images, the discriminator gives feedback to the generator about how to make the image and desired traits appear more realistic.

Despite the success of GANs in categorical image-to-image translation, they cannot perform continuous image modification without binarizing the task and have architectural downsides. GANs typically have many parameters and long training times. They are also sensitive to hyperparameter selection and the delicate balance between generator and discriminator training. Therefore, they can be difficult to train compared to a single-network model. Finally, they suffer from a lack of interpretability, offering no means of visualizing or understanding why the model makes the modifications it does.

In this work, we address these architectural concerns while designing a neural network to model and automatically modify continuous-scale face traits (rated from 1 to 9) in real face images. We create a sufficiently large dataset for training a generative model by combining CelebA images with a SOTA face impression predictive model (Liu, Luo, Wang, & Tang, 2015). Enabling interpretable bias visualization and controlled dataset creation for human face impressions, we introduce ModifAE. ModifAE (pronounced "modify") is a single-network image modification autoencoder.

Subjective Judgment Face Dataset

Building a Large Scale Facial Impression Dataset

To train a generative model on continuous face traits, we need a large and diverse dataset. We start with images from the CelebA dataset (Liu et al., 2015), which are annotated with binary categorical labels such as "wearing a hat" but lack continuous ratings of social impressions.

To generate continuous social impression ratings of these faces, we use our previous social impression predictive model (Song, Li, Atalla, & Cottrell, 2017). The model was trained on a smaller dataset (2,222 faces from the MIT 10k US faces dataset (Bainbridge, Isola, & Oliva, 2013b)) that had been annotated with ratings of 40 social traits on a scale from 1 to 9 by 15 raters for each face. Now, we focus on the subset of traits with the highest correlation between human judges: emotional, aggressive, trustworthy, responsible, attractive and intelligent. We apply this predictive model to about 190,000 faces from the CelebA dataset. Example faces and their predicted ratings are shown in Figure 1. Note that 6-8 are high ratings, and 2-4 are low ratings.

Validating the Algorithm-Augmented Dataset

Evaluating the effectiveness of this algorithm-augmented dataset, we collect human judgments of the model's predictions in two ways: pairwise comparison and single image ratings. All participants were recruited from Amazon Mechanical Turk (AMT).

Figure 1: CelebA faces and their predicted traits.

Table 1: Validation of the impression prediction model

Attribute	Accuracy	Attribute	Correlation
Aggressive	0.95	Aggressive	0.76^{***}
Emotional	0.92	Attractive	0.90^{***}
Trustworthy	0.88	Trustworthy	0.73^{***}
Responsible	0.78	Intelligent	0.62^{***}

For pairwise comparison, we test four attributes: aggressive, responsible, trustworthy and emotional. For each trait, we compose 40 pairs of images. Within each pair, one is from the 40 faces of highest scores, and the other is from the 40 faces of lowest scores, as predicted by the model. We then ask human participants which face better exemplifies the predicted trait. Each trait's 40 pairs are evaluated by 30 AMT workers. We then calculate the overall likelihood that the face of higher predicted score is chosen, which we call "accuracy." The results are shown in the left side of Table 1. The attributes predicted by the model align well with human judgments.

For the single-rating experiment, we examined four traits: attractive, aggressive, trustworthy and intelligent. For each trait, we chose 80 faces whose predicted scores are evenly spread across a range of predictions (i.e., from 2 to 8). Each participant is presented with a random sequence of 80 faces, and is asked to give each face a rating on a 1-9 scale for the specified trait. Every face is rated by 15 subjects, and we compute the average. Lastly, we compute the Spearman rank correlation between the average human ratings and the model's predictions of the same set of faces for each trait. For all four traits, human average ratings are significantly correlated with model predictions (*** indicates $p < 0.001$), as seen in the right side of Table 1.

Given the pairwise and single image rating results, we consider the predicted scores as roughly equivalent to human judgments. Hence, in the next section, we train our face mod-

ification model with these ratings.

ModifAE

ModifAE is a single network autoencoder which implicitly learns to modify continuous face impression traits in images (illustrated in Figure 2). Here, we elaborate on the architecture, training procedure, and mechanism of the ModifAE model.

Model Architecture

The ModifAE architecture consists of a single autoencoder with two (image and trait) sets of inputs which pass through an encoding stage, are fused (by averaging) in the middle of the network, and are then fed into an image decoder.

The image encoder and decoder are identical to the encode and decode portions of the StarGAN generator network, scaled to fewer channels (Choi et al., 2017). More specifically, the network has two downsampling convolutional layers with stride two, four residual blocks, a bottleneck with 16 channels, four more residual blocks, then two upsampling transposed convolutional layers with stride two (Choi et al., 2017). All layers have ReLU activation. We use the first half of this network (including the bottleneck) as the image encoder. We use the remainder of the network as the image decoder. Theoretically, this portion could consist of the encode and decode halves of any image autoencoder; we chose the architecture from StarGAN for the sake of comparability.

The trait encoder takes a 1-dimensional set of traits, feeds these into a single dense layer with Leaky ReLU activation, and reshapes the output to create a vector of the identical shape as the image encoder output. The outputs of the trait and image encoders are then combined into a single latent representation by averaging.

In order to encourage the model to encode the trait information, which is otherwise unnecessary to reproduce the image, 50% dropout is applied to the values from the image encoder. This is then averaged with the trait encoder output to arrive at the combined latent representation. The image decoder projects the representation back into image space, creating the single output image. The architecture is depicted in Figure 2, where "convs" refers to residual convolutional blocks from StarGAN.

Training Procedure

ModifAE is exclusively trained on an autoencoding task. We train ModifAE using the Adam optimizer (Kingma & Ba, 2014) and train for 100 epochs on CelebA images (Liu et al., 2015). The objective is to optimize a single loss function based on two terms. We use the L_1 loss on the image autoencoder. We also optimize the L_1 loss between the trait encoder and image encoder. The total loss is:

$$L = \frac{1}{N}\sum_{p=1}^{N} |x_p - AE(x_p)| + |E(x_p) - E(y_p)| \qquad (1)$$

where x_p is the p^{th} image example, y_p is its trait vector, $E(\cdot)$ is the result of the trait or image encoder, and $AE(\cdot)$ is the output of the full-architecture autoencoder. The second term in this loss function encourages the network to have a similar representation between the trait and the image encodings. The trait encoder obviously does not "know" what the image is, but this constrains the image encoding to include information about the trait.

Why the Model Learns Implicitly to Modify Images

Each image is encoded along with its predicted traits. The image encoder compresses the image down to a bottlenecked latent space, where higher level features about the image are encoded. Simultaneously, the trait encoder projects the given traits to the same latent space, creating an average face representation with those ratings.

Because dropout is applied to the face encoding, the decoder has to use the trait information to "fill in the gaps" in the face representation. Therefore, at training time, faithfully reconstructing the image is reliant on information coming from the trait encoder, and the trait encoder learns to mimic average latent distributions of images with the provided ratings.

At test time, an image can be passed in with any desired traits. The trait encoder estimates the latent space for images with those traits, and the decoder responds by altering the face image towards the encoded trait. Hence, the output image resembles the original but is changed according to the provided traits.

Experiments and Results

In this section, we provide examples of ModifAE's modifications and interpretable transformation maps. We also report an experiment which quantitatively compares the effectiveness of ModifAE and StarGAN with a user study, and we numerically compare the ModifAE architecture with other relevant systems.

Qualitative Evaluation

Multi-Trait Traversals Here, we show that ModifAE is capable of making continuous modifications on multiple traits with a single model (see Figure 3 and Figure 4). This enables ModifAE to modify some traits while holding others traits constant, which can be applied to creating datasets with controlled and isolated modifications for social psychology experiments.

For Figure 3, we trained ModifAE on two traits: "attractive" and "aggressive." The picture in the upper left corner is the original. At the (0,0) point in Figure 3 (unattractive and not aggressive) the man's mouth is fairly neutral, and his features are not very pronounced. As attractiveness and aggressiveness increase, the angles of the face become sharper, there is more definition of features like eyes and eyebrows, and the smile shrinks.

Figure 4 shows interpolations generated by two models. Each was trained on a social trait and a gender category from CelebA. Then, each trait was interpolated while holding the gender bias constant. The resulting figure shows how perception of "aggressiveness" may vary across genders. Likewise,

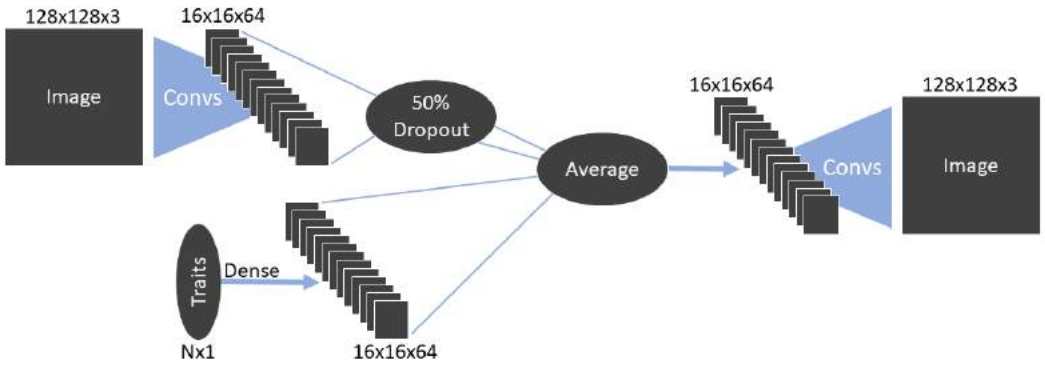


Figure 2: General illustration of ModifAE architecture.

this method can show how other traits may be less correlated with gender perception.

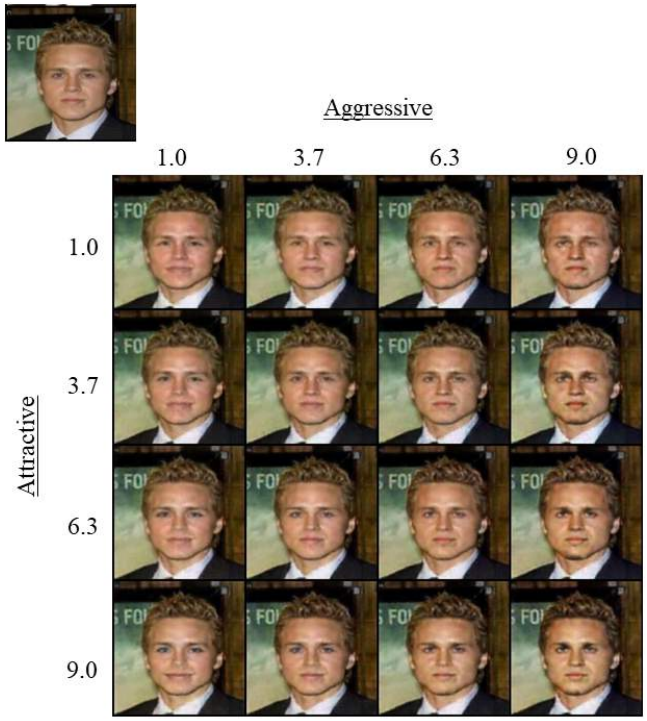


Figure 3: Continuous value, multi-trait image modification by ModifAE.

Qualitative Comparison to StarGAN Comparing our model to StarGAN (Choi et al., 2017), we binarize the continuous traits by doing a median split on the continuous-valued traits and train StarGAN on these two groups (low and high). This is necessary because StarGAN inherently only makes binary changes. The results are shown in Figure 5. While StarGAN produces high-resolution image reconstructions, they occasionally suffer from color distortions or lack of apparent changes. ModifAE makes subtle and reliable modifications to the original images, changing the way

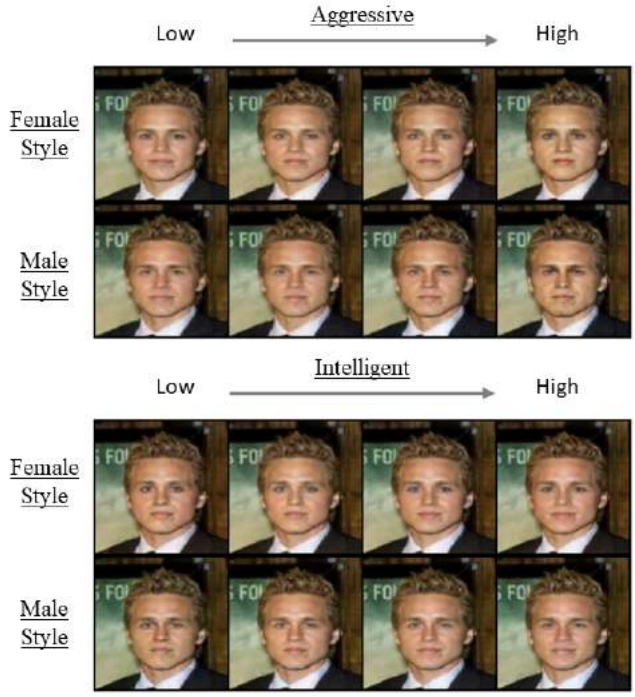


Figure 4: Continuous changes of a face while holding gender bias constant.

the social traits are perceived. In the images produced by ModifAE, more trustworthy faces smile more, and appear to have eyes set farther apart. The ModifAE attractive faces appear to smile more and notably have more well-defined eyes.

Interpretable Transformation Maps As mentioned above, ModifAE addresses the issue of interpretability in generative models. We provide a window into the model's representation of the traits by decoding the representation generated by the trait encoder without giving any actual image input. Figure 6 shows a traversal of the learned "trait faces" or "transformation maps" of attractiveness and intelligence. In this case, we trained the model on a combination

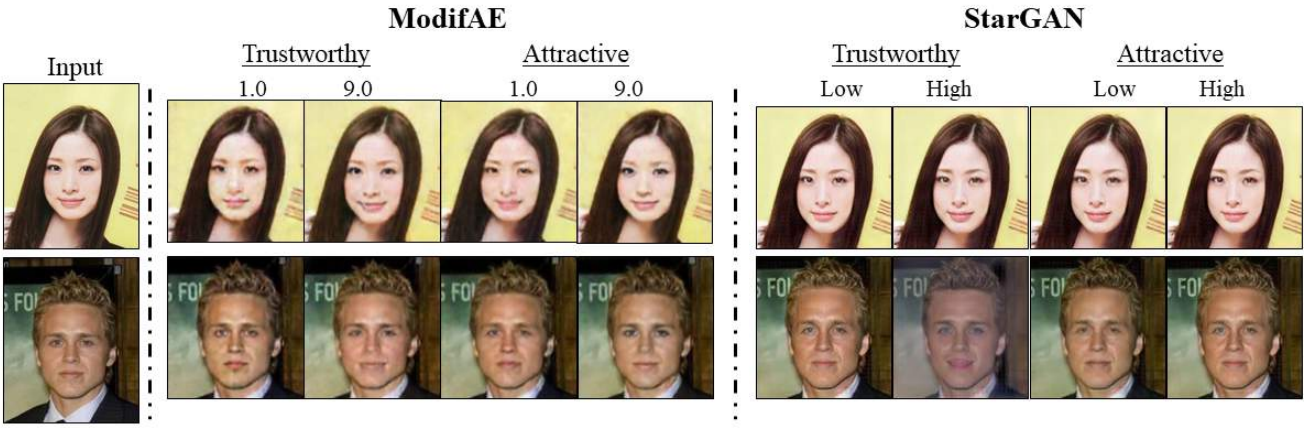


Figure 5: Comparison of ModifAE and StarGAN modifications.

of gender and the given trait, so we show a traversal of the model's representations for male and female faces separately.

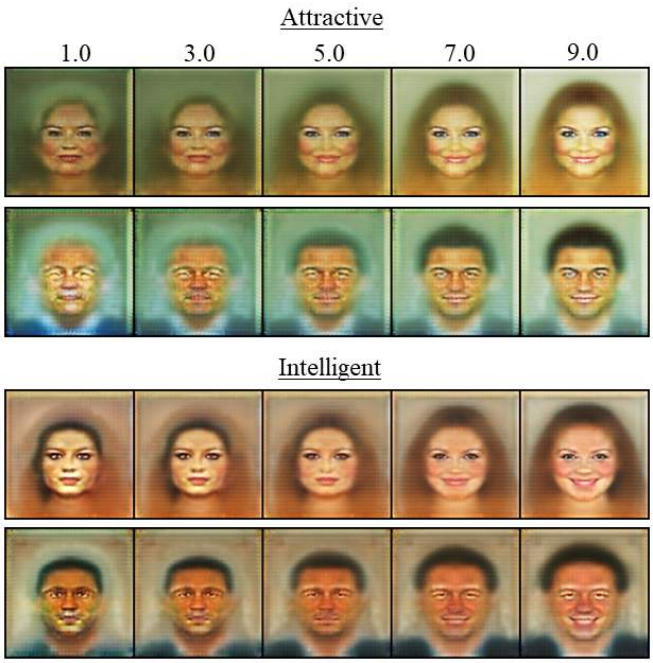


Figure 6: Visualization of model's internal perception of traits. Each is a traversal of a trait (increasing left to right) while gender is held constant.

Quantitative Evaluation

Quantitative Comparison with StarGAN To evaluate the quality of ModifAE's continuous subjective trait modifications, we perform Amazon Mechanical Turk (AMT) experiments on four traits: aggressive, attractive, trustworthy and intelligent. For each trait, we created 90 image pairs, of which 80 are the same identity modified to be at high and low values of each trait. For StarGAN, we used a median split of low and high rated traits to train the model. ModifAE was trained as previously described. For each model, then, faces were modified to be low or high on each trait. Subjects judged which face had more of the particular trait. 10 pairs were repeats in order to judge subject consistency, and 10 pairs were unmodified CelebA faces with high and low ratings. These latter we called "ground truth" pairs to test whether subjects were paying attention. Subjects whose ratings on these pairs were at chance or below were rejected.

Hence, for each trait, we present participants with a sequence of 100 image pairs, and participants are asked to pick which image most exemplifies the trait in each pair.[1] Each pair was evaluated by 15 subjects.

We calculate the fraction of pairs in which subjects chose the image with the higher modified trait across all participants and all pairs. If they choose the face that was modified to be higher in the trait, then they agree with the model's modifications. The results are shown in Table 2. We perform a binomial test to determine whether each trait's accuracy is significantly below or above chance ($^{***}p < 0.001$). Note that the fourth column "Ground Truth" indicates the overall accuracy of the unmodified "ground truth" pairs. Given the variance in human impression judgments, these numbers serve as a reference ceiling for how well the models can perform.

Evaluating ModifAE's Continuity Since ModifAE is able to generate continuous modifications, we evaluated this property by creating two more same-face pairs: Ones modified to have low values and middle values, and ones modified to have middle values and high values. We obtain human agreement (accuracy) over the Low-Mid and Mid-High pairs for each of the four traits. The results are shown in Table 3.

Model Size and Training Time

In contrast with GANs, ModifAE requires fewer parameters and less time to train. StarGAN takes about 24 hours to train on CelebA (Choi et al., 2017); ModifAE takes less than 11

[1] In a pilot experiment, we asked subjects to rate faces with different identities generated in a fine continuum, but found significant variance with no correlation to the intended scores, presumably because the images were not differentiable at that fine a grain.

Table 2: Comparison of ModifAE with StarGAN

Attribute	**ModifAE**	StarGAN	"Ground Truth"
Aggressive	0.68^{***}	0.72^{***}	0.90^{***}
Attractive	0.68^{***}	0.51	0.94^{***}
Trustworthy	0.63^{***}	0.40	0.87^{***}
Intelligent	0.68^{***}	0.58^{***}	0.81^{***}

Table 3: ModifAE Low-Mid-High Level Self-comparison

Attribute	Low-Mid	Mid-High	Low-High
Aggressive	0.60^{***}	0.52	0.68^{***}
Attractive	0.59^{***}	0.52	0.68^{***}
Trustworthy	0.61^{***}	0.53^{*}	0.63^{***}
Intelligent	0.60^{***}	0.50	0.68^{***}

hours. Table 4 shows the number of parameters required by different models trained on the CelebA dataset. The listed values are as reported in the original papers (Perarnau, van de Weijer, Raducanu, & Álvarez, 2016; Zhu et al., 2017) and in the parameter comparisons of Choi et al. (2017).

Note that the majority (over 40M) of StarGAN's parameters are in the discriminator network, and ModifAE uses a smaller version of the StarGAN generator. Also, ModifAE's relatively small trait encoder is the only part of the model which scales with supervising additional traits, so learning more traits with a single model is cheaper with ModifAE. Together, these properties mean that ModifAE takes over fifty times fewer parameters than any of the competing models.

Discussion

Quantitative Experiment Discussion

From Table 2, we can see that for all four traits, ModifAE produces pairs that yield above chance level human agreement. In three out of the four traits, ModifAE significantly outperforms StarGAN; whereas for the aggressive trait, StarGAN performs only slightly better than ModifAE. StarGAN is good at creating discrete changes in facial expressions, which accounts for this advantage.

From Table 3, we find that all the low-mid pairs yield significantly above chance accuracy, yet for mid-high level, only trustworthy pairs have accuracy slightly above chance ($p < 0.05^{*}$). This suggests that human psychological face space is nonlinear and has more differentiation towards the low- to mid-range of social dimensions. Another possibility is that when our model generates faces that are of more extreme scores (e.g. 8 or 9), the model is extrapolating, and produces artifacts that lead to that face being rejected. This speculation requires further analysis to be confirmed.

Interpreting Transformation Maps

The interpretability of the model may be useful in the field of social psychology, giving researchers new suggestions about what features of a face are most important for perceiving a given trait. It can also elegantly summarize the average opinions and biases of a group of raters who have created a dataset, or serve as a visual heuristic for understanding which traits are most similar to each other in human perception.

Table 4: Model size for learning seven traits

Model	CycleGAN	ICGAN	StarGAN	**ModifAE**
Parameters	736M	68M	53M	**1M**

The "intelligent" transformation map appears to show that bigger heads are rated as more intelligent (at least, pictures in which the head appears larger or closer). This suggests a bias that to our knowledge, has not been previously observed. Of course, in this case, it is simply faces that subtend a larger visual angle, rather than real-world head size. In further experiments, the head size should be normalized across images to avoid this potential bias. In addition, experiments could be run where image head size is systematically manipulated with the same face (judged by different subjects), to verify the bias.

The "intelligent" transformation map appears to show that bigger heads are rated as more intelligent (at least, pictures in which the head appears larger or closer). This suggests a bias that to our knowledge, has not been previously observed. Of course, in this case, it is simply faces that subtend a larger visual angle, rather than real-world head size. In further experiments, the head size should be normalized across images to avoid this potential bias. In addition, experiments where humans rate images with systematically manipulated head size could be run to verify the bias.

Conclusion

In this paper, we propose ModifAE: a single network autoencoder, which performs continuous image modification on subjective face traits in an interpretable manner. ModifAE does not require training multiple networks or designing hand-tailored features for image modification. Instead, a single network is trained to autoencode an image and its traits through the same latent space, implicitly learning to make meaningful changes to images based on trait values. Our experiments show that ModifAE requires fewer parameters and takes less training time than existing general methods. It also provides interpretable transformation maps of traits which demonstrably highlight biases in datasets and salient features in human perception of traits. Additionally, in this work, we compute and verify novel continuous subjective trait ratings for CelebA faces. Finally, we demonstrate that ModifAE makes more meaningful continuous image traversals than an equivalent SOTA method (Choi et al., 2017) and examine human agreement with ModifAE modifications in the subjective face trait space.

References

Bainbridge, W. A., Isola, P., & Oliva, A. (2013a). The intrinsic memorability of face photographs. *Journal of Experimental Psychology: General*, *142*(4), 1323.

Bainbridge, W. A., Isola, P., & Oliva, A. (2013b). The intrinsic memorability of face photographs. *Journal of Experimental Psychology: General*, *142*(4), 1323.

Brock, A., Donahue, J., & Simonyan, K. (2018). Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*.

Choi, Y., Choi, M., Kim, M., Ha, J., Kim, S., & Choo, J. (2017). Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. *CoRR*, *abs/1711.09020*.

Dumas, R., & Testé, B. (2006). The influence of criminal facial stereotypes on juridic judgments. *Swiss Journal of Psychology*, *65*(4), 237–244.

Eisenthal, Y., Dror, G., & Ruppin, E. (2006). Facial attractiveness: Beauty and the machine. *Neural Computation*, *18*(1), 119–142.

Falvello, V., Vinson, M., Ferrari, C., & Todorov, A. (2015). The robustness of learning about the trustworthiness of other people. *Social Cognition*, *33*(5), 368.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems 27: Annual conference on neural information processing systems 2014, december 8-13 2014, montreal, quebec, canada* (pp. 2672–2680).

Isola, P., Zhu, J., Zhou, T., & Efros, A. A. (2016). Image-to-image translation with conditional adversarial networks. *CoRR*, *abs/1611.07004*.

Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2017). Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*.

Khosla, A., Bainbridge, W. A., Torralba, A., & Oliva, A. (2013). Modifying the memorability of face photographs. In *International Conference on Computer Vision (ICCV-2013)* (pp. 3200–3207).

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *CoRR*, *abs/1412.6980*.

Lee, M., & Seok, J. (2017). Controllable generative adversarial network. *CoRR*, *abs/1708.00598*.

Leyvand, T., Cohen-Or, D., Dror, G., & Lischinski, D. (2008). Data-driven enhancement of facial attractiveness. *ACM Transactions on Graphics (TOG)*, *27*(3), 38.

Liu, Z., Luo, P., Wang, X., & Tang, X. (2015). Deep learning face attributes in the wild. In *Proceedings of international conference on computer vision (iccv).*

Mirza, M., & Osindero, S. (2014). Conditional generative adversarial nets. *CoRR*, *abs/1411.1784*.

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences*, *105*(32), 11087–11092.

Perarnau, G., van de Weijer, J., Raducanu, B., & Álvarez, J. M. (2016). Invertible conditional gans for image editing. *CoRR*, *abs/1611.06355*.

Song, A., Li, L., Atalla, C., & Cottrell, G. (2017). Learning to see people like people: Predicting social perceptions of faces. In *Proceedings of the 39th annual meeting of the cognitive science society, cogsci 2017, london, uk, 16-29 july 2017.*

Todorov, A., Olivola, C. Y., Dotsch, R., & Mende-Siedlecki, P. (2015). Social attributions from faces: Determinants, consequences, accuracy, and functional significance. *Annual Reviews of Psychology*, *66*(1), 519.

Vernon, R. J., Sutherland, C. A., Young, A. W., & Hartley, T. (2014). Modeling first impressions from highly variable facial images. *Proceedings of the National Academy of Sciences*, *111*(32), E3353–E3361.

Zhu, J., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE international conference on computer vision, ICCV 2017, venice, italy, october 22-29, 2017* (pp. 2242–2251). doi: 10.1109/ICCV.2017.244

Comparing Gated and Simple Recurrent Neural Network Architectures as Models of Human Sentence Processing

Christoph Aurnhammer (aurnhammer@coli.uni-saarland.de)
Department of Language Science and Technology, Saarland University, Campus C1, 66123 Saarbrücken, Germany
Centre for Language Studies, Radboud University, Erasmusplein 1, 6525 HT Nijmegen, The Netherlands

Stefan L. Frank (s.frank@let.ru.nl)
Centre for Language Studies, Radboud University, Erasmusplein 1, 6525 HT, Nijmegen, The Netherlands

Abstract

The Simple Recurrent Network (SRN) has a long tradition in cognitive models of language processing. More recently, gated recurrent networks have been proposed that often outperform the SRN on natural language processing tasks. Here, we investigate whether two types of gated networks perform better as cognitive models of sentence reading than SRNs, beyond their advantage as language models. This will reveal whether the filtering mechanism implemented in gated networks corresponds to an aspect of human sentence processing. We train a series of language models differing only in the cell types of their recurrent layers. We then compute word surprisal values for stimuli used in self-paced reading, eye-tracking, and electroencephalography experiments, and quantify the surprisal values' fit to experimental measures that indicate human sentence reading effort. While the gated networks provide better language models, they do not outperform their SRN counterpart as cognitive models when language model quality is equal across network types. Our results suggest that the different architectures are equally valid as models of human sentence processing.

Keywords: Surprisal; Gated Recurrent Neural Networks; Language Modeling; Sentence Processing; Sentence Reading; Self-paced Reading; Eye-tracking; Electroencephalography

Introduction

In psycholinguistics, the Simple Recurrent Network (SRN; Elman, 1990) has been a popular (and reasonably successful) neural architecture for modeling aspects of human sentence processing, and it remains so to this day (Brouwer, Crocker, Venhuizen, & Hoeks, 2017; Frank, Otten, Galli, & Vigliocco, 2015; Rabovsky, Hansen, & McClelland, 2018; Twomey, Chang, & Ambridge, 2014, to name just a few recent examples). However, it has been known since the late 1990s that the SRN struggles to integrate information over many classification steps, due to what is referred to as the vanishing gradient problem (Hochreiter, 1998).

This problem was addressed by neural network models containing recurrent units that have gates with trained weights, such as the Gated Recurrent Unit (GRU; Bahdanau, Cho, & Bengio, 2015) and the Long Short-Term Memory (LSTM; Hochreiter & Schmidhuber, 1997) network. The gating mechanism implemented in GRUs and LSTMs controls the flow of information in the recurrent cell, allowing the cells to memorise information over time, forget it when adequate, and to determine the weighting of old and new input. While the principles of the two architectures are similar, the GRU can be regarded as a more lightweight variation on the LSTM, making use of only two gates and a single hidden state, whereas the LSTM architecture provides three gates and introduces an additional memory state.

Gated networks outperform SRNs on several NLP tasks. For example, LSTMs perform more accurately than SRNs on number agreement (Linzen, Dupoux, & Goldberg, 2016) and conversational speech recognition (Xiong et al., 2017). In the current study, we investigate how well gated networks perform as cognitive models of human sentence processing compared to the traditional SRN. We model human word-level processing effort by using recurrent neural networks as probabilistic language models that estimate the predictability of words in context.

For the language modeling problem, the ability to make effective use of more of the words in the prior sequence can be expected to pose a crucial advantage of a gated recurrent network compared to the SRN. For instance, the processing of long-term dependencies has been proposed as one aspect of natural language processing addressed more adequately by gated networks than by SRNs (Bahdanau et al., 2015). Because gated networks are designed for long-distance encoding, they may also be superior cognitive models: The filtering mechanism implemented by the gates may mirror an aspect of human sentence processing. For example, it is known that humans read the word *or* faster when they processed the word *either* in the prior sequence of words, demonstrating their ability to remember dependencies between words across long spans (Staub & Clifton Jr., 2006). Gated networks may reflect this human behaviour more accurately than SRNs by assigning lower surprisal to the word *or* even when the corresponding *either* is distant.

Although LSTMs and GRUs have already been applied to account for human language performance measures (Futrell et al., 2019; Goodkind & Bicknell, 2018; Gulordava, Bojanowski, Grave, Linzen, & Baroni, 2018; Hahn & Keller, 2016; McCoy, Frank, & Linzen, 2018; Sakaguchi, Duh, Post, & Durme, 2017; Van Schijndel & Linzen, 2018a, 2018b), the question remains whether they form more accurate cognitive processing models than traditional SRNs, beyond what might be expected from their stronger language modeling abilities.

In the current study, we directly compare three recurrent neural network (RNN) language model architectures (SRNs, GRUs, and LSTMs) on their ability to predict human reading data collected in self-paced reading, eye-tracking, and electroencephalography experiments. If the mechanisms imple-

mented in GRUs and LSTMs correspond to cognitive mechanisms applied during sentence comprehension, we would expect predictions by these models to fit human reading data more closely than predictions by SRNs, over and above any advantage that GRUs/LSTMs might have because of their superiority as language models. Conversely, if the cognitive system does not apply anything like a gating mechanism, the SRN may simulate human language processing more closely than GRUs and LSTMs do. In that case, the SRN may predict human processing data more accurately than gated RNNs that are matched for language model quality.

Method[1]

To determine whether or not LSTMs and GRUs outperform SRNs as cognitive models of sentence processing, we train three different kinds of RNN language models, each using one of the three recurrent cell types. We evaluate the models by assessing the predictive power of the surprisal values they assign to stimuli used in three experiments of humans sentence reading.

Human processing data

We assess how well each RNN language model's word surprisal values predict human cognitive processing effort during sentence reading, as measured in self-paced reading (SPR), eye-tracking (ET), and electroencephalography (EEG) experiments. The SPR and ET data come from Frank, Monsalve, Thompson, and Vigliocco (2013) and the EEG data from Frank et al. (2015).

In all three experiments, participants read English sentences sampled from unpublished novels. All sentences are understandable out of their context in the novels. A subset of the sentences were used in the ET and EEG experiments; these were the shortest sentences (maximum length: 15 words) of those from the SPR study. Table 1 displays the numbers of participants and stimuli, along with ranges and means of sentence length for each of the three data sets. Importantly, we make sure that all word types in the stimuli are attested for in the training data, meaning that the language models do not encounter words for the first time when applied to the stimuli.

For this study, we select a single variable from each dataset that is indicative of human processing cost: Reading time (RT) from the SPR data, gaze duration (a.k.a. first-pass reading time) from the ET data, and N400 size from the EEG data set. We follow the insight that reading times reflect the cognitive effort the reader needs to employ during language processing (Levy, 2008). Reflecting this idea, the N400 event-related potential amplitude indicates processing effort on lexico-semantic levels (Kutas, Van Petten, & Kluender, 2006; Kutas & Federmeier, 2011). Earlier research has already demonstrated that these dependent variables, from these particular data sets, indeed correlate with word surprisal values (SPR: Monsalve, Frank, & Vigliocco, 2012; ET: Frank & Thompson, 2012; EEG: Frank et al., 2015).

Network architectures

Our RNN architecture consists of a 400-unit word embedding layer, a 500-unit recurrent layer, a 400-unit feed-forward layer with tanh activation function, and a final layer with log-softmax activation function, which maps to the vocabulary. We do not use pre-trained word embeddings. Rather, the weights of the embedding layer that transforms the vocabulary items to real-valued word vectors are learned during the next-word prediction task, along with the rest of the network weights. The model architectures only differ in that their recurrent layers use either SRN, LSTM, or GRU cells.

Training corpus

As training data for the language models we use section 13 of the English version of the Corpora from the Web (COW, 2014 version; Schäfer, 2015). This corpus consists of randomly ordered sentences collected from web pages. From this section, the 10,000 most frequent word types are selected as our model's vocabulary. One hundred and three word types that appear in the experimental stimuli (see Section on *Human Processing Data*) but are not yet covered in the vocabulary are added, resulting in a final vocabulary size of 10,103 word types. After determining the vocabulary, we select those sentences from the initial COW section that contain only in-vocabulary word types, thus also covering the low-frequency words in the experimental stimuli. We follow this strategy to avoid having to use a (cognitively implausible) UNKNOWN-type. Furthermore, we only keep sentences with a maximum length of 39 words, which corresponds to the longest sentence in the experimental stimuli (not counting punctuation as words). We remove a small number of sentences to arrive at a final selection that contains 6,470,000 training sentences and consists of 94,422,754 tokens in total.

Although this training set and vocabulary size is relatively small by current standards, note that our aim here is not to construct the best possible language model, and not even to provide the most accurate account of human sentence processing effort. Rather, we investigate whether RNN architectures differ in their ability to predict human data.

Network training

We train the networks on one sentence at a time to let model training resemble human language processing and acquisition. Further, we reset the hidden state of the recurrent cells to zero for each new sentence. From the network's log-probability output at each step, the loss function computes the negative log-likelihood. Based on this loss, we optimise the network weights using stochastic gradient descent with momentum (0.9) and an initial learning rate of 0.0025. After each third of the training data, we reduce the learning rate to half of its prior value. As precaution to the exploding gradient problem (Bengio, Simard, & Frasconi, 1994), we clip gradients at 0.25. The error is always back-propagated through the

[1] All code and data is available at `https://github.com/caurnhammer/AurnhammerFrank_CogSci2019`

Table 1: Numbers of participants, number of sentences, range of sentence length, mean sentence length, number of word tokens, and number of data points (after exclusion; see Section *Stage 1: Predicting human data from surprisal*) in the human sentence reading data sets. In the SPR experiment, each participant received a random subset of the 361 possible sentences (see Frank et al., 2013, for details).

Exp.	Part.	Sent.	Range sent. len.	Mean sent. len.	Tokens	Data points
SPR	54	361	5–39	14.1	4957	132,858
ET	35	205	5–15	9.4	1931	28,970
EEG	24	205	5–15	9.4	1931	24,618

entire sentence.

To account for random variation in model performance that is solely due the initial weights and training sentence presentation order, we train each RNN type six times, each time with different random initial weights (uniformly distributed between ± 0.1; with initial biases 0) and a different random order of sentence presentation. However, for each training repetition, the same initial weights (for connections that correspond between architectures) and the same presentation orders are applied across the three recurrent cell architectures. Hence, the *only* difference between the RNN types is in the architectures of their recurrent cells.

Language model evaluation

We evaluate the performance at the nine different training corpus sizes by computing the perplexity on the unseen experimental stimulus sentences. Perplexity is computed as

$$PPL = e^{-|W|^{-1} \sum_{w \in W} \log P(w)},$$

where $|W|$ is the number of word tokens in the experimental sentences. Lower perplexity results from language models that assign higher probabilities to the test data. Perplexity thus expresses the extent to which a language model captures the statistical structures of the data that are useful to predicting the next word, irrespective of the extent to which this is helpful for explaining human sentence processing measures.

Statistical model evaluation

The RNN models' ability to account for the human processing data is evaluated in two stages, as explained in more detail below. First, we compute surprisal for the experimental test items. Surprisal is computed as

$$\text{surprisal}(w_t) = -\log\ P(w_t | w_1, ..., w_{t-1}).$$

and formalises the extent to which occurrence of a word w_t is unexpected, given a sequence of preceding words $w_1, ..., w_{t-1}$ (Hale, 2001; Levy, 2008). The reading-time and N400 measures on each word are regressed on each model's surprisal estimates resulting in a collection of goodness-of-fit measures. Next, we assess the relation between each RNN type's goodness-of-fit and its quality as a language model.

Stage 1: Predicting human data from surprisal Each individual RNN generates surprisal estimates for each word of the 361 stimuli sentences. The surprisal values are obtained after training the network on 1K, 3K, 10K, 30K, 100K, 300K, 1M, 3M, and all 6.47M sentences. This procedure allows to observe how the goodness-of-fit to human data develops as a function of language model quality, which steadily increases with the amount of observed training data. In summary, we have 9 (points during training) $\times$ 6 (training repetitions) $\times$ 3 (RNN types) $=$ 162 sets of surprisal values to compare to the SPR times, gaze durations, and N400 sizes.

The predictive power of each set of surprisal values is assessed by means of linear mixed effects regression, using the `MixedModels` package[2] (v0.18.1) for `Julia` (Bezanson, Edelman, Karpinski, & Shah, 2017). First, a baseline model was fitted to each of the three human data sets. The aim of this baseline is to factor out the effects of the most imporant variables known to affect reading times and N400 sizes and thus be left with an effect of surprisal that is as isolated as possible.

The dependent variables self-paced reading times and gaze durations are log-transformed. In the EEG data, N400 size is analysed as defined by Frank et al. (2015): the average potential on central-parietal electrodes over a 300–500ms window after word onset.

The baseline models include as fixed effects: log-transformed word frequency in the training corpus, word length (number of characters) and word position in the sentence. For the SPR and ET data, we also enter the previous word's frequency and length into the analysis to account for spillover effects that are known to affect reading times (Rayner, 1998). Moreover, we add previous-word RT (log-transformed) to the SPR analysis to address the high correlation between consecutive word RTs that typically occurs in the SPR paradigm; and to the ET analysis we add a binary factor indicating whether the previous word was fixated. For the EEG analysis, we enter baseline activity (i.e., the average electrode potential in the 100ms leading up to word onset) into the regression. All interactions between the fixed effects are also included. Furthermore, there are by-subject and by-item (word token) random intercepts and by-subject random slopes of all fixed-effect predictors.

We exclude data on sentence-initial and -final words, words attached to a comma, and clitics. Furthermore, participants are removed from the analysis if they are not native English

[2] `github.com/dmbates/MixedModels.jl`

speakers or scored less than 80% correct on the yes/no comprehension questions that were presented for approximately half the sentence stimuli. In addition, SPR and ET data points are removed on words directly following a comma or clitic, and when reading times are below 50ms or over 3500ms. For the EEG data, we exclude artefacts as identified by Frank et al. (2015).

The goodness-of-fit of each set of surprisal values for each human data set equals the log-likelihood ratio (decrease in regression model deviance) between the baseline and a regression model that additionally includes surprisal as both a fixed effect and by-subject random slope. For the SPR and ET analyses, the previous word's surprisal is also added (again as fixed and random effects) in order to capture spillover effects. The resulting values are χ^2-statistics, with 2 degrees of freedom for the EEG data and 4 degrees of freedom for the two reading-time data sets. We further add a negative sign to the χ^2-statistics to indicate effects in the negative-going direction, that is, when higher surprisal results in shorter reading times or smaller (less negative) N400 size.

Stage 2: Predicting goodness-of-fit from language model accuracy Networks that form better language models tend to estimate surprisal values that fit human data better (Frank et al., 2015; Goodkind & Bicknell, 2018). In analysis Stage 2, we are interested in ascertaining whether the relation between language model accuracy and goodness-of-fit to human data differs between network architectures.

We quantify language model accuracy as the average log-probability (i.e., negative average surprisal) estimated over the experimental sentences, weighted by the number of times each word token takes part in the analysis described above, that is, for how many participants the data on this word was not excluded. Following this, we fit Generalized Additive Mixed Models (GAMMs), for each of the three RNN types and human data sets separately, to predict the goodness-of-fit measures (from analysis Stage 1) from the language model accuracies, with network training repetition as a random effect. This is done using the `R` package `mgcv` (Wood, 2004).

Results

Language modelling results

Figure 1 reports on the perplexities of the 18 individual language models at 9 different points during training. While the SRNs set in at lower perplexity than the gated networks early in training, the latter ultimately outperform the simple RNNs. Language model performance steadily increases throughout training but a saturation of the language model performance seems only to commence at the final training steps.

Statistial modelling results

Figure 2 displays the goodness-of-fit measures from analysis Stage 1 for each human data set, as well as the fitted curves relating goodness-of-fit to language model accuracy from analysis Stage 2. These plots clearly show that well-trained language models estimate surprisal values that account for reading times and N400 size, and that the goodness-of-fit generally improves as the language models more accurately capture the linguistic patterns. Interestingly, for lower levels of linguistic accuracy, corresponding to models trained on relatively few sentences, the effect of surprisal on gaze duration size is reversed, in that higher surprisal correlates with faster reading. The cause of this reversal remains to be identified.

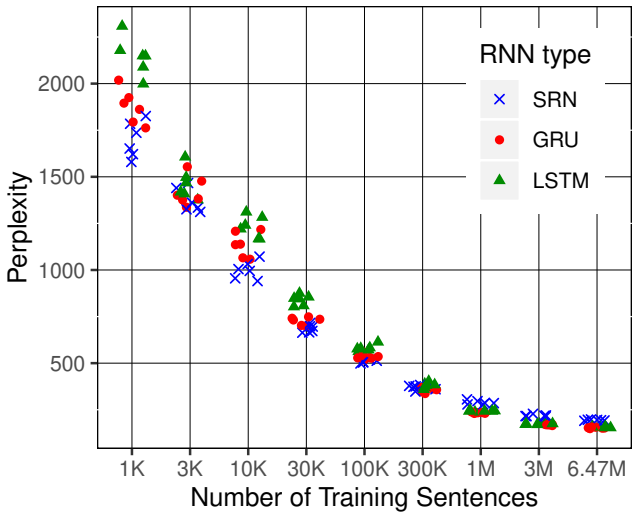


Figure 1: Perplexity on the experimental sentences for each of three RNN types at nine different training corpus sizes. At each training size there are six models of each type with different sentence orderings and initial weights. Data points are subjected to horizontal jitter to improve readability.

The gated RNN models reach higher levels of language model accuracy than the SRNs, which is why they can also outperform SRNs in terms of goodness-of-fit. For similar levels of language model accuracy, however, the three model types account for similar quantities of variance in the human processing data, as is evident from the largely overlapping confidence intervals of the fitted GAMM curves.

This does not imply that different network types make no independent contributions to human data prediction. To test whether the models differ qualitatively in that one RNN explains unique variance over and above the others, we average the surprisal values over the six fully trained versions of each network architecture. Next, we fit linear mixed models including the surprisals from two of the three RNN types and then test whether that regression model fits the data better than a regression with only a single set of surprisal values. That is, for each pair of RNN types we ask whether one explains human data over and above the other.

Table 2 shows model comparisons, testing for the significance of adding the surprisal from the models displayed in rows to the models in columns. The comparisons reveal statistically significant effects of GRU and LSTM surprisal over and above SRN surprisal in all three data sets. For the EEG data, SRN surprisal also explains variance not yet explained

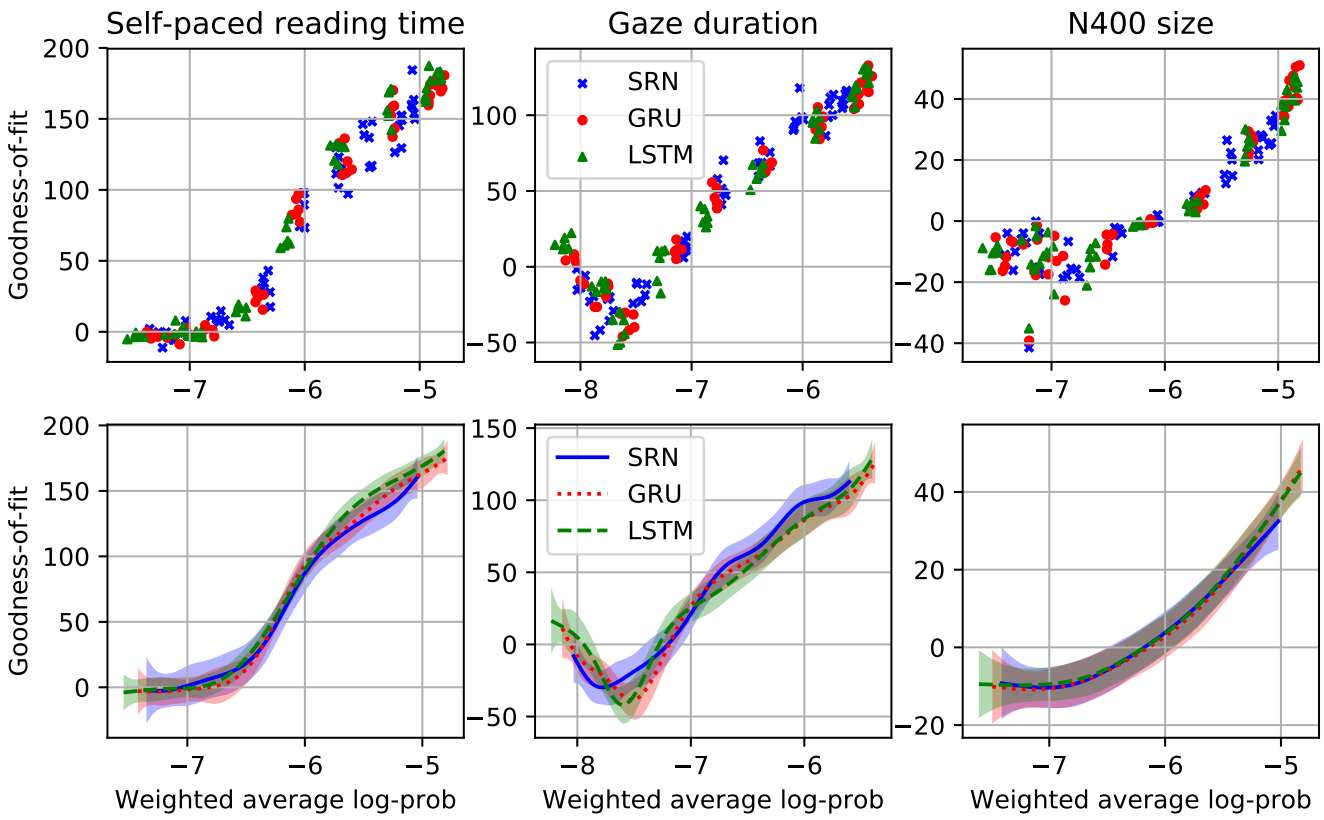


Figure 2: Top row: results from analysis Stage 1. The goodness-of-fit of surprisal to human data is plotted as a function of language model accuracy. Bottom row: results from analysis Stage 2. Plotted are the fitted GAMM curves relating goodness-of-fit to language model accuracy. Shaded areas indicate 95% confidence intervals. Panels on the left, middle, and right side are for SPR, ET, and EEG data, respectively.

by the gated networks.

Discussion

Our comparison of the abilities of SRNs, GRUs, and LSTMs to predict human reading-time and N400 measures (via the networks' word-surprisal estimates) do not reveal any large or reliable difference between the three RNN types, at least, not as long as the different networks' accuracies as language models do not differ. The two gated networks do form better language models than the SRN, resulting in more precise predictions of human data at the highest levels of language model accuracy. However, if the human cognitive system would employ mechanisms akin to the gates in GRU/LSTM recurrent cells, we would expect GRU/LSTM-based surprisal to show better fit than SRN-based surprisal to the human processing data, even without any difference in language model accuracy. Our analyses do not support this conclusion.

The gated RNNs explain variance over and above what is accounted for by the SRNs on SPR, ET, and EEG data. This is an expected effect, given that gated networks form better language models. Their ability to encode relations between word tokens along larger spans is likely giving them a clear advantage in accounting for human data. More surprisingly, on the EEG data the SRNs also explain a portion of variance that is distinct from the one explained by the gated networks. This finding may suggest a potential insensitivity of the N400 ERP component to long-distance dependencies, at least to the extent that N400 size reflects word predictability. Converging evidence for this interpretation is presented by Frank et al. (2015) who demonstrate that an n-gram language model with a context size of three words explains variance over and above an SRN on the same data set.

Conclusion

While gated recurrent neural networks provide better language models than simple recurrent networks, our investigations do not indicate that they have any substantial or reliable advantage as cognitive models of sentence reading, in addition to what is expected from their superior language modeling abilities. Nevertheless, gated networks consistently reached higher linguistic accuracy. This fact alone makes the use of gated RNN advisable not only from a language modeling point of view but also for psycholinguistics (and cognitive science more in general) when as much variance in human data as possible needs to be explained, for example when surprisal is used as a covariate in studies that aim to find a unique

Table 2: Results from regression model comparisons between RNN types. Each χ^2-statistic is the outcome of a log-likelihood ratio test for whether the network type in the table row accounts for variance in the human data over and above the network type in the table column. Asterisks indicate statistical significance level after multiple-comparison correction (Benjamini & Hochberg, 1995): * = $p < .05$; ** = $p < .01$; *** = $p < .001$.

Exp.		SRN	GRU	LSTM
SPR	SRN		$\chi^2(4) = 3.20$	$\chi^2(4) = 3.69$
	GRU	$\chi^2(4) = 12.7$*		$\chi^2(4) = 1.29$
	LSTM	$\chi^2(4) = 18.1$**	$\chi^2(4) = 6.22$	
ET	SRN		$\chi^2(4) = 6.18$	$\chi^2(4) = 8.70$
	GRU	$\chi^2(4) = 15.6$*		$\chi^2(4) = 0.46$
	LSTM	$\chi^2(4) = 22.5$***	$\chi^2(4) = 4.88$	
EEG	SRN		$\chi^2(2) = 10.9$*	$\chi^2(2) = 8.65$*
	GRU	$\chi^2(2) = 26.0$***		$\chi^2(2) = 3.26$
	LSTM	$\chi^2(2) = 21.7$***	$\chi^2(2) = 1.22$	

effects of some additional predictor.

Acknowledgments

The work presented here was funded by SFB 1102 "Information density and linguistic encoding", awarded by the German Research Foundation (DFG); and by the Netherlands Organisation for Scientific Research (NWO) Gravitation Grant 024.001.006 to the Language in Interaction Consortium.

References

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations.* San Diego, CA: ICLR.

Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, *5*(2), 157–166.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, *57*, 289–300.

Bezanson, J., Edelman, A., Karpinski, S., & Shah, V. B. (2017). Julia: A fresh approach to numerical computing. *SIAM Review*, *59*, 65–98.

Brouwer, H., Crocker, M. W., Venhuizen, N. J., & Hoeks, J. C. J. (2017). A neurocomputational model of the N400 and the P600 in language processing. *Cognitive Science*, *41*(S6), 1318-1352.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211.

Frank, S. L., Monsalve, I. F., Thompson, R. L., & Vigliocco, G. (2013). Reading time data for evaluating broad-coverage models of English sentence processing. *Behavior Research Methods*, *45*(4), 1182–1190.

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, *140*, 1–11.

Frank, S. L., & Thompson, R. L. (2012). Early effects of word surprisal on pupil size during reading. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 1554–1559). Austin, TX: Cognitive Science Society.

Futrell, R., Wilcox, E., Morita, T., Qian, P., Ballesteros, M., & Levy, R. (2019). Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of the 18th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.*

Goodkind, A., & Bicknell, K. (2018). Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics* (pp. 10–18). Salt Lake City, UT: CMCL.

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., & Baroni, M. (2018). Colorless green recurrent networks dream hierarchically. In *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* New Orleans, Louisiana: Association for Computational Linguistics.

Hahn, M., & Keller, F. (2016). Modeling human reading with neural attention. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 85–95). Austin TX: Association for Computational Linguistics.

Hale, J. T. (2001). A probabilistic Early parser as a psycholinguistic model. In *Proceedings of the second conference of the North American chapter of the Association for Computational Linguistics* (Vol. 2, pp. 159–166). Pittsburgh, PA: Association for Computational Linguistics.

Hochreiter, S. (1998). The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *6*(02), 107–116.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology*, *62*, 621–647.

Kutas, M., Van Petten, C. K., & Kluender, R. (2006). Psycholinguistics electrified II (1994–2005). In *Handbook of psycholinguistics (second edition)* (pp. 659–724). Elsevier.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.

Linzen, T., Dupoux, E., & Goldberg, Y. (2016). Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, *4*, 521–535.

McCoy, R. T., Frank, R., & Linzen, T. (2018). Revisiting the poverty of the stimulus: hierarchical generalization without a hierarchical bias in recurrent neural networks. In *Proceeding of the 40th Annual Conference of the Cognitive Science Society* (pp. 2093–2098). Madison, WI: Cognitive Science Society.

Monsalve, I. F., Frank, S. L., & Vigliocco, G. (2012). Lexical surprisal as a general predictor of reading time. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 398–408).

Rabovsky, M., Hansen, S. S., & McClelland, J. L. (2018). Modelling the N400 brain potential as change in a probabilistic representation of meaning. *Nature Human Behaviour*, *2*, 693–705.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*(3), 372–422.

Sakaguchi, K., Duh, K., Post, M., & Durme, B. V. (2017). Robsut wrod reocginiton via semi-character recurrent neural network. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence* (pp. 3281–3287). San Francisco, CA: AAAI.

Schäfer, R. (2015). Processing and querying large web corpora with the COW14 architecture. In P. Bański, H. Biber, E. Breiteneder, M. Kupietz, H. Lüngen, & A. Witt (Eds.), *Proceedings of the 3rd Workshop on the Challenges in the Management of Large Corpora* (pp. 28–34). Mannheim, Germany: Institut für Deutsche Sprache.

Staub, A., & Clifton Jr., C. (2006). Syntactic prediction in language comprehension: Evidence from either...or. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(2), 425–436.

Twomey, K. E., Chang, F., & Ambridge, B. (2014). Do as I say, not as I do: A lexical distributional account of English locative verb class acquisition. *Cognitive Psychology*, *73*, 41–71.

Van Schijndel, M., & Linzen, T. (2018a). Modeling garden path effects without explicit hierarchical syntax. In T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 2603–2608). Austin, TX: Cognitive Science Society.

Van Schijndel, M., & Linzen, T. (2018b). A neural model of adaptation in reading. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 4704–4710). Brussels: Association for Computational Linguistics.

Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, *99*(467), 673-686.

Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., ... Zweig, G. (2017). Toward human parity in conversational speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2410–2423.

(In-)definites, (anti-)uniqueness, and uniqueness expectations

Nadine Bade (nadine.bade@uni-tuebingen.de)
Collaborative Research Center 833, Nauklerstrasse 35
72074 Tuebingen, Germany

Florian Schwarz (florians@sas.upenn.edu)
Department of Linguistics, 3401-C Walnut St.
Philadelphia, PA 53706 USA

Abstract

Using "A" in noun phrases such as "A father of the victim" is odd, which is commonly explained by the principle *Maximize Presupposition*, requiring speakers to use the alternative with the strongest presupposition (here "The", given its uniqueness presupposition). This results in an anti-uniqueness inference for "A" (clashing with stereotypical expectations here), sometimes labelled as an 'anti-presupposition' (Percus, 2006), as it derives from reasoning over the presuppositions of alternative forms. We compare these inferences to the uniqueness inferences associated with definites, while manipulating uniqueness expectations in a picture manipulation task using visual world eye-tracking. This offers a minimal comparison of uniqueness-based inferences that are lexically encoded vs. pragmatically inferred, and furthermore tests the prediction that the accommodatability of the definite's presupposition plays a role in the derivation of anti-uniqueness inferences (Rouillard & Schwarz, 2017).

Keywords: presuppositions; visual world eye-tracking; definiteness; indefiniteness

Theoretical Background

There is a concurrent claim in the theoretical literature that definite descriptions, and presupposition triggers in general, have to be used if their presupposition (PSP) is fulfilled in the context. Since definite noun phrases come with a presupposition of uniqueness they must be used if this uniqueness presupposition is met in the context, see (1-a). The use of an indefinite noun phrase in (1-b) is infelicitous.

(1) a. The father of the victim came.
b. #A father of the victim came.

There are various theories explaining this effect by assuming that there is lexical competition between the presuppositionally stronger definite and presuppositionally weaker indefinite governed by the principle *Maximize Presupposition* (Heim, 1991; Percus, 2006; Sauerland, 2008; Chemla, 2008). Based on pragmatic reasoning over the stronger alternative – similar to the one used for the derivation of scalar implicatures – the indefinite yields the inference that the presupposition of the definite is false ('anti-uniqueness'). (1-b), for example, has the infelicitous anti-uniqueness inference that there is not exactly one father of the victim. The resulting inferences ('anti-presuppositions') are theoretically set apart from well-studied components of meaning like presuppositions and implicatures based on their weaker epistemic status, and their projection behaviour. Recently, the strength of the epistemic status has been argued to be dependent on accommodatability of the competing presuppositional statement, which is tied to the knowledge state of speaker and hearer (Rouillard & Schwarz, 2017). There is a less prominent alternative view according to which definites and indefinite both come with their own context restrictions, i.e. that the indefinite comes with a novelty condition (Heim, 1983) or its own presupposition of anti-uniqueness (Kratzer, 2005). These make different predictions for the processing profiles associated with anti-uniqueness.

According to the first view (theory A), in which anti-uniqueness is the result of reasoning over presuppositionally stronger alternatives, the consideration of the alternative and its subsequent negation to derive the inference should require extra processing costs, based on the observation that both presuppositions and negation have been independently shown to increase the cognitive load in processing (Schwarz, 2007; Tiemann, 2014; Kirsten et al., 2014; Carpenter, Just, Keller, Eddy, & Thulborn, 1999; Reichle, Carpenter, & Just, 2000; Herbert & Kübler, 2011). As this view works with alternatives it also predicts that drawing the inference might depend on the salience and accommodability of the competitor with "the". In the example in (1) this alternative is very salient and easy to accommodate as it is common knowledge that people have one (biological) father. However, in other examples where this is not the case the inference is not as strong. (2), for example, does not seem to have any implications as to how many pathologically noisy neighbours the speaker has, at least without any further knowledge about the likelihood about it being one or more.

(2) A pathologically noisy neighbour of mine broke into the attic. (Heim, 1991)

According to the second view (theory B), definites and

indefinites should have comparable processing costs, with minimal or no differences in processing patterns as both introduce their own restrictions that are part of their lexically endoed meaning. As a result, the salience of the definite as an alternative should not play a role in deriving any anti-uniqueness inferences.

Albeit the fact that inferences resulting from *Maximize Presupposition* and the principle itself have received a lot of attention in the theoretical literature, there is almost no experimental research on the topic, with few exceptions (Amsili, 2015; Eckardt, 2014; Bade, 2016). There is, however, some experimental research on definiteness versus indefiniteness discussed in the next section.

Previous experimental work

There have been previous experimental investigations of the difference between indefinite and definite determiners. One line of research which is of importance for the present discussion is the study of so called "bridging inferences" (Haviland & Clark, 1974). They describes the inference of unique entities in certain situations where such a referent is stereo-typically unique, e.g. "the bus driver" in a situation where someone is entering the bus. These inferences are associated with different processing costs depending on how easily the referents are accessible (Haviland & Clark, 1974). (Burkhardt, 2006) in an ERP study finds similar effects for both definites and indefinites if they follow contexts that do not explicitly mention the referent.

Many studies on the definite have found additional processing costs if the context does not furnish an appropriate (unique) referent (Tiemann et al., 2011; Kirsten et al., 2014; Tiemann, 2014). Additional processing costs have also been found for the indefinite, which has been attributed to it introducing a new discourse referent (Kirsten et al., 2014; Schumacher, 2009).

A set of studies which is of special interest for our discussion tested the use of definites versus indefinites for stereo-typically unique items in context in which they are typically unique (e.g. stove in a kitchen) or not (e.g. stove in an appliance store) (Clifton, 2013). Clifton found that interactions between contexts and determiner in reading times only emerged if the experiment involved a secondary arithmetic task, which is argued to lead to deeper processing resulting in participants forming a more complete situation model. This model required the accommodation of a unique referent in the mismatching condition of the definite, and the introduction of a new referent in the mismatching condition of the indefinite.

Experiment

The aim of the experiment was to test whether the potential theoretical distinction between (anti-uniqueness-) anti-presuppositions on the one hand, and (uniqueness-) presuppositions on the other hand is supported by processing measures. Additionally, we wanted to test the prediction of theory A that lexical alternatives, as well as the epistemic state of the speaker with regard to the truth of the inference, play a role in the derivation of anti-uniqueness inferences.

Design and Material

For that purpose, we created sentence materials that either contained the definite or indefinite determiner (first factor DETERMINER with two levels, +/-DEF) and combined them with either stereo-typically unique or non-unique nouns (second factor STEREO-TYPICALITY with levels +/-TYPICALLY UNIQUE, see an example in all 4 sentence conditions below.

(3) Someone spilled orange juice on. . .

a. a television +TYPICALLY UNIQUE, -DEF
b. the television +TYPICALLY UNIQUE, +DEF
c. a pillow -TYPICALLY UNIQUE, -DEF
d. the pillow -TYPICALLY UNIQUE, +DEF

. . . in the living room.

A given sentence was paired with two pictures providing settings where the referenced object was either typically unique or not, i.e. the sentences in (3-a) and (3-b) were paired with the two pictures in figures 3 and 4, and the sentences in (3-c) and (3-d) were paired with the two pictures in figures 1 and 2. As part of our task (described in more detail below) participants had to decide which of the two pictures the sentence was about, with the (anti-)uniqueness information conveyed by the respective determiners being key for the picture choice.

Figure 1: -TYPICALLY UNIQUE (*A/The pillow), target for unique Def (pic 1)*

Figure 2: -TYPICALLY UNIQUE (*A/The pillow), target for non-unique Indef (pic 2)*

Figure 3: +TYPICALLY UNIQUE (*A/The TV), target for unique Def (pic 1)*

Figure 4: +TYPICALLY UNIQUE (*A/The TV), target for non-unique Indef (pic 2)*

We created 24 sets of experimental stimuli, for counterbalanced presentation to participants of 6 items per experimental conditions. In addition,there were 24 filler items with the temporal connectives “before” and “after” (e.g. “Peter spilled coffee after doing the dishes") and 12 fillers with the quantifier “several” (e.g. “Peter spilled coffee on several chairs in the dining room").

Norming study

To determine what objects people considered to be stereo-typically unique in a given scenario, our first step was a norming study with 60 native speakers of English. They were asked to rate the typicality of uniqueness by being asked “How typical do you think it is that there is exactly one TV in a standard living room?”. We tested 48 different objects in the norming study. For the 24 critical items used in the study we took the 24 objects with the highest average rating and paired them with objects that received a very low average rating.

Main experimental task

We tested 77 native speakers of American English recruited through the SONA system of the University of Pennsylvania. They received course credit for their participation. Participants engaged in a simple game. In the main comprehension part of the experiment, they heard recordings of descriptions (containing indefinite or definite articles) of spills that happened in different rooms. They then had to try to best match the description they heard by dragging a splash representing a spilled beverage to one of two room settings, which differed in whether they contained one or two of the mentioned type of object (e.g., television or pillow). After going through a practice phase with 4 trials they went on to the 24 experimental trials (plus fillers). See Figure 5 for a screen shot of a sample trial (in a control condition, which sometimes depicted two different types of room settings).

Figure 5: Screen shot of a comprehension trial.

Picture choices, response times and eye-movements were tracked during the comprehension phase. Participants also went through a brief constrained production phase (9 trials), where they had to drag words to construct a sentence of the form above to describe a provided picture. This was intended to engage them with the task more by seeing both sides of the game, and to highlight the alternative choices between determiners in relation to the number of relevant objects in the picture. The determiner choices they were given were “A”, “The” and “Several”. The pictures were created so that each of these determiner would be chosen 3 times (given the theoretical assumptions, e.g. with the definite used for unique items and the indefinite for non-unique ones). There were two practice trials for the production task. A screen shot of a production trial is given in Figure 6 below. Production and comprehension block order was counterbalanced across participants.

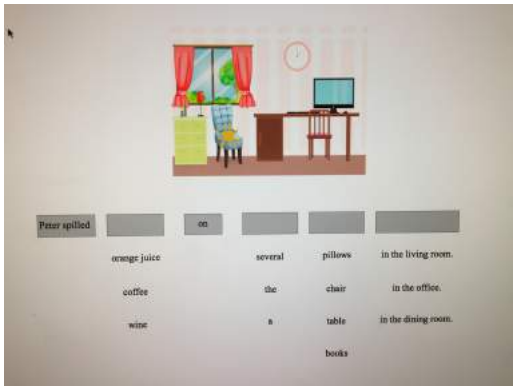


Figure 6: Screen shot of a production trial.

Predictions

Given the theoretical assumptions spelled out above we coded the pictures with two objects matching the description in the relevant noun phrase to be the target picture for the indefinite (pic 2), see again Figure 2 and 4 above. For the definite, we assumed the picture depicting exactly one object of the kind referred to in the sentence to be the target (pic 1), see Figure 1 and 3. Drawing the anti-uniqueness inference which is the basis for the target choice of the indefinite was predicted to be influenced by whether the referenced object is typically unique in the given setting by theory A. This theory predicts target choices to be higher for the indefinite in the non-typically unique condition than in the typically unique condition, whereas the definite should not be affected by this factor. As a result, we expected there to be an interaction between DETERMINER and STEREOTYPICALITY for picture choices. Moreover, based on theory A, anti-uniqueness anti-presuppositions are expected (or at least likely) to show a different pattern in eye-movements than the uniqueness presuppositions evoked by the definite. Again, since drawing the infer-

ence for the indefinite should be facilitated by the object being typically unique, we predict an interaction between DET and STEREO-TYPICALITY for measures reflecting looks to target, as well as a main effect of both factors. No such differences or interactions are predicted by theory B, according to which determiners should be more or less equally affected by stereo-typicality. According to theory A, but not B, the data should also be influenced depending on whether there was exposure to the alternative. Thus order effects of the tasks should be relevant for the processing associated with the indefinite following theory A but not B.

Analysis

Responses were analyzed using logistic mixed effect models as implemented in the glmer function in R (Bates, 2005). Reaction times and fixations on target were log transformed for analysis and analyzed using linear mixed effect models and the lmer function in R (R Core Team, 2017). Participant, condition and item were treated as random factors in both model types, and random effect slopes were included as model convergence allowed.

Results

Responses The anti-uniqueness inference of indefinites is less readily available, and less robust than the uniqueness inference of definites in our data, in line with previous results. This is witnessed by (i) low production 'accuracy' for indefinites when the production block came first, see Figure 7: in the condition where the displayed picture included multiple objects matching the noun phrase description, participants only chose an indefinite about half of the time in their sentences. There is a significant main effect of block order (p<.01) with more target choices for production when it followed the comprehension block and also a significant interaction between DETERMINER and BLOCK ORDER (p<.01) with accuracy of choice being less affected by block order if the target determiner was definite than when it was indefinite.

The weak status of the indefinite is also reflected in (ii) target choice rates in the comprehension task. Overall, there was a main effect of BLOCK ORDER, present at all levels, with overall more target choices when production came first (and comprehension second). Crucially for the present point, there is a bigger susceptibility of the indefinite to plausibility effects biasing against multiple instances of stereo-typically unique objects, especially in the initial comprehension block. This is reflected in an interaction between typical uniqueness and determiner in the 'comprehension first' block (p< .01), driven by more frequent target choices for Def in +typically unique condition (parallel simple effect also present for 'production first'), see Figure 8.

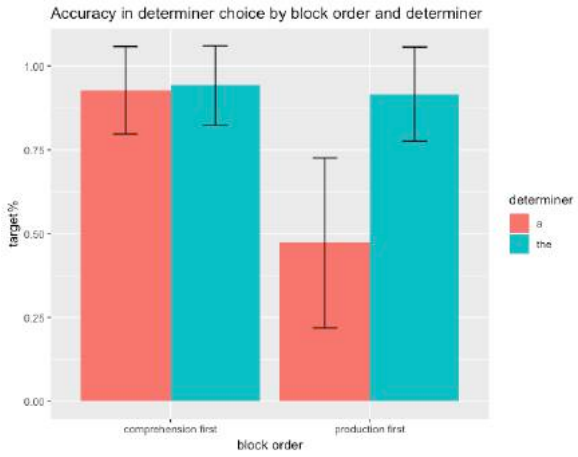


Figure 7: Target determiner choices for production task by block order

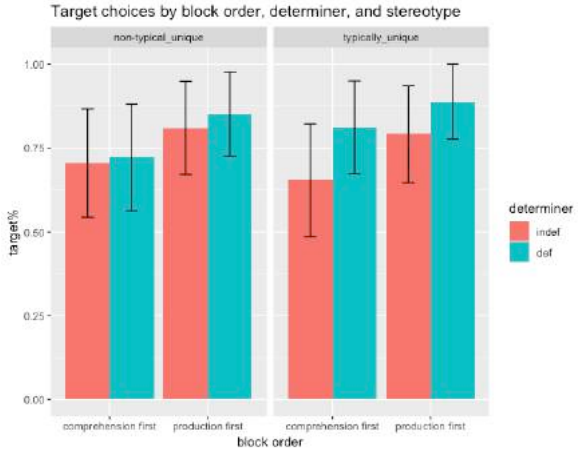


Figure 8: Target selections for comprehension task by block order and stereo-typical uniqueness

Eye movements For the analysis of the eye movements we considered time windows of 200ms between 100ms and 2300ms after noun phrase onset (which is roughly 4 seconds into the trial and the point where, on average, participants picked up the splash to place it on the picture). We looked at the fixations on the respective target for definite (pic 1) and indefinite (pic 2) relative to its competitor. The main dependent measure that we report on below is Target Advantage score, calculated by subtracting the proportion of fixations to the competitor from the proportion of fixations to the target.

In the first time window (100-300ms after noun phrase onset) we find a main effect of definiteness, with a higher target advantage for the indefinite. This effect is likely due to the target for the indefinite (pic 2) being the overall more unusual picture in the typically unique condition (e.g. with 2 TVs). This is also in line with the observa-

tion that in the typically unique cases, there is a significantly higher target advantage for the indefinite ($p<.01$) even before the noun phrase onset, see both graphs in Figure 9 and 10.

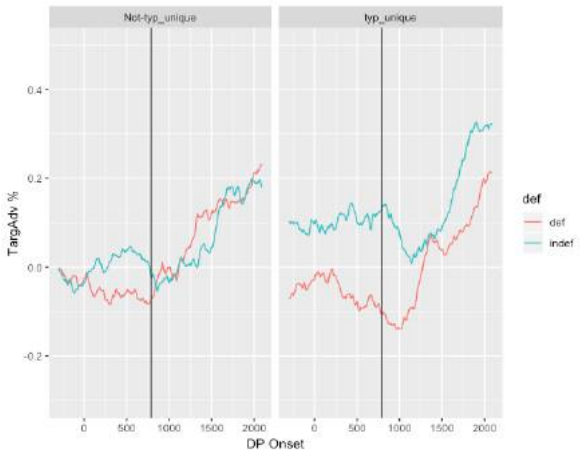


Figure 9: Target advantage by determiner and stereo-typicality for all trials (Black line indicates mean noun phrase-offset/PP-onset.)

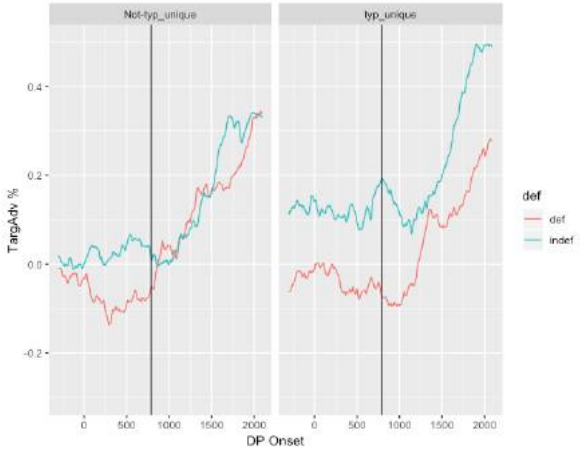


Figure 10: Target advantage by determiner and stereo-typicality for accurate trials (Black line indicates mean noun phrase-offset/PP-onset.)

Given this indication that pic 2, i.e. the target for the indefinite, has an advantage due to inherent properties of the picture, we also looked at the influence of both factors on the looks to pic 1 and pic 2, respectively, see figure 11. We find that in the early time window there is an interaction for looks to pic 1 between both factors: there are more looks to pic 1 for the definite if the item is typically unique and fewer looks to pic 1 for the indefinite if the target is non-typically unique. There are, however no main effects of the two factors. This together suggests that in the first time window the looks to the two pictures are guided by both determiner and stereo-typicality, and not oddness of the pic 2 picture alone.

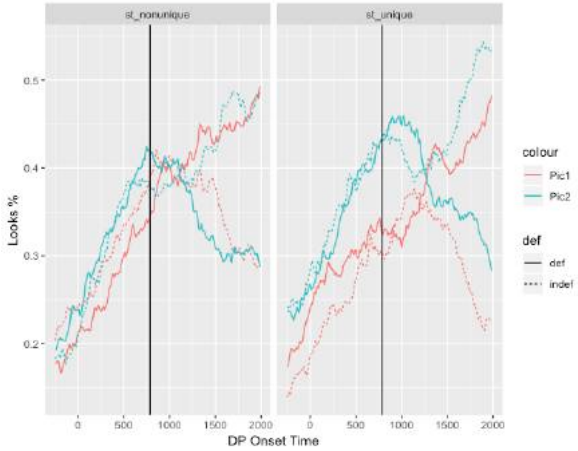


Figure 11: Looks to pic 1 and pic 2 by determiner and stereo-typical uniqueness, all trials)

In the time windows 300-500 and 500-700ms after noun phrase onset, there is a main effect of DEF ($p<.05$), but no interaction with STEREO-TYPICALITY. The effect is significant for the typically unique cases in both time windows ($p<.01$). But it is also marginally significant for the non-typically unique cases in time window 300-500ms ($p<.06$) and significant in the time window 500-700ms ($p<.05$). The interaction between both factors on looks to pic 1 is still marginal significant ($p<.06$), and significant when only looking at data from trials resulting in target choices, see figure 12.

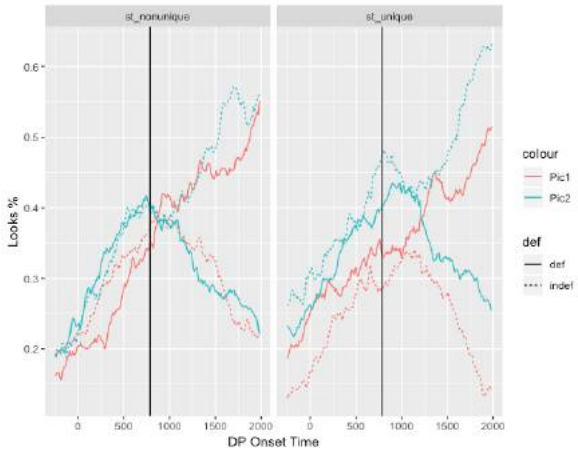


Figure 12: Looks to pic 1 and pic 2 by determiner and stereo-typical uniqueness, accurate trials)

In the time window 700-900 ms after noun phrase onset there is a significant interaction between STEREO-

TYPICALITY and DEFINITENESS in the predicted direction for Target Advantage: stereo-typicality has an effect on indefinites, but not on definites; as shown by the significant simple effect for the former ($p<.01$) and lack thereof for the latter. There is also a significant interaction between both factors in the time window 900-1100ms after noun phrase onset ($p<.01$). Looking at the simple effects shows that the interaction is due to stereo-typicality having a marginal effect on the definite ($p<.07$), but not the indefinite at this point, opposite to our predictions.

To better see potential differences from the properties of the respective target pictures for definites and indefinites, we also investigated looks to pic 1 and pic 2 separately for these time windows. There only are main effects of stereo-typicality for both pictures ($p<.05$ for pic 2, $p<.01$ for pic 1), suggesting that the interactions above are due to the pictures themselves attracting more attention depending on whether the item is typically unique.

No effects show up in time windows between 1100 and 1900 ms for target advantage. However, when considering looks to pic 1 and pic 2 we see a main effect of determiner for both pictures in all time windows between 1100 and 1900 ms after noun phrase onset ($p<.01$) for all of them).

There were no interactions between definite and time windows, and no interactions between block order and definite in all time windows.

Discussion

In combination, these results support the idea that, contrary to a view that sees anti-uniqueness as being based on a lexically encoded presupposition on par with the uniqueness presupposition of definites (theory B), the anti-uniqueness inference for indefinites is pragmatically derived from reasoning over the alternative expression (the definite) and its presupposition (uniqueness) (theory A). This is because, as predicted by theory A, the inference is not present as robustly from the start but is boosted by exposure to the alternatives and how they could matter: when the production follows comprehension, choice of the indefinite in the sentence construction phase increases for pictures with two objects of the relevant sort; and when the comprehension block follows the production block, choices of the picture with two of the relevant items increases for indefinite sentences.

With regards to the eye movement data, we find some effects that are at least in part complicated by the properties of the different target pictures. For the most part, except for very early time windows, the differences between definite and indefinite disappear when looking at only not-stereo-typically unique cases. This becomes especially apparent when looking at trials resulting in target choices, where there is no difference in time course between determiners, at least in the present task. We'd note, though, that the fixation shifts in our data are overall relatively late, which is likely at least partly due to the nature of the task requiring clicking and dragging splash-pictures around. Be this as it may, our data provides no evidence that the additional reasoning involved in deriving the inference by reasoning over the lexical presupposition of the alternative requires extra processing time if the conditions for this contextual reasoning over alternatives are met. This finding will need to be considered in relation to the complex empirical situation in the implicature processing literature, with some studies finding delays for implicature computation, and others not. The present results seem to constitute a case of a different type of a pragmatically derived inference that seemingly does not lead to any delays involved in accessing it. But further work is needed to try to establish this in task variations allowing for an earlier emergence of effects. Finally, a methodological lesson worth noting is that studying anti-uniqueness effects of indefinites experimentally requires careful fine-tuning of the task, as they can be evasive and are highly sensitive to various contextual factors.

References

Amsili, G., P. & Wintersein. (2015). Optionality in the use of too: The role of reduction and similarity. *Revista da Abralin (Associação Brasileira de Linguística)*, *XIV*, 229-252.

Bade, N. (2016). *Obligatory presupposition triggers in discourse - empirical foundations of the theories Maximize Presupposition and Obligatory Implicatures*. Unpublished doctoral dissertation, University of Tuebingen.

Bates, D. M. (2005). Fitting linear mixed models in r. *R News*, *5*, 27-30.

Burkhardt, P. (2006). Inferential bridging relations reveal distinct neural mechanisms: Evidence from event-related brain potentials. *Brain and Language*, *98*(2), 159 - 168. Retrieved from http://www.sciencedirect.com/science/article/pii/S0093934X06000733 doi: https://doi.org/10.1016/j.bandl.2006.04.005

Carpenter, P. A., Just, M. A., Keller, T. A., Eddy, W. F., & Thulborn, K. R. (1999, aug). Time course of fMRI-activation in language and spatial networks during sentence comprehension. *NeuroImage*, *10*(2), 216–224. doi: 10.1006/nimg.1999.0465

Chemla, E. (2008). An epistemic step for antipresuppositions. *Journal of Semantics*, *25*(2), 141-173.

Clifton, C. (2013). Situational context affects definiteness preferences: Accommodation of presupposi-

tions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(2), 487–501. doi: 10.1037/a0028975

Eckardt, R. (2014). *The semantics of free indirect discourse. how texts allow us to read minds and eavesdrop*. Brill.

Haviland, S. E., & Clark, H. H. (1974). What's new? acquiring new information as a process in comprehension. *Journal of Verbal Learning and Verbal Behavior*, *13*(5), 512 - 521. Retrieved from http://www.sciencedirect.com/science/article/pii/S0022537174800034 doi: https://doi.org/10.1016/S0022-5371(74)80003-4

Heim, I. (1983). On the projection problem for presuppositions. In D. P. Flickinger (Ed.), *Proceedings of WCCFL 2* (pp. 114–125). Stanford University, Stanford, California: CSLI Publications.

Heim, I. (1991). Artikel und definitheit. In A. Stechow & D. Wunderlich (Eds.), *Semantics: An international handbook of contemporary research* (p. 487-535). De Gruyter.

Herbert, C., & Kübler, A. (2011, oct). Dogs cannot bark: Event-related brain responses to true and false negated statements as indicators of higher-order conscious processing. *PLoS ONE*, *6*(10), e25574. doi: 10.1371/journal.pone.0025574

Kirsten, M., Tiemann, S., Seibold, V. C., Hertrich, I., Beck, S., & Rolke, B. (2014). When the polar bear encounters many polar bears: event-related potential context effects evoked by uniqueness failure. *Language, Cognition and Neuroscience*, *29*(9), 1147-1162.

Kratzer, A. (2005). Building resultatives. In C. Maienborn & A. Wöllstein-Leisten (Eds.), *Event arguments in syntax, semantics, and discourse* (pp. 177–212). Tübingen: Niemeyer.

Percus, O. (2006). Antipresuppositions. In U. Ueyama (Ed.), *Theoretical and empirical studies of reference and anaphora : Toward the establishment of generative grammar as empirical science.* Japan Society for the Promotion of Science.

R Core Team. (2017). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Reichle, E. D., Carpenter, P. A., & Just, M. A. (2000, jun). The neural bases of strategy and skill in sentence–picture verification. *Cognitive Psychology*, *40*(4), 261–295. doi: 10.1006/cogp.2000.0733

Rouillard, V., & Schwarz, B. (2017). Epistemic narrowing for maximize presupposition. In A. Lamont & K. A. Tetzloff (Eds.), *Proceedings of North East Linguistic Society 47* (p. 49-62).

Sauerland, U. (2008). On the semantic markedness of Phi-features. In D. Harbour, D. Adger, & S. Béjar (Eds.), *Phi theory* (pp. 57–82). Oxford: Oxford University Press.

Schumacher, P. B. (2009). Definiteness marking shows late effects during discourse processing: Evidence from erps. In S. Lalitha Devi, A. Branco, & R. Mitkov (Eds.), *Anaphora processing and applications* (pp. 91–106). Berlin, Heidelberg: Springer Berlin Heidelberg.

Schwarz, F. (2007). Processing presupposed content. *Journal of Semantics*, *4*(24), 373-416.

Tiemann, S. (2014). *The processing of wieder ('again') and other presupposition triggers*. Unpublished doctoral dissertation, University of Tuebingen.

Tiemann, S., Schmid, M., Bade, N., Rolke, B., Hertrich, I., Ackermann, H., ... Beck, S. (2011). Psycholinguistic evidence for presuppositions: On-line and off-line data. In I. Reich (Ed.), *Proceedings of sinn & bedeutung 15* (p. 581-595).

Fanning Creative Thought: Semantic Richness Impacts Divergent Thinking

Roger E. Beaty (rebeaty@psu.edu)
Department of Psychology, Pennsylvania State University
University Park, PA 16802 USA
Yoed N. Kenett (yoedk@sas.upenn.edu)
Department of Psychology, University of Pennsylvania
Philadelphia, PA 19104 USA
Richard W. Hass (Richard.Hass@jefferson.edu)
Department of Psychology, Thomas Jefferson University
Philadelphia, PA 19144 USA

Abstract

Creative thinking has long been associated with spreading of activation through concepts within semantic networks. Here we examine one potential influence on spreading activation known as the *fan effect*: increasing concept knowledge leads to increasing interference from related concepts. We tested whether cue association size—an index of semantic richness reflecting the average number of elements associated with a concept—impacts the quantity and quality of responses generated during the alternate uses task (AUT). We hypothesized that low-association AUT cues should benefit quality at the cost of quantity because such cues are embedded within a semantic network with fewer conceptual elements, thus yielding lesser interference from closely-related concepts. This hypothesis was confirmed in Study 1. Study 2 replicated the effect and found an interaction with fluid intelligence, indicating that cognitive control can overcome constraints of semantic knowledge. The findings thus highlight costs and benefits of semantic knowledge for creative cognition.

Keywords: Divergent Thinking; Fan Effect; Fluid Intelligence; Semantic Memory

Introduction

Divergent thinking (DT) is considered a crucial aspect of creative thinking (Runco & Acar, 2012). DT involves generating novel and appropriate responses to open-ended idea generation tasks. However, the basic cognitive processes involved in DT such as memory retrieval are far from understood (Volle, 2018). Recent research highlights both benefits and costs of semantic memory retrieval for creative thought (Beaty, Christensen, Benedek, Silvia, & Schacter, 2017; Hass, 2017a, 2017b; Kenett, 2019). Here, we borrow a classic experimental paradigm from cognitive science research on semantic memory—the fan effect (Anderson, 1974)—to further characterize the impact of semantic memory structure on creative idea generation.

A fan effect is an increase in response time (or error rates) on a recognition test with an increase in the number of associations with a concept in a memory probe (Anderson, 1974). According to Anderson and Reder (1999), the associations among concepts cause interference (i.e., the more association links fanning from a concept node, the greater the interference). Such interference occurs at retrieval (Anderson & Reder, 1999). Thus, during divergent thinking, the fan size of the target cue may relate to interference in retrieving creative responses. This hypothesis fits strongly with the associative theory of creativity.

According to the associative theory of creative thinking, creativity involves the connection of weakly related, remote concepts into novel and applicable concepts (Mednick, 1962). The farther apart the concepts are in semantic space, the more creative the new combination will be. For this new combination to be applicable (i.e., to make sense) a broad body of knowledge is required. While this theory is still debated, the importance of associative abilities in creative processing has been demonstrated (Benedek, Könen, & Neubauer, 2012). Furthermore, recent computational studies have provided further support for how individual differences in creative ability relates to differences in semantic memory structure (Kenett, 2019; Kenett & Faust, 2019).

However, more recent theories have argued that creative thinking is more broadly related to an interaction between semantic memory structure and cognitive control processes that facilitate guided search throughout memory. For example, Beaty and Silvia (2012) examined how fluid intelligence (*Gf*) relates to the serial order effect – the tendency for ideas to become increasingly more original over time during a DT task (Christensen, Guilford, & Wilson, 1957; Hass & Beaty, 2018). The authors found that participants scoring high on *Gf* showed less of a serial-order effect. That is, high *Gf* scores were associated with greater originality earlier in participants' response. Thus, the authors argue for an executive control process operating on semantic memory that facilitates avoidance of high-frequency concepts (i.e., low-originality). Hass (2017b) applied computational approaches to demonstrate that the semantic similarity of participants' DT responses non-linearly decreases as a function of response order. Furthermore, this study found that the semantic similarity of initial DT responses was lower for participants with higher *Gf* scores.

In the current series of studies, we present results from an ongoing project, where we examine for the first time the role of fan size on DT responses. As the fan effect is considered to be related to activation of multiple associations to a

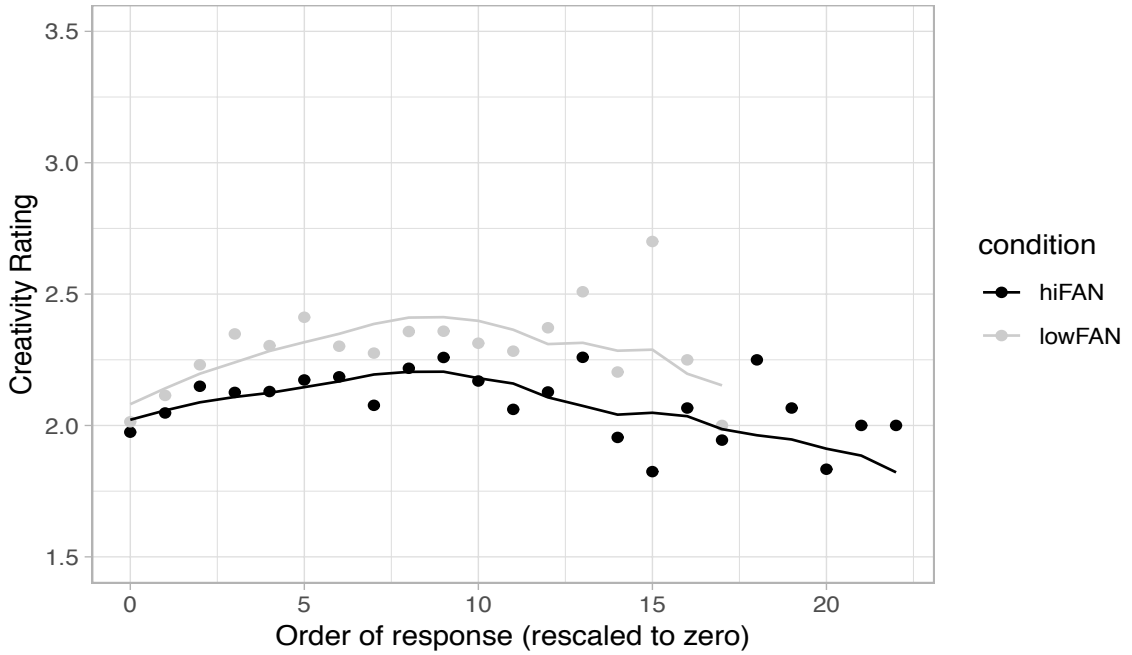


Fig. 1: Average creativity rating as a function of response order. The first response was rescaled to zero for the purpose of growth modeling. Lines represent fitted values for each condition based on the multilevel growth model.

concept that leads to interference, we predict that high-fan cues will lead to a higher number of DT responses but lower overall originality of these responses. Furthermore, we predict that fan size will interact with the serial order effect (Study 1) and with individual differences in *Gf* (Study 2).

Study 1

In Study 1, we sought to test for the existence of a fan effect in the context of performance on the Alternate Uses Task (AUT), a widely used assessment of DT (Runco & Acar, 2012). To this end, we assessed whether the quantity and creative quality of DT responses varied as a function of AUT cue association set size. We selected cue words for the AUT (i.e., common objects) with low- and high-fan size based on free association norms (Nelson, McEvoy, & Schreiber, 2004). Because high-fan cues are presumably embedded within denser semantic networks of highly-related concepts compared to low-fan cues, we hypothesized that participants would generate more AUT responses (i.e., higher fluency) but that this performance benefit would come at the cost of creative quality (i.e., lower originality) due to interference from salient concepts.

Participants

Fifty-five participants were recruited for the study via Amazon Mechanical Turk (AMT; Buhrmester, Kwang, & Gosling, 2011). Participants were offered $4.00 compensation for completion of all 10 tasks. No participants' work was rejected (i.e., all 55 workers were paid), however, a pre-analysis screening procedure identified 14 participants that failed to respond to all 10 cues and 1 participant that provided clearly random responses, and thus did not follow directions. The final sample size for analysis was 40 (30 female) with an average age of 38.1 years (SD = 12.07). A large majority of the sample identified as White/Caucasian (92.5%) with the remaining 7.5% identifying as either African American or "other". This study was approved by Jefferson University institutional review board.

Materials

Stimuli. We constructed low- and high-fan cues to be used in the DT task. Low- and high-fan cues were selected from the University of South Florida free association norms database, that includes norms for 5,000 cue words (Nelson et al., 2004). Importantly, for each of these cue words, the database lists the number and types of different associative responses that were generated to these cue words. The number of associative responses to a cue word was used as a proxy of fan size of the cue word (i.e., cue set size). Out of the 5,000 cue words, cue words of concrete objects that can be used in a DT task were manually selected. Finally, a list of five low-fan (clock, fork, lamp, lens, pen) and five high-fan (soap, rope, stick, marble, balloon) cues were selected. These cue words were matched on key linguistic variables: *frequency* (low-fan $M = 16.4$, $SD = 3.29$; high-fan $M = 21.3$, $SD = 10.97$; $t(8) = 1.00$, $p = .35$) and *concreteness* (low-fan $M = 5.88$, $SD = .67$; high-fan $M = 6.09$, $SD = .23$; $t(8) = .66$, $p = .53$). Critically, the average set size of the high-fan cues ($M = 22$, $SD = 1.22$) was significantly greater than the average set size of the low-fan cues ($M = 6.6$, $SD = 1.51$; $t(8) = 17.67$, $p < .001$).

Divergent thinking task. For each of the ten cue words (low- and high-fan), participants had three minutes to generate as many alternative uses as possible. Two main measures were computed from participants DT performance: originality and

fluency. For each response, *originality* was defined as the average of the originality ratings across independent raters for that response and *fluency* was defined as the sum of responses; we also logged *inter-response time* (the time between the first key strokes of successive responses) and the *order* of entry of each response. Participant-level variables were fluency and composite originality score (i.e., the average of the originality scores per person per prompt).

Procedure

A custom web application was created for administering the experimental tasks (Hass & Beaty, 2018). The interface consisted of an instructions page and a response-collection interface. The instructions page appeared before both blocks of trials (low- and high-fan) and, after reading instructions, participants proceeded to the tasks using a navigation button. The task interface appeared in an 800x600 pixel window and consisted of a text-display, which contained the object prompt for that trial and a text-entry field. The text-entry field allowed participants to edit responses prior to entering them and moving on to the next. Javascript code saved the first key press per response, the time at which the participant entered the response (by pressing ENTER or RETURN), and the text of the response itself. When ENTER was pressed, the text-field was cleared, and participants were not allowed to view previous responses.

First, participants provided consent to participate in the experiment. Following consent, participants were presented with an overall description of their task: that they would be prompted to generate ideas about specific prompts, along with some information about how long it would take, and that they should be ready to type. Participants then completed a practice trial to become acclimated to the typed entry interface which involved typing the names of colors that they knew for 30 seconds. Upon completion of practice, the first set of experimental trials started. They were informed that there would be five trials, each with a different object, and each lasting 3 minutes each. They then pressed a navigation button to continue. The order of trials within blocks and block presentation were randomized, and participants had a short break between blocks. Finally, participants completed a short demographic survey.

Results

Participants' responses were rated for originality on a 5-point scale designed for cognitive studies of DT (Hass, Rivera, & Silvia, 2018) by two research assistants and one AMT worker not involved in the experiment. Inter-rater reliability (ICC(2,3)) ranged from fair to good across the 10 cues (M = .47, SD = .15).

Participants generated a significantly higher number of responses to high-fan prompts (M = 9.17, SD = 3.42) than to low-fan cues (M = 8.1, SD = 3.1), $t(39) = 3.84$, $p < .001$). Furthermore, high-fan responses were rated significantly less original (M = 2.67, SD = 0.26) compared with low-fan responses (M = 3.03, SD = 0.26), $t(39) = 6.47$, $p < .001$. Together, these findings suggest that while high-fan cues afford more associative links (i.e., increased fluency), the effect may interfere with generating original responses (i.e., decreased originality) to them.

To investigate temporal effects of the fan manipulation, two response-level analyses were performed. First, inter-response times (IRTs) were compared across the two conditions with a mixed-effects regression model. In order to conform to model assumptions (namely normally distributed residuals), IRTs were log-transformed and regressed on 1) a fixed-effect of condition (low- vs. high-fan), 2) a random effect of participant, and 3) a random effect of cue. Though mean IRTs were shorter in the high-fan condition (M = 14.50s, SD = 13.99s) compared with the low-fan condition (M = 16.14s, SD = 16.63s), the fixed effect in the log-IRT model was not significant (b = .0004, $p = 0.55$).

Next, the relationship between response order and creativity rating was examined with a mixed-effects model. Prior results have illustrated a curvilinear relationship between serial order and creativity (Hass & Beaty, 2018) so linear and quadratic serial order terms were entered into the model. Interactions between condition (low- vs. high-fan) and both of the serial order terms were also modeled, along with random effects of participant and cue. There were significant linear (b = 0.039, $p < .001$) and quadratic trends (b = -0.016, $p = .02$), but the overall difference between low- and high-fan originality was not preserved in this model (b = 0.062, $p = .61$). Additionally, there was no difference in either the linear (b = 0.025, $p = 0.15$) or quadratic slopes (b = -0.016, $p = .23$) across the fan conditions (Fig. 1).

Discussion

The associative theory of creativity implicates spreading activation across concepts within semantic networks to generate novel ideas (Mednick, 1962). However, little is yet known about the benefits—and potential costs—of semantic memory in creative cognition. Here, we identify such benefits and costs of semantic knowledge to performance on a divergent thinking task. Participants generated more responses during the AUT when using high-fan cues compared to low-fan cues, suggesting that greater semantic content benefits ideational fluency. This benefit, however, came at the cost of originality: participants generated ideas that were rated as less original in the high-fan condition. This finding is consistent with the notion that salient conceptual information (e.g., high-fan associations) can constrain creative thought by acting as a source of interference that must be inhibited to establish more remote conceptual combinations (Beaty et al., 2017; Chrysikou, 2019). In sum, the results of Study 1 suggest that the structure and content of semantic knowledge impacts the quality and quantity of ideas generated during divergent thinking.

Study 2

In Study 2, we sought to replicate and extend the findings from Study 1. Specifically, we employed the same

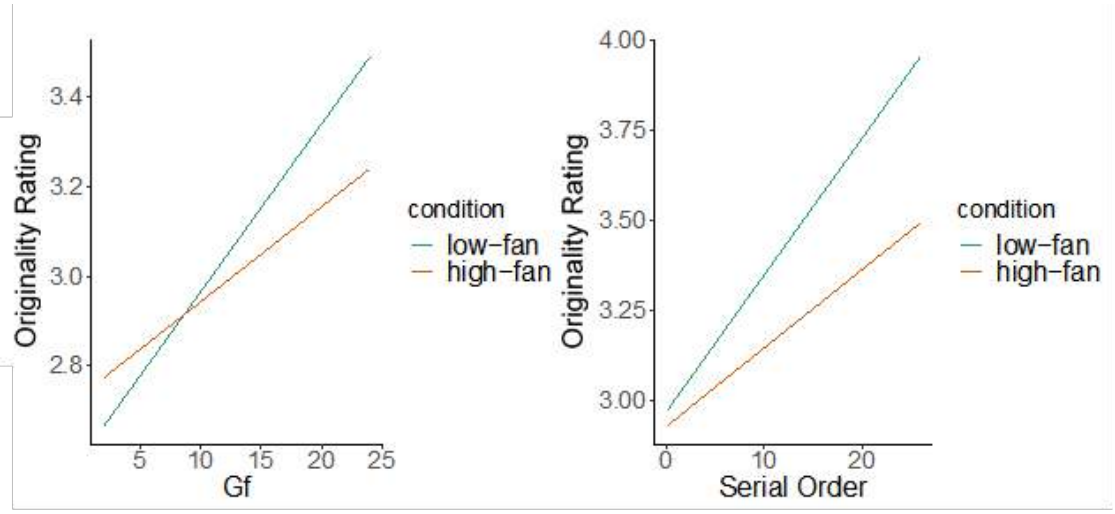


Fig. 2: Interaction effects between serial order and fan effect (left) and Gf and fan effect (right) on participant's originality ratings of their DT responses.

experimental paradigm—varying cue set size across AUT i tems—and further examined potential interactions with fluid intelligence (*Gf*), an individual difference variable with established links to divergent thinking (Beaty & Silvia, 2012). Although *Gf* has been shown to predict the creative quality of DT respons es, the cognitive contribution of *Gf* to creative performance remains largely uncharacterized. One possibility is that *Gf* supports inhibitory control processes, consistent with its strong association with cognitive control (Kane, Hambrick, & Conway, 2005) Thus, *Gf* may be more relevant for high-fan idea generation via the inhibition of salient conceptual knowledge (Beaty & Silvia, 2012). On the other hand, *Gf* may support low-fan idea generation by facilitating spreading activation within a relatively sparse semantic space. In addition to examining the role of *Gf*, we further probed temporal dynamics of the fan effect as a function of cue set size.

Participants

One hundred thirteen participants (50 females) were recruited from AMT. The average age of participants was 37.71 years (SD = 10.49). All participants were fluent in English and the majority (58%) were Caucasian. African Americans comprised 11% of the sample, Asian Americans comprised 5% of the sample, Hispanic Americans comprised 8% of the sample, Native Americans comprised 14% of the sample, while the remainder identified ethnicity as "other". Participants received $5.50 for their participation. Thirty-three participants were excluded from the analysis for failure to successfully complete all tasks or providing nonsensical answers to open-ended questions. The final sample size for the current analysis was $N = 83$. This study was approved by Jefferson University institutional review board.

Materials

Stimuli. The stimuli used in Study 2 were identical to those used in Study 1.

Divergent Thinking. The DT task used in Study 2 was identical to that used in Study 1.

Fluid Intelligence. Based on Kenett et al. (2016), *Gf* was assessed via three separate tasks: 1) The series task from the Culture Fair Intelligence Test (CFIT) which involves choosing an image that correctly completes a series of images (13 items, 3 min); 2) A letter-sets task, which presents a series of four-letter combinations and requires people to determine which set does not follow a rule governing the other four (16 items, 4 min); and 3) A number-series task in which participants complete a sequence of numbers by discovering a guiding rule (15 items, 5 min). To compute a general *Gf* score, we used principal component analysis. This composite *Gf* score was constructed as the sum of the multiplication of each independent *Gf* score by its weight of the first unrotated principal component (Kenett et al., 2016).

Procedure

The DT task was run similarly as in Study 1 and the *Gf* tasks were run via Qualtrics (www.qualtrics.com). Upon providing electronic consent, participants were presented with an overall description of their tasks: that they would be prompted to generate ideas about specific prompts for approximately 30 minutes, and they would then complete some IQ-based tasks for another 20 to 30 minutes. Participants then completed a practice idea-generation trial to become acclimated to the typed entry interface (naming colors). Upon completion of practice, the first set of experimental trials began. The order of trials within blocks and block presentation were randomized, and participants had a short break between blocks. Finally, participants completed a short demographic survey.

Results

Three raters scored responses for originality using the same scale used in Study 1 (Hass et al., 2018). Inter-rater reliability, assessed with interclass coefficients ICC(2,3), was generally high across the 10 cues (M = .68, SD = .12).

Analyzing the fluency and originality of participants responses, our results replicate the findings of Study 1: Participants generated a significantly higher number of responses to high-fan cues (M = 7.56, SD = 3.82) than to low-fan cues (M = 6.33, SD = 3.04), $t(82) = -4.64$, $p < .001$. Furthermore, high-fan responses were rated significantly less original (M = 3.04, SD = .33) compared with low-fan responses (M = 3.12, SD = .44),, $t(82) = 2.14$, $p < .035$. Together, these findings suggest that while high-fan concepts afford more associative links, these links may interfere with establishing more remote conceptual combinations.

Next, the relationship between response order and creativity rating was examined via a mixed-effects model. In our full model, *Gf*, fan, and serial order were assigned as independent measures, and the originality ratings as the dependent measure. Interactions between fan and *Gf*, interaction between fan and serial order, and interaction between *Gf* and serial order terms were also modeled, along with random effects of participant and cue (Table 1). We first compared this model to a model that only included the random effects and found that this model improved the fit to originality ratings, χ^2 (6, N = 83) = 105.52, $p < .001$. Specifically, we find a significant positive relation between each of the three main variables (*Gf*, Fan, and Order) on participant's originality scores. Thus, we replicate and strengthen the results found in Study 1, and replicate previous findings on the effect of *Gf* on DT (Beaty & Silvia, 2012). As for the effect of the interaction terms, we found significant negative relations between both interaction terms (*Gf**Fan and order*fan) on participant's originality scores (Fig. 2). However, due to high collinearity between the serial order variable and the interaction of *Gf* and serial order variable (r = -.71), the interaction effect of serial order and *Gf* was not significantly related to originality scores in this model.

Table 1: Linear mixed effect model of originality

Fixed Effects	*B*	SE	*p*
Intercept	2.28	0.18	< .001
Gf	0.05	0.01	< .001
Fan	0.19	0.10	.05
Order	0.05	0.01	< .001
*Gf**Fan	-0.02	0.00	< .001
Order*Fan	-0.02	0.01	< .001
Random Effects	Name	Variance	SD
Participant	Intercept	0.09	0.30
Cue	Intercept	0.01	0.09
Residual		0.65	0.80

Full model: Originality ~ *Gf* + Fan + Order + *Gf**Fan + Order*Fan + *Gf**Order (1|participant) + (1|cue)

Discussion

Study 2 replicated the findings of Study 1 and extended them by examining individual differences in *Gf* (Beaty & Silvia, 2012). As in Study 1, we found that, compared to low-fan cues, high-fan cues yielded increased fluency but decreased originality. Study 2 further examined temporal dynamics of this fan effect. Specifically, we replicated the serial order effect in divergent thinking—the tendency of idea originality to increase over time (Hass & Beaty, 2018)—and showed how this serial order effect interacted with both fan size and *Gf*. Although the 3-way interaction between serial order, fan size, and *Gf* was not significant, due to exceedingly high collinearity between these independent variables, we found that interaction effects of *Gf**Fan and Order*Fan explained significant variance in creativity ratings.

General Discussion

Divergent thinking tasks are widely used to assess creative thinking, but little is known about the basic cognitive processes underlying their performance. In two studies, we borrowed a classic experimental manipulation from cognitive research on semantic memory known as the fan effect (Anderson, 1974)—the tendency for increasing semantic associations to interfere with memory performance—and show that it similarly (but differentially) impacts the quality and quantity of divergent thinking responses. In Study 1, we found that although participants generated significantly more responses using high-fan cues compared to low-fan cues (i.e., increased fluency), these responses were rated as significantly less original. In Study 2, we replicated these findings and extended them by showing that the fan effect for originality varied as a function of individual differences in *Gf*: as *Gf* increased, so did originality ratings in the low-fan condition compared to the high-fan condition. Taken together, the results extend recent work on the dynamics of memory retrieval and cognitive control during creative idea generation (Benedek & Fink, 2019).

These findings inform a growing literature on the role of cognitive control in divergent thinking. Consistent with past work (Beaty & Silvia, 2012; Benedek, Jauk, Sommer, Arendasy, & Neubauer, 2014), Study 2 found that *Gf* predicted the creative quality of divergent thinking responses. Critically, we found that *Gf* interacted with the fan effect: higher-*Gf* benefited low-fan originality. From a semantic network perspective, the low-fan cues may be embedded in a less densely connected network, potentially blunting spreading activation to remote concepts due to less semantic scaffolding (Mednick, 1962). Thus, one possibility is that *Gf* compensates for such sparse semantic connectivity by driving search processes in a top-down fashion. In other words, when less is known about an object, cognitive control may facilitate strategic and deliberate conceptual combination.

On the other hand, one might predict *Gf* to benefit high-fan originality. Because the high-fan cues may be embedded within a relatively denser network of semantic associations—as reflected by higher ideational fluency in the high-fan

condition across both studies—these associations may have induced interference due to high salience and semantic relatedness. Prior research suggests that salient concepts can disrupt idea generation by priming what is already known and thus not original (Beaty et al., 2017). Thus, cognitive control could benefit high-fan cues via inhibitory mechanisms, i.e., suppressing dominant responses and redirecting search processes (Beaty & Silvia, 2012). Notably, however, Study 2 assessed *Gf*—a proxy measure of general cognitive control which shows strong correlation with executive processes such as inhibitory control (Kane et al., 2005) Future work might resolve this question by examining the contribution of specific executive functions to idea generation under similar semantic constraints.

The present research has potential implications for understanding the role of semantic knowledge in creative cognition (Kenett & Faust, 2019). Across both studies, we found a dissociation between the quantity and quality of ideas as a function of fan size: more ideas are generated when more was "known" about an object—as indexed via semantic associations—but these ideas were deemed to be of less creative quality. An interesting direction for future research would be to explore the extent to which this effect extends beyond "domain-general" creative performance to specific creative domains. Another outstanding question concerns whether the organization of semantic knowledge can be optimized for creativity through learning. We suspect that high creative ability is characterized by extensive domain-relevant knowledge, and superior access to that knowledge, via its hierarchical organization and top-down retrieval.

References

Anderson, J. R. (1974). Retrieval of propositional information from long-term memory. *Cognitive psychology, 6*(4), 451-474.

Anderson, J. R., & Reder, L. M. (1999). The fan effect: New results and new theories. *Journal of Experimental Psychology: General, 128*(2), 186-197.

Beaty, R. E., Christensen, A. P., Benedek, M., Silvia, P. J., & Schacter, D. L. (2017). Creative constraints: Brain activity and network dynamics underlying semantic interference during idea production. *NeuroImage, 148*, 189-196.

Beaty, R. E., & Silvia, P. J. (2012). Why do ideas get more creative over time? An executive interpretation of the serial order effect in divergent thinking tasks. *Psychology of Aesthetics, Creativity and the Arts, 6*(4), 309-319.

Benedek, M., & Fink, A. (2019). Toward a neurocognitive framework of creative cognition: the role of memory, attention, and cognitive control. *Current Opinion in Behavioral Sciences, 27*, I116-122.

Benedek, M., Jauk, E., Sommer, M., Arendasy, M., & Neubauer, A. C. (2014). Intelligence, creativity, and cognitive control: The common and differential involvement of executive functions in intelligence and creativity. *Intelligence, 46*, 73-83.

Benedek, M., Könen, T., & Neubauer, A. C. (2012). Associative abilities underlying creativity. *Psychology of Aesthetics, Creativity and the Arts, 6*(3), 273-281.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's mechanical turk a new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science, 6*(1), 3-5.

Christensen, P. R., Guilford, J. P., & Wilson, R. C. (1957). Relations of creative responses to working time and instructions. *Journal of Experimental Psychology, 53*(2), 82-88.

Chrysikou, E. G. (2019). Creativity in and out of (cognitive) control. *Current Opinion in Behavioral Sciences, 27*, 94-99.

Hass, R. W. (2017a). Semantic search during divergent thinking. *Cognition, 166*, 344-357.

Hass, R. W. (2017b). Tracking the dynamics of divergent thinking via semantic distance: Analytic methods and theoretical implications. *Memory & Cognition, 45*(2), 233-244.

Hass, R. W., & Beaty, R. E. (2018). Use or consequences: Probing the cognitive difference between two measures of divergent thinking. *Frontiers in Psychology, 9*(2327).

Hass, R. W., Rivera, M., & Silvia, P. J. (2018). On the dependability and feasibility of layperson ratings of divergent thinking. *Frontiers in Psychology, 9*(1343).

Kane, M. J., Hambrick, D. Z., & Conway, A. R. A. (2005). Working memory capacity and fluid intelligence are strongly related constructs: comment on Ackerman, Beier, and Boyle (2005). *Psychological Bulletin, 131*(1), 66-71.

Kenett, Y. N. (2019). What can quantitative measures of semantic distance tell us about creativity? *Current Opinion in Behavioral Sciences, 27*, 11-16.

Kenett, Y. N., Beaty, R. E., Silvia, P. J., Anaki, D., & Faust, M. (2016). Structure and flexibility: Investigating the relation between the structure of the mental lexicon, fluid intelligence, and creative achievement. *Psychology of Aesthetics, Creativity, and the Arts, 10*(4), 377-388.

Kenett, Y. N., & Faust, M. (2019). A semantic network cartography of the creative mind. *Trends in Cognitive Sciences, 23*(4), 271-274.

Mednick, S., A. (1962). The associative basis of the creative process. *Psychological Review, 69*(3), 220-232.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers, 36*(3), 402-407.

Runco, M. A., & Acar, S. (2012). Divergent thinking as an indicator of creative potential. *Creativity Research Journal, 24*(1), 66-75.

Volle, E. (2018). Associative and controlled cognition in divergent thinking: Theoretical, experimental, neuroimaging evidence, and new directions. In R. E. Jung & O. Vartanian (Eds.), *The Cambridge handbook of the neuroscience of creativity* (pp. 333-362). New York, NY: Cambridge University Press.

Relative Evaluation of Location: How Spatial Frames of Reference Affect What We Value

Andrea Bender (andrea.bender@uib.no)
Department of Psychosocial Science & SFF Centre for Early Sapiens Behaviour (*SapienCE*), University of Bergen, N-5020 Bergen, Norway

Sarah Teige-Mocigemba
Department of Psychology, Philipps University Marburg, D-35032 Marburg, Germany

Annelie Rothe-Wulf (annelie.rothe@psychologie.uni-freiburg.de)
Department of Psychology, University of Freiburg, D-79085 Freiburg, Germany

Miriam Seel (miriam.seel@googlemail.com)
Graduate School of Environmental Studies, Nagoya University, 464-8601 Nagoya, Japan

Sieghard Beller[†]
Department of Psychosocial Science, University of Bergen, N-5020 Bergen, Norway

Abstract

How we mentally represent spatial relations is known to have effects on cognitive processes such as inferences, co-speech gesture, or memorizing. In addition, spatial positions often serve as metaphors that carry valence. For instance, "moving up the social ladder, "getting it right", or being "in front" feels certainly better than "moving down", "having two left feet", or "lagging behind". Spatial position, however, depends on perspective, more concretely on which frame of reference (FoR) one adopts—and hence on cross-linguistically diverging preferences. What is conceptualized as "in front" in one variant of the relative FoR (e.g., *translation*) is "behind" under another variant (*reflection*), and vice versa. Do such diverging conceptualizations of an object's location also lead to diverging evaluations? We tested this with speakers of German, Chinese, and Japanese using an Implicit Association Test (IAT). Data from two studies suggest that across languages the object "in front of" another object is evaluated more positively than the one "behind", and that both location and evaluation depend on the adopted FoR. In other words: linguistically imparted FoR preferences appear to impact on evaluative processes.

Keywords: spatial cognition, frames of reference, valence, IAT, cross-linguistic comparison

Introduction

Space is of fundamental importance, not only for our very existence and survival—and hence for core cognitive activities devoted to them such as orientation and navigation (e.g., Hutchins, 1983; Golledge, 1999)—but also as a source of metaphors for grasping more abstract or elusive concepts such as number or time (Bender & Beller, 2014; Dehaene, 2003; Núñez & Cooperrider, 2013; Walsh, 2003). For instance, preferences for spatial representations seem to provide structure for how we represent temporal relations (Boroditsky, 2000; Boroditsky & Gaby, 2000).

A number of expressions points to the possibility that spatial representations may also provide metaphorical structure for evaluative judgments, especially along the vertical axis and the lateral axis, with *up* and *right* being predominantly linked to positive valence, and *down* or *left* to negative valence in various cultures (Keating, 1995; Lakoff & Johnson, 1999; Meier & Robinson, 2004). Expressions such as "being at the forefront" versus "lagging behind" do hint at corresponding associations along the sagittal axis as well.

The relationship between space and valence, however, is more complex than these examples suggest, and may be mediated by additional factors. For instance, the more positive evaluation of objects to the right than of those to the left is reversed in left-handers (Casasanto, 2009, 2011), and lateralization in terms of handedness even overrides strong cultural conventions (de la Fuente, Casasanto, Román, & Santiago, 2015). Yet, handedness only affects people's embodied experiences of their own right and left; it does not determine whether they mentally represent an object as being located to the right or left. Evaluations of objects are therefore directly dependent on location in space: If an object changes location, its evaluation changes. But what if it is not location in space that changes, but rather the mental representation of this location? Is the valence of objects also affected if relative positions themselves are conceptualized differently depending on a person's preference for referring to these positions? We addressed this question with a focus on the sagittal axis, for which space-valence associations have not been explored. At the same time, it is the only axis along which the conceptualization of location is affected in distinct ways by linguistic and cultural conventions and hence may vary in important ways (Beller, Singmann, Hüther, & Bender, 2015; Majid, Bowerman, Kita, Haun, & Levinson, 2004).

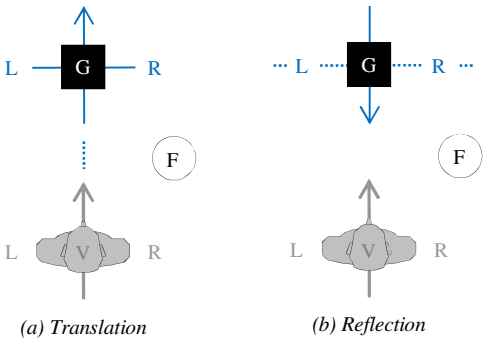


Figure 1:
Two variants of the relative FoR (Levinson, 2003); F: figure; G: ground; V: viewpoint of the observer; L/R: left/right

Indeed, what is assigned as FRONT or BACK along the sagittal axis depends on the frame of reference (FoR) one adopts. While all FoRs are coordinate systems that help to locate one object (the figure) in reference to another object (the ground), they differ with regard to where they are anchored (Levinson, 2003)[1]. The relative FoR relevant for our study is anchored in an observer. Therefore, to locate the figure in reference to the ground, the observer's coordinate system needs to be transferred to the ground. Crucially, this can be done in different ways—by shifting it to the ground (*translation*) or by mirroring it in the ground (*reflection*)—leading to opposing assignments of FRONT and BACK for the very same arrangement (see Figure 1): Whereas translation implies a further-away object to be conceptualized as "in front of" the ground and a nearer object as "behind", reflection implies the nearer object as "in front" and the further-away object as "behind".

Whether these diverging assignments of FRONT and BACK lead to diverging evaluations is the question we sought to answer. We assumed that, regardless of FoR preference, speakers of widely different languages evaluate objects more positively when conceptualizing them as "in front of" another object than those conceptualized as "behind". Since the object conceptualized as "in front" depends on FoR preference, speakers with a preference for *translation* should evaluate the further-away object more positively, whereas speakers with a preference for *reflection* should evaluate the nearer object more positively.

Study 1

In view of the cross-linguistic distribution of the relative FoRs, as obtained from language elicitation tasks (Beller & Bender, 2017; Beller et al., 2015), we recruited native speakers of German in which reflection is prevalent, and of Chinese and Japanese in which translation is more frequent. Introducing a novel approach into this field of research, we use an *Implicit Association Test* (IAT; Greenwald, McGhee, & Schwartz, 1998) to assess the positive versus negative valence of objects that the participants conceptualized as being "in front of" versus "behind" another object, depending on their preferred FoR.

Method

Participants The sample consisted of 43 native speakers of German (28 female; mean age 23 years, range: 18-35), 40 native speakers of Chinese (27 female; mean age 27 years, range: 22-38), and 40 native speakers of Japanese (22 female; mean age 19 years, range: 18-34). The Chinese participants were born in China to monolingual parents, had been living in Germany for 2.8 years on average (SD = 1.9), and reported excellent proficiency in Chinese (M = 5.0, SD = 0.2) compared to moderate levels of German (M = 3.1, SD = 1.3) and English (M = 3.5, SD = 0.9) on 5-point-rating scales. Data collection took place in Germany (for German- and Chinese-speaking participants) and Japan (for Japanese-speaking participants), and was conducted in the participants' mother tongue by native speakers of German, Chinese, or Japanese, respectively, as experimenters. Participation was voluntary, and was rewarded either with course credit or with 2 Euros or 400 Yen, respectively.

Materials In the IATs, participants discriminated stimuli according to either valence or space. For the standard valence discrimination task, six positive nouns (*health, happiness, smile, joy, peace, friend*) and six negative nouns (*agony, suffering, stench, mishap, illness, war*) had to be categorized as positive or negative. For the spatial discrimination task, twelve schematic drawings of two neutral objects were used. The objects were arranged on the front/back axis and were distinguishable by shape and color (blue/green). Counterbalanced across participants, the objects of one color were singled out as those to be categorized as "in front of" or "behind" the objects of the other color. If, for instance, the target color was green,

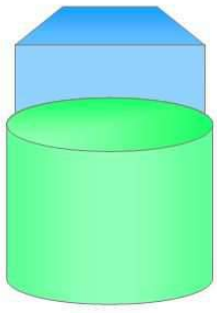


Figure 2:
Example of spatial stimuli used in Study 1

[1] Alternative terminologies are proposed, for instance, by Bohnemeyer and O'Meara (2012) and by Grabowski (1999).

Table 1: Task sequence and example of response key assignment.

			Example of response key assignment	
Block	*N* of trials	Task	D-key	L-key
1	26	Spatial discrimination	behind	in front
2	26	Valence discrimination	negative	positive
3	28	Initial combined task	behind/negative	in front/positive
4	52	Initial combined task	behind/negative	in front/positive
5	26	Reversed spatial discrimination	in front	behind
6	28	Reversed combined task	in front/negative	behind/positive
7	52	Reversed combined task	in front/negative	behind/positive

participants preferring translation (vs. reflection) would categorize the green cylinder in Figure 2 as "behind" (vs. "in front of") the blue cube.

Procedure The IATs were implemented as standard seven-block IATs (for details, see Table 1): Participants first completed two single-task practice blocks (one on spatial and one on valence discrimination). In Blocks 3 and 4, the two tasks were combined by mapping the four categories to two response keys (e.g., *in front*/*positive* on one key and *behind*/*negative* on the other). Block 5 was again a single-task block on spatial discrimination, but with the response key assignment reversed. In Blocks 6 and 7, this task was combined with the valence discrimination task of Block 2, thus mapping *behind*/*positive* on one key and *in front*/*negative* on the other.

The order of combined tasks was counterbalanced across participants (see Nosek, Greenwald, & Banaji, 2007). Stimuli were presented on a vertical computer screen, and responses were given by pressing the D- or L-key on the keyboard[2]. The intertrial interval was 500 ms. All blocks used warm-up trials with additional stimuli (excluded from the analyses), consisting of one trial per category that appeared within each block. Stimuli were presented randomly with the restriction that in the combined-task blocks, spatial and valence stimuli were presented in strictly alternating order.

The IAT effect is defined as the performance difference between the crucial blocks of combined tasks, and is interpreted as revealing the direction and strength of an association (here, between the space and valence categories). Typically, participants respond faster (and more accurately) when two associated categories share a response key than when they do not (Teige-Mocigemba, Klauer, & Sherman, 2010). Accordingly, if *in front* is evaluated more positively than *behind*, then responses should be faster in the *in front*/*positive*—*behind*/*negative* mapping than in the *in front*/*negative*—*behind*/*positive* mapping. If, by contrast, *in front* is perceived more negatively than *behind*, the response pattern should be reversed; and if no such link exists, response speed should not differ between mappings.

For all participants, IAT effects were coded such that positive values corresponded to the expected evaluation of *in front* as more positive than *behind*, independently of whether participants adopted translation or reflection to conceptualize where the target object is located. Assuming that all our participants evaluate objects more positively when conceptualizing them as "in front of" (than "behind") another object, we expected positive IAT effects. These effects may differ in size between samples, as there is no reason to assume that the space/valence associations should be of the exact same strength across cultures. What should differ significantly, subject to FoR preferences, is the object that is evaluated more positively: the further-away object under translation, and the nearer object under reflection.

Which of the two variants of the relative FoR a participant preferred was determined by assessing whether the figure presented in the IAT's practice block of the spatial discrimination task was categorized based on translation or reflection in the majority of trials. This assessment was necessary because adoption of a specific FoR is not determined by language, but based on a combination of (sub-)cultural conventions and individual preferences (Beller et al., 2015; Grabowski & Miller, 2000; Hill, 1982), and should therefore be gleaned from each participant's actual spatial discrimination decision.

Results and discussion

Using Tukey's (1977) criterion, we first examined whether any participant was an extreme outlier in terms of mean response latency in the combined tasks (i.e., with values three times the interquartile range below the first or above the third quartile). This led to the exclusion of two German participants, three Chinese participants, and five Japanese participants. Among the remaining participants, reflection was preferred by all 41 German participants (100%), by 28 Chinese participants (76%), and by 34 Japanese participants (97%), whereas translation was preferred by nine Chinese

[2] Keys were placed on the lateral instead of the sagittal axis to prevent confounding the very data we were interested in, namely on how FRONT and BACK are assigned along the sagittal axis.

Table 2: IAT effects in Study 1 and Study 2.

Study	Sample (*N*)	*M* (*SD*)	95% CI	*t*	*p*	Cohen's d_{D6}
Study 1	German (41)	231 (230)	[158, 303]	7.73	<.001	1.21
	Chinese (37)	146 (284)	[51, 241]	3.07	.004	0.50
	Japanese (35)	167 (207)	[105, 236]	5.48	<.001	0.93
Study 2	German (43)	292 (228)	[222, 362]	11.59	<.001	1.77
	Chinese (48)	153 (309)	[64, 243]	2.54	.015	0.37

participants (24%) and one Japanese participant (3%). Consistency in FoR adoption across the stimuli of the spatial discrimination task was high for all three samples and across FoR preferences, with $M > 94\%$ in each sub-group.

As recommended by Greenwald, Nosek, and Banaji (2003), IAT effects were calculated using the D6 scoring algorithm (used for inferential statistics only; for ease of interpretation, descriptive statistics are based on raw latencies). As expected, IAT effects were significant in all three samples of Study 1, $M \geq 146$ ms, $t \geq 3.07$, $p \leq .004$, indicating considerably faster responses to the *in front*/*positive*—*behind*/*negative* mapping than to the reversed mapping (Table 2).

Importantly, participants' evaluation of *in front* as more positive than *behind* was independent of their preferred variant of the relative FoR. Recall that nine of the 37 Chinese participants adopted translation. IAT effects for these participants were of the same size as those for participants preferring reflection, $t(35) = 1.06$, $p = .298$.

Participants thus evaluated *in front* more positively than *behind*—irrespective of their native language or cultural background. For the quarter of the Chinese participants preferring translation over reflection, the reversal of which object is conceptualized as "in front of" the other involved a corresponding reversal of evaluation of one and the same object: Further-away objects were more positive than nearer objects for participants preferring translation, but more negative for participants preferring reflection.

While the results of Study 1 are basically straightforward, the proportion of translational references among the Chinese- and Japanese-speaking participants was lower than anticipated. One reason could be that the nearer object partly occluded the further-away object, which may have highlighted the former at the cost of the latter (hence privileging reflection; cp. Bennardo, 2000; Grabowski, 1999; Hill, 1982). In addition, partially occluded objects may be devalued *a priori*. Since it was always the further-away object that was partially occluded, devaluation may have contributed to the more negative evaluation of this object by the majority of participants who preferred reflection and hence categorized the partially occluded further-away object as *behind*.

Study 2

To exclude partial occlusion as an alternative account, we repeated Study 1 with new spatial stimuli.

Method

Participants The new samples consisted of 50 native speakers each of German (35 female; mean age 22 years, *range*: 18-34) and Chinese (37 female; mean age 25 years, *range*: 18-33). Chinese participants were born in China to monolingual parents, had been living in Germany for 1.6 years on average ($SD = 1.4$), and reported excellent proficiency in Chinese ($M = 5.0$, $SD = 0.2$) compared to moderate levels of German ($M = 2.5$, $SD = 1.0$) and relatively good command of English ($M = 3.8$, $SD = 0.9$) on 5-point-rating scales. Data collection took place in Germany and was conducted in the participants' mother tongue.

Materials and Procedure For the spatial discrimination task, we now used photographs of real objects that were similar to the objects used in Study 1 both in shape and color, but differed in that no object was occluded (see Figure 3). In addition, an observer with the same looking

Figure 3:
Example of spatial stimuli used in Study 2

direction as that of the participant was inserted to emphasize perspective-taking. Apart from this, material and procedure were the same as in Study 1.

Results and discussion

The same exclusion criteria as in Study 1 led to the exclusion of seven German participants and two Chinese participants. Among the remaining participants, reflection was preferred by all 43 German participants (100%) and by 33 Chinese participants (69%), whereas translation was preferred by 15 Chinese participants (31%). Consistency in FoR adoption across stimuli was again high, with $M > 91\%$ in each sub-group.

IAT effects were computed as in Study 1 and were again significant in the two samples, $M \geq 153$ ms, $t \geq 2.54$, $p \leq .015$, indicating faster responses to the *in front*/*positive*—*behind*/*negative* mapping than to the reversed mapping (for details, see Table 2). Again, participants' evaluation of *in front* as more positive than *behind* was independent of their preferred FoR, as indicated by the non-significant difference between IAT effects for Chinese participants adopting translation versus reflection, $t(46) = 0.29$, $p = .773$.

As in Study 1, participants evaluated *in front* more positively than *behind*—irrespective of their native language or cultural background. And again, the reversal of which object is conceptualized as "in front of" the other involved a corresponding reversal of evaluation. Due to the modified stimuli used in this study, partial occlusion of the further-away object can be excluded as an explanation of its devaluation.

General Discussion

Does the way in which we evaluate objects depend also on how we *conceptualize* their location in space, rather than simply on where they *are* located? The work reported here suggests that this is indeed the case. Findings from two studies across three languages and cultural settings (with native speakers of German in Germany, of Chinese in Germany, and of Japanese in Japan) indicate that participants evaluate objects more positively when they conceptualize them as "in front of" another object than when they conceptualize them as "behind". Importantly, this positive evaluation holds for the further-away object when *translation* is adopted, yet for the nearer object when *reflection* is adopted.

The evidence is in line with the metaphor approach, according to which spatial concepts provide structure not only for more abstract domains, but also for evaluative judgments. While associations between space and valence have been described for the vertical axis (Keating, 1995; Meier & Robinson, 2004) and the lateral axis (Casasanto, 2009, 2011; de la Fuente et al., 2015), the present studies show these associations also for the sagittal axis. More concretely, they reveal that phrases such as "being at the forefront" versus "lagging behind" are not mere metaphorical expressions, but reflect a genuinely more positive evaluation of entities located "in front of" other entities. While the strength of this association differs somewhat across samples, with more pronounced effects for German participants than for the two East Asian groups (likely due to different strength of the association across cultures), its direction is the same in all three groups. This evidence is even more compelling in view of the fact that it was obtained with an *implicit* task specifically designed to tap into more automatic, rather than deliberate, processes.

Crucially, however, our findings also indicate that the association between location and valence is subject to linguistic and cultural conventions that affect how location is conceptualized—namely as *in front* or *behind*. Contingent on the adopted FoR, one and the same object in one and the same location may be evaluated as more positive or more negative: Under translation, the further-away object is regarded as the object *in front* and hence evaluated more positively, whereas under reflection, it is regarded as *behind* and hence evaluated more negatively.

In the current study, the proportion of translational references among the Chinese- (and Japanese-) speaking participants was lower than in previous surveys. While this lower proportion is disadvantageous for statistical power, it is not problematic per se, as such preferences are known to be subject to some variation depending on context (cf., Wilke, Bender, & Beller, 2019). A potentially more critical concern could be raised regarding the IAT itself. As this method assesses the link between relative location and valence by mapping category labels onto response keys, it might be suspected that participants could have used the category labels associated with the keys and their correspondence in polarity as a convenient short-cut (De Houwer, 2001; Proctor & Cho, 2006). However, the cognitive processing of stimuli required for the spatial task involved the computation of ternary relations between figure and ground from one's own viewpoint, which renders it unlikely that the observed effects were brought about by effects of labels or polarity only.

In conclusion, while previous work demonstrated that spatial representations have effects on cognitive processing (e.g., Bender & Beller, 2014; Haun, Rapold, Janzen, & Levinson, 2011; Levinson, Kita, Haun, & Rasch, 2002; Majid et al., 2004), here, we show that how we conceptualize the location of entities may even reverse the evaluation of these very entities. As conceptualizations of location are informed by diverging preferences for spatial FoRs across speech communities, their association with non-spatial conceptualizations and evaluations provides a promising new approach to explore effects of language and culture on cognition, which is a topic of key interest across several sub-disciplines of cognitive science. Opening up new avenues for investigation, implicit approaches like the one presented here could make it to the forefront in this contested field.

Acknowledgments

This work was supported by the Deutsche Forschungs-gemeinschaft DFG (Be 2451/13-1 and Be 2178/7-1) and in

part by the Research Council of Norway through the SFF *Centre for Early Sapiens Behaviour* (SapienCE), project number 262618. We are grateful to Christoph Klauer for support, to Lingyan Qian, Shixin Cheng, and Wenting Sun for their advice and assistance in data collection, and to Sarah Mannion de Hernandez for proofreading.

References

Beller, S., & Bender, A. (2017). How relative is the relative frame of reference? *Front* and *back* in Norwegian, Farsi, German, and Japanese. *Proceedings of the 39th Annual Conference of the Cognitive Science Society*. Austin: Cognitive Science Society.

Beller, S., Singmann, H., Hüther, L., & Bender, A. (2015). Turn around to have a look? Spatial referencing in dorsal vs. frontal settings in cross-linguistic comparison. *Frontiers in Psychology, 6:1283*.

Bender, A., & Beller, S. (2014). Mapping spatial frames of reference onto time: A review of theoretical accounts and empirical findings. *Cognition, 132,* 342-382.

Bennardo, G. (2000). Language and space in Tonga: "The front of the house is where the chief sits". *Anthropological Linguistics, 42,* 499-544.

Bohnemeyer, J., & O'Meara, C. (2012). Vectors and frames of reference: Evidence from Seri and Yucatec. In L. Filipović & K. M. Jaszczolt (Eds.), *Space and time in languages and cultures*. Amsterdam: John Benjamins.

Boroditsky, L. (2000). Metaphoric structuring: Understanding time through spatial metaphors. *Cognition*, *75*, 1-28.

Boroditsky, L., & Gaby, A. (2010). Remembrances of times East: absolute spatial representations of time in an Australian aboriginal community. *Psychological Science*, *21*, 1635-1639

Casasanto, D. (2009). Embodiment of abstract concepts: Good and bad in right-and left-handers. *Journal of Experimental Psychology: General*, *138*, 351-367.

Casasanto, D. (2011). Different bodies, different minds: The body specificity of language and thought. *Current Directions in Psychological Science, 20,* 378-383.

Dehaene, S. (2003). The neural basis of the Weber–Fechner law: A logarithmic mental number line. *Trends in Cognitive Sciences, 7,* 145-147.

De Houwer, J. (2001). A structural and process analysis of the Implicit Association Test. *Journal of Experimental Social Psychology, 37,* 443-451

de la Fuente, J., Casasanto, D., Román, A., & Santiago, J. (2015). Can culture influence body-specific associations between space and valence? *Cognitive Science, 39,* 821-832.

Golledge, R. G. (Ed.) (1999). *Wayfinding behavior: Cognitive mapping and other spatial processes*. Baltimore: JHU Press.

Grabowski, J. (1999). A uniform anthropomorphological approach to the human conception of dimensional relations. *Spatial Cognition and Computation, 1,* 349-363

Grabowski, J., & Miller, G. A. (2000). Factors affecting the use of dimensional prepositions in German and American English: Object orientation, social context, and prepositional pattern. *Journal of Psycholinguistic Research, 29,* 517-553.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology, 74,* 1464-1480.

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the Implicit Association Test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology*, *85*, 197-216.

Haun, D. B., Rapold, C. J., Janzen, G., & Levinson, S. C. (2011). Plasticity of human spatial cognition: Spatial language and cognition covary across cultures. *Cognition, 119,* 70-80.

Hill, C. A. (1982). Up/down, front/back, left/right. A contrastive study of Hausa and English. In J. Weissenborn & W. Klein (Eds.), *Here and there*. Amsterdam: Benjamins.

Hutchins, E. (1983). Understanding Micronesian navigation. In D. Gentner & A. L. Stevens (Eds.), *Mental models*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Keating, E. (1995). Spatial conceptualizations of social hierarchy in Pohnpei, Micronesia. In A. Frank & W. Kuhn (Eds.), *Spatial information theory*. Berlin: Springer.

Lakoff, G., & Johnson, M. (1999). *Philosophy in the flesh.* Chicago: University of Chicago Press.

Levinson, S. C. (2003). *Space in language and cognition.* Cambridge: Cambridge University Press.

Levinson, S. C., Kita, S., Haun, D. B., & Rasch, B. H. (2002). Returning the tables: Language affects spatial reasoning. *Cognition*, *84*, 155-188.

Majid, A., Bowerman, M., Kita, S., Haun, D. B. M., & Levinson, S. C. (2004). Can language restructure cognition? The case for space. *Trends in Cognitive Sciences, 8,* 108-114.

Meier, B. P., & Robinson, M. D. (2004). Why the sunny side is up: Associations between affect and vertical position. *Psychological Science, 15,* 243-247.

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2007). The Implicit Association Test at age 7: A methodological and conceptual review. In J. A. Bargh (Ed.), *Social psychology and the unconscious*. New York: Psychology Press.

Núñez, R. E., & Cooperrider, K. (2013). The tangle of space and time in human cognition. *Trends in Cognitive Sciences, 17,* 220-229.

Proctor, R.W., & Cho, Y.S. (2006). Polarity correspondence: A general principle for performance of speeded binary classification tasks. *Psychological Bulletin, 132,* 416-442.

Teige-Mocigemba, S., Klauer, K. C., & Sherman, J. W. (2010). Practical guide to Implicit Association Test and related tasks. In B. Gawronski & B. K. Payne (Eds.), *Handbook of implicit social cognition*. New York: Guilford Press.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Weasley.

Walsh, V. (2003). A theory of magnitude: Common cortical metrics of time, space and quantity. *Trends in Cognitive Science, 7,* 483-488.

Wilke, F., Bender, A., & Beller, S. (2019). Flexibility in adopting relative frames of reference in dorsal and lateral settings. *Quarterly Journal of Experimental Psychology*.

Building Individual Semantic Networks and Exploring their Relationships with Creativity

Matthieu Bernard* (matthieubernard@outlook.com)
Institut du Cerveau et de la Moelle épinière (ICM), UPMC UMRS 1127, Inserm U 1127, CNRS UMR 7225, Paris, France
Yoed N. Kenett* (yoedk@sas.upenn.edu)
Department of Psychology, University of Pennsylvania, Philadelphia, PA 19104, USA
Marcela Ovando-Tellez (marcela.ovandot@gmail.com)
Institut du Cerveau et de la Moelle épinière (ICM), UPMC UMRS 1127, Inserm U 1127, CNRS UMR 7225, Paris, France
Mathias Benedek (mathias.benedek@uni-graz.at)
Institute of Psychology, University of Graz, Graz, Austria
Emmanuelle Volle (emmavolle@gmail.com)
Institut du Cerveau et de la Moelle épinière (ICM), UPMC UMRS 1127, Inserm U 1127, CNRS UMR 7225, Paris, France
*** - Equal contributors**

Abstract

The associative theory of creativity suggests that creative abilities rely on the organization of semantic associations in memory. Recent research has demonstrated that semantic network methods allow testing this hypothesis. The aim of the current study was to investigate the properties of semantic networks at the individual level, in relation to creative abilities. Semantic judgement ratings were used to estimate individual semantic networks, whose topological properties measured by several graph metrics were correlated with individual creativity scores. We found a correlation between the theoretical semantic distance of our stimuli and the relatedness ratings given by the participants, demonstrating the validity of our approach. Importantly, we found a close relationship between creative abilities assessed by an achievement questionnaire and divergent thinking tasks and individual semantic network metrics, replicating and extending previous similar findings.

Keywords: creativity; semantic networks; network science; associative thinking

Introduction

The associative theory of creativity hypothesizes that creative abilities are related to individual differences in the organization of semantic associations in memory (Mednick, 1962). In support of this theory, several findings showed that more creative individuals had less common word associations or a less constrained organization of semantic associations (Beaty et al., 2014; Bendetowicz et al., 2017; Benedek et al., 2012; 2017; Kenett et al., 2014; Rossmann & Fink, 2010; Volle, 2018) and that in brain-damaged patients, rigid semantic associations were associated with poor creative abilities (Bendetowicz et al., 2018; Ovando-Tellez et al., 2019). Thus, the properties of semantic associations play a critical role in the cognitive processes that bring forth original ideas. Recently, computational methods exploring semantic memory structure in creativity are paving the way to uniquely study its role in creativity. One such computational approach is based on network science methodologies (Kenett, 2018; Kenett & Faust, 2019).

Network science is based on mathematical graph theory, providing quantitative methods to investigate complex systems, such as semantic memory, as networks (Siew et al., in press). In semantic networks, concepts or words are represented as nodes that are connected to each other by edges (denoting semantic similarity between concepts). The few studies that have applied semantic networks in the field of creative thinking indicate that studying the properties of semantic networks is a promising approach to explore creativity (Kenett, 2018; Kenett & Faust, 2019). For example, Kenett et al. (2014) investigated the semantic networks of low and high creative participants, based on free associations generated by both groups to a list of cue words. The authors showed that the semantic networks of low creative participants were less connected and more spread out compared to high creative participants.

However, aggregating over participants into groups may obscure individual differences related to creativity. To address this issue, Benedek et al. (2017) developed a new method to estimate individual semantic networks, based on semantic judgment ratings. Participants rated the relations between all possible pairs of 28 cue words, chosen to represent seven different categories. These relatedness ratings served as a proxy to the organization of these cue words in an individual's semantic memory. The authors showed how individual-based semantic network metrics correlated with individual-based creativity scores (Benedek et al., 2017) for specific types of filtered networks. However, in their study, the authors subjectively selected such cue words, and also applied a specific task, the Alternative Uses Task (AUT), to measure creative ability.

The general aim of the current study was to replicate and extend the relationships between the properties of individual semantic networks and creative abilities found by Benedek et al. (2017). Individual semantic networks were estimated using a modified version of Benedek et al. (2017) in which we controlled for the selection of the cue words based on a computational method. Participant's creativity was more extensively assessed via a creativity battery, including the

AUT used in the original study, a problem-solving task, and a creative achievement questionnaire. Specific network metrics of the individual semantic networks were computed and were correlated with the obtained creative scores.

Materials and Methods

Participants

Twenty-three healthy individuals aged between 22 and 37 years (26.96 ± 4.25) were included in the study. Participants were French-native speakers, right-handed with no neuropsychiatric disease. Two participants were excluded from the graph analysis because they rated >70% of word pairs as unrelated. This study was approved by the French ethical committee Sud Mediterrannée IV. Participants gave written consent and were paid for their participation.

General Overview

The study was composed of two parts. In the first part, the associative judgment task (AJT) was devised to estimate individual's semantic networks. The AJT was adapted from Benedek et al. (2017) by constructing new verbal material controlled for linguistic and semantic properties. In the AJT task, participants are asked to rate the semantic relatedness of pairs of words. In the second part, participants performed the AJT and a set of creativity tasks. AJT ratings were used to estimate the individual AJT-based semantic networks and network metrics were correlated with creative scores.

Part 1: The Associative Judgment Task (AJT)

We first used computational methods in order to develop and select a new set of cue words to be used in the AJT, accounting and controlling for semantic and linguistic properties. This was achieved by 1) estimating a large French semantic network, based on a large database of semantic association norms in French, and 2) by selecting a set of cue words, based on the properties of this network.

Creation of a French Semantic Network. To construct the French version of the AJT, we estimated a French semantic network of 1,081 words, based on French verbal association norms (Debrenne, 2011; http://dictaverf.nsu.ru/dictlist). This dataset was collected by asking French native speakers to provide the first word that came to mind after receiving a cue word. We selected words for which at least 400 participants provided a response. The final data contains 1,081 cue words and 26,268 responses from the participants.

The French semantic network was estimated using a network approach developed to analyze free association data (Kenett et al., 2014). According to this approach, each node represents a cue word and edges between nodes represent the association between these nodes. These associations represent the similarity profiles across any pair of cue words, i.e., the overlap of associative responses generated by the sample to each of the cue words.

The network was estimated in the following way: First the associative responses were preprocessed to standardize responses (correction of typos, elimination of non-words and articles, and spelling homogenization). Second, a data matrix was constructed such that each column is a cue word, and each row is a unique associative response. Thus, each cell denotes how many participants generated response *i* to cue word *j*. Third, the correlation between any pair of cue words was calculated using Pearson's correlation. This resulted in a 1,081 by 1,081 matrix where each cell denotes the semantic correlation between node *i* and node *j*. To minimize noise and possible spurious associations, we finally applied the planar maximally filtered graph filter (Kenett et al., 2014). To examine the structure of the networks, the edges were binarized so that all edges were converted to a uniform weight (i.e., 1). This allowed us to compute the shortest path between nodes in the network, serving as the theoretical semantic distance between them (Kenett et al., 2017).

Selection of AJT Stimuli. To select the verbal material to be used in the AJT, we developed a new computational method that allowed us to objectively select words with specific associative and linguistic properties from the French semantic network.

From the French semantic network, hierarchical tree structures were created recursively, using each node as a seed and searching for its neighbors. For each iteration, the neighbors of the neighbor's nodes were searched. In total we performed 4 iterations, considering that Kenett et al. (2017) demonstrated that most participants judged as unrelated the words separated by more than 4 steps in a force choice task. However, this tree procedure generated nodes that were separated by more than 4 steps when they belong to distinct branches, which allowed us to also generate word pairs that will be likely judged as unrelated. To avoid having one central node related to all the others by 4 steps or less, the initial seed node was removed.

The computation returned several solution trees among which one was selected for the AJT task based on the following criteria. First, for experimental reasons, the total number of nodes in the tree was limited to 35, i.e., 595 possible pair combinations between all words that had to be rated by the participants during the experiment. Second, we computed the *theoretical semantic distance* for all possible word pairs in term of the number of steps separating them in all of the trees. We selected the tree that optimized the proportion of pair words separated by 1, 2, 3, 4, or 5 or more steps. The selected tree contained a set of 35 words involving 595 possible word pairs with semantic distances distributed as follow: 10% of 1 step, 18% of 2 steps, 28% of 3 steps, 26% of 4 steps, 15% of 5 steps and 3% of 6 steps.

Part 2: AJT-based networks and creativity

Procedure of the AJT. Participants were presented successively with all the 595 combinations of pairs of the 35 selected words and were asked to rate their semantic relatedness, using a visual scale ranging from 0 (unrelated) to 100 (strongly related). Each trial started with the display of the word pair and a visual scale presented at the center of the

screen. After 2 seconds, the slider appeared in the middle of the scale. Participants could then freely move the cursor on the scale using a mouse and validated their response by a left click. They had to respond within 2 seconds. The final position of the slider in the scale after validation was considered as the semantic relatedness rating (Fig. 1).

In total, participants performed 6 different runs of 100 trials each (except the last run with 95 trials). Each run was composed of 4 blocks of 25 trials and separated by 20 seconds rest periods with a fixation cross. The trials were pseudo-randomly ordered within blocks with the constraint that each block contained a similar proportion of word pairs of each theoretical step. This order was fixed across participants. Before starting the task, participants performed a short practice. In addition, we checked that all participants were familiar with the 35 AJT words.

Relatedness ratings were coded for each participant and values were averaged separately for each theoretical distance and overall (see Fig. 2).

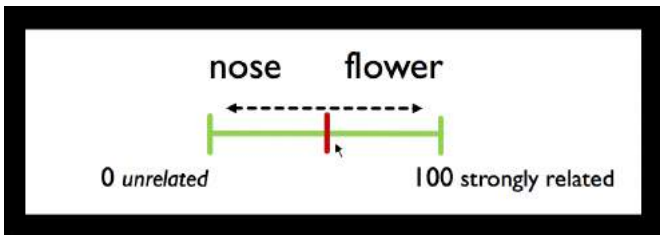


Fig. 1: Schematic representation of an AJT trial.

Estimating AJT-based individual semantic networks. Participant ratings of the word pairs during the AJT task served as a proxy of the organization of these words in their individual semantic network. We created a *n* by *n* matrix in which *n* represented the words used in the AJT task and each matching cell represented the semantic relatedness judgement given by the participant for these two words.

We employed two network filtering methods, one that had revealed significant relationships to creative abilities in previous work (Benedek et al., 2017) and a more conservative method that keeps more information in the network. In the first filtering method, we applied a fixed minimum relatedness threshold to the data and only edges with a weight of at least 50 were maintained. Since the value of 50 is the middle of the AJT scale, only edges corresponding to moderate to high semantic relatedness were kept and set to 1 whereas all the others were removed, resulting in an unweighted undirected network (UUN). In the second filtering method, all the edges were kept with their weight, so it preserved the variability in semantic judgments and resulted in a weighted undirected network (WUN). In this graph, each edge was weighted by the relatedness judgement given by the participant. For both networks, when the participant judged two words as unrelated (rating = 0), the two corresponding nodes had no edges linking them.

Based on the metrics previously related to creative abilities (Benedek et al., 2017; Kenett et al., 2014), we computed the following network metrics to characterize the structure of an individuals' semantic networks: the clustering coefficient (*CC*), the average shortest path length (*ASPL*), the diameter of the graph (*D*), smallworldness (*S*), betweenness centrality (*BC*), and modularity (*Q*). *CC* measures the degree to which nodes in a graph tend to cluster together. *ASPL* measures the average shortest number of steps that separate any pair of nodes, and *D* represents the longest path in the network. S is computed as a ratio between *CC* and *ASPL*. *BC* corresponds to the fraction of all shortest paths in the network that contain a given node. *Q* refers to the percentage of the network that is integrated into small-community structures. Analyses were performed with the Brain Connectivity Toolbox in Matlab (Rubinov & Sporns, 2010).

Creative Assessment. Creativity was assessed using the Combined Associates Task (CAT), the Alternative Uses Task (AUT) and the Inventory of Creative Activities and Achievements (ICAA).

The CAT is an adaptation of the Remote Associates Task (Mednick, 1962) developed by Bendetowicz et al. (2017; 2018) and assesses the ability to form new combinations between remotely associated words. In this task, participants are asked to find a word linked to three cue words with no apparent associations in a maximum of 30 seconds. CAT defines close and distant trials depending on the semantic distance between the cue words and the solution. 40 trials with an equal number of close and distant trials were administered. To quantify the data, four scores were analyzed. *CAT_Solving* is the sum of correct responses. *CAT_Close* and *CAT_Distant* correspond to the sum of correct responses in close and distant trials respectively, and *CAT_Index* corresponds to the difference in performance between distant and close trials, corrected by the averaged performance and was shown to reflect creative processes (Bendetowicz et al., 2017).

During ***the AUT*****,** participants were asked to generate original uses for a common object in three minutes. At the end of the three minutes, the participants selected their two most creative responses, as top-two scoring has been observed to be an effective approach to assess creativity (Benedek et al., 2013; Silvia et al., 2008). This procedure was repeated for three objects: tire, bottle and knife. The corresponding nouns naming the objects were presented on the screen during the 3 minutes. Scores for fluency and originality were assessed for each object. *AUT_Fluency* refers to the total number of ideas generated by the participant and *AUT_Originality* counts the number of infrequent ideas (given by less than 5% of the participants) among the top-two ideas of the participant.

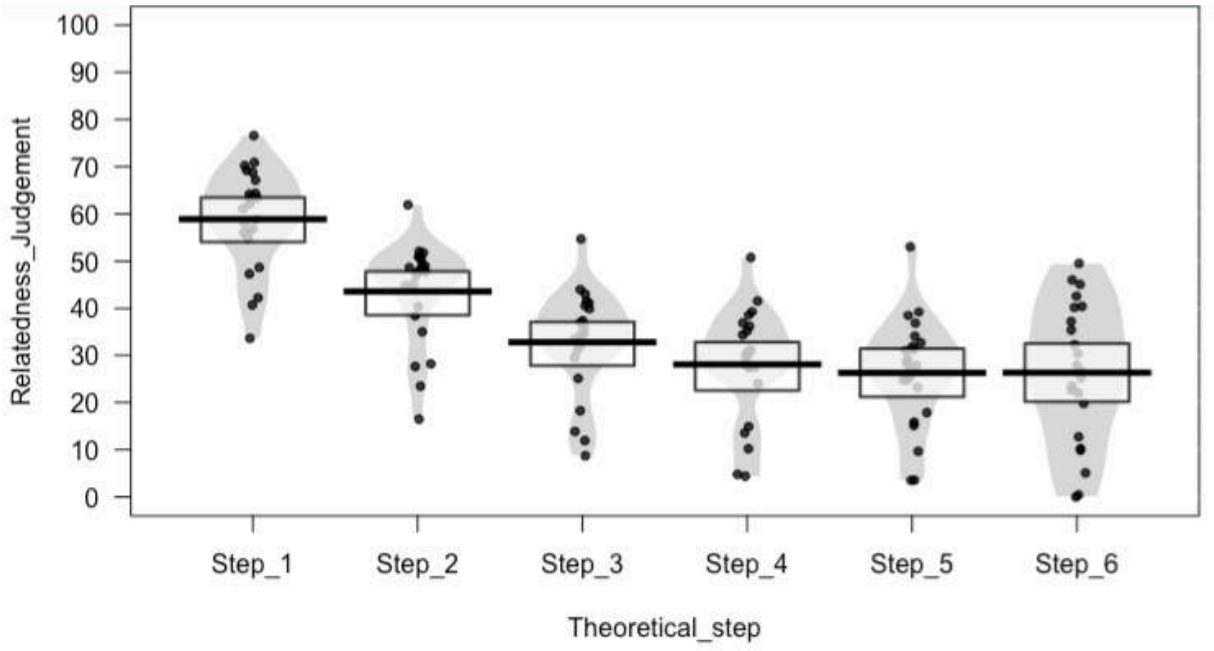


Fig. 2: AJT Task validation. Relatedness ratings of the participants are plotted against theoretical distance. X-axis - Theoretical distance according to the French semantic graph. Y-axis - Relatedness ratings given by the participants. Dots symbolize individual mean response ratings; bars show the mean across participants; white bands correspond to the inference representing the 95% of a Bayesian highest density interval; and the grey area displays the smooth density distribution.

The ***ICAA*** questionnaire was used to quantify everyday creative activities and achievements (Diedrich et al., 2018). This questionnaire contains two parts. In the first part, participants answered questions focused on 8 different specific domains. For each domain, the quantification considered aspects related to how many times the participant had carried out a certain activity over the last 10 years, the level of achievement they have attained in the domain and how many years they have engaged in the specific domain. In the second part, participants described the five most creative achievements in their life. The scores *ICAA_1* (creative activities) and *ICAA_2* (creative achievements) were obtained as the total score for part one and part two respectively.

Results

Relatedness Judgments and Theoretical Semantic Distance

Relatedness ratings within each participant ranged from 0 to 100 indicating that participants used the full scale to rate relationships. Overall mean relatedness ratings across participants ranged from 13.49 to 54.02, with a mean of 33.22 (± 8.66) and median of 34.08. For each participant, we found a significant negative correlation between the relatedness ratings and the theoretical distance ($p < .001$) with a correlation coefficient from -.2 to -.3 (Fig. 2).

AJT-based Network Metrics and Creativity

The network metrics were correlated to the creativity measures using Kendall Tau-b. These correlations were done separately for the WUN and UUN metrics. Fig. 3 shows an illustration of two WUN networks, from a high creative and a low creative participant, chosen among participants with respectively the highest vs poorest scores in both *AUT_Originality* and *ICAA_1*.

Significant correlations were found between several metrics from the WUN networks and creativity scores. *ICAA_1* negatively correlated with *D* ($\tau = -.32$, $p < .05$) and *ASPL* ($\tau = -.34$, $p < .05$) and positively with *S* ($\tau = .32$, $p < .05$). *AUT_Originality* negatively correlated with *D* ($\tau = -.45$, $p < .01$), *ASPL* ($\tau = -.41$, $p < .05$) and *BC* ($\tau = -.39$, $p < .05$) and positively correlated with *CC* ($\tau = .35$, $p < .05$). Similar correlations were found between several metrics from the UUN networks and creativity scores. *ICAA_1* correlated negatively with S ($\tau = -.41$, $p < .05$). *AUT_Originality* negatively correlated with *D* ($\tau = -.51$, $p < .01$), *ASPL* ($\tau = -.44$, $p < .05$), *BC* ($\tau = -.46$, $p < .01$), *S* ($\tau = -.49$, $p < .01$) and *Q* ($\tau = -.38$, $p < .05$). All *p*-values reported above are uncorrected and did not survive an FDR correction.

AJT behavior and Creativity Scores

To test whether creativity also relates more directly to AJT behavioral measures (Rossman & Fink, 2010), Pearson correlations were computed between the creativity scores and AJT relatedness ratings, overall and separately for each

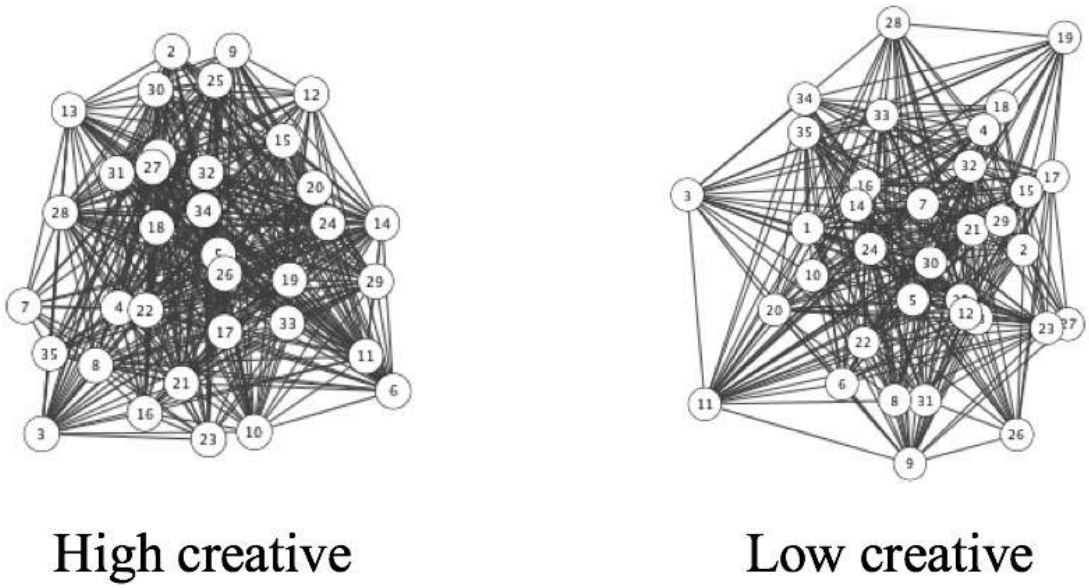


Fig. 3: Example individual semantic networks of a low and high creative participant (weighted undirected networks). Circles represent nodes (single words, labelled as numbers), grey lines represent the edges connecting the nodes, with higher weighted edges having a shorter length representing higher semantic relatedness. The high and low creative participants were chosen among participants with the highest and lowest scores in creativity assessments, respectively (*AUT_Originality* and *ICAA 1*).

theoretical step from 1 to 6. The mean AJT relatedness ratings correlated positively with *AUT_Originality*, $\rho(19) = .54$, $p < .05$: Participants judging word pairs as more related overall produced more original ideas at the AUT task. Correlations with the other creativity scores were not significant.

When analyzed separately for each theoretical distance, AJT relatedness judgement correlated with *AUT_Originality* for word pairs separated by 6 steps, $\rho(19) = .55$, $p < .01$, 5 steps, $\rho(19) = .46$, $p < .05$, 4 steps, $\rho(19) = .46$, $p < .05$, 3 steps, $\rho(19) = .49$, $p < .05$ and 2 steps, $\rho(19) = .56$, $p < .01$. Participants judging theoretically distant word pairs (step >= 2) as more related produced more original ideas at the AUT task. AJT relatedness ratings for close word pairs (1 step apart) correlated positively with *CAT_Solving*, $\rho(19) = .50$, $p < .05$, and *CAT_Close*, $\rho(19) = .55$, $p < .001$: Participants judging theoretically close word pairs (1 step apart) as more related were better at combining word associates and solved more CAT trials. No statistical results survive FDR correction for multiple comparisons.

Discussion

This study aimed to investigate the link between individual differences in the organization of semantic associations and creativity using computational methods based on graph theory. Individual semantic networks were estimated using an adapted version of the method from Benedek et al. (2017) by controlling the selection of the words based on computational methods. To this purpose, we first estimated a unique and large-scale semantic network in French. Next, we developed a method allowing to select a set of words in French while controlling for their semantic distance. Then, the selected words were used in a semantic relatedness judgement task and these relatedness ratings were used to estimate individual semantic networks. Several metrics characterizing the structure of these networks were computed and related to creative assessment scores.

Our results showed that the theoretical semantic distance correlated with the relatedness judgments of the participants, thus converging with the results of Kenett et al. (2017). Theoretical distance relies on the properties of a semantic network estimated from a free verbal association task submitted to a large number of independent volunteers and from the similarity between the generated associates of all cue words. This semantic network allows to measure a theoretical distance as the number of steps separating two nodes in the network. That this measure was strongly related to the subjective similarity judgement of our participants between these cue words validate the use of path length computed on such semantic network as a measure of semantic distance (also converging with results from Kenett et al., 2017). However, it is important to note that while the correlations were highly significant, the Kendall τ coefficients were of moderate size (mean of $\pm$ .22). One possibility would be that the relationship between the theoretical distance and rated distance between selected stimuli may not be linear across the full range of steps. In addition, other factors could impact these subjective relatedness ratings. For instance, subjective ratings showed a high inter-individual variability that could in part be explained by creative abilities, as indicated in the second part of our study.

The next step of the current study consisted of a behavioral experiment aiming to examine the relationships between the organization of semantic memory and creative abilities. The findings showed that some network metrics for both WUN and UUN networks were related to creativity measures including the originality of ideas generated during the AUT (*AUT_Originality*) and the creative activities in life assessed with ICAA (*ICAA_1*). However, those network metrics were not significantly correlated to the number of ideas generated in the AUT (*AUT_Fluency*) and creative achievements (*ICAA_2*) measured with the same tasks, nor to CAT scores.

Indeed, the results showed that participants with more original ideas in the AUT and/or more creative activities in their real life (*ICAA_1*), exhibit WUN networks that are less spread out (shorter *D* and *ASPL*), were more clustered (higher *CC*), showed greater small-world connective properties (higher *S*) and the nodes tended to have a more homogeneous connective role in the network (lower *BC*). Similarly, AJT-based UUN networks were also less spread out with shorter path length, less modular (lower *Q*) and with uniform nodes (lower *BC*) but with reduced small-world properties (*S*) in more creative participants.

Importantly, these findings replicate and expand the results from Benedek et al. (2017) who used UUN networks and showed similar correlations between *CC* and *ASPL* metrics and AUT; we additionally observed correlations between the AUT and other metrics (*BC*, *Q*, *D*). In WUN networks additional correlations were shown between network metrics (*D* and *ASPL*) and *ICAA_1*. These correlations indicate that the organization of semantic memory measured by network metrics is also a relevant factor in real life creativity.

Overall, the current findings suggest that more creative participants exhibit a more clustered and densely connected semantic network whereas less creative participants have a more spread out and fragmented network. These results are in line with the few previous studies that examined semantic memory and creativity (Benedek et al., 2017; Kenett et al., 2014; 2018; Kenett & Faust, 2019). Together these studies strengthen the view that exploring the organization of semantic associations using individual networks is both relevant and valuable for the neuroscience of creativity and support the associative theory of creativity.

Additionally, our method allows us to explore the relationships between AJT ratings and creativity measures. The AJT ratings averaged across all theoretical distances and separately for each theoretical distance greater than 1 was positively correlated with originality in AUT. This finding indicates that participants who produced more original ideas also identified word pairs as more related, especially for pairs of words being theoretically more distant. This result is consistent with Rossmann and Fink (2010) that showed a positive relationship between originality and the evaluated associative distance between unrelated word pairs. These findings suggest that creative people are able to perceive connections between concepts that others may not see. Conversely, the mean AJT rating for theoretically close word pairs (1 step apart) positively correlated with the total number of correct responses in CAT when considering all trials (*CAT_Solving*) or close trials only (*CAT_Close*). Participants who found close links in the CAT also judged theoretically close pairs as highly related. However, contrary to what was expected, the correlations with AJT ratings failed to reach significance when considering distant trials only (*CAT_Distant*) or the difference in performance between distant and close trials (*CAT_Index*). It is possible that distant trials involve additional processes that are less dependent on semantic associations (Bendetowicz et al., 2018). We cannot rule out that the small number of CAT trials used in this experiment may have influenced this result.

Finally, our results indicate that network metrics provide insight why people rate concepts as more or less related and that they are relevant quantitative measures to study creativity. However, statistical analyses revealed that no correlation with creativity scores survived the corrections for multiple comparisons. The small sample size may explain the lack of power, and more participants will be included in this study to address this issue. Nevertheless, the trends in the results and their consistency with previous studies are encouraging. Overall, the findings suggest that exploring individual semantic networks based on a controlled verbal material is a promising approach to study creativity.

Conclusions and Perspectives

To conclude, our data indicate a close relationship between the organization of semantic associations represented by semantic networks and creative abilities. Although the results will need to be confirmed in a larger sample, which is an ongoing project, the current study is consistent with previous studies performed in this field and offers improvements in semantic network methods. Our results are notably in agreement with previous studies that showed a link between creative abilities and the ability to make semantic connections between unrelated concepts. Developing new methods to measure the ability to make new semantic connections is an important challenge to better understand the mechanisms of creative cognition (Benedek & Fink, 2019). The analysis of individual semantic networks is one of the most promising approaches to achieve this goal.

The results of this study also offer interesting hypotheses to test regarding the brain substrates that underlie creative abilities. For instance, the same paradigm can be combined with functional MRI to explore how brain network activity and connectivity covary with the ability to connect distant concepts as measured by semantic network metrics. Moreover, graph theory can be used to study how brain network connectivity relates to the organization of semantic networks in the context of creativity.

Finally, the current study provides valuable insight regarding the fruitfulness of the newly created French semantic network. This network could be especially useful for measuring the semantic distance of words produced by participants in cognitive tasks or building new French task material in which semantic distance needs to be controlled.

References

Beaty R. E., et al. (2014). The roles of associative and executive processes in creative cognition. *Memory & Cognition*. 42: 1186–1197.

Bendetowicz, D., et al. (2017). Brain morphometry predicts individual creative potential and the ability to combine remote ideas. *Cortex*, 86, 216-229.

Bendetowicz, D., et al. (2018). Two critical brain networks for generation and combination of remote associations. *Brain*, 141, 217-233.

Benedek, M., & Fink, A. (2019). Toward a neurocognitive framework of creative cognition: The role of memory, attention, and cognitive control. *Current Opinion in Behavioral Sciences, 27*, 116-122.

Benedek, M. et al., (2012). Associative abilities underlying creativity. *Psychology of Aesthetics, Creativity, and the Arts*, 6(3), 273-281.

Benedek, M., et al. (2013). Assessment of Divergent Thinking by means of the Subjective Top-Scoring Method: Effects of the Number of Top-Ideas and Time-on-Task on Reliability and Validity. *Psychology of Aesthetics, Creativity, and the Arts*, 7, 341-349.

Benedek, M. et al. (2017). How semantic memory structure and intelligence contribute to creative thought: a network science approach. *Thinking & Reasoning*, 23, 158-183.

Debrenne, M. et al. (2011). Le dictionnaire des associations verbales du français et ses applications. Variétés, variations et forme. Éditions de l'École polytechnique p. 355-366.

Diedrich, J. et al. (2018). Assessment of real-life creativity: The Inventory of Creative Activities and Achievements (ICAA). *Psychology of Aesthetics, Creativity, and the Arts,* 12(3), 304-316.

Kenett, Y. N. (2018). Investigating creativity from a semantic network perspective. In Z. Kapoula, E. Volle, J. Renoult, & M. Andreatta (Eds.), *Exploring Transdisciplinarity in Art and Science* (pp. 49-75). Cham: Springer.

Kenett, Y. N., et al. (2014). Investigating the structure of semantic networks in low and high creative persons. *Frontiers in Human Neuroscience*, 8, 407.

Kenett, Y. N., & Faust, M. (2019). A semantic network cartography of the creative mind. *Trends in Cognitive Sciences*, 23, 271-274.

Kenett, Y. N., et al. (2017). The semantic distance task: Quantifying semantic distance with semantic network path length. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 43, 1470-1489.

Kenett, Y. N. et al. (2018). Flexibility of thought in high creative individuals represented by percolation analysis. *PNAS*, 115(5), 867-872.

Mednick, S. A. (1962). The associative basis of the creative process. *Psychological Review*, 69, 220-232.

Ovando Tellez, M.*, Bieth, T.*, Bernard, M., Volle, E. (2019). Contribution of the lesion approach to the neuroscience of creative cognition. *Current Opinion in Behavioral Sciences*, 27:100-108.

Rossmann, E. & Fink, A. (2010). Do creative people use shorter associative pathways? *Personality and Individual Differences*, 49, 891-895.

Rubinov M. & Sporns O. (2010). Complex network measures of brain connectivity: uses and interpretations. *NeuroImage*,52,1059–1069.

Siew, C. S. Q., Wulff, D., Beckage, N. M., & Kenett, Y. N. (in press). Cognitive network science: A review of research on cognition through the lense of network representations, processes, and dynamics. *Complexity*.

Silvia, P. J., Winterstein, B. P., Willse, J. T., Barona, C. M., Cram, J. T., Hess, K. I., ... & Richard, C. A. (2008). Assessing creativity with divergent thinking tasks: Exploring the reliability and validity of new subjective scoring methods. *Psychology of Aesthetics, Creativity, and the Arts, 2*(2), 68-85.

Volle, E. (2018). Associative and Controlled Cognition in Divergent Thinking: Theoretical Experimental, Neuroimaging Evidence, and New Directions. *In* R. Jung & O. Vartanian (Eds.). *The Cambridge Handbook of the Neuroscience of Creativity* (pp. 333-362). New York, NY: Cambridge University Press.

The Importance of Morally Satisfying Endings: Cognitive Influences on Storytelling in Gillian Flynn's *Gone Girl*

Sarah G. P. Binau (sarah.binau@pomona.edu)
Department of Linguistics and Cognitive Science, Pomona College
333 N. College Way, Claremont, CA 91711 USA

Robin J. Melnick (robin.melnick@pomona.edu)
Department of Linguistics and Cognitive Science, Pomona College
333 N. College Way, Claremont, CA 91711 USA

Jack I. Abecassis (jack.abecassis@pomona.edu)
Department of Romance Languages, Pomona College
333 N. College Way, Claremont, CA 91711 USA

Abstract

Peak End Rule (Kahneman, 1993; 2011) suggests that the average of the peak and end moments of an event disproportionately affect memory and thus perception of the experience. We investigate PER's application to the experience of reading fiction. Gillian Flynn's *Gone Girl* (2012) is an ideal case study because it is commercially popular but, unlike most popular novels, has a distinctly amoral ending. We hypothesize that humans expect moral payoffs at the end of narrative fiction, and that when these expectations are not met (i.e., pain at the end of the experience), as in the case of *Gone Girl*, readers' perceptions of the story will be influenced by this pain and manifest as disappointment and dislike. We reference existing models in evolutionary psychology, which seek to explain human altruism, and models in cognitive science, which seek to explain patterns in memory and assessment. To quantify disappointment and dislike, we conduct a programmatic corpus linguistic analysis of 40,000 web-scraped Amazon product reviews of *Gone Girl,* comparing them to reviews of eight other similarly popular novels from the same year. Results show that reader sentiments about *Gone Girl*, both the overall review ratings and analysis on a sentence-by-sentence basis, are more positive than for the comparison novels. When only reviews mentioning "end" are analyzed, however, the effect reverses, with a similar finding at the more granular level of sentences mentioning "end." These findings support our hypothesis that moral endings, or lack thereof, significantly shape reader perceptions of a novel.

Keywords: peak end rule; narrative endings; sentiment analysis; corpus linguistics; web scraping; Amazon product reviews; morality in narrative; evolutionarily stable systems; social cooperation; *Gone Girl*

Introduction

The principle of Peak End Rule (PER) suggests that a memory of an experience is influenced disproportionately by two key moments: the most intense moment of pain in the episode and the level of pain felt at the end of the episode (Kahneman et al., 1993). Multiple experiments have supported the notion of PER, utilizing a variety of methodologies (Fredrickson & Kahneman, 1993; Kahneman, 2011; Redelmeir & Kahneman, 1996). While Kahneman (2011) studied how subjects evaluate fictional stories of people's lives (ending happily or sadly) in a Peak End Rule framework, it remains to be explored how PER applies to readers' perceptions of literary fiction. In the present study, we explore such perceptions as expressed through reader book reviews. Here, the event to which PER applies is the reading of a novel. Since we assume that such review assessments are highly correlated to perceptions formed during the reading event, perception and memory here are then intertwined.

Literary critics working within an evolutionary framework believe fiction to be inherently moral (Gardner, 1978; Gottschall, 2012). Regardless of how much readers delight in highly problematic narratives, they nevertheless expect a moral payoff by way of closure (Carroll, 2011; Flesch, 2009; Gardner, 1978). Moreover, it has been shown that authors and filmmakers are acutely aware of this expectation and use cognitive biases in the construction of narratives to manipulate the effect on readers (Smith, 2015).

This expectation for moral endings may be rooted in the evolutionary advantages of altruism. Altruistic action is necessary in group systems because it ultimately promulgates group stability (Flesch, 2009). In Evolutionarily Stable Systems, people are cooperators, defectors, or punishers (Flesch, 2009). Cooperators constantly track the behavior of others in their group using gossip, honest signaling, and other social tools to ensure that no one is defecting (Dunbar, 2004; Flesch, 2009; Harari, 2015; Zahavi & Zahavi, 1997). If cooperators catch defectors, they seek to punish the latter by exposing their free-riding behavior to the rest of the group. This same human propensity for altruistic punishment has been demonstrated through the game theory constructs of the Ultimatum Game and the Prisoner's Dilemma (Fehr & Fischbacher, 2003; Güth et al., 1982; Tucker, 1983).

William Flesch argues that the moment the author allows moral characters to expose and punish immoral characters is the "pleasure of fiction" (Flesch, 2009). In fact, it is not just the authors' ability, but their social responsibility to provide this payoff (Carroll, 2011). Reader expectations for this payoff are also influenced by genre: when viewers watch more fictional television, they develop stronger Just World Beliefs, as opposed to when they watch news/infotainment

television (Appel, 2009). This could be because fictional narratives "tend to stimulate moral evaluation," whereas nonfictional television is saturated with immoral behaviors not necessarily followed by moral consequences (Appel, 2009). In this way, genre affects readers' expectations.

A case study: *Gone Girl*

Using the novel *Gone Girl* by Gillian Flynn (2012), we present an emblematic case study for moral endings in fictional narratives. *Gone Girl* is a useful case in that it is commercially popular, topping *New York Times* (NYT) bestseller lists and garnering over 40,000 product reviews on Amazon, but its ending is highly disturbing because of its lack of morality (Amazon Customer Reviews, 2016; New York Times Bestsellers, 2012, 2013).

For readers to engage with a book, they inherently need to trust the narrator (Carroll, 2011). Flynn expertly manipulates this trust to deliver a spectacularly amoral ending, thus experienced by the readers as particularly painful. With two unreliable narrators, Flynn's readers perceive their highly manipulated empathy for the two protagonists, Nick and Amy, as wasted. In exchange for this deception, readers expect Flynn to craft an even stronger, satisfying ending.

However, the novel's ending is both abrupt and lacking in a moral payoff. The book has no altruistic punishers. The defector, Amy, is not punished sufficiently; she succeeds in all of her free-riding efforts. As one reader comments: "I was disappointed in the ending. I was hoping for either justice on Nick's side or the demise of Amy. Ending just wasn't what I had hoped it would be." Given the principle of Peak End Rule, we would expect readers' perceptions and global assessments of the novel to be most influenced by the painful and unsatisfying ending (Kahneman, 2011).

To assess this and other related claims about moral closure in novels, this paper linguistically analyzes Amazon product reviews of *Gone Girl* (Amazon Customer Reviews, 2016). This methodology is derived from an increasing usage of online corpora to assess opinions of a product or work (Allington, 2016; Boot, 2013).

Hypotheses

Readers expect moral endings out of *New York Times* bestselling narrative fiction, and when these expectations are not met, as is the case in *Gone Girl*, readers will feel dissatisfied and disappointed because their unspoken contract with the author has been broken. Specifically, we may expect to observe the following:

(1) Readers of *Gone Girl* will more frequently discuss the ending, as compared to other best-selling literary novels.

(2) While *Gone Girl* is similarly popular (i.e., best-selling), with comparable overall Amazon review numerical ratings (scale: 1-5 "stars"), reviews discussing the "end" will garner significantly lower ratings for *Gone Girl*. Discussing "end" is not expected to similarly affect the comparison corpus.

(3) At the sentential level within *Gone Girl* reviews, mentioning "end" will be associated with significantly more negative surrounding sentiment, as judged by a machine-learning classifier. No similar effect is expected for the comparison corpus.

(4) Descriptively, adjectival collocates of "end" are expected to be substantially more negative within the *Gone Girl* corpus.

Methods

Two corpora were created and analyzed using a combination of web-scraping tools, custom scripting software, part-of-speech tagging, deep-learning sentiment classification, and the R statistics platform (version 3.5.1, R Core Team, 2018). Throughout this work, we consider references to the ending of a novel to be any of "end," "ends," "ended," or "ending," referring to this set collectively as simply "end."

Materials

Star-rated consumer reviews of novels were extracted from the Amazon website and organized into two corpora: one for the novel *Gone Girl* (Flynn, 2012) and one for a group of comparison works.

***Gone Girl* corpus** A total of 39,436 product reviews of the novel *Gone Girl* by Gillian Flynn were extracted in December 2016 using ParseHub web-scraping software. Data collected were: title of review, content of review, date published, and star rating. After a pilot study of the full corpus, final analysis was limited to the first 2,000 reviews chronologically following the book's release, all within 2012, in order to avoid any influence of the movie *Gone Girl* (Fincher, 2014), announced in 2013 and released in 2014.

Comparison corpus The comparison group of novels comprises all works that, like *Gone Girl,* appeared on the NYT bestselling fiction list for two or more consecutive weeks in 2012 or 2013. Our collected corpus contains product reviews for: *11/22/63* by Stephen King, *Fifty Shades of Grey* by E.L. James, *Reflected in You* by Sylvia Day, *The Racketeer* by John Grisham, *A Memory of Light* by Robert Jordan and Brandon Sanderson, *Until the End of Time* by Danielle Steel, *Inferno* by Dan Brown, and *The Cuckoo's Calling* by Robert Galbraith. Amazon reviews were extracted, collecting title of book, plus as before: title of review, content of review, date published, star rating.

Here, the number of reviews extracted for each novel was limited to either the first 2,000 chronologically or all reviews appearing within the first two years of release, whichever was less. As with *Gone Girl*, this was done to limit any potential influence from television or cinematic releases based on the books. Within these parameters, a total of 14,460 reviews were extracted, *11/22/63* and *Until the End of Time* via ParseHub, with the remainder collected using the webscraper.io utility (Balodis, 2018).

All tokens of the phrase "until the end of time," without regard to case, were excluded from analysis, as this is not only a somewhat fixed multi-word expression (MWE) in English, but here it is also specifically the title of one of our comparison novels. We assume the use of "end" in this expression does not actually refer to the ending of the narrative. A second fixed MWE, "loose ends," was similarly excluded.

Results

Use of *end*

Per hypothesis 1, reviewers of *Gone Girl* discuss the ending more. In the comparison corpus, 25.7% of the reviews mention "end," 3.7% of the review titles, and "end" compromises 0.21% of all words. The corresponding figures for *Gone Girl* are 52.4%, 8.4%, and 0.63%, more than double in each case (Figure 1).

Differences by "star" rating

Hypothesis 2 predicts that mentioning "end" will lower Amazon review numerical ratings for *Gone Girl* more so than for the comparison set. Figure 2 appears to bear this out.

To explore this further, we fit a linear mixed-effects regression model to predict review star rating from fixed effects for source corpus, length of review (in words, log-reduced to limit outlier effects), and whether or not each review included mention of "end" in its title or body text, with a random effect for individual book title.[1] We trialed all

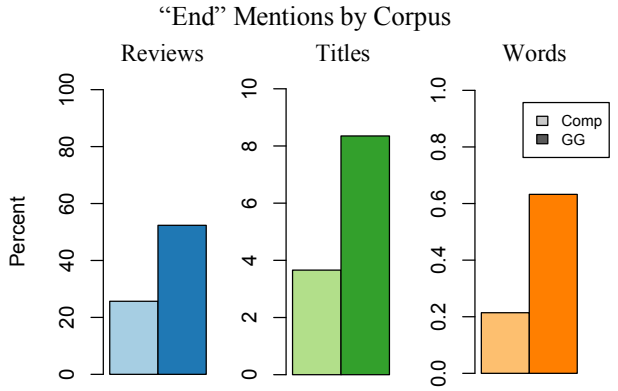


Figure 1: Percentages of "end" mentions for comparison corpus ("Comp", lighter bars) vs. *Gone Girl* corpus ("GG", darker bars).

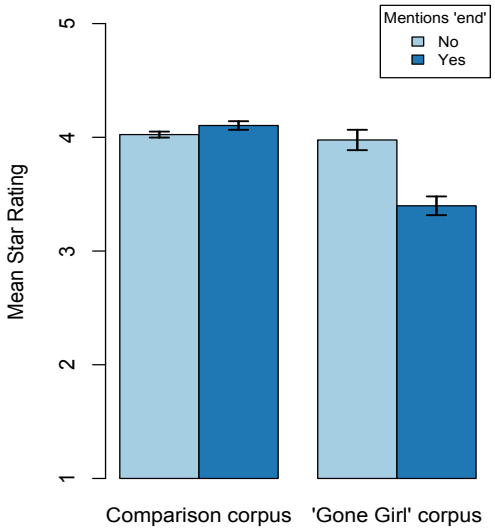


Figure 2: Effects of mentioning "end" on overall reviewer rating. (Comparison corpus *N*=14.5K, *Gone Girl N*=2K.)

possible two-way interactions among the fixed effects, as well as a maximal random effects structure (Barr et al., 2013). Stepwise model optimization preserves the random intercept for book title, the associated random slope for review length, and interactions of source corpus with both review length and "end" mentions (Table 1).[2]

As a baseline, there was no significant difference in overall star rating among these popular works when "end" was not mentioned (main effect of source corpus, $\hat{\beta} = 0.019$, $t = 0.078$, $p = 0.9403$), and longer reviews correlated with somewhat lower ratings overall ($\hat{\beta} = -0.283$, $t = -7.182$, $p = 0.0002$).[3] When "end" is mentioned, mean ratings

Table 1: Fixed effects and interactions from linear regression, with baseline *y*-intercept of about 4 out of 5 stars.

	Coef. $\hat{\beta}$	SE($\hat{\beta}$)	*t*	*Pr*(>|*t*|)
Intercept	3.964	0.081	49.181	< 0.0001
Corpus: *Gone Girl*	0.019	0.242	0.078	0.9403
Review length	-0.283	0.039	-7.182	0.0002
Mentions "end"	0.260	0.026	10.145	< 0.0001
Gone Girl × length	0.353	0.118	3.006	0.0197
Gone Girl × "end"	-0.870	0.064	-13.606	< 0.0001

[1] We do not also include a random effect for "participant" (i.e., reviewer) since, with rare exception, there is just one review per reviewer available within our pair of corpora—that is, there is no clustering to model.

[2] The model exhibits negligible collinearity, with condition number $\kappa = 1.41$ (where Cheney and Kincaid 2007 suggests 10+ is problematic) and largest variance inflation factor (VIF) for any single predictor = 1.31 (10+ again being high, per Hair et al. 1998).

[3] Inclusion of length in this analysis was motivated by an anonymous commenter's concern that lower ratings for *Gone Girl* reviews mentioning "end" might arise if these reviews tended to be longer, where longer reviews might in turn correlate with lower ratings. As reported above, the main effect of review length does turn out to be correlated with lower scores overall, but we find that greater length specifically among the *Gone Girl* reviews actually positively influences star rating (interaction of source corpus and length, $\hat{\beta} = 0.353$, $t = 3.006$, $p = 0.0197$). Further, *Gone Girl* reviews overall are not significantly longer than the comparison group ($\bar{x} = 101.5$ words *GG* vs. 100.6 *comp*, $t = 0.322$, $p = 0.7473$), and *Gone Girl* reviews mentioning "end" are actually significantly shorter on average than reviews mentioning "end" within the comparison corpus ($\bar{x} = 123.0$ vs. 160.4, $t = -7.027$, $p < 0.0001$).

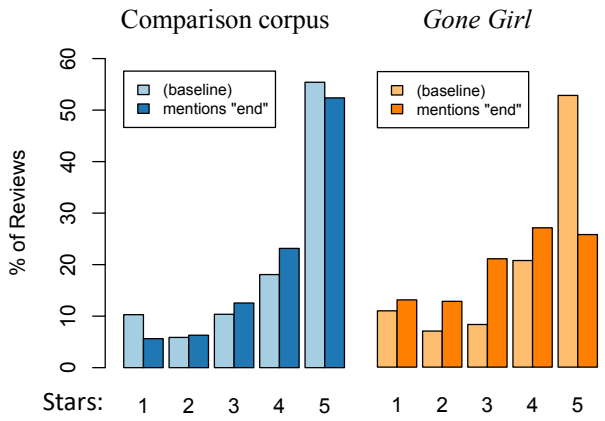


Figure 3: Star ratings by source, with and without "end."

actually increase overall ($\hat{\beta} = 0.260$, $t = 10.145$, $p < 0.0001$), but looking at the interaction terms, among *Gone Girl* reviews that mention "end," scores significantly decline ($\hat{\beta} = -0.870$, $t = -13.606$, $p < 0.0001$) as predicted.

Figure 3 further illustrates the distribution of star ratings with and without "end" mentions.

Sentiment analysis

The analyses detailed above offer a view of how mentioning "end" is reflected in an extrinsic measure of overall reviewer satisfaction—the star ratings. But a question remains with regard to what reviewers express more directly about the ending of the story.

To explore such an intrinsic measure, we began by splitting review text at punctuation points, a proxy for clause boundaries, yielding 136K such segments, including 12,277 for *Gone Girl*, of which 1,066 mention "end." We applied deep-learning sentiment annotation via the Stanford CoreNLP toolkit (Socher et al., 2013; Manning et al. 2014) to rate the emotional content of each segment of text ("clause") as "Very negative," "Negative," "Neutral," "Positive," or "Very positive."

Recasting these ratings as a continuous scale (-2 to +2), we once again fit a linear mixed-effects regression model, now predicting sentiment score from fixed effects for source corpus, clause length, and whether or not each such individual segment of text included mention of "end." To these, we added random effects, now both for book title and for individual review.[4] As before, we also model all possible two-way interactions of the fixed effects and the maximal random-effects structure supported by the data. Stepwise optimization this time preserved all main effects and their interactions (Table 2), as well as both of the random intercepts and a random slope for the separate effect of "end"-mention within each given review.[5]

Table 2: Fixed effects and interactions from regression, with baseline *y*-intercept = 0.02 (i.e., overall "Neutral" sentiment).

	Coef. $\hat{\beta}$	SE($\hat{\beta}$)	*t*	*Pr*(>|*t*|)
Intercept	0.020	0.038	0.515	0.6223
Corpus: *Gone Girl*	0.068	0.114	0.593	0.5720
Clausal length	-0.181	0.003	-61.121	< 0.0001
Mentions "end"	0.052	0.021	2.462	0.0139
Gone Girl × "end"	-0.351	0.041	-8.575	< 0.0001
Gone Girl × Length	0.025	0.010	2.406	0.0161
"End" × Length	-0.048	0.022	-2.149	0.0317

At this clausal level, there was again no significant main effect of source corpus, meaning segments of *Gone Girl* reviews overall were neither more positive nor more negative than those in the comparison corpus ($\hat{\beta} = 0.068$, $t = 0.593$, $p = 0.5720$). We also find that longer clauses are more likely to bear negative sentiment (main effect of length, $\hat{\beta} = -0.181$, $t = -61.121$, $p < 0.0001$), much as our previous analysis found that longer complete reviews received lower star ratings. This effect was, however, weakened within *Gone Girl* review prose (interaction term, $\hat{\beta} = 0.025$, $t = 2.406$, $p = 0.0161$). Among all clauses mentioning "end," mean sentiment increases overall ($\hat{\beta} = 0.052$, $t = 2.462$, $p = 0.0139$), much as did full-review star ratings. Critically, though, *Gone Girl* "end"-mentions are significantly more negative (interaction $\hat{\beta} = -0.351$, $t = -8.575$, $p < 0.0001$), just as we saw with lower star ratings for full reviews, again as predicted by Hypothesis 3.

Most common descriptive terminology

Finally, we examined adjectival collocates of "end" to see how reviewers specifically describe the respective endings. Here, we began by applying part-of-speech (POS) labels to all text, once again using Stanford CoreNLP (Toutanova et al., 2003; Manning et al., 2014), then analyzed adjectives appearing within a three-word window of "end," left or right, without crossing clausal or sentential boundaries. Negated contexts (e.g., "not good," "never disappointing") were excluded from this portion of the analysis (Pang et al., 2002).

The most frequent adjective used to describe "end" in the *Gone Girl* corpus was "disappointing," which in combination with the related form "disappointed" was almost twice as frequent as the next most used word to describe the ending of the story, "worst" (Table 3).

Examples from the *Gone Girl* reviews of the use of "disappointing" or "disappointed" in relation to "end" include: "the ending was disappointing to the point of making me wish I had not spent my time reading this," "it was a good book, but the ending was a disappointment," and "the big disappointment was the ending…this book ended horribly."

[4] Whereas our previous analysis—star ratings applied to the full reviews—did not include a per-review (i.e., "participant") effect.

[5] As with our earlier model, we find negligible collinearity, with $\kappa = 1.15$ and largest VIF for any single predictor = 1.77.

Table 3: Most common adjectives used in descriptions of "end" in the *Gone Girl* corpus. Coding for sentiment as rated by the Stanford CoreNLP toolkit (Socher et al., 2013; Manning et al. 2014): red/– = negative; green/+ = positive; gray/~ = neutral.

Adjective	Frequency (%)
– disappointing	5.08
– worst	3.81
+ good	3.60
– horrible	2.97
– unsatisfying	2.97
~ little	2.75
– awful	2.54
+ happy	2.54
– terrible	2.33
– bad	2.12
– disappointed	1.91

The adjective most frequently used to describe the ending in the comparison corpus was "great" (Table 4). Excerpts from reviews including use of "great" to describe "end" in the comparison corpus include: "great story...with a literally killer climax and great ending," "what a great ending to a terrific trilogy," "the ending was great and totally unexpected," and "awesome book, great ending to an epic saga."

While these respective sets of most frequent adjectives appear largely disjoint, we find a few terms in common among those frequently applied in both corpora, e.g., "good" and "little." To then further explore which terms most distinctively apply to "end" in one corpus as compared to the other, we examined relative frequency across the two data sets. We log-reduced values and *Z*-score normalized for comparison, then found the ratio of "end"-description frequency for *Gone Girl* over the comparison set. The largest ratios (Table 5) represent terms most frequently applied to "end" for *Gone Girl* vs. the comparison group.

Table 4: Most common adjectives used in descriptions of the endings of comparison novels. (Sentiment coding as in Table 3.)

Adjective	Frequency (%)
+ great	12.36
+ good	5.53
+ happy	3.76
+ perfect	2.69
+ wonderful	2.23
~ little	2.07
+ amazing	2.00
+ satisfying	2.00
+ fantastic	1.69
~ first	1.69
+ excellent	1.53

Table 5: Adjectives most distinctively applied to "end" in *Gone Girl* vs. comparison novels, as measured by relative frequency across the two. (Sentiment coding as in Table 3.)

Adjective	GG/Comp ratio
– awful	24.54
– worst	18.40
– anticlimactic	15.34
– flat	12.27
– bizarre	9.20
– disappointed	9.20
– unsatisfying	8.18
– ridiculous	7.67
– atrocious	6.13
– terrible	6.13
– horrible	6.13

Gone Girl corpus examples of "awful"—24.54 times more likely to be used to describe "end" for *Gone Girl* than for the comparison set—include: "slow pace and awful ending—don't waste your money," "great book, awful ending," "it could've been a perfect novel, but that God awful ending!" and "awful ending…makes you feel dirty for wasting so much time on a sick, twisted book."

Finally, we further visualized this notion of relative frequency ratio, focusing on the 20 adjectives with greatest average "end"-collocation frequency across the two corpora. Figure 4 graphs their respective (log-reduced, normalized) frequencies for the two data sets, with sentiment labeling once again as above in Table 3. Terms below the dividing line were more frequently "end" collocates for *Gone Girl,* none expressing positive sentiment. Those above the diagonal, none of which express negative sentiment, were more frequently applied to "end" for comparison works.

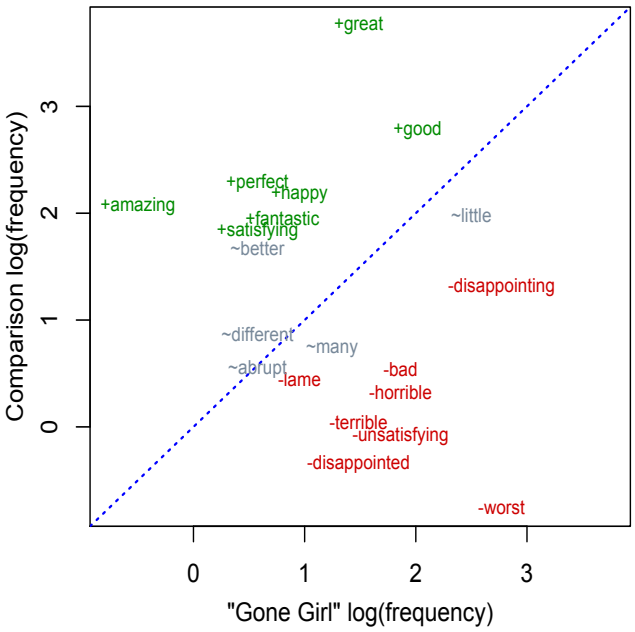


Figure 4: Relative frequency of collocates across corpora.

Discussion

Moral payoffs are closely related to the pleasure of fiction. Readers make an unspoken contract with the author before reading a book: I will read your book, expecting that the morally problematic content will be cleaned up by the end. When this baseline assumption is violated, as is the case in *Gone Girl*, the ending becomes a salient discussion topic. Our findings confirm this: mentions of "end" were twice as common in *Gone Girl* reviews than comparison novel reviews.

And readers do not merely discuss the ending: they vividly express their disappointment. In fact, "disappointed" was used almost 10 times more often in relation to the ending of *Gone Girl* than for the comparison novels. This adjective in particular is revealing in the context of the reader/author contract, because it signifies readers' expectations and their lack of fulfilment.

In contrast, the adjectives most frequently used to describe "end" in our comparison group of novels were generally positive, suggesting satisfaction with the ending and, therefore, fulfillment of the unspoken contract with the author.

Gone Girl is an important case study in the expectation for moral payoffs in novels because it lacks a moral ending, yet it is commercially popular. In fact, in terms of overall review scores, *Gone Girl* is just as highly rated as comparison best-selling novels. What stands out, however, is that when reviews are separated into those that mention "end" and those that do not, *Gone Girl* reviews significantly differ from comparison reviews. Crucially, we found that while including a mention of "end" in the comparison corpus does not significantly affect star rating, mentioning "end" in the *Gone Girl* corpus significantly lowers ratings. This suggests that *Gone Girl*'s amoral ending is a salient cause for overall dissatisfaction with the novel.

While this conclusion is made on the basis of reviewers' overall star ratings, an extrinsic measure, we found similar evidence in programmatic sentiment analysis of review prose. Here, the effect of mentioning "end" was even more pronounced: *Gone Girl* commentary about matters other than "end" was significantly more positive than in reviews of comparison works, while discussions of "end" were significantly more negative.

The fact that discussions of "end" were extremely negative in the *Gone Girl* corpus suggests that reviewers act as altruistic punishers in the framework of Evolutionary Stable Systems. Because Flynn controls the narration that explains the world of *Gone Girl*, readers expect her to act as the primary altruistic punisher in the novel's social ecosystem. If there is a defector within the novel, it is Flynn's job to guide the narration such that the defector is exposed and consequently punished. In the case of *Gone Girl*, there is more than one defector, and Flynn fails to craft the narration to punish any of them appropriately. Readers experience and react to Flynn's lack of punishment and become second-order altruistic punishers themselves; they go to Amazon and write product reviews for the novel, explaining their opinions of the ending and exposing Flynn herself as a defector for her failure to include a moral ending.

Therefore, we see that the evolutionary advantages of altruistic behavior not only guide immediate social groups, but literary communities as well. The principle of Peak End Rule further compounds this: when psychologically amoral, painful moments occur at the end of experiences or narratives, the impact is amplified. This heightening of moral discomfort is what drives readers to write over 40,000 product reviews for a novel, over half of which include a discussion of the "end," with "disappointing" as a key descriptor.

Future Directions

Given the wealth of data available via Amazon, further investigation using more novels and more reviews is an important next step. We would also like to consider reviews from other Internet sources, such as GoodReads.

Another possibility would be to examine responses to the movie *Gone Girl*, looking at how these compare with reactions to the novel. Although Flynn wrote the screenplay for the movie, its narration style inherently changes with the shift in medium, meaning that reactions to the ending could shift as well.

Conclusion

The Peak End Rule principle suggests that endings of experiences or narratives significantly affect a person's overall memory, or perception, of that event. Though first explored through cognitive research on human memory, the principle also applies to readers' perceptions of literary novels, and in particular, their endings. That is, moral endings, or lack thereof, as in the case of *Gone Girl*, have a strong effect on readers' perceptions of novels.

In light of the human propensity for cooperation, *Gone Girl*'s lack of a moral ending dramatically affects reader response, in comparison with other popular contemporary novels. This investigation demonstrates the profound effect endings have in shaping conception of stories, as well as our expectation for a moral payoff in literary novels—even disturbing ones.

Acknowledgements

We are grateful for the helpful comments from anonymous reviewers and from the attendees of a presentation of this work at Pomona College. Any errors remain our own. Statistical calculations and graphics were developed with R version 3.5.1 (R Core Team, 2018).

References

Allington, D. (2016). 'Power to the reader' or 'degradation of literary taste'? Professional critics and Amazon customers as reviewers of *The Inheritance of Loss*. *Language and Literature*. 25(3): 254–278.

Amazon Customer Reviews. (2016). *Gone Girl* by Gillian Flynn. Retrieved from http://a.co/d/4MPuLjL.

Appel, M. (2008). Fictional narratives cultivate just-world beliefs. *Journal of Communication*, 58: 62-83.

Balodis, M. (2018). Web Scraper [Computer software]. Retrieved from https://chrome.google.com/webstore

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3), 255-278.

Boot, P. (2013). The desirability of a corpus of online book responses. In *Proceedings of the Workshop on Computational Linguistics for Literature* (pp. 32–40). Association for Computational Linguistics.

Carroll, N. (2011). Narration. In P. Livingston & C. Plantinga (Eds), *The Routledge Companion to Philosophy and Film*. London: Routledge.

Cheney, E. W., & Kincaid, D. R. (2012). *Numerical Mathematics and Computing*. Cengage Learning.

Coplan, A. (2011). Empathy and character engagement. In P. Livingston & C. Plantinga (Eds), *The Routledge Companion to Philosophy and Film*. London: Routledge.

Dunbar, R. I. M. (2004). Gossip in evolutionary perspective. *Review of General Psychology, 8*(2), 100-110.

Fehr, E., & Fischbacher, U. (2003). The nature of human altruism. *Nature International Journal of Science,* 425, 85-791.

Fincher, D. (Producer/Director). (2014). *Gone Girl* [Motion picture]. United States: Twentieth Century Fox.

Flesch, W. (2009). *Comeuppance.* Harvard University Press.

Flynn, G. (2012). *Gone Girl: A Novel*. NYC: Crown.

Fredrickson, B. L., & Kahneman, D. (1993). Duration neglect in retrospective evaluations of affective episodes. *Journal of Personality and Social Psychology, 65*(1), 45-55.

Gardner, J. (1978). *On Moral Fiction*. NY: Basic Books.

Gottschall, J. (2012). Morality. *The Storytelling Animal: How Stories Make Us Human*. Boston: Houghton Mifflin Harcourt, 117-138.

Güth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization*, 3, 367-388.

Hair, J., Andreson, R., Tatham, R., & Black, W. (1998). *Multivariate Data Analysis* (5th ed.). Prentice-Hall, Inc.

Harari, Y. (2015). The tree of knowledge. In *Sapiens: A Brief History of Humankind*. Purcell, J., & Watzman, H., Trans. New York: Harper, 20-39.

Kahneman, D., Fredrickson, B., Schreiber, C., & Redelmeier, D. (1993). When more pain is preferred to less: Adding a Better End. *Psychological Science*, 4(6): 401-05.

Kahneman, D. (2011). *Thinking, Fast and Slow*. New York: Farrar, Straus, and Giroux.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55-60.

New York Times Bestsellers. (2013). *Fiction*. Retrieved from https://www.nytimes.com/books/best-sellers/.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing* (pp. 79-86). Association for Computational Linguistics.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Redelmeier, D., & Kahneman, D. (1996). Patients' memories of painful medical treatments: real-time and retrospective evaluations of two minimally invasive procedures. *Pain*. 66 (1): 3-8.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (pp. 1631-1642). Association for Computational Linguistics.

Smith, J. (2015). Filmmakers as folk psychologists: How filmmakers exploit cognitive biases as an aspect of cinematic narration, characterization, and spectatorship. In *The Oxford Handbook of Cognitive Literary Studies*. Oxford University Press.

Toutanova, K., Klein, D., Manning, C., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of HLT-NAACL 2003*, pp. 252-259.

Tucker, A. (1983). The mathematics of Tucker: A sampler. *Mathematical Association of America*. 14 (3): 228.

Zahavi, A., & Zahavi, A. (1997). *The Handicap Principle: A Missing Piece of Darwin's Puzzle*. New York: Oxford University Press.

Integrating Common Ground and Informativeness in Pragmatic Word Learning

Manuel Bohn
bohn@stanford.edu
Department of Psychology
Stanford University
LFE
Leipzig University

Michael Henry Tessler
tessler@mit.edu
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology

Michael C. Frank
mcfrank@stanford.edu
Department of Psychology
Stanford University

Abstract

Pragmatic inferences are an integral part of language learning and comprehension. To recover the intended meaning of an utterance, listeners need to balance and integrate different sources of contextual information. In a series of experiments, we studied how listeners integrate general expectations about speakers with expectations specific to their interactional history with a particular speaker. We used a Bayesian pragmatics model to formalize the integration process. In Experiments 1 and 2, we replicated previous findings showing that listeners make inferences based on speaker-general and speaker-specific expectations. We then used the empirical measurements from these experiments to generate model predictions about how the two kinds of expectations should be integrated, which we tested in Experiment 3. Experiment 4 replicated and extended Experiment 3 to a broader set of conditions. In both experiments, listeners based their inferences on both types of expectations. We found that model performance was also consistent with this finding; with better fit for a model which incorporated both general and specific information compared to baselines incorporating only one type. Listeners flexibly integrate different forms of social expectations across a range of contexts, a process which can be described using Bayesian models of pragmatic reasoning.

Keywords: Pragmatics; Word learning; Common ground; Bayesian models

Introduction

One of the most astonishing features of natural language is that it allows us to communicate precise meanings despite the fact that most utterances are inherently ambiguous. While the conventional mapping between sounds (words) and objects constrain what a speaker may mean by an utterance, the intended meaning of the utterance is not reducible to the words that are contained in it. It takes additional pragmatic inference to recover the intended meaning (Levinson, 2000).

Pragmatic inferences rest on a set of expectations that interlocutors bring to the table when entering a communicative interaction. On the one hand, speakers and listeners have the general expectation that their partner communicates in an informative and relevant way (Sperber & Wilson, 2001). Grice (1991) summarised this expectation via the *Cooperative Principle*: "Make your contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged." Importantly, the second half of the Cooperative Principle describes a second type of expectation: interlocutors expect each other to produce and interpret utterances *in light of the shared common ground between them* (H. H. Clark, 1996). Common ground refers to bits of information that are assumed to be shared, either because they were mentioned over the course of the conversation or grounded through some form of joint experience (Bohn & Koymen, 2018). Note that by its very nature, common ground may vary with the individuals involved in a particular conversation.

These same general and specific expectations can support children's word learning (E. V. Clark, 2009; Tomasello, 2009). On the one hand, children have been found to learn novel words by assuming speakers are generally informative (Frank & Goodman, 2014). That is, in the absence of any prior interaction with the speaker, children interpreted a novel word as referring to the most informative referent. On the other hand, children use conversation-specific common ground expectations to decide which object a specific speaker is referring to when they use a novel word (Akhtar, Carpenter, & Tomasello, 1996). For example, when a speaker expressed preference for a particular object, children expect a novel word from the same speaker to refer to the previously preferred object (Saylor, Sabbagh, Fortuna, & Troseth, 2009).

But how do listeners integrate general and common ground-related expectations during word learning? Are pragmatic inferences strengthened additively when both support a particular interpretation? How are they weighed when they are in conflict? The Rational Speech Act (RSA) framework (Frank & Goodman, 2012; Goodman & Frank, 2016) offers a formal framework for addressing this information integration problem. RSA models are characterized by a recursive structure in which a pragmatic listener tries to uncover a speaker's intended meaning by assuming the speaker chose their utterance in order to get a naive listener to recover their intended meaning. RSA models have made accurate quantitative predictions about various forms of pragmatic language use and word learning (Goodman & Frank, 2016). However, a comprehensive treatment of how general and common ground expectations are integrated is still missing.

Within RSA models, each agent in the recursion is modeled as a Bayesian reasoner; thus, information integration is treated as a process of probabilistic inference. The speaker-general informativeness expectation is already encoded in the structure of the model: Speakers produce utterances to aid the listener in disambiguating referents. We operationalize speaker-specific, common ground information as the shared prior probability of referents in the context of the utterance.

Thus, a natural locus for information integration within these models is the trade off between the prior probability of a particular referent and the likelihood of that referent given the current utterance.

Here we evaluate this rational, pragmatic account of information integration. We isolate speaker-specific and common-ground information experimentally, then test how adult listeners combine them in a word learning setting. In Experiments 1 and 2, we replicate findings showing that listeners expect speakers to a) produce informative utterances (Experiment 1) and b) communicate about things that are relevant to common ground (Experiment 2). Based on these results, we generate model predictions using the RSA framework about how these two components should be integrated. In Experiment 3, we test how listeners integrate their expectations and compare model predictions to empirical data. Experiment 4 replicates and extends Experiment 3 by varying the strength of common ground assumptions. For all experiments, we pre-registered the sample size, experimental design and the statistical analysis. For Experiment 3 and 4, we also registered the model structure and predictions (see [masked for peer review])

Method

General Design

Experiments were conducted online using Amazon's Mechanical Turk. Fig. 1 provides a schematic overview of the setup and experimental procedures. The instructions informed participants that they would see a number of animal characters asking for novel toys. Participants were asked to identify the toy being requested by a particular animal. The basic layout involved two tables with toys on them, located left and right of a little hill, on which the animal was standing. For each animal, we recorded a set of utterances (one speaker per animal) that were used throughout the experiments to provide information and make requests. At test, toys were requested using the following utterance: "Oh cool, there is a [non-word] on the table, how neat, can you give me the [non-word]?". Participants responded by clicking on one of the toys. Each experiment started with two training trials in which animals requested familiar objects (car and ball).

Experiment 1

Participants, Design and Procedure

All participants were recruited from Amazon Mechanical Turk and had US IP addresses. Sample size in each experiment was planned to be 120 data points per cell. Experiment 1 had 40 participants. In the test condition, one table contained one object of type A and the other table contained one object of type A and one of type B (see Fig. 1, left). On each trial, the animal introduced themselves (e.g. "Hi, I'm Dog"), turned towards the table with the two objects and made a request. If listeners expect speakers to produce informative utterances, they should select object B. This choice follows from the counterfactual inference that if the (informative) speaker would have wanted to request A, they would have turned to the table that only contained A. On the other hand, since B is only located on the table together with A, there was no alternative way to request B in a less ambiguous way. In the control condition, both tables contained two objects, one of which was randomly determined as the correct one. No inference was therefore licensed. Each participant received three trials in each condition for a total of six trials, presented in a randomized order.

Results and Discussion

Participants selected the less frequent object above chance in the test condition ($t(39) = 5.51$, $p < .001$, see Fig. 2) and did so more often compared to the control condition (generalized linear mixed model (GLMM[1]):$\beta = 1.28$, se = 0.29 $p < .001$). This result replicates earlier work (Frank & Goodman, 2014) and is consistent with the hypothesis that listeners expect speakers to communicate in an informative way.

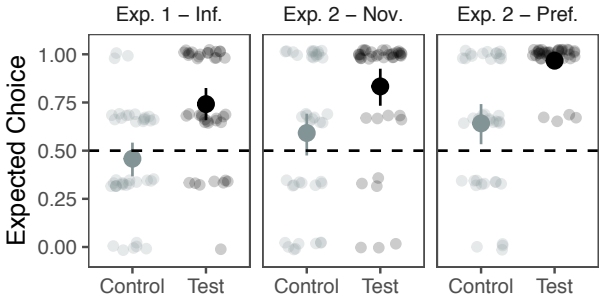


Figure 2: Results from Experiment 1 and 2. For preference and novelty, control refers to a different speaker (see Fig. 1). Transparent dots show data from individual participants, solid dots represent condition means, error bars are 95% CIs. Dashed line indicates performance expected by chance.

Experiment 2

We manipulated common ground expectations based on procedures that have successfully been used in developmental studies (e.g. Akhtar et al., 1996; Saylor et al., 2009). Speakers either expressed preference for one object or one object was new to the speaker.

Participants, Design and Procedure

We collected data from 80 participants, with 40 in each condition. In the preference condition, each table had a different object. In the beginning, the animal appeared on the hill and introduced themselves. Next, they turned to one of the tables and expressed either that they liked ("Oh wow, I really like that one") or disliked ("Oh bleh, I really don't like that one") the object before turning to the other side and expressing the respective other attitude. Then, the animal disappeared. After a short period of time, either the same or a different animal appeared and requested an object while facing straight ahead

[1]All models had maximal random effects structure conditional on model convergence.

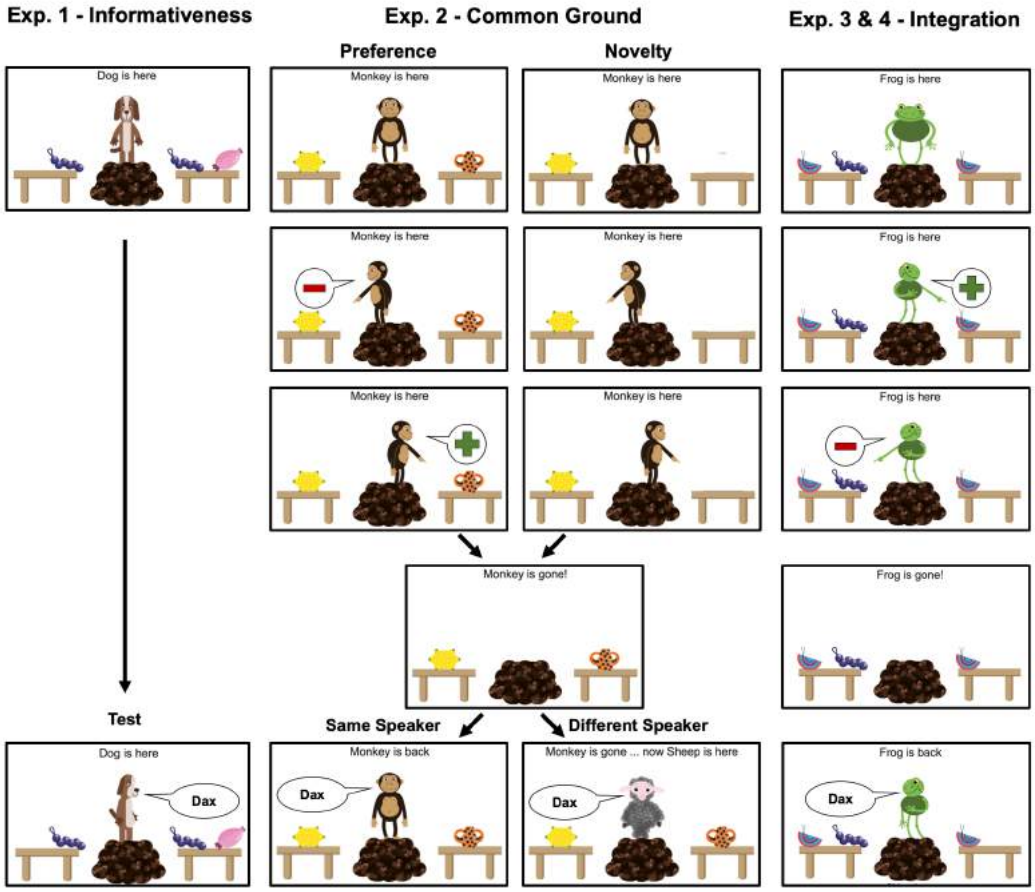


Figure 1: Schematic experimental procedure. In all conditions, at test (bottom), the speaker ambiguously requested an object using a non-word (e.g. dax). Participants clicked on the object they thought the speaker referred to. Informativeness (Experiment 1, left) translated to making one object less frequent in context. Common ground (Experiment 2, middle) was manipulated by making one object prefered by or new to the speaker. Green plus signs represent utterances that expressed preference and red minus of dispreference (see main text for details). Experiment 3 (right) combined manipulations. When expressing e.g. preference for an object on a table with two objects (panel 3 from top), the respective object was temporarily enlarged. Condition for Experiment 3 shown here: preference - same speaker - incongruent.

(see Fig. 1, middle). If participants took into account the information they gained about the speaker, they should select the previously preferred object if the returning animal was the same. If a different animal returned, they should choose randomly between objects.

In the novelty condition, one table was initially empty while there was an object on the other table (see Fig. 1). The animal turned to one of the sides and commented either on the presence ("Aha, look at that") or the absence of an object ("Hm, nothing there"). Next, the animal disappeared. The same animal re-appeared and the sequence above was repeated. When the animal disappeared for the second time, a second object appeared on the empty table while the animal was away. Like in the preference condition, we now manipulated if the same animal or a different one returned. In case of the same animal returning, listeners could infer the referent of the subsequent request by considering that one object was new to the speaker and therefore more likely to be of interest to them. However, no such inference was licensed when a different animal returned because both objects were novel.

Results and Discussion

Participants selected the preferred object above chance when the same animal returned ($t(39) = 29.14$, $p < .001$, see Fig. 2) and did so more often compared to trials in which a different animal returned (GLMM: $\beta = 2.92$, se = 0.56, $p < .001$). Thus, listeners inferred the referent of the utterance by considering previous interactions with the speaker. Interestingly, participants transferred preference to some extent from one animal to the other and selected the preferred object above chance when a different animal returned ($t(39) = 2.7$, $p = .01$). In sum, this study shows that adults make comparable infer-

ences to children (Saylor et al., 2009).

The novel object was selected above chance when the same animal returned (t(39) = 6.77, $p < .001$) but not when a different one appeared (t(39) = 1.49, $p = .144$, see Fig. 2). Furthermore, the two conditions differed in the expected direction (β = 6.27, se = 1.96, $p = .001$). Thus, like children (Akhtar et al., 1996), adults used their prior information about the speaker to resolve ambiguity in the utterance.

Experiment 3

In Experiment 3, we combined the expectations studied in Experiment 1 and 2 to see how listeners integrate them.

Participants, Design and Procedure

A total of 121 individuals participated in the experiment. The test situation was the same as in the test condition in Experiment 1 (see Fig. 1, right): One table with object of type A and the other with an object of type A and B. Again, the animal always turned to the table with two objects and ambiguously requested an object. In Experiment 1, the listener had no prior information about each object. In Experiment 3, however, we manipulated common ground expectations in the same way as in Experiment 2. For example, the animal would turn to the table with one object and express that they don't like object A, then turn to the other table and express that they like object B. Next, after quickly disappearing, they would reappear, turn to the table with two objects and make a request.

For each common ground condition, there were 4 conditions in Experiment 3 resulting from the cross of congruent/incongruent informativeness with same/different speaker. If the preferred/novel object was the less frequent one (object B), the two expectations were congruent. If the preferred/novel object was the more frequent one (object A), expectations were incongruent. For each type of expectation alignment, we varied if the same or a different animal returned. Participants either completed the preference or novelty version with two test trials in each of the four conditions. Before discussing the empirical results, we briefly discuss the model we used to predict expectation integration.

Model Predictions

To derive predictions, we used a probabilistic RSA model that simulates a pragmatic listener reasoning about a cooperative speaker who is trying to refer to an object (Frank & Goodman, 2012). The speaker chooses how to refer to the object by reasoning about a naive listener who does not know the labels for the object (Frank & Goodman, 2014). The conditional probability that the listener chooses a referent given an utterance is defined as follows:

$$P_L(r_s|u) \propto P_S(u|r_s)P_S(r_s)$$

Here, $P_S(u|r_s)$ is the likelihood that the speaker will use an utterance u to refer to a specific referent r. It is defined in terms of a utility function $U_S(u;s)$ consisting of the surprisal of u for a naive listener L_0, who interprets u according its literal semantics:

$$P_S(u|r_s) \propto exp(\alpha U_S(u;s))$$

The numerical strength of the expression above depends on a scalar value, α, which can be interpreted as an indicator of how rational the speaker is in choosing utterances (i.e. as α increases, the speaker is more likely to choose the most informative utterance).

The term $P_S(r_s)$ denotes the prior probability that a speaker will refer to a given referent. This probability represents the listeners expectations about the speaker depending on the manipulation (preference or novelty) and the identity of the speaker (same or different speaker).

We used the results from Experiment 1 and 2 to specify α as well as $P_S(r_s)$ in our model. We set α so that a model with uniform priors would predict the average proportion measured in Experiment 1. The prior probability for each object was set to be the proportion with which this object was chosen in Experiment 2[2]. Based on these parameter settings, we predicted the proportion with which listeners will choose the more informative object in each of eight unique conditions mentioned above (see also Fig. 3). We compared the fit of this pragmatic model to two alternative models using Bayes Factors (BF). The first alternative model ignored the speaker specific expectations (uniform prior model) while the second ignored the informativeness inference (prior only model). All models included a noise parameter, reflecting that participants may respond randomly instead of in line with the intended manipulation on a given trial. Noise parameters were estimated based on the data. They range between 0 and 1 and reflect the proportion of responses that are estimated to be random instead of following the pattern predicted by the model.

Results and Discussion

Results are discussed in the form of the proportion with which listeners chose the more informative object (i.e., the object that would be the more informative referent when only considering speaker general expectations). For a comparison to chance within each unique condition see Fig. 3. Combinations of alignment and speaker identity differed in how they influenced participants' responses (GLMM model term: `alignment*speaker`; β = -2.64, se = 0.48, $p < .001$). Fig. 4A shows the mean response in each unique condition compared to the pragmatics model. Model predictions and data were highly correlated (r = 0.96, $p < .001$). Model fit was much better in the model taking into account both types of expectations compared to the uniform prior (BF = 2e+79) or prior only model (BF = 1.8e+34). The inferred noise level in the pragmatics model was 0.27 (95% HDI: 0.21 - 0.34).

[2]Proportions were measured when participants chose between two objects. However, in Experiment 3, three objects were involved. For each object we used the proportion measured in Experiment 2 as the prior probability.This approached spread out the absolute probability mass for each object but conserved the relative relation between objects.

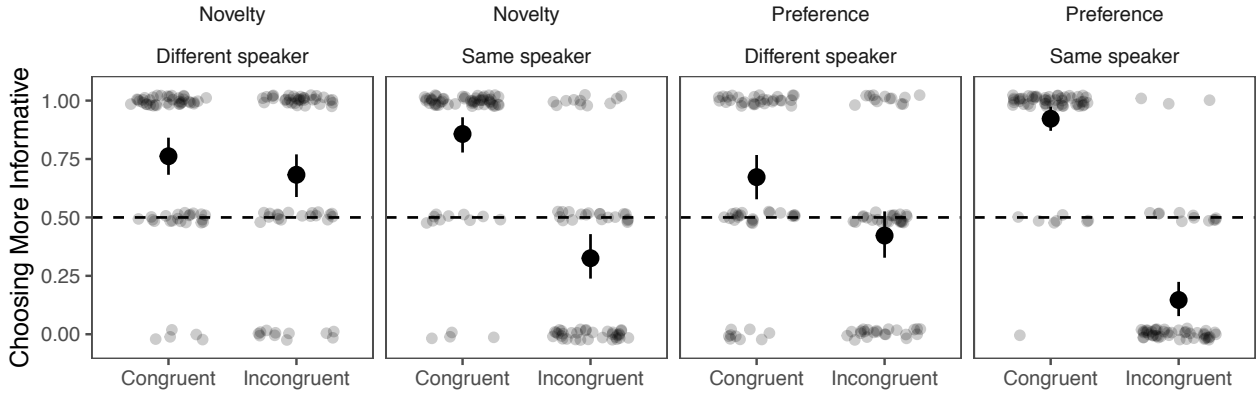


Figure 3: Results from Experiment 3. Dashed line indicates performance expected by chance. Plotting conventions are the same as in Fig. 2. All conditions in which CIs do not overlap with chance line are also statistically different from chance based on two-tailed Wilcoxon tests.

Interestingly, as in Experiment 2, there was a transfer of preference in the case of speaker change. Participants were at chance in the preference - different speaker - incongruent condition (see Fig. 3). If preference would have been specific to a particular individual, participants should have selected the less frequent object above chance (as they did in the corresponding condition with the novelty manipulation). Because it takes into account the measurement from the earlier experiment, our model predicts these results; future work might explicitly model generalization across speakers.

Experiment 4

Here we replicated and extended Experiment 3 by manipulating the strength of the common ground expectations.

Participants, Design and Procedure

This experiment had 453 participants. The structure of the experiment was the same as in Experiment 3. For each common ground expectation (preference and novelty), we intended to have a strong, a medium and a weak condition. The strength of each condition was determined by the proportion with which participants chose the preferred/novel object given the manipulation. We succeeded in generating quantitative variability for novelty. For preference we piloted a number of additional manipulations but did no find one that yielded a weaker preference compared to a medium condition.

The strong manipulations were identical to Experiment 3 and the results are therefore a direct replication (see Fig. 4C). For novelty, in the medium condition, the animal turned to each table only once before the test. In the weak condition, the animal only turned to the table with an object before the test (instead of turning to and commenting on both). In the medium condition for preference, the animal only expressed liking and did so in a more subtle way (saying only: "Oh, wow" while pointing to the object). Participants were assigned to one level of common ground expectation and completed two test trials in each of the four conditions (alignment x speaker change).

Model predictions were obtained in the same way as in Experiment 3; with α inferred from the data and $P_S(r_s)$ measured empirically (in a set of corresponding experiments parallel to Experiment 2).

Results and Discussion

As noted above, the strong prior condition was a direct replication of Experiment 3. Results from the two rounds of data collection were highly correlated (r = 0.97, $p < .001$, see Fig. 4C). Across levels of prior manipulation, the data from Experiment 4 were highly correlated to the corresponding model predictions (r = 0.91, $p < .001$, see Fig. 4B). Again, the pragmatics model provided a much better fit compared to the flat prior (BF = 4.4e+74) or prior only model (BF = 1.8e+84). The inferred noise level in the pragmatics model was 0.28 (95% HDI: 0.24 - 0.32).

Discussion

Language use and learning requires balancing different types of expectations about one's interlocutor - expectations about how speakers behave in general and expectations about how a particular speaker might behave in a particular context. Here we used a Bayesian pragmatics model to predict this integration process. Experiment 1 and 2 replicated previous studies showing that adult listeners expect speakers to produce utterances informatively and also with respect to common ground. We then combined the procedures from the first two experiments to study how listeners would integrate expectations. We used the results from Experiment 1 and 2 to specify model parameters that represented the two types of expectations, generating predictions about new behavior. Experiments 3 and 4 showed that both types of expectations influenced listeners inferences. Overall, listener behavior was accurately described by our model, suggesting that listeners trade-off flexibly between speaker specific and general pragmatic expectations.

Notably, Experiment 3 also included situations in which the two expectations were in conflict. For example, in some

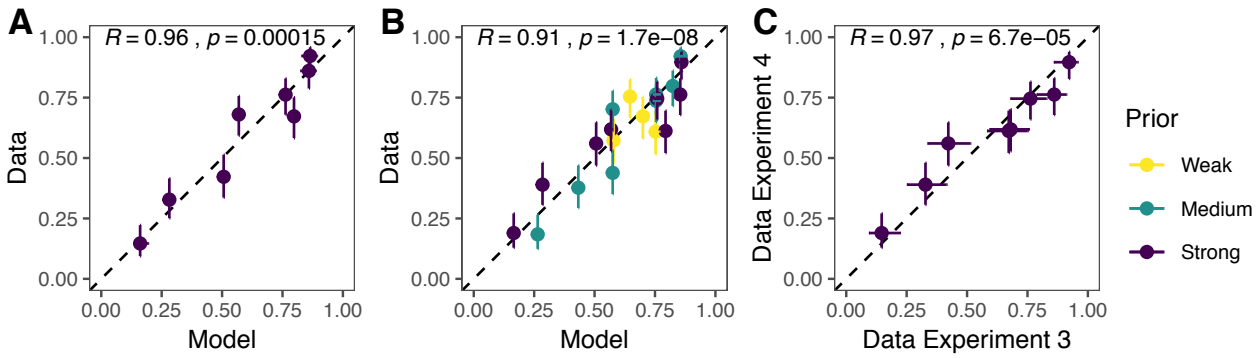


Figure 4: (A) Model predictions compared to data for Experiment 3 and (B) Experiment 4. (C) Data for strong prior manipulation in Experiment 3 and 4, providing a noise ceiling for the reliability of measurements. Error bars = 95% HDIs.

trials the speaker expressed preference for object A, which was also the more frequent (less informative) object. In these situations, a majority of participants chose the preferred object as the referent (see also Fig. 3 preference–same speaker–incongruent). A simple explanation for this pattern might be that common ground manipulations were simply "stronger", corroborated by the fact they produced higher rates of expected choice than the informativeness expectation when the two were presented in isolation (see Fig. 2). In Experiment 4, however, the medium manipulation for novelty yielded numerically weaker results compared to the informativeness expectation in Experiment 1, and yet participants still selected the novel object above chance when the expectations were in conflict. Why is this? Because common ground is represented in our model as the listener's prior distribution, speakers can reason about it in choosing their utterance. That is, in the mind of the listener, the speaker computes the effect of each utterance on a naive listener with shared common ground. Therefore, when prior interactions implicate one object as the more likely referent, the speaker reasons that this object will be the inferred referent of any semantically plausible utterance, even when the same utterance would point to a different object in the absence of prior information.

A range of probabilistic models have been used to model word learning (e.g. Fazly, Alishahi, & Stevenson, 2010; Frank, Goodman, & Tenenbaum, 2009; Xu & Tenenbaum, 2007). RSA models differ from these approaches in that they treat word learning as the outcome of a social reasoning process. In contrast to models for cross-situational word learning (Fazly et al., 2010; Frank et al., 2009), RSA models show how learning might occur in a one shot scenario based on pragmatic reasoning alone. While the ad-hoc informativeness inference characteristic for RSA would be predicted by the model of Xu and Tenenbaum (2007), in their work it follows from the "size principle" of generalization (Tenenbaum & Griffiths, 2001) and not from social reasoning. In contrast to RSA, this approach does not offer a straightforward way to incorporate other types of social information such as expectations following from common ground.

We treated common ground expectations as equivalent to more basic manipulations of contextual salience (e.g. in Frank & Goodman, 2012) and did not explicitly model the social-cognitive processes that give rise to these expectations. The interaction around the object prior to the test event simply increased the probability that this particular speaker will refer to the object subsequently. The same change could be brought about if one of the objects would be made perceptually more salient, for example by making it flash. In future work, it would be interesting to explore ways to model common ground expectations explicitly as well as to contrast perceptual and interactional salience.

Our work integrates different perspectives on the study of pragmatic inference. Previous work focused either on general or speaker specific expectations. The methodological approach taken here illustrates how computational and experimental approaches can be used in conjunction to explicate theories of language use and learning.

Corresponding data and code are available at
`https://github.com/manuelbohn/mcc`

Acknowledgements

MB received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 749229. MCF was supported by a Jacobs Foundation Advanced Research Fellowship and NSF #1456077.

References

Akhtar, N., Carpenter, M., & Tomasello, M. (1996). The role of discourse novelty in early word learning. *Child Development*, *67*(2), 635–645.

Bohn, M., & Koymen, B. (2018). Common ground and development. *Child Development Perspectives*, *12*(2), 104–108.

Clark, E. V. (2009). *First language acquisition*. Cambridge: Cambridge University Press.

Clark, H. H. (1996). *Using language*. Cambridge: Cam-

bridge University Press.
Fazly, A., Alishahi, A., & Stevenson, S. (2010). A probabilistic computational model of cross-situational word learning. *Cognitive Science*, *34*(6), 1017–1063.
Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998–998.
Frank, M. C., & Goodman, N. D. (2014). Inferring word meanings by assuming that speakers are informative. *Cognitive Psychology*, *75*, 80–96.
Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, *20*(5), 578–585.
Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, *20*(11), 818–829.
Grice, H. P. (1991). *Studies in the way of words*. Cambridge, MA: Harvard University Press.
Levinson, S. C. (2000). *Presumptive meanings: The theory of generalized conversational implicature*. Cambridge, MA: MIT press.
Saylor, M. M., Sabbagh, M. A., Fortuna, A., & Troseth, G. (2009). Preschoolers use speakers' preferences to learn words. *Cognitive Development*, *24*(2), 125–132.
Sperber, D., & Wilson, D. (2001). *Relevance: Communication and cognition* (2nd ed.). Cambridge, MA: Blackwell Publishers.
Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and bayesian inference. *Behavioral and Brain Sciences*, *24*(4), 629–640.
Tomasello, M. (2009). *Constructing a language*. Cambridge, MA: Harvard University Press.
Xu, F., & Tenenbaum, J. B. (2007). Word learning as bayesian inference. *Psychological Review*, *114*(2), 245.

Conversation Transition Times: Working Memory & Conversational Alignment

Julie E Boland
University of Michigan

Abstract

Fluent conversation is a marvel of multi-tasking within the language domain: listeners must simultaneously comprehend the speaker, predict a turn transition point, and plan a response. Experiment 1 used spontaneous conversation to investigate the apparent demands of conversation on working memory by manipulating the difficulty of a secondary task. The experiment found support for Load Theory's (e.g., Lavie et al. 2004) prediction that both conversational fluency and performance on a secondary task would decrease as working memory load increased. However, there was also some support for Pickering and Garrod's (2004, 2013) proposal that dialogue is facilitated by a collection of automatic cognitive operations when interlocutors are well-aligned (i.e., using the same words, phrases, and structures to discuss the same topics). Experiment 2 tested two claims motivated by this account: alignment is necessary for fluent turn transitions, and lexical repetition between speakers is an essential component of the alignment advantage. We found support for the former claim, but not the latter.

Keywords: Conversation, Dialogue, Working Memory

Introduction

When a conversation is fluent, the shift from one speaker to the next proceeds rapidly, usually with little or no overlap. In fact, the silent pauses between speakers (i.e., turn transitions) average less than 500ms across cultures, and around 200ms for English speakers in two-party conversations (e.g. Sacks, Schegloff, & Jefferson, 1974, Wilson & Wilson, 2005; Stivers et al, 2009). Levinson and Torreira note that it takes about 600ms to name a picture using a single word, and 425ms of that time is estimated to be necessary for the lexical retrieval and phonological encoding processes for a single word (Indefrey & Levelt, 2004). Thus, it is clear that the signal to begin preparing one's response cannot be the end of the current speaker's utterance, because turn transition times would be on the order of seconds, not milliseconds.

Current theories of conversation explain short transition times by positing multiple processing streams that allow the listener to prepare her response while simultaneously comprehending the current speaker and anticipating a turn transition point (Garrod & Pickering, 2015; Levinson & Torreira, 2015). This multi-tasking burden would seem to induce a heavy working memory load.

Many studies have reported a relationship between working memory and language processing (e.g., Danenman & Carpenter, 1980; DeDe, Caplan, Kemtes, & Waters, 2004; Lewis, Vasishth, & Van Dyke, 2007; Martin & Slevc, 2014). For example, Fedorenko, Gibson, and Rohde (2006) found that participants had difficulty comprehending complex sentences when they had to simultaneously remember three words that were semantically related to the words in the sentence, presumably because comprehending the sentence and remembering the words competed for the same working memory resources. In contrast, conversational participants manage to simultaneously comprehend the current speaker, predict when he will end his turn, and plan a response. All three tasks presumably use the same language system with no apparent interference and surprising efficiency. Thus, conversational fluency presents an interesting puzzle in light of established theories for how working memory supports language comprehension and language production.

If conversation places high demands on working memory, it should be difficult to converse while simultaneously doing another task. For example, Load Theory (Lavie, Hirst, de Fockert & Viding, 2004) predicts a decrease in processing fluency as working memory load is increased. The predictions of Load Theory, as applied to conversation, were supported by Boiteau, Malone, Peters, and Almor (2014), who found that conversation interfered with a simultaneous mouse-tracking task. In turn, the mouse-tracking task modulated speaking rate slightly, but did not increase the rate of disfluencies. Boiteau et al. did not examine fluency variables, such as turn transition time and turn length, nor did they manipulate the difficulty of the secondary task. Several other papers examined the relationship between language production and a secondary task, and also found trade-offs between the language and non-language tasks (Becic et al., 2010, Kemper, Herman, & Nartowicz, 2005; Sjerps & Meyer, 2015).

The literature supports the view that conversation carries a substantial working memory load, but in practice, people often converse while doing something else. In fact, Pickering and Garrod's (2004, 2013) Alignment account suggested that conversational fluency is attained via many automatic mechanisms, at least when interlocutors are well-aligned (i.e., using the same words, phrases, and structures to discuss the same topics). We test three claims from this account:

(i) Well-aligned conversation makes minimal demands on central resources

(ii) Topic alignment enhances conversational fluency
(iii) Lexical repetition enhances conversational fluency

Experiment 1 investigated (i) by manipulating the difficulty of a secondary task. Experiment 2 investigated claims (ii) and (iii) using a picture-description paradigm. The primary focus is on transition time, but speech rate, utterance length, turn type, and the occurrence of disfluencies were measured and analyzed as well, because there may be tradeoffs among these fluency measures.

Experiment 1

If the multi-tasking required for fluent conversation strains the working memory system, adding a secondary task should decrease conversational fluency. To test this, we had participants perform a letter version of the n-back task (Smith & Jonides, 1997) while carrying on a casual conversation with an experimenter. For the n-back task, participants saw a sequence of letters on a computer screen. Both lower-case and upper-case forms of a letter counted as the same letter, to encourage verbal encoding of the stimuli. In the 1-back condition, participants pressed a key if the current letter matched the previous letter. In the 2-back condition, participants pressed a key if the current letter matched the one two letters back. The Load Theory predicts a greater impact on conversational fluency in the 2-back condition, compared with the 1-back condition, and compared with conversation alone.

Method

Participants Forty undergraduates (9 male) received course credit for participation. All were native English speakers.

Procedure The experiment consisted of five experimental blocks: Conversation-Only, 1-back alone, 2-back alone, Conversation with 1-back, and Conversation with 2-back. Stimuli in the n-back consisted of upper and lower case tokens of 8 letters: A, F, J, K, L, O, S, U. The order of the blocks was rotated across five groups, so that each block occurred equally often in each serial position.

Participants were greeted by one of four native English speakers (two male, two female), who conducted the experiment and served as the other interlocutor in conversation blocks. Each experimenter ran two participants on each of the five block orders. The experimenter and the participant were separated by a cubicle barrier for all blocks. Before beginning the experiment, participants were first trained on the 1-back and 2-back tasks. After training, participants completed the experimental blocks. The conversation topics were always in the same order, regardless of block order: 1. life in a college town, 2. pop culture, 3. personal background. Each of the three conversation blocks was 8 minutes long.

The conversation blocks were audio-recorded. The middle 5 minutes were transcribed, with the onset and offset of each turn marked. These transcription records were used to code turn type and disfluencies, and to compute turn transition time, turn length, and speech rate. Alignment was (very roughly) estimated using Latent Semantic Similarity (LSS, Landauer and Dumais, 1997, online pairwise comparisons tool http://lsa.colorado.edu, settings: document to document, general reading up to 1st year college, maximum dimensions).

Results

Small differences in the participants' transition times were found as a result of the secondary tasks (see Figure 1). Interestingly, the longest transition times were observed for the experimenter, who had no secondary task other than to keep the conversation going.

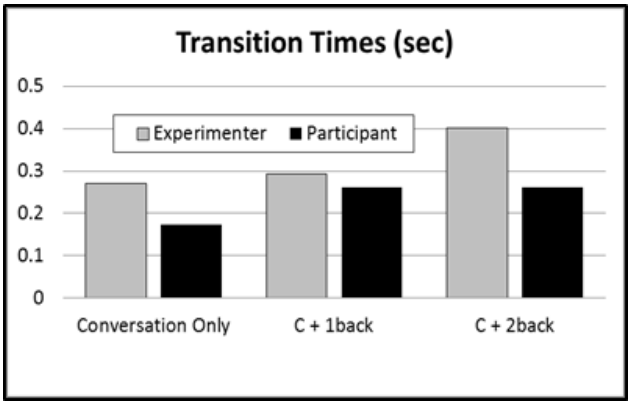


Figure 1: Transition time (in seconds) by task condition, for both participant and experimenter.

Table 1 summarizes the results of a linear mixed effects model on the participants' turn transition times: lmer(trans_time ~ Experimenter + Order + Turntype + Block + LSS + (1 + Block |subj)). The four experimenters, five block orders, and four most common turn types were used as control variables. To save space, only significant effects for control variables are included in Table 1.

Table 1. Analysis of Participant's Transition Time

	Estimate	t	p
Block C vs C1	.06	1.61	.11
Block C vs C2	.08	2.14	.04
Exp1 vs Exp4	-.23	-4.83	.00
Ord1 vs Ord3	.19	3.49	.00
Agree vs Answer	.23	6.92	.00
Agree vs Quest	.23	4.13	.00

There was no effect of the secondary task on participants' speech rate, but the task manipulation did impact both

utterance length and the probability of a turn-initial filled pause. As predicted by Load Theory, participants took longer turns and made fewer turn-initial filled pauses in the Conversation-Only block compared with blocks that combined conversation with the n-back task.[1]

Across all four dependent measures, the strongest predictor of the participant's conversational fluency was turn type, overshadowing the secondary task manipulation. The four most common turn types (agreement, answering a question, asking a question, or making a comment) made up 98% of the participant turns. As shown in Figure 2, type of utterance was a strong predictor of both transition time and utterance length.

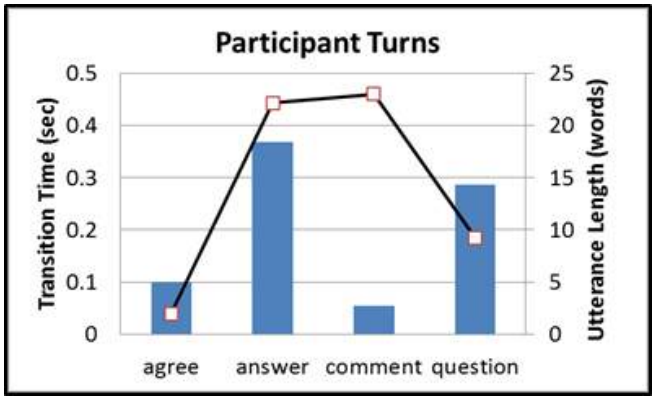


Figure 2: Participant's transition time (bars) and utterance length (line) by turn type.

As predicted by Load Theory, performance on the n-back task was worse in the blocks that required simultaneous conversation, especially in the 2-back condition (see Table 2). Participant means were submitted to a 2 (n-back level) by 2 (alone or w/conversation) by 5 (order) repeated measures ANOVA, with the third factor as a between-participants variable. Robust effects of n-back level [$F(1,35) = 139.23$, $p < .01$], conversation [$F(1,35) = 280.02$, $p < .01$], and their interaction [$F(1,35) = 98.54$, $p < .01$] were observed. No effect of order or interactions with order approached significance [all F's less than 1.9].

Table 2: Percent Correct (w/standard error) on n-back.

	alone	w/conversation
1-back	99 (.2)	91 (.8)
2-back	97 (1.5)	79 (1.1)

One-minute clips from the conversations were presented to 108 naive listeners, who judged whether the participant (always the first speaker in the clip) had been under no load, low load, or high load from a secondary task. As shown in Table 3 with correct responses highlighted, listeners were highly inaccurate. Their bias was to guess "none" or "low." This suggests that participants in the primary experiment were largely successful at maintaining fluency, despite the extra load from the secondary task.

Table 3. Percent load judgments by audio clip condition.

	C only	C+1-back	C+2-back
None	50	47	41
Low	40	37	42
High	10	16	17

Consistent with the Alignment hypothesis, there was some evidence that participants were more fluent when the alignment between speakers was highest, as illustrated in Figure 3: The higher the Latent Semantic Similarity (LSS) between speakers, the faster the participant's speech rate. The LSS was also a marginal predictor of transition time. To be sure, all of the conversations were at the high end of the LSS scale, which ranges from -1 to 1. Thus, there may have been insufficient variance to find a stronger correlation.

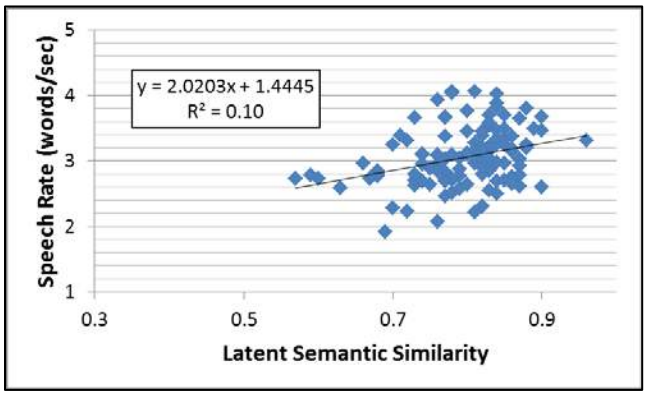


Figure 3: The relationship between participant speech rate and Latent Semantic Similarity.

Discussion

Load Theory was supported by the decrease in accuracy on the n-back task when it was paired with conversation, and by the increase in transition times with n-back load, While transition times remained within the normal range of around 200ms cited by Stivers et al. (2009) and others, participants seemed to use compensatory measures, such as shortening their turns and beginning their turns with a filled pause, in order to maintain short transition times when faced with a challenging secondary task.

While the experiment was not a clear test of Pickering and Garrod's (2004, 2013) Alignment account, the apparent robustness of conversational fluency to a secondary task (albeit with some modulations) is consistent with their account. Furthermore, the modest correlation between LSS and speech rate is suggestive. Unfortunately, there are no established, objective measures of conversational alignment,

[1] See Appendix for statistical support.

making it difficult to rigorously test Pickering and Garrod's predictions. Nonetheless, the strong effect of turn type (see Figure 2) seems problematic for their account. Turn type determines how quickly participants can begin planning their responses and how much planning is required. The robust turn-type effects suggest that the cognitive operations supporting utterance planning are less automatic than maintained by strong versions of the Alignment account (e.g., Pickering & Garrod, 2013).

Experiment 2

Because alignment can't be manipulated in spontaneous conversation, Experiment 2 used a picture description paradigm to test the effects of two aspects of alignment (shared topic and shared vocabulary). Topic was manipulated within-participants, such that each of the participant's picture descriptions was preceded by a pre-recorded sentence that was either a description of the same picture or a description of a different picture. Shared vocabulary was measured by counting the number of content words from the pre-recorded sentence that were repeated in the participant's picture description.

While it may seem odd to treat pre-recorded stimuli as a speaker in a conversation, this approach was successful in a recent experiment. Corps, Crossley, Gambi, and Pickering (2018) found that participants answered pre-recorded yes/no questions faster when the final word was predictable, with transition times averaging around 400ms. Participants were encouraged to respond quickly, answering "as soon as you expect the speaker to finish the question" (p. 83). While these transition times are slower than typical transition times in dyadic English conversations and the responses were very simple, the finding demonstrates that participants were actively predicting the content of the pre-recorded stimuli and using those predictions to prepare their own response during the other speaker's turn, analogous to conversation.

We encouraged participants to time their utterances to coincide with the offset of the pre-recorded stimuli through a scaffolded training procedure. However, this study did not use question/answer pairs, making the link to conversation somewhat more tenuous.

Method

Participants Twenty-nine undergraduates participated for course credit and were randomly assigned to one of two lists. All were native English speakers

Procedure On each of 36 trials during the experiment, participants looked at a line drawing of a complex scene while listening to an auditory sentence. Participants were instructed to describe the scene as soon as the auditory sentence ended. The participant was instructed to refer to the entity indicated by the arrow in their description of the picture (see Figure 4). In the Match condition, the auditory sentence was about the current image; in the Mismatch condition, an auditory sentence for a different image was substituted. Across the two lists, every picture occurred in both the Match and Mismatch conditions, and each participant received half of each type. After the participant finished their utterance, the next screen presented a printed word and participants judged whether it had been in the auditory sentence of the current trial (50% had been). This recognition probe encouraged attention to the auditory sentence. For the image in Figure 4, the matching sentence was “William was very pleased with himself for surprising his wife with an anniversary gift”, and the probe word was "pleased".

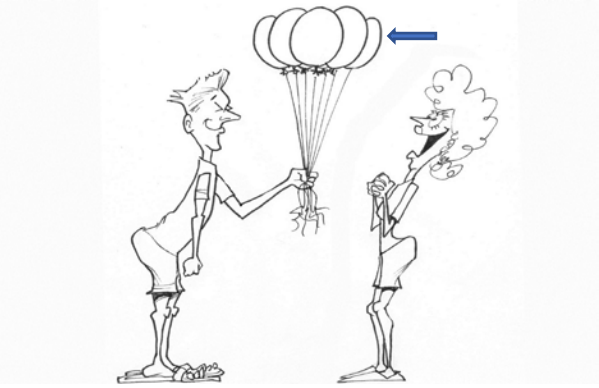


Figure 4: Example line drawing.

Prior to beginning the experiment, participants practiced each component of the task: describing the pictures while referring to the entity indicated by the arrow, timing their response to commence as closely as possible to the end of the auditory sentence, and answering yes/no to the recognition probe word.

Both the pre-recorded auditory stimuli and the participant’s picture descriptions were audio-recorded and transcribed as in Experiment 1, with the participant and the pre-recorded stimuli treated as different speakers.

Results

The participants' utterances were coded for transition time, speech rate, utterance length, the presence of disfluencies, and the number of content words repeated from the pre-recorded stimulus (see Table 4).

As predicted by the Alignment account, fluency was higher in the Match condition. There were shorter transition times and more succinct descriptions when the auditory stimulus sentence described the same image as the participant's sentence. The effect on transition time was confirmed in linear mixed effect model, summarized in Table 5: lmer(trans_time ~ Condition +RepeatWords.C +Accuracy + Utt_words.C + (1 + Condition|subj) + (1 + Condition|trial)). Accuracy on the recognition probe and the number of words in the participant's utterance were included

as control variables[2].

Table 4. Means (standard error) for Experiment 2

	Matched	Mismatched
Transition Time	515 ms (22)	575 ms (25)
No. of Words	10.01 (.17)	9.56 (.16)
Speech Rate	3.23 w/s (.04)	3.15 w/s (.04)
Disfluent %	33 (2)	33 (2)
No. Repeated Words	1.42 (.05)	0.12 (.02)
Probe Accuracy %	87 (1)	77 (2)

Table 5. Transition Time Analysis for Experiment 2

	Estimate	t	p
Intercept	.51	7.83	.00
Condition	**.08**	**2.55**	**.01***
RepeatWords.C	.03	1.25	.21
Accuracy	-.02	-.87	.38
Utt_words.C	**.14**	**4.51**	**.00***

Participants were more accurate overall on the recognition probe in the Match condition than in the Mismatch condition [2-tailed paired t-test: $t(29) = 4.54$, $p < .001$]. This could be because greater alignment eased overall processing load. Alternatively, higher accuracy on Match trials could reflect participants having produced the probe word themselves when describing the picture. This occurred on 9% of Match trials with a "yes" probe word and less than 1% of the time on Mismatch trials with a "yes" probe word. Not surprisingly, participants never used the probe word themselves on "no" trials, in which the probe word was not in the pre-recorded sentence. When analyzing only the "no" probe trials, the effect of Match remained robust [95% Match, 84% Mismatch condition, $t(28) = 4.15$, $p < .001$], consistent with the hypothesis that greater alignment eased overall processing load.

Contrary to the Alignment prediction, repetition of words did not increase fluency. Instead, the numerical trends went in the opposite direction (see Figure 5): the more content words the participant repeated from the auditory stimulus sentence, the longer the transition time and the more wordy the image description. This surprising pattern might arise if participants used a lot of pronouns in the Match condition, rather than repeating referring expressions from the auditory stimulus. This pattern was not found. Although, there was a slight numerical difference in pronoun usage (.66 pronouns per utterance in the Match condition, .60 in the Mismatch condition), it was not significant in a 2-tailed, paired t-test ($t(29) = 1.41$, $p > .10$), nor was there an effect of pronouns or an interaction between pronouns and Match/Mismatch, when pronoun usage was added to the statistical model used for Table 5.

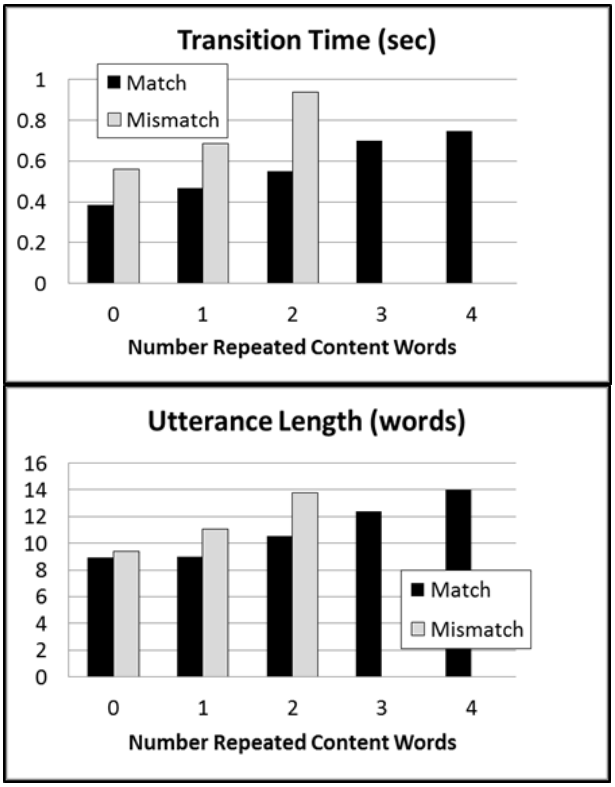


Figure 5: Transition times (upper bar graph) and utterance length (lower) in the Matched and Mismatched condition, when controlling for number of repeated words.

Discussion

Despite the artificiality of the paradigm, a shared topic decreased transition time, as predicted. This effect suggests that participants planned their utterance during the auditory stimulus, analogous to spontaneous conversation.[3] Furthermore, it provides some direct support for the Alignment account. However, the topic alignment advantage did not come from the most obvious source--lexical repetition of words in the auditory stimulus. Rather, the Match advantage in this paradigm must be due to other phenomena, such as more accurate prediction of the end of the auditory stimulus and semantic priming. It remains possible that lexical repetition plays a more important role in natural conversation than it did in this paradigm, but the

[2] To verify that the results remained the same when including only trials on which participants attended to the recorded sentence, an additional statistical model was run, including only trials on which participants responded accurately to the probe word. It was identical to the original model, except that accuracy was excluded. As expected, the same pattern of effects reported in Table 5 was obtained.

[3] In an earlier version of the experiment that did not include the pre-experiment training, mean transition time was well over a second, with no effect of match. Thus, participants may not multi-task in this paradigm (i.e., plan their utterance during the auditory stimulus while predicting its endpoint) unless explicitly encouraged to do so.

current experiment found no support for the lexical repetition prediction, motivated by the Alignment account.

General Discussion

Prior research indicated that conversation competes for central resources when paired with a secondary task from another domain, such mouse-tracking, walking with groceries, or driving (Becic et al., 2010; Boiteau et al., 2014, Kemper et al., 2005; Sjerps & Meyer, 2015). However, except for Boiteau et al., this research did not use natural, spontaneous conversation and examined relatively few measures of conversational fluency. Experiment 1 used spontaneous conversation to extend this finding to a secondary task (mixed-case letter n-back) that uses resources within the language domain. Consistent with prior research and with Load Theory, we found interference effects for both conversation fluency and the secondary task. We also found that turn type was a strong predictor of multiple conversational fluency variables, reflecting the differential processing demands of agreeing, questioning, answering, and commenting.

In addition, we explored some predictions of the Alignment theory using a picture description paradigm with pre-recorded stimuli instead of a live interlocutor. To increase the similarity with natural conversation, we trained participants to time their utterances to coincide with the offset of the auditory stimulus, while obeying other task-specific constraints. In this paradigm, participants initiated their own utterance closer to the offset of the auditory stimulus when both utterances shared the same topic (a co-present image). This finding is consistent with the Alignment account. However, the number of content words shared between the auditory stimulus and the participant's picture description was not related to conversational fluency in the direction predicted by the Alignment account.

In sum, we found considerable support for theories in which conversation consumes processing resources, such as working memory and attention. At the same time, we were surprised that participants were able to maintain typical transition times as working memory load increased. Our results from the two experiments suggest that this feat was possible because of shared topics across adjacent turns and due to compensatory mechanisms, such as making turns shorter or beginning turns with a filled pause.

Acknowledgments

This research was made possible by a team of undergraduates who helped create the experiments, collect the data, and participate in the conversations for Experiment 1. They and additional students transcribed coded the data: Kelly Kendro, Jocelyn Brickman, Meher Sabri, Madylin Eberstein, Yaozong Huang, Olivia Shulman, Maxwell Recknagel, Thinh Nguyen, Lindsey Harris, Katherine Hall, Hussam Hashem, Luciana Rosania, and Natalie Jackson.

References

Becic, E., Dell, G. S., Bock, K., & Garnsey, S. M. (2010). Driving impairs talking. *Psychonomic Bulletin & Review*, *17*, 15-21.

Boiteau T. W., Malone, P. S., Peters, S. A., & Almor, A. (2014). Interference between conversation and a concurrent visuomotor task. *Journal of Experimental Psychology: General. 143, 295-311*.

Corps, R. E., Crossley, A., Gambi, C., & Pickering, M. J. (2018). Early preparation during turn-taking: Listeners use content predictions to determine what to say but not when to say it. *Cognition*, *175*, 77-95.

Daneman M and Carpenter PA (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, *19*, 450-466.

DeDe, G., Caplan, D., Kemtes, K., & Waters, G. (2004). The relationship between age, verbal working memory, and language comprehension. *Psychology and Aging*, *19*, 601-616.

Fedorenko, E., Gibson, E., & Rohde, D. (2006). The nature of working memory capacity in sentence comprehension: Evidence against domain-specific working memory resources. *Journal of Memory and Language*, *54*, 541-553.

Indefrey, P. & Levelt, W.J.M. (2004). The spatial and temporal signatures of word production components. *Cognition*, 92, 101–144.

Kemper, S., R. Herman, R., & Nartowicz, J. (2005). Different Effects of Dual Task Demands on the Speech of Young and Older Adults. *Aging, Neuropsychology, and Cognition, 12, 340-258.*

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory. *Psychological Review*, *104*, 211-240.

Lavie, N., Hirst, A., de Fockert, J. W., & Viding, E. (2004). Load Theory of Selective Attention and Cognitive Control. *Journal of Experimental Psychology: General*, *133*, 339-354.

Levinson, S. & Torreira, F. (2015). Time in turn-taking and its implications for processing models of language. *Frontiers in Psychology*, *6*, 371.

Lewis, R. L., Vasishth, S., & Van Dyke, J. A. (2007). Computational principles of working memory in sentence comprehension. *Trends in Cognitive Sciences*, *10*, 447-454.

Martin, R. C. & Slevc, L. R. (2014). Language production and Working Memory. In *The Oxford Handbook of Language Production* (M. Goldrick, V. Ferreira, & M. Miozzo, Eds). Oxford University Press, pp 437-447.

Pickering, M. J. & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral & Brain Sciences*, *27*, 169-190.

Pickering, M. J. & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral & Brain Sciences*, *36*, 329-347.

Sacks. H., Schegloff, E. A., Jefferson, G. (1974). A simplest systematics for the organization of turn taking for conversation. *Language*, *50*, 696–735.

Sjerps, M. J. & Meyer, A. S. (2015). Variation in dual-task performance reveals late initiation of speech planning in turn-taking. *Cognition*, *136*, 304-324.

Smith, E. E. & Jonides, J. (1997). Working memory: A view from neuroimaging. *Cognitive Psychology*, *33*, 5-42.

Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., de Ruiter, J. P., Yoon, K-E, Levinson, S. C., & Kay, P. (2009). Universals and Cultural Variation in Turn-Taking in Conversation. *Proceedings of the National Academy of Sciences of the United States of America, 106* (26), 10587-10592.

Wilson & Wilson (2005). An oscillator model of turn-taking. *Psychonomic Bulletin & Review, 12,* 957-968.

Appendix

Linear mixed effect analyses for additional dependent variables from Experiment 1. The same statistical model was used for each: lmer(DV ~ Experimenter + Order + Turntype + Block + LSS + (1 + Block |subj)).

Table A1. Utterance length (words per turn)

	Estimate	t	p
Block C vs C1	-6.01	-3.25	.00
Block C vs C2	-7.70	-4.18	.00
Ord1 vs Ord3	-8.49	-2.36	.02
Agree vs Answer	21.37	19.35	.00
Agree vs Quest	6.80	3.72	.00
Agree vs Comm	19.09	16.97	.00

Table A2. Speech rate (words per second)

	Estimate	t	p
Block C vs C1	.00	0.02	.98
Block C vs C2	.09	1.54	.13
Exp1 vs Exp2	-.28	-2.11	.04
Exp1 vs Exp 3	-.31	-2.40	.02
Ord1 vs Ord3	-.47	-3.39	.00
Ord1 vs Ord5	-.33	-2.41	.02
Agree vs Answer	.36	6.76	.00
Agree vs Quest	.99	11.28	.00
Agree vs Comm	.80	14.87	.00
LSS	.18	2.85	.01

The probability of a sentence-initial filled pause not analyzed using the above statistical model due to the large number of 0's (no filled pause) across trials. Instead, the probability of a sentence initial filled pause for each participant, in each condition, was analyzed using repeated measures ANOVA. There was a main effect of block [$F(2, 76) = 6.79$, $p < .01$], with a probability of .09 in the Conversation-Only condition, .14 in the Conversation with 1-back condition, and .15 in the Conversation with 2-back condition.

An Insight into Language: Investigating Lexical and Morphological Effects in Compound Remote Associate Problem Solving

Alexander H. Bower (ahbower@uci.edu)
2275 Social & Behavioral Sciences Gateway Building
University of California
Irvine, CA 92697

Andrew Burton (ajburton@uci.edu)
402 Social Science Lab
University of California
Irvine, CA 92617

Mark Steyvers (mark.steyvers@uci.edu)
2201 Social & Behavioral Sciences Gateway Building
University of California
Irvine, CA 92697

William Batchelder
2129 Social Science Plaza A
University of California
Irvine, CA 92617

Abstract

Understanding the processes leading to insight has remained one of psychology's greatest challenges. In this study, we examined how different lexical properties affect cognitive processes involved in a popular class of insight problems: Compound Remote Associates (CRAs). These properties were familiarity, lexeme meaning dominance, and semantic transparency. We found that a higher proportion of problems were solved when they were presented beginning with the most familiar cues, but not when they began with right-headed dominant or the most semantically transparent cues. Further, we found that participants focused their efforts disproportionately on first and last cues, that subjective ratings of insight decreased as trial times elapsed, and that the magnitude of reported insight increased with the number of cues successfully solved. This suggests that participants can monitor their progress in such problems. These results contest longstanding assumptions of requisite periods of impasse and the absence of incremental progress in insightful problem solving.

Keywords: compound remote associates; insight; language and thought; problem solving

Introduction

Insight has sparked some of history's greatest accomplishments – from Einstein's special theory of relativity to Newton's universal law of gravitation. These sudden "aha!" moments also permeate our everyday lives – from practical household problems to puzzles in video games. However, our understanding of the processes underlying insight have remained subject to empirical gaps and theoretical debate (Batchelder & Alexander, 2012). Indeed, a prevailing assumption of the literature has been that insight occurs by merit of one solving an "insight problem" (Topolinski & Reber, 2010). To make meaningful progress toward understanding insight, we must first explore the cognitive mechanisms involved in problems in which it is reported.

One such class of problems are Compound Remote Associates (CRAs) (Bowden & Jung-Beeman, 2003). The CRA task was developed as a modified version of the Remote Associates Test (RAT) (Mednick, 1962), which has been correlated with performance in insight problems. The difference between the original RAT and CRAs is that the latter only uses structural associates based on syntax (Worthen & Clark, 1971). In CRAs, people are presented three cue words and must produce a solution word that is common to all three, forming compound words and phrases. For example, the solution to the triad "COTTAGE, SWISS, CAKE" is CHEESE (forming "COTTAGE CHEESE," "SWISS CHEESE," and "CHEESECAKE," respectively). The task is designed such that a solver must break free of high-frequency associations to access globally satisfactory solutions.

CRAs have many advantages over classic insight problems: 1) they have large, normed databases, 2) many can be completed in single, short experimental sessions, 3) they can be solved with and without insight, 4) people have reliably demonstrated that they can make subjective judgments of insight regarding them, 5) they can be used in neuroimaging studies to identify the neural correlates of insight, and 6) they can be supplemented with time-based measures of solution latencies. As a result, they have been widely used to explore various cognitive domains, such as intuition (Topolinski & Strack, 2008), sleep (Cai et al., 2009),

and computational/deep learning (Olteteanu, Gautam, & Famomir, 2015).

Much of the past research on the RAT and CRAs has defaulted to a correlational account that simply assumes insight and ignores the underlying processes that may drive it. This problem was highlighted by Topolinski and Reber (2010), who pointed out that many researchers neglect to explain the *phenomenology* of insight yet rely on it as a sufficient condition.

Recent studies have attempted to mend this by modelling CRA performance. Gupta, Jang, Mednick, and Huber (2013) were among the first to provide a formalized account of individual differences in CRA search behavior. They employed a norm-based model that defined the best guess at solutions based on the average of cues in the Word Association Space (WAS) (Steyvers, Shiffrin, & Nelson, 2005). This was contrasted with a frequency-biased model that assumes people's search is biased by word fluency, based on Griffiths, Steyvers, and Firl's (2007) work with PageRank and associative frequency. As predicted, they found that the probability of a given response is biased toward high-frequency words. Thus, people perform poorly if they're biased in favor of high-frequency incorrect words, precluding access to low-frequency correct responses.

This work was extended by Olteteanu and Falomir (2015), who developed the comRAT-C; a computational model that solves compound RAT queries, based on a cognitive theoretical framework for creative problem solving (CreaCogs) (Olteteanu, 2016). The knowledge base (KB) comprising the CRAs themselves used language data (2-grams pruned for relevance) from the Corpus of Contemporary English (COCA). They found that the comRAT-C used a convergence process similar to that of human solvers, and that the frequency of cues in the KB influences responses. The comRAT-C was able to correctly solve 64 of the 144 items in Bowden and Jung-Beeman's (2003) list of normed CRAs, in addition to suggesting unlisted, yet plausible solutions in more than 20 cases – suggesting its own form of creativity. Overall, their study laid a solid computational framework for formalizing the processes in CRA problem solving.

A promising experimental approach was taken by Smith, Huber, and Vul (2013), who used Latent Semantic Analysis (LSA) to evaluate the similarity between people's guesses, word cues, and answers. They accomplished this by having participants enter every word considered while searching for the answer, regardless of their correctness. By doing this, they focused on the search processes used when generating candidate answers through a probabilistic sampling framework. They found sequential dependencies between responses in a problem, with subjects generating semantically similar chains of responses. Additionally, people seemed to focus primarily on one cue at a time. However, their procedure assumes that guesses accurately reflect the implicit nature of the search, even though the very act of conscious report may alter the search process.

The main body of work on CRAs has focused on associative aspects - not the requisite that responses be syntactic *compounds* (with the notable exception of Olteteanu and Falomir (2015)). Indeed, research has largely ignored the morphological properties of the compounds themselves and how they affect performance and the likelihood of reported insight. We have thus failed to adequately address a critical aspect of their character. This approach has potentially restricted us from discovering how people attain insight in these problems. A look at the nature of compounds and their lexical elements is necessary to better understand the underlying cognitive processes involved in these problems.

Compound Word Research

Early work on compound words used a lexical decision paradigm (Taft & Forster, 1975), which measures peoples' response times (RT) in classifying words and nonwords. One such study found that only the lexical status of the first constituent word in a compound affects processing, with longer RT for word-word and word-nonword pairs (e.g., DUSTWORTH, FOOTMILGE) than nonword-word and nonword-nonword pairs (e.g., TROWBREAK, MOWDFLISK) (Taft & Forster, 1975). Thus, it appears that morphological decomposition takes place when processing compound words, instead of the words being stored and retrieved as a whole.

There has been considerable work on visual word recognition in recent years facilitated by databases containing lexical characteristics and behavioral data, such as latencies of word naming and lexical decisions for large sets of words (e.g., Balota et al., 2007) and investigations of word length (New, Ferrand, Pallier, & Brysbaert, 2006). Though initially focused on monosyllabic and monomorphemic words, this work has been extended to address processing in multisyllabic words (Yap & Balota, 2009) and English compound words.

Research suggests that English compounds are processed differently from length and frequency-matched monomorphemic words. For instance, both semantically-transparent compounds (e.g., ROSEBUD) and opaque compounds (e.g., HOGWASH) are processed more quickly than their monomorphemic counterparts (e.g., GIRAFFE) (Ji, Gagné, & Spalding, 2011). This sense of morphological complexity has ignited debate in the psycholinguistic literature, with competing perspectives on compound representation and processing (see Fiorentino & Poeppel, 2007).

The current study investigates the roles of three lexical properties involved in compound processing and, by extension, CRAs: word familiarity, semantic transparency, and lexeme meaning dominance. Thus, we investigated if and how they differentially affected CRA performance and the likelihood of insight. To do this, we used Juhasz, Lai, and Woodcock's (2015) database of 629 compound words to construct 21 novel CRA problems. This database, which adapted items from the English Lexicon Project (ELP: Balota

et al., 2007), compiled subjective ratings for six properties believed to affect morphological processing. The questionnaires used by these authors are available in their Supplementary Materials. We will now briefly explore each of these selected properties and justify their inclusion in the study.

Familiarity Whole word frequencies may be interpreted as analogous to whole word access and have thus been studied in compound word recognition (Juhasz, Lai, & Woodcock, 2015). However, English compound frequencies tend to be low relative to other languages, resulting in experimental challenges and a consequential gap compared to Dutch (Kuperman, Schreuder, Bertram, & Baayen, 2009) and Finnish (Kuperman, Bertram, & Baayen, 2008) counterparts. Rated familiarity can be regarded as a measure of subjective frequency and has been demonstrated to affect word recognition in English monomorphemic words. In particular, familiarity has been shown to influence eye fixation durations, along with word frequency (Juhasz & Rayner, 2003). This was further demonstrated in an experiment by Juhasz, White, Liversedge, and Rayner (2008), which found that familiarity affected gaze duration for both long (ten or more letters) and short (seven or fewer letters) English compound words.

We thus contend that ratings of familiarity can be used as a subjective proxy for word frequency and have a role in affecting morphological processing and CRA performance.

Semantic Transparency Semantic transparency also plays an important role in how compounds are processed and represented (Libben, 1998). A fully transparent compound is one in which both constituents contribute to the meaning of the compound word (e.g., SUNLIGHT), while a fully opaque compound is one in which neither constituent contributes to its meaning (e.g., FLAPJACK). There are also partially-opaque compounds, in which only one constituent contributes to the compound's meaning (e.g., JAYWALK, CHEAPSKATE) (Juhasz et al., 2015).

Libben (1998) proposed a model in which semantic transparency is represented in two distinct ways: the semantic relationship between the meaning of a constituent morpheme within a compound, and the meaning of the morpheme independent of it. For example, the opacity of the compound SHOEHORN results from HORN not being transparently related to the compound as a whole, whereas SHOE is fully transparent. Thus, it is classified as a T-O compound (wherein T = transparent, O = opaque). Compounds require some level of semantic transparency to be tied to semantic representations of their lexemes. Using a lexical decision task, Libben, Gibson, Yoon, and Sandra (2003) found that fully opaque and T-O compounds were responded to more slowly than other compound types, though there was a significant priming effect on all four compound types relative to neutral primes.

Research has demonstrated that semantically transparent compounds are especially susceptible to morphological decomposition, and that semantic priming only seems to occur when there is at least one transparent lexeme. Using Dutch compounds, Sandra (1990) used semantic associates of constituents as primes for transparent (e.g., BIRTHDAY primed by DEATH), opaque (e.g., SUNDAY primed by MOON), and pseudo-compounds (e.g., BOYCOTT primed by GIRL). Facilitatory priming effects were only observed for constituents in transparent compounds.

Lexeme Meaning Dominance Compared to other languages, English compound words tend to be right-headed (i.e., the second constituent word – or lexeme - is the semantic head of the compound). This lexemic dominance primarily defines the meaning of the compound. In a study by Inhoff, Starr, Solomon, and Placke (2008), location and word frequencies of lexemes were manipulated in lexical decision, naming, and sentence reading tasks. They found an effect for larger word frequency for the dominant lexeme in each task. Lexeme dominance also affected first fixations on compound words. These results suggest the headedness of a compound affects how it is recognized and subsequently processed.

Since all the word cues presented in the CRAs in this experiment are the second lexemes, their contribution to the overall meaning of the compound should affect the speed of access when solving each problem.

The Current Study

In accordance with the evidence above, we predicted that CRA problems beginning with word cues that 1) are the most familiar, 2) are the most semantically transparent, and 3) have right-headed lexeme dominance would result in the highest levels of performance and reporting of insight.

To test this, we staggered the presentation of word cues on-screen, with cues either increasing or decreasing in ratings for the relevant lexical domain. Thus, we actively constrained and manipulated the search processes used by solvers. As CRA triads are commonly presented at once, this presents an experimental departure that, we hypothesize, differentially affects performance and captures some of the latent features of this process. To our knowledge, this is one of the first studies to actively manipulate cue presentation in CRAs in such a way with precise behavioral predictions.

Another departure is in how problems are scored. CRAs are typically scored according to whether a submission 1) conforms to all three cue words and 2) conforms to the suggested response of the researchers, precluding other "incorrect", yet plausible responses. This does not allow for investigation of partially correct problems, in which fewer than three cues are satisfied by the solution candidate. We address this issue using a lexicon to test whether submitted responses form valid compounds against each individual cue presented, either as a prefix or suffix. This allows for a more comprehensive picture of the processes and strategies employed in such problems.

Together, this study contributes to the CRA literature in three major ways: 1) it uses a staggered presentation of word cues, facilitating semantic activation and lexical search

behavior, 2) it investigates the morphological properties of the compounds themselves, and 3) it uses partial scoring for each word cue. The goal of the study is to determine how the aforementioned lexical properties affect solution retrieval and if they influence performance and the probability of reported insight.

Methods

Participants

Each experimental condition was composed of two counterbalanced groups, comprising six groups total. All participants (n = 128) were University of California, Irvine undergraduate students, who were awarded course credit through the SONA system for their role in the study. The age distribution was 18-21 (n = 110), 22-25 (n = 11), 26-30 (n = 5), and 31-40 (n = 2). Everyone identified as a native English speaker, with 53 participants identifying as multilingual (though additional languages spoken were not specified).

In the *Familiarity* condition, Group 1 consisted of 23 participants (n = 21 females), and Group 2 consisted of 22 participants (n = 14 females).

In the *Lexeme Meaning Dominance* condition, Group 3 consisted of 21 participants (n = 14 females), and Group 4 consisted of 21 participants (n = 19 females).

Lastly, in the *Semantic Transparency* condition, Group 5 consisted of 21 participants (n = 16 females), and Group 6 consisted of 20 participants (n = 18 females).

Fourteen participants were excluded from the final analysis, as they did not meet the criteria of answering at least two of the three practice problems correctly.

Materials

We constructed 21 novel CRA problems from compounds that had at least three common stems (thus forming three cues with a common solution). For example, there are 10 compound words in Juhasz et al.'s (2015) database with the shared prefixed stem FOOT. The mean ratings for whatever variable was in question (on a 1-to-7 scale for familiarity and transparency and on a 1-to-10 scale for lexeme meaning dominance) were then sorted in descending order and the words with the highest and lowest values were selected. The mean of these two values was then calculated and the compound word with that value or its closest approximate was selected as the middle term. Using the same example of FOOT for the variable of familiarity: FOOTPRINT has the highest value at 7, FOOTPATH has the closest approximate to the mean with a value of 5.85, and FOOTHILL has the lowest value at 4.71. This forms the CRA problem "PRINT, PATH, HILL," with the solution FOOT. All compounds in this database begin with a prefixed solution stem. Thus, unlike other studies, the solution is always the first lexeme in the compound.

In the event of a tie between two compound word values, the compound with the closest letter length to the other two words was selected. If the competing compound had the same length, the tie was broken by identifying which one more closely matched the mean age of acquisition value of the other two compounds.

Due to the limited number of candidate items, some words were repeated in both problem and solution terms. For example, PORT occurs in the problems "PORT, BASE, SICK" and "FOOD, PORT, BOARD." There was also an instance of a having the same phonetic representation (WASTE and WAIST). Participants were told that words may occur more than once both as cues and as solutions.

Each condition was counterbalanced so that problems were presented in both ascending and descending order across two groups. This was done to control for potential order effects.

Procedure

Participants were given instructions and a working definition of "insight" (*Insight occurs when the answer suddenly pops into your head, accompanied by a strong burst of positive emotion ("aha!").*). They were then given an example CRA problem ("CREAM, SKATE, WATER," solution = ICE) and were asked to complete three practice problems with feedback. All four of these problems were pulled from the Bowden and Jung-Beeman's (2003) set of CRA norms and were the four easiest problems with uniformly prefixed solution stems.

The experiment was conducted using a MATLAB interface. Word cues were presented sequentially with 5-second delays between each cue. The first word cue appeared in the left-center of the screen, the second appeared in the center, and the third in the right-center. Cues remained on-screen after their presentation for the remainder of the trial. Figure 1 demonstrates the display of a typical problem trial.

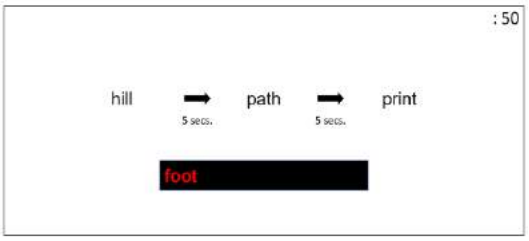


Figure 1: Example of problem trial.

Each trial lasted for one minute. A countdown timer appeared on the top right-hand corner of the screen when 50 seconds remained and turned red when 10 seconds remained. Participants typed their responses in a black box below the cues. They were encouraged to answer as quickly and as accurately as possible. They could submit their response at any time following the presentation of the third cue. Participants were forced to proceed after the minute had expired and whatever was typed into the solution box was accepted as the submitted response.

Following each problem trial, participants were asked to report the level of insight they experienced on a scale of 1 ("no insight") to 7 ("complete insight"). They were also

reminded of its operational definition on the bottom of the screen.

At the end of the experiment, participants were asked to provide a brief (150-word max.) description of what strategies they used to solve these problems. We also asked them to describe the difference they felt between solving problems with and without the feeling of insight. This was done to determine individual differences in reporting criteria and as a check for cross-validity with our definition. This data will also be evaluated to inform future, related experiments.

Participants were scored based on how quickly and accurately they responded to each problem.

Results and Discussion

First, we tested the hypothesis that presentation order of cues according to ratings in each lexical condition would affect performance. These results are shown in Figure 2. Note: "Direction" denotes whether the cue presentation sequentially increased or decreased for the lexical property in question (that is, "Down" indicates that the first cue had the highest rating for the property, while "Up" started with the lowest rating). It appears that the only observed difference was in familiarity, with a higher proportion of problems successfully solved when they began with the most familiar word cue (M = 0.383, SD = 0.126), rather than the least familiar cue (M = 0.301, SD = 0.161, $t(226) = 4.304$, $p < .001$, $d = 0.570$). The estimated Bayes factor suggested that the data were .001:1 in favor of the alternative hypothesis, suggesting decisive evidence for a presentation order effect (Jeffreys, 1961). While this finding was not shared by the other properties (lexeme meaning dominance and semantic transparency), there are several other important findings – some of which challenge widely-accepted assumptions regarding the "special process" view (Bowden, Jung-Beeman, Fleck, & Kounios, 2005) of the insight phenomenon.

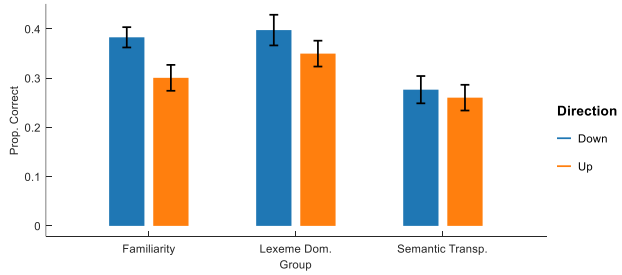


Figure 2: Differences in performance for each lexical property.

For further analysis, we used English compounds derived from the Touchstone Applies Science Associates (TASA) corpus and derived a lexicon of over 122,000 words, including hyphenated compounds. We used this lexicon to test whether a submitted response forms a valid compound against each individual cue presented, either as a prefix or suffix. The results of individual cue matches are shown in Figure 3, which demonstrates that the proportion correct for suffixes is smaller than that of prefixes. Further, submitted responses had a smaller likelihood of being valid prefixes for middle cues (M = 0.351, SD = 0.162) than for first cues (M = 0.411, SD = 0.182) and last cues (M = 0.402, SD = 0.174, $F(2,228) = 8.049$, $p < .001$). The estimated Bayes factor suggested that the data were .032:1 in favor of the alternative hypothesis, or rather, 31.25 times more likely to occur under the model including an effect for cue position than the model without it, providing strong evidence for its effect. This suggests that participants were alternating between cues when attempting to generate a solution, rather than using parallel processing. One possible explanation is that since cue presentation was staggered – and thus their search was guided – there may be primacy and recency effects whereby they were able to test and generate more candidate solutions following the first word cue, then worked backwards once all cues were presented using the third cue.

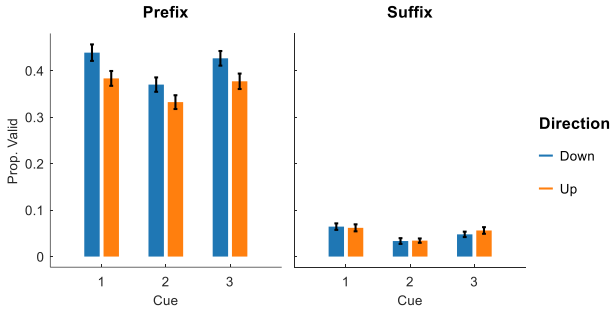


Figure 3: Proportion of valid prefixed (left) and suffixed (right) responses for each word cue position, according to lexicon.

Another interesting finding was that ratings of insight decreased as time elapsed throughout trials, as demonstrated in Figure 4. This finding holds for both correct and incorrect trials. This seemingly challenges the popular assertion that there must be a period of impasse, or mental block, preceding the experience of insight (Ohlsson, 1992). To the contrary, there were higher ratings of insight in the immediate time following the presentation of all three cues (i.e., 10-20 seconds) than in the time before the end of each trial (50-60 seconds). It is possible that participants simply rated solutions that they perceived to be correct as insightful *de facto* (hence, being submitted quickly), and correctly rejected the occurrence of insight for incorrect solutions proffered as a final guess before trials ended.

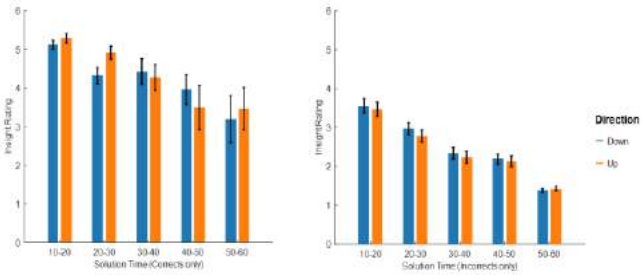


Figure 4: Insight ratings as a function of solution time (in secs.) for correct solutions (left) and incorrect solutions (right).

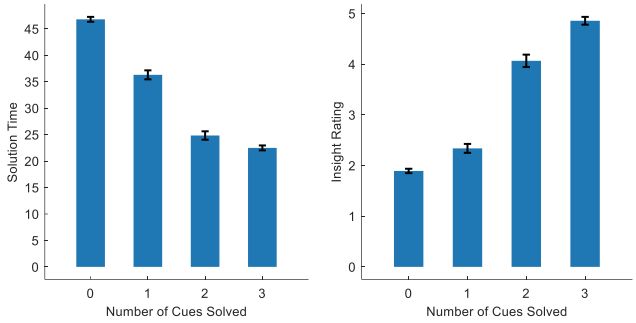


Figure 5: Solution time (in secs.) as a function of cues solved (left); Insight ratings as a function of cues solved (right).

Finally, there is the reporting of insight, itself. As demonstrated in Figure 5, fewer cues were likely to be solved as more time elapses in trials. The magnitude of reported insight also increased along with the number of cues correctly solved. Rather than an all-or-none experience – the "sudden, certain burst" frequently reported and used as a necessary criterion (Chronicle, MacGregor, & Ormerod, 2004) - it appears that participants used ratings of insight to indicate confidence in their answers. Indeed, these ratings increased as a function of the number of cues their proposed solution fit. There is not the presence of absolute insight for totally correct trials (in which all three cues are satisfied by the proposed solution), nor the absence of insight if this is not achieved. Rather, it exists on a continuum. This suggests more of an analytic approach, in which participants reliably monitor their progress in each problem and the likelihood of success using insight as a proxy for said progress. This contrasts previous research which states that incremental feelings of "warmth" do not precede moments of insight and are instead relegated to analytic or non-insightful problem solving strategies (Metcalfe & Wiebe, 1987). It should be noted again that a property of CRAs is that they can be solved with *or* without insight. What we argue here is the usefulness of insight ratings in CRAs to indicate perceived progress.

One limitation to the current work is that it used novel CRAs instead of those with established norms for difficulty and magnitude/frequency of reported insight (such as Bowden & Jung-Beeman, 2003). Applying lexical ratings to such a database for the dimensions present in Juhasz, Lai, and Woodock (2015) would be informative for future studies. Future research could also have subjects generate their own list of compounds given a set of word stems. Through doing this, researchers could collect latency data for how long people take to produce words, indicating their availability in memory. Researchers could also use LSA to analyze these participant-generated sequences of compounds to describe search behavior. These data could be applied across participants to establish cross-reliability and a more naturalistic set of items with norms.

All problems in the current study have suggested solutions that are the first lexeme in the compound. Since stem placement seems to matter in the processing of compounds (Taft & Forster, 1975), it may be beneficial to compile and use compounds with common prefixes and suffixes in future studies.

The self-identified magnitude of insight in our study was still based on subjective report. While this study focuses on the cognitive processes underlying these problems, rather than attempting to formalize insight in a significant manner, similar studies attempting to so may wish to include neural and/or physiological covariates to identify correlates of insight (e.g., EEG, fMRI, skin conductance, eye-tracking) (see Bowden et al., 2005 for suggested neurocognitive approaches). Future studies should also explore participants' differences in reporting thresholds, as one person may be more willing to identify the occurrence of insight than another. These individual differences could be applied to a signal detection theory model.

This study offers modest progress into understanding the linguistic contributors to CRA processing. There are other factors that should be investigated, such as if compounds with noun-noun links and adjective-noun links differentially affect performance. Other variables to investigate are word length effect (New et al., 2006), imageability, age of acquisition, sensory experience, or a combination of the above.

There may also be a reading direction effect present, as cue presentation always proceeded from left-to-right on the screen. To circumvent potential perceptual biases, future studies using a similar design may benefit from counterbalancing the order of reading direction, as well.

Lastly, it is important to be remain cognizant that not all insight problems are the same, and the phenomenology in CRAs may differ from that of other insight problems. It would be premature to make any sweeping statements about modeling insight from discoveries made in one class of problems.

Conclusion

If we are to solve the problem of insight, we must better understand the cognitive processes underlying the methods

we use to study it. Since we've largely neglected to explore these commonly-used procedures, we've defaulted to assumptions that they are "insight problems" simply because they elicit feelings of insight (based on the many and inconsistent criteria of researchers). While there have been both promising empirical and theoretical attempts to address this problem in recent years, much work remains. Better understanding the driving mechanisms, including lexical properties, within CRA problem solving will further inform us about how creativity is exercised and, perhaps, how insight is attained.

References

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., …. Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods, 39*, 445-459.

Batchelder, W. H., & Alexander, G. E. (2012). Insight problem solving: A critical examination of the possibility of formal theory. *The Journal of Problem Solving, 5*, 56-100.

Bowden, E. M., & Jung-Beeman, M. (2003). Normative data for 144 compound remote associate problems. *Behavior Research Methods, Instruments, and Computers, 35*, 634-639.

Bowden, E., Jung-Beeman, M., Fleck, J., & Kounios, J. (2005). New approaches to demystifying insight. *Trends in Cognitive Science, 9*, 322-328.

Butterworth, B. (1983). Lexical representation. In B. Butterworth (Ed.), Language production: Vol., 2. London: Academic Press.

Cai, D. J., Mednick, S. A., Harrison, E. M., Kanady, J. C., & Mednick, S. C. (2009). REM, not incubation, improves creativity by priming associative networks. *PNAS, 106*, 10130- 10134.444.

Chronicle, E. P., MacGregor, J. N., & Ormerod, T. C. (2004). What makes an insight problem? The roles of heuristics, goal conception, and solution recoding in knowledge-lean problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 14–27.

Corpus of Contemporary American English (COCA): http://corpus.byu.edu/coca

Fiorentino, R., & Poeppel, D. (2007). Compound words and structure in the lexicon. *Language and Cognitive Processes, 22*, 953-1000.

Griffiths, T. L., Steyvers, M., & Firl, A. (2007). Google and the mind: Predicting fluency with PageRank. *Psychological Science, 18*, 1069-1076.

Gupta, N., Jang, Y., Mednick, S. C., & Huber, D. E. (2012). The road not taken: Creative solutions require avoidance of high-frequency responses. *Psychological Science, 23*, 28-284.

Inhoff, A. W., Starr, M. S., Solomon, M., & Placke, L. (2008). Eye movements during the reading of compound words and the influence of lexeme meaning. *Memory and Cognition, 36*, 675-687.

Jeffreys, H. (1961). *Theory of probability* (3rd Ed.). Oxford, UK: Oxford University Press.

Ji, H., Gagné, C. L., & Spalding, T. L., (2011). Benefits and costs of lexical decomposition and semantic integration during the processing of transparent and opaque English compounds. *Journal of Memory and Language, 65*, 406-430.

Juhasz, B. J., White, S. J., Liversedge, S. P., & Rayner, K. (2008). Eye movements and the use of parafoveal word length information in reading. *Journal of Experimental Psychology: Human Perception and Performance, 34*, 1560-1579.

Juhasz, B. J., & Rayner, K. (2003). Investigating the effects of a set of intercorrelated variables on eye fixation durations in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*, 1312-1318.

Juhasz, B. J., Lai, Y. H., & Woodcock, M. L. (2015). A database of 629 English compound words: ratings of familiarity, lexeme meaning dominance, semantic transparency, age of acquisition, imageability, and sensory experience. *Behavior Research Methods, 47*, 1004-1019.

Kounios, J., & Beeman, M. (2014). The cognitive neuroscience of insight. *Annual Review of Psychology, 65*, 71-93.

Kuperman, V., Bertram, R., & Baayen, R. H. (2008). Morphological dynamics in compound processing. *Language and Cognitive Processes, 23*, 1089-1132.

Kuperman, V., Schreuder, R., Bertram, R., & Baayen, R. H. (2009). Reading polymorphemic Dutch compounds: Toward a multiple route model of lexical processing. *Journal of Experimental Psychology: Human Perception and Performance, 35*, 876-895.

Libben, G. (1998). Semantic transparency in the processing of compounds: Consequences for representation, processing, and impairment. *Brain and Language, 61*, 30-44.

Libben, G., Gibson, M., Yoon, Y. B., & Sandra, D. (2003). Compound fracture: The role of semantic transparency and morphological headedness. *Brain and Language, 84*, 50-64.

Mednick, S. A. (1962). The associative basis of the creative process. *Psychological Review, 69*, 220-232.

Metcalfe, J., & Wiebe, D. (1987). Intuition in insight an noninsight problem solving. *Memory and Cognition, 15*, 238-246.

New, B., Ferrand, L., Pallier, C., & Brysbaert, M. (2006). Reexamining the word length effect in visual word recognition: New evidence from the English Lexicon Project. *Psychonomic Bulletin & Review, 13*, 45-52.

Ohlsson, S. (1992). Information-processing explanations of insight and related phenomena. In M. T. Keane & K. J. Gilhooly (Eds.), *Advances in the psychology of thinking* (Vol. 1). New York: Harvester Wheatsheaf.

Olteteanu, A. M., Gautam, B., & Falomir, Z. (2015). Towards a visual remote associates test and its computational solver. In *AIC*.

Olteteanu, A. M., & Falomir, Z. (2015). comRAT-C: A computational Remote Associates Test solver based on language data and its comparison to human performance. *Pattern Recognition Letters*, *67*, 81-90.

Olteteanu, A. M. (2016). From Simple Machines to Eureka in Four Not-So-Easy Steps: Towards Creative Visuospatial Intelligence. In V. C. Müller (Ed.), Fundamental Issues of Artificial Intelligence. Cham: Springer International Publishing.

Sandra, D. (1990). In the representation and processing of compound words: Automatic access to constituent morphemes does not occur. *The Quarterly Journal of Experimental Psychology Section A*, *42*, 529-567.

Smith, K. A., Huber, D. E., & Vul, E. (2013). Multiply-constrained semantic search in the Remote Associates Test. *Cognition*, *128*, 64-75.

Steyvers, M., Shiffrin, R. M., & Nelson, D. L. (2005). Word association spaces for predicting semantic similarity effects in episodic memory. In A. F. Healy (Ed.), *Experimental Cognitive Psychology and its Applications*. Washington: American Psychological Association.

Taft, M., & Forster, K. I. (1975). Lexical storage and retrieval of prefixed words. *Journal of Verbal Learning and Verbal Behavior*, *14*, 638-647.

Topolinski, S., & Reber, R. (2010). Gaining insight into the "Aha" experience. *Current Directions in Psychological Science*, *19*, 402-405.

Topolinski, S., & Strack, F. (2008). Where there's a will – there's no intuition. The unintentional basis of semantic coherence judgements. *Journal of Memory and Language*, *58*, 1032- 1048.

Worthen, B. R., & Clark, P. M. (1971). Toward an improved measure of remote associational ability. *Journal of Educational Measurement*, *8*, 113-123.

Yap, M. J., & Balota, D. A. (2009). Visual word recognition of multisyllabic words. *Journal of Memory and Language*, *60*, 502-529.

Efficiency and Flexibility of Individual Multitasking Strategies - Influence of Between-Task Resource Competition

Jovita Bruening
Technische Universitat Berlin, Berlin, Germany

Marie Mckstein
Technische Universitat Berlin, Berlin, Germany

Dietrich Manzey
Technische Universitat Berlin, Berlin, Germany

Abstract

Evidence exists that individuals prefer distinguishable strategies for self-organized task scheduling in multitasking. They either prefer to work for long sequences on one task before switching to another (i.e., blocking), to switch repeatedly after short sequences (i.e., switching), or to process the current stimuli of two tasks before responding almost simultaneously (i.e., response grouping). We tested whether the strategies efficiency differs depending on the resource competition between tasks in a free concurrent dual-tasking paradigm and whether individuals adapt their strategies accordingly. Our results show that switcher and response grouper are more efficient than blocker during low than high resource competition between tasks. Comparably, more switchers shifted to a response grouping strategy than blockers towards a switching strategy. Overall, especially those individuals benefited from a lower resource competition, who already preferred a more flexible approach in dealing with the multitasking demand during high resource competition.

How Real is Moral Contagion in Online Social Networks?

Jason W. Burton (jburto03@mail.bbk.ac.uk)
Nicole Cruz (ncruzd01@mail.bbk.ac.uk)
Ulrike Hahn (u.hahn@bbk.ac.uk)

Department of Psychological Sciences, Birkbeck, University of London
Malet Street, London, WC1E 7HX, UK

Abstract

People increasingly turn to online social networks for information and debate. This means that the structures and properties of these networks, and the information they propagate, play crucial roles in the development of social beliefs, attitudes, and morals. Recently, research has shown that the presence of specific language drives the diffusion of moral messages, regardless of the informational quality, in a phenomenon dubbed *moral contagion* (Brady et al., 2017). Due to the widespread attention and implications of such findings for science and society, we investigate the presence of moral contagion across six sets of data that capture the communications of naturally-occurring networks on Twitter. Across a large corpus of diverse tweets (n = 525,229), we find moral contagion to be an inconsistent and often absent phenomenon that does not effectively predict message diffusion. The implications and reasons for this finding are discussed.

Keywords: moral contagion; social networks; social influence; computational social science; Twitter

Introduction

The advent of internet-based communications has permitted global connections at previously unfathomable speeds and volumes. While these connections present immense opportunities via global knowledge sharing and peer-to-peer collaboration, the magnitude of our interconnectedness also presents new liabilities, such as new forms of political persuasion.

The spread of psychological and behavioral phenomena is often likened to that of a pathogen moving from node to node, individual to individual, as a result of repeated exposure. Whereas concepts of "peer effects" and interpersonal influence have existed in psychology and sociology domains for quite some time (e.g., Allport, 1920; Redl, 1949), the formalization of *social contagion theory* has done well to shed light on the impacts that social networks have on everyday life. In a series of seminal studies, Christakis and Fowler (2007, 2008; Fowler & Christakis, 2009, 2010) utilized mass longitudinal datasets and network statistics to show that everything from obesity and smoking to happiness and cooperative behavior can cascade and cluster across social networks. From these findings, the development of collective behaviors, norms, and ideologies is understood to be a product of not only the aggregation of individuals, but also the topology of how individuals are arranged. For example, the proximity and volume of interpersonal ties of groups in social spaces, be it digital or not, increases the probability of both social information and behavior being transmitted amongst them. Simultaneously, this reinforcement of social homogeneity makes it difficult for intergroup connections to be made (i.e., echo chambers). As such, social contagion theory provides a lens through which the diffusion of information and the creation of collective intelligence can be examined.

Moral Contagion

In an interesting application of social contagion theory, Brady, Wills, Jost, Tucker, and Van Bavel (2017) present their conceptualization of *moral contagion*, which directly applies the process of social transmission to information diffusion. Extant literature suggests that morality is a powerful force in human reasoning and rationalization, with studies showing that one's moral beliefs are the foundation for one's ideology and political views (Graham, Haidt, & Nosek, 2009; Haidt, 2001). But how does an issue become moralized in the first place? Adopting the social intuitionist approach, Brady et al. (2017) explain that moral beliefs are less a product of private, individual reasoning and more the result of interpersonal processes and cultural norms (Haidt, 2001). What's more, they elaborate that the communication of moral ideas is tied to the use of emotion in social transmission. In other words, emotions, which serve as demonstrated contagions in social networks (e.g., Coviello et al., 2014; Ferrara & Yang, 2015; Kramer, Guillory, & Hancock, 2014), are highly associated with moral judgements and may serve as a segue to moralizing debates that would be otherwise nonmoral (Brady et al., 2017). In their analysis of a large corpus of Twitter communications, they find that not only does emotion drive the diffusion of moral content through social networks, but that the mere presence of moral-emotional words in a tweet increases its transmission by approximately 20% (Brady et al., 2017). Within the contemporary context of ideological polarization, the finding that moral-emotional language diffuses at such a high rate is concerning. As people increasingly rely on their online networks as news sources, blending spaces of socialization with information, the tendency of moral-emotional language to diffuse across networks means that feelings of outrage or disgust might be weaponized as tools of persuasion. Of course, there is a time and place for moralization, but the claims of Brady et al. (2017) suggest that their emotionally-driven moral contagion is highly impactful across domains, going so far

as to say that "it seems likely that politicians, community leaders, and organizers of social movements express moral emotions…in an effort to increase message exposure and to influence perceived norms within social networks" (p. 7316).

Because of the implications for society's ability to effectively reason and debate with contentious issues, the present study seeks to explore the prevalence of moral contagion across diverse, naturally-occurring social networks. More specifically, we aim to put the conclusions drawn by Brady et al. (2017) to the test by recreating their methodology and assessing whether moral-emotional language does in fact predict the diffusion of moral information regardless of quality or "truthiness."

Method

To investigate the presence of moral and emotional contagion in online social networks, an adaptation of Brady et al.'s (2017) methodology was employed. Specifically, we use the R programming language to recreate the main analysis strategy from Brady et al. (2017), reproduce their findings with their cleaned aggregated data, and then apply the analyses to five unique Twitter datasets that capture naturally-occurring social networks. Datasets and R scripts are made available at https://osf.io/943zm/.

Datasets

A total of six datasets were analyzed in this study. Four pre-existing datasets were obtained via the Open Science Framework (OSF) and Google's dataset search engine, which hosts links to a wide range of open data repositories. One dataset (*#MuellerReport*) was self-collected by connecting to the Twitter REST API with the `rtweet` package in R. While no specific dataset or topic was initially targeted, certain criteria were employed. To be considered for this study, datasets had to contain Twitter data (i.e., tweet messages and retweet counts), contain a significant number of messages written in English, and relate to a polarizing or morally-charged real-world issue, event, or social movement. Datasets were further narrowed by collapsing repeated messages into a single observation (to generate a composite diffusion count that combined the raw retweet count with the number of times the message appeared in the dataset) and removing non-English messages. Since the found datasets did not include language identifying metadata, the `textcat` package in R was employed to extract English tweets in these instances.

Brady et al. (2017) First and foremost, the present study drew directly from the recent study of moral contagion in social networks by Brady et al. (2017). Their data, which is generously shared on a public OSF project page, was thus crucial to the present study for both inspiration and corroboration. The data collected by Brady et al. (2017) focused on topical political issues in the United States: gun control, same-sex marriage, and climate change. Using the Twitter API and sets of topic-related filter words (e.g., *guns*, *gun control*, and *NRA* for the gun control topic), tweets and metadata were extracted between 30 October and 15 December 2015.

***#MeToo* Tweets** A second dataset comprised of Twitter messages containing the *#metoo* hashtag was obtained from the data.world repository. The raw dataset (n = 393,135) was extracted with the Twitter API between 29 November and 25 December 2017, little more than a month after the *#metoo* hashtag first appeared online in coordination with the "Me Too movement" (Turner, 2018). The "Me Too movement" is a movement against sexual harassment and assault. It was ignited by Hollywood sexual abuse allegations and has since become an international phenomenon garnering widespread media attention, support, and critique.

***#WomensMarch* Tweets** A third dataset with tweets containing the *#womensmarch* hashtag was also obtained from the data.world repository. Using the Twitter API, 15,000 messages were collected that referenced the pro-women's rights, and effectively anti-Trump, protest that took place in the wake of the presidential inauguration on 21 January 2017 (Adhokshaja, 2017). The Women's March has since become a worldwide movement with annual marches in late January to non-violently protest for women's reproductive rights, LGBTQ rights, immigration and healthcare reform, as well as racial, gender, and religious equality.

Post-Brexit Tweets A fourth dataset containing unfiltered tweets and metadata from the morning that Brexit was announced was obtained from the Mendeley Data repository. This unfiltered dataset (n = 17,998) was collected with NCapture from QSR, and employed a tight temporal parameter so as to capture the global public's reaction to the political event (Parker, 2017). Brexit refers to the result of the 2016 EU Referendum in the United Kingdom, and this dataset includes Twitter responses from across the globe.

Viral 2016 US Election Tweets A fifth dataset (n = 9,001) containing viral tweets (those with 1,000+ retweets) from the 2016 US Presidential Election was obtained from the Zenodo repository. The set of tweets was collected with the Twitter API and extracted messages that contained specific hashtags (*#MyVote2016*, *#ElectionDay*, and *#electionnight*) and/or user handles (*@realDonaldTrump* and *@HillaryClinton*) (Amador, Oehmichen, & Molina-Solana, 2017). This dataset was of special interest as it contained many "fake news" messages as coded by the curators, which one would might expect to use especially morally- and emotionally-charged language to garner extra attention.

***#MuellerReport* Tweets** A sixth dataset (n = 229,046) was collected by using the *#muellerreport* hashtag to retrieve

tweets from the Twitter API created between 23 and 25 March 2019 — the weekend during which US Attorney General William Barr released his summary of Special Counsel Robert Mueller's investigation into Donald Trump's 2016 presidential campaign. This corpus was of special interest because the Mueller Report has been (and at the time of writing, *still is*) a major source of controversy. While originally a non-polarized issue, the public opinion has divided overtime (Thomson-DeVeaux, 2019), meaning that moral-emotion might have played a part in moralizing conversations on Twitter.

Procedure and Analysis

All datasets were wrangled with R. Tweets were preprocessed with the `tm` and `dplyr` packages, and then a simple dictionary-based approach was employed to quantify the use of specific rhetoric in each message. To do so, the same three dictionaries used and validated by Brady et al. (2017) were used. One dictionary contains distinctly moral words and stems (n = 316; e.g., *fair*, *racism*, *family*), one contains distinctly emotional words and stems (n = 819; e.g., *panic*, *fear*, *heartwarming*), and one contains moral-emotional words and stems (n = 72; e.g., *shame*, *victimize*, *disgust*) that appeared in both of the original moral and emotional dictionaries (i.e., a subset of the moral and emotional dictionaries that was extracted to form the third unique dictionary). Through this categorization, "moral emotions" are considered distinct from "nonmoral emotions" because they are linked to triggers and functions specific to moral contexts, making them especially relevant to political debate (Haidt, 2003; Brady et al., 2017). For instance, outrage and disgust are often considered prototypical moral emotions because their expression can be elicited by perceiving a moral transgression, the breaking of some social axiom that threatens the collective order (e.g., infringement of human rights). In contrast, sadness is a nonmoral emotion because it can be triggered by nonmoral cues (e.g., the death of a loved one). The presence of these categorized words (moral, emotional, and moral-emotional) in each tweet was counted, so that each observation was coded with a discrete word count for each dictionary.

To accurately assess the degree to which each tweet in a given dataset diffused across the social media platform, the present study utilized a collapse-and-count scheme similar to that of Brady et al. (2017). Essentially, there are two measures of diffusion that can be calculated in an observational Twitter dataset: the retweet count displayed in collected metadata and the number of times a message appears in the dataset itself. Thus, the present study quantified diffusion by counting the presence of identical messages in each dataset, adding this count to the message's actual retweet count recorded in the metadata, and then collapsing repeated messages into a single observation.

To measure contagion effects, a negative binomial regression model was used. This model accounts for the overdispersion of data and effectively models count variables (i.e., discrete word counts and diffusion counts). In an effort to maintain consistency with Brady et al. (2017), incidence rate ratios (IRRs) were used as the ultimate indicator of the existence and magnitude of contagion effects. The `MASS` and `lmtest` packages were used for the main analysis.

Results

For the main analysis, negative binomial regression models with maximum likelihood estimation were fit onto each dataset to follow in line with the methodology of Brady et al. (2017), and to allow for a consistent measurement of moral, emotional, and moral-emotional contagion. The presence of contagion was determined by exponentiating the regression coefficients to generate incidence rate ratios (IRR) for each language dictionary, which were then used to plot diffusion prediction lines (Figure 1).

Brady et al. (2017) Across the corpus of 313,002 tweets, there was an average of 0.23 moral-emotional, 0.36 moral, and 0.69 emotional words per tweet. The main analysis indicated that moral contagion did indeed exist in the data collected by Brady et al. (2017). For distinctly moral language, there was a slight main effect (IRR = 1.02, $p <$ 0.05, 95% CI = 1.01, 1.04), and the same went for distinctly emotional language (IRR = 1.03, $p < 0.001$, 95% CI = 1.01, 1.04). Crucially, the presence of moral-emotional language appeared to have a strong effect on message diffusion (IRR = 1.21, $p < 0.001$, 95% CI = 1.19, 1.24). This result corroborates the statistics reported in Brady et al. (2017) to show that the use of moral-emotional language in tweets increases the likelihood of getting retweeted or otherwise shared among individuals in the social network platform by up to 21%. We also performed likelihood ratio tests to assess the statistical model's goodness of fit against nested univariate models that used *only* moral, emotional, *or* moral-emotional language as a predictor of diffusion (Table 1). These tests confirmed that the multivariate negative binomial regression model was effective for predicting message diffusion in the dataset.

***#MeToo* Tweets** After preprocessing the dataset, 151,572 unique tweets remained for analysis with an average of 0.21 moral-emotional, 0.30 moral, and 1.03 emotional words per tweet. The negative binomial regression model displayed a small but significant effect of distinctly moral language (IRR = 1.05, $p < 0.001$, 95% CI = 1.02, 1.09), as well as a significant effect of emotional contagion (IRR = 1.13, $p <$ 0.001, 95% CI = 1.11, 1.15) such that emotional language increased a message's diffusion by 13%. Curiously, while both moral and emotional language had a significant relationship with increased diffusion, moral-emotional language was significantly associated with message diffusion in a negative direction (IRR = 0.89, $p < 0.001$, 95% CI = 0.79, 0.85). Likelihood ratio tests confirmed that the multivariate model was the best fit for the dataset (Table 1).

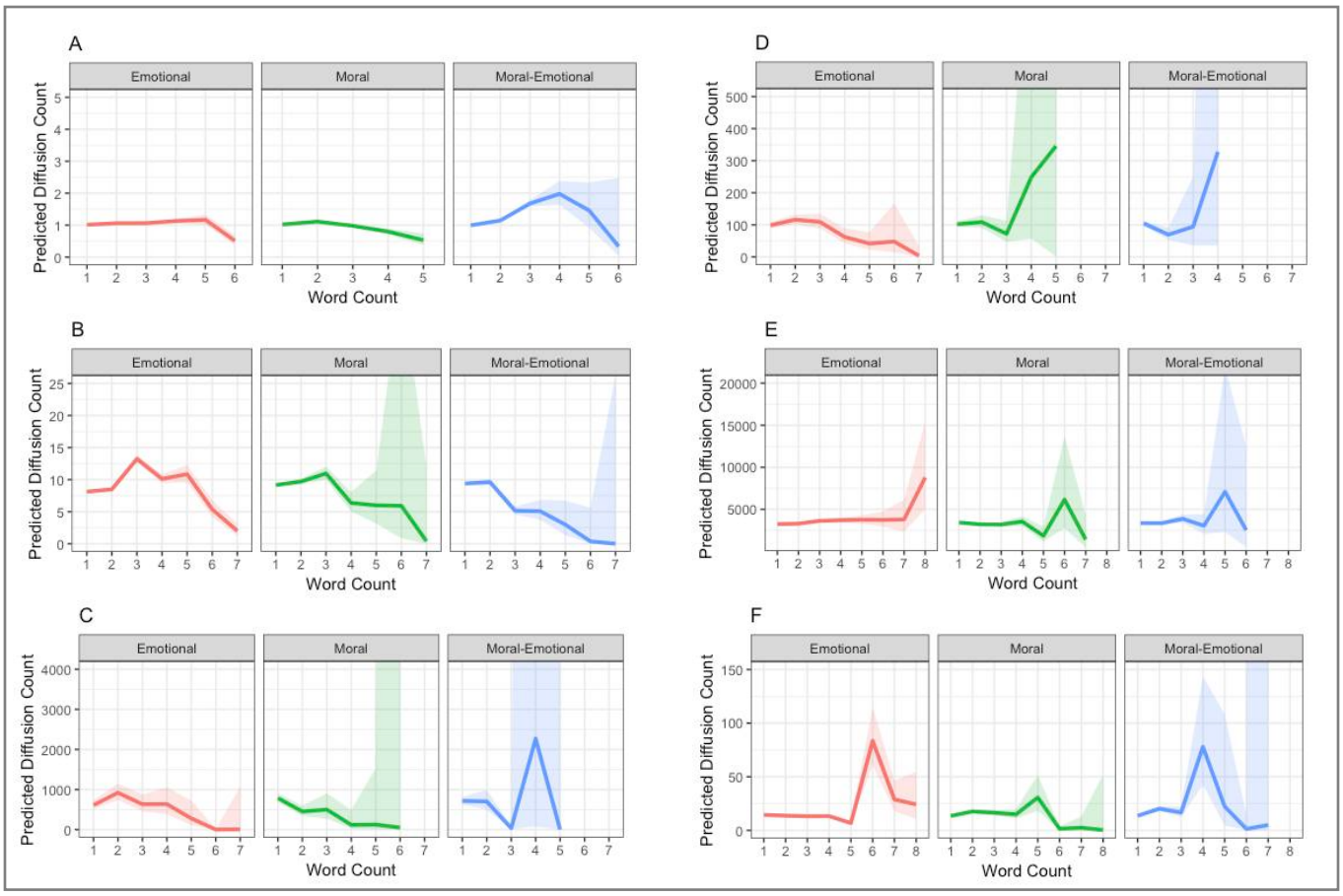


Figure 1: Predicted message diffusion trends as determined by multivariate negative binomial regression models. A = Brady et al. (2017) aggregate dataset; B = *#MeToo* dataset; C = *#WomensMarch* dataset; D = Post-Brexit dataset; E = Viral 2016 US election dataset; F = *#MuellerReport* dataset. 95% CIs are represented with the shaded areas. Note that scales vary widely as a result of the range of diffusion counts present in each dataset.

Post-Brexit Tweets Of the 7,124 analyzable tweets from the morning of the Brexit announcement, there was an average of 0.08 moral-emotional, 0.20 moral, and 0.69 emotional words per tweet. The regression model showed that there was no significant effect of moral language (IRR = 1.05, *n.s.*, 95% CI = 0.91, 1.22), emotional language IRR = 0.95, *n.s.,* 95% CI = 0.88, 0.1.04), or moral-emotional language (IRR = 0.86, *n.s.*, 95% CI = 0.71, 1.07) on message diffusion. While this goes against the claims of Brady et al. (2017), the large confidence intervals and low levels of moral, emotional, and moral-emotional language present in the dataset make the findings here generally inconclusive. In fact, a likelihood ratio test demonstrated that the main multivariate model was not suited to this corpus of tweets. Univariate models, where an isolated word dictionary alone is used to predict diffusion rates instead of the combined three, were slightly better at explaining the data (Table 1).

***#WomensMarch* Tweets** The 3,783 analyzable tweets, messages pertaining to the Women's March movement had an average of 0.17 moral-emotional, 0.31 moral, and 0.86 emotional words per tweet. Upon fitting the negative binomial regression model to the data, it was found that there was no effect of distinctly emotional (IRR = 0.98, *n.s.*, 95% CI = 0.86, 1.13) or moral-emotional language on diffusion (IRR = 0.95, *n.s.*, 95% CI = 0.72, 1.28). However, there was a significant negative effect of distinctly moral language on diffusion (IRR = 0.67, $p < 0.001$, 95% CI = 0.55, 0.82). This finding also contradicts that of Brady et al. (2017) and suggests that both emotional and moral contagion effects are domain specific. Likelihood ratio tests also indicated that a univariate model with *only* moral language was a better predictor of diffusion within the dataset (Table 1).

Viral 2016 US Election Tweets The 8,243 analyzable viral tweets from the 2016 US Presidential Election were found to have an average of 0.17 moral-emotional, 0.35 moral, and 0.98 emotional words per tweet. Analysis showed that there was indeed a small effect of emotional contagion (IRR = 1.05, $p < 0.001$, 95% CI = 1.03, 1.07), whereas the association with distinctly moral language decreased message diffusion (IRR = 0.96, $p < 0.01$, 95% CI = 0.93, 0.99). These findings set up what would have been an ideal

case for emotion to drive the diffusion of moral content, however the regression model showed that there was no significant relationship between moral-emotional language and message diffusion (IRR = 1.02, *n.s.*, 95% CI = 0.98, 1.06), despite hinting at an association in the expected positive direction. Nevertheless, likelihood ratio tests indicated that the multivariate model was the most appropriate predictor of the dataset, having outperformed each of the possible nested univariate models (Table 1).

***#MuellerReport* Tweets** In 41,505 unique analyzable tweets from the *#MuellerReport* corpus, an average of 0.18 moral-emotional, 0.47 moral, and 1.25 emotional words per message was found—the highest level of distinctly moral and distinctly emotional language of all datasets. Interestingly, a textbook moral contagion effect as per Brady et al. (2017) was found here. The negative binomial regression model showed that there was a significant effect of emotional contagion (IRR = 1.07, $p < 0.001$, 95% CI = 1.04, 1.09), and that the association between distinctly moral language and message diffusion was not statistically significant (IRR = 1.05, *n.s.*, 95% CI = 1.00, 1.11). And most importantly, there was a significant relationship between moral-emotional language and diffusion (IRR = 1.33, $p < 0.001$, 95% CI = 1.23, 1.46). This effect is even stronger than that of Brady et al. (2017), suggesting that the presence of moral-emotional language can increase a message's diffusion by 33%. Likelihood ratio tests supported the main multivariate model as the best explanation of the dataset (Table 1).

Aggregated Data Finally, we sought to rule out that the observed differences to Brady et al. (2017) were due to non-content differences in the samples, such as sample size. On top of analyzing each individual dataset, an aggregate dataset was compiled from the *#MeToo*, *#WomensMarch*, Post-Brexit, Viral 2016 US Election, and *#MuellerReport* datasets. This was done in an effort to present an analysis of a novel corpus that is similar in size to that addressed by Brady et al. (2017). However, it should be noted that statistics are skewed toward the *#MeToo* dataset as it is significantly larger than the others, comprising 71% of the aggregated data. Like the Brady et al. (2017) dataset, which captured discourse around multiple topics, this compilation of Twitter data also reaches a diverse range of contentious topics, as well as, in theory, a diverse range of individual Twitter users. This aggregation of five datasets into a single corpus (n = 212,227) displayed and average of 0.20 moral-emotional, 0.33 moral, and 1.05 emotional words per tweet. Analysis here showed that neither moral-emotional language (IRR = 0.90, $p < 0.001$, 95% CI = 0.87, 0.94), nor moral language (IRR = 1.00, *n.s.,* 95% CI = 0.97, 1.03), nor emotional language (IRR = 0.99, *n.s.,* 95% CI = 0.97, 1.00) predicted an *increase* in message diffusion. However, moral-emotional language was the only key variable that displayed a significant association with diffusion. This finding is reiterated by likelihood ratio tests, which showed that the multivariate model was slightly outperformed by a nested univariate model that used moral-emotional language only (Table 1). Further analysis with larger datasets, and examinations of specific moral-emotions (e.g., positive vs. negative affect; high- versus low-arousal) is planned for future studies in order to explore possible explanations.

Table 3: Likelihood ratio test statistics of deviance for goodness of fit. Significance indicates that the multivariate model (composed of moral-emotional language, distinctly moral language, *and* distinctly emotional language predicting diffusion) is the better fit for the dataset than the respective univariate model (composed of moral-emotional language, distinctly moral language, *or* distinctly emotional language predicting diffusion). The "Aggregate" column refers to the combined data from the *#MeToo*, *#WomensMarch*, Post-Brexit, Viral 2016 US Election, and *#MuellerReport* datasets. Significance codes: '***' < 0.001 '**' < 0.01 '*' < 0.05.

Univariate model	Dataset						
	Brady et al. (2017)	#MeToo	#Womens March	Post-Brexit	Viral US 2016 Election	#Mueller Report	Aggregate
Moral-emotional language only	30.07***	191.50***	13.06**	1.50	43.70***	27.73***	2.56
Distinctly moral language only	446.74***	212.59***	0.20	3.52	38.92***	86.72***	30.63***
Distinctly emotional language only	432.70***	43.62***	14.07**	2.12	9.62**	62.47***	26.85***

Discussion

Our results suggest that moral contagion driven by moral-emotional language is not as general a phenomenon as Brady et al. (2017) propose. In fact, the statistical models displayed no noteworthy effects of moral contagion in four of the six observational datasets analyzed. While the significant results of the likelihood ratio tests (Table 1) effectively link the use of moral, emotional, and moral-emotional language with information diffusion in most cases, the domain specificity of certain contagion effects in our results spurs a series of conceptual and methodological considerations.

Invoking morality in reasoning is known to harden existing belief structures, delegitimize authority, and, in extreme cases, dehumanize opposing perspectives (Ben-Nun Bloom & Levitan, 2011; Crockett, 2017). While morality can of course be a force for good—providing shared identities and guiding ethical behavior—the introduction of *unnecessary* morality and its emotional underpinnings can jeopardize rational debate. It is for this reason that moral justifications carry weight in some domains but not others. For example, loading an argument with moral-emotional language might be an effective strategy in discourse pertaining to human rights, yet that same strategy is likely to be penalized in an argument over mathematics. Sentiments about where morality is appropriate may be changing, and this may very well be a factor driving ideological polarization. But it seems unlikely a priori that moral language will be viewed the same in all domains. Our results are in keeping with such considerations.

There are also a number of methodological issues that potentially restrict the generality of studies such as this. Perhaps most conspicuous is the inability to parse true *causal* contagion from network homophily. The observational data used here and in Brady et al. (2017) fails to distinguish actual contagion (where exposure to a "contagious" condition has a causal effect on an individual's shift from state A to state B) from manifested homogeneity (where individuals with similar characteristics act in similar ways, irrespective of conditional exposure). It could be argued that the act of retweeting or sharing a message is a behavioral metric because it requires some motivated action. However, Brady et al. (2017) note that where moral contagion has been documented, it has been "bounded by [ideological] group membership" (Brady et al., 2017, p. 7313). This makes it important for future research to heed the substantial body of literature concerning the homophily-contagion problem (e.g., Aral, Muchnik, & Sundararajan, 2009; Shalizi & Thomas, 2011). Plus, Dehghani et al. (2016) specifically show that expressions of moral purity can predict the distance between users on Twitter, which further suggests that moral contagion may simply be an inadvertent measure of moral homophily. Along similar lines, the measurement of diffusion is also an imperfect operationalization of social influence. While collapsing repeated messages into a single observation ensures the language of a single repeated message does not skew analysis, it effectively penalizes unconventional retweeting (e.g., paraphrasing a message's content rather than clicking "retweet"), and is prone to overlook retweet chains (e.g., retweets of retweets) that might indicate true virality of a message (Brady et al., 2017). Crucially though, this imperfection applies to every dataset in the present study such that it cannot explain the discrepancies between datasets.

Needless to say, the use of social media analytics for investigations of the broader human condition has limitations with respect to external validity and representativeness (Tufekci, 2014). It is entirely possible that findings from studies conducted solely in the Twitterverse are in fact unique to the Twitterverse. Plus, the nature of Twitter metadata and correlational analyses like regression modelling mean that network agent variables (those pertaining to qualities of individual nodes/people) and structural variables like network topology are easily conflated. It may still be that human beings are susceptible to moral-emotionally-framed messages (Brady et al., 2017), but that unseen confounds, especially differences in network topology, can undermine contagion effects. Though the reverse is also possible, namely that contingent effects of topology may masquerade as a preference for moral-emotional language. Either way, the findings presented here demonstrate the need for a close partnership between descriptive accounts of "big data" analytics and controlled experimentation in order to draw confident conclusions about social rationality in the digitalized age.

Conclusion

Human reasoning is rarely, if ever, fully autonomous. We depend on our social environments for information and corroboration, and as these environments undergo digitalization, understanding how their evolution translates into new modes of influence is imperative for safeguarding spaces of rational debate. With high-profile papers (e.g., Brady et al., 2017) already pointing out concerning dynamics like moral contagion in real-world social networks, the present paper adds to this line of inquiry by illustrating the inconsistencies of such findings and offering theoretical and methodological explanations. Importantly, the results here indicate that given the diversity of naturally-occurring social networks, predicting the diffusion of information requires investigations of not only properties of the information itself, but also the domain specific topology of the networks through which it travels. Despite the limitations of current computational social science research, it is safe to say that exploring digital discourse can provide valuable insight into the state of human reasoning and argumentation in a time that has been labelled "post-truth."

References

Adhokshaja, P. (2017). #Inauguration and #WomensMarch. *Data.World*, (Version: a1689ca5). Retrieved from https://data.world/adhokshaja/inauguration-and-

womensmarch
Allport, F. H. (1920). The influence of the group upon association and thought. *Journal of Experimental Psychology*, *3*(3), 159–182. https://doi.org/http://dx.doi.org/10.1037/h0067891
Amador, J., Oehmichen, A., & Molina-Solana, M. (2017). Fakenews on 2016 US elections viral tweets (November 2016 - March 2017). *Zenodo*. https://doi.org/). http://dx.doi.org/10.5281/zenodo.1048826
Aral, S., Muchnik, L., & Sundararajan, A. (2009). Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences*, *106*(51), 21544–21549. https://doi.org/10.1073/pnas.0908800106
Ben-Nun Bloom, P., & Levitan, L. C. (2011). We're Closer than I Thought: Social Network Heterogeneity, Morality, and Political Persuasion. *Political Psychology*, *32*(4), 643–665. https://doi.org/10.1111/j.1467-9221.2011.00826.x
Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, *114*(28), 7313–7318. https://doi.org/10.1073/pnas.1618923114
Christakis, N. A., & Fowler, J. H. (2007). The Spread of Obesity in a Large Social Network over 32 Years. *New England Journal of Medicine*, *357*(4), 370–379. https://doi.org/10.1056/NEJMsa066082
Christakis, N. A., & Fowler, J. H. (2008). The Collective Dynamics of Smoking in a Large Social Network. *New England Journal of Medicine*, *358*(21), 2249–2258. https://doi.org/10.1056/NEJMsa0706154
Coviello, L., Sohn, Y., Kramer, A. D. I., Marlow, C., Franceschetti, M., Christakis, N. A., & Fowler, J. H. (2014). Detecting Emotional Contagion in Massive Social Networks, *9*(3), 1–6. https://doi.org/10.1371/journal.pone.0090315
Crockett, M. J. (2017). Moral outrage in the digital age. *Nature Human Behaviour*, 1–3. https://doi.org/10.1038/s41562-017-0213-3
Dehghani, M., Johnson, K., Hoover, J., Sagi, E., Garten, J., Parmar, N. J., … Graham, J. (2016). Purity Homophily in Social Networks. *Journal of Experimental Psychology: General*, *145*(3), 366. https://doi.org/http://dx.doi.org/10.1037/xge0000139 This
Ferrara, E., & Yang, Z. (2015). Measuring emotional contagion in social media. *PLoS ONE*, *10*(11), 1–14. https://doi.org/10.1371/journal.pone.0142390
Fowler, J. H., & Christakis, N. A. (2009). Dynamic spread of happiness in a large social network: Longitudinal analysis over 20 years in the Framingham Heart Study. *BMJ (Online)*, *338*(7685), 23–26. https://doi.org/10.1136/bmj.a2338
Fowler, J. H., & Christakis, N. A. (2010). Cooperative Behavior Cascades in Human Social Networks. *Proceedings of the Human Factors and Ergonomics Society*, *107*(12), 5334–5338. https://doi.org/10.1073/pnas.0913149107
Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and Conservatives Rely on Different Sets of Moral Foundations. *Journal of Personality and Social Psychology*, *96*(5), 1029–1046. https://doi.org/10.1037/a0015141
Haidt, J. (2001). The Emotional Dog and its Rational Tail: A Social Intuitionist Approach to Moral Judgment. *Psychological Review*, *108*, 814–834. https://doi.org/10.1016/j.bbrc.2005.09.115
Haidt, J. (2003). The Moral Emotions. In R. J. Davidson, K. R. Scherer, & H. H. Goldsmith (Eds.), *Handbook of affective sciences* (pp. 852–870). Oxford: Oxford University Press.
Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, *111*(24), 8788–8790. https://doi.org/10.1073/pnas.1412469111
Parker, C. (2017). Brexit Tweets from the morning of it's announcement. *Mendeley Data*, *v2*. https://doi.org/10.17632/x9wkrghz23.2
Redl, F. (1949). The phenomenon of contagion and "shock effect" in group therapy. In *Searchlights on delinquency* (K.R. Eissl, pp. 315–328). New York: International Universities Press.
Shalizi, C. R., & Thomas, A. C. (2011). Homophily and Contagion Are Generically Confounded in Observational Social Network Studies. *Sociological Methods & Research*, *40*(2), 211–239. https://doi.org/10.1177/0049124111404820
Thomson-DeVeaux, A. (2019). The Politics Surrounding Mueller Have Changed A Lot SInce He Started.
Tufekci, Z. (2014). Big Questions for Social Media Big Data: Respresentativeness, Validity and Other Methodological Pitfalls. In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media Big* (pp. 505–514). https://doi.org/10.1016/0022-5193(78)90170-4
Turner, A. (2018). 390,000 #MeToo Tweets. *Data.World*, (Version: 2249b175). Retrieved from https://data.world/balexturner/390-000-metoo-tweets

Politically Motivated Causal Evaluations of Economic Performance

Zachary A. Caddick (zacharycaddick@pitt.edu)
Benjamin M. Rottman (rottman@pitt.edu)
Department of Psychology, University of Pittsburgh
3939 O'Hara Street, Pittsburgh, PA 15260

Abstract

The current study seeks to extend research on motivated reasoning by examining how prior beliefs influence the interpretation of objective graphs displaying quantitative information. The day before the 2018 midterm election, conservatives and liberals made judgments about four economic indicators displaying real-world data of the US economy. Half of the participants were placed in an 'alien cover story' condition where prior beliefs were reduced under the guise of evaluating a fictional society. The other half of participants in the 'authentic condition' were aware they were being shown real-world data. Despite being shown identical data, participants in the Authentic condition differed in their judgments of the graphs along party lines. The participants in the Alien condition interpreted the data similarly, regardless of politics. There was no evidence of a 'backfire' effect, and there was some evidence of belief updating when shown objective data.

Keywords: motivated reasoning; politics; biases; reasoning; decision-making

Introduction

Previous research has shown that individuals often reason differently about information depending on whether it is congruent with their prior beliefs. Individuals tend to more easily accept information that is congruent with prior beliefs and desires and discount information that is incongruent with prior beliefs and desires. This process is known as motivated reasoning. In the current research, we studied the influence of political attitudes on how people interpret time series graphs of the economy. This research is at the intersection of two fields: causal reasoning about time series data, and motivated reasoning.

Motivated Reasoning and Causal Reasoning: Similarities and Differences

The fields of motivated reasoning and causal reasoning have long been intimately connected in certain ways, yet also distant in other ways. The current research aims to advance both of these fields, and to advance research on the intersection of the two.

In one aspect, these two fields have studied similar questions about the role of prior beliefs and desires on the acceptance or rejection of new information. On the causal reasoning side, there has been considerable research into how people incorporate new information with prior causal beliefs (e.g., Alloy & Tabachnik, 1984). Furthermore, many of the particular topics that have been studied in the field of motivated reasoning have had to do with causal or at least predictive relations. For example, in a seminal work on motivated reasoning, Kunda (1987) found that people tend to believe that other people who have attributes similar to themselves are less likely to get divorced than people with dissimilar attributes. Note how in this study, the attribute is as a potential cause or predictor of the effect (divorce). Other research on motivated reasoning that is less directly related to causation still often studies acceptance of causal-scientific explanations, for example, about global warming (Campbell & Kay, 2014).

On the other hand, there are also important differences between these fields. First, causal learning has traditionally been focused on the rational (Bayesian) updating of beliefs given new information, whereas motivated reasoning has focused on affective reasons for failing to update beliefs.

A second difference, more relevant to the current research, is that most research on causal reasoning has focused on the inferential process - how a learner infers a cause-effect relationship from a set of data. In contrast, research on motivated reasoning does not involve inference. Instead, participants are typically presented with a fact or a set of facts, and the question is whether participants accept or reject the facts (e.g., Ranney & Clark, 2016).

One recent study on motivated reasoning has investigated inference from data, similar to causal reasoning research. Kahan, Peters, Dawson, and Slovic (2017) presented participants with quantitative information in 2x2 contingency tables about the number of cities that did or did not ban handguns in public and whether there was an increase or decrease in crime, and participants were asked to infer the relation between gun bans and crime. Despite being presented with quantitative data, participants were more likely to make correct inferences when the data supported their prior attitudes about guns. The current research is in a similar vein–it investigates the role of political attitudes on inferences about economic trends.

Motivated Reasoning about Economics

The political arena is an especially ripe medium for motivated reasoning to occur, and has been one of the most studied types of motivated reasoning. Politically-relevant stimuli also provides a unique opportunity to study the intersection of motivated and causal reasoning about objective quantitative data that has high ecological validity.

Politicians often make competing statements about the credit or blame for the same economic outcomes. For example, in a speech to democratic supporters, former President Obama said: "…when you hear how great the economy's doing right now, let's just remember when this recovery started" (USA Today, 2018). In contrast, Kevin Hassett, the Chairman of The White House Council of

Economic Advisers, has stated "I can promise you that economic historians will 100 percent accept the fact that there was an inflection at the election of Donald Trump and a whole bunch of data items started heading north" (Horsley, 2018). Similarly, citizens also interpret the same economic outcomes based on political lenses. For example, Republicans interpret the 2017 tax bill as having more personal benefit than Democrats (Bump, 2018).

The current experiment is a controlled study to understand how people view the exact same economic data in different ways based on political orientation. There is little research into the cognitive processes engaged in motivated reasoning about objective economic data. In this study, we assessed politically-motivated reasoning before and after participants viewed economic time-series graphs, and after making judgments about the impact of each president.

Backfire Effects

One concern with the possibility of presenting participants with objective data is that it might actually produce a "backfire" effect in which the participant doubles-down and strengthen their prior belief. For example, Nyhan and Reifler (2015) found that participants who had previously held high levels of concern about potential side effects of flu vaccinations became less likely to get flu vaccinations after exposure to corrective information. However, evidence for the backfire is mixed. A more thorough investigation by Wood and Porter (2018) found no evidence of backfire effects. These two studies on backfire effects used text-based presentation of facts. In the current study we assess whether participants exhibit backfire effects when presented with economic time series data that require them to make an inference. Whether participants exhibit a backfire effect could help reveal whether such information might be useful for changing voters' opinions.

Current Study

In the current study, participants were shown time series graphs of economic variables, and the graphs denote the times when Democratic vs. Republican presidents held office. Participants were asked questions about whether Democrats or Republicans were better for the economy. This study allowed us to ask a number of questions that provide insight into motivated causal reasoning.

First, will people learn from the time series graphs and change their beliefs about which party is better, or will they exhibit a 'backfire' effect? This question is especially relevant for political campaigns wondering how objective economic data changes voters' opinions. One reason that a backfire effect could happen is because quantitative graphs always ignore some contextual information, and people may latch onto such factors to reinforce their prior beliefs. For example, in the current study, presidents only have limited control of the economy and there are other external factors (e.g., Congress, the Federal Reserve, international politics).

Second, to what extent do people engage in motivated reasoning even about highly objective, quantitative data? Participants were asked questions at multiple levels of granularity, from fairly general about Democratic vs. Republican presidents in general, to the influence of particular presidents, which could potentially show different degrees of motivated reasoning.

Third, the current research also provides a unique opportunity for research on causal reasoning. Recently there has been more research on causal reasoning about time series data (Rottman, 2016; Soo & Rottman, 2018). One of the challenges involved in making causal inferences in general, and from time series data in particular, is that the data are often ambiguous and can be interpreted in multiple ways. The current study extends prior research in two ways. First, it provides new methods for studying how people reason about real-world time-series data (as opposed to researcher-generated data). Second, it is the first causal reasoning study we know of that explicitly studies the role of motivated reasoning in causal attribution.

Methods

Participants

On November 5th, the day before the 2018 United States midterm election, 403 participants were recruited via Amazon's Mechanical Turk. They were paid $4 for participating in this study. Mechanical Turk premium qualifications were used to sample 200 individuals who had previously identified as liberals, and 200 who identified as conservatives. Three participants completed the study without accepting the HIT, resulting in 403 participants.

Stimuli and Design

Participants reasoned about time series graphs of four economic variables (Figure 1), within-subjects. Each graph depicted the period from 1977 through the most recent economic data when the study was conducted. We had to choose a year to begin the graphs. We wanted to include Reagan because of his important role in current political-economic debates, but we figured that the current electorate would probably have less partisan views about Carter and earlier presidents. Carter was included because one of the questions about Reagan requires having a trend line before Reagan took office. The graphs were accompanied with hyperlinks to the data sources to increase transparency. Unlike Figure 1, the colors of the two parties were red and blue for the Republican and Democratic parties. A brief explanation of each economic variable was included.

We desired to be able to compare participants' motivated reasoning against a more objective condition in which motivated reasoning is eliminated. To do this, half of the participants were presented with 'authentic' graphs like those in Figure 1. The other half saw graphs just like Figure 1, except the origin of the data was disguised; participants were told that the data came from a fictional alien society. Made-up alien names were used for the political parties and

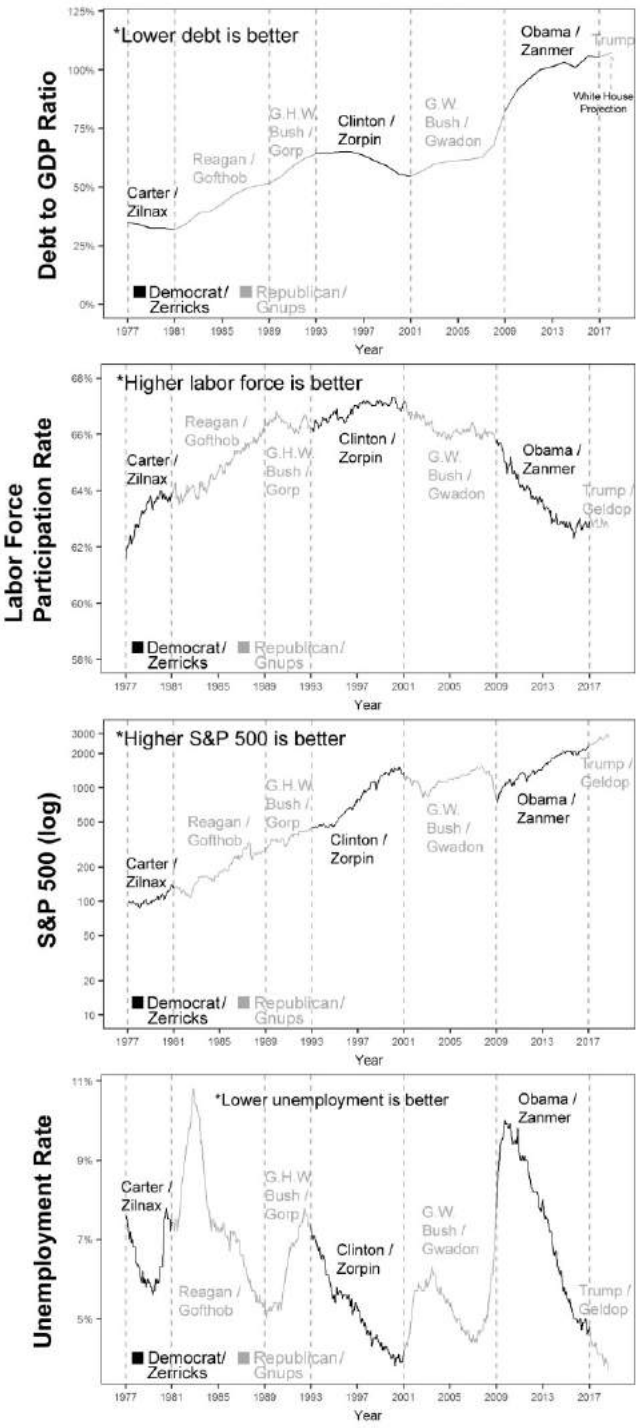


Figure 1: Stimuli for Authentic and Alien conditions combined together. The text "White House Projection" on Debt to GDP Ratio graph was only present for the Authentic condition. Alien graphs had a range of 3061-3621 years.

presidents. The alien graphs did not include the hyperlinks, and the colors of the two parties were green and orange (not red and blue) to reduce suspicion.

Procedure

Figure 2 provides a summary of the procedural flow of the

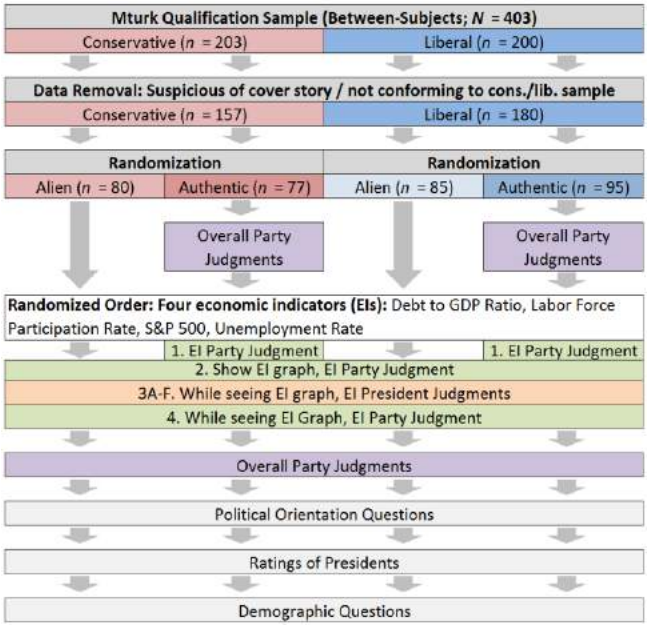


Figure 2: Study procedure flow. "EI" = economic indicator.

study. First, participants from both the liberal and conservative samples were randomized to either 'Authentic' or 'Alien cover story' conditions. Participants in the Alien condition were told to "*Please imagine that in roughly 1,000 years, there is an alien society on another planet. The government of this society has two political parties, Zerricks, and Gnups. Your goal for this study is to figure out whether the Zerricks or Gnups parties generally do a better job of handling the society's economy…*" Participants in the Authentic condition were not provided with an introduction, as we presumed that participants were acquainted with the major US political parties.

Next, participants in the Authentic condition completed an "Overall Party Judgment" (*"Which political party do you believe is better for the economy overall*?": 1 = strongly believe Rep., 4 = neutral, no opinion, 7 = strongly believe Dem.). Alien condition participants did not make this judgment as they had no prior beliefs about the fictional aliens.

Next, participants completed blocks of judgments about each of the four economic indicators in a random order. In Step 1, participants in the Authentic but not Alien condition made the Economic Indicator Party Judgment ("*Which political party do you believe is better for [econ. indicator]*?": 1 = strongly believe Rep., 4 = neutral, no opinion, 7 = strongly believe Dem.). Then, in Step 2, participants were shown an economic indicator graph like in Figure 1. With the graph still visible, they made another Economic Indicator Party Judgment.

In Step 3 (sub-steps: A-F), with the graph still visible, participants made judgments about the influence of each president (from Reagan to Trump; Figure 3). This 14-option question allowed participants to make precise judgments about the nature of the change in the trend line. If participants' judgments are still influenced by their political

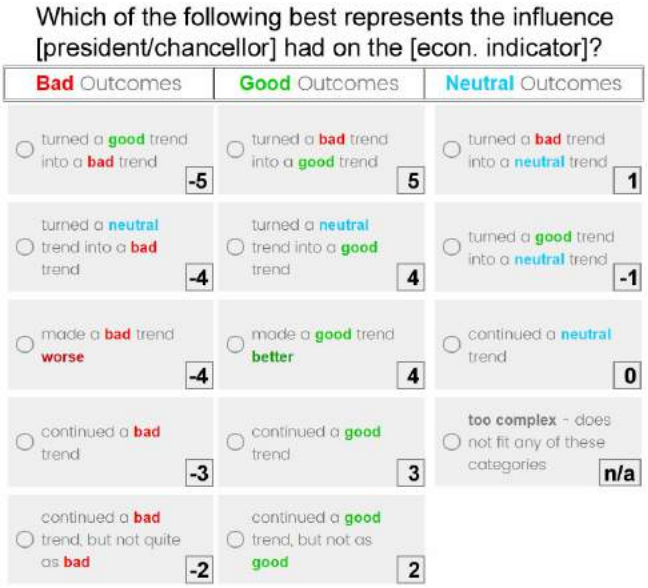


Figure 3: Economic Indicator President Judgments. The numerical coding scale was hidden during the study.

orientation, it would mean that motivated reasoning has an influence on even very low-level causal reasoning. Because prior research (Soo & Rottman, 2018) has found that people focus on changes in trends more than absolute levels, this question asked how trend in the variable changed during the president's time in office compared to before. To analyze these judgments, we turned the 14-option scale into a -5 to +5 scale, where +5/-5 means that the president had a very good/ bad influence on the trend. The numbers in Figure 3 display this scale mapping.

In Step 4, participants made the Economic Indicator Party Judgment one last time with the graph still present. The reason for asking this question three times was to see if participants' judgments become less biased with more exposure to the data and thinking about the data.

After completing the questions about the four economic indicators, participants made a final Overall Party Judgment without any graphs presented alongside this question.

Participants went on to complete four questions on political orientation. We used one of these questions (1 = extremely liberal, 4 = moderate/middle of the road, 7 = extremely conservative) to ensure that the participants' current political orientation matched the MTurk Qualification. Afterwards, participants rated how much they "liked" each of the presidents and completed demographics.

Results

Participants

Participants in the Alien condition were asked about degree of suspicion for the cover story after completing the study. Fifteen participants were dropped from analyses due to selecting that they "strongly suspected that the data reflected the United States." The remaining participants were included in analysis. Fifty-seven participants selected "I was a bit suspicious, but I didn't think much of it." The majority (n = 108) selected "No, I did not suspect that the data reflected the United States."

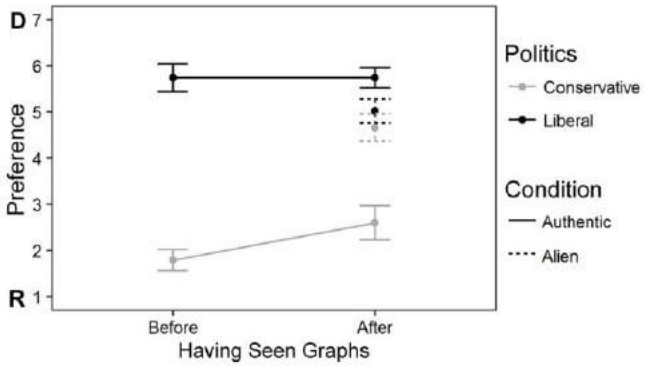


Figure 4: Overall party judgments before and after seeing time series graphs. Error bars are 95% confidence interval. D/R denote Dem./Rep. leaning judgments.

Another 53 participants were dropped because their responses to a question about whether they identified as conservative or liberal did not conform to how they previously identified according to the qualification. Thus, 337 participants were included in the analyses.

Conservative (M = 43; SD = 12) and liberal (M = 38; SD = 11) participants had similar ages. Conservative and liberal samples were predominately white (91% & 83%, respectively). Liberals were a bit more educated. The education breakdown was as follows for conservatives and liberals, respectively: high school or lower (19% vs. 9%), some college but no degree (23% vs. 27%), associate's or bachelor's (47% vs. 54%), and master's or higher (11% vs. 9%).

Statistics

For all the following analyses that used mixed-effect models, we used the R packages 'lme4', 'lmerTest' for p-values, and 'r2glmm' for R^2_{NSJ} effect sizes (Jaeger, 2017). Effects coding was also used for all mixed-effect models.

Overall Party Judgments

Figure 4 presents the overall party judgments. The overall impressions of the graph are as follows. First, there do not appear to be differences in the Alien condition by politics. We note that participants' ratings in the Alien condition are more favorable to Democrats (the means of the Alien condition are above the midpoint of the scale).[1] This suggests that we should expect to see more changes in beliefs for conservatives rather than liberals. Second, in the Authentic condition, there are large differences between

[1] The fact that participants in the Alien condition tended to believe that the Democrats (disguised as aliens) were better for the economy than Republicans (disguised as aliens) is intended merely as a summary of the stimuli used in this experiment, not as a political statement. There are many other economic indicators aside from these four, and there are historical events not depicted on the graphs that could affect their interpretation.

conservatives and liberals, though the difference appears to become somewhat smaller suggesting that participants are learning rather than having a backfire effect. Third, even after seeing the graphs, there still appear to be large differences by political orientation in the Authentic condition. We now assess these questions statistically.

First, overall party judgments between liberal and conservative participants were not significantly different after seeing the graphs for the Alien condition, $t(159.05) = 1.80$, $p = .074$, $d = .28$. The remaining analyses focus on the Authentic condition.

Second, we tested whether the judgments changed over time by doing a regression with time (before vs. after seeing the graphs), political orientation, and the interaction, and a by-subject random intercept. There was a main effect of politics ($\beta = 3.54$, $SE = .18$, $t = 19.71$, $p < .001$, $R^2_{NSJ} = .635$) implying strong politically-motivated reasoning. There was a significant effect of time ($\beta = .40$, $SE = .10$, $t = 4.08$, $p < .001$, $R^2_{NSJ} = .022$). Most importantly, there was a significant interaction between politics and time ($\beta = -.81$, $SE = .20$, $t = 4.08$, $p < .001$, $R^2_{NSJ} = .022$), implying that the two groups moved *closer together* after seeing the graphs. Third, even after seeing the graphs, overall party judgments between liberal and conservative participants were still significantly different, $t(125.66) = 14.33$, $p < .001$, $d = 2.13$.

Economic Indicator Party Judgments

Figure 5 shows graphs of the judgments of which party is better at controlling each of the four economic indicators. These judgments were made at three timepoints in the Authentic condition, and at two timepoints in the Alien condition. The overall impressions of the graphs are as follows. First, for the most part, the differences in the Alien condition by politics are small, if present at all.[2] Second, there appear to be some changes in beliefs after seeing the graph (Step 1 to Step 2), but there are few changes after making the president judgments (Step 2 to Step 4). For this reason, we just focus our analyses below on Steps 1 and 2. Third, even after seeing the graphs, there are still substantial differences between liberals and conservatives. We now test these impressions statistically.

First, at Step 2, we tested whether there are any differences based on political orientation within the Alien condition. We conducted a linear regression with a by-subject random intercept and a by-economic-indicator random intercept. There was no significant effect of politics ($\beta = .29$, $SE = .23$, $t = 1.28$, $p = .278$, $R^2_{NSJ} = .01$).

We then tested for motivated reasoning within the Authentic condition in Steps 1 and 2. We tested for the main effects and interaction of Time and Politics. We used a linear regression with by-subject random intercepts and

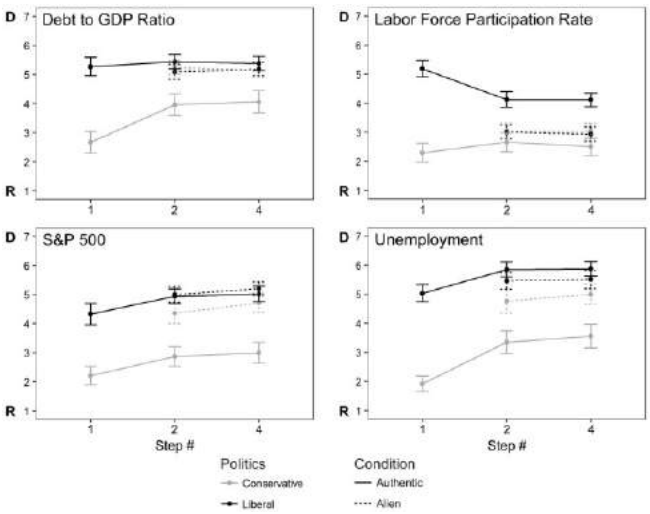


Figure 5: Mean economic indicator party judgments and 95% CIs. D/R denote Dem./Rep. leaning judgments.

slopes for time. The model also had by-economic-indicator random intercepts with random slopes for time and politics, and the interaction. We found significant differences in economic indicator party judgments based on participants' political orientation ($\beta = 2.28$, $SE = .22$, $t = 10.23$, $p < .001$, $R^2_{NSJ} = .344$), implying politically-motivated reasoning. There was not an effect of Time from Step 1 to Step 2 ($\beta = .54$, $SE = .32$, $t = 1.69$, $p = .188$, $R^2_{NSJ} = .028$). There was a marginal interaction between politics and Time ($\beta = -.80$, $SE = .32$, $t = 2.48$, $p = .082$, $R^2_{NSJ} = .016$). Perhaps there was a trend that conservatives and liberals' beliefs moved closer together after seeing the graphs, though this was only evident for some of the economic indicators (Figure 5).

Lastly, we tested for differences in economic indicator party judgments for the Authentic condition at Step 2 (after seeing graphs). This model included a by-subject random intercept and a by-economic-indicator random intercept and slope for politics. There was still a significant effect of politics on economic indicator party judgments ($\beta = 1.88$, $SE = .28$, $t = 6.71$, $p = .001$, $R^2_{NSJ} = .260$).

In sum, participants' judgments were biased by politics, and there was a trend of becoming less biased after seeing the graphs.

Economic Indicator President Judgments

Participants judged how each president influenced each EI. Because these judgments were very specific, they should be less open to interpretation than the other judgments. We wanted to test whether participants' political motivations would still affect these judgments. To test this, we reverse coded the judgments about Republican presidents. This means that judgments that are higher on the -5 to +5 scale are more positive towards Democrats, and judgments that are lower are more positive towards Republicans.

Participants made 24 judgments (6 presidents x 4 EIs). There were large differences across these 24 items because certain indicators performed very well or very poorly during

[2] Three of these economic indicators were more favorable for Democrats (the judgments in the Alien condition are higher than the midpoint of the scale for all but the Labor Force Participation Rate). This means that we would expect more changes in beliefs for Republicans than for Democrats in all but the Labor Force Participation Rate judgments.

certain presidencies. We used mixed effects models with by-item random intercepts and slopes for politics to account for the 24 items and by-subject random intercepts.

There was a significant effect of political orientation in the Authentic condition (β = .97, *SE* = .17, *t* = 5.87, *p* < .001, R^2_{NSJ} = .016). However, there was also a significant effect of political orientation in the Alien condition (β = .38, *SE* = .13, *t* = 2.96, *p* = .004, R^2_{NSJ} = .002). It is possible that some participants in the Alien condition realized that the alien chancellors were actually American presidents but did not report being highly suspicious.

To test whether there was more political bias in the Authentic condition than the Alien condition, we ran a model that also included condition and politics and their interaction as a by-item random slopes. There was no significant effect of condition (β = .17, *SE* = .11, *t* = 1.50, *p* = .138, R^2_{NSJ} = .001). We found a significant effect of politics; the judgments in the Liberal sample were a bit more favorable to Democrats (β = .68, *SE* = .12, *t* = 5.63, *p* < .001, R^2_{NSJ} = .008). Most importantly, there was a significant interaction between condition and politics, (β = .59, *SE* = .19, *t* = 3.14, *p* = .002, R^2_{NSJ} = .001). This suggests that economic indicator president judgments of conservatives and liberals were farther apart for the Authentic condition, and that there still is an effect of motivated reasoning even for judgments about specific presidents and specific economic indicators.

Discussion

Previous research has shown that individuals tend to preferentially view evidence congruent with prior beliefs and de-emphasize incongruent evidence. Our findings support that perceptions of quantitative information are influenced by the presence of prior beliefs. When prior beliefs were absent in the Alien condition, participants' judgments were largely in agreement with one another. However, when making judgments about US political parties, our participants' judgments were strongly influenced by their political beliefs.

The Economic Indicator President Judgments may offer the most supportive evidence of motivated reasoning as participants engaged in belief maintenance even when making very specific judgments (e.g., President X "changed a neutral trend into a bad trend"), implying that prior beliefs can influence even low-level perceptual judgments. However, the bias for these very specific judgments were not as strong as for the overall party judgments and economic indicator party judgments.

Despite the evidence of motivated reasoning, our participants did change their initial beliefs after viewing an objective graph, at least for the overall judgments. This suggests that presenting people with objective time series graphs of the economy might be a useful strategy for changing voters' minds. Perhaps another more radical strategy to change opinions is to show voters time series graphs with the political parties disguised, like in the Alien condition, to help them make judgments in a bias-free context, before revealing the political parties. In future research we plan to test whether this strategy is effective.

It is important to note that even though we have been calling the effects in the paper "politically motivated," the current results cannot distinguish between rational use of prior beliefs versus self-protective motivational forces. One view is that the liberal versus conservative participants have different prior knowledge (e.g., about other relevant economic factors that could have been causes of changes in the graphs), and interpret the graphs differently based on their different knowledge (Jern, Chang, & Kemp, 2014). The other view, which is traditionally called 'motivated reasoning' is that they interpreted the graphs differently simply to support self-serving desirable outcomes (i.e., protecting their political self-identity).

However, we believe the current results are still useful in that they show how disparate of views people can have making judgments from quantitative data (as opposed to prior research that used text-based stimuli). Another novel feature of this study is that it involved making inferences or generalizations, whereas prior research has focused simply on subjects acceptance of a textual argument.

More generally, given the current time of heightened polarization and misinformation, more research is needed to understand biased reasoning and find interventions to reduce biased reasoning about quantitative information.

References

Alloy, L. B., & Tabachnik, N. (1984). Assessment of covariation by humans and animals: The joint influence of prior expectations and current situational information. *Psychological review*, *91*(1), 112.

Bump, P. (2018, September 24). Republicans claim they saw a tax cut. Democrats claim they didn't. Retrieved January 10, 2019, from https://www.washingtonpost.com/politics/2018/09/24/republicans-claim-they-saw-payroll-tax-cut-democrats-claim-they-didnt

Campbell, T. H., & Kay, A. C. (2014). Solution aversion: On the relation between ideology and motivated disbelief. *Journal of Personality and Social Psychology*, *107*(5), 809–824.

Horsley, S. (2018, September 12). Fact Check: Who Gets Credit for the Booming U.S. Economy? Retrieved January 10, 2019, from https://n.pr/2p5Gntz

Jaeger, B. (2017). R2glmm: Computes R squared for mixed (multilevel) models. *R package version 0.1.2*.

Jern, A., Chang, K. K., & Kemp, C. (2014). Belief polarization is not always irrational. *Psychological Review*, *121*(2), 206–224.

Kahan, D. M., Peters, E., Dawson, E. C., & Slovic, P. (2017). Motivated numeracy and enlightened self-government. *Behavioural Public Policy*, *1*(01), 54–86.

Kunda, Z. (1987). Motivated inference: Self-serving generation and evaluation of causal theories. *Journal of personality and social psychology*, *53*(4), 636.

Nyhan, B., & Reifler, J. (2015). Does correcting myths about the flu vaccine work? An experimental evaluation

of the effects of corrective information. *Vaccine*, *33*(3), 459-464.

Ranney, M. A., & Clark, D. (2016). Climate change conceptual change: Scientific information can transform attitudes. *Topics in Cognitive Science*, *8*(1), 49–75.

Rottman, B. M. (2016). Searching for the best cause: Roles of mechanism beliefs, autocorrelation, and exploitation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(8), 1233.

Soo, K., & Rottman, B. M. (2018). Causal Learning from Trending Time-Series. In C. Kalish, M. Rau, J. Zhu, and T. T. Rogers (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

USA Today. (2018, September 08). Read transcript of former President Obama's speech, blasting President Trump. Retrieved January 10, 2019, from https://usat.ly/2wUHorw

Wood, T., & Porter, E. (2019). The elusive backfire effect: Mass attitudes' steadfast factual adherence. *Political Behavior*, *41*(1), 135-163.

Speech Processing does not Involve Acoustic Maintenance

Spencer Caplan
University of Pennsylvania, Philadelphia, Pennsylvania, United States

Alon Hafri
University of Pennsylvania, Philadelphia, Pennsylvania, United States

John Trueswell
University of Pennsylvania, Philadelphia, Pennsylvania, United States

Abstract

What happens to the acoustic signal after it enters the mind of a listener during real-time speech processing? Since processing involves extracting linguistic evidence from multiple, temporally distinct sources of information, successful communication relies on a listeners ability to combine these potentially disparate signals. Previous work has shown that listeners are able to maintain, and rationally update, some type of intermediate representations over time. However, exactly what type of information is being maintainedbe it acoustic-phonetic or rather a probability distribution over phonemeshas been underspecified. In this paper we present a perception experiment aimed at identifying the internal contents of intermediate representations in speech processing. Using an accent-adaptation paradigm, we find that listeners adapt to modulated acoustic signal when the corresponding orthography is provided before the audio, but not when audio follows the orthography. This supports the position that intermediate representations are uncertainty-distributions over discrete units (e.g. phonemes) and that, by default, speech processing involves no maintenance of the acoustic-phonetic signal.

The emergence of monotone quantifiers via iterated learning

Fausto Carcassi[1,*], **Shane Steinert-Threlkeld**[2,*], **Jakub Szymanik**[2]

fausto.carcassi@gmail.com; {S.N.M.Steinert-Threlkeld, J.K.Szymanik}@uva.nl

[1] School of Philosophy, Psychology and Language Sciences; University of Edinburgh
[2] Institute for Logic, Language and Computation; Universiteit van Amsterdam
[*] Co-first authors.

Abstract

Natural languages exhibit many *semantic universals*: properties of meaning shared across all languages. In this paper, we develop an explanation of one very prominent semantic universal: that all simple determiners denote monotone quantifiers. While existing work has shown that monotone quantifiers are easier to learn, we provide a complete explanation by considering the emergence of quantifiers from the perspective of cultural evolution. In particular, in an iterated learning paradigm, with neural networks as agents, monotone quantifiers regularly evolve.

Keywords: semantic universals; generalized quantifiers; monotonicity; iterated learning; neural networks

Introduction

While natural languages show great variability, there are features that they all appear to share. Linguists call these features linguistic *universals*. Universals have been found at all levels of linguistic structure: phonological, syntactic, and semantic.[1] Some universals might follow from constraints on what humans are physically capable of doing. For instance, there is no language whose prosody requires the production of ultrasounds. The reasons for other universals are harder to understand, leading to multiple proposed explanations.

One well-supported claim is that at least some universals are to be explained in terms of *learnability*.[2] More precisely, it is easier to learn a language that satisfies the universal than it is to learn a language that does not satisfy the universal, and this difference in the complexity of acquisition causes languages that satisfy universals to spread. In the case of universals of lexical semantics such as the one we focus on below, the learnability explanation says that lexical entries whose meaning satisfies the universal are easier to learn, and therefore more likely to be lexicalized. Complicated meanings can be obtained through complex grammatical constructions and compositional interpretation thereof.

The learnability explanation is an empirical, causal claim about the origins of linguistic universals. One way to support the learnability explanation for a specific universal is to provide a model of learning that is cognitively realistic and on which expressions that satisfy the universal are indeed easier to learn.

Finding an appropriate model of learning can however only partially explain a linguistic universal. Learnability is a fact about individual cognition, while a universal is a feature of a whole language. A second challenge consists in connecting these two levels, showing the effects of learnability on emerging language structure. This is the so-called problem of *linkage*.[3]

Iterated learning (IL) is a method that addresses the problem of linkage. In IL, parents teach children their language, who teach the next generation their language, and so on and so forth. The crucial insight of IL is that learning is not an inert process in cultural evolution, since the languages of a cultural child and its cultural parent are generally slightly different. The changes caused by learning are not random, but rather tend to be guided by the learner's cognitive biases. As a consequence, over time languages adapt better and better to the agents' cognitive biases. Learnability can then affect the frequency of different traits.[4]

Previous work has addressed the learnability challenge by showing that quantifiers, responsive predicates, and color terms that satisfy certain semantic universals are easier to learn for neural networks.[5] In this paper, we address the problem of linkage by building an iterated learning model of the evolution of the semantic structure of quantifiers. In particular, we will use neural networks as our agents and standard gradient descent as the learning method inside the context of iterated learning. The next section briefly reviews the theory of generalized quantification and the universal of *monotonicity*. After that, the following section presents the model of cognition and the iterated learning model, as well as an information-theoretic measure of the *degree of monotonicity* of a quantifier. Experiments with this model and their results are presented in the following section. Results are discussed in the final section, along with possible future directions.

[1] For some examples see, respectively, Hyman (2008), Newmeyer (2008), and Barwise and Cooper (1981).

[2] See, e.g., Steinert-Threlkeld and Szymanik (in press), Piantadosi, Tenenbaum, and Goodman (2013), and Peters and Westerståhl (2006).

[3] The problem of linkage was introduced in Kirby (1999).

[4] See, e.g., Tamariz and Kirby (2016); Culbertson and Kirby (2016); Kirby, Cornish, and Smith (2008) for discussions of the way individual cognition is reflected in language structure through IL and experimental evidence supporting the connection.

[5] See, respectively, Steinert-Threlkeld and Szymanik (in press); Steinert-Threlkeld (in press); Steinert-Threlkeld and Szymanik (2019).

Quantifiers and monotonicity

Determiners are expressions that take a common noun as an argument and return a Noun Phrase. Determiners can be grammatically simple—e.g. *some*, *few*, *most*—or complex—e.g. *fewer than three* or *at most five*.[6] Determiners express generalized quantifiers.[7] (Monadic) Generalized quantifiers are properties of sets of subsets of a domain of discourse. The generalized quantifiers expressed by natural language determiners are of type $\langle 1,1\rangle$, i.e. properties of exactly two sets. Equivalently, a quantifier of type $\langle 1,1\rangle$ takes (the characteristic function of) a set A and returns a function from (the characteristic function of) a set B to truth values. A is the *left argument* and B the *right argument* of the quantifier. For instance, the sentence "most As are B" is true if and only if the number of As that are B (cardinality of the intersection of A and B, i.e., $|A \cap B|$) is greater than the number of As that are not Bs (i.e., $|A - B|$), i.e.:

$$[\![most]\!] = \{(A,B) : |A \cap B| > |A \setminus B|\}$$

Various universals have been proposed about which generalized quantifiers are expressed by simple determiners. In the following, we focus on the *monotonicity* universal proposed by Barwise and Cooper (1981). This says that all simple determiners across all languages express monotone quantifiers. A quantifier is monotone iff it is *upward* monotone or *downward* monotone. A quantifier Q is upward monotone [downward monotone] iff for any three sets A, B and B', if $Q(A)(B)$ and $B \subseteq B'$ [$B' \subseteq B$] then $Q(A)(B')$. As an example, consider the upward monotone quantifier $[\![most]\!]$. Assume that the sentence "Most cats sleep" is true and that everything that sleeps is alive, i.e. $[\![sleep]\!] \subseteq [\![alive]\!]$. The monotonicity of $[\![most]\!]$ ensures then that "Most cats are alive" is true.

Monotonicity is an interesting universal because it is easy to imagine non-monotone quantifiers. Examples of non-monotone quantifiers abound among the meanings of complex determiners: "an even/odd number of" or "exactly 2", etc. The commonness of non-monotonicity among complex quantifiers makes the lack of simple non-monotone quantifiers especially puzzling and in need of an explanation. Previous work proposed to explain the universal of monotonicity in terms of the greater learnability of monotone quantifiers.

Steinert-Threlkeld and Szymanik (in press) propose to use neural networks in this context. A neural network is a computational device that can learn to approximate functions by observing tuples of inputs and relevant outputs, and progressively minimizing a suitably defined distance between the true output and the network's own prediction. In the case of a quantifier, the input is a structure where the relevant sets are specified and the output is 1 iff the structure verifies the quantifier and 0 otherwise. In practice, given a structure the neural network outputs a probability that can be interpreted as confidence that the structure verifies the quantifier.

[6] Exactly how to draw the distinction between simple and complex and whether, for instance, *most* is simple or complex, do not matter for present purposes.

[7] For more information on generalized quantifier theory from linguistic, computational, and cognitive perspectives, see also Peters and Westerståhl (2006) and Szymanik (2016).

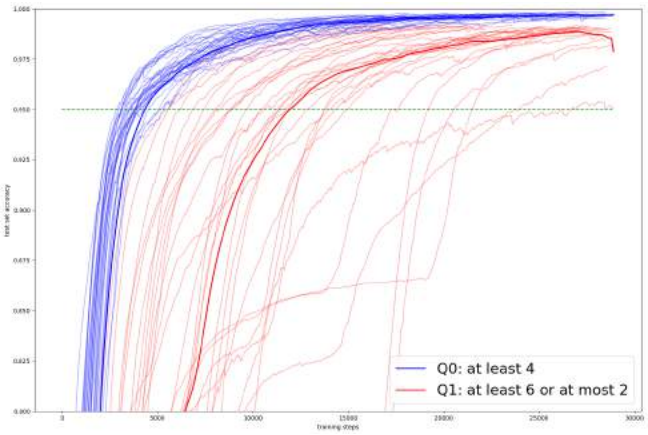


Figure 1: Learning curves on a neural network for the monotone *at least 4* (blue) versus *at least 6 or at most 2* (red). The x-axis is number of training steps; the y-axis is accuracy (percentage correct) on a test set of examples the network has not yet seen. This was Figure 4 in Steinert-Threlkeld and Szymanik (in press).

Data about how fast neural networks learn different kinds of quantifiers was produced with the following algorithms. First, two quantifiers are picked such that one satisfies the universal and the other does not. Then, the two quantifiers are taught to a neural network until it has accurately learned them. The crucial information is how long on average it takes neural networks to accurately learn quantifiers that satisfy the universal compared to ones that do not. Various universals were tested in this way. In the case of monotonicity, the data was produced both for a downward monotone and for an upward monotone quantifier. The neural networks were strikingly faster at learning monotone compared to non-monotone quantifiers. Figure 1 shows an example.

As discussed above, knowing that meanings with certain features can be learned more easily only goes some of the way in explaining the features' universality across various languages. A full explanation also needs to show that the structure can and eventually will be reached by processes of cultural evolution. In the rest of this paper, we develop an iterated learning model of the cultural evolution of quantifiers that embeds the learning model of neural networks, and show that monotonicity reliably emerges.

Methods

Iterated learning

IL models start with two groups of randomly initialized agents, the first and second generations. Each agent in the first generation—the *cultural parent*—is associated with one agent in the second generation—the *cultural child*. A set of

linguistic production data is generated for each cultural parent and used as input for the cultural child. The cultural child tries to approximate its cultural parent's language. In the following step, the process is repeated with agents in the second generation as cultural parents and the new agents in a third generation as cultural children. The cultural transmission process is iterated for some number of generations. Each cultural family line is called a *chain* of IL.

Crucially, the agents do not learn their parent's language perfectly. There can be various reasons for this. First, there can be a bottleneck in learning. This happens when the learner does not observe everything that is needed to perfectly reconstruct the language, and therefore has to guess some aspects of it. The number of data points given to the learners is fixed for all generations and agents and is called the *bottleneck size*. A second reason is that the agent might not have perfect memory or perfect reasoning abilities, and might therefore learn a language that does not perfectly conform to the given data. In this case, the more rational the agent, the closer the learned language will be to the teacher's language. A third reason is that the cultural parents might produce language in a way that is stochastic rather than deterministic. This can make the language harder to approximate and impossible to learn perfectly. For instance, a cultural parent might pick among the signals compatible with a certain observation according to a categorical distribution. The cultural child would need to infer the parameters of the distribution, a task which cannot in general be accomplished perfectly with a finite number of observations.

The changes introduced by each learner accumulate over generations. Since these changes are not completely random, but rather tend to be consistent across agents, the languages tend to change in the same way over time in different chains. In sum, IL is a way to study how the cognitive system of the learners determine which languages one should expect to see spoken in a population of such agents. The crucial individual level components of an IL model are the set of possible languages, and the way the agents learn them. We now explain these two components in turn.

Model of models, quantifiers, and language

Since the focus is on the evolution of monotonicity, we simplify the language model by assuming that the quantifiers are conservative and extensional.[8] This amounts to saying that the truth value of each quantifier only depends on the elements in A and $A \cap B$, and not on $\overline{A \cup B}$ or $B \setminus A$. Therefore, the truth of any quantifier depends only on which of the elements of A are also elements of B, and which are not. Assuming conservativity and extensionality both reduces the number of possible quantifiers that agents can speak and simplifies the model of each quantifier, since only A and $A \cap B$ need to be encoded. Moreover, we assume that the left argument of the quantifiers is fixed to some set A with cardinality n.

Assuming conservativity/extensionality and a fixed set A, we can represent the part of the world—called a *model*—that is relevant to determining the truth value of a quantifier as a bit vector of a fixed length n. Each element of the model represents an object in A. Each element has value 1 iff the object corresponding to that bit is also an element of B, and 0 otherwise. For instance, the vector $[0,1,1]$ would model a situation where $A = \{o_1, o_2, o_3\}$ and o_2, $o_3 \in B$. The set of models is the set of all binary strings of length n, i.e. the set of possible relations between a fixed A and any possible B. We call M' a *submodel* of a model M iff M' is 0 everywhere where M is 0. For instance, $[0,1,1,0,0]$ is a submodel of $[0,1,1,1,1]$. Note that each model is a submodel of itself.

We represent a *quantifier* as a function from models to single bits. An example of a quantifier is $Q(x) = 1$ if $\sum_{i=1}^{n} x_i > 2$ otherwise 0, meaning "more than two". Since for A of size n there are 2^n different models, each quantifier is a 2^n-sized bit vector. Each element of the quantifier vector corresponds to a model and has value 1 iff the model verifies the quantifier and 0 otherwise.

To see how this works in practice, consider a set A of size 3. There are 8 possible ways in which any other set B can overlap with A. Each of these is modelled as a bit vector of size 3. For instance, $[0,1,1]$ says that the second and third object of A are also elements of B, but the first is not. The English expression "all As are B" is modelled by a bit vector of size 8 that has value 1 at the index corresponding to the model $[1,1,1]$ and 0 otherwise. If the models are ordered lexicographically[9] and the last model is therefore $[1,1,1]$, then the quantifier corresponds to the vector $[0,0,0,0,0,0,0,1]$. We call a quantifier *degenerate* if and only if it corresponds to a vector of identical elements, 0s or 1s. A degenerate quantifier corresponds intuitively to a quantifier that is true or false of every model.

Each agent encodes a single quantifier. Agents do not encode the quantifiers directly. Rather, given a model they produce a truth value by using a neural network. The next two sections clarify the connection between the neural networks and the agent's behaviour.

Neural Networks

Because of the aforementioned learnability results of Steinert-Threlkeld and Szymanik (in press), the agents that make up the generations in our iterated learning setup are *neural networks*. Each network has n input neurons (one for each bit of a vector corresponding to a model) and one output neuron (how probable it thinks that the true output is a 1), with two hidden layers of 16 neurons each. We made this

[8] These, next to monotonicity, are two prominent semantic universals distinguishing natural language quantifiers from all logically possible quantifiers. Extensionality means that extending or shrinking the universe of discourse has no effect on the truth-value of the quantifier sentence as long as the left and right arguments are unchanged. Conservativity means that only the part of B that is common to A matters for the truth-value of the sentences. In other words, the elements in $B \setminus A$ can be safely ignored when determining the truth-value. See Peters and Westerståhl (2006) for definitions.

[9] In that case, lexicographic order is the dictionary order over sequences of letters from the alphabet $\{0,1)$ with 0 preceding 1 in the order.

choice so that the networks had enough expressive power to represent many quantifiers, including complex ones. Future work will analyze the effect of architecture choices on the results presented below. The networks and learning, which will be described in the next section, were implemented in PyTorch.[10]

Such a network learns from input/output pairs using a fancier version of gradient descent called Adam (Kingma & Ba, 2015). The network receives a number of true input/output pairs, which it iterates over in small batches. For each batch, it guesses the correct outputs for the inputs, and then updates its parameters (weights and biases connecting the neurons) in such a way that its future outputs are guaranteed to be closer to the truth.[11] Because this style of learning is fairly gradual, we introduce one more parameter to our simulations, namely *number of epochs*: this is how many times the network processes its training set in each generation. In other words, the network sees a portion of its parent's language, as determined by *bottleneck size*, but gets to learn from that portion number-of-epochs times.[12]

Model of the agents

Each agent plays two roles in an IL simulation. The first role is to learn a language given data from the previous generation. The second role is to produce data used to teach to the following generation. To produce this data, the agent is prompted with randomly chosen models.

In the learning phase, each agent receives learning data consisting of a set of tuples ⟨model, judgment⟩. The judgment is a single bit expressing whether the quantifier used by the agent is compatible or not with the model. This data is used to train the agent's neural network as described in the previous subsection.

Production works as follows. The agent feeds an observed model to its neural network. The neural network returns a number in the $[0,1]$ interval. Then, the agent rounds the number and returns it. The returned number expresses whether the agent's quantifier is compatible or not with the model that the agent observed. The production behaviour is deterministic, since an agent always produces the same bit given the same model.

Prompted with a string of 1s and 0s, agents produce a 1 or 0. The former models a state of the world, the latter models the compatibility of the agent's quantifier with the world state. However, nothing in the simulation implies that neural networks are interpreting 1 and 0 as True and False respectively in their input and output. Therefore, the output of an agent under-determines which quantifier the agent speaks, even when the output for all models is known. For instance, an agent that returns 1 for input $[0,0,1,1]$ can be interpreted as accepting the model where $B = \{o_3, o_4\}$ (if 1 is interpreted as True in the model and in the quantifier), as rejecting the model where $B = \{o_3, o_4\}$ (if 1 is interpreted as False in the quantifier and True in the models), as accepting the model where $B = \{o_1, o_2\}$ (if 1 is interpreted as True in the quantifier and False in the models), or as rejecting the model where $B = \{o_1, o_2\}$ (if 1 is interpreted as False in the quantifier and the models). Crucially, the interpretation of the bits has to be consistent across the models and across the quantifier judgments. Therefore, each agent can be interpreted as speaking four quantifiers, depending on whether 1 and 0 are interpreted as meaning true or false in the models and in the agent's output. We discuss below how we deal with underdeterminacy when it might make a difference to the interpretation of the results.

Measures of monotonicity

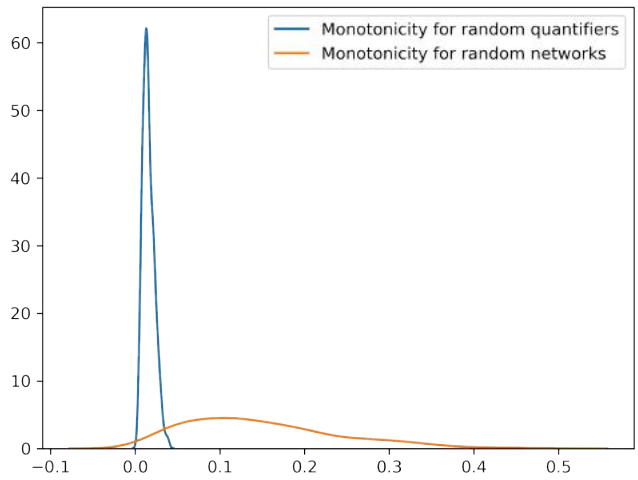


Figure 2: Kernel Density Estimation of the distribution of degrees of monotonicity from a sample of 300 completely random quantifiers and 300 random neural network agents. The x-axis is the measure of monotonicity we describe in the main text.

According to the standard definition, monotonicity is a binary property. A possible way of analyzing the results would be to find the proportion of monotone languages in every generation. However, some quantifiers are intuitively more monotone than other quantifiers. For instance, consider the three quantifiers "some", "between 3 and 5" and "an even number of". While "some" is monotone and the other two quantifiers are not, intuitively "an even number of" is the least monotone of the three. To track finer changes in monotonicity level over time, we define a graded measure of monotonicity.

We measure monotonicity in information-theoretic terms as the proportion of uncertainty in the output of a quantifier that is removed after knowing that there is a submodel where the quantifier is true (i.e. a 1). For a perfectly (upward) monotone quantifier Q, if a model M has a submodel to which the quantifier assigns 1 then Q will assign 1 to M. Therefore, for

[10] `http://pytorch.org`

[11] For general introductions, see Nielsen (2015); Goodfellow, Bengio, and Courville (2016).

[12] In some experimental literature — for example, Carr, Smith, Culbertson, and Kirby (2019) — this is also referred to as *exposures*.

a monotone quantifier all the uncertainty is removed and the measure has value 1.

More formally, first define the random variables $\mathbb{1}_Q$ and $\mathbb{1}_Q^{\prec}$ on the space of possible models as follows. $\mathbb{1}_Q$ is the value that Q assigns to the model M. $\mathbb{1}_Q^{\prec}$ is whether a model has a submodel that the quantifier considers true (assigns 1 to). The entropy of $\mathbb{1}_Q$, $H(\mathbb{1}_Q)$, quantifies the uncertainty about what truth value Q will assign to a model. The conditional entropy $H(\mathbb{1}_Q \mid \mathbb{1}_Q^{\prec})$ quantifies the uncertainty about what Q will assign to a model, given that one knows whether the model has a submodel that Q considers true (assigns 1 to). $H(\mathbb{1}_Q \mid \mathbb{1}_Q^{\prec})$ is minimized (attains value 0) for a perfectly monotone quantifier: if you know that a model has a true submodel, and the quantifier is upward monotone, you know the truth value of that model. The difference between the entropy and the conditional entropy between these variables is known as the mutual information:

$$I(\mathbb{1}_Q;\mathbb{1}_Q^{\prec}) := H(\mathbb{1}_Q) - H(\mathbb{1}_Q|\mathbb{1}_Q^{\prec})$$

This measures how much information $\mathbb{1}_Q^{\prec}$ provides about $\mathbb{1}_Q$. For a perfectly monotone quantifier, $H(\mathbb{1}_Q|\mathbb{1}_Q^{\prec}) = 0$, and so $I(\mathbb{1}_Q;\mathbb{1}_Q^{\prec}) = H(\mathbb{1}_Q)$. In other words: for a monotone quantifier, knowing which models have a true sub-model provides as much information as knowing the entire quantifier.

While this roughly captures what we want from a measure of monotonicity, it needs to be normalized to form a degree that applies across quantifiers, since $0 \leq I(\mathbb{1}_Q;\mathbb{1}_Q^{\prec}) \leq H(\mathbb{1}_Q)$. We do this by dividing by $H(\mathbb{1}_Q)$, moving the upper bound to 1. In total then, we measure monotonicity as

$$\begin{aligned}\mathsf{mon}(Q) &:= \frac{I(\mathbb{1}_Q;\mathbb{1}_Q^{\prec})}{H(\mathbb{1}_Q)} \\ &= \frac{H(\mathbb{1}_Q) - H(\mathbb{1}_Q|\mathbb{1}_Q^{\prec})}{H(\mathbb{1}_Q)} \\ &= 1 - \frac{H(\mathbb{1}_Q \mid \mathbb{1}_Q^{\prec})}{H(\mathbb{1}_Q)}\end{aligned}$$

To see how this measure tracks intuitions, consider the previous mentioned quantifiers "some", "between 3 and 5" and "an even number of". "Some" gets monotonicity 1.0, because knowing whether a model has a submodel that verifies "some" eliminates all uncertainty about the truth of the model. Recall that each agent can be interpreted as instantiating any of four quantifiers, which can be monotone to different degrees. This raises the question of which of the four degrees of monotonicity should be considered in the analysis of the results. The monotonicity of an agent's language is the highest among the degrees of the quantifiers compatible with the agent's language. For instance, an agent whose quantifier is "between 3 and 5" has degree 0.7517 and one with "an even number of" has degree 0.001.

We compare the results of the simulation to the distribution of the measure in randomly generated quantifiers. There are two different random distributions of quantifiers. On the one hand, there are the quantifiers instantiated by randomly initialized agents. On the other hand, there are the quantifiers sampled uniformly from the space of possible quantifiers. These two distributions are depicted in Figure 2. While the completely random quantifiers have a narrower distribution, both types of random distribution are very skewed towards low degree of monotonicity. This makes sense: monotonicity is a relatively rare property, and so should not be expected to appear randomly. We now turn to the results, showing that higher degrees do emerge via iterated learning.

Materials

For our experiments, we used a fixed model size of 10 (which, recall, is also the size of the input to the agents), with 10 agents in each generation, and varied the bottleneck size (200, 512, 715, 1024) and number of epochs (4 and 8). For each setting of those two parameters, we ran 20 trials.

The code, data, and instructions for running experiments may be found at `https://github.com/thelogicalgrammar/NeuralNetIteratedQuantifiers`.

Results

The first result is that monotone quantifiers evolve consistently and rapidly for some values of the simulation parameters. More specifically, the evolution of monotonicity depends on the bottleneck size and the number of epochs, i.e. how much of the parent's language is observed by the cultural child. See Figure 3 for the results. If the networks get too much input, they learn the quantifier accurately and change is very slow. If the networks get too little input, the learning has little effect and no pattern emerges. If languages are somewhat stable across generations, but enough variation is allowed by not over-training the cultural children, monotonicity evolves.

A second result is that the monotone quantifiers that emerge are in large part non degenerate. In Bayesian models that include a prior for simplicity, degenerate languages become widespread under pure IL (Kirby, Tamariz, Cornish, & Smith, 2015). Here, however, degenerate quantifiers are a small minority (about 0.005% of all quantifiers).

The third result is that most non-degenerate monotone quantifiers fall in one of a few types. About 79% of the perfectly monotone quantifiers show the following pattern: there is some index i such that the quantifier—call it Q_i—assigns 1 to a model iff the model is 1 at i (or an equivalent pattern obtained by switching 0 and 1 uniformly in the models and/or in the quantifier). Q_i is true iff o_i, the object represented by index i, belongs to the set B.[13]. Therefore $Q_i(A)$ functions like a proper noun for o_i. Just like "Anna is human" is true iff Anna belongs to the set of humans, "$Q_i(A)$ is B" is true iff o_i belongs to the set B.

[13] In set-theoretic terms, Q_i is a *principal ultrafilter* If U is a finite non-empty set, a set F is a principal ultrafilter on U if there is an $a \in U$ such that $F = \{B \in \mathcal{P}(U) | a \in B\}$. In the present model, Q_i is (the characteristic function of) a principal ultrafilter on B because it it contains every subset of B that contains i.

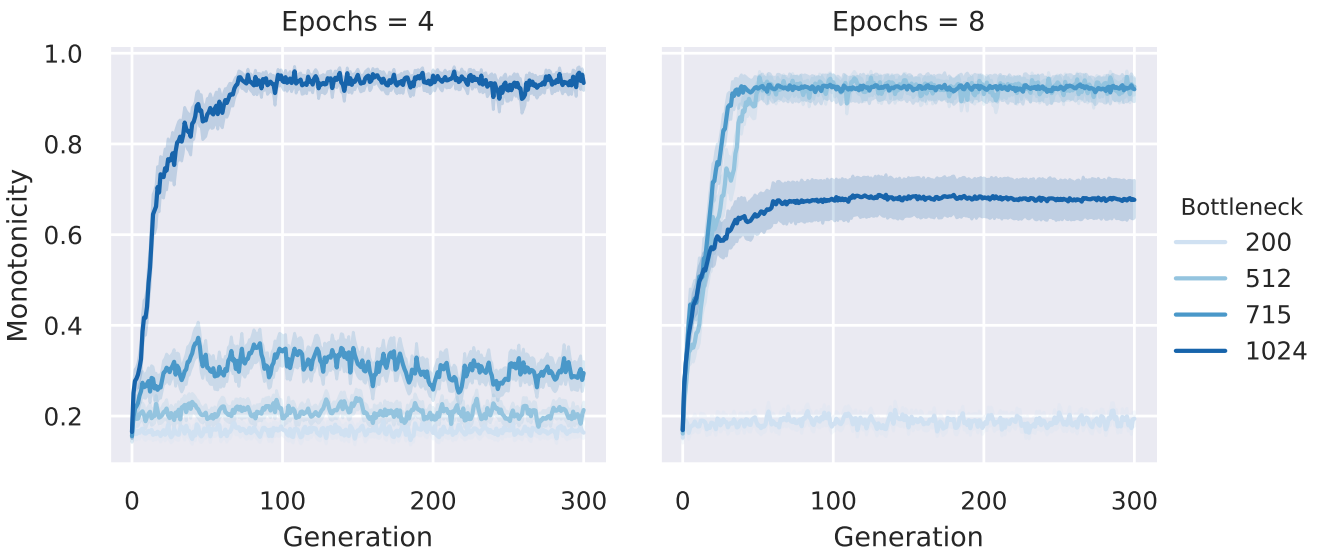


Figure 3: The simulation was ran 20 times for each combination of bottleneck size and number of epochs in a population of 10 agents and a maximum model size of 10. The plot shows how the average monotonicity level across all languages changes over 300 generations. Convergence to monotonicity depends on how much the learners' neural networks are trained, which itself depends on the number of epochs and the bottleneck size. With small bottleneck and few epochs, monotonicity does not evolve. With a bigger bottleneck size and more training epochs, monotone languages become widespread. However, increasing the training data further tends to impede the development of monotone languages.

For other monotone quantifiers $Q_{\{j,k\}}$, there are two indices j,k (with $j \neq k$) such that $Q_{\{j,k\}}$ assigns 1 to a model iff the model has value 1 at both j and k (or, again, an equivalent patterns obtained by switching 0 and 1 in the models and/or in the quantifier). $Q_{\{j,k\}}$ is true iff B contains two specific elements of A, and false otherwise.[14] It can be interpreted as the conjunction of two proper nouns. Like "Anna and Rob are human" is true iff Anna is human and Rob is human, "$Q_{\{j,k\}}(A)$ is B" is true iff o_j is B and o_k is B.

Discussion

The results we presented support the learnability account of the origins of semantic universals of quantification. While previous work compared quantifiers satisfying semantic universals to quantifiers that do not, we have presented a model where the former are selected out of all of the possible quantifiers by a process of cultural evolution. Moreover, the preference for monotone quantifiers is not a consequence of an explicitly coded bias for simplicity, but rather of an independently motivated, biologically plausible model of learning. The results therefore suggest that not only are monotone quantifiers easier to learn, but they are also widespread in language *because* of their learnability.

This model can be straightforwardly extended in various ways. The agents judged their quantifier compatible with a given model simply by rounding the output of their neural network. An alternative to this is for the agents to accept a model with a probability proportional to the network's output. Such so-called sample agents do not straightforwardly instantiate a quantifier, since they can produce inconsistent output when repeatedly prompted with the same model. However, preliminary results have shown that neural networks are capable of doing *statistical learning*: given enough data, they approximate not just whether their parents tend to reject or accept a model, but also the probability of acceptance.

While the quantifiers that emerge from our experiment are monotone, they are unnatural in certain respects. For instance, the proper-name-like quantifiers that emerge are not *quantitative*, i.e. their truth value depends not simply on the number of 1s and 0s, but on the identity of particular elements.[15]

To try and explain the emergence of quantifiers which are both monotone and quantitative, it might be necessary to make it more difficult for the networks to rely on the identity of particular objects by, for instance, shuffling the order of models in the parent and the teacher's inputs. Another pressure that might contribute to shape the meaning of quantifiers comes from communication (Kirby et al., 2015). While

[14] These are called in set-theoretic terms *principal filters*. They are not principal ultrafilters because their truth depends on more than one element.

[15] See Steinert-Threlkeld and Szymanik (in press) for the definition of and motivation for quantity, which generalizes the isomorphism/permutation constraint in generalized quantifier theory as discussed, for instance, in Peters and Westerståhl (2006).

some semantic universals of quantification might have an advantage in cultural evolution because they conform well with learning biases, other universals might evolve because they lead to more successful communication. Therefore, combining iterated learning with a pressure for accurate communication might help more natural quantifiers emerge. We leave all of these exciting possibilities to future work.

Acknowledgments

Shane Steinert-Threlkeld and Jakub Szymanik have received funding from the European Research Council under the European Unions Seventh Framework Programme (FP/20072013)/ERC Grant Agreement n. STG 716230 CoSaQ.

References

Barwise, J., & Cooper, R. (1981). Generalized Quantifiers and Natural Language. *Linguistics and Philosophy*, *4*(2), 159–219.

Carr, J. W., Smith, K., Culbertson, J., & Kirby, S. (2019). Simplicity and informativeness in semantic category systems. Retrieved from https://psyarxiv.com/jkfyx

Culbertson, J., & Kirby, S. (2016). Simplicity and Specificity in Language: Domain-General Biases Have Domain-Specific Effects. *Frontiers in Psychology*, *6*. doi: 10.3389/fpsyg.2015.01964

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. The MIT Press. Retrieved from https://www.deeplearningbook.org/

Hyman, L. M. (2008). Universals in phonology. *The Linguistic Review*, *25*(1-2), 83–137.

Kingma, D. P., & Ba, J. (2015). Adam: A Method for Stochastic Optimization. In *International Conference of Learning Representations (ICLR)*. Retrieved from https://arxiv.org/abs/1412.6980

Kirby, S. (1999). *Function, Selection, and Innateness: The Emergence of Language Universals*. Oxford; New York: OUP Oxford.

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, *105*(31), 10681–10686.

Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, *141*, 87 - 102. doi: https://doi.org/10.1016/j.cognition.2015.03.016

Newmeyer, F. J. (2008). Universals in syntax. *The Linguistic Review*, *25*(1-2), 35–82. doi: 10.1515/TLIR.2008.002

Nielsen, M. A. (2015). *Neural Networks and Deep Learning*. Determination Press. Retrieved from http://neuralnetworksanddeeplearning.com/

Peters, S., & Westerståhl, D. (2006). *Quantifiers in Language and Logic*. Oxford: Clarendon Press.

Piantadosi, S. T., Tenenbaum, J. B., & Goodman, N. D. (2013). Modeling the acquisition of quantifier semantics: a case study in function word learnability. Retrieved from http://colala.berkeley.edu/papers/piantadosi2012modeling.pdf

Steinert-Threlkeld, S. (in press). An Explanation of the Veridical Uniformity Universal. *Journal of Semantics*. Retrieved from https://semanticsarchive.net/Archive/DI5ZTNmN/UniversalResponsiveVerbs.pdf

Steinert-Threlkeld, S., & Szymanik, J. (2019). *Ease of Learning Explains Semantic Universals.* Retrieved from https://semanticsarchive.net/Archive/zM5ZGIxM/EaseLearning.pdf

Steinert-Threlkeld, S., & Szymanik, J. (in press). Learnability and Semantic Universals. *Semantics & Pragmatics*. Retrieved from http://semanticsarchive.net/Archive/mQ2Y2Y2Z/LearnabilitySemanticUniversals.pdf

Szymanik, J. (2016). *Quantifiers and Cognition. Logical and Computational Perspectives*. Springer.

Tamariz, M., & Kirby, S. (2016). The cultural evolution of language. *Current Opinion in Psychology*, *8*, 37-43. doi: 10.1016/j.copsyc.2015.09.003

"Natural concepts" revisited in the spatial-topological domain: Universal tendencies in focal spatial relations

Alexandra Carstensen[1], **George Kachergis**[1], **Noah Hermalin**[2], **Terry Regier**[2,3]
[1]Department of Psychology, Stanford University, Stanford, CA
[2]Department of Linguistics, University of California at Berkeley, Berkeley, CA
[3]Cognitive Science Program, University of California at Berkeley, Berkeley, CA

Abstract

It has long been noted that the best examples, or foci, of color categories tend to align across diverse languages (Berlin & Kay, 1969)—but there is limited documentation of such universal foci in other semantic domains. Here, we explore whether spatial topological categories, such as "in" and "on" in English, have focal members comparable to those in color. We document names and best examples of topological spatial relations in Dutch, English, French, Japanese, Korean, Mandarin Chinese, and Spanish, and find substantial consensus, both within and across languages, on the best examples of such spatial categories. Our results provide empirical evidence for focal best examples in the spatial domain and contribute further support for a theory of "natural concepts" in this domain.

Keywords: Language and thought; spatial cognition; categories; semantic universals.

The central role of foci

For decades, discussions of natural language categories such as "dog" or "blue" have emphasized prototypes, family resemblance, and fuzzy sets—all notions specifying relations between central cases and boundaries, and recognizing gradation in category membership. An especially well-studied and debated case is that of focal colors, or best examples of color categories (e.g. Berlin & Kay, 1969; Heider, 1972; Kay & McDaniel, 1978; Roberson et al., 2000; Regier et al., 2005; Abbott et al., 2016). Despite the ongoing debate, there is broad consensus that such best examples of color categories often (but not always) align across languages, and that languages sometimes have composite categories apparently organized around multiple foci—for example a composite green-blue or "grue" category.

Despite the attention given to focal colors, studies of categorization and semantic typology in many other semantic domains have not emphasized category best examples as prominently, but have instead tended to characterize categories as sets, such that an exemplar may simply be a member of the category or not. Within the domain of spatial topological relations, previous work has drawn on extensional patterns in naming as evidence for central exemplars and core meanings of categories like "in" and "on" (e.g., Levinson et al., 2003; Johannes, Wang, Papafragou, & Landau, 2015; Johannes, Wilson, & Landau, 2016; Landau, Johannes, Skordos, & Papafragou, 2017), but without directly querying speakers about best examples per se. Here, we employ empirical best example data to provide a long-overdue response to a call by Feist (2000: 236) to determine whether spatial relational categories, like colors, have focal members.

In what follows, we review key findings on focal colors and their relationship to color category semantics. We then describe parallels to color in the domain of spatial topological relations, and summarize an account (Levinson et al., 2003) of focal spatial relations that was developed and evaluated on the basis of spatial naming data, but without grounding in empirical best examples. We then present our study, which reexamines the hypotheses of this previous account using empirical best example data from seven languages. We explore three related questions about focal category members in the spatial domain:

1. Is there consensus within languages on focal spatial relations?
2. Is there consensus across languages on focal spatial relations?
3. Do spatial categories exhibit composite structure, with more than one focus per category?

To preview our results, we find initial evidence for universal tendencies in focal spatial relations, both within and across languages, based on naming and best example data from seven languages. We also find evidence for at least three composite spatial categories, where a single lexical category includes multiple foci. We conclude that focal spatial relations share some of the distinctive features of foci in the color domain.

Focal colors

Berlin and Kay (1969) proposed two key features of focal colors that we consider in the spatial domain: (1) a set of universal focal colors (red, green, yellow, blue, white, and black), and (2) an evolutionary sequence of color categories, by which languages follow a common hierarchy to successively partition color space, progressively subdividing the focal colors into categories. Kay and McDaniel (1978) elaborated this proposal, specifying multi-foci *composite categories* as shown in Figure 1.[1] By this model, the initial two-term category system represented as the first split in the diagram will group WHITE, RED, ORANGE, and YELLOW into a single "warm" category. Kay and McDaniel argued that

[1]Kay and McDaniel's (1978) proposal included two closely-related hierarchies, only one of which is shown here for illustration.

large categories like this in the early stages of the hierarchy are composite, and may be focused at any of their constituent foci. Accordingly, this "warm" category could be focused at WHITE, YELLOW, or RED but not ORANGE, as it is not one of the proposed universal color foci. Similarly, "grue" terms composed of GREEN and BLUE (the latter inclusive of PURPLE) could be focused at either of the two constituent foci, GREEN or BLUE.

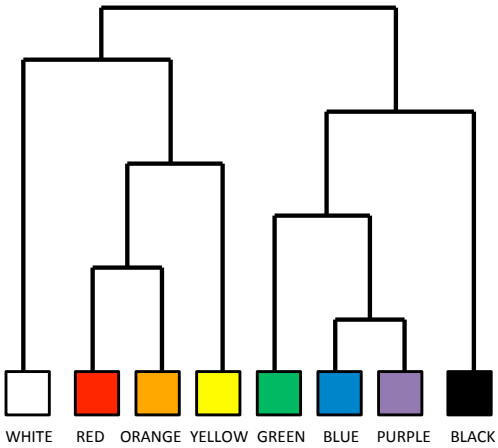


Figure 1: Kay and McDaniel's (1978) proposed evolutionary hierarchy of color terms.

Focal spatial relations?

In our analysis of spatial category best examples, we explore analogs to two distinctive focal color phenomena: cross-language agreement on specific focal colors, and the composite nature of categories spanning multiple foci. To do so, we draw on a proposal for spatial topological concepts by Levinson and colleagues (2003) that parallels much of Kay and McDaniel's (1978) characterization of color. Levinson et al. (2003) proposed an implicational hierarchy of spatial "natural concepts" (or notional clusters of related meanings) modeled on Kay and McDaniel's (1978) color hierarchy and based on a study of spatial semantics in a set of nine diverse languages. In their proposal, Levinson et al. suggest that spatial topological categories, as in color, tend to undergo successive subdivisions in which distinct focal senses of composite categories "split into primary (single-focus) categories over time" (Levinson et al., 2003: 512), as shown in Figure 2.[2]

The present study

To our knowledge, ours is the first study to document empirical best examples in the spatial topological domain. We ask whether speakers of seven languages (1) agree on best examples for common spatial terms in their language, (2) agree on

[2]We interpret Levinson et al.'s (2003) proposal to include two related hierarchies, one of which is shown here for illustration, and both of which are specified in Carstensen and Regier (2013).

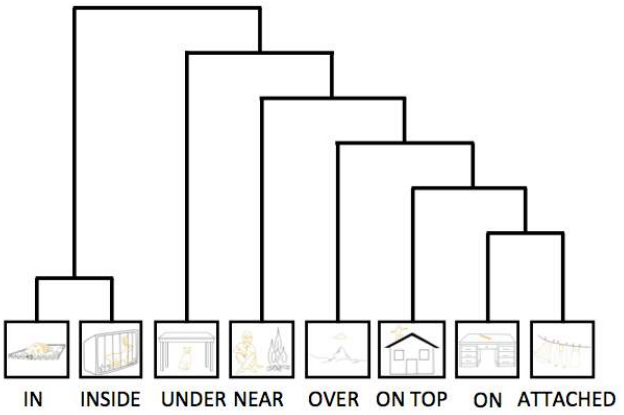


Figure 2: Levinson et al.'s (2003) proposed evolutionary hierarchy of topological spatial concepts, specifying a predicted order in which spatial categories in language will tend to subdivide distinct spatial notions as new terms emerge in the lexicon.

focal best examples *across* languages, and (3) demonstrate composite categories subject to successive differentiation of focal notions in keeping with Levinson et al.'s hypothesized spatial category hierarchy. If so, this finding would provide empirical evidence for focal best examples in the spatial domain that share key aspects with color foci, and contribute further support for Levinson et al.'s suggested "natural concepts."

Methods

In order to investigate whether spatial relations have focal members within and across languages, native speakers of seven languages (a convenience sample: English, Dutch, Spanish, French, Mandarin Chinese, Japanese, and Korean) were asked to name the spatial relation depicted in each of a set of cards, and then asked to select the best example, good examples, and all possible examples of the spatial terms they provided.

Participants The study included native speakers of 7 languages: 24 English, 29 Spanish, 18 French, 19 Japanese, 13 Dutch, 18 Korean, and 18 Mandarin Chinese speakers. All participants were native speakers of their respective language, and tasks were administered in that language by experimenters who were also native speakers.

Stimuli Stimuli were the 71 spatial scenes of the Topological Relations Picture Series (TRPS) by Bowerman and Pederson (1992). Scenes are line drawings showing an orange figure object located relative to a black ground object (e.g., a cup on a table; see Figure 2).

Procedure

Instructions and object labels for each of the TRPS scenes (e.g. cup, table) were translated from English to the study language and then backtranslated to ensure accuracy.

1. Scene naming. Participants were shown each of the spatial scenes in one of two fixed random orders, and asked to name the spatial relation in each. Each scene was shown above a fill-in-the-blank in the participant's native language with labels specifying the figure and ground objects, and the participant filled in the blank to complete a normal, everyday sentence answering the question "Where is the [figure]?" For example, the participant may see "The cup ____ the table," and respond "The cup is on the table."[3] The topological relation markers (prepositions or short phrases) supplied by each participant were sanitized by the experimenter, collapsing over responses that differed solely in components without spatial meaning (e.g., variation in verb tense).

2. Category mapping task. After the naming data was sanitized to produce a list of unique labels given by the participant, the experimenter provided an array (from Levinson et al. (2003) Figure 5) with all stimuli organized for contiguity in the spatial relations depicted. Participants were then asked, for each unique spatial category they had named, to first identify the TRPS scene that is the best example (BE) of that category by placing a large coin on the scene in the array, then to identify all good examples (GEs) of that category (with smaller coins, e.g. nickels), and finally to identify all exemplars (AEs) of that category (by placing small coins on each exemplar in the array to visually "map" the category).

Naming data. In total, participants used 55 unique spatial labels in English, 146 in Spanish, 22 in French, 29 in Japanese, 56 in Dutch, 149 in Korean, and 100 in Mandarin. We selected a subset of these responses for analysis by taking the label most frequently applied to each of the 71 TRPS scenes by speakers of each language (with ties broken randomly). This produced a total of 85 *modal categories* for further analysis (11 for English, 9 for Japanese, 9 for French, 9 for Spanish, 8 for Mandarin, 19 for Korean, and 20 for Dutch; see listing in Appendix, Table 1[4]).

Analysis and results

1) Is there consensus on focal spatial relations?

To determine whether speakers within each language share foci for common spatial categories in their language, we measure how well the speakers' choices of best examples align with each other. For each of the 85 spatial categories c, we created a 71-dimensional vector b_c representing the TRPS stimuli in which we tally the number of times speakers of that language chose each stimulus as a best example for category c. To measure how well speakers align with each other on the best examples for each category c, we use entropy (H), a measure of the uncertainty of a distribution:

$$H(b_c) = -\sum_{i=1}^{n} p(b_{c,i}) \cdot \log_2(p(b_{c,i})) \quad (1)$$

where $p(b_{c,i}) = b_{c,i}/\sum_j b_{c,j}$, that is, the proportion of a language's speakers that chose stimulus i as the best example of category c. Entropy is minimal (0) if all speakers choose the same best example (i.e., a Dirac distribution), and maximal ($\log_2(n)$, here $\log_2(71) = 6.15$) if the distribution of best examples is uniform across all stimuli. Thus, entropy is a measure of how flat or un-peaked a distribution is. The average entropy of these empirical best example distributions is $M_{emp} = 0.99$ ($SD = 0.70$), much lower than the entropy of a uniform distribution—but high enough to indicate variation in best example choices.

To determine if the amount of alignment within each category is greater than might be expected by chance, we modeled chance agreement as a scenario in which each participant randomly chose a best example from the set of scenes they had selected in the category mapping task as good or best examples of the category. Following this approach, we would expect to see peaks in each simulated best example distribution resulting from coincidences in random selection, but also as a result of varying categorization across participants: often one participant's good examples of "on" represent a subset of another participant's good "on" selections, creating peaked best example distributions in this simulation even when all members of a category have an equal probability of being selected as the best example. To model chance entropy values for each category, we used Monte Carlo simulations to create pseudo-random distributions of best examples for each of the 85 categories, and compared the empirical entropy of each category's best examples (BEs) to the entropy values of the simulated distributions. To create the simulated BE distribution for each category, we simulated each speaker choosing at random one of their best or good examples for that category. Thus, each simulated best example distribution $b_{c,sim}$ was comparable to the original in having the same number of votes as the empirical distribution, but chosen at random from each speaker's best *and* good examples.[5] For each of the 85 categories, 2,000 simulated best example distributions were created, and the entropy of each was calculated. We then measured where in this distribution of simulated entropies the empirical category's entropy fell. If speakers of each language agree *substantively* with each other (within languages) on the best examples for each category, then the entropy of the empirical best example distribution should be smaller than the entropies of more than 95% of the resampled distributions. Indeed, this was true for 76 of the 85 categories.[6]

Across all 85 categories, the entropy of the empirical best examples ($M_{emp} = 0.99$) is significantly lower than the mean entropy of 2000 example vectors randomly-sampled from participants' naming data for each category ($M_{sim} = 1.81$; paired $t(83) = 13.78$, $p < .001$). That is, empirical best examples of each category are significantly more peaked than they would be if chosen at random from all good and best

[3] Mandarin speakers filled in two separate blanks at the typical positions for verbs and prepositions, respectively.

[4] We render Korean in Hangul to avoid ambiguity across differing romanization schemes.

[5] This procedure was also performed using speakers' naming data instead, with very similar results.

[6] The 9 exceptions: SP 'cuelga,' FR 'dessus,' JP 'ni,' KO '나,' and '달려,' NL 'om,' 'hangen aan,' 'zitten om,' and 'zitten aan.'

examples selected by speakers as part of that category. Figure 3 shows the entropy of the empirical best example distribution for each of the 85 modal categories plotted against the mean of the 2,000 resampled entropies for each. Overall, the empirical best examples were more aligned than expected by chance in a majority of the categories, showing that speakers of a given language largely agree on focal spatial relations. We now turn to whether this alignment on spatial foci is also seen cross-linguistically.

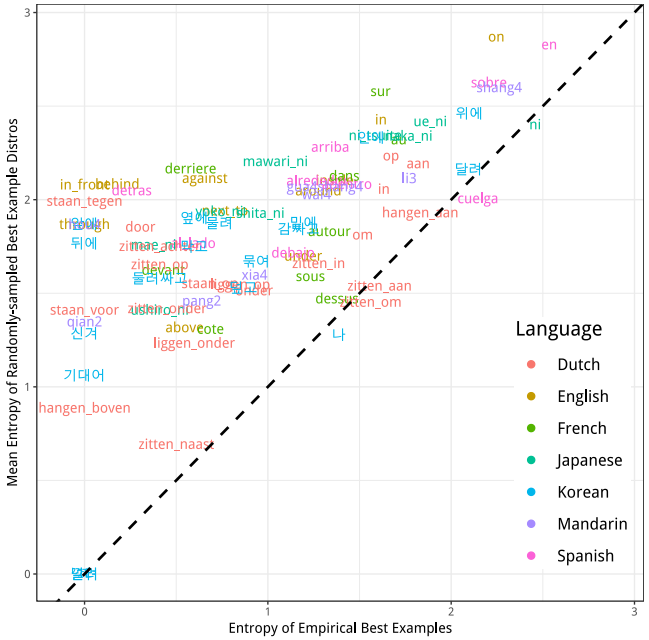


Figure 3: Consistency of best example choices across speakers for each category. Empirical entropy of the best example distributions of 85 spatial categories in 7 languages vs. the mean entropies of 2,000 randomly-chosen best example distributions created from each participant's chosen good and best examples of a category. The empirical best example distributions showed more alignment (lower entropy) than the resampled distributions for 80 of the 85 categories.

2) Does this consensus on focal spatial relations extend across languages?

We now investigate if there is consistency in the stimuli that get selected as best examples by speakers of different languages. In other words, we ask whether different languages align their best example choices on the same stimuli. To do so, we first tallied each language's best example distribution for the modal categories over all 71 TRPS stimuli, adding the b_c vectors for each language L into a single summed BE count vector per language, b_L. These summed BE counts, b_L, were then normalized $p(b_{L,i}) = b_{L,i}/\sum_j b_{L,j}$, meaning cell $p(b_{L,i})$ is the probability that stimulus i was selected as a best example for any of the modal categories of language L. The language-specific best example distributions $p(b_L)$ were then averaged together (with equal weight to each language) to obtain a cross-language BE distribution.

Figure 4 shows normalized best example distributions per language, as well as the cross-linguistic average ("all languages"). To determine how aligned the best examples are across languages, we compare the entropy of the cross-language distribution (3.70) to the distribution of entropies from a Monte Carlo simulation. For each language's summed BE distribution $p(b_L)$, the probabilities across stimuli were randomly permuted (swapping cells to preserve the overall structure of the distribution), and then the resulting normalized cross-language distribution was calculated (as above) on the permuted summed distributions for each language. The entropy of this pseudo-random cross-language BE distribution was found, and this procedure was repeated 10,000 times to generate a set of permuted entropies. The resulting distribution is shown in Figure 5. The empirical distribution's entropy (3.7, shown in red) was lower than all 10,000 entropies of the permuted distributions, which had a mean of ($M = 3.81$).

An additional, possibly more conservative, Monte Carlo simulation was also carried out. As before, the counts across stimuli for each language were randomly permuted, but this time only shuffling between stimuli that were selected by at least one speaker of the relevant language as a best example. Only permuting non-zero slots may increase the likelihood of chance alignment, depending on the number of such slots and their pre-existing cross-linguistic alignment. However, the empirical distribution's entropy was again lower than all 10,000 entropies of the randomly-permuted non-zero distributions, which had a mean of 3.82. These results confirm quantitatively that speakers of these seven languages share some consensus on foci for spatial relations, as is suggested qualitatively by inspection of Figure 4, where we have highlighted nine spatial scenes that were selected as best examples by a large proportion of participants across languages.

3) Do spatial categories exhibit composite structure, with more than one focus per category?

Finally, we consider three cases of composite spatial categories, analogous to "grue" in color, in which a single lexical category includes multiple foci. For this, we examine OVER/ON categories, at the third stage of the hierarchy in Figure 2. Levinson et al. (2003) propose that categories inclusive of OVER and ON senses are composites of four spatial foci: OVER, ON, ON-TOP ("location above eye-level"), and ATTACHMENT. In keeping with parallel work on color, Levinson et al. suggest that composite categories may or may not be focused at all of their constituent foci, so clustering of best example choices at OVER, ON, ON-TOP, ATTACHMENT, or any combination of these senses is consistent with this view. Alternatively, many classic models of central tendencies (e.g., mean, mode, prototype) would predict a single central focus. To the extent that the OVER, ON, ON-TOP, and ATTACHMENT senses are distinct from each other, a single-focus view suggests that a lexical category would be focused at only one of these four senses.

We will examine the best example distributions for three

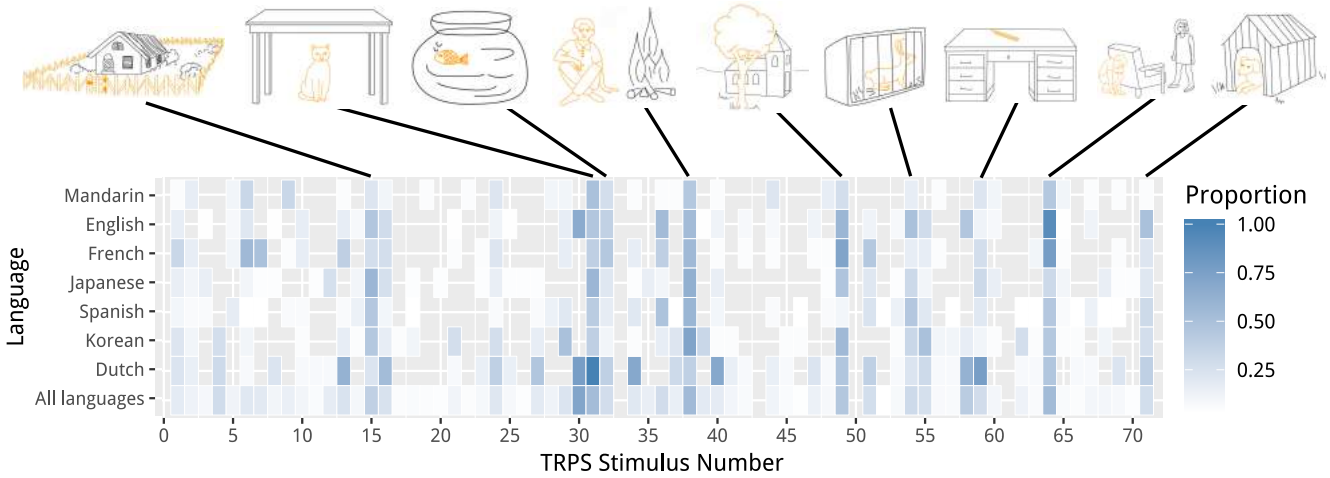


Figure 4: Focal spatial relations. Heatmap of the proportion of participants choosing each TRPS stimulus as a best example for the modal spatial categories in each language. The nine stimuli at the top are those that were selected as best examples in all seven languages, and selected by a large proportion of participants in all languages (best example frequency was greater than 1SD above the median for best examples in the "all language" average).

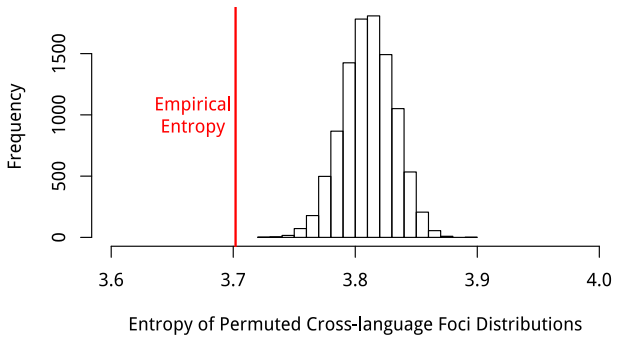


Figure 5: The entropy of the empirical cross-linguistic BE distribution was less than the entropy of all 10,000 randomly-permuted (by language) BE distributions ($p(b_L)$), whether all cells of $p(b_L)$ are permuted (shown) or only non-zero cells.

composite categories spanning these four predicted foci for evidence of composite (bi- or multi-modal) foci. To do so, we compare the best examples of OVER/ON composite categories in Mandarin, Korean, and Japanese to two smaller categories that represent the next stage of subdivision in Levinson et al.'s spatial hierarchy—OVER and ON (the latter inclusive of ON-TOP and ATTACHMENT)—using the closest corresponding modal categories in English, Spanish, French, and Dutch. If the composite categories in Mandarin, Korean, and Japanese have composite *foci*, we would expect their focus distributions to resemble combinations of the focus distributions for ON and OVER in languages that distinguish these senses (i.e., English, Spanish, French, and Dutch).

In this analysis, we measure the similarity of normalized BE distributions of individual spatial categories ($p(b_c)$) from different languages. Following the color literature, the similarity of two distributions will be measured using Jensen-Shannon Divergence (JSD), a finite-valued, symmetric measure of the difference between two probability distributions. JSD is minimal, 0, when the two distributions are identical and has a maximum of 1 in our comparisons.

The three composite categories we consider are Mandarin "shang4," Korean "위에," and Japanese "ue ni." Shown in Figure 6, the foci of these three categories closely correspond to each other (M-K JSD=.23; M-J JSD=.30; K-J JSD=.27). Mandarin's "shang4" corresponds well to the combined (averaged) category foci of two categories in the four other languages: English "above" and "on" (JSD=.35), Spanish "arriba" and "sobre" (JSD=.21), French "dessus" and "sur" (JSD=.21), and Dutch "hangen boven" and "op" (JSD=.46). Like "shang4", Korean "위에" corresponds similarly well to the same combined foci: English "above" and "on" (JSD=.21), Spanish "arriba" and "sobre" (JSD=.15), French "dessus" and "sur" (JSD=.11), and Dutch "hangen boven" and "op" (JSD=.35). Similarly, Japanese "ue ni" matches the averaged BE distribution of English "above" and "on" (JSD=.41), Spanish "arriba" and "sobre" (JSD=.32), French "dessus" and "sur" (JSD=.29), and Dutch "hangen boven" and "op" (JSD=.44) Importantly, these OVER-ON category pairs all have foci distributions that are more distant from each other: above-on JSD=1.0, arriba-sobre JSD=.50, dessus-sur JSD=.92, hangen boven-op JSD=1.0.[7] As shown in Figure 6, this suggests the existence of composite spatial categories with multiple distinct foci, analogous to "grue" cases in the color domain.

[7]The mean JSD of any category's foci to the average of any two other categories' foci is .91 (median=1), and the mean JSD of any two single categories' foci is 0.97 (median=1).

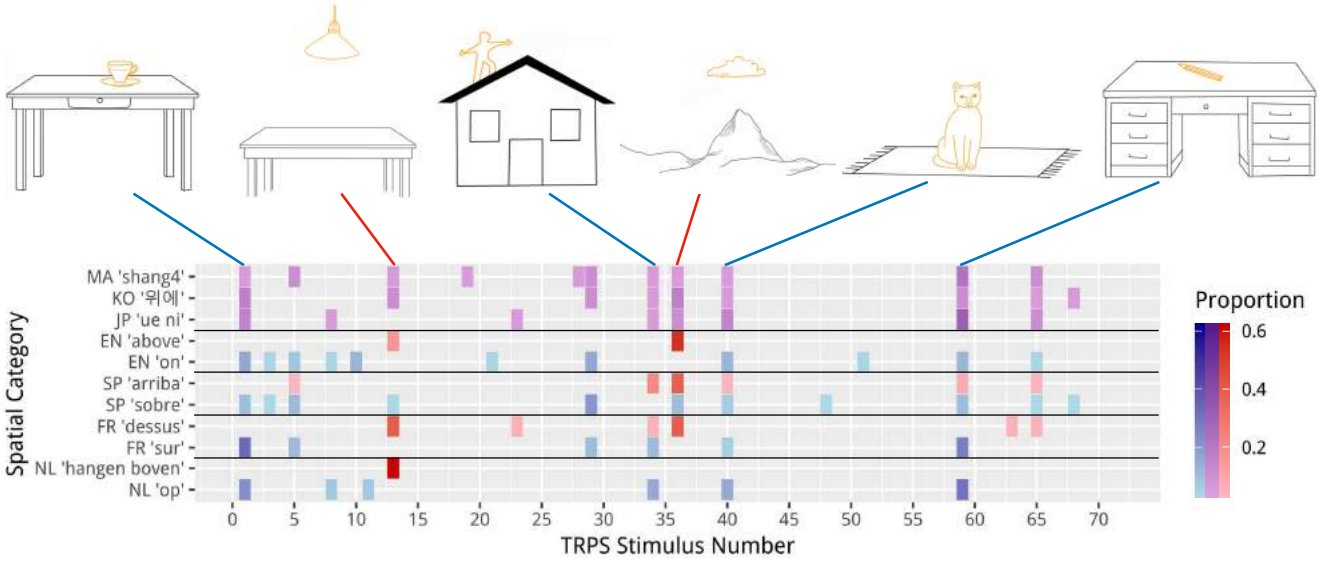


Figure 6: Cross-linguistic comparison of three expected composite categories: the best examples of Mandarin (MA) "shang4," Korean (KO) "위에," and Japanese (JP) "ue ni" span the best examples of separate OVER and ON categories in other languages (e.g., English (EN) "above" and "on"). The six scenes depict foci speakers align on, with red lines indicating OVER foci and blue lines indicating ON-TOP (man on house) and ON foci. Purple heatmap color coding indicates terms with composite extensions, red indicates OVER terms, and blue indicates ON terms.

Discussion

This study used empirical best example data from seven languages to explore whether spatial topological categories have focal members comparable to those in color. We documented names and best examples of topological spatial relations in Dutch, English, French, Japanese, Korean, Mandarin Chinese, and Spanish. To our knowledge, this is the first study to directly acquire and analyze best examples of spatial relations—although others e.g., Landau, Johannes, Skordos, and Papafragou (2017) have investigated related notions such as "core" spatial concepts.

In the first analysis, we considered whether there was consensus within languages on the best examples of spatial relations. Indeed, for the majority of categories speakers were significantly more aligned in their choice of best example than would be expected by chance (i.e., if they had drawn best examples merely from their chosen good or best examples). This demonstrates that within each of these seven languages, speakers tend to agree on focal spatial relations.

Our second analysis examined whether this consensus on focal spatial relations extended across languages. We found that the empirical cross-language distribution of best examples was significantly more aligned than would be expected by chance, confirming that speakers of these languages share some consensus on foci for spatial relations.

Finally, we investigated whether spatial categories reflect composite structure, with focal distributions organized around multiple distinct senses. For this, we examined the best examples of Mandarin "shang4," Korean "위에," and Japanese "ue ni," broad categories that encompass multiple predicted foci. We found evidence suggesting that these categories are indeed semantic composites, focused at multiple senses: the best examples of these large categories resembled combinations of best examples from distinct (and uncorrelated) categories, such as English "above" and "on." This finding supports a previous account of spatial topological semantics and may provide evidence for composite categories in the spatial domain comparable to "grue" in color.

However, there are grounds for caution in the interpretation of these findings. The classic composite category within the color domain, "grue," is evidenced by a focal distribution with both blue and green best examples, but where intermediate colors are not selected as best examples, making for two distinct peaks in the focal distribution. While the "above" and "on" foci selected as best examples of Mandarin "shang4," Korean "위에," and Japanese "ue ni," correspond to distinct attractors or "notional clusters" in Levinson et al.'s (2003) proposal, it is possible that "intermediate" spatial notions would also be selected as focal, making for a single focal peak that is inclusive of both "above" and "on" senses. Future work should examine possible composite categories with clear intermediate cases between the predicted foci to determine whether these senses are indeed distinct, exhibiting the double-peak structure seen in some "grue" cases.

This study offers empirical evidence for universal tendencies in spatial relations based on naming and best example data. Our findings provide evidence for focal best examples in the spatial domain and contribute further support for a theory of "natural concepts" in this domain.

Acknowledgments

We are very grateful to Aaliyah Aya Ichino, Jonatan Malis, Aagje van der Meer, Maggie Soun, Katie Chen, Vanessa Matalon, Ana Cuevas, Jongmin Jerome Baek, and Jae Hun Kim for data collection. We also thank Oana David, Iksoo Kwon, Yang Xu, and Christine Tseng for help with translation and data coding, and members of the Language and Cognition Lab at Stanford for helpful comments. This work was supported by NSF under grant SBE-1041707, the Spatial Intelligence and Learning Center (SILC), and under NSF Graduate Research Fellowship grant DGE 1106400 to AC.

References

Abbott, J. T., Griffiths, T. L., & Regier, T. (2016). Focal colors across languages are representative members of color categories. *Proceedings of the National Academy of Sciences*, *113*(40), 11178–11183. doi: 10.1073/pnas.1513298113

Berlin, B., & Kay, P. (1969). *Basic color terms: Their universality and evolution*. Univ. of California Press.

Bowerman, M., & Pederson, E. (1992). Cross-linguistic studies of spatial semantic organization. *In Annual Report of the Max Planck Institute for Psycholinguistics 1992*, 53-56.

Carstensen, A., & Regier, T. (2013). Individuals recapitulate the proposed evolutionary development of spatial lexicons. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th annual meeting of the cognitive science society* (p. 293-298). Austin, TX: Cognitive Science Society.

Feist, M. I. (2000). *On in and on: An investigation into the linguistic encoding of spatial scenes* (Unpublished doctoral dissertation). Northwestern University.

Heider, E. R. (1972). Universals in color naming and memory. *Journal of Experimental Psychology*, *93*, 10-20.

Johannes, K., Wang, J., Papafragou, A., & Landau, B. (2015). Similarity and variation in the distribution of spatial expressions across three languages. In *Proceedings of the 37th annual meeting of the cognitive science society* (p. 997-1002). Austin, TX.

Johannes, K., Wilson, C., & Landau, B. (2016). The importance of lexical verbs in the acquisition of spatial language: The case of *in* and *on*. *Cognition*, *157*, 174-189.

Kay, P., & McDaniel, C. K. (1978). The linguistic significance of the meanings of basic color terms. *Language*, *54*(3), 610–646.

Landau, B., Johannes, K., Skordos, D., & Papafragou, A. (2017). Containment and support: Core and complexity in spatial language learning. *Cognitive Science*, *41*(S4), 748-779.

Levinson, S. C., & Meira, S. (2003). Natural concepts in the spatial topological domain-adpositional meanings in cross-linguistic perspective. *Language*, *79*, 485-516.

Regier, T., Kay, P., & Cook, R. S. (2005). Focal colors are universal after all. *Proceedings of the National Academy of Sciences*, *102*(23), 8386–8391. doi: 10.1073/pnas.0503281102

Roberson, D., Davies, I., & Davidoff, J. (2000). Color categories are not universal: Replications and new evidence from a stone-age culture. *Journal of Experimental Psychology: General*, *129*(3), 369-398.

Appendix: Modal categories in spatial naming

Language	*Most Frequent Spatial Terms (N)*
English	in (24), on (24), around (23), behind (23), under (23), next to (19), inside (18), above (17), through (16), against (14), in front (14)
Japanese	naka ni (19), shita ni (19), ue ni (19), mawari ni (16), yoko ni (16), ni (13), ni tsuite (13), mae ni (8), ushiro ni (7)
French	au (18), autour (18), cote (18), dans (18), dessus (18), sur (18), derriere (17), sous (17), devant (13)
Spanish	alrededor (28), adentro (26), sobre (25) en (24), arriba (22), al lado (21), debajo (21), detras (14), cuelga (11)
Mandarin	shang4 (17), li3 (16), xia4 (15), wai4 (13), gua4 shang4 (13), pang2 (10), hou4 (8), qian2 (4)
Korean	안에 (17), 옆에 (16), 위에 (16), 밑에 (14), 달려 (12), 감싸고 (10), 묶여 (9), 앞에 (9), 뒤에 (8), 물려 (8), 둘러싸고 (7), 막고 (7), 덮고 (6), 나 (4), 신겨 (4), 기대어 (3), 깔려 (1), 껴 (1), 널려 (1)
Dutch	onder (13), op (13), aan (12), in (12), om (12), door (11), hangen aan (9) liggen op (9), staan op (9), hangen boven (8), staan tegen (8), liggen onder (7), zitten achter (7), zitten in (7), zitten op (7), staan voor (6), zitten om (6), zitten onder (6), zitten aan (5), zitten nast (5)

Table 1: The 85 modal spatial categories used in the analysis, organized by language. Numbers indicate how many participants produced each category label (e.g., all 24 English speakers produced “in”).

The shape of language experience in two traditional communities

Marisa Casillas
Max Planck Institute for Psycholinguistics, Nijmegen, Netherlands

Abstract

This study sketches the language environments of children ages 0;03;0 growing up in two traditional, indigenous communities: one Tseltal (Mayan) and the other Yl (Papuan). Past ethnographic work has suggested that caregivers' ideas about talking to young children differ greatly between these two communities. However, the present daylong recording analyses suggest that, in fact, children are rarely directly addressed in both places, with no age-related increase and with most child-directed speech coming from adults. Children's manual activities also suggest that child-carrying practices and cultural context moderate the extent to which children might use co-occurrence between held objects and ambient language to learn words.

The Role of Basal Ganglia Reinforcement Learning in Lexical Priming and Automatic Semantic Ambiguity Resolution

Jose M. Ceballos, Andrea Stocco, Chantel S. Prat
{josemceb, stocco, csprat}@uw.edu

Department of Psychology and
Institute for Learning & Brain Sciences,
University of Washington
119A Guthrie Hall, Seattle, WA 98195 USA

Abstract

The current study aimed to elucidate the contributions of the subcortical basal ganglia to human language by adopting the view that these structures engage in a basic neurocomputation that may account for its involvement across a wide range of linguistic phenomena. Specifically, we tested the hypothesis that basal ganglia reinforcement learning mechanisms may account for variability in semantic selection processes necessary for ambiguity resolution. To test this, we used a biased homograph lexical ambiguity priming task that allowed us to measure automatic processes for resolving ambiguity towards high frequency word meanings. Individual differences in task performance were then related to indices of basal ganglia functioning and reinforcement learning, which were used to group subjects by learning style: primarily from choosing positive feedback (Choosers), primarily from avoiding negative feedback (Avoiders), and balanced participants who learned equally well from both (Balanced). The pattern of results suggests that balanced individuals, whom learn from both positive and negative reward equally well, had significantly lower access to the subordinate homograph word meaning. Choosers and Avoiders, on the other hand, had higher access to the subordinate word meaning even after a long delay between prime and target. Experimental findings were then tested using an ACT-R computational model of reinforcement learning that learns from both positive and negative feedback. Results from the computational model confirm and extend the pattern of behavioral findings, and provide a reinforcement learning account of lexical priming processes in human linguistic abilities, where a dual-path reinforcement learning system is necessary for precisely mapping out word co-occurrence probabilities.

Keywords: language; semantics; lexical selection; ambiguity resolution; priming; reinforcement learning; basal ganglia; dopamine; cognitive modeling; ACT-R

Introduction

The field of the neurobiology of language has traditionally focused on the contributions of cortical structures to linguistic processes (Tremblay & Dick, 2016). However, research from different sub-fields suggests that the subcortical basal ganglia are an essential part of the neurobiological bases of human linguistic abilities (Crosson, 1985; Booth, Wood, Lu, Houk, & Bitan, 2007; Seo, Stocco, & Prat, 2018). To date, no existing account of the neurobiology of language is able to systematically explain what the role of these subcortical structures is across the many levels of linguistic processing. Thus, in its current stage, the field suffers from a limited understanding of the neural processes that give rise to language. A detailed whole-brain understanding of this human ability is key to inform robust models of language neurobiology and also to advance our understanding of language disabilities for translational purposes. In an effort to contribute to a whole-brain model of language functioning, this work focuses on understanding the role of the basal ganglia in language.

Given that the basal ganglia are some of the most neurobiologically ancient structures (Lieberman, 2001), it is reasonable to assume that their role in human linguistic abilities is analogous to the more general motor or cognitive functions observed in other species. Indeed, many prominent theories and models of basal ganglia functioning stem from observations of motor control (Mink, 1996) and extend these functions to non-motor and abstract cognitive processes spanning from cognitive control (Graybiel, 1995; Stocco, Lebiere, & Anderson, 2010) to working memory capacity (Hazy, Frank, & O'Reilly, 2007). Thus, the current research aims to understand basal ganglia contributions to language in the context of the already well-understood and well-established theories of selection and reinforcement learning (RL). To test the hypothesis space of basal ganglia selection processes in language, we turned to semantic processing as a model system for competition between multiple viable alternatives. Specifically, this work is grounded on models of models of semantic activation spreading (Collins & Loftus, 1975).

Semantic ambiguity (also referred to as lexical ambiguity) occurs when a word refers to multiple different concepts (Vitello & Rodd, 2015). For example, the word "hot" can refer either to temperature or to food spiciness. Cases of semantic ambiguity may arise in conversational settings, and are also more commonly encountered in written form such as news headlines, puns, poetry, and novels. The ability to properly disambiguate an input into the contextually appropriate represented meaning is key for listening and reading comprehension. More importantly, this process provides details on a fundamental neurocognitive mechanisms, such as the contextual integration of information, statistical learning, inhibition, and selection processes used to manage simultaneous and competing neural representations that are at odds with the task goal of accurate transfer of information in communicative settings.

Semantic ambiguity can arise in a variety of different ways. The first class of ambiguity arises from words that have different unrelated meanings. For example, "bark" can refer to the sound a dog makes, or the outermost layer of a tree. In this case, both meanings of "bark" constitute a true homonym,

but are also homographs and homophones (same spelling, and same sound, respectively). Furthermore, words can be encountered in contexts where only the written form is ambiguous (e.g., the homographs for "lead"), or only the spoken form is ambiguous (e.g., the homophones for "be/bee" or "seam/seem").

The cognitive mechanisms supporting the resolution of semantic ambiguities are best understood by exploring theories on the dynamics of semantic information and its representation. When a listener or reader first encounters a word with multiple meanings, all meanings are quickly activated and available in parallel. This refers to the automatic component in semantic processing. Furthermore, if encountered in isolation or in a highly ambiguous context, an ambiguous word will be automatically disambiguated towards the highest frequency meaning, reflecting another series of automatic selection processes. However, if an ambiguous word is encountered following a strong biasing context towards one specific meaning, only the contextually-relevant word meaning is available. This suggests that when ambiguous words are encountered, all meanings are initially activated, but this activation is modulated by multiple factors such as sentence context and meaning frequency.

While most research focused on understanding the neural mechanisms supporting lexico-semantic processing and ambiguity resolution has focused on cortical structures such as the left inferior frontal gyrus (for a review, see Vitello & Rodd, 2015), there is evidence suggesting a key involvement of subcortical structures in this process (e.g., Ketteler, Kastrau, Vohn, & Huber, 2008; Mason & Just, 2007). For example, a lexical priming investigation found that monolingual individuals experience abnormalities in the neurocognitive dynamics that shape lexical priming (Copland, Chenery, & Murdoch, 2001). Specifically, healthy participants show no traces of subordinate word activation following a long delay between prime and target, and thus reflect automatic semantic ambiguity resolution towards the dominant or highest frequency meaning. Parkinson's Disease (PD) patients, whom have decreased dopaminergic functioning resulting in a general hyperactivity of the basal ganglia indirect pathway, on the other hand, exhibit a longer-term activation of the multiple competing representations.

Although findings such as these have been traditionally framed under a selection and inhibition framework, we explore the hypothesis that the signature role of basal ganglia in RL may more accurately explain its role in semantic processing. In other words, the basal ganglia may be involved in statistical learning and predictive processing during language comprehension. Critical for the current investigation, the activity of the basal ganglia is often modeled as reflecting Temporal Difference (TD) learning. As it happens, TD-learning does not accurately reflect the computations of the basal ganglia, which are the result of the opposite contributions of two conflicting pathways. Their contribution have been modeled as the sum of competing RL systems (Frank, Seeberger, & O'reilly, 2004; Stocco, 2018). Individuals vary in the learning rates of the two pathways as a function of biological parameters (such as density of dopamine receptors: Frank, Moustafa, Haughey, Curran, & Hutchison, 2007) and external factors (administration of dopamine: Frank et al., 2004), and individual differences in the preponderance of each pathway can be indirectly measured through the PSS task (Frank et al., 2004; Stocco et al., 2017). Thus, the current investigation tests the hypothesis that individual differences in PSS task behavioral indices of basal ganglia pathways will be related to performance in a lexical prime style task. Furthermore, we make the prediction that a balance in functioning across both pathways is critical for optimal semantic processing and ambiguity resolution.

Methods

Participants

Informed consent was obtained from participants prior to the experiment, as outlined by the University of Washington Institutional Review Board. Participants were recruited using the Psychology Departments Participant Research Pool and all participants were compensated with course credit for their participation. Data were collected from 140 healthy monolingual participants (66 females, mean age = 19.4 years). Seven subjects were excluded from analyses due to low accuracy ($\leq$0.50) in the primary experimental task, the Word-Pair Task (WPT).

Tasks

All participants completed the following tasks in four pseudo-randomized orders to control for possible fatigue effects induced by the WPT and PSS task length.

Word-Pair Task Measures of lexical priming were collected using the WPT. This task was designed to measure the availability of dominant and subordinate word meanings following the presentation of primes with multiple meanings. The primes used shared both phonetic and orthographic forms across both word meanings, making them true homographs (e.g., "Bat"). The prime and target words were presented in the center of the screen, one at a time, separated by an inter-stimulus interval (ISI) of either 150 ms (short) or 850 ms (long). Prior to starting the task, participants were asked to place their right index finger on the "P" key of the keyboard, and their left index finger on the "Q" key of the keyboard. Participants were then asked to respond with a button press if the target word was related or unrelated to the prime. Key mappings for related and unrelated were counterbalanced.

There were two conditions of interest (1 and 2) and two control conditions (3 and 4): (1) homograph prime / dominant target, (2) homograph prime / subordinate target, (3) prime / related target, and (4) prime / unrelated target. These four conditions will be referred to as dominant, subordinate, related, and unrelated (respectively) form here on for simplicity purposes. Participants completed 100 total prime-target

pair trials, where 20 belonged to condition 1, 20 to condition 2, 30 to condition 3, and 30 to condition 4. Word frequency meanings were obtained from (Twilley, Dixon, Taylor, & Clark, 1994), and subordinate words were defined as having a relatedness frequency of less than 0.3 in a 0-1 scale, while dominant words had a relatedness frequency of greater than 0.7. Since each homograph prime is associated with two meanings, but each prime was presented once for each participant, two WPT versions were created. In one version, the dominant meaning of a homograph was used (e.g., version A contained "Bank" / "Money") while the other version used the subordinate meaning (e.g., version B contained "Bank" / "River"). The two lists were counter-balanced for word frequency, word length, and syllable length.

Probabilistic Stimulus Selection Task The PSS task is an iterative, two-alternative, forced-choice decision-making paradigm first introduced by Frank et al. (2007). In this task, participants are repeatedly asked to select one of two stimuli presented on the screen. Participants are also told that some of their choices would result in success, and some of them would result in failure, depending on which stimulus they choose. Feedback on the outcome of their decision is presented immediately after participants select a stimulus. To encourage participants to avoid explicit strategies (such as rote memorization of each stimulus history of successes), stimuli are implemented as complex shapes that are difficult to verbalize, typically Hiragana characters presented to non-Japanese speaking participants. Unbeknownst to participants, each stimulus has a predefined "success" probability. Six stimuli in total are used in the experiment, with success probabilities varying linearly from 80% to 20%. In the first phase, the stimuli are divided into fixed pairs, with the highest probability stimulus always paired with lowest probability one, then second higher stimulus paired with the second lowest one, and the third highest probability stimulus paired with the third lowest one.

Two values are calculated from the test phase of the PSS task: *Choose* accuracy, which represents the accuracy in choosing the most rewarding stimulus over others; and *Avoid* accuracy, that is, the proportion of times in which participants avoid the least rewarding stimulus. If we indicate the six stimuli with the letters $A, B \ldots F$, with A being the most rewarding stimulus and B the least rewarding one, then Choose and Avoid accuracies are calculated as the probability of choosing A when paired against C, D, E, and F, and the probability of choosing C, D, E, or F when they are paired with B, respectively.

Previous patients and genetic studies have demonstrated a functional connection between these two measures and the basal ganglia pathways. For example, Parkinson's patients, whose indirect pathway dominates over the direct one due to a loss of dopaminergic inputs from the substantia nigra pars compacta (SNc), have higher Avoid accuracy than Choose accuracy. Furthermore, this pattern is reversed when drugs are administered that overcompensate the direct pathway activity. Additionally, individuals with genetic alleles that cause a greater production of dopamine receptors in the direct pathway tend to be Choosers rather than Avoiders; conversely, individuals whose alleles cause greater number of dopamine receptors in the indirect pathway tend to be Avoiders (Frank et al., 2007; Frank & Hutchison, 2009).

Analyses

Behavioral Data Cleaning Target words in the WPT were cleaned on a by-participant basis for RT outliers, defined as trial RTs greater than or lower than three standard deviations from the mean.

Participant Groups Participant groups were created using PSS Choose and Avoid scores. Since one of the guiding assumptions of this investigation was that one's ability to learn from *both* positive and negative feedback, groups were created using a relative score where Avoid was subtracted from Choose, which resulted in scores between 100 and -100. Third-group splits were then used to separate individuals into participant groups. Thus, high values (approximately 33 to 100) reflected participants who learned primarily from positive feedback (Choosers), low values (approximately -100 to -33) reflected participants who learned primarily from negative feedback (Avoiders), and values around zero (-33 to 33) reflected individuals who learned equally as well from positive and negative feedback (Balanced). This resulted in 44 Choosers, 38 Avoiders, and 52 Balanced participants.

Analysis with Linear Mixed Effects Models The data were analyzed using linear mixed effects (LME) models, as this method has been previously shown to outperform the traditional procedures such as ANOVA (Kristensen & Hansen, 2004), and can adequately handle imbalances in group sizes. However, for validation purposes, the same results were reproduced using ANOVA (although not reported herein). LME models were specified using the R lme4 package (Bates, Mächler, Bolker, & Walker, 2015). The model was specified using the following formula:

$$\text{Target Accuracy} \sim \text{ISI} \times \text{Condition} \times \text{PSS Group} + (1 + \text{Condition} \mid \text{Participant})$$

where the dependent variable is Target Word accuracy, the fixed-effects term is the factors for ISI (short or long) $\times$ Condition (dominant or subordinate) $\times$ PSS Group (Choosers, Balanced, or Avoiders), and the random effects term allows for each participant to have a different slope (or effect) for Condition, while intercepts and slopes for each participant by Condition are allowed to be correlated (e.g., higher intercepts may also have steeper slopes). Finally, a type III ANOVA with Satterthwaite's method was used to test for significance between the factors of interest in the LME model.

Computational Model

A theoretical model was implemented to examine predictions on the relationship between reward learning and lexical re-

trieval[1]. The model was developed in the ACT-R cognitive architecture (Anderson, Fincham, Qin, & Stocco, 2008; Anderson et al., 2004), a general theory of cognition that enables the development of complete models capable of end-to-end simulations of a complete task while maintaining a high degree of psychological plausibility. The model described herein is based on a previously published model of the role of the basal ganglia in the PSS task (Stocco, 2018). According to this model, the conflict between the two pathways can be simulated in ACT-R as a conflict between the selection of opposite and symmetric *productions*, that is, state-action pairs that implement minimal cognitive steps. Productions representing the direct pathway implements "Go" actions, while those representing the indirect pathway represent opposite "No Go" actions. For example, the choice between two options in the PSS task, *A* and *B*, can be represented as the competition between two alternative pairs of productions, "Choose A" and "Avoid A" and "Choose B" and "Avoid A". In ACT-R, the competition between productions is resolved through a softmax algorithm that preferentially selects the actions with the highest estimated *utility*, a scalar quantity that depends on the history of previous successful uses of the production and is learned through a reinforcement-learning algorithm. Importantly, Stocco, 2018 noticed that both individual differences due to differential expressions of dopamine genes (Frank et al., 2007) and the effects of basal ganglia pathologies (Frank et al., 2004) can be successfully captured by differentially altering the learning rates of "Choose" and "Avoid" productions. The different learning rates will be indicated as α_C and α_A, respectively.

An ambiguity resolution experiment can also be understood as as a two-alternative forced choice (2AFC) task in the context of lexical retrieval. In essence, two homographs are competing for access to semantic retrieval. Consequently, for each choice, two competing selections are performed. Thus, if the two homographs are a dominant and a subordinate interpretation of the same written word, each of them will have two production rules associated with them, "Choose Dominant" and "Avoid Dominant", and "Choose Subordinate" and "Avoid Subordinate".

Contrary to traditional 2AFCs, in lexical access the two options are not equivalent in terms of response times. Selection of the dominant meaning is usually associated with much shorter retrieval times than selection of the non-dominant meaning. In our model, this was captured by forcing those production rules that select the subordinate meaning ("Avoid Dominant" and "Choose Subordinate") to have a longer execution time. As a consequence, under short ISI, the subordinate meaning is never successfully selected. Under longer ISIs, however, participants *do* have a chance to select these meanings, so that the eventual firing of productions that select the subordinate interpretation could result in the successful retrieval of the least common meaning of the homograph.

[1]Code for the model is available on our laboratory's GitHub repository: *http://github.com/UWCCDL/BAGELS_ACTR*

Finally, to derive predictions from the model, we conducted an extensive set of simulations of the utility values associated to production rules under different reward conditions, corresponding to different situations in which the selection of the dominant or subordinate meaning are correct. Specifically, we examined a hypothetical situation in which the dominant meaning is contextually correct 80% of the time and the subordinate 20% of the time. To simulate the large amount of experience with the occurrence statistics of different lexical items that is associated with adult native speakers, the model was let to learn the corresponding utility values until they reached asymptotic values.

Importantly, these simulations of language experience were conducted under different learning rate parameters. The parameter values were chosen to reflect the values that were found to best capture genetic variance of dopamine receptors in healthy adults in Stocco (2018). Specifically, we simulated three groups of individuals, exhibiting a preference to learn from positive feedback ($\alpha_C = 1.5, \alpha_A = 1.0$), a preference to learn from negative feedback ($\alpha_C = 1.0, \alpha_A = 1.5$), or no preference between the two ($\alpha_C = 1.5, \alpha_A = 1.5$). These parameters are associated with different performance profiles in the PSS task, corresponding to a preference for "Choose A", for "Avoid B", or for a balance between the two (Stocco, 2018).

Results

General Word-Pair Task Results

Mean accuracy for dominant trials (M = 0.90, SD = 0.14) was significantly higher than for subordinate trials (M = 0.55, SD = 0.15, t(138) = 22.17, $p < 0.0001$). Differences in mean RTs were also observed, faster for dominant trials (M = 832.04, SD = 194.71) than subordinate trials (M = 996.25, SD = 238.26, t(138) = -13.77, $p < 0.0001$).

General Probabilistic Stimulus Selection Task Results

Subjects performed similarly across Choose (M = 69.78, SD = 22.24) and Avoid (M = 67.99, SD = 22.22) trials. Furthermore, as in previous studies using the PSS Task (Stocco et al., 2017; Frank et al., 2007; Frank & Hutchison, 2009), Choose and Avoid trials were not correlated (r(138) = -0.12, p = 0.14).

Linear Mixed Effects Model Results: Relating WPT Performance and PSS Groups

The LME model predicting Target accuracy had a total explanatory power (conditional R^2) of 90.62%, in which the fixed effects explained 68.43% of the variance (marginal R^2). The model revealed a significant main effect of Condition ($F(1, 131) = 1096.33$, $p < 0.0001$). A significant two way interaction between Condition $\times$ ISI was also observed ($F(1, 262) = 6.47$, $p = 0.012$), alongside a significant three-way interaction between Condition $\times$ ISI $\times$ PSS Group ($F(2, 262) = 3.86$, $p = 0.022$). Marginal two-way interactions were observed for Condition $\times$ PSS Group ($F(2, 131) = 2.45$, p =

0.087) and also ISI × PSS Group (F(2, 262) = 3.00, p = 0.051). For details, see Figure 1.

A follow-up analysis using the orthogonal contrasts extracted from the LME model suggest that the three-way interaction between Condition × ISI × PSS Group is explained by higher accuracy to Target Words during the Subordinate condition observed in PSS Choosers (difference = 0.083, t(166.85) = 2.41, p = 0.028) and Avoiders (difference = 0.086, t(166.95) = 2.60, p = 0.017), relative to the Balanced group, during the long ISI.

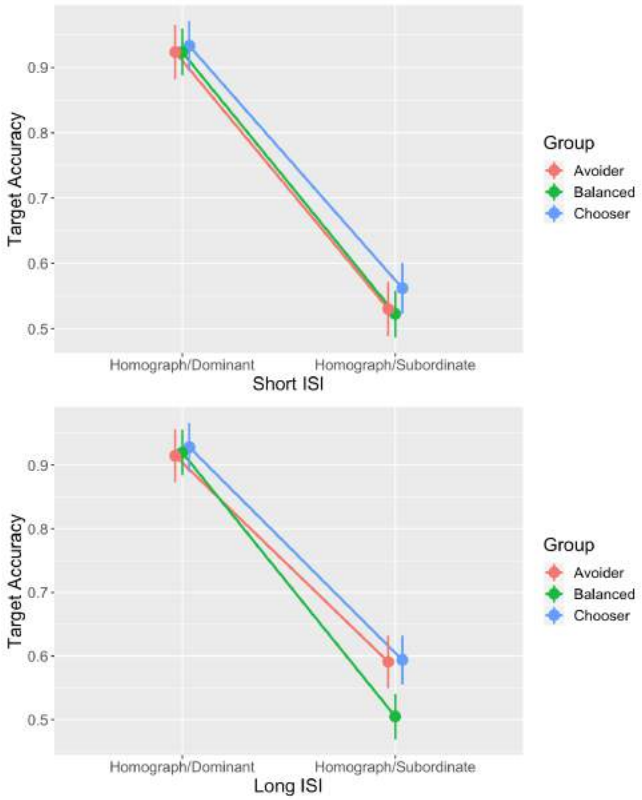


Figure 1: Top: Accuracy Dominant and Subordinate conditions for the short ISI. Bottom: Accuracy for Dominant and Subordinate conditions for the long ISI.

PSS Groups & Reading Experience Control Measure

The Author Recognition Test (Stanovich & West, 1989) score differences computed in order to ensure that differences in sensitivity to the subordinate word meaning observed was not driven by reading experience. There were no differences between Choosers and Balanced participants (t(94) = -0.44, p = 0.66), nor between Avoiders and Balanced (t(88) = -0.06, p = 0.95) that could account for the effect observed in the LME model results reported previously.

Computational Model Results

To generate predictions, the model was run for 1,000 times under the different values of α_C and α_A associated with Choosers, Avoiders, or Balanced individuals. The model predicts that, under short ISI, all three groups should perform at chance for the subordinate meaning, with no difference in performance. Under long ISI, however, the model predicts that Avoiders and Choosers should have greater than chance performance for the subordinate condition (62% and 63%, respectively), while Balanced individuals should still perform essentially at chance (55% accuracy). Note that these predictions are parameter-free, and come remarkably close to the actual results of our experiment. In the model, this asymmetry in behavior is due to the fact that different initial learning rates α_C and α_A result in biased estimates of success when selecting dominant and subordinate meanings, respectively. In particular, the model predicts that Choosers would tend to overestimate the probability of the subordinate meaning, while Avoiders would tend to underestimate the probability of the dominant meaning, with both cases resulting in a tendency to favor the selection of the subordinate meaning. Under balanced learning rates, instead, the model correctly estimates the rarity of the subordinate meaning and tends to select it significantly less often.

Discussion

The current project explored the hypothesis that human linguistic ability, and specifically semantic processing, is dependent on core basal ganglia RL mechanisms. The results provide evidence for the proposed hypothesis, and more specifically, suggest that individual differences in learning from positive or negative feedback are predictive of automatic semantic ambiguity resolution in context-free lexical ambiguity priming paradigm. Specifically, task performance was in line with behavioral predictions by the computational cognitive model, which predicted that action-selection in the basal ganglia for dominant and subordinate meanings would happen in line with an individual's estimate of success for choosing either meaning. To illustrate this, when a Balanced participant reads the word "bank" they co-activate the associated "money" and "river" meanings. Selection happens in line with their learned estimate that"river" rarely occurs following "bank," and this subordinate meaning is unavailable for the semantic relatedness judgment, resulting in poor task performance (for this condition, only). Thus, the signal generated by the basal ganglia during semantic selection can be seen as reflecting an individual's estimate of that word-meaning co-occurrence, or in other words, the individual's representation of relative frequency of a meaning associated with a lexical form.

Furthermore, findings from this investigation are compatible with the widely accepted view that prefrontal cortex (PFC) regions and specifically the left inferior frontal gyrus (LIFG), are involved in semantic selection processes (Vitello & Rodd, 2015). While the LIFG may very well be the primary driver of semantic selection, it is known to make use of biasing signals to rule out multiple competing representations (Schnur et al., 2009). This biasing signal is posited

to stem from the basal ganglia, as research on the functional and anatomical properties of the PFC-basal ganglia network has shown that two of the five main cortico-striatal-thalamo-cortical loops project directly to lateral prefrontal regions, including dorsolateral PFC and lateral orbitofrontal cortex (Alexander, Crutcher, & DeLong, 1991). Thus, the basal ganglia posses the functional, anatomical, and computational properties necessary to provide biasing signals to LIFG during semantic ambiguity resolution.

Interestingly, these results reproduce and extend, by artificially segmenting a continuum of basal ganglia-mediated Choose and Avoid learning in a healthy population, findings observed in clinical groups. As mentioned previously herein, PD patients show abnormal lexical priming effects, with disrupted automatic semantic ambiguity resolution and sustained multiple competing representations. Additionally, literature focusing on the cognitive effects of Huntingtons Disease (HD) a basal ganglia dysfunction characterized by hyper-dopaminergic signaling and thus a hyper-active direct pathway, reveals that HD patients also have an increased susceptibility to semantic priming (Randolph, 1991). Taken together, these findings highlight the importance of a competitive dual-path RL system that gives rise to learning from both positive and negative feedback.

Possible alternative explanations exist for the current set of experimental results. Many theoretical and computational models of basal ganglia functioning focus on its role as "gates" that modulate prefrontal cortex functioning through selection (or Choose) and inhibition (or Avoid) mechanisms. Thus, under this framework, we would anticipate to find that Choosers would manage conflict in multiple competing representations by selecting the relevant or dominant word meaning, while Avoiders would inhibit the subordinate meaning. This is, however, not what is observed in the behavioral results, where both Choosers and Avoiders show identical performance in the subordinate condition after the long delay. This pattern of results is most compatible with a RL explanation of statistical learning, where a one-path mechanism (akin to traditional TD-learning) would over-estimate the utility of the lower frequency meaning. In other words, it is possible that Choosers are overly sensitive to low frequency reward probabilities, while Avoiders are less sensitive to high frequency reward probabilities. This results in a misrepresentation of the relative frequency effect observed between the dominant and subordinate word meanings.

This proposed role of the basal ganglia in RL through statistical mapping of the rich and dynamic linguistic environment, and engaging in live predictive processing may ultimately account for its involvement across multiple language processing modalities. In fact, a great deal of work exists that discusses evidence of basal ganglia involvement in language through the lens of a pacemaker-like, live, temporal processing machine that synchronizes internal states with external inputs (Kotz, Schwartze, & Schmidt-Kassow, 2009). While this research has focused mostly on morphosyntactic processing, its framework is both compatible with the one proposed herein and can be extended to multiple processing domains, including those beyond linguistic processing (e.g., non-linguistic cognitive functioning and motor processing). We consider these exciting areas for future research.

Acknowledgments

This research was supported by a National Science Foundation Graduate Research Fellowship (DGE-1256082) awarded to Jose M. Ceballos and by an award from the Office of Naval Research (ONRBAA13-003) to Chantel S. Prat and Andrea Stocco.

References

Alexander, G. E., Crutcher, M. D., & DeLong, M. R. (1991). Basal ganglia-thalamocortical circuits: parallel substrates for motor, oculomotor,"prefrontal" and "limbic" functions. In *Progress in brain research* (pp. 119–146). Elsevier.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). *An integrated theory of the mind* (Vol. 111) (No. 4).

Anderson, J. R., Fincham, J. M., Qin, Y., & Stocco, A. (2008). A central circuit of the mind. *Trends in Cognitive Sciences*, *12*(4), 136–143.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Booth, J. R., Wood, L., Lu, D., Houk, J. C., & Bitan, T. (2007). The role of the basal ganglia and cerebellum in language processing. *Brain research*, *1133*, 136–144.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological review*, *82*(6), 407.

Copland, D. A., Chenery, H. J., & Murdoch, B. E. (2001, August). Discourse Priming of Homophones in Individuals With Dominant Nonthalamic Subcortical Lesions, Cortical Lesions and Parkinsons Disease. *Journal of Clinical and Experimental Neuropsychology*, *23*(4), 538–556.

Crosson, B. (1985). Subcortical functions in language: a working model. *Brain and language*, *25*(2), 257–292.

Frank, M. J., & Hutchison, K. (2009). Genetic contributions to avoidance-based decisions: striatal d2 receptor polymorphisms. *Neuroscience*, *164*(1), 131–140.

Frank, M. J., Moustafa, A. A., Haughey, H. M., Curran, T., & Hutchison, K. E. (2007). Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. *Proceedings of the National Academy of Sciences*, *104*(41), 16311–16316.

Frank, M. J., Seeberger, L. C., & O'reilly, R. C. (2004). By carrot or by stick: cognitive reinforcement learning in parkinsonism. *Science*, *306*(5703), 1940–1943.

Graybiel, A. M. (1995). Building action repertoires: memory and learning functions of the basal ganglia. *Current opinion in neurobiology*, *5*(6), 733–741.

Hazy, T. E., Frank, M. J., & O'Reilly, R. C. (2007). Towards an executive without a homunculus: computational models

of the prefrontal cortex/basal ganglia system. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *362*(1485), 1601–1613.

Ketteler, D., Kastrau, F., Vohn, R., & Huber, W. (2008, February). The subcortical role of language processing. High level linguistic features such as ambiguity-resolution and the human brain; an fMRI study. *NeuroImage*, *39*(4), 2002–2009.

Kotz, S. A., Schwartze, M., & Schmidt-Kassow, M. (2009, September). Non-motor basal ganglia functions: A review and proposal for a model of sensory predictability in auditory language perception. *Cortex*, *45*(8), 982–990.

Kristensen, M., & Hansen, T. (2004). Statistical analyses of repeated measures in physiological research: a tutorial. *Advances in physiology education*, *28*(1), 2–14.

Lieberman, P. (2001). Human language and our reptilian brain: The subcortical bases of speech, syntax, and thought. *Perspectives in Biology and Medicine*, *44*(1), 32–51.

Mason, R. A., & Just, M. A. (2007, May). Lexical ambiguity in sentence comprehension. *Brain Research*, *1146*, 115–127.

Randolph, C. (1991). Implicit, explicit, and semantic memory functions in alzheimer's disease and huntington's disease. *Journal of Clinical and Experimental Neuropsychology*, *13*(4), 479–494.

Schnur, T. T., Schwartz, M. F., Kimberg, D. Y., Hirshorn, E., Coslett, H. B., & Thompson-Schill, S. L. (2009). Localizing interference during naming: Convergent neuroimaging and neuropsychological evidence for the function of broca's area. *Proceedings of the National Academy of Sciences*, *106*(1), 322–327.

Seo, R., Stocco, A., & Prat, C. S. (2018). The bilingual language network: Differential involvement of anterior cingulate, basal ganglia and prefrontal cortex in preparation, monitoring, and execution. *NeuroImage*, *174*, 44–56.

Stanovich, K. E., & West, R. F. (1989). Exposure to print and orthographic processing. *Reading Research Quarterly*, 402–433.

Stocco, A. (2018). A biologically plausible action selection system for cognitive architectures: Implications of basal ganglia anatomy for learning and decision-making models. *Cognitive science*, *42*(2), 457–490.

Stocco, A., Lebiere, C., & Anderson, J. R. (2010). Conditional routing of information to the cortex: A model of the basal ganglias role in cognitive coordination. *Psychological review*, *117*(2), 541.

Stocco, A., Murray, N. L., Yamasaki, B. L., Renno, T. J., Nguyen, J., & Prat, C. S. (2017). Individual differences in the simon effect are underpinned by differences in the competitive dynamics in the basal ganglia: An experimental verification and a computational model. *Cognition*, *164*, 31–45.

Tremblay, P., & Dick, A. S. (2016). Broca and wernicke are dead, or moving past the classic model of language neurobiology. *Brain and language*, *162*, 60–71.

Twilley, L. C., Dixon, P., Taylor, D., & Clark, K. (1994). University of alberta norms of relative meaning frequency for 566 homographs. *Memory & Cognition*, *22*(1), 111–126.

Vitello, S., & Rodd, J. M. (2015). Resolving semantic ambiguities in sentences: Cognitive processes and brain mechanisms. *Language and Linguistics Compass*, *9*(10), 391–405.

Environmental effects on parental gesture and infant word learning

Rachael W. Cheung (r.w.cheung@lancaster.ac.uk)
Calum Hartley (c.hartley@lancaster.ac.uk)
Department of Psychology, Lancaster University, UK

Padraic Monaghan (p.j.monaghan@uva.nl)
Department of English Language and Culture, University of Amsterdam, The Netherlands
Department of Psychology, Lancaster University, UK

Abstract

How infants determine correct word-referent pairings within complex environments is not yet fully understood. The combination of multiple cues, including gestures, may guide learning as part of a communicative exchange between parent and child. Gesture use and word learning are interlinked, with early child gesture predicting later vocabulary size, and parental gesture predicting child gesture. However, the extent to which parents alter gesture cues during word learning according to referential uncertainty is not known. In this study, we manipulated the number of potential referents across conditions during a word learning task with 18–24-month-olds, and explored how changes in parental gesture use translated into infant word learning. We demonstrate that parents alter their gesture use according to the presence, but not the degree, of referential uncertainty. We further demonstrate that a degree of variability in the number of potential referents appears to benefit word learning.

Keywords: word learning; gesture; vocabulary development; parent-infant interaction

Introduction

Between 18–24 months of age, children's expressive vocabulary rapidly increases from approximately 90 to 320 words (Fenson et al., 1994). Children learn language in busy and variable environments containing multiple possible referents, but how they determine what the intended referent for a given word is remains under-investigated. Finding the correct word-referent pairing is a problem of substantial difficulty, as described by the well-known 'gavagai' problem (Markman, 1989; Quine, 1960), where a second language learner cannot know whether an unknown utterance – 'gavagai' – refers to a rabbit present in the scene, the rabbit bouncing, the rabbit's colour, or a range of other potential meanings. Infants face the same problem as the second language learner, with a further disadvantage – the lack of a first language to base their learning upon.

Recent attention has turned towards examining the multiple potential cues present in language learning environments that might help children to delineate referents of unfamiliar words (Monaghan, 2017). One of the earliest sources of information to support word-referent mappings is provided before children are able to speak: gestures in parent-child interactions. Within these interactions, gesture appears to be facilitative of word learning. For example, spontaneous pointing by the infant during a gaze-following task at 10–11 months predicted vocabulary growth at 24 months (Brooks & Meltzoff, 2008), and Fenson et al. (1994) found an increase in infants' gesture use between 8–16 months correlated with word comprehension. Parent and infant gesture use also appears to be reciprocal in nature. Rowe, Özçalışkan and Goldin-Meadow (2008) observed gesture use in parent-child dyads at four-month intervals between the ages of 14–34 months, then administered a vocabulary test at 42 months. They found that child gesture use at 14 months predicted vocabulary size at 42 months, and that parent gesture use predicted child gesture use at 14 months. Between 22–34 months, they found that child gesture use (number of gestures with or without speech) mirrored parent gesture use. Infant gesture therefore appears to predict language development and appears to be related to parental gesture use.

The nature of this relationship seems rooted in the informative role of gestures in word learning during active communication between parent and child, with gesture adding significant value to information exchange. The use of gesture as a response in perspective-taking tasks has demonstrated that infants use and adjust their gestures according to parent knowledge states. In a similar way to how older children (from 3.5-years) adjust their speech responses to actively incorporate a communicative partner's perspective (Nadig & Sevidy, 2002; Nilsen & Graham, 2009), when parents do not have the same information as infants, infants are more likely to gesture to support mutual understanding (O'Neill, 1996). Gesture thus may play a vital role in aiding effective communication when verbal ability is still being established. Infant gesture may also serve an interrogative function by acting as a signal to gain critical information from parents about a specific object (Iverson & Goldin-Meadow, 2005; Southgate, van Maanen & Csibra, 2007). Given that gestures are a vital means through which infants interact with and learn about their surrounding context, how might this assist children in navigating the complex environment surrounding word learning?

On the other side of this communicative partnership, gesture use by caregivers may provide valuable information about intended referents during rapid vocabulary development. In particular, parental gestures such as pointing serve as a useful tool for identifying a referent when learning word-referent pairs. Iverson et al. (1999) reported parental pointing during 15% of exchanges related to word learning. Furthermore, the quality of parental gesture appears to have

an effect on word learning. Cartmill et al. (2013) assessed parental input quality during parent-child interactions at 14–18 months by asking adult participants to guess words from muted observational videos. This provided a measure of input quality by indicating how informative non-verbal and gestural communication was in determining word meaning. When correlated with child vocabulary at 53 months, children whose parents produced higher quality input had higher receptive vocabulary.

Thus, both the frequency and quality of parental gesture are related to infant word learning and provide valuable cues that enable the child to predict the referents for their growing vocabulary. However, it is not yet known how adaptive parental gesture is to the information present in the environment, or what kind of gestures are most helpful to infants under conditions of differing referential uncertainty.

In verbal communication, we know that speakers adjust their phonology, prosody, word selection, and syntax in accordance with the context of communication and the listener's perspective (Bannard, Rosner, & Mathews, 2017; Brown-Schmidt & Duff, 2016; Gorman et al., 2013). We also know that children can adapt their gesture and speech to accommodate the perspective of adults (Bahtiyar & Küntay, 2009). Parents also adapt their spoken labelling behaviour according to infant familiarity with objects (Cleave & Kay-Raining Bird, 2006; Masur, 1997) and how conventional the label is (where conventionality refers to there being a culturally agreed referent for a specific word; Luce & Callanan, 2010). However, we do not know the extent to which parents adjust their gesture use contingently based on referential uncertainty during infant word learning.

In this study, we address this issue, testing whether parents would offer a higher number of gestural cues when a target item was amongst more, rather than fewer, distractor objects. Furthermore, we measured whether the type of gestures that occur, and their correspondence with speech, affected children's learning of novel words. Greater referential uncertainty, as determined by a higher number of potential referents for a novel label, has led to less reliable and slower word learning in previous studies (Smith, Smith, & Blythe, 2011; Trueswell et al., 2013). Consequently, we would expect parental gesture to play a stronger role in delineating referents when there is a higher degree of referential uncertainty. In a word learning task, we manipulated the number of potential referents for a novel word between one, two, and six referents. We hypothesised that parental gestural cues would increase with the frequency of potential referents from the one- to the six-referent condition, particularly for deictic cues (gestures directing attention to a specific object). We predicted the same pattern for the co-occurrence of speech with gesture, in particular for speech that used the target label. We then examined whether these cues translated into infant word learning accuracy by testing infants on their knowledge of the novel label. We hypothesised that infants of parents who offered more gestural cues would show higher word learning accuracy.

Method

Participants

Fifty-three monolingual English infants aged between 18–24 months-old (M = 20.9 months, SD = 1.7, 25 female) were recruited from a database of families who had registered interest in study participation at Lancaster University Babylab. Infants were from middle-class families (determined via parental education level). During training, six parent-infant dyads were excluded due to infant fussiness. Twenty-seven infants (M = 20.8 months, SD = 1.6, 14 female) also completed the testing phase, with the remaining sample excluded due to infant fussiness (n = 4) or incomplete trials (n = 16; less than 5 of 6 test trials).

Materials

Nine novel objects were used as referents for the novel words. Each novel object was a different colour and shape. Three novel words, selected from the NOUN database (Horst & Hout, 2016), were used as labels (*noop*, *darg*, and *terb*). Three objects were chosen as targets randomly for each participant, with all other objects serving as foils, and each novel label was randomly paired with each target per participant. Stimuli position, target, and condition order during training and testing were counterbalanced across participants using a Latin square. Parents also completed the UK-CDI (Alcock, Meints, & Rowland, 2017), a measure of receptive, expressive, and gesture (communicative and symbolic) vocabulary. Communicative gestures are declarative (deictic and imperative gestures) and symbolic gestures form a larger subset of actions with objects, games, and pretend play (representative gestures).

Procedure

Infants were seated on their parent's lap and viewed stimuli from 70 cm away. Each group of stimuli was presented for 30 seconds, with a moveable opaque screen shielding objects from view in-between trials. Parents were asked to imagine they were teaching real words for real objects. Familiarisation with the objects took place outside of the experimental room with the parent only. The labels and a three-word object description were visible to the parent throughout training to eliminate the need for parents to remember the novel label and paired target.

Participants began with one warm-up trial, where the experimenter placed a ball as a familiar object on a tray and instructed the parent to teach the infant the word as if it were novel. The aim was to familiarise parents with the procedure without increasing task demands. Parents then proceeded to the training phase, where they taught infants novel label-referent pairs with unfamiliar objects as stimuli. In the one-referent condition, only the target was presented; in the two-referent condition, one target and one foil were presented; and in the six-referent condition, one target and five foils were presented (see Figure 1). Each participant received each of the three conditions once.

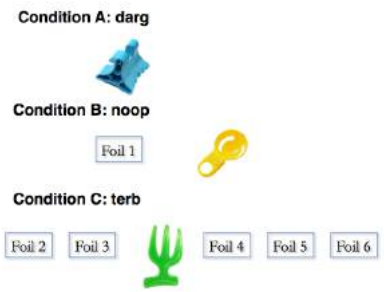


Figure 1: Training trials example.

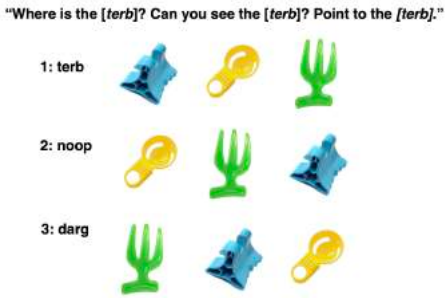


Figure 2: Testing trials example.

After completing all three training conditions, participants were then administered six testing trials, with each target word tested twice (see Figure 2). At the start of each trial, the infant was asked by the experimenter "Where is the [*target*]? Can you see the [*target*]? Point to the [*target*]." The trial ended when the infant made a response or the prompt had been repeated twice without a response.

Coding

All sessions were video-recorded and then coded for gestures and speech with gesture per utterance according to Rowe et al.'s (2008) coding scheme. A second coder coded 20% of the videos with an overall inter-rater reliability $\kappa = 0.78$ for gesture ($N = 284$) and $\kappa = 0.86$ for speech with gesture ($N = 160$).

Gesture types were split into three main groups (Rowe et al., 2008): *representational* gestures, indicating properties of the target referent such as size, shape, or function; *deictic* gestures, singling out the target referent by pointing with the arm and index finger extended or with the arm extended and the palm exposed and *other* gestures, which included all gestures not aimed towards the referent (those aimed at foils and related to caregiving interactions such as hugging).

The co-occurrence of speech that indicated properties of the target referent (e.g. size, shape, or function) with gesture was coded as *supplementary*. The co-occurrence of speech that singled out the target referent with gesture was coded as *complementary*. The frequency of referent label use was also recorded.

Results

A series of linear mixed effects models (lmer; lme4 in R, v3.4.1, 2017) were used to predict parents' use of gestures during training (gesture subtypes and co-occurring speech with gesture subtypes were dependent variables). These models were built up progressively with the addition of fixed effects of condition and child vocabulary (scores of communicative gesture, symbolic gesture and expressive subscales of CDI), comparing each model to a null model or previous best-fitting model using log-likelihood comparison after the addition of each new term (Barr et al., 2013). Random effects of subject and infant age were included in each analysis.

Environmental uncertainty effects on parental gesture use

The linear mixed effects models demonstrated a significant effect of condition on overall gesture count ($\chi^2(2) = 11.73$, $p = .003$). Consistent with our hypothesis, parents gave more gestural cues when they were faced with a higher number of potential referents (see Figure 3), with a significant difference between one- and two-referent conditions ($t(94) = 2.12$, $p = .037$), and one- and six-referent conditions ($t(94)=3.51$, $p = .001$), but not two- and six-referent conditions ($t(94) = 1.39$, $p = .167$). The addition of child vocabulary measures did not improve model fit (communicative gesture: $\chi^2(1) = 0.38$, $p = .539$; symbolic gesture: $\chi^2(1) = 0.28$, $p = .598$; expressive: $\chi^2(1) = 0.34$, $p = .560$). No significant interactions between fixed effects were found.

The relation between gestural cues and number of referents was particularly notable in deictic gestures, in-keeping with our hypothesis. There was a significant effect of condition on deictic gesture number ($\chi^2(2) = 8.35$, $p = .015$, see Figure 3), with significant differences between one- and two-referent conditions ($t(94) = 2.21$, $p = .030$), and one- and six-referent conditions ($t(94) = 2.80$, $p = .006$), but not two- to six-referents ($t(94) = 0.60$, $p = .553$). Adding child vocabulary did not improve model fit (communicative gesture: $\chi^2(1) = 0.001$, $p = .973$; symbolic gesture: $\chi^2(1) = 0.05$, $p = .832$; expressive: $\chi^2(1) = 0.01$, $p = .917$). No interactions between fixed effects were found.

For representational and other gestures, there were no significant effects or interactions.

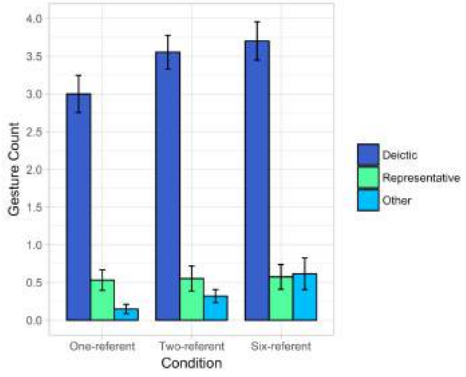


Figure 3: Mean count and standard error of gesture type generated by parents per condition.

Environmental uncertainty effects on co-occurring parental speech and gesture

When testing the co-occurrence of complementary speech with gesture, linear mixed effects models showed significant main effects of condition and child symbolic gesture vocabulary ($\chi^2(3) = 8.28$, $p = .041$; see Figure 4). There was a significant increase from one to two referents ($t(80) = 2.57$, $p = .012$), but no significant difference between one and six referents ($t(80) = 1.68$, $p = .096$) or two and six referents ($t(80) = -0.89$, $p = .376$). There were no other significant main effects of child vocabulary measures and no significant interactions between fixed effects.

When testing the co-occurrence of supplementary speech with gesture, we found condition was not significant as a main effect alone. There was a significant interaction between condition and child expressive vocabulary ($\chi^2(5) = 17.96$, $p = .003$), which showed that children with larger vocabularies were offered more information in the one- and two-referent conditions, but less in the six-referent condition, than children with smaller vocabularies. There were no other significant main effects or interactions between fixed effects.

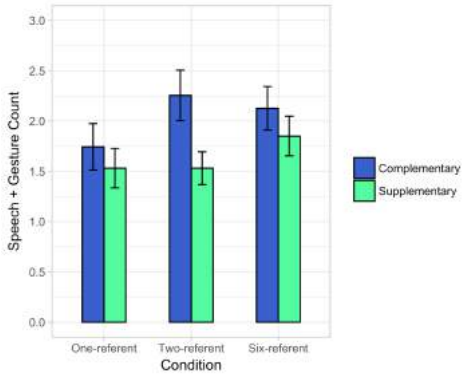


Figure 4: Mean count and standard error of types of speech with gesture generated by parents per condition.

Effect of parental gesture use on word learning

A series of binomial general linear mixed effects models (glmer; lme4 in R, v3.4.1, 2017) were used to predict accuracy. These models were built up progressively with the addition of fixed effects of condition and child vocabulary (scores of communicative gesture, symbolic gesture and expressive subscales of CDI), comparing each model to a null model or previous best-fitting model using log-likelihood comparison after the addition of each new term (Barr et al., 2013). Random effects of subject and infant age were included in each analysis.

Analysis using general linear mixed effects models revealed the addition of condition improved model fit ($\chi^2(2) = 6.08$, $p = .048$; see Figure 5), indicating a significant increase of accuracy from one to two referents ($\beta = 0.91$, $z = 2.19$, $p = .028$) and from one to six referents ($\beta = 0.86$, $z = 2.02$, $p = .044$), but no significant increase in accuracy from two to six referents ($\beta = -0.05$, $z = -0.13$, $p = .893$). However, this varied by parent, as the addition of a slope of condition per parent as a random effect removed the significant main effect of condition ($\chi^2(2) = 1.8$, $p = .406$).

Given that complementary speech with gesture was highest in the two-referent condition during training and accuracy was highest in this condition during testing (see Figure 5), we postulated that there might be some relationship between the two. However, the inclusion of total gestures, gesture subtype, and types of co-occurrence of speech with gesture did not improve model fit, suggesting there was no significant prediction of accuracy when these effects were taken into account. This did not support our hypothesis that increased parental gesture use during training would predict increased accuracy of infant word learning. A separate model examining these training response variables without an effect of condition did not demonstrate any significant improvement of model fit, suggesting any significant difference in accuracy was the result of differences in condition alone, without any demonstrable effects of training response.

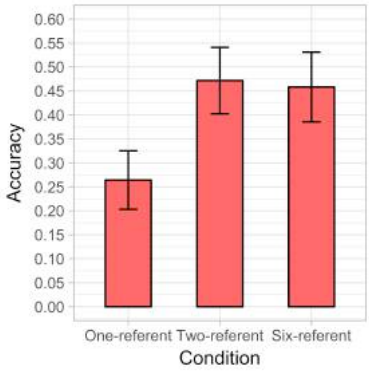


Figure 5: Mean infant word learning accuracy and standard error per condition.

Discussion

By varying referential uncertainty, we explored how parental gesture might aid infants in learning correct word-referent pairings within complex environments. Our training results demonstrated that parent gesture can be manipulated by altering the immediate environment around infant word learning. This was particularly notable in deictic gesture use. The results showed that parents use deictic gestures most in the presence of referential uncertainty, as parents gestured most in the six- and two-referent conditions compared to the one-referent condition. Deictic gestures have previously been found to be highly informative when determining word-referent pairs (Cartmill et al., 2013). Children also have been found to follow the direction of deictic gestural cues over linguistic cues in referent-selecting tasks (Grassmann & Tomasello, 2010). Thus, although it is possible that our findings related to deictic gesture use were influenced by task-demands (requesting parents to teach specific novel words for objects, and objects being out of reach), they are in line with other research that points towards the usefulness of

deictic gestures when delineating referents in naturalistic and laboratory settings.

The mechanism by which gesture adds information to speech may be a reduction of cognitive load for the infant, providing a visual component to learning resources alongside the verbal component (Goldin-Meadow, 2000; McGregor et al., 2009). This has been found particularly useful in situations of high task demands (McNeil, Alibali, & Evans, 2000) – consistent with parents using the most gestures in the six-referent condition. However, there was no significant difference in gesture use between the two-referent and six-referent conditions, which did not support our hypothesis that the higher the number of potential referents, the higher the number of parental gestures to assist the child in coping with referential uncertainty. One might also expect a higher number of potential referents to confer a higher task-demand, and thus perhaps a need for a greater reduction of cognitive load due to an increased amount of distracting information.

This result may demonstrate that the more important factor in referent-identification is whether there is referential uncertainty or not, rather than the degree of uncertainty. It is possible that the additional information conveyed in gesture is not as valuable in reducing cognitive load when there is more than one choice to be had. This interpretation is consistent with children's actual learning of novel words – infants demonstrated the highest accuracy in the two-referent condition, and performed marginally worse in the six-referent condition which had the highest frequency of parent gestural information and referent label use. Infants performed worst in the one-referent condition. This might be unexpected given the lack of referential uncertainty, although there was also the least amount of information available (provided in speech and gestural cues).

These learning results suggest that some referential uncertainty might actually be beneficial for learning, and that perhaps too much uncertainty begins to remove that benefit. In Monaghan's (2017) computational study of multiple cue integration in word learning, the model predicted that a small amount of variability in the cues available in the word learning environment yielded superior learning in comparison to conditions where cues were perfectly reliable and invariable. But when this variability became substantial, learning of novel words began to decline. In Monaghan et al. (2017), this prediction was supported in a study of adults learning novel word-referent mappings from multiple cues: variability was helpful. However, in these studies, the referential uncertainty was kept constant – in all cases, there were two possible referents from which to select. In the current study, we further show that a small degree of variability in referential uncertainty led to the best novel word learning.

The presence of two competing alternatives in the environment ties in with studies of children's application of mutual exclusivity (Markman & Wachtel, 1988). In these studies, children are shown to actively use a general principle of 'this, not that' to map unknown words to unknown objects in relation to known objects. Although this mechanism works primarily by prior knowledge, it is possible that having one choice enables some sorting of the available referents that makes word learning more efficient.

However, our results did not show a significant direct effect of parents' behaviour in driving children's word learning performance – the amount of gestural information with and without speech during training was not predictive of more accurate infant word learning as we had predicted. Any effect of condition on accuracy also disappeared with the addition of a random slope for condition per parent, suggesting that there was a high degree of variability in how parents used gestures across the conditions.

The lack of an effect may be partly due to limitations in our sample. All parents were of mid-socioeconomic status (SES), recruited from a database of families who had actively signed up to take part in child development studies. Families from mid to high SES backgrounds are known to use gesture more (Rowe & Goldin-Meadow, 2009). Kirk et al. (2013) suggest that the added benefit of gesture may be most prevalent in cases where there is general diminished parental input, providing a compensatory effect, and in mid to high SES families, parental input is less likely to be reduced. Gains in child vocabulary following training that involved increased gesture use have previously been found primarily in low SES environments (Hirsh-Pasek et al., 2015). Although parents in our study did gesture more with increased referential uncertainty, it is possible that any added benefit of gesture in this sample reached something of a ceiling effect when it came to word learning – infants were already subject to a level of parental input that meant gesture did not add to their learning.

Finally, given prior evidence that children's vocabulary and gesture use are positively related, and child gesture is linked to parental gesture (Rowe et al., 2008), child gesture vocabulary might be expected to have some effect on parental gesture use during training. However, this was not the case in our study. Our models of gesture alone did not identify an effect of child expressive and gesture vocabulary. We did find that these effects played a role in the amount of speech with gesture. We found that child gesture and expressive vocabulary were significant effects when referential uncertainty was increased. This may indicate that child gesture and expressive vocabulary are related to parental gestures co-occurring with speech, instead of parental gesture in isolation. This aligns with the idea of gesture playing a supplementary role to speech, rather than one supplanting the other (O'Neill, 1996; Iverson & Goldin-Meadow, 2005).

In summary, we found that referential uncertainty affected parents' gestures. Parents were affected by the number of potential referents in the environment, and adapted their gestures, and co-occurrences of gestures with naming of the target object, offering more cues when the child's environment became more complex. However, parental gesture use was only affected by whether there was referential uncertainty or not, rather than the degree of referential uncertainty. In terms of children's accuracy when testing their knowledge of novel labels, referential

uncertainty was again found to affect learning, and actually promoted it. The results add to a broad picture of communicative exchange where interlocutors are sensitive to the context and informational requirements of the situation, and also to growing evidence that variability, within speech and within the environment, is beneficial for learning.

Acknowledgments

This work was supported by the Leverhulme Trust (RWC, Leverhulme Trust Doctoral Scholar), and by the International Centre for Language and Communicative Development (LuCiD) at Lancaster University funded by the Economic and Social Research Council (PM and CH) [ES/L008955/1]. Thanks to Delyth Piper for aiding with video-coding.

References

Alcock, K.J., Meints, K., & Rowland, C.F. (2017) UK-CDI Words and Gestures – Preliminary norms and manual.

Bahtiyar, S., & Küntay, A. C. (2009). Integration of communicative partner's visual perspective in patterns of referential requests. *Journal of Child Language, 36(3)*, 529–555.

Bannard, C., Rosner, M., & Matthews, D. (2017). What's worth talking about? Information theory reveals how children balance informativeness and ease of production. *Psychological Science, 28(7),* 1–13.

Barr, D.J., Levy, R., Scheepers, C., & Tily, H.J. (2013). Random effects structure for confirmatory hypothesis testing: keep it maximal. *Journal of Memory and Language*, *68(3),* 255-278.

Brooks, R., & Meltzoff, A.N. (2008). Infant gaze following and pointing predict accelerated vocabulary growth through two years of age: a longitudinal, growth curve modeling study. *Journal of Child Language*, *35(1)*, 207-220.

Brown-Schmidt, S., & Duff, M. C. (2016). Memory and common ground processes in language use. *Topics in Cognitive Science, 8(4),* 722–736.

Cartmill, E.A., Armstrong-III, B.F., Gleitman, L.R., Goldin-Meadow, S., Medina, T.N., & Trueswell, J.C. (2013). Quality of early parent input predicts child vocabulary 3 years later. *PNAS*, *110(28),* 11278-11283.

Cleave, P.L. & Kay-Raining Bird, E. (2006) Effects of familiarity on mothers' talk about nouns and verbs. *Journal of Child Language*, 33, 661–676.

Fenson, L., Dale, P.S., Reznick, J.S., Bates, E., Thal, D.J., & Pethick, S.J. (1994) Variability in early communicative development. *Monographs of the Society for Research in Child Development, 59(5),* 1-185.

Goldin-Meadow, S. (2000). Beyond words: the importance of gesture to researchers and learners. *Child Development, 71(1),* 231-239.

Gorman, K. S., Gegg-Harrison, W., Marsh, C. R., & Tanenhaus, M. K. (2013). What's learned together stays together: Speakers' choice of referring expression reflects shared experience. *Journal of Experimental Psychology. Learning, Memory and Cognition, 39*, 843–853.

Grassman, S., & Tomasello, M. (2010) Young children follow pointing over words interpreting acts of reference. *Developmental Science, 13(1),* 252-263.

Hirsh-Pasek, K., Adamson, L.B., Bakeman, R., Owen, M.T., Golinkoff, R.M., Pace, A., … Suma, K. (2015) The contribution of early communication quality to low-income children's language success. *Psychological Science*, *26(7),* 1071-1083.

Horst, J.S., & Hout, M.C. The Novel Object and Unusual Name (NOUN) Database: a collection of novel images for use in experimental research. (2016) *Behavior Research Methods, 48(4),* 1393-1409.

Iverson, J.M., Capirici, O., Longobardi, E., & Caselli, M.C. Gesturing in mother-child interactions. (1999) *Cognitive Development*, *14*, 57-75.

Iverson, J.M., & Goldin-Meadow, S. (2005) Gesture paves the way for language development. *Psychological Science*, *16(5),* 367-371.

Kirk, E., Howlett, N., Pine, K.J., & Fletcher, B. (2013). To sign or not to sign? The impact of encouraging infants to gesture on infant language and maternal mind-mindedness. *Child Development*, *84(2),* 574-590.

Luce, M.R., & Callanan, M.A. Parents' object labeling: possible links to conventionality of word meaning? *First Language, 30(3),* 270–286.

Markman, E.M. (1989). Categorization and naming in children: problems of induction. Cambridge, MA: MIT Press.

Markman, E.M., & Wachtel, G.F. (1988). Children's use of mutual exclusivity to constrain the meaning of words. *Cognitive Psychology*, *20(2),* 121-157.

Masur, E.F. (1997). Maternal labelling of novel and familiar objects: implications for children's development of lexical constraints. *Journal of Child Language*, 24(2), 427-439.

McGregor, K.K., Rohlfing, K.J., Bean, A., & Marschner, E. (2009). Gesture as a support for word learning: the case of under. *Journal of Child Language*, *36(4),* 807-828.

McNeil, N.M., Alibali, M.W., & Evans, J.L. (2000) The role of gesture in children's comprehension of spoken language: now they need it, now they don't. *Journal of Nonverbal Behavior*, *24(2),* 131-150.

Monaghan, P. (2017). Canalization of language structure from environmental constraints: a computational model of word learning from multiple cues. *Topics in Cognitive Science, 9*, 21-34.

Monaghan, P., Brand., J., Frost, R.L.A., & Taylor, G. (2017). Multiple variable cues in the environment promote accuracy and robust word learning. In G. Gunzelman, A. Howes, T. Tenbrink, & E. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society (CogSci 2017, pp. 817-822).*

Nadig, A.S., & Sedivy, J.C.(2002). Evidence of perspective-taking constraints in children's online reference resolution. *Psychological Science*, *13(4),* 329-336.

Nilsen, E.S., & Graham, S.A. (2009). The relations between childrens communicative perspective-taking and executive functioning. *Cognitive Psychology*, *58(2),* 220-249.

O'Neill, D.K. (1996). Two-year-old's sensitivity to a parent's knowledge state when making requests. *Child Development*, *67(2),* 659-677.

Quine, W.V.O. (1960). Word and Object. Cambridge, MA: MIT Press.

Rowe, M.L., & Goldin-Meadow, S. (2009) Differences in early gesture explain SES disparities in child vocabulary size at school entry. *Science*, *323(5916),* 951-953

Rowe, M.L., Özçalışkan, S., & Goldin-Meadow, S. (2008) Learning words by hand: gesture's role in predicting vocabulary development. *First Language*, *28(2*), 182-199.

Smith, K., Smith, A.D.M., & Blythe, R.A. (2011) Cross-situational learning: an experimental study of word-learning mechanisms. *Cognitive Science*, *35(3),* 480-498.

Southgate, V., Maanen, C.V., & Gergley, C. (2007). Infant pointing: communication to cooperate or communication to learn? *Child Development*, *78(3),* 735-740.

Trueswell, J., Medina, T.N., Hafri, A., Gleitman, L.R. (2013) Propose but verify: fast mapping meets cross-situational word learning, *Cognitive Psychology, 66(1),* 126-156.

Task Goals Structure Conceptual Acquisition

Seth Chin-Parker (chinparker@denison.edu)
Eric Brown (brown_e6@denison.edu)
Department of Psychology, Denison University
Granville, OH 43023 USA

Abstract

The purpose of this study is to explore the role goals play in concept acquisition. Goals motivate and shape our interactions with items, so it stands to reason that they also impact the learning that occurs as a result of those interactions. There is abundant evidence that goals orient us to particular information about the items we encounter. A more speculative claim is that goals play a more integral role in the acquired concept in that they also help to structure and cohere the acquired conceptual knowledge. Using a novel concept learning paradigm, we examined participant knowledge of attributes of the items they interacted with in an experimental task. We found evidence that the interaction of the goal with the learning situation impacted the centrality of the attribute information within their conceptual knowledge. These results support the idea that conceptual knowledge is organized in terms of goals active during learning.

Keywords: categories; concepts; goals; conceptual acquisition

Conceptual knowledge plays an important role in human cognition. Concepts help to shape our perceptions and predictions as we move through the world, and they allow access to information about entities that are not immediately present. All facets of cognitive science (e.g. philosophy, psychology, computer science, anthropology) have engaged with questions concerning conceptual knowledge, but important questions remain. This study focuses specifically on the ways that goal-directed interactions with instances from a novel category of items shape the organization and content of the acquired conceptual knowledge.

Within the psychological research, there has been on-going study of concept acquisition for over a century. Machery (2007) notes that despite significant shifts in the theoretical perspectives as to what constitutes conceptual knowledge, there has been a noticeable lack of variation in how psychologists operationalize the acquisition of a concept. Related concerns have been raised about category learning research in that the experimental paradigms are limited and potentially restrict our understanding of the processes involved in concept acquisition and how they affect acquired knowledge (Markman & Ross, 2003; Ross, Chin-Parker, & Diaz, 2005). There are also questions as to how well those experimental paradigms reflect concept acquisition as it happens in everyday life (Murphy, 2005).

In response to these concerns, there have been intentional and systematic attempts to broaden the range of learning tasks in the study of concept acquisition. The rationale is that examining learning that occurs in the course of different kinds of interactions provides a richer and more applicable sense of what conceptual acquisition is really like. Out of this, a line of research has emerged that examines how the goal of the learner affects concept acquisition. If an individual interacts with a set of items in the course of working towards a particular goal, the conceptual knowledge acquired from those interactions should be tuned such that it supports that goal (Chin-Parker & Birdwhistell, 2017; Jee & Wiley, 2007; Love, 2005). The idea that goals meaningfully intersect with conceptual acquisition has existed in the literature for several decades (see Barsalou, 1995), but only relatively recently has it been formalized within concept acquisition studies.

A basic assumption of this approach is that the goal points the individual towards features of the items that are goal relevant. Jee and Wiley (2007) and Chin-Parker and Birdwhistell (2017) have found strong evidence for this *goal-orientation hypothesis*. When participants with different goals interact with same set of items, the conceptual knowledge acquired privileges access to the attributes of the items that were critical to completion of the goal. This idea fits well with learning theories that incorporate some means for the learner to adapt to the differential importance or salience of individual attributes (e.g. Kruschke, 2003; Le Pelley, et al., 2016).

It has been suggested that goals also play a role in the representation of conceptual knowledge. Because a concept provides information as to why the instances of the corresponding category belong together, Jee and Wiley (2007) propose that the goal acts as a glue that coheres the members of the category. This idea reflects earlier work on ad hoc categories (e.g. Barsalou, 1983).

Chin-Parker and Birdwhistell (2017) provide an account that focuses on how a goal plays a role in organizing the attribute information represented within the concept. They note that in any situation there are many possibilities as to how an individual might interact with the entities that constitute that context. However, having a specific goal means that each possible interaction within that situation can be defined in terms of its goal-relevance. An interaction that moves the individual closer to, or further from, the goal can be considered goal relevant. An interaction that does not do so would not be goal relevant. If a goal-relevant interaction involves a particular facet of the item at hand, that aspect of the item becomes defined

in terms of how it relates to the goal – the way in which it facilitates (or hinders) movement towards the goal. In this view, the attributes themselves become available through the interplay between what constitutes the items and the goal-directed behaviors. It is important to remember that these interactions are situated – what constitutes a goal relevant attribute and how that attribute relates to the goal may vary across situations. Through the goal-directed interactions, a structure emerges that reflects the goal-relevance of the various components of the situation – attributes that are more critical to completing the goal become more central within the concept. For instance, if an attribute was differentiated in relation to the goal, e.g. it offered a goal-relevant decision point, then the information about its differentiation with regards to the goal would also be captured within the acquired conceptual knowledge. The *goal-framework hypothesis* proposes that the goal is more integrated into the conceptual knowledge than is implied by the goal-orientation hypothesis.

The purpose of this study is to assess the goal-framework hypothesis. Participants interacted with a set of novel items in order to complete a particular task. These items represented two different types, although the participants were not told this: They were not asked to learn about the items or to do any explicit category-based work, only to use them to complete the task at hand. In both conditions the items had two primary attributes that were goal-relevant – the participant had to attend to both attributes to complete the task. However, we manipulated the task so the relationship of one of the attributes to the goal differed across conditions. We modified whether the specific shape of that attribute was *relevant*, i.e. that shape required a decision to made about how to proceed in the task, or if it was *irrelevant*, the interaction with that attribute occurred without any consideration of its specific shape.

Because the task required the participants to differentiate between the two types of items, we expected them to naturally recognize the two categories of items. We expected all participants should be able to assess class membership of the items based on the primary attributes and to make judgments based on those categorizations. Critically, we expected that our manipulation of the relevance of the shape of one of the attributes would affect later category-based judgments indicating that it had impacted the centrality of that attribute within the conceptual knowledge.

Experiment

Methods

Participants and Design Sixty-seven participants were randomly assigned to two experimental conditions: 33 participants were assigned to the *interior shape relevant condition* (ISR condition), and 34 to the *interior shape irrelevant condition* (ISI condition). All participants used the same set of items during the initial task and completed the same transfer tasks. The presentation order of items during the initial and transfer tasks was randomized.

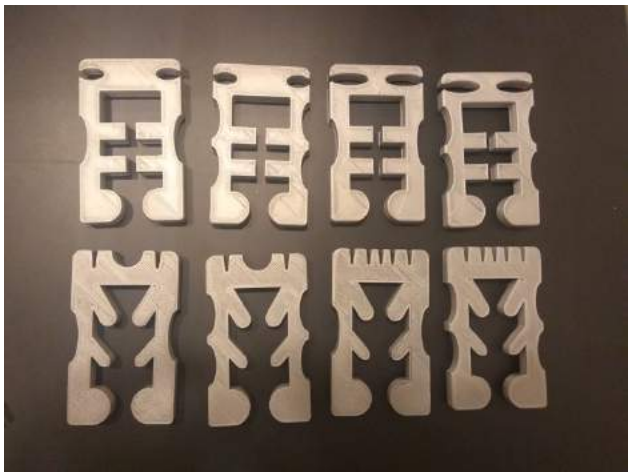


Figure 1: Keys used in the initial task. They were identified in the classification task as "alpha keys" (top row) and "zeta keys" (bottom row).

Materials and Procedure During the *initial task* of the experiment, all participants interacted with the same set of eight "keys" (see *Figure 1*). The keys had two primary attributes – the head shape and interior shape. These attributes co-varied so that there were two categories, or types, of keys defined by a particular combination of head and interior shape. The keys were made of ABS plastic and were created using a 3D printer. The keys were approximately 10 cm by 6 cm by 1 cm in size.

Participants used the keys to manipulate the task boards (see *Figure 2*). The boards were designed so the keys would be used as part of a two-step task. Each board featured a metal transport that could slide along the top surface of the board. The transport had an acrylic window that revealed a button the participant was instructed to press in order to complete the task.

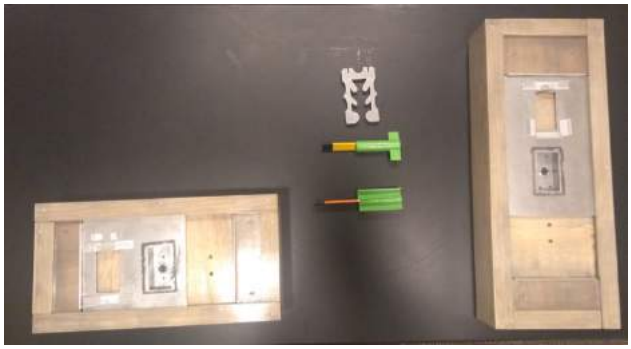


Figure 2: Set up of the initial task from the participant's perspective. The participant set the key onto the appropriate task board, slid the transport to reveal a button, then used one of the tools provided to press the button. The tools shown here were used in the ISR condition.

Figure 3: Close up of a "zeta" key placed on a task board. The shape of the key head (vertical slots on the top of the key) allowed it to fit onto the board (an "alpha" key with horizontal slots would not fit on this board). Here, the transport has been moved to its target position, so the button is accessible. To complete the second part of the task, a tool had to be inserted through the interior of the key to press the button.

The key frame on the transport was configured so that only one of the head shapes of the keys would fit into the frame for each board. Once a key had been properly placed into the frame, the transport could be moved to the target position (see *Figure 3*). This constituted the first part of the participant's task. When the transport was moved into the target position, the participant had access to the button through the interior of the key. The participant was instructed to press the button using a tool provided for that purpose (see *Figure 4*). Using one of the tools to press the button constituted the second part of the task. Once the participant pressed the button, lights built into the board turned on signaling that the task had been successfully completed.

As noted, the first part of the task required the participant to attend to and differentiate the head shape of the key being used so that the key could be correctly placed onto one of the task board transport frames. The second part of the task required the participant to attend to the interior shape of the key. In the ISI condition, the interior shape of the key was irrelevant in terms of which tool could be used. Participants in that condition could select either of the two tools available to press the button. In the ISR condition, the tools were designed so that each tool fit with the interior shape of one of the types of keys (either the alpha keys or zeta keys). Participants in that condition not only had to attend to the interior, they had to make a decision about which tool to use given the interior shape of the key.

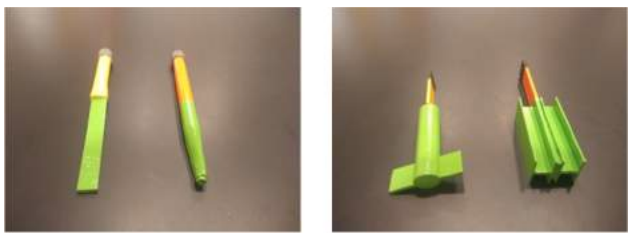


Figure 4: In the ISI condition (tools on left), either tool could be used with any key. In the ISR condition (tools on right), the tool used depended on the interior shape of the key.

At the start of the experiment, participants were introduced to the keys, task boards, and tools with a set of practice materials that allowed them to familiarize themselves with the basic aspects of the task (i.e. place a key onto the transport of one of the boards, move the transport, use a tool to press the button). However, the practice keys had different shape attributes and fit onto the transports differently (and the transports rotated on the surface of the board instead of sliding). Once the participant indicated they were comfortable with the basic idea of the task, the practice materials were replaced with the actual task boards and tools for the study. It is important to note that at no time during the initial task trials were the participants told that there were different types of keys – each trial featured one key and the instructions and communication with the participant focused solely on the completion of the task.

At the start of each trial during the initial task, the experimenter set a key (determined by a randomized order for each participant) on the table between the task boards. As described prior, the participant's task consisted of placing the key onto the proper transport, sliding the transport to its target condition, and then pressing the button using one of the tools. The participant handed the key back to the experimenter, the transport on the task board was returned to its initial position, and the trial ended. The keys were kept out of sight except for when they were being used during a trial. The participant completed two blocks of eight trials during the initial task. Each key was used once within each block.

After the initial task trials, the participants moved to a computer workstation. The computer tasks were designed and administered using PsychoPy software (Peirce, 2007). The images of the keys used in the computer tasks were created using the same computer aided design (CAD) software that was used in printing the physical keys. The 3D images of the keys used during the following tasks were identical to the physical keys the participants had used during the initial tasks, excepting for the modifications noted below.

The first task the participant completed on the computer was a *classification task*. The purpose of this task was to provide the participants with labels for the concepts they had acquired during their initial interactions with the keys. Chin-Parker and Birdwhistell (2017) showed that participants can make category-based judgments, e.g. sorting and similarity judgments, following goal-directed

tasks even without explicit labels, but as we planned to use a category-goodness rating task, the participants needed a way to explicitly differentiate the concepts. Because the general head and interior shape were perfectly correlated in the keys, the participants could use either, or both, of those attributes to guide their classification decisions. We expected the participants to look to whichever feature they already considered to be critical in terms of their knowledge of the keys, so the classification task should only reinforce the concepts they acquired during the initial task.

The initial screen of the classification task provided information about the task, and at this point the participants were explicitly told that there were two types of keys, identified for the classification task as "alpha keys" and "zeta keys". During each trial of the classification task, the image of a key was presented and the participant used the mouse to indicate whether they thought it was an alpha or zeta key. The participant received feedback on her classification, and the correct label for the key was shown with the key so that she could study it for two seconds before the next trial began. Each participant completed 16 classification trials comprised of two blocks of the eight keys used in the initial task.

After completing the classification task, the participant began the *category rating task*. Each trial consisted of an image of a key presented with a category label (see *Figure 5*). The participant was instructed to rate how good a member of the indicated category the key was. The participant could use the mouse to click on a rating scale that went from 0 (labeled with "definitely not this type of key") to 100 (labeled with "perfect example of this type of key"). The participant was encouraged to use the entire range of the scale in order to most accurately reflect her ratings of the keys shown.

Each participant completed 32 trials in the category rating task. There were eight types of items shown during the task, four instances of each type. They were balanced in terms of whether they represented alpha or zeta keys. In the *old match trials*, the key shown was one from the initial task and it was displayed with the correct category label. In the *old mismatch trials*, a key from the initial task was displayed with the incorrect category label. In the *new match trials*, new versions of keys that matched the keys from the initial task (and these keys would have functioned the same with the task boards and tools) were displayed with the appropriate label. In the *head violation trials*, the image of the key was modified so that it had the same basic shape, but it would no longer fit onto either key frame if it had been used in the initial task. In the *interior violation trials*, the interior of the key was modified so that it had a different shape that would keep any tool (from either the ISI or ISR conditions) from being able to reach the button in the initial task. The *head/interior mismatch* trials had items where the head from one type of key was matched with the interior from the other type of key, and these keys were presented with the label that matched the head of the key. Finally, there were *minor match trials* and *minor mismatch trials* where superficial aspects of the keys (e.g. whether the edges were rounded or squared off) were modified, and the key was presented with either the correct or incorrect label. These items were considered filler items.

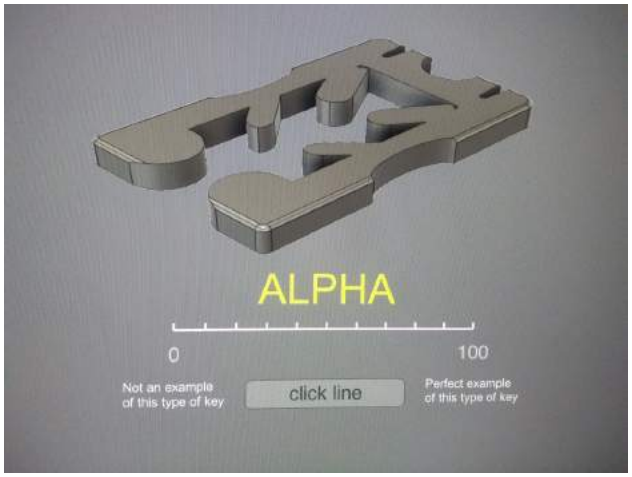


Figure 5: Screenshot of an Old Mismatch trial of the category rating task.

After the participant completed the category rating task, she was asked to move back to the task table. The task boards had been removed, and a pile of 16 "miniature" keys were in the center of the table. These keys were printed at ¼ scale and matched the keys from the initial task in terms of their attributes. Eight of these keys were identical to the initial task keys. The other eight keys were like the head/interior mismatch items from the category rating task – the head of one type of key was paired with the interior from the other type of key. The experimenter instructed the participant to "put these keys into groups that you think naturally reflect the types of keys you worked with today." The participant was free to sort the keys into any number of groups. Once the participant indicated she had completed the sorting task, the experimenter asked her to explain the sort.

Results

Initial Task Participants in the ISR condition (M = 221.03 secs, SD = 96.46) took longer to complete the first block of trials of the initial task than participants in the ISI condition (M = 162.24 secs, SD = 43.06), $t(65) = 3.24$, $p = .002$, $r_{pb}^2 = 0.14$. The difference persisted into the second block, but was much smaller in magnitude: ISR condition (M = 118.15 secs, SD = 21.59), ISI condition (M = 106.09 secs, SD = 19.82), $t(65) = 2.38$, $p = .02$, $r_{pb}^2 = 0.09$.

Classification Task There were no differences in the participants' ability to complete the classification task. The mean accuracy for the ISI condition (M = 0.95, SD = 0.07)

was nearly identical to the ISR condition (M = 0.96, SD = 0.07), $t(62) = 0.27$, $p = .78$, $r_{pb}^2 = 0.001$.

Category Rating Task The initial analysis of the category goodness ratings (see *Table 1*) was an omnibus test to determine whether the ISI and ISR conditions showed different patterns of ratings across the items. The category ratings were analyzed using a 2 (condition) X 8 (item type) mixed ANOVA. There was no main effect of the condition, $F(1, 65) = 0.07$, $p = .80$, $\eta_p^2 = .001$, but there was a significant main effect of the item type, $F(7, 455) = 78.04$, $p < .001$, $\eta_p^2 = .55$. Critically, there was a significant interaction between the condition and item type, $F(7, 455) = 2.16$, $p = .04$, $\eta_p^2 = .03$. Looking at the overall results, participants in both conditions provided similar ratings when both the head and interior of the key indicated the same category (i.e. the old match, new match, old mismatch, minor match, and minor mismatch items). The interaction appears to arise from a differential rating across items where the head and interior provided different information about the category membership.

Table 1: Mean Category Goodness Ratings (and Std. Error) Organized by Item and Condition

	Initial Task Condition	
Item	Interior Shape Irrelevant (ISI)	Interior Shape Relevant (ISR)
Old Match	85.13 (2.81)	85.52 (2.85)
New Match	73.10 (3.28)	77.61 (3.33)
Old Mismatch	10.33 (2.77)	13.77 (2.81)
Head Violation	40.38 (5.18)	50.27 (5.26)
Interior Violation	50.80 (5.87)	34.71 (5.96)
Head/Interior Mismatch	55.16 (6.47)	42.52 (6.57)
Minor Match	75.25 (3.81)	75.50 (3.86)
Minor Mismatch	13.57 (3.18)	13.82 (3.23)

As noted, we expected the manipulation of the relevance of the interior shape would affect the centrality of that attribute within the conceptual knowledge. To test this idea, we ran a more focused set of ANOVAs that assessed the conditions in terms of their ratings for the old items (as a baseline for the category ratings) compared to the items where the head and interior attributes provided different information about the category membership of the key (head violation, interior violation, and head/interior mismatch items). Across the analyses, there was a consistent effect of the item types (all ps < .001) because the old items were reliably rated as better members of the target categories compared to the items that contained inconsistent attributes. There was also no main effect of the condition in any of the analyses (all ps > .10). However, the interaction terms differed across the analyses. For the head violation items, $F(1, 65) = 1.19$, $p = .28$, $\eta_p^2 = .02$, and head/interior mismatch items, $F(1, 65) = 2.08$, $p = .15$, $\eta_p^2 = .03$, there was no interaction between the condition and item type. For the interior violation items, there was a significant interaction between the condition and item type, $F(1, 65) = 4.27$, $p = .04$, $\eta_p^2 = .06$. The interaction in the primary analysis appears to have occurred because the participants in the ISR condition dropped their ratings for the interior violation items in comparison to the old items more than the participants in the ISI condition did.

Sorting Task The participants in both groups created a variety of sorts for the miniature keys, and these sorts varied in terms of whether they reflected attention to a single attribute or multiple attributes. The sort by one participant in the ISR condition was not based on the physical attributes of the keys, so her data were removed from these analyses.

There was no difference in the number of groups created by participants in each condition (ISI condition: $m = 3.35$, $s = 2.28$; ISR condition: $m = 3.28$, $s = 1.99$), $t(64) = 0.13$, $p = .89$, $r_{pb}^2 = 0.001$. The proportion of participants that used information about the head of the keys (77% of ISI; 53% of ISR) differed between the conditions, $X^2_{(1, 64)} = 3.96$, $p = .04$, $v_c^2 = .06$. However, the proportion of participants that used information about the interior of the keys (53% of ISI; 66% of ISR) did not differ between the conditions, $X^2_{(1, 64)} = 1.10$, $p = .30$, $v_c^2 = .02$.

Forty-two participants (24 in the ISI condition, 18 in the ISR condition) sorted the items into only two groups. Those sorts provide a direct insight into what aspect of the keys was considered critical because the sort was based on a single attribute. Of this subset of participants, 66% of the participants in the ISI condition sorted the keys based on the head of the keys while 66% of the participants in the ISR condition sorted based on the interior. The primary attribute for the sort differed between the conditions, $X^2_{(1, 42)} = 4.58$, $p = .03$, $v_c^2 = .11$.

Discussion

The results of this study provide additional evidence for the goal-framework hypothesis. The participants in the two conditions were given equivalent tasks (and thus equivalent goals) to guide their interactions with the keys. In both conditions, the goals associated with their tasks oriented them to both of the critical attributes of the keys: the shape of the head and the shape of the interior. If the goal orientation hypothesis were sufficient to account for the role of the goal construct in the learning, the two conditions should have been largely equivalent in terms of how they organized their knowledge of the critical features of the keys. In some ways, they did show comparable learning. There is a striking similarity in terms of how the participants rated many of the items regardless of

condition. For instance, the mean ratings for the old match, new match, old mismatch, minor match, and minor mismatch items are nearly identical across the conditions. In all of those items, the head and interior attributes of the keys were in agreement with regard to the type of key shown. However, when the attribute information conflicted, the ratings differed between the conditions.

The participants in the ISI condition tended to consider the head shape as more critical when judging the category goodness of the items. Their ratings for the head violation items is lower than their ratings for the interior violation items. They also had a tendency to use the shape of the head of the key more consistently when organizing the keys during the final sorting task. The participants in the ISR condition tended to consider the interior of the keys as more critical to the category goodness, and they used the interior shape more consistently when sorting the keys.

As expected, the participants in the ISI condition acquired and used some knowledge of the interior shape in the category-based tasks. As noted, having to pay attention to the interior shape is sufficient to drive some learning. However, we would argue that their knowledge of the interior shape was less central to their concept of the key compared to the ISR condition, so it did not affect their ratings as much. A potentially important contribution of the notion of the goal-framework is that it explicitly connects the experiences of an individual, their interactions with objects in the world, to their conceptual knowledge.

It is not clear why the ISR participants showed less sensitivity to the head shape than the ISI participants. The head attribute played a comparable role in the task completion for both conditions. It is possible that having two goal-relevant attribute distinctions required some weighting of those attributes within the concept. This would fit with models of conceptual acquisition that have such a mechanism in place to account for the differential learning of attribute information. As suggested by an astute reviewer, it is also possible that the proximity of that part of the task to the completion of the goal might have privileged the knowledge of the interior shape in the concept. However, further study is necessary to determine why the ISR participants tended to emphasize the interior shape over the head shape.

The critical difference between the conditions was in the role the interior shape played in terms of how the participant could reach the goal. Both conditions had the same goal, to press the button, but the different tools during the second part of the task meant that the interior shape played a qualitatively different role in achieving that goal. In the ISI condition, the participant had to attend to the interior shape of the key in order to navigate the tool and press the button, but the shape of the interior of the key did not have relevance to the task beyond that. In the ISR condition, the shape of the interior was critical to differentiating the use of the tools to press the button. In this way, the differentiation of the shape of the interior was relevant to completing the task.

Chin-Parker and Birdwhistell (2017) posit that the learning process involves the development of the "goal framework" and that this framework reflects the structure that emerges as an individual interacts within a particular situation with a certain goal. In this way, the framework organizes the incoming information in terms of its goal-relevance providing structure to the acquired knowledge. They also propose that this framework is involved in organizing aspects of the basic perceptual experience of the individual when operating in a novel domain because there has to be some means to constrain the development of a feature language (see Landy & Goldstone, 2005). Although the current study was not designed to test these aspects of the goal-framework hypothesis, they fit within the experiences of the participants. When they had arrived for the study, they had no idea what the keys were or how to think about them. By the time they had completed the sorting task, they had a meaningful sense of what the keys were. As we develop this paradigm, we intend to revise the tasks so that we have the power to look at subtler indicators of the developing conceptual knowledge so we can assess these other claims.

This study examines concept acquisition in an arguably more naturalistic manner than most research in this this area. Our participants used the keys to complete an admittedly simple and arbitrary task, but in doing so, they had meaningful interactions with objects in a particular context in order to reach a goal. As a result, they developed useful ways to organize their knowledge of keys. This kind of experience invokes pragmatic constraints that are important to conceptual acquisition (Barsalou, 2017). The concept acquisition that occurs is not driven solely by the physical characteristics of the keys. Similarly, the goal of individual, in isolation, is unable to account for the conceptual acquisition. Instead, it is the interactions between the individual and environment that allow the useful structure to emerge.

Acknowledgments

This research was supported in part by a grant from the Denison University Research Foundation. Special thanks go to Christian Faur for his assistance with the 3D printing. Eric Gerlach and Zehui Xu assisted with data collection.

References

Barsalou, L. W. (1983). Ad hoc categories. *Memory & Cognition, 11, 211*-227.

Barsalou, L.W. (1995). Storage side effects: Studying processing to understand learning. In A. Ram & D. Leake (Eds.), Goal-driven learning (pp. 407-419). Cambridge, MA: MIT Press/Bradford Books.

Barsalou, L.W. (2017). Cognitively plausible theories of concept composition. In Y. Winter & J. A. Hampton (Eds.), Compositionality and concepts in linguistics and psychology (pp. 9-30). London: Springer Publishing.

Chin-Parker, S., & Birdwhistell, J. (2017). Diagnosticity and prototypicality in category learning: A comparison of inference learning and classification learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 216-226.

Jee, B. & Wiley, J. (2007) How goals affect the organization and use of domain knowledge. *Memory & Cognition, 35*, 837-51.

Kruschke, J. K. (2003). Attention in learning. *Current Directions in Psychological Science, 12*, 171-175.

Landy, D., & Goldstone, R. L. (2005). How we learn about things we don't already understand. Journal of Experimental and Theoretical Artificial Intelligence, 17, 343–369.

Le Pelley, M. E., Mitchell, C. J., Beesley, T., George, D. N., & Wills, A. J. (2016). Attention and associative learning in humans: an integrative review. *Psychological Bulletin*, *142*, 1111.

Love, B. C. (2005). Environment and goals jointly direct category acquisition. *Current Directions in Psychological Science, 14,* 829-835.

Machery, E. (2007). 100 years of psychology of concepts: The theoretical notion of concept and its operationalization. *Studies in History and Philosophy of Biological and Biomedical Sciences*, *38*, 63-84.

Markman, A. B., & Ross, B. H. (2003). Category use and category learning. *Psychological Bulletin*, *129*, 592.

Murphy, G. L. (2005). The study of concepts inside and outside the laboratory: Medin versus Medin. In W. Ahn, R. L. Goldstone, B. C. Love, A. B. Markman, & P. Wolff (Eds.), Categorization inside and outside the lab: Festschrift in honor of Douglas L. Medin. Washington, DC: APA.

Peirce, J. W. (2007). PsychoPy - Psychophysics software in Python. *Journal of Neuroscience Methods, 162,* 8-13.

Ross, B. H., Chin-Parker, S., & Diaz, M. (2005). Beyond classification learning: A broader view of category learning and category use. In W. Ahn, R. L. Goldstone, B. C. Love, A. B. Markman, & P. Wolff (Eds.), Categorization inside and outside the lab: Festschrift in honor of Douglas L. Medin. Washington, DC: APA.

The First Crank of the Cultural Ratchet: Learning and Transmitting Concepts through Language

Sahil Chopra (schopra8@stanford.edu)
Department of Computer Science, Stanford University

Michael Henry Tessler (tessler@mit.edu)
Department of Brain & Cognitive Sciences, MIT

Noah D. Goodman (ngoodman@stanford.edu)
Department of Psychology & Department of Computer Science, Stanford University

Abstract

Human knowledge accumulates over generations, amplifying our individual learning abilities. What is the mechanism of this accumulation? Here, we explore how language allows accurate transmission of conceptual knowledge. We introduce a novel experimental paradigm that allows direct comparison of learning from examples and learning from language. In our experiment, a *teacher* first learns a Boolean concept from examples; they then communicate this concept to a *student* in a free conversation; finally, we test both teacher and student on the same transfer items. We find that learning from language is both sufficient and efficient: Students achieve accuracy very close to their teachers, while studying for less time. We then explore the language used by teachers and find heavy reliance on generics and quantifiers. Taken together, these results suggest that cultural accumulation of conceptual knowledge arises from the ability of language to directly convey generalizations.
Keywords: concept learning; cultural ratchet; communication

Introduction

The human species is remarkable: We are able to learn by observing the world around us, forming new concepts that support prediction and manipulation (Tenenbaum, Kemp, Griffiths, & Goodman, 2011). Yet human concept learning has limits: Life is only so long and a person can only be in one place at a time. Individual learning from observations is thus unlikely to fully explain the ecological successes of our species (Henrich, 2015). If we are able to faithfully transmit our knowledge to the next generation, then limited individual learning can accumulate over generations to arrive at powerful systems of knowledge—a process termed the "cultural ratchet" (Tomasello, 1999). How does the ratchet work? What aspects of cognition support faithful transmission?

Cultural transmission has been often studied through the lens of imitation. This mechanism is particularly useful for learning procedural knowledge and rituals (Legare & Nielsen, 2015). Reproducing the behaviors of conspecifics, however, does not easily address ideas that go beyond the here-and-now: our generalizable knowledge and intuitive theories. Language, on the other hand, is a tool by which humans can convey abstract information. It allows us to transmit knowledge that would be otherwise difficult or unsafe to observe directly (e.g., which plants are poisonous; Gelman, 2009; Tessler, Goodman, & Frank, 2017).

Prior experimental work in cultural transmission has suggested that language may be a sufficiently expressive channel for conveying hard-to-discover knowledge (Beppu & Griffiths, 2009; Morgan et al., 2015). For example, Morgan et al. (2015) found that knowledge about stone flaking and tool making were best transmitted via verbal language. This work did not examine in detail, however, the kinds of natural language expressions utilized in the transmission of knowledge, nor relate it to the concepts being transmitted. In this paper we introduce a novel experimental paradigm that allows to explore how language can support the "first crank" of the cultural ratchet: how concepts learned from examples by one person are faithfully transmitted to a second via language.

Typical concept learning experiments are structured so that a single subject is presented with examples of objects that belong to (and don't belong to) a new category (Bruner, 1956; Piantadosi, Tenenbaum, & Goodman, 2016). We extend this paradigm by asking the initial learner to convey the concept to a second person. We allow them to do so freely using language. We then separately test the initial and secondary learner on the category. This allows us to explore detailed questions about whether and how language allows faithful transmission of these concepts: Is language sufficient for conveying concepts? How efficient is language compared to directly studying examples? What aspects of language are used to convey concepts?

In the remainder of the paper we introduce our experimental paradigm and then explore the resulting data with a variety of analyses. We find that language is sufficient and efficient for concept learning, and that certain linguistic forms seem to underlie this efficacy.

Methods

Participants

We recruited 224 participants from Amazon's Mechanical Turk (MTurk). This number was chosen to yield approximately 10 dyads per concept. Participants were restricted to those with U.S. IP addresses and at least a 95% work approval rating; in addition, participants who self-reported a native language other than English or failed to partake in the experiment (accepted the hit but then discussed matters entirely unrelated to the experiment) were excluded. In total, 11 pairs were excluded on this basis. The experiment took on average 15 minutes and participants were compensated $1.25 with an additional performance bonus (described below).

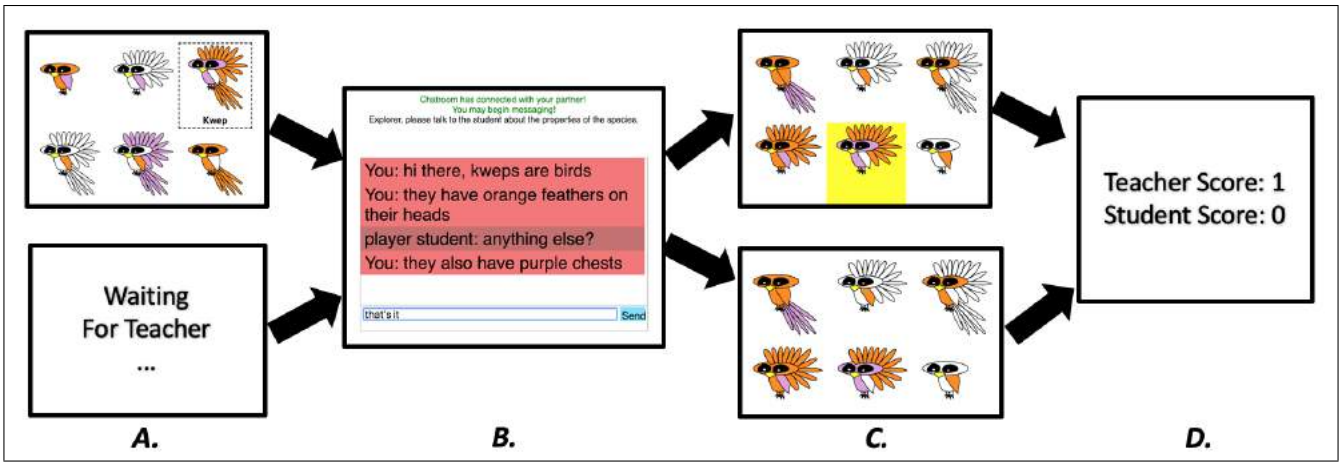


Figure 1: During the *concept learning* phase (**A**) the teacher (above) clicked through a grid of creatures, revealing the labels for creatures in the train set, while the student (below) waited. During the *concept communicating* phase (**B**), the teacher explained the concept to the student in a chatroom. During the *concept testing* phase (**C**), both participants were shown the same grid of held-out test creatures. Each selected the creatures (yellow) that they believed belonged to the concept. Finally, the participants were shown their scores for the round (**D**).

Concepts and stimuli

Participants learned concepts generated by 5 different *rules* (i.e. logical forms): Single Feature, Conjunction, Disjunction, Conjunctive Disjunction, and Disjunctive Conjunction. Rules were realized in specific *concepts* by varying Boolean properties of programatically generated images of creatures, from five different kinds: flowers, bugs, birds, fish, and trees (see Figure 1 for an example). Each kind had 5 to 7 Boolean features that we used to realize our concepts. Each of the 5 rules was realized twice in each creature kind, yielding a total of 50 concepts (listed on the axis of Figure 3). For each concept, we generated 100 specific creatures, split into 50 for training and 50 for testing. We ensured some positive examples of the concept even for very restrictive rules by first randomly selecting 6 positive instances of the concept and then adding 44 items chosen at random from all remaining items (i.e., according to the true concept base rate).

Procedure

Every pair of participants was placed in a game, where one was assigned the role of the "teacher" (initial learner) and the other was assigned the role of the "student" (secondary learner). Each game consisted of 5 rounds, each with a new concept from a new rule. Each of a game's 5 concepts used a different creature kind, and each concept was presented with a different nonce word as the species name. The ordering of concepts was randomized so that there was no standard ordering of rule types across the games.

On each round, participants went through three phases: *concept learning*, *concept communicating*, and *concept testing* (Figure 1). During the *concept learning* phase, the teacher was presented a grid of training creatures and was instructed to click on individual creatures to reveal whether or not they belonged to the species defined by the concept. Once the teacher clicked on every creature in the grid, they were presented a message advising them to review the creatures for as long as they needed. When the teacher ended the concept learning phase, they proceeded to the *concept communicating* phase, where they entered an online chatroom and were were instructed to teach the concept to the student. Participants were provided no additional instructions for the chatroom, and they were allowed to talk freely. In order to prevent a teacher from rushing through the chatroom without properly communicating with their student, only the student was given the ability proceed to the final *concept testing* phase. In the final phase both participants were (separately) given the same grid of test creatures and asked to tag the creatures that they believed belonged to the species. Neither participant had access to their chatroom messages during this phase.

Once both participants completed *concept testing* for a concept, they were provided feedback in the form of their own and their partner's score, computed as: # of hits − # of false alarms. We encouraged them to learn concepts thoroughly and communicate effectively with a monetary bonus equal to the sum of both players' scores (in cents). Participants were made aware of the task structure and bonus mechanic prior to starting the first round; they had to answer 5 comprehension questions correctly to begin to the game.

Analysis and Results

Our experiment yielded rich data for exploring whether and how concepts are learned from language, and how learning from language compares to learning from examples. We first examine performance during the *concept testing* phase for both the student and teacher participants. We then explore the time spent learning from each type of evidence. Finally,

we explore the actual language used to teach concepts in the *concept communicating* phase.

Concept learning performance

Participants assigned to be the *teacher* take part in a standard Boolean concept learning paradigm, and results are in accord with expectations. The five rule types we used in our experiment cover a range of complexity in terms of description length (Feldman, 2000), which manifests in variable performance in test accuracy across types (Figure 2).

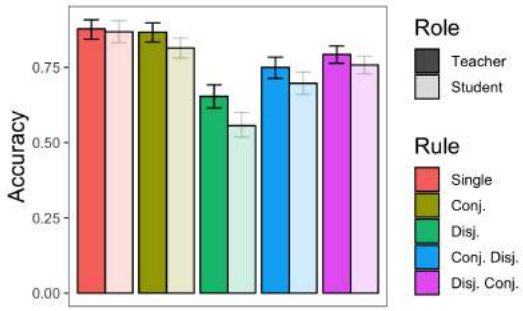


Figure 2: Average accuracy of teachers and students during the *concept communicating* phase of the experiment. Error bars denote bootstrapped 95% confidence intervals.

Students have access to the concept only through the language conveyed by their partner. The average student accuracy during the *concept testing* phase should thus be no greater than the average teacher accuracy, which appears to be true for our 5 rule types in Figure 2. To assess the potential accuracy differences between learning from examples vs. from language, we built a Bayesian mixed-effects model predicting whether or not a participant responded accurately during the *concept testing* phase as a function of the rule, the participant's role (teacher vs. student), and their interaction. We included random intercepts and effect of rule for participants and random intercepts and effect of role for each of the 50 concepts. All regression models were created in Stan (http://mc-stan.org/) accessed with the brms package (Brkner, 2017). We find a main effect of role such that students were less accurate than teachers (posterior mean and 95% credible interval: $\beta = -0.41(-0.69, -0.12)$). However, this effect is very small in absolute terms—the average difference in accuracy for students vs. teachers is just 5.3% (95% credible interval: 2.7%, 8.2%). Thus language appears to be sufficient to convey concepts; students are able to learn concepts from language, yielding performance very close to their teachers, who had access to the actual training examples.

Performance on individual concepts (rules reified in particular stimuli) reveal substantial variability in learning. Figure 3 shows the average performance of teachers and students for each of the 50 concepts along with the concept-specific chance accuracy[1]. Teachers perform above chance in all concepts, but there is significant variation in performance for concepts within a given rule. Such variation is expected given the known importance of feature salience and other stimulus properties on concept learning (Nosofsky, 1986). Notably, the gap between teacher and student performance also varies.[2] This variability cannot be attributed to stimulus features, which are shared between teacher and student, but rather reflect the language available for conveying different features. Inspection reveals that concepts with a large gap in teacher-student performance have a small number of teachers who used language in idiosyncratic ways. For example, one teacher described creatures belonging to the concept "bugs: no wings" as "like a worm ... [with a] straighten[ed] body". Another teacher described "flowers: purple petals OR thorns" as "no color ... a flower with sharp edge branches and some tails". In both cases the teacher fails to use a simple word for the relevant feature ("wings", "thorns") unlike most other participants. These cases may arise from particularly confused teachers, particularly difficult to describe features, or an interaction. We return to this question below.

Often, how well a person learns depends on the particular person they learned from. We find a strong linear relationship between average student accuracy and (corresponding) teacher accuracy across the 50 concepts ($r = .88$, $p < .001$; Figure 4). We further find that this correlation remains strong at the individual level ($r = .60$, $p < .001$; Figure 4).

While this suggests that students make mistakes when their teacher does, we may further ask whether they make the *same* mistakes. Since teachers and students are presented the same held-out test examples in the same order during the *concept testing* phase of the experiment, we can measure the similarity between teacher and student responses at the level of individual stimuli using Hamming distance (the total number of times the student and teacher responded differently). The average distance between teacher and student in our data set is 11.1 differences (out of 50 possible). To calibrate this number we computed a baseline by randomly permuting teacher-student pairings, which yields average distance 19.9 (95% CI $[19.84, 19.96]$). A second, tighter, baseline considers permutation of student-teacher pairs only within each concept (matching evidence seen by teachers). This yields average distance 13.53 (95% CI $[13.18, 13.91]$). Thus we can conclude that students' pattern of responses is more similar to their own teachers' responses than to other teachers in the same concept (and in the whole data set). Language seems to be sufficient to convey the concept as understood by the teacher, even when the teacher has learned the wrong thing.

We do not have a direct measure of teacher confidence that

[1]Chance is defined here as the accuracy achieved by guessing at random but with the base rate of positive examples shown for that concept. This is a stronger comparison than random guessing.

[2]Generally teachers do better than students. Ten concepts show the opposite trend, to varying extents. Three of these differences are driven by a few outliers where the teacher attained low accuracy in the final phase even though they properly communicated the concept. Seven of these concepts have students that are negligibly more accurate than teachers, i.e. correctly identify 1-2 more stimuli, of the 50 presented.

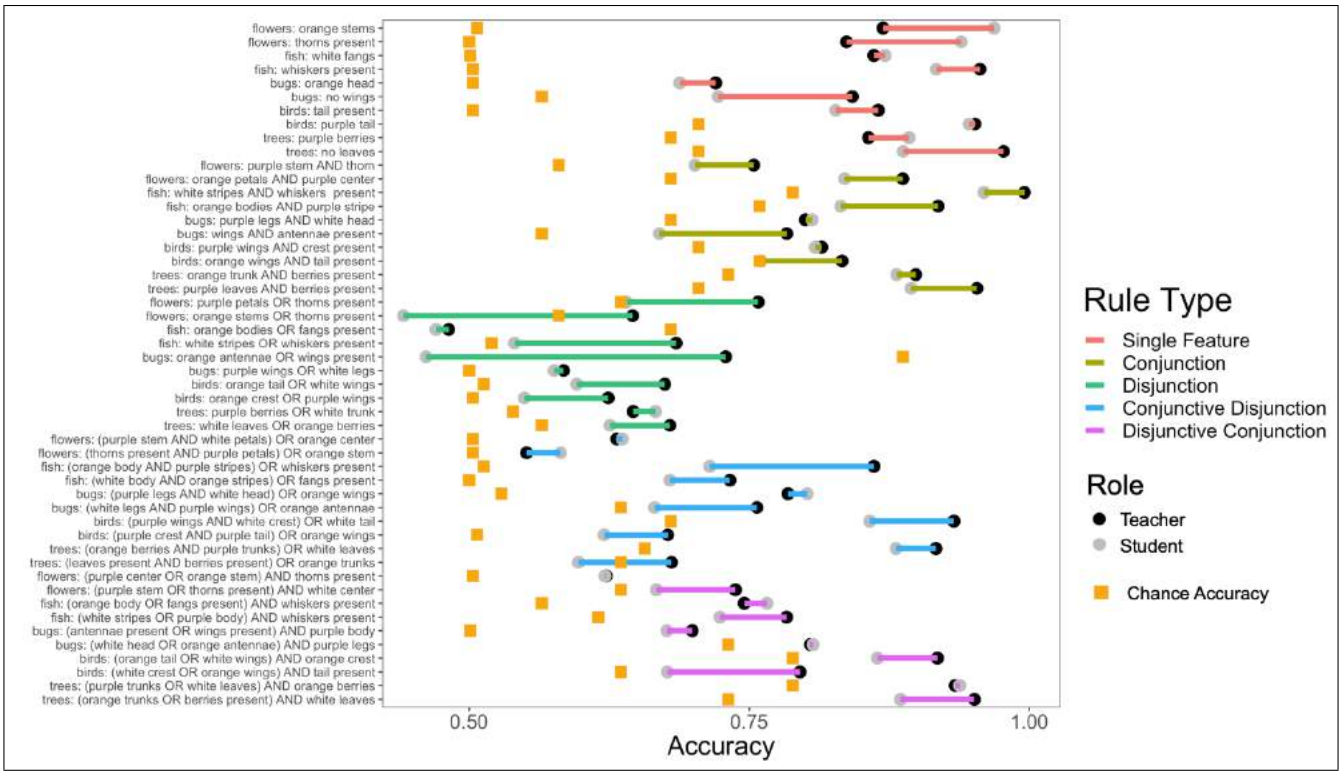


Figure 3: Accuracy on each concept. Black dots denote the average teacher accuracy on the test set; gray dots denote the average student accuracy. Gold squares denote chance accuracy.

we could use to explore the impact of confidence on the efficacy of language for transmission. Instead, we analyze an indirect measure of teacher confidence: the mean teacher accuracy within a concept. Figure 5 shows the relationship between teacher accuracy and distance from teacher to student responses. We find a strong relationship: language seems to yield stronger alignment between students and teachers when the teachers are (expected to be) confident in what they have learned ($r = -0.75, p < 001$).

Study time for observation vs. language

As we saw above, language appears to be relatively *sufficient* for conveying concepts, how *efficient* is language compared to directly learning from observed examples? We could consider efficiency in terms of amount of evidence required to learn or amount of effort required. In our experiment the amount of evidence was fixed in the *concept learning* phase, but the study time was controlled by participants. We thus consider study time as a proxy for learning effort. Since the amount of time spent in the *concept communicating* phase was similarly controlled by participants, we use time as a proxy also for effort required to convey a concept with language. Using time to measure learning effort makes it possible to directly compare effort required to learn from observations and from language.

For each concept, we recorded the amount of the time that teachers spent in the *concept learning* phase. During the *concept communicating* phase we recorded the time that elapsed between the moment a participant began typing a message into the chatbox and the moment they sent the message to their partner. Since some messages may have been unrelated to learning (e.g. pleasantries or commentary), we coded every message in the data set as "Informative", "Follow-Up", "Confirmation", "Miscellaneous", or "Social". Informative messages were those related to the concept that were sent by teachers without prompting from the student. Messages in the ensuing dialog that were relevant to the concept were labeled as Follow-Up. Social pleasantries ("hi", "hello", etc.) were labeled as Social, and messages that were unrelated to the current concept (e.g. commentary about performance on previous rounds) were labeled as "Miscellaneous". Overall, there were 1012 Informative, 1751 Follow-Up, 160 Social, and 300 Miscellaneous messages in the data set. For our time analysis, we only considered the concept-related messages: the Informative and Follow-Up messages that constituted the majority of participants' conversations.

To compare the study time of learning from examples vs. from language (Figure 6), we built a Bayesian mixed-effects model with fixed effects of rule, participant role, and their interaction; in addition, we included random intercepts and effects of rule for each participant and random intercepts and effects of participant role for each concept.[3] We observe

[3]The data was modeled as being generated from a lognormal dis-

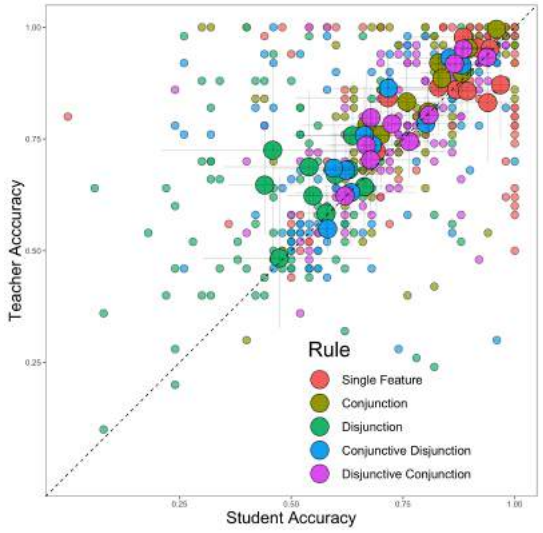


Figure 4: Accuracy of student-teacher pairs in *concept testing* phase. Small dots indicate individual teacher-student pairs, while larger dots indicate mean accuracy of teachers and students for a concept. Lines indicate bootstrapped 95% confidence intervals of teacher and student accuracy for a concept.

that the simplest rule (Single Feature) took the teacher substantially less time to study than average (comparison to the grand mean: $\beta = -0.25(-0.37, -0.12)$) and the most difficult rule took substantially more time to study than average ($\beta = 0.20(0.08, 0.31)$). Crucially, study time was systematically shorter when learning from language than from examples ($\beta = -0.64(-0.82, -0.48)$), which translated into an average 57 seconds (42, 75) less time for learning from language. There were no interactions between role and rule that were plausibly different from 0.

This suggests that learning from language may be *more*

tribution.

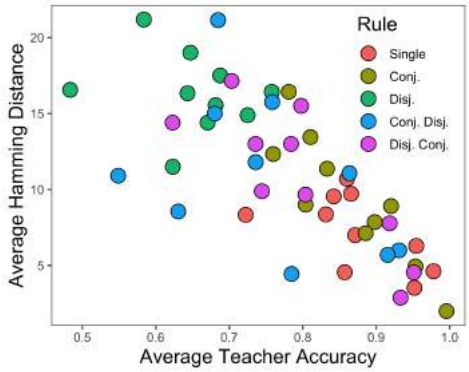


Figure 5: Average accuracy of the teacher versus the average hamming distance between student and teacher responses during *concept testing* phases of all 50 concepts.

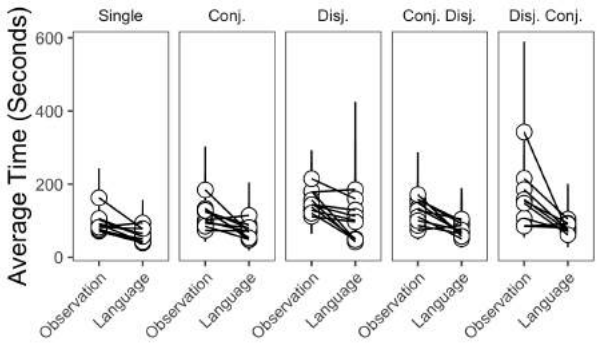


Figure 6: Time spent by teachers learning concepts from observation and time spent by teacher-student pairs communicating about concepts. Circles denote average time for a concept, error bars are bootstrapped 95% confidence intervals. Lines pair the same concept.

efficient than learning from observing examples. This conclusion warrants further study however, as our measures of study time likely depend on specific paradigm choices. For instance, teachers were forced to click on all 50 creatures during the *concept learning* phases of the experiment—it may be that not all of this time was needed for belief updating (as opposed to rote clicking of the stimuli).

Language used for knowledge transmission

We have seen that language is a sufficient and (probably) efficient means for transmitting concepts in our experiment. Now we turn to the question of what specific aspects of language were used by teachers to convey concepts. We first coded each of the messages in the game as Informative, Follow-Up, Social, or Miscellaneous, as described above. A vast majority of the messages (2763 of the 3223) were concept-relevant, i.e. Informative or Follow-up.

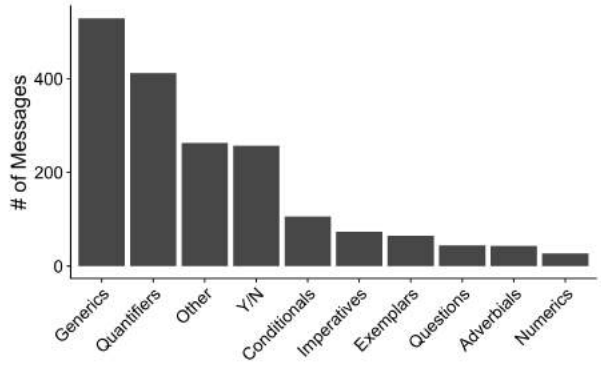


Figure 7: Distribution of concept-relevant messages.

When properties are predicated on categories, the resulting linguistic expression is typically a quantified sentence (e.g.,"All wugs have orange heads"; "Most feps have purple

Table 1: Utterance Categories & Examples

Category	Example(s)
Generics	"morseth[s] come in a variety of colors"
	"they have saber teeth"
Quantifiers	"some will have wings"
	"all morseth[s] have long whiskers"
Conditionals	"if the left wing is orange click it "
Exemplars	"12 white feathers ... no tail ..."
	"12 white feathers ... 5 white tail..."
Imperatives	"so click on the teeth!"
	"focus on orange fish..."
Adverbials	"usually their wing colors match their..."
	"stems are usually colored as well"
Numerics	"2/3 of them are the ones that qualify"
	"75% of what I clicked on was a zorb"
Yes/No	"nope", "k",
	"yes", "okay"
Other	"this sounds difficult"
	"okay idk what else to say"

wings") or a generic sentence which lacks explicit quantification (e.g., "Morseths have saber teeth" Carlson & Pelletier, 1995). Rather than talking about categories explicitly, participants could convey the actual examples they saw using numerical language (e.g., "4 of them ...") or describing individual exemplars (e.g., "white-tail with feathers, white-tail with no feathers, ...").

The first author first identified Generics consistent with other coding schemes used for generic sentences (Gelman, Goetz, Sarnecka, & Flukes, 2008), then identified Quantifiers, Numerics, and Exemplars, before grouping the remaining messages by the following linguistic constructs: Conditionals, Imperatives, Adverbials, and Yes/No statements. Remaining messages were grouped as "Other". See Table 1 for examples of messages across these categories.

Figure 7 shows label counts for concept-relevant messages in our data set. The majority of these messages use generics or quantifiers to convey information about the category, with generics being the most common. Examining this distribution within rules, we find that this pattern holds for all except disjunction, where quantifiers are more prevalent than generics. Additionally, we find that the number of generics (%G) and quantifiers (%Q) amongst concept-relevant messages does not vary appreciably across the rules: single features: (35%G, 21%Q); conjunction (34%G, 22%Q); disjunction (21%G, 26%Q), conjunctive disjunction (24%G, 21%Q); disjunctive conjunction (31%G, 22%Q).

The remainder of the responses are mostly made of other commentary about the concepts and Yes/No responses. It is important to note that teachers could have directly instructed the students what to choose (with Imperatives) or described their specific experience (e.g. "there were three morseths with blue wings"); they chose instead to use linguistic constructs that convey generalizations across categories.

Discussion

In this paper we introduce the first experimental paradigm that permits apples-to-apples comparison of learning concepts from examples and from language. We found that language is *sufficient* for faithful concept transmission, in the sense that the student who learns from language is nearly as accurate as the teacher who learned from examples (and as inaccurate, making similar mistakes). We have also seen preliminary evidence that language is *efficient* for concept transmission: that it may take less time to learn a concept from helpful language than to learn it from observations.

Most work on cultural transmission either investigates well-controlled experimental paradigms but with heavily restricted modes of transmission (e.g., sharing direct observations; Efferson et al., 2007; Kalish, Griffiths, & Lewandowsky, 2007; Griffiths, Lewandowsky, & Kalish, 2013; Kirby, Cornish, & Smith, 2008; Smith, Kalish, Griffiths, & Lewandowsky, 2008; Martin et al., 2014) or use open-ended modes of transmission (e.g., creating an instructional video) but on complex tasks where a ground-truth is difficult to establish (Muthukrishna, Shulman, Vasilescu, & Henrich, 2014; Caldwell & Millen, 2008; Morgan et al., 2015). In this paper, we chart a middle course: investigating a well-studied phenomenon (Boolean concept learning) with a relatively open-ended mode of transmission (free language production).

This allows us to perform parallel analyses of *what is being learned* and *how that knowledge is conveyed*. Recent advances in natural language processing have demonstrated potential in training and parameterizing classifiers according to language (Andreas, Klein, & Levine, 2017; Srivastava, Labutov, & Mitchell, 2018). Meanwhile, there has been a growing body of research aimed at understanding effective teaching and learning within Cognitive Science (Chi, Roy, & Hausmann, 2008; Chi, Siler, Jeong, Yamauchi, & Hausmann, 2001; Kapur, 2014). We believe that bringing the pedgagogical perspective to machine learning will be instrumental to improving models that learn from language. Importantly, our novel experimental method allows for scalable data collection of language-based instruction and provides a clear classification task, i.e. training models to learn from discourse and demonstrate understanding by predicting student responses.

In our experiment, we found substantial evidence that quantifiers and, especially, generics are used by teachers to convey their knowledge about concepts. In one sense this is unsurprising, as these linguistic constructs are *about* category generalization. Yet our results provide the first direct evidence for the connection between these aspects of language and cultural transmission of knowledge. This in turn provides initial support for a strong hypothesis about the mechanisms of knowledge accumulation: *The cultural ratchet arises specifically out of the ability of language to convey generalizations through generics and quantifiers*.

Acknowledgments

The research is (partially) based upon work supported by the Defense Advanced Research Projects Agency (DARPA), via the Air Force Research Laboratory (AFRL).

References

Andreas, J., Klein, D., & Levine, S. (2017). Learning with latent language. *arXiv preprint arXiv:1711.00482*.

Beppu, A., & Griffiths, T. L. (2009). Iterated learning and the cultural ratchet. In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st annual conference of the cognitive science society* (pp. 2089–2094).

Bruner, J. S. (1956). *A study of thinking*. New York, Wiley.

Brkner, P.-C. (2017). brms: An r package for bayesian multilevel models using stan. *Journal of Statistical Software*.

Caldwell, C. A., & Millen, A. E. (2008). Studying cumulative cultural evolution in the laboratory. *Philosophical Transactions of the Royal Society B: Biological Sciences*.

Carlson, G. N., & Pelletier, F. J. (1995). *The generic book*. University of Chicago Press.

Chi, M. H., Roy, M., & Hausmann, R. G. (2008, 03). Observing tutorial dialogues collaboratively: Insights about human tutoring effectiveness from vicarious learning. *Cognitive science*, *32*, 301-41.

Chi, M. H., Siler, S. A., Jeong, H., Yamauchi, T., & Hausmann, R. G. (2001). Learning from human tutoring. *Cognitive Science*, *25*(4), 471-533.

Efferson, C., Richerson, P. J., McElreath, R., Lubell, M., Edsten, E., Waring, T. M., ... Baum, W. (2007). Learning, productivity, and noise: an experimental study of cultural transmission on the bolivian altiplano. *Evolution and Human Behavior*, *28*(1), 11–17.

Feldman, J. (2000). Minimization of boolean complexity in human concept learning. *Nature*, *407*(6804), 630–633.

Gelman, S. A. (2009). Learning from others: Childrens construction of concepts. *Annual Review of Psychology*.

Gelman, S. A., Goetz, P. J., Sarnecka, B. W., & Flukes, J. (2008). Generic language in parent-child conversations. *Language Learning and Development*, *4*(1), 131.

Griffiths, T. L., Lewandowsky, S., & Kalish, M. L. (2013). The effects of cultural transmission are modulated by the amount of information transmitted. *Cognitive Science*, *37*(5), 953-967.

Henrich, J. (2015). Culture and social behavior. *Current Opinion in Behavioral Sciences*, *3*(1), 84–89.

Kalish, M. L., Griffiths, T. L., & Lewandowsky, S. (2007, Apr 01). Iterated learning: Intergenerational knowledge transmission reveals inductive biases. *Psychonomic Bulletin & Review*, *14*(2), 288–294.

Kapur, M. (2014, 03). Productive failure in learning math. *Cognitive science*, *38*.

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, *105*(31), 10681–10686.

Legare, C. H., & Nielsen, M. (2015). Imitation and innovation: The dual engines of cultural learning. *Trends in Cognitive Sciences*, *19*(11), 688–699.

Martin, D., Hutchison, J., Slessor, G., Urquhart, J., Cunningham, S. J., & Smith, K. (2014). The spontaneous formation of stereotypes via cumulative cultural evolution. *Psychological Science*, *25*(9), 1777–1786.

Morgan, T. J. H., Uomini, N. T., Rendell, L. E., Chouinard-Thuly, L., Street, S. E., Lewis, H. M., ... et al. (2015, Jan). *Experimental evidence for the co-evolution of hominin tool-making teaching and language.* Nature Publishing Group. Retrieved from https://www.nature.com/articles/ncomms7029

Muthukrishna, M., Shulman, B. W., Vasilescu, V., & Henrich, J. (2014). Sociality influences cultural complexity. *Proceedings of the Royal Society B: Biological Sciences*.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39-57.

Piantadosi, S. T., Tenenbaum, J. B., & Goodman, N. D. (2016). The logical primitives of thought: Empirical foundations for compositional cognitive models. *Psychological Review*, *123*(4), 392424.

Smith, K., Kalish, M. L., Griffiths, T. L., & Lewandowsky, S. (2008). Introduction. cultural transmission and the evolution of human behaviour. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1509), 3469-3476.

Srivastava, S., Labutov, I., & Mitchell, T. (2018). Zero-shot learning of classifiers from natural language quantification. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 306–316).

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*(6022), 1279-1285.

Tessler, M. H., Goodman, N. D., & Frank, M. C. (2017). Avoiding frostbite: It helps to learn from others. *Behavioral and Brain Sciences*, *40*.

Tomasello, M. (1999). *The cultural origins of human cognition*. Harvard University Press.

Generating normative predictions with a variable-length rate code

S. Thomas Christie (tchristie@umn.edu)
Cognitive Science, 75 E River Rd
Minneapolis, MN, 55455, USA

Paul R. Schrater
Cognitive Science, 75 E River Rd
Minneapolis, MN, 55455, USA

Abstract

Cognitive science is an archipelago of concepts and models, with cross-pollination between topics of interest often prohibited by incompatible approaches. Despite this, behavioral performance universally depends on information transmission between brain regions and is limited by physical and biological constraints. These constraints can be formalized as information theoretic constraints on transmission, which provide normative predictions across a surprising range of cognitive domains. To illustrate this, we describe a simple variable-length rate coding model built with Poisson processes, Bayesian inference, and an entropy-based decision threshold. This model replicates features of human task performance and provides a principled connection between a high-level normative framework and neural rate codes. We thereby integrate several disjoint ideas in cognitive science by translating plausible constraints into information theoretic terms. Such efforts to translate concepts, paradigms and models into common theoretical languages are essential for synthesizing our rich but fragmented understanding of cognitive systems.

Keywords: information theory; bayesian inference; rate coding; response time; learning

Introduction and Background

Cognitive science is home to almost as many models as phenomena they purport to describe. While this *sui generis* approach to each problem allows rich and flexible descriptions, it stands in sharp contrast to the physical sciences, in which scientists strive for and expect simple unifying principles, like Newton's axiomatic laws, from which individual phenomena arise as particular circumstantial manifestations (Chater & Brown, 2008). In cognitive science, we would say that Newton's laws are *normative*. But where are our first principles, from which we can hope to derive a coherent set of expectations about how cognition *should* operate? In this paper, we consider information transmission from the environment, through the brain, to behavior. By constraining both the channel code and each transmitted signal to be optimally inferred under normative assumptions, we can construct a message-transmission system that replicates the Hick-Hyman law (Hick, 1952; Hyman, 1953) and the Power Law of Practice (Newell & Rosenbloom, 1981), illuminates the connection between transmission rate and energy use, and produces human-like response time distributions. Our information-theoretic approach affords a principled way to connect levels of analysis (Marr, 1982) by integrating energetic resource availability, message encoding and decoding schemes, and task performance characteristics into a single framework.

Applying information-theoretic concepts to the study of cognition is not new. The years following Claude Shannon's 'A Mathematical Theory of Communication' (1948) produced a wealth of information-theoretic analyses of cognitive function, perhaps the most famous of which resulted in the Hick-Hyman law (Hick, 1952; Hyman, 1953). This mathematical approach merged with optimal control theory to become Cybernetics (Wiener, 1965), which promised to understand cognition and behavior as just another system of information transmission, feedback, and control, and subject to the same constraints. Despite their successes, enthusiasm about both information theory and cybernetics has not persisted to the present day, partly because cybernetics was abstracted away from biological and neurological characterizations, and partly because the cognitive revolution led to a focus on the nature and calculus of representation.

The development of cognitive architectures has resulted in highly successful models of a broad array of tasks (Sun, 2008; Anderson et al., 1997; McClelland, 2009). In parallel, architecture-free computational principles like Bayesian inference, prediction, credit assignment, and generalization bounds on learning have provided a rich framework for normative thinking (Shiffrin, 2010; Griffiths et al., 2008). Computational architectures form a possible hybrid (Chater & Brown, 2008), using normative computational principles to structure a cognitive architecture. However, these principles are often expressed in mathematical language disconnected from cognitive and neural architectures, leading to a pervasive difficulty in translating between mathematical formulation and plausible neural implementaiton.

The inability to translate between cognitive models directly results in a lack of knowledge transfer between domains (cognitive processes, language, tasks, etc) and levels of analysis (high-level models to low-level mechanistic details). For example, consider cognitive control as a case-in-point illustration. 'Cognitive control' refers to the deployment of attention and memory resources in the service of competing tasks. Each of these (control, attention, and working memory) are famously limited in capacity and inextricably intertwined in their roles in executive function. It is well-known that task practice lessens the effort required to do tasks, lessens attentional load, reduces response times, and decreases the amount of cognitive control required (Logan, 1985; Moors, 2016; Pierce & McDowell, 2017). These effects mirror practice

effects in perceptuo-motor skill acquisition, suggesting there should be some common principles, but it is currently difficult to transfer insights gained in the study of one phenomenon to the study of others. This lack of transferability means that cognitive science has developed a series of 'knowledge islands,' making it almost impossible to share insights across boundaries.

Even within a single topic, with the same underlying concepts, there are often many theories that are difficult to relate to each other. For example, under the shared working assumption that mental effort is treated as a cost in a cost-benefit analysis, there is considerable disagreement about the nature of the cost. Depending on the theory employed, it may represent an opportunity cost from foregone tasks, a loss of the intrinsic reward of cognitive leisure, the tendency of mental effort to discourage use of limited-capacity resources like working memory, or simply the effort of cognitive control as a cost per se. Although Shenhav et al. (2017) show that these ideas share common computational principles, they also leave it as an open question how to compare them directly. The difficulty is that cognitive costs are exogenous to the computational architecture, which means there are many non-equivalent ways to import them. Without further normative constraints, there are many rational ways to import computational modeling ideas (like costs), which means each new model multiplies the translational difficulties for integrating and relating existing models, constraints and concepts.

Like Shenhav et al. (2017), we take as foundational that the brain is an information-processing organ, ultimately transferring information from the environment via sensation, through the brain, and back into the environment as behavior. Although high level theories of behavior are most easily expressed in decision- and control-theoretic terms, re-expressing these theories in information-theoretic terms affords the incorporation of constraints on information processing, as illustrated by work on bounded rationality (Ortega et al., 2015). Biological constraints involving energy availability and noise, when translated into information theory formalism, become *normative bounds* on the ability to transfer information. Similarly, limitations on information available to the organism provide bounds on task performance. In essence, information theory provides a well-known, well-understood and sophisticated language for translating models and theories that has largely untapped potential. We illustrate this potential by demonstrating its capacity to use common computational principles to reveal relationships between the seemingly unrelated phenomena of learning rates, response time distributions, and energetic resource utilization.

Framework

Whatever the task at hand, neurons performing task-related computations must infer, in a continuous-time and streaming manner, which 'messages' are being transmitted from other brain regions (Rieke et al., 1999). This inference process is noisy, imperfect, and time-dependent. We model this process by performing continuous-time inference about the configuration of stochastic processes, with a stopping criterion based on a posterior entropy threshold. This approach produces normative predictions that match the behavioral characteristics so commonly observed in experimental paradigms, including the shape of response-time distributions and the decrease in response times and mental effort as a function of practice.

Characterizing the relationship between inferential constraints and transmission efficiency is the domain of information theory (Cover & Thomas, 2012). Information theory has been transformative in its applications to electronic communications, and has provided useful normative predictions for neural characteristics (Bialek, 2012). In particular, information theoretic constraints underlie the Efficient Coding Hypothesis (Barlow et al., 1961; Simoncelli & Olshausen, 2001), which suggests that neural connectivity is structured in such a way as to encode information from the natural environment with maximum efficiency. Despite widespread evidence for the general validity of this hypothesis in early sensory systems (e.g. Laughlin (1981); Vinje & Gallant (2000); Pitkow & Meister (2012)), there is still significant uncertainty as to whether information theoretic principles are relevant at the level of cognitive processing. Central to this reservation is a concern that Shannon's proofs of the existence of arbitrarily efficient binary codes rely on his use of 'block codes,' in which several messages are combined into a single string in a way that increases the likelihood of error-free transmission (Shannon, 1948; Cover & Thomas, 2012). For example, Luce (2003) writes "Shannon's way of defining the concept [of channel capacity] requires that not individual signals be transmitted but rather very long strings of them so as to be rid of redundancies. That is rarely possible within psychological experiments." Another recent paper raises similar concerns that Shannon's method of encoding "requires complex computation and long delays to encode and decode in ways that achieve optimality," and that it only "applies to settings of perfect signal recovery, which may not be possible or even desirable in biological settings" (Park & Pillow, 2017).

Concerns about the applicability of Shannon's proofs to information transmission in the brain confuse levels of analysis (Marr, 1982). It is true that Shannon's reliance on block-coding to achieve efficient information transmission is an implementation-level detail applicable to discrete-time codes and not to the communication of information between neurons. However, the core conceptual contribution of information theory lies not in coding techniques but in providing a method for quantifying uncertainty. More broadly, the theory serves to characterize the ways in which noise and redundancy affect the reliability, efficiency, and rate of inference. From this broader perspective, it is surely applicable to the study of cognitive function. That an understanding of these factors can lead to the design of optimal codes is important, but the specifics of code design in a discrete-time system do not invalidate the application of general principles to the study of cognition.

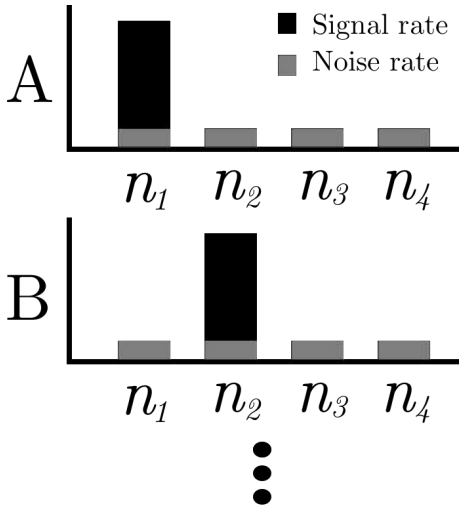


Figure 1: A codebook converts symbols A, B, etc. from a symbol alphabet into configurations of firing rates across Poisson processes $n_1, n_2, \ldots$. In this simple model, the codebook assigns a signal rate λ_S to a single Poisson process for a given symbol. Each Poisson process also emits spikes at a noise rate λ_N. As Poisson process rates are additive, this results in a total emission rate of $\lambda_N + \lambda_S$ for the 'activated' process.

In the remainder of this paper, we show an example of a continuous-time variable length coding mechanism, built using entropy and inference, that adheres to the principles of information theory while providing normative predictions of signal transmission time and accuracy. We emphasize that the continuous-time nature of the code means that signals are not discretized. Because of this, we are able to transmit messages such that transmission time is linearly related to message surprisal, replicating the Hick-Hyman law. By presenting such a code, we show that appropriate information-theoretic concepts can be applied to the study of neural information transmission.

Implementation

We model information transmission by having a sender encode a message into a configuration of Poisson process firing rates, and a receiver watch the generated spikes until they are confident about the configuration of underlying rates, and thus about the content of the encoded message (see Figure 2 for a schematic of the architecture). In more detail, the transmission mechanism consists of an encoder, a transmitter, a receiver, and a codebook. The transmitter is an array of Poisson processes, each continuously producing points or 'spikes' independently at a given noise rate λ_N. This can be viewed as a basic model of a neural rate code, as neural spikes trains are often modeled as Poisson processes (Rieke et al., 1999). The symbols to be communicated are taken from an alphabet of discrete symbols $\mathcal{A}$. The codebook describes a mapping between each symbol and a configuration of Poisson rates,

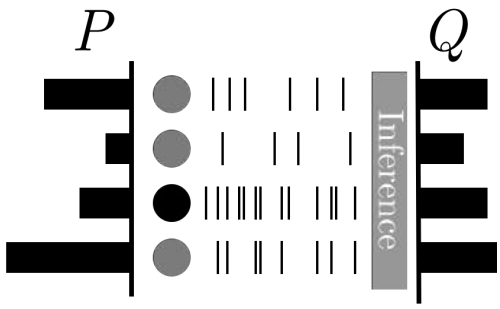


Figure 2: Messages are selected from a source distribution P. The codebook translates each message into a higher firing rate for a single process (a simplifying, but not restrictive, assumption). Poisson processes stochastically emit spikes, which are observed by the inference process. Bayesian inference combines the prior distribution Q with the likelihood of each message given the accumulated observations to produce a posterior distribution over possible messages.

and the mapping from a given symbol to rate configuration is carried out by the encoder. For the sake of expositional simplicity, we restrict the codebook to increasing the rate for a single Poisson process from the noise rate λ_N to a signal rate $\lambda_N + \lambda_S$, as shown in Figure 1. The neural analogue is that each Poisson process is 'tuned' to 'prefer' a particular symbol in a 1-hot manner, resulting in a sparse code.

The receiver observes the sequence of spikes emitting from each Poisson process and continuously attempts to infer which rate configuration is producing the spikes it observes, and thereby which symbol is being transmitted. We assume, again for simplicity and consistent with common information-theoretic analysis, that the receiver knows the values of both λ_N and λ_S. In standard binary or Gaussian channels, transmission is a discrete vector of amplitudes that takes a fixed time to transmit. Because of this, practitioners typically speak in terms of transmitting bits-per-signal, or bits-per-second (which are a constant multiple of each other). In our case, the receiver accumulates information about each transmission gradually, over time. In effect, observing for a longer period of time adds redundancy to the signal.

As observations continue, the receiver calculates and continuously updates a posterior probability distribution over possible messages, and stops decoding when the entropy of the posterior reaches a pre-specified stopping threshold. Let transmitted symbols be treated as realizations of a random variable X. The receiver begins each transmission at time $t = 0$ with an initial uncertainty $H_Q(X)$ regarding the symbol being transmitted, reflecting its prior distribution $Q(X)$ of the possible codewords. As time passes and observations $Y_t = \{y_1, \cdots, y_t\}$ are made, the receiver uses Bayesian inference to update the prior to obtain a posterior distribution $Q_t(X|Y_t)$ over messages according to Bayes rule, which yields an updated posterior entropy $H_{Q_t}(X|Y_t)$. The posterior entropy decreases non-linearly with time and reflects the de-

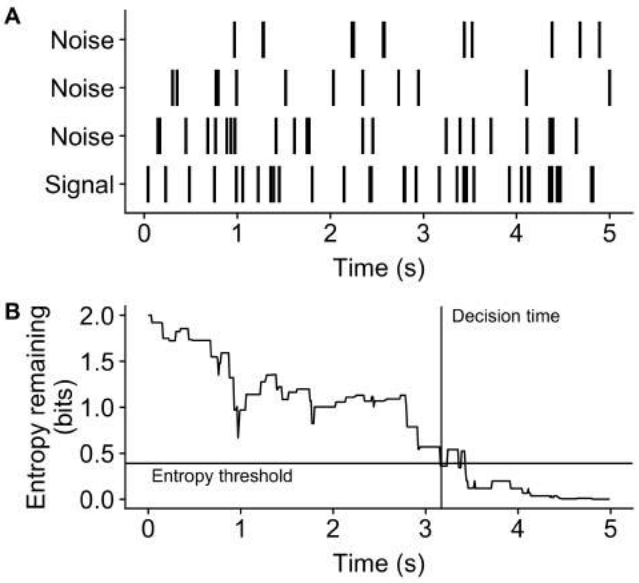


Figure 3: (A) Spikes are randomly emitted by each Poisson process as a function of time. The lower-most Poisson process is firing at a higher $\lambda_N + \lambda_S$ rate, while the others are firing at rate λ_N. (B) The receiver observes the spikes and infers which process is firing at rate $\lambda_N + \lambda_S$. The initial entropy is 2 bits, indicating a weak belief in equal probabilities for each of the 4 possible signals. The receiver's remaining entropy changes as the processes are observed and the posterior probability of each signal is calculated.

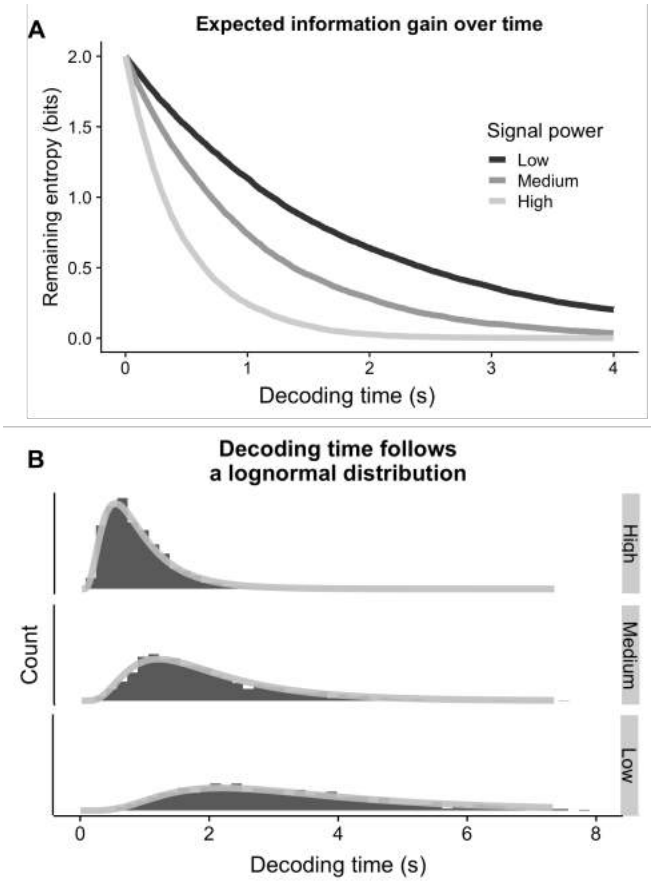


Figure 4: (A) The expected value of the receiver's entropy regarding four possible messages decreases as spikes are observed. Increasing the signal power λ_S changes the information transmission rate. (B) Response time distributions vary as a function of signal power λ_S, and in each case are well-fit by a log-normal distribution.

gree of confidence that a message has been correctly received. Transmission stops when $H_{Q_t}(X|Y_t)$ reaches a threshold. Figure 3 shows the change in posterior entropy over time for an example transmission.

Variable length transmissions

In the coding scheme introduced here, messages are *variable-length*: transmissions of messages with higher surprisal takes more time than messages with low surprisal, where surprisal is calculated using the prior probability distribution $Q(X)$ of the receiver. Recall that the surprisal $h(x)$ of a message x drawn from a distribution $P(X)$ is the logarithm of the inverse probability of the message, $h(x) = \log_2 \frac{1}{P(X=x)}$.

In 'entropy codes,' codeword length (and thus transmission time of each codeword) is roughly proportional to the surprisal of the encoded symbol in the absence of noise. When symbols are independently drawn according to a categorical probability distribution, this can manifest in two ways. In the first, increasing the number of possible symbols increases the surprisal of each individual symbol, and consequently the length of the code needed to encode its value. In the second, symbols drawn from a categorical distribution with unequal probabilities will have different surprisal values: more frequently transmitted messages will have lower surprisal and shorter codes than less frequent messages. We performed simulations to explore these scenarios in turn using our transmission model.

First, we varied codebook sizes and recorded transmission times using a fixed entropy threshold and a uniform source distribution. The nonzero entropy threshold occasionally results in transmission errors, as we see in human subjects. Information transmitted is thus less than the surprisal of each individual message, on average. We computed actual information transmitted by calculating the mutual information between transmitted symbols and received symbols, for each codebook size. The results are shown in Figure 5 and are a close qualitative match for the Hick-Hyman observations of human response times reported by Hick (1952) and Hyman (1953).

We next transmitted messages drawn from a non-uniform distribution $P(X)$ and measured transmission time for each message. For each transmission, we measured the information transmitted by comparing the receiver's prior probability distribution $Q(X)$ (which equals the source distribution

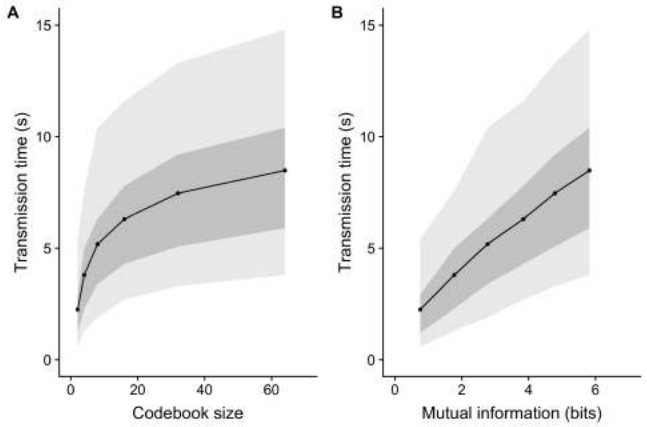


Figure 5: Mean transmission time increases logarithmically with codebook size and linearly with information transmitted, mirroring the Hick-Hyman law. Points represent mean transmission times and shaded regions represent the 50% and 90% high-density interval of the transmission time distribution. In each case, messages were transmitted according to a discrete uniform distribution $P(X)$ over messages, and the receiver maintained a uniform prior distribution $Q(X) = P(X)$ of the same dimensionality. For each transmission, an entropy threshold of 0.3 bits was used, with $\lambda_S = 4$ and $\lambda_N = 10$.

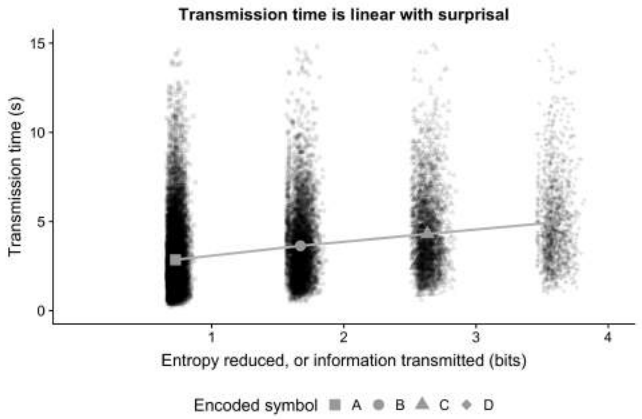


Figure 6: Mean transmission time increases linearly with actual information transmitted, echoing similar findings in humans by Hyman (1953). The quantity of information transmitted is calculated as the the KL-divergence between the prior distribution $Q(X)$ and the posterior distribution $P(X|Y)$ at decision time. Messages were drawn from a non-uniform source distribution $P(X)$. The receiver is assumed to know this source distribution and maintains a prior distribution $Q(X) = P(X)$. For each transmission, an entropy threshold of 0.3 bits was used, with $\lambda_S = 4$ and $\lambda_N = 10$.

$P(X)$, an assumption we relax below) with their posterior distribution $Q(X|Y)$ at decision time. We measured the difference in these distributions using the Kullback–Leibler divergence between the two distributions, $D_{KL}(Q(X|Y)||Q(X))$. The change between the receiver's prior and posterior distributions is equivalent to the decrease in the receiver's subjective uncertainty about which message is being transmitted. From the point of view of the receiver, this is equivalent to the amount of information transmitted, in bits. Figure 6 shows a linear relationship between message surprisal and transmission time, again qualitatively matching Hyman's reported results from human subjects.

Learning to efficiently transmit

As with source-coding systems, expected message transmission times are faster when more frequently transmitted messages are transmitted in less time than less frequently transmitted messages. In the our system, this is implemented by tailoring the receiver's prior distribution Q to match, as closely as possible, the source distribution P. This reveals an epistemic problem from the perspective of the receiver, which has no *a priori* knowledge of the source distribution: the prior must be learned and updated by observing message transmissions. The work of Hick and Hyman has been legitimately criticized for omitting this discussion (Laming, 2010).

Suppose we allow a receiver with an incorrect uniform prior message distribution Q_{init} to update its distribution to Q_{obs} in a Bayesian manner each time a message is received, so that the subsequent message transmission starts with the updated prior. As the receiver observes which messages are transmitted and at what relative frequency, Q_{obs} will become an ever-closer approximation to P, shrinking both $D_{KL}(P||Q_{obs})$ and the expected transmission times. Figure 7 shows message transmission times resulting from a uniform (naive) prior, a prior equal to the true source distribution, and an intermediate distribution, as might be expected to develop from a moderate level of experience with the task. In each case, response time is linearly related to message surprisal as calculated using Q. The slope depends on the amount of experience with the task: as experience accrues and Q_{obs} approaches P, response times more closely reflect the transmission frequencies of each message. The varying slopes are reminiscent of the subject-specific slope found by Hyman (1953).

As observations accumulate, the rate at which response times decrease as Q approaches P mirrors the Power Law of Learning (Newell & Rosenbloom, 1981). The Power Law of Learning is a ubiquitous finding that task response times have a power-law relationship with the number of practice episodes, when averaged across many subjects. We constructed a categorical source distribution P with $k = 16$ categories, but with most of the probability mass in two categories. We initialized Q_{init} to have a Dirichlet prior with concentration parameters 2, representing a weak prior belief that the source distribution is uniform. We simulated N message transmissions, for $N = 2$ to $N = 1024$, taken evenly in log space. For each value of N, we averaged the results across 1,000 simulated observers, resulting in an expected posterior

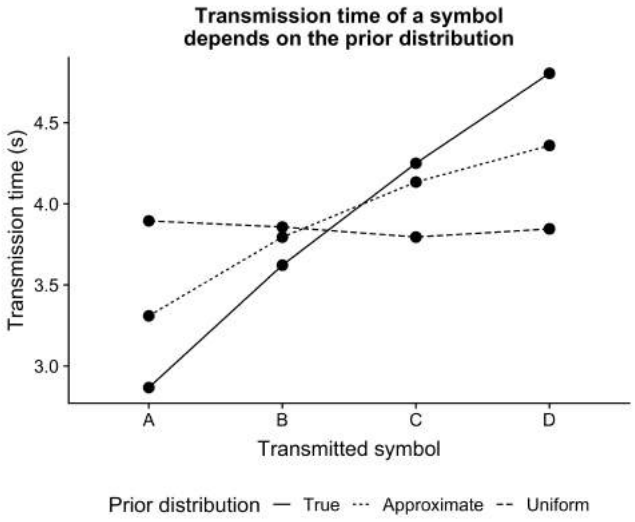


Figure 7: Mean transmission time is a function of the receiver's prior belief $Q(X)$ over messages, rather than the source distribution $P(X)$. In each case, messages were transmitted from the identical source distribution, where A was most frequent, followed by B, and so on. Each line connects response times arising from the same prior distribution. A uniform $Q(X)$ results in a flat line, while a $Q(X) = P(X)$ results in the steepest slope. In each case, the relationship between subjective surprisal and response time is approximately linear. For each transmission, an entropy threshold of 0.3 bits was used, with $\lambda_S = 4$ and $\lambda_N = 10$.

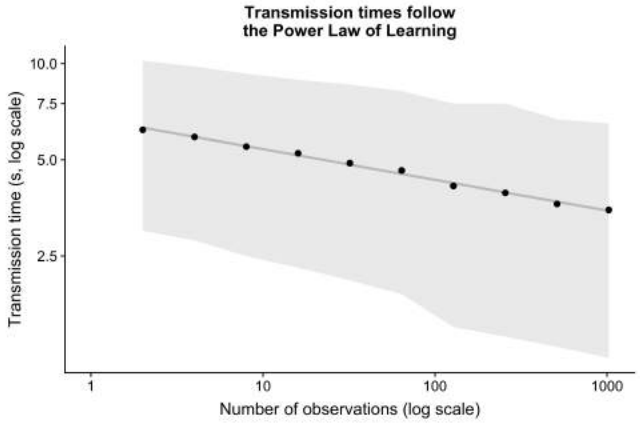


Figure 8: Simulated message transmission time decreases as a function of observations, as the prior Q approaches the source distribution P. Signals are transmitted with signal strength $\lambda_S = 4$, noise power $\lambda_N = 10$, and an entropy threshold of 0.3. Points represent mean transmission times, and the shaded region represents the 80% high-density interval of the response time distributions.

distribution Q_{obs} after N observations. For each Q_{obs} we then simulated more 2,000 message transmissions, with messages drawn with frequency defined by P, and calculated the transmission time for each. As illustrated in Figure 8, the relationship between observations N and transmission time is linear in log-log space, matching the Power Law of Learning.

The energy connection

Implicit in the above discussion is the notion that information transmission costs energy: transmission is initiated when an encoder assigns signal power λ_S to a Poisson process. If each spike costs energy, this implies a rate of energy expenditure. As shown in Figure 4, signal power has a direct effect on the rate of entropy decrease and the resulting transmission times. The framework introduced here allows us to explicitly describe the relationship between energy use (in terms of spikes), task novelty (in the form of naive Q estimates), task practice, and response times. If mental effort is a phenomenological correlate of signal transmission costs, it also provides a normative explanation for effort decrease as a function of practice, and provides weight to the currently tenuous relationship between mental effort and the utilization of metabolic resources.

Indeed, neural spikes are not free: an estimated 10% of an adult body's energy budget is allocated to neural information processing (Stone, 2018). In light of this, we might expect the brain to adopt a strategy of driving energetic efficiency by tailoring codes (represented by codebooks and Q distributions) to individual tasks. As stimulus distributions P are not equivalent between tasks, this would necessitate the creation and maintenance of a bank of task-specific codes, with a power-law response time trend repeated during the practice of each separate task (Newell & Rosenbloom, 1981). However, the power-law describes severely diminishing returns between task practice and transmission efficiency, and tasks in the world are not as discrete as in laboratory experiments. Because of this, in a naturalistic setting we might instead expect the brain to implement some 'universal' code (Cover & Thomas, 2012) that provides moderately efficient transmission across range of tasks (Vera et al., 2018). If this is the case, the brain would sacrifice efficiency to achieve flexibility, which is, after all, a chief characteristic of human cognition.

Conclusion

We have applied the principles of information theory to a simple rate-coding model of neural information transmission. We showed that placing normative bounds on the inference of both source distributions and the content of individual signals results in a coding mechanism that predicts the Hick-Hyman Law and the Power Law of Practice, describes a principled connection between information transmission and energy use, and produces realistic response-time distributions. By utilizing the information-theoretic principles relevant to a continuous-time system (in particular entropy and inference), and avoiding those that are not (block-coding), we have produced a simple and parsimonious explanation of a wide range of phenomena.

References

Anderson, J. R., Matessa, M., & Lebiere, C. (1997). Act-r: A theory of higher level cognition and its relation to visual attention. *Human-Computer Interaction*, *12*(4), 439–462.

Barlow, H. B., et al. (1961). Possible principles underlying the transformation of sensory messages. *Sensory communication*, *1*, 217–234.

Bialek, W. (2012). *Biophysics: searching for principles*. Princeton University Press.

Chater, N., & Brown, G. D. A. (2008). From universal laws of cognition to specific cognitive models. *Cognitive Science*, *32*(1), 36-67.

Cover, T. M., & Thomas, J. A. (2012). *Elements of information theory*. John Wiley & Sons.

Griffiths, T. L., Kemp, C., & Tenenbaum, J. B. (2008). Bayesian models of cognition.

Hick, W. E. (1952). On the rate of gain of information. *Quarterly Journal of Experimental Psychology*, *4*(1), 11-26. doi: 10.1080/17470215208416600

Hyman, R. (1953). Stimulus information as a determinant of reaction time. *Journal of experimental psychology*, *45*(3), 188.

Laming, D. (2010). Statistical information and uncertainty: A critique of applications in experimental psychology. *Entropy*, *12*(4), 72.

Laughlin, S. (1981). A simple coding procedure enhances a neuron's information capacity. *Zeitschrift für Naturforschung c*, *36*(9-10), 910–912.

Logan, G. D. (1985). Skill and automaticity: Relations, implications, and future directions. *Canadian Journal of Psychology/Revue canadienne de psychologie*, *39*(2), 367.

Luce, R. D. (2003). Whatever happened to information theory in psychology? *Review of general psychology*, *7*(2), 183.

Marr, D. (1982). *Vision: a computational investigation into the human representation and processing of visual information. w. h.* WH San Francisco: Freeman and Company.

McClelland, J. L. (2009). The place of modeling in cognitive science. *Topics in Cognitive Science*, *1*(1), 11–38.

Moors, A. (2016). Automaticity: Componential, causal, and mechanistic explanations. *Annual Review of Psychology*, *67*, 263–287.

Newell, A., & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. *Cognitive skills and their acquisition*, *1*(1981), 1–55.

Ortega, P. A., Braun, D. A., Dyer, J., Kim, K.-E., & Tishby, N. (2015). Information-theoretic bounded rationality. *arXiv preprint arXiv:1512.06789*.

Park, I. M., & Pillow, J. W. (2017). Bayesian efficient coding. *bioRxiv*. doi: 10.1101/178418

Pierce, J. E., & McDowell, J. E. (2017). Reduced cognitive control demands after practice of saccade tasks in a trial type probability manipulation. *Journal of cognitive neuroscience*, *29*(2), 368–381.

Pitkow, X., & Meister, M. (2012). Decorrelation and efficient coding by retinal ganglion cells. *Nature neuroscience*, *15*(4), 628.

Rieke, F., Warland, D., Steveninck, R. d. R. v., & Bialek, W. (1999). *Spikes: Exploring the neural code*. A Bradford Book.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell system technical journal*, *27*(3), 379–423.

Shenhav, A., Musslick, S., Lieder, F., Kool, W., Griffiths, T. L., Cohen, J. D., & Botvinick, M. M. (2017). Toward a rational and mechanistic account of mental effort. *Annual review of neuroscience*, *40*, 99–124.

Shiffrin, R. M. (2010). Perspectives on modeling in cognitive science. *Topics in cognitive science*, *2*(4), 736–750.

Simoncelli, E. P., & Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual review of neuroscience*, *24*(1), 1193–1216.

Stone, J. (2018). *Principles of neural information theory: Computational neuroscience and metabolic efficiency*. Sebtel Press.

Sun, R. (2008). Introduction to computational cognitive modeling. *Cambridge handbook of computational psychology*, 3–19.

Vera, M., Vega, L. R., & Piantanida, P. (2018). Compression-based regularization with an application to multitask learning. *IEEE Journal of Selected Topics in Signal Processing*, *12*(5), 1063–1076.

Vinje, W. E., & Gallant, J. L. (2000). Sparse coding and decorrelation in primary visual cortex during natural vision. *Science*, *287*(5456), 1273–1276.

Wiener, N. (1965). *Cybernetics, second edition: Or the control and communication in the animal and the machine*. The MIT Press.

The everyday statistics of objects and their names: How word learning gets its start

Elizabeth M. Clerkin (emclerki@indiana.edu)
Linda B. Smith (smith4@indiana.edu)
Department of Psychological and Brain Sciences, Indiana University
1101 E. 10th Street Bloomington, IN 47405

Abstract

A key question in early word learning is how infants learn their first object names despite a natural environment thought to provide messy data for linking object names to their referents. Using head cameras worn by 7 to 11-month-old infants in the home, we document the statistics of visual objects, spoken object names, and their co-occurrence in everyday meal time events. We show that the extremely right skewed frequency distribution of visual objects underlies word-referent co-occurrence statistics that set up a clear signal in the noise upon which infants could capitalize to learn their first object names.

Keywords: word learning; natural statistics; egocentric vision

Introduction

Infants begin learning object names before their first birthday. We know they can do this because infants look to the pictures of an object upon hearing the name (Bergelson & Swingley, 2012). By these measures, individual infants do not know many object names, and their knowledge is fragile. Still, it is clear that the start of learning object names begins well before infants produce words. By consensus, these novice learners must begin learning object names by linking the heard word to visually present objects. The ability to do this has been demonstrated in experimental studies (e.g., Smith & Yu, 2008). The problem is that the everyday visual world is much noisier and cluttered than the learning tasks presented in the laboratory (Clerkin, Hart, Rehg, Yu, & Smith, 2017). Laboratory studies also show that in the period just prior to the first birthday, infants have limited attention skills and quite limited memories for the learned object-word pairings taught in a single laboratory session (Vlach & Johnson, 2013). Accordingly, the field lacks a complete understanding of how object name learning gets its early start.

Learning depends on both the internal learning mechanisms and the data for learning. There are critical gaps in current knowledge about the everyday experiences that comprise the data for early object name learning. We know that parent-naming events are often ambiguous as the visual world is cluttered (Cartmill et al., 2013), parents often do not talk to the child in the home during natural activities (Tamis-LeMonda, Custode, Kuchirko, Escobar, & Lo, 2018), and parents only sometimes name the objects in the child's view during naturalistic play (Yurovsky, Smith, & Yu, 2013). Still, we know little about the statistical structure of everyday experiences across multiple naming events (but see Bergelson & Aslin, 2017 for recent work on this topic). Here we provide evidence-based estimates on three key statistical properties of the learning environment: the frequency distribution of heard object names, of seen visual objects, and their co-occurrence.

Rationale

The frequency distributions of words in parent talk to children are known to be extremely skewed with a small set of extremely frequent words and a much larger set of very rare words (Montag, Jones, & Smith, 2018). A small set of words that are heard pervasively – day in and day out – might define a constrained set upon which object name learning could get its start. Analyses of one large corpus of child-directed talk, however, suggests that the frequency distribution for object names in parent talk is not as skewed as other grammatical classes such that there are less dramatic differences between the most and least frequent object names (Sandhofer, Smith, & Luo, 2000). However, these analyses considered all parent talk – not talk within a particular context. Parent talk, and the words infants hear, are context bound (Montag et al., 2018). The child should be much more likely to hear the words "spoon" and "table" at mealtime than to hear the words "bat" or "ball." Thus, the key question for the role of very high frequency objects names at the start of object name learning may lie in the pervasiveness of a select set of objects names within a context.

There is very little evidence on the frequency distribution of visual objects in the natural environment generally or in infant everyday experiences in which these objects that must be linked to heard names. The evidence that does exist about the natural visual environment – from analyses of large corpora of photographs (Salakhutdinov, Torralba, & Tenenbaum, 2011) and from one analysis of head camera images collected by infants in their home (Clerkin et al., 2017) – suggests that the frequency distribution of object categories will be extremely skewed. The latter evidence further suggests that the very high frequency categories will correspond to the object names that are learned early by infants. Common sense and extant evidence from photography corpora (Sadeghi, McClelland, & Hoffman, 2015) also suggests that visual objects will be context dependent, with spoons and tables more likely in the immediate visual scene at mealtime than bats and balls.

For novice learners to learn object names, heard names must co-occur with referents in their experience. If a few object categories and their names are concurrently pervasive in infant everyday experiences, then there is a clear statistical solution to how object name learning starts – with the learning of the names of those few pervasive objects in infant experiences. Here we provide direct evidence on this possibility and show that the pervasive objects and pervasive

Figure 1. Example (non-consecutive) images from the videos recorded during infant mealtimes.

names in infant language learning environments *do not correspond* well, but that the learning environment offers a different statistical solution to the start of object name learning based on the 1) the skewed distributions of visual objects and 2) the quantity and quality of word-referent co-occurrences.

Method

The Corpus

We chose the mealtime context for three reasons: it is frequent, occurring on average 5 times a day for infants in this age group, the names for objects likely present at mealtimes are among the earliest learned concrete nouns by normative age of acquisition, and it is a potentially challenging context for learning given the sparsity of parent talk (Tamis-LeMonda et al., 2018) – very unlike contrived play contexts in laboratories. These mealtime events were selected from head cameras[1] embedded in hats worn by 14 infants aged 7 to 11 months at home as they went about their daily activities with no experimenters present (see Clerkin et al., 2017; Jayaraman, Fausey, & Smith, 2015 for details). Parents were not told specific activities to record and were told to record any and all activities during the times their infants were awake over a period of several days.

Figure 1 shows example images extracted from the video. Critically, the video collected from the head cameras is from the infants' ego-centric view. Thus, we have captured the visual environment directly in front of the infants' faces and the objects in it to which infants could be attending in any given moment. This ego-centric perspective is highly dependent on the infants' motor abilities, their interests, and their location and posture in any given moment. In sum, not only are we studying the natural word learning context at scale, but we are doing so with reference to the infants' own point of view.

Any video that included eating or meal preparation was included in the mealtime corpus which totaled 16.99 hours of footage and consisted of 344 mealtime events with 24.57 per subject on average (SD = 20.02).

Coding

Visual Objects Still images were down-sampled from the video recordings at a rate of 0.2hz (1 image every 5 seconds). The 11,549 down-sampled images were then coded by naïve adult coders for the 5 most obvious objects in the scene using basic level nouns; (see Clerkin et al., 2017 for more details). Each image was coded by 4 coders. These adult judgements of objects that are in view do not necessarily align with what the target infant's visual attention in the moment; however, we use these adult judgements as a way of describing the clutter of the natural environment from which infants are presumably visually sampling.

We chose to keep the coders' responses as intact as possible to avoid biasing the data; however, we did clean the data in the following ways. First, extraneous adjectives were removed (e.g., "baby spoon" was reduced to "spoon"); however, if an adjective-word combination was listed in the dictionary (e.g., "high chair"), it remained as a unique object. Also, different forms of the same object name were collapsed (e.g., "cup" and "cups" were both counted as instances of "cup"). Finally, words that were overly general (e.g., "food") or clearly did not refer to a concrete object (e.g., "color") were removed entirely. The frequencies of visual object categories are reported as the proportion of frames in which the object category occurred.

Object Names All speech in the target infants' environment was transcribed for each mealtime using Datatyvu (Datavyu Team, 2014). The audio data was broken down into 5 second intervals for ease of coding and to have an appropriate comparison to the visual data coded at 1 image every 5 seconds. It should be noted that infants this age do not talk, and thus none of the transcribed speech is the target infants' own vocalizations. Naming events (defined as any moment an object name was said) were extracted from the speech stream for object names that referred to objects which were reported as occurring at least once in the visual scenes. The speech transcripts were cleaned as described above for visual objects. The frequencies of object names are reported as the number of naming instances for each name across the corpus

[1] The Looxcie 2 weighs 22 grams and has a 75° diagonal field of view.

as a proportion of the number of 5 second intervals containing any speech.

Age of Acquisition Categories In order to understand how the statistics of the natural learning environment relate to learning first words, objects were broken down into two age of acquisition (AoA) categories. Objects in the First category were those named by nouns on the MacArthur Bates Communicative Developmental Inventory – MCDI (Fenson et al., 2007) and are present in the receptive vocabulary of 50% of 18-month-old children in the Wordbank repository of thousands of MCDI administrations (Frank, Braginsky, Yurovsky, & Marchman, 2016). Later objects were all other objects given by the coders.

Co-Occurrence Co-occurrence was coded by three trained raters in the laboratory. Each naming instance was located in the video, and if during the 5 second interval surrounding the naming instance an object which could be called by that name was visually present, then the coder recorded there was a co-occurrence. Approximately 20% of the naming instances were coded by all 3 coders. The final judgment for those instances was the response recorded by at least 2 of the 3 coders. The overall percentage agreement between the coders was 76.2%, but there were no naming instances on which at least 2 coders did not agree. Co-occurrence is reported as the proportion of naming instances during which a corresponding object was visually present.

Table 1: Summary of data coded.

	Num Frames	Num Speech Intervals
Total	11549	12237
With Talk	–	6833
Without Talk	–	5404

Table 2: Object and object name counts

	Num Unique Objects	Num Unique Object Names
Total	1095	350
First	118	97
Later	977	253

Results

Table 1 provides the number of fames coded for visual object and the total number of 5 second speech intervals, the number containing any speech, and the number containing no speech. Table 2 provides the number of unique visual objects and unique object names. As is apparent, there are many more objects than object names, showing considerable selectivity in parent talk relative to the wide variety of objects in view.

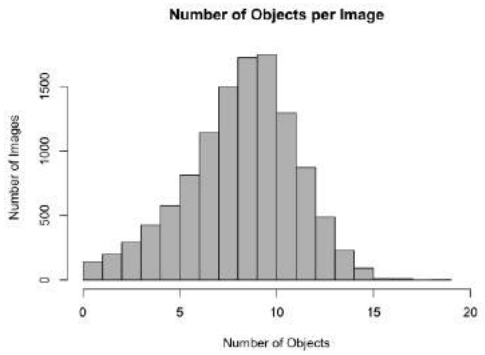


Figure 2. The frequency distribution of the number of objects per image across the corpus.

We consider the statistical regularities characterizing the visual objects, then the object names, and finally, their co-occurrence.

Visual Objects

The number of objects coded in each scene is an indication of the clutter present in the natural visual environment. Because 4 coders named a maximum of 5 objects each per image, the number of possible objects recorded as visually present in a scene ranged from 1 to 20. Figure 2 shows the frequency distribution of the number of objects per image. On average, images contained 8.63 objects (median = 9), which supports the long-held idea that the visual world is cluttered and that for most naming events there are multiple possible referents that a novice learner could consider.

In total, coders recorded 1,095 unique objects with a total of 97,407 object instances. Only 351 of these visual objects also occurred as object names in speech, and the reported analyses focus on these 351 objects that occurred in both modalities[2]. There were 72,446 total object instances for this smaller set. Figure 3a shows the proportion of images in which each object category appeared plotted against its rank frequency. As in Clerkin et al. (2017), visual objects occur in these natural scenes with a right skewed frequency distribution[3]. A small number of objects were pervasively present and a large number of objects occurred rarely with the 20 most frequent object categories (see table 3) accounting for 65.47% of all object tokens and the 37 most frequent object categories (that is, 10.5% of the 351 objects) accounting for 80.18% of all object tokens (see Figure 4).

Further, the AoA category of an object name is significantly related to the frequency of its corresponding visual object in the corpus. 97 of the visual objects reported by coders (that also appeared in the speech modality) were First objects and 253 were Later objects. Mann-Whitney-Wilcoxon tests were used to compare the frequencies of objects in these categories due to the non-normality of the

[2] All analyses follow the same statistics pattern when all 1,095 visual objects are analyzed.

[3] The distribution is referred to as right-skewed based on a histogram of the frequency distribution in which the placement of the most frequent objects is reversed on the x-axis as compared to Figure 3. We find the rank order plots better visualizations for our purposes.

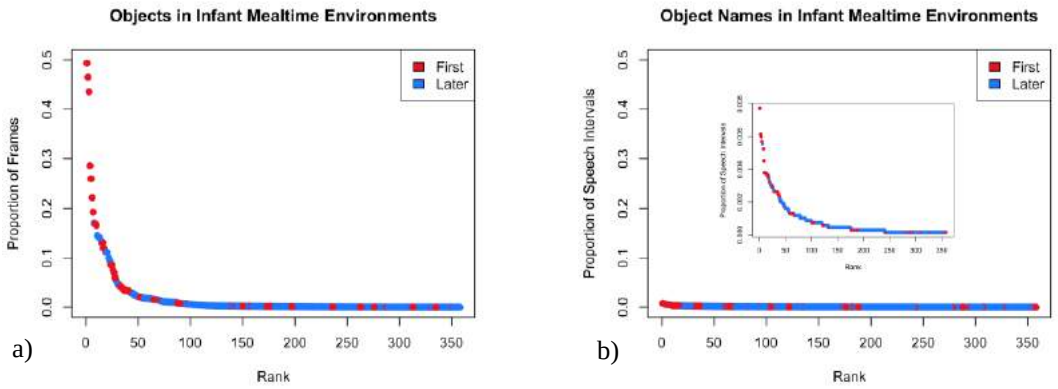


Figure 3. a) The proportion of images in which each visual object category appeared against its rank. b) The number of naming events for each object name as a proportion of the number of 5 second intervals containing any speech. Inset is the same information plotted with a smaller y axis.

data. Objects named by words in the First category (mean = 4.50%; Mdn = 0.55% of images) were significantly more frequent than objects named by Later words (mean = 0.75%; Mdn = 0.10% of images), U = 17609.5, $p < 0.0001$[4]. 11 of the 15 most frequent objects belonged to the First category. In sum, infants' visual experience during mealtime is dominated by a small set of objects named by very early learned words. These results suggest that day-in and day-out experience with these visual objects may be important for learning their names.

Object Names

Talk overall was extremely sparse in these mealtime scenes. Any speech, not just speech including object names, only occurred in 55.83% of the total video time (see table 1).

Table 1: Top visual objects, object names, and their AoAs.

Visual Objects	AoA
table	First
shirt	First
chair	First
window	First
bowl	First
cup	First
bottle	First
cabinet	Later
door	First
pants	First
picture	Later
counter	Later
tray	Later
spoon	First
toy	First

Object Names	AoA
egg	First
cheese	First
paper	First
book	First
camera	Later
water	First
juice	First
milk	First
paint	Later
spoon	First
table	First
dog	First
page	Later
plate	First
watch	First

Object names in speech occurred even more rarely; 117 mealtime events contained some speech but none of the target object names. The overall lack of talk and object names appears quite ordinary and typical when watching and listening to content these natural videos. These infants do not yet talk themselves, and the speech stream thus often contains terms of endearment and comments directed to the baby, talk between adults, and periods of silence as the parents and their infants go about their daily lives.

Nonetheless, 351 unique object names were said during mealtime activities across 1,941 naming events. It should be noted that only a small number of object names were said – about a third of those possible based on the list of visual objects. Figure 3b shows the number of naming instances as a proportion of the 5 second intervals containing any talk for each object name plotted against its rank frequency. Though the distribution of object talk is not uniform, it does not follow the pattern of extreme skewness as does the objects or might be predicted by the statistics of natural language more generally. The 40 most frequent object categories accounted for only 48.79% of all object name tokens, and the 123 most frequent object names (that is, 35.04% of all object names) were required to account for 80.06% of all object names tokens. Though object names do not appear equally frequently, there is not a clear set of object names that dominate talk about objects in this natural mealtime context. Note in Figure 4 the difference in the shapes of the curves for the proportions of unique visual objects and object names that account for all tokens.

A large proportion (97 out of 118) of the First words whose visual objects appeared in the images occurred in the auditory domain as well. Proportionally fewer of the possible Later objects had names that were said during mealtime; only 253 of the 977 Later object names were spoken in the corpus. As with the visual objects, object name frequency is significantly related to AoA. Object names from the First category (mean

[4] All reported p-values have been corrected for multiple comparisons using the Holm correction.

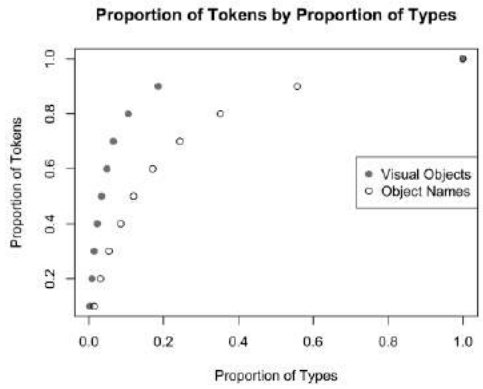


Figure 4. The proportion of types accounting for the proportion of tokens at intervals of 0.1.

= 0.14%; Mdn = 0.07% of speech intervals) were spoken more frequently than object names from the Later category (mean = 0.06%; Mdn = 0.03% of speech intervals), U = 16765.5, $p < 0.0001$. 12 of the 15 most frequent object names belonged to the First category. This, unsurprisingly, supports the idea that hearing objects names is important for learning them.

Correspondence and Co-occurrence

If the objects present most frequently in the visual environment were those whose corresponding names occur frequently in the environment, it would seem that the problem of breaking into learning first object names is solved. However, while there is a highly significant positive relationship between visual frequency and spoken frequency for object-name pairs, the relationship is very weak[5], τ_B = 0.17, $p < 0.0001$. As a demonstration, the 40 most frequent objects and the 40 most frequent object names only have 11

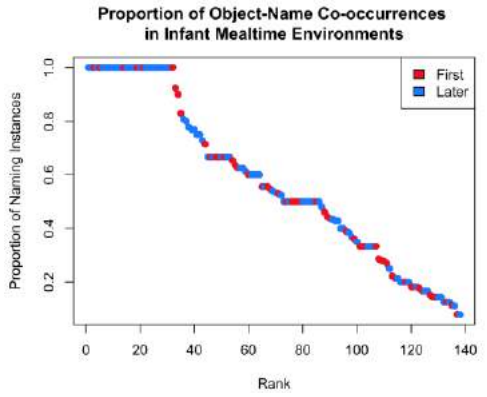


Figure 5. The proportion of naming instances in which a corresponding object was visually present. Only object-name pairs which ever co-occurred are shown.

items in common, and only 1 object name appears in the top 10 for frequency in both modalities. In sum, the pervasive visual objects are not named by words that are especially frequent in this context.

The number of co-occurrences between objects and their names even by our generous measure was very low. 213 of the 351 object-name pairs never occurred in the same 5 second interval. For the 138 that co-occurred at least once, the maximum number of co-occurrences was 34 (mean = 4.43; Mdn = 2). Because raw co-occurrence is so rare and the timescales of visual objects and spoken words are so different, we turned co-occurrence, reported here as the proportion of naming instances in which a corresponding object was visually present.

Figure 5 shows the co-occurrence proportion by rank order for the 138 object-name pairs that ever co-occurred. For co-occurrence, we do not find a right skewed frequency distribution but rather one that is bimodal. Further, co-occurrence proportion shows a statistical relationship with AoA that is opposite to those found for visual objects and object names individually. The co-occurrence proportion of the Later category (mean = 60.55%, Mdn = 62.95%) was significantly higher than that of the First category (mean = 48.83%, Mdn = 50%), U = 1662, $p < 0.01$. In fact, 26 Later objects co-occurred with their corresponding names 100% of the time whereas only 6 First objects did so. This result on its face seems surprising because there is a strong theoretical and empirical basis for the idea that co-occurrence is key to learning object-name mappings.

However, when the frequency of object names in the corpus are considered, it becomes clear why co-occurrence proportion was related to AoA in this direction. Co-occurrence proportion is in fact negatively correlated with word frequency, $\tau_B = -0.41$, $p < 0.0001$. This means that for many object names which were said perhaps only once, the corresponding visual objects were likely to be present during that naming instance. It makes sense that objects that are unusual in the context would be more likely to be present in the moment when those objects' names are said. For example, "fire extinguisher" (which is logically an unusual item for the mealtime context) was named once and the object was present, giving it a co-occurrence proportion of 1. This result suggests that it is important to consider not only the quantity of the co-occurrences (frequency) but also the quality of co-occurrence between object-names pairs (co-occurrence proportion) as it is unclear how much very young infants could learn from a single co-occurrence.

Quantity and Quality: Strength

To assess the quantity and quality of the co-occurrence of object-name pairs, we created a new compound measure of co-occurrence strength which was the proportion of naming instances during which the visual object was present - multiplied by the number of mealtime events in which both

[5] Kendall's rank correlation used instead of Pearson's product moment correlation due to the non-normality of the data.

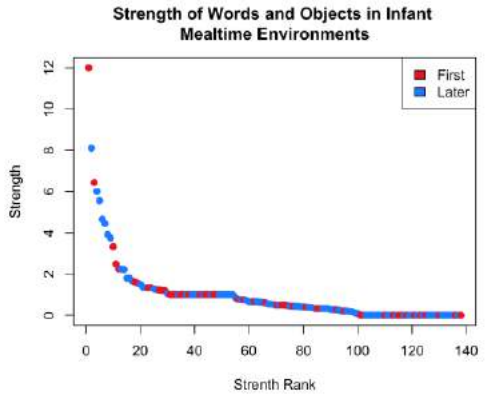


Figure 6. The strength of the co-occurrence of the object-word pair against its rank. Only object-name pairs which ever co-occurred are shown.

the visual object and the object name ever occurred. This measure allows us to approximate the potential learnability of an object-name pair.

Figure 6 shows the strength of the object-name pairing plotted against rank strength. Here we see again the skewed distribution with a small number of object-name pairings with relatively high strength values and a large number of object-name pairings with low strength values. We also again find a significant difference between strength of First object-name pairs (mean = 1.69, Mdn = 1) and Later object-name pairs (mean = 0.51, Mdn = 0.35), U = 3203, $p < 0.0001$. 13 of the 15 highest strength values belong to object-naming pairing for the First AoA category. This result suggests that the quality and quantity of co-occurrences between object-name pairs may be important for infants breaking into object name learning. Critically, the strength of the co-occurrence is underlain by the skewed frequency of visual objects.

Discussion

The results taken together do not support the hypothesis that many may have theorized: object names that occur frequently have pervasive referents, and these word-referent pairs occur frequently and simultaneously, thus providing a simple statistical solution for how infants can learn their first object names. Instead, the evidence of the present study suggests a different solution underlain by the pervasiveness of a few object categories in the visual environment. Object names that refer to visually pervasive objects may be said relatively rarely, but because the objects are visually pervasive, whenever the object name is said, the object is likely in the infants' view. The extremely skewed frequency distribution of objects in view in the mealtime context thus makes each naming event for those objects count – as demonstrated by our measure of co-occurrence strength.

Studies of the word-learning environment for children have typically focused on the frequency and diversity of the words (Hart, 1991; Montag et al., 2018). However, investigations of the natural environment including the visual domain are taking off with the advent of small, wearable cameras. Another recent at-home study which also examined the frequency of objects and their names in the natural environment found that the overall proportion of object-name co-presence predicts 6-month-olds' performance in an in-laboratory word comprehension task (Bergelson & Aslin, 2017). This result supports the idea that the statistical structure of the learning environment is directly related to word learning. Our results further suggest that the visual side of the learning problem specifically may be critical to the start of object name learning because it sets up the opportunities for learning moments.

The frequency distribution of visual objects during a particular context (mealtime in the present case) partitions potential referents into two potential classes for young learners –those that are typically present in this context and those that are not; classes that will be different for each context. Those that are persistently present provide a *selective* visual foundation to linking the objects to their referents.

The foundation for the *early* learning of object names may be contexts – such as mealtime, dressing, getting into the car – that occur day-in and day-out and are characterized by the same object categories repeatedly and pervasively present. These routines may bias the linking of even sparse naming events to those visually pervasive objects. Contexts that repeat in this way, along with the statistical structure of visual objects in those contexts, may be a critical contributing factor for early learning. This idea is consistent with the evidence on the value of repeatedly reading infants their favorite pictures books in supporting word learning (Horst, Parsons, & Bryan, 2011). For older children, the diversity of words in the learning environment may matter most for vocabulary development (Montag et al., 2018), but for the earliest learners, consistency of the visual content of repeated contexts may be the key.

In sum, the co-occurrence statistics of object names and their referents in the contexts comprising the early natural learning environment, as underlain by the extremely right skewed frequency distribution of visual objects, set up a clear signal in the noise which infants may use to learn their first object names.

Acknowledgements

We thank the families who participated in this study. We also thank Teagan Wilson, Remi Reich, Bryce Hockman, and Baker Nasser for in laboratory coding and our colleagues in the Cognitive Development Lab at Indiana University for helpful comments. This article was funded by NSF grant BCS-15233982, by NIH grants R01HD 074601, R01HD 28675, T32HD007475, and by Indiana University through the Emerging Area of Research Initiative - Learning: Brains, Machines, and Children.

References

Bergelson, E., & Aslin, R. N. (2017). Nature and origins of the lexicon in 6-mo-olds. *Proceedings of the National Academy of Sciences*, *114*(49), 12916-12921.

Bergelson, E., & Swingley, D. (2012). At 6–9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences, 109*(9), 3253-3258.

Cartmill, E. A., Armstrong, B. F., Gleitman, L. R., Goldin-Meadow, S., Medina, T. N., & Trueswell, J. C. (2013). Quality of early parent input predicts child vocabulary 3 years later. *Proceedings of the National Academy of Sciences, 110*(28), 11278-11283.

Clerkin, E. M., Hart, E., Rehg, J. M., Yu, C., & Smith, L. B. (2017). Real-world visual statistics and infants' first-learned object names. *Phil. Trans. R. Soc. B, 372*(1711), 20160055.

Fenson, L., Bates, E., Dale, P. S., Marchman, V. A., Reznick, J. S., & Thal, D. J. (2007). *MacArthur-Bates communicative development inventories*.

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2016). Wordbank: An open repository for developmental vocabulary data. *Journal of child language*, 1-18.

Hart, B. (1991). Input frequency and children's first words. *First Language, 11*(32), 289-300.

Horst, J. S., Parsons, K. L., & Bryan, N. M. (2011). Get the story straight: Contextual repetition promotes word learning from storybooks. *Frontiers in Psychology, 2*, 17.

Jayaraman, S., Fausey, C. M., & Smith, L. B. (2015). The faces in infant-perspective scenes change over the first year of life. *PloS one, 10*(5), e0123780.

Montag, J. L., Jones, M. N., & Smith, L. B. (2018). Quantity and diversity: Simulating early word learning environments. *Cognitive Science, 42*, 375-412.

Sadeghi, Z., McClelland, J. L., & Hoffman, P. (2015). You shall know an object by the company it keeps: An investigation of semantic representations derived from object co-occurrence in visual scenes. *Neuropsychologia, 76*, 52-61.

Salakhutdinov, R., Torralba, A., & Tenenbaum, J. (2011). *Learning to share visual appearance for multiclass object detection.* Paper presented at the Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on.

Sandhofer, C. M., Smith, L. B., & Luo, J. (2000). Counting nouns and verbs in the input: Differential frequencies, different kinds of learning? *Journal of child language, 27*(3), 561-585.

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition, 106*(3), 1558-1568.

Tamis-LeMonda, C. S., Custode, S., Kuchirko, Y., Escobar, K., & Lo, T. (2018). Routine Language: Speech Directed to Infants During Home Activities. *Child development*.

Datavyu Team. (2014). Datavyu: A video coding tool. Databrary Project, New York University. URL http://datavyu.org.

Vlach, H. A., & Johnson, S. P. (2013). Memory constraints on infants' cross-situational statistical learning. *Cognition, 127*(3), 375-382.

Yurovsky, D., Smith, L. B., & Yu, C. (2013). Statistical word learning at scale: The baby's view is better. *Developmental Science, 16*(6), 959-966.

Frequency Effects in Decision-Making are Predicted by Dirichlet Probability Distribution Models

Astin Cornwall, Hilary Don, & Darrell Worthy
Texas A&M University, Texas, USA

Abstract

Frequency of reward and average reward value are two types of reward information we utilize when making decisions between two alternative options. Often, these two pieces of information coincide with the highest value option, however, when a slightly less valuable option is presented more frequently, standard reinforcement learning models such as the Delta model can make incorrect predictions. This paper explores the discrepancy in these predictions by way of simulating relevant behavioral tasks with the Delta model, the Decay model, and a novel Bayesian model based on the Dirichlet distribution. We then compare model predictions to behavioral data from some of the same tasks that were simulated. The Delta model provides a poor fit to the data for each of the three presented tasks when compared to the Decay model and the two Bayesian learning models, because it predicts a bias toward options with higher average reward, while the Decay and Bayesian models predict a bias toward reward frequency. The Decay and Bayesian models show a distinct similarity in prediction and fits to the data for most of the tasks. This is because both models predict a bias toward reward frequency rather than average reward magnitude, despite different computational formalisms. However, we also note some interesting discrepancies between the Decay and Bayesian models which will show that in some cases, the frequency of reward may be more important than the reward value.

Keywords: Frequency Effect; Reinforcement Learning; Bayesian Learning

Introduction

A wide variety of decisions we make on a day-to-day basis are repetitive in the sense that we may choose one option over another fairly consistently. Whether these decisions are about choosing name-brand over store-brand items, restaurant A or B, or taking the freeway versus the side roads to work, it's possible that all of these decisions are computed by common algorithmic mechanisms. These decisions could be based on the average outcome of each option, for example, taking the freeway to work is nearly always faster than taking the side roads. However, supposing that a new bypass opens that is predicted to greatly reduce travel time, a person may still be inclined to choose the freeway since they have had many more experiences with the freeway being adequate enough.

Learning rules in formal models of cognition allow us to make sense of human decision-making processes and get a glimpse as to why people make the decisions they do in situations such as the examples above. In this paper, we compare the choice predictions of four learning models: the Delta rule, the Decay rule, and two Dirichlet distribution-based models, on a set of decision-making tasks.

The Delta rule, in particular, is a widely used learning rule across many domains of cognition. This model predicts that people will have a preference for options that have the greatest expected value, based on representations of the *average* reward for each option, amongst alternative options (e.g. Busemeyer & Stout, 2002; Daw et al., 2006; Gluck & Bower, 1988; Jacobs, 1988; Rescorla & Wagner, 1972; Rumelhart & McClelland, 1986; Sutton & Barto, 1981; 1998; Widrow & Hoff, 1960; Williams, 1992).

In contrast to the Delta rule's average value representations, the Decay rule learns to represent the *cumulative* value of each option based on the frequency with which it has been rewarded. Psychologically, the Decay model assumes that that decision outcomes are stored in working memory and decay over time. The Decay model utilizes a decay parameter which diminishes the expected value of each option at each timepoint. Therefore, the option with the greatest expected value in this model would be the option which is most frequently rewarded; in most cases (Erev & Roth, 1998; Yechiam & Busemeyer, 2005; Yechiam & Ert, 2007).

In a departure from these two standard learning models, this paper presents a Bayesian model which simply learns how many times each option has a positive outcome rather than learning expected values. The Dirichlet Probability Distribution (DPD) model holds in memory a representation of how many times each option has produced a reward, regardless of value. Each of these values are used as the concentration parameter values in the distribution which allocates more probability mass to the options which have been rewarded most frequently. Thus, when attempting to choose between options, the option more frequently rewarded will have a higher probability of being chosen.

The sole use of the Dirichlet distribution as the base for this model may seem atypical considering it is more often used in Bayesian data analysis for determining the clustering, or categorization, of data (e.g. Griffiths, Sanborn, Canini, & Navarro, 2008), or in Dirichlet Process or Mixture models (Navarro, Griffiths, Steyvers & Lee, 2006; Gershman & Blei, 2012; Sims, Neth, Jacobs, & Gray, 2013), or as the prior for

another Bayesian model. However, choice outcomes and options map nicely onto the Dirichlet distribution concentration parameter and categories respectively. Simply, the categories are effectively predetermined by the number of choices and the probability mass for each category is distributed as a function of the number of rewarding observations.

As an attempt to design a Bayesian analog to the Decay model, the DPD model was extended to include a decay parameter. The Dirichlet Probability Distribution Decay (DPD-Decay) model decays the memory representations of the total number of rewarded outcomes at each timepoint. Critically, this means that additional uncertainty is introduced into the probability distribution. As the memory of rewarded outcomes for each option tends towards 0, all options would have an equiprobable chance of being selected.

While Bayesian models have been criticized in prior research for being simple vote-counting models (Jones & Love, 2011), it's possible that, if each rewarding event is considered a vote, the DPD model will predict similar behavior as the Decay model. This could allow the DPD model to predict a bias toward frequency of reward rather than average reward magnitude, and recent work suggests that reward frequency exerts a larger effect on behavior than average reward magnitude (Worthy, Otto, Cornwall, Don & Davis, 2018). Thus, DPD models with sparse priors may represent a cognitive process of predicting the probability of a rewarding event, based solely on reward frequency. Our goal in the present work is to verify these predictions and examine the degree to which they are consistent with human behavior.

Difference Between Models

The key difference between the Delta model and the reward frequency models (Decay and Dirichlet) is in how each type of model utilizes reward information to make predictions about future choices. The Delta model uses average reward information whereas the Decay and Dirichlet models utilize the frequency of rewards to formulate a cumulative representation of reward. This is important as, per Estes (1976), probability judgements about the choices are heavily influenced by the frequency that each option produces a reward, rather than the average reward value. As such, it would be expected that the when rewarding options are shown in disproportionate frequencies, the predictions of the Delta model and the Decay and Dirichlet models will diverge. It would be expected that tasks which consist of rewards of varying frequency and value would show differences in each models' predictions.

To ascertain the general predictions of each model, and determine the differences therein, three tasks which have previously examined the effect of reward frequency and value were selected to be simulated using each model. To verify the predictions made by each model and task combination, each of the models were fit to human data collected from each of the three tasks.

Experimental Tasks

Iowa Gambling Task. The Iowa Gambling Task (IGT; Bechara, Damasio, Damasio, & Anderson, 1994) allows four options to be chosen from, each with their own reward schedule over the course of 100 total trials. The reward schedule for the IGT can be found in Table 1 below. Traditionally, the task consists of two options which result in a net loss of points, and two options which results in a net gain. Options A and B offer participants larger rewards on gain trials, but also larger losses on loss trials resulting in an overall net loss for both options. In contrast, Options C and D give smaller rewards and losses resulting in an overall net gain for these two options. Within each 10-choice block for each option, the frequency of gains differs between options. Options A and C show infrequent gains relative to Options B and D which are more consistent. Strictly looking at the net positive options, Options C and D should be the favored decks. However, as Bechara et al. observed, there is a preference for choosing Options A and B which have a higher frequency of larger rewards, but results in a net loss of points.

	A		B		C		D	
Trial	**IGT**	**SGT**	**IGT**	**SGT**	**IGT**	**SGT**	**IGT**	**SGT**
1	100	200	100	100	50	-200	50	-100
2	100	200	100	100	50	-200	50	-100
3	-50	200	100	100	0	-200	50	-100
4	100	200	100	100	50	-200	50	-100
5	-200	-1050	100	-650	0	1050	50	650
6	100	200	100	100	50	-200	50	-100
7	-100	200	100	100	0	-200	50	-100
8	100	200	100	100	50	-200	50	-100
9	-150	200	-1250	100	0	-200	50	-100
10	-250	-1050	100	-650	0	1050	-250	650
Net	**-250**	**-500**	**-350**	**-500**	**250**	**500**	**200**	**500**

Table 1: Reward schedules for both the IGT and SGT by Option Letter. This reward schedule is repeated over the total 100 trials.

Soochow Gambling Task. The Soochow Gambling Task (SGT; Chiu et al., 2008) is a task similar in procedure to the IGT aside from a change in the reward schedule of each option. The reward schedule for each option in the SGT can be found in Table 1 below. Similar to the IGT, over the course of 100 trials, participants are able to select one of four options. Options C and D are still the options with an overall net reward gain, and likewise with Options A and B having a net reward loss. Both Options A and B offer participants consistent gains of 200 or 100 points, respectively, followed by a large loss which results in a net loss for both options. Inversely, Options C and D show consistent losses followed by a large gain resulting in a net gain. The gains and losses shown in Options A and B are exactly opposite in terms of sign. Where A and B show consistent rewards followed by a large loss. Importantly, the best options according to overall gain are also the options with the most consistent losses.

Similar to the IGT, there is a large preference for Options A and B indicating that frequency of rewards, despite losses, is a good predictor of choice preference(Byrne & Worthy, 2016).

Binary Choice Task. This task, as presented by Worthy et al. (2018), assesses the effect of reward frequency in a different manner. The task consists of four options, A, B, C, and D, where each have a respective probability of giving a reward of .65, .35, .75, .25. The possible rewards for this task were binary in that the reward totals were either 1 or 0. The task pairs Options A and B, and Options C and D, together and presents them randomly interspersed during training. Importantly, there are 100 AB trials and 50 CD trials which creates a situation where frequency of reward and average reward are in opposition if it is learned that Option A and C are the most rewarding within the respective pairs. The task then consists of 25 transfer trials for each of the remaining pairs of A, B, C, and D and bars further reward feedback. Worthy et al. observed that human participants were more likely to prefer Option A over C on AC pairing trials indicating that despite having a smaller average reward, option A is preferred over option C because of more frequent, and therefore higher cumulative reward.

Method

Model Formalisms

The Delta and Decay rule used in this paper are identical to those described in Worthy et al. (2018). Reward (*r*) and the expected value (*EV*) is calculated for each *j* option on each *t* trial. The Delta rule is described in Equation 1 as:

$$EV_{j,t+1} = EV_{j,t} + \alpha \cdot (r_t - EV_{j,t}) \cdot I_j \quad (1)$$

Where I_j is a variable which indicates option choice via a value of 1 if *j* option is chosen on trial t, and 0 otherwise. This formulation ensures that only the expected value for the chosen option is updated, and the other options, whether seen or not, are not updated. Alpha (α) is denoted as a learning rate parameter where $\alpha \in (0,1)$. For the Delta model in particular, α modifies the $(r_t - EV_{j,t})$ prediction error by giving greater weight to more recent outcomes with higher α values, and lower α values giving less weight to recent outcomes and producing little change in the expected value on each trial.

Similarly, the Decay model tracks changes in expected value, but instead of updating the expected value by way of a prediction error the raw reward value is used. However, this does not mean that expected value consistently increases for each chosen option. On each trial, each *j* option will be modified by a decay parameter (*A*; $A \in (0,1)$) regardless of whether the *j* option was seen or chosen. Critically, this means that the expected value for each option will decay over time and only increase when a reward for that option is received. Thus, the more frequent the reward, the greater the expected value. The formula for computing the change in expected is described below in Equation 2:

$$EV_{j,t+1} = EV_{j,t} \cdot A + r_t \cdot I_j \quad (2)$$

As mentioned above, the DPD model focuses solely on the number of times each *j* option is rewarded (*r*) and uses that information to update a Dirichlet probability distribution. Simply, a Dirichlet distribution takes *k*, the total number of *j* options, and their respective number of rewarded trials (γ_j) and produces a probability density (x_j) for each *j* option where $x_j \in (0,1)\ and\ \sum_{j=1}^{k} x_j = 1$. In other words, the updating of the distribution occurs in two steps as described in Equations 3 and 4:

$$\gamma_{j,t+1} = \gamma_{j,t} + r_t \cdot I_j \quad (3)$$

$$f(x_{1,t+1} \ldots x_{k,t+1} | \gamma_{1,t+1} \ldots \gamma_{k,t+1}) = \frac{1}{B(\gamma)} \prod_{j=1}^{k} x_{j,t}{}^{\gamma_{j,t}-1} \quad (4)$$

$$\text{where } B(\gamma) = \frac{\prod_{j=1}^{k} \Gamma(\gamma_j)}{\Gamma(\sum_{j=1}^{k} \gamma_j)}$$

On each *t* trial, the reward value for one option is added to the chosen option which will distribute slightly more probability density to the chosen option. To determine choice with this model, a random sample is taken from the Dirichlet distribution which results in a simplex, or a vector of probabilities which sum to 1. Critically, this implies that as one option is rewarded more frequently, the probability value sampled from the distribution will tend to be of greater value, and thus the option is more likely to be chosen. Taking a single sample, rather than integrating over the posterior, was a decision made with the assumption that this would better reflect human performance as the beliefs surrounding each option is uncertain. As more information is learned about an individual option, the belief about the positive outcomes of that option will become more certain, and thus the probability of choosing that outcome will be more consistent.

An extension of the DPD model presented above, the DPD-Decay model includes the decay parameter (*A*) which decays the total number of rewarded trials (γ_j) for each option on each trial similar to how the Decay model functions. By decaying the rewarded trial values, the model increases the amount of uncertainty and allows a greater range of possible values to be randomly sampled. This also implies that the more frequently an option is seen the more likely it is to overcome the consistent decay, such that it is granted more probability density over time. Expressly, the decay parameter in this equation will weigh the model for or against more recent outcomes. In Equation 5, $\gamma_{j,t+1}$ is computed for every *j* option and are subsequently inserted into Equation 4.

$$\gamma_{j,t+1} = \gamma_{j,t} \cdot A + r_t \cdot I_j \quad (5)$$

For the Delta and Decay models, the predicted probability that any given option *j* is chosen *C* on a particular trial *t*, P($C_{j,t}$), is calculated by way of a Softmax choice function shown in Equation 6 below:

$$P\left|C_{j,t}\right| = \frac{e^{\beta \cdot EV_{j,t}}}{\sum_{1}^{N(j)} e^{\beta \cdot EV_{j,t}}} \quad (6)$$

Like the Yechiam & Ert (2007) Softmax application used in Worthy et al. (2018), $\beta = 3^c - 1; c \in (0,5)$, where c is an inverse temperature parameter which dictates how often the option with the higher expected value is chosen. When c approaches 0, choices are more random. Inversely, choices are weighted more heavily towards the options with the

highest expected value as c approaches 1. Simply, this choice function determines the probability of choice by computing the proportion of the scaled chosen option divided by the sum of the scaled choice and alternative choices.

Simulation and Behavioral Methods

For each task, 10000 simulated participant datasets were created with randomized model parameters of α, A, and c, for applicable models, for each participant. Each of these parameters were drawn from a uniform distribution: $U(0,1)$ for learning and decay rates, and $U(0,5)$ for the inverse temperature parameter. These parameters were kept consistent across models within each simulation, but each model ran independently in regard to the choices made and corresponding output. The output for each of these simulations was the probability of choosing each outcome, the expected value of each option, and the choices made on each trial.

For each task, human behavioral data was collected from an undergraduate population with sample sizes of ~50 for each task. Each participant completed the experiment in a Psychtoolbox 2.54 environment on a Windows computer running Matlab. The general procedures used in the simulations were identical to the computerized version of the tasks that participants completed, however graphical and counterbalancing considerations were needed for real participants that are detailed below for each experiment.

In both the IGT and SGT, the options were displayed onscreen as a deck of cards, each with their own random color. The onscreen location of each individual deck was displayed from left to right in a random arrangement of Options A-D for each participant. Upon selecting a deck, the participant would be shown the card being overturned and the amount of reward. Additionally, participants were given a set amount of points in an onscreen bank that would increase or decrease depending on the outcome.

For the Binary Choice Task, each of the four options were randomly assigned a fractal image randomly drawn from a pool of 12 images. Like the IGT and SGT, the order of the 4 selected images were randomly arranged on screen from left to right. However, Options AB and CD were always together as a pair, but the order of each pair varied for each participant. As an example, some potential orderings of the option could include: ABCD, CDAB, BACD, etc. Each selection of an option showed the option turning over to reveal the outcome of that trial. Importantly, and consistent with the simulations, reward feedback only occurred during the initial 150-trial training phase, but the transfer phase, participants were only shown a gray outline around the option they chose instead of the point value they would have seen on the training trials.

Results

Simulation Output

For the IGT and SGT, the simulation metric that will be reported is the overall performance on the task as computed by subtracting the sum of the net loss options from the sum of the net gain options: (A+B)-(C+D). For both the IGT and SGT, the performance of each model, and the actual participant data for comparison, is plotted over all 100 trials in Figures 1 and 2. In the IGT the Delta model was more likely to choose the net gain options over the more frequently rewarding net loss options. The Decay model also showed a preference for the net gain options overall. Both the DPD and DPD-Decay models showed no preference for either the net gain or loss options, but this behavior also seems to be reflected, albeit slightly, by the actual participant data which rapidly varies in preference for either the net gain or loss options over time.

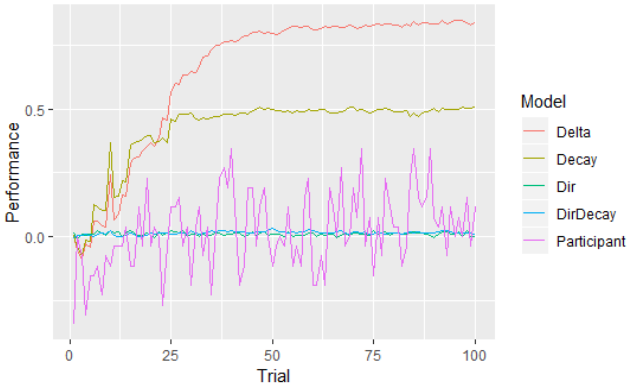


Figure 1: Average performance on the IGT by model and actual participant data.

In the SGT, the Delta model again showed a preference for the net gain options, but the Decay model now shows behavior that greatly reflects the behavior shown by actual participants. Both the human and simulated Decay model datasets showed an initial preference for the net loss options, but over time began to tend towards the net gain options which is consistent with prior research as previously discussed. The DPD and DPD-Decay model again showed similar results, but in the SGT, they show a large preference for the more frequently rewarding net loss options.

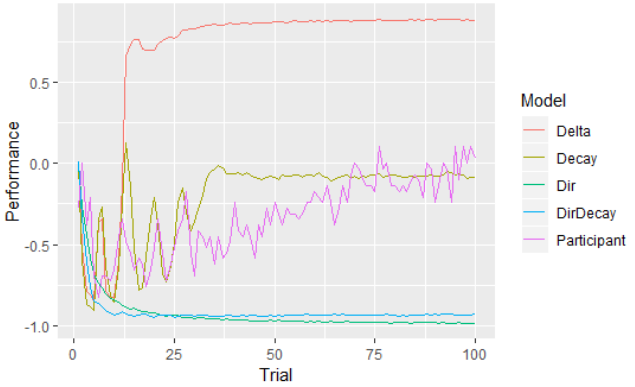


Figure 2: Average performance on the SGT by model and actual participant data.

For the Binary Choice Task, as shown in Figure 3A-B, each model was able to learn that there is a more rewarding option in each option pair. However, the rate at which the most rewarding, or best, option was identified and overall

preference for the best option differed between models. The Delta model showed the greatest preference for the best options out of the four models, followed by the DPD, Decay, and DPD-Decay. When solely learning which option has the largest average reward, it is no surprise that the Delta model outperforms the other three models. However, when looking at the choice predictions for the remaining option pairs, as shown in Figure 3C, a difference between the models emerge. The Delta model predicts more C choices, whereas the Decay, DPD, and DPD-Decay models all predict more A choices. The remaining option pairs showed relatively similar predictions since there was not as big of a discrepancy between an options' expected value and number of observations. The large peaks in the DPD model are indicative of the frequency of outcome observations for each pair. The more outcomes observed, the more likely the model will choose the same option.

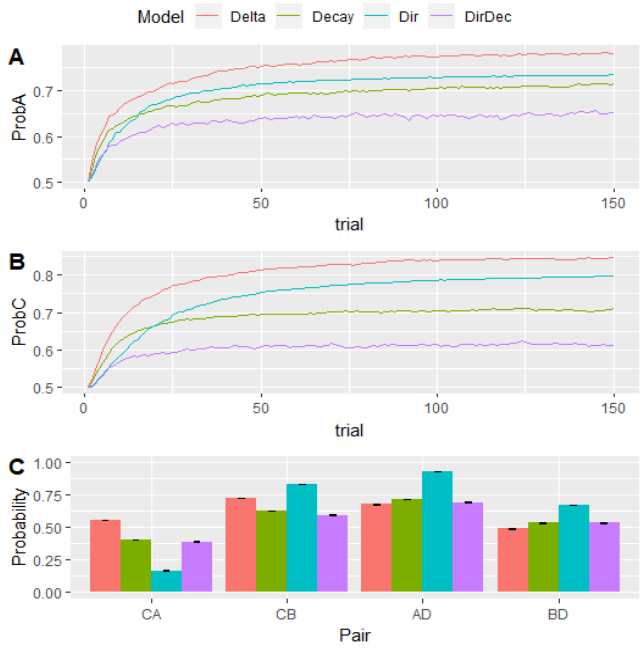


Figure 3: A and B show the probability of choosing the best option, either Option A or C respectively, over the course of 150 trials. C shows the predicted probability of choosing the best options if the simulated participant were to see the remaining pairings of options.

Behavioral Fits and Comparisons

Participants were independently recruited for one of the three tasks from an undergraduate sample. Each participant was reimbursed for their time with partial completion of course credit. For each task we recruited comparable sample sizes: 52 participants for the IGT; 58 participants for the SGT; 50 participants for the Binary Choice Task.

Each of the models were directly fit to the behavioral data by maximizing the likelihood of each model via the 'optim' function with a 'L-BFGS-B' method in R. The decay and inverse temperature parameters were included as free parameters for the respective models. The Delta and Decay models utilized two free parameters while the DPD-Decay model used only the decay parameter. No free parameters were used in the DPD model for the IGT and SGT. When fitting the Binary Choice Task data alone however, the DPD and DPD-Decay model included an inverse temperature parameter. In this task, for both models, the probability simplex was drawn from the Dirichlet distribution, as previously discussed, but the values relevant to the two observed options were used in the softmax function to compute a choice probability for each option which summed to 1.

A Bayesian Information Criterion (BIC; Schwarz, 1978) value was computed for each individual participant within each model and used to calculate the average BIC and subsequent BIC differences between each model. BIC was calculated by calculating the deviance of the model and adding additional error based on the number of free parameters k and number of trials t: $-2ln(L) + (k \cdot ln(t))$. Lower BIC values indicate a better fit to the behavioral data. As per Wagenmakers (2007), the BIC difference between the models can additionally be used to calculate a Bayes Factor which would show evidence for one model over another: $BF_{10,Model1}=exp((BIC_{model2}-BIC_{model1})/2)$.

Table 2 below details the BIC values of each model for each task along with the best fitting parameters for each model. For the IGT and Binary Choice Task, the Decay model shows an advantage over the other models. For the IGT, the next best fitting model was the DPD with a BIC of 268.9 which is shown to be significantly different from the Decay model with a Bayes Factor (BF) of 3.33. BFs with values greater than 3, or less than 1/3, are believed to have adequate evidence to reject the null hypothesis that the models are equal. The Decay model in the SGT was the next best fitting model behind the DPD-Decay model with BIC values of 269.8 and 267.8 respectively. This difference, with a BF of 2.7416, shows that both models are similar in their fits of the SGT data. In the Binary Choice Task, the Decay model BIC (279.7) is closely followed by both the DPD and DPD-Decay models; 282.8 and 280.5 respectively. The difference between the Decay and DPD model is significant with a BF of 5.1984, but there is not enough evidence to say that the Decay and DPD-Decay models are different, BF = 1.5115.

Table 2: Average Model Values

		Best a or A	Best c	BIC
IGT	Delta	.1009	.3756	278.0677
	Decay	**.6857**	**.00538**	**266.4658**
	DPD	N/A	N/A	268.8740
	DPD-D	.0218	N/A	273.8372
SGT	Delta	.4613	.3564	274.4863
	Decay	.5268	.0019	269.7714
	DPD	N/A	N/A	282.9299
	DPD-D	**.1454**	**N/A**	**267.7543**
Binary	Delta	0.3821	1.5120	296.2498
	Decay	**0.1765**	**0.4978**	**279.7178**
	DPD	N/A	1.3088	282.8315
	DPD-D	0.8770	1.5673	280.5440

It was also of interest to determine the proportion of participants whose data were best fit by each model. To do this, each non-redundant combination of models for each task was examined to figure how many participants' data were best fit by each model. Table 3 presents the proportion value for each model combination by task. The first model listed in the pair is the reference model. Values shown in bold represent that the reference model was the best fitting model of the pair.

As expected from the data in Table 2 for the IGT, the Decay and DPD model best fit the largest proportions of participants, and the DPD model showed the best fit overall. For the SGT, despite showing a large average BIC value, the Delta model showed a better fit slightly more participants than the Decay model, but not the DPD-Decay model. Additionally, the Decay model, rather than the DPD-Decay model, was the best fitting model for most participants. In the Binary Choice Task, the DPD models better fit more participants than the Decay model which showed the better fit on average. The DPD model showed the overall highest proportion best fit on this task as well.

Table 3: Proportion Best Fit

By Model	**IGT**	**SGT**	**Binary**
Delta<Decay	.44	**.54**	.38
Delta<DPD	.06	**.54**	.30
Delta<DPD-D	.44	.48	.38
Decay<DPD	.33	**.66**	.40
Decay<DPD-D	**.71**	**.62**	.44
DPD<DPD-D	**.83**	.47	**.70**
Overall	**IGT**	**SGT**	**Binary**
Delta	.05	.24	.26
Decay	.24	**.34**	.16
DPD	**.53**	.22	**.44**
DPD-D	.07	.19	.14

Discussion

The simulations and experiment presented in this paper examined the influence of reward frequency and probability on choices made in a decision-making task. Four models were compared that made both convergent and divergent predictions about which option was more valuable in three tasks which examine the effect of reward frequency. As similarly described by Worthy et al. (2018), there were divergent simulation predictions between the Delta rule and the reward frequency models where the Delta rule more often chose the options with the higher value rewards, whereas the reward frequency models, the Decay, DPD, and DPD-Decay, tended to choose the options which resulted in the most frequent rewards. The data from the experimental tasks showed that human participants more often chose the more frequent options in most cases. This behavior is in support of the predictions of all three of the reward frequency models. This is shown in which models where the best fitting model on average. For all three tasks, the best fitting model was a model which attended more towards the frequency of reward rather than the average value of reward. However, there also seems to be some individual differences in people who attend more towards average reward value instead of the frequency of reward. This can best be seen when looking at the SGT and Binary Choice Task. For both of these tasks, there was a sizable subset of participants who were best fit by the Delta model than the other three models.

There also exists some important differences in the reward frequency models despite their similarities. One of which is between the DPD and DPD-Decay models and the Decay model. When looking at Figure 1, the performance values for the DPD and DPD-Decay model are fairly constant about 0. This is most likely due to how the Dirichlet models compute reward. These models do not consider reward value, only the observation of a reward. Looking back to the reward schedules for the IGT, one net gain and one net loss option have fairly frequent rewards. With how the performance calculation considers the number of choices, and how the Dirichlet models determine choice by the number of observed rewards, you can begin to see how the number of net gain and loss choices would be about equal, and thus result in a performance of ~0. This can also be seen in the simulation of the SGT as well in Figure 2. The two Dirichlet models show an overwhelming preference for the net loss options. Again, looking at the reward schedule, the net loss options are the only options that have a frequent occurrence of reward as the net gain decks only give a reward every 5 successive picks. These two Dirichlet models may aid in making sense of the "Deck B" phenomenon in the SGT where people tend to choose the net loss options since the reward most frequently. However, the average fit for the DPD model was quite large. Which suggests that pure frequency of reward is not entirely predictive of choice on the SGT. With the DPD-Decay model showing the best average fit, this suggest that the frequency of reward is predictive, but that the overall representation of the total number of rewarded outcomes decays over time.

For the DPD model in particular, another difference be seen in Figure3C with the large peaks in the option pair predictions relative to the other models. Like detailed for the IGT and SGT, these peaks can be explained by looking at the rate of reward and frequency of observing the option pair. For these option pairs the best option is the one that is either the most frequently seen and/or rewarded. Thus, the model would be more likely to choose these options.

However, this also ties in to the major conclusion of this paper, that despite not utilizing any reward information, these Dirichlet models are able to fit human behavioral data on three tasks relatively well solely using a count of rewarding outcomes. Generally, choice selection may depend on reward value when all other factors are equal, but if rate of reward changes or if there is knowledge of number of previously rewarding outcomes, frequency of reward may take precedence over reward value. Though, like shown by the proportion of best fitting models, there may be a subset of people who focus on the overall reward value regardless of the frequency of the outcomes.

References

Annis, J., & Palmeri, T. J. (2018). Bayesian statistical approaches to evaluating cognitive models. *Wiley Interdisciplinary Reviews: Cognitive Science*, *9*(2)

Busemeyer, J.R., & Stout, J.C. (2002). A contribution of cognitive decision model to clinical assessment: Decomposing performance on the Bechara Gambling Task. *Psychological Assessment, 14,* 253-262.

Bechara, A., Damasio, A. R., Damasio, H., & Anderson, S. W. (1994). Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition*, *50*(1–3), 7–15.

Byrne, K. A., & Worthy, D. A. (2016). Toward a mechanistic account of gender differences in reward-based decision-making. *Journal of Neuroscience, Psychology, and Economics*,*9*(3–4),157–168.

Chiu, Y.-C., Lin, C.-H., Huang, J.-T., Lin, S., Lee, P.-L., & Hsieh, J.-C. (2008). Immediate gain is long-term loss: Are there foresighted decision makers in the Iowa Gambling Task? *Behavioral and Brain Functions*, *4*(1), 13. https://doi.org/10.1186/1744-9081-4-13

Daw, N., O'Doherty, J., Dayan, P., Seymour, B., & Dolan, R. (2006). Cortical substrates for exploratory decisions in humans. *Nature, 441,* 876-879.

Erev, I., & Roth, A.E. (1998). Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *American Economic Review*, *88*, 848-881.

Estes, W.K. (1976). The cognitive side of probability learning. *Psychological Review, 83,* 37-64.

Gershman, S. J., & Blei, D. M. (2012). A tutorial on Bayesian nonparametric models. *Journal of Mathematical Psychology*, *56*(1), 1-12.

Gluck, M.A., & Bower, G.H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General, 128,* 309-331.

Griffiths, T. L., Kemp, C., & Tenenbaum, J. B. (2008). Bayesian models of cognition.

Griffiths, T. L., Sanborn, A. N., Canini, K. R., & Navarro, D. J. (2008). Categorization as nonparametric Bayesian density estimation. *The probabilistic mind: Prospects for Bayesian cognitive science*, 303-328.

Jacobs, R.A. (1988). Increased rates of convergence through learning rate adaptation. *Neural Networks, 1,* 295-307.

Jacobs, R. A., & Kruschke, J. K. (2011). Bayesian learning theory applied to human cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*(1), 8-21.

Jones, M., & Love, B. C. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences*, *34*(4), 169-188.

Lee, M. D., & Wagenmakers, E. J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge university press.

Navarro, D. J., Griffiths, T. L., Steyvers, M., & Lee, M. D. (2006). Modeling individual differences using Dirichlet processes. *Journal of mathematical Psychology*, *50*(2), 101-122.

Otto, A. R., & Love, B. C. (2010). You don't want to know what you're missing: When information about forgone rewards impedes dynamic decision making. *Judgment and Decision Making*, *5*(1), 1–10.

Pang, B., Blanco, N. J., Maddox, W. T., & Worthy, D. A. (2017). To not settle for small losses: evidence for an ecological aspiration level of zero in dynamic decision-making. *Psychonomic bulletin & review*, *24*(2), 536-546.

Rescorla, R.A., & Wagner, A.R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. In A.H. Black & W.F. Prokasy (Eds.) *Classical conditioning II: Current research and theory.* New York: Appleton-Century-Crofts.

Rumelhart, D.E., McClelland, J.E. & the PDP Research Group. (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition,* vols. 1 and 2. Cambridge, MA: MIT Press.

Sims, C. R., Neth, H., Jacobs, R. A., & Gray, W. D. (2013). Melioration as rational choice: Sequential decision making in uncertain environments. *Psychological Review, 120*(1), 139.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, *6*(2), 461-464.

Wagenmakers, E.J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review, 14,* 779-804.

Widrow, B., & Hoff, M.E. (1960). Adaptive switching circuits. *1960 WESCON Convention Record Part IV,* 96104.

Williams, R.J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning, 8,* 229-256.

Worthy, D. A., Otto, A. R., Cornwall, A. C., Don, H. J., & Davis, T. A Case of Divergent Predictions Made by Delta and Decay Rule Learning Models. *Proceedings of the Cognitive Science Society.*

Yechiam, E., & Busemeyer, J.R. (2005). Comparison of basic assumptions embedded in learning models for experiencebased decision-making. *Psychonomic Bulletin & Review, 12,* 387-402.

Yechiam, E. & Ert, E. (2007). Evaluating the reliance on past choices in adaptive learning models. *Journal of Mathematical Psychology, 51,* 75-84.

Differences in learnability of pantomime versus artificial sign: Iconicity, cultural evolution, and linguistic structure

Tania Delgado (tadelgad@ucsd.edu)
Cognitive Science Department, 9500 Gilman Drive
La Jolla, CA 92093 USA

Seana Coulson (scoulson@ucsd.edu)
Cognitive Science Department, 9500 Gilman Drive
La Jolla, CA 92093 USA

Abstract

One of the central goals of language evolution research is to explain how systematic structure emerges. A cultural evolutionary approach proposes that the systematic structure of language arises from the use and transmission of language. Motamedi and colleagues (2016) investigated the influences of these forces on the evolution of language by generating an artificial sign language in the lab. Over several generations of new learners and their interactions, an initially unsystematic set of silent gestures developed markers for functional categories of person, location, object, and action. Here we describe results of two studies that compared the learnability of solo-produced pantomimes versus signals that had been transmitted and used by interlocutors. In these studies, participants saw an artificial sign and judged whether an English translation matched or mismatched the meaning of the sign. In an event-related potential (ERP) study, we found that mismatches elicited larger negativities in the ERP than matches. However, those effects were most reminiscent of the classic N400 response in the evolved signs. This study provides a clearer view on how the mechanisms that drive language evolution change language to adapt to a learner's brain.

Keywords: artificial language learning; gesture comprehension; iterated learning

Introduction

All languages demonstrate systematic structure. From the smallest units of sound, to words and phrases, the elements of language are not independent. These elements are part of a structured system that allows infinite expressive power through the reuse and recombination of those elements. Systematicity is a property found across the world's languages, but how does this systematic structure of language emerge?

One answer to this question appeals to the forces of cultural evolution. Languages, like species, change over time and are subject to similar evolutionary processes found in biological evolution, such as variation, selection, and inheritance. In this view, language is under selectional pressures from human cognitive biases and adapted to suit the human brain (Christiansen & Chater, 2008). The nature of linguistic structure would then be a product of the learning and processing constraints that derive from underlying neural mechanisms.

One avenue for investigating the emergence of linguistic structure is to examine natural languages in the early stages of linguistic development. Although most communities have long-established languages, emerging sign languages such as Nicaraguan Sign Language (NSL) provide us with the opportunity to observe how linguistic features arise in a new human communication system. In the 1970s, the Nicaraguan government established a school for deaf children. These children, who communicated with their families via idiosyncratic systems of home sign, were brought together and organically created a novel sign language (Kegl, 1994).

In the case of NSL, each incoming cohort to the school has shaped the language and furthered its development (Goldin-Meadow et al., 2014). One example of the emergence and development of grammatical structure in NSL can be found in the use of spatial modulation to mark semantic roles in sentences expressing events with both an agent and a patient. Senghas (2003) found that signers from the earliest generation did not use the direction of spatial modulation in their interpretation of such sentences, whereas signers from the next generation made systematic use of spatial location to determine who the patient of the event was. The properties and structure of NSL thus changed as a function of transmission to learners of the next generation, as well as its use between signers who had already acquired the rules of the grammar.

Recent laboratory studies of artificial languages likewise suggest that the cultural evolutionary mechanisms of transmission and interaction play pivotal roles in the emergence of language and its change over time (Kirby, Cornish & Smith, 2008; Kirby, Griffiths, & Smith, 2014;

Tamariz, Cornish, Roberts, & Kirby, 2012). Motamedi and colleagues (2016) investigate the impact of interaction and transmission on the evolution of language by generating an artificial sign language in the lab.

In their study, participants in an initial "seed" generation were asked to innovate gestures for concepts that vary across six themes and four functional dimensions. These concepts were selected to share salient semantic features across a thematic category (Figure 1). Signs from the initial seed generation demonstrate high iconicity, use a lot of space, require a lot effort, are redundant, and use similar salient features of the theme (e.g. handshape that represents scissors cutting). Moreover, the seed generation signs do not contain features to distinguish across functional categories within a theme.

Functional Dimension

Thematic Dimension	Person	Location	Object	Action
Food	chef	restaurant	frying pan	to cook
Church	priest	church	bible	to preach
Photography	photographer	dark room	camera	to take a photo
Concert	singer	concert hall	microphone	to sing
Hair	hairdresser	hair salon	scissors	to give a haircut
Police	police officer	prison	handcuffs	to make an arrest

Figure 1: Chart of the 24 concepts from Motamedi et al. (2016)

In an iterated language learning paradigm, new sets of participants came into the lab and were trained on the gestures produced by the seed generation, and played a communication game using those gestures. The signs produced by one of these participants in a dyad was then passed on to two new participants as the training set. The process was repeated for five generations in a transmission chain. This design was intended to create pressure for participants to develop a way to communicate the different dimensions of category structure. Concepts from within a thematic category were similar such that a pantomime of each might be difficult to distinguish across the functional categories.

Motamedi and colleagues (2016) show that under the pressures of communication and transmission, highly iconic and lengthy manual signals signs change to become more efficient and less iconic. After several generations of interaction, the authors also found the recycling of gestures within a theme. Most impressively, Motamedi and colleagues (2016) found the emergence and retention of functional markers that make it possible to distinguish between concepts within a theme. For example, in one dyad, signers pointed at themselves to indicate that the subsequent gesture depicted a person.

Despite their iconic origins, many of the functional markers are not transparent to new learners, and must be learned as arbitrary constructs. In one artificial sign system, for example, the marker for action involved the raising of the right hand with the palm facing out. The emergence of functional markers after several generations of learners in this study is used as a proxy of the emergence of systematicity in linguistic structure.

The Present Study

Here, we examine whether the communicative advantages of the final generation signs outweigh the benefit of the iconicity in the signs from the seed generation. Accordingly, we present videos of gestures from Motamedi et al. (2016) in a word learning task in which we compare participants' ability to learn the meanings of the iconic seed generation signs versus those of the more language-like final generation signs. We are interested in the processing and learning of language-like artificial signs, thus we applied methods typically used to study processing of natural languages.

In our study, participants viewed signals from the artificial sign language followed by English words that either match or mismatch the signal's meaning. We focus on two different ways in which the word presented can mismatch the meaning of the sign. A Thematic Mismatch is a violation of the thematic category, (e.g. present the sign for hairdresser, then display the word "chef" on the screen), whereas a Functional Mismatch is a violation of the functional category (e.g. present the sign for hairdresser, followed by the word "scissors").

In manipulating the generation that the sign comes from, we are able to see if there are differences between learning improvised pantomimes versus the signs evolved in the lab. We expect that identifying a mismatch in the thematic violation cases would not be difficult for either seed signs and evolved signs, as all signs displayed some degree of iconicity, and were readily distinguishable between thematic categories, (e.g. food versus photography). However, we expect that identifying a functional violation would be more difficult because signs within a theme share many iconic features associated with their thematic category, and may not provide features that would allow a learner to distinguish between the four potential meanings.

In Experiment 1, we measured response times and accuracies in a behavioural artificial language learning task. In Experiment 2, participants complete the same task as in Experiment 1, while we measure event-related potentials (ERP) time-locked to the onset of the English translation of the sign. We are particularly interested in the N400, ERP component known to index difficulty associated with meaning processing or retrieval from semantic memory, and is produced reliably across a range of stimuli (Kutas & Federmeier, 2011). Even within 14 hours of instruction, second language learners show larger N400 responses to pseudowords compared to real words that were semantically related or unrelated to primes, indicating that limited exposure is sufficient for new language learners to gain sensitivity to lexical status and word meaning (Mclaughlin, Osterhout, & Kim, 2004). ERP studies allow for real-time

indexing of brain activity and provide multidimensional data about stages of processing. Thus, the N400 component is an appropriate dependent measure to more precisely examine the learning of an artificial language, in such a way that is comparable to studies investigating the processing of natural language.

Experiment 1

In Experiment 1, we taught participants signs from the Motamedi et al. (2016) in an explicit language learning paradigm. We used a within-subjects design in which each participant learned 12 signs from the seed generation and 12 signs from the final generation. In this behavioural experiment, we measured accuracy and reaction time in making judgements about whether the sign and word presented on the screen matched. We predict that accuracy will be greater for final generation signs after participants have learned the mappings, and reaction times will decrease as participants learn the system. We also expect lower accuracy rates and slower response times for Functional Mismatches.

Methods

Participants

We recruited 38 healthy undergraduates (15 M, 23 F). All gave informed consent and received course credit for participating. English was the primary language of all participants. One participant was excluded for not completing the experiment.

Materials and Procedure

Each trial began with a fixation cross for 500ms, followed by the video that varied from 2 - 7 seconds depending on signal length. A word then appeared until a key press was made, with feedback displayed on the screen for 500ms until the next fixation cross. We used two different stimulus lists varied across participants so that each concept was conveyed once with a seed gesture, and once with a final gesture. Participants watched videos of signs from either the seed generation or final generation. After each video was played, a word was displayed on the screen. The word either matched or did not match the meaning of the previously shown sign. When the word was displayed on the screen, participants pressed a key to indicate whether or not the word matched the sign. Participants received immediate feedback after every response they made. Feedback was given by the words "correct" or "incorrect" presented on the screen, and an accompanying tone. The experiment comprised 4 blocks of 48 trials each.

Figure 2: Example of a single trial.

Results and Discussion

Accuracy

A mixed effects logistic regression model was used to analyze the accuracy rate data. Models were constructed with the *lme4* package in R (R Core Team, 2013; Bates et al., 2015). Analysis involved construction of a generalized linear model to predict accuracy with experimental Block (First, Second, Third, Fourth), Generation (Seed, Final), and Condition (Match, Thematic Mismatch, Functional Mismatch) as categorical predictors, and all interactions. Models were fit with random intercepts for participants and for items (i.e. the videos that were played). Mean accuracy rates in each experimental category are shown in Figure 3. Model estimates are listed in Table 1. Analysis suggests accuracy rates improved as the blocks progressed. Experimental condition also impacted performance as accuracy rates were highest for Thematic Mismatches, lower for Functional Mismatches, with intermediate performance on the matches. The interactions between Condition and Block result because the learning curve was steeper for the more difficult Functional mismatches than the Thematic mismatches.

Participants' performance show that Functional Mismatches are more difficult to judge as being mismatches. The signs within a thematic category share many of the same features with respect to handshape and movement, such that differentiating between signs within a theme is ambiguous. Initially, participants perform worse in trials with final generation signs, which suggests that the markers contained in these signs are not transparent to new learners. There appears to be more arbitrariness to the form of a marker, i.e. an open hand facing palm forward denoting an action would not be considered an obvious association. However, after several trials participants quickly learn to map the marker to action verbs, as demonstrated by the increase in accuracy by the second block.

Table 1: Mixed effects logistic regression for accuracy rates.

	Estimates	t-value
Mismatch Type:		
Functional	-0.251	-8.94
Thematic	0.186	6.86
Generation	0.0134	0.600
Block	0.0520	9.10
Functional:Block	0.0554	5.51
Thematic:Block	0.0456	-4.51

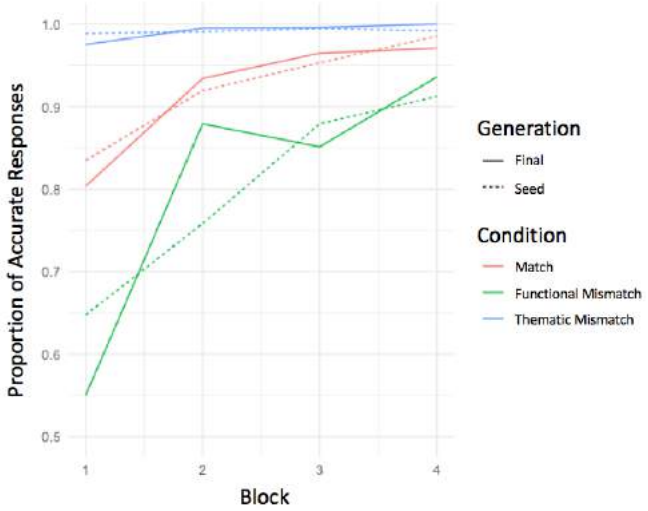


Figure 3: Accuracy rates between generation and condition across blocks.

Response Latency

To predict response latencies, we fit a linear mixed effects model in R (R Core Team, 2013) using the *lmer()* function of the *lme4* package (Bates et al., 2013). Predictor variables again included sign Generation (Seed, Final), Condition (Match, Thematic Mismatch, Functional Mismatch), and experimental Block (First, Second, Third, Fourth) and all interactions. Random intercepts were included for participants and item videos. Mean response latencies from each experimental category are shown in Figure 4 with model estimates listed in Table 2. Analysis revealed an interaction between Condition and Block, due to reaction times decreasing over the course of experimental blocks. Mismatch type also impacted performance as response latencies were consistently fastest for Thematic Mismatches, slower for Functional Mismatches, with intermediate performance for the Matches. Functional Mismatches also displayed the slowest average response latency across blocks, especially in the case of judging signals from the seed generation.

The results show that most of the learning of the mappings between sign and concept occurs during the first block of the experiment, as demonstrated by the slope of the response latencies from Block 1 to Block 2. As expected, participants respond faster to Thematic Mismatches since mismatches are easier to detect when the gestures produced clearly relate to different themes. Responding to seed signals is slower overall, which suggests that more processing occurs in deciding whether the signal matches the word presented. Seed signs are characterized as being longer in length, repetitive, pantomime-like, and lacking in defining features that would differentiate them from similar concepts. Between Blocks 3 and 4, there is a decrease in reaction time for decisions about final generation signals, suggesting that participants have mastered the meaning of the functional markers.

Table 2: Linear mixed effects model for response latency.

	Estimates	t-value
Condition	-0.573	-7.57
Generation	-0.573	0.726
Block	-0.208	-10.4
Condition:Block	0.132	4.91

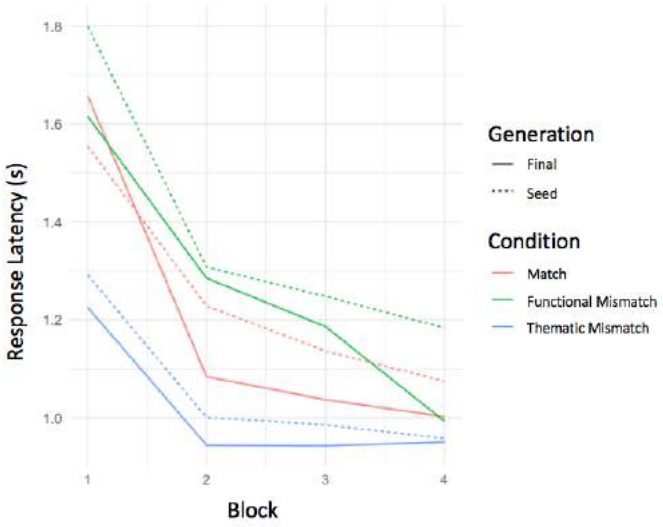


Figure 4: Mean response latencies between generation and mismatch type across experimental blocks.

Experiment 2

In Experiment 2, we measured neural responses in an artificial language learning paradigm. If the participant has learned the sign, we would expect mismatches to elicit a larger N400 response than matches. If final generation signs are indeed more learnable than those from the seed generation, we might expect to see larger amplitude N400 effects on words following the final generation signs than those following words from the seed generation.

Methods

Participants

We recruited 34 healthy undergraduates at UCSD (12 M, 22 F). All gave informed consent and received course credit for participating. English was the primary language of all participants. Two participants were excluded, one for excessive sneezing and sniffling, and one who was unable to complete the experiment within the allotted two hours. Participants completed surveys on handedness, neurological damage, and medication.

Materials and Procedure

Materials and procedure were adapted from the behavioural study outlined in Experiment 1.

EEG Data Collection

EEG was collected from 29 scalp sites using an ElectroCap mounted with electrodes. Scalp electrodes were referenced to the left mastoid. Blinks were monitored from an electrode below the right eye and referenced to the left mastoid. Horizontal eye movements were monitored via two electrodes placed beside each eye. Electrical impedance was reduced to less than 5 kohms. EEG was recorded and amplified using SA instrument bioelectric amplifier. The EEG was digitized at a sampling rate of 512 Hz. Recording took place in a dimly lit, sound-attenuated, electrically-shielded chamber. Participants were seated in front of a CRT monitor for stimulus presentation.

Results and Discussion

ERPs were time locked to the onset of potential meanings (viz. English words) presented after each signal. Mean amplitude was measured relative to a 100ms pre-stimulus baseline in two time windows: 300-500ms post-onset, intended to capture the N400 component, and 500-700ms post-onset, intended to capture the P600. In each interval, analysis involved repeated measures ANOVA with factors Condition (Match, Thematic Mismatch, Functional Mismatch), Generation (Seed, Final), Block (First, Second, Third, and Fourth), and two factors intended to capture the location of electrodes across the scalp, Hemisphere (Left, Right), and Region (Frontal, Frontocentral, Central, Centroparietal, Parietal, Occipital). Where relevant, the Greenhouse Geisser correction has been applied to p-values; however, for clarity, we report the original degrees of freedom.

Omnibus analyses revealed (among other effects) the presence of significant complex interactions with Block in both intervals (*N400:* Condition x Generation x Block x Hemisphere $F(6, 186) = 3.36$, $p < 0.05$; *P600:* Condition x Generation x Block $F(6, 186) = 2.76$, $p < 0.05$, Condition x Generation x Block x Hemisphere $F(6, 186) = 2.5$, $p < 0.05$), motivating separate follow-on analyses within each block.

N400 Analysis of ERPs in the first block revealed a main effect of Condition ($F(2, 62) = 8.2$, $p<0.05$), but no interaction with Generation ($F(2,62) = 1.03$, n.s.). By contrast, analysis of the second block suggested condition effects differed for signs from the seed versus the final generation (Condition, $F(2,62) = 18.4$ $p<0.001$; Generation, $F(1,31) = 4.2$, $p<0.01$; Condition x Generation, $F(2,62) = 3.26$, $p<0.05$; Condition x Generation x Hemisphere, $F(2, 62) = 7.8$, $p<0.01$). In the third block, Condition effects were present ($F(2,62) = 14.3$, $p<0.001$, but were similar for seed and final generation signals (Condition x Generation, $F(2,62) = 1.2$, n.s.). In the final block, Condition effects ($F(2,62) = 7.3$, $p<0.01$) displayed a different topographic profile following seed versus final generation signs (Condition x Generation x Region, $F(10,310) = 3.48$, $p<0.01$).

Figure 5 shows the topography of ERPs in the N400 interval for each type of mismatch following seed (upper panel) and final (lower panel) generation signs. Whereas the seed generation mismatches display a right frontal maximum reminiscent of ERPs to imageable words (see, e.g., Swaab, Baynes, & Knight, 2002), the topography of the final generation mismatches resembles the classic N400 that results from associative priming (e.g., Steinhauer, et al., 2017).

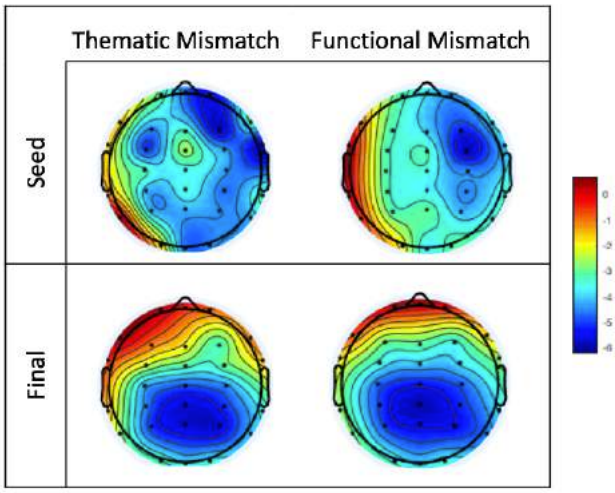


Figure 5: Difference in amplitude for latency between 300-500ms.

P600 Among other effects, follow up analyses revealed the presence of complex interactions between Condition, Generation, and topographic factors in blocks 1, 2, and 4 (*Block 1:* Condition x Generation, $F(2 ,62) = 4.55$, $p<0.01$; *Block 2:* Condition x Generation, $F(2,62) = 4.27$, $p<0.05$, Condition x Generation x Hemisphere, $F(2,62) = 7.23$, $p<0.001$; *Block 3:* Condition x Generation x Region $F(10,310) = 1.99$, n.s.; *Block 4:* Condition x Generation x Region, $F(10,310) = 6.96$, $p<0.001$). Figure 6 shows the topography of mismatch effects (match – mismatch) 500-700ms following seed and final generation gestures.

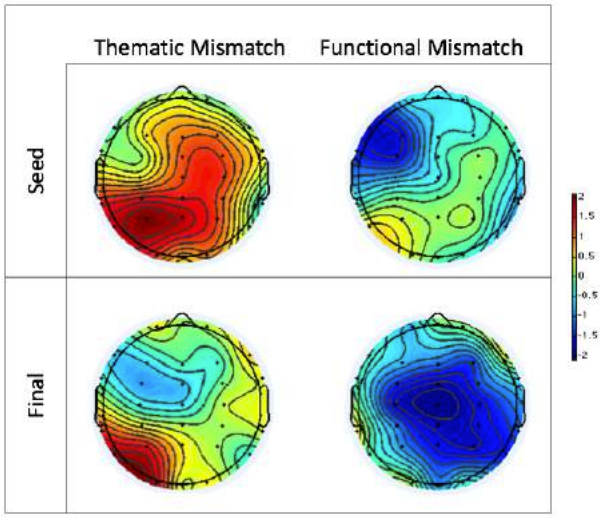


Figure 6: Difference in amplitude for each condition compared to the matches in the 500-700ms window

Figure 7 shows ERPs recorded at Pz, a parietal site where N400 and P600 are typically prominent. In the first half of the study, the N400 dominates the ERP response to these words, with more clear differentiation between the three conditions being evident in the seed generation signs. In the latter half of the study, N400 effects are overlapped by late positivities related to the task of classifying the word as a match or a mismatch. Following both seed and final generation signs, matches elicit a positivity that peaks earlier than the thematic mismatches (viz., seeing "chef" after the sign for hairdresser). For functional mismatches (viz., seeing "hair salon" after the sign for hairdresser), however, ERPs in the seed generation are more similar to the matches, whereas functional mismatches in the final generation are more similar to the thematic mismatches.

General Discussion

Here, we examined the ways in which a culturally-evolved artificial language may be advantageous to learn, in comparison to a system of individually iconic communication signals that lack internal systematicity. We found that signs from the more evolved system included both a consistent and concise iconic gesture to indicate thematic category, and a gesture that indicates whether the concept is a person, place, object, or action. Although the behavioral study suggested participants readily learned both the evolved final generation signs and the less systematic seed ones, the real-time brain response revealed processing differences for the two kinds of signs.

Our ERP study revealed a classic N400 response to signs from the final generation, indicative of semantic processing. By contrast, the iconic seed generation signs elicited concreteness effects that suggested participants exploited a learning strategy that involved mental imagery. Moreover, the brain response to final generation signs suggested participants could distinguish between closely related concepts such as hairdresser and scissors, whereas such concepts were treated identically in the seed generation.

Previous studies have also found that when used in a referential or communicative game, signs representing concepts from a set of shared semantic relations become more arbitrary, schematized, and systematized across dimensions (Theisen, Oberlander, & Kirby, 2010). We see that the introduction of a system of schematized signs influences how the meaning signs are retrieved from memory via ERPs to violations in signal-meaning pairings.

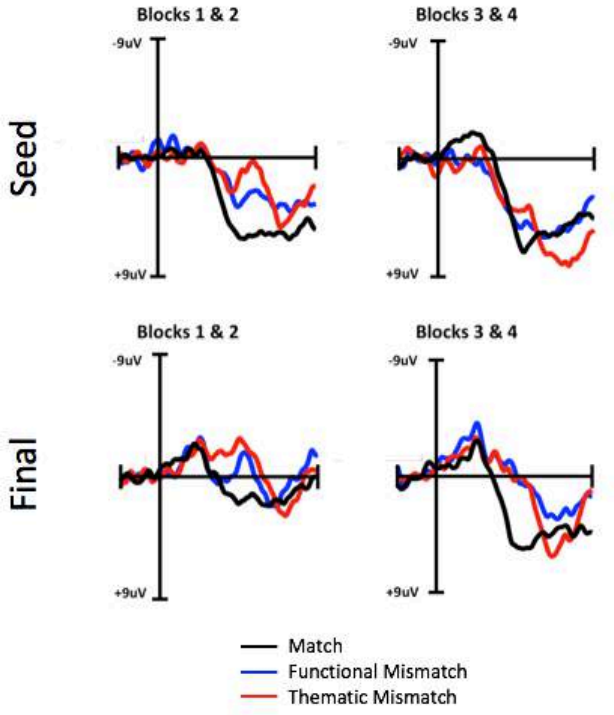


Figure 7: ERP waveforms recorded at electrode site Pz.

A recent study by Nölle and colleagues (2018) demonstrated how individuals use context and the environment to shape the signals they use together. The authors found that interlocutors were more likely to produce systematically-related signals rather than signals that refer to some idiosyncratic feature of the referent, even when both strategies were afforded by the environment. In the present study, we found that the brain's real time response displayed a greater sensitivity to subtle distinctions within a thematic domain for meanings conveyed by final generation signs that contained the functional markers. Systematic signs are easier to remember and rely more on abstraction to identify like features that can be referred to similarly.

Our current design adapts videos generated in a previous study as stimuli. This choice may introduce confounds relating to processing and working memory, as all seed generation signs were longer than the final generation signs. The seed signs are highly iconic pantomimes of actions associated with the theme, thereby resulting in longer signals that lack specificity. Consequently, the differences we found between learning seed and final signs might reflect differences in length of seed versus final generation signs, differences in the degree of structure, or some combination. Future research should seek to unconfound these factors.

Results of the present study support that artificial language shaped by interaction and transmission is more learnable. As such, it is in keeping with research that reports differences in the brain response in learners of another culturally-evolved artificial language (Verhoef, Walker, Marghetis, & Coulson, 2018). Signs evolved through interaction and transmission

display systematic structure, and this systematic structure better suits the learner's brain.

References

Christiansen, M. H. & Chater, N. (2008). Language as shaped by the brain. *Behavioral and Brain Sciences*, 31(5):489-509.

Goldin-Meadow, S., Brentari, D., Coppola, M., Horton, L. & Senghas, A. (2014). Watching language grow in the manual modality: Nominals, predicates, and handshapes. *Cognition*, 136, 381–395.

Kegl, J. (1994). The Nicaraguan sign language project: An overview. *Signpost*, *7*, 24-31.

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, *105*(31), 10681-10686.

Kirby, S., Griffiths, T., & Smith, K. (2014). Iterated learning and the evolution of language. *Current Opinion in Neurobiology*, *28C*, 108–114.

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). *Annual review of psychology*, *62*, 621-647.

McLaughlin, J., Osterhout, L., & Kim, A. (2004). Neural correlates of second-language word learning: Minimal instruction produces rapid change. *Nature neuroscience*, *7*(7), 703.

Motamedi, Y., Schouwstra, M., Smith, K., & Kirby, S. (2016). Linguistic structure emerges in the cultural evolution of artificial sign languages. In *The Evolution of Language: Proceedings of the 11th International Conference. New Orleans: EVOLANG11* (pp. 21-24).

Nölle, J., Staib, M., Fusaroli, R., & Tylén, K. (2018). The emergence of systematicity: how environmental and communicative factors shape a novel communication system. *Cognition*, *181*, 93-104.

Senghas, A. (2003). Intergenerational influence and ontogenetic development in the emergence of spatial grammar in Nicaraguan Sign Language. *Cognitive Development*, *18*(4), 511-531.

Steinhauer, K., Royle, P., Drury, J. E., & Fromont, L. A. (2017). The priming of priming: Evidence that the N400 reflects context-dependent post-retrieval word integration in working memory. *Neuroscience letters*, *651*, 192-197.

Swaab, T. Y., Baynes, K., & Knight, R. T. (2002). Separable effects of priming and imageability on word processing: an ERP study. *Cognitive Brain Research*, *15*(1), 99-103.

Tamariz, M., Cornish, H., Roberts, S., & Kirby, S. (2012). How generation turnover and intelocutor negotiation affect language evolution. In *The Evolution of Language: Proceedings of the 9th International Conference on the Evolution of Language: World Scientific* (pp. 555-556).

Theisen, C. A., Oberlander, J., & Kirby, S. (2010). Systematicity and arbitrariness in novel communication systems. *Interaction Studies*, *11*(1), 14-32.

Verhoef, T., Walker, E., Marghetis, T., & Coulson, S. (2018). Neural measures of sensitivity to a culturally evolved space-time language: shared biases and conventionalization. In *Proceedings of the 40th Annual Meeting of the Cognitive Science Society*. Madison, WI: Cognitive Science Society (pp. 1151-1156).

Contextualizing Conversational Strategies: Backchannel, Repair and Linguistic Alignment in Spontaneous and Task-Oriented Conversations

Christina Dideriksen (christina@cc.au.dk)
School of Communication and Culture
Aarhus University, Jens Chr. Skous vej 2, 8000 Aarhus C, Denmark

Riccardo Fusaroli (fusaroli@cc.au.dk)
Kristian Tylén (kristian@dac.au.dk)
School of Communication and Culture & The Interacting Minds Center
Aarhus University, Jens Chr. Skous Vej 2, 8000 Aarhus C, Denmark

Mark Dingemanse (m.dingemanse@let.ru.nl)
Centre for Language Studies, Radboud University & Max Planck Institute for Psycholinguistics
PO Box 310 6500 AH Nijmegen, The Netherlands

Morten H. Christiansen (morten@cas.au.dk)
Department of Psychology, Cornell University, Ithaca, NY 14853
School of Communication and Culture & The Interacting Minds Center
Aarhus University, Jens Chr. Skous vej 2, 8000 Aarhus C, Denmark

Abstract

Do people adjust their conversational strategies to the specific contextual demands of a given situation? Prior studies have yielded conflicting results, making it unclear how strategies vary with demands. We combine insights from qualitative and quantitative approaches in a within-participant experimental design involving two different contexts: spontaneously occurring conversations (SOC) and task-oriented conversations (TOC). We systematically assess backchanneling, other-initiated repair and linguistic alignment. We find that SOC exhibit a higher number of backchannels, a reduced and more generic repair format and higher rates of lexical and syntactic alignment. TOC are characterized by a high number of specific repairs and a lower rate of lexical and syntactic alignment. However, when alignment occurs, more linguistic forms are aligned. The findings show that conversational strategies adapt to contextual demands.

Keywords: conversational dynamics; common ground; interactive alignment; backchannels; repair

Introduction

How do we continuously update mutual knowledge and coordinate behaviors in conversations? The issue has been defined as grounding: a constant evaluation of whether we share mutual beliefs, knowledge, and understanding sufficient for the purpose of the situation. (Clark & Brennan, 1991). A diverse set of disciplines have approached this issue, from psycholinguistics to conversation analysis, highlighting several conversational strategies for coordinating interactions. Interlocutors might ensure common ground by subtly confirming their understanding (*backchanneling*), more explicitly signaling misunderstanding and correcting each other (*conversational other-repair*), or by re-using each other's linguistic forms (*linguistic alignment*). Even if such conversational strategies are often viewed as ubiquitous in interaction, one might expect them to vary considerably across individuals, and more importantly across different types of contexts. Conversations involve a plurality of linguistic and social games, from exchanging specific information to maintaining a social reputation; from coordinating decisions to ensuring a comfortable emotional environment (Fay et al., 2018; Fusaroli, Rączaszek-Leonardi, & Tylén, 2014; Fusaroli & Tylén, 2016). Thus, different contexts of conversation might afford different grounding strategies. For instance, conversations between pilots and airport control towers require a higher need of referential precision, and therefore more strict monitoring of the common ground (Prinzo & Britton, 1993). On the contrary, small talk and dinner conversations are arguably more focused on building and maintaining social relations, possibly not necessitating the same need for detailed and continuously monitored referential precision. Indeed in more casual chats people have been observed to not always realize that they are talking about different things, or even that they shift partners mid-conversation in an instant messaging system (Dunbar, Marriott, & Duncan, 1997; Galantucci & Roberts, 2014). Thus, different conversational contexts may require different dimensions and degrees of coordination.

Understanding how the joint use of multiple conversational strategies adapts to the activity at hand is crucial. Cognitive science, management, and other disciplines are increasingly focusing on how to promote effective team coordination and the conversational patterns underlying it (Fusaroli & Tylén, 2016; Pentland, 2012; Wiltshire, Butner, & Fiore, 2018). Clinical research is investigating how social impairment develops and unfolds across a wide spectrum of neuropsychiatric conditions, and atypical conversational strategies are likely to play a role in impaired social

functioning (Bolis, Balsters, Wenderoth, Becchio, & Schilbach, 2017; Lavelle, Healey, & McCabe, 2014; McCabe & Healey, 2018; Wadge, Brewer, Bird, Toni, & Stolk, 2018). Further, effective human-computer and human-robot interactions also require a detailed understanding of when and how grounding happens (Loth, Jettka, Giuliani, & De Ruiter, 2015). However, conversational strategies have traditionally been investigated across different disciplines, with varied methodological approaches and foci. Only recently the field has started combining qualitative insights with quantitative methodologies (De Ruiter & Albert, 2017; Dingemanse & Enfield, 2015) and systematically assessing the role of contexts and activities (Fusaroli et al., 2017; Healey, Purver, & Howes, 2014; Reitter & Moore, 2006). In the following, we define the relevant conversational strategies and how they might be adjusted to different contextual demands, and then present a study systematically assessing them across two different contexts: Spontaneously Occurring (SOC) and Task Oriented Conversations (TOC).

Conversational strategies and context

By conversational strategies, we here refer to a heterogeneous set of linguistic behaviors often investigated separately under the headline of *backchannels*, *repair* and *alignment*. Backchannels are defined as head nods or short utterances consisting of a word (e.g. *'uh-huh, 'yes', 'okay'*), or short sentences, often repeating the previous turn (e.g. A: *'let's meet Monday at 10', B: 'Monday at 10'*). They do not take the floor in a conversation, but are used to exhibit interest, understanding, and perhaps even agreement with the speaker's utterance. Thus, backchannels signal a shared common ground and that the speaker can continue with their speech turn (Bangerter & Clark, 2003; Jurafsky, Shriberg, Fox, & Curl, 1998; Schegloff, 1982; Yngve, 1970). Therefore, (Hypothesis 1, H1) we would expect that TOC involve a higher occurrence of backchannels than SOC. This hypothesis is supported by previous observational work (Fusaroli et al., 2017). Note that here we focus on vocal backchannels only.

The second strategy that we examine, conversational repair, also creates feedback on the level of mutual understanding between interlocutors. However, while backchannels mainly provide positive feedback, conversational repair works by providing negative feedback, signaling impending communicative trouble and a need to re-establish common ground. Repair can take different linguistic forms and levels of specificity in the feedback to the interlocutor. Here we focus on other-initiated repair, where a listener indicates trouble in hearing or understanding. The listener can use a repair request to signal that there is a problem with understanding, and thereby invite the speaker to clarify what was said and "repair" mutual understanding. Previous studies have defined three categories of repair (Dingemanse et al., 2015; Fusaroli et al., 2017). Open repair refers to problems on a general level of understanding, where the repair initiation does not specify what or where the problem is (e.g., *huh?*, *what?*, *what did you say?*). In contrast, restricted repair points to specific parts of the previous sentence that need clarification (e.g., *'who?'* or *'where*?'). Restricted suggestions are even more specific, pointing to the specific source of uncertainty and offering a suggestion as to how to repair it (e.g., '*did you say Monday?'* or '*X or what?*'). A previous study has found that repair is more frequent in TOC, due to the higher demand for precision in mutual understanding (Colman & Healey, 2011). Therefore, we also expect (H2) repair to occur more often in TOC than in SOC. Further, we expect (H3) the more specific forms of repair to be driven by the need for accuracy: restricted suggestion repairs should be higher in TOC, with restricted request and open repairs being frequent in SOC. Both hypotheses have preliminary support in a previous study (Fusaroli et al., 2017).

A third strategy is the reciprocal alignment of linguistic forms. As interlocutors hear each other using, for instance, specific words, they prime each other to re-use them. Linguistic alignment is argued to implicitly increase similarity of interlocutors' mental situation models, and thereby catalyze increased rapport and interpersonal coordination (Dale, Fusaroli, Duran, & Richardson, 2013; Pickering & Garrod, 2004). In particular, here we focus on lexical, syntactic and semantic alignment. Lexical alignment indicates the tendency to re-use an interlocutor's lexical choices ("do we *ignore it*?", "yes, let's *ignore it*"). Syntactic alignment indicates the tendency to reuse the interlocutor's syntactic constructions beyond lexical choices ("*you have passed* the youth hostel?", "yes and *I have reached* the top of the map", where the sequence of subject and verb forms are repeated). Semantic alignment indicates the tendency to keep talking about the same topics, beyond lexical and syntactic alignment. Since linguistic alignment is argued to facilitate joint task performance, we expect (H4a) it to be higher in TOC than SOC. Two studies support this hypothesis: syntactic alignment is observed to be higher in task oriented conversational corpora than in more spontaneous conversational corpora (Healey et al., 2014; Reitter & Moore, 2006). However, given that alignment is often associated with building and maintaining rapport (Ireland et al., 2011), one could also expect (H4b) it to be higher in SOC than TOC. Indeed, one previous study supports this second hypothesis (Fusaroli et al., 2017).

In summary, SOC and TOC involve different contextual demands and therefore may emphasize different aspects of the common ground. SOC have a more marked social function in the maintenance of relations, and a lower need for detailed and accurate referential understanding than TOC. We therefore expect TOC to display more precise building and assessment of shared situation models through backchannels, repair, and perhaps alignment.

The current study

While previous studies support at least some of our hypotheses, they mostly investigate one conversational strategy at a time, often with widely different methods or data. Further, all previous studies rely on cross-sectional

conversational corpora, that is, the individuals included do not overlap across corpora, making it harder to assess whether any observed differences are indeed a function of different contexts. Additionally, the corpora were often collected in a heterogeneous fashion, with varying numbers of interlocutors and different ongoing activities. Here we aim to more rigorously investigate the role of conversational strategies by controlling these contextual variables. We experimentally elicited multiple SOCs and TOCs in a within-subject design. This allowed us to get a more detailed picture of how participants use conversational strategies in different interactional settings, assessing both contextual, pair and individual variability, as well as their reciprocal relations.

Besides the hypotheses sketched above, we also explore how the three conversational strategies relate to each other, and how they are affected by the contrast between conditions. It has been argued that social interactions display signatures of self-organization dynamics, that is, conversational behaviors become interdependent within and between interlocutors, so that the increase in one strategy might result in decrease in another. For instance, a high rate of lexical alignment provides high informational redundancy and might make repairs less necessary. Further, these relations are likely to be modified by contextual demands, such that different types of conversations might result in very different interrelations. (Dale et al., 2013; Fusaroli et al., 2014).

Methods

We elicited conversations from 39 dyads (78 Danish individuals). All participants were native speakers of Danish (M age = 23.19, SD = 3.58, males = 33, females = 45). The dyads were composed of 15 female dyads, 9 male dyads, and 15 mixed-gender dyads. The participants in 8 of the 39 dyads knew each other prior to the experiment. Each dyad produced two SOCs and two TOCs. The members of each dyad were first offered a sheet of open-ended conversation prompts (e.g. "find two tv-series that your interlocutor would like to watch") and asked to freely chat while the experimenter was busy elsewhere (first SOC). Participants were then asked to engage in two joint problem-solving tasks: the map task (Anderson et al., 1991) and a categorization task (the alien game; Tylén, Fusaroli, Smith, & Arnoldi, 2016). Both tasks require participants to collaborate to solve the tasks effectively (2 TOCs). Finally, the participants were asked to freely chat again, or to use the conversation prompts (second SOC). Note that SOCs were elicited in an experimental context, which limits the degree of spontaneity of the conversations. They are nevertheless more spontaneous than TOCs. Each conversation lasted approximately 10 minutes, for a total of 40 minutes per dyad (2 SOCs and 2 TOCs). In the SOC condition, we had a total of 34,544 speech turns and an average of 443 speech turns per conversation. In the TOC condition, we had a total of 45,607 speech turns, with an average of 585 turns per conversation. All conversations were transcribed orthographically using ELAN (Brugman & Russel, 2004) and manually coded for backchannels and the three different types of repair by independent coders naïve to the purpose of the study (intercoder reliability: kappas > 0.6). Coding schemes were developed based on prior work (Dingemanse, Kendrick, & Enfield, 2016; Yngve, 1970) and are available at https://bit.ly/2LtUmax. Alignment was calculated as cosine similarity between successive conversational turns. Lexical alignment was based on lemmatized words, syntactic alignment on 2-grams of part-of-speech tags, and semantic alignment on FastText word2vec representations of Danish (Bojanowski, Grave, Joulin, & Mikolov, 2017; Duran, Paxton, & Fusaroli, accepted). While 2-grams are only a rough proxy for syntax, exploratory analyses of 3- and 4-grams yielded similar results. Previous work has also used comparisons to surrogate pair baselines, created by artificially interleaving the utterances of interlocutors from different pairs (Healey el al. 2014). Since the current work is focusing on a within pair manipulation, we leave such baselines for future work.

Bayesian multilevel models with weakly informative priors were used. Backchannel and repairs were modelled according to a Bernoulli likelihood function. Specific types of repairs were analyzed within the subset of repair utterances only. Alignments were fit to a Zero Inflated Beta likelihood function to account for the high amount of turns with no alignment (zero-inflation, see distribution plots here: https://bit.ly/2LtUmax). Note that we report alignment rate as the negative of the inflation term, thus indicating the log odds rate of any alignment instead of the rate of no alignment, and the alignment level as the log odds of the cosine similarity when there are occurrences of alignment.

All parameters were modelled as correlated and predicted by task, including varying effects by interlocutor nested within pair. LOOIC-based stacking weights assessed the relevance of the predictors (Vehtari, Gelman, & Gabry, 2017). Evidence ratio (ER) was used to test evidence in favor of our hypotheses. While ER is a continuous scale, indicative thresholds have been proposed with values of 3 corresponding to moderate evidence for the hypothesis and values of 0.3 to moderate evidence against the hypothesis. Full posterior distributions of varying effects were used to exploratorily estimate and visualize correlations between backchannels, repair and alignment as *r* Pearson coefficients[1]. Only correlations with an absolute coefficient above 0.2 and credibility intervals not overlapping with 0 were included. Global strength was calculated as the sum of the absolute value of all edges in the network. The implementation relied on brms, ggplot, igraph, Stan and R (Bürkner, 2017; Carpenter et al., 2017; Csardi & Nepusz, 2006).

[1] We also built a multivariate outcome model including all previous models and the correlation between outcomes in the same model. However, the model could not be fit due to its complexity.

Results

Estimates of the occurrence rate of the conversational strategies by condition are reported in Figure 1. Detailed results of the effects of task are reported in Table 1 and Figure 1.

Table 1: Conditional effects on conversational strategies. *Estimates represent the log-odds probability of the behavior in question, separately for each condition. Evidence ratio (ER) indicates the relative evidence in favor of our hypotheses as specified in "The current study" (against the alternative hypotheses). "Lex", "Syn" and "Sem" stand respectively for lexical, syntactic and semantic alignment.*

Outcome	Spontaneous	Task-Oriented	ER
Backchannel	-0.69 (-0.8, -0.58)	-1.45 (-1.57, -1.33)	*< 0.001*
Repair	-3.55 (-3.91, -3.17)	-2.86 (-3.11, -2.51)	*> 1000*
- Open	-0.79 (-1.08, -0.48)	-2.20 (-2.45, -1.97)	*> 1000*
- Restricted	-1.32 (-1.61 -1.02)	-1.80 (-2.05 -1.54)	= 132
- Suggestion	0.01 (-0.25 0.27)	1.02 (0.78 1.26)	> 1000
Lex Rate	0.22 (0.11, 0.33)	0.06 (-0.04, 0.16)	*= 332.3*
Lex Level	-0.79 (-0.83, -0.75)	-0.68 (-0.72, -0.64)	*> 1000*
Syn Rate	0.65 (0.52 0.77)	0.50 (0.40 0.59)	= 73.07
Syn Level	-0.78 (-0.83 -0.74)	-0.75 (-0.78 -0.73)	= 6.49
Sem Rate	1.47 (1.13 1.79)	1.61 (1.25 1.96)	= 2.93
Sem Level	0.36 (0.15 0.54)	0.34 (0.16 0.51)	= 1.3

Conversational strategies vary considerably across individuals and pairs. For instance, some pairs consistently show use of backchannels above average, while others consistently below. Analogously, some individuals consistently show higher use of, for instance, backchannels than their interlocutor. Further, the use of conversational strategies is interrelated, as shown in Figure 2. Alignment strategies are strongly and positively related (except for levels of lexical alignment), while repair and backchannel seem less related. Interestingly, the type of conversation seems to affect the relations between strategies, with task-oriented conversations displaying weaker relations (global strength of SOC = 2.91, TOC = 1.67).

Discussion

This study aimed to assess the impact of different contextual demands (SOC vs. TOC) on three conversational strategies and their interrelations. It used a more rigorous within-subject design than previous studies. We hypothesized that backchannel (H1) and repair (H2) would be used more in TOC than SOC, and that the specificity of the repair would also be higher in TOC than SOC (H3). We contrasted two hypotheses for linguistic alignment (H4a-b), as the previous evidence was contradictory, indicating sometimes higher alignment in SOC, sometimes in TOC.

We found high occurrence of backchannels, however, contrary to H1 this was higher in SOC (33.4% of utterances) than TOC (19%). While backchannels certainly play a role in grounding given their high occurrence in TOC, the findings suggest that they might be even more important in free conversation and thus related to, for instance, the maintaining

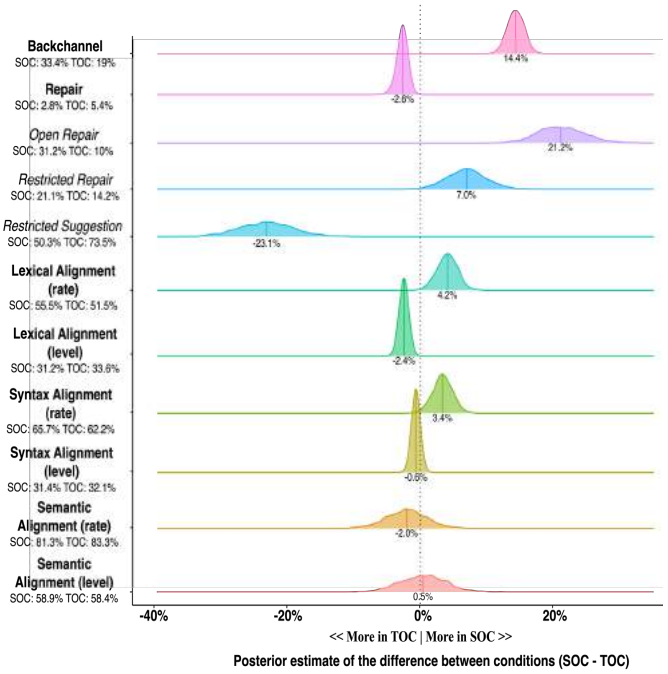


Figure 1 - Effects of task on the conversational strategies of interest. *The column on the left includes posterior estimates of the proportional use of the strategies across the two conditions, the ridge plots indicate the posterior estimates of the differences between conditions. Note that backchannels and repairs are calculated on the total number of speech turns, while repair types are calculated on repair turns.*

of social relations. Indeed, backchannels have been argued to strengthen the social relationship by making both interlocutors part of the conversation even though one speaker holds the floor (Duncan & Fiske, 1977), as well as by consistently displaying interest and attention to the speaker (Levinson, Brown, & Levinson, 1987). The large variation in the use of backchannels is in line with a previous study (Heldner, Hjalmarsson, & Edlund, 2013).

We found that conversational repair is less frequent than backchannelling. In line with H2 and H3, we found that repair was more frequent in TOC (5.4% of utterances) than in SOC (2.8% of utterances), and that the relative frequency of types of repair was affected by contextual demands. More specific repair was more frequent in TOC than SOC, while less specific repair was more common in SOC than TOC. This supports our prediction that in SOC, referential precision is less important than in TOC, where attention to details and accuracy of the information is crucial to solve the collective task at hand (Dingemanse et al., 2015; Fusaroli et al., 2017). Thus, in TOC interlocutors more carefully monitor potential misunderstandings and, given the higher attention to details, tend to prefer the strongest (in this case, most specific) repair form they are able to use in that situation (Clark & Schaefer,

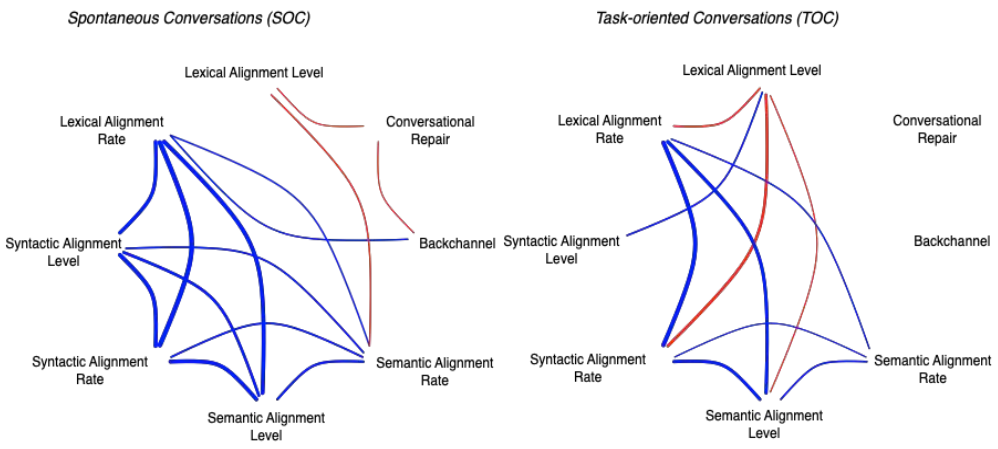


Figure 2 - Exploratory correlation network between individual usage of conversational strategies. Blue lines indicate positive correlations, red lines negative ones. Thickness of the line is proportional to the strength of the correlation, from 0.2 (backchannel and repair in SOC) to 0.9 (rates of syntactic and lexical alignment in SOC).

1987). Our findings suggest that interlocutors use repair as an adaptive strategy, adjusted to target the optimal way of grounding in order to meet specific contextual demands. We found partial support for both H4a and H4b. More than 50% of utterances displayed some lexical alignment, more than 60% syntactic alignment, and more than 80% semantic alignment. However, pervasive, the rate of lexical and syntactic alignment was affected by the type of conversation: higher in SOC than TOC (H4b). While the small effect size should warrant caution, alignment rate might be more related to building and maintaining social relationships. Previous studies have indeed argued that high rates of alignment can lead to informational redundancy, which might hamper task-oriented coordination, in situations that afford interlocutors to provide complementary information in a more synergistic fashion (Fusaroli et al., 2012; Fusaroli et al., 2014; Fusaroli & Tylén, 2016). When interlocutors do align, we find that a relatively high number of linguistic forms is aligned (alignment level): more than 30% of lexical and syntactic choices, and more than 50% of semantic variance. However, lexical and syntactic alignment levels are higher in TOC than in SOC (H4a), indicating that TOC presents fewer instances of alignment, but once alignment is there, more forms are aligned. This might again be due to the demand for precision affording more specific alignment strategies (as in the case of repair). Together with the zero-inflated distribution of alignment, this seems to suggest that we are confronted by two different alignment phenomena that have so far been conflated. The lack of clear differences in semantic alignment between SOC and TOC also seems to indicate that in both types of conversations interlocutors do speak about similar topics, although the specific strategies (lexical and syntactic choices) differ. Future work will include exploration of the effect of 3- and 4-grams on parts-of-speech tags and whether stratifying syntactic by lexical alignment preserves the same pattern of results.

The different conversational strategies, backchannels, repair and alignment, form a dense network of interdependencies (Figure 2). We show that the different forms of alignment are strongly related: interlocutors using alignment do so consistently for all forms of alignment, with the interesting exception of levels of lexical alignment (suggesting that verbatim lexical repetitions might play a different role; e.g., Fusaroli et al., 2014). However, contextual demands affect this network: while backchannels and repairs are negatively related to each other and connected to the rest of the network in SOC, in TOC they are isolated and generally the network displays sparser and weaker connections. This was an unexpected result that requires follow-up work.

In this study, we attempt to develop a more theory-driven and cumulative approach to the study of conversational strategies. Increased control enabled us to more robustly infer how contextual demands change the use of conversational strategies and affect their relations, with results at least partially different from previous less-controlled studies. This provides new insights into how we build effective coordination in conversations, and may illuminate some of the mechanisms underlying social impairment. We are aware of the limitations of this initial work. We have contrasted two macro-categories of conversation: task-oriented and spontaneously occurring. These categories are heterogeneous. The two tasks employed in TOC have somewhat different contextual demands. The two spontaneous conversations in SOC happen at the beginning and end of the experimental setup, influencing interlocutors' feeling of familiarity. Pairs display a high variability in joint performance in the tasks and rapport. This highlights the importance of considering differences across individuals and pairs in studies of conversations (which is also receiving increased attention in other areas of psycholinguistics; Kidd, Donnelly, & Christiansen, 2018). Future work will further explore these dimensions: analyzing differences between interlocutors who know or do not know each other in advance, introducing measures of performance and rapport, as well as accounting for progress familiarity and assessing how conversational strategies evolve over time as interlocutors become familiar with each other and with the tasks. A more direct manipulation of contextual demands, for instance, more continuously varying the need for accurate mutual understanding, is a necessary next step. Future work

will also further articulate the interrelations between strategies and contextual demands.

Conclusion

In a controlled within-participant design, we show that contextual demands afford diverse uses of conversational strategies. Spontaneously occurring conversations display higher frequencies of backchannels and a higher rate of lexical and syntactic alignment, possibly related to higher needs for relation building and maintenance. Task-oriented conversations display higher occurrence of repair (in particular of specific repairs) and levels of lexical and syntactic alignment, possibly related to needs for high precision. Our results suggest that backchannels, repair and alignment serve complementary functions, and that interlocutors flexibly adapt these grounding strategies contingent on current contextual demands. By focusing on how pairs and individuals adjust their strategies to contextual demands, we can better understand the patterns of effective communication, and how communication might fail. Future work might extend this approach to include measures of coordination success as well as investigation of these phenomena in contexts of social impairment.

Acknowledgments

We thank Katrine Garly and Jakob Steensig for their help in developing the coding schemes for repair and backchannel. The project was supported by the Danish Council for Independent Research (FKK) Grant DFF-7013-00074 awarded to Morten H. Christiansen and a seed grant from the Interacting Minds Center, Aarhus University.

References

Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., . . . Miller, J. (1991). The HCRC map task corpus. *Language, 34*(4), 351-366.

Bangerter, A., & Clark, H. H. (2003). Navigating joint projects with dialogue. *Cognitive Science, 27*(2), 195-225.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics, 5*, 135-146.

Bolis, D., Balsters, J., Wenderoth, N., Becchio, C., & Schilbach, L. (2017). Beyond autism: introducing the dialectical misattunement hypothesis and a bayesian account of intersubjectivity. *Psychopathology, 50*(6), 355-372.

Brugman, H., & Russel, A. (2004). *Annotating Multi-media/Multi-modal Resources with ELAN.* Paper presented at the Proceedings of LREC 2004, Fourth International Conference on Language Resources and Evalutation.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software, 80*(1), 1-28.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., . . . Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software, 76*(1).

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. *Perspectives on socially shared cognition, 13*, 127-149.

Clark, H. H., & Schaefer, E. F. (1987). Collaborating on contributions to conversations. *Language and cognitive processes, 2*(1), 19-41.

Colman, M., & Healey, P. (2011). *The distribution of repair in dialogue.* Paper presented at the Proceedings of the Annual Meeting of the Cognitive Science Society.

Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal, Complex Systems, 1695*(5), 1-9.

Dale, R., Fusaroli, R., Duran, N. D., & Richardson, D. C. (2013). The self-organization of human interaction. In *Psychology of learning and motivation* (Vol. 59, pp. 43-95): Elsevier.

De Ruiter, J., & Albert, S. (2017). An appeal for a methodological fusion of conversation analysis and experimental psychology. *Research on Language and Social Interaction, 50*(1), 90-107.

Dingemanse, M., & Enfield, N. J. (2015). Other-initiated repair across languages: towards a typology of conversational structures. *Open Linguistics, 1*(1).

Dingemanse, M., Kendrick, K. H., & Enfield, N. (2016). A coding scheme for other-initiated repair across languages. *Open Linguistics, 2*(1).

Dingemanse, M., Roberts, S. G., Baranova, J., Blythe, J., Drew, P., Floyd, S., . . . Manrique, E. (2015). Universal principles in the repair of communication problems. *PloS one, 10*(9), e0136100.

Dunbar, R. I., Marriott, A., & Duncan, N. D. (1997). Human conversational behavior. *Human nature, 8*(3), 231-246.

Duncan, S., & Fiske, D. W. (1977). *Face-to-face interaction: Research, methods, and theory*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Duran, N., Paxton, A., & Fusaroli, R. (accepted). ALIGN: Analyzing Linguistic Interactions with Generalizable techniques - a Python Library. *Psychological Methods*.

Fay, N., Ellison, T. M., Tylén, K., Fusaroli, R., Walker, B., & Garrod, S. (2018). Applying the cultural ratchet to a social artefact: The cumulative cultural evolution of a language game. *Evolution and Human Behavior, 39*(3), 300-309.

Fusaroli, R., Bahrami, B., Olsen, K., Roepstorff, A., Rees, G., Frith, C., & Tylén, K. (2012). Coming to terms: quantifying the benefits of linguistic coordination. *Psychological Science, 23*(8), 931-939.

Fusaroli, R., Rączaszek-Leonardi, J., & Tylén, K. (2014). Dialog as interpersonal synergy. *New Ideas in Psychology, 32*, 147-157.

Fusaroli, R., & Tylén, K. (2016). Investigating conversational dynamics: Interactive alignment,

Interpersonal synergy, and collective task performance. *Cognitive Science, 40*(1), 145-171.

Fusaroli, R., Tylén, K., Garly, K., Steensig, J., Christiansen, M. H., & Dingemanse, M. (2017). *Measures and mechanisms of common ground: backchannels, conversational repair, and interactive alignment in free and task-oriented social interactions.* Paper presented at the the 39th Annual Conference of the Cognitive Science Society (CogSci 2017).

Galantucci, B., & Roberts, G. (2014). Do we notice when communication goes awry? An investigation of people's sensitivity to coherence in spontaneous conversation. *PloS one, 9*(7), e103182.

Healey, P. G., Purver, M., & Howes, C. (2014). Divergence in dialogue. *PloS one, 9*(6), e98598.

Heldner, M., Hjalmarsson, A., & Edlund, J. (2013). *Backchannel relevance spaces.* Paper presented at the Nordic Prosody XI, Tartu, Estonia, 15-17 August, 2012.

Ireland, M. E., Slatcher, R. B., Eastwick, P. W., Scissors, L. E., Finkel, E. J., & Pennebaker, J. W. (2011). Language style matching predicts relationship initiation and stability. *Psychological Science, 22*(1), 39-44.

Jurafsky, D., Shriberg, E., Fox, B., & Curl, T. (1998). Lexical, prosodic, and syntactic cues for dialog acts. *Discourse Relations and Discourse Markers*.

Kidd, E., Donnelly, S., & Christiansen, M. H. (2018). Individual differences in language acquisition and processing. *Trends in cognitive sciences, 22*(2), 154-169.

Lavelle, M., Healey, P. G., & McCabe, R. (2014). Nonverbal behavior during face-to-face social interaction in schizophrenia: a review. *The Journal of nervous and mental disease, 202*(1), 47-54.

Levinson, P., Brown, P., & Levinson, S. C. (1987). *Politeness: Some universals in language usage* (Vol. 4): Cambridge university press.

Loth, S., Jettka, K., Giuliani, M., & De Ruiter, J. P. (2015). Ghost-in-the-Machine reveals human social signals for human–robot interaction. *Frontiers in psychology, 6*, 1641.

McCabe, R., & Healey, P. G. (2018). Miscommunication in Doctor–Patient Communication. *Topics in cognitive science, 10*(2), 409-424.

Pentland, A. (2012). The new science of building great teams. *Harvard Business Review, 90*(4), 60-69.

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences, 27*(2), 169-190.

Prinzo, O. V., & Britton, T. W. (1993). ATC/pilot voice communications-a survey of the literature: DTIC Document.

Reitter, D., & Moore, J. D. (2006). *Priming of syntactic rules in task-oriented dialogue and spontaneous conversation.* Paper presented at the Proceedings of the Annual Meeting of the Cognitive Science Society.

Schegloff, E. A. (1982). Discourse as an interactional achievement: Some uses of 'uh huh'and other things that come between sentences. *Analyzing Discourse: Text and talk, 71*, 93.

Tylén, K., Fusaroli, R., Smith, P., & Arnoldi, J. (2016). *The social route to abstraction.* Paper presented at the Cognitive Science 2016.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics Computing, 27*(5), 1413-1432.

Wadge, H., Brewer, R., Bird, G., Toni, I., & Stolk, A. (2018). Communicative Misalignment in Autism Spectrum Disorder.

Wiltshire, T. J., Butner, J. E., & Fiore, S. M. (2018). Problem-Solving Phase Transitions During Team Collaboration. *Cognitive Science, 42*(1), 129-167.

Yngve, V. (1970). *On getting a word in edgewise*, Papers from the Sixth Regional Meeting of Chicago Linguistic Society. Chicago Linguistic Society, Chicago.

The Goal Bias in Language and Memory: Explaining the Asymmetry

Monica L. Do (monicado@sas.upenn.edu)
Department of Psychology, Room 252, Stephen A. Levin Building, 425 S University Ave.
Philadelphia, PA 19104 USA

Anna Papafragou (papafragou@psych.udel.edu)
Department of Psychology, University of Delaware, 109 Wolf Hall
Newark, DE 19716 USA

John Trueswell (trueswel@psych.upenn.edu)
Department of Psychology, Room 254, Stephen A. Levin Building, 425 S University Ave.
Philadelphia, PA 19104 USA

Abstract

In language, speakers are more likely to mention the goals, or endpoints, of motion events than they are to mention sources, or starting points (e.g. Lakusta & Landau, 2005). This phenomenon has been explained in cognitive terms, but may also be affected by discourse-communicative factors: For participants in prior work, sources can be characterized as given, already-known information, while goals are new, relevant information to communicate. We investigate to what extent the goal bias in language (and memory) is affected when the source is or is not in common ground between speaker and hearer, and thus whether it is discourse-given or -new. We find that the goal bias in language is severely diminished when source and goal are discourse-new. We suggest that the goal bias in language can be attributed to discourse-communicative factors in addition to any cognitive goal bias. Discourse factors cannot fully account for the bias in memory.

Keywords: Source-Goal Asymmetry; Language Production; Goal bias; Discourse; Common Ground

Introduction

At their core, motion events involve movement of an object (i.e. the Figure) from a starting location (i.e. the Source) to an endpoint (i.e., the **Goal**; Talmy, 1983; cf. *A butterfly flew from a lamppost **to a chair***). Prior work has shown, though, that all these parts may not be "created equal". When talking about motion events, speakers are much more likely to mention the goal, or endpoint, of motion than they are to mention the source, or starting point (Lakusta & Landau, 2005, 2012; Papafragou, 2010; Regier & Zheng, 2007). This goal bias in language holds across ages (Papafragou, 2010; Lakusta & Landau, 2012; Lakusta, Muentener, Petrillo, Mullanaphy, & Muniz, 2016); different types of motion events (Lakusta & Landau, 2005, 2012); typologically different languages (e.g., Regier & Zheng, 2007; Johanson, Semilis, & Papafragou, in press); and even among deaf homesigners who lack exposure to conventional language (Zheng & Goldin-Meadow, 2002).

A similar goal bias has been shown in non-linguistic domains of cognition, such as memory, where goals have been shown to be more accurately encoded in memory than sources (e.g., Papafragou, 2010; Regier & Zheng, 2007; Regier, 1996). As in language, the goal bias in memory has been demonstrated across different types of motion events (Lakusta & Landau, 2012). And, has also been observed in pre-linguistic children (Lakusta, Wagner, O'Hearn, & Landau, 2007; Lakusta & Carey, 2015; Lakusta & DiFabrizio, 2017), suggesting that goals occupy a privileged, more salient status in non-linguistic as well as linguistic event representations.

Thus, in conjunction with a large body of work showing that infants attend to the goals or intentions of an event (e.g., Meltzoff, 1995; Bekkering, Wohlschläger, & Gattis, 2000), the presence of the goal bias in language and memory for *motion* events, also provides some basis to suggest that the linguistic bias has cognitive roots (e.g., Regier, 1996, Regier & Zheng, 2007; Srinavasan & Barner, 2013). Complicating this picture, though, is the fact that the goal bias in memory seems noticeably less robust compared to the goal bias in language. This is especially true when events no longer depict a prototypical *animate* agent moving from one inanimate reference point to another (Lakusta et al., 2007; Lakusta & Landau, 2012; Lakusta & Carey, 2015; Lakusta & DiFabrizio, 2017). In cases like these, some researchers have failed to find evidence of the goal bias in memory – even when the same studies have found a clear goal bias in language and even when the same materials have been used across linguistic and non-linguistic tasks (Lakusta & Landau, 2012).

The discrepancy between the strength of the goal bias in language and memory has been difficult to reconcile with claims that the goal bias is fundamentally rooted in the same (cognitive) mechanism in both domains. In particular, if the mechanism responsible for the goal bias in language is also responsible for the goal bias in memory, why doesn't the bias appear to work in precisely the same way across domains?

The present work proposes a novel explanation for the observation that the strength of the goal bias in linguistic production tasks is more robust than in non-linguistic tasks. We posit that the comparatively more robust goal bias in language may be attributable to an *additional* discourse/communicative asymmetry: When individuals are asked to describe video clips of simple motion events, the initial state of affairs – including the source (i.e., starting point) of the motion – is reasonably assumed to be given. By

contrast, the goal of the motion event (i.e., the endpoint) is considered 'the news' that is relevant to communicate. This makes sources less likely to be mentioned (see Lakusta & Landau, 2012 for a discussion of this possibility). To preview our results, we find evidence in support of this discourse/communicative account: Changing the discourse/communicative status of the source in motion events severely weakens (but does not eliminate) the goal bias in language. We conclude that in language, discourse/communicative factors operate over and above the more general cognitive factors that might drive the goal bias observed in memory.

The Current Study

Prior work on linguistic aspects of the goal bias has typically involved a single participant, who (i) sees a figure located at or near the source (i.e. starting point) of the motion event, (ii) presses a button to watch the event unfold, and then, (iii) describes the event out loud to either no one in particular or a physically co-present, but conversationally unengaged experimenter. Because motion clips in these paradigms typically begin with a scene that sets up the start of the event, the source can be considered already known, 'discourse-old', information, while the goal is considered the 'discourse-new', relevant piece of the event.

Given that speakers have a preference to mention discourse-new over discourse-old (i.e. "given") information in their utterances (Arnold, Wasow, Losongco, & Ginstrom, 2000), a consequence of this single-speaker paradigm may be that it inadvertently creates the conditions for a goal bias both in speakers' descriptions and their representations of motion events in memory. Specifically, participants who do not have to take into account the knowledge state of their interlocutor prioritize mentioning only what is new and relevant to themselves or a 'generic' addressee – in this case, the goal or endpoint of the motion event.

Unlike prior work, the current study asks participants to describe motion events to an attentive, engaged confederate addressee. The presence of an engaged addressee allows us to probe whether the goal bias in language can at least partially be attributed to an asymmetry in the *discourse/communicative status* of sources (typically presented as known, discourse-given entities) versus goals (typically unknown, discourse-new entities) in motion events. This is because the introduction of an addressee allows speakers to consider not only what is discourse-new to themselves, but also what is discourse-new (and presumably relevant to communicate) to their interlocutor.

This discourse/communicative account of the goal bias predicts that changes to the discourse status of the source should affect the magnitude of the goal bias in language. In particular, we expect the goal bias to weaken when sources are also made discourse-new. Alternatively, if the goal bias in language and memory is *purely* driven by a more general cognitive bias towards goals, then changing the communicative setting in which motion events are described should not affect the magnitude of the goal bias in language.

To test the discourse/communicative account, we manipulated the context in which participants described motion events. Participants in our Common Ground condition were asked to describe the motion event to a confederate addressee for whom information about the starting point of the motion was already known – that is, the source constituted discourse-given information. By contrast, participants in our No Common Ground condition were asked to describe the motion event to a confederate addressee that knew nothing about the upcoming motion event – that is, both the source and goal constituted discourse-new, relevant information to communicate about.

Following prior work, we investigated the goal bias in *language* by comparing the proportion of source versus goal mentions as participants describe motion events. We investigated the goal bias in *memory* using an adaptation of the change detection paradigm; we compared how accurately speakers remember sources versus goals *after describing events to an addressee*.

Methods

Participants Fifty-four native speakers of American English (mean age = 20; 28 male, 32 female) participated in the experiment for course credit or $10/hour – 27 in the Common Ground and 27 in the No Common Ground group. The number of participants was determined based on a power analysis of previously reported effects in the literature.

Materials We created 18 test clips, each of which depicted an animate entity moving from an inanimate source landmark (i.e. the starting point of motion) to an inanimate goal landmark (i.e. the end point of motion). (Clipart images were used. See Figure 1a for an example of a butterfly moving from a lamppost (the source) to a chair (the goal)). Each clip was roughly three seconds in length.

Clips were left-right counterbalanced such that half of our clips showed a figure moving from a source on the left to a goal on the right and the other half showed a figure moving from a source on the right to a goal on the left. Source and goal landmarks were also counter-balanced across lists such that objects which were the source of motion in one list were the goals of motion in another. This was done to ensure that our results would not be confounded by the inherent perceptual or conceptual salience (and by extension, salience in linguistic mentions or memory) of one landmark over the other. We also included 18 filler motion events, which did not involve motion between a source and a goal. These filler items were designed such that participants were not able to predict, based on the first frame of the video, whether the clip would eventually involve a source-to-goal motion event.

We probed speakers' encoding of these events in memory using a version of the change detection task used by prior work (Regier & Zheng, 2007; Lakusta & Landau, 2012; Papafragou, 2010). For this, we constructed a second set of videos that involved: (i) Changing the Source (ii) Changing the Goal; or (iii) No Changes (i.e., participants saw a video identical to the one they had previously described). Source and goal changes were always replaced with within-category

variants (e.g., the chair was changed to a different example of a chair; Figure 1) to control for the semantic distance between the original and changed object.

(Figure 1a)

(Figure 1b)

Figure 1. (a) Sample first frame of the 'butterfly flying from the lamppost to the chair' clip. In Common Ground conditions, both participant and addressee saw the first frame; in No Common Ground conditions, only participants saw this. Only participants saw events unfold. (b) Sample goal change in the memory task. The original chair was replaced with a slightly different chair.

Procedure Participants were told that they would be performing the experiment with a partner (in reality, a confederate addressee). Participants were told that they would be watching brief video clips and then describing them to their partner. Their partner would see a simple question about the clip on a separate screen and would answer those questions based on the participant's descriptions.

To demonstrate that the addressee was engaged in the experiment, participant and confederate addressee completed a Tower of Hanoi task together. Afterwards, participants performed two practice trials before moving on to the main experiment. Because prior work has shown that the level of engagement of an addressee can affect how much information speakers choose to include in their utterances (Clark & Wilkes-Gibbs, 1986; Bavelas, Coates, & Johnson, 2000) and may also affect speakers' later memory for the event (Pasupathi, Stallworth, & Murdoch, 1998), confederate addressees maintained eye-contact during event descriptions and verbally indicated when they were ready for the next trial (i.e., 'mhmm', 'ok', 'yup', 'I'm ready'). Critically, confederates maintained the same level of engagement in all conditions and used the same verbal indicators regardless of the utterance produced by participants.

Participants were seated in one of two experimental configurations. In the Common Ground condition (Figure 2), both speaker and confederate addressee were seated side-by-side in front of a centrally-located computer screen. Each trial began with the first frame of the video clip shown on this screen. Thus, both speaker and confederate addressee saw the figure's location relative to the source and the goal landmarks; more specifically, they saw where the animate figure started out in each clip. After briefly inspecting the scene, the addressee turned the participants' screen away so that the addressee was not able to watch the clip unfold.

In the No Common Ground condition (Figure 3), speaker and confederate addressee were seated across from each other so that neither could see each other's screens. Speakers were thus led to believe that addressees in this condition were unable to see *any* part of the video clip.

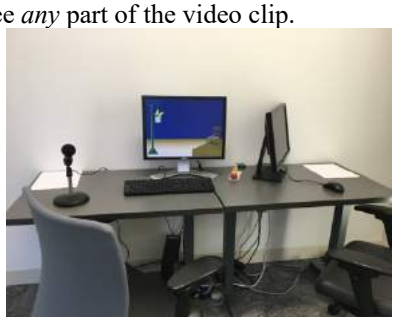


Figure 2. Common Ground configuration. Participants were always seated on the left, confederate addressees on the right. Confederate addressees were shown the first frame of the clip on the participant's screen before turning the screen away.

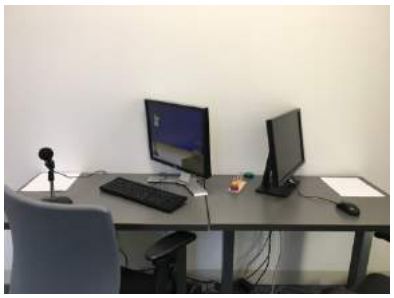


Figure 3. No Common Ground configuration. Participants were always seated on the left, confederate addressees on the right. In the No Common Ground condition, confederate addressees were not permitted to see any part of the participant's computer screen.

In both Common Ground and No Common Ground conditions, participants received the same set of video stimuli and participants performed the same tasks. They were told in all cases that confederate addressees would be answering a simple question about each video clip based on the speakers' description of what happened in each clip.

After finishing the description portion of the experiment, participants were separated from the confederate and participants were given a surprise memory task. During this portion of the experiment, participants were shown the (i) Source Change, (ii) Goal Change, or (iii) No Change variants of the test videos. This was a within-participant manipulation such that six items were randomly assigned the Source Change condition, six to the Goal Change, and six the no Change. Participants were told to circle 'Yes' on their answer sheet if the second video clip was 'exactly the same' as the clip that they had originally described; they were told to mark 'No' otherwise. Thus, correct responses in the Source and Goal Change conditions were always 'No' (i.e., they correctly rejected), but correct responses in No Change condition was always 'Yes' (i.e., they correctly failed to reject). Participants were only tested for memory of target items; clips in the memory portion of the study were presented in the same order as in the scene description portion

of the study.

Predictions

In the Common Ground condition, where the addressee was allowed to see the starting point of the motion event, the source (as in prior work) was discourse-given. However, in the No Common Ground condition, where the addressee was not privy to any information about the motion event, both source and goal were discourse-new.

On a purely cognitive account of the goal bias, the discourse status of entities in a motion event should not affect the frequency of mention of sources and goals. By contrast, on a discourse/communicative account of the goal bias, the goal bias in language should be severely weaker in the No Common Ground condition – where both sources as well as goals were discourse-new – than in the Common Ground condition – where only the goal was discourse-new.

A somewhat independent question is how linguistic descriptions of motion should affect later memory of that motion event. One possibility is that the generation of more informative linguistic representations implies the prior generation of more informative non-linguistic representations. If so, then we expect the patterns in the linguistic description portion of our study to largely correspond to the patterns that emerge in the memory portion of the study. That is, memory for sources should be more accurate in the No Common Ground condition than in the Common Ground condition, where relevant information was not just limited to the goal of motion. It is also possible, though, that there may be no direct relationship between what is mentioned in the motion event and what is subsequently remembered. For instance, even if speakers were more likely to talk about sources in the No Common Ground condition, memory for sources might nevertheless remain relatively impoverished compared to memory for goals.

Results

Language Productions We were primarily interested in how frequently speakers would mention sources relative to goals in their event descriptions. We coded whether each utterance included mention of the source and/or goal of the motion event. Following prior work (e.g. Lakusta and Landau, 2012; Papafragou, 2010), all mentions of sources and goals within (i) a prepositional phrase (e.g. 'from the chair'; 'off the chair'; 'to the chair'; etc.), (ii) within the verb + NP structure (e.g. 'left the cave'), or (iii) within a verb + particle structure (e.g. verb + 'away from the tree') were included.

Statistical analyses of the rate at which sources and goals were mentioned were done using a logistic mixed effect model. Ground Type (Common Ground vs. No Common Ground) and Mention Type (Source vs. Goal) were included as fixed effect factors. Mention Type was included as part of the by-subject and by-item random effects; Ground Type was only included as part of the by-item random effects. We simplified the model only if it failed to converge or if random effects did not significantly improve model fit.

As can be seen in Figure 4, in both the Common Ground and No Common Ground conditions, we replicated the goal bias observed in prior work (Lakusta & Landau, 2005, 2012; Papafragou, 2010). In the No Common Ground condition, though, the goal bias was severely weakened: The preference to mention the goal over the source was greater in the Common Ground condition than in the No Common Ground condition. Consistent with this, we detected significant main effects of Mention Type (ß= 3.26, SE= 0.59, $|z| = 5.51$, $p < .01$) and Ground Type (ß= 1.58, SE= 0.75, $|z| = 2.11$, $p < .05$), but these were modulated by a reliable Ground x Mention interaction (ß= -3.06, SE= 0.99, $|z| = 3.11$, $p < .01$).

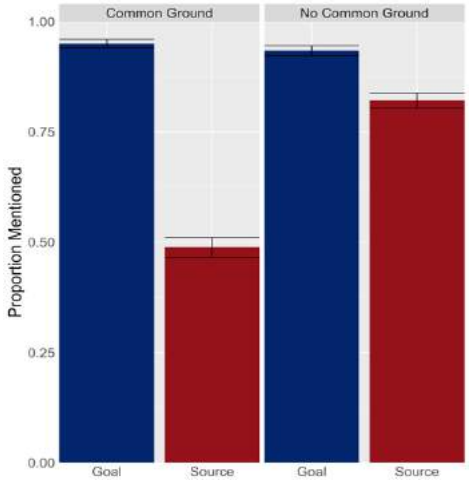


Figure 4. Proportion of Source (Red) and Goal (Blue) mentions in Common Ground and No Common Ground conditions. Error bars represent +/- 1 standard error.

Memory for Sources and Goal Accuracy in the memory task was analyzed using logistic mixed effects regressions. We included Ground Type and Change Type (Source Change or Goal Change) as fixed effects. Random effects were structured as before. The No Change condition was omitted from this analysis because it served only as an indicator of baseline performance and indeed, was similar in both Common Ground and No Common Ground Conditions (Figure 5; yellow).

Overall, participants were more accurate in the No Common Ground than in the Common Ground conditions, resulting in a significant main effect of Ground Type (Figure 5; ß = 1.07, SE = 0.38, $|z| = 2.82$, $p < 0.01$). Participants were significantly better at detecting changes to the Goal (Blue) than to the Source (Red) in both the Common Ground and No Common Ground conditions (ß = 0.90, SE = 0.22, $|z| = 4.19$, $p < .001$). However, the failure to detect a significant Ground x Change interaction (ß = -0.17, SE = 0.42, $|z| = 0.39$, $p = 0.69$) suggests that the strength of the goal bias in the Common Ground vs. No Common Ground conditions did not differ statistically. In other words, speaking to an addressee in the No Common Ground condition only had the effect of boosting speakers' memory for the event *more generally*. Unlike in the description task, it does not appear to weaken any goal bias that exists in memory encoding processes.

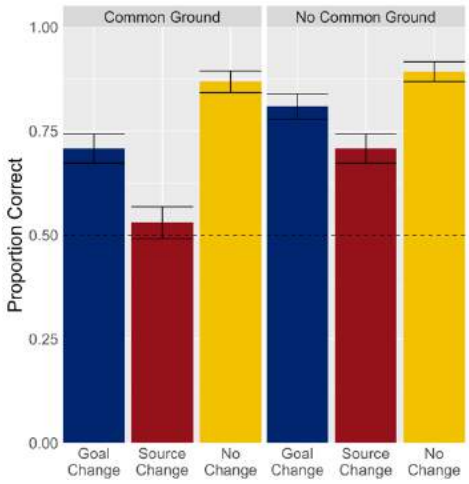


Figure 5. Proportion of Correct Responses in Memory Task. Error bars indicate +/- 1 standard error. Dashed horizontal line indicates chance performance in each condition.

The 'Source Mention Benefit' We used a logistic mixed effect regression to determine whether the rate at which speakers mentioned sources in the language task would predict how accurately they detected source changes in the later memory task. Ground Type and Source Mention (yes, no) were included as fixed effects. Source Mention was included in both by-subject and by-item random effects; Ground Type was only included by-items. Models were reduced and selected as before.

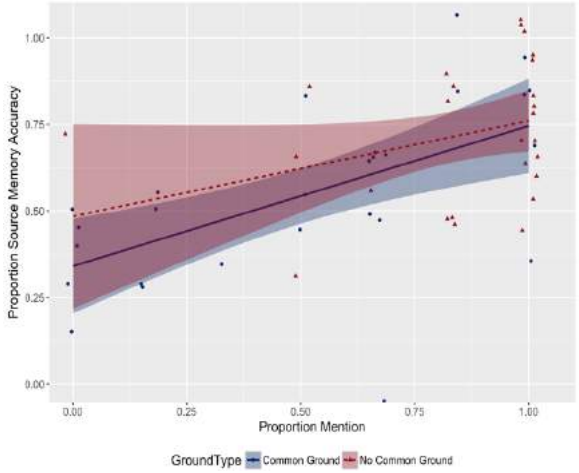
We found that speakers were more likely to accurately encode sources in memory if they had previously mentioned the source in their descriptions (Figure 6). This was indexed by a significant main effect of Source Mention (ß = 1.17, SE = 0.43, |z| = 2.73, p < .01). There was a marginally significant main effect of Ground Type (ß = 1.08, SE = 0.56, |z| = 1.94, p = .053), suggesting that speakers' memory for sources trended towards being more accurate in the No Common Ground than in the Common Ground condition. There was no significant Ground x Source Mention interaction (ß = 0.16, SE = 0.83, |z| = 0.20, p = 0.84). This latter finding suggests that mentioning the source provided the same benefit to source accuracy in the memory task, regardless of Common Ground or No Common Ground condition.

Recall that in the memory task participants were also better at detecting changes to the goal in the No Common Ground as compared to the Common Ground Condition (the two blue bars in Figure 5). This is surprising given that participants mentioned the goal to the same extent in the Common Ground and No Common Ground conditions. One possibility is that goals were remembered more accurately in the No Common Ground condition because mentioning the source (which happened more in this condition) helped to create a more coherent representation of the event as a whole.

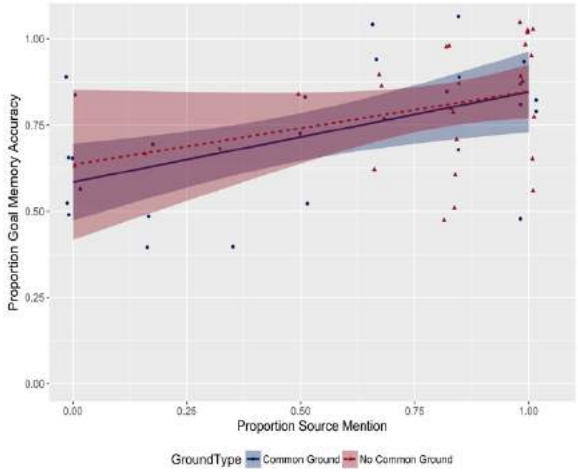
To investigate this, we analyzed whether the rate of source mention would also predict *goal* accuracy in the memory task. We found that participants who mentioned sources more frequently also tended to remember goals more accurately (Figure 6b; ß = 1.09, SE = 0.35, |z| = 3.10, p < .01). No other effects were significant, meaning that the magnitude of the source mention benefit on goal memory did not differ across Common Ground and No Common Ground conditions. Thus, mentioning the source had the secondary benefit of reinforcing memory for other aspects of the motion event – namely, the goal – as well.

Discussion

Prior work has shown a robust goal bias in language: Speakers are much more likely to mention the goal (i.e. endpoint) of a motion event than they are to mention the source (i.e. starting point) of that event. A goal bias has also



(Figure 6a)



(Figure 6b)

Figure 6. Performance in the description and memory task for each subject in the Common Ground (Blue circles) and No Common Ground (Red Triangle) conditions. The x-axis represents proportion of times sources were mentioned during the description task. The y-axis of Figures 6a and 6b show the proportion of accurate responses in the Source Change and Goal Change conditions, respectively. Shaded areas represent +/- 1 standard error.

been observed in non-linguistic cognitive domains, such as memory, suggesting that the goal bias may operate across domains of cognition. An open question and central challenge for such cognitive accounts, though, is how to account for the fact that the goal bias is much more robust across contexts in language than in memory.

We suggest that the discourse/communicative context in which motion descriptions were elicited in prior work exacerbated the goal bias in language: On top of any underlying cognitive goal bias, speakers were additionally more likely to mention the goal than the source because they were describing events in a discourse context that made mentioning the source unnecessary. In our Common Ground condition, where information about the source was discourse-given, we replicated the goal bias in prior work. When we equalized the discourse/communicative status of sources and goals (our No Common Ground condition), this goal bias was drastically weakened. These results are expected if the goal bias – at least in language – is not a reflection of cognitive factors alone.

It is work noting that other work (e.g., Stevenson, Crawley, & Kleinman, 1994) has independently reported a bias for the goal in transfer-of-possession events (i.e., *Leslie handed a book to Ann.*). There, work by Rodhe, Kehler, & Elman (2006) has similarly argued that discourse factors – like different types of coherence relations – can also modulate the goal bias in transfer-of-possession events. We do not manipulate factors like coherence here, but our results also demonstrate the way that discourse factors can interact with event representations in language. Further, our work suggests that (in addition to coherence relations) the goal bias in those cases may also be partially attributed to the givenness of the source in transfer-of-possession events.

Importantly, our results do not rule out the possibility of a cognitive bias towards goals/endpoints: Across tasks and conditions, we found a residual preference to mention goals over sources. Even in the No Common Ground condition, speakers were still more likely to mention goals over sources; and, though participants performed more accurately on the memory task in general, they nevertheless remembered goals more accurately than sources.

By contrast, the discourse status of the source did not *directly* influence the goal bias in memory. However, we did find evidence of an *indirect* 'source mention benefit' that affected how accurately goals, as well as sources, were encoded in memory: Speakers who were more likely to mention the source in their event descriptions were more accurate in remembering *both* goals and sources of motion events. For these participants, mentioning sources improved memory for sources themselves, but also helped to create a more accurate representation of the event more generally.

More broadly, we conclude that discourse/communicative factors should be incorporated into theories about the relationship between language and event cognition. Moreover, our results are consistent with prior work (e.g. Clark & Wilkes-Gibbs, 1986) showing that the extent of the addressee's knowledge has a direct effect on what information speakers choose to include in their utterance.

One question raised by our results is whether the discourse/communicative status of *goals* can modulate the mention of goal phrases in language. Is it possible, for instance, to reverse the goal bias (i.e., produce a source bias) strictly by manipulating the givenness vs. newness of goals? We are currently exploring this direction in ongoing work.

Acknowledgements

We thank Nathaniel Robinson and Victor Gomes for their help with experiment design, data collection, and coding the data. We also wish to thank the Language Development Lab at the University of Pennsylvania for their helpful feedback on this work. This project was funded by NSF grant #1632849 awarded to Anna Papafragou.

References

Arnold, J., Wasow, T., Losongco, A., & Ginstrom, R. (2000). Heaviness vs. newness: the effects of complexity and information structure on constituent ordering. *Language*, 76, 28-55.

Bavelas, J. B., Coates, L., & Johnson, T. (2000). Listeners as Co-Narrators. *Journal of Personality and Social Psychology*, 79, 941–952.

Bekkering, H., Wohlschläger, A, & Gattis, M. (2000). Imitation of gestures in children is goal-directed. *Quarterly Journal of Experimental Psychology*, 53, 153-164.

Clark, H. H. Clark & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, 22, 1-39.

Johanson, M., Selimis, S., & Papafragou, A. (in press). The Source-Goal asymmetry in spatial language: Language-general vs. language-specific aspects. *Language, Cognition and Neuroscience.*

Lakusta, L., & Carey, S. (2015). Twelve-month-old infants' encoding of goal and source paths in 'agentive' and 'non-agentive' motion events. *Language Learning and Development*, 11, 152–175.

Lakusta, L., & DiFrabrizio, S. (2017). And, the Winner Is…A Visual Preference for Endpoints over Starting Points in Infants' Motion Event Representations. *Infancy*, 23, 323–343.

Lakusta, L., & Landau, B. (2005). Starting at the end: The importance of goals in spatial language. *Cognition*, 96, 1–33.

Lakusta, L., & Landau, B. (2012). Language and Memory for Motion Events: Origins of the Asymmetry Between Source and Goal Paths. *Cognitive Science*, 36, 517–544.

Lakusta, L., Muentener, P., Petrillo, L., Mullanaphy, N., & Muniz, L. (2016). Does Making Something Move Matter? Representations of Goals and Sources in Motion Events With Causal Sources. *Cognitive Science*, 41, 1–13.

Lakusta, L., Spinelli, D., & Garcia, K. (2017). The relationship between pre-verbal event representations and semantic structures: The case of goal and source paths. *Cognition*, 164, 174-187.

Lakusta, L., Wagner, L., O'Hearn, K., & Landau, B. (2007). Conceptual foundations of spatial language: Evidence for

a goal bias in infants. *Language Learning and Development*, 3, 179–197.
Meltzoff, A. N. (1995). Understanding the intentions of others: Re-enactment of intended acts by 18-month-old children. *Developmental Psychology*, 31, 838–850.
Papafragou, A. (2010). Source-goal asymmetries in motion representation: Implications for language production and comprehension. *Cognitive Science*, 34, 1064–1092.
Pasupathi, M., Stallworth, L. M., & Murdoch, K. (1998). How what we tell becomes what we know: Listener effects on speakers' long-term memory for events. *Discourse Processes*, 26, 1–15.
Regier, T. (1996). *The human semantic potential: Spatial language and constrained connectionism.* Cambridge, MA: MIT Press.
Regier, T., & Zheng, M. (2007). Attention to endpoints: A cross-linguistic constraint on spatial meaning. *Cognitive Science*, 31, 705–719.
Rohde, H., Kehler, A. and Elman, J. (2006). Event Structure and Discourse Coherence Biases in Pronoun Interpretation. Proceedings of the Annual Meeting of the Cognitive Science Society, 28, 697-702.
Srinivasan, M., & Barner, D. (2013). The Amelia Badelia effect: World knowledge and the goal bias in language acquisition. *Cognition*, 128, 431-450.
Stevenson, R., Crawley R., & Kleinman, D. (1994). Thematic roles, focusing and the representation of events. *Language and Cognitive Processes*, 9, 519–548.
Talmy, L. (1983). How language structures space. In H. Pick & L. Acredolo (Eds.), *Spatial orientations: Theory, research, and application* (pp. 225–282). New York: Plenum Press.
Zheng, M. & Goldin-Meadow, S. (2002). Thought before language: how deaf and hearing children express motion events across cultures. *Cognition*, 85, 145-175.

A rational model of word skipping in reading: ideal integration of visual and linguistic information

Yunyan Duan (yduan@u.northwestern.edu)
Department of Linguistics, Northwestern University
Evanston, IL 60201 USA

Klinton Bicknell (klinton@duolingo.com)
Duolingo AI Research
Pittsburgh, PA 15206 USA

Northwestern University
Evanston, IL 60201 USA

Abstract

During reading, readers intentionally do not fixate a word when highly confident in its identity. In a rational model of reading, word skipping decisions should be complex functions of the particular word, linguistic context, and visual information available. In contrast, simple heuristic of reading only predicts additive effects of word and context features. Here we test these predictions by implementing a rational model with Bayesian inference, and predicting human skipping with the entropy of this model's posterior distribution. Results showed a significant effect of the entropy in predicting skipping above a strong baseline model including word and context features. This pattern held for entropy measures from rational models with a frequency prior but not from ones with a 5-gram prior. These results suggest complex interactions between visual input and linguistic knowledge as predicted by the rational model of reading, and a dominant role of frequency in making skipping decisions.

Keywords: eye movements; reading; word identification; rational analysis; skipping

Introduction

To achieve comprehension in reading, readers move their eyes across the text to obtain the information needed to identify the words. In the past decades, research on eye movements in reading has provided ample evidence that word identification can be seen as the primary driver of eye movements. The reasoning behind this conclusion, however, is based on relatively coarse observations, such as demonstrating that eye movements are sensitive to aggregate variables that are important in word identification (e.g., word length and frequency). Although such a coarse linking hypothesis between word identification and eye movements successfully predicts several reading behaviors, a model of reading that connects eye movements to ongoing language processing in a deeper way could lead to more precise predictions, improved data analysis, and an overall fuller utilization of the eye movement record to advance theories of sentence processing.

One promising model of this type comes from a perspective of rational analysis. The idea is to consider the reading process as one that combines information from various sources to identify words and then makes eye movement decisions to maximize identification efficiency (Bicknell & Levy, 2010, 2012; Legge, Klitz, & Tjan, 1997; Legge, Hooven, Klitz, Mansfield, & Tjan, 2002). In previous rational models of reading, text identification process is modeled using Bayesian inference that combines two sources of information: (1) probabilistic knowledge of the structure of the language, serving as the prior, and (2) uncertain visual evidence, serving as the likelihood. Given a prior and a particular set of visual evidence, probabilistic inference yields a posterior distribution on the text, which specifies the probability of each possible identity of the text. The role of eye movements in this analysis is to obtain particular pieces of visual evidence, and the most efficient, rational reading behavior will be to use the current posterior distribution to determine the most useful time and place to move the eyes next. Therefore, any eye movement behaviors explained by this model of reading can be seen as naturally born from one simple origin: the rational gathering of visual evidence for text identification.

In contrast, the dominant models of eye movement control in reading tend to use heuristic linking hypothesis between text identification and eye movements (e.g., E-Z Reader, Reichle et al., 2009; and SWIFT, Engbert et al., 2005). In these models, eye movements are driven by a word identification process that is represented with discrete states (e.g., not identified, partially identified that leads to saccade programming, fully identified), the transitions between which depend on a certain amount of durations computed from a few coarse visual and linguistic variables of the word. For example, in E-Z Reader the duration of L1 and L2 depend on a stochastic function of two linguistic variables, the word's frequency in the language and its predictability in context, and one visual variable, the average distance from each of its letters to the point of fixation. After spending the pre-computed duration needed to achieve a certain stage of word identification to begin programming a saccade and then achieve complete identification of the current word, the model moves eyes to (roughly) the center of the next word to be identified. The role of eye movements in this heuristic model is a direct reflection (with stochastic noise) of cognitive process identifying a word, the difficulty of which depends on coarse properties of the word as a whole.

There are situations where word identification can be completed with more fine-grained knowledge about the particular word than merely coarse information, and we would like to make precise predictions about eye movement behaviors accordingly. Consider situations where visual information about only the beginning of some words is enough for identification, e.g., seeing the initial letters 'xyl' of the word 'xylophone' (Hyönä, Niemi, & Underwood, 1989). Similarly, in certain linguistic contexts, a reader only needs to see a few of

the initial letters of a word to be confident in its identification, such as in 'The children went outside to pl…'. Do readers in fact combine more fine-grained information than simply word frequency and word length in the way as predicted by rational models of reading?

As illustrated in the preceding examples, an ideal testbed for these predictions of a rational model is when a word is identifiable with visual information about only part of the word. In natural reading, this situation occurs often in the eye movement behavior of skipping, when a reader move their eyes past a word without ever having directly fixated it. Intentionally skipping a word is generally modeled as a case in which the reader has identified the word (possibly incorrectly) while still looking at a prior word, and thus makes a saccade that takes the eyes past the word, skipping over it. Since this (implicit) decision about whether to skip the word is made when the reader is fixating a prior word, this is a case when the reader has high quality visual information about only some of the word's initial letters but does not yet have high quality visual information about the whole word. The amount of visual information the reader has at this time is a function of the *launch site*, the distance from the fixation position to the beginning of the word. In such a situation, both the rational model and the heuristic model predict that how likely a reader is to skip a word should be a function of launch site (amount of visual input), and also of linguistic knowledge (which words are common, and which words are likely in this position). The rational model alone additionally predicts that readers' likelihood of skipping the word will vary depending on the *particular* visual information obtained, and whether that information distinguishes it strongly from its (likely) visual neighbors. Therefore, skipping should be observed to be a complex function of the launch site, the particular word, and linguistic knowledge, in contrast to the heuristic model's predictions of skipping as well-described by coarse visual and linguistic information about the whole word.

Previous empirical research finds that readers' likelihood of skipping a word increases with short word length, close launch sites to the word, high word frequency, and high contextual predictability (Rayner, 1998). Regarding how different sources of information may interact in skipping, studies of skipping short words and especially the word *the* suggest that visual information and word frequency information trump information from the sentence context (Angele & Rayner, 2013; Angele, Laishley, Rayner, & Liversedge, 2014). Despite these findings, the fine-grained predictions of a rational model can be better tested with a set of eye movement decisions that happen in natural reading and that have wide variation in visual and linguistic information available to the reader. The goal of the current paper is to directly test the fine-grained predictions using word skipping, and to gain insights into the role of different sources of information in making skipping decisions.

Related work

Empirical findings about skipping

At the aggregate level, the effects of visual and linguistic variables on skipping are very robust. Word length is considered to play a more important role than any other factors, as found in a meta-analysis showing that word length explained more variance than word frequency and predictability in regression models predicting skipping rate (Brysbaert, Drieghe, & Vitu, 2005). The effect that close launch sites increase skipping rates is also strong and robust (Brysbaert et al., 2005). As for linguistic variables, there is abundant experimental evidence that skipping rate increases as word frequency increases (Rayner, Sereno, & Raney, 1996; Angele et al., 2014), and that high predictability leads to high skipping rate (Balota, Pollatsek, & Rayner, 1985; Rayner, Slattery, Drieghe, & Liversedge, 2011). Predictability is usually measured as cloze probability, varying across conditions either with different sentential frames or target words (Balota et al., 1985; Rayner et al., 2011). The effects hold in corpus analysis as well, as Luke and Christianson (2016) find that high target predictability lead to more word skipping for both content and function words. Kliegl, Grabner, Rolfs, and Engbert (2004) also find significant effect of predictability, word length, and word frequency on skipping rate using regression analyses on Potsdam Sentence Corpus, though not including any interactions among these factors.

Several studies have looked into the interactions between visual and linguistic factors on a coarse level. One approach is to analyze linguistic effects on data split in launch sites in post-hoc analysis. For example, Rayner et al. (1996) observe reliable frequency effect on skipping rate at near launch sites (> -5) but not at far launch sites, and White, Rayner, and Liversedge (2005) find significant interaction between predictability and word length preview overall, which diminish to non-effect for far launch sites (near launch sites are defined as those ≥ -3, while far launch sites are those ≤ -4). Another approach to study the interaction of visual and linguistic information is to manipulate parafoveal preview. A preview of the definite article *the* increases readers' skipping rate, even when syntactic constraints do not allow for articles to occur in that position (Angele & Rayner, 2013; Angele et al., 2014). Skipping rates are higher for the preview of a highly predictable word or its visually similar nonword counterpart than the preview of a low-predictability word (Balota et al., 1985), and for the preview of a predictable word than for a visually similar nonword (Drieghe, Rayner, & Pollatsek, 2005). Staub and Goddard (2019) observe that frequency effect on skipping rate is maintained with both valid and invalid preview, but predictability influences skipping only with valid preview. Additionally, English readers only benefit from the preview of a semantically similar neighbor in highly-constraining context but not in moderate-constraining context (Schotter, Lee, Reiderman, & Rayner, 2015).

In sum, previous research have identified visual and linguistic factors that influence skipping by conducting reading

experiments and corpus studies. There is also evidence for interactions between visual and linguistic factors, but they are constrained to a small set of well-controlled language materials and analyzed on a coarse level. A systematic analysis with skipping made for a variety of words in a variety of contexts with a variety of launch sites would help gain insights into how visual and linguistic variables interacts to identify a word before fixating it and skip at a fine-grained level.

Other instances of rational models of reading

Previous instances of the rational models of reading have provided explanation for several eye movement phenomena. For example, they explain why the initial fixation tends to land near word center and is affected by the launch distance (Legge et al., 2002), why readers often make regressions to previous words (Bicknell & Levy, 2010), and why high-frequency and low-surprisal words yield lower reading difficulty than low-frequency and high-surprisal words (Bicknell & Levy, 2012). In the field of single word identification, Duan and Bicknell (2017) implement a rational model of refixations, and find that readers rationally make refixations to seek visual information from parts of the word that the readers are uncertain.

The rational model of skipping presented in this paper has different focuses than previous models. Instead of setting the goal to be identifying a whole sentence, the rational model of skipping focus on identifying a single word before directly fixating it. Previously, the computational cost is high due to recomputing posterior beliefs about an entire sentence after each new piece of visual evidence. The model of skipping is computationally simple, enabling the incorporation of sophisticated models of language knowledge and visual evidence.

Rational model of skipping

Word identification as Bayesian inference

In our rational model of skipping, word identification uses Bayesian inference, in which a prior distribution over possible identities of the word given by the language model is combined with a likelihood term given by 'noisy' visual input conditional on the fixation position to form a posterior distribution over the identity of the word. Formalized with Bayes' theorem,

$$p(w|\mathcal{I}) \propto p(w)p(\mathcal{I}|w) \tag{1}$$

where the probability of the true identity of the word being w given uncertain visual input $\mathcal{I}$ is calculated by multiplying the language model prior $p(w)$ with the likelihood $p(\mathcal{I}|w)$ of obtaining this visual input from word w, and normalizing. Since the shape of the posterior distribution depends on the probability of each word relative to probabilities of other words in the vocabulary, it contains information about how well a word is distinguished from its neighbors.

In general, the prior $p(w)$ represents reader expectations for the next word, and for the present paper, we compare two representations of the prior: a word unigram model (i.e., using word frequency information), which ignores any context information, and a 5-gram model, which conditions on the previous four words of context. The likelihood $p(\mathcal{I}|w)$ represents how likely a piece of visual input is from a word w. For the present paper, we assume that all visual input is obtained only from the final fixation position prior to either fixating the word or skipping it (i.e., the launch site). The visual input obtained about a word consists of independent visual input obtained from each letter in it. Each letter is represented as a one-hot 52-dimensional vector (distinguishing 26 lower- and upper-case letters), with a single element being 1 and the rest being 0. Visual input about each letter is accumulated iteratively over time by sampling from a multivariate Gaussian distribution centered on that letter with a diagonal covariance matrix $\Sigma = \lambda^{-1}I$, where λ is the reader's visual acuity for that letter. Visual acuity depends on the location of the letter in relation to the point of fixation, or eccentricity, which we denote ε. Similar to Bicknell and Levy (2010), we assume that acuity is a symmetric, exponential function of eccentricity:

$$\lambda(\varepsilon) = \int_{\varepsilon-.5}^{\varepsilon+.5} \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{x^2}{2\sigma^2})dx \tag{2}$$

with $\sigma = 3.075$, the average of two σ values for the asymmetric visual acuity function ($\sigma_L = 2.41$ for the left visual field, $\sigma_L = 3.74$ for the right visual field) used in Bicknell and Levy (2010). In this paper, we take σ, the effective width of the visual field, as a free parameter, and experiment with a set of σ values. In addition, we introduce another free parameter Λ to scale the overall quality of visual information by multiplying it with each acuity λ (see the Experiment section below).

Single word belief updating

Given visual information and linguistic expectations, we may thus compute a posterior distribution over possible identities of the word. Since visual information arrives over time, this is a Bayesian belief updating process, where beliefs are updated as each new piece of visual information arrives. In the single word domain we study here, this Bayesian belief updating process turns out to be relatively computationally simple, and can be implemented as sampling from a multidimensional Gaussian distribution. Say we have a vocabulary of size v, where each word has dimensionality d (here $d = 52\times$ number of characters in the word), and we denote $\mathbf{y}_1$, $\mathbf{y}_2$, ..., $\mathbf{y}_v$ as the vector representations of all the words in the vocabulary. We can represent the current posterior over words at time step t by a $(v-1)$-dimensional log-odds vector $\mathbf{x}^{(t)}$, where each element $\mathbf{x}_i^{(t)}$ represents the log-odds of $\mathbf{y}_i$ relative to the final word $\mathbf{y}_v$. Working with beliefs in this format means that Bayesian inference is just additive in log-odds (no renormalization):

$$
\begin{aligned}
\mathbf{x}_i^{(t)} &= \log \frac{p(w_i|\mathcal{I}^{(0,\ldots,t)})}{p(w_v|\mathcal{I}^{(0,\ldots,t)})} \\
&= \log \frac{p(\mathcal{I}^{(t)}|w_i)p(w_i|\mathcal{I}^{(0,\ldots,t-1)})}{p(\mathcal{I}^{(t)}|w_v)p(w_v|\mathcal{I}^{(0,\ldots,t-1)})} \\
&= \log \frac{p(\mathcal{I}^{(t)}|w_i)}{p(\mathcal{I}^{(t)}|w_v)} + \log \frac{p(w_i|\mathcal{I}^{(0,\ldots,t-1)})}{p(w_v|\mathcal{I}^{(0,\ldots,t-1)})} \\
&= \Delta\mathbf{x}_i^{(t)} + \mathbf{x}_i^{(t-1)}
\end{aligned} \tag{3}
$$

That is, the log-odds posterior at time step t equals the log-odds posterior at time step $t-1$ (which serves as the prior at time step t) plus the log-odds likelihood. Thus, in an iterative belief-updating context, the log-odds vector begins at a value set by the prior, here the language model, $\mathbf{x}_i^{(0)} = \log p(w_i) - \log p(w_v)$. Then, as each piece of visual information $\mathcal{I}^{(t)}$ arrives, updating beliefs is as simple as adding to $\mathbf{x}^{(t-1)}$ the likelihood log-odds vector for this new piece of information $\Delta\mathbf{x}^{(t)}$, where each element $\Delta\mathbf{x}_i^{(t)}$ gives the likelihood log-odds for that word relative to the final word w_v. For a given true word, vocabulary, and eccentricity, the density function for the likelihood log-odds vector $\Delta\mathbf{x}^{(t)}$ is a $(v-1)$-dimensional multivariate normal distribution, as each element $\Delta\mathbf{x}_i$ is an affine transformation of $\mathcal{I}$, which is itself a multivariate Gaussian.

Experiment

To test whether readers display signatures of optimal integration across these contexts, we build a computational implementation of an ideal-integration model predicting identification confidence for each skipping decision. We show that these model predictions explain significant variance in human skipping rates when added to a strong baseline model.

Baseline model

Data The English part of the Dundee corpus contains eye movement records from 10 native English-speaking participants as they read through newspaper editorials (see Kennedy & Pynte, 2005, for further details.) We included 122,230 observations from the Dundee corpus if they were: 1) a word skipped on first pass (coded as a 1) or a word fixated on first pass (coded as a 0); 2) not adjacent to any blink; and 3) not the first or last fixation on a line. Further, the fixated/skipped word should not 1) contain any non-alphabetical character or be adjacent to punctuation, or 2) follow a word that was skipped or refixated. We excluded observations with far launch sites and long word lengths to ensure enough observations on every level of variations. In the final data, launch sites ranged between [-10, -1], with more than 1000 observations from each launch site, and word length ranged between [1, 8], with the skipping rate being higher than 9% for each word length. The overall skipping rate was 53.9%, resulting from the generally high skipping rate of Dundee corpus, which was over 40% (Demberg & Keller, 2008), and our criterion of requiring the previous word to be fixated, leading to a skipping rate even higher.

Table 1: Generalized additive mixed-effects regression model results of baseline model (note that random slopes for these fixed effects were not included in the model; the model included a random intercept over participants). The GAMM was fitted by REML, and p-values were reported using *summary.gam* function in *mgcv* package (Wood, 2011).

	χ^2	p-value
word length	6026.25	$< 2 \times 10^{-16***}$
launch site	9123.73	$< 2 \times 10^{-16***}$
frequency	527.94	$< 2 \times 10^{-16***}$
surprisal (5-gram)	38.40	$1.01 \times 10^{-6***}$
context entropy	71.16	$8.28 \times 10^{-11***}$
word length $\times$ frequency	89.06	$7.73 \times 10^{-16***}$
launch $\times$ frequency	36.09	$2.85 \times 10^{-5***}$
launch $\times$ surprisal	29.39	$1.13 \times 10^{-4***}$
launch $\times$ entropy	66.82	$2.24 \times 10^{-11***}$
word length (word N-1)	828.66	$< 2 \times 10^{-16***}$
frequency (word N-1)	54.11	$1.62 \times 10^{-9***}$
5-gram (word N-1)	127.22	$< 2 \times 10^{-16***}$
context entropy (word N-1)	31.68	$5.05 \times 10^{-5***}$
word length $\times$ frequency (word N-1)	84.69	$1.73 \times 10^{-14***}$

Model We analyzed first-pass skipping in the Dundee corpus with a generalized additive mixed-effects regression model (GAMM) predicting skipping from a wide range of variables previously shown to influence skipping, including word length, launch site, word frequency, surprisal, and contextual constraint. We estimated word frequency (log unigram probability) and 5-gram surprisal (log 5-gram probability) with n-gram models (Goodkind & Bicknell, 2018) trained on Google One Billion Word Benchmark (Chelba et al., 2013), and we measured contextual constraint as the entropy of the 5-gram probability distribution of words in a vocabulary of 20,001 words. We defined the vocabulary to include all words that were in both the Dundee corpus and our language modeling corpus, plus words with frequencies above a cutoff chosen such that the resulting total vocabulary would have about 20,000 words. We also controlled for the previous word's properties such as word length and frequency, and included a random intercept over participants. Crucially, this GAMM allowed for non-linear effects of each of these variables, providing a strong baseline. Table 1 shows all the fixed effects in the baseline model.

Rational model

Simulation For each observation in the dataset, we simulated 50 trials using the rational model of skipping for each parametrization of the model. In each trial, a piece of visual information from the launch site is sampled and combined with the linguistic information to generate a posterior distribution of possible identities of the word. As described above, the visual information in this model has two param-

Table 2: Significance of averaged entropy of a rational model's posterior distribution when added to the baseline model.

	Prior: Frequency		Prior: 5-gram	
(σ, Λ)	z-value	p-value	z-value	p-value
(1,5)	-2.99	2.78×10^{-3}**	1.23	0.22
(1,15)	-2.51	0.012*	1.43	0.15
(1,30)	-2.07	0.039*	2.27	0.024*
(2,5)	-4.49	7.26×10^{-6}***	1.15	0.25
(2,15)	-4.22	2.4×10^{-5}***	1.67	0.095·
(2,30)	-2.75	6.02×10^{-3}**	1.96	0.05·
(3,5)	-5.76	8.32×10^{-9}***	1.23	0.22
(3,15)	-4.92	8.75×10^{-7}***	1.56	0.12
(3,30)	-3.88	1.03×10^{-4}***	1.04	0.30
(4,5)	-5.98	2.27×10^{-9}***	1.16	0.25
(4,15)	-4.22	2.50×10^{-5}***	2.15	0.032*
(4,30)	-4.04	5.36×10^{-5}***	1.43	0.15
(5,5)	-5.58	2.37×10^{-8}***	1.14	0.26
(5,15)	-4.81	1.55×10^{-6}***	1.78	0.076·
(5,30)	-3.01	2.65×10^{-3}**	2.28	0.023*

eters: overall visual input quality Λ and the width of acuity function σ. We used fifteen sets of parameter pairs for the models; these parameters were chosen to be values that spanned a wide part of the parameter space while also respecting the trade-off between width of the acuity function and its overall quality.[1] The linguistic information (prior) in this model is given by either the word frequency (unigram) or 5-gram language models, as used in our baseline model.

Analysis From each trial, we extract the entropy of the posterior distribution (postH) and then calculate the average of postH from the 50 trials for each observation (for each model parametrization). For each parametrization, we add this average postH to our baseline model as a linear predictor. If human readers extract visual and linguistic information in a rational manner, we predict postH to show a significant effect predicting human skipping, even in a strong baseline model, such that skipping is more likely when the posterior entropy is low.

Results

Baseline model

GAMM results of the baseline model are summarized in Table 1. The results confirm previous findings that word length, launch site, frequency, surprisal, and contextual constraint significantly influenced human skipping. Moreover, this baseline model captures non-linear interactions among these predictors, indicating that different sources of information interactively guide skipping at an aggregated level.

[1]If the function is very wide and high quality, the model has too much information about the whole word, whereas if narrow and low quality, the model has almost no information.

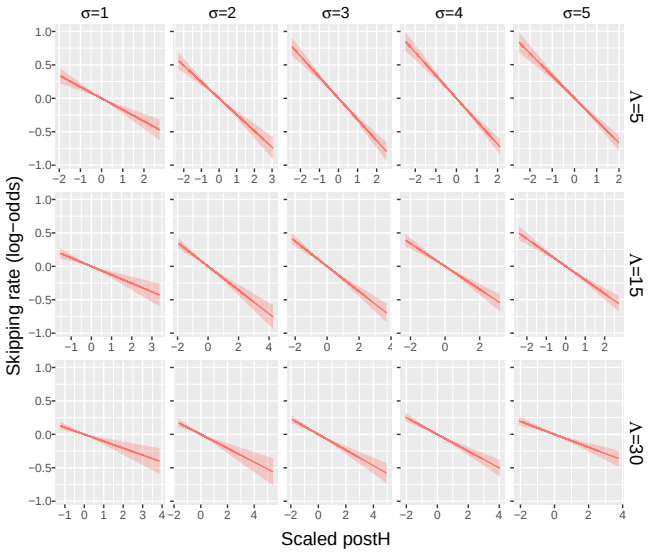


Figure 1: Partial effect of postH with a frequency prior in predicting skipping rate.

Rational model

The partial effects of postH computed from the GAMMs are visualized in Figure 1 (frequency prior) and Figure 2 (5-gram prior), after controlling for all variables in the baseline model and additionally a random slope of postH over participants. The significance of postH when added to the baseline model is reported in Table 2. For postH from rational models with a frequency prior, the effects are significant in the predicted direction: high postH indicates high uncertainty about the word's identification and is associated with lower skipping rates; these effects are robust to parameter choice and are significant for all parametrizations tested. For postH from rational models with a 5-gram prior, the effects are generally not significant, though they do all trend in the same direction and show the pattern that skipping rates increase as the uncertainty over the word's identity increases, opposite to the predicted direction.

Discussion

In this paper, we implemented a computational model of skipping that used Bayesian inference to combine visual and linguistic information. We then extracted the entropy of the posterior distribution as a measure of readers' confidence about word identification, and tested whether this measure improved the predictive power of a strong baseline model incorporating aggregate visual and linguistic factors known to influence skipping. Results showed that this postH measure had significant additional effect predicting skipping when extracted from rational models with a frequency prior, but generally not when extracted from rational models with a 5-gram prior. The direction of the effect of postH from models with a frequency prior is consistent with the prediction that low confidence about word identification leads to decreased skipping

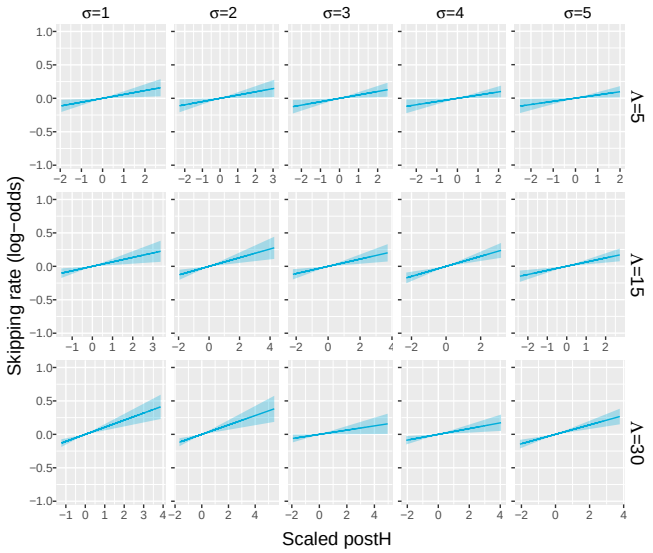


Figure 2: Partial effect of postH with a 5-gram prior in predicting skipping rate.

rate, while the trend of the effect of postH from models with a 5-gram prior is in an opposite direction.

These findings generally provide positive evidence for the rational model's prediction that readers' likelihood of skipping vary depending on the *particular* visual information obtained, and whether that information distinguishes it strongly from its (likely) visual neighbors as in linguistic knowledge. The predictor, postH, is computed from the posterior distribution of a Bayesian inference model with partial visual information about the word, and therefore captures how likely the word is differentiated from its neighbors in the vocabulary. If the true word is much more likely than its visually-similar neighbors, the postH should be low, while if the true word and its neighbors have similar probabilities, the postH should be high. Such a measure of reader's confidence about word identification is dynamic, innate, and hard to capture in factorial experiments, but can be approached through computational simulation. Its significant effect is not predicted by the heuristic model in principle, as postH is assumed to utilize information about how particular words relate to their neighbors regarding the specific visual information obtained about parts of the word.

The observation that postH from a frequency prior better predicts skipping than the 5-gram prior is potentially problematic for a fully rational model of skipping, though: a reader that maximize usage of all the information available should be better predicted by a model with 5-gram prior than one with frequency prior. Rather, this pattern lines up with previous findings on the skipping of *the*, which relies on visual and frequency information more than structural information (Angele et al., 2014). This pattern is also consistent with the finding that frequency effect but not predictability effect on skipping survives bad parafoveal visual input, which may be explained by different time course of frequency and contextual information in making eye movement decisions (Staub & Goddard, 2019). A possible reason of our finding is that skipping decisions may be made without full knowledge of the context, leading to the absence of effect from our measure (i.e. 5-gram) of contextual information. Specifically, since saccade programming takes a relatively long time and identification/processing of the fixated word continues during this lag, skipping decisions may be made before the previous word is fully identified and integrated into the context. In spite of this issue to be further examined, we find that the entropy of a posterior distribution from a frequency prior improves prediction of skipping with average variables, suggesting a complex combination of information sources as predicted by rational models of reading.

Acknowledgments

This research was supported by the National Science Foundation under NSF 1734217.

References

Angele, B., Laishley, A. E., Rayner, K., & Liversedge, S. P. (2014). The effect of high-and low-frequency previews and sentential fit on word skipping during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(4), 1181.

Angele, B., & Rayner, K. (2013). Processing the in the parafovea: Are articles skipped automatically? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(2), 649.

Balota, D. A., Pollatsek, A., & Rayner, K. (1985). The interaction of contextual constraints and parafoveal visual information in reading. *Cognitive psychology*, *17*(3), 364–390.

Bicknell, K., & Levy, R. (2010). A rational model of eye movement control in reading. In *Proceedings of the 48th annual meeting of the Association for Computational Linguistics* (pp. 1168–1178).

Bicknell, K., & Levy, R. (2012). Word predictability and frequency effects in a rational model of reading. In *Proceedings of the 34th annual conference of the Cognitive Science Society* (pp. 126–131).

Brysbaert, M., Drieghe, D., & Vitu, F. (2005). Word skipping: Implications for theories of eye movement control in reading. *Cognitive processes in eye guidance*, 53–77.

Chelba, C., Mikolov, T., Schuster, M., Ge, Q., Brants, T., Koehn, P., & Robinson, T. (2013). One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005*.

Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, *109*(2), 193–210.

Drieghe, D., Rayner, K., & Pollatsek, A. (2005). Eye movements and word skipping during reading revisited. *Journal of Experimental Psychology: Human Perception and Performance*, *31*(5), 954.

Duan, Y., & Bicknell, K. (2017). Refixations gather new visual information rationally. In *Proceedings of the 39th annual conference of the Cognitive Science Society* (pp. 301–306).

Engbert, R., Nuthmann, A., Richter, E. M., & Kliegl, R. (2005). SWIFT: a dynamical model of saccade generation during reading. *Psychological Review*, *112*(4), 777–813.

Goodkind, A., & Bicknell, K. (2018). Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th workshop on cognitive modeling and computational linguistics (cmcl 2018)* (pp. 10–18).

Hyönä, J., Niemi, P., & Underwood, G. (1989). Reading long words embedded in sentences: Informativeness of word halves affects eye movements. *Journal of Experimental Psychology: Human Perception and Performance*, *15*(1), 142.

Kennedy, A., & Pynte, J. (2005). Parafoveal-on-foveal effects in normal reading. *Vision Research*, *45*(2), 153–168.

Kliegl, R., Grabner, E., Rolfs, M., & Engbert, R. (2004). Length, frequency, and predictability effects of words on eye movements in reading. *European Journal of Cognitive Psychology*, *16*(1-2), 262–284.

Legge, G. E., Hooven, T. A., Klitz, T. S., Mansfield, J. S., & Tjan, B. S. (2002). Mr. chips 2002: New insights from an ideal-observer model of reading. *Vision Research*, *42*(18), 2219–2234.

Legge, G. E., Klitz, T. S., & Tjan, B. S. (1997). Mr. Chips: an ideal-observer model of reading. *Psychological Review*, *104*(3), 524–553.

Luke, S. G., & Christianson, K. (2016). Limits on lexical prediction during reading. *Cognitive psychology*, *88*, 22.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*(3), 372–422.

Rayner, K., Sereno, S. C., & Raney, G. E. (1996). Eye movement control in reading: a comparison of two types of models. *Journal of Experimental Psychology: Human Perception and Performance*, *22*(5), 1188–1200.

Rayner, K., Slattery, T. J., Drieghe, D., & Liversedge, S. P. (2011). Eye movements and word skipping during reading: effects of word length and predictability. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(2), 514.

Reichle, E. D., Warren, T., & McConnell, K. (2009). Using E-Z Reader to model the effects of higher level language processing on eye movements during reading. *Psychonomic Bulletin & Review*, *16*(1), 1–21.

Schotter, E. R., Lee, M., Reiderman, M., & Rayner, K. (2015). The effect of contextual constraint on parafoveal processing in reading. *Journal of memory and language*, *83*, 118–139.

Staub, A., & Goddard, K. (2019). The role of preview validity in predictability and frequency effects on eye movements in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(1), 110.

White, S. J., Rayner, K., & Liversedge, S. P. (2005). The influence of parafoveal word length and contextual constraint on fixation durations and word skipping in reading. *Psychonomic bulletin & review*, *12*(3), 466–471.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, *73*(1), 3-36.

If it's important, then I am curious: A value intervention to induce curiosity

Rachit Dubey (rdubey@princeton.edu)
Department of Computer Science, Princeton University

Thomas L. Griffiths (tomg@princeton.edu))
Department of Psychology and Computer Science, Princeton University

Tania Lombrozo (lombrozo@princeton.edu)
Department of Psychology, Princeton University

Abstract

Curiosity is considered essential for learning and sustained engagement, yet stimulating curiosity in educational contexts remains a challenge. Can people's curiosity about a topic be stimulated by evidence that the topic has potential value? In two experiments we show that increasing people's perceptions about the usefulness of a scientific topic also influences their curiosity and subsequent information search. Our results also show that simply presenting interesting facts is not enough to influence curiosity, and that people are more likely to be curious about a topic if they perceive it to be directly valuable to them. Given the link between curiosity and learning, these results have important implications for science communication and education more broadly.

Keywords: curiosity; intervention; education

Introduction

> *"Sometimes these dollars go to projects that have little or nothing to do with the public good. Things like fruit fly research. I kid you not."*
>
> – Sarah Palin, former Alaska Governor

In one of her first policy speeches, former Alaska Governor and Vice Presidential nominee Sarah Palin made the above remark to alert people to the alleged misuse of federal funds. Her comments were met with disappointment and dismay within the scientific community, and for good reason – fruit flies have been an essential part of biological research, and research on them has shed light on basic aspects of biology and prompted medical advance (Siegel, 2009).

Unfortunately, Palin's attack on fruit flies is not the first example of a politician deriding specific kinds of research (Kempner, 2008). While such statements could be motivated by various political or economic considerations, they also seem to reflect a lack of curiosity about the scientific topics in question. This highlights an important aspect of curiosity in that the *same* topic can elicit quite *different* levels of curiosity in different people. What accounts for this difference, and how might greater curiosity be induced?

Psychological accounts of curiosity posit that curiosity is piqued whenever people observe discrepancies (Berlyne, 1950, 1960), or perceive a "moderate" gap between their actual and desired knowledge state (Loewenstein, 1994). Based on these theories, many people's low curiosity for scientific topics (such as fruit flies) could be explained by a lack of (perceived) discrepancy and/or by inadequate prior knowledge, such that the information gap is too large. However, Palin's comments suggest an additional possibility: perhaps people simply fail to see any *value* in pursuing topics that seem to lack theoretical or practical implications. Indeed, a recent account of curiosity suggests that people's curiosity should be higher for information if they perceive that information to be important to them (Dubey & Griffiths, 2017). Furthermore, various studies from the education literature have shown that students' perceived utility value i.e., how valuable they think a task would be for future goals, correlates with their task enjoyment and engagement (Eccles & Wigfield, 1995; Hulleman et al., 2008). In line with these findings, education researchers have successfully used utility-value interventions to increase student's motivation and performance in various learning settings (Hulleman et al., 2010; Harackiewicz et al., 2012; Brown et al., 2015). However, this work has not investigated whether utility-value interventions can successfully induce *curiosity*.

The current work explores a novel way to stimulate curiosity – by manipulating the perceived value of a topic. More specifically, we explore whether changing the perceived value of a scientific topic can also affect people's curiosity about that topic. If such a value manipulation indeed affects curiosity, then interventions on value could not only have important implications for curiosity researchers, but also for science communicators and educators of all kinds.

The importance of perceived value

Motivation and value

A classic model of motivation is the expectancy-value theory (Atkinson, 1964; Wigfield & Eccles, 2000), which posits that motivation in educational contexts is determined by an individual's expected success (i.e., belief that one can succeed at an activity) and subjective task value. Studies based on this theory have primarily developed interventions that focus on the 'expected success' component – that is, on improving students' perceived ability to master tasks to improve their motivation and performance (Eccles & Wigfield, 1995; Wigfield & Eccles, 1994). More recently, a number of researchers have also developed interventions that focus on the 'subjective value' component. These interventions show that an increase in students' perception about the usefulness of a subject leads to enhanced motivation and improved performance in various learning settings (Hulleman et al., 2010;

Harackiewicz et al., 2012, 2014; Brown et al., 2015). Although curiosity is usually considered distinct from motivation, and possibly involves different computational and neural mechanisms, these findings provide a useful starting point for developing interventions on curiosity, and for considering why perceived value might play a role.

Curiosity and value

Although curiosity has long been recognized as an important aspect of cognition, there is no single, agreed-upon theory of curiosity (Kidd & Hayden, 2015). Instead, a number of theories have been proposed to explain curiosity (Berlyne, 1950, 1960; Schmidhuber, 1991; Loewenstein, 1994; Oudeyer et al., 2007). These theories link curiosity with various psychological factors, but none of them explicitly consider the potential role that value can play in influencing curiosity.

Dubey and Griffiths (2017) recently proposed an account that links curiosity to 'value of knowledge', which is a function of people's current understanding of a topic and the perceived usefulness of that topic. According to the theory, people's curiosity is evoked whenever they perceive an opportunity to increase the value of their knowledge (i.e., topics that either increase understanding or perceived usefulness). In essence, this model can be interpreted as providing a quantitative articulation of the expectancy-value theory. If people's curiosity is indeed driven by the perceived opportunity to increase the value of their knowledge, then this suggests that curiosity can be driven towards topics that seem initially unimportant if people come to perceive them as useful or otherwise valuable.[1]

Overview of experiments

In the current paper we ask whether manipulating perceived value can influence curiosity. Answering this question provides an opportunity to empirically evaluate theoretical claims related to the link between value and curiosity while also extending the rich literature in educational psychology on motivation.

To address this question, we report two experiments in which we present scientific topics to participants and have them indicate their curiosity about those topics. We then manipulate the perceived usefulness of those topics and record participants' change in curiosity. In Experiment 1, we manipulate how 'valuable' it would be for medical research to study fruit flies and rats, and we measure how participants' curiosity and information search is affected by this manipulation. In Experiment 2, we go one step further by considering what *kind* of value most effectively drives curiosity.

Experiment 1: Does value influence curiosity?

In Experiment 1, we investigated whether people's curiosity towards a scientific topic can be influenced by manipulating the perceived value of that topic, and whether this boost in curiosity affects subsequent information search. Participants read two short articles about two different scientific topics (one article for each topic). One of the two articles was 'high-value', and the other was 'low-value'. Participants' curiosity for the two scientific topics was recorded before and after they read the articles. Subsequently, participants had the choice to read some facts about the two scientific topics.

The experiment tested the following predictions: (1) Reading a high-value article will increase curiosity, and it will do so to a greater extent than reading a low-value article, (2) Participants will be more likely to read facts corresponding to the topic of the high-value article than those corresponding to the topic of the low-value article, and (3) The effect of the value manipulation on curiosity will be mediated by perceived value.

Participants

We recruited 240 participants from Amazon Mechanical Turk. They earned $1.00 for participating in a study that took approximately 7-8 minutes to complete.

Note that for both Experiments 1 and 2, sample sizes were determined prior to data collection; based on pilot data, we aimed to recruit at least 60 participants per condition (which required 240 in experiment 1, given two conditions with counterbalanced order).

Stimuli

The stimuli used in the experiment were two short articles describing the biology of fruit flies and two short articles describing the biology of rats. For each of the two topics (i.e., fruit fly and rat), one article was a 'high-value' article and the other was a 'low-value' article. The high-value article emphasized how research about that animal could be highly beneficial to medicine, while the low-value article raised questions about whether research concerning that animal could generate any medical benefits for humans. All four articles were otherwise matched in terms of length and, as much as possible, for general content and style (stimuli available at - `https://goo.gl/BNpHzU`).

Procedure

At the start of the experiment, participants were randomly assigned to one of two conditions. In condition 1, participants were assigned to the high-value article for fruit flies and to the low-value article for rats. In condition 2, participants were assigned to the low-value article for fruit flies and to the high-value article for rats.

Phase 1 At the beginning of the first phase, participants were presented with one of the two scientific topics, either 'biology of fruit flies' or 'biology of rats' (counter-balanced). After seeing the topic, participants were asked to respond to each of the following on a scale from 1-7:

1. *Usefulness:* "To what extent would knowing about this phenomenon be useful to you in the future?"

[1] Additionally, we note that although Loewenstein's theory of curiosity (Loewenstein, 1994) does not explictly consider value in its formal account, it does hypothesize that people will be more curious about topics that are important to them.

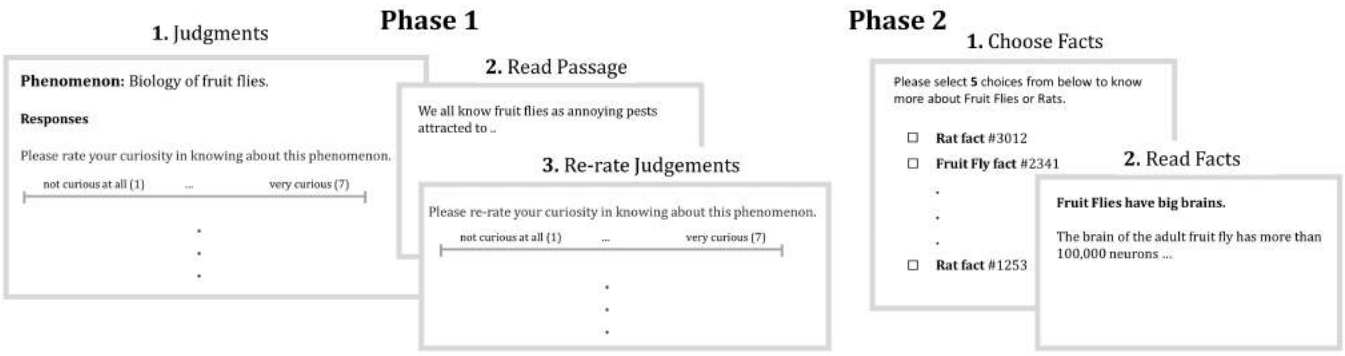


Figure 1: **Design of Experiment 1.** The experiment was divided into two phases. In Phase 1, participants were first presented with one of the two topics from our stimuli and asked to provide ratings for curiosity, understanding, and usefulness. They then read an article about that topic and once again rated curiosity, understanding, and usefulness. This procedure was then repeated for the second topic. In Phase 2, participants had the choice to reveal five out of eight facts presented to them (four facts from each topic). The chosen facts were then presented one by one. Note that instructions were provided before each phase

2. *Understanding:* "Please rate how well you feel you understand this phenomenon."

3. *Curiosity:* "Please rate your curiosity in knowing about this phenomenon."

The first question was to ensure that our manipulation of value was successful, and the second question was to ensure that the effect of value on curiosity couldn't be reduced to understanding. The third question i.e. participants' rating of their curiosity was the key variable of interest in Phase 1. After providing the ratings, participants were presented with the assigned article for that topic and they were instructed to read it as carefully as possible. After they finished reading the article, participants were asked to re-rate their understanding, perceived usefulness, and curiosity about that topic. Following this, the above procedure was repeated for the other topic (also refer to Figure 1).

Phase 2 In the second phase, participants were instructed that they would be presented with some facts about the two topics (four for each topic, eight in total), but that they only needed to read five of those facts. The eight fact choices were then presented (e.g., "Rat Fact 3201"), and participants indicated their five choices. The corresponding facts were shown to participants after they indicated their choices.

Results

For all analyses that follow, we compare participants' ratings for the low-value stimuli relative to the high-value stimuli across the two conditions.

Phase 1 We first investigated the change in participants' understanding ratings after reading the low- and high- stimuli. As shown in Figure 2(a), the mean understanding rating increased by 0.63 for the low-value stimuli and by 1.16 for the high-value stimuli. A mixed ANOVA revealed a significant interaction between time (pre and post ratings) and stimulus (low-value or high-value) on understanding, $F(1,239) = 29.1, MSE = 16.8, p < 0.001$. We next confirmed that our manipulation of value successfully manipulated perceived usefulness. As shown in Figure 2(b), the mean rating of value increased by 0.40 for the low-value stimuli and by 1.16 for the high-value stimuli. A mixed ANOVA again revealed a significant interaction between time (pre and post ratings) and item (low-value or high-value) on perceived value, $F(1,239) = 47.692, MSE = 35.3, p < 0.001$, indicating that our manipulation of value was effective. Finally, we tested whether our value manipulation influenced participants' curiosity. As shown in Figure 3(a), the mean curiosity rating increased by 0.44 for low-value stimuli and by 1.04 for high-value stimuli i.e. the increase for the high-value stimuli was 0.60 points higher than the increase for the low-value stimuli. A mixed ANOVA revealed a significant interaction between time (pre and post ratings) and item (low-value or high-value) on curiosity ,$F(1,239) = 32.69, MSE = 21.6, p < 0.001$, indicating that the manipulation of value had a significant effect on curiosity. A follow-up paired-samples t-test showed that the increase of curiosity was greater for the 'high-value' stimuli compared to the 'low-value' stimuli, $t(478) = -4.71, p < 0.001$.

We next considered whether understanding or perceived value mediated the effect of our value manipulation on curiosity. We first ran a linear regression to predict curiosity based on value manipulation (i.e. 'low-value' or 'high-value'); this yielded a significant and positive coefficient of 0.60, $t = 4.7, p < 0.001, 95\%$ CI$[0.35, 0.85]$. We then considered a regression predicting curiosity based on perceived value; yielding a significant and positive coefficient of 0.47, $t = 12.1, p < 0.001, 95\%$ CI$[0.39, 0.54]$. We also considered a regression predicting curiosity based on understanding; again yielding a significant and positive coefficient of 0.37, $t = 8.7, p < 0.001, 95\%$ CI$[0.28, 0.45]$. We then fit a multiple regression with both value manipulation and perceived value as predictors; this yielded coefficients of 0.26 and 0.44 respectively ($t = 2.21, p < 0.05, 95\%$ CI$[0.03, 0.49]$ and $t = 11.2, p < 0.001, 95\%$ CI$[0.37, 0.52]$),

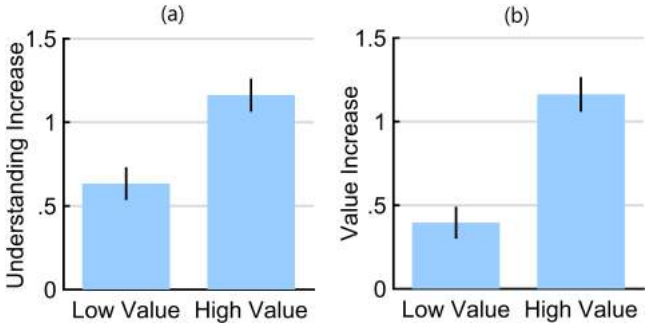


Figure 2: **Effect of value manipulation on understanding and perceived value (Experiment 1).** (a) Change in understanding ratings for participants who received the low-value and high-value stimuli before and after they read the corresponding articles. (b) Change in value ratings for participants before and after they read the corresponding articles.

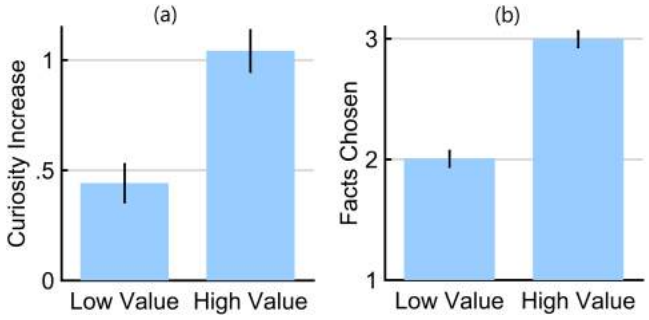


Figure 3: **Value influences curiosity (Experiment 1).** (a) Mean increase in participants' curiosity about a topic after reading a 'low-' or 'high-value' article. (b) Mean facts chosen after reading a 'low-' or 'high-value' article about that topic.

suggesting partial mediation. Finally, we fit a multiple regression with both value manipulation and understanding as predictors which yielded coefficients of 0.42 and 0.34 respectively ($t = 3.46, p < 0.001, 95\%$ CI$[0.18, 0.66]$ and $t = 7.98, p < 0.001, 95\%$ CI$[0.25, 0.42]$), again suggesting partial mediation.

Phase 2 We next investigated whether participants were more likely to reveal facts about the high-value stimuli compared to the low-value stimuli. As shown in Figure 3(b), participants indeed revealed more facts about the high-value stimuli (3 vs. 2). A paired-samples t-test found that this difference was significant, $t(478) = -10.6, p < 0.001$.

Discussion

Experiment 1 tested and found support for two of our three predictions about the effects of value on curiosity. First, results from phase 1 showed that participants became more curious about stimuli after reading information that suggested the topic was of high (vs. low) value. Second, results from phase 2 demonstrated that participants were more likely to reveal additional information about a topic after reading information suggesting it was of high value. We also found that our stimuli successfully manipulated perceived value, and that perceived value partially mediated the effect of our value manipulation on curiosity. However, the effect of our value manipulation on curiosity was also partially mediated by understanding. This raises the concern that perceived value is confounded with understanding, and that changes in understanding drove the effects of our manipulation on curiosity. We address this concern in Experiment 2.

Experiment 2: What influences value most?

Experiment 2 had two aims. First, the experiment aimed to test the influence of perceived value on curiosity while controlling for understanding. Second, the experiment aimed to investigate the effect of different kinds of information on people's perceived value and subsequently on curiosity. Participants were randomly assigned to three conditions in which they read a short article about the 'biology of fruit flies' and provided ratings before and after they read the article. In condition 1, the article presented interesting facts about *fruit fly reproduction*. In condition 2, the article showed how fruit flies are *valuable to the environment*. In condition 3, the article provided evidence that fruit flies are *valuable to medical research*. We hypothesized that participants' increase in understanding would be similar across the three conditions, but that perceived value would not be. Moreover, the contrast between conditions 2 and 3 allows us to test the hypothesis that perceived value would be especially sensitive to value with potential personal relevance.

More specifically, the experiment tested these predictions: (1) Participants' curiosity about fruit flies will increase most strongly in condition 3 (compared to conditions 1 and 2), and (2) the effect of perceived value on curiosity will not be reducible to other factors, such as understanding or surprise.

Participants

We recruited 203 participants from Amazon Mechanical Turk ($n = 67, 72$, and 64 for condition 1, 2, and 3 respectively). They earned \$0.35 for participating in a study that took approximately 2-3 minutes to complete.

Stimuli

The stimuli used in the experiment were three short articles describing the biology of fruit flies. The three articles varied in terms of their value to humans – the first article simply presented interesting facts about the reproductive cycle of fruit flies, the second article had facts about the importance of fruit flies for the ecosystem, and the third article provided facts about the importance of fruit flies for medical research. All three articles were matched for length and as much as possible for general content and style (stimuli available at - `https://goo.gl/BNpHzU`).

Procedure

At the start of the experiment, participants were randomly assigned to one of the three conditions. The three conditions followed the same procedure and differed only with respect to

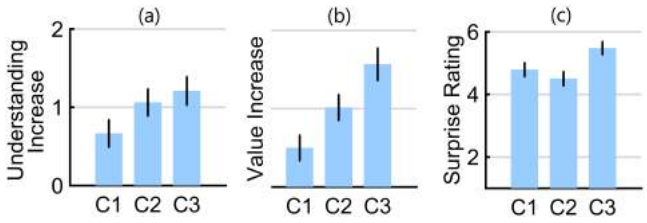


Figure 4: **Effect of value manipulation on understanding, value, and surprise (Experiment 2).** (a) Mean change in understanding ratings. (b) Mean change in value ratings. (c) Mean surprise ratings of participants in each condition.

which of the three articles the participants read. Participants were first presented with the scientific topic, 'biology of fruit flies', and were asked to rate their understanding, perceived value, and curiosity as in Experiment 1. After providing these ratings, participants were presented with the assigned article and they were instructed to read it as carefully as possible. After they finished reading the article, participants were asked to re-rate understanding, perceived usefulness, and curiosity about that topic. In addition to these ratings, after the participants read the article, they were also asked to respond to the following on a scale of 1-7 – "Please rate how surprising you found the previously shown information on fruit flies to be." This question on surprise was added as an additional control to ensure that any potential increase in curiosity was not caused simply by surprise.

Results

We first investigated how participants' understanding changed after they read the corresponding articles across the three conditions. As shown in Figure 4(a), participants' understanding ratings increased significantly after they read the article for all three conditions, $t(142) = -2.34, p < 0.05$ for condition 1, $t(126) = -3.57, p < 0.001$ for condition 2, and $t(132) = -3.88, p < 0.001$ for condition 3. Moreover, a one-way ANOVA revealed that these three groups were not significantly different from each other, $F(2,200) = 2.64, MSE = 5.5, p = 0.07$, indicating that understanding ratings increased the same across all three conditions. We next evaluated how participants' perceived value changed across the three conditions. Although participants' value ratings increased numerically for all three conditions (refer to Figure 4(b)), this increase was not significant for condition 1, $t(142) = -1.52, p = 0.13$. This suggests that simply presenting interesting facts about a topic was not enough to influence perceived value. Furthermore, a one-way ANOVA showed that the three groups differed significantly from each other, $F(2,200) = 9.25, MSE = 19.8, p < 0.001$, with condition 3 significantly higher than condition 2, $t(129) = 2.1, p < 0.05$, and condition 2 significantly higher than condition 1, $t(134) = 2.2, p < 0.05$. We next analyzed how much surprise each article evoked (refer to Figure 4(c)) and found that there was a significant difference for the surprise ratings across the three conditions, $F(2,200) = 5.12, MSE = 16.7, p < 0.01$. Specifically, condition 3 was significantly

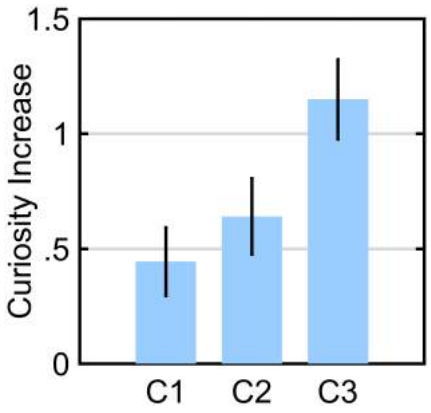


Figure 5: **People's curiosity is highest when they perceive something to be of direct value to them (Experiment 2).** Mean change in curiosity ratings for the three different conditions. Participants' curiosity increased the most in condition 3, in which they read an article that provided evidence that fruit flies are highly beneficial to medicine.

different than condition 2, $t(129) = 3.19, p < 0.05$, but condition 2 was not significantly different compared to condition 1, $t(134) = 0.92, p = 0.36$.

We next evaluated the change in participants' curiosity ratings and found that as per our hypothesis, curiosity ratings increased the most in condition 3 (by 1.15 points, also refer to Figure 5). Furthermore, similar to perceived value ratings, although participants' curiosity ratings increased for all three conditions, that increase was not significant for condition 1, $t(142) = 1.45, p = 0.15$. We also conducted a one-way ANOVA analysis and found that the three groups were significantly different from each other, $F(2,200) = 5.14, MSE = 9.1, p < 0.01$. Follow-up paired-samples t-tests showed that condition 3 was significantly different than condition 2 ,$t(129) = 2.13, p < 0.05$, but condition 2 was not significantly different compared to condition 1, $t(134) = 0.89, p = 0.38$. These results suggest that if people perceive stimuli to be less valuable to them, then they are less likely to become curious about them.

As in Experiment 1, we tested whether the effect of our value manipulation on curiosity was mediated by perceived value. First, a linear regression predicting curiosity from value manipulation (i.e. condition 1, condition 2, or condition 3) revealed a significant positive coefficient of 0.35,

Source	Effect Size	t	p-value	95% CI
condition 1	-0.30	-1.38	0.17	$[-0.73, 0.13]$
condition 2	-0.22	-1.01	0.31	$[-0.65, 0.21]$
understanding*	0.13	2.12	< 0.05	$[0.01, 0.25]$
value*	0.21	3.35	< 0.001	$[0.09, 0.34]$
surprise*	0.15	3.10	< 0.01	$[0.06, 0.25]$

Table 1: **Regression results (Experiment 2).** Regression results of the increase of curiosity ratings with condition 1, condition 2, understanding ratings increase, value ratings increase, and surprise; significant differences are starred.

$t = 3.1, p < 0.005, 95\%$ CI$[0.13, 0.57]$. A similar regression with perceived value as the predictor produced a coefficient of $0.33, t = 5.64, p < 0.001, 95\%$ CI$[0.21, 0.45]$. Next, a multiple regression with both value manipulation and perceived value resulted in a non-significant coefficient of 0.19 for value manipulation, $t = 1.72, p = 0.09, 95\%$ CI$[-0.03, 0.41]$, while perceived value remained significant at 0.3, $t = 4.91, p < 0.001, 95\%$ CI$[0.18, 0.42]$. This suggests that the effect of value manipulation on curiosity was fully mediated by perceived value. Finally, to confirm that the effect of perceived value on curiosity is not reducible to condition, understanding, or surprise, we conducted a linear regression to predict the increase of curiosity ratings with condition 1, condition 2, increase of understanding ratings, increase of value ratings, and surprise. We found a significant regression equation, $F(5, 197) = 10.61, p < 0.001$, with an R^2 of 0.192 and a significant effect of value on curiosity, greater than any of the other factors (refer to Table 1).

Discussion

The findings from Experiment 2 support both of our predictions. First, we found that not all kinds of value are equal: participants were more likely to become curious about a scientific topic if they learned of its direct value to them (condition 3 vs. 2). Second, we succeeded in identifying an effect of value that could not be explained by differences in understanding or surprise. Our results suggest that simply presenting interesting facts that have no direct value is not enough to induce curiosity (condition 1), even if those facts boost understanding and induce surprise.

General Discussion

The primary purpose of this research was to test whether curiosity can be influenced by manipulating people's perceptions of value. Across two experiments, we find that manipulating the perceived value of a topic influenced curiosity (Experiment 1 and 2), and this also influenced subsequent information search (Experiment 1). Results from Experiment 2 further demonstrated that the effects of our manipulation on curiosity were fully mediated by perceived value and cannot be reduced to understanding or surprise, which are both known to influence curiosity.

Our results have considerable theoretical implications as they demonstrate a link between value and curiosity. In doing so, our findings lend support to Dubey and Griffiths's (2017) theory of curiosity. They also challenge previous accounts of curiosity, such as the incongruity theory (Berlyne, 1960) and the information-gap theory (Loewenstein, 1994), insofar as those theories fail to incorporate an *explicit* role for value.

Despite the promise of our results, the significance of our study is limited by the nature of our stimuli, task, and our focus on short-term consequences of value on curiosity. Furthermore, several key theoretical questions about curiosity remain. For example, previous studies have shown that people become curious about completely irrelevant and sometimes even potentially harmful stimuli (Hsee & Ruan, 2016). Conversely, people are sometimes averse to information, even when that information is potentially useful to them (Sweeny et al., 2010). Understanding how curiosity interacts with value in these contexts is an important research question for future work.

Another limitation of our experimental manipulation is that the importance of the information is clearly spelled out to the participants especially in the high-value articles. Therefore, it is possible that the participants rate that information to be important even though they may not necessarily believe that to be the case (perhaps due to a social desirability bias). Future work will consist of conducting further experiments to rule out this possibility.

We also note that some theories stipulate that curiosity is an intrinsic drive and is not instrumental, thereby making our results seem counter-intuitive. On the other hand, even if the experience of curiosity is a drive for knowledge for its own sake, it is still possible that curiosity can be modulated by instrumental factors. Prior work has similarly pointed to the challenge of delineating extrinsic and intrinsic factors in various cases (Kidd & Hayden, 2015). For instance, what is intrinsic for one individual could be extrinsic for another.

Regardless of this debate, our work shows that self-reported curiosity (and information-seeking behavior) can be influenced by value and it sheds light on effective strategies to do so. The results from Experiment 2 suggest that simply presenting information that seems interesting is not effective in influencing value or curiosity (condition 1). Instead, a more effective way to stimulate curiosity is to present information in a way that allows people to directly see its value and relevance (condition 3). Perhaps fruit flies will never be welcome in our homes, but maybe people will become more curious about them – and more welcoming of basic research on them – once they find out how valuable they are to us.

References

Atkinson, J. W. (1964). An introduction to motivation.

Berlyne, D. E. (1950). Novelty and curiosity as determinants of exploratory behaviour. *British Journal of Psychology. General Section*, *41*(1-2), 68–80.

Berlyne, D. E. (1960). Conflict, arousal, and curiosity.

Brown, E. R., Smith, J. L., Thoman, D. B., Allen, J. M., & Muragishi, G. (2015). From bench to bedside: A communal utility value intervention to enhance students biomedical science motivation. *Journal of Educational Psychology*, *107*(4), 1116.

Dubey, R., & Griffiths, T. L. (2017). A rational analysis of curiosity. In *Proceedings of the 39th Annual Conference of the Cognitive Science Society.*

Eccles, J. S., & Wigfield, A. (1995). In the mind of the actor: The structure of adolescents' achievement task values and expectancy-related beliefs. *Personality and Social Psychology Bulletin*, *21*(3), 215–225.

Harackiewicz, J. M., Rozek, C. S., Hulleman, C. S., & Hyde, J. S. (2012). Helping parents to motivate adolescents in mathematics and science: An experimental test of a utility-value intervention. *Psychological Science*, *23*(8), 899–906.

Harackiewicz, J. M., Tibbetts, Y., Canning, E., & Hyde, J. S. (2014). Harnessing values to promote motivation in education. In *Motivational Interventions* (pp. 71–105).

Hsee, C. K., & Ruan, B. (2016). The Pandora effect: The power and peril of curiosity. *Psychological Science*, *27*(5), 659–666.

Hulleman, C. S., Durik, A. M., Schweigert, S. B., & Harackiewicz, J. M. (2008). Task values, achievement goals, and interest: An integrative analysis. *Journal of Educational Psychology*, *100*(2), 398.

Hulleman, C. S., Godes, O., Hendricks, B. L., & Harackiewicz, J. M. (2010). Enhancing interest and performance with a utility value intervention. *Journal of Educational Psychology*, *102*(4), 880.

Kempner, J. (2008). The chilling effect: how do researchers react to controversy? *PLoS Medicine*, *5*(11), e222.

Kidd, C., & Hayden, B. Y. (2015). The psychology and neuroscience of curiosity. *Neuron*, *88*(3), 449–460.

Loewenstein, G. (1994). The psychology of curiosity: A review and reinterpretation. *Psychological Bulletin*, *116*(1), 75.

Oudeyer, P.-Y., Kaplan, F., & Hafner, V. V. (2007). Intrinsic motivation systems for autonomous mental development. *IEEE Transactions on Evolutionary Computation*, *11*(2), 265–286.

Schmidhuber, J. (1991). Curious model-building control systems. In *Proceedings of 1991 IEEE International Joint Conference on Neural Networks* (pp. 1458–1463).

Siegel, V. (2009). I kid you not. *Disease Models and Mechanisms*.

Sweeny, K., Melnyk, D., Miller, W., & Shepperd, J. A. (2010). Information avoidance: Who, what, when, and why. *Review of General Psychology*, *14*(4), 340.

Wigfield, A., & Eccles, J. S. (1994). Children's competence beliefs, achievement values, and general self-esteem: Change across elementary and middle school. *The Journal of Early Adolescence*, *14*(2), 107–138.

Wigfield, A., & Eccles, J. S. (2000). Expectancy–value theory of achievement motivation. *Contemporary Educational Psychology*, *25*(1), 68–81.

A New Probabilistic Explanation of the Modus Ponens–Modus Tollens Asymmetry

Benjamin Eva (Benjamin.Eva@uni-konstanz.de)
Department of Philosophy, Univeristy of Konstanz, 78464 Konstanz (Germany)

Stephan Hartmann (S.Hartmann@lmu.de)
Munich Center for Mathematical Philosophy, LMU Munich, 80539 Munich (Germany)

Henrik Singmann (Henrik.Singmann@warwick.ac.uk)
Department of Psychology, University of Warwick, Coventry, CV4 7AL (UK)

Abstract

A consistent finding in research on conditional reasoning is that individuals are more likely to endorse the valid modus ponens (MP) inference than the equally valid modus tollens (MT) inference. This pattern holds for both abstract task and probabilistic task. The existing explanation for this phenomenon within a Bayesian framework (e.g., Oaksford & Chater, 2008) accounts for this asymmetry by assuming separate probability distributions for both MP and MT. We propose a novel explanation within a computational-level Bayesian account of reasoning according to which "argumentation is learning". We show that the asymmetry must appear for certain prior probability distributions, under the assumption that the conditional inference provides the agent with new information that is integrated into the existing knowledge by minimizing the Kullback-Leibler divergence between the posterior and prior probability distribution. We also show under which conditions we would expect the opposite pattern, an MT-MP asymmetry.

Keywords: conditional reasoning; probabilistic reasoning; Bayesian model; computational-level account

Introduction

Conditionals of the form "If A, then C" – for example, "If global warming continues, then London will be flooded" – are ubiquitous in everyday language and scientific discourse. One research question that has attracted a lot of attention is how individuals reason with conditionals. Usually four conditional inferences are studied, each consisting of the conditional as the major premise, a categorical minor premise, and a putative conclusion:

- *Modus Ponens* (MP): If A then C. A. Therefore, B.
- *Affirmation of the Consequent* (AC): If A then C. C. Therefore, A.
- *Denial of the Antecedent* (DA): If A then C. Not A. Therefore, not B.
- *Modus Tollens* (MT): If A then C. Not C. Therefore, not A.

According to classical logic MP and MT are valid (i.e., truth preserving) inferences and AC and DA are not valid. Early research with conditional inferences has emulated the inference process of classical logic; in the *abstract task*, inferences are presented with abstract content, participants are asked to treat the premises as true, and are asked to only accept necessary conclusions. Results generally showed that even untrained participants are able to distinguish valid from invalid inferences (i.e., they accept more valid than invalid inferences). However, their behavior is clearly not in line with the norms of classical logic. Whereas participants tend to unanimously accept the valid MP, the acceptance rates for the equally valid MT inference scheme is considerably lower. In a meta-analysis of the abstract task, Schroyens, Schaeken, and d'Ydewalle (2001) found acceptance rates of .97 for MP compared to acceptance rates of .74 for MT. This *MP-MT asymmetry* will be the main focus of the present manuscript.[1]

Research in recent years has moved away from the abstract task and its focus on logical validity towards tasks more akin to real-life reasoning within a *probabilistic framework* (Oaksford & Chater, 2007; Over, 2009). In the *probabilistic task*, inferences employ everyday content for which participant posses relevant background knowledge and they are usually asked for their subjective degree of belief in the putative conclusions. The degree of belief in the conclusion of course depends on the actual content (i.e., the probabilistic relationships among premises and conclusion), but there is still ample evidence for an MP-MT asymmetry that goes beyond what would be expected from existing probabilistic accounts. For example, Oaksford, Chater, and Larkin (2000) created materials for which their Bayesian model of conditional reasoning predicted participants to posses similar beliefs in MP and MT. Their results showed that, whereas this reduced the asymmetry, there were still differences such that participants expressed stronger beliefs in MP than MT. Similarly, Singmann, Klauer, and Over (2014) asked participants for their subjective degrees of belief, first in both premises, and then in the conclusion and showed that those only formed a coherent probability distribution "above chance" for MP, but not for MT. Essentially the same results were obtained by Evans, Thompson, and Over (2015). Together, these findings suggest a clear limit for simple probabilistic accounts of conditional reasoning.

Existing Accounts

To describe existing accounts and our new explanation, let us formalize the probabilistic structure of the reasoning problem. We consider an agent who entertains the propositions A (the antecedent) and C (the consequent) of a conditional

[1] Participants also tend to erroneously accept the invalid inferences AC and DA. Schroyens et al. (2001) report acceptance rates of .64 for AC and .56 for DA.

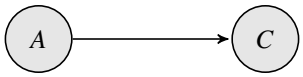


Figure 1: The Bayesian Network representation of the relation between A and C.

"If A, then C". To proceed, we introduce binary propositional variables A and C (in italic script) which have the values A and ¬A, and C and ¬C (in roman script), respectively. A prior probability distribution P is defined over these variables. It is represented by the Bayesian Network in Figure 1. For now, the exact parameterization of P is not yet relevant, however there exist several with three free parameters. In addition, we define the (absolute) *endorsement* of MP as $E_1 := P'(\mathrm{C})$. Similarly, we define the (absolute) endorsement of DA as $E_2 := P'(\neg\mathrm{C})$, the (absolute) endorsement of AC as $E_3 := P'(\mathrm{A})$, and the (absolute) endorsement of MT as $E_4 := P'(\neg\mathrm{A})$.

The original Oaksford et al. (2000) model makes two assumptions. First, it assumes that belief in the conclusion reflects the conditional probability of the conclusion given the minor premise. For example, $E_1 = P(\mathrm{C}|\mathrm{A})$ and $E_4 = P(\neg\mathrm{A}|\neg\mathrm{C})$. Second, it assumes that P is fixed throughout the reasoning process; that is, P is the same for responses to the four conditional inferences (Oaksford and Chater call this the "invariance assumption"). In other words, this model assumes that reasoning amounts to consulting ones fixed probability distribution and responding in line with it (e.g., by sampling from memory; Costello & Watts, 2014) . As shown by Oaksford and Chater (2007, ch.5) this model does a good job in accounting for many of the existing data from the abstract task, but underestimates the MP-MT asymmetry.

To account for the MP-MT asymmetry, the solution first proposed in Oaksford and Chater (2007, ch.5) and subsequently defended in (Oaksford & Chater, 2008, 2013) is to give up on the second of their assumptions, that P is fixed for responses to all four inferences. Specifically, they argue (e.g., Oaksford & Chater, 2013) that MP represents a special case that does not require changing P as it basically reflects the probabilistic information already present in the conditional (i.e., $P(\mathrm{C}|\mathrm{A})$). Thus, presenting MP does not allow the agent to learn new information about P. However, the other three inferences, MT, AC, and DA, present new information and thus require an updated probability distribution P', which individuals "learn". Practically, they did not specify many restriction of P', other than that $P(\mathrm{C}|\mathrm{A}) > P'(\mathrm{C}|\mathrm{A})$, which was primarily motivated by fitting their model to the extant data. From a statistical point of view, it not too surprising that a model that then essentially has one free parameter for fitting E_1 and three free parameters for fitting the remaining three observations (i.e., responses to the other three inferences, E_2, E_3, and E_4) does a relatively good job in accounting for the existing data.

Therefore, there are two main theoretical shortcomings in Oakford and Chater's approach. First, their revised model assumes that the endorsement for MP, E_1, comes from one probability distribution, P, whereas the endorsement for the other inferences, E_2 to E_4, comes from the updated probability distribution P'. This seems somewhat unsatisfactory from a rational Bayesian perspective and more of an ad-hoc solution than a principled argument. Second, the actual processes in which the agent updates P to arrive at P' are not specified well enough. What does it entail for the agent to learn the new information presented in MT? How can we characterize the cognitive processes involved in making a probabilistic MP or MT inference?

Our answer to these questions is based on Eva and Hartmann's (2018) recent Bayesian account of reasoning according to which "argumentation is learning". In line with Oaksford and Chater (2013), learning is specified as updating an agent's prior belief state, represented by P, in light of new information resulting in the posterior belief state P'. Specifically, the premises of an inference will affect specific parts of P (e.g., for MP, the agent learns the new values of both $P(\mathrm{C}|\mathrm{A})$ and $P(\mathrm{A})$). The novel assumption is that as a consequence, the agent needs to incorporate this new information into their existing beliefs which requires her to update potentially all parts of P. According to Eva and Hartmann (2018), this updating follows a well-defined Bayesian rule which generalizes conditonalization and Jeffrey conditionalization and requires that a suitably defined distance (or divergence) between P' and P is minimized. Eva and Hartmann (2018) argue that these divergencies should be members of the family of f-divergences. One important member of this family is the Kullback-Leibler (KL) divergence (Diaconis & Zabell, 1982), which we will use in the remainder. In this way, updating satisfies the constraints provided by the new information and is conservative (i.e., the changes are as minimal as possible). We will show that from this assumption, the typically found MP-MT asymmetry must appear for certain P. However, in some situations the opposite pattern (i.e., $E_4 > E_1$) should also be observed.

The Model

Our new explanation for the MP-MT asymmetry is based on the Bayesian Network in Figure 1 representing the prior probability distribution P. In addition, we assign

$$P(\mathrm{A}) = a, \tag{1}$$

for the prior probability of the antecedent and the conditional probabilities of the consequent C, given the values of its parent:

$$P(\mathrm{C}|\mathrm{A}) = p \quad , \quad P(\mathrm{C}|\neg\mathrm{A}) = q \tag{2}$$

With this, the joint prior probability distribution P over the variables A and C is given by

$$\begin{aligned} P(\mathrm{A},\mathrm{C}) &= a\,p \quad , \quad P(\mathrm{A},\neg\mathrm{C}) = a\,\overline{p} \\ P(\neg\mathrm{A},\mathrm{C}) &= \overline{a}\,q \quad , \quad P(\neg\mathrm{A},\neg\mathrm{C}) = \overline{a}\,\overline{q}, \end{aligned} \tag{3}$$

where we have used the shorthand notation $P(\mathrm{A},\mathrm{C})$ for $P(\mathrm{A}\wedge\mathrm{C})$ which we will use throughout this paper. We also use the shorthand $\overline{x}$ for $1-x$ and assume that $a,p,q\in(0,1)$.

Following the slogan "argumentation is learning", the agent then learns the premises of the argument. More specifically, she learns the major premise "If A, then C" and sets the new probability of $P'(\mathrm{C}|\mathrm{A})=p'=1$ in turn. This is the first constraint on P'. She also learns a minor premise: A in the case of MP, and ¬C in the case of MT. For completeness, we also consider AC and DA. In the case of AC, she additionally learns C, and in the case of DA she additionally learns ¬A. Following Eva and Hartmann (2018), we model this by assuming that the probability of the minor premise increases. This is the second constraint on P'. More specifically, we assume that the agent changes the probabilities of the minor premise in the following way:

$$\begin{aligned} P'_{MP}(\mathrm{A}) &= \lambda+\overline{\lambda}P(\mathrm{A}) \quad , \quad P'_{DA}(\mathrm{A}) = \overline{\lambda}P(\mathrm{A}) \\ P'_{AC}(\mathrm{C}) &= \lambda+\overline{\lambda}P(\mathrm{C}) \quad , \quad P'_{MT}(\mathrm{C}) = \overline{\lambda}P(\mathrm{C}) \end{aligned} \tag{4}$$

Here $\lambda\in(0,1]$ measures to what extent the agent changes the probability of the minor premise. For $\lambda\to 0$, the new probability of the minor premise does not change at all, and for $\lambda=1$ it goes to its maximal value, i.e. to 1.

To find the full new probability distribution P', we then minimize the KL-divergence between P' and P. This allows us to compute the new probability of the conclusion of the corresponding argument. For example, in the case of MP ($\mathrm{A},\mathrm{A}\to\mathrm{C}$, therefore C) the conclusion is C and the new probability of C, i.e. $P'(\mathrm{C})$ measures to what extent the agent endorses the corresponding inference pattern. More specifically, we define the (absolute) *endorsement* of MP as $E_1 := P'(\mathrm{C})$. As described above, we define the (absolute) endorsement of DA as $E_2 := P'(\neg\mathrm{C})$, the (absolute) endorsement of AC as $E_3 := P'(\mathrm{A})$, and the (absolute) endorsement of MT as $E_4 := P'(\neg\mathrm{A})$. Furthermore, we define the relative endorsement of inferences i and j as $\Delta_{ij} := E_i - E_j$ with $i<j$. These quantities will be calculated in the next section.

It is worth pausing here to note that these endorsement quantities should be conceptually distinguished from the corresponding acceptance rates discussed in the introduction. While the former are interpreted as representations of the extent to which a single idealised Bayesian agent will endorse an inference in a probabilistic reasoning task, the latter represent the relative frequency with which those inferences are accepted at the population level. There is no a-priori reason to expect a close correspondence between these two different quantities. In what follows, we try to explain the MP-MT asymmetry in terms of individual endorsement rates.

The Results

Our formal results can be summarized in the following two propositions (all proofs are in the Appendix):

Proposition 1 *An agent considers the binary propositional variables A and C with a probability distribution P defined over them. She then learns (i) the major premise of an argument and sets $P'(\mathrm{C}|\mathrm{A})=1$ and (ii) the minor premise and sets its new probability to a value according to eqs. (4) with $\lambda\in(0,1]$. To find the full new probability distribution P', we minimize the KL-divergence between P' and P. The (absolute) endorsements are then given by $E_1=\lambda+\overline{\lambda}P(\mathrm{A}\vee\mathrm{C})$, $E_2=\lambda P(\neg\mathrm{C}|\neg\mathrm{A})+\overline{\lambda}P(\neg\mathrm{A},\neg\mathrm{C})$, $E_3=\lambda P(\mathrm{A}|\mathrm{C})+\overline{\lambda}P(\mathrm{A},\mathrm{C})$ and $E_4=\lambda+\overline{\lambda}P(\neg\mathrm{A}\vee\neg\mathrm{C})$.*

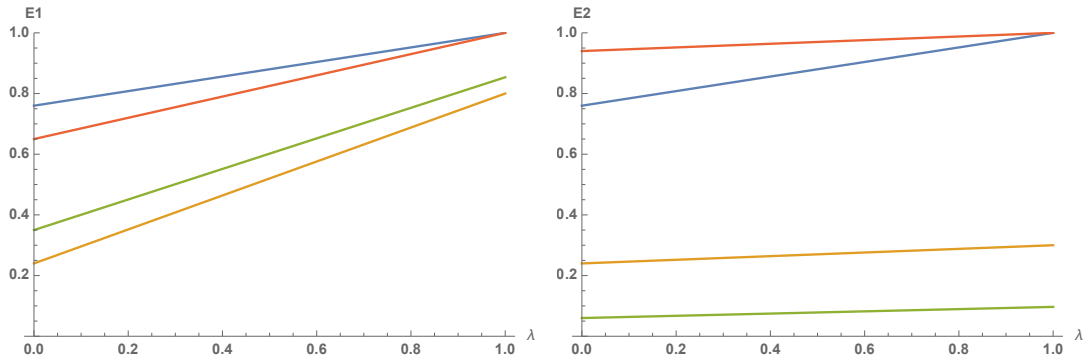
Proposition 2 *Proposition 1 implies the following statements: (i)* MP > AC. *(ii)* MT > DA. *(iii) If $P(\mathrm{A})\geq 1/2$, then* MP > DA *(iv) If $P(\mathrm{A},\mathrm{C})\geq P(\neg\mathrm{A},\neg\mathrm{C})$, then* MP > MT, AC > DA *and $E_1+E_2<E_3+E_4$. (v) If $P(\mathrm{A},\mathrm{C})\leq 1/2$, then* MT > AC. *(vi)* MP > MT *iff* AC > DA. *(vii) If $P(\mathrm{A}\vee\mathrm{C})\geq 1/2$, then* MP > DA.

Here we have used the notation MP > AC for $\Delta_{13}>0$ etc. Note that the assumptions stated in the various if-sentences in Proposition 2 are only sufficient conditions. It turns out that the respective consequents also hold in a large range of other contexts. These depend, however, on the value of both P and λ as shown in Figure 2.

The two left panels of Figure 2, panels (a) and (c), show a situation in which the probability of the antecedent is relatively high (i.e., large a), the conditional expresses a relationship with reasonable confidence (i.e., the conditional probability of the consequent given the antecedent, p, is at least .5), and exceptions are somewhat uncommon (i.e., relatively low conditional probability of the consequent given that the antecedent, q, does not hold). In this situation we see the typical MP-MT asymmetry pattern (as long as $\lambda<1$), when comparing the blue (MP) and red (MT) line. We also see that the degree of the MP-MT asymmetry crucially depends on λ and increases with decreasing λ. Furthermore, the degree of the MP-MT asymmetry also depends on the specific parameters of P. If the conditional expresses a more certain relationship, as in panel (c), the MP-MT asymmetry is larger than if the relationship expressed by the conditional is more uncertain, as in panel (a).

An interesting pattern is observed if the prior probability of the antecedent is low (i.e., $a<.5$), as shown in panels (b) and (d). We can see that in this case the sign of the MP-MT asymmetry flips. Now, we expect stronger endorsement to MT than to MP. However, as for the case in which the prior probability of the antecedent is relatively large, we see that the extent of this reversed asymmetry also depends on λ and the other parameters of P.

Figure 2 also shows the predicted endorsement for the other two inferences, AC and DA. Their ordering (i.e., whether endorsement is expected to be larger for AC or DA) follows the same general pattern also observed for MP and MT. For panels (a) and (c) we expect larger endorsement for AC and DA (as is commonly observed in the literature). However, if the prior probability of the antecedent is low, we expect the same flip; larger endorsement for DA than for AC. In addition, the figure shows another interesting empirical prediction. For certain values of P, see panel (c), we expect either AC > MT or MT > AC, depending on the value of λ



(a) $a = .7, p = .5$, and $q = .2$

(b) $a = .2, p = .3$, and $q = .7$

(c) $a = .8, p = .7$, and $q = .3$

(d) $a = .2, p = .7$, and $q = .3$

Figure 2: The absolute endorsements E_1 (MP, blue), E_2 (DA, orange), E_3 (AC, green) and E_4 (MT, red) as a function of λ for different prior probability distributions P.

(the qualitatively similar also holds between MP and DA, see panel (d)).

Discussion

These results show that the MP-MT asymmetry is predicted by the behavior of a rational agent who updates her belief after encountering new information that is part of the premises of a conditional inference under certain conditions. In contrast to previous probabilistic accounts (Oaksford & Chater, 2008, 2013), we do not need to assume two different probability distributions for MP and the other inferences. Instead, we describe a rational account of how agents update their beliefs in light of new information and use this updated probability distribution as the basis for her endorsement to the four conditional inferences. With this model, we can also make specific predictions when we would expect the opposite pattern, a MT-MP asymmetry.

Disabling Conditions

So far we have assumed that the agent only considers two propositions, i.e. A and C. In many cases, however, there are other relevant propositions and learning new information might affect them. This might have implications for the endorsement of the various inference patterns we have discussed. Consider the following case (Oaksford & Chater, 2008): Let A be the proposition "you turn the key of your car" and let C be the proposition "the car starts". You then learn the premises of a MT inference, i.e. $\mathrm{A} \rightarrow \mathrm{C}$ and $\neg\mathrm{C}$. In that case it seems reasonable to not infer $\neg\mathrm{C}$, but rather that the car is broken or, more generally, that a disabler is present (D). To model this situation, we consider the Bayesian Network in Figure 3 and assume that

$$P(\mathrm{A}) = a \quad , \quad P(\mathrm{D}) = d, \tag{5}$$

where a is large (you will be pretty certain that you turned the key of your car if you did so) and d is somewhat smaller, but it seems reasonable to take the possibility that the car might be broken into account before actually turning the key of the car.

Furthermore, we have to specify the likelihoods

$$\begin{aligned} P(\mathrm{C}|\mathrm{A},\mathrm{D}) &= \alpha \quad , \quad P(\mathrm{C}|\mathrm{A},\neg\mathrm{D}) = \beta \\ P(\mathrm{C}|\neg\mathrm{A},\mathrm{D}) &= 0 \quad , \quad P(\mathrm{C}|\neg\mathrm{A},\neg\mathrm{D}) = 0. \end{aligned} \tag{6}$$

Here we have assumed that the car does not start if the key is not turned. Note that the context suggests that $\beta > \alpha \approx 0$ although we will not need the left inequality. All we will need is that α is fairly small.

The agent then learns the conditional $\mathrm{A} \rightarrow \mathrm{C}$ which imposes the constraint $\beta' > \beta$ on P'.[2] The agent furthermore

[2] β' could be 1, but we will see that this does not matter. All we need is that $\beta, \beta' > 0$.

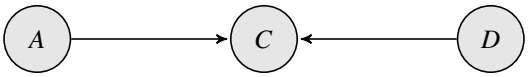


Figure 3: The Bayesian Network representation of the relation between A, C and D.

learns that $P'(\mathrm{C}) = 0$. (We set this value to 0 as it will be hard to doubt that the car did not start if in fact it did not start.) We then note that

$$P'(\mathrm{C}) = a'\,(d'\,\alpha' + \overline{d'}\,\beta') = 0, \tag{7}$$

where we have assumed that P' can be parameterized analogously to P. Given that $\beta' > \beta > 0$, eq. (7) has two solutions: (i) $a' = 0$ and (ii) $\alpha' = 0$ and $d' = 1$. Obviously, solution (i) corresponds to the proper MT inference. However, this inference is implausible in the present case. To explore this issue further, let us consider the KL-divergence between P' and P:

$$KL = \Phi_a + \Phi_d + a'\,d'\,\Phi_\alpha + a'\,\overline{d'}\,\Phi_\beta \tag{8}$$

We have to minimize KL with the constraint (7). Let us consider solution (i) first. Then $KL_1 = -\log\overline{a} + \Phi_d$. This expression minimizes for $d' = d$ and therefore $KL_1^{min} = -\log\overline{a}$. Next, consider (ii). Here $KL_2 = \Phi_a - a'\log\overline{\alpha} - \log d$. Minimizing this expression with respect to a' yields $a' = a\,\overline{\alpha}/(a\,\overline{\alpha} + \overline{a})$ and $KL_2^{min} = -\log((a\,\overline{\alpha} + \overline{a})\,d)$. Hence, $KL_2^{min} < KL_1^{min}$ iff $(a\,\overline{\alpha} + \overline{a})\,d > \overline{a}$ or $a\,\overline{\alpha} > \overline{a}\,\overline{d}$. This condition is fulfilled in the present case as $\alpha \approx 0$ and $a \approx 1$. (Note that the value of d does not matter too much here, but it should not be too low. If it is very low and the inequality is violated, then the agent should make a MT inference and infer $\neg$A.)

Conclusions

Our main goal was to provide a novel probabilistic explanation for the MP-MT asymmetry found in both the traditional abstract task as well as in probabilistic tasks with conditional inferences. In contrast to previous explanations within a probabilistic framework (Oaksford & Chater, 2007, 2013, 2008), our explanation is based on a principled approach of how agents update a probability distribution P in light of new information provided by the premises of a conditional inference resulting in a updated probability distribution P'. Following the idea that "argumentation is learning" (Eva & Hartmann, 2018), we propose that agents update their probability distributions in light of new information by minimizing the KL-divergence between the posterior and prior probability distribution. In this conceptualization, reasoning does not only amount to a read-out from memory, but requires the agent to actively integrate the new knowledge with the existing one. The exact cognitive processes how this is achieved (e.g., by creating new memory traces or overwriting existing ones), is an open question for future work. Our work provides a full *computational-level account* (in the sense of Marr, 1982) of conditional reasoning.

The theoretical results presented here provide evidence that the MP-MT asymmetry is a direct consequence from this Bayesian conceptualization of conditional reasoning. Specifically, it occurs if the prior probability of the conditional probability of C given A (i.e., the relationship expressed in the conditional) and the the prior probability of the antecedent is at least .5. In the case that these conditions do not hold, we expect the opposite pattern, an inverted MP-MT asymmetry.

Minimizing the KL-divergence, as proposed here, is one rational way for an agent to update her prior probability distribution in light of new information which implies Jeffrey conditionalization (Diaconis & Zabell, 1982). Importantly, the results shown here do not only apply to updating via minimizing the KL-divergence, but for updating based on minimizing the distance between P' and P for any divergence measure that is a member of the family of f-divergences. All these divergence metrics are rational in the same sense and also predict the MP-MT asymmetry under the same circumstances. This is an important aspect of our results in light of the findings of Singmann, Klauer, and Beller (2016). They have investigated the empirical adequacy of conditional reasoning based on KL-minimization between P' and P in a two-step conditional reasoning task – which allowed to obtain estimates of both P and P' in an independent manner – and found that it did not provide a very adequate account. However, as soon one is willing to give up the assumption that $P'(\mathrm{C}|\mathrm{A}) = p' = 1$ and assumes that $P'(\mathrm{C}|\mathrm{A}) < 1$ (as done in Singmann et al.'s study), different members of the family of f-divergences make different predictions. Preliminary work suggests that a more empirically adequate account of conditional reasoning is provided if we assume reasoners update their probability distribution by minimizing the inverse-KL divergence between prior and posterior distribution.

Proof of Proposition 1

We use the parameterization of the prior probability distribution P according to eqs. (1) and (2) and begin with MP and DA. Here we set the new value of the probability of the antecedent to a'. Disregarding constant terms, the KL-divergence is then given by $KL = \overline{a'}\,\Phi_q$ with

$$\Phi_x := x' \log\frac{x'}{x} + \overline{x'}\log\frac{\overline{x'}}{\overline{x}}\,. \tag{9}$$

Differentiating KL with respect to q' and setting the resulting expression equal to zero yields $q' = q$. Hence, $P'(\mathrm{C}) = a' + \overline{a'}\,q$. We now insert the appropriate values of c' from eqs. (4) and use the definition of the respective (absolute) endorsements to obtain

$$\begin{aligned} E_1 &:= P'(\mathrm{C}) \\ &= \lambda + \overline{\lambda}(a + \overline{a}\,q) \\ &= \lambda + \overline{\lambda}P(\mathrm{A} \vee \mathrm{C}) \\ E_2 &:= P'(\neg\mathrm{C}) \\ &= \lambda\,\overline{q} + \overline{\lambda}\,\overline{a}\,\overline{q} \\ &= \lambda P(\neg\mathrm{C}|\neg\mathrm{A}) + \overline{\lambda}P(\neg\mathrm{A}, \neg\mathrm{C}). \end{aligned}$$

Let us now consider AC and MT. In this case, learning the minor premise amounts to the constraint

$$a' + \overline{a'}\, q' = c', \tag{10}$$

with c' specified in eqs. (4). We therefore have to minimize the function

$$L = \Phi_a + a' \log \frac{1}{p} + \overline{a'}\, \Phi_q + \mu(a' + \overline{a'}\, q' - c'),$$

with the Lagrange multiplier μ.

Differentiating L with respect to q' and setting the resulting expression equal to zero yields

$$q' = \frac{1}{q + \overline{q}x}, \tag{11}$$

with $x := \exp(\lambda)$. Hence,

$$L = \Phi_a + a' \log \frac{1}{p} + \overline{a'} \log \frac{1}{q + \overline{q}x} + \mu \overline{c'}.$$

Differentiating this expression with respect to a' and setting the resulting expression equal to zero yields

$$a' = \frac{a\,p}{a\,p + \overline{a}\,(q + \overline{q}x)}. \tag{12}$$

From eqs. (10), (11) and (12), we then obtain

$$a' = \frac{a\,p\,c'}{a\,p + \overline{a}\,q}. \tag{13}$$

We now insert the appropriate values of c' from eqs. (4) in eq. (13) and use the definitions of the respective (absolute) endorsements to obtain

$$\begin{aligned}
E_3 &:= P'(\mathrm{A}) \\
&= \lambda \frac{a\,p}{a\,p + \overline{a}\,q} + \overline{\lambda}\, a\,p \\
&= \lambda P(\mathrm{A}|\mathrm{C}) + \overline{\lambda} P(\mathrm{A},\mathrm{C}) \\
E_4 &:= P'(\neg \mathrm{A}) \\
&= \lambda + \overline{\lambda}(1 - a\,p) \\
&= \lambda + \overline{\lambda} P(\neg \mathrm{A} \vee \neg \mathrm{C}).
\end{aligned}$$

This completes the proof of Proposition 1. ∎

Proof of Proposition 2

We use Proposition 1 to compute the relative endorsements:

$$\begin{aligned}
\Delta_{12} &= \lambda q + \overline{\lambda}\,[2(a + \overline{a}\,q) - 1] \\
\Delta_{13} &= \lambda \frac{\overline{a}\,q}{a\,p + \overline{a}\,q} + \overline{\lambda}(a\,\overline{p} + \overline{a}\,q) \\
\Delta_{14} &= \overline{\lambda}(a\,p - \overline{a}\,\overline{q}) \\
\Delta_{23} &= -(a\,p - \overline{a}\,\overline{q}) \cdot \left[\lambda \frac{q}{a\,p + \overline{a}\,q} + \overline{\lambda}\right] \\
\Delta_{24} &= -\lambda q - \overline{\lambda}(a\,\overline{p} + \overline{a}\,q) \\
\Delta_{34} &= -\lambda \frac{\overline{a}\,q}{a\,p + \overline{a}\,q} + \overline{\lambda}(2a\,p - 1)
\end{aligned}$$

From these results, the statements made in the proposition follow. For example, the third statement in (iv) follows by noting that $\Delta_{14} + \Delta_{23} < 0$. ∎

References

Costello, F., & Watts, P. (2014). Surprisingly rational: Probability theory plus noise explains biases in judgment. *Psychological Review*, *121*(3), 463–480. doi: 10.1037/a0037010

Diaconis, P., & Zabell, S. L. (1982). Updating Subjective Probability. *Journal of the American Statistical Association*, *77*(380), 822–830. doi: 10.1080/01621459.1982.10477893

Eva, B., & Hartmann, S. (2018). Bayesian Argumentation and the Value of Logical Validity. *Psychological Review*, *125*(5), 806–821. doi: 10.1037/rev0000114

Evans, J. S. B. T., Thompson, V. A., & Over, D. E. (2015). Uncertain deduction and conditional reasoning. *Cognition*, *6*, 398. doi: 10.3389/fpsyg.2015.00398

Marr, D. (1982). *Vision: a computational investigation into the human representation and processing of visual information*. San Francisco: W.H. Freeman.

Oaksford, M., & Chater, N. (2007). *Bayesian rationality: the probabilistic approach to human reasoning*. Oxford; New York: Oxford University Press.

Oaksford, M., & Chater, N. (2008). Probability logic and the Modus Ponens – Modus Tollens asymmetry in conditional inference. In N. Chater & M. Oaksford (Eds.), *The Probabilistic Mind:: Prospects for Bayesian cognitive science* (pp. 97–120). Oxford University Press.

Oaksford, M., & Chater, N. (2013). Dynamic inference and everyday conditional reasoning in the new paradigm. *Thinking & Reasoning*, *19*(3-4), 346–379. doi: 10.1080/13546783.2013.808163

Oaksford, M., Chater, N., & Larkin, J. (2000). Probabilities and polarity biases in conditional inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(4), 883–899. doi: 10.1037/0278-7393.26.4.883

Over, D. E. (2009). New paradigm psychology of reasoning. *Thinking & Reasoning*, *15*(4), 431–438. doi: 10.1080/13546780903266188

Schroyens, W. J., Schaeken, W., & d'Ydewalle, G. (2001). The processing of negations in conditional reasoning: A meta-analytic case study in mental model and/or mental logic theory. *Thinking & Reasoning*, *7*(2), 121–172. doi: 10.1080/13546780042000091

Singmann, H., Klauer, K. C., & Beller, S. (2016). Probabilistic conditional reasoning: Disentangling form and content with the dual-source model. *Cognitive Psychology*, *88*, 61–87. doi: 10.1016/j.cogpsych.2016.06.005

Singmann, H., Klauer, K. C., & Over, D. E. (2014). New normative standards of conditional reasoning and the dual-source model. *Frontiers in Psychology*, *5*, 316. doi: 10.3389/fpsyg.2014.00316

Children's overextension as communication by multimodal chaining

Renato Ferreira Pinto Junior (renato@cs.toronto.edu)
Department of Computer Science
University of Toronto

Yang Xu (yangxu@cs.toronto.edu)
Department of Computer Science
Cognitive Science Program
University of Toronto

Abstract

Young children often stretch terms to novel objects when they lack the proper adult words—a phenomenon known as overextension. Psychologists have proposed that overextension relies on the formation of a chain complex, such that new objects may be linked to existing referents of a word based on a diverse set of relations including taxonomic, analogical, and predicate-based knowledge. We build on these ideas by proposing a computational framework that creates chain complexes by multimodal fusion of resources from linguistics, deep learning networks, and psychological experiments. We test our models in a communicative scenario that simulates linguistic production and comprehension between a child and a caretaker. Our results show that the multimodal semantic space accounts for substantial variation in children's overextension in the literature, and our framework predicts overextension strategies. This work provides a formal approach to characterizing linguistic creativity of word sense extension in early childhood.

Keywords: language acquisition; linguistic creativity; overextension; word sense extension; multimodality; chaining; communication

Young children often stretch terms to describe novel objects when they lack the proper adult words, a phenomenon known as overextension (Clark, 1978). Overextension is a communicative strategy that draws on knowledge of diverse relations in the world. For instance, a child may use "dog" to refer to a *squirrel*, "ball" to refer to a *balloon*, or "key" to refer to a *door*. This creative use of words toward novel meanings, or *word sense extension*, is not only attested in child language acquisition, but it is also reflected in historical meaning change, e.g., we extended the meaning of "mouse" from a rodent to a computer device. We explore the origin of word sense extension by asking how the cognitive capacity of overextension in childhood can be characterized formally.

Early work by Vygotsky (1962) suggests that overextension relies on "chain complex", a critical element of concept formation in childhood. He demonstrated chain complex by a series of overextension cases from a child who extended the meanings of "quah" to wide-ranging things including a duck, water in a pond, liquids in general, an eagle on a coin, and any coin-like objects. Vygotsky's account resonates with work from philosophy and cognitive linguistics that suggest the complex structure of word meanings (e.g. Wittgenstein, 1953) is formed possibly due to a process of chaining (Lakoff, 1987), where one referent is linked to another forming a chain-like structure. More recent work has shown that chaining predicts word sense extension in the history of English (Ramiro, Srinivasan, Malt, & Xu, 2018) and other languages (Xu, Regier, & Malt, 2016). However, these works did not offer a formal account of how one might represent the rich knowledge in a chain complex.

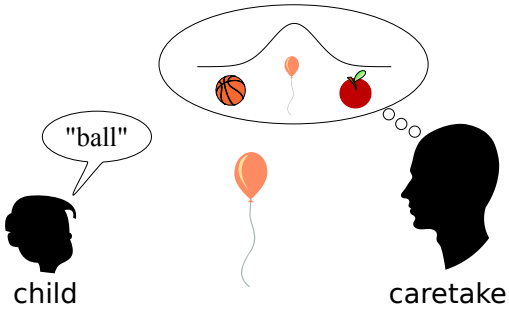


Figure 1: Overextension in child-caretaker communication.

Empirical work from Rescorla (1980) provided clues to the knowledge underlying children's overextension. Specifically, she identified three main types of relations between core and overextended meanings of a word, summarized as 1) *categorical* relation: overextension by linking objects within a taxonomy, e.g., "dog"→*squirrel*, 2) *analogy* or visual analogy: overextension by linking objects with shared perceptual properties, e.g., "ball"→*balloon*, and 3) *predicate-based* relation: overextension by linking objects that co-occur frequently in the environment, e.g., "key"→*door*. An open question we address in this work is how to combine these types of relations to predict overextension strategies in early childhood.

We propose a computational framework that considers overextension as a communicative game between a child and a caretaker, illustrated in Figure 1. The game involves a child and a caretaker in a situation where the child needs to refer to an out-of-vocabulary novel object. In this context, the child faces a *production problem*, where the goal is to extend a word from the existing vocabulary (e.g., "ball") to the novel object (e.g., a balloon). The caretaker instead faces a *comprehension problem*, where the goal is to guess the intended referent based on the child's utterance. Since "ball" does not typically map to balloons, we wish to reconstruct the cognitive processes that could have given rise to successful communication between the child and the caretaker in common cases of overextension. As such, our framework should support both strategic word choices for the child and prediction of intended referents for the caretaker.

Our communication-based framework relates to earlier work in overextension from the developmental literature. For example, Bloom (1973) argued that overextension is a performance error caused by vocabulary limitations, whereby a child may consciously use an incorrect word (from the adult perspective) as a strategy to convey the desired referent meaning. A related hypothesis poses overextension as a retrieval error (Fremgen & Fay, 1980; Gershkoff-Stowe, 2001; Huttenlocher, 1974; Thomson & Chapman, 1977), suggesting that children may overextend an earlier acquired word even if the correct adult word has been partially acquired (e.g., understood in comprehension), because the latter may be more difficult to produce. We explore evidence for a retrieval error hypothesis by evaluating whether children may favour words with higher usage frequencies in their overextended word choices. Another extensive line of research suggests that overextension arises from children's incomplete conceptual knowledge of the semantic features underlying different categories (Clark, 1973; Kay & Anglin, 1982; Mervis, 1987). While we do not directly test claims about children's conceptual space in this work, we show that a combination of semantic relations helps to explain overextension strategies, and may play an integral role in characterizing the mechanisms that subserve children's early word learning.

Our framework also draws on a multimodal space of semantic relations (cf. Rescorla, 1980) that serves as the knowledge engine for creating chain complexes in overextension. The notion of multimodality is motivated in part by work on visually-grounded word learning (e.g. Lazaridou, Chrupała, Fernández, & Baroni, 2016; Roy & Pentland, 2002; Yu, 2005), which shows that perceptual features play an important role in children's acquistion of core (or conventional) word meanings. Our focus is on investigating how by integrating diverse semantic relations one might account for word usage beyond the core meanings that children normally acquire. Our work thus differs from the extensive literature on cross-situational word learning (Fazly, Alishahi, & Stevenson, 2010; Frank, Goodman, & Tenenbaum, 2009; Kachergis, Yu, & Shiffrin, 2017; Siskind, 1996), where the emphasis has been typically on modeling children's behaviour in learning conventional word meanings, but not on how they extend existing terms to describe novel objects. Our work also extends existing computational studies that explore overextension in specific domains such as color terms (Beekhuizen & Stevenson, 2016) to more general cases of overextension that involve mappings across domain boundaries, e.g., "ball"→*balloon*.

Computational framework

We present our computational framework for overextension following two steps: 1) Specification of a probabilistic model that simulates child's word choices (production) and caretaker's inference of intended referents (comprehension); 2) Construction of a semantic space that supports multimodal chaining of word meanings, encapsulated in the same model.

For this work, we focus on overextension of nouns, but the general framework that we present can be used to explore other types of overextension (e.g., in verbs and adjectives).

Probabilistic formulation

We formulate overextension as communication between a child and a caretaker. In particular, the child wishes to refer to a novel object c in an environment E. The child does so by choosing (and stretching) a word w from her vocabulary V. We assume that the correct term for the novel object is not yet acquired by the child, hence $c \neq w$ and $c \notin V$. Based on the child's utterance w, the caretaker wishes to infer the referent c among possible referents in E. We then model the child's behaviour by a *production model* and pair it with a *comprehension model* for the caretaker's behaviour.

Production. We cast the production problem as probabilistic inference over existing words in the child's vocabulary given the probe novel object c, via Bayes' rule:

$$p_{\text{prod}}(w|c) \propto p_{\text{prod}}(c|w)p(w) \tag{1}$$

We define the prior $p(w)$ proportional to the logarithmic usage frequency of a word with add-one smoothing $p(w) \propto \log(1+\text{freq}(w))$. This formulation is consistent with the frequency effect found in the study of overextension in color terms (Beekhuizen & Stevenson, 2016). It captures the intuition that all things being equal, the child is more likely to choose a common word versus a rare word for overextension. We define the likelihood function $p_{\text{prod}}(c|w)$ by a meta similarity measure that encapsulates the three types of semantic relations reported by Rescorla (1980) which the novel referent c can bear with the existing referent c_w of word w:

$$\begin{aligned} p_{\text{prod}}(c|w) &\propto \text{sim}(c, c_w) \\ &= \exp\left(-\frac{d_c(c,c_w)+d_v(c,c_w)+d_p(c,c_w)}{h}\right) \end{aligned} \tag{2}$$

We take the exponential-decay form from the generalized context model (GCM) or exemplar model of categorization (Nosofsky, 1986), where the influence of each relational type is proportional to how similar c and c_w are under that relation. We represent similarity by inverse distance, where d_c, d_v, and d_p represent distances measured according to categorical relation, visual analogy, and predicate-based relation, respectively. We describe the construction of each of these relational features in the next section. To control for model sensitivity to these distance functions, we use a single parameter h that we estimate empirically from data. The magnitude of h determines how slowly the meta similarity or the likelihood function decreases with respect to the distance measured in the multimodal relations.

Comprehension. We pair the child's production model with a comprehension model for the caretaker. Specifically, the caretaker solves the inverse inference problem as the child by a probability distribution over the space of intended referents based on the child's utterance w:

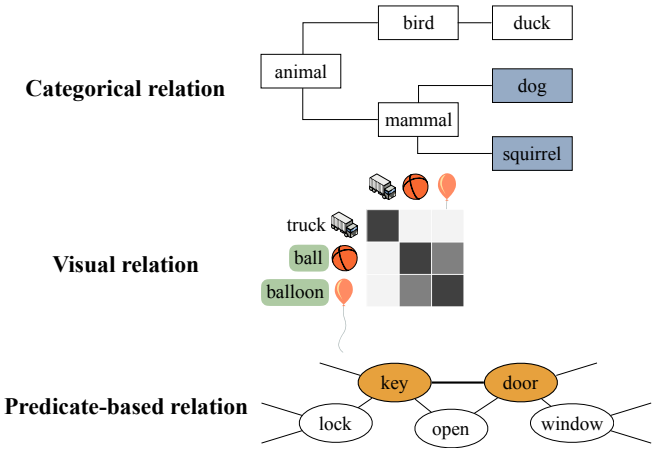


Figure 2: Types of semantic relations in multimodal space.

$$p_{\text{comp}}(c|w) \propto p_{\text{comp}}(w|c)p(c) \tag{3}$$

We consider the space of c to be all referents that appear in the communicative environment E where the child and the caretaker are situated in. We also assume a uniform prior $p(c)$ on possible referents in this environment, although it may be possible to enrich this prior by considering perceptual salience, eye gaze, pointing, and other pragmatic cues, which we do not model or have explicit access to in this work.

We define the likelihood function identical to the formulation used in the production model, under the assumption that both the child and the caretaker have knowledge of the multimodal relations:

$$p_{\text{comp}}(w|c) \propto \text{sim}(c_w, c) = \text{sim}(c, c_w) \tag{4}$$

Although it is possible to model perspective taking in a recursive way $p_{\text{comp}}(w|c) = p_{\text{prod}}(w|c) = p_{\text{prod}}(c|w)p(w)$ under the assumption that the caretaker takes into account the child's word choice in guessing the intended referent, e.g., similar to the rational speech act model (Goodman & Frank, 2016), we choose to work with the simplest version of this model that does not make any recursive assumption in the caretaker. We show in *Results* that our framework accounts for data well even without this assumption.

Multimodal semantic space

We define a multimodal semantic space that captures the three types of relational features in Rescorla (1980): categorical relation, visual analogy, and predicate-based relation. We construct these relational features using a fusion of resources drawn from linguistics, deep learning networks, and psychological experiments, as illustrated in Figure 2.

Categorical relation. We define categorical relation between two referents via a standard distance measure d_c in natural language processing by Wu and Palmer (1994), based on taxonomic similarity. Concretely, for two concepts c_1 and c_2 under a taxonomy T (i.e., a tree), the distance is:

$$d_c(c_1, c_2) = 1 - \frac{2N_{\text{LCS}}}{N_1 + N_2} \tag{5}$$

N_{LCS} denotes the number of shared parent nodes of the two concepts in the taxonomy. N_1 and N_2 denote the depths of the two concepts in the taxonomy. This distance measure is effectively the negated taxonomic similarity between c_1 and c_2, and is bounded between 0 and 1. Under this measure, concepts from the same semantic domain (such as *dog* and *squirrel*) should yield a lower distance than those from across domains (such as *ball* and *balloon*). To derive the categorical features, we took the taxonomy from WordNet (Miller, 1995) and annotated words by their corresponding *synset*'s in the database. We used the *NLTK* package (Bird & Loper, 2004) to calculate similarities between referents for this feature.

Visual analogical relation. We define visual analogical relation by cosine distance between vector representations of referents in visual embedding space. In particular, we extracted the visual embeddings from convolutional neural networks—VGG-19 (Simonyan & Zisserman, 2015), a state-of-the-art convolutional image classifier pre-trained on the ImageNet database (Deng et al., 2009)—following procedures from work on visually-grounded word learning (Lazaridou et al., 2016). Under this measure, concepts that share visual features (such as *ball* and *balloon*, both of which are round objects) should yield a relatively low distance even if they are remotely related in the taxonomy. To obtain a robust visual representation for each concept c, we sampled a collection of images $I_1, \ldots, I_k$ up to a maximum of 512 images from ImageNet. With each image I_j processed by the neural network, we extracted the corresponding visual feature vector from the first fully-connected layer after all convolutions: v_j^c. We then averaged the sampled k feature vectors to obtain an expected vector v^c for the visual vector representation of c.

Predicate-based relation. We define predicate-based relation by leveraging the psychological measure of word association. We assume that two referents that frequently co-occur together should also be highly associable, e.g., *key* and *door*. Specifically, we followed the procedures in De Deyne, Navarro, Perfors, Brysbaert, and Storms (2018) and took the "random walk" approach to derive vector representations of referents in a word association probability matrix. This procedure generates word vectors based on the positive pointwise mutual information from word association probabilities propagated over multiple leaps in the associative network. As a result, concepts that share a common neighbourhood of associates are more likely to end up closer together in the vector space. De Deyne et al. (2018) showed that this measure yields superior correlations with human semantic similarity judgements in comparison to other measures of association. We used word association data from the English portion of the Small World of Words project (De Deyne et al., 2018). The data is stored as a matrix of cue-target association probabilities for a total of 12292 cue words. We used the implementation provided by the authors (`https://github.com/SimonDeDeyne/SWOWEN-2018`)

to compute vector representations from the association probability matrix. We used cosine distance to compute predicate-based distances between pairs of referent vectors.

To ensure that the three types of relational features provide complementary information, we calculated their intercorrelations based on 66 concept pairs that we used for our analyses. Although correlations were significant ($p < .001$), all coefficients were low (category & visual: 0.179; category & predicate: 0.186; visual & predicate: 0.274).

Data

We collected linguistic data from three sources: 1) Metadata of child overextension from the literature; 2) Vocabulary of early childhood; 3) Text corpora of child-caretaker speech.

Metadata of child overextension. We performed a meta survey of 12 representative studies from developmental psychology and collected a total of 86 overextension example word-referent pairs. Each pair consists of an overextended word and the novel referent that word has been extended to. We kept word-referent pairs that overlapped with the available data from the three features we described, resulting in a total of 66 word-referent pairs. Table 1 shows examples from this meta dataset and their sources from the literature.

While the data we used for analysis may not constitute an unbiased sample of child overextension, two factors help to alleviate this concern. First, we followed a systematic approach in data collection by recording every utterance-referent pair in which both constituents could be denoted by one noun. Second, the diversity of the sources that we examined reduces the possibility of biasing our sample from any individual study.

Table 1: Examples of overextension data.

Uttered word → *Referent*	Source
"banana" → *moon*	Behrend, D. A. (1988)
"car" → *truck*	Fremgen, A., & Fay, D. (1980)
"apple" → *orange juice*	Rescorla, L. A. (1981)
"ball" → *bead*	Barrett, M. D. (1978)
"fly" → *toad*	Clark, E. V. (1973)
"cow" → *horse*	Gruendel, J. M. (1977)
"apple" → *egg*	Rescorla, L. A. (1980)

Vocabulary from early childhood. To approximate children's vocabulary in early childhood, we collected nouns reported to be produced by children of up to 30 months of age from the American English subset of the Wordbank database (Frank, Braginsky, Yurovsky, & Marchman, 2017). Because overextension has been typically reported to occur between 1;1 and 2;6 years (Clark, 1973) (that covers the range in Wordbank), we constructed a vocabulary V using all the nouns from Wordbank for which we could obtain the required semantic features. The resulting vocabulary includes 316 out of the 322 nouns from the database.

Corpora of child-caretaker speech. To evaluate our models in a realistic communicative context, we collected a large set of child-caretaker speech transcripts from the CHILDES database (MacWhinney, 2014), for child Eve (age 1;6 to 2;3) from the Brown corpus (Brown, 1973), Peter (1;9 to 3;1) from the Bloom70 corpus (Bloom, Hood, & Lightbown, 1974), and Nina (1;11 to 3;3) from the Suppes corpus (Suppes, 1974). We chose these children's data because their ages closely match the typical overextension period reported in child development. We considered each transcript as forming a communicative environment, and from each environment, we collected the set of all nouns uttered by the child and the caretaker for the analyses detailed in the next section. In total, we obtained 1586 communicative environments with a median of 139 distinct nouns per context.

Results

We assess our proposed framework in three aspects: 1) model accuracy in reconstructing child and caretaker strategies in overextension; 2) evidence for multimodal chaining in overextension; 3) model generation of chain complex.

Model reconstruction of overextension strategies

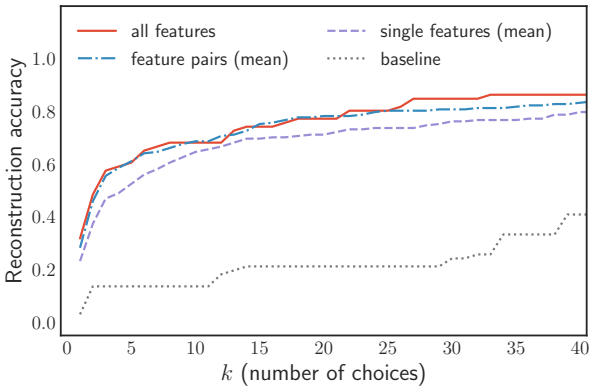
Production. We evaluated the child production model against the curated set of overextension word-referent pairs, $O = \{(w_i, c_i)\}$, with respect to all words in the child vocabulary V. For each pair, the model chooses the target word based on the given overextended sense c_i by assigning a probability distribution over words w in V. We assessed the model by finding the maximum *a posteriori* probability (MAP) of all the overextension pairs under the single sensitivity parameter h, which we optimized to the MAP objective function via standard stochastic gradient descent:

$$\max_h \prod_i p_{\text{prod}}(w_i|c_i;h,V) = \max_h \prod_i \frac{p_{\text{prod}}(c_i|w_i;h)p(w_i)}{\sum_{w\in V} p_{\text{prod}}(c_i|w;h)p(w)} \tag{6}$$

To assess the contribution of the three relational features, we tested this production model under single features and all possible combinations of features in pairs and triplets. We also compared these models under the frequency-based prior versus those under a uniform prior, along with a baseline model that chooses words only based on the prior distribution. We evaluated all models under two metrics: the Bayesian information criterion (BIC), which is a standard measure for probabilistic models that considers both degree of fit to data (i.e., likelihood) and model complexity (i.e., number of free parameters); a performance curve that measures model accuracy at different values of k, similar to the standard receiver-operating curve (ROC), where we assessed the predictive accuracy of each model from its choice of top k words for different levels of k, or the proportion of overextension pairs (w_i, c_i) for which the model ranks the correct production w_i among its top k predictions for referent c_i.

The left two columns of Table 2 summarize the BIC scores of the family of production models. We made three observations. First, models that incorporate features performed

(a) Production model



(b) Comprehension model

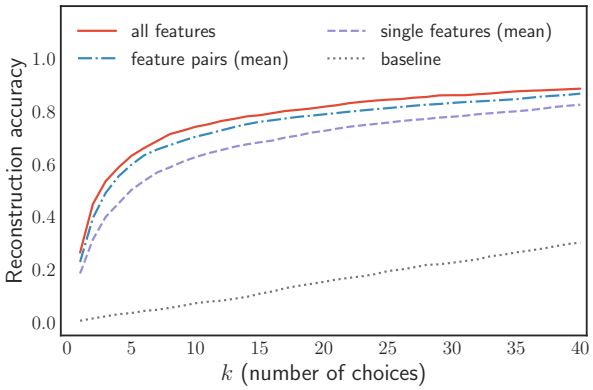


Figure 3: Performance curves for production and comprehension models.

Table 2: Bayesian Information Criterion (BIC) scores for production and comprehension (comp.) models.

	Production		**Comp.**
Model likelihood	freq. prior	unif. prior	unif. prior
Baseline	695	760	13268
category (cat.)	502	583	9663
visual (vis.)	496	574	10344
predicate (pred.)	499	582	9890
cat. + vis.	454	537	8949
cat. + pred.	457	544	8885
vis. + pred.	461	546	9261
all features	**439**	**526**	**8594**

better than the baseline (i.e., lower in BIC scores), suggesting that children overextend words by making explicit use of the semantic relations we considered. Second, models with the frequency-based prior performed dominantly better than those with the uniform prior, suggesting that children jointly consider word usage frequency (or effort) and semantic relations in overextension. Third, models with featural integration performed better than those with isolated features (i.e., all features<features pairs<single features in BIC score), suggesting that children rely on multiple kinds of semantic relations in overextensional word choices. Figure 3a further confirms these findings in performance curves that show the average predictive performance under the full range of k in top k modelled word choices: all features>features pairs>single features>baseline in the area under curves.

Comprehension. We next assessed the caretaker comprehension model by asking whether the model can retreive the intended referent from an uttered overextended word. Because we do not have the actual records of caretakers' inferences, we simulated a dataset for model evaluation by 1) identifying child-caretaker speech scripts that contain the overextended referents $\{c_i\}$ from our curated data; 2) replacing the correct word for a referent c_i (in the script) with the overextended word w_i reported in the literature. We then examined if the model is able to retrieve the correct referent c_i based on w_i among all other competing nouns in the communicative context of a script. As an example, knowing that "ball" has been reported to be overextended to *balloon*, we would identify child speech scripts that contain the word "balloon" and replace that word with "ball". We would then run our comprehension model and check if the top referents recovered by the model contain "balloon" among other nouns that appeared as context in that given script.

Similar to the case of production, we assessed the model by maximizing the posterior comprehension probability over all curated referents based on their appearances in the CHILDES transcripts. We optimized the MAP objective function under the sensitivity parameter h using stochastic gradient descent:

$$\max_h \prod_i p_{\text{comp}}(c_i|w_i;h,E_i) = \max_h \prod_i \frac{p_{\text{comp}}(w_i|c_i;h)p(c_i)}{\sum_{c\in E_i} p_{\text{comp}}(w_i|c;h)p(c)} \quad (7)$$

We used the same two metrics to evaluate the family of comprehension models and summarized the BIC-based results in the third column of Table 2 and ROC-based results in Figure 3b. We observed that results are qualitatively similar to those obtained in the production model: the rank order of performance among baseline model, models with single features, feature pairs, and all features, remains unchanged.

Evidence for multimodal chaining

To examine directly how the multimodal semantic space we constructed accounts for variation in the overextension data, we performed a logistic regression analysis. In particular, we considered two sets of data: the *attested set* overextension word-referent pairs, and a *control set* that shuffles the word-referent mappings from the attested set. We then performed a binary classification task via logistic regression to assess whether the attested pairs can be detected from the control pairs, given the same three relational features that we used for our previous analyses. The logistic model achieved 83% accuracy, compared to 50% chance. We also trained models on subsets of the feature space, achieving best feature pair performance of 82% and best single feature performance of 80%. This suggests that semantic relations provide significant predictability of concepts that might undergo overextension.

Figure 4 shows the distribution of dominant features across the 66 overextension pairs (we labelled each pair according to the top-scoring feature in the logistic regression model), along with a few examples that are best explained by each relational type. We observed that the contributions of these features are roughly even, providing support for the view that children rely on a combination of modalities in overextension.

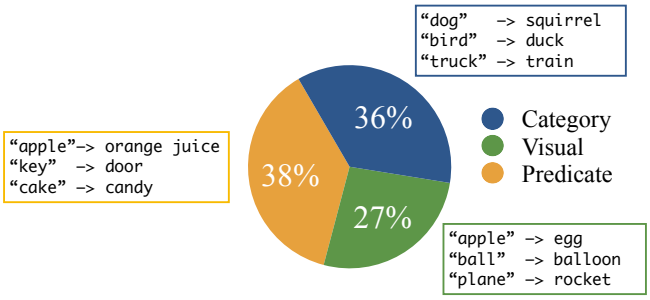


Figure 4: Percentage shares and examples explained by the three types of features from the curated overextension dataset.

Model simulation of chain complex

To illustrate how our model might simulate a chain complex, we applied an iterative scheme to sample a chain of concepts from the multimodal semantic space. Specifically, we began the chaining process with a seed word w_0 and initial chain $C^0 = \{w_0\}$. In the j-th iteration, we sampled word w uniformly from C^{j-1}, and word w' from children's vocabulary V according to probability distribution $p(w'|w) \propto \text{sim}(c_{w'}, c_w)$, where the similarity function is defined in Equation 2. We then added w' to the chain complex by linking it to w, hence extending the chain to $C^j = C^{j-1} \cup \{w'\}$. Figure 5 shows a chain complex sampled from seed concept "door". Similar to Vygotsky's "quah" example, it features referent-to-referent extensions that involve different types of relations, illustrating the thought processes that could have given rise to the diverse overextension patterns attested in young children.

While exploratory in nature, our simulation demonstrates the potential of a multimodal approach to capture the formation of chain complexes in child overextension. Future work should explore this generative aspect of the framework in more rigorous terms.

Discussion

We have presented a formal framework for characterizing children's overextension. We have shown that this framework yields good accuracies in reconstructing child-caretaker communication based on a relatively large set of overextension examples we curated from the developmental literature. Our results indicate that the diverse range of overextension patterns can be explained by our framework that encapsulates a multimodal representation of semantic relations with categorical, visual, and predicate-based features.

With respect to earlier work from developmental psychology, our results support the view that children's overextended

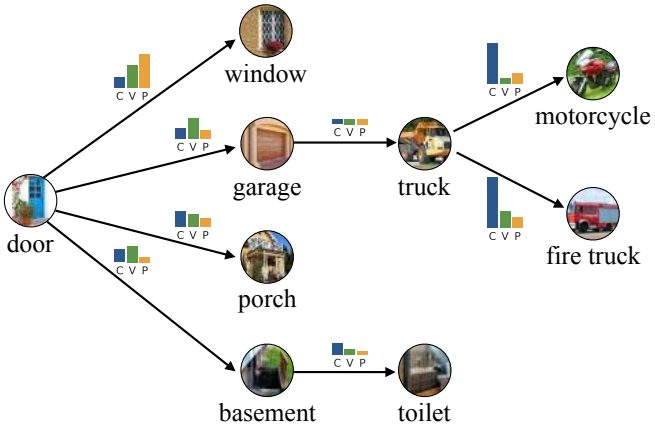


Figure 5: Chain complex sampled from the multimodal semantic space, and contributions of categorical (C), visual (V), and predicate-based (P) relations to chaining probabilities.

word choices reflect a communicative strategy under a limited vocabulary. Moreover, we have shown that children tend to favour high-frequency words in overextension, which provides evidence for the retrieval-error view of overextension. Future work should explore whether the current framework can explain overextension in children's language comprehension, as well as account for the later convergence to adult word usage.

We have shown the initial promise of a multimodal representational scheme toward a better characterization of the generative capacity for word sense extension in early childhood. Future work could explore the generality of this framework in accounting for overextension beyond nouns, as well as historical changes of word meaning.

Acknowledgements

We would like to thank Yu B Xia for helping with collection of child overextension data, Charles Kemp for reference to Vygotsky's work, and Suzanne Stevenson for constructive comments on the draft. We are also thankful to the members of the Computational Linguistics group at the University of Toronto for comments on an early version of this work. This research is supported by an NSERC DG grant and a Connaught New Researcher Award to YX.

References

Beekhuizen, B., & Stevenson, S. (2016). Modeling developmental and linguistic relativity effects in color term acquisition. In *CogSci 38*.

Bird, S., & Loper, E. (2004). Nltk: the natural language toolkit. In *ACL 42*.

Bloom, L. (1973). *One word at a time: the use of single word utterances before syntax*. Mouton.

Bloom, L., Hood, L., & Lightbown, P. (1974). Imitation in language development: If, when, and why. *Cognitive Psychology*, *6*(3), 380–420.

Brown, R. (1973). *A first language: The early stages*. Harvard U. Press.

Clark, E. V. (1973). What's in a word? on the child's acquisition of semantics in his first language. In *Cognitive development and acquisition of language*.

Clark, E. V. (1978). Strategies for communicating. *Child Development*, 953–959.

De Deyne, S., Navarro, D. J., Perfors, A., Brysbaert, M., & Storms, G. (2018). The small world of words english word association norms for over 12,000 cue words. *Behavior Research Methods*, 1–20.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *CVPR 2009*.

Fazly, A., Alishahi, A., & Stevenson, S. (2010). A probabilistic computational model of cross-situational word learning. *Cognitive Science*, *34*(6), 1017–1063.

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2017). Wordbank: An open repository for developmental vocabulary data. *J Child Lang*, *44*(3), 677–694.

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychol Sci*, *20*(5), 578–585.

Fremgen, A., & Fay, D. (1980). Overextensions in production and comprehension: A methodological clarification. *J Child Lang*, *7*(1), 205–211.

Gershkoff-Stowe, L. (2001). The course of children's naming errors in early word learning. *Journal of Cognition and Development*, *2*(2), 131–155.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, *20*(11), 818–829.

Huttenlocher, J. (1974). The origins of language comprehension. In *Theories in cognitive psychology: The loyola symposium* (pp. xi, 386–xi, 386).

Kachergis, G., Yu, C., & Shiffrin, R. M. (2017). A bootstrapping model of frequency and context effects in word learning. *Cognitive Science*, *41*(3), 590–622.

Kay, D. A., & Anglin, J. M. (1982). Overextension and underextension in the child's expressive and receptive speech*. *Journal of Child Language*, *9*(1), 83–98.

Lakoff, G. (1987). *Women, fire, and dangerous things: What categories reveal about the mind*. U Chicago Press.

Lazaridou, A., Chrupała, G., Fernández, R., & Baroni, M. (2016). Multimodal semantic learning from child-directed input. In *NAACL-HLT 15*.

MacWhinney, B. (2014). *The childes project: Tools for analyzing talk, volume ii: The database*. Psychology Press.

Mervis, C. B. (1987). Child-basic object categories and early lexical development. In *Concepts and conceptual development: Ecological and intellectual factors in categorization* (pp. 201–233).

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, *38*(11), 39–41.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39.

Ramiro, C., Srinivasan, M., Malt, B. C., & Xu, Y. (2018). Algorithms in the historical emergence of word senses. *Proceedings of the National Academy of Sciences*, *115*(10), 2323–2328.

Rescorla, L. A. (1980). Overextension in early language development. *J Child Lang*, *7*(2), 321–335.

Roy, D. K., & Pentland, A. P. (2002). Learning words from sights and sounds: A computational model. *Cognitive Science*, *26*(1), 113–146.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *ICLR 2015*.

Siskind, J. M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, *61*(1-2), 39–91.

Suppes, P. (1974). The semantics of children's language. *American Psychologist*, *29*(2), 103.

Thomson, J. R., & Chapman, R. S. (1977). Who is daddy revisited: the status of two-year-olds' over-extended words in use and comprehension. *Journal of Child Language*, *4*(3), 359–375.

Vygotsky, L. S. (1962). *Language and thought*. MIT Press.

Wittgenstein, L. (1953). *Philosophical investigations* (G. Anscombe, Trans.). Prentice Hall.

Wu, Z., & Palmer, M. (1994). Verbs semantics and lexical selection. In *ACL 32*.

Xu, Y., Regier, T., & Malt, B. C. (2016). Historical semantic chaining and efficient communication: The case of container names. *Cognitive Science*, *40*(8), 2081–2094.

Yu, C. (2005). The emergence of links between lexical acquisition and object categorization: A computational study. *Connection Science*, *17*(3-4), 381–397.

Do Children Ascribe the Ability to Choose to Humanoid Robots?

Teresa Flanagan (tmf87@cornell.edu)
Department of Human Development, Cornell University, 116 Reservoir Ave
Ithaca, NY 14853 USA

Joshua Rottman (jrottman@fandm.edu)
Department of Psychology, Franklin & Marshall College, P.O. Box 3003
Lancaster, PA 17604 USA

Lauren Howard (lauren.howard@fandm.edu)
Department of Psychology, Franklin & Marshall College, P.O. Box 3003
Lancaster, PA 17604 USA

Abstract

Investigating folk conceptions of choice and constraints have been problematic given that human actions are rarely considered constrained. In this paper, we utilize humanoid robots (more clearly influenced by determined programming) to empirically test children's developing concepts of choice and action. Using a series of agency attribution and choice prediction tasks, we examined whether children differentiate free will abilities between robots and humans. Results indicated that 5–7-year-old children similarly attributed the ability to choose to both a robot and human child. However, for moral scenarios, participants considered the robot's actions to be more constrained than the human. These findings demonstrate that children appear to hold a nuanced understanding of choice across agents and across context.

Keywords: choice attribution; human–robot interaction; free will; cognitive development

Introduction

Attributing another entity with the ability to make choices lies at the foundation of treating that individual as having moral responsibility and as being deserving of rights. As such it is critically important to understand when and why people attribute choice to others. Because it is difficult to imagine typical human actions being constrained or devoid of choice (Nichols, 2011), research on free will has been limited by participants' inabilities to conceive of a deterministic world (Sommers, 2010). In other words, since it is difficult for us to imagine our own actions as being constrained, it is similarly challenging to conceive of others' actions as being constrained. Investigating ideas of free will in non-human entities, such as robots, shows promise in terms of releasing participants from anecdotal notions of choice and constraint.

Modern technology has resulted in a growing presence of interactive robots (e.g., Siri, Alexa, Roombas), particularly in the environments of younger generations (Wei et al., 2011). Robots present an increasingly important category for which to investigate choice attribution, as they are known to be largely programmed by their designers. Previous research has indicated that adults are ambivalent about robots' capacities to make choices (Weisman, Dweck, & Markman, 2017). However, nothing is known about children's tendencies to attribute robots with the freedom to choose, despite children now growing up in technologically rich environments.

From early in life, children understand the possibility of completing "alternative" actions, denoting a basic grasp of free will. For example, 10-month-old infants expect human agents to use different actions to obtain an object depending on whether there are physical constraints present or absent (Brandone & Wellman, 2009), and toddlers use this understanding to differentially respond to agents who could have acted one way, but chose another (Behne et al., 2005; Dunfield & Kuhlmeier, 2010; Hamlin, Wynn, & Bloom, 2008). By the preschool years, children not only anticipate and react to alternative actions (e.g., Nichols, 2004), but are able to verbally generate alternative options when the main goal of an agent is constrained (Sobel, 2004). Thus, early in life, children show a relatively sophisticated understanding that human agents can choose to act in certain ways, and that these actions may be constrained by internal or external barriers.

This ability to entertain alternative actions suggests that children understand that agents can "choose to do otherwise", a hallmark for a mature understanding of free will (see Kushnir, 2018). However, a reliance on experiments involving human agents as targets of judgment means that the boundaries of children's free will ascriptions have not been fully charted. Though some work has shown that children assign more freedom of choice to human agents than inanimate objects (Nichols, 2004), none have explicitly examined attribution of free will to humanoid robots.

As in adults (e.g., Kahn et al., 2012b), research has shown that children ascribe a mixture of animate and inanimate characteristics to humanoid robots, suggesting an ontological category that is functionally separate from either (Kahn et al., 2012a; Severson & Carlson, 2010). For example, children may assume that robots hold a certain level of intelligence and some sensory abilities (e.g., can think, can see, can be tickled), but not emotions or biological capabilities (e.g., can feel happiness, needs sleep, can grow), though these ascriptions vary with both participant age and robot type (e.g., Bernstein & Crowley, 2008; Jipson & Gelman, 2007; Saylor, Somanader, Levin, & Kawamura, 2010). Furthermore, children often require prior information or experience with robots before they consider them as agentive

beings. For example, 18-month-old infants only follow the gaze of a robot they previously saw acting contingently with an adult (Meltzoff et al., 2010), 4- to 7-year-old children are more likely to assume a robot has intelligence if they have more exposure to robots (Bernstein & Crowley, 2008), and 5- to 7-year-old children are more likely to attribute emotional and physical characteristics to a robot that was previously framed as autonomous (Chernyak & Gary, 2016). This work highlights the ways in which robots straddle the animate and inanimate worlds, making them particularly interesting as a test case for children's ascriptions of free will.

Importantly, it appears that children's understanding of free will, even for human agents, is not monolithic, as children seem to struggle with understanding how alternatives can be applied in certain circumstances. For example, 4- to 5-year-old children seem to believe that it is not possible to act against desires even without physical constraints (e.g., wanting to eat a tasty cracker but choosing not to; Kushnir et al., 2015), and often choose to act in accordance with their desires at the expense of reaching a salient goal (Yang & Frye, 2018). Relatedly, 3- to 5-year-old children are likely to say that a choice is more moral if it is consistent with an agent's desires (e.g., cleaning up toys because they wanted to) versus conflicting with an agent's desires (e.g., cleaning up toys even if they wanted to go play outside), a pattern that is reversed in older children and adults (Starmans & Bloom, 2016). As such, there are certain scenarios, particularly those relating to internal desires or moral decisions, that appear to muddle children's understanding of free will for human actors.

In the current study, we asked whether 5- to 7-year-old children's predictions of action and choice varied across target agent (human child or robot) and constraint scenario (No Constraint, Moral Constraint, Rational Constraint). Children in this age range undergo relevant changes in their free will beliefs and their perceptions of robots (Bernstein & Crowley, 2008; Kushnir et al., 2015). During testing, both the human and robot agents were introduced as being similarly likely to make a particular choice when no constraints were present (i.e., was either 'programmed to' or 'born to" play a certain game). Within each scenario, we explored whether children predicted that the agent would follow the typical, default object choice or would respond to the constraints and pick an alternate object.

Based on previous work exploring children's trait attributions to robots and humans (Chernyak & Gary, 2016; Kahn et al., 2012), along with children's differential reactions to context and constraint (Kushnir et al., 2015; Nichols, 2011), we hypothesized the following: Without constraints, participants would predict the default action for both the robot and the human agent, and each would be significantly above chance in this choice. With rational constraints (the default action being impossible to completely fulfill), participants would predict the default action significantly more for the robot than the human agent. In the robot condition, participants would predict the default action above or at chance, and in the human condition, participants would predict the default action significantly below chance. With moral constraints (the default action causing harm), participants would predict the default action significantly more for the robot than the human agent. In the robot condition, participants would predict the default action above or at chance, and in the human condition, participants would predict the default action significantly below chance.

Method

Participants

The final sample consisted of 32 children, aged 5–7 years old (M_{age} = 5.72, SD_{age} = 0.68, 15 females, 26 White), who were recruited from a participant database and tested in a laboratory in a small city in the northeastern region of the United States. One additional child participated but was excluded due to a developmental disability.

Materials & Procedure

Participants were randomly assigned to one of two conditions (robot or human). In the robot condition, children were asked to watch and respond to the actions of a robot figure named Robovie. The robot was a black and white humanoid toy, approximately 35 cm tall. It was preprogrammed to complete a number of actions. In the human condition, participants watched and responded to the actions of a human child. The actor was a boy approximately the same age as participants, named Billy. All stimuli were pre-recorded and presented via video on a Dell laptop so that the agent's actions and perceived agency could be matched across conditions.

Regardless of condition, all participants proceeded through the same paradigm to assess their understanding of free will across ontological kinds. Children watched the video of either the robot or human, during and after which all participants were asked to predict the agent's actions (default or alternative object choice) and asked to attribute choice to the agent (did they "choose to" do the action or not, adapted from Kushnir et al., 2015). Answers to these questions indicated whether children believed the agent could act against its default choice and respond to constraints in a way that indicated free will.

Video Paradigm

Introduction Phase During the introduction phase of the video, participants watched a short clip (60 seconds long) that introduced the agent (robot or human) and showed the agent performing simple actions. The purpose of this introduction was to demonstrate that the agent was autonomous, intentional, and had some basic intelligence, as this has been found necessary for children to attribute agency (Chernyak & Gary, 2016; Meltzoff et al., 2010). The video consisted of a narrator first describing the agent ("This is Robovie, Robovie is a robot."/ "This is Billy. Billy is a kid") paired with a still picture of the agent in a children's room (see Figure 1). Then the agent performed two simple actions: dancing and

throwing a bucket. The next video clip presented the agent's actions as determined, stating that the agent was either programmed (robot condition, "Robovie is programmed to know a lot about science and can play science games") or born (human, "Billy's parents are scientists, so Billy knows a lot about science and plays science games") to play a certain type of game (science games). Following this presentation, it was reiterated that the agent *only* plays the science game, even if there are other games present. This conveyed to participants that the science game, given no other constraints, was the default choice for both agents. This also indicated that a choice to play an alternative game (e.g. a history game), given constraints, would require the agent to "override" their entrenched pattern of playing the default games.

Action Prediction After the introduction video, participants watched three further video segments that presented each of the constraint scenarios (No Constraint, Moral Constraint, Rational Constraint). These segments described the objects in the room (a science game and a history game), presented the relevant constraints on the agent's ability to play these games, and asked the participant to predict which game the agent would play within each of these scenarios. As explained in the introduction video segment, the science game should be the default game if no other constraints are present. The history game is the alternative game choice.

In the first video (No Constraint scenario), participants were asked to pick which of two games (default or alternative) the agent would play without any limitation. Since participants had previously been told that agent only plays the default game, we hypothesized that this question would elicit a default response without the need for inferring choice or free will. The second video (Moral Constraint scenario) was identical to the first, but with the limitation that the agent will be playing with another person and playing the default game would result in hurting that person's feelings. In this video, it would be wrong for the agent to play the default game, thus requiring them to play the alternative game if they wanted to stay within moral bounds. The third video (Rational Constraint scenario) asked the child to predict the game the agent would play if the default game was broken. In this video, it would be irrational to play the broken (unplayable) game, requiring them to play the alternative game in order to act rationally.

Figure 1: Screenshots of video stimuli used in the robot condition. Participants watched a short introduction video and then proceeded to view three Constraint scenario videos. Videos in the human condition were identical, with the exception of a child in place of the robot.

Choice Attribution After each video, the experimenter asked two follow-up questions to explore choice attribution, adapted from Kushnir et al. (2015). Specifically, participants were asked whether the agent "chose to" or "had to" play the default/alternative game, along with an open-ended prompt asking them why.

Coding

Children's action predictions were coded for each of the video constraints. In the No Constraint scenario, picking the alternative game (rather than the default game) clearly indicated choice, as it went against the default pattern of behavior. However, as there was no obvious reason to select the alternative game, the alternative game was not expected in either the human or the robot condition. In the Moral Constraint scenario, picking the alternative game indicated a choice that was driven by a consideration of others' feelings, whereas picking the default game indicated a disregard for others' feelings. Thus, we expected participants would predict the agent to play the alternative game in the human condition, but not in the robot condition. In the Rational Constraint scenario, picking the alternative game indicated a choice that was driven by a rational consideration of which game was possible to play, whereas picking the alternative game indicated a disregard for rational considerations. Therefore, in this scenario we expected participants would predict the agent to play the alternative game in the human condition, but not in the robot condition.

Children's responses to the choice question were coded for whether or not they responded that they agent "had to" or "chose to" play a certain game. The ability to choose was indicated by a response that the agent "chose to" play the game, regardless of which game the agent chose.

Results

Action Prediction Results

Across all three constraint scenarios, action predictions were explored by running an omnibus binomial test to determine whether the percentage of game prediction (default and alternative) differed from chance (50%). The percentage of the default game predicted was marginally lower than chance in the human condition ($p = .059$) and at chance in the robot condition ($p = .665$). A Mann-Whitney test indicated that the default game prediction did not differ by agent condition ($U = 1032$, $p = .301$, $r = .182$). Within each of the three constraint scenarios, action predictions were explored by running binomial tests to determine whether the percentage of game prediction (default and alternative) differed from chance (50%). Further, Mann-Whitney U tests were run to test for

differences between agents (human and robot) in the predicted percentage of default game prediction in each of the constraint scenarios. Frequencies are presented in Figure 2.

Action Prediction: No Constraint We ran binomial tests on participants' prediction of the game that would be played (default or alternative), for both the human and robot agents. Participants overwhelmingly tended to predict that both the human and the robot would play the default game (*ps* < .001). This demonstrates that participants predicted the agent to act in accordance with the ways that it had always acted in the past (i.e., playing the default game that it was programmed or born to play), indicating that participants understood that there was a strong likelihood for both the human and the robot to select the default game and confirming that the participants understood the introduction video. A Mann-Whitney test indicated that the default game prediction did not differ by agent condition ($U = 120$, $p = .317$, $r = .18$).

Action Prediction: Moral Constraint A binomial test indicated that the percentage of the default game predicted in the human condition was significantly lower than chance ($p < .001$). Thus, participants believed the human would go against his desires in order to act morally. In contrast, the percentage of the default game predicted in the robot condition was not significantly different from chance ($p = .804$). This demonstrates that participants were unsure if a robot would go against its programming in order to act morally. A Mann-Whitney test indicated that the prediction of the default game was significantly higher for the robot condition than the human condition ($U = 80$, $p < .05$, $r = .42$).

Action Prediction: Rational Constraint Binomial tests indicated that the percentage of the default game predicted in both the human condition and the robot condition was significantly lower than chance (*ps* < .001). This demonstrates that participants believed that both agents would go against their programming or desires in order to act rationally. A Mann-Whitney test indicated that action prediction did not differ by condition ($U = 120$, $p = .317$, $r = .18$).

Choice Attribution Results

In each scenario, participants were asked whether the agent "chose to" or "had to" play the predicted game, regardless of game type (default or alternative). Across all three constraint scenarios, choice attributions were explored by running an omnibus binomial test to determine whether the percentage of choice attribution ("choose to" and "have to") differed from chance (50%). The percentage of "choose to" responses did not differ from chance in either the human condition or the robot condition (human: $p = .312$; robot: $p = .193$). A Mann-Whitney test indicated that the choice attribution did not differ by agent condition ($U = 1128$, $p = .836$, $r = .037$). Within each of the three constraint scenarios, choice attributions were explored by running binomial tests to determine whether the percentage of choice attribution ("choose to" and "have to") differed from chance (50%). Further, Mann-Whitney U tests were run to test for differences between agents (human and robot) in the predicted percentage of default game prediction in each of the constraint scenarios. Frequencies are presented in Figure 3.

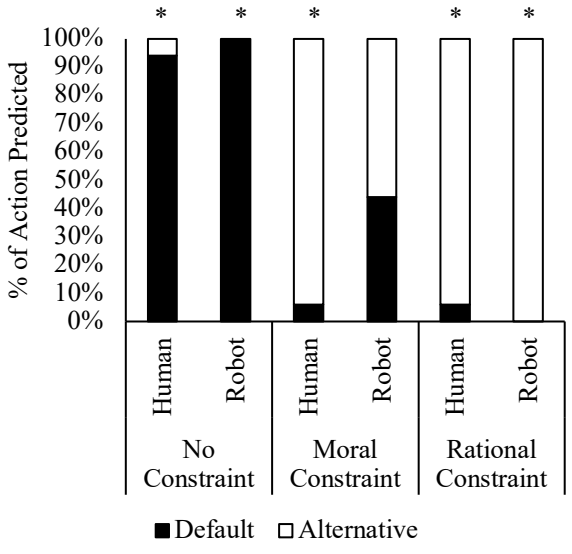


Figure 2: Percentage of the participants' action prediction across conditions for each constraint. Asterisks signify predictions that are significantly different than chance.

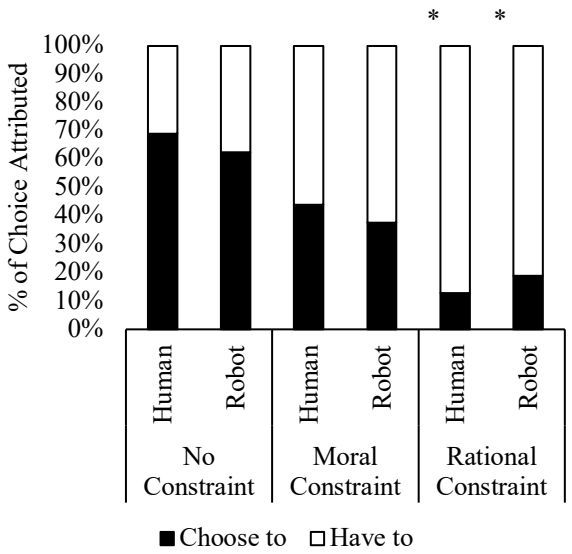


Figure 3: Percentage of participants' attribution of choice across conditions for each constraint. Asterisks signify attributions that are significantly different than chance.

Choice Attribution: No Constraint Binomial tests indicated that the percentage of "choose to" responses in both the human condition and the robot condition did not differ from chance (human: $p = .210$; robot: $p = .454$). A Mann-Whitney test indicated that choice attribution did not differ by condition ($U = 120$, $p = .714$, $r = .065$).

Choice Attribution: Moral Constraint Binomial tests indicated that the percentage of "choose to" responses in both the human condition and the robot condition did not differ from chance (human: $p = .804$; robot: $p = .454$). Mann-Whitney test indicated that choice attribution did not differ by condition ($U = 120$, $p = .723$, $r = .063$).

Choice Attribution: Rational Constraint Binomial tests indicated that the percentage of "have to" responses in both the human condition and the robot condition was significantly higher than chance ($ps < .05$). Mann-Whitney test indicated that choice attribution did not differ by condition ($U = 120$, $p = .632$, $r = .111$). These findings indicate that the participants believed acting in accordance with physical limitations was a constraint on choice regardless of the agent acting, such that free will was significantly limited by this rational constraint.

Discussion

This research investigated children's predictions and choice attributions about the actions of a robot or a human child. Results indicated that, overall, children tended to judge a humanoid robot as having a similar amount of freedom of choice as a human child. For example, without any constraints, or with a rational constraint (such as the default game being broken), participants predicted similar actions for both the human and the robot. However, when the robot had an opportunity to change its actions in order to avoid making another child sad, children judged the robot as less likely to act in this way as compared to the human.

Action Prediction

In the No Constraint scenario, participants significantly predicted the agent to play the default game, in both the human and robot condition. This demonstrates that children's basic understanding of the introduction video, where it was made clear that the agent only plays the science game. In the Rational Constraint scenario, results demonstrated that participants believed that both the robot and the human could act against its programming in order to act rationally (i.e., to play the alternative game). This may be an indication of choice attribution, as children believe a robot can be responsive to reasons and is not entirely constrained by its programming (Fischer, 2006). In the Moral Constraint, however, participants believed the human would be responsive to reasons and act morally (i.e., play the alternative game), but they were at chance in the robot condition. This demonstrates that in moral situations, children were unsure if the robot could go against its programming. This could be explained in a number of ways. Children may have thought that the robot did not "care" about moral reasons or they may have thought that the robot did not have the capacity to recognize moral reasons. The latter explanation could be due to the fact that children did not interact with the robot in this study, making the robot's social capacities were ambiguous. Future research could include various interaction components between the participant and the robot, which might unveil the types of social capacities that are required for a robot to appear responsive to moral reasons.

Choice Attribution

Overall, our results indicate that children's attribution of choice is not unitary across situations. Similar to previous work (Kushnir et al., 2015), it appears that children are sensitive to both the agent type and the context that an agent is presented in when attributing free will. Furthermore, these results suggest that robot programming is not *always* a constraint on freedom of choice for young children. Most importantly, these results demonstrate that children are sensitive to constraints, such that some constraints (e.g. physical impossibility) are more restrictive to an agent's choice than other constraints (e.g. desires, morals), and finding consistent for both the robot and the human agent.

In the No Constraint scenario, we were surprised to see that participants did not attribute choice above chance to the human agent. This may have been due to the presentation in the introduction video. For example, similar to the robot, we presented the human as having an entrenched disposition to play the default game. However, unlike the robot, who was programmed by scientists to choose this game, the human's "programming" was that his parents were scientists and that he played science games every day. Previous research has shown that child participants attribute this type of consistent (non-random) choice as denoting not just 'programming', but desires (Kushnir, Xu, & Wellman, 2010). Since participants were at chance in attributing choice to the human agent, this could mean that children varied in believing whether or not having a strong desire is a constraint on actions.

In contrast to the human condition, we did not anticipate that participants would attribute so *much* choice to the humanoid robot in the No Constraint scenario. Here, the percentage of children that said the robot "chose to" select the default game was not significantly different from chance, suggesting that approximately half of our participants gave some semblance of free choice to the robot agent. This could be due to the varied exposure children have to robots, which research has shown is correlated with children's propensity to attribute agency (Bernstein & Crowley, 2008). Alternatively, this could also be an indication of children's general understanding of choice under minimal constraints. Specifically, children varied in the amount of choice they thought an agent had if the agent was constrained to perform a certain action based on how the agent was programmed or raised. Future research should investigate what underlies this individual difference; for example, do children who have

more interactions with robots also attribute them more choice in unconstrained scenarios?

In the Moral Constraint scenario, children were at chance in attributing choice in both the human condition and robot condition. This demonstrates that participants were almost equally likely to say that a moral action was a constraint or a choice. These results dovetail with previous research that has shown that children under 7-years-old are less likely to say that people have a choice in moral actions in comparison to older children (Chernyak et al., 2013). For the Rational Constraint, results also followed previous research, such that children assumed the agent "had to" make a certain choice when there were physical constraints (Kushnir et al., 2015), regardless of agent type. While these results support previous work, they also extend previous findings to other agents. Since results were similar for the human condition and the robot condition, this demonstrates that children extend their existing beliefs of choice and constraint to a fundamentally different type of agent. Future research should investigate other agents, such as plants or animals, to see if choice attribution in light of moral and rational constraints is a general or agent specific attribution.

It is important to note that we don't fully understand the way children were interpreting some of our events, particularly those relating to programming ("Robovie is programmed to know a lot about science") or genetic inheritance ("Billy's parents are scientists so he knows a lot about science"). However, previous research has shown that, starting at 5-years-old, children display knowledge of biological inheritance (Gimenez & Harris, 2002) and children understand a robot is programmed (Bernstein & Crowley, 2008). Furthermore, all participants were told multiple times that the agent (whether robot or human) only played the default game. Future research should explicitly measure children's understanding of programming and inheritance in relation to their action prediction and choice attribution to an agent.

In sum, this research indicates that children are able to attribute choice to actions that are programmed or hard-wired. This suggests that "compatibilist" theories of free will – in which choice can be said to exist even in a fully determined universe (e.g., Fischer, 2006) – may be an intuitive aspect of folk psychology. This dovetails with previous studies that have advanced this claim, but which have based their conclusions on adults' assessments of complicated thought experiments (e.g., Nahmias et al., 2005). The present research has shown this to be true in children and for an everyday case of pre-determinism. New advances in technology, which have introduced robots into children's everyday environments, have not only improved quality of living but have also allowed for improvements in testing for folk attributions of choice to agents that straightforwardly exist in a constrained environment.

Acknowledgments

We would like to thank the Nissley Scholar Grant at Franklin & Marshall College for funding this project. We would like to thank Julia McAleer for her work as a research assistant. Also, we thank the families from the Lancaster community for participating in this project.

References

Behne, T., Carpenter, M., Call, J., & Tomasello, M. (2005). Unwilling versus unable: Infants' understanding of intentional action. *Developmental Psychology, 41*, 328-337.

Bernstein, D., & Crowley, K. (2008). Searching for signs of intelligent life: An investigation of young children's beliefs about robot intelligence. *The Journal of the Learning Sciences, 17*(2), 225–247.

Brandone, A. C, & Wellman, H. M. (2009). You can't always get what you want: Infants understand failed goal-directed actions. *Psychological Science, 20*(1), 85-91.

Chernyak, N., & Gary, H. E. (2016). Children's cognitive and behavioral reactions to an autonomous versus controlled social robot dog. *Early Education and Development, 27*(8), 1175–1189.

Chernyak, N., Kushnir, T., Sullivan, K. M., & Wang, Q. (2013). A comparison of American and Nepalese children's concepts of freedom of choice and social constraint. *Cognitive Science, 37*(7), 1343–1355.

Dunfield, K. A., & Kuhlmeier, V. A. (2010). Intention-mediated selective helping in infancy. *Psychological Science, 21*, 523-527.

Fischer, J. M. (2006). Responsiveness and moral responsibility. *My Way: Essays on Moral Responsibility*. New York, NY: Oxford University Press, 63-83.

Gimenez, M., & Harris, P. (2002). Understanding constraints on inheritance: Evidence for biological thinking in early childhood. *British Journal of Developmental Psychology, 20*, 307-324.

Hamlin, J. K., Wynn, K., & Bloom, P. (2008). Social evaluation by preverbal infants. *Nature, 450*, 557-559.

Jipson, J. L., & Gelman, S. A. (2007). Robots and rodents: Children's inferences about living and nonliving kinds. *Child Development, 78*(6), 1675–1688.

Kahn, P. H., Kanda, T., Ishiguro H., Freier, N. G., Severson, R. L, Gill, B. T., Rucket, J. H., & Shen, S. (2012a). "Robovie, you'll have to go into the closet now": Children's social and moral relationship with a humanoid robot. *Developmental Psychology, 48*(2), 303-314.

Kahn, P. H., Severson, R. L., Kanda, T., Ishiguro, H., Gill, B. T., Ruckert, J. H., … Freier, N. G. (2012b). Do people hold a humanoid robot morally accountable for the harm it causes? In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction - HRI '12* (p. 33). New York, New York, USA: ACM Press.

Kushnir, T. (2018). The developmental and cultural psychology of free will. *Philosophy Compass, 13*(11), 1-17.

Kushnir, T., Gopnik, A., Chernyak, N., Seiver, E., & Wellman, H. M. (2015). Developing intuitions about free will between ages four and six. *Cognition, 138*, 79-101.

Meltzoff, A., Brooks, R., Shon, A., & Rao, R. (2010). "Social" robots are psychological agents for infants: A test of gaze following. *Neural Networks, 23*(8-9), 966-972.

Nahmias, E., Morris, S., Nadelhoffer, T., & Turner, J. (2005). Surveying freedom: Folk intuitions about free will and moral responsibility. *Philosophical Psychology, 18*(5), 561–584.

Nichols, S. (2004). The folk psychology of free will: Fits and starts. *Mind & Language, 19*(5), 473–502.

Nichols, S. (2011). Experimental philosophy and the problem of free will. *Science, 331*, 1401-1403.

Saylor, M. M., Somanader, M., Levin, D. T., & Kawamura, K. (2010). How do young children deal with hybrids of living and non-living things: The case of humanoid robots. *British Journal of Developmental Psychology, 28*(4), 835-851.

Severson, R. L., & Carlson, S. M. (2010). Behaving as or behaving as if? Children's conceptions of personified robots and the emergence of a new ontological category. *Neural Networks, 23*(8–9), 1099–1103.

Sobel, D. M. (2004). Exploring the coherence of young children's explanatory abilities: Evidence from generating counterfactuals. *British Journal of Developmental Psychology, 22*, 37-58.

Sommers, T. (2010). Experimental philosophy and free will. *Philosophy Compass, 5*(2), 199–212.

Starmans, C., & Bloom, P. (2016). When the spirit is willing, but the flesh is weak: Developmental differences in judgments about inner moral conflict. *Psychological Science, 27*(11), 1498–1506.

Wei, C. W., Hung, I. C., Lee, L., & Chen, N. S. (2011). A joyful classroom learning system with robot learning companion for children to learn mathematics multiplication. *Turkish Online Journal of Educational Technology, 10*(2), 11–23.

Weisman, K., Dweck, C. S., & Markman, E. M. (2017). Rethinking people's conceptions of mental life. *Proceedings of the National Academy of Sciences*, 201704347.

Yang, F., & Frye, D. (2018). When preferences are in the way: Children's predictions of goal-directed behaviors. *Developmental Psychology, 54*(6), 1051-1062.

Children, more than adults, rely on similarity to access multiple meanings of words

Sammy Floyd (sfloyd@princeton.edu)
Department of Psychology, Princeton University, Princeton NJ 08544
Casey Lew-Williams (caseylw@princeton.edu)
Department of Psychology, Princeton University, Princeton NJ 08544
Adele E. Goldberg (adele@princeton.edu)
Department of Psychology, Princeton University, Princeton NJ 08544

Abstract

Past research has shown that adults can access multiple meanings for a word, but little work has examined how children process multiple meanings. We tested 48 4- to 7-year-old children and 48 adults in a touchscreen picture recognition task. Two meanings of the same word were displayed on successive trials, which varied according to whether the 2 meanings were unrelated (homonyms), related (polysemes), or repeated (same-meaning). Adults identified the second meaning more quickly than the first in all conditions and to the same extent. Children, however, identified the second meaning more quickly only on polysemy and same-meaning trials. This difference suggests that children are less capable of co-activating unrelated meanings, which raises the possibility that children must *learn* to do so over development. Despite the ubiquity of polysemy in language, our work is the first to show that children's processing of word representations is organized by similarity.

Keywords: polysemy, lexical processing, development, cognitive development, ambiguity

Introduction

Upon hearing a word like *bat,* which can refer to a flying mammal or a wooden stick, adults unconsciously activate both meanings, at least for a brief period when there is no biasing context and both meanings are equally frequent (Brocher, Koenig, Mauner, & Foraker, 2017; Onifer & Swinney, 1981; Swinney, 1979; Zwitserlood, 1989). This intriguing finding was initially used to argue for "exhaustive" lexical access during an early modular stage of processing (Fodor, 1985; Swinney, Plather, & Love 2000; but cf. Armstrong & Plaut, 2016). At the same time, a large and growing body of evidence indicates that people take advantage of communicative contexts to predict interpretation from the earliest stages of comprehension (Kintsch, 1988; Rubio-Fernandez, Mollica & Jara-Ettinger, 2018; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995; Yip & Zhai, 2018).

Due to the tension between evidence for exhaustive lexical access, on the one hand, and early contextual influences on the other, much work has varied task demands, relative frequencies of the two meanings, interstimulus intervals and degrees of contextual bias in order to predict the conditions under which multiple meanings of a word are accessed or recognized. Selective, rather than exhaustive, activation has been found to occur when the context is strongly biased toward a more frequent meaning (Meyer & Federmeier 2007; Sereno, Brewer, & O'Donnell 2003; Sereno, Pacht, & Rayner 1992; Simpson, 1981; Marslen-Wilson & Welsh, 1978).

Another factor that plays a role in lexical access and recognition is the degree of relatedness among a word's meanings. That is, there is a gradient distinction between meanings that are **homonymous** or unrelated to one another (e.g., a flying bat vs. baseball bat), and **polysemous** meanings, which are semantically related to varying degrees (Tuggy, 1993). For instance, the word *network* can be used to refer to a TV channel, a group of colleagues, or a graph (Lau, Cook, McCarthy, Gella, & Baldwin, 2014). While these meanings are distinct, they are to some extent related.

Relationships among meanings are relevant to the so-called "ambiguity advantage": Adults have been found to respond faster in lexical decision tasks to a meaning for an ambiguous word compared to a word with a single meaning (Jastrzembski, 1981; Rubenstein, Garfield, & Millikan, 1970). This effect has sometimes been found to be stronger for words with multiple *related* senses (Klepousniotou & Baum, 2007; Rodd, Gaskell, & Marslen-Wilson, 2002). In fact, Armstrong & Plaut (2008) and Rodd, Gaskell & Marslen-Wilson (2002) found that homonymous senses can *slow down* lexical access due to competition under higher of levels task difficulty (see also Brocher et al. 2016). Other evidence that ambiguous words compete in a way that polysemous meanings may not comes from an ERP study by Klepousniotou et al. (2012), who found a greater N400 was evoked by a less dominant meaning of homonymous words in a lexical-decision task, but no increase in the N400 for the less dominant meaning of polysemous words. On the other hand, Brocher et al.

(2017) found that both homonymous and polysemous meanings compete when words were equally biased toward both meanings. Thus, the work on access and recognition of ambiguous words has revealed a complicated picture, indicating that frequency, degree of contextual bias, timing, task demands, and semantic relatedness each influence lexical activation (Tabossi & Sbisá, 2001).

In order to clarify key influences on lexical access, the current work compares the behavior of children and adults on an identical task. A word repetition paradigm is used to detect whether witnessing one meaning of a word primes a second meaning of the word. Specifically, in a 2-alternative forced-choice picture identification task, adults and 4- to 7-year-old children were exposed to a word on each trial, and had to select which of two images corresponded to that word's meaning. On the immediately following trial, the same word was presented again. Across these key trials, the degree of relatedness between the first and second target meanings of words was systematically varied.

Of interest was whether reaction times decreased between the identification of the first and second meanings of words. If we do see priming effects for both homonymous and polysemous word meanings, it would be evidence that the two meanings are linked as is required for exhaustive access. This is expected in adults, at least if the time between trials is sufficiently brief. At longer inter-stimulus intervals (ISIs), we might expect the first meaning to interfere with the second meaning, which would predict an increase in reaction times to the second meaning.

If both children and adults display the same increase or decrease in reaction time when identifying the second meanings of ambiguous words, it would suggest that key aspects of lexical access are a developmentally stable. We know that children, like adults, comprehend language incrementally (Swingley, Pinto, and Fernald 1999; Fernald, Swingley, & Pinto 2001). Also, children, like adults, are subject to priming and plausibility effects when they need to disambiguate an intended meaning (Rabagliati, Pylkkänen & Marcus 2013). But we don't yet know whether children and adults will behave alike or differently under the identical task demands that require them to identify two familiar meanings of words in succession.

A significant difference between children and adults' behavior could shed light on the mechanisms involved in lexical access or on the way that lexical representations develop. If children show *stronger* evidence of exhaustive lexical access for ambiguous and polysemous words, it would be consistent with proposals that view selective access as requiring cognitive control (Balota, Cortese, & Wenke, 2001), since children's cognitive control is less well developed than adults (Bunge, et al. 2002). On the other hand, if children show *weaker* evidence of accessing multiple familiar meanings of words, it would suggest that they represent individual meanings more independently than adults do. This would suggest that word learning involves both acquisition of item-specific knowledge for each meaning *and* a protracted trajectory for linking among each word's meanings. This would indicate that children have to *learn* to co-activate multiple meanings based on experience, with potentially different trajectories for related versus unrelated meanings.

Some past work has investigated how children over the age of 8 activate the intended meaning of homonymous words, by focusing on cases of homonymy in which one meaning was dominant over others (Marmurek & Rossi, 1993; Simpson & Forster, 1986; Simpson et al. 1994). This research found relatively consistent results: older children are better at using contextual cues to activate less frequent homonymous meanings than younger children. Booth, Harasaki & Burman (2006) extended this work by comparing effects of sentence-level primes vs. lexical primes and found a more complex picture. Younger children or less skilled readers were less likely than older children to use a preceding lexeme to facilitate activation of a less-frequent homonymous meaning, while older children/high skilled readers facilitated and inhibited homonymous meanings using sentence-level information (Booth, Harasaki & Burman 2006).

The present work uses participants' reaction times to investigate how words with multiple meanings are processed in children and adults, when both meanings need to be identified in succession. By comparing performance on homonymous and polysemous meanings with a baseline condition, we can determine whether greater semantic similarity supports the co-activation of lexical representations. This would be evident if participants are faster to recognize the second meanings of polysemous words than homonymous words.

In the experiments reported below, we children and adults were presented with each of 18 target words twice in immediate succession: 6 words were paired with 2 unrelated meanings (homonymy condition); 6 words with 2 related meanings (polysemy condition); and 6 words were presented with different images which represented the same meaning (same-meaning trials). We also included 12 singleton filler trials to reduce the extent to which participants could rely on a repetition expectation to predict what they might hear and see next. For homonymous and polysemous trials, each word was presented with one target meaning on first exposure and a different target meaning on the second

Trial type	Word
6 same-sense repeated	*bowl, treehouse, ring, key, lantern, shelf, trunk*
6 polysemous senses	*cap, buttons, cone, glasses, shower, step*
6 homonymous meanings	*bow, ruler, pitcher, bat, calf, nail*
12 singletons	*basket*, *cake, crayon, feather, hood*, *lemon*, *ivy, log, playground, punch, wagon, bark*

Table 1: Items

exposure, with the order of target meanings counterbalanced across participants and sides of presentation randomized. Repeat same-meaning trials served as a window into baseline priming effects. The main prediction was that, if children's word representations are organized by similarity, they may be able to activate a second, distinct meaning quickly in the case of polysemy. In the case of homonymy, however, where semantic similarity is not available to help co-activate other meanings which share the same label, accessing both meanings should be slower. In the same paradigm, we also predict that adults should be able to activate multiple meanings equally well in both polysemy and homonymy, consistent both with prior findings and with the idea that we learn to access unrelated meanings, at least in certain contexts, through experience. Finally, adult participation allows us to verify that it is, in fact, possible for our paradigm and chosen items to elicit priming of unrelated meanings.

Method

Participants

48 adult participants [recruited online] and 48 children ages 4.5-7 (M=5.89; SD= 0.62). Children were given a book of their choosing and a small prize as thank-you gifts.

Procedure

Two initial training trials provided feedback if participants answered incorrectly, or took longer than 4500ms, ensuring that they understood the goal of the task was to answer accurately and quickly (see Figure 1). Between each trial (including between training trials), a pulsing blue dot appeared that participants had to press to advance to the next trial. This was to ensure that participants' hand positions were centered. Each participant responded to 48 trials including 6 homonym pairs (12 trials), 6 polysemy pairs (12 trials), 6 same-sense pairs (12 trials) and 12 singleton filler trials.

The design was 3 (condition) x 2 (1^{st} or 2^{nd} encountered meaning), within-subjects. We tested two groups (adults and children). Before each trial began, participants had to place their pointer finger on a dot in the middle of the screen. Overall order of stimuli (or stimuli pairs) from each of the four trial types (polysemy, homonymy, same-sense pairs, and fillers) was randomized across participants. The experiment was conducted on an iPad that recordedd participants' accuracy and reaction times to target. The key dependent measure was the difference in reaction time from the identification of first and second senses of words in the three experimental conditions.

On each trial, participants heard a word and had to choose the target image representing its meaning from a distractor image presented on the opposite side of the screen (screen side counterbalanced). For homonymous, polysemous and same-sense trials, the same word was repeated twice in succession with images corresponding to a second unrelated, related, or same sense, respectively.

The order of presentation within each word's pair of meanings as well as the order of trials in the experiment was counterbalanced across participants to avoid possible confounds of meaning familiarity or distractor salience. Moreover, since participants witnessed both ambiguous and polysemous trials, any significant difference in familiarity between the ambiguous items and the polysemous items should be evident in a comparison of response times to the first presentations across these conditions, which we also include as part of our analyses.

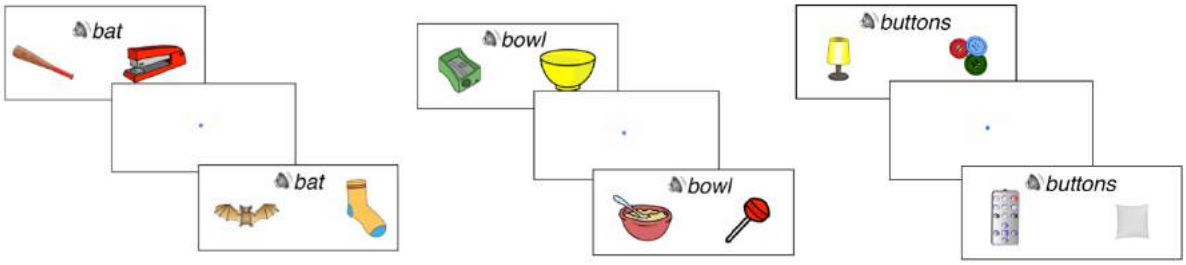


Figure 1: Representation of the experimental stimuli in the homonymy (*bat*), repeat (*bowl*) and polysemy (*buttons*) conditions.

Results

The data were log-transformed and analyzed using a multilevel linear model with condition (homonymy vs. polysemy vs. same meaning) and first vs. second meaning of each pair as fixed effects, and maximal converging random effect structure: here, random intercepts and slopes for subjects, presentation order, and items: Reaction Time ~ FirstOrSecond * Condition + (1 + FirstOrSecond | subject) + (1 + Condition | order) + (1 + FirstOrSecond | item).

Adults recognized the second sense of words more quickly after the initial exposure to that word, and facilitation was equally strong for unrelated (homonymy), related (polysemy), and same senses: (main effect of secondary sense response, β = -0.15515, p = 0.00129, with no significant interactions by condition (Figure 2).

Children, on the other hand, did not show significant facilitation when selecting the second sense of homonymous words, but did for polysemous words (β = -0.12798, p = 0.0333) and repeated meanings (β = -0.200424, p = 0.0122) (Figure 2). The difference between facilitation for polysemous and same-sense trials was not significantly different (β = -0.08638, p = 0.2189), suggesting that related senses were primed by one another to almost the same degree as a second instance of the same sense. Unlike results for adults, there was not even a numerical decrease in reaction time when the second presentation of a word was paired with an unrelated (homonymous) sense.

A concern worth addressing is whether children were less familiar with the meanings of the homonymous words. Indeed, we cannot expect facilitation for a second meaning if only one meaning was familiar to children. To ensure this did not account for our results, we excluded any trials in which children or adults had answered incorrectly on either trial for all analyses reported thus far. This issue can be further addressed by a comparison of accuracy in the polysemy vs. homonymy condition. We found that their accuracy was not significantly lower in homonymy than polysemy in a linear model with maximal converging random slopes and intercepts for subject and order (β -0.03, t= -1.410, p= 0.172), and neither were their reaction times slower to the first exposure in homonymy as compared to polysemy (β= 0.075, t= 0.844, p= 0.405). Since the order of presentation of the two meanings was counterbalanced across participants for each word, we can conclude that children were equally familiar with the senses of the homonymous, polysemous, and same-sense meanings, as intended.

Limitations

In our task, answers appeared on either side of the screen. In order to control for hand/mouse position effects, intervals between each trial required participants to press a central fixation, and the experiment did not advance to the next trial until participants did so. Because of this, inter-stimulus intervals (ISIs) were not controlled, and instead were determined by how long the participant took to press the central fixation. Importantly, prior work has shown that second senses of homonymous words become suppressed as quickly as a few syllables downstream, and early work in semantic priming did not reveal effects for priming across more than one intervening trial (Joordens & Besner, 1992), suggesting that we should not expect to see priming in the case of longer inter-stimulus intervals. Therefore, the ISIs observed in our experiment warrant further investigation.

To address this concern, we report average ISIs for the two groups, as well as a comparison of the two (adults: M = 1961ms, SD = 5196ms, children: M = 1147ms, SD = 701ms). We then entered log-transformed ISI lengths into a mixed effect model with age group (child vs. adult) as the fixed effect and maximal converging random structure including a random intercept and slope for subject and intercept for trial number (order), revealing no effect of the age group (child): β =-0.09643, p = 0.4). So, while average ISIs were longer than those used in traditional priming experiments, it is unlikely that the difference between

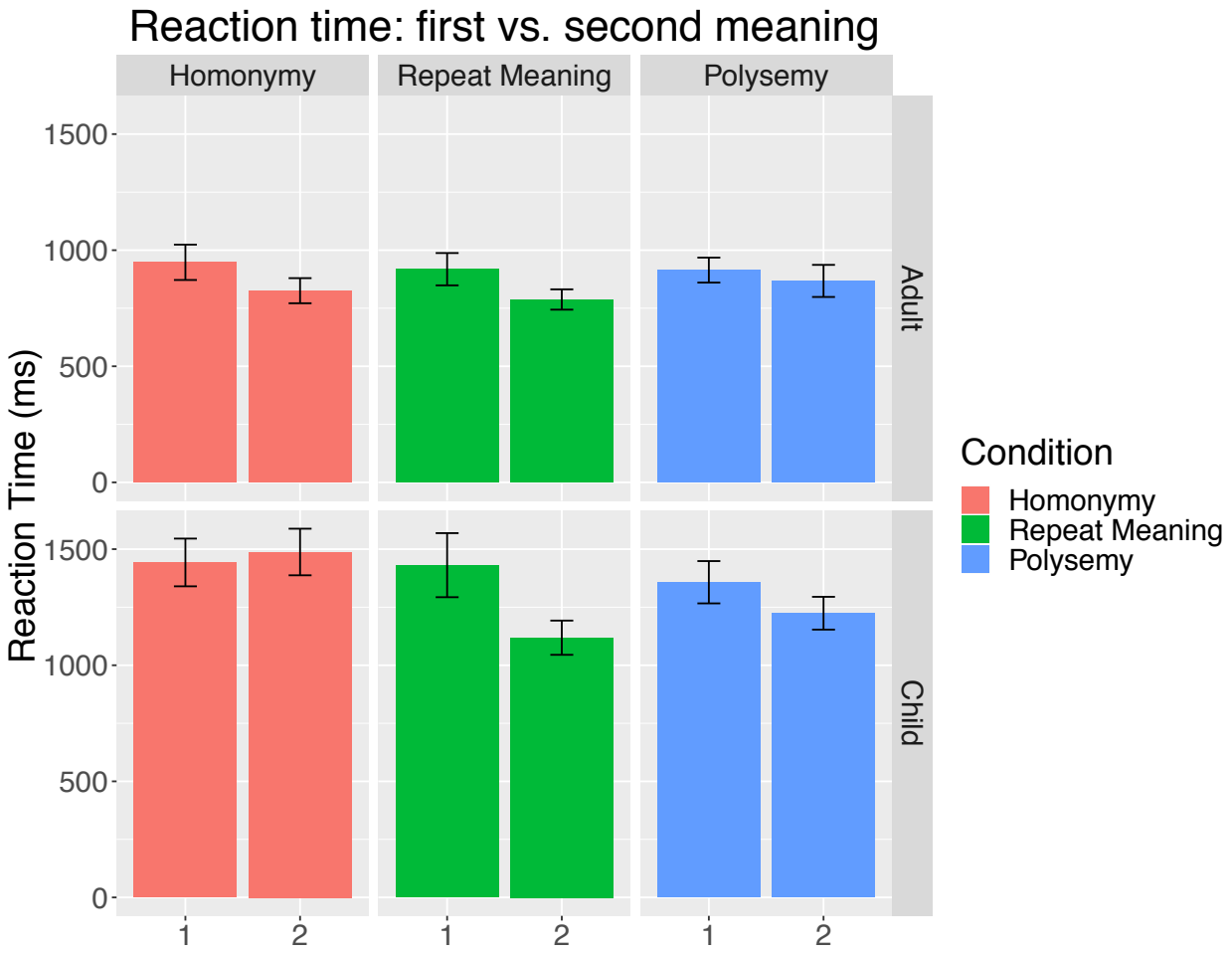


Figure 2: Raw reaction time data (analyses were done on log-transformed data). Error bars represent bootstrapped 95% confidence intervals.

ISIs is what drove our children to perform differently than adults. Past work with children has also used longer ISIs with children, such as 1,000ms (Booth, Harasaki & Burman, 2006), as compared to ISIs in adult lexical decision tasks.

Consistent with our results, Armstrong & Plaut (2016) emphasize that the timing of adult participants' suppression of irrelevant senses varies by task difficulty as well as latencies. Later work on semantic priming has in fact shown evidence for longer-term priming, across as many as 8 intervening items (Joordens & Becker, 1997). The implication of this work on our predictions instead suggests that participants may be expected to benefit from priming over longer periods of time.

Discussion and Conclusion

This investigation is the first, to our knowledge, to compare children's and adults' co-activation of related word meanings. Prior work has found that under certain conditions, adults access more than one meaning of a word, at least for a short period of time, unless one meaning is both more frequent and anticipated within the context. In the current study, the facilitation evident in adults' response times to second meanings demonstrates that, regardless of relatedness, adults are capable of accessing two meanings simultaneously or are at least able to anticipate a second meaning. To emphasize, adults displayed faster reaction times to a second meaning even when that meaning was entirely unrelated to the first (e.g., baseball *bat* following mammal *bat*).

Children, on the other hand, showed facilitation only when the second meaning was related or identical to the first. They showed no evidence that the recognition of one sense of a word facilitated the recognition of an unrelated meaning of that word. We addressed the possibility that children were less familiar with the meanings of the homonyms by observing that their

accuracy and response times on the first exposure of each word-type were not different. The current findings thus indicate that children's representations of a word's two unrelated meanings may not be linked together in the same way that adults' are. Instead, while children showed a facilitation effect in the recognition of a second *related* meaning, unrelated meanings were recognized as slowly as completely new words.

Given this, it may be that children must *learn* to activate multiple homonymous meanings across time. Intuitively, this makes sense: a spreading of activation between the mental representation of "bottle cap" and of "pen cap" may be a natural consequence of shared or similar features, while the representations of "baseball bat" and "flying bat" are likely to overlap very little, if at all. Yet again, ultimately speakers do eventually learn to access both meanings, at least or a brief period under certain task demands, as demonstrated by evidence co-activation both in our task and in previous work with adults (Brocher, Koenig, Mauner, & Foraker, 2017; Onifer & Swinney, 1981; Swinney, 1979; Zwitserlood, 1989). This raises the question as to *why* and *how* the ability to access unrelated meanings of a word develops.

Insofar as listeners cannot reliably predict which meaning of a word is intended, a degree of flexibility is advantageous in language processing to avoid being essentially garden-pathed by an unintended meaning. Indeed this type of flexibility may be advantageous in language learning as well, insofar as a more efficient ability to update predictions has been found to correlate with larger vocabulary size (Reuter, Emberson, Romberg, & Lew-Williams, 2018).

We can only speculate as to exactly *how* this ability to access secondary unrelated meanings of words increases after the age of 7. But presumably either links between two distinct representations are created or the representations of homonymous meanings come to share greater overlap. Stronger links between unrelated senses of homonymous words may be formed as a result of repeated misinterpretations that require learners to access an alternative sense as quickly as possible for the sake of comprehension. Alternatively, it is possible that links between meanings of homonymous words are formed on the basis of more explicit, metalinguistic knowledge. It is possible that co-activation is facilitated simply by an awareness that labels can refer to multiple meanings. On this interpretation, the information that the word *bat* as two unrelated meanings would be similar to learning that the word, *aunt* can be pronounced in two distinct ways.

A non-mutually exclusive possibility is that co-activation may be encouraged by learning to read. Specifically, a shared written form in combination with a shared auditory label can be expected lead to an increase in representational overlap between two meanings of a homonymous word. This would support the idea that representational overlap is required for co-activation. Future work can test this by comparing words that vary in whether they are spelled alike compared to words that are not (*bat vs. bat; flower vs. flour*). If the link between unrelated meanings is mediated via a shared visual form, we expect facilitation for homonyms that share the same spelling but not for homonyms that are spelled distinctly.

Armstrong, B. C., & Plaut, D. C. (2016). Semantic Ambiguity Effects in Lexical Processing: A Neural-Network Account Based on Semantic Settling Dynamics. *Unpublished Manuscript*.

Balota, D. A., Cortese, M. J., & Wenke, D. (2001). Ambiguity resolution as a function of reading skill, age, dementia, and schizophrenia: The role of attentional control. *On the Consequences of Meaning Selection: Perspectives on Resolving Lexical Ambiguity*, 87–102.

Beretta, A., Fiorentino, R., & Poeppel, D. (2005). The effects of homonymy and polysemy on lexical access: an MEG study. *Cognitive Brain Research*, *24*(1), 57–65. https://doi.org/10.1016/j.cogbrainres.2004.12.006

Booth JR, Harasaki Y& Burman, DD (2006). Development of lexical and sentence level context effects for dominant and subordinate word meanings of homonyms. Journal of Psycholinguistic Research, 35, 531-554.

Brocher, A., Foraker, S., & Koenig, J. P. (2016). Processing of irre- gular polysemes in sentence reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 42, 1798–1813. doi:10.1037/xlm0000271 Brysbaert,

Brocher, A., Koenig, J., Mauner, G., & Foraker, S. (2017). About sharing and commitment : the retrieval of biased and balanced irregular polysemes. *Language, Cognition and Neuroscience*, *0*(0), 1–24. https://doi.org/10.1080/23273798.2017.1381748

Bunge, S. A., Dudukovic, N. M., Thomason, M. E., Vaidya, C. J., & Gabrieli, J. D. (2002). Immature frontal lobe contributions to cognitive control in children: evidence from fMRI. *Neuron*, *33*(2), 301-311.

Davies, M. (2008). The corpus of contemporary American English (COCA): 385 million words, 1990-present. *On-line], Available: http://www. americancorpus. org*.

Fodor, J. A. (1985). Precis of the modularity of mind. *Behavioral and Brain Sciences*, *8*(1), 1–5.

Jastrzembski, J. E. (1981). Multiple meanings, number of related meanings, frequency of occurrence, and the lexicon. *Cognitive Psychology*, *13*(2), 278–305.

Joordens, S., & Besner, D. (1992). Priming effects that span an intervening unrelated word: Implications for models of memory representation and retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*(3), 483-491.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, *95*(2), 163.

Klepousniotou, E., & Baum, S. (2007). Clarifying further the ambiguity advantage effect in word recognition: Effects of aging and left-hemisphere damage on the processing of homonymy and polysemy. *Brain and Language*, *103*(1–2), 148–149. https://doi.org/10.1016/j.bandl.2007.07.089

Klepousniotou, E., Pike, G. B., Steinhauer, K., & Gracco, V. (2012). Not all ambiguous words are created equal: An EEG investigation of homonymy and polysemy. *Brain and language*, *123*(1), 11-21.

Lau, J. H., Cook, P., Mccarthy, D., Gella, S., & Baldwin, T. (2014). Learning Word Sense Distributions , Detecting Unattested Senses and Identifying Novel Senses Using Topic Models, 259–270.

Marslen-Wilson, W. D., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, *10*(1), 29–63.

Marmurek, HC & Rossi, M. (1993). The Development of Strategic Processing of Ambiguous Words: Riddles Versus Neutral Context. *Journal of Genetic Psychology - J GENET PSYCHOL. 154*. 475-486. 10.1080/00221325.1993.9914746.

Onifer, W., & Swinney, D. A. (1981). Accessing lexical ambiguities during sentence comprehension: Effects of frequency of meaning and contextual bias. *Memory & Cognition*, *9*(3), 225–236.

Pylkkänen, L., Llinás, R., & Murphy, G. L. (n.d.). *The Representation of Polysemy: MEG Evidence*.

Reuter, T., Emberson, L., Romberg, A., & Lew-Williams, C. (2018). Individual differences in nonverbal prediction and vocabulary size in infancy. *Cognition*, *176*, 215–219.

Rodd, J., Gaskell, G., & Marslen-Wilson, W. (2002). Making sense of semantic ambiguity: semantic competition in lexical access. *Journal of Memory and Language*, *46*, 245–266 ST–Making sense of semantic ambiguity:

Rubenstein, H., Garfield, L., & Millikan, J. A. (1970). Homographic entries in the internal lexicon. *Journal of Memory and Language*, *9*(5), 487.

Rubio-Fernandez, P., Mollica, F.& Jara-Ettinger, J. (2018). Why searching for a blue triangle is different in English than in Spanish.

Simpson, G. B. (1981). Meaning dominance and semantic context in the processing of lexical ambiguity. *Journal of Memory and Language*, *20*(1), 120.

Simpson, G. B., & Foster, M. R. (1986). Lexical ambiguity and children's word recognition. *Developmental Psychology, 22*(2), 147-154.

Simpson, G. B., & Adamopoulos, A. C. (2001). Repeated homographs in word and sentence contexts: Multiple processing of multiple meanings. *On the Consequences of Meaning Selection: Perspectives on Resolving Lexical Ambiguity*, 157–173.

Simpson, G. B., Krueger, M. A., Kang, H., & Elofson, A. C. (1994). Sentence context and meaning frequency effects in children's processing of ambiguous words. *Journal of Research in Reading, 17*(1), 62–72.

Swinney, D. A. (1979). Lexical access during sentence comprehension:(Re) consideration of context effects. *Journal of Verbal Learning and Verbal Behavior*, *18*(6), 645–659.

Swinney, D., Prather, P., & Love, T. (2000). The time course of Lexical Access and the role of context: converging evidence from normal and phasic processing. In *Language and the Brain: Representation and processing* (pp. 273–292). New York: Academic Press.

Tabossi, P., & Sbisá, S. (2001). Methodological issues in the study of lexical ambiguity resolution.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*(5217), 1632–1634.

Tuggy, D. (1993) Ambiguity, polysemy, and vagueness. *Cognitive Linguistics 4*.3: 273-90.

Yip, M. C. W., & Zhai, M. (2018). Context Effects and Spoken Word Recognition of Chinese: An Eye-Tracking Study. *Cognitive Science*, *42*, 1134–1153.

Zwitserlood, P. (1989). The locus of the effects of sentential-semantic context in spoken-word processing. *Cognition*, *32*(1), 25–64.

Metaphors we teach by: A method for mapping metaphorical lay theories

Stephen J. Flusberg (stephen.flusberg@purchase.edu)
SUNY Purchase College, Department of Psychology
735 Anderson Hill Road, Purchase, NY 10577, USA

Bridgette Martin Hard (bridgette.hard@duke.edu)
Duke University, Department of Psychology
417 Chapel Drive, Durham, NC 27708, USA

Abstract

People frequently use metaphors to communicate and reason about complex topics. However, many studies of metaphorical reasoning exclusively rely on researcher intuitions about different metaphors and their associated entailments. Here we describe a more principled method for mapping the structure of metaphorical lay theories, focusing on metaphors for teaching. Across two studies, we identified four common, apt metaphors for the teacher-student relationship and used factor analysis to explore whether these metaphors reflect systematically different intuitions about the qualities of college teachers. Our findings demonstrate that (1) people endorse a variety of different teaching metaphors, and (2) these metaphors bring to mind distinct, coherent clusters of teacher attributes. This work demonstrates a novel method for systematically mapping the structure of metaphorical lay theories and sets the stage for future research on metaphorical reasoning as well as innovative educational interventions centered on shifting lay theories of teaching.

Keywords: metaphors, lay theories, concepts, teaching

Introduction

According to a popular cognitive science metaphor, people are amateur scientists who actively explore the environment and develop intuitive theories for how the world works (Furnham, 1988; Gopnik, Meltzoff, & Kuhl, 1999). In turn, these intuitive, lay theories help people make sense of and respond to new experiences, guiding thought and action. For example, research over the last two decades suggests that students' lay theories about the malleability of intelligence (also known as *mindsets*) drive educational achievement (Yeager & Dweck, 2012). Students who think of intelligence as something that can improve through hard work—the *brain-as-muscle* metaphor—react better to performance setbacks, resulting in superior long-term learning outcomes (but see Sisk et al., 2018).

As we have just illustrated, one way that intuitive beliefs are conveyed is via metaphor. To take another example, when we describe a teacher as "*molding impressionable* students," we imply that the teacher is like a *sculptor* and students are like *clay*. Metaphors allow people to draw on familiar, common knowledge of a basic source domain (*building muscles; sculpting clay*) to communicate about a more complex or abstract target domain (*how brains learn; teaching*; Lakoff & Johnson, 1980). A large body of research finds that the metaphors people use to talk about complex issues both reflect and shape how they think about those topics (for review, see Thibodeau, Hendricks, & Boroditsky, 2017). The *teacher-as-sculptor* metaphor, for example, may reflect the intuitive belief that learning is passive and that the teacher (not the student) largely determines the learning outcomes.

How can we best understand the structure of these metaphorical lay theories and the extent to which they influence thought and action? An important first step is to map out the *entailments* of the metaphors; that is, the associated ideas and inferences licensed by the metaphorical comparison. For example, one potential entailment of the *brain-as-muscle* metaphor described earlier is that, while hard work may increase someone's intelligence or abilities, working too long on any one task may be cognitively exhausting (much like continuous physical exertion tires out muscles). These sorts of entailments, it is argued, provide critical insight into the mental model people use to represent the target domain and allow for empirically-informed predictions about how metaphors reflect and shape thinking.

The most common approach to mapping entailments is to examine the figurative language people use in everyday speech and apply a commonsense understanding of the observed source domains (as we have demonstrated in the preceding paragraph). Cognitive linguists have used this approach to isolate the structural schemas that underlie many fundamental concepts, from emotion to politics to time, across a variety of languages and cultures (Kövecses, 2005; Lakoff and Johnson, 1980).

However, there are several theoretical problems with relying on intuition and patterns of language alone to make inferences about underlying conceptual representations (Keysar & Bly, 1995; Casasanto, 2009; Murphy, 1996). For example, the meaning of common metaphorical expressions might seem obvious and intuitive but could also reflect a post-hoc rationalization based on one's preexisting understanding of the expression (Keysar & Bly, 1995). Indeed, some (non-linguistic) experiments have shown that metaphors in language do not always reflect how people mentally represent a given topic (see Casasanto, 2009). How, then, can researchers more reliably map the structural entailments of metaphorical concepts?

Another approach is to ask a set of naïve participants to freely generate the structural entailments of a source domain in order to derive a set of conceptually coherent metaphorical

entailments for the target domain. For example, for the metaphor "crime is a *virus,"* participants might be asked how they would solve a *literal* virus problem in their city. The responses would then be used to predict which solutions to a city's crime problem would be conceptually congruent with the *crime-as-virus* metaphor (Thibodeau & Boroditsky, 2011). To validate these experimenter intuitions, another set of participants might be asked to match specific solutions to a crime problem to specific metaphors (i.e., match a solution to a city's crime problem to either a *crime-as-virus* or *crime-as-beast* metaphor; Thibodeau & Boroditsky, 2013).

In this paper, we build on this approach and offer a systematic method for mapping the structural entailments of a complex metaphor, focusing specifically on metaphors for teaching. Our approach was inspired in part by traditional psychometric methods that have been used to uncover the dimensional structure of personality traits and other psychological constructs.

In Study 1, we first identified six common metaphors for the college teacher-student relationship, drawing on the literature in teacher education (e.g., Patchen & Crawford, 2011; Shaw, Berry, & Mahlios, 2008). We assessed the relative aptness of these metaphors by asking participants to rate their agreement with each metaphor, select the one they liked best, and explain their choice. In Study 2, we presented a new set of participants with one of the four most popular metaphors from Study 1. With this one metaphor in mind, participants rated the degree to which a wide range of statements describing the attributes of college teachers fit with the metaphor (as well as rating each metaphor on a few additional features). We used exploratory factor analysis to identify a smaller set of latent factors underlying the larger collection of teacher attributes. This allowed us to identify a distinct, coherent cluster of teacher attributes associated with each metaphor, providing a principled way of mapping the structure of metaphorical lay theories.

Study 1

Methods

Participants We recruited 119 participants to complete the survey through Amazon's Mechanical Turk. We required that participants be a current or former college student living in the U.S. or Canada, with an approval rating greater than 95% on at least 100 prior Turk tasks. We excluded data from nine participants: two that came from duplicate IP addresses, four that provided duplicated (i.e., copy and paste) responses to all free response items, and two who reported that they had never attended college.

Of the 110 participants included in the final data set, 54% were male, 83% identified solely as White, 7% as Black, 2% as Asian, 3% as Hispanic/Latino, and the remaining 4% as multiracial. About 71% had graduated from college, 15% were currently enrolled students, and 15% had attended college at one point but were not currently enrolled. About 43% had attended college less than 4 years ago, and 57% had been out of college for more than 4 years. Mean age was 35 (SD = 11) with a range of 20-74.

Materials & Procedure We designed a survey using Qualtrics online survey software in which participants considered six possible metaphors for the teacher-student relationship (see Table 1). We derived these metaphors based on qualitative findings of the metaphors that teachers use to describe their roles (e.g., Patchen & Crawford, 2011; Shaw, Berry, & Mahlios, 2008). Each metaphor described both the teacher and the student and suggested a relationship between them (e.g. "*A teacher is like a sculptor and students are like clay*").

Participants viewed all six metaphors in a random order and each was described as a metaphor for college teaching. Participants first rated their agreement with the metaphor on a 6-point scale (strongly disagree to strongly agree, with no neutral midpoint) and then freely explained their response by typing in a text box. After considering all six metaphors, participants considered the entire collection of metaphors and selected their favorite. They explained why they preferred this metaphor and, specifically, how it fit their experiences and views of college teaching. Finally, participants answered a series of basic demographic questions.

Results

What are the most popular metaphors? As shown in Table 1, participants leaned toward agreement (M > 3.5) for all the metaphors except the *app store* metaphor, but they agreed the most with the *gardener* metaphor, followed closely by the *coach*, *tour guide*, and *sculptor* metaphors. Participants' selection of their preferred metaphors showed a similar pattern of popularity. Of the six metaphors, the most popular was the *gardener* metaphor (32.7%), followed by the *tour guide* (19.1%) and *coach* metaphors (18.2%). The *sculptor* (10.9%), app store (9.1%) and ship captain metaphors (8.2%) were less favored.

An initial examination of participants' justifications for the metaphors shows that they frequently extended the metaphors in their free response descriptions. For example, the *gardener* metaphor prompted descriptions of teachers as "sowing information," "planting seeds of knowledge," and "cultivating students." The *coach* metaphor elicited descriptions of teachers working with students "toward the same goal," helping students "win and succeed," giving students "exercises," and creating "a plan of attack." The *tour guide* metaphor led to descriptions of teachers "showing students around", taking them into "the unknown," "showing the way," helping students "navigate" and taking them along "the path of learning." The *sculptor* metaphor prompted descriptions of teaching as "molding," "shaping," and being "hands-on," and described students as "impressionable," "raw material," and "undefined" but becoming "polished." This suggests that participants were actively using the metaphor to reason about the qualities of a teacher associated with a given metaphor (Thibodeau, Crow, & Flusberg, 2017).

Table 1. Rank-ordered ratings of agreement with each of the teaching metaphors on a 6-point scale. Higher numbers indicate higher levels of agreement.

Metaphor	*M*	*SD*
*"A teacher is like a **gardener** and students are like plants"*	4.36	1.29
*"A teacher is like a **coach** and students are like athletes"*	4.28	1.21
*"A teacher is like a **tour guide** and students are like tourists"*	4.19	1.31
*"A teacher is like a **sculptor** and students are like clay"*	4.15	1.25
*"A teacher is like the **captain** of a ship and students are like sailors"*	3.79	1.40
*"A teacher is like an **app store** and students are like smartphone users"*	2.95	1.56

Discussion

The findings from this initial study suggest that people with college experience endorse a variety of metaphors for the teacher-student relationship, but the most apt are the *gardener*, *coach*, *tour guide*, and *sculptor* metaphors. That participants spontaneously extended these metaphors in articulating their personal preferences provides some evidence that they were *thinking* about the nature of college teaching in terms of the metaphorical lay theory.

The primary goal of Study 2 was to identify and map out the conceptual entailments associated with each of the four most popular metaphors for the teacher-student relationship; that is, the associated beliefs and expectations that logically follow from the metaphor. Identifying the entailments of various metaphors is critical for making informed predictions about how metaphors may shape beliefs, attitudes, and behaviors. As a starting point, we examined entailments that focused on the characteristics of *teachers* (as opposed to students). Our key question was whether different metaphors would be reliably associated with distinct clusters of teacher attributes, and whether this can be measured in a systematic, principled way.

Study 2

Methods

Participants We recruited 201 participants to complete the survey through Amazon's Mechanical Turk using the same exclusion criteria as in Study 1. We excluded data from two participants who provided duplicated (i.e., copy and paste) responses to all free response items.

Of the 199 participants included in the final data set, 54% were male, 75.6% identified solely as White, 9% as Black, .5% as American Indian or Alaskan Native, 5% as Asian, 2.5% as Hispanic/Latino, and the remaining 7% as multiracial. About 70% had graduated from college, 13.4% were currently enrolled students, and 17% had attended college at one point but were not currently enrolled. Only 39% were recent college students who had attended college less than 4 years ago, and 61% had been out of college for more than 4 years. Mean age was 36 (*SD* = 11.6) with a range of 18-76.

Materials & Procedure Participants were randomly assigned to view one of the four most popular metaphors from Study 1: the *gardener*, *coach*, *tour guide*, and *sculptor* metaphors. Participants then viewed a list of 43 statements describing college teachers (e.g., "*Teachers transfer their knowledge to students*") and rated how well each item fit the metaphor that they were given (see Table 2). The statements were generated by consulting measures of teacher behavior (e.g., teacher behavior checklist, Keeley, Smith, & Buskist, 2006), examining free response data from Study 1, and the personal experience of the researchers). Participants were specifically instructed to rate how well each item agreed with the metaphor they received, not whether they personally believed each item was true.

Next, participants viewed all four of the metaphors and selected their personal favorite. Finally, participants answered four questions aimed whether the metaphors captured beliefs about teacher responsibility and power: They rated, according to the metaphor, (a) how much responsibility college teachers have for students' learning, (b) how much responsibility students have for their own learning in college, (c) how much power college teachers have to influence what students learn, and (d) how much power college teachers have to influence how students to develop as people. Each item was rated on a scale of 0 to 100, with 0 meaning "none at all" and 100 meaning "a great deal."

Results

Factor structure of teacher characteristic. To begin, we performed a principal axes factor analysis on the 43 different teacher attributes (oblimin rotation) to identify the latent variables underlying the attributes. Based on the eigenvalues, as well as the ability to meaningfully interpret the clustered attributes, a 7-factor solution provided the best fit for the data. The eigenvalues for the first seven factors were: 21.81, 3.02, 2.59, 1.63, 1.25, 1.00, and .82, with the eighth being .75. The rotated pattern matrix is shown in the Appendix. Based on the highest-loading items, we interpreted the factors as:

(1) **Community-building** (e.g., "Teachers encourage a sense of community")
(2) **Knowledgeable** (e.g., "Teachers are knowledgeable about their subject matter)
(3) **Authoritative** (e.g., "Teachers establish classroom rules")
(4) **Influencing** (e.g., "Teachers powerfully influence their students")
(5) **Philosophical** (e.g., "Teachers are abstract thinkers" and "Teachers provoke debate")
(6) **Informing** (e.g., "Teachers present information")
(7) **Nurturing** (e.g., "Teachers are sensitive to their students' needs")

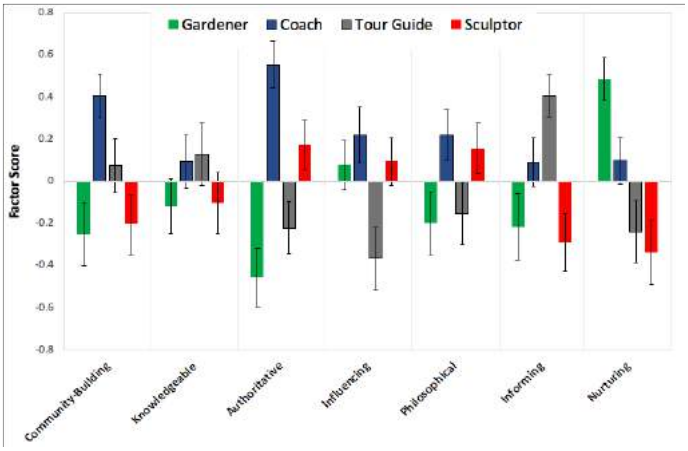


Figure 1: Mean factor scores for each of the seven factors, as a function of the conceptual metaphor that participants considered. The factor scores for the authoritative and nurturing factors are reversed for ease of interpretation.

Figure 1 shows the pattern of differences across the metaphors for each factor. Visual examination of the figure shows some striking differences in teacher characteristics across metaphors. The *gardener* metaphor entails that teachers are nurturing but low on other characteristics, especially authoritative traits. The *coach* metaphor entails that teachers build community, are authoritative, influencing, and even philosophical, in the sense of stimulating new knowledge and provoking debate. The *tour guide* metaphor entails that teachers are informing, but low on other characteristics, especially their ability to influence students. The *sculptor* metaphor entails that teachers are neither nurturing nor informing, but are somewhat authoritative, influencing, and philosophical.

How do teacher characteristics vary across metaphors? Once a 7-factor solution was applied, we saved the factor scores using the regression method. A multivariate ANOVA was performed on the seven factor scores with metaphor (gardener, coach, tour guide, or sculptor) as a between-subjects variable. Overall, metaphor condition showed a significant effect on the set of factors, $F(21, 552) = 6.98$, $p < .001$, $\eta^2 = .21$. Univariate analyses of variance revealed a significant effect of metaphor on the *community-building* ($F(3,188) = 5.01$, $p < .01$) *authoritative* ($F(3,188) = 12.72$, $p < .001$), *influencing* ($F(3,188) = 4.02$, $p < .01$), *informing* ($F(3,188) = 6.20$, $p < .001$), and *nurturing* factors ($F(3,188) = 8.23$, $p < .001$). There was only a marginal effect of metaphor on the *philosophical* factor ($F(3,188) = 2.43$, $p = .07$) and no effect on the *knowledgeable* factor ($F(3,188) < 1$).

Are different metaphors associated with different beliefs about responsibility? Metaphor had a small but reliable effect on ratings of how much responsibility college teachers had for students' learning, $F(3, 197) = 2.93$, $p < .05$, $\eta^2 = .04$. No pairwise comparisons across the conditions (using a Bonferroni adjustment) were significant. As shown in Figure 2, the *sculptor* metaphor promoted the highest rating of teacher responsibility, and the *coach* metaphor the least. Metaphor had a more dramatic effect on ratings of how much responsibility students had for their own learning, $F(3, 197) = 11.42$, $p < .001$, $\eta^2 = .15$. Pairwise comparisons indicated that the *coach* metaphor promoted the highest ratings of student responsibility, significantly more than the *gardener* or *sculptor* metaphors. As shown in Figure 2, the *sculptor* and *gardener* metaphors promoted the lowest ratings, significantly lower than the *coach* and *tour guide* metaphors, but not different from one another.

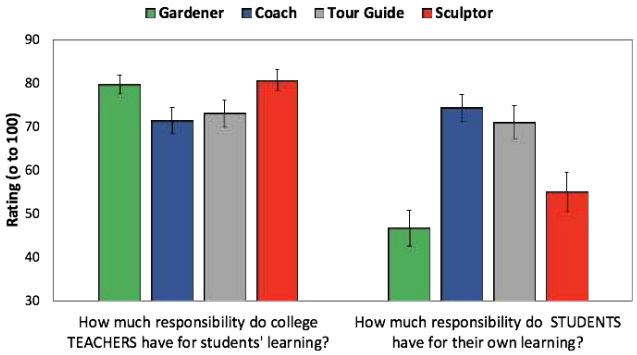


Figure 2. Effects of metaphor on ratings of how much responsibility teachers and students have for learning.

Are different metaphors associated with different beliefs about teacher's power? Metaphor condition had a small but reliable effect on ratings of how much power college teachers had to influence what students learn, $F(3, 197) = 3.02$, $p < .05$, $\eta^2 = .04$. No pairwise comparisons across the conditions were significant, however (all pairwise comparisons used a Bonferroni adjustment). As shown in Figure 3, the sculptor metaphor promoted the highest rating of teacher power over student learning, and the tour guide the least. Metaphor had a more dramatic effect on ratings of how much power college teachers had to influence how students develop as people, $F(3, 197) = 7.44$, $p < .001$, $\eta^2 = .10$. Pairwise comparisons indicated that the *sculptor* metaphor prompted the highest ratings of power to influence development, significantly more than either the *coach* or *tour guide* metaphor. The *gardener* metaphor did not significantly differ from the others.

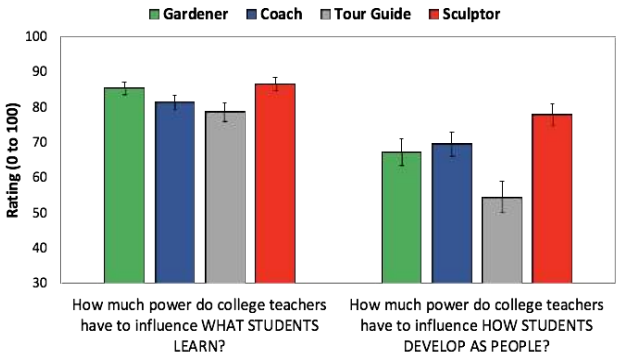


Figure 3. Effects of metaphor on ratings of how much power teachers have to influence what students learn and how they develop as people.

What are the most popular metaphors? Participants most preferred metaphors replicated the pattern found in Study 1. Of the four metaphors, the most popular was the *gardener* metaphor (26.9%), followed by the *coach* (25.9%) and *tour guide* metaphors (25.4%). The *sculptor* metaphor was the least favored (20.4%). Participants' preferred metaphor was not influenced by condition. That is, participants did not select a favorite metaphor that simply matched what they had been presented with in earlier questions. This suggests that people have stable preferred metaphors for thinking about the teacher-student relationship.

Discussion

The findings from this study suggest that different metaphors for teaching are associated with distinct entailments, specifically with respect to the characteristics of teachers. Different metaphors imply that teachers have different "profiles" or clusters of characteristics that can differ dramatically from one another. Notably, the metaphors differ in their implications for students' responsibility for their own learning. The *coach* and *tour guide* metaphors hold students more responsible for their own learning than do the *gardener* and *sculptor* metaphors. The metaphors also differ in how much power the teacher has to influence a student's general development. The *gardener* metaphor entails the most power to influence students' development, and the *tour guide* metaphor the least.

General Discussion

People use metaphors to express their lay beliefs about everything from the nature of intelligence to how the economy works. Though it seems quite natural to identify the conceptual entailments of such metaphors based on common sense knowledge, there are issues with theorizing about the structure of people's metaphorical lay theories based on patterns in language alone. In this paper, we aimed to provide a more principled method for mapping the structure of metaphorical lay theories, using metaphors for teaching as a case study. Our approach was inspired in part by psychometric methods used to uncover the latent structure of other theoretical psychological constructs.

In our first study, we used participant ratings to identify four common, apt metaphors for the college teacher-student relationship (*gardener*, *coach*, *sculptor*, and *tour guide*). In Study 2, participants were provided with one of these metaphors and rated the extent to which a large set of teacher attributes conceptually cohered with the metaphor. We then used exploratory factor analysis to uncover a small subset of meaningful dimensions underlying the larger set of teacher attributes. This revealed that our four teaching metaphors reflect systematically different intuitions about the qualities of college teachers, which can be captured by distinct, coherent clusters of teacher attributes. We contend that this method offers a useful, principled way for researchers interested in metaphorical reasoning to empirically derive the conceptual entailments of different metaphors.

In ongoing and future work, we are continuing to validate this approach by (1) replicating our findings in more representative samples of current college students, (2) applying the same factor analysis method to ratings of *student*, rather than teacher, attributes and (3) mapping the structure of these metaphors in the context of other types of teaching settings (e.g., high school or elementary school teaching). We also plan to measure the explanatory power of lay theories of teaching by examining whether the particular metaphor a student holds for the teacher-student relationship predicts their own attitudes and behaviors in the classroom. Based on the results of Study 2, for example, we would hypothesize that students who hold a *coach* metaphor should expect teachers to be more demanding and assertive than students who hold a *gardener* metaphor, and thus may expect higher and stricter standards in the classroom. Similarly, because the *gardener* and *sculptor* metaphors imply less responsibility on the part of the student, students who endorse these metaphors may hold a more passive view of the learning process and be less likely to engage in active-learning strategies (e.g., self-quizzing) than students who endorse the *coach* and *tour guide* metaphors.

Ultimately, this work could lay the foundation for novel educational interventions based around metaphor framing. Some studies have found that metaphors can shape student mindsets (Blackwell et al., 2007) and attitudes (Landau et al., 2014), but this work has not examined metaphors for teaching specifically. In addition, interventions aimed at changing intuitive lay theories of intelligence to improve student performance have generated inconsistent results (Sisk et al., 2018). One reason for this inconsistency may be that academic performance is shaped by *numerous* intuitive beliefs, not just about intelligence. It is possible that, because learning occurs in a social context that includes teachers and the broader classroom environment, the intuitive beliefs students hold about the nature of *teaching* will also influence academic behaviors and outcomes. As our current work offers a principled way to understand the structure of people's intuitive beliefs about teaching, it provides an important first step in developing such interventions.

References

Blackwell, L. S., Trzesniewski, K. H., & Dweck, C. S. (2007). Implicit theories of intelligence predict achievement across an adolescent transition: A longitudinal study and an intervention. *Child Development, 78*(1), 246-263.

Furnham, A. (1988). *Lay theories: Everyday understanding of problems in the social sciences*. Oxford, England: Pergamon Press.

Gopnik, A., Meltzoff, A. N., & Kuhl, P. K. (1999). *The scientist in the crib: Minds, brains, and how children learn*. William Morrow & Co.

Keeley, J., Smith, D., & Buskist, W. (2006). The Teacher behaviors checklist: Factor analysis of its utility for evaluating teaching. *Teaching of Psych*, *33*(2), 84–91.

Keysar, B., & Bly, B. (1995). Intuitions of the transparency of idioms: Can one keep a secret by spilling the beans? *Journal of Memory and Language*, *34*, 89-109.

Kövecses, Z. (2005). *Metaphor in culture: Universality and variation*. Cambridge University Press.

Landau, M. J., Oyserman, D., Keefer, L. A., & Smith, G. C. (2014). The college journey and academic engagement: How metaphor use enhances identity-based motivation. *Journal of Personality and Social Psychology, 106*(5), 679

Murphy, G. L. (1996). On metaphoric representation. *Cognition*, *60*, 173-204.

Patchen, T., & Crawford, T. (2011). From gardeners to tour guides: The epistemological struggle revealed in teacher-generated metaphors of teaching. *Journal of Teacher Education*, *62*(3), 286-298.

Shaw, D. M., Barry, A., & Mahlios, M. (2008). Preservice teachers' metaphors of teaching in relation to literacy beliefs. *Teachers and Teaching: theory and practice*, *14*(1), 35-50.

Sisk, V. F., Burgoyne, A. P., Sun, J., Butler, J. L., & Macnamara, B. N. (2018). To what extent and under which circumstances are growth mindsets important to academic achievement? Two meta-analyses. *Psychological Science, 29*(4), 549- 571.

Thibodeau, P. H., & Boroditsky, L. (2011). Metaphors we think with: The role of metaphor in reasoning. *PloS one*, *6*(2), e16782.

Thibodeau, P. H., & Boroditsky, L. (2013). Natural language metaphors covertly influence reasoning. *PloS one*, *8*(1), e52961

Thibodeau, P. H., Crow, L., & Flusberg, S. J. (2017). The metaphor police: A case study of the role of metaphor in explanation. *Psychonomic Bulletin & Review*, *24*(5), 1375-1386.

Thibodeau, P. H., Hendricks, R. K., & Boroditsky, L. (2017). How linguistic metaphor scaffolds reasoning. *Trends in Cognitive Sciences, 21*, 852-863.

Yeager, D. S., & Dweck, C. S. (2012). Mindsets that promote resilience: When students believe that personal characteristics can be developed. *Educational Psychologist, 47(4),* 302-314

Appendix

Pattern matrix with the loadings of each of the teacher characteristics on the 7 factors. Highest loading items for each factor (>.39) are marked in color and bolded.

Teachers...	Factor 1 *Community-Building* 50% of common variance	Factor 2 *Knowledgeable* 6.3% of common variance	Factor 3 *Authoritative* 5.3% of common variance	Factor 4 *Influencing* 3.0% of common variance	Factor 5 *Philosophical* 2.2% of common variance	Factor 6 *Informing* 1.6% of common variance	Factor 7 *Nurturing* 1.1% of common variance
encourage a sense of community	**.695**	.047	-.040	.058	.008	-.050	-.258
encourage students to get to know one another	**.557**	-.085	-.175	.041	.141	.024	-.223
are personable	**.554**	.127	.023	-.097	.166	.192	-.174
hold students' attention	**.515**	.203	-.083	.130	.061	.146	.136
encourage teamwork	**.485**	-.118	-.261	.136	.102	.088	-.239
encourage students to support one another	**.484**	-.172	-.281	.027	.110	.029	-.354
encourage students to actively participate in class	.311	.029	-.207	-.007	.214	.298	-.164
are knowledgeable about their subject matter	.014	**.668**	-.008	.049	.096	.293	.040
are experts in their field	.141	**.628**	-.035	.228	-.021	.126	.080
know what they are talking about	.210	**.626**	.022	.083	-.017	.123	-.100
know in advance what they are trying to accomplish	-.059	**.619**	-.217	.149	.003	-.010	-.164
are intelligent	.015	**.531**	.022	.004	**.483**	.026	-.106
work toward a clear goal	-.057	**.495**	-.179	.234	-.007	-.043	-.303
are prepared for class	-.051	**.481**	-.154	.001	.082	.323	-.205
are confident	.268	.376	-.262	.065	.259	.021	.206
establish classroom rules	.054	-.057	**-.648**	.049	.180	.218	-.076
are authority figures	.187	.366	**-.532**	-.041	.063	-.074	.160
command respect	.401	.129	**-.489**	.054	.134	-.032	.239
challenge students	.024	-.091	-.373	.272	.319	.188	-.092
powerfully influence their students	-.038	.100	-.140	**.632**	.025	-.120	-.032
stimulate students' thinking	.110	-.003	.311	**.630**	.267	.115	-.056
transfer their knowledge to students	-.028	.289	.036	**.399**	-.097	.379	-.008
motivate students to put effort into learning	.126	-.064	-.174	.376	.004	.307	-.247
give helpful feedback	.231	-.126	-.185	.307	.137	.280	-.176
are abstract thinkers	.027	.052	-.034	.017	**.756**	-.082	-.062
provoke debate	.045	-.074	-.018	.080	**.704**	.236	.011
question students' ideas	-.018	-.127	-.242	.104	**.625**	.139	.000
are creative	.168	.279	.013	.178	**.447**	-.207	-.167
promote class discussion	.297	-.018	.000	.066	**.440**	.342	-.041
have clear expectations for students	.070	.226	-.286	.187	.388	-.063	-.015
answer students' questions	.103	.106	.001	-.024	.125	**.600**	-.105
present information	-.082	.296	-.108	.028	.074	**.564**	-.001
communicate clearly	.268	.074	-.060	.089	.105	**.462**	-.094
engage students in conversation	.283	-.003	.001	-.012	.357	**.393**	-.109
are sensitive to their students' needs	.039	-.016	.066	-.040	.071	.052	**-.806**
adapt their teaching to different students' needs	.021	-.016	-.062	.174	.032	.085	**-.677**
care about students' well-being	.102	.001	.058	.314	-.012	-.099	**-.652**
listen to their students	.202	.107	.084	-.095	.160	.224	**-.544**
are understanding	.284	.025	.001	-.068	.285	-.001	**-.538**
are available when their students need help	.186	.214	-.038	.092	-.136	.212	**-.504**
put a lot of effort into teaching	-.182	.313	-.235	.214	.155	-.055	**-.456**
get to know their students	.248	.160	.025	.095	.209	-.027	**-.411**
create a positive classroom environment	.230	.254	.038	.089	-.057	.267	-.388

Phoneme learning is influenced by the taxonomic organization of the semantic referents

Abdellah Fourtassi (afourtas@stanford.edu)
Department of Psychology, Stanford University, USA

Emmanuel Dupoux (emmanuel.dupoux@gmail.com)
ENS/CNRS/EHESS/INRIA/PSL Research University, France

Abstract

Word learning relies on the ability to master the sound contrasts that are phonemic (i.e., signal meaning difference) in a given language. Though the timeline of phoneme development has been studied extensively over the past few decades, the mechanism of this development is poorly understood. Previous work has shown that human learners rely on referential information to differentiate similar sounds, but largely ignored the problem of taxonomic ambiguity at the semantic level (two different objects may be described by one or two words depending on how abstract the meaning intended by the speaker is). In this study, we varied the taxonomic distance of pairs of objects and tested how adult learners judged the phonemic status of the sound contrast associated with each of these pairs. We found that judgments were sensitive to gradients in the taxonomic structure, suggesting that learners use probabilistic information at the semantic level to optimize the accuracy of their judgements at the phonological level. The findings provide evidence for an interaction between phonological learning and meaning generalization, raising important questions about how these two important processes of language acquisition are related.

Keywords: language acquisition; phonological development; word learning; speech perception.

A crucial part of language acquisition is the mastery of the sound inventory, i.e., the set of atomic sounds of which words are made. The sound inventory is language-specific. English speakers, for instance, have to learn the distinction between the sounds /l/ and /r/ to differentiate minimal pairs such as *glass* and *grass*. In contrast, Japanese learners need not differentiate these sounds, which do not bring about difference in word meaning in their language. Crucially, even within the same language, learners have to distinguish the sounds that contrast word meaning (phonemic contrasts) from the sounds that do not (non-phonemic contrasts). For example, the aspirated and unaspirated versions of /p/ (which occur, respectively, in the first segment of the word *pin*, and the second segment of the word *spin*) belong to the same phonemic category. Another example, is the *cot-caught* merger whereby the vowels [ɑ]-[ɔ] have come to be treated by some English speakers as non-phonemic variations of the same sounds (Labov, 1991).

How do people learn when a sound contrast is phonemic and when it is a phonetic variation of the same sound category? Children start to show sensitivity to their native sounds at a very early age (e.g., Werker & Tees, 1984). Throughout development, they also learn to distinguish the subset of the native sounds that cue meaning (Dietrich, Swingley, & Werker, 2007; Seidl, Cristi, Onishi, & Bernard, 2009; Kazanina, Phillips, & Idsardi, 2006). These developmental facts have been documented in detail over the past few decades, but the mechanism of this learning is still poorly understood.

Most research has focused on exploring mechanisms which operate on the speech signal without any referential input (Peperkamp, Le Calvez, Nadal, & Dupoux, 2006; Maye, Werker, & Gerken, 2002; Vallabha, McClelland, Pons, Werker, & Amano, 2007; Swingley, 2009; Martin, Peperkamp, & Dupoux, 2013; Feldman, Myers, White, Griffiths, & Morgan, 2013; Dillon, Dunbar, & Idsardi, 2013). These mechanisms have been tested successfully with simplified input. However, they were not as successful when tested on more realistic acoustic data which is highly variable and noisy (e.g., Varadarajan, Khudanpur, & Dupoux, 2008; Fourtassi, Schatz, Varadarajan, & Dupoux, 2014; Jansen et al., 2013). Thus, though these mechanisms may play an important role, they are unlikely to account for the entire process of learning and refinement.

Learners are exposed to more than the speech signal. In particular, they usually have access to multimodal input which co-occur with speech. For example, the words *glass* and *grass* in English are typically associated with different visual input. Experimental data has shown that both children and adults can leverage such semantic/visual information to discriminate ambiguous sounds (Teinonen, Aslin, Alku, & Csibra, 2008; Yeung & Werker, 2009; Hayes-Harb, 2007).

Nevertheless, previous research has generally assumed–whether implicitly or explicitly–that learners have access, not only to the immediate visual input, but also to the entire meaning category intended by the speaker, e.g., the meaning of the word 'cow' is not limited to one specific cow–it includes cows of all shapes and colors, and it excludes instances of another category such as deer. Knowing the meaning's extension and boundary of a given word is crucial to the task of phoneme learning: If an ambiguous sound contrast is associated with two different objects (e.g., a cow and a deer), then in order to decide whether or not this contrast is phonemic, the learner has to determine first if the speakers' target meanings are two specific categories (cow and deer) or one broad category (e.g., animal). The contrast is phonemic in the former case, and non-phonemic in the latter.

Often, however, learners in the early stages of acquiring their first or second language, do not yet know the full mean-

ing extensions of the words they hear around them. Work in the word learning literature suggests that humans spontaneously restrict the set of possible extensions to taxonomic classes (Markman, 1989). For example, upon hearing the word 'cow' with an instance of, say, a brown cow, humans are unlikely to consider an extension that includes 'milk' or 'brown rice'. Though the taxonomic assumption simplifies the task, it still leaves a great deal of ambiguity regarding the level of generalization intended by the speaker. For example, the word 'cow' could have meant more abstract categories such as "mammal" or "animal".

The current work aims at studying how learners behave in a situation where there is uncertainty at both the phonological and semantic levels. We associate minimally different non-sense words (along the ambiguous sound contrast ɑ-ɔ) with pairs of objects that vary in their taxonomic proximity (Figure 1). Crucially, mere exposure to instances of the sound-object pairings is not enough to determine the exact meaning extension, leaving the participants in a situation of uncertainty similar to that faced by learners in the early stages of language acquisition. We are interested in the participants' subsequent judgment about the phonemic status of the pair of sounds.

There are several possible scenarios. For instance, participants may not be sensitive to the degree of taxonomic distance, treating all visual differences as equally indicative of a phonemic status. It is also possible that participants treat degrees of taxonomic distance in a categorical way, i.e., they may treat pairs of objects up to a certain taxonomic level as equally indicative of non-phonemicity, whereas they treat pairs of object beyond that level as equally indicative of phonemicity. Finally, participants may be sensitive to each gradient of taxonomic distance in their phonemic learning, in which case their judgements should be graded as well.

In what follows, we test these predictions with adults learning an alien language. In Experiment 1, we parametrize a subset of the semantic space, creating an evenly-spaced taxonomic scale, and we use this scale to explore the effect of different gradients of taxonomic distance on the phonemic status of the ɑ-ɔ contrast. In Experiment 2, we test whether results of Experiment 1 are due to interference from existing lexicalized categories in the first language. Finally, we discuss the implication of the findings on phoneme learning in the context of early language acquisition.

Experiment 1

The goal of this first experiment is to use a parameterized subset of the semantic space to test the effect of each gradient of taxonomic distance on learning the phonemic status of an ambiguous sound contrast. We use a between-subject design to avoid carry-over effects in the sound judgements.

Participants

152 Participants in total were recruited online through Amazon Mechanical Turk, restricting the pool to the United States residents. At the end of the experiment, participants were asked to rate the overall quality of the audio-visual stimuli on their local software/hardware. We excluded participants who judged this quality as medium or bad (N=26), keeping only those who rated the quality as good, that is, those for whom the experiment functioned correctly. We also excluded participants who took the experiment more than once (N=3), and participants who obtained less than 50% correct answers on the obvious filler questions (e.g., are the words "komi" and "pibu" different?) (N=5). We ended up with a sample size of 115 participants split across 5 groups.

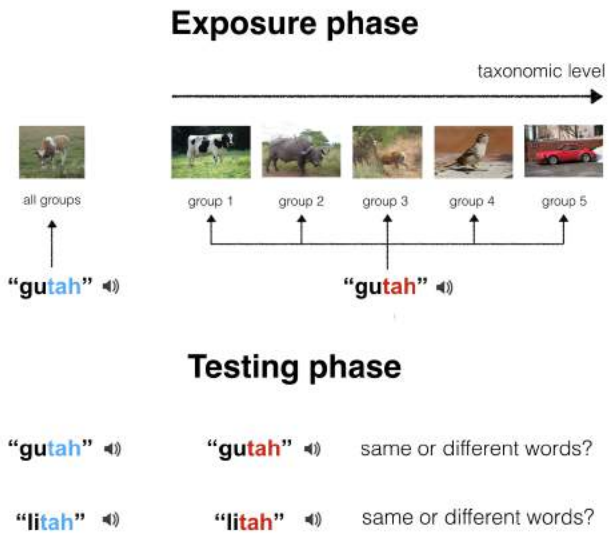


Figure 1: Overview of the task.

Stimuli

Objects The stimuli consist of a reference object (a cow), and five other objects which varied in their similarity to this reference. These objects were, in this order, another cow (with a different color), a buffalo, a deer, a bird and a car (Figure 2). To parmaterize the object stimuli in the taxonomic space, we recruited an additional N=30 participants online (through Amazon Mechanical Turk), and we asked them to rate the similarity of a series of pairs of objects in a 9 point-scale, 1 being "very similar" and 9 being "very different". The pairs of objects were formed by the pairwise combination of all six items described above. The order of trials was randomized across participants. We computed the average rating for each pair, which gave us a distance matrix.

Figure 2 (left) shows the taxonomic organization of the object stimuli, which we obtained via hierarchical clustering (using average linking) applied to participants' similarity data. Height indicates the average similarity within clusters at each hierarchical/taxonomic level. Figure 2 (right) shows a different visualization of the same data using bi-dimensional scaling. Both representations show that the way objects are organized around the reference (i.e. cow) corresponds to graded differences in the semantic space, and that these gra-

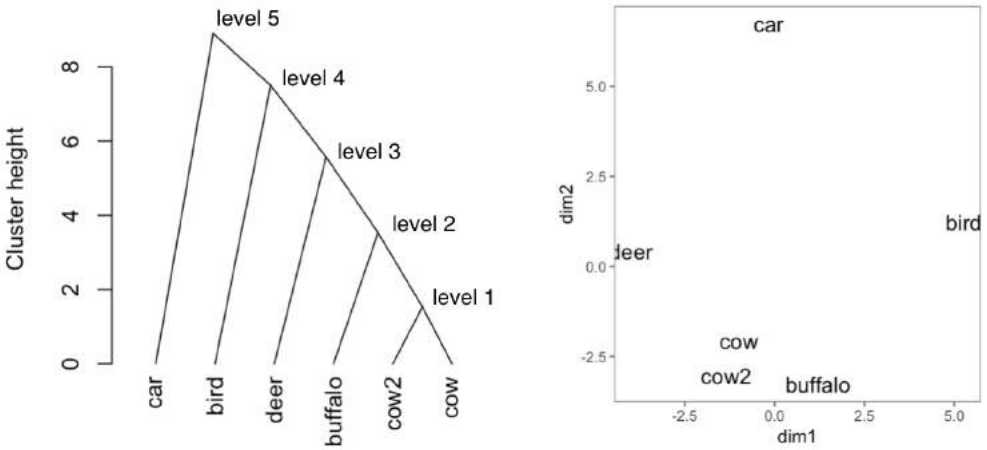


Figure 2: The graph on the left shows the taxonomic organization of the object stimuli obtained via hierarchical clustering of the participants' similarity ratings. The graph on the right shows another way of visualizing the same data via bi-dimensional scaling. Both representations show that the object stimuli lead to a graded and evenly-spaced semantic scale.

dients are quite evenly-spaced.

Sounds We followed Feldman et al. (2013) in using minimal pairs that vary along the vowel contrast [ɑ]-[ɔ]. This contrast is neither too acoustically similar, nor too different. In fact, depending on the dialect, these two vowels can be treated by English speakers as belonging to one or two categories (Labov, 1991). We chose such an ambiguous contrast in order to put the participants in a rather flexible situation where they can switch between phonological interpretations depending on the context. Two minimal pairs were constructed by concatenating two context syllables ([gu] and [li]) produced by a female native speaker of American English, with a target syllable contrast ([tɑ]-[tɔ]) produced by the same speaker. The resulting minimal pairs were [gutɑ]-[gutɔ] and [litɑ]-[litɔ]. For ease of presentation, we will refer to these minimal pairs by *gutah*/*gutaw* and *litah*/*litaw*. In addition, we used two artificial filler words [pibu], [komi] which we obtained by concatenating four vowels produced separately by the same speaker.[1]

Procedure

In order to avoid any carry-over effect on sound judgements, we used a between-subject design, i.e., each group of participants were exposed to only one degree of taxonomic distance. The minimal pair was paired with two objects whose similarity varied across five groups of participants (Figure 1). In all these groups, one member of the minimal pair (e.g., *gutah*) was paired with picture of a cow. The second member (i.e., *gutaw*) was paired with a referent whose similarity with the first referent varied on the five-step taxonomic scale.

The experiment had an exposure and a testing phase. In the exposure phase, participants heard a novel word in an alien language and saw the corresponding object simultaneously. In this phase subjects did not have to perform any specific task, but they were encouraged to listen carefully and try to learn the words. They were exposed to 3 series composed each of a randomized presentation of 4 word-object pairings: 2 target words (*gutah*/*gutaw*) whose referents similarity varied across groups of participants (Figure 2), and 2 filler words (*pibu* and *komi*) mapped invariably to two different objects. There were 12 trials in total, with each presentation lasting around 850 ms (i.e., the time it took the bi-syllabic word audio to complete).

In the testing phase, participants heard a series of trials composed of two word tokens, and were asked to judge if these tokens corresponded to different words in this artificial language, or if they represented a mere phonetic variation of the same word. We used a wording similar to the one used by Feldman et al. (2013)[2]. In this testing phase, subjects were encouraged to follow their intuition and think carefully before answering.

Half of the testing trials contained identical sounds ('same trials'), and the other half contained different sounds ('diff. trials') and were presented in a random order. The diff. trials were composed of the minimal pair used during the exposure phase (*gutah*/*gutaw*) plus a novel minimal pair containing the same syllable contrast (*litah*/*litaw*), and which we used only in the testing phase to investigate the ability of participants to generalize across the lexicon. There were 12 test trials in total: 4 for for the exposure word (2 same and 2 different), 4 for the generalization word (idem), and 4 for the fillers komi/pibu (idem). Participants were tested twice, once before exposure to referential data and once after the exposure.

[1] The audio stimuli were graciously provided by Naomi Feldman.

[2] "You will listen to pairs of words from an artificial language. You should decide if they are same or different. The words can be different in the language even if they are similar. Conversely, they can be same even if they are pronounced slightly differently."

Results

Figure 3 shows the proportion of times participants judged the minimal pairs as phonologically different as a function of group (i.e., taxonomic level), and as a function of the testing session, i.e., before or after exposure to referential data.[3] We show results for both diff. trials (e.g., "gutah"/"gutaw") and same trials ("gutah"/"gutah").

Note, first, that the proportion of 'different' answers for same trials was close to zero across groups, showing that participants were almost perfect in detecting same pairs. However, the proportion of 'different' on the diff. trials varied across groups, and this proportion was 50% in the 'before' session. These initial observations confirm our choice of the sound contrast, which was supposed to be perceptually distinguishable, but ambiguous in terms of its phonemic status, allowing participants to adjust their phonological interpretation depending on the referential context.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.07	0.41	-2.59	0.01
group	0.32	0.12	2.65	0.01
session	-0.86	0.35	-2.46	0.01
trial	-3.59	0.19	-18.55	<0.01
group:session	0.28	0.10	2.85	<0.01

Table 1: A mixed-effects logistic regression predicting participants' responses in a same-different task.

After exposure to referential data, we observed a graded effect of the objects' taxonomic distance on phonological judgment: Participants were more likely to judge the phonologically ambiguous contrast as different when this contrast corresponded to higher taxonomic levels (Figure 3). We fit a mixed-effect logistic regression which predicted the participants' response by the group (i.e., taxonomic distance), the session (before or after exposure the referential data), and the trial (same or diff. pairs). The model was specified as follows: `response` ∼ `group * session + trial + (1|Subj) + (1|item)`. The estimates are summarized in Table 1. Confirming our qualitative observations, the model shows that the type of the trial predicted the participants' responses (i.e., answering more 'same' on same pairs). Crucially, we also found an interaction between group and session, indicating that exposure to referential data influenced the participants' responses.

To further examine the influence of exposure on learning and generalization, we fit two simple mixed-effects logistic models to the diff. trials after exposure predicting responses as a function of group. We found an effect of object semantic distance on phonological judgment in both the exposure word ($\beta = 4.55$, $SE = 1.13$, $p < 0.01$) and the generalization item ($\beta = 2.58$, $SE = 1.00$, $p < 0.01$).

[3]Since answers do not vary across groups prior to the exposure phase, in the 'before' session we only show the average results where data were collapsed across groups.

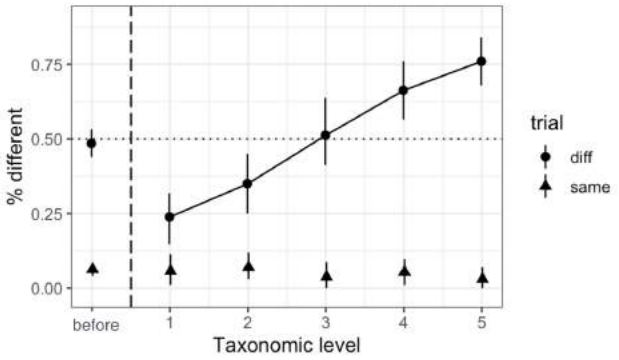


Figure 3: The points are the proportion of times participants judged the pair of sounds as 'different'. The triangles represent the judgments for exactly same pairs (e.g., gutah-gutah). The circles represent the judgments for different pairs (e.g., gutah-gutaw). Data on the left side of the vertical dashed line show the average responses before exposure to the referential data. The horizontal dotted line represents chance. Error bars represent 95% confidence intervals.

Discussion

Experiment 1 tested how gradients in the semantic space influence judgments about the phonemic status. It is possible, however, that participants relied, not on the taxonomic distance in the semantic space, but on available lexicalized concepts in their first language. Indeed, one could imagine that the more a common label is easily accessible for a pair of objects, the more participants answer "same" in the phonemic task. For example, if it is easier to access a common label in the case of *cow* and *deer* (e.g., "mammals"), than it is in the case of *cow* and *car* (e.g., "things"), then this difference may explain why participants judged the sound contrast more as phonemic in the latter. In Experiment 2 we explore whether such an account could explain the findings.

Experiment 2

We asked participants to provide common labels in English for each of the objects pairs used in Experiment 1, and we quantified the difficulty they had in generating these labels. If the phonemic judgements are driven by common labels in the first language, then the difficulty in accessing these labels should mimic closely the phonemic judgments (Figure 3).

Participants

40 participants were recruited online through Amazon Mechanical Turk, restricting the pool to the United States residents.

Stimuli

The same object stimuli used in Experiment 1.

Procedure

Participants were presented with pairs of objects, and were asked to type in, as fast as they could, the most specific la-

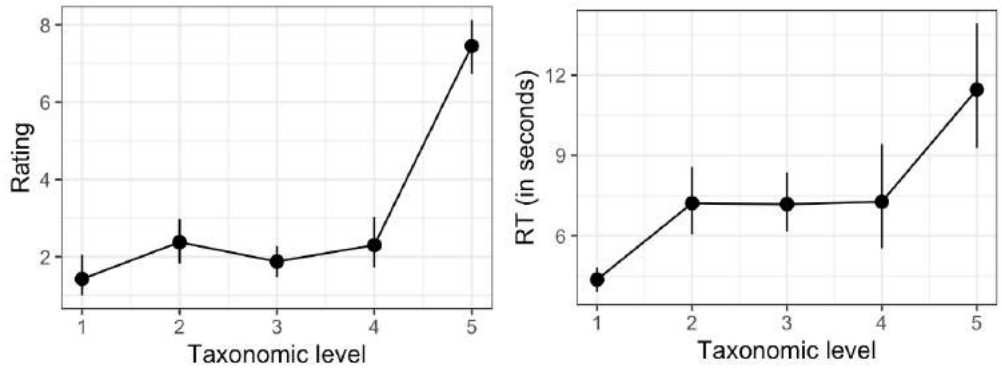


Figure 4: The graph on the left shows the average self-reported difficulty in generating a common English label for pairs of objects at different taxonomic levels (on a scale from 1 to 9). The graph on the right shows the average time it took participants to generate these labels.

bel in English that describes both objects (or "none" in the case they could not find a common label). We obtained both the labels and the reaction times, i.e., the time it took them from seeing the object to confirming their answer. Besides, participants were asked to evaluate the difficulty they had in generating the label on a scale from 1 to 9. The pairs were randomized across participants. To avoid carry-over effects, each participants saw the pairs only once.

Results

Results are shown in Figure 4. Overall, participants were faster and found it easier to generate a common labels when both objects were cows (the most frequent response was 'cow'). They were slower and found it difficult to generate a common labels when the pair was cow/car (most participants did not find a common label and typed 'none'). That said, they did not show any noticeable difference (neither in reaction times nor in subjective evaluation) for the intermediate cases. For all these cases, the most frequent response was 'animal'. Thus, though common labels in the first language may explain limit cases, it does not account for the entire pattern of graded responses obtained in Experiment 1.

General discussion

Previous research has suggested that semantic information can help with phoneme acquisition (Yeung & Werker, 2009; Hayes-Harb, 2007; Werker & Curtin, 2005). Nevertheless, learners often have to learn the phonemic status of the sounds they hear around them before they have determined the exact extension of the meaning intended by the speaker (e.g., a cow and a deer can be described by one or two words depending on the speaker's target level of taxonomy). The current work studied how the process of phoneme learning is influenced by such uncertainty at the semantic level.

More precisely, this study explored the effect of taxonomic distance on phonemic judgments. We associated minimally different word-forms with two semantic referents whose taxonomic distance varied across groups, and we asked participants in each group to judge the phonemic status of the corresponding sound contrast. We found that increasing the taxonomic distance induced graded judgments on the phonemic status, suggesting that learners are sensitive to the taxonomy of the referents when acquiring phonemes.

According to work in the word learning literature, humans have a bias towards extending the meaning of novel words to objects of similar kinds (Markman, 1989; Xu & Tenenbaum, 2007). In our case, this bias may have prompted participants to treat objects that were taxonomically similar (i.e., two cows with different colors, or cow/buffalo) as instances of the same meaning category, thus judging the sound variation as non-phonemic. In contrast, they may have treated objects of different kinds (cow/bird, or cow/car) as instances of different meaning categories, thus judging the corresponding sound variation as phonemic. Besides, the fact that participants provided graded–rather than stepwise–pattern of judgments mirroring the graded taxonomic distance suggests that they make use of probabilistic information at the semantic level to optimize the accuracy of their inference at the phonological level (see also Fourtassi & Frank, 2017).

How could the obtained relationship between taxonomic distance and phonemic judgements inform our understanding of development? First, this relationship may allow learners to collapse non-phonemic but perceivable sounds into the same phonemic category. This is crucial since the majority of sound contrasts in natural input consists of different pronunciations of the same word, rather than words that differ minimally (see Martin et al., 2013). Instances of the same words are likely to be associated with similar semantic information (i.e., at a similar taxonomic level), thus inducing a non-phonemic judgment for the corresponding contrasts.

As for true phonemic contrasts (e.g., *glass* vs. *grass*), sensitivity to the taxonomic structure will favor differentiation to the extent that minimal pairs have distant taxonomic distance in natural languages. Some research suggests that words that are similar phonologically tend to be similar semantically as well (Dautriche, Mahowald, Gibson, & Piantadosi, 2017).

However this research measured semantic similarity using a distributional model which relies on co-occurrence in a large corpus of text. It is possible that the type of semantic relationship that the model derived was thematic, rather than taxonomic. In fact, thematically related words can be taxonomically different (e.g., *cow* and *milk*). It has been shown that the nature of the semantic relationship depends on the model's parameter setting (Lenci, 2018). Further work on the semantic organization of minimal pairs is needed to elucidate this point.

The current study has some limitations. First, we only used the taxonomy of a subset of the conceptual space. To test the generality of the findings, future work will use different scales spanning several conceptual domains. Second, we only used familiar stimuli (real world objects and a native sound contrast). To completely rule out interference from categories in the native language, future work will seek to replicate the findings with non-native contrasts and with novel object stimuli.

To conclude, the current work showed that different degrees of taxonomic distance in the semantic space influence the acquisition of the phonemic status of sound contrasts. The findings show that learners make use of probabilistic information at the semantic level to optimize the accuracy of their phonemic judgments. More generally, this work suggests there to be an interaction between sound learning (phonemic judgment) and word learning (meaning generalization). Further work should aim at characterizing precisely this interaction and exploring its implications for both phonological and semantic development, two aspects of language development which have largely been studied separately.

All data and code for these analyses are available at
`https://github.com/afourtassi/top-down`

Acknowledgements

This work was supported by a post-doctoral grant from the Fyssen Foundation.

References

Dautriche, I., Mahowald, K., Gibson, E., & Piantadosi, S. (2017). Wordform similarity increases with semantic similarity: An analysis of 100 languages. *Cognitive Science*, *41*(8).

Dietrich, C., Swingley, D., & Werker, J. (2007). Native language governs interpretation of salient speech sound differences at 18 months. *Proceedings of the National Academy of Sciences*, *104*.

Dillon, B., Dunbar, E., & Idsardi, W. (2013). A single-stage approach to learning phonological categories: Insights from inuktitut. *Cognitive science*, *37*(2), 344–377.

Feldman, N. H., Myers, E. B., White, K. S., Griffiths, T. L., & Morgan, J. L. (2013). Word-level information influences phonetic learning in adults and infants. *Cognition*, *127*.

Fourtassi, A., & Frank, M. C. (2017). Word identification under multimodal uncertainty. In *Proceedings of the 39th annual meeting of the Cognitive Science Society.*

Fourtassi, A., Schatz, T., Varadarajan, B., & Dupoux, E. (2014). Exploring the Relative Role of Bottom-up and Top-down Information in Phoneme Learning. In *Proceedings of ACL.*

Hayes-Harb, R. (2007). Lexical and statistical evidence in the acquisition of second language phonemes. *Second Language Research*, *23*(1).

Jansen, A., Dupoux, E., Goldwater, S., Johnson, M., Khudanpur, S., Church, K., ... others (2013). A summary of the 2012 jhu clsp workshop on zero resource speech technologies and models of early language acquisition. In *2013 ieee international conference on acoustics, speech and signal processing* (pp. 8111–8115).

Kazanina, N., Phillips, C., & Idsardi, W. (2006). The influence of meaning on the perception of speech sounds. *Proceedings of the National Academy of Sciences*, *103*.

Labov, W. (1991). The three dialects of english. In P. Eckert (Ed.), *New ways of analyzing sound change.* New York, Academic Press.

Lenci, A. (2018). Distributional models of word meaning. *Annual review of Linguistics*, *4*.

Markman, E. M. (1989). *Categorization and naming in children: Problems of induction*. The MIT Press.

Martin, A., Peperkamp, S., & Dupoux, E. (2013). earning phonemes with a proto-lexicon. *Cognitive Science*, *37*.

Maye, J., Werker, J., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*.

Peperkamp, S., Le Calvez, R., Nadal, J., & Dupoux, E. (2006). The acquisition of allophonic rules: Statistical learning with linguistic constraints. *Cognition*, *101*.

Seidl, A., Cristi, A., Onishi, K., & Bernard, A. (2009). Allophonic and phonemic contrasts in infants' learning of sound patterns. *Language Learning and Development*, *5*.

Swingley, D. (2009). Contributions of infant word learning to language development. *Philosophical Transactions of the Royal Society B*, *364*.

Teinonen, T., Aslin, R., Alku, P., & Csibra, G. (2008). Visual speech contributes to phonetic learning in 6-month-old infants. *Cognition*, *108*.

Vallabha, G., McClelland, J., Pons, F., Werker, J., & Amano, S. (2007). Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences*, *104*(33), 13273.

Varadarajan, B., Khudanpur, S., & Dupoux, E. (2008). Unsupervised learning of acoustic sub-word units. In *Proceedings of the association for computational linguistics.*

Werker, J., & Curtin, S. (2005). Primir: A developmental framework of infant speech processing. *Language learning and development*, *1*.

Werker, J., & Tees, R. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the

first year of life. *Infant Behavior and Development*, *7*.
Xu, F., & Tenenbaum, J. (2007). Word Learning as Bayesian Inference. *Psychological Review*, *114*, 245.
Yeung, H., & Werker, J. (2009). Learning words' sounds before learning how words sound: 9-month-olds use distinct objects as cues to categorize speech information. *Cognition*, *113*, 234-243.

When Graph Comprehension Is An Insight Problem

Amy Rae Fox (amyraefox@ucsd.edu)
Department of Cognitive Science, University of California San Diego
The Design Lab—San Diego, California, USA

James D. Hollan
Department of Cognitive Science, University of California San Diego
The Design Lab—San Diego, California, USA

Caren M. Walker
Department of Psychology, University of California San Diego
Early Learning & Cognition Lab—San Diego, California, USA

Abstract

How do you make sense of an unconventional graph? Building on research demonstrating that prior knowledge of graphical conventions is difficult to overcome, we reconstrue graph reading as an insight problem. We hypothesize that imposing a *mental impasse* during a particular type of graph reading task will improve comprehension by inducing a sense of puzzlement, prompting learners to reconsider their interpretation. We find support for this proposal in a between-subjects experiment in which participants presented with an impasse-formulated version of graph reading questions are significantly more likely to correctly interpret a graph featuring an unconventional coordinate system. We characterize the differential patterns of mouse movements for learners between conditions and discuss implications for the use of novel graphical forms in science communication.

Keywords: graph comprehension; diagrammatic reasoning; insight; problem solving; representation; external representation; information visualization; mouse tracking

Introduction

The adage, "a picture is worth ten thousand words," surely applies to graphs. But what about a graph you don't know how to read? As Larkin and Simon note, "a representation is useful only if one has the productions that can use it," (1987, pg. 71). If we lack the ability to draw inferences from a graph, it is rendered useless. How is it then, that we develop such productions for new graphical forms?

Techniques for supporting graph comprehension have been a focus of research in the learning, cognitive and computer sciences for the past two decades. The most minimal interventions involve "graphical cues": visual elements that guide attention, akin to gesture and pointing. Acartürk (2014) investigated the influence of lines, arrows and point markers, finding that—used appropriately—such cues can help readers interpret the emphasis and temporal scope of a graph in alignment with a designer's intention. Kong and Agrawala (2012) surveyed the use of "graphical overlays" finding that reference structures (e.g. added gridlines), redundant encodings (e.g. data value labels), highlights, summary statistics, and annotations, are all commonly used to reduce cognitive load for particular graph reading tasks. Drawing inspiration from the literature in reading comprehension, Mautone & Mayer (2007) successfully demonstrated that animations, diagrams and drawings could help geology students connect the features of graphs to their geological referents. Each of these techniques serves to reinforce the semiotic connection between a graph, the world, and the reader's understanding, or to guide attention to information designers wish to emphasize. Importantly however, the techniques explored in this literature do not support learners in discerning *how* to read the graphs: the "rules" for their representational systems. Rather, it is assumed that the reader already has some familiarity with the type of graph being read (e.g. scatterplot, line graph, bar chart). In this way, the literature fails to differentiate between two types of prior knowledge brought to bear on a graph reading problem: knowledge of the domain, and knowledge of the graphical formalism.

In recent work (Fox & Hollan, 2018) we have taken up this challenge by investigating self-directed comprehension of an unconventional graph. In our paradigm, learners answer simple graph reading problems about a familiar domain—events in time—using an obscure graphical formalism. In an observational study, we found that readers struggled to make sense of the graph, misinterpreting the coordinate system as Cartesian. In a subsequent experiment, we evaluated four sets of instructional scaffolds aimed at overcoming the Cartesian misconception. We found that only an interactive version of the graph was effective for most learners. It seems that learners' expectations for the graphical formalism were so strong, even explicit text or image instructions failed to alert them to erroneous interpretations.

We argue this can be viewed as a sort of "graphical fixedness." Akin to Duncker's classic candle problem (1945), the learners in our previous studies were fixated on the conventional functions of the tools at their disposal: the marks on the page, and their assumptions about how axes and gridlines are meant to function. In the present work, we

reconceptualize our graph reading task as an insight problem. We test the hypothesis that intentionally inducing a state of puzzlement in learners—posing a *mental impasse*—will improve their ability to extract information from a simple unconventional graph.

Background

Graph Comprehension

Process models of graph comprehension describe an integration of top-down and bottom-up processing (Shah, Freedman, & Vekiri, 2005). Following the information processing tradition, these models invoke the concept of a *schema*: a structured representation of knowledge in long term memory that guides processing of new information in a "top- down" fashion (*see* Alba & Hasher, 1983; Anderson & Pearson, 1984). A number of theories describing graph comprehension have posited the existence of a graph schema that guides an individual's interpretation on the basis of their prior knowledge of similar external representations (Freedman & Shah, 2002; Pinker, 1990; Tabachneck-Schijf, Leonardo, & Simon, 1997).

Unsurprisingly, there is no consensus on the format or content of graph schemata. One important question that has been addressed is what features of a stimulus trigger activation of a particular graph schema. According to the "invariant structure view" certain general characteristics are shared across a number of graph types that then rely on a shared schema (Peebles & Cheng, 2003). Ratwani & Trafton (2008) proposed that the structural components of a graph that represent basic concepts and operations for extraction—the *graphical framework*—may be that invariant structure. In a scatterplot, for example, the graphical framework includes the x and y axes. From this formulation, one can predict that bar, line and scatterplot graphs (all relying on a Cartesian coordinate system) might invoke a single graph schema, while pie charts (relying on a polar coordinate system) might invoke a different schema. It is unclear what (if any) schema might be activated in order to comprehend a novel representation. Pinker (1990, p. 105) theorizes that upon encountering a novel graph, a reader will instantiate a "general graph schema", likely based on a combination of the graph's coordinate system and most predominate graphical forms (e.g. points, lines, bars, etc.) The exact mechanism of construction for this general schema is unknown, but Pinker suggests it may be related to the cognitive processes that represent abstract concepts like space and the movement of objects within it.

Prior Knowledge and Graphical Sensemaking

While the marks on a page invoke our prior knowledge of graphical formalisms, the context of the marks activate our knowledge of the domain (Shah & Hoeffner, 2002). We argue that scarcity of each type of prior knowledge impedes comprehension in different ways.

Limited prior knowledge. If presented with an unfamiliar graph depicting information in an unfamiliar domain, you will be unable to use knowledge of one to bootstrap inferences for the other. Imagine you are a novice physics student reading a Feynman diagram: without some understanding of particle physics, you cannot reverse-engineer the formalisms of the diagram. Without these formalisms, you cannot draw inferences about particle physics.

Limited prior *domain* knowledge. Alternatively, if presented with a familiar graph depicting data in an unfamiliar domain, you might draw on your knowledge of that graph type to learn something new about the content. If you know that a straight line represents a linear relationship, you can infer this relationship between unfamiliar variables connected by a straight line. It is this situation that we aim to optimize in STEM education. To this end, Mautone & Mayer (2007) demonstrated that animations, arrows, diagrams and photographs can all help students connect their prior knowledge of graphs to represented variables, improving their ability to draw inferences about related scientific concepts.

Limited prior *graphical* knowledge. Here, we are interested in the reciprocal case: an unfamiliar representation depicting information about a familiar domain; perhaps that strange-looking graph you saw in your favorite academic journal. Importantly, by "graphical knowledge", we are not referring to knowledge of graphs in general (graphicacy), but rather knowledge of the rules governing a *particular* graphic form. Can you figure out how to read the graph, if you know enough about the domain? (Test yourself! See Figure 1)

Reverse Engineering Formalisms. If the typical function of graphs is to use their formalisms as vehicles to learn something about the data (i.e. the domain) they represent, is the reverse also true? With sufficient domain knowledge, can readers reverse-engineer the formalisms governing a graph? Our data suggest this reciprocity of does not exist (Fox & Hollan, 2018). Despite extensive domain knowledge and personal experience with time, learners failed to correctly interpret the formalism of our graph with an unconventional coordinate system. Explicit instructions (text and images) were ineffective in supporting this reverse engineering, suggesting the need for a different scaffolding approach.

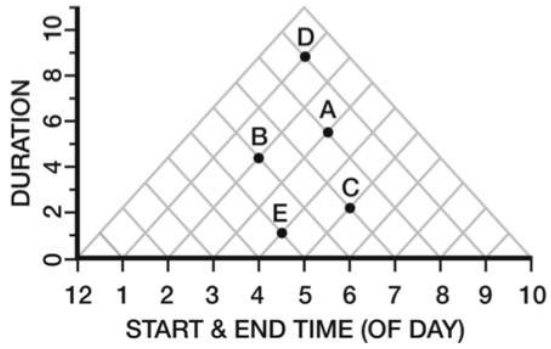
Problem Solving & Insight

In our earliest observational study with the Triangular Model graph (Figure 1), the vast majority of participants made the "Cartesian mistake": misinterpreting the graph as a Cartesian scatterplot (Fox & Hollan, 2018). However, for the few successful outliers, their production of the correct interpretation was accompanied by a protracted struggle, a

sudden clap of their hands and ecstatic exclamation, "Oh! *That's* how it works!"

What we observed were moments of *insight.* This insight came during the study debrief when we gave the learner feedback that their answers were incorrect. In some cases, this feedback alone was sufficient to produce a moment of insight. According to Ohlsson (1992), insight results when one breaks free from an impasse*:* "a mental state in which problem-solving has come to a halt; all possibilities have been exhausted and the problem-solver cannot think of any way to proceed" (pg. 4). But unlike traditional problems in the insight literature, the state of impasse in graph comprehension is not readily apparent. We must therefore craft the state of impasse to intentionally draw a learner's attention to their own misconception. The function of our feedback in the verbal debrief was to alert the learner to the fact they had made a mistake. While we cannot provide verbal feedback as a passive scaffold, we *can* indicate to learners that they've made a mistake by anticipating their mistaken response, and designing the graph reading question to exploit this error: relying on the convention that a multiple-choice question should have at least one response.

An Unconventional Graph: The Triangular Model of Interval of Relations



This graph depicts a schedule of events (A through E)
At what time does event B begin? [non-impasse]
What event(s) begin at 3? [impasse]
**see answers in acknowledgements*

Figure 1. A Triangular Model Graph (TM)

This line of research requires a very special stimulus: one that represents information about a familiar domain but is sufficiently obscure to be unrecognizable by most learners. We selected the Triangular Model graph (Figure 1) to depict information about schedules of events using a novel coordinate system. It has an informationally equivalent analogue, the Linear Model which, as the conventional external representation for intervals of time, is the basis for many scheduling artifacts including Gantt Charts. Both models indicate the start and end time, duration, and relations between intervals, which we present to participants as "events in time."

Based on work by Kulpa (2006) extended by (Qiang, Delafontaine, Versichele, De Maeyer, & Van de Weghe, 2012) the Triangular Model (*hereafter* TM) represents intervals as points in 2D metric space (Figure 1). Each point represents an interval of time. In the vertical dimension, the height of the point indicates its duration. The intersection of the point's triangular projections (using diagonally oriented grid lines) onto the *x*-axis indicate the start (leftmost) and end (rightmost) times. In this way, every interval is represented as a unique point in the 2D graph space, and each of its elementary properties are explicitly encoded by the location of the point. Although the graph's computational efficiency is best realized with a large number of data points, and tasks that require judgement about the relation between intervals (e.g. "starts-with", or "during" relations), first order readings (i.e. reading the start, end or duration) are readily available and directly reveal the reader's interpretation of the coordinate system. (See Qiang et. al (2012) for a thorough review of the computational efficiency of the Triangular Model, and elaboration of use cases for which it is preferable to more conventional interval graphics.)

A brief inspection of the TM by even the most experienced graph reader demonstrates its relative obscurity. However, while the coordinate system is unconventional, the graph depicts information about a domain in which we all share substantial prior knowledge: events in time.

The Present Study

Results of two prior studies (Fox & Hollan, 2018) give us reason to suspect that conventional graph knowledge may hinder comprehension of unconventional representations. In the case of the TM graph, Cartesian expectations for the structure of the coordinate system interfere with our ability to follow perceptual cues provided by the graph's diagonal gridlines. In the present study, we test the hypothesis that constructing a mental impasse will improve comprehension of this unconventional graph.

Methods

Participants and Design. Sixty (55% female) undergraduate STEM majors at a public American University participated in exchange for course credit (age: 18 - 33 years). We utilized a between-subjects design with two groups and one independent variable (scaffold: none [control] vs. impasse). Participants were randomly assigned to an experimental group, yielding 30 students per condition. Prior to analysis, data from six participants were excluded based on their failure to correctly answer an attention check question.

Procedure. Participants completed the study in person, seated at a desktop computer. After a brief introduction, they were randomly assigned to an experimental condition and

completed the Graph Reading Task, after which they received a short debrief. The session lasted approximately 30 minutes.

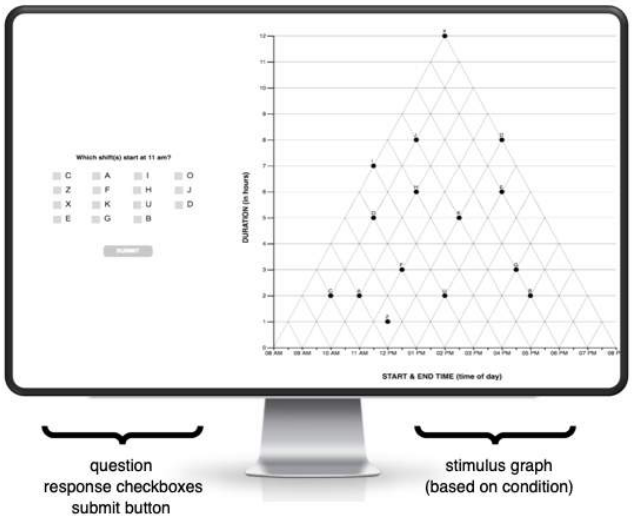
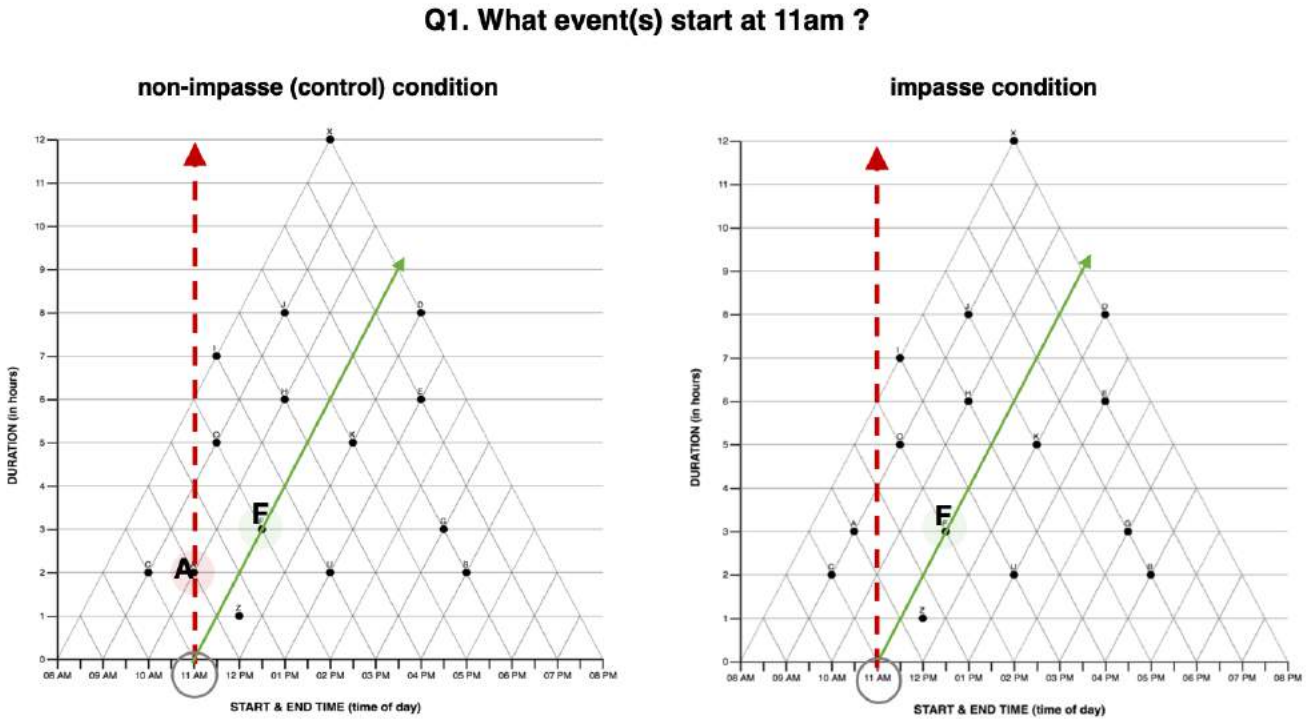


Figure 2. Sample stimulus

Materials. The Graph Reading Task consisted of a sequence of fifteen trials, each featuring a TM graph and multiple-choice question (Figure 2) about the temporal relationship between data points in the graph (i.e. "*Which event(s) start at 11am? At what time does event B end?*) Learners responded by clicking a checkbox corresponding to the data point(s) they wished to select. Trials were presented one at a time without feedback, in the same order for both conditions. Learners could not skip ahead nor return to previous questions. To assess the stability of student strategies over time, the first five trials included the assigned scaffold condition (none-control or impasse), while the following ten trials were identical (none-control). Questions were identical for both experimental conditions; however, the data sets rendered in the graph were slightly different for the first five trials. This allowed us to construct impasse problems with minimal differences between conditions. For each question in the non-impasse (control) condition, there was always a data point in the position where the participant would search if they interpreted the graph as Cartesian (Figure 3—*left*). Alternatively, in the impasse condition, the learner would find no data point in the expected position (Figure 3—*right*). For the final ten trials learners saw the same graph and questions. *See Figure 3 inset for a detailed description of the impasse structure.*



In both conditions, event F (solid green) is the single correct answer. In the non-impasse (control) condition at left, data point A crosses the Cartesian projection (dashed red line) from 11am, the most likely strategy taken by a learner misinterpreting the coordinate system. In this case, we expect the learner to select answer A. In the impasse condition (right) there is no data point intersecting the Cartesian projection. We expect this to pose a mental impasse. *note: extra lines and labels added for clarity, actual stimuli contain no such cues*

Figure 3. Experimental Conditions for Graph Reading Task Question #1

Data and Analysis. For each participant, we calculated a cumulative comprehension score [0-15], which served as the dependent variable. For further exploration of learner strategies, we integrated a JavaScript-based service (Mouseflow) to record all mouse-movements made by participants during the experiment session. Comprehension data were analyzed via inferential statistics, while mouse data were subject to exploratory qualitative analysis.

Results

Performance Accuracy. The mean comprehension score across the sample (n = 54) was approximately 6 points with a standard deviation of 0.68, and values ranging from 1 to 15 (max) points. On average, participants in the *impasse* group had higher scores (M = 7.6, SD = 5.2) than those in the non-impasse control group (M = 3.9, SD = 4.2), yielding a statistically significant difference $t(49.7) = -2.8$, $p = 0.006$; a moderate-sized effect $r = 0.37$.

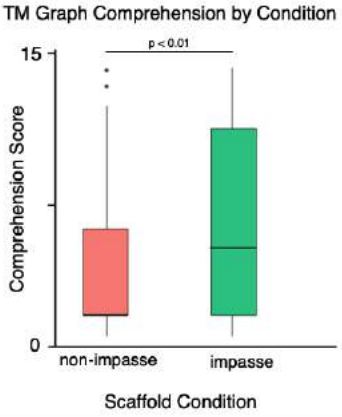


Figure 4. Results for graph reading task, by condition

Mouse Tracing Behavior. While raw comprehension scores can indicate whether learners correctly interpret the graph, they cannot reveal the strategies employed to answer the questions. To explore the mechanisms behind our results, we captured mouse tracing data. Similar to eye tracking data, mouse tracing provides an imperfect proxy for visual attention of the learner during the problem-solving session. This is a particularly rich source of insight for our graph reading problems as learners frequently used the mouse to navigate across the graph, the mouse acting like fingers tracing down or across gridlines. Of course, not all learners utilize the mouse to the same extent, and so we limit the present analysis to qualitative observation of gestalt patterns of graph traversal.

Figure 5 contains a set of heatmaps generated from raw path and dwell time data depicting the mouse movements of all participants on the *first* question of the Graph Reading Task. In the left column, we see data for learners in the control condition, and on the right, the impasse condition. The top row of heatmaps were generated from only those participants who correctly answered the question, while the bottom row from participants with a variety of incorrect answers. Visual inspection of these heatmaps reveal that across both conditions (top row), learners who correctly interpreted the coordinate system traversed the graph in a similar fashion, with the most prominent patterns following the relevant diagonal gridlines. Inspecting those with incorrect answers (bottom row), we see dramatically different patterns of tracing across conditions. While those in the control condition (bottom left) follow the expected Cartesian projection, learners in the impasse condition (bottom right) exhibit no single discernible pattern. While these learners did not arrive at the correct answer, their tracing behavior may be an indication of puzzlement.

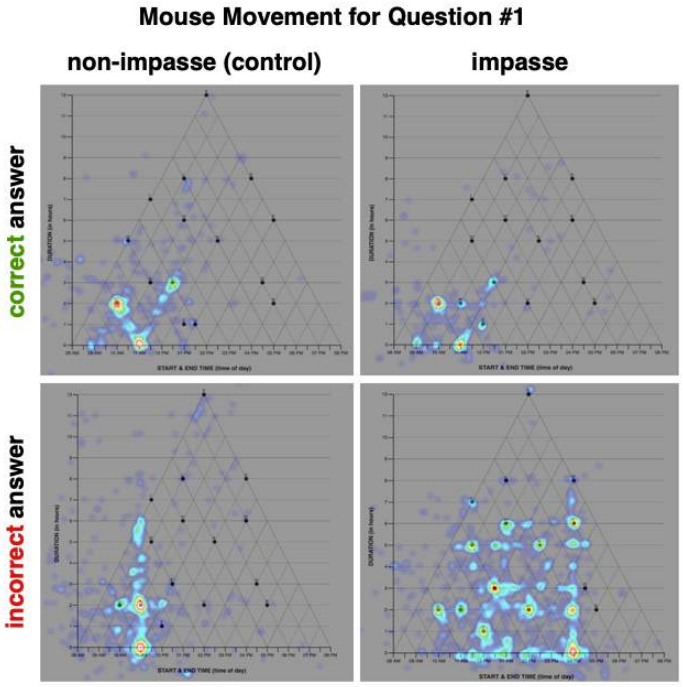


Figure 5. Mouse movement as heatmap for all participants, question #1

Discussion

The essence of functional fixedness, according to Ohlsson (1992), is that the experience of using an object in a particular way lowers the probability of finding a solution in which one uses the object in a different way. The strength of our association of the function to the object sets the strength of fixedness. In this light, we can see how substantial experience with common graphical forms serve to fix our expectations of axes, and coordinate systems in general, toward a Cartesian interpretation. The results of this study support our hypothesis that constructing a problem to present a learner with a mental impasse yields significantly better performance

on the unconventional graph reading task. Of course, not all graph reading tasks need be construed as insight problems. Most often the challenge we face concerns second-order readings—the inferences to be drawn from available information— and there is a close relationship between the nature of a graph-reading task and the suitability of the graph design (Shah & Hoeffner, 2002). However, we argue that these readings of trends and relationships between data points are unlikely to be made if the reader does not understand the nature of the graphical formalism itself, and this is where insight comes into play.

Lockhart, Lamon & Gick (1988) characterize difficulties in problem solving as a failure to *access* available information. This certainly seems applicable to the difficulties we observe with the Triangular Model graph where the reader need only perceive and recognize the importance of the diagonal gridlines to extract information from the graph (first-order readings). Lockhart et. al. propose that learners must often reconceptualize a problem in order to solve it, and simply giving students information may not be enough to achieve this effect. Presenting information in a form that induces puzzlement is significantly more effective in facilitating conceptual transfer and subsequent problem solving. We argue that the puzzlement induced by finding 'no available answer' in our impasse condition worked by leaving learners with no recourse but to reconsider their strategy (or give up). This conclusion is further supported by learners' failure to interpret this graph when provided with explicit information. While the text and image scaffolds in (Fox & Hollan, 2018) did not improve performance with the TM graph, a simple manipulation of the availability of answers to the first problems in a scenario for this study did.

We expect this technique should generalize to other representations with unconventional coordinate systems, though it is unclear whether the same attention-directing mechanisms would be appropriate for forms utilizing alternative markings. This is one of several open questions we are presently pursing. In ongoing analysis of mouse tracing data, we are exploring the strategies employed by learners in the impasse state and how they may reflect learner's graphical intuitions. In particular, we're interested in the strategies employed by learners in the impasse condition that provide non-Cartesian, but nonetheless incorrect responses. How are these learners reasoning about the graph elements, and does their behavior remain consistent after the scaffold phase (first 5 questions) when the remaining 10 questions have possible Cartesian answers? Based on ongoing analysis of the time course of response accuracy, we suspect that for impasse to be effective, the learner must confront the impasse in the initial phase of graph interpretation—when the graph schema is instantiated. In ongoing work, we address this question by varying the timing of impasse vs. non-impasse questions with analysis of the time course of correct and incorrect responses. We are also investigating which components of the design and layout of the graph are most influential in triggering a Cartesian interpretation, by manipulating the layout and saliency of axes, gridlines, and rotation of the figure in graph space.

While we hope this line of research will shed light on the elusive graph schema and how we develop graphical knowledge, the most immediate implications of our findings address the presentation of graphics in publications like this one. As communicators of science, we face an inevitable tension between communicating in what we believe to be the most revealing or expository fashion, and the way a community has come to expect. This makes innovation difficult. Nonetheless, the popularity of information visualization as a research area means that novel graphical forms are ever more present in our discourse. If you choose to utilize an unconventional representation in a traditional publication format, posing a carefully designed question (in perhaps, the figure caption) may aid the motivated reader to persevere in correctly reading the new graphic, and discovering your insights.

Acknowledgments

**answers to Figure 1: event B begins at 2; event A begins at 3.* Sincerest thanks are offered to Research Assistant Evan Barosay and The UCSD Design Lab. This work was supported in part by the United States Department of Defense (DoD) through the National Defense Science & Engineering Graduate Fellowship (NDSEG) Program.

References

Acartürk, C. (2014). Towards a systematic understanding of graphical cues in communication through statistical graphs. *Journal of Visual Languages and Computing*, *25*(2), 76–88. https://doi.org/10.1016/j.jvlc.2013.11.006

Alba, J., & Hasher, L. (1983). Is memory schematic? *Psychological Bulletin*, *93*(2), 203–231. https://doi.org/10.1037//0033-2909.93.2.203

Anderson, R., & Pearson, P. (1984). A schema-theoretic view of basic processes in reading comprehension. In P. Pearson (Ed.), *Handbook of Reading Research* (pp. 255–291). New York: Longman.

Duncker, K. (1945). "On problem solving". *Psychological Monographs, 58:5 (Whole No. 270).*

Fox, A. R., & Hollan, J. (2018). Read It This Way: Scaffolding Comprehension for Unconventional Statistical Graphs. In P. Chapman, G. Stapleton, A. Moktefi, S. Perez-Kriz, & F. Bellucci (Eds.), *Diagrammatic Representation and Inference* (pp. 441–457). Springer International Publishing.

Freedman, E. G., & Shah, P. (2002). Toward a model of knowledge-based graph comprehension. *Diagrammatic Representation and Inference*, 18–30. https://doi.org/10.1007/3-540-46037-3_3

Kong, N., & Agrawala, M. (2012). Graphical overlays: Using layered elements to aid chart reading. *IEEE Transactions on Visualization and Computer Graphics*, *18*(12), 2631–2638.

Kulpa, Z. (2006). A diagrammatic approach to investigate interval relations. *Journal of Visual Languages and Computing*, *17*(5), 466–502. https://doi.org/10.1016/j.jvlc.2005.10.004

Larkin, J., & Simon, H. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science*, *99*, 65–99.

Lockhart, R. S., Lamon, M., & Gick, M. L. (1988). Conceptual transfer in simple insight problems. *Memory & Cognition*, *16*(1), 36–44.

Mautone, P. D., & Mayer, R. E. (2007). Cognitive aids for guiding graph comprehension. *Journal of Educational Psychology*, *99*(3), 640–652. https://doi.org/10.1037/0022-0663.99.3.640

Ohlsson, S. (1992). Information-processing explanations of insight and related phenomena. In *Advances in the psychology of thinking* (pp. 1–44).

Peebles, D., & Cheng, P. C.-H. (2003). Modeling the effect of task and graphical representation on response latency in a graph reading task. *Human Factors*, *45*(1), 28–46. https://doi.org/10.1518/hfes.45.1.28.27225

Pinker, S. (1990). Theory of Graph Comprehension. In R. Freedle (Ed.), *Artificial Intelligence and the Future of Testing* (pp. 73–126). Hillsdale, NJ: Erlbaum.

Qiang, Y., Delafontaine, M., Versichele, M., De Maeyer, P., & Van de Weghe, N. (2012). Interactive Analysis of Time Intervals in a Two-Dimensional Space. *Information Visualization*, *11*(4), 255–272. https://doi.org/10.1177/1473871612436775

Ratwani, R. M., & Trafton, J. G. (2008). Shedding light on the graph schema: perceptual features versus invariant structure. *Psychonomic Bulletin & Review*, *15*(4), 757–762. https://doi.org/10.3758/pbr.15.4.757

Shah, Freedman, E. G., & Vekiri, I. (2005). The comprehension of quantitative information in graphical displays. In A. Miyake & P. Shah (Eds.), *Cambridge Handbook of Visuospatial Thinking*.

Shah, & Hoeffner. (2002). Review of Graph Comprehension Research: Implications for Instruction. *Educational Psychology Review*, *14*(1), 47–69.

Tabachneck-Schijf, H. J. M., Leonardo, A. M., & Simon, H. a. (1997). CaMeRa: A Computational Model of Multiple Representations. *Cognitive Science*, *21*(3), 305–350. https://doi.org/10.1207/s15516709cog2103_3

The interaction between structure and meaning in sentence comprehension: Recurrent neural networks and reading times

Stefan L. Frank (s.frank@let.ru.nl)
Centre for Language Studies, Radboud University
Erasmusplein 1, 6525 HT Nijmegen, The Netherlands

John C. J. Hoeks (j.c.j.hoeks@rug.nl)
Faculty of Arts, University of Groningen
Oude Kijk in 't Jatstraat 26, 9712 EK Groningen, The Netherlands

Abstract

Recurrent neural network (RNN) models of sentence processing have recently displayed a remarkable ability to learn aspects of structure comprehension, as evidenced by their ability to account for reading times on sentences with local syntactic ambiguities (i.e., garden-path effects). Here, we investigate if these models can also simulate the effect of semantic appropriateness of the ambiguity's readings. RNN-based estimates of surprisal of the disambiguating verb of sentences with an NP/S-coordination ambiguity (as in 'The wizard guards the king and the princess *protects* ...') show identical patters to human reading times on the same sentences: Surprisal is higher on ambiguous structures than on their disambiguated counterparts and this effect is weaker, but not absent, in cases of poor thematic fit between the verb and its potential object ('The teacher baked the cake and the baker *made* ...'). These results show that an RNN is able to simultaneously learn about structural and semantic relations between words and suggest that garden-path phenomena may be more closely related to word predictability than traditionally assumed.

Keywords: garden-path sentences; self-paced reading; reading time; thematic fit; recurrent neural network; LSTM; surprisal

Introduction

Garden-path phenomena, in which a local structural ambiguity results in comprehension difficulty upon disambiguation, have been studied extensively in psycholinguistics. Traditionally, the garden-path effect has been explained in terms of syntactic structure building: When the ambiguity is encountered, the parser chooses the structure that later turns out to be incorrect, triggering a process of syntactic reanalysis (e.g., Frazier & Rayner, 1982). Nowadays, this process is often expressed in probabilistic terms: The syntactic interpretation of the sentence-so-far takes the form of a probability distribution over (all) possible structures, and processing a word comes down to redistributing the probability mass in light of the incoming linguistic information. In case of a garden-path sentence, the incorrect reading of the ambiguity receives a (much) higher probability than the correct one, which means that a lot of probability mass needs to be redistributed upon encountering the disambiguating word (Brouwer, Fitz, & Hoeks, 2010; Hale, 2001; Levy, 2008). This corresponds to high cognitive processing load.

In the probabilistic account of sentence processing sketched above, the amount of update in the probability distribution due to processing a word can be shown to equal the word's *surprisal*, which has therefore been proposed as relevant measure of cognitive processing difficulty during incremental language comprehension (Hale, 2001; Levy, 2008). Indeed, word surprisal correlates with word reading time in general, as long as it is estimated by an accurate-enough probabilistic language model. The model's underlying architecture does not appear to matter much: It can be a probabilistic grammar (Boston, Hale, Patil, Kliegl, & Vasishth, 2008; Demberg & Keller, 2008), a recurrent neural network (Goodkind & Bicknell, 2018; Monsalve, Frank, & Vigliocco, 2012), or even a simple *n*-gram model (Frank, 2017; Smith & Levy, 2013). However, it stands to reason that surprisal must be estimated by a model that builds syntactic structure (like a probabilistic grammar does) if it is to account for the garden-path phenomenon. After all, the garden-path effect is (allegedly) caused by structural reanalysis. Hence, a model that does not engage in structure building should not be able to explain the effect.

Recent results from Long Short-Term Memory models (LSTM; Hochreiter & Schmidhuber, 1997) cast doubt on this assumption. An LSTM is a recurrent neural network in which the flow of activation is controlled by gates with learned weights, making it better at learning long-distance dependencies than Elman's (1990) well-known Simple Recurrent Network. LSTMs have shown remarkable capability to deal with long-term structure (Gulordava, Bojanowski, Grave, Linzen, & Baroni, 2018), including correct predictions of reading-time effects in garden-path sentences. Van Schijndel and Linzen (2018) had an LSTM estimate surprisal on the disambiguating verb phrase in sentences such as 'The employee understood [that] the contract *would be* ...' (NP/S ambiguity; critical word in italics) and 'Even though the girl phoned[,] the instructor *was* ...' (NP/Z ambiguity). They found higher surprisal in the locally ambiguous sentences than in their unambiguous counterparts.[1] Futrell, Wilcox, Morita, and Levy (2018) show that an LSTM model can account for the garden-path effect in sentence pairs such as 'The witness [that was] examined *by the lawyer*' (RR/MV ambiguity). Moreover, the model correctly predicts the absence of a garden-path effect when the subject noun is inanimate, as in 'The evidence [that

[1] Futrell et al. (2019), however, report that LSTMs predict a weaker NP/Z garden-path effect when the ambiguous region is longer, contrary to what has been observed in human readers (Tabor & Hutchins, 2004).

was] examined *by the lawyer*'. This suggests that the LSTM learned not only the relative frequencies of the different structures but also how these frequencies interact with a lexical semantic property.

The current study goes beyond this work by looking at a different local structural ambiguity and, more importantly, its interaction with the thematic fit between an action (i.e., verb) and its potential patient (i.e., syntactic object). That is, we investigate the sensitivity of LSTMs to a semantic *relation* as opposed to a single word's semantic property.

Garden paths and thematic fit

Sentence (1a) is structurally ambiguous when the third noun phrase ('the princess') in encountered: It can be understood as part of the larger NP 'the king and the princess' or as the beginning of a new sentence clause. This is known as the NP/S-coordination ambiguity. The upcoming verb ('protects') disambiguates towards the S-coordination structure, which causes comprehension difficulty compared to the unambiguous variant (1b). In other words, (1a) is a garden-path sentence because readers initially prefer the NP-coordination reading (Frazier, 1987).

(1a) The wizard guards the king and the princess *protects* the prince with her life.

(1b) The wizard guards the king, and the princess *protects* the prince with her life.

Sentence pairs (2a) and (2b) are structurally identical to (1a) and (1b) but differ in an important respect: The NP-coordination reading, in which the teacher bakes both the cake and the baker, is semantically anomalous: Bakers are not usually baked objects. Would such poor thematic fit lead to an immediate S-coordination interpretation and, consequently, remove any comprehension difficulty in (2a) compared to (2b)?

(2a) The teacher baked the cake and the baker *made* twelve breads for the coming holidays.

(2b) The teacher baked the cake, and the baker *made* twelve breads for the coming holidays.

In an eye-tracking experiment, Hoeks, Hendriks, Vonk, Brown, and Hagoort (2006) investigated the processing of sentences with NP/S coordination ambiguities in Dutch, which is structurally identical to English in this respect. They found the expected garden-path effect in the Good Fit condition: Reading times were longer on sentences such as (1a) than on (1b). When thematic fit was poor (sentence pair 2a/b) the picture was less clear, but the authors concluded that there is also a garden-path effect in this condition, albeit weaker than that for the sentences with good thematic fit.

However, the reliability of this result is questionable because the garden-path effect on Poor Fit sentences never reached statistical significance on any of the investigated reading time measures; it was at best marginally significant for total reading time. Hoeks et al.'s conclusion was based on the presence of a main effect of Ambiguity (i.e., whether or not the sentence had a comma) in combination with the *absence* of a significant interaction with Thematic Fit. Hence, the claim that the garden-path effect also occurred in the Poor Fit sentences is in fact based on accepting the null hypothesis that there is no interaction.

The current study

We trained LSTM models on Dutch text corpora after which they estimated surprisal of the critical words in the experimental sentences of the Hoeks et al. (2006) study. In addition, we analysed unpublished self-paced reading data on these same sentences. Bayesian mixed-effects regression analyses revealed similar patterns for the surprisal values and reading times (RTs): They are larger in the locally ambiguous than unambiguous sentences and this difference is smaller (but not zero) in case of poor thematic fit than for sentence with good thematic fit. These findings demonstrate that poor thematic fit indeed reduces, but not completely removes, the garden-path effect caused by the NP/S-coordination ambiguity; and that these effects can be explained by the statistical word-order patterns that recurrent neural networks are able to learn from text corpora.[2]

Method

Self-paced reading experiment

Stimuli The stimulus set was identical to that of Hoeks et al. (2006). It consisted of 120 experimental sentences with a local NP/S coordination ambiguity. In 60 of the 120 sentences, the two nouns of the (potential) NP coordination were animates, making them semantically plausible objects of the verb. These were the Good Thematic Fit sentences (Example 1a, translated from Dutch). In the 60 Poor Fit sentences, in contrast, the verb had a strong selectional preference for an inanimate object and only the first noun of the potential NP coordination was inanimate (see 2a). Items were not matched between the Good Fit and Poor Fit conditions.

The sentence's critical word was the second verb (italicized in Examples 1 and 2), which always disambiguated towards the S-coordination reading. Unambiguous versions of the sentences were constructed by simply introducing a comma after the second noun (Examples 1b and 2b).

In addition to the experimental sentences, there were 200 filler sentences, 80 of which had unambiguous conjoined object NPs. In half of these fillers sentences, both object nouns were animate; in the other half the first object noun was inanimate and the second one animate, mimicking the order of inanimate/animate nouns in the Poor Fit condition. The other 120 fillers contained relative clauses.

Forty items were paired with a simple comprehension question in the form of a statement about the sentence. These were intended to ensure participants would read for comprehension.

[2]The LSTM models, self-paced reading data, and analysis code are available from `https://osf.io/npzc7`.

Participants One hundred and three native Dutch speaking undergraduate students from Radboud University participated in the experiment. The data from seven participants were excluded from analysis because they answered more than 20% of the comprehension questions incorrectly.[3] This left 96 participants with analysed data.

Procedure Each participant read 120 experimental sentences, 30 in each of the 2×2 (Ambiguity × Thematic Fit) conditions. Stimuli were presented using word-by-word, non-cumulative, moving window self-paced reading. The sentence appeared when the participant pressed a button, but only the first word was visible initially. All other characters (including the comma but excluding spaces and the end-of-sentence period) were replaced by hyphens. On each subsequent button press, the next word would be revealed and the previous word changed back to hyphens. If, after completing the sentence, a comprehension question appeared, the participant had to indicate by button press whether or not the statement was correct.

Neural network models

Training corpus Training sentences were selected from the NLCOW2014 corpus (Schäfer, 2015) which contains individual Dutch sentences crawled from the web. It is divided into seven slices with approximately 37 million sentences each. NLCOW14 treats punctuation marks as individual tokens, meaning that they are separated from the preceding and the following word. Because this is incorrect in case of the apostrophe, we preprocessed the corpus, reattaching apostrophes to the word to which they belong.[4]

For each slice, we extracted the 20,000 most frequent words without distinguishing between upper- and lower-case and ignoring any string containing a non-letter other than the hyphen or apostrophe. Next, this frequent-word list was joined with the set of word types in the Hoeks et al. (2006) stimuli. We then selected only and all corpus sentences that contain only words from the combined word list.[5] These sentences form the training data from that slice. The seven training sets comprised between 8.57 and 9.00 million sentences (108 to 115 million tokens) each.

Model architecture We trained one LSTM network on each of the seven training data sets for two epochs. All networks had a 300-dimensional input embedding layer, a 600-unit recurrent layer, a 300-unit non-recurrent layer between the recurrent and output layers, and softmax output layer with one unit for each word type in the training set. No attempt was made to optimize this architecture. The seven networks differed only in their output layer sizes and random initial connection weights.

After processing the first $t-1$ words of a sentence, the network's output activation for word unit w is its estimate of $P(w_t|w_{1\ldots t-1})$: the probability that word w will occur at position t given the word sequence (sentence context) w_1 to w_{t-1}. The surprisal of the actually occurring next word is defined as the negative logarithm of its occurrence probability: $\text{surprisal}(w_t) = -\log P(w_t|w_{1\ldots t-1})$.

Test sentences All seven networks estimated surprisal on all experimental sentences in both the Ambiguous (comma absent) and Unambiguous (comma present) condition. However, in spite of the training sentence selection method described above, 22 of the 120 experimental stimuli sentences contained one or more words not present in all seven training data sets. We replaced these words by semantically congruent words from the same syntactic category that did occur in all training sets. For example, in *De politie traceerde de dief* ('The police traced the thief') the verb *traceerde* was changed to *achtervolgde* ('chased').

Data analysis

We analysed the effect of Ambiguity on surprisal and RT by fitting Bayesian mixed-effects regression models using the R package `brms` (Bürkner, 2018). A positive regression coefficient for Ambiguity (i.e., $\beta_{\text{ambiguity}} > 0$) indicates higher surprisal or RT on Ambiguous than Unambiguous sentences, that is, a (predicted) garden-path effect.

The prior for $\beta_{\text{ambiguity}}$ was an improper flat distribution over the real numbers, as is the default in `brms`. For the RT analysis, it would have been justified to have the prior be informed by the Hoeks et al. (2006) results. However, we opted for a flat prior so that exactly the same analysis could be run for surprisal as for RT. The dependent variable was normalized so the intercept of the regression line is guaranteed to be 0. Hence, we set the strong prior of $\mathcal{N}(0, 0.1)$ over the intercept. We chose the Exponentially modified Gaussian family because of the positive skew in the dependent variables' distributions. The regression model included as random effects by-network and by-item random intercepts and random slopes of Ambiguity. Random-effect priors were the `brms` defaults.

Separate analyses were run for the Good and Poor Fit conditions, in addition to analyses including the factors Ambiguity, Fit, and their interaction. Both the Ambiguity and Fit factors were effect coded (± 0.5) with positive values for the Ambiguous and Good Fit conditions. Priors on the Fit and interaction coefficients were the default improper flat distribution.

Because self-paced reading often leads to so-called spillover effects, where comprehension difficulty on a word results in reading slowdown at a later word, we analysed RT on both the critical word and the immediately following word.

[3] There were in fact two versions of the experiment, which differed only in whether or not the comprehension questions were presented. Fifty-five of the the 103 participants took part in the version that included the questions. The data from the two experiment versions are combined in our analysis.

[4] In Dutch orthography, apostrophes can occur in the plural suffix *-'s* and in unstressed forms of pronouns (e.g., *m'n*, 'my') and determiners (e.g., *'n*, 'a').

[5] Single-word sentences and sentences containing over 50 words were excluded, as were sentences containing a punctuation token other than the period, comma, exclamation point, and question mark.

For completeness, we did the same for the surprisal analysis even though there is no reason why surprisal effects would spill over to the next word.

RTs below 50ms or over 4000ms were considered outliers and removed from analysis, but there were only four such data points (three on the critical word, one on the post-critical word).

Results

Effects of ambiguity and thematic fit

The two upper panels of Figure 1 show the posterior probability densities for the effect of Ambiguity on RT, in the Good and Poor Fit conditions. The reading time pattern is consistent with the conclusions Hoeks et al. (2006) draw from eye-tracking data on the same items: The ambiguity leads to a garden-path effect that is stronger in the Good than Poor Fit condition. The latter is apparent from the fact that, in Poor Fit sentences, the effect of Ambiguity occurs only on the critical word whereas it spills over to (and is even stronger on) the following word of Good Fit sentences. Table 1 presents the probability that there is indeed a garden-path effect in each of the Thematic Fit conditions, as well as the probability of an interaction such that the Ambiguity effect is larger in the Good Fit than Poor Fit condition.

This RT pattern is correctly predicted by the LSTM, as can be seen in the lower panels of Figure 1 as well as in Table 1. There is a clear effect of Ambiguity on surprisal in both the Good and Poor Fit conditions, and the evidence for an interaction between Ambiguity and Fit is very strong. Surprisal effects appear on the critical word rather than the post-critical word, which supports the claim that the post-critical RT effects are due to spillover of comprehension difficulty that arises at the critical word.

Effect of network training

As shown in Figure 2, it takes about 1 to 3 million training sentences for the garden-path effect and its interaction with thematic fit to appear. These effects continue to grow in size with additional training.

Table 1: Posterior probabilities of positive coefficients (i.e., $P(\beta > 0)$) of Ambiguity and its interaction with Thematic Fit.

			Word position	
Coefficient	Fit	Dep. Var.	Critical	Post-crit.
$\beta_{\text{ambiguity}}$	Good	RT	.98	$> .99$
		surprisal	$> .99$	.18
	Poor	RT	.93	.69
		surprisal	$> .99$	.32
$\beta_{\text{ambiguity}\times\text{fit}}$		RT	.78	$> .99$
		surprisal	$> .99$	.36

Item-level analysis

To investigate whether LSTM surprisal accounts for garden-path effects at the item level, surprisal was averaged per sentence over the seven fully trained networks, and log-transformed RTs were averaged per sentence over participants as well as over the critical and post-critical words. Figure 3 shows a scatter plot of average surprisal against average log-RT, excluding the 22 sentences that were adapted for LSTM processing. Clearly, the surprisal estimates are unable to explain garden-path effects at the level of individual sentences.

Discussion

Surprisal and reading time

Patterns of surprisal on the critical word matched the self-paced-reading results (as well as Hoeks et al.'s, 2006, eye-tracking data) albeit not at the individual item level. When comparing between experimental conditions, surprisal was higher when the sentence contained a local ambiguity (e.g., the LSTMs predict a garden-path effect) and this effect of Ambiguity was reduced (but not absent) when poor thematic fit between the verb and a following noun made the correct S-coordination reading more semantically appropriate before the disambiguating word. These results again demonstrate the power of RNNs to learn fairly subtle structural and semantic aspects of language, and thereby account for human processing behaviour.

The absence of effects on surprisal at the post-critical word supports the interpretation that the effect on RT here is caused by spillover from the critical word, as Hoeks et al. (2006) also conclude on the basis of their eye-tracking data. In that study, the authors found the garden-path effect to be more short-lived on Poor compared to Good Thematic Fit sentences. Our analysis of self-paced RTs shows the same pattern, in that the effect has disappeared on the post-critical word in the Poor Fit but not in the Good Fit condition. This suggests there may be a qualitative difference in the garden-path effects between the Thematic Fit conditions, that is not captured by the unidimensional surprisal measure.

Structural processing in RNNs

As explained in the Introduction, garden-path effects have been explained in terms of syntactic reanalysis, or probabilistically in terms of the redistribution of probability mass over syntactic structures. However, RNNs do not encode syntactic structure, at least not explicitly, so why did our networks correctly predict the garden-path effect?

One possibility is that the correspondence between surprisal and reading time is just an artefact of the experimental items. Possibly, the mere presence of a comma speeds up reading at the critical word, but less so in the Poor Fit Sentences, without any relation to the garden-path phenomenon. However, even if this is the case, it leaves unexplained why at least three other garden-path effects have been explained by

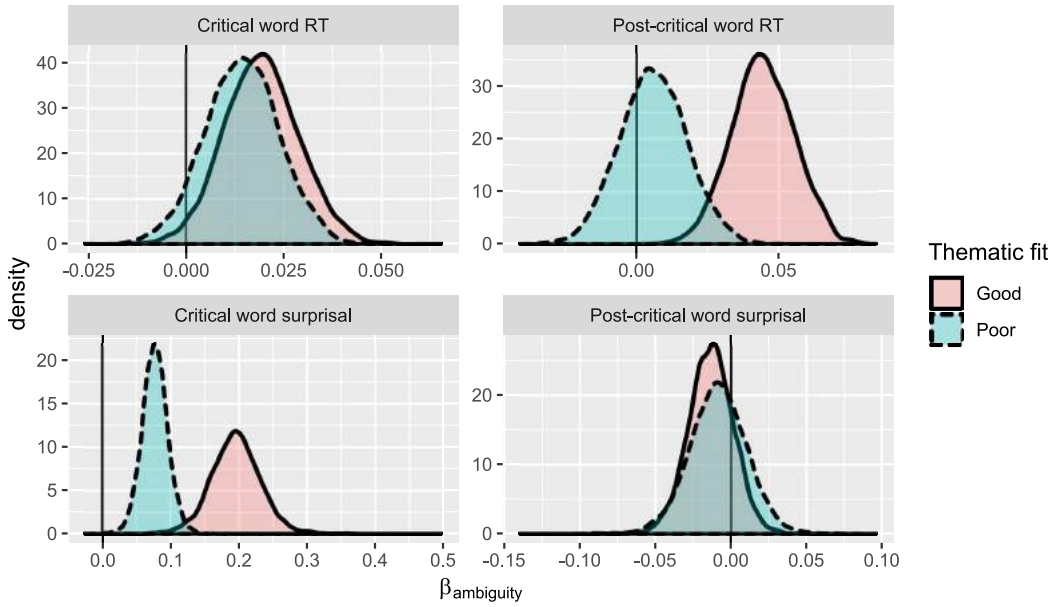


Figure 1: Posterior probability densities of the Ambiguity coefficient. Top: effect on RT; bottom: effect on word surprisal. Left: effect at critical word; right: effect at post-critical word.

Figure 2: Estimated coefficient of Ambiguity at the critical verb in the surprisal analysis, as a function of number of training sentences and thematic fit. Shaded areas represent 95% Credible Intervals.

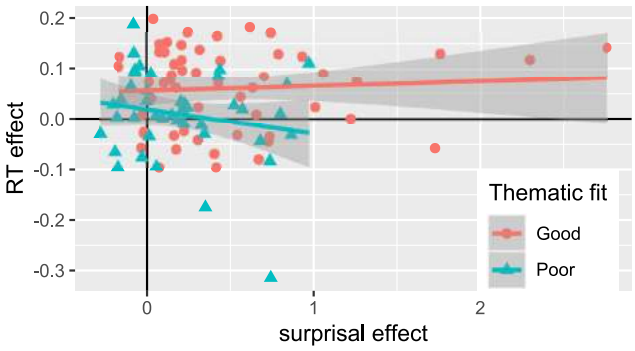


Figure 3: Garden-path effects in surprisal estimates and log-transformed RT, with regression line per Thematic Fit condition.

LSTM surprisal (Futrell et al., 2018; Van Schijndel & Linzen, 2018).

Alternatively, garden-path effects could be merely due to incorrect next-word prediction, as reflected in high surprisal on the disambiguating word. This would imply that there is no qualitative difference between comprehension difficulty due to a garden-path and due to an unlikely word co-occurrence. However, this seems implausible considering that ERP studies have shown that garden-pathing leads to a P600 effect (Osterhout & Holcomb, 1992; Osterhout, Holcomb, & Swinney, 1994) while higher surprisal in non-garden-path sentences corresponds to a stronger N400 component (Delaney-Busch, Morgan, Lau, & Kuperberg, 2019; Frank, Otten, Galli, & Vigliocco, 2015). Moreover, the initially preferred, but incorrect, reading of the ambiguity in a garden-path sentence can 'linger' (Christianson, Hollingworth, Halliwell, & Ferreira, 2001; Patson, Darowski, Moon, & Ferreira, 2009) which shows that such an interpretation was indeed entertained.

Possibly, being led up the garden path *also* results in incorrect next-word prediction and the *reading time effect* that comes with garden pathing actually reflects the resulting surprisal increase rather than the update of a structure or interpretation. However, this is not a particularly satisfying explanation as it would mean that the cognitive work of reanalysis

is itself not reflected in longer reading time.

Hence, we tentatively conclude that the LSTMs learn representations that capture relevant aspects of sentence structures/interpretations. As words come in, the network performs probabilistic, incremental reinterpretation, and generates word surprisal values that reflect the amount of representation update required to incorporate the word into the sentence representation under construction.

Conclusion

Word surprisal values estimated by LSTM models mirrored human reading times on garden-path sentences, predicting both the garden-path effect itself and its interaction with the manipulation of thematic fit between a verb and its potential object noun. This finding yet again demonstrates LSTMs' ability to extract structural aspects of language by learning to do next-word prediction in flat, unannotated text. Investigations of the neural networks' internal state are needed to substantiate this claim. If such an investigation fails to reveal evidence of structure representations in the networks, this would raise doubt about the necessity for structure building and revision in an explanation of garden-path phenomena.

Acknowledgments

The work presented here was funded by the Netherlands Organisation for Scientific Research (NWO) Gravitation Grant 024.001.006 awarded to the Language in Interaction Consortium.

References

Boston, M. F., Hale, J., Patil, U., Kliegl, R., & Vasishth, S. (2008). Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus. *Journal of Eye Movement Research*, *2*, 1–12.

Brouwer, H., Fitz, H., & Hoeks, J. (2010). Modeling the noun phrase versus sentence coordination ambiguity in Dutch: Evidence from surprisal theory. In *Proceedings of the 2010 Workshop on Cognitive Modeling and Computational Linguistics* (pp. 72–80). Uppsala, Sweden: Association for Computational Linguistics.

Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, *10*, 395–411.

Christianson, K., Hollingworth, A., Halliwell, J. F., & Ferreira, F. (2001). Thematic roles assigned along the garden path linger. *Cognitive Psychology*, *42*, 368–407.

Delaney-Busch, N., Morgan, E., Lau, E., & Kuperberg, G. R. (2019). Neural evidence for Bayesian trial-by-trial adaptation on the N400 during semantic priming. *Cognition*, *187*, 10–20.

Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, *109*, 193–210.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*, 179–211.

Frank, S. L. (2017). Word embedding distance does not predict word reading time. In *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 385–390). Austin, TX: Cognitive Science Society.

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, *140*, 1–11.

Frazier, L. (1987). Syntactic processing: evidence from Dutch. *Natural Language and Linguistic Theory*, *5*, 519–559.

Frazier, L., & Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology*, *14*, 178–210.

Futrell, R., Wilcox, E., Morita, T., & Levy, R. (2018). RNNs as psycholinguistic subjects: Syntactic state and grammatical dependency. *arXiv preprint*. Retrieved from `http://arxiv.org/abs/1809.01329`

Futrell, R., Wilcox, E., Morita, T., Qian, P., Ballesteros, M., & Levy, R. (2019). Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of the 18th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.*

Goodkind, A., & Bicknell, K. (2018). Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics* (pp. 10–18). Salt Lake City, UT: Association for Computational Linguistics.

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., & Baroni, M. (2018). Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1195–1205). New Orleans, LA: Association for Computational Linguistics.

Hale, J. T. (2001). A probabilistic Early parser as a psycholinguistic model. In *Proceedings of the Second Conference of the North American Chapter of the Association for Computational Linguistics* (Vol. 2, pp. 159–166). Pittsburgh, PA: Association for Computational Linguistics.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*, 1735–1780.

Hoeks, J., Hendriks, P., Vonk, W., Brown, C., & Hagoort, P. (2006). Processing the noun phrase versus sentence coordination ambiguity: Thematic information does not completely eliminate processing difficulty. *The Quarterly Journal of Experimental Psychology*, *59*, 1581–1599.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*, 1126–1177.

Monsalve, I. F., Frank, S. L., & Vigliocco, G. (2012). Lexical surprisal as a general predictor of reading time. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 398–408). Avignon, France.

Osterhout, L., & Holcomb, P. J. (1992). Event-related brain

potentials elicited by syntactic anomaly. *Journal of Memory and Language*, *31*, 785–806.

Osterhout, L., Holcomb, P. J., & Swinney, D. A. (1994). Brain potentials elicited by garden-path sentences: Evidence of the application of verb information during parsing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 786–803.

Patson, N., Darowski, E., Moon, N., & Ferreira, F. (2009). Lingering misinterpretations in garden-path sentences: evidence from a paraphrasing task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 280–285.

Schäfer, R. (2015). Processing and querying large web corpora with the COW14 architecture. In P. Bański, H. Biber, E. Breiteneder, M. Kupietz, H. Lüngen, & A. Witt (Eds.), *Proceedings of the 3rd Workshop on the Challenges in the Management of Large Corpora* (pp. 28–34). Mannheim, Germany: Institut für Deutsche Sprache.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*, 302–319.

Tabor, W., & Hutchins, S. (2004). Evidence for self-organized sentence processing: Digging-in effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 431–450.

Van Schijndel, M., & Linzen, T. (2018). Modeling garden path effects without explicit hierarchical syntax. In T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 2603–2608). Austin, TX: Cognitive Science Society.

Subjectivity-based adjective ordering maximizes communicative success

Michael Franke (mchfranke@gmail.com)
Institute of Cognitive Science, University of Osnabrück

Gregory Scontras (gscontra@uci.edu)
Department of Language Science, University of Californa, Irvine

Mihael Simonič (smihael@gmail.com)
Jožef Stefan Institute, Ljubljana

Abstract

Adjective ordering preferences (e.g., *big brown bag* vs. *brown big bag*) are robustly attested in English and many unrelated languages (Dixon, 1982). Scontras, Degen, and Goodman (2017) showed that adjective subjectivity is a robust predictor of ordering preferences in English: less subjective adjectives are preferred closer to the modified noun. In a follow-up to this empirical finding, Simonič (2018) and Scontras, Degen, and Goodman (to appear) claim that pressures from successful reference resolution and the hierarchical structure of modification explain subjectivity-based ordering preferences. We provide further support for this claim using large-scale simulations of reference scenarios, together with an empirically-motivated adjective semantics. In the vast majority of cases, subjectivity-based adjective orderings yield a higher probability of successful reference resolution.

Keywords: adjective ordering, subjectivity, reference resolution, hierarchical modification

Introduction

When speakers use two or more adjectives to modify a noun, they exhibit robust preferences in the relative order of the adjectives (e.g., *big brown bag* vs. *brown big bag*). Using a series of behavioral and corpus experiments, Scontras et al. (2017) demonstrated that adjective order in multi-adjective strings is reliably predicted by the subjectivity of the adjectives involved: less subjective adjectives are preferred closer to the modified noun, and the strength of the preference is modulated by the subjectivity differential between the adjectives. Thus, speakers strongly prefer *big brown bag* over *brown big bag*, as *brown* is much less subjective than *big*.

The question that immediately arises is why subjectivity should play the role it does in adjective ordering preferences. The current work follows Simonič (2018) and Scontras et al. (to appear) in advancing the claim that pressures from successful reference resolution deliver subjectivity-based ordering preferences. In certain cases of restrictive modification which proceed incrementally based on syntax-driven meaning composition, adjectives that compose with the nominal later will classify a smaller set of potential referents (e.g., the set of bags vs. the set of brown boxs). We demonstrate that, in order to avoid alignment errors where a listener might mischaracterize the intended referent, it is, when averaging over many contexts of use, a better strategy to introduce the more error-prone (i.e., more subjective) adjectives later in the hierarchical meaning composition; the structure linearizes such that subjectivity decreases the closer you get to the modified noun. We build on the work that precedes ours by making minimal assumptions about online processing (cf. Scontras et al., to appear) and by assuming a more principled implementation of adjective subjectivity within an empirically-motivated semantics (cf. Simonič, 2018).

The paper is structured as follows. First, we review the empirical generalization concerning subjectivity-based preferences, together with the proposals offered to account for this generalization. Then, we consider empirical work on adjective semantics, which serves as inspiration for our own proposal. We demonstrate, using Monte Carlo simulation, how a minimal set of independently-motivated assumptions leads to a ready explanation for subjectivity-based ordering preferences: ordering adjectives with respect to decreasing subjectivity has a higher probability of successful reference resolution, when averaging across many contexts of use.

Background

Given the robustness of adjective ordering preferences within and across languages, there has been no shortage of proposals meant to account for the regularities in adjective ordering. Some have offered grammatical proposals that attend to semantic composition or articulated syntactic hierarchies (e.g., Cinque, 1994; Scott, 2002; McNally & Boleda, 2004; Truswell, 2009). Others have advanced more psychological proposals built around notions like inherentness or accessibility (e.g, Whorf, 1945; Ziff, 1960; Martin, 1969). Recently, Scontras et al. (2017) synthesized several proposals that preceded them and advanced the hypothesis that adjective subjectivity predicts ordering preferences (see also Quirk, Greenbaum, Leech, & Svartvik, 1985; Hetzron, 1978; Dixon, 1982; Tucker, 1998; Hill, 2012).

In order to test the subjectivity hypothesis, Scontras et al. (2017) first had to determine what the ordering preferences were. They established a behavioral measure of the preferences whereby experimental participants indicated the preferred ordering of multi-adjective strings that differed only in the relative order of the adjectives involved (e.g., *the big brown bag* vs. *the brown big bag*). Scontras et al. (2017) then validated their behavioral measure by comparing it with naturalistic productions from corpora. They found a high correlation between the behavioral and corpus measures ($r^2 = .83, 95\%$ CI $[.63, .90]$), suggesting that the behavioral measure was successful in capturing the preferences speakers use

when forming multi-adjective strings.

Next, Scontras et al. (2017) measured adjective subjectivity. They started by simply asking participants how "subjective" a given adjective was (e.g., "How subjective is *brown*?"). Wary of how naive participants might interpret the word "subjective," the authors validated their subjectivity measure by comparing it with faultless disagreement scores (Kölbel, 2004; Barker, 2013; Kennedy, 2013; MacFarlane, 2014). In a faultless disagreement task, participants observe a disagreement between two speakers about whether an adjective applies to some object (e.g., whether or not a table is brown). The task is to decide whether the two speakers can both be right while disagreeing, or whether one of them must be wrong; to the extent that both speakers can be right, the adjective admits that degree of faultless disagreement. Scontras et al. (2017) found an extremely high correlation between the raw "subjectivity" scores and the faultless disagreement measure ($r^2 = .91, 95\%$ CI $[.86, .94]$), suggesting that they had a reliable measure of adjective subjectivity.

Comparing the ordering preferences with adjective subjectivity, Scontras et al. (2017) found that subjectivity accounts for 85% of the variance in the ordering preferences ($r^2 = .85, 95\%$ CI $[.75, .90]$) for 26 different adjectives from seven semantic classes. The authors then looked at every multi-adjective string in the Switchboard corpus of English, finding that subjectivity accounts for 61% of the variance in ordering preferences ($r^2 = .61, 95\%$ CI $[.47, .71]$) for 74 unique adjectives from 13 semantic classes. In other words, the authors found strong support for their hypothesis that subjectivity predicts adjective ordering preferences. The question that immediately presents itself, however, is why subjectivity should matter in adjective ordering. Scontras et al. (2017) gesture toward an answer to this question—less subjective adjectives are more useful for establishing reference—but their suggestion is purely speculative.

Using a model of probabilistic utterance choice (e.g., *big brown bag* vs. *brown big bag*), Simonič (2018) systematically explored the idea that subjectivity-based ordering preferences arise under pressure from successful reference resolution. The utterance choice model was formulated within the Rational Speech Act modeling framework (e.g., Franke & Jäger, 2016; Goodman & Frank, 2016; Scontras, Tessler, & Franke, n.d.).[1] To model adjective subjectivity, the speaker is taken to assumes that the listener might have a different lexical meaning for each adjective. If $L^{S,C}_{adj}$ is the speaker's lexical entry for adjective adj in context C, the speaker believes that the listener has lexical entry $L^{L,C}_{adj}$ with probability:

$$P(L^{L,C}_{adj} \mid L^{S,C}_{adj}) \propto \begin{cases} 1 & if L^{S,C}_{adj} = L^{L,C}_{adj} \\ \varepsilon_{adj} & \text{otherwise} \end{cases} \quad (1)$$

[1] See Hahn, Degen, Goodman, Jurafsky, and Futrell (2018) for a different approach to modeling adjective ordering within the Rational Speech Act framework. Their model defines speaker utility not in terms of referential success, but rather in terms of communicating subjective opinions about objects.

The more subjective the adjective, the higher the error probability ε_{adj}. With these beliefs about lexical divergence, Simonič shows that the subjectivity-based ordering *big brown bag* is a more rational choice for the speaker than *brown big bag* in a wide range of randomly-generated contexts. However, Simonič did not explicitly quantify the extent to which one ordering of adjectives is better than another, when averaging over many contexts.

Scontras et al. (to appear) pursue a similar explanation. They treat adjective subjectivity as potential noise in the semantics of an adjective, similar to Simonič, but they assume that, based on a ground-truth of objective adjective meaning, each agent (speaker or hearer) will incorrectly classify each potential referent in the current context C with an error rate ε_{adj}, which, again, indexes adjective subjectivity:

$$[\![\text{ADJ}]\!]^C = \lambda x \in C.\ \text{if } \texttt{ADJ}(x) \text{ then } \texttt{flip}(1-\varepsilon_{adj}), \quad (2)$$
$$\text{else } \texttt{flip}(\varepsilon_{adj})$$

This move allows Scontras et al. to treat deviations from the ground truth as gradient: greater deviation is increasingly less likely. Scontras et al. further assume that each object classification requires some processing cost. As a result, the error probability ε_{adj} is assumed to increase with the size of context C. Based on these assumptions, Scontras et al. demonstrate how subjectivity-based ordering preferences can maximize the probability of correctly classifying the intended referent. The authors explored 103,740 cases of multi-adjective modification and found that subjectivity-based ordering behaved as expected in 93% of those cases.

In sum, both Simonič (2018) and Scontras et al. (to appear) demonstrate how subjectivity-based adjective ordering serves successful referential communication. However, both accounts involve non-trivial and potentially controversial assumptions. Simonič's definition in (1) of the speaker's beliefs about the listener's lexicon are not very intuitive: why would the speaker believe that a small deviation from his own lexicon is equally likely as a massive deviation? Scontras et al. (to appear) likewise merely stipulate that error of classification ε_{adj} in (2) is a function of context size C. It would be much more desirable to derive divergences between the speaker's and listener's semantic classifications from more fundamental assumptions, first and foremost by a more explicit view of what the underlying semantics of adjectives is. Consequently, our aim here is to build on these previous accounts by showing how subjectivity-based ordering serves successful referential communication. However, rather than making what are now rather stipulative assumptions about the misalignment of semantic representations, we will show how these misalignments can arise from a generally plausible context-dependent semantics. It is to one such semantics that we turn next.

Semantic assumptions

Schmidt, Goodman, Barner, and Tenenbaum (2009) built their study of adjective meaning on the observation that gradable adjectives mean different things depending on the nouns

they modify: what counts as big for a mouse diverges drastically from what counts as big for an elephant. The question is what serves as the core meaning of a gradable adjective, such that speakers can determine its contextual extension?

To answer this question, Schmidt et al. collected human judgments about what counts as "tall" for different sets of objects. They then compared these judgments with the predictions from a number of semantic models that use various strategies to determine tallness in context. The strategies considered fell into one of two classes. The first class computed the tallness threshold directly, using various parametric and non-parametric procedures to compute a height cutoff above which objects count as tall. The second class inferred the tallness threshold on the basis of category membership, first performing a clustering analysis on the set of objects and then identifying as tall those objects that belonged to the cluster with the tallest object.

Two models outperformed the rest. The simplest was a threshold-computing model that sets the threshold on the basis of relative height by range: any object that fell within the top $k\%$ of the range of heights in context C counts as tall in C. Formally, the set $[\![\text{tall}]\!]^C$ of objects in C that count as tall in C is (where $\texttt{tall}(o)$ is the tallness of object o, $\texttt{max}$ is the tallness of the tallest object in C, and $\texttt{min}$ that of the smallest):

$$[\![\text{tall}]\!]^C = \{o \in C \mid \texttt{tall}(o) \geq \texttt{max} - \theta \cdot (\texttt{max} - \texttt{min})\}, \quad (3)$$
$$\text{where } \theta = k/100.$$

So, if the maximum object height is 10 on the relevant scale and the minimum height is 2, a k of 50% would set the tallness threshold at 6; that is, an object with a height of at least 6 would count as tall in that context. Notably, the more complex clustering model performed no better than this threshold model when it came to predicting human judgments. We will therefore use this simple but empirically-motivated threshold semantics in the reasoning that follows, treating the threshold θ as a free model variable.

Following Simonič (2018) and Scontras et al. (to appear), we assume that iterated adjectival modification triggers *sequentially intersective context updates*. Later adjectives (syntactically farther from the modified noun) are interpreted relative to contexts that are already restricted by previous adjectives. For example, the denotation of the phrase "$[\text{adj}_\text{i}\ [\text{adj}_\text{j}\ N]]$" given a shared context C of potential referents is:

$$[\![[\text{adj}_\text{i}\ [\text{adj}_\text{j}\ N]]]\!]^C = [\![\text{adj}_i]\!]^{[\![\text{adj}_j]\!]^{C \cap [\![N]\!]}} \quad (4)$$

In words, a string like "big brown bag" characterizes the set of all bags in context C that count as brown (in the set of bags in C) and that count as big (in the set of bags that count as brown in the set of bags in C). Each adjective is therefore interpreted relative to its local context of incremental compositional semantic interpretation, so to speak. The effect is that adjectives closer to the noun will operate over a larger context (i.e., one that is less restricted); paired with a context-dependent semantics as in (3), it is conceivable that the ordering of adjectives matters for referential success.

Figure 1: Illustration of subjective agent representations.

Motivating example

For the discussion that follows, we use "brown" and "big" as mnemonic labels for any two adjectives that are, respectively, less and more subjective. Our goal is to demonstrate why an utterance of "big brown X"—that is, a multi-adjective string ordered with respect to decreasing subjectivity—is communicatively more efficient on average than an utterance of "brown big X"—an utterance not ordered with respect to decreasing subjectivity. An utterance's average communicative success is spelled out here as the *expected utility* in a situation where the speaker wants to refer to an object; this value is specified as the average probability of the listener choosing the intended referent on the basis of that utterance.

We first need to make some assumptions about the effects of adjective subjectivity on our mental representations—representations that will be relevant to referential communication. Figure 1 gives a concrete example to illustrate the main idea. Suppose that the speaker and listener share access to a context of four bags that differ only with respect to color and size. Depending on their different perceptual angles, different background knowledge, or differences in previous experiences, the speaker and listener might represent the context differently: their impressions of object size and object color could deviate from the ground truth.

Here is where subjectivity comes in: we assume that more subjective properties are more likely to lead to deviation between the ground truth (i.e., the true context) and an agent's representation of the property. Crucially, by deviating from the ground truth, these more subjective properties are also more likely to lead to deviations between two agent representations (e.g., between the speaker's and listener's representations in Figure 1); these deviations *and our awareness of their potential* are what lead to perceived subjectivity as measured by a faultless disagreement task. Language users are aware that their representations might deviate from each other's, and the potential for deviation is different for different properties. We illustrate this tendency in Figure 1, where the agent representations of size deviate more from the ground truth than their representations of color.

We now ask: if the speaker wants to describe a bag that is both big and brown according to her subjective representation

of the context, would it be better, on average, to describe it as "big brown bag" or "brown big bag", if the listener would interpret either phrase from his own subjective perspective? Concretely, suppose the speaker wants to refer to bag 4 in Figure 1, which is both brown and big from her subjective point of view. If the listener hears "big brown bag", he tries to find the speaker-intended referent by incrementally restricting the set of possible referents according to the interpretation rule in (4), applying the context-dependent semantics in (3) to his own subjective representation of the objects in question. For the example from Figure 1 and assuming that $\theta = 0.5$ in (3), the phrase "brown bag" would make the listener consider only bags 2 and 4. Of these, only bag 4 is in the top 50% along the range of size in this context set. So, the interpretation of "big brown bag" is successful; the listener recovers the speaker-intended referent uniquely. In contrast, for the expression "brown big bag", the listener first looks at the bags that count as big, which rules out only bag 2, since it is the only bag whose size is in the lower 50% of the range of sizes. Among the remaining bags (1, 3 and 4), bag 3 is clearly not brown. For the sake of this informal example, assume that the listener therefore considers both bags 1 and 4 as possible referents when hearing "brown big bag". The chance of referential success (i.e., choosing bag 4)—neglecting salience or other factors—would be $1/2$, which is lower than the certain communicative success when interpreting "big brown bag".

Computing average communicative success

We use a Monte Carlo simulation to estimate the difference in expected referential success between phrases "big brown bag" and "brown big bag"; we calculate this value by averaging over many different contexts with different numbers of objects and varying degrees of subjectivity for the properties involved. In this way, we are not assuming that agents themselves necessarily reason actively about the stochastic misalignment of semantic judgements, or that they always choose expressions that are optimal with respect to these calculations in each context. (We will come back to this issue in the final discussion.) We merely compute the average communicative success of, say, a fictitious community of agents who would use "big brown bag" (i.e., subjectivity-based ordering) and compare their average communicative success to that of a different community that uses "brown big bag" instead.

A single run of the Monte Carlo simulation proceeds as follows:[2]

1. We first sample a number n of bags in the current context uniformly at random from 4 to 20.
2. We then sample the degree to which each object is brown and the degree to which it is big. Samples are independent draws from a standard normal distribution. This yields a representation of the *actual context* C as an $n \times 2$ matrix of feature values for the n objects. The probability of sampling context C for fixed n is
$$P(C \mid n) = \prod_{i=1}^{n} \prod_{j=1}^{2} \mathcal{N}(C_{ij} \mid \mu = 0, \sigma = 1)\,.$$
3. Agent X's (speaker's or listener's) subjective representation C^X of C is derived from C by assuming normally distributed noise around the property degrees in C, with a fixed standard deviation for each adjective. The probability of obtaining a subjective representation C^X from true C is
$$P(C^X \mid C) = \prod_{i=1}^{n} \prod_{j=1}^{2} \mathcal{N}(C^X_{ij} \mid \mu = C_{ij}, \sigma = \sigma_j)\,.$$
The standard deviations $\sigma_{1,2}$ are obtained by sampling two numbers uniformly from the interval $[0; 0.5]$ and assigning the higher number to the more subjective ("tall") and the lower to the less subjective adjective ("brown").
4. A *semantic threshold* θ is sampled uniformly at random from the unit interval. We apply the context-dependent threshold semantics in (3) from Schmidt et al. (2009) with the incrementally intersective context update in (4), using each agent's context representation, to yield each agent's subjective interpretation of each referential phrase.
5. We then sample the *speaker-intended referent object* i^* randomly from the set $[\![\text{adj}_1]\!]^{C^S} \cap [\![\text{adj}_2]\!]^{C^S}$ (i.e., an object that is both brown and big from the point of view of the speaker). If there is no such object, the run is discarded.
6. If the listener's interpretation of the phrase "[adj_i [adj_j]]" from his subjective point of view is $I = [\![[\text{adj}_\text{i}\ [\text{adj}_\text{j}]]]\!]^{C^L}$, the probability of recovering the intended referent is $|I|^{-1}$ if $i^* \in I$ and 0 otherwise. We record the probability of recovery for both adjective orders and evaluate their distribution over all samples obtained in this way.

Results

Based on 10^6 Monte Carlo samples from the process outlined above, we estimate the expected probability of recovering the speaker's intended referent with the subjectivity-based ordering "big brown bag" as 0.54, compared to 0.49 for the reverse ordering "brown big bag". The obtained samples of expected utilities for each ordering appear to indeed be different (paired t-test, $t \approx 19.261$, $p < 10e^{80}$). The direction of this difference lends credence to the general idea that, on average, ordering adjectives by subjectivity does affect average referential success, and that using the less subjective adjective early in sequential interpretation is communicatively beneficial. In other words, ordering adjectives with respect to decreasing subjectivity increases the probability of communicative success.

To understand these results better, Figure 2 shows results from Monte Carlo simulations for a small selection of the parameter values we investigated. We limit our focus to values for standard deviations $\sigma_{\text{brown}} \in \{0.1, 0.2\}$ and $\sigma_{\text{big}} \in$

[2] Code to reproduce this simulation can be found at `https://github.com/michael-franke/adjective_order`.

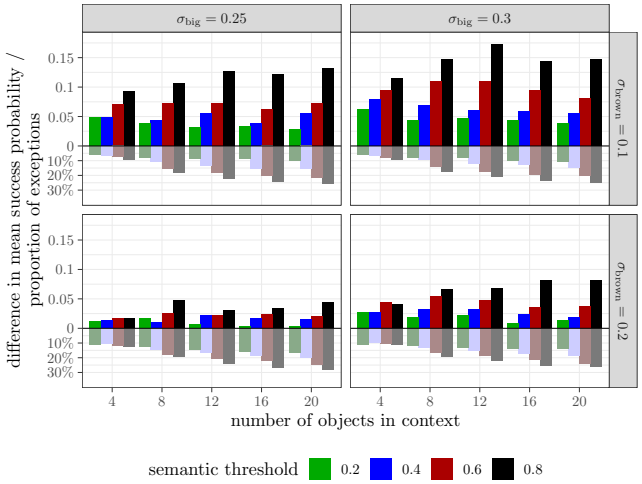


Figure 2: Results from Monte Carlo simulation with fixed values of σ_{brown}, σ_{tall} and θ. Above the 0-mark, the vertical axis shows the mean expected success of "big brown bag" minus that of "brown big bag". Below the 0-mark it shows the percentage of simulation runs where the latter ordering had a higher (however small) expected success.

$\{0.25, 0.3\}$ for the subjective agent representations; we consider semantic threshold values $\theta \in \{0.2, 0.4, 0.6, 0.8\}$. For each combination of these values, we ran 10,000 simulations following the procedure outlined above. The vertical axis in Figure 2 plots two measures. Upward from the 0-mark is the difference in mean communicative success between "big brown bag" and "brown big bag". We see that all mean values are positive, which signals that for all parameter constellations picked out here, the phrase "big brown bag" was indeed estimated to be communicatively more successful in each case. Below the 0-mark in Figure 2, we see the percentage of simulation runs in which the reverse ordering "brown big bag" had a higher expected utility. This shows that the communicative advantage of one adjective ordering over another is not absolute: there are exceptions. However, when averaging over all cases, there is nonetheless a clear communicative benefit of "big brown bag" over "brown big bag".

Discussion

The results of our simulation suggest that a simple, empirically-motivated adjective semantics can lead to increased communicative success when multi-adjective strings are ordered with respect to decreasing subjectivity. We thus have an answer for the question of why subjectivity should matter in adjective ordering: subjectivity matters because ordering adjectives by decreasing subjectivity increases communicative success. Importantly, we arrive at this conclusion without the potentially controversial assumptions from previous work (cf. Simonič, 2018; Scontras et al., to appear). However, our model is not without its own assumptions. In what follows, we revisit the critical assumptions that led to our findings.

From a theoretical standpoint, there are three important assumptions implemented by our model. While each of these assumptions may be challenged, they serve to deliver an articulated hypothesis concerning the interpretation of multi-adjective strings—a hypothesis that offers a plausible explanation for the role of subjectivity.

First, we here operationalize the subjectivity of property *A* as the degree to which, on average, listeners and speakers will have diverging (meaning-relevant) representations of the same object's property *A*. It bears noting that the subjective property representations we assume are not (necessarily) the same as the formal linguist's notion of degree. For us, these representations serve as an abstract way of implementing divergences in truth-value assignments. As modeled here, stochastic misalignments can arise from the particulars of perception in context, but these misalignments could also arise from differing general dispositions toward classifying an object as having the property *A* when paired with random other objects.

Second, we assume that adjectival modification is, at least sometimes (see below), incrementally intersective. Moreover, we assume that meaning composition follows the hierarchical syntactic structure, rather than the linear order of the relevant string. We share this assumption with both Simonič (2018) and Scontras et al. (to appear). This assumption—that the construction of a multi-adjective nominal proceeds outward from the modified noun—ostensibly stands at odds with findings concerning the linear uptake of information in adjectival modification (e.g., Eberhard, Spivey-Knowlton, Sedivy, & Tanenhaus, 1995; Sedivy, Tanenhaus, Chambers, & Carlson, 1999). However, this assumption is common to semantic analyses of modification and necessary in many cases of multi-adjective modification (e.g., "Minnesotan wild rice" or "angry bad apple"; McNally & Boleda, 2004).

The final critical assumption we make is that adjectives have, at least sometimes (see below), a meaning that is determined at least in part by the local context that they modify. In other words, we assume that it is possible to interpret the meaning of "big" in the phrase "big brown bag" as "big for the brown bags". This assumption is the primary driver of the increased communicative success for subjectivity-based orderings: placing more subjective adjectives farther from the modified noun means that they modify a smaller context, which means that there are fewer opportunities for the listener's subjective representation to deviate from the speaker's. While some adjectives are surely less likely to have variable meanings of this sort (e.g., "cardboard", "four-legged"), the presence of any such adjectives in a multi-adjective string will lead to the pressures summarized above, which means that they will lead to pressure toward subjectivity-based orderings.

When we combine these three assumptions, which appear necessary for the obtained results, we can see more clearly what the sources of assumed inter-subjectively divergent representations of objects might (not) be. For example, it is a nat-

ural idea to conceive of inter-subjective differences in judgements of whether a given object has property *A* as the result of inter-subjectively different beliefs about the comparison class against which *A*-hood of the object is evaluated. Concretely, agents might interpret "big" as "big for boxes from country *X*" where *X* is their, say, home country. Differences in the statistical distribution of sizes of boxes in different countries would then lead to inter-subjective disagreement about whether a given box might be "big" or not (Qing & Franke, 2014; Lassiter & Goodman, 2015). While we do not deny that this kind of inter-subjective divergences in comparison-class relative evaluations may exist, they are not, at least not straightforwardly, the kind of inter-subjective difference that drives the results of the present simulation study. This is because, as stressed above in connection with the third and final assumption, the presented setup requires inter-subjective differences that are affected by the current local context. This is compatible with the idea of differential beliefs about the comparison class. But if we wanted to say that diverging beliefs about the relevant comparison class are the main or sole factor that explains adjective ordering preferences based on the mechanism proposed here, we would have to spell out precisely how local contexts affect truth-value judgements in interaction with beliefs about the comparison class. — An interesting challenge for future work.

We conclude by considering the implications of our findings for our understanding of how adjective ordering preferences might develop over time. First, a note on the limitations of our findings. Our simulations, while extensive and systematic, have looked at a narrow sample of properties and scale types. We have begun to explore the predictions for other scale types (i.e., closed scales for adjectives like "full" or "safe"); however, a systematic investigation awaits future research. Still, we have demonstrated a clear communicative benefit of subjectivity-based orderings. Perhaps more importantly, we have demonstrated that this benefit does not apply universally to every possible multi-adjective string. Some parameter settings lead to exceptions where the reverse of subjectivity-based ordering yields a higher probability of communicative success.

The presence of exceptions suggests that speakers' robust, subjectivity-based adjective ordering preferences arise not out of active rational deliberation about the optimal ordering in context, but rather evolved gradually as speakers increasingly took notice of the communicative successes and failures associated with their utterances. In this way, the communicative pressures that favor subjectivity-based orderings in the majority of cases could have strengthened into the robust preferences we observe today. This sort of reasoning calls into question the nature of our knowledge of these preferences. It seems less likely that speakers represent this knowledge as a subjectivity-based heuristic that gets applied in the construction of multi-adjective strings, and more likely that the knowledge is a reflection of the statistical regularities of our linguistic experience.

Other potential explanations for subjectivity-based ordering preferences are conceivable. A prominent example is the recent explanation put forward by Hahn et al. (2018) who, unlike here, focus on non-restrictive uses of multi-adjective strings and communicative benefit related to exchanging subjective opinions about objects, which they show can be related to surface order and its impact on memory. We believe that this approach is perfectly compatible with our approach here. Both factors can play a role in supporting subjectivity-based adjective orderings. Even more usage-types of adjectival modification can and should be considered. Seen in this light, the present contribution is but a first step. It highlights that under one specific kind of use—albeit arguably the most fundamental information conveying mode of language: referential communication—a general benefit accrues for ordering adjectives by subjectivity in the way widely observed in many of the world's languages.

References

Barker, C. (2013). Negotiating Taste. *Inquiry*, *56*(2-3), 240–257. doi: 10.1080/0020174X.2013.784482

Cinque, G. (1994). On the evidence for partial N-movement in the Romance DP. In R. Kayne, G. Cinque, J. Koster, J.-Y. Pollock, L. Rizzi, & R. Zanuttini (Eds.), *Paths towards Universal Grammar. Studies in honor of Richard S. Kayne* (pp. 85–110). Washington DC: Georgetown University Press.

Dixon, R. M. W. (1982). *Where have all the adjectives gone? And other essays in semantics and syntax*. Berlin: Mouton.

Eberhard, K. M., Spivey-Knowlton, M. J., Sedivy, J. C., & Tanenhaus, M. K. (1995). Eye movements as a window into real-time spoken language comprehension in natural contexts. *Journal of Psycholinguistic Research*, *24*(6), 409-436. doi: 10.1007/BF02143160

Franke, M., & Jäger, G. (2016). Probabilistic pragmatics, or why Bayes' rule is probably important for pragmatics. *Zeitschrift für Sprachwissenschaft*, *35*(1), 3–44.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, *20*(11), 818-829.

Hahn, M., Degen, J., Goodman, N. D., Jurafsky, D., & Futrell, R. (2018). An information-theoretic explanation of adjective ordering preferences. In *Proceedings of the 40th annual conference of the Cognitive Science Society.* London: Cognitive Science Society.

Hetzron, R. (1978). On the relative order of adjectives. In H. Seiler (Ed.), *Language universals* (pp. 165–184). Tübingen: Narr.

Hill, F. (2012). Beauty before age? Applying subjectivity to automatic English adjective ordering. In *NAACL HLT 2012 Student Research Workshop* (pp. 11–16).

Kennedy, C. (2013). Two sources of subjectivity: Qualitative assessment and dimensional uncertainty. *Inquiry*, *56*(2-3), 258–277. doi: 10.1080/0020174X.2013.784483

Kölbel, M. (2004). Faultless Disagreement. *Proceedings of the Aristotelian Society*, *104*, 53–73. doi: 10.1111/j.0066-7373.2004.00081.x

Lassiter, D., & Goodman, N. D. (2015). Adjectival vagueness in a Bayesian model of interpretation. *Synthese*, *194*(10), 3801-3836.

MacFarlane, J. (2014). *Assessment Sensitivity*. Oxford: Clarendon Press.

Martin, J. E. (1969). Semantic determinants of preferred adjective order. *Journal of Verbal Learning and Verbal Behavior*, *8*, 697–704.

McNally, L., & Boleda, G. (2004). Relational adjectives as properties of kinds. *Empirical Issues in Formal Syntax and Semantics*, *5*, 179–196.

Qing, C., & Franke, M. (2014). Gradable adjectives, vagueness, and optimal language use: A speaker-oriented model. In J. Grieser, T. Snider, S. D'Antonio, & M. Wiegand (Eds.), *Proceedings of SALT 44* (pp. 23–41). elanguage.net.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A Comprehensive Grammar of the English Language.* London: Longmans.

Schmidt, L. A., Goodman, N. D., Barner, D., & Tenenbaum, J. B. (2009). How tall is *tall*? Compositionality, statistics, and gradable adjectives. In *Proceedings of the 31st annual conference of the Cognitive Science Society.*

Scontras, G., Degen, J., & Goodman, N. D. (2017). Subjectivity predicts adjective ordering preferences. *Open Mind: Discoveries in Cognitive Science*, *1*(1), 53-65. doi: 10.1162/opmi_a_00005

Scontras, G., Degen, J., & Goodman, N. D. (to appear). On the grammatical source of adjective ordering preferences. *Semantics and Pragmatics*.

Scontras, G., Tessler, M. H., & Franke, M. (n.d.). *Probabilistic language understanding: An introduction to the Rational Speech Act framework.*

Scott, G.-J. (2002). Stacked adjectival modification and the structure of nominal phrases. In G. Cinque (Ed.), *The cartography of syntactic structures, Volume 1: Functional structure in the DP and IP* (pp. 91–120). Oxford: Oxford University Press.

Sedivy, J. C., Tanenhaus, M. K., Chambers, C. G., & Carlson, G. N. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition*, *71*, 109-147. doi: 10.1016/S0010-0277(99)00025-6

Simonič, M. (2018). *Functional explanation of adjective ordering preferences using probabilistic programming* (Unpublished master's thesis). University of Tübingen.

Truswell, R. (2009). Attributive adjectives and nominal templates. *Linguistic Inquiry*, *40*, 525-533. doi: 10.1162/ling.2009.40.3.525

Tucker, G. (1998). *The lexicogrammar of adjectives: A systemic functional approach to lexis*. London: Cassell Academic.

Whorf, B. L. (1945). Grammatical Categories. *Language*, *21*(1), 1–11. doi: 10.2307/410199

Ziff, P. (1960). *Semantic Analysis*. Ithaca, NY: Cornell University Press.

Simulating Explanatory Coexistence: Integrated, Synthetic, and Target-Dependent Reasoning

Scott E. Friedman (friedman@sift.net)
SIFT, 319 N 1st Ave.
Minneapolis, MN 55401 USA

Micah B. Goldwater (micah.goldwater@sydney.edu.au)
The University of Sydney, School of Psychology
Brennan MacCallum Building (A18)
NSW 2006 Australia

Abstract

Understanding the cognitive structure of explanations— and the cognitive processes that assemble them— is a milestone for understanding how people learn and communicate. Recent research on *explanatory coexistence* suggests that people's causal beliefs are less globally coherent than previously thought: people use seemingly-competing supernatural and biological causes to explain different aspects of the same phenomenon, or they assemble supernatural and biological causes into single, coherent explanations (Legare & Gelman, 2008; Legare & Shtulman, 2018; Shtulman & Lombrozo, 2016). This coexistence— and unexpected coherence— of diverse causal mechanisms poses interesting questions about the role of coherence and fragmentation in people's mental models and explanations. This paper presents a computational model of explanatory coherence in the well-characterized domain of disease transmission, extending a previous cognitive model of explanation-based conceptual change (Friedman, Forbus, & Sherin, 2018). Our approach (1) retrieves diverse causal model fragments based on the phenomenon to explain, (2) assembles coherent causal models using relevance-directed abductive reasoning, and (3) selects explanatory paths that support within-explanation and within-scenario coherence. Our model simulates the three different types of explanatory coexistence detailed in the literature.

Keywords: cognitive modeling; explanatory coexistence; AI; abductive reasoning; explanation

Introduction

The cognitive process of explanation has been a central focus of cognitive science since its inception, and it has broad implications for communication, instruction, and conceptual change (Chi, De Leeuw, Chiu, & LaVancher, 1994; Vosniadou, 1994; diSessa & Sherin, 1998; Shtulman & Lombrozo, 2016; Friedman et al., 2018). The more recent focus on *explanatory coexistence*, whereby people utilize diverse— and seemingly incompatible— causal mechanisms in their explanations (Legare & Shtulman, 2018), poses additional questions about how people construct and consider explanations, how explanations are structured, and how explanations cohere with other beliefs.

This paper presents a computational cognitive model of explanation, building on previous cognitive models of conceptual change (Friedman et al., 2018). We apply our cognitive model to simulate human subjects' explanatory coexistence in the domain of disease, as characterized by Legare and Gelman (2008) and later by Legare and Shtulman (2018).

Our cognitive model assembles situation-specific causal models from smaller, generic *model fragments* (i.e., causal knowledge units). Given a new situation to explain, the model explains the situation by:

1. Retrieving causal model fragments based on the situation.
2. Traversing backwards recursively, instantiating model fragments within the situation in an relevance-directed beam search, assuming entities and relations as necessary.
3. Identifying the causal path(s) that maximize an objective coherence function with respect to global assumptions, coverage over the situation, and *presupposition* beliefs.

This model assumes that intuitive and culturally-acquired knowledge coexists, and that the process of assembling explanations is biased principally by coherence. This means that scientific and supernatural causal mechanisms can coexist in the same explanation, e.g., so that supernatural events might cause a biological event that leads to a viral infection, assuming the causal knowledge is primed and applicable.

Our simulation results demonstrate that that our model (1) simulates the three categories of explanatory coexistence in the literature and (2) varies its choice of explanation according to priming in a manner similar to human subjects.

We continue with an overview of explanatory coexistence and computational methods used in our cognitive model. We then describe our approach, present our simulation results, and conclude with a discussion of our results, key psychological assumptions, and directions for future work.

Background

We describe psychology research on explanatory coexistence, and then we review computational modeling techniques relevant to our simulation.

Explanatory Coexistence

There are scientific and religious or supernatural explanations for the same natural phenomena (e.g., creation of the universe, death, disease transmission). It is intuitive that learning scientific explanations for natural phenomena would replace previously learned supernatural explanations; however, evidence over the past decade suggests the opposite: scientific explanations replacing supernatural explanations is the

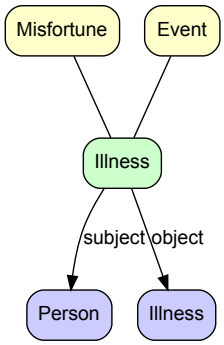


Figure 1: Model fragment for *Illness*.

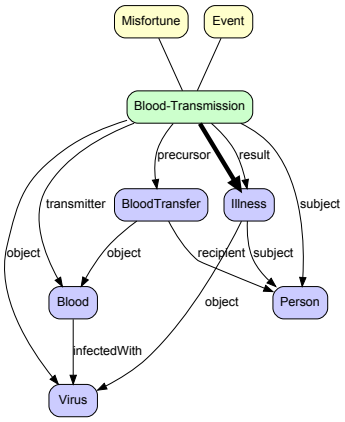


Figure 2: Model fragment for *Blood-Transmission* of disease via *BloodTransfer*.

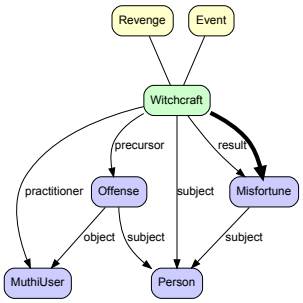


Figure 3: Model fragment for *Witchcraft* causing misfortune after an offense.

exceptional case. More frequently, people utilize both explanations (Shtulman & Lombrozo, 2016).

Legare and Gelman (2008) examined the specific case of explaining HIV transmission in South Africa. Before educational interventions focusing on the biological transmission of HIV, AIDS symptoms were explained as the result of witchcraft. Legare and Gelman (2008) showed that the educational interventions did not replace the bewitchment explanations; instead, both biological and bewitchment explanations coexist. For example, a man may have contracted HIV from sexual intercourse, but was attracted to a woman with HIV because of witchcraft.

Legare and Shtulman (2018) acknowledge the following categories of explanatory coexistence, all of which we simulate in this work:

1. **Integrated** reasoning combines seemingly-incompatible causal mechanisms into a coherent causal structure. For instance, bewitchment could cause somebody to choose a sexual partner who has AIDS, and intercourse with that partner causes disease transmission.
2. **Synthetic** reasoning invokes multiple causal mechanisms without articulating hierarchical or temporal precedence to any, possibly due to competing explanations.
3. **Target-dependent** reasoning applies different mechanisms to distinct aspects of a situation, in a highly-contextualized fashion. The various mechanisms do not participate in the *same* explanation.

Compositional Modeling

Simulating people's causal mental models requires expressive knowledge representation and reasoning (KR&R). An approach using only atomic logical propositions is not expressive enough to suit the mental model literature (Vosniadou, 1994; Chi et al., 1994; diSessa & Sherin, 1998; Gentner & Stevens, 1983) or the analogy literature (Friedman, Barbella, & Forbus, 2012), and an approach using only neural networks does not support sufficient interpretability.

Previous KR&R research on *compositional modeling* (Falkenhainer & Forbus, 1991) provides (1) representations for modeling the structure and continuous processes of dynamic systems, and (2) algorithms for composing these models on-the-fly for novel situations. Structure-behavior-function models (Goel, Rugaber, & Vattam, 2009) expand on this formalism to capture teleology, and have been used to simulate people's mental models.

Following recent cognitive modeling work (see the **Assembled Coherence** subsection), we simulate people's mental models using compositional modeling semantics extended with more expressive event structure (Pustejovsky, 2013). We represent each causal mechanism as a generic *model fragment* that can compose with others into large situation-specific explanations. Each model fragment describes:

- **Categories** that it instantiates, from general (e.g., *Misfortune*) to specific (e.g., *Sexual Transmission* [of a virus]).
- **Participants** are the entities or events that interact within the described mechanism. Each participant has one or more categories of its own. Model fragments with the same binding of participants are semantically equivalent.
- **Constraints** are *existence* conditions specified over the participants. If the constraints hold over participants in a situation, the model fragment may be *instantiated*.
- **Consequences** are functional or behavioral representations specified over the participants. They are asserted into the situation when the model fragment is instantiated.[1]

We include diagrams of three of the ten model fragments used in this simulation domain: Figure 1 shows a simple fragment describing an *Illness* state: a *subject* participant of type *Person*; a *object* participant of type *Disease*; and super-categories of *Event* and *Misfortune*.

Figure 2 shows the more complex fragment *Blood Transmission* [of disease], including sub-events of *Blood Transfer*

[1] A consequence may have *conditions* that must hold in the situation for them to be asserted, but in this paper each conclusion is a causal relationship without conditions.

and *Illness*, as well as a conclusion stating that, if instantiated, the *Blood Transmission* is a cause of the *Illness*. Importantly, this model fragment constrains its own participants: its subject (the *Person* at lower-right) is the subject of the *Illness* and the *Blood Transfer*. These constraints are required for the fragment— and its causal structure— to be realized.

Finally, Figure 3 illustrates a simple *Witchcraft* fragment, with a conclusion stating that an instantiated *Witchcraft* can cause a *Misfortune* of its subject. Per Figures 1 and 2, both *Illness* and *Blood Transmission* are types of *Misfortune*, so they can be directly caused by *Witchcraft*. This allows assembly of larger causal models, provided the situation (or explicit assumptions) satisfies the fragments' constraints. We describe assumptions below, and their affect on explanation quality.

Scalability of compositional modeling is a key consideration as the number of fragments grows, since the deductive closure of possible models can grow geometrically. We later describe a relevance-based heuristic in our approach that jointly (1) reduces the compositional modeling search space drastically and (2) helps model priming effects.

Abductive Reasoning

Abductive reasoning generates multiple explanations for observations— potentially generating assumptions along the way— and then selects the *"best"* explanation and its constituent assumptions as inferences or rationale for the observations. Previous computational approaches have modeled explanation quality as numerical cost (Charniak & Shimony, 1994) or as likelihood maximization with Bayesian approaches (Raghavan & Mooney, 2010). Our approach uses cost-based abductive reasoning to select explanations built from model fragments, and could be extended to use Bayesian approaches if we had estimates of subjects' beliefs of prior probability distributions. To improve scalability over previous abduction approaches— since the search space can grow geometrically (Poole, 1993)— our approach uses a relevance-based heuristic to guide its search for explanations.

Assembled Coherence

This paper extends recent work on the *assembled coherence (AC) theory* (Friedman et al., 2018) of mental models and conceptual change.

AC theory proposes that fragmented knowledge is assembled into larger, coherent mental models through the process of *abductive reasoning* (i.e., reasoning to the best explanation). Once assembled, these mental models are evaluated against a network of *presupposition* beliefs and then reused in novel situations by partial reformulation or by analogy (Friedman et al., 2012). This incorporates ideas from both the knowledge-in-pieces (diSessa & Sherin, 1998) and framework theory (Vosniadou, 1994) perspectives of mental models, and postulates that the two perspectives are compatible and complementary.

AC theory has been implemented in computational cognitive models to simulate explanation-based conceptual change in the domains of force dynamics (Friedman & Forbus, 2010), the day-night cycle (Friedman et al., 2012), the human circulatory system (Friedman & Forbus, 2011), and seasonal change (Friedman et al., 2018).

Approach

Our computational model generates a causal explanation by (1) retrieving model fragments based on the scenario to explain, (2) instantiating causal model fragments in an effect-to-cause beam search prioritized by relevance, (3) scoring coherent explanatory paths for coherence, and (4) selecting the most optimal explanatory path. We describe each of these processes below.

Retrieving causal knowledge. Given a new situation to explain, the system retrieves its model fragments (i.e., causal mechanisms) based on the categorical and relational overlap of the situation with those of its model fragments.

Specifically, given a situation s and a model fragment m, we compute relevance $Rel(m,s)$ with respect to the model fragment's participant categories C_m and constraint relations R_m and the situation's categories C_s and relations R_s. We use a simple Jaccard distance as a relevance function:

$$Rel(m,s) = \frac{|C_m \cap C_s| + |R_m \cap R_s|}{|C_m \cup C_s| + |R_m \cup R_s|} \quad (1)$$

This relevance function is a very coarse estimate of a model fragment's applicability to a situation, and we use it for simplicity: a model fragment's relevance strictly increases with situation-shared categories (e.g., *Person*, *Blood*, *SexualIntercourse*) and relations (e.g., *infectedWith*, *knows*, *motherOf*), and its relevance decreases monotonically relative to its total number of categories and relations. This approach is similar to performing spreading activation (Crestani, 1997) from categories and relations to relevant model fragments but allows indexing for scalability.

We discuss other plausible retrieval and salience factors in the conclusion of this paper.

Relevance-directed beam search. Given its relevance over causal model components, the system performs an incremental backward search through the space of possible causal models. This process is given an *explanandum* (i.e., event or assertion to explain), such as the illness of an individual, and then performs the following recursive operations for its explanation queue.

For each item x in its queue, it finds *applicable* model fragments that have x's type as a habitat consequence, e.g., if x is an *Illness*, then *BloodTransmission* (Figure 2) and *Witchcraft* (Figure 3) both apply. It selects applicable model fragments within the top 10% relevance window and attempts to compose the retrieved model fragment(s), constraining them by binding x as the consequent participant. The composition algorithm may assume any participants necessary to compose at least one instance, provided it obeys the input binding(s). The system then adds these new instances (e.g., *BloodTransmission* or *Witchcraft*) to the queue and will focus on those next, repeating.

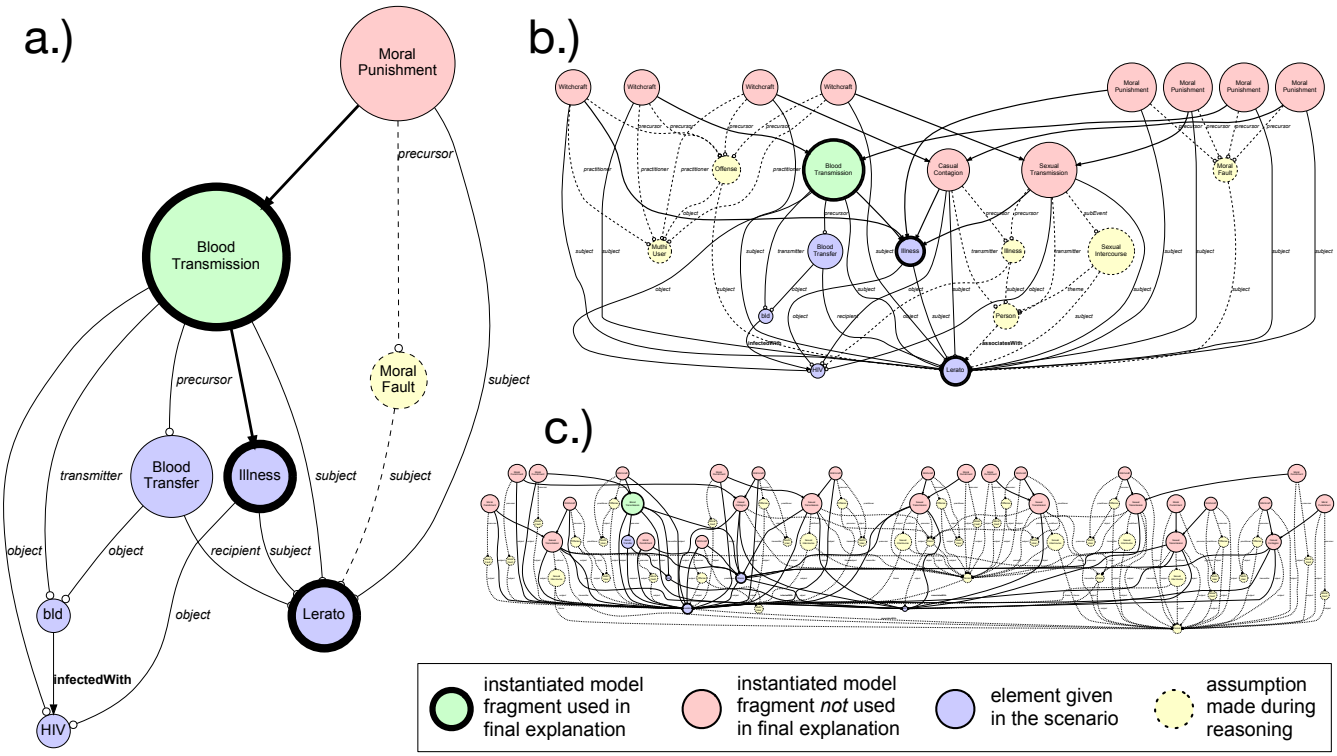


Figure 4: Explanations generated and selected for the same prompt of Lerato's AIDS after a blood transfer: (a) with relevance-directed causal search; (b) with undirected causal search; and (c) with exhaustive forward-chaining. All approaches utilize the same causal knowledge and result in the same final explanation, but with orders of magnitude difference in computation.

Explanation structure. Relevance-directed beam search produces a network of model fragment instances, as illustrated in Figure 4(a). In Figure 4(a), we provided the system the situation plotted in blue: Lerato has HIV and was the recipient of a blood transfer. Lerato and her *Illness* instance are outlined in bold for clarity. As with the human subjects of Legare and Gelman (2008), Lerato's illness is the explanandum in every simulation in this paper, but we vary the details of the situation to model priming effects.

The green (e.g., *Blood Transmission*) and red (e.g., *Moral Punishment*) elements in Figure 4(a) are instances of model fragments that were retrieved and assembled by this algorithm. The difference is that the green instances were chosen as part of the *best explanation* (described below), and the red instances were assembled and considered by the system, but were not ultimately included in the best explanation.

The yellow elements (e.g., *Moral Fault*) were *assumed* during the course of instantiation in order to satisfy model fragment participants and constraints.

In summary, Figure 4(a) shows that the system assembled a *Blood Transmission* event as a cause of Lerato's HIV, given that Lerato was the recipient of a blood transfer that was infected with HIV. It explained the *Blood Transmission* with a possible *Moral Punishment*, and assumed that Lerato committed some *Moral Fault* in the course of instantiating the *Moral Punishment*. The *Moral Punishment* (in red) was not included in the best explanation due to the additional assumption, since this reduces the coherence score (described below). All of our simulation results use this color-coding.

For reference, we contrast the Figure 4(a) explanation structure resulting from relevance-directed beam search with two other (less efficient) explanation-assembly algorithms to characterize the strength of our approach:

- Figure 4(b) illustrates the same situation and explanandum (i.e., blue nodes) using a backward search *without* relevance as a heuristic: it regresses from effects to causes, but tries *all* causes rather than those primed by the situation.
- Figure 4(c) illustrates the same situation and explanandum using exhaustive forward search. This instantiates all applicable events and then repeats.

Neither of these graphs' structure are legible, but we include them to visualize the difference in computation across approaches. Both of these alternative approaches select the exact same final explanation (in green) as the more efficient relevance-directed beam search in Figure 4(a). This suggests that relevance from the situation is a useful heuristic for approximating coherence while assembling explanations in a large space of possible explanatory paths.

These plots also demonstrate that our qualitative models are capable of expressing a wide range of explanations, many of which are incoherent and not employed by people.

This means we have not trivially *"baked in"* the explanation within the knowledge representation; rather, it is the product of assembly and assumption (described above) and coherence assessment, which we describe next.

Scoring explanations for coherence. After assembling explanation structure from model fragments— and potentially making assumptions in the process— the system traverses the explanation structure to select a *best explanation*. This is the culmination of *abductive reasoning* (i.e., inference to the best explanation), which has been formulated as likelihood maximization (Raghavan & Mooney, 2010), simplicity, and other measures of explanation quality (Lombrozo, 2007).

Our system scores explanations by (1) identifying connected causal subgraphs of at least one cause (i.e., of Lerato's *Illness*), (2) scoring those subgraphs for coherence, where larger scores indicate greater coherence, and (3) selecting the highest-scoring subgraph as the best explanation.

The coherence score is the sum of epistemic *features* of a causal graph, where features positively or negatively contribute to coherence. Each feature is scored once for each causal graph, so many model fragment instances can rely on one assumption and incur the cost once. We employ a simple order-of-magnitude scoring technique over these features:

- *Model Fragments* (-1) penalize for increasing complexity.
- *Assumptions* (-10) penalize for increasing complexity.
- *Situation premises* (10) are situation events and entities that participate in model fragments, increasing explanatory inclusion (i.e., coherence) over the stated situation.
- *Causal associations* (100) are presuppositions that associate categories of causes and effects, e.g., *witchcraft* causally contributes to *illness*.
- *Causal dissociations* (-100) are presuppositions that dissociate categories of causes and effects, e.g., *witchcraft* does not cause *physical effects*.

These features coarsely quantify coherence: within-explanation coherence, explanation-to-situation coherence, and explanation-to-presupposition coherence. Following Vosniadou (1994), we model presuppositions as overarching belief-level constraints on people's explanations acquired culturally or via observation. We do not believe our list of is complete, since factors like analogical structure, narrative structure, likelihood, and other factors all contribute to people's explanatory preferences (Lombrozo, 2007).

Simulation

Our simulation setup is a variation of a human experiment by Legare and Gelman (2008): as exemplified in the previous section, we prompt the system to explain how Lerato contracted HIV. We use priming conditions from their study— *biological* priming, *bewitchment* priming, *neither* priming, and *both* types of priming— by varying the information we provide about Lerato. We provide two alternative types of biological priming: sexual intercourse and blood transfer. We also provide a "moral" priming condition, since other results from Legare and Gelman (2008) suggests that some subjects believe immoral behavior can cause illness.

Legare and Gelman (2008) report that 60% to 70% of their subjects exhibited some case of explanatory coexistence, where both supernatural and biological mechanisms (a) explained aspects of the scenario (*target-dependent*); (b) were juxtaposed (*synthetic*); or (c) coexisted in a causal chain (*integrated*). Subjects were sensitive to priming effects: biological and bewitchment priming was associated with more of those mechanisms appearing in explanations. We next review our simulation results for seven priming conditions, shown in the Figure 5 explanation graphs.

Target-dependent reasoning. Graphs (a-e) are all evidence of target-dependent reasoning. Graph (a) is *no priming*, where the system assumes immoral behavior as a simple cause for the disease. Graph (b) is immoral priming, which removes the need for the assumption of immorality. Graph (c) is bewitchment priming, mentioning a practitioner who knows Lerato, which results in assuming an offense, and also considering *Moral Punishment*, but ultimately choosing *Witchcraft* as an explanation. Graph (d) is biological priming with mention of receiving infected blood, resulting in a *Blood Transmission* explanation. Graph (e) is biological priming with mention of sexual intercourse with an HIV-infected partner, resulting in a *Sexual Transmission* explanation, but considering that the illness or the sexual transmission might have been caused by immoral behavior.

Synthetic reasoning. Graph (g) demonstrates one possible example of synthetic reasoning, where a presupposition causally associates *Witchcraft* with *Illness*, and we prime both biological and witchcraft causes. In this case, the *Sexual Transmission* fragment coheres with the situation (i.e., it requires no assumptions), and the *Witchcraft* fragment coheres with the presupposition (rendered in black), so the union of those causes of the illness is higher-scoring than either alone. Our system has no hard constraint to select single causes at causal junctions; however, selecting two causes— when either alone is sufficient— is counter-intuitive. The "synthetic reasoning" category of explanatory coexistence is not as well-specified as the other two, and could plausibly represent multiple sub-strategies, e.g., where subjects integrate causes in parallel, mention multiple salient or competing causes, or are vaguely verbalizing a more integrated causal chain (as below). This suggests further research with human subjects.

Integrated reasoning. Graphs (f) and (h) are evidence of integrated reasoning. Graph (f) is priming of both bewitchment and biology, resulting in an integrated explanation: witchcraft caused the sexual transmission of HIV during the sexual encounter. Graph (h) is priming of moral and biology, resulting in another integrated explanation: immoral behavior caused transmission of HIV during a transfer of blood. Legare and Gelman (2008) did not explicitly attempt the priming condition in Graph (h), but our model predicts that an integrated explanation is plausible for these mechanisms.

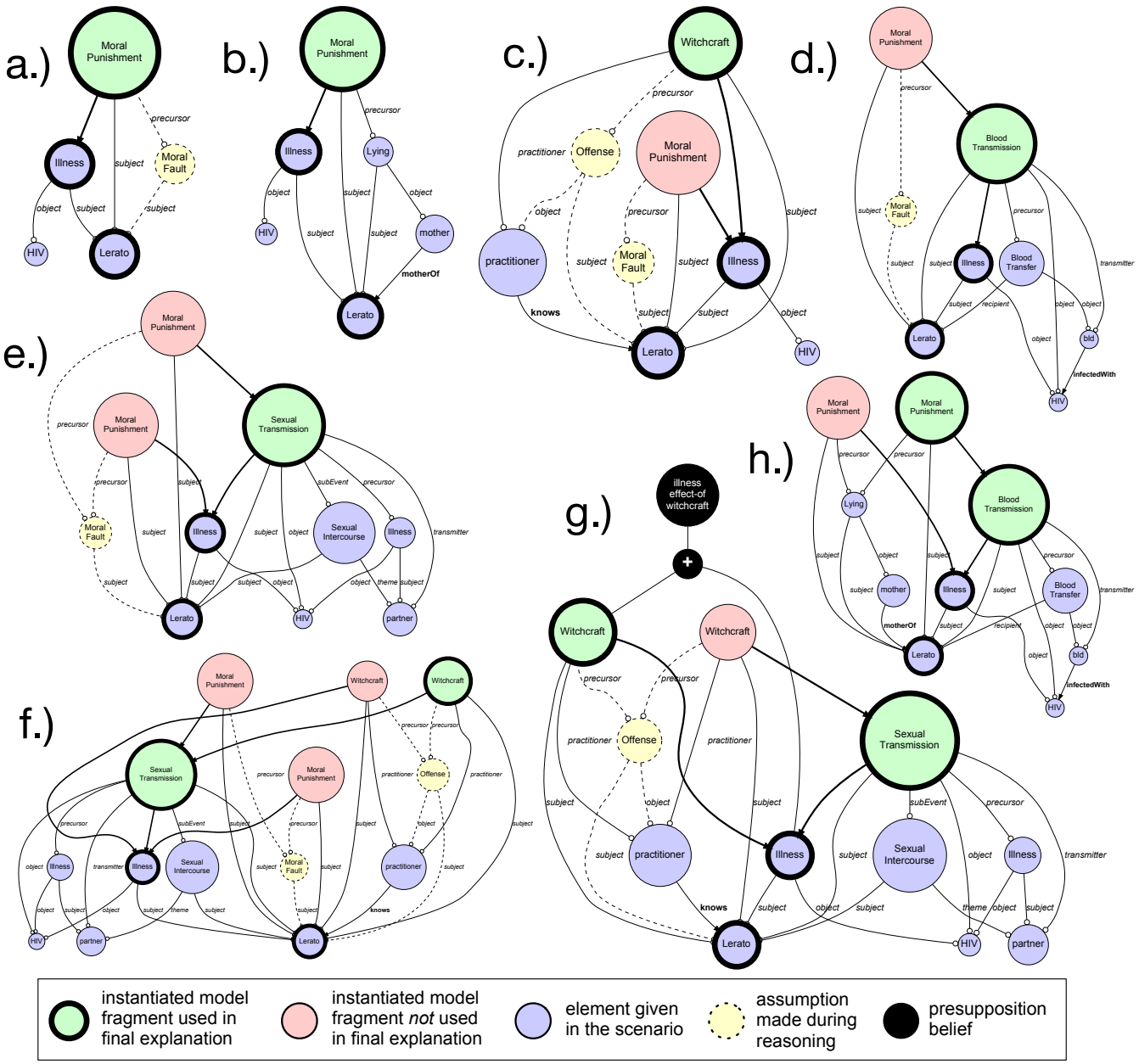


Figure 5: Simulation results with identical causal knowledge, by varying priming: (a) no priming; (b) priming immoral behavior; (c) priming with witchcraft practitioner; (d) priming with blood transfer; (e) priming with sexual encounters; (f) priming with *both* sexual encounters and witchcraft; (g) same priming but including a presupposition that illness is caused by witchcraft; and (h) priming with *both* immorality and blood transfer.

Conclusion

This paper presents a computational cognitive model that simulates all three categories of explanatory coexistence (Legare & Shtulman, 2018; Legare & Gelman, 2008), using the same psychological assumptions as previous models of conceptual change and self-explanation (Friedman et al., 2018). Our computational model retrieves diverse causal model fragments based on relevance to the scenario, and then assembles and evaluates explanations that may integrate both biological and supernatural causes.

We simulated different explanatory coexistence outcomes by varying high-level presuppositions and priming effects; we did *not* vary any causal models, likelihood values, or retrieval parameters across trials. The simulations demonstrate that the model's explanation-assembly is sensitive to priming effects, similar to people (Legare & Gelman, 2008). We showed that salient high-level beliefs— which have been termed *presuppositions* (Vosniadou, 1994)— bias the system to prefer explanations that cohere with their constraints.

Psychological claims and assumptions. This model and simulation support the claims that (1) explanatory coexistence may be the rule rather than the exception (Legare & Shtulman, 2018) and (2) explanatory coherence is a secondary property of assembling and assessing fragmentary, reusable causal knowledge (Friedman et al., 2018). These claims must be framed within the assumptions and limitations of our computational cognitive model.

Our model does not explicitly represent prior probabilities or joint probabilities across causal mechanisms. On the one hand, this allows it to flexibly assemble human-like causal explanations with diverse, seemingly-conflicting mechanisms; however, it could produce uncharacteristic explanations in other domains. This is an empirical question we will investigate in future work, described below. Although we did not encode likelihoods in this work, our model is compatible with Bayesian and statistical relational learning: its situation-specific explanation structure supports statistical inference (Raghavan & Mooney, 2010), and its coherence score could inform likelihood judgments in absence of prior probabilities.

Our model simplifies the psychological processes of knowledge activation and explanation assessment. Some activation and assessment factors *not* modeled here include structural similarity to previous situations, prior likelihood estimates for any given causal mechanism, and probability distributions over causal mechanisms conditionalized on the situation. Implementing these factors would increase the power of our model, but at the expense of interpretability: these factors make additional assumptions about the belief state of each simulated subject, such as their episodic knowledge and the likelihood they ascribe to each causal mechanism.

Future work. In addition to the domain of disease, people's explanatory coexistence has been characterized in the domains of death and human origins (Legare & Shtulman, 2018). Simulating these domains will provide additional empirical evidence of our model's generality.

In addition to other domains, running this model of explanation on other explanation tasks will help qualify its broader psychological plausibility. Also, applying this model of explanation within larger models of explanation-based learning and conceptual change will help us refine the model's parameters knowledge representations.

Finally, this paper's simulations utilized a purely qualitative comparison between human and machine explanations as a proof of concept, but we plan to model quantitative properties, such as subjects' reaction time, in future work.

Acknowledgments

We acknowledge CogSci reviewers for their insightful reviews, and we thank Cristine Legare and Andrew Shtulman for helpful discussions on the direction of this work.

References

Charniak, E., & Shimony, S. (1994). Cost-based abduction and MAP explanation. *Artificial Intelligence*, *66*, 345–374.

Chi, M. T., De Leeuw, N., Chiu, M.-H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive science*, *18*(3), 439–477.

Crestani, F. (1997). Application of spreading activation techniques in information retrieval. *Artificial Intelligence Review*, *11*(6), 453–482.

diSessa, A. A., & Sherin, B. L. (1998). What changes in conceptual change? *International journal of science education*, *20*(10), 1155–1191.

Falkenhainer, B., & Forbus, K. D. (1991). Compositional modeling: finding the right model for the job. *Artificial Intelligence*, *51*(1-3), 95–143.

Friedman, S. E., Barbella, D., & Forbus, K. (2012). Revising domain knowledge with cross-domain analogy. *Advances in Cognitive Systems*, *2*, 13–24.

Friedman, S. E., Forbus, K., & Sherin, B. (2018). Representing, running, and revising mental models: A computational model. *Cognitive Science*, *42*(4), 1110–1145.

Friedman, S. E., & Forbus, K. D. (2010). An integrated systems approach to explanation-based conceptual change. In *Proceedings of AAAI 2010* (pp. 1523–1529).

Friedman, S. E., & Forbus, K. D. (2011). Repairing incorrect knowledge with model formulation and metareasoning. In *Proceedings of ijcai 2011* (Vol. 22, pp. 887–893).

Gentner, D., & Stevens, A. L. (1983). *Mental models*. Psychology Press.

Goel, A., Rugaber, S., & Vattam, S. (2009). Structure, behavior, & function of complex systems: The structure, behavior, & function modeling language. *AI EDAM*, *23*, 23–35.

Legare, C. H., & Gelman, S. A. (2008). Bewitchment, biology, or both: The co-existence of natural and supernatural explanatory frameworks across development. *Cognitive Science*, *32*(4), 607–642.

Legare, C. H., & Shtulman, A. (2018). Explanatory pluralism across cultures and development. *Metacognitive Diversity: An Interdisciplinary Approach*, 415.

Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive psychology*, *55*(3), 232–257.

Poole, D. (1993). Probabilistic horn abduction and bayesian networks. *Artificial intelligence*, *64*(1), 81–129.

Pustejovsky, J. (2013). Dynamic event structure and habitat theory. In *Proceedings of the 6th international conference on generative approaches to the lexicon* (pp. 1–10).

Raghavan, S., & Mooney, R. J. (2010). Bayesian abductive logic programs. In *Statistical relational artificial intelligence* (pp. 82–87).

Shtulman, A., & Lombrozo, T. (2016). Bundles of contradiction: A coexistence view of conceptual change. *Core knowledge and conceptual change*, 49–67.

Vosniadou, S. (1994). Capturing and modeling the process of conceptual change. *Learning and instruction*, *4*(1), 45–69.

Stereotypes of Transgender Categories: Attributes and Lay Theories

Natalie Gallagher
Northwestern University, Evanston, Illinois, United States

GALEN BODENHAUSEN
Northwestern University, Evanston, Illinois, United States

Abstract

What is the descriptive content and guiding lay theory of transgender stereotypes? The recent rise in public visibility and the numeric minority of this gender group make this an opportunity to understand not only the content of stereotypes applied to transgender individuals today, but also the ontology of gender guiding the content of these stereotypes. Using convergent methods, we measure the descriptive content of transgender stereotypes and assess the role of essentialist beliefs in guiding that content. We show that transgender categories are perceived less positively than cisgender categories, and that while perceptions of cisgender men and women differ sharply, those of transgender men and women show striking similarity. Essentialist beliefs about gender exaggerate these patterns.

Incorrect Guesses Boost Retention of Novel Words in Adults but not in Children

Chiara Gambi
Cardiff University, Cardiff, United Kingdom

Martin J. Pickering
University of Edinburgh, Edinburgh, United Kingdom

Hugh Rabagliati
University of Edinburgh, Edinburgh, United Kingdom

Abstract

What is the mechanism by which linguistic knowledge is updated over time? In six experiments, we asked whether error-driven learning can explain how adults and children add new words to their vocabulary. Participants were exposed to novel object labels that were more or less unexpected given participants linguistic knowledge. Two-to-four-year-olds were strongly affected by expectations based on contextual constraint when choosing the referent of a new label. However, while adults formed stronger memory traces for novel words that violated a stronger prior expectation, childrens memory was unaffected by the strength of their prior expectations. We conclude that the encoding of new words in memory follows the principles of error-driven learning in adults, but not in preschoolers.

Sleep Does not Help Relearning Declarative Memories in Older Adults

Emilie Gerbier (emilie.gerbier@univ-cotedazur.fr)
Université Côte d'Azur, CNRS, BCL, France,
24 avenue des Diables Bleus, 06357 Nice Cedex 4, France

Guillaume T. Vallet (guillaume.vallet@uca.fr)
Université Clermont Auvergne, LaPSCo (UMR CNRS 6024)
17, rue Paul Collomp, 63037 Clermont-Ferrand, France

Thomas C. Toppino (thomas.toppino@villanova.edu)
Department of Psychological & Brain Sciences, Villanova University
800 Lancaster Avenue, Villanova, PA 19085, USA

Stéphanie Mazza (stephanie.mazza@univ-lyon1.fr)
Université Lyon 1, CRNL, INSERM U1028, HESPER
Hôpital Neurologique, 59 Boulevard Pinel, 69677 Bron, France

Abstract

How sleep affects memory in older adults is a critical topic, since age significantly impacts both sleep and memory. For declarative memory, previous research reports contradictory results, with some studies showing sleep-dependent memory consolidation and some other not. We hypothesize that this discrepancy may be due to the use of recall as the memory measure, a demanding task for older adults. The present paper focuses on the effect of sleep on relearning, a measure that proved useful to reveal subtle, implicit memory effects. Previous research in young adults showed that sleeping after learning was more beneficial to relearning the same Swahili-French word pairs 12 hours later, compared with the same interval spent awake. In particular, those words that could not be recalled were relearned faster when participants previously slept. The effect of sleep was also beneficial for retention after a one-week and a 6-month delay. The present study used the same experimental design in older adults aged 71 on average but showed no significant effect of sleep on consolidation, on relearning, or on long-term retention. Thus, even when using relearning speed as the memory measure, the consolidating effect of sleep in older adults was not demonstrated, in alignment with some previous findings.

Keywords: sleep-dependent memory consolidation; ageing; learning; relearning; repeated practice

Sleep, Memory, and Age

The importance of sleep in cognitive functioning is now well established. For instance, sleep has been shown to benefit the consolidation of declarative memories acquired during the day. It is usually observed that sleeping after a learning episode improves recall performance during a test compared to the same delay without sleep (see Rasch & Born, 2013, for a comprehensive review). Such memory consolidation is thought to originate from two complementary processes. First, the reactivation or "replay" of recent memory traces during slow-wave sleep (SWS) causes declarative memories that are initially hippocampal-dependent to become increasingly dependent on the prefrontal cortex (Takashima et al., 2009). In parallel, a downscaling process during sleep leads to the recalibration of synaptic connections that were modified during learning (Tononi & Cirelli, 2014). The overnight memory improvement has been clearly demonstrated in young adults but is more controversial in older adults (Harand et al., 2012).

Sleep undergoes both quantitative and qualitative changes over the course of ageing. The most obvious change in sleep in healthy older adults is a decrease in total sleep time (TST) induced by an increase in the time spent awake both after bedtime (i.e., sleep latency) and during the night (Carrier et al., 1997). As a result, sleep efficiency (i.e., ratio TST / time in bed) declines and reaches 79% or less at age 70 (Bliwise, et al., 2005) compared to about 90% in healthy adults. Sleep architecture is also modified, with a reduction in SWS (Lombardo et al., 1998), whereas time in lighter stages (non-Rapid Eye Movement nREM1 and nREM2) increases (Ohayon et al., 2004). However, the time spent in REM sleep remains relatively unchanged. Modifications in sleep microstructure are also observed. A reduced spectral EEG power is observed, in particular for slow-wave activity (SWA) during SWS, especially over the prefrontal cortex (Dubé et al., 2015). Spindle density, frequency, and duration are also diminished (Crowley et al., 2002; Guazzelli et al., 1986), as well as the density of phasic REM phases (Darchia, Campbell, & Feinberg, 2003). Thus, many sleep components underpinning sleep-related memory consolidation in young adults, such as SWA and spindle activity, are reduced with increasing age (Carrier et al., 2011; Martin et al., 2013).

Aging is also associated with declarative memory changes. Overall, older adults exhibit poorer memory recall while their recognition performance remains relatively spared (e.g. Danckert & Craik, 2013; see Grady & Craik, 2000 for review). They also exhibit an increasing difficulty to learn new material compared to the younger adults, because of poor encoding strategies (see Craik & Rose, 2012, for review). These effects might be mainly explained by an impairment in forming associations (Old & Naveh-Benjamin, 2008).

It has been proposed that changes in sleep may contribute to age-related memory impairments (Buckley & Schatzberg, 2005; Hornung, Danker-Hopfe, & Heuser, 2005; Mander et al., 2014; see Scullin & Bliwise, 2015). A proposed mechanism (Mander et al., 2013) is that reduction of gray matter volume associated with aging impedes the generation of SWA and spindles during sleep, thereby impairing memory consolidation.

Despite these age-related changes in sleep, some research reported the persistence of sleep-dependent memory consolidation in the older adult (Aly & Moscovitch, 2010; Sonni & Spencer, 2015; Wilson et al., 2012), whereas others failed to demonstrate it or demonstrated a lesser consolidation compared to young adults (Baran, Mantua, & Spencer, 2016; Cherdieu et al., 2014; Mander et al., 2013, 2014; Mary, Schreiner, & Peigneux, 2013; Scullin, 2013; Scullin et al., 2017). Other research described a more complex picture (Jones et al., 2016).

Most of these studies used free- or cued-recall performance as a measure of consolidation. Such measures may be suboptimal to reveal consolidation and may thus contribute to the contradictory findings, especially since recall is a demanding task specifically impaired in the older adult (e.g. Troyer, Graves, & Cullum, 1994; Danckert & Craik, 2013). In addition, these measures are binary (i.e., correct/incorrect) and are poorly informative as to the strength of a memory trace. It seems possible to assess memory retention in an alternative, more implicit manner by measuring how fast participants relearn the information (i.e., savings in relearning). In a recent study, Mazza et al. (2016) found that young adults significantly benefitted from sleep to improve their memory performance in both relearning and retention of word pairs. The Sleep group, who slept during the 12-hour interval between the learning and the relearning sessions, displayed a faster rate of relearning compared to the Wake group that did not sleep. This was true even after controlling the recall performance just before relearning. In other words, those words that could not be recalled before the relearning session were relearned faster. The present study consists of a close replication of the study by Mazza et al. (see Figure 1) with older adults and using the same type of material. We tested the hypothesis that sleep does still favor consolidation in older adults, as evidenced when relearning speed is measured instead of recall.

Method

Participants

Forty French healthy participants completed the study. They were aged between 65 and 80 and had normal sleep and cognitive abilities. Participants with sleep problems, as assessed by a score of 8 or above on the Pittsburgh Sleep Quality Index (PSQI; Buysse et al., 1989), or presenting altered cognition, as assessed by a score below 27 in the Mini-Mental State Examination (MMSE; Folstein, Folstein, & McHugh, 1975), were excluded. Health information was gathered from all participants during an interview about their medical history and medication. Participants with a medical history and/or taking medications with known sensory or neurological effects were also excluded. All participants in the study were native French-speaking and present normal or corrected-to-normal vision. The present study was in accordance with the ethical standards of the responsible committee on human experimentation of the Helsinki Declaration of 1975, as revised in 2000. Informed written consent was obtained from each participant.

Furthermore, all participants underwent a neuro-psychological and sleep assessment with sleep quality (PSQI), circadian topology (the Horne and Ostberg morning/evening questionnaire; Horne & Ostberg, 1976), level of sleepiness (Epworth Sleepiness Scale; Johns, 1991), basic long-term and short-term memory capacity (subtests from the Wechsler Adult Intelligence Scale IV, Wechsler, 2008, and from the Wechsler Memory Scale III; Wechsler, 1997), global cognitive ability (MMSE), and anxiety and depression levels (The Hospital Anxiety and Depression Scale; Zigmond & Snaith, 1983).

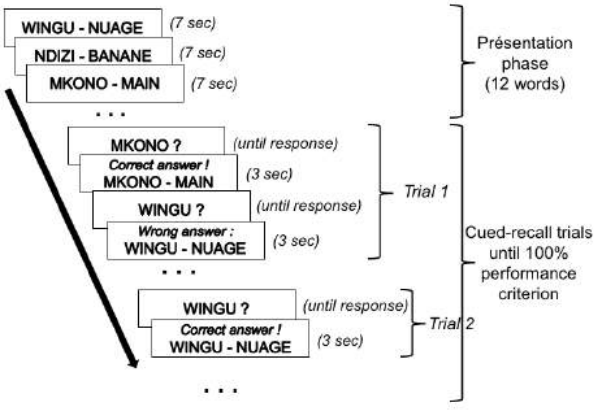
The participants were randomly assigned to the Sleep or the Wake group. Three participants did not pursue the experiment after having spent too much time completing the learning session. Eventually, the data from 19 participants in the Sleep and 18 in the Wake group were included for analysis. Their age ranged from 65 to 80 (mean age 71.6 +/- SE) with 20 women in total. The Sleep and Wake groups did not significantly differ on any of the following variables: gender, age, number of attended school years, sleep quality, circadian topology, level of sleepiness, basic long-term and short-term memory capacity, global cognitive ability, and anxiety and depression levels. An additional 6 participants started but did not complete the study for various reasons (e.g., not available for the upcoming session).

Material and Procedure

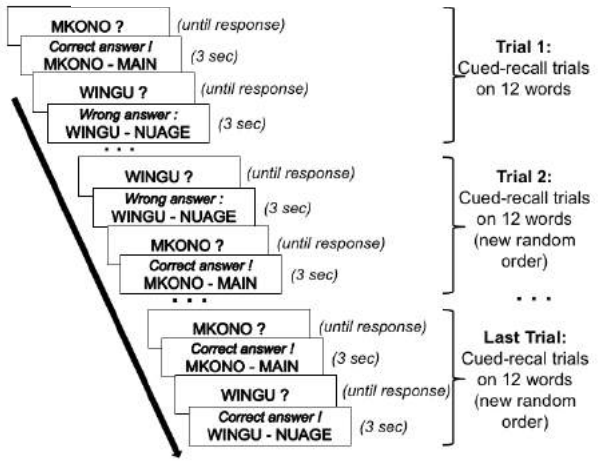
During the first session, participants were trained to learn the French translation of 12 Swahili words (e.g., nyanya-tomate), using repeated tests with feedback. The number of pairs was decreased from 16 in Mazza et al. (2016) to 12 here to adjust to the memory difficulties encountered by the older adult. A perfect learning criterion was adopted in which pairs were tested until they received a correct answer (Figure 1A). Twelve hours later, in the second (relearning) session, they

had to relearn the pairs to the criterion of correctly answering the 12 items in a row (Figure 1B). This equated the participants' performance at the end of the relearning session. Initial performance on the first trial for the 12 items and the number of trials necessary to attain the relearning criterion were measured.

The Wake group performed the learning session at 9:00 a.m. and the relearning session at 9:00 p.m. the same day (Figure 2). They did not sleep between the two sessions, as was instructed. The Sleep group performed the learning session at 9:00 p.m. and the relearning session at 9:00 a.m. the following day. They experienced a night of sleep between the two sessions, during which actimetry (Actiwatch system, CamNtech, Cambridge, UK) was used to quantify TST. One week and six months after the relearning session, the retention of the material was further assessed for each group using a cued-recall task without feedback.



B. Learning session



A. Relearning session

Figure 1. Procedure for the Learning (A) and Relearning (B) sessions.

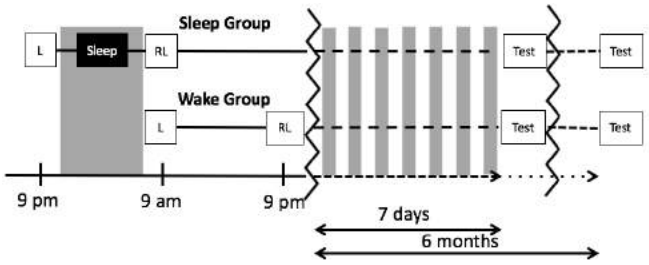


Figure 2. Temporal description of the experimental design. L = Learning session. RL= Relearning session. Nights are represented by grey areas.

Results

Based on the results obtained in the young adults in Mazza et al. (2016), it was expected that the Sleep group would start the relearning session with an advantage over the Wake group, would relearn faster, and would perform better after one week and six months. The overall results are presented in Figure 3.

Along with inferential frequentist statistics (two-tailed Student's *t-tests*) using a critical *p*-value of .05 and effect sizes reported using Cohen's *d*, dependent variables were also submitted to two-tailed Bayesian *t-tests* (Rouder, et al., 2009) comparing the Wake and the Sleep groups. Because we observed weak effects, we will present the BF_{01}, that is, the odds ratios in favor of the null hypothesis H_0 (i.e., no difference between the means) against the alternative hypothesis (i.e., a difference between the means). They were computed using JASP (JASP software, 2016) and will be considered according to the following scale: values inferior to 3 as anecdotal evidence, values ranging from 3 to 10 as substantial evidence, and values above 10 as strong evidence in favor of H_0, with higher values indicating gradually increasing confidence (Jarosz & Wiley, 2014; Jeffreys, 1961).

The learning session was performed similarly by the two groups with respect to the proportion of correct answers provided at the first trial ($M = 0.21$, $SE = 0.04$ in both groups). The estimated Bayes factor ($BF_{01} = 3.14$) suggested that the data were 3.14 times more likely in favor of H_0 than of the alternative hypothesis, i.e., indicative of anecdotal to substantial evidence in favor of the absence of a difference. The number of trials necessary to achieve the learning criterion was not significantly different (Sleep: $M = 5.79$, $SE = 0.41$; Wake: $M = 6.00$, $SE = 0.45$; $t(35) = 0.35$, $p = .73$), with an anecdotal evidence in favor of H_0 ($BF_{01} = 3$).

During the relearning session, the proportion of correct answers on the first trial did not differ in the Sleep ($M = 0.46$, $SE = 0.04$) and in the Wake group ($M = 0.40$, $SE = 0.05$; $t(35) = 0.93$, $p = .36$, Cohen's $d = 0.31$, $BF_{01} = 2.23$, anecdotal evidence in favor of H_0). Both groups needed an equivalent number of trials to achieve the relearning criterion (Sleep: $M = 7.63$, $SE = 0.76$; Wake: $M = 8.00$, $SE = 1.16$; $t(35) = 0.27$, $p = 0.79$; $d = 0.09$; $BF_{01} = 3$, anecdotal evidence in favor of H_0). The *relearning speed* was computed by dividing the number of unrecalled items at the first trial by the number of

trials necessary to complete the session. This was a way to control for the influence of initial performance on relearning speed. Indeed, the more the participants initially recalled, the fewer items remained to be relearned. The relearning speed did not differ between the Sleep (M = 0.96, SE = 0.09) and the Wake group (M = 1.06, SE = 0.08, $t(35) = 0.78$, p = .44; d = 0.26; BF_{01} = 2.5, anecdotal evidence in favor of H_0). Thus, contrary to our expectation, there was no clear-cut indication of a consolidating effect of sleep, since neither the initial retrieval performance nor the relearning speed varied between the two groups.

After one week, there was no significant difference between the performance in the Sleep (M = .75, SE = .04) and the Wake group (M = .64, SE = .06; $t(35) = 1.59$, p = .12; d = .52; BF_{01} = 1.19, anecdotal evidence in favor of H_0). In addition, these scores were overall relatively high, indicating that the specific relearning method used was quite efficient at inducing long-term retention. For the six-month delay, the data from 4 participants in the Sleep group and from 2 participants in the Wake group could not be obtained. There was no significant difference between the groups (Sleep: M = .42, SE = .08; Wake: M = .30, SE = .06; $t(29) = 1.22$, p =.23; d = .44; BF_{01} = 2.35, anecdotal evidence in favor of H_0).

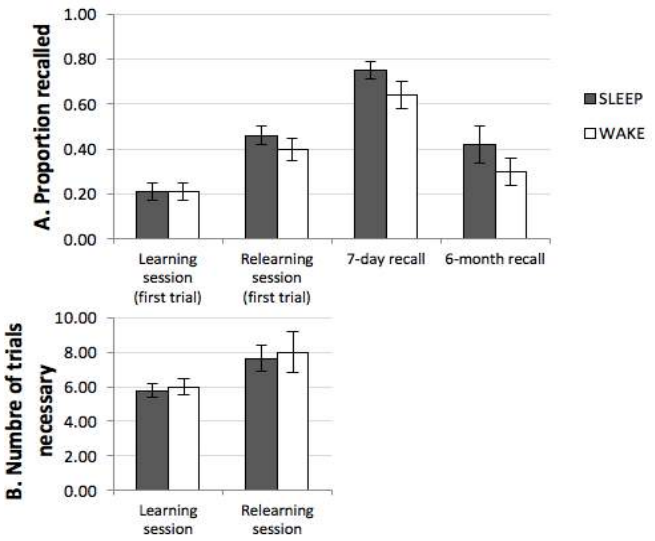


Figure 3: Upper part: Mean proportion correct in the first trial of the learning and relearning sessions, and in the 7-day and 6-month recall sessions.
Lower part: Number of trials necessary to achieve the criterion in the learning and the relearning session.
The error bars represent standard errors of the mean.

The actimetric data was lost for 4 participants in the Sleep group. The Sleep group (N = 15) exhibited a mean TST of 402 min (6 hours and 42 minutes; 308 - 469 min; SE = 12.1). The correlation between TST and initial performance on the relearning session was not significant (r = .40; df = 13; p = .14). The correlation between TST and the number of trials needed at the relearning session was surprisingly positive but not significant either (r = .45, p = .09). In addition, the correlation between TST and the relearning speed was significant and negative (r = -.74, p = .0015), indicating that the more the participants slept, the slower they were to relearn. Finally, there was no correlation between TST and the performance after one week (r = -.03, p = .92).

Discussion

The present results display major differences from those of the Mazza et al. study that used similar methods and sample sizes (n=20 in each group) with younger adults. In the present study with older adults, an episode of sleep did not significantly boost subsequent recall 12 hours after learning, compared to wakefulness, Thus, contrary to our hypotheses, even with more fine-grained and implicit measures of declarative memory, we did not show significant benefits of sleep on relearning in this population. The effect on the long term, however, is more ambiguous, with a potential weak benefit of interpolated sleep after 7 days and 6 months, which is not significant in the present study, most likely due to lack of power. In any case, the size of this benefit is far weaker in the older than in the younger adults.

Our hypothesis that the sleep-dependent processes that consolidate memories are subtle in the older population and therefore require more implicit measures to be shown was not validated. These results are however consistent with other studies that did not show any sleep-dependent benefit for declarative memory in the older adult (see Gui et al., 2017 for a review). Contrary to younger adults, recommendations such as "you should sleep between learning and relearning" does not seem as relevant for the aging population, although it does not seem detrimental either (especially for long-term retention).

One surprising finding is the negative correlation between TST and the relearning speed, indicating that the more the participants slept, the more trials they needed in order to reach the relearning criterion. This is in contradiction with results from Aly & Moscovitch (2010) but is consistent with that of Scullin (2013) and Tsapanou et al. (2017). A possibility is that people with poorly efficient sleep with respect to consolidation and maybe other functions tend to sleep more to compensate.

How to reconcile these results with those indicating clear sleep-dependent consolidating effects of sleep in the older adult (e.g., Aly & Moscovitch, 2010; Sonni & Spencer, 2015; Wilson et al., 2012)? A first limit might be that the task was too difficult and not well calibrated for the participants, therefore impeding the emergence of potential effects of sleep. However, the relatively high scores observed after one week go against this limitation. Another limit could be that the criteria of perfect-performance during learning and relearning did not leave enough possibility for sleep-dependent improvement. However, the relatively low performance at the beginning of relearning and the fact that the paradigm was identical to that in Mazza et al.'s study go against this limitation. Finally, the discrepancies could be linked to the material used. Our study consisted of learning

pseudowords which are verbal representations that do not yet exist in the mental lexicon and need to be integrated into it. This integration has been shown to require time (e.g., Dumay & Gaskell, 2007). Moreover, in the present study, in addition to creating such a novel verbal representation, participants needed to associate it to a known French word in order to succeed at the cued-recall task. Such associative learning is impaired in aging (e.g., Service & Craik, 1993). Therefore, it would be interesting to examine in future research whether the absence of sleep-induced benefits would also be observed when older adults learn and relearn word pairs instead of pseudoword-word pairs (see Kurdziel, Mantua, & Spencer, 2017 for an overall review of the effect of sleep on word learning in the older adult).

Quality of sleep and memory performance are critical issues in modern societies, especially for older people. Their functional relationship needs to be investigated further (Scullin & Bliwise, 2015). In particular, an intriguing issue is whether improving sleep quality could efficiently improve memory functioning. Such possibility could lead to potential practical applications for improving the aging population's quality of life.

References

Aly, M., & Moscovitch, M. (2010). The effects of sleep on episodic memory in older and younger adults. *Memory*, 18(3), 327-334. doi : 10.1080/09658211003601548

Baran, B., Mantua, J., & Spencer, R. M. C. (2016). Age-related Changes in the Sleep-dependent Reorganization of Declarative Memories. *Journal of Cognitive Neuroscience*, 28(6), 792-802. doi: 10.1162/jocn_a_00938

Bliwise, D. L., Ansari, F. P., Straight, L.-B., & Parker, K. P. (2005). Age Changes in Timing and 24-Hour Distribution of Self-Reported Sleep. *The American Journal of Geriatric Psychiatry*, 13(12), 1077-1082.

Buckley, T. M., & Schatzberg, A. F. (2005). Aging and the Role of the HPA Axis and Rhythm in Sleep and Memory-Consolidation. *The American Journal of Geriatric Psychiatry, 13*(5), 344-352. doi: 10.1097/00019442-200505000-00002

Buysse, D. J., Reynolds, C. F., Monk, T. H., Berman, S. R., & Kupfer, D. J. (1989). The Pittsburgh sleep quality index: A new instrument for psychiatric practice and research. *Psychiatry Research, 28*(2), 193-213.

Carrier, J., Monk, T. H., Buysse, D. J., & Kupfer, D. J. (1997). Sleep and morningness-eveningness in the 'middle' years of life (20–59y). *Journal of Sleep Research, 6*(4), 230-237. doi: 10.1111/j.1365-2869.1997.00230.x

Carrier, J., Viens, I., Poirier, G., Robillard, R., Lafortune, M., Vandewalle, G., … Filipini, D. (2011). Sleep slow wave changes during the middle years of life. *European Journal of Neuroscience, 33*(4), 758-766. doi: 10.1111/j.1460-9568.2010.07543.x

Cherdieu, M., Reynaud, E., Uhlrich, J., Versace, R., & Mazza, S. (2014). Does age worsen sleep-dependent memory consolidation? *Journal of Sleep Research, 23*(1), 53-60. doi: 10.1111/jsr.12100

Craik, F. I. M., & Rose, N. S. (2012). Memory encoding and aging: A neurocognitive perspective. *Neuroscience & Biobehavioral Reviews, 36*(7), 1729-1739. doi: 10.1016/j.neubiorev.2011.11.007

Crowley, K., Trinder, J., Kim, Y., Carrington, M., & Colrain, I. M. (2002). The effects of normal aging on sleep spindle and K-complex production. *Clinical Neurophysiology, 113*(10), 1615-1622.doi: 10.1016/S1388-2457(02)00237-7

Danckert, S. L., & Craik, F. I. M. (2013). Does aging affect recall more than recognition memory? *Psychology and Aging, 28*(4), 902-909. doi: 10.1037/a0033263

Darchia, N., Campbell, I. G., & Feinberg, I. (2003). Rapid Eye Movement Density is Reduced in the Normal Elderly. *Sleep, 26*(8), 973-977. doi: 10.1093/sleep/26.8.973

Dubé, J., Lafortune, M., Bedetti, C., Bouchard, M., Gagnon, J. F., Doyon, J., … Carrier, J. (2015). Cortical Thinning Explains Changes in Sleep Slow Waves during Adulthood. *Journal of Neuroscience, 35*(20), 7795-7807. doi: 10.1523/JNEUROSCI.3956-14.2015

Dumay, N., & Gaskell, M. G. (2007). Sleep-Associated Changes in the Mental Representation of Spoken Words. *Psychological Science, 18*(1), 35-39. doi: 10.1111/j.1467-9280.2007.01845.x

Folstein, M., Folstein, S., & McHugh, P. (1975). « Mini-mental state ». A practical method for grading the cognitive state of patients for the clinician. *Journal of psychiatric research, 12*(3), 189-198.

Grady, C. L., & Craik, F. I. (2000). Changes in memory processing with age. *Current Opinion in Neurobiology, 10*(2), 224-231.

Guazzelli, M., Feinberg, I., Aminoff, M., Fein, G., Floyd, T. C., & Maggini, C. (1986). Sleep spindles in normal elderly: comparison with young adult patterns and relation to nocturnal awakening, cognitive function and brain atrophy. *Electroencephalography and Clinical Neurophysiology, 63*(6), 526-539. doi: 10.1016/0013-4694(86)90140-9

Gui, W.-J., Li, H.-J., Guo, Y.-H., Peng, P., Lei, X., & Yu, J. (2017). Age-related differences in sleep-based memory consolidation: A meta-analysis. *Neuropsychologia, 97*, 46-55. doi: 10.1016/j.neuropsychologia.2017.02.001

Harand, C., Bertran, F., Doidy, F., Guénolé, F., Desgranges, B., Eustache, F., & Rauchs, G. (2012). How Aging Affects Sleep-Dependent Memory Consolidation? *Frontiers in Neurology, 3*. doi: 10.3389/fneur.2012.00008

Horne, J. A., & Östberg, O. (1976). A self-assessment questionnaire to determine morningness-eveningness in human circadian rhythms. *International Journal of Chronobiology, 4*, 97-110.

Hornung, O. P., Danker-Hopfe, H., & Heuser, I. (2005). Age-related changes in sleep and memory: commonalities and interrelationships. *Experimental Gerontology, 40*(4), 279-285. doi: 10.1016/j.exger.2005.02.001

Jarosz, A. F., & Wiley, J. (2014). What Are the Odds? A Practical Guide to Computing and Reporting Bayes Factors. *The Journal of Problem Solving, 7*(1). doi: 10.7771/1932-6246.1167

JASP Team. (2016). Jasp. Version 0.8. 0.0. Software.

Jeffreys, H. (1961). *Theory of probability* (3rd Ed.). Oxford, UK: Oxford University Press.

Johns, M. W. (1991). A New Method for Measuring Daytime Sleepiness: The Epworth Sleepiness Scale. *Sleep, 14*(6), 540-545. doi: 10.1093/sleep/14.6.540

Jones, B. J., Schultz, K. S., Adams, S., Baran, B., & Spencer, R. M. C. (2016). Emotional bias of sleep-dependent processing shifts from negative to positive with aging. *Neurobiology of Aging, 45*, 178-189. doi: 10.1016/j.neurobiolaging.2016.05.019

Kurdziel, L. B. F., Mantua, J., & Spencer, R. M. C. (2017). Novel word learning in older adults: A role for sleep? *Brain and Language, 167*, 106-113. doi: 10.1016/j.bandl.2016.05.010

Lombardo, P., Formicola, G., Gori, S., Gneri, C., Massetani, R., Murri, L., … Salzarulo, P. (1998). Slow wave sleep (SWS) distribution across night sleep episode in the elderly. *Aging Clinical and Experimental Research, 10*(6), 445-448. doi: 10.1007/BF03340157

Mander, B. A., Rao, V., Lu, B., Saletin, J. M., Ancoli-Israel, S., Jagust, W. J., & Walker, M. P. (2014). Impaired Prefrontal Sleep Spindle Regulation of Hippocampal-Dependent Learning in Older Adults. *Cerebral Cortex, 24*(12), 3301-3309. doi: 10.1093/cercor/bht188

Mander, B. A., Rao, V., Lu, B., Saletin, J. M., Lindquist, J. R., Ancoli-Israel, S., … Walker, M. P. (2013). Prefrontal atrophy, disrupted NREM slow waves and impaired hippocampal-dependent memory in aging. *Nature Neuroscience, 16*(3), 357-364. doi: 10.1038/nn.3324

Martin, N., Lafortune, M., Godbout, J., Barakat, M., Robillard, R., Poirier, G., … Carrier, J. (2013). Topography of age-related changes in sleep spindles. *Neurobiology of Aging, 34*(2), 468-476. doi: 10.1016/j.neurobiolaging.2012.05.020

Mary, A., Schreiner, S., & Peigneux, P. (2013). Accelerated long-term forgetting in aging and intra-sleep awakenings. *Frontiers in Psychology, 4*. doi: 10.3389/fpsyg.2013.00750

Mazza, S., Gerbier, E., Gustin, M.-P., Kasikci, Z., Koenig, O., Toppino, T. C., & Magnin, M. (2016). Relearn Faster and Retain Longer Along With Practice, Sleep Makes Perfect. *Psychological Science, 27*(10), 1321-1330. doi: 10.1177/0956797616659930

Ohayon, M. M., Carskadon, M. A., Guilleminault, C., & Vitiello, M. V. (2004). Meta-Analysis of Quantitative Sleep Parameters From Childhood to Old Age in Healthy Individuals: Developing Normative Sleep Values Across the Human Lifespan. *Sleep, 27*(7), 1255-1273.

Old, S. R., & Naveh-Benjamin, M. (2008). Memory for people and their actions: further evidence for an age-related associative deficit. *Psychology and Aging, 23*(2), 467-472. doi: 10.1037/0882-7974.23.2.467

Rasch, B., & Born, J. (2013). About Sleep's Role in Memory. *Physiological Reviews, 93*(2), 681-766. doi: 10.1152/physrev.00032.2012

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*(2), 225-237. doi: 10.3758/PBR.16.2.225

Scullin, M. K. (2013). Sleep, memory, and aging: The link between slow-wave sleep and episodic memory changes from younger to older adults. *Psychology and Aging, 28*(1), 105-114. doi: 10.1037/a0028830

Scullin, M. K., & Bliwise, D. L. (2015). Sleep, Cognition, and Normal Aging: Integrating a Half Century of Multidisciplinary Research. *Perspectives on Psychological Science, 10*(1), 97-137. doi: 10.1177/1745691614556680

Scullin, M. K., Fairley, J., Decker, M. J., & Bliwise, D. L. (2017). The Effects of an Afternoon Nap on Episodic Memory in Young and Older Adults. *Sleep, 40*(5). doi: 10.1093/sleep/zsx035

Service, E., & Craik, F. I. M. (1993). Differences Between Young and Older Adults in Learning A Foreign Vocabulary. *Journal of Memory and Language, 32*(5), 608-623. doi: 10.1006/jmla.1993.1031

Sonni, A., & Spencer, R. M. C. (2015). Sleep protects memories from interference in older adults. *Neurobiology of Aging, 36*(7), 2272-2281. doi: 10.1016/j.neurobiolaging.2015.03.010

Takashima, A., Nieuwenhuis, I. L. C., Jensen, O., Talamini, L. M., Rijpkema, M., & Fernández, G. (2009). Shift from Hippocampal to Neocortical Centered Retrieval Network with Consolidation. *Journal of Neuroscience, 29*(32), 10087-10093. doi: 10.1523/JNEUROSCI.0799-09.2009

Tononi, G., & Cirelli, C. (2014). Sleep and the Price of Plasticity: From Synaptic and Cellular Homeostasis to Memory Consolidation and Integration. *Neuron, 81*(1), 12-34. doi: 10.1016/j.neuron.2013.12.025

Troyer, A. K., Graves, R. E., & Cullum, C. M. (1994). Executive functioning as a mediator of the relationship between age and episodic memory in healthy aging. *Aging, Neuropsychology, and Cognition, 1*(1), 45-53.

Tsapanou, A., Gu, Y., O'Shea, D. M., Yannakoulia, M., Kosmidis, M., Dardiotis, E., … Scarmeas, N. (2017). Sleep quality and duration in relation to memory in the elderly: Initial results from the Hellenic Longitudinal Investigation of Aging and Diet. *Neurobiology of Learning and Memory, 141*, 217-225. doi: 10.1016/j.nlm.2017.04.011

Wechsler, D. (1997). *Wechsler Memory Scale–Third Edition*. San Antonio, TX: Psychological Corp.

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale–Fourth Edition*. San Antonio, TX: Pearson.

Wilson, J. K., Baran, B., Pace-Schott, E. F., Ivry, R. B., & Spencer, R. M. C. (2012). Sleep modulates word-pair learning but not motor sequence learning in healthy older adults. *Neurobiology of Aging, 33*(5), 991-1000. doi: 10.1016/j.neurobiolaging.2011.06.029

Zigmond, A. S., & Snaith, R. P. (1983). The hospital anxiety and depression scale. *Acta psychiatrica scandinavica,* 67(6), 361-370.

At the Zebra Crossing: Modelling Complex Decision Processes with Variable-Drift Diffusion Models

Oscar Giles (o.t.giles@leeds.ac.uk)
Institute for Transport Studies and School of Psychology, University of Leeds, LS2 9JT, Leeds, United Kingdom

Gustav Markkula (g.markkula@leeds.ac.uk)
Institute for Transport Studies, University of Leeds, LS2 9JT, Leeds, United Kingdom

Jami Pekkanen (j.j.o.pekkanen@leeds.ac.uk)
School of Psychology and Institute for Transport Studies, University of Leeds, LS2 9JT, Leeds, United Kingdom

Naoki Yokota (ujaeu-gfuir4@a5.keio.jp)
Faculty of Science and Technology, Keio University, Yokohama 223-8522, Japan

Naoto Matsunaga (pinen07mn@a2.keio.jp)
Faculty of Science and Technology, Keio University, Yokohama 223-8522, Japan

Natasha Merat (n.merat@its.leeds.ac.uk)
Institute for Transport Studies, University of Leeds, LS2 9JT, Leeds, United Kingdom

Tatsuru Daimon (daimon@keio.jp)
Faculty of Science and Technology, Keio University, Yokohama 223-8522, Japan

Abstract

Drift diffusion (or evidence accumulation) models have found widespread use in the modelling of simple decision tasks. Extensions of these models, in which the model's instantaneous drift rate is not fixed but instead allowed to vary over time as a function of a stream of perceptual inputs, have allowed these models to account for more complex sensorimotor decision tasks. However, many real-world tasks seemingly rely on a myriad of even more complex underlying processes. One interesting example is the task of deciding whether to cross a road with an approaching vehicle. This action decision seemingly depends on sensory information both about own affordances (whether one can make it across before the vehicle) and action intention of others (whether the vehicle is yielding to oneself). Here, we compared three extensions of a standard drift diffusion model, with regards to their ability to capture timing of pedestrian crossing decisions in a virtual reality environment. We find that a single variable-drift diffusion model (S-VDDM) in which the varying drift rate is determined by visual quantities describing vehicle approach and deceleration, saturated at an upper and lower bound, can explain multimodal distributions of crossing times well across a broad range vehicle approach scenarios. More complex models, which attempt to partition the final crossing decision into constituent perceptual decisions, improve the fit to the human data but further work is needed before firm conclusions can be drawn from this finding.

Keywords: complex decision making; road crossing; variable-drift diffusion models

Introduction

Sensorimotor decision making, how people decide what motor actions to take and when, has been a key object of research over the past hundred years in the psychological sciences. One area of particular progress has been in the development of mathematical models which predict action choices and reaction times. In particular, drift diffusion models (DDMs) and various related models, which describe the decision making process as a noisy accumulation of sensory information to a bound, have been found to very successfully capture behavioral data across a plethora of experimental tasks (Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006; Ratcliff, Smith, Brown, & McKoon, 2016) and have shown success in bridging the gap between neurophysiological and behavioral data (Purcell et al., 2010).

DDMs and related models have most commonly been applied to two alternative force choice (2-AFC) tasks, in which people make a decision between two alternative choices based on perceptual information. Quintessential among these is the kinematogram task in which people decide the direction of a random flow of dots (Ratcliff et al., 2016). DDMs have also been successfully applied to more complex sensorimotor tasks, such as determining the action intentions of other people (Koul, Soriano, Tversky, Becchio, & Cavallo, 2019). However, standard DDMs and related models of the evidence accumulation type typically assume that the drift rate (i.e., the rate at which evidence accumulates to a bound), is set to a fixed value. Yet many sensorimotor decisions take place in the context of a continuous stream of varying sensory information. Models with variable drift rate, which we will refer to here as variable-drift diffusion models (VDDMs), have been successful in the vehicle driving context, accounting well for driver brake responses to the time varying visual looming of an approaching vehicle (Xue, Markkula, Yan, & Merat, 2018) as well as for steering responses during lane-keeping (Markkula, Boer, Romano, & Merat, 2018).

However, further generalization to more complex real world decisions brings additional challenges. Firstly, more

complex decisions may depend on multiple types of sensory cues, raising the question of how different cues should contribute to the drift rate. In this paper, we will consider a pedestrian's decision of when to cross at a zebra crossing with an approaching vehicle, a decision relying on at least two types of cues (Rasouli, Kotseruba, & Tsotsos, 2017): (1) Cues regarding own affordances, for example in terms of the time to arrival (TTA) of the approaching vehicle, in relation to the width of road to be crossed. (2) Cues regarding the action intention of the vehicle driver, in the form of kinematic cues (e.g., vehicle deceleration) and/or communicative cues (e.g., flashing headlights).

Secondly, when the sensory inputs to the model vary over a large magnitude, this may result in undesirable model behavior. For example, when a vehicle decelerates to a stop, its perceptually estimated TTA will go to infinity. If this is used as a model input then the accumulator will be guaranteed to reach its threshold (and initiate a crossing) immediately when the vehicle stops, when in fact people show a probabilistic delay in crossing times.

Finally, it remains unclear how complex decisions, like the zebra crossing decisions, are structured in practice. Is the overt behavior the result of only a single action decision ("I am crossing now"), or is that action decision underpinned by separate, purely perceptual decisions about the affordances and action intentions mentioned above (e.g., "I can make it across before the car"; "The car is stopping for me")? There are many examples in the broader literature of psychological, cognitive, and robotics models where multiple parallel units of activation dynamics akin to evidence accumulators have been interconnected to produce more complex emergent behavior (e.g., Cooper & Shallice, 2000; Sandamirskaya, Richter, & Schöner, 2011), but DDM type decision models have seemingly not been previously generalized in this direction.

In the current study we wished to test three novel VDDMs, which aim to address the above three challenges. Firstly we wished to test a model recently proposed by Markkula, Romano, et al., (2018), which we refer to as the connected variable-drift diffusion model (C-VDDM). The C-VDDM models action decisions and perceptual decisions as separate but interconnected accumulator units as discussed above (see Figure 1, top), where the drift rate of each perceptual unit is a function of a time varying sensory input. In turn, the drift rate of the action unit is a function of the current activation levels of each of the two perceptual units. The activation of each perceptual unit is bounded to ± 1 which ensures that large perceptual inputs do not immediately lead to the action unit reaching threshold. Markkula, Romano, et al., (2018) showed that this model could qualitatively account for bimodal distributions of crossing decision times, as reported for human pedestrians, but did not formally test or fit the model with human data.

We also wished to test a simplification of the C-VDDM model, in which a single perceptual unit has a drift rate which varies as a function of a linear combination of multiple sensory cues (see Figure 1, middle), in turn

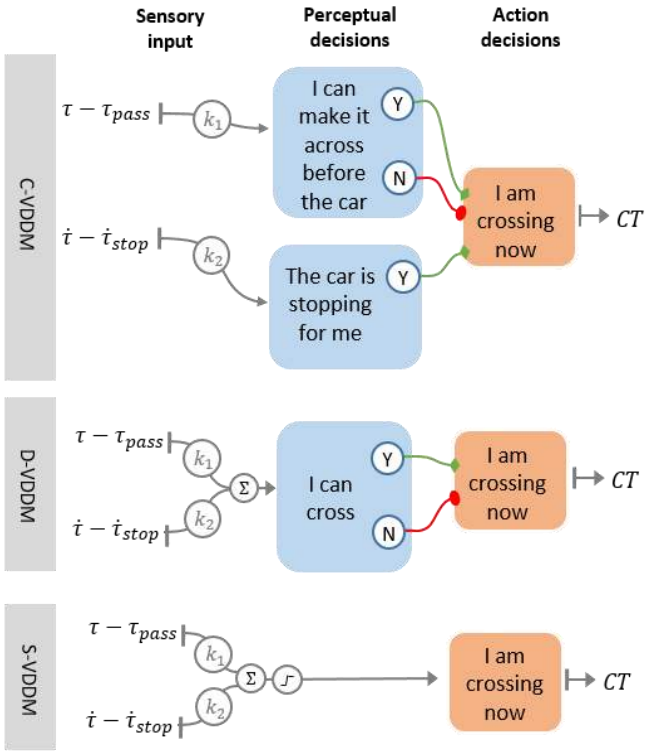


Figure 1: The three variable-drift diffusion models (VDDMs).

modulating the activation level of an action unit. We refer to this as the dual variable-drift diffusion model (D-VDDM). Like the C-VDDM, this model ensures that large sensory inputs do not result in the reaction time distribution collapsing to a spike. However, unlike the C-VDDM, this model does not independently represent different underlying decision processes.

Finally, we also wanted to test a model consisting of a single accumulator unit with a drift rate that varies as a function of a linear combination of sensory cues (see Figure 1, bottom). Instead of the D-VDDM's intermediate accumulator unit, to ensure that high input values did not result in rapid termination of the accumulation process, we limited the drift rate of this model by saturating the perceptual input, such that it could not have a magnitude greater than a certain value. We refer to this model as the S-VDDM.

To compare the three models we collected data on pedestrian crossing times, using a virtual reality (VR) headset. VR allowed us to carefully control the experimental stimulus (i.e., vehicle approach trajectories) and avoid confounding variables that may be present when observing crossing behavior in the real world (e.g., effects of other pedestrians or additional vehicles on crossing behavior). We used a large range of vehicle approach trajectories, which were specifically chosen with the aim of creating different types of situations with respect to pedestrian affordances and vehicle action intentions.

Virtual Reality Road Crossing Task

Participants

Twenty participants (age 24-60, average 27.9 years; 11 male) took part in the study and were recruited from a University participant pool. All participants provided informed consent, and the study was approved by the University Research Ethics committee.

Materials and Design

Participants wore an HTC Vive Virtual Reality headset while standing. All stimuli were created in Unity 2018. The stimuli consisted of a straight two lane road (width: 5.85 m) with a zebra crossing and pavements on either side. The trial started with the participant standing at the edge of the zebra crossing, looking directly across it. To start the trial the participant turned their head to the right, which (unbeknownst to the participants) instantiated the approaching car at its initial position and speed for the scenario in question. For increased experimental control and simplicity, the participants did not physically walk across the VR pedestrian crossing, but instead pressed a trigger button on an HTC Vive controller when they decided it was safe to cross, and the participant's view point in the virtual world then translated across the crossing at 1.31 $\mathrm{m\,s^{-1}}$. Once the participant had crossed the road in VR, the trial ended. The time at which the participant initiated the crossing, measured from the point at which the vehicle began moving, was the primary outcome measure.

Scenarios

To preserve as much as possible a natural road-crossing behavior, the number of trials per participant was limited to 16. Each of these trials used a different vehicle approach scenario, presented in a pseudo-randomized order to the participants. The scenarios were defined so as to elicit a broad range of different crossing situations, and were of three general types, with parameters as listed in Table 1:

"*Constant velocity*" (6 scenarios): The vehicle appeared at distance D_{init} from the pedestrian, and maintained a constant velocity v_{init}, i.e., it had an initial time to arrival TTA_{init}.

"*Decelerate to a stop*" (8 scenarios): The vehicle appeared at distance D_{init} from the pedestrian, with initial speed v_{init}, and immediately decelerated at a constant rate so as to reach zero speed at distance D_{stop}.

"*Decelerate without stopping*" (2 scenarios): The vehicle appeared at distance D_{init} at speed v_{init} and immediately decelerated at a constant rate until distance D_{stop}, where it continued to travel at a final speed of 5 km/h.

Table 1: Scenario parameters

Scenario type	v_{init} (km/h)	D_{init} (m)	TTA_{init} (s)	D_{stop} (m)
Constant velocity	25	15.90	2.29	N/A
	50	31.81	2.29	N/A
	25	31.81	4.58	N/A
	50	63.61	4.58	N/A
	25	47.71	6.87	N/A
	50	95.42	6.87	N/A
Decelerate to a stop	25	15.90	2.29	4
	50	31.81	2.29	4
	50	31.81	2.29	8
	25	31.81	4.58	4
	50	63.61	4.58	4
	50	63.61	4.58	8
	25	47.71	6.87	4
	50	95.42	6.87	4
Decelerate w/o stopping	50	27.78	2	8
	50	41.67	3	8

Variable-Drift Diffusion Models

We developed three models to capture the road crossing times (CT) of pedestrians in the VR study, as illustrated in Figure 1. All models received the same perceptual inputs. As in Markkula, Romano et al., (2018) the first input was based on the instantaneous apparent time to arrival (TTA) of the vehicle, disregarding any deceleration. This apparent TTA is visually available, as the relative rate of optical expansion τ (Lee, 1976). The model input was given by $\tau - \tau_{pass}$, where $\tau_{pass} = 2.46$ (the time it took to cross the VR road). Thus the model input was positive when it was possible to make it across the road before the vehicle (based on apparent TTA), and negative when it was not. The second model input was based on the derivative of the vehicle's apparent TTA, $\dot{\tau}$. The input was defined as $\dot{\tau} - \dot{\tau}_{pass}$, with $\dot{\tau}_{pass} = -0.5$, corresponding to the vehicle stopping to just exactly touch the participant (Lee, 1976). Thus, the input was positive when the vehicle was decelerating so as to stop before the participant, and negative when not.

For the C-VDDM model these inputs were fed into two separate "perceptual decision" units. For the D-VDDM and S-VDDM model these were linearly combined and fed into a single accumulator unit. For the S-VDDM this combined weighted input was also limited such that it could not exceed a certain magnitude.

Model Specification

The models were all specified on the same general form, following Markkula, Romano et al., (2018), of which a brief summary is provided here. At any point in time t, the activation level of each of the model's accumulator units is described by the vector, $\boldsymbol{A}_t = \left[A_{1,t}, A_{2,t}, \ldots, A_{U,t}\right]^T$, where U is the number of accumulator units, and each unit's activation is limited to $-1 \leq A_{i,t} \leq 1$, with 1 and -1 signifying "yes" and "no" decision states, respectively. At each simulation time step, the activation levels are updated according to,

$$\frac{d}{dt}\boldsymbol{A}_t = -\frac{1}{T}\boldsymbol{A}_t + f_c(\boldsymbol{W}_I D(\boldsymbol{K})\boldsymbol{I}_t, \eta) + \boldsymbol{W}_Y D(\boldsymbol{Y}) f_Y(\boldsymbol{A}_t) + \boldsymbol{W}_N D(\boldsymbol{N}) f_N(\boldsymbol{A}_t)$$

$$\boldsymbol{A}_{t+dt} \sim MultiNorm\left(\boldsymbol{A}_t + d\boldsymbol{A}_t, \boldsymbol{\Sigma}\sqrt{dt}\right),$$

where $\boldsymbol{I}_t = \left[\tau_t - \tau_{pass}, \dot{\tau}_t - \dot{\tau}_{stop}\right]^T$, is a vector of perceptual inputs. $\boldsymbol{K} = [k_1, k_2]^T$ is a vector of relative weights for these two perceptual inputs, $\boldsymbol{Y}$ and $\boldsymbol{N}$ are vectors of connection weights for the "yes" and "no" accumulator output connections respectively and $D(\boldsymbol{x})$ is a diagonal matrix with diagonal $\boldsymbol{x}$. $\boldsymbol{W}_I$, $\boldsymbol{W}_N$ and $\boldsymbol{W}_Y$ are design matrices which specify accumulator inputs and connections, with elements $\boldsymbol{W}_{[j,k]} \in \{0, 1\}$. The function $f_c(\boldsymbol{W}_I D(\boldsymbol{K})\boldsymbol{I}_t, \eta)$ limits the perceptual inputs to the accumulators between $\pm\eta$. In the C-VDDM and D-VDDM η was fixed at infinity (and so had no effect), while in the S-VDDM it was a free parameter. This allowed the S-VDDM's activation to gradually rise to 1, even when the inputs were at large values. The function $f_Y(x)$ limits the input between 0 and 1, thus returning $f_Y(A_{i,t}) = 1$ for an accumulator activation $A_{i,t} = 1$ (a "yes" state), while $f_N(x) = f_Y(-x)$, such that $f_N(A_{i,t}) = 1$ for $A_{i,t} = -1$ (a "no" state). $\boldsymbol{\Sigma}$ is a covariance matrix with all off diagonal elements set to 0, and all diagonal elements sharing the same value, σ^2, representing noise in the decision process.

When the activation of the action decision accumulator reaches a value of 1, a decision to cross the road is made, and the time at which this occurs is the crossing time, CT_m.

Model Fitting

To simplify notation, here we denote all the parameters of a given VDDM model as θ. Fitting to the VR dataset is made challenging as calculating the likelihood function, $P(CT|\theta)$, involves computing a high dimension integral.

Instead we estimated the likelihood function using a large number of data simulations, referred to as the pseudo-likelihood estimation, $\hat{P}(CT|\theta)$. For each trial scenario we generated 5000 simulated crossing times, CT_m, from the model being fitted. We then calculated a numerical probability distribution $\boldsymbol{b}$ over 80 bins equally spaced between 0 and 20 seconds, where $\boldsymbol{b}$ is a vector where each element, b_i, is the relative frequency of CT_m falling into the ith bin. $\hat{P}(CT|\theta)$ was then estimated as the value of $\boldsymbol{b}$ for the bin corresponding to CT.

Due to the finite number of model simulations, with this method it is possible that a bin is assigned zero probability (no values of CT_m fell within that bin), despite the model having support over this region. If CT falls within such a bin then $\hat{P}(CT|\theta) = 0$, which can cause issues for the model fitting. To avoid this, we ensured that all bins had a non-zero probability by adjusting $\boldsymbol{b}$ by a constant $\boldsymbol{z}$, to $\boldsymbol{b}\lambda + \boldsymbol{z}(1-\lambda)$, where $\lambda = .98$. $\boldsymbol{z}$ was set as the probability of drawing a value from any given bin when sampling from a uniform distribution with bounds 0 and 20. In practice this had almost no discernible effect on the estimate of $\hat{P}(CT|\theta)$, but ensured non-zero support over all values of CT. Finally we removed the first "decelerate without stopping" trial from the analysis. This was because many participants began crossing while the vehicle was still in front of them, which the models were not designed to capture.

We used PSO (Wahde, 2008) to fit the models using the pseudo-likelihood estimation method described above. A swarm of 50 particles was used and optimized for 50 iterations. In all cases the algorithm appeared to converge to some local optimum (pseudo log-likelihood estimates stopped increasing) before the 50th iteration.

Table 2: Log likelihood and Akaike information criterion (AIC) for each of the models. *indicates the model with highest log likelihood estimate

Model	$log\ P(CT\|\hat{\theta})$	N param	AIC
C-VDDM	-953.72	7	1921.4
D-VDDM*	-871.90	6	1755.8
S-VDDM	-882.04	5	1774.0

Table 3: Estimated parameter values for each model. Fixed parameters are shown in italics.

Param	C-VDDM	D-VDDM	S-VDDM
T	0.67	0.26	0.34
$\boldsymbol{K}$	[4.35, 0.46]	[0.66, 0.42]	[0.47, 0.19]
$\boldsymbol{Y}$	[*0*, 0.44, 1.83]	[*0*, 3.25]	N/A
$\boldsymbol{N}$	[*0*, 0.76, *0*]	[*0*, 10.0]	N/A
σ	0.87	1.03	1.05
η	N/A	N/A	2.5

Results

Table 2 shows the pseudo log-likelihood estimate and AIC of the VR crossing time data for each of the three models. We can see that the D-VDDM captured the data the best (highest log likelihood) and had the lowest AIC value. The S-VDDM performed slightly worse, while the C-VDDM performed poorer than both. The parameters returned by the PSO algorithm are shown in Table 3.

To explore the model fits in more detail we simulated 5000 crossing times (CT_m) for each vehicle approach scenario and each fitted models. The left panel of Figure 2 shows the real CT (top panel), and simulated CT_m (bottom panel) for one of the "constant velocity" scenarios. We also plot the model activations for the S-VDDM (black traces, bottom panel). In this trial the vehicle starts far enough away that the participant has time to successfully cross the road, if this decision is made relatively quickly. However, the vehicle soon comes too close for a successful crossing to take place. Some participants crossed early in the vehicle's trajectory, while some waited for the vehicle to pass. All of the models were able to capture this trend. However, it appears that the S-VDDM (blue line; bottom panel) and

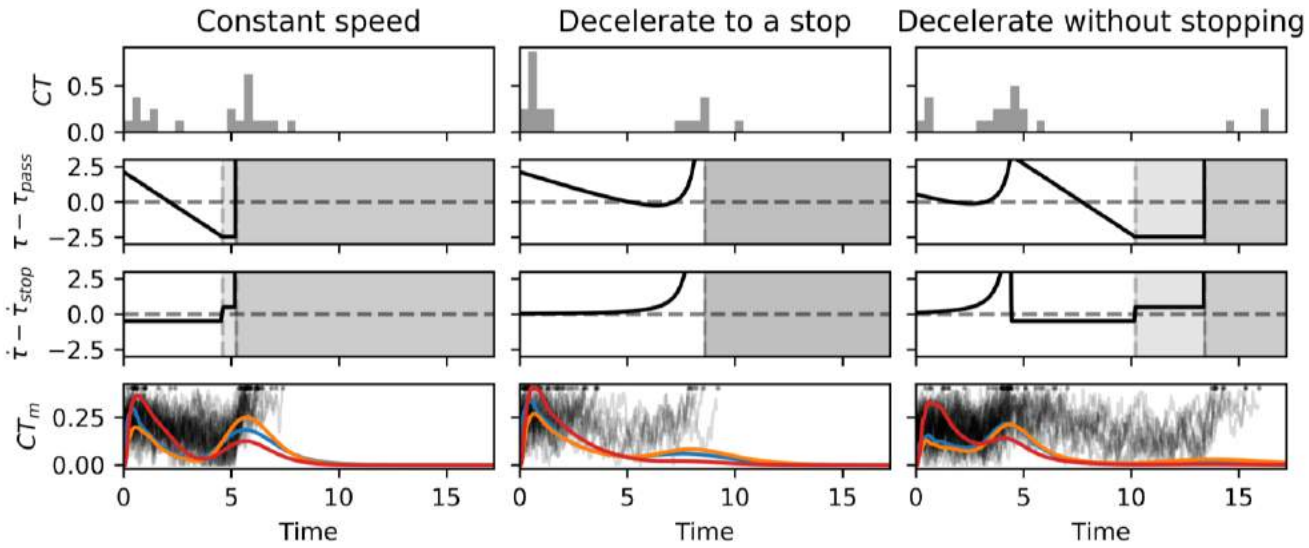


Figure 2: Human and model behavior in three example scenarios. Top panels show the observed human crossing times (CT) in the virtual reality experiment, and the second and third rows of panels show the sensory input cues to the models. Dark grey regions indicate that the vehicle has come to a stop or has passed the participant. Light grey regions indicate that the vehicle is passing the participant. The bottom panels show the simulated crossing time (CT_m) for the C-VDDM (red lines), D-VDDM (orange lines), and S-VDDM (blue lines). The black traces show example activations of the S-VDDM accumulator unit.

especially the C-VDDM (red line; bottom panel) showed a larger peak around the early crossings, while the D-VDDM (orange line; bottom panel) showed a larger peak after the vehicle had crossed, which better matched the participants' behavior.

The middle panels of Figure 2 shows the same plots for one of the "decelerate to a stop" scenarios. Again, we observed a bimodal distribution of crossing times (top panel), with some participants crossing early in the vehicle's trajectory, and others waiting until the vehicle had nearly or completely stopped. Here both the D-VDDM and S-VDDM captured this trend rather well, with a larger mode at the early crossing times and a smaller mode after the vehicle stopped. However, the C-VDDM was not able to capture the later crossing mode.

The right panels of Figure 2 show the same plots for one of the "decelerate without stopping" trials. Here, beyond the bimodal pattern already described for the "decelerate to a stop" scenario, a small number of participants also waited for the vehicle to completely pass before crossing. Thus the observed CT showed a tri-modal distribution. Both the D-VDDM and S-VDDM models reproduced these three modes well (the third mode is rather flat, but its presence can be seen from the black activation traces in bottom panel), while again the C-VDDM appeared to place too much weight over the initial mode, and predicted close to zero participants crossing after the vehicle had passed. Figure 3 shows the observed CT and model simulations, CT_m, for all scenarios.

However, we were concerned that the more complex C-VDDM's poor performance might be caused by the PSO algorithm getting stuck in a local optimum. Indeed, rerunning the fitting of the different VDDMs with new initial random seeds, and/or additional constraints on the parameter search range, we obtained slightly different parameterizations, but for the C-VDDM these never performed better than either the D-VDDM or S-VDDM.

Assuming that the C-VDDM's poor relative performance is not the result of challenges in finding the global optimum, we wondered whether one issue might be that the connected accumulator models all share a single σ parameter. Thus we refit the C-VDDM model with a separate σ parameter for each accumulator unit. This improved the model fit, achieving a log likelihood of -874.17 (AIC 1766.3), a better fit than for the S-VDDM and approaching the performance of the D-VDDM. For completeness, we also tested a version of the D-VDDM with separate σ parameters for its two accumulator units, achieving a log likelihood of -924.66 (AIC 1863.3), i.e., a worse fit than the single-σ D-VDDM. This is clearly a local optimum, since the better-performing single-σ D-VDDM is actually present in the parameter search space of the dual-σ VDDM (along the line where both σ are equal).

Discussion

Here we explored the ability of variable-drift diffusion models (VDDMs) to capture complex sensorimotor decisions based on a continuous stream of multiple sensory cues. Our initial hypothesis was that a complex model consisting of several parallel VDDM processes (the C-VDDM) would be needed to capture the multimodal decision time distributions exhibited by humans in the zebra crossing situation. Instead, we found that a relatively simple model with just a single VDDM unit (the S-VDDM) and five free parameters was able to reproduce multimodal probability distributions of human crossing times, across 15 separate scenarios with a diverse range of vehicle approach

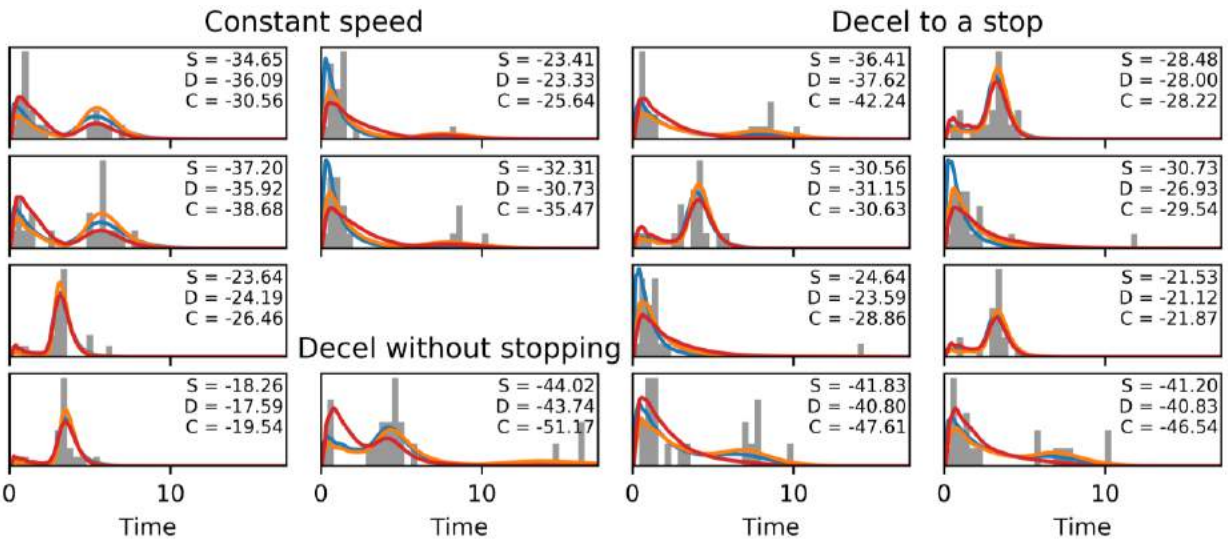


Figure 3: Observed crossing times versus predictions by the C-VDDM (red lines), D-VDDM (orange), and S-VDDM (blue) for all scenarios. Y-axis scale varies between panels. The text shows density estimate log-likelihoods for the three models.

trajectories. This is arguably the most striking finding from this work.

One important insight here, and seemingly a main reason behind the good performance of the S-VDDM, is that the τ variable (the apparent time to arrival; TTA), can in itself help explain the observed human behavior to a large extent. As seen in Figure 2, with more positive $\tau - \tau_{pass}$, participants became more likely to initiate crossing, and the non-trivial variation of τ over time during each scenario seemed to drive the number and location of peaks in the crossing time distribution. The VDDM provides a potential mechanistic explanation for how the observed crossing time distribution arises from this time-varying perceptual input.

Another critical aspect of the S-VDDM model was that while the drift rate was allowed to vary as a function of the perceptual inputs, we also limited its magnitude with a saturation threshold parameter. This ensured that large inputs, arising when the vehicle decelerated to a stop, did not result in the drift rate immediately trending to a very large value. This enabled the model to capture the distribution of crossing times that are observed after a vehicle comes to a stop or passes.

With respect to the more complex model variants, it is difficult to draw firm conclusions from the present results. If the C-VDDM model had been able to capture qualitative aspects of the human data that the S-VDDM was unable to, this could have been taken as tentative evidence for the C-VDDM's hypothesized partition of the decision process into constituent perceptual and action decisions. However, since the best version of the C-VDDM, with three separate σ parameters, simply improved the goodness of fit without changing the qualitative nature of the model behavior, it cannot be excluded that the added model complexity simply led to overfitting to the present data. To further investigate whether there is some merit to the hypotheses behind the C-VDDM, larger datasets with even more diverse scenarios would be useful, and more stringent methods than AIC for controlling for overfitting, such as hold-out validation on parts of the dataset.

Exactly the same argument applies to the D-VDDM, which was the model for which the overall best fit was obtained. The D-VDDM was adopted here as an intermediate-complexity model, in practice replacing the static input saturation step of the S-VDDM with a time-dynamic accumulator. Again, for the same reasons as mentioned above, further work is needed to shed light on whether the improved fits for this model over the S-VDDM have some theoretical relevance.

These difficulties in drawing conclusions from the fits of the more complex models are exacerbated by the apparent tendency of the PSO algorithm to get stuck in local optima. This was evidenced clearly when the PSO found a provably suboptimal parameterization for the two-σ D-VDDM, but may also be part of the reason for the somewhat surprising finding that the relatively complex single-σ C-VDDM yielded the poorest goodness of fit across all tested models. Existing methods for efficient DDM fitting are based on the conventional assumption of constant drift rate (e.g., Vandekerckhove, Tuerlinckx, & Lee, 2011); good methods for fitting also VDDMs would be a valuable future pursuit.

In summary, we demonstrate that already simple VDDMs are able to capture sensorimotor decision making behavior in a task that is more complex, and arguably of higher applied relevance, than the laboratory decision-making tasks typically modelled with DDMs. We suspect that VDDMs could be applied to a wide range of non-trivial real world sensorimotor decision making tasks, but methodological developments are needed to efficiently and reliably fit these models to data.

Acknowledgments

This work is part of the interACT project, funded by the European Union's Horizon 2020 research and innovation program under grant agreement 723395.

References

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review*, *113*(4), 700–765.

Cooper, R., & Shallice, T. (2000). Contention scheduling and the control of routine activities. *Cognitive Neuropsychology*, *17*(4), 297-338.

Koul, A., Soriano, M., Tversky, B., Becchio, C., & Cavallo, A. (2019). The kinematics that you do not expect: Integrating prior information and kinematics to understand intentions. *Cognition*, *182*(October 2018), 213–219.

Lee, D. N. (1976). A theory of visual control of braking based on information about time to collision. *Perception*, *5*(4), 437–459.

Markkula, G., Boer, E., Romano, R., & Merat, N. (2018). Sustained sensorimotor control as intermittent decisions about prediction errors: computational framework and application to ground vehicle steering. *Biological Cybernetics*, *112*(3), 181–207.

Markkula, G., Romano, R., Madigan, R., Fox, C. W., Giles, O. T., & Merat, N. (2018). Models of Human Decision-Making as Tools for Estimating and Optimizing Impacts of Vehicle Automation. *Transportation Research Record*, *2672*(37), 153–163.

Purcell, B. A., Heitz, R. P., Cohen, J. Y., Schall, J. D., Logan, G. D., & Palmeri, T. J. (2010). Neurally constrained modeling of perceptual decision making. *Psychological Review*, *117*(4), 1113–1143.

Rasouli, A., Kotseruba, I., & Tsotsos, J. K. (2017). Agreeing to cross: How drivers and pedestrians communicate. In *2017 IEEE Intelligent Vehicles Symposium (IV)* (pp. 264–269). IEEE.

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion Decision Model: Current Issues and History. *Trends in Cognitive Sciences*, *20*(4), 260–281.

Sandamirskaya, Y., Richter, M., & Schöner, G. (2011, August). A neural-dynamic architecture for behavioral organization of an embodied agent. In *2011 IEEE International Conference on Development and Learning (ICDL)* (Vol. 2, pp. 1-7). IEEE.

Vandekerckhove, J., Tuerlinckx, F., & Lee, M. D. (2011). Hierarchical diffusion models for two-choice response times. *Psychological Methods*, *16*(1), 44–62.

Wahde, M. (2008). *Biologically inspired optimization methods: an introduction.* WIT Press.

Xue, Q., Markkula, G., Yan, X., & Merat, N. (2018). Using perceptual cues for brake response to a lead vehicle: Comparing threshold and accumulator models of visual looming. *Accident Analysis & Prevention*, *118*(March), 114–124.

Evidence of error-driven cross-situational word learning

Chris Grimmick[1], Todd Gureckis[1], and George Kachergis[2]
chrisgrimmick@gmail.com, todd.gureckis@nyu.edu, gkacherg@stanford.edu
[1]Department of Psychology, New York University, New York, NY, USA
[2]Department of Psychology, Stanford University, Stanford, CA, USA

Abstract

One powerful way children can learn word meanings is via cross-situational learning, the ability to discern consistent word-referent mappings from a series of ambiguous scenes and utterances. Various computational accounts of word learning have been proposed, with mechanisms ranging from storing and testing a single hypothesized referent for each word, to tracking multiple graded associations and selectively strengthening some of them. Nearly all word learning models assume storage of some feasible word-referent mappings from each situation, resulting in a degree of learning proportional to the number of co-occurrences. While these accumulative models would generally predict that incorrect co-occurrences would slow learning, recent empirical work suggests these accounts are incomplete: paradoxically, giving learners incorrect mappings early in training was found to boost performance (Fitneva & Christiansen, 2015). We test this finding's generality in a new experiment with more items, consider system- and item-level explanations, and find that a model with error-driven learning best accounts for this benefit of initially-inaccurate pairings.

Keywords: cross-situational word learning; error-driven associative learning model; word learning;

Introduction

Among the many challenging aspects of learning a language is the problem of determining which words pick out which referents in our environment. When we encounter a new word, there is rarely an explicit explanation of its meaning, and the context it appears in may present any number of possible referents. While any given situation may present a high degree of ambiguity with many possible referents, if learners are able to roughly track words and referents that often co-occur, they may learn word meanings cross-situationally (Gleitman, 1990). Both infants and adults have been found capable of cross-situationally learning names for novel objects in the laboratory (Smith & Yu, 2008; Yu & Smith, 2007; Kachergis & Yu, 2013), and such learning may be one important means of acquiring the meanings of nouns (Smith, 2000).

It is generally assumed that learners accomplish cross-situational learning by tracking the co-occurrence of each uttered word with a subset of the visible referents in a scene. A variety of biases have been proposed that could enable the learner to restrict the number of word-referent mappings they must attend to and remember. For example, the learners have been shown to exhibit a mutual exclusivity bias, preferring to map each word to one referent–and vice-versa (Markman & Wachtel, 1988; Markman, Wasow, & Hansen, 2003; Ichinco, Frank, & Saxe, 2009). Despite a variety of proposed biases and constraints, considerable debate remains about the exact mechanisms underlying this ability.

Models of Cross-situational Word Learning

A variety of computational models have been proposed, ranging from models that store and test a single hypothesized referent per word (Trueswell, Medina, Hafri, & Gleitman, 2013), to Bayesian models, (Frank, Goodman, & Tenenbaum, 2009), to associative learning models (Kachergis, 2012; Fazly, Alishahi, & Stevenson, 2010). Typically, most models can match overall human learning performance in several experiments, and can be hard to distinguish on the basis of goodness of fit. However, detailed modeling of human learning trajectories (Kachergis & Yu, 2017) and performance in systematically varied conditions (e.g., repetitions and context diversity: (Kachergis, Yu, & Shiffrin, 2016); repetitions and number of distractors: (Yurovsky & Frank, 2015) have revealed interacting memory and attentional constraints that help differentiate models.

In many accounts of cross-situational learning, it is assumed that forming an association (or hypothesis) between a word and referent makes future exposures more valuable, as the familiar trace will draw more attention if confirmed. This advantage for prior knowledge (i.e., "rich-get-richer") is present both in hypothesis-testing accounts such as the propose-but-verify model (Trueswell et al., 2013), as well as in associative accounts that allocate more attention to pre-existing associations (Kachergis, 2012).

However, errors also play an important role in a variety of types of learning. For example, in motor control learning is thought to be based on a mismatch between predicted sensory outcomes of an action and the actual sensation (Seidler, Kwak, Fling, & Bernard, 2013). Similarly, models of animal and human conditioning experiments (Kamin, 1968; Rescorla & Wagner, 1972; Kruschke, 2011) adjust associations based on how surprising an outcome is when given particular cues. A classic example of a prediction error-based learning mechanism is the (Rescorla & Wagner, 1972) model in which the amount of learning on a trial is proportional to the amount of prediction error (i.e., surprise at an outcome). When there is a large difference between the actual outcome and predicted outcome, a large change in the predictive value of a stimulus results. Applied to cross-situational learning, surprise will be generated by the failure of a word and referent to appear together when they have been previously associated. This surprise, generated by the difference between the expectation of the word, given that object, and the actual outcome (failure of the word to appear), results in a higher learning rate for a new word to be associated with that object. Despite the widespread evidence of prediction error-based learning in the animal kingdom, empirical investigations of its role in

word learning have been limited (though see Ramscar, Dye, and McCauley, 2013). Most cross-situational word learning experiments do not facilitate continuing prediction errors: in most designs, each time a word is heard its intended referent is visible, and thus as learning proceeds, the surprise that is initially generated due to the discrepancy between the words a learner predicts and what they actually hear will only decrease.

Findings from two recent empirical studies investigating erroneous mappings early in learning suggest that greater prediction error may play an important role in cross-situational word learning (Fitneva & Christiansen, 2011, 2015). In Fitneva and Christiansen (2011), eye-tracking during cross-situational learning was used to investigate the performance of learners who by chance initially looked longer or shorter at the correct referent when a word was heard. Participants were trained on 24 word-referent pairs in four blocks, seeing two referents on each trial while hearing two sequential pseudowords. A post-hoc median split based on location of longest fixation when a word was first heard was used to place participants into High and Low Initial Accuracy conditions (HIA and LIA, respectively). Thus, participants in the HIA condition happened *by chance* to look at more of the intended referents upon each word's first occurrence than the LIA participants, who happened to look more at the incorrect referents. The accuracy of each trial in the first block (i.e., initial accuracy) was determined by the fixation time on each referent after each pseudoword was displayed. A subset of 12 of the words was used for 2-alternative forced choice (2AFC) test, in which a participant heard a word and selected the better of two referents. Participants in the LIA condition outperformed the HIA group at test. Additionally, eye tracking data provided implicit evidence for increased learning among LIA participants. In instances where the correct referent was the first location of fixation, the LIA group took longer to look away than the HIA group, and when the location of first fixation was inaccurate the LIA group was quicker to move their gaze. Proportion of time spent fixated on the accurate referent increased in LIA participants, past that of HIA participants.

A follow-up study used a "familiarization" phase before a similar cross situational learning task to induce differences in initial accuracy, and tested three age groups: 4 year-olds, 10 year-olds, and adults (Fitneva & Christiansen, 2015). In the familiarization phase, 10 unambiguous word-object pairs were serially presented to participants. However, four of these pairs would be switched in the subsequent cross-situational training for participants randomly assigned to the HIA condition (60% initial accuracy), while six of the 10 pairs would be switched in the LIA condition (40% initial accuracy). This exposure was meant to seed more (LIA) or fewer (HIA) inaccurate hypotheses/associations before the subsequent cross-situational training, which presented 15 2x2 trials (i.e., two word-referent pairs per trial). Adult participants in the LIA condition again showed higher performance than those in the HIA condition, in line with the prior results. Notably, the initial accuracy of an item within a given condition seemed to have no significant effect on performance. (Fitneva & Christiansen, 2015) interpreted this lack of an item-level effect as evidence of a 'system-level' effect, meaning that "the effect emerges from the cognitive resources recruited by initially inaccurate items affecting initially accurate items as well" (p. 5). Interestingly, four year-olds showed an opposite effect of condition, with HIA participants performing better, and 10 year-olds showed only an effect of item category, performing better on initially accurate items in both conditions.

Fitneva and Christiansen (2015) suggest that the lack of item-level effects of initial inaccuracy in adults (and in 4-year-olds) may be taken as evidence of system-driven learning: rather than individual initially-inaccurate items garnering extra attention (compared to IA items), more cognitive effort is expended overall by adults in the Low IA condition, triggered by the many inaccuracies. The present study again considers system-level vs. item-level effects of IA in adults by conducting an experiment with more to-be-learned items than Fitneva and Christiansen (2015) (18 vs. 10), and with a more sensitive 19-alternative forced choice (19AFC) test. A potential concern about finding item-level effects of IA in Fitneva and Christiansen (2015) is that adults had quite high performance in the task, which tested half of the 10 studied words using a 2AFC test. In addition, the difference between the HIA and LIA conditions was one of only two words (6 out of 10 and 4 out of 10 accurate, respectively).

The superior performance on initially inaccurate items in both experiments may be accounted for with a prediction error mechanism. An additional attentional account may be able to account for the overall difference in performance between the HIA and LIA conditions. Thus, the present design offers a stronger manipulation, more data per participant, and a more sensitive test, while addressing the same underlying issue of the effects of initial accuracy on learning. We then present modeling in an associative learning framework to determine if learning behavior is better accounted for by an attentional (system-level) mechanism, or by a prediction error-based (item-level) mechanism.

Experiment

To investigate the robustness of the effect of low initial accuracy observed in Fitneva and Christiansen (2015) in a setting with more to-be-learned items and a consequently longer training period, we use a similar 2x2 procedure with a "familiarization phase". However, in our design, we used studied 18 stimulus pairs (vs. 10), and a greater degree of difference between between high and low initial accuracy (12 vs. 6 of 18 pairs switched instead of 6 vs. 4 of 10 pairs switched). This presents a stronger manipulation of initial accuracy: 66.6% vs. 33.3% in the current study, compared to 60% vs. 40% in Fitneva and Christiansen (2015). In addition, at test we presented the full array of possible referents for each word (18 studied + 1 unstudied: 19AFC vs. 2AFC), and tested all 18 words (vs. 5 of 10).

Methods

Participants Participants were 45 people recruited online who completed the experiment in their web browser through Amazon Mechanical Turk. All participants completed the entire experiment, and were paid $1.50 for their participation. Participants were randomly assigned to either the High or Low Initial Accuracy condition (23 and 22 participants, respectively).

Stimuli Stimuli consisted of images of uncommon real-world objects and mono- and bisyllabic nonce words. Each participant was given a random selection of 18 images and 18 nonce words from a collection of 72 images and words.

Procedure The experiment consisted of three phases: familiarization, study, and test. The familiarization phase was one block of 18 trials. Participants were told they would be shown examples of the type of objects and words they would be learning. Each trial showed one object-pseudoword pair for 3 s with a 1 s interstimulus interval (ISI), with each of the 18 pairs being shown once, in a randomized order.

For participants in the Low Initial Accuracy (LIA) condition, 12 of the 18 pairs were switched (inaccurate) in the subsequent study phase, yielding 33.3% initial inaccuracy. In the High Initial Accuracy (HIA) condition, 6 of the 18 pairs were switched at study, yielding 66.6% initial accuracy. For each participant, the total number of objects was constant, such that when word-object pairs were switched at study, both the word and object had been seen in the familiarization phase.

On each trial during the study phase, two word-object pairs were shown simultaneously for 3 s, with a 1 s ISI. Objects were shown side by side, with words vertically arrayed in the center, below the object images. The location of objects and words was randomized to ensure participants could not reliably determine the pairing of stimuli by their location with respect to one another. Trial order, along with which word-object pairs were shown on a given trial, was randomized with the constraint that each word-object pair was presented once before being shown again. Thus, there were three (contiguous) blocks of 9 trials, for a total of three presentations per pair.

In the test phase, each trial displayed an array of all 18 studied objects along with one novel object (the same across all test trials) and a single pseudoword from the study. For each pseudoword, participants were instructed to click on the corresponding object. In addition to testing each of the 18 studied pseudowords, a trial with a novel pseudoword was added, to determine if participants were able to fast-map this novel word to the novel object in the array. The order of test trials was randomized for each participant.

A post-test questionnaire asked participants how many words they thought they mapped correctly (0-19), their rating of the engagement and difficulty of the task on scales of 1-7, and whether they used any external memory aids (Yes/No; indicating that they would still be paid, regardless).

Results

Participant's item-level accuracy data for each studied item were subjected to a logistic mixed-effects regression with condition (High Initial Accuracy (HIA) or Low Initial Accuracy (LIA)) as a between-subjects factor and item category (Initially Accurate or Initially Inaccurate) as a within-subject factor. Mixed-effects regression is more appropriate for forced-choice data than ANOVAs, especially for experiment designs with imbalanced cells such as this one (Jaeger, 2008). The analysis was conducted using the `afex` `R` package (Singmann, Bolker, Westfall, & Aust, 2018). This analysis indicated a significant main effect of item category (F(1,43.7)=42.19, $p < .001$), and no significant main effect of condition (F(1,43.7)=0.86, $p = .36$). Learners had higher performance for items that were initially accurate (M=.59, SD=.31) than for items that were initially inaccurate (M=.35, SD=.30). There was a marginal interaction of condition and item category (F(1,43.7) = 3.50, $p = 0.07$). Shown in Figure 1, accuracy on initially inaccurate items was higher in the LIA condition (M=.42, 95% CI=$[.30, .55]$) than in the HIA condition (M=.28, CI=$[.15, .40]$), but lower than initially accurate items, which were similarly high in both conditions (IA: LIA M=.60, CI=$[.46, .73]$; HIA M_{HIA}=.59, CI=$[.47, .72]$). Overall, in both conditions participants learned on average the same proportion of the 18 items (M_{HIA}=.49; M_{LIA}=.48). Finally, for the novel word presented at test, 47% of participants chose the unstudied test object.

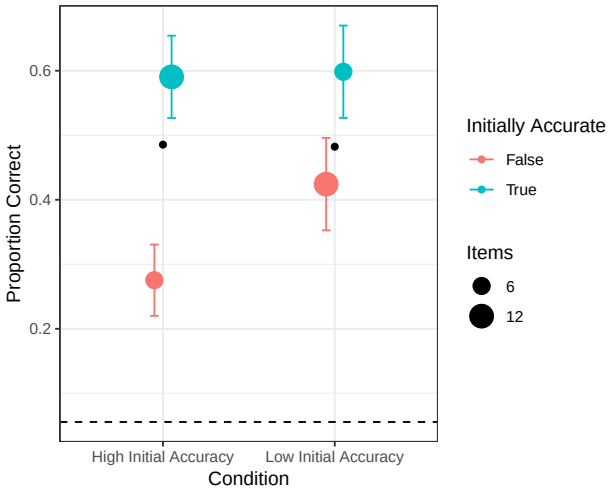


Figure 1: Participants' mean accuracy at test in each condition by item category, with dot size denoting the number of items that could be learned in that category. Black dots show mean performance per condition. Dotted line shows chance (1/18). Error bars represent ± 1 SE.

Post-test Questionnaire To investigate metacognitive awareness, individuals' performance was correlated with post-test questionnaire results. Participants in both conditions

were aware of their performance level, with significant correlations between their actual and estimated number of learned items they learned (HIA $r(20) = 0.67$, $p < .001$) and LIA ($r(18) = 0.60$, $p < .01$)). Rank-order tests were used to investigate the relationship of engagement and difficulty ratings with performance at test. Strong negative relationships between performance at test and difficulty ratings were found in both HIA ($r_s = -0.73$, $p < .001$) and LIA ($r_s = -0.77$, $p < .001$) conditions. A system-level account of benefits of initial inaccuracy might predict that difficulty would be higher in the LIA condition, but that engagement would be higher. However, there was no difference in participant's perceived level of difficulty in the two conditions ($t(42.7) = 0.56$, $p = .57$), and participants in the HIA condition trended toward being more engaged than LIA participants ($t(38.6) = 1.87$, $p = .07$)–the opposite of what might be predicted by a system-level account.

Discussion

Our results differ somewhat from those of Fitneva and Christiansen (2015) in that we do not find an overall advantage for the Low Initial Accuracy condition. Moreover, we find in both conditions that initially accurate items are learned more often than initially inaccurate items, in agreement with conventional assumptions. However, we do find that initially inaccurate items are learned at a greater rate when they are a greater proportion of the study items (i.e., in the LIA condition). Given that this experiment has a stronger manipulation of initial accuracy (66% vs. 33% instead of 60% vs. 40%), includes more studied items (18 instead of 10), and tests all 18 of them (instead of half) with a more sensitive test (19AFC instead of 2AFC), we contend that it makes a stronger case for the influence of varying initial inaccuracy on cross-situational learning. In the following, we explore what mechanisms account for these effects.

Models

To determine whether the effect of initial accuracy implies novel system-level or item-level learning mechanisms, we first test whether the biased associative model (Kachergis, Yu, & Shiffrin, 2012), with competing attentional biases for existing associations and for attending to stimuli with uncertain associates, is able to account for the effect of varying initial accuracy. This model, explained in detail below, has successfully captured human behavior in a variety of cross-situational learning experiments (Kachergis & Yu, 2017; Kachergis et al., 2016; Kachergis, 2012; Kachergis & Yu, 2013). We also test two modified versions of this model, representing the two theories of why forming initial inaccurate associations may improve overall learning. In the system-level variant, the learning rate on each trial was scaled by the model's relative uncertainty about the words for each presented referent, representing the theory proposed in (Fitneva & Christiansen, 2015) that learners may be more alert in the Low Initial Accuracy condition. In the item-level variant, we add a simple prediction error-based learning mechanism borrowed from Rescorla and Wagner (1972).

Biased Associative Model

The biased associative model (Kachergis et al., 2012) assumes that learners do not attend equally to all possible word-object pairings. Thus, although all co-occurrences are registered to some extent in associative memory (a word $\times$ object association matrix), greater attention and storage is directed to pairings that have previously co-occurred. Moreover, this bias for familiar pairings competes with a bias to attend to stimuli that have no strong associates (e.g., novel stimuli). In addition to familiar associations being reinforced, attention is also pulled individually to novel stimuli because of the high uncertainty of their associations (i.e., they have diffuse associations with several stimuli). Uncertainty is tracked by the entropy of a stimulus' association strengths, and attention is allocated to a stimulus in proportion to this entropy.

Formally, given n words and n objects to be learned over a series of trials, let M be an n word $\times$ n object association matrix that is built incrementally during training. Cell $M_{w,o}$ will be the strength of association between word w and object o. Strengths are augmented by viewing the particular stimuli. Before the first trial, M is empty. On each training trial t, a subset S of m word-object pairings appears. If there are any new words and objects are seen, new rows and columns are first added. The initial values for these new rows and columns are k, a small constant (here, 0.01).

Association strengths are allowed to decay, and on each new trial a fixed amount of associative weight, χ, is distributed among the associations between words and objects, and added to the strengths. The rule used to distribute χ (i.e., attention) balances a preference for attending to unknown stimuli with a preference for strengthening already-strong associations. When a word and referent are repeated, extra attention (i.e., χ) is given to this pair—a bias for prior knowledge. Pairs of stimuli with no or weak associates also attract attention, whereas pairings between uncertain objects and known words, or vice-versa, do not attract much attention. To capture stimulus uncertainty, strength is allocated using entropy (H), a measure of uncertainty that is 0 when the outcome of a variable is certain (e.g., a word appears with one object, and has never appeared with any other object), and maximal ($log_2 n$) when all of the n possible object (or word) associations are equally likely (e.g., when a stimulus has not been observed before, or if a stimulus were to appear with every other stimulus equally). In the model, on each trial the entropy of each word (and object) is calculated from the normalized row (column) vector of associations for that word (object), $p(M_w, \cdot)$, as follows:

$$H(w) = -\sum_{i=1}^{n} p(M_{w,i}) \cdot \log(p(M_{w,i})) \quad (1)$$

The update rule for adjusting the association between a given word w and object o on a given trial is:

$$M_{w,o} = \alpha M_{w,o} + \frac{\chi \cdot e^{\lambda \cdot (H(w)+H(o))} \cdot M_{w,o}}{\sum_{w \in W} \sum_{o \in O} e^{\lambda \cdot (H(w)+H(o))} \cdot M_{w,o}} \quad (2)$$

In Equation 2, α is a parameter governing forgetting, χ is the weight being distributed, and λ is a scaling parameter governing differential weighting of uncertainty ($H(\cdot)$; roughly novelty) and prior knowledge ($M_{w,o}$; familiarity). As λ increases, the weight of uncertainty (i.e., the exponentiated entropy term, which includes both the word and object's association entropies) increases relative to familiarity. The denominator normalizes the numerator so that exactly χ associative weight is distributed among the potential associations on the trial. For stimuli not on a trial, only forgetting operates. As each word w is tested, learners choose referent o from the m alternatives in proportion to associative strength $M_{w,o}$.

Biased Associative Model with Attention

To capture the theory proposed in (Fitneva & Christiansen, 2015) that learners may be more alert in the Low Initial Accuracy condition, we scale the learning rate used on each trial by the mean entropy of the objects on a given trial, relative to the overall entropy of all associations. Thus, trials with more uncertain items–as in the Low IA condition, and particularly for initially inaccurate items–will have a higher learning rate.

Predictive Biased Associative Model

This model differs from the original Biased Associative Model in two ways. First, for the cues (objects) on the trial, let the prediction of each outcome (word w) be

$$V_w = \sum_{o \in O} M_{w,o} \quad (3)$$

V_w was added to the update equation for on-trial word-object associations as a prediction error term

$$M_{w,o} = \alpha M_{w,o} + \chi \cdot e^{\lambda \cdot (H(w)+H(o))} \cdot M_{w,o} \cdot (\beta - V_w) \quad (4)$$

where β is the maximum association value (here, 1), and as before α is a memory fidelity parameter, χ is a learning rate, and λ is relative novelty/familiarity focus. The second difference is the removal of the denominator, which makes it possible for the predictive model to distribute different amounts of associative weight per trial. Thus, the amount of adjustment for a particular association $M_{w,o}$ is scaled not only by the current strength of that association and the uncertainty (entropy) of w and of o, but also proportional to the prediction error of w from the sum of all associations involving w and objects on that trial.

Model Fitting

All models were fit hierarchically: first, differential evolution optimization (Ardia, Mullen, Peterson, & Ulrich, 2015) was used to find best-fitting parameters for each individual, and then optimization was run again with a regularization term to penalize parameter values far from the medians of the group's parameter values.[1]

Model Results

The best-fitting performance achieved by each model, along with Mean Squared Error (MSE) are shown in Figure 2, as well as in Table 1. All variants of the Biased Associative Model (BAM) match performance well in the HIA condition, for both initially accurate and inaccurate items. However, in the LIA condition, both the original BAM and BAM + Attn underestimate human performance on initially inaccurate items and overestimate learning of initially accurate items, while the Predictive BAM fits well.

Condition	High IA		Low IA			
Initially Accurate	False	True	False	True	MSE	r^2
Human	.28	.59	.42	.60	–	–
Biased Assoc.	.26	.60	.30	.66	.026	.939
Biased Assoc. + Attn	.27	.60	.34	.67	.025	.941
Predictive Biased Assoc.	.27	.59	.40	.63	.008	.983

Table 1: Human performance vs. best-fitting models.

Model Discussion

All models match human performance well in the High Initial Accuracy (HIA) condition, and predict slightly higher than observed performance for initially accurate items in the Low IA condition. Both the original Biased Associative Model (Kachergis et al., 2012) and the variant with a learning rate scaled to the uncertainty (i.e., entropy) about items on the current trial are unable to match human performance for initially inaccurate items in the Low IA condition. However, the variant of the Predictive Biased Associative Model does match human performance, suggesting that learners allocate more attention to associations involving words from initially inaccurate items as a result of prediction error.

Discussion

Similar to earlier studies of the effects of initial accuracy on cross-situational word learning (Fitneva & Christiansen, 2011, 2015), our findings show that experiencing a single initial inaccurate mapping of more word-object pairs selectively benefits the later learning of those initially mismatched pairs. However, in contrast to prior research, which found overall higher learning in the Low Initial Accuracy condition, we found the benefit was not conferred on initially accurate items. Rather, performance on initially accurate items in our experiment was similarly high in both conditions–and higher than initially inaccurate items in either condition. As mentioned earlier, the present experiment presents a stronger test of the effects of initial accuracy due to the stronger manipulation, the larger number of studied and tested words, and the

[1] This approximates Gaussian L1-regularization.

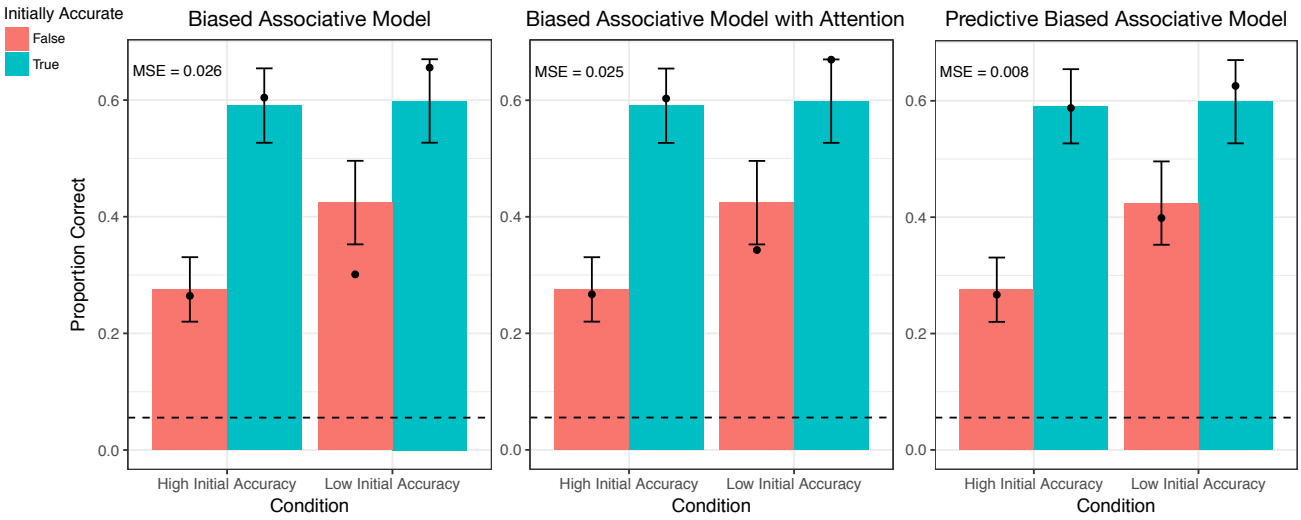


Figure 2: The Biased Associative Model (left) fails to show much benefit for initially inaccurate items in the Low IA condition, unlike people. This is largely true for the variant of the model that gives greater attention to trials with high-entropy stimuli (middle), while the associative model with prediction-error based learning (right) fits quite well (MSE=.008).

lower chance rate of the test format. However, we should note that our stronger manipulation resulted in a slightly higher proportional imbalance of item types per condition (67% vs. 33%)–although we did have more items of both types per condition. It's possible that this greater number of items influenced participants' awareness and thus their treatment of initially inaccurate items. However, we note that difficulty was similarly rated similarly in both groups, and engagement trended higher in the HIA group–opposite to what might be expected if more attention was drawn by the LIA condition.

Our item-level results show that initial accuracy predicts a greater chance of remembering the pairing of that item, which is in accordance with conventional assumptions. However, in the LIA condition initially inaccurate pairs are ultimately more likely to be learned. One explanation may be that errors draw attention selectively to initially inaccurate items. Analysis of fixation times in the eye-tracking experiment of Fitneva and Christiansen (2011) indeed suggests that attention to targets overall increases with greater error. Additionally, the pattern of our results suggests an attention effect: the HIA condition may have included too low a proportion of inaccurate items to draw attention attention away from the majority accurate items. In the LIA group, if there was an overall increase in attention, we should expect to see an increase in performance for initially-accurate (IA) items as well, especially as there were only six IA items in this condition. Our modeling results are consistent with this idea, as without a learning rate proportional to item-level prediction error the fit is notably poor for initially inaccurate items in the LIA condition.

Together with Fitneva and Christiansen (2011, 2015), our results suggest that cross-situational word learning is subject to prediction error-based learning. Our account suggests that when learners see referents they may predict which words will be heard. Subsequently, they allocate attention based on competing biases toward known associations and referents with uncertain associations (Kachergis et al., 2012), and learn at a rate proportional to their surprisal at hearing each word with the given referents. Further research is needed to determine whether this item-level prediction error-based learning mechanism accounts for human behavior—both in typical research settings which offer few inaccurate mappings, as well as in more naturalistic scenarios—to help us further understand the domain-generality of error-based learning.

References

Ardia, D., Mullen, K. M., Peterson, B. G., & Ulrich, J. (2015). *'DEoptim': Differential Evolution in 'R'*. Retrieved from `http://CRAN.R-project.org/package=DEoptim`

Fazly, A., Alishahi, A., & Stevenson, S. (2010, May). A Probabilistic Computational Model of Cross-Situational Word Learning. *Cognitive Science*, *34*(6), 1017–1063.

Fitneva, S. A., & Christiansen, M. H. (2011). Looking in the Wrong Direction Correlates With More Accurate Word Learning. *Cognitive Science*, *35*(2), 367–380.

Fitneva, S. A., & Christiansen, M. H. (2015). Developmental changes in cross-situational word learning: The inverse effect of initial accuracy. *Cognitive Science*, 367–380.

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2009, May). Using Speakers' Referential Intentions to Model Early Cross-Situational Word Learning. *Psych. Science*, *20*(5), 578–585.

Gleitman, L. (1990). The structural sources of word meaning. *Language Acquisition*, *1*, 3–55.

Ichinco, D., Frank, M., & Saxe, R. (2009). Cross-situational word learning respects mutual exclusivity. In *Proceedings of the 31st Annual Meeting of the Cognitive Science Society* (pp. 1–6).

Jaeger, T. F. (2008). Categorical data analysis: Away from anovas (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*, 434–446.

Kachergis, G. (2012). Learning nouns with domain-general associative learning mechanisms. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proc. of CogSci 34* (p. 533-538).

Kachergis, G., & Yu, C. (2013). More naturalistic cross-situational word learning. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proc. of CogSci 35* (pp. 527–532).

Kachergis, G., & Yu, C. (2017). Observing and modeling developing knowledge and uncertainty during cross-situational word learning. *IEEE Transactions on Cognitive and Developmental Systems*. doi: 10.1109/TCDS.2017.2735540

Kachergis, G., Yu, C., & Shiffrin, R. M. (2012). An associative model of adaptive inference for learning word–referent mappings. *Psychonomic Bulletin and Review*, *19*(2), 317–324.

Kachergis, G., Yu, C., & Shiffrin, R. M. (2016). A bootstrapping model of frequency and contextual diversity effects in word learning. *Cognitive Science*. doi: 10.1111/cogs.12353

Kamin, L. J. (1968). "Attention-like" processes in classical conditioning. In M. R. Jones (Ed.), *Miami symposium on the prediction of behavior, 1967: Aversive stimulation* (pp. 9–31). Coral Gables, FL: University of Miami Press.

Kruschke, J. K. (2011). Models of attentional learning. In E. M. Pothos & A. J. Wills (Eds.), *Formal approaches in categorization.* Cambridge University Press.

Markman, E. M., & Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology*, *20*, 121–157.

Markman, E. M., Wasow, J. L., & Hansen, M. B. (2003, November). Use of the mutual exclusivity assumption by young word learners. *Cognitive Psychology*, *47*(3), 241–275.

Ramscar, M., Dye, M., & McCauley, S. (2013). Error and expectation in language learning: The curious absence of 'mouses' in adult speech. *Language*, *89*(4), 760-793.

Rescorla, R. A., & Wagner, A. R. (1972). *A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement.* New York, NY: Appleton Century Crofts.

Seidler, R. D., Kwak, Y., Fling, B. W., & Bernard, J. A. (2013). Neurocognitive mechanisms of error-based motor learning. *Advances in Experimental Medicine and Biology*(782).

Singmann, H., Bolker, B., Westfall, J., & Aust, F. (2018). afex: Analysis of factorial experiments [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=afex` (R package version 0.22-1)

Smith, L. B. (2000). How to learn words: An associative crane. In R. Golinkoff & K. Hirsh-Pasek (Eds.), *Breaking the word learning barrier* (pp. 51–80). Oxford: Oxford University Press.

Smith, L. B., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*, 1558–1568.

Trueswell, J. C., Medina, T. N., Hafri, A., & Gleitman, L. R. (2013). Propose but verify: Fast mapping meets cross-situational word learning. *Cognitive Psychology*, *66*(1), 126–156.

Yu, C., & Smith, L. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psych. Science*, *18*, 414–420.

Yurovsky, D., & Frank, M. C. (2015). An integrative account of constraints on cross-situational word learning. *Cognition*, *145*, 53–62.

A comprehensive examination of preschoolers' probabilistic reasoning abilities

Samantha Gualtieri (sgualtieri@uwaterloo.ca) & Stephanie Denison (stephanie.denison@uwaterloo.ca)
Department of Psychology, University of Waterloo
Waterloo, Ontario, Canada

Abstract

Historically, research on preschool-aged children's probabilistic reasoning abilities has yielded mixed results. Although some findings have suggested that young children can successfully evaluate probabilities, others have suggested that they may use strategies that only approximate true probabilistic inference and therefore sometimes make errors (e.g., Girotto et al., 2016; Piaget & Inhelder, 1975). To explore the factors that affect young children's probabilistic reasoning, we developed a battery of problems that contained features that affect the ease with which a problem is evaluated, and the types of alternative strategies that can be applied to solve them. The current experiments (total $N = 124$) assessed 3- and 4-year-old children's probabilistic reasoning using an experimental paradigm tailored to this age group. Results from both experiments suggest that young children are able to engage in true probabilistic inference, as they performed well-above chance on each problem. Nuances in children's performance are discussed, along with possibilities for future research.

Keywords: probabilistic reasoning; cognitive development; decision making

Introduction

Our ability to make inferences under uncertainty is critical to learning and decision-making, mimicking the contexts in which every day reasoning tends to occur. That is, we are typically in situations where we only have access to probabilistic information. Although sensitivity to base-rates is evident very early in development, questions remain regarding young children's strategies for using base-rates in their inferences, which can be diagnosed by asking children to make more complex proportional comparisons.

Non-human primates, infants, toddlers, and preschoolers correctly infer that an object of the majority type is most likely to be randomly sampled from simple probabilistic distributions (Denison, Konopczynski, Garcia, & Xu, 2006; Denison & Xu, 2010; Denison & Xu, 2014; Eckert, Call, & Rakoczy, 2017; Goldberg, 1966; Kushnir, Xu, & Wellman, 2010; Ma & Xu, 2011; Rakoczy et al., 2014; Téglás, Girotto, Gonzalez, & Bonatti, 2007; Téglás et al., 2011; Xu & Garcia, 2008; Yost, Siegel, & Andrews, 1962). For example, if a distribution has more red than white balls (e.g., 80 red and 20 white), they infer that a small sample taken from that distribution should also have more red than white balls. Although young children and non-human primates perform above chance on many probability problems, poor performance has been observed in some experiments, particularly in the 3- and 4-year-old age group (Girotto, Fontanari, Gonzalez, Vallortigara, & Blaye, 2016; Girotto & Gonzalez, 2008; Piaget & Inhelder, 1975). The current experiments explore whether some of this variability in performance is due to differences in problem difficulty by manipulating features of the problem that diagnose strategy use. We used a paradigm designed specifically for 3- and 4-year-olds to ensure that their abilities were not masked by difficulties with, or lack of engagement in, the task itself.

When adapting a task for a particular population, it is important to ensure that the paradigm is suitable to their abilities and still captures the essential aspects of the skill of interest. Issues regarding task-appropriateness have arisen throughout the course of research on children's probabilistic reasoning. Though Piaget's seminal work provides one of the first analyses of children's probabilistic reasoning abilities (Piaget & Inhelder, 1975), younger children's performance may have suffered due to the very high verbal demands of the task. Participants were asked which color item the experimenter was most likely to obtain on a random draw (Yost et al., 1962). Children's responses were then coded as correct based on their explicit reference of probabilistic concepts. From this work, it was concluded that children younger than 12 years of age struggled with probabilistic concepts. Conversely, presenting preschoolers with a choice paradigm suitable for infants and primates (e.g., Denison & Xu, 2014; Rakoczy et al., 2014) also appears to hinder their performance. When designing tasks for pre-verbal infants, experimenters use prompts that provide general encouragement (i.e., infants are told, "You can do it! Get the one you like!"). However, this prompt could make the task unclear to a preschooler with more advanced cognitive and linguistic abilities because these instructions are misleading. Children might recognize that when they choose something in a probabilistic context, they cannot guarantee that they will "get the one they want", they can only make a best guess. When this prompt was used with preschoolers, 3- and 4-year-olds' performance suffered (Girotto et al., 2016, Expt. 2). We used an age-appropriate method in the current experiments by asking children to provide a forced-choice response to a direct but simple probability question (see Procedure).

Moreover, there is considerable variability in the types of problems that have been presented to children in this age group. Falk, Yudilevich-Assouline, and Elstein (2012) outline this important point in their comprehensive assessment of school-aged children's probabilistic reasoning. Children were asked to choose between two small populations of items, each including a proportion of target and non-target items, to sample from in order to maximize their chances of obtaining a target item on a blind draw. The authors note that much previous research has overlooked the importance of manipulating numerical features of the presented problems when examining children's overall performance. Without manipulating these features across a variety of problems, it is difficult to know whether heuristic reasoning or true probabilistic inference led to correct responses in previous experiments. To combat this problem,

Falk et al. developed a battery of diagnostic problems that could not be solved using simple heuristic strategies (see Denison & Xu, 2014, for a similar approach with infants). For instance, in probability problems, children can use a heuristic in which they only compare the number of target items across populations, and thus ignore proportions. One can diagnose whether children are using this strategy by presenting them with problems that contain an equal number of target objects across two populations (e.g., 12 targets and 4 non-targets vs. 12 targets and 48 non-targets), and asking them to choose a population to draw from for the best chance of obtaining a target. This allows researchers to diagnose use of a strategy that solely focuses on choosing the population with more target objects, because children would be unable to solve such a problem if they tried to apply this strategy. One can also include problems in which there are more non-target objects in the more probable population to diagnose an avoidance strategy. Thus, children cannot succeed by simply choosing the population with more target items, or by choosing the population with fewer non-targets.

Notably, Falk et al. (2012) included problems in their experiment that assessed use of a good versus bad label shortcut. That is, instead of discerning the proportion of objects in each population, children could use a simpler shortcut that focuses on the majority type of objects in each population but does not require comparing proportions *across* populations. Many studies have presented children with a choice between, for example, a 75% target population and a 25% target population. A child could solve this problem by labelling the 75% population as "good", because the target objects are in the majority, and the 25% population as "bad", because the non-target objects are in the majority. This would lead them to approach the "good" population without carefully discerning and comparing the proportion of objects in each population. To assess use of this heuristic, Falk et al. included problems that were on the same side of ½. If a child who uses the good versus bad label shortcut was presented with two populations on the same side of ½, such as 75% and 95%, they would be unable to solve such a problem because both populations would receive the same label.

We attempted to tease apart preschoolers' use of true proportional reasoning from use of heuristics that approximate probabilistic inference. Because we were presenting these problems to children younger than those tested by Falk et al. (2012), we included problems that diagnosed use of simpler heuristics that may be used by preschoolers, as well as some of the more advanced ones described above. We included problems with more target objects in the less probable population and problems with an equal number of target objects in both populations. These features allowed us to examine if young children solely focus on target objects. Additionally, we included problems where the more probable population contained more non-target objects to examine if children attempted to avoid this option. We also included problems on the same side of ½ to gauge children's use of a shortcut that involves focusing on the majority of objects in individual populations.

Finally, closer, rather than more disparate, relative likelihoods (sometimes referred to as the "ratio of ratios") can make problems more difficult to evaluate. That is, when the likelihoods of each population are closer together, the problem can be more difficult to solve than when they are further apart because the populations themselves are more difficult to visually discriminate. For example, if Problem 1 contained a comparison between 80% and 75% targets, and Problem 2 contained a comparison between 90% and 60% targets, Problem 1 would be more difficult to solve because the relative likelihoods are closer and are more difficult to discriminate. Previous investigations of preschooler's probabilistic reasoning have not examined the impact of relative likelihood on their responses (but see Hoemann & Ross, 1971, for a similar manipulation using a spinner task), so we include a manipulation of this feature in the current experiments. Thus, for each problem type we included two versions, denoted 1 and 2, to mark, respectively, closer and further relative likelihoods.

Experiment 1

In Experiment 1, we presented 3- and 4-year-old children with a battery of probabilistic reasoning problems using a two-alternative forced choice procedure in a gumball machine paradigm. Children were tasked with selecting the population that was more likely to yield a blue object. We developed a set of problems to assess use of different strategies (see Figure 1). Problems A1 and A2 presented children with populations on the same side of ½. These problems also included more targets and more overall objects in the less probable population, so a child would not succeed on these problems if they were drawn to these features. Because this problem is challenging, the more probable population only contained target items, and thus the outcome was deterministic. Problems B1 and B2 were simple probabilistic comparisons that could be solved with multiple shortcuts. Although these simpler problems do not diagnose use of these shortcuts, we included them in our problem set to gauge the effectiveness of our paradigm with this age group, as 3- and 4-year-old children have solved these very simple problems in previous experiments. Problems C1 and C2 prevented children from selecting the population with more target objects, because the number of target objects was the same in both populations. Problem D presented children with two uniform populations in which one population only contained targets, and the other contained only non-targets to assess the effectiveness of the paradigm. This problem was always presented second to last, allowing us to gauge whether most children were following the task through such a large number of problems. Problem E was the inverse of Problem A1 and was included to diagnose whether children might use an avoidance strategy to solve problems (i.e., choosing a population that has fewer non-targets).

We included two versions of Problems A, B, and C to determine if probabilities that had higher relative likelihoods (i.e., problems that were further apart in probability, which were labeled with a 2), were easier for children to evaluate.

Methods

Participants Data from 50 3- and 4-year-olds were included in analyses (*mean age* = 4;2 [years;months]; *range* = 3;3 to 4;11). The sample size for the experiment was determined based on a power analysis for a larger study. An additional five children were tested and were excluded from analyses due to parental report of atypical development (n = 1), parental report of very low English exposure (i.e., hearing English less than 50% of the time; n = 2), and not finishing the task (n = 2). Participants were recruited from a database of families and received a small gift for their time.

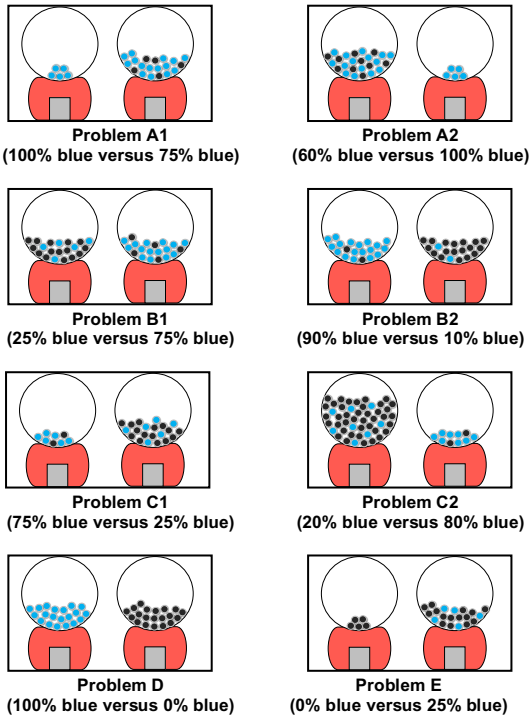


Figure 1: Probability problems presented in Experiment 1.

Materials and Procedure Participants were tested individually in a quiet room. The child and experimenter were seated together at a table. If the child's parent was present in the room, they were seated across the table from the child, unable to see the iPad's screen, and were asked to refrain from commenting or influencing their child's responses during the experiment. After the probability task, children completed measures assessing individual differences in their cognitive abilities (i.e., executive function and receptive vocabulary skills) as part of a larger study. Because the probability task was completed first, the additional measures had no impact on performance.

Probability problems were presented on an iPad, using a gumball machine paradigm. Prior to the test trials, the experimenter explained how gumball machines worked by showing participants two machines filled with a mixture of gumballs of various colors. To discourage children from focusing on the objects that were closer to the opening, the gumballs were then mixed and appeared in different positions in the machine, illustrating that any gumball in the population, regardless of its initial position, could be sampled. After mixing three times, each machine yielded one gumball. The experimenter told participants they had to choose between two machines and reiterated that each machine would only yield one gumball. Participants were told that they would receive a sticker if they chose a machine that yielded a blue gumball. Children then completed eight probability trials and were asked to choose the gumball machine that gave them the best chance of obtaining a blue gumball. On each trial, the machines always produced the more probable color gumball.

In populations that contained both colors, a blue and black gumball were positioned near the opening to ensure that children did not solely focus on the objects that were situated closer to the opening. The side of the correct gumball machine and the order each problem was presented were counterbalanced. Problems A, B, and C were counterbalanced in two blocks, with half of the participants completing version 1 in the first block. Problems D and E were always presented as problems 7 and 8, respectively. Problem D was presented second to last to so that we could assess whether children remained motivated throughout the task. Problem E was presented last; children were given a sticker for either choice, as black was the more likely outcome in both populations. Thus, it was presented last to ensure children did not expect to receive a sticker for a black gumball on subsequent problems.

Results and Discussion

Children received a score of 1 on each problem if they chose the machine that contained the higher proportion of blue (see Table 1 for means, standard deviations, and significance tests against chance for all problems).

Table 1: Children's performance in Experiment 1.

Problem	*M*	*SD*
A1	.82	.39
A2	.82	.39
B1	.90	.30
B2	.94	.23
C1	.82	.39
C2	.76	.43
D	.88	.32
E	.92	.27
Overall	.86	.17

Note: Individual problems were analyzed using binomial probabilities, overall score analyzed using single-sample t-test. All p values for the above analyses were ≤ .001.

We examined if children found some of the critical problem types (A through C) more difficult than others, and if they found problems with higher relative likelihoods easier to evaluate. To investigate this, we conducted a repeated-measures ANOVA with the critical problem types (A, B, C) and version (1, 2) as a within-subjects factor and child's age (younger half versus older half) as a between-subjects factor. There was a main effect of age, $F(1, 48) = 7.06$, $p = .01$, η^2_p

= .13, and problem type, $F(2, 96) = 4.15$, $p = .02$, $\eta^2_p = .08$, on children's scores. On average, the older children in the sample scored higher than the younger children (older children: $M = .91$, $SE = .04$; younger children: $M = .77$, $SE = .04$; $Mean_{Difference} = .14$, $p = .01$). Problem type B ($M = .92$, $SE = .02$) was significantly easier than problem type A ($M = .82$, $SE = .04$; $Mean_{Difference} = .10$, $p = .04$) and problem type C ($M = .79$, $SE = .05$; $Mean_{Difference} = .13$, $p = .01$). Problem version (i.e., relative likelihood) and all interactions were non-significant. Because problems D and E did not include these critical features and did not have a complement problem, we did not include them in these analyses. However, both problems were solved well-above chance (see Table 1). Performance on Problem D indicates that most children could still follow the task after they completed a number of more difficult problems. Moreover, the successful performance on Problem E suggests that children are not simply choosing the population with fewer non-targets.

To examine whether children's strong performance on probability problems was driven by learning over the course of the experiment (as the machines produced the more probable color on each problem type), we ran an additional repeated measures ANOVA with trial order (problem presented in place 1, 2, 3, 4, 5, 6) as a within-subjects factor. This analysis indicated that trial order did not significantly impact children's performance, $F(5, 240) = 1.22$, $p = .30$. Regardless of problem type, children performed well-above chance on trial 1 ($M = .90$, $SD = .3$, binomial, $p < .001$), which also suggests no effect of learning.

To summarize, Experiment 1 established that young children are able to solve probabilistic reasoning problems at rates well-above chance. Although the older children in our sample performed significantly better than the younger children, both age groups successfully solved the problems. Children performed significantly better on problem type B than types A and C, which is unsurprising due to the number of shortcuts they could have used to solve problems B1 and B2. Nevertheless, children still performed well on the more difficult problem types, suggesting that they do not solely rely on these heuristics.

Experiment 2

In Experiment 2, we presented a second group of children with more difficult problems to further test their use of various strategies. Because the children in Experiment 1 performed very well on our problems, we wanted to further explore their performance with a battery that contained some more challenging features (see Figure 2). Problems A1 and A2 presented children with two populations on the same side of ½, in which there were more targets and more overall objects in the less probable population. Problem types B and C presented children with two populations that had an equal number of target objects and more overall objects in the less probable population. Problems D1 and D2 contained more target objects and more overall objects in the less probable population. In this experiment, Problem E presented children with two uniform populations in which one population only contained blue gumballs, and the other contained only black (see Figure 1, Problem D). Because children in this experiment were presented with a more difficult set of problems, this problem was included again to gauge children's ability to follow the task. We included two versions of Problems A, B, C, and D to determine if probabilities that had higher relative likelihoods (i.e., problems that were further apart in probability, which were labeled with a 2), were easier for children to evaluate.

Methods

Participants Data from 74 3- and 4-year-olds were included in analyses (*mean age* = 4;2; *range* = 3;7 to 4;11). Again, the sample size was determined based on a power analysis for the larger study. An additional seven children were tested but were excluded from analyses due to parental report of atypical development ($n = 2$), parental report of very low English exposure (i.e., hearing English less than 50% of the time; $n = 2$), and not finishing the task ($n = 3$). Participants were recruited from a database of families and a daycare in the region. Children received a small gift for their time.

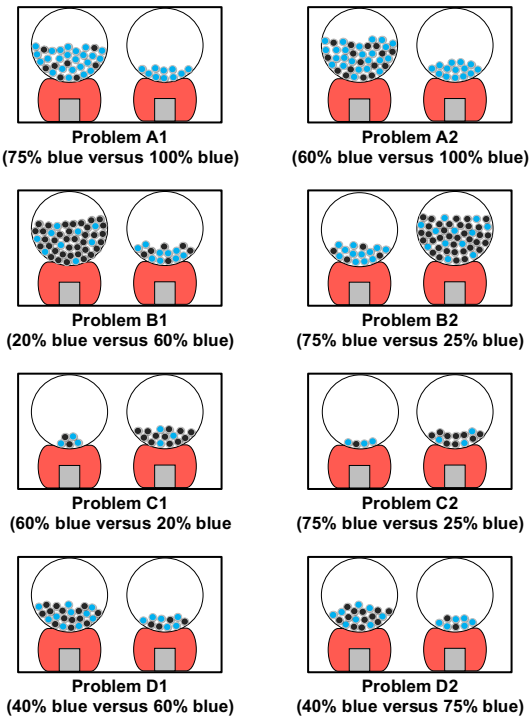


Figure 2: Probability problems presented in Experiment 2. *Note*: Problem E (not shown) was identical to Problem D in Experiment 1 *(see Figure 1)*

Materials and Procedure Participants were tested individually in a quiet room in the lab or at their daycare. The procedure was identical to Experiment 1, with the exception of the new battery of problems.

The probability problems were presented in the same manner as in Experiment 1. The side of the correct gumball machine and the order that each problem was presented were counterbalanced. Problems A, B, C, and D were

counterbalanced in two blocks, with half of the participants completing version 1 in the first block. To ensure that children remained motivated and followed the instructions throughout the task, Problem E was always presented last.

Results and Discussion

Children received a score of 1 on each problem if they chose the machine that contained the higher proportion of blue (see Table 2 for means, standard deviations, and significance tests against chance for all problems).

Table 2: Children's performance in Experiment 2.

Problem	*M*	*SD*
A1	.88	.33
A2	.88	.33
B1	.64	.49
B2	.76	.43
C1	.74	.44
C2	.84	.37
D1	.73	.45
D2	.78	.41
E	.93	.25
Overall	.80	.17

Note: Individual problems were analyzed using binomial probabilities, overall score analyzed using single-sample t-test. All p values for the above analyses were ≤ .001, with the exception of B1 (p = .03).

Similar to Experiment 1, we explored if children found some of the critical problem types (A through D) more difficult than others, and if they found problems with higher relative likelihoods easier to evaluate. To examine this, we conducted a repeated-measures ANOVA with the critical problem types (A, B, C, D) and version (1, 2) as a within-subjects factor and child's age (younger half versus older half) as a between-subjects factor. There was a main effect of problem type, $F(3, 216) = 5.36$, $p = .001$, $\eta^2_p = .07$, and version, $F(1, 72) = 4.65$, $p = .03$, $\eta^2_p = .06$, on children's scores. Problem type A ($M = .88$, $SE = .03$) was significantly easier than problem type B ($M = .70$, $SE = .04$; $Mean_{Difference} = .18$, $p < .001$) and problem type D ($M = .76$, $SE = .04$; $Mean_{Difference} = .12$, $p = .007$). Problem type C ($M = .79$, $SE = .03$) was marginally more difficult than problem type A ($Mean_{Difference} = -.09$, $p = .07$) and marginally easier than problem type B ($Mean_{Difference} = .10$, $p = .06$). Problems labeled with 2 ($M = .81$, $SE = .03$), comparisons that were further apart in relative likelihood, were significantly easier than problems labeled with 1 ($M = .75$, $SE = .03$; $Mean_{Difference} = -.07$, $p = .03$). Children's age and all interactions were non-significant. Because Problem E did not include these critical features and did not have a complement problem, it was not included in this analysis. However, as seen in Table 2, this problem was again solved well-above chance, indicating that most children still followed the task after they completed a number of more difficult problems.

To examine learning over the course of the experiment, we ran an additional repeated measures ANOVA that included counterbalanced trial order (problem presented in place 1, 2, 3, 4, 5, 6, 7, 8) as a within-subjects factor. There was an effect of trial order, $F(7, 511) = 2.31$, $p = .03$, $\eta^2_p = .03$, on children's scores, with scores improving over the session. Though this effect of order suggests that some learning may have occurred throughout the experiment, children performed well above chance on trial 1 ($M = .77$, $SD = .42$, binomial $p < .001$), indicating that learning did not entirely account for the strong performance.

In Experiment 2, we presented children with more challenging probabilistic reasoning problems. Although they were presented with this more difficult battery, children still performed at rates well-above chance across the age group. Problem type A was relatively easy for participants to solve, possibly because the correct option only contained target gumballs. Compared to the other problems, children found problem types B and D more difficult. On those problems, children were unable to rely on a number of cues, including the number of targets and the number of overall objects. We also found that children performed better on problems labelled with 2, which had relative likelihoods that were further apart and thus were easier to visually discriminate. Finally, although children performed above chance on the first trial, we observed an effect of trial order on performance, suggesting that learning may have contributed to performance.

General Discussion

In two experiments, we established that 3- and 4-year-old children are able to reason about probabilities at rates well-above chance. Though the older children in our sample performed significantly better than the younger children in Experiment 1, we did not find any age differences in performance in Experiment 2. Problems that contained multiple shortcuts or a deterministic outcome were easier for children to solve, and relative likelihoods impacted performance in Experiment 2 with our more difficult set of problems. Though children in both experiments performed above chance on the first trial, feedback may have affected children's scores over the course of Experiment 2.

Differences between our design and those of previous experiments may have facilitated performance in our paradigm. Children in our experiments were asked an age-appropriate question about probability. The verbal demands of the task affected preschoolers' performance in the past, as their performance suffered in paradigms with very high and very low verbal demands. In contrast to using verbal explanations as a dependent measure (as in Piaget & Inhelder, 1975), or using verbal cues that might have been too general (as in Girotto et al., 2016), children provided a forced-choice response to a simple, explicit question about probability. This method appears to have suited their abilities.

Children may have also found our gumball machine paradigm, which was presented on an iPad, engaging, and this may have helped maintain their interest over a number of trials. This design allowed us to display the contents of the gumball machine clearly, and the objects remained in view while the child made their choice. In some previous experiments, the experimenter sampled a hidden object from

each population and would ask the child to choose between the two hidden samples. Displaying the populations during the child's choice may have eliminated a working memory demand, because children did not have to maintain a representation of the populations during the sampling process. To disentangle the influence of these features, future work could again present children with two gumball machines on an iPad, though the populations would be covered while a hidden object is drawn from each machine. This would help us determine if clearly displaying the objects aids performance, and if hiding the objects during the sampling procedure creates a working memory demand.

We also provided children with feedback for their performance on each trial, and they were shown the most probable outcome from both populations after they made their choice. We used feedback to help sustain motivation over the course of the experiment because we were presenting very young children with multiple trials. Although we found no evidence of learning in Experiment 1, we found an effect of trial order on performance in Experiment 2. Nevertheless, children in both experiments performed above chance on the first trial prior to receiving any feedback. To further investigate learning in this context, future work could test the effectiveness of feedback at combating the use of overlearned strategies that approximate probabilistic inference. In turn, this work would shed light on how more sophisticated probabilistic reasoning strategies are acquired and fine-tuned with practice.

Moreover, the current experiments explored various strategies that children could use to approximate probabilistic inference. Though older children are drawn to populations with more target objects (i.e., denominator neglect; Falk et al., 2012), preschoolers in our experiments performed well on problems in which the less probable population contained more target objects, and when the number of target objects were equated. One notable difference between our problems and those that older children struggled with is that older children are typically presented with more difficult problems, in which the relative likelihoods are more difficult to discriminate. In our problems, the relative likelihoods were more distinct, making the problems easier overall. Surprisingly, preschoolers were drawn to populations with more *overall* objects (that is, target plus non-target). In both experiments, children's performance was slightly worse on problems where the less probable population noticeably contained more objects. Though older children are not drawn to populations with more objects (Falk et al., 2012), the current findings suggest that the overall number of objects is a salient feature for preschoolers. Because of this somewhat surprising finding, we are currently developing a battery of problems to further clarify how features of the problem, such as overall objects and number of targets, affect young children's probabilistic reasoning performance. Future work with a larger age range could also investigate how use of different strategies varies over the course of development.

We presented preschoolers with problems that were on the same side of ½ to explore nuances in their ability to compare proportions. Children in both experiments were able to make these comparisons and considered the proportion of objects in each population, even though the less probable population contained more target objects. Because we were unsure if preschoolers could solve these more difficult problems, the more probable population was uniform and only contained target objects. Inclusion of the uniform population allowed for a straightforward assessment of children's reasoning abilities, serving as a first step in pitting true proportional reasoning against a heuristic that focuses on the absolute number of target objects. Although this first step established that they are able to make these comparisons, future work should present preschoolers with two *probabilistic* populations (i.e., both contain target and non-target objects) on the same side of ½. This comparison is more difficult, because children are comparing two probabilistic populations and, by the nature of this design, the relative likelihoods are closer together. Though relative likelihood did not influence performance on Experiment 1, it impacted preschoolers' responses on the more difficult battery in Experiment 2. Thus, future work should continue to test the impact of relative likelihood on preschooler's performance, notably in cases where both populations are on the same side of ½.

Finally, we used two sets of problems to assess preschoolers' probabilistic reasoning in the current experiments. Though both batteries indicated that children could successfully reason about probability, one may wonder which battery would best provide an overall assessment of a child's abilities. For space and intended focus of the current paper, we did not report the results of a large set of individual difference measures that were collected with the children that assessed their executive function and receptive vocabulary abilities. These measures tend to correlate well with children's quantitative and general reasoning abilities during early childhood. However, the battery used in Experiment 1 correlated well with these measures, while the battery in Experiment 2 did not show as strong of a relationship. Therefore, at the present time, the problems in Experiment 1 might be the best set to use when gauging children's abilities in future work. We are currently working on another battery of problems, which include some problems from Experiments 1 and 2, and some additional problems of even greater difficulty to continue refining the set.

In sum, preschool children in both experiments solved probabilistic reasoning problems at rates above chance. The current findings illustrate the importance of using an age-appropriate paradigm when establishing the abilities of a particular population. Though children did not rely solely on erroneous strategies, future work is needed to explore how features of probabilistic problems impact performance.

Acknowledgments

These experiments were conceived of in collaboration with Tara McAuley and Bethany Nightingale as part of a larger project and we thank them for their insights. We thank members of the Developmental Learning Lab for help with data collection. Special thanks to parents and children for

participating. This research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada to S. D.

References

Denison, S., Konopczynski, K., Garcia, V., & Xu, F. (2006). Probabilistic reasoning in preschoolers: Random sampling and base rate. *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (pp. 1216–1221).

Denison, S., & Xu, F. (2010). Twelve- to 14-month-old infants can predict single-event probability with large set sizes. *Developmental Science, 13*(5), 798-803.

Denison, S., & Xu, F. (2014). The origins of probabilistic inference in human infants. *Cognition, 130*(3), 335-347.

Eckert, J., Rakoczy, H., & Call, J. (2017). Are great apes able to reason from multi-item samples to populations of food items? *American Journal of Primatology, 79*(10).

Falk, R., Yudilevich-Assouline, P., & Elstein, A. (2012). Children's concept of probability as inferred from their binary choices—revisited. *Educational Studies in Mathematics, 81*(2), 207-233.

Girotto, V., Fontanari, L., Gonzalez, M., Vallortigara, G., & Blaye, A. (2016). Young children do not succeed in choice tasks that imply evaluating chances. *Cognition*, *152*, 32–39.

Girotto, V., & Gonzalez, M. (2008). Children's understanding of posterior probability. *Cognition, 106,* 325 – 344.

Goldberg, S. (1966). Probability judgments by preschool children: Task conditions and performance. *Child Development*, 157-167.

Hoemann, H. W., & Ross, B. M. (1971). Children's understanding of probability concepts. *Child Development*, 221-236.

Kushnir, T., Xu, F., & Wellman, H. M. (2010). Young children use statistical sampling to infer the preferences of others. *Psychological Science, 21*(8), 1134–1140.

Ma, L., & Xu, F. (2011). Young children's use of statistical sampling evidence to infer the subjectivity of preferences. *Cognition, 120*(3), 403-411.

Piaget, J., & Inhelder, B. (1975). *The origins of the idea of chance in children.* New York, NY: Norton & Company.

Rakoczy, H., Clüver, A., Saucke, L., Stoffregen, N., Gräbener, A., Migura, J., & Call, J. (2014). Apes are intuitive statisticians. *Cognition, 131*(1), 60-68.

Téglás, E., Girotto, V., Gonzalez, M., & Bonatti, L. L. (2007). Intuitions of probabilities shape expectations about the future at 12 months and beyond. *Proceedings of the National Academy of Sciences, 104*(48), 19156-19159.

Téglás, E., Vul, E., Girotto, V., Gonzalez, M., Tenenbaum, J. B., & Bonatti, L. L. (2011). Pure reasoning in 12-month-old infants as probabilistic inference. *Science, 332*(6033), 1054-1059.

Xu, F., & Garcia, V. (2008). Intuitive statistics by 8-month-old infants. *Proceedings of the National Academy of Sciences, 105*(13), 5012-5015.

Yost, P. A., Siegel, A. E., & Andrews, J. M. (1962). Nonverbal probability judgments by young children. *Child Development, 33*(4), 769-780.

Looking Patterns during Analogical Reasoning: Generalizable or Task-Specific?

Abstract

Given the importance of developing analogical reasoning to bootstrapping children's understanding of the world, why is this ability so challenging for children? Two common mechanisms have been implicated: 1) children's inability to prioritize relational information during initial problem solving; 2) children's inability to disengage from salient distractors. Here, we use eye tracking to examine children and adults' looking patterns when solving scene analogies, allowing for differentiation between attention to relations versus to featurally salient distractors. In contrast to a recent study with propositional analogies, our data suggest prioritization of source information does not differ between adults and children, nor is it predictive of performance; however, children and adults attend differently to distractors, and this attention predicts performance. These results suggest that feature-based distraction is a key way children and adults differ during analogical reasoning, and that the analogy problem format should be taken into account when considering children's analogical reasoning.

Keywords: analogy, attention, eye tracking, reasoning, pattern recognition

Introduction

Analogical reasoning involves identifying higher order similarities in relational structure shared between representations. This form of reasoning is used in many contexts, and is predictive of academic and professional success (Richland & Burchinal, 2013). Yet, analogical reasoning proficiency develops over time: young children often struggle to notice or extract deep underlying structures from comparison opportunities (Gentner & Smith, 2013). Given the importance of developing this ability, researchers have asked why analogical reasoning problems are challenging for children. Two common explanations in the literature implicate: 1) children's inability to prioritize attending to relational information during initial problem solving, or 2) children's inability to disengage from featurally salient distractors.

Prioritization of Relational Information

One explanation for children's difficulty with analogical reasoning problems is that they fail to attend to relational information that is crucial for correctly solving problems. Much of this work has used propositional analogies, in the format A:B::C:D. In these analogies, participants select from four choices a D item that is relationally similar to the C item in the same way that A and B are similar. For example, if A and B are both triangles, with B being a stretched version of A, and C is a square, the correct choice for D would be a stretched square. A featural distractor in the response choices might be a diamond of the same color as the square – color being a salient perceptual feature that could distract from the deeper, structural relation between C and the stretched square.

From eye tracking work, we know that adults generally attend to the A:B pair before fixating on C and the response choices, showing that they can maintain the overarching goal (i.e. find the picture that goes with C in the same way that A goes with B) (Starr, Vendetti, & Bunge, 2018). In contrast, 5- and 6-year-old children ignore the A:B items, and focus their attention on C and the response choices (Glady, French, & Thibaut, 2017; Thibaut & French, 2016). This suggests that children do not extract relational information before considering response options, instead focusing on the immediate task goal (i.e. find the picture that goes with C).

In support of this idea, using linear discriminant analysis, French and Thibaut (2014) found that children's visual attention during the first third of the trial can predict with 64% accuracy whether or not the problem would be answered correctly. This is especially true if attention is focused on the A:B pair. Glady and colleagues (2017) have shown that guiding children's attention to the A:B pair during initial problem solving significantly improved children's performance.

Featurally Salient Distractors

While attention during the task may be important, an alternative explanation for children developing proficiency on analogy problems emphasizes the effect of featurally salient distractors. In many situations that require analogical reasoning, the visual scene is complex. Although a higher order relational structure is present, children are more likely to make judgments based on mere appearance or surface-level similarities between representations – attending to items that are perceptually or semantically related to the item in question, rather than structurally related. Young children are particularly susceptible to this type of error, tending to shift from more object-based similarity matching to more relational reasoning over time, defined as the *relational shift* (Gentner, 1988). Adults also appear to make relational shifts when reasoning about information for which they have low knowledge, yet children tend to make featural errors even when reasoning about relations that are familiar (Richland, Morrison & Holyoak, 2006). This finding has led researchers to suggest that the inability to disregard salient featural information in favor of relational information may be, at least in part, attributed to still developing executive function (EF) resources, and that gains in EF allow children to increasingly manipulate complex relations in working memory and direct attention toward relevant aspects of an analogy (Richland et al., 2006; Simms, Frausel & Richland, 2018).

Behaviorally, this explanation has been supported using a variety of analogical reasoning tasks (i.e. scene analogy and propositional analogy paradigms). For example, Richland and colleagues (2006) asked children to identify relational similarities between two scenes (e.g. a source and target scene), while ignoring items with featural similarities. In their task, the goal was to identify something in a target scene that

corresponded relationally to a prompted item in a source scene. Importantly, a featural distractor, an item in the target scene that was not incorporated in the relation of focus and had great surface similarity to the prompted item in a source scene, was sometimes present in the target scene (Richland et al., 2006). For example, a pair of scenes might depict a relation of a dog chasing a cat (source scene) and a man chasing a woman (target scene). If the dog was prompted, the correct choice would be the man and the incorrect featural choice would be a perceptually similar dog in the target scene. For children ages 3-4, the perceptually similar match was an effective featural distractor, such that accuracy for the problems with distractors was 15% less than that for the problems without distractors. Individual differences in children's EF (working memory in particular) explained these patterns of performance (Simms et al., 2018). Further, these behavioral findings have been complemented by modeling work: Simulations in the LISA computational model of analogy (Hummel & Holyoak, 1997, 2003) suggest that changes in inhibition levels, along with relational knowledge accretion, account for young children's difficulty when reasoning analogically (Morrison, Doumas, & Richland, 2011; see also Doumas, Morrison, & Richland, 2018). Using the same task, the model replicates the experimental findings of Richland and colleagues (2006), such that the model was more likely to choose a featurally similar distractor object than an analogically correct choice.

Thibaut and colleagues (2010) demonstrated a similar effect of featural distractors using propositional analogies. Similar to scene analogy paradigms, correct responses require inhibition of salient features and a focus on common relational structure. As with scene analogies, children were more prone to errors when featural distractors were present. Indeed, later work using eye tracking revealed a negative association between the amount of time looking to a distractor and performance, such that the more time children spent looking at the distractor the worse they performed (Thibaut & French, 2016).

Distractor versus Prioritization

Whereas the majority of previous literature has considered these two mechanisms separately, Starr and colleagues (2018) examined both how looking to featural distractors and focusing on source relational information affected children's ability to solve propositional analogies. They argued that children's poor performance was due to an inability to prioritize attending to the A:B relation when initially processing an analogy, rather than an inability to disengage from perceptual lures. What is unknown is whether this finding is unique to propositional analogies, or consistent across all analogy types.

Current Study

Here, we examined visual attention while children and adults solved scene analogy problems similar to those used by Richland and colleagues (2006). If the main factor underlying children's poor performance on scene analogy problems is their non adult-like looking patterns (characterized in propositional analogies as a prioritization of relational information – A:B pair – during early problem solving) we should find that adults show greater attention to the source scene and key relationship than children, especially early in problem solving. We should also find that attending to the source relation predicts performance. Indeed, we already know that adults initially focus on the relations within a source scene – prioritizing the existing structural relation before considering the target scene (Gordon & Moser, 2007). However, if we do not find this difference in visual attention between adults and children, this would suggest that whereas adults may have a systematic approach to solving all analogy problems, the format of the problem may have a strong influence on how children solve these problems. In this case, Starr and colleagues' findings would be specific to propositional analogy problems.

The scene analogy task also allows us to measure looking to the featural distractor, determining whether children's looking patterns appear similar to or systematically different from adults'. Thus we will examine both looking to the source relation, as well as attention to featural distractors to assess which of these attentional mechanisms best explain children's developmental trajectory in solving scene analogies.

Methods

Participants

Data from 57 4- and 5-year-old children (29 females, M_{age} = 4.88, SD_{age} = 0.47) and 45 adults (37 females, M_{age} = 19.45, SD_{age} = 0.99) were analyzed for the present study[1]. Participants represented a diverse sample from a large metropolitan city. Children were recruited from schools and participated individually in one experimental session during a regular school day. Children were compensated with stickers and a certificate noting their participation in a research study. Adults were recruited from a participant pool at a university and participated individually in a lab setting.

Materials

Stimuli. Participants were shown scene analogies adapted from Richland et al. (2006). Each stimulus included a pair of scenes presented simultaneously on a *15-inch* Dell laptop. Pairs of scenes depicted one of two *relation categories* (i.e. chasing or reading) occurring between items (i.e. animals or people) within the scenes. Source scenes contained five

[1] Data from 57 children and 60 adults was collected. Although all children were included in analyses, a subset of data from particular timepoints were excluded from 8 children based on insufficient usable eye tracking data. Five adult participants were excluded for having lacking sufficient eye tracking data. For adult participants to be included in the sample, they must have > 75% accuracy. This was to ensure that we had a measure of successful, mature visual attention patterns. Ten adult participants were excluded for having < 75% accuracy across trials.

items: the two items within the relation that participants were to attend to, and three additional items (i.e., neutral inanimate objects). Target scenes also contained 5 items: the two items within the relation, two additional items, and a featural distractor.

Figure 1a. shows an example of a "*chasing"* relation depicted in both a source and a target scene. The source scene on the left shows a tiger *chasing* a woman (items within the chasing relation), and the corresponding target scene on the right side shows a lion *chasing* a horse (items within the chasing relation). Target scenes also contained a featural distractor that was featurally similar to the prompted source-scene item. In Figure 1a., the tiger in the target scene serves as the distractor because the tiger is prompted (i.e. circled) in the source scene. To maintain the same number of items across scenes, additional items were included. These items were *neutral*, meaning they were not involved in the chasing relations and were not the distractor (in Figure 1a, source scene: dog house, jeep, and plant, target scene: barn and soccer ball). Importantly, the distractor is never involved in the relation within the scene. Distractors were centrally located, increasing the likelihood that participants would notice them.

Figure 1b. shows an example of a "*reading"* relation in both a source and a target scene. Items depicting the reading relation were oriented towards each other with one character reading to the other character. In all source scenes, one of the two items within the relation of chasing or reading was prompted with a circle. The directionality of relations within a pair of scenes was reversed to avoid children making choices based on spatial location alone. For example, in Figure 1a., if chasing is depicted between characters to the left in a source scene, the chasing would then be depicted to the right in the target scene.

Eye Tracker. Eye tracking data were collected via corneal reflection using a TobiiPro X3-120 remote eye tracker affixed to a *15-inch* Dell laptop. Tobii software was used to perform a 5-point calibration procedure using standard animation blue dots. This step was followed by the collection and integration of gaze data with the presented instructional videos (described below) using Tobii Studio (Tobii Technology, Sweden). All gaze data was extracted from Tobii Studio Software for each participant.

Procedure

For the purpose of the present question, we considered a subsection of data from a longer study: eye tracking data during which children and adults visually attended to scene analogy problems without any training on how to solve them. For children, the data came from 12 pretest problems (6 chasing; 6 reading), after which children received training on how to solve scene analogies and completed 12 posttest problems. For adults, the data came from 24 problems (12 chasing; 12 reading). Items included in a child's pretest and posttest were counterbalanced, and all items were shown to adults.

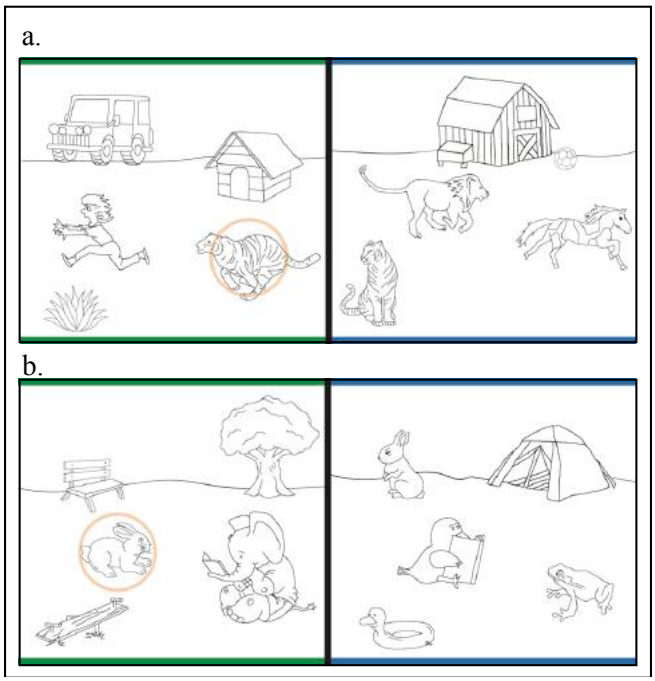


Figure 1. a. Example trial of *chasing* relation category. b. Example trial of *reading* relation category.

Introduction to Task and Calibration. Participants were told they were going to play a picture game and shown one example trial, orienting them to the layout of test trials (i.e., two pictures with different colored borders), and their task (i.e., that for each set of scenes their job was to figure out the pattern in the pictures). The experimenter described the chasing relation and asked the participant to solve the relation. For children, the explanation was repeated until they chose the correct item. This introduction ensured that when children incorrectly answered a trial, it was not because of a misunderstanding about the goal of the task.

Next, calibration on the eye tracker was completed: participants were seated approximately 40 cm in front of the laptop, familiarized with the eye tracker, and told it was important to remain still throughout the session.

Task. Participants completed a set of scene analogies while their visual attention was monitored. Participants were instructed to respond verbally to "*Which thing in the picture with the blue edges is in the same part of the pattern as the circled thing in the picture with the green edges?*". The task was self-paced, but if no response was given after a few seconds, the experimenter re-prompted. Responses were recorded for each trial.

Results

Areas of interest (AOIs) were generated for the items within the scene pairs using Tobii Studio (i.e. each trial had 10 AOIs, 5 in each scene). The remaining spaces outside of these AOIs were collapsed into an "Other" AOI. For analyses, we

considered visual attention to 1) the source relation (comprised of two relational items and analogous to the A:B items in propositional analogies) and 2) the distractor (analogous to a choice item in propositional analogies). Data were extracted and processed so that the AOI a participant fixated could be determined at 8 msec intervals across the entire length of each problem. Proportion of time spent looking to each AOI was calculated using the total gaze duration of a given trial (e.g., 1000 msec), and the amount of time spent looking at a given AOI during a particular trial. All analyses considered visual attention patterns and accuracy at the trial level, not aggregated across trials for a given participant.

Because prior work (Starr et al., 2018; French & Thibaut, 2014) suggests visual attention during initial solving has the most predictive power for whether a participant will arrive at the correct answer, we consider proportion of looking to these AOIs across the entire trial, as well as proportion of looking during an initial segment of each problem. In prior work, participants had set time limits for solving problems, thus, researchers could consider a set amount of time (e.g., the first third or fourth of a trial) when examining attention at the beginning of problem solving. Here, we used a self-paced design, which resulted in variability of trial length both across and within participants. Therefore, we considered the first 5 fixations of each trial as the first interval of problem solving.

Prioritization of Relational Information

Our primary goal was to establish whether visual attention to the source relation during scene analogy problems differed between age groups in the same way as for propositional analogies. Figure 2 shows the proportion of visual attention allocated to AOIs for both children and adults. In contrast to previous work using propositional analogies, both children and adults attended to the source relation about one-third of the time, across the entire solution time (adults: $M = 0.34$, $SD = 0.05$; children: $M = 0.33$, $SD = 0.66$). A generalized linear model supported the interpretation that attention to the source relation did not differ by age group ($\beta = -0.01$, $SE = 0.01$, $t = -0.39$, $p = 0.70$). Focusing just on initial solution time revealed that a higher proportion of looking to these items occurred when participants first viewed these problems than across the entire trial: adults spend nearly half of early problem-solving time focused on this relation ($M = 0.52$, $SD = 0.10$), and children allocated just under half of their attention to these items ($M = 0.44$, $SD = 0.18$). Again, there was no significant difference in this looking pattern between groups ($\beta = 0.00$, SE = 0.01, $t = 0.24$, $p = 0.81$).

In order to make conclusions about whether looking to the source relation supports successful reasoning, we must assess the relation between performance and visual attention patterns. Unsurprisingly, children performed poorly on scene analogies, answering less than one-third of the problems correctly ($M = 0.30$, $SD = 0.26$), whereas adults were much more accurate ($M = 0.92$, $SD = 0.06$). Because adults performed nearly at ceiling, we only assess whether looking patterns predict accuracy for child participants.

Binomial generalized linear models, with accuracy on each problem (0, 1) as the dependent measure, were used to determine whether looking to source relation is predictive of behavioral performance. In contrast to prior work, we found no relation between performance and looking to the source relation across the entire trial ($\beta = 0.30$, SE = 0.82, t = 0.34, p = 0.71), or during initial problem solving ($\beta = 0.05$, SE = 0.33, t = 0.16, p = 0.87) and learning. Overall, these results challenge previous work suggesting that children's lower performance on analogy problems can be explained by failures to attend adequately to the source relationship.

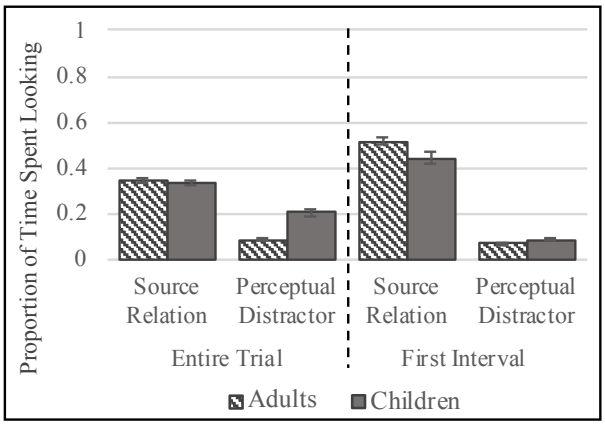


Figure 2. Proportion of time looking to source relations and distractors across entire trials and the first interval of trials.

Featurally Salient Distractors

When children and adults were not looking at the source relation, how did they allocate their attention? In line with previous work assessing children's visual attention during analogical reasoning, children spent roughly 10% percent more of their time looking to featural distractors ($M = 0.20$, $SD = 0.10$) as compared to adults ($M = 0.09$, $SD = 0.02$) across the entire problem-solving time. A generalized linear model indicated that children spend reliably more time looking towards the distractor than did adults ($\beta = 0.09$, $SE = 0.01$, $t = 6.31$, $p < .001$). However, during initial problem solving, both children and adults spent less than 10% of their time looking to the distractor (adults: $M = 0.07$, $SD = 0.03$; children: $M = 0.09$; $SD = 0.06$). A generalized linear model supported a lack of difference between age groups ($\beta = -0.05$, $SE = 0.03$, $t = -1.75$, $p = 0.08$). This indicates that, at first, children and adults explore the distractor equally, but children continue assessing the distractor throughout the trial.

Finally, we asked whether behavioral accuracy was predicted by looking to the distractor. Interestingly, in line with previous work, children performed better if they spent less time looking to the distractor across the entire problem-solving time ($\beta = -9.21$, $SE = 1.29$, $t = -7.16$, $p < .001$). However, when considering initial looking times only, there was no relation between looking to the distractor and performance ($\beta = 0.88$, $SE = 0.62$, $t = 1.41$, $p = 0.16$),

suggesting that initial attention to the scenes was not the key differentiating period.

Discussion

While previous work seems to be at a consensus about the differences between mature and immature visual attention patterns while solving propositional analogies, the current study was conducted to see if these patterns hold when children and adults solve scene analogies, which are arguably more similar to real world analogies. Specifically, previous work shows that when solving propositional analogies, children look less to the source relation (A:B) than adults, and that prioritizing attention to relational information early in analogical problem solving is predictive of later accuracy (Glady et al., 2017; Starr et al., 2018; Thibaut & French, 2016). Here, we do not find differences in patterns of attention to the source relation when solving scene analogies between children and adults, neither in their initial attention nor in the full problem-solving period. Furthermore, we do not find that children's attention to relational information is predictive of their performance, even though we replicate the pattern that children perform significantly worse than adults. We do, however, find that attention to the featural distractor predicted accuracy in children. Together, these results suggest the format of the problem influences attentional patterns and that prioritization of relational information is not always critical for successful problem solving across all analogy paradigms.

While both propositional and scene analogies require processing relational information in order to arrive at a correct solution, their structures differ significantly. This difference in structure may account for why children approach these problems in different ways. In analogies of A:B::C:D format, children who are not skilled at analogical reasoning seem to overlook the relational information contained in the A:B pair, and focus on 'C' and the response options, because they interpret the task as 'match 'C' to something' and treat A:B as irrelevant. In the example used previously from Thibaut & French (2016), children might ignore that 'shape' is the relational structure, such that B is a stretched version of A, and be more likely to pick an option that is similar to C on another dimension, such as 'color'. It seems that the salience of the A:B pair is not great enough to warrant attention from those children who do not understand the task goal. In contrast, in a scene analogy, children's visual attention is still drawn to the source relation initially, perhaps due to the circled item. Based on our results, the salience of the circled item draws both children and adults' attention equally at first, but unlike adults, children less often utilized that information to correctly solve the problem. Furthermore, the presence of a distractor lowered children's performance and drew children's attention. While looking to the source relation is obligatory for children and adults because of the circled item's salience, this looking pattern does not uniformly result in successful analogical reasoning.

Previous work has consistently demonstrated that adults prioritize relational information during initial problem solving, characterized by looking to the source relation (Gordon & Moser, 2007) or the A:B pair (Starr et al., 2018; Thibaut & French, 2016). Perhaps because adults understand that they need to identify the deeper structure in these problems, and therefore, are more proficient analogical reasoners, they are not restricted by the structure of the problem. Adults can organize their visual search in a particular way despite analogy format, whereas children's visual search during analogical reasoning is strongly influenced by problem structure, and, as will be discussed next, the presence of a featural distractor.

The secondary goal of this work was to ask if looking patterns to featural distractors are comparable between scene analogies and propositional analogies (Thibaut et al., 2010; Thibaut & French, 2016). Across the entire problem-solving episode, children allocated more of their attention to the distractor than adults, and this was negatively related to behavioral performance. This corroborates previous work using propositional analogies (Thibaut & French, 2016). However, when we only considered the first interval of problem-solving time, children and adults allocated an equal amount of time to the distractor, and this was *not* predictive of performance. This differs from previous work that has stressed the importance of initial looking patterns for predicting accuracy.

Based on these results, we can suggest that when solving scene analogies, adults and children both consider the distractor, but children continue their examination of the distractor across the entire trial. It is this continued focus on the distractor that leads to poor behavioral performance – initial consideration may be indicative of children and adults processing the items that appear in the source and target scenes before working to solve the problem.

Overall, incorporating our findings about children and adult's visual attention across the problem-solving process and during initial solving, and attention towards the source relation and distractor, we can conclude that children and adults organize their visual search in different ways when solving analogical reasoning problems: In processing analogies, adults begin by identifying the relational information necessary to understand the structure of the analogy. In contrast, children have more disorganized looking patterns, such that their visual search is dependent upon analogy format, rather than the overarching goal to identify relational structure. The consistent effect of featural distractors on children's visual attention, across analogy formats, lends further support for the conclusion that children have inefficient looking patterns. While consistently looking to distractors could be considered an 'organized looking pattern' because they perform this behavior somewhat reliably, in this case, it demonstrates children's difficulty attending to underlying relational structure. This conclusion is in line with the work of Glady and colleagues (2010), who found a clear difference between adults and children's visual strategies when solving analogy problems, such that adults have more organized search patterns (Glady, Thibaut, & French, 2010).

Although our work adds an important piece to understanding the development of analogical reasoning ability, it should be noted that one limitation of this work lies in the restricted comparisons that can be made between propositional and scene analogies. Specifically, previous analogical reasoning research has made strong conclusions about the differences between age groups in terms of looking to the C item in propositional analogies (A:B::C:D), such that children look more to the C item earlier in the problem solving process and focus their search around C, whereas adults will search in a more organized way by first examining the A:B pair and then looking at the C item and the possible answers (Starr et al., 2018; Thibaut & French, 2016). Unfortunately, there is not a functionally comparable item to C in a scene analogy. Items C and D are already in relation with one another in a scene analogy, whereas D must be chosen from multiple options by the participant in a propositional analogy. Therefore, in this study, we cannot make conclusions about looking to the C item. This, again, speaks to the structural difference between propositional analogies and scene analogies.

Overall, our results suggest that while there are some generalizable differences between how adults and children process analogies regardless of their format, there are other aspects of how attention is allocated that are dependent upon analogy type. These results allow us to resolve inconsistencies in previous work by identifying exactly how children's visual attention differs across analogy formats. Gaining a better understanding about these differences across the domain of analogical reasoning will better elucidate the attentional mechanisms underlying learning in this domain and inform teaching techniques. Determining how children view analogy problems will help us understand what underlies this behavioral ability in children and adults, and could lead to evidence-based practices for teaching analogical reasoning through guided looking. This work, and future work in this field, can begin to inform practical instructional techniques by helping educators design instruction that reaches diverse classrooms of learners, as they struggle to develop this difficult, yet important ability: analogical reasoning.

References

Doumas, L.A.A. Morrison, R.G. & Richland, L.E. (2018). Individual differences in relational learning and analogical reasoning: A computational model of longitudinal change. *Frontiers in Psychology*. doi: 10.3389/fpsyg.2018.01235

French, R. M., & Thibaut, J.P. (2014). Using eye-tracking to predict children's success or failure on analogy tasks. *Proceedings of the Thirty-Sixth Annual Meeting of the Cognitive Science Society*, 2222–2227.

Gentner, D. (1988). Metaphor as Structure Mapping: The Relational Shift. *Child Development*, *59*, 47–59.

Glady, Y., French, R. M., & Thibaut, J. P. (2017). Children's failure in analogical reasoning tasks: A problem of focus of attention and information integration? *Frontiers in Psychology, 8*, 1–13. doi.org/10.3389/fpsyg.2017.00707

Glady, Y., Thibaut, J.P., & French, R. (2010). Visual strategies in analogical reasoning development: a new method for classifying scanpaths. *Proceedings of Thirty-Fifth Annual Meeting of the Cognitive Science Society*, 2398–2403. doi.org/10.13140/2.1.4107.1365

Gordon, P. C., & Moser, S. (2007). Insight into analogies: Evidence from eye movements. *Visual Cognition, 15*, 20–35. doi.org/10.1080/13506280600871891

Gentner, D., & Smith, L. A. (2013). Analogical learning and reasoning. In Reisberg, D. (Ed.), *The Oxford handbook of Cognitive Psychology* (668-681). New York, NY: Oxford University Press.

Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review, 104,* 427-466.

Hummel, J. E., & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review, 110,* 220-264.

Morrison, R. G., Doumas, L. A. A., & Richland, L. E. (2011). A computational account of children's analogical reasoning: Balancing inhibitory control in working memory and relational representation. *Developmental Science, 14*, 516–529. doi.org/10.1111/j.1467-7687.2010.00999.x

Richland, L. E., & Burchinal, M. R. (2013). Early executive function predicts reasoning development. *Psychological Science, 24*, 87–92. doi.org/10.1177/0956797612450883

Richland, L. E., Morrison, R. G., & Holyoak, K. J. (2006). Children's development of analogical reasoning: Insights from scene analogy problems. *Journal of Experimental Child Psychology, 94*, 249–273. doi.org/10.1016/j.jecp.2006.02.002

Simms, N., Frausel, R. R., Richland, L. E. (2018), Working Memory Predicts Children's Analogical Reasoning. *Journal of Experimental Child Psychology,* 166, 160–177.

Starr, A., Vendetti, M. S., & Bunge, S. A. (2018). Eye movements provide insight into individual differences in children's analogical reasoning strategies. *Acta Psychologica, 186*, 18–26. doi.org/10.1016/j.actpsy.2018.04.002

Thibaut, J. P., & French, R. M. (2016). Analogical reasoning, control and executive functions: A developmental investigation with eye-tracking. *Cognitive Development, 38*, 10–26. doi.org/10.1016/j.cogdev.2015.12.002

Thibaut, J. P., French, R., & Vezneva, M. (2010). The development of analogy making in children: Cognitive load and executive functions. *Journal of Experimental Child Psychology, 106*, 1–19. doi.org/10.1016/j.jecp.2010.01.001

The Social Network Dynamics of Category Formation

Douglas Guilbeault
The University of Pennsylvania, PHILADELPHIA, Pennsylvania, United States

Andrea Baronchelli
City University, London, London, United Kingdom

Damon Centola
The University of Pennsylvania, PHILADELPHIA, Pennsylvania, United States

Abstract

How do societies develop categories for continuous sets of novel phenomena, as in the domains of art and technology? Seminal work in the nativist tradition argues that given the same stimuli, people can independently produce the same categories as a result of universal cognitive constraints. These constraints are said to account for cross-group coherence, where distinct communities and cultures have been shown to arrive at highly similar categories. Cross-group coherence is widely seen as incompatible with functionalism, which holds that categories are defined through communication, leading to divergent category systems. Here, we use an experiment to demonstrate that communication can generate either the divergence or convergence of category systems, depending on the size of the social network (2, 6, 8, 24, and 50). We find that large social networks amplify population biases, where a subset of slightly more frequent words become exponentially more likely to spread as network size increases.

Evaluating Models of Human Adversarial Behavior Against Defense Algorithms in a Contextual Multi-Armed Bandit Task

Marcus Gutierrez (mgutierrez22@miners.utep.edu)
Computer Science Department, University of Texas at El Paso

Jakub Černý (cerny@disroot.org)
Computer Science Department, Nanyang Technological University

Noam Ben-Asher (noam.ben.asher@gmail.com)
Army Research Laboratory

Efrat Aharonov (efrat.aharonov@gmail.com)
Department of Social & Decision Sciences, Carnegie Mellon University

Branislav Bošanský (branislav.bosansky@agents.fel.cvut.cz)
Agent Technology Center, Computer Science Department, Czech Technical University in Prague

Christopher Kiekintveld (cdkiekintveld@utep.edu)
Computer Science Department, University of Texas at El Paso

Cleotilde Gonzalez (coty@andrew.cmu.edu)
Department of Social & Decision Sciences, Carnegie Mellon University

Abstract

We consider the problem of predicting how humans learn interactively in an adversarial Multi-Armed Bandit (MAB) setting. In a cybersecurity scenario, we designed defense algorithms to assign decoys to lure attackers. Humans play the role of cyber attackers in an experiment to try to learn the defense strategy after repeated interactions. Participants played against one of three defense algorithms: a stationary strategy, a static game-theoretic solution, and an adaptive MAB strategy. Our results show that humans have the most difficulty learning against the adaptive defense. We also evaluated five different models of attack behavior and compared their predictions against human data. We show that a modified version of Thompson Sampling and a cognitive model based on Instance-Based Learning Theory are the best at replicating human learning against defense strategies. We discuss how these models of human attacker can inform future cyberdefense tools.

Keywords: Cognitive Modeling; Reinforcement Learning; Intelligent Agents; Decision Making; Cybersecurity

Introduction

With the popularity of autonomous systems, the question of how humans interact with these systems becomes increasingly important (Gershman, Horvitz, & Tenenbaum, 2015). Humans are imperfect agents, but they are capable of learning and in some settings able to adapt to novel situations. Our ability to anticipate human behavior, to represent human decision making computationally, and to use these predictions to improve autonomous agents is critical to making autonomous systems more capable and secure.

We study an adversarial decision making setting framed in the context of cybersecurity. Humans attackers try to compromise a network while automated defender algorithms deploy decoys in the network (i.e., honeypots) to detect and thwart attackers. Honeypots are designed to waste the attacker's resources and provide information to the defender (Spitzner, 2003). Attackers try to avoid detection by honeypots. Deploying a fixed configuration of honeypots (i.e., a static defense) may capture an attacker in a single interaction. However, an adaptive attacker may learn the static honeypot defenses and actively avoid them in future interactions. A defender who can predict this attack learning dynamic should be able to deploy defensive strategies that are harder to learn and defeat over the long term. Our goal is to determine how human attackers behave against defense algorithms of various complexities, and to test cognitive models of adversarial behavior against other common behavioral models.

We model a cybersecurity scenario as a repeated Multi-Armed Bandit task (MAB) where a human attacker plays against an automated defender. MAB tasks have been useful in the study of human decision making, characterizing the common exploration-exploitation tradeoff (e.g., (Steyvers, Lee, & Wagenmakers, 2009)). However, our goal is to determine how a human attacker is able to learn the defender's deception strategy and avoid honeypots based on previous experience.

In a standard MAB, a decision maker select arms on a "slot machine" in each round and observes the outcome, typically with the value of each arm in the range $[0,1]$. The adversarial MAB considers an adversary (i.e., the algorithmic defender) who has control over the rewards of each node. Here, we consider a variation of the MAB in which each node i has bounded support interval $\{-c_i^a, v_i - c_i^a\}$. This allows the MAB agent to make more informed decisions in earlier

rounds. This maps naturally to an attacker who has probed the network prior to making an attack, and it relates to recent approaches to study learning and decision making under contextual MAB, where information about rewards is provided.

Learning in Multi-Armed Bandits

In a MAB, individuals learn by repeatedly choosing among multiple options that have varying probabilities of different rewards that are observed through immediate feedback after a choice. In theories of decisions from experience, two-arm bandit problems are a classical research paradigm used for modeling human decisions and learning from experience (e.g., (Gonzalez & Dutt, 2011)).

Experiments of human behavior have demonstrated that humans are able to learn in MABs by gradually transitioning from exploration of the available alternatives to exploitation of the most rewarding options while learning from feedback and experience (Gonzalez & Dutt, 2016; Mehlhorn et al., 2015). Sripa et al. notably ran an experiment with 451 human participants playing the MAB (Sripa et al., 2009), and applied a Bayesian learning model to explain the human data. Zhang et al. extended this work by improving the participant behavioral prediction with a Knowledge Gradient model (Zhang & Angela, 2013). Our current work differs from these works in that we consider differences in reward distributions. Specifically, the previously mentioned authors address human performance in stochastic settings. In this work, we consider humans in static, stochastic, and adversarial MABs settings and analyze the effects of each environment. Furthermore, we provide context to the human decision makers by advertising the potential gains and losses of each arm of the MAB.

Recent research has shown that humans are able to learn well in contextual MABs, and various algorithms have been used to replicate this human behavior, including Thompson sampling (Agrawal & Goyal, 2012; Speekenbrink & Konstantinidis, 2015). In contrast to these models often used in MAB tasks, cognitive models of human behavior represent the cognitive mechanisms (e.g, memory, learning, forgetting) which are essential elements for human learning (Gonzalez, Lerch, & Lebiere, 2003). We offer a unique paradigm to test cognitive models of human learning and decision making and pair them against other representations of behavior in MAB tasks, playing against defense algorithms of various complexities.

Honeypot Cybersecurity Game

In the Honeypot Cybersecurity Game (HCG) a defender places decoys to protect network resources (nodes) and the attacker aims to capture those resources. A screenshot of the user interface shows a network with 5 nodes (Figure 1). Each node in the network has the following values: v_i is the value of node i, c_i^a is the cost to attack node i, and c_i^d is the cost to defend node i. The reward $v_i - c_i^a$ for attacking a non-honeypot appears as a positive number on top of each node. The cost for attacking a honeypot $-c_i^a$ appears as a negative number at the bottom of the node.

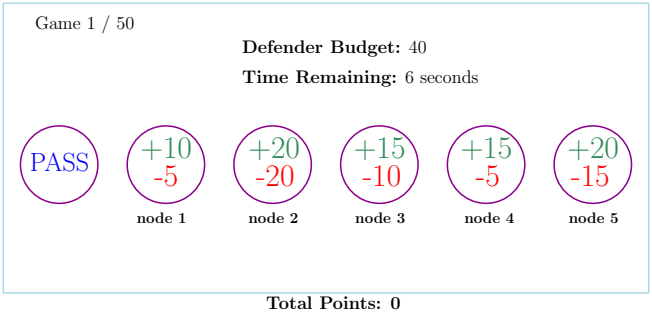


Figure 1: User interface for the HCG.

Table 1 shows the specific values used in the HCG for our experiments. We designed the node values to fit common risk-reward archetypes (e.g., low-risk/low-reward, high-risk/high-reward, low-risk/high-reward). The explicit values shown in each node give an attacker the possibility of making informed decisions that will be combined with experiential decisions as in (Lejarraga, Dutt, & Gonzalez, 2012).

	pass	node 1	node 2	node 3	node 4	node 5
v_i	0	15	40	35	20	35
c_i^a	0	5	20	10	5	15
c_i^d	0	10	20	15	15	20

Table 1: Node parameters for online human experiment.

At the beginning of each round, the defender spends her budget D to turn some subset of the nodes into honeypots, such that the total cost is $\leq D$. Once the defender deploys honeypots, the attacker selects a node to attack or passes. If the attacker's chosen node i is not a honeypot, the attacker receives the reward $v_i - c_i^a$, and the defender receives a reward of 0. If the attacker's chosen node i was a honeypot, the attacker receives the negative reward $-c_i^a$, and the defender receives the positive reward v_i[1]. At the end of a round with n trials, the game resets and a new round begins. The attacker and defender are only informed of the rewards they receive after each action, and do not directly observe the other player's choices (known as incomplete or semi-bandit feedback).

Defender Algorithms

We consider 3 different defender algorithms to investigate their impact on human adversarial decision making and learning. We expect these to create varying levels of difficulty for the human attackers to learn the defense policy.

The *Static Pure Defender* algorithm employs a "set and forget," defense that implements an unchanging, greedy strategy that spends the budget to protect the highest valued nodes. For the scenario in Figure 1, the defender always sets nodes 2 and 5 as honeypots, leading to nodes 3 and 4 being the optimal ones to attack. Against this defender, the attacker can gain a maximum of 750 total points in this specific scenario by always attacking node 3 or 4 for all 50 rounds.

[1]We assume $v_i \geq c_i^a$ and $\sum_{i \in N} c_i^d > D$.

The *Static Equilibrium Defender* plays according to a fixed probability distribution over the possible combinations of nodes to be honeypots. A new combination is selected randomly each round according to the distribution shown in Table 2. This is a game-theoretic Mixed Strategy Nash Equilibrium that optimizes the defender's expected utility assuming a single, non-repeated interaction against a fully rational attacker. The optimal strategy for the attacker against this strategy is to attack node 4, with an expected total value of ≈ 447 points for the attacker.

defended nodes	{1,3,4}	{2,3}	{2,5}	{3,5}
probability	≈ 0.303	≈ 0.095	≈ 0.557	≈ 0.0448

Table 2: Static Equilibrium Defender probabilistic strategy.

Algorithm 1 Learning with Linear Rewards (LLR)

If $\max_a |\mathcal{A}_a|$ is known, let $L = \max_a |\mathcal{A}_a|$; else, $L = N$
for $t = 1$ to N **do**
  Play any action a such that $t \in \mathcal{A}_a$
  Update $\left(\hat{\theta}_i\right)_{1\times N}$, $\left(m_i\right)_{1\times N}$ accordingly
end for
for $t = N+1$ to ∞ **do**
  Play an action a which solves the maximization:

$$a = \arg\max_{a\in\mathcal{F}} \sum_{i\in A_a} a_i\Big(\hat{\theta}_i + \sqrt{\frac{(L+1)\ln n}{m_i}}\Big), \qquad (1)$$

  Update $\left(\hat{\theta}_i\right)_{1\times N}$, $\left(m_i\right)_{1\times N}$ accordingly
end for

The *Adaptive Learning with Linear Rewards Defender (LLR)* (Gai, Krishnamachari, & Jain, 2012) plays an adaptive, learning defense strategy that tries to maximize reward by balancing exploration and exploitation using an approach designed for MAB learning. $\mathcal{A}_a$ in LLR is the set of all individual actions (nodes to defend). In the scenario from Figure 1, $\mathcal{A}_a$ is the set containing all 5 nodes. LLR uses a learning constant L, which we set to $L = 3$ since this is the maximum number of nodes we can play in a defense. LLR has an initialization phase for the first $N = 5$ rounds where it guarantees playing each node at least once. $\left(\hat{\theta}_i\right)_{1\times N}$ is that vector containing the mean observed reward $\hat{\theta}_i$ for all nodes i. $\left(m_i\right)_{1\times N}$ is the vector containing m_i, or number of times node i has been played. The vectors are updated after each round.

After the initialization phase, LLR solves the maximization problem in equation 1 and deterministically selects the subset of nodes that maximizes the equation each round until the end of the game. The algorithm tries to balance between nodes with high observed means (i.e., have captured the attacker often in the past) and exploring less frequently played nodes (which the attacker may move to in order to avoid capture). While LLR has no concept of an opponent, it indirectly adapts to the attacker based on the observations of previous rewards that depend on the attacker's strategy.

In this scenario, it is difficult for the attacker to fully exploit the strategy of the defender due to incomplete information. When facing a static defender in a static environment, the optimal node(s) will remain the same, but when facing LLR or another adaptive defender the node(s) providing the highest expected value may change from round to round.

Experimental Design

We recruited 304 human participants on Amazon's Mechanical Turk (AMT) where 130 reported female and 172 reported male with 2 participants reporting as other. All participants were above the age of 18, and the median age was 32. Participants interacted with one of the 3 defense algorithms for 50 rounds. 101 participants played against the Static Pure Defender; 100 played against the Static Equilibrium Defender; and 103 played against the LLR defender. Participants took roughly 10 minutes from start to finish. They were paid US $1.00 for completing the experiment and were given a bonus payment proportional to their performance in the 50 round game, ranging from US $0 to an extra US $3.25.

This task did not require cybersecurity knowledge and participants were given detailed instructions and definitions of the concepts needed to perform the task (e.g., honeypot). Participants were told that the defender has a budget $D = 40$ that limits the number of honeypot configurations (i.e., combinations of defended nodes). In each round, the participant attacks a node and receives either a positive reward $v_i - c_i^a$ or a negative reward $-c_i^a$ depending on the defender's action. The setup in Figure 1 was the same for every participant.

We analyzed 4 measures associated with participants' performance, and we compared predictive algorithms using the same measures. **Switching** is a common measure of exploration used in human decision-making and learning studies (Gonzalez & Dutt, 2016; Todd & Gigerenzer, 2000). High switching indicates high exploration and low switching indicates exploitation in the case of a static defender and static environment. **Switching with Honeypot** is a measure of switching after attacking a honeypot (i.e., receiving a negative reward). This corresponds with the "Lose-Shift" aspect of Win-Stay-Lose-Shift (WSLS) (Robbins, 1985), a common strategy studied in economics. **Switching without Honeypot** measures switching after attacking a real node (i.e., receiving a positive reward). This opposes the "Win-Stay" aspect of the WSLS (i.e., "Win-Shift"). Finally, **Optimal Play** is the fraction of decisions that have the actual highest expected value.

Behavioral Results

The results for the 4 dependent measures are shown in Figure 2. The rightmost graph in Figure 2 shows the frequency of optimal decisions over the 50 rounds. We note that participants playing against the static pure defender learn very early to play optimally and significantly improve over time, while the difference between the static equilibrium defender and adaptive LLR defenders is not clear early on. A significant advantage for LLR only emerges after at least 20 rounds.

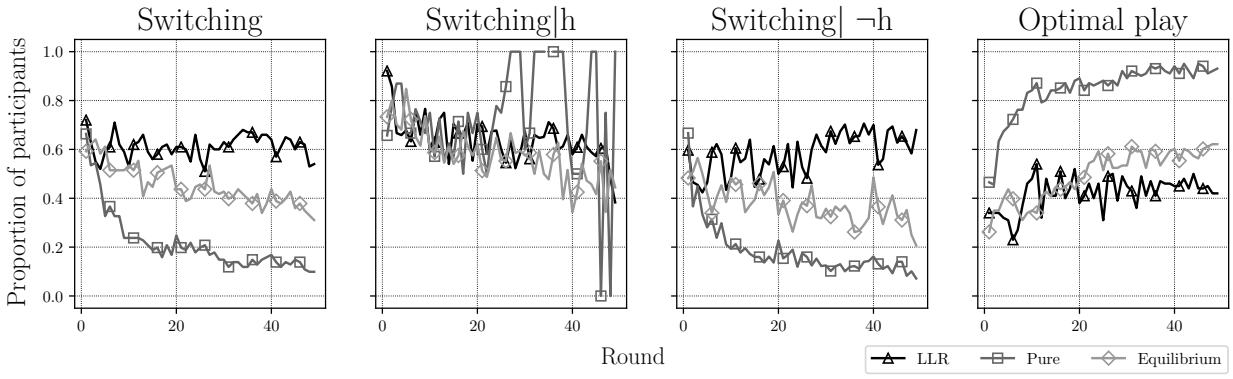


Figure 2: The proportions of participants switching nodes and playing optimally over time. The high switching after triggering a honeypot seen in round 26 from participants facing the static pure defender is a small portion of the population.

We also observe in the leftmost graph in Figure 2 that the overall proportion of switching decreases over time, particularly when participants face the static pure defender. When the participants face the adaptive LLR defender, they seem to have a high proportion of switching throughout the 50 rounds.

The middle left graph in Figure 2 describes the participants' switching behavior after triggering a honeypot. For the static pure defender, the attackers show noticeable spikes because only a few participants attacked the 20 point nodes (triggering the honeypots), upon which the players immediately switched. There are few differences between switching behavior when triggering honeypots of the participants who faced the equilibrium defender and those who faced adaptive LLR. We see a downward trend, hinting that the participants are moving from an early exploratory state to a more exploitative state. Since adaptive LLR updates it's beliefs about a node's expected payoff after playing it, if it captures an attacker that node will be more likely to be selected in the immediate future. Due to this adaptive behavior, switching when triggering a honeypot against adaptive LLR will be more beneficial than against the static equilibrium defender. When facing the static equilibrium defender in this scenario the attacker should always attack node 4, regardless of triggering a honeypot or not.

The middle right graph in Figure 2 shows distinct differences when the attackers did not trigger a honeypot (i.e., received a positive reward). Concerning the static pure defender and static equilibrium defender, decreases in switching demonstrate a move towards a more exploitative strategy and understanding of the static defense. Compare this with participants who faced the adaptive LLR defender where the switching remains high in comparison to the defenders. In general, adaptive LLR tries to react to the observed rewards and slowly moves from exploration to exploitation over time. High switching and remaining mobile is a good strategy against adaptive LLR. However, when we compare the participants' switching behavior with their performance versus adaptive LLR, it appears the participants were largely unable to learn the LLR strategy.

Overall, the pure defender predictably performed the worst (best for the human attackers), yielding an average score of 611.93 points. The equilibrium defender performed significantly better, yielding an average of 247.81 points. Finally, LLR was the most resilient defender against the human attackers with an average of 172.6 points yielded to the participants. Table 3 shows the aggregate statistics of the human attacker performance in terms of end-game attacker points.

	average	std. dev.	median	min	max
Pure	611.93	168.88	675	-375	750
Equ.	247.81	149.60	290	-185	570
LLR	172.6	123.02	160	-85	640

Table 3: Aggregate data of participants' end-game attacker points.

Adversarial Models

We evaluated 4 behavioral models and one cognitive model (IBL) (Gonzalez et al., 2003) to emulate participants' performance in the experiment. These models can give insights into the underlying mechanisms that influence decision making and support the development of better defense algorithms that hinder human attacker learning in cybersecurity settings. The models selected below are representatives of behavioral predictors that have been known to capture human performance in numerous MAB settings (Sripa et al., 2009; Zhang & Angela, 2013; Agrawal & Goyal, 2012).

Win-Stay-Lose-Shift: WSLS plays uniform randomly on the first round. If WSLS receives a positive reward, it attacks the same node again in the next round. Otherwise, it attacks another node uniform randomly. The "pass" action does not count as a positive reward.

ε-Greedy: This model addresses the exploration-exploitation dilemma directly with the parameter $\varepsilon \in \{0,1\}$. With probability ε, ε-Greedy attacks uniform randomly (exploration) and with probability $(1-\varepsilon)$, attacks the node with the highest observed average reward (exploitation).

ε-Greedy Decreasing: ε-Greedy Decreasing dynamically changes the parameter ε in order to prefer exploitation to-

	LLR				Pure				Equilibrium			
	Sw	Sw| h	Sw|¬h	OP	Sw	Sw| h	Sw|¬h	OP	Sw	Sw| h	Sw|¬h	OP
ε-G 0.2	0.317	0.258	0.353	0.153	0.146	0.325	0.121	0.163	0.189	0.245	0.164	0.138
ε-GD	0.236	0.173	0.309	0.205	0.39	0.259	0.392	0.239	0.211	0.179	0.25	0.159
WSLS	0.221	0.364	0.486	0.190	0.211	**0.079**	0.191	0.254	**0.104**	0.434	0.26	0.285
TS	**0.091**	0.121	0.140	0.137	0.210	0.318	0.21	0.076	0.124	**0.156**	**0.123**	**0.070**
IBL	0.109	**0.118**	**0.139**	**0.127**	**0.084**	0.347	**0.094**	**0.057**	0.136	0.163	0.164	0.152

Table 4: The distances of the predictions of individual predictors or IBL models from human data, calculated using RMSE metric. The measures we use are switching (Sw), switching after triggering a honeypot (Sw|h), switching after not triggering a honeypot (Sw|¬h) and optimal play (OP). Bold font indicates the lowest value in each column.

wards the end of the interaction. The predictor starts with $\varepsilon = 1$ and decreases it linearly towards $\varepsilon = 0$ at the end of the interaction, given a known finite horizon.

Thompson Sampling (TS): We follow the description of the TS algorithm as detailed by Agrawal and Goyal for Bernoulli Bandits (2012). We extend this version of the TS algorithm for the Bernoulli MAB by incorporating a support function $W_i(\theta_i)$ instead of selecting the action i with the maximum sample θ_i as described by Agrawal and Goyal. For this setting, we use a support function $W_i(\theta_i) = v_i \cdot \theta_i - c_i^a$ where $\theta_i \sim Beta(S_i + 1, F_i + 1)$ samples from a Beta distribution, thus the algorithm favors successes (S_i) over failures (F_i).

Instance-Based Learning: An IBL model (Gonzalez & Dutt, 2011) describes a learning attacker with an ability to recall and identify similar "instances" of past decisions using memory. An IBL instance represents a decision made in a specific situation, and the outcome feedback. The feedback here is the net payoff calculated as a difference between a successful and a failed attack, i.e., $v_i - 2c_i^a$. The IBL decision process has three main parameters: (1) decay, d, which specifies how past experiences are considered in current decisions based on time; (2) noise parameter σ, capturing random variability between experiences; and (3) the similarity, S, capturing the influence of the past on the present based on the similarity of the situations.

In the HCG game, an attacker can observe two possible outcomes of an attack on node i: a positive reward ($v_i - c_i^a$) when she attacks a real resource (success s_i) or a negative reward ($-c_i^a$) if the target is a honeypot (failure f_i). We denote an instance in memory representing a combination of situation, decision and outcome that was experienced in the past as $o(t') \in \bigcup_{i \in N} \{s_i, f_i\}$. In round t, an attacker targets a node i_t^* which maximizes a blended value (BV) as follows:

$$i_t^* \leftarrow \arg\max_{i \in N} BV_t(i) \tag{2}$$

$$BV_t(i) = (v_i - c_i^a) \frac{e^{A_t(s_i)}}{e^{A_t(s_i)} + e^{A_t(f_i)}} - c_i^a \frac{e^{A_t(f_i)}}{e^{A_t(s_i)} + e^{A_t(s_i)}} \tag{3}$$

$$A_t(o_i) = \ln \sum_{t' \in \{1,\ldots,t-1\}: o(t') = o_i} (t - t')^{-d} - \\ - S \sum_{i' \in N} (sim(i, i')) - \sigma \ln \frac{1-\gamma}{\gamma}, \tag{4}$$

where $\gamma \in (0, 1]$ is uniformly randomly sampled and *sim* is a similarity function. We used a linear similarity function that normalizes the net payoff from a decision based on the maximal payoff of 20 and is calculated as $sim(i, i') = 1 - |(v_i - 2c_i^a) - (v_{i'} - 2c_{i'}^a)|/20$.

We fit a separate IBL attacker model to human data when playing against each of the algorithmic defenders. We calibrated parameters values using exhaustive search over a wide range of values for each parameter with 350 repetitions for each combination. We used a multiobjective optimization minimizing average RMSE (see Equation 5) of all measures. The resulting three sets of parameters were: $(\sigma = 0.2, d = 0.1, S = 0.6)$ for the LLR defender, $(\sigma = 0.35, d = 1.2, S = 0.4)$ for the Pure defender and $(\sigma = 1.4, d = 0.5, S = 0.5)$ for the Equilibrium defender.

Simulation Results

To analyze the predictors' effectiveness in emulating human behavior we did a simulation with identical settings to the human experiment. Each predictor played against the 3 defenders in the same scenario 100 times. We consider the same performance measures as before. How well a predictor approximates human behavior is determined by a distance of a prediction $\{p\}_{t=1}^T$ from human data $\{hd\}_{t=1}^T$, calculated using the RMSE metric below where where m is a performance measure and T is a number of rounds.

$$RMSE_m(p, hd) = \sqrt{\frac{\sum_{t=1}^{50} (m(p_t) - m(hd_t))^2}{T}} \tag{5}$$

In Table 4, IBL accounted for the "most human" behavior on most of the measures when playing against the Pure and LLR defenders. In contrast, TS plays most closely to human performance when playing against the Equilibrium defender. This may be because the static equilibrium defender most closely reflects the standard stochastic MAB setting that TS was designed for. ε-Greedy, ε-Greedy Decreasing and WSLS perform poorly in general as predictors of human behavior.

However, these observations may only paint part of the picture. The actual overall point performance of human participants versus LLR is much lower than the 4 behavioral predictors as shown in Table 5. Nearly all 4 of the behavioral predictors double the median score of the human participants when facing the LLR defender. In contrast, the IBL model plays the most closely to human performance versus LLR. The IBL model comes rather close to the human data in relation to the

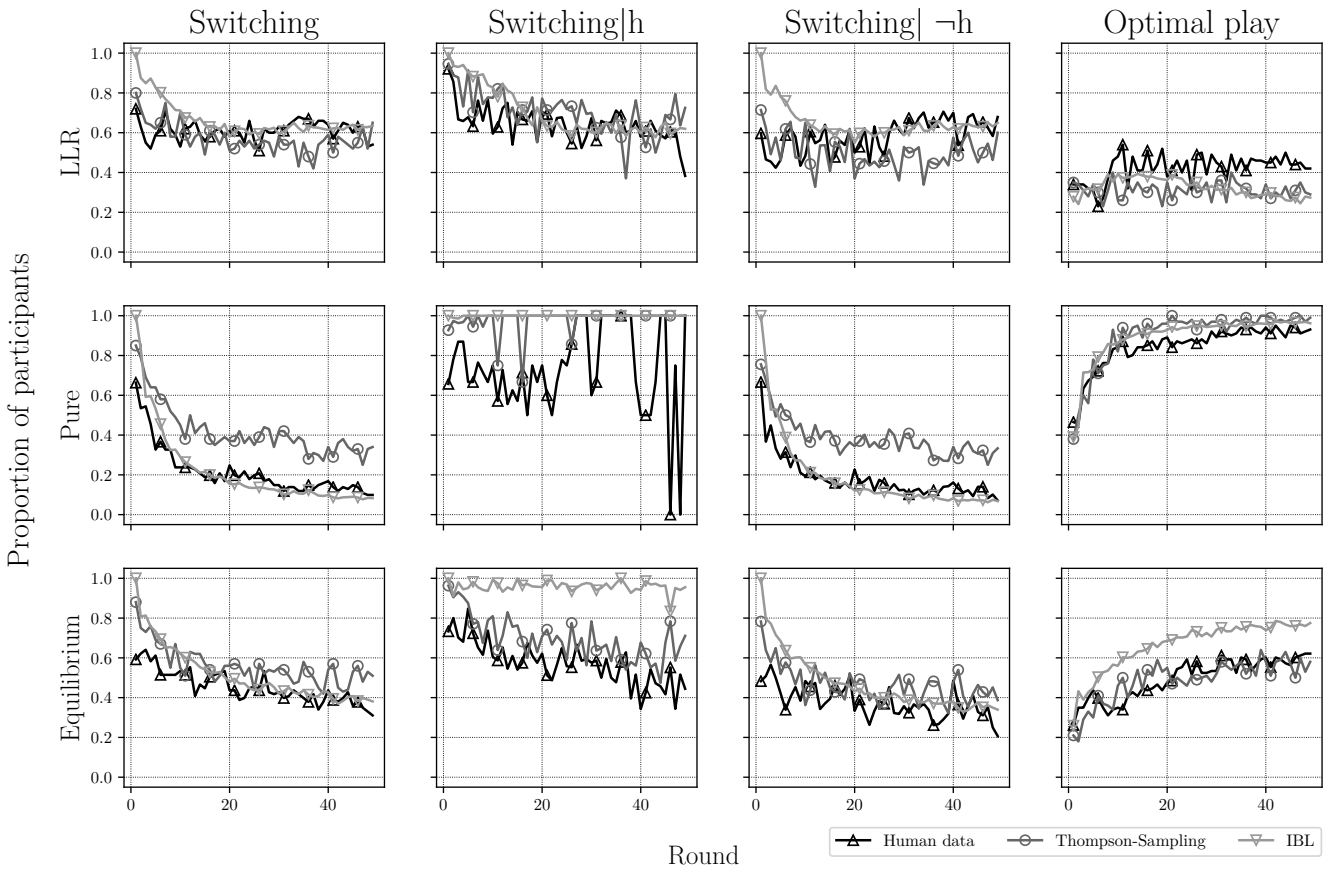


Figure 3: Comparison of the strategy predictions of TS and IBL with human data.

average and median scores. When considering all this information, it appears that the adaptive LLR defender exploited the human participants' learning mechanisms as well as IBL predicts. We can also see that humans may adopt different strategies depending on an opponent's strategy. Thus, when choosing a modeling approach there is a need to carefully select the granularity level at which predictions are needed: aggregate or individualized behavior. The IBL model can produce predictions at both levels.

	μ	σ	median	min	max
Human	172.6	123.02	160	-85	640
ε-G 0.2	303.9	140.3559	320	-75	640
ε-GD	265.1	99.55705	275	-115	480
TS	332	109.6275	330	90	585
WSLS	292.4	114.2686	287.5	35	590
IBL	198.9	193.44	220	-335	685

Table 5: Performance of predictors against the LLR defender in attacker points.

Conclusion

We study how humans learn in a novel version of an adversarial, contextual multi-armed bandit scenario motivated by a real-world cybersecurity scenario where defenders use deceptive decoys and attackers must learn to avoid them. We evaluated three different types of defensive strategies and showed that an adaptive defensive strategy was clearly the strongest against human players, and the hardest for them to learn. We also made novel comparisons between predictive models for emulating how humans learn in this type of adversarial setting, comparing leading models from both the MAB literature and cognitive science. We find that the best models (Thompson Sampling and IBL) are able to predict human behavior quite effectively, but that human attackers use different strategies depending on the adversary they are up against, and the best predictor may depend on this context. There are many interesting opportunities to improve both types of models especially in making personalized predictions for individuals and specialized context. However, the results so far have immediate practical implications for how we can design better strategies for deploying decoy systems to enhance cybersecurity. In particular, these systems must be adaptive to prevent attackers from easily learning the defensive strategy. The predictive models of attacker learning we have developed will also allow us to develop defenses that actively mitigate the ability of attackers to learn the defensive strategy.

Acknowledgements

This research was sponsored by the Combat Capabilities Development Command Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-13-2-0045 (ARL Cyber Security CRA). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Combat Capabilities Development Command Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes not withstanding any copyright notation here on. The authors also acknowledge the support of the OP VVV MEYS funded project CZ.02.1.01/0.0/0.0/16_019/000 0765 Research Center for Informatics. The authors thank Orsolya Kovacs from the Dynamic Decision Making Laboratory at Carnegie Mellon University for her help with data collection.

References

Agrawal, S., & Goyal, N. (2012). Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on learning theory* (pp. 39–1).

Gai, Y., Krishnamachari, B., & Jain, R. (2012). Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations. *IEEE/ACM Transactions on Networking (TON)*, *20*(5), 1466–1478.

Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, *349*(6245), 273–278.

Gonzalez, C., & Dutt, V. (2011). Instance-based learning: Integrating sampling and repeated decisions from experience. *Psychological review*, *118*(4), 523.

Gonzalez, C., & Dutt, V. (2016). Exploration and exploitation during information search and experimential choice. *Journal of Dynamic Decision Making*, *2*(1).

Gonzalez, C., Lerch, J. F., & Lebiere, C. (2003). Instance-based learning in dynamic decision making. *Cognitive Science*, *27*(4), 591–635.

Lejarraga, T., Dutt, V., & Gonzalez, C. (2012). Instance-based learning: A general model of repeated binary choice. *Journal of Behavioral Decision Making*, *25*(2), 143–153.

Mehlhorn, K., Newell, B. R., Todd, P. M., Lee, M. D., Morgan, K., Braithwaite, V. A., ... Gonzalez, C. (2015). Unpacking the exploration–exploitation tradeoff: A synthesis of human and animal literatures. *Decision*, *2*(3), 191.

Robbins, H. (1985). Some aspects of the sequential design of experiments. In *Herbert robbins selected papers* (pp. 169–177). Springer.

Speekenbrink, M., & Konstantinidis, E. (2015). Uncertainty and exploration in a restless bandit problem. *Topics in cognitive science*, *7*(2), 351–367.

Spitzner, L. (2003). *Honeypots: tracking hackers* (Vol. 1). Addison-Wesley Reading.

Sripa, B., Mairiang, E., Thinkhamrop, B., Laha, T., Kaewkes, S., Sithithaworn, P., ... Bethony, J. M. (2009). Advanced periductal fibrosis from infection with the carcinogenic human liver fluke opisthorchis viverrini correlates with elevated levels of interleukin-6. *Hepatology*, *50*(4), 1273–1281.

Steyvers, M., Lee, M. D., & Wagenmakers, E.-J. (2009). A bayesian analysis of human decision-making on bandit problems. *Journal of Mathematical Psychology*, *53*(3), 168–179.

Todd, P. M., & Gigerenzer, G. (2000). Précis of simple heuristics that make us smart. *Behavioral and brain sciences*, *23*(5), 727–741.

Zhang, S., & Angela, J. Y. (2013). Forgetful bayes and myopic planning: Human learning and decision-making in a bandit setting. In *Advances in neural information processing systems* (pp. 2607–2615).

Character-based Surprisal as a Model of Reading Difficulty in the Presence of Errors

Michael Hahn (mhahn2@stanford.edu)
Department of Linguistics, Stanford University
Margaret Jacks Hall, Stanford, CA 94305, USA

Frank Keller (keller@inf.ed.ac.uk)
School of Informatics, University of Edinburgh
10 Crichton Street, Edinburgh EH8 9AB, UK

Yonatan Bisk (ybisk@cs.washington.edu)
Paul G. Allen School of Computer Science & Eng.
University of Washington
185 E Stevens Way NE, Seattle, WA 98195, USA

Yonatan Belinkov (belinkov@seas.harvard.edu)
John A. Paulson School of Eng. & Applied Sciences,
Harvard University, and Computer Science and Artificial
Intelligence Laboratory, MIT, Cambridge, MA, USA

Abstract

Intuitively, human readers cope easily with errors in text; typos, misspelling, word substitutions, etc. do not unduly disrupt natural reading. Previous work indicates that letter transpositions result in increased reading times, but it is unclear if this effect generalizes to more natural errors. In this paper, we report an eye-tracking study that compares two error types (letter transpositions and naturally occurring misspelling) and two error rates (10% or 50% of all words contain errors). We find that human readers show unimpaired comprehension in spite of these errors, but error words cause more reading difficulty than correct words. Also, transpositions are more difficult than misspellings, and a high error rate increases difficulty for all words, including correct ones. We then present a computational model that uses character-based (rather than traditional word-based) surprisal to account for these results. The model explains that transpositions are harder than misspellings because they contain unexpected letter combinations. It also explains the error rate effect: expectations about upcoming words are harder to compute when the context is degraded, leading to increased surprisal.

Keywords: human reading, eye-tracking, errors, computational modeling, surprisal, neural networks.

Introduction

Human reading is both effortless and fast, with typical studies reporting reading rates around 250 words per minute (Rayner, White, Johnson, & Liversedge, 2006). Human reading is also adaptive: readers vary their strategy depending on the task they want to achieve, with experiments showing clear differences between reading for comprehension, proofreading, or skimming (Kaakinen & Hyönä, 2010; Schotter, Bicknell, Howard, Levy, & Rayner, 2014; Hahn & Keller, 2018).

Another remarkable aspect of human reading is its robustness. A lot of the texts we read are carefully edited and contain few errors, e.g., articles in newspapers and magazines, or books. However, readers also frequently encounter texts that contain errors, e.g., in hand-written notes, emails, text messages, and social media posts. Intuitively, such errors are easy to cope with and impede understanding only in a minor way. In fact, errors often go unnoticed during normal reading, which is presumably why proofreading is difficult.

The aim of this paper is to experimentally investigate reading in the face of errors, and to propose a simple model that can account for our experimental results. Specifically, we focus on errors that change the form of a word, i.e., that alter a word's character sequence. This includes letter transposition (e.g., *innocetn* instead of *innocent*) and misspellings (e.g., *inocent*). Importantly, we will not consider whole-word substitutions, nor will we deal with morphological, syntactic, or semantic errors.

We know from the experimental literature that letter transpositions cause difficulty in reading (Rayner et al., 2006; Johnson, Perea, & Rayner, 2007; White, Johnson, Liversedge, & Rayner, 2008). However, transpositions are artificial errors (basically they are an artifact of typing), and are comparatively rare.[1] It is not surprising that such errors slow down reading. This contrasts with misspellings, i.e., errors that writers make because they are unsure about the orthography of a word. These are natural errors that should be easier to read, because they occur more frequently and are linguistically similar to real words (*inocent* conforms to the phonotactics of English, while *innocetn* does not). This is our first prediction, which we will test in an eye-tracking experiment that compares the reading of texts with transpositions and misspellings.

Readers' prior exposure to misspellings might explain why reading is mostly effortless, even in the presence of errors. The fact remains, however, that all types of errors are relatively rare in everyday texts. All previous research has studied isolated sentences that contain a single erroneous word. This is a situation with which the human language processor can presumably cope easily. However, what happens when humans read a whole text which contains a large proportion of errors? It could be that normal reading becomes very difficult if, say, half of all words are erroneous. In fact, this is what we would expect in expectation-based theories of language processing, such as surprisal (Levy, 2008): the processor constantly uses the current context to compute expectations for the next word, and difficulty ensues if these expectations turn out to be incorrect. However, if the context is degraded by a large number of errors, then it is harder to compute expectations (and they become less reliable), and reading should slow down. Crucially, we expect to see this effect on all words, not

[1] For example, in the error corpus we use (Geertzen, Alexopoulou, & Korhonen, 2014) only 11% of the errors are letter swaps or repetitions, see Table 1.

just on those words that contain errors. This is the second prediction that we will test in our eye-tracking experiment by comparing texts with high and low error rates.

In the second part of this paper, we present a surprisal model that can account for the patterns of difficulty observed in our experiment on reading texts with errors. We start by showing that standard word-based surprisal does not make the right predictions, as it essentially treats words with errors as out of vocabulary items. We therefore propose to estimate surprisal with a character-based language model. We show that this model successfully predicts human reading times for texts with errors and accounts for both the effect of error type and the effect of error rate that we observed in our reading experiment.

Eye-tracking Experiment

The aim of this experiment was to determine how human reading is affected by errors in the input. As explained in the introduction, we expected different error types to affect reading differentially, as error types can differ in familiarity. In addition, we predicted the overall number of errors in a text to have an effect on reading behavior, because a high error rate degrades word context, which is crucial for computing expectations about upcoming material.

The experiment used a two-by-two factorial design, crossing error type (transpositions vs. misspellings) with error rate (10% of all words contain errors vs. 50%). Both of these variables were administered as between-text factors, i.e., we created four versions for each text, one with 10% transpositions, one with 10% misspellings, one with 50% transpositions, and one with 50% misspellings.

The two experimental factors were administered within participants, i.e., all participants read all our texts, each of them presented in one of the four versions. Versions were distributed across participants using a Latin square design, so as to ensure that every version was seen by the same number of participants.

Methods

Participants Sixteen participants took part in the experiment after giving informed consent. They were paid £10 for their participation, had normal or corrected-to-normal vision, and were self-reported native speakers of English.

Materials We used the materials of Hahn and Keller (2018), but introduced errors into the texts. These materials contain twenty newspaper texts from the DeepMind question answering corpus (Hermann et al., 2015). Ten texts were taken from the CNN section of the corpus and the other ten texts from the Daily Mail section. Texts were comparable in length (between 149 and 805 words, mean 323) and represent a balanced selection of topics. Two additional texts were used as practice items.

Each text comes with a question and a correct answer. The questions are formulated as sentences with a blank to be completed with a named entity so that a statement implied by the text is obtained. Three incorrect answers (distractors) are included for each question; these are also named entities, chosen so that they closely match the correct answer (e.g., if the correct answer is *Minnesota*, then the distractors are also US states).[2]

phonetics	deletion	swap/repeat	keyboard	insertion	other
36.2	16.7	11.0	10.5	8.3	17.3

Table 1: Percentages of different types of misspellings in the natural error condition.

We introduced errors into the materials of Hahn and Keller (2018) following the method suggested by Belinkov and Bisk (2018). These errors are automatically generated and are either transpositions (i.e., two adjacent letters are swapped) or natural errors that replicate actual misspellings. For the latter, we used a corpus of human edits (Geertzen et al., 2014), and introduced errors in our experimental materials by replacing correct words with known misspellings from our edit corpus. The percentages of different types of misspellings are listed in Table 1. By generating texts with errors automatically we were able to ensure that both error conditions (transpositions or misspellings) contain the same percentage of erroneous words for the two error rates (10% or 50% erroneous words).

Procedure Participants received written instructions, which mentioned that they would be reading texts with errors. They first went through two practice trials whose data was discarded. Then, each participant read and responded to all 20 items (texts with questions and answer choices); the items were presented in a new random order for each participant. The order of the answer choices was also randomized.

In each trial, the text was displayed over one or more pages (max 5, mean 2.1 pages), where each page contained up to eleven lines with about 80 characters per line. To get to the next page, and at the end of the text, participants again had to press a button. After the last page, the question was displayed, together with the four answer choices, on a separate page. Participants had to press one of four buttons to select an answer.

Eye-movements were recorded using an Eyelink 2000 tracker (SR Research, Ottawa). The tracker recorded the dominant eye of the participant (as established by an eye-dominance test) with a sampling rate of 2000 Hz. Before the experiment started, the tracker was calibrated using a nine-point calibration procedure; at the start of each trial, a central fixation point was presented. Throughout the experiment, the experimenter monitored the accuracy of the recording and carried out additional calibrations as necessary.

[2] We used the no questions preview condition of Hahn and Keller (2018), i.e., the questions were shown only after participants had read the whole text. The original paper also had a question preview condition, in which participants were shown the questions before they read the text.

	Hahn & Keller	This experiment	
		No error	Error
First fixation	221.3	211.8	225.1
First pass	260.7	242.5	265.2
Total time	338.0	306.9	342.1
Fixation rate	0.50	0.45	0.48
Accuracy	70%	72%	

Table 2: Left: per-word reading times, fixation rates, and question accuracies in the experiment of Hahn and Keller (2018), right: same measures for our experiments (same texts, but some of the words contain errors).

Data Analysis For data analysis, each word in the text was defined as a region of interest. Punctuation was included in the region of the word it followed or preceded without intervening whitespace. If a word was preceded by a whitespace, then that space was included in the region for that word. We report data for the following eye-movement measures in the critical regions: *First fixation duration* is the duration of the first fixation in a region, provided that there was no earlier fixation on material beyond the region. *First pass time* (often called gaze duration for single-word regions) consists of the sum of fixation durations beginning with this first fixation in the region until the first saccade out of the region, either to the left or to the right. *Total time* consists of the sum of the durations of all fixation in the region, regardless of when these fixations occur. *Fixation rate* measures the proportion of trials in which the region was fixated (rather than skipped) on first-pass reading. For first fixation duration and first pass time, no trials in which the region is skipped on first-pass reading (i.e., when first fixation duration is zero) were included in the analysis. For total time, only trials with a non-zero total time were included in the analysis.

Due to space limitations, we will only present analyses of the first pass time and fixation rate data in the remainder of this paper.

Results

In Table 2, we present some basic reading measures for our experiments, and compare these to the reading experiments of Hahn and Keller (2018), which used the same texts, but did not include any errors (the data is taken from their no question preview condition, which corresponds to our experimental setup, see Footnote 2). Even for words with errors, the reading measures in our experiments are similar to the ones reported by Hahn and Keller (2018). For words without errors, we find slightly faster reading times and lower fixation rates than Hahn and Keller (2018). Also the accuracy (which can only be measured on the text level, hence we do not distinguish words with and without errors) is essentially unchanged. This provides good evidence for the claim that human readers cope well with errors in text: they take longer to read words with errors and fixate them more compared to words without errors, but this this is a comparatively small effect. Overall, reading times, fixation rates, and question accuracy are very similar to those found in texts without any errors (such as the ones used by Hahn & Keller, 2018).[3]

In the following, we analyze two reading measures in more detail: first pass time and fixation rate. We analyzed per-word reading measures using mixed-effects models, considering the following predictors:

1. ERRORTYPE: Does the text contain mispellings (-0.5) or transpositions ($+0.5$)?
2. ERRORRATE: Does the text contain 10% (-0.5) or 50% ($+0.5$) erroneous words overall?
3. ERROR: Is the word correct (-0.5) or erroneous ($+0.5$)?
4. WORDLENGTH: Length of the word in characters.
5. LASTFIX: Was the preceding word fixated ($+0.5$) or not (-0.5)?

All predictors were centered. Word length was scaled to unit variance. We selected binary interactions using forward model selection with a χ^2 test, running the R package `lme4` (Bates, Mächler, Bolker, & Walker, 2015) with a maximally convergent random effects structure. We then re-fitted the best model with a full random effects structure as a Bayesian generalized multivariate multilevel model using the R package `brms`; this method is slower but allows fitting large random effects structures even when traditional methods do not converge. Resulting Bayesian models are shown in Table 3. We used the `brms` default priors (Bürkner, 2017), with four chains with 1000 samples each (and 1000 warmup iterations). The $\hat{R}$ values (≤ 1.01) indicated that the models had converged.[4]

The main effects of WORDLENGTH replicate the well-known positive correlation between word length and reading time (see Demberg & Keller, 2008, and many others). We also find main effects of ERROR, indicating that erroneous words are read more slowly and are more likely to be fixated. The main effects of ERRORRATE show that higher text error rates lead to longer reading times and higher fixation rates for all words (whether they are correct or erroneous). Additionally, we find a main effect of ERRTYPE in fixation rate, showing that transposition errors lead to higher fixation rates. This is consistent with our hypothesis that misspellings are easier to process than transpositions, as they are real errors that participants have been exposed in their reading experience.

Figure 1 graphs mean first pass times and fixation rates by error type and error rate. The most important effect is that

[3] Note that participants are not performing at ceiling in question answering; our pattern of results therefore cannot be explained by asserting that the questions were too easy.

[4] An analogous analysis for log-transformed first-pass times led to the same pattern of significant effects and their directions.

	First Pass	Fixation Rate
(Intercept)	248.41 (6.34)***	−0.16 (0.12)
ErrType	1.41 (1.32)	0.08 (0.02)***
ErrRate	7.20 (1.60)***	0.16 (0.02)***
Error	23.77 (4.12)***	0.21 (0.07)***
WLength	22.18 (2.02)***	0.83 (0.04)***
LastFix	3.10 (4.18)	0.22 (0.18)
ErrRate × LastFix	6.71 (2.77)*	0.16 (0.04)***
Error × LastFix	—	0.26 (0.10)**
WLength × LastFix	—	0.74 (0.10)***

$Pr(\beta < 0)$: *** < 0.001, ** < 0.01, * < 0.05

Table 3: Bayesian generalized multivariate multilevel models for reading measures with maximal random-effects structure. Each cell gives the coefficient, its standard deviation, and the estimated posterior probability that the coefficient has the opposite sign.

error words take longer to read and are fixated more than non-error words. The effect of error rate is also clearly visible: the 50% error condition causes longer reading times and more fixations than the 10% one, even for non-error words. We also observe a small effect of error type.

Turning now to the interactions, we found that ErrorRate and LastFix interact in both reading measures, which indicates that reading times and fixation rates increase in the high-error condition if the previous word has been fixated.

Only in fixation rate, there was also an interaction of Error and LastFix, indicating that fixation rate goes up for error words if the preceding word was fixated, presumably because of preview of the erroneous words, which is then more likely to be fixated in order to identify the error.

For fixation rate, WordLength interacts with LastFix: longer words are more likely to be fixated if the preceding word was fixated; again, this is likely an effect of preview. While Figure 1 seems to suggest an interaction of Error and Error Type, this was not significant in the mixed model.

Discussion

We have found four main results: (1) Erroneous words show longer reading times and are more likely to be fixated. (2) Higher error rates lead to increased reading times and more fixations, even on words that are correct. (3) Transpositions lead to an increased fixation rate compared to misspellings. (4) Whether the previous word is fixated or not modulates the effect of error and error rate.

However, it is conceivable that the effects of error and error rate are actually artifacts of word length. All else being equal, longer words take longer to read and are more likely to be fixated. So if error words and non-error words in our texts differ in mean length, then that would be an alternative explanation for the effects that we found.

For transposition errors, error words by definition have the same length as their non-error versions. For misspellings,

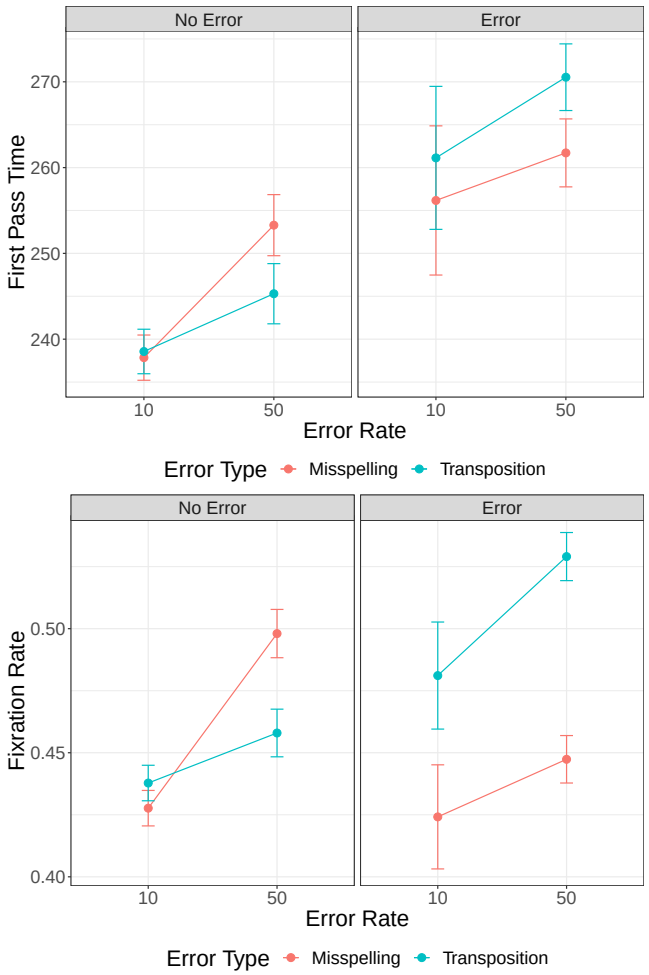


Figure 1: First pass time (top) and fixation rate (bottom) when reading texts with transposition errors or misspelling.

a mixed-effects analysis with word forms as random effects showed no significant difference in the lengths of error words and their correct versions (mean difference −0.011, SE 0.029, $t = -0.393$). Comparing the erroneous words of the two error types, we found that they differ in mean length (misspellings 5.44, transpositions 6.06 characters); however this difference was not significant in a mixed-effects analysis predicting word length of erroneous words from error types, with items as a random effect (mean difference 0.015, SE 0.010, $t = 1.449$).

Surprisal Model

Most models of human reading do not explicitly deal with reading in the face of errors. In fact, reading models that use a lexicon to look up word forms (e.g., to retrieve word frequencies) cannot deal with erroneous words without further assumptions. We can use the surprisal model of processing difficulty (Levy, 2008) to illustrate this: in its original, word-based formulation, surprisal is forced to treat all error words as out of vocabulary items; it therefore cannot distin-

guish between different types of errors or between different error rates.

Intuitively, a more fine-grained version of surprisal is required that computes expectations in terms of characters, not words. In such a setting, the word *inocent* would be more surprising than *innocent* in the same context, but not as surprising as a completely unfamiliar letter string. In other words, the surprisal of the same word with and without misspellings or letter transpositions would be similar but not the same. To achieve this, we can use character-based language models, which are standard tools in natural language processing for dealing with errors in the input (e.g., the work by Belinkov & Bisk, 2018, on errors in machine translation).

Crucially, once we have a character-based surprisal model, we can derive predictions regarding how errors should affect reading. We predict that transpositions should be more surprising than misspellings, as they involve character sequences that are unfamiliar to the model (e.g., *innocetn* contains the rare character sequence *tn*). Also, we predict that words that occur in texts with a high error rate are more difficult to read than words in texts with a low error rate: if the context of a word contains few errors, then we are able compute expectations for that word confidently (resulting in low surprisal). If the context contains lots of errors then expecations are difficult to compute and they become unreliable (resulting in high surprisal). We will now test these predictions regarding error type and error rate using a character-based version of surprisal.

Methods

We trained a character-based neural language model using LSTM cells (Hochreiter & Schmidhuber, 1997). Such models can assign probabilities to any sequence of characters, and thus are capable of computing surprisal even for words never seen in the training data, such as erroneous words. For training, we used the Daily Mail portion of the DeepMind corpus. We used a vocabulary consisting of the 70 most frequent characters, mapping others to an out-of-vocabulary token.

The hyperparameters of the language model were selected on an English corpus based on Wikipedia text.[5] We then used the resulting model to compute surprisal on the texts used in the eye-tracking experiment for each experimental condition.

The model estimates, for each element of a character sequence, the probability of seeing this character given the preceding context. We compute the surprisal of a word as the sum of the surprisals of the individual characters, as prescribed by the product rule of probability. For a word consisting of characters $x_t \dots x_{t+T}$ following a context $x_1 \dots x_{t-1}$, its surprisal is:

$$-\log P(x_t \dots x_{t+T} | x_1 \dots x_{t-1}) = \sum_{i=t}^{t+T} -\log P(x_i | x_1 \dots x_{i-1}) \quad (1)$$

In this computation, we take whitespace characters to belong to the preceding word.

To control for the impact of the random initialization of the neural network at the beginning of training, we trained seven models with identical settings but different random initializations.

The quality of character-based language models is conventionally measured in Bits Per Character (BPC), which is the average surprisal, to the base 2, of each character. On held-out data, our model achieves a mean BPC value of 1.28 (SD 0.025), competitive with BPC values achieved by state-of-the-art systems of similar datasets (e.g., Merity, Keskar, & Socher, 2018, report a BPC value of 1.23 on Wikipedia text).

In the introduction we predicted that word-based surprisal is not able to model the reading time pattern we found in our eye-tracking experiment. In order to test this prediction, we compare our character-level surprisal model to surprisal computed using a conventional word-based neural language model. Word-based models have a fixed vocabulary, consisting of the most common words in the training data; a typical vocabulary size is 10,000. Words that were not seen in the training data, and rare words, are represented by a special out-of-vocabulary (OOV) token. From a cognitive perspective, this corresponds to assuming that all unknown words (whether they contain errors or not) are treated in the same way: they are recognized as unknown, but not processed any further. We used a vocabulary size of 10,000. The hyperparameters of the word-based model were selected on the same English Wikipedia corpus as the character-based model.[6]

Results and Discussion

In this section, we show that surprisal computed by a character-level neural language model (CHARSURPRISAL) is able to account for the effects of errors on reading observed in our eye-tracking experiments. We compute character-based surprisal for the texts used in our experiments, and expect to obtain mean surprisal scores for each experimental condition that resemble mean reading times. We will also verify our prediction that word-based surprisal (WORDSURPRISAL) is not able to account for the effects observed in our experimental data, due to the way it treats unknown words.

Figure 2 shows the mean surprisal values across the different error conditions. We note that the pattern of reading time predicted by CHARSURPRISAL (solid lines) matches the first-pass times observed experimentally very well (see Figure 1), while WORDSURPRISAL (dotted line) shows a clearly divergent pattern, with error words showing *lower* surprisal than non-error words. This can be explained by the fact that a word-based model does not process error words beyond recognizing them as unknown; the presence of an unknown word itself is not a high-surprisal event (even without errors, 17 %

[5] 1024 units, 3 layers, batch size 128, embedding size 200, learning rate 3.6 with plain SGD, multiplied by 0.95 at the end of each epoch; BPTT length 80; DropConnect with rate 0.01 for hidden units; replacing entire character embeddings by zero with rate 0.001.

[6] 1024 units, batch size 128, embedding size 200, learning rate 0.2 with plain SGD, multiplied by 0.95 at the end of each epoch; BPTT length 50; DropConnect with rate 0.2 for hidden units; Dropout 0.1 for input layer; replacing words by random samples from the vocabulary with rate 0.01 during training.

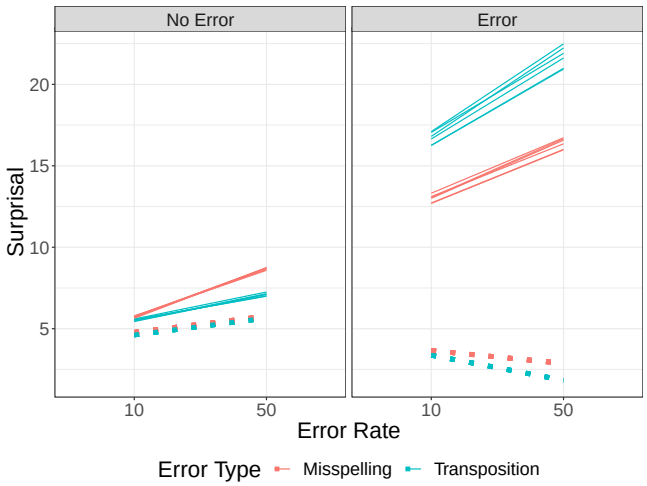


Figure 2: CHARSURPRISAL (full lines) and WORDSURPRISAL (dotted lines) as a function of error type and error rate, for correct (left) and erroneous (right) words. For CHARSURPRISAL, we show the means of all seven random initializations of our neural surprisal model.

of the words in our texts are unknown to the model, given its 10,000-word vocabulary).

To confirm this observation statistically, we fitted linear mixed-effects models with CHARSURPRISAL and WORDSURPRISAL as dependent variables. We enter the seven random initializations of each model as a random factor, analogously to the participants in the eye-tracking experiment. We use the same predictors that we used for the reading measures, except for LASTFIX. This predictor is not available: suprisal models compute a difficulty measure for each word (viz., its surprisal), but they are not able to predict whether a word will be skipped or not.

The results of the mixed model with CHARSURPRISAL as the dependent variable (see Table 4) replicated the effects of ERRORRATE, ERROR, and WORDLENGTH found in first pass and fixation rate, as well as the effect of ERRORTYPE found only in fixation rate (see Table 3). The same mixed model with WORDSURPRISAL as the dependent variable (see again Table 4), however, does not yield the correct pattern of results: Crucially, the coefficients of ERROR and ERRORTYPE have the opposite sign compared to both CHARSURPRISAL and the experimental data (though both effects are small, see dotted lines in Figure 2).

We have shown that character-based surprisal computed on the texts used in our experiment is qualitatively similar to the experimental results. As a next step we will test its quantitative predictions, i.e., we will correlate surprisal scores with reading times. For this, we performed mixed-effects analyses in which first-pass time and fixation rate are predicted by WLENGTH, LASTFIX, and character-based surprisal residualized against word length (RESIDCHARSURP).[7] Note that

[7]The correlation between word length and raw surprisal is 0.26.

	CHARSURPR	WORDSURPR
(Intercept)	10.47 (0.09)***	5.06 (0.07)***
ERRTYPE	1.27 (0.02)***	−0.40 (0.02)***
ERRRATE	1.57 (0.02)***	0.01 (0.00)***
ERROR	13.88 (0.03)***	−2.96 (0.02)***
WLENGTH	3.02 (0.05)***	0.25 0.01 ***

$Pr(\beta < 0)$: *** < 0.001, ** < 0.01, * < 0.05

Table 4: Models of character-level and word-level surprisal with random effects for model runs and items. Each cell gives the coefficient, its standard deviation and the estimated posterior probability that the coefficient has the opposite sign.

	First Pass	Fixation Rate
(Intercept)	248.73 (5.55)***	−0.15 (0.09)
WLENGTH	22.22 (0.79)***	0.75 (0.01)***
LASTFIX	2.65 (1.34)	0.22 (0.02)***
WLENGTH × LASTFIX	—	0.60 (0.19)***
RESIDCHARSURP-ORACLE	9.89 (0.78)***	0.09 (0.01)***
RESIDCHARSURP	13.82 (0.66)***	0.14 (0.01)***
ΔAIC	−273.88	−205.83
ΔBIC	−273.88	−205.83

$Pr(\beta < 0)$: *** < 0.001, ** < 0.01, * < 0.05

Table 5: Models for reading measures with surprisal predictors. We compare model fit between a model with character-based surprisal (RESIDCHARSURP) and character-based oracle surprisal (RESIDCHARSURPORACLE), both residualized against word length.

we did not enter the error factors (ERRORTYPE, ERRORRATE, ERROR) into this analysis, as we predict that surprisal will simulate the effect of errors in reading.

It is known that surprisal predicts reading times in ordinary text not containing errors (Demberg & Keller, 2008; Frank, 2009); thus, it is important to disentangle the specific contribution of modeling errors correctly from the general contribution of surprisal in our model. We do this by constructing a baseline version of character-based surprisal that is computed using an oracle (RESIDCHARSURPORACLE). For this, we replace erroneous words with their correct counterparts before computing surprisal, and again residualize against word length.[8] If RESIDCHARSURP correctly accounts for the effects of errors on reading, then we expect that RESIDCHARSURP – which has access to the erroneous word forms – will improve the fit with our reading data compared to RESIDCHARSURPORACLE.

For RESIDCHARSURPORACLE, we use the same seven models as for RESIDCHARSURP, only exchanging the char-

[8]The correlation between word length and unresidualized oracle surprisal is 0.47.

acter sequences on which surprisal is computed. This ensures that any difference in model fit between the two predictors can be attributed entirely to the way RESIDCHARSURP is affected by the presence of errors in the texts.

The resulting models are shown in Table 5. For WLENGTH and LASTFIX, we see the same pattern of results as in the experimental data (see Table 3). Furthermore, regular surprisal (RESIDCHARSURP) and oracle surprisal (RESIDCHARSURPORACLE) significantly predict both first pass time and fixation rate. This is in line with the standard finding that surprisal predicts reading time (Demberg & Keller, 2008; Frank, 2009), but has so far not been demonstrated for texts containing errors. We compare model fit using AIC and BIC. Both measures indicate that RESIDCHARSURP fits the experimental data better than RESIDCHARSURPORACLE. Thus, character-level surprisal provides an account of our data going beyond the known contribution of ordinary surprisal to reading times, and correctly predicts reading in the presence of errors.

Conclusion

We investigated reading with errors in texts that contain either letter transpositions or real misspellings. We found that transpositions cause more reading difficulty than misspellings and explained this using a character-based surprisal model, which assigns higher surprisal to rare letter sequences as they occur in transpositions. We also found that in texts with a high error rate, all words are more difficult to read, even the ones without errors. Again, character-based surprisal explains this: computing word expectations is harder when the context of a word is degraded by errors, resulting in increased surprisal.

In future work, we plan to integrate character-based surprisal with existing neural models of human reading (Hahn & Keller, 2018). Models at the character level are necessary not only to account for errors, but also to model landing position effects, parafoveal preview, and word length effects, all of which word-based models are unable to capture.

Acknowledgements

Y.B. was supported by the Harvard Mind, Brain, and Behavior Initiative. F.K. was supported by the Leverhulme Trust through International Academic Fellowship IAF-2017-019.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using **lme4**. *Journal of Statistical Software*, *67*(1), 1–48.

Belinkov, Y., & Bisk, Y. (2018). Synthetic and natural noise both break neural machine translation. In *Proceedings of the 6th International Conference on Learning Representations.* Vancouver, Canada.

Bürkner, P.-C. (2017). brms: An r package for bayesian multilevel models using stan. *Journal of Statistical Software*, *80*(1), 1–28.

Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, *109*(2), 193–210.

Frank, S. L. (2009). Surprisal-based comparison between a symbolic and a connectionist model of sentence processing. In N. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st annual conference of the cognitive science society* (pp. 1139–1144). Amsterdam.

Geertzen, J., Alexopoulou, T., & Korhonen, A. (2014). Automatic Linguistic Annotation of Large Scale L2 Databases: The EF-Cambridge Open Language Database (EFCamDat). In R. T. Miller (Ed.), *Selected Proceedings of the 2012 Second Language Research Forum* (pp. 240–254).

Hahn, M., & Keller, F. (2018). *Modeling task effects in human reading with neural attention.* (arXiv:1808.00054)

Hermann, K. M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., & Blunsom, P. (2015). Teaching machines to read and comprehend. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 28* (pp. 1693–1701).

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Johnson, R. L., Perea, M., & Rayner, K. (2007). Transposed-letter effects in reading: Evidence from eye movements and parafoveal preview. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(1), 209–229.

Kaakinen, J. K., & Hyönä, J. (2010). Task effects on eye movements during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(6), 1561–1566.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.

Merity, S., Keskar, N. S., & Socher, R. (2018). *An analysis of neural language modeling at multiple scales.* (arXiv:1803.08240)

Rayner, K., White, S. J., Johnson, R. L., & Liversedge, S. P. (2006). Raeding wrods with jubmled lettres: There is a cost. *Psychological Science*, *17*(3), 192–193.

Schotter, E. R., Bicknell, K., Howard, I., Levy, R., & Rayner, K. (2014). Task effects reveal cognitive flexibility responding to frequency and predictability: Evidence from eye movements in reading and proofreading. *Cognition*, *131*(1), 1–27.

White, S. J., Johnson, R. L., Liversedge, S. P., & Rayner, K. (2008). Eye movements when reading transposed text: The importance of word-beginning letters. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(5), 1261–1276.

Idea Generation and Goal-Derived Categories

Richard W. Hass (Richard.Hass@jefferson.edu), J. Colin Long, Joshua Pierce
College of Humanities and Sciences, Thomas Jefferson University
Philadelphia, PA 19144 USA

Abstract

Semantic search and retrieval of information plays an important role in creative idea generation. This study was designed to examine how semantic and temporal clustering varies when asking participants to generate ideas about uses for objects compared with generating members of goal-derived categories. Participants generated uses for three objects: brick, hammer, picture frame, and also generated members of the following goal-derived categories: things to take in case of a fire, things to sell at a garage sale, and ways to spend lottery winnings. Using response-time analysis and semantic analysis, results illustrated that all six prompts generally led to exponential cumulative response-time distributions. However, the proportion of temporally clustered responses, defined using the slope-difference algorithm, was higher for goal-derived category responses compared with object uses. Despite that, overall pairwise semantic similarity was higher for object uses than for goal derived exemplars. The effect of prompt on pairwise semantic similarity is likely the result of context-dependency of exemplars from goal-derived categories. However, the current analysis contains a potential confound such that special instructions to give "common and uncommon" responses were provided only for the object-uses prompts. The confound is likely minimal, but future work is necessary to verify that these results would hold when the confound is removed.

Keywords: Creativity; Divergent Thinking; Goal-Derived Categories; Latent Semantic Analysis; Semantic Memory

Creative cognition researchers often highlight the contributions of memory structure and process to creative idea generation. Though theories vary widely in explaining how existing knowledge is actually used to support the generation of creative ideas and products, there is sufficient evidence to suggest that in both laypeople (Ward, 2008), and in eminent creators (Weisberg, 2006) creative thinking operates within the bounds of an individual's system of knowledge. This study was designed to extend recent work (Hass, 2017a) exploring the degree of semantic clustering found among ideas generated when participants complete divergent thinking tasks. Divergent thinking tasks are heavily used as a proxy for creative thinking in a variety of behavioral (Snyder, Hammond, Grohman, & Katz-Buonincontro, 2019) and neuroscientific (Dietrich & Kanso, 2010) studies. There is a general consensus that dynamic interplay among executive search and control processes and semantic memory organization enables the generation of creative ideas (cf. Abraham & Bubic, 2015; Beaty, Christensen, Benedek, Silvia, & Schacter, 2017; Chrysikou & Thompson-Schill, 2011).

The central aim of the study was to extend prior results (e.g., Hass, 2017a; Hass & Beaty, 2018) by comparing semantic processing during object-uses generation to the generation of exemplars from goal-derived categories (Barsalou, 1985). Generating uses for objects is the core feature of the Alternative Uses task (Wilson, Guilford, Christensen, & Lewis, 1954), one of the most popular divergent thinking tasks, and its validity as a psychometric measure of creative thinking is enhanced by illuminating the underlying cognitive processes operating while people perform it. Creative thinking has also been described as related to goal-derived knowledge (Chrysikou, 2006), so it is natural to explicitly examine potential similarities between generating uses for objects and generating exemplars of goal-derived categories. The paper is structured as follows: first, research on the relationship between semantic memory retrieval and idea generation will be reviewed, along with a brief discussion of how divergent thinking tasks like object-uses generation relate to goal-derived category recall or generation tasks. Then, the analysis is presented in three phases: an analysis of cumulative response-time functions across conditions, an analysis of temporal clustering of responses, and an analysis of the semantic similarity of pairs of responses across two prompts, one from each condition.

Knowledge and creative generation

As mentioned, cognitive accounts of creativity tend to differentially emphasize the importance of associative processes of semantic organization and executive control of thought (cf. Chrysikou & Thompson-Schill, 2011; Mednick, 1962). In an early theoretical account, (Mednick, 1962) suggested that creative idea generation is underpinned by associative networks that afford more remote connections among concepts. Recent studies of creative thinking have shown support for this view, illustrating that individuals with flexible semantic networks tend to perform better on creative cognitive tasks and report a greater number of creative achievements (e.g. Kenett, Beaty, Silvia, Anaki, & Faust, 2016). Additional studies have highlighted the influence of executive control on the remote association process. For example (Beaty, Silvia, Nusbaum, Jauk, & Benedek, 2014) showed that the fluency and originality of uses for objects was almost equally well predicted by measures of remote association and associative flexibility, the latter thought to be an index of executive control over lexical association. Similarly, (Hass, 2017b) showed that the degree to which creative uses for objects were seman-

tically distant from the core of the prompt object concept was positively related to fluid intelligence.

Goal-derived Categories A key aspect of this study was to propose that goal-derived category exemplar generation can serve as a basis for understanding object-uses generation. Goal-derived categories constructed during goal-directed activities (Barsalou, 1983, 1985), like deciding which chair to sit on in a cafe or which personal belongings to keep from a childhood home. These categories can be distinguished from "natural" or "taxonomic" categories in several ways, though, we focus on two. First, goal-derived categories are constructed when performing decision-making tasks; defined by personal objectives and constrained by the environment or immediate context. Second, it can be argued that goal-derived categories such as "things to take with you in case of a fire" are not as well-established in memory as categories such as breakfast foods (Barsalou, 1983). Omelets and pancakes are within the same category (breakfast foods) because they are edible, made with eggs, served warm, are eaten in the morning, and relatively straightforward to cook. Attributes like those, such as times of consumption and ingredients, which are the basis of category discrimination (Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976), do not co-occur as frequently in goal-derived categories. In addition, goal-derived categories are not as well used in the literature on semantic search and retrieval, so this analysis provides novel information about memory search dynamics when people name members of goal-derived categories. More importantly for this analysis, Barsalou (1985) suggested that retrieval of conceptual information for category processing involves generation of multiple conceptual representations, each held in working memory, when the category is encountered in normal life. This reliance on multiple conceptual representations could account for the effects reviewed above relating object-uses generation to fluid intelligence. Thus, comparing goal-derived category search to object uses search serves the dual purpose of exploring how context-dependent organization and executive control might interact during idea generation.

The current study

The primary focus of the current analysis was on semantic clustering. Because there are no established category norms for the prompts used in this study (cf. Troyer, Moscovitch, & Winocur, 1997), clusters were first identified using the slope-difference algorithm (Gruenewald & Lockhead, 1980). Latent semantic analysis (Landauer & Dumais, 1997) was then used to quantify the semantic similarity among sequential pairs of responses. The slope-difference algorithm identifies potential semantic clusters in terms of the difference between an actual IRT and the expected IRT given a mathematical relation between response time and output total. It was expected that slope-difference clusters would be more prevalent in the goal-derived response arrays, and that the pairwise semantic similarity of within-cluster responses would be higher in goal-derived response arrays. The reasons to expect that goal-derived response arrays would be more clustered than object-uses arrays are two-fold. First, response totals are usually quite low when people generate uses for objects, and though clusters appear, the number of responses per cluster is usually small. As cluster size decreases, output total should follow (Herrmann & Pearle, 1981), and the lack of success in finding newly retrieved clusters will ultimately lead to search termination (Raaijmakers & Shiffrin, 1981). Second, memory search is often described as a multiply-constrained problem (e.g., Polyn, Norman, & Kahana, 2009; Smith, Huber, & Vul, 2013), with multiple sources of information vying for attention in the process. It is plausible that goal-derived category generation is less constrained than object-use generation, such that a single context-dependent goal (e.g., "items to sell at a garage sale") remains in mind. This should enable the integration of contextual and semantic information in more efficient manner than in object-use generation, where the goal may change from response to response (e.g., "use a brick as a weight" $\rightarrow$ "use a brick as a pencil holder").

As will be described, the prompt used for object-uses in the current study was to "think of common *and* uncommon uses", designed to provide a more natural comparison to the generation of category exemplars (i.e., the word "creative" was not used in the instructions). That is, several studies have shown that instructing participants to "be creative" decreases fluency (output total), while increasing the average originality of their responses (Forthmann et al., 2016; Nusbaum, Silvia, & Beaty, 2014). Since the primary interest in the current study was the nature of the category itself (e.g., use of an object v. goal-derived category) and *not* whether participants were trying to engage in creative thought, we felt the special instruction was warranted. However, as we discuss, the inclusion of this "common and uncommon" instruction was not used in the goal-derived conditions, which presents a confound. The nature of our results do not suggest the confound is serious, it is important to keep in mind.

Method

Participants

A total of 32 participants were recruited from undergraduate psychology courses. Participants were offered extra credit or chocolate in compensation for their time. Participants ranged in age from 18 to 25 years old, and the demographics were consistent with a traditional undergraduate university in the northeastern United States. All recruitment and consent procedures were approved by the university's Institutional Review Board.

Materials

Participants completed the tasks using a custom Matlab interface on an Apple iMac. Instructions and prompts appeared as text on white background above a text-box where participants entered responses. Instructions were displayed and read to participants prior to each of three task blocks, the first

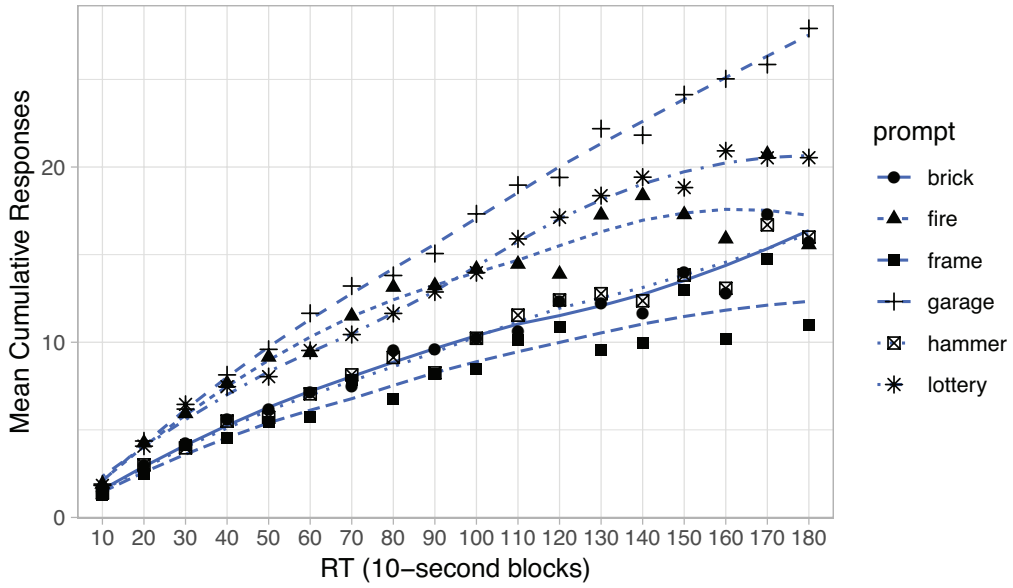


Figure 1: Mean number of cumulative responses per 10-second blocks for each of the six prompts. Object-uses prompts: brick, (picture) frame, hammer; Goal-derived prompts: fire = things to take in case of fire, garage = things to sell at a garage sale, lottery = ways to spend lottery winnings. Note: lines represent loess fits, not the exponential fits used in the next section.

of which was a practice block (naming colors). Instructions were not visible during response generation, but the prompt remained displayed for the entire duration of each response-generation interval. Demographic information was obtained using a pencil-and-paper survey after the experiment finished.

Procedure

Participants were greeted by the experimenter, and were told that the experiment was designed to test memory. The experimenter read general instructions about how the computer system worked and instructed participants to type responses on the computer keyboard, and to enter responses by pressing the return button. Participants then practiced this by naming colors (at least 3) for 30 seconds. The experimenter then answered any questions before the experiment began. Matlab recorded the time of the first keypress of each response, the time between the first key-stroke and the response entry, and the actual response.

The tasks were presented in two blocks of 3 prompts each, with a break in between each block. Both the order of the blocks and the order of presentation of prompts within the blocks were randomized by Matlab code. All participants responded to each prompt in each block. Each response interval was three minutes in length to permit valid comparison among the two prompt conditions (goal-derived categories, and object-uses prompts). The goal-derived category prompts began with "name examples of" and ended with one of the three categories: things to spend lottery winnings on, things to take from your house if it caught on fire, and things to sell at a garage sale. The object-use prompts began with "name common and uncommon uses for a" and ended with one of the three prompt objects: brick, hammer, and picture frame.

The entire prompt phrase remained on the screen above the text-entry box for the entire 3 minutes. When 3 minutes expired, the screen displayed a message indicating that the next prompt was loading for 5 seconds before the next prompt appeared. After the first and second blocks, instructions for the next block appeared on the screen, and the participant was given a 1-2 minute break before beginning the next block. After the final block, a thank-you message appeared and the participant filled out the demographic questionnaire, and the experimenter answered any questions the participant may have had. The entire process lasted between 20 and 25 minutes for each participant.

Analyses and Results

All analyses were conducted using the R Statistical Programming Language, and all data and algorithms are available for download (https://osf.io/fvne2/). Response times were defined in terms of the time (since presentation of the prompt) of the first key-press of each response, to be consistent with studies using voice-keyed response recording. Prior to analysis, data were examined for repeated responses and malfunctions in Matlab's execution of the experiment. Three participants were excluded due to Matlab malfunctions reducing the final sample size to 29. Repeated responses were those that were identified as the same response given more than once by the same individual to a specific prompt. When repeats were identified, the RTs for those responses were removed from the data set. Participants gave a total of 1746 responses to the three goal-derived prompts after the removal of 23 repeated responses. Finally, participants gave a total of 1012 responses

Table 1: Average response totals per category and intercorrelations (Spearman's ρ). Object Uses prompts are in the top half of the table. All correlations are significant, with $p \leq .01$, except the correlation between garage sale and hammer totals ($p = .06$).

			ρ				
Prompt	M	SD	1	2	3	4	5
1. Brick	12.76	4.66	-				
2. Hammer	11.86	5.09	.64	-			
3. Frame	10.28	4.40	.72	.78	-		
4. ... sell at garage sale	24.17	7.40	.35	.67	.46	-	
5. ... take from fire	17.17	5.33	.67	.52	.48	.53	-
6. ... do with lottery winnings	18.86	6.69	.50	.57	.46	.65	.55

to the object-use prompts after the removal of 11 responses.

Fluency and Cumulative Response Times

The mean number of cumulative responses was computed in 10-second blocks and plotted in Figure 1. Clearly there is nonlinearity, and not surprisingly, fluency is higher for the goal-derived prompts compared with the object-use prompts. The shape of the distributions in these plots is consistent with those found in normal memory retrieval studies. Table 1 further illustrates that response totals are uniformly lower for objects uses prompts, and that there is a relatively large degree of correlation among output totals.

Clustering

Clusters were identified using a modification the Slope Difference Algorithm (Gruenewald & Lockhead, 1980), that uses an exponential function rather than the hyperbolic function used by Gruenewald and Lockhead:

$$R(t) = N(1 - e^{-\frac{t}{\tau}}) \quad (1)$$

This is the "two parameter" exponential, with N being the estimated asymptote, or number of responses generated with an unlimited amount of time, and τ being the inverse of the rate parameter λ in an exponential distribution. Thus, τ is parameterized as the estimated mean response time.

The Slope-Difference algorithm works as follows: given the estimated N and τ parameters for each participant, calculate the difference between the predicted and observed instantaneous rates of change in responding. The predicted rate of change is just the derivative of Equation 1 calculated with each participant's parameters and the cumulative response times of that participant. The observed instantaneous rate of change is just the reciprocal of each inter-response time (IRT) (i.e., for R = cumulative number of responses, $\frac{\Delta R}{\Delta t} = \frac{1}{IRT}$). Gruenewald and Lockhead (1980) provided support for the validity of the algorithm, such that large, positive differences between observed and predicted rates were indications that responding was faster than predicted, and thus, faster than expected responses qualify as being within clusters. The threshold for slope-differences being categorized as "switches" was .10, which is the same as used by Grunewald and Lockhead.

To obtain slope-difference values, Equation 1 was first fit to each participant's cumulative response-time distribution *per prompt*, using ordinary nonlinear least-squares estimation. Parameter values along with response times were used to compute predicted rates of change to be differenced from the actual rates of change. Clusters were then identified as any IRT with a slope difference value less than .10, the threshold used by Gruenewald and Lockhead (1980). Exponential parameter estimates were not optimal for 1-3 participants per prompt, and data from those participants were excluded for the cluster analysis of each prompt.

Figure 2 shows that the proportion of responses identified as within cluster was significantly greater for goal-derived categories compared with object uses, $\chi^2(1) = 48.27, p < .001$. Of the 998 object-use responses, 18.8% were identified as within-cluster, while 31.3% of the 1567 goal-derived responses were identified as within-cluster.

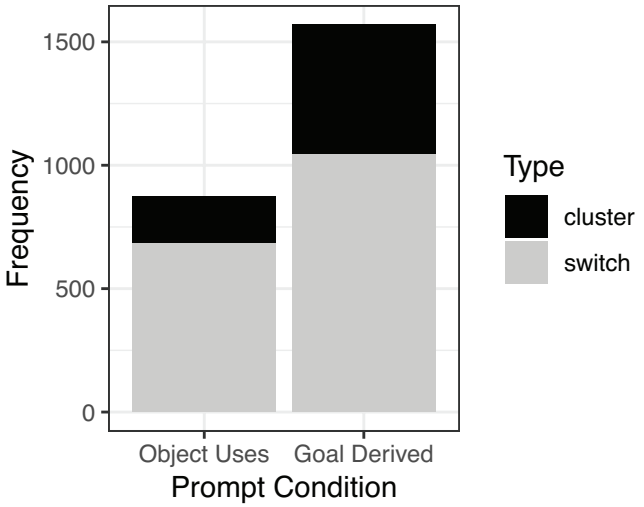


Figure 2: Number of responses classified as within cluster, or as a switch between clusters by the slope difference algorithm for the two types of prompts. See text for proportions.

Pairwise Semantic Similarity

The validity of the slope difference algorithm rests on further semantic analysis of its results. Here, the central question was whether responses in clusters corresponding to goal-derived category were more semantically similar than those in clusters corresponding with object-use generation. Pairs of sequential responses were analyzed for semantic similarity using the tools at the UC Boulder website (lsa.colorado.edu). The General Reading corpus, with 300 factors, was chosen as the basis for comparisons, and the term-to-term comparator was used.

Mixed-effects regression was used to examine the main-effects of clustering (within cluster v. between cluster response) and prompt-type (goal-derived v. object use) on LSA-derived cosine similarities, and the interaction of the two fixed effects. A random intercept term was added to account for participant variation, and another to account for variations across the 6 prompts. Table 1 contains the results of the analysis including 95% confidence intervals for the fixed and random effects terms. Rather surprisingly, on average the pairwise similarity of responses to the goal-derived prompts was less than the average pairwise similarity of object-uses responses. However, the slope difference algorithm seems to distinguish between semantic clusters such that on average, within-cluster responses were less similar (in terms of pairwise similarity) than between cluster responses. Figure 3 illustrates that there may be a small interaction between prompt type and clustering, and in Table 2, the estimate is a slightly smaller difference in similarity of clustered and non-clustered responses for object uses compared with goal derived categories, though zero remains a plausible value for the interaction.

A slightly different result is obvious if one plots pairwise similarity as a function of IRT. Figure 4 shows that, at the level of individual pairs of responses, the relationship between IRT and similarity is not linear, and that for a great many pairs of responses on all six prompts, there is a substantial degree of variability in pairwise similarity for short IRTs. A closer look at Figure 4 reveals that the garage prompt has the highest concentration of low-similarity pairs. This an interesting result on its own and is likely the result of context dependency for that prompt, as will be discussed next.

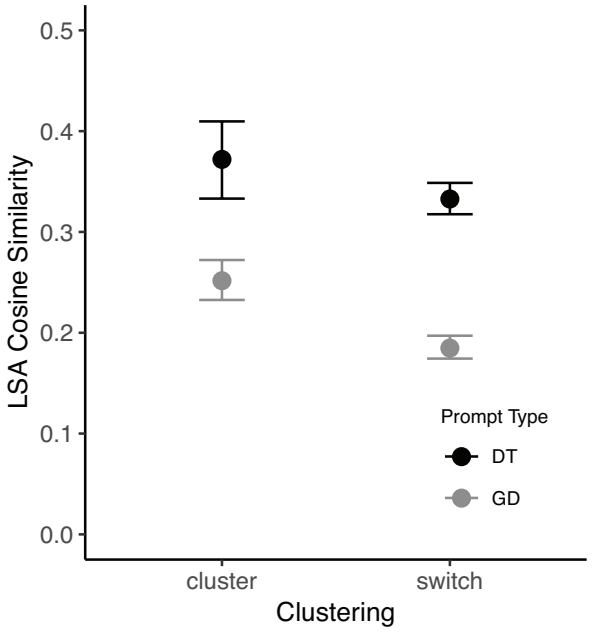


Figure 3: Average pairwise semantic similarity per prompt and per cluster category (in cluster v. switch). Bars are bootstrapped 95% confidence limits.

Discussion

The purpose of this study was to probe the differences between object-uses generation and goal-derived category exemplar generation in terms of semantic search and retrieval. Using measures of clustering and similarity, this analysis illustrated that there may be two key differences between responding to these two types of prompts. First, output totals for goal-derived categories were much higher than object uses, and also contained a greater proportion of faster than expected IRTs, identified by the slope difference algorithm.

Though semantic analysis of adjacent pairs of responses showed that the slope-difference clusters are indeed semantic clusters, pairwise similarity among goal-derived exemplars was surprisingly less than the similarity among adjacent pairs of object-use responses. One explanation for this result is that semantic relationships *across* clusters of goal-derived exemplars may be minimal because of the dependence of semantic similarity on context (e.g., Barsalou, 1982). Of course, that characterization might also be said of object-uses. More importantly, Hass (2017a) illustrated that LSA-derived cosine similarities may not accurately represent context-dependent relationships between object uses. Indeed, the main difference between the two types of prompts is that goal-derived prompts identify a context (e.g., a garage sale), which all items must relate to in some way, while object-uses prompts identify an exemplar (e.g., a brick) to which responses must relate. While object-uses responses will likely have context-dependency, it is also likely that context dependency will be greater among goal-derived categories such as those in this study, as the context itself is the main constraint on conceptual activation. That is, consider the example discussed in the introduction: electronics to sell at a garage sale. Say a participant activates electronics as a concept and exploits it for a bit, what is the likelihood that the next conceptual representation activated will be highly similar to electronics in a context-independent fashion? Contrast that with the activation of the attribute "heavy" in generating uses for a brick. What is the likelihood that the next conceptual representation used after "heavy" is going to be semantically similar to "heavy" in a context-independent sense. It seems plausible that the semantic similarity among all activated concepts

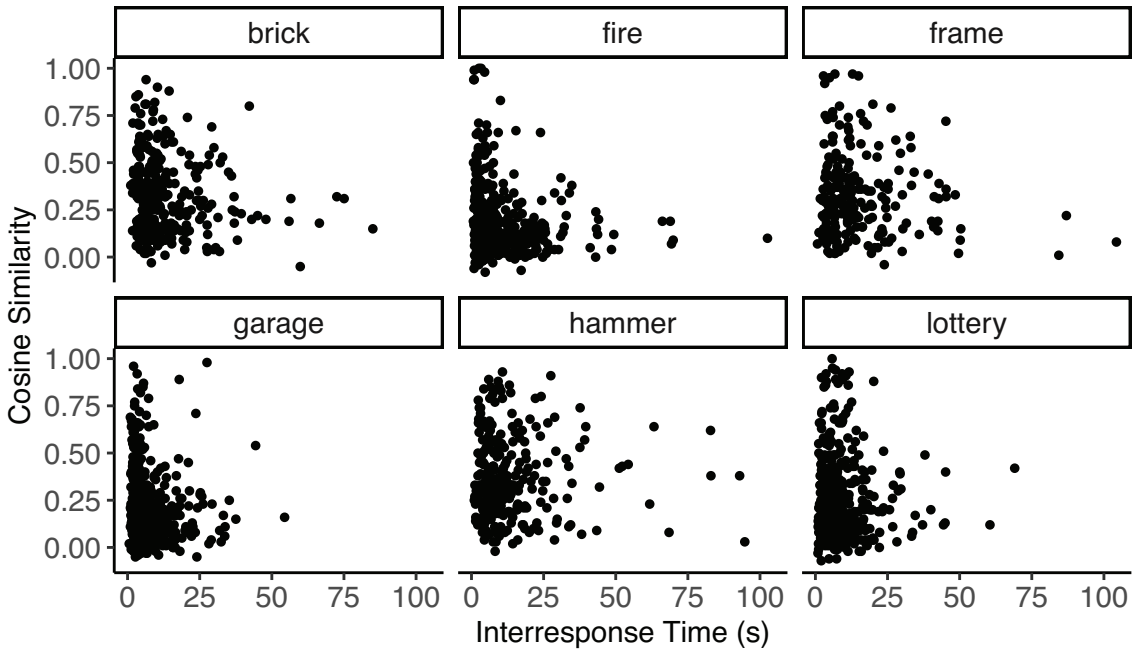


Figure 4: Scatterplots of the IRT-similarity relationship across all size prompts. Object uses prompts are in the top row, goal-derived prompts are in the bottom row

Table 2: Results of the mixed-effects regression with pairwise similarity as the dependent variable. The baseline prompt-type condition was object-use, the baseline cluster condition was Within Cluster.

			Confidence Interval	
Fixed Effects	b	*t*	2.5%	97.5%
Intercept	0.38	14.65	0.33	0.42
Prompt Type	-0.12	-3.88	-0.18	-0.06
Cluster	-0.04	-2.24	-0.08	-0.005
prompt × switch	-0.03	-1.49	-0.07	0.01
Random Effects	σ		2.5%	97.5%
Participant	0.04		0.03	0.06
Prompt	0.03		0.01	0.05
Residual	0.20		0.19	0.20

in the garage context might be lower than the semantic similarity among activated concepts in the context of a use for a brick because of the dependence on the garage context. The latter conclusion is still highly speculative, but it suggests that this is a fruitful avenue for future research to follow, as it will likely illuminate how semantic information is organized and used in both kinds of tasks.

Limitations and future directions

In this study, participants were explicitly instructed to *think of common and uncommon uses for objects* in an effort to obtain a greater total number of responses generated across the object uses prompts (i.e., the word "creativity" was not present in the instructions). In the recent study by Hass (2017a), participants were instructed to think of *creative* uses for objects, and indeed, their response totals were, on average, lower than the current study (about 7 responses). So it is likely that the instruction to be creative may limit the semantic similarity of clustered output when generating object uses. Indeed, the major motivation for the choice to avoid the word creative was to provide a baseline for future studies that would vary instructions, including "be-creative" conditions (e.g., Forthmann et al., 2016), and strategy inductions (e.g., Unsworth, Brewer, & Spillers, 2013). However, since participants were only given the "common and uncommon" instructions in one condition, the effect of prompt type on semantic similarity is confounded by the differing instructions. Specifically, our use of the phrase "common *and* uncommon" uses in the object use condition may have confused participants, or led some participants to approach the task differently from others, with some potentially assuming that they should be creative, or only think of uncommon uses. We believe that this can be remedied in future studies by changing all prompts to be of the form, "think of things to ..." and then appending the prompt (e.g., ... to sell at a garage sale; ... to do with a brick). Participants can then be instructed to perform the two tasks in the ways just mentioned (e.g., creatively, or using a certain search strategy), without the confound currently present. However, the current results are still informative, and it is likely that the confound presented by the "common or uncommon" phrasing was minimal.

References

Abraham, A., & Bubic, A. (2015, March). Semantic memory as the root of imagination. *Frontiers in Psychology*, *6*, 1–5.

Barsalou, L. W. (1982, January). Context-independent and context-dependent information in concepts. *Memory & Cognition*, *10*(1), 82–93.

Barsalou, L. W. (1983). Ad hoc categories. *Memory & Cognition*, *11*, 211–227.

Barsalou, L. W. (1985, October). Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*(4), 629–654.

Beaty, R. E., Christensen, A. P., Benedek, M., Silvia, P. J., & Schacter, D. L. (2017, March). Creative constraints: Brain activity and network dynamics underlying semantic interference during idea production. *NeuroImage*, *148*(C), 189–196.

Beaty, R. E., Silvia, P. J., Nusbaum, E. C., Jauk, E., & Benedek, M. (2014, June). The roles of associative and executive processes in creative cognition. *Memory & Cognition*, *42*(7), 1186–1197.

Chrysikou, E. G. (2006). When Shoes Become Hammers: Goal-Derived Categorization Training Enhances Problem-Solving Performance. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(4), 935–942.

Chrysikou, E. G., & Thompson-Schill, S. L. (2011, April). Dissociable brain states linked to common and creative object use. *Human Brain Mapping*, *32*(4), 665–675.

Dietrich, A., & Kanso, R. (2010). A review of EEG, ERP, and neuroimaging studies of creativity and insight. *Psychological Bulletin*, *136*(5), 822–848.

Forthmann, B., Gerwig, A., Holling, H., Celik, P., Storme, M., & Lubart, T. (2016, July). The be-creative effect in divergent thinking: The interplay of instruction and object frequency. *Intelligence*, *57*, 25–32.

Gruenewald, P. J., & Lockhead, G. R. (1980, May). The free recall of category examples. *Journal of Experimental Psychology: Human Learning and Memory*, *6*(3), 225–240.

Hass, R. W. (2017a, September). Semantic search during divergent thinking. *Cognition*, *166*, 344–357.

Hass, R. W. (2017b, October). Tracking the dynamics of divergent thinking via semantic distance: Analytic methods and theoretical implications Richard W. Hass. *Memory & Cognition*, *45*, 233–244.

Hass, R. W., & Beaty, R. E. (2018). Use or consequences: Probing the cognitive difference between two measures of divergent thinking. *Frontiers in psychology*, *9*, 2327.

Herrmann, D. J., & Pearle, P. M. (1981). The proper role of clusters in mathematical models of continuous recall. *Journal of Mathematical Psychology*, *24*(2), 139–162.

Kenett, Y. N., Beaty, R. E., Silvia, P. J., Anaki, D., & Faust, M. (2016). Structure and Flexibility: Investigating the Relation Between the Structure of the Mental Lexicon, Fluid Intelligence, and Creative Achievement. *Psychology of Aesthetics, Creativity, and the Arts*, *10*, 377–388.

Landauer, T. K., & Dumais, S. T. (1997, April). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211–240.

Mednick, S. A. (1962). The associative basis of the creative process. *Psychological Review*, *69*(3), 220–232.

Nusbaum, E. C., Silvia, P. J., & Beaty, R. E. (2014). Ready, Set, Create: What Instructing People to "Be Creative" Reveals About the Meaning and Mechanisms of Divergent Thinking. *Psychology of Aesthetics, Creativity, and the Arts*, *8*(4), 423–432.

Polyn, S. M., Norman, K. A., & Kahana, M. J. (2009). A context maintenance and retrieval model of organizational processes in free recall. *Psychological Review*, *116*(1), 129–156.

Raaijmakers, J. G., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review*, *88*(2), 93–134.

Rosch, E. H., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*, 382–439.

Smith, K. A., Huber, D. E., & Vul, E. (2013, July). Multiply-constrained semantic search in the Remote Associates Test. *Cognition*, *128*(1), 64–75.

Snyder, H. T., Hammond, J. A., Grohman, M. G., & Katz-Buonincontro, J. (2019). Creativity measurement in undergraduate students from 1984–2013: A systematic review. *Psychology of Aesthetics, Creativity, and the Arts*, *13*(2), 133–143.

Troyer, A. K., Moscovitch, M., & Winocur, G. (1997, January). Clustering and switching as two components of verbal fluency: evidence from younger and older healthy adults. *Neuropsychology*, *11*(1), 138–146.

Unsworth, N., Brewer, G. A., & Spillers, G. J. (2013, June). Strategic search from long-term memory: An examination of semantic and autobiographical recall. *Memory*, *22*(6), 687–699.

Ward, T. B. (2008, November). The role of domain knowledge in creative generation. *Learning and Individual Differences*, *18*(4), 363–366.

Weisberg, R. (2006). *Creativity: Understanding innovation in science, problem solving, and the arts*. Wiley.

Wilson, R. C., Guilford, J. P., Christensen, P. R., & Lewis, D. J. (1954). A factor-analytic study of creative-thinking abilities. *Psychometrika*, *19*(4), 297–311.

Disentangling contributions of visual information and interaction history in the formation of graphical conventions

Robert X. D. Hawkins*
Department of Psychology
Stanford University
rxdh@stanford.edu

Megumi Sano*
Department of Psychology
Stanford University
megsano@stanford.edu

Noah D. Goodman
Department of Psychology
Stanford University
ngoodman@stanford.edu

Judith E. Fan
Department of Psychology
UC San Diego
jefan@ucsd.edu

Abstract

Drawing is a versatile technique for visual communication, ranging from photorealistic renderings to schematic diagrams consisting entirely of symbols. How does a medium spanning such a broad range of appearances reliably convey meaning? A natural possibility is that drawings derive meaning from both their visual properties as well as shared knowledge between people who use them to communicate. Here we evaluate this possibility in a drawing-based reference game in which two participants repeatedly communicated about visual objects. Across a series of controlled experiments, we found that pairs of participants discover increasingly sparse yet effective ways of depicting objects. These gains were specific to those objects that were repeatedly referenced, and went beyond what could be explained by task practice or the visual properties of the drawings alone. We employed modern techniques from computer vision to characterize how the high-level visual features of drawings changed, finding that drawings of the same object became more consistent within a pair of participants and divergent across participants from different interactions. Taken together, these findings suggest that visual communication promotes the emergence of depictions whose meanings are increasingly determined by shared knowledge rather than their visual properties alone.

Keywords: alignment; coordination; iconicity; sketch understanding; visual communication

Introduction

From ancient etchings on cave walls to modern digital displays, visual communication lies at the heart of key human innovations (e.g., cartography, data visualization) and forms a durable foundation for the cultural transmission of knowledge and higher-level reasoning. Perhaps the most basic and versatile technique supporting visual communication is drawing, the earliest examples of which date to at least 40,000-60,000 years ago (Hoffmann et al., 2018). What began as simple mark making has since been adapted to a wide array of applications, ranging from photorealistic rendering to schematic diagrams consisting entirely of symbols.

Even in the relatively straightforward case of drawing from life, there are countless ways to depict the same object. How does a communication medium spanning such a broad range of appearances reliably convey meaning? On the one hand, prior work has found that semantic information in a figurative drawing, i.e., the object it represents, can be derived purely from its visual properties (Fan, Yamins, & Turk-Browne, 2018). On the other hand, other work has emphasized the role of socially-mediated information for making appropriate inferences about what even a figurative drawing represents (Goodman, 1976).

How can these two perspectives be reconciled? Our approach is to consider the joint contributions of visual information and social context in determining how drawings derive meaning (Abell, 2009), and to propose that a critical factor affecting the balance between the two may be the amount of shared knowledge between communicators. Specifically, we explore the hypothesis that accumulation of shared knowledge via extended visual communication may promote the development of increasingly schematic yet effective ways of depicting an object, even as these *ad hoc* graphical conventions may be less readily apprehended by others who lack this shared knowledge.

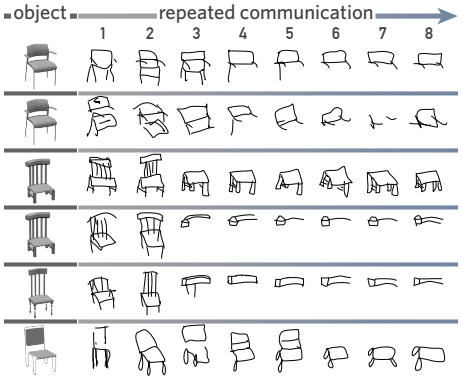


Figure 1: Repeated visual communication depicting the same object.

To investigate this hypothesis, we used an interactive drawing-based reference game in which two participants repeatedly communicated about visual objects. We examined both how their task performance and the drawings they produced changed over time (see Fig. 1). Our approach was inspired by a large literature that has explored how extended interaction influences communicative behavior in several modalities, including language (Clark & Wilkes-Gibbs, 1986; Hawkins, Frank, & Goodman, 2017), gesture (Goldin-Meadow, McNeill, & Singleton, 1996), and drawings (Garrod, Fay, Lee, Oberlander, & MacLeod, 2007; Galantucci, 2005). There are three aspects of the current work that advance our prior understanding: *first*, we include a control set of objects that were not repeatedly drawn but only shown at the beginning and end of the interaction, allowing us to measure the specific contribution of repeated reference vs. general practice effects; *second*, we measure how strongly the visual properties of drawings drive recognition in the absence of interaction history for naive viewers, while equating other task variables; and *third*, we employ recent advances in computer vision to quantitatively characterize changes in the high-level visual properties of drawings across repetitions.

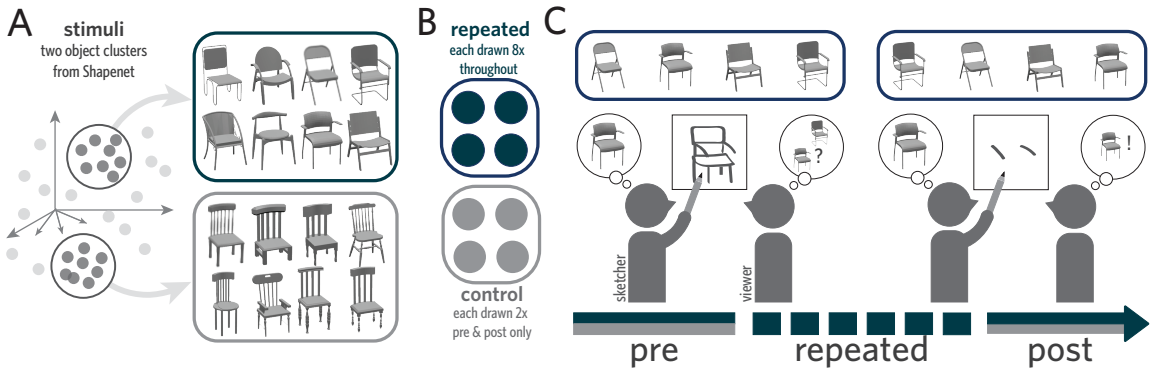


Figure 2: (A) Stimuli from ShapeNet. (B) Each pair of participants was randomly assigned two sets of four objects, each set from one of the two categories. (C) Repeated objects drawn eight times throughout; control objects drawn once at the beginning and end of each interaction.

Part I: How does repeated reference support successful visual communication?

Our first goal was to understand how people learn to communicate about visual objects across repeated visual communication. To accomplish this, we developed a drawing-based reference game for two participants. On each trial, both participants shared a *communicative context*, represented by an array of four objects. One of these objects was privately designated the 'target' to the sketcher. The sketcher's goal was to draw the target so that the viewer could select it from the array as quickly and accurately as possible. We hypothesized that learning would be *object-specific*: that over repeated visual reference to a particular object, participants would discover ways of depicting that object more effectively relative to non-repeated control objects.

Methods: Visual communication experiment

Participants We recruited 138 participants from Amazon Mechanical Turk, who were grouped into 69 pairs (Hawkins, 2015). Within each experimental session, one participant was assigned the sketcher role and the other the viewer role, and these role assignments remained the same throughout the experiment. Data from two pairs were excluded due to unusually low performance (i.e., accuracy $<$ 3 s.d. below the mean). In this and subsequent experiments, participants provided informed consent in accordance with the Stanford IRB.

Stimuli In order to make our task sufficiently challenging, we sought to construct communicative contexts consisting of objects whose members were both geometrically complex and visually similar. To accomplish this, we sampled objects from the ShapeNet (Chang et al., 2015), a database containing a large number of 3D mesh models of real-world objects. We restricted our search to 3096 objects belonging to the `chair` class, which is among the most diverse and abundant in ShapeNet. To identify groups of visually similar chairs, we first extracted high-level visual features from 2D renderings of each object using a deep convolutional neural network (DCNN) architecture, VGG-19 (Simonyan & Zisserman, 2014). This network had been previously trained to recognize objects in photos from the ImageNet database (Deng et al., 2009), containing 1.2 million natural photographs of 1000 different object classes. Trained DCNN models have been shown to predict human perceptual similarity judgments about objects (Kubilius, Bracci, & de Beeck, 2016; Peterson, Abbott, & Griffiths, 2018), as well as neural population responses in visual cortex during object recognition (Yamins et al., 2014; Güçlü & van Gerven, 2015). As such, they provide a principled choice of encoding model for extracting high-level visual information from images. Following previous work that has employed DCNN models to evaluate perceptual similarity (Peterson et al., 2018; Kubilius et al., 2016), for each image we extract a 4096-dimensional feature vector reflecting activations in the second fully-connected layer (i.e., `fc6`) of VGG-19, a higher layer in the network. We then applied dimensionality reduction (PCA) and k-means clustering on these feature vectors, yielding 70 clusters containing between 2 and 80 objects each. Among clusters that contained at least eight objects, we manually identified two visual categories containing eight objects each (Fig. 2A).

Task Procedure On each trial, both participants were shown the same set of four objects in randomized locations. One of the four objects was highlighted on the sketcher's screen to designate it as the target. Sketchers drew using their mouse cursor in black ink on a digital canvas embedded in their web browser (300×300 pixels; pen width = 5px). Each stroke was rendered on the viewer's screen in real time and sketchers could not delete previous strokes. The viewer aimed to click one of the four objects as soon as they were confident of the identity of the target, and participants received immediate feedback: the sketcher learned when and which object the viewer had clicked, and the viewer learned the true identity of the target. Both participants were incentivized to perform both quickly and accurately. They both earned an accuracy bonus for each correct response, and the sketcher was required to complete their drawings in 30 seconds or less. If the viewer responded correctly within this time limit, participants also received a speed bonus inversely proportional to the time taken until the response.

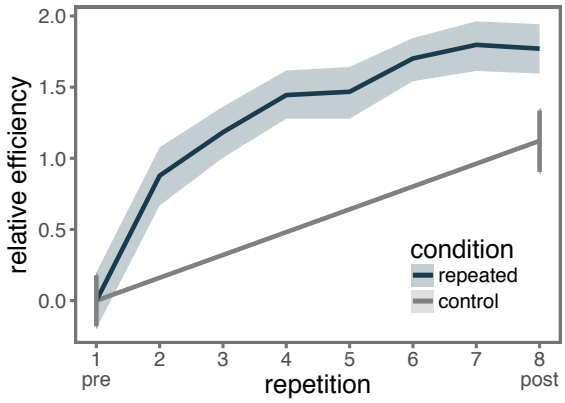


Figure 3: Communication efficiency across repetitions. Efficiency combines both speed and accuracy, and is plotted relative to the first repetition. Error ribbons represent 95% CI.

Design For each pair of participants, two sets of four objects were randomly sampled to serve as communication contexts: one was designated the *repeated* set while the other served as the *control* set (Fig. 2B).[1] The experiment consisted of three phases (Fig. 2C). During the repeated reference phase, there were six repetition blocks of four trials, and each of the four *repeated* objects appeared as the target once in each repetition block. In a pretest phase at the beginning of the experiment and a posttest phase at the end, both repeated and control objects appeared once as targets (in their respective contexts) in a randomly interleaved order.

Results

Because objects were randomly assigned to repeated and control conditions, we expected no differences in task performance in the pretest phase. We found that pairs identified the target at rates well above chance in this phase (75.7% repeated, 76.1% control, chance = 25%), suggesting that they were engaged with the task but not at ceiling performance. We found no difference in accuracy across conditions (mean difference: 0.3%, bootstrapped CI: $[-7\%, 7\%]$).

In order to measure how well pairs learned to communicate throughout the rest of their interaction, we used a measure of communicative efficiency (the *balanced integration score,* Liesefeld & Janczyk, 2018) that takes both accuracy (i.e., proportion of correct viewer responses) and response time (i.e., latency before viewer response) into account. This efficiency score is computed by first z-scoring accuracy and response time across repetitions within an interaction to map values from different interactions to the same scale, and then subtracting the standardized response time from standardized accuracy. It is highest when pairs are both fast and accurate, and lowest when they make more errors and take longer, relative to their own performance on other trials.

[1] In half of the pairs, the four control objects were from the same stimulus cluster as repeated objects; in the other half, they were from different clusters. The rationale for this was to support investigation of between-cluster generalization in future analyses. In current analyses, we collapse across these groups.

To evaluate changes in communicative efficiency, we fit a linear mixed-effects model including random intercepts, slopes, and interactions for each pair of participants. We found a main effect of increasing communicative efficiency for all targets between the *pre* and *post* phases ($b = 1.45$, $t = 14.3$, $p < 0.001$), reflecting general improvements due to task practice. Critically, however, this analysis also revealed a reliable interaction between phase and condition: communicative efficiency improved to a greater extent for repeated objects than control objects ($b = 0.648$, $t = 3.09$, $p = 0.003$; see Fig. 3). Analysis of changes in raw accuracy yielded a similar result: performance on repeated objects improved by 14.5%, while performance on control objects only improved by 7.1%. Together, these data show that there are benefits of repeatedly communicating about an object that accrue specifically to that object, suggesting the formation of object-specific graphical conventions.

Part II: What explains gains in efficiency?

Our visual communication experiment established that pairs of participants coordinate on more efficient and *object-specific* ways of depicting targets. This raises the question: to what extent do these gains in efficiency reflect the accumulation of *interaction-specific* shared knowledge between a sketcher and viewer, as opposed to the combination of task practice and the inherent visual properties of their drawings?

To disentangle the contributions of these different factors, we conducted two control experiments to estimate the how recognizable these drawings were to naive viewers outside the social context in which they were produced. Participants in one control group were shown a sequence of drawings taken from a single interaction, closely matching the experience of viewers in the communication experiment. Participants in a second control group were instead shown a sequence of drawings pieced together from many different interactions, thus disrupting the continuity experienced by viewers paired with a single sketcher. Insofar as interaction-specific shared knowledge contributed to the efficiency gains observed previously, we hypothesized that the second group would not improve as much over the course of the experimental session as the first group would.

Methods: Recognition Control Experiments

Participants We recruited 245 participants via Amazon Mechanical Turk. We excluded data from 22 participants who did not meet our inclusion criterion for accurate and consistent response on attention-check trials (see below).

Task, Design, & Procedure On each trial, participants were presented with a drawing and the same set of four objects that accompanied that drawing in the original visual communication experiment. They also received the same accuracy and speed bonuses as viewers in the communication experiment. To ensure task engagement, we included five identical attention-check trials that appeared once every eight trials. Each attention-check trial presented the same set of

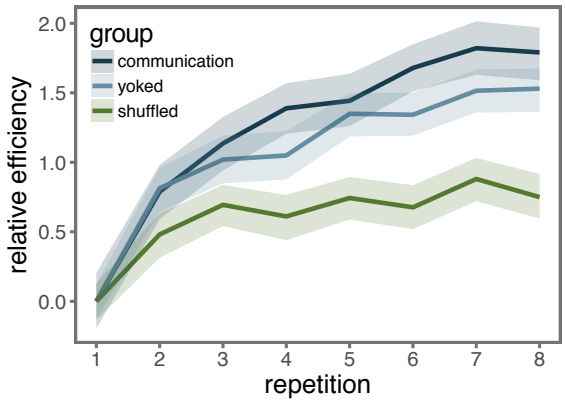


Figure 4: Comparing drawing recognition performance between viewers in communication experiment with those of yoked and shuffled control groups. Error ribbons represent 95% CI.

objects and drawing, which we identified during piloting as the most consistently and accurately recognized by naive participants. Only participants who responded correctly on at least four out of five of these trials were retained in subsequent analyses.

Each participant was randomly assigned to one of two conditions: a *yoked* group and a *shuffled* group. Each yoked participant was matched with a single interaction from the original cohort and viewed 40 drawings in the same sequence the original viewer had. Those in the shuffled group were matched with a random sample of 10 distinct interactions from the original cohort and viewed four drawings from each in turn, which appeared within the same repetition block as they had originally. For example, if a drawing was produced in the fifth repetition block in the original experiment, then it also appeared in the fifth block for shuffled participants.

At the trial level, groups in both conditions thus received exactly the same visual information and performed the task under the same incentives to respond quickly and accurately. At the repetition level, both groups received exactly the same amount of practice recognizing drawings. Thus any differences between these groups are attributable to whether drawings came from the same communicative interaction, which would support the accumulation of interaction-specific experience, or from several different interactions, where such accumulation would be minimal.

Results

Interaction-specific history enhances recognition by third-party observers We compared the yoked and shuffled groups by measuring changes in recognition performance across successive repetitions using the same efficiency metric we previously used. We estimated the magnitude of these changes by fitting a linear mixed-effects model that included group (yoked vs. shuffled), repetition number (i.e., first through eighth), and their interaction, as well as random intercepts and slopes for each participant. While we found a significant increase in recognition performance across both groups ($b = 0.18,\ t = 12.8,\ p < 0.001$), we also found a large and reliable interaction: yoked participants improved to a substantially greater degree than shuffled participants ($b = 0.10,\ t = 4.9,\ p < 0.001$; Fig. 4). Examining accuracy alone yielded similar results: the yoked group improved to a greater degree across the session (yoked: +15.8%, shuffled: +5.6%). Taken together, these results suggest that third-party observers in the yoked condition who viewed drawings from a single interaction were able to take advantage of this continuity to more accurately identify what successive drawings represented. While observers in the shuffled condition still improved over time, being deprived of this interaction continuity made it relatively more difficult to interpret later drawings.

Viewer feedback also contributes to gains in performance Unlike viewers in the interactive visual communication experiment, participants in the yoked condition made their decision based only on the whole drawing and were unable to interrupt or await additional information if they were still uncertain. Sketchers could have used this feedback to modify their drawings on subsequent repetitions. As such, comparing the yoked and original communication groups provides an estimate of the contribution of these viewer feedback channels to gains in performance (Schober & Clark, 1989). In a mixed-effects model with random intercepts, slopes, and interactions for each unique trial sequence, we found a strong main effect of repetition ($b = 0.23,\ t = 12.8,\ p < 0.001$), as well as a weaker but reliable interaction with group membership ($b = -0.05,\ t = -2.2,\ p = 0.032$, Fig. 4), showing that the yoked group improved at a more modest rate than viewers in the original communication experiment had.

To better understand this interaction, we further examined changes in the accuracy and response time components of the efficiency score. We found that while viewers in the communication experiment were more accurate than yoked participants overall (communication: 88%, yoked: 75%), *improvements* in accuracy over the course of the experiment were similar in both groups (communication: +14.5%, yoked: +15.8%). The interaction instead appeared to be driven by differential reductions in response time between the first and final repetitions (communication: 10.9s to 5.84s; yoked: 4.66s to 3.31s). These reductions were smaller in the yoked group, given that these participants did not need to wait for each stroke to appear before making a decision, and thus may have already been closer to floor.

Part III: How do visual features of drawings change over the course of an interaction?

The results so far show that repeated visual communication establishes object-specific, interaction-specific ways of efficiently referring to objects. An intriguing implication is that interacting pairs achieved this by gradually forming *ad hoc* graphical conventions about what was relevant and sufficient to include in a drawing to support rapid identification of the target object. Here we explore this possibility by examining how the drawings themselves changed throughout an interac-

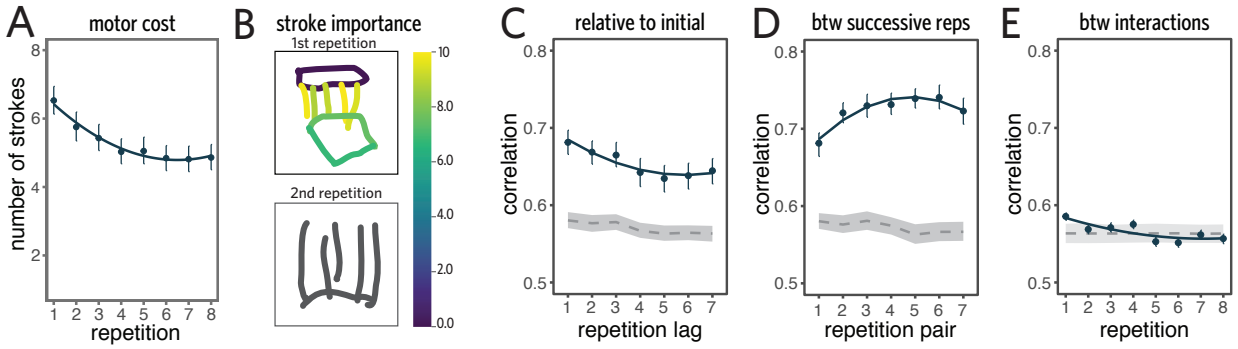


Figure 5: (A) Sketchers use fewer strokes over time. (B) Visualizing importance of individual strokes in successive drawings. (C) Drawings become increasingly dissimilar from initial drawing. (D) Drawings become more consistent from repetition to repetition. (E) The same object is drawn increasingly dissimilarly by different sketchers. Error ribbons represent 95% CI, dotted lines represent permuted baseline.

tion. Concretely, we investigated four aspects that would reflect the increasing contribution of interaction-specific shared knowledge: *first*, decreasing number of strokes used (i.e., reducing motor cost of each drawing); *second*, increasing dissimilarity from the initial drawing produced (i.e., cumulative drift from the starting point); *third*, increasing similarity between successive drawings (i.e., convergence on internally consistent ways of depicting objects within an interaction); *fourth*, increasing dissimilarity between drawings of the same object produced in different interactions (i.e., discovery of multiple viable solutions to the coordination problem).

Measuring visual similarity between drawings

Measuring visual similarity between drawings depends upon a principled approach for encoding their high-level visual properties. Here we capitalize on recent work validating the use of deep convolutional neural network models to encode such perceptual content in drawings (Fan et al., 2018). As when identifying clusters of similar object stimuli, we again used VGG-19 to extract 4096-dimensional feature vector representations for drawings of every object, in every repetition, from every interaction. Using this feature basis, we compute the similarity between any two drawings as the Pearson correlation between their feature vectors (i.e., $s_{ij} = {}^{cov(\vec{r}_i,\vec{r}_j)}/_{\sqrt{var(\vec{r}_i)\cdot var(\vec{r}_j)}}$).

Results

Fewer strokes across repetitions A straightforward explanation for the gains in communication efficiency observed in Part I is that sketchers were able to use fewer strokes per drawing to achieve the same level of viewer recognition accuracy. Indeed, we found that the number of strokes in drawings of repeated objects decreased steadily as a function of repetition in a mixed-effects model ($b = -0.216$, $t = -6.00$; Fig. 5A), suggesting that pairs were increasingly able to rely upon shared knowledge to communicate efficiently. This result raises a question about *which* strokes are preserved across successive repetitions during the formation of graphical conventions. In ongoing work, we are using a lesion method to investigate the "importance" of each stroke within a drawing for explaining similarity to the next repetition's drawing of that object. We re-render the drawing without each stroke and compute the similarity, yielding a heat map across strokes (see Fig. 5B for an example visualization). The more dissimilar the lesioned drawing without a particular stroke is to an intact version of the next repetition's drawing, the more "important" we consider that stroke to be.

Increasing dissimilarity from initial drawing Mirroring the observed reduction in the number of strokes across repetitions, we hypothesized that there was also cumulative change in the visual content of drawings across repetitions. Concretely, we predicted that drawings would become increasingly dissimilar from the initial depiction. We tested this prediction in a mixed-effects regression model including linear and quadratic terms for repetition as well as intercepts for each target and pair. We found a significant decrease in similarity to the initial round across successive repetitions, ($b = -0.62$, $t = -5.59$; Fig. 5C), suggesting that later drawings had moved to a different region of visual feature space. However, since the entire distribution of drawings may have drifted to a different region of the visual feature space for generic reasons (i.e., because they were sparser overall), we conducted a stricter permutation test. We scrambled drawings across pairs but within each repetition and target and re-ran our mixed-effects model. The observed effect fell outside this null distribution ($CI = [-3.53 - 0.88]$, $p < .001$), showing that successive drawings by the same sketcher deviated from their own initial drawing to a greater degree than would be expected due to generic differences between drawings made at different timepoints in an interaction.

Increasing internal consistency within interaction As sketchers modified their drawings across successive repetitions, we additionally hypothesized that they would converge on increasingly consistent ways of depicting each object. To test this prediction, we computed the similarity of successive drawings of the same object made in the same interaction (i.e. repetition k to $k+1$). A mixed-effects model with random intercepts for both object and pair showed that similarity between successive drawings increased substantially

throughout an interaction ($b = 0.53$, $t = 5.03$; Fig. 5). Again, we compared our empirical estimate of the magnitude of this trend to a null distribution of slope t values generated by scrambling drawings across pairs. The observed increase fell outside this null distribution ($CI = [-3.21, -0.60], p < .001$), providing evidence that increasingly consistent ways of drawing each object manifested only for series of drawings produced within the same interaction.

Increasingly different drawings across interactions Our recognition control experiments suggested that the graphical conventions discovered by different pairs were increasingly opaque to outside observers. This effect could arise if early drawings were more strongly constrained by the visual properties of a shared target object, but later drawings diverged as different pairs discovered different equilibria in the space of viable graphical conventions. Under this account, drawings of the same object from different pairs would become increasingly dissimilar from each other across repetitions. We tested this prediction by computing the mean pairwise similarity between drawings of the same object within each repetition index, but produced in different interactions. Specifically, for each object, we considered all interactions in which that object was repeatedly drawn. Then, for each repetition index, we computed the average similarity between drawings of that object. In a mixed-effects regression model including linear and quadratic terms, as well as random slopes and intercepts for object and pair, we found a small but reliable negative effect of repetition on between-interaction drawing similarity ($b = -1.4$, $t = -2.5$; Fig. 5E). We again conducted a permutation test to compare this t value with what would be expected from scrambling sketches across repetitions for each sketcher and target object. We found that the observed slope was highly unlikely under this distribution ($CI = [-0.57, 0.60]$, $p < 0.001$), even if the similarity at each round was not so unlikely.

Discussion

In this paper, we investigated the joint contributions of visual information and social context to determining the meaning of drawings. We observed in an interactive Pictionary-style communication game that pairs of participants discover increasingly sparse yet effective ways of depicting objects over repeated reference. Through a series of control experiments, we demonstrated that these conventionalized representations were both object-specific and interaction-specific: drawings were harder for independent viewers to recognize without sharing the same interaction history. Furthermore, by analyzing the high-level visual features of drawings, we found that they became increasingly consistent within an interaction, but that different pairs discover different equilibria in the space of viable graphical conventions. Taken together, our findings suggest that repeated visual communication promotes the emergence of depictions whose meanings are increasingly determined by interaction history rather than their visual properties alone.

A key experimental design choice was the use of visual objects as the targets of reference, by contrast with the verbal labels or audio clips used in prior work (Galantucci & Garrod, 2011; Fay, Garrod, Roberts, & Swoboda, 2010). As such, communication between the sketcher and viewer was grounded in the same visual information about the appearance of these objects, encouraging the production of more 'iconic' initial drawings that more strongly resembled the target object (Verhoef, Kirby, & de Boer, 2016; Perlman, Dale, & Lupyan, 2015). As their communication became increasingly efficient across repetitions, their drawings became simpler and apparently more 'abstract'. An exciting direction for future work is to develop robust and principled computational measures of the degree of visual correspondence between any drawing and any target object, thereby shedding light on the nature of visual abstraction and iconicity.

A second important design choice was the use of a speed bonus incentivizing participants to complete trials quickly. What role do such incentives play in the formation of graphical conventions? Recent computational models of visual communication have found that both how costly a drawing is to produce (i.e., time/ink) and how informative a drawing is in context are critical for explaining the way people spontaneously adjust the level of detail to include in their drawings in one-shot visual communication tasks (Fan, Hawkins, Wu, & Goodman, 2019). The consequences of this intrinsic preference for less costly drawings may be compounded across repetitions, as the accumulation of interaction history allows people to be equally informative with fewer strokes (Hawkins et al., 2017). The magnitude of these intrinsic costs may vary across individuals, however, and the speed bonus made them explicit.

A major open question raised by our work concerns how people decide what information to preserve or discard across repetitions. One possibility is that successful viewer comprehension is attributed to the most recent strokes produced, leading these to be more strongly preserved. For example, if the viewer was able to correctly identify the target only after the backrest was drawn, the sketcher may continue to selectively draw this part. Another possibility is that sketchers preserve what they judge to be the most diagnostic information about the target, regardless of when the viewer made their response. For example, sketchers may focus on drawing the backrest if it strongly distinguishes the target from distractors in context. Future work should disentangle these possibilities empirically and via development of computational models of visual communication that can learn from task-related feedback, as well as judge which strokes would be most diagnostic.

Visual communication is a powerful vehicle for the cultural transmission of knowledge. Over time, advancing our knowledge of the cognitive mechanisms underlying the formation of graphical conventions may lead to a deeper understanding of the origins of modern symbolic systems for communication and the design of better visual communication tools.

Acknowledgments

Thanks to Mike Frank and Hyo Gweon for helpful discussion. RXDH was supported by the E. K. Potter Stanford Graduate Fellowship and the National Science Foundation Graduate Research Fellowship (DGE-114747). MS was supported by the Masason Foundation Scholarship and the Center for the Study of Language and Information at Stanford.

All code and materials available at:
`https://github.com/cogtoolslab/graphical_conventions`

References

Abell, C. (2009). Canny resemblance. *Philosophical Review*, *118*(2), 183–223.

Chang, A. X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., ... others (2015). Shapenet: An information-rich 3D model repository. *arXiv preprint arXiv:1512.03012*.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, *22*(1), 1–39.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 ieee conference on computer vision and pattern recognition* (pp. 248–255).

Fan, J. E., Hawkins, R. X. D., Wu, M., & Goodman, N. (2019). Pragmatic inference and visual abstraction enable contextual flexibility during visual communication. *arXiv preprint arXiv:1903.04448*.

Fan, J. E., Yamins, D. L. K., & Turk-Browne, N. B. (2018). Common object representations for visual production and recognition. *Cognitive Science*.

Fay, N., Garrod, S., Roberts, L., & Swoboda, N. (2010). The interactive evolution of human communication systems. *Cognitive Science*, *34*(3), 351–386.

Galantucci, B. (2005). An experimental study of the emergence of human communication systems. *Cognitive science*, *29*(5), 737–767.

Galantucci, B., & Garrod, S. (2011). Experimental semiotics: a review. *Frontiers in Human Neuroscience*, *5*, 11.

Garrod, S., Fay, N., Lee, J., Oberlander, J., & MacLeod, T. (2007). Foundations of representation: where might graphical symbol systems come from? *Cognitive Science*, *31*(6), 961–987.

Goldin-Meadow, S., McNeill, D., & Singleton, J. (1996). Silence is liberating: removing the handcuffs on grammatical expression in the manual modality. *Psychological Review*, *103*(1), 34.

Goodman, N. (1976). *Languages of art: An approach to a theory of symbols*. Hackett.

Güçlü, U., & van Gerven, M. A. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, *35*(27), 10005–10014.

Hawkins, R. X. D. (2015). Conducting real-time multiplayer experiments on the web. *Behavior Research Methods*, *47*(4), 966-976.

Hawkins, R. X. D., Frank, M. C., & Goodman, N. D. (2017). Convention-formation in iterated reference games. In *Proc. of the 39th Annual Meeting of the Cognitive Science Society.*

Hoffmann, D., Standish, C., García-Diez, M., Pettitt, P., Milton, J., Zilhão, J., ... others (2018). U-th dating of carbonate crusts reveals neandertal origin of iberian cave art. *Science*, *359*(6378), 912–915.

Kubilius, J., Bracci, S., & de Beeck, H. P. O. (2016). Deep neural networks as a computational model for human shape sensitivity. *PLoS Computational Biology*, *12*(4), e1004896.

Liesefeld, H. R., & Janczyk, M. (2018). Combining speed and accuracy to control for speed-accuracy trade-offs. *Behavior Research Methods*.

Perlman, M., Dale, R., & Lupyan, G. (2015). Iconicity can ground the creation of vocal symbols. *Royal Society open science*, *2*(8), 150152.

Peterson, J. C., Abbott, J. T., & Griffiths, T. L. (2018). Evaluating (and improving) the correspondence between deep neural networks and human representations. *Cognitive Science*, *42*(8), 2648–2669.

Schober, M. F., & Clark, H. H. (1989). Understanding by addressees and overhearers. *Cognitive Psychology*, *21*(2), 211–232.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Verhoef, T., Kirby, S., & de Boer, B. (2016). Iconicity and the emergence of combinatorial structure in language. *Cognitive science*, *40*(8), 1969–1994.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 201403112.

Efficient use of ambiguity in an early writing system: Evidence from Sumerian cuneiform

Noah Hermalin[1] **(nmhermalin@berkeley.edu)**
Terry Regier[1,2] **(terry.regier@berkeley.edu)**

[1]Department of Linguistics and [2]Cognitive Science Program
University of California, Berkeley, CA 94720 USA

Abstract

Ambiguity has often been viewed as a hindrance to communication. In contrast, Piantadosi et al. (2012) argued that ambiguity may be useful in that it allows communication to be efficient, and they found support for this argument in the spoken forms of modern English, Dutch, and German. The historical origins of this phenomenon cannot be probed in the case of spoken language, but they can for written language, as it leaves an enduring trace. Here, we explore ambiguity and efficiency in one of the earliest known written forms of language: Sumerian cuneiform. Sumerian cuneiform exhibits extensive ambiguity, and for that reason it has been considered to be poorly suited for communication. We find, however, that ambiguity in Sumerian cuneiform supports efficient communication, mirroring the earlier findings for spoken English, Dutch, and German. Thus, the early stages of human writing exhibit evidence suggesting pressure for communicative efficiency.

Keywords: efficient communication; ambiguity; writing systems; cuneiform; information theory

Introduction

Ambiguity in language is often considered to be communicatively disadvantageous, because it can make a speaker's intention unclear to a listener. However, it has been argued (Zipf, 1949) that a certain amount of ambiguity in language is inevitable given the competing needs of speakers and listeners. Piantadosi, Tily, and Gibson (2012) pursued this idea further, and argued that ambiguity may be useful in that it can support efficient communication. They showed empirically that patterns of ambiguity in the spoken forms of modern English, Dutch, and German are consistent with pressure for efficiency in communication.

Given this finding, it is natural to wonder about the historical origins of this phenomenon. How quickly do languages come to exhibit efficient use of ambiguity? Was this phenomenon present near the beginning of language use? Such questions are unanswerable for the spoken form of language, which leaves no lasting trace — but they can be addressed with respect to written language, which does leave such a trace.

Sumerian cuneiform is one of the earliest known writing systems, and is one of the four 'pristine' writing systems of the world, meaning that its origins are not traceable to borrowing or influence from any previously existing writing system (Woods, 2015a). It is also known to be highly ambiguous, such that a given character often has numerous distinct semantic and/or phonological values (Cooper, 1996). Additionally, the distribution of meanings across forms in written Sumerian was not simply a straightforward reflection of spoken Sumerian; this means that any finding of efficiency with respect to the writing system cannot be dismissed as entirely derivative of the corresponding spoken language. Finally, Sumerian is unrelated to the languages studied earlier by Piantadosi et al., which are closely related to each other. For these reasons, Sumerian cuneiform suggests itself as a natural case study for probing the historical origins of the efficient use of ambiguity, in the accessible case of written language.

In what follows, we first provide a brief introduction to Sumerian cuneiform, and its relevance to the question of ambiguity and efficiency. We then restate the argument and results of Piantadosi et al. on modern spoken languages. Then, in three studies, we apply the logic and methods of Piantadosi et al. to the problem of assessing efficiency in Sumerian cuneiform. We find that ambiguity in Sumerian cuneiform bears the same signatures of efficiency as were found in modern spoken languages. We conclude that pressure for efficient communication may have been present near the earliest stages of human writing, and we discuss the implications of this conclusion.

Sumerian cuneiform

Cuneiform writing developed in southern Mesopotamia throughout the 4th millennium BC; first used for linguistic writing by the 31st century, the system survived roughly three thousand years, over which it was adapted into various languages of the Middle East (Veldhuis, 2012). The first language for which cuneiform was used was most likely Sumerian (Veldhuis, 2012), an agglutinative language with mild nominal morphology (case-marking suffixes) and rich verbal morphology, including a plethora of tense-aspect-mood and agreement affixes (Michalowski, 2004).

Cuneiform tablets compartmentalized text into columns, which were further divided into lines/cells, somewhat similar in layout to a modern-day spreadsheet; smaller items would only have one column (see Figure 1 for an example). The amount of information contained within a cell of a text had some degree of variation, but was at least at the level of a word and typically at the level of a phrase. Earlier scribal practice was not always concerned with preserving a consistent linear order of characters within a cell. By c. 2400 BC, however, scribes adhered to fairly strict and consistent linearity in spellings (Michalowski, 2004).

Figure 1: A sample text in Sumerian cuneiform. Since the text is small, it only has one column, which is divided into seven lines. Image from the Cuneiform Digital Library Initiative (2016), CDLI #102525. Image reprinted with permission of Robert K. Englund.

Spelling	Transliteration	Meaning
	paŋ	'breathe'
	asag	'demon'
	pa	'branch'
	ugula	'overseer'
	sag	'beat'
	rig	'boil down'
	rig	'donate'
	ensi	'governor/ruler'
	maškim	'administrator'
	*munsub*	'shepherd$_1$'
	*sipad*	'shepherd$_2$'

Table 1: Non-exhaustive list of words that contain the character in their spelling.

Written Sumerian was primarily logographic: the level of linguistic representation for a given graphical unit would usually be the morpheme (or some sub-morphemic, non-phonemic unit of information), although it also made use of phonography to some degree, with characters sometimes mapping more directly to (strings of) sounds, usually at the level of the syllable (Civil, 1973). A major feature of the system was its extensive use of ambiguity: any given character could have numerous distinct semantic and/or phonological values (Cooper, 1996). A non-exhaustive list of words containing the character can be found in Table 1; this list serves as an example of how a single character may occur in the spellings of words which do not all share semantic or phonological information. The table also demonstrates the lack of strict isomorphism between written form and corresponding spoken form, either in terms of phonemes, syllables, or morphemes. This point is important for our purposes because it means that if written Sumerian bears signs of efficiency, that efficiency cannot have been entirely inherited from spoken Sumerian.

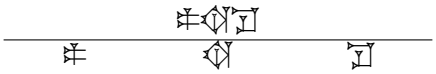
Two open questions concerning efficiency and ambiguity emerge from this overview. First, and most centrally: is the ambiguity of Sumerian cuneiform communicatively harmful, as might be expected given its extensiveness — or is it instead consistent with pressure for efficiency in communication? Second: is the shift to greater linearity in writing attributable to pressure for efficiency? We pursue these questions below.

The argument of Piantadosi et al. (2012)

To address these questions, we draw on the logic and methods of an earlier study that focused on modern spoken languages. Piantadosi et al. (2012) argued that "ambiguity is a functional property of language that allows for greater communicative efficiency" (p. 280). Their argument coheres naturally with a classic functionalist view that seeks to explain language structure and use in terms of efficient communication, and a growing body of recent research that pursues that idea with respect to various aspects of language (e.g. Aylett & Turk, 2004; Ferrer i Cancho & Solé, 2003; Piantadosi, Tily, & Gibson, 2011; Fedzechkina, Jaeger, & Newport, 2012; Kirby, Tamariz, Cornish, & Smith, 2015; Kemp, Xu, & Regier, 2018).

Piantadosi et al. (2012) pursued this argument as follows. First, they argued that context has the potential to resolve ambiguity. The communicative problem posed by ambiguity is that of the listener's (or reader's) uncertainty about the meaning of a given form, and they engaged this problem in information-theoretic terms, casting uncertainty as entropy. They noted that if context is informative about meaning, context will necessarily reduce uncertainty (entropy) about meaning. This means that context has the potential to alleviate the problem posed by ambiguity: a form that may be highly ambiguous in isolation may be much clearer when considered in context. A central assumption of their paper is that context is in fact informative about meaning, and therefore does help to disambiguate.

Piantadosi et al. then pursued the hypothesis that ambiguity in language is deployed in a manner that increases efficiency. The core idea is that if ambiguity is resolved by context, forms are free to take on multiple meanings — and the efficient way to do this would be to preferentially re-use forms that are low-cost, so as to minimize overall cost, or effort (Zipf, 1949). Forms may be low-cost in various ways: they may be short or otherwise simple; they may be frequent and therefore processed more quickly, and so on. Their paper considered several measures of form cost, and asked to what extent each predicts ambiguity of form. Specifically, using data on the spoken forms of German, Dutch, and English, they conducted quasi-Poisson regressions to establish the relationship between various count measures of form ambiguity and three properties of form cost: length, frequency (as negative log probability), and phonotactic surprisal. They found that in general, greater ambiguity was predicted by lower form cost (with the possible exception of phonotactic surprisal). Thus, shorter and more frequent forms were more ambiguous in



(RULER, 1) (RULER, 2) (RULER, 3)

Table 2: Example morpheme, meaning 'ruler'. The top row shows how this morpheme would be spelled in characters in the original text. The bottom two rows show the characters that appear in the spelling of this morpheme, each paired with the value of that character with respect to this morpheme, as defined in Equation 1.

German, Dutch, and English — consistent with the expectation that low-cost forms are preferentially re-used, as predicted by pressure for efficiency.

The present studies

We applied an analogous line of investigation to the question of ambiguity in Sumerian cuneiform.[1] Ambiguity arises when a form has more than one value, or symbolic function. Thus, to explore ambiguity in Sumerian cuneiform, we need to specify the relevant unit of form, and the corresponding values. It is natural to take the character as the relevant unit of form in Sumerian cuneiform, as it is characters that are often considered to be ambiguous. And given that characters are not defined either purely semantically or purely phonologically, but are used to specify morphemes, it is natural to define the values of a character in terms of that character's role in identifying morphemes, i.e. the character's role in spelling morphemes. A morpheme can have more than one spelling, so we first define the spellings $S(m)$ of a morpheme m to be the set of character strings that spell out that morpheme in Sumerian cuneiform. We then define the values *V(x)* of a character x as:

$$V(x) = \{\ (m,i)\ \mid\ x \text{ is the } i^{\text{th}} \text{ character in } s \in S(m)\ \} \quad (1)$$

That is, the values of character x are the set of all (morpheme, index) pairs (m,i) such that x is the i^{th} character in one of the spellings s of morpheme m. For example, the values of the character 𒁹 include the pairs (RULER, 1), (BRANCH, 1), and (DEMON, 2), among others. Table 2 illustrates a spelling of a specific morpheme, and the determination of character values from that spelling.[2]

Data

The data we used were from ORACC, the Open Richly Annotated Cuneiform Corpus (Tinney & Robson, 2014), an open-access corpus of cuneiform texts which is, to the best of our knowledge, the largest open-access corpus for Sumerian texts that has POS tagging and morphological annotation. Specifically, we used the texts in the Ur III Administrative Documents corpus within ORACC; this corpus is roughly 5.5 million cuneiform characters in length, and it consists of various administrative and transactional documents from the Ur III period (c. 2112-2004 BC). This corpus was chosen because it is the largest single-genre morphologically annotated corpus of third millennium Sumerian texts.

Substantial parts of the corpus had to be discarded. We omitted tokens that were damaged or for which the reading was unknown. In addition, most proper nouns had to be omitted.[3] The resulting cleaned data had roughly 3.3 million character tokens. We refer to this cleaned corpus as the 'dataset'.

Overview of the present studies

We conducted three studies to test whether ambiguity in Sumerian cuneiform is consistent with pressure for efficient communication. Piantadosi et al. (2012) assumed that much ambiguity could be resolved by context; we wished to test this question directly, so Study 1 asks to what extent context resolves ambiguity in Sumerian cuneiform. Study 2 asks whether context disambiguates more effectively in Sumerian cuneiform than it does in a number of plausible hypothetical variants of it; in so doing, this study explores whether increasing linearity in Sumerian writing may have resulted from pressure for efficiency. Finally, Study 3 applies the analyses of Piantadosi et al. (2012) to Sumerian cuneiform, to determine whether the signatures of efficiency they found in modern spoken languages are also found in cuneiform.

Study 1: Does context disambiguate?

To what extent does context resolve ambiguity in Sumerian cuneiform? We considered a simple version of this general question. We first determined the uncertainty concerning which value a character has when the reader knows only the current character (unigram condition). We then compared this to the uncertainty when the reader knows not just the current character but also the preceding character (bigram condition).

We took uncertainty concerning character values to be the conditional entropy of values V conditioned on context C:

$$H(V|C) = -\sum_{c \in C} P(c) \sum_{v \in V} P(v|c) \log_2 P(v|c) \quad (2)$$

Lower conditional entropy denotes greater certainty concerning character value.

We calculated $H(V|C)$ over the entire dataset, once taking C to be the current character alone (unigram), and once again taking C to be the current and preceding characters together (bigram). The results are shown in the top two lines of Table 3. Conditional entropy in the bigram condition is much lower than in the unigram condition. This demonstrates not only that context disambiguates, but also that just a single

[1] We believe we are the first to treat Sumerian in this way. However Civil (1973) informally explored the possibility of examining Sumerian cuneiform through the lens of information theory.

[2] We also ran all of the analyses using an alternate definition of a character's values: $V(x) = \{\ m\ \mid\ x \text{ is present in } s \in S(m)\ \}$. By this definition, a character x's values are simply the set of morphemes that contain x anywhere in any of their spellings. The results using this definition of character values were qualitatively the same as the results reported here.

[3] Proper nouns had no morphological annotation. Among other problems, this meant that inflectional morphology was not automatically separable from the rest of the word for proper nouns, as it was for other words in the corpus.

Table 3: Conditional entropy $H(V|C)$ of character values V given one character (unigram) vs. two characters (bigram) of text C, on attested and hypothetical data. Study 1: One added character of context results in a sharp decrease in uncertainty in attested data. Study 2: context disambiguates more effectively in attested Sumerian cuneiform than it does in some hypothetical variants of it. Value for WLSS is the average $\pm$ 1 SD, over 500 systems.

Study	Condition	$H(V\|C)$
1	Unigram, attested data	1.5281
1	Bigram, attested data	0.4584
2	Bigram, BWS	0.4719
2	Bigram, WLSS	0.9796 ($\pm$ 0.0004)

additional preceding character of context suffices to substantially reduce uncertainty. Since much more context than this would be available to readers, it is reasonable to expect that a competent reader of Sumerian would be able to infer with high certainty which value a given character was intended to have, in context. We conclude from this finding that context does effectively disambiguate in Sumerian cuneiform.

Study 2: Comparison with hypothetical systems

Given that context disambiguates in Sumerian cuneiform, we ask the follow-up question of whether plausible hypothetical variants of the system exhibit better, worse, or comparable results. Study 2 tested whether the consistency of spellings and strict linearity of Sumerian cuneiform demonstrate advantages over hypothetical competitors with regards to certainty of decoding character values in context.

We considered two hypothetical variants of Sumerian cuneiform. The first variant is 'backwards Sumerian' (BWS): this is a hypothetical variant of Sumerian in which the entire corpus is spelled backwards. Effectively this means that when considering a character in context, we take as context what would have been the following character in actual Sumerian, rather than the preceding character as in Study 1. The other hypothetical variant is 'within-line shuffled Sumerian' (WLSS): this is a system that is derived from our Sumerian cuneiform dataset by randomly shuffling the order of characters within a line. In this case, a neighboring character taken as context could be any other character within the same line in the original dataset. The latter hypothetical variant is motivated to some extent by actual scribal practices in earlier periods, in which characters were not always arranged in linear order. It is known that written Sumerian shifted towards more consistent linearity over time (Michalowski, 2004), and these hypothetical variants allow us to test the hypothesis that the greater linearity that we see in Ur III written Sumerian (the period of our dataset) may have aided disambiguation.

We first calculated $H(V|C)$ over the BWS datset. We then generated 500 WLSS datasets by randomly reordering characters and their respective values within each line, and calculated $H(V|C)$ over each resulting WLSS dataset. We considered only the bigram condition (in which C is the current character together with an immediately preceding character), because the unigram condition would yield identical results in the attested and hypothetical systems.

The results are shown in Table 3. Bigram conditional entropy is very slightly higher for BWS than it is for the attested system; thus, following context may serve as a marginally weaker disambiguator than preceding context, but the difference is small. Bigram conditional entropy is substantially higher for the WLSS systems than it is for the attested system, demonstrating that consistent linearity of spelling does confer an advantage on an ambiguous, logographic system such as Ur III written Sumerian, at least with respect to determining a given character's value based on immediately neighboring context. These results elaborate those of Study 1, and show that context disambiguates more effectively in Sumerian cuneiform than it does in at least some hypothetical variants of that system.

Study 3: Is ambiguity used efficiently?

We have seen that the ambiguity of written Sumerian is much reduced by contextual information, and that this is more true of actual Sumerian than it is of some possible variants of it. This sets the stage for a question directly parallel to that posed by Piantadosi et al.: given that context disambiguates, do languages use ambiguity efficiently, by reusing low-cost (simple, frequent) forms for a large number of meanings, thereby reducing system-wide cognitive costs?

We addressed this question in a manner that mirrors that of Piantadosi et al.: by asking whether the number of values associated with a specific character was predicted by the character's frequency of occurrence in the dataset, and by its simplicity.[4] Our measure of complexity (the opposite of simplicity) for a cuneiform character was stroke count: the number of strokes or wedges required to produce the canonical form of the character. For example, the character 𒐼 has 3 strokes. Stroke counts were coded manually by the first author based on forms in the Electronic Pennsylvania Sumerian Dictionary Project (ePSD; Tinney, 2009), an open-access online dictionary. Following Piantadosi et al., we transformed character frequency to negative log (unigram) probability, using add-one smoothing so that no character had frequency zero.

Figure 2 plots the number of values a character has (its character valence, $|V(x)|$, which is the size of the set $V(x)$), as a function of negative log probability based on unigram frequency, and as a function of stroke count. In both cases it appears qualitatively that lower-cost (more frequent, simpler) characters tend to have more values, consistent with pressure for efficiency.

To probe this pattern quantitatively, we conducted a quasi-Poisson regression to predict the number of values $|V(x)|$ associated with each character x, from that character's negative log probability and stroke count. We standardized the two pre-

[4]The third predictor considered by Piantadosi et al., phonotactic surprisal, is not applicable to Sumerian cuneiform.

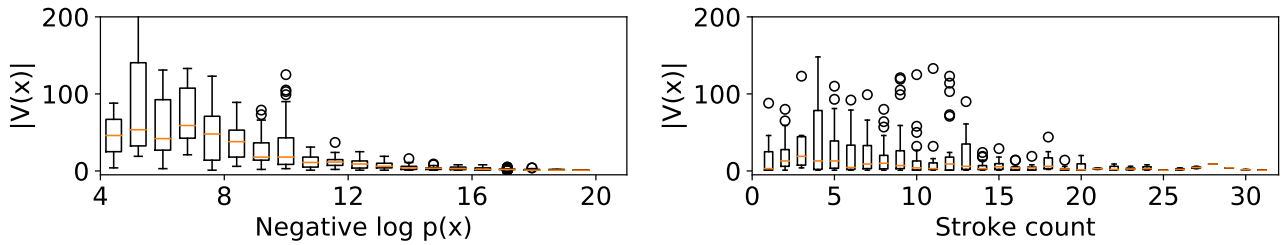


Figure 2: Box plots showing character valence $|V(x)|$ as a function of (left panel) negative log probability based on character frequency, and (right panel) character stroke count. Boxes extend from lower to upper quartiles; orange lines denote median; whiskers extend to 1.5 times the interquartile range beyond the lower and upper quartiles; all data points not in this range are treated as outliers and shown as empty circles. Both panels suggest that low-cost forms are preferentially re-used: more frequent (lower $-\log p(x)$) characters tend to have more values, and simpler (lower stroke count) characters tend to have more values.

dictor variables: for each variable, we subtracted the mean of that variable and divided by one standard deviation. This regression revealed significant effects both of negative log probability and of stroke count. Higher negative log probability (lower frequency) was negatively associated with number of values ($\beta = -1.249, t = -19.792, p < 0.001$), meaning that higher frequency characters were associated with more values. To make this outcome concrete, consider that the most frequent 180 characters, which make up only 27% of the total number of character types in the dataset, bear 66% of all values. Thus, a reader only needs intimate familiarity with a modest number of characters in order to be fairly literate. Higher stroke count was also negatively associated with number of values ($\beta = -0.127, t = -2.072, p < 0.05$), meaning that simpler characters (those with fewer strokes) were associated with more values.

Thus, characters with more values tend to be both more frequent and graphically simpler, as predicted by the hypothesis of efficiency: Sumerian cuneiform exhibits preferential re-use of low-cost material.

Discussion

The traceable origins and early years of written language offer a unique window into the role that pressure for efficient communication can play in shaping linguistic systems. For this reason, the present study has explored efficiency in one of the earliest known writing systems: Sumerian cuneiform, the written form of the Sumerian language.

We have seen that written Sumerian bears signs that are consistent with the hypothesis of pressure for efficient communication. Despite the high degree of ambiguity in written Sumerian, we have seen that a reader would only need a small amount of additional context to be able to decode a character's value with high certainty (Study 1). We have also seen that a comparison with hypothetical alternate systems which deviate from canonical linearity suggests that the system may have gravitated towards a more consistent linearity of spelling in a way that allowed for increased certainty of decoding (Study 2). Finally, we have seen that since context serves to reliably disambiguate character values, the system was able to use a single given form for several different values without sacrificing system informativeness — and that it appears to have done so in an efficient manner, preferentially re-using low-cost forms (Study 3). Taken as a whole, this evidence shows that written Sumerian was not an inefficient system.

Several general implications can be drawn from this observation. One of these concerns efficiency in writing systems generally. While factors such as medium (e.g. Woods, 2015b) and societal pressures (e.g. Veldhuis, 2012) are undoubtedly relevant to the development of a written language, our results demonstrate that pressures of communicative efficiency have acted on written systems since the earlier days of writing itself. Despite the relative disconnect between written Sumerian and its corresponding spoken language in terms of how values are distributed across contrastive units, the same signature of efficiency that Piantadosi et al. (2012) observed in three spoken languages in is also found in Ur III written Sumerian. This suggests that pressure for efficient communication is not unique to spoken or signed language, but is present in written language as well — critically, even when the written language does not closely mirror a corresponding spoken language. Thus, communicative efficiency may be viewed as a general principle of linguistic communication independent of medium or modality.

Another potential implication concerns the time course of the presumed cultural evolutionary process that produces efficiency in linguistic systems. The fact that our results were obtained in a linguistic system as young as 1000 years old suggests that these pressures may act upon a system from its inception and guide it toward greater efficiency within a comparatively short period of time. Since our analyses do not include actual data from periods earlier than Ur III we cannot be completely sure that earlier periods would have been less efficient. However, the fact that our hypothetical shuffled system performed poorly relative to the Ur III corpus is at least suggestive that earlier texts, which were analogously less consistent with their linearity, may not have evolved the specific communicatively useful features we have documented for Ur III Sumerian. Settling this question more definitively would require a thorough comparison of efficiency across ear-

lier time periods, tracking the progression of written Sumerian toward the system we have investigated.

In addition to a direct comparison of written Sumerian across earlier timer periods, future work on this topic would benefit from a more thorough consideration of the psycholinguistic evidence regarding recognition, decoding, and storage of graphical units. While we considered stroke count as a measure of visual complexity (which can be detrimental towards character recognition and processing, especially at lower frequencies; see e.g. Tamaoka & Kiyama, 2013), we did not consider other factors such as visual similarity between characters. Finally, future work could usefully consider the consequences of using the same (or similar) characters for phonologically or semantically related morphemes.

Firmer, broader, and more detailed conclusions will have to await the outcome of such possible future research. For now, however, we can conclude on the basis of the evidence we have seen here that one of the earliest known writing systems exhibits patterns of ambiguity that are consistent with pressure for efficient communication.

Acknowledgments

We thank Robert K. Englund for giving us permission to reprint the image shown in Figure 1. This study was supported in part by the Defense Threat Reduction Agency; the content of the study does not necessarily reflect the position or policy of the U.S. government, and no official endorsement should be inferred.

References

Aylett, M., & Turk, A. (2004). The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, *47*, 31–56.

Civil, M. (1973). The Sumerian writing system: Some problems. *Orientalia*, *42*, 21–34.

Cooper, J. S. (1996). Sumerian and Akkadian. In P. T. Daniels & W. Bright (Eds.), *The world's writing systems* (pp. 37–72). New York: Oxford University Press.

Cuneiform Digital Library Initiative. (2016). Retrieved from `http://cdli.ucla.edu`

Fedzechkina, M., Jaeger, T. F., & Newport, E. L. (2012). Language learners restructure their input to facilitate efficient communication. *Proceedings of the National Academy of Sciences*, *109*, 17897-17902.

Ferrer i Cancho, R., & Solé, R. V. (2003). Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences*, *100*, 788–791.

Kemp, C., Xu, Y., & Regier, T. (2018). Semantic typology and efficient communication. *Annual Review of Linguistics*, *4*, 109–128.

Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, *141*, 87-102.

Michalowski, P. (2004). Sumerian. In R. D. Woodard (Ed.), *The Cambridge encyclopedia of the world's ancient languages* (pp. 19–59). Cambridge: Cambridge University Press.

Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, *108*, 3526-3529.

Piantadosi, S. T., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, *122*, 280–291.

Tamaoka, K., & Kiyama, S. (2013). The effects of visual complexity for Japanese kanji processing with high and low frequencies. *Reading and Writing*, *26*, 205–223.

Tinney, S. (2009). *Electronic Pennsylvania Sumerian Dictionary project.* Retrieved from `http://psd.museum.upenn.edu`

Tinney, S., & Robson, E. (2014). *ORACC: The Open Richly Annotated Cuneiform Corpus.* Retrieved from `http://oracc.museum.upenn.edu`

Veldhuis, N. (2012). Cuneiform: Changes and developments. In S. D. Houston (Ed.), *The shape of script: How and why writing systems change* (pp. 3–23). Santa Fe, NM: School of Advanced Research Press.

Woods, C. (2015a). The earliest Mesopotamian writing. In C. Woods, G. Emberling, & E. Teeter (Eds.), *Visible language: Inventions of writing in the ancient Middle East and beyond.* (pp. 33–50). Chicago: The Oriental Institute of the University of Chicago.

Woods, C. (2015b). Visible language: The earliest writing systems. In C. Woods, G. Emberling, & E. Teeter (Eds.), *Visible language: Inventions of writing in the ancient Middle East and beyond.* (pp. 15–25). Chicago: The Oriental Institute of the University of Chicago.

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Reading, MA: Addison-Wesley.

Productivity depends on communicative intention and accessibility, not thresholds

Alexia Hernandez (alexiamh@princeton.edu)
Department of Psychology, Princeton University, Princeton NJ 08544
Sammy Floyd (sfloyd@princeton.edu)
Department of Psychology, Princeton University, Princeton NJ 08544
Adele E. Goldberg (adele@princeton.edu)
Department of Psychology, Princeton University, Princeton NJ 08544

Abstract

When do children extend a construction ("rule") productively? A recent *Threshold* proposal claims that a construction is productive if and only if it has been witnessed applying to a sufficient proportion of cases and sufficiently few exceptions. An alternative proposal, *Communicate and Access* (C&A), argues that children extend a construction productively because they wish to express an intended message and are unable to access a "better" (appropriate and more conventional) way to do so. Accessibility, in turn, is negatively affected by interference from competing alternatives. In a preregistered experiment, 32 4-6-year-old children were provided with exposure to 2 mini-artificial languages for which the two proposals make opposite predictions. Results support the C&A proposal: children were more productive after witnessing 3 rule-following cases than after 5, due to differences in interference. We conclude that productivity is encouraged by a desire to communicate a message and is constrained by accessibility and interference.

Keywords: productivity, communication, accessibility, Tolerance Principle, Sufficiency Principle

Introduction

When children learn a new noun, *wug*, they are quite adept at producing its plural, *wugs* (Berko, 1958). On the other hand, the *-th* nominalizing suffix (*warmth, width)* is not generally added to new cases *(?coldth; ?oldth*) outside the domain of ordinal numbers (*gazilionth*). A recent *Threshold* proposal has attempted to predict when rules "go" productive and when they do not (Yang, 2016). In particular, in order for a rule to be productive, a Tolerance Principle offers a ceiling on the number of witnessed exceptions and a Sufficiency Principle suggests a floor on the number of cases witnessed following the rule. The required calculations are based on the following 3 numbers:

1) # of cases which potentially follow a rule: N
2) # of witnessed exceptions to a rule: e
3) # of witnessed rule-following cases: *M*

Specifically, the upper bound on exceptions and lower bound on rule-following cases have been proposed according to the thresholds in (1) and (2) (Yang 2016):

(1) Tolerance Principle (TP): $e \leqslant \mathrm{N}/\ln\mathrm{N}$
(2) Sufficiency Principle (SP): $\mathrm{M} \geq \mathrm{N} - \mathrm{N}/\ln\mathrm{N}$

For instance, in a domain of size 9, for a rule to be used productively, the minimum number of cases that must be witnessed following a rule is 5 (Sufficiency Principle) and the maximum number of exceptional cases is 4 (Tolerance Principle) (Table 1; Yang 2016; Schuler, Yang, & Newport 2016).

Table 1: The Threshold numbers predicted by Sufficiency and Tolerance Principles (Yang, 2016; SYN '16).

Domain Size (N)	Minimum # of rule-following cases (M): N- N/lnN	Maximum # of exceptions (e): N/ln N
9	**5**	**4**

A prior study (Schuler, Yang, & Newport, 2016, hereafter SYN '16), aimed to test the predictions in Table 1, but as explained below, the results are open to a different interpretation. The alternative proposal, which we refer to as Communicate and Access (C&A), takes as its starting point the idea that learners aim to convey their messages while obeying the conventions of the language as best they can (Goldberg, 2019). In order to *be able to* use a new language to express an intended message in an appropriate way, children need to be able to *access* the appropriate form. Accessibility is positively affected by the availability of a target form (Bybee, 2010) and is negatively affected by interference from contextually relevant competitors (Bates & MacWhinney 1987; Montag et al. 2017). We report new data involving two new experimental conditions that unconfound the predictions of the two proposals.

SYN '16 aimed to test the predictions in Table 1 by exposing 5-8-year-old children to a rule that could potentially apply to 9 cases in one of two conditions. In a **5R/1-1-1-1E** condition, the rule applied to 5 cases and 4 other cases were witnessed being exceptional, with each exception being exceptional in its own way. In this case, the domain size (N) was 9, the number of cases witnessed following the rule (M) was 5, and the number of exceptions (e) was 4. Because each exceptional case was unique, we represent the 4 exceptions here as 1-1-1-1. This 5R/1-1-1-1E condition satisfied both the Tolerance and Sufficiency principles and, as predicted by SYN '16, children treated the rule as fully productive. In a **3R/1-1-1-1-1-1E** condition, children saw a rule applied to 3 cases and 6 other cases were witnessed being exceptional. Here the Sufficiency Principle was violated (at least 5 rule-following cases should be required for productivity), and there were more exceptions than allowed by the Tolerance Principle. As predicted by SYN '16, children did *not* extend

the rule to new cases in this 3R/1-1-1-1-1-1-1E condition (see Figure 1).

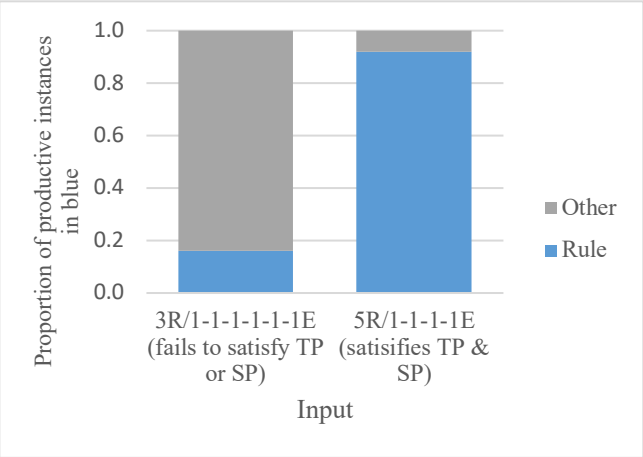


Figure 1: Data reported by SYN (2016): Proportion of productive rule-following (blue) and Other (grey) responses. Children exposed to a rule with 3 rule-following and 6 unique exceptions (left) or 5 rule-following and 4 unique exceptions (right).

Critically, the pattern of results reported by SYN '16 is equally consistent with the Communicate & Access proposal. That is, children extend a construction productively when they wish to express an intended message and are unable to access a "better" (appropriate and more conventional) way to do it. From this perspective, productivity is the effect of producing a "good enough" option when no conventional form exists, or when none is sufficiently accessible at the moment of speaking. Accessibility, in turn, is negatively affected by interference from competing alternatives, which themselves may be more or less accessible (Harmon & Kapatsinski, 2017; Macdonald, 2013; Montag, Matsuki, Kim, & Macdonald, 2017).

If we compare the 5R/1-1-1-1E and 3R/1-1-1-1-1-1-1E "rules" which children were exposed to in SYN '16, the C&A proposal likewise predicts that the 5R/1-1-1-1E rule should be more productive than the 3R/1-1-1-1-1-1-1E rule, but for different reasons than the Threshold proposal suggests.

Instead of viewing language as requiring abstract rules that are subject to numerical thresholds which render the rules either categorically productive or unproductive, the C&A approach predicts that learners record imperfect (lossy) memory traces that relate linguistic words and phrases to their meanings in context. Therefore, in what follows we refer to emergent generalizations as *constructions* instead of rules in describing the C&A perspective. Other things being equal, a construction is more accessible in memory after being witnessed with a greater variety of distinct cases because variability increases accessibility within the range of witnessed variability, and decreases it outside the range of witnessed exemplars (Tenenbaum & Griffiths, 2001; Suttle & Goldberg, 2011). This follows from the fact that memory is associative and content-addressable. The fact that memory is associative entails that new memories are integrated with existing memories; the fact that memory is content-addressable means that existing clusters of memories are activated to the extent that they are relevantly similar for the purpose of task demands.

Conversely, accessibility is negatively impacted by interference from competing constructions, with interference increasing as the accessibility of the competing constructions increases: witnessing 6 exceptional alternative cases *interferes* with a construction more than witnessing only 4 exceptional cases. Since other things were held constant in SYN '16, the availability of the construction was higher, and interference was lower in the 5R/1-1-1-1E condition relative to the 3R/1-1-1-1-1-1-1E condition. Therefore, the C&A proposal concurs that the 5R/1-1-1-1E condition should be more productive.

To summarize, the results reported by SYN '16 cannot distinguish between the proposal based on thresholds as determined by Tolerance and Sufficiency Principles, on the one hand, and the Communicate and Access proposal, on the other (Table 2).

Table 2: Convergent predictions are made by Threshold and C&A proposals for the productivity of an unconditioned "rule" with domain size of 9 in conditions tested by SYN '16 on 5-8-year-olds.

M vs. *e*, M = #Rule-following cases *e* = exceptional cases		**(Shared) Predictions and Results (SYN '16):**
3R/ 1-1-1-1-1-1-1E	Threshold: Neither TP nor SP are satisfied	No systematic productivity
	C&A: Tentative constructional generalization competes with many alternatives: no clear winner emerges	No systematic productivity
5R/ 1-1-1-1E	Threshold: TP and SP are satisfied	Productivity
	C&A: Constructional generalization is more accessible than any alternative	Productivity

In order to compare the two proposals directly, we report a new experiment for which they make *opposing* predictions. Specifically, we exposed a group of 4-6-year-old children to 2 new mini-artificial languages. In a **3R/0E condition**, a novel "rule" was witnessed applying to 3 out of 9 cases with 0 exceptions. The Threshold proposal predicts that children in this condition will not use the rule productively because an insufficient number of rule-following cases are witnessed: recall that in a domain of 9, a minimum number of *5* cases is required for productivity. The C&A proposal predicts, on the other hand, since 0 exceptions were witnessed, there should be no competition. Therefore, the C&A proposal predicts that as long as children understand the function of the construction and are able to access it, a construction that is

witnessed applying to 3 cases and 0 exceptions *will* be used productively.

In a separate **5R/4E condition**, a second novel rule was witnessed applying to 5 out of 9 cases, while 4 other cases were exceptional. The only difference between this 5R/4E condition and the 5R/1-1-1-1E condition in SYN '16 is that here the 4 exceptional cases behaved alike. In both cases, there were 5 rule following cases and 4 non-rule following cases. Therefore, the Threshold proposal predicts the 5R/4E rule should be as categorically productive as the 5R/1-1-1-1E rule of SYN '16 was. The C&A proposal, on the other hand, predicts that the "exceptional" construction—which was applied to 4 entities—should interfere with learners' ability to access the higher type frequency construction—which was applied to 5 entities. Because there is only a 25% difference in availability (and interference) between the two patterns, and no conditioning factors that could systematically distinguish the two, interference should render the (slightly) more dominant construction—the one witnessed applying to 5 entities—less than fully productive. C&A further predicts that when the more dominant construction is not used, the competing, less dominant construction will be used instead.

To summarize, the Threshold proposal predicts that when exposed to the 3R/0E rule, children should treat it as completely unproductive, and when exposed to the 5R/4E rule, they should treat it as completely productive. The C&A proposal, on the other hand, predicts that the 3R/0E pattern should be productive because it has no competition. As long as children are able to understand the task, they should use the pattern productively. In the 5R/4E condition, C&A predicts that the dominant pattern should be subject to interference from the less dominant pattern and should therefore be less than fully productive. Again, if children fail to use the dominant construction, C&A predicts that they will use the competitor, less-dominant construction instead. The predictions of the two proposals are represented in Table 3.

Table 3: Predictions of the Threshold and Communicate and Access proposals.

NEW CONDITIONS: M vs. e, (M= #Rule-following cases e = exceptional cases)		**PREDICTIONS:**
3R/0E	**Threshold:** SP is not satisfied	$Rule_3$ should not be productive
	C&A: Tentative constructional generalization has no interference from alternatives	$Construction_{(3)}$ should be productive
5R/4E	**Threshold:** TP and SP are satisfied	$Rule_{(5)}$ should be productive
	C&A: Dominant construction is only 25% more accessible than interchangeable alternative construction	$Construction_{(5)}$ and $Construction_{(4)}$ should compete

Experiment

Preregistration at Open Science Framework (OSF). We preregistered a plan to test 16 children without counterbalancing the constructions across conditions, and to use *t*-tests against full and 0 productivity (following SYN '16). We subsequently preregistered a second design with another 16 children in order to counterbalance the constructions (plural vs. classifier) and in order to preregister a more appropriate mixed model (glmer) analysis. Data was collected for each experiment only after it was preregistered. Results are combined below, as is appropriate, but both groups of participants were also analyzed separately (the first group with and without the 5 additional children tested with slightly different instructions). The pattern of results reported below remain unchanged in these subgroups.

Methods

Participants

32 children between the ages of 4 and 6 (M = 56 months) are analyzed below. We changed the instructions after an initial 5 children were tested and these children are excluded from analysis. All but one child provided four critical responses, two in each condition. One child opted out after the first condition (which happened to be 3R/0E for this child). All children were tested at the Princeton University Baby Lab, two were bilingual English speaking and the rest were monolingual English speakers. All had normal hearing and vision and were born at full term (38+ week gestation)**.** After each question, children received a sticker regardless of their response, and after the study, each child received a book and a prize, and the family received $10**.**

Procedure

The design was within-participants. In each of the two conditions, children were exposed to a mini-language that included 1 or 2 novel words, and 9 familiar English words naming each of 9 distinct kinds of animals or crayon colors. In the 3R/0E condition, a single novel form (*po*) was witnessed being used with 3 out of 9 items. In the 5R/4E condition, one form (*dax* or *fep)* was randomly assigned to 5 of the 9 items, and the other form was assigned to the remaining 4 items (see Figure 2).

The following were counterbalanced (in a nested fashion) across participants:

- order: whether children witnessed the 3R/0E or the 5R/4E condition first
- function: whether the rule/construction tested had a plural function or was used as a classifier
- item: whether the 9 items (or pairs of items) in the domain were crayons or animals
- dominant form: whether *dax* or *fep* was dominant form in the 5R/4E condition (*po* was consistently used in the 3/0 condition).

In each condition, the choice of which individual items (animals or crayons) was witnessed in the target

construction was randomly determined for each child, as was the order of presentation of items.

Pretest before each condition. Children were asked to count the 9 distinct entities in order to ensure that they recognized that the relevant domain size was 9. Children were then asked to name each distinct animal or crayon color. After each response, children received a sticker. All children succeeded in both tasks.

Exposure to a potential rule and exceptions. Children were then introduced to a puppet, Mr. Chicken, who, they were told, spoke a different language. Each child took part in both the 5R/4E and 3R/0E conditions as follows:

5R/4E condition: each child witnessed the rule applied once to each of 5 unique cases. 4 other cases were witnessed that were exceptional in that they did not follow the rule.

- When the novel forms were **classifiers,** Mr. Chicken picked up each of the objects and named the entity in "chicken language," saying the name of the entity immediately followed by a novel classifier, 5 of which followed the dominant pattern and 4 of which followed the exceptional pattern, ordered randomly. (*e.g., lion fep, monkey dax, zebra fep, giraffe dax…).* There were no conditioning factors that determined which novel classifier was used with each animal. Children were asked to repeat every novel form witnessed.
- When the novel forms were **plurals,** Mr. Chicken picked up one of each type of object, said its name and then picked up two of the same type, and used a novel suffix as a plural marker (e.g., *lion,* picking up one lion, *lion dax,* picking up two lions). Children repeated each singular and plural form. 5 entities were pluralized with one morpheme (*dax* or *fep*, counterbalanced) and the other 4 were pluralized with the other form. The items assigned to each novel plural marker were selected randomly, so there were no conditioning factors that determined which novel plural was used. Items were selected in random order (e.g., *lion, lion dax; monkey, monkey fep)*

3R/0E condition: each child witnessed the rule applied once to each of 3 unique cases. The other 6 entities were not witnessed either following the rule or being exceptional.

- When the novel form was a **classifier**, Mr. Chicken picked up 3 animals (or crayon colors) and named them with a novel classifier, *po* (e.g., *lion po, zebra po…*). Which animals were named was random for each child. Children were asked to repeat each label after hearing it.
- When the novel form was a **plural**, Mr. Chicken picked up and named one animal or crayon (e.g., *lion*) and then picked up two of the same animals or crayons which were labeled with the name and the plural morpheme, *po* (e.g., *lion po*). This was done for 3 types of animals or crayon colors, selected randomly. Children were asked to repeat each label after hearing it.

Production task

After initial exposure, children were asked to label another item the way Mr. Chicken would. Then, children were exposed again in the same way to the same condition and were asked to label a different item. This provided two responses for each condition. In the 5/4 condition, children were asked to label two never-before-seen items. In the plural condition, children labeled one of the remaining 6 items they hadn't heard labeled. Thus, children provided four critical responses, two in each condition.

Figure 2: Sample stimuli. 3R/0E condition with animals and 5R/4E condition with crayons. The two functions of the constructions, and animals vs. crayons were separately counterbalanced across

Results

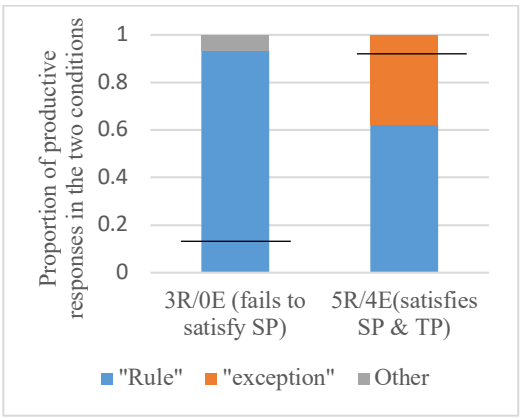


Figure 3: Proportion of responses in 3R/0E and 5R/4E conditions. Dominant form ("rule," blue); less dominant form (orange; relevant in 5R/4E condition), or other (gray). Domain size = 9. Black lines indicate Threshold predictions for height of the rule-following cases (in blue).

The values indicated by blue are the proportion of cases in which the rule was used productively. The Threshold

proposal's predictions for the expected proportion of rule-following cases (in blue) are indicated by the black lines in Figure 2; they are predicted to be identical to those in Figure 1. However, in the 3R/0E condition, the proportion of responses in which children used the novel form productively was near ceiling (M = .935; Figure 2, left). Specifically, 29 out of the 32 children consistently used the novel form productively, 2 children used it on one out of two trials, and only 1 child failed to use it at all.

In the 5R/4E condition, 14 out of 31 children consistently used the dominant form productively for new novel objects, 5 children consistently used the slightly less dominant form productively, and another 12 children produced both of the novel forms (one on each trial) (M = .625; Figure 2, right).

We analyzed the data using the glmer package with Condition (3R/0E or 5R/4E) as the predictor and by-subject random intercepts and slopes, and random intercept for Function-Assignment (plural or classifier first): Correct ~ Condition + (1 + Condition |Subject) + (1|Function-Assignment), family=binomial, data). Recall that other random factors were counterbalanced. The Threshold proposal predicted that the 3R/0E condition should be categorically unproductive and the 5R/4E condition should be categorically productive. However, results show a significant difference between the two conditions in the *opposite* direction ($\beta = 8.824, z = -2.918, p = 0.0035$). This is consistent with the Communicate and Access proposal which predicted that children should be productive in the 3R/0E as long as they understood the task, since there was no interference from a competing form; and children should be markedly less productive in the 5R/4E condition since the 2 forms witnessed would compete with one another, as there were no conditioning factors available to distinguish them.

Prior work has found children over-rely on *either* of two options when the difference in type frequency is not overwhelming (Hudson Kam & Newport 2005; 2009; Schwab, Lew-Williams and Goldberg 2018), and this occurred in the 5R/4E condition. In fact, 19 out of 31 children used only one form or the other: 14 children only used the more dominant form, and another 5 only used the less-dominant form. For this reason, it is not particularly meaningful to compare children's performance in the 5R/4E condition to chance. A majority of children chose one of the two options and simply repeated that option for all cases. But it also not appropriate to describe children's behavior as treating the more dominant form as a rule, given that fewer than half of the children consistently used the dominant form (14/31). Moreover, the remaining 12 children used one of each form, which is a pattern of behavior regularly seen in adults, when two options are witnessed with nearly equal type frequency (Hudson Kam & Newport 2005). In a comparison of the age of the 19 children who used a single form and the 11 children who used both forms, we find on average, that the latter group was 4 months older (M = 54 vs. 59 months). Using a 1-tailed t-test, this result is marginally significant (t = 1.61, p = .059).

Discussion

Critically, children were more productive in the 3R/0E condition than in the 5R/4E condition, directly contradicting the Threshold proposal's predictions, while being consistent with the predictions of Communicate & Access. Moreover, the Threshold proposal makes clear predictions that were disconfirmed in each condition considered separately.

In the 3R/0E condition, the Threshold proposal predicted that children should have been completely *un*productive, as they witnessed fewer rule-following cases than the number demanded by the Sufficiency Principle, given the domain size of 9: i.e., they only witnessed 3 cases, when 5 is predicted to be the minimum number required. Nevertheless, children overwhelmingly used the novel construction productively. The Sufficiency Principle has generally been argued to require an unrealistically high number of rule-following cases be witnessed in order for productivity to be realized (Goldberg 2018, 2019), and children's behavior in the 3R/0E condition confirms this. It is highly unlikely that children misjudged the size of the domain of the construction, since there were exactly nine entities (or pairs of entities) in the display (Figure 2) and children accurately counted them at the beginning of each condition. In fact, if children had assumed that the domain only included the three items that had been witnessed in the novel construction, with the other cases falling outside of the construction's domain, then the construction should not have been applicable to the other cases, and yet children overwhelmingly *did* extend it to the randomly selected new entities at test.

In the 5R/4E condition, the Threshold proposal predicted full productivity of the dominant form (which was witnessed with 5 out of 9 cases), as both the Tolerance and Sufficiency principles were satisfied. While 45% did use the dominant form productively, another 16% used the "exceptional" form productively. The rest, 39% of children, used both forms, one with each of the new entities. Defenders of the Threshold proposal might argue that the last group of children interpreted the input as evidence for *two* distinct and exceptionless rules, one of which applied to 4 cases and the other of which applied to 5 cases. However, this would require distinct domains for the two rules, and yet no conditioning factors were provided. Recall that instances that appeared with the dominant form and instances that appeared with the less dominant form were selected at random and differed across children. And, although the difference in type frequency between the dominant and less dominant constructions was close (5:4), it falls squarely within the thresholds that were proposed for the more dominant construction to become productive as children had done in the 5R/1-1-1-1 case reported by SYN '16.

Is it possible to defend the Threshold proposal on the grounds that the children in the current experiment were more adult-like? That is, the Threshold proposal is specifically aimed at young children's behavior rather than adults', since adults are recognized to behave somewhat differently than children in artificial language paradigms (Boyd & Goldberg 2012; Hudson Kam & Newport 2005, 2009), perhaps relying

on strategies or metalinguistic awareness that is unavailable to children as they learn their first language. Notably, however, the children in the current work were almost 3 years younger than those tested by SYN '16 (56 *vs.* 90 months).

Results in both conditions are consistent with the Communicate and Access proposal. In the 3R/0E condition, only one option was provided and so there was no interference from any competitors. C&A predicts that as long as children are able to appreciate the convention and access the form, they should use the form productively for new cases, as they overwhelmingly did. The results of the 5R/4E condition are also consistent with the C&A proposal. Since there were no conditioning factors to distinguish the two constructions, and since the forms were nearly equal in dominance (type frequency), C&A predicted that children would have no good way to resolve the competition between them. In fact, 14 children consistently used the more dominant option, while 5 used the less dominant form. This over-reliance on a single form recalls prior work that investigated children's productions when faced with unconditioned variation (Kam & Newport, 2005; Singleton & Newport, 2004), or when faced with variation that is conditioned, but by factors that the children fail to recognize (Schwab, Lew-Williams, & Goldberg, 2018). In those studies, children tended to rely on a single option in production tasks, but recognized both forms as acceptable in judgment tasks. The discrepancy between production and judgment tasks suggests that the over-reliance on one form during production results from the challenge of accessing and choosing between multiple forms without any reason to prefer one over the other (Harmon & Kapatsinski 2017; Schwab, Lew-Williams, and Goldberg, 2018).

Recall that the Communicate and Access proposal takes as its starting point the idea that learners aim to convey their messages *while obeying the conventions of their language as best they can*. While it is simpler to over-rely on one option in the face of unconditioned variation, it is more conventional to use both options, since both options were witnessed. As expected, then adults should be more likely to match the relative frequencies in the input even when faced with unconditioned variation between two alternatives, because they are better able to access both forms and choose between them. And in fact adults do tend to be more successful than children at matching the input veridically in mini-artificial language experiments (Kam & Newport 2005; 2009; SYN '16). We see evidence that an over-reliance on a single form is *simplification* in the current work, in that 12 out of 31 children used both novel forms in the 5R/4E condition. Moreover, the children who used both forms in their own productions were marginally older than those who over-relied on a single option, by an average of four months. We take that as an indication that children attempted to successfully produce both options, with older children simply being more successful.

We therefore conclude then that interference—the nature of the exceptional cases—played a key role in whether a competing form was used productively. That is, the difference between the 5R/4E condition here and 5R/1-1-1-1 condition of SYN '16 is that the current class of exceptions all occurred with the same form, making the "exceptional" form itself accessible. And since the exceptional case was just as appropriate for expressing the intended message (i.e., there were no conditioning factors that made either more appropriate), and the "exceptional" cases were nearly as accessible as the "rule," the C&A proposal predicted that the exceptions would interfere with the productive use of the rule. And this is evident in the current results in that children were significantly less productive in the 5/4 condition than in the 3/0 condition.

Conclusion

The present work investigated the factors that underlie children's productive use of a novel rule or construction. We compared two proposals that make contrasting predictions. The first, a Threshold proposal, argues that rules are used productively as long as two thresholds are met: the proportion of potential cases that are witnessed obeying a rule must cross a threshold in order to satisfy a Sufficiency principle and the proportion of potential cases that are witnessed behaving exceptionally must remain below a threshold in order to satisfy a Tolerance principle (Yang, 2016). A Communicate and Access proposal instead appeals to the idea that a speaker's goal is to convey her intended message while obeying the conventions of her language as best she can. On this view, children extend constructions in new ways when they need to express a given message and they are unable to access a more conventional or better match. Accessibility of a construction increases as the variability of witnessed cases increases; and accessibility of the construction decreases as the accessibility of a competing construction increases.

In the current experiments, 4-6-year-old children were exposed to 2 mini-artificial languages. Each language provided exposure to a potentially productive rule, which was assigned a plural or classifier function. In one condition, a novel construction was witnessed applying to 3 out of 9 cases and 0 exceptions. The Threshold proposal predicted that children would *not* use this 3R/0E rule productively, as too few instances were witnessed to satisfy the Sufficiency principle. The Communicate and Access proposal predicted that children *would* use the construction productively because there was no better way to communicate their intended message; i.e., there was no interference from any competing alternative. As predicted by the C&A proposal, the construction was overwhelmingly used productively.

The other condition exposed children to 5 out of 9 cases following a rule, the 4 other cases being exceptions to that rule. The Threshold proposal predicted that in this 5R/4E condition, children should be fully productive, since a sufficient number of rule-following cases was witnessed, and a low enough number of exceptions was witnessed. Unlike in previous work (SYN '16), here the 4 exceptional cases all behaved alike. The Communicate and Access proposal predicted that there would be competition between the two constructions, and that this would interfere with the productivity of both. In fact, there was markedly less

productivity in the 5R/4E condition than in the 3R/0E condition, counter to what the Threshold proposal predicted and consistent with the C&A proposal. Children in the 5R/4E condition over-relied on the dominant construction (45%), *or* on the less dominant construction (16%), *or* they used both constructions (39%).

To summarize, our preregistered experiment contradicts the Threshold proposal while being consistent with Communicate and Access. We conclude that productivity is encouraged by the desire to communicate a message while obeying the conventions of the language. On this perspective, we do not extend a construction productively unless we are unable to access a "better" (more conventional and appropriate) way to express our intended message. Productivity of a construction is contrained by the accessibility of the construction, and accessibility is affected by both the variability of witnessed exemplars and interference from a competing construction (Goldberg, 2019). When there is no better way or when we are unable to access a better way at the moment of speaking, we have no choice but to extend appropriate constructions that *can* be accessed.

The C&A proposal takes a different perspective on prior findings that children tend to "regularize" their input, making it more systematic and therefore in some sense better. The C&A proposal suggests that "regularization" arises from a failure to successfully access a more conventional and appropriate alternative. C&A takes the position that both children and adults aim to conform to the conventions used by others who are considered to be knowledgeable. Adults are more successful at reflecting the input veridically given very limited exposure, but children aim to--and ultimately do--learn the conditioning factors of the constructions they are exposed to, and to a remarkable extent, successfully conform to the speech patterns used in their language communities. In fact, we saw adult-like behavior in a subset of (somewhat older) children in the current experiment who, in the 5R/4E condition, used both options.

The Threshold proposal faces other outstanding issues that are not addressed here. For example, exceptions are assumed to be searched serially and before rule-following cases, despite a lack of psycholinguistic evidence for this claim (Hernandez, 2019; Wittenberg & Jackendoff 2018; Kapatsinki 2018). The proposal assumes that exceptions are listed in order of frequency so that neither exceptions nor rule-following cases are allowed to cluster within our associative memory as proposed by the C&A and other accounts (Ambridge et al. 2018; Bybee 2010; Goldberg 2019; Kapatsinki 2018; McClelland& Patterson 2002). Without allowing instances to cluster in memory, it is entirely unclear how children are able to determine the domain of a rule, let alone calculate the size of the domain, as is required for the Threshold proposal to make any predictions at all.

To summarize, constructions (or "rules") do not "go productive" by crossing predetermined numerical thresholds. Rather, people extend constructions for new uses when doing so provides an accessible way to best express their intended messages.

References

Ambridge, B., Barak, L., Wonnacott, E., Bannard, C., & Sala, G. (2018). Effects of both preemption and entrenchment in the retreat from verb overgeneralization errors. *Collabra: Psychology*, *4*(1).

Bates, E., & MacWhinney, B. (1987). Competition, variation, and language learning. In *Mechanisms of Language Acquisition* (pp. 157–193). L. Erlbaum Ass.

Berko, J. (1958). The child's learning of English morphology. *Word*, *14*(2–3), 150–177.

Boyd, J. K., & Goldberg, A. E. (2012). Young children fail to fully generalize a novel argument structure construction when exposed to the same input as older learners. *Journal of Child Language*, *39*(3), 457-481.

Bybee, J. (2010). *Language, usage and cognition*. Cambridge University Press.

Goldberg, A. E. (2018). The sufficiency principle hyperinflates the price of productivity. *Ling. Approaches to Bilingualism*, *8*(6), 727-732.

Goldberg, A.E. (2019) *Explain me this: creativity, competition, and the partial productivity of constructions*. Princeton University Press.

Freudenthal, D., Pine, J. M., & Gobet, F. A 2018. Computational Model of the Acquisition of German Case. *Proc. of Cog. Sci. Conference*. Madison, Wisc.

Harmon, Z., & Kapatsinski, V. (2017). Putting old tools to novel uses . *Cog. Psych. 98*, 22–44.

Hernandez, Alexia. (2019). Rule Generalization in Children: Testing a Threshold Proposal. Ling. Senior Thesis,Princeton.

Kam, CH & Newport, EL (2005). Regularizing unpredictable variation. *Lang. Learn.& Dev. 2*(2), 151–195.

Kam, C. L. H., & Newport, E. L. (2009). Getting it right by getting it wrong. *Cognitive Psychology*, *59*(1), 30-66.

Kapatsinski, V. (2018). On the intolerance of the Tolerance Principle. *Ling. Approaches to Bilingualism*, *8*(6), 738-742.

Kapatsinski, V. (2018). *Changing minds changing tools.* MIT Press.

MacDonald, M. C. (2013). How language production shapes language form & comprehension. *Frontiers in Psych. 4*: 1–16.

McClelland, J. L., & Patterson, K. (2002). Rules or connections in past-tense inflections. *TiCS 6*(11), 465-472.

Montag, JL, Matsuki, K., Kim, JY, & MacDonald, MC (2017). Language Specific and Language General Motivations of Production Choices *Collabra: 3*: 1–22.

Schuler, K. D., Yang, C., & Newport, E. L. (2016). Testing the Tolerance Principle . *Proc. of Cog Sci.*

Schwab, JF, Lew-Williams C, & Goldberg, AE (2018). When regularization gets it wrong. *JCL* 1–19.

Singleton, JL & Newport, EL(2004). When learners surpass their models. *Cog. Psych.49*(4), 370–407.

Suttle, L., & Goldberg, A. E. (2011). The partial productivity of constructions as induction. *Linguistics*, *49*(6), 1237-1269.

Tenenbaum, JB & Griffiths, TL (2001). Generalization, similarity, and Bayesian inference, *BBS:* 629–640.

Wittenberg, E. & Jackendoff, R. (2018). Formalist modeling and psychological reality. *Ling. Approaches to Bilingualism*, *8*(6), 787-791.

Yang, C. (2016). *Price of Linguistic Productivity.* MIT Press

Linguistic syncopation: Alignment of musical meter to syntactic structure and its effect on sentence processing

Courtney Hilton (courtney.hilton@sydney.edu.au)
School of Psychology, The University of Sydney, Sydney, Australia
Centre for Research on Learning and Innovation, Sydney, Australia
Micah Goldwater (micah.goldwater@sydney.edu.au)
School of Psychology, The University of Sydney, Sydney, Australia

Abstract

Language and music are structured at multiple temporal scales and have been characterized as having meter: a hierarchical and periodic alternation of the prominence of syllables/beats. Meter is thought to emerge from the entrainment of neural oscillators, affording temporal expectations and selective attention. Higher-levels of a metric hierarchy also tend to track syntactic phrase structure, however, it is not clear within the framework of temporal attending why this would be advantageous. Neural oscillations have recently been shown to also track syntactic phrases. We propose that meter aligns to phrase structure so as to make syntactic processing more efficient. In two experiments (both visual and auditory language), we show that certain alignments of meter to syntax influence sentence comprehension and we suggest potential mechanisms for why certain alignments tend to be preferred. Our results underline the rhythmicity of not only low-level perception but also of higher-level cognitive processing of syntactic sequences.

Keywords: Language, time, oscillations, musical meter, syntax, merge

Introduction

Music and spoken language present similar challenges to a listener in that structure at multiple temporal scales must be decoded from a continuous sound signal in real-time. It is, therefore, no surprise that there are some parallels in how this is achieved and, as such, also parallels in musical and linguistic structure that bear the mark of these shared processing means. One such parallel is metrical structure. Meter generally refers to the perceived hierarchical alternation of stress in syllables in speech (Port, 2003), or beats in music (for more detail, see: Lerdahl & Jackendoff, 1983). In this paper, we motivate a view of meter as being something emerging from, on the one hand, the computational problem of extracting a discrete structured representation from a continuous signal, and on the other hand, an algorithmic solution that fits within the implementational oscillatory-constraints of neuro-computation (Rimmele et al, 2018a).

In explaining what meter affords its perceiver, the predominant theory has been that it is a system for *predicting when* and that these temporal expectations then in turn afford the dynamic allocation of attention to expected points in time to optimize processing (Jones, 1976; Pitt & Samuel, 1990). These theories of 'dynamic attending' have been formalized in models using coupled neural oscillators to explain how meter is flexibly entrained to a signal and how the dynamics of hierarchical perceived stress emerge naturally from this mechanism (Large & Jones, 1999; Port, 2003).

More generally, there is an attractive isomorphism between the temporally multi-scaled structure of language and music, and the multi-scaled oscillatory paradigm of neural processing in the brain. The consensus seems to be that the entrainment of one to the other—'tuning the inside to the outside'—is, at least, important if not necessary to both basic perception and perhaps even to deeper analysis and comprehension. As such, the concepts of oscillation and entrainment have become central in recent cognitive and neuroscientific theories of language processing (Giraud & Poeppel, 2012), and in theories of music processing for both rhythm/meter (Large & Kolen, 1994) and tonality (Large et al, 2016).

Specifically for the case of speech, it is proposed that the auditory cortex entrains a cascade of oscillatory sampling windows to the speech envelope: phonemes sampled with gamma oscillations (>30hz), syllables with theta (3-8hz), and intonational phrases with delta (<3hz). And while delta-oscillations are normally observed to follow prosody (Bourguignon et al, 2013) they have recently been shown to track syntactic phrases, even in the absence of prosodic cues (Ding et al, 2016), thus demonstrating top-down linguistic knowledge. Meyer and colleagues (2017) additionally showed that when prosody and syntax are misaligned, delta tracks the syntactic rather than the prosodic phrase. How should this all be interpreted?

One consideration that has been neglected is how meter figures into this: perhaps what delta is really tracking here is meter. This is especially important as the paradigms used to show delta-tracking of syntax employ a frequency-tagging approach where the speech must be presented isochronously and thus may be particularly likely to induce a subjective percept of meter. Indeed, similar paradigms have also been used to show oscillatory tracking of meter where delta too tracks higher-metric levels not present in the acoustic signal (Nozaradan et al, 2011).

While the precise rhythmicity of naturalistic speech is still hotly debated (for example, two contrasting positions: Nolan & Jeon, 2014; Brown Pfordresher, Chow, 2017), and thus the extent to which strict parallels between speech and musical

rhythm/meter are valid, it is certainly clear that there many special cases of speech where the parallel with musical meter is clear, such as poetry and song (Lerdahl, 2001). And perhaps more importantly, it seems that rhythm and meter are especially important cues during early language development, in line with the framework of prosodic bootstrapping where infants rely on prosodic information to segment input. In line with this, it has been shown that the perception of meter is present in the first year of infancy and that meter supports the learnability of other structure in a signal such as rhythm and melody (Hannon & Johnson, 2004). More generally, this idea may explain the clear metric structures of nursery rhymes and in children's literature (Breen, 2018; Fitzroy & Breen, 2019).

In recent years, there has also been growing interest in the relationship between musical experience and language abilities (for recent meta-analyses see: Gordon et al, 2015a; LaCroix et al, 2015), with the underlying rationale of some overlap in neural implementation and that music may have certain properties that enable to it to preferentially strengthen these networks (Patel, 2011). Specifically, it seems that the subcomponent of meter is particularly crucial in mediating the transfer of musical abilities to the processing of speech and syntax in language (Gordon et al, 2015b; Jung et al, 2015). Relatedly, the syntactic deficits observed in Parkinson's Disease patients may actually be more to do with a deficit in the ability to process the meter of language than a deficit to syntax directly (Kotz & Schmidt-Kassow, 2008).

These findings are surprising: why should meter support syntax? While it is conceivable that the dynamic allocation of attention could support the processing of speech under noisy conditions (where signal and noise are time delimited) such as in 'cocktail party' paradigms where this oscillatory entrainment mechanism is implicated (Zion Golumbic et al, 2013). It is not clear how this mechanism would support syntactic processing specifically. Some recent work, however, has started to provide clues. Rimmele et al (2018b) have shown that delta is involved in chunking auditory short-term memory. And relatedly, the BUMP model (Hartley et al, 2016) has provided a mechanism by which entrained oscillators support auditory short-term memory for serial order, further supported by Gilbert et al (2017) who provided empirical support for a shared resource underpinning this aspect of short-term memory and temporal precision. In summary, metrical structure (especially in the delta-range) may support aspects of short-term memory, which would then in turn, support syntactic processing.

Another not mutually exclusive possibility is suggested by Nelson and colleagues (2017). Using intracranial electrophysiological recordings, they observed fine-grained neural dynamics of syntactic structure building. Specifically, they observed a monotonic ramping of activity for each new word presented in a sentence until a syntactic constituent could be formed, at which time there is a spike of activity proportional to the number of words then a sudden decrease of activity reflecting the freeing of working-memory resources. They interpreted this in terms of a Chomskian/Minimalist merge operation (see Friederici et al, 2017), however, these observations can also be interpreted in less theoretically committal 'chunking' terms. Regardless, this result captures real-time dynamics of processing demands that relate to syntactic structure building, and that these demands stack-up toward ends of phrases where ramping of activity reaches its summit and where there is a 'spike' of activity that merges/chunks the information. And importantly, these demands are time localized. Therefore, if the 'strong' and 'weak' of meter relate to oscillatory fluctuations in neural excitability then perhaps meter may also function to temporally align neural resources with these processing demands.

Some evidence linking delta-oscillations with such an idea is suggested by Meyer & Gumbert (2018), who found that the phase of delta-oscillations tends to align excitability with phrase-endings (however, they interpreted this as aligning delta with syntactic informativeness, nonetheless, their data are consistent with our idea here).

In summary, a more general way to make sense of this relationship between meter and syntax is in terms of prediction and efficiency of processing (Gibson et al, 2019), and how this is constrained by the oscillatory nature of the brain (Rimmele et al, 2018a). Syntax gives top-down prediction of "what next" (Levy, 2008) and meter/neural-resonance gives bottom-up prediction of "when next" (Large & Kolen, 1994). However, together they are more flexibly able to entrain to not just low-level acoustics but also higher-level structures, and thus enable more efficient processing of syntactically structured sequences as in language and music.

We now explore this idea in two experiments that manipulate the alignment of meter and syntax and measure the effect of this alignment on comprehension.

Experiment 1

Method

Our central hypothesis that is tested in both of the following experiments is that comprehension (measured by probe accuracy and response-times) is highest when the strong-beat of meter aligns most often with phrase-boundaries (see Figure 1). We also manipulate syntactic complexity by using both subject-extracted and object-extracted relative-clause sentence structures. The difference in complexity here is defined in terms dependency locality theory (Gibson, 1998), where object-extracted sentences require integrating over a greater number of words and thus pose a greater strain on resources. We predict an interaction between sentence complexity and congruency on the grounds that better-aligned resource allocation may be more needed in sentences that integrate over more words. Thus, this yields a 2 X 2 factorial design, manipulating syntactic complexity (subject- vs object-extracted relative-clause) and congruency (congruent, incongruent). While we also manipulate which clause of the sentence is probe (main or relative), we have no theoretical prediction about this other than the main clause would have higher accuracy.

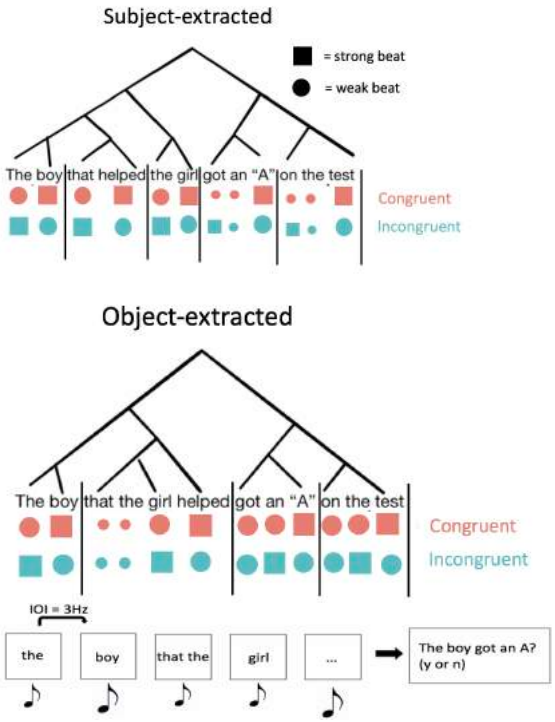


Figure 1, Top & Middle: examples of syntactic tree-structure for a sentence and possible alignment with meter defining congruencies. Bottom: trial presentation schematic

Participants 40 native English speakers (20 female) from the Sydney area took part in this study and were naive to its purpose.

Materials The language materials consisted of 48 sentences composed of largely monosyllabic words, each 12 words long. Each sentence had subject- or object-extracted versions. We used an additional 25 filler-sentences of assorted structure and length. Each sentence has an accompanying comprehension probe, which was balanced within participants as to probing either the main- or relative-clause and whether the correct answer was "yes" or "no". For example, if the sentence was "The boy that the girl helped got an A on the test", the probe was either "The boy/girl got an A?" or "The boy/girl helped the boy/girl?". The congruency manipulation was achieved by shifting the phase of the metrical pattern relative to the language presentation such that the strong beats fell on different positions in each phrase, e.g. "the BOY" or "THE boy" (see Figure 1). To fit the structure of the subject or object extracted forms, these sentences appeared in either binary or ternary meters respectively. Conditions were randomized over the sentences for each participant and presented in a random order.

The auditory materials were generated using a Python script and consisted of a 333Hz pure tone in which a 3Hz beat was induced by amplitude-modulating the signal with an asymmetric Hanning window with 80% depth and a 19:1 ratio of rise-to-fall time. Metrical accents were then applied by a 50% volume increase every 2 (binary) or 3 (ternary) tones.

Procedure The experiment was self-paced, and after an initial practice block, was completed in a single block where the participant was encouraged to take short breaks between trials. The experiment was run using software written in Python, using the PsychoPy library. Each trial begins with one full-bar of the meter (three strong beats) while a fixation-cross is shown center screen, after which the words begin appearing in the place of the fixation cross synchronized to the auditory tones. At the end of the sentence, the probe question appears center screen, and the participant is prompted to respond as quickly as possible with either "y" or "n" keys on a keyboard. If participants take longer than 5 seconds to respond, they will be prompted to speed up on the next trial. The participant also receives corrective feedback after each trial and is encouraged to balance speed with accuracy.

Results

Comprehension data were analyzed using a mixed-effects logistic regression including fixed-effects for congruency (congruent, incongruent), syntactic complexity (subject-RC, object-RC), probed clause (main-clause, relative-clause), and the interaction between congruency and syntactic complexity. We also included random intercepts for participants and items. Response times (RTs) were analyzed using a linear mixed-effects regression with the same structure. All analyses were done in R.

As seen in Figure 2, participants made fewer comprehension mistakes in the congruent conditions (χ^2 = 7.99, p = .005), fewer mistakes for the subject-RC sentences over the object-RC ones (χ^2 = 26.21, p = <.001) and fewer mistakes when the main clause is probed rather than the relative clause (χ^2 = 40.03, p = <.001). There was, however, no significant interaction between congruency and syntactic complexity (χ^2 = 0.43, p = .513). There was also no significant effect of congruency on reaction times (χ^2 = 1.20, p = 0.273). However, there were significant effects of syntactic complexity and probed-clause on RTs (χ^2 = 16.314, p = <.001; χ^2 = 35.796, p = <.001).

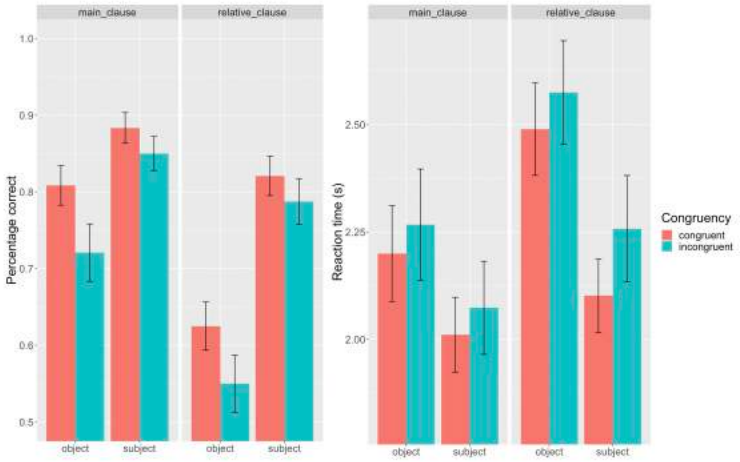


Figure 2, Experiment 2 results. Left: accuracy results as percentage correct. Right: response-time results in seconds.

Discussion

In line with our main prediction, congruency affected comprehension accuracy, however, there was no significant difference for response times. Against our initial prediction, there was also no significant interaction between congruency and syntactic complexity.

One limitation of the design was that the meter was induced passively with the auditory stimuli. Thus, we do not know to what degree participants actually interpreted the stimuli according to this meter. And although there was no significant interaction between congruency and syntactic complexity, one potential problem of the design was that object-RC sentences always had a ternary meter and subject-RC sentences always had a binary meter. This issue is an inevitable consequence of the phrase lengths in these respective sentence types, however, it may complicate the interpretation of an interaction. Further, reading sentences presented in an RSVP format is not a naturalistic way of processing language.

Experiment 2

Method

In our second experiment we wanted to build on the results of the first, replicate the congruency effect on accuracy, and address some of its limitations. Notably, we presented the sentence stimuli as auditory speech to make it more naturalistic and to check the robustness of the congruency effect to stimulus modality. As opposed to Experiment 1, where meter was induced *passively,* in Experiment 2 we induced meter *actively* by asking participants to tap on a drum-pad in time with the strong metric-beats while they listen to the speech stimuli (see for how tapping/motor actions entrain auditory attention: Morillon & Baillet, 2017). This also allows us to use tapping consistency as a DV, thus, we add the prediction that tapping will be most consistent in congruent trials (consistency being defined as the standard deviation of their accuracy). Finally, both subject and object extracted RC sentences were presented to the same ternary meter. This allows us to discount the possible meter by syntactic-complexity confound. Although, in order to make the subject-RC sentence fit a ternary meter, we had to introduce a new potential confound of inserting silences as in Figure 3. The main reason why we opted for a ternary meter, however, was to enable us to have three levels of the congruency condition for each sentence type (congruent, incongruent-1, incongruent-2). This came with the additional hypothesis that incongruent-1 would be the most incongruent metric alignment. That is, according to our delta-oscillation hypothesis, while incongruent-1 & 2 are both equidistant from the 'merge' position of the phrase, for incongruent-2, the merge would occur while attentional resources are rising, whereas for incongruent-1, the merge would occur while this attentional energy is falling (Figure 5).

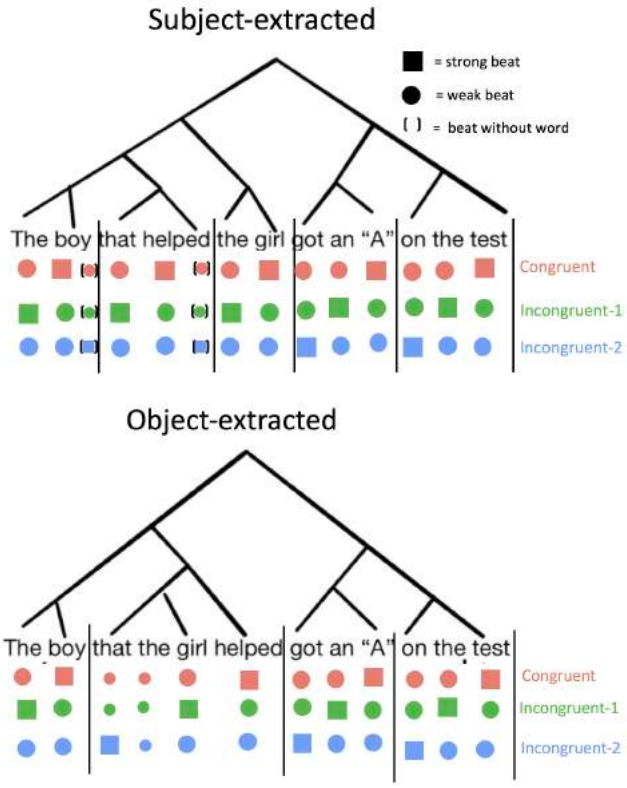


Figure 3, Congruency examples Experiment 4: syntactic tree diagrams for subject and object extracted sentences with accompanying alignments to meter defining the congruencies.

Participants (same specification as Experiment 1)

Materials Extending the 48 sentences and probes from Experiment 1, we created a further 24 of the same constraints, yielding a total of 72 sentences and probes. Speech stimuli were then generated and preprocessed from these sentence materials using a custom Python script, using Google's text-to-speech API to generate audio-files for each word individually. These stimuli were then volume normalized and cut and stretched to 2.5Hz (this new presentation-rate was based on piloting), then assembled into the sentences. Like Experiment 1, each trial starts with one full-bar of the tones to set the metric context. In a departure from Experiment 1, however, the tones drop-out when the speech stimuli start.

Procedure To ensure an active percept of the meter in Experiment 2, participants were required to tap on a drumpad in time with the strong-beats while listening to the speech (tapping once every three words). Participants used their right index finger to tap on a pressure sensitive MIDI drumpad. Before the main section of the experiment, participants completed a 'tapping-only' trial-block which estimated their tapping consistency without any language stimuli. This was then followed by practice trials for the language section and then the main trial block. Otherwise, the trial design followed that of Experiment 1.

Results

Comprehension and response-time data were analyzed using logistic and linear mixed-effects models as in Experiment 1, with the only difference being three levels of the congruency fixed-effect (congruent, incongruent1, incongruent2).

The results replicate the main effect from Experiment 1, showing that congruency significantly affected comprehension (incongruent1: χ^2 = 13.23, p = <.001, incongruent2: χ^2 = 8.30, p = .004). While the incongruent2 condition had a smaller cost on comprehension that incongruent1 (as predicted), the difference between these predictors was not significant (χ^2 = 0.533, p = 0.465). As before, there was also a significant difference between which clause is probed (χ^2 = 24.641, p = <.001). Surprisingly, however, there was no significant effect of syntactic-complexity (χ^2 = 0.101, p = .750). We believe this is a likely consequence of the added rhythmic complexity required to make subject-RC sentences fit a ternary meter.

As with Experiment 1, there was no significant effect of congruency on RTs (incongruent-1: χ^2 = 1.371, p = 0.242, incongruent-2: χ^2 = 0.837, p = 0.360) although syntactic-complexity and clause-probed had strong effects (χ^2= 19.900, p = <.001; χ^2= 49.801, p = <.001).

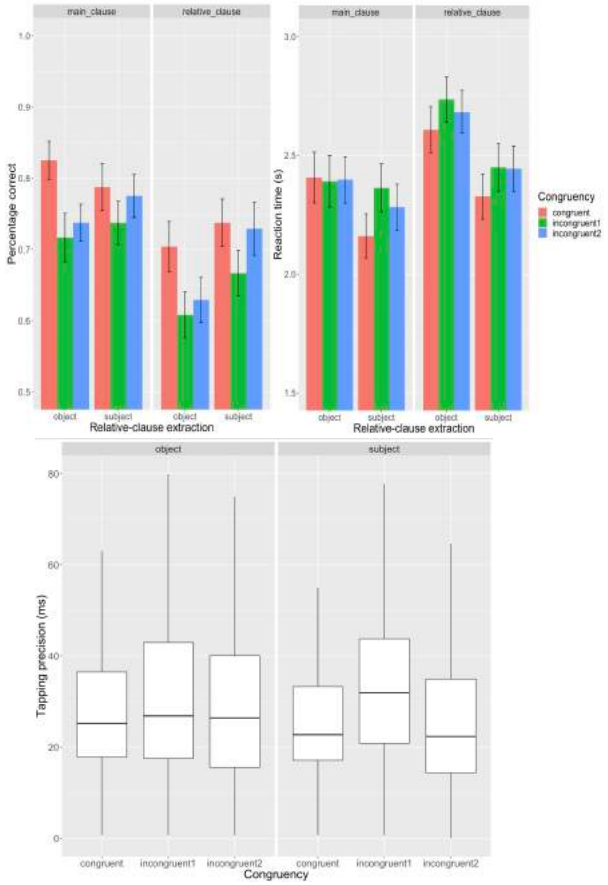


Figure 4: Comprehension results from Experiment 2. Left: accuracy. Right: response times. Bottom: tapping accuracy (standard deviation of asynchrony between tap and target tone)

For the tapping data during sentence processing, we also used a linear-mixed effects regression with congruency and sentence-extraction, and their interaction, as fixed effects, and participants as a random-effect.

Congruency had a significant effect on tapping precision in the incongruent1 conditions (χ^2= 10.27, p = .001) and only marginally significant for incongruent2 (χ^2= 3.25, p = .071). There was also a significant interaction between incongruent1 and syntactic complexity (χ^2= 12.72, p = <.001). However, it is problematic to interpret this interaction in light of the above-mentioned issues with rhythmic complexity, so we do not interpret it further.

General discussion

The results of both experiments support our hypothesis that sentence comprehension is optimal when metrical strong-beats align with phrase boundaries. We interpret this as supporting the more geneal idea that meter, and its alignment to phrase structure, plays a role in syntactic processing.

Jung and colleagues (2015) showed a similar effect of temporal expectancy on syntactic processing by having key words arrive early or late compared to an established rhythm. Kotz & Schmidt-Kassow (2015) showed that the syntactic deficit of Parkinson's Disease patients was actually due to a deficit processing the meter/timing of speech. In our study, however, each word was perfectly predictable from a rhythmic standpoint, and participants had no generalized timing deficit, however, we showed that the hierarchical distribution of attention embodied by meter, and its alignment with syntactic structure, was sufficient to show differences in comprehension.

We did not, however, find significant differences in response times. Although it is worth noting that the direction of the RT-effect was consistent with our hypothesis in both experiments. One possible explanation for this null-result is that top-down endogenous attention tends to affect accuracy, while bottom-up exogenous attention affects reaction-times (Prinzmetal et al, 2005). While lower levels of meter may be driven by bottom-up cues in the signal (Large & Kolen, 1994), it is likely that higher-levels increasingly rely on top-down phase-resetting of oscillations in response to syntactic structure or other structural cues (Rimmele et al, 2018a). Thus, this could explain why congruency had a stronger effect on accuracy over RTs. However, it may have also been that the effect was too small to detect for our sample size.

Comparison across the two studies also shows that this congruency effect is robust to modality (visual presentation in Experiment 1 and auditory presentation in Experiment 2), and thus is not specific to speech rhythms. A hypothesis that would need further experimentation to explore would be that any sequential stimulus that must incrementally form hierarchical structures would be influenced by this metric congruency effect. If this hypothesis were confirmed, it would suggest that meter is part of a more general cognitive strategy for the timely allocation of resources to process

structural relations, and that music and language are just the most prominent domains in which this plays out.

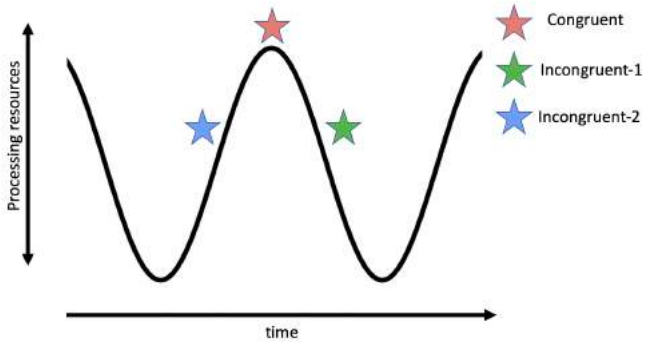


Figure 5: hypothetical oscillation of attentional resources. Stars represent the position of the phrase-ending word relative to this oscillation in each congruency condition.

Experiment 2 showed suggestive behavioral evidence of behavioral oscillations at the delta rate. Specifically, incongruent-1 conditions had a larger effect on comprehension than incongruent-2, in line with the differences in neuronal excitability predicted if the meter did phase-align delta to strong-beats (Figure 5). Although it is important to stress that this difference was not statistically significant, it was however consistent in direction across all conditions, including in the tapping data. It is also important to acknowledge that preferences for the alignment of meter to phrase structure may also be mediated by cultural factors relating to differences in language structure and preferences for grouping (Iversen et al, 2008). Future research is planned to further test these possibilities, including with more sensitive electrophysiological paradigms.

More generally, the idea that the analysis of syntactic phrases is somehow constrained by oscillatory processes is in line with the average phrase duration of speech at 2-3seconds (Vollrath, 1992), fitting within the delta-range. This may also be a neuronal constraint that results in Uniform Information Density (Levy & Jaeger, 2007), which stipulates that we, as rational communicators, attempt to spread information across a signal in a uniform way, and in a way that makes the most of our capacity (i.e. not undershooting). In other words, part of what defines our capacity to process linguistic information is these rhythmic processing constraints and the extent to which we can optimally entrain our internal rhythms to the rhythms of the information in the signal. Thus, delta sampling of syntactic phrases may define a crucial biologically grounded bottleneck that explains these patterns in human communication.

It is also likely that meter serves a similar function for the processing of harmonic syntax in music to the function articulated here for linguistic syntax (Patel, 2003). This would be consistent with some recent studies showing that harmonic structure is a strong cue for metrical strength (White, 2017). This would also accord with data showing an interaction between the processing of linguistic and musical syntax (Fedorenko et al, 2009).

Conclusion

We have shown that the alignment of meter to syntactic structure influences sentence comprehension. We have also discussed possible mechanisms from which this effect arises, namely, how delta-oscillations facilitate aspects of short-term memory processing that in turn allow for syntactic structure-building. These results imply that entraining the 'inside to the outside' may allow for more efficient processing of syntactic sequences. Future work will be required to further pick-apart the details of these ideas, ground them in neural measurement, and to explore their generality cross-linguistically and to other syntactically structured domains such as music and mathematics. In general, this work supports a co-dependency of "what" and "when" predictions, and grounds this in the biological implementational constraints of the rhythmic brain.

References

Bourguignon, M., De Tiège, X., De Beeck, M. O., Ligot, N., Paquier, P., Van Bogaert, P., … Jousmäki, V. (2013). The pace of prosodic phrasing couples the listener's cortex to the reader's voice. *Human Brain Mapping*, *34*(2), 314–326.

Breen, M. (2018). Effects of metric hierarchy and rhyme predictability on word duration in The Cat in the Hat. *Cognition*, *174*(January), 71–81.

Brown, S., Pfordresher, P. Q., & Chow, I. (2017). A musical model of speech rhythm. *Psychomusicology: Music, Mind, and Brain*, *27*(2), 95–112.

Ding, N., Melloni, L., Zhang, H., Tian, X., & Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neuroscience*, *19*(1), 158–164.

Fedorenko, E., Patel, A., Casasanto, D., Winawer, J., & Gibson, E. (2009). Structural integration in language and music: Evidence for a shared system. *Memory and Cognition*, *37*(1), 1–9.

Fitzroy, A. B., & Breen, M. (2019). Metric Structure and Rhyme Predictability Modulate Speech Intensity During Child-Directed and Read-Alone Productions of Children's Literature. *Language and Speech*.

Friederici, A. D., Chomsky, N., Berwick, R. C., Moro, A., & Bolhuis, J. J. (2017). Language, mind and brain. *Nature Human Behaviour*, *1*(10), 713–722.

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, *68*(1), 1–76.

Gibson, E., Futrell, R., Piantadosi, S. T., Dautriche, I., Bergen, L., & Levy, R. (2019). How Efficiency Shapes Human Language. *Trends in Cognitive Sciences*, 1–40.

Gilbert, R. A., Hitch, G. J., & Hartley, T. (2017). Temporal precision and the capacity of auditory–verbal short-term memory. *Quarterly Journal of Experimental Psychology*, *70*(12), 2403–2418.

Giraud, A.-L., & Poeppel, D. (2012). Cortical oscillations and speech processing: emerging computational principles and operations. *Nature Neuroscience*, *15*(4), 511–517.

Gordon, R. L., Shivers, C. M., Wieland, E. A., Kotz, S. A., Yoder, P. J., & Devin Mcauley, J. (2015). Musical rhythm discrimination explains individual differences in grammar skills in children. *Developmental Science*, *18*(4), 635–644.

Gordon, R. L., Fehd, H. M., & McCandliss, B. D. (2015). Does music training enhance literacy skills? A meta-analysis. *Frontiers in Psychology*, *6*(DEC), 1–16.

Hannon, E. E., & Johnson, S. P. (2005). Infants use meter to categorize rhythms and melodies: Implications for musical structure learning. *Cognitive Psychology*, *50*(4), 354–377.

Hartley, T., Hurlstone, M. J., & Hitch, G. J. (2016). Effects of rhythm on memory for spoken sequences: A model and tests of its stimulus-driven mechanism. *Cognitive Psychology*, *87*, 135–178.

Iversen, J. R., Patel, A. D., & Ohgushi, K. (2008). Perception of rhythmic grouping depends on auditory experience. *The Journal of the Acoustical Society of America*, *124*(4), 2263–2271.

Jung, H., Sontag, S., Park, Y. S., & Loui, P. (2015). Rhythmic effects of syntax processing in music and language. *Frontiers in Psychology*, *6*(NOV), 1–11.

Jones, M. R. (1976). Time, out Lost dimension. *Psychological Review*, *83*(5).

Kotz, S. A., & Schmidt-Kassow, M. (2008). Event-related Brain Potentials Suggest a Late Interaction of Meter and Syntax in the P600 Impact of social interaction on second language learning. *Journal of Cognitive Neuroscience*, 1693–1708.

Kotz, S. A., & Schmidt-Kassow, M. (2015). Basal ganglia contribution to rule expectancy and temporal predictability in speech. *Cortex*, *68*, 48–60.

LaCroix, A. N., Diaz, A. F., & Rogalsky, C. (2015). The relationship between the neural computations for speech and music perception is context-dependent: an activation likelihood estimate study. *Frontiers in Psychology*, *6*(August), 1–19.

Large, E. W., & Kolen, J. F. (1994). *Resonance and the perception of musical meter*. *Connection Science*, 6

Large, E. W., & Jones, M. R. (1999). The dynamics of attending: How people track time-varying events. *Psychological Review*, *106*(1), 119–159.

Large, E. W., Kim, J. C., Barucha, J. J., & Krumhansl, C. L. (2016). A Neurodynamic Account of Music Tonality. *Music Perception*, *33*(3), 319–331.

Lerdahl, F., & Jackendoff, R. (1983). *A generative theory of tonal music*. Cambridge, MA: MIT Press.

Lerdahl, F. (2001). The sounds ofpoetry viewed as music. Ann. N.Y. Acad. Sci. 930, 337–354.

Levy, R. and Jaeger, T.F. (2007) Speakers optimize information density through syntactic reduction. In Adv. Neural Inf. Process. Syst. (Jordan, M.I. et al., eds), Adv. Neural Inf. Process. Syst. 849–856 72.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.

Meyer, L., Henry, M. J., Gaston, P., Schmuck, N., & Friederici, A. D. (2017). Linguistic bias modulates interpretation of speech via neural delta-band oscillations. *Cerebral Cortex*, *27*(9), 4293–4302.

Meyer, L., & Gumbert, M. (2018). Synchronization of Electrophysiological Responses with Speech Benefits Syntactic Information Processing. *Journal of Cognitive Neuroscience*.

Morillon, B., & Baillet, S. (2017). Motor origin of temporal predictions in auditory attention. *Proceedings of the National Academy of Sciences*, 201705373.

Nelson, M. J., El Karoui, I., Giber, K., Yang, X., Cohen, L., Koopman, H., … Dehaene, S. (2017). Neurophysiological dynamics of phrase-structure building during sentence processing. *Proceedings of the National Academy of Sciences*, *114*(18), 3669–3678.

Nolan, F., & Jeon, H. S. (2014). Speech rhythm: A metaphor? *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1658).

Nozaradan, S., Peretz, I., Missal, M., & Mouraux, A. (2011). Tagging the neuronal entrainment to beat and meter. *The Journal of Neuroscience*, *31*(28), 10234–10240.

Patel, A. D. (2003). Language, music, syntax and the brain. *Nature Neuroscience*, *6*(7), 674–681.

Patel, A. D., & Morgan, E. (2016). Exploring Cognitive Relations Between Prediction in Language and Music. *Cognitive Science*, *41*, 1–18.

Pitt, M. A., & Samuel, A. G. (1990). The use of rhythm in attending to speech. Journal of Experimental Psychology: Human Perception and Performance, 16, 564–573.

Port, R. F. (2003). Meter and speech. *Journal of Phonetics*, *31*(3–4), 599–611.

Prinzmetal, W., McCool, C., & Park, S. (2005). Attention: Reaction time and accuracy reveal different mechanisms. *Journal of Experimental Psychology: General*, *134*(1), 73–92.

Rimmele, J. M., Morillon, B., Poeppel, D., & Arnal, L. H. (2018a). Proactive Sensing of Periodic and Aperiodic Auditory Patterns. *Trends in Cognitive Sciences*, *22*(10), 870–882.

Rimmele, J. M., Poeppel, D., Ghitza, O. (2018b). Accuracy in chunk retrieval is correlated with the presence of acoustically driven delta brain waves. (poster*) Society for Neuroscience*.

Vollrath, M. (1992). A universal constant in temporal segmentation of human speech. *Naturwissenschaften*, 10.

White, C. (2017). Relationships Between Tonal Stability and Metrical Accent in Monophonic Contexts. *Empirical Musicology Review*, (1983), 2–5.

Zion Golumbic, E. M., Ding, N., Bickel, S., Lakatos, P., Schevon, C. A., McKhann, G. M., … Schroeder, C. E. (2013). Mechanisms underlying selective neuronal tracking of attended speech at a “cocktail party.” *Neuron*, *77*(5), 980–991.

Iconicity and Structure in the Emergence of Combinatoriality

Matthias Hofer (mhofer@mit.edu), Roger Levy
Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology,
43 Vassar St, Cambridge, MA 02139 USA

Abstract

One design feature of human language is its combinatorial phonology, allowing it to form an unbounded set of meaningful utterances from a finite set of building blocks. Recent experiments suggest how this feature can evolve culturally when continuous signals are repeatedly transmitted between generations. Because the building blocks of a combinatorial system lack independent meaning, combinatorial structure appears to be in conflict with iconicity, another property salient in language evolution. To investigate the developmental trajectory of iconicity during the evolution of combinatoriality, we conducted an iterated learning experiment where participants learned auditory signals produced using a virtual slide whistle. We find that iconicity emerges rapidly but is gradually lost over generations as combinatorial structure continues to increase. This suggests that iconicity biases, whose presence was revealed in a signal guessing experiment, manifest in nuanced ways. We discuss implications of these findings for different ideas about how biases for iconicity and combinatoriality interact in language evolution.

Keywords: phonology; language evolution; combinatorial structure; iterated learning; iconicity

Introduction

Combinatorial phonology is an important design feature of human language, allowing it to form an unbounded set of novel, meaning-bearing words from a small set of building blocks. How did it emerge in language? As part of a larger research program that attempts to explain linguistic properties through biases operating during language acquisition and use (Christiansen & Chater, 2016; Kirby, Cornish, & Smith, 2008), recent laboratory experiments have suggested how combinatorial structure could have arisen from continuous signals through a process called iterated learning (Verhoef, Kirby, & de Boer, 2014; Giudice, 2012). But while combinatorial structure might confer a range of advantages to language, it appears to be in conflict with another salient feature of communication systems: iconicity. In order to participate freely as primitives in larger composite forms that carry arbitrary meanings, the building blocks of a combinatorial system should be meaningless (Dingemanse, Blasi, Lupyan, Christiansen, & Monaghan, 2015). Iconic signs, on the other hand, are motivated by properties of the meanings they refer to.

Evidence suggests that iconicity plays an important role in bootstrapping communication. In a study where subjects had to develop novel communication systems, Fay, Arbib, and Garrod (2013), found that gesture was preferentially adopted over speech, and explained their findings in terms of gesture's

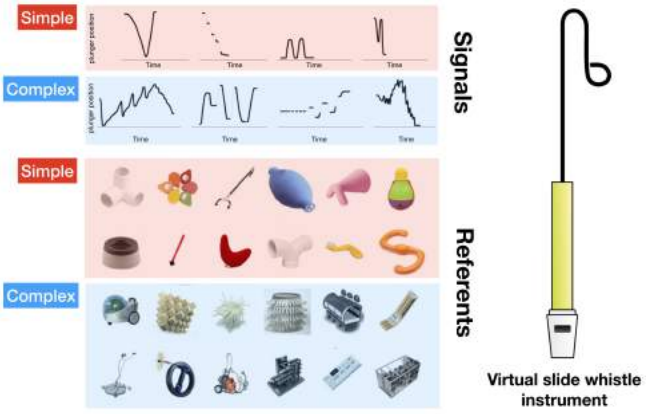


Figure 1: Depiction of stimuli and virtual slide whistle used in the iterated learning experiment to investigate the relation between iconicity and combinatoriality. Visual referent stimuli are from Lewis and Frank (2016).

stronger affinity for iconic representation. On the above account, signals tend to eventually lose these iconic origins as they develop into combinatorial systems. Goldin-Meadow and McNeill (1999) have similarly argued that iconicity is the default strategy and that combinatoriality is not adopted for the benefits it provides but to compensate when iconicity is not available. Consistent with this account, Verhoef, Kirby, and de Boer (2016) found that the onset of combinatorial structure in an iterated learning experiment was delayed when signal/referent mappings were scrambled between generations, making it harder for iconicity to develop, relative to a condition where mappings were kept intact.

Despite the possible loss of iconicity in the emergence of combinatorial phonology, many familiar forms of iconicity such as onomatopoeia or sound symbolism continue to play an important role in language (Dingemanse et al., 2015). How do these various forms of iconicity develop as signals undergo their transition from holistic to combinatorial structure? To reconcile the existence of iconicity at different stages of language evolution, we focus on a more subtle form of iconicity, recently described by Lewis and Frank (2016), that exists between word length and conceptual complexity. In their analysis of monosyllabic words across 80 languages, the authors

found that longer words are systematically associated with more complex meanings. Whereas the languages considered by Lewis and Frank (2016) are already fully combinatorial, we examine whether this form of iconicity also arises in continuous signal spaces and use it to address questions about the relationship between iconicity and combinatoriality. To this end, we conducted an iterated learning experiment where subjects evolved a set of signals through iterated reproduction. Participants learned artificial languages consisting of whistled signals that were paired with referents taken from Lewis and Frank (2016). Both signals and referents varied in complexity (Figure 1) but were paired in such a way that there was no systematic relationship between simple and complex items in the beginning. Based on the literature presented above, we predicted that such a relationship, indicative of iconicity, would emerge but eventually disappear as the communication systems become more combinatorial.

Using the languages resulting from the iterated learning study, we present results from a series of experiments designed to answer the following questions:

1. Did the languages evolve combinatoriality? This was assessed by asking subjects to rate the amount of combinatorial structure that existed in the languages.

2. Did the signals evolve iconicity? Iconicity, defined in terms of congruent complexity associations between signals and referents, was measured by collecting complexity judgments for the evolved signals.

3. Which underlying cognitive structures support our inferences about iconicity? Previous studies suggest the existence of strong biases for the development of combinatoriality and iconicity (Lewis & Frank, 2016; Verhoef et al., 2016). To better understand the role of iconicity biases and how they manifest in our experiment, we devised a guessing game where naive listeners were asked to choose the most likely referent for each signal.

After presenting our results, we close by discussing how our findings relate to different ideas about the evolution of iconicity and combinatoriality.

Experiments

To investigate how iconicity develops during the emergence of combinatorial structure, we conducted an iterated learning experiment. Miniature artificial languages were repeatedly acquired and subsequently transmitted by one 'generation' of subjects to the next. This took place across several independent transmission chains. We adopted the signal space used in Verhoef et al. (2014), in which subjects produced signals using a slide whistle instrument. Since we conducted the experiment online, we developed an on-screen, virtual version of the instrument, depicted in Figure 1. Pitch was controlled by moving the plunger up and down using the mouse. Sounds were produced by pressing down the space bar and continued until the space bar was released. Before the experiments started, participants were given an opportunity to familiarize themselves with this interface. Using the languages that evolved during iterated learning, we subsequently conducted four additional experiments to address the aforementioned questions about the emergence of combinatorial structure and iconicity.

Iterated learning experiment

The iterated learning experiment consisted of 15 independent chains, each consisting of 10 generations. Per chain and generation, a single subject learned and later reproduced an artificial language. The first subject in each chain was given a language constructed according to principles described below, while subsequent generations learned the language produced by the previous generation.

Materials Each language consisted of eight whistled sounds paired with different referents. Figure 1 shows which signals were used to initialize each experimental chain. The signals were obtained from whistles recorded and subsequently rated for their complexity in a pilot experiment. The signals were paired with unfamiliar visual objects selected from a stimulus set used in Lewis and Frank (2016), which was normed for complexity. Referents were categorized as either simple or complex. For each chain, four simple and four complex referents were selected at random from the stimuli depicted in 1 and assigned to signals with the constraint of counterbalancing between signal and referent complexity classes (half of the complex signals were paired with complex referents and while the other half was paired with simple ones and vice versa). This procedure ensured that the relation between signal and referent complexity was initially fully unsystematic.

Procedure Subjects were told that they had to learn an artificial language produced using a slide whistle with the goal of teaching the language to a computer program. After familiarizing themselves with the instrument, subjects engaged in five learning blocks, where they were shown each of the eight signal/referent pairs in random order. Each trial first displayed the visual referent, then the slide whistle playing back the corresponding signal. The whistle then stayed on screen and participants were instructed to repeat the signal. No feedback was given during learning. Subjects were admitted to the reproduction phase if they reached a learning criterion to assess how well they learned the language. The criterion test consisted of eight 2-AFC trials. Each of the eight signals was played to subjects once and they had chose the correct referent from a set of two. The distractor item was sampled from among the remaining three items of the same complexity class, preventing participants to identify the correct referent based on referent complexity alone. No feedback was given during these trials. To advance to the final stage of the experiment, participants had to correctly identify at least six of the eight items. Participants that reached the learning criterion advanced to the reproduction phase, which

was framed as a computer teaching paradigm. Subjects were asked to record each signal for a computer program that will attempt to learn the language from them. All referent stimuli were presented simultaneously on screen and subjects could chose the order in which they recorded signals by clicking on the corresponding item.

While chosen to prevent subjects from producing the same signal multiple times, the framing of the task as a teaching paradigm did not fully prevent a loss of expressivity (see, e.g., Kirby et al., 2008). Throughout the experiment, 7% of signals were identified as duplicates using a dynamic time warping-based similarity measure, and replaced with signal versions produced during learning with the constraint of being sufficiently distinct from the remaining test phase signals. This approach is conservative since signals produced during learning typically very closely resemble the input the participant was given and thus limit the amount of change (relative to the input) experienced by the next participant.

Subjects A total of 382 subjects were recruited on Amazon's Mechanical Turk. 250 subjects passed a preliminary headphone check (Woods, Siegel, Traer, & McDermott, 2017), implemented to ensure consistent listening conditions, and were admitted to the main experiment. Of those, 164 subjects reached the learning criterion. Data from 14 subjects was due to technical reasons, leaving us with 150 subjects, one subject per chain and generation.

Quantification of combinatorial structure

To answer whether signals evolved combinatoriality, we conducted a rating experiment. Naive participants were asked to rate the amount of structure present in the languages from the iterated learning experiment. Participants saw languages from either one of two conditions: In the intact condition, languages were randomly selected from across chains and generations in the iterated learning experiment. Participants were blind to which generation or chain a language came from. In the scrambled condition, participants were shown languages where signals from the different chains of each generation were randomly mixed together. Including this baseline condition allows us to assess to what extent combinatoriality judgment are about properties of the signals in the context of the language they evolved in, or simply about the structure of signals irrespective of their relation to the other signals in their language.

Subjects and Procedure Subjects were told that they had to rate the amount of structure of different newly discovered whistle languages. Structure was described as the existence of building blocks or principles that are shared among the signals in a language. Signals were presented visually in a presentation format similar to Figure 3 (but in a single row). This allowed subjects to make holistic judgments and facilitated comparisons between items in the language. Subjects were asked to report how structured a given language was, ranging from least to most structured, using a continuous slider. A total of 314 subjects took part in the rating experiment and each participant rated 24 items.

Quantification of signal complexity and iconicity

To address our second question, it was necessary to quantify signal complexity in order to assess if signals developed to match the conceptual complexity of their referents. Two experiments were conducted, one where signals were presented visually and one where they were presented auditorily.

Visual complexity ranking Similar to the previous experiment, languages were presented in the form of a visual array. Subjects were instructed to sort the signals from least to most complex. Complexity was defined as signals that have many parts and that are difficult to memorize or reproduce. 374 subjects took part in this part of the experiment and each subject rated 16 items. After realizing that effect sizes of the iconicity measure are likely too small and that the noise introduced from using a perceptual modality different from the original, auditory modality could potentially mask important differences, we conducted a second rating experiment.

Auditory complexity rating experiment Focusing on just the first five generations of iterated learning, in the second complexity rating experiment, signals were presented similar to the main experiment, with the slide whistle playing back the recorded signals. Signals were randomly selected from across chains and generations, which enabled us to obtain absolute complexity judgments (compared to the rank-level judgments obtained in the visual experiment). Subjects judged the complexity of each signal from least to most complex using a slider. 175 subjects took part in the experiment and each subject rated 16 items.

Evaluation of signal iconicity

Finally, do people exhibit iconicity biases that explain their productions during iterated learning? To develop further insight into the nature of the biases that support iconicity, we conducted a guessing game where subjects were presented all eight signals from a language (in random order) and had to identify the most likely referent per signal. Subjects were instructed that they should always choose the referent that they thought most likely belonged to the signal, and that the same referent could be chosen more than once. This allowed us assess the existence, and to quantify the strength of iconicity biases that exist in signal interpretation. 218 subjects took part in the experiment and each subject rated 8 languages.

Results

Emergence of combinatorial structure

The first question we address in our analysis is whether languages developed combinatorial structure, as described in prior experiments (e.g., Verhoef et al., 2014; Giudice, 2012). In summary, we observe that signals in all chains develop combinatorial structure. Figure 3 depicts a representative example language from the final generation of chain 1, giving

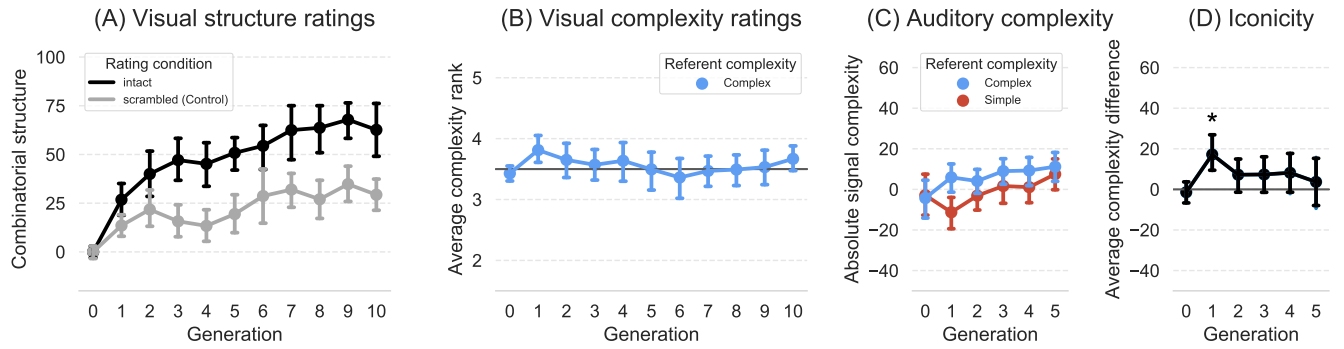


Figure 2: Results from the three experiments that we conducted to assess signal structure and complexity (visual and auditory), and the derived iconicity measure. Error bars are 95% confidence intervals.

a qualitative impression of the emergence of shared building blocks. Figure 2A shows the results from our combinatoriality measure. Across both the intact ($t(14) = 8.06, p < 0.001$) as well as the scrambled condition ($t(14) = 5.59, p < 0.001$), languages are judged to increase in structure over generations[1], but languages in the intact condition are judged to increase more ($t(14) = 2.5, p = 0.02$). This difference can only be explained by assuming that signals in the intact condition are structurally more similar to each other, which results in higher combinatoriality ratings compared to mixing languages across chains as in the scrambled condition.

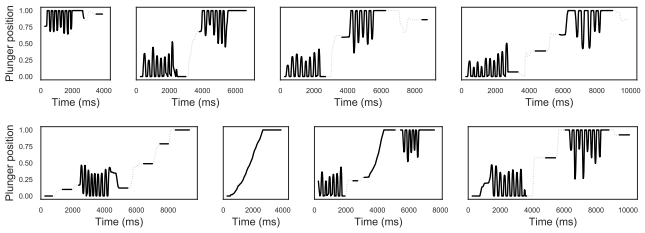


Figure 3: Set of stimuli from the last generation of chain 1. The language appears to consists of three building blocks: a short beep occurring at different pitch values, a long pitch sweep, and a wiggly segment of alternating pitches.

Development of iconic signal structure

Turning to our second question, Figure 2B depicts the results of the visual complexity ranking experiment where participants were asked to order signals from least to most complex (coded as 0 to 7). The depicted mean complexity rank represents the average rank of the four signals that were associated with complex referents. An at chance association between signals and referents corresponds to a mean rank of 3.5. While nearly all chains in the first generation have an average complexity rank of greater than 3.5, quantitatively this difference does not reach significance after correcting for multiple comparisons ($t(14) = 2.60, p = 0.02$ before, $p = 0.21$ after *Holm-Bonferroni* correction). Two features of the experimental measure could potentially mask this effect: the ranking score only captures ordinal differences and not differences in magnitude. Secondly, measurements may be noisy because the experiment was conducted in the visual instead of the auditory modality. While not posing a problem to the combinatoriality measure reported earlier (because of the larger effect sizes), this might hinder detection of iconicity in the languages.

To address these points, a second experiment collected complexity ratings in the auditory domain, restricting ourselves to the first five generations of iterated learning. Figure 2C shows the resulting ratings, grouped and averaged by associated referent complexity. These data were used to derive an iconicity measure, depicted in Figure 2D, by subtracting the average complexity of signals associated with simple referents from signals associated with complex referents. Positive values indicate the presence of iconicity in a congruent direction. As suggested earlier, we find that iconicity emerges immediately after initialization in generation one ($t(14) = 3.71, p = 0.002$ before, $p = 0.01$ after correction). While the return of the iconicity measure to chance is not significant within the first five generations of the auditory measure ($t(14) = -1.56, p = 0.14$), the visual complexity measure from 2C strongly suggests that iconicity drops back to chance in subsequent generations and thus, taken together, licenses the inference that iconicity eventually disappears from the languages.

Relationship between iconicity and structure

The previous analyses have demonstrated that languages develop both combinatoriality and iconicity over the course of the experiment. To develop insight into state-dependent trade-offs between iconicity and combinatoriality, we now look at the development of iconicity as a function of combinatorial structure instead of generation. Figure 4A shows the evolutionary trajectories of all fifteen languages in terms of their combinatorial structure and their iconicity (based on Figures 2A and D). The plots suggest that languages, while

[1] Analyses compare the regression coefficients fitted to the fifteen chains to a zero slope.

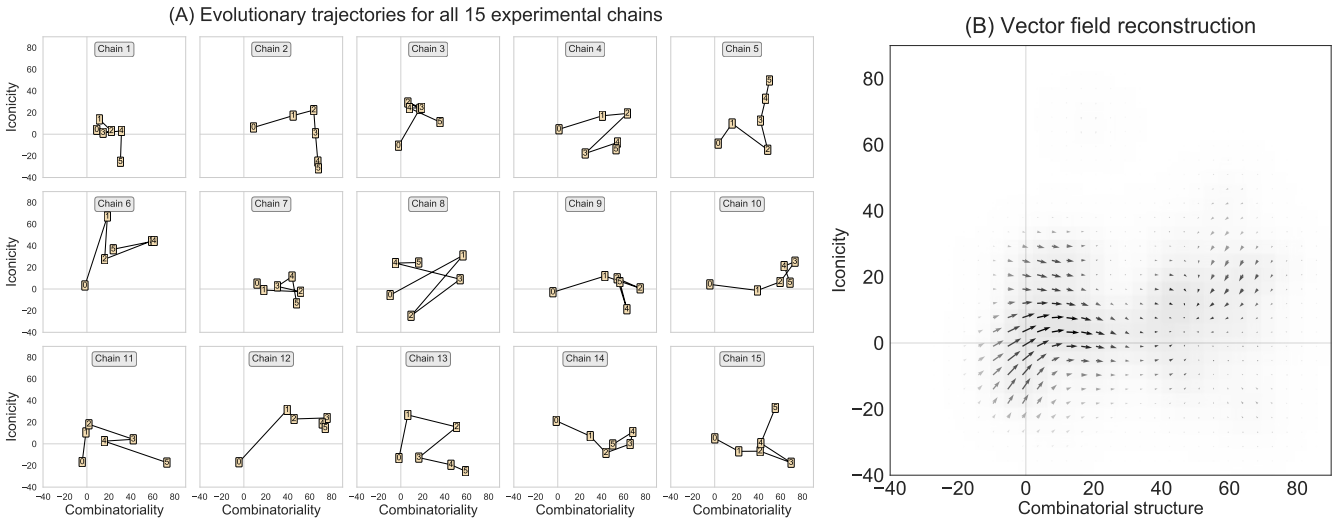


Figure 4: Visual representation of the evolutionary trajectories of all fifteen experimental chains, on the basis of which we constructed a vector field representation that shows inter-generational parameter changes for observed and hypothetical languages.

displaying common patterns, can vary substantially in terms of their developmental time constant. We summarize these data by constructing a vector field that shows extrapolated average magnitude and direction of inter-generational changes on a grid (based on the behavior of nearby languages). The model is constructed by considering all 75 vectors in Figure 4A that represent transitions from one generation to the next. For each grid point, an average magnitude and direction is estimated by computing the weighted linear combination of vectors using their distance, obtained with a multivariate Gaussian kernel centered around the grid point, as weights. Figure 4B shows the resulting vector space model. The total sum of Gaussian weights per grid point, superimposed in grey, corresponds to the number of vectors nearby that were used to construct the estimate. The model summarizes in which direction, and how much, hypothetical languages would change in terms of combinatoriality and iconicity, based on the observed data. Adding to the results reported above, gains in iconicity or maintenance of already existing iconicity is only observed when languages are still relatively unstructured. Languages loose their iconic structure as combinatoriality increases further. More sporadically observed 'extreme' levels of iconicity and combinatoriality appear to be unsustainable and eventually revert to lower levels.

Inductive biases for iconicity

Which underlying cognitive structures support our inferences about iconicity? We asked naive subjects in a guessing game to pick the most likely referent for each signal. Figure 5 shows the probability of listeners choosing a complex referent as a function of signal complexity, indicating a strong tendency for choosing referents that match the perceived complexity of the signal. (Note that the ground truth referent information in Figure 5 is displayed as additional information and not part of the reported analysis.) Crucially, this bias allowed subjects to reliably identify the correct referent complexity class for signals that exhibited the most iconicity ($t(9) = 2.77, p = 0.024$). We compared the probability of choosing the correct referent class with chance performance for the ten languages that scored highest in iconicity (measure taken from Figure 2D).

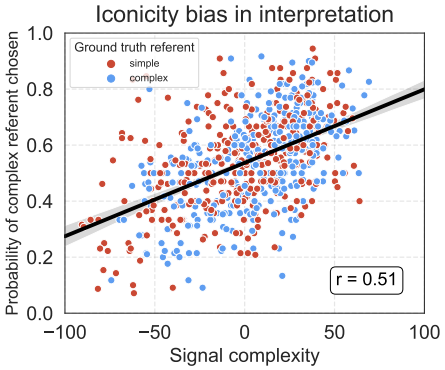


Figure 5: Results from the guessing game, conducted to quantify the strength of iconicity biases for the evolved signals.

Discussion and Conclusion

We conducted an iterated learning experiment to investigate how iconicity develops when combinatoriality emerges in initially unstructured, continuous signals. Signals gradually became more combinatorial over the course of the experiment. The emergence of iconicity, measured in terms of signals matching the conceptual complexity of their referents, was shown to develop immediately, but iconicity eventually disappeared while combinatorial structure continued to increase. This result is particularly strong because languages were initialized at a point of complete arbitrariness.

The largest increases in combinatoriality were seen in generation one. This is consistent with the idea that signal structure is a consequence of cognitive biases for combinatoriality. Because of having selected a diverse signal set for initialization, memory demands during learning and reproduction were arguably the highest in the first generation. Since this affords the greatest potential for prior biases to manifest, we would expect the largest increase in structure here.

The loss of iconicity is consistent with the hypothesis that iconicity is in complementary distribution with combinatorial structure (Verhoef et al., 2016; Goldin-Meadow & McNeill, 1999; Roberts, Lewandowski, & Galantucci, 2015), since the building blocks of such a system must be stripped of their iconicity when they participate in larger meaning-bearing units. It is important to note, however, that it is not clear why the particular kind of iconicity we investigated here must be lost in order for combinatorial structure to arise. In the transition from holistic to combinatorial structure, complexity in the signal domain could be expressed equally well in terms of number of building blocks (Lewis & Frank, 2016). The observation that iconicity is nevertheless lost could, however, provide important insight into the nature of the transition process. Zuidema and de Boer (2018) recently distinguished between analytic and synthetic routes to combinatoriality. In the synthetic route, preexisting signals are combined to form larger combinatorial signals, while in the analytic route, potentially overlapping parts of preexisting signals are used to form new signal. The present findings are consistent with the holistic account, which predicts that productive recombination leads to new signals that are composite, therefore complex, irrespective of the complexity of their referent but simply due to the mechanics of recombination.

While our guessing game suggests that people have strong biases for iconicity, our results indicate that these biases manifest in subtle ways. Smith et al. (2017) presented evidence that the strength with which cognitive biases manifest in cultural evolution depend on a number of factors. The authors focused on properties of the transmission paradigm, such as how many different agents subjects learn from, which shapes the input to learning. In the present study, we found evidence that the manifestation of otherwise strong cognitive biases, such as a bias for iconicity, can also depend on properties of the input more directly, for instance, on how much structure signals exhibit. Understanding how properties of the input can modify the expression of biases more broadly is an avenue for future research. In addition, the novel vector field analysis we present suggests the possibility of testing specific combinations of iconicity and combinatoriality to develop a more complete picture of trade-offs in parameter space.

One further aspect that is not addressed in our study is the role of modality on the form of iconicity studied here. In work that explored the structure of signals that subjects created when more or less signal dimensions were available, Little, Eryılmaz, and de Boer (2017) found strong modality effects mediating the relationship between iconicity and combinatoriality. Future work is needed to explore how our findings generalize to other signal modalities.

Finally, we note that the combinatoriality measure obtained via subject ratings is only an approximation to signal structure that emerged in the experiment. To better understand the patterns that exist in the evolved signals and how they are used productively, it is necessary to develop computational models. In ongoing work, we are developing statistical models of signal structure and learners' underlying inductive biases that will allow us to test more specific hypotheses about the evolution of iconicity and combinatoriality in language.

References

Christiansen, M. H., & Chater, N. (2016). *Creating language: integrating evolution, acquisition, and processing*. Cambridge, MA: The MIT Press.

Dingemanse, M., Blasi, D. E., Lupyan, G., Christiansen, M. H., & Monaghan, P. (2015). Arbitrariness, Iconicity, and Systematicity in Language. *Trends in Cognitive Sciences*, *19*(10), 603–615.

Fay, N., Arbib, M., & Garrod, S. (2013). How to Bootstrap a Human Communication System. *Cognitive Science*, *37*(7), 1356–1367.

Giudice, A. D. (2012). The emergence of duality of patterning through iterated learning: Precursors to phonology in a visual lexicon. *Language and Cognition*, *4*(04), 381–418.

Goldin-Meadow, S., & McNeill, D. (1999). The Role of Gesture and Mimetic Representation in Making Language the Province of Speech..

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, *105*(31), 10681–10686.

Lewis, M. L., & Frank, M. C. (2016). The length of words reflects their conceptual complexity. *Cognition*, *153*, 182–195.

Little, H., Eryılmaz, K., & de Boer, B. (2017). Signal dimensionality and the emergence of combinatorial structure. *Cognition*, *168*, 1–15.

Roberts, G., Lewandowski, J., & Galantucci, B. (2015). How communication changes when we cannot mime the world: Experimental evidence for the effect of iconicity on combinatoriality. *Cognition*, *141*, 52–66.

Smith, K., Perfors, A., Fehér, O., Samara, A., Swoboda, K., & Wonnacott, E. (2017). Language learning, language use and the evolution of linguistic variation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1711), 256-278.

Verhoef, T., Kirby, S., & de Boer, B. (2014). Emergence of combinatorial structure and economy through iterated learning with continuous acoustic signals. *Journal of Phonetics*, *43*, 57–68.

Verhoef, T., Kirby, S., & de Boer, B. (2016). Iconicity and the Emergence of Combinatorial Structure in Language. *Cog-*

nitive Science, *40*(8), 1969–1994.

Woods, K. J. P., Siegel, M. H., Traer, J., & McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, & Psychophysics*, *79*(7), 2064–2072.

Zuidema, W., & de Boer, B. (2018). The evolution of combinatorial structure in language. *Current Opinion in Behavioral Sciences*, *21*, 138–144.

Separating object resonance and room reverberation in impact sounds

Jennifer Hu
Massachusetts Institute of Technology, Cambridge, Massachusetts, United States

James Traer
Massachusetts Institute of Technology, Cambridge, Massachusetts, United States

Josh McDermott
Massachusetts Institute of Technology, Cambridge, Massachusetts, United States

Abstract

Everyday hearing requires inferring the causal factors that produce a sound, as when we separate the acoustic effects of the environment (reverberation) from those of sound sources. Here we consider perceptual inferences from impact sounds, in which the resonance of a struck object provides cues to its material, but via acoustic effects that might be nontrivial to disentangle from reverberation. We investigated whether and how humans separate the effects of object resonance and reverberation in a material classification task. For comparison, we implemented a Bayesian observer that inferred material from a generative model of object sounds without reverberation. Humans were robust to reverberation, whereas the model was not. However, human robustness was specific to reverberation consistent with the statistics of natural environments. The results suggest that humans use internal models of room and object acoustics to determine their respective contributions to sound, providing an example of causal inference in audition.

Dark Forces in Language Comprehension: The Case of Neuroticism and Disgust in a Pupillometry Study

Isabell Hubert (isabell.hubert@ualberta.ca)
Department of Linguistics, University of Alberta
Edmonton, AB, Canada T6G 2E7

Juhani Järvikivi (jarvikivi@ualberta.ca)
Department of Linguistics, University of Alberta
Edmonton, AB, Canada T6G 2E7

Abstract

We report on initial findings from a pupillometry study that investigated the influence of two extra-linguistic variables, namely Neuroticism and Disgust Sensitivity, on auditory language comprehension in adults. Results suggest that: (1) Language comprehension is influenced by extra-linguistic variables and individual differences; (2) the processing of different kinds of linguistic errors, as opposed to clashes with an individual's value or belief system, are influenced by different extra-linguistic variables; and that (3) Disgust Sensitivity at least partially predicts pupillary responses to utterances clashing with an individual's belief system. Results are discussed with regards to linguistic anticipation, cognitive effort and arousal, and resource allocation.

Keywords: psycholinguistics; extra-linguistic information; individual differences; pupillometry; language comprehension; personality; Disgust; neuroticism

Introduction

The field of linguistics has not traditionally focused on what is known as individual differences, or *hot cognition* - for example, emotion or personality. Instead, the focus has been on "abstracting away" or "averaging over" individual differences to be able to make inferences about a population. However, listeners appear to use the preceding discourse and their knowledge of the world immediately to interpret language (Van Berkum, Brown, Zwitserlood, Kooijman, & Hagoort, 2005), and extra-linguistic influences, such as personality, mood, or accent, are not just "noise," but can help reveal new information about language comprehension processes (Van den Brink et al., 2010). An individual's personality has also been found to influence aspects of general cognition and daily life, such as academic motivation and the choice of learning style (Busato, Prins, Elshout, & Hamaker, 1998; Jensen, 2015); use of language (Oberlander & Gill, 2004; Pennebaker, Mehl, & Niederhoffer, 2003); response to written errors (Boland & Queen, 2016); speech production in both native speakers and second language learners (Dewaele & Furnham, 2000); and the use of online social media (Park et al., 2015; Wehrli et al., 2008).

Results from experimental psycho-linguistic studies indicate that utterances such as "the girl comforted the clock" can be non-anomalous if the context warrants such an interpretation, from which Nieuwland and Van Berkum (2006) conclude that context can overrule grammatical violations. This is not strictly possible in a purely bottom-up model of language comprehension, where integration with the real world is thought to happen at a later stage. Research instead suggests that the language comprehension process involves at least some level of top-down processing, with contextual information rapidly being integrated into language comprehension (Kamide, Altmann, & Haywood, 2003; Levy, 2008; Sedivy, Tanenhaus, Chambers, & Carlson, 1999; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995; Traxler, 2014; Van Berkum et al., 2005).

Van Berkum, Van den Brink, Tesink, Kos, and Hagoort (2008) analyzed ERP responses to statements colliding with a speaker's perceived identity, such as an adult male announcing that he wished he looked like Britney Spears, and found that such statements, clashing with Dutch stereotypes based on age, class, or gender, reliably elicited an N400 component. This component is generally elicited by all content words, but is significantly larger in amplitude for items that are difficult to integrate into the preceding context (Allen, Badecker, & Osterhout, 2003; Kutas & Federmeier, 2007).

Van Berkum, Holleman, Nieuwland, Otten, and Murre (2009) found that statements clashing with an individual's value system, such as "I think euthanasia is an acceptable..." when the participant opposed this practice, elicited a distinctive ERP response just 200-250ms after the onset of the critical word, in addition to an N400 component. These results suggest that, in addition to inferences about the speaker, the *listener's* values and beliefs also play a role in language comprehension.

Van den Brink et al. (2010) found that listeners with high empathy levels showed a significantly larger N400 component in response to socially contradictory information than those with low empathy scores, reasoning that the ability to empathize to a higher degree may encourage a more top-down behaviour, engaging in more prediction based on inferences about the speaker – and hence experiencing surprisal at an unexpected item.

Mood, a more transitory state than personality traits, was also found to affect language comprehension in an implicit causality experiment (Van Berkum, De Goede, Van Alphen, Mulder, & Kerstholt, 2013). Implicit causality verbs, such as "praise" or "apologize," bias participants as to which of the noun phrases in the sentence is the likely "cause" of an event (Pyykkönen & Järvikivi, 2010). A good mood caused listeners to engage in more prediction as to what the referent might be. This was reflected in a distinctive ERP component

in response to a bias-inconsistent continuation; a bad mood, on the other hand, effectively stifled anticipation. Even a *simulated* mood (Havas, Glenberg, & Rinck, 2007) appears to affect processing speed, such that processing is faster when an individual's simulated facial expression matched the valence of the sentence.

Summing up, recent research suggests that individual differences such as personality, mood, and world view affect language processing from a very early stage, and not only at a later stage, in what used to be considered a secondary step, referred to as "real-world integration."

In this paper, we report on initial findings from a pupillometry study that investigated auditory language comprehension in adults, correlating their pupil sizes in response to sentences (non-anomalous. vs. those containing errors or clashes) with measures of Neuroticism and Disgust Sensitivity. Pupil size is considered a non-invasive measure of autonomous nervous system activity (Partala & Surakka, 2003) that is especially responsive to – beyond ambient light levels – cognitive effort, mental workload, attention, and arousal (Gingras, Marin, Puig-Waldmüller, & Fitch, 2015; Goldinger & Papesh, 2012; Just & Carpenter, 1993; Rondeel, Van Steenbergen, Holland, & van Knippenberg, 2015), and that is largely free of task effects. In an auditory experiment with linguistic stimuli, pupil dilation can thus be used as an indicator of the intelligibility and complexity of an utterance (Ben-Nun, 1986; Lõo, van Rij, Järvikivi, & Baayen, 2016; Zekveld, Kramer, & Festen, 2010).

Disgust Sensitivity has, to our knowledge, not yet been investigated with regards to language comprehension. However, being considered one of the most primitive emotions that, for example, serves to protect and organism from novel pathogens, has been found to be strongly linked to feelings of morality, purity, political orientation, and voting behaviour (Inbar, Pizarro, Iyer, & Haidt, 2011; Smith, Oxley, Hibbing, Alford, & Hibbing, 2011). Higher Disgust Sensitivity is generally linked to a more conservative approach, relying more on established socio-cultural stereotypes rather than novel, more liberal ideas.

A proposed tie-in of language processing with cognition more generally comes from Havas et al. (2007), who relate their results to theories in which emotions are assumed to change affordances, the links between perception and action: In this view, a positive mood prepares the body to approach, whereas a negative mood prepares the body to avoid. Under this account, mood and personality could be assumed to influence how strongly a human engages in "approaching" or "exploring", or how much they stay put and rely on bottom-up information. A related take can be found in the *bio-energetic account*, which suggests that emotional states signal the amount of cognitive resources available for more "costly" behaviours, such as exploration and anticipation (Zadra & Clore, 2011; Van Berkum et al., 2013).

We show below that both Disgust Sensitivity and Neuroticism, as two extra-linguistic variables and components of an individual's world view that are not typically investigated in regards to language processing, indeed influence automatic language comprehension processes even in the absence of a conscious judgment or task.

Main Experiment

240 sentences were constructed, 32 of which were unrelated filler sentences. Clashes were distributed among the following conditions (examples are given in Table 1):

MO: 56 sentences total; 28 of which violated subject-verb agreement, resulting in a morpho-syntactic error;

SE: 32 sentences total; 16 of which created a semantic mismatch between the verb and the object, resulting in a semantic error;

SC: 120 sentences total; 60 of which clashed with established gender stereotypes, and as such the speaker's perceived identity; resulting in a socio-cultural clash (Van Berkum et al., 2008; Van den Brink et al., 2010).

Clash type	Example
MO	She usually **drive** her car slowly in the snow.
SE	People often read **heads** for pleasure at night.
S-C	♂I buy my **bras** at Hudson's Bay.

Table 1: The template used for item construction, with three example sentences.

All stimuli followed the same syntactic pattern to ensure comparability. Frequency of the critical region, i.e. the main verb plus the direct or oblique object directly following the verb, was controlled for frequency via the *Corpus of Contemporary American English (COCA)* (Davies, 2008).

Items were then recorded by one male and one female native speaker of Canadian English and distributed across four lists of just over 130 items each, counterbalanced for error condition and speaker gender.

82 participants, recruited from the university's undergraduate Linguistics pool and from the general population, participated in this experiment. Data from six participants (7% of all participants) was removed as their comprehension question accuracy rates were below 80%, and comprehension or attention to the experiment could hence not be guaranteed; or as information given on the language background questionnaire precluded their data from inclusion in the analyses. Data from 728 trials (8.6% of all trials) was removed due to issues during recording that resulted in more than 33% of sampling points on a given trial being recorded as NA. Thus, analyses below are based on the data from 76 participants (males/females = 18/58; native/non-native speakers of English = 61/15; age = 1783; mean [SD] = 25 [12.6] years).

Each participant was presented with one list and, accordingly, each item only once, in just one condition and spoken by one of the speakers. All items were previously rated for acceptability in a separate experiment, by a separate set of

participants, with the resulting average per-item ratings being fed into the statistical models below as a numerical predictor. The distribution of Big Five trait scores within the participants (raters) in this separate norming study was found to be in line with several others reported in the literature, such as those found in Srivastava, John, Gosling, and Potter (2003) and Schmitt and Shackelford (2008).

Each trial began with a one-point drift correct, and, immediately after, the display of a fixation cross at the centre of the screen. The size of the participant's right pupil was recorded at 250Hz, using an EyeLink 1000 system on a desktop PC, from that the start of the fixation cross onwards. 2000ms later, the audio stimulus began to play, with the latter half of this interval used to create participant-by-trial baselines. Pupil size was recorded until 500ms after audio offset. After a three-second break, in which participants were able to rest their eyes, the next trial began. Attention and comprehension were assessed via simple questions after approximately every fourth trial, and participants were given longer breaks approximately every thirty-five trials.

Post-Tests

Personality traits were assessed via the *Big Five* (John, Donahue, & L., 1991) personality inventory. The Big Five Inventory was chosen for its frequent and continued use in psychological research, and because it assesses various aspects of an individual's personality rather than just providing one overall score. Of special interest for this paper is the Neuroticism subscale, where high scores are typically associated with a higher tendency to feel anxious, nervous, or tense, and where low scores in contrast are associated with a more even temper (John et al., 1991), as these variables have traditionally been underresearched in regards to language processing.

The participants' Disgust Sensitivity was assessed via the Disgust Scale - Revised (*DS-R*) (Haidt, McCauley & Rozin, 1994, modified by Olatunji et al., 2007), also used in Ahn et al. (2014). Special interest is given to these two particular scales as prior research has largely focused on the "lighter," more positively loaded aspects of human personality and cognition, such as empathy. Data on the participants' language background was collected via a pen-and-paper questionnaire that included questions on items such as age, gender, and languages spoken.

Prior research has reported systematically higher Disgust Sensitivity among women as compared to men (Al-Shawaf, Lewis, & Buss, 2018; Sparks, Fessler, Chan, Ashokkumar, & Holbrook, 2018). In this study, only a non-significant tendency in this same direction was found in a two-sample t-test ($mean_{male} = 1.78, SD_{male} = 0.68; mean_{female} = 2.06, SD_{female} = 0.58; t(28.678) = -1.62, p = 0.12$).

Results

The raw pupillometry data was first downsampled to 125 Hz and then preprocessed in `R`. Blinks and the 20 adjacent data points were removed using Jacolien van Rij's `removeBlinks()` function.

Pupil sizes as the response variable were modelled using generalized additive mixed effects modelling (*GAM modelling*, or *GAMM*) with the `itsadug` (van Rij, Wieling, Baayen, & van Rijn, 2016) package in R. All models included random slopes for participant-by-time, and random intercepts by item, to account for individual differences within the stimuli, and for random variance between participants *beyond* the factors of interest. This makes the analyses markedly different from e.g. Van den Brink et al. (2010); Van Berkum et al. (2008), as GAM modelling can capture non-linear interactions between continuous predictors; as it does not assume linear relationships, an assumption that is often unwarranted (Tremblay & Newman, 2015); and as it allows to control for random participant and item effects. Additionally, GAMMs can comfortably model continuous measurements data, such as those obtained in pupillometry studies, without losing information in time-binning or averaging. GAM modelling has been used successfully in experimental psycholinguistics to model the influence of listener experience and the perception of foreign accents (Porretta, Tucker, & Järvikivi, 2016), and pupillary responses in a naming task (Lõo et al., 2016).

Data in a time window from 500ms before clash onset to 2000ms after clash onset was analyzed, and models included variables such as speech rate and the participant's progress in the experiment as control variables. Additionally, all numerical predictors were normalized and centered to avoid effects of differential order-of-magnitude scaling between predictors.

Morpho-Syntactic & Semantic Errors

While neither Neuroticism or Disgust were found to significantly influence the processing of semantic errors, Neuroticism was a significant individual difference predictor in a three-way interaction with time and item rating in the morpho-syntactic error model (dev. explained = 9.94%; see

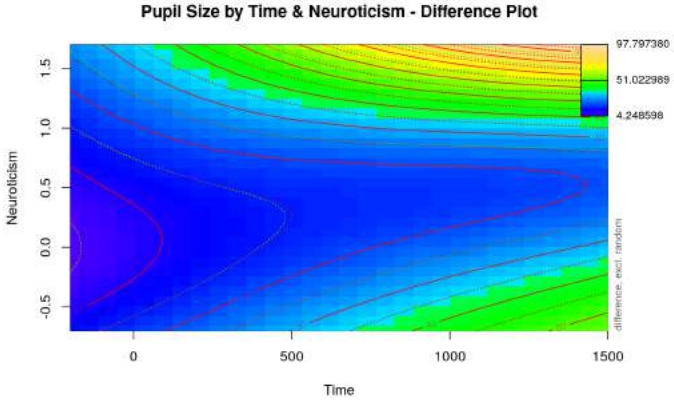


Figure 1: Difference in pupil size between the correct and clashing conditions in response to morpho-syntactic errors.

Parametric coefficients	*Estimate*	*Std. Error*	*t-value*	*p-value*
(Intercept)	-38.8152	21.1778	-1.8328	0.0668
Smooth terms	*edf*	*Ref.df*	*F-value*	*p-value*
Speech rate	2.9885	2.9998	261.8398	< 0.0001
Trial count	2.9952	3.0000	4055.6671	< 0.0001
Item rating	2.9942	2.9999	476.5416	< 0.0001
Neuroticism	1.0000	1.0000	0.0515	0.8205
Time x rating	15.9104	15.9977	255.7946	< 0.0001
Neur. x time	1.0025	1.0030	0.2502	0.6172
Neur. x rating	15.7562	15.9621	108.7371	< 0.0001
Neur. x time x rating	62.3328	63.7663	88.6382	< 0.0001
Random structure				
Participant x time	673.7516	682.0000	537.2140	< 0.0001
Item	101.6891	102.0000	279.8801	< 0.0001

Table 2: Model output for morpho-syntactic errors. Note that all numerical predictors, except time, were scaled and centered.

Parametric coefficients	*Estimate*	*Std. Error*	*t-value*	*p-value*
(Intercept)	-25.4912	17.3276	-1.4711	0.1413
B. Smooth terms	*edf*	*Ref.df*	*F-value*	*p-value*
Speech rate	2.9859	2.9998	259.0271	< 0.0001
Trial count	2.9981	3.0000	4287.4004	< 0.0001
Item rating	2.9976	2.9999	673.4162	< 0.0001
Disgust	1.0049	1.0049	0.1383	0.7087
Time x rating	15.8359	15.9884	397.1878	< 0.0001
Disgust x time	3.0779	3.1676	1.6172	0.1174
Disgust x rating	15.8861	15.9860	168.5608	< 0.0001
Disgust x time x rating	62.6458	63.7768	94.6083	< 0.0001
Random structure				
Participant x time	647.7298	666.0000	662.3551	0.0116
Item	101.7037	102.0000	343.1954	< 0.0001

Table 3: Model output for socio-cultural clashes. Note that all numerical predictors, except time, were scaled and centered.

also Table 2).[1]

This three-way interaction shows that different Neuroticism scores are correlated with different changes in pupil sizes, which differ further between correct and anomalous items. Specifically, our findings indicate that high Neuroticism scores led to a much stronger pupillary response to morpho-syntactic errors as compared to low scores on this scale (cf. Fig. 1. Like all surface plots in this paper, this plot visualizes the difference in pupil size by time since clash onset (on the x-axis) and Neuroticism scores (on the y-axis) between the clashing and correct conditions. Pupil size is represented as colour on the z-axis: A blue/green hue indicates smaller a smaller difference in pupil sizes, and yellow/orange indicates larger dilation in the clashing compared to the correct condition).

Further significant main effects in this model include those of speech rate (faster $\rightarrow$ larger dilation), progress made in the experiment (early trials $\rightarrow$ larger dilation), item rating (lower $\rightarrow$ larger dilation), and Neuroticism (higher $\rightarrow$ larger dilation). It should be noted that the significant three-way interaction between Neuroticism, item rating, and time was found to be significant *beyond* these main effects, and beyond the random effects included in the model.

Socio-Cultural Clashes

In the modelling of socio-cultural errors, Disgust Sensitivity was found to be the single best individual difference predictor tested in an interaction with time and item rating (dev. explained = 9.65%; see also Table 3):[2] High values, indicating high Disgust Sensitivity, were found to correlate with the

[1]The remaining Big Five traits were tested as well; while elaborating on all results goes beyond the scope of this current paper, Agreeableness was found to be a significant predictor in this same three-way interaction in a model of equally good fit, with low Agreeableness associated with larger differences in pupil sizes.

[2]All Big Five traits were tested here as well; for socio-cultural clashes, Disgust emerged as the single best extra-linguistic predictor.

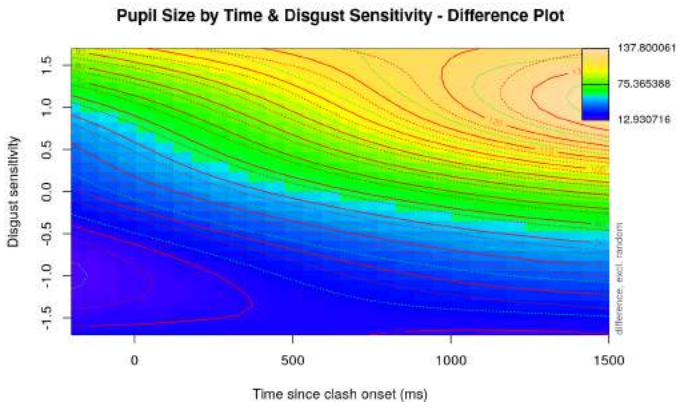


Figure 2: Difference in pupil size between the correct and clashing conditions in response to socio-cultural clashes.

largest pupil dilations in response to statements clashing with socio-cultural stereotypes (cf. Fig. **??**).

Further significant main effects include those of speech rate (faster → larger dilation), progress made in the experiment (early trials → larger dilation), item rating (lower → larger dilation), and Disgust Sensitivity (higher → larger dilation). Again, the significant three-way interaction between Disgust Sensitivity, item rating, and time was found to be significant *beyond* the main effects.

Discussion

Our results, specifically the three-way interactions including one of the two extra-linguistic variables, suggest that the processing of morpho-syntactic errors on the one hand, and stereotype-based clashes on the other, are influenced by individual differences and extra-linguistic information. Patterns of influence are not the same across the board, but instead are distinct between different types of errors and clashes. As an example, Neuroticism seemed to only influence the processing of morpho-syntactic errors, whereas Disgust Sensitivity best predicted pupillary responses to socio-cultural clashes. Neither of those two negatively loaded variables of individual differences were found to be significant predictors of pupil size in response to semantic errors.

These results lend further support to theories of language comprehension in which extra-linguistic information is considered early in the comprehension process (Kamide et al., 2003; Levy, 2008; Sedivy et al., 1999; Tanenhaus et al., 1995; Traxler, 2014; Van Berkum et al., 2005), and are not explained well within purely bottom-up theories: Larger pupil dilations for Disgust-sensitive individuals in response to socio-cultural clashes suggest that a statement that is at odds with one's expectations of purity and morality, and that hence triggers a visceral reaction, results in higher levels of arousal and/or requires more cognitive resources to "unpack." In this reading, extra-linguistic variables internal to the listener, such as feelings towards or the desire for purity and morality, affect the comprehension process right from the start, instead of being integrated with the sentence in a later step.

Considering the effect of Neuroticism on the processing of simple morpho-syntactic errors, our results add further support to models that include a top-down component; They also support the notion that, very generally, one's personality affects language comprehension, and that language comprehension does not take place in a vacuum (Van Berkum et al., 2008, 2009). Specifically, our results suggest that individuals that are more prone to feelings of anxiety or nervousness may experience greater distress when experiencing a simple grammatical error. Of note is that morpho-syntactic errors do not clash with experiences or value systems as such, but only violate intra-linguistic rules; This suggests that the listener's personality seems to affect linguistic processing even when the utterance in question does not directly require value judgments or beliefs to process.

Building on Ahn et al. (2014); Inbar et al. (2011); Smith et al. (2011), our results suggest that Disgust Sensitivity at least partially correlates with sensitivity towards stimuli that, as per existing cultural stereotypes, may be associated more with a progressive and liberal view of the world, and that may trigger stronger reactions in conservative individuals. In addition to further supporting models of language comprehension in which context and experience factor significantly early on, this also meshes with the idea of Disgust serving as a mechanism protecting the individual from novel pathogens carried by members of an out-group: Utterances indicative of out-group status appear to trigger higher levels of arousal, and/or demand more cognitive resources to process.

Within the context of affordances and the bio-energetic account (Havas et al., 2007; Zadra & Clore, 2011; Van Berkum et al., 2013), our results do not neatly tie in with previous research: They suggest that higher Disgust Sensitivity and higher Neuroticism scores may be associated with *more* context-based prediction and anticipation, and hence more surprisal at an unexpected continuation, than lower scores on these scales. These somewhat counter-intuitive results warrant further research, as prior studies have generally found *positive* emotions and moods, such as empathy or an elevated mood, to be associated with more resource availability, prediction, and exploration (Van den Brink et al., 2010).

It should be noted that our results should not necessarily be interpreted as a causal relationship, in that different values of Neuroticism or Disgust Sensitivity "trigger" differences in processing. It is conceivable that a common underlying variable, relating to e.g. resource allocation or to a general predisposition towards other-ness, is causing the effects.

In this fairly new field of research, there is lots of room for both broader and more targeted investigations; we are currently investigating the effects of other extra-linguistic variables, such as the remaining Big Five traits, on language comprehension.

More broadly, future research could, for example, assess the effects of extra-linguistic variables using additional

methodologies, or clashes with different aspects of the listener's identity. Research along those lines may be able to form a more coherent picture, for example in regards to whether it is anticipation or prediction that is modulated by a certain variable, or whether there may be an additional underlying variable that influences both a listener's personality and Disgust Sensitivity, *and* their linguistic processing at the same time.

Summing up, our results add further support to models of language comprehension that include a top-down component, and to extra-linguistic information and individual differences factoring in language comprehension from a very early stage; and they assessed the influence of Disgust Sensitivity as a "darker" cognitive force on language comprehension for the first time.

Acknowledgments

This research was supported by a Social Sciences and Humanities Research Council of Canada (http://www.sshrc-crsh.gc.ca/) Partnership grant (*Words in the World*, 895-2016-1008).

References

Ahn, W. Y., Kishida, K. T., Gu, X., Lohrenz, T., Harvey, A., Alford, J. R., ... Montague, P. R. (2014). Nonpolitical images evoke neural predictors of political ideology. *Current Biology*, *24*(22), 2693–2699. Retrieved from http://dx.doi.org/10.1016/j.cub.2014.09.050 doi: 10.1016/j.cub.2014.09.050

Allen, M., Badecker, W., & Osterhout, L. (2003). Morphological analysis in sentence processing: An erp study. *Language and Cognitive Processes*, *18*(4), 405–430.

Al-Shawaf, L., Lewis, D. M., & Buss, D. M. (2018). Sex Differences in Disgust: Why Are Women More Easily Disgusted Than Men? *Emotion Review*, *10*(2), 149–160. doi: 10.1177/1754073917709940

Ben-Nun, Y. (1986). The use of pupillometry in the study of on-line verbal processing: Evidence for depths of processing. *Brain and Language*, *28*(1), 1–11.

Boland, J. E., & Queen, R. (2016). If youre house is still available, send me an email: Personality influences reactions to written errors in email messages. *PloS one*, *11*(3), e0149885.

Busato, V. V., Prins, F. J., Elshout, J. J., & Hamaker, C. (1998). The relation between learning styles, the big five personality traits and achievement motivation in higher education. *Personality and individual differences*, *26*(1), 129–140.

Davies, M. (2008). *The corpus of contemporary american english: 520 million words, 1990-present.* Available online at http://corpus.byu.edu/coca/.

Dewaele, J.-M., & Furnham, A. (2000). Personality and speech production: a pilot study of second language learners. *Personality and Individual differences*, *28*(2), 355–365.

Gingras, B., Marin, M. M., Puig-Waldmüller, E., & Fitch, W. (2015). The eye is listening: Music-induced arousal and individual differences predict pupillary responses. *Frontiers in human neuroscience*, *9*.

Goldinger, S. D., & Papesh, M. H. (2012). Pupil dilation reflects the creation and retrieval of memories. *Current Directions in Psychological Science*, *21*(2), 90–95.

Haidt, McCauley & Rozin, 1994, modified by Olatunji et al. (2007). *The ds-r.*

Havas, D. A., Glenberg, A. M., & Rinck, M. (2007). Emotion simulation during language comprehension. *Psychonomic Bulletin & Review*, *14*(3), 436–441.

Inbar, Y., Pizarro, D., Iyer, R., & Haidt, J. (2011). Disgust Sensitivity, Political Conservatism, and Voting. *Social Psychological and Personality Science*. doi: 10.1177/1948550611429024

Jensen, M. (2015). Personality traits, learning and academic achievements. *Journal of Education and Learning*, *4*(4), 91.

John, O. P., Donahue, E. M., & L., K. R. (1991). *The Big Five Inventory–Versions 4a and 54*. University of California, Berkeley, Institute of Personality and Social Research. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0092656603000461 doi: 10.1016/S0092-6566(03)00046-1

Just, M. A., & Carpenter, P. A. (1993). The intensity dimension of thought: pupillometric indices of sentence processing. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *47*(2), 310.

Kamide, Y., Altmann, G. T., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and language*, *49*(1), 133–156.

Kutas, M., & Federmeier, K. D. (2007). Event-Related brain potential (ERP) studies of sentence processing. In *Oxford handbook of psycholinguistics* (pp. 385–406). doi: 10.1093/oxfordhb/9780198568971.013.0023

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.

Lõo, K., van Rij, J., Järvikivi, J., & Baayen, H. (2016). Individual differences in pupil dilation during naming task. In *Proceedings of the 38th annual conference of the cognitive science society, a. papafragou, d. grodner, d. mirman, and j. trueswell, eds. austin, tx: Cognitive science society* (pp. 550–555).

Nieuwland, M. S., & Van Berkum, J. J. (2006). When peanuts fall in love: N400 evidence for the power of discourse. *Journal of cognitive neuroscience*, *18*(7), 1098–1111.

Oberlander, J., & Gill, A. J. (2004). Individual differences and implicit language: personality, parts-of-speech and pervasiveness. In *Proceedings of the cognitive science society* (Vol. 26).

Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., ... Seligman, M. E. (2015).

Automatic personality assessment through social media language. *Journal of personality and social psychology*, *108*(6), 934.

Partala, T., & Surakka, V. (2003). Pupil size variation as an indication of affective processing. *International journal of human-computer studies*, *59*(1), 185–198.

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology*, *54*(1), 547–577.

Porretta, V., Tucker, B. V., & Järvikivi, J. (2016). The influence of gradient foreign accentedness and listener experience on word recognition. *Journal of Phonetics*, *58*, 1–21.

Pyykkönen, P., & Järvikivi, J. (2010). Activation and persistence of implicit causality information in spoken language comprehension. *Experimental Psychology*, *57*(1), 5–16.

Rondeel, E. W., Van Steenbergen, H., Holland, R. W., & van Knippenberg, A. (2015). A closer look at cognitive control: differences in resource allocation during updating, inhibition and switching as revealed by pupillometry. *Frontiers in human neuroscience*, *9*.

Schmitt, D. P., & Shackelford, T. K. (2008). Big Five Traits Related to Short-Term Mating: From Personality to Promiscuity across 46 Nations. *Evolutionary Psychology*, *6*(2), 147470490800600. doi: 10.1177/147470490800600204

Sedivy, J. C., Tanenhaus, M. K., Chambers, C. G., & Carlson, G. N. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition*, *71*(2), 109–147.

Smith, K. B., Oxley, D., Hibbing, M. V., Alford, J. R., & Hibbing, J. R. (2011). Disgust Sensitivity and the Neurophysiology of Left- Right Political Orientations. *PLoS ONE*, *6*(10). doi: 10.1371/journal.pone.0025552

Sparks, A. M., Fessler, D. M. T., Chan, K., Ashokkumar, A., & Holbrook, C. (2018). Disgust as a mechanism of decision making under risk. *Emotion*, *18*(7), 942–958.

Srivastava, S., John, O. P., Gosling, S. D., & Potter, J. (2003). Development of Personality in Early and Middle Adulthood: Set Like Plaster or Persistent Change? *Journal of Personality and Social Psychology*, *84*(5), 1041–1053. doi: 10.1037/0022-3514.84.5.1041

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 1632–1634.

Traxler, M. J. (2014). Trends in syntactic parsing: Anticipation, bayesian estimation, and good-enough parsing. *Trends in cognitive sciences*, *18*(11), 605–611.

Tremblay, A., & Newman, A. J. (2015). Modeling nonlinear relationships in erp data using mixed-effects regression with r examples. *Psychophysiology*, *52*(1), 124–139.

van Rij, J., Wieling, M., Baayen, R. H., & van Rijn, H. (2016). *itsadug: Interpreting time series and autocorrelated data using gamms.* (R package version 2.2)

Van Berkum, J. J., Brown, C. M., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating upcoming words in discourse: evidence from erps and reading times. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(3), 443.

Van Berkum, J. J., De Goede, D., Van Alphen, P. M., Mulder, E. R., & Kerstholt, J. H. (2013). How robust is the language architecture? the case of mood. *Frontiers in psychology*, *4*.

Van Berkum, J. J., Holleman, B., Nieuwland, M., Otten, M., & Murre, J. (2009). Right or wrong? the brain's fast response to morally objectionable statements. *Psychological Science*, *20*(9), 1092–1099.

Van Berkum, J. J., Van den Brink, D., Tesink, C. M., Kos, M., & Hagoort, P. (2008). The neural integration of speaker and message. *Journal of cognitive neuroscience*, *20*(4), 580–591.

Van den Brink, D., Van Berkum, J. J., Bastiaansen, M. C., Tesink, C. M., Kos, M., Buitelaar, J. K., & Hagoort, P. (2010). Empathy matters: Erp evidence for inter-individual differences in social language processing. *Social cognitive and affective neuroscience*, *7*(2), 173–183.

Wehrli, S., et al. (2008). Personality on social network sites: An application of the five factor model. *Zurich: ETH Sociology (Working Paper No. 7)*.

Zadra, J. R., & Clore, G. L. (2011). Emotion and perception: The role of affective information. *Wiley interdisciplinary reviews: cognitive science*, *2*(6), 676–685.

Zekveld, A. A., Kramer, S. E., & Festen, J. M. (2010). Pupil response as an indication of effortful listening: The influence of sentence intelligibility. *Ear and hearing*, *31*(4), 480–490.

Detecting social transmission in the design of artifacts via inverse planning

Ethan Hurwitz (ehurwitz@ucsd.edu),
Timothy F. Brady (timbrady@ucsd.edu),
Adena Schachner (schachner@ucsd.edu)
University of California, San Diego, Department of Psychology
9500 Gilman Drive M/C 0109, San Diego, CA 92093-0109 USA

Abstract

How do people use human-made objects (artifacts) to learn about the people and actions that created them? We test the richness of people's reasoning in this domain, focusing on the task of judging whether social transmission has occurred (i.e. whether one person copied another). We develop a formal model of this reasoning process as a form of rational inverse planning, which predicts that rather than solely focusing on artifacts' similarity to judge whether copying occurred, people should also take into account availability constraints (the materials available), and functional constraints (which materials work). Using an artifact-building task where two characters build tools to solve a puzzle box, we find that this inverse planning model predicts trial-by-trial judgments, whereas simpler models that do not consider availability or functional constraints do not. This suggests people use a process like inverse planning to make flexible inferences from artifacts' features about the source of design ideas.

Keywords: social cognition; Bayesian inference; explanation; social transmission; imitation; artifact; design; inverse planning

Introduction

We live surrounded by human-made objects, or artifacts. These artifacts are crucial to our lives not only as tools, but also as an omnipresent source of social information. Based on the objects a person owns, people make quick and accurate judgments about a person's traits, interests, and social affiliations (Gosling, 2008; Richins, 1994). The artifacts a person creates - like novel tools, art, music, or text - provide particularly rich information about the person and actions that created them (Gosling, 2008).

How do people reason about other individuals from the artifacts they create? Here we explore the nature of this reasoning, a form of *intuitive archeology*. In the same sense that archeologists use objects to make inferences about the people and cultures that created them, we propose that people also infer complex social-causal information from the design of artifacts, by integrating their mental theories of the physical-mechanical world with their theories of the social world (e.g. Battaglia, Hamrick & Tenenbaum, 2013; Gopnik, 2012; Baker, Saxe & Tenenbaum, 2009) to infer the most probable explanation for an objects' features.

Intuitive Archeology as Inverse Planning

Previous work in the domain of action understanding has proposed that people make inferences about the goals of others' actions based on a process of 'inverse planning' (Baker, Saxe, & Tenenbaum, 2009; Liu, Ullman, Tenenbaum & Spelke, 2018). The idea of inverse planning is that people have knowledge of the generative process behind actions from planning their own – and this planning process allows them to know what a rational agent would do, given the same goals and environmental constraints. Therefore, when reasoning about others' actions, people invert this generative process to infer the goals of another agent from its observed behaviors. Here we propose that a fundamentally similar inverse planning processing explains how we reason about the artifacts people create: People use their own generative model of how they would construct an artifact under a given set of constraints to infer the goals and decisions that led another person to create this artifact and its features. Such a reasoning process would allow people to flexibly infer a variety of social-causal information about others from the physical features of artifacts they create.

We focus on a foundational inference in this domain: Inferring whether *social transmission of ideas* has occurred (i.e. imitation, copying), or whether a particular aspect of a design was *generated independently* by an individual. The interaction of these two basic processes, termed imitation and innovation, account for cultural evolution of artifacts' designs over human history (Henrich, 2015; Tomasello, 1999; Legare & Neilsen, 2015). This inference also has real-world applications for understanding plagiarism detection – and what can be reasonably expected of jurors in plagiarism cases as they consider two designs and determine the likelihood that copying has occurred. Lastly, this inference is foundational to understanding how people infer social-causal information from artifacts, since designs that were created independently license different inferences than those that were copied. For example, a highly functional, complex design that was independently generated may tell you about the intelligence or creativity of a designer (Gosling, 2008), whereas a design that was copied may instead be informative about the designer's social history and cultural group (their source of shared knowledge; e.g. Schachner et al., 2018; Soley & Spelke, 2016). Thus, in the current work, we model and test how people infer whether or not copying (social transmission) occurred in the design of an artifact.

Inverse Planning, Or a Simpler Cognitive Process?

A natural alternative theory exists to the rich and structured explanation-based reasoning process proposed by inverse planning models. People may infer that copying occurred

using a simple heuristic based on perceptual similarity: If two things are more perceptually similar, then copying is more likely to have occurred. Notably, past work on detection of copying in music has relied on this type of simple similarity metric in formal models, to predict jury decisions in music plagiarism cases (Savage, Cronin, Müllensiefen, & Atkinson, 2018).

In contrast to these straightforward similarity-based models, other work has provided initial evidence that people detect copying via a more complex process of inverse planning or explanation- based reasoning (Schachner et al., 2018). In particular, this work found that people expect others to have a preference for efficiency, and factor this in when making inferences about copying. Thus, when two characters create identical train track designs that are also highly efficient ways to achieve the intended goal, observers use efficiency to 'explain away' the similarity – and thus judge copying less likely for identical efficient tracks than they would otherwise.

While this work is suggestive of a system of inverse planning, it is possible (and even plausible) that understanding of efficiency is unique and privileged in people's reasoning. Reasoning about efficiency, and expecting others to act rationally by moving efficiently toward their goals, is thought to be foundational to cognition: It develops early in infancy (Gergely, Nádasdy, Csibra, & Bíró, 1995, Skerry, Carey & Spelke, 2013), is shared with other species (Hauser & Wood, 2010), and is a foundation for the entire domain of action understanding (Dennett, 1987; Baker et al. 2009). Thus, rather than showing a rich and flexible process of reasoning that takes into account a wide variety of alternative explanations (as proposed by inverse planning models), the evidence thus far is consistent with a much simpler system, in which similarity metrics are selectively overridden by privileged efficiency-based explanations.

The Current Work

In the current work, we test whether people use a rich and flexible process of inverse planning that takes into account alternative explanations that go beyond efficiency. In particular, we ask whether people rationally consider two factors: the range of materials available to build with, which we term the *availability constraint*; and whether each of the available materials would function or fail to function to solve the problem at hand, which we term the *functional constraint*. Rationally speaking, if a larger set of materials are available to choose from, similarity should be seen as stronger evidence of copying than if there is a smaller set of materials available to choose from (as the probability of selecting the same item by chance is lower; similar to the suspicious coincidence mechanism sometimes referred to as the 'size principle'; Tenenbaum & Griffiths, 2001). Similarly, if many of these materials would solve the problem, similarity is more indicative of copying than if only one or a few of the options would solve the problem at hand – as clearly non-functional materials are unlikely to be used. We first formalize these constraints and then experimentally test their usage when people make copying inferences.

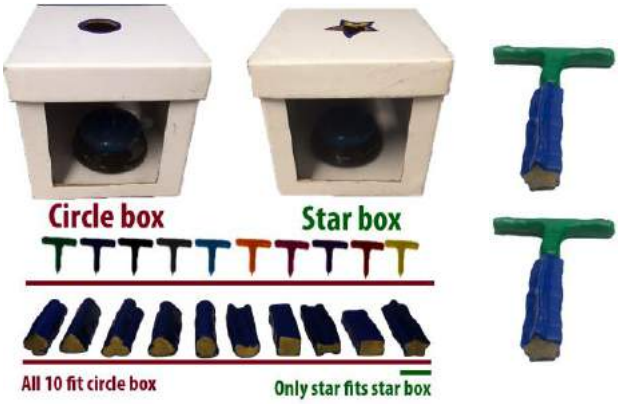


Figure 1: Left: Tool selection task with example handles (which differ in color), and rods (which differ in shape and therefore functionality). Right: Example of two identical tools people might be shown on a particular trial.

An Inverse Planning Model of Copy Detection

To provide a clear test of the inverse planning account, and tease it apart from simpler alternatives, we model and test a simple artifact-building task which crucially involved both availability and functional constraints. Consider a scenario where one is asked to solve a puzzle: A button is out of reach in a box, with the front covered by glass, so only the hole in the top allows access. You must build a tool to reach the button. To do so, you are given two sets of pieces: 10 handles, which differ by color; and 10 rods, which differ by shape. You can connect one handle to one rod to form a two-part tool (see Figure 1).

You may be asked to solve one of two puzzle boxes, which differ in one respect: How many of the rods would work to solve them. In particular, for one box, all of the 10 rods would fit through the box's circular hole and solve the puzzle (*unconstrained; circle box)*. In the other case, only 1 of the 10 rods fits (only the star-shaped rod fits into the star-shaped hole), and so only 1 of the 10 rods can be used to solve the puzzle (*constrained; star box*). This box thus introduces a functional constraint that applies selectively to rods, and not handles (which would all function in both cases).

Now, you observe two tools that other people have made: for example, two people built the same tool, choosing the same star-shaped rod and the same red handle. How likely are they to have copied each other? This task provides a simple instantiation of relevant issues people confront when making complex decisions about copying through inverse planning: Reasoning about the range of materials available to the builders; which pieces would work; and a multi-part decision process (choose a handle, choose a rod).

Formally, we can think of this task as having the following structure: You see a tool built by person 1, and a second tool built by person 2, in order to solve a puzzle box. You wish to

infer whether person 2 copied the tool's design from person 1, or independently created it.

Each tool consists of two pieces linked together – a rod, *r*, and a handle, *h* – each of which was selected from the set of available options. Formally, you are asked to make an inference, where if c indicates whether person 2 copied person 1, you wish to infer the probability of copying $P(c|r_1, h_1, r_2, h_2)$, given the observed rod and handle of person 1's tool (r_1, h_1) and the observed rod and handle of person 2's tool (r_2, h_2). Taking only the case of a rod being copied, and assuming copying judgments depend only on the rod and handle being identical or different (e.g., a binary notion of similarity), the posterior on copying is:

$$P(c \mid r_1, r_2) = \frac{P(c_r)\, P(r_2 == r_1 \mid c)}{P(r_2 == r_1)}$$

This is the probability that copying has occurred, given your prior likelihood on copying and the relative likelihoods that such an overlapping design would be generated under each of the possible mechanisms (copying, *c*, vs. independent creation, $\neg c$), where:

$$P(r_2 == r_1) = P(c_r)\, P(r_2 == r_1 \mid c) + \big(1 - P(c_r)\big)\, (r_2 == r_1 \mid \neg c)$$

In the current task this depends not only on the rod but on both the rod and handle, such that, when the rod is identical but the handle is not identical, this posterior on copying depends on $P(r_2 == r_1 \mid c, r_1)$, $P(r_2 == r_1 \mid \neg c)$, $P(h_2 \neq h_1 \mid c)$, and $P(h_2 \neq h_1 \mid \neg c)$. This has the structure of a Bayes net, including the key concept of explaining away: A given aspect of the design can be generated either via copying or independently, and evidence for one provides evidence against the other. Thus, if two people create identical tool designs, but this design is also likely to be created independently (due to either availability constraints or functional constraints), this provides weak evidence of copying despite the identical tools.

To make this model concrete, we need to specify 5 things:

(1) $P(c_r), P(c_h)$ - the a priori estimate of how likely person 2 was to have copied either the rod or handle (unconditional on the data; i.e. before we see either of the built objects). This depends for example on how close or distant the two people are from one another (Schachner et al., 2018). We assume the chance of copying is identical and independent for both rods and handles, e.g. $P(c_r) == P(c_h)$, and refer to this as $P(c)$, the prior on copying.

(2) $P(r_2 == r_1 \mid c)$ - the likelihood of the particular rod being used by person 2 matching that of person 1, given that person 2 was in fact copying from person 1's object. We formalize this as perfect copying plus a small error rate term, *e*, to account for the rate at which an individual might intend to copy but ultimately select a different rod: $P(r_2 == r_1 \mid c) = 1 - e$. Therefore $P(r_{2 \neq} r_1 \mid c) = e$.

(3) $P(r_2 == r_1 \mid \neg c)$ - the likelihood of rod r_2 being the same as r_1, given that person 2 was NOT copying from person 1's object, and independently generated the object with no reliance on r_1. When all pieces would function, this is simply *1/R*, where *R* is the total number of rod choices available. However, functional constraints also affect this factor: When only a subset of pieces will function, this effectively reduces the number of reasonable options. Accordingly, in the context of a functional constraint, the model treats only the functional pieces as options, reducing the value of R to the number of functional options (if only one rod functions, R=1).

4) $P(h_2 == h_1 \mid c)$ - the likelihood of the particular handle being generated by person 2, given that person 2 was in fact copying from person 1's object, and given h_1. This again is based on the same error rate *e*.

(5) $P(h_2 == h_1 \mid \neg c)$ - the likelihood of handle h_2 being the same as h_1, given that person 2 was NOT copying from person 1's object, and independently generated the object with no reliance on h_1. In contrast to the rods above, the handles differ only in color rather than shape; thus, all handles function equally well in both the *unconstrained (circle box)* condition, and the *functionally constrained (star box)* condition. This is therefore simply 1/H, where H is the number of handle options.

Comparing to Simpler Alternatives

This model of inference as inverse planning posits that people consider both the number of available options and the functional constraint of the puzzle box when judging whether copying occurred. To test whether each of these components are needed to predict participants' judgments, we compared this model to three simpler models.

These models followed a 2x2 structure, either taking into account or not taking into account the availability constraints (+/- availability) or the functional constraints (+/- functional). For example, the model that considers availability constraints but ignores functional constraints does not take into consideration the functional constraint of the star box, e.g., assumes people choose among all rods even in the star box condition. The model which ignored availability constraints did not take into account the number of pieces available in a flexible way. Instead, this model posited that people had a fixed a-priori idea of the number of pieces available to choose from, and that this number did not change based on the situation presented. Thus, rather than choose a rod with 1/R, where R is the number of options, a parameter N quantified this fixed number of imagined choices (e.g., regardless of how many were present). This model *did* take into account the functional constraint of the star box (assuming people only choose the star rod in this case). A final simplified model ignored both functional and availability constraints, and thus effectively instantiated a simple perceptual similarity heuristic. This model only took into account the extent to which the pieces were similar, without taking into consideration either functional constraints or availability constraints.

Testing the Models' Predictions

These models make quantitative predictions about the likelihood of copying for any given pair of tool designs, in a wide range of contexts. We next aimed to test how well the various models predict human behavior. The inverse

planning model predicts that for two identical tools, people will infer that copying is more likely to have occurred when (a) there were more pieces available as options to build with, thus creating more of a suspicious coincidence that the same piece was chosen twice; (b) there were no functional constraints on which pieces would work or not work, thus allowing all of the available pieces to serve as equally good options. By contrast, the simplest perceptual similarity model predicts that any identical objects will lead people to infer copying. Thus, we focused our data collection on these and other particularly informative trials.

Method

Full study design/analysis plan including model code was preregistered on the Open Science Framework (OSF), and is available at https://osf.io/y8u7t.

Participants

Using a pre-registered design, N=108 adults from the U.S. (57 male, 50 female, 1 other gender identity; *M* age=37.9, *SD*=10.9, range=20-72) were recruited through Amazon's Mechanical Turk. Sample size was preregistered and determined from power analysis of a pilot dataset with a slightly different design (N=20; tested a subset of the current test trials; with each subject completing all trials). The R "pwr" package was used to conduct a paired t-test power calculation on participant-level BICs with the goal of 90% power (Champely et al., 2018). Based on pre-registered exclusion criteria, additional participants were excluded due to: 1. Appearing to be non-native English speakers or a bot (n=13; determined by 2 independent coders' rating of free-response text answers) 2. Incorrectly answering any memory check question (n=49) 3. Incorrectly answering 50% or more of the attention check questions (n=12). The number of participants failing the preregistered memory check questions was higher than expected, thus we reanalyzed the data with these participants included, and found that our model results and conclusions remain unchanged in this case (see *Results*).

Design

Participants were shown tools that two target individuals designed, and were asked to judge whether or not one of those individuals copied the other's tool. Across trials we manipulated (1) the number of rod options available (2 versus 10); (2) the number of handle options available (2 versus 10); (3) The presence or absence of a functional constraint, i.e. whether they were trying to solve the circle or star puzzle box; (4) The extent of similarity of the two tools that were built (both rod and handle identical, one part identical and one part different, or both rod and handle different). As all designers were assumed to have successfully solved the puzzle, we did not include trials in the star box condition which had different rods, as this would involve building a tool that would not function. Thus in total there were 24 unique test trials. Because of the possibility of demand characteristics if all participants saw the full design, each participant completed only a randomly-selected subset of 4 trials, resulting in 18 unique participants completing each trial.

Procedure

Participants first received instructions regarding the puzzle-box task, and that they would see pairs of tools that people had built to reach the button. Instructions described an ambiguous situation, where copying may or may not have occurred ("While designing the tools the people were in the same room, facing away from each other"). They were instructed that different pairs of people had different numbers of handles and rods to choose from (10 or 2), received either the circle box or star box to solve, and that only one of the rod pieces could fit into the star-shaped opening.

On each trial, participants saw (1) the two tools that the two people had built; (2) which puzzle box the people were trying to solve; (3) the materials they had available to build with. Participants were asked to judge as a 2-alternative forced choice: Do you think someone copied, or they made them independently?

After each trial, an attention check question asked either what puzzle box was present, the number of rod options, or number of handle options. At the end of the task, memory check questions asked participants to select which rods would work, and which handles would work, to successfully solve each of the two puzzle boxes. Lastly, participants were asked to describe what they did in the experiment and guess the point of the study in free-response format, and complete demographics questions.

Analysis Plan

For each model, the best fitting parameters and likelihood of our data given those parameters were assessed via maximum likelihood estimation (MLE). We decided a priori that the prior on copying (range: 0-1) and number of imagined choices (for models that do not use the real number that participants were presented with; range 0-infinity) should be fully free to vary, while the copying error rate *e* was bounded from 0 to a maximum of 0.1. For all models, using this a priori specification, the MLE-derived value for the copying error rate was at max (0.1). To make sure this boundedness was not responsible for our findings, we also reran analyses letting the error rate parameter vary (0-1), and found the same results for comparative model fits in this case. To compare models, we use BIC (Schwarz, 1978), which penalizes models for complexity according to their number of parameters. We used bootstrapping to calculate standard errors (SEs) for each BIC.

Results

We first checked that participants took into account the perceptual similarity of designs in their assessments of copying, as predicted by all four models. As expected, participants inferred copying most often when the two tool designs were identical (M=51.4%, SEM=9.8%), and least often when the two tools were most different (M=5.6%, SEM=2.3%; $p<.01$).

Table 1: Maximum likelihood parameters for each model

Model	Copying Prior, p(*c*)	Error Rate	Imagined # Options
+Availability +Functional	0.09	0.10	
+Availability -Functional	0.06	0.10	
-Availability +Functional	0.09	0.10	5.31
-Availability -Functional	0.11	0.10	2.76

We next compared the fit of the four alternative models. The full model out-performed all competing models, with a difference in BIC of 35 (∓ SEM: 11-25) in comparison to the next-best-fit model and >400 to the other models (Table 2). Approximately the same results held when including individuals who failed the memory check: difference in BIC of 38 to next-best-fit model and >700 to the other models. In addition, the full model provided a good overall fit to participants' responses across trials (R^2= 0.75, Fig. 2A).

Note that while the model is relatively straightforward to specify, the predictions it makes are quite nuanced: because the model weighs and combines several factors, it predicts a continuous gradient of how likely copying should be, rather than simply saying people should never assume copying took place if there is any alternative explanation. The model thus goes well beyond verbal theories.

Use of Availability Constraints

Participants' judgements showed sensitivity to availability constraints (i.e. the number of pieces available to build with), and the use of availability constraints as an alternative explanation for similarity. For example, on trials where two people made identical tools and no functional constraint was present, participants judged copying more likely as the number of available options increased (circle box condition: 2 rods; 2 handles: 33% judged copied; 2 rods, 10 handles: 72%; 10 rods, 2 handles: 72%; 10 rods, 10 handles: 83%).

Use of Functional Constraints

Participants also showed sensitivity to functional constraints, and used functional constraints as an alternative explanation for similarity. In particular, on trials where two people used identical *rods*, participants judged copying less likely on trials where they were solving the star box (which added a functional constraint; Mean copied=21.5%), vs. when they were solving the circle box (Mean copied=52.8%, *p*=0.02, 2 tailed t-test). In contrast, on trials where two people used identical *handles*, participants' judgements did not differ for the star vs. circle box (Star box: Mean copied=36.8%, Circle Box: Mean copied=37.5%; *p*=0.97, 2 tailed t-test), as predicted since all handle pieces would function equally well for both puzzle boxes. Although the model without functional constraints did not perform that poorly as measured by BIC, it did systematically miss this aspect of the data (see also deviations of this model in Figure 2).

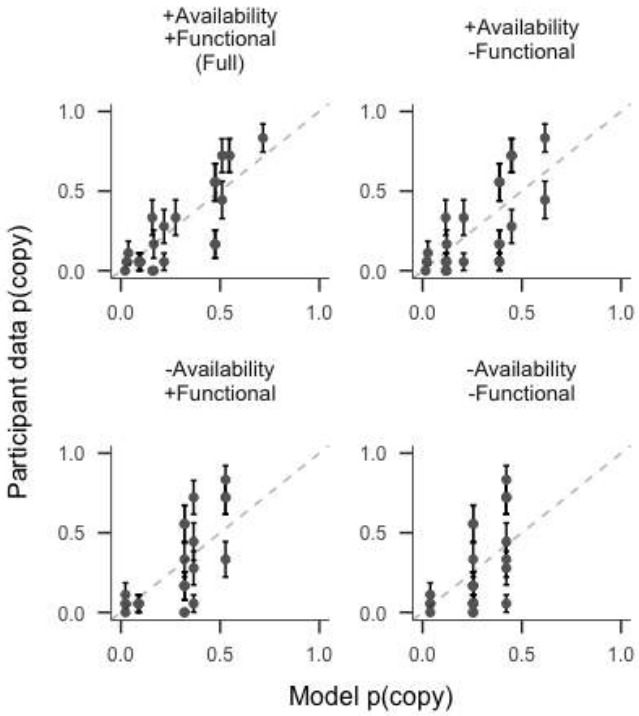


Figure 2: Fit of models' predictions to participants' ratings of whether copying occurred; each point represents one trial. The full inverse planning model appears top left; other plots show three simpler alternative models that do not consider either the availability constraints (-availability) or the functional constraints (-functional).

Table 2: Difference in BIC from best fitting model (higher BIC indicates worse fit)

Model	BIC Δ to full model	∓ SEM
+Availability -Functional	35	11 - 25
-Availability +Functional	467	422 - 490
-Availability -Functional	491	468 - 512

Participants' judgments deviated slightly from the full model's predictions in one regard: Participants appeared to under-weight the similarity of the handles, relative to the rods. For instance, the largest deviations between participants' judgements and the full model's predictions came on trials when the tools had different rods, but the same handle. To demonstrate this differential weighing of the rod vs. handle, consider trials where there are an equal number of rod and handle options, no functional constraint, and the built tools had only one similar piece. On these trials, people were considerably more likely to say the design was copied if the rod was similar than if the handle was (2 options: 0% vs.17%; 10 options: 17% vs. 56%). Thus, participants seemed to overweight evidence from the functionally-relevant component of the tool, even when functional constraints were not present.

Overall, however, the good fit of the inverse planning model – and the continuous range of predictions it makes –

supports the idea that participants use an inverse planning strategy in judging copying from artifacts.

Discussion

We find strong evidence that when reasoning about artifacts, people use a rich, flexible system of explanation-based reasoning to infer whether a design idea was copied or generated independently. We formalized such reasoning in a Bayesian model as a form of inverse planning. We compared this model to three simpler alternatives in a task where participants had to judge whether a pair of artifacts was copied or designed independently, to test whether each component of the full model was needed to predict judgments.

We found that the full inverse planning model best predicted participants' judgments of whether copying had occurred. In line with the model, we found that people considered two broad classes of alternative explanations for artifacts' similarity: the range of materials available to build with (availability constraints), and which of these materials would work to solve the problem (functional constraints). Both of these constraints 'explained away' similarity, making similarity weaker evidence of copying. This pattern of responses is the signature pattern of a Bayesian reasoner, in which a design can have different alternative explanations, and evidence for one provides evidence against the other (e.g., Gopnik et al. 2004).

The success of this model suggests people use a process of inverse planning to infer the source of design ideas from artifacts' features. In other words, people consider the generative processes involved in building the artifacts, including what the goal would be, what constraints they would be subject to, and what (as a result) they would be likely to build. By inverting this generative process, people rationally infer the source of other people's design ideas, taking into account goals and multiple kinds of constraints.

These findings show that inferences about the source of design ideas do not boil down to various simpler heuristics, or more limited systems of reasoning. First, copying judgments are not just based on the extent of perceptual similarity of the two objects, but take into account rational explanations for this similarity. This has implications for understanding how laypeople detect plagiarism in court cases, which has been previously formalized as a process of simple similarity detection (Savage et al., 2018).

Second, we show that this system of reasoning goes beyond efficiency: People can take into account multiple types of constraints as explanations for similarity, and are not limited only to reasoning about design efficiency as the only, privileged type of alternative explanation. This simpler efficiency-only account was consistent with previous findings, and plausible given the foundational role of efficiency in reasoning about intentional action (Schachner et al., 2018). The current data falsify this simpler account, showing that people flexibly take into account the materials available and the functional constraints of the puzzle boxes, which do not map to an efficiency metric (e.g. the length of a train track from A to B, used in Schachner et al., 2018).

More broadly, we provide evidence for a novel theoretical and formal framework for artifact cognition, as a form of inverse planning. Previous work has shown that people use inverse planning to understand the causal processes underlying others' actions (Baker et al., 2009; Liu et al. 2018). The current work extends this framework by conceptualizing artifacts as the products of intentional action. We suggest that people use fundamentally the same inverse planning process to understand artifacts as they do to understand actions themselves. Specifically, they rationally take into account people's goals and constraints not only when observing actions, but also when observing artifacts generated by these actions – even when the actions themselves are not observed. This work thus links together artifact cognition and theories of action understanding in a new way, points to a deep connection between reasoning about actions and artifacts, and provides a foundation for formalizing the processes underlying a domain of 'intuitive archeology' – social-causal reasoning about artifacts, as products of intentional action.

Acknowledgements

We thank Michelle Lee for her help with qualitative coding and stimulus design, as well as Carissa Jantz and Kimberly McGee for stimulus preparation. This material is based upon work supported by the National Science Foundation Grant No. BCS-1749551 to AS and TFB.

References

Baker, C.L., Saxe, R., & Tenenbaum, J.B. (2009). Action understanding as inverse planning. *Cognition*, 113, 329-349.

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, 110 (45) 18327-18332.

Dennett, D.C. (1987). *The Intentional Stance*. MIT Press, Cambridge, MA.

Gergely, G., Nádasdy, Z., Csibra, G., & Bíró, S. (1995). Taking the intentional stance at 12 months of age. *Cognition*, 56(2), 165-193.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: causal maps and Bayes nets. *Psychological Review*, *111*(1), 3-32.

Gopnik, A. (2012). Scientific Thinking in Young Children: Theoretical Advances, Empirical Research, and Policy Implications. *Science*, *337*(6102), 1623–1627.

Gosling, S. (2008). *Snoop: What your stuff says about you*. Profile Books.

Hauser, M., & Wood, J. (2010). Evolving the capacity to understand actions, intentions, and goals. *Annual Review of Psychology*, 61, 303-324.

Henrich, J. (2015). *The secret of our success: How culture is driving human evolution, domesticating our species, and making us smarter*. Princeton University Press.

Legare, C. H., & Nielsen, M. (2015). Imitation and innovation: The dual engines of cultural learning. *Trends in Cognitive Sciences*, 19(11), 688-699.

Liu, S., Ullman, T. D., Tenenbaum, J. B., & Spelke, E. S. (2017). Ten-month-old infants infer the value of goals from the costs of actions. *Science*, *358*(6366), 1038-1041.

Richins, M. L. (1994). Valuing things: The public and private meanings of possessions. *Journal of Consumer Research*, *21*(3), 504–521.

Savage, P. E., Cronin, C., Müllensiefen, D., & Atkinson, Q. D. (2018). Quantitative evaluation of music copyright infringement. In *Proceedings of the 8th International Workshop on Folk Music Analysis* (FMA2018), 61–66.

Schachner, A., Brady, T.F., Oro, K., & Lee, M. (2018). Intuitive archeology: Detecting social transmission in the design of artifacts. In C. Kalisch, M. Rau, T. Rogers, & J. Zhu, *Proceedings of the 40th Annual Conference of the Cognitive Science Society*.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, *6*(2), 461–464.

Skerry, A. E., Carey, S. E., & Spelke, E. S. (2013). First-person action experience reveals sensitivity to action efficiency in prereaching infants. *Proceedings of the National Academy of Sciences*, 110(46):18728-33.

Soley, G., & Spelke, E. S. (2016). Shared cultural knowledge: Effects of music on young children's social preferences. *Cognition*, *148*, 106–116.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, 24(4), 629-640.

Tomasello, M. 1999. *The Cultural Origins of Human Cognition.* Cambridge, MA: Harvard University Press.

Individual Differences in Judging Similarity Between Semantic Relations

Nicholas Ichien[1] ichien@ucla.edu
Hongjing Lu[1,2] hongjing@ucla.edu
Keith J. Holyoak[1] holyoak@lifesci.ucla.edu

[1] Department of Psychology
[2] Department of Statistics
University of California, Los Angeles
Los Angeles, CA 90095 USA

Abstract

The ability to recognize and make inductive inferences based on relational similarity is fundamental to much of human higher cognition. However, relational similarity is not easily defined or measured, which makes it difficult to determine whether individual differences in cognitive capacity or semantic knowledge impact relational processing. In two experiments, we used a multi-arrangement task (previously applied to individual words or objects) to efficiently assess similarities between word pairs instantiating various abstract relations. Experiment 1 established that the method identifies word pairs expressing the same relation as more similar to each other than to those expressing different relations. Experiment 2 extended these results by showing that relational similarity measured by the multi-arrangement task is sensitive to more subtle distinctions. Word pairs instantiating the same specific subrelation were judged as more similar to each other than to those instantiating different subrelations within the same general relation type. In addition, Experiment 2 found that individual differences in both fluid intelligence and crystalized verbal intelligence correlated with differentiation of relation similarity judgments.

Keywords: relational reasoning, similarity, semantic cognition, fluid intelligence, crystallized intelligence

Introduction

A house key and an email password are intuitively similar. This similarity is not based on any common attributes or constituent properties of individual objects; rather, it seems to be based on some common *relation* that a house key and an email password respectively bear to a house and to an email account (roughly, *providing access*). The ability to grasp and exploit similarity based on a wide variety of relations is an important and distinguishing trait of human intelligence (Penn, Holyoak, & Povinelli, 2008). This ability underlies much of human thought, including aspects of language (Gentner & Namy, 2006), categorization (Gentner & Kurtz, 2005; Goldwater & Schalk, 2016), and perhaps most prominently, analogical reasoning (Holyoak, 2012). The explicit representation of abstract relations is an indispensable explanatory construct in major computational accounts of human analogical reasoning (Doumas, Hummel, & Sandhofer, 2008; Falkenhainer, Forbus, & Gentner, 1989; Halford, Wilson, & Phillips, 1998; Hummel & Holyoak, 2003; Lu, Chen, & Holyoak, 2012; Lu, Wu, & Holyoak, 2019; Petrov, 2013). Empirical work on relational reasoning has provided compelling evidence that humans store representations of semantic relations in memory (Estes & Jones, 2006; Popov, Hristova, & Anders, 2017; Spellman, Holyoak, & Morrison, 2001).

A number of important research questions depend on finding an effective method to assess human judgments of relational similarity. A major source of complexity stems from evidence that relations are not represented as discrete all-or-none concepts, but rather exhibit internal variability. Just as instances of natural and functional object categories differ in typicality (Rosch, 1975), so too people reliably judge word pairs to be better or worse instantiations of relations (Jurgens, Mohammad, Turney, & Holyoak, 2012). For example, *fail:succeed* is considered to be a better example of the relation *reverse* than is *eat:starve*.

Given such variations in intra-relation "goodness", it is natural to hypothesize that inter-relation similarity will also have a graded structure. Indeed, a recent theory of relation learning (*Bayesian Analogy with Relational Transformations*, BART) claims that the specific relation between a pair of words corresponds to a distributed representation over multiple relations, each of which the pair instantiates with some probability (Lu et al., 2019). For example, *lid:bottle* seems to instantiate the relations *part-whole*, *on-top-of*, and *closure-of*. BART can be used to derive theoretical predictions about the degree of similarity between a wide range of word pairs that collectively instantiate multiple relations.

It would clearly be desirable to obtain reliable human judgments of relational similarity, which might then be compared to theory-based predictions. Such data could also be used to assess potential individual differences in relation representations. A great deal of research indicates that complex relational reasoning depends on working memory capacity and other aspects of fluid intelligence (for a review see Holyoak, 2012). In particular, there is evidence that performance on analogical reasoning tasks is positively related to fluid intelligence as measured by tests such as the Raven's Progressive Matrices (RPM; Gray & Holyoak, 2018). It is possible that fluid intelligence plays a role in maintaining and comparing relations in working memory in order to differentiate among relations that overlap in meaning. Similarly, crystalized verbal intelligence seems to play an important role in comprehending metaphors (Stamenković, Ichien, & Holyoak, 2019), and may be related to the differentiation of relational concepts in semantic memory.

A reliable measure of human judgments of relation similarity would clearly be very useful for testing theories of relation representation. However, in practice it is difficult to

find an efficient procedure to elicit similarities among large sets of items (since the number of pairwise comparisons becomes prohibitively large when the number of items is substantial). Here we explore the use of a *multi-arrangement* method (adapted from previous work on assessing object similarity; Kriegeskorte & Mur, 2012) for obtaining judgments that can be used to efficiently generate a map of the psychological similarity space for abstract semantic relations.

The present paper aims to offer a first step in the exploration of relational similarity, assessing the validity and reliability of a new method for collecting human judgments of relational similarity and conducting preliminary analyses of these similarity judgments. Experiment 1 sets the stage by testing whether the method can generate sensible patterns of relation similarity. Experiment 2 then extends the method to more fine-grained semantic distinctions among relations to examine potential gradations in relational similarity. Further, Experiment 2 assesses the potential association between judgments of relation similarity and individual differences in both fluid and crystalized intelligence.

Experiment 1

The major goal of Experiment 1 was to determine whether a novel method for eliciting human judgments of relation similarity is able to capture broad distinctions among semantic relations that have been posited on the basis of previous theoretical and empirical investigations.

Method

Participants

20 participants (mean age = 19.05 years; 17 female) were recruited from the Psychology Department subject pool at the University of California, Los Angeles (UCLA). All participants were self-reported fluent English speakers. Participants provided verbal consent in accordance with the UCLA Institutional Review Board and were compensated with course credit.

Stimuli

All stimuli were word pairs taken from the SemEval-2012 Task 2 dataset (Jurgens et al., 2012), which is in turn based on a taxonomy of abstract semantic relations developed by Bejar, Chaffin, and Embretson (1991). Word pairs in this dataset express one of 79 specific relations, each falling into one of 10 general types of relations. Experiment 1 tested examples drawn from relations in each of three different general relation types (*similar*, *contrast*, and *cause-purpose*). We will refer to the examples in Experiment 1 by the names of the specific relations: *synonymy*, *contrary*, and *cause:effect* (see Table 1). Each relation included 16 word pairs, consisting of one paradigm exemplar (a seed used by Jurgens et al. to define the relation) and the 15 most prototypical word pairs for that relation. Pairs were unique in that they did not include inversions of one another. Table 1 provides examples of the word pairs used in the experiment.

Relation types	Word pair examples
synonymy	car:auto
contrary	old:young
cause:effect	joke:laughter

Table 1. Relations and examples of word pairs used in Experiment 1.

Procedure

We acquired human similarity judgments of semantic relations by asking participants to perform a multi-arrangement task, a method for efficiently eliciting similarity judgments, especially for large sets of items (Kriegeskorte & Mur, 2012). The method, which can be viewed as an inverse of standard multidimensional scaling (Shepard, 1962), has previously been successfully used for judgments of object similarity (Kriegeskorte & Mur, 2012; Mur et al., 2013; Jozwik, Kriegeskorte, Storrs, & Mur, 2017). Here we extend it to judgments of relation similarity.

On each trial, participants were presented with a subset of the 48 word pairs on a computer screen. They were asked to first identify the relation between words in each pair, and then use a mouse to arrange word pairs in a two-dimensional circular space according to the similarity of their relations (see Figure 1). Participants were told, "word pairs that involve similar relations should be placed close together," "word pairs that involve very different relations should be placed far apart," and "the distance between two word pairs should represent how different their relations are." Participants were also instructed to use the entire space to arrange word pairs on each trial.

We aimed to obtain similarity judgments from each participant relating each of the 48 item pairs to each other (a total of 1128 pairwise measurements). Estimates of similarity were based on the *relative* on-screen distances between word pairs as arranged by participants on each trial. These estimates were calculated by scaling the distances between items arranged on a single trial to match a weighted average of these distances calculated across trials. This weighted average was iteratively recomputed until convergence.

On a given trial, participants were presented with a maximum of 20 word pairs. The multi-arrangement task involves adaptively selecting stimuli to present on each trial. On the first trial, participants arranged a random subset of 20 items from the entire set of 48 items. On subsequent trials, participants arranged a subset of 20 or fewer items selected based on item pairs with the weakest similarity evidence (see Kriegeskorte & Mur, 2012, for an extended discussion).

Previous uses of the multi-arrangement task have involved 1-hour sessions (e.g., Kriegeskorte & Mur, 2012; Mur et al., 2013; Jozwik et al., 2017), but these studies all asked participants to do a relatively easier task of arranging individual objects according to their similarity. Due to the higher demand on working memory in arranging word pairs according to their relational similarity, pilot experiments suggested that a 1-hour session length would likely result in fatigue and disengagement for naïve participants. Accordingly, we limited session length to 30 minutes.

Figure 1. Example trial of the multi-arrangement task used to generate a semantic space for relations.

Participants were allowed to spend as long as they needed to complete each trial. On average, participants completed 28.5 trials (SD = 11.86, range = 4-44) within the 30-minute experimental duration.

Results

All but five participants provided a full set of pairwise similarity judgments between all combinations of the 48 word pairs. Of the five who failed to complete all possible comparisons, four provided judgments for 98% of the pairwise combinations of word pairs. The fifth participant provided judgments for just 57% of the combinations; this individual's data were excluded from analyses.

We assessed the inter-subject reliability of our relational similarity judgments by calculating the Pearson correlation coefficients between individual participants' distance matrices. The mean correlation between any two participants' distance matrices was .50 (range = .11 to .83).

We then examined whether the multi-arrangement task provided a reliable measure of relation similarity (assuming that greater inter-pair distance implies lower similarity). The results showed that participants generated smaller distances between word pairs within a relation compared to distances between word pairs instantiating a different relation. Figure 2 depicts a mean distance matrix obtained by averaging across distance matrices generated by individual participants performing the multiple-arrangement task.

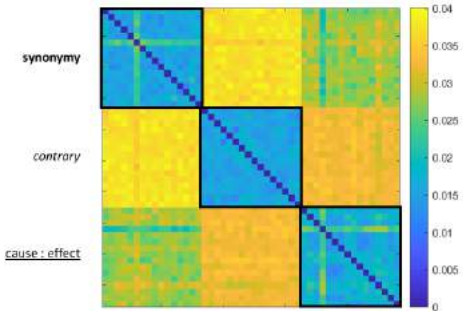


Figure 2. Mean distance matrix between the 48 word pairs used in Experiment 1. Cold colors represent smaller distances (i.e., greater pairwise similarity); hot colors represent greater distances (i.e., lesser pairwise similarity). Boxed regions represent pairwise distance measures between word pairs instantiating the same relation.

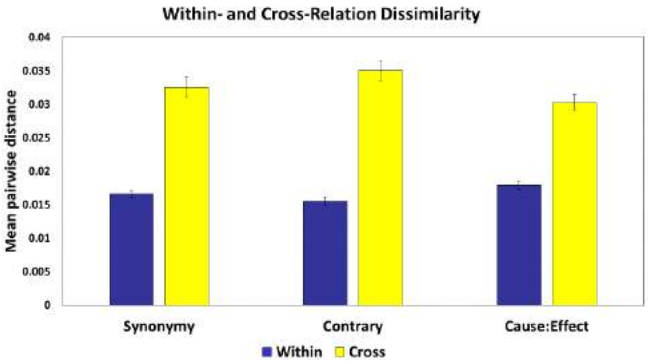


Figure 3. Mean within- and cross-relation distance measures for pairs instantiating each relation (Experiment 1). Higher bars indicate greater distance (i.e., lower similarity). Error bars indicate +/- 1 standard error of the mean.

We compared the mean distances between word pairs instantiating different relations (i.e., cross-relation distances) to the mean distances between word pairs instantiating the same relation (i.e., within-relation distances). To perform this analysis, we first calculated within- and cross-relation distances for each word pair for each individual participant. Next, we found the mean value of both of these distance measures averaged across word pairs within each relation. As depicted in Figure 3, cross-relation distances were greater than within-relation distances for each relation: for *synonymy* ($t(18) = 8.66$, $p < .001$, Cohen's $d = 1.99$); for *contrary* ($t(18) = 10.26$, $p < .001$, Cohen's $d = 2.35$); for *cause:effect* ($t(18) = 8.91$, $p < .001$, Cohen's $d = 2.5$). These findings thus establish that the multi-arrangement task is an effective method to obtain human judgments of relation similarity.

Experiment 2

Experiment 2 aimed to determine whether human judgments of relational similarity are sensitive to more fine-grained distinctions among relations than those examined in Experiment 1. In addition, we investigated whether relation judgments are systematically influenced by individual differences in cognitive capacity and/or semantic knowledge. To assess fluid intelligence, we administered a short version of the RPM (Arthur, Tubre, Paul, & Sanchez-Ku, 1999) adapted for computer administration using Matlab software. Participants are presented with a 3x3 grid of items with the item in the bottom right corner missing. They are asked to use the pattern instantiated by the presented items to select the most appropriate item to fill that bottom right corner from a set of 8 options. Prior research has shown that superior performance on this test is correlated with performance on tests of analogical reasoning (Vendetti, Wu, & Holyoak, 2014; Kubricht, Lu, & Holyoak, 2017). We hypothesized that the RPM measure would be associated with the degree to which people are able to differentiate word pairs that instantiate distinct relations.

In addition to fluid intelligence, the ability to differentiate among semantic relations may vary with crystalized verbal intelligence, particularly knowledge of semantic relations. As

a measure of semantic knowledge, we administered the Semantic Similarities Test (SST). This test was designed to be similar to the Similarities subscale of the Weschler Adult Intelligence Scale (WAIS), and is correlated with the Vocabulary subtest (Stamenković et al., 2019). Participants are presented with 20 pairs of verbal concepts and asked to describe how the concepts in each pair are similar. The concept pairs span a broad range of similarities: some are fairly specific (e.g., *bird-airplane*, which both fly), some are more general (e.g., *tavern-church*, which are both public buildings), and some are more metaphorical (e.g., *marriage-alloy*, which are both bonds between elements). Because the identification of more specific and fine-grained relations likely depends on greater semantic knowledge, we hypothesized that superior performance on the SST would also be correlated with greater differentiation of similarities among semantic relations.

Method

Participants

93 new participants (mean age = 20.17 years; 69 female) were recruited from the UCLA Psychology Department subject pool. All participants had normal or corrected-to-normal vision and were self-reported fluent English-speakers. Participants provided verbal consent in accordance with the UCLA Institutional Review Board and were compensated with course credit.

Stimuli

The multi-arrangement task in Experiment 2 used 27 word pairs drawn from the same norms as in Experiment 1 (Jurgens et al., 2012). Three word pairs were chosen from each of three specific subrelations of three general relation types (see Table 2). Note that the three relations used in Experiment 1 were included as specific subrelations used in Experiment 2. Whereas Experiment 1 did not manipulate the level of relation abstraction, Experiment 2 did. Specifically, Experiment 2 examined whether similarity judgments not only reflect broad distinctions at a high level of abstraction (i.e., between general relation types), but also fine distinctions at a lower level of abstraction (i.e., between specific subrelations within general relation types). Word pairs drawn from different subrelations of the same general type (e.g., *car:auto* instantiates *synonymy* and *rake:fork* instantiates *attribute similarity*, two subrelations of the relation type *similar*) are differentiated on the basis of relatively subtle relational differences. Each set of three unique word pairs consisted of one paradigm exemplar and the third and sixth most prototypical unique word pairs for that subrelation in the SemEval-2012 Task 2 norms (Jurgens et al., 2012).

Procedure

All participants completed three tasks in the following order: the multi-arrangement task, the Raven's Progressive Matrices (RPM) and the Semantic Similarities Test (SST).

General relation types	Specific subrelations	Word pair examples
similar	*synonymy*	car:auto
	attribute similarity	rake:fork
	change	discount:price
contrast	*contrary*	old:young
	directional	east:west
	pseudoantonym	right:bad
cause-purpose	*cause:effect*	joke:laughter
	cause: compensatory action	hunger:eat
	action/activity: goal	flee:escape

Table 2. General relation types, three specific subrelations chosen to exemplify each, and examples of word pairs used in Experiment 2.

Results

All 93 participants completed the multi-arrangement task. On average participants completed 19.51 trials (SD = 9.70, range 2-55). All but one participant provided pairwise similarity judgments for all 27 word pairs (351 pairwise comparisons). That one participant provided judgments for 86% of the pairwise combinations. Due to program failures, only 88 participants completed the SST, and 90 participants completed the RPM.

We again assessed the inter-subject reliability of our relational similarity judgments by calculating the Pearson correlation coefficients between individual participants' distance matrices. The mean correlation between any two participants' distance matrices was .38 (range = -.09 to .88).

Figure 4 depicts the mean distance matrix for all word pairs. We compared the mean distances of word pairs drawn from different general relation types (i.e., cross-type distances) to mean distances of word pairs within the same relation type (i.e., within-type distances). As depicted in Figure 5, cross-type distances were greater than within-type distances for each relation type: for *similar* ($t(92) = 10.53$, $p < .001$, Cohen's $d = 1.09$); for *contrast (*$t(92) = 18.32$, $p < .001$, Cohen's $d = 1.90$); for *cause-purpose (*$t(92) = 17.06$, $p < .001$, Cohen's $d = 1.77$).

To examine whether participants were sensitive to differences between specific subrelations within the same relation type, we compared the mean distances of word pairs instantiating different subrelations within the same general relation type (i.e., cross-subrelation distances) to the mean distances of word pairs instantiating the same subrelations (within-subrelation distances). For each relation type, mean cross-subrelation distances were greater than mean within-subrelation distances: for *similar* ($t(92) = 13.17$, $p < .001$, Cohen's $d = 1.37$); for *contrast (*$t(92) = 12.95$, $p < .001$, Cohen's $d = 1.34$); for *cause-purpose (*$t(92) = 7.35$, $p < .001$, Cohen's $d = 0.76$). These findings indicate that participants were not only able to differentiate between general relation types but were also sensitive to much more fine-grained distinctions within the same relation type. Further, these findings provide evidence of graded similarity structure among semantic relations. Specifically, word pairs instantiating the same general relation type were judged as

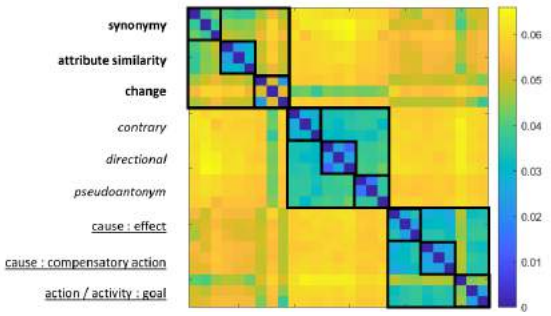


Figure 4. Mean distance matrix from Experiment 2. Cold colors represent smaller distances (i.e., greater pairwise similarity), whereas hot colors represent greater distance (i.e., lesser pairwise similarity). Larger boxed regions represent pairwise distance judgments between word pairs instantiating the same general relation type. Smaller boxed regions represent pairwise distance judgments between word pairs instantiating the same specific subrelation within a common relation type.

more similar to each other than those instantiating different general relation types, and word pairs instantiating the same subrelation were judged as more similar to each other than those instantiating different subrelations within the same general relation type.

Next, we performed analyses to determine whether individual differences in cognitive capacity (as assessed by the RPM) and semantic knowledge (as assessed by the SST) were associated with participants' sensitivity to differences among relations. Two independent raters scored the SST based on the criteria summarized by Stamenković et al. (2019). We assessed the reliability of these raters' scores by testing the average measure intraclass correlation coefficient across scores using a two-way mixed model (ICC = .971, $F(19,19) = 44.72$, $p < .001$, with a 95% confidence interval from .899 to .990). Given the reliability of these scores, we used the average score across these two raters in the following analyses.

In order to estimate individual differences in sensitivity to broad distinctions relation types, we computed a relation type discriminability index for each participant using the following steps. First, we found each participant's cross-type distance by calculating the mean distance for pairwise comparisons between word pairs instantiating different general relation types (e.g., *old:young* instantiates the relation type *contrast*, while *car:auto* instantiates the relation type *similar*). Second, we found each participant's within-type distance by calculating the mean distance for pairwise comparisons between word pairs instantiating the same general relation type (e.g., *old:young* and *east:west* both instantiate the relation type *contrast*). Third, we computed each participant's discriminability index by dividing that participant's cross-type distance by their within-type distance (range = 1.01 to 2.60). This relation type discriminability index reflects how well a participant discriminated between relation types in their similarity judgments. An index of 1 indicates complete lack of discriminability between word pairs instantiating different relation types and those instantiating the same relation type, whereas higher indices indicate judgments of greater similarity between word pairs instantiating the same relation type than between word pairs instantiating different relation types.

These discriminability indices for relation types were significantly correlated with RPM scores (Pearson's $r = .33$, $p = .005$, power = .90) and also with SST scores (Pearson's $r = .30$, $p = .014$, power = .82). Partial correlations revealed that these discriminability indices were significantly correlated with RPM scores after residualizing out SST scores (Pearson's $r = .236$, $p = .028$, power = .61), and that they were significantly correlated with SST scores after residualizing out RPM scores (Pearson's $r = .236$, $p = .028$, power = .61). These results indicate that there is an association between the discrimination of general relation types both with cognitive capacity and with semantic knowledge.

In order to estimate each participant's sensitivity to more fine-grained distinctions between specific subrelations within general relation types, we also computed a subrelation discriminability index using the following steps. First, we found each participant's cross-subrelation distance by calculating the mean distance for pairwise comparisons between word pairs instantiating different subrelations within the same general relation type (e.g., *old:young* instantiates the subrelation *contrary*, and *east:west* instantiates the subrelation *directional*, where both instantiate the relation type *contrast*). Second, we found each participant's within-subrelation distance by calculating the mean distance for pairwise comparisons between word pairs instantiating the same subrelation (e.g., *old:young* and *black:white* both instantiate the subrelation *contrary*). Third, we computed each participant's subrelation discriminability index by dividing each participant's cross-subrelation distance by their within-subrelation distance (range = .96 to 2.74). This subrelation discriminability index reflects how well a participant was able to discriminate between specific subrelations within a relation type in their similarity judgments. An index of 1 would indicate a complete lack of discriminability between word pairs instantiating different subrelations and those instantiating the same subrelation, whereas higher indices indicate judgments of greater similarity between word pairs instantiating the same subrelation than between word pairs instantiating different subrelations.

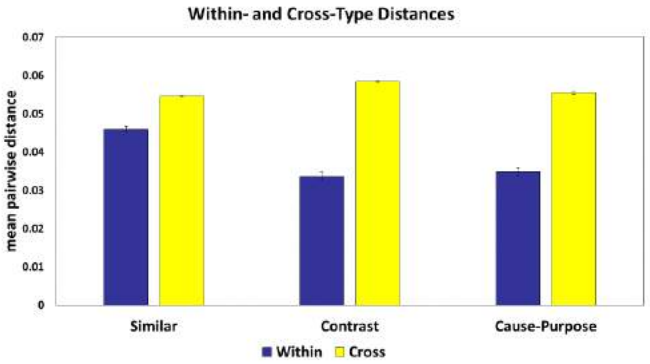


Figure 5. Mean within- and cross-type distances for each general relation type in Experiment 2. Error bars indicate +/- 1 standard error of the mean.

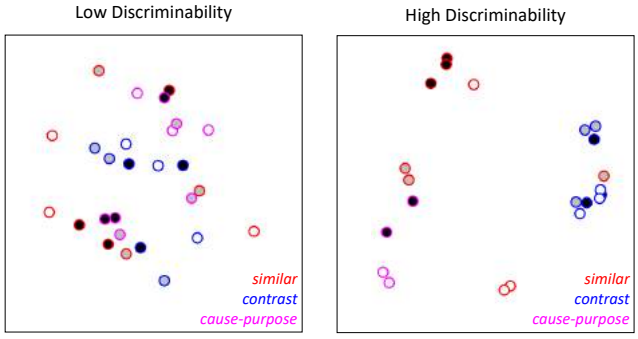


Figure 6. Visualization of relation similarities from two representative participants. Left: MDS solution for a participant with low discriminability indices (relation type discriminability index = 1.02; subrelation discriminability index = .98). Right: solution for a participant with high discriminability indices (2.08 and 2.74, respectively). Each marker indicates a single word pair. Marker outline color indicates word pair relation type, and marker shading indicates subrelation within relation type.

These fine-grained discriminability indices for subrelations showed a significant correlation with RPM scores (Pearson's r = .35, p = .003, power = .93), and also with SST scores (Pearson's r = .30, p = .014, power = .82). Partial correlations revealed that these discriminability indices were significantly correlated with RPM after residualizing out SST scores (Pearson's r = .291, p = .006, power = .79), but that they were not correlated with SST scores after residualizing out RPM scores (Pearson's r = .090, p = .408). These results indicate that there is a stronger association between the discrimination of specific subrelations within relation types with cognitive capacity than with semantic knowledge.

To provide a visualization of the difference between high and low discriminability, Figure 6 presents multidimensional scaling (MDS) solutions (Shepard, 1962) for the distance matrices of a participant with both a low relation type and a low subrelation discriminability index (left) and of a participant with both a high relation type and a high subrelation discriminability index (right). The latter solution shows a much greater degree of clustering into distinct relation types as well as into subrelations.

General Discussion

Across two experiments, we showed that a multi-arrangement task can be used to efficiently assess judgments of similarity among semantic relations. Human judgments obtained using this method have a clear interpretation. Judged similarity reflects not only broad distinctions between relation types, but also finer distinctions between subrelations within relation types. Moreover, the degree to which a participant differentiated between pairs from the same versus different relation types was positively correlated with measures of both fluid and verbal crystallized intelligence. At the more detailed level of subrelations, only fluid intelligence was a reliable predictor of discriminabilty. Future work should examine these associations further and assess directions of causality.

The present findings add to mounting evidence that semantic relations do not have discrete, all-or-none representations. Previous work has shown that word pairs instantiating a particular relation vary systematically in their *typicality* (Jurgens et al., 2012; Popov et al., 2017), much like instances of object categories (Rosch, 1975). Our findings reveal that *similarities* between relation examples (within and across subrelations) also vary in a graded fashion. In addition, the present study establishes that similarity gradients for relations show reliable individual differences across people who vary in either cognitive capacity or semantic knowledge of relations.

Note typicality judgments are importantly distinct from similarity judgments. Specifically, typicality is a relation between entities at different levels of abstraction (i.e., exemplar and category), and the typicality of a word pair is necessarily defined with respect to a particular relation. For example, *up:down* is typical of the relation *opposite*. In contrast, similarity is generally a relation between entities at the same level of abstraction (i.e., exemplar and exemplar), and relational similarity of a word pair can be defined with respect to another word pair. For example, *up:down* is similar to *light:dark*., Notably, whereas typicality judgments can be used to evaluate relational semantic representations *within* relations, similarity judgments can be used as a more holistic evaluation *across* relations.

This emerging picture of human relation concepts is consistent with models of relation learning and analogical reasoning that assume relations are coded by distributed representations (e.g., Lu et al., 2019). More generally, judgments of relation similarity provide a rich source of potential data that can be used to evaluate computational models. Specifically, a relation distance matrix generated from a theoretical model can be compared to a distance matrix obtained from human judgments of relation similarity, as described here. To the extent that a model-generated distance matrix approximates a human-generated distance matrix, the model's representation of semantic relations is descriptive of human semantic cognition. The same logic can be applied to test computational models as predictors of relational priming (Estes & Jones, 2009; Popov et al., 2017; Spellman et al., 2001), and of neural responses to relation processing (Kriegeskorte, Mur, & Bandettini, 2008).

The multi-arrangement method of collecting similarity judgments for relations may also prove useful in guiding studies of educational interventions (Goldwater & Schalk, 2016). The type of MDS solution that can be derived from similarity judgments can be related to the well-known technique of using "concept maps" to teach systematically related concepts. The degree of match between the clusters identified in an MDS solution obtained for an individual learner may provide a useful index of how well that learner's internal representation of a set of concepts maps onto the organization the teacher aimed to convey.

Acknowledgements

We thank Ali Hepps, Anvita Diwan, Lina Chan, and Zhibo Zhang for assistance in data collection. This research was supported by NSF Grant BCS-1827374.

References

Arthur, P. L., Tubre, T. C., Paul, D. S., & Sanchez-Ku, M. L. (1999). College-sample psychometric and normative data on a short form of the Raven Advanced Progressive Matrices Test. *Journal of Psychoeducational Assessment*, *17*, 354-361.

Bejar, I. I., Chaffin, R., & Embretson, S. E. (1991). *Cognitive and psychometric analysis of analogical problem solving*. New York: Springer-Verlag.

Doumas, L. A. A., Hummel, J. E., & Sandhofer, C. M. (2008). A theory of the discovery and predication of relational concepts. *Psychological Review*, *115*, 1-43.

Estes, Z., & Jones, L. L. (2009). Integrative priming occurs rapidly and uncontrollably during lexical processing. *Journal of Experimental Psychology: General*, *138*(1), 112-130.

Falkenhainer, B., Forbus, K. D., & Gentner, D. (1989). The structure-mapping engine: Algorithm and examples. *Artificial Intelligence*, *41*(1), 1–63.

Gentner, D., & Kurtz, K. (2005) Relational categories. In W. K. Ahn, R. L. Goldstone, B. C. Love, A. B. Markman, & P. W. Wolff (Eds.), *Categorization inside and outside the lab*. Washington, DC: APA.

Gentner, D., & Namy, L. L. (2006). Analogical processes in language learning. *Current Directions in Psychological Science*, *15*, 297-301.

Goldwater, M. B., & Schalk, L. (2016). Relational categories as a bridge between cognitive and educational research. *Psychological Bulletin*, *142*, 729-757.

Gray, M. E., & Holyoak, K. J. (2018). Individual differences in relational reasoning. In C. Kalish, M. Rau, J. Zhu & T. T. Rogers (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 1741-1746). Austin, TX: Cognitive Science Society.

Halford, G. S., Wilson, W. H., & Phillips, S. (1998). Processing capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology. *Behavioral and Brain Sciences*, *21*(6), 803-864.

Holyoak, K. J. (2012). Analogy and relational reasoning. In K. J. Holyoak & R. G. Morrison (Eds.), *Oxford handbook of thinking and reasoning* (pp. 234-259). New York: Oxford University Press.

Hummel, J. E., & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review*, *110*(2), 220–263.

Jozwik, K. M., Kriegeskorte, N., Storrs, K. R., & Mur, M. (2017). Deep convolutional neural networks outperform feature-based but not categorical models in explaining object similarity judgments. *Frontiers in Psychology*, *8*, Article ID 1726. DOI: 10.3389/fpsyg.2017.01726

Jurgens, D. A., Mohammad S. M., Turney P. D., & Holyoak K. J. (2012) SemEval-2012 Task 2: Measuring degrees of relational similarity. *Proceedings of the First Joint Conference on Lexical and Computational Semantics (*SEM)*, 356-364.

Kriegeskorte, N., & Mur, M. (2012). Inverse MDS: Inferring dissimilarity structure from multiple item arrangements. *Frontiers in Psychology*, *3*. DOI: 10.3389/fpsyg.2012.00245

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in System Neuroscience*, *2*(4). DOI: 10.3389/neuro.06.004.2008

Kubricht, J. R., Lu, H., & Holyoak, K. J. (2017). Individual differences in spontaneous analogical transfer. *Memory & Cognition*, *45*, 576-588.

Lu, H., Chen, D., & Holyoak, K. J. (2012). Bayesian analogy with relational transformations. *Psychological Review*, *119*, 617-648.

Lu, H., Wu, Y. N., & Holyoak, K. (2019). Emergence of analogy from relation learning. *Proceedings of the National Academy of Sciences, USA*, *116*, 4176-4181.

Mur, M., Meys, M., Bodurka, J., Goebel, R., Bandettini, P. A., & Kriegeskorte, N. (2013). Human object-similarity judgments reflect and transcend the primate-IT object representation. *Frontiers in Psychology*, *4*, 128. DOI: 10.3389/fpsyg.2013.00128

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, *104*(3), 192-233.

Penn, D. C., Holyoak, K. J., & Povinelli, D. J. (2008). Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. *Behavioral and Brain Sciences*, *31*(2), 109-130.

Petrov, A. A. (2013). *Associative Memory-based Reasoning: A computational model of analogy-making in a decentralized multi-agent cognitive architecture*. Saarbrücken, Germany: Lambert Academic.

Popov, V., Hristova, P., & Anders, R. (2017). The relational luring effect: Retrieval of relational information during associative recognition. *Journal of Experimental Psychology: General*, *146*(5), 722-745.

Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika*, *27*, 125-40.

Spellman, B. A., Holyoak, K. J., & Morrison, R. G. (2001). Analogical priming via semantic relations. *Memory & Cognition*, *29*, 383-393.

Stamenković, D., Ichien, N., & Holyoak, K. J. (2019). Metaphor comprehension: An individual-differences approach. *Journal of Memory and Language*, *105*, 108-118.

Vendetti, M., Wu, A., & Holyoak, K. J. (2014). Far-out thinking: Generating solutions to distant analogies promotes relational thinking. *Psychological Science*, *25*(4), 928-933.

The impact of anecdotal information on medical decision-making

Sara Jaramillo (sdjarami@asu.edu)
School of Social and Behavioral Sciences, Arizona State University
Phoenix, AZ, USA

Zachary Horne (Zachary.Horne@asu.edu)
School of Social and Behavioral Sciences, Arizona State University
Phoenix, AZ, USA

Micah Goldwater (micah.goldwater@sydney.edu.au)
School of Psychology, The University of Sydney
Camperdown, Australia

Abstract

In prior research, arguments using both anecdotal and statistical evidence are more persuasive than arguments using either alone (Allen, Bruflat, Fucilla, Kramer, McKellips, Ryan, & Spiegelhoff, 2000; Hornikx, 2005). However, it is less clear how people integrate information when the statistics and the anecdotes present conflicting information. In three preregistered experiments, we tested how people integrate conflicting information to judge the efficacy of a medicine in a clinical trial. Participants read either an anecdote from someone in the trial, summary statistics about the trial, or both types of information. We found that reading an anecdote from a member of the trial for whom treatment was ineffective reduced people's beliefs in a medical treatment even when participants received strong evidence that the treatment was effective. In Experiment 3, we found that introducing icon arrays increased the perceived efficacy of the treatment but did not eliminate the effect of the anecdote.

Keywords: anecdotal reasoning; medical decision-making; open science

Introduction

Making decisions about medical treatments can be a difficult and stress inducing process. When the decision concerns those we love, or those who are vulnerable, the stakes can make even obvious decisions seem paralyzing. People are inundated with popular press reports about medical research concerning what's healthy, get advice from doctors, and hear personal anecdotes from friends, relatives, and the media. How can people make appropriate medical decisions under these conditions? It might seem obvious that people's beliefs should reflect the scientific consensus, but when our own and our families' health is at stake, a compelling narrative or personal anecdote can be hard to ignore. For instance, vaccine hesitancy has been found to be driven by reliance on anecdotal evidence about the side effects of vaccines spread throughout online communities even though vaccines are among the safest medical treatments (Powell, Weisman, & Markman, 2018). Altogether making a medical decision is no easy feat, even for the epistemically diligent.

Prior research suggests that although people are capable of correctly integrating statistical information to make informed medical decisions (e.g., Allen & Preiss, 1997; Allen et al., 2000; Hornikx, 2005), they may nonetheless improperly attend to irrelevant anecdotal information, particularly when that evidence is salient and relates to uncertainty and risk (e.g., Allen et al., 2000; Shen, Sheer, & Li, 2015). Some researchers suggest that narratives are more engaging and comprehensive (Dahlstrom, 2014), but when learning about new scientific information, anecdotal information can distract from making proper scientific judgments (Rodriguez, Rhodes, Miller, & Shah, 2016). What remains unclear is how people integrate anecdotes *with* statistical information. When people are presented with both statistical summary information and anecdotes, how do they reason on the basis of this information? Can positive anecdotal information aid in the integration of statistical information when in concert with each other? Some research suggests that anecdotes do not impact the integration of statistical evidence about government policy (Hornikx 2018), but there is little research on this question in the domain of medical decision-making, where the stakes are high and thus anecdotes might may exhibit stronger effects.

In the present studies, we examined the effect of anecdotes on medical decision-making. We investigated the ways in which anecdotal information influences how people interpret a study describing the efficacy of a novel medical treatment (Experiments 1 and 2), and what other factors may weaken the effect of anecdotes on reasoning (Experiment 3).

General Methods

Preregistration We preregistered the data collection plan, analyses, and predictions for all three experiments. Experimental scripts, full analytic results, and supplementary online materials (SOM) are available on the Open Science Framework at https://osf.io/dkcwv/.

Analytic Approach We performed Bayesian estimation using the R package `brms` (Bürkner, 2018). We set regularizing priors for all population-level effects in our models, which we detail below. These priors are recommended because they provide conservative effect size estimates and reduce the likelihood of overfitting (McElreath, 2016). Following the recommendations of Liddell & Kruschke (2018), Likert data were modeled with a cumulative probability distribution.

Experiment 1

Experiment 1 examined how anecdotes affect people's reasoning about medical information. We sought to avoid polarizing medical treatments because beliefs about these topics may be particularly intransigent. Consequently, we focused on a plausible but relatively unknown medical treatment that people would not have strong beliefs about. Specifically, we examined people's beliefs about B-12 injections as means for treating chronic headaches.

Participants We recruited 497 participants through Amazon's Mechanical Turk (47% women, M_{age}= 38 years old). Participants were paid $0.50 for participating in a five-minute study. After excluding participants who missed questions checking their attention, 431 participants remained in our sample. Our exclusion criteria were determined a priori and were in accordance with our study preregistration.

Procedure In Experiment 1, we presented participants with either statistical evidence, anecdotal evidence, or the combination of both types of evidence about a medical trial testing the effectiveness of B-12 injections on chronic headaches. The study consisted of three parts: a pretest questionnaire, an intervention, and then a posttest questionnaire. After completing this portion of the study, participants completed medical individual differences measures and demographic questions. We describe each component below.

Pretest Questionnaire Participants answered a brief questionnaire examining their familiarity with B-12 injections, whether they are currently receiving or have received B-12 injections, and whether they are considering receiving B-12 injections as a medical treatment. After responding to these questions, participants were then asked on a five-point Likert scale whether they believe B-12 injections are an effective medical treatment (1 = "Not effective at all", 5 = "Extremely effective").

Conditions After completing the B-12 pretest questionnaire, participants were randomly assigned to one of four conditions: the Statistics condition, the Positive Anecdote condition, the Statistics + Positive Anecdote condition, or the Statistics + Negative Anecdote condition.

In the Statistics condition, participants were shown a description with summary statistics about a clinical trial examining the effects of B-12 injections on patients with chronic headaches. Namely, participants read that in a clinical trial with 1,000 subjects, B-12 injections were 87.3% effective as a medical treatment for chronic headaches.

In the Positive Anecdote condition participants did not receive the statistical information but were told "Jamie's [the protagonist in the anecdote] doctor recommended that she participate in a new clinical trial that was examining the effects of B-12 on headaches" and then were told that Jamie decided to receive B-12 and subsequently experienced a reduction in her symptoms.

In the Statistics + Positive Anecdote condition, participants first read the summary statistics demonstrating the efficacy of B-12 injections (that is, the only material presented in the Statistics condition). They were then told that they would read about the experience of one of the subjects in the study, after which they were presented with the anecdote from the Positive Anecdote condition.

Participants in the Statistics + Negative Anecdote condition were given the same materials as participants in the Statistics + Positive Anecdote condition, but now the anecdote is from a member of the trial for whom treatment was ineffective. Participants learned that "Jamie received a B-12 injection and her headaches, lack of energy, and inability to focus persisted." Critically, however, Jamie was not described as experiencing any side-effects as a consequence of her treatment.

Two design decisions are important to highlight: First, the anecdote contains no new information in the conditions that paired an anecdote with a statistic. This is because summary statistics already capture the success or failure of B-12 injections in the clinical trial and the anecdote concerns someone who was in the clinical trial. In other words, the anecdote contains no *additional* information over and above the statistic – the anecdote either describes the treatment as effective or ineffective and no other relevant information beyond this.

This point is related to a second design decision: Namely, in the Negative Anecdote condition, B-12 was described as failing as a treatment but not introducing any unwanted side-effects. Together, then, the negative anecdote should not affect participants' interpretation of the statistical information presented to them.

Posttest Questionnaire After completing the intervention portion of the task, participants completed a posttest questionnaire in which they were asked whether they believed B-12 injections were an effective medical treatment.

As noted, one possibility is that when the stakes are high for a given medical decision, people may be more susceptible to anecdotal information leading them to ignore strong statistical information. To this end, we also included two additional questions in the posttest questionnaire. First, participants were asked how likely it was they would try B-12 injections on a five-point Likert scale. Second, they were asked how likely they were to give B-12 injections to their child (if applicable). It's possible that a negative anecdote would exhibit a stronger negative effect on people's reasoning about their child compared to themselves because people are more risk averse when it comes to making decisions that impact their children's health (e.g., Brody, Annett, Scherer, Perryman, & Cofrin, 2005; Johnson, Özdemir, Mansfield, Hass, Siegel, & Sands, 2009).

Predictions

We predicted that participants in the Statistics + Positive Anecdote condition would be most likely to think that B-12

injections were effective as a treatment for chronic headaches—the positive anecdote would make salient the statistical summary information. This outcome would suggest that health communication experts could include similar positive anecdotes to increase people's uptake of statistical information (Allen et al., 2000). In contrast, we were unsure whether the Statistics condition or the Positive Anecdote condition would differ from each other, though the Statistics condition objectively contains much stronger evidence.

Of particular interest was how participants would respond to the negative anecdote in the Statistics + Negative Anecdote condition. One possibility is that presenting participants with a negative anecdote could raise the salience of the inefficacy of B-12 injections. However, we were unsure to what extent a single negative anecdote could impact people's use of the statistical summary information.

Results

We tested our predictions by fitting a Bayesian multivariate ordinal regression model regressing B-12 beliefs (i.e., efficacy beliefs, willingness to try B-12, and willingness to give these injections to their children) on Condition (Reference = Positive Anecdote condition) and pretest beliefs about the efficacy of B-12. Following the recommendations of Bürkner and Charpentier (2018), we modeled pretest as a monotonic effect because the ordinal nature of this predictor. The model is specified below in brms syntax:

```
mvbind(B12,TryB12,ChildB12) ~ Condition
+ mo(Pretest) + (1|p|Subject)
```

Bayesian analyses formulate model parameters as probability distributions wherein the posterior distribution for a parameter θ is computed via the prior and the likelihood of θ. To model the joint probability distribution of participants' responses, we specified the following regularizing priors over the possible effects each parameter could have on the response variable:

Experiment 1 - Priors

$\beta_{Intercept[1]} \sim \mathcal{N}(.5,.5)$
$\beta_{Intercept[2]} \sim \mathcal{N}(1.09,.5)$
$\beta_{Intercept[3]} \sim \mathcal{N}(2.94,.5)$
$\beta_{Intercept[4]} \sim \mathcal{N}(4.59,.5)$
$\beta_{Pretest} \sim \mathcal{N}(2,4)$
All remaining β were distributed as $\mathcal{N}(0,1)$
$\Omega_k \sim LKJ(1)$ where Ω_k is a correlation matrix of group-level parameters
Group-level parameters were distributed as $t(3,0,10)$

This model revealed that the Positive Anecdote, Statistics, and Statistics + Positive Anecdote conditions did not materially differ from each other (see Figure 1). However, the negative anecdote in the Statistics + Negative Anecdote condition caused participants to ignore the statistical information, despite the fact that (1) the statistic already summarizes the information contained in the negative anecdote and (2) the negative anecdote in no way suggests that the protagonist suffered a side-effect as a result of taking B-12 injections, $b_{B12} = -1.28$, 95% CI [−1.83, −0.73]; $b_{Try} = -0.79$, 95% CI [−1.44, −0.15]; $b_{Child} = -1.39$, 95% CI [−2.07, −0.72]. A subsequent analysis interacting pretest beliefs with condition provided no evidence for an interaction between these predictors.

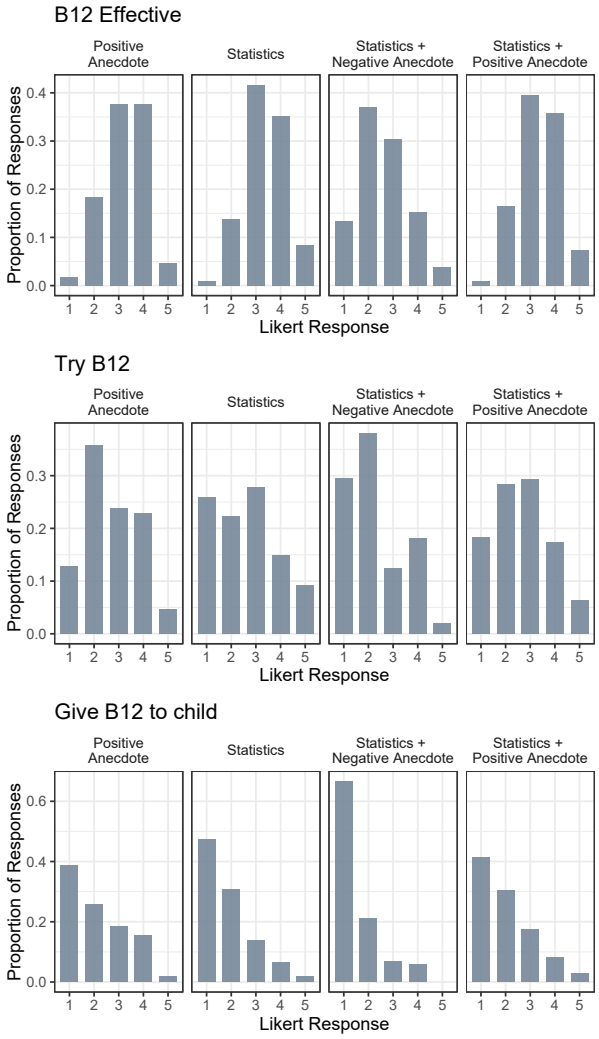


Figure 1: B-12 injection beliefs across conditions in Experiment 1. Higher Likert scale choices indicate more favorable attitudes towards B12 injections. The figure indicates that participants in the Statistics + Negative Anecdote condition had less favorable attitudes towards B12 injections after the intervention relative to participants in the other three conditions.

Altogether, these findings suggest that (negative) anecdotal information affected participants' beliefs. A single positive anecdote carried the same evidential weight as a study describing a double-blind clinical trial with 1,000 participants, though it appears that it did not affect beliefs

additively—the Statistics + Positive Anecdote condition did not differ from the Positive Anecdote condition nor the Statistics condition. More worrisome was the effect of the negative anecdote on participants' reasoning about compelling statistical evidence. One negative anecdote, in effect, caused people to dismiss strong statistical evidence, even though the anecdote implied no negative side effects and contained no additional information over and above the information carried by the statistics.

Experiment 2

Experiment 1 suggested that people's beliefs about the efficacy of B-12 injections are affected by anecdotal information. In Experiment 2, we sought to further understand the impact of anecdotes on medical decision-making. Given that a single negative anecdote can undo, as it were, strong statistical evidence, we sought to determine what would reduce the impact of this negative anecdote. Consequently, we tested whether presenting participants with both a positive and negative anecdote paired with statistical information would lead participants to primarily attend to the statistical information about the efficacy of B-12 injections in treating chronic headaches. Reading contradictory anecdotal information should indicate to participants that a different evidence source (in this case, the statistics) is needed to come to an informed belief about B-12 injections.

Method

Participants

We recruited 492 participants through Amazon's Mechanical Turk (50% women, M_{age}= 36 years old). Participants were paid $0.50 for participating in the study. After excluding participants who missed questions checking their attention, 431 participants remained in our sample. Our exclusion criteria were determined a priori and were in accordance with our study preregistration.

Procedure

The procedure of Experiment 2 was similar to Experiment 1, with the exception of the conditions participants were assigned to. Namely, we replaced the Positive Anecdote condition with a Statistics + Positive & Negative Anecdotes condition to determine whether including a positive anecdote in conjunction with a negative anecdote would lead participants to focus on summary statistics.

We made two other changes in Experiment 2. First, participants in the Statistics condition were explicitly told *both* the inefficacy and efficacy rates of B-12 injections in treating chronic headaches. We did this to better equate the salience of the inefficacy rate in the Statistics condition to the conditions in which the negative anecdote appeared. Specifically, participants read that "After a two-year trial with 1,000 participants, their study revealed that B-12 injections failed to work for 12.7% of participants and worked for 87.3%." Second, we changed the Likert scale for our posttest questions regarding the likelihood of trying B-12 injections and giving B-12 to one's child. These were changed to a six-point Likert scale which ranged from 1 = "Very unlikely" to 6 = "Very likely".

Predictions

As in Experiment 1, we predicted that participants in the Statistics + Positive Anecdote condition would tend to have the most positive beliefs towards B-12 injections. We predicted that when participants in the Statistics condition are explicitly presented with the rate of ineffectiveness, this would raise the salience of the inefficacy of B-12 injections. In turn, this may reduce overall endorsement of the efficacy of B-12 injections relative to the Statistics + Positive Anecdote condition. Finally, we sought to examine whether inclusion of the positive anecdote with the negative anecdote in the Statistics + Positive & Negative Anecdotes condition would cause participants to primarily attend to the statistical information they received. We suspected that the presence of the positive anecdote would not entirely undercut the effect of the negative anecdote on participants' judgments.

Results

As in Experiment, we fit a multivariate regression model regressing B12 attitudes on Condition (Reference = Statistics condition) and Pretest beliefs. We set priors on intercepts based on posterior estimates from Experiment 1.

Experiment 2 - Priors

$\beta_{Intercept[1]} \sim \mathcal{N}(-1.38, .5)$
$\beta_{Intercept[2]} \sim \mathcal{N}(1.09, .5)$
$\beta_{Intercept[3]} \sim \mathcal{N}(2.19, .5)$
$\beta_{Intercept[4]} \sim \mathcal{N}(4.59, .5)$
$\beta_{Pretest} \sim \mathcal{N}(2, 4)$
All remaining β were distributed as $\mathcal{N}(0, 1)$
$\Omega_k \sim LKJ(1)$
Group-level parameters were distributed as t(3,0,10)

These analyses replicated the effects of Experiment 1, showing that (1) the Statistics and Statistics + Positive Anecdote conditions did not differ from each other and (2) that participants in the Statistics + Negative Anecdote condition were more likely to discount the statistical evidence from the clinical trial (see Figure 2), b_{B12} = −1.34, 95% CI [−1.89, −0.80]; b_{Try} = −0.45, 95% CI [−1.22, 0.28]; b_{Child} = −0.94, 95% CI [−1.75, −0.19]. The positive anecdote in the Statistics + Positive and Negative Anecdotes condition, however, did not consistently improve participants' integration of the statistical information, and in some cases, did not differ at all from when participants only received the negative anecdote (see Figure 2), b_{B12} = −1.40, 95% CI [−1.93, −0.87]; b_{Try} = −0.22, 95% CI [−0.96, 0.52]; b_{Child} = −0.02, 95% CI [−0.74, 0.74]. These effects again did not interact with people's pretest attitudes towards B12 vaccines.

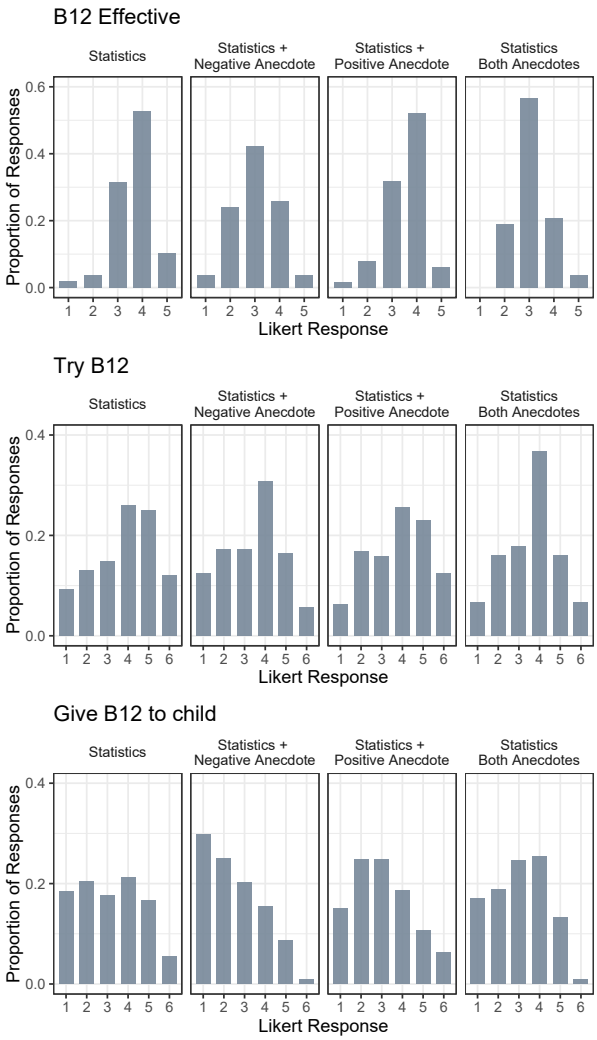


Figure 2: B-12 injection beliefs across conditions in Experiment 2. Higher Likert scale choices indicate more favorable attitudes towards B12 injections. The figure indicates that participants in the Statistics + Negative Anecdote and Statistics + Both Anecdotes conditions had less favorable attitudes towards B12 injections after the intervention relative to participants in the other conditions.

Experiment 3

Experiment 2 revealed that presentation of a negative anecdote raises the salience of the inefficacy of B-12 injections. This effect was not consistently negated by positive anecdotal information, raising the question of what means could undercut negative anecdotal information.

Experiment 3 sought to address two questions. First, we addressed the possibility that participants did not realize the anecdote they read was about a person in the study. Our hope was that by visually showing participants that the anecdote they read was about a person in the study we could rule out the possibility that a negative anecdote had its effects just in virtue of it being *new*, negatively-valenced information. Second, inspired by recent work, Experiment 3 tested whether a visual aid would reduce the impact of the negative anecdote on participants reasoning by making the strength of the summary statistics more salient. Several recent studies suggest that icon arrays, for instance, can improve understanding of scientific consensus (Lewandowsky, Gignac, & Vaughan, 2013; Nyhan & Reifler, 2018). Thus, Experiment 3 used an icon array to reduce the effect of the negative anecdote on people's beliefs.

Participants

We recruited 1,622 participants through Amazon's Mechanical Turk (54% women, M_{age}= 38). Participants were paid \$0.50 for participating in the study. After excluding participants who missed questions checking their attention, 1,539 participants remained in our sample. Our exclusion criteria were determined a priori and were in accordance with our study preregistration.

Procedure

The procedure of Experiment 3 was similar to that of Experiments 1 and 2. Participants were randomly assigned to one of four conditions in a 2 (Icon Array: Present or Absent) $\times$ 2 (Negative Anecdote: Present or Absent) between-subjects design. All four conditions included the summary statistical information from the Statistics condition in Experiment 1, allowing us to internally replicate our results in a larger sample.

In the Icon Array only condition, participants first read the statistic about the efficacy of B-12 injections as a medical treatment. They were then shown an icon array showing the success rate of B-12 in 100 people. Participants were then told:

> "This image is a depiction of the effectiveness of B-12 as a medical treatment. Imagine 100 people received B-12 injections. The blue figures represent participants that would benefit from the B-12 injections. The green figures represent participants who would fail to benefit from the B-12 injections."

In the Icon Array + Negative Anecdote condition, participants received the same information as the Icon Array only condition but were then told they would read about the experience of one of the subjects in the study and an icon array was displayed with one of the participants circled (see Figure 3), clearly indicating that the anecdote was from someone who participated in the clinical trial.

Predictions

We predicted that we would replicate the effect of negative anecdotes on participants' acceptance of strong statistical evidence, as we found in Experiments 1 and 2. We also predicted that in the Icon Array + Negative Anecdote condition, the presence of the icon array would weaken the effect of the negative anecdote (indicating an Icon $\times$ Anecdote interaction).

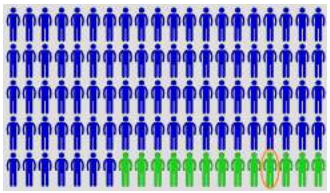


Figure 3: The icon array used in the Icon Array + Negative Anecdote condition in Experiment 3. For the Icon Array only condition, the same array was presented, but without the orange circle.

Results

We fit a Bayesian multivariate regression model regressing B12 attitudes on the interaction between Icon Array (Reference = No Array) and Anecdote (Reference = No Anecdote), controlling for pretest beliefs:

```
mvbind(B12,TryB12,ChildB12) ~
Anecdote*Array + mo(Pretest) +
(1|p|Subject)
```

Experiment 3 - Priors

$\beta_{Intercept[1]} \sim \mathcal{N}(-1.38, .5)$
$\beta_{Intercept[2]} \sim \mathcal{N}(1.09, .5)$
$\beta_{Intercept[3]} \sim \mathcal{N}(2.19, .5)$
$\beta_{Intercept[4]} \sim \mathcal{N}(4.59, .5)$
$\beta_{Pretest} \sim \mathcal{N}(4, 2)$
All remaining β were distributed as $\mathcal{N}(0, 1)$
$\Omega_k \sim LKJ(1)$
Group-level parameters were distributed as $t(3,0,10)$

We replicated the effects of Experiments 1 and 2, showing that a negative anecdote affected participants' integration of statistical information, b_{B12} = −1.49, 95% CI [−1.83, −1.16]; b_{Try} = −1.21, 95% CI [−1.74, −0.70]; b_{Child} = −0.87, 95% CI [−1.31, −0.45]. Consistent with prior work, we also found that providing an icon array improved people's integration of statistical information (b_{B12} = 1.02, 95% CI [0.69, 1.35]; b_{Try} = 0.89, 95% CI [0.39, 1.42]; b_{Child} = 0.86, 95% CI [0.42, 1.30]), but we observed little evidence for an interaction between these factors, b_{B12} = 0.00, 95% CI [−0.46, 0.45]; b_{Try} = −0.10, 95% CI [−0.77, 0.58]; b_{Child} = −0.33, 95% CI [−0.89, 0.24]. These results suggest that the negative anecdote nonetheless impacted people's reasoning even when an icon array was present and removed all ambiguity that the anecdote concerned someone who was in the clinical trial.

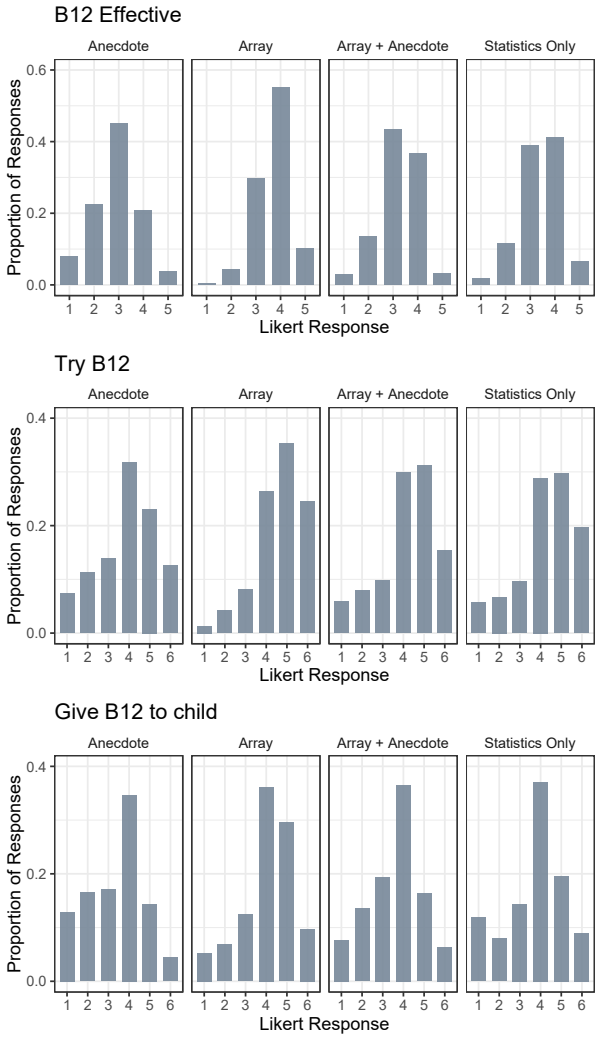


Figure 4: B-12 injection beliefs across conditions in Experiment 3. Higher Likert scale choices indicate more favorable attitudes towards B12 injections. The figure indicates that participants in the Anecdote condition had less favorable attitudes towards B12 injections than participants in the other three conditions.

Discussion

People have access to more medical information than ever before. From journal articles to online forums, people must determine what information is relevant and reliable to make medical decisions. How do people make these decisions? In three experiments, we tested how people reason about a medical treatment when provided with statistical or anecdotal information. In Experiment 1, we found that a negative anecdote caused people to ignore strong statistical information even though the anecdote involved no negative side effects—indeed, the information presented in the anecdote was already captured by the summary statistics presented to participants. In Experiment 2, we explored whether providing a positive anecdote in addition to a negative anecdote would counteract the effect of the negative anecdote. We found that emphasizing a positive outcome of a clinical trial did not consistently undo the effect of the negative anecdote. In Experiment 3, we found that introducing icon arrays improved integration of statistical information overall, but even in this case, anecdotal information negatively impacted people's beliefs. This

suggests that a single negative anecdote can carry substantial, unwarranted weight when making a medical decision.

It is striking that a negative anecdote led people to discount strong summary statistics even though the patient was described as suffering no negative side effects because of their treatment. Indeed, we were careful to describe B-12 injections as failing to benefit people who participated in the clinical trial. In reality, many medical treatments involve an element of risk, and some treatments can even involve severe side effects. In these situations, anecdotes that contain new information and highlight side effects would, if anything, yield a larger negative impact on people's ability to properly integrate statistical information. We can see evidence of these effects today: In 2019, vaccine hesitancy was listed as one of the top ten threats to global health (World Health Organization, 2019). In 2018 only 91.1% (compared to the recommended 95%) of children in the United States who were eligible for vaccines received the MMR (measles, mumps, and rubella) vaccine (Centers for Disease Control and Prevention, 2019). 2018 saw the second-highest number of measles cases since 2000. In part, vaccine hesitancy is a consequence of (1) people relying on discredited research linking vaccines to autism and (2) improper reliance on anecdotal information spread in forums purporting to demonstrate the side effects vaccines can wreak on young children (Powell et al., 2018). Our findings can help us make sense of this tendency. Anecdotes carry more weight than they should, as evidenced by the fact that they affected people's reasoning even when they were captured by summary statistics and involved no side effects, highlighting a serious obstacle to public health and demanding new interventions to overcome people's tendency to rely on anecdotal reasoning more than they should.

References

Allen, M., Bruflat, R., Fucilla, R., Kramer, M., McKellips, S., Ryan, D., & Spiegelhoff, M. (2000). Testing the persuasiveness of evidence: Combining narrative and statistical evidence. *Communication Research Reports*, *17*(4), 331–336.

Allen, M., & Preiss, R. (1997). Comparing the persuasiveness of narrative and statistical evidence. *Communication Research Reports*, *14*(2), 125–131.

Brody, J., Annett, R., Scherer, D., Perryman, M., & Cofrin, K. (2005). Comparisons of adolescent and parent willingness to participate in minimal and above-minimal risk pediatric asthma research protocols. *Journal of Adolescent Health*, *37*(3), 229–235.

Bürkner, P., & Charpentier, E. (2018). Monotonic effects: A principled approach for including ordinal predictors in regression models. *PsyArXiv*.

Bürkner, P. C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, *10*(1), 395–411.

Centers for Disease Control and Prevention. (2019). Measles cases and outbreaks. Retrieved from https://www.cdc.gov/measles/cases-outbreaks.html.

Dahlstrom, M. (2014). Using narratives and storytelling to communicate science with nonexpert audiences. *Proceedings of the National Academy of Sciences*, *111*(4), 13614–13620.

Hornikx, J. (2005). A review of experimental research on the relative persuasiveness of anecdotal, statistical, causal, and expert evidence. *Studies in Communication Sciences*, *5*(1), 205–216.

Hornikx, J. (2018). Combining anecdotal and statistical evidence in real-life discourse: Comprehension and persuasion. *Discourse Processes*, *55*(3), 324–336.

Johnson, F., Özdemir, S., Mansfield, C., Hass, S., Siegel, C., & Sands, B. (2009). Are adult patients more tolerant of treatment risks than parents of juvenile patients? *Risk Analysis*, *29*(1), 121–136.

Lewandowsky, S., Gignac, G. E., & Vaughan, S. (2013). The pivotal role of perceived scientific consensus in acceptance of science. *Nature Climate Change*, *3*(4), 399 – 404.

Liddell, T., & Kruschke, J. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology*, *79*, 328–348.

McElreath, R. (2016). *Statistical rethinking: A bayesian course with examples in R and Stan*. Boca Raton, FL: CRC Press.

Nyhan, B., & Reifler, J. (2018). The roles of information deficits and identity threat in the prevalence of misperceptions. *Journal of Elections, Public Opinions and Parties*, 1 – 23.

Powell, D., Weisman, K., & Markman, E. (2018). Articulating lay theories through graphical models: A study of beliefs surrounding vaccination decisions. *Proceedings of the 40th Annual Conference of the Cognitive Science Society*.

Rodriguez, F., Rhodes, R., Miller, K., & Shah, P. (2016). Examining the influence of anecdotal stories and the interplay of individual differences on reasoning. *Thinking & Reasoning*, *22*(3), 274–296.

Shen, F., Sheer, V., & Li, R. (2015). Impact of narratives on persuasion in health communication: A meta-analysis. *Journal of Advertising*, *44*(2), 105–113.

World Health Organization. (2019). Ten threats to global health in 2019. Retrieved from https://www.who.int/emergencies/ten-threats-to-global-shealth-in-2019.

Controlling Attention To Solve Working Memory Tasks Using a Memory-Augmented Neural Network

T.S. Jayram[1], **Younes Bouhadjar**, **Tomasz Kornuta**[2], **Ryan L. McAvoy**, **Alexis Asseman**, and **Ahmet S. Ozcan**[3]

IBM Research AI, Almaden Research Center, San Jose, CA 95120

Abstract

We introduce a memory-augmented neural network, called Differentiable Working Memory (DWM), that captures some key aspects of attention in working memory. We tested DWM on a suite of psychology inspired tasks, where the model had to develop a strategy only by processing sequences of inputs and desired outputs. Thanks to novel attention control mechanisms called *bookmarks*, the model was able to rapidly learn a good strategy—generalizing to sequence lengths even two orders of magnitude larger than that used for training—allowing it to retain, ignore or forget information based on its relevance. The behavior of DWM is interpretable and allowed us to analyze its performance on different tasks. Surprisingly, as the training progressed, we observed that in some cases the model was able to discover more than one successful strategy, possibly involving sophisticated use of memory and attention.

Introduction

Keeping information in mind after it is no longer present in the environment is critical for all higher cognitive behaviors. *Working memory* (WM) is the term used for this ability, which is distinct from the storage of vast amount of information in long-term memory (Baddeley, 2003; Oberauer, 2009). The two main distinguishing characteristics of WM are the limited capacity (3-5 items) (Cowan, 2001) and temporary retention (secs-minutes). Hence, WM is not a storage per se, but a mental workspace utilized during planning, reasoning and solving problems. Most psychologists differentiate WM from "short-term" memory because it can involve the manipulation of information rather than being a passive storage (Cowan, 2017). Along the same lines, Engle, Tuholski, Laughlin, and Conway (1999) argued that WM is *all about the capacity for controlled, sustained attention in the face of interference or distraction.* Attention-control is a fundamental component of the WM system and probably the main limiting factor for capacity (Conway & Engle, 1994; Engle & Kane, 2004). Consequently, the inability to effectively parallel process two-attention demanding tasks limits our multitasking performance severely.

Over the past several decades psychologists have developed tests to measure the individual differences in WM capacity and better understand the underlying mechanisms. These tests have been carefully crafted to focus on the specific aspects of WM such as task-driven attention control, interference and capacity limits (Oberauer & Lin, 2017). The best known and successfully applied class of tasks for measuring WM capacity is the "complex span" paradigm. The challenge presented by complex span tasks is recalling the list of items, despite being distracted by the processing task. Studies show that individuals with high WM capacity are less likely to store irrelevant distractors (Vogel, McCollough, & Machizawa, 2005) and they are better at retaining task-relevant information (Maxcey-Richard & Hollingworth, 2013). Developing task-driven strategies for cognitive control are essential for the effective use of WM.

In the past there were several attempts to build computational models that mimic the operation of a human working memory (Henson, 1998; Farrell & Lewandowsky, 2002; Oberauer & Lewandowsky, 2011; Lemaire & Portrat, 2018). For example, Burgess and Hitch (1999, 2005) used a shallow neural network and put the emphasis on Hebbian-like learning rules that enabled the model to achieve similar behavior to the one achieved by human subjects. In those works the experimental paradigm was the serial recall task, which is limited in testing the complex processing and attention component of WM. One notable exception was (Oberauer, Lewandowsky, Farrell, Jarrold, & Greaves, 2012), where the authors focused on the complex span task.

The power of maintaining information over time has also been recognized by the AI community. Starting with the basic recurrent neural network architectures (Elman, 1990; Hopfield & Tank, 1986) followed by the introduction of gating mechanisms (Hochreiter & Schmidhuber, 1997), the research has recently moved onto more complex architectures with memories (Graves, Wayne, & Danihelka, 2014; Joulin & Mikolov, 2015; Weston, Chopra, & Bordes, 2015; Graves et al., 2016; Santoro, Bartunov, Botvinick, Wierstra, & Lillicrap, 2016; Gulcehre, Chandar, & Bengio, 2017). These models are typically applied to tasks (e.g. associative recall, bAbI QA (Weston, Bordes, Chopra, & Mikolov, 2015)) that require a complex mixture of long-term memory (episodic and semantic) and working memory. In the human brain, these kinds of memory systems are distinct: working memory is instantiated in multiple interconnected areas with the prefrontal cortex playing a major role (Constantinidis & Klingberg, 2016), whereas for episodic memory the hippocampus is the critical structure (Fortin, Agster, & Eichenbaum, 2002). Studying these mechanisms separately is necessary to disen-

[1] jayram@us.ibm.com. Primary contact author.

[2] tkornut@us.ibm.com

[3] asozcan@us.ibm.com

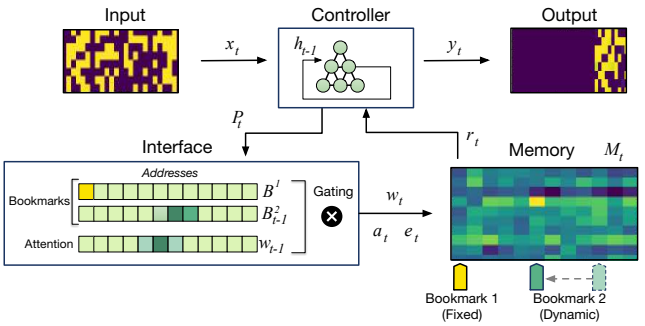


Figure 1: Illustration of the operation of the DWM Model

tangle the contributions of each memory system and develop a detailed understanding of human intelligence.

In this work, we take inspiration from biological and computational models of working memory to develop an *artificial* neural network model augmented with external memory, called *Differentiable Working Memory (DWM)*. We provide the DWM model a set of generic mechanisms to encode inputs, access the memory, and control its attention over the memory contents. Key to this design is a new attention control mechanism in memory, called *bookmarks*, which helps in dealing with interference. In contrast to previous works, we applied our model to a variety of psychometry-inspired tasks, each requiring the system to control its attention in a slightly different way. We show that the DWM is capable of solving these diverse tasks by looking only at input-output pairs using the provided attention mechanisms. The model is easy to train and accurately generalizes to sequences two orders of magnitude longer than the training data. We also describe the strategies that the DWM develops during training and demonstrate that the bookmarks can effectively deal with interference in complex tasks. Finally, we show that the DWM is also capable of learning multiple strategies during training and, moreover, develop better strategies in the presence of memory scarcity.

Differentiable Working Memory (DWM)

The operation of Differentiable Working Memory (DWM) model is presented in Fig. 1. As a memory-augmented neural network (MANN), the DWM has three main components: a *controller*, an *external memory* and an *interface* between the two (Zaremba, Mikolov, Joulin, & Fergus, 2016). The interface is composed of several *attention* mechanisms that the controller learns to use by generating appropriate parameters for accessing the external memory. The procedure is sketched in Algorithm 1. We describe the main steps, Lines 4–7, below in order of significance.

Attention control. The memory consists of N addresses, each storing a vector of real numbers of length L. Thus the memory contents are given by an $L \times N$ matrix of real numbers. The read and write operations share a single attention

Algorithm 1 Operation of Differentiable Working Memory

1: Initialize:
- the hidden state h_0 and memory array M_0
- the read/write attention vector w_0
- bookmarks $\{B_0^i : i = 1,2,\ldots,K\}$

2: **for** $t \in \{1,2,\ldots,T\}$ **do**
3: **Memory read**: $r_t \leftarrow M_{t-1} w_{t-1}$
4: **Controller**: $h_t, P_t \leftarrow \phi(x_t, r_t, h_{t-1})$
5: **Memory update**: $M_t \leftarrow \mathsf{update}(w_{t-1}, P_t, M_{t-1})$
6: **Attention control**: $w_t, \{B_t^i\} = \mathsf{attn}(w_{t-1}, \{B_{t-1}^i\}, P_t)$

mechanism. Further, we use *soft addressing*: let w denote a non-negative weight vector of dimension N whose components sum up to 1. Each component indicates the *relative strength* with which a value (i.e. a vector of dimension L) will be read/written at the corresponding address.

The behavioral studies indicate that people can access memories sequentially (Singh, Tiganj, & Howard, 2018). For that reason we have decided to add a mechanism based on circular convolution, similar to the one used in Neural Turing Machine (NTM) (Graves et al., 2014), enabling it to shift attention over memory:

$$w_t = \text{convolution}(w_t^g, s_t), \tag{1}$$

where w_t^g and w_t are the vectors of attention weights over cells in memory at time t before and after shifting, and s_t is a shift vector outputted by the controller. We also apply a weight sharpening step typically used after the shifting; as we observed, it seemed to be crucial for models using circular convolution to converge properly.

The Embedded-Processes Framework (Cowan, 1988) assumed the presence of Focus of Attention (FOA). In this model the items in the FOA were interpreted as pointers to the representations stored in the long-term memory rather then being the actual representations themselves. Inspired by this concept, we introduced a new attention mechanism called *bookmarks* that store the system's attention at previous time steps. This is recorded in K bookmark vectors $\{B_t^i : i = 1,2,\ldots,K\}$ at time t. The first bookmark $B^1 := B_t^1$ has a time-independent *fixed* attention to a single address so that the model maintains a reference frame for memory. The remaining bookmarks are *dynamic*: at time t, the DWM must decide whether to remember its previous (read/write) attention w_{t-1} by recording it in a bookmark, as:

$$B_t^i = g_t^i w_{t-1} + (1 - g_t^i) B_{t-1}^i, \quad i = 2,3,\ldots K, \tag{2}$$

where the gating parameter g_t^i is emitted by the controller. As discussed later, we found in our experiments that even limiting to only two bookmarks (one *fixed* and one *dynamic*), the model could still solve all tasks.

The DWM must also decide before moving sequentially whether it wishes to return to a previous bookmark. For this

purpose we once again use a gating mechanism, this time in a slightly more sophisticated form:

$$w_t^g = \delta_t^0 w_{t-1} + \sum_{i=1}^{K} \delta_t^i B_{t-1}^i, \tag{3}$$

where $\delta_t^i, i = 1,2,\ldots K$ are gating parameters emitted by the controller. These gating parameters are scalars, normalized using a softmax function.

The DWM attention control incorporates the presented mechanisms by applying equations (3), (2), and (1) in order.

Memory read and update. We use the standard formula for soft attention, e.g., (Weston, Chopra, & Bordes, 2015), that computes the read vector r_t from memory M_{t-1}:

$$r_t = M_{t-1} w_{t-1} \tag{4}$$

For memory update, we decided to use the simple erase-add scheme derived from NTM (Graves et al., 2014):

$$M_t = M_{t-1} \circ (E - e_t \otimes w_t) + a_t \otimes w_t, \tag{5}$$

where E is a matrix of all ones, e_t and a_t are vectors of content to be erased and added to memory, respectively. The parameters e_t and a_t are emitted by the controller.

Controller. The controller's role is to process inputs so as to produce outputs as well as interface parameters. In DWM we use a single-layer recurrent neural network controller:

$$h_t = \sigma(W_h[x_t, h_{t-1}, r_t]), \tag{6}$$

where x_t denotes the current input and h_{t-1} and r_t are the hidden state and vector read from memory in the previous time step, respectively. To prevent the controller from acting as a separate working memory, the hidden state size is chosen to be smaller than that of a single input vector (in all of our experiments it was set to 5). The output logits, y_t and interface vector P_t are produced similarly as:

$$y_t = W_y[x_t, h_{t-1}, r_t] \tag{7}$$

$$P_t = W_P[x_t, h_{t-1}, r_t] \tag{8}$$

W_h, W_y and W_P are the only trainable parameters of our DWM model. The interface vector P_t contains all of the parameters that control reading, writing, and the attention mechanisms. Denoting the unprocessed parameters from the interface with a hat, the full list of parameters in P_t is as follows:

- The write vector $a_t \in \mathbb{R}^{N_M}$
- The erase vector $e_t = \sigma(\hat{e}_t) \in [0,1]^{N_M}$
- The shift vector $s_t = \text{softmax}(\text{softplus}(\hat{s})) \in [0,1]^3$
- The bookmark update gates $g_t^i = \sigma(\hat{g}_t^i) \in [0,1]^{K-1}$
- The attention update gate $\delta_t^i = \text{softmax}(\hat{\delta}_t^i) \in [0,1]^{K+1}$
- The sharpening parameter $\gamma = 1 + \text{softplus}(\hat{\gamma}) \in [1,\infty]$

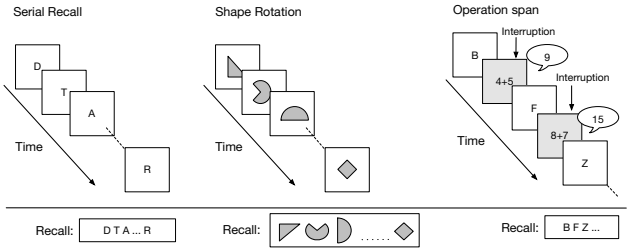


Figure 2: Exemplary tasks for testing the performance of human working memory

Psychometric tasks for working memory

Over last several decades cognitive psychologists have developed many psychometric tests (Conway et al., 2005) to measure the performance of human WM (See Fig. 2 for examples). These tasks are mainly sequential and typically divided into verbal and visuospatial domains. Given that diversity and various categorizations by different researchers, we built a taxonomy of tasks (Fig. 3) and carefully selected tasks that seem to be the most representative for a given category. First order categorization is based on the number and complexity of tasks. For simple tasks, the presence of data manipulation is the next level sub-category, with Serial Recall being a prime example of a task without manipulation. The tasks requiring manipulation we further categorized into spatial and temporal domains. The complex tasks involve multiple sequential inputs or sub-tasks but not necessarily imply "multi-tasking". We follow the framework of Clapp, Rubens, and Gazzaley (2009) to distinguish the sources of goal interference, i.e. Distraction (to-be-ignored) and Interruption (i.e. multi-tasking). For example, in Operation Span (Fig. 2c) the subjects had to attend and process the summation (Interruption) even though they did not need to recall the results afterwards, whereas in Reading Span (Daneman & Carpenter, 1980) subjects had to read sentences and recall the last word of each one. In addition to the classical psychometric tasks, we introduced several tasks testing the effectiveness of attention control in memory (Ignore, Forget and Scratch Pad). As a result, a suite of tasks presented in Table 1 emerged.

The input to every task is a sequence of items. As we wanted to be agnostic to audio/visual preprocessing, we have implemented those tasks using sequences of randomly generated *binary patterns* (vectors of bits) as items (instead of words/images). At a higher level, we view the input as a *concatenation* of various subsequences that represent different functional units of processing. For all simple tasks, there is only one type of subsequence, and the output will be reproduced from the memory with or without manipulation. The complex tasks may involve a secondary set of subsequences, optionally requiring immediate output as indicated in the Forget and Operation Span tasks.

Additionally, we use a constant-sized set of special items (called *command markers*) to both mark the beginning of a subsequence as well as indicate its functional type. It is im-

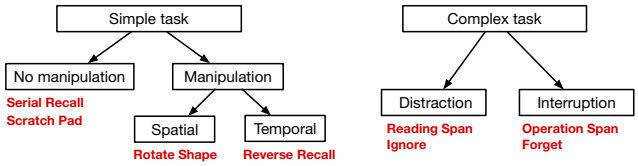


Figure 3: Taxonomy of working memory tasks

portant to note that the system does not know a priori what kind of operation is associated with a given type of marker and must learn that from data. We ignored markers in Table 1 to keep the description simple. Also, note that such markers are also commonly employed in the psychometric tests, e.g., see McNab and Klingberg (2008).

Experimental results

We evaluated the performance of DWM on the proposed tasks and compared it to two models: LSTM (Long Short-Term Memory) (Hochreiter & Schmidhuber, 1997), considered as a classical baseline for sequential problems, and DNC (Differentiable Neural Computer) (Graves et al., 2016) being one of the state-of-the-art MANN models. In our implementation we used the MI-Prometheus (Kornuta et al., 2018) framework built on top of PyTorch (Paszke et al., 2017). During training we used the Adam (Adaptive Momentum) optimizer (Kingma & Ba, 2014) and (average) binary cross-entropy as the loss function. We apply early stopping based on validation loss (10^{-4}). Additionally, we terminate training when the number of training episodes reach 100,000 where a single episode involves processing a batch of sequences. The size of batch was a hyper-parameter that was tuned along with training rate for each model using validation loss as the reference.

	Task		(I)input/(O)output sequences	
Simple	Serial Recall	I:	$x_1 x_2 \ldots x_n$	$\_\_\ldots\_$
		O:	$\_\_\ldots\_$	$x_1 x_2 \ldots x_n$
	Scratch Pad	I:	$\mathbf{x}_1 \mathbf{x}_2 \ldots \mathbf{x}_k$	$\_$
		O:	$\_\_\ldots\_$	$\mathbf{x}_k$
	Reverse Recall	I:	$x_1 x_2 \ldots x_n$	$\_\_\ldots\_$
		O:	$\_\_\ldots\_$	$x_n x_{n-1} \ldots x_1$
	Rotate Shape	I:	$x_1 x_2 \ldots x_n$	$\_\_\ldots\_$
		O:	$\_\_\ldots\_$	$x_1^{\circlearrowleft} x_2^{\circlearrowleft} \ldots x_n^{\circlearrowleft}$
Complex	Reading Span	I:	$\mathbf{x}_1 \mathbf{x}_2 \ldots \mathbf{x}_k$	$\_\_\ldots\_$
		O:	$\_\_\ldots\_$	$z_1 z_2 \ldots z_k$
	Ignore	I:	$\mathbf{x}_1 \mathbf{y}_1 \ldots \mathbf{x}_k \mathbf{y}_k$	$\_\ldots\_$
		O:	$\_\_\ldots\_\_$	$\mathbf{x}_1 \ldots \mathbf{x}_k$
	Operation Span	I:	$\mathbf{x}_1 y_1 \_\ldots \mathbf{x}_k y_k \_$	$\_\ldots\_$
		O:	$\_\_ y_1^{\circlearrowleft} \ldots \_\_ y_k^{\circlearrowleft}$	$\mathbf{x}_1 \ldots \mathbf{x}_k$
	Forget	I:	$\mathbf{x}_1 \mathbf{y}_1 \_\ldots \mathbf{x}_k \mathbf{y}_k \_$	$\_\ldots\_$
		O:	$\_\_ \mathbf{y}_1 \ldots \_\_ \mathbf{y}_k$	$\mathbf{x}_1 \ldots \mathbf{x}_k$

Table 1: Working Memory Tasks. A bold letter denotes a subsequence of items. The | sign indicates delay between input and output of the primary subsequence(s). Above, $x_i^{\circlearrowleft}$ denotes the circular shift of x_i by half its bitlength. In the Reading Span task, z_i is the last item of $\mathbf{x}_i$.

Task	Validation Accuracy Seq. Size 100 [%]			Test Accuracy Seq. Size 1000 [%]		
	LSTM	DNC	DWM	LSTM	DNC	DWM
Serial	53.3*	100	100	50.2*	64.6	100
Scr. Pad	71.3*	100	100	70.0*	75.0	100
Reverse	53.0*	50.6*	100	50.4*	50.2*	99.8
Rot. Shape	52.2*	100	100	50.2*	60.9	100
Read. Span	50.9*	53.4*	100	50.4*	49.0*	91.9
Ignore	56.1*	69.3*	100	50.9*	50.0*	90.0
Op. Span	58.2*	79.2*	99.9	51.3*	53.6*	99.6
Forget	55.9*	69.4*	98.9	50.5*	49.9*	94.1

Table 2: Summary of experimental results. The first column is the average of validation accuracies achieved by the models for 10 training runs on each task. The second column is the average of test accuracies achieved by models that converged during training. For the majority of tasks, the DNC and LSTM models did not converge. In those cases (indicated with *) we report scores of the best (even though diverged) model.

As stated in the introduction, the main question we wanted to answer was whether a model can learn an algorithm to solve a task. In case of tasks presented in Table 1, this implies that the model should generalize over the sequence lengths. For that reason, our methodology assumed that we will use different lengths of sequences for training (up to 10), validation (exactly 100) and testing (exactly 1000). Although human WM does not have the capacity to handle 1000 items, our goal was to show that the model truly generalizes in that actually develops an effective memory strategy, i.e., it learns an algorithm to solve the task.

All models achieved perfect accuracies on training sequences. However, as presented in Table 2, LSTM and DNC struggled with generalization to longer sequences. Besides, the DWM models converged faster, requiring less than 5000 episodes in most cases (exemplary convergence plot is presented in Fig. 4). The convergence speed is associated with number of trainable parameters of those models (the DWM controller had 1066, the DNC had 4,792, whereas for LSTM baseline we used stacked LSTM with 3 layers and over 5 million trainable weights). Please note that *fair comparison* simply made no sense, as the LSTM and DNC models with less trainable parameters could not even learn to generalize over short (i.e. training) sequences. Aside of that, we hypothesize that the DNC had problems with convergence because of the

complexity of its attention and memory management mechanisms (the *Temporal Link Matrix*, in particular).

Analysis of strategies for solving tasks

The proposed tasks require the models to develop different strategies for solving them. For example, ignoring distractions without encoding them in the memory is arguably the best strategy to minimize memory consumption. On the other hand, for a complex task with an interruption (i.e. multitasking), the secondary task cannot be ignored and may require extensive memory usage. In this case, the best strategy might be to forget (e.g. erase or overwrite) the secondary information as soon as possible in order to maintain sufficient memory capacity for the main task.

During the training and testing of all of the tasks reported in Table 2 we provided sufficient memory size, so that the system could store all the encoded input items in the memory (if it has chosen to). However, limitation of the memory size can force the system to develop more memory efficient strategies, thus we decided to investigate that issue further.

Strategies for the Scratch Pad task The goal of the Scratch Pad task is to recall only the last input subsequence.

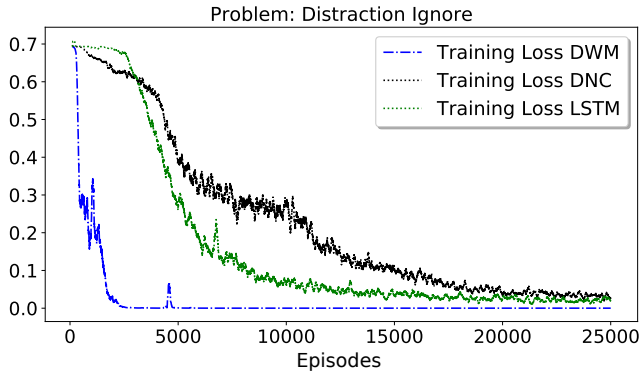


Figure 4: Convergence of the best models on Ignore task

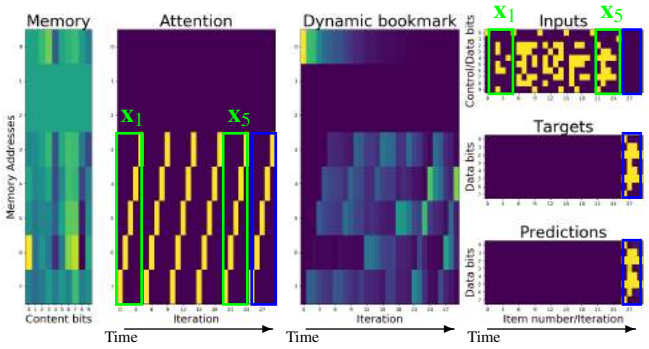


Figure 5: Overwrite strategy developed by DWM for Scratch Pad (episode 969). Memory plot contains a snapshot of the memory content from the last iteration, whereas the other ones present concatenation of states from consecutive iterations (evolution in time)

Given the DWM mechanisms, we expect two possible strategies for the model to learn in order to solve this task.

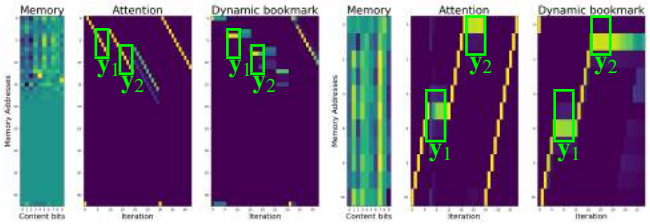
The "Expand" strategy exploits the fact that memory can be used in a similar way to a circular buffer, storing each consecutive subsequences one after the other in the memory. In this case the model should write each subsequence, then place the *dynamic bookmark* at the start of given subsequence, and then update the bookmark position to the beginning of the next subsequence. Finally, when the model receives a command marker indicating it needs to recall, it should recall the attention associated with that *dynamic bookmark* and then retrieve consecutive items one by one by shifting.

The "Overwrite" strategy for the Scratch Pad relies on the fact that when a new subsequence appears, the elements from the previous one can be discarded. The model could exploit this by learning to recall attention stored in the *fixed bookmark* every time it processes a command marker denoting the next subsequence, which will result in overwriting the previous subsequences until the system is told to recall. This strategy may be interpreted as memory saving, as the system reuses the same addresses and overwrites them repeatedly.

To our (initial) surprise, the model *always* developed the Overwrite strategy, irrespective of the memory size (i.e. as long as the memory size was sufficient to fit all the encoded items of a single subsequence). An exemplary run of an early training episode is presented in Fig. 5. Note that memory addresses 1 and 2 remain unchanged and the model stores consecutive items of subsequences $\mathbf{x}_1$ to $\mathbf{x}_5$ in the same addresses 3-7. After analyzing several runs, we hypothesize that overwriting was simpler to learn for this task because: a) both for storing and recalling the command markers, the model had to learn exactly the same behavior: recalling the attention stored in the *fixed bookmark*, b) for every other input item it had to shift by one address location with the circular convolution. As a result, it could converge rapidly by disregarding the control (update, recalling) of the *dynamic bookmark* (in the later training episodes the *dynamic bookmark* was typically "following" the current attention, despite it wasn't recalled at all).

Strategies for the Ignore task The main goal of the Ignore task is to test the retention capabilities of the system in the presence of distractors. For this task the input consists of two types of subsequences $\mathbf{x}$ and $\mathbf{y}$, where the system is supposed to ignore all $\mathbf{y}_i$ and at the end recall $\mathbf{x}_i$ one by one in the order of their appearance. The task can be solved with two strategies which we call "Overwrite" and "Skip".

The "Overwrite" strategy involves overwriting of the distractors, similarly to the "Overwrite" from Scratch Pad task. It assumes that model will store the consecutive items in memory and use the bookmark for moving its attention to the first address containing $\mathbf{y}$ to be overwritten. The difference is, however, that in here the model must learn to use the *dynamic bookmark* for that purpose. Our experiments with sufficient memory have shown that the system can learn this strategy. Exemplary plot from one of the final training episodes (Fig. 6a) shows that the *dynamic bookmark* re-



(a) Overwrite strategy (episode 17000)

(b) Skip strategy (limited memory, episode 5614)

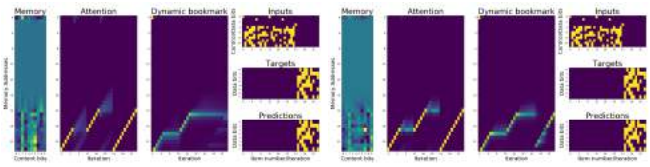
Figure 6: Strategies developed by DWM for solving the Ignore task (two different training runs)

tains its attention while processing items from $\mathbf{y}_1$ and $\mathbf{y}_2$. As soon as the command marker indicating $\mathbf{x}$ appears, the model jumps back its attention to the *dynamic bookmark* and starts to overwrite memory content. Finally, when the recall marker appears, it recalls the attention stored in the *fixed bookmark*.

The "Skip" strategy involves ignoring elements within the $\mathbf{y}$ subsequences, i.e. *skipping* writing them to memory. Our experiments with limited memory have shown that the model could also learn this strategy. Exemplary plots from the final episode from one of the training runs are presented in Fig. 6b. Note that in this case the model has learned to keep its attention focused on a single address for all items of $\mathbf{y}_i$ and shift attention only for items belonging to $\mathbf{x}_i$.

That behavior of the model that mastered the "Skip" strategy seems to be more difficult from the operational point of view. In the "Overwrite" strategy the system develops a reactive behavior, i.e. it always performs convolutional shift except for the rare cases when it hits the command marker – at that point it has to retrieve attention from one or the other bookmark. In the "Skip" strategy the command markers for $\mathbf{x}$ and $\mathbf{y}$ activate one of two distinct operation modes that will be executed for the whole subsequence until hitting the next marker, i.e. for $\mathbf{x}$ attention is supposed to move to the next address, whereas for $\mathbf{y}$ it is supposed to stay at the same position. The only way to perform this is that the controller must learn how to *carry the information about the current operation mode* from one iteration to another in its hidden state, which is more difficult to learn.

We performed several experiments to support that hypothesis. In Fig. 7 we present two episodes from one of the training runs when the operation of the system seems to be evolving from one strategy to the other. At the early stages of the training (Fig. 7a) we can observe that the attention shift with the circular convolution is active for both types of input subsequences, whereas the dynamic bookmark already learned how to *follow* attention for $\mathbf{x}$ and *freeze* for $\mathbf{y}$. As learning to shift attention is crucial for learning both storing and recall, model has to master that first. However, once achieved, it seems to switch to different operation mode. Obviously, learning two modes is simpler for *dynamic bookmark*, as it possesses simpler gating mechanism and cannot shift its attention. As the training progresses (Fig. 7b) the model finally learns to freeze its attention when processing $\mathbf{y}$ subsequences.



(a) episode 1757: utilization of Overwrite strategy

(b) episode 17685: utilization of Skip strategy

Figure 7: Evolution of the strategy developed by DWM, when learning the Ignore task (during a single training run, intentionally used the same verification sequence in both cases)

Conclusion

We have demonstrated that DWM has the appropriate attention mechanisms to tackle psychology-inspired tasks. When compared to existing models such as DNC, LSTM it appeared to manage generalization to much longer sequences. Besides, after careful step-by-step analysis we discovered that the model is able to develop more than one strategy to control attention and use its memory resources for a given task. While some strategies are harder to learn, DWM can develop them by first finding *any* working strategy and then gradually modifying it towards a different one as learning progresses. Why the model seems to prefer some strategies is intriguing and worth further investigation. Another direction is to incorporate this mechanism into a larger system in order to solve tasks that require both working and long-term memory.

References

Baddeley, A. (2003). Working memory: Looking back and looking forward. *Nature Reviews Neuroscience*, *4*(10), 829–839.

Burgess, N., & Hitch, G. (1999). Memory for serial order: a network model of the phonological loop and its timing. *Psychological review*, *106*(3), 551.

Burgess, N., & Hitch, G. (2005). Computational models of working memory: putting long-term memory into context. *Trends in cognitive sciences*, *9*(11), 535–541.

Clapp, W., Rubens, M., & Gazzaley, A. (2009). Mechanisms of working memory disruption by external interference. *Cerebral Cortex*, *20*(4), 859–872.

Constantinidis, C., & Klingberg, T. (2016, may). The neuroscience of working memory capacity and training. *Nature Reviews Neuroscience*, *17*, 438.

Conway, A., & Engle, R. (1994). Working memory and retrieval: A resource-dependent inhibition model. *Journal of Experimental Psychology: General*, *123*(4), 354.

Conway, A., Kane, M., Bunting, M., Hambrick, D., Wilhelm, O., & Engle, R. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic bulletin & review*, *12*(5), 769–786.

Cowan, N. (1988). Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information-processing system. *Psychological bulletin*, *104*(2), 163.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*(1), 87–114.

Cowan, N. (2017). The many faces of working memory and short-term storage. *Psychonomic Bulletin and Review*, *24*(4), 1158–1170.

Daneman, M., & Carpenter, P. (1980). Individual differences in working memory and reading. *Journal of verbal learning and verbal behavior*, *19*(4), 450–466.

Elman, J. L. (1990). Finding structure in time. *Cognitive science*, *14*(2), 179–211.

Engle, R., & Kane, M. (2004). Executive attention, working memory capacity, and a two-factor theory of cognitive control. *Psychology of learning and motivation*, *44*, 145–200.

Engle, R., Tuholski, S., Laughlin, J., & Conway, A. (1999). Working memory, short-term memory, and general fluid intelligence: a latent-variable approach. *Journal of experimental psychology: General*, *128*(3), 309.

Farrell, S., & Lewandowsky, S. (2002). An endogenous distributed model of ordering in serial recall. *Psychonomic bulletin & review*, *9*(1), 59–79.

Fortin, N. J., Agster, K. L., & Eichenbaum, H. B. (2002, mar). Critical role of the hippocampus in memory for sequences of events. *Nature Neuroscience*, *5*, 458.

Graves, A., Wayne, G., & Danihelka, I. (2014). Neural turing machines. *arXiv preprint arXiv:1410.5401*.

Graves, A., Wayne, G., Reynolds, M., Harley, T., Danihelka, I., Grabska-Barwińska, A., ... others (2016). Hybrid computing using a neural network with dynamic external memory. *Nature*, *538*(7626), 471.

Gulcehre, C., Chandar, S., & Bengio, Y. (2017). Memory augmented neural networks with wormhole connections. *arXiv preprint arXiv:1701.08718*.

Henson, R. (1998). Short-term memory for serial order: The start-end model. *Cognitive psychology*, *36*(2), 73–137.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Hopfield, J. J., & Tank, D. W. (1986). Computing with neural circuits: A model. *Science*, *233*(4764), 625–633.

Joulin, A., & Mikolov, T. (2015). Inferring algorithmic patterns with stack-augmented recurrent nets. In *Advances in neural information processing systems* (pp. 190–198).

Kingma, D., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kornuta, T., Marois, V., McAvoy, R. L., Bouhadjar, Y., Asseman, A., Albouy, V., ... Ozcan, A. S. (2018). Accelerating machine learning research with mi-prometheus. In *NeurIPS 2018 MLOSS Workshop.*

Lemaire, B., & Portrat, S. (2018). A computational model of working memory integrating time-based decay and interference. *Frontiers in psychology*, *9*, 416.

Maxcey-Richard, A. M., & Hollingworth, A. (2013). The strategic retention of task-relevant objects in visual working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(3), 760.

McNab, F., & Klingberg, T. (2008). Prefrontal cortex and basal ganglia control access to working memory. *Nature Neuroscience*, *11*(1), 103–107.

Oberauer, K. (2009). *Chapter 2: Design for a working memory* (1st ed.). Elsevier.

Oberauer, K., & Lewandowsky, S. (2011). Modeling working memory: A computational implementation of the time-based resource-sharing theory. *Psychonomic Bulletin & Review*, *18*(1), 10–45.

Oberauer, K., Lewandowsky, S., Farrell, S., Jarrold, C., & Greaves, M. (2012). Modeling working memory: An interference model of complex span. *Psychonomic bulletin & review*, *19*(5), 779–819.

Oberauer, K., & Lin, H.-y. (2017). An interference model of visual working memory. *Psychological Review*, *124*(1), 1–39. doi: 10.1037/rev0000044

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., ... Lerer, A. (2017). Automatic differentiation in pytorch. In *NIPS Autodiff Workshop.*

Santoro, A., Bartunov, S., Botvinick, M., Wierstra, D., & Lillicrap, T. (2016). Meta-learning with memory-augmented neural networks. In *International conference on machine learning* (pp. 1842–1850).

Singh, I., Tiganj, Z., & Howard, M. (2018). Is working memory stored along a logarithmic timeline? converging evidence from neuroscience, behavior and models. *Neurobiology of learning and memory*.

Vogel, E. K., McCollough, A. W., & Machizawa, M. G. (2005). Neural measures reveal individual differences in controlling access to working memory. *Nature*, *438*(7067), 500.

Weston, J., Bordes, A., Chopra, S., & Mikolov, T. (2015). Towards ai-complete question answering: A set of prerequisite toy tasks. *CoRR*, *abs/1502.05698*.

Weston, J., Chopra, S., & Bordes, A. (2015). Memory networks. In *International conference on learning representations (ICLR).*

Zaremba, W., Mikolov, T., Joulin, A., & Fergus, R. (2016). Learning simple algorithms from examples. In *International conference on machine learning* (pp. 421–429).

Pedagogical Questions Empower Exploration

Anishka Jean[1], Emily Daubert[1], Yue Yu[2], Patrick Shafto[1], & Elizabeth Bonawitz[1]

[1] Rutgers University, Newark NJ 07102
[2] National Institute of Education, Singapore

Abstract

Children are motivated to explore and learn about the world, but they vary in their degree of perseverance during exploration. A growing body of literature suggests that is malleable from an early age. Here, we ask whether pedagogical questions empower children to persevere during a difficult problem-solving task with a blicket detector machine. Previous research has shown that when presented with a blicket detector, asking children "pedagogical questions" promotes more exploratory behaviors compared to direct instruction. A pedagogical question is a question asked by a knowledgeable person, whose intention is to teach rather than to seek an answer to that question. The current study examines whether pedagogical questions influence the amount of time children spend problem-solving before seeking help, compared to direct instruction, overheard pedagogical questions, and overheard questions asked by a naive other. We predicted that children who were asked a pedagogical question prior to having the opportunity to play with a machine would persevere longer in trying to make it work, and would be less likely to ask for help. Results suggest that pedagogical questioning encourages children to attempt more hypothesis-test interventions in an effort to make the machine work. Results will be discussed in terms of the role of pedagogical questioning in promoting perseverance during problem-solving.

Keywords: pedagogy; pedagogical question; perseverance

Introduction

Young children are curious and creative problem-solvers. They are motivated to explore and learn about how things work, why they work, and, if necessary, how to fix them. However, there is a great deal of variation in children's perseverance during problem solving, and this characteristic is malleable. Children's perseverance during exploration is likely influenced by a number of factors, including the nature of their interactions with adults. For example, children are more likely to persevere during a difficult task after watching an adult model persevere (Leonard, Lee, & Schulz, 2017). Here, we investigate the particular qualities of adult instruction that may promote children's perseverance during exploration and learning.

Previous research suggests that when children are faced with a difficult task, they rely on their interactions with and observations of adults to guide their exploration and problem solving efforts. For example, preschool-aged children readily detect and utilize pedagogical cues (e.g., the teacher's knowledgeability; the intentionality of the teacher's demonstration; the social context of the learning scenario; etc.) to guide deductive reasoning, exploration, and learning about the world (Bonawitz, Shafto et al., 2011; Buchsbaum, Gopnik, Griffiths, & Shafto, 2011; Butler & Markman, 2014). For example, in Bonawitz, Shafto et al. (2011) children were assigned to one of a few conditions that differed in the social presentation of information. In the Pedagogical condition, a knowledgeable and helpful adult demonstrated one function on a complex-looking toy. After this, children were presented with the toy and allowed to explore. The Pedagogical condition was contrasted with several other conditions including an Accidental condition in which a naive demonstrator accidentally elicited the function, and an Interrupted condition in which the demonstrator was interrupted before it was clear they were completed. Results showed that children in the Pedagogical condition explored less than children in the other conditions, consistent with the explanation that the pedagogical demonstrations lead to high confidence that there was little to be learned beyond the demonstrated function.

Thus, children rely on adults' pedagogical cues to guide learning. What are the particular qualities of these cues that might be most relevant to perseverance during exploration? One pedagogical tool whose efficacy has been of particular interest as of late is *pedagogical questioning*. A pedagogical question is a question asked by a knowledgeable person, whose intention is to teach rather than to seek an answer to that question. Recent research indicates that pedagogical questioning yields effective knowledge transmission, while also promoting exploration (Yu, Landrum, Bonawitz, & Shafto, 2018). This is in contrast to *direct instruction*, another common pedagogical tool. In direct instruction, information is communicated directly from a knowledgeable teacher to a naïve learner. Past research shows that while direct instruction is beneficial for sharing information, this can come at the expense of children's subsequent exploratory learning (Bonawitz & Shafto et al., 2011). This is likely due to the expectation that is often brought into pedagogical learning scenarios that good teachers should provide all the necessary evidence for the learner to be able to solve the problem (Shafto & Goodman, 2008). However, these implications for exploration appear not to be induced by pedagogical questioning. For instance, in Yu et al. (2018), children were shown a novel toy, which had many possible functions, and were told that the experimenter knows all about the toy and how it works. In the direct

instruction (DI) condition, the experimenter told children to push the button on the novel toy; in the pedagogical question (PQ) condition, children were asked to think about "what does this button do?" While children were equally likely to discover the key function (i.e., the button) in both conditions, children also spent longer playing with the toy and discovered more additional functions in the PQ condition (Yu et al., 2018). These results support the claim that pedagogical questions empower children to engage in exploratory behaviors, while direct instruction may constrain exploration.

One limitation of this study however, is that children in both the PQ and DI conditions were at ceiling in their ability to discover the target function of the toy, so it is difficult to assess the extent to which pedagogical questions might differentially influence the pursuit of learning about queried information. In other words, when children are tasked with a simple problem, which was readily solvable (the button immediately generated the effect), it was not possible to explore the potentially different influence of PQs and DI on persistence for learning *targeted information.* Of course, once the goal was complete (in this case discovering the buttons function), it remained important to ask what next steps children would take with the toy. In Yu's paper, children in the PQ condition then went on to discover significantly more functions of the toy as compared to the DI condition, providing important insights into the power of PQs in supporting longer term exploration. Nonetheless, it remains unclear if pedagogical questions also empower children to persevere and engage in exploratory behaviors in service of the queried information. The current study addressed this question by presenting children with a more challenging problem (i.e., an unsolvable problem). Children were tasked with discovering how to make a (deactivated) blicket detector machine work, a procedure inspired by past literature (Gweon & Schulz, 2011).

We hypothesized that that pedagogical questions might promote persistence in problem-solving during a difficult task for a few reasons. First, pedagogical questions that are directed to the child may empower them to feel as though the expectation is that they can figure the machine out on their own, rather than having to seek help from an adult. That is, by asking "what do you think?" a Pedagogical Question could imply that the questioner believes the child can discover the answer. Second, pedagogical questions may encourage children to engage in exploratory behaviors during a difficult problem-solving task because questions do not limit the number or nature of potential solutions to the problem at hand. In contrast, direct instruction may hinder children's creative exploration of potential solutions by "over focusing" children in on the directed content.

One alternative to the claim that pedagogy is the driving factor behind pedagogical questions, is the possibility that *any* kind of question might lead to greater perseverance. In order to control for this possibility, we included a condition in which children overheard a *naive* confederate asking a question to an experimenter (Overheard Naive Question condition; ONQ). In this way, the exact language of the question is matched, but the crucial difference is that in the ONQ condition, the question-asker was not knowledgeable (i.e., was known by the child to have no knowledge of how to make the machine work), where as in the PQ condition, the question-asker was knowledgeable. A person who does not know the answer is not naturally thought of as a having the goal of teaching the outcome because they do not know the outcome. Thus, in the current study, any potential effects could be attributed to the pedagogical nature of the question, and not just the question itself.

Another alternative to the claim that pedagogy is the driving factor behind pedagogical questions, is the idea that any pedagogical question, no matter who it is being directed to, might promote greater perseverance. In order to control for this possibility, we included another condition in which children overheard an experimenter asking a pedagogical question to a confederate (Overheard Pedagogical Question; OPQ). In this way, the exact language of the pedagogical question is matched, but the pedagogical question is not child-directed This condition allows us to isolate the influence of the child-directed nature of the pedagogical question asked in the PQ condition form the mere influence of overhearing a pedagogical question as in the OPQ condition. . Although pedagogical questions have been found to promote exploratory behaviors in children, these questions may only influence exploration if they are child-directed.

Thus, we hypothesized that pedagogical questions, compared to direct instruction and overheard naive questions, would encourage children to persevere and play with the machine longer before seeking help than children in the DI, OPQ, and ONQ conditions, although we expect no condition differences in the amount of time it takes children to recognize that there is something wrong with the machine. We might expect that pedagogical questions promote perseverance in problem-solving during a difficult task for a few reasons. First, pedagogical questions may empower children to feel as though the expectation is that they can figure the machine out on their own, rather than having to seek help from an adult (i.e., the experimenter). The amount of time the child spends engaging with the machine before reaching out for help and the number of hypothesis tests the child performs during exploration could make this claim evident. Second, pedagogical questions may encourage children to engage in meaningful exploratory behaviors during a difficult problem-solving task because questions do not limit the number or nature of potential solutions to the problem at hand. The number of unique actions and the variability in the nature of hypothesis tests performed on the toy could make this claim evident. In contrast, DI may hinder children's creative exploration of potential solutions. We also predicted that children in all conditions would be equally quick to notice that something was wrong with the machine. The time to first look could make this claim evident.

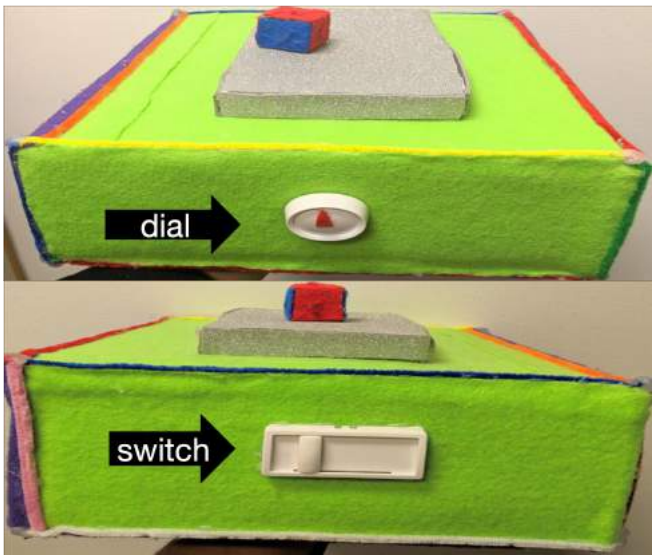


Figure 1. Blicket detector machine used by children in the experiment. The switch and dial "sliders" are highlighted. Another feature of the machine is the platform on the top of the machine. The multicolored block is the accessory that appears to activate the machine when placed on the platform, even though it's the remote control transmitter that's actually activating the machine.

Method

Participants

Participants were 100 4- to 6-year-old children (M_{age} = 57.91 months, *SD* = 6.62 months; Range = 48.26 - 78.77 months, 53% female) recruited from preschools and public sites located in Essex County, New Jersey, one of the most racially and socioeconomically diverse counties in the United States. Prior to the study, consent from the sites, participating families, and the Rutgers University - Newark internal review board were obtained. Based on a power analysis, and as preregistered[1], we recruited 25 children per condition. Children were only included in the final sample if they met the following criteria: 1) English was their primary language, and 2) there was no outside interference during the testing session. Participants were randomly assigned to one of four conditions: Pedagogical Question (PQ); Overheard Pedagogical Question (OPQ), and Overheard Naïve Question (ONQ), and Direct Instruction (DI). Age was matched across conditions.

Materials

A novel "blicket detector" machine that was approximately 13.25'' X 10.5'' X 5'' with a switch on one side and a dial on the other was created (Figure 1). A wireless doorbell receiver hidden inside the machine produced a C major arpeggio when the experimenter surreptitiously pressed the button on the remote control transmitter which connected to the wireless doorbell. The switch and dial "sliders" on either side of the machine were inert. There was also a multi-colored block whose sole purpose was to "activate" the machine when placed on a shiny platform on top.

Procedure

Children were introduced to the machine by the experimenter and told that "the way the toys works is, the way the toy works is, when the block is on the platform and the toy is all set up right, the toy goes". In all conditions, a confederate was seated next to the child during the introduction to the machine and the demonstration of its use. In the PQ and DI conditions, the experimenter first demonstrated directly to the child how the block can activate the machine (but did not show whether, and which of, the switch and the dial should be positioned for it to work.) The experimenter then showed the machine separately to the confederate in another location in the room, so that the child could not see what was being changed on the machine. In the OPQ and ONQ conditions, the child also observed the block activating the machine (but was not shown the role of the switch or dial like in the PQ and DI conditions), however in the OPQ and ONQ conditions, the confederate did not observe the role of the switch and dial. Specifically, the experimenter walked away from the child and confederate to a corner of the room to activate the machine while verifying with the confederate that they cannot see what the experimenter was doing with the machine. Thus, in the OPQ and ONQ conditions, the confederate was never shown how the machine worked with respect to the switch and dial. In all four conditions, the experimenter explained that something about the machine had been changed so that now the block would not activate the machine.

Critically, the prompt given to the child prior to the free play period varied by condition. In the PQ condition, the experimenter asked the child, "what happens if you change *these sliders?*" while moving the sliders (the switch and dial) on either side of the machine. In the DI condition, the experimenter instructed the child by telling them to "change *these sliders* to see what happens" while moving the sliders. In the OPQ condition, the experimenter asked the confederate, "what happens if you change *these sliders?*" while moving the sliders, controlling for the pedagogical nature of the question. In the ONQ condition, the confederate picked up the machine and asked the experimenter, "what happens if you change *these sliders?*" while moving the sliders, controlling for children's awareness of a knowledgeable other by having the confederate ignorant to how the machine works. Immediately, following these prompts, children engaged in a free play period described below.

Free Play

The child was then given five min to play with the machine and was informed that the confederate would be there if they needed anything. Specifically, children in all conditions

[1] Link to preregistration: https://aspredicted.org/j3ah7.pdf

were told, "You can go ahead and play with this. I have to go over there to write something down for a couple minutes, but [confederate's name] will be here if you need anything!" Then,the experimenter sat behind the child (out of sight) and pretended to write, while the confederate sat next to the child and pretended to read a book. During this play period, the machine would not activate at all, regardless of whether the child followed the experimenters' instructions or suggestions. Playtime was ended once one of the following occurred: five minutes elapsed, the child verbally requested help, the child did not interact with the machine for 15 consecutive s twice in a row, or the child asked to stop playing. All sessions were video-recorded.

Outcome Measures

All videos were coded by a trained coder for three outcome measures: (1) time to first look, (2) time before help-seeking, (3) number of unique actions, (4) number of hypothesis-test interventions and (5) variability of hypothesis-test interventions.

Time to first look The time to first look was the amount of time the child spent attempting to activate the machine before looking at the confederate or experimenter for the first time.

Time Before Help-seeking The time to help-seeking was the amount of time child spent attempting to activate the toy on their own before verbally requesting help from either the confederate or the experimenter.

Number of hypothesis tests (Perseverance) The number of hypothesis tests was the number of times a child performed any other intervention on the machine that may involve the traditional use of the switch or dial "sliders" right before attempting activation with the block on the machine. An intervention required a manipulation of some factor of the toy or block and critically an attempted activation that immediately followed whereby the child placed the block on the machine.

Number of unique actions The number of unique actions was the number of unique manipulations to the machine that did not involve the traditional use of the switch or dial "sliders."

Variability of hypotheses tested The variability of hypotheses tested was the number of unique hypothesis-tests performed on the machine. For example, if the child tried adjusting the dial and then placed the block on the activator, that would count as a single hypothesis test. However, a second manipulation of the same dial with a following block test would not count as a unique intervention and so would not additionally increase the total variability score beyond the initial attempt.

Table 1
Unique Actions

Unique Actions		
Placing the multicolored block on the platform	Placing the block and/or hand inside the toy	Flicking the block on top of the platform
Changing the colored side of the block to a different color	Removing the dial "slider" from the toy	Placing another toy/object on top of the platform
Placing the block on other parts of the toy	Removing the switch "slider" from the toy	Rolling the block like a dice
Moving the block on the platform without lifting up the block	Switching the tune of the toy by pressing the battery inside the toy	Placing the block on 4 corners of the platform
Rubbing the block on the shiny platform of the toy	Placing the switch "slider" on the platform of the toy	Lifting up the shiny platform foam paper
Lifting the toy up	Placing the dial "slider" on the platform of the toy	Shaking the toy
Placing the toy on its side	Playing with the block itself	Shaking the block
Putting the toy upside down	Touching the hot glue gun balls on a corner of the toy	

Results

Time to first look. Our first question concerned whether children were equally likely to visually "check-in" with the adults during the free play period. The rationale for this measure was that we hypothesized children might initially look to the confederate or experimenter when it became apparent that the machine was no longer activating as expected. Indeed, in all four conditions, on average, children looked to an adult within the first minute of play (PQ: 25 s; DI: 44 s; OPQ: 27 s; ONQ: 26 s), and there were no significant differences in the total amount of time before this first look, $F(3, 96) = 1.12$, $p = .347$, suggesting that children were equally capable of detecting the activation issue across conditions.

Overall time playing. Our second question pertained to whether children would stop playing with the toy earlier in the DI, OPQ and ONQ conditions. We hypothesized that children in the PQ condition might play with the machine longer than children in the DI, OPQ, and ONQ conditions based on previous research (Yu et al., 2018), in which children explored the novel (functioning) machine longer in the Pedagogical Question condition. However, we observed no significant differences between conditions in the amount of time children spent playing with the machine, $F(3, 96) = 2.53$, $p = .062$, indicating that children in all conditions played with the machine for approximately the same amount of time ($M_{PQ} = 240.72$; $SD_{PQ} = 95.78$; $M_{DI} = 188.16$; $SD_{DI} = 106.13$; $M_{OPQ} = 205.56$; $SD_{OPQ} = 112.03$; $M_{ONQ} = 159.52$; $SD_{ONQ} = 112.32$). However, the trend here for children to play longer in the PQ conditions is suggestive.

Number of hypothesis tests (Perseverance) Third, we asked whether the number of hypothesis-test interventions during children's play differed significantly between conditions. Specifically, if Pedagogical Questions both empower children to pursue a relevant learning goal (in this case to discern why the machine is failing to activate) in the face of repeated failure then we would expect children in the

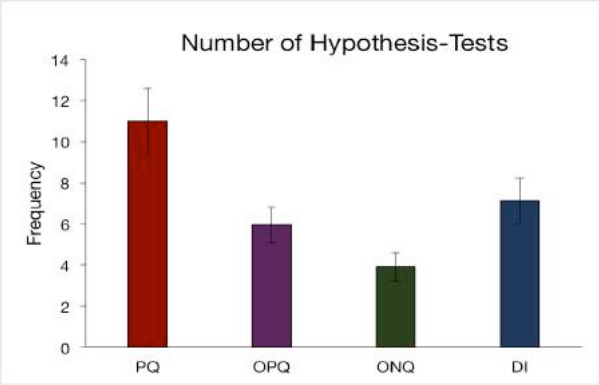


Figure 2. Children in the Pedagogical Question condition performed significantly more hypothesis tests during the play period than children in the Overheard Pedagogical Question, the Overheard Naive Question, and the Direct Instruction conditions.

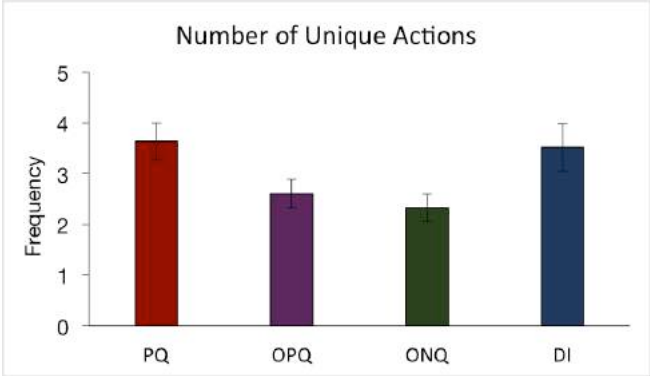


Figure 3. Children in the Pedagogical Question and Direct Instruction conditions performed significantly more unique actions during the play period than children in the Overheard Pedagogical Question and the Overheard Naive Question conditions.

PQ condition to perform more interventions on the machine. Indeed, the number of hypotheses tested significantly differed across conditions, $F(3, 96) = 7.03$, $p < .001$. Specifically, planned contrasts1 revealed that children in the PQ condition conducted significantly more hypothesis tests ($M = 11.00$; SD = 8.00) than children in the DI condition ($M = 7.12$; SD = 7.12), OPQ condition ($M = 5.96$; SD = 4.36) and ONQ condition ($M = 3.92$; SD = 3.49). There was no difference in the number of hypothesis tests between the children in the DI condition and children in the OPQ condition, $p = .417$, and there was no difference in the number of hypothesis tests between the children in the OPQ and ONQ conditions, $p = .074$. Thus, even though on average, children in all conditions were equally quick to notice something was wrong with the machine, and played with the machine for approximately the same amount of time, children in the PQ condition engaged in more hypothesis testing during this time, suggesting that PQs might both empower children to persevere in their exploratory causal testing attempts during play.

Number of unique actions Next, we asked if children were more likely to explore more different features of the machine overall depending on the type of instruction they were given. By virtue of pedagogical questions being questions, the variability in children's exploratory actions is not limited, leading us to predict children to show more variable exploration in the three question conditions. There were significant differences in the number of unique actions by condition, $F(3, 96) = 3.36$, $p = .022$. Planned contrasts revealed that children in the PQ ($M = 3.64$; $SD = 1.85$) and DI ($M = 3.52$; $SD = 2.37$) conditions performed more unique actions than children in the OPQ ($M = 2.60$; $SD = 1.44$) and ONQ ($M = 2.32$; $SD = 1.35$) conditions, $p = .003$. There were no significant differences in the number of unique actions between the PQ and DI conditions, $p = .842$, and there were no significant differences in the number of unique actions between the OPQ and ONQ conditions, $p = .481$. Contrary to our hypothesis, this suggests that the *child-directed* nature of pedagogical questions (and direct instruction), rather than the inquisitive nature of the input appears to promote variability during play. However, given that there were relatively few actions that might be attempted with the toy (unlike the Yu et al, 2018 novel toy study), this result should be interpreted with caution.

Variability of hypotheses tested Finally, we asked whether the variability of hypotheses tested specifically during children's play differed significantly between conditions. That is, if pedagogical questions both empower the pursuit of a relevant learning goal (in this case to discern why the machine is failing to activate), then we would expect children in the PQ condition to perform more different types of interventions on the machine. Overall, the number of different hypotheses tested significantly differed across conditions, $F(3, 96) = 4.08$, $p = .009$. Specifically, children in the PQ ($M = 2.28$; $SD = 1.21$) and DI ($M = 2.08$; $SD = 1.12$) conditions performed more variable hypothesis-tests than children in the OPQ ($M = 1.52$; $SD = .77$) and ONQ ($M = 1.52$; $SD = .65$) conditions, $p = .001$. There was no difference in the variability of hypothesis tests between the PQ and DI conditions, p = .546, and there was no difference in the variability of hypothesis tests between the OPQ and the ONQ conditions, p = .999 Again, *child-directed* conditions led to more variable exploration during play time.

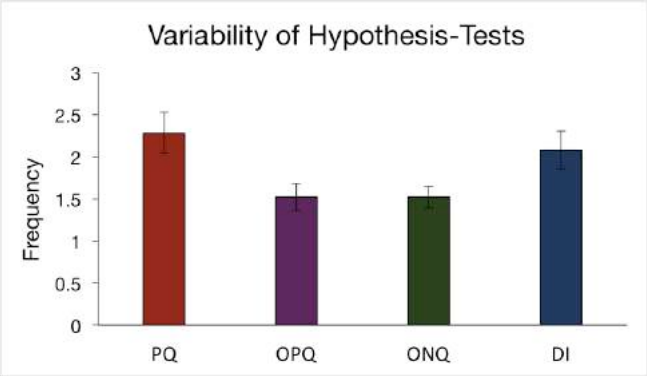


Figure 4. Children in the Pedagogical Question and Direct Instruction conditions demonstrated significantly more variability in their hypothesis tests during the play period than children in the Overheard Pedagogical Question and the Overheard Naive Question conditions.

Discussion

This study examined the effect of Pedagogical Questions on young children's perseverance during a difficult problem-solving task. First, we found that children in all conditions were equally likely to visually check-in with the adults during the free play period, thus children are equally quick to recognize that the machine was not functioning as expected. Second, we found that there were no differences in how long children played with the machine before reaching out for help. Despite recognizing a problem with the machine at equal rates and spending the same amount of time playing with the machine, , children in the PQ condition performed significantly more hypothesis tests, suggesting that prompting children with a Pedagogical Question may lead to their independently persevering through more failed attempts at problem solving before looking to others for help. Our results point to both the pedagogical nature of the question (rather than this effect being about questioning generally) as children in the ONQ condition demonstrated significantly fewer hypothesis tests prior to turning for help, and the child-directed nature of the question as children in the OPQ condition demonstrated significantly fewer hypothesis tests. Additionally, two surprising, but interesting findings indicate that when it came to promoting more variable exploration, as measured by the number of unique actions and the variability in hypothesis tests, the child-directed nature of the pedagogical input was crucial, as children in the two child-directed conditions(PQ and DI) demonstrated more variability during play time.s.

This study extends our understanding of the role of Pedagogical Questions in the preschool years by examining how pedagogical questions affect perseverance and variability during exploration when children are presented with a difficult problem. In the current study, there was a more obvious and specific goal for learners in contrast to Yu et al. (2018), which examined what additional, unbounded exploration children pursued after the initial goal was quickly completed. Classic debates contrast instruction with exploration in terms of their ability to foster learning (Bruner, Jolly, & Sylva, 1976; Csibra & Gergeley, 2009; Piaget, 1929; Singer, Golinkoff, & Hirsh-Pasek, 2008; Tomasello & Barton, 1994; Vygotsky, 1978). However, learning in the real world depends on myriad factors beyond learning content. Often learning comes down to hard work and trying many possible solutions. Whereas these previous debates centered around the material to be learned, at least as important is the effort required. Effective methods of promoting learning in the real world will engage both.

Pedagogical questions are particularly promising in this respect. Bonawitz and colleagues (2011) showed that instruction, though powerful for ensuring specific information is learned, has negative consequences for future learning by reducing exploration. Yu et al. (2018) showed that pedagogical questions offer a potentially promising solution by achieving the benefits of direct instruction without restricting exploration following completion of a goal. Here we have shown that pedagogical questions additionally foster learning by increasing the children's persistence in pursuit of solutions.

Pedagogical questions, questions asked by a knowledgeable person for the purpose of teaching, are a surprisingly simple approach. Demonstrations are easily converted into such questions. Given their simplicity, and the relevance to literatures in education and in question asking, it is interesting that they do not appear to have been explored previously. One possible reason is that these literatures tend to focus on behaviors that are easy to see. Pedagogical questions by definition depend on inferences about the questioner's knowledge and intent. For this reason, comparison with overheard questions is an important control and a powerful demonstration of the importance of latent social variables in understanding learning.

Our work on Pedagogical Questions is part of a broader movement beyond simple dichotomies such as direct instruction versus exploration. Recent research has proposed Guided Learning as a framework for considering learning as a dynamic, interactive, social activity (e.g. Hirsh-Pasek et al., 2015). Many aspects of this framework remain to be formalized; however, we believe Pedagogical Questions provide one compelling example of guidance. Pedagogical Questions foster learning not by telling the learner the answer, but by offering the learner strong guidance toward the answer. Many open questions remain regarding when Pedagogical Questions are most effective and how they fit into the broader Guided Learning framework. We leave these to future work.

In sum, this study supports the view that pedagogical questions promote learning. Children who are asked pedagogical questions persevere in service of a specific goal. These findings are particularly relevant for educators who can use pedagogical questions in their classrooms to enhance children's perseverance during challenging problem-solving activities.

References

Bonawitz, E., Shafto, P., Gweon, H., Goodman, N.D., Spelke, E., & Schulz, L. (2011). The double-edged sword of pedagogy: Instruction limits spontaneous exploration and discovery. *Cognition,* 120(3), 322-330. doi:10.1016/j.cognition.2010.10.001

Buchsbaum, D., Gopnik, A., Griffiths, T.L., & Shafto, P. (2011). Children's imitation of causal action sequences is influenced by statistical and pedagogical evidence. *Cognition, 120*(3):331-40. doi: 10.1016/j.cognition.2010.12.001

Butler, L.P., & Markman, E. M. (2014). Preschoolers use pedagogical cues to guide radical reorganization of category knowledge. *Cognition, 130*(1), 116-127. doi:10.1016/j.cognition.2013.10.002

Leonard, J.A., Lee, Y., & Schulz, L.E. (2017). Infants make more attempts to achieve a goal when they see adults persist. *Science,* 357(6357), 1290-1294. Doi:10.1126/science.aan2317

Shafto, P. & Goodman, N. (2008). Teaching games: Statistical sampling assumptions for learning in pedagogical situations. Proceedings of the 30th annual conference of the Cognitive Science Society.

Yu, Y., Landrum, A.R., Bonawitz, E., & Shafto, P. (2018). Questioning supports effective transmission of knowledge and increased exploratory learning in pre-kindergarten children. *Developmental Science,* 21(6). doi:10.1111/desc.12696

Bruner, J.; Jolly, A.; Sylva, K. (1976). Play-Its role in development and evolution. Basic Books Inc; New York.

Csibra G, Gergely G. (2009). Natural Pedagogy. *Trends in Cognitive Sciences, 13*(4):148–153.

Hirsh-Pasek, K., Zosh, J.M., Golinkoff, R.M., Gray, J.H., Robb, M.B., & Kaufman, J. (2015). Putting Education in "Educational" Apps: Lessons From the Science of Learning. *Psychological Science in the Public Interest, 16*(1) 3–34. doi:
10.1177/1529100615569721

Piaget, J. (1929). *The Child's Conception of the World.* New York: Harcort, Brace.

Singer, DG.; Golinkoff, MR.; Hirsh-Pasek, K. (2008). Play = Learning: How play motivates and enhances children's cognitive and social-emotional growth. New York: Oxford University Press.

Tomasello M, Barton M. (1994). Learning words in nonostensive contexts. *Developmental Psychology, 30*,639–650.

Vygotsky, L. (1978). Mind in Society. Cambridge, MA: Harvard University Press.

Targeted Mathematical Equivalence Training Lessens the Effects of Early Misconceptions on Equation Encoding and Solving

Kristen Johannes (kjohann@WestEd.org),
WestEd, 2470 Mariner Square Loop, Alameda, CA 94501 USA

Jodi Davenport (jdavenp@WestEd.org)
WestEd, 2470 Mariner Square Loop, Alameda, CA 94501 USA

Abstract

Many students fail to develop adequate understanding of mathematical equivalence in early grades, with detrimental consequences for later algebra learning. The change resistance account (McNeil, 2014) proposes that students struggle with equivalence because traditional arithmetic practice overexposes students to mathematical expressions where all the operators are on the left of the equal sign. Students erroneously believe the equal sign means to "do something" or "give the answer" – and fail to see equations as relations between two expressions. These operations-based misconceptions affect how they perceive, conceptualize, and approach math problems and interfere with developing correct understandings of equivalence. The current paper explores 1) are these misconceptions evident as encoding errors in second graders? 2) do item properties make specific error types more or less likely? 3) do misconceptions in encoding impact solving performance? and 4) can targeted training mitigate the effects of prior misconceptions on both equation encoding and solving? We identify a category of misconception-based encoding errors that negatively impacts equation solving and replicate findings that a conceptually rich research-based intervention program is maximally effective in training students to overcome problematic misconceptions.

Keywords: Mathematical representations; relational reasoning; mathematics education; randomized control trial

Introduction

How do early conceptions about equivalence impact children's ability to correctly encode, and later solve, arithmetic equations? Research suggests that understanding mathematical equivalence is a critical component of algebraic reasoning (Carpenter, Franke, & Levi, 2003; Charles, 2005; Knuth, Stephens, McNeil, & Alibali, 2006). However, the majority of US students fail to reason with and apply concepts of equivalence (McNeil & Alibali, 2005), making encoding errors when reconstructing mathematical equations (e.g., McNeil & Alibali, 2004), and interpreting the equal sign to mean "calculate the total" rather than "two amounts are the same" (e.g., Behr, Erlwanger, & Nichols, 1980).

McNeil and Alibali (2005; McNeil 2014) proposed a change-resistance account of children's difficulty with mathematical equivalence. Traditional arithmetic instruction, which focuses on procedures (i.e., solving problems such as 7 + 2 = _), reinforces a *misconception* of the equal sign as a request for an answer, which, in turn, interferes with the development of relational concepts. Most arithmetic problems in early elementary math curricula show operations (e.g., addition and subtraction) on the left of the equal sign and the "answer" on the right (Seo & Ginsburg, 2003; McNeil, 2008). Children detect and extract patterns from these examples and ultimately construct long-term memory representations. McNeil and Alibali characterize these representations as "operational patterns" as they reflect an understanding of arithmetic that focuses on the operators (e.g., +, -, ×, ÷) rather than the relational nature of mathematical expressions. Although default representations typically speed computation in the problem-solving contexts that children encounter most frequently, these representations may lead to difficulties when operational patterns are mistakenly transferred to similar, but non-applicable, problem types (e.g., Bruner, 1957).

Alibali and colleagues (Crooks & Alibali, 2013; McNeil & Alibali, 2004, 2005) have identified three different sub-patterns, described below, that reflect a distorted view of arithmetic and hinder conceptual understanding of equivalence and underlying mathematics. Once entrenched, children rely on these potentially misleading patterns when encoding, interpreting, and solving novel mathematics problems. In the current study, we group these three types of errors as "misconception errors" (see Table 2) to differentiate them from errors believed to stem from working memory constraints or performance demands.

Perceptual pattern errors. Through over-exposure to traditional arithmetic problems, children learn to expect math problems to have **all operations on the left side of the equal sign**, with the equal sign immediately before the answer blank on the right, an "operations = answer" problem format (McNeil & Alibali, 2004, Carpenter et al., 2003). Students who expect all problems to have operations on the left fail to correctly encode the problem before them. For instance, after briefly viewing the problem "7 + 4 + 5 = 7 + __" children who rely on their representations of the "operations = answer" problem format erroneously remember the problem as "7 + 4 + 5 + 7 = __" (McNeil & Alibali, 2004).

Conceptual pattern errors. Children learn to interpret the equal sign operationally as a symbol to do something (Baroody & Ginsburg, 1983; Behr et al., 1980). When asked

to define the equal sign—even in the context of a mathematical equivalence problem—many children treat it like an arithmetic operator (like + or -) **that means they should calculate the total of everything on the left side of the equal sign** (McNeil & Alibali, 2005).

Procedural pattern errors. Through early practice with traditional problems (e.g., all operations on the left of '='), children learn to **perform all of the listed operations on all given numbers in a math problem** (e.g., add up all the numbers in an addition problem, McNeil & Alibali, 2004, 2005). This incorrect representation of equations misleads students to solve the problem "7 + 4 + 5 = 7 + __" by performing all given operations on all given numbers and put 23 (instead of 9) in the blank (McNeil, 2007; Rittle-Johnson, 2006, Falkner et al., 1999).

A history of findings supports the hypothesis that children's difficulties with mathematical equivalence are partially due to inappropriate knowledge of the perceptual structure, conceptual meaning, and procedural routine associated with encoding and solving equations. The change-resistance account further suggests that these faulty representations are derived from overly narrow experience with traditional arithmetic. Recent studies have documented the effects of incorrect representations of equivalence in fourth-graders (McNeil & Alibali, 2004) and have induced similar error patterns in adults (Crooks and Alibali, 2013). We build on the work of McNeil, Fyfe, and Dunwiddie (2015), who examined the impact of an early intervention on second-graders multi-faceted understanding of equivalence, replicate and extending these findings to more closely examine the nature of early equivalence encoding and its relationship to equation solving in a large representative sample of students.

In the current study, we sought to more deeply examine the nature of second-grade students' encoding responses, looking for evidence of the misconception-based (i.e., perceptual, conceptual, and procedural) error patterns that have been theorized in past work from McNeil, Alibali, and others (McNeil et al., 2019, McNeil & Alibali, 2005), and induced in adults by Crooks and Alibali (2013).

We further explore the relationship between encoding and solving of equivalence problems, asking whether the specific misconceptions identified through encoding errors are predictive of equation solving performance. We then examine the impacts of research-based equivalence training activities on encoding and solving accuracy. Specifically, we randomly assigned classrooms to training using an intensive treatment intervention or an active control condition consisting solely of non-traditional mathematical practice and measured the training impact on students' ability to encode equations and solve equivalence problems post-training. We organize our findings to explore four related questions:

Do second-grade students make encoding errors consistent with overgeneralizing patterns from early arithmetic?

Do encoding errors systematically vary across items with different structure and length? How does the frequency of different types of encoding errors change with targeted training?

How do misconception-based errors in students' equation encoding predict equation solving?

Does targeted, conceptually rich equivalence training impact encoding and equation solving?

Measuring Equation Encoding and Solving. We assessed second-grade students' ability to correctly encode and solve non-traditional equivalence problems before and after the intervention training using the same measures of equation encoding, equation solving sign used in previous work by McNeil and colleagues (Johannes et al., 2017; McNeil et al., 2012; McNeil & Alibali, 2005b).

Equation encoding. The equation encoding measure consisted of recalling four math expressions (e.g., 2 + 6 =2 + _) presented one at a time. Each expression was visible for five seconds and students were instructed to remember and write down exactly what they saw. Responses were coded as correct if the student wrote the equation exactly as shown (i.e., the correct numbers and symbols in the correct order). We discuss the coding of relevant erroneous response types in the results.

Equation solving. The equation solving measure consisted of eight equations with operations on both sides of the equal sign (e.g., 3 + 5 + 6 = 3 + __). For a response to be coded as correct, a student needed to write the value that would make the equivalence relation hold.

Our sample of encoding and solving items is listed in Table 1. All items included one addend and a blank on the *right* side of the equal sign. The items varied on two dimensions: the number of addends (two or three) on the *left* side of the equal sign, and the position of the blank (at the end of the equation or directly after the equal sign).

Table 1. Equation encoding and solving items administered pre- and post-intervention

Addends	Position of blank	Encoding items	Solving items
Two	End of equation	4+5=3+_	3+7=3+_ 2+7=6+_
	After '='	7+1=_+6	5+3=_+3 8+2=_+6
Three	End of equation	2+3+6=2+_	3+5+6=3+_ 6+2+8=4+_
	After '='	3+5+4=_+4	7+2+4=_+4 7+4+6=_+3

ICUE: Improving Children's Understanding of Equivalence Intervention

As current math practice seems to promote the development of faulty representations, the change resistance account of "operational patterns" offers design principles for instruction to improve students' understanding of equivalence. Initially, researchers hypothesized that greater exposure to "non-traditional arithmetic" problems (e.g., presenting operations on the right side of the equation, "_ = 2 + 4" and using relational phrases such as "is equal to" instead of the equal sign in practice problems) may prevent students from developing operational patterns (McNeil et al., 2011). Though practice with non-traditional arithmetic led to improved outcomes over traditional instruction, a number of students failed to reach proficiency (McNeil, Fyfe, & Dunwiddle, 2015).

To further promote mastery of equivalence, McNeil and colleagues added additional design features beyond non-traditional arithmetic practice. The current version of the materials, dubbed Improving Children's Understanding of Equivalence (ICUE), consists of second grade student activities that reduce reliance on operational patterns and promote deep understanding of mathematical equivalence through four key components, outlined below, that have independently been shown to be effective. Multiple pilot studies have since found that the ICUE treatment intervention is successful in improving student understanding of mathematical equivalence (Byrd et al., 2015; Johannes et al. 2017).

1. Nontraditional arithmetic practice (McNeil, Fyfe, & Dunwiddle, 2015, Chesney et al., 2012),
2. Lessons that first introduce the equal sign outside of arithmetic contexts (e.g., "28 = 28") before introducing arithmetic expressions (e.g., Baroody & Ginsburg, 1983).
3. Concreteness fading exercises in which concrete, real-world, relational contexts (e.g., sharing stickers, balancing a scale) are gradually faded into the corresponding abstract mathematical symbols (e.g., Fyfe, McNeil, Son, & Goldstone, 2014), and
4. Activities that require students to compare and explain different problem formats and problem-solving strategies (e.g., Carpenter, Franke, & Levi, L. 2003).

Methods

Design

We used a cluster-randomized control trial design to examine the impacts of the ICUE intervention training relative to an active control program. Teachers were randomly assigned to use the either the ICUE Treatment intervention or Active Control materials. The Active Control consisted of workbook activities to control for time on task and contained non-traditional arithmetic practice but not the additional components present in the Treatment ICUE condition, described above.

Participants. 44 second-grade teachers (24 treatment, 20 control) used the activities in their classrooms in California. Class sizes ranged from 18 to 25, and we analyzed data from 482 students who completed the Treatment activities and 406 students who completed the Control activities and measures.

Procedure and Materials

The procedure for ICUE Treatment and Active Control conditions were identical, differing only in the content of the materials used by teachers and students. Each teacher received training on the study purpose, features of the activities, and strategies for integrating the activities into their typical mathematics curriculum.

Prior to starting the study, participating teachers completed online surveys assessing their mathematics teaching experience and classroom structure and dynamics.

After administering a pre-test, teachers used the study materials for approximately 15 minutes twice each week for 16 weeks. In both conditions, teachers were asked to use the study materials to supplement, rather than replace current math instruction, and to limit the duration of the activities to 20 minutes per session.

After completing the 32 sessions, teachers administered the same pre-intervention measure of mathematical equivalence understanding, which included the equation encoding and solving items reported here, along with an item prompting children to name and define the "=" symbol, not reported. Teachers administered additional post-intervention measures of transfer and computation fluency, we do not report these here.

Active Control. Teachers in the Active Control condition received a set of student workbooks and a teacher guide.

ICUE Treatment. Teachers in the ICUE Treatment condition received a set of student workbooks, a teacher guide, a set of classroom manipulatives including balance scales and flashcards.

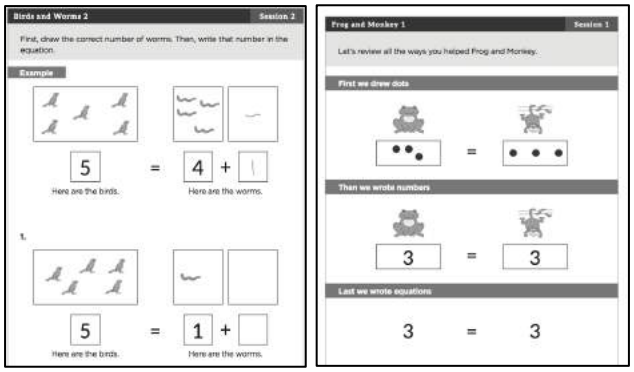


Figure 1. Sample workbook page from the Active Control (left) and ICUE Treatment (right) condition materials.

Results

Do second graders make misconception-based encoding errors?

Crooks and Alibali (2013) identified three categories of errors in encoding and solving, reviewed above, which stem from different types of knowledge and misconceptions. We asked whether, after only one year of formal mathematics instruction, second graders produced erroneous encoding responses that align with these any of these three related facets of misunderstanding, grouping them together as "misconception errors". We examined the frequency with which theoretically-relevant types of encoding errors, which can be induced in adults (Crooks & Alibali, 2013) naturally occur in young students. We differentiated these misconception-based errors from other types of errors, including performance errors, which we hypothesize stem from memory-based constraints in this population.

We assessed students' accuracy in encoding four different equivalence problems (see Table 1 for items) and examined the frequency with which they made different types of errors. Student in both the Control and Treatment groups produced a range of responses for each encoding item and made multiple types of errors, including misconception-based errors, with different frequency. Examples and overall frequency of response types are listed in Table 2.

Students produced the misconception-based errors of interest in approximately 20% of their responses overall. The majority of misconception-based errors produced in both conditions aligned with the perceptual error type identified by Crooks and Alibali (2013); conceptual and procedural errors types were produced relatively infrequently.

Table 2. Response types and examples for encoding item 2+3+6=2+_, with overall frequency of response pre- and post-training.

Response type	*Examples*	*Control Pre/Post*	*Treatment Pre / Post*
Correct	2+3+6=2+_	0.25/0.47	0.35/0.56
Misconception errors	**2+3+6+2=_ 2+3+6=11+2 2+3+6=2+13**	**0.23/0.21**	**0.24/0.17**
Memory error	2+3+6	0.06 / 0.08	0.06/ 0.15
Other errors	2+3+7=6	0.39 / 0.21	0.28/ 0.09
No response	no response	0.07 / 0.02	0.05/ 0.03

How does the equation structure influence encoding errors?

We chose to focus on misconception-based and memory-based encoding error patterns and explored variation in error rates across the four encoding items that varied in A. the number of addends, and B. the position of the blank in the equation (see Table 1 for items). The larger number of addends was predicted to increase the working memory demands of the problem. The position of the blank at the end of the equation (e.g., 4+5=3+_) was predicted to increase the likelihood of perceptual pattern errors as these items are most perceptually similar to traditional arithmetic problems (e.g., 4+5+3=__), and may trigger operational, instead of relational, interpretations of the equal sign (e.g., as a symbol that means give the answer') that give rise to erroneous arithmetic procedures (e.g., add up all numbers and write the sum in the blank; see Crooks & Alibali, 2013; McNeil et al., 2011).

Students' pre-intervention encoding error frequency is displayed by item in Figure 2. The frequency of misconception- and memory-based errors varied based on both the position of blank (at the end of the equation – first and third items - or directly after the '=' sign – second and fourth items in Table 1) and the number of addends on the left side of the equation (two – first two items - or three – last two items in Table 1).

In line with our predictions, regression models confirmed that students in both conditions produced a reliably greater number of perception-based errors for items with the blank at the end of the equation (β=0.852, SE =0.12, p<.01), and this interacted with the number addends, such that students produced the greatest number of misconception errors for the three-addend item with the blank at the end: "2+3+6=2+_" (β=0.534, SE =0.09, p<.05).

Finally, students made a reliable number of memory-based errors, but only for items with three addends (β=0.472, SE =0.11, p<.05).

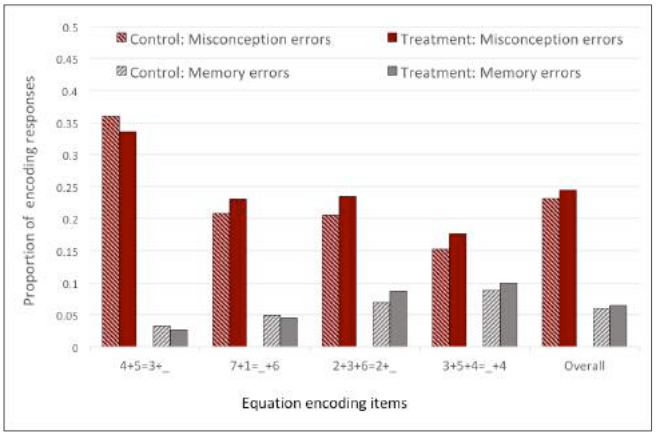


Figure 2. Pre-intervention patterns of misconception- and memory-based error responses for Treatment and Control students. Misconception errors were greatest for items with a blank space at the end of the equation (first and third items); memory errors were greatest for items with three addends (second and fourth items).

How does conceptually rich equivalence training change students' encoding responses?

We next examined the impact of the ICUE Treatment and Active Control training on misconception- and memory-based encoding errors. We asked whether exposure to non-traditional arithmetic, through the Active Control condition, was sufficient to maximally reduce these encoding errors, or whether more conceptually rich training, found in the ICUE Treatment intervention, would lead to greater error reduction. The change in students' encoding errors from pre- to post-intervention is displayed in Figure 3. Students in the Treatment condition showed a greater reduction, post-intervention, in misconception-based errors compared to students in the Control condition (β=0.921, SE =0.10, p<.01), and this reduction was greatest for items with a blank space at the end of the expression (β=0.633, SE =0.09, p<.05). Thus, for encoding, we find that conceptually rich training leads to greater reduction misconception-based errors. Training type did not significantly impact the frequency of memory errors for three-addend encoding items.

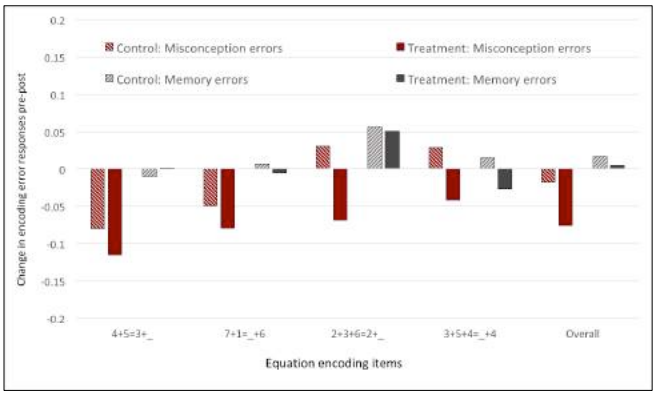


Figure 3. Pre- to post-intervention changes in error responses for Treatment and Control students. Students in the Treatment condition showed a greater reduction in misconception-based errors (solid red bars), which was greatest for items with a blank space at the end of the equation.

How do misconception-based errors in students' equation encoding impact equation solving?

Turning to equation solving, we tested whether perceptual errors in equation encoding reliably predicted students' equation solving performance before students had received any training through the Treatment or Control interventions. For each item type, students completed one encoding item and two solving items (see Table 1). Thus, for each type of item, a student could solve both solving items correct, one correct, or zero correct. We used ordinal regression models to capture this ordering and tested encoding performance (i.e., whether a student encoded that type of item correctly), error types (misconception- and memory-based), and item properties (number of addends and location of blank space) as predictors.

Pre-intervention solving performance was predicted by multiple aspects of encoding responses: students were more likely to solve an equivalence problem correctly if they had accurately encoded the same type of item correctly (β=0.778, SE=0.121, p<.05), and students were less likely to solve a problem correctly if they had produced a misconception-based error for that type of item on the encoding measure (β= -0.420, SE =0.097, p<.01).

Performance was also predicted by properties of the items: items with two addends were more likely to be solved correctly than those with three addends (β=0.360, SE =0.079, p<.01) and items with a blank space directly after "=" were more likely to be solved correctly than those with a blank at the end of the equation (β=0.226, SE =0.079, p<.05). However, pre-intervention solving performance was not reliably predicted by memory-based errors, or assigned condition (β=-0.061, SE =0.078, *ns*).

How does equivalence training impact equation solving?

We used a similar ordinal regression model to test the effect of training condition on students' post-intervention solving performance. Performance was best predicted by a combination of intervention condition and encoding responses; item properties (position of blank, number of addends) were not reliable predictors in the best-fitting model of solving performance. The strongest single predictor of post-intervention solving performance was training condition: students in the Treatment condition were more likely to correctly solve items post-intervention, compared to students in the Active Control condition (Figure 4; (β= 1.267, SE =0.074, p<.01)). As in the case of encoding, we found that conceptually richer training led to more accurate solving performance.

Students were also more likely to solve one or both equation solving problem correctly on the post-intervention measure if they had solved one correctly on the pre-intervention measure (β=0.636, SE =0.12, p<.01) and were increasingly likely to solve both items correctly if they had solved both items correctly pre-intervention (β=1.053, SE =0.15, p<.01).

Finally, as in the pre-intervention model, post-intervention solving performance was predicted by encoding responses: students were more likely to solve an equivalence problem correctly post-intervention if they had accurately encoded the same type of item correctly post-intervention (β=1.493, SE =0.08, p<.01), and students were less likely to solve a problem correctly if they had produced a misconception-based error for that type of item on the encoding measure (β= -0.749, SE =0.06, p<.05).

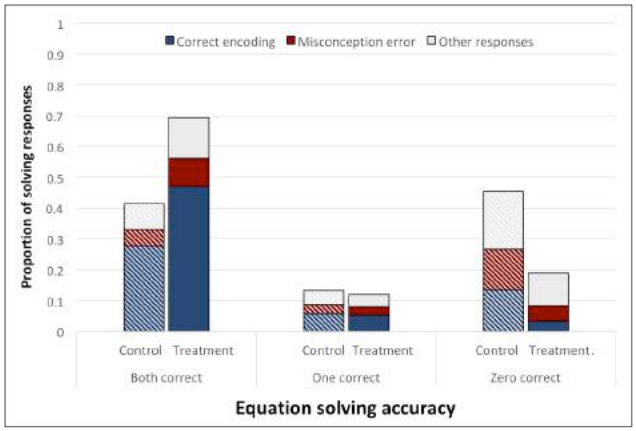


Figure 4. Treatment and Control students' post-intervention solving performance broken down by frequency of correct and misconception-based encoding responses. Conceptually rich training in the Treatment condition led to the greatest improvement in solving performance; students from both conditions were more likely to solve an item correctly if they had encoded the same type of item correctly, and less likely if they had made a misconception-based error in encoding.

Considering both our encoding and solving results, we found that, while manipulated item properties impacted students' ability to correctly encode and solve non-traditional equations pre-intervention, the magnitude of this impact was reduced, for encoding, and eliminated, for solving, by targeted conceptually rich training in mathematical equivalence. Specifically, students in the Treatment condition showed a greater reduction, post-intervention in misconception-based encoding errors compared to Control students, and this reduction was greatest for items with a blank space at the end of the equation (i.e., items that are perceptually most similar to traditional arithmetic problems). Treatment students were also more accurate on a post-intervention equation solving task (with no reliable condition differences pre-intervention) and, while encoding responses were predictive of solving performance, manipulated item properties (number of addends and position of blank space) were not.

Conclusions

Understanding equivalence is key for later mathematical understandings. The change-resistance account suggests that students fail to develop appropriate representations of equations and equivalence because instruction with traditional arithmetic problems encourages students to develop ineffective representations of problems.

In the current study, we explored the relationship between problematic representations and students' ability to accurately encode and later solve non-traditional equivalence problems. We examined encoding and solving abilities in second-grade students and found that a single year of formal instruction (i.e., first grade) with traditional arithmetic practice was sufficient to reliably lead to misconception-based errors at encoding, which predominantly consisted of perceptual pattern errors, in the framework of Crooks and Alibali (2013). Baseline performance on both tasks worsened when target problems perceptually resembled traditional arithmetic problems (i.e., when a blank was at the end of an equation), and when working memory load (number of addends) was increased. Misconception errors at encoding were predictive of solving performance, both at baseline (pre-intervention) and at post-intervention, suggesting that students who make these misconception-based errors at encoding may be activating similar faulty representations during the solving task. Finally, training improved both encoding and solving performance, demonstrating that erroneous response patterns can be overcome with intervention. However, students in the Treatment condition showed greatest improvements on both tasks, suggesting that deeper conceptual learning is required to resolve what, at first glance, might be thought of as a perceptual bias towards traditional arithmetic problem structure.

Interestingly, while manipulated properties of the encoding and solving items (see Table 1) predicted students' errors in the encoding task, these properties were only predictive of *pre-intervention* solving performance. Students' post-intervention solving performance was predicted by their post-intervention encoding responses, but not directly predicted by item properties, suggesting that any relationship between these manipulated item properties, such as the number of addends and position of blank space, and solving performance is potentially mediated by encoding. This is consistent with a mediation analysis performed by Crooks and Alibali (2013), in which the authors demonstrated that the impact of priming incorrect representations on adults' equation solving performance was mediated by problem reconstruction (or encoding). Future work will explore this relationship in more depth by using a greater number of items and possible combination of manipulated item properties.

Even after training, students in both conditions did not reach ceiling performance in either encoding or solving non-traditional equations. On the one hand, the equation encoding and solving assessment items were chosen to leave room for improvement and to avoid ceiling effects. However, in future work, we plan to explore how individual students resolve or persist in error patterns with training. We further plan to test whether different encoding errors give rise to specific solving responses, or whether any error simply creates noise in students' equation solving processes. Our preliminary findings suggest that misconception-based errors in encoding lead to greater error rates in solving, but a larger sample of items and responses may be required to support fine-grained conclusions about the nature of this relationship and, specifically, how perceptual, conceptual, and procedural misconceptions individually and collaboratively impact equation encoding and solving.

Acknowledgments

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A150088 to WestEd. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

References

Alibali, M. W., Phillips, K. M., & Fischer, A. D. (2009). Learning new problem-solving strategies leads to changes in problem representation. *Cognitive Development*, *24*, 89-101.

Baroody, A. J., & Ginsburg, H. P. (1983). The effects of instruction on children's understanding of the "equals" sign. *Elementary School Journal*, *84*, 199-212.

Behr, M., Erlwanger, S., & Nichols, E. (1980). How children view the equal sign. *Mathematics Teaching*, *92*, 13-15.

Bruner, J. S. (1957). On perceptual readiness. *Psychological Review*, *2*, 123-152.

Byrd, C. E., McNeil, N. M., et al. (2015). Pilot Test of a Comprehensive Intervention to Improve Children's Understanding of Math Equivalence. AERA, Chicago, IL.

Carpenter, T. P., Franke, M. L., & Levi, L. (2003). *Thinking mathematically: Integrating arithmetic and algebra in elementary school.* Portsmouth, NH: Heinemann.

Charles, R. I. (2005). Big ideas and understandings as the foundation for elementary and middle school mathematics. *NCSM: J. of Mathematics Education Leadership*, 8(2), 9-24.

Chesney, D. L., McNeil, N. M., Petersen, L. A., & Dunwiddie, A. E. (2012). *Arithmetic practice that includes relational words promotes conceptual understanding and computational fluency*. Poster presented at the APS Annual Convention, Chicago, IL.

Crooks, N.M., and Alibali, M. W. (2013). Noticing relevant problem features: Activating prior knowledge affects problem solving by guiding encoding. *Frontiers Psychoogy. 4, 884*, 1-10.

Falkner, K. P., Levi, L., & Carpenter, T. P. (1999). Children's understanding of equality: A foundation for algebra. *Teaching Children Mathematics*, *6*, 232-236.

Fyfe, E. R., McNeil, N. M., Son, J. Y., & Goldstone, R. L. (2014). Concreteness fading in mathematics and science instruction: A systematic review. *Educational Psychology Review*, *26*, 9-25.

Johannes, K., Davenport, J., Kao, Y, Hornburg, C.B., & McNeil, N.M. (2017). Promoting children's relational understanding of equivalence. *Proc. of the 39th Annual Meeting of the Cognitive Science Society,* 600-605.

Jones, I., Inglis, M., Gilmore, C., & Dowens, M. (2012). Substitution and sameness: Two components of a relational conception of the equals sign. *J. of Experimental Child Psychology*, *113*, 166-176.

Knuth, E. J., Stephens, A. C., McNeil, N. M., & Alibali, M. W. (2006). Does understanding the equal sign matter? Evidence from solving equations. *J. for Research in Mathematics Education, 37*(4), 297-312.

McNeil, N. M. (2007). U-shaped development in math: 7 year-olds outperform 9-year-olds on equivalence problems. *Developmental Psychology*, *43*, 687-695.

McNeil, N. M. (2008). Limitations to teaching children 2 + 2 = 4: Typical arithmetic problems can hinder learning of mathematical equivalence. *Child Development*, *79*, 1524-1537.

McNeil, N. M. (2014). A "change-resistance" account of children's difficulties understanding mathematical equivalence. *Child Development Perspectives*, *8*, 42-47.

McNeil, N. M., & Alibali, M. W. (2004). You'll see what you mean: Students encode equations based on their knowledge of arithmetic. *Cognitive Science*, *28*, 451-466.

McNeil, N. M., & Alibali, M. W. (2005). Why won't you change your mind? Knowledge of operational patterns hinders learning and performance on equations. *Child Development*, *76*, 883-899.

McNeil, N. M., Chesney, D. L., et al. (2012). Organizing addition knowledge around equivalent values facilitates understanding of math. equivalence. *J. of Educational Psychology*. *104*, 1109-1121.

McNeil, N. M., Fyfe, E. R., Petersen, L. A., Dunwiddie, A. E., & Brletic-Shipley, H. (2011). Benefits of practicing 4 = 2 + 2: Nontraditional problem formats facilitate children's understanding of mathematical equivalence. *Child Development*, *82*, 1620-1633.

McNeil, N. M., Fyfe, E. R., & Dunwiddie, A. E. (2015). Arithmetic can be modified to promote understanding of mathematical equivalence. *J. of Educational Psychology, 107,* 423-436.

McNeil, N. M., Hornburg, C. B., Brletic-Shipley, H., & Matthews, J. M. (2019). Improving children's understanding of math equivalence via an intervention that goes beyond non-traditional arithmetic practice. *J. of Educational Psychology, 111*.

Rittle-Johnson, B. (2006). Promoting transfer: The effects of self-explanation and direct instruction. *Child Development, 77*, 1-15.

Seo, K.-H., & Ginsburg, H. P. (2003). "You've got to carefully read the math sentence...": Classroom context and children's interpretations of the equals sign. In A. J. Baroody & A. Dowker (Eds.), *The development of arithmetic concepts and skills: Constructing adaptive expertise* (pp. 161-187). Mahwah, NJ: Erlbaum.

Moral Reputation and the Psychology of Giving: Praise Judgments Track Personal Sacrifice Rather Than Social Good

Samuel G. B. Johnson (sgbjohnson@gmail.com)
School of Management, University of Bath, Bath, BA2 7AY UK

Abstract

Do we praise altruistic acts because they produce social benefits or because they require a personal sacrifice? On the one hand, utilitarianism demands that we maximize the social benefit of our actions, which could motivate altruistic acts. On the other hand, altruistic acts signal reputation precisely because personal sacrifice is a strong, costly signal. Consistent with the reputational account, these studies find that in the absence of reputational cues, people mainly rely on personal cost rather than social benefit when evaluating prosocial actors (Study 1). However, when reputation is known, personal cost acts as a much weaker signal and play a smaller role in moral evaluations (Study 2). We argue that these results have far-reaching implications for the psychology and philosophy of altruism, as well as practical import for charitable giving, particularly the effective altruism movement.

Keywords: Moral psychology; reputation; decision-making; prosocial behavior; altruism

Introduction

Moral philosophers, as well as our inner ethicists, often recommend altruism as an essential component of moral behavior. Altruistic acts have a dual character—an altruistic act requires a *personal cost* and produces a *social benefit*. The most plausible arguments for the morality of altruism seem to place the emphasis on the social benefit. Consequentialism tells us that we should act to produce the greatest good for the greatest number (e.g., Bentham, 1907/1789; Mill, 1998/1861), which often entails altruistic acts. For example, if you live in a rich country, you probably gain far less from $20 than would a family in a poor country, suggesting that the moral act is to donate the $20 (Singer, 2015).

Because human survival depends on coordinated social activity, we have moral intuitions which sometimes appear to track the conclusions of moral and legal philosophy (e.g., Mikhail, 2007); moreover, people sometimes behave like intuitive consequentialists, particularly when they have time to reflect (Greene, Morelli, Lowenberg, Nystrom, & Cohen, 2008). So perhaps our intuitive praise for altruistic agents stems from the same underlying consequentialist psychology that animates moral philosophers.

But do we really value altruistic acts in proportion to the social benefit they produce? Or do we rely primarily on the personal cost entailed? In many situations this a moot question, since personal sacrifice and social good are often highly correlated. This is true in both our ancestral environment, where presumably our moral intuitions evolved, and our modern environments where our moral intuitions guide behavior. If you spend two hours gathering berries, you gather more than if you spend one hour; if you give $100 rather than $50 to Oxfam, the charity can accomplish double. But these dimensions are not always so tightly correlated. The *effective altruism* movement focuses on maximizing the social good accomplished per dollar donated, since effectiveness varies hugely across different causes (MacAskill, 2015; Singer, 2015). For example, one can prevent blindness in dozens of children in the developing world for the cost of training one service dog in the developed world. To effective altruists, a donation's *quality* is at least as important as its *quantity*.

Do ordinary people, like effective altruists, prioritize social good over personal sacrifice in evaluating prosocial acts? We propose that, conversely, personal sacrifice usually looms larger. Many argue that our intuitive morality evolved to induce cooperative behavior (Goodwin, Piazza, & Rozin, 2014; Haidt, 2007; Nowak & Sigmund, 2005; Sperber & Baumard, 2012; Uhlmann, Pizarro, & Diermeier, 2015). Praise by one's social group rewards prosocial behaviors while blame penalizes anti-social behaviors, and these judgments reflect changes in moral reputation. For example, people blame others for harmless actions accompanied by "wicked desires" (Inbar, Pizarro, & Cushman, 2012) because such desires can signal poor moral character. Character judgments depend mainly on intentions, not outcome (Cushman, 2008), serving to track reliable individual differences in social behavior. Beliefs about moral reputation even have an identifiable neural basis (Delgado, Frank, & Phelps, 2005), speaking to their psychological fundamentality.

On this view, praise judgments flow from evidence of good moral character. Personal sacrifice is a stronger signal of character than social good for two reasons. First, personal sacrifice is under an actor's direct control, whereas social good depends partly on uncontrollable factors. One can write a check to Oxfam for any amount, but what the charity accomplishes depends on their decisions and on luck. Inferences based on personal sacrifice avoid such sources of noise. Second, personal sacrifice is directly observable, whereas social good is often unobservable. We see the number on the Oxfam check, but usually not how many people were helped.

Given that our interest here is in people's moral evaluations of prosocial behaviors, the most directly relevant literature would seem to be moral judgment. However, this research has focused primarily on factors influencing blame for negative acts rather than praise for positive acts (e.g., Cushman, 2008; Inbar et al., 2012;

Mikhail, 2007). Research on charitable giving does provide some hints, however. People view prosocial acts unfavorably when those acts also benefit the actor (Ariely, Bracha, & Meier, 2009; Barasch, Levine, Berman, & Small, 2014; Newman & Cain, 2014), and these perceptions have negative downstream consequences for actual prosocial behavior (Ariely, Bracha, & Meier, 2009). This suggests that personal sacrifice is a necessary condition for positive evaluations of prosocial behavior. However, presumably sacrifice alone is not sufficient—it seems doubtful that purely self-sacrificial acts would be seen as praiseworthy in the absence of some broader social benefit. The film *The Seventh Continent* depicts a middle-class family in modern Europe that destroys itself for no apparent reason, flushing their money down the toilet and committing suicide. To this audience these acts are puzzling and horrifying, not praiseworthy. Thus, the prior literature together with common intuition suggests that some degree of personal cost and some degree of social benefit are required for a prosocial act to be praised; indeed, this may be part of the very concept of altruism.

Both cost and benefit appear to track judgments of praise when we are comparing some versus none. But would they track praise when comparing a larger amount to a smaller amount? In prior work, highly prosocial acts are not seen as more praiseworthy than slightly prosocial acts, although, interestingly, people were sensitive to the degree of harm in assigning blame (Klein & Epley, 2014). However, the effects of prosocial benefits versus costs have not been teased apart in observer's moral evaluations—people may well be insensitive to the degree of cost as well as the degree of benefit in evaluating prosocial acts. On the actor side, people are more moved to donate by the plight of one than of many (Small, Loewenstein, & Slovic, 2007) and are largely indifferent to the number of individuals helped (Kahneman & Knetsch, 1992). Conversely, people are likelier to donate money when paired with a painful sacrifice (explaining, arguably, the prevalence of charity runs; Olivola & Shafir, 2013). These results again are suggestive of possible insensitivity to the degree of benefit, but do little to clarify how prosocial actors respond to the degree of cost. Moreover, it is unclear whether these results would generalize to moral evaluations rather than prosocial behaviors themselves or when cost versus benefit are pitted against one another directly.

Overall, prior work does not tell us whether moral judgments of altruist acts track personal sacrifice or social benefit. We know that some amount of personal sacrifice and social benefit are necessary conditions, but not whether one of these factors has an outsized influence compared to the other, when pitted against one another directly. This issue is critical to understanding the psychological basis of moral praise (utilitarian admiration vs. character signaling) and likely has implications for the design of charitable appeals.

Thus, the current studies investigate this issue by testing judgments of praise in response to charitable donations. The studies orthogonally manipulate the amount of personal sacrifice (size of donation) and social good (number of individuals helped), measuring judgments of praise and character. In Study 2, independent reputational cues are available, whereas in Study 1 they are not. When strong reputational cues attest to a donor's robust character, personal sacrifice is uninformative about moral character and therefore should not influence praise; however, sacrifice should have a large effect when other reputational cues are absent.

Study 1

Study 1 tested whether, absent further information about a person, judgments of prosocial behavior depend mainly on the degree of *personal sacrifice*, but not *social good*.

Method

A total of 598 American participants (56% female, M_{age} = 37.4) were recruited for Studies 1A and 1B through Mechanical Turk. Participants were excluded if they failed an attention check (see below; N = 65).

Participants read about a charitable donation benefiting people in a developing country. The charities focused on blindness, hunger, education, or disaster relief. The donations involved a *low*, *moderate*, or *high* monetary contribution (to manipulate personal sacrifice), and were *low* or *high* in effectiveness (to manipulate social good), with both manipulations between-subjects. These conditions always differed from one another by one order of magnitude (a factor of 10). For two of the vignettes, the beneficiary was an individual in the low-effectiveness condition and a small group in the high-effectiveness condition. For example:

> Julia decided to make a donation to charity. She donated [$20/$200/$2000] to a charity focused on international health. Her donation was used to cure [a child's/10 children's] blindness in Ethiopia.

For the other two vignettes, the beneficiary was a small group in the low-effectiveness condition and a large group in the high-effectiveness condition. For example:

> Rob decided to make a donation to charity. He donated [$12.50/$125/$1250] to a charity focused on disaster relief. His donation was used to provide basic shelter to [10/100] people for one month after a hurricane in Guatemala.

On the same screen, participants rated the praiseworthiness of the action ("Please rate the moral praiseworthiness of Julia's action") on a scale from 0 ("Not very praiseworthy") to 10 ("Extremely praiseworthy"), and the actor's character ("Please rate Julia's moral character") on a scale from 0 ("Ordinary moral character") to 10 ("Saint-like moral character").

After the main task, participants checked whether each of the four donation targets was mentioned in the study; participants making any incorrect answers were excluded.

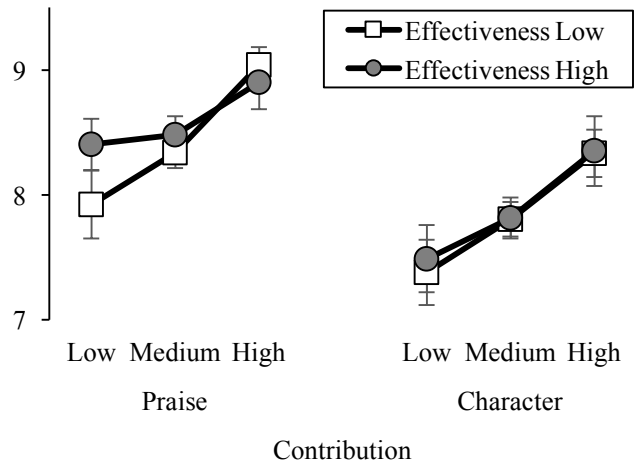


Figure 1: Results of Study 1 (character information absent). Bars represent 1 SE.

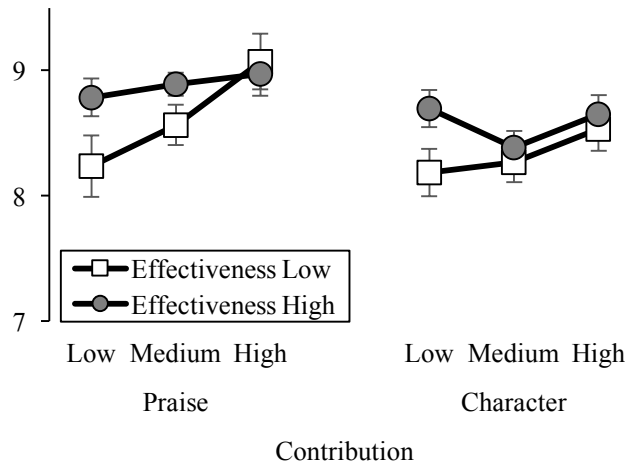


Figure 2: Results of Study 2 (character information present). Bars represent 1 SE.

Studies 1A and 1B were identical, except Study 1A included the low- and medium-contribution conditions, while Study 1B included the medium- and high-contribution conditions. Thus, both studies used 2 (personal sacrifice) x 2 (social good) designs. Given that the designs differed only in contribution levels, we combine them for analysis, to maximize statistical power and facilitate comparisons across studies.

Results

Overall, participants used the degree of personal sacrifice, but not of social good, to inform judgments of moral praise and character. The means are plotted in Figure 1.

Since contribution condition is equal-interval in log scale, it was coded as a continuous variable, (–1 = low, 0 = medium, 1 = high); effectiveness condition was contrast-coded (–1 = low, 1 = high). A linear regression was conducted, predicting moral judgments from contribution, effectiveness, and their interaction. There was a significant main effect of contribution, $b = 0.40$, $SE = 0.11$, 95% CI[0.18,0.61], $p < .001$, indicating that greater degrees of sacrifice were viewed as more morally praiseworthy. However, there was no effect of effectiveness, $b = 0.08$, $SE = 0.07$, 95% CI[–0.06,0.22], $p = .26$, nor a significant interaction, $b = –0.16$, $SE = 0.11$, 95% CI[–0.37,0.06], $p = .15$. Thus, people did not take account of social benefit in evaluating the moral praiseworthiness of the donations. Moreover, this effect did not depend on whether the less-effective donations benefited an individual or a small group: Adding this variable and its interactions to the regression model did not improve fit, $F(529,4) = 1.40$, $p = .23$. (Adding a factor for vignette also did not improve fit, indicating that there are no reliable differences in the effects across vignettes.)

The results were similar for character judgments. A regression analysis parallel to the above revealed a significant effect of contribution, $b = 0.45$, $SE = 0.12$, 95% CI[0.21,0.69], $p < .001$, but not of effectiveness, $b = 0.02$, $SE = 0.08$, 95% CI[–0.14,0.18], $p = .82$, or the interaction between these variables, $b = –0.02$, $SE = 0.12$, 95% CI[–0.27,0.22], $p = .84$. Once again, adding the individual vs. small-group dummy-code and its interactions to the model did not improve fit, $F(4,529) = 0.62$, $p = .65$, indicating that the effect does not depend on whether the less-effective donation benefitted an individual or small group.

Discussion

Praise judgments track the amount of money sacrificed by donors, but not the social good produced by those donations. This is consistent with a signaling theory of moral praise, which assumes (i) that moral praise derives from evidence of character and (ii) that personal sacrifice is a stronger (costly, controllable, and observable) signal of character.

This mechanism will be tested more directly in Study 2. Before doing so, however, let us consider a possible boundary condition: Whether the individuals helped are closer or farther within one's moral circle (Singer, 1981). People are parochial about their charitable giving (Baron & Szymanska, 2011; see also Nagel & Waldmann, 2013), favoring causes that benefit their in-group. Perhaps altruistic acts done to benefit others in distant countries are viewed as altruistic mainly due to the signaling value (i.e., their cost), but those done to benefit one's own society are seen in a more utilitarian way. It is plausible that one would praise an altruistic act to the extent that it helped one personally, so if one identifies with one's in-group, the effectiveness of in-group help may impact praise judgments.

To test this, a replication of Study 1 was conducted (Johnson, 2018), identical except for replacing the beneficiaries living in the developed world with beneficiaries living in America (e.g., a hurricane in South Carolina rather than Guatemala). This study found a very

similar pattern of results to Study 1: Contribution was a large and robust predictor of praise, $b = 0.39$, $p < .001$, while effectiveness had a small and marginal effect, $b = 0.11$, $p = .09$. Combining the data from this follow-up with the Study 1 data, there was no significant interaction between beneficiary (in-group vs. out-group) and either contribution or effectiveness on praise. This both replicates the results of Study 1 and suggests that parochialism is not a boundary condition on the findings.

Study 2

Personal sacrifice is typically under the actor's personal control and is typically visible; therefore it can be an informative, costly signal of moral reputation. In contrast, social good is less controllable and less visible. If this drives attention to costs rather than benefits, then independent evidence of an actor's pristine moral character should decrease the relevance of individual prosocial acts for evaluating character and attenuate the effect of personal sacrifice.

Method

A total of 600 American participants (57% female, $M_{age} = 36.6$) were recruited for Study 2. Participants were excluded if they failed the same attention check used in Study 1 ($N = 91$).

Studies 2A and 2B were identical to Studies 1A and 1B, respectively, except that the vignettes were altered to include information establishing the actor's altruistic moral character. For example:

> Rob works as a receptionist, earning about $31,000 per year. He donates about 30% of his salary each year to a variety of charitable causes.
>
> One of the donations Rob decided to make this year was [$12.50/$125/$1250] to a charity focused on disaster relief. His donation was used to provide basic shelter to [10/100] people for one month after a hurricane in Guatemala.

The moral judgment question was rephrased so it was clear that it referred to this *specific* donation, rather than the pattern of charitable donations (e.g., "Please rate the moral praiseworthiness of Rob's [$12.50/$125/$1250] donation"). Rephrasing this question to emphasize the contribution's magnitude should, if anything, *increase* the salience of this factor, working against the hypothesis.

Results

The effects of sacrifice on perceptions of moral judgment and character were less pronounced in Study 2, when the donor's strong moral character was established, compared to Study 1, where it was not. Figure 2 plots the means.

Effects of contribution and effectiveness. Conditions were coded following the same procedure as Study 1. A linear regression was used to predict moral judgments from contribution, effectiveness, and their interaction.

For character judgments, there were no significant effects for any of the variables—neither contribution, $b = 0.07$, $SE = 0.10$, 95% CI[–0.12,0.26], $p = .48$, nor effectiveness, $b = 0.10$, $SE = 0.07$, 95% CI[–0.03,0.23], $p = .13$, nor their interaction, $b = –0.10$, $SE = 0.10$, 95% CI[–0.29,0.09], $p = .32$ reached significance. This is essentially a manipulation check, demonstrating that the manipulation successfully eliminated the diagnosticity of the specific donation for character.

For praise judgments, there was a significant effect of contribution, $b = 0.25$, $SE = 0.10$, 95% CI[0.06,0.44], $p = .009$, albeit weaker than in Study 1 (see moderated mediation analysis below). Thus, moral judgments were more positive for actors making larger contributions, but this effect was less pronounced in Study 2, where moral character was established through independent evidence, compared to Study 1.

Interestingly, there was also a modest effect of effectiveness on moral judgments, $b = 0.14$, $SE = 0.07$, 95% CI[0.01,0.27], $p = .039$, driven particularly by differences between effectiveness conditions when sacrifice was low. (This interaction, however, did not reach significance, $b = –0.16$, $SE = 0.10$, 95% CI[–0.35,0.03], $p = .10$.) This was not predicted *a priori* and should be taken with caution. One possibility is that if one is known to have a strong reputation, it may require considerable evidence to revise this default belief. When a donation is low in both magnitude and effectiveness, the combination of these two cues may provoke a negative revision to beliefs about that actor's character. A second possibility is that personal sacrifice "crowds out" social benefit when reputation is unknown, but that there is room for social benefit to play a role when there is no need to establish reputation. However, these speculations are not tested directly, and these small, unpredicted effects should be interpreted cautiously until replicated.

Moderated mediation. To test whether differences in character inferences accounted for the difference across Studies 1 and 2, a moderated mediation analysis (PROCESS Model 7; Hayes, 2013) was conducted on the combined dataset ($N = 1042$).

As shown in Figure 3, character (the mediator) was predicted by contribution, $b = 0.26$, $SE = 0.08$, $p = .001$, 95% CI[0.11,0.41] and by character information (–1 = Study 1, 1 = Study 2), $b = 0.29$, $SE = 0.05$, $p < .001$, 95% CI[0.19,0.39]. Importantly, the interaction was significant, $b = –0.19$, $SE = 0.08$, $p = .015$, 95% CI[–0.34,–0.04], as contribution was a stronger predictor when character information was absent. Bootstrapping revealed that there was an indirect effect of contribution on praise judgments via character judgments for Study 1, $b = 0.23$, $SE = 0.07$, 95% CI[0.10,0.36], but not Study 2, $b = 0.03$, $SE = 0.04$, 95% CI[–0.05,0.12]. This led to a significant index of moderated mediation, $b = –0.20$, $SE = 0.08$, 95% CI[–0.35,–0.05]. Thus, character judgments mediate the effect of contribution magnitude on praise judgments only when the actor's moral reputation is unknown.

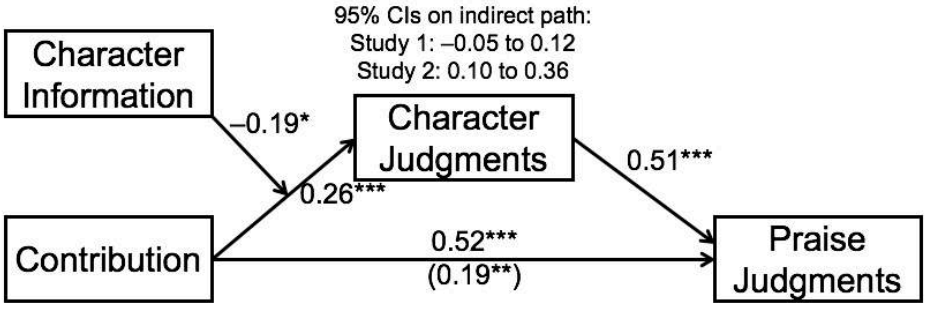


Figure 3: Moderated mediation model for Studies 1 and 2.

Discussion

Together, Studies 1 and 2 tell a clear story about moral evaluations of prosocial acts. Without evidence of reputation, prosocial behaviors are evaluated mainly by considering their personal sacrifice, rather than the social benefit. This occurs because personal sacrifice is a controllable and visible signal of cooperativeness and thus a useful input to reputational judgments. Thus, when reputation is available, personal sacrifice is less relevant to moral evaluations

One possible concern about the character information manipulation differentiating Studies 1 and 2 is that this manipulation also introduced a reference point (the donor's salary in dollars) in addition to establishing the donor's generosity. However, since the key finding of Study 2 is that people rely *less* on personal sacrifice (in terms of dollars), it seems unlikely that this result would be explained by introducing a reference point. If anything, a reference point should make contribution amounts more salient and more readily comparable to the reference point, leading people to rely on contribution more rather than less. Nonetheless, future work might further rule out this concern by manipulating character in other ways (e.g., mentioning that the donor also volunteers her time or has dedicated her career to prosocial causes).

General Discussion

Do we admire altruists because they make personal sacrifices or because they help others? The present studies found that, for altruistic donations of money, moral praise is driven almost entirely by sacrifice (Study 1). This occurs because personal sacrifice, but not social good, is taken as a signal of moral character (Study 2).

These results have implications for the psychology, philosophy, and practice of altruism. The findings are consistent with evolutionary accounts of moral psychology, according to which our moral faculties evolved to facilitate cooperation by tracking others' reputations and creating social rewards for those willing to act for the group's benefit (Nowak & Sigmund, 2005; Sperber & Baumard, 2012). If this is true, then our evaluations (e.g., praise and blame) of prosocial behaviors would track changes to the moral reputation or character of the actor. Since personal cost, but not social good, is usually under the actor's direct personal control, the former is a more reliable signal of cooperativeness.

How far would we expect these effects to generalize beyond this task? As discussed in conjunction with Study 1, the results do not seem to depend on the fact that the beneficiaries live in distant countries, as similar results are observed when the beneficiaries are Americans (Johnson, 2018). Other boundary conditions, however, may be plausible.

For example, the donors in the current studies may be seen as "outsourcing" the effectiveness of their charity to experts, and would thus not be seen as responsible for the outcome (e.g., Erat, 2013). This may be plausible, given previous work finding that people sometimes attribute more responsibility to individuals later in a causal chain (e.g., Brickman, Ryan, & Wortman, 1975; Spellman, 1997) as well as research documenting intransitivity beliefs about causal judgments (i.e., X causes Y and Y causes Z, but X does not cause Z; Johnson & Ahn, 2015). In that case, people may value effectiveness more when a prosocial agent contributes directly rather than indirectly. A more specific version of this possibility is that people think differently about the effectiveness of time- versus money-donations. Previous work has indeed documented differences in how people think about donations of money versus time (Johnson & Park, 2019; Liu & Aaker, 2008; Reed, Aquino, & Levy, 2007). Would effectiveness also loom larger for time-donations?

To test this, a replication of Study 1 was conducted, replacing the money-donations with time-donations (Johnson, 2018). The effects found in Study 1 were indeed reversed: Effectiveness but not sacrifice drove praise judgments. That is, unlike Study 1, contribution magnitude did not predict praise judgments, $b = 0.09$, $p = .32$, whereas effectiveness did, $b = 0.17$, $p = .007$. Thus, donation type (time vs. money) appears to be a boundary condition, such that effectiveness matters for time- but not for money-donations.

This is broadly consistent with the causal responsibility account, according to which effectiveness is only deemed irrelevant when it is outsourced to others. Indeed, low effectiveness in time-donations may signal incompetence as much as prosociality. This account alone does not

easily explain why personal sacrifice was not also used when evaluating time-donations. One possibility is that people place a greater psychological value on money than on other resources (Johnson, Zhang, & Keil, 2018), so that the sacrifice only looms large for money- but not time-donations. Future work might directly test these proposed mechanisms—the competence-signaling value of time effectiveness and the valuation difference between time and money sacrifices—in prosocial contexts.

These results are mainly bad news for effective altruism, whose *raison d'être* is improving the quality of prosocial acts, not merely their quantity. Effective altruists may receive no more social praise than *ineffective* altruists who make comparably large donations, even if the former do far more good for the world. This compounds a related problem, that people often view the importance of various causes as subjective, rather than objectively measurable (Berman et al., 2018). However, people may well be able to account for effectiveness in their moral evaluations when this factor is more salient and the causes are easily comparable. Websites like givewell.com, which directly compare charities in terms of metrics such as dollars per life saved, may be an important front on the battle for effective giving. More broadly, interventions that make both the quantity and quality of donations publicly observable may help to incentivize effective prosocial behavior.

Acknowledgements

I thank Josh Knobe, Seo Young Park, Yvetta Simonyan, and the members of the Bath Behavioural Science Lab for useful discussion.

References

Ariely, D., Bracha, A., & Meier, S. (2009). Doing good or doing well? Image motivation and monetary incentives in behaving prosocially. *American Economic Review*, *99*, 544–555.

Barasch, A., Levine, E. E., Berman, J. Z., & Small, D. A. (2014). Selfish or selfless? On the signal value of emotion in altruistic behavior. *Journal of Personality and Social Psychology*, *107*, 393–413.

Baron, J., & Szymanska, E. (2011). Heuristics and biases in charity. In D. M. Oppenheimer & C. Y. Olivola (Eds.), *The science of giving: Experimental approaches to the study of charity* (pp. 215–235). New York, NY: Psychology Press.

Bentham, J. (1907). *An introduction to the principles of morals and legislation*. Oxford, UK: Clarendon Press. (Original work published 1789.)

Berman, J. Z., Barasch, A., Levine, E. A., & Small, D. A. (2018). Impediments to effective altruism: The role of subjective preferences in charitable giving. *Psychological Science*, *29*, 834–844.

Brickman, P., Ryan, K., & Wortman, C. B. (1975). Causal chains: Attribution of responsibility as a function of immediate and prior causes. *Journal of Personality and Social Psychology, 32*, 1060–1067.

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, *108*, 353–380.

Delgado, M. R., Frank, R. H., & Phelps, E. A. (2005). Perceptions of moral character modulate the neural systems of reward during the trust game. *Nature Neuroscience*, *8*, 1611–1618.

Erat, S. (2013). Avoiding lying: The case of delegated deception. *Journal of Economic Behavior & Organization*, *93*, 273–278.

Goodwin, G. P., Piazza, J., & Rozin, P. (2014). Moral character predominates in person perception and evaluation. *Journal of Personality and Social Psychology*, *106*, 148–168.

Greene, J. D., Morelli, S. A., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2008). Cognitive load selectively interferes with utilitarian moral judgment. *Cognition*, *107*, 1144–1154.

Haidt, J. (2007). The new synthesis in moral psychology. *Science*, *316*, 998–1002.

Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. New York, NY: Guilford.

Inbar, Y., Pizarro, D. A., & Cushman, F. (2012). Benefiting from misfortune: When harmless actions are judged to be morally blameworthy. *Personality and Social Psychology Bulletin*, *38*, 52–62.

Johnson, S. G. B., & Ahn, W. (2015). Causal networks or causal islands? The representation of mechanisms and the transitivity of causal judgment. *Cognitive Science, 39*, 1468–1503.

Johnson, S. G. B., & Park, S. Y. (2019). *Moral evaluations of time versus money donations.* Available at SSRN.

Johnson, S. G. B., Zhang, J., & Keil, F. C. (2018). Psychological underpinnings of zero-sum thinking. In T. T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 566–571). Austin, TX: Cognitive Science Society.

Johnson, S. G. B. (2018). *Dimensions of altruism: How do evaluations of charitable behavior track prosocial benefit or personal sacrifice?* Available at SSRN.

Kahneman, D., & Knetsch, J. (1992). Valuing public goods: The purchase of moral satisfaction. *Journal of Environmental Economics and Management*, *22*, 57–70.

Klein, N., & Epley, N. (2014). The topography of generosity: Asymmetric evaluations of prosocial actions. *Journal of Experimental Psychology: General*, *143*, 2366–2379.

Liu, W., & Aaker, J. (2008). The happiness of giving: The time-ask effect. *Journal of Consumer Research*, *35*, 543–557.

MacAskill, W. (2015). *Doing good better: How effective altruism can help you make a difference*. New York,

NY: Penguin.
Mikhail, J. (2007). Universal moral grammar: Theory, evidence, and the future. *Trends in Cognitive Sciences*, *11*, 143–152.
Mill, J. S. (1998). *Utilitarianism*. Oxford, UK: Oxford University Press. (Original work published 1861.)
Nagel, J., & Waldmann, M. R. (2013). Deconfounding distance effects in judgments of moral obligation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*, 237–252.
Newman, G. E., & Cain, D. M. (2014). Tainted altruism: When doing some good is evaluated as worse than doing no good at all. *Psychological Science*, *25*, 648–655.
Nowak, M. A., & Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature*, *437*, 1291–1298.
Olivola, C. Y., & Shafir, E. (2013). The martyrdom effect: When pain and effort increase prosocial contributions. *Journal of Behavioral Decision Making*, *26*, 91–105.
Reed, A., Aquino, K., & Levy, E. (2007). Moral identity and judgments of charitable behaviors. *Journal of Marketing*, *71*, 178–193.
Singer, P. (1981). *The expanding circle: Ethics, evolution, and moral progress*. Princeton, NJ: Princeton University Press.
Singer, P. (2015), *The most good you can do: How effective altruism is changing ideas about living ethically*. New Haven, CT: Yale University Press.
Small, D. A., Loewenstein, G., & Slovic, P. (2007). Sympathy and callousness: The impact of deliberative thought on donations to identifiable and statistical victims. *Organizational Behavior and Human Decision Processes, 102*, 143–153.
Spellman, B. A. (1997). Crediting causality. *Journal of Experimental Psychology: General, 126*, 323–348.
Sperber, D., & Baumard, N. (2012). Moral reputation: An evolutionary and cognitive perspective. *Mind & Language*, *27*, 495–518.
Uhlmann, E. L., Pizarro, D. A., & Diermeier, D. (2015). A person-centered approach to moral judgment. *Perspectives on Psychological Science*, *10*, 72–81.

Predictions from Uncertain Moral Character

Samuel G. B. Johnson[1], Gregory L. Murphy[2], Max Rodrigues[3] & Frank C. Keil[3]
(sgbjohnson@gmail.com, gregory.murphy@nyu.edu, mrod1791@gmail.com, frank.keil@yale.edu)
[1]School of Management, University of Bath, Bath, BA2 7AY UK
[2]Department of Psychology, New York University, New York, NY 10003 USA
[3]Department of Psychology, Yale University, New Haven, CT 06520 USA

Abstract

People assess others' moral characters to predict what they will do. Here, we study the computational mechanisms used to predict behavior from *uncertain* evidence about character. Whereas previous work has found that people often ignore hypotheses with low probabilities, we find that people often account for the possibility of poor moral character even when that possibility is relatively unlikely. There was no evidence that comparable inferences from uncertain non-moralized traits integrate across multiple possibilities. These results contribute to our understanding of moral judgment, probability reasoning, and theory of mind.

Keywords: Moral psychology; theory of mind; prediction; causal reasoning; categorization

Introduction

People intensely scrutinize others' moral characters. Is Hillary Clinton a bastion of moral sanity or a devious opportunist? Is Donald Trump a man of the people or a corrupt plutocrat? Is your neighbor Todd a good person because he donates 20% of his income to charity, or a bad person because he received a citation for reckless driving?

This obsession with moral character makes good evolutionary sense: People track reputation to assess who will cooperate (Sperber & Baumard, 2012). For this reason, some have argued that our moral judgments about individual acts are primarily determined by what those acts reveal about the actors' character, rather than the intrinsic properties of those acts (Goodwin et al., 2014; Uhlmann et al., 2015). This explains our interest in intentions when judging an act's wrongness (Cushman, 2008). More broadly, knowing another's character allows us to predict what he or she will do, just as knowing a thing's category tells you about its properties. A bird is likely to fly; a snake is likely to be venomous. A good person may lend a helping hand; a bad person may stab you in the back.

But often we do not know someone's moral character with any certainty. Todd gives money to charity, but might the charity be a money-laundering operation? He was driving at a reckless speed, but what if he may have been doing so because he needed to perform an emergency surgery? How do we predict Todd's actions when we cannot be sure of his intentions?

In this paper, we ask what computational principles govern our predictions of others' actions from uncertain beliefs about moral character. This work falls in a research tradition studying predictions from uncertain categories and uncertain beliefs about causation. For example, if you have uncertain evidence leading you to identify a bird as a heron with 65% probability and a crane with 35% probability, then when you predict the bird's behavior, you may assume it is definitely a heron, without hedging for the possibility it is a crane (Malt et al., 1995; Murphy & Ross, 1994). If you think there's a 75% chance that the Fed chair's statement implies a tightening of the money supply but a 25% chance it does not, you will act as though the Fed is certainly tightening the money supply when you are predicting the stock market (Johnson & Hill, 2017; Johnson et al., 2018). People often ignore uncertainty because it is computationally difficult to consider two possibilities simultaneously, considering the implications of each and integrating across those two possible worlds.

But perhaps people *would* integrate across possibilities when reasoning about moral character. First, we may have encapsulated, module-like mechanisms for aspects of mental-state understanding (Leslie, 1995) and moral judgment (Mikhail, 2007). Perhaps such domain-specific mechanisms perform more efficiently than domain-general mechanisms (Cosmides, 1989). Second, people *do* seem to integrate across possibilities for categories when one of the categories is dangerous (a shark) rather than neutral (a school of fish) (Zhu & Murphy, 2013).

By analogy, if you think an act is probably caused by a morally neutral motive, and then encounter evidence that the motive may actually have been immoral, you might take account of that motive when making further predictions about the person's future behavior. But if you instead encounter evidence that the motive may have been some *other* morally neutral motive, you may very well ignore that possibility when predicting behavior.

The Current Studies

Participants read scenarios about various actions taken by characters. For example, in one item, a driver struck a bicyclist while heading the wrong way on a one-way street. For each action, there were three possible explanations. One explanation was neutral (e.g., the driver did not know that the street was one-way; hypothesis H_{Neut}), one implied poor moral character (e.g., the driver hit the bicyclist deliberately to teach him a lesson; hypothesis H_{Imm}), and one implied that the person had some other, non-moralized trait, such as forgetfulness, risk-aversion, or poor eyesight (e.g., the driver had forgotten to turn her headlights on; hypothesis H_{Other}).

We developed a set of actions predicted by H_{Imm} or

H_{Other}. If H_{Imm} were true (the driver hit the bicyclist on purpose), then her immoral character would suggest other immoral actions (driving with expired registration; prediction Z_{Imm}). If H_{Other} were true (she forgot to turn her lights on), then her trait (forgetfulness) would suggest other related actions (leaving her windows open before a rainstorm; prediction Z_{Other}). In Pretest B, we obtained judgments of how likely each prediction was given each hypothesis. For example, we measured $P(Z_{Imm}|H_{Imm})$, the probability the driver would drive with an expired registration given that she had hit the bicyclist on purpose. We also measured $P(Z_{Imm}|H_{Neut})$, $P(Z_{Imm}|H_{Other})$, $P(Z_{Other}|H_{Neut})$, $P(Z_{Other}|H_{Imm})$, and $P(Z_{Other}|H_{Other})$.

For the Main Study, we were interested in predictions about these actions (Z_{Imm} and Z_{Other}) when participants had evidence rendering her motives (H_{Neut}, H_{Imm}, H_{Other}) uncertain, and how these predictions from uncertain motives would compare to the predictions from certain motives from Pretest B. We constructed two versions of each scenario. In one version (uncertain evidence U_{Imm}), the neutral explanation H_{Neut} was presented as most likely, but the immoral explanation H_{Imm} was also introduced as possible.[1] For example, the driver probably didn't know the street was one-way, but possibly hit the bicyclist on purpose. In Pretest A, we ensured that participants viewed the neutral intention (H_{Neut}) as likelier than the immoral intention (H_{Imm}) given the uncertain evidence—that she really *was* likelier to have forgotten about the one-way street—but also that H_{Imm} was still reasonably likely.

In the other-trait version (uncertain evidence U_{Other}), H_{Neut} was presented as most likely, but the other-trait explanation H_{Other} was also introduced as possible. For example, the driver probably didn't know the street was one-way, but possibly forgot to turn on her lights. Pretest A ensured that people viewed the neutral explanation as likelier than the other-trait explanation. Thus, Pretest A overall elicited judgments of $P(H_{Imm}|U_{Imm})$, $P(H_{Neut}|U_{Imm})$, $P(H_{Other}|U_{Other})$, and $P(H_{Neut}|U_{Other})$.

Our Main Study then tested whether people account for uncertainty about the actor's character given uncertain evidence (U_{Imm} and U_{Other}) when they are making predictions, measuring $P(Z_{Imm}|U_{Imm})$ and $P(Z_{Other}|U_{Imm})$. Would participants think the driver is likelier to perform an immoral act like driving with an expired registration (Z_{Imm}) when she possibly hit the bicyclist on purpose (U_{Imm}) than when she definitely did not (H_{Neut})? If people focus on the most likely hypothesis (i.e., ignore uncertainty about character), then they should view these immoral acts as equally likely regardless of whether there is a chance the driver behaved immorally. Moreover, they should view Z_{Imm} as much less likely if it is merely possible that the driver has poor moral character (U_{Imm}) compared to knowing this for sure (H_{Imm}). That is:

$$P(Z_{Imm}|H_{Neut}) = P(Z_{Imm}|U_{Imm}) < P(Z_{Imm}|H_{Imm})$$

Likewise, if people ignore uncertainty about non-moralized traits, the driver should be seen as equally likely to do other forgetful things regardless of whether it is possible that she forgot to turn on her lights:

$$P(Z_{Other}|H_{Neut}) = P(Z_{Other}|U_{Other}) < P(Z_{Other}|H_{Other})$$

But as mentioned earlier, people might attend to the lower-probability trait when it is moralized, but not when it is non-moralized. If so, people would think the driver likelier to commit other immoral acts even if it is merely possible that she hit the bicyclist on purpose. But people would not consider the driver likelier to commit other forgetful acts if it is merely possible that she forgot to turn on her lights:

$$P(Z_{Imm}|H_{Neut}) < P(Z_{Imm}|U_{Imm}) < P(Z_{Imm}|H_{Imm})$$
$$P(Z_{Other}|H_{Neut}) = P(Z_{Other}|U_{Other}) < P(Z_{Other}|H_{Other})$$

To test these hypotheses, Pretest A normed judgments of $P(H_i|U_k)$—inferences of character from uncertain evidence of intentions—and Pretest B normed judgments of $P(Z_i|H_j)$—predictions of future actions from certain knowledge of character. In the Main Study, we tested judgments of $P(Z_i|U_k)$—predictions of future actions from uncertain evidence of intentions, using the pretest norms to generate normative predictions for comparison.

Pretest A: Intentions from Uncertain Evidence

We first sought to norm the values of $P(H_{Neut})$ and $P(H_{Imm})$ given evidence U_{Imm} and the values of $P(H_{Neut})$ and $P(H_{Other})$ given evidence U_{Other}. That is, we evaluated the scenarios we constructed to be sure that readers interpreted them as intended. This serves two purposes. First, for an item to be included, we need the neutral explanation to be deemed likelier than the moral or non-moral trait explanations—that is, $P(H_{Neut}|U_{Imm}) > P(H_{Imm}|U_{Imm})$ and $P(H_{Neut}|U_{Other}) > P(H_{Other}|U_{Other})$. Second, these estimates are needed to compute normative responses in the Main Study.

Method

We recruited 100 participants from Amazon Mechanical Turk (50% female, M_{age} = 36.9). Participants were excluded if they incorrectly answered more than 30% of a set of 10 check questions (N = 9).

Each participant read eight items, with each item in one of two versions. In one version (U_{Imm}), the evidence suggested two possible explanations, H_{Neut} and H_{Imm}, with H_{Neut} designed to be more plausible. For example:

> Navigation through Tabbsboro is complicated by a set of one-way streets, which were put into place because the streets are too narrow to allow parking and traffic in both directions. The police recently reported on an accident that happened in one of them. One late afternoon, Cindy Harlan

[1] That is, U_X refers to a case in which the neutral explanation of the person's behavior is offered as likely, but X is mentioned as a less likely explanation. H_X refers to cases in which only explanation X is offered.

struck a bicyclist who was riding towards her car. The bicyclist was in her way and injured his hip in the accident. He was taken away in an ambulance.

[H_{Neut}] The police questioned Cindy, and she denied knowing that it was a one-way street. This was the first time she had driven on this road. There was no sign near the driveway, and she had not noticed that it was one-way when she arrived there.

[H_{Imm}] The reporting officer noted that the bicyclist, who was a teenager, had seen Cindy earlier that day, and that she seemed irritated when he and his friends didn't get out of the street fast enough when Cindy was driving to her acquaintance's home. The officer asked if Cindy went the wrong way down the road because she saw the bicyclist playing in the street again and wanted to teach him a lesson. Cindy denied this and said that she simply didn't know it was one-way.

Participants then estimated the probabilities of H_{Neut} ("Cindy hit the bicyclist because she didn't know she was driving the wrong way down a one-way street") and H_{Imm} ("Cindy hit the bicyclist because she was trying to teach the teenager a lesson"). These judgments were entered in separate text boxes on scales from 0 to 100. Since the hypotheses were not strictly exhaustive, the judgments did not have to sum to 100 (though most did).

In the other-trait version of each item (U_{Other}), the evidence suggested two possibilities, H_{Neut} and H_{Other}, with H_{Neut} again more plausible. For the item above, the last paragraph of the U_{Imm} version was replaced with:

[H_{Other}] The reporting officer noted that Cindy's lights were not on. He asked if she might not have seen the bicyclist because she had forgotten to turn her lights on. Cindy pointed out that the accident had happened almost an hour ago, when it was light out.

Participants then judged H_{Neut} and H_{Other} ("Cindy hit the bicyclist because she didn't have her lights on").

Participants saw one version of each item, with half of the items in version U_{Imm} and half in version U_{Other}, counterbalanced across participants. The order of the probability judgments was randomized for each item.

Results

All eight items met our desired conditions (see Appendix). Mean judgments of $P(H_{Neut}|U_{Imm})$ and $P(H_{Neut}|U_{Other})$ ranged from 65% to 82% across items (Ms = 75.1% and 74.3%, respectively), and judgments of $P(H_{Imm}|U_{Imm})$ and $P(H_{Other}|U_{Other})$ ranged from 17% to 32% (Ms = 24.2% and 25.3%, respectively).

Pretest B: Predictions from Certain Intentions

Next, we normed the values of the predictions, $P(Z_{Imm})$ and $P(Z_{Other})$, given certain intentions H_{Neut}, H_{Imm}, and H_{Other}. Once again this has two purposes. First, an inclusion criterion: We want the immoral prediction Z_{Imm} to be more plausible given the immoral than the neutral intention—that is, $P(Z_{Imm}|H_{Imm}) > P(Z_{Imm}|H_{Neut})$—and likewise for the prediction Z_{Other} to be more plausible given the other-trait than the neutral intention—that is, $P(Z_{Other}|H_{Other}) > P(Z_{Other}|H_{Neut})$. For example, since the immoral prediction Z_{Imm} in our example was driving with an expired registration, we needed to ensure that people agree that someone who hits a bicyclist on purpose (H_{Imm}) is more likely to drive with an expired registration (Z_{Imm}) than someone who hit the bicyclist accidentally (H_{Neut}). Second, these values—predictions given certain intentions—are needed for comparison with the Main Study, which measured predictions given uncertain intentions.

Method

We recruited 149 participants from Mechanical Turk (29% female, M_{age} = 33.9). Participants were excluded using the same criterion as Pretest A (N = 25).

Each participant read eight items, with each item in one of three versions. In one version, H_{Neut} was true. For the Tabbsboro example, the first paragraph was the same as in Pretest A, and the remainder of the item read:

Cindy didn't realize that it was a one-way road. This was the first time she had driven on this road. There was no sign near the driveway, and she had not noticed that it was one-way when she arrived there.

In a second version, H_{Imm} was true:

Cindy had pulled out of an acquaintance's driveway and turned left, even though that was the wrong way for this one-way street. She went the wrong way because she saw several kids who had irritated her earlier in the day for not getting out of the street fast enough, so when she saw them again, she wanted to drive close to them to teach them a lesson.

Finally, in a third version, H_{Other} was true:

Cindy had pulled out of an acquaintance's driveway and turned left, but she didn't see the bicyclist because she had forgotten to turn her lights on.

Participants then estimated five probabilities for each item. We included two versions each of Z_{Imm} ("What is the probability that Cindy would drive her car with an expired vehicle registration?") and Z_{Other} ("What is the probability that Cindy would forget to shut her window the night before a thunderstorm?"). For the Main Study, we chose the best version for each item to maximize the chance we could use a given item. We also included a filler item ("What is the probability that the city will install a clearer sign in the next week?") which would not necessarily vary based on Cindy's intention. These judgments were made using the same procedure as Pretest A.

Participants saw one version of each item, with the eight items distributed about evenly across the three versions, counterbalanced across participants. The order of the probability judgments was randomized for each item.

Results

The probability of Z_{Imm} was consistently judged higher given H_{Imm} than given H_{Neut}; that is, $P(Z_{Imm}|H_{Imm}) > P(Z_{Imm}|H_{Neut})$ for all items (Ms = 65.6% and 26.4%, respectively), with the difference between these conditional probabilities ranging from 26.7% to 53.4% across items (see Appendix). $P(Z_{Other}|H_{Other})$ was higher than $P(Z_{Other}|H_{Neut})$ for all items but one (Ms = 60.7% and 49.8%, respectively), with the difference between these probabilities varying from –1.5% to 23.2% across items.

Since all items satisfied the desired conditions for Z_{Imm}, we did not exclude any items for the Main Study. However, these results suggest two caveats. First, it is difficult to compare participants' inferences about moral versus other kinds of traits, since the morally laden predictions (Z_{Imm}) were much more responsive to knowledge of intentions compared to the non-morally laden predictions (Z_{Other}—see later discussion). Given this limitation, any conclusions about moralized versus non-moralized character traits must be provisional. Second, because some of these items were not very robust for the non-morally laden predictions, we repeat key analyses on individual items.

Main Study: Predictions from Uncertain Evidence

The Main Study tested inferences about people's future actions based on uncertain knowledge of their intentions. Participants saw the evidence normed in Pretest A, making predictions about the actions normed in Pretest B.

Method

We recruited 99 participants from Mechanical Turk (54% female, M_{age} = 37.5). Participants were excluded using the same criterion as in the pretests (N = 1).

Participants read the eight vignettes used in Pretest A, each in one of the two versions (either U_{Imm} or U_{Other}). For each item, participants were asked questions across two pages (with the vignette text displayed on the screen for both). On the first page, participants made predictions of Z_{Imm} and Z_{Other}, using the phrasing normed in Pretest B. On the second page, participants indicated which intention they thought was likelier. For the U_{Imm} version of the item, participants chose between H_{Neut} ("Cindy hit the bicyclist because she didn't know she was driving the wrong way down a one-way street") and H_{Imm} ("Cindy hit the bicyclist because she was trying to teach the teenager a lesson"); for the U_{Other} version, participants chose between H_{Neut} and H_{Other} ("Cindy hit the bicyclist because she didn't have her lights on").

Results

Overall, participants tended to place positive weight on the immoral explanation H_{Imm} when making predictions, even when they acknowledged that the neutral explanation H_{Neut} was likelier. This is a departure from most previous studies of predictions from uncertain beliefs. On the other hand, there was little evidence that participants placed any weight on the other-trait explanation H_{Other} when making predictions, which raises the possibility that people might attend selectively to evidence of immoral intentions. Finally, there was modest evidence that people underweighted H_{Imm} relative to normative standards, and considerable evidence for underweighting H_{Other}.

We tested reliance on H_{Imm} in two ways. First, we conducted an item-level analysis, averaging probability judgments across all participants (see Appendix). Unsurprisingly, mean judgments of $P(Z_{Imm}|U_{Imm})$ (31.9%) were lower than $P(Z_{Imm}|H_{Imm})$ in Pretest B (65.6%), $t(7) = 8.80$, $p < .001$, $d = 3.11$, reflecting the fact that H_{Imm} had a low prior probability given evidence U_{Imm}. More interestingly, mean judgments of $P(Z_{Imm}|U_{Imm})$ (31.9%) in this study were higher than $P(Z_{Imm}|H_{Neut})$ in Pretest B (26.4%), $t(7) = 3.37$, $p = .012$, $d = 1.19$. That is, when the evidence is uncertain between H_{Neut} and H_{Imm}, predictions of Z_{Imm} fall between predictions made when either H_{Neut} or H_{Imm} is certain. This shows that people take both H_{Imm} and H_{Neut} into account when predicting Z_{Imm}. (In English: When there are two possibilities, the induction will take both into account and therefore lie in between the predictions given either possibility alone.)

We can also use the data from Pretests A and B to calculate normative values of $P(Z_{Imm}|U_{Imm})$:

$$P(Z_{Imm}|H_{Neut})P(H_{Neut}|U_{Imm}) + P(Z_{Imm}|H_{Imm})P(H_{Imm}|U_{Imm})$$

These normative responses are given in the Appendix for each item (M = 35.7%). Participants' actual judgments (M = 31.9%) were marginally more conservative, $t(7) = 1.99$, $p = .087$, $d = 0.82$, compared to the normative responses, suggesting that participants underweighted H_{Imm}. Although statistically not very robust, this would be consistent with previous studies, finding that people underweight unlikely hypotheses, even when they do not ignore them entirely (Johnson, Merchant, & Keil, 2018).

This item analysis, however, can be criticized because it lumps together participants who agreed that H_{Neut} was likelier than H_{Imm} (which should be the dominant belief, based on Pretest A), with those who believed the converse. In fact, about 19% of responses disagreed with our assumption that H_{Neut} was likelier. Thus, the analysis above could be lumping together two populations: Those who believed H_{Neut} was likelier and assigned no weight to H_{Imm}, and those who believed H_{Imm} was likelier and assigned no weight to H_{Neut}. The item means would look like both hypotheses are being considered, but this is an illusion due to mixing two populations (Malt et al., 1995).

Thus, our second approach analyzed the data at the level of individual participants, including only participants for each item who agreed that H_{Neut} was the likelier than H_{Imm}. Using this approach, participants rated $P(Z_{Imm}|U_{Imm})$ numerically higher than the average pretest ratings of $P(Z_{Imm}|H_{Neut})$ for 6 of the 8 items, significantly so for three of the items (items 2, 5, and 6; ps < .02); one

item was significant in the opposite direction (item 7; p = .026). Overall, this evidence is moderately consistent with the idea that people often place weight on H_{Imm} even when they view H_{Neut} as likelier.

We also used this approach to test whether people would *underweight* H_{Imm} even when they assigned positive weight to it. For this analysis, we included all participants (even those indicating that H_{Imm} was likelier than H_{Neut}) because the estimates from Pretest A, used to calculate normative values, average across both kinds of participants. Participants rated $P(Z_{\text{Imm}}|U_{\text{Imm}})$ numerically lower than its normative value for 5 of the 8 items, but significantly for only one item (item 7); no items were significant in the opposite direction.

The above analyses focused on the U_{Imm} condition. What about the U_{Other} condition? The item analysis found that judgments of $P(Z_{\text{Other}}|U_{\text{Other}})$ (M = 45.2%) were lower than $P(Z_{\text{Other}}|H_{\text{Other}})$ from Pretest B (M = 60.7%), $t(7)$ = 2.85, p = .025, d = 1.01. But unlike the U_{Imm} condition, there was no evidence that people took account of H_{Other}, since $P(Z_{\text{Other}}|U_{\text{Other}})$ judgments (45.2%) did not differ from $P(Z_{\text{Other}}|H_{\text{Neut}})$ judgments from Pretest B (49.8%), and indeed were in the wrong direction on average, $t(7)$ = –0.86, p = .42, d = –0.31. That said, these judgments also did not differ significantly from their normative values (52.6%), $t(7)$ = 1.47, p = .18, d = 0.50 (although the normative values themselves did differ significantly from $P(Z_{\text{Other}}|H_{\text{Neut}})$; p = .023). These inconclusive results are probably due to the poor diagnosticity of intentions for predicting Z_{Other}, shown in Pretest B.

Since the diagnosticity differed across items, it is useful to conduct subject-level analyses for each item, as we did for the U_{Imm} condition. Looking at just those who agreed that H_{Neut} was likelier than H_{Other}, ratings of $P(Z_{\text{Other}}|U_{\text{Other}})$ were higher than the average pretest ratings of $P(Z_{\text{Other}}|H_{\text{Neut}})$ for only 3 out of the 8 items, with only one item reaching significance (item 1; p = .037), and three items reaching significance in the opposite direction (items 6, 7, and 8; ps < .001). Conversely, looking at all participants, ratings of $P(Z_{\text{Other}}|U_{\text{Other}})$ were lower than the normative scores for 5 out of the 8 items, with 4 items reaching significance (items 2, 6, 7, and 8; ps < .02). These results cast doubt on the idea that people place positive weight on the other-trait hypotheses when making predictions, suggesting that people underweight such hypotheses. However, this conclusion must be provisional given the poor diagnosticity of some of the non-moralized traits.

Discussion

Judgments of moral character are central to social life. They guide our decisions about who we interact with, inform our beliefs about what others are thinking, and help us to predict what others are going to do. But moral character is often ambiguous, since we often cannot know others' intentions with certainty. How do we predict others' behavior when their character is uncertain?

First, in contrast to other studies of predictions from uncertain beliefs (Johnson et al., 2018; Malt et al., 1995), we find that people have at least some ability to account for the possibility of immoral character, even when it is relatively unlikely. In vignettes where characters were assigned a 25% probability of a nefarious motive, people took this motive into account when predicting other immoral behaviors. Although this result was not consistent across all of our items, it was statistically robust for some of them and was significant overall.

Second, it is possible that this ability to account for uncertain traits is specific to moral traits. There was little evidence that participants weighted uncertain non-moral traits (e.g., poor eyesight) in predictions. This result is limited by the relatively poorer quality of our non-moral than of our moral items, suggesting the need for future research with more directly comparable items.

This problem, however, may reflect a real issue in making predictions about human behavior. When someone makes a mistake of some kind (as all these examples are), there are many factors that could be involved, and it may be hard to rule any out. If someone makes a wrong turn while driving, the person could well have not been paying attention, the sign might not have been very clear, the traffic might have been distracting, and so on. The presence of one of these explanations does not greatly reduce the possibility that one of the others also applied. Thus, such explanations based on non-moral character traits may not be very diagnostic about future actions. Morality, in contrast, may be of special interest to people precisely because it is thought to be diagnostic.

Third, we compared judgments to normative benchmarks. There was a trend toward underweighting the less-likely hypothesis. For the moral traits, this trend reached only marginal significance overall because participants' judgments were actually quite close to the normative benchmarks; only one individual item revealed significant evidence of underweighting. For the non-moralized traits, there was less room for reliable differences to emerge between actual and normative judgments overall, given the poor quality of some of the non-moral items. But there was considerable evidence for underweighting for half of the individual items. Thus, there seems to be more robust underweighting of unlikely non-moral traits than of unlikely moral defects.

These results contribute to several conversations in cognitive and social psychology. First, they add to our understanding of when people account (or fail to account) for uncertainty in probabilistic reasoning (e.g., Johnson et al., 2018; Zhu & Murphy, 2013). Second, they help to elucidate the mechanisms by which we compute others' mental states (Jara-Ettinger et al., 2016; Leslie, 1995). Finally, they sharpen our understanding of the interplay between domain-specific and domain-general computational principles in moral judgment (Cosmides, 1989; Mikhail, 2007). Moral reasoning may be more than just a special case of general-purpose thought.

References

Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition*, *31*, 187–276.

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, *108*, 353–380.

Goodwin, G. P., Piazza, J., & Rozin, P. (2014). Moral character predominates in person perception and evaluation. *Journal of Personality and Social Psychology*, *106*, 148–168.

Jara-Ettinger, J., Gweon, H., Schulz, L. E., Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in Cognitive Sciences*, *20*, 589–604.偏

Johnson, S.G.B., & Hill, F. (2017). Belief digitization in economic prediction. In *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 2314–2319). Austin, TX: Cognitive Science Society.

Johnson, S.G.B., Merchant, T., & Keil, F.C. (2018). *Belief digitization: Do we treat uncertainty as probabilities or as bits?* Available at SSRN.

Leslie, A. M. (1995). A theory of agency. In *Causal cognition: A multidisciplinary debate* (pp. 121–149). New York: Oxford University Press.

Malt, B.C., Ross, B.H., & Murphy, G.L. (1995). Predicting features for members of natural categories when categorization is uncertain. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 646–661.

Mikhail, J. (2007). Universal moral grammar: Theory, evidence, and the future. *Trends in Cognitive Sciences*, *11*, 143–152.

Murphy, G.L., & Ross, B.H. (1994). Predictions from uncertain categorizations. *Cognitive Psychology*, *27*, 148–193.

Sperber, D., & Baumard, N. (2012). Moral reputation: An evolutionary and cognitive perspective. *Mind & Language*, *27*, 495–518.

Uhlmann, E.L., Pizarro, D.A., & Diermeier, D. (2015). A person-centered approach to moral judgment. *Perspectives on Psychological Science*, *10*, 72–81.

Zhu, J., & Murphy, G.L. (2013). Influence of emotionally charged information on category-based induction. *PLoS ONE*, *8*, e54286.

Appendix

		Pretest A			
		$P(H_{Neut}\|U_{Imm})$	$P(H_{Imm}\|U_{Imm})$	$P(H_{Neut}\|U_{Other})$	$P(H_{Other}\|U_{Other})$
Item 1	Hitting bicyclist	65.4	31.5	73.4	29.4
Item 2	Taking someone's umbrella	78.7	21.4	72.5	23.6
Item 3	Assigning jobs	75.7	24.8	67.5	32.1
Item 4	Staring at student	69.4	29.8	73.1	26.2
Item 5	Writing essay	81.0	17.9	74.2	25.4
Item 6	Hitting sports opponent	81.4	18.2	75.1	25.0
Item 7	Child's eye injury	73.3	25.8	82.0	16.9
Item 8	Medical recommendation	76.0	24.3	76.8	23.5
Mean		75.1	24.2	74.3	25.3
		Pretest B			
		$P(Z_{Imm}\|H_{Imm})$	$P(Z_{Imm}\|H_{Neut})$	$P(Z_{Other}\|H_{Other})$	$P(Z_{Other}\|H_{Neut})$
Item 1	Hitting bicyclist	55.3	28.6	45.9	23.3
Item 2	Taking someone's umbrella	71.4	35.2	66.5	43.3
Item 3	Assigning jobs	84.3	31.0	56.2	54.4
Item 4	Staring at student	72.1	36.6	67.2	64.8
Item 5	Writing essay	60.2	9.0	72.9	57.8
Item 6	Hitting sports opponent	63.1	18.2	35.9	14.9
Item 7	Child's eye injury	72.0	33.3	71.1	67.9
Item 8	Medical recommendation	46.3	19.5	70.3	71.8
Mean		65.6	26.4	60.7	49.8
		Main Study			
		Actual		**Normative**	
		$P(Z_{Imm}\|U_{Imm})$	$P(Z_{Other}\|U_{Other})$	$P(Z_{Imm}\|U_{Imm})$	$P(Z_{Other}\|U_{Other})$
Item 1	Hitting bicyclist	37.6	35.1	37.3	29.7
Item 2	Taking someone's umbrella	43.0	39.5	42.9	49.0
Item 3	Assigning jobs	37.3	61.6	44.1	55.0
Item 4	Staring at student	41.7	62.7	47.3	65.4
Item 5	Writing essay	16.2	27.2	18.2	61.6
Item 6	Hitting sports opponent	27.1	25.7	26.4	20.2
Item 7	Child's eye injury	28.1	55.6	43.4	68.4
Item 8	Medical recommendation	24.2	54.7	26.0	71.4
Mean		31.9	45.2	35.7	52.6

Note. Entries are the mean probability judgments for each item (expressed as percentages), averaged across participants.

Individual Differences in Self-Recognition from Body Movements

Akila Kadambi (akadambi@ucla.edu)[1]

Hongjing Lu (hongjing@ucla.edu)[1,2]

[1] Department of Psychology, University of California, Los Angeles
[2] Department of Statistics, University of California, Los Angeles

Abstract

Since we rarely view our own body movements in our daily lives, understanding the recognition of self-body movement can shed light on the core of self-awareness and on the representation of actions. We first recorded nine simple and nine complex actions performed by individual participants, who also subsequently observed nine videos displayed on the screen and imitated these actions. After a delay period of 35-40 days, participants were asked to identify their self- body movements presented as point-light displays amongst three other actors who performed the same actions. Participants were able to recognize themselves solely based on kinematics in point-light displays. However, self-recognition accuracy varied according to the complexity of performed actions, with more accurate self-recognition for complex than simple actions. The ability of self-recognition with simple actions showed a significant relation with autistic traits (negative relation: poorer self-recognition accuracy with more autistic traits), schizophrenic traits (quadratic non-linear relation, participants with the median degree of schizophrenia traits performed better than participants at the extremes), and with imitation actions and motor imagery traits (linear relation: increased self-recognition accuracy with greater motor imagery). We also found that participants did not recognize actions that only required visual experience but could identify their self-generated actions that required motor experience, underscoring the importance of motor experience to the representation of self-body movements.

Keywords: self-recognition, body movement, action, individual differences

Introduction

Of the fundamental prerequisites of human existence, the recognition of the "self" is a crucial pre-reflective, automatic process, underlying human perception and cognition. The ability to self-recognize is fundamental to the construction of an identity, agency, self-awareness, and self-consciousness (Gallup, 1970), and impairments in self-recognition ability can impact the quality of social interaction and communication (Ornitz & Ritvo, 1968)

Constructing the "self" is complex, accounted for by various disciplines all attempting to instantiate a definition. For example, examining a singular construct such as self-consciousness, has been extensively studied in humans, other primates, dolphins, and even extended to non-human agents, such as robots. Importantly, most of these accounts of self-processing are rooted in recognition-based self-face processing (e.g., Uddin, Iacoboni, Lange, and Keenan, 2007), famously standardized by Gallup (1970) in his prototypical mirror mark test. However, only relying on self-face recognition as an index for identifying the self is limited to serve as a general account for the integrated self-processing based holistically on face, body, voice and even body movements.

Given the dynamism of our everyday environment and lack of privileged access to viewing our bodies in motion, movements of our own body may serve as a good measure without relying on rich visual experiences of the self. In this vein, several studies extended self-recognition from static faces to whole-body movements, with visual input reduced to dynamic dot movements, as in point-light displays. After participants' body movements were recorded with a motion capture system, participants were still able to recognize their own action, even with scant visual information (Cutting & Kowlowski, 1977). Such above-chance performance for self-recognition extracted from predominantly from body kinematics was found for many different actions that varied in complexity (Loula et al., 2005; Burling, Kadambi, Safari, & Lu, 2018).

The impact of intrinsic traits to self-recognition ability, on the other hand, is less studied in the literature. There are a number of reasons as to why it is important to measure individual difference traits in self-body recognition. First, the unique contribution of various individual difference measures can uncover critical information that could potentially be lost through group-level averaging (Peterzell, 2016). Additionally, self-recognition is a complex process, with its investigation particularly hampered by its own operationalization and resulting lack of objectivity (consisting of no clear-cut computational investigation).

What individual differences might impact self-recognition from body movements? The joint contribution of both action perception and understanding likely recruits a distinct neural system, with the most prominent account surrounding the mirror neuron system. The mirror neuron account of action understanding suggests that perception and action are tightly linked through a "mirroring", simulation-based mechanism that allows humans to understand the kinematic goals of actions (Rizzolatti & Craighero, 2004). Impairments in this mirroring mechanism may underlie social perception disorders such as Autism and Schizophrenia. Consistent with this view, previous behavioral research in biological motion perception has shown that individuals with Autism (Blake et al., 2003, Moore et al., 1997) and individuals with Schizophrenia demonstrate impairments in biological motion

perception, such as in discriminating communitive actions from non-communicative actions presented in point-light displays (Okruszek et al., 2015).

The ability to interpret social actions not only shows impairment in individuals clinically diagnosed with Autism Spectrum Disorder and Schizophrenia, but also individual differences amongst typical populations in those with varied degrees of autistic traits (Miller & Saygin, 2013; Ahmed & Vander Wyk, 2013; van Boxtel, et. al., 2017), as well as schizophrenic traits, which impacts self-face processing (Platek & Gallup, 2002). Given the individual differences in biological motion perception in the general population, it is possible that people may show differing ability in self-recognition of own body movements. To date, only one other study (Burling et al., 2018) has compared self-recognition performance of body movements between people with high degree of autistic traits and people with low autistic traits. This study found a significant difference at the performance level between the two groups of participants. However, no study has systematically mapped out any other individual difference measures and run a large sample of participants to examine the individual differences in self-recognition from body movements.

In the present study, we included three different individual difference measures: autistic traits, schizophrenic traits, and motor imagery traits, all of which are linked to both social perception and likely functions of the mirror neuron system. Three main research questions were addressed. First, how well can people identify themselves from only the kinematics of body movements, and does the performance of self-recognition depend on the complexity of performed actions? Second, to what extent does the interplay between motor (more mirror-based) and visual experience (more perception-based) determine the performance of self-recognition from actions? Finally, how do individual differences in the ability to recognize own-actions displayed in point-light stimuli relate to motor imagery ability and distinct socio-cognitive traits (autistic and schizophrenic)?

Experiment

Method

Participants 71 undergraduate students (M_{age} = 20.98) were recruited through the Subject Pool at the University of California, Los Angeles. The study was approved by the UCLA Institutional Review board. All participants were provided course credit for their participation, and were naïve to the purpose of the study. Participants had normal or corrected-to-normal vision and no physical disabilities.

Procedure The experiment was split into two sessions: motion recording and action recognition. The first phase consisted of a motion recording session, where participants performed various actions and were recorded with a motion capture system. The second phase, consisted of two action recognition components. The first component, the self-recognition session, occurred after a delay period of 30 – 45 days. The stimuli were first generated in the action recording session and subsequently tested in the self-recognition task. Immediately, following the self-recognition task, participants completed the final action recognition task, consisting of a visual recognition" task.

Materials

Apparatus Participants' body movements were recorded using the Microsoft Kinect V2.0 and Kinect SDK in a quiet testing room. Here, participants were instructed to perform the actions in a rectangular 2.5 x 5 ft space, in order to provide flexibility to perform the action, while remaining within recording distance. The Kinect was placed 5 ft above the floor and 8.5 ft away from the participant. The three-dimensional (X-Y-Z) coordinates of the key joints were extracted at a rate of approximately 33 frames per second and later used to generate point-light displays of actions (see Figure 1). Customized software developed in our lab was utilized to enhance movement signals, and to carry out additional processing and trimming for actions presented later in the testing phase (van Boxtel & Lu, 2013).

Stimuli Generation For each participant, 27 point-light displays performing different actions were captured based on their body movement recordings. The first nine actions were simple actions which included *grab, jump, wave, lift, kick, hammer, push, point, punch.* The next nine actions were complex actions, which included: *argue, macarena, wash windows, play baseball, get attention, hurry up, fight, stretch, and play guitar.* These actions were selected in part based on a previous self-recognition study (Burling et al., 2018), but four actions (*macarena, wash windows, play baseball, play guitar*) were modified to be more constrained from their original actions (*dance, clean, play sport, play instrument*) in order to reduce the impact of memory cues. The actions varied in complexity in order to characterize a broad range of common movements in daily life. During action selection, simple and complex actions were determined by whether the action was a simple goal (e.g., wave) or a complex goal (e.g., argue), and all actions were selected to be commonly encountered actions.

The final nine actions were labeled imitation actions, which included *jumping jacks, basketball, bend, direct traffic 1, direct traffic 2, conversation, laugh, digging a hole, and chopping wood.* The nine imitation actions were selected from the Carnegie Mellon Graphics Lab Motion Capture Database available online (http://mocap.cs.cmu.edu) and also captured a broad range of variability and goal-directed actions. Some imitation videos were easily recognizable to subject (e.g., basketball), while others were unclear in what they conveyed (e.g., directing traffic). Each video displayed a stick figure performing one of the imitation actions and was presented in three different angles to the subject, either to the right or left (+/- 45°) or facing forward (0°) by rotating the horizontal axis. The varying viewpoints were included in order to assess the inherent viewpoint dependence in self-recognition. Each imitation action was recorded twice: once after viewing the three different angles, and once more after viewing only the forward-facing angle. In the self-recognition phase, the first imitation recording served as

practice, and only the second imitation recording was utilized.

Following action recording and prior to the self-recognition session, we filtered noise from the movements by applying a double exponential adaptive smoothing filter (LaViola, 2003), in order to remove recording errors from the Kinect system (e.g., missing a joint due to occlusion or small jitter for some joints). Additionally, the stimuli were trimmed in order to display the point light-displays (van Boxtel & Lu, 2015) with their segmented action recording, which would be reiteratively looped in the self-recognition session.

Procedure

Motion Recording Session

For the 18 simple and complex actions in the first recording session, participants were provided verbal instruction and instructed to perform the actions as naturally as possible. As a result, the action was open to interpretation, in order to emphasize the lack of a systematic way to perform the action. For the remaining nine imitation actions, all the participants were naïve to the name of the action. Instead, participants were visually instructed to *imitate* the actions (however they chose to imitate), in order to emphasize their naturalistic response to "imitation." After completing the action recording, participants completed two questionnaires: Schizotypal Personality Questionnaire (SPQ) and the revised Vividness of Motor Imagery Questionnaire (VMIQ-2). The SPQ was administered to assess degrees of schizotypal traits among individuals in the typical population. The VMIQ-2 was included to assess motor imagery differences as a potential source of variability in biological motion perception. Since perception and motor imagery representations presumably share common resources, we hypothesized that there may be correlations between the two abilities (Miller & Saygin, 2013; Iacoboni & Dapretto, 2006).

Recognition Session: Self-Recognition Task

In the subsequent self-recognition task, participants returned after a delay period of 30-45 days later in order to minimize the effect of memory on performance. Participants were seated 2.5 feet in front of a monitor in a dimly lit room and were asked to select their own action amongst three other distractor actions spread out horizontally along the center of the screen, as shown in Figure 1.

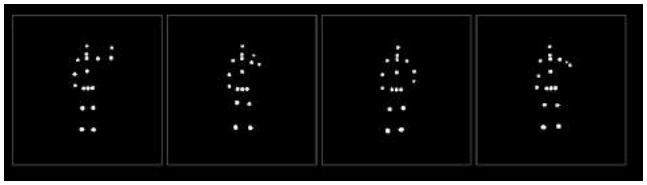


Figure 1. Illustration of a sample trial showing wave action (wave). One point-light display is the participant's action, while the other three point-lights are distractor actions normalized for gender, width, and height.

Each action was presented with 17 point-lights located at key joints, in three different orientations (rotated around the vertical axis 0°, (facing front), 45° (facing right), 225° (facing left), for a total of 81 trials. However, all of the actions within a trial displayed the same orientation. The actions were looped until the participant selected one of the four boxes, or until a time period of 30 seconds. Participants were not provided any feedback. Participants were instructed to select their own point-light action amongst four displays. The four animations included their own action and the corresponding actions performed by three other distractor actions that were normalized for height and gender.

Recognition Session: Visual Recognition Task

44 of the participants also participated in an additional visual recognition task consisting of nine trials displaying only the forward-facing imitation actions. The order of presentation of the visual recognition task was counterbalanced to either follow or precede the self-recognition task. Since imitation is a unique behavior that consists of both action observation and action performance, this additional task was included to assess whether performance would differ from the self-recognition task, and to understand the contribution of motor experience to self-recognition accuracy. Including this task could potentially allow us to contrast action observation in conjunction with execution (self-recognition task) with solely action observation (visual recognition task). Participants were instructed to identify the actor previously shown during the imitation recording amongst three other actors who performed the same action. Importantly, while the visual layout of the task was identical to the self-recognition session, the participants' own action was replaced by the original imitation actor from the Carnegie Mellon Database. As a result, participants' own point-light display was never amongst the four actions displayed on the screen. The remaining three distractor actions were maintained from the self-recognition session.

Following testing in the self-recognition and visual recognition task, participants were asked to complete an Autistic Quotient (AQ) questionnaire to assess the degree of Autistic traits (Baron, Cohen et al., 2001).

Individual Difference Measures

Autistic Quotient The Autism-Spectrum Quotient (AQ) questionnaire consists of 50 questions and is the most commonly used method to measure self-reported autistic traits (Baron-Cohen et al., 2001). Recent evidence has identified an overlapping genetic and biological etiology underlying ASD and autistic traits (Bralten et al., 2017) in addition to behavioral overlap (Baron-Cohen et al., 2001). Several studies of biological motion perception have reported an association between AQ scores and performance on various tasks (Miller & Saygin, 2013; Ahmed & Vander Wyk, 2013; van Boxtel et al., 2017). The AQ measures five different subtypes (social skill, attention switching, attention to detail, communication, and imagination).

Schizotypal Personality Questionnaire The Schizotypal Personality Questionnaire (SPQ) is a 74-question survey, designed to screen for schizotypal personality disorder in the general population. The SPQ is administered to assess

degrees of schizotypal traits among individuals in the typical population. It measures three constructs of schizotypy: cognitive, perceptual dimension (positive schizotypy), interpersonal dimension (negative schizotypy), and disorganized feature dimension (odd behavior, speech) based on DSM-IV criteria (Raine, 1991). The SPQ consists of nine different subtypes (ideas of reference, social anxiety, odd beliefs, unusual perceptual experiences, eccentric behavior and appearance, no close friends, odd speech, constricted affect, and suspiciousness/paranoid ideation).

Vividness of Motor Imagery Questionnaire The VMIQ-2 (Roberts, 2008) is designed to measure vividness of imagery in kinesthetic (movement simulation), internal (first person simulation), and external (third person simulation) visual imagery of 12 different actions in a series of three separate sections. Vividness of motor imagery is rated on a five-point Likert scale for each of the 12 actions in each of the three sub-areas (lower scores indicate more vivid images). According to simulation theory, perception and motor imagery representations share common resources (Miller & Saygin, 2013; Iacoboni & Dapretto, 2006). Therefore, the VMIQ-2 was included to assess motor imagery differences as a potential source of variability in biological motion perception.

Results

Self-recognition from body movements

Average self-recognition accuracy was .46 (SD = .12), significantly above chance level of .25 ($p < .001$), indicating that participants were able to self- recognize primarily on the kinematics of their body movements. As shown in Figure 2, participants were able to recognize all actions significantly above chance performance: for simple actions with verbal instruction (M = .40, SD = .15), for complex actions with verbal instruction (M = .56, SD = .16), and for imitated actions with visual display (M = .41, SD = .16). One-way ANOVA results revealed a significant main effect of action type (simple, complex, and imitation) on self-recognition performance, $F(2, 140) = 44.66$, $p < .001$, $\eta_p^2 = 0.389$. Specifically, self-recognition was more accurate for complex than simple actions ($t(70) = 9.026$, $p < .001$) and imitation actions ($t(70) = 7.749$, $p < .001$).

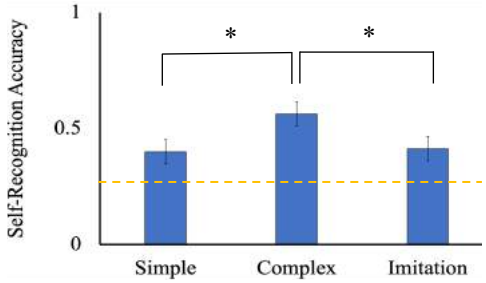


Figure 2. Self-recognition accuracy by the type of Action. Dashed line indicates chance performance (0.25). The error bars indicate standard error of means.

To examine whether the visual representation of own-body movements was viewpoint- invariant or viewpoint-specific, we conducted a one-way ANOVA consisting of orientation (facing left: 225°, front: 0°, right: 45°) on self-recognition performance $F(2, 140) = .335$, $p = .716$. We found that people recognized their own actions equally well from different viewpoints, suggesting a viewpoint-invariant representation of self-generated actions. A previous study similarly found that recognition of walking patterns from self-generated point-light displays was independent of the viewing angle. This is likely due to simulating the motor action through referring to three-dimensionally stored motor representations (Jokisch, Daum, & Troje, 2004).

We compared recognition of imitation actions from motor experience (as in the self-recognition task), and recognition of imitation actions from the visual experience task (where subjects had to identify the imitation action they observed but was not their own). We found people recognized actions less accurately from visual experience (M = .239) than from self-generated (M = .404) actions ($t(43) = 4.987$, $p < .001$). Due to around-chance performance for identical actions with only visual experience, prior visual experience does not appear to be sufficient for self- recognition. This suggests that motor experience may constrain visual experience and is critical to the recognition of one's own action. Importantly, every individual has experience with their own motor actions. Identifying oneself may require the ability to simulate the action onto one's own motor system, with self-recognition in turn dependent on a matching process- matching simulated action to performed action.

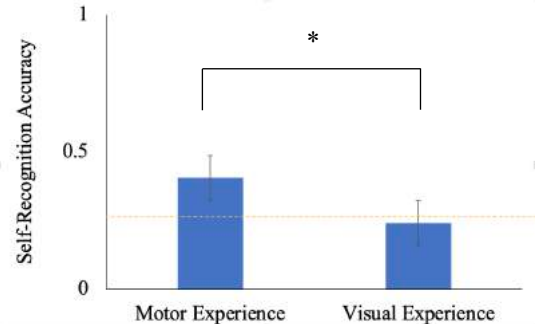


Figure 3. Self-recognition accuracy by experience type (visual vs motor). Significantly worse performance for imitation actions from visual experience than for self-recognition from performed actions. Dashed line indicates chance performance (0.25). Error bars indicate standard error of means.

Relations between self-recognition and individual difference measures

We did not find any significant correlations between self-recognition performance for complex actions and the individual difference measures. However, we found significant relations between self-recognition performance for simple actions with various individual difference measures. As shown in the top panel of Figure 3, a significant relationship was revealed between overall motor imagery

ability and self-recognition performance for imitation actions (spearman ρ = -.241, p = .043). For simple actions, a significant negative relationship emerged (Figure 3, middle) between the degree of autistic traits (AQ score) and self-recognition performance (spearman ρ = -.244, p = .040), revealing that people with more autistic traits performed less accurately in self-recognition with simple actions. To further probe the impact of autistic traits on self-recognition performance, we examined specific subtypes of the Autistic Quotient. We found a significant correlation between simple actions and the communication AQ subscale scores (spearman ρ = -.316, p = .007), but not with other subscale scores. For individual differences in schizophrenia traits, as shown in Figure 3 bottom plot, the trend analysis revealed a significant quadratic relationship between schizophrenia traits (SPQ score) and self- recognition performance, ($F(2,68) = 4.166$, p = .020) , with participants scoring near the median of SPQ scale performing better than participants at the extremes in self-recognition. More discussion about the non-linear relation is included in the discussion section.

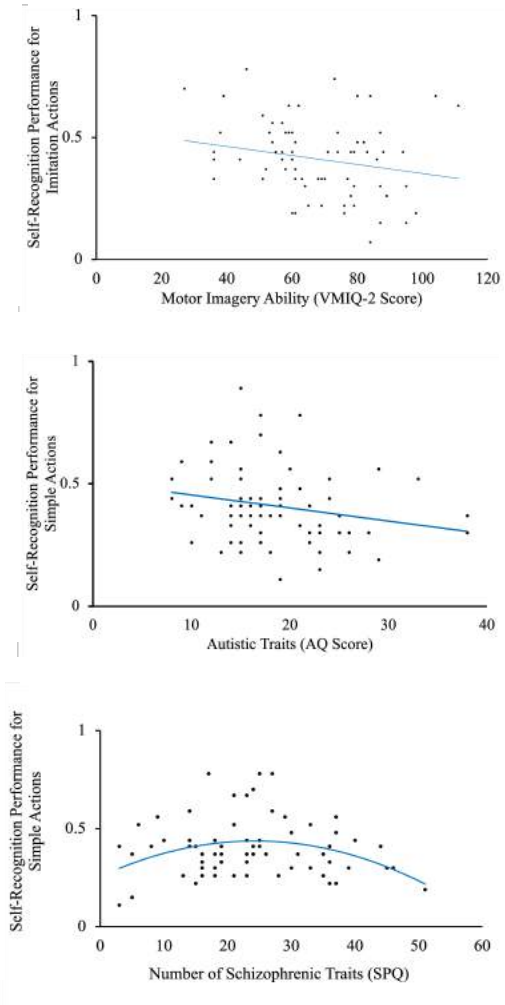


Figure 4. Relations between self-recognition performance and individual difference measures. Top: Positive relationship between motor imagery simulation and self-recognition for imitation actions. Middle: Negative relationship between autistic traits and self-recognition for simple actions. Bottom: Quadratic relationship between schizophrenic traits and self-recognition for simple actions (worse self-recognition at the extreme scores)

Discussion

The ability to self-recognize is integral to the construction of oneself as a unique entity, separate from the external world. Utilizing dynamic actions construed through self-generated point-light displays is a significant improvement over prototypical indices of self-recognition. Therefore, in the present study, we adopted the motion capture paradigm to examine how well people can identify themselves from only the kinematics of body movements from a range of commonly encountered actions. We found that participants were able to reliably self-recognize solely based on kinematics in point-light displays, in line with previous findings (Burling et al., 2018; Loula et al., 2005; Cutting & Kowlowski, 1977). Self-recognition accuracy also varied according to the complexity of performed actions, with more accurate self-recognition for complex than for simple actions, also corroborating a recent study (Burling et al., 2018). Since the complex and simple actions differed based on their variability, greater self-recognition for complex actions may be driven by the unique movement signatures available from these actions and increased motor planning (lack of automaticity) while performing complex actions. Importantly, the biometric identity cues in simple actions (e.g. walking) may not be readily apparent to the human visual system to recognize and differentiate these actions involving little variability (Dittrich, 1993; Loula et al., 2005). Therefore, participants exhibited greater self-recognition performance for the rich visual input conveyed by complex action sequences.

To assess the mechanisms underlying self-action recognition, we examined the contribution of visual and motor experience. Previous literature has indicated that people rely on motor experience when recognizing their own-body actions, as evidenced by greater recognition performance for self-generated point-light displays (reliant on motor experience) over close friends (reliant on visual experience) and strangers, presumably due to an internal simulation of the action (Loula et al., 2005). Conceptually, this is straightforward, as humans generally do not have privileged access to observe own locomotion movements from a third-person perspective, and consequently, experience little visual feedback (Jokisch, Daum, & Troje, 2004).

Therefore, to systematically contrast the relative importance of visual versus motor experience, we included an additional visual recognition task, wherein participants were asked to identify the imitation action they observed in the action recording session. We found that participants did not recognize actions that only required visual experience (actions they previously imitated, but that were not their own). Instead, participants were only able to identify their self-generated actions that required motor experience,

underscoring the importance of motor experience to the representation of self-body movements.

Finally, we measured individual differences in self-recognition performance. We looked at three correlates of variability in the general population: motor imagery (as measured by the VMIQ-2) and two social perception traits (autistic and schizophrenic traits). Both Autism and Schizophrenia are linked to dysfunctions of the mirror neuron system and impairments in social perception. Because action perception is presumed to involve an internal simulation on one's own motor repertoire, we hypothesized reduced simulation ability in individuals high on the Autistic Quotient and Schizophrenic Quotient.

Success in self-recognition with simple actions showed a significant relation with autistic traits (negative relation: poorer self-recognition accuracy with more autistic traits), schizophrenic traits (quadratic non-linear relation: participants with the median degree of schizophrenia traits performed better than participants at the extremes), and motor imagery traits (linear relation: increased self-recognition accuracy for imitation action with greater motor imagery).

We found that self-recognition performance for simple actions was affected by the participant's degree of autistic traits, in line with results from a recent study by Burling and colleagues (2018). One possible explanation could be due to a general processing style in autism, as decreased attention directed toward social stimuli in high-AQ individuals (see Chevellier et al., 2012) or weakened top-down influence (Lu, Tjan, Liu, 2006) and adaptability to social environment in autism (Thurman, et. al., 2016, van Boxtel, et. al., 2013). Although typical human adults are sensitive to social information in actions (Thurman & Lu, 2014; Su, van Boxtel & Lu, 2016), such ability is impaired in autism which could result in the worse performance in self-action recognition for people with high degree of autistic traits. Another explanation may pertain to a specific and mechanistic account, an underlying dysfunction in the mirror neuron system, with an impairment in self to other matching. A useful indicator related to the simulation-component of the mirror neuron system, is motor imagery, presumably reliant on an internal simulation of one's own motor system of the activated action (Jeannerod, 2001; Miller & Saygin, 2013). Specifically, the relationship between poorer self-recognition performance for simple actions and individuals with high autistic traits may be linked to worse motor imagery ability, as we found greater self-recognition accuracy with increased motor imagery ability. Additionally, in the clinical population, a previous study (Conson et al., 2013) found that subjects with Autism Spectrum Disorder exhibited alterations in mental hand rotation, specifically linked to impairments in motor action simulation. Further characterizing the link between motor imagery deficits and autistic traits in the general population may shed light on the underlying mechanisms of motor imagery and mirror neuron impairments in Autism.

We conjecture that worse performance for individuals with high schizophrenic traits may be due to over-simulation and motor imagery deficits (Sack et al., 2005), leading to delusions and hallucinations- a mark of positive schizotypy. For worse performance on simple actions with a low degree of schizophrenic traits, we hypothesize a lack of motor imagery ability as vividness of motor imagery is theorized to be an independent trait marker of Schizophrenia and simple actions may require a greater degree of simulation to dissociate between distractors (Sack et al., 2005).

Our study did not reveal any significant correlations between complex actions and the individual difference measures. Since complex actions may rely more on distinctive movement cues customized for different individuals, or long-term memory (specifically memory of how one would perform the action), it is likely that participants need not rely on motor simulation.

Collectively, the present results demonstrate that motor experience is an important component to understanding the core of self-body processing. Importantly, the perceptual representation of self-generated actions is affected by the degree of three key individual difference measures linked to the action understanding account of the mirror neuron system: autistic traits, schizophrenic traits, and motor imagery traits.

Acknowledgements

We thank Tabitha Safari, Aya Strauss, Liz Soltelo, Justin Azarian, Nazar Flome, Marian Spannowsky, and Sunhee Jin, for assistance in data collection. We thank Joseph Burling and Steve Thurman for help with the motion capture system. This research was supported by NSF Grant BCS-1655300.

References

Ahmed A. A., Vander Wyk B. C. (2013). Neural processing of intentional biological motion in unaffected siblings of children with autism spectrum disorder: an fMRI study. Brain Cogn. 83, 297–306. 10.1016/j.bandc.2013.09.007

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001). The autism-spectrum quotient (AQ): Evidence from asperger syndrome/high-functioning autism, malesand females, scientists and mathematicians. Journal of autism and developmental disorders, 31(1), 5-17.

Bralten, J., van Hulzen, K. J., Martens, M. B., Galesloot, T. E., Vasquez, A. A., Kiemeney, L. A., ... & Poelmans, G. (2017). Autism spectrum disorders and autistic traits share genetics and biology. Molecular psychiatry.

Burling, J. M., Kadambi, A., Safari, T., & Lu, H. (2018). The Impact of Autistic Traits on Self-Recognition of Body Movements. *Frontiers in Psychology*, *9*.

Cutting, J. E., & Kozlowski, L. T. (1977). Recognizing friends by their walk: Gait perception without familiarity cues. *Bulletin of the Psychonomic society*, *9*(5), 353-356.

Chevallier, C., Kohls, G., Troiani, V., Brodkin, E. S., & Schultz, R. T. (2012). The social motivation theory of autism. *Trends in cognitive sciences*, *16*(4), 231-239.

Conson, M., Mazzarella, E., Frolli, A., Esposito, D., Marino, N., Trojano, L., ... & Grossi, D. (2013). Motor imagery in Asperger syndrome: testing action simulation by the hand laterality task. *PLoS One*, *8*(7), e70734.

Dittrich, W. H. (1993). Action categories and the perception of biological motion. *Perception*, *22*(1), 15-22.

Dapretto, M., Davies, M. S., Pfeifer, J. H., Scott, A. A., Sigman, M., Bookheimer, S. Y., & Iacoboni, M. (2006). Understanding emotions in others: mirror neuron dysfunction in children with autism spectrum disorders. Nature neuroscience, 9(1), 28.

Gallup, G. G. (1970). Chimpanzees: self-recognition. Science, 167(3914), 86-87.

Jeannerod, M. (2001). Neural simulation of action: a unifying mechanism for motor cognition. *Neuroimage*, *14*(1), S103-S109.

Jokisch, D., Daum, I., & Troje, N. F. (2004). Self recognition versus recognition of others by biological motion: Viewpoint-dependent effects. *Journal of Vision*, *4*(8), 237-237.

LaViola J. J. (2003). An experiment comparing double exponential smoothing and Kalman filter-based predictive tracking algorithms, in Virtual Reality, 2003. Proceedings IEEE (Los Angeles, CA: IEEE;), 283–284.

Loula, F., Prasad, S., Harber, K., & Shiffrar, M. (2005). Recognizing people from their movement. *Journal of Experimental Psychology: Human Perception and Performance*, *31*(1), 210.

Lu, H., Tjan, B. S., & Liu, Z. (2006). Shape recognition alters sensitivity in stereoscopic depth discrimination. *Journal of Vision*, *6*(1),

Miller, L. E., & Saygin, A. P. (2013). Individual differences in the perception of biological motion: links to social cognition and motor imagery. Cognition, 128(2), 140-148.

Okruszek, Ł., Haman, M., Kalinowski, K., Talarowska, M., Becchio, C., & Manera, V. (2015). Impaired recognition of communicative interactions from biological motion in schizophrenia. PLoS One, 10(2), e0116793.

Ornitz, E. M., & Ritvo, E. R. (1968). Neurophysiologic mechanisms underlying perceptual inconstancy in autistic and schizophrenic children. Archives of General Psychiatry, 19(1), 22-27.

Platek, S. M., & Gallup Jr, G. G. (2002). Self-face recognition is affected by schizotypal personality traits. Schizophrenia Research, 57(1), 81-85.

Peterzell, D. H., & Kennedy, J. F. (2016). Discovering sensory processes using individual differences: a review and factor analytic manifesto. Electronic Imaging, 2016(16), 1-11.

Raine, A. (1991). The SPQ: a scale for the assessment of schizotypal personality based on DSM-III-R criteria. Schizophrenia bulletin, 17(4), 555-564.

Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annu. Rev. Neurosci.*, *27*, 169-192.

Sack, A. T., Van De Ven, V. G., Etschenberg, S., Schatz, D., & Linden, D. E. J. (2005). Enhanced vividness of mental imagery as a trait marker of schizophrenia?. Schizophrenia Bulletin, 31(1), 97-104.

Su, J., van Boxtel, J. J., & Lu, H. (2016). Social interactions receive priority to conscious perception. *PloS one*, *11*(8), e0160468.

Thurman, S. M., & Lu, H. (2014). Perception of social interactions for spatially scrambled biological motion. *PLoS One*, *9*(11), e112539.

Thurman, S. M., van Boxtel, J. J., Monti, M. M., Chiang, J. N., & Lu, H. (2016). Neural adaptation in pSTS correlates with perceptual aftereffects to biological motion and with autistic traits. *Neuroimage*, *136*, 149-161.

Uddin, L. Q., Iacoboni, M., Lange, C., & Keenan, J. P. (2007). The self and social cognition: the role of cortical midline structures and mirror neurons. *Trends in cognitive sciences*, *11*(4), 153-157.

van Boxtel, J. J., & Lu, H. (2013). A biological motion toolbox for reading, displaying, and manipulating motion capture data in research settings. Journal of vision, 13(12), 7-7.

van Boxtel, J. J., & Lu, H. (2015). Joints and their relations as critical features in action discrimination: Evidence from a classification image method. Journal of vision, 15(1), 20-20.

van Boxtel, J. J., Peng, Y., Su, J., & Lu, H. (2017). Individual differences in high-level biological motion tasks correlate with autistic traits. Vision research, 141, 136-144.

van Boxtel, J. J., & Lu, H. (2013). Impaired global, and compensatory local, biological motion processing in people with high levels of autistic traits. *Frontiers in psychology*, *4*, 209.

Statistical Learning Supports Word Learning and Memory

Ferhat Karaman
Usak University, Usak, Turkey

Jill Lany
University of Liverpool, Liverpool, United Kingdom

Jessica Hay
University of Tennessee, Knoxville, Knoxville, Tennessee, United States

Abstract

Learning new words does not only require infants to find words in continuous speech, but also be remember recently segmented words and link them to meaning. Prior research has shown that statistical learning supports word learning. However, as infant statistical learning was typically tested immediately after familiarization with a speech stream, we know very little about whether infants experience with statistical regularities supports long-term memory and future word learning. The current study was designed to shed light on the relationship between statistical learning, word learning, and memory. We found that while both co-occurrence statistics and syllable frequency information support word learning in the moment, co-occurrence information alone supports long-term memory for recently segmented candidate object labels.

How do infants start learning object names in a sea of clutter?

Hadar Karmazyn Raz (hkarmazy@iu.edu)
Drew H. Abney (dhabney@indiana.edu)
David Crandall (djcran@indiana.edu)
Chen Yu (chenyu@indiana.edu)
Linda B. Smith (smith4@indiana.edu)

Department of Psychological and Brain Sciences
Indiana University, Bloomington, IN 47405 USA

Abstract

Infants are powerful learners. A large corpus of experimental paradigms demonstrate that infants readily learn distributional cues of name-object co-occurrences. But infants' natural learning environment is cluttered: every heard word has multiple competing referents in view. Here we ask how infants start learning name-object co-occurrences in naturalistic learning environments that are cluttered and where there is much visual ambiguity. The framework presented in this paper integrates a naturalistic behavioral study and an application of a machine learning model. Our behavioral findings suggest that in order to start learning object names, infants and their parents consistently select a set of a few objects to play with during a set amount of time. What emerges is a frequency distribution of a few toys that approximates a Zipfian frequency distribution of objects for learning. We find that a machine learning model trained with a Zipf-like distribution of these object images outperformed the model trained with a uniform distribution. Overall, these findings suggest that to overcome referential ambiguity in clutter, infants may be selecting just a few toys allowing them to learn many distributional cues about a few name-object pairs.

Keywords: infancy; early word learning; machine learning; Zipfian distribution.

Introduction

The natural environment is visually cluttered with multiple namable objects in view (Clerkin, 2017). To learn their first object names, infants must link a heard object name to the referent object (Bloom, 2000). But for any heard object name, from the infant's perspective, there are multiple potential referents in view. This referential ambiguity has defined a major theoretical problem to be solved in early word learning (Quine, 1960). Despite a sea of clutter, infants already know the names of many objects by the time of their first birthday. We know this because they look to the named objects in laboratory tests (Bergelson, 2012; Swingley & Aslin, 2000) and because they begin to say object names in the contexts of those objects (Fenson et al, 1994). How does this work? The current paper integrates behavioral and modeling frameworks to explore how infants learn object names despite the referential ambiguity in their natural learning environments.

One explanation for solving referential ambiguity is the distributional cues in the language and visual input (Aslin, 2017). According to this explanation, infants track the frequencies of word-object co-occurrences to aggregate the most likely referent (Smith, Smith, & Blythe, 2011; Kachergis, Yu, & Shiffrin, 2017). A large collection of laboratory paradigms has demonstrated that infants can rapidly learn from distributional cues of visual and auditory input (e.g., Cartwright & Brent, 1997; Mintz, 2003; Mintz, Newport, & Bever, 2002; Reeder, Newport, & Aslin, 2013). However, it is still unclear how learning from distributional cues of words and objects in laboratory settings transfers to the distributional cues in the natural environment. Laboratory paradigms are typically highly controlled, presenting uniform word-object frequencies (Aslin, Saffran, & Newport, 1998; Kurumada et al., 2013). In contrast, for natural languages, word frequencies are known to follow a Zipfian distribution, in which a small number of words occur very frequently (e.g. boy, car), while many words occur rarely (Zipf, 1965). These so-called Zipfian distributions, are universal across human languages (Zipf, 1949; Piantadosi, 2014), including nouns and all words in infant-directed speech (Hendrickson & Perfors, 2018). Furthermore, recent studies show that even the distribution of objects in infants' natural visual environments follow a Zipfian distribution, where a few objects appear highly frequently and most objects are rare (Clerkin et al., 2017).

Nevertheless, the sensitivity of infants and adults to distributional cues highlights an intriguing, but as of yet untested, benefit for learning from Zipfian distributions. Theoretically, learning from a Zipfian distribution should be more difficult than a uniform distribution, as there is not enough information in a Zipfian distribution to link the referents for words that occur rarely (Blythe et al., 2010; Vogt, 2012; Reisenauer et al., 2013; Blythe et al., 2016). However, a recent adult study demonstrated that adults learn word-object links more easily from Zipfian distributions than from uniform distributions of word-object occurrences (Hendrickson & Perfors, 2018). Those results suggest that Zipfian distributions improve adults' learning by providing more statistical cues for the highly-frequent words, which in turn reduces the referential uncertainty associated with the unknown rare words. Yet, how infants learn from Zipfian distributions is still unknown.

The approach in this paper is that the natural training data for learning new object names are generated by the behaviors of the learner from the mature social partner who provides the

name. Here we demonstrate that during infant-parent interactions, objects being handled and named by the parent create Zipfian frequency distributions, in which very few events occur very frequently, forming a very small set for learning.

Recent studies of infant naturalistic environments suggest that Zipfian distributions provide a balance between *consistency* of a few high-frequent events with *diversity* of rare events (Clerkin et al., 2017; Smith & Slone, 2018; Montag, Jones, & Smith, 2017). Here we hypothesized that the parent and infant consistently select to play with a few objects, generating a training data set balanced with rare exploration of diverse objects. In other words, parents' and infants' selective and exploratory behaviors naturally generate name-object experiences that form Zipfian distributions, which is hypothesized to reduce ambiguity and optimize learning.

In this study we demonstrate infant naturalistic learning, while infants and parents were engaging with objects in a cluttered and an unstructured environment. The play sessions were recorded from the infant's egocentric perspective. From these visual experiences we report the frequency distributions of the objects with which infants and parents engaged during toy play. We subsequently applied a machine learning model to evaluate the structure of the visual "training data" produced from these play experiences. The model was tested for detecting the play objects with a training dataset of uniform distributions compared to Zipf-like distributions of infants' egocentric object views.

Behavioral Methods

To evaluate how infants learn early object names in a naturalistic environment, we conducted a toyplay experiment allowing infant-parent dyads to freely engage with object toys.

Participants

The final sample included 16 infant-parent dyads with 12 month-old infants (8 female) ranging from 12.2 to 12.5 months (*M*=12.3, *SD*=1.12) were included in the final sample.

Stimuli and Experimental Setup

Parents and their infants were invited to play in a naturalistic setting for a duration of approximately 10 minutes. Parents and infants both sat on a carpeted floor in a playroom environment. To create an unstructured environment, a random assortment of 33 toys were randomly distributed on the floor (see Figure 1). The toy objects' themes were not related in any particular way; thus, any selection and exploration of objects emerged naturally from infant and parent behaviors. The same toys were used in each session. The instructions were to play freely as they normally would at home.

Figure 1: (A) Stimuli set. (B) Experimental setup (left to right) infant wearing a Looxcie camera, infant ego-centric views from camera during toyplay.

Egocentric View

To collect infant egocentric view, we used a commercially available, lightweight (22 g) wearable camera (Looxcie). The camera was secured to a hat that was custom fit to the infant so that when the hat was securely placed on the infant's head, the lens was centered above the nose and did not move (see Figure 1).

The head camera captured the scene in front of the viewer but did not provide direct gaze information, which in principle could be outside of the head camera image (Smith et al., 2015). However, head mounted eye-tracking studies have demonstrated that under active viewing conditions, human observers, including infants, typically turn both heads and eyes in the same direction and align heads and eyes within 500 ms of a directional shift to maintain head and eye alignment when sustaining attention (Yoshida & Smith, 2008; Smith, Yu, & Pereira, 2011). Therefore, it can be expected that a high proportion of gaze during active viewing is highly concentrated in the center of the head camera image (Yoshida & Smith, 2008).

Data Processing

The raw videos were coded using Datavyu by sampling frames at 0.2Hz (1 frame every 5 sec; 2,008 frames total). To describe the dyadic behaviors of engagement with objects, the corpus of frames was coded for (1) objects in view, (2) objects handled by the infant or parent and (3) objects named by the parent. The *objects in view* were defined as the number of objects in the field of view from the infant's perspective. The *objects handled* were coded for both the parents and infant and defined as any hand contact with objects. Finally, parents' speech was manually transcribed and then annotated for *objects naming* when objects were named explicitly (*N*=580).

We measured the statistics for objects in parents' and infants' hands and parent naming events over the entire 10-minute period. Since there were individual differences among the infants in terms of the objects they played with, we constructed rank-ordered frequency histograms (see

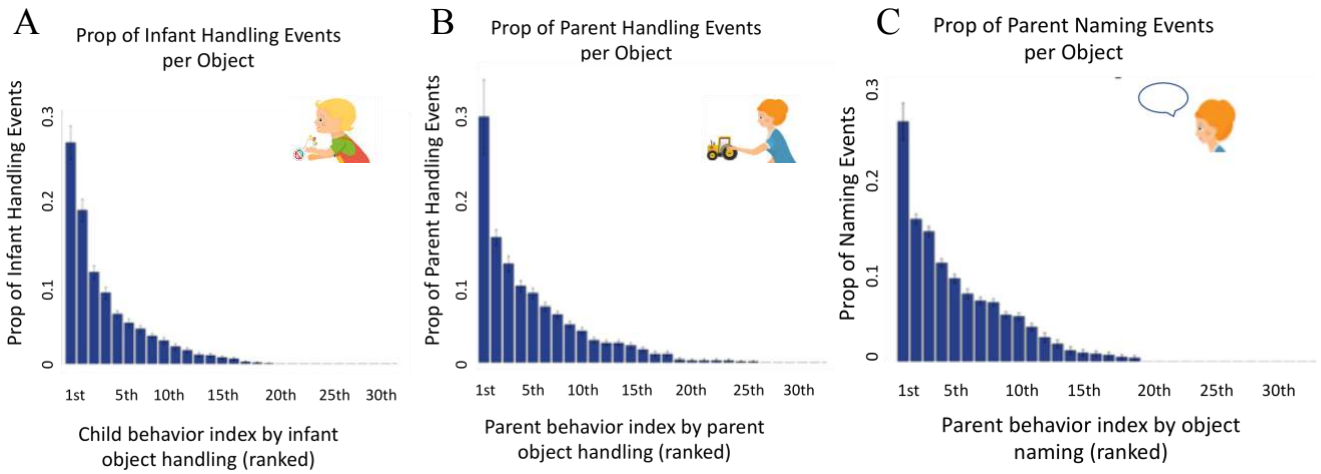


Figure 2: Ranked order histograms of the objects infant and parent handled and named (objects' ranks across these histograms are not necessarily the same). (A) histogram of infant object handling, showing the proportion of instances infants handled each object. (B) histogram of parent object handling, showing the proportion of instances parents handled each object. (C) histogram of parent object naming, showing the proportion of instances parents named each object. Error bars indicate 95% confidence intervals.

Figure 2). The rank-ordered frequency was measured independently for each infant and then combined since the most frequent object in view, handled or named differed for different infants. The objects are annotated by their rank-ordered frequency ranging from the 1st most frequent to least 33rd frequent object.

Behavioral Results

The experimental setup of dyadic play was visually cluttered as there were on average 13 objects in view every frame (*Min*=1, *Max*=30, *M*=13.18, *SD*=6.16). Figure 2 shows ranked order histograms of infant and parent object handling proportions, as well as parent naming proportions. The histograms display a Zipf-like pattern which is indicative of behavioral selectivity of objects in the scene. Specifically, these zipf-like distributions follow an approximate power-law, in which a small set of objects are handled and named very frequently and most objects are rare. Six objects (*Min*=9, *Max*=19, *M*=13.70, *SD*=3.13) account for over 80% of the total proportion of infant's object handling events. Six objects (*Min*=9, *Max*=24, *M*=15.5, *SD*=3.8) account for over 80% of the total proportion of parents' object handling events. Eight objects (*Min*=10, *Max*=48, *M*=24.20, *SD*=10.23) account for over 80% of the total proportion of parents' object naming events. These Zipf-like distributions likely reflect a balance between the parents' and infants' stability and exploration of objects, which may benefit object learning.

Modeling Methods

A machine learning model was used to test whether the distributional properties of infants' visual object experience impacted learning. In particular we wanted to understand the learning mechanism by which infants learn new object names in clutter environments.

The data

The collected corpus of 2,008 infants' egocentric views were used to construct two different toy object training sets, as detailed in Table 1. Six of the objects were selected for our machine learning study: baby doll, ball, chair, bucket, boat, and duck (see Figure 1). One of the two training datasets had a uniform frequency distribution of object images, and the other was with a Zipf-like frequency distribution. There were a few reasons for only using six specific objects for the modeling framework. First, from the raw corpus of images, only 1,200 images included at least 1 of the 6 objects for detection in the scene. Second, Bounding boxes indicating the objects' location and label were annotated for the set of 6 objects intended for detection. Note that some images were removed from the corpus due to low image quality such as high blur. The corpus was augmented by 180-degree rotation and horizontally flipped, yielding a final corpus of about 3,000 images: 2400 split into training and validation and 600 for testing.

Figure 3: Example of cluttered training images, including multiple objects for detection, labeled and annotated with a bounding box.

Each training dataset (the uniform and Zipf-like) was formed by a subset of the 2400 images for training. Due to the nature of the cluttered scenes, many images included more than 1 detectable object as seen in Figure 3. These images that included multiple objects for detection were counted toward the frequencies of more than one object when forming the datasets of uniform and Zipf-like frequency distributions (see Table 1). The final training data sets included 2,154 images each. In the *uniform* dataset, each object was present in 400 images. The *Zipfian* dataset included high frequency and low frequency images of objects. In the *Zipfian* dataset the baby doll had the highest frequency (1000 images), and the duck was the rarest (100 images).

Table 1: Distribution of object images among the uniform and right-skewed datasets for training

Images Per Object	Uniform	Zipf-like
Baby Doll	400	1000
Ball	400	600
Chair	400	320
Bucket	400	240
Boat	400	140
Duck	400	100
Total	2,154	2,154

Model Parameters

The applied machine learning model was the Faster R-CNN, Region-based Convolutional Neural Network (Ren, Girshick & Sun, 2015), a well-known, state-of-the-art machine learning model for object detection. The model is essentially a network composed of three main components: a feature extractor, a region proposal network (RPN), and a classifier. First, for the feature extraction part, we adapted a pretrained CNN VGG16 on the ImageNet data set which includes approximately 1.2 million images (Russakovsky et al., 2015). The model has 16 layers and classifies images into 1000 object categories (e.g. keyboard, mouse, coffee mug, pencil). The input images (size 224X224) are inputted into the VGG16 network. The network evaluates the distinctive visual features for the whole image, which allows us to detect multiple objects in each image. Second, after feature extraction the regions are proposed, therefore only running one CNN over the entire image instead of multiple CNN's for each proposed region. Finally, a single softmax layer, outputs the class probabilities directly for each region. The last fully connected layer and classification layer were adjusted for the number of classes in the data set applied in this framework (N classes= 7, including 'background').

Training

The Fast-RCNNs were trained with two different datasets that varied in the frequency distributions of object images: the Zipf-like and uniform distributions. The network was trained for 1000 epochs (iterations).

Modeling Results

To determine whether infants' selective behavior benefits learning, we applied a machine learning model trained with Infants' egocentric views. We compared the model's performance of object recognition when trained with a Zipf-like vs. a uniform distribution of infants' egocentric views as seen in Table 2. Overall, the model trained with the Zipf-like dataset had a significantly higher (t=-3.35, p<0.05) mean average precision (mAP=40%) compared to the model trained with the uniform dataset (mAP=23%). This pattern of results suggests that a Zipf-like distribution of data yields higher accuracy and benefits learning.

To further evaluate whether the Zipf-like distribution of objects in infants' egocentric views reduce ambiguity we evaluated the average precision of each object (see Table 2). For the baby doll and the ball there were more images in the Zipf-like dataset relative to the uniform distribution and there was accordingly a higher average precision for these objects in the Zipf-like trained model (AP=56% and 48%, respectively). The chair had a similar number of images in both datasets and had a similar average precision in the Zipf-like (AP=23%) and uniform datasets (AP=21%). For the rest of the objects (the bucket, boat and the duck), there were less images in the Zipf-like dataset, yet a higher average precision in the Zipf-like trained model (AP= 29% and 38% and 43%, respectively) compared with the model trained on the uniform distribution. These patterns of results, where there is higher precision despite less training images suggests that information has been shared among objects reducing likely competing objects and reducing ambiguity.

Discussion

In the real world, the sea of visual clutter provides multiple competing referents for every heard object name. This paper explored how infants solve this referent ambiguity in a cluttered environment to learn first object names. Here we presented a behavioral study of infants and their parents playing freely with objects and applied a learning model to explore the 'behind the scene' learning machinery.

To observe how infants learn object names in a cluttered environment, we recorded the play from infants' egocentric view. In order to describe experiences relevant for object name learning, we reported the frequency distributions of the

Object AP	Baby Doll	Ball	Chair	Bucket	Boat	Duck	mAP
Uniform	0.25	0.23	0.21	0.19	0.28	0.24	0.23
Zipf-Like	0.56	0.48	0.25	0.29	0.38	0.43	0.40

Table 2: Model test results: average precision per object for the uniform and Zipf-like datasets

objects infants and parents handled, as well as the objects parents named. We found that the frequency distributions of objects handled and named approximated a right-skewed Zipf-like distribution with few highly frequent objects along with many low frequency objects. This finding suggests that in a cluttered environment, infants and parents consistently select a set of a few objects for learning and rarely explored the other objects. The consistent object handling and naming behaviors during early word learning offers repetition, a key component for learning (Hintzman & Block, 1971, Vlach, 2014).

The infant-parent dyads consistently created datasets that were highly selective and focused on just a few objects. These dynamic patterns of selection may be due to the influence of other systems such as human memory or attention which decays in a power-law pattern (e.g., Wixted & Ebbesen, 1991; Wixted, 2004; Baronchelli, Ferrer-i-Cancho, Pastor-Satorras, Chater, & Christiansen, 2013). These non-uniform distributions have been shown as optimal conditions for adults and may help solve learning problems across many domains (Schuler, Reeder, Newport & Aslin, 2017; Hendrickson & Perfors, 2018; Caron & Vincent, 2002; Salakhutdinov, Torralba, & Tenenbaum, 2011).

As we could not directly observe infants' learning machinery, we applied a machine learning model to explore how infants may be learning from a Zipfian distribution. The application of the model was weaved with the behavioral study by using infants' egocentric object views from the behavioral study of play as the training images for the learning model. The learning machinery from Zipfian distributions was evaluated by comparing a training apparatus of a Zipf-like and a uniform frequency distribution of object images. The testing demonstrated that the training using a Zipf-like distribution yielded higher accuracy than a uniform distribution. Interestingly, the testing also demonstrated that low frequency objects were learned at higher rates when trained in the Zipf-like distribution.

The Zipf-like model's patterns of results were consistent with machine learning and adult studies of Zipfian learning (Schuler, Reeder, Newport & Aslin, 2017; Hendrickson & Perfors, 2018; Caron & Vincent, 2002; Salakhutdinov, Torralba, & Tenenbaum, 2011). These studies suggested that the learned features of highly frequent items are shared with the low frequency items to reduce referent ambiguity. For example, when learning to recognize a rare vehicle such as 'bus', the exemplar shares features of wheels and window shields from an already learned 'car', a highly frequent vehicle. It has also been suggested that low frequency items such as 'napkin' may benefit from co-occurrences with high frequency objects such as 'bowl'. These model's results coincide with infants laboratory studies demonstrating that infant early word learning is tuned to statistical cues of word-object co-occurrences. These findings also suggest that infants may be able to learn object names not only from uniform distributions of word-object occurrences but also from a Zipfian distributions.

Beyond previous early word learning studies, the findings in this paper suggest that infants solve referential ambiguity in a sea of clutter by consistently selecting a few objects and rarely exploring a large subset of objects. This behavior may benefit learning and reduce ambiguity in clutter by allowing to learn a lot of statistical cues about a few objects. Finally, this paper offers a methodological framework of incorporating behavioral paradigms with computational modeling to stretch our understanding of cognition.

Acknowledgements

This research was supported in part by NSF grant BCS-1523982, NICHD T32HD007475-22 and F32HD093280, and by Indiana University through the Emerging Area of Research Initiative – Learning: Brains, Machines, and Children.

References

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. Psychological Science, 9, 321–324. doi:10.1111/1467-9280.00063

Aslin, R. N. (2017). Statistical learning: a powerful mechanism that operates by mere exposure. *Wiley Interdisciplinary Reviews: Cognitive Science, 8*(1-2), e1373.

Baronchelli, A., Ferrer-i-Cancho, R., Pastor-Satorras, R., Chater, N., & Christiansen, M. H. (2013). Networks

in cognitive science. *Trends in Cognitive Sciences*, *13*(7), 348–360.

Bergelson, E., & Swingley, D. (2012). At 6–9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences*, *109*(9), 3253-3258.

Blythe, R., Smith, A., & Smith, K. (2016). Word learning under infinite uncertainty. *Cognition*, *151*, 18-27.

Blythe, R., Smith, K., & Smith, A. (2010). Learning times for large lexicons through cross-situational learning. *Cognitive Science*, *34*(4), 620–642.

Bloom, P. (2000). *How children learn the meanings of words*. Cambridge, MA: MIT Press.

Clerkin, E. M., Hart, E., Rehg, J. M., Yu, C., & Smith, L. B. (2017). Real-world visual statistics and infants' first-learned object names. *Phil. Trans. R. Soc. B*, *372*(1711), 20160055.

Caron, Y., Makris, P., & Vincent, N. (2002). *A method for detecting artificial objects in natural environments.* Paper presented at the Pattern Recognition, 2002. Proceedings. 16th International Conference on.

Cartwright, T. A., & Brent, M. R. (1997). Syntactic categorization in early language acquisition: Formalizing the role of distributional analysis. Cognition, 63, 121–170. doi:10.1016/S0010-0277(96)00793-7

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., Pethick, S. J., ... & Stiles, J. (1994). Variability in early

communicative development. *Monographs of the society for research in child development*, i-185.

Hendrickson, A., & Perfors, A. (2018). Cross-situational learning in a Zipfian environment.

Hintzman, D. L., & Block, R. A. (1971). Repetition and memory: Evidence for a multiple-trace hypothesis. *Journal of Experimental Psychology*, *88*(3), 297.

Kachergis, G., Yu, C., & Shiffrin, R. M. (2017). A bootstrapping model of frequency and context effects in word learning. *Cognitive science*, *41*(3), 590-622.

Kurumada, C., Meylan, S., & Frank, M. C. (2013). Zipfian frequency distributions facilitate word segmenta-

tion in context. *Cognition*, *127*, 439–453.

Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. Cognition, 90, 91– 117. doi:10.1016/S0010-0277(03)00140-9

Mintz, T. H., Newport, E. L., & Bever, T. G. (2002). The distributional structure of grammatical categories in speech to young children. Cognitive Science, 26, 393–424. doi:10.1207/s15516709cog2604_1

Montag, J. L., Jones, M. N., & Smith, L. B. (2018). Quantity and diversity: Simulating early word learning environments. *Cognitive science*, *42*, 375-412.

Piantadosi, S. (2014). Zipf's word frequency law in natural language: A critical review and future directions.

Psychonomic Bulletin and Review, *21*, 1112–1130.

Quine, W. (1960). *Word and object*. Cambridge, MA: MIT Press.

Reisenauer, R., Smith, K., & Blythe, R. (2013). Stochastic dynamics of lexicon learning in an uncertain and

nonuniform world. *Physics Review Letters*, *110*(258701).

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems* (pp. 91-99).

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... & Berg, A. C. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*(3), 211-252.

Salakhutdinov, R., Torralba, A., & Tenenbaum, J. (2011, June). Learning to share visual appearance for multiclass object detection. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on* (pp. 1481-1488). IEEE.

Schuler, K. D., Reeder, P. A., Newport, E. L., & Aslin, R. N. (2017). The effect of Zipfian frequency variations on category formation in adult artificial language learning. *Language Learning and Development*, *13*(4), 357-374.

Smith, L. B., Yu, C., & Pereira, A. F. (2011). Not your mother's view: The dynamics of toddler visual experience. *Developmental science*, *14*(1), 9-17.

Smith, L. B., & Slone, L. K. (2017). A developmental approach to machine learning?. *Frontiers in psychology*, *8*, 2124.

Smith, K., Smith, A. D., & Blythe, R. A. (2011). Cross-situational learning: An experimental study of word-learning mechanisms. *Cognitive Science*, *35*(3), 480-498.

Smith, L. B., Yu, C., Yoshida, H., & Fausey, C. M. (2015). Contributions of head-mounted cameras to studying the visual environments of infants and young children. *Journal of Cognition and Development*, *16*(3), 407-419.

Swingley, D., & Aslin, R. N. (2000). Spoken word recognition and lexical representation in very young children. *Cognition*, *76*(2), 147-166.

Vlach, H. A. (2014). The spacing effect in children's generalization of knowledge: allowing children time to forget promotes their ability to learn. *Child Development Perspectives*, *8*(3), 163-168.

Vogt, P. (2012). Exploring the robustness of cross-situational learning under Zipfian distributions. *Cognitive*

Science, *36*, 726–739.

Wixted, J. (2004). The psychology and neuroscience of forgetting. *Annual Review of Psychology*, *55*, 235-269.

Wixted, J., & Ebbesen, E. (1991). On the form of forgetting. *Psychological Science*, *2*, 409–415.

Yoshida, H., & Smith, L. B. (2008). What's in view for toddlers? Using a head camera to study visual experience. *Infancy*, *13*(3), 229-248.

Zipf, G. (1949). *Human behavior and the principle of least effort*. New York, NY: Addison-Wesley.

Zipf, G. (1965). Human behavior and the principle of least effort: An introduction to human ecology. New York, NY: Hafner.

Do people use gestures differently to disambiguate the meanings of Japanese compounds?

Kei Kashiwadate (18rmd06@ms.dendai.ac.jp)
Graduate School of Tokyo Denki University
Ishizaka, Hatoyama-machi, Hiki-gun, Saitama 350-0394
Tetsuya Yasuda (t-yasuda@mail.dendai.ac.jp)
Tokyo Denki University
Harumi Kobayashi* (h-koba@mail.dendai.ac.jp)
Graduate School of Tokyo Denki University

Abstract

Spoken language often includes ambiguity in meaning. Compounds such as "green teacup" can be interpreted with two different meanings: "green colored teacup" and "cup for green tea." We can assume there are two different underlying syntactic structures. Phonetic aspects have been studied in the disambiguation process of such ambiguous phrases, but the roles of nonlinguistic information such as gestures have not been explored yet. We investigated whether people use gestures differently when they were asked to describe the meanings of Japanese compounds that can be interpreted as two different meanings. We found that the偏ming of gestures in relation to the target words of accompanying speech was different between right branching compounds and left branching compounds. Gestures seem to be used to suggest upcoming two words (adjective and noun) as a unit in branching. Gestures can be a useful means to disambiguate the meanings of compounds.

Keywords: Gestures; Disambiguation; Branching; Compounds

Introduction

A phrase consists of concatenation of words that are produced sequentially. It is known that compounds can be interpreted to have multiple meanings. For example, "green teacup" can be interpreted either as a green-colored teacup or as a cup of green tea. A phrase structure with the meaning "teacup for green tea" can be classified as left branching (LB); that is, "green" and "tea" are first grouped to "green tea" and then together play an adjective role in "cup." The phrase structure with the meaning of "green-colored teacup" can be classified as right branching (RB); that is, "tea" and "cup" are first grouped together, and the word "green" plays an adjective role in "teacup" (Figure 1). Because speech is produced sequentially, the surface structure does not have enough information to show the underlying syntactic structure. Therefore, phrases inevitably have ambiguity in meaning. Nevertheless, people usually seem to have little difficulty in discerning the meanings of such phrases. Humans may use some disambiguation cues to resolve ambiguities in such ambiguous structures.

Previous studies have focused on prosodic cues as a means of disambiguation (Ito, Arai, & Hirose, 2015; Hirose & Mazuka, 2015; Venditti, 1994). Native Japanese speakers prefer LB interpretation over RB interpretation for slightly simpler Japanese compound constructions and to make RB interpretation more accessible. A clear prosodic demarcation that raises the pitch range of the second word has been found effective (Ito, et al, 2015; Hirose & Mazuka, 2015; Venditti, 1994). However, the exact disambiguation cues are still unknown. In the present study, we focused on nonverbal cues, in particular, gestures that have not been examined yet in the disambiguation mechanism of syntactic structures.

Figure 1: Two different syntactic structures, left branching (left) and right branching (right), in the compound "green teacup"

Gestures play an important role in communication. Humans simultaneously use gestures and language to convey information to others. Gestures are usually produced slightly earlier than associated speech, and this can make the hearer anticipate the information in the upcoming speech (MacNeill, 1987). Iconic gestures (e.g., depicting objects by movement trajectories) and pointing gestures can reflect aspects of the speaker's nonlinguistic spatial representations (Majit, Bowerman, Kita, Haun & Levinson, 2004). Gestures can spontaneously accompany speech and make communication smooth (Kita & Saito, 2002). Representational gestures (i.e., iconic and deictic gestures) can express spatial contents or metaphorically express temporal concepts (Kita, 2009). Additionally, gestures express information even when it is difficult to express in language (Alibali, Evans, Hostetter, Ryan & Mainela-Arnold, 2009). Various functions are known about gestures, but the topic of whether gestures can contribute disambiguation mechanisms of syntactic structures has been largely unexplored.

Previous studies on interpretation of compounds of possibly different branching structures have showed that people prefer a certain branching over other branching when two (or more than two) different branchings are possible (e.g., Ito et al., 2015). In our study (accepted) on Japanese participants' interpretation of Adjective_1 + Noun_1 + Adjective_2 + Noun_2 compounds, we found that some adjectives are interpreted more dominantly than other adjectives for certain nouns. For example, "long" can be a

typical adjective for "tail," but an atypical one for "cat." It is possible that "long cat" may mean that the cat's body is long, but this sounds somewhat strange. The typicality of the adjective + noun combination may affect the predominant interpretation.

The present study investigated whether gestures are used as a clue to resolve ambiguities in branching structures of Japanese compounds. We examined whether the productions of participants' gestures differ in the case of compounds of either LB or RB. Our prediction was that participants might make gestures with different timings when verbally producing the compounds of LB or RB.

We also examined whether people may exaggerate their gestures by taking more time for relevant gestures when they are aware that more than one interpretation is possible for ambiguous compounds. To examine this exaggeration aspect, we compared a one-picture condition (Alone condition) that denoted either LB or RB meaning and a two-picture condition (side-by-side condition) that denoted both LB and RB meanings side by side so that people could more easily notice the different interpretations.

Further, we also examined another source for possible exaggeration, the combination of nouns and adjectives. We decided to compare the two adjectives "big" and "long." The adjective "big" can be typically applied to "cat" (big cat) or "tail" (big tail), whereas the adjective "long" can be typically applied to only "tail" (long tail) and not "cat" (??long cat). We expected the participants to feel less ambiguity when they interpreted phrases with "long" rather than phrases with "big," so the participants would take less time for "long" condition and the timing of the gestures may also be different between the "big" and "long" conditions.

Method

Participants

Sixteen Japanese monolingual students who spoke Japanese as a first language participated (M age = 21.6, SD = 1.32; 1 female). This study was approved by the ethics committee of the participants' university.

Stimuli

A total of 32 slides were prepared using Adobe Illustrator. Sixteen slides were prepared for the side-by-side picture condition, and another 16 slides were prepared for the alone picture condition. In the side-by-side condition, two comparable objects were drawn side by side in each slide (Figure 2). One was the object (animal) according to LB interpretation, and the other was the object (animal) according to RB interpretation. The slide in the side-by-side condition consisted of one target phrase on the top, two illustrations (i.e., LB and RB interpretations) in the middle, and explanatory notes for each illustration on the bottom (Figure 3). A compound had two possible interpretations: an LB interpretation and RB interpretation. For example, [Kuroi] [Sippono] [Ookina] [Neko] in Japanese (i.e., [Black] [Tailed] [Big] [Cat]) can be interpreted either as "a big cat with black tail" (LB) or as "a black cat with a big tail" (RB)[1]. The difference of meaning can be explained as follows: in the case of LB, the "tailed" branch connects to "black" branch. In the case of RB, the "tailed" branch connects to the "big cat" branch (Figure 4). The position of the LB object and RB object in each slide was counterbalanced. In the alone condition, there was only one object of either LB interpretation or RB interpretation on each slide. There were eight side-by-side slides and eight alone slides. The slide in the alone condition consisted of one target phrase on the top, one illustration in the middle, and an explanatory note for the illustration on the bottom.

In the stimulus compounds, 16 slides included the adjective "big," and another 16 slides included the adjective "long."

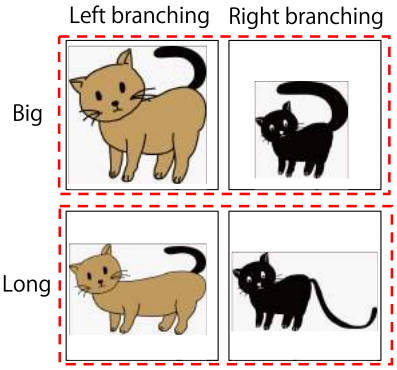


Figure 2: An example of object sets

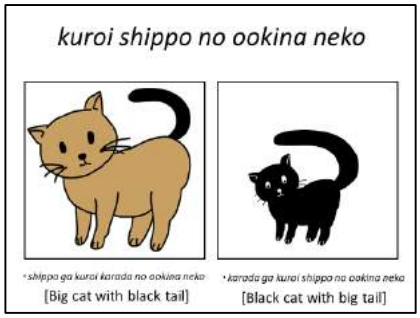


Figure 3: An example of side-by-side slide

1

Japanese	|*kuroi*	|*shippo*	|*no*	|*ookina*	|*neko*
Word class	|[adjective]	|[noun]	|[particle]	|[adjective]	|[noun]
English	|black	|tail		| big	|cat

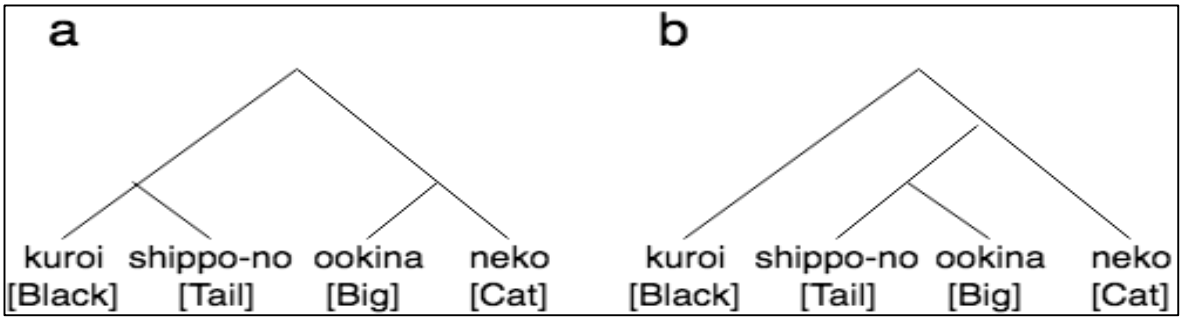


Figure 4: Branching structure. (a): LB interpretation; (b): RB interpretation

Procedure

The participants were divided into two experimental groups: alone slide group and side-by-side slide group. Each group looked at eight slides on a computer monitor.

After filling in the consent form, the participants were seated in front of a monitor (Figure 5).

The participants took part in one practice trial to be familiarized with the task, and then the experiment was started. On each trial, a fixation cross appeared in the center of the monitor. After the cross was fixated for one second, a slide appeared for 10 seconds. Then, only the top phrase was displayed. At that moment, participants were asked to make gestures to describe the presented picture while verbally producing the phrase (Figure 6).

Participants' gestures and utterances participants were recorded by a video recorder (Microsoft LifeCam). In the side-by-side slides, one of the two objects was presented with a surrounding red frame, and the participants were asked to describe the indicated object.

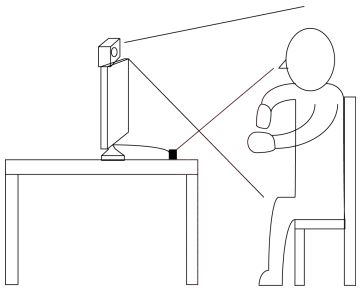


Figure 5: Experimental layout

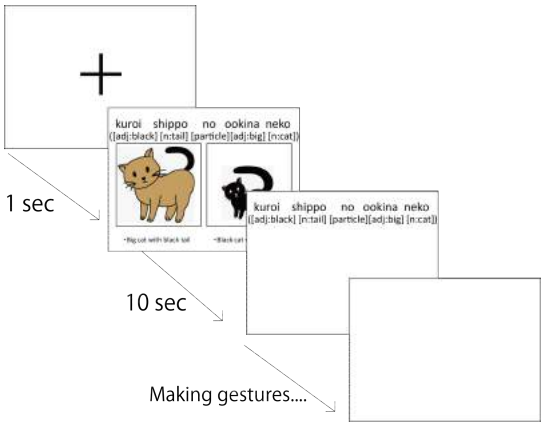


Figure 6: Flowchart of slides

Coding

We annotated utterances and gestures using ELAN 2017 (Version 5.1). The timing and duration of each gestures were recorded (Figure 7). We used the coding scheme modified version of Kita, Gijn, and Hulst's (2014) gesture coding. In this scheme, a gesture consists of a preparatory movement, followed by a stroke, and then finally, a finishing movement. We recorded the time of onset and end of each gesture stroke to determine the timing and duration of each gesture. The third word was the critical adjective "big" or "long" that was grouped with either the second word (e.g., "tail") or the final word (e.g., "cat").

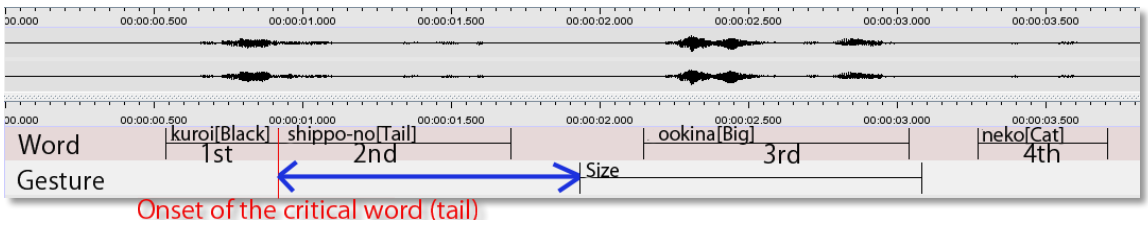


Figure 7: An example of the annotation using ELAN

Results

Timing of the first gesture

To find out whether the onset of gestures differed between LB and RB, between the alone slide and side-by-side slide, and between the adjectives "big" and "long," the difference in time (seconds) between the onset of the critical word "shippo (tail)" on the branching point and the onset of the first gesture in each slide in the participant's performance was measured and taken as a dependent measure. A 2 (Slide: Alone, Side-by-side) × 2 (Adjective: Big, Long) × 2 (Branching: LB, RB) three-way ANOVA was performed on the measure. Slide (Alone, Side-by-side) was the between-participants variable. Adjective (Big, Long) and branching (LB, RB) were the within-participants variables.

There was a marginally significant effect of branching ($F(1,14) = 3.6925$, $p = .075$, $\eta^2 = 0.0956$). This meant that the onset of the first gesture was earlier in RB (M= -0.55) than in LB (M= -0.07). Furthermore, there was a significant Slide × Adjective × Branching interaction ($F(1,14) = 6.3258$, $p < .05$). To explore the significant Slide × Adjective × Branching interaction, the simple interaction effects of Slide, Adjective, and Branching within each condition were calculated (Figure 8).

The simple main effect of branching in the "long" condition was marginally significant ($F(1,14) = 3.1574$, p = .097, $\eta^2 = 0.0842$). This meant that when the third word was "long," the onset of the first gesture tended to be earlier in RB (M= -0.49) than in LB (M= -0.08).

There was a simple Slide × Branching interaction in the "long condition" ($F(1,14) = 4.9359$, $p < .01$). Simple-simple main effects of Slide and Branching within the "long" condition were calculated. There were simple-simple main effects of branching ($F(1,7) = 7.0215$, $p < .05$) and slide ($F(1,14) = 4.4872$, $p = .052$, $\eta^2 = 0.2427$). The simple-simple main effect of slide was marginally significant. It meant that when the third word was "long" and the slide was "alone," the onset of the first gesture was earlier in RB (M= -0.88) than in LB (M= 0.04). When the third word was "long" and branching was RB, the onset of the first gesture was more behind when slide was side by side (M= -0.11) than when slide was alone (M= -0.88).

Total duration of the gestures

Using the video recordings, we calculated the total duration of gestures (seconds) produced in each slide. We predicted that the duration of gestures was different between LB and RB.

A 2 (Slide: Alone, Side-by-side) × 2 (Adjective: Big, Long) × 2(Branching: LB, RB) three-way ANOVA was performed on the total duration of gestures.

There was a marginally significant effect of slide ($F(1,14) = 3.935$, $p = .067$, $\eta^2 = 0.0512$). The total duration time of gestures was longer when the slide was alone (M=2.16) than when the slide was side by side (M=1.82). There was also a significant Slide ×Adjective × Branching interaction ($F(1,14) = 15.9588$, $p < .01$). To explore the significant Slide × Adjective × Branching interaction, simple interaction effects of slide, adjective, and branching within each condition were calculated.

There was a simple main effect of Slide in the "big" condition ($F(1,14) = 5.6515$, $p < .05$) and adjective in the alone condition ($F(1,7) = 4.8931$, $p = .062$, $\eta^2 = 0.0171$). The simple main effect of the adjective in the alone condition was marginally significant. When the third word was "big," the total duration time of gestures was longer when the slide was alone (M= 2.27) than when the slide was side-by-side (M= 1.70). When the slide was alone, the total duration time of gestures was longer when the third word was "big" (M=2.27) than when the third word was "long" (M=2.06).

There were simple Adjective × Branching interactions in the alone condition ($F(1,7) = 6.9771$, $p < .05$, $\eta^2 = 0.0368$) and in the side-by-side condition ($F(1,7) = 9.1492$, $p < .05$, $\eta^2 = 0.1196$). The simple-simple main effects of adjective and branching within the alone condition and side-by-side condition were calculated. There were simple-simple main effects of the branching in the alone condition ($F(1,7) = 33.2153$, $p < .001$, $\eta^2 = 0.1386$) and in the side-by-side condition ($F(1,7) = 7.6180$, $p < .05$, $\eta^2 = 0.2495$). When the slide was alone and branching was LB, the total duration time of gestures was longer when the third word was "big" (M= 2.39) than when the third word was "long" (M= 1.89). When the slide was side by side and the third word was "long," the total duration time of the gestures was longer in LB (M= 2.32) than in RB (M= 1.55).

There was a simple Slide × Branching interaction in the "long" condition ($F(1,14) = 5.0550$, $p < .05$). The simple-simple main effect of slide and branching within the "long" condition was calculated. There was a simple-simple main effect of branching that was marginally significant ($F(1,14) = 4.0335$, $p < .05$). When the third word was "long," and branching was RB, the total duration time of the gestures was longer when the slide was alone (M= 2.23) than when the slide was side by side (M= 1.55).

There was a simple Slide × Adjective interaction in the LB condition ($F(1,14) = 7.8598$, $p < .05$). The simple-simple main effect of slide and adjective within the LB condition was calculated. There was a simple-simple main effect of slide ($F(1,14) = 5.9902$, $p < .05$). When branching was LB and the third word was "big," the total duration time of the gestures was longer when the slide was alone (M= 2.39) than when the slide was side by side (M = 1.60).

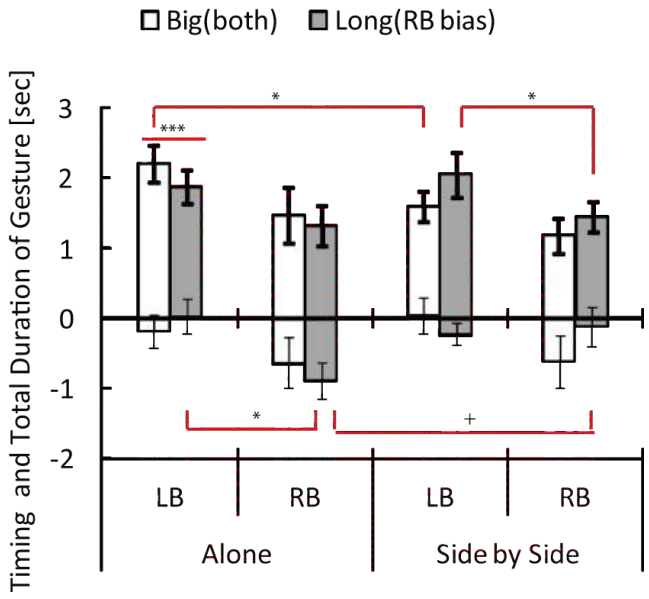


Figure 8: The start time point, the end time point, and the duration of the gestures in each condition. Zero denotes the starting time point of the target word "shippo (tail)" in accompanying speech. The thin error bars denote standard errors in the onset of gesture. The thick error bars denote standard errors in the duration of gesture.

Discussion

The most important finding was that the onset of the first gesture in relation to the critical word (the second word "tail") was earlier in the RB condition than in the alone slide. This means that the participants started a gesture earlier when they wanted to mean "a black cat with a big tail" (RB) than they wanted to mean "a big cat with a black tail" (LB). We interpreted this result as follows: in RB, the meaning "the tail is big" is important to convey, so participants started the gesture earlier to easily bring "tail" and "big" together as words that belonged to the same branch. Usually, the participants' gesture in this RB slide involved extending and moving their arm horizontally to describe a long big tail. Thus, the participants seemed to emphasize "big tail" by this gesture. In contrast, the gesture used in LB slide involved moving both hands widely up and down to describe the shape of the big cat. Thus, the participants seemed to emphasize "big cat." The point is that the critical adjective "big" must be grouped with either "tail" (RB) or "cat" (LB). As for the duration data, we found that the duration of the gestures was shorter in the side-by-side slides than in the alone slides. The participants might have thought that certain gestures would be enough for disambiguation when an alternative picture was explicitly presented. In the side-by-side/long/RB condition, participants might not have thought that disambiguation was necessary because "long cat" (LB interpretation) sounded too atypical compared with "long tail" (RB interpretation).

In conclusion, using Japanese compounds, we found that the timing and duration of gestures in relation to the target words in accompanying speech were different between RB compounds and LB compounds. Gestures seemed to be used to suggest two upcoming words in speech (adjective and noun) as a unit in branching. Gestures can be a useful means to disambiguate the meaning of compounds.

References

Alibali, M., Evans, J., Hostetter, A., Ryan, K. & Mainela-Arnold, E. (2009). Gesture – speech integration in narrative: Are children less redundant than adults? Gesture 9 (3), 290 – 311. doi: 10.1075/gest.9.3.02ali

Hirose, Y. & Mazuka, R. (2015). Predictive processing of novel compounds: Evidence from Japanese. Cognition, 136, 350–358. doi: 10.1016/j.cognition.2014.11.033

Ito, K., Arai, M., & Hirose, Y. (2015). The interpretation of phrase-medial prosodic prominence in Japanese: Is it sensitive to visual and discourse context? Language, Cognition and Neuroscience , 30, 167-196. doi: 10.1080/01690965.2013.864778

Kita, S. (2002). Jesuchaa: Kangaeru karada [Gesture: The body that thinks]. Tokyo: Kaneko Shobo.

Kita, S. (2009). Cross-cultural variation of speech accompanying gesture: A review. Language and Cognitive Processes, 24 (2), 145167. doi: 10.1080/01690960802586188

Kita, S., & Sito H. (2002). Jesuchaa・Koui・Imi [Gesture, Action, and Meaning]. Tokyo: Kyoritu Shuppan.

Kita, S., Gijn, I., & van der Hulst, H. (2014). The non-linguistic status of the symmetry condition in signed languages: Evidence from a comparison of signs and speech-accompanying representational gestures. Sign Language & Linguistics, 17, 209-232. doi: 10.1075/sll.17.2.04kit

MacNeill, D. (1987). Psycholinguistics: A New Approach. Inc. New York, New York, U.S.A: Harper & Row.

Majid, A., Bowerman, M., Kita, S., Haun, D. B. M. & Levinson, S. C. L. (2004). Can language restructure cognition? The case of space. Trends in Cognitive Sciences, 8, 108-114. doi: 10.1016/j.tics.2004.01.003

Venditti, J. J. (1994). The influence of syntax on prosodic structure in Japanese. In J. J. Venditti (Ed.), Papers from the Linguistics laboratory, Ohio State working papers in Linguistics, 44 (pp. 191–223). Ohio State University, Department of Linguistics, Columbus OH.

Acknowledgments

We would like to thank all our participants. We would like to thank Professor Sotaro Kita of the University of Warwick for his invaluable suggestions. We would also like to thank Editage (www.editage.jp) for English language editing. This study was supported by JSPS/MEXT KAKEN JP17H06382 (H.K.) and JP16K04318 (H.K.)

The Decision Science of Voting: Behavioral Evidence of Factors in Candidate Valuation

Janne Kauttonen (janne.kauttonen@haaga-helia.fi)
Haaga-Helia University of Applied Sciences, FI-00520 Helsinki, Finland
NeuroLab, Laurea University of Applied Sciences, Vanha maantie 9, 02650 Espoo, Finland

Jyrki Suomala (jyrki.suomala@laurea.fi)
NeuroLab, Laurea University of Applied Sciences, Vanha maantie 9, 02650 Espoo, Finland

Abstract

Despite decision science have increased our understanding of human decision-making in different contexts, voters' decision has been studied less from this point of view. Therefore, we investigated, how electorate- and candidate-related factors affect electorate's (N=1334) valuation to the Prime Minister candidates (N=11) on the multiparty democracy. Electorates valuated candidates individually and through pairwise candidate comparison. We collected the data by using anonymous questionnaire and sent it via mass emailing and social media. We applied linear mixed-effects and Bayesian network models to analyze the data. Electorate-related variable Valence and candidate-related variables Trustworthiness and Righteousness was found as the strongest main effects. The pairwise analysis comparison highlighted voters' personal characteristic. In particular, the interactions associated to valence, arousal and gender had high effect only in pairwise comparisons. Our results suggest that the pairwise comparisons - which is typical for elections, e.g., in USA - highlights the importance of emotional and gender-related factors.

Keywords: decision making; politics: valuation; voting; linear mixed-effects model; Bayesian networks

Introduction

Mainstream scholarly research assumes that voting decision is driven by rational preferences over policy proposals offered by political parties (Bischoff, Neuhaus, Trautner, & Weber, 2013; Hibbing, Smith, & Alford, 2014; Knutson, Wood, Spampinato, & Grafman, 2006). However, recent decision science studies have suggested that decision involves, besides explicit processes, psychological, social and cultural processes (Blouw, Solodkin, Thagard, & Eliasmith, 2016; Tymula & Glimcher, 2016). Whereas these studies have increased our understanding about human decisions in the marketing-, social- and risks contexts (Tymula & Glimcher, 2016), voters' decision have been less studied from decision science point of view. In addition, the multiparty democracies have been less studied compared to two-party democracies, especially USA (Walther, 2015). Therefore, we investigated, how electorate-related and Prime Minister candidate-related factors affect electorate's valuation. We chose eleven Prime Minister candidates (three females) on the multiparty democracy which were valuated using judgements of each candidates' directly and in pairwise comparison between candidates. We used linear mixed-effect models (Gelman & Hill, 2007) and Bayesian networks (Borgelt, Steinbrecher, & Kruse, 2009) to test statistical dependencies between candidate valuation and battery of ratings for features of both candidates and the rater himself/herself. Below we describe these dimensions more specifically.

Electorate-related Factors and Voting Decision

Political orientation has been studied with The Big-Five framework (Gosling, Rentfrow, & Swann, 2003; Hibbing, Smith, & Alford, 2014). Current study (Sibley, Osborne, & Duckitt, 2012) found, that political conservatism had negative correlation to Openness to Experience and positive correlation of Conscientiousness variables. In the same vein, Carney et al. (2008) showed that both low Openness to Experience and high Conscientiousness were associated with participants' self-reported conservatism. Thus, conservatives are more orderly, conventional, and better organized, whereas liberals are more open-minded, creative, curious, and novelty seeking (Carney, Jost, Gosling & Potter, 2008).

People with different political orientations have been found to resolve risk-decisions different ways (Hibbing, Smith, & Alford, 2014). Relative to liberals, politically conservative individuals are remembered which stimuli have bad value and pursued a more risk-avoidant strategy to the game. On the contrary, Liberals have greater tendency to explore, take more risk by choosing more unknown possibilities than Conservatives have (Shook & Fazio, 2009).These studies indicate that Conservatives show greater sensitivity to threatening stimuli in the environment than Liberals and have to tendencies to behave without risk-taking.

Prime Minister Candidate-related Factors and Voting Decision

In most of democracies the party leaders are also prime minister candidates and influential electoral force in election campaigns (Bean & Mughan, 1989). This candidate-centered politics (Garzia, 2011; Wattenberg, 1991) is accompanied by a great importance of leaders' personal characteristics in the eyes of voters. Thus, this study concentrates electorate's opinions about politicians' leadership skills and their opinions about the suitability of these candidates to the prime minister in the multi-party democracy country.

Previous studies have found that trustworthiness is one of the most important attribute for a political leader (Barisione, 2009; McAllister, 2000; Rule et al., 2010) as well as communication and collaboration skills (Barisione, 2009). In addition, voters want that political leader is one of them and works for their benefits (Garzia, 2011). Moreover, the voters give values for the fair leaders as well as "traditional" hard leadership skills like the capacity to make decisions (Bean & Mughan, 1989; Rule et al., 2010).

Second important dimension is electorate's emotional reactions to politicians' faces. Valence and arousal are two independent dimensions of emotion. When subjects anticipate pleasurable events, positive arousal increases, and when they anticipate unpleasant event, negative arousal increases. Studies have found that positive arousal has important effect on people's behavior towards the issues, which trigger these positive arousal (Knutson & Greer, 2008). Thus, we measured participants' valence and arousal as they imagined each candidate as a prime minister. We used above described individual and Prime Minister candidate–related factors as the framework for questionnaire. The faces of politicians have many learned symbolic and cultural meanings (Knutson et al., 2006). Therefore, we used politicians' faces as basic stimuli in order to clarify how much each politician's face can produce emotional reactions. Judgements of each candidates' direct valuation and pairwise candidate comparison were used as dependent variables.

Methods

Participants

Participants were recruited via mass emailing and social media to participate in the research. Total 1653 full responses were received over 4 months from which we removed 50 responses with missing/corrupted data, 9 duplicates (same subject), 176 responses with unrealistically fast response times (median time <7s per page) and 84 responses with zero of very low response variance. This resulted in 1334 responses (503 males) in final analysis. Filling the full questionnaire allowed participants to join lottery of 20 gift cards (each worth 25 euros).

Questionnaire Procedure

In the questionnaire, electorate-related variables included gender, age-group (between 18 and 60+) and eight self-spaced personal qualities. Variables dependable/self-disciplined and disorganized/careless measure characteristic conscientiousness, whereas variables open to new experiences and conventional/uncreative measure characteristic openness to experiences from Big Five personality scale. In addition, participants' opinions about his/her level of conservatism and level of liberalism were measured separately. Finally, participants' risk-sensitivity was measured by using social and investment risk variables from Weber et al. (2002) risk-attitude scale.

Prime Minister candidate-related variables included candidate's gender, candidate's familiarity and candidate's leadership skills. Leadership skills included variables trustworthiness, communication skills, fairness, tendency to work for nation, and decision skills. All candidates were established figures for their parties, i.e., the name and face were familiar to majority of people on national level. In addition, the emotional components valence and arousal were measured by showing candidates face with his name and party. Below of the face was two statements "She/He has just been elected Prime Minister of Nation X. What is the emotion (valence) of the choice in you? How intensive this emotion is (arousal)?"

In summary, the questionnaire contained four mandatory sections with following questions (variable labels in parenthesis):

1. Responder's background (x_{1-10}^{b}): Gender [binary], age [Likert scale; 1-7] and 8 personal qualities [1-7].
2. *Individual* candidate valuation (x_{1-8}^{r}): candidate gender [binary], 5 ratings, familiarity and suitability scores [1-7].
3. Emotion ($x_{1,2}^{e}$): Valence and arousal assuming the candidate was chosen as a Prime minister [1-7].
4. *Pairwise* candidate valuation (x^{c}): Preference between two randomly chosen candidates [-4-4].

Suitability score (x_{8}^{r}) and pairwise comparison score (x^{c}) were considered as the *responses* (valuations). Variables x_{1-7}^{r} encoded the *feature vector* of a candidate (1334 vectors in total, one from each subject). Candidate's order was randomized in all parts of the survey. In part 4, out of the pool of 55 possible candidate pairs, we presented randomly chosen 20 (randomized for each subject). In the analysis, genders (responders and candidates) were one-hot encoded using "female" label as the (arbitrary) reference level.

Data Analysis

Linear Mixed-effect Models First we fitted linear mixed-effects models (Gelman & Hill, 2007; Wu, 2009) to the data using Matlab (R2018a). Subject *id* and response date (*month*) were set as random effects of no interest. We fitted total of 4 models; two for the direct valuation and two for the pairwise valuation. Two of these models contained all variables (*full models*) and the remaining two (*reduced models*) did not include valence (x_{1}^{e}). Valence was highly correlated with valuations, hence it was deemed useful to repeat fitting without it. As there was no variation in background variables

(x_i^b) within a subject, those were entered into models through interactions.

For the individual valuation, using Wilkinson's notation (Wilkinson & Rogers, 1973), the formula of the full model was:

$$x_8^r \sim 1 + (x_1^r + \cdots + x_7^r + x_1^e + x_2^e) : (1 + x_1^b + \cdots + x_{10}^b) + (1|id + month),$$

where the total number of non-constant fixed terms (aka *predictors*) was 99 with 1338 random-effects intercepts. We used maximum likelihood criterion to fit parameters (Wu, 2009). The equation for the reduced model was similar, but without the valence term (88 fixed-effects terms).

For the pairwise valuations, the formulas were identical, but as the valuation was indirect, the features were transformed into differences, i.e., $x_i^r := x_{i,A}^r - x_{i,B}^r \;\; \forall\, i = 1, \ldots, 7$ (same for $x_{1,2}^e$ and x^c), where A and B correspond to two candidates in comparison. In this case the random-effects term id also covers the randomness related sampling of candidate pairs. Note that a linear model is invariant for the order of candidates in the differencing, i.e., flipping the order also flips the predictors and response. As a result, interpretation of the coefficients remains similar to direct valuation.

Statistical significance of linear models and their predictors were estimated using permutation testing scheme where responses were randomly shuffled while preserving subject-level grouping hierarchy. Original, un-shuffled t-values of each predictor were compared against distributions of 10.000 t-values obtained via permutation. False Discovery Rate (FDR; Benjamini & Hochberg, 1995) was applied to adjust for multiple comparisons over fixed-effects predictors. Overall model performance was measured with Mean Squared Error (MSE) compared against constant-only null models (with MSE_{null}) and those obtained via permutations.

Bayesian Network Models Next we dropped the assumption of the linearity and fitted Bayesian network probabilistic graphical model to the data (Borgelt, Steinbrecher, & Kruse, 2009; Nagarajan, Scutari, & Lèbre, 2013). For this, we used *bnlearn*[1] toolbox. Bayesian network models allow estimation of a full probability distribution via Directed Acyclic Graph (DAG) structure that represents relationships between data variables (nodes in the graph). Here we were mainly interested in the structure of DAGs and causal relationships between variables.

We adopted the approach of Scutari et. al (2017) with network bootstrapping and cross-validation to estimate DAGs and the quality of models. The aim was to find networks that fit the data best. We used Tabu and Hill-Climbing (HC) structure search algorithms with Akaike and Bayesian Information Criteria (AIC and BIC) scoring, which allow both fast computations and are robust in modeling real data (Beretta, Castelli, Gonçalves, Henriques, & Ramazzotti, 2018; Olmedilla, Rubio, Fuster-Parra, Pujals, & García-Mas, 2018). By varying scores and search methods, we build 1200 candidate networks using bootstrapped dataset by keeping 80% of all samples in each iteration. We restricted the size of network search space by blacklisting total 137 causally unfeasible directed edges. Variables related to subject's background were allowed to be parents for the candidate-related choices. All variables related to age and gender were only allowed to serve as parents. After model bootstrapping, we varied the edge frequency threshold and estimated the classification accuracy of the resulting DAG for the responses (individual or pairwise) using 10-fold cross validation.[2] For the model inference, we used maximum likelihood criterion and in validation we used posterior classification error loss (Nagarajan, Scutari, & Lèbre, 2013). Above steps were repeated separately for individual and pairwise response data. All variables, including valence, were kept in the data in this analysis.

Results

The relative valuation scores of candidates' for individual and pairwise valuation methods and pooled over all subjects are depicted in Fig. 1. Individual scores were computed by averaging over all ratings (x_8^r) for each candidate. Pairwise scores were computed by averaging over rows of an anti-symmetric pairwise rating matrix where each element was the sum of pairwise ratings (x^c) for all 55 combinations of candidates. As the scale of the scores was arbitrary, score distributions were standardized before plotting. Distributions were highly similar (Pearson correlation 0.958), thus confirming that both methods resulted in similar relative valuation of candidates.

From now on, as we report the modeling results, all variables (predictors) are referred with their alphabetic abbreviations. Variables x^r and x^e, which we consider as *main-effects*, are capitalized. The alphabetic abbreviations for the responses were SUITABILITY for x_8^r and SELECTION for x^c.

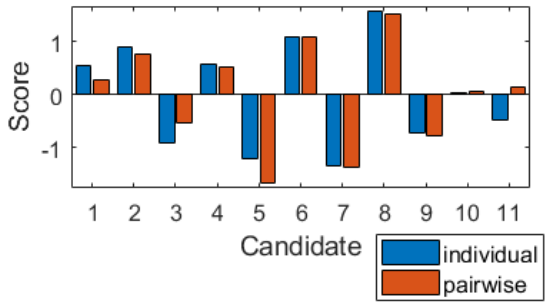


Figure 1: Mean valuation scores of all 11 candidates measured by individual (direct) and pairwise (indirect) method

[1] http://www.bnlearn.com for R (ver. 3.4).

[2] Note that until this point all nodes were equal and no "response" nodes were specified during bootstrapping

Linear Mixed-effect Models Results for the full and reduced linear models for the individual and pairwise responses are listed in Table 1. Positive t-values indicate increase of valuation (and vice versa). Total 6+7 individual and 24+26 pairwise fixed-effects terms surpassed p<0.05 (FDR adjusted over 99 and 88 terms) for full and reduced models. All predictors that were significant for at least one of the four models are shown in table (total 44 terms). Total 14 predictors were significant for at least two of the four models. Three of these were the main effects including variables TRUSTWORTHINESS, VALENCE and RIGHTEOUSNESS. Models reached MSE/MSE$_{nul}$ ratios 0.292-0.405 (smaller better) in 10-fold cross-validation. All models were also significant at p<0.0001 against permutations. Raw Pearson correlation between valence and responses were 0.802 (x_8^r) and 0.813 (x^c), which accounted lots of the variation in the full models.

		Full		Reduced	
		Ind.	Pair.	Ind.	Pair.
	MSE/MSE$_{null}$	0.292	0.300	0.401	0.405
	FAMILIARITY	-0.64	2.03	1.37	**3.52****
	TRUSTWORTHINESS	**3.95****	**3.92****	**3.35***	**3.29****
	RIGHTEOUSNESS	0.91	**7.00****	2.40	**7.69****
	NATIONAL_VALUE	2.64	-0.81	**3.91****	1.61
	VALENCE	**5.01****	**7.20****		
	AROUSAL	0.59	-2.33	-0.40	**-4.22****
	GENDER[male]	-0.40	**2.60***	-1.72	-0.06
age :	FAMILIARITY	**3.00***	0.07	2.51	0.41
age :	NATIONAL_VALUE	**-3.05***	0.73	**-4.06****	-1.03
age :	VALENCE	**4.60****	0.98		
carelesness :	AROUSAL	-1.76	**2.61***	-1.39	0.98
carelesness :	CO-OPERATION	2.75	-0.57	2.89*	-0.06
carelesness :	GENDER[male]	1.67	-1.42	0.17	**-3.81****
carelesness :	NATIONAL_VALUE	0.17	2.47	0.70	**3.30****
conservativeness :	CO-OPERATION	0.77	**2.94***	0.94	**2.48***
conservativeness :	DECISIONMAKING	-0.22	2.20	0.56	**3.93****
conservativeness :	FAMILIARITY	1.35	-1.34	-0.01	**-3.55****
conservativeness :	GENDER[male]	1.27	-0.95	**2.90***	1.98
conservativeness :	NATIONAL_VALUE	-0.40	0.61	1.36	**3.26****
conservativeness :	RIGHTEOUSNESS	-0.52	**-6.88****	-2.85	**-8.23****
conservativeness :	VALENCE	1.43	**3.43****		
discipline :	FAMILIARITY	-0.18	2.06	0.23	**2.84***
discipline :	VALENCE	**3.13***	**4.25****		
gender[male] :	AROUSAL	-0.28	**-6.19****	-1.62	**-5.20****
gender[male] :	CO-OPERATION	-0.90	-0.71	-2.83	**-4.75****
gender[male] :	FAMILIARITY	2.23	**3.12***	2.68	**4.32****
gender[male] :	GENDER[male]	-0.55	**3.06***	2.25	**5.19****
gender[male] :	NATIONAL_VALUE	2.88	**4.21****	**4.53****	**4.88****
gender[male] :	VALENCE	2.13	**-5.51****		
liberalness :	AROUSAL	0.12	**-2.75***	-0.21	-2.34
liberalness :	GENDER[male]	-0.62	**-2.77***	-0.08	**-1.72**
liberalness :	RIGHTEOUSNESS	0.31	**-3.08***	0.03	-2.40
liberalness :	VALENCE	2.10	**2.93***		
openness :	CO-OPERATION	-0.03	2.26	0.06	**2.72***
openness :	DECISIONMAKING	0.13	1.54	0.84	**2.53***
openness :	FAMILIARITY	-2.25	-1.81	**-3.02***	**-2.71***
openness :	RIGHTEOUSNESS	-1.50	**-5.28****	-1.05	**-4.44****
ordinality :	AROUSAL	-2.30	**3.37****	-1.38	**3.39****
ordinality :	TRUSTWORTHINESS	-1.16	**-2.90***	1.11	0.72
riskyinvestor :	GENDER[male]	0.09	0.98	1.48	**2.98***
riskyinvestor :	VALENCE	-1.35	**-2.57***		
SoMe_activity :	DECISIONMAKING	-1.76	**-4.47****	-2.51	**-4.25****
SoMe_activity :	RIGHTEOUSNESS	1.71	1.70	1.90	**2.60***
SoMe_activity :	TRUSTWORTHINESS	-1.52	**-3.19****	-1.47	**-2.82***

Table 1: T-values of the linear mixed-effects model using individual (Ind.) and pairwise (Pair.) valuation for full and reduced models. The main effects are capitalized and interactions (if any) marked with "**:**". Here * and ** indicate p<0.05 and p<0.01 (both FDR adjusted).

Bayesian network models In the Bayesian network analysis we found no major differences between search methods (HC or Tabu). AIC scoring, which tends to add more edges, resulted in generally smaller classification losses (i.e., better models). In general, higher edge density (bootstrapping frequencies <50%) resulted in higher classification accuracies. Here we present results obtained with Tabu and AIC. Results of bootstrapping are depicted in Fig. 2. Fig. 2a show all (undirected) edges with at least 5% frequency (i.e., 0.05) where upper triangular part is for individual and lower triangular for pairwise valuation. Weight 1.00 indicates very strong causal connection. The difference of the two triangular matrices is depicted in Fig. 2b (no thresholding), where all positive values correspond to higher frequency obtained for the pairwise valuation. The results indicate that most direct connections were within main effects (20 and 28) and subject-dependent characteristics (38 and 41), than between the two (only 6 and 14). While the individual valuation resulted in more subject-to-candidate edges (14 vs. 6), the

edges were generally weaker (<0.5) than those for the pairwise valuation (three edges with weight 1.0).

Finally, an example DAG for the SELECTION response is depicted in Fig. 3 with edge weight threshold 0.5 (with AIC and Tabu). The edge line weight corresponds to frequencies between 0.5 and 1.0 (thicker line = higher value). Node size indicates total number of incoming and outgoing edges (here between 2 and 9). The classification accuracy loss for this network was 0.581, while the (adjusted) baseline accuracy loss was 0.828. The *Markov blanket* for SELECTION included eight variables: gender, age, conservativeness, GENDER, VALENCE, TRUSTWORTHINESS, CO-OPERATION and RIGHTEOUSNESS. In other words, the SELECTION had direct causal connection with three background variables.

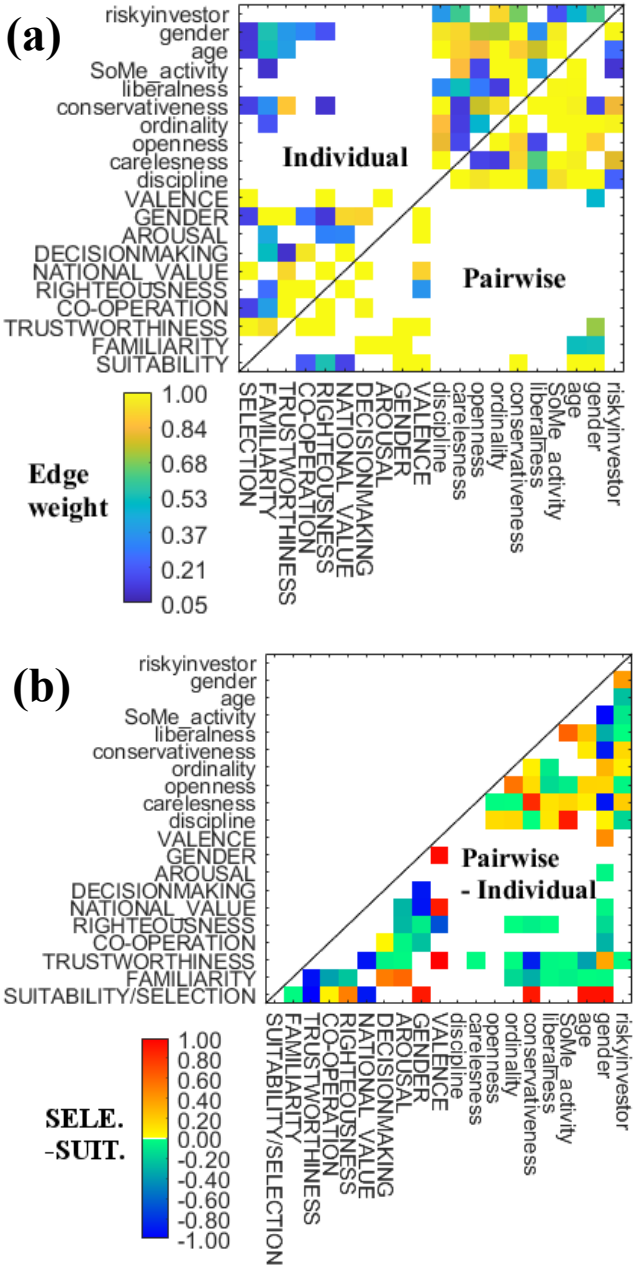


Figure 2: Bayesian network bootstrapping results for individual and pairwise valuation. **(a):** Occurrence rate of edges for individual (upper triangular) and pairwise valuation (lower triangular), only edges with at least 0.05 frequency are shown. **(b):** Difference of the two matrices (both unthresholded).

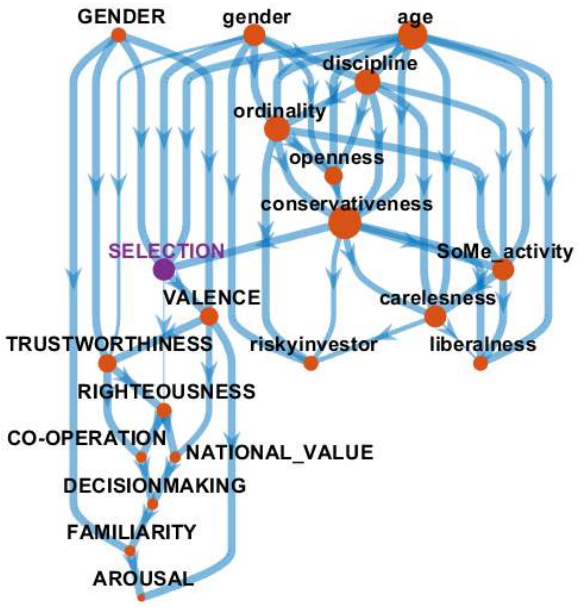


Figure 3: An example of a directed acyclic graph with 50 ed ges (at density >0.50) estimated using pairwise valuations of candidates. Line widths correspond to bootstrapping strength and node size to total number of connections.

Discussion

We collected behavioral questionnaire data on how voters valuate and judge politicians and their presumed suitability to serve as Prime Ministers. Aim was to pinpoint candidate and subject dependent factors that influence the valuation. We used linear mixed-effects models and Bayesian networks to analyze the data. We build two flavors of models; one for direct candidate valuation and the other for indirect valuation based candidate pairwise comparison. Although the average valuation scores of candidates were similar between direct and indirect approaches (Fig. 1), the models revealed differences in how the subjects arrived in their valuations.

In linear models, the pairwise valuation emphasized between individual- and candidate -related interactions with higher t-values magnitudes (Table 1). While the results for the main effects were similar (both highlighted trustworthiness, righteousness and valence), pairwise analysis resulted in more interaction terms surpassing significance (by the factor 3). While this can partly result from differences in number of samples (20 pairwise vs. 11 individuals per subject), it also reflects the difference in valuation processing when forced to choose between two choices. In particular, the interactions associated to emotion (valence and arousal) and gender (both candidate and subject)

had high impact in pairwise comparisons. Male responders favored male candidates and national value score of the candidate.

In order to complement our linear models, we also applied Bayesian network analysis. This framework allowed building full (nonlinear) probabilistic models for the data; however, here, we mainly used it as an exploratory tool to pinpoint causal connections between variables. The analysis also resulted in notable differences between individual and pairwise valuation (Fig. 2). In comparison to linear models, the candidate-related variable valence had direct causal effect only with electorate-related gender, but only for pairwise valuation. For individual valuation, causal connection between candidate valuation and electorate-related variables were more numerous (14 vs. 6), but were generally weaker. The strongest causal connection with the valuation score were found with conservativeness, age and gender of the electorate. These three had direct connections also with various other candidate-related properties, e.g., trustworthiness and familiarity.

In conclusion, we found that the background factors with strongest effect on the valuation of candidates were conservativeness, gender, age, ordinality and activity in social media of the voter. Emotion, especially valence, was strongly associated with valuation both directly and via interactions with voters' conservativeness, gender and ordinality. For males, higher arousal and valence strongly reduced the valuation. Emotion was found generally more important in pairwise candidate valuation.

Our results highlight the importance of how one measures the valuation of candidates (individual vs. pairwise) and how one analyzes such data (linear vs. nonlinear). Multiple views related to the data and methods are needed in pinpointing the most relevant effects. Previous studies have shown, that stimuli which trigger positive arousal increases the probability that people will behave according to the stimuli's suggestions in the future. Our results suggest that pairwise comparison – which is typical in USA elections – could enhance emotional and gender-related valuation of candidates.

Acknowledgments

This work is part of the Confidence AI -project funded by Helsingin Sanomat Foundation through "The post-truth era" research program.

References

Barisione, M. (2009). Valence Image and the standardisation of Democratic Political Leadership. *Leadership*, *5*(1), 41–60.

Bean, C., & Mughan, A. (1989). Leadership Effects in Parliamentary Elections in Australia and Britain. *The American Political Science Review*, *83*(4), 1165.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, *57*(1), 289–300.

Beretta, S., Castelli, M., Gonçalves, I., Henriques, R., & Ramazzotti, D. (2018). Learning the Structure of Bayesian Networks: A Quantitative Assessment of the Effect of Different Algorithmic Schemes. *Complexity*, *2018*, 1–12.

Bischoff, I., Neuhaus, C., Trautner, P., & Weber, B. (2013). The neuroeconomics of voting: Neural evidence of different sources of utility in voting. *Journal of Neuroscience, Psychology, and Economics*, *6*(4), 215–235.

Blouw, P., Solodkin, E., Thagard, P., & Eliasmith, C. (2016). Concepts as Semantic Pointers: A Framework and Computational Model. *Cognitive Science*, *40*(5), 1128–1162.

Borgelt, C., Steinbrecher, M., & Kruse, R. R. (2009). *Graphical Models: Representations for Learning, Reasoning and Data Mining.* Hoboken: John Wiley & Sons, Ltd.

Carney, D. R., Jost, J., Gosling, S. D., & Potter, J. (2008). The secret lives of liberals and conservatives: Personality profiles, interaction styles, and the things they leave behind. *Political Psychology*, *29*(6), 807–840.

Garzia, D. (2011). The personalization of politics in Western democracies: Causes and consequences on leader–follower relationships. *The Leadership Quarterly*, *22*(4), 697–709.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge ; New York: Cambridge University Press.

Gosling, S. D., Rentfrow, P. J., & Swann, W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, *37*(6), 504–528.

Hibbing, J. R., Smith, K. B., & Alford, J. R. (2014). Differences in negativity bias underlie variations in political ideology. *Behavioral and Brain Sciences*, *37*(03), 297–307.

Knutson, B., & Greer, S. M. (2008). Anticipatory affect: neural correlates and consequences for choice. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1511), 3771–3786.

Knutson, K. M., Wood, J. N., Spampinato, M. V., & Grafman, J. (2006). Politics on the Brain: An fMRI Investigation. *Social neuroscience*, *1*(1), 25–40.

McAllister, I. (2000). Keeping Them Honest: Public and Elite Perceptions of Ethical Conduct among Australian Legislators. *Political Studies*, *48*(1), 22–37.

Nagarajan, R., Scutari, M., & Lèbre, S. (2013). *Bayesian networks in R with applications in systems biology*. New York, N.Y: Springer.

Olmedilla, A., Rubio, V. J., Fuster-Parra, P., Pujals, C., & García-Mas, A. (2018). A Bayesian Approach to Sport Injuries Likelihood: Does Player's Self-

Efficacy and Environmental Factors Plays the Main Role? *Frontiers in Psychology*, *9*.
Rule, N. O., Ambady, N., Adams, R. B., Ozono, H., Nakashima, S., Yoshikawa, S., & Watabe, M. (2010). Polling the face: Prediction and consensus across cultures. *Journal of Personality and Social Psychology*, *98*(1), 1–15.
Scutari, M., Auconi, P., Caldarelli, G., & Franchi, L. (2017). Bayesian Networks Analysis of Malocclusion Data. *Scientific Reports*, *7*(1), 15236.
Shook, N. J., & Fazio, R. H. (2009). Political ideology, exploration of novel stimuli, and attitude formation. *Journal of Experimental Social Psychology*, *45*(4), 995–998.
Sibley, C. G., Osborne, D., & Duckitt, J. (2012). Personality and political orientation: Meta-analysis and test of a Threat-Constraint Model. *Journal of Research in Personality*, *46*(6), 664–677.
Tymula, A. A., & Glimcher, P. W. (2016). Expected Subjective Value Theory (ESVT): A Representation of Decision Under Risk and Certainty. *SSRN Electronic Journal*.
Walther, D. (2015). Picking the winner(s): Forecasting elections in multiparty systems. *Electoral Studies*, *40*, 1–13.
Wattenberg, M. P. (1991). *The rise of candidate-centered politics: presidential elections of the 1980s*. Cambridge, Mass.: Harvard University Press.
Weber, E. U., Blais, A.-R., & Betz, N. E. (2002). A domain-specific risk-attitude scale: measuring risk perceptions and risk behaviors. *Journal of Behavioral Decision Making*, *15*(4), 263–290.
Wilkinson, G. N., & Rogers, C. E. (1973). Symbolic Description of Factorial Models for Analysis of Variance. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *1973*(3), 392–399.
Wu, L. (2009). *Mixed Effects Models for Complex Data*. CRC Press.

Season naming and the local environment

Charles Kemp (c.kemp@unimelb.edu.au)
School of Psychological Sciences
University of Melbourne, 3010, Australia

Alice Gaby (alice.gaby@monash.edu)
School of Languages, Cultures and Linguistics
Monash University, 3800, Australia

Terry Regier (terry.regier@berkeley.edu)
Department of Linguistics and Cognitive Science Program
University of California, Berkeley, CA 94720 USA

Abstract

Seasonal patterns vary dramatically around the world, and we explore the extent to which systems of season categories support efficient communication about the local environment. Our analyses build on a domain-general information-theoretic model of categorization across languages, and we identify several qualitative predictions that emerge when this model is applied to season naming, including the prediction that systems with even numbers of categories should be more common than systems with odd sizes. We test the model quantitatively using a collection of season systems drawn from the linguistic and anthropological literature and data specifying temperature and precipitation in locations associated with these systems. Our results support the predicted even-odd asymmetry, and we also find that the model makes a number of successful predictions about the locations of boundaries between seasons.

Keywords: categorization; efficient communication; information theory

Imagine an alien geographer who has detailed knowledge about the natural environment in one part of our planet. The geographer knows how temperature, rainfall, humidity, wind speed, and wind direction vary over the course of the year. The geographer knows about clouds, fog, dew, storms, and lightning, and about the water levels in local streams, rivers and lakes. The geographer is intimately familiar with the flowering patterns of local plants and the breeding and migration patterns of local animals. In all of these cases the geographer knows about long-run averages as well as the variability that can be expected year to year. Before meeting any of the local people, what predictions could the geographer make about the categories named in their language? We focus on a special case of this question, and consider the extent to which named seasons reflect properties of the local environment. For example, we ask whether the geographer could predict how many seasons the local people might recognize, and where the boundaries between these seasons might lie.

Our approach builds on a growing body of work that explores ways in which languages support efficient communication (Rosch, 1978; Corter & Gluck, 1992; Gibson et al., 2019). Particularly relevant to our approach are information-theoretic accounts of variation in named categories across languages (Baddeley & Attewell, 2009; Kemp, Xu, & Regier, 2018). Regier, Kemp and colleagues have developed an information-theoretic formulation of the idea that named categories achieve a near-optimal tradeoff between complexity and communicative cost, and have applied it to domains including color (Zaslavsky, Kemp, Regier, & Tishby, 2018) and kinship (Kemp & Regier, 2012). Here we use the same formal framework to study season naming across languages.

Our work addresses an important question that is largely absent from previous formal treatments of categorization and efficient communication. The information theoretic framework that we adopt allows for different languages to reflect different communicative priorities. For example, the framework allows that systems of color categories may vary in part because speakers of different languages are embedded in environments (e.g. desert vs rainforest) with very different colour distributions, which may produce different local communicative needs. Previous authors acknowledge this point but typically implement models that assume that speakers all around the world encounter the same distributions over colors (Zaslavsky et al., 2018), kin types (Kemp & Regier, 2012), and other elements of their environments.

A notable exception is a project that explored words for ice and snow, and found that languages with a term that covers both of these concepts tend to be found in warm regions (Regier, Carstensen, & Kemp, 2016). That work focused specifically on environmental variation, but the naming behavior considered is extremely simple (one term versus two for frozen precipitation). Here we focus on environmental variation in a domain that offers the potential to make detailed predictions about not just the number of categories, but the locations of the boundaries between these categories.

Season naming has previously been studied by researchers from disciplines including linguistics, anthropology and geography. In a pioneering project Orlove (2003) compiled systems of season terms from twenty eight languages, and used them to document general tendencies in season naming. For example, Orlove suggests that seasons are usually characterized in terms of atmospheric phenomena such as rainfall, wind, and temperature. In some cases, however, seasons are based on changes related to plants (e.g. the flowering of a certain species), animals (e.g. the first appearance of a given species), or water levels in local rivers and lakes. Our approach builds on the work of Orlove and others by using computational methods to probe the relationship between season naming and the local environment.

A small amount of previous work has taken a computational approach to season naming. Hatfield-Dodds (2016) gives a detailed description of Yolngu seasons from the north east Arnhem land in Australia, and describes a computational model that uses climate data to detect when the seasons start and begin. Our work provides much less detail about any single language, but complements the approach of Hatfield-Dodds by using computational methods to explore season naming across a relatively large set of languages.

Previous authors have also discussed the notion of an optimal set of seasons for a given area. Entwisle, for example, proposes a set of five seasons for southeastern Australia that fits the local climate better than the four traditional European seasons (Entwisle, 2014). Proposals like these are often based in part on climate data, but are not typically derived from computational models. Our work builds on these approaches by connecting season naming with a domain-general account of categorization across languages.

Theoretical framework

This section introduces an information-theoretic approach that measures the extent to which a system of season terms supports informative communication about the environment. Consider a speaker who is talking about an event that falls within a standard year of 365 days. Let d indicate the day of the event. The prior distribution $p(d)$ captures the probability that the speaker will talk about an event that occurs on day d. For simplicity we assume that $p(d)$ is uniform.

Each day is associated with a distribution $p(\vec{s}|d)$ over a vector of season variables. We will consider three—precipitation (s^p), temperature (s^t), and temporal location within the year (s^y)—so that $\vec{s} = [s^p, s^t, s^y]$. Many other factors are relevant to season naming, and in principle we would like to include additional variables that capture information about the local climate, food sources, and bodies of water. In future it may be possible to include some of these variables, but for now we work with two climate variables (precipitation and temperature) that are readily available for locations all around the world.

Each day is also associated with a distribution $p(w|d)$ over words for seasons. The speaker labels day d by sampling from the distribution $p(w|d)$. After hearing the label the listener uses Bayesian inference to compute a distribution over the season variables:

$$p(\vec{s}|w) = \sum_d p(\vec{s}|d)p(d|w) \propto \sum_d p(\vec{s}|d)p(w|d)p(d).$$

We assume that communication succeeds to the extent that the speaker distribution $s = p(\vec{s}|d)$ resembles the listener distribution $l = p(\vec{s}|w)$, and formalize this idea using the same information-theoretic measure of communication cost used by previous work on domains including color and kinship (Kemp & Regier, 2012; Zaslavsky et al., 2018). Communication cost is defined as the Kullback-Leibler divergence $KL[s||l]$ from the speaker distribution s to the listener distribution l, and is low when the distributions are similar to each other. This cost measure can be used to assess the overall communication cost associated with an entire system of season terms. This overall cost is defined as the expected cost when the speaker communicates about an event:

$$\text{system cost} = \sum_d P(d)KL[s||l] = \sum_d P(d)KL[p(\vec{s}|d)||p(\vec{s}|w)].$$

There is a tradeoff between the communication cost of a system of categories and its complexity. Complexity can be formalized in different ways (Kemp & Regier, 2012; Zaslavsky et al., 2018) and here we define the complexity of a system as the number of terms that it contains. A system with many terms (high complexity) can allow the listener to reconstruct the speaker distribution very precisely (low communication cost), but a system with few terms (low complexity) means that the listener is typically able to approximate the speaker distribution only roughly.

Previous work suggests that systems of kinship terms (Kemp & Regier, 2012) and color terms (Zaslavsky et al., 2018) are efficient in the sense that they achieve near-optimal tradeoffs between communicative cost and complexity. An optimal tradeoff is achieved if the communicative cost of a system cannot be reduced without increasing the system's complexity, and vice versa. We will explore the extent to which attested season systems support efficient communication by comparing them to hypothetical systems of equal complexity.

Synthetic climate data

To illustrate some qualitative predictions of the model we first apply it to a simple synthetic data set that specifies how a single climate variable s^c varies over a hypothetical 48 day year. Fig 1a shows a climate variable s^c that rises smoothly then falls over the course of the year, as temperature does in many parts of the world. We combined this climate variable with a temporal variable s^y so that $\vec{s} = [s^c, s^y]$. Fig 1a includes season systems that minimize communication cost for different levels of complexity. For example, when $n = 2$ the optimal categories divide the year into days when $s^c < 0.5$ and days when $s^c > 0.5$

Although the model allows categories to be disconnected the categories in these optimal systems are always connected regions of the year. This result emerges because connected categories ensure that category members have similar values of the two season variables s^c and s^y.

A second qualitative result is that the turning points of the climate variable (i.e. the peak and trough) always lie within a category rather than at a category boundary. Because the days on either side of a turning point have similar values of s^c and s^y, assigning them to the same category minimizes communication cost. A related but more subtle result is that for a fixed value of the system size n, categories containing turning points are longer than categories without turning points. For example, when $n = 4$ the categories that contain the peak and trough have 13 days each, and the remaining categories have 11 days each. In general, combining two intervals of length k

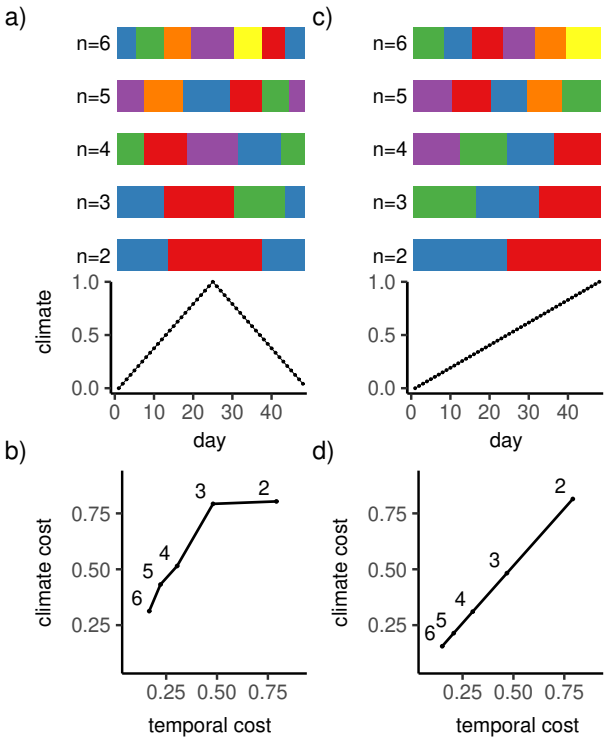


Figure 1: Analyses of synthetic climate data over a 48 day year. (a) Optimal systems of different sizes given a climate variable that varies smoothly. (b) Communication costs relative to the climate and temporal variables for the optimal systems in a. Labels indicate the sizes n of the 5 systems. (c), (d) Analogous results when the climate variable has a discontinuity at the end of the year.

that lie on either side of a turning point produces a category of size $2k$ that has the same coherence (in s^c) as a category of size k in a region without a turning point.

A final qualitative result is that *even systems* (i.e. systems with even numbers of categories) are more effective than *odd systems* at capturing information about the climate variable. The $n = 2$ system in Fig 1a distinguishes naturally between low and high values of s^c, but distinguishing between low, medium and high values turns out to be less straightforward. If only three categories are used, then the medium category must have two disconnected components (not shown in the figure). If all categories are connected regions of the year than four categories (as for the $n = 4$ system in Fig 1a) are actually needed to distinguish between low, medium and high values of s^c. More generally, if categories are connected then at least $2k-2$ categories are needed to distinguish k levels of s^c. As a result, distinguishing between levels of s^c in a parsimonious way naturally leads to an even system.

Fig 1b compares the even and odd systems in Fig 1a by plotting communication cost with respect to variables s^c (climate cost) and s^y (temporal cost). Although communication cost was defined earlier with respect to the entire set of season variables $\vec{s}$, here we use the same approach to define communication cost with respect to one variable at a time. Fig 1b shows that moving from 2 to 3 categories produces a relatively small improvement in climate cost, but moving from 3 to 4 categories produces a relatively large improvement. A similar but less pronounced kink in the curve is visible when moving from 4 to 5 to 6 categories. Moving from 2 to 3 categories does allow a speaker to convey additional information about s^y, but Fig 1b shows that this increase in complexity provides little additional information about s^c.

Most of the qualitative results just discussed depend critically on the assumption that s^c varies smoothly over time. Figs 1c and 1d show analogous results if s^c increases smoothly over the year then drops very sharply to its original value before the year starts again. In this case optimal categories are still connected regions, but the turning point always lies at a category boundary, the categories within each system have equal sizes, and there is no even-odd asymmetry.

The simulated environment in Fig 1a is simple and highly stylized, and it is not clear whether qualitative results like the even-odd asymmetry still apply if the climate variable rises and falls at different speeds, or if additional climate variables are added. Even so, we propose that seasonal variation in real-world climates is more like Fig 1a than Fig 1c. Our analyses therefore identify several characteristics of real-world systems that might be expected purely on the basis that these systems support communication about periodic variables that vary smoothly through time.

Season naming data

We next evaluated the model using real-world naming and environmental data. Orlove's (2003) ethnoclimatology database was not available and we therefore consulted the primary literature to assemble our own data set.

The data set includes 53 languages in total. For 25 of these languages the set of season terms was described in enough detail to be roughly positioned relative to the Western calendar year, and the data set includes season boundaries for each season in each of these systems. Four examples of systems with boundaries are shown in Fig 2. For the remaining 28 languages the data set specifies only the number of season terms in each language. Our data have a strong Australian focus because our two biggest sources are collections of indigenous Australian seasonal calendars compiled by the Commonwealth Bureau of Meteorology and the CSIRO.[1]

The data set inevitably reflects a number of decisions that are somewhat arbitrary. There is no universally accepted definition of a season, and it is likely that our sources adopted slightly different notions of what qualifies as a season. Some of the systems are hierarchies with two levels: they include a number of major seasons which are in turn divided into minor

[1]Unless specified otherwise, all season systems discussed in this paper (including three of the four in Fig 2) are drawn from one of these resources (`http://www.bom.gov.au/iwk/index.shtml` and `https://www.csiro.au/en/Research/Environment/Land-management/Indigenous/Indigenous-calendars`).

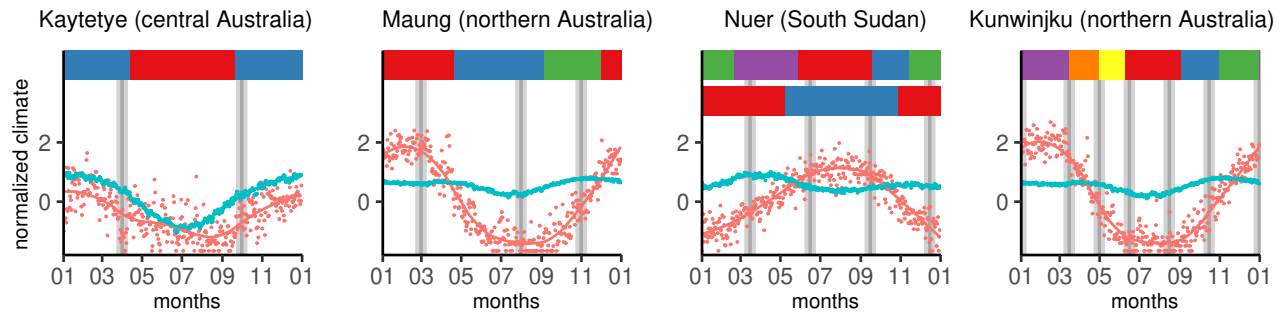


Figure 2: Climate data, seasons and optimal systems for four languages. Precipitation and temperature are shown using pink and cyan respectively. A cube-root transform has been applied to precipitation, and both variables are normalized to have zero mean and unit variance across the entire data set. Empirical season boundaries are shown using vertical lines. The colored bars across the top show optimal systems according to the model with sizes matched to the linguistic systems. Two systems are shown for Nuer for comparison to the two major seasons and the four minor seasons recognized in this language.

seasons. Judgments about subjective seasons are likely to be especially subjective. For example, the Tiwi system includes three major seasons and thirteen minor overlapping seasons, including *Kurukurari* ("season of the mangrove worm", when these worms are easy to find) and *Tawutawungari* ("season of the clap sticks," when special yam ceremonies are held). It seems likely that some of the languages in our data set have minor seasons that are not documented in our sources.

Although 25 of the languages in the data set include season boundaries, our sources repeatedly stress that mappings of indigenous seasons onto the calendar year are approximate only. Seasons are often fuzzy categories with no sharp boundaries, and the boundaries between seasons often shift from year to year as a result of variability in the local climate and other factors.

Some of our sources describe overlapping seasons, and this overlap is preserved in our data set. When seasons overlapped the distribution $P(w|d)$ over season terms for a given day was taken to be uniform over all seasons including that day. None of our sources describes gaps (i.e. unnamed periods) between seasons, and as a result each system in our data assigns each day to at least one season.

Among our systems with season boundaries, seasons always correspond to connected regions of the year, but exceptions are known outside our data set. For example, Rukiga has two words for seasons, *orugazi* (rainy season) and *ekyanda* (dry season), but these seasons may alternate over the course of a calendar year so that there are two rainy seasons and two dry seasons (Orlove, 2003). For languages included in our data set, season terms may pick out disconnected regions of the year when actually applied by native speakers. For example, if an unusually cold spell occurred during the summer months, a Yolngu speaker might say that one season had "interrupted" another (Hatfield-Dodds, 2016). These interruptions mean that seasons can occur in different orders during the year, and that a particular season could occur multiple times. For all of these reasons the representations in our data set are best viewed as crude approximations of bodies of knowledge that are both rich and subtle.

Season variables

The precipitation (s^p) and temperature (s^t) variables are based on global gridded data available from the Climate Prediction Center (CPC) in the USA.[2] Our analyses used daily precipitation and daily temperature averaged over the period from 1979 to 2005 and excluding leap years. Following a common practice in climate modeling we applied a cube-root transform to the precipitation data. We then normalized both variables to have zero mean and unit variance; normalized variables for four locations are shown in Fig 2.

We assigned Glottocodes manually to each language in the data set then retrieved the position (i.e. latitude and longitude) associated with each language in the Glottolog data base (Hammarström, Forkel, & Haspelmath, 2018). We then used these positions to extract precipitation and temperature data for each language from the CPC data.

The distribution $p(\vec{s}|d)$ for a given day and location was defined as a multivariate Gaussian distribution over a three-dimensional space. Two of the dimensions were the normalized precipitation and temperature dimensions already described, and the temporal dimension ran from 1 to 365 days and wrapped around so that day 366 was identical to day 1. The covariance was an axis-aligned distribution with standard deviation of 0.1 along the precipitation and temperature dimensions and standard deviation of 40 along the temporal dimension. The relative magnitudes of these standard deviations capture assumptions about the extent to which season categories should be informative about the three dimensions. For example, increasing the standard deviation along the temporal dimension would mean that there is less pressure for season categories to convey precise information about the location of an event within a year. As a result the temporal dimension would become less important and precipitation and temperature would effectively become more important. The numerical parameters used in our analyses (i.e. 0.1 and 40) were intended to give precipitation and temperature equal

[2]CPC data provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA, from their Web site at `https://www.esrl.noaa.gov/psd/`

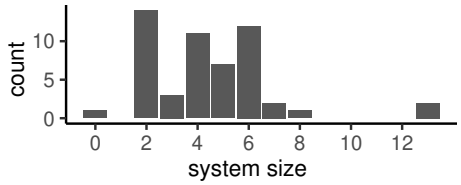


Figure 3: Distribution of system sizes.

weight, and to capture the idea that season categories locate events within the calendar year only roughly.

Because climate data are noisy we smoothed the distributions $p(\vec{s}|d)$ using a linear kernel with a width of 9 days. This smoothing process meant that the distribution $p(\vec{s}|d)$ for a given day (e.g. Jan 15) was defined as a weighted average of distributions for Jan 11 through 19. As a final step we discretized these distributions over a regular grid for use in our information-theoretic analyses.

Analysis of system sizes

Fig 3a shows the distribution of system sizes across our data set. For languages with hierarchical systems, the system size is defined as the number of seasons at the finest level of resolution. The system of size zero corresponds to the Grand Valley Dani, who constitute "a significant exception to [the general statement] that all cultures recognize seasons" (Heider, p 212). The two systems of size 13 represent Tiwi (described earlier) and Ngan'gi, and both feature ecological seasons defined with respect to the local plants and animals. Other languages in the data set almost certainly have ecological seasons that were not documented in our sources, and our Fig 3 therefore likely exaggerates the difference between Tiwi and Ngan'gi and the other languages in our data set.

As suggested earlier the model predicts a preference for systems with even sizes, and Fig 3a reveals that 2, 4 and 6 are the most common sizes. Leaving aside the three systems with sizes of zero or 13, 38 out of 50 systems or 76% have even sizes. We evaluated the significance of this result using a Bayesian mixed effects binomial model based on the `rstanarm` package and its default priors (Goodrich, Gabry, Ali, & Brilleman, 2018). The binary outcome variable indicated the parity (even or odd) of a system, and we included both a fixed intercept and a random intercept for language family to acknowledge genetic relatedness between languages.[3] The median of the fixed intercept indicates a probability of 0.77 that a random system would have an even size, and the 95% posterior credible interval ($[0.59, 0.94]$) excludes the probability (0.5) that makes even and odd systems are equally likely. Our data therefore support the conclusion that even systems are more common than odd systems.

Orlove (2003) previously noted that systems with odd sizes are rare, and in his data 23 out of 28 systems, or 82% have an even size. He did not offer an explanation for this asymmetry, but we have argued that it emerges from a pressure for season systems to support informative communication about factors that vary smoothly over time.

[3]The model call was `stan_glmer(parity ~ 1 + (1|language_family), family='binomial')`. Language families (e.g. Pama-Nyungan) were extracted from Glottolog.

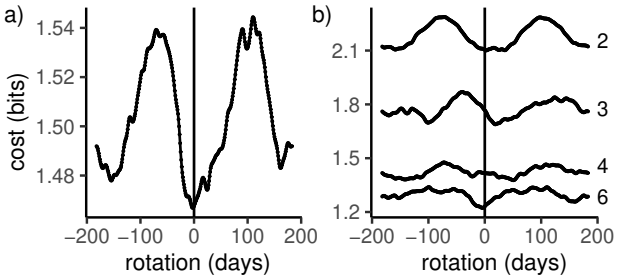


Figure 4: (a) Average rotation curve across all languages. (b) Averages across 2-term systems (4 languages), 3-term systems (3 languages), 4-term systems (8 systems) and 6-term systems (10 languages). The curve for 5-term systems (2 languages) is not shown because it overlaps the 6-term curve.

Analysis of season boundaries

Our remaining analyses focus on the 25 languages for which we have season boundaries. Four of these languages are shown in Fig 2 along with optimal systems according to our model.

Kaytetye has two seasons — *Watangka* (hot season) and *Yurluurp* (dry season) — and the boundaries between these categories roughly match the model predictions. Maung has three seasons: *Walmatpamalat* (heavy rain), *Wumulukuk* (cold weather) and *Kinyjapurr* (hot and humid). The model predicts three categories of roughly the right duration—in particular, the category that includes the steep increase in precipitation is shorter than the other two. The predicted season boundaries, however, are all shifted later in the year relative to the Maung system. Fig 2 also suggests that two of the Maung season boundaries lie close to simultaneous turning points in both temperature and rainfall. The Maung system therefore challenges the qualitative prediction that turning points in the climate data should lie within categories rather than at category boundaries.

Nuer has two major seasons: *tot* (mid-March to mid-September) and *mei* (mid-September to mid-March), each of which is divided into two minor seasons. The Nuer system provides additional evidence that season boundaries can be aligned with turning points in the climate data. Evans-Pritchard (1939, p 191) notes that "the *mei* season commences at the decline of the rains—not at their cessation." At the beginning of *mei* the Nuer start to anticipate the life they will lead when the dry weather arrives, and Evans-Pritchard (p 191) writes that their classification of seasons "aptly summarizes their way of looking at the movement of time, direction of attention in marginal months being as significant as actual climatic conditions."

Kunwinjku has six terms: *Kudjewk* (monsoon season), *Bangkerreng* (knock'em down storms), *Yekke* (start of dry time), *Wurrkeng* (cool weather time), *Kurrung* (hot, dry time), and *Kunumuleng* (humidity builds). The boundaries

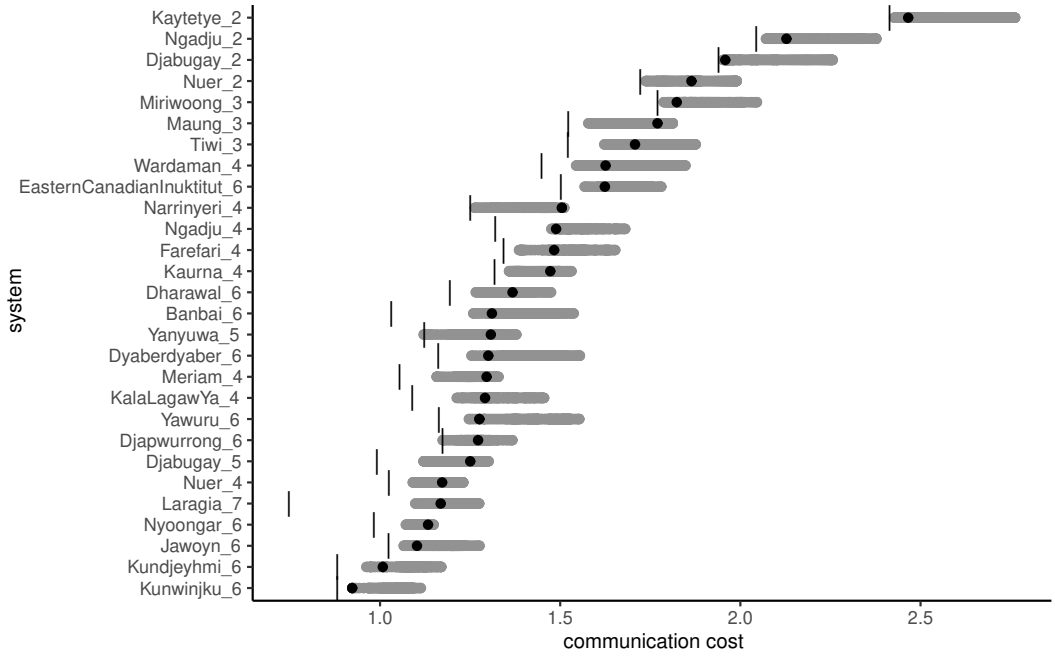


Figure 5: Attested systems (black dots) compared to rotations (grey dots) and size-matched optimal systems (black lines).

in the model system roughly match the linguistic data, and the model correctly predicts that there should be two short adjacent seasons during the part of the year when rainfall is declining sharply. Much more could be said about Kunwinjku and about each specific language in Fig 2, but we now turn to analyses that range more broadly across the entire data set.

A first general question is how closely season categories are aligned with variation in the two environmental variables (temperature and precipitation) included in our model. If a given system is closely aligned with the environmental variables, then rotating the system through the calendar year (i.e. incrementing all season boundaries by a constant while allowing for wrap around) should disrupt this alignment. Fig 4a plots communication cost against rotation size, and suggests that attested systems (i.e. systems rotated by zero days) tend to achieve lower communication cost than rotations of these systems. As shown in Fig 4, 0 day rotations score better than 99% of the 365 possible rotations. Fig 4b shows separate rotation curves for systems of size 2, 3, 4 and 6. The 2 term systems make an especially large contribution to the average result in Fig 4a, but a clear trough at zero days is visible also for the systems of size 6.

Fig 5 summarizes rotation results for individual languages. The three languages with hierarchies are included twice in the plot, once for each level of the hierarchy. Some systems (black dots) score better than most of their rotations (gray bar), including Kaytetye and Kunwinjku from Fig 2), but others (in particular Narrinyeri) do not. On average, each system scores better than 64% of its rotations.

Fig 5 also compares each system to the optimal system according to our model. Again, the pattern of results is mixed. Some systems (including Kaytetye and Kunwinjku) achieve scores close to the optimum, but others (including Laragia) do not. A likely explanation is that our model was given only two environmental variables even though language groups around the world use many markers of seasonal transitions other than changes in precipitation and temperature. For example, Narrinyeri seasons are distinguished by factors including "the growth of particular plants" and the "appearance of various creatures" (Berndt et al, 1993, p 76), and the lack of these factors in our analyses may explain why Narrinyeri achieves a sub-optimal score in Fig 5.

Conclusion

We developed a computational model that assumes that systems of season terms are near-optimal at conveying information about the local environment. The model helps to explain why systems with odd numbers of terms are relatively rare, and makes a number of successful predictions about the locations of season boundaries.

Our results do not provide strong support for claims about optimality but nevertheless demonstrate the value of the efficient-communication approach to naming and categorization. Most interesting to us are the qualitative issues exposed by the model. We have touched on some of them already, including the even-odd asymmetry, and the relationship between season boundaries and turning points in environmental variables. Many others arise: for example, our approach could be used to test the hypothesis that systems with large numbers of terms are especially likely to be found in regions with variable climates, and the hypothesis that boundaries are more likely to be aligned with sharp transitions (e.g. the first major rainfall of the year) than gradual changes in variables such as temperature. Although our current model is extremely simple, we have found it to be a useful conceptual tool for thinking about season naming across languages.

Acknowledgments

TR's work on this study was supported in part by the Defense Threat Reduction Agency; the content of the study does not necessarily reflect the position or policy of the U.S. government, and no official endorsement should be inferred.

References

Baddeley, R., & Attewell, D. (2009). The relationship between language and the environment: Information theory shows why we have only three lightness terms. *Psychological Science*, *20*(9), 1100–1107.

Berndt, R. M., Berndt, C. H., & Stanton, J. E. (1993). *A world that was: The Yaraldi of the Murray River and the Lakes, South Australia*. UBC Press.

Corter, J. E., & Gluck, M. A. (1992). Explaining basic categories: feature predictability and information. *Psychological Bulletin*, *111*(2), 291–303.

Entwisle, T. (2014). *Sprinter and Sprummer: Australia's changing seasons*. CSIRO.

Evans-Pritchard, E. E. (1939). Nuer time-reckoning. *Africa*, *12*(2), 189–216.

Gibson, E., Futrell, R., Piantadosi, S. T., Dautriche, I., Mahowald, K., Bergen, L., & Levy., R. (2019). How efficiency shapes human language. *Trends in Cognitive Sciences*.

Goodrich, B., Gabry, J., Ali, I., & Brilleman, S. (2018). *rstanarm: Bayesian applied regression modeling via Stan.* Retrieved from `http://mc-stan.org/` (R package version 2.17.4)

Hammarström, H., Forkel, R., & Haspelmath, M. (2018). *Glottolog 3.3.* Max Planck Institute for the Science of Human History. Jena. Retrieved from `https://glottolog.org/`

Hatfield-Dodds, Z. (2016). *Integrating understandings of a Yolngu seasonal calendar* (Honours Thesis). Australian National University.

Heider, K. G. (1970). *The Dugum Dani: A Papuan culture in the highlands of West New Guinea*. Wenner-Gren Foundation.

Kemp, C., & Regier, T. (2012). Kinship categories across languages reflect general communicative principles. *Science*, *336*(6084), 1049–1054.

Kemp, C., Xu, Y., & Regier, T. (2018). Semantic typology and efficient communication. *Annual Review of Linguistics*, *4*, 109-128.

Orlove, B. (2003). How people name seasons. In S. Strauss & B. S. Orlove (Eds.), *Weather, climate, culture.*

Regier, T., Carstensen, A., & Kemp, C. (2016). Languages support efficient communication about the environment: Words for snow revisited. *PLOS ONE*, *11*(4).

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 27–48). New York: Lawrence Erlbaum Associates.

Zaslavsky, N., Kemp, C., Regier, T., & Tishby, N. (2018). Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences*, *115*, 7937-7942.

Tuning to Multiple Statistics: Second Language Processing of Multiword Sequences across Registers

Elma Kerz (elma.kerz@ifaar.rwth-aachen.de)
Department of English Linguistics, RWTH Aachen University, Karmanstr. 17-19
52064 Aachen, Germany

Daniel Wiechmann (d.wiechmann@uva.nl)
Institute for Logic, Language and Computation, University of Amsterdam, Spuistraat 134
Amsterdam, 1012 VT Netherlands

Morten H. Christiansen (christiansen@cornell.edu)
Department of Psychology, Cornell University, Uris Hall
Ithaca, New York 14853

Abstract

A substantial body of research has demonstrated that children and adults (both native and non-native speakers) are sensitive to the statistics of multiword sequences (MWS) and rely on knowledge of such statistics to facilitate their language processing and boost their acquisition. However, this research was primarily aimed at determining whether and to what extent speakers can develop sensitivity to MWS statistics of a single type of linguistic input: that of spoken language. Recently, there has been a growing awareness of the key role of written input in the development of linguistic knowledge, as it provides a source of substantial change in the statistics of an individual's language experience. The present study reports on a series of experiments designed to determine whether second language learners of English are able to develop sensitivity to distributional statistics of MWS inherent in different (register-specific) input types.

Keywords: life-long learning; multiword sequences; second language processing; statistical learning

Recent theoretical approaches have highlighted the key importance of linguistic experience to the acquisition and processing of language. This broad class of approaches, commonly referred to as 'emergentist' approaches,[1] put the emphasis on usage and/or experience with language and assume a direct and immediate relationship between processing and learning, conceiving of them as inseparable rather than governed by different mechanisms ('two sides of the same coin'). Language acquisition is viewed as learning how to process linguistic input efficiently (e.g., Chang, Dell & Bock, 2006; Chater & Christiansen, 2018). In the emergentist perspective, language learning does not result in the establishment of a static knowledge system. Rather, as long as there is exposure to linguistic input, an individual's knowledge of a language is subject to constant change. Learning about the statistical regularities and distributional patterns inherent in linguistic input is viewed as a continuous process that does not end at some discrete point in time in ontogenetic development but instead takes place across the lifespan (e.g., Armstrong et al., 2017; Saffran & Kirkham, 2018; Seidenberg & MacDonald, 2018). This lifelong process brings about changes in language representations in response to the statistics in linguistic input. These experientially-driven adaptive processes are shown to occur across multiple linguistic levels and apply to the acquisition of new structures, the modification and/or adjustment of already learned representations or changes in accessibility of learned representations.

Moving away from the traditional 'words-and-rules' approach (e.g., Pinker, 1999), emergentist accounts have developed an increasing interest in the role of multiword sequences (MWS), often defined as variably-sized continuous or discontinuous recurring strings of words. This interest stems from an extensive body of evidence demonstrating that children and adults (both native and non-native speakers) are sensitive to the statistics of MWS and rely on knowledge of such statistics to facilitate their language processing and boost their acquisition (e.g. Shaoul & Westbury, 2011; N. Ellis, 2011; see Arnon & Christiansen, 2017, for a recent review). Sensitivity to the statistics of MWS facilitates chunking - required to integrate the greatest possible amount of available information as fast as possible so at to overcome the fleeting nature of linguistic input and the limited nature of our memory for sequences of linguistic input (Now-or-Never bottleneck, see Christiansen & Chater, 2016). Processing a MWS as a chunk will minimize memory load and speed up integration of the MWS with prior context (see, e.g., a chunk-based computational model of early language acquisition presented in a recent study by McCauley & Christiansen, 2019).

Frequency estimates obtained from corpora of actual language use have been shown to be robust predictors of language behavior across different types of experimental designs, as evidenced by higher accuracy rates, faster reaction times, and fewer and faster fixations. These effects have been shown in both child and adult populations as well as second-language learner populations. While earlier studies on the processing of MWS have used a threshold-approach to test whether MWS are stored and processed as holistic units (Biber & Conrad, 1999), more recent studies have incorpo-

[1] Following the literature (see, e.g., Kidd et al. 2018), we use the term 'emergentist' to refer to a broad class of approaches - usage-based (a.k.a. experience-based) models, complex dynamic systems theory, constraint-based approaches, exemplar-based models and connectionist models - that share a number of key tenets, for more details (see, e.g., Beckner et al. 2009; Daelemans & van den Bosch, 2005; McClelland et al. 2010)

rated further methodological improvement by testing these effects across the frequency continuum after controlling for substring frequency (for studies in child language acquisition, see, Bannard and Matthews, 2008; Matthews and Bannard, 2010; for studies on adult – both first and second language – processing see, Arnon, McCauley & Christiansen, 2017; Arnon and Snider, 2010; Hernandez et al., 2016, Kerz & Wiechmann, 2017, Yi et al. 2017).

This line of research has also shown that while being the most robust statistic, frequency is not the only kind of distributional information to which language users are sensitive. For example, in a study of MWS repetition in children, Matthews and Bannard (2010) showed that MWS with a high slot entropy value have increased uncertainty for what word occur in that slot and that such sequences were easier to generalize and hence easier to process for children than MWS with lower slot entropy.

The prior studies reviewed here have made important theoretical and methodological contributions to research on MWS. However, they have primarily focused on examining sensitivity to the frequencies derived from corpora representing spoken language (i.e., spontaneous conversations). In contrast to early child language acquisition (prior to literacy), where children are mainly exposed to the statistics of the spoken linguistic input (i.e., to child-directed speech), the role of written language becomes increasingly more important in later stages of learning which also sees increased demands on literacy. Indirect support for this assumption comes from a growing number of studies indicating that written language constitutes a key input type in the development of linguistic knowledge, as it provides a source of substantial change in the statistics of an individual's language experience (e.g., Montag & MacDonald, 2015; Seidenberg & MacDonald, 2018). Language users are thus faced with the challenge of keeping track of the ever-changing statistics of these two main types of language input. This challenge is exacerbated by considerable variability in the distributional properties of linguistic patterns at multiple levels of linguistic structure *within* these two input types (Roland, Dick & Elman, 2007; see also work on register/genre[2] variation by Biber and colleagues, e.g. Biber et al. 1999, Biber & Conrad, 2009).

In light of the lifelong nature of language learning highlighted in emergentist accounts, there is an apparent need not only to characterize the statistical learning processes in early stages of child language development, but also to understand how language users develop sensitivity in later stages of learning to the multiple kinds of statistics found in written language. This issue is of particular importance for second language (L2) learners, who are likely to get a lot of their language from written sources. Using a within-subject design, the present study sets out to investigate whether and to what extent language users can develop sensitivity towards the multiple statistics of MWS. We perform analyses of large samples of corpus data representing four registers and use the results from these analyses to make predictions about language users' performance in a MWS decision task. We predict faster response latencies for more frequent MWS (after controlling for all part frequencies) across the registers/genres investigated here. In addition to determining the effects of frequency ('more simple' distributional statistics), the study also investigated whether and to what extent language users are sensitive to 'more complex' distributional statistics (entropy) that captures the variability of MWS. The effects of frequency and entropy were investigated in a L2 learner population by conducting four reaction time experiments where processing latencies of MWS are compared for pairs of MWS that differ in sequence-frequency and entropy of their final slot.

[2]In the present paper, the terms 'registers' and 'genres' are used interchangeably in Biber's sense (2006:11) as referring to "situationally-defined varieties described for their characteristic distributions of linguistic structures and patterns."

Methods

Participants

Sixty advanced learners of English participated as a part of a larger project (34 female and 26 male, M = 23.56 years, SD = 4.52). All participants were college students recruited from the RWTH Aachen University studying either towards an BA or an MA at the time of testing. Participants were asked to fill out the Language Experience and Proficiency Questionnaire (LEAP-Q, see, Lemhofer & Broersma, 2012), a questionnaire used to obtain general demographic information and more specific information on self-rated proficiency for three language areas (reading, understanding and speaking) and self-rated current knowledge of L2 English and exposure to the L2. The data gathered from the LEAP-Q instrument are reported in Table 1, showing means, standard deviations and ranges of our L2 group.

Materials

The current study follows the general methodological approach described in the previous studies reviewed above that used carefully chosen stimuli, controlled for substring frequency. Following these studies, we chose pairs of four-word sequences as stimuli that differed only in the final word and in overall MWS frequency (high vs. low) but were matched for substring frequency (e.g. *to justify the cost* vs. *to justify the effort* from the newspaper register; e.g., *is beyond the scope* vs. *is beyond the boundaries* from the academic register). We constructed a total of 240 experimental items, 60 for each of four registers. The items were constructed using the Corpus Contemporary American English (COCA; Davies, 2008), a 560 million words corpus with approximately equal-sized subcomponents representing the statistics of MWS from the four target registers: (1) spoken (118 million words), (2) fiction: (113 million words), (3) newspaper (114 million words and (4) academic journals (112 million words). In a first step, all COCA text files were preprocessed using the sentence splitting (`ssplit`) and tokenization (`PTNTokenizer`) components from the Stanford CoreNLP toolkit V.3.2.9 (Manning et

Table 1: Self-report information on English acquisition, exposure, and proficiency

	mean	sd	obs. range
English acquisition (years)			
Age start acquisition	8.46	2.23	6–22
Age became fluent	14.63	3.9	8–23
Current exposure to English			
Friends (0-10)	4.63	3.1	0–10
Family (0-10)	1.36	2.6	0–10
Reading (0-10)	7.64	2.25	1–10
Class instruction (0-10)	5.48	3.43	0–10
Self instruction (0-10)	4.86	2.81	0–10
Watch (0-10)	7.64	2.72	0–10
Listening music (0-10)	7.39	2.84	0–10
Social Media (0-10)	7.39	2.68	0–10
Immersion (month)			
English speak. country	2.96	3.46	0–11
Self-rated proficiency			
Speaking (0-10)	7.25	1.69	1–10
Listening (0-10)	8.49	1.38	5–10
Reading (0-10)	7.86	1.58	1–10

al., 2014). In a second step, we extracted frequencies for all n-grams of orders 1 to 4 using Java scripts. N-grams with a frequency of one (so-called 'hapax legomena') were discarded. These two steps were performed on the RWTH Aachen University high-performance computing cluster. In a third step, four-grams that had a function word as their last word were filtered out to ensure that the position at which entropy was measured was filled by a lexical word.[3]. For all remaining four-grams the Shannon entropy H was computed for their final word slot, which is given in (1), where X is the final slot of the MWS, each x is a word that appears in that slot, and $p(x)$ is the probability of seeing each x in that position. All conditional word probabilities needed to compute entropy scores were estimated using second-order Markov models (cf. Willems, Frank, Nijhof, Hagoort, and van den Bosch, 2015).

$$H(X) = -\sum_{x} p(x) \log p(x) \quad (1)$$

We next identified all sequences of four words that began with the same first three words (i.e. shared the same pattern). Within each set of these sequences, a frequency difference score (FDS) was computed for a given sequence in relation to the most frequent sequence in that set.[4] We then ordered the sequences according to their FDS and explored how FDS scores related to entropy using a moving window approach/technique. A window with a size that corresponded to a predefined FDS was moved over the entire candidate-item pool to bin all four-gram into groups with similar FDS (see Figure 2 for a visualization). Inspection of these data indicated that four-grams with small differences in FDS tended to exhibit low entropy scores. Based on these observations, we restricted our candidate pool to four-grams that had entropy scores between 0 and 3 and a difference in log normalized frequency between 6.5 and 30. From this candidate pool, we randomly sampled, from each register, a total of 60 experimental item pairs: 20 pairs from each of three entropy ranges ('low': H(X) between 0 and 1, 'mid': H(X) between 1 and 2, 'high': H(X) between 1 and 2). Applying these filters meant that the log frequencies of our items ranged between 0.69 and 6.85 (spoken $0.69-6.07$, fiction: $0.69-6.85$; news $0.69-5.97$, academic $0.69-5.87$).

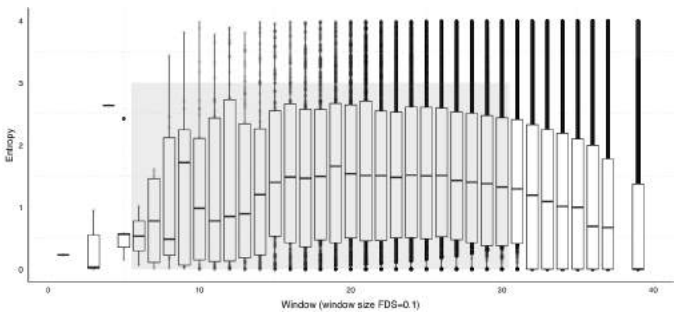


Figure 1: Distribution of entropy scores across the 'frequency difference score' (FDS) range for the academic subcomponent of COCA. The shaded area represents the range from which experimental items were sampled.

Procedure

Four separate MWS experiments, one for each register, were conducted as part of a larger project at two different days with two registers tested per day (day 1: academic and fiction; day 2: news and spoken). Each experiment was divided into two blocks of about 7 minutes each, which were separated by intervening tasks assessing individual differences in L2 experience and another task assessing a cognitive individual differences variable (not investigated here). The 120 MWS from a given register were distributed across two lists that each contained one of the two variants of a given pair, so that in a given experimental run participants would never see both variants of a pair. In addition to the experimental items, the lists also contained 60 ungrammatical items, which were incorrect due to scrambled word order. The order of presentation of the blocks was counterbalanced between participants. Participants were asked to judge if a four-word sequence that appeared on the screen was a possible sequence in English or not. They were given no information about the fact that MWS were extracted from different registers. Each trial began with the presentation of a fixation point for 500 ms. Phrases appeared at once in the middle of the screen and participants were instructed to respond as quickly and accurately as possible using the keyboard. The MWS was then presented and stayed visible on the screen until participants responded or

[3]The stop-list for function words was derived from the 'Essential Word List' https://www.edu.uwo.ca/faculty-profiles/docs/other/webb/essential-word-list.pdf

[4]FDS were expressed in terms of as the absolute of the $log10$ of the normalized frequency of a four-gram minus the $log10$ of the normalized frequency of the most frequent four-gram sharing the first trigram.

until 3000 ms had passed. The task was run using PsychoPy v3.0 (Peirce, 2007).

Results

Responses under 200 ms and over two standard deviations from the mean were excluded. This resulted in loss of small percentage of data for each register ($< 9\%$). Accuracy for target items was near ceiling ($> 92\%$ correct) for all registers. On average, participants were faster in responding to MWS from the spoken and fiction registers (mean response latencies spoken and fiction = 1.44 seconds, $SD = 0.5$) than in responding to MWS from the academic and news registers (mean response latencies: academic = 1.55 seconds, $SD = 0.51$; news = 1.57 seconds, $SD = 0.5$). The results were analyzed using mixed-effect linear regression models. To determine to what extent L2 learners can develop sensitivity to the two distributional statistics of MWS (log MWS frequency, slot entropy) inherent in the four registers investigated in the study, separate models were fitted to the data from each of the four experiments.[5] All analyses were carried out using the `lme4` package (v 1.1-17, Bates et al., 2015) in `R` (version 3.5.0; R Core Team, 2017). Log response times were used as the predicted variable to reduce the skewness in the distribution of response times. In a first step, we fitted models containing all control variables, i.e. LENGTH (in number of characters), two substring frequency measures (LOG FREQUENCY OF THE FINAL UNIGRAM and LOG FREQUENCY OF THE FINAL BIGRAM)[6], BLOCK ORDER (first vs. second), and PAIR VARIANT (high-low frequency variant of a pair). All continuous predictors were mean centered prior to analysis. We then added our key predictors, LOG PHRASE FREQUENCY and SLOT ENTROPY (high, mid, low), to examine their predictive value over and above our controls, using likelihood ratio tests to compare nested models. In all models, we used the maximal random effect structure justified by the data, which included random intercepts for participants and items and by-subject random slopes for log MWS frequency and entropy. To compare the effects of (log) MWS frequency on (log) reaction times across the four registers, standardized coefficients as well as marginal and conditional pseudo-R2 were computed (cf. Nakagawa and Schielzeth, 2013).[7]

In a next step, we tested for a potential interaction between our two key predictors. To this end, we conducted model comparisons between a model containing only the main effects and a corresponding model that also included the two-way interaction between MWS frequency and slot-entropy using Akaike's Information Criterion (AIC). The results of the final best-fitting model for each register are presented in Table 2 below. Likelihood ratio tests comparing models including LOG MWS FREQUENCY with a model that included only the control variables revealed that – after statistically controlling for the effects of length and frequency-related control variables – MWS frequency exerted a significant effect on (log) reaction times for all registers except fiction (spoken: $\chi(1) = 18.53$, $p < .0001$; fiction: $\chi(1) = 2.28$, $p = 0.51$; academic: $\chi(1) = 13.31$, $p = 0.004$; news: $\chi(1) = 59.46$, $p < .0001$). The frequency effect was found to be strongest in the spoken and news registers (both standardized $\beta = -0.13$), followed by academic language (standardized $\beta = -0.11$). A significant main effect of slot entropy was observed for the spoken and academic register (spoken: $\chi(1) = 40.03$, $p < 0.001$; fiction: $\chi(1) = 5.77$, $p = 0.58$; academic: $\chi(1) = 14.23$, $p = 0.047$; news: $\chi(1) = 12.42$, $p = 0.061$), such that that mean response times were significantly faster for MWS with higher slot entropy. There was also a significant interaction effect between MWS frequency and entropy in the spoken register ($\beta = -0.046$, $SE = 0.016$, $t = -2.79$), indicating that the frequency effect was more pronounced in high-entropy MWS (see Figure 2). The effects of the length and frequency related control variables were in the predicted directions - with longer MWS being read more slowly on average and more frequent final words leading to faster response times - but these effects were significant in only some of the registers. Significant effects of block order were observed for two of the four registers (see Table 3 for details).

Discussion and Conclusions

In this paper we reported a series of experiments with English L2 learners designed to determine to what extent the multiple distributional statistics of MWS inherent in register/genre-specific linguistic input (as estimated using a large corpus of actual language use) would affect the processing latencies of the MWS. We found the MWS frequency effect in three out of four registers investigated (all with the exception of fiction), i.e. our participants responded faster to higher frequency MWS, even after controlling for the effects of substring frequency. The finding that our participants showed MWS frequency effects in the spoken register is in line with the results of previous studies on adult native speakers and non-native speakers (e.g., Arnon & Snider, 2010; Tremblay et al., 2012; Hernandez et al. 2016). Importantly, our findings extend this

[5] Two anonymous reviewers recommended to pool the data from the four experiments and report on the interaction effects between our key predictors (log MWS frequency, slot entropy) and a 'register' variable. We have computed such a 'global' using orthogonal contrasts for the 'register' variable. This model revealed a significant effect of log MWS frequency ($= -0.026$, $SE = 0.006$, $t = -4.169$) but no significant interactions between log MWS frequency and register. Since we aimed to test whether our participants can detect and adapt to the changing statistics of multiple input types, we decided to report on four separate models in the study.

[6] To avoid overfitting resulting from multiple substring frequency control variables, we followed the procedure used in Arnon & Snider (2010) and first ran a model with all substring frequency controls and then removed the variables whose standard error was greater than the value of their coefficient in the model. All reported models had low collinearity (all $VIFs < 1.8$).

[7] Standardized beta coefficients indicate how many standard deviations a dependent variable will change, per standard deviation increase in the predictor variable. Pseudo-R2 for generalized mixed-effect models (GLMM) can be categorized into two types: Marginal R^2 represents the variance explained by fixed factors. Conditional R^2 is interpreted as variance explained by both fixed and random factors (i.e. the entire model).

Table 2: Results from the mixed effects regression models fitted to the data from the four experiments.

	Register comparison			
	Log. Reaction Time			
	Spoken	Fiction	Academic	News
Constant	0.286^{**}	0.752^{***}	0.447^{***}	0.532^{***}
	(0.069, 0.503)	(0.469, 1.036)	(0.302, 0.592)	(0.300, 0.763)
log MWS frequency	$B = -0.037^{**}$	$B = -0.013$	$B = -0.025^{**}$	$B = -0.035^{**}$
	(−0.061, −0.013)	(−0.039, 0.013)	(−0.042, −0.008)	(−0.060, −0.009)
	$\beta = -0.13$	$\beta = -0.06$	$\beta = -0.11$	$\beta = -0.13$
slot entropy (low to mid)	$B = 0.009$	$B = 0.019$	$B = -0.046^{**}$	$B = 0.037$
	(−0.066, 0.085)	(−0.039, 0.078)	(−0.078, −0.014)	(−0.012, 0.086)
	$\beta = 0.01$	$\beta = 0.03$	$\beta = -0.08$	$\beta = 0.11$
slot entropy (low to high)	$B = 0.069$	$B = -0.010$	$B = 0.006$	$B = -0.0003$
	(−0.013, 0.152)	(−0.067, 0.046)	(−0.029, 0.041)	(−0.050, 0.049)
	$\beta = 0.1$	$\beta = 0.02$	$\beta = -0.02$	$\beta = 0.04$
log MWS freq.:entropy (mid)	$B = -0.013$			
	(−0.044, 0.017)			
log MWS freq.:entropy (high)	$B = -0.041^{*}$			
	(−0.081, −0.001)			
log final bigram	$B = 0.002$	$B = 0.005$	$B = 0.004$	$B = -0.003$
	(−0.007, 0.011)	(−0.006, 0.016)	(−0.003, 0.011)	(−0.012, 0.005)
log final word	$B = 0.011$	$B = -0.031^{**}$	$B = -0.020^{***}$	$B = -0.010$
	(−0.002, 0.024)	(−0.050, −0.011)	(−0.031, −0.009)	(−0.027, 0.007)
length (char)	$B = 0.003$	$B = 0.002$	$B = 0.008^{***}$	$B = 0.008^{***}$
	(−0.004, 0.010)	(−0.005, 0.009)	(0.005, 0.012)	(0.003, 0.013)
pair variant (low to high)	$B = -0.065^{*}$	$B = 0.008$	$B = -0.028$	$B = -0.036$
	(−0.118, −0.012)	(−0.055, 0.071)	(−0.070, 0.014)	(−0.093, 0.022)
block order	$B = -0.025$	$B = -0.097^{***}$	$B = 0.010$	$B = -0.062^{***}$
	(−0.060, 0.009)	(−0.140, −0.054)	(−0.016, 0.035)	(−0.097, −0.028)
Conditional R^2	0.41	0.39	0.30	0.40
Marginal R^2	0.02	0.03	0.03	0.05

Note: *p<0.05; **p<0.01; ***p<0.001
Numbers in parentheses indicate 95% confidence intervals.
B indicate unstandardized beta coefficients
β indicate standardized beta coefficients

prior research by demonstrating that language users' ability to track the statistics in the input is not limited to the spoken conversational language but it can be observed in written registers/genres. In addition to the MWS frequency effects, the significant main effect of entropy found in the register of academic writing indicated that our participants were able to develop sensitivity to more complex distributional statistics, i.e. they showed faster response latencies with higher slot entropy. The direction of the entropy effect is consistent with the finding of Matthews & Bannard's (2010) study demonstrating that 2-3 years old children were more accurate in repeating MWS with higher slot entropy. Additional support for the facilitatory effect of more complex distributional statistics on the processing of MWS comes from the significant interaction between frequency and entropy indicating that the effect of MWS frequency increased with increasing degrees of MWS entropy.

To our knowledge, this is the first study to show that language users (whether native or non-native speakers) are able to tune to the multiple distributional statistics inherent in register-specific input types within a single language. The findings from this study provide a key contribution to a growing body of research that explore statistical learning through the lens of multilingual acquisition. This research has explored the consequences of accruing statistics in multilanguage input and has typically been conducted using artificial stimulus-sequences (cf., Bulgarelli, Lebkuecher & Weiss, 2018, for a recent overview). Our study has demonstrated how the acquisition of multiple statistics can be investigated on the basis of stimuli constructed from large corpora of au-

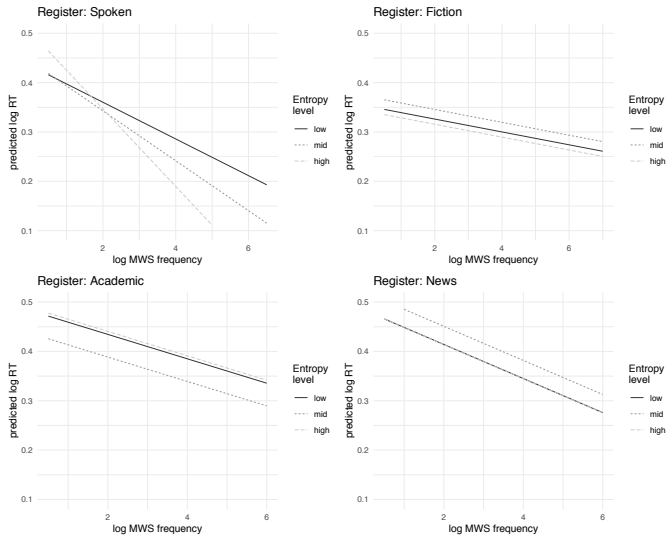


Figure 2: Effects of log MWS frequency by entropy level (high, mid, low) across registers.

thentic language data.

Some of the questions left open by the current study may provide interesting avenues for future work. First, we investigated sensitivity to the register-specific multiple statistics in adult second language learners. The question arises whether similar results could be obtained for adult native speakers. Second, in the light of the lifelong nature of language learning, it is of special importance important to track the developmental progression in response to the changes in the distributional properties of the linguistic input across the lifespan. This involves understanding not only the developmental progression during early stages of child language acquisition (prior to literacy) but also understanding the nature of such progression during later stages of language development, which is strongly driven by the distributional statistics of written input (Seidenberg & MacDonald, 2018). And, third, it would be important to determine whether the ability to tune to multiple statistics is subject to individual variability, and if so, to what extent this variability is linked to other cognitive, affective and environmental factors.

References

Arnon, I., & Christiansen, M. H. (2017). The role of multiword building blocks in explaining l1–l2 differences. *Topics in Cognitive Science*, *9*(3), 621–636.

Arnon, I., McCauley, S. M., & Christiansen, M. H. (2017). Digging up the building blocks of language: Age-of-acquisition effects for multiword phrases. *Journal of Memory and Language*, *92*, 265–280.

Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language*, *62*(1), 67–82.

Bannard, C., & Matthews, D. (2008). Stored word sequences in language learning: The effect of familiarity on children's repetition of four-word combinations. *Psychological Science*, *19*(3), 241–248.

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., ... Grothendieck, G. (2015). Package 'lme4'. *Convergence*, *12*(1).

Beckner, C., Blythe, R., Bybee, J., Christiansen, M. H., Croft, W., Ellis, N. C., ... Schoenemann, T. (2009). Language is a complex adaptive system: Position paper. *Language Learning*, *59*(s1), 1–26.

Biber, D. (1999). A register perspective on grammar and discourse: variability in the form and use of english complement clauses. *Discourse Studies*, *1*(2), 131–150.

Biber, D., & Conrad, S. (1999). Lexical bundles in conversation and academic prose. *Language and Computers*, *26*, 181–190.

Biber, D., & Conrad, S. (2009). *Register, genre, and style*. Cambridge University Press.

Bulgarelli, F., Lebkuecher, A. L., & Weiss, D. J. (2018). Statistical learning and bilingualism. *Language, speech, and hearing services in schools*, *49*(3S), 740–753.

Chang, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic. *Psychological Review*, *113*(2), 234.

Chang, F., Janciauskas, M., & Fitz, H. (2012). Language adaptation and learning: Getting explicit about implicit learning. *Language and Linguistics Compass*, *6*(5), 259–278.

Chater, N., & Christiansen, M. H. (2018). Language acquisition as skill learning. *Current Opinion in Behavioral Sciences*, *21*, 205–208.

Christiansen, M. H., & Chater, N. (2016a). *Creating language: Integrating evolution, acquisition, and processing*. Cambridge, MA: MIT Press.

Christiansen, M. H., & Chater, N. (2016b). The Now-or-Never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, *39, e62*.

Daelemans, W., & Van den Bosch, A. (2005). *Memory-based language processing*. Cambridge University Press.

Davies, M. (2009). The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics*, *14*(2), 159–190.

Ellis, N. C. (2012). Formulaic language and second language acquisition: Zipf and the phrasal teddy bear. *Annual Review of Applied Linguistics*, *32*, 17–44.

Hernández, M., Costa, A., & Arnon, I. (2016). More than words: multiword frequency effects in non-native speakers. *Language, Cognition and Neuroscience*, *31*(6), 785–800.

Kerz, E., & Wiechmann, D. (2017). Individual differences in L2 processing of multi-word phrases: Effects of working memory and personality. In R. Mitkov (Ed.), *Computational and corpus-based phraseology. EUROPHRAS 2017* (pp. 306–321).

Kidd, E., Donnelly, S., & Christiansen, M. H. (2018). Individual differences in language acquisition and processing. *Trends in Cognitive Sciences*, 154–169.

Lemhöfer, K., & Broersma, M. (2012). Introducing lextale: A quick and valid lexical test for advanced learners of english. *Behavior Research Methods*, *44*(2), 325–343.

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the Association for Computational Linguistics: system demonstrations* (pp. 55–60).

Matthews, D., & Bannard, C. (2010). Children's production of unfamiliar word sequences is predicted by positional variability and latent classes in a large sample of child-directed speech. *Cognitive Science*, *34*(3), 465–488.

McCauley, S. M., & Christiansen, M. H. (2019). Language learning as language use: A cross-linguistic model of child language development. *Psychological Review*, *126*(1), 1–51.

McClelland, J. L., Botvinick, M. M., Noelle, D. C., Plaut, D. C., Rogers, T. T., Seidenberg, M. S., & Smith, L. B. (2010). Letting structure emerge: connectionist and dynamical systems approaches to cognition. *Trends in Cognitive Sciences*, *14*(8), 348–356.

Montag, J. L., & MacDonald, M. C. (2015). Text exposure predicts spoken production of complex sentences in 8-and 12-year-old children and adults. *Journal of Experimental Psychology: General*, *144*(2), 447-468.

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining r2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*(2), 133–142.

Peirce, J. W. (2007). Psychopy—psychophysics software in python. *Journal of neuroscience methods*, *162*(1-2), 8–13.

Roland, D., Dick, F., & Elman, J. L. (2007). Frequency of basic english grammatical structures: A corpus analysis. *Journal of Memory and Language*, *57*(3), 348–379.

Saffran, J. R., & Kirkham, N. Z. (2018). Infant statistical learning. *Annual Review of Psychology*, *69*, 181-203.

Seidenberg, M. S., & MacDonald, M. C. (2018). The impact of language experience on language and reading: A statistical learning approach. *Topics in Language Disorders*, *38*(1), 66–83.

Shaoul, C., & Westbury, C. (2011). Formulaic sequences: Do they exist and do they matter? *The Mental Lexicon*, *6*(1), 171–196.

Team, R. C., et al. (2013). R: A language and environment for statistical computing.

Willems, R. M., Frank, S. L., Nijhof, A. D., Hagoort, P., & Van den Bosch, A. (2015). Prediction during natural language comprehension. *Cerebral Cortex*, *26*(6), 2506–2516.

Yi, W., Lu, S., & Ma, G. (2017). Frequency, contingency and online processing of multiword sequences: An eye-tracking study. *Second Language Research*, *33*(4), 519–549.

Comparing Alternative Computational Models of the Stroop Task Using Effective Connectivity Analysis of fMRI Data

Micah Ketola (ketolm@uw.edu)
Department of Psychology, University of Washington
Campus Box 351525, Seattle, WA 98195 USA

Linxing Preston Jiang (prestonj@cs.washington.edu)
Paul G. Allen School of Computer Science and Engineering, University of Washington
Campus Box 351525, Seattle, WA 98195 USA

Andrea Stocco (stocco@uw.edu)
Department of Psychology and Institute for Learning and Brain Sciences (I-LABS), University of Washington,
Campus Box 351525, Seattle, WA 98195 USA

Abstract

Methodological advances have made it possible to generate fMRI predictions for cognitive architectures, such as ACT-R, thus expanding the range of model predictions and making it possible to distinguish between alternative models that produce otherwise identical behavioral patterns. However, for tasks associated with relatively brief response times, fMRI predictions are often not sufficient to compare alternative models. In this paper, we outline a method based on *effective connectivity*, which significantly augments the amount of information that can be extracted from fMRI data to distinguish between models. We show the application of this method in the case of two competing ACT-R models of the Stroop task. Although the models make, predictably, identical behavioral and BOLD time-course predictions, patterns of functional connectivity favor one model over the other. Finally, we show that the same data suggests directions in which both models should be revised.

Keywords: ACT-R, Dynamic Causal Modeling, Cognitive Science

Introduction

One of the traditional problems in the field of cognitive modeling is deciding which of two alternative models provides best explains a phenomenon. Traditionally, the most common approach has been to compare models using null-hypothesis testing procedures. In essence, conditions are identified in which the two models make qualitatively different predictions, and the hypothesized pattern is tested using classical statistical testing techniques. While more sophisticated approaches have been proposed (Pitt, Kim, Navarro, & Myung, 2006), this approach remains the *de facto* standard of the field.

The search for conditions in which two models differ is sometimes strenuous, as the same external behavior can occasionally be obtained through different possible internal processes and model parameters. By shedding light on more direct correlates of cognitive processes, neuroimaging data provides a potential way to distinguish between otherwise behaviorally identical results (Sohn et al., 2004). For this reason, procedures have been devised to derive neuroimaging predictions from computational models, most commonly in the domain of fMRI (Anderson, Fincham, Qin, & Stocco, 2008; Borst, Nijboer, Taatgen, van Rijn, & Anderson, 2015).

While the use of fMRI has greatly expanded upon the possible predictions that can distinguish between the two models, a number of limitations still exist. A main limitation arises from poor temporal resolution of fMRI. The BOLD signal that is recorded in MRI scanners is extremely sluggish, and peaks approximately five seconds after an event. This poses a problem for resolving cognitive processes that occur quickly in time.

Other neuroimaging methods, such as EEG and MEG, offer much greater temporal resolution, but they trade off this advantage with much lower spatial resolution. Furthermore, the oscillatory nature of EEG and MEG signals further complicates the process of deriving predictions from models, as changes in raw signals can occur at different frequency bands (van Vugt, 2014).

Even if these technical issues could be solved, a deeper problem is that the most common methods devised to compare models against neuroimaging data focus on accounting for the common time course of brain activity and model computations. But models, by their very nature, usually make richer predictions about the internal dynamics that lead to either brain activity or behavioral responses. For example, models often make specific assumptions about the *directionality* of an effect, or about how different model components interact with each other. These predictions cannot be tested by simply correlating neuroimaging time series with the order of computations.

In this paper, we describe and demonstrate an alternative and novel method to test models using neuroimaging data. This method is based on patterns of *effective connectivity* between brain regions. "Effective connectivity" is an umbrella term to characterize the functional exchange of information between two brain regions, based on the analysis of their respective time series. Because effective connectivity provides measures of directional communication between two regions, it can be used to examine the internal dynamics of a computational model. Furthermore, because effective connectivity can be estimated from either fMRI or EEG data, it expands the dimensions across which models can be compared with-

out requiring collecting additional data.

In the remainder of this paper, we will outline our method and apply it to a specific, and exquisitely cognitive case, namely, determining which of two prominent computational explanations for the Stroop interference best explains the data.

ACT-R

Although our method could be applied to any computational model, for convenience, it will be demonstrated with two models developed in the Adaptive Control of Thought–Rational (ACT-R) cognitive architecture (Anderson et al., 2004). This choice was made for three reasons. First, ACT-R is the most successful and widespread architecture, having been used in hundreds of publications since its inception, and by far the most popular in the field of cognitive research (Kotseruba & Tsotsos, 2018); Thus, it provides an excellent domain in which to demonstrate the procedure. Second, ACT-R already provides well-tested mappings between architectural components and brain regions with established procedures to predict fMRI activity from model simulations. Therefore, these assumptions can be adopted without the need to provide additional justifications. Finally, the assumptions of ACT-R provide a reasonable mechanism to translate model activity into effective connectivity. As it will be shown, this is based on the functional requirements of the procedural module, which have been examined and discussed in the past.

ACT-R represents knowledge in two formats, *declarative* and *procedural*. Declarative knowledge is made of record-like structures, called *chunks*, which capture semantic memories, perceptual inputs, and motor commands. Procedural knowledge consists of production rules (or simply "productions"), state-action pairs that encode the specific policy to perform a task. In summary, chunks represent information, and productions act upon them.

Chunks are processed by functionally specialized modules. For instance, perceptual modules create new chunks to represent the contents of the outside world, and a memory module maintains chunks in long-term memory. Each module contains one or more buffers, limited-capacity stores that contain at most one chunk. Buffers are the only mechanisms through which chunks and productions interact: Chunks can be inspected, copied, and modified by productions when exposed into buffers.

As noted above, much work has been dedicated to map ACT-R modules to corresponding neural circuits. This work has yielded a number of reliable functional mappings, including the association between anterior cingulate cortex and the goal buffer in the goal module, between the lateral prefrontal cortex and the retrieval buffer of the long-term memory module, between posterior parietal cortex and the imaginal buffer of working memory, between the fusiform gyrus and the visual buffer in the visual module, and between the primary motor cortex and the manual buffer in the motor module (Fincham & Anderson, 2006; Sohn, Albert, Jung, Carter, & Anderson, 2007; Danker, Gunn, & Anderson, 2008; Anderson et al., 2004, 2008). These five modules will be the focus of this paper.

Dynamic Causal Modeling

To estimate effective connectivity, we adopted a framework known as Dynamic Causal Modeling (DCM) (Friston, Harrison, & Penny, 2003). In essence, DCM is procedure to model the time-course of in brain activity in a set of brain regions through a dynamical system of other brain regions and event vectors. Specifically, the time course of activity of a region i is expressed as a bilinear state equation:

$$\dot{\boldsymbol{y}} = \boldsymbol{A}\boldsymbol{y} + \sum_i x_i \boldsymbol{B}(i)\boldsymbol{y} + \boldsymbol{C}\boldsymbol{x} \tag{1}$$

where $\boldsymbol{y}$ are the time series of neuronal activities and $\boldsymbol{x}$ are the time series of the events. $\boldsymbol{A}$ defines intrinsic connectivity between different regions (fixed connectivity), $\boldsymbol{C}$ defines effects by task inputs, and $\boldsymbol{B}$ defines the modular effects that task conditions have on the connectivity between regions (modulation of connectivity).

ACT-R Predictions for Effective Connectivity

Because effective connectivity can be interpreted as directional effects between cortical regions, a direct link can be made between this measure and the nature of ACT-R computations. As discussed above, ACT-R works by firing one production at a time during its cognitive cycle; this production, in turn, changes the state of the system by modifying or copying information from one buffer to the other. For example, in what is perhaps the most common operation in ACT-R models, a production rule extracts values from the slots of chunks placed in either the imaginal or the visual buffer (to extract contextual task information) and places them in the retrieval buffer, so that they function as cues for retrieving relevant information from long-term memory. In fact, production rules are the only way information is exchanged between modules.

Given their role in coordinating module-to-module communication, we made the assumption that patterns of effective connectivity can be derived by the analysis of information transferred carried by out in the sequence of production rules firing.

On the surface, this idea runs against the established identification between production rules (and their associated procedural module) and the activity of the basal ganglia (Anderson et al., 2004; Anderson, 2007). The two interpretations, however, are not incompatible with each other. Anderson et al. (2008) had previously suggested that common functional connectivity patterns in the brain reflect the ubiquity of common operations that exchange information between different buffers; the example production given above is one of those put forward by the authors. It has also been noted before that the function of the basal ganglia is to direct inputs to cortical regions, a role that is both compatible with the procedural module and with the proposed interpretation of effective connectivity (Stocco, Lebiere, & Anderson, 2010). Finally, a recent study that combined ACT-R modeling and Transcranial

Magnetic Stimulation (Rice & Stocco, 2019) has provided evidence that production rules do not only reflect the activity of the basal ganglia but also, more generally, the direct exchange of information between cortical regions. Thus, we believe that the hypothesis that production rules could be used to estimate effective connective is a plausible one.

In this study, the relationship between production rules and effective connectivity was operationalized in the following, simple algorithm. First, an $N \times N$ squared matrix $\boldsymbol{E}$, with N being the number of buffers examined, is generated and initialized to zeros. Then, the target model is run and its trace is segmented into epochs of interest (e.g., all the trials of the same conditions). The structure of each production rule firing within that epoch is then examined. For each variable in the production rule, the *source* buffer S at which the variable is introduced (or, technically, bound to a value) in the left-hand side and the *target* buffer T in which the bound value is placed are recorded. The value of the matrix cell $\boldsymbol{E}_{S,T}$ is then incremented by one. If a variable appears in multiple source buffers $S_1, S_2 \ldots S_N$ or target buffers $T_1, T_2 \ldots T_N$, then all the cells $\boldsymbol{E}_{i \in N, j \in N}$ are updated. When all the productions have been examined, $\boldsymbol{E}$ is taken to represent the predicted effective connectivity for that particular condition.

An Application of the Method: ACT-R Models of the Stroop Task

This method was demonstrated using two competing models of the Stroop task. In the Stroop task, participants are shown a colored character string and asked to report the color of the character string. The character string can either be congruent with the color ("RED" printed in red), incongruent ("BLUE" printed in green), or neutral ("CHAIR" printed in blue). The typical finding is that reaction times in each condition are significantly different from one another, with congruent trials being the fastest, incongruent trials being the slowest, and neutral trials in between (Bugg, McDaniel, Scullin, & Braver, 2011). This difference in reaction times between trial types is referred to as Stroop interference.

The two models were adapted versions of two previously proposed models of the Stroop task, authored by Lovett (2005) and by Altmann and Davidson (2001), respectively. Since both models were published before ACT-R was modified to account for neuroimaging data, they had to be re-implemented in the most recent version of ACT-R (version 7.6). This processes also ensured that the two models interacted with the task using the same sensorimotor mechanisms, i.e. visual objects and responses were given in the same way. From now on, we will refer to these two models as the Altmann-like model and the Lovett-like model.

The re-implemented models maintained the underlying assumptions of their original versions. Specifically, the two models provide different explanations about the nature of Stroop interference. In the Altmann model (Figure 1A), Stroop interference is driven by interference at the lemma layer. When a word is processed, it has direct access to its' lemma, or conceptual representation. Access to the lemma of a color is indirect, requiring an extra retrieval not seen with words. The model assumes that the word dimension of the Stroop stimulus is automatically processed first, therefore activating the lemma attached to the word dimension of the stimulus. As it tries to process the color dimension of the stimulus, the word-lemma is active and can either facilitate or inhibit retrieval of the correct color-lemma. In cases of facilitation, activation from the word-lemma spreads to the coinciding color-lemma, increasing the likelihood of correct retrieval on congruent trials. Oppositely, on incongruent trials, this activation spreads to the incorrect color-lemma, creating increased competition between color-lemmas and introducing ambiguity. For neutral trials, the word-lemma has no corresponding color-lemma, resulting in neither facilitation nor inhibition. The color-lemma is compared to visual cues and re-selected if inconsistent or otherwise used in further processing. A manual response is then retrieved using the color-lemma, and used to press a key on the keyboard.

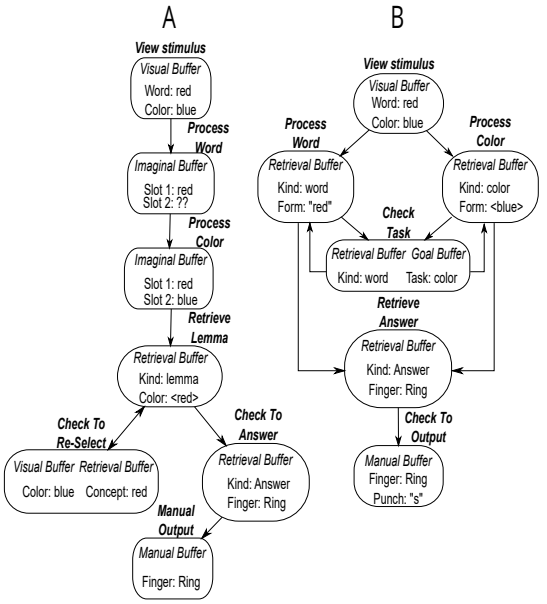


Figure 1: Flow-chart representation of the strategies used by the Altmann model ("A", left) and the Lovett model ("B", right) when processing Stroop trials.

In the Lovett model (Figure 1B), Stroop interference is driven by the competition between alternative *word-association chunks*, linking a word to its' conceptual representation, and *color-association chunks*, linking a color to its' conceptual representation. The idea is similar to lemmas from the Altmann model, but in this case both types of dimension-associated chunks need to be retrieved. The Lovett model accounts for individual differences by supporting various strategies to complete the task. In contrast to the Altmann model, this model allows for processing of either stimulus dimension first, but is highly biased towards the word dimension. From either path, chunks associated with the processed dimension

are retrieved. Processing can maintain with the retrieval of an answer directly, or the task is checked. Answering directly allows for incorrect answers on incongruent trials and fast responses on congruent trials. When the task is checked, the model compares the dimension of the processed chunk to the goal, which for our purpose is to always respond according to the color of the stimulus. If there is a mismatch, processing continues with the alternative stimulus dimension. Now when retrieving the alternative dimension-associated chunk, the previously retrieved chunk has the same effect as in the Altmann model, facilitating retrieval on congruent trials, having no effect on neutral trials, and inhibiting retrieval on incongruent trials. Notably, this does not necessarily happen on every trial, as there are alternate pathways and strategies, and the model will not retrieve the wrong answer at this point. The base-level activations are set in such a way that incongruent chunks slow retrieval of the correct chunk, and congruent chunks facilitate retrieval of the correct chunk. Once the correct dimension-associated chunk is retrieved, a manual answer is retrieved using the matching chunk, and used to press a key on the keyboard.

The two models offer an ideal comparison for several reasons. First, they deal with an experimental paradigm that is representative of research in cognitive neuroscience. Second, although they embody different and opposing views about the nature of Stroop interference, they are equally successful at predicting the canonical response time effects in the Stroop task (Lovett, 2005; Altmann & Davidson, 2001). Most importantly, these two models exemplify the limits of model identification using behavioral and fMRI data. The two models make use of the same five buffers (visual, motor, goal, imaginal, retrieval). When considering the time needed for perceptual and visual processes (identical in the two models), the difference between the two models is concentrated in a 300 ms window in which different interactions between imaginal, goal, and retrieval buffers are posited. Because the BOLD responses recorded in fMRI are much more sluggish and extend for multiple seconds after a point event, it is reasonable to assume that the two models would make almost identical neuroimaging predictions.

To confirm this suspicion, ACT-R's canonical BOLD-response prediction tools were used to simulate the neuroimaging responses for the the various experimental conditions in the two models. Fig 2 illustrate the case for incongruent trials. For the sake of illustration, the amplitudes of the BOLD curves were fit so that they would have the same height[1]. It is immediately apparent that the different inter-module dynamics of the two models are lost in the neuroimaging data; all the BOLD curves for all modules are largely overlapping within and between models.

Crucially, although these different interactions produce indistinguishable BOLD traces, they *do* produce different *effective connectivity* matrices. And, as the next sections will show, these matrices *do* provide evidence in favor of one model over the other.

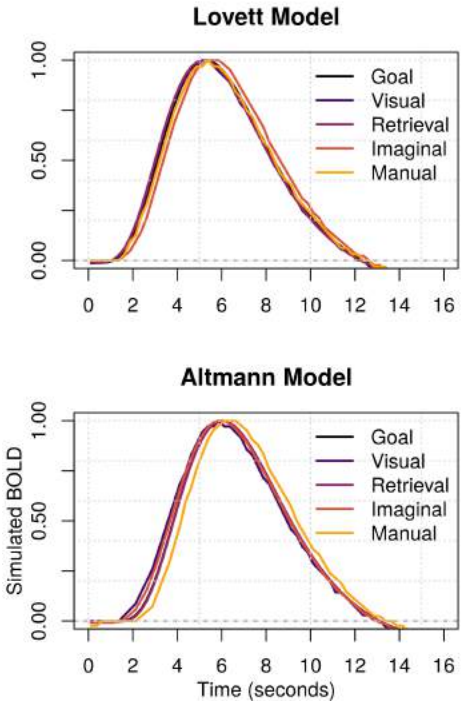


Figure 2: Normalized BOLD-response predictions for incongruent Stroop trials across five different modules in the Lovett (top) and Altmann (bottom) models (See Fig. 1)

Materials and Methods

Experimental Dataset

In this analysis, we used fMRI data publicly available from an open repository[2]. The original data was collected at Carnegie Mellon University and published by Verstynen (2014).

Participants

The dataset contained data from $N = 30$ participants (10 female), aged 21–45 (mean 31). The recruitment procedures can be found in the original publication (Verstynen, 2014).

Experimental Task

Participants performed a manual-response version of the Stroop task (Stroop, 1935), during which the subjects were asked to indicate the color of a written word presented in the center of the screen. Stimuli could be congruent ("RED" printed in red), incongruent ("RED" printed in green), or neutral ("CHAIR" printed in red). Participants responded by indicating the colors red, green, and blue using the right index, middle, and ring fingers, respectively. Each session consisted

[1]The amplitude of the BOLD response is a free parameter that can be separately fit for every module; thus, our procedure does not lose generality

[2]The data is available on OpenNeuro at the following URL: https://openneuro.org/datasets/ds000164/versions/00001

of 120 trials (42 congruent, 42 neutral, 36 incongruent) in randomized order.

Image Acquisition and Preprocessing

As described in Verstynen (2014), the original raw data was acquired using a Siemens Verio 3T system in the Scientific Imaging and Brain Research (SIBR) Center at Carnegie Mellon University with a 32-channel head coil. Functional images were collected using gradient echoplanar pulse sequence with TR = 1,500 ms, TE = 20 ms, and a 90 flip angle. Each volume acquisition consisted of 30 axial slices, each of which was 4 mm thick with 0-mm gap and an in-plane resolution of 3.2 $\times$ 3.2 mm. A T1-weighted structural image was also acquired for each participant in the same space as the functional images, but consisting of 176 1-mm slice with with an in-plane resolution of 1 $\times$ 1 mm.

For the purpose of our analysis, the original raw data was processed in SPM12 (Wellcome Department of Imaging Neuroscience, `www.fil.ion.ucl.ac.uk/spm`) following the exactly same preprocessing pipeline as the one indicated in the original publication. Images were corrected for differences in slice acquisition time, spatially realigned to the first image in the series, normalized to the Montreal Neurological Institute (MNI) ICBM 152 template, resampled to 2 $\times$ 2 $\times$ 2 mm voxels, and finally smoothed with a 8 $\times$ 8 $\times$ 8-mm full-width-at-half-maximum Gaussian kernel to decrease spatial noise and to accommodate individual differences in anatomy.

Regions of Interest

DCM analysis is performed on fMRI time-series extracted from specific ROIs. In our case, the ROIs correspond to the specific brain regions that have been previously identified as corresponding to ACT-R buffers. The Talairach coordinates used for each module in the brain followed the convention used by Anderson et al. (2008). The algorithms described in Lacadie, Fulbright, Rajeevan, Constable, and Papademetris (2008) were used to convert Talairach coordinates to Montreal Neurological Imaging Institute (MNI) coordinates. The ROI mask files were created through FSL (Woolrich et al., 2009) of size 16 mm (125 voxels in total) then used to extract fMRI time series from each voxel in each ROI. Principal Component Analysis was then applied on all the extracted time series to identify the time series that best characterized each ROI. The largest principle component was used to project the original data to the new space with more than 75% of the variance explained in each module.

Dynamic Causal Modeling Analysis

Because DCM is a model-based technique, estimates of connectivity can only derived from parameters corresponding to the specified connectivity between ROIs. To gather complete estimates of connectivity, an unconstrained, fully connected model was generated, in which any ROI was bidirectionally connected to all the others. Furthermore, to identify different patterns of connectivity between conditions, both matrices B and C were used. Specifically, matrix C was used to specify the onset and offset of stimuli, and drive the activity of the "visual" ROI, thus initiating trial-specific activity in the network. In addition, we used the modulatory matrix B to specify modulatory effects of condition-specific trials (congruent, neutral, and incongruent) and the ROI connectivity parameters $\boldsymbol{A}$. Thus, the effective connectivity matrix E_k specific to task condition k can be expressed as the element-wise product of $\boldsymbol{A}$ and the modulatory effects of condition k $\boldsymbol{B}_k$, namely:

$$\boldsymbol{E}_k = \boldsymbol{A} + \boldsymbol{A} \odot \boldsymbol{B}_k \tag{2}$$

As it is common in DCM, all the parameters were identified using an Expectation-Maximization procedure.

Results

Figure 3 illustrates the results of the effective connectivity analysis of the fMRI data and the corresponding model predictions. In the figure, columns correspond to the three experimental conditions of the Stroop task (congruent, incongruent, and neutral trials), while the rows correspond to either the predictions of the models (Lovett model, top row; Altmann model, middle row) or the empirical data (bottom row). The reported values of effective connectivity were generated by performing a Bayesian parameter averaging procedure (Kasess et al., 2010) over the individual connectivity matrices generated for each individual participant. Because, in DCM, self-connectivity values need to be set to negative values to ensure the stability of the dynamic state equation (1), the corresponding values were ignored in the analysis and set to zero in Figure 3. Note that the reason we chose Frobenius norm instead of correlation as the metric is that we are interested in the **absolute** measurement of the effective connectivity, not the relative scale between modules. For example, two connectivity vectors of $[1,1,1,2,1]$ and $[-2,-2,-2,-1,-2]$ would have perfect correlation ($r = 1$), yet they represent opposite connectivity effects (excitatory vs. inhibitory) in all modules. The scale of the values between real fMRI data and ACT-R models may be different, but since all ACT-R models are on the same scale, the differences are still comparable across models.

In general, the connectivity patterns predicted by the two models are much less rich and interconnected than what was measured in the data (Figure 3). This is not unexpected, given the high level of neural abstraction that characterizes ACT-R models (Figure 1). Critically, and as expected, the two models do make different predictions in terms of effective connectivity. To compare the degree of similarity between each model's predictions and the data, we calculated the Frobenius distance of the difference between the predicted ($\boldsymbol{P}$) and the empirical data matrix ($\boldsymbol{D}$) for each condition k, i.e. $||\boldsymbol{P}_k - \boldsymbol{D}_k||_F$. This measure can be interpreted as a dissimilarity metric; the smaller the difference between two matrices, the smaller the norm. The results of these comparisons are shown in Figure 4. As shown, the Lovett model yields consistently smaller norm values, and is therefore more similar to the data, across all three conditions.

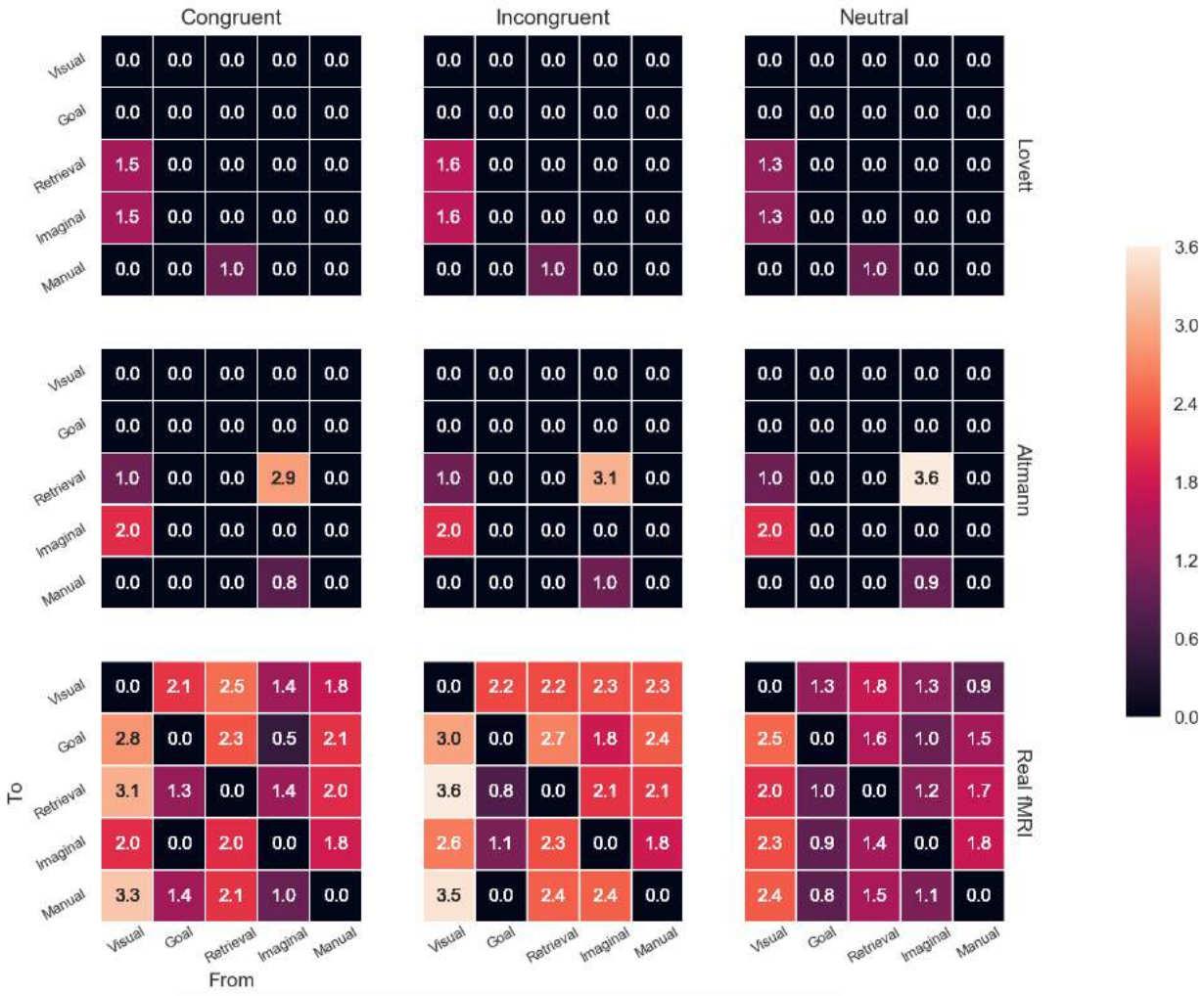


Figure 3: Effective connectivity analysis of the fMRI data and corresponding model predictions (rows), divided by experimental condition (columns).

Discussion

This paper has provided a proof of concept of how analysis of effective connectivity can be used to supplement traditional, GLM-based analysis of neuroimaging data in distinguishing between alternative models. While effective connectivity analysis has been used in cognitive neuroscience for more than a decade, this is the first time, to the best of our knowledge, that this method is used in conjunction with a cognitive modeling approach, and with cognitive architectures in particular. In outlining our method, we choose ACT-R as a modeling paradigm and DCM as a technique to estimate effective connectivity. Neither of these choices, however, are absolute requirements. Connectivity estimates can be gathered from many types of models; the procedure described in this paper certainly applies to other production system-based architecture, like Soar and EPIC, as well. Similarly, although connectivity was estimated with DCM, other methods could be possibly used. For example, Granger Causality. Thus, although we made specific implementation choices, our methods could be instatiated in multiple ways.

Despite encouraging results, a number of limitations need to be acknowledged. First, our method for deriving effective connectivity predictions from ACT-R models is still preliminary. While we believe that it is reasonable, other procedures could be envisioned. For example, operations such as buffer status checks and buffer harvesting could be included in generating our matrices. It is plausible that richer prediction schemes could lead to more realistic connectivity matrices that the ones in Figure 3. It is also plausible that better similarity metrics than Frobenius distance could be used to compare predictions.

These limitations notwithstanding, we see our method as having potential for future modeling research. In particular, we believe that the connectivity matrices obtained from the data can be used to inform model development as well as for model comparison. It is apparent that neither the Lovett nor the Altmann model provide good fits to the data. Because the differences correspond to variables in production rules, the

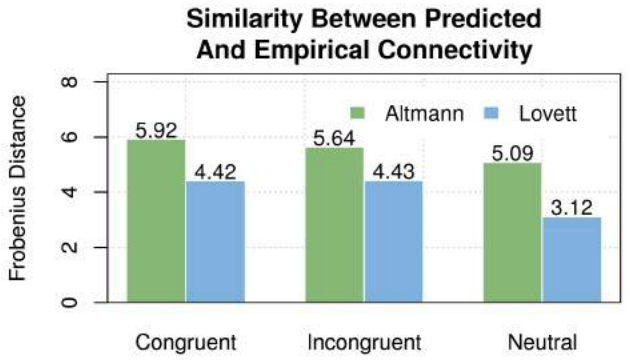


Figure 4: Similarity (Frobenius distance) bewteen predicted and empirical effective connectivity for the Altmann and the Lovett models.

comparison suggests which other production rules or variable bindings could be taking place in the model. In theory, and provided reasonable task constraints, an analysis of the effective connectivity matrices might be used to automatically generate production rules that would match the data. We see this an exciting opportunity for future research.

References

Altmann, E. M., & Davidson, D. J. (2001). An integrative approach to stroop: Combining a language model and a unified cognitive theory. In *Proceedings of the 23rd Annual Meeting of the Cognitive Science Society* (pp. 21–26).

Anderson, J. R. (2007). *How can the human mind occur in the physical universe?* Oxford University Press.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, *111*(4), 1036–1060.

Anderson, J. R., Fincham, J. M., Qin, Y., & Stocco, A. (2008). A central circuit of the mind. *Trends in Cognitive Sciences*, *12*(4), 136–143.

Borst, J. P., Nijboer, M., Taatgen, N. A., van Rijn, H., & Anderson, J. R. (2015). Using data-driven model-brain mappings to constrain formal models of cognition. *PLoS One*, *10*(3), e0119673.

Bugg, J. M., McDaniel, M. A., Scullin, M. K., & Braver, T. S. (2011). Revealing list-level control in the stroop task by uncovering its benefits and a cost. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(5), 1595.

Danker, J. F., Gunn, P., & Anderson, J. R. (2008). A rational account of memory predicts left prefrontal activation during controlled retrieval. *Cerebral Cortex*, *18*(11), 2674–2685.

Fincham, J. M., & Anderson, J. R. (2006). Distinct roles of the anterior cingulate and prefrontal cortex in the acquisition and performance of a cognitive skill. *Proceedings of the National Academy of Sciences*, *103*(34), 12941–12946.

Friston, K., Harrison, L., & Penny, W. (2003). Dynamic Causal Modelling. *NeuroImage*, *19*(4), 1273–1302.

Kasess, C. H., Stephan, K. E., Weissenbacher, A., Pezawas, L., Moser, E., & Windischberger, C. (2010, feb). Multi-subject analyses with dynamic causal modeling. *NeuroImage*, *49*(4), 3065–3074.

Kotseruba, I., & Tsotsos, J. K. (2018). 40 years of cognitive architectures: core cognitive abilities and practical applications. *Artificial Intelligence Review*, 1–78.

Lacadie, C. M., Fulbright, R. K., Rajeevan, N., Constable, R. T., & Papademetris, X. (2008, aug). More accurate talairach coordinates for neuroimaging using non-linear registration. *NeuroImage*, *42*(2), 717–725.

Lovett, M. C. (2005). A strategy-based interpretation of stroop. *Cognitive Science*, *29*(3), 493–524.

Pitt, M. A., Kim, W., Navarro, D. J., & Myung, J. I. (2006). Global model analysis by parameter space partitioning. *Psychological Review*, *113*(1), 57.

Rice, P., & Stocco, A. (2019). The role of dorsal premotor cortex in resolving abstract motor rules: Converging evidence from Transcranial Magnetic Stimulation and cognitive modeling. *Topics in cognitive science*, *11*(1), 240–260.

Sohn, M.-H., Albert, M. V., Jung, K., Carter, C. S., & Anderson, J. R. (2007). Anticipation of conflict monitoring in the anterior cingulate cortex and the prefrontal cortex. *Proceedings of the National Academy of Sciences*, *104*(25), 10330–10334.

Sohn, M.-H., Goode, A., Koedinger, K. R., Stenger, V. A., Fissell, K., Carter, C. S., & Anderson, J. R. (2004). Behavioral equivalence, but not neural equivalenceneural evidence of alternative strategies in mathematical thinking. *Nature Neuroscience*, *7*(11), 1193.

Stocco, A., Lebiere, C., & Anderson, J. R. (2010). Conditional routing of information to the cortex: A model of the basal ganglias role in cognitive coordination. *Psychological Review*, *117*(2), 541.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*(6), 643–662.

van Vugt, M. K. (2014). Cognitive architectures as a tool for investigating the role of oscillatory power and coherence in cognition. *NeuroImage*, *85*, 685–693.

Verstynen, T. D. (2014, nov). The organization and dynamics of corticostriatal pathways link the medial orbitofrontal cortex to future behavioral responses. *Journal of Neurophysiology*, *112*(10), 2457–2469.

Woolrich, M. W., Jbabdi, S., Patenaude, B., Chappell, M., Makni, S., Behrens, T., ... Smith, S. M. (2009, mar). Bayesian analysis of neuroimaging data in FSL. *NeuroImage*, *45*(1), S173–S186.

Modeling individual performance in cross-situational word learning

Yung Han Khoe (y.khoe@student.ru.nl)
Centre for Language Studies
Radboud University Nijmegen

Amy Perfors (amy.perfors@unimelb.edu.au)
School of Psychological Sciences
University of Melbourne

Andrew T. Hendrickson (a.hendrickson@tilburguniversity.edu)
Department of Cognitive Science & Artificial Intelligence
Tilburg University

Abstract

What mechanisms underlie people's ability to use cross-situational statistics to learn the meanings of words? Here we present a large-scale evaluation of two major models of cross-situational learning: associative (Kachergis, Yu, & Shiffrin, 2012a) and hypothesis testing (Trueswell, Medina, Hafri, & Gleitman, 2013). We fit each model individually to over 1500 participants across seven experiments with a wide range of conditions. We find that the associative model better captures the full range of individual differences and conditions when learning is cross-situational, although the hypothesis testing approach outperforms it when there is no referential ambiguity during training.

Keywords: Cross-situational word learning; language acquisition; Zipfian distributions

Introduction

The ability to acquire language is not only a fundamental part of what makes us human, but a mystery: how do we accomplish it given the complexity of the learning task? Even in an apparently simple task like word learning, many real-world contexts involve multiple possible referents for any one label (Pinker, 1984). How can a learner figure out which referent corresponds to which label? One suggestion is that people can leverage the statistics that come from observing multiple ambiguous presentations of words and objects. This sort of cross-situational word learning has been demonstrated in both children and adults (Yu & Smith, 2007; L. Smith & Yu, 2008). However, there is still considerable debate about what mechanisms underlie cross-situational word learning and what representations are learned (Kachergis & Yu, 2018).

One major theory of cross-situational learning, known as the associative framework, proposes that people track detailed word-object co-occurrence statistics across many presentations (Vouloumanos, 2008; Yu & Smith, 2007). By contrast, the hypothesis testing framework suggests that people track at most one word-object pair theory for each word (or object) and update these hypotheses during learning (Medina, Snedeker, Trueswell, & Gleitman, 2011). A number of computational models have been developed based on both frameworks but no consensus has emerged about which account better describes people's learning. We argue that this has occurred, at least in part, because of a focus on modeling aggregate rather than individual data, and because existing experiments have not varied the range and variety of learning conditions sufficiently to differentiate the models.

Here we present and analyze data from seven different experiments with over 1500 participants that vary on a number of factors including vocabulary size, level of ambiguity, length of training, distributional structure, and task. We fit each person's data to both the associative and hypothesis testing models described in the following sections. Our results suggest that associative accounts provide the best fit in almost all cases, unless there is no ambiguity during learning and the learning is thus no longer strictly cross-situational.

Associative framework

Associative models propose that people learn word meanings by tracking the frequency with which words and objects co-occur across multiple ambiguous presentations. The representation is a large word-object matrix in which each cell contains the associative strength between one word and one object (Vouloumanos, 2008; Yu & Smith, 2007). This basic framework has been applied widely, and the model we implement here is one of the most widely used (Kachergis et al., 2012a). It provides a compelling account of human behavior across studies that vary the number of late repetitions (Kachergis et al., 2012a) and if learning is passive or active (Kachergis, Yu, & Shiffrin, 2012b; Kachergis & Yu, 2018).

Formally, the goal of the model is to update the association strength between a word (w) and object (o) within an association matrix ($M_{w,o}$). It incorporates several psychologically-motivated parameters that specify the total amount of updating on each trial (χ), memory fidelity (α), and the bias towards updating the association strength of uncertain versus already familiar words and objects (λ) with uncertainly of an item quantified as the entropy across all association strengths for that item.

Hypothesis testing framework

As an alternative to the memory-intensive associative framework, Medina et al. (2011) outlined a more minimalistic approach based on storing only a single hypothesis for each word. The hypothesis represents a guess about the referent of the word, and is replaced if it is inconsistent with new training trials or fails to be recalled when the word is present.

Vocabulary	Ambiguity	Guessing	Presentations	Distribution	Length Relationship	N	Source
12	3	Yes	108	Uniform	Only one syllable	48	H&P (2018) Exp 1
12	3	Yes	108	Zipfian	Only one syllable	72	H&P (2018) Exp 1
32	4	Yes	244	Uniform	Uniform	79	H&P (2018) Exp 2
32	4	Yes	244	Zipfian	Correlated	81	H&P (2018) Exp 2
32	4	Yes	244	Zipfian	Random	80	H&P (2018) Exp 2
32	1	NA	244	Uniform	Uniform	74	H&P (2018) Exp 3
32	1	NA	244	Zipfian	Correlated	77	H&P (2018) Exp 3
32	1	NA	244	Zipfian	Random	86	H&P (2018) Exp 3
28	4	Yes	240	Uniform	Uniform	171	Exp 1
28	4	Yes	240	Zipfian	Random	166	Exp 1
40	4	Yes	240	Uniform	Uniform	71	Exp 2
40	4	Yes	240	Zipfian	Random	90	Exp 2
28	4	No	240	Uniform	Uniform	82	Exp 3
28	4	No	240	Zipfian	Correlated	84	Exp 3
28	1	NA	240	Uniform	Uniform	159	Exp 4
28	1	NA	240	Zipfian	Correlated	151	Exp 4

Table 1: **Overview of experimental structure**. This table describes all of the experiments whose data we fit. *Vocabulary* indicates the number of unique word-object pairs to be learned (which also corresponds to the number of objects present during the test phase). *Ambiguity* indicates the number of objects present on each training screen. *Guessing* indicates whether learning was passive (just watching) or active (if participants were required to submit a guess after each word during training). *Presentations* indicates the total number of training trials. *Distribution* indicates the frequency distribution of the words and objects across the experiment. *Length Relationship* indicates the relationship between the length of words and their frequency during training, with more frequent words being shorter in the *Correlated condition*. *N* indicates number of complete participants. *Source* indicates the source of the data set: H&P (2018) denotes Hendrickson and Perfors (2018).

We evaluate the Propose-but-Verify hypothesis testing model (Trueswell et al., 2013), a popular extension of the original Medina et al. (2011) formulation, which captures children's word learning behavior well (Woodard, Gleitman, & Trueswell, 2016; Aravind et al., 2018). The model involves a two-stage process. Upon initially being exposed to a word, the model chooses an object from the as-yet-unmapped objects in that trial and maps it to that word to form a word-object hypothesis. The initial probability of later recalling that mapping is denoted by a free parameter $\alpha_{initial}$. On each subsequent exposure to the word, if the model recalls the hypothesis and the corresponding object is present, the probability is updated to a different memory strength indicated by another free parameter, $\alpha_{confirmed}$. If the hypothesis fails to be recalled or the corresponding object is not present, a new hypothesis is established with an unmapped object.

Model comparisons

Many previous papers have compared these two modeling approaches in terms of how well they fit experimental data (e.g., K. Smith, Smith, & Blythe, 2009; Kachergis et al., 2012b; Rasilo & Räsänen, 2015; Kachergis & Yu, 2018; Aussems & Vogt, 2018; Stevens, Gleitman, Trueswell, & Yang, 2017). Despite this effort, no consensus has emerged. One reason may be the focus on modeling aggregate performance using one optimal set of parameter values per model for all learners, which ignores individual differences. This approach may favor highly stochastic models that can fit different people's responses with a single parameter, rather than models that can fit the behavior of more people using individual parameter values. Moreover, comparison studies commonly fit these models to experiments that involve relatively few learners, and have a small number of conditions which do not capture the variation across conditions in the literature. Finally, such studies tend to use uniform word frequencies that do not reflect the highly-skewed distribution of words in natural language, which limits the generalizability to real-world word learning (Hendrickson & Perfors, 2018).

In this paper we address these issues by evaluating a hypothesis testing model and an associative learning model against experimental data involving over 1500 participants and spanning the broad range of conditions shown in Table 1. We varied the distribution of the words and objects, the size of the vocabulary to be learned, whether the task was passive or active, the number of presentations during training, the level of ambiguity during learning, and the relationship between the length of the word and word frequency. We fit parameter values for each learner by optimizing the log-likelihood of model response probability for each of the word-object test trials. When comparing models, we penalize for additional parameters by converting the log likelihood to AIC values (Akaike, 1974).

Experiments

The empirical data that we use for model evaluation includes data from the eight conditions from Hendrickson and Perfors (2018) in addition to eight additional new conditions. We describe each in turn.

Hendrickson and Perfors (2018)

The goal of the work in Hendrickson and Perfors (2018) was to explore cross-situational learning when the words followed either a ZIPFIAN or a UNIFORM distribution. The first experiment involved presenting participants with a small vocabulary of words in one of the two distributions. The second increased the vocabulary and ambiguity level, while adding a condition in which the length of the word was negatively correlated with word frequency (shorter words were more frequent). The third evaluated the effect of removing ambiguity during training.

Procedure. Each experiment consisted of a training phase and a test phase, though Experiment 1 repeated these phases multiple times. During training, participants viewed either 3, 4, or 1 objects on the screen at once while they heard the words for each object presented one at a time in random order. In experiments with ambiguity during training, participants were asked to guess which object each word matched. At test, people were shown all of the items at once and asked to select the matching object for each word. They were not given feedback during training or test.

Conditions. In the UNIFORM conditions the words and objects all occurred with the same frequency, while in the ZIPFIAN conditions a few words and objects occurred very frequently and many words occurred very infrequently or only once across training. The pairing of words and objects and trial order was randomized across participants.

Materials. Words varied in length from one to three syllables and were designed to sound English-like as well as be maximally distinct from each other. They were generated by the AT&T Natural Voices Text-to-Speech tool (Crystal voice). The objects were selected from the NOUNS image corpus (Horst & Hout, 2015) and each image was 150x150 pixels displayed against a white background. Hendrickson and Perfors (2018) contains the full set of stimuli.

Participants. Their 597 participants were recruited from Amazon Mechanical Turk (AMT). Our four additional experiments (with 974 participants) were also run on AMT, paying US$3.25 for the $\sim$20 minute task.

Experiment 1

Experiment 1 provides a near replication of Experiment 2 of Hendrickson and Perfors (2018), a design aimed to approximate learning conditions when the meaning of words is ambiguous. There were two minor differences. First, their experiment included four single-presentation items in order to check for participant cheating; we omitted those in order to ensure that the UNIFORM distribution contained no low-frequency items. As a result, we had 240 rather than 244 total presentations and 28 rather than 32 test items. Second, we did not include their second ZIPFIAN condition, in which the length of the word and word frequency was correlated. 337 individuals provided complete data, half in the UNIFORM condition and half in the ZIPFIAN condition.

Experiment 2

A number of simulation studies have suggested that increasing the number of items to be learned should be particularly challenging for learners in Zipfian environments (Vogt, 2012; Reisenauer, Smith, & Blythe, 2013). In Experiment 2 we therefore replicated the design of Experiment 1 but with 40 unique word-object pairs instead of 28. We presented each word slightly less frequently in order to match the total number of word-object presentations in Experiment 1. In the test phase 40 rather than 28 objects were displayed, resulting in a more difficult test. Complete data was collected from 161 individuals, with roughly half assigned randomly to the UNIFORM and ZIPFIAN conditions.

Experiment 3

In all of the experiments so far, participants have been required to respond by selecting a best-guess object after each word was presented during training. However, recent work has suggested that forcing people to guess may influence the representation that they learn (Aussems & Vogt, 2018). We address this possibility in Experiment 3, which is identical to Experiment 1 but removes the obligation to guess during training. Instead of waiting for a guess after each word, the next word is played automatically after 2000 ms.

The other difference from Experiment 1 is that the length of each word was correlated with its frequency in the ZIPFIAN condition, as is found in natural language and Hendrickson and Perfors (2018). Complete data was collected from 166 individuals, with roughly half assigned randomly to the UNIFORM and ZIPFIAN conditions.

Experiment 4

Experiment 4 provides a near replication of Experiment 3 of Hendrickson and Perfors (2018), whose goal was to approximate learning when the meaning of words was unambiguous. The only differences, as in Experiment 1, were that we removed the four "cheating check" items and thus had 240 presentations and 28 test items, and we had only one ZIPFIAN condition in which word length was correlated with frequency. Additionally, since there was no ambiguity during training, participants were not required to guess. Instead, the timing of item presentation matched Experiment 3. Complete data was collected from 310 individuals, with roughly half assigned to the UNIFORM and ZIPFIAN conditions.

Model Fitting

Both models were fit to the individual data of each person independently by minimizing the negative log likelihood across all responses in the test phase. Every person participated in exactly one condition and thus parameters were not constrained across conditions in any way. For the associative model, the likelihood of a correct answer was determined for each word by dividing the associative mass on the correct object by the total associative mass across all objects. For

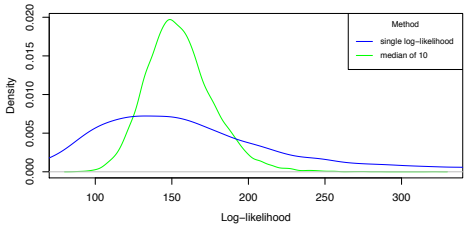


Figure 1: **Distribution of log-likelihood vs. median log-likelihood values for the hypothesis testing model.** The density plot shows the different distributions of 10,000 simulations produced by the hypothesis testing model for one individual participant with a constant set of parameters when using either a single log-likelihood value or the median of 10 such values. Using the median of 10 values results in a substantial increase in stability.

the hypothesis testing model, the likelihood was given by the stored probability value for a correct pairing, smoothing zero probability values to 0.0001.

Interestingly, the models differed widely from each other in the variability of the responses probabilities predicted at test. Given a fixed training trial order and a single set of parameter values, the associative model has no stochastic aspect to how the representation is formed. Therefore, the likelihood of a set of responses given a set of parameters is stable and parameter estimation was straightforward.[1]

In contrast, the hypothesis testing model is decidedly random about which words are paired with objects when forming hypotheses. This results in the production of markedly different representations and thus likelihoods from one simulation to the next, even when the training trials and parameter values are constant across runs. In order to address this issue, we performed ten simulations for each set of parameters during the optimization process and used the median log likelihood across the simulations. This required the use of a particle swarm optimization algorithm to determine the optimal parameters, which is more robust to less smooth optimization problems.[2] It was notable that across the range of 10 likelihood values for a set of parameters, the best value was markedly better than the median likelihood value, suggesting that optimization routines that rely on the best likelihood given a set of parameters can overestimate the expected fit to data of a set of parameters, especially for the hypothesis testing model. In addition, the median likelihood from 10 simulations of the hypothesis testing model produces considerably more stable estimates relative to a single simulation (Figure 1).

The number of parameters differ between the two models, with the associative model containing three (α, χ and λ) and the hypothesis testing model containing only two ($\alpha_{initial}$ and $\alpha_{confirmed}$). We therefore penalized for model complexity by converting the log likelihood scores to AIC values; lower AIC scores indicate a better fit to the data after taking the number of free parameters into account.

[1]The optimal parameter values were derived using the default settings for the optimize function in the SciPy package in Python 3.

[2]Fitting was done using the PSO package in Python 3 using a swarm size of 1,000 and a maximum of 50 iterations.

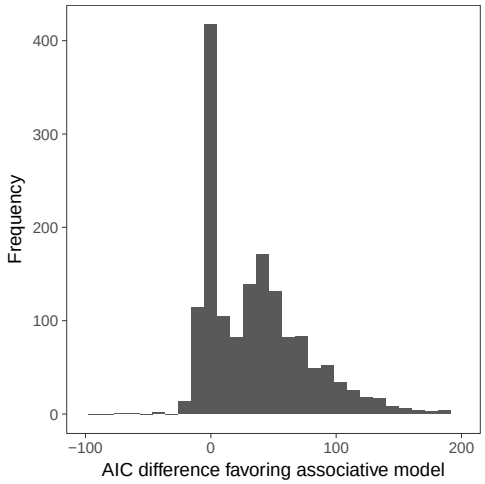


Figure 2: **Overall model performance.** The histogram shows the difference between the AIC score for the hypothesis testing model and the AIC score for the associative model for each person across all datasets. Positive scores (which made up 74% of the data) indicate that the associative model had a lower AIC score than the hypothesis testing model and thus accounted for the data better.

Results

Figure 2 shows the overall performance of the two models across all individuals. The AIC scores of the associative model are lower than the hypothesis testing model for 74% of participants. Since a lower AIC indicates better fit, this suggests that for most people the associative model provides a better account of their performance. Next, we turn to exploring exactly when and where each model does best.

Ambiguity. As the top row of Figure 3 illustrates, the performance of the two models strongly depends on the degree of ambiguity during training. In conditions with any degree of ambiguity during training (by presenting three or four items on each training screen, rather than individually), the associative model is a better fit in virtually all cases (97% of participants). However, the opposite is true when there is no ambiguity: when only one item was shown at a time, the hypothesis testing model was favored for 68% of participants.

Word frequency. The near unanimous advantage for the associative model in ambiguous learning conditions suggests that the only differences in model performance due to word frequency might occur in the conditions without ambiguity. In the conditions with only one item per screen (top left panel of Figure 3), the hypothesis testing model (red points) is highly preferred (96% of the time) when the distribution is UNIFORM. When the word frequency distribution is ZIPFIAN (blue points), the two models are roughly even (52% of participants are better fit by the associative model).

Vocabulary size. The impact of vocabulary size on model performance differs based on the ambiguity of word meaning during training (bottom row, left and center panels in Figure 3). When word meaning is ambiguous during training, the hypothesis testing model does particularly poorly as the vocabulary size increases. (Note that we do not show the 3-item

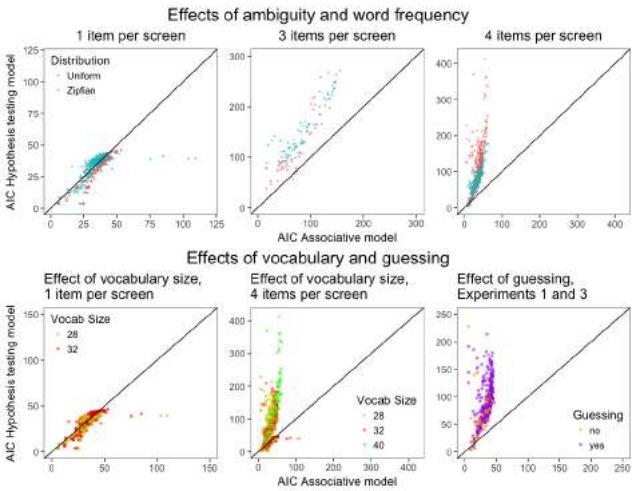


Figure 3: **Model comparisons for different effects.** All panels show the AIC scores for the hypothesis testing model (y axis) plotted against the AIC scores of the associative model (x axis). Values above the identity line indicate that the AIC score of the associative model is better (lower). *Top:* A comparison of the effects of ambiguity and word frequency. Each red dot shows a participant in the UNIFORM condition; blue dots represent people from the ZIPFIAN condition. Each panel shows a different level of ambiguity during training (1, 3, or 4 items on screen at once). The associative model does much better whenever there is ambiguity (and regardless of distribution), while the hypothesis testing model does slightly better when there is only one item on screen during learning. *Bottom left and center:* Evaluation of the effects of vocabulary size, broken down by degree of ambiguity. When training is unambiguous, there is no consistent effect of vocabulary size on performance; when it is ambiguous, the hypothesis training model appears to perform especially poorly when there are more words to be learned. *Bottom right:* Regardless of whether participants were passive or active learners, the AIC favored the associative model.

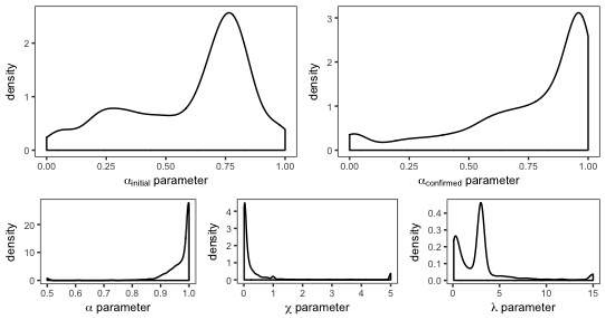


Figure 4: **Distribution of best fitting parameters across all participants**. *Top row*: Hypothesis testing model parameters $\alpha_{initial}$ (initial probability of recalling a mapping) and $\alpha_{confirmed}$ (later probability of recalling a mapping). *Bottom row*: Associative model parameters α (memory decay rate), χ (amount of updating for each trial), and λ (bias towards updating uncertain words and objects).

ambiguous case because vocabulary size in those conditions was constant at 12 items). Within the unambiguous training conditions there does not appear to be any consistent effect of vocabulary size on model performance.

Guessing. In order to evaluate the prediction that forcing participants to guess during training can bias participants to adopt hypothesis testing representations (Aussems & Vogt, 2018), the bottom right panel of Figure 3 shows the performance of both models as a function of whether participants had to guess or not. Here we include only data from Experiments 1 and 3, similar experiments that differ in whether guessing occurs. Both have a vocabulary size of 28 and ambiguous training (although the relationship between word length and word frequency differs between the two ZIPFIAN conditions). Across all conditions the associative model consistently outperforms the hypothesis testing model.

Fitted Parameters

In addition to being useful for model comparison, the best-fitting parameters across all participants (shown in Figure 4) provide us with several deeper insights. First, the distribution of parameter values can tell us something about the distribution of individual differences across the population. For example, the distributions of α and χ for the associative model are highly skewed and show ceiling and floor effects, which suggest these parameters might not capture meaningful variation across individuals. By contrast, the distributions of the memory strength parameters for the hypothesis testing model both display a unimodal peak around relatively high values with a long tail of low values for some participants. This suggests a high level of population variance or reflects the inherent stochasticity of the hypothesis testing model.

It is also useful to compare our best-fit parameters to the reported values from other studies, which generally fit aggregate data or use other fitting metrics. The distribution of our fitted values for the two hypothesis testing model parameters are generally higher than those reported by Trueswell et al. (2013) in their two experiments: $\alpha_{initial}$ values of 0.26 and 0.60, and $\alpha_{confirmed}$ values of 0.71 and 0.81. On average, our best-fit parameter values were higher and show less difference between the initial and confirmed memory strength. It remains an open question if this difference is due to modeling individual and aggregate performance or a shift in strategy due to experimental conditions.

In contrast to the hypothesis testing model, the values reported for the associative model by Kachergis et al. (2012b) are largely consistent with our results. Their optimized values, fit to aggregate data, are $\alpha = 0.97$, $\chi = 0.05$, and $\lambda = 1.74$; values quite similar to the peaks of our distributions.

Finally, some model parameters strongly depend on the experimental condition. For example, the multimodal distribution of λ values (bottom right panel of Figure 4) suggests a mixture of different strategies across participants. We investigate this in Figure 5, which separates the best-fit λ values according to the ambiguity during training. It is evident that as learning conditions are more ambiguous, the λ value decreases. Since λ affects the weight assigned to novel words relative to familiar words, one interpretation of this is that the level of ambiguity during training has a strong impact on the extent to which novel items are emphasized during learning.

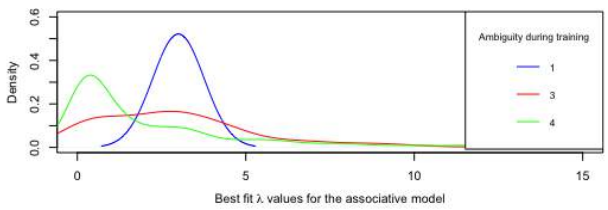


Figure 5: **Distribution of λ by ambiguity.** The best-fit λ values for the associative model (x axis) are plotted as a function of the level of ambiguity during training. The distribution of λ when there is no ambiguity (blue) has higher average values. As ambiguity increases (red, then green) the estimated λ values get smaller.

Discussion

In this work we investigated which of two computational models of cross-situational word learning offers a better account of word learning by individual participants across a wide range of conditions. For most people, the associative model (Kachergis et al., 2012a) outperforms the hypothesis testing model (Trueswell et al., 2013).

The advantage for the associative model is most pronounced in conditions in which the meaning of words is ambiguous during training, where it provides a better account for nearly all people. However, in conditions without ambiguity of word meaning, the hypothesis testing model outperformed the associative model for over 60% of participants. This advantage for the hypothesis testing model in unambiguous training conditions occurred for nearly every participant who experienced a uniform frequency distribution of words, but for participants in conditions with a Zipfian word frequency distribution the associative and hypothesis testing models provide the best account equally often.

The impact of other aspects of the learning environment on the relative performance of the two models was less striking. The total number of unique words present did not seem to influence which model was preferred, though there was some suggestion that the participants whose AIC was worst for the hypothesis testing model were in the conditions with the largest vocabulary size. Finally, manipulating if participants were required to guess during training had no effect on model preference as all relevant conditions were ambiguous during training and thus nearly all participants were best fit by the associative model.

Why the hypothesis testing model, despite multiple studies showing support for the model, performed consistently worse in the ambiguous learning contexts that require cross-situational learning is perhaps the biggest open question raised by these results. One possibility is that the hypothesis testing model, though designed to account for individual learning behavior, is not sufficiently flexible to account for the variation across participants. Restricting the memory strength to two possible values might provide a good account of aggregate data but be too rigid for matching individual behavior.

Another possible explanation for the worse performance of the hypothesis testing model is that even if people do form hypotheses about word-object pairs, they are also incorporating some co-occurrence information to shape their representations. This class of hybrid learning mechanisms, which incorporate both hypothesis testing and associative learning mechanisms (Yurovsky & Frank, 2015), provide a suggestion of additional types of models that might better capture the range of learning behavior in the ambiguous conditions. Similarly, Pursuit (Stevens et al., 2017), a recent variant of the Propose-but-Verify model that retains disconfirmed meanings and counts of referential success, might also improve on the performance of the earlier hypothesis testing model by finding a balance between testing hypotheses and gathering some co-occurrence information.

A final explanation of this effect may be due to specific aspects of the model fitting in this study. These choices include how the hypothesis testing model was extended to produce probability distributions across responses, the 10-fold simulation of parameter values to compute the median log likelihood, or the choice of AIC for model comparison instead of measures that have higher penalties for model complexity (e.g. BIC) or flexibility (Navarro, Pitt, & Myung, 2004).

Despite the clear advantage across many conditions for one model in this comparison, further work is clearly needed to fully understand the learning mechanisms and representations that underlie word learning. These include evaluating alternative models (e.g. Yu & Smith, 2012; Yurovsky & Frank, 2015; Stevens et al., 2017), expanding the range of evaluation techniques, and constraining models with additional data (e.g. Kachergis & Yu, 2018) or conditions (e.g. Hendrickson & Perfors, 2018).

References

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, *19*(6), 716–723.

Aravind, A., de Villiers, J., Pace, A., Valentine, H., Golinkoff, R., Hirsch-Pasek, K., ... Wilson, M. S. (2018). Fast mapping word meanings across trials: Young children forget all but their first guess. *Cognition*, *177*, 177–188.

Aussems, S., & Vogt, P. (2018). Adults use distributional statistics for word learning in a conservative way. *IEEE Transactions on Cognitive and Developmental Systems*.

Hendrickson, A., & Perfors, A. (2018, Nov). *Cross-situational learning in a zipfian environment.* PsyArXiv. Retrieved from `psyarxiv.com/6jumv`

Horst, J., & Hout, M. (2015). The novel object and unusual name (NOUN) database: A collection of novel images for use in experimental research. *Behavior Research Methods*, *48*(4), 1393-1409.

Kachergis, G., & Yu, C. (2018). Observing and modeling developing knowledge and uncertainty during cross-situational word learning. *IEEE Transactions on Cognitive and Developmental Systems*, *10*(2), 227–236.

Kachergis, G., Yu, C., & Shiffrin, R. M. (2012a). An associative model of adaptive inference for learning word–

referent mappings. *Psychonomic Bulletin & Review*, *19*(2), 317–324.

Kachergis, G., Yu, C., & Shiffrin, R. M. (2012b). Cross-situational word learning is better modeled by associations than hypotheses. *IEEE Conference on Development and Learning*, 1–6.

Medina, T. N., Snedeker, J., Trueswell, J. C., & Gleitman, L. (2011). How words can and cannot be learned by observation. *PNAS*, *108*, 9014–9019.

Navarro, D. J., Pitt, M. A., & Myung, I. J. (2004). Assessing the distinguishability of models and the informativeness of data. *Cognitive Psych.*, *49*(1), 47–84.

Pinker, S. (1984). *Language learnability and language development*. Cambridge, MA: Harvard University Press.

Rasilo, H., & Räsänen, O. J. (2015). Computational evidence for effects of memory decay, familiarity preference and mutual exclusivity in cross-situational learning. In *CogSci.*

Reisenauer, R., Smith, K., & Blythe, R. (2013). Stochastic dynamics of lexicon learning in an uncertain and nonuniform world. *Physics Review Letters*, *110*(258701).

Smith, K., Smith, A. D., & Blythe, R. A. (2009). Reconsidering human cross-situational learning capacities: A revision to yu & smiths (2007) experimental paradigm. In *CogSci.*

Smith, L., & Yu, C. (2008). Infants rapidly learn word/referent mappings via cross-situational statistics. *Cognition*, *106*, 1558–1568.

Stevens, J. S., Gleitman, L. R., Trueswell, J. C., & Yang, C. (2017). The pursuit of word meanings. *Cognitive Science*, *41*, 638–676.

Trueswell, J. C., Medina, T. N., Hafri, A., & Gleitman, L. (2013). Propose but verify: Fast mapping meets cross-situational learning. *Cognitive Psych.*, *66*, 126–156.

Vogt, P. (2012). Exploring the robustness of cross-situational learning under Zipfian distributions. *Cognitive Science*, *36*, 726–739.

Vouloumanos, A. (2008). Fine-grained sensitivity to statistical information in adult word learning. *Cognition*, *107*, 729–742.

Woodard, K., Gleitman, L. R., & Trueswell, J. C. (2016). Two-and three-year-olds track a single meaning during word learning: Evidence for propose-but-verify. *Language Learning and Development*, *12*(3), 252–261.

Yu, C., & Smith, L. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psych. Science*, *18*, 414–420.

Yu, C., & Smith, L. (2012). Modeling cross-situational word-referent learning: Prior questions. *Psych. Review*, *119*(1), 21–39.

Yurovsky, D., & Frank, M. C. (2015). An integrative account of constraints on cross-situational learning. *Cognition*, *145*, 53–62.

A Unified Model of Fatigue in a Cognitive Architecture: Time-of-Day and Time-on-Task Effects on Task Performance

Ehsan B. Khosroshahi[a] (ehsanebk@drexel.edu)
Dario D. Salvucci[a] (salvucci@drexel.edu)
Glenn Gunzelmann[b] (glenn.gunzelmann@us.af.mil)
Bella Z. Veksler[c] (bellav717@gmail.com)

[a] Department of Computer Science, Drexel University, 3675 Market St.
Philadelphia, PA 19104, United States

[b] Warfighter Readiness Research Division, Air Force Research Laboratory, 2620 Q St.
Wright Patterson Air Force Base, OH 45433, United States

[c] Tier1 Performance Solutions, 100 E. Rivercenter Blvd., Suite 100
Covington, KY 41011, United States

Abstract

Capturing the effects of fatigue and, more generally, the effects of physical and mental states on human performance has been a topic of research for many years. Recent models, especially those developed in a cognitive architecture, have shown great promise in capturing these effects by providing insight into the specific cognitive and other components involved in task performance (like perception and motor movement). In particular, separate models have been developed to account for both time-of-day and time-on-task effects related to fatigue. In this paper, we present a novel unified model, developed in the ACT-R cognitive architecture, that captures both time-of-day and time-on-task effects with a single set of mechanisms and parameters. We demonstrate how this unified model accounts for quantitative and qualitative aspects of fatigued performance from two experiments, one focused on time-on-task effects under conditions of moderate fatigue, the other focusing on time-of-day effects under conditions of severe fatigue in a study of long-term (88-hour) sleep deprivation.

Keywords: Fatigue; sleep deprivation; cognitive architectures

Introduction

One of the most significant physiological states that affects human cognition is fatigue. Decades of research have investigated the effects of fatigue, sleep deprivation, and time-on-task in a number of important areas, including industrial disasters (e.g. Mitler et al.,1988), transportation accidents (e.g. Lauber & Kayten, 1988; Dinges, 1995), and motor vehicle crashes (e.g. Horne & Reyner, 1999; Pack et al., 1995). These studies have explored in depth the question of how fatigue modulates cognition and performance, and how we might quantify the effects of fatigue using mathematical or computational models and formalisms.

Of the many aspects of cognitive fatigue, there are two main factors that affect sustained attention and task performance: (1) sleep-related factors which are a function of sleep history and the time of the day when the task is being performed (circadian rhythm); and (2) task-related factors which are a function of the type of the task and how long the person has been doing the task, or *time-on-task* (Figure 1). Fatigue can also vary widely in its level of intensity: mild to moderate time-of-day or time-on-task effects may affect performance significantly (e.g., Pattyn et al, 2008; Bakan, 1955; Mackworth, 1948; Parasuraman, 1979), but severe fatigue that occur with long-term sleep deprivation can have even more drastic impacts on performance (e.g., Doran, Van Dongen, Dinges, 2001; Dorrian, Rogers, & Dinges, 2005).

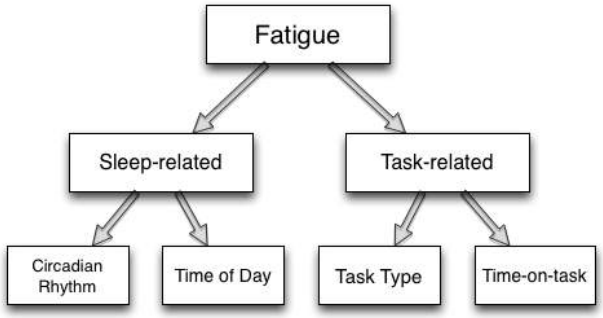


Figure 1: Main factors contributing to fatigue in sustained-attention tasks.

Mathematical models of fatigue come in various forms, and can provide very good insight into the fluctuations in overall performance, accounting for moderate time-of-day and time-on-task effects (e.g. Fisk & Schneider, 1981; Giambra & Quilter, 1987; Mackworth, 1964) as well as effects of long-term sleep deprivation (e.g., Achermann, 2004; Borb & Achermann, 1999; Hursh et al., 2004; Jewett & Kronauer, 1999; McCauley et al., 2013). Such mathematical models aim to model the overall level of fatigue

at different points in time, but do not provide a detailed account of cognitive (and other) processes involved. Building on this work, recent models have focused on modeling fatigue within a computational cognitive architecture (e.g. French & Morris, 2003; Jones, Laird, & Neville, 1998; Gunzelmann, Gross, Gluck, & Dinges, 2009; Gunzelmann, Moore, Salvucci, & Gluck, 2011; Walsh, Gunzelmann, & Van Dongen, 2017; Veksler & Gunzelmann, 2018) to offer deeper insight into relationship between fatigue and the basic information processing mechanisms inherent to a task.

In this paper, we present a new unified model of fatigue that accounts for both sleep-related and task-related factors, and accounts for performance under both moderate and severe fatigue. In particular, we extend the work of Veksler and Gunzelmann (2018) and Walsh et al. (2017) by first examining the underlying theoretical foundations of both types of fatigue based on recent empirical work. We then utilize these ideas to propose an updated formulation of fatigue within the ACT-R cognitive architecture, testing its predictions against two data sets that demonstrate the benefits of a unified model.

Theoretical Foundations

The central idea of our modeling work is the concept of *microlapses*, introduced by Gunzelmann et al. (2009) to account for changes in behavioral performance related to fatigue. Microlapses can be viewed as an implementation of the "state instability" hypothesis (Doran et al., 2001): that a person's fatigue may be characterized as the switching between sleep and awake states, which may fluctuate second by second and can eventually progress to a physiological sleep state. Microlapses, however, incorporate the idea that switches between sleep and awake states may be more rapid (i.e., tens of ms), with transitions into, and remaining within, a "sleep" state becoming more likely as fatigue increases.

The concept of microlapses relies on the computational mechanisms of a procedural system in a cognitive architecture. A procedural system implemented as a production system is the central core of most well-established cognitive architectures like ACT-R (Anderson, 2007). ACT-R's production system implements a serial bottleneck in cognitive processing, representing cognition as a sequence of recognize-decide-act cycles that require about 50 ms each to execute. Under fatigue, microlapses cause the execution phase of the cycle to fail, leading to delays in completing, or even failure to complete, a task. As we will see, this theoretical foundation allows for an elegant model of fatigue that can account for both sleep- and task-related factors, and for performance across a range of degrees of fatigue.

Modeling Time-of-Day Effects

The first building block for our unified model is the model of sleep-related fatigue described in Walsh et al. (2017). Their model relied on ACT-R's concept of *utility*, namely that each production rule (effectively a 50-ms unit of action) has an associated utility that determines its usefulness in being activated, and this utility can be compared to those of other rules to determine the next action. By manipulating the utility of the productions and the utility threshold, the system is able to produce microlapses: if the utility U_i of the selected production is less than a set utility threshold UT, a microlapse occurs. Because U_i values are noisy, changes in U_i and UT thus influence the probability of microlapses occurring.

To account for sleep-related factors, Walsh et al. (2017) used a biomathematical model to quantify the overall impact of time awake and circadian rhythms. First, let us assume that we have a biomathematical model value $B^S(t)$ that, given a sleep schedule S (i.e., the prior hours for which the person was asleep and awake), provides the level of fatigue at a given time of day t. As mentioned earlier, several such models have been developed in the past; Gunzelmann et al. use the formulation provided by McCauley et al. (2013), which we include here as well. Using this value, we can specify a fatigue scale factor $F_{bio}(t)$ that will scale a production's overall utility proportionally based on the biomathematical model's predictions:[1]

$$F_{bio}(t) = 1 - c_{bio} * B^S(t)$$

We also include a fatigue constant c_{bio} to scale the biomathematical value, and we will consider this constant as one parameter to estimate in our model fitting later.

The next component of the model represents the accumulated effect of microlapses, and incorporates the fact that when a microlapse occurs, another microlapse is more likely to occur immediately after. This component is formulated as follows:

$$F_{dec}(n) = (c_{dec})^n$$

Here, n is the number of consecutive microlapses that have occurred—thus, $n = 0$ after a normal production has fired, but would increase by 1 for each consecutive microlapse thereafter until another normal production firing. c_{dec} is assumed to be a constant between 0 and 1, and thus the value $F_{dec}(n)$ is also a value between 0 and 1 that decreases with larger values of n. As described by Walsh et al. (2017), $F_{dec}(n)$ can quickly decay to the point that will be too low to fire any production; however, there is a counterbalancing effect that resets $F_{dec}(n)$ by setting $n = 0$ (akin to awakening the model) when a stimulus is presented.

Integrating these factors together, following Gunzelmann et al. (2009), Walsh et al. (2017) defined a fatigued utility $FU_i(t,n)$ as a modified value of production i's base utility $U_i(t)$ scaled by both $F_{bio}(t)$ and $F_{dec}(n)$:

$$FU_i(t,n) = F_{bio}(t) * F_{dec}(n) * U_i(t) + \epsilon$$

The final term ϵ adds noise to the final fatigued utility, where the noise is sampled from a logistic distribution. This component is carried over from the standard utility function

[1] The names of some variables and constants have been changed from the original formulation for increased clarity.

in ACT-R, which includes this parameter to generate stochasticity in model behavior. Once this fatigued utility is computed, its value is compared to a utility threshold $UT(t)$, computed using the biomathematical model and a specified initial utility threshold UT_0:

$$UT_{bio}(t) = 1 - d_{bio} * B^S(t)$$

$$UT(t) = UT_{bio}(t) * UT_0$$

These equations introduce another constant, d_{bio}, that scales the biomathematical model value.

Modeling Time-on-Task Effects

As an extension to the above model of time-of-day effects, Veksler and Gunzelmann (2018) developed a model to capture the effects of time-on-task. Using the same core mechanisms as Walsh et al. (2017) described earlier, they replaced the biomathematical factor $F_{bio}(t)$ with a time-on-task factor $F_{\text{tot}}(T)$ defined as follows:

$$F_{tot}(T) = (1 + T)^{c_{tot}}$$

$$FU_i(t, T, n) = F_{tot}(T) * F_{dec}(n) * U_i(t) + \epsilon$$

Here, T represents the total time-on-task, or time spent performing the same task. Veksler et al. used a similar formulation to revise the computation of the utility threshold:

$$UT_{tot}(T) = (1 + T)^{d_{tot}}$$

$$UT(T) = UT_{tot}(T) * UT_0$$

The constants c_{tot} and d_{tot} are assumed to be between –1 and 0, and thus their respective functions decrease as the time-on-task T increases.

A Unified Model of Fatigue

The foundational components above provide the basis for our own unified model, and at first glance, one might expect that we could simply combine the equations and have a unified account directly. Unfortunately, a simple combination does not work well either theoretically or experimentally. We thus explore how we might combine these accounts and then proceed with a specification of the final unified model.

Developing a Unified Model

Examining the formulations for the time-of-day and time-on-task models above, the most straightforward approach to a unified model would be to simply multiple the respective factors together—that is, computing fatigued utility as:

$$FU_i(t, T, n) = F_{bio}(t) * F_{tot}(T) * F_{dec}(n) * U_i(t) + \epsilon$$

This approach multiples the biomathematical component $F_{bio}(t)$ with the time-on-task component $F_{tot}(T)$ to derive the total fatigued utility. In fact, this formulation has been tried with limited success in earlier work: Khosroshahi et al. (2016) used it to account for time-of-day effects on performance in psychomotor vigilance and driving.

Unfortunately, however, we have attempted to use this formulation to account for a broader set of time-of-day and time-on-task effects (discussed more later), and found this approach lacking for several reasons. Using this formulation, it was impossible to find a set of parameter values that produces acceptable results simultaneously for both time-on-task and time-of-day effects—especially when the latter is drawn out to long periods of sleep deprivation. For example, consider how the model might account for lapses in the psychomotor vigilance task (PVT), where participants simply see a visual stimulus and press a button in response, and where a *lapse* is defined as a response time greater than 500 ms. Using the formulation above, the model can nicely fit the number of lapses in the early stages of fatigue, namely during the first day or two without sleep; however, this produces a model that rarely suffers the sleep attacks (response times greater than 30 s) suffered by humans after 48-88 hours of sleep deprivation. On the flip side, if the model parameters were fitted to produce a human-like frequency of sleep attacks, the lapses under moderate fatigue would be much too large.

In summary, this was not an issue of parameter fitting—the model formulation itself was fundamentally flawed. Closer analysis of the model revealed its theoretical flaw: increasing values of the biomathematical model $B^S(t)$ over time would actually scale down the time-on-task effect—effectively making the time-on-task effects *smaller* as the model became more fatigued. This effect is counterintuitive, and indeed, we did not find any evidence to support it in our available data or in the literature. In addition, in their study of time-on-task effects, Veksler et al. (2018) found no correlation between either prior night's sleep or wake-up time and the difference in response times between the first and last blocks of a 35-minute task—indicating an additive, not multiplicative, relationship between time-of-day and time-on-task (see Kribbs & Dinges, 1994; Gunzelmann et al., 2010).

Yet another observation about the naïve combined model, and about the earlier time-on-task model, relates to the model's $F_{dec}(n)$ equation. Recall that this factor incorporates the idea of cascading microsleeps, such that when a microsleep occurs, another is more likely to happen in the subsequent cycle. In the original formulation, because $F_{dec}(n) = (c_{dec})^n$ and $0 < c_{dec} < 1$, there is a rapid initial drop for small n followed by a leveling off to an asymptote near zero. Instead, based on our observations of sleep attacks, a better formulation would allow for only a slight drop for small n, but as n gets larger, the microsleeps would rapidly deteriorate into a sleep attack.

The Unified Model

Given the reasoning above, we created our unified model based on the earlier models of time-of-day and time-on-task while reflecting the evidence above. In particular, we modified the formulations of several equations as follows. First, we changed the decrement factor to a negated exponential function to introduce a steep drop in fatigue as microsleeps accumulate:

$$F_{dec}(t,n) = -(e)^{c_{dec}*n} + 2$$

Next, we modified the biomathematical factor to eliminate the initial 1 in a way that forces it to reduce the overall utility:

$$F_{bio}(t) = -c_{bio} * B^S(t)$$

We then introduced the additive effect between time-of-day and time-on-task into the computation of fatigued utility:

$$F_{tot}(T) = (1+T)^{c_{tot}}$$

$$FU_i(t,T,n) = F_{dec}(t,n) * [F_{bio}(t) + F_{tot}(T) + U_i(t) + \epsilon]$$

Analogous changes were applied to the utility threshold:

$$UT_{bio}(t) = -d_{bio} * B^S(t)$$

$$UT_{tot}(T) = (1+T)^{d_{tot}}$$

$$UT(t,T) = UT_{bio}(t) + UT_{tot}(T) + UT_0$$

These changes all together represent our unified model that accounts for both time-of-day and time-on-task effects. The next section aims to validate this model across two experimental data sets.

Model Evaluation

To validate our model, we rely on two studies that employ arguably the most common task in fatigue-related studies, namely the psychomotor vigilance task (PVT: Dinges and Powell, 1985). As mentioned, the PVT involves an extremely simple stimulus-response. PVT has been used extensively in sleep-related studies because of its sensitivity to sleep and circadian-based fatigue and its procedural simplicity and the consistency of individual performance (e.g., Gunzelmann, Moore, Gluck, Van Dongen, Dinges, 2008; Dorrian et al., 2005). PVT is thus a highly sensitive sustained attention task which can be an independent measure of fatigue (Van Dongen et al., 2011).

A typical PVT trial lasts 10 minutes and requires a button response every 2-10 seconds. The visual stimulus is a millisecond counter displayed on the screen, which starts at 0 at stimulus onset and counts forward as time passes; when the person presses the response key, the counter stops, thus providing feedback for performance. The main dependent measure in the PVT is the number of *lapses*, where a lapse is defined as a reaction time of more than 500 ms. Researchers have also measured the *median response time* (RT) of alert responses (reaction times between 150 and 500 ms), *false starts* (incorrect keypresses or reaction times less than 150 ms), and *sleep attacks* where the participant does not respond for 30 seconds or more.

It is worth noting that we used a single set of parameter values for the models in both studies. Our unified model contains 7 free parameters in total (see Table 1). Another parameter that was treated as a free parameter in previous models is *cycle time*, which controls the amount of time to evaluate and select a production during each cognitive cycle.

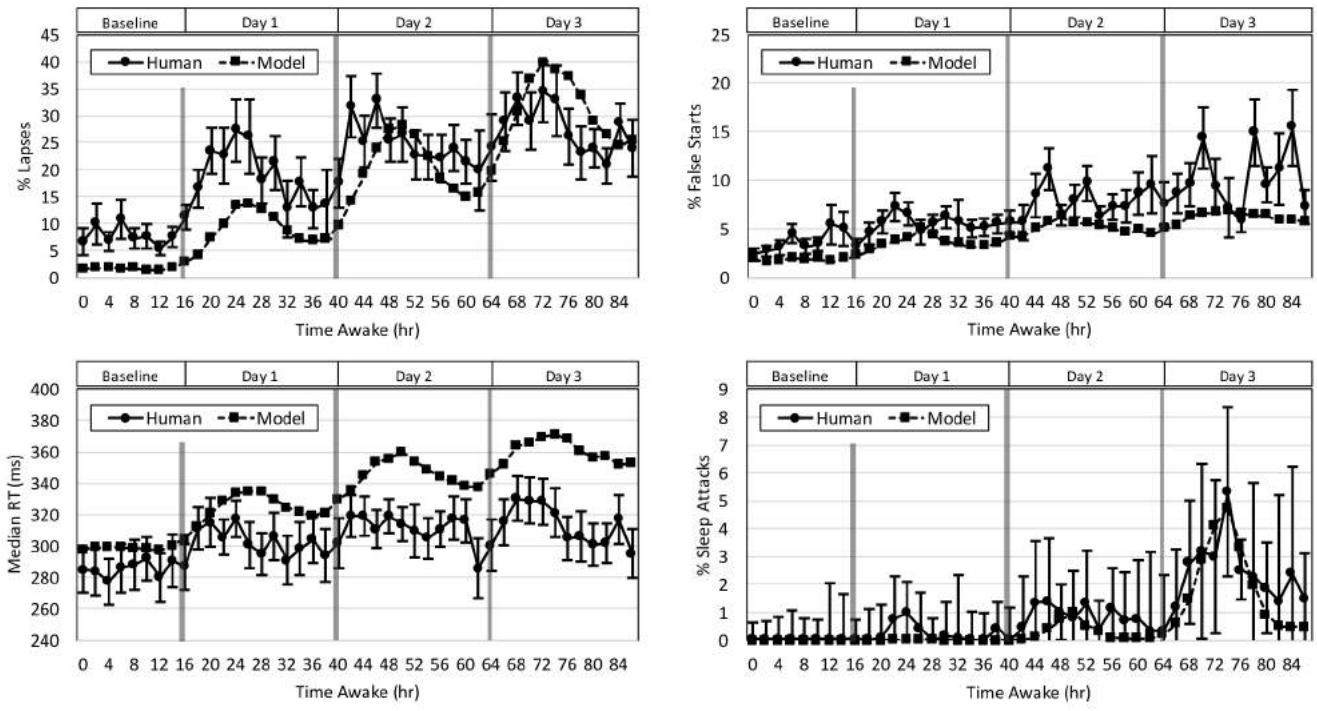


Figure 2: Human and model results for the PVT across 88 hours of sleep deprivation (Study 1).

We used the default value of 50 ms (Anderson, 2007) to keep the consistency with ACT-R theory. To reduce the chance of overfitting, we searched the parameter values to first fit parameters related to the time-on-task effect and then used the same values to fit parameters related to the time-of-day effect. Table 1 shows the list of free parameters and our best estimates for each parameter.

Table 1: ACT-R unified fatigue model free parameters and their estimations.

Parameter	Definition	Estimates
c_{dec}	Utility decrement factor	.006
c_{bio}	Utility biomathematical factor	.028
c_{tot}	Time-on-task decrement factor	.12
U_i	Base utility[2]	1.56
d_{bio}	Threshold biomathematical factor	.01
d_{tot}	Threshold time-on-task decrement factor	.04
UT_0	The initial threshold	1.15

PVT Model

Because the fatigue mechanisms described here are general to any production system or task, we require a model specifically of the PVT to test the fatigue mechanisms. For this purpose, we developed an ACT-R model that performs the PVT in as straightforward a manner as possible, with three main production rules, following the original model by Walsh et al. (2017):

1. *Attend*: shift visual attention to the stimulus
2. *Encode-and-Respond*: completes the visual encoding of the stimulus and initiates the response keypress
3. *Wait*: wait for the next stimulus

To capture the false starts in the PVT model, Walsh et al. (2017) used procedural partial matching: when enabled, productions whose conditions do not perfectly match the current state get a chance to be selected with a similarity difference (a negative value) added to their utility:

$$U_i' = U_i + SD_i + \epsilon$$

SD_i is the similarity difference which is added to the utility value when the conditions for the production are not met. At each cycle, the production with the greatest value U_i is selected when its utility exceeds the utility threshold. By enabling the procedural partial matching, Walsh et al. (2017) eliminated the need of a separate production (false-response); encode-and-respond can be selected at any time and when it is selected before the stimulus appears, false starts occur (which happens rarely because of the similarity difference added to it).

In the design of PVT in Walsh et al. (2017), U_i' was treated as a single free parameter meaning that one value was estimated and used for all the productions. The ACT-R's procedural learning (Anderson, 2007) was also disabled due to the nature of PVT and similar sustained attention tasks (Van Dongen et al., 2003) and the similarity difference was set to negative value of the production utility to simplify matters. Here we follow a similar design to stay consistent with earlier studies.

Study 1: Time-of-Day Experiment and Results

The first study for our model evaluation is a study of long-term sleep deprivation conducted by Doran et al. (2001). The study included 13 healthy participants who experienced 88 hours of total sleep deprivation. During periods of wakefulness for the duration of the study, participants

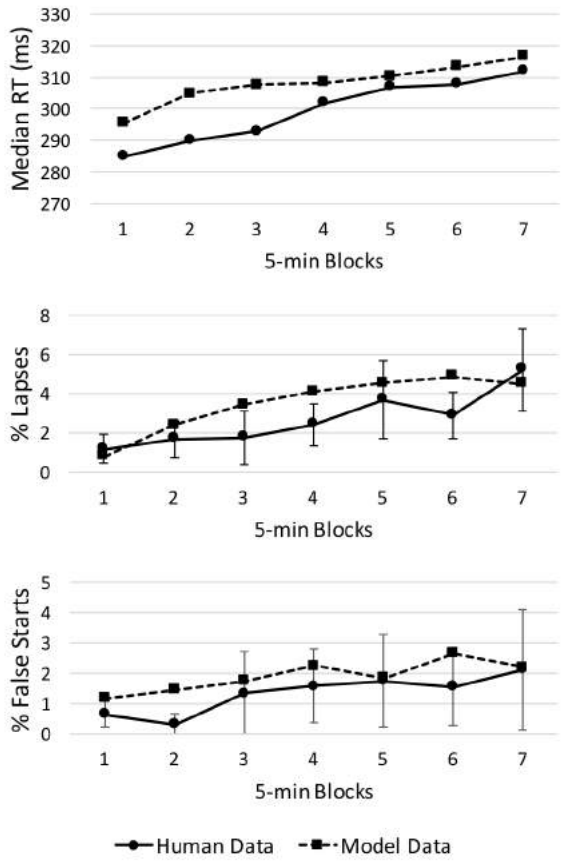


Figure 3: Human and model PVT results across the 5-min blocks of the 35-min experiment (Study 2).

[2] Base utility is defined as the standard utility value for all productions.

completed a battery of performance evaluation tasks every 2 hours, including a 10-minute PVT. Although Gunzelmann et al. (2009) modeled the same experiment, our effort here is different in two ways: (1) the older model used different parameter values for each day, whereas we are constraining our model to a single set of parameters; and (2) the older model did not include the time-on-task factor; although time-on-task was not a focus of the study, it is important for us to know that the model can produce a good fit while incorporating this factor (see also Gunzelmann et al. 2011).

To model this study, we ran iterations of the PVT model for 10-minute periods and matched the sleep schedule of the model to the 88-hour sleep deprivation experimental protocol. Parameters were estimated to produce the best fit across the four PVT measures; a single set of parameter values was used across the entire experiment. Figure 2 shows the human data and model's performance for all four measures: lapses ($R^2 = 0.68$, $RMSE = 7.92$), median reaction times ($R^2 = 0.55$, $RMSE = 34.41$), false starts ($R^2 = 0.56$, $RMSE = 3.42$), and sleep attacks ($R^2 = 0.77$, $RMSE = 0.64$). Overall, the model accounted for all the major aspects of the data; it slightly underpredicted lapses in days 1-2, and slightly overpredicted RT in days 4-5, but in general, the model captured most of the fluctuations in performance across all four measures.

Study 2: Time-on-Task Experiment and Results

The second study for our model evaluation is a study examining time-on-task effects conducted by Veksler and Gunzelmann (2018). In the study, 20 participants performed a 35-minute PVT instead of the usual 10 minutes; by extending the typical PVT duration, they were able to draw out how the effects of time-on-task on PVT are similar to those of sleep loss. As mentioned earlier, Veksler and Gunzelmann modeled the time-on-task effects in this experiment, but at the time did not incorporate the biomathematical model, and used a different set of parameters than earlier models. To include biomathematical modeling in our simulations, we assumed 8 hours the night before the experiment, waking at 7:30am and performing the experiment at 10:00am.

The results of the model compared to the human data are shown in Figure 3. For this evaluation, we compared the performance of the model with the experimental results across seven 5-minute blocks of PVT. The model was able to capture the changes across the blocks for median reaction times ($R^2 = 0.85$, $RMSE = 9.67$), lapses ($R^2 = 0.53$, $RMSE = 1.27$), and false starts ($R^2 = 0.55$, $RMSE = 0.69$). The model shows a slight overprediction of lapses in the middle blocks, but in general, the model performs well for these three measures, especially considering that this is the same model with the same parameters as the previous study.

General Discussion

In this paper, we introduce a unified computational model that accounts for two of the most important aspect of fatigue, namely time-of-day and time-on-task effects on behavior and performance. Our result once again accounts for the microlapse hypothesis (Gunzelmann et al. 2009) and the fact that microlapses could account for both sleep loss and time-on-task effects in sustained attention (following Veksler et al., 2018). We were also able to capture both the time-of-day and time-on-task effects with the same parameters; going forward, we are interested in understanding how these parameters might generalize to other tasks, and how they might vary across individuals. It is also notable that the mechanisms here are complex, with a number of free parameters that are sensitive to changes in setting. Nevertheless, we believe that as we continue to fit additional experiments with this unified model, we can reduce the space of free parameters and can find parameter values that cut across a variety of task domains, providing an even more general model with easier estimation of parameters.

In conclusion, by validating that the unified model can account for the negative consequences in behavioral performance of both time-of-day and time-on-task effects, we have demonstrated that both phenomena have similar natures and as a result could be modeled with a single set of mechanisms. Although PVT as a testbed for our modeling seems to be a simple task, this research will give us a strong foundation to expand the model to more complex domains. We are also interested in extending this model beyond the sleep-loss and time-on-task to moderate levels of fatigue (e.g., sequential sleep limitation), which would further bolster the model's generalizability to complex real-world task domains.

Acknowledgments

This work was funded in part by a grant from the Air Force Research Laboratory (#FA8650-15-2-6603).

References

Achermann, P. (2004). The two-process model of sleep regulation revisited. *Aviation, Space, and Environmental Medicine*, *75*, A37-A43.

Anderson, J. R. (2007). *How Can the Human Mind Occur in the Physical Universe?* New York: OUP.偏

Bakan, P. (1955). Discrimination decrement as a function of time in a prolonged vigil. *Journal of Experimental Psychology*, *50*, 387.

Borb, A. A., & Achermann, P. (1999). Sleep homeostasis and models of sleep regulation. *Journal of Biological Rhythms*, *14*, 559-570.

Dinges, D. F., & Powell, J. W. (1985). Microcomputer analyses of performance on a portable, simple visual RT task during sustained operations. *Behavior Research Methods, Instruments, & Computers*, *17*, 652-655.

Dinges, D. F. (1995). An overview of sleepiness and accidents. *Journal of Sleep Research*, *4*, 4-14.

Doran, S. M., Van Dongen, H. P. A., & Dinges, D. F. (2001). Sustained attention performance during sleep deprivation:

evidence of state instability. *Archives Italiennes de Biologie*, *139*, 253-267.

Dorrian, J., Rogers, N. L., & Dinges, D. F. (2005). Psychomotor vigilance performance: Neurocognitive assay sensitive to sleep loss. In C. A. Kushida (Ed.), *Sleep Deprivation: Clinical Issues, Pharmacology and Sleep Loss Effects* (pp. 39–70). New York: Marcel Dekker.

Fisk, A. D., & Schneider, W. (1981). Control and automatic processing during tasks requiring sustained attention: A new approach to vigilance. *Human Factors*, *23*, 737-750.

French, J., & Morris, C. S. (2003). Modeling fatigue degraded performance in artificial agents. In *Proc. of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 47, No. 3, pp. 307-310). Los Angeles, CA: SAGE.

Giambra, L. M., & Quilter, R. E. (1987). A two-term exponential functional description of the time course of sustained attention. *Human Factors*, *29*, 635-643.

Gunzelmann, G., Gross, J. B., Gluck, K. A., & Dinges, D. F. (2009). Sleep deprivation and sustained attention performance: Integrating mathematical and cognitive modeling. *Cognitive Science*, *33*, 880-910.

Gunzelmann, G., Moore, L. R., Gluck, K. A., Van Dongen, H. P., & Dinges, D. F. (2008). Individual differences in sustained vigilant attention: Insights from computational cognitive modeling. In *Proc. of the 30th Annual Meeting of the Cognitive Science Society* (pp. 2017-2022).

Gunzelmann, G., Moore, L. R., Gluck, K. A., Van Dongen, H. P., & Dinges, D. F. (2010). Fatigue in sustained attention: Generalizing mechanisms for time awake to time on task. *Cognitive fatigue: Multidisciplinary perspectives on current research and future applications* (pp. 83-101). Washington, DC: American Psychological Association.

Gunzelmann, G., Moore Jr., L. R., Salvucci, D. D., & Gluck, K. A. (2011). Sleep loss and driver performance: Quantitative predictions with zero free parameters. *Cognitive Systems Research*, *12*, 154-163.

Horne, J., & Reyner, L. (1999). Vehicle accidents related to sleep: A review. *Occupational and Environmental Medicine, 56*, 289-294.

Hursh, S. R., Redmond, D. P., Johnson, M. L., Thorne, D. R., Belenky, G., Balkin, T. J., ... & Eddy, D. R. (2004). Fatigue models for applied research in warfighting. *Aviation, Space, and Environmental Medicine*, *75*, A44-A53.

Jewett, M. E., & Kronauer, R. E. (1999). Interactive mathematical models of subjective alertness and cognitive throughput in humans. *Journal of Biological Rhythms*, *14*, 588–597.

Jones, R. M., Laird, J. E., & Neville, K. (1998). Modeling pilot fatigue with a synthetic behavior model. In *Proc. of the 7th Conference on Computer Generated Forces and Behavioral Representation* (pp. 349-356).

Khosroshahi, E. B., Salvucci, D. D., Veksler, B. Z., & Gunzelmann, G. (2016). Capturing the effects of moderate fatigue on driver performance. In *Proc. of the 14th Intl. Conference on Cognitive Modeling* (pp. 163-168).

Kribbs, N. B., & Dinges, D. (1994). Vigilance decrement and sleepiness. In R. D. Ogilvie & J. R. Harsh (Eds.), *Sleep onset: Normal and abnormal processes* (pp. 113-125). Washington, DC: APA.

Lauber, J. K., & Kayten, P. J. (1988). Sleepiness, circadian dysrhythmia, and fatigue in transportation system accidents. *Sleep: Journal of Sleep Research & Sleep Medicine*, *11*, 503-512.

McCauley, P., Kalachev, L. V., Mollicone, D. J., Banks, S., Dinges, D. F., & Van Dongen, H. P. (2013). Dynamic circadian modulation in a biomathematical model for the effects of sleep and sleep loss on waking neurobehavioral performance. *Sleep*, *36,* 1987–1997.

Mackworth, N. H. (1948). The breakdown of vigilance during prolonged visual search. *Quarterly Journal of Experimental Psychology*, *1*, 6-21.

Mackworth, J. F. (1964). Performance decrement in vigilance, threshold, and high-speed perceptual motor tasks. *Canadian Journal of Psychology*, *18*, 209.

Mitler, M. M., Carskadon, M. A., Czeisier, C. A., Dement, W. C., Dinges, D. F., & Graeber, R. C. (1988). Catastrophes, sleep, and public policy: Consensus report. *Sleep*, *11*, 100-109.

Pack, A. I., Pack, A. M., Rodgman, E., Cucchiara, A., Dinges, D. F., & Schwab, C. W. (1995). Characteristics of crashes attributed to the driver having fallen asleep. *Accident Analysis & Prevention*, *27*, 769-775.

Parasuraman, R. (1979). Memory load and event rate control sensitivity decrements in sustained attention. *Science*, *205*, 924-927.

Pattyn, N., Neyt, X., Henderickx, D., & Soetens, E. (2008). Psychophysiological investigation of vigilance decrement: boredom or cognitive fatigue? *Physiology & Behavior*, *93*, 369-378.

Van Dongen, H. P. A., Belenky, G., & Krueger, J. M. (2011). A local, bottom-up perspective on sleep deprivation and neurobehavioral performance. *Current Topics in Medicinal Chemistry*, *11*, 2414-2422.

Van Dongen, H. P. A., Maislin, G., Mullington, J. M., & Dinges, D. F. (2003). The cumulative cost of additional wakefulness: Dose-response effects on neurobehavioral functions and sleep physiology from chronic sleep restriction and total sleep deprivation. *Sleep, 26*, 117–126.

Veksler, B. Z., & Gunzelmann, G. (2018). Functional equivalence of sleep loss and time on task effects in sustained attention. *Cognitive Science, 42*, 600-632.

Walsh, M. M., Gunzelmann, G., & Van Dongen, H. P. A. (2017). Computational cognitive models of the temporal dynamics of fatigue from sleep loss. *Psychonomic Bulletin & Review, 24*, 1785-1807.

Congenitally Blind Individuals Theories and Inferences About Object Color

Judy Kim
Johns Hopkins University, Baltimore, Maryland, United States

Lindsay Yazzolino
Johns Hopkins University, Brookline, Massachusetts, United States

Brianna Aheimer
Johns Hopkins University, Baltimore, Maryland, United States

Vernica Montan Manrara
Johns Hopkins University, Baltimore, Maryland, United States

Marina Bedny
Johns Hopkins University, Baltimore, Maryland, United States

Abstract

Locke argued that persons born blind do not possess true knowledge about color. While prior studies find some knowledge of color among blind individuals, questions remain about the depth of this knowledge. Do blind individuals merely learn inferentially shallow verbal associations (e.g., bananayellow)? We hypothesized instead that blind individuals are more likely to acquire causally-relevant color information. Blind (n=20) and sighted adults (n=20) reported colors of natural kinds (e.g. banana) and artifacts (e.g. car) and judged the likelihood that two instances of a type have the same color. Relative to the sighted, blind participants were less likely to know specific object colors (e.g. banana-yellow), but made identical inferences about color consistency (more consistent colors for natural kinds). Inferences were similar across groups even for novel objects. Further, blind individuals gave detailed and coherent causal explanations of color origins. Inferentially rich knowledge of sensory categories can develop without first-person experience.

I know what you did last summer (and how often). Epistemic states and statistical normality in causal judgements

Lara Kirfel (ucjulki@ucl.ac.uk)
David Lagnado (d.lagnado@ucl.ac.uk)

Department of Experimental Psychology, University College London
26 Bedford Way, London WC1H 0AP England

Abstract

When several causes contributed to an outcome, we often single out one causal factor as being "more of a cause" than others. What explains this selection? Existing research suggests that people's judgements of actual causation can be influenced by the degree to which they regard certain events as norm-deviant, or "abnormal" (Hart & Honoré, 1963; Kahneman & Miller, 1986; Hitchcock & Knobe, 2009; Halpern & Hitchcock 2015). In this paper, we argue that statistical abnormality influences causal judgements about human agents by changing the agents' epistemic states (*Epistemic Hypothesis*). In Experiment 1, we replicate previous findings that people assign more causal strength to a statistically abnormally acting agent, but show that they also assign them more knowledge about the behaviour of their peers. In Experiment 2, we show that in case of equal epistemic uncertainty, people do not differentiate between statistically abnormal and normal causal agents. In Experiment 3, we explore the difference between type and token abnormality, and find that a token abnormal, but type normal behaviour still influences causal judgments, with people's epistemic judgments mirroring these causal judgments. We discuss the implications of this research for current norm-frameworks in causal cognition.

Keywords: statistical norms, normality, causal judgment, counterfactual reasoning, epistemic states

Our ability to form causal judgements plays a fundamental role in human cognition. In everyday life, we encounter situations that demand an explanation of why something happened, how it happened, or how it could have been prevented. Fortunately, our environment is rich in statistical information. Statistical patterns have been shown to be a reliable cue in guiding people's causal inferences and judgements (Cheng, 1997). The co-variation of cue and outcome, their proximity in space and time or the temporal order in which events occur have been shown to inform assumptions about *causal structure*, i.e. the existence of a causal relation between cue and outcome, as well as *causal strength*, i.e. the degree of a causal relation between cue and outcome (Lagnado, Waldmann, Hagmayer & Sloman, 2007).

Recent research suggests that the influence of statistical information on causal cognition goes even further. The *statistical normality* of a causal factor, i.e. how likely, typical or frequent it is perceived, can make a difference to people's causal judgement about this factor over and beyond its actual causal contribution (Cheng & Novick, 1991; Hitchcock & Knobe, 2009; Samland & Waldman, 2016; Kominsky, Phillips, Gerstenberg, Lagnado & Knobe, 2015; Icard, Kominsky, Knobe, 2017). In a range of empirical studies, people have been shown to differentiate between causal factors according to their statistical features, even when both factors are necessary for the outcome to occur (Hitchcock & Knobe, 2009; Icard et al., 2017; Gerstenberg & Icard, n.d.).

Most prominently, this research suggests that deviations from statistical normality increases the causal strength assigned to a cause. Specifically, people are more inclined to judge that *C* causes *E* when *C* is perceived to be statistically "abnormal", i.e. unlikely, infrequent or atypical manner, rather than when C is perceived to be statistically normal. This holds even when in both cases, *C* is known to have the same actual causal contribution to the effect. These findings raise the question of why people take statistical features into account even when these features do not function as supplementary cues to causal structure or strength. What makes people prefer abnormal causal candidates?

Normality matters – but why?

A prominent line of research argues that norms or normality influence causal judgments by changing the relevance or propensity to consider counterfactual possibilities (Kahneman & Miller, 1986; Hitchcock & Knobe, 2009, Icard et al., 2017). A statistical norm violation increases the likelihood of thinking about an alternative scenario in which the norm-violation is replaced by norm-conforming behaviour. A typical test case in this research is causation in a conjunctive causal structure, where two causes are each necessary to produce an outcome. When both C_{normal} and $C_{abnormal}$ together bring about outcome E, people will be more likely to envisage a counterfactual scenario in which $C_{abnormal}$ is absent, rather than a counterfactual in which C_{normal} is absent. According to the counterfactual account, imagining a counterfactual alternative in which normality, or norm-conformity, is restored highlights the causal role of the abnormal causal factor for the outcome, compared to that of the normal causal factor (Kahneman & Miller, 1986; Hitchcock & Knobe, 2009, Icard et al., 2017).

Counterfactual accounts of norm effects in causal cognition have gained increasing popularity. On the one hand, they have integrated norms into formal causal frameworks that can explain a variety of norm effects on causal judgments, such as "causal superseding" (Kominsky

et al., 2017) or "abnormal deflation" (Icard et al., 2017; 2018). On the other hand, they not only predict the influence of statistical norms on people's causal judgements, but also the impact of other kind of norms, such as prescriptive norms (Hitchock & Knobe, 2008) or norms of proper functioning (Phillips & Kominsky, 2018). Recently, it has been suggested that the influence of both prescriptive and statistical norms on causal judgements can be explained by a single normality concept (Bear & Knobe, 2017).

Knowing me, knowing you

The majority of studies supporting the counterfactual account has been conducted using vignette stories in which participants rate the causal impact of human agents who differ in certain aspects of normality. This has led some to argue that the influence of moral abnormality on causal judgements in the context of human agents reveals something about people's blame responses, rather than a difference in counterfactual and causal reasoning (Samland & Waldman, 2016, Alicke, Rose & Bloom, 2012). However, most research argues that statistical norms influence the underlying process of causal judgement. When it comes to statistical norms, it is the abnormality itself that leads people to judge a causal difference between an abnormally and a normally acting causal agent.

In this paper, we propose an alternative hypothesis. While we agree that statistical likelihoods can have an impact on people's causal judgements about events or objects, we think that in the context of human agents, there is another important factor to consider. Epistemic states, i.e. the knowledge an agent has about their environment, have been shown to influence how we evaluate the causality of their actions (Lagnado & Channon, 2008). Whether an agent engages in a frequent of typical action, or an infrequent or atypical action, will likely change their epistemic states about the consequences of this action. In particular, in the case of conjunctive causal structures, an abnormally acting agent seems to have an epistemic advantage over the normally acting agent in knowing or expecting the outcome to happen. We believe it is the epistemic advantage that arises from a statistically abnormal action, rather than the abnormality per se, that drives the main difference in people's judgements about causal agents. We call this the *Epistemic Hypothesis* (EP). We conducted three experiments to investigate this hypothesis. In Experiment 1, we replicate previous literature by showing that people assign more causal strength to a statistically abnormally acting agent. In Experiment 2, we show that in case of equal epistemic uncertainty, people do not make a causal difference between abnormal and normal causal agents. In Experiment 3, we find that a token abnormal, but type normal behaviour still influences causal judgments, with people's epistemic judgments mirroring these causal judgments. We discuss the implications of this research for current norm-frameworks in causal cognition.

Experiment 1

The term "statistical abnormality" has been used broadly in the causal cognition literature, referring to actions or events that are unlikely, rare or atypical. In our experiments we have concentrated on statistical normality in the sense of the frequency of an action. We follow the current paradigm of assessing causal ratings of two causal agents in a conjunctive causal structure, while varying the statistical normality of their actions. In order to focus our investigation, we deviate from the current experimental paradigms in two aspects. Instead of descriptive vignettes ("Agent X frequently does action Y"), we use sequential animated video scenes in order to represent action frequencies more naturalistically. Furthermore, previous literature has suggested that the co-variation between cause and effect influences causal considerations (Harinen, 2017, Cheng 1997, Kirfel & Lagnado, 2018). Current experimental studies are ambiguous about the statistical normality of the effect, which is why we decided to employ a causal structure which allows us to control the frequency of the outcome.

Participants[1]

176 participants were recruited for this online study via Amazon Mechanical Turk. 10 participants were excluded for answering more than one check question wrong, leaving a final sample of 166 (M_{age} = 37.19, SD_{age}=11.24, *age range*= [20-77]; 101 male, 64 female, 1 N.A.) They were paid £0.70 upon completion of the study (Ø 8.06min).

Design

We manipulated two factors in a two-agent-scenario: the statistical normality of an action (*frequent* vs. *infrequent action*) and the type of scenario (*microwave* vs. *coffee machine*). Statistical normality, i.e. frequency of actions was manipulated for one agent (Agent 2: *varied agent*) while holding the frequency of actions fixed on the second agent (Agent 1: *fixed agent*). The scenario type was manipulated between-participant, while the statistical normality was manipulated within-participant. Participant saw two video clips ("*frequent", "infrequent"*) from one of the two scenario types, presented in randomized order. Names of the agents were varied across all conditions.

Material

The frame story consists of two co-workers in a shared office. Depending on the scenario type, the office has either two coffee machines or two microwaves that the employees can use. For energy saving purposes, the company introduces the "Green Friday" on which the building is switched into a power-saving mode. As a result, the use of more than one coffee machine (microwave) on Fridays will lead to a power failure in the building. All workers are aware of the Green Friday.

[1] The material and data for all experiments are available under: https://osf.io/zhvsb/

Figure 1. Scenario "Coffee Machine" with Fixed Agent ("Henry") and varied Agent ("James").

Response Measures

Causal Rating. After each video clip participants were asked to express their agreement with statements about the causal contribution of each agent to the outcome ["Agent 1 (2) has caused the power failure."] on a 7-point Likert scale [1 – 'Strongly disagree' to 7 – 'strongly agree']. Questions were presented in randomized order.

Manipulation Checks. In two subsequent manipulation check questions, participants were asked about their understanding of the action frequency in the scenario ["Who used a coffee machine frequently (rather than infrequently) this week?" – 'Agent 1', 'Agent 2'; multiple answers possible] and the causal structure ["The use of how many coffee machines does it take to produce a power failure on Friday? – 'One coffee machine', 'Two coffee machines']. At the end of the survey, i.e. after watching both videos and answering the causal rating questions, participants were asked to express their opinion about the epistemic states of the agents in both videos ["Agent 1 (2) knew that Agent 2 (1) would use a microwave on Friday."] on a 7-point Likert scale [1 – 'strongly disagree' to 7 – 'strongly agree']. By this, we wanted to check for people's assumptions of the agent's epistemic states.

Results

A Mixed ANOVA for participant's agreement ratings about the causal statements revealed a significant interaction for Frequency × Agent, $F_{(1,164)} = 29.05$, $p < .001$, $\eta_p^2 = .15$. While people judge no difference between the causal contribution of the agents when both of them have frequently performed the action, an agent whose action is rare is seen as more causal ($M = 5.52$, $SD = 1.63$, 95% CI [5.27, 5.78]) then a frequently acting agent ($M = 4.54$, $SD = 1.97$, 95% CI [4.24, 4.84]).

There was no effect for scenario type ($p = .653$). A Mixed ANOVA for agreement ratings about the agent's epistemic states revealed a significant interaction for Frequency × Agent $F_{(1,164)} = 291.60$, $p < .001$, $\eta_p^2 = .64$. When the two agents differ in the frequency of their actions, people express more agreement with the proposition that the agent acting for the first time on Friday knows that their (frequently acting) coworker would act ($M = 5.83$, $SD = 1.74$, 95% CI [5.54, 6.07]), than vice versa ($M = 2.37$, $SD = 1.54$, 95% CI [2.11, 2.64]).

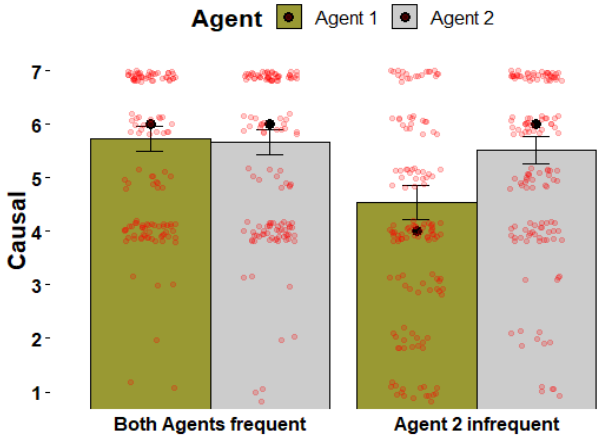


Figure 2. Mean agreement ratings (scale 1-7) for causal statement. Error bars represent ±1 SE mean, black points represent the median.

Discussion

In this experiment we found that when a frequently and infrequently acting agent together cause an outcome, people judge the agent who has acted infrequently to be of greater causal strength than the frequently acting agent. Our findings are in line with the literature in causal cognition showing that people tend to assign more causal strength to abnormal causes (Hitchcock & Knobe, 2009, Icard et al., 2017). In our study, we manipulated the statistical normality among agents' actions. However, in a two-agent conjunctive structure, acting abnormally gives the agent a better chance of foreseeing the consequences of their action. This is because the infrequent worker has witnessed the frequent worker acting on multiple occasions, whereas the frequent worker has never seen the infrequent worker act. In accordance with this prediction, we found that people assigned more knowledge about the co-worker's behaviour to the abnormally acting agent. This leaves open the question whether it was the epistemic advantage of the abnormally acting agent, or the abnormality of their action, that led people to make a causal difference. For our second experiment, we therefore examined whether abnormality still influences causal judgements when there is no such epistemic advantage.

Experiment 2

In the second experiment, we aimed to investigate the effect of statistical normality on causal judgments when neither agent knows about the frequency of the other's actions.

Participants

171 participants were recruited for this online study via MTurk; 19 were excluded for answering more than one check question wrong (N=152, M_{age} = 38.22, SD_{age} =11.25, *age*

range= [19-71]; 81 male, 79 female,). They were paid £0.70 upon completion of the study (Ø 8.61min).

Design & Material

The experiment was designed as Experiment 1, with the difference that the two agents are shown as working in separate offices on different floors. The agents are introduced as co-workers who "[despite] working for the same company, do not know each other and have never met or seen each other." (https://youtu.be/dYaXueuGOoA).

Response Measures

We used the same Causal Rating Measures and Manipulation Checks as in in Experiment 1.

Figure 3. Scenario "Coffee Machine" with fixed Agent 1 ("Henry") and varied Agent 2 ("James").

Results

A Mixed ANOVA for participant's agreement ratings about the causal statements revealed a main effect for Frequency $F(1,150) = 9.96$, $p = .002$, $\eta_p^2 = .06$. Higher causal ratings are given when both agents act frequently ($M = 5.07$, $SD = 1.87$, 95% *CI* [4.87, 5.28]), compared to the case in which only one has acted frequently ($M = 4.72$, $SD = 2.06$, 95% *CI* [4.60, 4.93]).

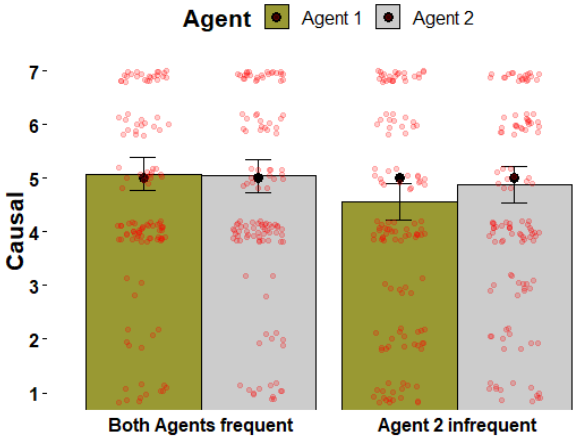


Figure 4. Mean agreement ratings (scale 1-7) for causal statement. Error bars represent ±1 SE mean, black points represent the median.

There was no interaction effect of Frequency × Agent ($p = .118$), and no effect of scenario type ($p = .441$).

A Mixed ANOVA for agreement ratings about the agent's epistemic states revealed a significant interaction for Frequency × Agent $F_{(1,150)} = 4.83$, $p = .029$, $\eta_p^2 = .03$.

Discussion

In our second experiment, we investigated whether statistical normality influences causal judgments when neither agent knows about the other's behaviour. We found that people do not differentiate between a frequently and rarely agent when neither agent knows or observes the other's behaviour. When both agents operate out of sight from each other, people do not judge the abnormally acting agent as contributing more to the joint outcome of their actions. However, the epistemic manipulation check questions revealed that our manipulation of epistemic uncertainty was only partly successful. Although both agents were introduced as working from different offices and not knowing each other, participants still assumed a very small epistemic difference when they differ in their action frequency. Compared to Experiment 1, however, the epistemic difference is negligible ($MD_{EXP1} = 3.36$, $MD_{EXP2} = 0.16$) and rated at the bottom of the 7 point Likert scale [1 – 'strongly disagree' to 7 – 'strongly agree'] (*Frequent Agent*: $M = 1.24$, *infrequent Agent* $M = 1.40$).

As a result, people overall disagreed with the statement that the agents had knowledge of each other. Our experiment shows that the general reduction in the agents' knowledge about each other led to an absence of influence of statistical normality. If an agent has not secured knowledge about the behaviour of their peers, people do not take into account the statistical normality of the agent's behaviour when making causal judgements. Our second experiment therefore shows that in case of epistemic uncertainty, i.e. when acting abnormally does not generate an epistemic advantage, statistical normality does not affect causal judgement.

Type and Token Normality

Our two experiments so far confirm the hypothesis that statistical normality influences causal judgments by giving an epistemic advantage. However, there is another interesting case to consider. Statistical abnormality does not necessarily need to lead to an epistemic advantage when agents, despite differing in their action frequency, can still predict the general outcome-causing behaviour. This case might be hard to experience naturally, because it is exactly the unpredictability of abnormal behaviour that makes it difficult for other agents to foresee it, leading to an epistemic asymmetry. However, when the agent acts for the first time, but their specific action has been performed frequently before by someone else, the agent's behaviour is still abnormal, but others might have been able to foresee the occurrence of this type of action. Strictly speaking, in such a case the abnormality of the behaviour is abnormal only in a limited sense. The agent is abnormal on an "agent-token" level, i.e. *this* particular agent performing action φ, but normal on an "agent-type" level, i.e. *an agent* performing action φ. In their

paper "Two types of typicality", Sytsma, Livengood and Rose (2015) reassess the role of statistical normality by distinguishing between agent-level and population-level statistical norms. They find that agent-level statistical normality has an influence on causal attributions, while deviating from a populational-level norm does not affect people's causal judgements.

For our third experiment, we adopted a similar paradigm as used by Sytsma et al. (2015). We introduced a third 'auxiliary' agent who uses one of the outcome triggering devices regularly during the week before the abnormally acting agent uses it on Friday. By this, we were interested whether an action that is token abnormal, but type normal, still influences causal judgment. Crucially, we assumed that introducing type normality might also make a difference to the agents' epistemic states. That is, in contrast to Experiment 1, here we would expect the token normally acting agent to have certain foreseeability that *someone* performs the causally relevant action on Friday (even though on that day, this happens to be a different agent than expected). The manipulation of epistemic states in Experiment 3 however is much noisy and occurs indirectly through the manipulation of type normality. In line with EP, we predict that if people continue to judge the token abnormal agent to be more causal for the outcome, this would again be tracked by a perceived epistemic asymmetry between these agents.

Experiment 3

In the third experiment, we aimed to investigate the effect of statistical normality on causal judgments when an agent acts statistically abnormal, but their action has been performed before by others.

Participants

180 participants were recruited for this online study via Amazon Mechanical Turk; 26 were excluded for answering more than manipulation wrong (N=154, M_{age} = 38.47, SD_{age} =12.16, *age range* = [19-72]; 90 male, 62 female, 1 2.A). They were paid £0.70 upon completion of the study (Ø 8.64min).

Design & Material

We used the same scenarios as in Experiment 1, but added a third causally irrelevant agent, Agent 3. The statistical normality of the agents who are causing the final outcome was manipulated as before, i.e. varied for one agent and held fixed for the other (*Agent 1*: fixed agent; *Agent 2*: varied agent). In the condition in which both Agent 1 and Agent 2 behave statistically normal, both of them use a coffee machine (microwave) from Monday to Friday, with Agent 3 simply being present and not acting (https://youtu.be/Tsxt1peUA74). In the condition in which Agent 2 acts abnormally, Agent 2 uses the coffee machine (microwave) on Friday, but Agent 3 uses that exact same coffee machine (microwave) the days before, i.e. from Monday to Thursday (https://youtu.be/k2wE52iZPKY).

Response Measures

We used the same Causal Rating Measures as in Experiment 1, but for the sake of completeness, added a Causal Rating for Agent 3 which we did not include in our analysis. We added a Manipulation Check Question to test whether people correctly perceived who had acted on the final day of the outcome ["Who used a microwave on Friday?" 'Agent 1', 'Agent 2', 'Agent 3', multiple answers possible]. We changed our Epistemic Question into a question about i) the type of behaviour "Agent 1 (2) knew that the other coffee machine (microwave) would be used by someone on Friday", and ii) the behaviour of the specific agent "Agent 1 (2) knew that Agent 2 (1) would use the other coffee machine (microwave) on Friday" [1 – 'strongly disagree' to 7 – 'strongly agree'].

Figure 5. Scenario "Coffee Machine" with fixed Agent 1 ("Dan"), varied Agent 2 ("Eddie") and 'auxiliary' Agent 3 ("Sam").

Results

A Mixed ANOVA for participant's agreement ratings about the causal statements about Agent 1 (fixed) and Agent 2 (varied) revealed an interaction effect for Frequency × Agent $F_{(1,152)} = 9.89$, $p = .002$, $\eta_p^2 = .06$.

When Agent 1 and 2 differ in the frequency of the actions that they perform on Friday, people agree more with the statement that the infrequently acting Agent 2 caused the outcome (M = 5.05, SD = 1.99, 95% CI [4.73, 5.36]), than that the frequently acting Agent 1 (M = 4.48, SD = 2.11, 95% CI [4.15, 4.82]).

A Mixed ANOVA for agreement ratings about agent's epistemic states for the type of behaviour revealed a significant interaction for Frequency × Agent $F_{(1,152)} = 10.82$, $p = .001$, $\eta_p^2 = .07$. The infrequently acting agent is judged to have more certainty that *someone* would use the other device on Friday (M = 4.73, SD = 2.01, 95% CI [4.41, 5.05]), than the frequently acting agent (M = 4.23, SD = 2.12, 95% CI [3.89, 4.56]). A Mixed ANOVA for ratings on the agent's assumptions about the *specific agent* using the other device also revealed a significant interaction for Frequency × Agent $F_{(1,152)} = 110.01$, $p < .001$, $\eta_p^2 = .42$. Participants agreed substantially more with the statement that the infrequently acting agent knows that the frequently acting agent would be

using the other relevant device on Friday (M = 4.61, SD = 2.01, 95% CI [4.29, 4.93]), than vice versa (M = 2.53, SD = 1.83, 95% CI [2.24, 2.82]).

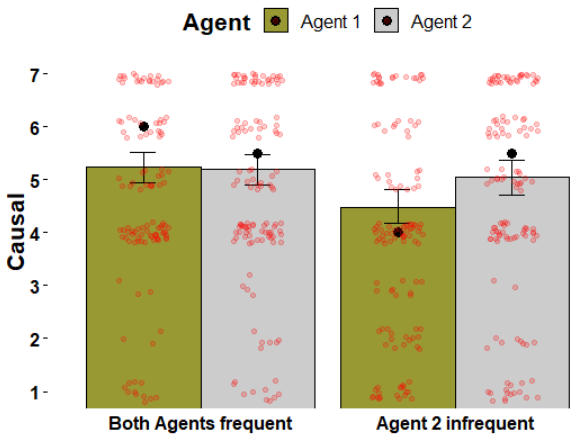


Figure 4. Mean agreement ratings (scale 1-7) for causal statements. Error bars represent ±1 SE mean, black points represent the median.

Subgroup Analysis. We conducted an additional analysis for the causal agreement ratings of the subgroup of people who rated the type behaviour expectations of normal and abnormal agent as equal (n=98). Here, we found no significant interaction for Frequency × Agent ($F_{(1,96)}$ = 1.6, p = .147) ($MD_{\text{Abnormal-Normal}}$=0.29, SD_{MD}=2.0).

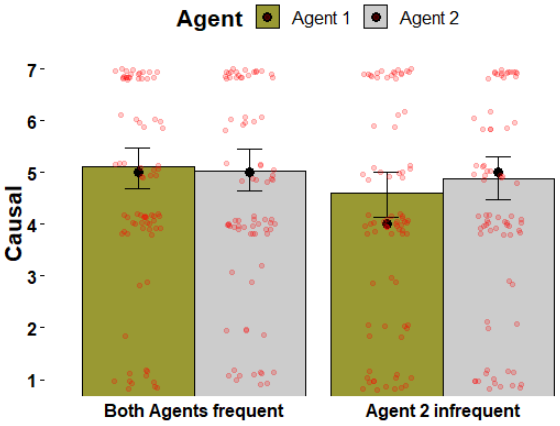


Figure 5. Mean agreement ratings (scale 1-7) for causal statements. Error bars represent ±1 SE mean, black points represent the median.

Discussion

In our third experiment, we found that an action that is token abnormal, but type normal, still influences causal judgments,. However, the judged difference between token normal and abnormal agent is significantly smaller than in Experiment 1. In addition, we found that people thought that the normally acting is less certain that the abnormally acting agent would act, but also less certain that *someone else* would act. This result comes as a surprise, given that both focal agents should have been able to expect *an agent* to act in the final scenario. While we assessed the focal agents' expectations towards the general type and each other's token behaviour, we did not assess their predictions about the behaviour of the third 'auxiliary' agent. It is therefore likely that some people might have assumed Agent 1 (and/or Agent 2) to have expected Agent's 3 omission. In consequence, it might be that the difference in action type expectations comes about as a difference in expectations about who *in fact* acted on Friday. This, again, leaves the normal agent with an epistemic disadvantage. However, a subgroup analysis showed that participants who assumed that both agents had equal behaviour type knowledge, i.e. that both agents were equally expecting that someone would act on Friday, did not judge a significant causal difference between abnormal and normal agent.

General Discussion

In three experiments, we investigated what we call the *Epistemic Hypothesis* (EP), the hypothesis that statistical abnormality will influence causal judgments via generating an epistemic asymmetry. In our first experiment, we showed that an abnormally acting agent is seen as more causally effective for an outcome, but also as more knowing about the behaviour of their normal counterpart. In accordance with EP, we found that in the case of mutual ignorance about each other, statistical abnormality does not influence causal judgements. Finally, we found that token abnormal, but type normal behaviour still influences causal judgments. At the same time, people's epistemic judgments about type and token behaviour mirror these causal judgments.

What role do epistemic states play in the influence of normality on causal judgements? Samland and Waldmann (2016) have shown that the mental states of agents can affect whether people's judgements about their causal contribution are influenced by prescriptive abnormality. They found that people do not take prescriptive norms into account for their causal judgments when the norm-violating agent is unaware of their norm transgression. Counterfactual accounts leave open under which circumstances people start to perceive a behaviour as "abnormal" (Phillips & Kominsky, 2018). Therefore, an agent's lack of knowledge or awareness of existing norms might determine whether the behaviour is perceived as norm-violating or abnormal in the first place. However, we think that in case of statistical normality, an agent can assess the normality status of their behaviour relative to their own action history, their agent-level normality. In consequence, the assessment of statistical normality is not necessarily conditional on the knowledge about external factors, such as rules or laws, or the behaviour of other people. In this paper, we aim to make different claim. We argue that it is the epistemic state that occurs qua the normality or abnormality of an action that drives the difference in people's causal judgements (Kirfel & Lagnado, 2017; Kirfel & Lagnado, 2018). Our experiments support this hypothesis. Hence, we argue that current norm incorporating causal frameworks are in need of a firm theory of epistemic states in order to explain their influence on norm-based causal cognition.

References

Alicke, M. D., Rose, D., & Bloom, D. (2011). Causation, Norm Violation, and Culpable Control. *Journal of Philosophy, 108*(12), 670–696.

Bear, A., & Knobe, J. (2017). Normality: Part descriptive, part prescriptive. Cognition, 167, 25–37.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review, 104*(2), 367–405.

Cheng, P.W., & Novick, L.R. (1991). Causes versus enabling conditions. *Cognition*, 40, 83-120

Gerstenberg, T. & Icard, T.F. (submitted). Expectations affect physical causation judgments. Pre-print.

Hagmayer, Y., Sloman, S.A., Lagnado, D.A. and Waldmann, M.R. (2007): Causal Reasoning through Intervention. In: Gopnik, A. and Schulz, L., Eds., *Causal Learning: Psychology, Philosophy, and Computation*, Oxford University Press, Oxford, 86-100.

Halpern, J. & Hitchcock, C. (2015). Graded Causation and Default. *British Journal for the Philosophy of Science*, 66 (2):413-457.

Harinen, T. (2017). Normal Causes for Normal Effects: Reinvigorating the Correspondence Hypothesis About Judgments of Actual Causation. *Erkenntnis 82*(6) 1–22.

Hart, H. L. A., & Honoré, T. (1963). Causation in the Law, 2nd ed., Oxford: The Clarendon Press.

Hitchcock, C., & Knobe, J. (2009). Cause and Norm. *Journal of Philosophy*, 106(11), 587–612.

Icard, T. F., Kominsky, J. F., & Knobe, J. (2017). Normality and actual causal strength. *Cognition*, 161, 80–93.

Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review, 93*(2), 136–153.

Kirfel, L. & Lagnado, D. (2017). Oops, I did it again. The impact of frequency on causal judgements. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society*. (pp. 2420-2425). Austin, TX: Cognitive Science Society

Kirfel, L. & Lagnado, D. (2018). Statistical Norm Effects in Causal Cognition. In T.T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society*. (pp. 615-620). Austin, TX: Cognitive Science Society.

Lagnado, D. A., Waldmann, M. R., Hagmayer, Y., & Sloman, S. A. (2007). Beyond covariation: Cues to causal structure. In A. Gopnik & L. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation* (pp. 154-172). New York, NY, US: Oxford University Press.

Lagnado, D. A., & Channon, S. (2008). Judgments of cause and blame: The effects of intentionality and foreseeability. *Cognition, 108*, 754–770.

Lagnado, D. A., Gerstenberg, T., & Zultan, R. (2013). Causal responsibility and counterfactuals. *Cognitive Science, 37*(6), 1036–1073.

Phillips, J. S., & Kominsky, J. F. (2018). Immoral professors and malfunctioning tools: Counterfactual relevance accounts explain the effect of norm violations on causal selection. *Pre-print*.

Kominsky, J. F., Phillips, J., Gerstenberg, T., Lagnado, D., & Knobe, J. (2015). Causal superseding. *Cognition , 37*, 196–209.

Samland, J., & Waldmann, M. R. (2016). How prescriptive norms influence causal inferences. *Cognition, 156*. 156, 164-176.

Sytsma, J., Livengood, J., & Rose, D. (2012). Two types of typicality: rethinking the role of statistical typicality in ordinary causal attributions. *Studies in History and Philosophy of Biological and Biomedical Sciences*, 43(4), 814-820.

Modelling Emotion Based Reward Valuation with Computational Reinforcement Learning

Can Koluman (can.koluman@city.ac.uk)
Department of Computer Science, City, University of London
Northampton Square, London EC1V 0HB, UK

Christopher Child (C.Child@city.ac.uk)
Department of Computer Science, City, University of London
Northampton Square, London EC1V 0HB, UK

Tillmann Weyde (T.E.Weyde@city.ac.uk)
Department of Computer Science, City, University of London
Northampton Square, London EC1V 0HB, UK

Abstract

We show that computational reinforcement learning can model human decision making in the Iowa Gambling Task (IGT). The IGT is a card game, which tests decision making under uncertainty. In our experiments, we found that modulating learning rate decay in Q-learning, enables the approximation of both the behaviour of normal subjects and those who are emotionally impaired by ventromedial prefrontal lesions. Outcomes observed in impaired subjects are modeled by high learning rate decay, while low learning rate decay replicates healthy subjects under otherwise identical conditions. The ventromedial prefrontal cortex has been associated with emotion based reward valuation, and, the value function in reinforcement learning provides an analogous assessment mechanism. Thus reinforcement learning can provide a good model for the role of emotional reward as a modulator of the learning rate.

Keywords: reinforcement learning; Q-learning; learning rate decay; Iowa Gambling Task; ventromedial prefrontal impairment

Introduction

According to psycho-evolutionary theorists, emotions assist the organism in maintaining homeostasis relative to its behavioural and survival goals (Plutchik, 2003). The emotion feedback mechanism solves problems without the need for higher cognitive analysis (Damasio, 2006).[1] Rolls (2013, Ch. 4) proposes that emotions regulate instrumental learning and influence contingent outcome-action selection.

The pre-frontal cortex and its regions play a key role in goal directed learning and behaviour (Miller & Cohen, 2001). Ventromedial prefrontal cortex (VMF) lesions produce a characteristic learning deficit, where the subject, while retaining good intellectual function and understanding, is no longer able to learn from real life mistakes. Wallis (2007) has argued that the VMF provides emotion valuation input critical for good decision making.

The Iowa Gambling Task (IGT) was the first clinical test, which identified VMF impairment in human trials (Bechara, Damasio, Damasio, & Anderson, 1994). In the IGT, subjects need to choose a card from one of four decks. There are two 'good,' and two 'bad' decks, but the 'bad' decks start with positive rewards. Once penalties set in on the bad decks, subjects should adjust the choice of decks accordingly. Fellows and Farah (2005, 2004) present a re-shuffled variation of the original IGT, where penalties start earlier in the bad decks. While VMF impaired subjects fail the original IGT, they pass the re-shuffled variant. On the basis of these differing test results, Fellows and Farah (2005, 2004) link VMF impairment to reversal learning deficit.

[1] First published in 1994 by G.P. Putnam's Sons, New York, USA.

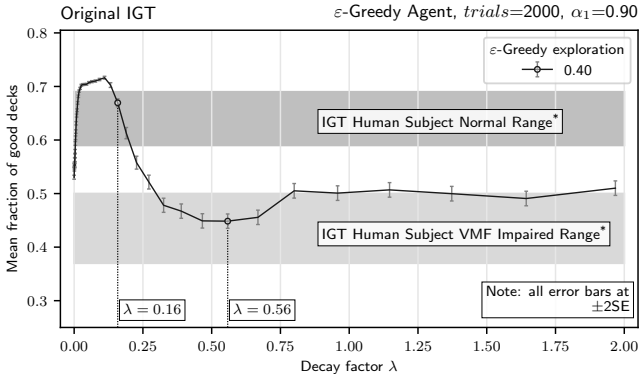


Figure 1: The ε-Greedy agent applied to the original IGT data with different learning rate decay values. A low decay rate yields normal behaviour, whereas a high decay rate reproduces VMF impairment. *See text and Table 4 for details.

Computational reinforcement learning methods approximate an optimal decision policy by iteratively aggregating time-contingent reward values (Sutton & Barto, 2018). For example, reinforcement learning techniques may be used to calculate a suitable path for escaping a maze (Osmankovic & Konjicija, 2011).

Watkins (1989) developed, the Q-learning framework, a reinforcement learning model, which, in addition to the discount rate, uses a single novel parameter denoted by α, known as the learning rate. The learning rate determines the relative contribution of current yield to accumulated value. We add to the Q-learning model a decay factor parameter λ, which produces exponential decay of the learning rate (Powell, 2011, pp. 427). We show that Q-learning, with the addition of learning rate decay, reproduces the clinical results of the original and re-shuffled IGT variants.

Figure 1 presents our key result for the original IGT variant. For low learning rate decay, the ε-Greedy agent performs in the range of normal human subjects. As learning rate decay increases, agent performance reduces to the range of VMF impaired subjects. The dark and light gray zones mark the mean fraction of cards chosen from the good decks, reported in the literature for normal and VMF impaired subjects respectively.

In the remainder of this paper, we first review related literature. Then we discuss the IGT in detail, develop the computational treatment, and summarize the experimental design and the results. These are followed by the discussion and conclusion.

Literature Review

According to Dalgleish (2004), the prefrontal cortex constitutes a primary anatomical locus for animal and human behaviour attributed to emotion. Dalgleish's prefrontal cortex includes the areas (Krawczyk, 2002, pp. 633-635) others have called the VMF (Bechara, 2004) or OFC (Rolls, 2000; Wallis, 2007). This paper uses the term VMF to refer to the area of the prefrontal cortex involved in valuation by emotion. However, some studies prefer the term orbitofrontal cortex, or OFC. This section retains the respective authors' original use of the terms OFC or VMF.

VMF impaired patients, can recognise poor decisions and describe good decision making strategies, but exhibit a distinctive inability to learn from their mistakes (Bechara et al., 1994). In IGT studies, this inability applies to negative (Bechara et al., 1994) and positive rewards (Bechara, Tranel, & Damasio, 2000).

To explain VMF impaired deficits, Damasio (1998) proposes the Somatic Marker Hypothesis: an involuntary feedback mechanism where a physical or virtual body sensation is associated with a particular emotion. VMF impairment disrupts somatic marker pathways, and the affected individual remains in a slow, logic based decision making paradigm (Bechara, 2004; Damasio, 1998, 2006). Others have instead advanced the view that VMF impairment leads to loss of reversal learning ability (Dunn, Dalgleish, & Lawrence, 2006; Maia & McClelland, 2005; Fellows & Farah, 2003, 2005, 2004). Reversal learning ability is the facility to unlearn a stimulus-response-association, which had previously produced favourable emotion-valued outcomes.

The VMF is also associated with emotion (Krawczyk, 2002; Hornak et al., 2003; Rolls, 2000). Modelling emotion in learning and decision making has been challenging (Volz & Hertwig, 2016). Without using emotion, the Rescorla-Wagner classical conditioning model presents a learning rule for assessing the pre and post trial associative strength of a new stimulus (Rescorla & Wagner, 1972). TD(λ) reinforcement learning methods extend the Rescorla-Wagner model and enable intra-trial assessment of an associative stimulus (Sutton & Barto, 2018, pp. 350-357). Contingent stimulus-response animal studies also inspired Q-learning. However, unlike Rescorla-Wagner, Q-learning does not explain the conditioning mechanism, but instead develops a decision theoretic learning framework (Watkins, 1989). Q-learning remains one of the most successful machine learning algorithms, especially as the feedback stage for deep neural networks (Mnih et al., 2015).

Puviani and Rama (2016) propose a complex, neurologically motivated emotion learning framework, which models both the OFC and the Amygdala. However, typically computational emotion synthesis employs more abstract, behaviourally driven approaches based on varied psychological views. Recently, reinforcement learning approaches incorporating emotion have been receiving increased attention. Reinforcement learning can produce lightweight models, has close ties to optimal control, and provides an intuitive approach for aggregating contingent values (Powell, 2011; Sutton & Barto, 2018).

Moerland, Broekens, and Jonker (2018; 2017) identify and survey 52 papers published from 1998 to 2016 relating to emotion and reinforcement learning. They report four common methods for eliciting emotion: homeostatic targets, introspective appraisal, value function or reward modulation, and, sensor or sense driven. Emotions influence rewards, contingencies, modulate the exploitation versus exploration trade-off, and sometimes directly act on action selection. Typically, the value function itself aggregates emotion modulated inputs into an action selection mapping. We believe that emotion modulated reinforcement learning thus aims to encapsulate the functionality of the VMF.

While developing our model, the Moerland et al. (2018; 2017) survey had not yet come out. However, we had considered Broekens, Jacobs, and Jonker (2015), where joy, distress, hope, and fear act as value inputs into TD(0) computational reinforcement learning. In contrast to Moerland et al. (2018; 2017) and Broekens et al. (2015), our model does not need an emotion generation layer. In the context of the discussed models, our model re-interprets the Q-value function as a single aggregated emotion signal. While our learning rate is modulated by another hyper-parameter, the decay factor, we do not synthesize emotions to modulate these hyper-parameters. Instead, we use an external search grid to assess the end-effect of learning rate changes, which we hypothesize might result from VMF impairment.

Our learning rate decay law does not satisfy the well-known statistical convergence requirement that the sum of the learning weights must be infinite (Robbins & Monro, 1951; Spall, 2003). In practice, fully proving theoretical statistical convergence is difficult (Spall, 2003, p. 122), and proof of theoretical convergence does not automatically ascertain good model performance (Powell, 2011, p. 450). Moreover, an individual organism and its decision making mechanisms possess a finite lifespan. Therefore we think it is valid to investigate finite term, periodic decisions with tools where statistical convergence is not theoretically guaranteed. We propose that our method of simulating human behaviour with

learning rate decay could form a useful baseline for generalised reinforcement learning solutions. We focus here on the empirical effect of learning rate decay on decision quality and learning.

The Iowa Gambling Task

The original (Bechara et al., 1994) and re-shuffled (Fellows & Farah, 2005, 2004) Iowa Gambling Task (IGT) variants form the basis of this paper and we explain them here in more detail.

Description

The IGT is a card game where the participant receives a loan, and should maximize profit including repayment of any loans. The card game consists of four decks: A, B, C, and D. The participants are told that "some decks are worse than others." (Bechara et al., 2000, p. 2192) In each turn, the participant draws one card from any deck. For each draw, the participant then receives a fixed reward, and occasionally has to pay a fine. Decks C and D, known as the 'good' decks, give low fixed rewards, low fines, and, on average, yield net gains. The remaining two 'bad' decks, A and B, produce high rewards, but even higher losses, and, on average, produce a net loss.

The game stops after 100 turns, when the dealer announces the end. However, the participant does not know when the game will end. If the participant runs out of money, additional loans are available. The hypothesis is that the participants discover the 'good,' low risk decks and choose accordingly. A score of more than 50 draws from the good decks is defined as a normative pass by Fellows and Farah (2005, 2004).

While the original IGT lasts 100 turns, Bechara et al. (1994) only predefine a 40-draw sequence for each deck. They do not discuss whether any participants drew more than 40 cards from the same deck, and in the provided example draws, human participants do not draw more than 40 cards from the same deck. In our implementation, we use the published 40-draw predefined sequences. However, to ensure that a software agent could potentially draw more than 40 consecutive cards from the same deck, we loop at the end of each deck to the beginning of the deck.

Original and Re-shuffled Card Deck Differences

In the original IGT, the 'bad' decks, A and B, each start with an eight card long 'special' sequence, where the player receives positive net gains. Consequently, at the beginning of the task, the 'bad' decks appear 'good.' However, in each bad deck, the 'special sequence' is immediately followed by one or more high fines, causing the player, on subsequent selections, to lose all gains and move into debt.

In the re-shuffled variant, Fellows and Farah (2005, 2004) move the first 8 cards in each original deck to the end. This removes the initial confounding conditioning sequence, and players experience, across all decks, fines relatively quickly.

The full details of the original and re-shuffled decks can be found in Bechara et al. (1994, p. 9) and Fellows and Farah (2005, 2004, p. 59) respectively.

ε-Greedy Q-Learning with Learning Rate Decay

This section motivates and develops our Q-learning model with learning rate decay.

Computational Background

The IGT constitutes a version of the n-armed bandit problem (Ross, 1983, pp. 131-151): there are four processes, of which only one can be operated at any one time. The software agent devises a policy for gaining information (exploring), for assessing (scoring), and then choosing the most advantageous process (exploiting). Kuleshov and Precup (2000) present various classic computational techniques for scoring, and for balancing exploration versus exploitation. We employ Q-learning because it is simple and permits investigation of learning rates which vary from $1/n$ and its derivatives.

Single State Q-learning

We model the IGT as a single state environment with four card decks and four actions. We do not fully implement Q-learning as proposed by Watkins (1989) where the current contribution to the Q-factors uses off-policy updating. Instead, we apply on-policy value function updates as suggested by Sutton and Barto (2018, p. 32).

Given an action a, let $Q(a)$ be an unknown value function, and let $Q_n(a)$ denote the n^{th} iterative approximation. Then we write the computational estimation problem as:

$$Q_n(a) = \alpha_n r_n^a + (1-\alpha_n)\gamma Q_{n-1}(a) \tag{1}$$

where $r_n^a = reward_n^a - fine_n^a$ is the net reward for action a at iteration n , γ is the discount rate, and α_n is the learning rate at iteration n. The discount rate γ, when set to less than 1, is used to devalue future yields r_n^a. We assume that the length of the card game, although unknown, is not long enough to create a preference for present rewards. Consequently, we set $\gamma = 1$.

Learning Rate Decay

A rapidly decaying learning rate sequence, $\{\alpha_n\}$, can get close to 0 prior to some final period T and effectively curtail learning. We consider a geometric-decay learning rate sequence of the form (Powell, 2011, pp. 427):

$$\alpha_n = \Lambda\alpha_{n-1} \tag{2a}$$

$$\Lambda = 2^{-\lambda/ln2} \tag{2b}$$

where $\lambda \in [0,\infty)$ is the *decay factor*, and $ln2$ is a normalizing constant used to rescale to natural logarithms in the computations.

Given equation (2b), $\{\alpha\}_n$ only satisfies the theoretical statistical convergence requirement $\sum_n \alpha_n = \infty$ (Powell, 2011, pp. 274-285), when $\lambda = 0$.

However, equations (2a) and (2b) always guarantee, in a finite number of iterations, computational convergence in the sense of $|Q_n - Q_{n-1}| < \varepsilon$ for some $n \ll \infty$ and $\varepsilon > 0$. In practice, our approach produces good approximations to normal as well as VMF impaired behaviour.

Table 1: Methodology, Simulation Parameter Summary

Trials, N	2000
Initial Learning Rate, α_1	0.05 to 1 by 0.05 steps
*Decay Factor, λ	$\lambda_i = \lambda_{max} 2^{-ir/ln2}$
ε-Greediness, ε	0.00 to 0.50 by 0.10 steps

*With $\lambda_{max} = 3.3765$, $r = 0.012$, $i = 0, 1, 2, ...$

Table 2: Original IGT Test, Pixel Match Computed Means $\pm$ SE for Fraction of Cards Chosen from the Good Decks reported in the IGT Literature

Subjects	Study	N	Mean fraction of good decks
Controls	Bechara et al. (1994)	44	0.69 ± 0.015
	Bechara et al. (1998)	21	0.62 ± 0.032
	Bechara et al. (2000)*	20	0.59 ± 0.019
	Farah et al. (2004)	14	0.63 ± 0.023
VMF Impaired	Bechara et al. (1994)	6	0.37 ± 0.055
	Bechara et al. (1998)	9	0.40 ± 0.035
	Bechara et al. (2000)*	10	0.45 ± 0.028
	Farah et al. (2004)	9	0.50 ± 0.020

*Results reported in 20 draw blocks. Calculation of 100 draw values assume no inter-block covariance.

The ε-Greedy Agent

For most of the time, the ε-Greedy agent exhibits unconstrained maximizing behaviour, and at any iteration n, picks the deck with the highest attributed value:

$$Q_n^* = \max_a Q_n(a),\ a \in \{A, B, C, D\} \tag{3}$$

To ensure exploration, occasionally the ε-Greedy agent chooses an action randomly. Consequently, the agent's decision making rule is:

$$Q_{n,\varepsilon}^* = \begin{cases} Q_n^*, & \text{with probability } 1-\varepsilon, \\ \text{choose } a \text{ randomly} & \text{with probability } \varepsilon \end{cases} \tag{4}$$

where $\varepsilon \in [0, 1]$ indicates the probability of exploration.

Experimental Design and Results

Simulations consist of multiple trials of 100 draws. All cross-section comparisons are conducted at the 100^{th} draw, which corresponds to the duration of the clinical tasks. Table 1 summarizes the parameter values used in this paper. We assess the parameter space with brute-force, grid-based searches.

As the original test data (Bechara et al., 1994, 2000; Fellows & Farah, 2005, 2004; Bechara, Damasio, Tranel, & Anderson, 1998) was not available, we converted the graphical presentations into numerical format using pixel matching. For each study, Tables 2 and 3 summarize, for normal and VMF impaired subjects, the pixel match calculated original and re-shuffled IGT test results respectively, reported in terms of the fraction of cards chosen from the good decks.

Table 3: Re-shuffled IGT Test, Pixel Match Computed Means $\pm$ SE for Fraction of Cards Chosen from the Good Decks reported in the IGT Literature

Subjects	Study	N	Mean fraction of good decks
Controls	Farah et al. (2004)	17	0.72 ± 0.038
VMF Impaired	Farah et al. (2004)	9	0.67 ± 0.078

Table 4: Original and Re-shuffled IGT Mean Fraction Good Deck Ranges Used for Comparing ε-Greedy Agent and Literature Results

IGT Variant	Original	Re-shuffled
Pixel matched studies Comparison Rule	4 Table 2 Minimum and Maximum	1 Table 3 $\pm$ 2 SEs
Normal Match Range	0.59 to 0.69	0.64 to 0.80
VMF Impaired Match Range	0.37 to 0.50	0.51 to 0.83

Table 4 shows the pixel matched ranges of fraction of good decks we derived from IGT literature results and use to compare to the ε-Greedy agent results.

Results

We found that, given appropriate standard values for initial learning rate and exploration, learning rate decay λ proves to be the key variable, which determines the ε-Greedy agent's degree of success. We first present the results obtained from learning rate decay and exploration variations, and then discuss the effects of the initial learning rate.

The Effects of Learning Rate Decay and Exploration Fig. 2 shows, given exploration, the strong effect of learning rate decay on mean fraction of good decks. For the original IGT, as the decay factor increases, the mean fraction good decks achieved by the agent decreases; and, eventually approaches a value close to or below 0.5, the IGT fail criterion. But for the re-shuffled IGT, as the decay factor increases, mean fraction of good decks scores remain above 0.5.

Figure 2 also shows that for the original and re-shuffled decks, at $\varepsilon = 0.40$, the ε-Greedy agent matches actual IGT test subject behaviours: control subject behaviour is matched at a learning rate decay factor of $\lambda = 0.16$ (15% per period learning rate decay), and VMF impaired subject behaviour is matched at $\lambda = 0.56$ (43% decay).

$\varepsilon = 0.40$ constitutes the first exploration value at which we obtain a match for healthy and VMF impaired human performance zones. Further match candidates exist for $\varepsilon =$

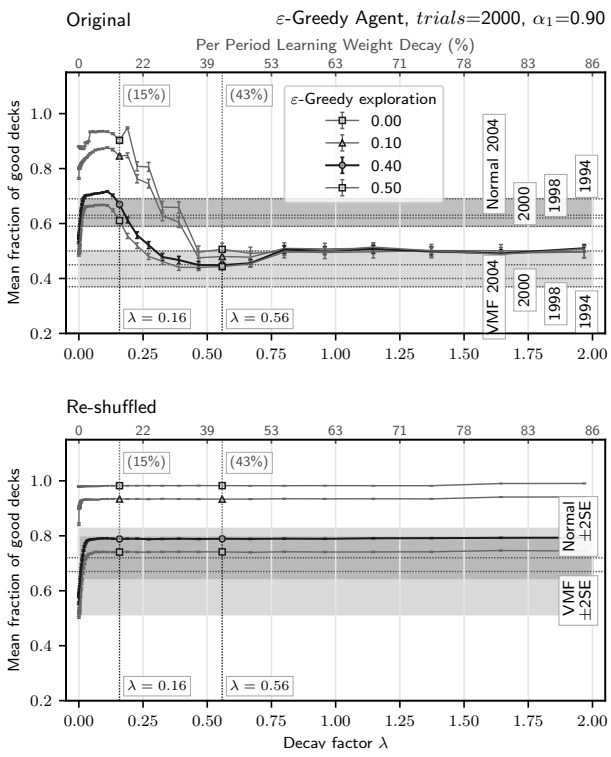


Figure 2: Learning rate decay and ε-Greedy agent IGT performance. Dark lines approximate human IGT behaviour. With high exploration, for the Original IGT, at the lower decay factor, the agent matches control subject performance; and, at the higher decay factor, the agent matches VMF impaired subject results. With high exploration, for the re-shuffled IGT, at the lower and higher decay factors, the agent matches human subject performance.

$0.40 - 0.50$, where the values of the agent's mean fraction of good decks are inside the match ranges for the corresponding values reported in the literature for human subjects.

At $\varepsilon = 0.50$, the agent explores 50% of the time. 50% exploration seems high. However, it constitutes a targeted strategy, for example, compared to always choosing lottery numbers randomly. We can also see that agents, which do not explore at all ($\varepsilon = 0.0$), or explore just a little ($\varepsilon = 0.10$), substantially exceed human performance. We discuss this result later.

Table 3 shows that the re-shuffled deck VMF impaired match range is derived from a single study with 9 participants. In Fellows and Farah (2005, 2004, p. 60, Figure 4), VMF impaired subject performance includes a high performace cluster of 3 subjects with a pixel matched cluster mean of 0.95. These 3 VMF impaired subjects achieve a re-shuffled deck test result approximated by the performance of our ε=0.10 agent, which achieves across all decay factors a mean fraction of good decks score of 0.92.

Having only a single re-shuffled deck study makes interpreting the statistical context of this high performance cluster difficult. Therefore in Table 4, we construct re-shuffled deck VMF performance match ranges using ± 2 standard errors, which produces approximately a 92% confidence interval (two-sided p-value: 0.080516). Our match range can be interpreted as the smallest match range based on the availability of a single study.

With re-shuffled decks, the decay factor λ influences the mean fraction of good decks by very little. This result appears to be driven by card sequencing. To test the effect of card sequencing, we created a new deck environment, where cards are drawn randomly, without replacement, from the original IGT decks. This new random draw card environment produces plots, which display a pattern similar to that of the original decks in Figure 2, except that as the decay factor increases, mean fraction of good decks decreases towards but remains above 0.5. Therefore relative to randomly ordered decks, both the original and re-shuffled decks create sequencing biases, which put different demands on learning: the original decks tax re-learning, while the re-shuffled decks teach via 'early punishment.' It would be interesting to test whether both normal and VMF impaired subjects pass the random draw version of the IGT as predicted by our simulation.

Finally, increasing exploration leads to a steady downward shift of the mean fraction of good decks plots with little effect on contour shaping. In contrast, learning rate decay λ appears key for determining agent behaviour; and increasing learning rate decay approximates the behaviour of normal and VMF impaired IGT participants.

The Effects of the Initial Learning Rate Unlike learning rate decay, the initial learning rate α_1, like exploration, only has a mild effect on the mean fraction of good decks.

Figure 3 shows the effect of the initial learning rate α_1 on mean fraction good decks at the 100th draw for the ε-Greedy agent with $\varepsilon = 0.40$. For the the original and re-shuffled decks, mean fraction of good decks scores vary little along the initial learning rate axis. In contrast, increasing learning rate decay leads to normative IGT fail (i.e., mean fraction of good decks ≤ 0.50) for the original decks; but not for the re-shuffled decks, thereby inducing agent behaviour to match human trial performance.

Discussion

In our Q-learning IGT simulations, learning rate decay λ constitutes a critical parameter. Increasing learning rate decay generates the observed behaviour of human IGT participants. For low learning decay factors, the ε-Greedy agent passes both the original and re-shuffled IGT. As we increase the learning decay factor, the agent fails the original test, while continuing to pass the re-shuffled variant. Therefore, increasing the decay factor leads to the learning behaviour of VMF impaired IGT participants.

In reinforcement learning, the software agent's internal valuation produces action selection. Rolls and others have ar-

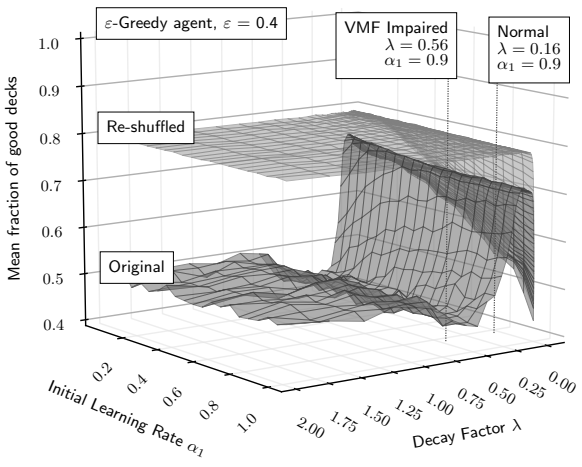


Figure 3: ε-Greedy agent with $\varepsilon = 0.40$: fraction of good decks by initial learning rate α_1, and decay factor λ. Initial learning rate α_1 variations exert a mild influence, while learning rate decay λ variations exert a strong influence on mean fraction of good decks. Normal and VMF impaired match values are marked accordingly for $\varepsilon = 0.40$.

gued that emotions result from reward assessment in the VMF (Krawczyk, 2002; Hornak et al., 2003; Rolls, 2000, 2013). We draw parallels between VMF provided reward values and the reinforcement learning process. The Q-value function encapsulates reward information. Learning rate decay λ can elicit progressive decay in current reward contribution. If learning rate decay is very high, then current reward value contribution decreases rapidly, and this leads to quick computational convergence. This effect produces two impediments, which may mimic VMF impairment: the value function not only 'finalises' too quickly, but also is itself dominated disproportionately by initial experiences.

Consequently with high learning rate decay, early and high 'bad' deck payoffs in the original IGT produce an incorrectly learned policy response: the 'bad' decks appear to be good. The ε-Greedy agent's beliefs, once established, even when presented with current information to the contrary, can no longer be modified. If emotion impairment due to VMF lesions removes the ability to unlearn previously learned responses, then in reinforcement learning, this behavioural effect can be achieved via high learning rate decay.

Conclusion

Bechara et al. (1994, in title) state that VMF impaired patients suffer from an "insensitivity to future consequences." Our simulated VMF impaired original IGT results suggest that this insensitivity comes from remaining mired in the past, and appears consistent with loss of the ability to reverse learning.

Interestingly, at lower exploration values, the ε-Greedy agent achieves mean fractions of good decks that are better than those achieved by human subjects. To match actual test subject behaviours, exploration has to be set at a high level.

It is not clear why agent behaviour, while qualitatively mirroring human behaviour, achieves better than human results. A number of possibilities could explain this finding. A reformulated model with decaying ε-Greediness may provide additional insight into the exploration versus exploitation tradeoff. Human behaviour may initially have higher exploration, which then progressively decreases with learning. In this paper, to keep the parameter count low, to avoid over-fitting, and to focus on the decay factor λ, we have not added any additional parameters for modelling variable exploration.

Alternatively, given the lack of full-knowledge, human behaviour may be more cautious. Human level learning has evolved for a wide variety of tasks, and therefore may perform optimally at other tasks for which Q-learning is less well suited. In contrast, grid search allows the searcher to become all-knowing with respect to the parameter space. For humans with incomplete information, keeping exploration high may make sense, just in case a deck would produce some unexpected yields later in the task.

Finally, it is also possible that the calculations performed by reinforcement learning agents are too hard for mental arithmetic and that the lack of precise calculations leads to sub-optimal decisions.

In a psycho-evolutionary context, emotions provide a flexible mechanism for establishing homoeostatis under environmental uncertainty (Plutchik, 2003; Rolls, 2013). If this environmental uncertainty fulfils certain regularity conditions, such as distributional full, or bounded, time-invariance, existence of the mean, or high-yield state correlation, then there could be high survival value to speculative learning; that is, deriving a working decision policy from just a few samples. From short learning bursts, the organism, or agent, could converge, to a long-term optimal decision rule. Emotions (via learning rate decay) could be responsible for opening and closing a short learning window. It is possible that the VMF driven emotion mechanism has evolved to produce the ability for organisms to learn efficiently from just a few samples.

Humans have evolved as generalised decision learners. In many machine learning tasks, only a narrow range of hyperparameter values produce a coherent result. Therefore the addition of a learning decay factor, which mimics human learning could provide an ideal starting point over a number of tasks for computational learning. Overall, our results indicate that computational reinforcement learning may be used as the basis for modelling emotion based learning. The results are encouraging for further investigation into more complex forms of learning and emotions.

References

Bechara, A. (2004). The role of emotion in decision-making: Evidence from neurological patients with orbitofrontal damage. *Brain and Cognition*, *55*(1), 30–40.

Bechara, A., Damasio, A. R., Damasio, H., & Anderson, S. W. (1994). Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition*, *50*(1–3), 7–15.

Bechara, A., Damasio, H., Tranel, D., & Anderson, S. W. (1998). Dissociation Of Working Memory from Decision Making within the Human Prefrontal Cortex. *Journal of Neuroscience*, *18*(1), 428–437.

Bechara, A., Tranel, D., & Damasio, H. (2000). Characterization of the decision-making deficit of patients with ventromedial prefrontal cortex lesions. *Brain*, *123*(11), 2189–2202.

Broekens, J., Jacobs, E., & Jonker, C. M. (2015). A reinforcement learning model of joy, distress, hope and fear. *Connection Science*, *27*(3), 215-233.

Dalgleish, T. (2004). Timeline: The emotional brain. *Nature Reviews Neuroscience*, *5*(7), 583–589.

Damasio, A. R. (1998). The somatic marker hypothesis and the possible functions of the prefrontal cortex. In A. C. Roberts, T. W. Robbins, & L. Weiskrantz (Eds.), *The prefrontal cortex: Executive and cognitive functions.* Oxford: Oxford University Press.

Damasio, A. R. (2006). *Descartes' Error: Emotion, Reason and the Human Brain*. London: Vintage.

Dunn, B. D., Dalgleish, T., & Lawrence, A. D. (2006). The somatic marker hypothesis: A critical evaluation. *Neuroscience and Biobehavioral Reviews*, *30*(2), 239–271.

Fellows, L. K., & Farah, M. J. (2003). Ventromedial frontal cortex mediates affective shifting in humans: evidence from a reversal learning paradigm. *Brain*, *126*(8), 1830–1837.

Fellows, L. K., & Farah, M. J. (2005, 2004). Different Underlying Impairments in Decision-Making Following Ventromedial and Dorsolateral Frontal Lobe Damage in Humans. *Cerebral Cortex*, *15*(1), 58–63.

Hornak, J., Bramham, J., Rolls, E. T., Morris, R. G., O'Doherty, J., Bullock, P. R., & Polkey, C. E. (2003). Changes in emotion after circumscribed surgical lesions of the orbitofrontal and cingulate cortices. *Brain*, *126*(7), 1691–1712.

Krawczyk, D. C. (2002). Contributions of the prefrontal cortex to the neural basis of human decision making. *Neuroscience and Biobehavioral Reviews*, *26*(6), 631–664.

Kuleshov, V., & Precup, D. (2000). Algorithms for Multi-Armed Bandit Problems. *Journal of Machine Learning Research*, *1*, 1–48.

Maia, T. V., & McClelland, J. L. (2005). The somatic marker hypothesis: still many questions but no answers. *Trends in Cognitive Sciences*, *9*(4), 162–164.

Miller, E. K., & Cohen, J. D. (2001). An Integrative Theory of Prefrontal Cortex Function. *Annual Review of Neuroscience*, *24*(1), 167–202.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529–533.

Moerland, T. M., Broekens, J., & Jonker, C. M. (2018; 2017). Emotion in reinforcement learning agents and robots: a survey. *Machine Learning*, *107*(2), 443–480.

Osmankovic, D., & Konjicija, S. (2011). Implementation of Q - Learning algorithm for solving maze problem. In *Proceedings of the 34th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)* (pp. 1619–1622). USA: IEEE.

Plutchik, R. (2003). *Emotions and Life : Perspectives from Psychology, Biology, and Evolution*. Washington, DC: American Psychological Association.

Powell, W. B. (2011). *Approximate Dynamic Programming: Solving the Curses of Dimensionality* (2nd ed.). Hoboken, N.J: Wiley.

Puviani, L., & Rama, S. (2016). A System Computational Model of Implicit Emotional Learning. *Frontiers in Computational Neuroscience*, *25*(56), 1–26.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of pavlovian conditioning: Variations on the effectiveness of reinforcement and non-reinforcement. In A. H. Black & W. F. Prokasy (Eds.), . New York: Appleton-Century-Crofts.

Robbins, H., & Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, *22*(3), 400-407.

Rolls, E. T. (2000). The Orbitofrontal Cortex and Reward. *Cerebral Cortex*, *10*(3), 284–294.

Rolls, E. T. (2013). *Emotion and Decision-Making Explained* (First ed.). Oxford New York, NY: Oxford University Press.

Ross, S. M. (1983). *Introduction to Stochastic Dynamic Programming*. New York, New York: Academic Press, Inc.

Spall, J. C. (2003). *Introduction to stochastic search and optimization: Estimation, simulation, and control*. Chichester; Hoboken, N.J: Wiley-Interscience.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction* (Second ed.). London; Cambridge, Massaschusetts: MIT Press.

Volz, K. G., & Hertwig, R. (2016). Emotions and Decisions: Beyond Conceptual Vagueness and the Rationality Muddle. *Perspectives on Psychological Science*, *11*(1), 101–116.

Wallis, J. D. (2007). Orbitofrontal Cortex and Its Contribution to Decision-Making. *Annual Review of Neuroscience*, *30*(1), 31–56.

Watkins, C. (1989). *Learning From Delayed Rewards*. Doctoral dissertation, King's College, London, UK.

The Effects of Embodiment and Social Eye-Gaze in Conversational Agents

Dimosthenis Kontogiorgos, Gabriel Skantze, Andre Pereira, and Joakim Gustafson
(diko@kth.se, gabriel@speech.kth.se, atap@kth.se, jocke@speech.kth.se)
KTH Royal Institute of Technology
Stockholm, Sweden

Abstract

The adoption of conversational agents is growing at a rapid pace. Agents however, are not optimised to simulate key social aspects of situated human conversational environments. Humans are intellectually biased towards social activity when facing more anthropomorphic agents or when presented with subtle social cues. In this work, we explore the effects of simulating anthropomorphism and social eye-gaze in three conversational agents. We tested whether subjects' visual attention would be similar to agents in different forms of embodiment and social eye-gaze. In a within-subject situated interaction study (N=30), we asked subjects to engage in task-oriented dialogue with a smart speaker and two variations of a social robot. We observed shifting of interactive behaviour by human users, as shown in differences in behavioural and objective measures. With a trade-off in task performance, social facilitation is higher with more anthropomorphic social agents when performing the same task.

Keywords: human-computer interaction, social agents, conversational artificial intelligence, smart speakers, social robots

Introduction

With conversational AI and domestic technology on the rise, several questions remain open on how humans engage with interactive agents when in different forms of embodiment and social behaviour. A wide range of interaction modalities has been researched for agents, that come in various forms such as smart speakers (Alam, Reaz, & Ali, 2012), and social robots (Breazeal, Dautenhahn, & Kanda, 2016). However, by design, social robots provide additional modes of pragmatic communication. Social robots can express their internal state using not only speech but also non-verbal behaviour. By generating multimodal communicative behaviours (Breazeal & Fitzpatrick, 2000; Mizoguchi, Sato, Takagi, Nakao, & Hatamura, 1997), social robots enable different manifestations of interaction similar to how humans interact with each other (Shibata, Tashima, & Tanie, 1999).

In the fields of human-computer interaction and human-robot interaction, anthropomorphism is often leveraged as a way to make machines more comfortable to use. The additional comfort comes from ascribing human features to machines with the aim to simplify the complexity of technology (Marakas, Johnson, & Palmer, 2000; Moon & Nass, 1996). While interactions between humans include many subtle social cues that we take for granted, 'face-to-face' interactions are still considered to be the gold standard of communication when interacting with either humans or conversational agents (Adalgeirsson & Breazeal, 2010). Therefore, agents need to employ anthropomorphic designs and a rich set of social behaviours to be considered as socially intelligent partners in interactions.

Figure 1: Situated interaction with a human-like social robot.

Many social robots do employ these elements, especially the ones with a human-like design, and provide the possibility of generating non-verbal social behaviours in their interactions with humans (Fong, Nourbakhsh, & Dautenhahn, 2003). Many of these behavioural elements are subtle social cues (e.g. gaze shifts and facial expressions), that are highly important for situated human conversational environments. One reason why face-to-face interaction is preferred is that a lot of familiar information is encoded in the non-verbal cues that are being exchanged. However, generating and interpreting these cues, induces higher levels of cognitive load (Torta, Oberzaucher, Werner, Cuijpers, & Juola, 2013) and may therefore increase interaction time. This suggests that human-like conversational agents that can express patterned non-verbal behaviours can cause social facilitation in users, but may be less efficient in task performance.

In this paper we contribute to this emerging field with a two-fold empirical evaluation of the elements of: 1) *anthropomorphic design* and 2) *non-verbal social behaviour* in conversational agents. We study whether a human-like face (i.e. a social robot), capable of displaying non-verbal cues, shifts interactive behaviour in comparison to a voice-only conversational agent (i.e. a smart speaker), that does not employ these multimodal features. Our contribution consists of a user study that was conducted with participants interacting with a smart speaker and a social robot collaborating in dialogue. To comprehend the effects of the comparison further, we test whether it is the anthropomorphic face or the social eye-gaze features that contribute to the perceived differences and remove the non-verbal behaviour of the social robot in a third condition.

The aim of the study was therefore to investigate the following research question:

- What are the effects in human behaviour when simulating anthropomorphism and social eye-gaze in conversational agents?

Related work

Conversational agents have become ubiquitous, and they are embedded in various forms and embodiments, from smart phones to voice-based smart speakers such as Amazon Echo and Google Home. There seems to be an interest in literature on how different representations of physical embodiment and anthropomorphic features affect the perception of social presence and facilitation in agents. Studies have compared agents in digital screens to social robots (Torta et al., 2013; Kidd & Breazeal, 2008) and have shown that anthropomorphic agents that are physically co-located are generally preferred and perceived to be more socially present than their virtually embodied versions (Kennedy, Baxter, & Belpaeme, 2015; Lee, Jung, Kim, & Kim, 2006; Kidd & Breazeal, 2004; Bainbridge, Hart, Kim, & Scassellati, 2008; Koda & Ishioh, 2018; Jung & Lee, 2004; Thellman, Silvervarg, Gulz, & Ziemke, 2016), or remote video representations of the same agents (Powers, Kiesler, Fussell, Fussell, & Torrey, 2007; Wainer, Feil-Seifer, Shell, & Mataric, 2006). Other studies have shown that social robots' perceived situation awareness is higher (Luria, Hoffman, & Zuckerman, 2017), and by adding non-verbal cues, the same agent is perceived more socially present (Pereira, Prada, & Paiva, 2014; Goble & Edwards, 2018).

Anthropomorphic agents take advantage of design elements afforded in their shape and movements (Gomez, Szapiro, Galindo, & Nakamura, 2018). Social robots in particular, raise expectations on how sophisticated they are in their actions and how socially intelligent they are perceived. A very human-like agent will make humans expect a higher degree of interaction and *social facilitation*, which is essential when designing the physical appearance of a social agent. However, it is not just the physical embodiment of the robot that has implications on its perceived social presence, but the behaviour and actions of the robot as well (Straub, 2016).

Socially interactive agents that make use of social behaviour features promise an opportunity to bring social values into computing and help coordination between humans and machines by taking advantage of their social cues and intentions (Dourish, 2004). While conversational interfaces manifest intent recognition using language and dialogue, social robots as embodied interfaces, communicate intentions with the use of multimodal cues, and additionally encourage users to anticipate joint actions and shared intent in the same physical space (Luria et al., 2017).

Non-verbal behaviour is used for communication and social coordination. The more human-like the agents' responses, the more they are attributed as social actors (Nass & Steuer, 1993). Social eye-gaze in particular, refers to the communicative cues of eye contact between humans and is classified to 4 main archetypes (Admoni & Scassellati, 2017): 1) *Mutual gaze* where interlocutors attention is directed at each other, 2) *Joint attention* where interlocutor's focus their attention on the same object, 3) *Referential gaze* which is directed to an object and often comes with referring language, and 4) *Gaze aversions* that typically avert from the main direction of gaze -i.e. the interlocutors face.

(a) Smart Speaker (b) Social Robot

Figure 2: The conversational agents used in the study.

The current work differs and in part extends the discussed studies. First, we simulate both anthropomorphism and non-verbal behaviour in the same study, and second we apply the comparison in only physically present voice-based agents, where we discuss the implications of *social eye-gaze* against *task performance*. Is a human-like face sufficient to cause social facilitation or is non-verbal social behaviour also needed when interacting with conversational agents?

Method

In order to investigate the impact of anthropomorphism and social eye-gaze in this study, we chose three conversational agents and a human trial. All agents engaged in human-agent interaction using the same dialogue policy and simulated situation awareness of human actions (Figure 1).

Experimental conditions

1. The *Human Agent (H)*. In order to avoid any misunderstandings on the task and the subjects' role, we began the interactions with a control trial with a human instructor. That way, subjects got familiar with the task and we were able to reduce the learning curve.

2. The *Smart Speaker (SS)* is an embodied conversational agent (Figure 2a) that can only interact with speech. We used a first generation Amazon Echo smart speaker, which was connected via Bluetooth and a Text-to-speech (TTS) service similar to the default Echo TTS was selected to send pre-scripted voice commands.

3. The *AnthropoMorphic Robot (AMR)* is an embodied conversational agent (Figure 2b) in the form of a robotic head with a human-like face, that as the SS uses only speech to interact and no other modalities. We used a back-projected human-like robotic head with three degrees of freedom called

Furhat. The robot was stationary and did not use any head movements, but statically looked at the subject. The robot had a TTS of equivalent quality to Echo, speaking the same pre-scripted utterances. The reason for choosing a robotic head instead of a full-body embodied robot is that it limits the modalities of communication, making it easier to control for comparison to a smart speaker.

4. The *AnthropoMorphic Social Robot (AMSR)* is the same robotic head as AMR that also uses voice for interaction and additionally generates a set of social eye-gaze behaviours using head movement. These included task-based functional behaviours such as gazing to the ingredient during a referring expression and a turn-taking gaze mechanism.

Hypotheses

Towards answering the research question defined above, we posed the following hypotheses:

- ☐ **H1.** We expected that a robot with non-verbal social behaviour will be perceived to be more socially present (Pereira et al., 2014). *The AMSR will cause more social facilitation with human users than the SS and the AMR.*
- ☐ **H2.** While non-vebal behaviour should cause more social facilitation, a human-like design without non-verbal cues should not induce the same differences. *Differences in social facilitation will not apply between the SS and AMR.*
- ☐ **H3.** *There will not be any difference in task performance across the agents.*
- ☐ **H4.** As a conversational partner, *the AMSR will generally be preferred for the task.*

Experimental design

A within-subject design was used in a study (Kontogiorgos, Pereira, Andersson, et al., 2019) where participants interacted with all four agents. To test our hypotheses, we manipulated two independent variables [*embodiment* and *social eye-gaze*], in three conditions [*Smart Speaker (SS)*, *AntropoMorphic Robot (AMR)*, *AntropoMorphic Social Robot (AMSR)*], presented in different orders to participants using a Latin Square, and a *human trial* that was always first.

Task

We asked subjects to cook 4 variations of fresh spring rolls without providing the recipes; they had to find out the recipes while interacting with the agents. Different varieties of ingredients and amounts were used. The setup also included ingredients not used in any of the recipes, encouraging participants to interact with the agents to find out the correct ingredients for each recipe. The task was the same in each condition, but different recipes were used (varied across conditions).

To ensure participants would engage with the agents more, they were told that if they followed the recipe with the correct ingredients and amounts, they would take the food with them at the end of the experiment. Counting the time participants took to cook the recipe served as a measure of the time they spent engaging with each agent. We had a total of 20 ingredients and a recipe typically included 7 ingredients to prepare.

All agents used a combination of nouns, adjectives and spatial indexicals as linguistic indicators to identify ingredients on the table, "The *cucumber* is the *green* thing *on the right*". AMSR however, also gazed at the referent ingredients (typically 0.5s prior to the reference). The agent's role in the task was therefore to instruct and the subject's role was to assemble the ingredients together.

Dialogue policy

All agents followed the same dialogue policy and interaction protocol, which was defined upon a set of *dialogue acts* within the action space of the interaction (Kontogiorgos, Pereira, & Gustafson, 2019). Given a human action or utterance, an appropriate response was selected from the dialogue policy. Driven by the possible set of actions, agent utterances are aggregated to higher level *dialogue acts* that describe the current state of the conversation. The dialogue acts model user actions, user utterances and any changes in the environment. An example dialogue:

USER : [FINISHED ACTION] What's next?
AGENT : [INSTRUCTION] Next, take three pieces of lettuce and put it in the spring roll.
USER : [CLARIFICATION-Q] Where is the lettuce?
AGENT : [CLARIFICATION-A] The lettuce is the green thing in the middle.
USER : [STARTED ACTION] Uh, yes!

To dismiss potential problems in speech recognition and language understanding, we used a human wizard (WoZ) to control the behaviours of the agents in timings and decision making. The social behaviours were designed to maintain a socially contingent interaction with the subjects, and in order to keep the dialogues between the subjects and the agents consistent for comparison across conditions. The human wizard had to select the appropriate agent response, triggered only by the state of the environment and user actions. The wizard application and dialogue acts were the same across all conditions. For every dialogue act, a set of predefined utterances was available, that the system would choose at random to generate, given the current dialogue act. The wizard therefore indicated only the current dialogue act in conversation.

Gaze for facilitating turn-taking

Gaze has been shown to be important for regulating conversational turn-taking, as people look towards the listener at the end of their utterances to indicate they have finished their turn (Kendon, 1967). Employing such a behaviour in agents, leads to human-like conversational turn-taking where each participant waits for the speaker's utterance before taking an action (Skantze, Hjalmarsson, & Oertel, 2014). In order to facilitate natural turn-taking mechanisms from the agent, we defined a heuristic gaze model on timings for turn-taking gaze and referential gaze to objects.

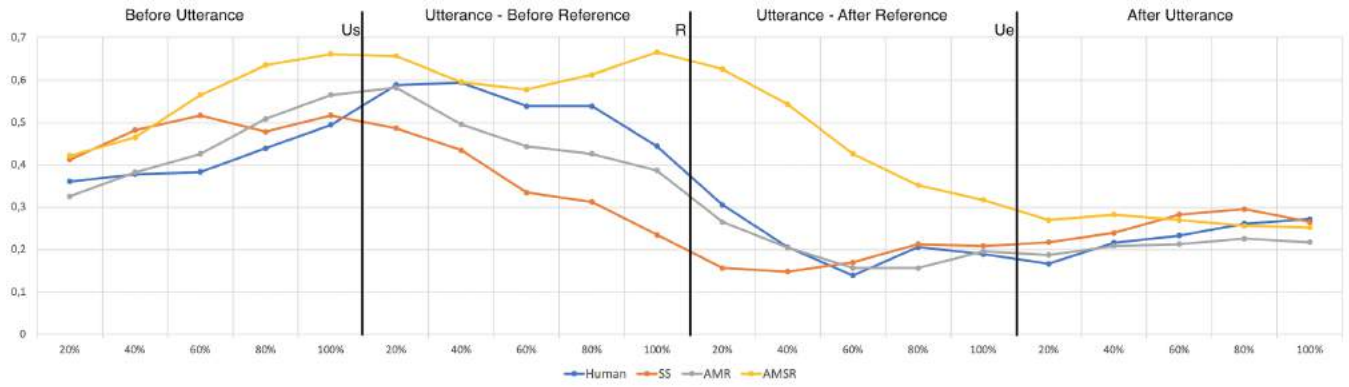


Figure 3: Gaze proportion to the agent during an agent instruction. Each phase of the instruction is normalised in time: Before the utterance - During the Utterance and before the reference - During the utterance and after the reference - After the utterance. The x axis shows the relative time of the instruction and the y axis the eye-gaze proportion across all participants per condition.

The AMSR agent engaged in *mutual gaze* and *joint attention* with the subjects during the interactions. Before an utterance, the agent made a gaze shift to the subject to establish attention, followed by deictic gaze to a referent object indicating it is keeping the floor, and at the end of the utterance a gaze shift back at the subject to establish the end of the turn and pass the floor to the subject. The agent gazed at referent objects right before they were mentioned (500ms before).

Experimental procedure

Participation in the study was individual and the experiment was divided in 3 phases. First, participants filled a demographics questionnaire and then cooked the first recipe with a human instructor. In the second phase, they cooked a recipe with the help of an agent. They repeated that phase 3 times with a new agent every time (counter-balanced). In the third phase, participants filled an exit questionnaire. During the agent trials, participants were alone in the room, and a human wizard was monitoring their actions using a ceiling camera with a live feed of the room (Figure 1). Participants were not told that the agents were controlled by a human wizard.

The human instructor throughout the trial was kept the same for all subjects, and followed the same behaviour and dialogue policy as the agents. The subjects stood in front of a table, with a cutting board and ingredients prepared and laid out in front of them. The agents were situated on the sides of the table, with only the agent relevant to the task visible.

Participants

Participants were compensated with a cinema ticket and the food they cooked during the study. We recruited 30 participants (18 female and 12 male) with ages in range 19-42 and mean 24.2 (stdev=5). The experiment was in English, and all participants were fluent (mean 5.8, stdev=0.7). 17 had interacted with a robot before and 20 had interacted with smart speaker. 13 had interacted with both a smart speaker and a robot before, while 6 with none of the two. Overall, their experience with digital technology was 4.8 (stdev=1.6) and their cooking skills were 5.0 (stdev=1.2). 24 had never cooked spring rolls recipes before. All scales above are 1-7.

Results

We present the main findings along two main themes: a) *visual attention*, and b) *interaction time*. Repeated measures analyses of variance (ANOVA) and post-hoc tests with Bonferroni corrections were carried out to test statistical differences across conditions. We report the behavioural and objective measures on visual attention during the agent utterances, interaction times (Table 2), and finally, notable insights from qualitative data.

Visual attention

Using motion capture, we detected subjects' head pose over time and measuring their visual angle (Kontogiorgos et al., 2018), we extracted proportional eye-gaze to the agent and the task table during different phases of the robot's utterances: a) before the robot speaks an utterance, b) during the utterance right before a reference to an object is uttered, c) during an utterance right after the reference has occured, and d) after the utterance. The four phases of proportional eye-gaze to the agent are presented in figure 3. Before and after the utterance phases are in 2 second intervals.

Each phase is first normalised per subject to reduce subject variability and then, each phase interval mean is used for comparison (Table 1). It is important to note that while agent conditions were counter-balanced in order, the human trial was always first to get familiar with the task.

Eye-gaze to the agent before the utterance. A repeated measures ANOVA to test the effect of gaze before the robot instruction showed a significant main effect, Wilks' Lambda = .674, $F(3,27) = 4.35$, $p = .013$). Post-hoc tests with a Bonferroni correction, and p value adjusted for multiple comparisons, revealed that gaze towards AMSR is statistically different than gaze to the AMR condition (p=.022) and to the Human trial (p=.029). No other statistical differences were found in pairwise comparisons.

	Human Trial		SS		AMR		AMSR	
	Mean	*SD*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
Before Utterance	.4008	.1822	.4740	.2083	.4476	.2083	**.5472**	.2028
During Utterance (Before Reference)	.5402	.2131	**.3609**	.2321	.4659	.2057	**.6239**	.2114
During Utterance (After Reference)	.2119	.1749	.1836	.1605	.2006	.1320	**.4514**	.1851
After Utterance	.2273	.1570	.2524	.1372	.2072	.1301	.2599	.1633

Table 1: Mean eye-gaze to agent in different phases of the agent utterances. Each phase is normalised per subject and each phase interval mean is used for comparison.

	SS	AMR	AMSR
Task time (sec)	212.6 ± 7.93	217.2 ± 7.75	232.9 ± 8.52

Table 2: Interaction time in seconds: Each cell shows mean and standard error of the mean.

Eye-gaze to the agent during utterance (before the reference). A repeated measures ANOVA on the gaze before the reference showed a significant main effect, Wilks' Lambda = .483, F(3,27) = 9.65, $p < .001$). Post-hoc tests with a Bonferroni correction and p value adjusted for multiple comparisons revealed that gaze towards AMSR is statistically different than gaze to SS ($p<.001$) and AMR (p=.001). SS was also different than the Human trial (p=.022). No other statistical differences were found in pairwise comparisons between the rest of the conditions.

Eye-gaze to the agent during utterance (after the reference). A repeated measures ANOVA on the gaze after the reference revealed a significant main effect, Wilks' Lambda = .316, F(3,27) = 19.49, $p < .001$). Post-hoc tests with a Bonferroni correction and p value adjusted for multiple comparisons showed that gaze towards AMSR is statistically different than gaze to all other conditions ($p<.001$) and to the Human trial ($p<.001$). Here as well, no other statistical differences were found in pairwise comparisons between the rest of the conditions.

Eye-gaze to the agent after the utterance. A repeated measures ANOVA on the gaze after the robot instruction showed no statistically significant difference across conditions, Wilks' Lambda = .859, F(3,27) = 1.47, p = .244).

Interaction time

As indicators to task performance we measured *task time* (time from first to last agent action). We tested for comparison in time within the sequence of the conditions, and no statistical difference was found, meaning the condition sequence did not affect task performance (subjects were not significantly faster in the last trial). When compared across conditions however, a repeated measures ANOVA revealed a significant main effect, Wilks' Lambda = .739, F(2,28) = 4.94, p = .014). Post-hoc tests with a Bonferroni correction and p value adjusted for multiple comparisons showed a significant effect between AMSR to SS (p=.041) and AMR (p=.023) but there was no evidence of a difference between SS and AMR (Table 2).

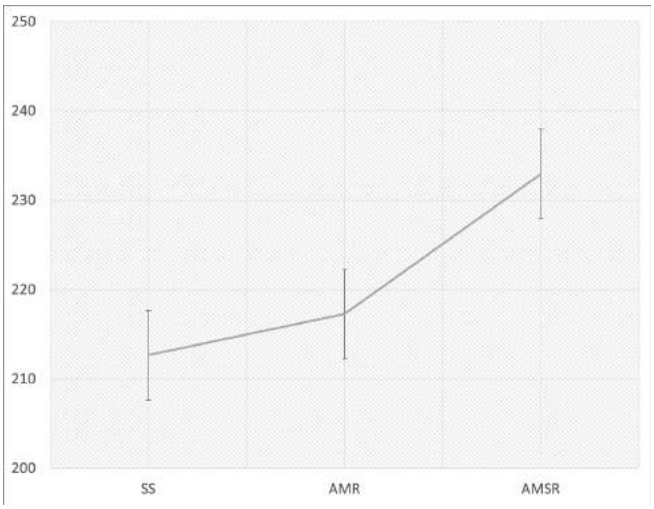


Figure 4: Mean *interaction time* per condition. Error bars indicate standard error of the mean (n = 30).

Qualitative data

The post-experimental questionnaire included asking participants to choose their preferred agent for the task and questions to elaborate on the preference. Participants were also asked to identify the differences of the three agents to understand if they are aware of what is tested in the experiment.

Perceived differences between agents. Out of the 30 participants, 18 replied this question. While the differences in the agent embodiment were obvious between [SS] and [AMR/AMSR], 66% of the participants did not notice a difference between [AMR] and [AMSR]. Asking participants further, we found they identified that there was head movement from the social robots, but were not aware that only one of them [AMSR] employed that behaviour.

Preferred agent. 69% of the participants preferred AMSR, while 24% preferred AMR and 7% preferred SS (χ^2 = 17.862, $p < .001$). Looking further at the participants who did not notice a difference between AMR and AMSR, 2/3 chose AMSR as the preferred robot for the task. However, from 1/3 of the participants who identified the difference in gaze, therefore less sensitive to our manipulation, all preferred AMSR.

Discussion

In an experiment with human subjects, we found a lack of positive effects in task performance on interactions with anthropomorphic social agents. Nonetheless, our findings show a higher degree of social facilitation in conversation with AMSR, as determined by subjects' visual attention and agent preference. Our strongest finding therefore is a trade-off between interaction time and social facilitation.

Anthropomorphism

The agents we compared represent different levels of embodiment in conversational agents. The most preferred agent for

the task had an anthropomorphic embodiment and a set of social eye-gaze behaviours. While AMSR was preferred, task time was increased by 10% with this agent in comparison to the less anthropomorphic in physical embodiment SS. We saw that participants looked at AMSR longer after the referent word was uttered and started following up on the agent's instruction close to the end of its turn. Intuitively, a turn-taking gaze mechanism invokes subjects a greater feeling of social facilitation, assuming they attribute that agent the role of a more socially present partner in conversation.

Non-verbal social behaviour

AMSR has joint attention afforded as an embodied phenomenon in its actions. Eye-gaze here is attributed as a social function where it regulates turn-taking, closer to how humans do when they interact with each other. AMSR therefore gave the impression that it is aware of the situatedness of the task.

In cases, it is possible the user may be distracted from the task through agent social behaviour because more attention is required to the agent's behaviour. While face-to-face collaboration is favourable due to its natural mediated channels of communication, interpreting social cues and maintaining attention is a timely and cognitively demanding process.

Social behaviour and task performance

Social behaviour is timely and counter-intuitive to task performance with more attentive agents. Nevertheless, task performance is certainly dependent on the nature of the task; in more task-oriented domains, such as emergency management, interactions may be more efficiency-prone. A human user may want to get the task done as quickly as possible, and get frustrated when having to interact longer than necessary. However, other tasks such as in the home-care domain are very dependent on social cues and interaction value.

As mentioned, referring expressions to objects did not contain any ambiguities in language (i.e. "this one here"). Therefore gaze from AMSR did not add value to task success but was attributed to a social function, as humans typically gaze at objects before mentioning them in language. Our purpose in the gaze condition (AMSR) was therefore not to increase task performance but to observe the social functions of gaze behaviour across agents.

We were able to verify hypothesis **[H1]**, that AMSR will cause social facilitation, as shown in the visual attention and preference dimensions, however with the cost of task performance. We did verify **[H2]** in the assumption that SS and AMR will not be different in social facilitation. In fact, a human-like design is not enough to establish rapport with human users; human-like behaviour may be expected too, when more anthropomorphic designs are manifested. The results also suggest that smart speakers, while embodied, do not facilitate the same turn-taking mechanisms as social robots do, likely due to the lack of non-verbal behaviours.

The results support **[H4]** reflecting that AMSR would be preferred for the task. We saw a wide difference between AMSR and SS, however AMR was also rated higher than SS, which may align with the fact that there is a relation to anthropomorphic agents with human-like designs, in terms of natural means of communication.

Most participants were more familiar with smart speakers than with social robots, which may indicate a novelty effect in the agent preference. Social robots are at time of writing emerging platforms and not as common and commercially available as smart speakers. However, we found that 2/3 of the participants were not able to identify the difference between AMR and AMSR, while they still preferred AMSR for the task. This indicates that the non-verbal behaviours used were subtle and asserted familiarity with the device.

Finally, we reject hypothesis **[H3]** reflecting that no differences would be found in task performance. Our assumption is that anthropomorphic facial features, without non-verbal behaviours is not enough to create more socially contingent interactions than SS: it is a combination of the two features that creates social facilitation to users.

Conclusion

In this paper, we presented a trade-off between task performance and social behaviour with conversational agents. Our contribution lies on an empirical evaluation of the anthropomorphic and non-verbal behaviour parameters of agents in task-oriented dialogues. This is particularly important to applications in which agents engage in a variety of tasks, and depending on the nature of the task, may need more or less social facilitation versus the value of task performance.

Not every agent needs to be anthropomorphised or to communicate with nonverbal behaviour; teasing out these variables and how they affect interaction time and social behaviour is the focus of this paper. To fully address the aspect of a potential novelty effect, longitudinal studies need to be designed where users' experiences are tested in long-term interactions with social robots and smart speakers. Potentially, increased familiarity with AMSR could decrease gaze time to levels similar to a more familiar social agent (i.e. another human).

To understand which of the independent variables contributed to the general preference of the robot, we concluded that while an anthropomorphic physical embodiment affects social behaviour, a set of non-verbal behaviours also increase the interaction time with the agent. Further research should be conducted in a variety of HRI scenarios, to investigate variability in the nature of the task and its relation to social facilitation between human users and agents. In sum, despite the task performance shortcomings of social and situation aware robots, they do hold a good interaction paradigm for enabling social facilitation with users.

Acknowledgements

We would like to acknowledge the support from the Swedish Foundation for Strategic Research project FACT (GMT14-0082). We would also like to thank the anonymous reviewers for their valuable comments earlier versions of this paper.

References

Adalgeirsson, S. O., & Breazeal, C. (2010). Mebot: a robotic platform for socially embodied presence. In *International conference on human-robot interaction.*

Admoni, H., & Scassellati, B. (2017). Social eye gaze in human-robot interaction: a review. *Journal of Human-Robot Interaction*.

Alam, M. R., Reaz, M. B. I., & Ali, M. A. M. (2012). A review of smart homes - past, present, and future. *IEEE Transactions on Systems, Man, and Cybernetics*.

Bainbridge, W. A., Hart, J., Kim, E. S., & Scassellati, B. (2008). The effect of presence on human-robot interaction. In *Ro-man.*

Breazeal, C., Dautenhahn, K., & Kanda, T. (2016). Social robotics. In *Springer handbook of robotics.*

Breazeal, C., & Fitzpatrick, P. (2000). That certain look: Social amplification of animate vision. In *Aaai.*

Dourish, P. (2004). *Where the action is: the foundations of embodied interaction*.

Fong, T., Nourbakhsh, I., & Dautenhahn, K. (2003). A survey of socially interactive robots. *Robotics and autonomous systems*.

Goble, H., & Edwards, C. (2018). A robot that communicates with vocal fillers has... uhhh... greater social presence. *Communication Research Reports*.

Gomez, R., Szapiro, D., Galindo, K., & Nakamura, K. (2018). Haru: Hardware design of an experimental tabletop robot assistant. In *International conference on human-robot interaction.*

Jung, Y., & Lee, K. M. (2004). Effects of physical embodiment on social presence of social robots. *Proceedings of PRESENCE*.

Kendon, A. (1967). Some functions of gaze-direction in social interaction. *Acta psychologica*.

Kennedy, J., Baxter, P., & Belpaeme, T. (2015). Comparing robot embodiments in a guided discovery learning interaction with children. *International Journal of Social Robotics*.

Kidd, C. D., & Breazeal, C. (2004). Effect of a robot on user perceptions. In *Iros.*

Kidd, C. D., & Breazeal, C. (2008). Robots at home: Understanding long-term human-robot interaction. In *Iros.*

Koda, T., & Ishioh, T. (2018). Analysis of the effect of agent's embodiment and gaze amount on personality perception. In *4th international workshop on multimodal analyses enabling artificial agents in human-machine interaction.*

Kontogiorgos, D., Avramova, V., Alexandersson, S., Jonell, P., Oertel, C., Beskow, J., ... Gustafsson, J. (2018). A multimodal corpus for mutual gaze and joint attention in multiparty situated interaction. In *Lrec.*

Kontogiorgos, D., Pereira, A., Andersson, O., Koivisto, M., Gonzalez Rabal, E., Vartiainen, V., & Gustafson, J. (2019). The effects of anthropomorphism and non-verbal social behaviour in virtual assistants. In *International conference on intelligent virtual agents.*

Kontogiorgos, D., Pereira, A., & Gustafson, J. (2019). The trade-off between interaction time and social facilitation with collaborative social robots. In *The challenges of working on social robots that collaborate with people, chi 2019.*

Lee, K. M., Jung, Y., Kim, J., & Kim, S. R. (2006). Are physically embodied social agents better than disembodied social agents?: The effects of physical embodiment, tactile interaction, and people's loneliness in human–robot interaction. *International Journal of Human-Computer Studies*.

Luria, M., Hoffman, G., & Zuckerman, O. (2017). Comparing social robot, screen and voice interfaces for smart-home control. In *Chi conference on human factors in computing systems.*

Marakas, G. M., Johnson, R. D., & Palmer, J. W. (2000). A theoretical model of differential social attributions toward computing technology: when the metaphor becomes the model. *International Journal of Human-Computer Studies*.

Mizoguchi, H., Sato, T., Takagi, K., Nakao, M., & Hatamura, Y. (1997). Realization of expressive mobile robot. In *Robotics and automation.*

Moon, Y., & Nass, C. (1996). How "real" are computer personalities? psychological responses to personality types in human-computer interaction. *Communication research.*

Nass, C., & Steuer, J. (1993). Voices, boxes, and sources of messages: Computers and social actors. *Human Communication Research*.

Pereira, A., Prada, R., & Paiva, A. (2014). Improving social presence in human-agent interaction. In *Sigchi conference on human factors in computing systems.*

Powers, A., Kiesler, S., Fussell, S., Fussell, S., & Torrey, C. (2007). Comparing a computer agent with a humanoid robot. In *International conference on human-robot interaction.*

Shibata, T., Tashima, T., & Tanie, K. (1999). Emergence of emotional behavior through physical interaction between human and robot. In *Robotics and automation.*

Skantze, G., Hjalmarsson, A., & Oertel, C. (2014). Turn-taking, feedback and joint attention in situated human–robot interaction. *Speech Communication*.

Straub, I. (2016). 'it looks like a human!' the interrelation of social presence, interaction and agency ascription: a case study about the effects of an android robot on social agency ascription. *AI & society*.

Thellman, S., Silvervarg, A., Gulz, A., & Ziemke, T. (2016). Physical vs. virtual agent embodiment and effects on social interaction. In *International conference on intelligent virtual agents.*

Torta, E., Oberzaucher, J., Werner, F., Cuijpers, R. H., & Juola, J. F. (2013). Attitudes towards socially assistive robots in intelligent homes: results from laboratory studies and field trials. *Journal of Human-Robot Interaction*.

Wainer, J., Feil-Seifer, D. J., Shell, D. A., & Mataric, M. J. (2006). The role of physical embodiment in human-robot interaction. In *Roman 2006.*

Illusory Body Perception and Experience in Furries

Alexander Kranjec[1,2] (kranjeca@duq.edu)

Louis Lamanna[1] (lamannal1@duq.edu)

Erick Guzman[1] (guzmane@duq.edu)

Courtney Plante[3] (plantec5@macewan.ca)

Stephen Reysen[4] (stephen.reysen@tamuc.edu)

Kathy Gerbasi[5] (kgerbasi@niagaracc.suny.edu)

Sharon Roberts[6] (sharon.roberts@uwaterloo.ca)

Elizabeth Fein[1] (feine@duq.edu)

[1] Psychology Department, Duquesne University, 600 Forbes Avenue, Pittsburgh, PA, 15282, USA
[2] Center for the Neural Basis of Cognition, 4400 Fifth Avenue, Suite 115 Pittsburgh, PA, 15213, USA
[3] Psychology Department, MacEwan University, Room 6-329, City Centre Campus, 10700 – 104 Avenue, Edmonton, AB T5J 4S2, CA
[4] Psychology and Special Education, Texas A&M University-Commerce, P.O. Box 3011, Commerce, TX, 75429, USA
[5] Psychology Department, Niagara County Community College, 3111 Saunders Settlement Rd., Sanborn, NY 14132, USA
[6] Social Development Studies, Renison University College at the University of Waterloo, 240 Westmount Road North, Waterloo, ON, N2L 3G4, CA

Abstract

The Rubber Hand Illusion (RHI) is an illusion of body ownership. This study investigates the RHI in *furries*: people who manifest interest in anthropomorphic animals through various combinations of costuming, roleplay, identification with a *fursona*, and unusual bodily experiences. Furry culture suggests two ways furries could differ from non-furries in their RHI experience: (1) furries' malleable perception of bodily self and identity may result in stronger feelings of illusory experience; alternatively, (2) furries' identification with non-human animals may result in weaker feelings of self-ownership for a human prosthetic. Results support the latter hypothesis; furries felt less subjective embodiment compared to non-furries. Moreover, proprioceptive drift was predicted by the extent individual furries valued humanity and their human bodies. The less esteem furries had for humanity and their human form, the less drift toward the human rubber hand was observed. These findings suggest how embodiment is related to subjectivity, identity, and practice.

Keywords: Rubber Hand Illusion; Embodiment; Body Perception; Culture; Identity

Introduction

Embodiment has been defined as the subjective awareness of, and self-coincidence with, one's own body (Longo et al., 2008). Research suggests that this pre-reflexive, bodily self-consciousness is constituted and undergirded by complex processes of bottom-up and top-down modulation of multisensory integration (Tsakiris, 2010). Previous studies have investigated the influences of these processes by using the Rubber Hand Illusion (RHI), a bodily illusion in which participants experience a sense of ownership for a prosthetic human hand (Botvinick & Cohen, 1998). To perform the RHI, a prosthetic hand is placed inside the participant's peripersonal space in a position congruent with their real hand. Participants are then instructed to look at the rubber hand while it and the real hand are stroked synchronously with a paintbrush. When these incongruous visual and tactile stimuli are integrated, participants report experiencing the rubber hand as their own. Additionally, when the illusion of ownership of the rubber hand is successfully induced, participants exhibit proprioceptive drift, a tendency to perceive the location of their real hand as closer to the rubber hand than it actually is. Longo et al.'s (2008) principal component analysis of RHI questionnaire data found evidence for three dissociable subcomponents influencing the experience of embodiment of the rubber hand: "ownership"; "location"; and "agency". The two subcomponents of "ownership" and "location" were significantly correlated with increased levels of proprioceptive drift in the RHI, suggesting that both top-down "body-representation" and

bottom-up "body schema" influences converged to structure the experience of embodiment of the rubber hand.

Further experimental research on the RHI has supported the dissociation of these influences by either highlighting significant group differences in experience of the RHI or manipulating the RHI procedure itself. Findings supporting the influence of "body-representation" on RHI experience have found that incongruent positioning, shape, texture (Haans et. al, 2008; Tsakiris & Haggard, 2005) and skin color (Lira et al., 2017) of the rubber hand attenuate RHI experience. Findings supporting the influence of "body schema" on RHI experience have found that asynchronous stimulation of the rubber hand attenuates the strength of the illusion significantly more than incongruence with body-representation (Armel & Ramachandran, 2003) and that populations with increased body-schema plasticity or flexibility such as individuals who are diagnosed with anorexia nervosa (Keizer et al., 2014), susceptible to out-of-body experiences (Braithewaite et al., 2017), hemiparetic (Llorens et al., 2017), psychosis-prone (Germine et al., 2012), or under the influence of dexamphetamine (Albrecht et al., 2011) demonstrate higher susceptibility to the RHI.

If previous research suggests (1) decreased illusory effects when individuals identify less with the form of the rubber hand and (2) increased illusory effects for individuals with greater body-schema flexibility, could the RHI be used to test hypotheses that interrogate the nature of body perception and experience in a unique population? *Furries* are self-identified fans of media featuring non-human animal characters who have been imbued with human-like traits (e.g., speech and bipedal walking; Gerbasi et al., 2008). As is typical with other media-based fandoms (e.g., science fiction; Jenkins, 1992), furries are both avid consumers and creators of fan-made artwork, animation, and writing (Plante, Roberts, Reysen, & Gerbasi, 2016a). They often share this interest with other fans, congregating primarily online, but also in-person at local meet-ups and at large-scale fan conventions (Mock, Plante, Reysen & Gerbasi, 2013). Illustrating the scope of these meetups, conventions such as Anthrocon, one of the world's largest furry conventions, regularly attract more than 5,000 furries.

A subset of the furry fandom (approximately 20%) also expresses their interest through *fursuiting*, the wearing of elaborate, mascot-style foam-and-fabric costumes of furry-themed characters (Plante, et al., 2016b). Fursuiting is somewhat analogous to the practice of cosplaying among anime fans, who invest considerable time and effort into dressing up and interacting with other fans as their favorite character from a show (Reysen, et al., 2018). Unlike cosplay, however, fursuiting tends to involve characters of furries' own creation.

One of the most universal activities in the furry fandom is the creation of a *fursona* – a non-human animal avatar imbued with human traits. Fursonas are used by furries as a representation of themselves within fandom spaces. Virtually all furries have a fursona, usually consisting of one or more non-human species, a name, and physical and personality traits (Plante et al., 2016b). Furries spend a great deal of time creating, thinking about, and interacting with others in the fandom through their fursonas, with which they strongly identify (Plante et al., 2016b). This suggests the possibility that many furries may have a relatively malleable perception of self and body. For example, a furry may spend an hour or two per day interacting with other furries as their fursona, whose species differs from their own (i.e., not human), whose personality may differ from their own (e.g., more gregarious), and whose appearance, gender, and age may differ from their own. Given that prior research has shown that furries have fairly active imaginations and spend a great deal of time engaging in fantasy-themed activities (e.g., role-playing games and online roleplaying; Plante et al., 2016b), furries' perception of bodily self and identity may be influenced by spending time engaged in furry-themed activities.

Speaking to this possibility, research suggests that some furries are likely to think of themselves as less than fully human and identify, at least in part, with non-human animals (Roberts et al., 2015). Furry conventions are also often attended by *therians* and *Otherkin,* those who have human bodies but experience themselves as something other than human (Gerbasi, Fein, Reysen, Plante, & Roberts, 2017). In contrast to non-therian furries, who may identify *with* a non-human species but usually understand themselves to be fundamentally human, therians identify *as* a non-human animal that exists or has existed on earth, such as a bear or a mammoth, while Otherkin identify *as* a creature usually considered to be mythological or fantasy-based, such as a fairy or unicorn. (*Note*: Although there are many therians and Otherkin who do not identify as furries, all therians and Otherkin in the current study also identified as furries.) Therians and Otherkin often report experiencing unusual bodily experiences, such as feeling phantom limbs belonging to the creature they identify as (such as claws, tails, or wings), and/or "shifts" into a mental state that they associate with their identified species. Many therians and Otherkin report experiences of deep discomfort with their human bodies and/or a desire to be in the body of the species with which they identify (Grivell, Clegg, & Roxburgh, 2014).

Presently, the RHI allows for testing between two competing hypotheses. If furries identify less with the human form of the rubber hand as compared to a control population then they should exhibit decreased RHI experience as a group. However, if furries have relatively greater body-schema flexibility, they should exhibit increased RHI experience. Results can inform our general understanding for how embodiment relates to identity, subjectivity, and practice.

Methods

Participants

All participants were recruited and tested at Anthrocon 2018 in a quiet, private room. Of the 57 participants tested, two early participants' data were not analyzed because they were recorded as having worn a ring or band-aid during the

procedure; one participant dropped out before completion; four other participants did not self-identify as furries in the subsequent survey. This left 50 furries for analyses (M_{age}=26.77; 11 female/36 male/3 NA; $M_{education\ years}$=15.55). For a comparison group of non-furries, we used raw data from 131 participants previously published in Longo et al. (2008).

Procedure

Rubber Hand Illusion We used the procedure described in Longo et al. (2008) as a model to carry out the RHI in the current study but using only a right rubber hand (there was no effect of handedness in the original study) and an occluder box described by the JoVE Science Education Database (2019). (In this version, the participant can view the experimenter.) Participants sat across from the experimenter with their hand hidden inside the occluder and the rubber hand placed congruently in view. There were two blocks. At the beginning of each block, participants estimated the location of the tip of their occluded right index finger by reporting the corresponding number on a ruler with a variable random offset (to prevent participants from using a remembered numeric label rather than a perceived location on subsequent trials). Following the pre-test location judgment, a 60s induction phase consisted of the visible rubber hand and occluded real hand being stroked with 2 identical paint brushes. In the *synchronous* block, individual fingers on each hand were brushed simultaneously; in the *asynchronous* block they were brushed 180° out of phase. (The asynchronous condition is frequently conceptualized as a kind of control or placebo, although subjective *deafference* scores have been observed to be higher in this condition.) Block order was randomized. After the induction phase in each block, participants were again asked to estimate the location of their index finger. Upon completion of the post-induction location judgement, participants filled out a

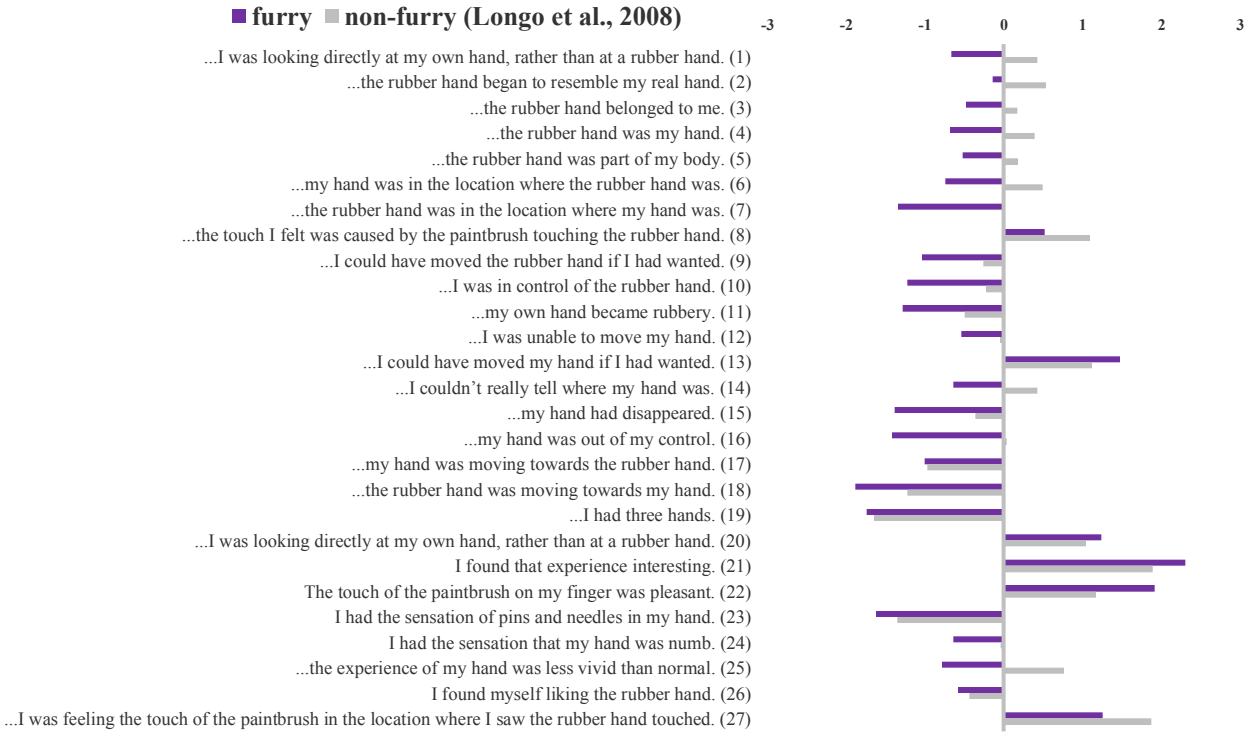


Figure 1. RHI embodiment survey items and response means. 7-point agreement scale; -3=strongly disagree; 0=neither agree or disagree; 3=strongly agree. (From Longo et al., 2008)

questionnaire assessing their subjective experience of the illusion (see Figure 1). This questionnaire, developed by Longo et al. (2008), measures 5 principal components (*embodiment of rubber hand, loss of own hand, movement, affect and deafference*) and three subcomponents of embodiment (*ownership, location, and agency*).

Experiential Survey After completion of both Rubber Hand Illusion blocks, participants filled out a survey with items designed to measure a number of variables related to their identity, experience, and attitudes, including questions about sexual identity, time since identifying as a furry and/or therian, and beliefs about being other than 100% human. The survey also asked systematic questions about the Duration ("How long ago did you start …"), Frequency ("How often do you have…"), and Intensity ("How intense are…") of relevant experiences and practices including: *Fursuiting*, *Role-Playing*, and *Online Interaction* with other furries. Two attitude scales were included as well (below).

Humanity-esteem version of the Rosenberg Self-Esteem Scale (Luke & Maio, 2009). This scale measures the extent to which participants think humanity is bad or good ("Human Value"). Example questions include, "I feel that the human species is very valuable, at least on an equal plane with other species in the universe;" "I feel that human beings have a number of very good qualities;" "All in all, I am inclined to regard the human species as a failure."

Identity version of the Transgender Congruence Scale. We modified a previously validated gender congruence scale (Kozee, Tylka, & Bauerband, 2012) to measure the extent to which furry participants feel comfortable with the match between their identity and human body ("Human Body Image"). Example questions include, "My outward appearance represents my identity;" "I experience a sense of unity between my identity and my body;" "My physical appearance adequately expresses my identity."

Results

Subjective Results Between Groups (Figure 2A.) Compared to the non-furry population reported in Longo et al. (2008), furries appear to experience several critical principal components of the RHI to a lesser extent when asked identical questions [MANOVA: $F(5, 175) = 9.72$, $p < .0005$; Wilk's $\Lambda = 0.783$].

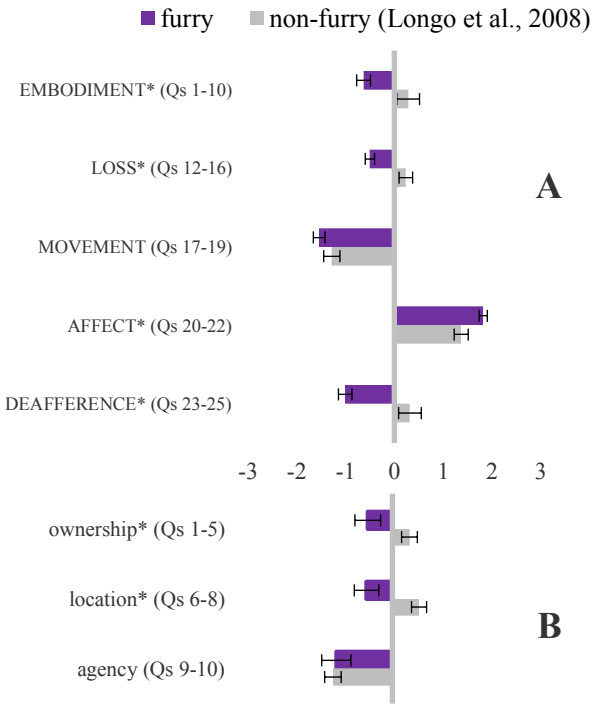


Figure 2. **(A)** Mean scores for Rubber Hand Survey principal components and **(B)** Mean scores for Embodiment item subcomponents. (*) Indicates significant differences between furry and non-furry groups. All survey scores shown were recorded after synchronous condition except deafference scores which indicate responses recorded after asynchronous condition.

Post-test ANOVAs indicate that furries reported significantly weaker experiences associated with the principal components of *embodiment* [$F(1, 179) = 11.22$, $p < .001$] and *loss* [$F(1, 179) = 16.94$, $p < .0005$], (during synchronous condition) and *deafference* [$F(1, 179) = 24.16$, $p < .0005$] (during asynchronous condition; See Longo et al., 2008 for more detailed explanation). Furries exhibited higher scores for *affect* [$F(1, 179) = 7.69$, $p < .01$] (regardless of condition; i.e., in synchronous and asynchronous blocks) suggesting furry participants enjoyed the experience of being brushed irrespective of any illusory effects. There was no significant difference for *movement* (as non-furries also reported negative scores). (Figure 2B.) Furries indicated weaker subjective feelings for the critical embodiment subcomponents [$F(3, 177) = 6.45$, $p < .0005$; Wilk's $\Lambda = 0.901$.] of *ownership* [$F(1, 179) = 7.71$, $p < .0005$] and *location* [$F(1, 179) = 13.38$, $p < .01$], but no difference for *agency*.

Proprioceptive Results Between Groups Proprioceptive drift is the tendency for participants to perceive the location of their real hand as closer to the rubber hand than it actually is. It is calculated by subtracting post-induction index finger location judgments from pre-induction location judgments in the synchronous block. Despite numerically smaller numerical averages for furries vs. non-furries, proprioceptive drift did not differ significantly for either condition.

Table 1. Proprioceptive Drift (cm) Between Groups and Conditions

Condition		Synchronous		Asynchronous	
	N	*M*	*SD*	*M*	*SD*
Furry	**50**	**0.76**	**3.73**	**0.05**	**3.57**
Non-furry	120	1.34	3.22	0.30	2.69

Between Group Results Summary Negative average values on relevant components, significantly lower than a large control sample, indicated lower subjective embodiment, loss, and deafference scores for furries. This suggests that identifying or role-playing as somewhat less, or other than human may be mitigating the strength of the RHI. That is, furries may experience the illusion to a lesser extent because they identify less with the human rubber hand.

These results simultaneously appear to argue against the alternative hypothesis; that furry participants who actively move between human and non-human roles in terms of distinct individual identities and practices, might have a more plastic body schema as compared to non-furries. This hypothesis predicts larger RHI effects in furries when compared to a typical population. This was not the result.

Analyses focused on variability within the furry sample could sharpen our explanation. If furries are experiencing the illusion to a lesser extent because they identify less with the human rubber hand, then we should expect that the strength of the illusion for furries could be predicted by the extent to which individual participants value humanity and feel comfort in their human bodies

		Prop Drift	Human Value	Body Image	Furry Time	Human %	Wear Freq	Wear Intens	Wear Time	Role Freq	Role Intens	Role Time	Online Freq	Online Intens	Online Time
Prop Drift	**R**		**.315***	**.434****	**-0.092**	**0.192**	**-0.047**	**-0.087**	**-0.13**	**-.389***	**-.441***	**-0.267**	**-0.009**	**-0.002**	**-0.068**
	p		*0.026*	*0.002*	*0.529*	*0.201*	*0.781*	*0.614*	*0.366*	*0.028*	*0.011*	*0.084*	*0.957*	*0.992*	*0.65*
	N		50	49	49	46	37	36	50	32	32	43	43	43	47
Human Value	**R**			**.649****	**0.188**	**0.157**	**-0.228**	**0.086**	**-0.003**	**-0.099**	**-0.236**	**-0.124**	**0.159**	**-0.078**	**-0.057**
	p			*<0.001*	*0.195*	*0.297*	*0.175*	*0.616*	*0.986*	*0.591*	*0.193*	*0.429*	*0.308*	*0.621*	*0.703*
	N			49	49	46	37	36	50	32	32	43	43	43	47
Body Image	**R**				**-0.038**	**0.28**	**-0.319**	**0.09**	**-0.008**	**-0.116**	**-.364***	**-0.187**	**0.005**	**-0.075**	**-0.259**
	p				*0.8*	*0.062*	*0.058*	*0.6*	*0.956*	*0.528*	*0.041*	*0.229*	*0.975*	*0.636*	*0.082*
	N				48	45	36	36	49	32	32	43	42	42	46

Table 2. Correlations (**R**) 2-tailed significance values (*p*) and sample sizes (N) for relations between **Proprioceptive Effects** (Prop Drift), **Human Value and Body Image Questionnaires**, and **Furry Experience Data**. Furry Experience Data includes (1) Frequency in terms of hours per day (Freq) (2) Intensity of Experience (Intens) and (3) Time in Months since beginning a particular kind of practice, including (A) Fursuiting (Wear), (B) Role-Playing (Role) and (C) Online Interaction with other furries (Online). Also shown are correlations for time in months since first identifying as a furry (Furry Time) and the relative extent in percentage terms that participants identify as Human/Non-human (Human %). * Indicates correlation is significant at the 0.05 level; ** correlation is significant at the 0.01 level.

Proprioceptive Results Within Group In furry participants, the extent of drift toward the rubber hand was positively correlated with individual scores on the **Humanity-esteem version of the Rosenberg Self-Esteem scale** ("Human Value") and the **Identity version of the Transgender Congruence Scale** ("Body Image"), which were highly correlated with each other. This suggests that among furries, lower esteem for humanity and feelings of incongruence between one's identity and human body is predictive of less proprioceptive drift towards the rubber hand (see Table 2 and Figure 3).

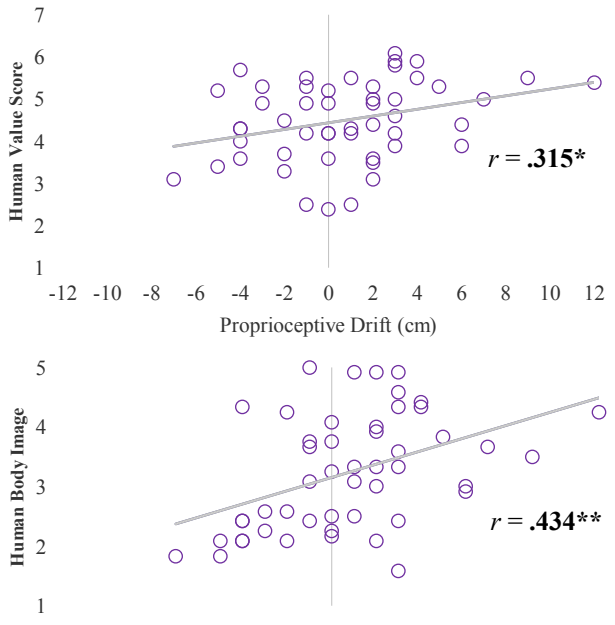


Figure 3. Correlations between individual proprioceptive drift scores (cm) and Human Value (7pt. scale, top) and Human Body Image (5pt. scale, bottom) scores. See Table 2.

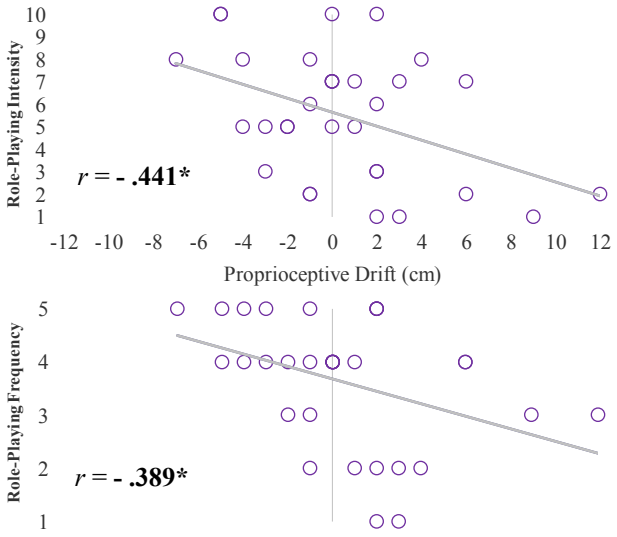


Figure 4. Correlations between individual proprioceptive drift scores (cm) and Role-Playing Intensity (10pt. scale, top) and Frequency (5pt. scale, bottom). See Table 2.

There were significant negative correlations between individual proprioceptive drift scores and Role-Playing Intensity and Frequency Scores from the **Experiential Survey**. This suggests that among furries, more frequent and intense role-playing in a fursona is predictive of less proprioceptive drift towards the rubber hand (see Table 2 and Figure 4).

Therian or Otherkin vs. Non-Therian Furries (Table 3.) There were significant differences (MANOVA and post-test ANOVAs) between non-therian and therian/Otherkin participants and survey scores (Human Value and Body Image) predictive of Proprioceptive drift (which showed a marginal difference between these groups). Therians had lower Human Value and Body Image scores and predictably self-identified as less human than non-therian furries.

Table 3. Non-Therian Furries vs. **Therian Furries**.

	Human %**	Prop Drift	Human Value*	Body Image*
Non-therian (*N*=38)	98.19	1.32	4.71	3.43
Therian (*N*=12)	**61.50**	**-1.00**	**3.83**	**2.64**
p	*<.0001*	*=.06*	*<.01*	*<.05*

General Results Summary The overall pattern of results supports the hypothesis that furries are experiencing a mitigated RHI because they identify less with the human rubber hand. **(1)** Compared to a control sample, furries exhibit lower average and subjectively negative scores for relevant principal components of a validated RHI survey. **(2)** The extent of proprioceptive drift, which can be regarded as a more objective illusion index, is predicted by Human Value, Body Image, and Role-Playing scores in furries. **(3)** Therians and Otherkin, i.e., furries who identify *as* non-human, exhibit lower survey scores associated with human esteem and marginally lower scores for proprioceptive drift as compared to non-therian furries.

Discussion & Conclusion

The present study suggests ways that illusions of body ownership can be used to test distinct hypotheses - that make opposite predictions - within unique populations. Lira et al. (2017) found that individual differences in implicit racial bias modulated proprioceptive drift (and other measures of RHI magnitude). That is, higher racial bias in white participants mitigated drift toward a black rubber hand, suggesting that within-subject attitudinal differences can reduce proprioceptive effects. Elsewhere it has been suggested (Dempsey-Jones & Kirikos, 2014) that proprioceptive effects are relatively impervious to top-down modulation, suggesting that neurocognitive group differences in body-perception may be driving results in other groups (e.g., in autism) that show reduced RHI effects. While the results of this study suggest that furries are less likely to identify with a human hand, they cannot determine if, broadly speaking, top-down or bottom up processes better describe why this is the case.

Despite these limitations (based principally on using a comparison data set from a previous study and correlational methods) the present study seeks to broaden the range of salient identity categories to studies of cognitive difference, joining the growing body of literature that explores the implications of variability in particular populations. We investigated a subculture whose membership is defined through a powerful and often embodied experience of affinity with a particular symbolic form – in this case, anthropomorphic animals. Our findings suggest that the kind of cultural differences that may not be visible to the eye or reportable on a typical demographic questionnaire, but manifest instead in self-identification with a particular community, subjective experience of difference, and ongoing participation in patterned cultural practices (Roepstorff, Niewöner, & Beck, 2010), may be profoundly related to body perception. These findings argue for a broader conceptualization of cultural and identity difference than is often found in cross-cultural cognitive research – one deeply grounded in both subjectivity and practice.

Acknowledgements

The authors would like to thank the organizers of Anthrocon for generously providing participant testing space over several days of the 2018 conference, and Matthew Longo for sharing raw data from a previously published study (Longo, et al., 2008).

This research was supported by the Social Sciences and Humanities Research Council of Canada.

Social Sciences and Humanities Research Council of Canada | Conseil de recherches en sciences humaines du Canada | Canada

References

Albrecht, M. A., Martin-Iverson, M. T., Price, G., Lee, J., Iyyalol, R., & Waters, F. (2011). Dexamphetamine effects on separate constructs in the rubber hand illusion test. Psychopharmacology, 217(1), 39-50.

Armel, K. C., & Ramachandran, V. S. (2003). Projecting sensations to external objects: evidence from skin conductance response. Proceedings of the Royal Society of London B: Biological Sciences, 270(1523), 1499-1506.

Botvinick, M., & Cohen, J. (1998). Rubber hands 'feel' touch that eyes see. Nature, 391(6669), 756.

Braithwaite, J. J., Watson, D. G., & Dewe, H. (2017). Predisposition to out-of-body experience (OBE) is associated with aberrations in multisensory integration: Psychophysiological support from a "rubber hand illusion" study. Journal of Experimental Psychology: Human Perception and Performance, 43(6), 1125.

Dempsey-Jones, H., & Kritikos, A. (2014). Higher-order cognitive factors affect subjective but not proprioceptive aspects of self-representation in the rubber hand illusion. Consciousness and cognition, 26, 74-89.

Gerbasi, K. C., Fein, E., Plante, C. N., Reysen, S., & Roberts, S. E. (2017). Furries, therians and otherkin, oh my! What do all those words mean, anyway? In T. Howl (Ed.), *Furries among us 2: More essays on furries by furries* (pp. 162-176) Lansing, MI: Thurston Howl Publications.

Gerbasi, K. C., Paolone, N., Higner, J., Scaletta, L. L., Bernstein, P. L., Conway, S., & Privitera, A. (2008). Furries from A to Z (anthropomorphism to zoomorphism). *Society & Animals, Journal of Human-Animal Studies, 16*(3), 197-222.

Grivell, T., Clegg, H., & Roxbaugh, E.C. (2014). An interpretive phenomenological analysis of identity in the therian community. *Identity: An International Journal of Theory and Research. 14*(2), 113-135.

Germine, L., Benson, T. L., Cohen, F., & Hooker, C. I. L. (2013). Psychosis-proneness and the rubber hand illusion of body ownership. Psychiatry research, 207(1-2), 45-52.

Haans, A., IJsselsteijn, W. A., & de Kort, Y. A. (2008). The effect of similarities in skin texture and hand shape on perceived ownership of a fake limb. Body Image, 5(4), 389-394.

Jenkins, H. (1992). *Textual poachers: Television fans & participatory culture*. London: Routledge.

JoVE Science Education Database (2019). *Sensation and Perception.* The Rubber Hand Illusion. JoVE: Cambridge, MA.

Keizer, A., Smeets, M. A., Postma, A., van Elburg, A., & Dijkerman, H. C. (2014). Does the experience of ownership over a rubber hand change body size perception in anorexia nervosa patients? Neuropsychologia, 62, 26-37.

Kozee, H. B., Tylka, T. L., & Bauerband, L. A. (2012). Measuring transgender individuals' comfort with gender identity and appearance: Development and validation of the Transgender Congruence Scale. *Psychology of Women Quarterly*, *36*(2), 179-196.

Longo, M. R., Schüür, F., Kammers, M. P., Tsakiris, M., & Haggard, P. (2008). What is embodiment? A psychometric approach. *Cognition*, *107*(3), 978-998.

Luke, M. A., & Maio, G. R. (2009). Oh the humanity! Humanity-esteem and its social importance. *Journal of Research in Personality*, *43*(4), 586-601.

Lira, M., Egito, J. H., Dall'Agnol, P. A., Amodio, D. M., Gonçalves, Ó. F., & Boggio, P. S. (2017). The influence of skin colour on the experience of ownership in the rubber hand illusion. *Scientific Reports*, *7*(1), 15745.

Llorens, R., Borrego, A., Palomo, P., Cebolla, A., Noé, E., i Badia, S. B., & Baños, R. (2017). Body schema plasticity after stroke: subjective and neurophysiological correlates of the rubber hand illusion. Neuropsychologia, 96, 61-69.

Mock, S. E., Plante, C. N., Reysen, S., & Gerbasi, K. C. (2013). Deeper leisure involvement as a coping resource in a stigmatized leisure context. *Leisure/Loisir, 37*(2), 111-126.

Plante, C. N., Reysen, S., Groves, C. L., Roberts, S. E., & Gerbasi, K. (2017). The fantasy engagement scale: A flexible measure of positive and negative fantasy engagement. *Basic and Applied Social Psychology, 39*(3), 127-152.

Plante, C. N., Reysen, S., Roberts, S., & Gerbasi, K. (2017). 'Welcome to the jungle': Content creators and fan entitlement in the furry fandom. *The Journal of Fandom Studies, 5*(1), 63-80.

Plante, C. N., Roberts, S. E., Reysen, S., & Gerbasi, K. C. (2016a). *FurScience A summary of five years of research from the International Anthropomorphic Research Project.* Waterloo, ON: FurScience.

Plante, C. N., Roberts, S. E., Reysen, S., & Gerbasi, K. C. (2016b). "By the numbers": Comparing furries and related fandoms (pp. 106-126). In T. Howl (Ed.), *Furries among us: Essays on furries by the most prominent members of the fandom*. Nashville, TN: Thurston Howl Publications.

Reysen, S., Plante, C. N., Roberts, S., & Gerbasi, K. (2018). Motivations of cosplayers to participate in the anime fandom. *The Phoenix Papers, 4*(1), 29-40.

Roberts, S., Plante, C., Gerbasi, K., & Reysen, S. (2015b). The anthrozoomorphic identity: Furry fandom members' connections to non-human animals. *Anthrozoös, 28*(4) 533-548.

Roepstorff, A., Niewöhner, J., & Beck, S. (2010). Enculturing brains through patterned practices. Neural Networks, 23(8-9), 1051-1059.

Tsakiris, M. (2010). My body in the brain: a neurocognitive model of body-ownership. *Neuropsychologia*, *48*(3), 703-712.

Implicit Evaluations Reflect Causal Information

Benedek Kurdi
Harvard University, Cambridge, Massachusetts, United States

Adam Morris
Harvard University, Cambridge, Massachusetts, United States

Fiery Cushman
Harvard University, Cambridge, Massachusetts, United States

Abstract

Evaluations along a positivenegative dimension can be measured either explicitly (via self-report) or implicitly (via response interference tasks). Whether implicit evaluations encode relational information (e.g., A causes B) or only co-occurrence information (AB) has been debated extensively. 1,082 participants observed a machine being activated by causally responsible stimuli and dispensing rewards in the presence of merely associated, but not causal, stimuli. Evaluations of causally responsible vs. associated stimuli were measured implicitly and explicitly. Explicit and implicit evaluations of causally responsible stimuli were more positive than those of associated stimuli, both in the presence (Study 1) and absence (Study 2) of verbal instructions about the operation of the machine. Study 3 eliminated temporal primacy and overshadowing as explanations of the effect. Supporting propositional theories, these findings suggest that implicit evaluations are sensitive not only to co-occurrence but also to relational information, whether conveyed verbally or learned solely from experience.

Unexpectedness makes a sociolinguistic variant easier to learn: An alien-language-learning experiment

Wei Lai (weilai@sas.upenn.edu)
Department of Linguistics, University of Pennsylvania
Péter Rácz (RaczP@ceu.edu)
Cognitive Development Center, Central European University
Gareth Roberts (gareth.roberts@ling.upenn.edu)
Department of Linguistics, University of Pennsylvania

Abstract

We report two artificial-language-learning experiments investigating if the acquisition of sociolinguistic associations is facilitated by two kinds of expectation violation: encountering a variant (a) for the first time or (b) in an ungrammatical context. Participants learned an artificial language with two dialects, each spoken by one of two alien species: *Gulus* and *Norls*. The two dialects differed with regard to a plural suffix: Gulus mostly used *-dup*, and Norls mostly used *-nup*. In the first learning phase, participants learned the language without aliens; in the second learning phase, they were exposed to it with alien interlocutors. In Experiment 1 we manipulated whether *-nup* occurred in the first learning phase; in Experiment 2 we manipulated linguistic constraints on its occurrence. The acquisition of sociolinguistic association was evaluated by asking participants to select suffixes given aliens and vice versa. We found that sociolinguistic acquisition was facilitated in Experiment 1, but not Experiment 2. In Experiment 2, however, a post hoc analysis revealed that participants who had learned the grammatical context of the linguistic conditioning did experience facilitation, while those who had not did not. Our results provide laboratory evidence that unexpectedness facilitates the learning of sociolinguistic variation.

Keywords: artificial-language learning; social meaning; sociolinguistics; salience; surprisal

Introduction

The role of *salience* in the acquisition and propagation of linguistic variants has long been documented in classic sociolinguistic research (Labov, 1972). Variants with higher salience are encoded with more attention and higher meta-linguistic awareness, leading them to be more easily recognized and retained than other variants with equal frequency, resulting in an acquisition bias that cannot be explained by frequency of exposure alone (Jaeger & Weatherholtz, 2016).

In this study we investigated the role of salience in facilitating the learning of sociolinguistic meaning (i.e., the association of particular linguistic variants with particular social groups). We focused in particular on the effect of previous experience on salience. Previous work has paid much attention to the role of certain kinds of *non-linguistic* experience such as social and developmental experience (Foulkes & Docherty, 2006) or social stereotypes (Levy, 2008), but *linguistic* experience is relatively understudied. In particular, there is very little work on how the perceived salience of a sociolinguistic variant is affected by prior experience of that variant in other contexts. Jaeger and Weatherholtz (2016, p. 1) proposed that salience related to language experience can be understood in terms of expectation violation, analogous to the well-attested novelty bias effect: Novel items and events that we do not expect tend to stand out. Jaeger and Weatherholtz (2016) argued that this might occur for linguistic variants that a listener has not encountered before and might thus lead to surprisal. The salience generated by surprisal may facilitate learning the variant and its socioindexical meaning.

Although Jaeger and Weatherholtz's (2016) approach to experience-based salience seems appealing for its operationalization of expectation-related salience in an information-theoretic framework (Shannon, 1948; Hale, 2001; Levy, 2008), it is not yet supported by linguistic data. Several experimental studies on language processing show that less expected words and structures take longer to process and at greater cost (McRae, Spivey-Knowlton, & Tanenhaus, 1998; McDonald & Shillcock, 2003), with similar effects observed in comprehension tasks (Kaschak & Glenberg, 2004; Squires, 2014; Fraundorf & Jaeger, 2016). However, additional processing for novel variants in comprehension does not necessarily result in better performance in noticing or memorizing these variants or in associating them with the right social group.

The present study

The present study investigates the hypothesis that experience-dependent salience can arise from expectation violation, and cause a sociolinguistic variant to be more learnable. We used an "alien language" learning paradigm in which participants first learned a miniature artificial language and were then exposed to it in a simple social context with "alien interlocutors". We investigated two kinds of expectation violation, hypothesizing that participants would be more likely to learn a sociolinguistic association if (Experiment 1) they had not encountered it before and (Experiment 2) they had encountered it before, but subject to grammatical constraints that now appear to be violated.

As an example of the first kind of violation, one might imagine an American English speaker visiting Liverpool and hearing, for the first time, *book* pronounced with a final velar fricative [x] (as in German *Buch*) instead of the expected velar stop [k]. As an example of the second kind of violation, consider a speaker who has heard *-th* pronounced as [f], but only at the end of syllables (as in [boʊf] for *both* or [ˈɛfnɪk] for *ethnic*). For a speaker who had acquired syllable-finality as a constraint on this variant, hearing [fɪnk] for *think* would likely be relatively salient.

In both our experiments a certain variant of the alien language was associated predominately (but not necessarily exclusively) with a particular species of alien. We evaluated whether participants had learned this sociolinguistic association pattern by asking them at the end of the experiment to (a) select the variant that a given alien would most likely produce, and (b) select the alien most likely to produce a given variant. We predicted that increased salience, via expectation violation, would lead participants to do better at both tasks (though, because we did not make the sociolinguistic relationship categorical, we did not expect that the response to the two tasks would be identical).

Finally, we predicted that listeners who learned a sociolinguistic association in the experiment could generalize that relationship to new words.

Experiment 1: First encounter

Experiment Overview

In Experiment 1, we investigated whether encountering a linguistic variant for the first time in a social context would facilitate the sociolinguistic learning of that variant (we will term this hypothesized effect "first-encounter facilitation"). Participants were trained on an alien language with two dialects, each used by a different alien species, the *Gulus* or the *Norls* (Fig. 1). The dialects differed with regard to a plural suffix: Gulus used *-dup* as the only form of the plural suffix whereas Norls sometimes used *-dup* but mostly used *-nup*.

The experimental procedure consisted of three phases: Participants were first trained on the language without seeing any aliens, which was intended to establish *prior experience* with the language; then (having been introduced to the two alien species) they were exposed further to the language with alien interlocutors, which allowed them to learn associations between plural suffixes and alien species. In the third and final phase, acquisition of sociolinguistic variants was evaluated on the basis of whether participants could infer which alien might have used a given suffix and, conversely, which suffix a given alien might have used.

Crucially, we manipulated participants' *prior experience* with the variant *-nup* such that half the participants would never be exposed to it in the first phase, encountering only *-dup* (*NoExposure* condition), whereas the other half would see both suffixes in every phase (*Exposure* condition). We predicted that participants with no experience of *-nup* in the first learning phase would find it more salient in the second phase and better learn to associate it with Norls.

Method

Participants 100 participants completed Experiment 1 online within the specified amount of time (1.5 hours). After excluding participants whose duration was below the 2.5% quantile or above the 97.5% quantile of all participants, we used the data of the remaining 93 participants. There were 51 female and 43 male participants, aged 17–73 (mean: 28.9) years. 30 of them were recruited from the University of Pennsylvania subject pool (in return for course credit) and 64 were recruited through the Prolific Academic website (and were paid $5 each). 49 participants were in the *Exposure* condition and 45 in the *NoExposure* condition.

Alien language The artificial language was composed of 14 word stems, as shown in Table 1, and a plural suffix with two variants, *-dup*, *-nup*.

Table 1: 14 Stem Words in the Alien Language

nesel, laniz, firot, hiwen, maqub, jemulok, gequzis
tugan, nuwik, falon, wumos, wukin, sehilod, takoles

The 14 stem words were randomly generated by combining one or two CV syllables with a word-final CVC syllable from a segment pool of five vowels /a e i o u/ and 12 consonants /k g q h m n t s z j l w/.

Aliens The language was used by two alien species: *Gulus* and *Norls*. The stem forms were the same across dialects, but the suffix variants had different distributions: Gulus attached *-dup* to all 14 words to signal plurality, whereas Norls attached *-dup* to only four of the words (*hiwen*, *wukin*, *jemulok* and *wumos*) and *-nup* to the remaining eight words. Put differently, Gulus used the *-dup* variant 100% of the time as their plural suffix whereas Norls used *-dup* and *-nup* at a ratio of 71% to 29%. Within each alien species, six idiosyncratic aliens were designed in order to ensure that the linguistic variation on the group level wouldn't be mistaken for variation on the individual level (See Fig. 1 for examples).

Figure 1: Alien Species: Gulus (left) and Norls (right)

Procedure The experimental task was composed of two learning phases, in which participants were trained on the alien words through passive exposure to word-object[1] pairs and multiple-choice exercises with feedback, and a test

[1]We thank Professor Janet Pierrehumbert for making images of the objects available for use. The artworks are copyrighted to Northwestern University and used with permission.

(a) (b)

Figure 2: Example trials in learning phase 1: (a) a passive exposure trial; (b) a forced-choice trial

(a) (b)

Figure 3: Example trials in learning phase 2: (a) a passive exposure trial; (b) a forced-choice trial

phase that evaluated how well participants had learned the association between plural variants and alien species.

First learning phase: Learning without aliens. The experiment started with a learning phase that exposed participants to the words of the language without any aliens. This was designed to give participants exposure to the language before introducing it in a social context. It consisted of a series of trials, with two kinds of trial, as shown in Fig. 2. In passive exposure trials (Fig. 2a) a word was paired with an image of the object(s) it referred to. Participants were instructed to memorize the word and its meaning before proceeding to the next trial. In forced-choice trials (Fig. 2b) participants had to choose the correct word to go with an image; there were always two options to choose from, one correct and one a foil generated by changing one or two segments in the stem of the correct form. Participants received one point for each correct response and no point for an incorrect one (maximum: 168 points). Feedback on the correct form and the point received for each question was provided immediately afterwards.

Participants were trained on 28 alien words (14 singular, 14 plural), which were divided into seven sets of four words each, with the constraint that each set contained two singular words and two plural words that all had different stems. For each set, a participant would see a passive exposure trial for each word in turn; then they would see a forced-choice trial for each of the same four words. Then they would proceed to the next set. The order of the seven word sets was randomized, as was the order of the four trials within each passive exposure section and within each forced-choice section. The whole process was repeated once participants had completed training on all seven word sets. In total, participants were exposed to 14 words $\times$ 2 forms (singular and plural) $\times$ 2 trial types (exposure, forced choice) $\times$ 2 repetitions = 112 trials.

Alien introduction. After the first learning phase, the aliens were introduced. Participants were first presented with images of Gulus and Norls, each labeled with the species name; then the labels were removed and participants were instructed to drag and drop each alien into one of the the two boxes labeled *Gulu* and *Norl*. Feedback was provided after the drag-and-drop.

Second learning phase: Learning with aliens. After the aliens had been introduced, the second learning phase started. This phase resembled the first phase in its structure, except that each trial (whether passive-exposure or forced-choice trial) included a picture of an *alien interlocutor*, as shown in Fig. 3. Participants saw both the Gulu and the Norl form of every word, so the second learning phase was twice as long as the first learning phase (with each set of trials containing eight words rather than four). In total each participant was exposed to 14 words $\times$ 2 forms (singular, plural) $\times$ 2 species (Gulu, Norl) $\times$ 2 trial types (exposure, forced choice) $\times$ 2 repetitions = 224 trials.

Test phase: Measuring acquisition. After the second learning phase, the test phase began, which evaluated the extent to which participants had established associations between alien groups and plural suffixes. The test phase contained two tasks: a suffix-identification task in which participants had to choose which form might be used based on the presented alien interlocutor, and an alien-identification task in which participants had to choose which alien was most likely to have said a prompt word. Trials in these tasks contained both *old word* stimuli from the learning phase and *new word* stimuli that participants had never seen, in order to evaluate the generalization of sociolinguistic associations to novel items. Trial order was randomized for each participant, and the order of the two options within each trial was counterbalanced. No feedback was provided.

In ***suffix identification***, trials on old words worked like forced-choice trials in the second learning phase (Fig. 3b), except that the optional answers had identical stems and different suffixes (i.e., the reverse of the situation in the learning phases). Participants were instructed to choose the form the pictured alien would likely use. Trials on new words were different: Participants were presented with a singular word, an image of the object it referred to, and an alien interlocutor; they were required to choose between a *dup*-ending word and a *nup*-ending word as the plural form (Fig. 4).

In all, the task included 56 trials on old words (14 words $\times$ 2 species $\times$ 2 repetitions = 56 trials), 24 trials on

Figure 4: Example suffix-identification trial with a new word

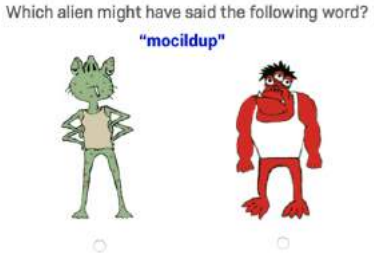


Figure 5: Example alien-identification trial

new words (6 new words × 2 species × 2 repetitions = 24 trials), and 34 filler trials, which tested participants on either singular words or plural words with incorrect stems.

In ***alien identification*** (Fig. 5) participants were given a plural word and had to choose between a Gulu and a Norl as the likely speaker of the word. The idiosyncratic aliens were kept consistent throughout the whole task, but whether they appeared on the left or the right was counterbalanced across questions. The stimulus words were generated by affixing the 14 old words and the six new words once each with *-dup* and once each with *-nup*, so that there were 40 trials (14 old words × 2 suffixes + 6 new words × 2 suffixes) in total.

Experimental conditions. Participants were randomly assigned to two experimental conditions: the *NoExposure* condition and the *Exposure* condition. Fig. 6 shows the distribution of variants in the two learning phases in different experimental conditions.

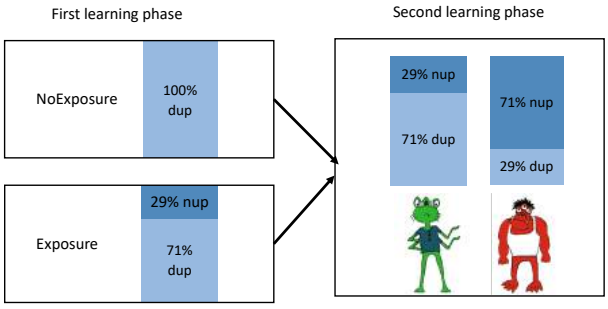


Figure 6: Variant distribution in the learning phases of Experiment 1

The two conditions differed with respect to the presence or absence of the variant *-nup* in the first learning phase: For participants in the *NoExposure* condition, plural words in this phase would always be affixed with *-dup*, whereas participants in the *Exposure* condition would see ten instances of plurals with *-dup* (71%) and four with *-nup* (29%). The two conditions were identical in the second learning phase: Gulus exclusively used *-dup* while Norls used *-nup* 71% of the time and *-dup* 29% of the time.

Results

Analyses were conducted using the R Statistical environment (R Core Team, 2014); linear models were run using the lme4 library (Bates, Mächler, Bolker, & Walker, 2015), and plots were created using ggplot (Wickham, 2016).

On average, it took participants (outliers excluded) 52 minutes (*sd* = 14) to complete the whole experiment. Out of a maximum of 168 points, participants achieved an average score of 153 (*sd* = 13).

Fig. 7 shows the aggregate results for suffix identification (left) and alien identification (right). The left panel shows how often participants selected the *-nup* suffix for a given alien the suffix-identification task. Consistent with our predictions, participants in the *NoExposure* group were more inclined to choose a *-nup* word for a Norl than those in the Exposure condition. Notably, the *-nup* response ratios given a Norl were relatively low in both conditions, nowhere matching the 71% in the input. The right panel shows what proportion of the time participants selected a Norl for a given suffix in the alien-identification task. Again, consistent with the hypothesis, participants in the *NoExposure* condition were more inclined to choose a Norl given a *-nup* word and to choose a Gulu given a *-dup* word, compared with those in the Exposure condition, who chose Norl interlocutors for both *-dup* and *-nup* at chance level.

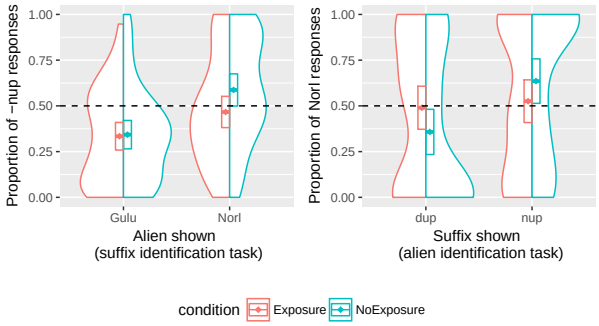


Figure 7: Proportion of *-nup* responses in suffix identification and Norl responses in alien identification (including 95% confidence interval). Dotted line indicates chance level.

Mixed-effects logistic regression models were fit on the two tasks, with Response as the dependent variable, Condition (Exposure as the intercept), Stimulus (Norl as the intercept in suffix identification; *-dup* as the intercept

in alien identification) and their interactions as independent variables, and Participant and Word as random factors. Both models revealed a significant Condition effect ($\beta = 0.57, p = 0.001$ in suffix identification; $\beta = -0.54, p < 0.001$ in alien identification) and its Interaction with Stimulus ($\beta = -0.48$ in suffix identification, $\beta = 1.00$ in alien identification, $p < 0.001$ in both cases). A stimulus effect was found only in suffix identification ($\beta = 0.62, p < 0.001$), not in alien identification ($\beta = 0.14, n.s.$).

Novel Stimuli We hypothesized that participants would apply the sociolinguistic association they had learned in the second training phase to novel words they had never seen before. The results show that identification with old and new words strongly mirrored each other in both conditions and both tasks. A mixed-effects model was fit on each of the two tasks, with Response (Suffix or Alien) as the dependent variable, Participant and Word as random factors, and Condition, Stimuli (either Alien or Suffix) and Novelty as fixed effects. The results showed no significant Novelty effect in suffix identification ($\beta = 0.23, n.s.$) and alien identification ($\beta = 0.10, n.s.$). These results indicate that the acquired sociolinguistic association could be generalized to new lexical items, and that first-encounter facilitation applies to both familiar and unfamiliar words.

Summary Our prediction concerning first-encounter facilitation was supported. That is, participants in the *NoExposure* condition were more likely to acquire the association between *-nup* and the Norl species than participants in the Exposure conditions, suggesting that the first encounter with a novel variant facilitated the acquisition of sociolinguistic variants of that variant. We also found that this effect extended to previously unseen words.

Experiment 2: Constraint violation

Experiment 2 used a similar paradigm to Experiment 1, but we modified the suffixation patterns to investigate a different source of surprisal. Instead of surprisal caused by encountering a variant for the first time, Experiment 2 investigated whether surprisal caused by encountering a linguistic variant in an apparently ungrammatical context (i.e., where it violated a grammatical constraint) would also facilitate the acquisition of sociolinguistic associations. We will term this constraint-violation facilitation.

Method

Participants 103 participants completed Experiment 2 online within 1.5 hours. After excluding participants whose duration was below the 2.5% quantile or above the 97.5% quantile of all participants, there were 97 participants left whose data were used for the final analysis. They were 69 females and 28 males, aged 17–78 (mean: 29.3) years. 28 of them were recruited from the University of Pennsylvania subject pool (and rewarded with course credit), and the remaining 69 were recruited through the Prolific Academic website (and paid $5 each). There were 48 participants in the Conditioned condition and 49 in the Unconditioned condition.

Materials and Procedure The same words and aliens were used in Experiment 2 as in Experiment 1. The procedure was also the same, consisting of two learning phases and a test phase with two tasks.

Experimental Conditions and Predictions There were two between-subjects conditions based on the linguistic environment for the suffix *-nup*, which is shown in Fig. 8.

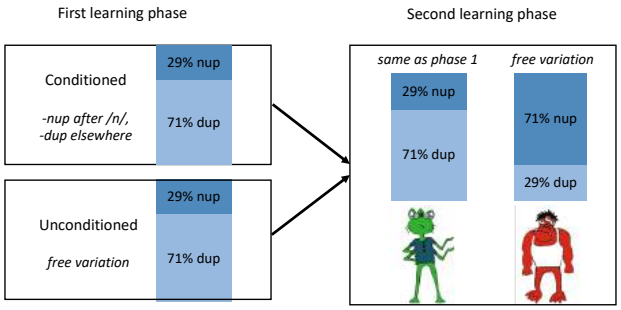


Figure 8: Variant distribution in the learning phases of Experiment 2

In the *Conditioned* condition participants in the first learning phase only ever saw *-nup* attached to the four nasal-ending stems (i.e., *falon*, *hiwen*, *tugan* and *wukin*), while *-dup* was attached to the 10 stems that did not end in a nasal. This implied a grammatical constraint on the distribution of *-nup* (i.e., that it only occurs after nasals). By contrast, participants in the *Unconditioned* condition were exposed to the two suffix variants in free variation (i.e., both *-nup* and *-dup* occurred with both nasal and non-nasal stems), though the variants still occurred at a ratio of ten (*-dup*) to four (*-nup*) – or 71% to 29% – just as in the *Conditioned* condition. In the second learning phase, Gulus exhibited precisely the suffixation pattern of the first phase, whereas Norls used the two suffixes freely across contexts at a ratio of four (*-dup*) to ten (*-nup*).

Similar to Experiment 1, we predicted that participants in the *Conditioned* condition would experience greater surprisal when they saw Norls using the two variants, especially *-nup*, in an ungrammatical way, and would be facilitated by this surprisal in learning the association between *-nup* and the Norl species, compared with those in the *Unconditioned* condition.

Results

On average, participants took 51 minutes (sd = 12) to complete the experiment and achieved a mean score of 152 ($sd = 13$).

Fig. 9 shows the aggregate results for suffix identification (left) and alien identification (right). The results do not appear to exhibit the predicted between-group difference in learning.

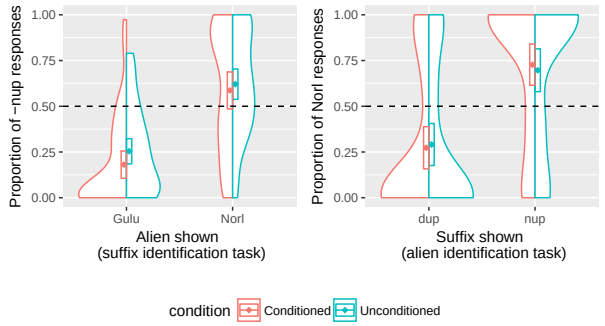


Figure 9: Proportion of *-nup* responses in suffix identification and Norl responses in alien identification (including 95% confidence interval). Dotted line indicates chance level.

A mixed-effects logistic model with Participant and Word as random factors, and Condition, Alien and their Interaction as independent variables revealed a significant Alien effect (Norl as the default, $\beta = -2.08, p < 0.001$) and a significant interaction ($\beta = 0.28, p = 0.012$), but no effect of Condition ($\beta = 0.28, n.s.$). In alien identification, a mixed-effects logistic model showed a significant effect of Suffix (*-nup* as the default, $\beta = 1.96, p < 0.001$), but no effect of Condition ($\beta = 0.06, n.s.$) and the Interaction ($\beta = -0.18, n.s.$).

Learning Proficiency It is possible that the absence of facilitation in Experiment 2 was due to variation in learning performance. The predicted facilitation depends on surprisal due to the apparent violation of a grammatical constraint. It therefore seems *a priori* clear that our predicted effect should occur only if participants learned the grammatical constraint. If they did not, violation of the constraint should not generate surprisal. To evaluate this possibility, we conducted a post hoc analysis in which we took participants' scores in the learning phase as a proxy for their learning performance. In particular, we divided participants into *good learners* and *poor learners* within each condition, according to whether their score was above or below the group mean. We then investigated whether constrain-violation facilitation could be found among good learners but not poor learners.

Fig. 10 shows the results for the 47 *good* and 50 *poor* learners. First, good learners showed a higher *-nup* rate for Norls and a lower *-nup* rate for Gulus in suffix identification, as well as a higher Norl rate for *-nup* and a lower one for *-dup*, compared with poor learners, indicating a better alignment between their responses and the pattern in the input, compared with poor learners. Second, the predicted learning facilitation is exhibited in the results of good learners, in that participants in the *Conditioned* condition exhibited a lower *-nup* rate for Gulus in suffix identification, and exhibited a higher Norl identification rate for *-nup* words and a lower Norl rate for *-dup* words in alien identification, compared with those in the *Unconditioned* condition.

We fit a mixed-effects logistic regression respectively on

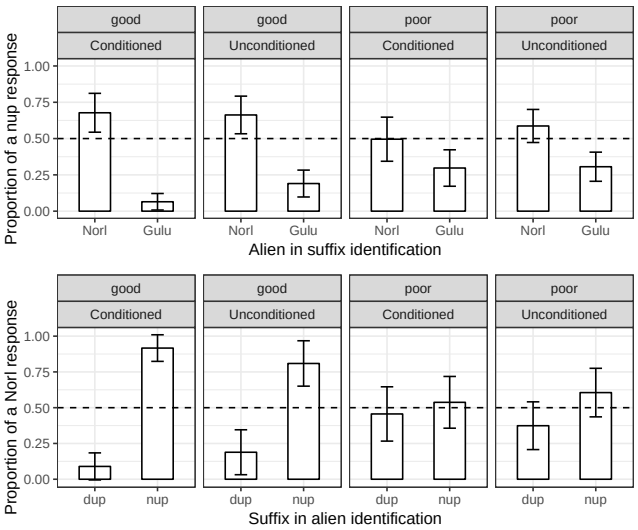


Figure 10: Proportion of *-nup* responses in suffix identification (top) and Norl responses in alien identification (bottom) by *good* and *poor* learners (including 95% confidence interval) Dotted line indicates chance level.

the data of good learners and poor learners. For good learners, the results showed significant effects of Alien ($\beta = 4.5$, $p < 0.001$), Condition ($\beta = 1.9$, $p < 0.001$) and their interaction ($\beta = -1.99$, $p < 0.001$) in suffix identification, as well as significant effects of Suffix ($\beta = 5.28$, $p < 0.001$), Condition ($\beta = 1.16$, $p < 0.001$) and Interaction ($\beta = -2.27$, $p < 0.001$) in alien identification. For poor learners, however, the results of suffix identification showed a main effect of Alien ($\beta = -0.93$, $p < 0.001$) and significant interaction between Alien and Condition ($\beta = -0.38$, $p = 0.008$), but no main effect of Condition ($\beta = 0.42$, $n.s.$). In alien identification, both factors of Suffix ($\beta = 0.33$, $p = 0.012$) and Condition ($\beta = -0.34$, $p = 0.007$) are significant, as is their interaction ($\beta = 0.62$, $p < 0.001$). Interestingly, however, the learning difference associated with the Suffix factor is the opposite of what was predicted: Learners in the *Unconditioned* condition did a better job in associating Norls to *-nup* than those in the *Conditioned* condition.

Novel Stimuli In evaluating whether acquisition effects were generalized to new words, we examined good and poor learners separately given their different patterns in acquisition. The results showed that although learners with different performance showed distinct patterns from each other, the behaviors with seen and unseen stimuli were highly consistent within each of the two learner groups. Good learners showed the correct alien-language association as well as additional facilitation from rule violation with both old and new words. Poor learners also showed consistent behaviors across old and new words, although behavior was mostly near chance level. Two mixed-effects models, one fit on each task, with Response as the dependent variable, Participant and Word as random factors, and Condition,

Stimuli, Novelty and their interactions as mixed effects, showed no significant Novelty effect or Novelty-relevant interactions.

Summary There was no evidence for constraint-violation facilitation in the aggregate results. However, post hoc analysis revealed that there was such an effect among "good learners", participants who performed above the mean in training. This is consistent with the hypothesis, as constraint violation should facilitate learning only among individuals who have learned the constraint. Finally, the results of Experiment 2 replicate those of Experiment 1 in showing an ability to generalize acquired patterns, whether accurate or inaccurate, to new words.

General Discussion

We hypothesized that violation of expectation would cause a linguistic variant to be more salient and, as a result of this, that an association between this variant and a particular social group would be easier to learn. We tested this hypothesis in two experiments, each investigating a different kind of expectation violation.

The first experiment investigated exposure to a previously unencountered variant while the second investigated exposure to a variant that had previously occurred within a narrower grammatical context. In the first experiment the expectation violation had the predicted effect: Participants were more likely to associate the new suffix with the correct alien species (and the correct alien species with the new suffix) when the suffix had not been encountered in the initial learning phase. We also found that this effect extended to previously unseen words.

In the second experiment, we found the predicted effect, but only for *good learners*. While the division of Experiment 2 participants into good and poor learners was not planned and should therefore be taken with caution, the distinction has a clear precedent in earlier work (Rácz, Hay, & Pierrehumbert, 2017) and makes good theoretical sense. We should not expect violation of a grammatical rule to be salient to participants who have not learned that rule. Indeed, it would have been inconsistent with our hypothesis if we had found such an effect for participants who had not learned the rule.

Taken together, our results suggest that unexpectedness increases the salience of variants and makes their social distribution easier to learn, deepening our understanding of the role of individual language experience in the acquisition of sociolinguistic meaning.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Foulkes, P., & Docherty, G. (2006). The social life of phonetics and phonology. *Journal of phonetics*, *34*(4), 409–438.

Fraundorf, S. H., & Jaeger, T. F. (2016). Readers generalize adaptation to newly-encountered dialectal structures to other unfamiliar structures. *Journal of memory and language*, *91*, 28–58.

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American chapter of the association for computational linguistics on language technologies* (pp. 1–8).

Jaeger, T. F., & Weatherholtz, K. (2016). What the heck is salience? How predictive language processing contributes to sociolinguistic perception. *Frontiers in Psychology*, *7*.

Kaschak, M. P., & Glenberg, A. M. (2004). This construction needs learned. *Journal of Experimental Psychology: General*, *133*(3), 450.

Labov, W. (1972). *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.

McDonald, S. A., & Shillcock, R. C. (2003). Low-level predictive inference in reading: The influence of transitional probabilities on eye movements. *Vision Research*, *43*(16), 1735–1751.

McRae, K., Spivey-Knowlton, M. J., & Tanenhaus, M. K. (1998). Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language*, *38*(3), 283–312.

R Core Team. (2014). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org/`

Rácz, P., Hay, J. B., & Pierrehumbert, J. B. (2017). Social salience discriminates learnability of contextual cues in an artificial language. *Frontiers in Psychology*, *8*.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379–423.

Squires, L. (2014). Social differences in the processing of grammatical variation. *University of Pennsylvania Working Papers in Linguistics*, *20*(2), 20.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. New York: Springer-Verlag. Retrieved from `http://ggplot2.org`

Human few-shot learning of compositional instructions

Brenden M. Lake[1,2], **Tal Linzen**[3], **and Marco Baroni**[2,4]
[1]New York University, [2]Facebook AI Research, [3]John Hopkins University, [4]ICREA

Abstract

People learn in fast and flexible ways that have not been emulated by machines. Once a person learns a new verb "dax," he or she can effortlessly understand how to "dax twice," "walk and dax," or "dax vigorously." There have been striking recent improvements in machine learning for natural language processing, yet the best algorithms require vast amounts of experience and struggle to generalize new concepts in compositional ways. To better understand these distinctively human abilities, we study the compositional skills of people through language-like instruction learning tasks. Our results show that people can learn and use novel functional concepts from very few examples (few-shot learning), successfully applying familiar functions to novel inputs. People can also compose concepts in complex ways that go beyond the provided demonstrations. Two additional experiments examined the assumptions and inductive biases that people make when solving these tasks, revealing three biases: mutual exclusivity, one-to-one mappings, and iconic concatenation. We discuss the implications for cognitive modeling and the potential for building machines with more human-like language learning capabilities.

Keywords: concept learning; compositionality; word learning; neural networks

People use their compositional skills to make critical generalizations in language, thought, and action. Once a person learns a new concept "photobombing", she or he immediately understands how to "photobomb twice", "jump and photobomb", or "photobomb vigorously." This example illustrates systematic compositionality, the algebraic capacity to understand and produce an infinite number of utterances from known components (Chomsky, 1957; Montague, 1970; Fodor, 1975). This ability is central to how people can learn from limited amounts of experience (Lake, Ullman, Tenenbaum, & Gershman, 2017), and uncovering its computational basis is an important open challenge.

There have been dramatic advances in machine language capabilities, yet the best algorithms require tremendous amounts of training data and struggle with generalization. These advances have been largely driven by neural networks, a class of models that has been long criticized for lacking systematic compositionality (Fodor & Pylyshyn, 1988; Marcus, 1998; Fodor & Lepore, 2002; Marcus, 2003; Calvo & Symons, 2014). Neural networks have developed substantially since these classic critiques, yet recent work evaluated contemporary neural networks and found they still fail tests of compositionality (Lake & Baroni, 2018; Bastings, Baroni, Weston, Cho, & Kiela, 2018; Loula, Baroni, & Lake, 2018). To evaluate compositional learning, Lake and Baroni (2018) introduced the SCAN dataset for learning instructions such as "walk twice and jump around right," which were built compositionally from a set of primitive instructions (e.g., "run" and "walk"), modifiers ("twice" or "around right"), and conjunctions ("and" or "after"). The authors found that modern recurrent neural networks can learn how to "run" and to "run twice" when both of these instructions occur in the training phase, yet fail to generalize to the meaning of "jump twice" when "jump" but not "jump twice" is included in the training data.

Classic arguments about the human ability to generalize have mostly rested on thought experiments. The latter, however, might underestimate facilitating factors, such as our knowledge of English, on which we are undoubtedly relying when interpreting "photobombing twice". In this paper, we study the scope and nature of people's compositional learning abilities through artificial instruction learning tasks that minimize reliance on knowledge of a specific language. The tasks require mapping instructions to responses, where an instruction is a sequence of pseudowords and a response is a sequence of colored circles. These tasks follow the popular sequence-to-sequence (seq2seq) framework and studied in Lake and Baroni (2018) and used to great effect in recent machine learning (e.g., machine translation; Sutskever, Vinyals, & Le, 2014). Seq2seq tasks require a learner to first read a sequence of input symbols, and then produce a sequence of output symbols (Fig. 1), whereby the input and output sequences can have different lengths. This framework allows us to directly compare humans and recent recurrent neural network architectures, while providing enough flexibility and richness to study key aspects of compositional learning. Moreover, the seq2seq problems investigated here present a novel challenge for both human and machine learners: unlike standard seq2seq benchmarks, which provide the learner with thousands of paired input and output examples, our "few-shot learning" paradigm provides the learner with only a handful of training examples.

Our tasks differ from the artificial grammar learning (Reber, 1967; Fitch & Friederici, 2012), rule learning (Marcus, Vijayan, Bandi Rao, & Vishton, 1999), and program learning (Stuhlmuller, Tenenbaum, & Goodman, 2010) paradigms in that we do not ask participants to implicitly or explicitly determine if items are grammatical. Instead, we ask them to process input sequences in a pseudo-language in order to generate output sequences ("meanings"). Asking participants to associate new words or sentences with visual referents is a standard practice in psycholinguistics (e.g., Bloom, 2000; Wonnacott, Boyd, Thomson, & Goldberg, 2012, and references there). Some of this work is particularly close to ours in that it studies the biases underlying linguistic generalization (e.g., Hudson Kam & Newport, 2009; Fedzechkina, Newport, & Jaeger, 2016). However, we are not aware of other studies that adopted the sequence-to-sequence language-to-meaning paradigm we are proposing here. Moreover, the biases studied in the earlier miniature language literature are more specific to grammatical phenomena attested in

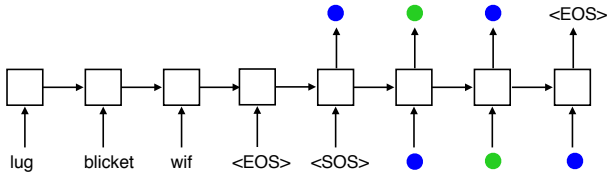


Figure 1: A sequence-to-sequence (seq2seq) recurrent neural network applied to few-shot instruction learning. Instructions are provided in a novel language of pseudowords and processed with an encoder network (in this case, the instruction is "lug blicket wif"), in order to generate an output sequence using a decoder network ("BLUE GREEN BLUE"). The symbols <EOS> and <SOS> denote end-of-sentence and start-of-sentence, respectively. The encoder (left) ends with the first <EOS> symbol, and the decoder (right) begins with <SOS>.

language (e.g., pertaining to linguistic syntax and morphology) than the basic generalization preferences we are exploring here.

Experiment 1: Few-shot instruction learning

Participants were asked to learn novel instructions from limited demonstrations. The task was inspired by the SCAN dataset for evaluating compositional learning in machines (Lake & Baroni, 2018), adapted to be novel and tractable for human learners in the lab. Instead of following instructions in English, participants learned to interpret and execute instructions in a novel language of pseudowords (e.g., "zup blicket lug") by producing a sequence of abstract outputs (a sequence of colored circles; Fig. 2). Some pseudowords were primitive instructions that correspond to a single output symbol, while other pseudowords are interpreted as functions that need to be applied to arguments to construct the output. As in SCAN, one primitive ("zup") is only presented in isolation during study but is evaluated compositionally during test, appearing in each test instruction. To perform well, participants must learn the meaning of each function from just a small number of demonstrations, and then generalize to new primitives and more complex compositions than previously observed.

Stimuli. The instructions consisted of seven possible pseudowords and the output sequences consisted of four possible response symbols (Fig. 2). Four primitive pseudowords are direct mappings from one input word to one output symbol (e.g., "dax" is "RED" and "wif" is "GREEN"), and the other pseudowords are functional terms that take arguments. To discourage a strategy based on word-to-word translation into English, the functional terms could not be easily expressed by single-word modifiers in English; they also formed phrases whose order would be unnatural in English.

The meanings of the functions were as follows. Function 1 ("fep" in Fig. 2) takes the preceding primitive as an argument and repeats its output three times ("dax fep" is "RED RED RED"). Function 2 ("blicket") takes both the preceding primitive and following primitive as arguments, producing their outputs in a specific alternating sequence ("wif blicket dax" is "GREEN RED GREEN"). Last, Function 3 ("kiki") takes both the preceding and following strings as input, processes them, and concatenates their outputs in reverse order ("dax kiki lug" is "BLUE RED"). We also tested Function 3 in cases where its arguments were generated by the other functions, exploring function composition ("wif blicket dax kiki lug" is "BLUE GREEN RED GREEN"). During the study phase (see Methods below), participants saw examples that disambiguated the order of function application for the tested compositions (Function 3 takes scope over the other functions).

Methods. Thirty participants in the United States were recruited using Amazon Mechanical Turk and the psiTurk platform (Gureckis et al., 2015). Participants were informed that the study investigated how people learn input-output associations, and that they would be asked to learn a set of commands and their corresponding outputs. Learning proceeded in a curriculum with four stages, with each stage featuring both a study phase and a test phase. In the first three stages, during the study phase participants learned individual functions from just two demonstrations each (Functions 1 through 3; Fig. 2). In the final stage, participants learned to interpret complex instructions by combining these functions (Function compositions; Fig. 2).

Each study phase presented participants with a set of example input-output mappings. For the first three stages, the study instructions always included the four primitives and two examples of the relevant function, presented together on the screen. For the last stage, the entire set of study instructions was provided together in order to probe composition. During the study phases, the output sequence for one of the study items was covered and participants were asked to reproduce it, given their memory and the other items on the screen. Corrective feedback was provided, and participants cycled through all non-primitive study items until all were produced correctly or three cycles were completed. The test phase asked participants to produce the outputs for novel instructions, with no feedback provided. The study items remained on the screen for reference, so that performance would reflect generalization in the absence of memory limitations. The study and test items always differed from one another by more than one primitive substitution (except in the Function 1 stage, where a single primitive was presented as novel argument to Function 1). Some test items also required reasoning beyond substituting variables, and in particular understanding longer compositions of functions than were seen in the study phase.

The response interface had a pool of possible output symbols which could be clicked or dragged to the response array. The circles could be rearranged within the array or cleared with a reset button. The study and test set only used four output symbols, but the pool provided six possibilities (that is, there were two extra colors that were not associated to pseudowords), to discourage reasoning by exclusion. The assignment of nonsense words to colors and functions was randomized for each participant (drawn from nine possible nonsense words and six colors), and the first three stages were

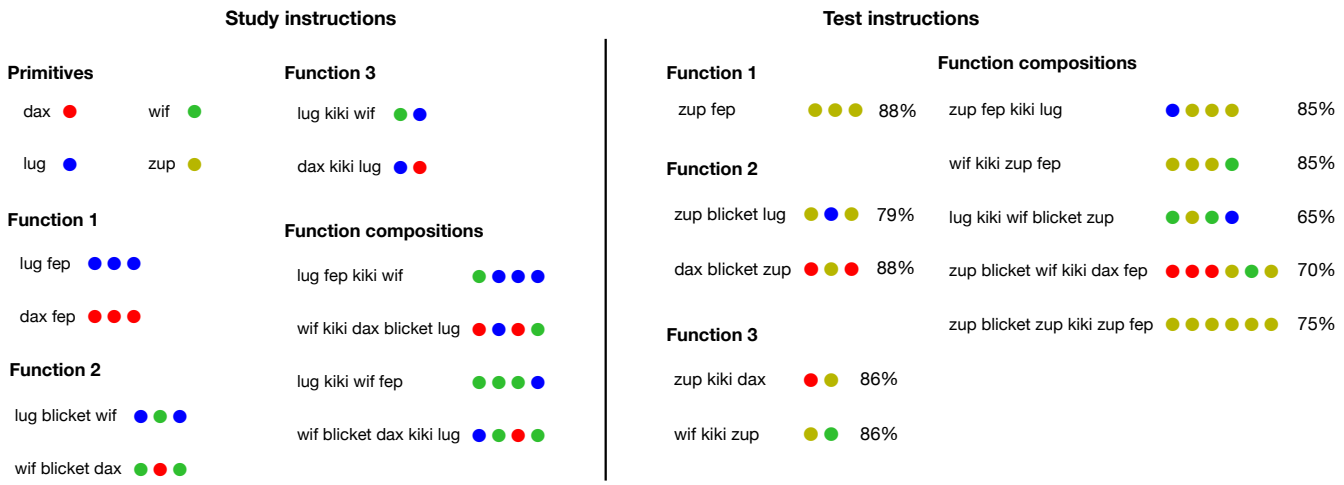


Figure 2: Few-shot learning of instructions in Exp. 1. Participants learned to execute instructions in a novel language of pseudowords by producing sequences of colored circles. Generalization performance is shown next to each test instruction, as the percent correct across participants. The pseudowords and colors were randomized for each participant; the figure illustrates an example of such an assignment.

presented in random order.

We used several strategies to ensure that our participants were paying attention. First, before the experiment, participants practiced using the response interface and had to pass an instructions quiz; they cycled through the quiz until they passed it. Second, catch trials were included during the test phases, probing the study items rather than new items, with the answers clearly presented on the screen above. There was one catch trial per stage (except the last stage had two); a participants' test data was excluded if the participant missed two or more catch trials ($n = 5$). Finally, test phases were also excluded if the corresponding study phases were not passed in the allotted time (13% of remaining data).

Recurrent neural networks. Standard sequence-to-sequence recurrent neural networks (RNNs; Fig. 1) failed to generalize from the study set to the test set. RNNs were trained using supervised learning on the 14 study instructions and evaluated on the test instructions (Fig. 2), using the best overall architecture from Lake and Baroni (2018) on the related SCAN benchmark (2-layer LSTM encoder and decoder, 200 hidden units per layer, a dropout probability of 0.5, no attention). This network (Fig. 1) consists of two neural networks working together: an encoder RNN that processes the instruction and embeds it as a vector, and a decoder RNN that decodes this vector as a sequence of output symbols. Another top architecture from Lake and Baroni was also evaluated (1-layer LSTM encoder and decoder, 100 hidden units per layer, dropout 0.1, with attention). The training setup mimicked Lake and Baroni but with 10,000 instruction presentations, corresponding to about 700 passes through the training data (epochs). Several variants of the architectures were also trained, repeatedly reducing the number of hidden units by half until there were only three hidden units per layer. Averaged across five random seeds, no architecture generalized better than 2.5% correct on the test instructions, confirming Lake and Baroni's conclusion that seq2seq RNNs struggle with few-shot learning and systematic generalization.

Results. Human participants showed an impressive ability to learn functions from limited experience and generalize to novel inputs, as summarized in Fig. 2. In the first three stages, performance was measured separately for each functional term after exclusions through the above attention criteria. Average performance across participants was 84.3% correct ($n = 25$), counting sequences as correct only if every output symbol was correct. Measured for individual functions, accuracy was 88.0% ($n = 25$) for Function 1, 83.3% ($n = 24$) for Function 2, and 86.4% ($n = 22$) for Function 3.[1]

Participants were also able to compose functions together to interpret novel sequences of instructions. In the final stage, accuracy on complex instructions was 76.0% ($n = 20$). People could generalize to longer and more complex instructions than previously observed, an ability that seq2seq neural networks particularly struggle with (Lake & Baroni, 2018). During the study phase, the most complex instruction consisted of five input pseudowords requiring two function compositions, producing four output symbols. At test, most participants could successfully go beyond this, correctly processing six input pseudowords requiring three function compositions, producing six output symbols (72.5% correct).

The pattern of errors showcases intriguing alternative hypotheses that participants adopted. Some errors were suggestive of inductive biases and assumptions that people bring to the learning task—principles that are reasonable a priori and consistent with some but not all of the provided demonstrations. For instance, many errors can be characterized by a bias we term "one-to-one," the assumption that each input

[1]The number of participants varies since data was included on the basis of passing the study phase. For comparison, the overall accuracy with no exclusions at all was 72.0%.

symbol corresponds to exactly one output symbol, and that inputs can be translated one-by-one to outputs without applying complex functional transformations. This characterized 24.4% of all errors.[2] Other errors involved misapplication of Function 3, which required concatenating its arguments in reverse order. When participants made an error, they often concatenated but did not reverse the argument (23.3% of errors for instructions using Function 3), a bias we term "iconic concatenation," referring to a preference for maintaining the order of the input symbols in the order of the output symbols. Forms of iconic concatenation are widely attested in natural language, and constitute important biases in language learning (Haiman, 1980; Goldin-Meadow, So, Özyürek, & Mylander, 2008; de Ruiter, Theakston, Brandt, & Lieven, 2018).

In sum, people learn in several ways that go beyond powerful seq2seq neural networks. People can learn novel functions from as few as two examples and generalize in systematic ways, appropriately applying the functions to previously unused input variables. People can also compose these novel functions together in ways not observed during training. Finally, people appear to bring strong inductive biases to this learning challenge, which may contribute to both their learning successes and failures.

Experiment 2: Inductive biases in instruction learning

This experiment investigated the inductive biases that appeared to influence the previous task. We devised a new set of seq2seq problems that were intentionally ambiguous and compatible with a number of possible generalizations, related to the "poverty of the stimulus" paradigm in experimental linguistics (Wilson, 2006; McCoy, Frank, & Linzen, 2018). These problems provide a more direct window into people's inductive biases because the information provided is insufficient for deducing the correct answer. The design also parametrically varied the context under which the biases were evaluated to better understand their structure and scope.

This experiment studies the one-to-one and iconic concatenation biases identified above, as well as the mutual exclusivity (ME) bias that has been studied extensively in the developmental literature. Classic studies of ME present children with a familiar and an unfamiliar object (e.g., a ball and a spatula; Markman & Wachtel, 1988), or two unfamiliar objects in which one is familiarized during the experiment (Diesendruck & Markson, 2001). When given the instruction "show me the zup," children typically understand "zup" to refer to the novel object rather than acting as a second name for the familiar object. In our instruction learning paradigm, ME is operationalized as the inference that if "dax" means "RED", then "zup" is likely another response besides "RED." Although Exp. 1 did not naturally lend itself to probing the effect of the ME bias, we conjecture that it is because of the latter that participants rapidly eliminated many degenerate solutions (such as all strings referring to the same output item) in virtually any word learning experiment. We thus want to study the impact of ME more explicitly.

Methods. Twenty-eight participants in the United States were recruited using Mechanical Turk and psiTurk. The instructions were as similar as possible to the previous experiment. In contrast, the curriculum of related stages in the previous experiment was replaced by 14 independent trials that evaluated biases under different circumstances. Each trial provided a set of study instructions (input-output mappings) and asked participants to make a judgment about a single new test instruction. To highlight the independence between trials, the pseudoword and colors were re-randomized for each trial from a larger set of 20 possible pseudowords and 8 colors. To emphasize the inductive nature of the task, participants were told that there were multiple reasonable answers for a given trial and were instructed to provide a reasonable guess.

The trials were structured as follows. Six trials pertain to ME and whether participants are sensitive to counterevidence and the number of options in the response pool (e.g., Fig 3A left and middle columns). Three trials pertain to iconic concatenation and how participants concatenate instructions together in the absence of demonstrations (e.g., Fig 3A right column). Three additional trials pertain to how people weigh ME versus one-to-one in judgments that necessarily violate one of these biases (not shown in figure). Finally, two catch trials queried a test instruction that was identical to a study instruction. The design minimized the risk that the biases could be learned from the stimuli themselves. None of the study instructions demonstrated how to concatenate, facilitating a pure evaluation of concatenation preferences. In the novel test trials, 6 instructions supported ME and 6 violated it, although both catch trials also supported ME. We did not explicitly control for the one-to-one bias. Missing a catch trial was the only criterion for exclusion ($n = 6$). There was no memory quiz for the study items since each contained just a few instructions.

Results. There was strong evidence for each of the three inductive biases. The classic mutual exclusivity (ME) effect was replicated within our seq2seq learning paradigm. If "dax" means "RED", what is a "zup"? As shown in the top-left cell of Fig 3A, most participants (18 of 22; 81.8%) chose a single "BLUE" symbol as their response if the pool provided only "RED" and "BLUE" as options, and a larger fraction (20 of 22; 90.9%) followed ME by choosing a (possibly multi-element) meaning different from "RED."

While the ME effect was robust, it was sensitive to context and was not rigidly applied. The other ME trials examined the influence of two additional factors (Fig 3A left and middle columns): the number of contradictory examples provided (0–2; Fig 3A rows) and the number of output symbols available in the response pool (2 vs. 6; Fig 3A columns). With these two variables as fixed effects, we fit a logistic

[2]These errors are defined as responses such that the input and output sequence have the same length, and each input primitive is replaced with its provided output symbol. Function words are replaced with an arbitrary output symbol.

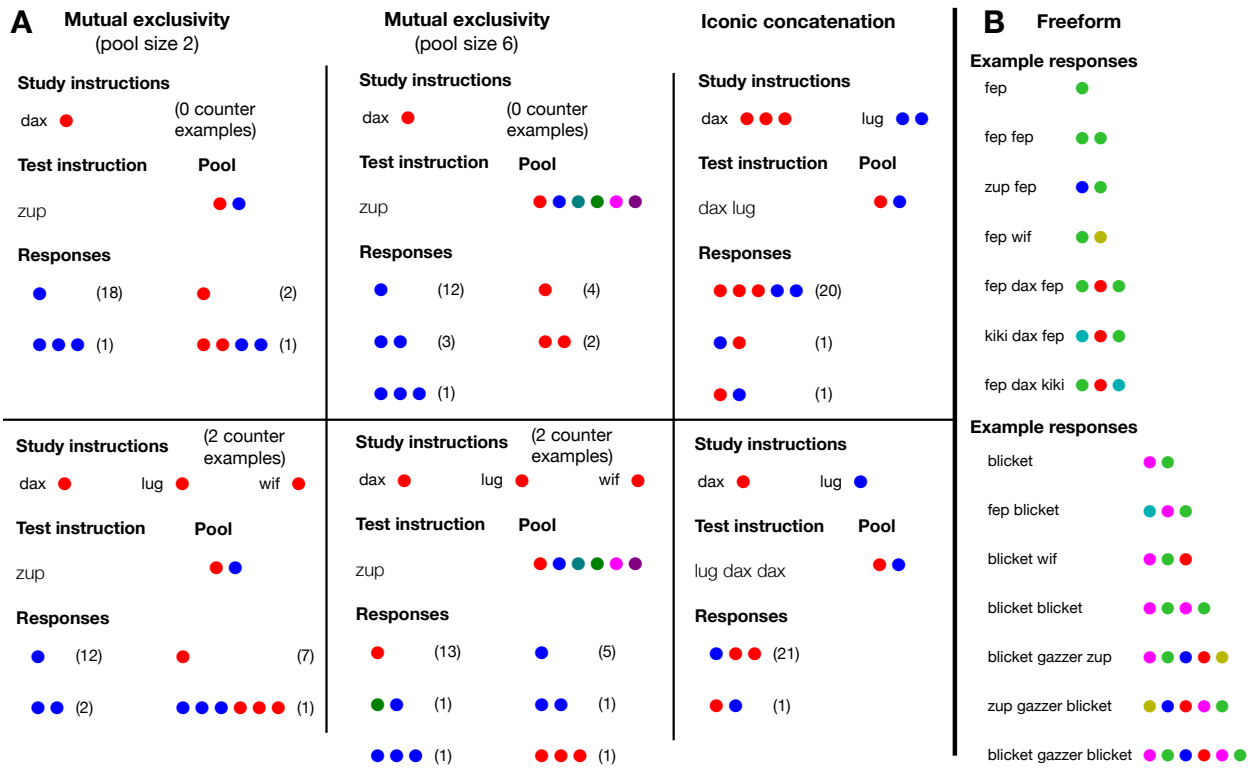


Figure 3: Inductive biases in seq2seq word learning from Exp. 2 and 3. A: In Exp. 2, Participants were asked to respond to the Test instruction given the Study instructions, using only the symbols in the Pool. Shown are four examples trials (left and middle columns) examining mutual exclusivity with varying counter-evidence (varied across rows) and pool sizes (varied across columns), and two example trials (right column) examining iconic concatenation. All unique participant responses are shown with their frequency in parentheses. A canonical assignment of pseudowords and colors was used to aggregate the data, but it was randomized in the experiment. B: Responses from two participants in the Exp. 3 free-form task. The top participant was consistent with ME, one-to-one, and iconic concatenation, while the bottom participant was missing the one-to-one bias. For part B the words and colors are as-seen in the experiment.

mixed model predicting whether or not a response was consistent with ME. Both the number of contradictory examples ($\beta = 1.76$, $SE = 0.483$, $Z = 3.64$, $p < 0.001$) and pool size ($\beta = 2.05$, $SE = 0.698$, $Z = 2.93$, $p < 0.01$) were significant predictors, indicating that people were willing to override or weaken ME when faced with more ME counter-evidence (or equivalently in our case, positive evidence that "RED" is the right answer), or when more output symbols were available in the pool (Fig. 4). The second effect is intriguing. Although we leave a detailed analysis to future work, we conjecture that it stems from pragmatic reasoning on behalf of the participants: When five yet-to-be-named objects are in the pool, ME is such a weak heuristic that participants might conclude that the experiment is not asking them to rely on it.

There was strong confirmatory evidence for iconic concatenation. Across three trials that examined this bias in various forms, we found that 93.9% ($n = 22$) of responses were consistent with iconic concatenation, even though no examples of concatenation were provided during this experiment (Fig. 3A right column). In three trials where all of the output symbols in the pool were already assigned to unique pseudowords, participants had to choose between violating ME by reassigning an output symbol, or violating one-to-one by choosing a more complex functional or multi-element meaning. Interestingly, the responses were evenly split (50.0%) between following one principle versus the other.

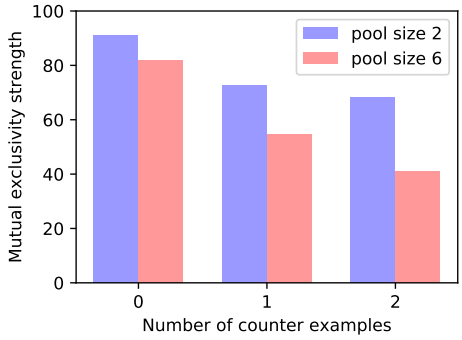


Figure 4: The proportion of responses consistent with mutual exclusivity (y-axis) declines with the number of contradictory examples and the number of output symbols available in the response pool.

Taken together, there was substantial support for three inductive biases in how people approach compositional learning in sequence-to-sequence mapping problem, confirming our hypotheses from Exp. 1. A drawback of this experiment's within-subjects design was the risk of judgments interfering with one another. The experiment used heavy randomization

and mitigated the risk that the biases could be learned from the aggregate statistics of the stimuli, but these controls were not perfect. The next experiment addresses these concerns.

Experiment 3: Inductive biases in free-form response

In this experiment, participants responded to novel instructions without receiving any demonstrations, e.g., making plausible guesses for the outputs of instructions "fep", "fep fep," and "fep wif" and how they relate to one another. This design offers the purest examination of people's assumptions since they have no relevant evidence about how to respond.

Methods. Thirty participants in the United States were recruited using Mechanical Turk and psiTurk. The instructions were similar as possible to the previous experiments, using Exp. 2's wording emphasizing there are multiple reasonable answers and to provide a reasonable guess. Participants produced the output for seven novel instructions utilizing five possible pseudowords (Fig. 3B). Responses were entered on a single page, allowing participants to edit and maintain consistency. Participants also approved a summary view of their responses before submitting. There were six pool options, and the assignment of pseudowords and item order were random. One participant was excluded because she or he reported using an external aid in a post-test survey.

Results. The results provide strong confirmatory evidence for the three key inductive biases: ME, iconic concatenation, and one-to-one. Although the task was highly underdetermined, there was a substantial structure in the responses, unlike an untrained seq2seq recurrent neural network which would respond arbitrarily. The majority of participants (17 of 29; 58.6%) responded in an analogous way to the participant shown at the top of Fig. 3B. This set of responses is perfectly consistent with all three inductive biases, assigning a unique output symbol to each input symbol and concatenating to preserve the input ordering. Other participants produced alternative hypotheses that followed some but not all the inductive biases. Overall, 23 of 29 participants (79.3%) followed iconic concatenation, assigning consistent (but possibly multi-element) output sequences to individual input words (e.g., Fig. 3B bottom). In all but one of these cases, each input word was assigned a unique output sequence, abiding by mutual exclusivity (22 of 23; 95.7%).

Discussion and Conclusions

People learn in fast and flexible ways not captured by today's algorithms. After learning how to "dax", people can immediately understand how to "dax slowly" or "dax like you mean it." These types of inferences are critical to language learning and understanding, yet modern recurrent neural networks struggle to generalize in similarly systematic ways (Lake & Baroni, 2018; Loula et al., 2018). To study these distinctively human abilities, we examined people's compositional skills in novel language-like instruction learning problems. The tasks followed the popular sequence-to-sequence (seq2seq) framework from machine learning, allowing humans and machines to be compared side-by-side. Experiment 1 examined how people learn novel instructions from examples, asking participants to interpret sequences of pseudowords by producing sequences of abstract output symbols. People could learn new functions from just two examples and successfully applied them to new inputs, while standard seq2seq recurrent neural networks (RNNs) failed to generalize. People could also handle longer sequences that require more compositions than previously observed, again surpassing the skills of powerful neural networks. Inspired by the errors participants made, Experiments 2 and 3 investigated inductive biases that constrain human learning, revealing that human learners draw upon mutual exclusivity (ME), iconic concatenation, and one-to-one in seq2seq word learning tasks.

More than a source of error, these biases provide important inductive constraints. If people interpreted the instruction as unanalyzable wholes, they would have no basis for generalization. Instead, people facilitate generalization by favoring hypotheses that assign unique and consistent meanings to individual words and follow certain input/output ordering constraints. As the final experiment shows, participants assume these characteristics before observing any data. The assumptions turn out to be powerful, characterizing most of the word meanings in Exp. 1 and the related SCAN benchmark, even though neither was designed with these biases in mind. Notably, the biases can mislead when learning function words; this was the case in many of the errors made in Exp. 1.

Future work should investigate the origin and scope of these biases through other compositional learning tasks. To the extent that our tasks evoke language learning, they could recruit biases known in the developmental literature such as mutual exclusivity (Markman & Wachtel, 1988). If the outputs are viewed as objects, one-to-one is related to the whole object assumption in word learning (Macnamara, 1982). Alternatively, if the outputs are viewed as events or actions, iconic concatenation could be justified by aligning a description with its content in time (de Ruiter et al., 2018). Another important line of future work should be providing a more explicit account of how the biases, which we observed emerging in participants' errors, are also aiding faster learning of the correct generalizations.

These insights from human learning could be fruitfully incorporated into machine learning. These biases could facilitate learning of seq2seq problems such as machine translation and semantic parsing, or related image2seq problems such as caption generation. Powerful seq2seq models do not have these inductive biases, suggesting a path to building more powerful and human-like learning architectures by incorporating them.

Acknowledgments

We thank the NYU ConCats group, Michael Frank, Kristina Gulordava, Germán Kruszewski, Roger Levy, and Adina Williams for helpful suggestions.

References

Bastings, J., Baroni, M., Weston, J., Cho, K., & Kiela, D. (2018). Jump to better conclusions: SCAN both left and right. In *Proceedings of the emnlp blackboxnlp workshop* (pp. 47–55). Brussels, Belgium.

Bloom, P. (2000). *How children learn the meanings of words*. Cambridge, MA: MIT Press.

Calvo, P., & Symons, J. (Eds.). (2014). *The architecture of cognition: Rethinking Fodor and Pylyshyn's systematicity challenge*. Cambridge, MA: MIT Press.

Chomsky, N. (1957). *Syntactic structures*. Berlin, Germany: Mouton.

de Ruiter, L., Theakston, A., Brandt, S., & Lieven, E. (2018). Iconicity affects children's comprehension of complex sentences: The role of semantics, clause order, input and individual differences. *Cognition*, *171*, 202–224.

Diesendruck, G., & Markson, L. (2001). Children' s Avoidance of Lexical Overlap: A Pragmatic Account. *Developmental Psychology*, *37*(5), 630–641.

Fedzechkina, M., Newport, E., & Jaeger, F. (2016). Miniature artificial language learning as a complement to typological data. In L. Ortega, A. Tyler, H. Park, & M. Uno (Eds.), *The usage-based study of language learning and multilingualism* (pp. 211–232). Washington, DC: Georgetown University Press.

Fitch, T., & Friederici, A. (2012). Artificial grammar learning meets formal language theory: An overview. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1598), 1933–1955. doi: 10.1098/rstb.2012.0103

Fodor, J. (1975). *The Language of Thought*. Harvard University Press.

Fodor, J., & Lepore, E. (2002). *The compositionality papers*. Oxford, UK: Oxford University Press.

Fodor, J., & Pylyshyn, Z. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*, 3–71.

Goldin-Meadow, S., So, W. C., Özyürek, A., & Mylander, C. (2008). The natural order of events: How speakers of different languages represent events nonverbally. *Proceedings of the National Academy of Sciences*, *105*(27), 9163–9168.

Gureckis, T. M., Martin, J., McDonnell, J., Alexander, R. S., Markant, D. B., Coenen, A., ... Chan, P. (2015). psiTurk: An open-source framework for conducting replicable behavioral experiments online. *Behavioral Research Methods*.

Haiman, J. (1980). The iconicity of grammar: Isomorphism and motivation. *Language*, *56*(3), 515–540.

Hudson Kam, C., & Newport, E. (2009). Getting it right by getting it wrong: When learners change languages. *Cognitive Psychology*, *59*(1), 30–66.

Lake, B. M., & Baroni, M. (2018). Generalization without Systematicity: On the Compositional Skills of Sequence-to-Sequence Recurrent Networks. In *International Conference on Machine Learning (ICML)*.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*, E253.

Loula, J., Baroni, M., & Lake, B. M. (2018). Rearranging the Familiar: Testing Compositional Generalization in Recurrent Networks. *arXiv preprint*. Retrieved from http://arxiv.org/abs/1807.07545

Macnamara, J. (1982). *Names for things: A study in human learning*. Cambridge, MA: MIT Press.

Marcus, G. F. (1998). Rethinking Eliminative Connectionism. *Cognitive Psychology*, *282*(37), 243–282.

Marcus, G. F. (2003). *The Algebraic Mind: Integrating Connectionism and Cognitive Science*. Cambridge, MA: MIT Press.

Marcus, G. F., Vijayan, S., Bandi Rao, S., & Vishton, P. M. (1999). Rule Learning by Seven-Month-Old Infants. *Science*, *283*(5398), 77–80.

Markman, E. M., & Wachtel, G. F. (1988). Children's Use of Mutual Exclusivity to Constrain the Meanings of Words. *Cognitive Psychology*, *20*, 121–157.

McCoy, R. T., Frank, R., & Linzen, T. (2018). Revisiting the poverty of the stimulus: Hierarchical generalization without a hierarchical bias in recurrent neural networks. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 2093—2098).

Montague, R. (1970). Universal Grammar. *Theoria*, *36*, 373–398.

Reber, A. (1967). Implicit learning of artificial grammars. *Verbal Learning and Verbal Behavior*, *5*(6), 855–863.

Stuhlmuller, A., Tenenbaum, J. B., & Goodman, N. D. (2010). Learning Structured Generative Concepts. In *Proceedings of the Thirty-Second Annual Conference of the Cognitive Science Society.*

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks. In *Advances in Neural Information Processing Systems (NIPS).*

Wilson, C. (2006). Learning phonology with substantive bias: An experimental and computational study of velar palatalization. *Cognitive Science*, *30*(5), 945–982.

Wonnacott, E., Boyd, J., Thomson, J., & Goldberg, A. (2012). Input effects on the acquisition of a novel phrasal construction in 5 year olds. *Journal of Memory and Language*, *66*, 458–478.

On Formal Verification of ACT-R Architectures and Models

Vincent Langenfeld (langenfv@tf.uni-freiburg.de)
Department of Computer Science, Albert-Ludwigs-Universität Freiburg

Bernd Westphal (westphal@tf.uni-freiburg.de)
Department of Computer Science, Albert-Ludwigs-Universität Freiburg

Andreas Podelski (podelski@tf.uni-freiburg.de)
Department of Computer Science, Albert-Ludwigs-Universität Freiburg

Abstract

Subject of this article is the question whether the potential for automatic defect analysis for symbolic timed ACT-R models as demonstrated in earlier work can be developed into a scalable and comprehensible technique. We present a formal, operational model of an ACT-R architecture and a translation scheme of ACT-R models into timed automata. We have applied this translation to ACT-R models and report on scalability experiments with automatic defect analysis.

Keywords: ACT-R; Cognitive Architecture; Formal Methods; Timed Automata; Modelling

Introduction

ACT-R (Anderson, 1983, 2009) is a cognitive architecture (an implementation of a unified theory of cognition) that is widely used in cognitive modelling to validate psychological theories. A psychological theory is a hypothesis on how a given task is solved by humans. Psychological theories can be validated by constructing an ACT-R model that implements the psychological theory and comparing the model's predictions to experimental data.

The work (Langenfeld, Westphal, Albrecht, & Podelski, 2018) points out that it is critical for this approach that an ACT-R model *correctly* implements the psychological theory because an incorrect ACT-R model (wrt. the psychological theory) may lead to a false rejection or false acceptance of an invalid theory as follows. An incorrect ACT-R model may give predictions that *do not match* experimental data although the theory's predictions would (thus false rejection), or the model may give predictions that *do match* the experimental data although the theory's predictions would not (thus false acceptance).

(Langenfeld et al., 2018) introduce the notion of *model defect* in general, i.e., any kind of programming error in production rules like simple typing errors, forgotten conditions or requests, etc. They formally study the following three model properties that can indicate the presence of errors. The first considered model property is called *deadlock*, a situation where model execution cannot continue although the model is not in a final state. The second property is *correctness of the mental model*, that is, the questions whether it is possible to observe expected chunks (as defined by the psychological theory) during model executions and whether it is impossible to observe unexpected chunks. The third property is *timing feasibility*, that is, whether an ACT-R model is principally able to reproduce timing aspects that are observed in experimental data (e.g. given a model, is it possible to complete the necessary computation steps of the ACT-R model within the time frame observed with human participants). Using an abstract, formal semantics of ACT-R (Albrecht & Westphal, 2014b), it is principally possible, effective, and useful from a modeller's point of view to automatically and exhaustively analyse ACT-R models for the absence of defects (Langenfeld et al., 2018). Spotting such errors by simulation alone is, in contrast, tedious and time consuming in general.

Subject of this article is the question whether the potential for automatic defect analysis for symbolic timed ACT-R models as mentioned above can be developed into a scalable and comprehensible technique. To this end, we present a formal, operational model of an ACT-R architecture and a translation scheme of ACT-R models into the same formalism of timed automata (Alur & Dill, 1994), that allows us to easily model the discrete, timing, and concurrency aspects of ACT-R and that is well supported by existing analysis tools (Behrmann, David, & Larsen, 2004). We have applied this translation to artificial ACT-R models as well as an ACT-R model from the research literature and we have found the analysis of the resulting network of timed automata (using existing tools) to scale well, both in number of chunks and number of production rules. By using the formalism of timed automata, we obtain a comprehensible model including all architecture aspects, hence the potential to analyse theories regarding architectures in addition to psychological theories.

Related Work. Formalisations of the ACT-R semantics appear in (Albrecht & Westphal, 2014b) and later in (Gall & Frühwirth, 2014; Gall & Frühwirth, 2018), and are used towards comparing cognitive architectures (Ragni et al., 2018).

Preliminary results on the formal analysis of ACT-R models for defects have been presented in (Albrecht & Westphal, 2014a) and elaborated in (Langenfeld et al., 2018). The feasibility of such analyses is investigated by encoding simplified fragments of ACT-R architecture and model aspects and selected rules into logical formulae that can effectively be analysed for satisfiability. This work, in contrast, supports a wider range of analysis goals and aims at a comprehensible model of an architecture and a complete ACT-R model.

An analysis procedure for confluence of ACT-R models can be obtained by encoding an ACT-R architecture and models in constraint handling rules and solving the confluence problems in this domain (Gall & Frühwirth, 2017).

Preliminaries

F-ACT-R. The formal description of ACT-R (Albrecht, 2013; Albrecht & Westphal, 2014a) differentiates between a syntactical

description of the ACT-R model and description of the semantics assigned to the constructs of the model by a cognitive architecture.

The abstract syntax of ACT-R defines a model over a set of module signatures. A module signature consists of a finite set of buffers B, a finite set of module queries Q, and a finite set of action symbols A. A production rule r is a pair of a precondition and an action. A precondition is a proposition over buffer slots and module queries. An action is a set of similar propositions together with a buffer and an action symbol. An ACT-R model is a finite set of production rules $R = \{r_1,...r_n\}$ and a finite set of chunks $\{c_0,...,c_n\}$.

An ACT-R architecture consists of a interpretation function for the action symbols of modules and a production rule selection mechanism. A cognitive state is a function γ from buffers to pairs (c,d) where c is a chunk and d is a time delay. A pair in γ thus describes buffer contents (if $d = 0$) and buffer assignments in the future (if $d > 0$). The ACT-R architecture works in cycles of production rule selection and execution. Cognitive states γ,γ' are in a successor relation ($\gamma \xrightarrow{(r,t)} \gamma'$), if the selection mechanism chooses production rule r (consuming time t) whose precondition is fulfilled by the current cognitive state γ and γ' is the result of applying the interpretation of every action symbol of r to γ.

Running Example. The addition model from the ACT-R tutorial (Bothell, 2017b), Unit 1.7.1, models the addition of two numbers by counting up from the first number in as many steps as given by the second number. To implement counting, the model uses rules whose preconditions match the current number and retrieve the corresponding count fact, i.e. a pair of a number and its direct successor, from declarative memory. In our examples we use the production rule *initialize-addition*, which is only applicable at the beginning of the computation. Its precondition requires an empty goal buffer slot `sum`. To start the addition, its action assigns the first number of the addition to goal buffer slot `sum`, and 0 to the `count` slot that tracks counting of the second number. Then a retrieval for the successor of `sum` is started.

Timed Automata. Timed automata (Alur & Dill, 1994) are a formal, operational model of real-time systems, i.e., systems that have to compute outputs within certain time intervals. In the simplest case, a *timed automaton* $\mathcal{A}$ is a tuple $(L,B,\mathbb{X},I,E,\ell_{ini})$ comprising a finite set of *locations* L (including the *initial location* ℓ_{ini}), a set of *channels* B, and a set of *clocks* $\mathbb{X}$. Function I labels each location with a clock constraint (called location *invariant*), and E is a finite set of edges. An edge $(\ell,\alpha,\varphi,\rho,\ell')$ comprises source and destination location ℓ and ℓ', action α (which can be the internal action τ, or an *output* $b!$ or *input* $b?$ on a channel $b \in B$), clock constraint φ as *guard*, and the *update* $\rho \subseteq \mathbb{X}$ that denotes the set of clocks to be reset.

The operational semantics of a *network of timed automata* $\mathcal{N} = \mathcal{A}_1 \| \cdots \| \mathcal{A}_n$ ('$\|$' denoting parallel composition) is a labelled transition system over *configurations* $\langle \vec{\ell}, \nu \rangle$ where $\vec{\ell}_i$ is the *current location* of automaton $\mathcal{A}_i$ and $\nu : \mathbb{X} \to \mathbb{R}_0^+$ is a valuation of the clocks. Two configurations are in transition relation $\langle \vec{\ell},\nu \rangle \xrightarrow{\lambda} \langle \vec{\ell}',\nu' \rangle$ if and only if $\lambda \in \mathbb{R}_0^+$, $\vec{\ell} = \vec{\ell}'$, and $\nu' = \nu + \lambda$ satisfies the invariants of all locations in $\vec{\ell}$ (delay transition), or there is an edge $(\ell,\tau,\varphi,\rho,\ell') \in E_i$ such that $\vec{\ell}_i = \ell$, $\vec{\ell}'_i = \ell'$, φ is satisfied in ν, and ν' is obtained from ν by resetting the clocks in ρ to zero (internal transition), or there are two edges enabled in two different automata in $\mathcal{N}$ with complementary input and output actions (rendezvous transition). A *computation path* of a network of timed automata is a sequence of configurations starting with $\langle \vec{\ell}_0,\nu_0 \rangle$ where $\vec{\ell}_0$ comprises the initial locations, and ν_0 assigns value 0 to all clocks (and satisfies all location invariants), and subsequent configurations are in transition relation.

The modelling, simulation, and model-checking tool Uppaal (Behrmann et al., 2004) extends the simple case by features such as data variables, broadcast channels, and committed locations where no delay is possible. In the remaining article, we use a graphical representation of timed automata (see, e.g., Figure 1) where the double outline location is initial, locations marked by a 'C' are committed locations, and invariants (if any) are shown in purple. Edges are annotated with action (in cyan), guard (in green), and updates (in blue).

TA-ACT-R

In this section, we describe how we represent ACT-R models by networks of timed automata that can then be analysed for ACT-R model defects. Recall that an ACT-R model R is a finite set $\{r_1,...,r_n\}$ of production rules that has computations *on* an architecture A.

Given an ACT-R model R and an architecture A, we construct networks $\mathcal{N}^R$ and $\mathcal{N}^A$ of timed automata such that we can conclude from analysis results of the network $\mathcal{N}^R \| \mathcal{N}^A$ to the presence or absence of model defects in the ACT-R model on architecture A. Constructing the network $\mathcal{N}^A$ can be considered a one-time effort: In the case described below, we consider timed automata that follow the production rule selection mechanism and the behaviour of models in the ACT-R tool. Network $\mathcal{N}^A$ can be composed with any $\mathcal{N}^R$, i.e., with any network of a specific ACT-R model, as long as model R is compatible with the modules offered by A.

In the following paragraphs, we first describe the construction of the timed automata in $\mathcal{N}^A$ that model module behaviour, here on the example of the declarative module. Then we describe the construction of production rule automata to obtain $\mathcal{N}^R$, and conclude with the production rule selection mechanism in $\mathcal{N}^A$. Figure 5 visualises the overall structure and potential communication between the timed automata in $\mathcal{N}^R \| \mathcal{N}^A$ over shared channels.

Chunks and the Declarative Module. The declarative module is responsible for memory management, i.e. to maintain and recall chunks of previously learned information. Actions of production rules can initiate a recall of information, e.g., the successor of a number in the addition model, and the declarative module delivers one (of possibly many) matching chunks or none at all. Recalling information takes time: In ACT-R, the declarative module takes a certain amount of time to recall information (retrieval delay) or considers a recall failed after a time limit (retrieval threshold).

Our TA-ACT-R model of the declarative module is a set of timed automata that realise the behaviour described above. It comprises one timed automaton $\mathcal{A}^D$, and one timed automaton $\mathcal{A}^{ch}$

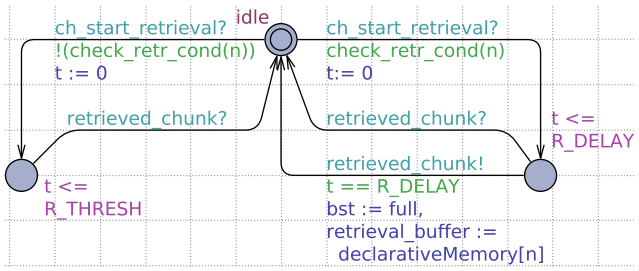


Figure 1: TA-ACT-R chunk automaton $\mathcal{A}^{ch}$.

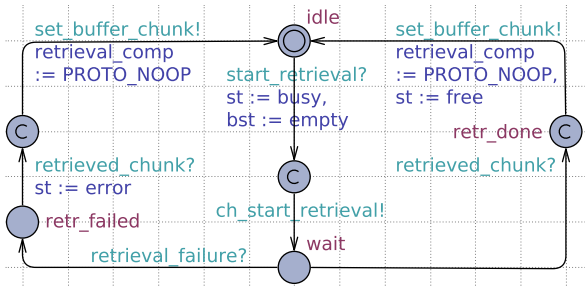


Figure 2: TA-ACT-R automaton $\mathcal{A}^{D}$ (declarative).

for each chunk in memory. The idea of this structure is that, for each recall action, $\mathcal{A}^{D}$ stimulates all chunk automata at once, and each chunk automaton with a chunk matching the current recall action offers its chunk to $\mathcal{A}^{D}$. Selection between matching chunks is non-deterministic (and exhaustively considered in analysis).

Figure 1 shows the TA-ACT-R chunk automaton $\mathcal{A}^{ch}$. From the initial location `idle` there are two possible ways of handling a recall action: Either the chunk managed by $\mathcal{A}^{ch}$ matches the request (continue to the right) or not (continue to the left). The distinction between the two cases is made by function `check_retr_cond()`, which hides the details of comparing the request (as specified by shared variables). Both edges (to the right and to the left of `idle`) reset clock t to 0. By the invariants of the bottom right (or left) locations (shown in purple), either an amount of time units corresponding to retrieval delay or retrieval threshold pass. In case of a match (bottom right location), automaton $\mathcal{A}^{ch}$ can send or receive on the broadcast channel `retrieved_chunk`. If multiple chunks match, exactly one chunk automaton non-deterministically acts as sender and all others receive simultaneously. In any case (including no matching chunk), the synchronisation moves the automaton back to location `idle` and the recall action is completed from the perspective of the chunk automaton. Only the chunk automaton acting as sender writes its chunk into the shared variable `retrieval_buffer` that models the retrieval buffer of the declarative module and sets the buffer flag `bst` to indicate that there is a chunk in the retrieval buffer. This update corresponds to placing the retrieved chunk in the declarative module's retrieval buffer.

That is, the timing behaviour of the declarative module is modelled in the chunk automata. Note that our model can support any number of chunks in memory, yet an upper bound on the number of $\mathcal{A}^{ch}$ automata needs to be fixed before analysing the model. This constraint corresponds to the observation that the majority of ACT-R models considers cognitive tasks that are solved in bounded time, and there is the assumption that only finitely many chunks can be used in bounded time. Yet an analysis of a TA-ACT-R model can detect if a given upper bound is sufficient to support a given ACT-R cognitive model. If not, the upper bound on the number of chunks can be increased and the analysis restarted, which in particular allows us to analyse the maximum number of chunks actually considered in a given ACT-R model. Note, that the specified number of chunks only restricts memory size but not memory content. Chunk content may be set in advance modelling pre-existing declarative knowledge (as in the addition model) and content may be acquired by chunk automata during run time (modelling learning of declarative knowledge).

Checking for matching chunks, retrieving one matching chunk (if any), and reporting the result of the recall action is organised by the timed automaton $\mathcal{A}^{D}$ shown in Figure 2.

In the TA-ACT-R model, the recall action is started by a synchronisation on channel `start_retrieval` with a rule automaton (see below) and taking the edge from location `idle` downwards. Without intermediate delay, the module flags are updated to indicate that the declarative module is busy and then the chunk automata are triggered by sending on channel `ch_start_retrieval` (cf. Figures 2 and 1) and changing to location `wait`. From location `wait`, there are two cases: either at least one chunk matched or none. The first case is handled by the sequence of edges to the right of `wait` (by synchronising with the chunk automata as explained above) and updating some more shared variables of the network, so that other automata in the complete network can access the recalled chunk. The second case (no chunk matched) is handled by the mostly symmetric sequence of edges to the left of `wait`, which sets the module's flags accordingly. In both cases, before returning to location `idle` and being ready for the next recall action, the procedural automaton is notified of completion by synchronisation on channel `set_buffer_chunk`.

Note that Figure 1 shows a simplified chunk automaton for brevity of this presentation. In general, the declarative module is able to learn new chunks. Further note that TA-ACT-R models the purely symbolic variant of ACT-R declarative memory, that is, fulfilling a request is a sufficient condition for a chunk being retrieved from memory. In general, retrieval through the declarative module is affected by an *activation* value that models the effect of frequent usage of a chunk and prevents chunks with an activation value below a given threshold from being retrieved from memory. The analysis of the TA-ACT-R model presented here hence detects model errors like deadlock or (in)correctness of the mental model under the assumption of perfect memory. These errors do not disappear when considering activation hence such errors should be removed before considering more expensive analysis with activation (which is future work).

Production Rules. Given an ACT-R model R consisting of the rules $r_1,\ldots,r_n$, the network $\mathcal{N}^R$ is the parallel composition of n rule automata, i.e. $\mathcal{N}^R = \mathcal{A}^{r_1} \| \cdots \| \mathcal{A}^{r_n}$.

Figure 3 shows a concrete rule automaton to illustrate the principal construction of rule automata. Each rule automaton has two

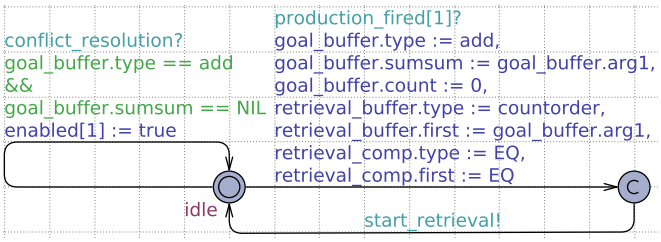


Figure 3: TA-ACT-A automaton $\mathcal{A}^r$ of the production rule *initialize-addition* of the addition model.

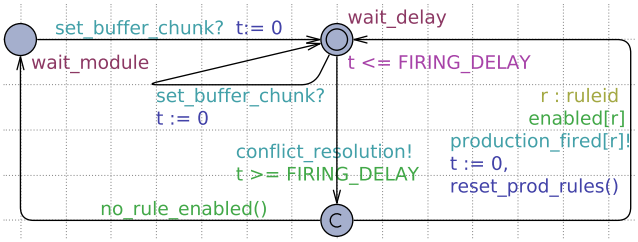


Figure 4: TA-ACT-R automaton $\mathcal{A}^P$ (procedural).

cycles (or phases) starting in the initial location `idle`. To the left is a single edge that synchronises with the procedural module automaton (see below) to determine the currently enabled production rules. The principle is similar to the selection of a matching chunk from the chunk automata by the declarative module automaton $\mathcal{A}^D$ in that it uses a broadcast channel (here `conflict_resolution`).

If the procedural module automaton sends on `conflict_resolution`, all rule automata simultaneously take the left edge from `idle` to `idle` if the guard is satisfied. Automaton $\mathcal{A}^r$ writing a value 1 into their position of the shared array variable `enabled` indicates that it is possible to fire rule r under the current module configuration.

If rule r is selected by the procedural module, the latter sends on the rendezvous channel `production_fired`$[r]$such that the rule automaton takes the sequence of edges to the right of `idle` (cf. Figure 3). The first edge in the sequence has updates according to the actions of the rule, followed by a sequence of edges that trigger activities of modules (in this example, of the declarative module discussed above).

Figure 3 actually shows the rule automaton of the production rule *initialize-addition* (cf. Preliminaries). The precondition of this rule, r_{ia} for short, is satisfied if the goal buffer does not yet hold an intermediate or final result. This precondition has a direct translation to the guard (shown in green) of the left edge in Figure 3. The action of r_{ia} updates the goal buffer and prepares the buffer of the declarative module for a chunk retrieval. This action directly translates to the update (shown in blue) of the right edge in Figure 3.

The general translation of a rule from an ACT-R model uses the structure shown in Figure 3. A translation of the rule's precondition becomes the guard of the left edge and a translation of the rule's action becomes the update of the right edge, followed by a sequence of synchronisations to initiate behaviour of the modules referred to in the rule.

Procedural Module. Figure 4 shows the automaton that realises the behaviour of the procedural module which selects enabled rules for execution. As explained with the rule automaton above, there are two phases. A rule execution cycle starts in location `wait_delay` by sending on channel `conflict_resolution` on the downward edge. Rules whose preconditions are fulfilled receive, and update the shared array `enabled` accordingly. If at least one rule is enabled, the lower location is exited to the right. One enabled rule is selected non-deterministically and sending on the corresponding `production_fired` channel triggers the execution of the action of the selected rule. The shared array `enabled` is also reset on the right edge back to `wait_delay`. In location `wait_delay`, the invariant and the guard of the outgoing edge ensure that the next rule is executed at least `FIRING_DELAY` time units later. In case that no rule is enabled, the procedural module automaton waits for any module to change state, since only a change in the cognitive state makes it necessary to check again for an enabled rule (Anderson, 2009; Bothell, 2017a). The self loop on the location `wait_delay` ensures that the firing delay is observed if the cognitive state changes during the procedural module waiting for the next rule execution cycle to start.

The Rule Execution Cycle. Figure 5 shows everything put together. The network $\mathcal{N}^A$ modelling the (ACT-R model independent) architecture is the parallel composition $\mathcal{A}^P \| \mathcal{A}^D \| \cdots$ of all module automata (as discussed above). The network $\mathcal{N}^R$ representing the behaviour of the considered ACT-R model $R = \{r_1,...,r_n\}$ is the parallel composition $\mathcal{A}^{r_1} \| \cdots \| \mathcal{A}^{r_n} \| \mathcal{A}^{ch}_1 \| \cdots \| \mathcal{A}^{ch}_m$ of the rule automata for R with m chunk automata (memory size).

The cognitive architecture of ACT-R interprets ACT-R models by repetitive application of the following, 3-step rule execution cycle: 1.) wait a fixed time, 2.) check rules' preconditions on the current cognitive state to determine the set of enabled rules, and 3.) executing the action of an enabled rule (if any; otherwise wait for a change of the cognitive state). Modules may work during the waiting time in Step 1. The same execution cycle is directly visible in our TA-ACT-R model where Steps 1 to 3 are driven by automaton $\mathcal{A}^P$ and the steps are conducted in cooperation with the rule automata. The basic rule execution cycle is controlled by $\mathcal{A}^P$ (acting as sender on different channels), yet during its waiting time module automata may work concurrently.

Execution of the TA-ACT-R addition model would start by $\mathcal{A}^P$ waiting for the fixed time in location `wait_delay` (Step 1; cf. Figure 4) and then triggering each rule automaton to update their enabled flag (Step 2; cf. Figure 4 and 3). In our TA-ACT-R addition model, the shared variables are initialised such that the initialisation production rule (as shown in Fig. 3) is enabled. Hence $\mathcal{A}^P$ would then trigger this rule automaton (Step 3; cf. Figure 4 and 3). The rule automaton executes the actions of its rule (possibly in cooperation with module automata) while $\mathcal{A}^P$ is already back in location `wait_delay`, that models Step 1. The retrieval action started by the rule automaton is processed (including the retrieval delay) by the declarative module and

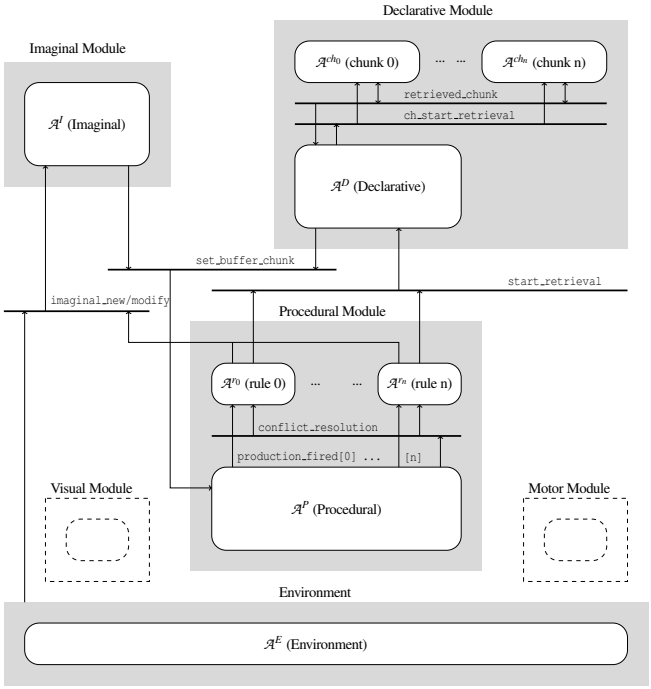


Figure 5: Structure of TA-ACT-R. Each white rectangle represents one timed automaton in the TA-ACT-R model. Arrows show potential synchronisation and are directed from senders to receivers and labelled with the channel name. The grey boxes group together those timed automata that together model an architecture module. Note that the environment does not directly interact with the imaginal module in ACT-R, but through, e.g., the visual module. For our experiments, we have abstracted from this indirection, a timed automaton model of, e.g., the visual module would be placed in the structure as shown by the dashed boxes.

chunk automata (cf. Figure 2 and 1). After completion of this retrieval, $\mathcal{A}^P$ is notified about the changed cognitive state and commences the next rule execution cycle (Step 1).

Formally, we observe sequences of timed automata configurations that are related by delay or synchronisation transitions. In these transition sequences of the TA-ACT-R network, we can clearly identify those configurations that correspond to situations *right before* starting a new rule execution cycle. In the more abstract F-ACT-R semantics, a rule execution cycle basically corresponds to one transition between two cognitive states, namely $\gamma_0 \xrightarrow{(r_1,50)} \gamma_1$ where $\gamma_0 = \{\text{goal} \mapsto (c_0,0), \text{retrieval} \mapsto (\bot,0)\}$ is the initial cognitive state and $\gamma_1 = \{\text{goal} \mapsto (c_0,0), \text{retrieval} \mapsto (c_2,50)\}$ is the cognitive state at the end of the rule execution cycle yet waiting for the retrieval action to complete.

The abstract, F-ACT-R computation paths of a given ACT-R model are hence refined by TA-ACT-R computation paths (one transition in the F-ACT-R model is related to a sequence of transitions in the TA-ACT-R model), which in turn is refined by computations of the ACT-R tool. In all three cases, we can clearly pinpoint the configuration right before the next rule execution and thus conclude from, e.g., an analysis of a TA-ACT-R model to the reachable cognitive states in the more abstract F-ACT-R view.

Discussion. Figures 1 to 4 show an abstract, comprehensible, readable and simulatable *model* of an ACT-R architecture. Using this architecture model, it becomes remarkably easy to evaluate ACT-R models under different architecture assumptions of a much wider range than the parameters of the ACT-R simulator allow. For example, other retrieval delays are obtained by redefining constants in $\mathcal{A}^D$ (cf. Fig.2); counting presentations (to support activation values) can be realised by increasing a counter in the successful case of a chunk automaton; unsuccessful retrieval of chunks in memory (sporadic forgetting) can be realised by removing the left edge from `idle` in the chunk automaton; etc.

By using a formal modelling language like timed automata, we obtain a precisely defined semantics. In contrast to a textual description of ACT-R's behaviour, it is unambiguously determined which delays or edges are possible in each model configuration. The Uppaal tool uses this fact to offer a convenient simulation environment that shows, in each configuration, the enabled edges and allows a user to choose the next one. If a model analysis finds a defect, the simulator can be used to inspect one computation path that exhibits the defect.

From these two aspects, we also envision a use of our TA-ACT-R models in teaching ACT-R: We see our model to fill a gap between a slide presentation of the concepts and principles of ACT-R and the ACT-R tool. Instructors could use the timed automata simulator in order to present the dynamic behaviour of the ACT-R architecture from rule selection to module activities before referring students to the ACT-R tool.

Evaluation

A highly relevant question on model analysis techniques and tools is about scalability. To be practically useful, a tool needs to be able to analyse ACT-R models that are used in cognitive science research.

Our investigation of the scalability of our TA-ACT-R-based approach to model analysis considers the following three research questions: (1) How does the number of chunks in the declarative memory affect the consumption of computational resources? (2) How does the length of the cognitive computation path affect the consumption of computational resources? (3) How does the number of rules in the ACT-R model affect the consumption of computational resources?

Addition Model. We have investigated the scalability of our approach using a parameterised ACT-R model of the addition task. Table 1a reports measurements of the classical addition model with four rules that we apply to a given number of count order chunks in declarative memory. The goal, that is, the number of count steps necessary to complete the addition, is fixed and thereby we isolate the effect on computational resource consumption to the number of chunks. The analysis checks that for each TA-ACT-R computation path, we finally observe the correct result in the goal buffer. Table 1a shows that the analysis of this parameterised addition model easily scales to 1,000 chunks considered for retrieval, while the length of the TA-ACT-R model compu-

Decl.	Time	States	Memory
25	0.09 s	165	8.1 MiB
50	0.17 s	165	8.9 MiB
100	0.29 s	165	10.7 MiB
500	1.20 s	165	24.5 MiB
1,000	2.70 s	165	40.8 MiB

(a) Computational resources used with increasing number of chunks in the declarative memory (Decl.) for fixed addend 9.

Decl.	Time	States	Memory
25	0.1 s	681	8.4 MiB
50	0.7 s	1,431	10.6 MiB
100	2.5 s	2,931	17.8 MiB
500	65.2 s	14,931	177.6 MiB
1,000	254.8 s	29,931	653.9 MiB

(b) Resource consumption with increasing number of chunks (Decl.), with highest possible addend (chunk number minus 1).

Proc.	Time	States	Memory
100	0.8 s	11,806	9.0 MiB
1,000	11.9 s	17,230	58.1 MiB

(c) Resource consumption with increasing number of production rules (Proc.), with highest possible addend.

Table 1: Evaluation results for time and memory consumption of an exhaustive analysis of addition models with `verifyta` 4.1.19 (Behrmann et al., 2004). Column 'States' gives the number of reachable configurations of the network of timed automata (cf. Preliminaries). The figures given above are averaged over ten runs (i7-6500/2.5 GHz, 8 GiB, Windows 10/64bit laptop).

tation path remains constant as expected from the fixed goal. Time consumption increases about linearly because each step of the analysis algorithm needs to check each chunk automaton for whether it offers a matching chunk; the reason for increased memory consumption is that the number of automata in the network uniformly increases the size of each TA-ACT-R configuration.

Table 1b reports measurements from the same model discussed above but with increasing addition goal. The model is supposed to apply the highest number of count steps possible with the given chunks, i.e. the instance with 1,000 chunks is supposed to conduct 999 count steps. The time needed for the analysis in the table scales roughly linearly in both, number of chunks and length of computation; the numbers of reachable TA-ACT-R configurations in the table grow linearly in the length of the computation. Table 1b shows that an exhaustive analysis of the model with a few hundred chunks takes not much more than a minute. With an analysis time in this low order of magnitude, we anticipate that our TA-ACT-R analysis can be effectively used during the process of cognitive modelling, that is, to analyse an ACT-R model for common errors, and, in case errors are found, to fix these errors and re-run the analysis. For large chunk numbers and computation lengths, the time needed to complete the analysis becomes more noticeable. We suggest to value the computation time wrt. the obtained outcome: After (in case of the addition model) about 4 minutes, *all possible computations* of the cognitive model have been considered.

Table 1c reports measurements from a different addition model where each count fact is modelled as its own production rule. That is, in order to, e.g., do 100 count steps, there are 100 different rules. Table 1c shows that the analysis of this parameterised addition model easily scales to 1,000 rules.

Preferred Mental Model Theory. To evaluate the performance of our TA-ACT-R-based approach on a cognitive model from the research literature, we have considered the PMMT[1] model that has been used in (Langenfeld et al., 2018) to illustrate the usefulness of checking models for the absence of deadlocks (a deadlock is a cognitive state where no production rule is able to fire while the end of the modelled behaviour has not been reached).

The considered ACT-R model of the PMMT is technically non-trivial as it makes use of multiple modules (often in the same rule) and depends on complex preconditions including buffer requests and module queries. From its design parameters (about 40 production rules, less than 10 learned chunks), we would have expected an exhaustive analysis of the computational space of the TA-ACT-R model to take at most one second considering the figures in Table 1. In fact, the analysis was much faster: The analysis tool `verifyta` (Behrmann et al., 2004) reported the absence of deadlocks for every possible combination of two premises and a conclusion within 146*ms* (storing 701 TA-ACT-R states in 8.7 MiB of memory). Thus there are complex ACT-R research models that can be very efficiently analysed for the absence of model defects (Langenfeld et al., 2018).

Conclusion and Future Work

As future work we will automate the translation of the production rules of an ACT-R model to the according automata to enable the analysis of models without manual translation. We will also extend TA-ACT-R by hybrid processes like chunk activation and retrieval delays. We will also integrate hybrid processes (e.g. calculation of chunk activation and retrieval delays) into TA-ACT-R to replace the non-deterministic sub symbolic layer for more precise analysis of ACT-R models.

In this article we investigated the potential for automatic defect analysis of ACT-R models. We developed a formal but easy to comprehend model of the ACT-R architecture and a translation scheme for ACT-R models. Benchmark results show, that the analysis of useful properties scales well for high numbers of chunks and production rules so that it can be applied during model development.

Acknowledgements

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) PO 279/2-1.

[1]The preferred mental model theory (PMMT; Ragni, Knauff, & Nebel, 2005; Ragni & Knauff, 2013) is the most recent refinement of the established mental model theory (MMT; Johnson-Laird, 1980), that aims to explain human spatial reasoning.

References

Albrecht, R. (2013). *Towards a Formal Description of the ACT-R Unified Theory of Cognition*. Unpublished master's thesis, Albert-Ludwigs-Universität Freiburg.

Albrecht, R., & Westphal, B. (2014a). Analysing Psychological Theories with F-ACT-R. *Cogn. Processing*, *15*, 77–79.

Albrecht, R., & Westphal, B. (2014b). F-ACT-R: Defining the Architectural Space. *Cogn. Processing*, *15*, 79-81.

Alur, R., & Dill, D. L. (1994). A theory of timed automata. *Theoretical Computer Science*, *126*(2), 183–235.

Anderson, J. R. (1983). *The Architecture of Cognition*. Psychology Press.

Anderson, J. R. (2009). *How can the human mind occur in the physical universe?* Oxford University Press.

Behrmann, G., David, A., & Larsen, K. G. (2004). *A Tutorial on Uppaal*. In *SFM* (Vol. 3185, p. 200-236). Springer.

Bothell, D. (2017a). *ACT-R 7 reference manual.* Retrieved from `http://act-r.psy.cmu.edu/actr7`

Bothell, D. (2017b). *ACT-R Tutorial.* Retrieved from `http://act-r.psy.cmu.edu/actr7/.`

Gall, D., & Frühwirth, T. (2018). An operational semantics for the cognitive architecture ACT-R and its translation to constraint handling rules. *ACM TCL*, *19*(3), 22:1–22:42.

Gall, D., & Frühwirth, T. W. (2014). A formal semantics for the cognitive architecture ACT-R. In *LOPSTR* (Vol. 8981, pp. 74–91). Springer.

Gall, D., & Frühwirth, T. W. (2017). A decidable confluence test for cognitive models in ACT-R. In *RuleML+RR* (Vol. 10364, pp. 119–134). Springer.

Johnson-Laird, P. (1980). Mental models in cognitive science. *Cognitive Science*, *4*, 71–115.

Langenfeld, V., Westphal, B., Albrecht, R., & Podelski, A. (2018). But does it really do that? Using formal analysis to ensure desirable ACT-R model behaviour. In *CogSci 2018* (pp. 659–664).

Ragni, M., & Knauff, M. (2013). A theory and a computational model of spatial reasoning with preferred mental models. *Psychological review*(3), 561–588.

Ragni, M., Knauff, M., & Nebel, B. (2005). A Computational Model for Spatial Reasoning with Mental Models. In *Proc. of the 27th annual Cog. Sci. Conf.* (pp. 1064–1070).

Ragni, M., Sauerwald, K., Bock, T., Kern-Isberner, G., Friemann, P., & Beierle, C. (2018). Towards a formal foundation of cognitive architectures. In *CogSci 2018* (pp. 2321–2326).

Without Conceptual Information Children Miss the Boat: Examining the Role of Explanations and Anomalous Evidence in Scientific Belief Revision

Nicole E. Larsen[†] (nicole.larsen@mail.utoronto.ca), Vaunam P. Venkadasalam[†] (vaunam.venkadasalam@mail.utoronto.ca) & Patricia A. Ganea (patricia.ganea@utoronto.ca)
Department of Applied Psychology & Human Development,
University of Toronto, 252 Bloor St West, Toronto, Ontario, M5S 1V6
[†]*Authors contributed equally.*

Abstract

In this study we investigated the role of conceptually rich explanations and anomalous evidence in children's scientific belief revision. We also explored whether the order in which children experience these two learning opportunities influences their belief revision ability. Five-year-old children were assigned to one of two conditions, where they either first received conceptual explanations about buoyancy and then observed anomalous data in a guided activity (*Explanation-First*), or the reverse (*Anomalies-First*). Results showed that (1) conceptually rich explanations lead to more accurate predictions about which objects sink and which float than anomalous data presentation, and (2) when explanations and anomalous data were combined, children's correct predictions increased significantly from pre-test to post-test when they received the conceptual information before the anomalous evidence (*Explanation-First*), but not in the opposite order condition (*Anomalies-First*). These results suggest that children are more likely to maintain their misconceptions when exposed to anomalies without prior instruction involving conceptually rich explanations.

Keywords: cognitive development; belief revision; scientific reasoning

Supporting Scientific Belief Revision

Scientific beliefs have their foundations in early development. Much of children's early science learning is informal, and the intuitive theories they build through daily observation and cultural learning are frequently at odds with accurate scientific theories (Kuhn, 1989; Vosniadou & Brewer, 1992). Children's naïve misconceptions are often resistant to change (Vosniadou, 2002) and some persist into adulthood (Coley & Tanner, 2012; Pine, Messer, & St. John, 2001; Shtulman & Valcarcel, 2012). Conceptual change is the process of restructuring naive theories to include counter-intuitive concepts, which for some scientific domains can be a lengthy and arduous process (Vosniadou, 2013).

The process of early scientific reasoning has been compared to formal scientific theory change, in which children get to formulate, test, and revise hypotheses based on evidence and observations (Gopnik, 2012; Gopnik & Wellman, 2012). As part of this process, experiencing anomalous evidence that contradicts existing naïve theories is an important driver of belief revision. For example, some existing research suggests that, depending on which type of anomalous data they observe, preschool children either explain away or change their naïve theories about how objects balance (Bonawitz, Van Schijndel, Friel, & Schulz, 2012).

However, anomalous data may not always be sufficient for facilitating belief revision. Research about causal systems has indicated that when children are shown novel causal systems, they can theorize about the causal relation in these systems, and are subsequently resistant to changing these theories, even when immediately presented with new anomalous data (Schauble, 1990; Schulz, Goodman, Tenenbaum, & Jenkins, 2008). This is compounded by the fact that although providing children with anomalous data presents an opportunity for belief revision, children often make errors during the observation, interpretation, generalization, or retention stages of science activities when they encounter anomalous evidence (Chinn & Malhotra, 2002).

In the case of existing misconceptions, children's tendency to hold onto naïve theories may be even more pronounced as these theories are more entrenched. Children's difficulty in making inferences from evidence that is in conflict with their naïve theories may result from the absence of a viable alternative theory (Chinn & Brewer, 1993). If children are provided with alternative explanations, they may be better equipped to interpret the anomalous data they encounter and as a result be more likely to engage in belief revision. Thus, combining anomalous evidence with correct conceptual explanations may be particularly effective for belief revision and science learning more generally (Koslowski, 1996).

Current Study

The goal of the current study was twofold. First, we examined the role of conceptually rich explanations and anomalous evidence in children's ability to revise an existing naïve scientific belief. Second, we explored whether the order in which children experience these two learning opportunities influences their belief revision ability. Five-year-old children were provided with conceptually rich information about buoyancy (during a brief picture book reading session) either before or after they had the opportunity to observe anomalous examples (i.e., heavy objects floating) in a guided play activity. We selected to deliver the conceptually rich explanations in a picture book format not only because picture-book reading is an

enjoyable activity for many young children, but also because research has shown that young children can learn scientific information from picture books, even in cases where they hold misconceptions (Kelemen, Emmons, Seston Schillaci, & Ganea, 2014; Venkadasalam & Ganea, 2018). We presented anomalous evidence through a guided activity, which allowed for active engagement with real, physical objects (Nayfeld, Brenneman, & Gelman, 2011; Peterson & French, 2008). Here, we were interested in how children interpret and generalize from real-life anomalous evidence. In guided activities, adults plan an activity with a learning goal, and scaffold this learning, which allows children to maintain an active role in the process (Weisberg, Hirsh-Pasek, & Golinkoff, 2013). Thus, guided activities provided children with hands on opportunities to interact with anomalous evidence in an engaging way, but also ensured that they were able to produce such evidence with guidance.

We examined belief revision in children's acquisition of a physical science concept (buoyancy), a concept with common misconceptions. Buoyancy is the upward force on objects in a liquid. An object floats if the buoyant force is equal to the force of gravity, and an object sinks when the gravitational force is stronger. Sinking and floating are concepts taught throughout science education (Kallery, 2015; Selley, 1993) and ones for which children often hold misconceptions (Hardy, Jonen, Möller, & Stern, 2006; Yue, Tomita, & Shavelson, 2008). One difficulty young children have is that they often conflate density with weight (Wilkening & Cacchione, 2011), which is problematic when children have to compare the relative densities of the objects and water, often leaving them with the misconception that heavy objects sink and light objects float (Lehrer, Schauble, Strom, & Pligge, 2011; Smith, Carey, & Wiser, 1985). When 5-year-old children notice anomalies to their intuitive theories ("heavy objects sink and light objects float") they sometimes hypothesize about the material of the objects (e.g., wooden objects float). A focus on material is a promising step in children's ability to think about density because some materials are less dense than others and therefore sink at different rates. However, to fully understand what makes objects sink or float, children also have to consider how the mass is distributed and therefore take into account the shape of the object as well. Here we explore the effect of pairing anomalies with conceptually rich explanations to promote children's ability to dissociate the objects' behavior in water from their weight and recognize the role of air-filled cavities and surface tension in explaining why objects sink or float.

The study was designed using a pre-, mid-, and post-test to measure children's belief revision. In each test phase, we examined differences in children's predictions of whether objects would sink or float as a function of the order of instructional methods used (conceptual information or anomalous data). Children's predictions were chosen as an implicit measure of learning. The pre-test allowed us to control for children's previous knowledge. The mid-test allowed us to determine the role of each learning opportunity (explanations or anomalous evidence) on children's belief revision in isolation. Finally, the post-test was used to determine whether the order in which children received the two learning opportunities mattered when they were combined.

We expected that, when compared to pre-test scores, children's predictions at mid-test about which objects float or sink would be significantly higher in the *Explanation-First* condition but not the *Anomalies-First* condition. There is previous research showing that children are able to learn scientific information from conceptually rich explanations (Kelemen et al., 2014; Venkadasalam & Ganea, 2018) and we expected to find the same type of evidence here. However, given the existing research with adults on the use of anomalous evidence indicating that individuals often make errors in the interpretation and generalization of this evidence (Chinn & Malhotra, 2002), we expected that the exposure to anomalies alone will not lead to a change in children's misconceptions.

With respect to the order in which children receive the conceptually rich information and the anomalous evidence, we considered the possibility that children who received the anomalous data first may make comparable gains at post-test after receiving the conceptual information. However, although possible this is not very likely, because without an alternative theory to explain the anomalies, children could appeal to extraneous variables to fit the anomalous evidence into their naïve theory, therefore strengthening it. As a result, the hypothesized difference at mid-test would remain significant at post-test, even after exposure to an alternative theory. The alternative order of presentation (explanations followed by anomalies) might be more effective, because children could rely on the conceptual information provided to interpret the anomalous evidence.

Methods

Participants

Ninety-six 5-year-old children (M = 5.49; range: 5.03- 5.99, 48 males) participated in this study. Equal numbers of children were randomly assigned to one of two conditions: *Explanation-First* (n = 48, M_{age} = 5.50, 24 males, 24 females), and *Anomalies-First* (n = 48, M_{age} = 5.49, 24 males, 24 females). Within these conditions, children were read one book and completed one activity. We developed two books and two guided activities to teach children about buoyancy. This was done to ensure that differences in learning did not arise from the type of book the child read or the activity the child completed. All 16 combinations of the books and activities were included and were counterbalanced, such that 6 children received each possible combination. No differences between the two types of books and activities were expected.

Procedure

There were five phases in this study: a pre-test, a learning phase 1 (depending on the condition, *Explanation-First* or *Anomalies-First*), mid-test phase, learning phase 2 (Explanation/Anomalies, depending on the condition), and the post-test. The session was video-recorded and lasted 40 minutes to 1 hour.

Test Phase. To measure children's belief revision a pre-, mid- and post-test were administered. The procedure for the pre-, mid-, and post-test was identical. The materials for each of the 3 test phases included 4 pairs of objects, for a total of 12 objects pairs. Within each set of 4 pairs, two pairs of objects were the same weight, and two were different weights. Two of the pairs of objects were made of the same material and two were made of different materials. Materials included: metal, plastic, rubber, and glass. For each test phase, children received a different object set, but the order in which children received these sets was counterbalanced across the pre-, mid- and post-tests.

Children were given objects in pairs to inspect. The experimenter told children what each object was made of so there was no ambiguity. Children were then provided with a scale and prompted to weigh the object pairs so they could definitively identify which object was heavier and which was lighter. To avoid differences in response patterns within the sample, within each object set the pairs of objects were presented in the same order to each participant: different weight/different material, same weight/same material, same weight/different material, and different weight/same material.

After children were given time to inspect and feel the objects, weigh them, and were told what they were made of, children were asked the test question: "If I took these two objects and put them into the water, which one would float on the top and which one would sink to the bottom?". Their predictions were recorded. Children received neutral feedback ("Thank you") after answering each question.

Learning Phase. In this phase, we used picture books to deliver the conceptually rich explanations and guided activities to present children with the anomalous evidence. We developed two books and two guided activities to ensure that differences in learning did not arise from the type of book or activity used. For the picture books, we created an informational, non-fiction book, and a narrative, fiction book which contained the same conceptual information about buoyancy. Given previous work reporting no differences in children's learning based on book genre, we did not expect to find differences between these two book types (Venkadasalam & Ganea, 2018).

In the first guided activity, Activity One, children made predictions about whether 12 different objects would sink or float. The objects in this activity varied in weight and material. Children then tested these objects in water to see if their predictions were correct. The second activity, Activity Two, involved children manipulating a piece of clay into shapes that either floated or sank. Children then tested these shapes in water, demonstrating that an object with a constant weight can both sink and float. No differences were expected between activities as children were guided through each activity to ensure the production and observation of anomalous evidence and both activities were designed to demonstrate the same type of anomalies.

Children in the *Explanation-First* condition were read the book prior to the activity, whereas children in the *Anomalies-First* condition were read the book following the activity. During the book reading, the experimenter read either the non-fiction or the fiction book to each child aloud. In the activity, the experimenter guided children through different instances where they could compare objects sinking and floating, either with the 12 objects in Activity One, or the pieces of clay in Activity Two. The books and activities were structured to be analogous in terms of their content. The goal of both learning phases was explicitly identified as teaching children about why objects sink or float. However, no mention of the book was made during the activity, and likewise no mention of the activity was made during the book.

Coding

Children's predictions for which object would sink and which one would float in each pair were scored. Children who correctly identified which object in the pair would sink and which would float received a score of 1. A score of 0 was assigned if children incorrectly identified the sinker and the floater in the object pair or if they said both objects would sink or both objects would float. Two research assistants coded 100% of the children's responses from the video recordings. The coders were blind to the hypotheses of the study, the condition and test phase. There was high interrater reliability determined by Cohen's $\kappa = .91$, $p < .001$, a 95.66% agreement rate. The coders resolved disagreements through discussion.

Results

In preliminary analyses, we ensured there were no differences between the scores at mid- and post-test as a result of the two types of books and activities used. A Mann Whitney U-test found that scores for mid- and post-test were similar for both books (*ps* > .84), and both activities (*ps* > .12). As there were no significant differences between the type of books and activities used in the intervention, these factors were collapsed in the following main analyses. We also examined differences between children's knowledge across the two conditions at pre-test. A Mann Whitney U-test found that the pre-test scores were similar across conditions at baseline, $U = 1017$, $z = -1.02$, $p = .31$ with a mean rank pre-test score of 45.69 for the *Explanation-First* condition and 51.31 for the *Anomalies-First* condition. Additionally, Wilcoxon Signed-ranks tests revealed that the pre-test scores were significantly lower than chance responding, indicating that children held misconceptions at pre-test (*Explanation-First*: $Z = -3.35$, $p =$

.001; *Anomalies-First:* $Z = -2.32$, $p = .021$). Table 1 displays the proportion of correct responses across the three test phases for both conditions.

A generalized estimating equation (GEE) analysis with multinomial distributions and cumulative logit link functions was conducted to investigate whether children correctly predicted which object would sink and which would float. This type of analysis was selected to accommodate the ordinal dependent variable and the presence of a within-subject factor (pre-, mid- and post-test scores) in the data.

Table 1: Percent Correct Responses Across Test Phases by Condition.

Score	Test Phase					
	Pre-Test		Mid-Test		Post-Test	
	Anom-First	Expl-First	Anom-First	Expl-First	Anom-First	Expl-First
0/4	27%	40%	27%	19%	25%	8%
1/4	19%	19%	21%	0%	15%	15%
2/4	33%	19%	27%	27%	25%	8%
3/4	13%	17%	15%	27%	21%	38%
4/4	8%	6%	10%	27%	15%	31%

Notes. Anom-First stands for the *Anomalies-First* condition. Expl-First stands for the *Explanation-First* condition. The percentages are calculated out of 48 total responses for each condition per test phase.

There was no effect of condition, ($p = .31$), nor a difference between pre- and mid-test ($p = .80$), nor pre- and post-test ($p = .10$). However, there was a significant interaction between condition and test phase. From pre- to mid-test children in the *Explanation-First* condition were more likely to answer more test questions correctly, Wald $\chi 2(1) = 19.87$, $p < .001$, $b = 1.51$, $SE = .34$, compared to the *Anomalies-First* condition. Children in the *Explanation-First* condition (Exp(B) = 4.51, 95% CI = [2.33, 8.75]) were approximately four and a half times more likely to answer the test questions correctly at mid-test in comparison to the *Anomalies-First* condition.

Additionally, from pre- to post-test children in the *Explanation-First* condition were more likely to answer more test questions correctly, Wald $\chi 2(1) = 14.66$, $p < .001$, $b = 1.53$, $SE = .40$, compared to the *Anomalies-First* condition. Children in the *Explanation-First* condition (Exp(B) = 4.62, 95% CI = [2.11, 10.09]) were approximately four and a half times more likely to answer the test questions correctly at post-test in comparison to the *Anomalies-First* condition.

Post-hoc Wilcoxon Signed Rank Tests were conducted using a Bonferroni correction to account for multiple comparison (alpha = .008). There was a significant increase in children's score in the *Explanation-First* condition between pre- and mid-test ($z = 4.45$, $p < .001$) and pre- and post-test ($z = 4.86$, $p < .001$), but not between mid- and post-test ($z = 1.63$, $p = .10$). In the *Anomalies-First* condition there was no significant increase in scores for any of the test phases ($ps > .13$); see Figure 1.

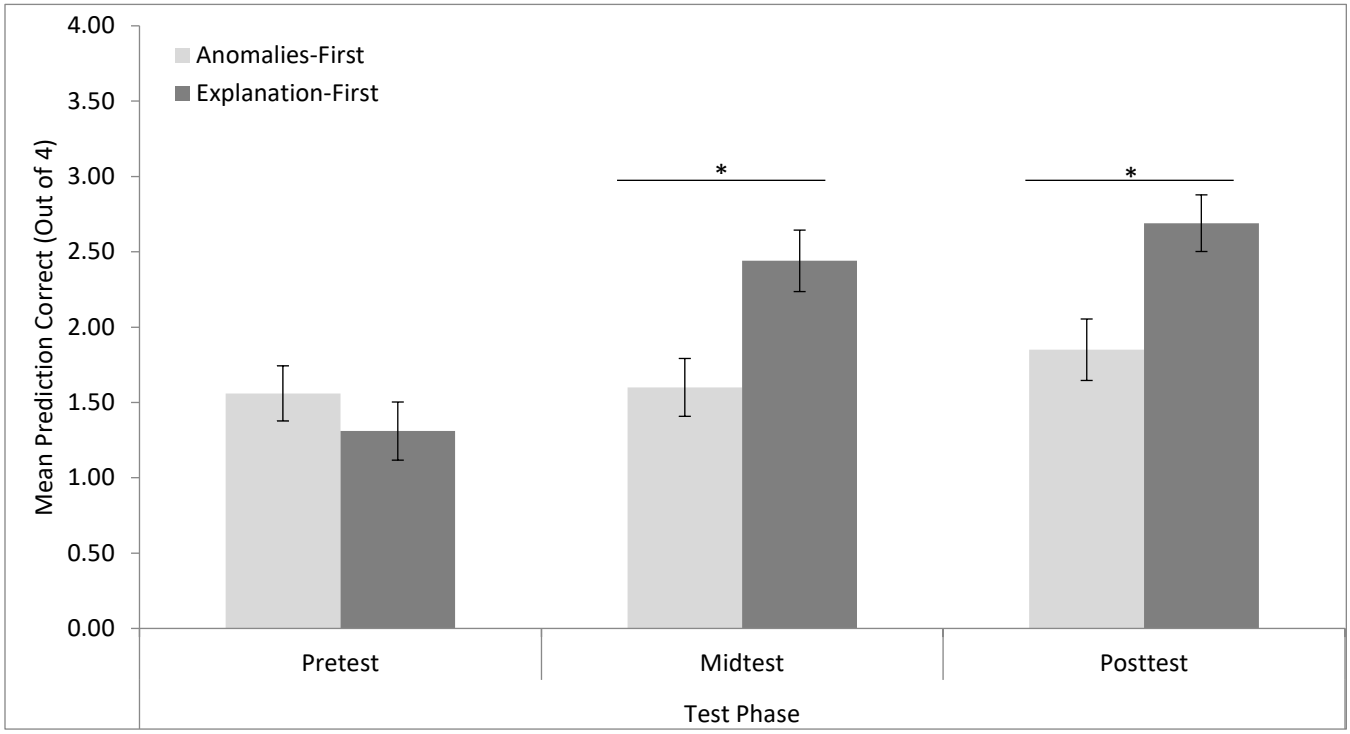


Figure 1: Mean Predictions Correct across Test-Phase by Condition

Discussion

This study investigated the role of conceptually rich explanations and anomalous evidence in children's revision of physical science misconceptions. We first found that children can revise their belief about what makes objects float or sink when provided only with a conceptual explanation but not when only witnessing anomalies. Compared to pre-test scores, children in the *Explanation-First* condition made significantly better predictions at mid-test after they received the conceptual information only. However, the children in the *Anomalies-First* condition did not make significantly better predictions after observing anomalies. This indicates that children maintain their misconceptions when they have exposure to anomalies.

We also found that the order of instructional methods affects children's belief revision in the context of a physical science concept. Children performed better at post-test when presented with rich conceptual explanations *prior to* observing anomalous evidence. The *Explanation-First* condition facilitated greater revision of beliefs than the *Anomalies-First* condition. Of note, even in the *Explanation-First* condition, when children observed anomalous evidence after they received the explanations, the anomalies did not lead to any significant changes. In contrast, results from the *Anomalies-First* condition showed that observing the anomalies first subsequently interfered with children's ability to incorporate and apply the conceptual information they received from the picture book.

Together these findings provide evidence that using anomalous data to promote belief revision can be challenging, as children are biased to rely on their own theories, and resistant to setting aside this prior knowledge when confronted with counter-evidence (Chinn & Brewer, 1993; Kuhn et al., 1988). Despite observing counterexamples, children may ignore them, or even find a way to fit the anomalies within their existing theoretical framework, thereby strengthening their naïve misconceptions. However, when children have access to a viable, alternative explanatory framework, they can then activate this alternative theory to interpret the anomalous evidence. Thus, the present findings indicate that supplementing prior beliefs with an alternative conceptual explanation before anomalous evidence is observed may be particularly effective for promoting knowledge revision.

Further work is needed to determine if the addition of anomalous evidence affects retention after a delay. That is, while we found no positive effects of the anomalous evidence above and beyond what the explanations provided, observing anomalous evidence after receiving the correct explanation, may lead to greater retention of the new theory than receiving only the explanation alone.

Another consideration for future work is that explicit connections were not made between the book and the activity. While these learning phases were built to be highly analogous, and the same content goal was verbally specified for both, no explicit connection was made between them. It is possible that with an explicit connection between the two, children may achieve higher performance across conditions.

Additionally, in the current study the explanations were presented through a picture book. It is an open question whether results would be similar if both the explanations and anomalies are presented in a similar manner. Currently, we are exploring whether pairing live anomalous evidence with verbal explanations will have positive effects on learning. Further work should also explore the applicability of these findings to different scientific concepts. Sinking and floating are complex physical concepts, particularly for young children to grasp. It is possible that presenting children with anomalies only or first may be equally effective for belief revision for simpler concepts.

The current results can inform our theories about the process of conceptual change and optimal science instruction. This study demonstrates the importance of critically examining not just *what* we teach children, but *how* we teach them, and in particular the order in which instruction is delivered. Presenting children with content information before they observe anomalous data prevents children from fitting anomalies into their naïve schema, giving them an alternative viewpoint from which to interpret this evidence. Therefore, this order of presentation better facilitates the revision of children's misconceptions. Providing children with comprehensive explanations of phenomena they observe is a promising educational technique to improve their scientific reasoning and literacy.

Acknowledgments

We are grateful to the children who participated in this research and we thank the families who made this possible. We would like to thank Angela Nyhout, Myrto Grigoroglou and Begum Ozdemir for feedback on a previous draft of this paper. This research was supported by a SSHRC Insight grant to Patricia Ganea.

References

Bonawitz, E. B., Van Schijndel, T. J. P., Friel, D., & Schulz, L. (2012). Children balance theories and evidence in exploration, explanation, and learning. *Cognitive Psychology*, *64*, 215–234. https://doi.org/10.1016/j.cogpsych.2011.12.002

Chinn, C. A., & Brewer, W. F. (1993). The Role of Anomalous Data in Knowledge Acquisition: A Theoretical Framework and Implications for Science Instruction. *Review of Educational Research*, *63*(1), 1–49. https://doi.org/10.3102/00346543063001001

Chinn, C. A., & Malhotra, B. A. (2002). Children's responses to anomalous scientific data: How is conceptual change impeded? *Journal of Educational Psychology*, *94*(2), 327–343. https://doi.org/10.1037/0022-0663.94.2.327

Coley, J. D., & Tanner, K. D. (2012). Feature Approaches to Biology Teaching and Learning Common Origins of Diverse Misconceptions: Cognitive Principles and the Development of Biology Thinking. *Life Sciences Education*, *11*, 209–215.

https://doi.org/10.1187/cbe.12-06-0074

Gopnik, A. (2012). Scientific Thinking in Young Children: Theoretical Advances, Empirical Research, and Policy Implications. *Science*, *337*(6102), 1623–1627.

Gopnik, A., & Wellman, H. M. (2012). Reconstructing constructivism: Causal models, Bayesian learning mechanisms, and the theory theory. *Psychological Bulletin*, *138*(6), 1085–1108. https://doi.org/10.1037/a0028044

Hardy, I., Jonen, A., Möller, K., & Stern, E. (2006). Effects of instructional support within constructivist learning environments for elementary school students' understanding of "floating and sinking." *Journal of Educational Psychology*. https://doi.org/10.1037/0022-0663.98.2.307

Kallery, M. (2015). Science in early years education: introducing floating and sinking as a property of matter. *International Journal of Early Years Education*, *23*(1), 31–53. https://doi.org/10.1080/09669760.2014.999646

Kelemen, D., Emmons, N. A., Seston Schillaci, R., & Ganea, P. A. (2014). Young Children Can Be Taught Basic Natural Selection Using a Picture-Storybook Intervention. *Psychological Science*, *25*(4), 893–902. https://doi.org/10.1177/0956797613516009

Koslowski, B. (1996). *Learning, development, and conceptual change. Theory and evidence: The development of scientific reasoning.* Cambridge, MA, US: The MIT Press.

Kuhn, D. (1989). Children and Adults as Intuitive Scientists. *Psychological Review*, *96*(4), 674–689. https://doi.org/10.1037/0033-295X.96.4.674

Kuhn, D., Amsel, E., O'Loughlin, M., Schauble, L., Leadbeater, B., & Yotive, W. (1988). *The development of scientific thinking skills. The development of scientific thinking skills.* San Diego, CA, US: Academic Press.

Lehrer, R., Schauble, L., Strom, D., & Pligge, M. (2011). Similarity of form and substance: Modeling material kind. In S. M. Carver & D. Klahr (Eds.), *Cognition and instruction : twenty-five years of progress* (pp. 39–74). Psychology Press.

Nayfeld, I., Brenneman, K., & Gelman, R. (2011). Science in the Classroom: Finding a Balance Between Autonomous Exploration and Teacher-Led Instruction in Preschool Settings. *Early Education and Development*. https://doi.org/10.1080/10409289.2010.507496

Peterson, S. M., & French, L. (2008). Supporting young children's explanations through inquiry science in preschool. *Early Childhood Research Quarterly*. https://doi.org/10.1016/j.ecresq.2008.01.003

Pine, K., Messer, D., & St. John, K. (2001). Children's Misconceptions in Primary Science: A Survey of teachers' views. *Research in Science & Technological Education*, *19*(1), 79–96. https://doi.org/10.1080/02635140120046240

Schauble, L. (1990). Belief revision in children: The role of prior knowledge and strategies for generating evidence. *Journal of Experimental Child Psychology*, *49*(1), 31–57. https://doi.org/10.1016/0022-0965(90)90048-D

Schulz, L. E., Goodman, N. D., Tenenbaum, J. B., & Jenkins, A. C. (2008). Going beyond the evidence: Abstract laws and preschoolers' responses to anomalous data. *Cognition*, *109*(2), 211–223. https://doi.org/10.1016/j.cognition.2008.07.017

Selley, N. (1993). Why do things float? A study of the place for alternative models in school science. *School Science*, *74*(269), 55–61.

Shtulman, A., & Valcarcel, J. (2012). Scientific knowledge suppresses but does not supplant earlier intuitions. *Cognition*, *124*(2), 209–215. https://doi.org/10.1016/j.cognition.2012.04.005

Smith, C., Carey, S., & Wiser, M. (1985). On differentiation: A case study of the development of the concepts of size, weight, and density. *Cognition*, *21*(3), 177–237. https://doi.org/10.1016/0010-0277(85)90025-3

Venkadasalam, V. P., & Ganea, P. A. (2018). Do objects of different weight fall at the same time? Updating naive beliefs about free-falling objects from fictional and informational books in young children. *Journal of Cognition and Development*, *19*(2), 165–181. https://doi.org/10.1080/15248372.2018.1436058

Vosniadou, S. (2002). On the Nature of Naïve Physics. In M. Limon & L. Mason (Eds.), *Reconsidering Conceptual Change: Issues in Theory and Practice.* (pp. 61–76). Dordrecht: Springer.

Vosniadou, S. (2013). Conceptual Change in Learning and Instruction. In S. Vosniadou (Ed.), *International Handbook of Research on Conceptual Change* (p. 768). New York, NY: Routledge.

Vosniadou, S., & Brewer, W. F. (1992). Mental models of the Earth: A study of conceptual change in childhood accepted information that the earth is a sphere. *Cognitive Psychology*, *585*(24), 535–585. https://doi.org/10.1016/0010-0285(92)90018-W

Weisberg, D. S., Hirsh-Pasek, K., & Golinkoff, R. M. (2013). Guided Play: Where Curricular Goals Meet a Playful Pedagogy. *Mind, Brain, and Education*, *7*(2), 104–112. https://doi.org/10.1111/mbe.12015

Wilkening, F., & Cacchione, T. (2011). Children's intuitive physics. In U. C. Goswami (Ed.), *The Wiley-Blackwell handbook of childhood cognitive development* (pp. 473–496). Maiden: Wiley-Blackwell.

Yue, Y., Tomita, M. K., & Shavelson, R. J. (2008). Diagnosing and Dealing with Student Misconceptions: Floating and Sinking. *Science Scope*, *34*(1), 34–39.

Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. *Developmental Review*, *27*(2), 172–223. https://doi.org/10.1016/j.dr.2006.12.001

Children Learn Words Better in Low Entropy

Ori Lavi-Rotbain (oriedit.lavi@mail.huji.ac.il)
Edmond & Lily Safra Center for Brain Sciences, Edmond J. Safra Campus,
The Hebrew University Jerusalem, 9190401, Israel

Inbal Arnon (inbal.arnon@mail.huji.ac.il)
Department of Psychology, The Hebrew University of Jerusalem,
Mount Scopus, Jerusalem 91905, Israel

Abstract

During their first year, infants learn to name objects. To do so, they need to segment speech, extract the label and map it to the correct referent. While children successfully do so in the wild, previous results suggest they struggle to simultaneously learn segmentation and object-label pairings in the lab. Here, we ask if some of children's difficulty is related to the uniform distribution they were exposed to, since it differs from that of natural language, and has high entropy (making it less predictable). Will a low entropy distribution facilitate children's performance in these two tasks? We looked at children's (mean age=10;4 years) simultaneous segmentation and object-label mapping of words in an artificial language task. Low entropy (created by making one word more frequent) facilitated children's performance in both tasks. We discuss the importance of using more ecologic stimuli in the lab, specifically- distributions with lower entropy.

Keywords: Statistical learning; Multi-modal cues; Word segmentation; Word learning; Entropy; Children.

Introduction

During the first year of life, infants make their initial steps in learning language. One ability they acquire is naming objects. To do so, infants need to extract the segmented labels and map them onto the correct object. While infants learn some object-label mappings early on (Bergelson & Swingley, 2012), even older children seem to struggle with simultaneously learning segmentation and object-label pairings in the lab. Previous work examined children's ability to perform both tasks at the same time in a statistical learning paradigm (Lavi-Rotbain & Arnon, 2017). Children were exposed to an unsegmented speech stream where transitional probabilities served as a cue for word boundary (as in Saffran, Aslin, & Newport, 1996). The language had an additional visual cue to segmentation: each word was matched to an image of an object that appeared for the duration of the word (e.g., 'dukame' → blue star). The prediction was that the visual cue will assist segmentation and allow children to learn the object-label pairings, illustrating their ability to integrate multimodal cues. As predicted, the results showed that after a short exposure (under two minutes) 10;6-year-olds managed to learn both aspects (segmentation and object-label mapping). However, while children showed some learning of the object-label pairing (they were above chance, M=34.4%, chance=25%), their learning was relatively poor. Younger children (mean age: 7;8 years) did not learn the pairings at all (M=25.96%%, chance=25%), even though they are clearly capable of relating labels to objects in natural language. Why then do children struggle with this task in the lab? And what can we learn from their difficulty about the factors that impact children's language learning? Here, we ask how the distributional properties of the language may have impeded learning. In particular, we focus on the use of a uniform distribution – where all items are equally frequent. This was the distribution used in the previous study, and one that is used in most statistical learning studies.

A uniform distribution of stimuli, where every element (e.g. word) is presented the same number of times, differs from what is found in natural language. Words in natural language have a Zipfian distribution (Zipf, 1936) with few very frequent words, and most words having low frequency. The Zipfian distribution is a highly skewed distribution, with a narrowed peak for the small number of very frequent words, and a long tail for the rest of the words. Words show a Zipfian distribution across many languages, in both adult-to-adult speech (Zipf, 1936; Piantadosi, 2014) and child directed speech (Hendrickson & Perfors, 2019; Lavi-Rotbain & Arnon, submitted). Other aspects of language, like grammatical categories, also show a Zipfian distribution (Piantadosi, 2014; Lavi-Rotbain & Arnon, submitted). Interestingly, the objects that infants see also show a Zipfian distribution (Clerkin, Hart, Rehg, Yu, & Smith, 2017). That is, using a uniform distribution does not accurately reflect the distribution of words (or objects) that children are exposed to.

Moreover, uniform distributions are also less predictable than non-uniform distributions. One way to quantify the difference between them is to use Shannon's Entropy (Shannon, 1948). Entropy quantifies how unpredictable a distribution is as a whole, with higher entropy assigned to less predictable distributions. The uniform distribution is the least predictable - it is hard to guess which word will appear next when they all have equal probabilities - and consequently has high entropy. Non-uniform distributions, such as Zipfian distributions, are more predictable, and have lower entropies: it is easier to guess the next word when only a few are highly probable.

Here, we ask if children's simultaneous learning of segmentation and object-label pairings will be facilitated when using a distribution with low unigram entropy. Such a finding would have several important implications. First, it would indicate that children are sensitive to entropy, thereby expanding our understanding of the distributional properties

that impact learning. Second, it would highlight the importance of using stimuli that are more ecologically valid in their distributional properties: If children show better learning from a low entropy distribution, then previous conclusions about their ability to use multimodal cues may not be accurate. Under more natural conditions, children may show learning that was not previously detected. To give an example from another domain, children's knowledge of irregular plurals is much better when they are produced in familiar frames, as they are often produced in natural language (e.g. children produce "teeth" more accurately after "brush your---" compared to on its own, Arnon & Clark, 2011). Assessing children's morphological knowledge using single word elicitation under-estimated their true abilities, and could lead to inaccurate conclusions (e.g., that they have not learned the correct irregular form yet). Similarly, performance in artificial language learning studies improves when there are multiple cues to segmentation, as is found in natural language. Visual cues for word boundaries improve segmentation in adults (Cunillera, Camara, Laine, & Rodriguez-Fornells, 2010), as does the use of one-to-one mappings between words and objects (children: Lavi-Rotbain & Arnon, 2017; adults: Thiessen, 2010). Under these conditions, children and adults show better learning.

Will a reduction in entropy have a similar facilitative effect on learning? Looking at another domain, adults' cross-situational learning of novel object-label mappings was facilitated after exposure to a Zipfian distribution (with low entropy) compared to a uniform distribution (with high entropy) (Hendrickson & Perfors, 2019, Experiment 2). This facilitative effect was not found when words and labels were presented one at a time: the authors propose that Zipfian distributions are beneficial only when learners are faced with ambiguity. In such cases, the very frequent word can be learned early on and used to disambiguate later trials. Another reinforcement to the potential advantage of low entropy distributions comes from word segmentation studies: children's and adults' segmentation is facilitated when the input had low entropy (entropy was reduced by making one word more frequent than the other, Lavi-Rotbain & Arnon, 2018, 2019), and when it has a Zipfian distribution (Kurumada, Meylan, & Frank, 2013).

Here, we expand on these findings to look at the effect of reduced entropy on word learning in children: will lower entropy facilitate learning in a task that involves both segmentation and object-label mapping? The segmentation task is inherently ambiguous: since learners are exposed to an unsegmented stream, successfully segmenting one word can help in segmenting the rest. An additional facilitative effect can come from the overall greater predictability of the input: non-uniform distributions are more predictable and have lower entropy. If learners are sensitive to such measures of the environment, then learning may be facilitated even in non-ambiguous situations. Since both factors are relevant for the segmentation task, we hypothesized that segmentation will be better under low entropy. The predictions are less clear about learning the object-label mappings. On the one hand, this task does not involve ambiguity: the same object is always presented with the same label. At the same time, the overall predictability of the mappings is greater in the non-uniform distribution. If there is an effect of reduced entropy regardless of ambiguity, we should see a facilitative effect here as well. We hypothesized that learning the object-label mappings will also be facilitated under low entropy.

The current study

In the current study, we use the same artificial language learning paradigm used previously to examine children's learning of multimodal information (Lavi-Rotbain & Arnon, 2017). Children are exposed to an unsegmented speech stream containing four novel words, with consistent word-object pairings (each word is paired with an object: e.g., 'dukame' with a blue star). We ask if children will show better learning of both segmentation and object-label pairings when exposed to low entropy input compared to high entropy input. We focus our inquiry on words that have lower frequency. Frequency is known to affect word learning during infancy with more frequent words learned earlier (Goodman, Dale, & Li, 2008). Frequency, however, does not account for all the variance in a words' age-of-acquisition. It is easy to find examples of low frequency words among the early acquired ones: for example, the word ' cheek' is learned at 22 months (Frank, Braginsky, Yurovsky, & Marchman, 2017), but appears only 18 per million. Could the low entropy found for words in natural language help children learn low frequency words? Finding such a pattern in our experimental manipulation would open up new avenues for understanding how low frequency words are acquired.

We manipulate entropy by making one word much more frequent: in the high entropy condition, all words appeared an equal number of times (each word appeared 32 times). In the low entropy condition, one word was much more frequent (appearing 214 times), while the other three appeared 19 times (half of the frequency of the words in the high entropy condition). We compare segmentation and word-object pairings for the low frequency words from the low entropy condition (which appeared 19 times) with the words from the high entropy condition (which appeared 32 times). If children are mostly sensitive to frequency, then learning should be better in the high entropy condition. However, if children are sensitive to more than mere frequency, in particular to the entropy of the distribution, than learning of the low frequency words should be better in the low entropy condition. If this happens regardless of ambiguity, then we should see better performance due to entropy reduction for both segmentation and learning the correct object-label pairings.

Method

Participants

61 children took part in this Experiment (age range: from 9;0 to 12;0 years, mean age: 10;4 years; 27 boys, 34 girls). We chose this age range since it matches the one used in the older

group of Lavi-Rotbain & Arnon (2017), where children showed poor learning of the object-label pairings. Participants were visitors at the Bloomfield Science Museum in Jerusalem and were recruited for this study as part of their visit to the Living Lab. Parental consent was obtained for all participants. None of the children had known language or learning difficulties and all were native Hebrew speakers. Each child received a small prize for their participation.

Materials

Auditory stimuli

All participants were exposed to a familiarization stream corresponding to the condition they were assigned to. All streams were composed of the same four unique tri-syllabic synthesized words: "dukame", "nalubi", "kibeto", and "genodi". We used only four words to make the task learnable for children. As the results show, even this number proved challenging for children. The twelve different syllables making up the words were taken from Glicksohn & Cohen (2013). The syllables were created using the PRAAT synthesizer (Boersma & van Heuven, 2001) and were matched on pitch (~76 Hz), volume (~60 dB), and duration (250–350 ms). The four words were created by concatenating the syllables using MATLAB to ensure that there were no co-articulation cues to word boundary. The words were matched for length (average word length- 860ms, range=845-888ms). The words were then concatenated together using MATLAB in a semi-randomized order to create the auditory familiarization streams. Importantly, there were no breaks between words and no prosodic or co-articulation cues in the stream to indicate word boundaries. The only cue for word boundaries was transitional probabilities (TP's): TP's between words were lower compared to TP's within words.

Experimental conditions

We created two auditory sequences, corresponding to two levels of entropy: high and low. In the high entropy level, words followed a uniform distribution with each word appearing 32 times in a semi-randomized order (no word appeared twice in a row). The sequence had 128 tokens and lasted 1:50 minutes. TP's within a word were 1, and TPs between words were 0.333. In the low entropy level, words appeared with a skewed distribution: one word appeared 80% of the time (214 appearances) while each of the other three words appeared only 7% of the time (19 appearances for each word). The sequence had 271 tokens and lasted 3:50 minutes. The identity of the frequent word was counterbalanced across subjects in the low entropy condition to prevent item-specific effects. TP's within a word were 1, but the TP's between words varied depending on the next word (since the frequent word in this condition was more likely to occur). See Table 1 for full details of the experimental conditions.

Visual stimuli

While listening to the audio stream, participants saw shapes on the screen whose appearance was synchronized with word boundaries. Shapes appeared at word onset and remained onscreen for the duration of the word. Each word appeared always with the same shape and vice versa ("dukame": blue star, "nalubi": green hexagon, "kibeto": purple heart, and "genodi": orange diamond). In the low entropy condition, when the same word appeared twice in a row, the shape disappeared briefly (for 200 ms) at the end of the first occurrence and reappeared with the second occurrence onset. The visual stimuli is modelled on the regular condition from Thiessen (2010) and Lavi-Rotbain & Arnon (2017), which was shown to facilitate segmentation in both adults and children. See Fig. 1 for an illustration.

Table 1: Different experimental conditions

	High entropy (Uniform)	**Low entropy**
Exposure length [minutes]	1:50	3:50
Number of tokens	128	271
Tokens per word	32	Frequent: 214 Infrequent: 19
Unigram entropy [bits]	2	1.1
TP's between words	0.33	For the frequent word: 0.75 For infrequent words: 0.08

Segmentation test

This test asked how well children segmented the continuous stream into words using 16 two alternative forced-choice trials. The visual stimuli did not appear on screen during test. Participants heard two words and were asked to decide which belonged to the language they heard. We used non-words as foils ("dunobi", "nabedi", "kilume", and "gekato", average length: 860ms; range 854-868ms), created by taking three syllables from three different words, while keeping their original position. We used non-words (instead of part-words) as foils since these are easier to distinguish from 'real' words. Since children struggle with this task, we chose to focus only on the "easier" non-word vs. word distinction. Each of the four words appeared once with each of the four foils to create 16 trials (in a random order, with the constraint that the same word/foil did not appear in two consecutive trials). The order of words and foils was counter-balanced so that in half the trials, the real word appeared first and in the other half, the foil appeared first.

Fig. 1: Audio-video illustration

Word–shape correspondence test

This test asked how well children learned the correspondence between the words and the shapes. In each trial, children saw the four shapes on the screen and heard one of the four words. Then, they had to choose the shape corresponding to the word. Each word was repeated four times on non-consecutive trials, to create 16 trials that appeared in a random order between subjects.

Procedure

After receiving parental consent, children were seated in front of a computer station with a noise-blocking headset next to an experimenter. The children were told they are about to hear an alien language, and that they need to pay attention to what they will see and hear and try to learn it as best as they can. Each child was randomly assigned to one of the two experimental conditions. After the exposure phase, children completed a segmentation test and a word-shape correspondence test. The instructions were identical in all conditions.

Results

Children were divided as follows between the two conditions: high entropy, N=28; low entropy, N=33. Age did not differ across entropy conditions (F(1)=0.195, p=0.66). In the low entropy condition, the frequent word was counterbalanced across subjects. A one way ANOVA revealed that performance was not impacted by which word was the frequent one in the segmentation test (F(3)=2.326, p=0.1), or in the recognition test (F(3)=0.52, p=0.67). Consequently, in all subsequent analyses we collapsed the data across the different frequent words.

Segmentation analysis

Children showed learning (were above chance) in both conditions (low entropy condition: M=73.9%, t(32)=7.69, p<0.001; high entropy condition: M=65.2%, t(27)=4.83, p<0.001). However, this success rate includes both the frequent and the infrequent word for the low entropy condition. Since the frequent word had much higher frequency (214 appearances) than the other words (19 appearances) in the low entropy condition, it does not make sense to include the frequent word in our analysis. In order to examine the effect of entropy on low frequency words alone, we looked only at trials where the correct answer was a low frequency word (appearing 19 times during exposure). This left 12 trials per participant. In this subset of the segmentation test, children showed learning above chance of the infrequent words (low entropy condition: M=73.0%, t(32)=7.0, p<0.001) (see Fig. 2). We will now compare this mean to the one from the high entropy condition (73.0% versus 65.2% respectively).

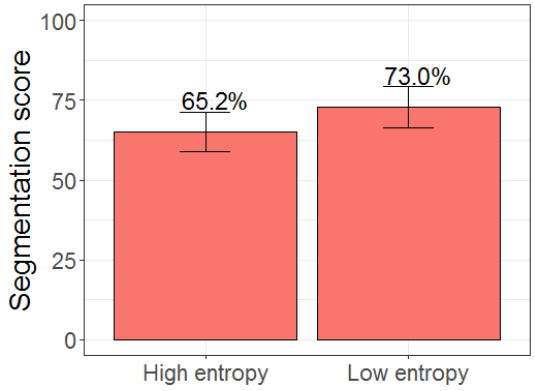


Fig. 2: Mean segmentation score of low frequency words by entropy condition with 95% confidence intervals

We used mixed-effect linear regression model to examine the effect of entropy level on segmentation of infrequent words. Following Barr et al. 2013, the models had the maximal random effect structure justified by the data that would converge. Our dependent binominal variable was success on a single trial of the segmentation test. We had entropy condition (high entropy condition as baseline) as a fixed effect, as well as: age-in months (centered); gender; trial number (centered); and order of appearance in the test (word-first trials vs. foil-first trials). The model had random intercepts for participants and for item (Table 2). To examine the overall effect of entropy, we used model comparisons.

As predicted, entropy level impacted segmentation of low frequency words (chi(1)=3.2, p=0.07). Participants showed better segmentation of low frequency words in the low entropy condition compared to the high entropy condition, despite appearing half the times (β=0.42, SE=0.23, p=0.07). Order of appearance in the test significantly affected segmentation, with better accuracy on trials where the word appeared before the foil (β=0.39, SE=0.16, p<0.05), as has been found in previous studies (Lavi-Rotbain & Arnon, 2017; Raviv & Arnon, 2018). Since the order of presentation of

Table 2: Mixed-effect regression model for segmentation of infrequent words. Variables in bold were significant. Significance obtained using the lmerTest function in R.

	Estimate	**Std. Error**	**z value**	**p-value**
(Intercept)	0.69584	0.21439	3.246	**<.01 ****
Age (centered)	0.23516	0.14220	1.654	=0.098 .
Low entropy condition	0.42079	0.23321	1.804	**=0.07 .**
Gender (male)	-0.03948	0.11747	-0.336	>.1
Trial number (centered)	-0.02390	0.01705	-1.401	>.1
Order of appearance (word)	0.19385	0.07844	2.471	**<.05 ***

words and foils was counter-balanced this could not reflect a preference for pressing 1 or 2, and is in line with the "interval bias" which is often found in 2AFC tests (Yeshurun, Carrasco, & Maloney, 2008). Age almost reached significance: older children were slightly better than younger ones (β=0.24, SE=0.14, *p*=0.098). Trial number and gender did not affect segmentation (trial number: β= -0.02, SE=0.02, *p*>0.1; gender: β= -0.08, SE=0.23, *p*>0.1).

The beneficial effect of low entropy on segmenting low frequency words cannot be attributed to learning only the frequent word, and ruling out foils due to syllables they share with the frequent word. To see if there is a difference between trials in the low entropy condition where the foil shared one syllable with the frequent word (M=72.7%) and trials where it didn't (M=77.3%) we used a linear regression model with success on a single trial as the dependent binominal variable, and "is foil frequent" (assigned '1' for trials in which the foil shared any of its three syllables with the frequent word and '0' when it didn't) as a fixed effect, as well as log frequency (centered), gender, trial number (centered); and order of appearance in the test. The model had random intercepts for subjects and for items. "Is foil frequent" was not a significant predictor of accuracy (β= -0.27, SE=0.25, *p*>0.1), neither all the other fixed effects. That is, children in the low entropy condition indeed learned the low frequency words better.

Recognition analysis

Children showed learning of object-label pairings (were above chance) in the low entropy condition (M=49.3%, chance=25%, t(32)=5.16, *p*<0.001). In contrast, they were not above chance in the high entropy condition (M=32.7%, chance=25%, t(27)=1.66, *p*=0.11). The accuracy in the high entropy condition is similar to that from Lavi-Rotbain & Arnon (2017), using the same task and uniform distribution (M=34.4%). While children did show learning in the previous study (just above chance), their performance was still poor, indicating difficulty in learning the mappings from a uniform distribution. How well did children learn the infrequent words in the low entropy condition? As in the segmentation test, the mean accuracy in the low entropy includes also recognition of the frequent word. In order to look at recognition of low frequency words, we looked only at trials where the correct word was a low frequency word (12 trials per child). Since children learned the frequent word quite well (M=64.6%), we assume that chance level on each trial is not 25% but 33% (since they could rule out the shape corresponding to the frequent word). Note however that this is a quite rigid assumption: children did not show complete learning of the frequent word (they were incorrect 35% of the time), and there was very large variance in accuracy (SD=37.9%), meaning that for some trials they were picking between four options. Nevertheless, we put our prediction to a stringent test and assume that chance is 33% for the low frequency words. As predicted, children learned the infrequent words above chance in the low entropy condition (M=44.0%, chance=33%, t(32)=2.14, p<0.05) (see Fig. 3). This means that while in the uniform condition children did not show learning of the object-label pairings (were not above chance), children in the low entropy condition did show learning even of the infrequent words, despite appearing half the number of times and despite the rigid chance level.

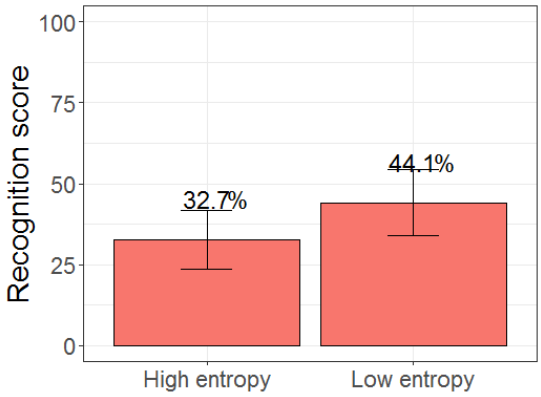


Fig. 3: Mean recognition score of low frequency words by entropy condition with 95% confidence intervals

Is there a correlation between segmentation and recognition scores? Previous results showed positive correlation in adults' performance between these two tasks, indicating that better segmentation went along with better word learning (Lavi-Rotbain & Arnon, 2017; Thiessen, 2010). However, such a correlation was absent in children's performance (Lavi-Rotbain & Arnon, 2017). Here, we found a positive correlation only in the low entropy condition: children who performed well in the segmentation test, were also good at mapping labels to objects (R^2=0.4, t(31)=2.42, *p*<0.05), highlighting the connection between segmentation and word learning found in natural language. Such a correlation was not found in the high entropy condition (R^2=0.21, t(26)=1.12, *p*>0.1).

Discussion

We set to ask whether children's ability to segment and learn object pairings for low frequency words will be better when learning from low entropy input compared to high entropy input. To do so, we examined children's performance in an artificial language across two levels of entropy (high and low), in two tasks: segmentation and object-label pairing. Entropy was reduced by making one word more frequent than the rest, so that it appeared 80% of the time. We focused on children's performance on low frequency words (that appeared only 19 times in the low entropy condition, versus 32 in the high entropy condition). Our results show that entropy reduction is beneficial for children's segmentation, (see also Lavi-Rotbain & Arnon, 2018, 2019), as well as for their learning of object-label mapping. In addition, we found a positive correlation between segmentation and mapping only under the low entropy condition. Based only on findings from the uniform conditions from this study and from Lavi-Rotbain & Arnon (2017), one could conclude that children are not able to simultaneously learn segmentation and object-

label mapping (at least in lab conditions). However, the low entropy condition offers an alternative explanation: when exposed to more predictable and ecological input, children show evidence of learning both tasks at the same time. Importantly, children's object-label accuracy was still not good, raising the need to find ways to make the task easier: we predict that the effect of reduced entropy will be stronger once that is done. We are currently running a similar study with the younger age group (that showed no learning of the object-label mappings in the previous study), to see if entropy reduction will have a similar facilitative effect on this age group and will enable them to learn both the segmentation and object-label pairing.

Why did the low entropy condition facilitate learning? Several inherent properties of low entropy distributions may be facilitative for learning. First, creating a low entropy in the way we did drastically increases the frequency of one or more of the words. These highly frequent words can be learned relatively early on and later serve as an anchor for learning other words, similar to presenting words in isolation prior to presenting the unsegmented stream (Cunillera, Càmara, Laine, & Rodríguez-Fornells, 2010). In addition, TP's between the frequent and infrequent words can be lower and hence be more salient for learning. However, we suggest that there is more to the low entropy condition that facilitates learning than anchoring and lower TP's. Language learners may be sensitive to the overall predictability of the input, and learn better from input with lower entropy. Such an account predicts that entropy reduction will also facilitate learning when there is less ambiguity. Our results provide some support for this, by showing that learning was improved also for the non-ambiguous object-label pairings (contra the prediction made in Hendrickson & Perfors, 2019). This prediction is also supported by findings showing that adults' word segmentation is facilitated in a low entropy condition compared to a medium entropy one, despite both having similar anchoring and TP cues (Lavi-Rotbain & Arnon, 2019). Further work is needed to understand what exactly about low entropy is facilitative and how that relates to the input that children are actually exposed to.

From a methodological perspective, our results highlight the importance of creating experimental stimuli that better reflect the input children hear. In particular, most SL studies use a uniform distribution during exposure, although the distribution itself is not relevant for their research question. However, by doing so, we may be introducing unnecessary difficulties for our participants that may interfere with our assessment of their abilities. For children, who find artificial language learning experiments harder to begin with, such factors may impact performance more, and more easily. Theoretically, the findings point to the importance of studying the impact of entropy on language learning. Entropy has been studied across many domains of language, including language processing, use and structure (e.g., Cohen Priva, 2017; Linzen & Jaeger, n.d.; Piantadosi, Tily, & Gibson, 2011). For example, there is evidence that the entropy of single words is restricted to a small range of values across many languages, suggesting that speakers have similar preferences for how predictable their languages are (Bentz, Alikaniotis, Cysouw, & Ferrer-i-Cancho, 2017). In addition, there is a trade-off between unigram and trigram entropy over time across many languages, indicating that speakers maintain a relatively constant information rate (Cohen Priva & Gleason, 2016). Children also show sensitivity to such measures: two-year-olds show better repetition of unfamiliar four-words sequences when the final word "slot" has higher entropy (Matthews & Bannard, 2010). These studies show that language users are sensitive to entropy and other information-related measures, and suggest that languages are shaped by constrains arising from these measures. However, the role of entropy on language learning is understudied. Our results show that entropy effects are found in children and impact learning of both segmentation and word labels.

Our results may offer a possible explanation for how children acquire low frequency words at a relatively early age. Words in natural language show a Zipfian distribution, in which most of the words have low frequencies. Under a low entropy distribution, such as the Zipfian distribution, the disadvantage of low frequency can turn into an advantage: the few frequent words can serve as an anchor for learning the low frequency ones. We are currently conducting a series of studies to examine the role of entropy in natural language learning, and in predicting variance in age-of-acquisition.

Acknowledgments

We wish to thank Zohar Aizenbud for her help with creating the study; the Living Lab staff and the Bloomfield Science Museum in Jerusalem and Noam Siegelman for his help in analyses. The research was funded by the Israeli Science Foundation grant number 584/16 awarded to the second author.

References

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. Journal of Memory and Language, 68(3), 255–278.

Bentz, C., Alikaniotis, D., Cysouw, M., & Ferrer-i-Cancho, R. (2017). The entropy of words-Learnability and expressivity across more than 1000 languages. Entropy, 19(6), 1–32.

Bergelson, E., & Swingley, D. (2012). At 6-9 months, human infants know the meanings of many common nouns. Proceedings of the National Academy of Sciences, 109(9), 3253–3258.

Boersma, P., & van Heuven, V. (2001). Speak and unSpeak with Praat. Glot International, 5(9–10), 341–347.

Clerkin, E. M., Hart, E., Rehg, J. M., Yu, C., & Smith, L. B. (2017). Real-world visual statistics and infants' first-learned object names. Philosophical Transactions of the Royal Society B: Biological Sciences, 372(1711), 20160055.

Cohen Priva, U. (2017). Not so fast: Fast speech correlates with lower lexical and structural information. Cognition,

160, 27–34. Retrieved from
Cohen Priva, U., & Gleason. (2016). Simpler structure for more informative words : a longitudinal study. Proceedings of the 38th Annual Conference of the Cognitive Science Society, (2012), 1895–1900.
Cunillera, T., Camara, E., Laine, M., & Rodriguez-Fornells, A. (2010). Speech segmentation is facilitated by visual cues. Quarterly Journal of Experimental Psychology (2006), 63(2), 260–274.
Cunillera, T., Càmara, E., Laine, M., & Rodríguez-Fornells, A. (2010). Words as anchors: Known words facilitate statistical learning. Experimental Psychology, 57(2), 134–141.
Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2017). Wordbank: An open repository for developmental vocabulary data. Journal of Child Language, 44(3), 677–694.
Glicksohn, A., & Cohen, A. (2013). The role of cross-modal associations in statistical learning. Psychonomic Bulletin & Review, 20(6), 1161–1169.
Goodman, J. C., Dale, P. S., & Li, P. (2008). Does frequency count? Parental input and the acquisition of vocabulary. Journal of Child Language, 35(3), 515–531.
Hendrickson, A. T., & Perfors, A. (2019). Cross-situational learning in a Zipfian environment. 189(May 2017), 11–22.
Kurumada, C., Meylan, S. C., & Frank, M. C. (2013). Zipfian frequency distributions facilitate word segmentation in context. Cognition, 127(3), 439–453.
Lavi-Rotbain, O. & Arnon, I. (submitted). Zipf's Law in Child-Directed Speech.
Lavi-Rotbain, O. & Arnon, I, (2018, November). Frequency or predictability? The effect of entropy on statistical learning in children and adults. Poster session presented at the Boston University Conference on Language Development, Boston, MA.
Lavi-Rotbain, O., & Arnon, I. (2019). Low Entropy Facilitates Word Segmentation in Adult Learners, Proceedings of the 41st Annual Conference of the Cognitive Science Society. Cognitive Science Society
Lavi-Rotbain, O., & Arnon, I. (2017). Developmental Differences Between Children and Adults in the Use of Visual Cues for Segmentation. Cognitive Science.
Linzen, T., & Jaeger, T. F. (2014). Investigating the role of entropy in sentence processing. ACL 2014, 10.
Matthews, D., & Bannard, C. (2010). Children's production of unfamiliar word sequences is predicted by positional variability and latent classes in a large sample of child-directed speech. Cognitive Science, 34(3), 465–488.
Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: A critical review and future directions. Psychonomic Bulletin & Review, 21(5), 1112–1130.
Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. Proceedings of the National Academy of Sciences of the United States of America, 108(9), 3526–3529.
Raviv, L., & Arnon, I. (2018). The developmental trajectory of children's auditory and visual statistical learning abilities: modality-based differences in the effect of age. Developmental Science, 21(4), 1–13.
Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. Science (New York, N.Y.), 274(5294), 1926–1928.
Shannon, C. E. (1948). A mathematical theory of communication. The Bell System Technical Journal, 27, 379–423.
Thiessen, E. D. (2010). Effects of Visual Information on Adults' and Infants' Auditory Statistical Learning. Cognitive Science, 34(6), 1093–1106.
Yeshurun, Y., Carrasco, M., & Maloney, L. T. (2008). Bias and sensitivity in two-interval forced choice procedures: Tests of the difference model. Vision Research, 48(17), 1837–1851.
Zipf, G. (1936). The Psychobiology of Language. London: Routledge.

Active Learning for a Number-Line Task with Two Design Variables

Sang Ho Lee
The Ohio State University, Columbus, Ohio, United States

Dan Kim
The Ohio State University, Columbus, Ohio, United States

John Opfer
The Ohio State University, Columbus, Ohio, United States

Mark Pitt
Ohio State University, Columbus, Ohio, United States

Jay Myung
The Ohio State University, Columbus, Ohio, United States

Abstract

The number-line task is a widely used task in diverse fields of study. In the task, a given number that varies every trial is estimated on a continuum flanked with 0 and an upper-bound number. An upper-bound of a number-line is often arbitrarily selected by researchers, although this design variable has been shown to affect the non-linearity in estimates. Examining estimates of varying given numbers (design variable 1) with varying upper-bound numbers (design variable 2) can be costly because adding a new design dimension into a number-line task could drastically increase the number of trials required for examining the underlying representation of number. The present study aims to conduct a number-line task with the given number and the upper-bound being the design variables. A design optimization algorithm, Gaussian Process Active Learning (GPAL), made this new paradigm feasible without increasing the number of trials, by presenting only the most informative combinations of the design variables every trial. Our experimental data showed that the non-linearity of the number-line estimates increases with the upper-bound of the number line. The degree of non-linearity could predict a math skill (i.e., addition proficiency), but only when the upper-bound was relatively large. The observed range-dependency of the number-line estimates would not be fully explored without systematically manipulating the upper-bound as an additional design variable. As in the present number-line task, GPAL would be a useful tool for the research problems that require multidimensional design experiments to be solved.

Who is better? Preschoolers infer relative competence based on efficiency of process and quality of outcome.

Julia A. Leonard[1] (jlnrd@sas.upenn.edu), Grace Bennett-Pierre[2] (gbp@stanford.edu), Hyowon Gweon[2] (hyo@stanford.edu)
[1]Department of Psychology, 3720 Walnut Street Philadelphia, PA 19104 USA
[2]Department of Psychology, 450 Serra Mall, Jordan Hall Stanford, CA 94025 USA

Abstract

The ability to reason about our own and others' competence informs our everyday decisions. However, competence is an abstract concept which manifests in the objective properties of the task completed by an agent (i.e., task-based features, such as quality of outcome or task difficulty) as well as the subjective properties of the agent (i.e., agent-based features, such as dexterity, speed, focus). Thus, acquiring an integrated notion of competence may be a nontrivial challenge for young children. Prior work on children's understanding of competence has often used explicit verbal cues to describe the relevant features, or experimental tasks that confounded these features. Here we examine how preschool-aged children evaluate the relative competence of two agents by systematically manipulating task-based and agent-based features without explicit linguistic or gestural support. We find that 4- and 5-year-olds readily use perceptual cues to task-based (i.e., task difficulty) and agent-based (i.e., agent speed) features to infer competence (Exp.1-3) but not when when these perceptual cues are closely matched (Exp.4). These results suggest that a basic understanding of relative competence emerges earlier than previously believed, but an abstract, adult-like concept of competence may continue to develop throughout childhood.

Keywords: Social Cognition, Competence, Ability

Introduction

Beliefs about our own and others' competence are deeply ingrained in our everyday lives; we think about it, talk about it, and use it to guide our daily decisions. Even young children prefer agents who are perceived as more competent (Jara-Ettinger, Tenenbaum, & Schulz, 2015), and consider their own and others' competence to decide how to allocate tasks that vary in difficulty (Magid, DePascale, & Schulz, 2018). More generally, the way we perceive our own competence influences our motivation to learn and to choose challenging goals (Dweck & Leggett, 1988; Nicholls, 1984; Stipek & Iver, 1989; Wentzel & Wigfield, 1998).

However, it is often difficult to generate a clear definition of what it means to be *good* (competent) at something, or what makes some people *better* (more competent) than others. In fact, the meaning of competence seems to change depending on the task domain or the nature of the activity. When we say "Sally is good at math" or "Sally is a good pianist", we are referring to different dimensions on which we evaluate other people - her intellectual abilities in one, and her finesse in playing an instrument in the other. Even within the realm of sports, saying "Sally is good at gymnastics" or "Sally is a good swimmer" refers to physical abilities that vary along several dimensions, such as strength, agility, or speed. Thus, acquiring an integrated concept of competence that incorporates a coherent relationship between these dimensions may be a formidable challenge for young children.

There are broadly two ways in which one's competence can manifest. First, a competent agent might be capable of achieving goals or outcomes that are costlier, or more effortful, than what others can achieve (i.e., more difficult, more complex, or more elaborate). In this case, competence is marked by an objective property of the task or the quality of outcome achieved by the said agent (henceforth task-based features). Second, a competent agent might achieve the same outcome on the same task as others but more efficiently (i.e., spending less time, less physical effort, or less mental effort such as care or attention). In this scenario, competence manifests as a property of the agent who completes the task (henceforth agent-based features). Although there are cases in which we can expect someone to be competent even before observing anything he or she does (e.g., if someone went to Julliard, they presumably can play an instrument quite well), when we are trying to learn about an agent's competence based solely on their actions or outcomes, we usually attend to these task-based and agent-based features.

Prior work suggests that young children readily use explicit task-based features (e.g., clearly good or poor performance; frequency of successes and failures) to infer agents' competence. For instance, 3-year-olds judge that an agent who made a "tastier" cake is better at baking than an agent who made a "yucky" cake (Yang & Frye, 2016). However, a coherent, theory-like understanding of competence that integrates task-based and agent-based features seems to emerge relatively late in childhood. Given a video of two children, one of whom worked diligently on a math problem and the other who "goofed off" and worked only intermittently, children under age 7 say that the one who worked harder is more competent even if they both got the same score (matched outcome, different efficiency; Nicholls, 1978). Strikingly, it was not until age 12 that children showed an adult-like understanding that one's competence is inversely related to the total amount of effort invested when outcome is matched.

One way to interpret these results is that young children consider competence as a globally positive construct, and do not yet understand the specific relationship between task-based and agent-based features in reasoning about an agent's competence. Instead, young children might resort to explicit verbal or perceptual cues, and associate anything positive (e.g., better outcome, higher effort, being more diligent) with higher competence. Consistent with this idea, some studies

suggest that focusing on task-based features such as quality of outcome can lead children astray. Heyman, Gee, & Giles (2003) found that after being told stories about characters that varied in how much they tried and how well they did on schoolwork, children were more likely to remember situations where high effort led to a positive academic outcome than a poor academic outcome. Similarly, 3-4-year-olds only attended to positive task outcome (e.g., who won the race), and not process (e.g., the faster agent tripped on an obstacle), to infer future competence (Yang & Frye, 2016).

However, another possibility is that young children do possess a coherent yet preliminary understanding of competence that manifests only with additional contextual support. For instance, one study used a paradigm similar to Nicholls (1978) but with explicit labeling (e.g. this person is "lazy", this person paints "very well") and found that even 4-year-olds have a mature understanding of the causal relationship between ability, effort, and outcome: They understand that an agent with high ability and a poor outcome probably didn't try hard, whereas an agent with low ability and a good outcome probably did try hard (Wimmer, Wachter, & Perner, 1982). In a similar paradigm, Heyman & Compton (2006) found that 5-year-olds correctly inferred that a faster agent was smarter, but only when primed to focus on difficulty (whether actors thought the test was hard or easy), and not effort (whether the actors tried hard or not).

These two possibilities have been challenging to tease apart particularly because prior work has used different ways of presenting information about competence, making it difficult to compare results across studies. Many studies used narratives that are rather high in verbal or working memory demands, or required extensive domain knowledge about what constitutes better quality or outcome. Thus, studies that are high in verbal or memory demands might have underestimated children's abilities. On the other hand, there are reasons to believe that some of the tasks used in prior work provided ample (and rather generous) behavioral and linguistic cues that are superficially associated with competence. For instance, some prior studies simply required mapping valenced cues of quality (e.g., success vs. failure; good vs. bad outcome) to agents' competence on a similarly valenced scale (e.g., who is smarter?).

However, the quality of outcomes in many real-world activities are usually not clearly marked nor necessarily positive or negative. Therefore, a test of a genuine understanding of competence must ask whether children can integrate information about the expected time or effort required to complete a given task (i.e., difficulty) and the actual time or effort an agent needed to complete the task. For instance, if two people took the same amount of time to build two block towers, one of which clearly looks harder to build than the other (e.g., a tall vs. short tower), a child might simply associate the agent who built the taller tower with higher competence. If, however, the towers are the same height and shape but nonetheless vary in the actual effort required for building (e.g., one is made of many more smaller blocks than the other), judging the competence of agents requires an abstract understanding of competence that goes beyond the use of perceptual cues. Whether children have such an abstract notion of competence remains an open question.

Here, we ask whether 4- and 5-year-olds use task-based features (i.e., difficulty of the completed goal) and agent-based features (i.e., speed) to infer an agent's underlying competence. While competence can be assessed in a variety of domains, our experiments focus on children's inferences about agent competence in building block structures. We choose this domain because 1) previous work has shown that even 4-year-olds can accurately judge the relative difficulty of building different block structures (Gweon, Asaba, & Bennett-Pierre, 2017) and 2) unlike more abstract forms of competence (e.g., mental ability, intelligence), agents' competence on physical tasks often manifests in ways that are more concrete and visually accessible. Thus, we can use simple perceptual features to manipulate task-based and agent-based indicators of competence without relying on explicit verbal cues. Across four experiments, we systematically vary a task-based feature (task difficulty, marked by perceptual properties of the block structures) and an agent-based feature (building speed, marked by duration of total build time) to see if 4- and 5-year-olds can use these features to infer others' relative competence.

Experiment 1

Methods

Participants & Materials We recruited 30 4- to 5-year-old children at a local children's museum (mean: 59.70 months (range: 48 - 70), 47% girls). One additional child was excluded from analyses due to failing the practice question (n = 1, see Procedure). Participants viewed laminated pictures and watched videos on a laptop.

Procedure Children were tested individually in a private testing room. To make sure that children understood the word "better", the experimenter first asked children "Who is better at writing letters - you or your parents?" and then "Who is better at playing on the playground – you or your parents?". If children answered incorrectly (i.e., choosing themselves for writing, their parents for playing), they were corrected. Next, children were given a detailed explanation with visual aids of what would happen in the following videos.

Children first watched a practice video where two agents drew shapes (a star and a flower) and finished at different times (counter-balanced for side; agents throughout were matched on ethnicity and physical build). While the agents were drawing, a screen was lowered to block visual access to their progress. One of the agents indicated she was done drawing (saying "all done" with her hands raised above the screen and then moving to the side of the screen to read a book). A few seconds later, the other agent indicated she was done in the same manner. Then the screen lifted to reveal what they made. Children were asked to indicate which agent

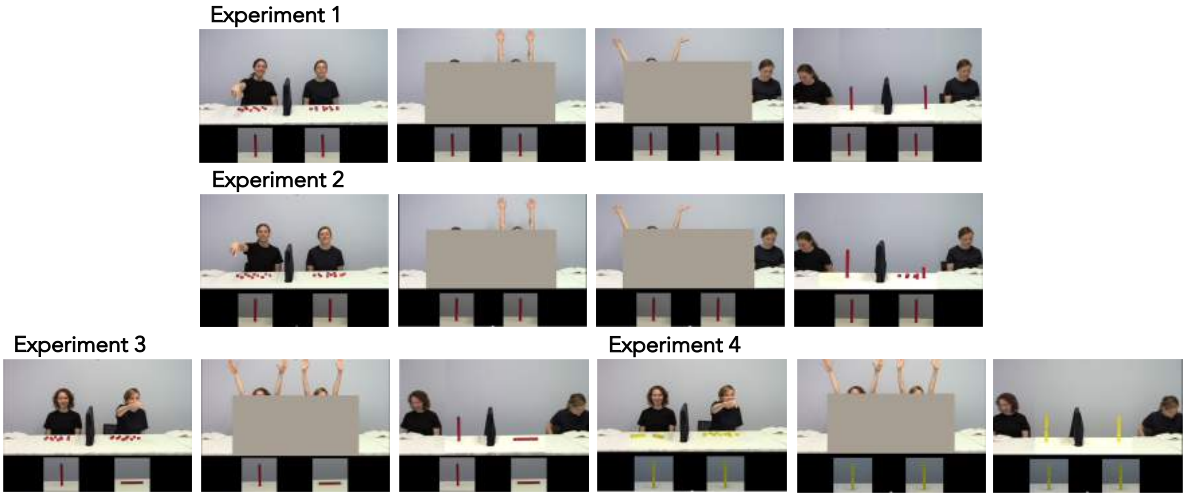


Figure 1: Still images from video stimuli for Experiments 1 through 4. In the first frame of each experiment, agents take turns pointing to the picture below them and saying "I'm going to make this". In the next frame, visual access to their building progress is blocked by an occluder. In Experiments 1 and 2, the agents finish at different times (marked by reaching up their hands, saying "all done", and moving to the side to read a book). In Experiments 3 and 4 agents finish building at the same time. Finally in the last frame, the occluder is lifted to reveal what each agent made.

finished first. If they answered this question incorrectly, they were excluded from analyses.

In the test video, children watched two agents build block structures. Below each agent was a picture of a 10-block vertical tower (see Figure 1, Experiment 1). The agents first indicated that they wanted to build the tower in the picture: One agent first said, pointing to the picture below her, "I'm going to make *this*", then the other agent repeated the same action. Next, the agents said "Ready? Go!" and began to build at the same time. As in the practice video, a screen was lowered to block the child's visual access to the agents' building actions. One agent finished building the tower in 10 seconds, indicating she was done in the same way as in the practice video (raising hands, saying "all done", and moving to the side of the screen to read); 5 seconds later, the other agent indicated she was done in the same manner (building time 15 seconds; side counter balanced). Critically, there were no verbal cues to indicate that one agent was "faster", or "found building easier" - children had to infer the agents' relative competence solely from the perceptual information in the video. The screen then lifted up to reveal what each agent made. At the end children were asked the critical test question, "Who is better at building blocks?"

Results & Discussion

Children's performance on the test question was significantly above chance (binomial test against chance (50%): 86.7% correct, CI = [76.7%, 100%][1], $p < .001$). Thus, children were able to understand that if two agents build the same structure, the agent that completes it earlier is more competent than the one who finishes later. This suggests that, when outcomes were matched (i.e., task-based feature kept constant), children can use differences in building time (an indicator of an agent's speed, an agent-based feature) to infer relative competence. However, another possibility is that children simply associated being faster with being better, without considering outcomes. In Experiment 2, we sought to rule out this alternative. If the agent who finishes first actually does not complete her goal, a simple association would still favor this agent as more competent. However, if children consider speed as an indicator of competence only when the agents have achieved the same goals, they should favor the agent who completed her goal even though she finished later.

Experiment 2

Methods

Participants Using data from Experiment 1, we ran a simulated power analysis using 10,000 binomial tests (bootstrapped samples with replacement) and set the sample size at n = 20 for a simulated power of .96.[2]. We recruited 4- and 5-year-olds at a children's museum (mean: 59.90 months (range: 48 - 71), 50% girls); 5 additional children were excluded from analyses due to failing the practice question (n = 2) or the inclusion criteria question (n = 3, see Procedure).

Procedure The procedure was similar to Experiment 1, except for the final outcome revealed at the end of the test video. While both agents indicated that they'd build a 10-block vertical tower, the agent who finished first (in 10 seconds) actually only built a 3-block tower whereas the agent who finished

[1] All reported CIs are 95% confidence regions estimated through a basic non-parametric bootstrap of the data using 500,000 samples.

[2] Experiments 2 and 3 were pre-registered on Open Science Framework (OSF): https://osf.io/pc945/registrations.

second (in 15 seconds) completed her goal (10-block tower). The test question was the same. To ensure children remembered the key event in the video, we asked: "Which agent didn't finish making her tower?" Children who gave incorrect answers were excluded from analyses.

Results & Discussion

Children's performance on the test question was significantly above chance (binomial test against chance = 50%; 90% correct, CI = [80%, 100%], $p < .001$). These results suggest that children do not indiscriminately use speed or time-to-completion as a cue to competence; when one person did not complete her goal, children resisted saying she was more competent even though the agent claimed to be done before the other agent. This complements our finding from Experiment 1, providing evidence that children's successful use of time-to-completion is not based on a simple heuristic "faster = better".

In Experiment 3, we now ask whether children can use a task-based feature (task difficulty) to infer relative competence when an agent-based feature (time to completion) is held constant. Critically, going beyond prior work that provided children explicit verbal cues to the task difficulty or outcome, we had children simply observe two agents building two different structures–10 blocks stacked vertically vs. lined up horizontally–and use the inferred difficulty of the two tasks to reason about competence. We chose these structures based on findings from (Gweon et al., 2017) showing that 4-year-olds readily judge the 10-block vertical structure as harder to build than the 10-block horizontal structure based on static pictures of the initial states (i.e., scattered blocks) and final states (finished towers), without seeing the building process. Given these results, we predicted that 4- and 5-year-olds would be able to use their understanding of task difficulty to infer the relative competence of two agents, even when total building time is matched.

Experiment 3

Methods

Participants We preregistered this experiment using the same power analysis as in Experiment 2 (see Footnote 2). We recruited 30 4- and 5-year-old children at a local children's museum (mean: 62.25, months (range: 49 - 71) 50% girls); 10 additional children were tested but excluded due to failing the practice question (n = 3, see Procedure) and inclusion criteria question (n = 7, see Procedure).

Procedure The procedure was similar to Experiment 1 with a few changes. To help children understand that the two agents might complete different goals, they were asked to indicate if the agents drew the same or different pictures after watching the practice videos. For the test video, agents had pictures of different block structures below them; one agent had a picture of a 10-block vertical tower and the other had the picture of a 10-block horizontal tower. As in Experiment 1, the agents pointed to the picture and said "I'm going to build this"; however, it was clear that agents were simply pointing to the structure that was depicted below them rather than making an active choice about which one to build. Furthermore, they never explicitly mentioned the physical properties of the structures nor their expected difficulty. Critically, the agents finished building their structures at the same time. Children were asked: "Who is better at building blocks?" followed by the inclusion question "Which tower is better?" Those who answered the inclusion question inaccurately[3] were excluded from analyses.

Results & Discussion

Children's performance on the test question was significantly above chance (95%, CI = [90%, 100%], $p < .001$). This result held even after including the 7 children who failed to answer the inclusion question accurately (74%, CI = [60%, 93%], $p = .02$). Thus, children were able to tell that when two agents take the same amount of time to build block structures, the agent that built the more difficult structure is more skilled. Critically, children were able to do so from their own assessment of the tasks, in the absence of any explicit information about the task difficulty.

While task difficulty was never mentioned explicitly, one might wonder if children still picked up on the fact that the 10-block vertical tower is taller than the 10-block horizontal structure, and simply associated building a"taller" tower with being "better" at building. Prior work provides some evidence against this possibility, showing that simple heuristics such as height or size do not fully explain children's inferences about task difficulty on a range of structures that vary along different dimensions. (Gweon et al., 2017).

However, whether children can infer the relative competence of two agents in the absence of any physical cues for agent-based (Experiment 1) or task-based features (Experiment 3) remains an open question. Experiment 4 provides a test of this ability, by asking children to judge the relative competence of two agents who take equal amounts of time to make towers that look identical in overall shape and height; critically, despite their near identical appearances, the towers differ in their building difficulty because one is made of 10 cubes and the other is made of 2 long blocks (and thus takes fewer steps, and is easier to build; see Figure 1). We chose these structures because Gweon et al. (2017) have shown that 4- and 5-year-olds can reliably identify the 10-block vertical structure as harder to build than the two-block structure given static pictures of the initial and final states, even without seeing the intermediate building process.

Unlike Experiments 1 - 3 where we hypothesized successful performance given explicit perceptual cues, one might entertain different predictions for Experiment 4. To succeed in this task, children must first infer that one tower is harder than the other, and spontaneously use this understanding to reason

[3]While the correct answer was the vertical 10-block tower, the wording of this question was confusing and potentially problematic; we thus also present results that include these children. In Experiment 4, we used a different question.

about the agents' competence; both of these inferences must be made based on the initial and the final states of the towers, without direct visual access to the actual building process. Thus, on the one hand, 4- and 5-year-olds may struggle with this task; prior work suggests that an abstract, coherent understanding of competence does not emerge until later in childhood, and our stimuli provide no superficial perceptual cues that children could use to judge relative competence. On the other hand, given that our task involves minimal verbal and memory demands, children might show an earlier success than previously believed. Thus, we did not preregister this experiment, allowing ourselves to explore a broader age range.

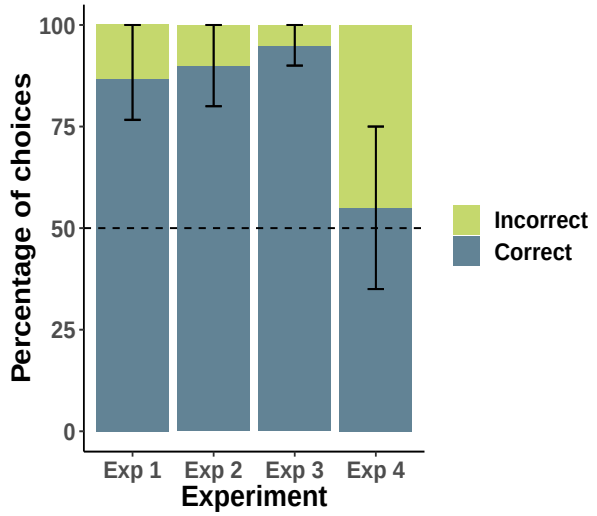


Figure 2: Results from Experiments 1 - 4: The percentage of children who correctly chose the more competent agent Error bars represent bootstrapped 95% confidence intervals. Dashed line indicates chance performance.

Experiment 4

Methods

Participants We used the same sample size as Experiments 2 and 3 (n=20). However, given the more difficult nature of this task, we extended our age range to include 6-year-olds (mean: 62.25 months (range: 48 - 83) 30% girls). Seven additional children were excluded due to failing inclusion criteria question (see Procedure). We also tested 25 adults on Amazon Mechanical Turk.

Procedure The procedure was similar to Experiment 3, with a few changes to minimize task demands. First, we removed the practice trial (more than 93% of children passed in Experiments 1 - 3) because this might bias children to focus on completion time. Additionally, to ensure that children paid attention to the fact that the two agents used different blocks to make similar-looking final structures, children were presented with physical examples of the 10 cubes and 2 long blocks, which were placed next to the side of the screen that matched the tower on the screen. In the test video, one agent had 10 cubes in front of her and pointed to the vertical tower below to indicate that she wants to build that tower; the other agent had 2 long blocks and also pointed to the tower below her (as in other experiments, the pointing was casual and did not indicate any active choice to construct a particular tower). Critically, the agents finished building their structures at the same time. Again, children were asked, "Who is better at building blocks?" followed by a inclusion question, "Which tower is harder to make?" Children were excluded if they incorrectly said that the 2-block tower was harder than the 10-block vertical tower.

Results & Discussion

We first verified that adults can infer relative competence accurately from these videos: 100% of the adults said the agent who built the 10-block tower was more competent than the agent who built the 2-block tower. However, children's performance was not significantly different than chance (55% correct, CI = [35%, 75%], $p = .82$), suggesting that when perceptual markers of difficulty and completion time are matched, children do not distinguish the agent who built the 10-block tower from the agent who built the 2-block tower. However, there was some evidence for a developmental change: Proportionally more 6-year-olds (6/7) than 5-year-olds (3/8) and 4-year-olds (2/5) answered the test question correctly. A logistic regression found a trend for an effect of age in years on children's success on this task ($B = 1.08$, $p = .1$).

General Discussion

Here we asked whether preschool-aged children can use a task-based feature (i.e., difficulty of the task) and an agent-based feature (i.e., agent speed) to infer the relative competence of agents. Critically, these cues were never verbally communicated by the experimenter or the agents in the video. As is the case in many real world situations, children had to spontaneously pick up on these cues and use them to infer relative competence. The difficulty of the tasks had to be inferred from the visual properties of the block structures (such as size or height), and the agents' speed or efficiency had to be inferred from their completion time on a given task. Our results suggest that children not only detect the perceptual cues that signal both types of features, but also readily use them to draw accurate judgments about the relative competence of two agents.

We found near-ceiling performance in 4- and 5-year-old children when one feature was matched and the other clearly varied across agents, marked by explicit perceptual cues. If two agents made the same block tower, the agent who completed her tower first was judged as more competent (Experiment 1), but not when this agent did not complete her goal (Experiment 2). If both agents completed their towers at the same time, the agent who built the more difficult tower was judged as the more competent agent (Experiment 3). However, in a more conservative test where the two agents

spent the same amount of time building towers that varied in difficulty but were matched in their final shape and height, children's accuracy dropped to chance-level (Experiment 4). While there is suggestive evidence that 6-year-olds are able to respond accurately in this scenario, overall children struggled without clear perceptual cues.

What explains children's difficulty with Experiment 4, given their robust success in Experiments 1-3? It is unlikely that children's failure is due to their inability to infer task difficulty; the structures used in the stimuli here have been verified to elicit accurate judgments of difficulty among 4-year-olds in prior work (Gweon et al., 2017). Furthermore, we only included children who were accurately able to tell which tower was "harder". One possibility is that children's success in Exp.1 - 3 simply reflects their use of superficial cues associated with "being better", such as one person finishing the task earlier than the other (Experiment 1) or one tower being larger than the other (Experiment 3). By contrast, Experiment 4 required integrating time and task difficulty in the absence of these cues. Some anecdotal support comes from pilot data for Experiment 4 where children were asked both (1) which tower was "harder" and (2) which tower was "better". While 4- and 5-year-olds correctly judged the 10-block tower as "harder" than the 2-block tower, they did not judge this tower as "better". By contrast, most children in Experiment 3 picked the vertical 10-block tower as "better" than the horizontal 10-block tower, suggesting that children relied primarily on perceptual cues such as relative time or size/height to judge who (or what) is "better".

One way to interpret these results is that children's concept of competence is quite different from that of adults, and that it continues to develop beyond age 5. This interpretation is largely consistent with what previous studies have proposed (Heyman et al., 2003; Nicholls, 1978; Yang & Frye, 2016). However, another possibility is that children's failure on Experiment 4 reflects the developmental change in the semantics of "better", rather than a genuine conceptual change in their understanding of competence. If children strongly associate the word "better" with positive perceptual features of objects or agents, this might bias children's judgments of "who's better at bulding" to whoever finishes first, or whoever builds something larger. When these explicit cues don't differ between the two agents, as in Experiment 4, children are thus at chance.

The current study cannot tease apart these possibilities, as the critical test question involves verbally asking children "who is better". Thus, it still leaves open the possibility that children do have an abstract understanding of competence as a subjective quality that is determined by both task-based and agent-based features. One promising future direction is to try eliciting competence judgments without using the word "better". In addition to non-verbal measures, future work might capitalize on a previous finding that toddlers' friend choice reflects representations of agents' competence (Jara-Ettinger et al., 2015).

Despite the limitation of using a verbal prompt in our outcome measure, our stimuli had lower verbal demands relative to earlier work that involved heavy-handed manipulations of competence with explicit verbal information. The words used in these tasks often implied evaluative judgments (e.g., "lazy", "smart", see Heyman et al., 2003; Heyman & Compton, 2006; Nicholls, 1978), raising the possibility that children in these studies succeeded by matching the valence of these words with "being better", instead of engaging in genuine inference based on the features of the event. On the other hand, while verbal cues may help make these features easier to detect, verbally presented scenarios can also hinder performance by increasing processing demands, taxing verbal knowledge and working memory. This may have led to either underestimation or overestimation of children's understanding of competence depending on the paradigm (Nicholls, 1984; Heyman et al., 2003; Yang & Frye, 2016), producing discrepant findings across studies and age ranges. The fact that children in Experiments 1-3 successfully used task-based and agent-based features suggests that young children are adept at picking up on non-verbal cues embedded in observed events to infer relative competence, in addition to using verbal cues (Wimmer et al., 1982; Heyman & Compton, 2006).

While not quite at the level of adults (note that adults are near-ceiling on Experiment 4), children's robust performance on most of these experiments suggests that some basic notion of competence based on quality and efficiency may emerge early in life. Indeed, infants have a sophisticated understanding of physical events (Stahl & Feigenson, 2015) as well as agent's actions and outcomes (Liu, Ullman, Tenenbaum, & Spelke, 2017). Furthermore, infants can use information about other's effortful actions to inform their own (Leonard, Lee, & Schulz, 2017). In order to employ this sort of social learning about effort, children presumably need some basic understanding of how effort relates to outcomes, scales with difficulty, and is constrained by competence. The ability to go beyond superficial cues to infer who is more competent than others can be especially beneficial for early learning, as the learner can make better decisions about who to learn from or ask for help. Future work could further explore when children begin to use task-based and agent-based features using similar stimuli as the current study with nonverbal dependent measures in a younger age range.

An open question is whether young children's inferences about competence generalize to domains outside of physical ability. One possibility is that children develop a stronger sense of physical competence before mental competence, due to its overt perceptual cues and children's more salient experience in this domain early in life. Furthermore, the paradigms tested here only looked at how task-based and agent-based features relate to inferences about competence, yet many other features are surely involved in this calculation. For example, if someone was unmotivated to play basketball and failed to shoot a 3-pointer, we wouldn't necessarily conclude

that they were unskilled. In other words, we also consider how much people *want* to achieve their goals when inferring their competence. Future work should probe the range of additional features that affect competence judgments broadly.

While we show that preschoolers are fairly accurate at reasoning about other's competence (at least with adequate perceptual cues), a great deal of work has shown that young children are out of touch, and in fact overly optimistic, about their own competence. However, most of these studies looked at how children predict how they would do in the future given their past performance, which might have led to wishful thinking (Schneider, 1998; Parsons & Ruble, 1977; Harter, 2012). Just as children were able to use observed evidence to infer others' competence, they may similarly evaluate their own competence based on observed outcomes. In fact, recent work suggests that children are even sensitive to the discrepancy between their own belief about their actual competence (i.e., the child successfully activates a toy after a few failures) vs. others' beliefs (i.e., an adult only observed the child's failures) and demonstrate their success to others to change these beliefs (Asaba & Gweon, 2018). Collectively these results are consistent with the recent proposal that children's understanding of competence is not "irrationally" optimistic (Cimpian, 2017), and calls for better tasks that tap into their underlying cognitive processes.

More generally, this work highlights the importance of reexamining old topics in a new light. The current work conceptually replicates prior results (including some from the 70's) while also raising new questions about what these results mean. Children's perception of competence in the early years is crucial as it likely informs their achievement beliefs and mindsets, which in turn impacts their academic outcomes (e.g. Dweck, 2006). Thus, understanding the ways in which children conceptualize competence early in life allows us to potentially help set children on the path towards a learning-focused mindset even before they enter formal schooling. We hope a new wave of interest from the broader community will shed more light on this important topic.

References

Asaba, M., & Gweon, H. (2018). Look, i can do it! young children forego opportunities to teach others to demonstrate their own competence. In *Proceedings of the 40th annual conference of the cognitive science society.*

Cimpian, A. (2017). Early reasoning about competence is not irrationally optimistic, nor does it stem from inadequate cognitive representations. In A. J. Elliot, C. S. Dweck, & D. Yeager (Eds.), *Handbook of competence and motivation, second edition: Theory and application* (p. 387-407). New York, New York: The Guilford Press.

Dweck, C. S. (2006). *Mindset: The new psychology of success*. New York, New York: Random House.

Dweck, C. S., & Leggett, E. L. (1988). A social-cognitive approach to motivation and personality. *Psychological review*, *95*(2), 256.

Gweon, H., Asaba, M., & Bennett-Pierre, G. (2017). Reverse-engineering the process: Adults' and preschoolers' ability to infer the difficulty of novel tasks. In *Proceedings of the 39th annual conference of the cognitive science society.*

Harter, S. (2012). *The construction of the self: Developmental and sociocultural foundations*. New York: The Guilford Press.

Heyman, G. D., & Compton, B. J. (2006). Context sensitivity in children's reasoning about ability across the elementary school years. *Developmental Science*, *9*(6), 616–627.

Heyman, G. D., Gee, C. L., & Giles, J. W. (2003). Preschool children's reasoning about ability. *Child Development*, *74*(2), 516–534.

Jara-Ettinger, J., Tenenbaum, J. B., & Schulz, L. E. (2015). Not so innocent: Toddlers' inferences about costs and culpability. *Psychological Science*, *26*(5), 633–640.

Leonard, J. A., Lee, Y., & Schulz, L. E. (2017). Infants make more attempts to achieve a goal when they see adults persist. *Science*, *357*(6357), 1290–1294.

Liu, S., Ullman, T. D., Tenenbaum, J. B., & Spelke, E. S. (2017). Ten-month-old infants infer the value of goals from the costs of actions. *Science*, *358*(6366), 1038–1041.

Magid, R. W., DePascale, M., & Schulz, L. E. (2018). Four- and 5-year-olds infer differences in relative ability and appropriately allocate roles to achieve cooperative, competitive, and prosocial goals. *Open Mind*, *1*(4), 194–207.

Nicholls, J. G. (1978). The development of the concepts of effort and ability, perception of academic attainment, and the understanding that difficult tasks require more ability. *Child development*, *49*(3), 800–814.

Nicholls, J. G. (1984). Achievement motivation: Conceptions of ability, subjective experience, task choice, and performance. *Psychological review*, *91*(3), 328.

Parsons, J. E., & Ruble, D. N. (1977). The development of achievement-related expectancies. *Child Development*, *48*, 1075-–1079.

Schneider, W. (1998). Performance prediction in young children: Effects of skill, metacognition and wishful thinking. *Developmental Science*, *1*(2), 291–297.

Stahl, A., & Feigenson, L. (2015). The development of achievement-related expectancies. *Science*, *348*(6230), 91–94.

Stipek, D., & Iver, D. M. (1989). Developmental change in children's assessment of intellectual competence. *Child development*, 521–538.

Wentzel, K. R., & Wigfield, A. (1998). Academic and social motivational influences on students' academic performance. *Educational Psychology Review*, *10*(2), 155–175.

Wimmer, H., Wachter, J., & Perner, J. (1982). Cognitive autonomy of the development of moral evaluation of achievement. *Child Development*, 668–676.

Yang, F., & Frye, D. (2016). Early understanding of ability. *Cognitive Development*, *38*, 49–62.

Algebraic Patterns as Ensemble Representations

Anna Leshinskaya (alesh@sas.upenn.edu) and **Sharon L. Thompson-Schill (sschill@psych.upenn.edu)**
Department of Psychology, University of Pennsylvania
Stephen A. Levin Building, 425 S. University Ave
Philadelphia, PA 19104

Enoch Lambert (elambe03@tufts.edu)
Center for Cognitive Studies, Tufts University
115 Miner Hall
Tufts University
Medford, MA 02155 USA

Abstract

Observers rapidly extract summary statistics from sets of visually presented items, like the mean size of a set of circles, or the mean expression of a set of faces. Their excellent ability to report summary statistics stands in contrast to near-chance representation of any of the individuals. Here we asked to what extent this 'ensemble perception' signature extends to a more abstract property: *relations* among elements. Participants watched ten unique animations of visually patterned objects (hereafter, 'shapes') colliding with each other and producing a new shape. Collisions conformed to ABA patterns, such that the result shape always matched one of the collider shapes. Recognition tests showed that participants accurately recalled the collisions they saw, but also falsely accepted foils which conformed to the ABA pattern but which were not in fact specifically seen (were rearrangements of the original shapes across collisions). On the other hand, they were much less likely to accept foils which did not conform to the pattern, but were equally distinct rearrangements (e.g., AAB). This suggests that participants represented the overall, common pattern better than the specifics of what they saw; the superior encoding of the summary relative to the individuals thus applies to summaries of relations. However, in contrast to prior findings with visual features, we did not find that recall of individual patterns was entirely at chance. Our paradigm offers a way to pursue future questions such as the pressures and motivations which might govern the trade-off between summarizing evidence vs. retaining individual experiences.

Keywords: ensemble perception; artificial grammar learning; pattern recognition; episodic memory; semantic memory

Introduction

Rather than encoding experiences in perfect detail, the mind naturally uses regularities and summary statistics to compress them. We can keep more items in working memory if items are predictive of each other (Brady, Konkle, & Alvarez, 2009); it becomes faster to find images in search display if they appear in predictable spatial configurations (Chun & Jiang, 1998); and we spontaneously and obligatorily register the mean orientation of sets of gabor patches (Parkes, Lund, Angelucci, Solomon, & Morgan, 2001). Although contingencies and averages are distinct statistics, all of these cases demonstrate that we spontaneously compress experiences, by encoding a summary of what is common across them.

In principle, representational systems can differ in the extent to which they compute summaries and discard individual observations (Dennett, 1991). Curiously, human participants sometimes represent summaries better than the observations composing them. When we see sets—like a series of differently-sized circles—we recall their mean (here, size) substantially better than we can recall any particular individual (Ariely, 2001; Chong & Treisman, 2003; Haberman & Whitney, 2009). This is true even when the number of items is relatively small (4) and when the items are presented sequentially. This suggests that we compute summaries and update them rapidly, discarding the items that went into this computation along the way. This 'ensemble perception' signature is true for visual properties like size, orientation, or facial expression (see Alvarez, 2011 and Whitney & Yamanashi Leib, 2018 for reviews). Here we asked whether this signature also applies to a property which is not a visual feature, but rather an abstract rule.

Algebraic rules (Marcus, 2001) are patterns based on relations among elements, such as *same* and *different*. For example, triplets of syllables can be readily seen as belonging to patterns like ABA— "ga di ga", "ku la ku", or "do re do"—vs AAB—"ga ga di". Learners (adults or infants) can recognize such patterns even with entirely distinct syllable sets and in both auditory and visual modalities (Ferguson, Franconeri, & Waxman, 2018; Marcus, Vijayan, Rao, & Vishton, 1999; Saffran, Pollak, Seibel, & Shkolnik, 2007). Algebraic rules are hallmarks of relational thinking, requiring relatively advanced computational architecture (Marcus, 2001; Overlan, Jacobs, & Piantadosi, 2017). They are also excellent compressions: recognizing that the last element always matches the first reduces the number of bits needed to represent the triplet by 1/3. Thus, despite the possible computational cost, encoding relations among stimuli is adaptive for circumventing limited memory capacity.

Here we asked how a representation of a shared algebraic pattern relates to the representation of the diverse individuals exhibiting the pattern. Specifically, we asked whether we would see the signature of ensemble perception. If so, participants should not only recognize that the set of items tends to follow an ABA pattern, but they should find it easier to recall that abstract pattern than the particular

items they specifically saw (for instance, “ga di ga”, but not “ga ku ga”). Alternatively, due to the computational

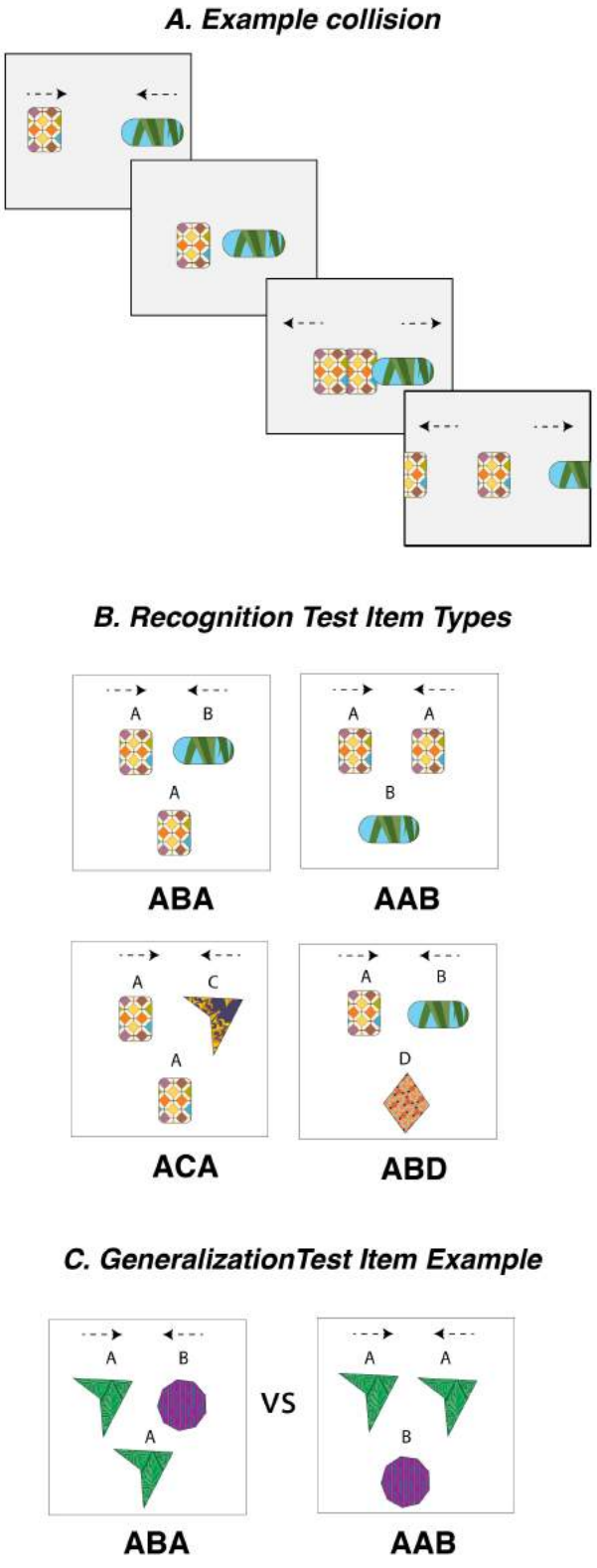


Figure 1. Top: Example of a collision between an A and B shapes and an A shape result. Bottom: Foil types for this collision, represented schematically. C shapes were taken from other collisions presented during the demonstration.

demands of inferring algebraic patterns, participants may be less reliable at recognizing the common ABA pattern across items, failing to summarize the data, and may be better at discerning which individual items they specifically saw.

We used a novel paradigm in which participants watched pairs of novel shapes collide with each other and ‘produce’ a third shape (Figure 1). Participants’ task was to watch the collisions and see how many they could remember. As there was no instruction to look for patterns, the choice to summarize or not had to be intrinsically motivated.

Each collision was in fact governed by an ABA pattern: two distinct shapes, A and B, collided, producing another A. (We use ABA to denote the abstract elements, and lowercase *aba* to denote specific shapes). After watching 10 unique ABA collisions once each, participants performed a recognition test where each test item was either a collision they had seen, or one of three kinds of foils (unseen collisions). ACA foils swapped shapes between collisions: if specific collisions *aba* and *dcd* were shown, foils were *aca* and *dbd*. AAB foils were rearrangements of the same shapes in seen collisions, so that two A’s collided to produce a B.

We reasoned that if participants recalled the common pattern better than the individual items, they should accept ACA foils at a higher rate than AAB foils. This is because ACA is pattern-consistent while AAB is not, though in terms of individual shapes composing the collisions, ACA is in fact more different from the original. ABD foils were also used as these were equal in the number of element-wise changes from ABA as ACA, but were also pattern-inconsistent.

We were also able to ask whether participants recalled *only* the summary pattern, and lost all item representations, by seeing whether they accept ACA foils at the same rate as ABA correct items. In ensemble perception, tests of individual recognition are often at chance (Ariely, 2001; Haberman & Whitney, 2009). Finally, a forced-choice test with new items directly tested whether learners represented the pattern in generalizable form.

Methods

Participants

30 participants were recruited and tested via Amazon Mechanical Turk. Participants provided electronic consent and procedures were approved by the Institutional Review Board of the University of Pennsylvania. Compensation was $2. Three participants were excluded for failing an attention measure, and one for missing data. The included sample had 15 females and 15 males, with age M = 37, range 21 – 64). The task took an average of 15.62 minutes.

Stimuli

Stimuli were animated shape collisions (Figure 1). In each animation, two shapes approached each other from the left and right sides of the screen, met in the middle, and a third

'result' shape appeared between them as they moved away. Individual frames were created in Adobe Illustrator and concatenated into GIFs. Each GIF was composed of 23 frames shown at a 180 ms framerate and 4.14 s duration. GIFs were interspersed with 660 ms of blank screen for a 4.8 s total ISI. The majority of shapes used in the displays is shown in the Appendix.

Procedure

The task was presented to participants using a custom JavaScript webpage. It began with a *demonstration phase.* Participants were shown the following instructions: "You will play a game where you will see pairs of shapes collide with each other and see how many you can remember." They then watched 10 unique demonstration collisions in randomized order (lasting ~ 1 minute). Each of these collisions followed an ABA pattern: the two collider shapes, A and B, were distinct, and the result shape was a duplicate of A. A total of 20 different shapes were used, so that no shapes were repeated across collisions.

They then saw the *specific recognition test*. On each trial, a collision was shown, and participants had to decide whether or not they had seen it in the demonstration phase, by clicking 'yes' or 'no' after it ended. They were allowed to replay the collision. Apart from all 10 demonstration collisions, test items also showed three types of foils, created by rearranging the shapes across or within the demonstration collisions. ACA foils swapped the 'B' shapes between two different collisions, so that if specific shapes *aba* and *dcd* had been shown, foils were *aca* and *dbd*. AAB foils rearranged the shapes within an original collision, so that now two A's collided to produce B. ABD foils produced a result shape taken from another collision. The swaps were selected by pairing the 10 collisions into 5 foil-pairs. One of each of the three foil types was shown for each of the 10 original collisions; thus, there were 10 data points for each participant for each test item type.

We also added three attention check items, which showed previously unseen shapes in which two of the same shape collided, producing another duplicate (i.e., an AAA pattern). Participants had to respond 'no' to all three attention items to be included in the further analyses. Overall, there were 43 specific recognition test trials, shown in randomized order. There was no trial-level feedback, but an overall score was shown at the end of the test.

Participants were then given the *generalization test*. The instructions read, "The collisions you first watched followed certain patterns or rules. Now you will see new collisions and be asked to decide which ones follow similar patterns or rules." A two-alternative forced-choice test asked them to choose between pairs of collisions, shown one at a time, side by side. We used previously unseen shapes to create two new sets of ABA, AAB, and ABD items. Critical questions asked participants to choose between a pattern-consistent collision (ABA) and one of the two foils (AAB or ABD). Filler items showed the two foils, in order to balance the number of times each collision was shown overall. Each question type was shown once for each novel shape set, creating a total of 8 trials.

Finally, we asked participants whether or not they took any notes during the task. No participant reported taking notes.

Results

Specific Recognition Test

We computed the percent acceptance rate ('yes' response) for each type of test item; results are shown in Figure 2. The correct test item (ABA) was identical to the collision previously shown; this was (correctly) accepted at a high rate (M = 85%, SE = 0.05%). The ACA foil item maintained the pattern but its middle shape was swapped across previously seen collisions; this was (falsely) accepted at a high rate (M = 73%, SD = 0.05%). The AAB foil item was accepted at a low rate (M = 15%, SE = 0.06%) as was the ABD foil item (M = 13%, SE = 0.04%).

A 4- way ANOVA indicated a significant effect of item type, $F(75,3) = 59.43$, $p < .001$. Planned t-tests were used to probe these differences pairwise. We found that ACA foils were accepted at a higher rate than AAB foils, $t(25) = 6.77$, $p < .001$, CI [40 75] and ABD foils, $t(25) = 6.16$, $p < .001$, CI [42 86], indicating that participants indeed represented the pattern better than the specifics. Nonetheless, we also found higher acceptance rates for the correct (ABA) items than the ACA foils, $t(25) = 3.30$, $p = .002$, CI [4 19], indicating that item information was not completely lost.

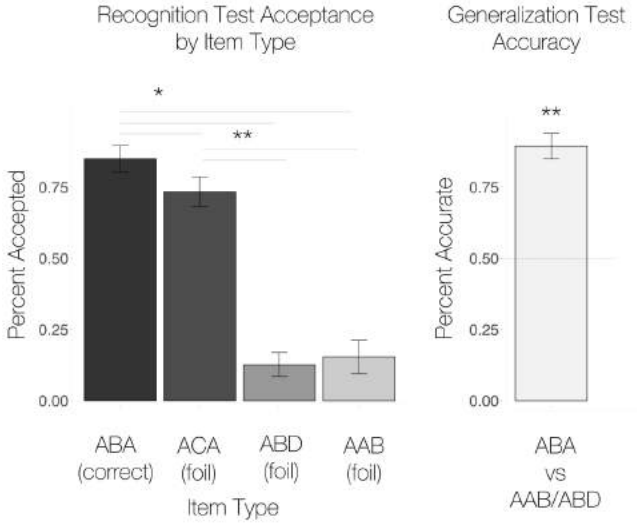


Figure 2. Rate of acceptance on the specific recognition test for each type of test item, and accuracy on the generalization test. Statistical comparisons are shown with * indicating $p < .01$ and ** indicating $p < .001$.

One potential account of these effects is that individual item representations decayed more rapidly or were more susceptible to interference from the question presentations. It should be noted that the majority of test items are also pattern-inconsistent, and so the amount of interference

should be equal for both specific and pattern recall; however, it could still be that the item representations are more susceptible. We thus looked at responses from the first 10 test items only (thus, a random subsample from each participant). The critical effect of greater ACA than AAB acceptance was still significant, $t(25) = 7.10$, $p < .001$, CI [45 82], and the difference between ABA and ACA was marginal, $t(25) = 1.89$, $p = .07$, CI [-1 28].

Generalization Test

Participants were reliably above chance on choosing the pattern-consistent, entirely novel ABA collision relative to both foils; AAB: $M = 89\%$, $SE = 0.05\%$; $t(25) = 7.63$, CI [78 99], $p < .001$; ABD: $M = 90\%$, $SE = 0.05\%$; $t(25) = 8.38$, CI [81 100], $p < .001$ (Figure 2). They were thus highly reliable in learning a generalizable representation of the algebraic rule. As there was no difference between the two foil types ($t < 1$), accuracies for both were collapsed into a composite generalization score ($M = 89\%$). We found that this composite accuracy was not significantly different from the rate at which participants accepted the correct ABA items on the specific recognition test ($t < 1$), indicating that the representation of the abstract pattern was no worse than specific recall. We also found that accuracy on the generalization test was substantially higher than participants' ability to accurately reject the ACA foils (i.e., inverse of their acceptance rate; $t(25) = 7.45$, CI [47 82], $p < .001$. This is in line with the findings from the specific recognition test that the representation of the abstract pattern was superior to specific recall.

Discussion

We investigated the relationship between the ability to recall specific items (unique collisions of three shapes) and to identify and recall the common pattern governing them (here, an ABA algebraic rule). We found that analogously to signatures in ensemble perception, participants recalled the common pattern substantially better than the specifics of the individual items. Nonetheless, some memory of the individuals persisted, in contrast to certain findings with visual feature ensembles.

Our results indicate that the core signature seen in ensemble perception—superior fidelity of summary statistics over individual items—generalizes beyond visual features like size, facial expression, or line orientation (Whitney & Yamanashi Leib, 2018) and similarly applies to relational properties over visual events, like algebraic rules. This substantially extends the repertoire where such ensemble signatures might be found.

Our findings also speak to the question of how much a pattern-based summary relies on the representation of the individuals being summarized. Individual items must of course be processed at *some* level, but showing that their details can be quickly forgotten in spite of near-ceiling summary representations suggests that this level is relatively minimal. Because items were shown sequentially, and were short-lived, learners had to encode the pattern and update the summary with each subsequent representation—otherwise, it would be too late. It could therefore be the case that the item representation is discarded almost immediately after it is perceived.

The literature on episodic memory has similarly investigated whether summary recall is dependent on item recall, and has separated out these representations using delay paradigms and studies of amnesia. With multi-day delays, animals' reliance on the locations of specifically experienced platforms in a water maze declines, and is replaced by a representation of their mean location (Richards et al., 2014). Patients with amnesia (impairment to episodic memory) are as able as controls to extract patterns in artificial grammar learning studies, but unlike them, fail on recognition tasks of individual items from which they learned that grammar (Knowlton, Ramus, & Squire, 1992). Here we offer an elegant way to show this dissociation in healthy participants within a few minutes of testing, and to directly quantify the amount of information preserved about the individual items and the overall patterns. This opens an avenue of research investigating the circumstances and pressures that may motivate our cognitive system to rely on one or the other.

What might such pressures be? If learning is an attempt to infer the underlying model that generates observations, specific experiences serve as evidence towards hypotheses about that model—for example, a mean value or an underlying structure (Tenenbaum, Kemp, Griffiths, & Goodman, 2011). Once their beliefs over relevant models are updated, data points could be discarded (Nagy, Török, & Orbán, 2018). In this light, it has been argued that the choice to update a model vs. keep the data may be informed by factors such as the number of relevant models or how likely the relevant model to update might eventually change (Nagy et al., 2018; Richards & Frankland, 2017). Here, the right model was the ABA pattern, which explained all observations reliably. We might predict that if the pattern sometimes changed, recalling the specifics of all collisions might be enhanced, as this suggests to the learner that the model may not tell the full story or might change. We plan to test this in future work.

Another factor may be the computational cost of that update. Representing items in terms of their relations may be inferentially complex (Frank & Tenenbaum, 2011; Kuehne, Gentner, & Forbus, 2000; Overlan et al., 2017) and appears optional: one could perceive and remember a specific collision without ever representing the relations among its elements. If hypotheses about relations are computationally costly to update, the compression benefit of computing a relation may not outweigh the costs. The qualitative divergence we saw between algebraic patterns here vs. visual features in the past is consistent with this possibility: in the case of algebraic patterns, representations of individuals were not entirely lost, while for visual feature summaries, they often are (Ariely, 2001; Haberman & Whitney, 2009). If visual features require fewer inferential steps to encode than relational patterns, this could be consistent with that idea. Our paradigm offers a way to test some of these questions directly in future work.

Acknowledgments

This work was supported by NIH grant R01DC015359 to S.L.T-S. This project was also made possible through the support of a grant from the John Templeton Foundation, to A.L, E.L., and S.L.T-S. The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the John Templeton Foundation.

References

Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition. *Trends in Cognitive Sciences*, *15*(3), 122–131.

Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological Science*, *12*(2), 157–162.

Brady, T. F., Konkle, T., & Alvarez, G. A. (2009). Compression in visual working memory: Using statistical regularities to form more efficient memory representations. *Journal of Experimental Psychology: General*, *138*(4), 487–502.

Chong, S. C., & Treisman, A. (2003). Representation of statistical properties. *Vision Research*, *43*(4), 393–404. https://doi.org/10.1016/S0042-6989(02)00596-5

Chun, M. M., & Jiang, Y. (1998). Contextual cueing: Implicit learning and memory of visual context guides spatial attention. *Cognitive Psychology*, *36*(1), 28–71.

Dennett, D. C. (1991). Real patterns. *The Journal of Philosophy*, *88*(1), 27–51.

Ferguson, B., Franconeri, S. L., & Waxman, S. R. (2018). Very young infants learn abstract rules in the visual modality. *PLoS ONE*, *13*(1), 1–14.

Frank, M. C., & Tenenbaum, J. B. (2011). Three ideal observer models for rule learning in simple languages. *Cognition*, *120*(3), 360–371.

Haberman, J., & Whitney, D. (2009). Averaging facial expression over time. *Journal of Vision*, *9*(2009), 1–13.

Knowlton, B. J., Ramus, S. J., & Squire, L. R. (1992). Intact artificial grammar learning in amnesia: Dissociation of category-level knowledge and explicit memory for specific instances. *Psychological Science*, *3*(3), 172–179.

Kuehne, S., Gentner, D., & Forbus, K. (2000). Modeling infant learning via symbolic structural alignment. *Proceedings of the twenty-second annual conference of the cognitive science society*, 286–291.

Marcus, G. F. (2001). *The Algebraic Mind*. Cambridge, MA: MIT Press.

Marcus, G. F., Vijayan, S., Rao, B., & Vishton, P. (1999). Rule learning by seven-month-old infants. *Science*, *283*(January), 77–80.

Nagy, D. G., Török, B., & Orbán, G. (2018). Semantic compression of episodic memories. *ArXiv Preprint ArXiv:1806.07990.*

Overlan, M. C., Jacobs, R. A., & Piantadosi, S. T. (2017). Learning abstract visual concepts via probabilistic program induction in a Language of Thought. *Cognition*, *168*, 320–334.

Parkes, L., Lund, J., Angelucci, A., Solomon, J. A., & Morgan, M. (2001). Compulsory averaging of crowded orientation signals in human vision. *Nature Neuroscience*, *4*(7), 739–744.

Richards, B. A., & Frankland, P. W. (2017). The persistence and transience of memory. *Neuron*, *94*(6), 1071–1084.

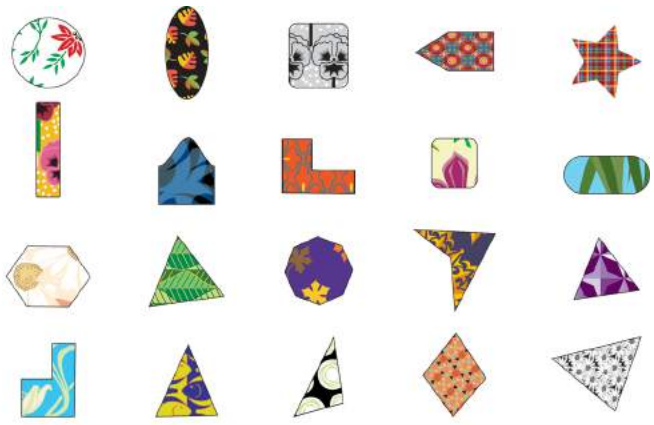
Richards, B. A., Xia, F., Santoro, A., Husse, J., Woodin, M. A., Josselyn, S. A., & Frankland, P. W. (2014). Patterns across multiple memories are identified over time. *Nature Neuroscience*, *17*(7), 981–6.

Saffran, J. R., Pollak, S. D., Seibel, R. L., & Shkolnik, A. (2007). Dog is a dog is a dog: Infant rule learning is not specific to language. *Cognition*, *105*(3), 669–680.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: statistics, structure, and abstraction. *Science*, *331*(6022), 1279–85.

Whitney, D., & Yamanashi Leib, A. (2018). Ensemble perception. *Annu. Rev. Psychol*, *69*, 12.1-12.25.

Appendix



Parents Calibrate Speech to Their Children's Vocabulary Knowledge

Ashley Leung, Alexandra Tunkel, and Daniel Yurovsky
{ashleyleung, aetunkel, yurovsky}@uchicago.edu
Department of Psychology
University of Chicago

Abstract

Young children learn language at an incredible rate. While children come prepared with powerful statistical learning mechanisms, the statistics they encounter are also prepared for them: Children learn from caregivers motivated to communicate with them. Do caregivers modify their speech in order to support children's comprehension? We asked children and their parents to play a simple reference game in which the parent's goal was to guide their child to select a target animal from a set of three. We show that parents calibrate their referring expressions to their children's language knowledge, producing more informative references for animals that they thought their children did not know. Further, parents learn about their children's knowledge over the course of the game, and calibrate their referring expressions accordingly. These results underscore the importance of understanding the communicative context in which language learning happens.

Keywords: parent-child interaction; language development; communication

Introduction

Children learn language at astonishing rates, acquiring thousands of words by the time they are toddlers. How do children learn so many words before they know how to dress themselves? One account for children's rapid language acquisition is statistical learning. Young children can attend to the distributional structure of language, learning to discriminate words and identify word order from speech streams (Saffran, 2003; Saffran, Aslin, & Newport, 1996). Statistical learning can be a powerful tool for early language learning, and showcases the ability that children have to harvest information from their surroundings. However, the particular structure of children's language environments may also play a role in supporting language development.

The way we speak to children often differs from the way we speak to adults. Child-directed speech (CDS) exists across cultures, and is characterized by higher pitches and more exaggerated enunciations when compared to adult-directed speech (ADS) (Cooper & Aslin, 1990; Grieser & Kuhl, 1988). Not only do children prefer CDS over ADS, CDS is also a better predictor for language learning than overheard ADS (Shneidman, Arroyo, Levine, & Goldin-Meadow, 2013). CDS does not only differ from ADS in prosodic features- the structural qualities of CDS make speech segmentation and word learning easier (Thiessen, Hill, & Saffran, 2005; Yurovsky, Yu, & Smith, 2012). While children live in the same physical environments as adults, their *language environments* contain specific types of input that facilitate early language learning.

Children's language environments are not only suited for their abilities; they also change across development. Parents play a role in changing their children's language environment, and there is evidence suggesting that these changes aid language development. Parents use simpler, more redundant language when talking to toddlers, and more complex syntactic structures when speaking with school-aged children (Snow, 1972). Importantly, sensitive modification of parent response shapes language learning in children (Hoff-Ginsberg & Shatz, 1982; Tamis-LeMonda, Kuchirko, & Song, 2014).

Why do parents modify the way they speak according to their children? One possible explanation is that parents are actively teaching their children. Indeed, some have posited that CDS is an ostensive cue for social learning, and that infants are born prepared to attend to these cues (Csibra & Gergely, 2009). While it may be true that parents hope to impart knowledge to their children, we argue that effective communication is the proximal goal. The field of linguistics has long established that adults communicate in ways that are efficient. Grice's (1975) maxim of quantity states that speech should be as informative as necessary, and no more. Adults are able to adhere to these maxims, adapting speech according to conversational partners' knowledge as needed for successful communication (Clark & Wilkes-Gibbs, 1986). We argue that the parent's goal to communicate with their child drives the change in language use. Specifically, parents adapt their speech according to their children's language abilities.

Parents modify their language as a *means* to achieve successful communication. Research show that parents use simpler language and are more linguistically aligned with their younger children, and these patterns of speech change as their children develop (Snow, 1972; Yurovsky, Doyle, & Frank, 2016). Parents are also sensitive to children's vocabulary knowledge, and the way they refer to objects change markedly depending on whether they are novel, comprehended, or familiar to their children (Masur, 1997). These changes in parent speech may indicate adaptations that are aimed at fulfilling the goal of effective communication, and that the language necessary to fulfill that goal changes as children develop.

Based on work by Masur (1997), we developed a study to investigate how parents adapt their speech according to

their children's vocabulary knowledge. Masur's study involved parents and children engaging in unstructured free play, and parents reported their children's vocabulary knowledge after the session. Our study uses a structured interactive game that allows us to control for the amount and type of stimuli presented to the parent-child dyads, and parent-reported vocabulary measures are collected before the study. Our paradigm also introduces a communicative goal within a structured game, which also allows parent utterances to be more comparable across dyads.

We designed an interactive iPad game in which parents verbally guide their children to select animals on an iPad. Each animal in the game appeared as a target twice. We predicted that parents would modify their speech based on their beliefs about their children's vocabulary knowledge. Specifically, we predicted: (1) Parents should use shorter referring expressions when describing animals that they believe their children know, and (2) Upon the second appearance of an animal, parents would adapt the length of their referring expression according to whether the child responded accurately on the first appearance of the animal.

Method

Participants

Toddlers (aged 2.0 to 2.5 years) and their parents were recruited from a database of families in the local community or approached on the floor of a local science museum in order to achieve a planned sample of 40 parent-child dyads. A total of 46 parent-child pairs were recruited, but data from six pairs were dropped from analysis due to experimental error or failure to complete the study. The final sample consisted of 40 children aged 2.02 to 2.48 years ($M = 2.17$), 20 of whom were girls.

Stimuli

Eighteen animal images were selected from the Rossion & Pourtois (2004) image set, which is a colored version of the Snodgrass & Vanderwart (1980) object set. Animals were selected based on age of acquisition (AoA), using data from WordBank (Frank, Braginsky, Yurovsky, & Marchman, 2017). The AoA of the selected animals ranged from 12 to 31 months. Half of the animals had lower AoA (12-20 months), and the other half had higher AoA (25-31 months). Each trial featured three animals, all from either the low AoA or high AoA category.

A modified version of the MacArthur-Bates Communicative Development Inventory (CDI; Fenson et al., 2007), a parent-reported measure of children's vocabulary, was administered before the testing session via an online survey. The selected animal words were embedded among the 85 words in the survey. Two of the animal words–one in the early AOA and one in the late AOA category–were accidentally omitted, so trials for those words were not included in analysis.

Figure 1: Example iPad screens for the child (top) and parent (bottom) during the experiment.

Design and Procedure

Each parent-child pair played an interactive game using two iPads. Children were given two warm-up trials to get used to the iPads. The practice and experimental trials began after the warm-up. On each trial, three images of animals were displayed side by side on the child's screen, and a single word appeared on the parent's screen (Figure 1). Parents were instructed to communicate as they normally would with their child, and encourage them to choose the object corresponding to the word on their screen. The child was instructed to listen to their parent for cues. Once an animal was tapped, the trial ended, and a new trial began. There was a total of 36 experimental trials, such that each animal appeared as the target twice. Trials were randomized for each participant, with the constraint that the same animal could not be the target twice in a row. Practice trials followed the same format as experimental trials, with the exception that images of fruit and vegetables were shown. All sessions were videotaped for transcription and coding.

Results

The data of interest in this study were parent utterances used during the interactive game and parents' responses on the adapted CDI. Transcripts of the videos were analyzed for length of referring expressions. We measured the length of parents' referring utterances as a proxy for amount of information given in each utterance. Parent utterances irrelevant to the iPad game (e.g. asking the child to sit down) were not analyzed. Children's utterances were coded when audible, but were not analyzed.

Word difficulty. We first confirm that the animals predicted be later learned were less likely to be marked known by the parents of children in our studies. As predicted, animals in the

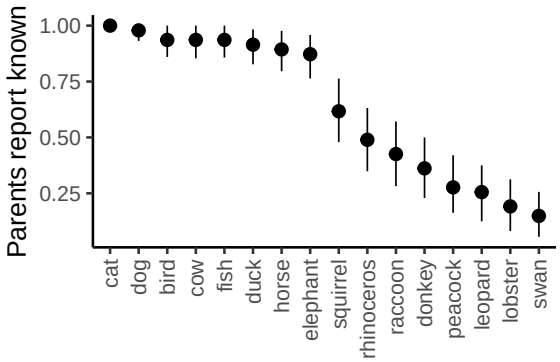


Figure 2: Proportion of parents who reported that their child understood the word for each of our target animals. Error bars indicate 95% confidence intervals computed by non-parametric bootstrap.

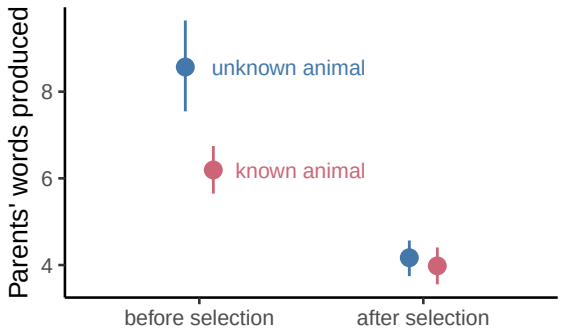


Figure 3: Length of parents' references before and after their child selected a target animal. Points indicate means, error bars indicate 95% confidence intervals computed by non-parametric bootstrapping.

early AoA category were judged to be understood by 93% of parents, and items in the late AoA category were judged understood by 35%.

The difference between these groups was confirmed statistically with a logistic mixed effects regression with a fixed effect of AoA type and random effects of participants. The late AoA items were judged known by a significantly smaller proportion of parents (β = -5.49, t = -11.22, $p < .001$). Parents' judgments for each target word are shown in Figure 2.

Length of referring expressions. If parents calibrate their referential expressions to their children's linguistic knowledge, they should provide more information to children for whom a simple bare noun (e.g. "leopard") would be insufficient to identify the target. Parents did this in a number of ways: With one or more adjectives (e.g., "the spotted, yellow leopard"), with similes (e.g., "the one that's like a cat"), and with allusions to familiar animal exemplars of the category. In all of these cases, parents would be required to produce more words. Thus, we analyzed the length of parents' referential expressions as a theory-agnostic proxy for informativeness.

We predicted that parents should produce more informative–and thus longer–referring expressions to refer to animals that they thought their children did not know. We divided every trial of the game into phases: The time before a child selected an animal, and the time following selection until the start of the next trial. Figure 3 shows the number of words that parents produced to refer to animals that they believe their children know versus those they believe their children do not know–both before their children selected an animal and after. In line with our prediction, parents produced significantly longer referring expressions when talking about animals that they believe their children do not know. However, once the child had selected an animal, the expressions that followed did not differ between known and unknown animals.

We confirmed this result statistically, predicting number of words from a mixed effects model with fixed effects of phase and animal knowledge and their interaction, and random effects of participant and item. In this and all future models, we analyzed the number of words on a log scale as that improved model fit, but results are qualitatively similar when raw number of words was the dependent variable. Phase and the interaction of phase and knowledge were significant: Parents produced fewer words after selection (β = -0.51, t = -13.16, $p < .001$), and when the animal was known, (β = -0.21, t = -6, $p = < .001$), but the change was smaller for known animals (β = 0.08, t = 1.61, p = .107). In the remainder of our analyses, we focus on utterances in the pre-selection phase of each trial as the post selection phase did not vary across trial targets.

Although each parent only gave a single bit of information about each animal–whether they thought their child knew it or not–we pooled these judgments across parents to estimate a continuous measure of difficulty (Figure 2). If parents' referring utterances reflect a sensitivity to this continuous difficulty, the length of their referring expressions should vary smoothly with the difficulty of words. Figure 4 shows this relationship, which was confirmed by a mixed effects model predicting length from fixed effects of difficulty and animal knowledge, and random effects of subject and trial target. Referring expressions were reliably longer for more difficult animals (β = 0.2, t = 2.63, p = .012), over and above the increase for unknown animals (β = 0.14, t = 3.05, p = .002)

We then tested our second hypothesis: Parents should modify their productions over the course of the experiment as they obtain evidence about their children's knowledge. Because each animal was the target twice, parents could use their children's selection on the first appearance of the animal to inform their referential expressions on the second appearance. Figure 5 shows the length of parents' referring expressions as a function of their prior belief about their children's knowledge and their children's selection on the first appearance of the target animal. As predicted, parents who thought their children knew an animal, but who observed evidence that

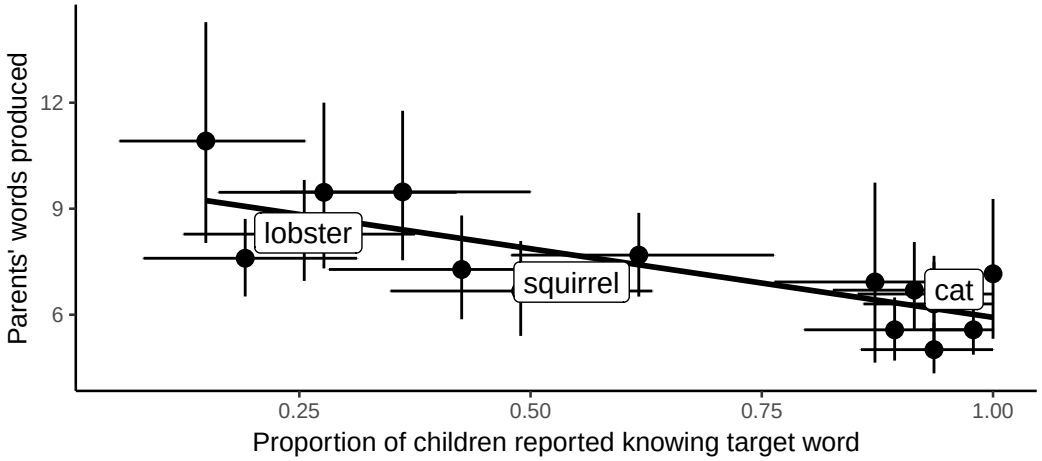


Figure 4: Number of words in parents' referential expressions as a function of the proportion of children reported to know the word for target animal. Points show group averaged proportions, error bars show 95% confidence intervals computed by non-parametric bootstrap.

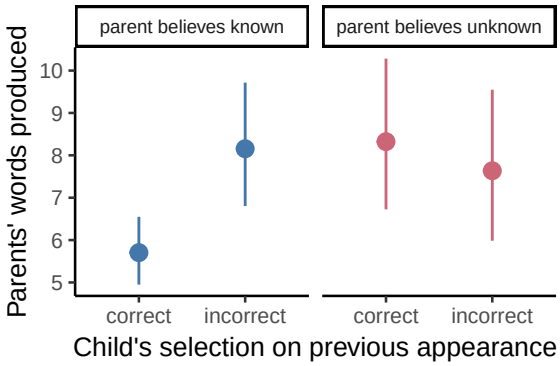


Figure 5: Length of parents' referring expressions on the second appearance of each animal. Points show group averaged proportions; error bars show 95% confidence intervals computed by non-parametric bootstrap.

they didn't (i.e. their children selected the wrong animal), lengthened their referring expressions on its second appearance. Parents who thought their children did not know an animal before the start of the game did not shorten their referring expressions if their children were correct the first time. We cannot say definitively why their referring expressions do not change in length, but one likely explanation is that the references that lead to success the first time were heavily scaffolded and may not even have contained the animal's canonical label (e.g. "the one that looks like a cat" for leopard). We confirmed these results with a mixed effects model predicting length of expressions from parents' prior beliefs, their children's selection on the first trial, and their interaction. We found only the interaction to be significant: References were not reliably longer when parents thought their children did not know the animal ($\beta = 0.28$, $t = 4.14$, $p = < .001$), nor when the children were incorrect on the previous trial ($\beta = 0.27$, $t = 3.82$, $p = < .001$, but only when the parent thought

term	estimate	t-value	p-value
intercept	3.10	4.29	$< .001$
length (log)	-1.34	-2.53	.011
unknown	-3.06	-3.07	.002
second appearance	-0.18	-1.06	.288
trial number	0.01	1.00	.317
length * unknown	1.39	1.88	.061

Table 1: Coefficient estimates for a mixed-effects logistic regression predicting children's success in selecting the target animal. The model was specified as `correct` $\sim$ `log(length) * unknown + appearance + trial + (1|subj) + (1|animal)`.

their children did not know the animal and their children were incorrect on the previous trial ($\beta = -0.44$, $t = -4.29$, $p = < .001$).

Children's selections. Overall, children performed significantly above chance for both low AoA and high AoA trials. In our previous analyses, we showed that parents calibrated the length of their referring expressions to their beliefs about their children's knowledge. They did this both in response to their prior beliefs (Figure 3), and their in-game observations of their children's knowledge (Figure 5). In our final analyses, we asked whether this mattered for children's selections. Are children more likely to succeed in the task when parents provide well calibrated utterances? We asked this question by predicting children's selection trial by trial from a mixed effects logistic regression with fixed effects of parents' prior beliefs about children's knowledge of the target animal, whether the trial was the first or second appearance of the the target animal, the length of parents' referring expressions, and the interaction of parents' prior beliefs and the length of their ex-

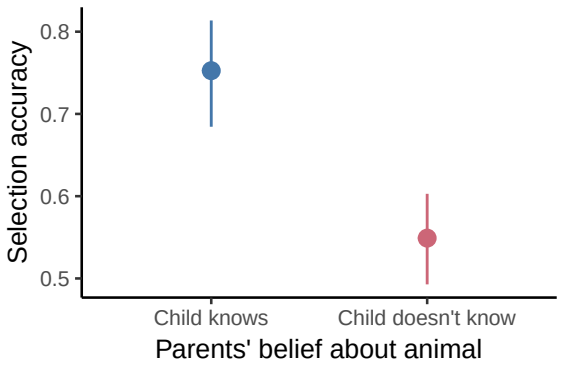


Figure 6: Children's accuracy at selecting both known and unknown animals. Points indicate means, error bars indicate 95% confidence intervals computed by non-parametric bootstrapping.

pressions, as well as random effects of subject and trial target. Children were more likely to be correct when their parents produced longer references, but only for animals that their parents believed that they did not know. Thus, parents' informative references to unknown animals did appear to be supporting successful communication of the target animal. Table 1 shows coefficient estimates for all parameters.

Discussion

Parents have a wealth of knowledge about their kids, including their linguistic development (Fenson et al., 2007). Do they draw on this knowledge when they want to communicate? In a referential communication task, we showed that parents speak differently depending on their beliefs about their children's vocabulary knowledge. Specifically, they produce shorter, less informative expressions to refer to animals that they believe their children know relative to animals that they think their children do not know. Further, parents update their beliefs during the course of the task, producing more informative expressions on the second appearance of an animal they previously thought their children knew if they observed evidence to the contrary (i.e. when children selected the wrong animal). We further found that more informative referring expressions were associated with increased likelihood of successful communication: Children were more likely to correctly select animals whose names they did not know if their parents produced longer utterances to refer to them. We leveraged length as a proxy for informativeness in parents' expressions in the service of quantitative, theory-agnostic predictions. In ongoing work, we are analyzing *how* parents succeed on these trials, and investigating whether different strategies lead to different levels of success.

In general, communicative success was high. Children selected the correct animal at above chance levels, even for targets whose names their parents thought they did not know. Because easy and hard animals appeared on separate trials, children's high accuracy in selecting unfamiliar animals is unlikely to be due to the use of strategies like mutual exclusivity (Markman & Wachtel, 1988). Instead, parents must have produced sufficient information for their children to find the correct target. Taken together with our finding that parents used longer sentences for words they think their children do not know, our results suggest that parents modified their speech as a means to communicate.

Our proposed explanation for these results is that they are produced by a pressure for effective communication: Parents need to produce sufficient information for their children to understand their intended meaning. That is, parents design their utterances for their children's benefit (speaker-design, Jaeger, 2013). It could be instead that these utterances reflect pressure from speaking itself. For example, length of parents' utterances may reflect their difficulty in retrieving certain animal words (MacDonald, 2013). We find this explanation unlikely given that parents were given the target words in written form on their iPad, essentially eliminating retrieval problems (Wingfield, 1968). The fact that parents are using long and short referring expressions depending on their beliefs about children's vocabulary knowledge suggests that they are calibrating to their children.

It is important to note that our current results do not rule out the possibility that parents are engaging in pedagogy. Parents may be using longer referring expressions because they wish to teach their children certain words, and this could potentially explain why parents use longer references for words they believe their children do not know. To understand the motivations behind long and short utterances, we are currently analyzing the content of parents' speech. Preliminary qualitative analysis shows that parents use more adjectives on trials where they believe their children do not know the target word (e.g. "Pick the red lobster" instead of "Pick the lobster"). The use of adjectives on these trials may reflect an intention to teach children about a certain animal, but it could also indicate a pressure to communicate effectively. In the lobster example, the color "red" is likely a helpful cue for children, and parents may be using adjectives as a way to help children select the correct target quickly. While our current findings do not allow us to distinguish between the pedagogical and communicative hypotheses, we hope that further analysis of parents' speech will help us differentiate the two accounts.

Our work contributes to the current literature on parent-child interaction, and forms the basis for further experimental work examining the influences that parent speech has on children's language development. In line with Masur (1997), our findings provide evidence that parents calibrate speech sensitively to their children's vocabulary knowledge. These results are important in light of previous work suggesting that parent responsiveness and sensitivity shape the way young children learn language (Hoff-Ginsberg & Shatz, 1982; Tamis-LeMonda et al., 2014). Furthermore, we propose that parents are modifying their speech as a means to communicate, and that communicative intent shapes the language environ-

ments children experience. Further qualitative analysis of our dataset will shed light onto the characteristics of parent-child communication that are helpful for language acquisition.

Finally, this study highlights the importance of studying the parent-child pair as a unit, rather than viewing children as isolated learners: both parents and children contribute to the process of language development (Brown, 1977; Hoff-Ginsberg & Shatz, 1982). Focusing on the interactive and communicative nature of language captures a more realistic picture of children's language environments: The input that children receive is not random – it is sensitive to their developmental level.

All code for these analyses are available at `https://github.com/ashleychuikay/animalgame`

Acknowledgements

This research was funded by a James S. McDonnell Foundation Scholar Award to DY.

References

Brown, R. (1977). Introduction. In C. E. Snow & C. A. Ferguson (Eds.), *Talking to children: Language input and interaction*. Cambridge, MA.: MIT Press.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, *22*, 1–39.

Cooper, R. P., & Aslin, R. N. (1990). Preference for Infant-Directed Speech in the First Month after Birth. *Child Development*, *61*(5), 1584–1595.

Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in Cognitive Sciences*, *13*(4), 148–153.

Fenson, L., Marchman, V. A., Thal, D. J., Dale, P. S., Reznick, J. S., & others. (2007). *MacArthur-bates communicative development inventories: User's guide and technical manual*. Baltimore, MD: Brookes.

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2017). Wordbank: An open repository for developmental vocabulary data. *Journal of Child Language*, *44*(3), 677–694.

Grice, H. P. (1975). Logic and conversation. *1975*, 41–58.

Grieser, D. A. L., & Kuhl, P. K. (1988). Maternal Speech to Infants in a Tonal Language: Support for Universal Prosodic Features in Motherese. *Developmental Psychology*.

Hoff-Ginsberg, E., & Shatz, M. (1982). Linguistic input and the child's acquisition of language. *Psychological Bulletin*, *92*(1), 3–26.

Jaeger, T. (2013). Production preferences cannot be understood without reference to communication. *Frontiers in Psychology*, *4*, 230.

MacDonald, M. C. (2013). How language production shapes language form and comprehension. *Frontiers in Psychology*, *4*, 226.

Markman, E. M., & Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology*, *20*(2), 121–157.

Masur, E. F. (1997). Maternal labelling of novel and familiar objects: implications for children's development of lexical constraints. *Journal of Child Language*, *24*, 427–439.

Rossion, B., & Pourtois, G. (2004). Revisiting Snodgrass and Vanderwart's object pictorial set: The role of surface detail in basic-level object recognition. *Perception*, *33*, 217–236.

Saffran, J. R. (2003). Statistical language learning: Mechanisms and constraints. *Current Directions in Psychological Science*, *12*(4), 110–114.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928.

Shneidman, L. A., Arroyo, M. E., Levine, S. C., & Goldin-Meadow, S. (2013). What counts as effective input for word learning? *Journal of Child Language*, *40*(3), 672–686.

Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory*, *6*(2), 174.

Snow, C. E. (1972). Mothers' Speech to Children Learning Language. *Child Development*, *43*(2), 549–565.

Tamis-LeMonda, C. S., Kuchirko, Y., & Song, L. (2014). Why Is Infant Language Learning Facilitated by Parental Responsiveness? *Current Directions in Psychological Science*, *23*(2), 121–126.

Thiessen, E. D., Hill, E. A., & Saffran, J. R. (2005). Infant-Directed Speech Facilitates Word Segmentation. *Infancy*, *7*(1), 53–71.

Wingfield, A. (1968). Effects of frequency on identification and naming of objects. *The American Journal of Psychology*, *81*, 226–234.

Yurovsky, D., Doyle, G., & Frank, M. C. (2016). Linguistic input is tuned to children's developmental level. In *Proceedings of the 38th annual meeting of the cognitive science society* (pp. 2094–2098).

Yurovsky, D., Yu, C., & Smith, L. B. (2012). Statistical speech segmentation and word learning in parallel: Scaffolding from child-directed speech. *Frontiers in Psychology*, *3*.

A Conceptual Model of Self-Adaptive Systems based on Attribution Theory

Nianyu Li (li_nianyu@pku.edu.cn)
Key Laboratory of High Confidence, Peking University, China

Zhengyin Chen (chenzy512@pku.edu.cn)
Key Laboratory of High Confidence, Peking University, China

Zi-Long Li (zl.li@imt-atlantique.net)
IMT Atlantique, France

Wenpin Jiao (jwp@pku.edu.cn)
School of Electrical Engineering and Computer Science, Peking University, China

Abstract

The development of self-adaptive systems has attracted lots of attention as they can adapt themselves autonomously to environmental dynamics and maintain user satisfaction. However, there are still tremendous challenges remained. One major challenge is to guarantee the reusability of the system and extend the adaptability with the changing deployment environments. Another challenge is to ensure the adaptability coping with the open and complex environments with the existence of unknown. To solve these problems, we introduce a conceptual self-adaptive model, decoupling the environment with the system. This model is a two-layer structure, based on internal causes and external causes from attribution theory. The first layer, determining how the internal causes affect the adaptation behaviors, is independently designed and reusable; while the second layer, mapping the relationship between external causes with internal causes, is replaceable and dynamically bound to different deployment environments.

Keywords: Self-Adaptation; Attribution Theory; Reusability

Introduction

Current society extensively relies on software systems to achieve specific goals. However, achieving those required goals is a tremendous challenge (Cheng, de Lemos, & et al., 2009) since there are lots of uncertainties that developers have not considered or cannot fully understand during design time, and the changing environment leads to costly reconfiguration and time-consuming maintenance tasks (de Lemos, Giese, & et al., 2010). Therefore, there is a high demand for managing complexity reduction and achieving desired goals within a reasonable cost and timely manner. Self-adaptation is generally considered as one of the most promising approaches to manage the uncertainties of modern software systems since it enables a system to adapt itself autonomously to user requirements or environmental dynamics to continuously achieve system goals including performance, security, fault management, etc (Sawyer, Bencomo, & et al., 2010).

In the existing literatures, most of the adaptation behaviors are triggered by events in the environment (Salehie & Tahvildari, 2009; Shevtsov, Berekmeri, & et al., 2018; Filieri, D'Ippolito, & et al., 2017; Modoni, Trombetta, Veniero, Sacco, & Mourtzis, 2019). That is to say, the main adaptability of a self-adaptive software system is the internal response to the changes in the external environmental factors. Accordingly, the whole lifecycle of the adaptive system, including design time and run-time, is always associated with the environment where the system is deployed. In the design phase, system environment, as well as the mechanisms of perceiving and effecting environment are modeled and implemented. And the set of adaptation policies, tightly binding to this specific environment, are defined and customized. Then at run-time, adaptation behaviors could be achieved by implementing the activities of a well recognized feedback control loop called MAPE-K (Monitoring, Analysis, Planning, and Execution with Knowledge).

One of the disadvantages of current method is that these adaptation policies bound tightly to a specific environment will inevitably limit the adaptivity of the system to various deployment environments. Take a robot system as the example. In a wood floor environment, there could be policies describing how fast the robot should move forward to reach its destination as soon and as safe as possible; or how many angles it shall turn when encountering obstacles. However, the value of speed and angles will be very different in a more slippery tile floor or on a rough cement road. Therefore, for an adaptive system, in addition to being able to adapt in the specific deployment environment, it should have a wide range of applications (i.e., being deployed in a variety of environments). The other disadvantage is that current adaptive systems cannot cope with the increasing complexity and openness of the environment. It is basically impossible to pre customize a complete environment model since the developer cannot fully understand or have considered at design time. Inevitably, new environmental factors might exist and appear at run-time and the system is not reliable to recognize or predict those unforeseen. For example, when designing the adaptive strategies for a robot avoiding obstacles, it is necessary to know in advance what kind of obstacles it might encounter, and then to specify how to deal with obstacle A, obstacle B, etc. Obviously, obstacles in the environment could be infinite. New and unexpected obstacles will emerge constantly in the practical environment, which leads to the inadequate capacity of the existing system.

The fundamental reason for these disadvantages lies in the tight bound between the specific environment and the system. To deal with current challenges, this paper comes up with a novel approach based on attribution theory. Philosophically, the internal causes are the fundamental reasons for the change or the development of things while the external causes are merely the conditions and become operative through internal causes. In other words, it is when the external causes lead to the changes of internal causes that the adaptation behaviors could be triggered. Therefore, the basic idea behind our approach is to decouple the environment with the system both at two stages: independent design and run-time binding. In the design stage, how the adaptation behaviors of the system are determined by the internal causes is focused and emphasized, which makes the design and development of software are independent of the practical environment. In the run-time stage, the relationships between environmental factors (i.e., environmental events as external causes) and state of the system (i.e., internal causes) are dynamically established, thus binding the system to the specific application (i.e., deployment) and realizing the environmental-related adaptability.

The main contributions of our research is summarized as follows:

- We propose a new conceptual model of designing adaptive systems based on the attribution theory;
- We describe a two-layer structure in accordance with the conceptual model. The first layer is the independent design with decisive adaptation policies pertaining the relation between internal causes to adaptation behaviors; while the second layer is the dynamic bound with influential adaptation policies connecting the external causes to internal causes.

Approach Overview

So fundamental is the process of asking and answering "why" questions – trying to figure out what caused something else – that it has been characterized as a basic human activity, and a family of theories has been developed to illumine how and why things happen as they do. This set of theories, collectively called Attribution Theory initiated by Fritz Heider(Heider, 1958) and further advanced by Harold Kelley and Bernard Weiner(Kelley, 1967), attempts to describe and explain the processes involved in everyday explanations, most typical explanations of individual behaviors and events. An interesting example that someone is angry could be attached to the causes of bad-tempered characteristics or something bad happened.

There are a number of definitions for attributions, but a common way to define attributions is as the internal and external process of interpreting and understanding what is behind individual behaviors. External attribution, also called situational attribution, refers to interpreting the causes of behaviors to the situational or environment features outside a person's control. Internal attribution is the process of assigning the causes to some personality traits, rather than to outside forces.

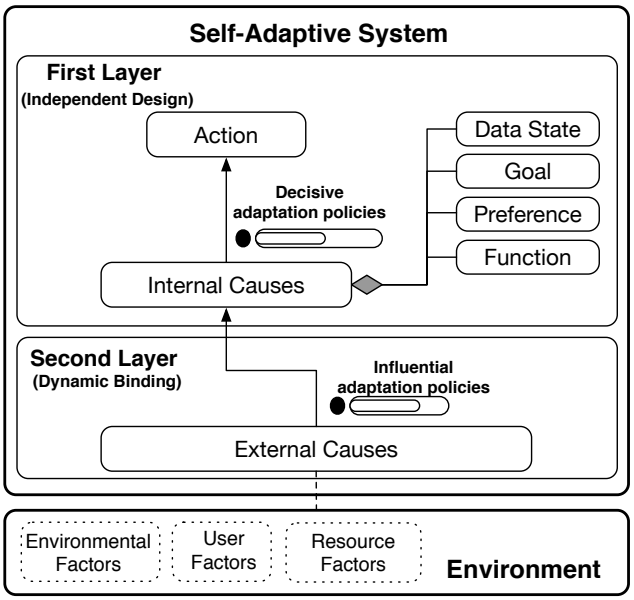


Figure 1: Conceptual Model of Attribution-Based Self-Adaptive System.

Similarly, adaptive behaviors of self-adaptive systems are the reactions to the external causes, i.e., changes. It is important to identify the reason for an adaptation: *why do we have to adapt?*. This is the central question influencing the reaction. In general, the reasons for an adaptation could be i)a change in the technical resources, e.g., the availability of an alternative network connection; ii)a change in the environmental variable, e.g., the workload for a website changed; and iii)a change regarding the user, e.g., a change in the user goal or the user preferences(Krupitzer, Roth, & et al., 2015). Users and operative technical resources could be regarded as a part of the environment and together with the environmental factors form the periphery of adaptive systems (Jiao & Sun, 2016). They are the external conditions of the existences and referred to as external causes.

Factors or events in the environment are not the necessary conditions that systems could execute reactive behaviors. For example, when the number of active users increases, some of the websites may saturate while others may not. In fact, whether an application deployed in the cloud is saturated and then allocated with more resources depends on whether the response latency (the time elapsed from sending the first byte of the request to receiving the last byte of the response) is long, and one will not do so if the latency is within a satisfactory range even if this application is with a huge number of users. In other words, the change of the user number does not determine the adaptive behaviors on an allocation of the resource; instead, the influence could take effect only when the change affects system internal states, which further im-

pair system goals. On the contrary, internal changes on the states of the system itself are the intrinsic reasons for adaptive behaviors.

Figure 1 provides a birds-eye view of our conceptual model of self-adaptive systems. An adaptive system can be divided into two layers, each corresponding to the internal and external attribution process. The first layer is composed of internal causes (including data state, function, preference, and goal) and adaptation behaviors (i.e., actions). Decisive adaptation policies, how to cope with changes in internal causes, determine the relations between internal causes and actions. This layer is independently designed and fixed even with the changing deployment or the extended environment. Note that the changing deployment is the change from one specific environment to another; the extended environment is the open environment with environmental factors from unknown to known. In the second layer, the relationship between external causes and internal causes is denoted in the influential adaptation policies. The second layer is a replaceable component and dynamical bound when the running environment is determined.

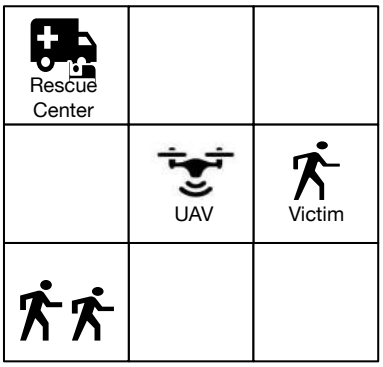


Figure 2: Running Example

Running Example. As a motivating example showcasing our approach, consider a *disaster relief* mission in adaptive system domains (Li, Jiao, & et al., 2018). In such a scenario, communication infrastructure is disabled in a city due to disasters; parts of the city may be unsafe. Figure 2 visualizes a possible configuration of a part of the city (i.e., a district). The rescue center is the safe zone in charge of the district's safety. The district then is divided into several blocks. The victims are spread in different blocks and have no idea of where the rescue center is. The unmanned aerial vehicle(abbr. as UAV) will be arranged to search and guide victims to the rescue center. In the process of searching for the victims, the UAV should not only guarantee the search and rescue task quick and thorough (search and rescue all victims in an acceptable time), but also ensure its own safety (no crash) and energy storage (avoid battery depletion). Furthermore, we expect that this UAV system can participate in missions in various disaster environments, such as fire, floods, earthquakes etc.

Formal Definition of the Conceptual Model

Inspired by the attribution theory, external causes take effects on the system through the internal causes, instead of directly affecting or determining the behaviors of the system. To this end, the conceptual model (M) of a self-adaptive system is defined as a tuple $\mathtt{M} = (\mathtt{IC}, \mathtt{DAP}, \mathtt{IAP})$, where:

- $\mathtt{IC}$ are the internal causes (i.e., intrinsic reasons for adaptation behaviors) which can be further specified as a tuple $\mathtt{IC} = (\mathtt{Data\ State}, \mathtt{Goal}, \mathtt{Preference}, \mathtt{Function})$;
- $\mathtt{DAP}$ are the decisive adaptation policies, which define how the internal causes of the system determine the adaptation behaviors and are generally expressed as rules of "internal causes – actions";
- $\mathtt{IAP}$ are influential adaptation policies, denoting how events in the environment affect the changes of internal causes with the form of "external causes – internal causes".

Internal Causes and System State

Data State The remembered information of the system determined by a set of attribute values. Let $(Attr = a1, ..., an)$ be the attribute set of the system, and $(Dom = dom(a1), ..., dom(an))$ be the set of domains of these attributes. Then the data state of the system is the mapping of these attributes to their values. In the motivation example, UAV system needs to maintain certain data, such as current location, residual power, flight height, searched blocks, unsearched blocks, hazard blocks, status(i.e., cruise or guidance). The data state of UAV is defined by the value of these kinds of information.

Goal the data state that the system expects to achieve or maintain. Generally, goals can be classified into three categories (Filieri et al., 2017). One type of goal is a reference value, called setpoint, to track. In this case, the objective is to keep a measurable quantity as close as possible to the setpoint. The second category of the goals is the variation of the classic setpoint-based goal where the goal resides in a specific range of interest with confidence intervals. The third category of goals concerns the minimization (or maximization) of a measurable quantity of the system. In substance, these goals can be regarded as functions of data states:

$$\mathtt{Goal} = \{\ \mathtt{g} \mid \mathtt{g} \in 2^{\mathtt{Data}} \wedge \mathtt{eval}(\mathtt{g}) = 1\}. \tag{1}$$

In general, the system is considered completely achieving a goal if it enters the target data state; however, in some cases, the system can only (infinitely) get close to the target but not reach. For example, it is impossible to require the UAV's flighting speed "to maintain exactly at 2 meters per second", then the system is said to be achieving the goal to some extent. Therefore, a goal is usually associated with an evaluation function which determines whether the goal has been achieved or how far it has been achieved. For

instance, in this mission scenario, should all the blocks had been searched and rescued, the evaluation result of this goal is one. If it is not, the evaluation function is $\texttt{eval}(\texttt{g}) = \texttt{num}(\texttt{searched blocks}) \div \texttt{num}(\texttt{total blocks})$, and pertains that goal satisfaction is directly proportional to the number of blocks cruised.

Preference The data state that the system is more interested in. Contrary to the goals that the system must be achieved or maintained, the preferences are not necessary must-to-do, but the performance of the system would be better if they are met. For example, the UAV is not only expected to complete the search and rescue goal but also with economical (less electricity consumption) and fast preference (all the victims should be searched as soon as possible). Similar to the goal, the preferences are associated with the utility functions that measure the satisfaction of preferences. For example, a utility function is present: $\texttt{u} = \texttt{residual power} \div \texttt{total energy storage}$, illustrating the tendencies of less energy consuming.

$$\texttt{Preference} = \{\ \texttt{p} \mid \texttt{p} \in 2^{\texttt{Data}} \wedge \texttt{util}(\texttt{p}) \in [0,1]\}. \quad (2)$$

Function The means or methods of achieving the goals. The functions are essentially changing or maintaining system status through manipulating controllable variables. For example, the UAV has the functions of take-off, flying, landing, direction change, etc.

$$\texttt{Function} = \{\texttt{f} \mid \texttt{f} : \texttt{Data} \rightarrow \texttt{Data}\}. \quad (3)$$

System State The state of the system is determined by the internal factors. In other words, data state, the available functions, and the satisfaction of goals and preferences together define the system state. Note that a function is not always valid (sometimes not working); the valid data of the system is the combination of attribute values and functions(i.e., $\texttt{Data} = \{\texttt{s} | \texttt{s} \in (\texttt{Attr} \times \texttt{Dom}) \cup 2^{\texttt{Function}}\}$)

Adaptation Policies

In the complex and uncertain environment, many kinds of environmental factors (discovered and to be discovered) exist, thus resulting in complicated influences on the internal causes. It is not necessarily true that all the changes of environmental factors will affect the system thereupon triggering adaptation behaviors. Only those leading to the changes of internal causes have an impact on the self-adaptive system.

Influential Adaptation Policies The IAP describe how the external causes especially environmental factors affect the internal causes of the system. These external causes directly influence the data state, which will further affect the functions, preferences and goals. In a fire scenario, the environmental factors for the motivation example could be the magnitude of the fire, which inevitably results in different data states for the UAV system. For example, the detection of a serious situation (i.e., high magnitude) for a block would render the UAV marking it as hazard and UAV will try to avoid this block cruise in a certain amount of time for its own safety. However, if in an earthquake scenario, the obstacles from the ground would probably not affect the mark of the block which is stored as an internal data since it is not a threat to high altitude flying UAV.

$$\texttt{IAP} = \{\texttt{p}_\texttt{i} \mid \texttt{p}_\texttt{i} : \texttt{ExtFactors} \rightarrow \texttt{Data State} \times \texttt{Function} \times \texttt{Preference} \times \texttt{Goal}\}. \quad (4)$$

Decisive Adaptation Policies The DAP characterize how the internal causes determine the self-adaptive actions of the system. An action is the operation of a function, which means that taking an action is to perform a function. $\texttt{Action} = \{\texttt{a} | \texttt{f} \in \texttt{Function}, \texttt{a} = \texttt{Do}(\texttt{f})\ \}$. The factors that determine the adaptive actions may involve system states, functions that the system possesses, preferences and goals. For the current location of UAV as shown in Figure 2, without detected victims to be guided to the rescue center, actions of four direction changes (North, South, East, West) are available if all corresponding blocks have not been cruised before and no hazard mark for the time-being.

$$\texttt{DAP} = \{\texttt{p}_\texttt{d} \mid \texttt{p}_\texttt{d} : \texttt{Data State} \times \texttt{Function} \times \texttt{Preference} \times \texttt{Goal} \rightarrow \texttt{Action}\}. \quad (5)$$

Through this conceptual model with a tuple structure, adaptation behaviors are achieved by explicitly defining internal causes and reasoning about the influences of external events on internal causes via IAP, upon which reactive actions are acquired by DAP. Concretely, for the motivation example, the UAV can infer the influences of environmental factors on its internal causes; and then the UAV can reason about its DAP to decide its adaptation behaviors. With the two layer structure, this conceptual model is supposed to be characteristic with applicability and reusability. Applicability entails the appropriateness of our attribution theory based conceptual model to design the self-adaptive systems and to capture dynamics of the environment, triggering IAP and DAP while maintaining satisfaction on system goals. Reusability describes usability of the DAP without modification, especially coordinating with various alterable dynamic binding IAP to different deployment environments and extended environments allowing continuously gaining knowledge.

Implementation Model

This section provides an implementation model of our attribution based approach. Adaptation builds on adaptation policies characterizing casual relationships between external and internal causes of a system, and between internal causes and actions in the knowledge. Adaptation behaviors are achieved by

implementing the activities of the MAPE (Monitoring, Analysis, Planning, Execution) loop (Kephart & Chess, 2003). Analysis and Planning are responsible for identifying possible requirements violations and generating an adaptation strategy, respectively, while Monitoring and Execution are responsible for enacting it at runtime.

Knowledge

To achieve self-adaptation, the system needs to be tailored, mainly regarding to the adaptation policies in the knowledge. This component Knowledge (K) is shared by all MAPE components. Ideally, all of the knowledge should be reusable across the same class of systems, i.e., these systems can adapt to all kinds of deployments and achieve user goals. However, the generality of this component comes at a cost:

- A significant amount of system-specific knowledge needs to be specified and maintained to apply the system to different deployment environments.

- Should the need for changing the K arises with the gain of information from the environment, the whole component shall be revised separately and deployed to aid (correctly) user expectation.

To support the extendability of new information and reusability of adaptation policies across different systems, we separate system-specific knowledge from the environment-specific part, echoing the two-layer structure in the conceptual model. System-specific knowledge, denoted as the DAP and instructing the behaviors to certain states of the system, is fixed and reusable between systems in similar functions. Meanwhile, environment-specific knowledge for defining how events in the deployment environment (external causes) affect system state in the form of IAP, is alterable as the deployment changes or discovery of additional environmental factors impacting system state. This is faithful to the principle of separation of concerns – the principle for separating a design into distinct sections, such that each section addresses a separate concern (Dijkstra & W, 1982).

MAPE Loop

For a specific deployment environment, adaptation behaviors are achieved by following the widely adopted mechanism of MAPE loop, which is shown in the implementation model in Figure 3.

Monitor Events generated in the environment indicating the execution of system actions or natural changes in the external factors are received. Component Monitor (M) gathers or synthesizes particular data through probes (or sensors) from the environment, and saves data in the knowledge in the form of external causes. For our example, events can indicate a serious fire detected by the cruising UAV.

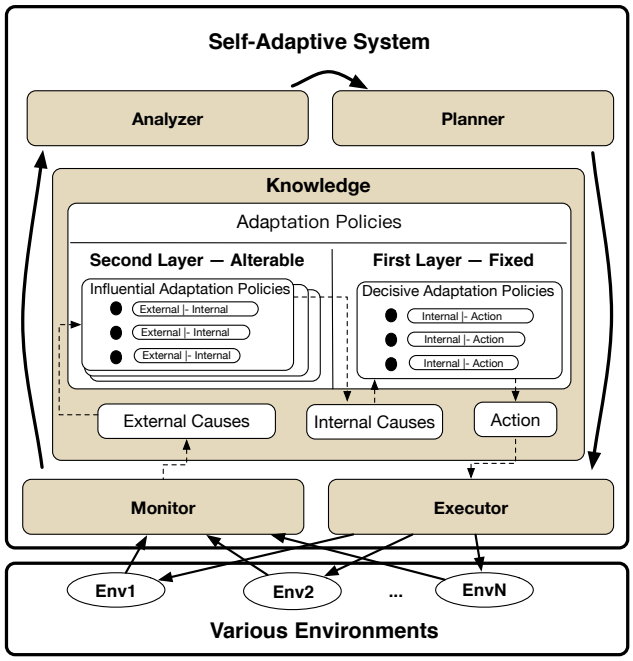


Figure 3: Implementation Model of Attribution-Based Self-Adaptive System.

Analyzer During speculative analysis, conditions of the environment representing violations of goals or better satisfaction of goals which can arise when active entities perform actions are identified. The component Analyzer (A), with the input of external causes from the Monitor, performs analysis by starting adaptation policies engine and reasoning about `IAP` to acquire the data state of the system. On this basis, analyzer further checks whether the goal is fulfilled; preference is satisfied; an adaptation is required. A typical example could be a new mark of hazard blocks resulted from the fire situation which might endanger its own safety.

Planner Component Planner (P) composes a workflow of adaptation actions aiming to counteract violations of system goals or better achieving goals. It consists of one or a set of actions to be enacted inferring from `DAP` in the adaptation policies engine receiving internal causes as input. For each situation, it identifies a policy if one exists, or prompts for a change in the design of the system if the violation cannot be handled and the system goal cannot be satisfied. Direction changing or safe landing could be feasible actions for UAV facing with a dangerous situation.

Executor During execution, the action from the `DAP` is enacted on the system by the component executor (E) through effectors (or actuators). This activity receives as input the current conditions of the environment from the monitoring activity, and identifies if a specific state in the adaptation policy is reached. If that is the case, it enacts specific action indicated in the adaptation strategy.

Discussion

Self-adaptation has been growing increasingly important. Though numerous excellent research efforts have been put into this area, self-adaptation as a field is still in its infancy, and existing knowledge and approaches are not adequate enough to address today's ever-expanding and ever-changing various environments. In this paper, we mainly focus on a novel conceptual model to design self-adaptive systems for various and open environments based on attribution theory, offering reusability engineering by decoupling the environment with the system. Accordingly, the related work will be classified into two categories. First, we look into the mechanisms of reusability in adaptive systems, positioning our work. Then, we discuss cross approaches among different disciplines that play an important role in the construction of self-adaptive systems.

Reusability has always been a concern in adaptive systems field. Generally, research has focused on providing frameworks for adaptation, such as rainbow (Garlan, Cheng, & et al., 2004) monitoring the executing system in the system-layer through an abstract model in the architecture layer which interacts with system layer through a translation layer, and Hogna, a platform for deploying self-managing web applications on cloud (Barna, Ghanbari, & et al., 2015), allowing developers to customize each phase of the feedback loop without having to implement the entire layer themselves. Autonomic Software Product Lines (ASPL) is a strategy for developing self-adaptive software systems with systematic reuse by integrating a domain-independent managing system into a domain-specific software product line (Abbas & Nadeem, 2018). Besides, different patterns that can be reused have been proposed facilitating the development of dynamic adaptive systems (Ramirez & Cheng, 2010); other techniques such as bidirectional transformations, a mechanism of synchronization, have been applied to ensure the correctness of reusability in adaptive systems (Colson, Dupuis, & et al., 2016). Though our approach divides the system framework into two levels like most of the reusable approaches, the basis for this division is the attribution to either environment or system itself facilitating the reusability in various deployment environments, not the structure to be reused in different systems with similar functions.

Adaptive system is an interdisciplinary research field. The concept of self-adaptation, derived from biology, is the characteristics of a creature changing its habits to adapt to a new environment (Longman, 1994). Biological approaches in computer science have emerged with the study of collective behavior in natural multi-agent systems by Parunak (Parunak, 1997). Other mechanisms in biology, such as flocking, nest building, molding (Mamei, Menezes, & et al., 2006) and human immune system (Hart, McEwan, & et al., 2011) has been adopted in self-organizing systems and can be transferred to self-adaptive systems. Besides that, it is important to learn and borrow from other fields of knowledge that are working or have been working in the development and study of similar systems, or have already contributed solutions that fit for the purpose of self-adaptive systems. Researches from chemical have been gradually applied. Viroli et al. propose a coordination model for self-organizing systems based on biochemical tuple spaces and chemical reactions (Viroli, Mirko, & et al., 2009). In the physical field, Weyns et al. employ field-based mechanisms for adaptive task assignment in multi-agent systems. Social area concentrates on market and auction mechanisms and as an example, coordination in multi-agent systems is based on social conventions (Salazar, Rodríguez-Aguilar, & et al., 2010). To the end, our approach is inspired by the research findings from psychology, emphasizing that the influence on adaptation behaviors comes from two aspects, the external environment and the internal system. It decouples the system with a specific environment and brings a new perspective in the construction of self-adaptive systems.

In our future research, we plan to further elaborate on the work presented in this paper by applying the method to practical scenarios to strengthen the applicability. In addition, the mapping relations between external factors and internal causes are complicated and changeable due to open and various environments. More efforts would be put into investigating the automatic acquisition of influential adaptation policies, such as machine learning in response to uncertain environmental changes and reinforcement learning method constantly adjusting to new environments.

Acknowledgments

We would like to thank the anonymous reviewers for their valuable comments. This work is partially sponsored by the National Basic Research Program of China (973) (2015CB352200), and the National Natural Science Foundation of China (61620106007).

References

Abbas, & Nadeem. (2018). *Designing self-adaptive software systems with reuse*. Unpublished doctoral dissertation, Linnaeus University Press.

Barna, C., Ghanbari, H., & et al. (2015). Hogna: A platform for self-adaptive applications in cloud environments. In *10th IEEE/ACM international symposium on software engineering for adaptive and self-managing systems, SEAMS 2015, florence, italy, may 18-19, 2015.*

Cheng, B. H. C., de Lemos, R., & et al. (2009). Software engineering for self-adaptive systems: A research roadmap. In *Software engineering for self-adaptive systems [outcome of a dagstuhl seminar]* (pp. 1–26).

Colson, K., Dupuis, R., & et al. (2016). Reusable self-adaptation through bidirectional programming. In *Proceedings of the 11th international symposium on software engineering for adaptive and self-managing systems, seams@icse 2016, austin, texas, usa, may 14-22, 2016.*

de Lemos, R., Giese, H., & et al. (2010). Software engineering for self-adaptive systems: A second research roadmap.

In *Software engineering for self-adaptive systems II - international seminar, dagstuhl castle, germany, october 24-29, 2010 revised selected and invited papers* (pp. 1–32).

Dijkstra, & W, E. (1982). On the role of scientific thought. In *Selected writings on computing: a personal perspective* (pp. 60–66). Springer.

Filieri, A., D'Ippolito, N., & et al. (2017). Control strategies for self-adaptive software systems. *TAAS*.

Garlan, D., Cheng, S., & et al. (2004). Rainbow: Architecture-based self-adaptation with reusable infrastructure. *IEEE Computer*.

Hart, E., McEwan, C., & et al. (2011). Advances in artificial immune systems. *Evolutionary Intelligence*.

Heider, F. (1958). *The psychology of interpersonal relations*. New York: Wiley.

Jiao, W., & Sun, Y. (2016). Self-adaptation of multi-agent systems in dynamic environments based on experience exchanges. *Journal of Systems and Software*.

Kelley, H. H. (1967). Attribution theory in social psychology. *Nebraska Symposium on Motivation*, *15*, 192-238.

Kephart, J. O., & Chess, D. M. (2003). The vision of autonomic computing. *IEEE Computer*.

Krupitzer, C., Roth, F. M., & et al. (2015). A survey on engineering approaches for self-adaptive systems. *Pervasive and Mobile Computing*, *17*, 184–206.

Li, N., Jiao, W., & et al. (2018). *2018 IEEE international conference on software quality, reliability and security, QRS 2018, lisbon, portugal, july 16-20, 2018.* IEEE.

Longman, A.-W. (1994). Adaptive control. In *Publishing co., inc. boston, ma, usa.*

Mamei, M., Menezes, R., & et al. (2006). Case studies for self-organization in computer science. *Journal of Systems Architecture*.

Modoni, G. E., Trombetta, A., Veniero, M., Sacco, M., & Mourtzis, D. (2019). An event-driven integrative framework enabling information notification among manufacturing resources. *Int. J. Computer Integrated Manufacturing*, *32*(3), 241–252. Retrieved from https://doi.org/10.1080/0951192X.2019.1571232 doi: 10.1080/0951192X.2019.1571232

Parunak, H. V. D. (1997). "go to the ant": Engineering principles from natural multi-agent systems. *Annals OR*.

Ramirez, A. J., & Cheng, B. H. C. (2010). Design patterns for developing dynamically adaptive systems. In *2010 ICSE workshop on software engineering for adaptive and self-managing systems, SEAMS 2010, cape town, south africa, may 3-4, 2010.*

Salazar, N., Rodríguez-Aguilar, J. A., & et al. (2010). Robust coordination in large convention spaces. *AI Commun.*.

Salehie, M., & Tahvildari, L. (2009). Self-adaptive software: Landscape and research challenges. *TAAS*, *4*(2), 14:1–14:42.

Sawyer, P., Bencomo, N., & et al. (2010). Requirements-aware systems: A research agenda for RE for self-adaptive systems. In *RE 2010, 18th IEEE international requirements engineering conference, sydney, new south wales, australia, september 27 - october 1, 2010* (pp. 95–103).

Shevtsov, S., Berekmeri, M., & et al. (2018). Control-theoretical software adaptation: A systematic literature review. *IEEE Trans. Software Eng.*, *44*(8), 784–810.

Viroli, Mirko, & et al. (2009). Biochemical tuple spaces for self-organising coordination. In *Coordination models and languages.* Springer Berlin Heidelberg.

Inquiry, Theory-Formation, and the Phenomenology of Explanation

Emily G. Liquin (eliquin@princeton.edu)
Tania Lombrozo (lombrozo@princeton.edu)
Department of Psychology, Princeton University
Princeton, NJ 08540 USA

Abstract

Explanations not only increase understanding; they are often deeply *satisfying*. In the present research, we explore how this phenomenological sense of "explanatory satisfaction" relates to the functional role of explanation within the process of inquiry. In two studies, we address the following questions: 1) Does explanatory satisfaction track the epistemic, learning-directed features of explanation? and 2) How does explanatory satisfaction relate to both antecedent and subsequent curiosity? In answering these questions, we uncover novel determinants of explanatory satisfaction and contribute to the broader literature on explanation and inquiry.

Keywords: explanation; curiosity; theories; inquiry; learning

Humans have an insatiable drive to explain the world around them, and this drive plays an important role in supporting our amazing capacity to learn (Lombrozo, 2012, 2016). In fact, some have suggested that explanation is to theory-building as orgasm is to reproduction (Gopnik, 2000): the phenomenological sense of satisfaction that accompanies an explanation motivates theory-building, just as orgasm motivates reproduction. In the present research, we investigate this hallmark phenomenological component of explanation ("explanatory satisfaction"). What makes an explanation satisfying, and how does this phenomenological sense function to support learning and theory-formation?

Following Gopnik (2000), we assume explanations are comprised of two elements: an epistemic element and a phenomenological element. The epistemic element of an explanation is straightforwardly related to the process of theory-formation and inquiry: broadly, an explanation includes theory-relevant information, which enables learning and facilitates future prediction and intervention. The phenomenological element, on the other hand, is best characterized as an affective response (Gopnik, 2000), and its role in the process of theory-formation and inquiry is less clear. In the present research, we address two questions that situate explanatory satisfaction within this broader process.

First, how does the phenomenological component of explanation relate to the epistemic component of explanation? If explanatory satisfaction plays a functional role in the process of theory-building and inquiry, we might expect explanations to be found more satisfying when they possess features that suggest the epistemic function of explanation has been achieved. We refer to such features as "learning-directed," as they relate to the epistemic role explanation plays in learning. For instance, we might expect explanations to be deemed more satisfying when they identify novel, useful, and generalizable patterns in the environment, or when they possess explanatory virtues (such as simplicity and breadth) that support correspondingly simple and broad theories. Our first research question is whether explanatory satisfaction is indeed influenced by these learning-directed features.

Second, how does explanatory satisfaction relate to *curiosity*, another affective state that often drives explanation-seeking and exploration? Does curiosity about the answer to a given question increase the explanatory satisfaction experienced upon receiving the answer? Do satisfying explanations terminate inquiry by satisfying curiosity, or do they stimulate further inquiry by prompting curiosity about related matters?

In addressing these questions, our studies are among the first to consider explanatory satisfaction within a broader process of inquiry and theory-building, tying the phenomenological component of explanation to its epistemic role ("learning-directed" considerations), and linking it to other affective states that influence learning (namely curiosity). We briefly review prior work on explanatory satisfaction and curiosity before presenting two novel studies.

Prior Work on Explanatory Satisfaction

Research on explanatory preferences and judgments of explanation quality has shown that people prefer explanations that are simple in the sense that they appeal to few unexplained causes (Bonawitz & Lombrozo, 2012; Lombrozo, 2007; Pacer & Lombrozo, 2017; see also Thagard, 1989), and broad in two senses: in that they explain all the relevant features of what's currently being explained (Johnson, Johnston, Toig, & Keil, 2014; Pennington & Hastie, 1992; Thagard, 1989), and in that they explain additional phenomena as well (Preston & Epley, 2005). Other research has found that people prefer explanations with reductive mechanism information (Hopkins, Weisberg, & Taylor, 2016), that appeal to the function of the thing being explained (Kelemen & Rosset, 2009), that have a narrow "latent scope" (Khemlani, Sussman, & Oppenheimer, 2011), and that cite information "inherent" to what is being explained (Cimpian & Salomon, 2014).

There is also evidence that explanations are favored when they are believed to be generalizable and well-suited to future goals. For example, people find functional explanations more acceptable when they appeal to a generalizable causal process (Lombrozo & Carey, 2006). They also judge such explanations better (relative to

category-based or mechanistic explanations) when they anticipate making future inferences on the basis of information about an entity's function as opposed to information about its category membership or the mechanism by which it operates (Vasilyeva, Wilkenfeld, & Lombrozo, 2017). Additionally, it has been proposed that the "explanatory virtues" that have been tied to explanatory satisfaction—simplicity and breadth—are important exactly because they point to the value of an explanation in guiding future inference and action (Lombrozo, 2016; Pacer & Lombrozo, 2017; see also Vasilyeva, Blanchard, & Lombrozo, 2018). Consistent with this idea, research finds that prompts to explain make children and adults more likely to discover simple and broad patterns, improving learning under some conditions (for a review, see Lombrozo, 2016).

Taken together, this work suggests that explanatory satisfaction may be driven in part by features of explanations relevant to learning and theory-formation. Very little work, however, has investigated the relationship between judgments of explanatory satisfaction and learning-directed considerations more directly. In one study, Zemla, Sloman, Bechlivanidis, and Lagnado (2017) presented participants with explanations drawn from an on-line forum, and had them rate the explanations on several dimensions, including what they called novelty ("I learned something new from this explanation"), generality ("This explanation appeals to a general principle [that is, a general rule that applies to many things]"), perceived expertise ("This explanation was written by an expert in this topic"), and quality ("This is a good explanation"). Novelty and generality were moderately correlated with quality, though these correlations were not significant after correcting for multiple comparisons. There was also evidence of a preference for *complexity* over simplicity: participants favored explanations involving multiple causal mechanisms. These findings hint at possible relationships between learning-directed considerations and judgments of explanation quality, but many questions remain open. In particular, which learning-directed features might predict explanatory satisfaction, and when and why is simplicity versus complexity favored? In Studies 1-2, we consider how judgments of learning, utility, simplicity, complexity, expertise, and breadth relate to explanatory satisfaction.

Prior Work on Curiosity and Epistemic Emotions

Recent work on *explanation-seeking curiosity* has investigated what triggers curiosity about why something is the case, motivating a learner to seek an explanation (Liquin & Lombrozo, 2018). In this work, participants received explanation-seeking questions posed in an on-line forum, and rated the questions along a variety of dimensions, including curiosity ("How curious are you about the answer to this question?"). Anticipated learning, generality, and future utility were among the strongest predictors of curiosity. Complexity and expertise were also found to be positive predictors of curiosity. However, it is not known whether curiosity about an explanation affects the perceived quality of or rated satisfaction with that explanation once obtained. In Studies 1-2, we consider whether antecedent curiosity predicts explanatory satisfaction. In Study 2, we additionally consider how explanatory satisfaction affects curiosity for further inquiry.

One reason it is valuable to relate explanation to curiosity is because doing so helps bridge the epistemic role of explanation with the affective and motivational factors that guide (epistemic) behavior. Curiosity is often characterized as an *epistemic feeling* or *emotion* (alternatively referred to as a noetic feeling; Arango-Muñoz, 2014; de Sousa, 2009; Dokic, 2012; Morton, 2010): one of a class of evaluative appraisals of one's own knowledge, which have a distinctive phenomenology and guide epistemic action (de Sousa, 2009). While a full treatment of epistemic emotions is beyond the scope of this paper, linking explanatory satisfaction to curiosity and learning is a step towards a more complete account of how the phenomenological and epistemic roles of explanation function together to support effective learning.

Study 1

In Study 1, we present participants with why-questions and their corresponding answers. In addition to having them indicate the extent to which they find each answer satisfying ("explanatory satisfaction"), we have them rate each answer along a variety of epistemically-relevant dimensions. We also have them rate their curiosity about the answer to each question prior to receiving it. This design allows us to address two related questions.

First, we ask about the role of learning and theory-building considerations in determining explanatory satisfaction. To do so, we have participants indicate the extent to which each explanation teaches them something new, and whether the information it offers is useful and generalizable. We also ask them to evaluate the extent to which each explanation is simple, broad (in the sense of applying beyond what is being explained), and required expertise to produce. We can then evaluate whether and how strongly these factors predict explanatory satisfaction.

Second, we ask how curiosity about an explanation affects explanatory satisfaction. Specifically, are the explanations offered in response to questions that elicit high levels of curiosity judged more satisfying than those offered in response to questions that elicit lower levels of curiosity?

By answering these questions, we shed light on how explanatory satisfaction relates to the epistemic features of explanations and to curiosity, another epistemic emotion that drives inquiry.

Method

Participants Participants in Study 1 were 159 adults (77 male, 78 female, 2 other, and 1 prefer not to specify, ages 19-68) recruited from Amazon Mechanical Turk. Participation was restricted to MTurk workers in the United States, who had completed at least 1000 prior tasks with a

Table 1: Items (each rated on a seven-point scale) for explanatory satisfaction and learning-directed features in Studies 1-2.

Dimension	Full text of item
Satisfaction	How satisfying do you find the answer to this question?
Actual Learning	To what extent has the answer to this question taught you something new?
Learning Potential	Do you think there is something to be learned from the answer to this question (even if you yourself already knew the answer)?
Expertise	Do you think that answering this question required special expertise in some domain?
Simplicity	Do you think the answer to this question is simple or complex?
Breadth	Do you think the answer to this question is narrow (only applies to what is being explained) or broad (also applies to other similar cases)?
Future Utility	To what extent will the answer to this question be useful to you in the future?
Regularity	Do you think the answer to this question helps reveal a genuine pattern, structure, or regularity?

minimum approval rating of 99%. Forty additional participants completed the study but were not included in analyses because they did not pass two attention checks.

Materials Fifty-six questions and answers were selected from the book *1000 Questions & Answers Factfile* (Kerrod, Madgwick, Reed, Collins, & Brooks, 2006). For example, the question "Why do some stars explode?" was answered with the following explanation: "Massive stars explode when they come to the end of their lives. They swell up into huge supergiants. Supergiants are unstable, so they collapse and blast into pieces in an explosion called a supernova. Supernovae are the most intense explosions in the universe, as bright as billions of suns put together."

Procedure Each participant saw four questions randomly selected from the 56 questions described above. Participants first rated their curiosity about each question ("Consider the following: *[question premise]*. How curious are you about why this is the case?"). Participants also rated seven items that are not relevant to the present research and are not reported here. Next, participants completed seven arithmetic problems; those who did not correctly respond to at least five items were excluded.[1] After this task, participants read the answer to each of the four questions, and rated each answer on explanatory satisfaction and several learning-directed features (see Table 1). Finally, participants completed a memory check, which required selecting four of the questions presented during the rating tasks from a list with four distractor questions. Participants were given one point for each correct response (hit or correct rejection), and those who scored fewer than six points were excluded.

Results

Due to the nested structure of the data, all analyses used a mixed-models approach, with random intercepts for participant and item in all models. Standardized regression coefficients are reported; all reported coefficients reached significance at the $p < .05$ level using likelihood ratio tests. In addition to the results reported here, all regression analyses were repeated controlling for the length of the explanation in number of words, as prior work has shown that longer explanations tend to be more satisfying (Weisberg, Taylor, & Hopkins, 2015). Controlling for explanation length had no effect on our results.

Learning-Directed Features First, we tested the role of learning-directed considerations in predicting explanatory satisfaction. To do so we fit a regression model predicting satisfaction with all learning-directed considerations entered simultaneously as fixed effects. Only actual learning, β = 0.26, 95% CI [0.17, 0.34], learning potential, β = 0.18, 95% CI [0.10, 0.26], and future utility, β = 0.12, 95% CI [0.04, 0.21], explained unique variance in satisfaction holding all other measures fixed (see Figure 1). However, as many of the measures were modestly correlated with each other (see Figure 2), potentially affecting the robustness of the coefficient estimates reported above, we also fit a separate regression model for each measure. Actual learning, β = 0.39, 95% CI [0.32, 0.47], learning potential, β = 0.39, 95% CI [0.31, 0.46], expertise, β = 0.31, 95% CI [0.24, 0.39], simplicity, β = -0.23, 95% CI [-0.30, -0.15], breadth, β = 0.13, 95% CI [0.05, 0.20], future utility, β = 0.25, 95% CI [0.17, 0.33], and regularity, β = 0.17, 95% CI [0.10, 0.25], were all significant predictors of explanatory satisfaction (see Figure 1).

Antecedent Curiosity Next, we tested whether curiosity about the anticipated answer to a question predicted explanatory satisfaction. We found that curiosity was a significant (though modest) predictor, β = 0.19, 95% CI [0.11, 0.26], and that the model including curiosity as a fixed effect was a significant improvement upon the null model, $\chi^2(1) = 23.12$, $p < .001$.

[1] This attention check may assess numeracy, which could lead to unnecessary exclusions that are irrelevant to successful completion of our task. However, when the participants who failed this task are included in all analyses (for Studies 1 and 2), all results remain unchanged.

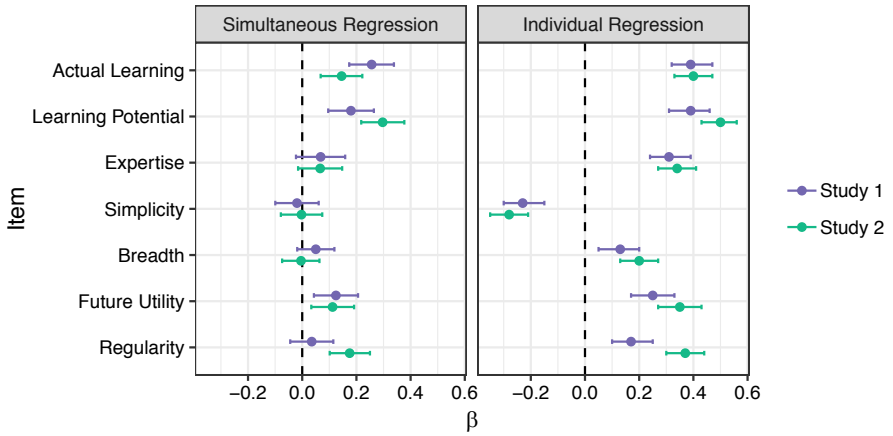


Figure 1: Study 1 and Study 2 standardized regression coefficients for each measure predicting explanatory satisfaction, in a simultaneous regression model (left panel) and in individual regression models (right panel). Study 2 regression coefficients control for interest and knowledge. Error bars = 95% CI.

Discussion

The findings from Study 1 are largely consistent with previous research on the role of breadth (Johnson et al., 2014; Preston & Epley, 2005), future utility (Vasilyeva et al., 2017), and generalizability (Lombrozo & Carey, 2006) in driving explanatory satisfaction. However, one important qualification is that *complexity*, rather than *simplicity*, led to higher ratings of explanatory satisfaction. This is surprising in light of prior work documenting a preference for simpler explanations when using well-controlled stimuli (where, for example, probability is matched; Lombrozo, 2007, Pacer & Lombrozo, 2017), but is consistent with prior work using more naturalistic stimuli, such as those employed here (e.g., Zemla et al., 2017).

Our findings go beyond prior work in identifying an important role for our new learning-directed measures of actual learning, learning potential, and expertise. In fact, these were among the strongest predictors of explanatory satisfaction. We also found a modest role for antecedent curiosity, in that greater curiosity about the answer to a question predicted greater satisfaction with the answer. While this has not (to our knowledge) been tested in prior research, there is evidence that the gap between curiosity about the answer to a trivia question and the satisfaction upon receiving the answer predicts later memory for the answer (Marvin & Shohamy, 2016). This suggests that how much is learned from an explanation could be a function of *both* antecedent curiosity and the explanatory satisfaction experienced from the explanation itself.

These findings highlight the value of approaching the study of explanatory satisfaction through the lens of theory-formation and inquiry. In particular, if achieving explanatory satisfaction effectively motivates learning and theory-formation, then we should expect a close correspondence between explanatory phenomenology and the epistemic functions of explanation. Our findings provide initial support for this correspondence.

Study 2

In Study 2, we replicate the key findings from Study 1, while controlling for two potentially relevant factors: participants' a priori interest in and knowledge about the topics the explanations address. We also investigate how explanatory satisfaction relates to the ongoing process of inquiry (for a discussion, see Danovitch & Mills, 2018) by considering how explanatory satisfaction affects subsequent curiosity. We propose two competing hypotheses: First, it is possible that the receipt of a satisfying explanation will halt further inquiry. Supporting this hypothesis, Frazier, Wellman, and Gelman (2009) found that preschoolers in both naturalistic and experimental settings were less likely to re-ask a question following an explanation (vs. a non-explanation) from an adult, suggesting that the receipt of an explanation stopped further inquiry, at least concerning the topic in question. Relatedly, Mills, Sands, Rowles, and Campbell (2019) found that children were more likely to request additional information in response to explanations that they rated as less-complete answers to the relevant question, relative to more-complete explanations.

However, it is also possible that receiving a satisfying explanation could *promote* further inquiry. Even a satisfying explanation will often highlight new things the learner does not yet know, promoting further exploration and information search. For example, Liquin and Lombrozo (2017) found that generating explanations during learning increased information search in the face of surprising evidence (see also Legare, 2012). Moreover, some theories of curiosity posit that curiosity peaks when there is a modest "gap" between a learner's current and desired knowledge state, resulting in an inverted-U-shaped relationship between prior knowledge and curiosity (Loewenstein, 1994). For learners on the ascending side of the "U," a satisfying explanation could result in *greater* curiosity.

To distinguish between these hypotheses, we ask participants to rate their curiosity about several follow-up questions in response to an explanation, after completing the

same ratings as in Study 1. If explanatory satisfaction halts inquiry, we would expect greater satisfaction to predict lower curiosity about follow-up questions. By contrast, if explanatory satisfaction promotes inquiry, we would expect greater satisfaction to predict greater curiosity about follow-up questions.

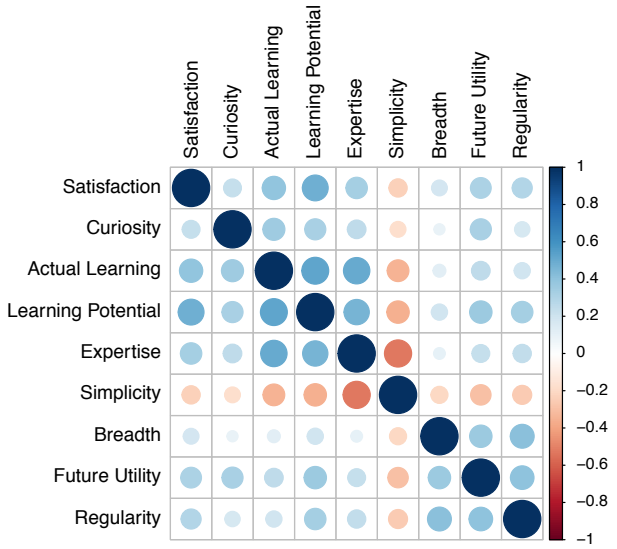


Figure 2: Matrix of pairwise correlation magnitudes for all measures of interest (collapsed across Study 1 and Study 2 data).

Method

Participants One hundred seventy-one adults (96 male and 75 female, ages 21-69) from Amazon Mechanical Turk participated in Study 2. Participation was restricted to MTurk workers in the United States who had completed at least 1000 HITs with a minimum approval rating of 99%. Twenty-nine additional participants completed the study but were excluded from analysis because they did not pass the same attention checks used in Study 1.

Materials Of the 56 questions and answers used in Study 1, twenty were randomly selected for use in this study. An initial sample of 48 MTurk participants read random samples of five question-answer pairs and wrote between 3 and 10 follow-up questions in response to each answer. From this set of follow-up questions, we randomly selected 10 for each question-answer pair. Thus, the materials used in this study were 20 question-answer pairs from *1000 Questions & Answers Factfile* (Kerrod et al., 2006), with 10 follow-up questions in response to each. Additionally, each question was classified into a "topic area," based loosely on the chapter and page topics in the *1000 Questions & Answers Factfile* book. The 20 questions fell into 14 distinct topic areas (e.g., "dinosaurs," "stars," "Ancient Egypt").

Procedure First, each participant rated their interest in and knowledge about each of the 14 topic areas described above. Then, each participant saw four questions randomly selected from the 20 questions. For each question, they completed the initial curiosity rating, followed by the arithmetic distractor/attention task, as in Study 1. Then, two tasks were presented in a randomized order: the answer ratings, as described in Study 1, and the follow-up question task. For the latter task, participants saw a random sample of five of the ten follow-up questions for each of the four questions (presented with answers) that they had seen previously. For each follow-up question, participants rated how curious they were about the answer to that question on a seven-point scale. These five ratings were averaged within each of the four questions, creating a "follow-up curiosity" scale for each question rated by each participant (Cronbach's α = 0.85).

Results

Results were analyzed as in Study 1, using a mixed-models approach. Again, all results remained unchanged when controlling for explanation length.

Replications of Study 1 First, we repeated all analyses from the previous study, but controlling for interest in and knowledge of the topics corresponding to the question-answer pairs. In a simultaneous regression model, actual learning, β = 0.14, 95% CI [0.07, 0.22], learning potential, β = 0.30, 95% CI [0.22, 0.38], future utility, β = 0.12, 95% CI [0.04, 0.20], and regularity, β = 0.18, 95% CI [0.10, 0.25], explained unique variance in satisfaction holding all other measures fixed (see Figure 1). In separate regression models, actual learning, β = 0.40, 95% CI [0.33, 0.47], learning potential, β = 0.50, 95% CI [0.43, 0.56], expertise, β = 0.34, 95% CI [0.27, 0.41], simplicity, β = -0.28, 95% CI [-0.35, -0.21], breadth, β = 0.20, 95% CI [0.13, 0.27], future utility, β = 0.35, 95% CI [0.27, 0.43], and regularity, β = 0.37, 95% CI [0.30, 0.44], were all significant predictors of explanatory satisfaction (see Figure 1). Curiosity was also a significant predictor of explanatory satisfaction, controlling for interest and knowledge, β = 0.19, 95% CI [0.11, 0.26].

Satisfaction and Inquiry Next, we tested the relationship between explanatory satisfaction and subsequent curiosity. To do so, we compared a model predicting average follow-up curiosity with satisfaction as a fixed effect to a null model with no fixed effects. Satisfaction was a significant predictor of follow-up curiosity, $\chi^2(1) = 45.20$, $p < .001$. Critically, the relationship between satisfaction and follow-up curiosity was positive, β = 0.22, 95% CI [0.16, 0.29], indicating that explanatory satisfaction, at least in this context, *encourages* rather than halts ongoing inquiry.

Finally, we repeated this analysis, but controlling for interest in and knowledge of the topics corresponding to the question-answer pairs. Satisfaction remained a significant predictor of follow-up curiosity, β = 0.21, 95% CI [0.14, 0.27], $\chi^2(1) = 40.45$, $p < .001$.

Discussion

In Study 2, we replicated the results of Study 1 while

controlling for topic knowledge and interest, again demonstrating that learning-directed features predict explanatory satisfaction, and that curiosity about an explanation-seeking question predicts satisfaction with the answer.

Study 2 also investigated how explanatory satisfaction relates to subsequent curiosity. We found support for the hypothesis that explanatory satisfaction encourages rather than halts inquiry—that is, the more satisfied a participant was with a given explanation, the more curious they were about several follow-up questions. This is in contrast to past work demonstrating a negative relationship between explanation *completeness* and subsequent information search (Frazier et al., 2009; Mills et al., 2019). This could reflect methodological differences in what was evaluated (explanatory completeness versus satisfaction), or in the opportunities for further inquiry that were offered. For instance, we might expect inquiry concerning the original explanandum to cease after obtaining a satisfying explanation, but for inquiry concerning related matters to be piqued. These questions merit further research.

General Discussion

Explanations play an important role in the process of inquiry: they contribute to learning and theory-building, which in turn support predictions, interventions, and understanding. Explanations also have a unique phenomenology that may motivate this theory-building behavior. However, most research on explanation has not directly addressed how this phenomenology relates to the functional role of explanation within the process of theory-formation and inquiry. In the present research, we addressed two questions: 1) To what extent is explanatory satisfaction driven by features of an explanation that support learning and theory-formation? and 2) How does explanatory satisfaction relate to *curiosity*, another epistemic emotion that motivates inquiry?

In response to the first question, we find that several learning-directed features (such as actual learning, learning potential, future utility, and regularity) are related to explanatory satisfaction, even when controlling for interest in and knowledge of the topics addressed by the explanation. Answering the second question, we find that antecedent curiosity predicts satisfaction with a subsequent explanation to a modest degree, and that explanatory satisfaction in turn predicts curiosity about follow-up questions in response to an explanation, thus encouraging further inquiry.

These studies build upon previous research on explanatory satisfaction and explanation-seeking behavior. In particular, we replicate previous research on the role of breadth, generalizability, and future utility in explanatory satisfaction, and we find several additional predictors of explanatory satisfaction that have not previously been explored—or for the case of learning, that have not previously found strong support (Zemla et al., 2017). Additionally, we add to recent research on curiosity (Liquin & Lombrozo, 2018; Marvin & Shohamy, 2016), demonstrating a systematic relationship to explanatory satisfaction throughout the process of inquiry.

Several limitations of these studies must be noted. First, future work should explore a broader range of materials, including "everyday" questions and explanations from more ecologically-valid settings. Second, the findings we report are all correlational, so it remains to be seen whether (for example) curiosity about an explanation *causes* satisfaction with the later-received explanation. More critically, these studies do not cleanly disentangle the phenomenological component of explanation from the epistemic component. That is, participants' ratings of explanatory satisfaction likely reflected affective responses (perhaps in contrast to ratings of goodness, quality, or completeness, which have often been used in past research; e.g., Mills et al., 2019; Vasilyeva et al., 2018; Zemla et al., 2017), but also evaluation of (epistemic) quality, which may not have been accompanied by any particular phenomenology. For our purposes, the key question is whether and how explanatory satisfaction motivates inquiry, so it is notable that in Study 2, there was a positive relationship between explanatory satisfaction and ongoing curiosity. Future work should explore the relationship between explanatory satisfaction and subsequent epistemic *behaviors*, such as information search, as well as epistemic *consequences*, such as learning.

Another possible limitation of this work is that participants only read a single explanation in response to each question, while previous work on explanatory preferences (e.g., Lombrozo, 2007; Pacer & Lombrozo, 2017) has typically used *comparative* judgments of explanation quality between two competing explanations. As a result, satisfaction judgments in the present research may reflect satisfaction that an explanation *exists*, rather than satisfaction that this explanation fulfills certain explanatory desiderata relative to other possible explanations. Future work should explore whether different criteria are used to evaluate explanations presented simultaneously versus in isolation.

Despite these limitations, these studies are among the first to approach explanatory satisfaction in terms of its functional role within a broader process of inquiry, providing new insights into the determinants of explanatory satisfaction and the importance of this phenomenology in driving ongoing inquiry and theory-building.

Acknowledgements

This work was supported by an NSF Graduate Research Fellowship to EL [DGE 1656466]. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Arango-Muñoz, S. (2014). The nature of epistemic feelings. *Philosophical Psychology*, *27*(2), 193–211.

Bonawitz, E. B., & Lombrozo, T. (2012). Occam's rattle: Children's use of simplicity and probability to constrain inference. *Developmental Psychology*, *48*(4), 1156–1164.

Cimpian, A., & Salomon, E. (2014). The inherence heuristic: An intuitive means of making sense of the world, and a potential precursor to psychological essentialism. *Behavioral and Brain Sciences*, *37*(5), 461–480.

Danovitch, J. H., & Mills, C. M. (2018). Understanding when and how explanation promotes exploration. In M. M. Saylor & P. A. Ganea (Eds.), *Active Learning from Infancy to Childhood: Social Motivation, Cognition, and Linguistic Mechanisms*. Springer International Publishing.

de Sousa, R. (2009). Epistemic feelings. *Mind and Matter*, *7*(2), 139–161.

Dokic, J. (2012). Seeds of self-knowledge: Noetic feelings and metacognition. In M. J. Beran, J. Perner, & J. Proust (Eds.), *Foundations of metacognition*. Oxford University Press.

Frazier, B. N., Gelman, S. A., & Wellman, H. M. (2009). Preschoolers' search for explanatory information within adult-child conversation. *Child Development*, *80*(6), 1592–1611.

Gopnik, A. (2000). Explanation as orgasm and the drive for causal knowledge: The function, evolution, and phenomenology of the theory formation system. In F. C. Keil & R. A. Wilson (Eds.), *Explanation and cognition*. Cambridge, MA: The MIT Press.

Hopkins, E. J., Weisberg, D. S., & Taylor, J. C. V. (2016). The seductive allure is a reductive allure: People prefer scientific explanations that contain logically irrelevant reductive information. *Cognition*, *155*, 67–76.

Johnson, S. G. B., Johnston, A. M., Toig, A. E., & Keil, F. C. (2014). Explanatory scope informs causal strength inferences. *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 707–712). Austin, TX: Cognitive Science Society.

Kelemen, D., & Rosset, E. (2009). The human function compunction: Teleological explanation in adults. *Cognition*, *111*(1), 138–143.

Kerrod, R., Madgwick, W., Reed, S., Collins, F., & Brooks, P. (2006). *1000 Questions & Answers Factfile*. Boston, MA: Kingfisher.

Khemlani, S. S., Sussman, A. B., & Oppenheimer, D. M. (2011). Harry Potter and the sorcerer's scope: Latent scope biases in explanatory reasoning. *Memory & Cognition*, *39*(3), 527–535.

Legare, C. H. (2012). Exploring explanation: Explaining inconsistent evidence informs exploratory, hypothesis-testing behavior in young children. *Child Development*, *83*(1), 173–185.

Liquin, E. G., & Lombrozo, T. (2017). Explain, explore, exploit: Effects of explanation on information search. *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 2598–2603). Austin, TX: Cognitive Science Society.

Liquin, E. G., & Lombrozo, T. (2018). Determinants and consequences of the need for explanation. *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 696–701). Austin, TX: Cognitive Science Society.

Loewenstein, G. (1994). The psychology of curiosity: A review and reinterpretation. *Psychological Bulletin*, *116*(1), 75.

Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive Psychology*, *55*(3), 232–257.

Lombrozo, T. (2012). Explanation and abductive inference. In K. J. Holyoak & R. G. Morrison (Eds.), *Oxford handbook of thinking and reasoning* (pp. 260–276).

Lombrozo, T. (2016). Explanatory preferences shape learning and inference. *Trends in Cognitive Sciences*, *20*(10), 748–759.

Lombrozo, T., & Carey, S. (2006). Functional explanation and the function of explanation. *Cognition*, *99*(2), 167–204.

Marvin, C. B., & Shohamy, D. (2016). Curiosity and reward: Valence predicts choice and information prediction errors enhance learning. *Journal of Experimental Psychology: General*, *145*(3), 266.

Mills, C. M., Sands, K. R., Rowles, S. P., & Campbell, I. L. (2019). "I want to know more!": Children are sensitive to explanation quality when exploring new information. *Cognitive Science*, *43*(1).

Morton, A. (2010). Epistemic emotions. In P. Goldie (Ed.), *The Oxford handbook of philosophy of emotion*. Oxford, UK: Oxford University Press.

Pacer, M., & Lombrozo, T. (2017). Ockham's razor cuts to the root: Simplicity in causal explanation. *Journal of Experimental Psychology: General*, *146*(12), 1761.

Pennington, N., & Hastie, R. (1992). Explaining the evidence: Tests of the Story Model for juror decision making. *Journal of Personality and Social Psychology*, *62*(2), 189.

Preston, J. L., & Epley, N. (2005). Explanations versus applications: The explanatory power of valuable beliefs. *Psychological Science*, *16*(10), 826–832.

Thagard, P. (1989). Explanatory coherence. *Behavioral and Brain Sciences*, *12*(03), 435–467.

Vasilyeva, N., Blanchard, T., & Lombrozo, T. (2018). Stable causal relationships are better causal relationships. *Cognitive Science*, *42*(4), 1265–1296.

Vasilyeva, N., Wilkenfeld, D., & Lombrozo, T. (2017). Contextual utility affects the perceived quality of explanations. *Psychonomic Bulletin & Review*, *24*, 1436–1450.

Weisberg, D., Taylor, J., & Hopkins, E. (2015). Deconstructing the seductive allure of neuroscience explanations. *Judgment and Decision Making*, *10*(5), 429–441.

Zemla, J. C., Sloman, S., Bechlivanidis, C., & Lagnado, D. A. (2017). Evaluating everyday explanations. *Psychonomic Bulletin & Review*, *24*(5), 1488–1500.

Hard choices: Children's understanding of the cost of action selection

Shari Liu (shariliu01@g.harvard.edu)
Fiery Cushman (cushman@fas.harvard.edu)
Sam Gershman (gershman@fas.harvard.edu)
Wouter Kool (wkool@fas.harvard.edu)
Elizabeth Spelke (spelke@wjh.harvard.edu)
Department of Psychology, Harvard University,
Cambridge, MA 02143 USA

Abstract

When predicting or explaining another person's actions, we often appeal to the physical effort they require; a person who works hard for something, for instance, must really like it (Liu, Ullman, Tenenbaum, & Spelke, 2017). But people are not only motivated to avoid physical effort; they also seek to avoid mental effort (Shenhav et al., 2017; Kool & Botvinick, 2018). Here, we ask whether mental effort enters into preschoolers' understanding of other people's actions. Across 4 experiments (N=112), we presented 4- and 5-year-old children with an agent (naive in Exp 1, 2 and 4, and knowledgeable in Exp 3) who can either move through a simple or complex maze environment with a specific goal (in Exp 1-3, to reach a play structure beyond the mazes, and in Exp 4, to practice solving the mazes). We found that children were sensitive to the physical and mental effort associated with more complex mazes, and to the trade-offs between effort and gain in skill. The intuition that choices impose costs on our bodies and minds appears to guide children's understanding of other people.

Keywords: intuitive psychology; cognitive development; decision-making

Introduction

Observing other people try hard tells us something about their desires, beliefs, competence, and what is worth trying for ourselves. All of these abilities rely on the the basic intuition that actions carry cost in the first place. This intuition is an early-emerging component of our human social intelligence: Infants, children, and adults consider the physical effort behind other people's actions as one variable in their plans to maximize utility (Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016; Gergely & Csibra, 2003; Baker, Saxe, & Tenenbaum, 2009), and use how hard people try to infer their goals, beliefs, competence, and the value of effort in general (Jara-Ettinger, Tenenbaum, & Schulz, 2015; Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017; Leonard, Lee, & Schulz, 2017; Liu et al., 2017).

Is our understanding of action cost restricted to the physical exertions of body, or does it also encompass the costs of mental exertion? Everyday activities like thinking, writing, and learning are not physically costly (in fact, our bodies are usually still when engaging in them), but they incur a similar subjective disutility–in other words, a sense of exhaustion. More specifically, cognitive operations like loading information into working memory, transforming and maintaining it over long delays, and task switching–in other words, all the elements of rational planning–carry an intrinsic cost (Kool & Botvinick, 2018; Shenhav et al., 2017; Westbrook & Braver, 2015). Because of this cost, we do not always engage in rational planning; sometimes we use computationally cheaper heuristics, such as selecting actions proportional to their historical rewards. Experiments show that while people often often avoid costly rational planning, they become more likely to bear this cost when it is associated with a sufficiently large prospect of reward (Kool, Gershman, & Cushman, 2017).

What is the role of mental effort in our analysis of other people's actions? Do we assume that mental effort is costly? Do we assume that others would seek to avoid mental effort, all else being equal? Some recent research offers circumstantial evidence in adults (Gershman, Gerstenberg, Baker, & Cushman, 2016): When participants are asked what someone with a strong habit (e.g., to take a certain route to work, or turn a doorknob clockwise) will do in a new situation, they respond that the person is likely to rely on habit, especially under time pressure. This is consistent with the possibility that adults associate cognitive effort with model-based control, and use this association to predict and explain other people's actions.

There is, however, strong reason to believe that the ability to represent and reason about mental effort develops slowly over childhood. Although preschool-aged children understand that other people have emotional states, perceptions, beliefs, and knowledge (Wellman, 2002), they do not reliably know when people are thinking, struggle to make reasonable inferences about what they might be thinking about, and do not reliably report the content of their own thoughts (Flavell, Green, & Flavell, 1995). Furthermore, children are relatively poor at monitoring their own comprehension, memory, and learning, at least in ways that can be measured through explicit questioning (Flavell, Friedrichs, & Hoyt, 1970). If the ability to monitor one's own cognition develops slowly, then children may come to reason about the role of mental effort in others' plans at a later age than they reason about physical effort in these plans.

This paper presents a case study of the developmental origins of reasoning about mental effort. Specifically, we ask whether children understand that making choices can lead to both physically and mentally costly outcomes, and whether they understand the trade-offs people make between effort and reward in the context of learning.

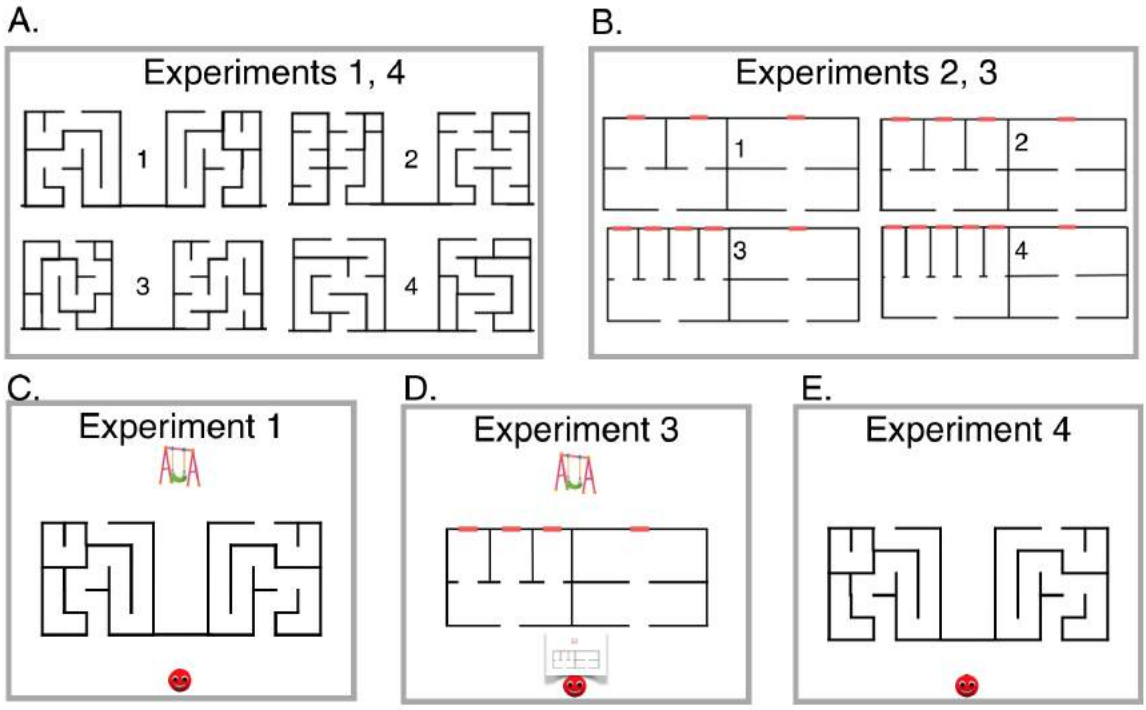


Figure 1: All mazes shown to participants on test trials in (A) Experiments 1 and 4, and (B) Experiments 2 and 3, as well as examples of individual trials from (C) Experiment 1, where a naive agent had the goal of getting to a specific location (D), Experiment 3, where a knowledgeable agent had the goal of getting to a specific location, and (E) Experiment 4, where a naive agent had the goal of getting better at solving mazes.

Experiment 1

In Experiment 1, we investigated whether young children whose attention is drawn to the difficulty of various mazes will choose easier rather than harder mazes for another agent to navigate.

Methods

Participants N=32 children (20 girls, Mean age = 58.94 months, range = 49.63-70.67 months) were included in our final sample of participants. All were recruited through a database of participants in the Boston area, participated at the Harvard Lab for Developmental Studies with the written consent of their parents, and received a small gift and travel compensation for their participation. One participant was excluded and replaced in our sample due to experimenter error. All data collection methods and procedures were approved by the Committee on the Use of Human Subjects at Harvard University. We chose our sample size based on a power analysis from a pilot study. For a pre-registration of the methods and analysis for this experiment, see https://osf.io/fx8yt/.

Materials and Procedure We built our maze stimuli using an online maze generator (http://www.mazegenerator.net/), using a width and height of 5 (4 mazes from test trials) or 6 (1 maze from introduction), an inner width and height of 0, an E-value of 50 (parameter that controls length of solution, relative to size of maze), and an R-value of 50 (parameter that controls length of dead ends subpaths). For the test trials, we selected 4 mazes that had at least 1 wrong turn, with 1 and only 1 solution, hereafter the harder' or more complex' mazes. To generate the simpler versions of these mazes, we added walls blocking all wrong turns, leaving only one available path through the maze. Throughout the experiment, each complex maze was always presented with the simpler version of itself flipped across the vertical axis. See Figure 1A. We presented all experimental materials using Keynote.

During the introduction to the experiment, children saw an animated agent, Bob, travel through an easier and harder maze. The agent's actions were realistic: he traveled through the easier maze without pause, but reached 2 dead ends in the harder maze before finding the solution. Children were asked Which maze took longer to go through?" and Which maze was harder to go through?" with feedback to check and reinforce their understanding of these scenes (e.g., "Yup, that one is harder!" or "Actually, *this* one is harder because it has more paths and ways to get lost").

In Experiment 1, children were told the following cover story: *Bob is at a playground and needs to go through mazes to get to things he wants to play with. He wants to play with as many things as possible before having to go home. He needs your help because he doesn't know anything about these mazes*.

On each test trial, Bob faced a choice between an easier and harder maze that lead to a piece of playground equipment (swings, monkey bars, slide, and a seesaw). Children were first asked to identify the easier (2 trials) or harder (2 trials) maze with feedback. Then, children were asked to help the agent choose which way to go. After participants chose a maze, they were asked to provide explanations for their response. Children viewed a bouncing animation of the agent next to his goal after every test trial, regardless of how they answered, and did not receive feedback for their choice.

We counterbalanced the order of the 4 maze pairs and the left-right position of the easier/harder maze, resulting in 8 different conditions of the procedure. The experiment lasted about 5 minutes.

Data and analysis All comprehension checks and test trials were coded on-line, and then checked offline from videos of the testing session.

We used the lme4 package (Bates, Mächler, Bolker, & Walker, 2015) in R (Team, 2015) to implement all generalized linear mixed effects models (GLMMs). All models with repeated measures included a random intercept for participant identity and maze identity. We used the ggplot2 package (Wickham, 2009) to produce Figure 2. The results sections of this paper were written in R Markdown (Allaire et al., 2014) to enhance reproducibility.

Results

In the introductory phase, prior to any feedback, children correctly identified the more difficult maze at a rate of 0.688, and the maze that would take longer to travel through at a rate of 0.969. During the test phase, which included feedback, children correctly identified the more difficult room at a rate of 0.586.

Our main question was whether children would preferentially choose the easier maze for the agent to travel through. We found that during test trials, children were more likely to select the easier maze than the hard maze, 95% confidence interval (CI) [2.262,10.05], B(SE)=4.932(2.526), z=1.952, p=0.026, one-tailed, OR=138.64, model syntax: `response ~ 1 + (1|subj) + (1|maze)`. Removal of influential cases yielded similar results. See Figure 2.

Discussion

Building on previous findings that infants and children expect agents to minimize the physical effort of their actions (Gergely & Csibra, 2003; Liu et al., 2017), the results of Experiment 1 suggest that children choose lower-effort tasks for others. Nevertheless, the question remains whether children were responding to the physical or the mental effort demands of the complex mazes. The more complex mazes presented a greater planning challenge for the mind, but were also associated with greater travel time and distance (variables that determine physical effort). We conducted Experiment 2 to ask whether children understand that actions can impose cognitive effort in the absence of differences in physical effort.

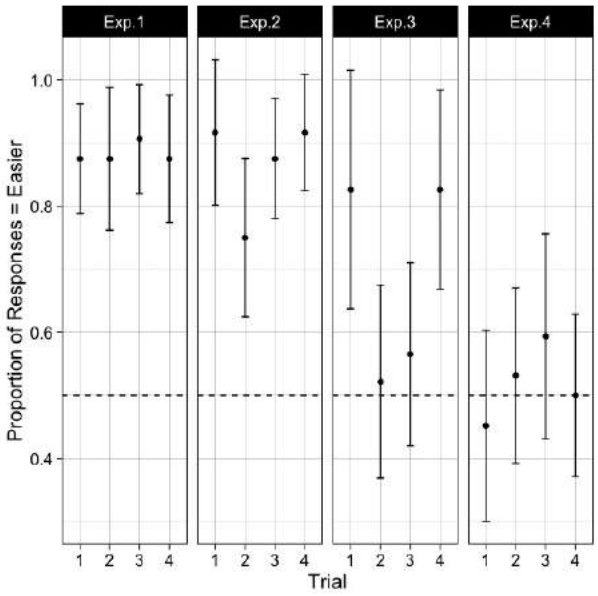


Figure 2: Proportion of choices for the easier maze or room during test trials (4 trials per participant; N=447 responses) across Experiments 1-4. Error bars indicate within-subjects 95% confidence intervals.

Furthermore, Experiment 1 asked children to choose a maze for an agent who was unaware of the effort involved in each choice, but did not ask children to predict which maze a knowledgeable agent would choose for themselves. We conducted Experiment 3 to ask whether children predict the actions of knowledgeable agents the same way they choose to help naive agents.

Lastly, Experiment 1 leaves open the question of whether children always regard mental and physical effort as negative, or whether they understand that harder actions can sometimes generate positive value. Thus, we conducted Experiment 4 to ask whether children appreciate the value of effort in the context of learning.

Experiment 2

In Experiment 2, we asked whether children appreciate that making a decision carries a unique cost, even when equating for physical effort across decision contexts. For a preregistration of the methods and analysis of this experiment, see https://osf.io/9dr7m/.

Methods

Participants N=24 children (14 girls, Mean age = 61.09 months, range = 48.5-71.43 months) were included in our final sample of participants. This sample size was chosen based on a power analysis from Experiment 1. One participant was excluded and replaced in the final sample due to parental interference.

Materials, Procedure, and Analysis Experiment 2 differed from Experiment 1 in three ways. First, the mazes from Experiment 1 were replaced with rooms (see Figure 1B). Pairs of rooms differed in the number of choices available: The more complex room featured multiple hallways for the agent to choose from, and the simpler room consisted of only one path. To equate for the dead ends, we designed these rooms so that all hallways were direct exits; regardless of whether the agent chose the easier or harder room, the agent would exit the first hallway she chose, reaching her goal. To prevent children from reasoning about the agent's line of sight through the rooms, we covered each outlet with a door, which opens only when the agent approaches it. Like in Experiment 1, children were told that the agent was naive about the contents of the rooms, and had the goal of reaching something beyond them. Second, during the introduction of the experiment, the agent moved through the easier and harder room in exactly the same way. This differed from Experiment 1, where the agent took several wrong turns in the harder maze. Third, before each test trial, children were asked to point at the room the agent thinks is harder or easier (2 questions of each kind), and which room the agent thinks has more or less choices (2 questions of each kind) with feedback.

Results

During the introduction to the experiment, prior to any feedback, children correctly identified the more difficult room at a rate of 0.667, and correctly identified the room with more choices at a rate of 0.917. During the test phase, which included feedback, children correctly identified the harder/easier room at a rate of 0.896, and the room with more/less choices at a rate of 0.667.

As in Experiment 1, children were more likely to select the easier room for the agent to travel through, 95% CI [2.278,13.514], B(SE)=7.432(2.871), z=2.588, p=0.005, one-tailed, OR=1689.047[1]. Removal of influential cases yielded similar results. Children's responses did not differ between Experiments 1 and 2, 95% CI [-3.057,2.364], B(SE)=-0.35(1.249), z=-0.28, p=0.779, two-tailed, OR=0.705[2]. See Figure 2.

Discussion

In Experiment 2 we asked whether children appreciated differences in decision complexity between two situations matched for physical path features like travel length and dead ends. As in Experiment 1, children discriminated between these decision structures and chose the simpler option for the naive agent. Together, Experiments 1-2 show that children appreciate the cognitive cost that enters decision-making. Nevertheless, it is less clear whether children expect other agents to willfully minimize their own mental effort, when asked to make a prediction about what a knowledgeable agent would do. Experiment 3 addresses this question.

[1]model syntax: `response ~ 1 + (1|subj) + (1|maze)`

[2]model syntax: `response ~ experiment + (1|subj) + (1|maze)`

Experiment 3

In Experiment 3, children predicted the choice of a knowledgeable agent in the same physical situations as in Experiment 2. For a pre-registration of the methods and analysis of this experiment, see https://osf.io/jyag8/.

Methods

Participants N=24 children (9 girls, Mean age = 60.27 months, range = 48.83-70.7 months) were included in our final sample of participants. This sample size was chosen based on a power analysis from Experiment 2. One participant was excluded due to experimenter error.

Materials, Procedure, and Analysis Experiment 3 was identical to Experiment 2 except that instead of helping the agent, children were asked to predict which room the agent will pick to go through in order to reach the goal, given that he knows everything about both of the rooms. To convey that the agent was knowledgeable, the agent on each test trial always had a map of the two rooms, and children were told explicitly that he knows everything about these rooms". See Figure 1D.

Results

In the introduction to the experiment, prior to any feedback, children correctly identified the more difficult room at a rate of 0.375[3], and the room with more choices at a rate of 0.792. During the test phase, which included feedback ("Yup that's right!" or "Actually, *this* room is easier/harder because he doesn't have to think about where to go") children correctly identified the more difficult room at a rate of 0.781 and the room with more/less choices at a rate of 0.698. As in Experiments 1 and 2, children were more likely to select the easier room for the agent to travel through, 95% CI [-0.296,2.642], B(SE)=1.026(0.598), z=1.716, p=0.043, one-tailed, OR=2.789[4]. Removal of influential cases yielded the same results. However, this effect was significantly weaker than the responses of children from Experiment 2, 95% CI [-4.7,-0.833], B(SE)=-2.397(0.914), z=-2.623, p=0.009, two-tailed, OR=0.091[5]. See Figure 2.

Discussion

In Experiment 3, we asked whether children expect knowledgeable agents to choose to minimize the mental effort of their actions. While we found a positive result, this effect was weaker than when children were asked to help a naive agent in identical environments. There are several possible interpretations of this finding. First, the agent's knowledge about the environments in Exp 3 could have affected children's responses: If a rational agent faces a false choice and knows it, and has a map of the rooms and has already analyzed the

[3]We too are puzzling over why this rate was lower than .5, and lower than in the other experiments.

[4]model syntax: `response ~ 1 + (1|subj) + (1|maze)`

[5]model syntax: `response ~ experiment + (1|subj) + (1|maze)`

choice structure of the two rooms, she may choose randomly. It is also possible that children's expectations about how others spend their mental effort is truly noisier than their intuitions about what is optimal. Regardless of these open questions, Experiments 2-3 provide evidence that children expect other agents to minimize the mental effort of their actions, both when predicting their actions, and when recruited to help them choose an action.

Experiment 4

In Experiment 4, we ask whether children appreciate the tradeoffs between mental effort and information gain. In other words, do children understand that sometimes, it is worthwhile to think and work hard? For a preregistration of the methods and analysis of this experiment, see https://osf.io/w3kh9/.

Methods

Participants N=32 children (17 girls, Mean age = 61.72 months, range = 50.0-71.0 months) were included in our final sample of participants. This sample size was chosen based on a power analysis from Experiment 1. Two participants were excluded and replaced in the final sample, 1 for not responding to any test trial questions, and 1 for experimenter error.

Materials, Procedure, and Analysis Experiment 4 was identical to Experiment 1, except that children were told a different cover story: *Bob wants to learn as much as he can about mazes. Which maze should he go through if he wants to practice solving mazes?* Children were asked what they thought the word practice' meant (19/32 produced passable definitions, like learning something you don't know how to do" and doing something until you know it so much"), and all children were told that to practice meant to try and try again so that you can get better at something". All goals were removed from test trials, and on each trial, and as in Exp 1-2, children were asked which way Bob should go.

Results

During the introduction to the experiment, prior to any feedback, children correctly identified the more difficult maze at a rate of 0.969, and correctly identified the maze that took a longer time to navigate at a rate of 0.875. During the test phase, which included feedback (e.g., "Yup, that one is harder!" or "Actually, *this* one is harder because it has more paths and ways to get lost"), children correctly identified the harder/easier room at a rate of 0.7086.

In contrast to Experiment 1, children in Experiment 4 did not preferentially choose the harder or easier room for the agent, 95% CI [-0.627,0.999], B(SE)=0.15(0.354), z=0.424, p=0.672, two-tailed, OR=1.162[6]. Removal of influential cases yielded similar results. As predicted under the hypothesis that children understand that effort trades off against increases in skill, their tendency to choose the easier maze in Experiment 4 was substantially lower than in Experiment 1, 95% CI [-4.368,-1.503], B(SE)=-2.936(0.731), z=-4.016, p<.001, one-tailed, OR=0.053[7]. See Figure 2.

Results, Experiments 1-4

Effects of experimental manipulations First, we asked which manipulations affected children's responses across all experiments. We found that children chose the easier vs harder action at comparable rates when shown the mazes from Experiments 1 and 4, and the rooms from Experiments, 2 and 3, 95% CI [-0.57,0.87], B(SE)=0.15(0.367), z=0.409, p=0.682, two-tailed, OR=1.162, that children were more likely to choose harder environments for a naive (Exp 3) than a knowledgeable agent (Exp 1, 2, 4), 95% CI [-1.566,-0.117], B(SE)=-0.842(0.369), z=-2.278, p=0.023, two-tailed, OR=0.431. Finally, we found that children were more likely to choose the harder environment when the agent had a learning goal (Exp 4) than an efficiency goal (Exp 1-3), 95% CI [0.793,2.122], B(SE)=1.457(0.339), z=4.298, p<.001, two-tailed, OR=4.293[8].

Role of feedback To address a concern that children's response to the test questions were influenced by the feedback they received during comprehension checks, we asked whether children's comprehension in Experiments 1-4 was different before they received any feedback (during the introduction) and after they began receiving feedback (during test trials 1-4). We found that children responded similarly prior to and after feedback (and if anything, performed less well with feedback), 95% CI [0.845,2.26], B(SE)=-0.243(0.186), z=-1.306, p=0.192, two-tailed, OR=0.785[9].

We also asked whether children's response to the main test question changed across the 4 trials of the experiment. If their responses were influenced by reinforcement during the comprehension checks, these responses should shift towards the direction of the hypothesis over the 4 trials. We tested this by fitting a model using Helmert contrasts, comparing children's responses on each test question (Which way should / will Bob go?") with their average responses on all preceding trials. Relative to all preceding trials, children did not clearly shift their response on trials two 95% CI [-0.669,0.036], B(SE)=-0.316(0.18), z=-1.76, p=0.078, two-tailed, OR=0.729, three 95% CI [-0.145,0.263], B(SE)=0.059(0.104), z=0.567, p=0.57, two-tailed, OR=1.061, or four 95% CI [-0.061,0.235], B(SE)=0.087(0.076), z=1.149, p=0.251, two-tailed, OR=1.091[10]. See Figure 1.

Discussion

Across Experiments 1 and 4, we found that children were more likely to choose a costly action in a context where the

[6]model syntax: `response ~ 1 + (1|subj) + (1|maze)`

[7]model syntax: `response ~ experiment + (1|subj) + (1|maze)`

[8]model syntax: `response ~ maze.or.room + knowledge + goal + (1|subj) + (1|maze) + (1|experiment)`

[9]model syntax: `response ~ phase + (1|experiment)+ (1|subj)`

[10]model syntax: `response ~ trial + (1|subj) + (1|experiment)`

actor's goal was to improve their planning abilities, versus when their plans were means to an end. Our findings show that children appreciate the trade-off between effort and information gain that working and thinking hard can generate.

General Discussion

Across four experiments, we asked whether children are sensitive to the mental and physical consequences of action selection in the context of mazes and rooms. Building on previous evidence that young children expect other people to minimize the physical cost of their actions (Gergely & Csibra, 2003; Liu & Spelke, 2017), we found that children assume complex maze environments are costly (relative to simpler ones), and that having to make choices is costly (relative making no choices). We also found that children do not expect agents to minimize effort in all situations, but instead appear to understand that trying hard is more likely to generate increases in knowledge and skill.

Within the limits of our experimental context, these results begin to reveal how young children reason about other's subjective mental effort costs. Specifically, in these experiments, children appear to place a cost on the process of action selection. This comports with a large literature showing that action selection by planning is, indeed, experienced by most people as costly (Kool & Botvinick, 2018; Westbrook & Braver, 2015; Shenhav et al., 2017). Nevertheless, the mechanisms by which children read out judgments of difficulty and use them to make predictions are not explored in this paper. In the domain of physical effort, past work suggests that even young infants represent action cost as force applied over a path, rather than as any single perceptual feature that correlates with more or less effortful actions (Liu et al., 2017). What information supports similar judgments in the domain of mental effort? Furthermore, it is unclear how much or how little children rely on processes of simulation to solve the tasks in our experiment. Most of the preschoolers in our sample probably came into the lab with prior experience solving mazes, and many of them traced paths through the mazes as part of their explanations for why they answered the way they did. Thus, one important remaining question is what role our experiences of choosing, thinking, and learning play in the development of our understanding of mental effort.

Of course, action selection is not the only costly step of rational planning, or the only difference between habits and plans. Our results thus suggest important new directions for future research. For instance, do children understand that the closer 2 options are in utility, the harder it is to choose between them, or that habits are lower in cost than plans? Our findings also opens the door to studies of children's intuitive theories of other people's and their own knowledge and learning. For instance, do children understand that learners have an optimal zone of task difficulty in which to gain knowledge? Future work in this area can address the many open questions regarding how we conceptualize the mental lives of other people, and its development.

Acknowledgments

Many thanks to the families who volunteered to participate, to the Harvard Lab for Developmental Studies for discussion and feedback, and to Aracely Aguirre, Akshita Srinivasan, Nensi Gjata, and Caitlin Connolly for help with data collection. This material is based upon work supported by the Center for Brains, Minds and Machines (CBMM), funded by National Science Foundation STC award CCF-1231216, and by a National Science Foundation Graduate Research Fellowship under grant DGE-1144152.

References

Allaire, J. J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., ... Chang, W. (2014). *rmarkdown: Dynamic documents for R.*

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, *1*(March), 0064.

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329–349.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.*, *67*(1).

Flavell, J. H., Friedrichs, A. G., & Hoyt, J. D. (1970, October). Developmental changes in memorization processes. *Cogn. Psychol.*, *1*(4), 324–340.

Flavell, J. H., Green, F. L., & Flavell, E. R. (1995). Young children's knowledge about thinking. *Monogr. Soc. Res. Child Dev.*, *60*(1), 1–96; discussion 97–114.

Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: The naïve theory of rational action. *Trends Cogn. Sci.*, *7*(7), 287–292.

Gershman, S. J., Gerstenberg, T., Baker, C. L., & Cushman, F. A. (2016, September). Plans, habits, and theory of mind. *PLoS One*, *11*(9), e0162246.

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends Cogn. Sci.*, *20*(8), 589–604.

Jara-Ettinger, J., Tenenbaum, J. B., & Schulz, L. E. (2015). Not so innocent: Toddlers' inferences about costs and culpability. *Psychol. Sci.*.

Kool, W., & Botvinick, M. (2018, September). Mental labour. *Nature Human Behaviour*.

Kool, W., Gershman, S. J., & Cushman, F. A. (2017, September). Cost-Benefit arbitration between multiple Reinforcement-Learning systems. *Psychol. Sci.*, *28*(9), 1321–1333.

Leonard, J. A., Lee, Y., & Schulz, L. E. (2017, September). Infants make more attempts to achieve a goal when they see adults persist. *Science*, *357*(6357), 1290–1294.

Liu, S., & Spelke, E. S. (2017). Six-month-old infants expect agents to minimize the cost of their actions. *Cognition*, *160*, 35–42.

Liu, S., Ullman, T. D., Tenenbaum, J. B., & Spelke, E. S. (2017, November). Ten-month-old infants infer the value of goals from the costs of actions. *Science*, *358*(6366), 1038–1041.

Shenhav, A., Musslick, S., Lieder, F., Kool, W., Griffiths, T. L., Cohen, J. D., & Botvinick, M. M. (2017, July). Toward a rational and mechanistic account of mental effort. *Annu. Rev. Neurosci.*, *40*, 99–124.

Team, R. D. C. (2015). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing.

Wellman, H. M. (2002, January). Understanding the psychological world: Developing a theory of mind. In U. Goswami (Ed.), *Blackwell handbook of childhood cognitive development* (pp. 167–187). Malden, MA, USA: Blackwell Publishers Ltd.

Westbrook, A., & Braver, T. S. (2015, June). Cognitive effort: A neuroeconomic approach. *Cogn. Affect. Behav. Neurosci.*, *15*(2), 395–415.

Wickham, H. (2009). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag.

People's perception of others' risk preferences

Shari Liu (shariliu01@g.harvard.edu) and **Tomer Ullman (tullman@fas.harvard.edu)**
Department of Psychology, Harvard University
Cambridge, MA 02143 USA

John McCoy (jpmccoy@wharton.upenn.edu)
The Wharton School, University of Pennsylvania
Philadelphia, PA 19104 USA

Abstract

Our everyday decisions are driven by costs, risk, and reward. How do people take these factors into account when they predict and explain the decisions of others? In a two-part experiment, we assessed people's perceptions of other people's risk preferences, relative to their own. In Part 1, participants reported their relative preference between a guaranteed payout and lotteries with various probabilities and payouts, and made predictions about other people's preferences. In Part 2, participants estimated the lottery payout that generated a given relative preference between a guaranteed payout and a lottery, both for themselves and others. We found considerable individual variability in how people perceive the risk preferences of others relative to their own, and consistency in people's perceptions across our two measures. Future directions include formal computational models and developmental studies of how we think about our own and each other's decision-making.

Keywords: intuitive psychology; decision making; risk

Introduction

Humans are social beings, who spend much of their time attempting to predict what decisions others will make, and explain why others chose as they did. Adults, and even infants, make predictions about what another person will do based on their beliefs about the person's mental state, and also make inferences about someone's mental state after observing their behavior (Epley, 2015; Kushnir, Xu, & Wellman, 2010; Repacholi & Gopnik, 1997).

Recent computational accounts of such abilities see people as performing Bayesian inference using a model of others as rational planners or intuitive utility maximizers who take actions to maximize their expected reward relative to their incurred cost (Baker, Saxe, & Tenenbaum, 2009; Lucas et al., 2014; Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016; Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017). Previous work has shown that such rewards and costs are early-emerging, separate targets of inference (Jara-Ettinger, Gweon, Tenenbaum, & Schulz, 2015; Liu, Ullman, Tenenbaum, & Spelke, 2017). Here, we study a related variable at the heart of other people's expected utility: the probability of the outcome. Specifically, we study how people perceive and reason about other people's risk preferences, especially compared to their own.

The central role of risk in decision making has been long appreciated (Bernoulli, 1738). For example, many people prefer a 50/50 chance of losing $200 to losing $100 for sure, and prefer gaining $100 for sure over a 50/50 chance of gaining $200, even though the expected value of the options are equal in each case. Under expected utility theory, decision makers weight probabilities linearly, and risk aversion is measured and explained by the curvature of the utility function (Pratt, 1964; Arrow, 1965). Rabin's Calibration Theorem illustrates the difficulties with this approach (Rabin, 2000). A large body of work by psychologists and behavioral economists has shown that decision making under risk involves non-linear weighting of probabilities (Kahneman & Tversky, 1979; Tversky & Kahneman, 1992; Wakker, 2010; Dhami, 2016).

Research has also examined how people perceive the risk sensitivity of others. Some previous work finds that people perceive others as more risk-seeking than themselves (Hsee & Weber, 1997), while a different set of studies finds that, on average, people assume others are more risk-averse (Eckel & Grossman, 2008), although the focus of this study and others (Siegrist, Cvetkovich, & Gutscher, 2002) was the role of gender stereotypes in risk perception. Differences in cross-national risk perceptions have also been explored (Hsee & Weber, 1999). This paper differs from previous work in a number of ways. First, the previous literature used group-based analyses that collapsed across people and a small number of gambles, and so could not determine whether the average results reflected homogeneous perceptions across individuals, whereas we additionally consider individual level perceptions. Second, participants in past studies made predictions about binary choices between lotteries, whereas in our study participants give more fine-grained predictions about their degree of relative preferences for a lottery over a sure thing. Third, participants previously only made predictions about the decisions of others, whereas we additionally have participants estimate the monetary value of gambles that would cause a particular preference in other people.

We use a two-part experiment to study how people perceive the risk preferences of others. In Part 1, we present participants with choices between $100 for sure and a lottery, with eight levels of payout and five levels of probability. For each choice, participants reported their own preferences and predicted the preferences of others, using a five point Likert scale. In Part 2, we ask the same participants to estimate the (unseen) payout that led others to report a specific preference, and that would lead themselves to report the same preference. We then relate the judgments of participants across the two

parts of the experiment.

We had two main research questions. First, we were interested in the distribution of people's perception of their own risk preferences relative to others. Second, we investigated the consistency of people's perceptions about their own and others' risk preferences across two tasks, one that asked people to make predictions about preferences, and the other that asked people to make inferences about lottery payouts given preferences.

Experiment

Participants

We recruited 205 participants on Amazon Mechanical Turk, restricted to the United States. Of these, we excluded 33 participants for (1) failing to pass an attention check, or (2) providing the same answer for all questions in Parts 1 or 2, or (3) taking less than 5 minutes to complete the experiment, or (4) giving payout judgments larger than $2000 in Part 2. These criteria were specified ahead of data analysis but not ahead of data collection. After exclusion our sample consisted of 172 participants (median age=34 years, median annual income=$47,000, 75 female, 96 male, 1 other). All participants gave informed consent prior to participating. All recruitment and study procedures were approved by the MIT Committee on the Use of Humans as Experimental Subjects.

Methods

Participants were presented with a series of hypothetical choices between lotteries and $100 for sure. Each lottery consisted of a random draw from a box of 10 balls. If a player were to enter the lottery, a ball would be drawn at random from the box, and the player would win the amount of money on the ball. For example, a lottery where a player has a 50-50 chance to win $500 would contain 5 balls worth $0 each, and 5 balls worth $500 each.

In Part 1, participants saw 40 trials, each involving a choice between $100 for sure, or a [.1, .3, .5, .7, or .9] chance of winning [$100, $150, $200, $300, $400, $600, $800, or $1000]. For each decision, participants gave their own preference, and predicted the preference of an average other player, on a 5-point Likert scale (1=$100 is a lot better, 2=$100 is somewhat better, 3=$100 and lottery are equally good, 4=lottery is somewhat better, and 5=lottery is a lot better). We note that these Likert ratings do not express participants' valuation of the lottery itself, but rather differences between the utility of the lottery and the utility of the sure reward. See Figure 1.

In Part 2, participants saw 5 trials, each involving a choice between $100 for sure, or a 50-50 lottery to win some other amount of money, this time unknown to the participant (Figure 1, bottom). On each trial, participants were informed that, on average, other players rated the lottery one of the five possible levels of the Likert scale (i.e., on the first trial participants were told that other players on average strongly preferred the $100, on the second trial that other players slightly preferred the $100, etc.). Participants were asked to estimate how much money was at stake in the lottery given this preference, and gave their response using a freeform text field.

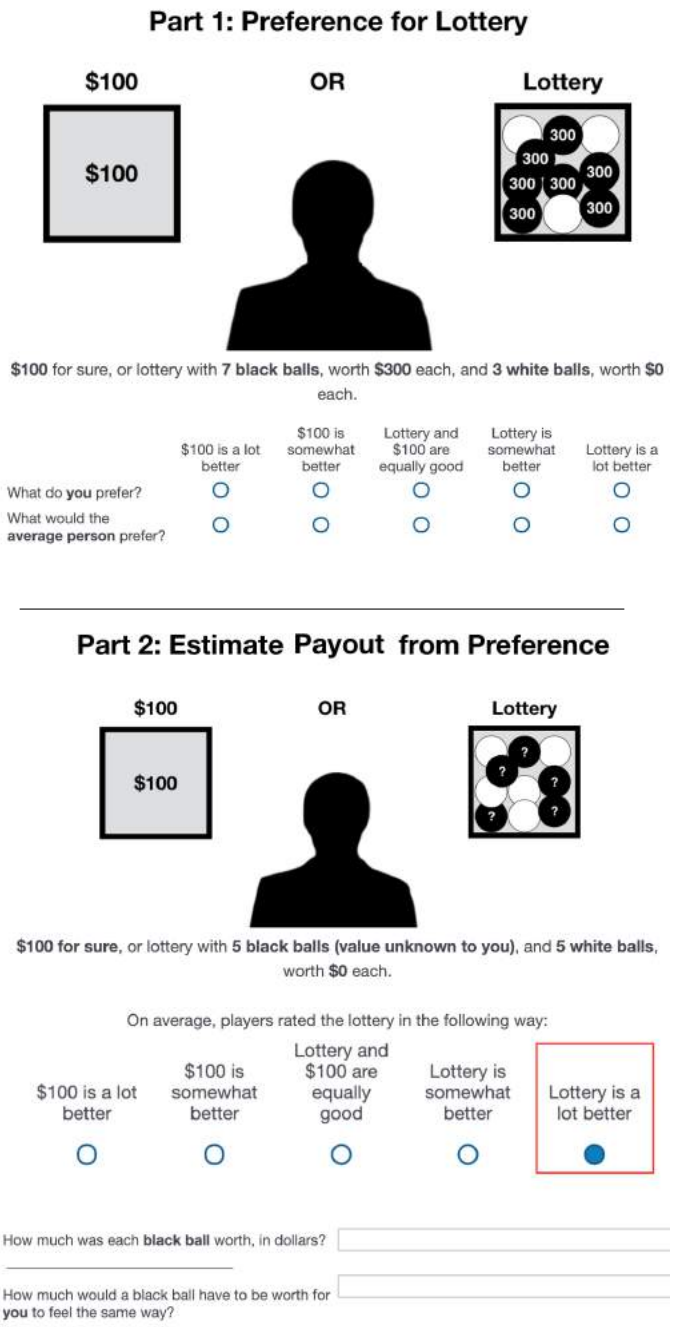


Figure 1: Example trials from the experiment. In Part 1, participants rated their own preferences between $100 for sure and a lottery, and predicted the preference of others. In Part 2, participants were told the preference of another person and both estimated the payout of the lottery, and judged how much money would have to be at stake for them to feel the same way.

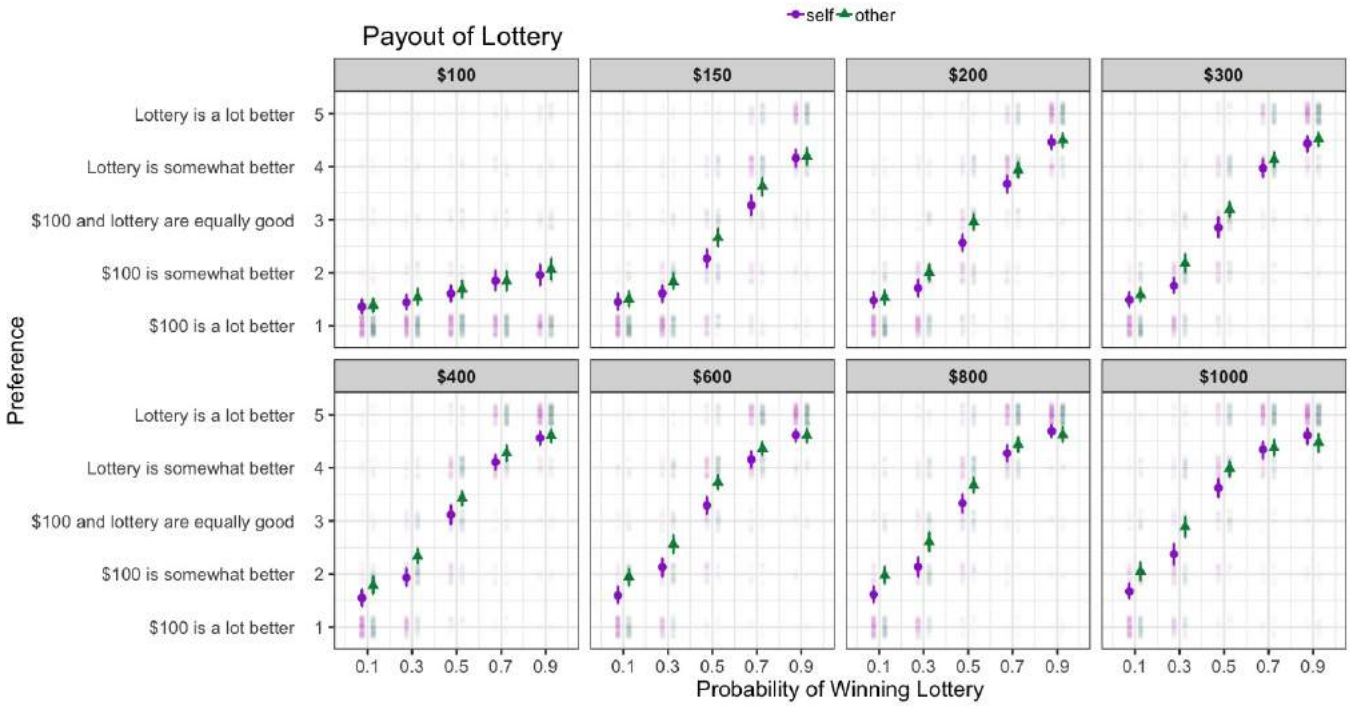


Figure 2: Participant Likert ratings indicating preference for lottery or guaranteed $100, and their predictions for the average other player, across all probabilities and payouts. Opaque points indicate mean Likert ratings with bootstrapped 95% confidence intervals. Translucent points indicate raw data plotted with vertical jitter.

Participants then estimated how much money would have to be at stake in the lottery for they themselves to give the same rating. See Figure 1.

In Parts 1 and 2, trials were presented in a random order, and the left-right orientation of the lottery vs $100 and the anchors for the Likert scale were consistent within participants, but randomized across participants.

Results

The data were analyzed using mixed effects linear models (Bates, Mächler, Bolker, & Walker, 2015; Team, 2015), unless noted otherwise. All models included random intercepts for participant identity (i.e. responses are nested within participants), and for trial number (i.e. responses are nested within linear trial order). We report coefficients from modeling the Likert rating as continuous for ease of interpretation, but fitting a Cumulative Link Model yields similar results. Bracketed values indicate 95% confidence intervals of unstandardized coefficients (e.g. the effect of increasing the stake of the lottery by $1 on preferences for the lottery in Likert ratings), and p-values are all two-tailed. Participant gender and annual household income are included as regressors.

Part 1: Preferences between a lottery and sure thing

People's own risk preferences. Before turning to our first question concerning how people perceive the risk sensitivity of others compared to their own risk sensitivity, we conducted a basic analysis of the data to confirm that 1) people more strongly preferred the lottery as its probability and payout increased, and 2) whether people, on average, were risk averse. As expected, across all 40 trials, participants' preference for the lottery increased as the payout increased ([1.2e-3,1.4e-3], $p<.001$), and as the probability of winning increased ([3.456,3.637], $p<.001$), see Figure 2.[1]. To measure participants' level of risk aversion, we examined the two trials that included lotteries equal in expected value to receiving a guaranteed $100 (i.e. the 50-50 lottery with $200 payout, and the 10-90 lottery with $1000 payout). In both of these trials, people preferred the guaranteed $100 over the lottery (Likert mean=2.56, median=2 for 50-50 lottery, $p<.001$; mean rating=1.67, median=1 for 10-90 lottery, $p<.001$, one-sample t-test against $\mu=3$).

Perceptions of risk preferences of others. We repeated the same basic analyses as reported above, this time on people's judgments of others. Across all 40 trials, participants' estimates of others' Likert ratings increased as the payout increased ([1.3e-3,1.5e-3], $p<.001$), and as the probability of winning increased ([3.171, 3.353], $p<.001$). We again analyzed the two trials that included lotteries with an expected value of $100. In the trial with the 50-50 lottery, participants predicted that others would be indifferent between the

[1]Model formula: `response ~ payout + probability + gender + log(income) + (1|participant) + (1|trial)`

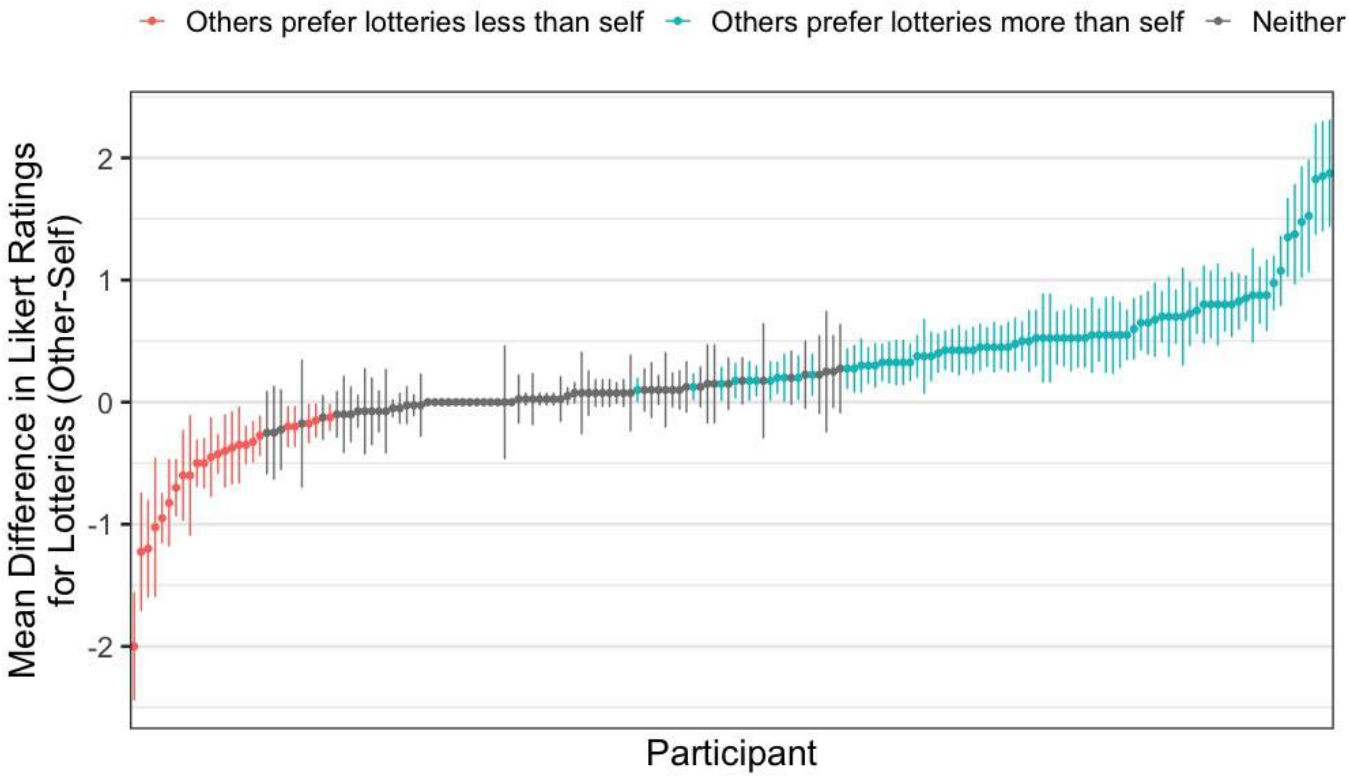


Figure 3: Differences in people's own preference for the lottery and their prediction of the preference of others, shown for each participant and averaged across trials. Each point indicates the mean difference between a single participant's Likert ratings for themselves and predicted ratings for others (40 pairs per participant), estimated by a paired-samples t test. Error bars indicate 95% confidence intervals around the mean. Participants without confidence intervals gave the same rating for themselves and others on every trial (but different ratings across trials). Of 172 participants, 81 (47%, blue) judged that others were significantly more risk-seeking than themselves, 24 (14%, red) judged that others were significantly more risk-averse than themselves, and 67 (39%, gray) gave similar ratings for themselves and others as discussed in the text.

50-50 lottery and the guaranteed \$100 (mean rating=2.96, median=3, p=.602, one-sample t-test against μ=3). In the trial with the 10-90 lottery, participants predicted that others would slightly prefer the \$100 (mean rating=2.05, median=2, $p<0.001$, one-sample t-test against μ=3).

Comparing risk preferences for self and other. Our first main question is how people perceive the risk preferences of other people, relative to their own. A group level analysis indicated that participants predicted other people to be more risk-seeking than themselves. That is, they expected others to prefer the lottery (rather than the guaranteed \$100) more than themselves, across payout amounts and lottery probabilities ([0.086,0.123], $p<.001$)[2]. See Figure 2.

A group level analysis, however, can obscure important individual heterogeneity. Figure 3 shows, for each participant, the average difference between the participant's prediction of how much other people prefer the lottery and how much they themselves prefer the lottery. While these differences clearly fall on a continuum, we were interested in what proportion of participants judged that others were more risk averse or risk seeking, relative to themselves. By this measure, 47% of participants believed that others were more risk-seeking than themselves (by a paired sample t-test on each participant, the estimated average difference for these participants had confidence intervals strictly above 0), 14% believed others were less risk-seeking (confidence intervals strictly below 0), and 39% did not show a significant difference between their preferences and those they predicted for others (confidence intervals crossed 0). These results are consistent with a paired sample sign test on each participant, which identifies 43% of participants who believe that others are more risk-seeking, 9% who believe that others are more risk-averse, and 48% who do not show a significant difference between the ratings of themselves and others (all at the $p<.001$ level).

[2]Model formula: `response ~ probability + payout + agent + gender + log(income) + (1|participant) + (1|trial)`

Part 2: Estimating the payout of lotteries given a preference

As a reminder, in Part 2 we asked participants what payout of a 50-50 lottery would cause themselves and others to have a given preference for the lottery (from much preferring the guaranteed $100, to much preferring the lottery).

Estimates of lottery payouts for self. As in our analysis of the data from Part 1, we first conducted a basic analysis to examine people's inferences about the lottery conditional on choices for themselves. Across all trials, participants reasonably believed that a greater preference for the lottery meant that the payout of the lottery was higher ([112.65,131.60], $p<.001$)[3]. When the Likert rating was 3 (indifferent between the lottery and guaranteed $100), participants judged that the 50-50 lottery payout must exceed $200 for them to have given this rating (mean=$267, median=$200, $p<.001$, one-sample t-test against μ=$200), indicating risk aversion.

Estimates of lottery payouts for others. We repeated the same analyses as reported above, this time on people's judgments of others. Across all five trials, participants judged that other people having a greater preference for the lottery was caused by a higher lottery payout ([110.15,127.90], $p<.001$). When the average Likert rating reported by others was 3 (indifferent between the lottery and guaranteed $100), participants estimated that the 50-50 lottery payout for other people was no different than the risk-neutral value of $200 (mean=$210, median=$200, p=.384, one-sample t-test against μ=$200).

Comparing payout estimates for self and other. Our first main research question concerns differences in participant's estimates for the lottery payment for themselves and others. Across all participants and trials, for a given Likert rating participants judged that the estimated lottery payout was lower for other people than for themselves (agent coefficient was significantly negative, [-24.983,-6.939], p=0.001)[4].

Since there are only five trials per participant in Part 2 (5 paired estimates for self and other), to assess individual level differences, we computed for each participant the mean difference between the payout that they believed would be required to make the lottery equally attractive to themselves and other people: 51% of participants gave higher estimates, on average, for themselves than others (i.e. believed that others were more risk-seeking), 29% gave lower estimates, on average, for themselves compared to others (i.e. believed that others were more risk-averse), and 20% gave, on average, equal estimates for themselves and others (i.e. believed that they and others had the same risk preference).

Predictions of preferences and estimates of lottery payouts given a preference Our second main research question was whether people's judgments were consistent across our two tasks. We found that participants' average difference in Part 1 between their own preferences and their ratings of the preferences of others corresponded to the average difference between their payout estimates for themselves and others in Part 2 ([46.2, 102.1], $p<.001$).[5]. That is, the more a participant judged that others would prefer the lottery more than themselves in Part 1, the more lottery payout that participant needed to give the same rating as others in Part 2.[6]. See Figure 4.

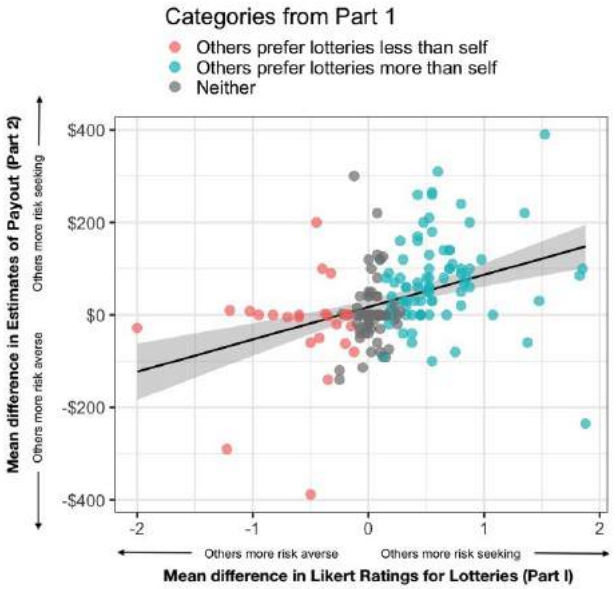


Figure 4: Relating individual differences in relative risk preference as measured in Part 1 and Part 2 of the experiment. Each point indicates one person's mean difference in preferences for 40 lotteries for themselves vs others (x-axis), and the mean difference in their estimated payout for themselves vs. others over 5 preferences for a 50-50 lottery vs a guaranteed $100 (y-axis). Solid line indicates regression between these values with 95% confidence interval. Colors of points indicate classification of participants based on Part 1, as in Figure 3 and the main text.

Discussion

Risk matters, both for our own decisions, and in our reasoning about the decisions of others. We presented participants with choices between lotteries and guaranteed payouts, and used prediction and estimation measures to explore individual variability in people's beliefs about the risk preferences of others. Across both measures, we found two large subsets of participants: participants who believed that others were more risk seeking than themselves, and participants who believed that other people exhibited roughly the same degree of risk sensitivity as themselves. People's beliefs about how their

[3]Model formula: `response ~ likert + gender + log(income) + (1|participant) + (1|trial)`

[4]Model formula: `response ~ probability + payout + agent + gender + log(income) + (1|participant) + (1|trial)`

[5]Model formula: `diffpart2 ~ diffpart1 + gender + log(income)`

[6]Performing the same comparison using Spearman's rank correlation yielded similar results, ρ=0.472, $p<.001$

own risk sensitivity compared to other people were fairly stable across the two parts of the experiment.

Our findings are consistent with, but also complicate, the framework of Bayesian Theory of Mind. This framework models people's reasoning about others by assuming that others are carrying out a rational planning procedure to achieve goals given constraints (Baker et al., 2009, 2017; Jara-Ettinger et al., 2016). While most previous work in this literature assumed, for simplicity, that people reason about others as maximizing expected value, behavioral economics has long highlighted how people deviate from simple expected value (for example, by being risk-averse) (Dhami, 2016). Recent work has investigated deviations from optimal rational planning and the use of bounded agents in Bayesian Theory of Mind and Inverse Reinforcement Learning, for example by replacing the ideal rational planner with an agent that has false beliefs and exhibited temporal inconsistency (Evans, Stuhlmüller, & Goodman, 2016). Along the same lines, one could replace the rational planner with an agent that displayed either risk-seeking or risk-averse behavior, for example either by manipulating the agent's utility function or its probability weighting function. We are currently pursuing this direction so as to explore the cognitive processes underlying the results presented in this paper.

Another future direction suggested by the results in this paper are the downstream consequences of differences in people's own risk sensitivity and their perception of the risk sensitivity of others. For example, do people use their own or their perception of others' risk preference when making decisions on behalf of others? What do people expect others to do, when others are assigned to make decisions on their behalf?

Our experiment focused on risk of a specific kind, but risk may not be a unified concept (Loewenstein, Weber, Hsee, & Welch, 2001; Wallach & Wing, 1968). Moreover, most situations are ambiguous rather than simply risky - people are not confronted with explicit, known probabilities, but must instead act in the face of uncertainty given their beliefs. Similar experiments could examine how people perceive the degree to which other people exhibit ambiguity aversion, relative to their own ambiguity preferences.

While all our participants were adults, it is interesting to consider perceptions of other's risk sensitivity through the lens of development. Infants and children are sensitive to other people's preferences (Woodward, 1998; Jara-Ettinger et al., 2015), and the probabilities of events (Téglás et al., 2011; Xu & Garcia, 2008, 2008). Recent studies suggest that children use probability (Denison & Xu, 2010, 2014) and reward (Feigenson, Carey, & Hauser, 2002) to make decisions and analyze the decisions of others (Wellman, Kushnir, Xu, & Brink, 2016; Lucas et al., 2014). But these experiments leave open when children become sensitive to risk in their own decisions, and when they understand others as risk-sensitive.

In this paper, we examined risk in the context of a series of simple lotteries. This is a common laboratory paradigm, but is less common in real life. Outside the lab, risk is a major force in consequential decisions, from starting wars, to developing new technologies, to making medical decisions for ourselves and our loved ones. Such decisions are not made in isolation, but in consultation, collaboration, and competition with other people. Thus, studies of risk—a fundamental component of our decisions and social lives—bear on all of these situations, by revealing the nature of how we represent other people's decisions, and our own.

Acknowledgments

This material is based upon work supported by the Center for Brains, Minds and Machines (CBMM), funded by National Science Foundation STC award CCF-1231216, and by a National Science Foundation Graduate Research Fellowship under grant DGE-1144152.

References

Arrow, K. J. (1965). Aspects ofthe theory of risk-bearing. *Helsinki: Academic Bookstore*.

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, *1*(March), 0064.

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329–349.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.*, *67*(1).

Bernoulli, D. (1738). Specimen theoriae novae de mensura sortis. *Commentarii Academiae Scien-tiarum Imperialis Petropolitanae (Volume V)*. (Translated by L. Sommer as Exposition of a New The- ory on the Measurement of Risk, Econometrica, Jan. 1954, 22(1), pp. 23-36)

Denison, S., & Xu, F. (2010, September). Twelve- to 14-month-old infants can predict single-event probability with large set sizes. *Dev. Sci.*, *13*(5), 798–803.

Denison, S., & Xu, F. (2014). The origins of probabilistic inference in human infants. *Cognition*, *130*(3), 335–347.

Dhami, S. (2016). *The foundations of behavioral economic analysis*. Oxford University Press.

Eckel, C. C., & Grossman, P. J. (2008, October). Forecasting risk attitudes: An experimental study using actual and forecast gamble choices. *J. Econ. Behav. Organ.*, *68*(1), 1–17.

Epley, N. (2015). *Mindwise: How we understand what others think, believe, feel, and want*. Vintage.

Evans, O., Stuhlmüller, A., & Goodman, N. D. (2016). Learning the preferences of ignorant, inconsistent agents. In *Aaai* (pp. 323–329).

Feigenson, L., Carey, S., & Hauser, M. (2002). The representations underlying infants' choice of more: Object files versus analog magnitudes. *Psychol. Sci.*, *13*(2), 150–156.

Hsee, C. K., & Weber, E. U. (1997). A fundamental prediction error: Self-others discrepancies in risk preference. *J. Exp. Psychol. Gen.*, *126*(1), 45–53.

Hsee, C. K., & Weber, E. U. (1999). Cross-national differences in risk preference and lay predictions. *J. Behav. Decis. Mak.*, *12*(2), 165–179.

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends Cogn. Sci.*, *20*(8), 589–604.

Jara-Ettinger, J., Gweon, H., Tenenbaum, J. B., & Schulz, L. E. (2015). Children's understanding of the costs and rewards underlying rational action. *Cognition*, *140*, 14–23.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263–291.

Kushnir, T., Xu, F., & Wellman, H. M. (2010). Young children use statistical sampling to infer the preferences of other people. *Psychological science*, *21*(8), 1134–1140.

Liu, S., Ullman, T. D., Tenenbaum, J. B., & Spelke, E. S. (2017, November). Ten-month-old infants infer the value of goals from the costs of actions. *Science*, *358*(6366), 1038–1041.

Loewenstein, G. F., Weber, E. U., Hsee, C. K., & Welch, N. (2001, March). Risk as feelings. *Psychol. Bull.*, *127*(2), 267–286.

Lucas, C. G., Griffiths, T. L., Xu, F., Fawcett, C., Gopnik, A., Kushnir, T., ... Hu, J. (2014). The child as econometrician: A rational model of preference understanding in children. *PLoS One*, *9*(3).

Pratt, J. W. (1964). Risk aversion in the small and in the large. *Econometrica*, *32*(1/2), 122–136.

Rabin, M. (2000). Risk aversion and expected-utility theory: a calibration theorem. *Econometrica*, *68*(5), 1281.

Repacholi, B. M., & Gopnik, A. (1997). Early reasoning about desires: evidence from 14-and 18-month-olds. *Developmental psychology*, *33*(1), 12.

Siegrist, M., Cvetkovich, G., & Gutscher, H. (2002, January). Risk preference predictions and gender stereotypes. *Organ. Behav. Hum. Decis. Process.*, *87*(1), 91–102.

Team, R. D. C. (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Téglás, E., Vul, E., Girotto, V., Gonzalez, M., Tenenbaum, J. B., & Bonatti, L. L. (2011). Pure reasoning in 12-month-old infants as probabilistic inference. *Science*, *332*(6033), 1054–1059.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *J. Risk Uncertain.*, *5*(4), 297–323.

Wakker, P. P. (2010). *Prospect theory: For risk and ambiguity*. Cambridge university press.

Wallach, M. A., & Wing, C. W. (1968, May). Is risk a value? *J. Pers. Soc. Psychol.*, *9*(1), 101–106.

Wellman, H. M., Kushnir, T., Xu, F., & Brink, K. A. (2016, September). Infants use statistical sampling to understand the psychological world. *Infancy*, *21*(5), 668–676.

Woodward, A. L. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition*, *69*(1), 1–34.

Xu, F., & Garcia, V. (2008). Intuitive statistics by 8-month-old infants. *Proc. Natl. Acad. Sci. U. S. A.*, *105*(13), 5012–5015.

Verb Frequency Explains the Unacceptability of Factive and Manner-of-speaking Islands in English

Yingtong Liu[1,2] (y_liu@g.harvard.edu), Rachel Ryskin[2,4] (ryskin@mit.edu), Richard Futrell[3] (rfutrell@uci.edu) , Edward Gibson[2] (egibson@mit.edu)

[1] Department of Linguistics, Harvard University, Cambridge, MA 02138 USA
[2] Department of Brain and Cognitive Sciences, MIT, Cambridge, MA 02139 USA
[3] Department of Language Science, University of California, Irvine, CA 92697 USA
[4] Department of Speech, Language, and Hearing Science, Boston University, Boston, MA 02215 USA

Abstract

The unacceptability of wh-extraction (e.g., question formation) out of certain syntactic structures, known as 'island' effects, has been a central topic in theoretical syntax for many years (Ross, 1967; Chomsky, 1973). A prominent example of islands is that extraction out of a sentential complement introduced by factive and manner-of-speaking verbs ('What did John *know/whisper* that Mary bought?') is less acceptable than extraction from a clause introduced by "bridge" verbs ('What did John *say* that Mary bought?'). We aimed to replicate Ambridge and Goldberg (2008) who argued that extraction from a sentential complement is unacceptable in proportion to its discourse salience. We failed to replicate their results and found that there is no true island effect for such structures: instead there are separate, additive penalties based on two factors: (a) verb-frame frequency (cf. Dabrowska, 2008), and (b) the presence of extraction. These penalties give rise to apparent island effects as a result of the nonlinear relationship between true acceptability and acceptability ratings as measured in Likert scales and forced-choice tasks.

Keywords: Sentence Processing; Frequency Effect; Acceptability of Sentences; Long-distance Dependencies

Introduction

An important feature of human languages is that they contain constructions that license long-distance dependencies--so-called "filler-gap" constructions, such as wh-questions, relative clauses, clefts and topicalization, among others. For example, the declarative form of a simple clause is provided in (1a), along with a wh-question version of this information in (1b), where the patient (object) is extracted. A corresponding relative clause is provided in (1c) and a cleft is in (1d)[1]:

(1) a. Mary bought some apples.
b. wh-question: What$_i$ did Mary buy __$_i$?
c. relative clause: The apple that$_i$ Mary bought __$_i$
d. cleft: It was the apple that$_i$ Mary bought __$_i$

[1] Following standard notation in the linguistics literature, we will notate the position in the declarative that is associated with fronted element with an empty element "__". We provide a subscript such as "i" to the fronted element (the "filler") and the empty position.

Some long-distance extractions are allowed (1), but others are not (2)&(3) (Ross, 1967):
(2) a. * What$_i$ did [$_S$ you hear [$_{NP}$ the statement that Jeff baked __$_i$]] ?
b. * Who$_i$ do [$_S$ you think [$_{NP}$ the gift from __$_i$] prompted the rumor] ?

(3) (relative clause versions of 2):
a. * The bread that$_i$ [$_S$ you heard [$_{NP}$ the statement that Jeff baked __$_i$]]
b. * The politician who$_i$ [$_S$ you think [$_{NP}$ the gift from __$_i$] prompted the rumor].

The unacceptable versions in (2) and (3) have been called 'islands' to extraction: unacceptable long-distance filler-gap constructions. The major theoretical interest in island phenomena began with Chomsky (1973), where he argued for a pure structural account, *Subjacency*: noun phrase (NP) and sentence (S) syntactic nodes are bounding nodes for extraction. Extraction across two bounding nodes was proposed to be ungrammatical. Consequently, extractions across the NP and S nodes in (2ab) and (3ab) result in an unacceptable form. Furthermore, Chomsky argued that these constraints must be innate and unlearnable, because (a) the unacceptable extractions occur independent of the meaning of the constructions involved; and (b) a child would not be exposed to the right input across all the different constructions in which they hold (see Schütze et al., 2015, for a summary).

In this paper we focus on extractions out of sentence-complements (S-complements) of factive and manner-of-speaking verbs, as in (4). Researchers have long noted that extraction out of sentence-complements taken by factive verbs – such as "know" (4b), "regret", and "notice", whose S-complements are presupposed (Kiparsky and Kiparsky, 1971) – and manner-of-speaking verbs – such as "whisper" (4c) "mutter", and "mumble", which describe physical characteristics of the speech act (Zwicky, 1971) – are less acceptable than extraction out of "bridge" S-complement taking verbs (4a) (e.g., Erteschik-Shir. 1973; Snyder, 1992; Ambridge & Goldberg, 2008). Note that the definition of a "bridge" verb is not independently defined. A bridge verb is simply one for which extraction from its S-complement is

possible – such as "say" or "think", which makes the 'bridge' baseline of the previous accounts unclear. That is, the notion of 'bridge' is not defined in terms of the meaning of the verb, and thus immediately calls into question a potential meaning basis for an observed difference.

(4) a. Bridge verb
What did John **say** that Mary bought?
b. Factive verb
??What did John **know** that Mary bought?
c. Manner-of-speaking verb
??What did John **whisper** that Mary bought?

Previous and Current Theories:

Syntactic Accounts: In order to explain the difference between extraction across bridge verbs (4a) on the one hand and extraction across factive and manner verbs (4b/c) on the other, a syntactic account requires different syntactic structures for bridge verbs compared to the other two kinds of verbs. For instance, it has been claimed that bridge verbs take embedded clauses as arguments, while embedded clauses of manner-of-speaking verbs and factive verbs contain extra covert structures at an abstract level ('deep structure' of the Chomskyan framework), such as an invisible complex NP (Kiparsky & Kiparsky, 1971; Snyder, 1992; Stowell, 1981; Stoica, 2016). In this way, the unacceptability of extraction across factive and manner-of-speaking verbs could be captured by syntactic constraints of extraction such as *Subjacency*. However, there are no independent reasons to propose these covert complex structures.

Discourse Accounts: The fundamental idea of discourse accounts is that grammatical constructions specify certain parts of a sentence as 'foreground' or 'background', and the gap in a filler-gap construction can't fall within a backgrounded domain. In this spirit, Ambridge & Goldberg (2008) (henceforth A&G) proposed an account they call Backgrounded Constituents are Islands (BCI), arguing that extraction from an S-complement is unacceptable in proportion to its 'backgroundedness'. The more backgrounded the embedded clauses, the less acceptable the extraction.

Frequency Accounts: Frequency accounts link extraction difficulties to low exposure: less frequent or unpredictable extractions are more difficult to process (cf. Hale, 2001; Levy, 2008). Dabrowska (2008) proposed that speakers store prototypical templates corresponding to frequent combinations they have encountered in their experience such as 'Wh-word *do you think/say* S-complement?'. Filler-gap constructions that are more similar to the prototypical constructions are more acceptable.

We will propose a different generalization of Dabrowska's account, following the results of Exp 1 (presented below):

***The verb-frame frequency hypothesis*:** The acceptability of an utterance is best captured by 2 independent, separate effects: (i) the frequency or the type of the construction (e.g., wh-questions vs. declaratives) and (ii) the frequency of the verb head-structure.

Extractions are rated less acceptable than declaratives, because extractions are less common compared to declaratives in communication, or they require more cognitive resources. As for acceptability variance within declaratives or extractions out of S-complements, the major determining factor is the frequency of the matrix verbs taking S-complements (P (matrix verb, S-complement)). This account does not predict an interaction ('island') effect between the acceptability of declaratives and extractions. (The interaction obtained in previous works may be a result of applying linear models to non-linear acceptability.)

Following this new *verb-frame frequency hypothesis,* manner-of-speaking wh-questions such as (4c) are less natural, because the manner verbs rarely take S-complements. Factive verbs that take S-complements with a similar frequency to bridge verbs should form equally good wh-questions. A major outlier to our account, the verb '*know*', may be explained by pragmatic factors.

Predictions of The Three Theories on Factive and Manner-of-speaking Islands:

Prediction of the Syntactic Accounts: All factive and manner-of-speaking wh-questions should be less acceptable than all the bridge ones due to categorically distinct covert structures (e.g., Kiparsky & Kiparsky, 1971; Snyder, 1992), as in Fig.1a.

Prediction of the BCI Account (A&G 2008): There should be a correlation between the acceptability of wh-questions and the backgroundedness of the S-complements taken by the verbs, as shown in Fig.1b. Factive verbs take presuppositions, the most backgrounded constituents, and therefore should form the most unnatural wh-questions. Manner-of-speaking verbs should form less strong islands, while bridge verbs form fully acceptable wh-questions.

Prediction of Verb-frame Frequency Hypothesis: The acceptability of extraction out of SC verbs should depend primarily on the frequency of those verbs taking S-complements, and the effect of frequency should be similar on both wh-questions and declaratives (no 'island' effect), as plotted in Fig.1c.

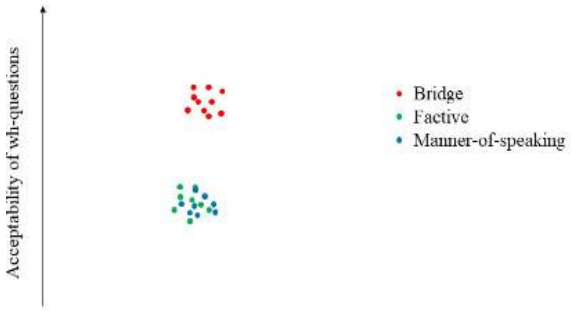


Figure 1a: The prediction of the syntactic accounts

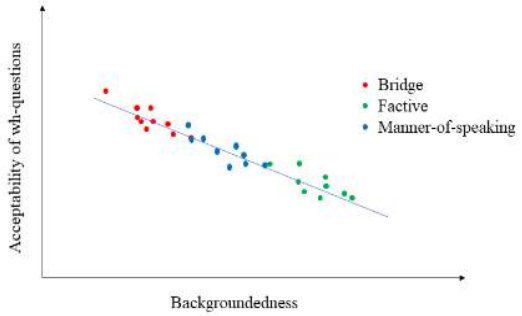


Figure 1b: The prediction of the BCI account.

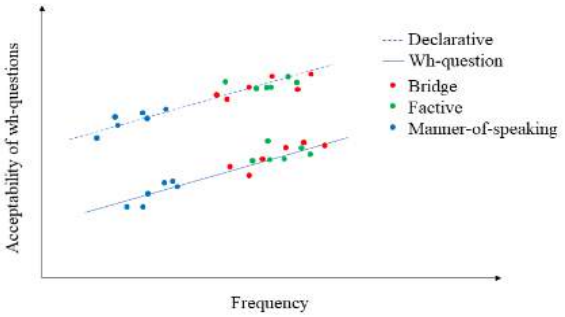


Figure 1c. The prediction of our frequency account.

Experiment 1: Replication of Ambridge and Goldberg (2008)

Experiment 1 is an attempted replication and extension of A&G (2008) using an expanded set of 24 verbs in the 3 categories (A&G tested 12 verbs). There were 2 sub-experiments. Experiment 1a: acceptability judgements of wh-questions formed by the 3 groups of verbs and their corresponding declarative controls. Experiment 1b: negation test to measure the backgroundedness of S-complements of those verbs where extraction appeared. The BCI account predicts a negative correlation between the backgroundedness of the extraction domain and the acceptability of the wh-questions (A&G, 2008).

This experiment also tests the syntactic theories via collecting acceptability ratings of wh-questions formed by the 3 groups of verbs.

Methods

Participants: 180 subjects participated in this experiment via Amazon Mechanical Turk in exchange for $2 each: 120 participants answered acceptability questions for wh-questions and declarative clauses. Another 60 subjects completed the negation task.

In all the experiments reported here, data from participants who did not self-report themselves as native speakers of American English or didn't answer all the comprehension questions with at least 85% accuracy were excluded. Responses from 116 participants in the acceptability task and 49 participants in the negation task were analyzed.

Design and Materials: The acceptability and negation tasks were constructed for 24 sentence complement (SC) verbs of the 3 categories, as listed in (6)[2].

(6) a. Bridge verbs: **say, decide, think, believe**, feel, hope, claim, report, declare

b. Factive verbs: **know, realize, remember, notice,** discover, forget

c. Manner-of-speaking verbs: **whisper, stammer, mumble, mutter,** shout, yell, scream, murmur, whine

In the acceptability task, wh-questions and their corresponding declarative sentences were designed as (7a) and (7b) respectively. 96 pairs of wh-questions and declaratives were constructed, and each of the 24 tested verbs in (6) formed 4 pairs. In each pair of wh-question and declarative control, NP1 and NP2 are common names, V1 comes from (6), and V2 is the past tense form of one of the frequently used 25 verbs (*like, eat, buy, build, cook, destroy, dislike, drink, draw, fix, find, know, learn, lose, make, mention, need, see, sell, steal, take, teach, throw, want, write*). To reduce the possibility of semantic plausibility confounds, we used 'something' instead of a specific NP as the embedded object.

(7) a. What did [NP1] [V1] [[that] [NP2] [V2]]?
e.g., What did Susan know that Anthony liked?
b. [NP1] [V1] [that] [[NP2] [V2+something]]
e.g., Susan knew that Anthony liked something

The 96 pairs were split across 2 lists: each list contained 2 declaratives and 2 different wh-questions per verb. Each participant saw 96 sentences (from 1 list) in a random order. They were asked to rate how natural each sentence was with a rating scale from 1 (extremely unnatural) to 5 (extremely natural). Each sentence was followed by a comprehension question about the content of the preceding sentence to check if participants were paying attention to the task.

In the negation-test task (from A&G, 2008), each trial included a negated complex sentence (8a) and a negated

[2] Verbs in bold are those tested in A&G (2008). The labeling of a verb as 'bridge' was obtained from previous literature, such as Erteschik-Shir (1973), Snyder (1992), A&G (2008).

simple sentence (8b) which is the negated version of the S-complement in (8a).

(8) a. [NP1] didn't [V1] [that] [NP2] [V2+Appropriate NP]
e.g., Susan didn't know that Anthony liked the cake.
b. [NP2] didn't [V2+Appropriate NP]
e.g., Anthony didn't like the cake.

Participants were asked to rate how true they thought the second sentence was, given the first sentence, with a scale from 1 (false) to 5 (true). A&G proposed that these negation scores reflect how "backgrounded" the information in the S-complement is.

Results and Discussion:

A&G (2008) calculated the difference scores between the ratings of wh-questions and declarative clauses as the measurement for acceptability of those wh-questions, and they found a strong correlation between these difference scores and negation scores ($r=-0.83$, $p<0.001$; see Fig.2a). We applied the same analysis to our data. The obtained correlation in our data was in the same direction as in A&G (2008) but the effect was smaller both in the 12 verbs they tested ($r=-0.39$, $p=0.2$) and in the full set of 24 verbs ($r=-0.31$, $p=0.13$; see Fig.2b). Further, we found overlap between acceptabilities for factive and bridge verbs, contradicting the syntactic accounts, which predict non-overlapping acceptability between factive and bridge wh-questions given their distinct covert deep structures.

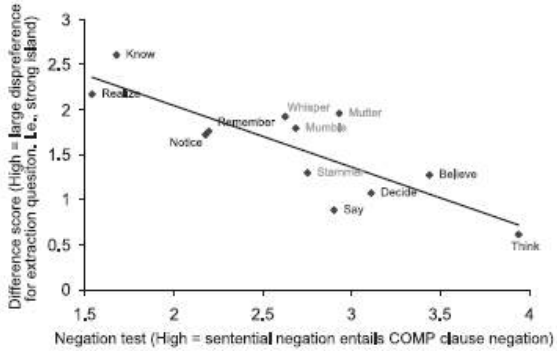


Figure 2a: A&G (2008) - correlation between mean difference scores and mean negation test scores by verb (12 verbs)

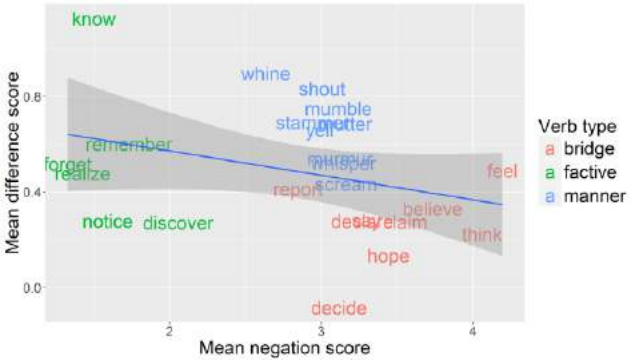


Figure 2b: Our study - correlation between mean difference scores and mean negation test scores by verb (24 verbs).

In a post-hoc analysis, we collected the frequency of the 24 verbs followed by the complementizer 'that' from the Google books corpus as a proxy for relative verb frame frequency. Acceptability ratings for wh-question forms were significantly correlated with verb frame frequency ($rho=0.72$, $p<0.001$), as plotted in Fig.3, as were the corresponding declaratives ($rho=0.76$, $p<0.001$). Furthermore, 74.6% of ratings were between 4/5 and 5/5 for both the wh-questions and declaratives of verbs, suggesting that participants were not using the full range of the scale.

Thus, we propose *the verb-frame frequency hypothesis*: the acceptability of an utterance is best captured by 2 independent, separate effects: (i) the frequency of the type of construction (e.g., wh-questions vs. declaratives) and (ii) the frequency of the verb head-structure- the frequency of the matrix verbs taking S-complements P(matrix verb, S-complement). This hypothesis suggests the impact of verb frame frequency on filler-gap constructions should be similar to that on declaratives.

An outlier to this account is the verb 'know'. We hypothesize that the idiosyncratic behavior of 'know' was due to pragmatic factors in the wh-question: a question is a request for knowledge but a question with 'know' implies that the speaker already has the knowledge. We hypothesize that 'know' might not be an outlier in other extraction constructions whose meaning does include implicit knowledge of the interlocutor, such as clefts, which is tested in Exp3.

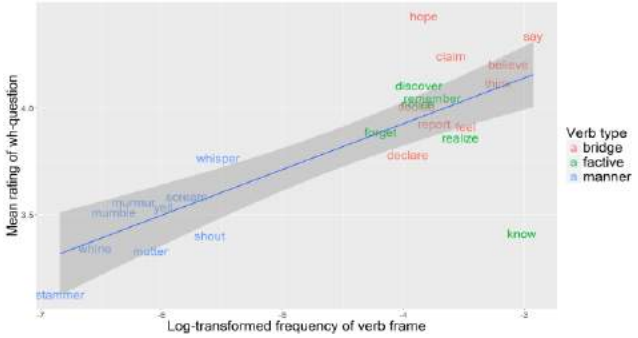


Figure 3: Correlation between mean ratings of wh-questions and log-transformed frequencies by verb ($rho=0.72$, $p<0.001$).

Experiment 2: Wh-questions with 48 Verbs

The goal of this experiment was to test the frequency account with more matrix verbs beyond the 3 categories (bridge, factive, manner-of-speaking). The *verb-frame frequency hypothesis* predicts that the verbs that frequently take S-complements would be more acceptable as wh-questions and declaratives, regardless of the verb category. The syntactic accounts make no predictions for verbs outside the 3 categories.

Given that the 5-point scale does not seem to be appropriate for measuring the acceptability of these sentences, we performed a forced-choice binary acceptability judgment task in this experiment and applied mixed-effects logistic regression to the data.

Methods:

Participants:120 people participated via MTurk, and each was paid $2. Responses from 110 participants were included in the analysis.

Design and Materials: The design was similar to Experiment 1a, with 48 verbs. The verbs included 8 for each of the 3 categories and another 24 outside the 3 categories, as listed in (9). The 24 'other' type verbs were not clearly categorized in the previous literature.

(9) **Matrix verbs**:

Bridge (8): feel, say, believe, hope, think, report, declare, claim,

Factive (8): know, remember, realize, notice, discover, forget, learn, hate

Manner (8): whisper, mumble, murmur, mutter, whine, shout, yell, scream

Other (24): hear, recall, blab, conjecture, conceal, proclaim, hint, remark, infer, confirm, deny, guess, confide, maintain, testify, reveal, suspect, verify, prove, insist, guarantee, presume, hypothesize, complain

Wh-questions and declaratives were constructed for the 48 matrix verbs with 6 items for each verb (288 items in total). The design of the items is the same as Experiment 1a.

As in Expt 1, participants were assigned to 1 of 2 lists made up of 3 declaratives and 3 wh-questions for each of the 48 verbs. Each participant saw 288 sentences in a random order. Participants were asked to rate each sentence using a binary scale (acceptable vs. unacceptable) based on how natural they thought the sentence was. Each sentence was also followed by a comprehension question.

Results and Discussion:

Acceptability judgments were analyzed with a mixed-effects logistic regression using the *lme4* package in *R. Sentence type (declarative vs. wh-question)*, *log-transformed frequency of the verb frame* and their interaction were entered as predictors. The model was fit with the maximum random effect structure which contained random by-*subject* and by-*verb* intercepts as well as slopes for *sentence type*frequency* by-subjects and slopes for *sentence type* by-verb.

The results were in line with the *verb-frame frequency hypothesis*. Wh-questions and declaratives formed by verbs that frequently take S-complements were significantly more acceptable (β=0.58, z=3.98, p<0.001). There was also a significant main effect of sentence type: declaratives were rated more acceptable than wh-questions (β=-3.27, z=-2.924, p<0.004). No interaction was found (p>0.4), suggesting no island effect was present. The log-odds of an 'acceptable' response for a given verb frame frequency are plotted in Fig.4a. Note that an island theory would predict the effect of frequency would have a steeper slope for wh-questions than declaratives, but Fig.4a shows the opposite (non-significantly). A pattern resembling a spurious interaction ('island' effect) shows up when log-odds are converted into probabilities of acceptance, as shown in Fig.4b.

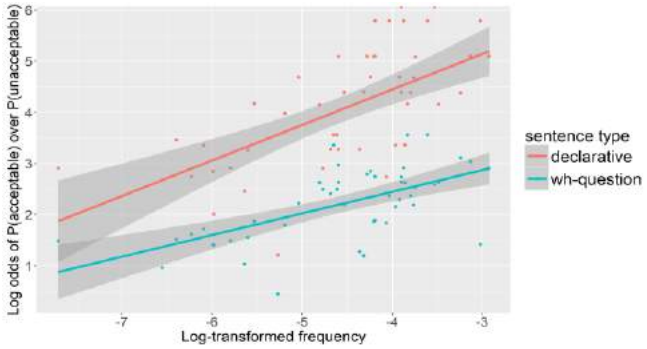


Figure 4a: Log-odds of 'acceptable' response for wh-questions and declarative clauses against log-transformed frequencies by verb (48 verbs).

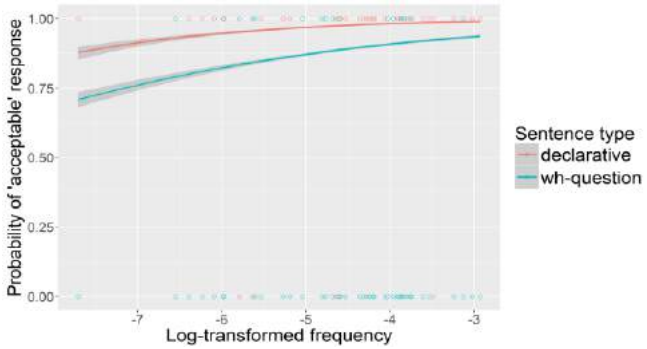


Figure 4b: Probability of 'acceptable' response for wh-questions and declaratives against log-transformed frequencies by verb (48 verbs)

Experiment 3: Cleft Structure

Experiment 3 aims to further test the *verb-frame frequency hypothesis* and check if 'know' is always idiosyncratic in filler-gap constructions with respect to the frequency account. The syntax-based theories claimed that extractions obey the same set of constraints regardless of construction, which indicates extraction difficulties across different verbs should be the same across constructions (e.g., in wh-

questions and cleft structures). However, an alternative is that the unusual behavior of 'know' in Experiment 1 might be related to the 'information-obtaining' function of wh-questions. If so, then 'know' should not be an outlier in cleft structures, because cleft structures are modifications of an NP and not associated with 'knowing'. We thus propose that, beyond verb frame frequency, extraction difficulties may differ depending on the meaning and function of the specific construction (Abeillé et al., 2018).

Methods:

Participants: Data from 120 participants were collected via MTurk, and each was paid $2. Responses from 104 participants were analyzed.

Design and Materials: Cleft structures and their corresponding declarative sentences were designed as in (10a) and (10b) respectively. 96 pairs of clefts and declaratives were constructed, and each of the 24 tested verbs in (6) formed 4 pairs as in Exp 1a. Participants were asked to rate each sentence with a binary rating scale. Each sentence was followed by a comprehension question.

(10) a. It was the pie that Angela mumbled that Kevin liked
b. Angela mumbled that Kevin liked the pie.

Results and Discussion:

Acceptability responses were analyzed in the same way as in Exp 2. Sentences with higher frequency verb frames were significantly more acceptable (β=1.2, z=2.7, p<0.01) and cleft structures were less likely to be acceptable (β=-14.6, z= 2.5, p<0.011). The interaction of sentence type and frequency was not significant (β=-0.87, z=-0.9, p=0.34), thus providing no evidence of an island effect (Fig. 5). These data are best explained by positing that verb frame frequency and extraction have independent, additive effects in log-odds space, as predicted by the *verb-frame frequency hypothesis*.

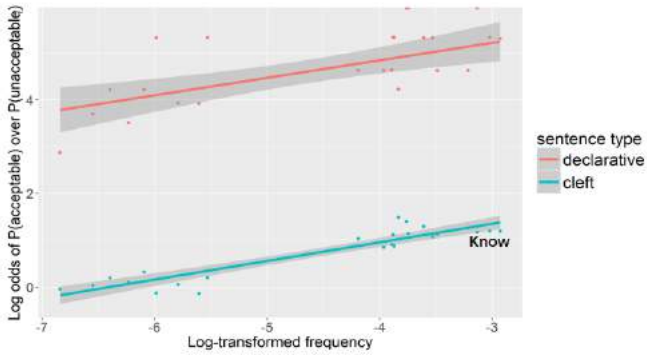


Figure 5: Log-odds of 'acceptable' response for clefts and declaratives against log-transformed frequencies (24 verbs)

As predicted by the meaning-based approach to long-distance dependency acceptability, 'know' is not an outlier for the frequency account in the cleft structure. The idiosyncratic behavior of 'know' seems to have been due to pragmatic factors in the wh-question: a question is a request for knowledge but a question with 'know' implies that the speaker already has the knowledge. If the long-distance dependency structure does not involve the meaning of '*know*' (as in clefts), then extraction out of S-complements of '*know*' is acceptable. Such distinct behaviors of 'know' in wh-questions and cleft structures suggest extractions vary across constructions, due to their meaning differences.

General Discussion

The results of all three experiments show that the amount of exposure is a key determining factor for the acceptability of filler-gap constructions formed by various SC verbs, including factive and manner-of-speaking verbs. The apparent interaction ('island' effect) may be a false positive caused by the use of linear models with ordinal acceptability ratings.

Interestingly, we also found that island constraints are not the same across constructions. Though different extractions may share similar cognitive processes, variation across constructions does exist and can be attributed to different meanings or functions associated with those different types of extractions. Though we didn't find strong evidence for the discourse-based accounts in the phenomena investigated here, frequency and discourse accounts are not necessarily mutually exclusive in capturing filler-gap constructions (and other phenomena) in general (Abeillé et al., 2018).

Our results suggest that (un)acceptable filler-gap constructions could potentially be learnable via exposure. Although direct negative evidence is missing especially for such complex structures, it is likely that children could use indirect negative evidence to acquire long-distance dependencies. Children may draw statistical inference from the input and regard the absence of a type of extraction in the input as evidence of its unacceptability or ungrammaticality.

Reference

Abeillé A., B. Hemforth, E. Winckel, E. Gibson. (2018). A construction-conflict explanation of the subject-island constraint, CUNY Conference, UC Davis

Ambridge, B., & Goldberg, A. (2008). The island status of clausal complements: Evidence in favor of an information structure explanation. *Cognitive Linguistics*, *19*, 349–381

Chomsky, N. (1973). Conditions on transformations. In S. Anderson, & P. Kiparsky (Eds.), *A Festschrift for Morris Halle*. New York: Holt, Rinehart, & Winston.

Dabrowska, E. (2008). Questions with long-distance dependencies: A usage-based perspective. *Cognitive Linguistics*, *19*(3), 391–425.

Erteschik-Shir, N. (1973). *On the nature of island constraints.* Ph.D. dissertation, MIT, Cambridge, MA.

Hale, J. (2003). *Grammar, Uncertainty and Sentence Processing*. Ph.D. dissertation, John Hopkins University, Baltimore.

Kiparsky, P. & Kiparsky, C. (1971). Fact. In M. Bierwisch & K. Heidolph (Eds.), Progress in Linguistics. The Hague: Mouton

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*,106(3), 1126–1177.

Ross, J. R. 1967. *Constraints on variables in syntax.* Ph.D. dissertation, MIT, Cambridge, MA.

Snyder, W. (1992). *Wh-extraction and the lexical representation of verbs*. Unpublished manuscript, MIT, Cambridge, MA.

Zwicky, A. M. (1971). In a manner of speaking. *Linguistic Inquiry*, *11*, 223-233.

Unflinching Predictions: Anticipatory Crossmodal Interactions are Unaffected by the Current Hand Posture

Johannes Lohmann (johannes.lohmann@uni-tuebingen.de)
Chair of Cognitive Modeling, Sand 14
Tübingen, 72076 Germany

Martin V. Butz (martin.butz@uni-tuebingen.de)
Chair of Cognitive Modeling, Sand 14
Tübingen, 72076 Germany

Abstract

According to theories of anticipatory behavior control, action planning and control is realized by activating desired goal states. From an event-predictive perspective, this activation should focus sensorimotor processing on expected, upcoming event boundaries. Previous studies have shown that peripersonal hand space (PPHS) is remapped to the future hand location in a grasping task before the movement commences. Here, we investigated if the current hand posture interferes with the anticipatory remapping of PPHS. Participants had to grasp virtual bottles from two differently oriented starting postures. During the prehension, they received a vibrotactile stimulus on their right index finger or on their thumb, while a visual stimulus appeared at the bottle, either matching the future finger position, or not. Participants had to name the stimulated finger. While the hand posture affected verbal response times, the anticipatory remapping remained unchanged. Apparently, the predictive processes that realize the anticipatory remapping, generalize over initial hand postures.
Keywords: Event Predictive Cognition; Anticipatory Behavioral Control; Peripersonal Space; Virtual Reality

Introduction

According to theories of anticipatory behavior control, the initiation of goal-directed actions requires the activation of event-predictive structures or *schemata* (EPSs; e.g. Butz, 2016; Butz & Kutter, 2017; Hommel, Müsseler, Aschersleben, & Prinz, 2001; Hoffmann, 2003; Zacks, Speer, Swallow, Braver, & Reynolds, 2007; Richmond & Zacks, 2017). These EPSs are considered to encode the final outcome of an action, but also the sensorimotor changes that usually unfold during an action, as well as the situational pre-conditions of successful action execution. In relation to free-energy based theories of cognition (Friston, 2009), EPSs are assumed to be involved in the more general active inference process that realizes action planning, decision making, and control (Butz, 2016). This perspective is closely related to the *ideomotor principle* (Greenwald, 1970) from cognitive psychology and essentially states that anticipated final outcomes and sensorimotor dynamics are activated before actual goal-directed motion takes place.

Empirical evidence for the assumed active inference process comes from eye-tracking studies, showing that the fixation pattern on a grasping target depends on the interaction goal (Belardinelli, Stepper, & Butz, 2016). Apparently, visual processing was tuned to those spatial locations which were critical for a successful object interaction. Considering the multisensory information, which is expected to be represented in EPSs, predictive processing should not be limited to eye-movements, but should also involve other action relevant representations. One example fur such representations are spatial body representations, like the peripersonal hand space (PPHS). PPHS seems crucial for successful object interactions and tool-use (Graziano & Cooke, 2006) . PPHS has also been found to be highly flexible in adapting to interaction possibilities (Holmes, 2012). Furthermore, PPHS enforces multisensory processing (Holmes & Spence, 2004; Bernasconi et al., 2018). According to the outlined theory, one would expect that PPHS is involved in predictive processing and might be remapped towards the grasping target during action planning. If this is the case, typical PPHS-related effects should be observed at the grasping target before the actual hand arrives. One typical indicator of PPHS is the selective interaction between vision and touch, which can be assessed by means of the crossmodal congruency paradigm (Spence, Pavani, Maravita, & Holmes, 2004).

In crossmodal congruency tasks, participants have to indicate the position of a tactile stimulation. Task-irrelevant visual stimuli occurring close to the stimulated body part can interfere with tactile perception. For instance, participants are slower to identify whether thumb or index finger received a tactile stimulation, if a LED is flashed at the non-stimulated finger (incongruent), whereas a flash at the location of the stimulated finger prompts a faster response (congruent). Previous studies indeed showed that interference between vision and touch can occur in object interaction tasks at the target object location even before movement initiation (Brozzoli, Pavani, Urquizar, Cardinali, & Farnè, 2009; Brozzoli, Cardinali, Pavani, & Farnè, 2010). This implies an anticipatory crossmodal congruency effect (aCCE), which can be used to investigate the anticipatory remapping of PPHS. In more recent studies (Belardinelli, Lohmann, Farnè, & Butz, 2018; Lohmann, Belardinelli, & Butz, 2019; Patané et al., 2018), it was shown that the aCCE can be observed on a trialwise basis without explicit instruction of a certain grasping type. These results imply that the aCCE indeed reflects an adaptive remapping due to action planning instead of a general shift in spatial attention. Apparently, PPHS is involved in the guidance of goal-directed actions, by providing a mapping between the space that can be interacted with and the according actions (Bufacchi & Iannetti, 2018).

While these results imply that PPHS is engaged in predic-

tive processing, some aspects of this mechanism remain illusive. For instance in the studies of Belardinelli et al. (2018) and Lohmann et al. (2019), the orientation of the final grasp modulated the strength of the aCCE. In case of underhand grasps, the aCCE was smaller compared to overhand grasps. This might be due to the fact that underhand grasps are less frequent in object interactions, rendering the planning more difficult. However, since the initial hand posture in these experiments was closer to the overhand grasp, this effect could also imply that the assumed prediction process does not completely generalize over the initial hand posture. This would dovetail with previous results from research on motor imagery, especially on mental rotation, which showed a strong interaction between ongoing motor planning and the actual posture (Parsons, 1987; Qu, Wang, Zhong, & Ye, 2018). Hence, our main aim in the present study was to investigate whether the aCCE depends on a postural match between initial and future hand position. If the aCCE would be affected by variations in the initial posture, this would imply that the sensorimotor changes assumed to be encoded in EPSs are less general than expected. If not, this would corroborate further evidence for the assumption that the aCCE is indeed an indicator for a general movement planning mechanism.

We conducted a behavioral study to discern these alternatives. Participants performed a grasp-and-carry task in VR, interacting with a virtual bottle. At different times before and during the interaction, participants received a tactile stimulation at the thumb or index finger. Concurrently, a visual stimulus appeared at the left or right side of the bottle, either matching the future location of the stimulated finger, or not. Participants had to respond as fast as possible, by verbally naming the finger that was stimulated. A typical aCCE would be reflected by faster responses if the visual stimulus matched the future finger position. The starting position of the hand varied from trial to trial, participants had either to start from a more clockwise, or more counterclockwise rotated starting posture. The main question was whether aCCEs would be modulated by this trialwise variation of the hand orientation.

Method

Participants

Twenty-four students from the University of Tübingen participated in the experiment (ten females). Their age ranged from 19 to 26 years (M = 21.2, SD = 1.9). All but one participant were right-handed and all participants had normal or corrected-to-normal vision. Participants provided informed consent and received either course credit or a monetary compensation for their participation. Two participants had difficulties with the virtual grasping procedure and could not complete the experiment. The respective data were not considered in the analysis.

Apparatus

Participants were equipped with an Oculus Rift© DK2 stereoscopic head-mounted display (Oculus VR LLC, Menlo Park, California). Motion tracking of hand movements was realized with a Leap Motion© near-infrared sensor (Leap Motion Inc, San Francisco, California, SDK version 3.2.1). The Leap Motion© sensor provides positional information regarding the palm, wrist, and phalanges. This data can be used to render a hand model in VR. Participants responded verbally to the tactile stimulation. In order to so, participants were equipped with a headset. Speech recognition was implemented by means of the Microsoft Speech API 5.4. The whole experiment was implemented with the Unity® engine 2017.4.5f1 using the C# interface provided by the API. During the experiment, the scene was rendered in parallel on the Oculus Rift and a computer screen, such that the experimenter could observe and assist the participants.

Tactile stimulation was realized by means of two small (10 mm $\times$ 3.4 mm) shaftless vibration motors attached to the tip of the thumb and the index finger of the participants. The motors were controlled via an Arduino Uno microcontroller (Arduino S.R.L., Scarmagno, Italy) running custom C software. The microcontroller was connected to the computer via an USB port, which could be accessed by the Unity® program. The wiring diagram as well as additional information regarding the components can be found at the first author's webpage. [1]

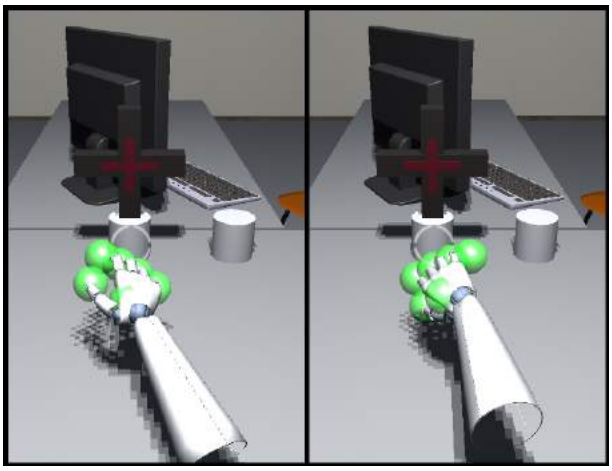


Figure 1: The VR scene with the clockwise oriented (left panel) and the counterclockwise oriented (right panel) starting postures. Participants had to grasp a bottle which appeared on the central pedestal and place it upright onto the right pedestal. The bottle could be either upright, or rotated.

Virtual Reality Setup

The VR setup put participants in an office-like room. Centered about 50 cm in front of them, a pedestal was placed, where, during the trials, the target object appeared. The target was always a 3D model of a plastic bottle either oriented upright or upside down. The bottle was 15 cm in height, sub-

[1] https://uni-tuebingen.de/de/26084

tending a visual angle of 17.1° at the initial location. A second pedestal, 15 cm to the right of the first one, served as the target location (see Fig. 1). The positions of the pedestals were marked with actual cardboard boxes providing haptic feedback regarding the bounds of the task space (participants were seated in a way that they had to stretch their arm to reach the pedestals). Instructions and feedback were presented in different text-fields, aligned at eye-height. At the beginning of a trial, a fixation cross appeared at the initial location of the target bottle (see Fig. 1). The fixation cross was 10 cm wide and 10 cm high, subtending a visual angle of 11.4°. The visual distractor was realized by means of a red, spherical flash with a diameter of 8 cm (equal to a visual angle of 8°) appearing at the left or right side of the bottle.

Procedure

At the beginning of the experiment, participants received a verbal instruction regarding the VR equipment. Then they were equipped with vibration motors and familiarized with the tactile stimulation. Participants were then seated comfortably on an arm chair and put on the HMD. Before the actual experiment, participants performed a grasping training and trained the verbal response until they felt comfortable with both tasks. In the grasp training, participants performed the grasp-and-carry task without receiving a tactile stimulation. Furthermore, participants could familiarize themselves with the two different starting positions. In the verbal response training, participants did not perform a grasping movement, but remained with their hand in the starting position.

The actual experiment combined both tasks in a dual-task paradigm. At the beginning of each trial, participants had to move their right hand into a designated starting position, consisting of red, transparent spheres indicating the required positions of the fingers and the palm. There were two possible variations of the starting position. One was tilted by 15° clockwise in the frontal plane, and one was tilted by 15° counterclockwise in the frontal plane. Accordingly, this required participants' to rotate their hands either clockwise or counterclockwise. The spheres turned green when the respective fingers were in position (see Fig. 1). Furthermore, participants had to maintain a stable looking direction on a fixation cross. Once both requirements were met for 1000 ms, the fixation cross as well as the visible markers of the initial position disappeared and a bottle appeared on the central pedestal. The bottle was either oriented upright, or upside down. Participants were instructed to grasp the bottle with a power grasp, and put it in an upright orientation within the target location. We did not explicitly instruct a underhand grasp in case of upside down bottles, however, all participants performed this kind of grasp. The initial hand postures were close to the respective grasping hand posture for the upright oriented bottle (clockwise hand posture), or the upside down bottle (counterclockwise hand posture).

Besides the grasp-and-carry task, participants had to discriminate which finger received a vibrotactile stimulation and to report the stimulated finger as fast as possible (by saying "index or "thumb, i.e., in German "Zeigefinger or "Daumen) upon vibration detection. The onset of the tactile stimulation varied from trial to trial. A visual distractor appeared at the same time at either the right or the left side of the bottle. Depending on the bottle orientation, this was expected to yield different congruent and incongruent conditions with respect to the aCCE (see Fig. 2).

The experiment consisted of 480 trials, presented in a single block. The experiment was self-paced and participants could pause between trials. The whole procedure took between 90 and 120 minutes, including preparation and training.

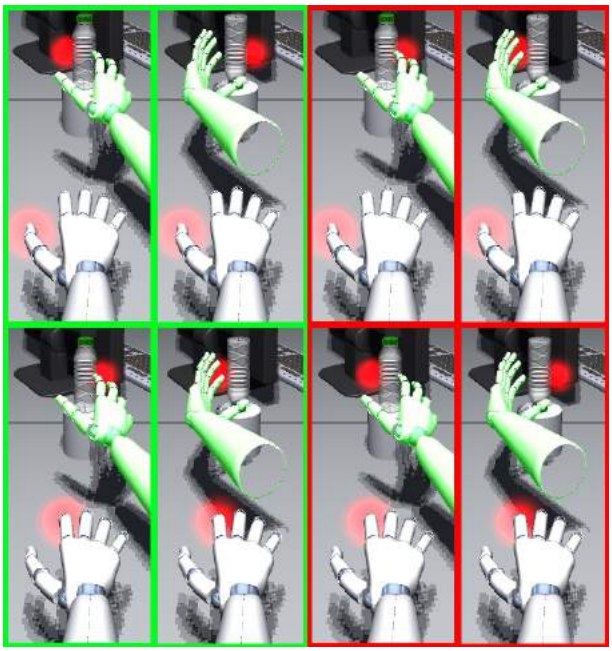


Figure 2: The different congruency conditions with respect to the future hand position (transparent green hand), depending on bottle orientation. The stimulated finger is indicated by a red flash, this was done for the sake of visibility, the participants received no visual cue regarding the tactile stimulation. Red frames indicate incongruent conditions, congruent conditions are marked with a green frame. Please note that the initial hand posture was different from the one shown in this image (cf. Fig 1), the flat hand posture here was used for the sake of visibility.

Factors, Measures, Data Treatment

We varied five factors across trials. First, the target bottle could be oriented upright or upside down (*orientation*). Second, the visual distractor could appear either on the left or the right side of the bottle (*distractor*). Third, the tactile stimulation could be applied either to the thumb or to the index finger (*stimulation*). Fourth, we varied the initial hand posture - clockwise or counterclockwise - which participants had to maintain to start the trial (*posture*). Fifth, we varied the on-

set of the tactile stimulation and the visual distractor (*SOA*): 250 ms after presentation of the bottle (SOA1), at movement onset (SOA2), or after the hand traveled half-way to the bottle (SOA3). We repeated the 2 (distractor) × 2 (stimulation) × 2 (orientation) × 2 (posture) × 3 (SOA) factor combinations ten times, yielding 480 trials. The primary dependent measure were the verbal response times for naming the stimulated finger. Data from error trials (wrong or no verbal response, 1.8% of the trials) were excluded from the response time analyses. Furthermore, we analyzed the error data using a mixed effects logistic regression.

Congruency

For our hypothesis, possible aCCE's were most relevant. aCCE's are reflected by three-way interactions between the factors orientation, distractor, and stimulation (cf. Fig. 2) . For instance, in the case of an upright bottle a tactile stimulation of the index finger along with a visual distractor on the right side of the bottle is congruent. To focus the analysis, we recoded the data accordingly and obtained a congruency factor, combining the visual distractor and tactile stimulus factor. For the response times, we report an analysis of the respective differences (incongruent - congruent) with a 2 (orientation) × 2 (posture) × 3 (SOA) ANOVA. In this analysis a significant, positive intercept would indicate a significant aCCE (faster responses in congruent as opposed to incongruent conditions).

Results

Verbal response times from the 22 considered participants were analyzed with a 2 (congruency) × 2 (orientation) × 2 (posture) × 3 (SOA) repeated measures ANOVA. Verbal response times differences between incongruent and congruent conditions were further analyzed with a 2 (orientation) × 2 (posture) × 3 (SOA) repeated measures ANOVA. Only correct trials were included in the RT analysis. All reported post-hoc comparisons were submitted to a Holm-Bonferroni correction. The analyses were carried out with R (R Core Team, 2016) and the `ez` package (Lawrence, 2015). In case of violations of the assumption of sphericity, p-values were submitted to a Greenhouse-Geisser adjustment. Error rates were analyzed with mixed effects logistic regression, using the `lme4` package (Bates, Mächler, Bolker, & Walker, 2015).

Verbal Response Times

The 2 (congruency) × 2 (orientation) × 2 (posture) × 3 (SOA) repeated measures ANOVA yielded significant main effects for orientation ($F(1,21) = 7.63$, $p = .012$, $\eta_p^2 = .27$), congruency ($F(1,21) = 32.57$, $p < .001$, $\eta_p^2 = .61$), and SOA ($F(1,21) = 28.04$, $p < .001$, $\eta_p^2 = .57$), as well as significant interactions between orientation and SOA ($F(2,42) = 8.74$, $p = .001$, $\eta_p^2 = .29$), orientation and posture ($F(1,21) = 5.54$, $p = .028$, $\eta_p^2 = .21$), orientation and congruency ($F(1,21) = 9.55$, $p = .006$, $\eta_p^2 = .31$), SOA and congruency ($F(2,42) = 10.39$, $p = .001$, $\eta_p^2 = .33$), as well as a three-way interaction for orientation, congruency, and SOA ($F(2,42) = 7.32$, $p = .002$, $\eta_p^2 = .26$; all remaining p's $\geq .168$).

Participants responded faster to upright bottles ($M_{upright}$ = 700 ms vs. $M_{rotated}$ = 713 ms), and in case of congruent stimulation ($M_{congruent}$ = 691 ms vs. $M_{incongruent}$ = 722 ms). Verbal RTs decreased with SOA (M_{SOA1} = 742 ms, M_{SOA2} = 710 ms, M_{SOA3} = 669 ms; all respective p's $<.001$). Regarding the interaction between orientation and SOA, participants responded faster to bottles oriented upright at SOA1 ($t(21) = 3.21$, $p = .016$) and SOA2 ($t(21) = 4.02$, $p = .004$), for SOA3, this difference was no longer significant ($t(21) = -0.54$, $p = .595$). Post-hoc analyses of the orientation × posture interaction showed that participants responded faster to upright than to upside down bottles when starting in a clockwise posture ($t(21) = 3.59$, $p = .010$). The respective difference was not significant for the counterclockwise posture. Furthermore, response times in case of upright bottles and a clockwise posture were significantly faster than response times in the other three conditions (all respective p's $<.04$).

To further analyze the interactions involving the congruency factor, we analyzed the RT differences between incongruent and congruent conditions with a 2 (orientation) × 2 (posture) × 3 (SOA) ANOVA. The analysis yielded a significant intercept ($F(1,21) = 47.00$, $p < .001$, $\eta_p^2 = .69$), significant main effects of SOA ($F(2,42) = 20.43$, $p < .001$, $\eta_p^2 = .49$) and orientation ($F(1,21) = 16.43$, $p = .001$, $\eta_p^2 = .44$), as well as a significant interaction between orientation and SOA ($F(2,42) = 9.20$, $p = .002$, $\eta_p^2 = .30$). No further main effects or interactions reached significance (remaining p's $\geq .230$).

The congruency effect was significantly larger at SOA3 compared to SOA1 and SOA2 (ΔM_{SOA1} = 17 ms, ΔM_{SOA2} = 20 ms, ΔM_{SOA3} = 60 ms; all respective p's $<.001$). For bottles presented upright, the congruency effect was larger than for upside down bottles ($\Delta M_{upright}$ = 54 ms vs. $\Delta M_{rotated}$ = 11 ms). Regarding the interaction between orientation and SOA, after adjusting for multiple comparisons, the only significant difference between upright and upside down bottles was found at SOA3 ($\Delta M_{upright}$ = 98 ms vs. $\Delta M_{rotated}$ = 23 ms; $t(21) = 4.60$, $p < .001$).

To further probe the significance of the aCCE, all of the 2 (orientation) ×2 (posture) × 3 (SOA) mean differences were tested against a true mean of 0. The results are shown in Fig. 3.

Error Rates

Both error and correct trials of all participants, except the trials without response (65 out of 10560 trials) were coded as 0 (error) or 1 (correct) and entered into a mixed effects logistic regression analysis with a binomial distribution. We compared models of increasing complexity with likelihood ratio tests to determine whether the factors orientation, posture, congruency, and SOA were required to account for the error pattern. We kept the error structure simple, applying only a random intercept per participant. After the identification of the null model, we added fixed effects for the ex-

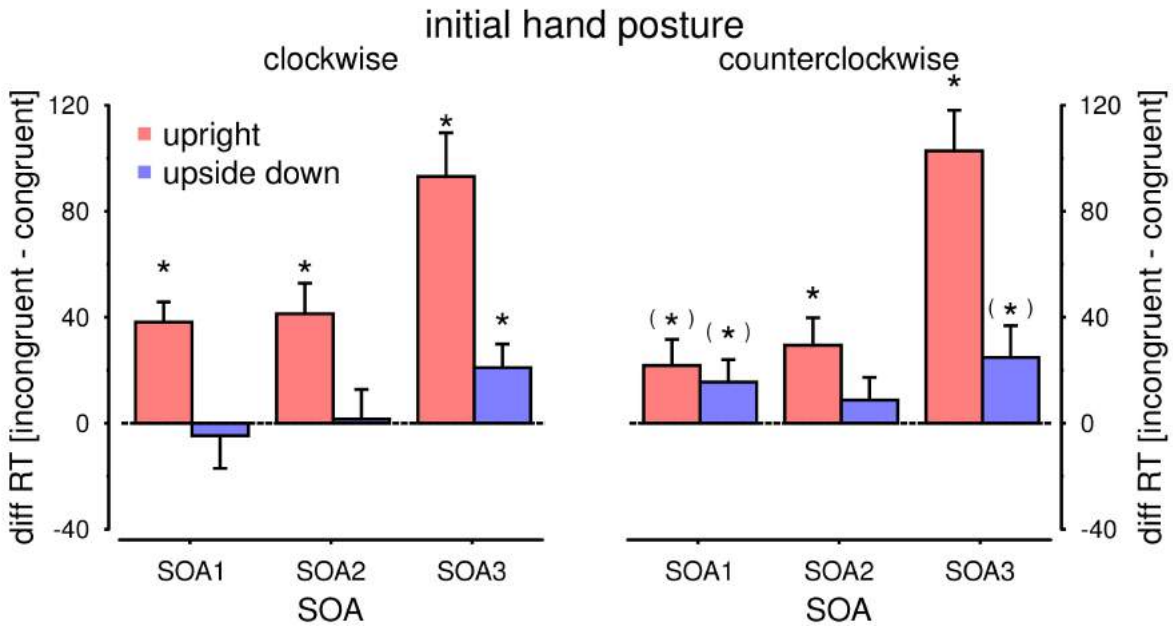


Figure 3: The aCCE – measured as the response time difference between congruent and incongruent trials – including its temporal dynamics and dependency on bottle orientation and initial hand posture. Significant differences from 0 are indicated with an asterisk. Asterisks in brackets indicate comparisons which failed significance after adjusting for multiple comparisons. Error bars indicate the standard error of the mean.

perimental factors to the model as long as the likelihood ratio test between the simpler and the more complex model yielded significant results (with α = .05). We only compared nested models differing with respect to one factor. Models with a single fixed effect were compared with the null model, models with two fixed effects were compared with models with one fixed effect and so on. The best fitting model involved fixed effects for the factors SOA, orientation and congruency, as well as the interaction between congruency and orientation (see Tab. 1, only significant effects are included)[2]. The error risk is increased by a factor of 3.4 in case of later SOAs compared to earlier ones. The error risk decreases by a factor of 0.28 in case of rotated compared to upright bottles. This pattern is further modified by the interaction between congruency and orientation. For upright bottles, the error risk increases in case of incongruent stimulation by a factor of 8.5, for upside down bottles, there is no difference in the error risk for congruent and incongruent stimulation.

Table 1: Effect estimates for the best fitting binomial mixed effects logistic regression model regarding the error rates ($df = 7, logLik = -803.3, BIC = 1671.4$). The logit estimates have been transformed to odds, only significant effects (α = .05) are shown. Z statistics for the Wald test and according p-values are presented in the last two columns.

fixed effect	**odds**	**95% CI**	**Z**	**p**
intercept	0.005	[0.002 , 0.009]	-15.91	$< .001$
SOA3	3.364	[2.295, 4.930]	6.22	$< .001$
orientation	0.288	[0.155, 0.534]	-3.95	$< .001$
orientation $\times$ congruency	8.540	[4.150, 17.574]	5.83	$< .001$

Discussion

We aimed at investigating the mechanism of anticipatory remapping of PPHS in advance of a prehension movement. In order to do so, we investigated anticipatory cross-modal congruency effects (aCCEs) during virtual grasping movements. Participants had to grasp virtual bottles with their right hand, while receiving a tactile stimulation on thumb or index finger of that hand along with a visual stimulation close to one of the future finger positions. The visual distractor could either match the future finger location or not. In line with earlier findings (Belardinelli et al., 2018; Brozzoli et al., 2009, 2010; Lohmann et al., 2019; Patané et al., 2018), we observed dynamic aCCEs, which were more pronounced at later SOAs. To probe whether the strength of the aCCE depends on the match between current and future hand posture, we varied the starting posture of the participants' hands from trial to trial. Participants started either with a clockwise (matching the grasp for an upright bottle), or counterclockwise (matching the grasp for a upside down bottle) posture. While we observed response time differences for the clockwise posture (faster responses for upright bottles, delayed responses for upside down bottles), the congruency effect itself remained unaffected by the initial hand posture. Also with respect to the

[2]Please note that the model assuming the three-way interaction between all factors provided a slightly better fit, however, the respective BIC was much larger than the one of the selected model.

error rates, the initial hand posture yielded no significant influence. A closer inspection of the response time differences for incongruent compared to congruent stimulation implied a small increase in the congruency effect for upside down bottles in case of the counterclockwise posture, while at the same time slightly decreasing the congruency effect for upright bottles. However, these effects seem too small to become significant with the applied sample size. In general, congruency effects were more pronounced for upright compared to upside down bottles (with respect to both RTs and errors). Since this was still the case for the counterclockwise posture, this difference seems not to be due to an initial mismatch between current and future hand position. It rather implies a planning advantage for canonical object orientations.

While the observed interaction between bottle orientation and hand posture dovetails with findings that the current body posture can indeed interfere with mental imagery processes (Parsons, 1987; Qu et al., 2018) and has a significant weighted impact on the actual chose hand grasp posture (Herbort & Butz, 2012), this modulation did not apply to the congruency effect itself. Apparently, the anticipatory control process that gives rise to the aCCE generalizes over the actual hand posture, remapping PPHS towards the future goal, irrespective of the current hand posture. In general, the reported results on the aCCE provide support for theories of probabilistic, event-oriented, active inference (Butz, 2016; Butz & Kutter, 2017): the results confirm that PPHS is adaptively remapped onto future event boundaries during the preparation and for the control of goal-directed behavior.

However, the understanding of the remapping mechanism requires further investigation. In our data, as well as in the results reported by Belardinelli et al. (2018), and Lohmann et al. (2019), the aCCE was much more pronounced for bottles presented upright. It seems that the remapping works more efficient in case of canonical object orientations. As it was pointed out by Bufacchi and Iannetti (2018), measures of PPHS like the aCCE are not only modulated by proximity, but by many other factors like learning, stimulus valence, and environmental characteristics. Hence, a modulation of the aCCE by familiarity seems plausible, but a systematic comparison between bottles and objects with a less pronounced canonical orientation is pending. Moreover, there is still much further light to be shed on the dynamics of this process and its dependency on event-predictive precision estimates. From the event-predictive, anticipatory behavioral control perspective, it can be expected that the future horizon will reach the deeper into the future, the more precise the predictive model estimates are expected to be. That is, the higher our confidence about the upcoming environmental events and sequences thereof, the more we will look ahead and act in a more versatile and flexible goal-directed manner.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi: 10.18637/jss.v067.i01

Belardinelli, A., Lohmann, J. Y., Farnè, A., & Butz, M. V. (2018). Mental space maps into the future. *Cognition*, *176*, 65–73.

Belardinelli, A., Stepper, M. Y., & Butz, M. V. (2016). It's in the eyes: Planning precise manual actions before execution. *Journal of Vision*, *16*(1), 18. doi: 10.1167/16.1.18

Bernasconi, F., Noel, J.-P., Park, H. D., Faivre, N., Seeck, M., Spinelli, L., ... Serino, A. (2018). Audio-tactile and peripersonal space processing around the trunk in human parietal and temporal cortex: an intracranial eeg study. *Cerebral Cortex*, *28*(9), 3385–3397.

Brozzoli, C., Cardinali, L., Pavani, F., & Farnè, A. (2010). Action-specific remapping of peripersonal space. *Neuropsychologia*, *48*(3), 796–802.

Brozzoli, C., Pavani, F., Urquizar, C., Cardinali, L., & Farnè, A. (2009). Grasping actions remap peripersonal space. *Neuroreport*, *20*(10).

Bufacchi, R. J., & Iannetti, G. D. (2018). An action field theory of peripersonal space. *Trends in Cognitive Sciences*, *22*, 1076–1090.

Butz, M. V. (2016). Toward a unified sub-symbolic computational theory of cognition. *Frontiers in Psychology*, *7*, doi: 10.3389/fpsyg.2016.00925.

Butz, M. V., & Kutter, E. F. (2017). *How the mind comes into being: Introducing cognitive science from a functional and computational perspective*. Oxford, UK: Oxford University Press.

Friston, K. (2009). The free-energy principle: a rough guide to the brain? *Trends in Cognitive Sciences*, *13*(7), 293–301.

Graziano, M. S., & Cooke, D. F. (2006). Parieto-frontal interactions, personal space, and defensive behavior. *Neuropsychologia*, *44*(6), 845–859.

Greenwald, A. G. (1970). Sensory feedback mechanisms in performance control: with special reference to the ideomotor mechanism. *Psychological review*, *77*(2), 73–99.

Herbort, O., & Butz, M. V. (2012). The continuous end-state comfort effect: Weighted integration of multiple biases. *Psychological Research*, *76*, 345-363. doi: 10.1007/s00426-011-0334-7

Hoffmann, J. (2003). Anticipatory behavioral control. In M. V. Butz, O. Sigaud, & P. Gerard (Eds.), *Anticipatory behavior in adaptive learning systems: Foundations, theories, and systems* (pp. 44–65). Berlin: Springer-Verlag.

Holmes, N. P. (2012). Does tool use extend peripersonal space? a review and re-analysis. *Experimental Brain Research*, *218*(2), 273–282.

Holmes, N. P., & Spence, C. (2004). The body schema and multisensory representation(s) of peripersonal space. *Cognitive Processing*, *5*(2), 94–105.

Hommel, B., Müsseler, J., Aschersleben, G., & Prinz, W. (2001). The theory of event coding (TEC): A framework for perception and action planning. *Behavioral and Brain*

Sciences, *24*, 849–878.

Lawrence, M. A. (2015). ez: Easy analysis and visualization of factorial experiments [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=ez (R package version 4.3)

Lohmann, J., Belardinelli, A., & Butz, M. V. (2019). Hands ahead in mind and motion: Active inference in peripersonal hand space. *Vision*, *3*(2), 15. doi: 10.3390/vision3020015

Parsons, L. M. (1987). Imagined spatial transformations of one's hands and feet. *Cognitive Psychology*, *19*(2), 178–241.

Patané, I., Cardinali, L., Salemme, R., Pavani, F., Farnè, A., & Brozzoli, C. (2018). Action planning modulates peripersonal space. *Journal of Cognitive Neuroscience*, 1–14.

Qu, F., Wang, J., Zhong, Y., & Ye, H. (2018). Postural effects on the mental rotation of body-related pictures: An fmri study. *Frontiers in psychology*, *9*. doi: 10.3389/fpsyg.2018.00720

R Core Team. (2016). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Richmond, L. L., & Zacks, J. M. (2017). Constructing experience: Event models from perception to action. *Trends in Cognitive Sciences*, *21*(12), 962–980. doi: 10.1016/j.tics.2017.08.005

Spence, C., Pavani, F., Maravita, A., & Holmes, N. (2004). Multisensory contributions to the 3-d representation of visuotactile peripersonal space in humans: Evidence from the crossmodal congruency task. *Journal of Physiology*, *98*(1), 171–189.

Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., & Reynolds, J. R. (2007). Event perception: A mind-brain perspective. *Psychological Bulletin*, *133*(2), 273–293. doi: 10.1037/0033-2909.133.2.273

Developmental changes in the ability to draw distinctive features of object categories

Bria Long
Department of Psychology
Stanford University
450 Serra Mall
Stanford, CA 94305
bria@stanford.edu

Judith E. Fan
Department of Psychology
UC San Diego
9500 Gilman Drive
La Jolla, CA 92093
jefan@ucsd.edu

Zixian Chai
Department of Psychology
Stanford University
450 Serra Mall
Stanford, CA 94305
zchai14@stanford.edu

Michael C. Frank
Department of Psychology
Stanford University
450 Serra Mall
Stanford, CA 94305
mcfrank@stanford.edu

Abstract

How do children's visual concepts change across childhood, and how might these changes be reflected in their drawings? Here we investigate developmental changes in children's ability to emphasize the relevant visual distinctions between object categories in their drawings. We collected over 13K drawings from children aged 2-10 years via a free-standing drawing station in a children's museum. We hypothesized that older children would produce more recognizable drawings, and that this gain in recognizability would not be entirely explained by concurrent development in visuomotor control. To measure recognizability, we applied a pretrained deep convolutional neural network model to extract a high-level feature representation of all drawings, and then trained a multi-way linear classifier on these features. To measure visuomotor control, we developed an automated procedure to measure their ability to accurately trace complex shapes. We found consistent gains in the recognizability of drawings across ages that were not fully explained by children's ability to accurately trace complex shapes. Furthermore, these gains were accompanied by an increase in how distinct different object categories were in feature space. Overall, these results demonstrate that children's drawings include more distinctive visual features as they grow older.

Keywords: object representations; child development; visual production; deep neural networks

Introduction

Children draw prolifically, providing a rich source of potential insight into their emerging understanding of the world (Kellogg, 1969). Accordingly, drawings have been used to probe developmental change in a wide variety of domains (Fury, Carlson, & Sroufe, 1997; Karmiloff-Smith, 1990; e.g., Piaget, 1929). In particular, drawings have long provided inspiration for scientists investigating how children represent visual concepts (Minsky & Papert, 1972). For example, even when drawing from observation, children tend to include features that are not visible from their vantage point, yet are diagnostic of category membership (e.g., a handle on a mug) (Barrett & Light, 1976; Bremner & Moore, 1984).

As children learn the diagnostic properties of objects and how to recognize them, they may express this knowledge in their drawings of these categories. Indeed, children's visual recognition abilities have a protracted developmental trajectory: configural visual processing—the ability to process relationships between object parts (Juttner, Muller, & Rentschler, 2006; Juttner, Wakui, Petters, & Davidoff, 2016)—may mature slowly throughout childhood, as does the ability to recognize objects under unusual poses or lighting (Bova et al., 2007).

Inspired by this prior work, our goal is to understand the relationship between developmental changes in how children draw visual concepts and their representations of these visual concepts. In particular, we hypothesize that children's drawings become more recognizable in part because children learn the distinctive features of categories that set them apart from other similar categories (Figure 1). If so, we would expect an increase in the distinctiveness of children's drawings across childhood that is not explained by improvements in children's visuomotor ability. However, this goal poses several methodological challenges to overcome.

First, it requires a principled and generalizable approach to encoding the high-level visual properties of drawings that expose the extent to which they contain category-diagnostic information (Fan, Yamins, & Turk-Browne, 2018). This approach stands in contrast to previous approaches, which have relied upon provisional criteria specific to each study (e.g., handles for mugs) (e.g., Barrett & Light, 1976; Goodenough, 1963), which limited their ability to make detailed predictions on new tasks or datasets. Recently, deep convolutional neural network (DCNN) models that have been trained on challenging object recognition tasks have been shown to extract high-level visual information from images (Yamins et al., 2014). As these models have been directly optimized to recognize objects in photographs, features in higher layers of these networks represent high-level visual information that is important for distinguishing between object categories. We thus meet this challenge by capitalizing on prior work validating the use of these higher-layer features to analyze the high-level visual information in drawings (Fan et al., 2018; Long, Fan, & Frank, 2018). In particular, we investigate the extent to which children include distinctive features in their drawings by assessing how well these visual features can be used to identify the category (e.g., dog, bird) that children were intending to draw.

Second, it requires a large sample of drawings collected under consistent conditions from a wide range of participants to identify robust developmental patterns (e.g., M. Frank et al., 2017). This is in contrast to the relatively small samples that have characterized classic studies in this domain (Bremner & Moore, 1984; Karmiloff-Smith, 1990). To meet this challenge, we installed a free-standing drawing station in a local science museum, allowing us to collect a large sample of drawings (N = 13205 drawings) of 23 object categories

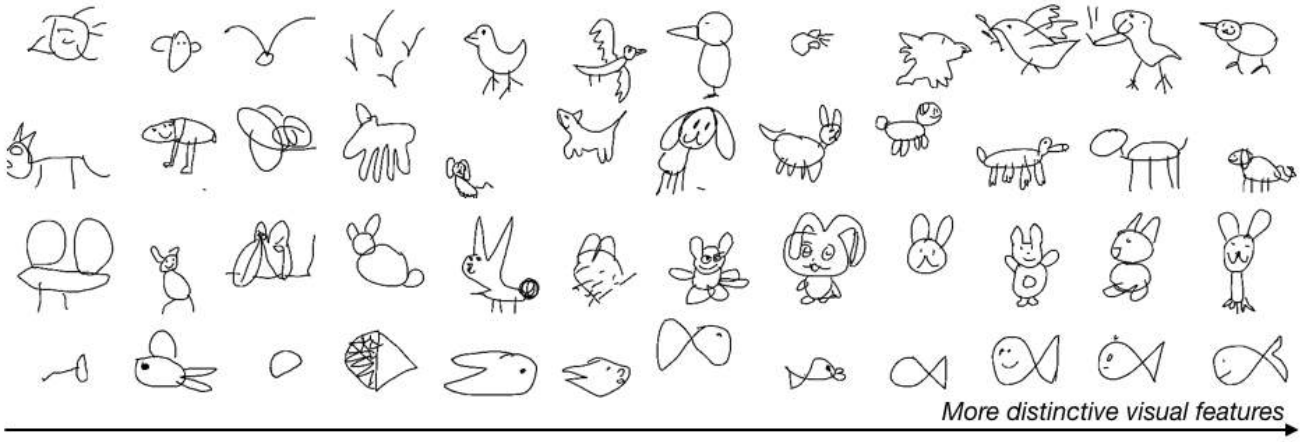


Figure 1: Examples of drawings that have increasingly more distinctive visual features of their categories, making them more easily recognizable. These examples are generated from the results of the classification process outlined below.

over a wide developmental age range (i.e., 2-10 years) under consistent task conditions.

Third, it requires simultaneous and detailed measurement of developmental changes in other cognitive and motor abilities that may influence children's ability to include relevant information in their drawing (Freeman, 1987; Rehrig & Stromswold, 2018). For example, children's developing visuomotor abilities may limit their ability to include the diagnostic visual features in their drawings. In this paper, we focus on visuomotor control, operationalized as performance on shape tracing tasks, because they share many of the same demands on controlled, visually-guided movement with our primary object drawing task. Critically, because we collected both tracings and drawings from every participant in our dataset, we are able to model the contribution of both individual and age-related variation in tracing task performance for explaining how well children produce recognizable drawings.

In sum, our paper provides an advance over our prior work investigating developmental change in drawing behavior (Long et al., 2018) in three ways: first, we build a free-standing drawing station to continually crowdsource children's drawings under consistent conditions, enabling the collection of a substantially larger dataset; second, we exploit this larger dataset to characterize the category-level distinctiveness inherent to children's drawings across a wide range of ages; and third, we develop an automated procedure for analyzing concurrent changes in visuomotor control using a tracing task.

Methods

Dataset

Drawing Station We installed a drawing station that featured a tablet-based drawing game in a local science museum. Each participant sat in front of a table-mounted touchscreen tablet and drew by moving the tip of their finger across the display. Participants gave consent and indicated their age (in years 2-10 or adult) via checkboxes and no other identifying information was collected; our assumption was that parents would navigate this initial screen for children. To measure fine visuomotor control, each session began with two tracing trials, followed by a copying trial. On each tracing trial, participants were presented with a shape in the center of the display. The first shape was a simple square, and the second was a more complex star-like shape (Figure 2). On the subsequent copying trial, participants were presented with a simple shape (square or circle) in the center of the display for 2s, which then disappeared. They then were asked to copy the shape in the same location it had initially appeared. Next, participants completed up to eight object drawing trials. On each of these trials, participants were verbally cued to draw a particular object category by a video recording of an experimenter (e.g., "What about a dog? Can you draw a dog?"). On all trials, participants had up to 30 seconds to complete their tracing, copy, or drawing. There are 23 common object categories represented in our dataset, which were collected across three bouts of data collection focused on 8 of these objects at a time. These categories were chosen to be familiar to children, to cover a wide range of superordinate categories (e.g., animals, vehicles, manipulable objects), and to vary in the degree to which they are commonly drawn by young children (e.g., trees vs. keys).

Dataset Filtering & Descriptives Given that we could not easily monitor all environmental variables at the drawing station that could impact task engagement (e.g., ambient noise, distraction from other museum visitors), we anticipated the need to develop robust and consistent procedures for data quality assurance. We thus adopted strict screening procedures to ensure that any age-related trends we observed were not due to differences in task compliance across age. Early on, we noticed an unusual degree of sophistication in 2-year-

old participants' drawings and suspected that adult caregivers accompanying these children may not have complied with task instructions to let children draw on their own. Thus, in later versions of the drawing game, we surveyed participants to find out whether another child or an adult had also drawn during the session; all drawings where interference was reported were excluded from analyses. Out of these 2685 participants, 700 filled out the survey, and 156 reported interference from another child or adult (5.81%). Raw drawing data (N = 15594 drawings) were then screened for task compliance using a combination of manual and automated procedures (i.e., excluding blank drawings, pure scribbles, and drawings containing words), resulting in the exclusion of 15.3% of all drawings (N = 13205 drawings after exclusions). After filtering, we analyzed data from 2443 children who were on average 5.28 years of age (range 2-10 years).

Measuring Tracing Accuracy

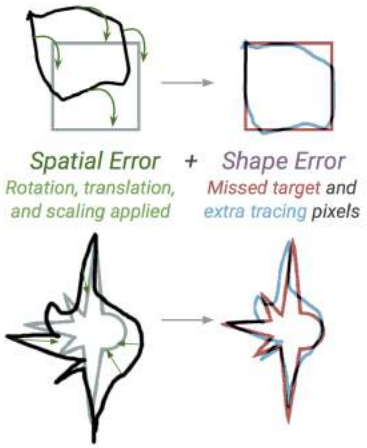


Figure 2: Measurement of tracing task performance reflects both spatial and shape error components. Left: The grey shape is the target; the black shape is the raw tracing. After applying affine image registration, the spatial error reflects the extent of translation, rotation, and scaling transformation required to minimize shape error. Right: Shape error reflects how closely the contour of the transformed tracing aligns with the target.

We developed an automated procedure for evaluating how accurately participants performed the tracing task, validated against empirical judgments of tracing quality. We decompose tracing accuracy into two components: a shape error component and a spatial error component. Shape error reflects how closely the participant's tracing matched the contours of the target shape; the spatial error reflects how closely the location, size, and orientation of the participant's tracing matched the target shape (Figure 2).

To compute these error components, we applied an image registration algorithm, AirLab (Sandkhler, Jud, Andermatt, & Cattin, 2018), to align each tracing to the target shape, yielding an affine transformation matrix that minimized the pixel-wise correlation distance between the aligned tracing, T, and the target shape, S: $Loss_{NCC} = -\frac{\sum S \cdot T - \sum E(S)E(T)}{N \sum Var(S)Var(T)}$, where N is the number of pixels in both images.

The shape error was defined by the final correlation distance between the aligned tracing and the target shape. The spatial error was defined by the magnitude of three distinct error terms: location, orientation, and size error, derived by decomposing the affine transformation matrix above into translation, rotation, and scaling components, respectively. In sum, this procedure yielded four error values for each tracing: one value representing the shape error (i.e., the pixel-wise correlation distance) and three values representing the spatial error (i.e., magnitude of translation, rotation, scaling components).

Although we assumed that both shape and spatial error terms should contribute to our measure of tracing task performance, we did not know how much weight to assign to each component to best predict empirical judgments of tracing quality. In order to estimate these weights, we collected quality ratings from adult observers (N=70) for 1325 tracings (i.e., 50-80 tracings per shape per age), each of which was rated 1-5 times. Raters were instructed to evaluate "how well the tracing matches the target shape and is aligned to the position of the target shape" on a 5-point scale.

We fit an ordinal regression mixed-effects model to predict these 5-point ratings, which contained correlation distance, translation, rotation, scaling, and shape identity (square vs. star) as predictors, with random intercepts for rater. This model yielded parameter estimates that could then be used to score each tracing in the remainder of the dataset (N=3242 tracings from 1886 children). We averaged scores within session to yield a single tracing score for each participant (2245 children completed at least one tracing trial).

Measuring Object Drawing Recognizability

We also developed an automated procedure for evaluating how well participants included category-diagnostic information in their drawings by examining classification performance on the features extracted by a deep convolutional neural network model.

Visual Encoder To encode the high-level visual features of each sketch, we used the VGG-19 architecture (Simonyan & Zisserman, 2014), a deep convolutional neural network pretrained on Imagenet classification. We used model activations in the second-to-last layer of this network, which contain more explicit representations of object identity than earlier layers (Fan et al., 2018; Long et al., 2018; Yamins et al., 2014). Raw feature representations in this layer consist of flat 4096-dimensional vectors, to which we applied channel-wise normalization.

Logistic Regression Classifier Next, we used these features to train an object category decoder. To avoid any bias due to imbalance in the distribution of drawings over cate-

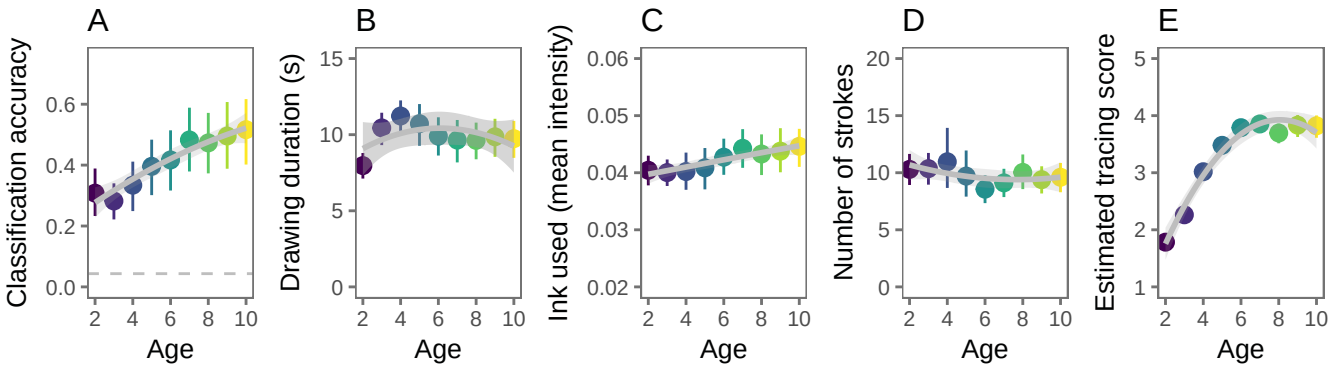


Figure 3: (A) Leave-one-out classification accuracy (grey dotted line indicates chance) (B) the amount of time spent drawing in seconds, (C) the amount of ink used (i.e., mean intensity of the drawings), (D) the number of strokes used, and (E) the average estimated tracing scores are plotted as a function of childrens age.

gories (since groups of categories ran at the station for different times), we sampled such that there were an equal number of drawings of each of the 23 categories (N=8694 drawings total). We then trained a 23-way logistic classifier with L2 regularization under leave-one-out cross-validation to estimate the recognizability of every drawing in our dataset.

Predicting Object Drawing Recognizability If children's drawings contain more features that are diagnostic of the drawn categories, then these visual features (estimated via VGG-19) should lead to greater classification accuracy. However, we anticipated that classification accuracy may also vary with children's tracing abilities as well how much time and effort children invested in their drawings; we thus recorded how much time was taken to produce each drawing, how many strokes were drawn, and the proportion of the drawing canvas that was filled. Our main statistical model was then a generalized linear mixed-effects model predicting classification accuracy from the category decoder, with scaled age (in years), tracing score (averaged over both trials), and effort cost variables (i.e., time, strokes, ink) modeled as fixed effects, and with random intercepts for each child and object category.

Measuring Category Distinctiveness To investigate changes in the underlying feature representation of children's drawings that may help explain variation in classification accuracy, we computed a measure of pairwise category distinctiveness D_{ij} for each pair of categories i, j within each age. This metric is a higher-dimensional analog of d-prime that incorporates both the distance between each pair of categories as well as the dispersion within each category. We first computed the category centers as the mean feature vector for each category, $\vec{r}_i$ and $\vec{r}_j$. The distance between each pair of categories i, j was then taken as the Euclidean distance between their category centers, $\|\vec{r}_i - \vec{r}_j\|_2$. The dispersion for each category was computed as the root-mean-squared Euclidean distance of each individual drawing vector from the category center vector $\vec{r}$ and is expressed as s. By direct analogy with d-prime, we compute the distinctiveness D_{ij} of each pair of categories i, j by dividing the Euclidean distance between category centers by the quadratic mean of the two category dispersions, $D_{ij} = \frac{\|\vec{r}_i - \vec{r}_j\|_2}{\sqrt{\frac{1}{2}(s_i^2 + s_j^2)}}$.

Results

Overall, drawing classification accuracy increased with age (Figure 3A), validating our basic expectation that older children's drawings would be more recognizable. Our mixed-effects model on drawing classification revealed that this age-related gain held when accounting for task covariates—the amount of time spent drawing, the number of strokes, and total ink used (Figure 3B,C,D)—and for variation across object categories and individual children. All model coefficients can be found in Table 1.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.714	0.274	-2.606	0.009
Tracing	0.311	0.034	9.141	0.000
Age	0.282	0.033	8.499	0.000
Draw Duration	0.136	0.034	3.976	0.000
Avg Intensity	-0.064	0.033	-1.910	0.056
Num. Strokes	-0.034	0.034	-1.009	0.313
Tracing*Age	0.011	0.029	0.357	0.721

Table 1: Model coefficients of a GLMM predicting the recognziability of each drawing

We next examined the relationship between children's ability to trace complex shapes and the subsequent recognizability of their drawings. Tracing abilities increased with age (Figure 3E) and individual's tracing abilities were good predictors of the recognizability of the drawings they produced. This main effect of tracing ability also held when accounting for effort covariates (number of strokes, time spent drawing, ink used). However, children's tracing abilities did not interact with the age-related gains in classification we observed (Figure 4) and we observed age-related classification gains at each level of tracing ability.

To examine the contributions of age and tracing ability to

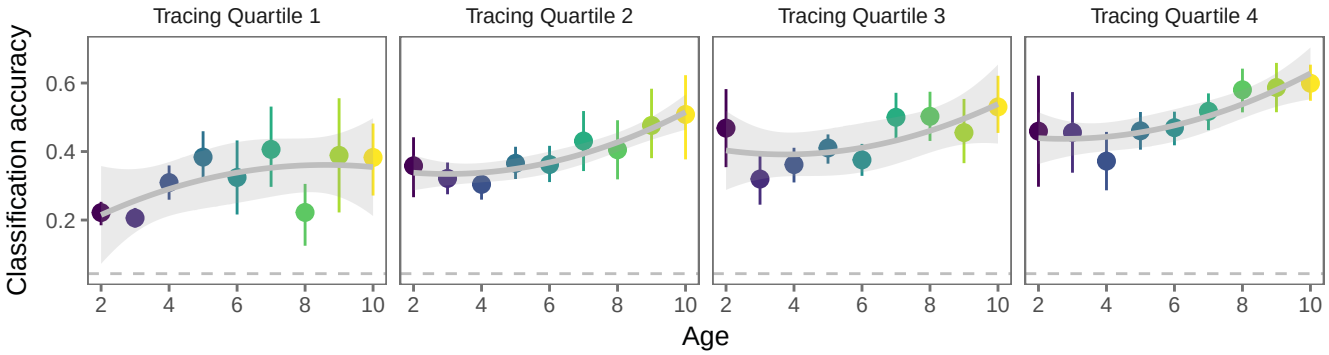


Figure 4: Data are divided into four quantiles based on the distribution of tracing scores in the entire dataset; these divisions represent the data in each panel. In each panel, the average classification accuracy is plotted as a function of childrens age. Error bars represent 95% CIs bootstrapped within each age group and subset of tracing scores; grey dotted lines indicate chance.

recognizability, we also fit reduced versions of the full model and examined the marginal R^2 (Nakagawa & Schielzeth, 2013). The fixed effects in a null model without tracing or age (which mainly captures drawing effort) accounted for very little variance (marginal R^2 = 0.004). Adding only children's age to the model increased R^2 (marginal R^2 = 0.037) as did only adding tracing (marginal R^2 = 0.039). Adding both factors without their interaction (marginal R^2 = 0.05) had a similar effect to adding both factors and their interaction (marginal R^2 = 0.05). Attesting to the immense variability between individuals and categories, adding random effects (and many more parameters) accounted for a much larger amount of variance (conditional R^2 for full model = 0.403). Finally, as we had many more younger participants in our dataset, we also repeated these analyses with a subset of the dataset that was balanced across both children's age and category (N=2691 drawings), and found the same pattern of results.

These age-related changes in classification accuracy show that the underlying feature representations of older children's drawings were more linearly discriminable. This finding led us to investigate a potential source of this enhanced discriminability: that drawings from different categories were spread further apart in feature space, while drawings within a category were clustered closer together. To evaluate this possibility, we used a measure of pairwise category distinctiveness D_{ij} that accounts for both the distance between each pair of categories, as well as the dispersion within each category. We found that category distinctiveness increased consistently with age (Figure 5).

Taken together, these results reveal developmental changes in how well children are able to emphasize the relevant distinctions between object categories in their drawings that thereby support recognition. Moreover, they show that these age-related gains in classification are not entirely explained by concurrent development in visuomotor control.

General Discussion

How do children represent different object categories throughout childhood? Drawings are a rich potential source of information about how visual representations change over development. One possibility is that older children's drawings are more recognizable because children are better able to include the diagnostic features of particular categories that distinguish them from other similar objects. Supporting this hypothesis, the high-level visual features present in children's drawings could be used to estimate the category children were intending to draw, and these classifications became more accurate as children became older. These age-related gains in classification were not entirely explainable by either low-level effort covariates (e.g., amount of time spent drawing, average intensity, or number of strokes) or children's tracing abilities. In addition, these gains in classification were paralleled by an increase in the distinctiveness between the categories that children drew (Figure 5).

Taken together, these results suggest that children's drawings contain more distinctive features as they grow older, perhaps reflecting a change in their internal representations of these categories. While children could simply be learning routines to draw certain categories—perhaps from direct instruction or observation, our results held even when restricted to a subset of very rarely drawn categories (e.g., couch, scissors, key) arguing against a simple version of this idea.

Nonetheless, there are limitations on the generalizability of these findings due to the nature of our dataset. First, while this dataset is large and samples a heterogenous population, all drawings were collected at a single geographical location, limiting the generalizability of these results to children from other diverse cultural or socioeconomic backgrounds. Second, while we imposed strong filtering requirements on the dataset, we were not present while the children were drawing and thus cannot be sure that we've eliminated all sources of noise or interference. At the same time, additional interference would only generate extra noise in our data rather than the observed age-related trends. In any case, these correlational results call for validation in more carefully controlled

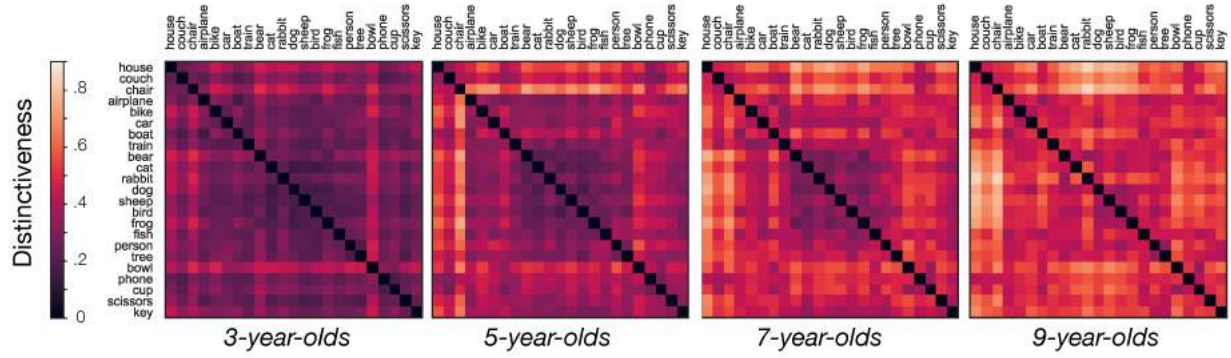


Figure 5: Pairwise category distinctiveness for drawings made by 3-, 5-, 7-,and 9-year-olds; darker (vs. lighter) values represent pairs of categories that have more overlapping (vs. distinctive) representations.

contexts and across more diverse populations.

Furthermore, they open the door for future empirical work to establish causal links between children's drawing behavior and their changing internal representation of visual concepts. For example, it would be valuable to explore the extent to which a child's ability to include the most distinctive visual features in their drawings of object categories predicts their ability to perceptually discriminate those object categories. Another promising direction would be to investigate the relationship between children's general ability to retrieve relevant information from semantic memory (e.g., that a rabbit has long ears and whiskers), and their ability to produce recognizable drawings of those categories. Insofar as such retrieval mechanisms are engaged during drawing production, developmental changes in semantic memory systems may also explain an important portion of the age-related variation in drawing behavior.

Overall, we suggest that children's drawings change systematically across development, and that they contain rich information about children's underlying representations of the categories in the world around them. A full understanding of how children's drawings reflect their emerging perceptual and conceptual knowledge will allow a unique and novel perspective on the both the development and the nature of visual concepts—the representations that allow us to easily derive meaning from what we see.

Acknowledgements

We thank the San Jose Children's Discovery Museum for their collaboration and for hosting the drawing station where these data were collected. We are also grateful to members of the Stanford Language and Cognition lab for their feedback. This work was funded by an NSF SPRF-FR Grant #1714726 to BLL and a Jacobs Foundation Fellowship to MCF.

References

Barrett, M., & Light, P. (1976). Symbolism and intellectual realism in children's drawings. *British Journal of Educational Psychology*, *46*(2), 198–202.

Bova, S. M., Fazzi, E., Giovenzana, A., Montomoli, C., Signorini, S. G., Zoppello, M., & Lanzi, G. (2007). The development of visual object recognition in school-age children. *Developmental Neuropsychology*, *31*(1), 79–102.

Bremner, J. G., & Moore, S. (1984). Prior visual inspection and object naming: Two factors that enhance hidden feature inclusion in young children's drawings. *British Journal of Developmental Psychology*, *2*(4), 371–376.

Fan, J., Yamins, D., & Turk-Browne, N. (2018). Common object representations for visual production and recognition. *Cognitive Science*, *0*(0).

Frank, M., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., ... others. (2017). A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. *Infancy*, *22*(4), 421–435.

Freeman, N. H. (1987). Current problems in the development of representational picture-production. *Archives de Psychologie*.

Fury, G., Carlson, E. A., & Sroufe, A. (1997). Children's representations of attachment relationships in family drawings. *Child Development*, *68*(6), 1154–1164.

Goodenough, F. L. (1963). *Goodenough-harris drawing test*. Harcourt Brace Jovanovich New York.

Juttner, M., Muller, A., & Rentschler, I. (2006). A developmental dissociation of view-dependent and view-invariant object recognition in adolescence. *Behavioural Brain Research*, *175*(2), 420–424.

Juttner, M., Wakui, E., Petters, D., & Davidoff, J. (2016). Developmental commonalities between object and face recognition in adolescence. *Frontiers in Psychology*, *7*.

Karmiloff-Smith, A. (1990). Constraints on representational change: Evidence from children's drawing. *Cognition*, *34*(1), 57–83.

Kellogg, R. (1969). *Analyzing children's art*. National Press Books Palo Alto, CA.

Long, B., Fan, J., & Frank, M. C. (2018). Drawings as a window into developmental changes in object representations. In *Proceedings of the 40th annual meeting of the cognitive*

science society.

Minsky, M., & Papert, S. A. (1972). Artificial intelligence progress report.

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining r2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*(2), 133–142.

Piaget, J. (1929). The child's concept of the world. *Londres, Routldge & Kegan Paul*.

Rehrig, G., & Stromswold, K. (2018). What does the dap: IQ measure?: Drawing comparisons between drawing performance and developmental assessments. *The Journal of Genetic Psychology*, *179*(1), 9–18.

Sandkhler, R., Jud, C., Andermatt, S., & Cattin, P. C. (2018). AirLab: Autograd image registration laboratory. *ArXiv Preprint ArXiv:1806.09907*.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *ArXiv Preprint ArXiv:1409.1556*.

Yamins, D., Hong, H., Cadieu, C., Solomon, E., Seibert, D., & DiCarlo, J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624.

Unconscious Number Discrimination in the Human Visual System

Che Lucero
Cornell University, Ithaca, New York, United States

Geoffrey Brookshire
University of Chicago, Chicago, Illinois, United States

Roberto Bottini
University of Trento, Trento, Italy

Susan Goldin-Meadow
University of Chicago, Chicago, Illinois, United States

Edward Vogel
University of Chicago, Chicago, Illinois, United States

Daniel Casasanto
Cornell University, Ithaca, New York, United States

Abstract

How do humans compute approximate number? According to one influential theory, approximate number representations arise in the intraparietal sulcus and are amodal (independent of any sensory modality). Alternatively, approximate number may be computed initially within sensory systems. We tested for approximate number representations in the visual system using steady state visual evoked potentials (SSVEPs). We recorded EEG from human subjects while they viewed dotclouds presented at 30 Hz. Alternating the dotcloud numerosity at 15 Hz evoked a 15 Hz SSVEP detectable over the occipital lobe (Oz). The SSVEP amplitude increased as the numerical difference between dotclouds increased, indicating that subjects visual systems were differentiating dotclouds on the basis of their numerical ratios. Critically, subjects were unable to consciously discriminate dotcloud numerosity, indicating the rapid presentation disrupted reentrant feedback to visual cortex. Approximate number appears to be computed within the visual system, independently of higher-order areas such as the intraparietal sulcus.

Limits on the Use of Simulation in Physical Reasoning

Ethan Ludwin-Peery[1] (elp327@nyu.edu), **Neil R. Bramley**[2] (Neil.Bramley@ed.ac.uk)
Ernest Davis[3] (davise@cs.nyu.edu), **Todd M. Gureckis**[1] (todd.gureckis@nyu.edu)
[1]Department of Psychology, NYU, New York, [2]Department of Psychology, University of Edinburgh, Edinburgh, Scotland,
[3]Department of Computer Science, NYU, New York

Abstract

In this paper, we describe three experiments involving simple physical judgments and predictions, and argue their results are generally inconsistent with three core commitments of probabilistic mental simulation theory (PMST). The first experiment shows that people routinely fail to track the spatio-temporal identity of objects. The second experiment shows that people often incorrectly reverse the order of consequential physical events when making physical predictions. Finally, we demonstrate a physical version of the conjunction fallacy where participants rate the probability of two joint events as more likely to occur than a constituent event of that set. These results highlight the limitations or boundary conditions of simulation theory.

Keywords: intuitive physics; mental simulation; inference; conjunction fallacy

Introduction

Successful interaction with our environment often requires reasoning about the physical world (e.g., predicting if a stack of books on a desk is unstable), but the mental processes that support this ability remain poorly understood. Simulation is a technique used for physical reasoning in many applications ranging from modeling molecular interactions to designing realistic video game physics engines. In a simulation, a program starts with an initial state, and then applies the relevant dynamic laws of physics to compute what will happen over a series of (typically short) time steps; in effect, computing a "movie" of how the scenario progresses.

Some researchers have recently argued that humans use cognitive strategies analogous to computer simulation when intuitively reasoning and making predictions about the physical world (Battaglia, Hamrick, & Tenenbaum, 2013; Hamrick, Battaglia, Griffiths, & Tenenbaum, 2016; Hamrick, Smith, Griffiths, & Vul, 2015; Smith, Dechter, Tenenbaum, & Vul, 2013; Smith & Vul, 2017; Ullman, Spelke, Battaglia, & Tenenbaum, 2017). In order to account for the more imprecise and qualitative nature of human physical reasoning, they propose that multiple simulations are run from a range of different initial configurations. For instance, consider a person asked to predict whether a tower of wooden blocks will fall over. A *probabilistic* mental simulation begins by assuming that each observer has an imperfect perception of the positions of the blocks (i.e., their precise locations in physical space) owing to perceptual limitations and occlusion. Based on this uncertain percept, the simulator samples a number of slightly different towers, each altered according to random (perceptual) noise. According to the theory, a reasoner might start with, for instance, ten initial towers and run a (possibly noisy) physics simulation forward until some termination point with the resulting outcomes driving their stability judgment. For example, if 8 of the 10 simulated towers fall over then a reasoner might estimate a 0.8 probability that the structure is unstable (Battaglia et al., 2013). We refer to this approach as "probabilistic mental simulation theory" (PMST).

PMST has been found to approximate human judgments in a diverse set of tasks, including judging how a 3-D tower of blocks will collapse (Battaglia et al., 2013), predicting the destination of a virtual ball on a 2-D bumper table (Smith et al., 2013), and predicting the proportion of a poured liquid that will end up on either side of a divider (Bates, Yildirim, Tenenbaum, & Battaglia, 2015), among others. However, this theory has been contested (cf., Davis & Marcus, 2015, 2016). Criticisms of this theory include the incompatibility of an accurate physical simulation engine with decades of psychological work documenting human errors in simplified physical reasoning tasks (Hegarty, 2004; Kubricht, Holyoak, & Lu, 2017; McCloskey, Caramazza, & Green, 1980; Proffitt, Kaiser, & Whelan, 1990; Siegler, 1976). In addition, in many situations, simulation would be computationally inefficient or impossible. For instance, if a closed can half full of sand is shaken, simulation would require calculating all the collisions of all the grains of sand (e.g., Kubricht et al., 2016) but if the goal is just to predict whether the sand remains in the can, that can be done through the application of a simple rule (Smith et al., 2013)

The goal of the present paper is to provide a strong empirical test of PMST. We begin by describing three core tenets of PMST that transcend specific applications of the theory and make important testable claims about human physical reasoning. We then describe three novel experiments that test these principles by setting up pre-registered (here) edge-cases where we might expect the predictions made by PMST to fail.

Three key principles of probabilistic mental simulation theory (PMST)

An agent using a probabilistic simulation of the physical world to solve physical reasoning problems should adhere to the following three principles. While we accept that PMST may include limitations and shortcuts (Ullman et al., 2017), the principles outlined here are necessary for simulation to be a viable strategy.

Object Persistence A reasoner using PMST is required to maintain interacting objects within all simulations/samples. Objects occupy particular locations within space and time and a mental simulation must encode these relative spatiotemporal positions and update them according to the rules of physics. This is a core aspect of the theory, because drop-

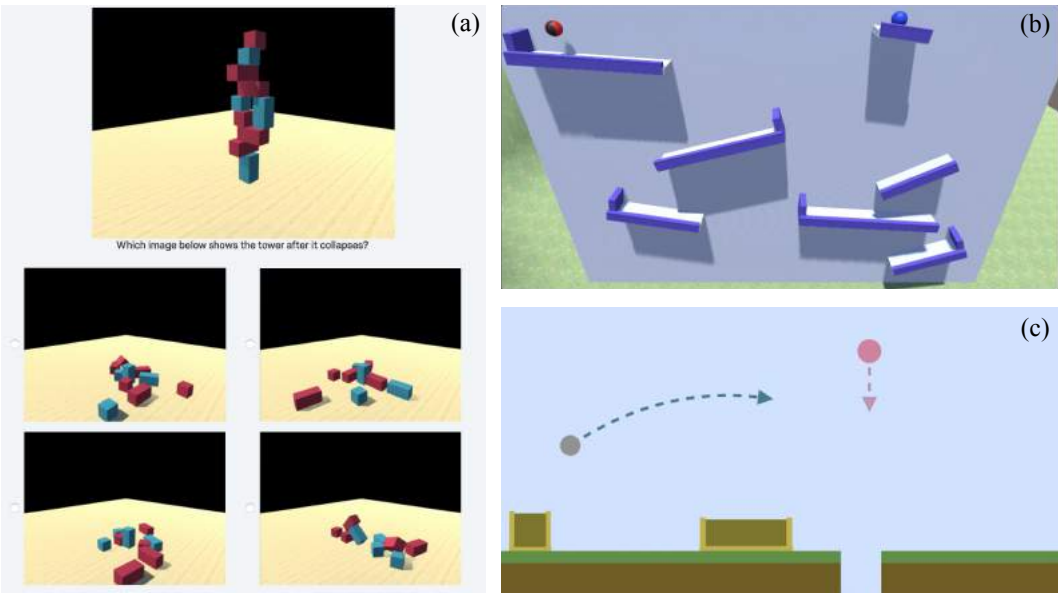


Figure 1: (a) A "Minus One" block tower question, as it appeared to participants. The answer in the upper left is correct; the other three answers are each missing one block. (b) A "marble run"-type temporal consistency question. (c) An example probabilistic coherence scene. Dotted arrows indicate approximate motion over the 2/3 second clip.

ping an object from a simulation or deleting it (as is possible in a video game engine) would radically alter possible outcomes and subsequent predictions. For example, imagine a person thinking about a table. If their mental simulation accidentally deleted its representation one of the table legs (even temporarily), the result would be a major disruption to the simulation; e.g., the otherwise static table might begin to fall. Keeping track of the location of objects in space and time and accounting properly for their movements is fundamental to what it means to "simulate" a physical scene. Any plausible physics engine must keep track of all the interacting objects involved in order to maintain coherent predictions about the future. Physics engines do make mistakes and approximations, but deleting or radically altering objects is not the kind of mistake they make.

There are, of course, some cases where objects may be ignored. In video game physics engines, for example, objects at rest are often put to "sleep" to save on computation, and it has been suggested that mental simulation might make use of this trick as well (Ullman et al., 2017). When a physics engine puts an object to sleep, however, this simply means that the engine assumes that the object is stationary, and it does not mean that the physics engine forgets that the object exists.

This leads to the key prediction tested in our first experiment: in interacting multi-object scenes, every object from the initial percept will be represented in each simulation's final state, because every object is necessarily represented and tracked throughout each simulation.

Temporal Consistency Building upon the first principle, an iterative simulation must advance all interacting objects simultaneously. This step-by-step, synchronous nature of the simulation ensures that the order of events is preserved. Preserving the order of events is important for generating accurate simulations and using them to make decisions. When two processes might interact, it is necessary for their simulations to be properly synchronized in order to predict whether and how they interact. Consider a case where a bottle is rolling towards the front door of a house, which is slowly closing. To predict if the bottle ends up inside or outside the house, the simulator has to represent whether the door will swing shut before the bottle gets there. Time cannot for instance run faster for the bottle than for the door if reasoning is to be coherent. A synchronous approach does this, and ensures that there is no way for one event to get ahead of, or fall behind another, because they share a common timeline.

We refer to this property of mental simulation as *temporal consistency*. A person using PMST to reason about a physical scene should preserve the temporal order of events.

Probabilistic Coherence According to PMST, after running multiple (noisy) simulations, the final scene configurations from each simulation are used to make predictions and inferences about the physical world. A variety of ways of aggregating across these simulations have been proposed. For example, in Battaglia et al. (2013), the output of the model was the average proportion of towers that fell across the set of the simulations. In this example, PMST uses the Monte Carlo principle to estimate probabilities. An event that is almost certain to occur will occur in all the simulations while a more uncertain event (or one more sensitive to perceptual noise) will occur less frequently. Although approximate, computing probabilities from samples or simulations still conforms to the axioms of probability theory. Indeed, this is a key virtue of the approach, and helps to relate the theory to existing Bayesian theories of human inference.

One classic signature of coherent probabilistic reasoning is that the probability of a conjunction of multiple events must always be less than the probability of any component (i.e., $P(A \wedge B) \leq P(A)$). However, people in many cases will estimate conjunctions to be more likely than one of their components (known as the "conjunction fallacy", Tversky & Kahneman, 1983). While they have been found in a number of domains including social reasoning, conjunction fallacy errors have not (to our knowledge) been observed in reasoning about physical outcomes, and all of the methods PMST proposes for estimating probability from the results of simulation predict that conjunction fallacies should not regularly occur. If likelihood is calculated by tallying the relative outcomes of different random simulations, the conjunction rule will not be systematically violated, because it is impossible for a sample to have more outcomes that include a conjunction than outcomes that include one of the constituent elements.

Study 1: Object Persistence

The first experiment tested the principle of Object Persistence. In particular, we tested if people are able to keep track of the number, size, and color/identity of a relatively small number of objects when predicting the future state of a simple scene. If people fail consistently at this task, it calls into question a key assumption of PMST; that simulations preserve objects over time. The assumption that we make in designing this test is that if an object is represented and tracked in each simulation, then it should be available for other judgements such as being identified/recognized. If people are limited in this regard, it calls into question if people use PMST or, alternatively, refines this theory by pointing out the cognitive inaccessibility of object-level details from the mental representation of the scene.

The experiment builds upon the block tower designs first used by Battaglia et al. (2013). Rather than asking participants to make predictions about the collapse of a standing tower, we showed participants one standing tower (the target) and then a set of four collapsed towers and asked them to judge which collapsed option was the result of the target tower falling according to gravity (with one of the four options being the ground truth of running the standard tower through a physics engine). Given the target tower, a simulation based reasoner could simply simulate the standing tower forward to generate one or a set of collapsed tower states. The actual result of the tower collapsing should be similar to several results generated by the simulation.

Method

Participants We ran groups of 9 at a time until the number of participants who meet the criteria reached or exceeded the planned number of participants, which was 100[1]. We recruited 201 participants (71 female, mean age = 33.9, SD = 9.8) on Amazon Mechanical Turk (AMT). Participants could earn a bonus of $3 depending on the accuracy of their predictions. Of these, 101 participants were eligible for our analysis. We analyzed the first 100 (39 female, mean age = 34.0, SD = 10.0). This collection plan and all criteria were outlined in our preregistration (here).

Stimuli The stimuli were still images of standing but unstable block towers (targets). Each target tower consisted of 10 blocks, similar to what has been used in previous research (e.g. Battaglia et al., 2013; Hamrick et al., 2016). The blocks came in three colors (red, blue, and green) and in three dimensions (the "cube" in 1x1x1, the "brick" in 1x1x2, and the "plank" in 1x0.5x2; units are relative).

For each target tower, there were four still images of possible resting states, i.e., what the tower might look like once it had collapsed under gravity. One of the resting states was always the real result of the target tower collapsing in the physics engine we used to create the stimuli.[2] The other three were incorrect and impossible in one of the following ways. In "Change Type" questions, one of the blocks was replaced with a block of different dimensions. In "Change Color" questions, one block was switched to a different color. In "Swap Color" questions, two-color towers swapped the colors of all blocks; e.g. all red blocks would become blue and all blue would become red. In "Plus One" questions, an additional block was included. In "Minus One" questions, one block was missing (e.g. Figure 1a). In "Minus Two" questions, two blocks were missing. In "Minus Three" questions, three blocks were missing.

The impossible endstates were created by changing the original tower (e.g. deleting, adding, or changing the properites of one or more blocks), adding some noise (so that all the incorrect answers were not identical), and then allowing the simulation to run to rest. Materials were created until there were three impossible endings that had no blocks that fell outside the viewing area nor were entirely obscured by other blocks.

Procedure Participants read a detailed description of the task. This included several example videos generated from the PhysX materials, and example images like those that appeared in the main body of the task. Participants were asked to watch each video a few times so that they would know how the blocks act when they fall.

The main body of the study consisted of 14 4AFC trials randomly intermixed with 10 easy trials. The easy trials were designed so that the correct answer would be obvious to a participant who was paying attention. Trial order was randomized. When choosing between the four fallen towers the original tower of blocks was still visible on the screen (see

[1] In an earlier preregistration (here), we allowed for a small number of exclusions. However, when we began collecting data for this study we realized that the exclusion rate was much higher than expected. As a result, we stopped data collection and developed a new protocol with a fixed n per experiment *after* exclusions. See Kennedy, Clifford, Burleigh, Waggoner, and Jewell (2018) for discussion of why the exclusion rate may have been unusually high during the summer of 2018, when the majority of these data were collected. In addition, due to space limitations we can report only the key planned analyses in this conference paper.

[2] The PhysX physics engine, through the Unity interface (*Unity*, n.d.).

Figure 1a).

Results

In accordance with our preregistered analysis plan, we pooled the number of correct answers participants gave on the 14 critical items, and used both a two-tailed one-sample *t*-test and the one-sample *"Bayesian Estimation Supersedes the* t-*Test"* or BEST (Kruschke, 2013) to estimate credible intervals for overall performance. The average number of correct answers was 6.33 (SD = 2.37). We calculated a 99% confidence interval of [5.75, 7.00], and the one-sample BEST gave a 99% credible interval of [5.72, 6.98]. Performance at this simple physical reasoning task is thus exceedingly poor; this contrasts sharply with the high performance at predicting whether towers are unstable found by Battaglia et al. (2013)

These errors varied by trial type. The mean number correct (out of 2) were 0.38 for Change Type items, 0.79 for Change Color items, 0.59 for Swap Color items, 0.65 for Plus One items, 1.18 for Minus One items, 1.36 for Minus Two items, and 1.38 for Minus Three items. We calculated confidence intervals corrected for multiple comparisons (Bonferroni with .05/7 = 0.00714) for all items. Intervals for Swap Color and for Plus One were consistent with a null of 0.50. For these item types, participants perform as poorly as if they were given no information at all. The 99.29% interval for the items with the highest accuracy, Minus Three, was [1.19, 1.59], the upper limit being just less than 80% accuracy. Notably, 39 of the 100 participants gave the correct answer to fewer than half of the items. Only 3 participants made no errors at all.

We included a free-response question after all trials, asking participants: "Roughly speaking, how did you try to solve the problems? Please tell us a little about your approach below." Three coders who had not been involved in the design of the study or the collection of data coded the free responses into the following categories: 0) No response, Nonsensical response, or "Other" strategy, 1) Simulation, Visualization, or Imagination, 2) Heuristics or Rules, 3) Both Simulation & Heuristics. To conduct subgroup analyses, we used a best 2 out of 3 approach to resolve disagreements among the coders, and had the three coders manually resolve disagreement for the small number of self-reports where all three coders coded the response differently. The ratings had a Cronbach's alpha of 0.85, indicating acceptable agreement (Kline, 2013).

When participants were asked to describe the way they completed the relevant tasks, 19 gave answers that suggested a simulation or visualization approach, 50 said they used specific rules or heuristics, 20 said that they used both simulation and heuristics, and the remaining 11 gave an uninterpretable answer. Results did not differ between participants who reported using different strategies.

Discussion

Reasoning about sets of 10 simple objects should be well within the abilities of a person using PMST (Battaglia et al., 2013; Hamrick et al., 2016). Despite this, performance was remarkably poor.

This behavioral result seems very unlikely if participants were tracking every block, which in causally-bound systems is a requirement of PMST. While it is possible that simulators might not always keep track of things like color, tracking shape is necessary to predict object interactions, and tracking every object is fundamentally necessary for the task. Because of this requirement, every object will end up in the end states of every simulation. It would seem trivial then to detect a mismatch between the end state of a mental simulation and a provided image of such a final scene. Alternatively, if one retained the spirit of the PMST approach, this result significantly constrains the availability of particular information within a mental simulation. Introducing this new constraint seems hard to reconcile with the ability of people to judge if the tower will fall via simulation because it would imply someone could answer the falling question ("will this block tower fall over?") but not a question about an individual block within a tower (e.g., "will the long red block remain standing when the tower falls over?").

Study 2: Temporal Consistency

PMST conducts simulations in an iterative fashion. At every time-step, the system applies elementary physical rules to each object in the simulation. This is done recursively; once every object has been updated at time t, the system moves on to time t+1, updates all objects again, and so on (Battaglia et al., 2013). This ensures that events will generally occur in the correct order, as long as the approximate trajectory is clear. In this study, we assessed if people have difficulty predicting the order in which events occur, for physical events with reliable trajectories.

The materials for this study consisted of video clips of events in a simple 3-D world. Participants viewed the first two seconds of several short clips of physical scenes in which two independent physical processes unfolded. For example, the physical processes might be two balls, each rolling down its own series of ramps (see Figure 1b), or they might be two lines of dominoes falling over. Each physical process followed a predictable trajectory, and we informed participants of this fact.

In each scene we identified one object in each process (usually "the red ball" and "the blue ball"), and participants were asked to predict which of the two objects would hit the ground first. Participants did not see the outcomes of the video clips, so they had to engage in prospective reasoning in order to make this judgment. The key dependent variable was the proportion of scenes for which participants thought the wrong event would occur first.

Method

Participants As above, our stopping rule was designed to collect a fixed number of participants *after* exclusions. We collected 78 participants (29 female, mean age = 35.1, SD = 9.8) in groups of 9 at a time on Amazon Mechanical Turk. Participants could earn a bonus of $3 depending on the accuracy of their predictions. Of these participants, 63 met our exclusion criteria, and we analyzed only the first 60 partici-

pants (22 female, mean age = 36.4, SD = 10.2), as stated in our preregistration.

Stimuli The main stimuli were video clips (example clip) showing the first two seconds of a scene (full version of same scene). Each scene included two key objects, one red and one blue, each involved in its own causal chain, which would eventually lead to each object colliding with the ground.

Each scene was designed to make the outcome that would occur second seem, at the end of the 2-second clip, more likely to occur first. The object that would actually strike the ground second was moving faster, had gone further, had fewer "obstacles" in its way, etc., or some combination of these factors. We iterated the design of the scenes based on these heuristics until we believed that pausing at the two-second mark would lead to incorrect judgment of the conclusion. PMST predicts that no such items should exist, as long as the trajectories are clear.

In the full scenes, the first object always struck the ground at least 2/3 of a second before the second one did, sometimes much earlier. The full scenes took about 10s to complete.

Procedure Participants read a detailed description of the task which included several example videos of the physics engine we used, and example clips similar to those that appeared in the main body of the survey. Participants were assured that the simulations were designed to be as much like real physics as was possible, that both critical objects would always eventually reach the ground, that there were no hidden objects or forces that would interfere, and that everything relevant to the scene was readily visible in the video clips.

In the main body of the study, participants viewed several video clips of the first two seconds of a physical scene where two independent chains of events unfold. In each case there were two items of interest, one red and the other blue, and participants judged which of the two would reach the ground (indicated by a grass texture) first.

The study presented four questions each of three types ("Marble Run", "Parthenon", and "Domino"), for a total of twelve critical questions. There were also four filler scenes, which were designed to be trivially easy.

Results

We used both a two-tailed one-sample t-test and one-sample BEST to determine if, on average, accuracy was different from chance. Participants answered a mean of 4.77 questions correctly (SD = 2.55), which was less than chance (6), according to both a t-test, $t(59) = -3.75$, $p < .001$, 95% confidence interval [4.11, 5.42] and a one-sample BEST, 95% Credible Interval: [4.11, 5.45].

In answering the 12 critical questions, 56.7% of the participants gave the incorrect answer to more than half of the questions. Every participant made at least two errors. The highest level of performance was ten of twelve correct, achieved by only two participants. Further, 3.3% of the participants gave the wrong answer on all twelve trials.

The same three coders coded free response reports of strategy according to the system described above. The ratings had a Cronbach's alpha of 0.74, indicating acceptable agreement (Kline, 2013). When participants were asked to describe the way they completed the relevant tasks, 8 gave answers that suggested a simulation or visualization approach, 35 said they used specific rules or heuristics, 8 said that they used both simulation and heuristics, and the remaining 9 gave no answer or an uninterpretable answer. Results did not differ between participants who reported using different strategies.

Discussion

In this study, participants saw two processes with predictable trajectories, and were asked to estimate which process would complete first. Overall, participants reversed the order of the events in their predictions, predicting that the event that truly occurred second would occur first, and did so more often than chance. Admittedly, the scenarios used in this study were deliberately designed to be adversarial. If we were to imagine the (hypothetical) space of all possible scenes, it is likely that few cases would prompt the reversals in judgment we observed. However, PMST suggests that *no items* showing such reversals should exist, barring major uncertainties in trajectory, etc. That there exist any items where this kind of reversal is consistently found is evidence that PMST is not the approach being used to make these judgments.

Study 3: Probabilistic Coherence

When making predictions about a physical scene, a key claim of PMST is that judgments reflect probabilistic inference, estimated via repeated stochastic runs of the simulation (Battaglia et al., 2013). As such, people's physical judgments should approximately obey the laws of probability theory.

Conjunction fallacy errors are cases where people rate a joint probability (A & B both occur) as more likely than the marginal probability of one component (e.g. A occurring at all). This is logically contradictory because there is no way for the joint probability to be larger than either of its components. At most, it will be equal to the smaller component.

In the cognitive domain, this is often known as the "Linda Problem", because of a well-known example in which participants judged a hypothetical individual named Linda as more likely to be both a bank teller and a feminist than to be a bank teller in general (Tversky & Kahneman, 1983). To test this commitment of PMST, in this study we assessed if people fall prey to conjunction fallacy-style judgment errors for physical reasoning problems.

Methods

Participants We collected data from 90 participants (28 female, mean age = 33.6, SD = 9.8) on Amazon Mechanical Turk (AMT). Following the criteria outlined in our preregistration, we analyzed only the first 60 participants (18 female, mean age = 34.2, SD = 9.7) of 62 eligible.

Stimuli The main stimuli were video clips, 2/3 of a second long, in which two round objects (a pink "sphere" and a gray "cannonball") interacted in a 2-dimensional world (example video here). This world included gravity and some stationary

objects. There was always "ground" on the bottom edge of the scene, with a green section representing grass on top, and usually one or more boxes resting on the grass. There was always a hole in the ground, wide enough for either object to potentially fall into.

Over the course of each clip, the gray cannonball would travel in a parabola across the screen, while the pink sphere would fall under the influence of gravity (see Figure 1c). The cannonball always traveled toward the sphere, in a way that suggested that the two might collide. Each video ended after approximately 700 ms, well before the cannonball could intersect the pink sphere's path, leaving ambiguity about the outcome of the scene.

Procedure Participants read a detailed description of the task. This included several example simulation videos from the physics engine we used (PhysX) and example clips like those that appeared in the main body of the task. The example videos included many forms of inter-object interactions, including collisions, and participants were allowed to watch these videos as many times as they wanted.

In the main body of the study, participants saw several simple physical scenes. For each scene, participants estimated the likelihood of a particular prompted outcome (e.g., "How likely is it that the pink sphere will end up on the grass?"), as a percentage ranging from 0% to 100% in 1% increments.

Eight of the scenes were considered "critical", and the answers to these provided our primary dependent measure. Unknown to participants, each critical scene appeared twice, for a total of 16 critical trials.

For each scene that appeared twice, in one appearance participants were asked the question, "How likely is it that the pink sphere will end up on the grass?" and in the other, "How likely is it that the cannonball will hit the pink sphere, and then the pink ball will end up on the grass?" Scenes did not repeat until after several filler scenes were presented.

Results

We averaged the difference scores (conjunction probability - sole probability) for each participant for each of the eight critical scenes. Positive values on these difference scores indicate that participants rated a conjunction as more likely than the constituent sole probability, which is a form of the conjunction fallacy. The average rating difference score was 7.29 (SD = 13.07), which was reliably greater than zero, according to both a t-test, $t(59) = 4.32$, $p < .001$, 95% confidence interval [3.92, 10.67], and a one-sample BEST (Kruschke, 2013), 95% Credible Interval: [4.06, 10.79]. This suggests that, on average, participants were inclined to commit the conjunction fallacy in a physics domain.

In rating conjunction and sole probabilities on critical trials, 72% percent of the participants show a bias toward the conjunction event. In addition, 62% percent of subjects committed the conjunction fallacy for more than half of the pairs.

The same three coders coded free response reports of strategy according to the system described above. The ratings had a Cronbach's alpha of 0.75, indicating acceptable agreement (Kline, 2013). When participants were asked to describe the way they completed the relevant tasks, 23 gave answers that suggested a simulation or visualization approach, 17 said they used specific rules or heuristics, 12 said that they used both simulation and heuristics, and the remaining 8 gave no answer or an uninterpretable answer. Somewhat surprisingly, we found that participants actually made *more* extreme conjunction fallacy errors when they reported using a simulation approach, $F(3, 56) = 3.90$, $p = 0.013$.

Discussion

Participants making judgments about outcomes in physical processes routinely predicted that conjunctions were more likely than one of their constituent events. PMST states that judgments about the outcomes of physical processes are made by aggregating over the result of multiple noisy runs of a simulation, and so conjunction fallacy errors contradicts this aspect of the theory.

General Discussion

Simulation has been argued to be an important and effective way in which people reason about the physical world. In this paper we ask about the limits on the use of simulation as a strategy. Across three studies, we found empirical contradictions to the natural predictions made by PMST.

First, when trying to identify the resting state for an unstable tower of 10 blocks, participants have great difficulty distinguishing between the true set of blocks and sets that differ because of changes of color, changes of dimensions, additions, or deletions. PMST suggests that this should not happen without significant additional assumptions about the content and accessibility of particular features of simulated representations.

Second, when judging the order of events in a scene with highly predictable trajectories, participants consistently make incorrect predictions about the order of events. Although the examples were designed to be adversarial, PMST does not admit the existence of such examples because judgments are made using an iterative simulation where every object is advanced synchronously in each unit of time.

Third, participants consistently commit the conjunction fallacy (Tversky & Kahneman, 1983) when reasoning about simple physical scenes, a result that contradicts the claims of PMST about how estimated judgments of physical scenes are made by aggregating across probabilistic samples.

The design of our experiments tried to mimic many of the empirical studies which have supported PMST in complexity and content. Thus we believe they represent an interesting test bed for the generalization of the theory.

As the field tries to grapple with these complex questions, we argue that any complete account of human physical reasoning must contend with both the cases where people appear to do well and the situations where they apparently are limited or deceived. As a result, experiments exposing the limits of simulation can be as informative as those that show the successes.

Acknowledgments The authors thank Ellie Robbins, Michael Lepori, Xuechen Sheryl Zhang, and Adi Kwiatek for help with this research and Gregory L. Murphy for helpful comments.

References

Bates, C. J., Yildirim, I., Tenenbaum, J. B., & Battaglia, P. W. (2015). Humans predict dynamics using probabilistic simulation. In *Proceedings of the 37th annual conference of the cognitive science society.*

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013, November). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, *110*(45), 18327–18332.

Davis, E., & Marcus, G. (2015, June). The scope and limits of simulation in cognitive models. *arXiv preprint arXiv:1506.04956*.

Davis, E., & Marcus, G. (2016, April). The scope and limits of simulation in automated reasoning. *Artificial Intelligence*, *233*, 60–72.

Hamrick, J. B., Battaglia, P. W., Griffiths, T. L., & Tenenbaum, J. B. (2016, December). Inferring mass in complex scenes by mental simulation. *Cognition*, *157*, 61–76.

Hamrick, J. B., Smith, K. A., Griffiths, T. L., & Vul, E. (2015). Think again? The amount of mental simulation tracks uncertainty in the outcome. In *Proceedings of the 37th annual conference of the cognitive science society.*

Hegarty, M. (2004, June). Mechanical reasoning by mental simulation. *Trends in Cognitive Sciences*, *8*(6), 280–285.

Kennedy, R., Clifford, S., Burleigh, T., Waggoner, P., & Jewell, R. (2018). The shape of and solutions to the mturk quality crisis. *Unpublished manuscript*.

Kline, P. (2013). *Handbook of psychological testing*. Routledge.

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, *142*(2), 573–603.

Kubricht, J. R., Holyoak, K. J., & Lu, H. (2017, October). Intuitive Physics: Current Research and Controversies. *Trends in Cognitive Sciences*, *21*(10), 749–759.

Kubricht, J. R., Jiang, C., Zhu, Y., Zhu, S.-C., Terzopoulos, D., & Lu, H. (2016). Probabilistic simulation predicts human performance on viscous fluid-pouring problem. In *Proceedings of the 38th annual conference of the cognitive science society* (pp. 1805–1810).

McCloskey, M., Caramazza, A., & Green, B. (1980). Curvilinear motion in the absence of external forces: Nave beliefs about the motion of objects. *Science*, *210*(4474), 1138–1141.

Proffitt, D. R., Kaiser, M. K., & Whelan, S. M. (1990, July). Understanding wheel dynamics. *Cognitive Psychology*, *22*(3), 342–373.

Siegler, R. S. (1976). Three aspects of cognitive development. *Cognitive psychology*, *8*(4), 481–520.

Smith, K. A., Dechter, E., Tenenbaum, J. B., & Vul, E. (2013). Physical predictions over time. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 35).

Smith, K. A., & Vul, E. (2017). Thinking inside the box: Motion prediction in contained spaces uses simulation. In *Proceedings of the 39th annual meeting of the cognitive science society.*

Tversky, A., & Kahneman, D. (1983, October). Extensional Versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment. *Psychological Review*, *90*(4), 23.

Ullman, T. D., Spelke, E., Battaglia, P., & Tenenbaum, J. B. (2017, September). Mind Games: Game Engines as an Architecture for Intuitive Physics. *Trends in Cognitive Sciences*, *21*(9), 649–665.

Unity. (n.d.). Retrieved from `https://unity3d.com`

Cognitive Aging Effects on Language Use in Real-Life Contexts: A Naturalistic Observation Study

Minxia Luo[a,b], (m.luo@psychologie.uzh.ch)
Gerold Schneider[c,d] (gschneid@cl.uzh.ch)
Mike Martin[a,b] (m.martin@psychologie.uzh.ch)
Burcu Demiray[a,b] (b.demiray@psychologie.uzh.ch)

[a] Department of Psychology, University of Zurich
[b] University Research Priority Program "Dynamics of Healthy Aging", University of Zurich
[c] English Department, University of Zurich
[d] Institute of Computational Linguistics, University of Zurich

Abstract

This study examined age effects on real-life language use and within-person variations in language use across social contexts. We used the Electronically Activated Recorder (i.e., a portable audio recorder that periodically records sound snippets) to collect over 31,300 snippets (30 seconds long) from 61 young and 48 healthy older adults in Switzerland across four days. We examined vocabulary richness and grammatical complexity across the social contexts of (a) activities (i.e., socializing, working); and (b) conversation types (i.e., small talk, substantive conversation). Multilevel models showed that vocabulary richness and grammatical complexity increased during socializing and substantive conversations, but decreased in small talk. Moreover, young adults produced shorter clauses at work than not at work. Furthermore, compared with young adults, older adults used richer vocabulary and more complex grammatical structures at work; and used richer vocabulary in small talk. In contrast, young adults used richer vocabulary than older adults during non-socializing and non-working occasions, such as watching TV and exercising. Results are discussed in the context of cognitive aging research with a novel emphasis on context.

Keywords: vocabulary richness; grammatical complexity; social context; cognitive behavior; electronically activated recorder (EAR); naturalistic observation method

Introduction

Real-life language use is mostly embedded in social interactions and conversations (e.g., Clark, 1996). While effects of social context on language use have been widely acknowledged in sociolinguistics, linguistic ethnography, and social psychology (e.g., Finkbeiner, Meibauer, & Schumacher, 2012), they have been underrepresented in cognitive aging research (e.g., Horton, Spieler, & Shriberg, 2010). Rooted in laboratory experiments, cognitive aging research assumed that cognitive change with aging was the primary determinant of variations in language use (Burke & Shafto, 2008). However, unlike in the laboratory, where the upper limits of one's abilities are tested (Baltes, Dittmann-Kohli, & Dixon, 1984), in real life, contexts should also play a role in influencing behaviors (Lewin, 1951). Although some cognitive aging studies have controlled for the effects of social context in their examination of age and real-life language use, they have not treated social context as an essential determinant in their theoretical frameworks (Meylan & Gahl, 2014; Moscoso del Prado Martín, 2016). Furthermore, past studies, focusing on comparisons of different speakers in different contexts (i.e., between-person differences), were limited in inferring how the same individuals varied their language across contexts (i.e., within-person variations; Hamaker, 2012). Moreover, many real-life speech samples in the literature have been collected via telephone conversations between strangers, which may not be representative of naturally occurring language use. In sum, only one recent study has combined cognitive aging effects with within-person variations across social contexts in the investigation of language use in real life (Luo, Robbins, Martin, & Demiray, under review).

The current study used a naturalistic observation method to collect speech samples in real life and examined age effects in language use across different social contexts. Using the Electronically Activated Recorder (EAR; Mehl, Pennebaker, Crow, Dabbs, & Price, 2001), a digital recorder which periodically and unobtrusively captures ambient sounds in natural environments, we assessed language use and social contexts by examining speakers' moment-to-moment conversations. Vocabulary richness and grammatical complexity are related to cognitive changes with age (e.g., Horton, et al., 2010). We examined vocabulary richness and grammatical complexity across two types of social contexts that have been shown relevant to language use: (a) activities (i.e., socializing, working); and (b) conversation types (i.e., small talk, substantive conversations; Levinson, 1992). Our goals were to examine (1) whether individuals changed their language across real-life social contexts; and (2) whether age effects on language use differed across social contexts. Thus, this study is the first to examine cognitive aging effects on real-life language use in relation to within-person variations across different activities and conversation types.

Cognitive Aging Effects in Language Use

The differences in language use between young and older adults have been associated with cognitive changes with age. For example, the observations of older adults using richer

vocabulary than young adults have been explained as due to lifelong vocabulary accumulation in old age (e.g., Horton, et al., 2010). Moreover, the findings of older adults producing simpler grammatical structures than young adults have been interpreted as due to decreasing working memory in old age (e.g., Cheung & Kemper, 1992). Majority of these findings came from laboratory tasks, which asked participants to describe a novel picture, an important person, or a historical event (e.g., Cheung & Kemper, 1992). These studies assumes that cognition was the primary determinant of vocabulary richness and grammatical complexity and that participants' speech reflected their cognitive abilities in a controlled and consistent environment.

In theory, behavior is conceptualized as the interactions between personal characteristics and different supporting or impeding contexts (e.g., WHO, 2015; Verhaeghen, Martin, & Sędek, 2012). That is, in real life, where the environment is more diverse than in the laboratory, contextual effects should be taken into account. In order to improve the generalizability of their findings, some researchers examined speech outside of the laboratory, such as in telephone conversations (e.g., Horton, et al., 2010). These studies examined age effects on language use and controlled for contextual factors (e.g., talking with different conversional partners; Meylan & Gahl, 2014; Moscoso del Prado Martín, 2016). However, these studies examined between-person differences, instead of within-person variations across contexts (Hamaker, 2012). Additionally, the telephone conversations between strangers may not represent naturally occurring conversations.

In sum, some studies have identified effects of social context on language use, but they have not considered social context as an essential determinant of language use in their theoretical frameworks. Additionally, past studies have not examined contextual effects on vocabulary richness and grammatical complexity in naturally occurring language use with a within-person research design. Amid the growing interest in examining age effects on language use in real life, it is necessary to understand contextual effects on language use with data that properly capture within-person variations in language use in naturally occurring conversations.

Contextual Effects in Language Use

Social context is an important construct in the theoretical frameworks of language use in social psychology, sociolinguistics and linguistic ethnography (e.g., Clark, 1996; Finkbeiner, et al., 2012). There are substantial variations in language use across different social contexts, such as types of activities (i.e., socializing, working; Levinson, 1992). For example, speakers use more swearing words in leisure activities than at work (Cameron, 1969). Speakers refer to themselves more often in socializing and entertaining activities than while working (Mehl & Pennebaker, 2003). Furthermore, types of conversations (e.g., small talk, substantive conversation) also have effects on language use. Conversation topics and discourse markers (e.g., "anyway" and "you know") are different in small talk versus formal conversations, and the differences influence the degree of trust among speakers (Bickmore & Cassell, 2001). In addition, how speakers engage in small talk and substantive conversations is associated with their well-being (Mehl, Vazire, Holleran, & Clark, 2010).

Past studies have shown that the contexts of activity types and conversation types have effects on language use. However, majority of these studies have explained effects of social context from the perspective of social role and social identity and have not linked their findings to cognitive effects (e.g., Mehl & Pennebaker, 2003). In fact, cognitive-biological and socio-cultural determinants of language use are intertwined and inseparable (e.g., Gerstenberg, & Voeste, 2015). Furthermore, variations in language use across social contexts are likely to differ between young and older adults (e.g., Adams, Smith, Pasupathi, & Vitolo, 2002).

In sum, research that identifies effects of contexts on language use has highlighted the importance of understanding variations in language use across contexts. Thus, it is important to consider cognitive and contextual effects in the examination of real-life language use.

The Current Study

This study used the EAR to periodically and unobtrusively capture ambient sounds and speech in real life. The intensive and repeated sampling approach of the EAR captures multiple observations from each participant and, thus, allows us to analyze within-person variations in language use across social contexts. We treated social contexts and age as two important concepts in our theoretical model and inspected their joint effects on real-life language use.

The first goal of our study was to examine contextual effects on real-life language use. We focused on vocabulary richness and grammatical complexity that are associated with cognitive aging. We examined the contexts that have been found to have effects on language use: (a) activities (i.e., socializing, working); and (b) conversation types (i.e., small talk, substantive conversation). If activities and conversation types had effects on language use, we considered it in line with our assertion that contextual factors should be examined in the understanding of real-life language use. However, as there was a lack of evidence on how these social contexts would influence vocabulary richness and grammatical complexity, we refrained from forming hypotheses about the directionality of contextual effects. The second goal of our study was to explore whether age effects on real-life language use varied across different social contexts. If age effects on language use differed across different contexts, we considered it offered support for our anticipation that age effects on language use would be influenced by contexts.

Method

Participants

Our sample included over 31,300 sound files collected from 48 healthy older adults (62-83 years, M = 70.5, SD = 4.7; 22 men, 26 women) and 61 young adults (19-31 years, M = 23.0, SD = 3.10; 24 men, 37 women). Participants were recruited

via the participant pool of our department, via flyers in university buildings and advertisements in a local newspaper, and through snowball sampling used by a research assistant. All participants were local residents and spoke Swiss German.

Older participants were healthy with no record of neurological or psychiatric illness and lived independently. Their years of education ranged from seven to 25 (M = 10.55, SD = 3.02). Five of them were working part-time or full-time. They were compensated with 50 Swiss Francs. Young participants were mostly university students, whose years of education ranged between three and 17 years (M = 12.35, SD = 2.41). Eight of them had a part-time or full-time job. They could choose between 50 Swiss Francs and research credits for compensation.

Procedure

The study included an introduction session, a four-day EAR observation period, and a feedback session. In the introduction session, participants were given instructions on the study. They were asked to sign an informed consent form and to complete questionnaires including demographic and psychological measures. Next, participants received an iPhone with the EAR application installed. Participants were informed that the EAR would randomly record 30 seconds of ambient sounds. They were told that they would not notice when the EAR was recording, so that they could continue their normal lives. They were informed that they would have the opportunity to review and delete any sound files at the end of the study, before anyone listened to them.

After the introduction session, participants carried the EAR with them for four consecutive days. Additionally, they kept a diary every evening about their hour-by-hour activities of that day. Finally, participants met with the researchers again for a feedback session, in which they returned the phone and completed further questionnaires. They evaluated their experience with carrying the phone. They were given a password-protected CD containing all of their sound files to review. All procedures were approved by the local ethics committee.

EAR We provided each participant with an iPhone 4S, where the EAR application was installed (version 2.3.0). We programmed the EAR to record 30-second sound files at random times throughout the day. It was set to record 72 sound files per day (a total of 288 sound files per participant). We set a blackout period between midnight and 6 AM, when the EAR was inactive. We turned on the "Airplane mode" of the iPhone and locked it with a screen-lock password. Thus, participants could not access the EAR settings or use the phone for other purposes. We set a reminder in the phone calendar to automatically beep every evening at 9 PM to remind the participants to charge the iPhone overnight.

Linguistic Measures

All utterances of the participants captured by the EAR were transcribed. A research assistant created the transcripts, which were then checked and corrected by a second research assistant. Swiss-German dialect was translated word-by-word into standard written German and then transcribed. The utterances of interlocutors or bystanders were not transcribed due to ethical reasons.

We used the the TreeTagger (Schmid, 1999) via the R package of "koRpus" version 0.10-2 (Michalke, 2018) to process the transcripts. First, we identified each word according to its grammatical class (e.g., a noun, a verb), a process called *part-of-speech tagging*. We also turned each word to its lemma form, a process called *lemmatization*. For example, we turned *isst* ("eats"), *aß* ("ate"), and *gegessen* ("eaten") to the lemma form of *essen* ("eat"). Subsequently, we calculated the following two linguistic measures.

Vocabulary Richness: Entropy. Vocabulary richness was calculated with Shannon entropy measure, representing the diversity of words (e.g., Moscoso del Prado Martín, 2016). We calculated the frequency of occurrence of each word based on its lemma form and part-of-speech tag. Afterwards, we calculated the Shannon entropy of each sound file using the frequency. We used the R package of "entropy" (version 1.2.1; Hausser & Strimmer, 2018) to calculate Shannon entropy and corrected the results with Chao-Shen estimator, according to Moscoso del Prado Martín (2016). Higher scores of entropy indicate higher usage of unique words.

Grammatical Complexity: Clause Length. Clause length is the word count in a clause, representing the complexity of grammatical structures (e.g., Horton, et al., 2010). We used the German Pro3Gres parser (Sennrich, Schneider, Volk, Warin, 2009) to identify the following patterns as clauses: (a) a root element, i.e. the top element of a sentence, typically the inflected verb; (b) a relative clause (which is attached with the label rel to the NP it modifies); (c) a subordinated adjunct clause (label neb); (d) a subordinated complement clause (label objc or subjc); and (e) coordination at clause level (kon); (f) a fragmented or complete sentence (label s; Foth, 2005). Finally, we calculated word count per clause in each sound file.

EAR Coding

Every sound file has been manually coded for the participant's momentary (a) activity (i.e., socializing, working); and (2) conversation types (i.e., small talk, substantive conversations). More specifically, *socializing* refers to when the participant is doing something to socialize or entertain with others. *Working* refers to doing paid work. *Small talk* refers to any conversation that is completely non-instrumental, with no (or very trivial) information being exchanged. Finally, *substantive conversation* is any conversation that serves the purpose of exchanging information and ideas about a topic, e.g., news, politics.

All coding categories were dichotomous, indicating the presence (1) or absence (0) of the targeted item within a sound file. Trained coders coded these categories by listening to the pitch of the participants' voice, ambient sounds, and conversation topics in each sound file, and by referring to the

adjacent sound files. The coders also verified their coding with the participants' diaries. Note that the coders were not aware that vocabulary richness and grammatical complexity would be analyzed. Thus, the coders coded for activity and conversation types without referring to these linguistic measures. Social contexts were coded by only one research assistant, because the reliability of the coding of these contextual variables is found to be high in past EAR studies (e.g., Mehl, & Pennebaker, 2003; Mehl, et al., 2001).

Results

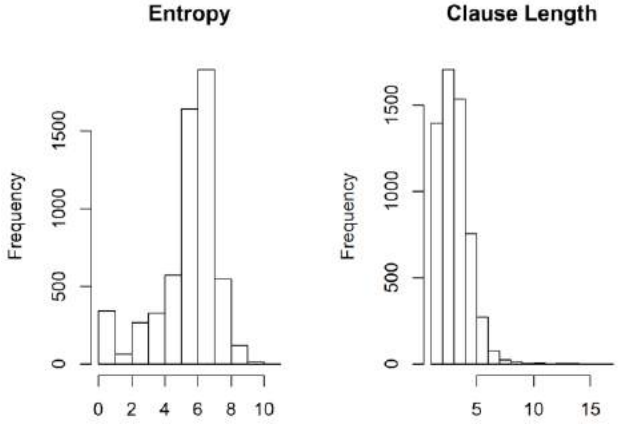
Preliminary Analyses

We collected over 31,300 sound files. For privacy reasons, 15 participants deleted 133 sound files, ranging from 1 to 40 sound files per person. From the remaining sound files, 6,542 included participant speech, ranging from 2 to 158 per participant (*M* = 60.02, *SD* = 32.09). That is, participants were talking, on average, in 21% of the sound files.

Young and older participants reported that the EAR did not affect their daily activities or way of speaking, in line with past EAR studies (e.g., Mehl, et al., 2001). Additionally, the proportion of the sound files in which the participants mentioned the EAR was low (only 0.8% of all sound files that included speech).

Out of the 6,542 sound files, 778 were deleted, as the participants' speech was unclear or included another language than German. This resulted in a final sample of 5,764 sound files. There were over 140,000 spoken words. The average score of entropy was 5.36 (*SD* = 1.9, Range: 0.00-10.24), and the average length of clauses was 3.09 words (*SD* = 1.39, Range: 1-17). The word count in sound files ranged from 1 to 123 words (*M* = 24.34, *SD* = 21.36). Figure 1 shows the histograms of entropy and clause length.

Figure 1. Histograms of Linguistic Measures.



Averaging across participants, in young adults, 7% of sound files (*SD* = 6%, Range: 0-23%) have been coded as including socializing, 2% (*SD* = 5%, Range: 0-19%) included working, 1% (*SD* = 1%, Range: 0-5%) included small talk, and 12% (*SD* = 7%, Range: 0-33%) included substantive conversation. In older adults, 6% of sound files (*SD* = 6%, Range: 0-18%) included socializing, 1% (*SD* = 4%, Range: 0-27%) included working, 2% (*SD* = 1%, Range: 0-7%) included small talk, and 12% (*SD* = 9%, Range: 0-38%) included substantive conversation.

Analytical Approach

The sound files (level 1) are nested within individuals (level 2). We analyzed these hierarchical data with multilevel models, which simultaneously examine between-persons and within-person variances (Bolger & Laurenceau, 2013). We estimated separate models for the two linguistic measures and for the different social contexts. In each model, we first estimated effects of age group and social context, and then added Age Group × Social Context interactions. More specifically, the full model is specified as follows:

Level 1: $\text{Language}_{ti} = \beta_{0i} + \beta_{1i}(\text{Context}_{ti} - \text{Context}_i) + e_{ti}$

Level 2: $\beta_{0i} = \gamma_{00} + \gamma_{01}(\text{AgeGroup}_i) + U_{0i}$

$\beta_{1i} = \gamma_{10} + \gamma_{11}(\text{AgeGroup}_i) + U_{1i}$

where *i* indexes individuals and *t* indexes sound files. At level 1, Language_{ti} represents the linguistic variable. β_{0i} is the random intercept, and β_{1i} represents within-person effects of contexts. The contextual variables were coded such that a non-event served as the reference group (i.e., socializing versus non-socializing, working versus non-working, small talk versus non-small talk, substantive conversation versus non-substantive conversation). This contrast scheme was used in line with the dichotomous nature of the contextual variables (coded as 0 vs. 1). e_{ti} represents the unexplained within-person context-to-context differences in language use. At level 2, β_{0i} represents the intercept of each age group and is modelled in detail through the level-1 model. β_{1i} is the slope of each age group. γ_{00} represents the grand mean of outcomes over all of the participants. γ_{10} represents the grand mean of slopes over all of the participants. γ_{01} and γ_{11} represented effects of age group, where young adults were the reference group. U_{0i} represents the random intercepts of individuals. U_{1i} represents the random slopes of individuals.

We decomposed each dummy-coded contextual variable into between-persons variance and within-person variance (Bolger & Laurenceau, 2013). More specifically, we firstly calculated the average score of context of each participant (Context_i). Afterwards, we deducted the score of context in each sound file from the mean score of context of each participant (Context_{ti} - Context_i; i.e., within-person contextual effect). The within-person contextual variables were our contextual predictors. Finally, we controlled for sex and education in each model.

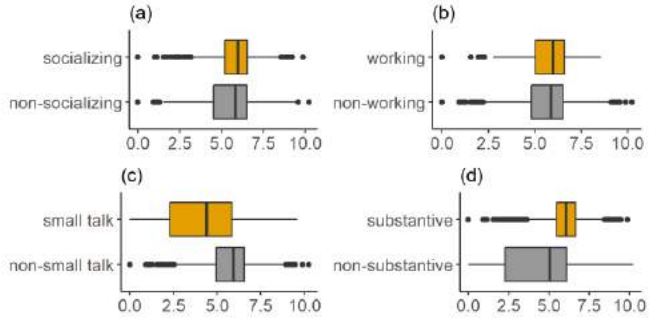
We used the R package "lme4" (version 1.1-17) in R (version 3.5.2) to estimate the models and the 95% confidence intervals (CI). We estimated the models with full information maximum likelihood estimation method, which treated incomplete data as missing at random and adjusted for unbalanced data (Singer & Willett, 2003). We additionally calculated p-values with R package "lmerTest" (version 3.0-1) and considered $p < .05$ as significant.

Major Analyses

Our first research goal was to examine contextual effects on language use. Thus, we estimated models with effects of age group and social context. We, then, added Age Group × Social Context interaction to the model for the second research goal: exploring whether age effects on language use were influenced by different social contexts. Due to their non-significant effects, we dropped sex and education from our final models. Additionally, we dropped the random slope effects from the models of socialization, working, and small talk in vocabulary richness, because the random intercept and slope models did not fit better than the random intercept models.

Vocabulary Richness: Entropy In the model of socialization, as shown in Figure 2 (a), participants used richer vocabulary while socializing than non-socializing (M = 0.32, p = <.001, 95% CI [0.20, 0.44]).[1] As shown in Figure 3 (a), young adults used richer vocabulary than older adults during non-socializing (M = -0.23, p =.014, 95% CI [-0.41, -0.05]). However, there was no age group difference during socializing (M = 0.17, p = .155, 95% CI [-0.07, 0.41]).

Figure 2. Entropy by Contexts

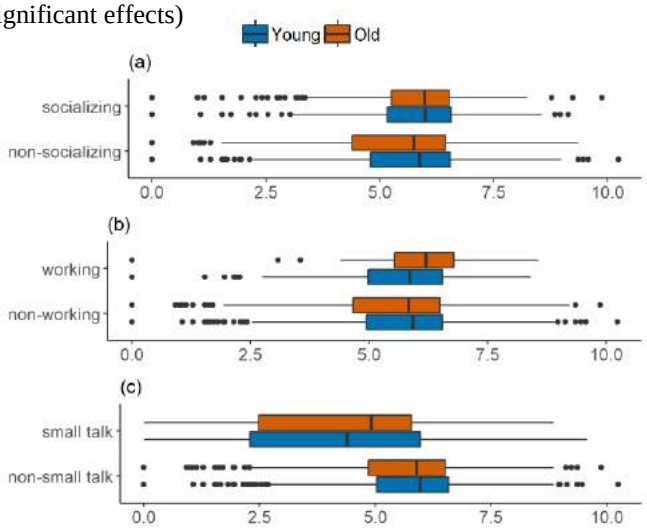


In the model of working, as displayed in Figure 2 (b), there was no significant difference in vocabulary richness between working and non-working occasions (M = 0.17, p = .291, 95% CI [-0.14, 0.47]). As presented in Figure 3 (b), young adults used richer vocabulary than older adults in non-working occasions (M = -0.21, p = .026, 95% CI [-0.38, -0.03]). In contrast, older adults used richer vocabulary than young adults at work (M = 0.96, p = .030, 95% CI [0.09, 1.82]).

Figure 2 (c) shows that participants used richer vocabulary in non-small talk than in small talk (M = -1.31, p < .001, 95% CI [-1.51, -1.11]). Figure 3 (c) shows that there was no significant age group difference in non-small talk (M = -0.18, p = .050, 95% CI [-0.36, 0.00]). However, in small talk, older adults used richer vocabulary than young adults (M = 0.42, p = .041, 95% CI [0.02, 0.82]).

Figure 2 (d) shows that participants used richer vocabulary in substantive conversations than in non-substantive conversations (M = 1.57, p < .001, 95% CI [1.47, 1.67]). There was no significant age group difference in non-substantive conversations (M = -0.14, p = .376, 95% CI [-0.44, 0.17]) or in substantive conversations (M = 0.15, p = .348, 95% CI [-0.17, 0.47]).

Figure 3. Entropy Across Age Groups and Contexts (significant effects)



Grammatical Complexity: Clause Length In the model of socializing (Figure 4 [a]), participants uttered longer clauses while socializing than non-socializing (M = 0.18, p = <.001, 95% CI [0.09, 0.27]). There was no age group difference in non-socializing (M = -0.13, p = .086, 95% CI [-0.27, 0.02]) or socializing occasions (M = 0.11, p = .234, 95% CI [-0.07, 0.28]).

In the model of working (Figure 4 [b]), there was no significant difference in grammatical complexity between working and non-working occasions when examining both older and young adults (M = -0.18, p = .112, 95% CI [-0.41, 0.04]). However, young adults produced shorter clauses at work than not at work (M = -0.31, p = .013, 95% CI [-0.55, -0.07]). As shown in Figure 5, age group difference was non-significant in non-working occasions (M = -0.12, p = .101, 95% CI [-0.27, 0.02]), but was significant at work (M = 0.86, p = .008, 95% CI [0.23, 1.49]). That is, older adults used longer clauses than young adults at work.

Figure 4 (c) shows that participants produced shorter clauses during small talk than in non-small talk (M = -0.60, p < .001, 95% CI [-0.74, -0.45]). There was no age group difference in non-small talk (M = -0.10, p = .170, 95% CI [-0.24, 0.04]) or in small talk (M = 0.01, p = .954, 95% CI [-0.47, 0.50]).

As depicted in Figure 4 (d), participants produced longer clauses than in non-substantive conversations (M = 0.77, p < .001, 95% CI [0.70, 0.85]). There was no age group difference in non-substantive conversations (M = -0.05, p

[1] While our analyses focused on within-person variations in each participant, for simplicity, the figures show within-person variations across all participants.

= .607, 95% CI [-0.24, 0.14]) or in substantive conversations (*M* = 0.00, *p* = .966, 95% CI [-0.21, 0.20]).

Figure 4. Clause Length by Contexts

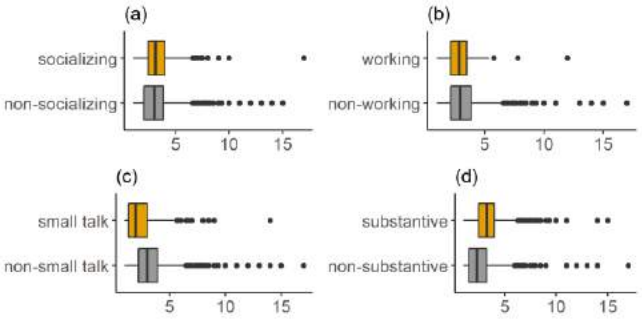
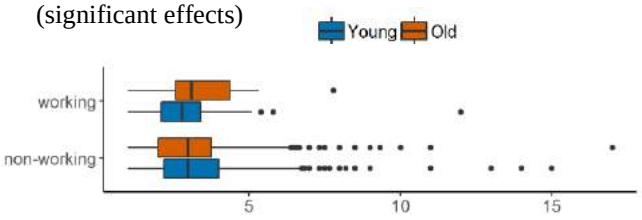


Figure 5. Clause Length Across Age Groups at Work (significant effects)



Discussion

Using a naturalistic observation method, we examined age group differences in language use across social contexts in real life. We found that for both young and older adults, vocabulary richness and grammatical complexity increased while socializing and during substantive conversations. These findings indicate that participants activated richer vocabulary and produced more complex grammar to communicate information in these social contexts. Moreover, for both young and older adults, vocabulary richness and grammatical complexity decreased during small talk. These findings suggest that small talk includes routine and probably repetitive information. Furthermore, young adults produced shorter clauses at work than not at work. Young adults may have been inexperienced at the workplace and thus grammatical complexity differed at work versus not.

Additionally, older adults used richer vocabulary and more complex grammatical structures than young adults at work; they also uttered richer vocabulary in small talk. Older adults may be more inclined to use formal language than young adults in professional settings or in small talks, e.g., greeting the others. In contrast, we found that young adults used richer vocabulary than older adults during non-socializing and non-working occasions, such as doing housework, watching TV, exercising, or commuting in a bus.

Although vocabulary richness and grammatical complexity have been shown to be associated with cognitive abilities in past cognitive aging studies (e.g., Cheung & Kemper, 1992), our findings indicate that age effects can vary depending on the contexts in real life. In other words, unlike in laboratory studies that are designed to test the upper limits of cognitive abilities (Baltes, et al., 1984), in real life, variations in language use are likely to be associated with not only age, but also social contexts.

In cognitive aging and gerontology research, behavior is conceptualized as determined by the interactions between personal characteristics and contexts (e.g., WHO, 2015; Verhaeghen, et al., 2012). Our findings offer evidence for the effects of context on vocabulary richness and grammatical complexity, in addition to age. This perspective is particularly useful when there is a growing interest in collecting "big data" and understanding cognitive behaviors in real life (e.g., Demiray, Mischler, & Martin, 2017; Demiray, Mehl & Martin, 2018; Luo, et al. under review).

Limitations and Future Work

Despite the novel approach that we contributed to the literature, this study has limitations. First, the small number of observations for working and small talk could have influenced statistical estimations. Although multilevel models adjusted for unbalanced data, it is still worthy to prolong the data collection period in future research to obtain more observations. Second, even though the models' fit seemed passable (i.e., the residuals of the models' estimation looked normal), the distributions of the linguistic measures were not bell-shape normal. Limited by the capacity of the lme4 package, we treated these variables as normal distributions. Future studies could use other estimation approaches, e.g., Bayesian method to estimate the linguistic measures. Third, we observed that language use varied across different social contexts and offered speculative explanation for different contextual effects. Future studies should try to incorporate momentary self-reports from participants to understand the subjective perceptions of participants during language use across different contexts. Fourth, this study included only young and old age groups. Future studies should include middle-aged adults to understand language use across the whole adult lifespan.

Conclusion

We contributed to the literature by using a novel approach to unobtrusively collect thousands of sound files in natural environments and by examining age effects on language use with a focus on context. We found that (1) social contexts had effects on language use; and (2) age effects on language use varied across social contexts. Our findings showed that both personal (i.e., age) and contextual factors (i.e., social contexts) are important determinants in the understanding of real-life language use. We offer a new perspective for understanding age effects on real-life language use, or more generally real-life behavior, in the context of cognitive changes with age.

Acknowledgments

This research was supported in part by Hedwig Widmer Stiftung awarded to Minxia Luo, and Velux Stiftung (Grant 917) awarded to Mike Martin. The authors declared no conflict of interest.

References

Adams, C., Smith, M. C., Pasupathi, M., & Vitolo, L. (2002). Social context effects on story recall in older and younger women: Does the listener make a difference? *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *57*(1), 28-40.

Baltes, P. B., Dittmann-Kohli, F., & Dixon, R. A. (1984). New perspectives on the development of intelligence in adulthood: Toward a dual-process conception and a model of selective optimization with compensation. In Baltes, P. B. & Brim, O. G. Jr. (Eds.), *Life-span development and behavior* (Vol. 6, pp. 33-76). Orlando, Florida: Academic Press, Inc.

Bickmore, T., & Cassell, J. (2001, March). Relational agents: a model and implementation of building user trust. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 396-403). ACM.

Bolger, N., & Laurenceau, J. P. (2013). *Intensive longitudinal methods*. New York, NY: The Guilford Press.

Burke, D. M., & Shafto, M. A. (2008). Language and aging. In Craik, F. I. M. & Salthouse, T. A. (Eds.), *The handbook of aging and cognition* (3th ed.) (pp. 373-443). New York, NY: Psychology Press.

Cameron, P. (1969). Frequency and kinds of words in various social settings, or what the hell's going on?. *Pacific Sociological Review*, *12*(2), 101-104.

Cheung, H., & Kemper, S. (1992). Competing complexity metrics and adults' production of complex sentences. *Applied Psycholinguistics*, *13*(1), 53-76.

Clark, H. H. (1996). *Using language*. New York, NY: Cambridge University Press.

Demiray, B., Mehl, M. R., & Martin, M. (2018). Conversational Time Travel: Evidence of a Retrospective Bias in Real Life Conversations. *Frontiers in Psychology*, *9*. Article 2160.

Demiray, B., Mischler, M., & Martin, M. (2017). Reminiscence in everyday conversations: a naturalistic observation study of older adults. *The Journals of Gerontology: Series B*, *00*(00), 1–11.

Finkbeiner, R., Meibauer, J., & Schumacher, P. B. (Eds.). (2012). *What is a context?: linguistic approaches and challenges* (Vol. 196). John Benjamins Publishing.

Foth, K.A. (2005). *Eine umfassende Contraint-Dependenz-Grammatik des Deutschen*. University of Hamburg.

Gerstenberg, A., & Voeste, A. (Eds.). (2015). *Language development: The lifespan perspective* (Vol. 37). Amsterdam: John Benjamins Publishing Company.

Hamaker, E. L. (2012). Why researchers should think "within-person": A paradigmatic rationale. In Mehl, M. R. & Conner, T.S. (Eds.), *Handbook of research methods for studying daily life* (pp. 43-61). New York, NY: The Guilford Press.

Hausser, J., & Strimmer, K. (2018). Estimation of Entropy, Mutual Information and Related Quantities. R package version 1.2.1. https://cran.r-project.org/web/packages/entropy/entropy.pdf.

Horton, W. S., Spieler, D. H., & Shriberg, E. (2010). A corpus analysis of patterns of age-related change in conversational speech. *Psychology and Aging*, *25*(3), 708-713.

Levinson, S. C. (1992). Activity types and language. In *Talk at work: Interaction in institutional settings* (pp. 66-100). Cambridge University Press.

Lewin, K. (1951). *Field theory in social science*. New York, NY: Harper & Brothers.

Luo, M., Demiray, B., Robbins, M. L., & Martin, M. (under review) *Real-life Language Use Across Different Interlocutors: A Naturalistic Observation Study of Adults Varying in Age*.

Mehl, M. R., Pennebaker, J. W., Crow, D. M., Dabbs, J., & Price, J. H. (2001). The Electronically Activated Recorder (EAR): A device for sampling naturalistic daily activities and conversations. *Behavior Research Methods*, *33*(4), 517-523.

Mehl, M. R., & Pennebaker, J. W. (2003). The sounds of social life: A psychometric analysis of students' daily social environments and natural conversations. *Journal of Personality and Social Psychology*, *84*(4), 857-870.

Mehl, M. R., Vazire, S., Holleran, S. E., & Clark, C. S. (2010). Eavesdropping on happiness: Well-being is related to having less small talk and more substantive conversations. *Psychological science*, *21*(4), 539-541.

Meylan, S., & Gahl, S. (2014, January). The divergent lexicon: Lexical overlap decreases with age in a large corpus of conversational speech. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 36, No. 36).

Michalke, M. (2018, March). *"Entschuldigen Sie, dass ich Ihnen einen komplizierten Artikel schreibe, für einen lesbaren habe ich keine Zeit" -- Textanalyse mit den R-Paketen koRpus & tm.plugin.koRpus*. Paper at the Tagung experimentell arbeitender Psychologen (TeaP), Marburg.

Moscoso del Prado Martín, F. (2016). Vocabulary, grammar, sex, and aging. *Cognitive Science*, *41*(4), 950-975.

Schmid, H. (1999). Improvements in part-of-speech tagging with an application to German. In *Natural language processing using very large corpora* (pp. 13-25). Springer, Dordrecht.

Sennrich, R., Schneider, G., Volk, M., & Warin, M. (2009). "A New Hybrid Dependency Parser for German". *From Form to Meaning: Processing Texts Automatically*. *Proceedings of the Biennial GSCL Conference 2009*, 115-124.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford university press.

Verhaeghen, P., Martin, M., & Sędek, G. (2012). Reconnecting cognition in the lab and cognition in real life: The role of compensatory social and motivational factors in explaining how cognition ages in the wild. *Aging, Neuropsychology, and Cognition*, *19*(1-2), 1-12.

World Health Organization. (2015). *World report on ageing and health*. Geneva: WHO Press.

Role of Working Memory on Strategy Use in the Probability Learning Task

Mahi Luthra (mkluthra@iu.edu), Peter M. Todd (pmtodd@indiana.edu)
Psychological and Brain Sciences Department and Cognitive Science Program, Indiana University
1101 E. 10th Street, Bloomington, IN 47408 USA

Abstract

Extensive research on probability learning has reported on the ubiquity of the probability matching strategy—choosing options in proportion to their probability of being correct. The current paper explores why the optimal strategy in this task (always choosing the higher probability option) is not intuitive for participants, by examining their decisions in relation to their working memory capacities. We hypothesize that probability matching is a by-product of an automatic recency-based strategy produced by limits in working memory storage and that deliberate strategizing mediated by working memory processing can override recency in favor of optimal responding. A variant of the Expectancy-Valence Learning Model is fit to participant data from a two-choice probability learning task using hierarchical Bayesian modelling. Point estimates of the best-fitting parameter values are then correlated with working memory measures. Results indicate close relations between them, providing support for our hypothesis.

Keywords: working memory; probability learning; recency

Introduction

Decisions in life often condense into simple binary choices—to react or not to react, to speak or not to speak, to do or not to do. An important factor influencing such decision making is the outcomes of previous similar decisions. However, our abilities to integrate the histories of outcomes is strongly constrained by the attentional and processing limits of our working memory (WM), thus compromising the quality of our decision making. Indeed, several researchers have focused on differences in decision making between situations when information is gathered over sequential experience (where the narrow window of WM is likely to have an impact) and when it is obtained from simultaneous description (which is relatively uninfluenced by WM capacity; Hertwig, Barron, Weber, & Erev, 2004). The former is more typical of real-life, emphasizing the importance of examining the role of WM limits. In the current paper, we investigate how limits in storage and processing mechanisms of WM influence behavior on binary choices through the probability learning task.

Probability Learning Task

The probability learning task is a simple experimental paradigm involving multiple trials of choosing between two mutually exclusive and exhaustive outcomes (Vulkan, 2000). For instance, in each trial, participants may be asked to predict which of two presented light bulbs will turn on (Humphreys, 1939). Typically, the two options have pre-determined and unequal probabilities of occurring—e.g., Bulb A will turn on with 0.7 probability, and Bulb B with 0.3 probability. Each trial is independent; hence the optimal strategy is to choose the higher probability side (once it has been identified) 100% of the time. This is known as probability maximizing—in our example such a strategy would lead to 70% accuracy.

However, participants rarely perform this relatively simple strategy of exploring for the high payoff option and then exploiting via probability maximizing. Rather, a typically observed behavior is probability matching—choosing options in proportion to their probability of occurrence. Participants therefore tend to choose Bulb A 70% of the time and Bulb B 30%, leading to a lower accuracy level of 58% (.7×.7 + .3×.3). This behavior typically persists even after enough samples have been drawn to identify the higher probability option with at least some level of certainty (Arrow, 1958). Probability matching has been given wide attention as a supposed lapse of judgement for which several explanations have been proposed, without much consensus regarding the underlying mechanism (Feher da Silva, Victorino, Caticha, & Baldo, 2017).

Working Memory and Probability Matching

One of the primary explanations of probability matching is the recency effect. Human short-term retention abilities are limited, creating a narrow window of recent experience which makes information highly susceptible to time-based decay (Kareev, 1995). In the current task, this constraint encourages decisions to be based on smaller samples of information (most likely the very recent samples), which, given the law of large numbers, is likely to produce probability matching behavior (Plonsky, Teodorescu, & Erev, 2015; Rakow & Newell, 2010). For example, if participants retain only one previous trial in their short-term window and make utility calculations and decisions based on this previous trial, they would exhibit perfect matching. Several studies have fit such one-outcome-based win-stay-lose-shift strategies to decision making with surprising success despite their relative simplicity (Nowak & Sigmund, 1993). More sophisticated reinforcement learning models (such as the EVL and PVL models; Busemeyer & Stout, 2002; Erev & Roth, 1998) also incorporate a recency weighting which discounts the influence of older outcomes. Such findings suggest that probability matching behavior could be a result of overweighting recent outcomes, produced by their higher activation in the attentional window.

It must be noted that most studies find that probability matching does not persist—when enough trials are

presented, participants are often able to switch to the optimal strategy of maximizing. For instance, Restle (1961) found that probability matching disappeared after 1000 trials. Other studies have emphasized that switching to the optimal strategy is more likely if participants are provided with higher monetary payoffs, regular feedback, and more intense practice (Shanks, Tunney, & Mccarthy, 2002). An interpretation of this is that probability matching (produced e.g. by short-term recency) is a default response, which can be overridden in favor of maximizing through conscious deliberation. This dual process hypothesis is supported by correlations between SAT scores and maximizing on a descriptive version of this task (West & Stanovich, 2003).

These features of probability learning behavior—recency-based responding and deliberate strategy shift to maximizing—are likely to be mediated by WM capacity. Several models of WM consist of two core functions, storage and processing (frequently known as the span and control of attention respectively; Cowan, 2008). Here, we refer to storage as the ability to temporarily hold information in an active attentional state, protected from time-based decay and other interference. Decay in storage capacity is likely to produce recency-based performance in the probability learning task, as it constrains the number of previously observed outcomes that are in a readily accessible state when making a new decision (Ricker, Vergauwe, & Cowan, 2016). The processing component of WM directs attentional use, focusing it on goal-relevant information. An important function of WM processing is the inhibition of automatic but incorrect responding, as suggested by correlations with performance on the antisaccade and Stroop tasks (Kane & Engle, 2003; Unsworth, Schrock, & Engle, 2004). In our task, this component is perhaps responsible for resisting convenient recency-based responding and deducing the optimal strategy by steering and focusing attention toward task-relevant information (which could include independence of trials and the existence of a higher probability option).

Based on this previous research, in our study, we hypothesize the following to be correlated: (1) recency-based responding and WM storage capacity, and (2) strategy shift to maximizing and WM processing abilities.

Previous Studies and the Current Experiment

Several experiments have previously linked WM with performance on probability learning or other similar tasks (Gaissmaier, Schooler, & Rieskamp, 2006; Kareev, 1995; Rakow & Newell, 2010). These studies have reported mixed results—some have found positive correlations between WM capacity and maximizing, while others have reported the opposite. Through this paper, we attempt to resolve this debate. Further, unlike previous studies, our primary motivation is to model the interaction of the two WM components in producing recency-based responding and suppressing it in favor of the optimal strategy.

For our task, we used the light bulb setting described earlier. Participants chose between two bulbs and received feedback (i.e., which bulb lit up) after each trial. To model probability learning behavior, we used the Strategy-Shift Expectancy-Valence Learning (SS-EVL) model–a variant of the original EVL model (Busemeyer & Stout, 2002). Recency and strategy shift parameters extracted from this model were correlated with WM scores. Since such statistical analysis is likely to be noisy, our study has a larger sample size than that of previous experiments.

Methods

Participants

One hundred and thirty-one undergraduate students of Indiana University served as participants and were compensated with course credit. Of these, data of eight participants was excluded due to failure to perform at least one of the tasks.

Tasks and Procedure

The experiment consisted of five computer-based tasks (four WM and one probability learning). Each session lasted around 60 minutes and began with administration of the WM tasks.

Memory tasks. Participants performed four WM tasks in the following order: symmetry span, digit span, visual array, and operation span.

WM storage was measured with the digit span and visual array tasks. The digit span is a simple number recall task classically used as a measure of short-term memory (method similar to Quinn, Tuci, Harvey, Di Paolo, & Wood, 2005). The visual array task requires detecting rapid color changes in an array of 4, 6, 8, or 10 colored squares (method similar to Cowan, Fristoe, Elliott, Brunner, & Saults, 2006). Here, task performance depends on temporary storage of colors, and has been frequently used as a measure of storage (Cowan et al., 2006; Shipstead, Redick, Hicks, & Engle, 2012).

The symmetry span and operation span tasks require simultaneous usage of memory and processing and were used as measures of WM processing (methods similar to Oswald, Mcabee, Redick, & Hambrick, 2014). The memory component of these tasks involves the retention of presented items (spatial positions of colored squares for symmetry span and letters for operation span). Memory items are interpolated with processing components (symmetry or arithmetic accuracy judgements respectively) that interfere with rehearsal of memory items.

These specific working-memory tasks were selected because they not only represent the functional components of working memory (i.e., storage and processing), but also use different content modalities—symmetry span and visual array are visuo-spatial tasks, while digit span and operation span are verbal-numeric tasks.

Probability learning task. Participants performed three probability learning games, each involving 100 trials. During each game, participants were presented with an

image of a 'bulb-box', a device containing two lightbulbs (Bulb A and Bulb B). Participants were informed that on each trial one of the two bulbs would turn on and it was their task to guess the correct bulb. For every correct guess, participants gained one point and for every incorrect guess, they lost one point. Number of points won by participants was revealed at the end of each game. To motivate participants to aim for higher points and achieve optimal decisions, participants were rewarded with between 0 to 3 nutrition bars based on performance. The probability with which the two bulbs lit up remained constant within each game but varied from game to game. Three probability contingencies were used—0.60, 0.70, and 0.80—the order of which was determined randomly. Bulb A or Bulb B was set as the more frequent bulb in each game with equal probability. Participants were informed that each 'bulb-box' game had a different underlying 'program' controlling it to minimize tendencies of using previous games as priors for future ones. To further combat this, the color of the lightbulbs was changed from game to game.

Results

Probability matching (selecting the bulbs in proportion to how often they light up) was observed in the aggregated data of participants, decreasing with successive games (Figure 1).

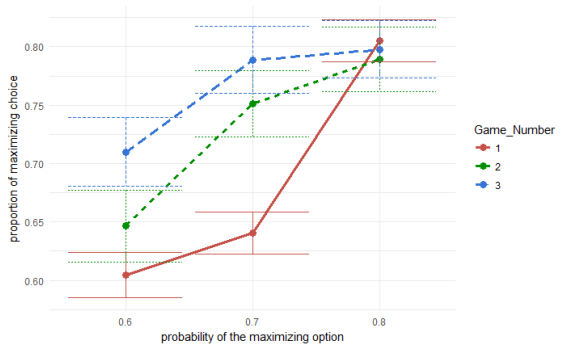


Figure 1: Proportion of maximizing choices averaged across trials (data for all game and probability contingencies)

Further, we found that participants were more likely to choose the maximizing option as the number of trials played increased within each game (Figure 2).

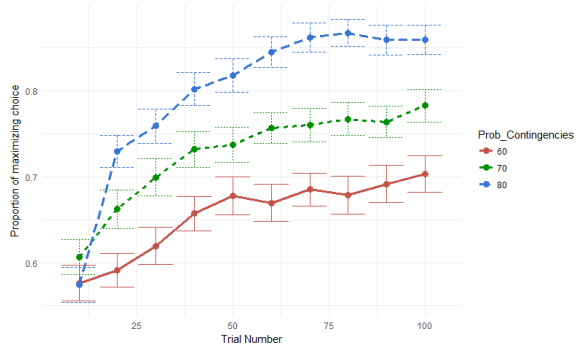


Figure 2: Averaged proportion of maximizing responses across trials

We then calculated correlations between WM scores and frequency of maximizing responding. Maximizing responding was calculated as the proportion of times the maximizing option was selected in a game. Significant correlations were obtained for scores on visual array ($r(123) = 0.2$, $p=.03$) and spatial span tasks ($r(123) = 0.19$, $p=.04$), while correlations with digit span ($r(123) = 0.09$, $p=.36$) and operation span ($r(123) = 0.19$, $p=.07$) were weaker. Stronger correlation with the visuo-spatial WM tasks (as opposed to the verbal ones) could arise if participants were retaining previous outcomes as visuo-spatial information (e.g. *left* bulb, *right* bulb, *right* bulb…).

These positive correlations between WM and optimal responding are in line with our hypotheses. They are consistent with results from some previous studies on WM and probability learning (Rakow & Newell, 2010; West & Stanovich, 2003); but contradict others which have found negative correlations (Gaissmaier et al., 2006; Kareev, 1995).

Modelling

Correlation measures provide us a small peak into the relationship between WM capacity and probability matching. However, they do not reveal the relation between WM capacity and the use of recency or strategy shift to maximizing. We therefore modelled the data using a modified EVL model and correlated parameters with WM scores. We also employed a Baseline Bernoulli model for comparison.

Model Descriptions

Strategy-Shift Expected-Valence Learning Model (SS-EVL). Variants of the EVL model have been previously used to model probability learning (Feher da Silva et al., 2017; Schulze, van Ravenzwaaij, & Newell, 2015) and other reinforcement learning tasks (such as the Iowa and Soochow Gambling Tasks; Ahn, Busemeyer, Wagenmakers, & Stout, 2008). Its parameters typically include consistency c and recency A. In our version of the model, we accommodate a strategy shift toward maximizing through a third parameter—timepoint of shift T.

The model assumes that on every trial, participants assign a utility value to the two lightbulbs—1 if it is correct on that trial, and 0 otherwise. Therefore, in a trial, utility $u(t)$ gained from bulb j based on outcome x is defined by:

$$u_j(t) = \begin{cases} 1 & if\ x(t) = j, \\ 0 & if\ x(t) \neq j \end{cases}$$

This utility is then incorporated into the running expected utility E_j of the two options using a weighted utility updating rule (Rescorla & Wagner, 1972) which discounts older outcomes with a recency parameter A. Larger the value of A, greater is the influence of older outcomes:

$$E_j(t) = A \cdot E_j(t-1) + (1-A) \cdot u(t)$$

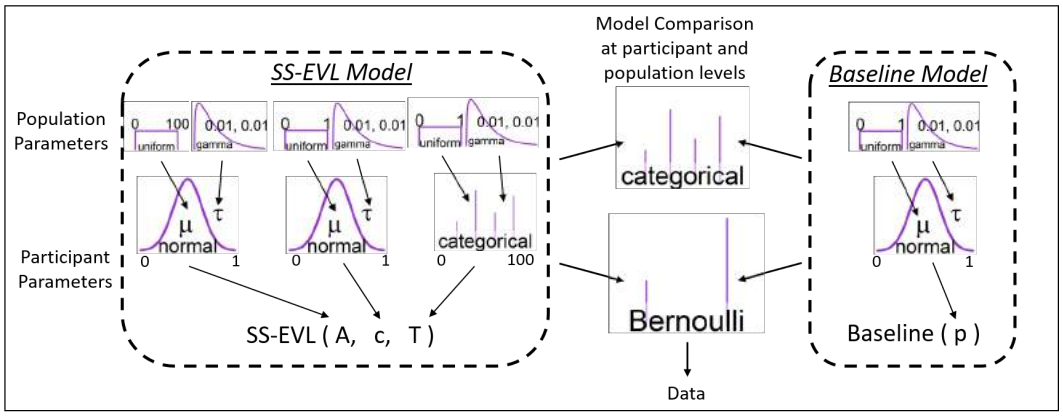


Figure 3: Structure and priors of the hierarchical Bayesian model

The expected utility calculations are then used to make a choice decision D based on Luce's choice rule (Luce, 1959) incorporating exploration θ:

$$\Pr[D(t+1)=j]=\frac{e^{\theta(t)\cdot E_j(t)}}{\sum_{k=1}^{2} e^{\theta(t)\cdot E_k t)}} \;;\quad \theta(t)=(\frac{t}{10})^{c}$$

$\theta(t)$ represents the extent to which participants make choice decisions based on calculated utilities. If $\theta(t)=0$, decisions are random and as $\theta(t)$ increases, decisions are highly sensitive to utilities. The value of θ is dependent on the free consistency parameter *c,* which is constrained between 0 and 1. Though we do not use this parameter for future WM analysis, it is essential to incorporate it in the model—it provides for a cleaner estimate of recency by accounting for the influence of exploration in participant data.

Finally, we assume that at some trial *T*, participants identify and shift to the maximizing strategy. Therefore, from this trial onward, the expected utilities of the maximizing and non-maximizing options are set to 1 and 0 respectively. Hence, the running utility E_j is revised such that:

$$\textbf{\textit{for } t > T}: \quad E_j(t)=\begin{cases} 1 \;\; if\; j = maximizing\; option, \\ 0\; if\; j \neq maximzing\; option \end{cases}$$

Baseline Model. A simple Bernoulli baseline model was also fit to data. The Baseline model has only one parameter—probability that participants choose the maximizing option, $p(j=1)$. Therefore, the model predicts unequal probabilities of choosing between the two bulbs, which are independent of outcomes observed by participants and constant across trials.

In our task, participants could be using varied strategies (e.g., looking for patterns in outcomes or random guessing). This model serves to filter out such participants who are better modelled by a random Bernoulli process than by a recency model which assumes positive dependency on observed outcomes. Thus, this model is not intended to be a process model of the underlying mechanism, but rather a useful cache for unaccounted strategies. If a larger number of participants are better fit by this model than the SS-EVL, it suggests that our proposed mechanism of probability matching is not dominant in the population.

Model Fitting

We used Bayesian hierarchical modelling for parameter fitting and model comparison (see Figure 3 for details about prior and multilevel structure). We combined the two models into a single hyper-model and employed a categorical distribution to determine the strategy used by each participant— on each MCMC timestep, for each trial, it sampled from one or the other model based on its probability of being the true process underlying that participant's data. In a similar way, we also estimated the population level posterior probability for each model. The analysis was implemented on JAGS via R. We drew 200,000 samples via three MCMC chains. Inspection of diagnostic plots indicated convergence for most parameters.

Here we only fit data from the first probability learning game of each participant because of considerable order and practice effects in future games.

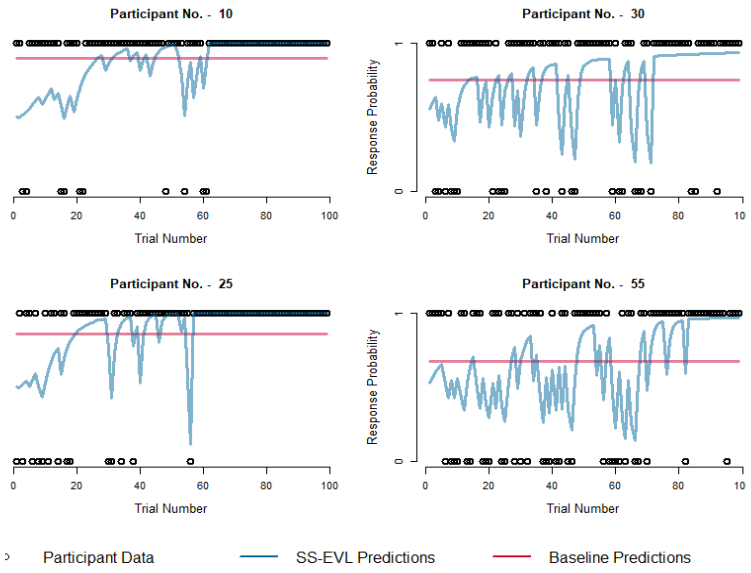


Figure 4: Model fitting of individual participants

Model Comparisons

Overall, the SS-EVL model outperformed the Baseline, with a posterior probability *P(model=SS-EVL|D)* of 0.71. Further, 85 out of 123 participants were categorized as employing an SS-EVL strategy (for examples of individual fit, see Figure 4). SS-EVL also better captured the participants' average pattern of performance across trials (Figure 5).

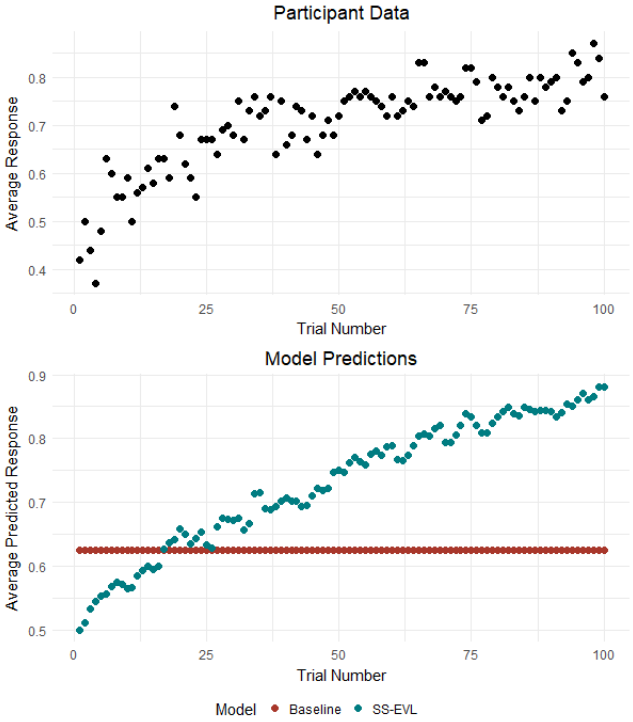


Figure 5: Average participant data and model predictions across trials

Correlations with WM scores

To test our hypothesis that WM components correlated with strategy use, we analyzed those participants who were better fit by the SS-EVL model. Point estimates (modes) of two SS-EVL parameters were correlated with WM scores—recency *A* and timepoint of strategy shift *T* (Table 1). As in the behavioral correlations reported above, visual array and symmetry span were more strongly correlated than other measures. Of the two measures of WM storage, only visual array showed indication of correlation with recency, providing partial support for hypothesis 1 that overweighting of recent outcomes is a by-product of WM storage limits. As predicted in hypothesis 2, measures of WM processing shared a significant negative correlation with timepoint shift—participants with higher WM processing abilities were likely to shift toward the maximizing strategies within fewer trials.

Discussion

Our study demonstrates the process by which WM components work together to produce typical probability learning behaviour. The picture that emerges suggests that the limits of the WM store intensify weighting of recent events, producing default responses that require greater WM processing to inhibit them in favor of the optimal strategy. In the real world, such a tendency toward recency makes sense as it allows us to adapt to our dynamic and temporally autocorrelated environment, where making decisions based on older information is often unsuccessful and recent events are a good indicator of the current state of the world (Plonsky et al., 2015). It appears that the two components of WM thus work together to produce appropriate everyday behavior—limits in the WM store allow for quick recency-based responses to environmental stimuli while WM processing acts as a correctional mechanism, stepping in to replace the recency-based strategy if an optimal strategy is found.

It would therefore be hasty to call probability matching a lapse in judgement (Vulkan, 2000)—participants do not fail to arrive at successful decisions in the probability learning task because of some cognitive failure. Rather, they do not always use the optimal strategy because the task itself is not representative of natural environments: unlike typical real-world situations, here the event probabilities are stationary across trials, and the trials are independent of one another. Participants therefore must deploy deliberate processing to resist responding automatically based on assumed environmental structures where recency would be best. While binary decisions may be common to our everyday life, the probability structure underlying this task is not, making the optimal strategy unintuitive. Future work can examine participant performance using real-world probability structures—for instance having the probabilities of the bulbs shift or be autocorrelated across trials (Gaissmaier & Schooler, 2008).

As mentioned earlier, previous studies have found mixed

Table 1: Correlations between WM obtained parameter values

	WM storage measures		WM processing measures	
	Visual Array	Digit Span	Symmetry Span	Operation Span
Recency *(A)*	0.19^{+}	0.08	0.18^{+}	0.11
Timepoint of shift *(T)*	−0.16	−0.11	-0.24^{*}	-0.20^{+}
$^{+}p<.1$. $^{*}p<.05$. $^{**}p<.01$. $^{***}p<.001$ N=85				

results when relating WM to performance in similar tasks—some have obtained positive correlations, providing support to our results (e.g., Rakow & Newell, 2010; West & Stanovich, 2003), while others have obtained the opposite (e.g., Gaissmaier et al., 2006; Kareev, 1995). While, the differing results could be due to difference in task structure—the studies reporting negative correlations use a correlation-detection task, which involves estimating two probabilities and not just one (for details of the task, refer to Kareev, 1995)—this is an unlikely explanation since our model would still predict positive correlations for such a task structure. Therefore, a more likely possibility is that participants employ different strategies (such as pattern matching, random responding etc.), producing different results. In the current paper, we only focused on recency-based responding—the SS-EVL model fit participants for this specific strategy and our results suggested that it was the dominant strategy in our sample when compared to a Bernoulli baseline. We then correlated the obtained parameter estimates for participants best fit by this model with WM scores, therefore excluding any effect of other strategies. However, future work must model other possible strategies, determine their frequency in the sample and their relation to WM capacity.

Further work must also be done to narrow in on the mechanisms underlying these decisions. While our model estimates the timepoint at which the strategy-shift toward maximizing occurs, it does not uncover the mechanism that produces this shift. Our correlational evidence argues that this mechanism is associated with the processing component of WM, but we do not know what operation within this component leads to optimal strategizing and why it reaches a threshold at a particular timepoint. Identifying the likely mechanisms at work in making decisions based on recent and older information will help us understand the role of limited WM storage and processing in these common choice settings.

References

Ahn, W.-Y., Busemeyer, J. R., Wagenmakers, E.-J., & Stout, J. C. (2008). Comparison of Decision Learning Models Using the Generalization Criterion Method. *Cognitive Science*, *32*, 1376–1402. https://doi.org/10.1080/03640210802352992

Arrow, K. J. (1958). Utilities, attitudes, choices: A review note. *Econometrica*, *26*(1), 1–23.

Busemeyer, J. R., & Stout, J. C. (2002). A Contribution of Cognitive Decision Models to Clinical Assessment: Decomposing Performance on the Bechara Gambling Task The Bechara Gambling Paradigm The gambling task involves the use of four decks of cards (labeled here. *Psychological Assessment*, *14*(3), 253–262. https://doi.org/10.1037/1040-3590.14.3.253

Cowan, N. (2008). What are the differences between long-term, short-term, and working memory? *Progress in Brain Research*, *169*, 323–338. https://doi.org/10.1016/S0079-6123(07)00020-9

Cowan, N., Fristoe, N. M., Elliott, E. M., Brunner, R. P., & Saults, J. S. (2006). Scope of attention, control of attention, and intelligence in children and adults. *Memory & Cognition*, *34*(8), 1754–1768.

Erev, I., & Roth, A. E. (1998). Predicting How People Play Games: Reinforcement Learning in Experimental Games with Unique, Mixed Strategy Equilibria. *The American Economic Review*, *88*(4), 848–881.

Feher da Silva, C., Victorino, C. G., Caticha, N., & Baldo, M. V. C. (2017). Exploration and recency as the main proximate causes of probability matching: a reinforcement learning analysis. *Scientific Reports*, *7*(1), 15326. https://doi.org/10.1038/s41598-017-15587-z

Gaissmaier, W., & Schooler, L. J. (2008). The smart potential behind probability matching. *Cognition*, *109*(3), 416–422. https://doi.org/10.1016/j.cognition.2008.09.007

Gaissmaier, W., Schooler, L. J., & Rieskamp, J. (2006). Simple predictions fueled by capacity limitations: When are they successful? *Journal of Experimental Psychology: Learning Memory and Cognition*, *32*(5), 966–982. https://doi.org/10.1037/0278-7393.32.5.966

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Is reading about the kettle the same as touching it? Decisions from experience and the effects of rare events in risky choice. *Psychological Science*, *15*(8), 534–539. https://doi.org/10.1093/geronb/gbt081

Humphreys, L. G. (1939). Acquisition and extinction of verbal expectations in a situation analogous to conditioning. *Journal of Experimental Psychology*, *25*(3), 294–301. https://doi.org/10.1037/h0053555

Kane, M. J., & Engle, R. W. (2003). Working-memory capacity and the control of attention: the contributions of goal neglect, response competition, and task set to Stroop interference. *Journal of Experimental Psychology. General*, *132*(1), 47–70.

Kareev, Y. (1995). Through a narrow window: Working memory capacity and the detection of covariation. *Cognition*, *56*, 263–269.

Luce, R. D. (1959). On the possible psychological laws. *The Psychological Review*, *66*(2), 81–95. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.800.3821&rep=rep1&type=pdf

Nowak, M., & Sigmund, K. (1993). A strategy of win-stay, lose-shift that outperforms tit-for-tat in the Prisoner's Dilemma game. *Nature*, *364*(6432), 56–58. https://doi.org/10.1038/364056a0

Oswald, F. L., Mcabee, S. T., Redick, T. S., & Hambrick, D. Z. (2014). The development of a short domain-general measure of working memory capacity. https://doi.org/10.3758/s13428-014-0543-2

Plonsky, O., Teodorescu, K., & Erev, I. (2015). Reliance on Small Samples, the Wavy Recency Effect, and Similarity-Based Learning. *Psychological Review*, *122*(4), 621–647. https://doi.org/10.1037/a0039413

Quinn, M., Tuci, E., Harvey, I., Di Paolo, E., & Wood, R.

(2005). Evolutionary robotics: A new scientific tool for studying cognition. *Artificial Life*, *11*(1–2), 79–98.

Rakow, T., & Newell, B. R. (2010). The role of working memory in information acquisition and decision making : Lessons from the binary prediction task. *The Quarterly Journal of Experimental Psychology*, *63*(7), 1335–1360. https://doi.org/10.1080/17470210903357945

Rescorla, R., & Wagner, A. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical Conditioning: Current Research and Theory, Vol. 2*, (May), 64–99. https://doi.org/10.1101/gr.110528.110

Restle, F. (1961). *Psychology of judgment and choice: A theoretical essay. Psychology of judgment and choice: A theoretical essay.* Oxford, England: Wiley.

Ricker, T. J., Vergauwe, E., & Cowan, N. (2016). Decay theory of immediate memory: From Brown (1958) to today (2014). *Quarterly Journal of Experimental Psychology (2006)*, *69*(10), 1969–1995. https://doi.org/10.1080/17470218.2014.914546

Schulze, C., van Ravenzwaaij, D., & Newell, B. R. (2015). Of matchers and maximizers: How competition shapes choice under risk and uncertainty. *Cognitive Psychology*, *78*, 78–98. https://doi.org/10.1016/J.COGPSYCH.2015.03.002

Shanks, D. R., Tunney, R. J., & Mccarthy, J. D. (2002). A Re-examination of Probability Matching and Rational Choice, *250*(March), 233–250.

Shipstead, Z., Redick, T. S., Hicks, K. L., & Engle, R. W. (2012). The scope and control of attention as separate aspects of working memory. *Memory*, *20*(6), 608–628. https://doi.org/10.1080/09658211.2012.691519

Unsworth, N., Schrock, J. C., & Engle, R. W. (2004). Working Memory Capacity and the Antisaccade Task: Individual Differences in Voluntary Saccade Control. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(6), 1302–1321. https://doi.org/10.1037/0278-7393.30.6.1302

Vulkan, N. (2000). An Economist's Perspective on Probability Matching. *Journal of Economic Surveys*, *14*, 101–118.

West, R. F., & Stanovich, K. E. (2003). Is probability matching smart? Associations between probabilistic choices and cognitive ability. *Memory & Cognition*, *31*(2), 243–251. https://doi.org/10.3758/BF03194383

Sensorimotor Norms: Perception and Action Strength norms for 40,000 words

Dermot Lynott (d.lynott@lancaster.ac.uk)
Department of Psychology, Lancaster University, Lancaster, UK
Louise Connell (l.connell@lancaster.ac.uk)
Department of Psychology, Lancaster University, Lancaster, UK
Marc Brysbaert (Marc.Brysbaert@UGent.be)
Department of Experimental Psychology, Ghent University, Belgium.
James Brand (james.brand@canterbury.ac.nz)
New Zealand Institute of Language Brain and Behaviour, University of Canterbury, New Zealand
James Carney (James.Carney@brunel.ac.uk)
Department of Arts and Humanities, Brunel University, London, UK.

Abstract

Sensorimotor information plays a fundamental role in cognition. However, datasets of ratings of sensorimotor experience have generally been restricted to several hundred words, leading to limited linguistic coverage and reduced statistical power for more complex analyses. Here, we present modality-specific and effector-specific norms for 39,954 concepts across six sensory modalities (touch, hearing, smell, taste, vision, and interoception) and five action effectors (mouth/throat, hand/arm, foot/leg, head excluding mouth, and torso), which were gathered from 4,557 participants who completed a total of 32,456 surveys using Amazon's Mechanical Turk platform. The dataset therefore represents one of the largest set of semantic norms currently available. We describe the data collection procedures, provide summary descriptives of the data set, demonstrate the utility of the norms in predicting lexical decision times and accuracy, as well as offering new insights and outlining avenues for future research. Our findings will be of interest to researchers in embodied cognition, cognitive semantics, sensorimotor processing, and the psychology of language generally. The scale of this dataset will also facilitate computational modelling and big data approaches to the analysis of language and conceptual representations.

Keywords: embodied cognition; semantics; norms

Background

Sensorimotor information is central to how we experience and navigate the world. We acquire information through our senses, while our bodies provide feedback, as we physically interact with objects, people, and the wider environment. Many theoretical views of cognition describe a fundamental role for such sensorimotor knowledge (e.g., Barsalou, 1999; Connell & Lynott, 2014; Smith & Gasser, 2005), with numerous empirical demonstrations supporting such claims (e.g., Connell, Lynott & Dreyer, 2012; Kaschak et al., 2006; Matlock, 2004; Zwaan & Taylor, 2006).

In order to test such embodied (or grounded) theories of cognition, researchers need appropriate stimuli for empirical tests and for developing mathematical or computational models. Lynott and Connell (2009, 2013) developed a set of modality-specific sensory norms for concepts where each sensory modality (e.g., auditory, gustatory, haptic, olfactory, visual) maps onto distinct cortical regions (e.g., gustatory cortex, auditory cortex etc.). By having individuals provide ratings for each modality separately, the norms capture the extent to which something is experienced across different sensory modalities, without risk of ignoring or distorting the role of particular modalities (Connell & Lynott, 2016). Subsequent empirical studies have found that such modality-specific measures are good predictors of people's performance across a range of cognitive tasks (e.g., lexical decision, word-naming) and often out-performed long-established measures such as concreteness and imageability (e.g., Connell & Lynott, 2012; 2014). For example, in examining performance on lexical decision and word naming (reading aloud) tasks, Connell and Lynott (2012) found that modality-specific experience (and specifically the highest level of perceptual experience on any modality for a given concept, or "max strength") was a more reliable predictor of performance than either concreteness and imageability.

An added advantage of using measures of sensory experience for specific modalities is that it allows researchers to tap into effects that relate to particular modalities and not others. Connell & Lynott (2010) showed how a processing disadvantage for tactile stimuli observed during perceptual processing (Spence, Nichols & Driver, 2001) was also observed when processing modality-specific words. Connell and Lynott (2014) derived contrasting modality-specific predictions relating to lexical decision and reading aloud for individual words. Thus, for lexical decisions, a visually-focussed task, strength of perceptual experience in the visual modality (but not the auditory modality) was a reliable predictor of performance. By contrast, reading aloud, requires additional attention on the auditory modality (as participants must monitor their speech output to ensure correctly articulated responses). Consistent with this idea, both strength of auditory experience and strength of visual experience were reliable predictors for performance for the reading aloud task. Other semantic measures (such as concreteness or imageability) could not have been used as they do not offer sufficient granularity in terms of sensory experience. Thus, modality-specific measures of sensory experience provide the capacity to generate and test novel predictions related to modality-specific processing and representations.

More recently, Connell, Lynott and Banks (2018) showed that interoception (i.e., sensations inside the body) also plays an important role in semantic representations, and could be a primary grounding mechanism for abstract concepts. It was found that strength of interoceptive experience was higher for abstract concepts, such as *hungry* and *serenity*, compared to more concrete concepts like *capacity* or *rainy*. What's more, interoceptive experience was found to be most important for emotion concepts, especially for negative emotions such as *fear* and *sadness,* with interoceptive experience found to be just as important as other sensory modalities in capturing semantic knowledge.

Finally, speaking to the utility of sensory norms and broader interest in this area, several research groups have extended this earlier work, either by developing modality-specific norms in other languages, including Russian, Serbian, Dutch and Mandarin (Miklashevsky, 2018; Đurđević et al., 2016; Speed & Majid, 2017; Chen et al., 2017), or by applying these norms in novel ways. For example, the original modality-specific norms have been used to examine stylistic differences of authors (Kernot, Bossomaier & Bradley, 2019), test models of lexical representations (Johns & Jones, 2012), and evaluate the iconicity of words (Winter et al., 2017).

Nonetheless, a notable gap in the work discussed above is that it focuses solely on sensory experience, and has not included parallel measures of action or effector-specific experience. However, there is good evidence for the relevance of action experience to people's semantic representations of concepts (e.g., Glenberg & Gallese, 2012; Hauk, Johnsrude & Pulvermuller, 2004). For instance, manual action verbs like *throw* activate some of the same motor circuits as moving the hand (Hauk et al., 2004), and their processing is selectively impaired in patients with Parkinson's disease, which entails neurodegeneration of the motor system (Boulenger et al., 2008; Fernandino et al., 2013). Critically, the motor basis to semantic knowledge is specific to the bodily effector used to carry out a particular action. Applying transcranial magnetic stimulation (TMS) to hand and leg areas of the motor cortex differentially influences processing of hand- and leg-action words: hand area TMS facilitates lexical decision of hand-action words like *pick* compared to leg-action words like *kick*, whereas this effect is reversed with leg-area TMS (Pulvermueller, Hauk, Nikulin & Ilmoniemi, 2005). Such double dissociations in motor-language facilitation underscore the importance of individually examining separate action effectors when norming the motor basis of words and concepts.

Some existing measures have attempted to capture action knowledge, but have alternatively used feature production methods as opposed to rating dimensions of action (e.g., where people verbally list features associated with concepts: McRae et al., 2005; Vinson & Vigliocco, 2008), focused on generalised action (e.g., body-object interaction: Tillotson, Siakaluk, & Pexman, 2008; relative embodiment: Sidhu, Kwan, Pexman, & Siakaluk, 2014; see Connell & Lynott, 2015, for review), or on a restricted subset of action types (e.g., graspability: Amsel, Urbach, & Kutas, 2012: actions associated with lower limb, upper limb, or head: Binder et al., 2016) that omits other parts of the body involved in action. For example, the action of *pushing* can also involve the torso (Moody & Gennari, 2010), and mouth actions are cortically distinct from other actions of the face (Meier, Aflalo, Kastner, & Graziano, 2008). To our knowledge, therefore, there is no large-scale set of norms that taps into a comprehensive range of effector-specific action experience. In the present work, we address this gap by collecting effector-specific action strength norms for a large number of concepts.

Here, we present sensorimotor norms collected across 11 dimensions for approximately 40,000 concepts, comprising 6 modality-specific dimensions of perceptual strength (auditory, gustatory, haptic, olfactory, visual, interoceptive) and 5 effector-specific dimensions of action strength (head, arm/hand, mouth/throat, leg/foot, torso).

Study 1: Sensorimotor Norms

Method

Participants A total of 4,557 unique participants completed 32,456 surveys via Amazon's Mechanical Turk platform (M = 7.12 samples per participant). Data for perceptual strength ratings and action strength ratings were gathered separately. Participants were self-selecting and had English as their first language. We recruited only experienced MTurk users who had already completed over 100 HITS, and high-quality participants who had >97% HIT approval. Participants were remunerated at a rate above minimum wage in the US.

Materials Perceptual and action ratings were collected for a total of 39,954 words. These words were taken from Brysbaert, Warriner, & Kuperman's (2014) work on concreteness ratings, which included 37,058 English lemmas and 2,896 two-word expressions. These words were split into 832 lists of 48 items, along with 5 calibrator words and 5 control words occurring in each. Responses to controls and calibrators (selected for being highly familiar, and low in variance based on previous norms) were used for quality checks, which we describe below in the subsection on Data Quality and Exclusions. Lists were populated to provide words that varied in terms of familiarity ("percentage known" in Brysbaert et al's study) and concreteness.

Procedure Using Qualtrics survey software, a template survey was created that followed procedures developed in Lynott & Connell (2009, 2013). At the start of the survey participants read an information sheet, and indicated their informed consent to continue with the study. Specifically, each concept in a 58-word sample was presented individually on a screen (order randomised by participant) followed by question text. For perceptual strength ratings, the text was "To what extent do you experience WORD," where WORD was replaced with the concept in question. Underneath were six rating scales, one for each of the perceptual modalities under investigation, labelled "By feeling through touch", "By hearing", "By sensations inside your body", "By smelling" and "By tasting,". The order of the ratings scales was randomised by sample.

For the action strength ratings, the text read "To what extent do you experience WORD by performing an action with," followed by choices of "Foot / leg", "Hand / arm", "Head excluding mouth", "Mouth / throat", "Torso". For these ratings, each scale also contained an image of a body avatar that highlighted the body part relevant to each intended effector. Participants were asked to rate the extent to which they experience each concept through each of the named senses or effectors; both the sensory and motor components had 6-point scales ranging from 0 (not at all) to 5 (greatly).

There was no default value selected on the scale and participants clicked on a button under the relevant value to select or change their response. Participants were explicitly told there were no right or wrong answers and they should use their own judgment; they were also instructed to select the "I don't know the meaning of this word" option if the word was unfamiliar to them. Progress to the next item could only occur if values were selected for all perceptual senses or action effectors or the "I don't know the meaning of this word" option was checked. The study was self-paced and timed to last 18-20 minutes.

Data quality and exclusions In order to ensure the data collected is of sufficiently high quality, we instituted a number of checks, in terms of individual performance, item performance, and agreement for each list of words. Overall, only 0.8% of all responses were removed following data checks. Participants whose scores exhibited a Pearson's r < 0.2 with the controls or who responded 'don't know the meaning of this word' for more than five control and calibrator words, were dropped from the sample. Additionally, there were a small number of participants who completed the same sample of words more than once, when this happened only the earliest submitted responses were retained. Cronbach's alphas (Cronbach, 1951) were calculated for each modality for all other participants; results were only retained when the mean alpha for all samples was >= 0.8.

Norms Data The final set of norms, results, analyses, and scripts are available on the project's Open Science Framework page: https://osf.io/7emr6/

Results

Summary statistics were calculated for all valid samples, with 39,707 words included in the overall norms, following exclusion criteria. Each word in a sample is represented by a row that contains ratings for each of the 11 dimensions. Each dimension has separate values for mean score, standard deviation, median score, trimmed mean, trimmed standard deviation by modality/effector, and the percentage of participants who knew the word. Inter-rater reliability by modality/effector was high for both perceptual and action ratings: mean Cronbach's alphas for perceptual modalities were: auditory 0.93, gustatory 0.96, haptic 0.92, interoceptive 0.92, olfactory 0.94 and visual 0.90; for action effectors mean alphas were: foot 0.93, hand 0.91, head 0.85, mouth 0.92 and torso 0.89.

Following Lynott and Connell (2009; 2013), additional variables of interest were calculated for each of the words in the sensorimotor norms. This included: Exclusivity scores (i.e., a measure of the extent to which a particular concept is experienced through a single dimension, calculated per word as the rating range divided by the sum of the ratings, and extending from 0%, for completely multidimensional, to 100%, for completely unidimensional); separate exclusivity scores were calculated for the perceptual (6 modalities) and action components (5 effectors), in addition to scores calculated across all 11 dimensions. Similarly, each concept was assigned a dominant dimension (i.e., the dimension that had the highest mean rating), for the perceptual, action and the full sensorimotor norms. When the highest mean rating was found in more than 1 dimension (Perceptual: N = 593; Action: N = 706; Sensorimotor: N = 478), a random dimension was assigned.

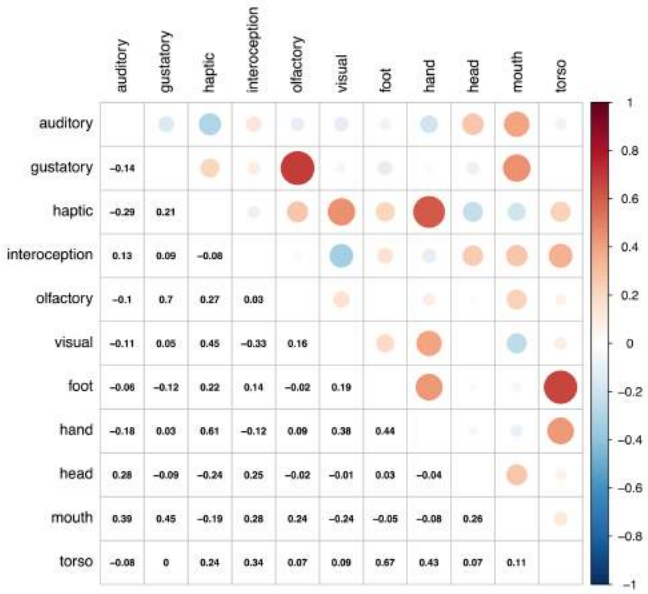


Figure 1 Correlation matrix plot between 11 dimensions for mean ratings of the sensorimotor strength norms (N = 39,707). Larger circles indicate stronger correlations, with red shades being positive, and blue shades being negative.

The norms confirm previous reports that we predominantly experience the world perceptually through our visual modality (Lynott & Connell, 2009; 2013; Winter, Perlman & Majid, 2018 – See Table 1), with the head emerging as the primary action effector. The least prominent dimensions were gustation and olfaction, highlighting the fact that only a small subset of the conceptual system is experienced strongly through these modalities. For the action norms, the head was observed to be the dominant effector and the torso had the least dominance.

Bayesian correlation analysis (Figure 1) between the dimensions showed that almost all the dimensions were significantly correlated with one another, with the exception of gustation~torso, as well as head~vision, with correlations approaching zero. It should be noted however, that a large number of the correlations were very weak, which is to be expected as each dimension is tapping into different aspects of sensorimotor experience. In some cases of course, certain dimensions often co-occur in our sensorimotor experience,

with notable relationships found between gustation~olfaction (r = .70), foot/leg~torso (r = 0.67) and hand/arm~haptic (r = 0.62).

	M	*SD*
Perceptual Modality		
Auditory	1.514	0.991
Gustatory	0.324	0.697
Haptic	1.074	0.934
Interoceptive	1.032	0.880
Olfactory	0.390	0.619
Visual	2.897	0.902
Action Effector		
Foot/leg	0.807	0.750
Hand/arm	1.447	0.907
Head	2.276	0.719
Mouth/throat	1.257	0.903
Torso	0.816	0.670

Table 1 Mean Strength Ratings (0–5) and Standard deviations (SD) per Sensorimotor Dimension

Study 2: Modelling Lexical Decision

In Study 2, we address three issues. First, we determine what is the best composite variable (i.e., single value) for representing a concept's sensorimotor profile. Second, we wish to replicate the utility of perceptual strength ratings in modelling people's performance in cognitive tasks (e.g., Connell & Lynott, 2012), and establish the independent utility of action strength as a performance predictor. Third, we will check the generalisability of the findings, by examining performance across two different data sets (i.e., English Lexicon Project, British Lexicon Project).

While an 11-dimension sensorimotor profile is a rich source of semantic information about a particular concept, it can nonetheless be somewhat unwieldy for some uses. It is often useful to aggregate multiple dimensions into a single composite variable, such as for use as a predictor in regression analyses without unnecessarily inflating the number of parameters. A single variable would also facilitate comparisons with other single-variable measures of people's experience (e.g., concreteness, valence etc.) There are many different methods of creating a composite variable. Previous work on perceptual strength has used strength of the dominant modality (i.e., maximum perceptual strength rating across all modalities) as the preferred composite variable (e.g., Connell & Lynott, 2016; Connell, Lynott, & Banks, 2018), finding it offered a better fit than alternatives to visual word recognition performance (Connell & Lynott, 2012). However, work in Serbian (Đurđević, Stijačić & Karapandžić, 2016) found the best fit emerged from summed perceptual strength (i.e., sum of perceptual strength ratings across all modalities) or vector length (i.e., Euclidean distance of the multidimensional vector of perceptual strength ratings from the origin). It is difficult to be certain whether this variability is due to language differences (i.e., English vs Serbian) or sampling differences (i.e., hundreds of words with limited overlap).

We therefore sought to empirically determine the best single composite variable for the 11-dimension sensorimotor profile using a much larger and more representative sample of concepts in English. As with previous studies (e.g., Connell & Lynott, 2012), we judge the "best" variable to be the one that offers the best fit to lexical decision latency, a task where semantic facilitation emerges from automatic and implicit access to the sensorimotor basis of the concept.

Method

Materials A total of 22,297 words were collated, representing the intersection of data available between the sensorimotor strength norms and lexical decision data from the English Lexicon Project (Balota, et al., 2007). A separate set of 11,768 words was also collated from the British Lexicon Project (Keuleers, Lacey, Rastle & Brysbaert, 2012).

Candidate Composite Variables Composite variables were calculated separately for sensorimotor (all 11 dimensions), perception (6 dimensions) and action (5 dimensions) dimensions. Most of the candidate variables we tested are distance metrics in vector space of a particular concept (i.e., an 11-dimension vector) from the origin. Minkowski distance (with exponent parameter *m*) is a generalisation of these distance metrics: roughly speaking, the highest-value dimension always contributes to the calculated distance, and *m* determines the extent to which the other dimensions contribute according to how close their values are to the highest-value dimension. That is, low-value *m* means that all dimensions make noticeable contributions to the calculated distance, whereas high-value *m* means only the highest-value dimension(s) make noticeable contributions to the calculated distance.

For example, for Minkowski 10 distance (Minkowski distance at m = 10 of the vector from the origin), theoretically, it represents sensorimotor strength of the dominant dimension plus an attenuated influence of any other dimensions that are nearly as strong as the dominant dimension. By contrast Minkowski 3 distance represents sensorimotor strength in all dimensions but the influence of weaker dimensions is attenuated.

Our set of candidate variables comprises: maximum strength, Minkowski 3, Minkowski 10, Euclidean vector length, Summed strength, and single PCA component.

Design and Analysis We performed Bayesian linear regressions predicting the dependent variable of zRT (i.e., standardised Lexical Decision RT per participant) and accuracy from 2 datasets: Elexicon (ELP) and the British Lexicon Project (BLP). First, for each dependent variable we built a null model of lexical predictors (log SUBTLEX word frequency, number of letters, number of syllables, orthographic Levenstein Distance), all of which are known to reliably predict lexical decision performance. In subsequent models, we then added one of the candidate composite variables to the model and use Bayes Factors to quantify the evidence in favour of each. In Table 2, we report R-squared change for each model comparison, to allow comparisons with other megastudies in the literature (e.g., Pexman, Muraki, Sidhu, Siakuluk & Yap, 2019).

Results

Overall, we found that the sensorimotor norms reliably predicted lexical decision performance for both response times and accuracy, and in both the English Lexicon and British Lexicon datasets. Each of the six composite measures accounted for a significant amount of additional variance, over and above the basic model of lexical variables (see Table 2). Log Bayes Factors for each variable (ranging from 50 to 228 for zRT, and from 29 to 138 for accuracy) revealed very strong support for their inclusion in the models. Minkowski 3 was the best performing measure in both the ELP and BLP datasets, while PCA, although still considerably improving model fit over the basic model, was the weakest performing composite measure.

Subsequently, using Minkowski 3 as the best predictor, we also found that the inclusion of action effector ratings improved model fit over and above perceptual ratings alone (see Table 3). Furthermore, adding action effector ratings provided better model fit for both ELP and BLP datasets, across both reaction time and accuracy measures.

In summary, these findings replicate the finding that perceptual information is a good predictor of people's performance in lexical decision tasks, provides new support for the utility of action effector experience in modelling cognitive performance, and shows that the findings generalise over more than one largescale data set.

General Discussion

We present a set of almost 40,000 words, normed for perceptual and action strength across 11 dimensions. The first study shows that these sensorimotor norms provide a rich dataset, with the data revealing complex patterns between various dimensions. The second study provides support for the utility of modality-specific and effector-specific sensorimotor information in modelling human performance in classic psycholinguistic tasks.

While these norms extend earlier modality-specific norms, they also quantify important new relations, such as between specific effectors and particular perceptual modalities, as well as including often ignored perceptual dimensions, such as interoception Connell, Lynott & Banks, 2018). What's more, we show that effector-specific information is also predictive of data from lexical decision tasks, over and above using perceptual-specific information alone. These findings provide evidence for a broad role for perceptual and action information in terms of their possible involvement in conceptual representations and their recruitment during cognitive processes.

A notable difference in the new set of norms is the identification of a different single composite variable that could be used in place of the full multi-dimensional vector. In the previous sets of norms (Connell & Lynott, 2012), Maximum Perceptual Strength (i.e., the highest value of any single dimension) was identified as the best single value predicting lexical decision data. In the current analyses, although max strength continued to perform very well, it was outperformed by the Minkowski 3 measure. This is an interesting pattern to emerge, as Minkowski 3 has previously been identified as an optimal parameter when modelling the integration of multiple perceptual cues (To, Baddeley, Troscianko, & Tolhurst, 2011), suggesting greater weighting to higher value dimensions. To and colleagues provided evidence that Minkowski values around 3 actually represent a general principle for perceptual integration, and may reflect the summation of neural responses to perceptual stimuli.

The current norms provide a rich source of information, and provide lexical coverage that reflects a grown adult's conceptual system. As such, we hope that they will provide many avenues for further research. There is much scope for combining the current norms with other data sets to provide even broader coverage of the human conceptual system. These and other data could then be useful for predicting human performance in a diverse array of cognitive tasks. With the increased size of the norms, they may be amenable to some machine learning techniques, for example to acquire semantic representations that could be used in robotics, or perhaps as diagnostic tools (as has been used by Kernot, Bossomaier & Bradley, 2019). Those interested in linguistics, could further investigate the role of grammatical differences in people's sensorimotor experience, and there are also opportunities to extend these norms to other languages and populations, which will enable researchers to consider cross-cultural similarities and individual differences.

Acknowledgments

This research was supported by the Leverhulme Trust, UK (RPG-2015-412, to DL, LC & MB) and the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement No. 682848 to LC. We thank Rosie Meers and Peidong Mei who provided help with data collection and quality control during the project, and David Balota for permitting us to include data from the English Lexicon Project in our deposited analysis files.

			Elexicon Project (N=22,297)			British Lexicon Project (N=11,768)		
Measurement	Stage	Specification	$logBF_{10}$	R^2	ΔR^2	$logBF_{10}$	R^2	ΔR^2
RT	0	Lexical predictors		0.591			0.485	
	1	PCA	54.532	0.593	0.002	50.747	0.490	0.005
		Summed strength	142.813	0.596	0.005	136.651	0.497	0.012
		Max strength	165.420	0.597	0.006	175.912	0.501	0.016
		Minkowski 10	177.86	0.598	0.007	193.544	0.502	0.017
		Euclidean	192.129	0.598	0.007	210.165	0.503	0.018
		Minkowski 3	**202.551**	**0.598**	**0.007**	**228.285**	**0.505**	**0.020**
Accuracy	0	Lexical predictors		0.237			0.286	
	1	PCA	43.377	0.241	0.004	29.613	0.290	0.004
		Summed strength	69.975	0.242	0.005	92.785	0.298	0.012
		Max strength	88.240	0.244	0.007	105.252	0.299	0.013
		Euclidean	93.375	0.244	0.007	131.714	0.302	0.016
		Minkowski 10	93.828	0.244	0.007	114.398	0.300	0.014
		Minkowski 3	**102.067**	**0.245**	**0.008**	**138.688**	**0.303**	**0.017**

Table 2 Bayesian linear regression results for Elexicon and British Lexicon Projects lexical decision data (Study 2). Lexical predictors were added to a null model at Stage 0 (LogSUBTLEX-US word frequency, orthographic length, number of syllables and orthographic Levenshtein distance)

			Elexicon Project (N=22,297)			British Lexicon Project (N=11,768)		
Measurement	Model	Specification	$logBF_{10}$	R^2	ΔR^2	$logBF_{10}$	R^2	ΔR^2
RT	Null	Lexical predictors		0.591			0.485	
	BF_{10}	Minkowski 3 perception	141.337	0.596	0.005	182.839	0.501	0.016
	BF_{20}	Minkowski 3 action	149.563	0.597	0.006	138.873	0.497	0.012
	BF_{21}	Comparison of simple effects	8.226			-43.966		
	BF_{31}		65.757	0.599		47.366	0.505	
Accuracy	Null	Lexical predictors		0.237			0.286	
	BF_{10}	Minkowski 3 perception	66.951	0.242	0.005	98.251	0.298	0.012
	BF_{20}	Minkowski 3 action	81.215	0.243	0.006	105.771	0.299	0.013
	BF_{21}	Comparison of simple effects	14.264			7.520		
	BF_{31}		37.815	0.245		46.023	0.304	

Table 3 Bayesian linear regression results for Elexicon and British Lexicon Projects lexical decision data. As above, Lexical predictors were added to a null model at Stage 0.

References

Amsel, B. D., Urbach, T. P., & Kutas, M. (2012). Perceptual and motor attribute ratings for 559 object concepts. *Behavior Research Methods*, *44*(4), 1028-1041.

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., ... & Treiman, R. (2007). The English lexicon project. *Behavior research methods*, *39*(3), 445-459.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences, 22*, 577–660.

Binder, J. R., Conant, L. L., Humphries, C. J., Fernandino, L., Simons, S. B., Aguilar, M., & Desai, R. H. (2016). Toward a brain-based componential semantic representation. *Cognitive neuropsychology*, *33*, 130-174.

Boulenger, V., Hauk, O., & Pulvermüller, F. (2008). Grasping ideas with the motor system: semantic somatotopy in idiom comprehension. *Cerebral cortex, 19(8)*, 1905-1914.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. Behavior Research Methods, 46(3), 904–11.

Chen, I. H., Chao, Q., Wang, S., Long, Y., & Huang, C. R. (2017). Exclusivity and competition of sensory modalities: evidence from Mandarin synaesthesia. *International Cognitive Linguistics Conference*, Tartu, Estonia.

Connell, L., & Lynott, D. (2010). Look but don't touch: tactile disadvantage in processing modality-specific words. *Cognition, 115*, 1-9.

Connell, L., & Lynott, D. (2012). Strength of perceptual experience predicts word processing performance better than concreteness or imageability. *Cognition*, *125*(3), 452-465.

Connell, L., & Lynott, D. (2014). I see/hear what you mean: Semantic activation in visual word recognition depends on perceptual attention. *Journal of Experimental Psychology: General, 143*, 527-533.

Connell, L., & Lynott, D. (2015). Embodied semantic effects in visual word recognition. *Foundations of embodied cognition*, *2*, 71-89.

Connell, L., & Lynott, D. (2016). Do we know what we're simulating? Information loss on transferring unconscious perceptual simulation to conscious imagery. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(8), 1218.

Connell, L. M., Lynott, D. J., & Banks, B. (2018). Interoception: the forgotten modality in perceptual grounding of abstract and concrete concepts. *Philosophical Transactions B: Biological Sciences*.

Đurđević, D. F., Popović Stijačić, M., & Karapandžić, J. (2016) A quest for sources of perceptual richness: Several candidates. In S. Halupka-Rešetar and S. Martínez-Ferreiro (Eds.) *Studies in Language and Mind* (pp. 187-238). RS, Novi Sad.

Fernandino, L., Conant, L. L., Binder, J. R., Blindauer, K., Hiner, B., Spangler, K., & Desai, R. H. (2013). Where is the action? Action sentence processing in Parkinson's disease. *Neuropsychologia, 51(8),* 1510-1517.

Hauk, O., Johnsrude, I. & Pulvermüller, F. (2004). Somatotopic representation of action words in human motor and premotor cortex. *Neuron, 41*, 301-307.

Kaschak, M. P., Zwaan, R. A., Aveyard, M., & Yaxley, R. H. (2006). Perception of auditory motion affects language processing. *Cognitive Science*, *30*(4), 733-744.

Kernot, D., Bossomaier, T., & Bradbury, R. (2019). The stylometric impacts of ageing and life events on identity. *Journal of Quantitative Linguistics*, *26*(1), 1-21.

Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2012). The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior research methods*, *44*(1), 287-304.

Lynott, D. & Connell, L. (2009). Modality exclusivity norms for 423 object properties. *Behavior Research Methods, 41*, 558-564.

Lynott, D., & Connell, L. (2013). Modality exclusivity norms for 400 nouns: The relationship between perceptual experience and surface word form. *Behavior Research Methods, 45*, 516-526.

Meier, J. D., Aflalo, T. N., Kastner, S., & Graziano, M. S. (2008). Complex organization of human primary motor cortex: a high-resolution fMRI study. *Journal of neurophysiology, 100(4)*, 1800-1812.

Miklashevsky, A. (2018). Perceptual Experience Norms for 506 Russian Nouns: Modality Rating, Spatial Localization, Manipulability, Imageability and Other Variables. *Journal of psycholinguistic research*, *47*, 641-661.

Pexman, P. M., Muraki, E., Sidhu, D. M., Siakaluk, P. D., & Yap, M. J. (2019). Quantifying sensorimotor experience: Body–object interaction ratings for more than 9,000 English words. *Behavior research methods*, 1-14.

Pulvermüller, F., Hauk, O., Nikulin, V. V., & Ilmoniemi, R. J. (2005). Functional links between motor and language systems. *European Journal of Neuroscience*, *21*, 793-797.

Smith, L., & Gasser, M. (2005). The development of embodied cognition: Six lessons from babies. *Artificial life*, *11*(1-2), 13-29.

Speed, L. J., & Majid, A. (2017). Dutch modality exclusivity norms: Simulating perceptual modality in space. *Behavior research methods*, *49*(6), 2204-2218.

Spence, C., Nicholls, M. E., & Driver, J. (2001). The cost of expecting events in the wrong sensory modality. *Perception & Psychophysics*, *63*(2), 330-336.

Tillotson, S. M., Siakaluk, P. D., & Pexman, P. M. (2008). Body—object interaction ratings for 1,618 monosyllabic nouns. *Behavior Research Methods*, *40*(4), 1075-1078.

To, M. P. S., Baddeley, R. J., Troscianko, T., & Tolhurst, D. J. (2010). A general rule for sensory cue summation: evidence from photographic, musical, phonetic and cross-modal stimuli. *Proceedings of the Royal Society B: Biological Sciences*, *278*(1710), 1365-1372.

Vinson, D. P., & Vigliocco, G. (2008). Semantic feature production norms for a large set of objects and events. *Behavior Research Methods*, *40*(1), 183-190.

Winter, B., Perlman, M., Perry, L. K., & Lupyan, G. (2017). Which words are most iconic?. *Interaction Studies*, *18*(3), 443-464.

Winter, B., Perlman, M., & Majid, A. (2018). Vision dominates in perceptual language: English sensory vocabulary is optimized for usage. *Cognition*, *179*, 213-220.

Zwaan, R. A., & Taylor, L. J. (2006). Seeing, acting, understanding: motor resonance in language comprehension. *Journal of Experimental Psychology: General, 135*, 1-11.

Does predictive processing imply predictive coding in models of spoken word recognition?

James S. Magnuson (james.magnuson@uconn.edu)

Monica Li (monica.li@uconn.edu)

Sahil Luthra (sahil.luthra@uconn.edu)

Heejo You (hee_jo.you@uconn.edu)

Rachael Steiner (rachael.steiner@uconn.edu)

Psychological Sciences & CT Institute for the Brain and Cognitive Sciences, U. Connecticut, Storrs, CT 06269-1020

Abstract

Pervasive behavioral and neural evidence for *predictive processing* has led to claims that language processing depends upon *predictive coding*. In some cases, this may reflect a conflation of terms, but predictive coding formally is a computational mechanism where only deviations from top-down expectations are passed between levels of representation. We evaluate three models' ability to simulate predictive processing and ask whether they exhibit the putative hallmark of formal predictive coding (reduced signal when input matches expectations). Of crucial interest, TRACE, an interactive activation model that does not explicitly implement prediction, exhibits both predictive processing and model-internal signal reduction. This may indicate that interactive activation is functionally equivalent or approximant to predictive coding, or that caution is warranted in interpreting neural signal reduction as diagnostic of predictive coding.

Keywords: prediction; predictive coding; language; computational modeling; neural networks

Prediction in spoken language processing

Listeners often predict upcoming information in spoken language. They anticipate upcoming phonemes based on lexical expectations (Grosjean, 1980; Allopenna, Magnuson, & Tanenhaus, 1998), and upcoming words based on lexical, syntactic, and/or discourse expectations (Altmann & Kamide, 2007; Magnuson et al., 2008; Strand et al., 2018). There is also neural evidence consistent with prediction. Indeed, many ERP studies test the magnitude and timing of responses to expectation violations, including responses that precede complete bottom-up specification. Despite difficulties replicating one classic example (Delong, Urbach, & Kutas, 2005 vs. Nieuewland et al., 2018), a large number of studies support varying degrees of prediction (for reviews, see Kuperberg & Jaeger, 2015; Hickock, 2012).

Evidence for predictive *processing* (PP) is often considered evidence for *predictive coding* (PC), and there may be instances where these terms are conflated and treated synonymously. PC, however, is a computational formalism enaling efficient coding by comparing bottom-up inputs to predictions from a top-down model and passing forward (and backward) only *deviance from prediction* (Rao & Ballard, 1999). This deviance is the novel *information*; sending bottom-up details would be redundant when predicted by higher-level expectations. Thus, formal PC predicts reduced feedforward and feedback signal when inputs conform to top-down expectations. In light of several reports of neural signal reduction when word-level expectations are met (e.g., Blank & Davis, 2016; Gagnepain, Henson, & Davis, 2012), we next consider what evidence for PP and PC implies for models of spoken word recognition (SWR).

Implications for models of spoken word recognition

First, even without considering sentence-level contexts (beyond the scope of current models), models of SWR must be able to simulate attested word level PP. Intuitively, some models might do this readily (e.g., a simple recurrent network [SRN; Elman, 1990] trained to predict the next phoneme given the current phoneme), while others may not. For example, Gagnepain et al. (2012) suggest that the interactive activation model, TRACE (McClelland & Elman, 1986), may be inconsistent with PC because they describe its primary mode as *competitive* rather than *predictive*.

Second, given growing neural evidence consistent with formal PC (reduced neural signal when expectations are confirmed vs. violated; e.g., Sohoglu & Davis, 2016) we can also ask whether a model of SWR exhibits this hallmark of PC: internal signal reduction when expectations are confirmed. This leads us to two questions for models of SWR. (1) Do models with explicit prediction (e.g., SRNs) and without explicit prediction (e.g., TRACE) simulate PP? (2) If so, do they show hallmarks of formal PC (model-internal signal reduction when expectations are confirmed)? To address these questions, we will compare three models.

Model comparisons

Our simulations are based on human experiments by Gagenpain et al. (2012). In those experiments, there were three critical stimulus types: an **Original** word (e.g., *formula*), a **Trained** nonword (e.g., *formubo*), and an **Untrained** nonword (e.g., *formuty*). In the examples, we have underlined letters corresponding to the critical phonemes. Prior to training, both Trained and Untrained nonwords differ from expectations at 1-3 phonemes from offset; this position follows the *deviation point*. The critical question is how the system responds at the phoneme(s) following the deviation point before a training phase and after. In the training,

participants get extensive exposure to the *trained* nonwords. Prior to training, Gagnepain et al. (2012) found reduced neural activity in left superior temporal gyrus following the deviation point for Original items vs. both types of nonwords. Following training (and sleep), Trained items showed the same reduction relative to Untrained items as real (Original) words. In the following sections, we examine whether each of 3 models is able to simulate PP (sensitivity to expectations at the deviation point) in simulations of this paradigm, and whether they exhibit the hallmark of PC: signal reduction when top-down lexical expectations are met.

Model 1: Predictive Cohort

Gagnepain et al. (2012) used a simple mathematical model to generate predictions for their experiment. We call their model "predictive cohort" because it simply looks up the set (cohort) of words that remain consistent with the phoneme-by-phoneme input for each item. For example, given *formula*, at position 1, all /f/-initial words are possible and the prediction for position 2 is the frequency-weighted probability distribution of each phoneme following /f/. As input progresses, probability distributions narrow. For *formula* (/formjul^/), by /u/ at position 6, very few possibilities remain (*formula, formulaic, formulation*) and all predict /l/. Gagnepain et al. (2012) derived positional prediction error for the three item types. Given a prediction of 1.0 for /l/ at position 7, *formula* would garner zero prediction error, while prediction error would be high for *formubo* and *formuty*.

The logic is that a formal PC implementation would pass back a prediction of /l/ at position 7 given the input for positions 1-6, and therefore pass forward a very weak signal given *formula,* where the prediction error is low, compared to the nonword cases. Note that prediction error is not an internal signal in this model; it is a derived term meant to stand in for computations that would occur in formal PC.

Methods

Materials We implemented predictive cohort as described by Gagnepain et al. We selected 37.6k words ≤ 12 phonemes long from the English Lexicon Project (ELP; Balota et al., 2007). Critical items were 54 Original-Trained-Untrained triples from Gagnepain et al. (mean length: 6.3 phonemes). Deviation points were 1-3 positions before offset.

Procedure We conducted two suites of simulations with all 54 x 3 = 162 items. In *pretraining*, the lexicon was restricted to 37.6k real words; thus, the Original items were words, and the Trained and Untrained items were nonwords. *Post-training*, Trained items were simply added to the lexicon, changing the positional probability distributions embedded in the lexicon (as done by Gagnepain et al.). For each simulation, we computed predicted probability distributions at each position, and calculated implied prediction error.

Results are presented in Fig. 1. Consistent with PP, the probability for Original items continues to increase beyond the deviation point at Pretraining, and probabilities also increase for Trained items Post-training. Because error is summed over all phonemes, the maximum is 2.0 (e.g., if predicted values for /l/ and /b/ were 0.8 and 0.0, but the input were 0.0 for /l/ and 1.0 for /b/, summed error would be 1.8). Error plots do *not* reflect model-internal information. Rather, error is meant to approximate what the forward signal would be *if a formal PC model were implemented.* Thus, while the predictive cohort model is able to exhibit PP, it does not inherently exhibit PC. Of course, a fully-implemented PC model would show such signal reduction.

Model 2: Simple Recurrent Network

The second model we tested was a Simple Recurrent Network (SRN; Elman, 1990). An SRN would seem likely to naturally produce PP, given that an SRN is typically trained to predict

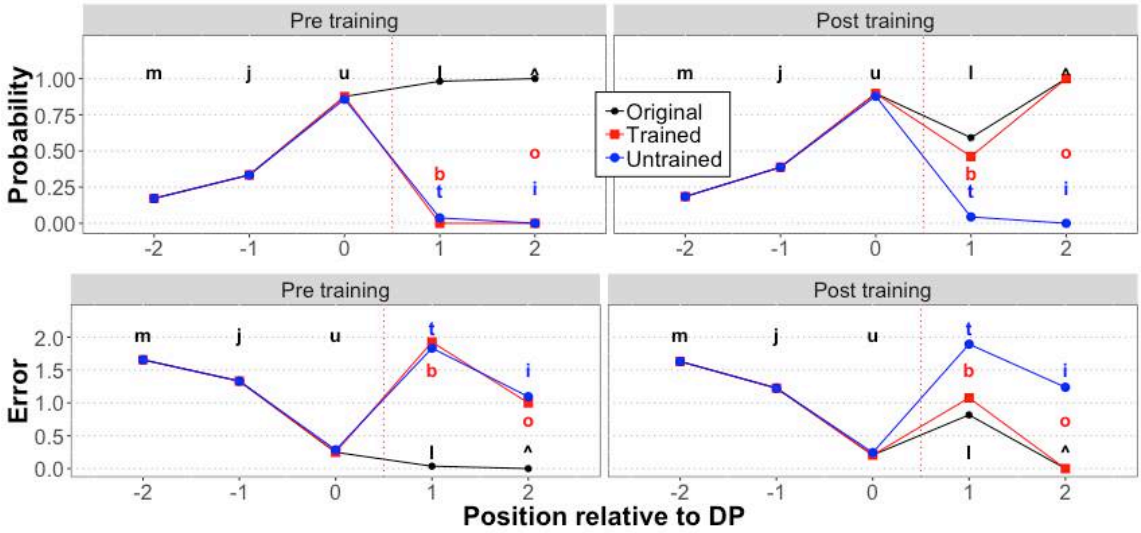


Figure 1: Predicted phoneme-by-phoneme probabilities (top) and derived errors (bottom), pre- (left) and post- (right) training for the Predictive Cohort model. The X-axis is position relative to the deviation point (allowing us to align results for all items). The dashed lines between positions 0 and 1 indicate the deviation point.

the next item in a series. We created an SRN with localist phonemic inputs (41 nodes, one for each phoneme) with forward connections to 200 hidden units with forward connections to 41 localist phonemic output nodes. The hidden nodes feed an exact copy of their states with a 1-cycle time delay to *context* nodes, which feedback to hidden nodes (providing a memory sensitive to multi-step contingencies).

Methods

Materials We used the same materials as for Model 1.

Procedure The model was presented with a continuous series of phonemes constructed by randomizing the order of the 37.6k words and presenting each phoneme-by-phoneme, without any break or indication of a word boundary. The network was trained using backpropagation of error to predict the next phoneme. At each time step, output activations were compared to the desired output pattern (1.0 for the following phoneme, 0.0 for all others). Backpropagation allows "blame" to be assigned to all connections in the network (i.e., to calculate how small changes to all weights could alter the network such that if the same input sequence were applied again, the network would come closer to the target pattern).

After approximately 2000 epochs (each epoch is 1 pass through all 37.6k words in random order), error plateaued (aggregated over small batches of words). This does not mean error rate was uniform. Rather, output activations come to resemble the probability distributions calculated by the predictive cohort model (Model 1). Thus, error is relatively high near word onset and diminishes as the input progresses.

For the *pretraining* test of the model, only the 37.6k words selected from the ELP for Model 1 were included. Because the SRN is a learning model, we were able to actually train the model on Trained items. The 54 Trained items were presented in novel random orders for 50 epochs. This number of instances was sufficient for the model to achieve Original-level accuracy with Trained items without impairing the model's ability to process items already in its lexicon.

Results are in Fig. 2, and are similar to those from Model 1, but with output activations for relevant phonemes. Error indicates the summed error over all 41 output phoneme units. Like Model 1, the SRN exhibits PP pre- and post-training in that phonemes from trained items become more probable after training. Also like Model 1, though, note that error is not a model-internal value; it is calculated externally. Model-internal signals (here, activations) *do not exhibit the reduced-signal hallmark of PC*. Instead, activations are *higher* when expectations are met (when the input sequence corresponds to a word in the lexicon). Thus, even the most intuitively predictive model of SWR one might propose (short of a formal PC model) – an SRN – does not inherently exhibit PC.

Some might disagree with this analysis, since SRNs are trained using backpropagation of error, and these error terms could be considered to be passed back through the model, even if error is typically not passed during tests and is not necessary for a trained SRN to function. We might counter that backpropogation is model-external (the procedure is not part of the network dynamics of an SRN; adjustments to weights are imposed on the network, rather than an emergent property. One might contrast this with Hebbian learning, where weight changes occur through biologically-inspired interactions among nodes. On the other hand, while backpropogation may not have a direct analog in biology, functionally-equivalent, neurally-plausible mechanisms are not far-fetched (Lillicrap & Santoro, 2019). It may be sensible, then, to consider the error signal in an SRN as a feedback signal, in which case SRNs show the PC hallmark of relative signal reduction when inputs match expectations.

Model 3: TRACE

TRACE (McClelland & Elman, 1986) is an interactive activation model: a neurally-inspired, parallel-distributed processing model with feedforward connections from inferior to superior levels (features→phonemes→words) and lateral inhibition within levels. It also has feedback from words to constituent phonemes. As mentioned earlier, TRACE may

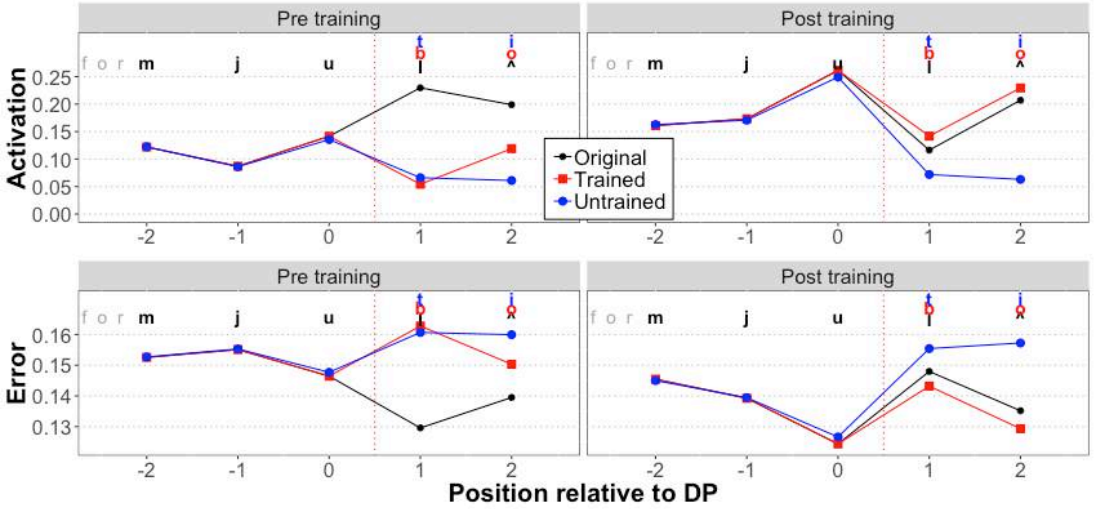


Figure 2: Phoneme-by-phoneme SRN output activations indicating how strongly the model predicted each upcoming phoneme (top) and those activations converted to error scores over time (bottom).

not seem to be a predictive model (e.g., Gagnepain et al. [2012] describe TRACE as having a primarily competitive mode of processing, dominated by lateral inhibition). However, word→phoneme feedback in TRACE provides a generative model (McClelland, 2013; Magnuson et al., 2018): as features and phonemes consistent with a specific word are presented, the lexical node for that word form sends increasingly strong feedback to *all* its constituent phonemes, including those that have not yet occurred. This allows graded pre-activation of phonemes (as a function of how strongly *expected* they are due to feedback from one or more words). But is there any possibility that TRACE exhibits the hallmark of PC – reduced signal when expectations are confirmed vs. violated?

Methods

Materials We used the original 212-word TRACE lexicon, wherein we identified 15 six-phoneme words on which to base item sets. From each set, we created 2 nonwords by changing the final two phonemes. For example, from the Original /art^st/ (*artist*) we created /art^da/ and /art^pi/.

Procedure Simulations were conducted with all 15 (set) x 3 (item type) items. We tracked activations of phonemes and words over time as well as the total amount of activation (and inhibition) flow between and within levels during each simulation. For pre-training, the lexicon consisted only of the TRACE lexicon, including the 15 Original items. For post-training, the 15 Trained items were added to the lexicon.

Results We begin by comparing lexical activations for each item type (Original, Trained, Untrained) pre- and post-training (Fig. 3). Pre-training, we see significantly weaker Original activation when input ends with final phonemes of Trained or Untrained items. Post-training, we see a decrease in Original activation given Untrained input, and a massive decrease given Trained input. This is because Trained items are now words in the lexicon; with clear input, Trained items strongly activate and inhibit their Original counterparts. The post-training panel in Fig. 3 includes a red line marked with an open red square; this indicates activation of Trained items given corresponding input. This line is directly on top of the Original line; since both items are words in the lexicon, clear corresponding input drives both similarly.

Next, consider the phoneme level (Fig. 4). Activations of phonemes one position beyond the deviation point are plotted for the Original word, as well as replaced phonemes in the case of the Trained and Untrained nonwords. In Fig. 4, we can see differences in the lines with open symbols that achieve high activation. These correspond to activations of replaced phonemes (/d/ in /art^da/ or the /p/ in /art^pi/). Pre-training, the highest activation is achieved for the phoneme in penultimate position in the Original word, thanks to support from both bottom-up input and top-down lexical feedback. There is only a slight disadvantage for the replaced phonemes; given clear bottom-up input, phonemes will be strongly activated, even in the absence of lexical support. Post-training, with Trained items added to the lexicon, the 'replaced' phoneme in a Trained item achieves nearly identical activation as a phoneme in an Original item, since both receive lexical support.

To address PP, Fig. 5 zooms in on the regions delineated with dashed squares in Fig. 4. Pre-training, the activation of the Original phoneme is higher than that for replaced phonemes beginning ~12 cycles prior to the deviation point. Phoneme activations from cycles ~18 to ~33 (just past the deviation point, indicated by the dashed vertical line) are driven nearly exclusively by top-down feedback. Bottom-up input begins to override feedback just after the deviation point. At this point, when the input has a replaced phoneme (one of the nonworeds), the activation of the Original phoneme drops, while activation of the replaced phonemes when they are actually the input (dashed lines, open symbols) jumps dramatically. Post-training, we see a lexical advantage for phonemes after the deviation point for both Original and Trained items (for Trained items, activations after the deviation point is slightly less due to a small trend for those items to have lower transitional probability in the lexicon, even when they have been added to the lexicon). In summary, training elicits clear PP: increased activation of critical phonemes prior to the deviation point.

Next, let's consider PC, which could manifest as reduced feedforward or feedback signal when expectations are confirmed; to be fully consistent with PC, both the feedforward and feedback signal would have to be reduced when expectations are met. However, the standard in many cognitive neuroscience studies is that any evidence of signal reduction is taken as evidence for PC. We therefore tracked the total amount of activation flowing between levels

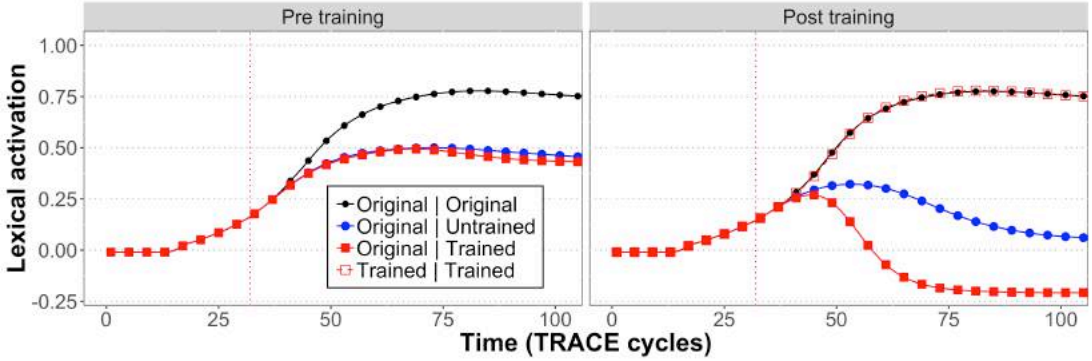


Figure 3: Lexical activations in TRACE before and after 'training'. Note that 'Trained | Trained' is only valid post-training.

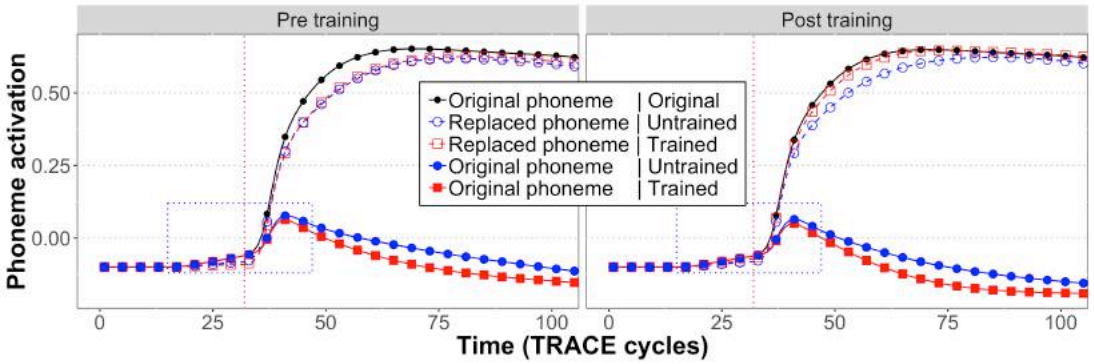


Figure 4: Activations of critical phoneme (following deviation point) in TRACE.

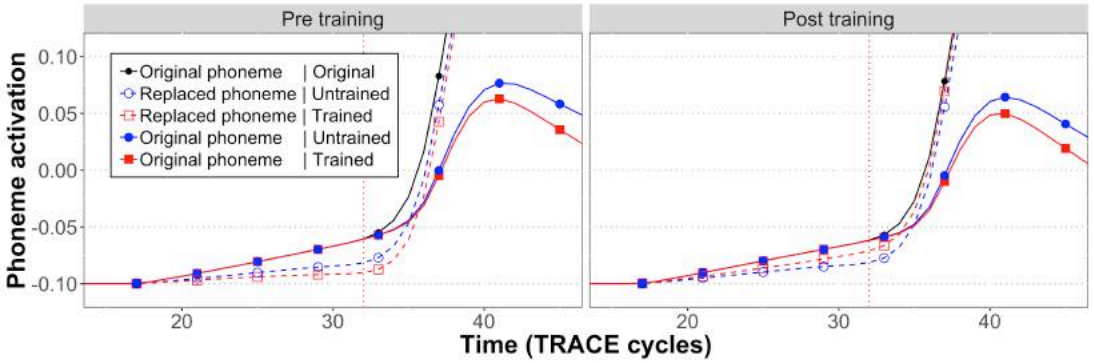


Figure 5: Zoomed view of critical time period from Figure 4.

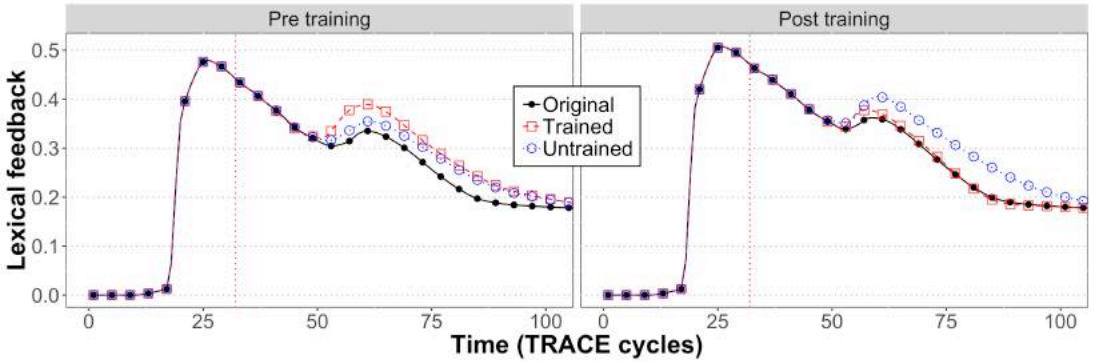


Figure 6: Total lexical feedback over time in TRACE, showing robust signal reduction when expectations are met.

(forward and backward) and within levels (lateral inhibition), looking for any signal reduction.

Two activation indices were reduced when expectations were met: word→phoneme feedback (Fig. 6) and lateral inhibition (absolute value). The latter shows virtually the same pattern as Fig. 6, but we omit it due to length constraints and challenges in interpreting a reduction in a signal with negative valence. In Fig. 6, total lexical feedback is *lowest* when expectations are met (for Original words pre- or post-training, as well as for Trained items post-training). This is because when an unexpected phoneme occurs, Original items already have strong support and continue to send substantial feedback. Additional feedback comes from words partially activated by replaced phonemes (any word unit containing the unexpected phoneme *aligned* with the unexpected phoneme[s] would get activated; e.g., a word unit for *piano* aligned at position 5 overlaps with the /pi/ of /art^pi/]). This follows from the *total* amount of feedback actually being less when one word can strongly dominate and inhibit other words; there can actually be *more* total feedback when many words are weakly activated. Thus, only TRACE, the model one might have predicted to be least likely to exhibit PC, shows a *model-internal signal reduction* often considered *diagnostic* of PC in cognitive neuroscience.

Discussion

All three models tested – predictive cohort, an SRN, and TRACE – exhibit PP. The first two showed model-internal signal *increases* when expectations were met. While these increases can be converted to predicted error, this takes place outside the current instantiation of these models (though see our earlier discussion of backpropagated error in SRNs). TRACE shows model-internal signal *reduction* when expectations are confirmed, in the form of lesser top-down

lexical→phoneme feedback.

This raises the possibility that interactive activation (as implemented in TRACE) may provide a generative model that is functionally equivalent (or functionally approximant) to a Bayesian generative model (McClelland, 2013) or even PC. Addressing this question will require the development of explicit, formal PC models of SWR based on formalisms like those introduced by Rao and Ballard (1999). This is a tall order; such a model must work on over-time inputs (if not real speech), must be validated with a moderately large lexicon (at least hundreds of words), and must be comprehensively compared to other models, such as TRACE.

There are promising starts in this direction. For example, Yildiz et al. (2013) have reported a PC model of SWR that operates on real speech. However, this model was limited to a 10-word vocabulary (names for the digits 0 to 9). Another promising example comes from Blank and Davis (2016), who implemented simple network models of SWR with lexical→phoneme feedback that was either multiplicative (as in TRACE) or subtractive (one possible interpretation of PC). Both models correctly simulated one experiment, but their subtractive feedback model correctly predicted neural signal reduction in a second experiment where the multiplicative model predicted signal increase (but with radical parameter changes required to fit the two experiments; in one, they ran models for more than 300 cycles, while for the second, they ran models for only 1 cycle). This sort of work, along with comprehensive tests of models on at least moderately large vocabularies (to verify that the models are consistent with known facts about SWR), are needed to advance understanding of the potential role for PC in SWR.

In the absence formal PC models, we must exercise caution when interpreting neural signal reduction. Though our results indicate that TRACE exhibits model-internal signal reduction, it remains an open question whether interactive activation is indeed functionally equivalent or approximant to PC. Similarly, it may be premature to consider evidence of a reduction in neural signal when expectations are met as *diagnostic* of PC.

Acknowledgments

Supported by NSF 1754284, NSF IGERT 1144399, & NSF NRT 1747486 (PI: J.S.M.).

References

Allopenna, P.D., Magnuson, J.S. & Tanenhaus, M.K. (1998) Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language* 38, 419-439.

Altmann, G.T.M. & Kamide, Y. (2007). The real-time mediation of visual attention by language and world knowledge. *J. Memory & Language,* 57, 502-518.

Blank, H., & Davis, M. H. (2016). Prediction errors but not sharpened signals simulate multivoxel fMRI patterns during speech perception. *PLoS Biology*, *14*: e1002577.

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, *8*(8), 1117-1121.

Elman, J. (1990). Finding structure in time. *Cognitive Science*, *14*, 179-211.

Frauenfelder, U. H. & Peeters, G. (1998). Simulating the time course of spoken word recognition: an analysis of lexical competition in TRACE. In J. Grainger & A. M. Jacobs (Eds.), *Localist connectionist approaches to human cognition* (pp. 101-146). Mahwah, NJ: Erlbaum.

Gagnepain, P., Henson, R.N., Davis, M.H. (2012) Temporal predictive codes for spoken words in human auditory cortex. *Current Biology, 22*(7), 615-621.

Grosjean, F. (1980). Spoken word recognition processes and the gating paradigm. *Percep. & Psychophyss, 28,* 267-283.

Hickock, G. (2012). The cortical organization of speech processing: Feedback control and predictive coding the context of a dual-stream model. *Journal of Communication Disorders, 45,* 393-402.

Kuperberg, G.R. & Jaeger, T.F. (2015). What do we mean by prediction in language comprehension? *Language & Cognitive Neuroscience, 31(1)*, 32-59.

Lillicrap, T.P. & Santoro, A. (2019). Backpropogation through time and the brain. *Current Opinion in Neurobiology, 55*, 82-89.

Magnuson, J. S., Mirman, D., Luthra, S., Strauss, T., & Harris, H. (2018). Interaction in spoken word recognition models: Feedback helps. *Frontiers in Psychology*, 9:369.

Magnuson, J.S., Tanenhaus, M.K., & Aslin, R.N. (2008). Immediate effects of form-class constraints on spoken word recognition. *Cognition, 108(3),* 866-873.

McClelland, J. L. (2013). Integrating probabilistic models of perception and interactive neural networks: A historical and tutorial review. *Frontiers in Psychology, 4*, 503.

McClelland J.L. & Elman, J.L. (1986) The TRACE model of speech perception. *Cognitive Psychology* 18, 1-86.

Nieuewland, M.S., Politzer-Ahles, S., Heyselaasr, E., Segaert, K., Darley, E., Kazanina, N. … Huettig, F. (2018). Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. *eLife, 7*: e33468. doi:10.7554/eLife.33468.

Rao, R. & Ballard, D. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience, 2(1),* 79-87.

Sohoglu, E. & Davis, M.H. (2016). Perceptual learning of degraded speech by minimizing prediction error. *Proc. Nat'l Academy Sci., 113(12)*, E1747-56.

Strand, J., Brown, V., Brown, H., & Berg, J. (2017). Keep listening: Grammatical context reduces but does not eliminate activation of unexpected words. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 44,* 962-973. doi: /10.1037/xlm0000488

Yildiz, I.B., von Kriegstein, K., & Kiebel, S.J. (2013). From birdsong to human speech recognition: Bayesian inference on a hierarchy of nonlinear dynamics systems. *PLoS Computational. Biology, 9(9)*: e1003219.

Individual differences in reading experiences: The roles of mental imagery and fantasy

Marloes Mak
Radboud University, Nijmegen, Netherlands

Roel M. Willems
Radboud University, Nijmegen, Netherlands

Abstract

It is well established that readers form mental images when reading a narrative. The influence of mental imagery on the way people experience stories is however still unclear. In two experiments reported here, participants received instructions aimed at encouraging or discouraging mental imagery before reading literary short stories. After reading, participants answered questions about their reading experiences. The results from the first experiment suggested an important role of mental imagery in determining reading experiences. However, the results from the second experiment showed that individual trait differences in how imaginative participants are predicted reading experiences much better than guided mental imagery. Moreover, the role of mental imagery did not extend to aspects of the reading experience other than mental imagery. The implications of these results for the relationship between mental imagery and reading experiences are discussed.

Hands in mind: learning to write with both hands improves inhibitory control, but not attention

Mukesh Makwana (mukesh@cbcs.ac.in), **Biswajit Boity** (bib231@mail.harvard.edu), **Prasanth P.** (prasanth@students.iiserpune.ac.in), **Amogh Sirnoorkar** (sirnoorkaramogh@gmail.com), **and Sanjay Chandrasekharan** (sanjay@hbcse.tifr.res.in)

The Learning Sciences Research Group, Homi Bhabha Centre for Science Education,
Tata Institute of Fundamental Research, Mumbai, 400088, India

Abstract

Embodied cognition theories predict that changing motor control would change cognitive control, as cognition is considered to emerge from action in this theoretical approach. We tested this prediction, by examining the attention and cognitive control capabilities of a group of school students (12-13-year-olds) trained to write using both hands (experimental group, N=28), compared to a group of age-matched children (control group, N=33) who did not receive such training. The key tasks used were the attentional network test (ANT) task and the hearts and flowers (HF) task. Results from the ANT task showed that there was no significant difference in the three attentional networks between the groups. However, results from the HF task showed that the experimental group had better inhibitory control. This second result provides support to the embodied cognition prediction that cognitive control and motor control are related, and the former can be changed to some extent by changing the latter.

Keywords: Embodied Cognition; Handedness; Executive Functions; Motor Control.

Introduction

Embodied theories of cognition argue that cognitive processes are shaped by the way the body interacts with the environment (Glenberg et al., 2013). This is because the brain evolved to control coordinated actions in multicellular creatures, and cognitive and affective processes evolved later, to guide action. This evolutionary view is partly based on the work of Rudolfo Llinas (2001), who argues that "A nervous system is only necessary for multicellular creatures (not cell colonies) that can orchestrate and express active movement -- a biological property known as "motricity"".

The embodied cognition position would thus predict that changes in motor control would lead to changes in cognition and affect, as the latter are derivative systems. Supporting this view, a series of studies have linked the manipulation of motor system with changes in executive functions of children as young as 5-year-olds (Stein et al., 2017; Rueda et al., 2012). Motor functions have also been shown to influence inhibition and cognitive flexibility (Livesey et al., 2006). Further, executive functions have been shown to be related to physical activity (Campbell et al., 2002; Becker et al., 2014), and motor functions (Livesey et al., 2006; Davis et al., 2011) in both kindergartners and older children (Stein et al., 2017). These effects of motor functions on cognitive functions are supported by the fact that the biological development of both motor and cognitive functions are closely related (e.g., Sibley and Etnier, 2003), and cognitive functions are stimulated and required when learning and executing new motor skills (Best, 2010; Diamond, 2000).

A related empirical thread has examined the role of handedness, and lack of consistent handedness, on cognitive and affective abilities (Casasanto, 2009; Coren, 1992). It has been shown that handedness (ranging from strongly right-handed to strongly left-handed) predicts whether electrical excitation via transcranial direct current stimulation causes an increase or decrease in the experience of approach-related emotions. (Brookshire and Casasanto, 2012). In such studies, handedness is typically considered a marker of motor training, and thus not explored further. The development of handedness and its results, particularly how training to use both hands at a young age affects cognitive abilities such as attention and executive functions, has not been much explored.

To understand the relation between handedness development and cognitive abilities, we conducted a study based on the Attentional Network Test (ANT) Task and Hearts and Flowers Task (HF). Both ANT and HF tasks are standard psychological tasks that reliably provide independent measures for different attentional networks (i.e. alerting, orienting, and executive control; Fan et al., 2002; Rueda et al., 2004) and executive function components (i.e. working memory, inhibition, and flexibility; Davidson et al., 2006). These tasks were selected as they tap into different types of inhibitory control. The ANT task involves resolving conflict of the stimulus-stimulus type (e.g. both the target and the distractors are visual stimuli in a flanker task). The HF task involves resolving conflict of the stimulus-response type (e.g. overcome the default propensity to make a response matching the stimulus location). These tasks were administered to two student groups (experimental, control) from two schools. The experimental group students studied in a school that provided a school-wide basic training to write using both hands. The control group studied in a school that had no such training.

Methods

Demographics Owing to the uniqueness (only 2 identified schools in India) of the experimental group

school[1] in imparting training to write with both hands, this school was assigned to the experimental group by default. Another school which was similar in all other aspects (other than the training) was assigned as the control group school. The criteria for selecting a relevant and comparable control group were multifold: similar parents' profession and annual incomes (migrant laborers), similar school infrastructure (both low-income private schools), similar curricular and extracurricular aspects (including training on physical activities), same age groups (12-13-year-olds), and an equivalent number of languages exposed to the students (each group was familiar with at least three different languages). Apart from the training to write with both hands, the other major difference between the groups was the location: the experimental group school was a village school while the control school was in a slum in the heart of a metropolitan city. The experimental group students were familiar with Kannada, English, and Hindi while the control group students had familiarity with Hindi, English, and Marathi.

Training Process An ethnographic study of the experimental group school showed that students start their training by using their dominant hand to write during the first six months after being admitted into the school. Thus, students may start the training as early as 3 years (kindergarten) or as late as 12 years (7th grade) depending on when they join the school. They are then instructed to use the non-dominant hand for the next six months. The training starts with making lines and curves, then progresses to writing alphabets and numbers, and concludes with words and sentences. Instruction is given in small, often mixed-age groups (4-5 students per group). A teacher first demonstrates the techniques by writing on a blackboard. She then allows students to practice on the board and on their notebooks. After 3rd grade, however, these practice sessions are considered an extracurricular activity (optional) that students are free to pursue before/after school hours. Students who participated in our experiment had an average of 2.1 years (S.D. = 0.69 years) experience writing with both hands.

Attention Network Task (Child Version)

Participants 27 students (Mean age = 12.5 years, S.D. = 0.57, 12 male, 15 female) from the experimental group and 32 students (Mean age = 12.8 years, S.D. = 0.80, 19 male, 13 female) from the control group participated in the experiment. The school principal and teachers were communicated in advance about the purpose and nature of the study. Participants were explained in detail about the consent process (including the option to discontinue whenever they wanted) and the tasks, following which signed consent was obtained from each participant and school principal prior to the study. All communication between participants and experimenters was in the language that participants understood most clearly (Kannada for the experimental group and Hindi for the control group).

Stimuli and apparatus The stimuli were presented using Inquisit 5's ANT (Child version), a commercial application by Millisecond[2], run on laptops (all 15.6-inch screens: 34.54 cm x 19.41 cm) with Windows 10 OS. Participants viewed the screen from a distance of 53 cm (approx.), and responded to the stimuli by pressing two keys on the laptop keypad.

The stimuli consisted of a central fixation (+ sign) that appeared at the beginning of each trial, presented against a constant blue-green (0, 255, 255) screen background. This was followed by one of four warning cue conditions: no-cue, center-cue, double-cue, or spatial-cue. A black dot (cue) appeared in the center instead of the + sign for the center-cue condition. The double-cue condition involved the cue being presented on target locations both above and below the + sign. In the spatial-cue condition, the cue appeared either above or below the + sign. The no-cue condition did not provide any warning about the forthcoming stimulus, while the center-cue and double-cue conditions warned the participants when the target will appear. The spatial-cue condition alerted as well as indicated the locations of the target stimulus (see Fan et al., 2002; Rueda et al., 2004 for more details). (See link for a schematic diagram)

The target stimuli were a yellow color-filled line drawing of either a single fish or an array of five fish that appeared above or below the central fixation. Each fish projected a visual angle of 1.6° and the contours of adjacent fish were at a distance of 0.06° from each other. The total visual angle projected by the array of 5 fish was 8.4°. The target stimuli were presented at 1.08° above or below the central fixation.

Procedure Participants were instructed to focus on the hungry central fish and feed them by pressing the "E" (when the fish facing left) or "I" (when the fish facing right) key. While receiving the instructions, participants were asked clarifying questions to ensure that they understood the context and task requirements.

Each session lasted ~30 minutes and consisted of one practice block (24 trials) and three experimental blocks (48 trials). The trials in the experimental blocks had one of the following combinations: 4 cue conditions (no-cue, center-cue, double-cue, spatial-cue) x 3 flanker conditions (congruent, incongruent, neutral) x 2 target stimuli positions (up, down) x 2 target stimuli directions (left, right). (See link for a schematic diagram). The order of the trials was random.

Each trial sequence had the following trial structure: fixation period with randomly chosen presentation time (between 400-1600 ms), followed by a warning cue for 100 ms, followed by a fixation period of 400 ms after the disappearance of the cue, and concluding with the appearance of the target stimulus, either alone or along with

[1] https://www.youtube.com/watch?v=PDVDw60sG5c

[2] https://www.millisecond.com/

flankers for 1700 ms. Participants had to respond within this 1700 ms duration, after which the stimulus disappeared. The inter-trial interval was set at 1000 ms. Participants received audio-visual feedback for both correct and incorrect responses during the practice block. The experimental blocks did not have any feedback.

Hearts and Flowers Task

Participants The participants for this experiment were the same as those in the ANT experiment.

Stimuli and apparatus The stimuli were presented through Inquisit 5's Hearts and Flowers Task (Child-friendly version). Other apparatus remained the same as the previous task.

The stimuli consisted of a central fixation (+ sign) followed by a heart or a flower (target stimuli) that appeared on the left or the right of the fixation cross. The fixation sign was constantly present on the white background screen while the stimuli appeared in red. The hearts/flowers appeared at a visual angle of 5.6° to the left or right of the central fixation. A heart subtended a visual angle of 2.04° whereas a flower subtended a visual angle of 2.16° (Davidson et al., 2006). (See link for a schematic diagram)

Procedure Each session lasted for ~20 minutes and consisted of three sequential blocks: Congruent-only block (Hearts as stimulus) followed by Incongruent-only (Flowers as stimulus) followed by Mixed (both Hearts and Flowers as stimulus). Each block had 8 practice trials and 20 experimental trials. The experimental trials were initiated only if participants reached an accuracy of minimum 75% in the practice trials. Participants received audio-visual feedback during the practice trials for both correct and incorrect responses. The experimental trials did not have any audio-visual feedback.

Each trial sequence in the experiment block started with the presentation of the target stimulus. The maximum response time was 5000 ms (for congruent-only and incongruent-only) and 6000 ms (for mixed block). The inter-trial interval was set at 1000 ms. Participants were required to press "A" for a heart appearing on the left of the + sign and "L" for a heart appearing to the right of the + sign (congruent trials). For the flower stimulus, participants had to press the "A" key for a flower appearing on the right of the + sign and "L" for the flower appearing to the left of the + sign (incongruent trials). Both congruent-only and incongruent-only blocks had the stimulus on the left of + sign for ten trials and on the right for the remaining ten, appearing in random order. In the mixed block, there were 10 hearts (5 right, 5 left) and 10 flowers (5 right, 5 left) that appeared in a random order, with the following constraint: a maximum of 3 trials of the same type (congruent or incongruent) could be run consecutively, and the number of switch trials (i.e. from congruent to incongruent and vice-versa) would vary from trial to trial (with a minimum of 6 per trial).

Edinburgh Handedness Inventory (EHI)

26 participants (Mean age = 12.5 years, S.D. = 0.58, 11 male, 15 female) from the experimental group and 28 (Mean age = 12.9, S.D. = 0.85, 18 male, 10 female) from the control group were provided with the EHI questionnaire (Oldfield, 1971). Participants were asked to respond orally to a 12-item questionnaire, using one of five responses: always right, usually right, both equally, usually left, always left. Since the participants were not familiar with surveys, concrete everyday examples were provided for clarification of each questionnaire item, along with the response categories. Participants were asked to act out how they would perform each of the items in the questionnaire while reporting their response. The Laterality Quotient (LQ) score for each participant was calculated as below:

$$LQ = 100\frac{(\sum\text{number of positives} - \sum\text{number of negatives})}{(\sum\text{number of positives} + \sum\text{number of negatives})}$$

Where "always right" was assigned ++ (2 positives), "usually right" was + (1 positive), "both equally" was +- (one positive, one negative), "usually left" was - (1 negative), and "always left" was -- (2 negatives). An LQ score closer to +100 denoted strongly right-handed, -100 denoted strongly left-handed, and a 0 represented an equal preference for both hands in the tasks. Scores other than the above represent the use of both hands but not in an equal measure.

Results

One-way ANCOVA using group (experimental, control) as the fixed factor and age and gender as covariates on LQ scores showed a significant main effect of group [$F(1,50) = 6.481$, $p = 0.014$, $\eta_p^2 = 0.115$]. Results revealed that the experimental group ($M = 65.62$, $S.D. = 23.63$) had significantly lower LQ score compared to control group ($M = 83.14$, $S.D. = 13.75$) (see Fig 1a). This suggests that training to use both hands might have influenced the participants to use both their hands for motor activities other than writing, as the LQ score in EHI is calculated by taking into consideration the handedness preference in various everyday general motor activities.

Attention Network Task

Overall Accuracy Analysis An 80% overall accuracy criterion led to the elimination of five participants (1 in experimental and 4 in the control group), giving 54 participants' data for further analysis. JASP software was used to perform statistical analysis. One-way ANCOVA using group (experimental, control) as the fixed factor and age and gender as the covariate on accuracy showed a main effect of group [$F(1,50) = 6.52$, $p = 0.014$, $\eta_p^2 = 0.115$] while the effect of gender and age were not significant, suggesting that the experimental group ($M = 95.78$, $S.D. = 2.90$) had significantly higher overall accuracy in the ANT task, compared to control group ($M = 92.71$, $S.D. = 5.44$)

(see Fig. 1b). However, when participants' LQ score was used as a covariate, it could explain the difference in overall accuracy between the two groups [$F(1,51) = 4.446$, $p = 0.04$, $\eta_p^2 = 0.08$]. To further understand the relationship between LQ scores and overall accuracy scores, Pearson's correlation analysis was performed. Results indicated a significant negative association between LQ score and overall accuracy ($r(52) = -0.383$, $p = 0.004$), suggesting that participants with low LQ scores performed better compared to those with high LQ scores. Low LQ scores indicate more usage of both hands for everyday motor activities, whereas high LQ scores indicate more usage of a single or dominant hand.

Flanker type x Cue type x Group Analysis We performed 3 (*Flanker type:* congruent, incongruent, neutral) x 4 (*Cue type:* no-cue, center-cue, double-cue, spatial-cue) x 2 (*Group:* experimental, control) mixed ANOVA with flanker type and cue type as within subject factors and group as between subject factor on the median RTs. The main effect of flanker type [$F(1.46, 76.27) = 77.67$, $p < .001$, $\eta_p^2 = 0.599$] and cue type [$F(2.56, 133.55) = 66.30$, $p < .001$, $\eta_p^2 = 0.56$] were found to be significant. However, the main effect of group was not significant [$F(1, 52) = 0.701$, $p = 0.406$, $\eta_p^2 = 0.013$]. Planned comparisons showed that participants were significantly faster [$t(53) = 3.52$, $p < 0.001$] in the congruent flanker condition ($M = 625.33$ ms, $S.E. = 15.23$) compared to the incongruent one ($M = 697.46$ ms, $S.E. = 16.71$), showing the standard flanker effect (Ericksen & Ericksen, 1974).

Planned comparisons for different cue conditions showed that participants were significantly [$t(53) = 4.026$, $p < 0.001$] faster in the double-cue condition ($M = 633.08$ ms, $S.E. = 15.39$) compared to the no-cue condition ($M = 675.97$ ms, $S.E. = 15.08$) demonstrating the typical alerting effect of the cue on RT. Also, the difference between center-cue ($M = 645$ ms, $S.E. = 14.38$) and spatial-cue ($M = 596.63$ ms, $S.E. = 13.86$) was significant [$t(53) = 4.541$, $p < 0.001$], demonstrating the orienting effect of the spatial-cue. The difference between center-cue and no-cue was also significant [$t(53) = 2.907$, $p < 0.05$] suggesting that even the single cue had an alerting effect, though the magnitude was less compared to the double cue. There was a significant interaction between cue type and group [$F(2.56, 133.55) = 3.04$, $p = 0.031$, $\eta_p^2 = 0.055$]. Post-hoc analyses showed no significant difference between experimental and control group for each cue type.

Alerting, Orienting, and Conflict Analysis The measures of effects for the three networks were calculated by subtracting different cue type and flanker type conditions. The altering effect was calculated by subtracting the double-cue condition RT from the no-cue condition RTs. The orienting effect was calculated by subtracting the spatial-cue condition RT from center-cue condition RTs. The conflict or executive function effect was calculated by subtracting congruent flanker condition RT from incongruent flanker condition RT (Rueda et al., 2004).

Pearson's correlation analysis revealed that there was no significant correlation between any of these three networks [alerting and orienting, $r(52) = 0.081$, $p = .562$; alerting and conflict, $r(52) = 0.173$, $p = .211$; orienting and conflict, $r(52) = 0.212$, $p = .124$], thus supporting the finding in previous studies that the three networks are independent.

A series of one-way ANOVA were performed to examine the effect of group, age, and gender on the mean of median RTs and errors for the alerting, orienting and conflict quotients. None of the comparisons reached significance, except for a group difference in percentage error for alerting quotients [$F(1,49) = 4.891$, $p = 0.032$, $\eta_p^2 = 0.091$].

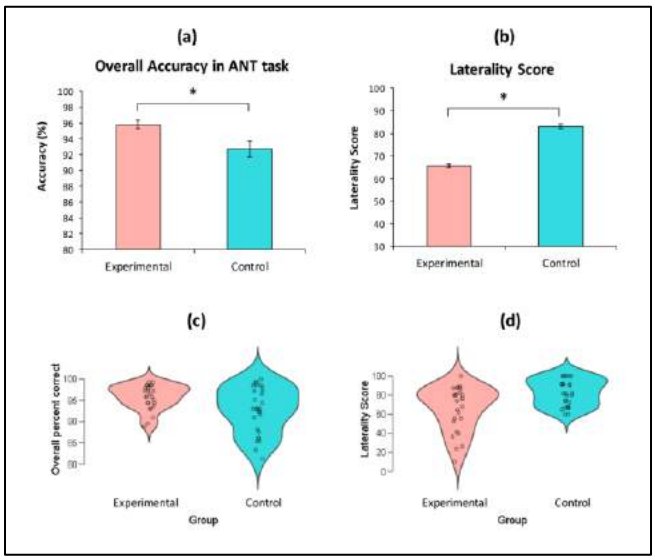


Figure 1. The top panel shows the bar plot for (a) overall accuracy in the ANT task, and (b) LQ score for both groups. The bottom panel shows the corresponding violin plot. Error bar represents S.E. of mean. * indicates <0.05

Hearts and Flowers Task

Overall Accuracy and Reaction Time An 80% overall accuracy criterion led to the elimination of five participants (3 in experimental, 2 in control). Additionally, two participants from the control group didn't complete the task. This resulted in a total of 52 participants' data for further analysis. One-way ANCOVA using group (experimental, control) as the fixed factor and age, gender, and LQ score as covariates showed no significant difference in overall accuracy between the two groups [experimental group 92%, control group 91.4%; $F(1, 48) = 0.225$, $p = 0.637$, $\eta_p^2 = 0.005$]. Similarly, there was no significant difference in overall mean RT as a function of group [$F(1, 48) = 0.388$, $p = 0.536$, $\eta_p^2 = 0.008$; experimental group, $M = 636.3$ ms, $S.E. = 28.11$; control group, $M = 660.9$ ms, $S.E. = 26.48$].

Block x Group Analysis Two-way 3 (*block type:* congruent, incongruent, mixed) x 2 (g*roup:* experimental, control) mixed ANOVA with block type as within subject factor and group as between subject factor on the mean RTs showed a significant main effect of block type [$F(1.64, 83.83) = 164.33$, $p < .001$, $\eta_p^2 = 0.763$]. However, the main effect of group and interactions were not significant. Planned comparisons showed the expected significant differences between congruent, incongruent and mixed block types. That is, participants were significantly faster in the congruent block ($M = 497.1$ ms, *S.E.* = 18.20) compared to both incongruent block [$M = 590.22$ ms, *S.E.* = 20.26; $t(52) = 3.642$, $p < .001$] and mixed blocks [$M = 899.194$ ms, *S.E.* = 31.51; $t(52) = 15.72$, $p < .001$]. Also, the difference between incongruent block and mixed block was significant [$t(52) = 12.08$, $p < .001$].

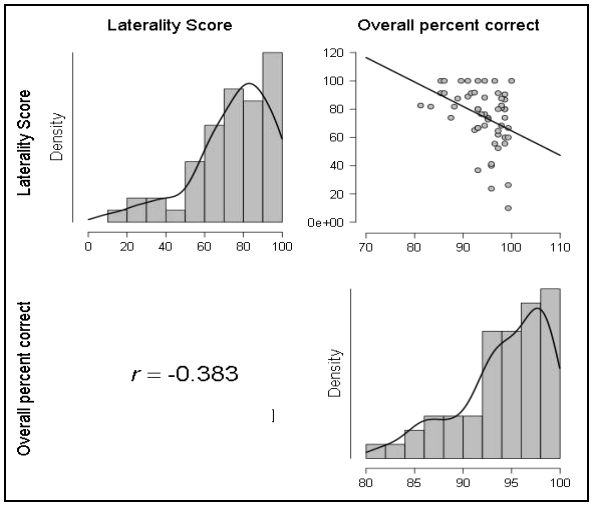


Figure 2. The correlation matrix between overall accuracy in ANT and LQ score.

Another similar 3 (block) x 2 (group) mixed ANOVA on the mean accuracy revealed a significant main effect of block type [$F(1.41, 71.99) = 54.02$, $p < .001$, $\eta_p^2 = 0.514$]. However, the main effect of *group* [$F(1, 51) = 0.119$, $p = 0.731$, $\eta_p^2 = 0.002$] and interaction [$F(1.41, 71.99) = 2.704$, $p = 0.091$, $\eta_p^2 = 0.05$] were not significant. Planned comparisons for the different block types revealed significant differences between all the groups [congruent vs. incongruent, $t(52) = 2.933$, $p < 0.05$; congruent vs. mixed, $t(52) = 8.624$, $p < 0.01$; incongruent vs. mixed, $t(52) = 5.69$, $p < 0.01$]. Results replicated the expected effects of block type on reaction time and accuracy, wherein participants became slower and less accurate as the task demand increased from congruent to incongruent to mixed block.

Inhibitory Control and Cognitive Flexibility To measure inhibitory control, the congruent block RTs (working memory) were subtracted from the incongruent block (working memory + inhibition control), and to measure cognitive flexibility the incongruent block RTs were subtracted from the mixed block (working memory + inhibition + cognitive flexibility). The switching score was obtained by subtracting the non-switch trials from the switch trials in the mixed block.

Pearson's correlation analysis was used to evaluate the association between inhibitory control, cognitive flexibility, and switching scores. Results showed a negative association ($r(52) = -0.464$, $p < .001$) between inhibitory control and cognitive flexibility, whereas other correlations were not significant [inhibitory control and switching, $r(52) = 0.162$, $p = 0.247$; cognitive flexibility and switching, $r(52) = -0.138$, $p = 0.325$]. Further analysis is needed to understand this relationship. (See link for correlation matrix)

We performed a series of one-way ANCOVAs for all the three subtraction scores, with group as the between subject variable and laterality as covariate. The only significant main effect of group was in the inhibition control for both reaction time [$F(1, 51) = 8.749$, $p = 0.005$, $\eta_p^2 = 0.146$] and accuracy [$F(1, 51) = 6.431$, $p = 0.014$, $\eta_p^2 = 0.112$]. These results suggest that participants with training to write with both hands were better in inhibitory control, compared to participants in the control group (see figure 3). (See http://handedness.surge.sh/ for more tables, figures, and a detailed analysis)

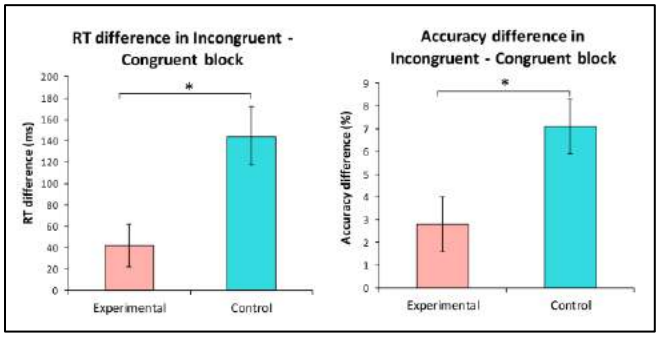


Figure 3. The bar plot displaying the difference in RT between incongruent and congruent block (left), and difference in accuracy between incongruent and congruent block (right).

Discussion

In both ANT and HF tasks, standard effects were observed, suggesting that the tasks were executed successfully. The central result was the significant group difference observed in inhibitory control, as measured by the HF task. The ANT results revealed that training to write with both hands improved overall accuracy, without significantly hampering response time. However, participants from the experimental group were slower in response (though not significantly) compared to the control group, suggesting a kind of speed-accuracy trade-off. This group difference in accuracy covaried with the differences in the LQ score. Laterality and overall accuracy were negatively correlated, suggesting that participants with low LQ scores had higher accuracy, while participants with high LQ scores had lower accuracy.

Overall, the writing training did not significantly improve performance in any of the attentional networks. Similar results were obtained by Rueda et al., (2012), where computer-based attentional training provided to school children did not significantly improve the alerting and orienting networks. However, there was an enhancement in the executive network, which overlapped with the domain of training provided to participants. Similarly, even though we did not find any significant group difference for alerting, orienting and conflict networks in ANT, we observed a significant group difference in inhibitory control as measured by the HF task.

The Hearts and Flowers task can be viewed as a child version of the Simon task, which tracks a standard tendency to inhibit the prepotent impulse when the stimulus location overlaps with the response side. People with better inhibitory control would be able to resolve this conflict faster, and would thus be less prone to the Simon effect. It has been shown that playing video games, but not visual training, improves inhibitory control, and reduces the cost of Simon effect (Hutchinson, Barrett, Nitka, &Raynes, 2016).

These results suggest that learning to write with both hands could be understood as leading to the improvement of inhibitory control. However, it is not clear how this improvement in inhibitory control is related to writing with both hands. One possibility is a heightened activation model, where writing with both hands leads to both hands getting activated by motor plans for writing, and active inhibition of one is required to write with the other. This process requires, and improves inhibitory control.

This model fits well with our ethnographic data, which showed that when students were asked to write a novel paragraph using both their hands, they did so with only with one hand at a time i.e., they did not write *simultaneously* with both hands. Some students wrote one character with one hand and the next with the other. Others wrote a word or multiple characters of a word with one hand before moving to write the next word or the remaining characters with the other hand. Based on this data and the heightened activation model, learning to write with both hands could be understood as having effects similar to learning to speak in more than one language, where all the known languages get activated when planning to speak. The speaker thus needs to inhibit the other activated languages when choosing to speak in one, and also when trying to understand speech, as many candidate words will be activated. This choosing process requires, and supports, heightened inhibitory control, whose effects would be seen in other control situations. Supporting this model, bilingualism studies show that executive function improves through learning more than one language (see Bialystok, 2001, 2011). Although these studies show that bilingualism improves cognitive control (the "bilingual advantage"), there exists a debate regarding the main effect (Anton et al., 2014). Some studies show that bilingual training only provides a domain-specific advantage (i.e. improves inhibition and control of perceptual or stimulus-stimulus type representations), and no drastic improvement in inhibition and control of motor or habitual or stimulus-response type representations (Blumenfeld & Marian, 2014; Martin-Rhee & Bialystock, 2008; Poarch, 2018). This fits well with our findings, as which show benefits of motor training on inhibitory control in the HF task but not in the ANT task. The growing literature on the cognitive control effects of *changes* in motor control (Stein et al., 2017; Stuhr et al., 2018; for a review see Diamond & Ling, 2016) -- to which our study contributes -- shows that training motor control abilities might have global effects, which are reflected in tasks wider than the immediate context of training. Apart from our results on inhibitory control, our Edinburgh Handedness Inventory (to determine handedness or the level of hand preferences for various everyday motor tasks) also found that participants who had received training to write with both hands used both hands for other everyday motor tasks, suggesting that hand preferences change in a global fashion with such training. These, and related results showing the role of action in language and imagination (Pulvermuller, 2001; Glenberg, 1997), open up the possibility of using the motor system as an intervention channel, particularly to change higher-order cognitive and affective systems.

However, this intervention possibility needs to be approached with caution, as the relationship between higher-order systems (such as imagination and language) and motor control is not straightforward, as higher-order systems typically draw on, *and recombine*, many networks, including from frontal regions of the brain. Further, tasks in higher-order cognition, such as physics problem-solving, requires bringing together many cognitive components, such as reading, imagining, calculating, reasoning, etc. Whether these processes and their integration, are improved by motor control is currently unclear.

Thus, even though schools that train students to write with both hands do so with possible educational effects in mind, the results related to wider control capabilities we report here cannot be taken as an indication of training to write with both hands improving problem-solving abilities. Further studies need to be done to investigate whether such improvements could follow from motor training. While this study leveraged the opportunity provided by a particular school that trains students to write with both hands, future studies would benefit if the above experiments are conducted as part of a controlled intervention study. This study provides a good starting point in demonstrating the effect of bimanual writing on cognitive flexibility. However, a more extensive and controlled study is required to replicate as well as extend the results.

Acknowledgements

The authors thank the school founder Mr. Goudish Biradar and the Sri Ram Kannada Convent School, Kalaburagi, Karnataka and the school principal Mr. Vinoth Raj Velayuthan of Royal City English High School, Mumbai, Maharashtra along with the teaching staff of both the schools for granting permission and supporting the

studies. We also thank Prof. Geeta R. M., Sharvani Shahapurkar, Soujanya Ganig, Suman Barua, Harshit Agrawal, Durga Prasad Karnam, and Ganesh Shinde for helping with various aspects of the studies.

References

Antón, E., Duñabeitia, J. A., Estévez, A., Hernández, J. A., Castillo, A., Fuentes, L. J., & Carreiras, M. (2014). Is there a bilingual advantage in the ANT task? Evidence from children. *Frontiers in psychology, 5*, 398.

Becker, D. R., McClelland, M. M., Loprinzi, P., & Trost, S. G. (2014). Physical activity, self-regulation, and early academic achievement in preschool children. *Early Education & Development, 25*(1), 56-70.

Best, J. R. (2010). Effects of physical activity on children's executive function: Contributions of experimental research on aerobic exercise. *Developmental Review*, 30(4), 331-351.

Bialystok, E. (2001). *Bilingualism in development: Language, literacy, and cognition*. Cambridge University Press.

Bialystok, E. (2011). Reshaping the mind: the benefits of bilingualism. *Canadian Journal of Experimental Psychology*, *65*(4), 229.

Blumenfeld, H. K., & Marian, V. (2014). Cognitive control in bilinguals: Advantages in Stimulus–Stimulus inhibition. *Bilingualism: Language and Cognition, 17*(3), 610-629.

Brookshire, G., & Casasanto, D. (2012). Motivation and motor control: hemispheric specialization for approach motivation reverses with handedness. *PLoS One*, 7(4), e36036.

Campbell, D. W., Eaton, W. O., & McKeen, N. A. (2002). Motor activity level and behavioural control in young children. *International Journal of Behavioral Development*, *26*(4), 289-296.

Casasanto, D. (2009). Embodiment of abstract concepts: good and bad in right-and left-handers. *Journal of Experimental Psychology: General*, 138(3), 351.

Coren, S. (1992). Handedness, traffic crashes, and defensive reflexes. *Am. journal of public health*, 82(8), 1176-1177.

Diamond, A. (2000). Close Interrelation of Motor Development and Cognitive Development and of the Cerebellum and Prefrontal Cortex. *Child Development, 71*(1), 44–56.

Diamond, A., & Ling, D. S. (2016). Conclusions about interventions, programs, and approaches for improving executive functions that appear justified and those that, despite much hype, do not. *Developmental cognitive neuroscience, 18*, 34-48.

Davis, E. E., Pitchford, N. J., & Limback, E. (2011). The interrelation between cognitive and motor development in typically developing children aged 4–11 years is underpinned by visual processing and fine manual control. *British Journal of Psychology, 102*(3), 569-584.

Davidson, M. C., Amso, D., Anderson, L. C., & Diamond, A. (2006). Development of cognitive control and executive functions from 4 to 13 years: Evidence from manipulations of memory, inhibition, and task switching. *Neuropsychologia*, 44(11), 2037-2078.

Fan, J., McCandliss, B. D., Sommer, T., Raz, A., & Posner, M. I. (2002).Testing the efficiency and independence of attentional networks. *Journal of cognitive neuroscience*, 14(3), 340-347.

Glenberg, A. M., Witt, J. K., & Metcalfe, J. (2013). From the revolution to embodiment: 25 years of cognitive psychology. *Perspectives on Psychological Science*, 8(5), 573-585.

Glenberg, A. M. (1997). What memory is for. *Behavioral and brain sciences*, 20(1), 1-19.

Hutchinson, C. V., Barrett, D. J., Nitka, A., & Raynes, K. (2016). Action video game training reduces the Simon Effect. *Psychonomic bulletin & review*, 23(2), 587-592.

Inquisit 5 [Computer software]. (2016).

JASP Team (2018). JASP (Version 0.9).

Livesey, D., Keen, J., Rouse, J., & White, F. (2006). The relationship between measures of executive function, motor performance and externalizing behaviour in 5-and 6-year-olds. *Human Movement Science*, 25(1), 50-64.

Llinás, R. R. (2001). *I of the vortex: From neurons to self*, Cambridge, MA: MIT press.

Martin-Rhee, M. M., & Bialystok, E. (2008). The development of two types of inhibitory control in monolingual and bilingual children. *Bilingualism: language and cognition, 11*(1), 81-93.

Oldfield, R. C. (1971). The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia*, 9(1), 97-113.

Poarch, G. J. (2018). Multilingual language control and executive function: a replication study. *Frontiers in Communication*, *3*(46), 10-3389.

Pulvermüller, F. (2001).Brain reflections of words and their meaning. *Trends in cognitive sciences*, *5*(12), 517-524.

Rueda, M. R., Checa, P., & Combita, L. M. (2012). Enhanced efficiency of the executive attention network after training in preschool children: immediate changes and effects after two months. *Developmental cognitive neuroscience*, 2, S192-S204.

Rueda, M. R., Fan, J., McCandliss, B. D., Halparin, J. D., Gruber, D. B., Lercari, L. P., & Posner, M. I. (2004). Development of attentional networks in childhood. *Neuropsychologia*, 42(8), 1029-1040.

Sibley, B. A., & Etnier, J. L. (2003). The relationship between physical activity and cognition in children: a meta-analysis. *Pediatric exercise science*, 15(3), 243-256.

Stein, M., Auerswald, M., & Ebersbach, M. (2017). Relationships between motor and executive functions and the effect of an acute coordinative intervention on executive functions in kindergartners. *Frontiers in psychology*, 8, 859.

Stuhr, C., Hughes, C. M. L., & Stöckel, T. (2018). Task-specific and variability-driven activation of cognitive control processes during motor performance. *Scientific reports, 8*(1), 10811.

Something about *us*: Learning first person pronoun systems

Mora Maldonado (Mora.Maldonado@ed.ac.uk)
Centre for Language Evolution
University of Edinburgh

Jennifer Culbertson (Jennifer.Culbertson@ed.ac.uk)
Centre for Language Evolution
University of Edinburgh

Abstract

Languages partition semantic space into linguistic categories in systematic ways. In this study, we investigate a semantic space which has received sustained attention in theoretical linguistics: person. Person systems convey the roles entities play in the conversational context (i.e., speaker(s), addressee(s), other(s)). Like other linguistic category systems (e.g. color and kinship terms), not all ways of partitioning the person space are equally likely. We use an artificial language learning paradigm to test whether typological frequency correlates with learnability of person paradigms. We focus on first person systems (e.g., 'I' and 'we' in English), and test the predictions of a set of theories which posit a universal set of features (±exclusive, and ±minimal) to capture this space. Our results provide the first experimental evidence for feature-based theories of person systems.

Keywords: artificial language learning; categorization; person systems; extrapolation; typology; linguistic universals

Introduction

One of the fundamental goals of cognitive science is to understand how human languages carve up semantic space into linguistic categories. Research on the typology of categorization systems, from colour names, to noun classification and kinship terms suggests that not all systems are equally likely.

For example, despite some cross-linguistic variation, certain ways of carving up the continuous color space into linguistic categories are much more common than others. This has been argued to provide evidence for a universal basis for color categorization, reflecting properties of the human perceptual system (Kay & Regier, 2007; Zaslavsky, Kemp, Tishby, & Regier, 2018; Gibson et al., 2017). Similar arguments have been made to explain the distribution of kinship systems across languages (Kemp & Regier, 2012; Kemp, Xu, & Regier, 2018).

Here, we focus on a semantic space which has garnered substantial attention in theoretical linguistics: person systems (e.g., Zwicky, 1977; Harley & Ritter, 2002; Harbour, 2016; Ackema & Neeleman, 2018). Such systems–exemplified in pronoun paradigms (e.g. 'me', 'you', 'her')–describe how languages categorize entities as a function of their role in the context of a speech event (i.e., speaker(s), addressee(s), other(s)). Like color and kinship systems, person systems have long been observed to exhibit constrained variation.

The person space

Research on the typological distribution of person systems has hypothesized an inventory of four discrete categories: first exclusive (speaker only), first inclusive (speaker and addressee), second (addressee) and third (other) (Harley & Ritter, 2002; Cysouw, 2003; Bobaljik, 2008). The interaction with number multiplies the possible distinctions.

Here, we focus specifically on *first* person systems, as they allow us to investigate a contrast that is not instantiated by English (1^{st} inclusive vs. 1^{st} exclusive). Theories of first person systems have posited two binary features, one for person (±addressee) and one for number (±minimal) (Bobaljik, 2008; Cysouw, 2011; Harley & Ritter, 2002).[1] This two-feature system is designed to instantiate all first person categories, as illustrated in Figure 1.[2]

A language which takes advantage of the maximal 4-way contrast will have a person *paradigm* with 4 distinct forms (e.g., Ilocano pronouns). Alternatively, the contrast between some cells can be neutralized within a paradigm, in which case different cells will use the same form. Such paradigms exhibit *homophony*.

Homophony which neutralizes one of the two hypothesized features–person or number–has been called *systematic homophony* (Harbour, 2008; Baerman, Brown, Corbett, et al., 2005). For example, a paradigm that neutralizes only the person contrast (keeping the number one) would have just two pronominal forms, one for both minimal inclusive and exclusive, and another for

[1] The ±minimal feature encodes an asymmetry in status between the minimal group consisting of the speaker and addressee, and a larger group including others. This is used rather than the more intuitive singular/plural contrast to distinguish between the two inclusive categories.

[2] The two-feature system in Figure 1 is a simplification of current proposals for the complete person space (i.e. including 2^{nd} and 3^{rd} persons). Most approaches rely on the existence of at least two different person features and three number features (Bobaljik, 2008; Harbour, 2016; Bobaljik & Sauerland, 2018).

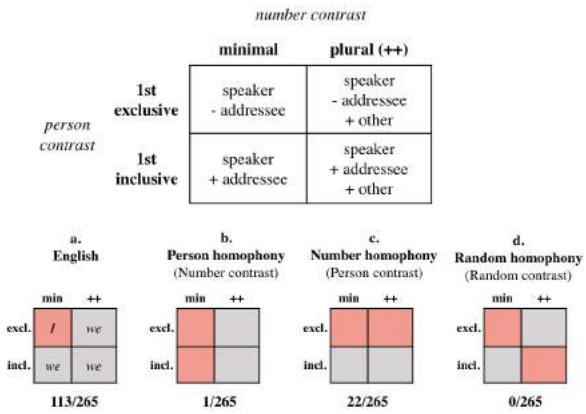


Figure 1: First person system (top) and four possible paradigms obtained by homophony (bottom, a-d) along with typological counts (Cysouw, 2003).

the two non-minimal (plural) categories ('Person homophony', Figure 1b). A paradigm that neutralizes only number (keeping the person contrast) would have one inclusive and one exclusive form ('Number homophony', Figure 1c). Homophony of both features is also possible, as in English ('we' for inclusive and exclusive plural, and minimal inclusive). Finally, a paradigm can partially neutralize one feature, for example, number homophony in the inclusive, but two distinct exclusive forms.

Random homophony patterns, not based on feature neutralization, are in principle also possible. For example, minimal exclusive and plural inclusive could share the same form, minimal inclusive and plural exclusive another ('Random homophony', Figure 1d).

Feature-based theories of person systems (cf. Figure 1) predict that systematic homophony is a natural consequence of feature-neutralization (or loss), and should arise regularly and be (easily) learnable. By contrast, they argue that there is no linguistic basis for random homophony, which is expected to arise only by historical accident, and be less readily learnable. Intuitively, there is nothing which ties together homophonous cells in a random homophony paradigm, therefore they should be less natural for learners. Notably, these theories are formulated on the basis of typological samples of person paradigms (the largest of which include <300 languages). Interestingly, while their predictions hold when considering complete paradigms, it is less clear for first person systems. According to Cysouw (2003), most of the possible paradigms for the 4-cell 1^{st} person space have not been documented. Among those that are attested, the skew is zipfian: the English-like pattern (Figure 1a) is by far the most frequent, the next most common systems have partial or complete number homophony (e.g., Figure 1c). Unexpectedly, both random *and* person (only) homophony appear to be very rare (see also Sauerland & Bobaljik, 2013; Baerman et al., 2005).

Experimental goals and predictions

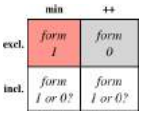
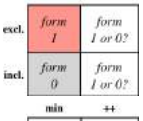
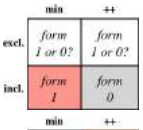
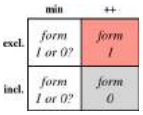
The principal goal of this paper is to set out a method for investigating person systems experimentally. The first step we take here is to test whether some first person paradigms are more natural than others. Our measure of naturalness will be learners' likelihood of inferring the relevant paradigm. We will test three main hypotheses: the first is a sanity-check, and the second two are derived from the theories outlined above in combination with the typology.

The first hypothesis is that, all things equal, learners generally assume a new language to have the same structure as their own. Learners in our experiment are native English speakers, therefore this predicts that they will be most likely to infer a first person paradigm that is English-like in its homophony pattern. The second hypothesis is that typologically frequency is correlated with learnability (Culbertson, 2018). This predicts that learners will be more likely to infer a paradigm characterized by number homophony than person or random homophony.[3] The third hypothesis is that there is a universal set of person/number features, as in (3), which learners are sensitive to regardless of their native language. This predicts that natural homophony patterns–which neutralize one specific feature–should be more likely to be inferred by learners than random homophony.

To test these predicted patterns of inference, we use an artificial learning paradigm in which learners are required to generalize (or extrapolate) from ambiguous evidence (a.k.a 'Poverty-of-the-Stimulus design, Wilson, 2006; Culbertson & Adger, 2014). Participants are trained on two cells of a first person paradigm, and must then use the forms they have learned to express all the cells in the paradigm. In other words, they must extrapolate the forms they have learned to the remaining two categories. For example, if a learner is trained on two distinct forms for exclusive minimal (speaker only) and exclusive plural (speaker plus others), they will be tested on the two remaining categories that include the addressee. If they use the plural form for both new categories, then they have inferred an English-like paradigm. Different patterns of extrapolation would indicate person or random homophony (as described in detail in Table 1).

[3] In principle this also predicts that learners should be most likely to infer an English-like paradigm, since this pattern of person *and* partial-number homophony is much more common. However, we cannot test this prediction with English-speaking learners.

Table 1: Summary of conditions.

Condition		Critical training set	Critical held-out set	Compatible paradigms
(1)		excl.min, excl.++	incl.min, incl.++	**English-like**, Person Hom., Random Hom.
(2)		excl.min, incl.min	excl.++, incl.++	**English-like**, Number Hom., Random Hom.
(3)		incl.min, incl.++	excl.min, excl.++	**Person Hom.**, Random Hom.
(4)		excl.++, incl.++	excl.min, excl.min	**Number Hom.**, Random Hom.

Methods

This experiment, including all hypotheses, predictions, and analyses, was preregistered.[4]

Participants

A total of 332 English-speaking adults were recruited via Amazon Mechanical Turk (female = 152). Participants were paid 2 USD for their participation which lasted approximately 15 mins. Per our pre-registered plan, participants were excluded if (a) their accuracy rates during exposure training were below 80%, or (b) their accuracy rates for trained cells during the test phase were below 66%. This resulted in analysis of 181 participants (Conditions 1: 46; Condition 2: 50; Condition 3: 49; Condition 4: 36).[5]

Design

Participants were randomly assigned to one of four possible conditions, summarized in Table 1. Conditions differed in which subset of two first person categories was trained (*critical training set*) and held-out (*critical held-out set*). This determines which alternative full paradigms are consistent with the two categories participants have learned. Conditions 1 and 2 are consistent with an English-like pattern (or systematic homophony). Conditions 3 and 4 are each consistent with one type of systematic homophony, and random homophony.

All participants were additionally exposed to another four pronominal forms which mapped into the second and third person singular and plural categories. These forms were used as controls.

Table 2: Highlighted family members for each category.

Category	Highlighted set
1st excl.min	speaker
1st incl.min	speaker, addressee
1st excl.pl	speaker, other(s)
1st incl.pl	speaker, addressee, other(s)
2nd sg.	addressee
2nd pl.	addressee, other(s)
3rd sg.	one other
3rd sg.	multiple others

Materials

The language consisted of 6 different pronoun forms, used for the control categories (2nd sg/pl, 3rd sg/pl), plus the critical first person forms. For each participant, these 6 lexical items were randomly drawn from a list of 8 CVC non-words created following English phonotactics: 'kip', 'dool', 'heg', 'rib', 'bub', 'veek', 'tosh', 'lom'. Items were presented orthographically.

To express the pronoun meanings, we commissioned a cartoonist to draw scenarios involving a family of three sisters and their parents. Each family member has a clearly-defined role in the conversational context. The two older sisters are speech act participants (in all scenarios they are either speaker or addressee). The third (little) sister was spatially close, but never a speech act participant. The parents were seated in the background (serving as additional others).

Pronouns were used as one-word answers to questions like 'Who will be rich?'. Meanings were expressed by highlighting subsets of family-members, as in Table 2.[6] An example illustrating 1st incl.min is provided in Figure 2. All questions were English interrogative sentences of the form 'Who will...?', which were randomly drawn from a list of 60 different tokens.

Procedure

Participants were first introduced to the family, including the names of the sisters, and were told they were going to see the sisters playing with a hat that had two magical properties: whoever wore it could see the future but would also talk in a mysterious ancestral language. Participants were instructed to figure out the meanings of words in this new language. They were given a hint that the words were not names, and an example trial with an English pronoun ('her').[7]

[4]Maldonado, M., & Culbertson, J. (2019, January 29). Extrapolation to bipartitions. https://doi.org/10.17605/OSF.IO/J2RCN.

[5]High accuracy rates on trained critical items were required because extrapolation of these forms is not interpretable if participants have not learned them.

[6]To ensure that forms were not associated with specific quantities, in all non-minimal categories, pronouns randomly referred to two or three individuals. Third person singular meanings were always expressed with a female other.

[7]In addition, the speaker and addressee roles switched during the experiment to highlight that the words were de-

Figure 2: Trial example in the test phase for the inclusive minimal category.

The experiment had two phases. In the training phase, participants were taught the pronouns in the control and training sets (6 person categories). Each training trial had two parts: a scene where a question is asked, and a scene where the question is answered with a pronoun form in the language (cf. Figure 2). There were 12 training trials (2 repetitions per form). Participants were given feedback on their answers.

After this initial training phase, participants were given an initial test of the trained forms. Each trial consisted of a question and answer scene, as in training, followed by a 'what if?' scene in which a new set of individuals was highlighted. They were asked to pick the correct word for that meaning among two options. There were 16 such trials (2 repetitions per control form, 4 per critical training form). Participants were given feedback on their answers. The test phase involved a similar procedure but included trials for the two remaining critical categories, i.e. the held-out set. This phase consisted of 48 trials (6 repetitions per form). Participants received no feedback during this phase.

The experimental session lasted approximately 15 minutes. The order of presentation of meanings was fully randomized within training and test phases for each participant.

Results

Recall that participants were taught two pronominal forms (coded as forms 1 and 0), which they had to use to describe both a critical trained set of first person meanings, and a held-out set. Figure 3 shows the proportion of trials on which participants chose the 'form 1' (pronoun) for each first person category during the test phase. Choice of the same form across categories indicates homophony. A visual inspection of Figure 3 suggests that participants in Conditions 1 and 2 are consistently using one form for 1st excl.min., and the other for the remaining three categories: this indicates inference of an English-like paradigm. Participants in Conditions 3 and 4 appear somewhat noisier in their responses, however, distinct patterns are evident. In Condition 3, one form is used for the two minimal categories, and the other for the plurals (consistent with person homophony). In Condition 4, one form is used for the two exclusive categories, and, at least for some participants, the other form is used for the two inclusive categories (consistent with number homophony).

Following our pre-registered plan, we conducted three analyses to evaluate these patterns statistically[8]:

Prediction 1: Preference for L1 pattern Figure 3 suggests that participants in Conditions 1 and 2 are more likely to infer a stable pattern, as predicted if an L1-like pattern is easier to learn. To test this, we used *joint entropy* of the two held-out categories to measure how variable participants' are in their mapping of the taught forms for the two held-out categories. The entropy value for a given category indicates the degree of uncertainty or variability in the responses. The joint entropy will therefore reveal the level of variability or uncertainty for each of the two held-out categories, with higher joint entropy values for participants who are less consistent in their answers. We fit a simple linear regression model predicting joint entropy by Condition (4 levels, treatment coded, Condition 1 as baseline). No random effects were included in the model, as each participant had a single joint entropy value associated. As predicted, joint entropy rates were significantly higher for Conditions 3 and 4 (intercept= .28; vs. 3: $\beta = .639$

pendent on contextually-determined speech-act roles.

[8] All analyses used the lme4 package in R (Bates, 2010). The data and analyses script can be found here.

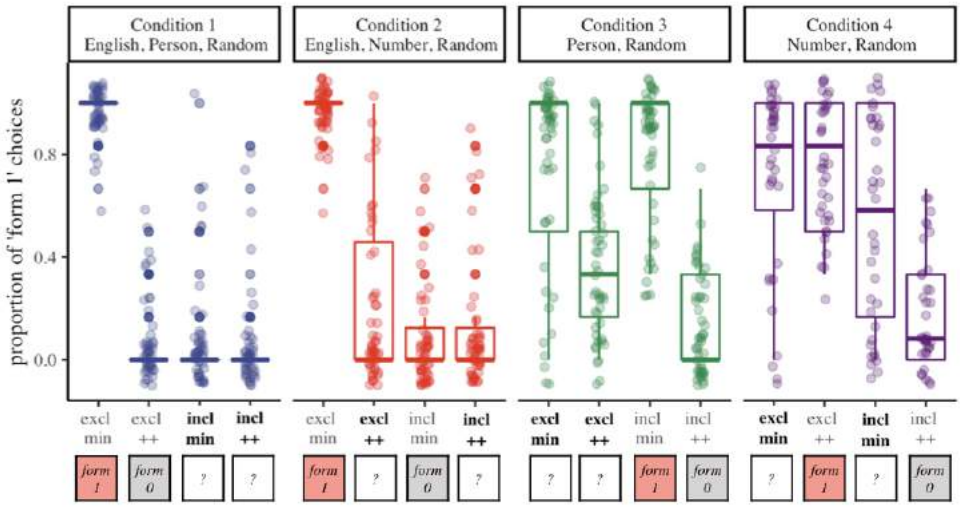


Figure 3: Proportion of 'form 1' choices for each first person category during the test phase. The held-out set for each condition is highlighted in bold-face. Choice of the same form (1 or 0) across categories indicates homophony. Dots are means of individual participants. Boxplots show by-participant means, quartiles, and range.

$\pm.12$, $p < .001$; vs. 4: $\beta = .447 \pm .13$, $p < .001$). Joint entropy rates for Condition 2 were marginally higher than Condition 1 ($\beta = .232 \pm .12$, $p = .055$).

Prediction 2. Preference for number over person homophony Figure 3 provides some evidence that participants in Conditions 3 and 4 are inferring paradigms with person and number homophony. Based on typological frequency, we predicted that number homophony should be more readily inferred than person homophony. If this is the case, then we should see a higher overlap in forms that share number in Condition 3 (columns in table 1) than forms that share clusivity in Condition 4 (rows). We measured this degree of homophony using the joint entropy between the relevant cells. For person homophony, we merged cells within a column and calculated joint column entropy. For number homophony, we merged cells within a row and calculated joint row entropy. The lower the joint entropy levels for a given homophony type, the more likely it is that participants are inferring a paradigm which neutralized that distinction. A simple linear regression model predicting joint entropy by Condition (2 levels, treatment coding, Condition 3 as baseline) revealed a marginally significant difference ($\beta = .207 \pm .11, p = .068$), with higher rates of person homophony (Condition 3) than number homophony (Condition 4). This fails to confirm our prediction.

Prediction 3. Preference for systematic over random homophony Finally, are participants in Conditions 3 and 4 in fact more likely to infer systematic rather than random homophony (as suggested by Figure 3)? To test this, the joint column/row entropy scores for systematic homophony computed above were compared to a random homophony score: the joint entropy of *all* alternative two-category combinations.[9] We ran separate mixed-effects models for Conditions 3 and 4, predicting entropy by homophony type (systematic vs. random) and including random intercepts per subject. We used likelihood ratio tests to compare these models to models with no fixed-effects. In both cases, entropy score for the systematic homophony pattern was significantly different from the random homophony score (person vs. random in Condition 3: $\chi^2 = 171.6$, $p < .001$; number vs. random in Condition 4: $\chi^2 = 84.4$, $p < .001$). This confirms that participants are more likely to use forms in a way that is consistent with systematic, not random homophony.

Discussion

In this experiment, we exposed English-speaking learners to sub-paradigms expressing person categories in a new language. We focused on first-person systems, which have been argued to have a universal basis in two features, encoding person and number. Participants were taught labels for two first person meanings, and asked to extrapolate to the two remaining meanings. We tested three hypotheses, designed to evaluate (1) whether learners were most likely to infer an English-like paradigm; (2) whether number homophony was more likely than person homophony (expected based on typological frequency); and (3) whether systematic homophony was more likely than random homophony

[9]For example, the joint column entropy in Condition 3 was compared to the joint entropy of each pair of diagonal and horizontal cells.

(predicted by feature-based theories).

Our results confirm that learners' are indeed highly likely to infer an English-like pattern when their training is consistent with this, producing systematic patterns of extrapolation from trained forms to new meanings. This result functions as a sanity check: it shows that participants are indeed understanding the stimuli in terms of a pronominal system.[10]

Our results also indicate that systematic homophony–which neutralizes either the number or person feature–is more natural than random homophony. This supports the claim that learners perceive the first person space as based on these two distinct features. Importantly, this finding cannot be accounted for solely based on experience with English. Inferring a person homophony pattern requires making a productive use of the ±minimal distinction. This is not the same number contrast made in English pronouns, which distinguish atomic (speaker only) and non-atomic entities (i.e., the more familiar singular/plural distinction). Similarly, inferring a number homophony pattern requires participants to learn and generalize the ±exclusive contrast, which is completely absent in English.

As for the typological difference between number (more common) and person (very rare) homophony, this does not appear to correlate with a learning difference in our task. Learners were, if anything, marginally more likely to infer paradigms characterized by person (Condition 3) rather than number homophony (Condition 4). One possibility is that, unlike random homophony, the rarity of person homophony in first person systems cross-linguistically is purely accidental, or reflects low sampling numbers. Indeed, person homophony is found for other parts of the person space (e.g., homophony of 1st and 2nd person in some languages). However, it may also reflect participants' experience of person homophony in English. Assuming that English encodes an atomic/non-atomic number distinction, it is possible to characterize English as a case of (only) person homophony (Harbour, 2016). In other words, English speakers have more experience with distinctions in number than in clusivisty. Indeed, a *posthoc* analysis shows that accuracy rates on trained categories (before exclusion) are higher in Condition 3 than 4 ($p < .001$), suggesting that the person distinction was harder to learn than the ±minimal distinction.

Finally, it is worth noting that differential sensitivity to person and number may also explain the marginal difference between Conditions 1 and 2. Both of these conditions allowed participants to generalize to an English-like paradigm, but they differed in whether a person or a number contrast was learned during the training phase (cf. Table 1). It could be that learning a new or unexpected distinction–between inclusive and exclusive minimal forms–led learners to be less likely to neutralize this feature in the plural.

[10]This is further confirmed by a debrief questionnaire, in which most participants reported having understood the new words as pronouns. For example, participants in Condition 4 have described the meaning of form 1 as 'Me or us not including you' and the meaning of form 0 as 'Us including you'.

Conclusion

In this study, we present the first experimental evidence for differences in learnability between alternative person paradigms. This was prompted by recent research in cognitive science on semantic spaces, and a lively literature in theoretical linguistics on the universal basis of person systems. We find, perhaps unsurprisingly, that English learners have a strong bias for first person paradigms that resemble their native language. They are more likely to infer paradigms analogous to English, and show a greater tendency to neutralize features that English also neutralizes (i.e. person). However, we also find that participants are sensitive to contrasts not found in their native language. Learners make productive use of both the ±minimal and ±exclusive distinctions, neither of which is present in English. Importantly, as predicted by feature-based theories of first person systems, learners were more likely to infer patterns which neutralized these features as compared to patterns in which featurally-unrelated cells were randomly homophonous. These initial results suggest that the paradigm we have developed can answer theoretically-motivated about how languages carve up the person space. Future work will target the full person paradigm, and incorporate recent insights about the potential role generally cognitive biases, such as simplicity, and communicative pressures like need probability (Kay & Regier, 2007; Kemp & Regier, 2012).

Acknowledgments

This project has received funding from the European Research Council (ERC) under the European Unions Horizon 2020 research and innovation programme (grant agreement No 757643).

References

Ackema, P., & Neeleman, A. (2018). *Features of person: From the inventory of persons to their morphological realization* (Vol. 78). MIT Press.

Baerman, M., Brown, D., Corbett, G. G., et al. (2005). *The syntax-morphology interface: A study of syncretism* (Vol. 109). Cambridge University Press.

Bates, D. M. (2010). *lme4: Mixed-effects modeling with R Springer* (Tech. Rep.).

Bobaljik, J. D. (2008). Missing persons: A case study in morphological universals. *Linguistic Review*, *25*(1-2), 203–230.

Bobaljik, J. D., & Sauerland, U. (2018). *ABA and the combinatorics of morphological features. *Glossa: a journal of general linguistics*, *3*(1), 1–34.

Culbertson, J. (2018). Artificial language learning. In *Oxford handbook of experimental syntax* (pp. 1–26).

Culbertson, J., & Adger, D. (2014). Language learners privilege structured meaning over surface frequency. *Proceedings of the National Academy of Sciences*, *111*(16), 5842–5847.

Cysouw, M. (2003). *The Paradigmatic Structure of Person Marking* (Oxford University Press, Ed.).

Cysouw, M. (2011). The expression of person and number: A typologist's perspective. *Morphology*, *21*(2), 419–443. doi: 10.1007/s11525-010-9170-5

Gibson, E., Futrell, R., Jara-Ettinger, J., Mahowald, K., Bergen, L., Ratnasingam, S., … Conway, B. R. (2017). Color naming across languages reflects color use. *Proceedings of the National Academy of Sciences*.

Harbour, D. (2008). On homophony and methodology in morphology. *Morphology*, *18*(1), 75–92.

Harbour, D. (2016). *Impossible persons*. MIT Press.

Harley, H., & Ritter, E. (2002). A Feature-Geometric Analysis. *Language*, *78*(3), 482–526.

Kay, P., & Regier, T. (2007). Color naming universals: The case of Berinmo. *Cognition*, *102*(2), 289–298.

Kemp, C., & Regier, T. (2012). Kinship Categories Across Languages Reflect General Communicative Principles. *Science*, *336*(6084).

Kemp, C., Xu, Y., & Regier, T. (2018). Semantic typology and efficient communication. *Annual Review of Linguistics*, *4*, 109–128.

Sauerland, U., & Bobaljik, J. D. (2013). Syncretism Distribution Modeling: Accidental Homophony as a Random Event. In *Glow in asia* (pp. 1–27).

Wilson, C. (2006). Learning phonology with substantive bias: An experimental and computational study of velar palatalization. *Cognitive Science*.

Zaslavsky, N., Kemp, C., Tishby, N., & Regier, T. (2018). Color naming reflects both perceptual structure and communicative need.

Zwicky, A. M. (1977). Hierarchies of person. In *Papers from the 13th regional meeting of the chicago linguistic society* (pp. 714–733). Chicago.

The Acquisition of French *Un*

Elisabeth Marchand (emarchan@ucsd.edu)
David Barner (dbarner@ucsd.edu)

University of California, San Diego
Department of Psychology
9500 Gilman Drive, La Jolla, CA 92093-0109, USA

Abstract

How does cross-linguistic variation in grammatical structure affect children's acquisition of number words? In this study, we addressed this question by investigating the case study of young speakers of French, a language in which the number *one* and the indefinite article *a* are phonologically the same (i.e., *un*). We tested how French-speaking children interpret *un,* and whether it more closely resembles the English word *a* or *one.* We found that French-speaking children almost always accepted sets of 1 for *un,* but that their responses for sets of 2 were more equivocal, with many children saying "Oui" (Yes) when asked whether there was *un.* Overall, French children's interpretation of *un* differed from how English-speaking interpret both *a* and *one*. This suggests that French-speaking children's interpretation of *un* reflects the ambiguity of the input that they are exposed to. We conclude that French morphological structure may pose a challenge to French-speaking children in acquiring an exact numerical meaning for the word *un,* potentially causing a delay in number word learning.

Keywords: Number; language; cognitive development

Introduction

How does the grammatical structure of a language affect children's acquisition of number words? By some accounts, morphology plays a central role in the acquisition of number words as it provides a conceptual framework for understanding small number words prior to conceptualizing them in terms of positive integers (Carey, 2004; Sarnecka, Kamenskaya, Yamana, Ogura, & Yudovina, 2007). The acquisition of these number words is progressive and follows a specific order (Le Corre & Carey, 2007; Sarnecka & Carey, 2008; Wynn, 1990). First, children learn the meaning of *one*, such that when they are asked to provide one object, they are able to correctly give one object and avoid giving one for other number requests. At this stage, these children are called "one-knowers". Then, children learn an exact meaning for *two*, and can give one or two when asked for one and two objects, but provide an incorrect response for other numbers. At this stage, children are called "two-knowers". Following the same pattern, children become "three-knowers" and sometimes "four-knowers". Finally, sometime after learning these number words, children seem to realize that they can use the count list to generate and give sets of any cardinality and for this reason, they are referred to as "Cardinal-Principle-knowers" (CP-knowers).

According to Carey (2009) and Sarnecka et al. (2007), morphology occupies a central role in the numerical acquisition process as children initially interpret *one*, *two* and *three* as markers of grammatical number categories. On this hypothesis, when children hear the word *one* in their input, it frequently occurs with singular agreement (e.g., *one cat*), whereas larger number words typically occur with plural nouns (e.g., *two cats*). Such cues might speed learning, allowing children to "bootstrap" number word meanings from grammatical morphology, such that, initially "one" is assigned a meaning similar to "a", and "two" is interpreted like a plural (Barner & Bachrach, 2010; Bloom & Wynn, 1997; Clark & Nikitina, 2009). Compatible with this, children learning languages like English, which has a grammatical singular/plural distinction, learn the meaning of *one* earlier than children exposed to languages that lack obligatory singular/plural marking, such as Japanese and Mandarin (Barner, Libenson, Cheung, & Takasaki, 2009; Le Corre, Li, Huang, Jia, & Carey, 2016; Sarnecka et al., 2007). Additionally, 2- to 4-year-old children learning Slovenian and Saudi Arabic, languages that have singular/dual/plural systems, acquire the meanings of *one* and *two* earlier than children exposed to any other previously tested language, despite being less familiar with counting overall (Almoammer et al., 2013; Marusic et al., 2016).

While previous tests of the relation between morphology and number word learning have focused mainly on how differences in grammatical morphology across languages might impact number words, few studies have asked whether the grammatical form of the numbers themselves might impact learning. Although children might initially interpret "a" and "one" similarly in English to learn a preliminary meaning of "one", they differentiate these words by at least 2 years of age: when children are shown a plate with two strawberries and are asked, "Is there *a* strawberry on the plate?" and, "Is there *one* strawberry on the plate?", 2-year-olds answer "Yes" for *a* strawberry but "No" for *one* strawberry, and do so as soon as they become one-knowers (Barner, Chow, & Yang, 2009). This suggests that *a* receives a purely existential interpretation (compatible with sets of 2 objects), while *one* receives an exact interpretation (compatible with sets of only 1 object). This suggests that, to acquire an exact meaning of "one", English children's input for "a" and "one" must differ.

Interestingly, however, other languages, like French and German feature the same phonological representation for both "a" and "one", a fact which might make it more difficult for them to determine whether to assign an existential or exact meaning to any particular instance of the word. For

example, in French, the word *un* is used both as an indefinite article and as a numeral. Consequently, French learners are presented with a potentially difficult learning problem, since the same phonological form is associated with both exact and non-exact meanings in their input.

Here, we investigated how French-speaking children interpret *un* – whether it resembles more closely the English *a* or *one* – and whether their interpretation of *un* differs based on the context of the task and whether surrounding test items are numerals or are restricted to non-exact quantifiers like "some" and "all". To further understand the impact of the ambiguity of the French morphological structure on children's interpretation of *un*, we compared the French-speaking children's interpretation of *un* to those of English-speaking children for *a* and *one* (obtained from Barner, Chow, & Yang, 2009).

Method

Participants

In total, 63 French monolingual children, aged 2;4 to 4;5-year-old were included in the study (M = 42.6 months). An additional 13 were excluded from analysis because of failure to complete all 3 tasks (n=2), bilingual status (n=7) or because they were not yet one-knowers or greater (non-knower; n=4). Participants were recruited from preschools in Québec (Canada). Informed consent was obtained from the parents. The study received approval by the ethics committee of UCSD.

Materials and procedure

Participants were tested at their preschool in a quiet corner of their classroom. Each session lasted approximately 15 min and included (1) a Truth-Value Judgement task, (2) the Give-a-Number task and (3) the Highest Count task. All participants were administered the tasks in this order. Children received a small prize for their participation at the end of the session.

Truth-Value Judgement Task (TVJ). This task was adapted from Barner, Chow, and Yang (2009) and its goal was to measure children's comprehension of the quantity terms: *un, des, deux, tous* (i.e., one/a, some, two, all) by asking them questions like, "Est-ce qu'il y a un canard dans la maison?" (Is there a/one duck in the house). Stimuli consisted of a drawing of a farmhouse and a forest, as well as three sets of small plastic animals (i.e., cats, pigs, and ducks). These animals were chosen as they are denoted by masculine nouns in French and therefore accompanied by the masculine form of the quantifier *un* (in contrast to the feminine *une*), which is the same form that typically corresponds to the number one (*un*). Animals were presented in separate piles organized by kind (Figure 1). Children were presented with the following instructions: "*Ça c'est la maison des animaux et ça, c'est la forêt. Moi je vais mettre des animaux dans la maison puis je vais te poser des questions. Toi, tu dois me répondre par oui ou non, ok?*" (i.e., "This is the animals' house and this is the forest. I will put animals in the house and ask you some questions. You need to answer by yes or no, ok?"). For each trial, the experimenter moved a certain number of animals into the farmhouse and asked the child a yes/no question. The animals were returned to their original piles after each trial. Children were randomly assigned to one of two conditions that differed with respect to the filler items that they included: (1) the Number condition, (2) the Quantifier condition. Children in the Number condition were presented with *un* and, as filler items, the number word *deux* (two), as well as the quantifier *tous* (all). Children in the Quantifier condition were also presented with *un*, in addition to the quantifiers *des* (some) and *tous* (all). Each item was presented with two different sets of animals. In both conditions, *un* was presented with sets of 1 and 2 objects. Children in the Number condition were asked questions with *deux* in the presence of 2 and 3 objects, to check whether they would interpret the number as exactly two and not compatible with larger sets (even by one object). Children in the Quantifier condition were questioned about *des* with sets of 1 and 2 objects, to check whether they would have a plural interpretation of *des*. In both conditions, *tous* was presented with sets of 3 and all 4 objects, to ensure that children had an interpretation of *tous* that was compatible with only all objects being present. Each combination of item and set was presented three times, for a total of 18 critical trials. The order of critical trials was counterbalanced across subjects.

Figure 1: Material used in the TVJ task.

Give-a-Number Task (Give-N). This task was adapted from Wynn (1990) and its goal was to evaluate children's understanding of number words. Stimuli consisted of a puppet, a red plastic plate, and 10 foam paper cookies. Children were asked to put a certain number of cookies into the plate (e.g., "*Peux-tu mettre trois biscuits dans l'assiette?*" i.e., "Could you put three cookies into the plate?"). After this first prompt, children were asked to count to verify that they had provided N, and if they had chosen to fix their answers, only their final responses were recorded. Each child was given 15 trials: three trials for each of the numbers 1, 2, 3, 4, 6. Order of trials was counterbalanced across children. Children were credited as N-knowers (e.g., two-knowers) if they correctly gave N cookies two out of three times when asked for N, and failed to give the correct N two out of three

times for N+1. In addition, to be classified as an N-knower, children could not use N more than 50% of the time for requests other than N. Finally, children were credited as CP knowers if they could correctly give six, two out of three times.

Highest Count Task (HC). Participants were asked to count as high as they could. The last number reached before making an error was taken as the highest count.

Results

Our primary question of interest was how French-speaking children interpret *un* and whether their interpretation differs according to the presence of other exact expressions in the context. In our first set of analyses, we tested whether performance in both conditions (Number and Quantifier) differed in terms of knower-levels (Give-N), Age, and Highest Count. Then, we conducted a series of analyses on *tous*, *des,* and *deux* to ensure that children either performed similarly across conditions (i.e., tous) or as expected given their respective conditions (i.e., des/deux). In our third set of analyses, we addressed our principal question of interest: Whether acceptance of sets of 1 or 2 objects for *un* differed across conditions. Finally, in a fourth set of analyses, we assessed how the acceptance rates for *un* compare to those for *a/one* in English by statistically comparing previously published data from English-speaking children (obtained from Barner, Chow, & Yang, 2009).

Preliminary Analyses

Knower-Levels. Table 1 shows the distribution of knower-levels across Conditions (Number vs Quantifier). Aside from a slightly greater number of one-knowers in the Quantifier condition, the conditions were similar in terms of their representation of each knower-level. Knower-level was not included as a factor in subsequent analyses comparing conditions both because our hypothesis is neutral to differences in knower level, and because such analyses require very substantial sample sizes to obtain adequate power.

Table 1: Distribution of Knower-Levels in the Number and Quantifier condition

	1K	2K	3K	4K	CP
Number	8	6	6	3	5
Quantifier	13	6	8	4	4
Total	21	12	13	7	9

Table 1: Here, 1K refers to one-knower, 2K to two-knower, 3K to three-knower, 4K to four-knower and CP to cardinal-principle-knower.

Highest Count. On average, children had difficulty counting to "dix/ten" (M = 6.30; SD = 4.80). The average Highest Count did not differ between the Number condition (M = 6.50; SD = 4.39) and the Quantifier condition (M = 6.14; SD = 5.18; p = 0.77).

Age. There was no difference in age between children in the Number condition (M = 42.71 months; SD = 6.86) and the Quantifier condition (M = 42.43; SD = 6.84; p = 0.87).

Truth-Value Judgment Task

Preliminary analysis of Tous, Des, Deux. In total, there were 35 children in the Quantifier condition and 28 in the Number condition. Figures 2 and 3 show the percentage of ''oui/yes" responses for each quantity term in each condition. As a first control check, we considered whether conditions differed in their acceptance of *tous* when controlling for Age and Highest Count. To do this, we performed a logistic mixed-effects model comparison,[1] using lme4 and car packages in R (Bates, Maechler, Bolker, & Walker, 2015; Fox & Weisberg, 2011).

In our first model, we predicted acceptance (coded as *yes* or *no*) from Age and Highest Count (HC), with participant as a random factor. In our second model, we added the main effects and interaction of Condition (Number vs. Quantifier) and Set Size (3 or all 4 objects) to the first model. In this model, we expected only a main effect of Set size and no difference between Conditions or interaction between Conditions and Set Size. The models were significantly different ($\chi^2(3) = 315.33$, $p = <.0001$). As expected, in our second model, the only significant predictor was Set Size ($\chi^2(1) = 30.75$, $p = <.0001$). This suggests that, in both conditions (Number and Quantifier), children were more likely to accept sets containing all 4 objects (Number: M = 0.99, SD = 0.11; Quantifier: M = 1.00, SD =0.00) compared to sets of 3 objects (Number: M = 0.25, SD = 0.44; Quantifier: M = 0.18, SD = 0.39).

As our second control check, we asked whether children in the Number condition accepted sets of 2 more often than sets of 3 objects when presented with *deux* (two), controlling for Age and Highest Count. In a model predicting acceptance from Age, Highest Count, and Set Size (with participant as a random factor), only Set Size was a significant predictor ($\chi^2(1) = 21.26$, $p = <.0001$). As expected, children accepted sets of 2 (M = 0.92, SD = 0.28) more often than sets of 3 (M = 0.28, SD = 0.45). Finally, as our last control check, we asked whether children in the Quantifier condition accepted sets of 2 more often than sets of 1 object when presented with *des*, after controlling for Age and Highest Count. Similar to the analysis with *deux*, in a model predicting acceptance from Age, Highest Count and Set Size (with participant as a random factor), only Set Size was a significant predictor ($\chi^2(1) = 15.89$, $p = <.0001$). In this context, children answered

[1] For *tous*, the first model specification was: Acceptance ~ HC + Age + (1|subject). The second was: Acceptance ~ HC + Age + Set Size * Conditions + (1|subject). For *des* and *deux*, the first model specification was: Acceptance ~ HC + Age + (1|subject). The second model was: Acceptance ~ HC + Age + Set Size + (1|subject).

"yes" more often when presented with sets of 2 (M = 0.93, SD = 0.26) compared to sets of 1 object (M = 0.69, SD = 0.47), though acceptance was high overall across these cases (compatible with past findings in English; Barner et al., 2009, and formal semantic analyses of the plural; see Bale, Gagnon, & Khanjian, 2011; Krifka, 1989; Sauerland, Anderssen, & Yatsushiro, 2005; Spector, 2007). Overall, the preliminary analyses combined confirmed that both conditions 1) did not differ in terms of Age and HC, 2) elicited interpretations of *tous*, *des* (Quantifier) and *deux* (Number) that the task was designed to induce.

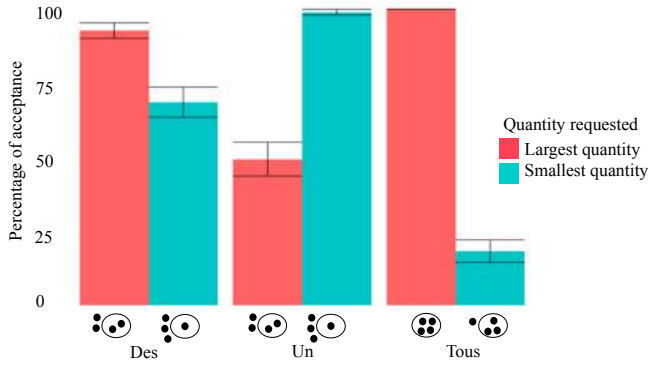


Figure 2: Children's percent saying "Oui/Yes" responses for each quantity term in the Quantifier condition in the Truth-Value Judgment Task. For the quantifier *des*, the largest quantity that was presented is 3 objects while the smallest is 2 objects. For *un*, the largest quantity that was presented is 2 objects while the smallest is 1 object. For *tous*, the largest quantity corresponds to all 4 objects while the smallest quantity consists of 3 objects. Error bars indicate standard error of the mean.

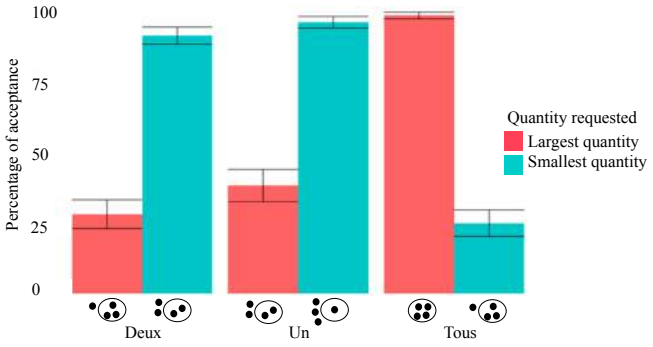


Figure 3: Children's percent saying "Oui/Yes" responses for each quantity term in the Number condition in the Truth-Value Judgment Task. For the numeral *deux*, the largest quantity that was presented was 3 objects while the smallest was 2 objects. For *un*, the largest quantity presented was 2 objects and the smallest 1. For *tous*, the largest quantity corresponded to all 4 objects while the smallest quantity was 3 objects. Error bars indicate standard error of the mean.

Children's interpretation of *Un*. In our main set of analyses, we addressed the question of how French-speaking children interpret *un* and whether their interpretation was exact (compatible with only sets of 1 object), inexact (compatible with sets of both 1 and 2 objects) or ambiguous. To do this, similar to the preliminary analysis, we performed a logistic mixed-effects model comparison.[2] In Model 1, we predicted acceptance from Age and Highest Count, with participant as a random factor, and in Model 2, we added the main effects and interaction of Set Size (1 or 2 objects) and Condition (Number vs. Quantifier) to Model 1. The presence of a main effect of Set Size would indicate of an interpretation of *un* that is either exact or ambiguous. Furthermore, adding Condition to our second model allowed us to test the question of whether acceptance rates differed based on the context of the game. If children have access to two different meanings for *un* that can be triggered by the pragmatic context of the game, then we should expect a significant interaction between Condition and Set Size, and specifically, that the acceptance rate for 2 objects should be higher in the Quantifier condition compared to the Number condition. Models 1 and 2 were significantly different ($\chi^2(3) = 164.79$, $p = <.0001$). However, in Model 2, the only significant predictor was Set Size ($\chi^2(1) = 50.24$, $p = <.0001$). As can be seen in Figures 2 & 3, children in both conditions accepted sets of 1 object (Number: $M = 0.96$, $SD = 0.19$; Quantifier: $M = 0.99$, $SD = 0.10$) more often than sets of 2 (Number: $M = 0.38$, $SD = 0.49$; Quantifier: $M = 0.49$, $SD = 0.50$). These results suggest that French-speaking children almost always accept sets of 1 for *un*, but when presented with sets of 2 objects, their interpretation of *un* is uncertain, hovering around 50% chance of saying "Oui/Yes", regardless of the context in which *un* is embedded.

Thus far, our data are more compatible with the third alternative presented: that French-speaking children have an ambiguous interpretation of *un* that is compatible with sets of 2 objects. In addition, despite the lack of significant difference between conditions, the question of whether French-speaking children have access to one or two meanings for *un* remains open. Indeed, our data are compatible with different interpretations: first, it is possible that French-speaking children only have access to a fuzzy representation of *un* - i.e., one that is neither exact like the English *one*, but not fully inexact like *a*. Second, it is possible that French-speaking children have access to two meanings for *un* but that the context of the task couldn't trigger the different interpretations. In order to further shed light on these possibilities, we compare these French-speaking children's acceptance rates to those of an English-speakers sample.

[2] The first model specification was: Acceptance ~ HC + Age + (1|subject). The second was: Acceptance ~ HC + Age + Set Size * Conditions + (1|subject).

Comparison with English data

To assess how these results compare to English, we obtained previously published data from English-speaking children's performance on the same task (from Barner, Chow, & Yang, 2009) and compared them to our sample of French-speaking children. The English sample included 31 participants of the same age (M = 45.3 months) as the French-speaking children in our study. The original task in Barner et al. (2009) included trials with *a, some, most, all, none, one, two*, with different options of set sizes, but only trials that tested *a* (sets of 1 and 2), *some* (sets of 1 and 2), *two* (sets of 2 and 3), *one* (sets of 1 and 2) and *all* (sets of 3 and all 8 objects) were selected for the current analyses.

First, we compared the acceptance rates for *some/des*, *two/deux,* and *all/tous* across French and English using mixed-effects model comparisons.[3] This was used as a control check to ensure that the linguistic groups didn't differ in the way that they understood and responded to the task. In all model comparisons, we first predicted acceptance from Set Size (with participant as a random factor) and then, added Language (English vs French) and the interaction between the two terms in the second model. There was no difference across languages for *some/des* and *two/deux* (both $ps > 0.01$). There was, however, a significant difference across languages and Set Size for *all/tous* ($\chi^2(1) = 7.13$, $p = <.001$) revealing that French children were more likely to accept sets containing all objects (M = 0.99; SD = 0.07) when asked about all objects compared to English speakers (M = 0.91; SD = 0.30)[4].

Our primary question of interest was whether French- and English-speaking children differed in their acceptance rate of *a, one,* and *un*. This question is also closely related to the question of whether French-speaking children have one or two meanings for *un*. Our prediction was that if French-speaking children had access to two interpretations for *un*, the context could be manipulated to favor one interpretation over the other, and we expected specifically that children in the Number condition would be more likely to have an interpretation of *un* close to *one* but not *a*, while children in the Quantifier condition would have an interpretation of *un* close to *a* but not *one*. To foreshadow, we found that children in the Number condition had an interpretation of *un* that was similar to *one* but not *a* and that children in the Quantifier condition interpreted *un* somewhat closer to *one* but differently than *a*. We obtained these results by performing 4 model comparisons contrasting the acceptance rate of: (1) the Number condition's interpretation of *un* to English speakers' interpretation of *one*, (2) the Number condition's interpretation of *un* to English speakers' interpretation of *a*, (3) the Quantifier condition's interpretation of *un* to English speakers' interpretation of *one*, (4) the Quantifier condition's interpretation of *un* to English speakers' interpretation of *a.* In all our first models, we predicted acceptance from Set Size (with participant as a random factor) and then, added Language (English vs French) and the interaction between the two terms in the second model.[5] When comparing (1) the Number condition's interpretation of *un* to English speakers' interpretation of *one,* we found only a main effect of Set Size ($\chi^2(1) = 2.15$, $p = <.001$) suggesting that all children were more likely to say "Yes" when presented with sets of 1 object (M = 0.96; SD = 0.19) compared to sets of 2 objects (M = 0.31; SD = 0.47), regardless of their linguistic group. Next, we looked at whether (2) children in the Number condition interpreted *un* differently than English speakers' interpretation of *a*. Here, our analysis revealed a main effect of Set Size ($\chi^2(1) = 20.95$, $p = <.001$) and an interaction between Set Size and Language ($\chi^2(1) = 7.42$, $p = <.01$) suggesting that English speakers where more likely to accept sets of 2 when asked for *a* (M = 0.78; SD = 0.42) compared to French speakers asked for *un* (M = 0.38; SD = 0.49). We then turned to children in the Quantifier condition and looked at how their interpretation of *un* compared to English. We first checked whether (1) the Quantifier condition's interpretation of *un* differed from English *one.* Here, we found a main effect of Language ($\chi^2(1) = 8.01$, $p = <.01$) and of Set Size ($\chi^2(1) = 30.22$, $p = <.001$), but the interaction between the two terms was not significant. This suggests that all children were more likely to accept sets of 1 object (M = 0.98; SD = 0.12) compared to sets of 2 objects (M = 0.39; SD = 0.49) and that French-speaking children (M = 0.77; SD = 0.42), on average, were more likely to say "Yes" compared to English-speaking children (M = 0.51; SD = 0.50). Finally, we looked at (4) how interpretation of *un* in the Quantifier condition compared to English speakers' interpretation of *a*. Here, we found a significant effect of Set Size ($\chi^2(1) = 17.47$, $p = <.001$), but most importantly a significant interaction between Set Size and Language ($\chi^2(1) = 6.77$, $p = <.01$), driven by the fact that French-speaking children were less likely to accept sets of 2 objects for *un* (M = 0.49; SD = 0.50) compared to English-speaking children for *a* (M = 0.78; SD = 0.42). Overall, these results suggest that children in the Quantifier condition interpreted *un* differently from English-speaking children's *a* and that children in the Number condition had an interpretation of *un* that was similar to *one* but not *a*.

[3] In all model comparisons, the first model specification was: Acceptance ~ Set Size + (1|subject). The second was: Acceptance ~ Set Size * Language + (1|subject).

[4] The difference between French- and English-speaking children could be explained by the fact that the English speakers, unlike the French speakers, when asked for *tous*, were presented with sets of no object at all in addition to the sets of 3 and all objects. English speakers could have accepted the sets of 3 objects more often simply due to the fact that it was already closer to "all objects" compared to the sets with no object at all. Regardless, the acceptance for sets of 3 in was still lower than 50% and for this reason, was not considered in our next analyses.

[5] In all model comparisons, the first model specification was: Acceptance ~ Set Size + (1|subject). The second was: Acceptance ~ SetSize * Language + (1|subject).

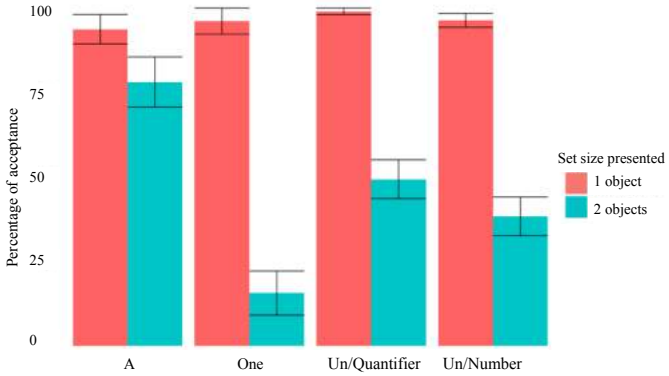


Figure 4: Comparison of English- and French-speaking children's acceptance rates for *a*, *one*, *un* in the Truth-Value Judgment Task (data from Barner, Chow, & Yang, 2009). "Un/Quantifier" represent the performances of children in the Quantifier condition while Un/Number represent the performances of children in the Number condition. Each term was presented with sets of one and two objects. Error bars indicate standard error of the mean.

Discussion

The goal of this study was to investigate the role of the morphological structure on children's acquisition of number words, via the case study of French-speaking children. Specifically, we investigated (1) how French-speaking children interpret *un,* and whether it is interpreted exactly, non-exactly or ambiguously, and (2) whether they have access to different meanings for *un* that can be triggered by the context of a task: an exact meaning that closely resembles English *one* and an inexact meaning similar to English *a*. When comparing acceptance rates across conditions, we found that children almost always accepted sets of 1 for *un,* but that their responses for sets of 2 were more varied, with many children saying "Oui/yes" when asked whether there was *un*, regardless of whether they were in the Number or Quantifier condition. However, an interesting mixed pattern emerged when comparing these acceptance rates to those of English-speaking children of the same age: children in the Number condition interpreted *un* as English-speaking children interpreted *one* (i.e., an exact interpretation compatible with only sets of 1 object), but children in the Quantifier condition interpreted *un* in a way that was not close to English-speaking children's *a*.

Overall, our results suggest that the morphological structure of French has an impact on children's learning. Specifically, our findings support the view that the homophony of *un*, compatible with both an exact and inexact interpretation, matters for the acquisition of the number word *one* as it creates a communicative problem. This homophony of *un* may provide more variable input to French-speaking children, leading to an ambiguous interpretation of the word. The contrast between English- and French-speaking children's interpretations of *un* vs *one/a* also suggests that French-speaking children not only need to learn that *un* can bear different meanings (exact and inexact) but also that meanings are affected by the context.

It has to be noted that the homophony of *un* is not the only aspect of French's morphology that could have an impact on children's acquisition of *un* as a numeral. Indeed, French's plural morphology is less salient in verbal communication compared to other languages like English. For example, in spoken French, very few nouns and verbs mark the singular/plural distinction, with the result that most nouns lack an audible word-final *s* to mark the plural like in English. This lack of salient plural agreement might exacerbate the challenge faced by French-speaking children to acquire an exact interpretation for *un*, but also other number words. Indeed, in English, children could in theory quickly start to notice a distinction between *one* and *two* as *one* always receives the singular agreement while *two* receives the plural. However, in French that distinction is unavailable for children (e.g., chat – i.,e., cat – is pronounced the same way regardless of whether it is presented with *un* or *deux*). As a consequence, French-speaking children might need to rely on more complex syntactic structures to pick up the distinction between *un* and *deux* (e.g., un chat dort/sleeps *vs* deux chats dor*ment*/sleep) and may need a significantly larger amount of input compared to English speakers – as not all verbs change phonetic forms based on plural agreement.

Another interesting aspect of these results is the apparent discrepancy between French children's performance for *un* in the Give-N task and the TVJ task. As a reminder, we excluded children who were not at least classified as One-knower at the Give-N task. This implies that when asked to provide *un biscuit* (i.e., *one* cookie), all children were able to provide exactly 1 object at least 2 out of 3 times. However, from the TVJ task, we can see that these same children still accepted sets of 2 objects as compatible with *un* around 50% of the time. Nonetheless, these results are not necessarily in contradiction. Indeed, these results are compatible with previous accounts which posit that though words like *a* and *one* may be associated with cardinal values of 1, they may not be pragmatically "strengthened" to exclude larger sets, especially in young children (Barner & Bachrach, 2010; Sauerland et al., 2005; Spector, 2007). According to these theories, if a child knows the meaning for *one* and knows that other numerals don't refer to set of 1 object, it would be infelicitous to provide more than 1 object when asked for *one*. For example, if a person asks to provide a fork, it would be pragmatically odd to give 2 or 3 forks. However, it would be more natural to have an existential interpretation of *one/a* in the context of a question – e.g., it seems less odd to say "yes, there is a fork in the bag" when asked whether there is *a/one* fork in bag and there is in fact 2 forks.

Taken together, our results raise the possibility that the ambiguity of French morphological structure poses a challenge to French-speaking children in acquiring an exact numerical meaning for the word un, potentially causing a delay in number word learning. Studies are currently in progress to test the possibility of a delay in the acquisition of early number words in French-speaking children.

Acknowledgments

We would like to thank members of the Language and Development Lab for their support during this project and the reviewers for their helpful comments and feedback.

References

Almoammer, A., Sullivan, J., Donlan, C., Marušič, F., O'Donnell, T., & Barner, D. (2013). Grammatical morphology as a source of early number word meanings. *Proceedings of the National Academy of Sciences*, 201313652.

Bale, A., Gagnon, M., & Khanjian, H. (2011). On the relationship between morphological and semantic markedness. *Morphology*, *21*(2), 197-221.

Barner, D., & Bachrach, A. (2010). Inference and exact numerical representation in early language development. *Cognitive psychology*, *60*(1), 40-62.

Barner, D., Chow, K., & Yang, S. J. (2009). Finding one's meaning: A test of the relation between quantifiers and integers in language development. *Cognitive psychology*, *58*(2), 195-219.

Barner, D., Libenson, A., Cheung, P., & Takasaki, M. (2009). Cross-linguistic relations between quantifiers and numerals in language acquisition: Evidence from Japanese. *Journal of experimental child psychology*, *103*(4), 421-440.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1–7. 2014.

Bloom, P., & Wynn, K. (1997). Linguistic cues in the acquisition of number words. *Journal of Child language*, *24*(3), 511-533.

Carey, S. (2004). Bootstrapping & the origin of concepts. *Daedalus*, *133*(1), 59-68.

Clark, E. V., & Nikitina, T. V. (2009). One vs. more than one: Antecedents to plural marking in early language acquisition. *Linguistics*, *47*(1), 103-139.

Fox, J., & Weisberg, S. (2011). Multivariate linear models in R. *An R Companion to Applied Regression. Los Angeles: Thousand Oaks*.

Krifka, M. (1989). Nominal reference, temporal constitution and quantification in event semantics. *Semantics and contextual expression*, *75*, 115.

Le Corre, M., & Carey, S. (2007). One, two, three, four, nothing more: An investigation of the conceptual sources of the verbal counting principles. *Cognition*, *105*(2), 395-438.

Le Corre, M., Li, P., Huang, B. H., Jia, G., & Carey, S. (2016). Numerical morphology supports early number word learning: Evidence from a comparison of young Mandarin and English learners. *Cognitive psychology*, *88*, 162-186.

Marušič, F., Plesničar, V., Razboršek, T., Sullivan, J., & Barner, D. (2016). Does grammatical structure accelerate number word learning? Evidence from learners of dual and non-dual dialects of Slovenian. *PloS one*, *11*(8), e0159208.

Sarnecka, B. W., & Carey, S. (2008). How counting represents number: What children must learn and when they learn it. *Cognition*, *108*(3), 662-674.

Sarnecka, B. W., Kamenskaya, V. G., Yamana, Y., Ogura, T., & Yudovina, Y. B. (2007). From grammatical number to exact numbers: Early meanings of 'one','two', and 'three'in English, Russian, and Japanese. *Cognitive psychology*, *55*(2), 136-168.

Sauerland, U., Anderssen, J., & Yatsushiro, K. (2005). The plural is semantically unmarked. *Linguistic evidence: Empirical, theoretical, and computational perspectives*, 413-434.

Spector, B. (2007). Aspects of the pragmatics of plural morphology: On higher-order implicatures. In *Presupposition and implicature in compositional semantics* (pp. 243-281). Palgrave Macmillan, London.

Wynn, K. (1990). Children's understanding of counting. *Cognition*, *36*(2), 155-193.

The complex system of mathematical creativity: Modularity, burstiness, and the network structure of how experts use inscriptions

Tyler Marghetis[1] tyler.marghetis@gmail.com
Kate Samson[1] knsamson1293@gmail.com
David Landy[1,2] dhlandy@gmail.com

[1] Indiana University, Bloomington, IN
[2] Netflix, Los Gatos, CA

Abstract

One of the pinnacles of human cognition is the creative insight of expert mathematics. While its concepts are abstract, the actual practice of mathematics is undeniably material and embodied. Mathematicians draw, sketch, write; having created these inscriptions, they interact with them. This iterated process of inscription is the engine of mathematical discovery. But how does this engine work? Here, using a new video corpus of mathematical experts working on proofs, and deploying tools from network and complexity science, we characterize the structure and temporal dynamics of how mathematical experts create and interact with blackboard inscriptions. We find regularities in the structure of this activity (e.g., emergent 'communities' of inscriptions) and its temporal dynamics (e.g., 'bursty' shifts in attention). By characterizing this activity, we gain a better understanding of the distributed ecosystem in which mathematical creativity occurs — including the ways that mathematicians actively construct their own notational niches.

Keywords: mathematical cognition; networks; complex systems; inscription; distributed cognition; embodiment

Introduction

One of the pinnacles of human cognition is the creative insight of expert mathematics. Often working alone, sometimes for years, mathematicians generate new knowledge about completely abstract objects, from infinite sets to imaginary numbers. The actual practice of mathematics, on the other hand, is undeniably concrete, material, and embodied. Mathematicians draw. They sketch. They write out derivations, erase them, start again. Having created these inscriptions, mathematicians interact with them: shifting their attention, talking about and gesturing at them, elaborating them further. This iterated process of inscription is the engine of mathematical discovery.

But how does this engine work? While philosophers, historians, and sociologists have argued that notations, diagrams, and the process of inscription are central to mathematical practice (Barany & MacKenzie, 2014; Mialet, 2012; Muntersbjorn, 2003), we know surprisingly little about the details of this process. Here, we use tools from network and complexity science to characterize the structure and temporal dynamics of expert mathematical activity—in particular, the process by which experts create and interact with inscriptions while working on mathematical proofs.

Notations in mathematical cognition

Past work in a range of disciplines has explored the role of notations and inscription in mathematical reasoning. Within mathematics education, for instance, it has long been recognized that choosing the right notation is often half the battle (Polya, 2004). This is true among experts just as much as it is true for schoolchildren (Muntersbjorn, 2003). Indeed, there is now a growing body of qualitative and theoretical research on the centrality of inscription in mathematical reasoning (Barany & MacKenzie, 2014; Greiffenhagen, 2014; Muntersbjorn, 2003; Roth & McGinn, 1998).

More controlled, quantitative studies have established that notations are a critical part of the distributed system of mathematical reasoning. In particular, there are bidirectional influences between, on the one hand, the specific notations used to solve mathematics problems, and, on the other, the psychological processes used to solve problems (Goldstone, Marghetis, Weitnauer, Ottmar, & Landy, 2017). Both undergraduate students and more expert reasoners, for instance, rely on the correspondence between spatial proximity and algebraic precedence in standard algebraic notation; algebraic performance is improved when this correspondence is maintained, harmed when it is violated (Landy & Goldstone, 2007). Conversely, experience with mathematical notations can reshape the psychological processes used to interact with them. Marghetis and colleagues (2016) found that, among adults who had mastered the syntax of algebra, the visual system had learned to perceive syntactically-related elements as unified visual objects. How we think about a mathematical domain shapes how we interact with inscriptions, and interacting with those inscriptions shapes how we see the problem.

Most of this past work, however, has focused on contexts where the notations are *supplied* rather than created by the participant. In real-world mathematical activity, by contrast, the reasoner must often explore multiple approaches to representing a problem—sketching out specific examples, pursuing different algebraic derivations, drawing a variety of different graphs—before settling on the final approach. Focusing only on the end product of this practice hides the dynamic messiness of mathematical reasoning. As mathematician Reuben Hersh put it, this confuses the clear, organized, pristine 'front stage' of

published or textbook mathematics for the messy, dynamic 'backstage' of real mathematical practice (Hersh, 1991).

Describing expert inscription activity

In this paper, we zoom in on this messy 'backstage' of mathematical reasoning, to try to characterize the dynamic contexts of mathematical creativity. To do so, we draw on tools from network science and complex systems. We describe a video corpus of mathematical experts working on non-trivial mathematical proofs. While this corpus offers endless possibilities for qualitative analysis, here we adopt a quantitative approach that allows us to measure how experts create and interact with mathematical inscriptions by identifying the 'inscription objects' that each expert created (e.g., equations, graphs, etc.) and then creating a timeseries of when, exactly, the expert attended to these objects, from their first creation to their final glance. We use this dense timeseries to describe the structure and dynamics of inscription.

To do so, we adapt tools from network science to offer a new methodology for studying situated cognitive activity: representing each expert's activity as a *directed network*, in which individual inscription objects are represented as nodes, and transitions between objects (e.g., shifting attention from one graph to another) are represented as directed edges (see Methods and Figure 1). This approach is a way to 'coarse-grain' the messy, chalk-covered reality of expert inscription, to better reveal the deeper regularities that characterize expert notational practices.

Methods

Corpus

We created a video corpus of experts solving non-trivial mathematics problems in a naturalistic setting (total corpus length: 4 hours and 40 minutes). Doctoral students in mathematics (N = 7, 4 men and 3 women) were recruited through the website of the mathematics department at a major research university and compensated $10/hour.

These experts solved up to three non-trivial problems in a natural setting: either their own office or a nearby seminar room within the mathematics department. They were encouraged to talk out loud as they solved the problems. All participants made ample use of the blackboard.

Videos were recorded with a Sony HDR-CX405 high-definition digital. The camera was positioned such that the board and the participant were visible.

Mathematics problems

Problems were drawn from the William Lowell Putnam Mathematics competition, an annual mathematics competition for undergraduate students. These problems are typically too difficult for even advanced undergraduate students, but tractable for mathematics experts at the doctoral level or above. Problems were selected to include a range of content areas (i.e., set theory, geometry, analysis):

(1) Find an uncountable subset, S, of the power set of a countable set, such that the intersection of each pair of elements in S is finite.

(2) Let $f : R^2 \rightarrow R$ be a function such that $f(x, y) + f(y, z) + f(z, x) = 0$ for all real numbers x, y, and z. Prove that there exists a function $g: R \rightarrow R$ such that $f(x, y) = g(x) - g(y)$ for all real numbers x and y.

(3) Let d_1, d_2, … d_{12} be real numbers in the interval (1, 12). Show that there exist distinct indices i, j, k such that d_i,d_j, d_k are the side lengths of an acute triangle.

Each participant worked for approximately an hour on the problems, depending on their availability. Most participants were only able to complete two of the problems in that time.

Video Coding

Each participant created dozens of inscriptions on the blackboard and then interacted with those inscriptions—by talking about them, gesturing towards them, or elaborating them with further inscriptions. We conducted a fine grained coding of the video corpus, at a nearly frame-by-frame resolution, to track the creation of and interaction with 'inscription objects' on the blackboard.

Blackboard inscriptions naturally clustered together into objects. For instance, a graph of a function might consist of two axes, labels for those axes ('x,' 'y'), and then a line representing the function. Each of those components, however, naturally cluster together in both meaning (they are all part of the same graph) and in spatial location (they are all located close together, with only minimal blank space in between). We used these two criteria—semantic relatedness and spatial proximity—to identify cohesive 'inscription objects' on the blackboard.

A coder viewed each video and annotated the onset and offset of inscription events: either the creation of a new inscription object, or subsequent interactions with that object (via talk, gaze, gesture, further elaboration, or erasing). This generated a timeseries of events for each inscription object, from its initial creation to the final time that the expert attended to it. For instance, if an expert created a graph at the very start of a session, the timeseries would include the onset and offset times for that process of initially drawing the graph; if the expert later looked at the graph, the timeseries also included that event. All coding was conducted in ELAN, software designed for annotating audio and video (Lausberg & Sloetjes, 2009).

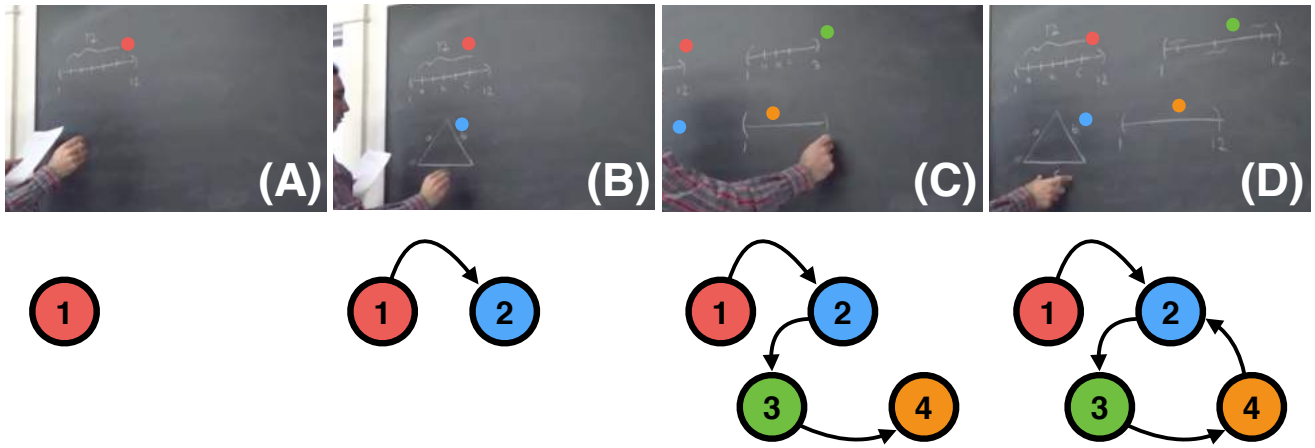


Figure 1. Illustration of a network representation of inscription activity. Blackboard images (top row) capture four consecutive stages in the process of developing a mathematical proof. The network representation of this process (bottom row) includes a node for each inscription object and an edge for transitions in attention from one object to another. We have added colored dots to the blackboard to indicate the location of each inscription object. (Node locations do not correspond to the objects' spatial locations.)

Network representation of inscription activity

To characterize the structure and temporal dynamics of experts' inscription activity, we used tools from network science. For each attempt to solve a problem, we used the timeseries of inscription events to generate a directed network, in which nodes represent inscription objects and directed edges represent transitions between objects.

As a simplified illustration, consider a scenario where an expert begins to solve a problem by creating and interacting with four inscription objects: three number-lines and one triangle (Fig. 1). The final network representation of this inscription activity would consist of four nodes, one for each inscription object, with a directed edge between two nodes whenever the expert attended first to one object and then to the other. For instance, if the expert started by creating a number-line (Fig. 1A), before abandoning that number-line to draw a triangle (Fig. 1B), the network representation of their activity up to that point would consist of two nodes — one for each inscription object — and a single directed edge from the first object to the second. As the expert creates new objects on the blackboard, the network grows (Fig. 1C, D), with their shifts in attention represented by directed edges between nodes 1 and 2, then from 2 to 3, then from 3 to 4, and then back to 2 again as they return their attention to an earlier inscription. This abstract graphical representation thus captures how the expert created and shifted their attention between inscription objects over the course of solving the problem.

Results

We first describe the network *structure* that emerged from the experts' inscription activity, then the *temporal dynamics* of their inscription activity, and finally the relations between the structure and dynamics of their notational activity.

Network structure of inscription activity

Experts created 360 distinct inscription objects, which they interacted with 4718 times. On average, solving an individual problem involved creating 24 inscription objects (SD = 16) and interacting with them 315 times (SD = 191).

Despite working on the same set of problems, experts in the corpus exhibited considerable variability in how they created and then shifted their attention among inscription objects. Figure 2, for instance, illustrates two different approaches to solving the same problem (problem #3, quoted above). For one individual (left), edges between nodes are distributed more or less randomly; nodes do not group together into interconnected clusters. By contrast (right), another individual interacted with inscriptions in interconnected clusters, and were much more likely to transition from one object to another within these clusters. This reflects a strategy where attention is likely to move to another inscription within the same cluster, creating pockets of activity wherein attention jumps between the same subset of inscriptions.

To identify these "communities" of inscriptions, we used the Girvan–Newman algorithm for community detection (Girvan & Newman, 2002), which identifies highly interconnected clusters of nodes using "edge betweenness" — the number of shortest paths between pairs of nodes that go through the edge — to identify highly central edges. In Figure 2, a node's community is identified by its color.

One way of describing the structure of inscription activity, therefore, is by how strongly shifts of attention defined communities of highly interconnected nodes — that is, the *modularity* of the network of inscription activity (Clauset, Newman, & Moore, 2004). Overall, inscription activity was significantly modular (M = 0.18, t_{14} = 4.1, p = .001; positive values indicate modularity, while 0 indicates

no modularity). Modularity exhibited both diversity and regularity. The participant illustrated on the left in Figure 2, for instance, generated a network with below average modularity compared to other experts who solved the same problem (modularity = 0.14), while the participant illustrated on the right had the second highest (modularity = .47). The modularity of individuals' activity varied considerably between problems (correlation in modularity between problems: r = 0.31); the individual who had the most modular activity on one problem, for instance, had the second-lowest modularity on another. By contrast, the two problems completed by most of the experts elicited reliably different modularity in inscription activity (M_{triangle} = 0.26 vs. M_{function} = 0.08, t_{13} = 2.4, p = .03). While modular clustering of inscriptions seems to be a recurring pattern in inscription activity, therefore, the precise amount of modularity likely reflects both the demands of the particular problem and stochastic, situated decisions.

In addition to the reliably modular structure of inscription activity overall, we also found finer grained regularities in the structure of communities themselves. Among communities, we observed two recurring 'motifs' or subgraph structures. One such motif was the 'cluster' motif, in which most nodes within a community were connected to each other (Fig. 3, right). These 'clusters' captured cases where a subset of inscription objects were all 'in conversation' with each other, with the expert shifting their attention among all inscriptions within that community. In contrast to these clusters, other 'loop' communities consisted entirely of a single, recurring route from one node, to another, to another, etc. in a straight, non-branching path (Fig. 3, left). These loops reflect inscriptions with a canonical pathway of attention—such as an algebraic derivation, where the experts attention would typically flow from the first expression to the last, in a set order.

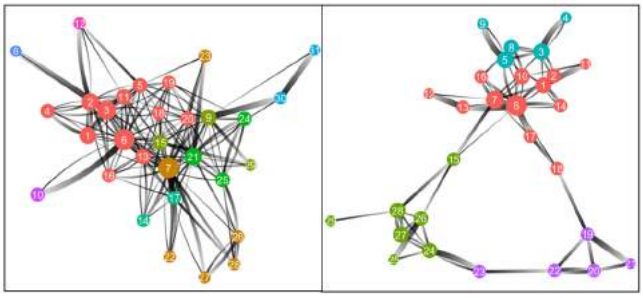


Figure 2. Different approaches to solving the same problem. Two different experts (left and right) solved the same problem, using approximately the same number of inscription objects (nodes). However, they interacted with those inscriptions in different ways, producing networks with different topological properties (see text). (Edge thickness indicates transition probabilities. Node color indicates community membership, as detected using the Girvan–Newman algorithm.)

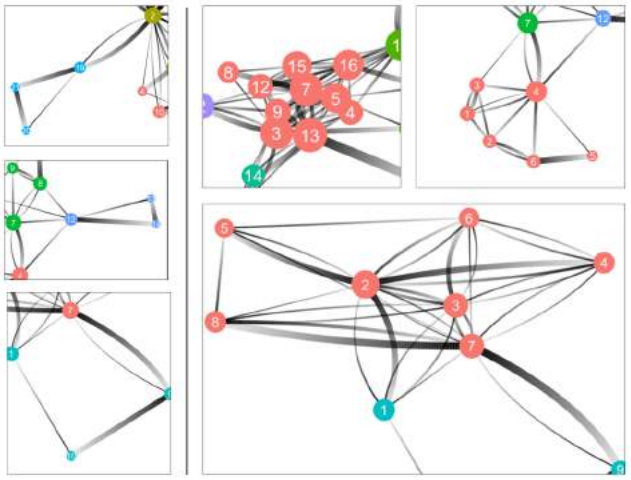


Figure 3. Recurring motifs in network communities. (left) Multiple communities involved only a single, recurring route from one node, to another, to another, etc. in a straight, non-branching path. (right) Other communities were highly interconnected, with most nodes connected to most other nodes.

Temporal dynamics of shifts in attention

We next characterized the temporal dynamics of inscription activity. To do so, we focused on the sequence of inter-event intervals — that is, the amount of time between the onset of attention towards one inscription object and the onset of attention towards the next object. A similar approach has been used to study the temporal dynamics of other human and human-technical systems (Barabasi, 2005; Goh & Barabási, 2008), such as email wait-times and dynamics of phone calls.

Here, we use a measure that have been used previously to characterize complex systems of social and cognitive activity, and to distinguish those systems from natural (e.g., earthquakes) and autonomous physiological activity (e.g., heartbeats): the 'burstiness' of the activity (Goh & Barabási, 2008). Past work has established that human activity systems often exhibit heavy-tailed dynamics — for instance, bursts of high activity followed by long periods of inactivity, with longer periods of inactivity than expected (e.g., assuming a Poisson or Gaussian distribution). This 'burstiness' may reflect interaction-dominant dynamics, with multiple processes combining in non-additive ways, or by an underlying process that involves priority queuing (Barabasi, 2005). The burstiness of a distribution of inter-event intervals {t} is typically measured by:

$$B = \frac{\sigma - M_t}{\sigma + M_t}$$

where σ is the standard deviation and M_t is the mean inter-event interval. More recently, this measure has been found to be sensitive to the size of finite samples, and the following elaboration has been adopted:

$$B = \frac{\sqrt{n+1}r - \sqrt{n-1}}{(\sqrt{n+1}-2)r - \sqrt{n-1}}$$

where r is the coefficient of variation, σ / M_t, and n is the sample size. Both these measures are designed to equal 0 for random, Poisson distributions; -1 for regular, periodic distributions; and +1 for bursty distributions. Past work has found that human activity systems typically exhibit significant burstiness (Goh & Barabási, 2008).

Overall, inscription activity was significantly bursty (B = 0.17, $p < .0001$, bootstrapped with n = 1000 samples). The burstiness was even more pronounced when we considered the distribution of times spent within a community of inscriptions — that is, the time between attending to one inscription object that belonged to a new community, and attending to a new object that belongs to a new community. This timeseries of inter-community dynamics was extremely bursty (B = .47, $p < .0001$, bootstrapped with n = 1000 samples), comparable to the most bursty human systems (Goh & Barabási, 2008). Inscription activity, therefore, was marked by with long periods of time spent within a community of inscriptions, followed by 'bursty' periods with rapid transitions between communities.

Relationship between structure and dynamics

Finally, we sought to characterize the relationship between the topological structure of experts' inscription activity (e.g., community structure and modularity) and the temporal dynamics of that activity.

First, we examined when, exactly, experts transitioned from one community of inscriptions to another. To do so, we took our timeseries of inscription object attention and determined whether the new object of attention belonged to the same or a different community— that is, a community transition. We then tried to predict the transition to a new community, using a generalized linear mixed-effects model of whether the new object belonged to a different community. We included as fixed effects the cumulative time spent on the problem; the amount of time spent attending to the current object; and, critically, the amount of time spent in the current community since most recently beginning to attend to that community ('sticking time'). We included random intercepts and slopes by participants, and random intercepts by problem.

There was no reliable relationship between the cumulative amount of time spent on the problem and the probability of transitioning to a new community of inscriptions (b = 0.22 ± 0.30 SEM, p = .46). By far the strongest predictor, however, was the amount of time spent within the current community, which had a large and negative relationship to the probability of transitioning to a new community (b = -2.29 ± 0.40 SEM, $p < .0001$). In other words, communities of inscriptions were themselves 'sticky,' so that the longer an expert spent within a community of inscriptions, the more likely they were to stay there going forward. This thus helps explain the highly bursty dynamics of transitions between communities, reported above: experts become fascinated with a particular cluster of inscriptions and spend considerable time, before suddenly transitioning to different inscription, and perhaps then undergoing a 'bursty' period of rapid transition between communities.

Finally, we looked at the relationship modular structure and between bursty dynamics. We used a linear mixed-effects model of the burstiness of the inscription activity used to solve each problem, and included predictors for the problem, the total number of events, the total number of inscription objects, the mean duration of an inscription event, a measure of the 'memory' of the activity dynamics (Goh & Barabási, 2008), and, crucially, our measure of modularity. The predictor with the largest relationship to burstiness, and the only one that was statistically significant was modularity (b = 1.29 ± 0.52 SEM, $p < .04$). More modular inscription activity—with communities of densely interconnected inscriptions—was associated with more bursty temporal dynamics (Fig. 4).

Discussion

Drawing on a corpus of mathematical experts working on non-trivial problems, and deploying tools from network and complexity science, we set out to characterize the 'manual labor' of mathematics (Marghetis, Edwards, & Núñez, 2014). We found that expert mathematical practice involved actively creating dozens of inscriptions and navigating between them, shifting attention from one to another. These shifts in attention were not random, however, but exhibited systematic modularity; inscriptions clustered together into 'communities,' subgroups of inscriptions that were likely to follow each other in a cascade of attention. This *structure* of inscription activity was related to the *temporal dynamics* of inscription, with a systematic relationship between inscription modularity and temporal burstiness (a hallmark of complex human activity). Overall, our network analysis of mathematical activity revealed both diversity and regularity in the inscription activity of experts.

The complex 'ecosystem' of cognition

By transforming raw video of situated problem solving into a directed network of inscription activity, we created a tractable representation of an otherwise prohibitively nuanced practice. This allowed us to adopt a quantitative approach without sacrificing a systems-level analysis. This approach shifts the focus away from individuals and skull-confined brain, and toward the ecosystem of mathematical practice, spanning brains, bodies, and blackboards.

From this perspective, the engine of mathematics is not the mathematicians' brain, locked away inside their skull. The brain is undeniably part of that engine. But equally important is the system of notations to which the mathematician has recourse, the particular inscriptions she creates in the moment, and the way her body allows her to

bring all those parts into coordination—by looking, pointing, sketching. The mathematician's creative insights are the product, not of solitary brains, but of a socially and materially distributed cognitive system (Hutchins, 1995). Indeed, this was true even of the physicist Stephen Hawking, who was famously confined to a wheelchair; instead of creating his own inscriptions, he worked closely with his able-bodied students, who created inscriptions on his behalf (Mialet, 2012).

This shift away from traditional intracranial processes to the larger complex system involved in creative mathematics puts a new emphasis on the material context of mathematical discovery. How should we characterize the endless inscriptions that mathematicians produce daily? How do those inscriptions change over the course of their mathematical training? How important are the inscriptions that a mathematician produces for herself, compared to those produced by her colleagues?

These questions suggest an analogy with another context of insight and learning: The early development of infants' visual and linguistic systems. Recent work has begun to characterize the rich visual and linguistic input that is received by the developing child—including how the child actively shapes that input to facilitate learning (e.g., Smith, Jayaraman, Clerkin, & Yu, 2018). Understanding the larger ecosystem in which learning occurs—whether by a pre-verbal child or a highly trained mathematician—will be critical to understanding how, exactly, that learning occurs.

Indeed, this analogy with early child learning highlights another critical component of situated mathematical practice: Mathematicians do not receive carefully formed representations of their problems. They must figure out how to represent their ideas. In this way, the mathematician is like the child who actively shapes their visual input. Children shape their visual context to facilitate learning. Mathematicians transform their material context to facilitate creative insight. By sketching, drawing, graphing, and writing various algebraic expressions, they engage in a form of niche construction: 'notational niche construction.'

Limitations

Our analysis has a number of limitations.

For one, naturally occurring inscriptions need not necessarily cluster into objects defined by semantic relatedness and spatial proximity. In our corpus, however, the inscriptions did typically fall into unambiguous clusters, and the few unclear cases were resolved through discussion among the authors (e.g., deciding whether a vertical stack of equations should count as one object or multiple, distinct objects). Second, one modality by which experts could engage with an object was through gaze; however, since our data consisted only of a single camera, it was not always possible to determine where a participant was looking. Gaze toward an object was only coded when there was unambiguous evidence that the participant had shifted their gaze toward an object, such as when they turned their entire head to look at an inscription that was relatively isolated on the blackboard. As a result of this conservative approach, we may have underestimated the number of transitions between objects. To address both these issues, future work will need to establish the reliability of the coding scheme by using multiple coders and calculating inter-coder reliability.

Third, the methodology introduced here is very time-consuming, both when initially collecting the data (which requires recruiting highly trained experts) but especially when coding the video data afterwards. As a result, the current corpus consists of hundreds of inscriptions and thousands of interactions, but these were drawn from the activity of only seven experts. Were are currently working to expand our corpus in order to investigate the generality of the current findings.

Future Directions and Conclusions

We have not even begun to look at how the structure and dynamics of inscription might change over the course of a problem solving episode. For instance, are there distinct phases of activity—perhaps early exploration of different inscriptions, followed by later exploitation of successful ones?

Relatedly, we have yet to investigate the association between the structure and dynamics of inscription and various other outcome measures. For instance, does the network structure of inscription activity predict the creativity or completeness of the final proof? On a more granular level, what happens immediately before the expert has a sudden insight—can we predict the onset of a critical transition in understanding (e.g., Setzler, Marghetis, & Kim, 2018; Stephen, Boncoddo, Magnuson, & Dixon, 2009)?

Third, future analyses will look in more detail at the *kinds* of inscriptions that experts are creating. Does inscription activity differ between, say, algebraic equations versus Cartesian plots? Might the structure and dynamics of inscription offer insights into how individuals tend to use different kinds of inscriptions—or perhaps reveal that superficially dissimilar inscriptions are actually treated similarly by experts?

Finally, we are curious about which aspects of inscription activity are specific to highly-trained mathematical experts, and which might also occur among novices. Work on other complex systems has found that the temporal dynamics of a complex system can predict the system's health or resilience (Kleiger, Miller, Bigger Jr, & Moss, 1987). One possibility, for instance, is that bursty dynamics during inscription is diagnostic of mathematical expertise.

Answering these questions will bring us closer to understanding how one of the most *abstract* forms of human understanding is so undeniably *concrete:* Covered in chalk, gesturing emphatically at the blackboard, the thinking mathematician is engaged in manual labor — and then, suddenly, she understands infinity.

Acknowledgments

We are thankful for the generous feedback from members of the Landy Lab and four anonymous reviewers. This work is supported by Indiana University Bloomington through the Emerging Area of Research initiative — Learning: Brains, Machines, and Children.

References

Barabasi, A.-L. (2005). The origin of bursts and heavy tails in human dynamics. *Nature*, *435*(7039), 207.

Barany, M. J., & MacKenzie, D. (2014). Chalk: Materials and concepts in mathematics research. *Representation in Scientific Practice Revisited*, 107–130.

Clauset, A., Newman, M. E., & Moore, C. (2004). Finding community structure in very large networks. *Physical Review E*, *70*(6), 066111.

Girvan, M., & Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, *99*(12), 7821–7826.

Goh, K.-I., & Barabási, A.-L. (2008). Burstiness and memory in complex systems. *EPL (Europhysics Letters)*, *81*(4), 48002.

Goldstone, R. L., Marghetis, T., Weitnauer, E., Ottmar, E. R., & Landy, D. (2017). Adapting Perception, Action, and Technology for Mathematical Reasoning. *Current Directions in Psychological Science*, *26*(5), 434–441.

Greiffenhagen, C. (2014). The materiality of mathematics: Presenting mathematics at the blackboard. *The British Journal of Sociology*, *65*(3), 502–528.

Hersh, R. (1991). Mathematics has a front and a back. *Synthese*, *88*(2), 127–133.

Hutchins, E. (1995). *Cognition in the Wild*. MIT press.

Kleiger, R. E., Miller, J. P., Bigger Jr, J. T., & Moss, A. J. (1987). Decreased heart rate variability and its association with increased mortality after acute myocardial infarction. *The American Journal of Cardiology*, *59*(4), 256–262.

Landy, D., & Goldstone, R. L. (2007). How abstract is symbolic thought? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(4), 720.

Lausberg, H., & Sloetjes, H. (2009). Coding gestural behavior with the NEUROGES-ELAN system. *Behavior Research Methods*, *41*(3), 841–849.

Marghetis, T., Edwards, L. R., & Núñez, R. (2014). More than mere handwaving: Gesture and embodiment in expert mathematical proof. *Emerging Perspectives on Gesture and Embodiment in Mathematics*, 227–246.

Marghetis, Tyler, Landy, D., & Goldstone, R. L. (2016). Mastering algebra retrains the visual system to perceive hierarchical structure in equations. *Cognitive Research: Principles and Implications*, *1*(1), 25.

Mialet, H. (2012). *Hawking incorporated: Stephen Hawking and the anthropology of the knowing subject*. University of Chicago Press.

Muntersbjorn, M. M. (2003). Representational innovation and mathematical ontology. *Synthese*, *134*(1–2), 159–180.

Polya, G. (2004). *How to solve it: A new aspect of mathematical method*. Princeton university press.

Roth, W.-M., & McGinn, M. K. (1998). Inscriptions: Toward a theory of representing as social practice. *Review of Educational Research*, *68*(1), 35–59.

Setzler, M., Marghetis, T., & Kim, M. (2018). Creative leaps in musical ecosystems: early warning signals of critical transitions in professional jazz. *The 40th Annual Meeting of the Cognitive Science Society*.

Smith, L. B., Jayaraman, S., Clerkin, E., & Yu, C. (2018). The developing infant creates a curriculum for statistical learning. *Trends in Cognitive Sciences*.

Stephen, D. G., Boncoddo, R. A., Magnuson, J. S., & Dixon, J. A. (2009). The dynamics of insight: Mathematical discovery as a phase transition. *Memory & Cognition*, *37*(8), 1132–1149.

Navigating the "chain of command": Enhanced integrative encoding through active control of study

Douglas B. Markant (dmarkant@uncc.edu)
Department of Psychological Science
University of North Carolina at Charlotte
9201 University City Blvd., Charlotte, NC 28223 USA

Abstract

A growing body of research indicates that "active learning" improves episodic memory for material experienced during study. It is less clear how active learning impacts the integration of those experiences into flexible, generalizable knowledge. This study used a novel active transitive inference task to investigate how people learn a relational hierarchy through active selection of premise pairs. Active control improved memory for studied premises as well as transitive inferences involving items that were never experienced together during study. Active learners also exhibited a systematic search preference, generating sequences of overlapping premises that may facilitate relational integration. Critically, however, advantages from active control were not universal: Only participants with higher working memory capacity benefited from the opportunity to select premise pairs during learning. These findings suggest that active control enhances integrative encoding of studied material, but only among individuals with sufficient cognitive resources.

Keywords: active learning; transitive inference; information search; integrative encoding

Introduction

How does the opportunity to control a learning experience alter subsequent memory of it? Recent research has shown that active control over learning enhances episodic memory for experienced material compared to passive observation of the same information (Markant, DuBrow, Davachi, & Gureckis, 2014; Voss, Gonsalves, Federmeier, Tranel, & Cohen, 2011). This enhancement can arise from a number of mechanisms, including improved attentional coordination, metacognitive monitoring, or enriched encoding associated with volitional control (Markant, Ruggeri, Gureckis, & Xu, 2016).

Less is known about how active control affects the integration of studied material into flexible, generalizable knowledge. Other work has revealed benefits from active information selection when learning categorical rules (Markant & Gureckis, 2014) or causal structures (Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003). However, these studies have not examined a crucial question about improvements in generalization following active control of study: Do they reflect better memory for experienced information itself (which then supports generalization later on) or the formation of relational knowledge during encoding that abstracts away from that experience? Following Zeithamova, Schlichting, and Preston (2012), these alternatives can be mapped onto two types of memory formation: *elemental encoding* of stimuli or associations that are directly experienced during study, and *integrative encoding* through which disparate study episodes are bound together into a unified representation. Whereas existing research has established that active control enhances elemental encoding in a variety of contexts, its relationship to integrative encoding remains unclear.

The present study examined the effects of active control in a well-known example of relational generalization: transitive inference (TI). In TI people learn about an ordered hierarchy (e.g., A $<$ B $<$ C) by studying premises comprised of adjacent items (e.g., A $<$ B, B $<$ C). They are then tested on their memory for studied pairs (*recall trials*; e.g., A ? B) and their ability to infer relationships between items that were never experienced together (*inference trials*; e.g., A ? C).

Transitive inference is a fundamental form of reasoning and has been the subject of a wealth of past research, but has always been studied under passive conditions in which control over the study experience is absent. This study introduces a novel *active transitive inference* task in which participants choose which premises to study during learning. Based on prior evidence that active selection improves episodic memory, active selection was expected to improve recall of studied premises relative to passive study. Active control was also predicted to improve transitive inference, but this advantage might arise from two distinct mechanisms. Enhanced elemental encoding of premises should bolster retrieval at the time of test, allowing participants to make transitive inferences by reasoning across overlapping pairs. Alternatively, active control may enhance integrative encoding during study, aiding the formation of a unified representation of the hierarchy. Importantly, these processes predict distinct relationships between performance and the distance between test items (see below), making TI well-suited to examine how learner control changes the representation of studied material.

Elemental vs. integrative encoding in transitive inference

Transitive inference involves comparing items that have never been experienced together but are linked by one or more studied pairs. TI may be supported by a number of alternative processes which can be distinguished by their dependence on elemental or integrative encoding. Elemental encoding-based inference occurs by reactivating studied premises at the time of test and reasoning across overlapping relations (Kumaran & McClelland, 2012). In this case, successful inference hinges

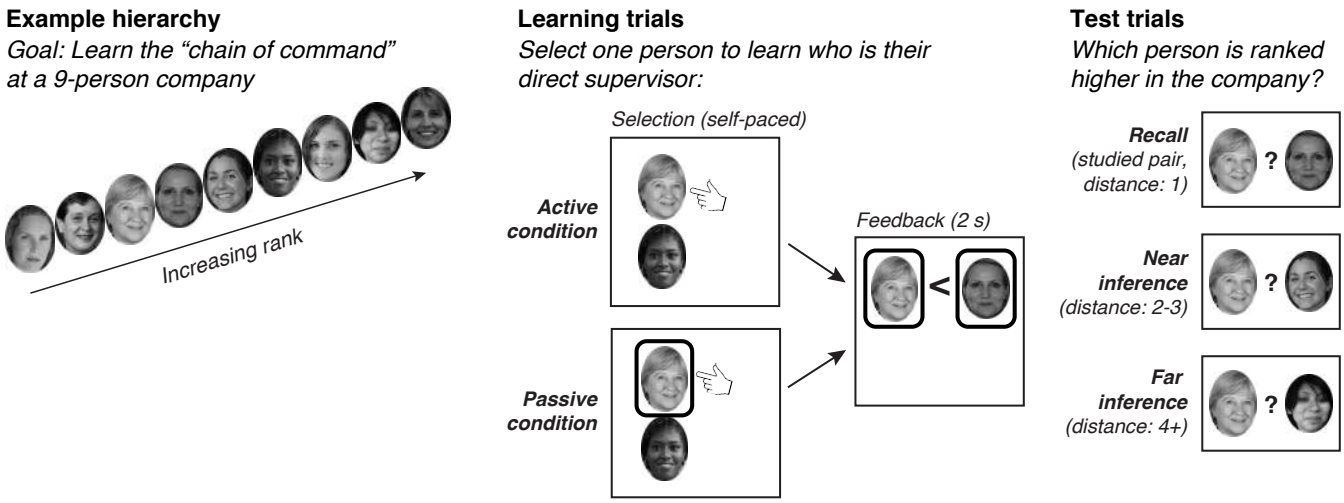


Figure 1: Depiction of the transitive inference task.

on robust encoding of studied pairs to ensure later retrieval. This process implies that large distances between test items (e.g., A ? E) will be associated with lower accuracy since there are more opportunities for retrieval errors along the way.

In contrast, integrative encoding-based accounts of TI postulate the formation of a unified, ordinal representation during study (De Soto, London, & Handel, 1965; Hummel & Holyoak, 2001; Shohamy & Wagner, 2008; Trabasso, Riley, & Wilson, 1975). Inference then entails comparing the positions of any two items along that dimension. Importantly, this process implies accuracy should *increase* with inferential distance, as items that are further apart on that latent dimension are easier to distinguish. Such *symbolic distance effects* are a hallmark of integrative encoding (Moyer & Landauer, 1967).

Although alternative forms of associative or reinforcement learning may also support TI (Frank, Rudy, Levy, & O'Reilly, 2005), the construction of an integrated representation during encoding is especially likely when participants are aware there is an underlying hierarchy to be learned (Greene, Spellman, Levy, Dusek, & Eichenbaum, 2001; Lazareva & Wasserman, 2010). Accuracy is higher among participants who report post-task awareness of the hierarchy (Martin & Alsop, 2004), who are informed about it prior to training (Greene et al., 2001; Smith & Squire, 2005), or when stimuli evoke hierarchical schemas (Kumaran, 2013). Reliance on integrated representations also appears to depend on working memory capacity (WMC) (Titone, Ditman, Holzman, Eichenbaum, & Levy, 2004; Fales et al., 2003). Thus, while constructing an integrated representation is typically associated with superior generalization, it may also depend on explicit awareness of the hierarchical organization of items and incur greater cognitive costs.

Learner control and integrative encoding

Since elemental and integrative encoding predict distinct relationships between inferential distance and performance, TI can be used to examine whether active control has broader benefits for memory formation beyond improved episodic memory for studied pairs. One reason to expect enhanced integrative encoding is that the opportunity to select premises may encourage learners to construct an integrated representation as they learn, which can then guide selection decisions (e.g., allocating study to items from less familiar portions of the hierarchy). At the same time, this process might involve additional demands on aspects of executive functioning such as working memory. To evaluate this possibility an assessment of WMC (operation span) was included in addition to the TI task in the experiment below.

In addition to the main goal of identifying any effect of active control on integrative encoding, the TI task was designed to explore information search during active study. Passive training in TI is often scaffolded such that overlapping pairs are experienced in direct succession (e.g., A $<$ B, B $<$ C, ...), which leads to faster learning than random sequences (Halford, 1984). If studying overlapping premises aids relational integration, active learners may prefer to select such options when possible. Each selection therefore involved a choice between a *near* and *far* option which differed in their distance from the pair studied on the previous trial. This made it possible to identify any search preference during active study and its relationship to inferential accuracy.

Experiment

Participants and Materials

N = 100 participants (60 women; age: M = 21.94 years, SD = 5.60) were recruited from the student population at UNC Charlotte. Participants received either course credit or \$8 (\$4 per session), as well as a \$0–\$5 incentive based on their performance in the first test session. N = 62 participants returned for the second session.

Face stimuli for the TI task were obtained from the 10k US Adult Faces Database (Bainbridge, Isola, & Oliva, 2013). For

Table 1: Estimated fixed effects from mixed effects logistic regression model of test accuracy.

Predictor	OR	95% CI-lower	95% CI-upper	Wald z	p
(Intercept)	4.07	3.30	5.07	12.36	0.00
Condition [passive]	0.76	0.67	0.85	-5.19	0.00
Session [retest]	0.85	0.76	0.98	-2.67	0.01
Distance	1.10	1.02	1.16	3.00	0.00
Operation span	2.06	1.67	2.58	6.47	0.00
Condition [passive] x Session [retest]	0.80	0.68	0.94	-2.68	0.01
Condition [passive] x Distance	0.94	0.87	1.02	-1.33	0.18
Condition [passive] x Operation span	0.55	0.49	0.60	-12.49	0.00

each sex, the stimulus set was filtered to include only faces that were non-famous and which had mean ratings within a 1-point interval centered on the midpoint of the rating scale for perceived age, emotional affect, and memorability. Thirty-six images (18 male, 18 female) were manually chosen from the filtered set to ensure high image quality and the absence of other distinctive features (e.g., jewelry, background objects).

Procedure

There were two sessions. The first session included the TI task followed by the operation span assessment. The second session occurred 6-8 days later and included a second run of the test phases from the TI task.

The TI task (Figure 1) used a within-subjects design with two rounds. Participants were tasked with learning the "chain of command" at two companies. Each participant learned about one 9-item hierarchy in the active condition and a second 9-item hierarchy in the passive condition. Each hierarchy was composed of all female faces or all male faces in order to reduce interference between conditions. The order of conditions and mapping of stimulus set to condition were counterbalanced across participants. Each round was comprised of a learning phase (56 trials) followed by a test phase (72 trials).

The instructions included an example of a 3-item hierarchy in which participants learned about two premise pairs (person A $<$ person B, person B $<$ person C) and were asked to infer the transitive relation (person A $<$ person C). All participants were therefore aware of the hierarchical nature of the stimuli and were explicitly instructed to learn to judge the relative rank of any two individuals in a given company.

Learning phase. The learning phase involved a series of choices between two non-adjacent items in the present hierarchy (excluding the highest-ranking item which was never presented as a choice option). The options on the first learning trial were two non-adjacent items sampled at random. On all subsequent trials, options differed in their distance from the item selected on the previous trial: Each option set included a *near* option that was 1–2 positions away from the item selected on the previous trial, and a *far* option that was 3 or more positions away from the item selected on the previous trial. This manipulation of option distance was designed to test whether participants in the active condition preferred to select items based on their distance. In the passive condition selections were evenly divided between near and far options.

Active study condition. Each trial began with the presentation of the two options in a vertical array in random order (Figure 1, middle). Participants were instructed to select an option at their own pace in order to learn that person's direct supervisor. Following their choice the unselected option disappeared and the premise pair (selected item and superordinate feedback item) was displayed for 2 s.

Passive study condition. In the passive condition participants did not decide which option to select. As in the active condition, the trial began with the presentation of two options, one of which was already highlighted with a red border. Participants were instructed to select the highlighted option at their own pace, at which point the trial proceeded in the same manner as in the active condition.

Test phase. In each test trial, two items were presented side-by-side and the participant clicked on the person they judged to be ranked higher in the hierarchy. The test phase was comprised of three trial types (Figure 1, right): *recall* trials involving a choice between studied premise pairs (e.g., A ? B), *near inference* trials involving items that were 2–3 positions apart (e.g., A ? C), and *far inference* trials involving items that were 4 or more positions apart (e.g., A ? E). In the second session, participants completed a second run of the same test phases experienced during the first session, with test pairs presented in a new random order.

Operation span. In the operation span task, participants attempt to hold a sequence of items in memory while judging the validity of interleaved math operations (Unsworth, Heitz, Schrock, & Engle, 2005). At the end of a trial involving multiple such steps, participants recall the sequence of digits in the same order as they appeared. The set size (number of operations/digits) ranged from 2–7, presented in increasing order, with three trials completed for each set size. Participants were highly accurate at evaluating the validity of the math operations (judgment accuracy M = 0.92, SD = 0.06). Operation span was scored according to the summed number of digits recalled in the correct order for those trials in which no

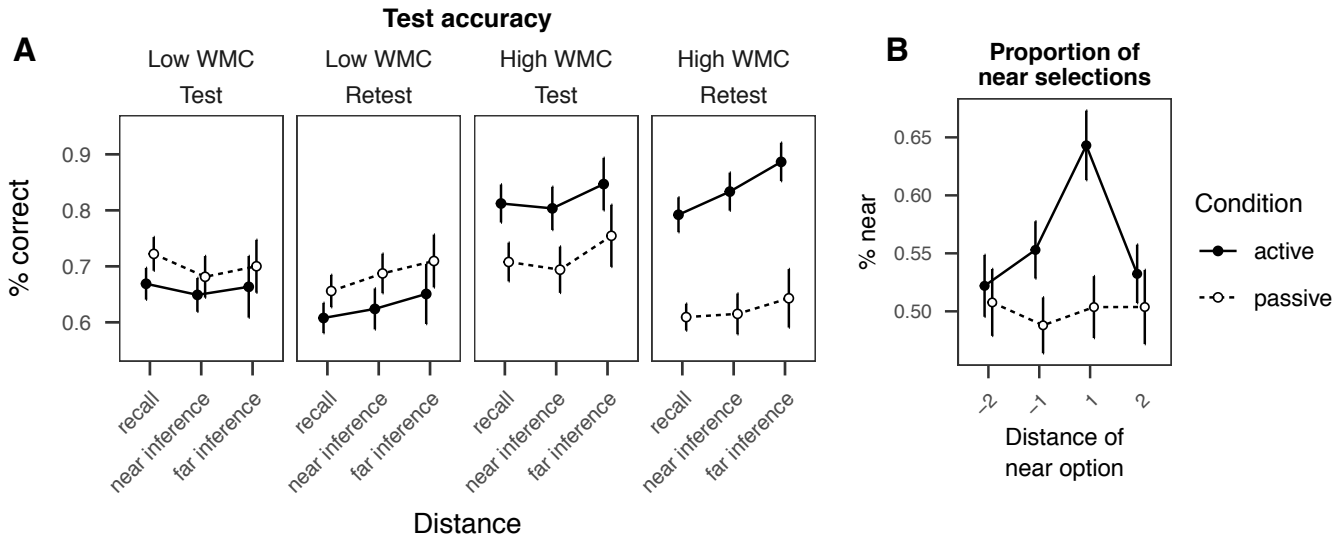


Figure 2: **A:** Test accuracy after median split on operation span, for both the immediate test and delayed retest. Performance is shown as a function of study condition (active/passive) and inferential distance (recall/near inference/far inference). **B:** Proportion of trials in which the *near* option was selected as a function of its distance from the item chosen on the previous trial. All error bars indicate within-subjects 95% confidence intervals.

errors were made (M = 14.86, SD = 11.29, median = 12.50).

Results

Test accuracy. Test responses were scored according to whether participants correctly identified the superordinate item in each test pair (0 = incorrect, 1 = correct). Test trials involving either endpoint of the hierarchy were excluded from analysis since participants could rely on non-transitive strategies to respond. Accuracy was modeled using mixed effects logistic regression (Table 1). The model included fixed effects for condition (active/passive), session (test/retest), distance (recall/near inference/far inference), and operation span (continuous), as well as pairwise interactions between condition and the other predictors. Random intercepts were included for participants and stimuli in each test pair.

Active performance was higher than passive performance in both the immediate test (active: $M = 0.74$, $SD = 0.21$; passive: $M = 0.71$, $SD = 0.21$; $OR = 1.32$, $CI = [1.13, 1.54]$, $z = 5.19$, $p < .001$) and in the retest (active: $M = 0.73$, $SD = 0.21$; passive: $M = 0.65$, $SD = 0.21$; $OR = 1.64$, $CI = [1.37, 1.98]$, $z = 7.64$, $p < .001$). Accuracy declined from the immediate test to the retest in both the active condition ($OR = 0.85$, $CI = [0.71, 1.01]$, $z = -2.67$, $p = 0.007$) and the passive condition ($OR = 0.68$, $CI = [0.57, 0.81]$, $z = -6.48$, $p < .001$).

There was a symbolic distance effect, such that accuracy increased with inferential distance, in the active condition ($OR = 1.10$, $CI = [1.00, 1.20]$, $z = 3.00$, $p = 0.003$) but not the passive condition ($OR = 1.04$, $CI = [0.95, 1.13]$, $z = 1.22$, $p = 0.22$). In addition, operation span was positively related to accuracy in the active condition ($OR = 2.06$, $CI = [1.50, 2.84]$, $z = 6.47$, $p < .001$) but not the passive condition ($OR = 1.13$, $CI = [0.82, 1.54]$, $z = 1.09$, $p = 0.28$). Figure 2A shows test accuracy in each condition following a median split on operation span. Active control of study had markedly different consequences depending on participants' operation span, with active control leading to a large, persistent advantage over passive study only among higher WMC participants.

Selections during learning. The next analysis examined participants' selections during learning and whether they could account for differences in test performance described above. Study condition was not related to item selection frequency (multinomial logistic regression, likelihood ratio test: $\chi^2_{(1,7)} = 7.20$, $p = 0.41$), indicating that the aggregate distribution of experienced premise pairs was comparable across active and passive study.

Each learning trial involved a choice between a near option (1–2 positions away from the option selected on the previous trial) and a far option (3+ positions away). By design, near and far options were chosen with equal frequency during passive study. In the active condition participants had a small but significant preference for selecting the near option ($M = 0.56$, $SD = 0.07$; $OR = 1.30$, $CI = [1.21, 1.40]$, $z = 6.86$, $p < .001$).

Near selections may be especially useful if they cause overlapping premise pairs to be experienced in successive trials, which could facilitate integrative encoding when representations of overlapping premise pairs are simultaneously active. I next examined whether the preference to select near items depended on the distance between the near option and the item selected on the previous trial ($dist_{\text{near}} \in \{-2, -1, +1, +2\}$). When $dist_{\text{near}} = +1$, the near option was immediately superordinate to the previously selected item; that is, the near option had appeared as the feedback in the

previous trial.

Figure 2B shows the proportion of near selections as a function of near option distance. In the active condition, the proportion of near selections did not differ from the passive condition when $dist_{\text{near}} = -2$ (OR = 1.05, CI = [0.86, 1.29], z = 0.63, p = 0.53) or $dist_{\text{near}} = +2$ (OR = 1.15, CI = [0.94, 1.41], z = 1.73, p = 0.08). However, there was a higher proportion of near selections when $dist_{\text{near}} = -1$ (OR = 1.29, CI = [1.07, 1.56], z = 3.48, $p < .001$) or $dist_{\text{near}} = +1$ (OR = 1.75, CI = [1.45, 2.11], z = 7.52, $p < .001$). Within the active condition, the proportion of near selections was markedly higher for $dist_{\text{near}} = +1$ than $dist_{\text{near}} = -1$ options (OR = 1.45, CI = [1.20, 1.75], z = 4.95, $p < .001$). In the active condition participants therefore preferred the near option when it was adjacent to the item selected on the previous trial, and this preference was strongest when the option had appeared as feedback in that trial. Although the aggregate frequency of item selection was similar across conditions, this result suggests that active participants generated study sequences in which overlapping premise pairs were more likely to be experienced in successive trials.

Can this tendency to select overlapping items account for the performance benefit in the active condition? A new model of test accuracy was fit for the active condition which included predictors for the proportion of near selections at each level of $dist_{\text{near}}$. There were no significant relationships between accuracy and the proportion of near selections at any distance ($dist_{\text{near}} = -2$: OR = 1.01, CI = [0.70, 1.44], z = 0.06, p = 0.95; $dist_{\text{near}} = -1$: OR = 1.30, CI = [0.92, 1.85], z = 1.87, p = 0.06); $dist_{\text{near}} = +1$: OR = 0.96, CI = [0.67, 1.38], z = -0.27, p = 0.79; $dist_{\text{near}} = +2$: OR = 1.30, CI = [0.90, 1.89], z = 1.75, p = 0.08). The proportion of near selections at any distance was also unrelated to operation span ($dist_{\text{near}} = -2$: OR = 0.93, CI = [0.80, 1.07], z = -1.34, p = 0.18; $dist_{\text{near}} = -1$: OR = 0.98, CI = [0.85, 1.12], z = -0.38, p = 0.70; $dist_{\text{near}} = +1$: OR = 1.07, CI = [0.93, 1.22], z = 1.19, p = 0.23; $dist_{\text{near}} = +2$: OR = 0.98, CI = [0.85, 1.14], z = -0.30, p = 0.77). Thus, the preference to select overlapping options was a general one and could not on its own account for the gap between active and passive performance.

Discussion

This study used a novel TI task to examine whether active control aids the integration of relational knowledge during study. Control over the selection of premise pairs improved performance relative to passive study in both an immediate test and a retest one week later. Symbolic distance effects observed in the active condition strongly imply that this benefit resulted from enhanced integrative encoding, such that active learners relied on an integrated representation of the hierarchy rather than sequential reactivation of premise pairs at test (Acuna, Sanes, & Donoghue, 2002; Zeithamova, Schlichting, & Preston, 2012). The absence of such effects following passive study suggests that integrative encoding was less prevalent when the same participants lacked the opportunity to select premises for themselves.

Active control did not benefit all learners, however, as working memory capacity strongly predicted accuracy in the active condition. Among higher WMC participants, active control produced a $\sim$10% initial advantage over passive study (increasing to $\sim$20% in the retest) and sustained performance across sessions. WMC was unrelated to accuracy in the passive condition, a finding that conflicts with reports that WMC moderates TI under experimenter-controlled conditions (Fales et al., 2003; Libben & Titone, 2008; Titone et al., 2004). This discrepancy may be due to the relative difficulty of passive study in the present task. Previous studies have typically involved smaller hierarchies and scaffolded training sequences in which participants are likely to experience overlapping premises (e.g., Libben & Titone, 2008). With larger hierarchies and greater distances between successive premises, the passive condition used here may have been especially difficult even for participants with higher WMC. An important next step is to evaluate whether the large disadvantage from passive study among higher WMC persists when observing more useful sequences of premises (e.g., when yoked to participants' selections in the active condition).

This study provides the first evidence of systematic search in active TI: Participants strongly preferred to select options that appeared as feedback on the previous trial ($dist_{\text{near}} = +1$). They thereby naturally generated "chained" sequences of overlapping pairs which tend to improve performance in passive conditions relative to random presentation (Halford, 1984). This preference was widespread: 73 of 100 participants chose the $dist_{\text{near}} = +1$ option in more than half of trials in which one appeared, and the proportion of near selections was unrelated to WMC. Although selection of overlapping pairs should facilitate integrative encoding, not everyone benefited from it. One possibility is that only higher WMC individuals capitalize on chained sequences because they maintain representations of premises from trial to trial. Alternatively, higher WMC individuals may be more likely to use an integrated representation of the hierarchy to decide which option to study next (e.g., choosing to learn about the option whose rank is more uncertain). Further work is necessary to determine whether this goal-directed evaluation of options' usefulness during selection contributes to the active advantage among higher WMC individuals.

Finally, it is important to note that participants in this study were aware that there was an underlying hierarchy to be learned. Awareness influences strategy use in TI (Smith & Squire, 2005) and it is unknown how active control might affect performance in its absence. It is likely that active control would enhance elemental encoding in such conditions, perhaps due to the mere opportunity for volitional control (Murty, DuBrow, & Davachi, 2015) or additional metacognitive processing (Kornell, Klein, & Rawson, 2015). An intriguing further possibility is that active control increases the likelihood of becoming aware of an underlying hierarchy by focusing attention on abstract relationships across

study episodes (Henriksson & Enkvist, 2016). This would lend support to the broader notion that active learning not only enriches memory for experienced materials, but also fosters self-directed discovery of abstract, relational knowledge.

References

Acuna, B. D., Sanes, J. N., & Donoghue, J. P. (2002). Cognitive mechanisms of transitive inference. *Experimental Brain Research*, *146*(1), 1–10.

Bainbridge, W. A., Isola, P., & Oliva, A. (2013). The intrinsic memorability of face photographs. *Journal of Experimental Psychology: General*, *142*(4), 1323.

De Soto, C. B., London, M., & Handel, S. (1965). Social reasoning and spatial paralogic. *Journal of Personality and Social Psychology*, *2*(4), 513.

Fales, C. L., Knowlton, B. J., Holyoak, K. J., Geschwind, D. H., Swerdloff, R. S., & Gonzalo, I. G. (2003). Working memory and relational reasoning in Klinefelter syndrome. *Journal of the International Neuropsychological Society*, *9*(6), 839–846.

Frank, M. J., Rudy, J. W., Levy, W. B., & O'Reilly, R. C. (2005). When logic fails: Implicit transitive inference in humans. *Memory & Cognition*, *33*(4), 742–750.

Greene, A. J., Spellman, B. A., Levy, W. B., Dusek, J. A., & Eichenbaum, H. B. (2001). Relational learning with and without awareness: Transitive inference using nonverbal stimuli in humans. *Memory & cognition*, *29*(6), 893–902.

Halford, G. S. (1984). Can young children integrate premises in transitivity and serial order tasks? *Cognitive Psychology*, *16*(1), 65–93.

Henriksson, M. P., & Enkvist, T. (2016). Learning from observation, feedback, and intervention in linear and nonlinear task environments. *The Quarterly Journal of Experimental Psychology*, 1–57.

Hummel, J. E., & Holyoak, K. J. (2001). A process model of human transitive inference. In *Spatial schemas in abstract thought* (pp. 279–305).

Kornell, N., Klein, P. J., & Rawson, K. A. (2015). Retrieval attempts enhance learning, but retrieval success (versus failure) does not matter. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(1), 283.

Kumaran, D. (2013). Schema-driven facilitation of new hierarchy learning in the transitive inference paradigm. *Learning & Memory*, *20*(7), 388–394.

Kumaran, D., & McClelland, J. (2012). Generalization through the recurrent interaction of episodic memories: A model of the hippocampal system. *Psychological Review*, *119*(3), 573.

Lazareva, O. F., & Wasserman, E. A. (2010). Nonverbal transitive inference: Effects of task and awareness on human performance. *Behavioural Processes*, *83*(1), 99–112.

Libben, M., & Titone, D. (2008). The role of awareness and working memory in human transitive inference. *Behavioural processes*, *77*(1), 43–54.

Markant, D., DuBrow, S., Davachi, L., & Gureckis, T. M. (2014). Deconstructing the effect of self-directed study on episodic memory. *Memory & Cognition*, *42*(8), 1211–1224.

Markant, D., & Gureckis, T. M. (2014). Is it better to select or to receive? Learning via active and passive hypothesis testing. *Journal of Experimental Psychology: General*, *143*(1), 94–122.

Markant, D., Ruggeri, A., Gureckis, T. M., & Xu, F. (2016). Enhanced memory as a common effect of active learning. *Mind, Brain, and Education*, *10*(3), 142–152.

Martin, N., & Alsop, B. (2004). Transitive inference and awareness in humans. *Behavioural processes*, *67*(2), 157–165.

Moyer, R. S., & Landauer, T. K. (1967). Time required for judgements of numerical inequality. *Nature*, *215*, 1519–1520.

Murty, V. P., DuBrow, S., & Davachi, L. (2015). The simple act of choosing influences declarative memory. *The Journal of Neuroscience*, *35*(16), 6255–6264.

Shohamy, D., & Wagner, A. D. (2008). Integrating memories in the human brain: hippocampal-midbrain encoding of overlapping events. *Neuron*, *60*(2), 378–389.

Smith, C., & Squire, L. R. (2005). Declarative memory, awareness, and transitive inference. *Journal of Neuroscience*, *25*(44), 10138–10146.

Son, J. Y., Smith, L. B., & Goldstone, R. L. (2011). Connecting instances to promote children's relational reasoning. *Journal of experimental child psychology*, *108*(2), 260–277.

Steyvers, M., Tenenbaum, J., Wagenmakers, E., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science*, *27*(3), 453–489.

Titone, D., Ditman, T., Holzman, P. S., Eichenbaum, H., & Levy, D. L. (2004). Transitive inference in schizophrenia: impairments in relational memory organization. *Schizophrenia research*, *68*(2-3), 235–247.

Trabasso, T., Riley, C. A., & Wilson, E. (1975). The representation of linear order and spatial strategies in reasoning: A developmental study. *Reasoning: Representation and process in children and adults*, 201–229.

Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior research methods*, *37*(3), 498–505.

Voss, J., Gonsalves, B., Federmeier, K., Tranel, D., & Cohen, N. (2011). Hippocampal brain-network coordination during volitional exploratory behavior enhances learning. *Nature Neuroscience*, *14*(1), 115–120.

Zeithamova, D., Schlichting, M. L., & Preston, A. R. (2012). The hippocampus and inferential reasoning: building memories to navigate future decisions. *Frontiers in Human Neuroscience*, *6*.

Model-based Approach with ACT-R about Benefits of Memory-based Strategy on Anomalous Behaviors

Shota Matsubayashi (shota.matsubayashi@nagoya-u.jp)
Kazuhisa Miwa (miwa@is.nagoya-u.ac.jp)
Graduate School of Informatics, Nagoya University
Furo-cho, Chikusa, Aichi, 464-8601, Japan

Hitoshi Terai (terai@fuk.kindai.ac.jp)
Faculty of Humanity-Oriented Science and Engineering, Kindai University
Kaya no Mori 11-6, Iizuka, Fukuoka, 820-8555, Japan

Abstract

Users sometimes face anomalous behaviors of systems, such as machine failures and autonomous agents. Predicting such behaviors of systems is difficult. We investigate the benefits of the memory-based strategy, which focuses on memorization of instances to predict anomalous and regular behaviors of the system, with ACT-R simulations with a cognitive model. In this study, we presumed the parameters defining the encoding processes on anomalous instances and regular instances in the model of the memory-based strategy and performed simulations to verify how these two parameters influence prediction performance. The results of simulations showed that (1) regular instances are not encoded as default values in the memory-based strategy and that (2) such inactivity on regular instances suppresses commission errors of regular instances and does not suppress commission errors of anomalous instances nor omission errors.

Keywords: memory-based strategy; prediction; anomalous behavior; regular behavior; ACT-R

Introduction

There are many various systems around us, and users often predict their behaviors. It is relatively easy for users to predict systems' stationary behaviors by applying schemas (henceforth referred to as "regular behaviors"). However, users sometimes observe that systems' behaviors deviate from regular behaviors (henceforth referred to as "anomalous behaviors"). Predicting anomalous behaviors is effortful (e.g., Besnard & Bastien-Toniazo, 1999; Casner, Geven, & Williams, 2013) and requires users to execute much cognitive processing such as reallocation of cognitive resources (Meyer, Reisenzein, & Schützwohl, 1997). Therefore, it is necessary to process anomalous behaviors and regular behavior differently in order to predict systems' behaviors precisely.

One of the strategies to predict systems' behaviors is the "inference-based strategy," which focuses on inferences and understandings regarding the causal structure from systems' behaviors. The literature from various areas of research show that users apply the inference-based strategy spontaneously when encountering anomalous instances (e.g., Baker et al., 2009; Clary & Tesser, 1983; Howard & Holcombe, 2010; Tremoulet & Feldman, 2000, 2006). Inferences contribute to users' understanding of systems, but these inferences include advanced integration processes of the knowledge and the environment (Darabi, Nelson, & Palanki, 2007); therefore, the inference-based strategy is not always effective for highly complex systems.

We define the "memory-based strategy," which focuses on memorization of instances to predict systems' behaviors without understandings regarding causal structure. A knowledge base, such as a database of prior failure instances, is an example of the memory-based strategy. Experimental studies have demonstrated that the benefits of the memory-based strategy appear in the test situations, which is the same as the learning situations (e.g., Lane, Mathews, Sallas, Prattini, & Sun, 2008).

Our previous study reveals that the memory-based strategy is effective in a high-complexity task and the inference-based strategy is effective in a low-complexity task (Matsubayashi, Miwa, & Terai, in press). This study indicates that the benefits of the memory-based strategy are likely to be provided by the activity in which the instances representing the regular behaviors (henceforth referred to as "regular instances") are not encoded as default values, whereas the instances representing the anomalous behaviors (henceforth referred to as "anomalous instances") are intentionally encoded. In this study, we investigate these features of the memory-based strategy in detail, with a cognitive model.

First, we review our argument that regular instances are not encoded in the memory-based strategy by reproducing the human data in the psychological experiment. We presume the two parameters defining the encoding processes on anomalous instances and regular instances, and then examine whether the simulated data with inactivity of encoding regular instances provide a good fit to the human data. Second, we reveal why the benefits of the inactivity of encoding regular instances appear by confirming the performance with settings of two encoding parameters. Specifically, when the parameters are set for encoding not only anomalous instances but also regular instances, what happens to the simulated performance data?

Experimental Task

Stimulus

The experimental task required participants to predict the final position of the ball based on its observed movement.

The screen used in this task comprises a visible region and an invisible region (see Figure 1). A hidden object is placed in the invisible region. If the ball makes contact with the object, it changes its direction, whose trajectory is defined as an anomalous instance. Conversely, a regular instance is generated when the ball goes straight without direction changes. The ball is ejected from a certain initial position in the outer frame and at a certain angle. The ball is temporarily invisible while it passes through the invisible region. The ball becomes visible again when it enters the visible region. The ball's movement finally stops at the outer frame. Hereafter, an initial position and an initial angle of the trajectory are defined as "input," and a final position and a final angle are defined as "output."

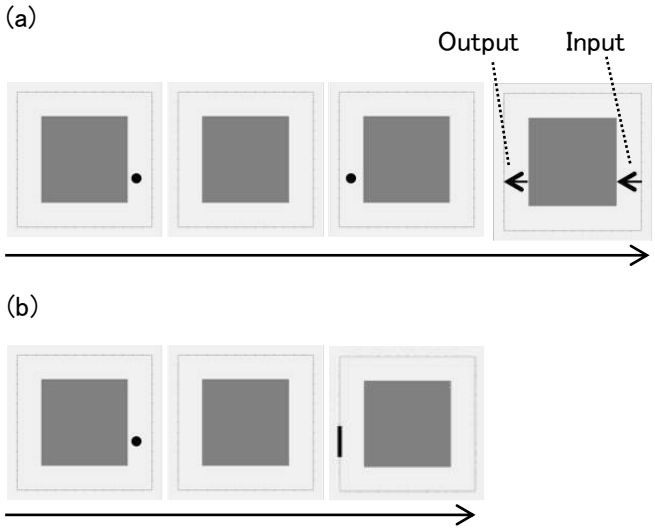


Figure 1: Overview of the task. (a) In the observation phase, the movement of the ball is presented and then the confirmation screen is displayed. (b) In the test phase, participants can move the paddle.

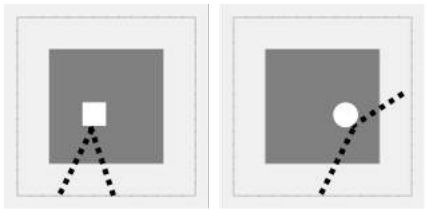


Figure 2: Difficulty settings of the task. Shapes of hidden objects and examples of trajectories in a low-complexity task (left) and in a high-complexity task (right).

Table 1: Composition of the trials in blocks 2−5.

Phase	Instance	Experience
Observation	Anomalous (3)	
	Regular (9)	
Test	Anomalous (6)	Novel (3)
		Experienced (3)
	Regular (6)	Novel (3)
		Experienced (3)

The observation phase and the test phase are alternated repeatedly in this task. In the observation phase, participants observe the movement of the ball from its ejection (i.e., input) until its stoppage in the outer frame (i.e., output). Participants are also shown the confirmation screen with two arrows representing the input and the output (see Figure 1a).

In the subsequent test phase, the ball stops as soon as it enters the invisible region, and a paddle is also displayed (see Figure 1b). To predict the final position of the ball and catch it with the paddle, participants are required to move the paddle with a left click button and determine its position with a right click button. The paddle is displayed at the same location in which the ball would arrive if it went straight without direction changes. In other words, it is not necessary to move the paddle in regular instances but is necessary to move the paddle in anomalous instances to catch the ball. The number of correct trials in which the range of the paddle includes the genuine final position of the ball is regarded as the prediction performance. No feedback on the predictions is provided to participants.

The shapes of the hidden objects in the invisible region determine the complexity of the tasks (see Figure 2). Anomalous instances follow a simple trajectory in a low-complexity task with a square-shaped object and a complex trajectory in a high-complexity task with a circular object.

Procedure

Prior to the observation phase, participants were informed that they were required to predict, as precisely as possible, the final position of the ball in the test phase. Participants were instructed to focus on and memorize the two arrows representing the input and the output in the confirmation screen in the observation phase. Participants are expected to use the memory-based strategy and encode the combination composed of an initial position, an initial angle, and a final position as an instance comprised of the input and the output.

The movement of the ball constituted one sequence, and each sequence constitutes one trial. A block comprised 12 trials in the observation phase and 12 trials in the test phase. All trials in block 1 corresponded to regular instances in the observation phase and in the test phase. Trials comprised three anomalous instances and nine regular instances in the observation phase in blocks 2−5. In the test phase, trials comprised six anomalous instances and six regular instances. In addition, each trial in the test phase comprised three novel instances, which were shown only at this time, and three experienced instances, which had been shown in the previous observation phase (see Table 1).

Participants implemented a 5-block low-complexity task and a 5-block high-complexity task. The positions and the shapes of hidden objects are consistent throughout all the trials in each task.

Summary of Psychological Experiment Results

Overall, the data of 24 participants were analyzed. A summary of the results is described here, and the details are mentioned with the simulation results (see Figure 3).

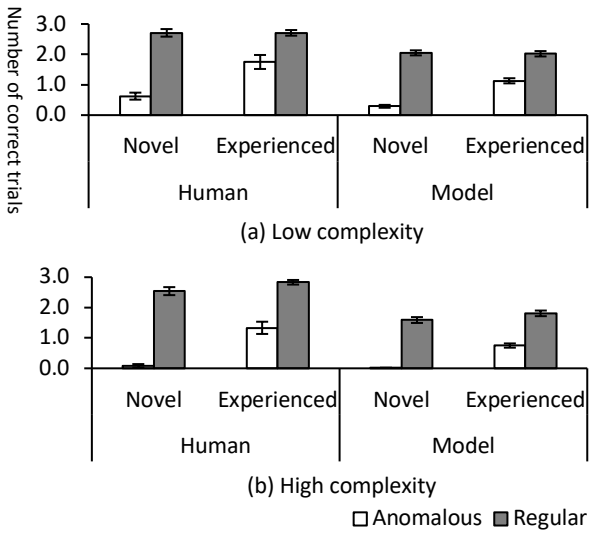


Figure 3: Prediction performance in block 5. Error bars represent standard errors.

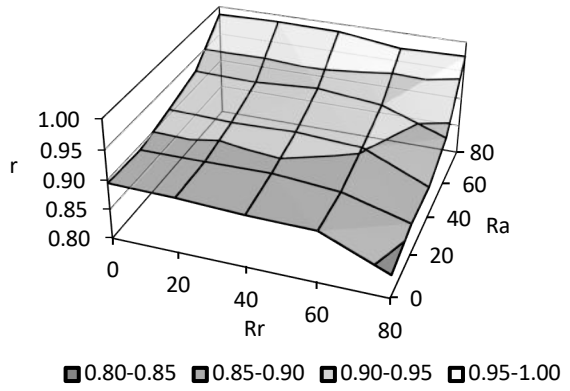


Figure 4: Correlation coefficients on prediction performance in block 5 between the simulated data and human data. Ra represents the rehearsal probability of anomalous instances and Rr represents the rehearsal probability of regular instances.

Statistical results show that the interaction of the instance factor (anomalous/regular) and the experience factor (novel/experienced) was significant for the prediction performance in each task in block 5 (low-complexity: $F(1, 23) = 13.8$, $p < .005$, $\eta^2 = .60$; high-complexity: $F(1, 23) = 10.7$, $p < .005$, $\eta^2 = .47$). Specifically, the performances for anomalous-experienced instances are higher than those for anomalous-novel instances (low-complexity: $F(1, 46) = 27.6$, $p < .001$; high-complexity: $F(1, 46) = 41.6$, $p < .001$). This result indicates that anomalous instances were encoded in the observation phase. Additionally, no differences are observed in the performance for regular-novel instances and for regular-experienced instances (low-complexity: $F(1, 46) = 0.0$, $p = 1.0$, $r = .00$; high-complexity: $F(1, 46) = 2.2$, $p = .13$, $r = .32$). This result indicates that regular instances were not encoded in the observation phase; therefore, they were not retrieved even for regular-experienced instances in the test phase.

Simulations with Cognitive Model

This study adopts ACT-R simulations (Anderson, 2007) with a cognitive model to investigate the details of processing. Two retrieval errors critical to the memory-based strategy are available in ACT-R, that is, commission errors representing that wrong instances are retrieved and omission errors representing that encoded instances are not retrieved.

In this study, we examine the following two points with simulations. First, we reveal the features of the memory-based strategy by performing simulations with two parameters defining the encoding processes of anomalous instances and regular instances. If the simulated data with the parameters meaning inactivity of encoding regular instances provide a good fit to the human data, our argument regarding such an inactivity is supported. Second, we reveal the reason why the benefits of inactivity of encoding regular instances appear in the memory-based strategy. Two research questions are drawn: How does the parameter on encoding regular instances decrease the prediction performance? What type of retrieval error is the cause of such decline in performance?

Simulation Settings

The following is the outline of the memory-based strategy model. In the observation phase, the model detects an input arrow and an output arrow, reads the position and the angle of each arrow, and then encodes them as a chunk in the declarative memory. This chunk comprises three slots—the initial position, the initial angle, and the final position. Next, the model runs rehearsals by repeating retrievals of the chunk. The rehearsal probability parameters determine whether the model continues to run a rehearsal on every rehearsal. There are two types of rehearsal probability parameters. If an input angle is different from an output angle, the model regards this trial as "an anomalous instance" and runs rehearsals on the basis of the rehearsal probability of anomalous instances (henceforth referred to as "Ra"). Additionally, if an input angle is the same as an output angle, the model regards this trial as "a regular instance" and runs rehearsals on the basis of the rehearsal probability of regular instances (henceforth referred to as "Rr").

In the subsequent test phase, the model reads the position and the angle of an input arrow and makes a retrieval request to declarative memory with them as a clue. If the model fails to retrieve an instance or the retrieved final position is included in the range of the paddle, the model does not move the paddle. Otherwise, the model moves it to the retrieved final position with left click button. After that, the model confirms the position of the paddle with a right click button.

Making retrieval errors on two adjacent initial positions is likely because these two positions are highly similar and difficult to distinguish from each other. Therefore, the similarity parameters between two adjacent initial positions are set to −0.5. Other similarity parameters are set to −1.0 as default.

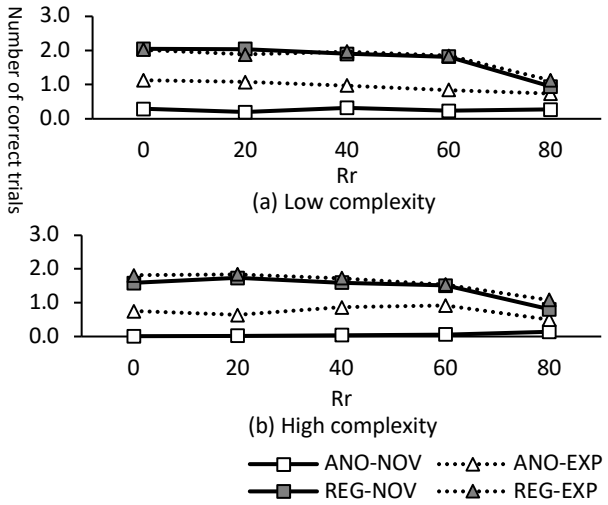


Figure 5: Variations of prediction performance on a function of the rehearsal probability of regular instances (Rr). ANO, REG, NOV, and EXP represent anomalous, regular, novel, and experienced respectively.

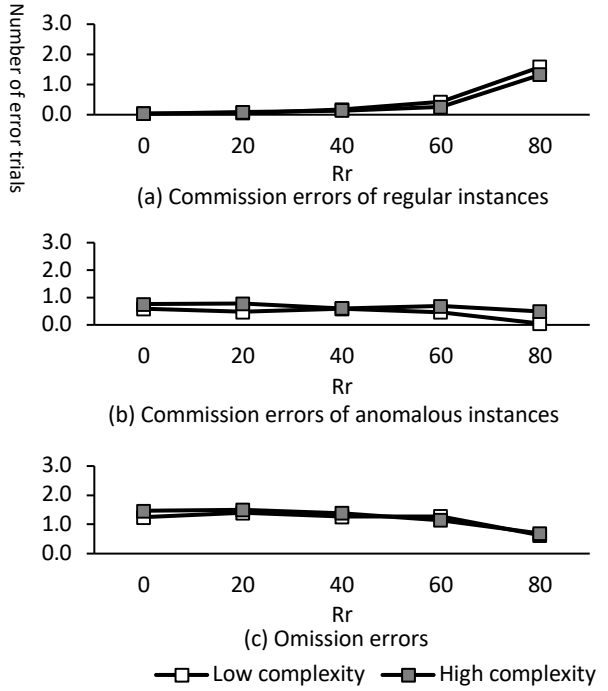


Figure 6: Variations of retrieval errors on anomalous-experienced instances on a function of the rehearsal probability of regular instances (Rr).

Each Ra and Rr has five levels; therefore, 25 parameter combinations are simulated. The five levels of rehearsal probability correspond to 0%, 20%, 40%, 60%, and 80%, that is, the expected values of the number of rehearsals are 0.00, 0.25, 0.67, 1.50, and 4.00 respectively.

Results of Simulations

Best Parameters First, in order to investigate the features of the memory-based strategy, we calculate correlation coefficients between the simulated data and the human data on prediction performance in block 5. Figure 4 shows that the simulated data in which anomalous instances are encoded sufficiently and regular instances are not encoded provide a best fit to the human data. Prediction performance at Ra 80% and Rr 0% is reproduced well. Specifically, there is no difference in the performance for regular-experienced instances and for regular-novel instances, and the performance for anomalous-experienced instances is higher than that for anomalous-novel instances (see Figure 3). These results support our argument that regular instances are not encoded and anomalous instances are encoded. Notably, the simulated data are wholly lower than the human data. We will discuss this topic in Discussion and Conclusion.

Effects of Encoding Regular Instances Second, we investigate the reason why the benefits of the inactivity of encoding regular instances appear in the memory-based strategy. What happens to the prediction performance when the Rr parameter is set to 20% or higher?偏

Figure 5 represents the variations of the prediction performance based on a function of Rr. The results show that the performances for anomalous-experienced instances decrease gradually as Rr increases and that the performances for regular instances decrease rapidly when Rr increases to 80%.

We verify what retrieval error is the cause of decline in performance. There are three types of errors on anomalous instances—commission errors in which regular instances are retrieved incorrectly, commission errors in which another anomalous instances are retrieved, and omission errors, in which no instance is retrieved. On the other hand, there are three types errors on regular instances—commission errors in which another regular instances are retrieved, commission errors in which anomalous instances are retrieved incorrectly, and omission errors. However, the omission errors on regular instances do not correspond to retrieval errors because participants can catch the ball even if they do not move the paddle and such trials are regarded as successful prediction. Therefore, we verify the only two commission errors on regular instances as possible causes of decline in performance for regular instances.

Figure 6 represents the variations of retrieval errors on anomalous-experienced instances. The results show that commission errors of regular instances increase as Rr increases to 80%. Additionally, there is no change on commission errors of anomalous instances and on omission errors from Rr 0% to 60%, but rapid drops appear in these errors at Rr 80% in each task. We found that the cause of declines in performance for anomalous-experienced instances is the commission errors in which regular instances are retrieved inappropriately.

Subsequently, Figure 7 represents the transitions of retrieval errors on regular instances. The results show that commission errors of regular instances increase as Rr increases to 80% and that commission errors of anomalous instances decrease at 80%. As a result, the cause of declines in performance for regular instances is the commission errors in which another regular instance is retrieved.

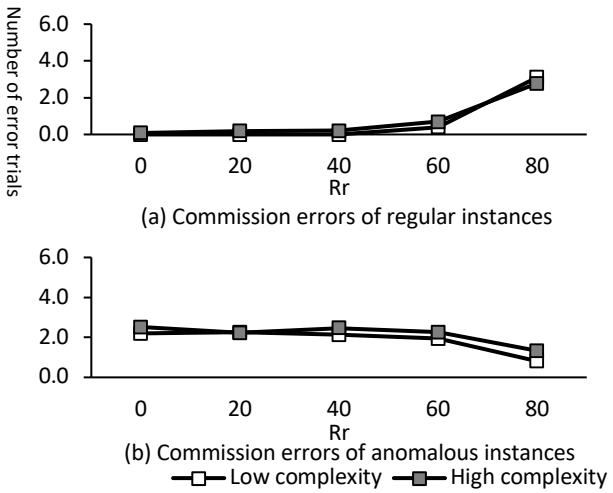


Figure 7: Variations of retrieval errors on regular instances on a function of the rehearsal probability of regular instances (Rr).

In summary, we found that the commission errors of regular instances contribute to the declines in performances on anomalous-experienced instances and on regular instances. That is, encoding regular instances in the memory-based strategy leads to increases in retrieval of inappropriate regular instances. In other words, participants using the memory-based strategy are likely to inhibit the commission errors of regular instances by not encoding regular instances. Additionally, and notably, commission errors of anomalous instances and omission errors do not increase according to Rr and decrease at Rr 80%. We will discuss this topic in Discussion and Conclusion.

Discussion and Conclusion

In this study, we performed the simulations of the processing of the memory-based strategy with a cognitive model and revealed the following two points in the context of the prediction on anomalous behaviors. First, by reproducing the human data, we found that the results support our argument that regular instances are not encoded as default value, and anomalous instances are encoded in the memory-based strategy. Second, the simulations in prediction performance with settings of encoding parameters show that the benefits of the memory-based strategy appear when such inactivity on regular instances inhibits commission errors of inappropriate regular instances and does not inhibit commission errors of anomalous instances nor omission errors.

Processes of Memory-based Strategy

We found that the simulated data in which regular instances are not encoded provide a best fit to the human data. This result confirms our argument that regular instances are not encoded in the memory-based strategy. Additionally, this result corresponds to the results in our previous experiment about participants' subjective evaluations toward anomalous instances and regular instances (Matsubayashi et al., in press).

Although the tendencies on prediction performance in simulations are reproduced well, the simulated data are wholly lower than the human data. This result indicates that participants in the memory-based strategy could perform other additional processing than the encoding processing that we presumed in the current model when they observed various instances. For example, participants might integrate some similar instances into one chunk, make an inference regarding the causal structure through the anomalous trajectories, or revise relevant schema (Meyer et al., 1997).

The studies on category learning have presumed the models that implement multiple processing when observing an instance (Nosofsky, Palmeri, & McKinley, 1994). Furthermore, our previous study indicates that participants adopt the inference-based strategy and the memory-based strategy when not provided explicit instructions about strategies (Matsubayashi et al., in press). The human data cited in this article correspond to the data when participants were urged to use the memory-based strategy, but we cannot dismiss the possibility that the participants use the inference-based strategy alongside. However, the inference-based strategy is possible to consume much cognitive resources (Darabi et al., 2007); therefore, using both strategies could reduce prediction performance. Notably, the trade-off between the costs and the benefits on two strategies must be verified for future work.

Benefits of Memory-based Strategy

The benefits of the memory-based strategy appear because of the inhibition of retrieval errors of inappropriate regular instances. The inactivity on regular instances inhibits commission errors in which regular instances are retrieved incorrectly on anomalous instances and commission errors in which another inappropriate regular instances are retrieved on in-situ regular instances. In summary, such inactivity of the memory-based strategy has a critical role in preventing confusion in encoded instances when they are retrieved and in saving cognitive resources to encode instances. The results of the simulations show that when regulars are encoded as frequently as anomalous instances are, more commission errors of regular instances occur, which indicates that it is critical not to encode regular instances in the memory-based strategy.

On the other hand, the commission errors of anomalous instances or the omission errors do not increase even if regular instances are encoded. Furthermore, we found that these two errors decrease only if regular instances are encoded as frequently as anomalous instances are. These decreases seem to occur, confounded with the effect of the increase in the commission errors of regular instances. If regular instances are encoded as frequently as anomalous instances are, the current model stores three anomalous instances and nine regular instances in the declarative memory in each block, with similar activation levels. Consequently, regular instances are more likely to be retrieved than anomalous instances, which results in a relative decrease in the commission errors of anomalous instances

and omission errors. However, the benefits of encoding regular instances do not appear because the whole prediction performance decreases even if these two errors decrease.

The features of cognitive processing on anomalous instances have been verified with visual search tasks. Studies have revealed that the objects incongruent with the schema of the scene are difficult to identify (Mudrik, Deouell, & Lamy, 2011) and these objects are represented internally prior to the objects congruent with the schema (Hollingworth & Henderson, 2000). Our findings that there are no benefits of encoding regular instances are not contradictory to such studies. Furthermore, our study reveals the cognitive processing on regular instances, which are congruent with the schema, while other studies have referred to that on anomalous instances, which are incongruent with the schema. Model-based approaches can clarify the internal cognitive processes that are difficult to observe and have been used in various areas, such as category learning (Erickson & Kruschke, 1998). Particularly, studies on the cognitive model about instance-based learning have revealed decision making processes from experience (Gonzalez & Dutt, 2011; Paik & Pirolli, 2013). Our findings regarding regular instances could not have been obtained without the simulations with a cognitive model.

In this study, we performed simulations of the processing of the memory-based strategy with a cognitive model from a perspective of predicting anomalous behaviors. First, by reproducing the human data, we found the results that support our argument that regular instances are not encoded as default values and anomalous instances are encoded in the memory-based strategy. Second, simulations in performance with encoding parameters clarified that the benefits of the memory-based strategy appear when such inactivity on regular instances inhibits the commission errors of inappropriate regular instances and does not inhibit the commission errors of anomalous instances nor the omission errors.

Acknowledgments

This research was supported by the JST-Mirai Program from Japan Science and Technology Agency.

References

Anderson, J. R. (2007). *How Can the Human Mind Occur in the Physical Universe?* NY: Oxford University Press.

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329-349.

Besnard, D., & Bastien-Toniazzo, M. (1999). Expert error in trouble-shooting: An exploratory study in electronics. *International Journal of Human Computer Studies*, *50*(5), 391-405.

Casner, S. M., Geven, R. W., & Williams, K. T. (2013). The effectiveness of airline pilot training for abnormal events. *Human Factors*, *55*(3), 477-485.

Clary, E. G., & Tesser, A. (1983). Reactions to Unexpected Events: The Naive Scientist and Interpretive Activity. *Personality and Social Psychology Bulletin*, *9*(4), 609-620.

Darabi, A. A., Nelson, D. W., & Palanki, S. (2007). Acquisition of troubleshooting skills in a computer simulation: Worked example vs. conventional problem solving instructional strategies. *Computers in Human Behavior*, *23*(4), 1809-1819.

Erickson, M. A., & Kruschke, J. K. (1998). Rules and Exemplars in Category Learning. *Journal of Experimental Psychology: General*, *127*(2), 107-140.

Gonzalez, C., & Dutt, V. (2011). Instance-Based Learning: Integrating Sampling and Repeated Decisions From Experience. *Psychological Review*, *118*(4), 523-551.

Hollingworth, A., & Henderson, J. M. (2000). Semantic informativeness mediates the detection of changes in natural scenes. *Visual Cognition*, *7*(1–3), 213-235.

Howard, C. J., & Holcombe, A. O. (2010). Unexpected changes in direction of motion attract attention. *Attention, Perception, & Psychophysics*, *72*(8), 2087-2095.

Lane, S. M., Mathews, R. C., Sallas, B., Prattini, R., & Sun, R. (2008). Facilitative interactions of model- and experience-based processes: Implications for type and flexibility of representation. *Memory & Cognition*, *36*(1), 157-169.

Matsubayashi, S., Miwa, K., & Terai, H. (in press). Empirical investigation of memory-based strategy on anomalous behavior. *Japanese Journal of Psychology*.

Meyer, W.-U., Reisenzein, R., & Schützwohl, A. (1997). Toward a Process Analysis of Emotions: The Case of Surprise. *Motivation and Emotion*, *21*(3), 251-274.

Mudrik, L., Deouell, L. Y., & Lamy, D. (2011). Scene congruency biases Binocular Rivalry. *Consciousness and Cognition*, *20*(3), 756-767.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, *101*(1), 53-79.

Paik, J., & Pirolli, P. (2013). An ACT-R Model of Sensemaking in Geospatial Intelligence Tasks, *Proceedings of the 22nd Annual Conference on Behavior Representation in Modeling and Simulation*.

Tremoulet, P. D., & Feldman, J. (2000). Perception of animacy from the motion of a single object. *Perception*, *29*(8), 943-951.

Tremoulet, P. D., & Feldman, J. (2006). The influence of spatial context and the role of intentionality in the interpretation of animacy from motion. *Perception and Psychophysics*, *68*(6), 1047-1058.

Modeling Children's Early Linguistic Productivity Through the Automatic Discovery and Use of Lexically-based Frames

Stewart M. McCauley (stewart-mccauley@uiowa.edu)
Department of Communication Sciences and Disorders, University of Iowa, Iowa City, IA 52242

Morten H. Christiansen (christiansen@cornell.edu)
Department of Psychology, Cornell University, Ithaca, NY 14853

Abstract

A central question for cognitive science is whether children's linguistic productivity can be captured by item-based learning, or whether the learner must be guided by abstract, system-wide principles governed by innate constraints. Here, we present a computational model of early language acquisition which learns to discover and use lexically-based frames in a fully incremental, on-line fashion. The model is rooted in simple prediction- and recognition-based processes, subject to the same memory limitations as language learners. When exposed to English corpora of child-directed speech, the model is able learn developmentally plausible frames and use them to capture over 70% of the utterances produced by target children aged 2 to 5. Across a typologically diverse range of 29 languages, the model is able to capture over 68% of child utterances. Together, these findings suggest that much of children's early linguistic productivity can be captured by item-based learning through computationally simple mechanisms.

Keywords: language learning; language acquisition; usage-based approaches; computational modeling; chunking

Introduction

By four years of age, most children have mastered the basic grammatical structures of their native language, an achievement marking the transition to a seemingly unbounded capacity for communicating novel information. But how is such open-endedness possible, given the finite, noisy nature of the input? This is among the foundational questions of cognitive science. For over half a century, researchers have argued that children's learning is guided by system-wide, abstract principles and constrained by innate biases (e.g., Chomsky, 1965). In recent decades, an alternative perspective has emerged in the form of usage-based approaches, which hold that children's linguistic productivity emerges gradually as a process of storing and abstracting over the input (e.g., Tomasello, 2003). In this framework, children's earliest steps towards unbounded productivity come in the form of lexically-based frames: through knowledge of partially overlapping sequences, children form schemas with slots that are filled according to semantic, pragmatic, or phonological constraints (e.g., Braine, 1963).

Among the earliest quantitative studies offering evidence for lexical frames was that of Lieven, Behrens, Speares, and Tomasello (2003), who used a technique known as the "traceback method" to analyze the speech of a single child during its second year. Lieven et al. found that a high proportion of the child's linguistic productivity—utterances which went beyond frozen or recycled sequences to feature novel word combinations—could be explained in terms of lexically-specific frames, such as "*there's the* __ ." Subsequent work improved on the original traceback method and yielded similar findings (e.g., Lieven, Salomo, & Tomasello, 2003).

As highlighted by other researchers (e.g., Kol, Nir, & Wintner, 2014) the traceback method is not automated and is therefore severely limited in terms of the range of corpora and languages to which it can be applied. Moreover, the lack of a computationally explicit formulation means that the general approach does not make specific commitments to the types of learning mechanisms or representations that allow productivity to emerge from lexically-based representations.

This problem highlights a general lack of computational work examining item-based learning as a starting point for linguistic abstraction, which is reflected in the imprecise language with which usage-based theory is often discussed. For instance, researchers have appealed to complex psychological constructs such as analogical reasoning to explain lexically-based frames (e.g., Gentner & Namy, 2006; Tomasello, 2003). Even computational studies examining the transition from item-based learning to abstraction have appealed to analogy while remaining agnostic as to the lower-level mechanisms supporting it (e.g., Bod, 2009).

By contrast, we aim to provide an account of early abstraction which is rooted in basic processes of prediction and recognition. Moreover, we wish to capture such learning in a way that is consistent with the myriad sensory and memory limitations imposed on the learner (as discussed in Christiansen & Chater, 2016). This requires a fully incremental and on-line learning model, in line with memory constraints that force reliance on local rather than global syntactic information. It also means capturing learning in a way that is fully usage-based in the sense that all learning takes place in the context of specific processing events.

Modeling Children's Discovery and Use of Lexically-based Frames

Here, we seek to model children's discovery and use of lexical frames by modifying an existing usage-based computational framework, known as the Chunk-Based Learner (CBL; McCauley & Christiansen, 2014, 2019). Inspired by the aforementioned memory constraints, the CBL model aims to recreate individual children's utterances by learning from the linguistic input to which they have been exposed. The model offers strong performance across a typologically diverse range of languages (McCauley & Christiansen, 2019) while capturing psycholinguistic data from both children (McCauley & Christiansen, 2014) and adults (Grimm, Cassani, Gillis, & Daelemans, 2017). Importantly, previous research has only used CBL to model the discovery and use of concrete multiword units. In the present study, we implement this pre-existing model and modify it to support the incremental, on-line discovery and use of lexically-based frames.

In what follows, we describe the basic workings of the CBL model as well as the modifications we applied to enable the learning of lexical frames. Next, we examine qualitative and quantitative properties of the frames discovered by the model when exposed to corpora of English child-directed speech. We also evaluate the model's ability to use these frames in a sentence production task, exploring the extent to which they can support early linguistic productivity. Finally, we look at the model's ability to use frames in this sentence production task across a typologically diverse array of 29 different languages.

Experiment 1: Modeling the Development of Lexically-based Frames in English

The CBL Model

The model has been described in detail in previous work (e.g., McCauley & Christiansen, 2019). We therefore briefly provide sufficient information to implement the model. The model processes the input corpus on a word-by-word basis, tracking low-level frequency information for words and word pairs (bigrams). This information is used on-line to calculate the backward transition probability (BTP) between words. By maintaining a running average of BTP over previously seen word pairs and using it as a threshold, the model classifies BTPs linking words as either *high* or *low*. High BTPs are used to group words together to form part of a chunk, while low probabilities are used to define chunk boundaries. When a boundary is placed, the preceding word(s)—there is no *a priori* limit on the size of a chunk—are placed as a unit in the model's *chunk inventory*. When the model encounters a previously-discovered chunk in the input, its frequency count is incremented by 1. The resulting chunk inventory thus contains a mix of single-word and multiword units. The model maintains frequency counts for pairs of chunks occurring together, which supports the incremental construction of utterances during production.

The model also uses its chunk inventory on-line while processing the input. Through a combination of prediction- and recognition-based processing, knowledge of previously discovered chunks can assist in further discovery: when a word-pair is encountered, if it has occurred at least twice as part of an existing chunk, it is automatically grouped together (regardless of BTP). Otherwise, the BTP is evaluated against the running average threshold as described above.

A record of the model's on-line chunking of utterances is maintained for later evaluation against the output of a parser. CBL's ability to approximate the output of shallow parsers cross-linguistically has been suggested to capture key aspects of comprehension (cf. McCauley & Christiansen, 2019). The model also aims to capture key aspects of production: as the model makes its way through a corpus of child-directed speech, it encounters utterances produced by the target child of the corpus, at which point the production side of the model comes into play. The model must produce its own utterance by generalizing from the chunks and statistics it has learned up to that point in the simulation. This task is used to evaluate our version of the model and is described below in the subsection entitled Sentence Production Task.

Modifications to the CBL Model

To enable the on-line discovery and use of lexical frames, we made some slight changes to the original CBL implementation. When the model has discovered 5 or more multiword chunks which overlap in all but one position, it creates a lexical frame—a chunk with an empty slot—and stores it in the chunk inventory. When chunks matching this frame are encountered, the frame's frequency count is incremented, as are the counts of matching chunks. The 5+ criterion was selected in light of previous corpus studies of evidence for lexical frames in child-directed speech (e.g., Cameron-Faulkner, Lieven, & Tomasello, 2003, who used a criterion of 4+ in their analyses). As the original version of CBL already uses its chunk inventory during on-line processing, we felt this change was in keeping with the model's intended psychological features.

As an example of frame creation, consider an instance in which the model has already discovered and used the chunks *in the box*, *in the tub, in the bag,* and *in the chair.* When the model discovers the chunk *in the cup*, it also discovers the frame *in the* __ as an automatic generalization over the previous multiword chunks. Both *in the cup* and *in the* __ are initialized in the chunk inventory with counts of 1, the starting frequency value for newly-discovered chunks. The frame's count is then incremented by 1 when the model later encounters *in the box*, a previously discovered chunk, as is the count for that chunk. The frame's count is also incremented by 1 again when the model discovers a new chunk, *in the sink*, and so forth.

As described in the below section entitled Sentence Production Task, the model can rely on its knowledge of lexical frames during production.

Input Corpora

Rather than aggregate across multiple corpora, each of our simulations involved exposing the model to a single corpus of child-directed speech. We selected, from the English language portion of the CHILDES database (MacWhinney, 2000) all corpora meeting the following three criteria: i) contained at least 50,000 words; ii) featured a multiword child-to-adult utterance ratio of at least 1:10; and iii) spanned at least a 6-month period in terms of the target child's age across the corpus. These criteria were met by individual corpora for 43 English-learning children (US: 25; UK: 18). Tags and punctuation were removed from the corpora, leaving, for each utterance, only speaker identifiers and the original sequence of words.

Learning Lexical Frames in English

Across the entire set of 43 simulations, the model discovered a mean of 14 lexical frames per 10,000 words of input. Rather than leading to a combinatorial explosion of units—as might suggest psychological implausibility, or coverage due to trivial factors in subsequent evaluation tasks--frames made up just 5% of the total chunk inventory for the simulation involving the largest corpus (*Thomas*; Maslen, Theakston, Lieven, & Tomasello, 2004), with smaller percentages for smaller corpora (3% on average).

To offer a sense of the qualitative nature of the model's lexical frames, we show, for the largest corpus (Thomas), a range of frequent frames as well as less-frequent but developmentally interesting frames.

Table 1: Frequent and Developmentally Interesting Frames Learned from the Thomas (Dense) Corpus and Corresponding Counts in the Chunk Inventory

Frequent Frames		**Developmentally Interesting Frames**	
the __	56117	*a little __*	2131
a __	42937	*what's __*	2122
your __	8366	*a big __*	1401
in the __	7718	*are you going to __*	1196
on the __	6950	*what do you __*	945
this __	6742	*more __*	837
that __	6343	*I want to __*	427
very __	4911	*on __ own*	228
I don't __	3386	*the red __*	120
going to __	3348	*more __*	103

As can be seen in Table 1, even though slots are allowed anywhere in a chunk, the vast majority of lexical frames featured a slot in the final position. Across all the English corpora, slot-final frames accounted for a large percentage of overall frames utilized, ranging from 85% to 98%.

There is good overlap between the frames appearing in Table 1 and frames postulated by other researchers on the basis of corpus analyses, including some of the earliest to advance the notion of lexical frames: for instance, *more __* is one of the first frames identified in Braine (1963).

Next, we turn to the question of whether the lexical frames discovered by the model can offer insights into the nature children's early productivity. To this end, we evaluate the model according to its ability to capture children's actual utterances in these corpora, and measure the extent to which the model's lexical frames can support production above and beyond concrete multiword chunks.

Sentence Production Task

The sentence production task was based on the bag-of-words incremental generation task first described by Chang, Lieven, and Tomasello (2008). The task rests on the simplifying assumption that the overall message the child wishes to convey can be—very roughly—approximated by treating the utterance as an unordered bag-of-words. When the model encounters a multiword utterance produced by the target child of a corpus, its task is to sequence the items in the bag to produce its own utterance, using only the words and statistics it has discovered prior to that point.

We used a nearly identical version of the task to that described by McCauley and Christiansen (2019): following psycholinguistic evidence for children's use of multiword units (see above), the model was allowed to draw upon previously discovered chunks to populate the bag-of-words. To produce an utterance, the model begins by selecting from the bag the word or chunk with the highest transition probability given the start-of-utterance marker (a marker preceding every line in the corpus). At each subsequent time step the model removes and produces the word or chunk with the highest probability *given the most recently placed chunk*. This process continues until the bag is empty.

Thus, production is implemented as fully incremental, chunk-to-chunk process, relying entirely on local information. In other words, there is no global whole-sentence optimization. In this sense, the model captures the sorts of memory limitations described in the introduction.

Where our version of the task differed from that described by McCauley and Christiansen (2019) was in the additional use of lexical frames: if the model lacked experience of a given sequence in the child's utterance, but had learned a lexical frame capable of fitting that sequence, it was allowed to utilize the frame in the bag-of-words task. Consider the model's attempt to produce the child utterance: "*red one stuck in the jam.*" In a case in which the model has discovered the lexical frame *in the __* but has never encountered the sequence *in the jam* in the input, the model is allowed to use the lexical frame to complete this pattern. Statistics are then calculated over the frame itself, as if it were a fully concrete chunk.

Gold Standard for Sentence Production Task Following each production attempt, the model's utterance is scored against the child's original utterance according to an all-or-nothing scoring metric: if the two utterances do not match completely, a score of 0 is assigned. Otherwise, a score of 1 is given. Thus, the overall accuracy of the model

across a corpus can be calculated as a percentage of correctly produced multiword utterances (single-word utterances are excluded to avoid inflating performance). We call this the *Sentence Production Accuracy* (SPA) score.

Alternate Distributional Models We evaluate the model against two baseline models: the first is the basic version of CBL used as a starting point for the present study (described above; cf. McCauley & Christiansen, 2019). The second is a standard trigram model; this approach was selected as a baseline due to its widespread use and generally robust performance as a probabilistic language model across a range of genres (Manning & Schütze, 1999).

Results and Discussion Across all 43 English simulations, the lexical frames version of the CBL model (CBL+LF) achieved a median Sentence Production Accuracy of 71.3% (mean: 69.5%). This is compared to a median score of 58.5% (mean: 57.8%) for the original CBL model and a median score of just 45.7% (mean: 45.1%) for the trigram (3G) baseline. The distribution of scores for each model are shown in Figure 1.

A linear mixed-effects model fit using logit-transformed SPA scores, with child as a random factor, confirmed that both the CBL+LF model (β=0.53, t=22.8, $p < 0.001$) and the 3G model (β=-0.53, t=-22.6, $p < 0.001$) differed significantly from the original CBL model, in opposite directions.[1]

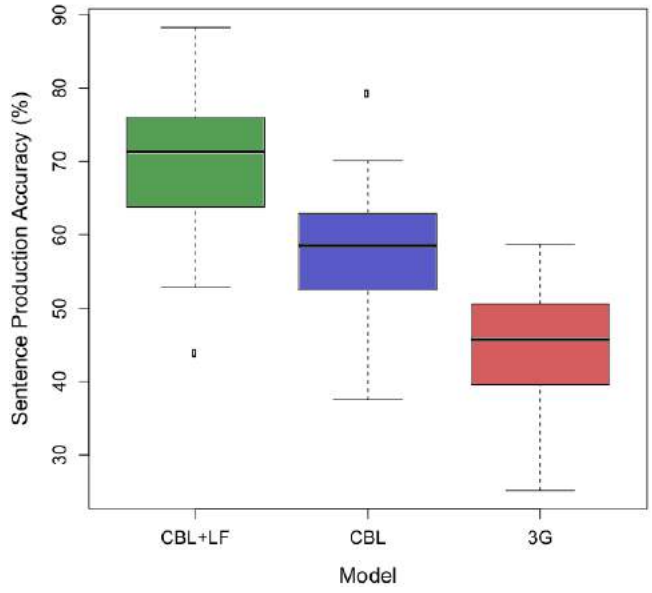


Fig. 1: Box and whisker plots depicting English Sentence Production Accuracy (%) for the model and its baselines.

Thus, in addition to discovering developmentally and psychologically plausible lexical frames, the CBL+LF model was able to use these units to improve upon the CBL model's production performance by nearly 12 percentage-points, surpassing the performance of a standard trigram model by nearly 25 percentage-points.

[1]All p-values computed via Satterthwaite approximation.

Experiment 2: Modeling the Development of Lexically-based Frames Across Typologically Diverse Languages

The vast majority of computational modeling work in the study of language acquisition has focused on English. It is crucial, however, to determine whether the types of linguistic representations and learning mechanisms we ascribe to children can plausibly accommodate languages with typological features that differ greatly from those of English. In the case of the present model, which uses multiword units as much of the basis for learning and processing, morphological features are of particular interest.

A previous study using the CBL model has demonstrated that multiword units do indeed facilitate production for typologically diverse languages, including morphologically rich languages (McCauley & Christiansen, 2019). Here, we ask the question of to what extent limited productivity based on lexical frames can improve the ability of CBL to capture the utterances of children learning a typologically diverse set of languages, above and beyond what can be captured through learning tied to concrete chunks.

Corpora

We selected from the CHILDES database (MacWhinney, 2000) corpora involving single target children, rather than aggregating data across multiple corpora. Due to limitations on the number of corpora for several of the languages in CHILDES, these were selected according to slightly relaxed criteria: each corpus contained at least 10,000 words, at least 1,000 multiword child utterances, and a child-to-adult utterance ratio of no less than 1:20.

These criteria were met by corpora for 160 additional target children from 28 different languages (Afrikaans: 2, Cantonese: 8, Catalan: 4, Croatian: 3, Danish: 2, Dutch: 12, Estonian: 3, Farsi: 2, French: 15, German: 22, Greek: 1, Hebrew: 6, Hungarian: 4, Indonesian: 8, Irish: 1, Italian: 8, Japanese: 10, Korean: 1, Mandarin: 7, Polish: 11, Portuguese: 2, Romanian: 1, Russian: 2, Sesotho: 3, Spanish: 11, Swedish: 5, Tamil: 1, Welsh: 6). Table 2 lists some basic typological properties of these languages.

To get a rough quantitative measure of morphological complexity for child-directed speech in each language, we calculated word type/token ratios (following the reasoning and methods of Chang et al., 2008). We refer to this as the *Morphological Complexity Score*.

Sentence Production Task

We used the same sentence production task as in Exp. 1.

Results and Discussion

Across all 29 languages and 200+ corpora, the lexical frames version of CBL achieved a mean SPA score of 68.4%, compared to 55.3% for the original CBL model and just 45.9% for the trigram model. Means for each language are shown in Figure 2. By discovering and utilizing lexical frames, the model was able to reproduce the majority of the

child utterances in every language, with mean scores ranging from 55% (Swedish) to 81% (Romanian).

A linear mixed-effects model was fit to logit-transformed SPA scores with language and child as random effects, and a by-language random slope of model. This confirmed that the CBL+LF model (β=0.58, t=26.8, $p < 0.001$) and the trigram model (β=-0.32, t=-8.6, $p < 0.001$) differed significantly from the original CBL, in opposite directions.

Because previous work with the original version of CBL demonstrated that the model's performance decreased as a function of morphological richness (McCauley & Christiansen, 2019), we compared CBL performance to CBL+LF in order to determine whether this effect was reduced by the use of lexical frames. Figure 3 depicts the relationship between the CBL+LF model and Morphological Complexity Score.

Table 2: Typological Properties of the 29 Languages

Language	Family	Genus	Word Order	# Cases
Irish	Indo-European	*Celtic*	VSO	2
Welsh	Indo-European	*Celtic*	VSO	0
English	Indo-European	*Germanic*	SVO	2
German	Indo-European	*Germanic*	N.D.	4
Afrikaans	Indo-European	*Germanic*	N.D.	0
Dutch	Indo-European	*Germanic*	N.D.	0
Danish	Indo-European	*Germanic*	SVO	2
Swedish	Indo-European	*Germanic*	SVO	2
Greek	Indo-European	*Greek*	N.D.	3
Farsi	Indo-European	*Iranian*	SOV	2
Romanian	Indo-European	*Romance*	SVO	2
Portuguese	Indo-European	*Romance*	SVO	0
Catalan	Indo-European	*Romance*	SVO	0
French	Indo-European	*Romance*	SVO	0
Spanish	Indo-European	*Romance*	SVO	0
Italian	Indo-European	*Romance*	SVO	0
Croatian	Indo-European	*Slavic*	SVO	5
Russian	Indo-European	*Slavic*	SVO	7
Polish	Indo-European	*Slavic*	SVO	7
Estonian	Uralic	*Finnic*	SVO	10+
Hungarian	Uralic	*Ugric*	N.D.	10+
Sesotho	Niger-Congo	*Bantoid*	SVO	0
Hebrew	Afro-Asiatic	*Semitic*	SVO	0
Tamil	Dravidian	*S. Dravidian*	SOV	7 or 8
Indonesian	Austronesian	*Malayic*	SVO	0
Cantonese	Sino-Tibetan	*Chinese*	SVO	0
Mandarin	Sino-Tibetan	*Chinese*	SVO	0
Korean	Korean	*Korean*	SOV	7
Japanese	Japanese	*Japanese*	SOV	9

Note: Information from Haspelmath et al. (2005)

Though Morphological Complexity Score was indeed a predictor of CBL+LF performance (β=-2.01, t=-3.5, $p < 0.001$, r=0.23), we found that the presence of lexical frames reduced this effect in comparison to that observed for the original CBL model (β=-2.4, t=-4.1, $p < 0.001$, r=0.27), as confirmed by a significant interaction between model and Morphological Complexity Score (β=0.15, t=3.03, $p < 0.01$) in a linear mixed model which included model as a categorical factor.

A close inspection of the lexical frames discovered by the model when exposed to English revealed that they were both psychologically and developmentally plausible, but we currently lack the cross-linguistic expertise to offer a detailed analysis of lexical frames for the 28 additional languages. Nevertheless, these simulations offer clear evidence that, in principle, the same types of representations and mechanisms can support the discovery of lexical frames across a typologically diverse range of languages. Indeed, for all the 29 languages, lexical frames capture early linguistic productivity above and beyond what can be achieved through concrete words and chunks: CBL+LF lead to a 13 percentage-point improvement over mean CBL performance and a 23 percentage-point improvement over trigram models.

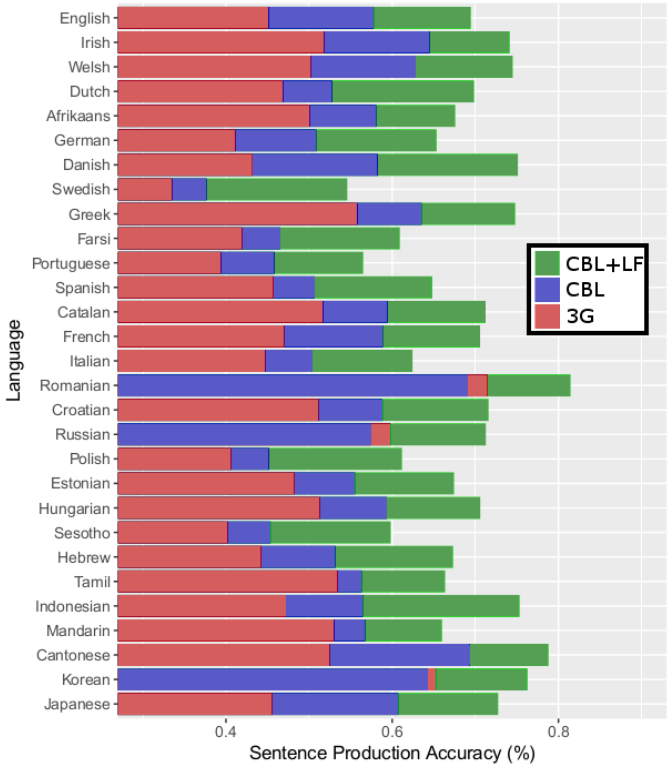


Fig. 2: Sentence Production Accuracy (%) for the model and its baselines across 29 languages. Bars are overlapping.

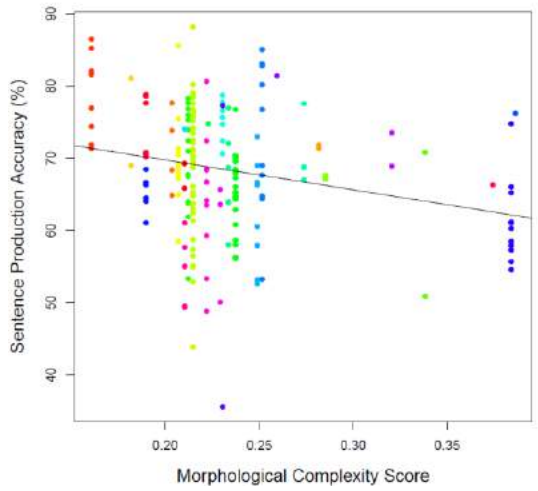


Fig. 3: CBL+LF SPA scores across all children and languages plotted against Morphological Complexity Score. Different colors denote distinct languages. Trendline from simple linear regression.

General Discussion

In this paper, we have demonstrated that a simple, developmentally-motivated model rooted in concrete prediction- and recognition-based processes can discover lexically-based frames that are not only psychologically plausible but also can capture a significant amount of children's early linguistic productivity. In 200+ simulations of individual children across a typologically diverse array of languages, the CBL+LF model was able to capture a significantly higher proportion of child utterances than a version of the model relying solely on concrete words and chunks, offering an even larger improvement over trigram models. Moreover, this was achieved while accommodating the sorts of memory limitations that drive children (and adults) to rely on local information during comprehension and production (e.g., Christiansen & Chater, 2016).

In contrast to previous quantitative studies examining evidence for lexical frames in child speech (e.g., those using the traceback method of Lieven et al., 2003), we 1) capture the actual learning of frames during comprehension, as well as their use in production, and 2) do this for children beyond their second year, with corpora covering child productions during the third and fourth year.

Nonetheless, the CBL+LF approach is not without limitations. Firstly, frames operate on the level of words appearing within chunks; chunks themselves are not yet able to appear in slots. By overcoming this limitation in a principled way, a wider variety of linguistic phenomena could be captured. For instance, non-adjacent dependencies can be learned in the current version of the model: frames like *this __ one* and *those __ ones* capture a number dependency. Extending the model to allow entire chunks in slots will be a necessary subsequent step towards capturing more abstract processing of long-distance dependencies.

A more serious limitation of the present work is that it does not incorporate the learning or use of semantic information. The semantic properties of words and frames are needed to provide constraints on which items that can appear in lexical frames. The learning of such information is crucial for moving towards a framework capable of producing utterances based on meaning representations (and forming meaning representations during comprehension). To this end, ongoing work aims to simulate the learning of lexical semantics, semantic roles, and argument structures through the use of automatically generated, idealized "visual scenes" which are paired with utterances in corpora.

More generally, the promise of item-based computational approaches for tracing a path to more sophisticated forms of linguistic abstraction is great: previous work has shown that the systematic use of pseudographs to align and compare the sentences in a text can give rise to complex context-free grammars (Solan, Horn, Ruppin, & Edelman, 2005). Bayesian induction of item-based grammars from the speech of single target children has also yielded good coverage of those children's increasing productivity (in a manner akin to the traceback method; Bannard, Lieven, & Tomasello, 2009). However, these models are not subject to memory limitations, and involve computations beyond what children are capable of. A motivation for the current approach, therefore, was to take initial steps towards modeling increasingly productive linguistic representations in a way that is psychologically motivated, incremental, and on-line.

Acknowledgments

Thanks to Erin Isbilen for helpful comments and feedback. Thanks to Nick Chater, Padraic Monaghan, Colin Bannard, and Ben Ambridge for helpful discussion.

References

Bannard, C., Lieven, E., & Tomasello, M. (2009). Modeling children's early grammatical knowledge. *Proceedings of the National Academy of Sciences, 106*, 17284–17289.

Bod, R. (2009). From exemplar to grammar: A probabilistic analogy-based model of language learning. *Cognitive Science, 33*, 752-793.

Braine, M. D. S. (1963). The ontogeny of English phrase structure: The first phrase. *Language, 39*, 1-13.

Cameron-Faulkner, T., Lieven, E., & Tomasello, M. (2003). A construction-based analysis of child directed speech. *Cognitive Science, 27*, 843-873.

Chang, F., Lieven, E., & Tomasello, M. (2008). Automatic evaluation of syntactic learners in typologically-different languages. *Cognitive Systems Research, 9*, 198-213.

Chomsky, N. (1965). *Aspects of a theory of syntax.* Cambridge: MIT Press.

Christiansen, M. H., & Chater, N. (2016). The Now-or-Never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences, 39*, e62.

Gentner, D., & Namy, L. L. (2006). Analogical processes in language learning. *Current Directions in Psychological Science*, *15*, 297-301.

Grimm, R., Cassani, G., Gillis, S., & Daelemans, W. (2017). Facilitatory effects of multi-word units in lexical

processing and word learning: A computational investigation. *Frontiers in Psychology, 8:555.*

Haspelmath, M., Dryer, M. S., Gil, D., & Comrie, B. (2005). *The world atlas of linguistic structures.* Oxford, UK: Oxford University Press.

Kol, S., Nir, B., & Wintner, S. (2014). Computational evaluation of the Traceback Method. *Journal of Child Language, 41*, 176-199.

Lieven, E., Salomo, D., & Tomasello, M. (2009). Two-year-old children's production of multiword utterances: A usage-based analysis. *Cognitive Linguistics, 20*, 481-507.

Lieven, E., Behrens, H., Speares, J., & Tomasello, M. (2003). Early syntactic creativity: A usage-based approach. *Journal of Child Language, 30*, 333–370.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing.* Cambridge: MIT.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk, Volume II: The Database.* Mahwah, NJ: Lawrence Erlbaum Associates.

Maslen, Theakston, Lieven, & Tomasello (2004). A dense corpus study of past tense and plural overregularization in English. *Journal of Speech, Language, and Hearing Research, 47*, 1319-1333.

McCauley, S.M. & Christiansen, M.H. (2014). Acquiring formulaic language: A computational model. *Mental Lexicon, 9*, 419-436.

McCauley, S.M. & Christiansen, M.H. (2019). Language learning as language use: A cross-linguistic model of language development. *Psychological Review, 126*, 1-51.

Solan, Z., Horn, D., Ruppin, E., & Edelman, S. (2005). Unsupervised learning of natural languages. *Proceedings of the National Academy of Sciences, 102*, 11629-11634.

Tomasello, M. (2003). *Constructing a language.* Cambridge: Harvard University Press.

Multiword Units Predict Non-inversion Errors in Children's *Wh*-questions: "What Corpus Data Can Tell Us?"

Stewart M. McCauley (stewart-mccauley@uiowa.edu)
Department of Communication Sciences and Disorders, University of Iowa, Iowa City, IA 52242

Colin Bannard (Colin.Bannard@liverpool.ac.uk)
Department of Psychological Sciences, University of Liverpool, Liverpool, UK, L3 5TR

Anna Theakston (anna.theakston@manchester.ac.uk)
Michelle Davis (michelle.davis@manchester.ac.uk)
Division of Human Communication, Development & Hearing, University of Manchester, Manchester, UK, M13 9PL

Thea Cameron-Faulkner (T.Cameron@manchester.ac.uk)
Department of Linguistics and English Language, University of Manchester, Manchester, UK, M13 9PL

Ben Ambridge (Ben.Ambridge@liverpool.ac.uk)
Department of Psychological Sciences, University of Liverpool, Liverpool, UK, L3 5TR

Abstract

Subject-auxiliary inversion in interrogatives has been a topic of great interest in language acquisition research, and has often been held up as evidence for the structure-dependence of grammar. Usage-based and nativist approaches posit different representations and processes underlying children's question formation and therefore predict different causes for these errors. Here, we explore the question of whether input statistics predict children's spontaneous non-inversion errors with *wh*- questions. In contrast to previous studies, we look at properties of the non-inverted, errorful forms of questions. Through a series of corpus analyses, we show that the frequency of uninverted subsequences (e.g., "*she is going"* in "*what she is going to do?*"*) is a good predictor of children's errors, consistent with recent evidence for multiword units in children's comprehension and production. This finding has implications for the types of mental representations and cognitive processes researchers ascribe to children acquiring a first language.

Keywords: language acquisition; interrogatives; corpora; corpus analyses; usage-based approach; chunking

Introduction

Whether the input available to children is sufficient to explain their emerging language abilities is a fundamental question in cognitive science (Chomsky, 1957; Skinner, 1957). Central to the ongoing discussion are tensions between the view of grammar as the result of gradual abstraction over the input (e.g., Lieven, Salomo & Tomasello, 2009; Tomasello, 2003), and approaches in which the acquisition process is guided by innate, language-specific biases (e.g., Pinker, 1999; Fisher, 2002).

In the realm of theoretical linguistics, work in support of the latter approach has focused on specific linguistic phenomena, such as interrogatives. A topic of particular interest is that of subject-auxiliary inversion, which has been held up as evidence for the structure-dependence of grammar (e.g., Crain, 1991; Berwick, Pietroski, Yankama, & Chomsky, 2011), and is often still discussed in the same terms as it was half a century ago (Chomsky, 1968).

In developmental psycholinguistics, a great deal of work has also focused on interrogatives, in part because they represent some of the few sentence types for which English-speaking children reliably make errors involving word order (e.g., Klima & Bellugi, 1966; Stromswold, 1990). Moreover, these sentence types provide a means to evaluate subject-auxiliary inversion as evidence for structure-dependence within a developmental framework. This applies to *wh*-questions especially: as both the *wh*-word and the auxiliary are fronted, it has been argued that they are structurally more complex than *yes/no* questions (e.g., Pozzan & Valian, 2017; Jakubowicz, 2011); and unlike *yes/no* questions, children rarely encounter *wh*-questions in uninverted form as part of the input, yet still make errors of uninversion as in (1).

*(1) What they are doing over there ? **

Thus, *wh*- questions represent an ideal case for mediating between nativist and constructionist approaches, as each posit different representations and processes underlying children's errors and therefore predict different error properties. While the former emphasizes abstract structural considerations, the latter perspective stresses the importance of input frequency in supporting lexically-specific representations.

In line with structure-dependence accounts, a number of researchers have argued for earlier acquisition of argument *wh*-questions than adjunct *wh*-questions, based on their structural properties (e.g., Stromswold 1990, de Villiers 1991). Consistent with this, Pozzan and Valian (2017) report higher non-inversion rates for adjunct than for argument *wh*- questions, a finding they argue to be independent of input frequencies (as might be predicted under usage-based approaches). However, frequency is not rigorously controlled for in the design of the stimulus items themselves, nor is the frequency of substrings beyond the

wh-word/auxiliary combination considered (in the following subsection, we discuss why this may be of importance).

Initial support for usage-based approaches to subject-auxiliary inversion came from a corpus analysis of one child's early *wh*- questions (Rowland & Pine, 2000). The authors found that the frequency of specific *wh*-word + auxiliary combinations reliably predicted non-inversion rates. Ambridge, Rowland, Theakston, and Tomasello (2006) extended this finding with an elicited production study in which *wh*-word + auxiliary combinations predicted non-inversion rates in children aged 3;6 to 4;6. Moreover, *wh*-word alone was not found to predict errors, in contrast to structure-dependence accounts (e.g., Pozzan & Valian, 2017). Rather, the pattern of results was consistent with the notion of lexically-specific representations driving performance with particular question types.

In a further elicitation study, Ambridge and Rowland (2009) investigated a wider range of question types, including negative polarity questions, replicating the finding that *wh*-word + auxiliary frames predicted error rates. Though the relevant frequency dimensions were not controlled for in a rigorous way, Ambridge and Rowland also found initial support for the notion that patterns learned from declarative utterances may also shape errors. It is to this possibility that we turn in the present study.

A Role for Multiword Units in Predicting Non-inversion Errors

A serious limitation of previous work on subject-auxiliary inversion is that only the distributional properties of *correct* forms have been taken into account. This partly stems from the lingering influence of theoretical frameworks in which individual words are viewed as the fundamental units over which language processing take place (e.g., Pinker, 1999). After all, the correctly inverted and errorful, non-inverted forms of a question contain the same set of words; only the word order differs. Thus, if words are the fundamental units of language, we would not expect the distributional properties of an errorful form to play a role in question formation.

Recent years, however, have seen an explosion of psycholinguistic data suggesting that language users are not only sensitive to the properties of compositional multiword sequences, but—in some sense—store and actively utilize such sequences in comprehension and production, as linguistic units in their own right. The frequency of such multiword units—or "chunks"—has been shown to facilitate processing in adult comprehension (e.g., Arnon & Snider, 2010; Bannard, 2006; Reali & Christiansen, 2007) as well as production (e.g., Janssen & Barber, 2012). These findings have received further support from event-related brain potentials (Tremblay & Baayen, 2010) and eye-tracking data (Siyanova-Chanturia, Conklin, & van Hueven, 2011).

Importantly, these findings are mirrored in psycholinguistic work with children (see Theakston & Lieven, 2017 for an overview). Bannard and Matthews (2008) found that, when controlling for substring frequency, overall sequence frequency predicted the speed and accuracy with which 2- and 3-year-olds produced compositional phrases. Arnon and Clark (2011) report evidence that multiword chunk frequency intersects with morphological development: errors of noun plural overregularization were significantly reduced when irregular plurals were produced in the context of more frequent sequences. Moreover, multiword units exhibit the same type of age-of-acquisition (AoA) effects as do individual words, when AoA is determined by either subjective ratings or by corpus-based metrics (Arnon, McCauley, & Christiansen, 2017). Taken together, these findings underscore the possibility that multiword chunks serve as building blocks for language learning.

The importance of these findings to more general theoretical debates is further highlighted by computational modeling work which has shown that abstraction over stored sequences can lead to a considerable amount of linguistic productivity (e.g., Solan, Horn, Ruppin, & Edelman, 2005). Even models lacking abstraction have served to demonstrate that associative learning of chunks from naturalistic input can account for a substantial portion of children's language production (McCauley & Christiansen, 2019).

Therefore, if children are sensitive to the properties of multiword sequences, we might expect such information to play a role in *wh*-question formation. Take, for instance, the following correctly inverted and non-inverted (errorful) forms (2-3):

(2) What <u>is she going</u> to do ?

*(3) What <u>she is going</u> to do ? **

If the uninverted strings "*she is going*" or "*is going*" are highly frequent in the child's input, we might expect—given evidence that multiword chunks play a role in learning and processing—that the child is more likely to produce the errorful form. By the same token, we might expect the frequency of "*is she going*" or "*she going*" to alter this likelihood in the opposite direction. From this perspective, chunks from both the correctly inverted and non-inverted forms might be seen as competing. In other words, multiword sequence frequencies from the correctly inverted and non-inverted forms are both important, insofar as they relate to one another.

The Present Study

If such a relationship exists at all, it is likely to be a complex one, mediated by a host of distributional, pragmatic, and semantic factors. In the present study, we take an initial step towards disentangling these factors by considering, simultaneously, the many distributional factors at play. Not only have the frequencies of individual *wh*-words and auxiliaries been argued to shape errors, but also the frequencies of distinct *wh*-word/auxiliary combinations

themselves (e.g., Rowland & Pine, 2000). Given the perspective we have put forth regarding a role for multiword sequences stretching beyond the *wh*-word and auxiliary, it is necessary to consider the distributional properties of individual words and higher-order *n*-grams for both the correctly inverted and uninverted forms of questions, simultaneously.

In the present study, we evaluate the role of multiword units in early *wh*-question production by using distributional statistics from child-directed speech to predict children's *spontaneous* uninversion errors. Using the entire English portion of the CHILDES database (MacWhinney, 2000), we collect distributional statistics for words and higher-order *n*-grams, which are then used to construct a logistic regression model of children's correctly inverted and errorful (uninverted) questions across the 12 most question-rich corpora. Thus, we are able to test whether, and to what extent, frequencies for individual words and multiword combinations predict spontaneous error rates. Moreover, this allows us to evaluate the role played by multiword sequences from the uninverted forms of questions while controlling for the statistics of the correctly inverted forms, and vice-versa.

In this context, usage-based approaches make predictions that are separable and distinct from those made by theories emphasizing abstract, system-wide principles: if children are forming questions based on structural properties, we would not expect to see a role for uninverted *n*-gram statistics in predicting uninversion errors. Moreover, we would expect structural differences in question type (e.g., argument questions vs. adjunct questions) to be better predictors of correct inversion than frequency (e.g., Pozzan & Valian, 2017). By contrast, usage-based approaches would predict experience with particular *wh*-words, auxiliaries, and even specific subjects/verbs to be robust predictors of error rates, and would quite naturally accommodate findings that *n*-gram sequences from the uninverted forms predict error rates. Under such a view, abstract grammatical constructions tied to questions would emerge gradually as a process of abstracting over stored sequences, and this would be reflected in the probabilities with which children fail to correctly invert certain sentences.

Methods

The corpus analysis consisted of three general phases: extraction of all child-produced *wh*- questions from a set of target corpora, followed by semi-automated identification of uninversion errors; collection of *n*-gram statistics for child-directed speech in English; and mixed-effects logistic regression modeling to determine which *n*-gram statistics predicted uninversion errors in the extracted questions.

Corpus Selection and Preparation

We began by extracting the 12 corpora with the highest number of *wh*- questions from the English language portion of the CHILDES database (MacWhinney, 2000). Each corpus followed a single target child and spanned at least one year of development; the age range and nationality for each target child is shown in Table 1 alongside citation information.

Table 1: Details of CHILDES Corpora Used in Analysis of Uninversion Errors

Target Child	Corpus	Age Range
Abe	Kuczaj, 1977	2;04-5;00
Adam	Brown, 1973	2;03-5;02
Eleanor	Lieven et al., 2009	2;00-3;00
Ethan	Demuth & McCullough, 2009	0;11-2;11
Fraser	Lieven et al., 2009	2;00-3;01
Laura	Braunwald, 1976	1;05-7;00
Lara	Rowland & Fletcher, 2006	1;09-3;03
Lily	Demuth & McCullough, 2009	1;01-4;00
Naima	Demuth & McCullough, 2009	0;11-3;10
Ross	MacWhinney, 1991	1;04-7;08
Sarah	Brown, 1973	2;03-5;01
Thomas	Maslen et al., 2004	2;00-4;11

Each corpus was then prepared for analysis using an automated procedure which removed codes, tags, and punctuation, leaving only speaker identifiers and the original sequence of words. Lines consisting solely of morphological tags (included as standard in CHILDES corpora) were unaffected by this procedure and were retained for later use in extracting uninversion errors.

As part of this procedure, contractions were split into their component words: e.g., "what's he doing" was re-coded as "what is he doing." As corpus annotation differs in terms of how contractions are transcribed (leading to arbitrary noise), this step ensured that modeling work reflected accurate *n*-gram frequencies for *wh*- words and auxiliaries across all questions. As a further step we collapsed the pronouns "she" and "he" into a single form to control for individual differences across children's exposure to gender pronouns.

Wh- Question and Uninversion Error Candidate Extraction and Coding

Child-produced *wh*- questions were automatically extracted from the target corpora by utilizing the standard default morphological tagging included in CHILDES. All extracted questions featured a *wh*- word in the first position, followed immediately by an auxiliary. This yielded approximately 13,000 child-produced *wh*- questions across the 12 corpora.

For the purpose of automatically identifying possible uninversion errors, we extracted, from the full corpora, all child questions which featured a *wh*- word in the initial position which was not immediately followed by an auxiliary. These candidate items were then manually coded for error type by the first author, yielding a total of 300 identified uninversion errors produced across the target children. *wh*- questions featuring an error type other than

uninversion (such as doubling or omission errors) were excluded from our dataset. Importantly, our analyses were restricted to questions produced before the age of five years.

N-gram Data Collection

In order to capture *n*-gram statistics which accurately reflected the nature of child-directed speech in the English language, we gathered *n*-gram frequencies for the entire English (UK and US) portion of the CHILDES database. This allowed us to overcome issues of data sparseness arising from corpus size (Manning & Schütze, 1999).

The aggregated corpus was prepared for data collection following the same procedure described in the above subsection. Frequencies were then collected for unigrams (single words), bigrams (word pairs), and trigrams (word triplets), which were then applied to each of the *wh*-questions extracted for the 12 target child corpora. To this end, *n*-gram statistics were calculated for each question (separate unigram counts for each word, separate bigram counts for each word pair, and so forth). Thus, for the question "what is that," three unigram counts (one for each of three word positions), two bigram counts (one for each of two word pair positions), and one trigram count (for the single word triplet position) were available.

Because our statistical analyses aimed to explore the role of multiword chunk frequency in shaping children's uninversion errors, we sought to directly compare the correctly inverted "target question" for children's uninversion errors to the correctly inverted questions which made up the rest of the dataset. To achieve this, we calculated *n*-gram frequencies for the correctly inverted forms of the uninverted questions identified by the earlier procedure. Uninversion errors were "corrected" by hand in order to achieve this.

By the same token, we also sought to explore the role of multiword sequence frequencies for the relevant uninverted question forms in determining error rates. For this, we retained the original child uninversion errors and employed an automated procedure to produce the errorful, uninverted form corresponding to each correctly inverted question in the corpus. The second and third words could not simply be swapped because a large number of questions featured multiword subject noun phrases, such as "where is my red ball?" Thus, to automatically achieve a realistic uninverted form across such a large number of questions, we first chunked utterances using a shallow parser (Punyakanok & Roth, 2001). Shallow parsers are widely used tools in the field of natural language processing which segment out the non-overlapping, non-embedded phrases in a text. For instance, the shallow parser output for the previous example would be: "[where] [is] [my red ball]." After submitting all correctly inverted questions to the shallow parser, we merely switched the second and third chunks, yielding the relevant, uninverted errorful forms, such as "where my red ball is?"

Thus, we collected unigram, bigram, and trigram statistics for each position across all correctly inverted questions (and, in the case of uninversion errors, the correctly inverted target questions), alongside a separate set of *n*-gram statistics for the uninversion errors (and, in the case of correctly inverted questions, the relevant errorful form).

Analysis

In order to evaluate the predictive relationship between multiword chunk frequency and uninversion errors, we used mixed-effects logistic regression modeling (cf. Agresti, 2002). We carried out a set of model comparisons to determine which *n*-gram frequencies were uniquely predictive of the relationship. This involved selecting predictors at each *n*-gram level separately, starting at the unigram level before moving to the bigram level, followed by the trigram level.

Questions originally produced by the target children in their correctly inverted form were coded as 0, while questions produced in an errorful, uninverted form were coded as 1. *N*-gram frequencies were then used as predictors for this binary variable. All models included a random intercept for child, to reflect the fact that the 12 target children may differ in the extent to which their errors could be predicted by *n*-gram frequencies. By-child random slopes were also included where they improved fit.

Our model comparisons sought to evaluate *n*-gram frequencies of both the correctly inverted question and their corresponding uninverted (errorful) forms as predictors of child uninversion error. The model comparison procedure was designed such that the risk of false positives for higher-order *n*-grams was insignificant, as we conservatively prioritized lower-order *n*-grams in the selection process. Importantly, all predictors were log-transformed and scaled. All model comparisons were carried out using log-likelihood ratio tests.

Starting at the unigram level, we used a leave-one-out procedure to determine which predictors explained variance over and above that explained by any other variable. The full baseline model at this level included random effects of the first 5 unigrams (by child) as well as fixed effects for these 5 unigrams. This was then compared to five subsequent models, each leaving out the fixed effect term for a different unigram (random effects by child were included for every unigram in each model). Removal of only the first two unigrams harmed model fit to a significant extent, according to log-likelihood tests. Thus, these two unigrams were held over for the next level of model comparisons.

The same procedure described for unigrams was then carried out for the first four bigrams, but with random (by child) and fixed effects for the first two unigrams also included in each model (as unigrams are identical across the inverted and uninverted forms, only one set was included in the previous step). Importantly, bigrams from both the correctly inverted and the corresponding errorful forms were included at this second step.

For correctly inverted question forms, removal of the third and fourth bigrams harmed model fit to a statistically

significant extent, according to the log-likelihood tests, while for the uninverted forms, removal of the second, third, and fourth bigrams harmed model fit. Thus, in addition to the first two unigrams from the previous step, the third and fourth bigrams from the correctly-inverted question forms and the second, third, and fourth bigrams from the errorful (uninverted) forms were held over for the final set of model comparisons.

For the first three trigrams, the same procedure was followed once more (with random and fixed effects for the first two unigrams and first two bigrams). Only removal of the second and third trigrams from the uninverted/errorful question forms harmed model fit to a significant extent.

Thus, the final set of predictors included the first two unigrams, the third and fourth bigrams from the correctly inverted forms, the second, third, and fourth bigrams from the uninverted forms, and the second and third trigrams from the uninverted forms.

Results

Our model comparison procedure (as described above) yielded a model with 9 *n*-gram predictors: the first two unigrams, third and fourth bigrams from the correctly inverted question forms; and the second, third, and fourth bigrams as well as the second and third trigrams for the errorful (uninverted) question forms. The log-likelihood, chi-squared value, and p-value for each model comparison is shown in Table 2.

Table 2: Results of Model Comparisons

Left-out Predictor	Log-likelihood	χ^2	*p*-value
Unigram (full/baseline)	-702.13	-	-
Unigram 1	-705.6	6.95	0.00 **
Unigram 2	-707.16	10.07	0.00 **
Unigram 3	-702.27	0.29	0.59
Unigram 4	-702.13	0.00	0.97
Unigram 5	-702.20	0.14	0.71
Bigram (full/baseline)	-626.40	-	-
Bigram 1	-627.28	1.76	0.19
Bigram 2	-627.20	1.59	0.21
Bigram 3	-631.41	10.01	0.00 **
Bigram 4	-632.68	12.55	0.00 ***
Trigram (full/baseline)	-614.62	-	-
Trigram 1	-615.44	1.641	0.2002
Trigram 2	-615.69	2.141	0.1434
Trigram 3	-614.67	0.103	0.748
Uninverted Bigram (full/baseline)	-626.40	-	-
Uninverted Bigram 1	-626.42	0.02	0.88
Uninverted Bigram 2	-634.79	16.77	0.00 ***
Uninverted Bigram 3	-634.87	16.94	0.00 ***
Uninverted Bigram 4	-632.5	12.19	0.00 ***
Uninverted Trigram (full/baseline)	-614.62	-	-
Uninverted Trigram 1	-614.87	0.505	0.4772
Uninverted Trigram 2	-617.55	5.874	0.02 *
Uninverted Trigram 3	-618.41	7.582	0.01 **

To help understand the relationship of these *n*-gram frequencies with child uninversion errors, we constructed non-partial (single-predictor) models for each of the final variables, as reported in Table 3. Each model included a random intercept for target child and a random effect (by child) for the relevant predictor as well as the fixed effect. This procedure was preferred as, in a multi-predictor model, estimates may change sign based on the relative strength of predictor correlations with the dependent variable (cf. Wurm & Fisicaro, 2014).

The first and second unigram frequencies (corresponding to the *wh*- word and auxiliary) were significant predictors with negative estimates, indicating lower likelihood of an uninversion error with more frequent items. Importantly, for higher-order *n*-gram predictors drawn from the errorful, uninverted question forms, the estimate was positive. This means that the higher the *n*-gram frequency was for the uninverted form of a question, the more likely it was for that question to have been produced in its uninverted form.

Table 3: Results of Non-partial Models

N-gram	β	Std. Error	*Z*	*p*-value
Uni 1	-0.792	0.27	-2.91	0.004 **
Uni 2	-0.634	0.11	-5.34	0.000 ***
Bi 3	0.031	0.11	0.25	0.795
Bi 4	0.239	0.14	1.64	0.100
Bi 2 (*uninv.*)	0.328	0.11	2.89	0.004
Bi 3 (*uninv.*)	0.563	0.13	4.24	0.000 ***
Bi 4 (*uninv.*)	0.207	0.16	1.26	0.207
Tri 2 (*uninv.*)	0.462	0.10	4.44	0.000 ***
Tri 3 (*uninv.*)	0.454	0.11	4.03	0.000 ***

General Discussion

The corpus analyses presented here represent, to our knowledge, the most rigorous attempt to control for input frequency in analyzing non-inversion errors to date. We find that, when *n*-gram frequencies from both the correctly-inverted, "target" form of a question, and the non-inverted, "errorful" form of a question are considered in parallel, frequency is a robust predictor of when non-inversion errors will occur. Moreover, the frequencies of higher-order *n*-

grams from the non-inverted form are shown to be more robust predictors than frequencies from the correctly inverted form.

This finding appears to stem from children's use of multiword units in production (e.g., Bannard & Matthews, 2008). Consider the effect of the (non-inverted) second trigram in the context of the following non-inversion error: "*where we can go today?**" The more heavily *we can go* holds together as a unit in the child's language experience, the less likely the child will be to break up the sequence by fronting the auxiliary *can* (e.g., by relying on a lexical frame for *what can*). Similar reasoning can be applied to the effect of the non-inverted third bigram (*can go,* in this example). Errors caused by the intrusion of overlearned sequences occur in all kinds of human action (Bannard et al., in press).

Thus, our findings weigh in favor of previous proposals that children rely on lexically-based representations in question formation (e.g., Rowland & Pine, 2000) and support the proposal that material learned from declarative utterances can drive systematic errors (Ambridge & Rowland, 2009). Our findings are inconsistent, however, with structure-dependent accounts of children's *wh*-questions (e.g., de Villiers, 1991).

The present study, therefore, offers an interesting additional line of evidence supporting usage-based approaches, especially accounts of language development which stress the importance of multiword units (e.g., Theakston & Lieven, 2017; McCauley & Christiansen, 2019) including exemplar-based approaches (Ambridge, 2018).

References

Ambridge, B. (2018, July 25). Against stored abstractions: A radical exemplar model of language acquisition. https://doi.org/10.31234/osf.io/gy3ah

Ambridge, B., Rowland, C.F., Theakston, A.L., & Tomasello, M. (2006). Comparing different accounts of inversion errors in children's non-subject *wh*-questions: "What experimental data can tell us?." *Journal of Child Language, 33*, 519-557.

Ambridge, B., & Rowland, C.F. (2009). Predicting children's errors with negative questions: Testing a schema-combination account. *Cognitive Linguistics, 20*, 225-266.

Agresti, A. (2002). *Categorical data analysis.* Hoboken, NJ: Wiley.

Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multiword phrases. *Journal of Memory and Language, 62*, 67–82.

Arnon, I., McCauley, S.M. & Christiansen, M.H. (2017). Digging up the building blocks of language: Age-of-acquisition effects for multiword phrases. *Journal of Memory and Language, 92*, 265-280.

Bannard, C. (2006). *Acquiring phrasal lexicons from corpora*. Doctoral dissertation: University of Edinburgh.

Bannard, C., Leriche, M., Bandmann, O., Brown, C., Ferracane, E., Sánchez-Ferro, A., Obeso, J., Redgrave, P., & Stafford, T. (in press). Reduced habit-driven errors in Parkinson's Disease. *Nature Scientific Reports.*

Bannard, C., & Matthews, D. (2008). Stored word sequences in language learning. *Psychological Science, 19*, 241.

Berwick, R. C., Pietroski, P., Yankama, B., & Chomsky, N. (2011). Poverty of the stimulus revisited. *Cognitive Science, 35,* 1207-1242.

Chomsky, N. (1957). *Syntactic structure.* Hague: Mouton.

Chomsky, N. (1968). *Language and mind.* New York: Harcourt, Brace, Jovanovitch.

Crain, S. (1991). Language acquisition in the absence of experience. *Behavioral and Brain Sciences, 14,* 597-612.

de Villiers, J. (1991). Why question? In T. L. Maxfield & B. Plunkett (eds.), *Papers in the acquisition of wh: Proceedings of the UMASS Roundtable*, May 1990. Amherst, MA: University of Massachusetts Occasional Papers.

Fisher, C. (2002). The role of abstract syntactic knowledge in language acquisition: A reply to Tomasello (2000). *Cognition, 82*, 259-278.

Jakubowicz, C. (2011). Measuring derivational complexity: New evidence from typically developing and SLI learners of L1 French. *Lingua, 121*, 339-351.

Klima, E.S., & Bellugi, U. (1966). Syntactic regularities in children's speech. In J. Lyons & R. Wales (Eds.), *Psycholinguistic papers* (pp. 183-208). Edinburgh: Edinburgh University Press.

Lieven, E., Salomo, D., & Tomasello, M. (2009). Two-year-old children's production of multiword utterances: A usage-based analysis. *Cognitive Linguistics, 20,* 481-507.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge: MIT press.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk, Volume II: The Database.* Mahwah, NJ: Lawrence Erlbaum Associates.

McCauley, S.M. & Christiansen, M.H. (2019). Language learning as language use: A cross-linguistic model of child language development. *Psychological Review, 126,* 1-51.

Pinker, S. (1999). *Words and rules: The ingredients of language.* New York: Basic Books.

Pozzan, L. & Valian, V. (2017). Asking questions in child English: Evidence for early abstract representations. *Language Acquisition, 24,* 209-233.

Punyakanok, V., & Roth, D. (2001). The use of classifiers in sequential inference. In *Proceedings of NIPS 2001* (pp. 995-1001).

Reali, F. & Christiansen, M.H. (2007). Word-chunk frequencies affect the processing of pronominal object-relative clauses. *Quarterly Journal of Experimental Psychology, 60*, 161-170.

Rowland, C.F. & Pine, J.M. (2000). Subject–auxiliary inversion errors and *wh*-question acquisition: 'What children do know?*' Journal of Child Language, 27*, 157–81.

Siyanova-Chanturia, A., Conklin, K., & Schmitt, N. (2011). Adding more fuel to the fire: An eye-tracking study of idiom processing by native and non-native speakers. *Second Language Research, 27*, 251-272.

Skinner, B. F. (1957). *Verbal behavior.* New York: Appleton-Century-Crofts.

Solan, Z., Horn, D., Ruppin, E., & Edelman, S. (2005). Unsupervised learning of natural languages. *Proceedings of the National Academy of Sciences, 102*, 11629-11634.

Stromswold, K. J. (1990*). Learnability and the acquisition of auxiliaries*. Doctoral dissertation: Massachusetts Institute of Technology.

Tomasello, M. (2009). *Constructing a language.* Cambridge: Harvard University Press.

Theakston, A. & Lieven, E. (2017), Multiunit sequences in first language acquisition. *Topics in Cognitive Science, 9*, 588-603.

Tremblay, A., & Baayen, R. H. (2010). Holistic processing of regular four-word sequences. In D. Wood (Ed.) *Perspectives on formulaic language* (pp. 151-173). London: Continuum International Publishing Group.

Wurm, L. H., & Fisicaro, S. A. (2014). What residualizing predictors in regression analyses does (and what it does not do). *Journal of Memory and Language, 72*, 37-48.

Applying Deep Language Understanding to Open Text: Lessons Learned

Marjorie McShane (margemc34@gmail.com)
Stephen Beale (stephenbeale42@gmail.com)
Irene Nirenburg (irene_bn@yahoo.com)

Cognitive Science Department, Rensselaer Polytechnic Institute
Troy, NY 12180 USA

Abstract

Human-level natural language understanding (NLU) of open text is far beyond the current state of the art. In practice, if deep NLU is attempted at all, it is within narrow domains. We report a program of R&D on cognitively modeled NLU that works toward depth and breadth of processing simultaneously. The current contribution describes lessons learned – scientifically and methodologically – from an exercise in applying deep NLU to open-domain texts. An overarching lesson was that although learning to compute sentence-level semantics seems like a natural step toward computing full, context-sensitive, semantic and pragmatic meaning, corpus evidence underscores just how infrequently semantics can be cleanly separated from pragmatics. We conclude that a more comprehensive methodology for automatic example selection and result validation is needed as prerequisite for success in developing NLU applications operating on open text.

Keywords: natural language understanding; cognitive modeling; language-endowed intelligent agents

Introduction

Operationalizing human-level natural language understanding (NLU) in computer systems has been a goal of AI since its inception. People want intelligent agents to understand not only what they say but what they mean, taking into account the linguistic and real-world context, shared background knowledge, the interlocutors' mutually understood plans and goals, and even their mental, physical, and emotional states. All of these considerations explain why human-level NLU is an AI-complete problem.

It is difficult to carve out a program of R&D for AI-complete problems. With respect to natural language, the field has responded in five broadly-defined ways.[1] **(1)** *Avoid meaning.* For the past 25 years, mainstream NLP has chosen to pursue so-called *knowledge-lean* methods, i.e., the statistical processing of big data with little to no computation of meaning. This has proven useful for certain applications but is not moving toward explainable, human-level NLU in service of intelligent agents. **(2)** *Address select aspects of meaning.* Computing individual aspects of meaning has improved the quality of some primarily knowledge-lean systems. Topics addressed include, e.g., case-role identification, speech act detection, textual coreference resolution, and the semantic clustering of word strings using distributional semantics (Jurafsky and Martin, 2009). **(3)** *Pursue deep NLU in a (very) narrow domain.* This provides systems with the kinds of knowledge and reasoning capabilities that people leverage when interpreting language (e.g., Allen et al. 2007; Lindes and Laird, 2016). **(4)** *Build theories without systems.* Such work anticipates that prerequisites – such as NLU – will be eventually be fulfilled externally, and is typical in the fields like computational formal semantics and machine reasoning. **(5)** *Build extensive theories but implement and evaluate just a subset*. This appears to be the choice of the dialog specialist David Traum (compare Traum 1994 for scientific work with Nouri et al. 2011 for application-oriented work).

The program of R&D described here – developing Language-Endowed Intelligent Agents (LEIAs) within the OntoAgent cognitive architecture – offers a sixth approach to attacking the AI-complete problem of human-level NLU (McShane, Nirenburg and English, 2018). It pursues **depth of analysis and breadth of coverage concurrently**, but with appropriately flexible expectations about the coverage, quality, and confidence of analyses depending on the correlation of text inputs with knowledge bases. It focuses on the *actionability* of language interpretations, as judged by the agent systems that use them.

Of the many theoretical and methodological issues at the core of this program of work (McShane, Nirenburg and Beale, 2016), the following are particularly relevant for this discussion.

1. LEIAs are modeled after humans. Like humans, they do not need to understand everything their interlocutors say and mean; instead, they need to achieve actionable interpretations, defined as interpretations that are sufficient to support reasoning about action.
2. The same knowledge that allows LEIAs to function intelligently in their domain of expertise supports language-oriented reasoning in that domain. Full NLU is not possible without such knowledge.

[1] This is a thumbnail sketch of a long history and extensive literature. See Nirenburg and McShane (2016) for a more in-depth treatment.

3. For both theoretical and methodological reasons, NLU is best implemented as a series of layers of ever-deeper analysis, resulting in ontologically-grounded text meaning representations (TMRs) that are well-suited to agent reasoning.
4. Most narrow-domain approaches seek to avoid disambiguation, one of the hardest problems of NLU; however, such approaches will not attain a human level of understanding until this problem is solved and agents function with a realistic-sized lexicon.
5. A very large number of linguistic phenomena (to name just a few: nominal compounding; all aspects of reference resolution, including fragments and ellipsis; non-literal language; indirect modification; indirect speech acts; implicatures) must be handled by LEIAs no matter their domain of specialization (McShane and Nirenburg, *forthcoming*). The computational microtheories accounting for these phenomena are best investigated using open text.

The original hypothesis underlying the work described here was that we could *quickly* validate many of the implemented microtheories of NLU for LEIAs using an open corpus. Why an open corpus? As discussed in more detail later, this method a) provides useful fodder for improving microtheories, b) makes the work "real" in the eyes of the mainstream NLP community, and c) shows how the analysis capabilities can be usefully applied to open texts.

In formulating the reported exercise, we assumed that a corpus would contain a sufficient number of sentences that could be automatically interpreted using general linguistic and world knowledge, without the need for the finer-grain knowledge resources supporting agent-reasoning capabilities that are available only in narrower domains. Such sentences would be similar in nature, but methodologically preferable, to the invented examples we use to test out individual microtheories.

We further assumed that a simple, automatic method of extracting examples would serve the purpose. However, this experience has shown that, in order to sufficiently evaluate all of the microtheories contributing to the system, we need a more sophisticated example extraction methodology operating over a larger corpus, as well as more human effort devoted to reviewing results. However, rather than change the original hypothesis by allocating more time and effort to data collection, we heeded the lessons learned from the Reproducibility Project (Open Science Collaboration, 2015) and its analytical wake: It is not appropriate to tweak hypotheses or results until they achieve the envisioned threshold. Research habitually involves things not going to plan, and the associated lessons learned are central to progress in the field. This paper focuses on lessons learned. But we must begin with the briefest introduction to the NLU environment at hand.

The OntoAgent Cognitive Architecture

The OntoAgent cognitive architecture underlying LEIAs includes the modules of perception, reasoning and action. Language is one of the perception modes of a LEIA. Language inputs are analyzed into disambiguated, ontologically-grounded meaning representations. For example, the bare-bones basic TMR for *I knocked on the door* (stripped of metadata and calls to the procedural semantic routines for coreference resolution) is as follows:

```
(HIT-1 (AGENT HUMAN-1)
       (THEME DOOR-1)
       (INSTRUMENT HAND (OPENNESS 0))
       (TIME < find-anchor-time)) ; indicates past tense
```

The fact that the instrument is a closed hand is provided by the lexical description of the selected sense of *knock* in the system's lexicon, which also expects the object of the preposition to refer to, among other possibilities, a door.

Although we cannot adequately familiarize readers with the theory of Ontological Semantics, the agent applications that this approach to NLU has supported, the knowledge bases employed, or how the analysis process works (see, e.g., Nirenburg and Raskin, 2004; McShane, Nirenburg, and English, 2018; Nirenburg, McShane and Beale, 2008), the following facts will serve as orientation. The lexicon contains ~30,000 word senses, which are comprised of linked syntactic and semantic representations and, whenever necessary, calls to procedural semantic routines (for example, to resolve coreferences). Argument-taking words, multiword expressions, and polysemy are richly represented. The semantic descriptions are written in an unambiguous ontological metalanguage. The ontology contains ~9,000 concepts (~145,000 RDF triples), mostly from the general domain. Concepts are described using attributes and relations. Scripts detailing complex events are available in select domains.

The lexicon and ontology were mostly compiled through a modest, short-term effort around 25 years ago in service of interligua-based machine translation and have been only minimally modified since. They were not modified at all for the reported exercise. The key benefit of our lexicon is that it is far from toy and, therefore, allows us to develop and test the essential capability of lexical disambiguation. All parts of speech include polysemous entries, and light verbs such as *have, make,* and *do* have dozens of senses, many of which involve multi-word expressions or constructions. The ontology, for its part, provides selectional constraints on case-roles that support disambiguation, as well as a substrate for various types of language-oriented reasoning, such as topic/domain detection based on ontological distance.

Although these resources have served our research goals quite well, their insufficiencies are relevant to the current report. We estimate that the lexicon would need to be around ten times larger to provide baseline coverage of open text, with the necessary acquisition including a large percentage of multi-word expressions and constructions. An

acquisition effort of this size is, we estimate, no more labor-intensive than some of the well-known corpus annotation efforts in service of supervised machine learning.

NLU by LEIAs is reasoning-intensive. The overall process is modeled as two types of incrementality: *horizontal incrementality* involves analyzing elements of input as they become available to the agent (essentially, word by word); *vertical incrementality* involves applying, on an as-needed basis, increasingly sophisticated methods of analysis to the given state of input, be it a fragment, a complete utterance, or a multi-sentence text. Agents dynamically decide how deeply to process chunks of input as they are perceived.

There are six stages of vertical incrementality, described in greater detail in (McShane and Nirenburg, *forthcoming*): **1.** Preprocessing and syntactic parsing, for which we use the CoreNLP toolset (Manning et al. 2014). **2.** Integrating these results into our environment, which includes recovering from unexpected syntax as well as the initial stage of learning new words. **3.** Basic semantic analysis, which uses lexical and ontological knowledge for disambiguation and semantic dependency analysis. This includes such advanced capabilities as the detection and resolution of many types of ellipsis and learning the semantics of unknown words. **4.** Aspects of reference resolution that do not require full contextual grounding. These include resolving textual coreference, identifying which referring expressions do not require a coreferent and why, establishing reference relations that are not *co*reference (e.g., bridging constructions), and reconsidering upstream lexical disambiguation decisions based on coreference relations. **5.** Extended semantic analysis, which treats select instances of residual ambiguities and incongruities using additional general-purpose rule sets. These include, e.g., ontological patterns for interpreting nominal compounds, rules for interpreting metonymies, and dialog-analysis strategies for integrating the meaning of fragmentary utterances into the discourse. **6.** Situated NLU, which applies all of an agent's domain-specific and situational knowledge and reasoning to resolve residual ambiguities and incongruities, and anchors newly learned knowledge to agent memory.

If it sounds like this system is claiming to do *everything,* that is, in a certain sense, correct. The overall challenges of NLU must be addressed in an integrated system, within an architecture and theory that reserves a place for each component microtheory. The microtheories must be crafted as components of such an overall analysis system. This approach avoids the two most serious problems of strictly modular or limited-scope research: the assumption that prerequisites for one's own work will be provided externally; and the avoidance of all cross-modular phenomena.

Stages 1-5 involve what some call *semantic* meaning, as contrasted with *pragmatic* (discourse, situational) meaning. This level of meaning should be understandable at the sentence level, outside of context – even if some expressions (e.g., pronouns) remain underspecified. Following this expectation, individual sentences outside of their context were the focus of the reported exercise. Given that the ~30,000-sense lexicon contains over 1,600 verb senses, and that the system can process proper nouns and learn new words, we projected that there would be plenty of appropriate sentences to seed our exercise. As concerns Stage 6 of processing, it cannot be validated using individual sentences outside context; we are working on that separately, within a robotic application (Nirenburg et al., 2018).

Methodology

Our initial goal was to focus on *validating* our system rather than formally *evaluating* it in the way that has become standard in the field of natural language processing (NLP). That methodology is of no use for systems that seek human-level understanding of language. It is not, therefore, surprising that mainstream NLP has all but officially placed our area of R&D beyond the boundaries of the discipline. For example, in their chapter on "Evaluation of NLP Systems" in *The Handbook of Computational Linguistics and Natural Language Processing* (Clark, Fox and Lappin, 2010), Resnik and Lin do not even address the evaluation of cognitively-oriented systems that integrate scientific and technological goals. They write: "such scientific criteria [involving, e.g., the cognitive modeling of human language processing] have fallen out of mainstream computational linguistics almost entirely in recent years in favor of a focus on practical applications, and we will not consider them further here." (p. 271) So, we need an alternative validation/evaluation methodology.

There is no truly fast, easy, and complete way to validate (no less evaluate) a large and complex knowledge-based system, nor can the full set of options be fleshed out in this short space. As a starting point, consider just a few of the options. **(1)** *Invent test inputs guided by the knowledge bases and system capabilities*. This gives credit for what *does* work but rarely uncovers unexpected phenomena and is viewed skeptically by the field at large. **(2)** *Use inputs limited to a narrowly-defined domain.* This, too, usually involves manual example creation since 'narrowly-defined' must be enforced; moreover, it fails to give the system or component microtheories credit for their applicability across domains. **(3)** *Use randomly selected inputs from the open domain*. Although this is a cornerstone of statistical NLP, it is inapplicable to deep NLU given that the environment is known to have limited lexical coverage. **(4)** *Focus on full sentences from open text that the system analyzes perfectly.* This approach tasks the system with extracting from open text, and processing, only those sentences it hypothesizes it can analyze correctly. During validation, people inspect only the highest-quality results – i.e., those for which exactly one TMR achieves the highest score, and that score reflects high confidence. This is the approach we used for the current exercise. Its insufficiencies underlie many of the lessons learned from this exercise, as discussed in the next section. **(5)** *Focus on subsentential chunks of text from the open domain that the system analyzes perfectly.* Such chunks can represent propositions, individual phenomena (e.g., nominal compounds, instances of verb phrase ellipsis),

or sentences for which all aspects but one – e.g., an unknown adverb – are correctly understood. We have used this method in past formal evaluations (Nirenburg et al., 2018) and have found it useful for vetting individual microtheories. The problem is that it is time-consuming to formulate a vetting regimen for even a single microtheory, let alone the dozens that the system currently comprises, or the interactions among them. Additionally, any of the above methods can also involve inspecting outputs that are partially correct, residually ambiguous, etc.

Results

As should be clear by now, this vetting exercise was primarily intended to guide our continued R&D effort. It did – but more through lessons learned than from compiling examples that work. That being said, we do want to present some examples to show that our NLU system *can,* in fact, work on open text.

To further specify the set-up: The system extracted examples from two randomly selected excerpts of the COCA corpus (Davies 2008), one literary and the other journalistic. It extracted sentences that included a maximum of one unknown word, with "known" implying that the lexicon contained an entry with the necessary part of speech. No other extraction filters were applied. The system processed the sentences into TMRs using Stages 1-5 of our NLU system. We manually reviewed only those results that seemed promising. For example, we did not inspect the TMRs for sentences that were incomprehensible outside of context, or that required knowledge or reasoning beyond that available in Stages 1-5 of NLU.

We spent just a few person-weeks on the exercise, much of which involved code debugging (after all, the exercise was primarily in service of R&D). However, the examples we cite as "correct" were correct *before* any system modifications. No amendments to the knowledge bases were made. It did not take long to determine that we had learned what we could from this exercise, and we, therefore, did not prolong it to collect more working examples.

Unless otherwise noted, all examples presented in this section were analyzed perfectly. Any incorrect portions are indicated by strikethroughs or explanatory text. Every input required disambiguation decisions, in some cases, from a large choice space: e.g., *He looked for the creek* disambiguates between 16 senses of *look*, and *I went into the bathroom* disambiguates between 54 senses of *go*. The examples below are grouped by the specific phenomena they illustrate.

Complex semantic descriptions. For example, the TMR for *I knocked on the door* includes a hand as the instrument, and the TMR for *I pointed at the blood* includes a finger as the instrument.

Disambiguation of highly polysemous particles and prepositions: *She rebelled against him; He stared at the ceiling; She jokes with him; She switched on the light; He passed through the entrance; I called for a blanket; I thought about Amalia; He talked about Leona***.**

Modification and sets: *An old white couple lived in a trailer.*

Multiword expressions: *He took me by surprise*,

Verbal disambiguation using a specificity preference. For example, in *I do not know Dave,* three senses of *know* (glossed as *be acquainted with, be aware of,* and *be able to identify*) formally match the case-role constraints. The sense *be acquainted with* fulfills the tightest case-role constraints, so it wins. This example also shows the correct processing of the modality indicated by negation.

Dynamic sense bunching. This allows the system to underspecify an interpretation rather than end up with competing analyses. E.g., ~~*No, and*~~ *I didn't ask him* does not permit disambiguation between three senses of *ask* – those encoded using the ontological concepts REQUEST-INFO, REQUEST-ACTION and PROPOSE – so the system bunches these into their closest common ontological ancestor, ROGATIVE-ACT, whose case-roles are correctly understood as AGENT and THEME.

Lateral selectional constraints for disambiguation. E.g., in *I heard the hands on the clock* ~~*move*~~*, clock* was correctly used to disambiguate *hands* (but since the CoreNLP misidentified "clock move" as a nominal compound, that aspect of the analysis was wrong). Similarly, in *The arm jerked, eyelids* ~~*rose*~~*,* the meaning of *eyelids* was correctly used to disambiguate *arm* between body part and furniture part (but *rise* as applied to eyelids was misanalyzed).

New word learning. An example of new noun learning is 'uncle' in *The uncle said something to him*, which is understood as referring to a HUMAN since the AGENT slot of ASSERTIVE-ACT must be filled by a HUMAN. The results of learning are understood as provisional, and values of properties of the newly learned concept are expected to be added opportunistically as a side effect of continued processing of input – or, alternatively, by a knowledge acquirer. An example of new property learning is *inconsiderate* in *Burying Leora* ~~*in Pittsburgh*~~ *is inconsiderate*. The system represents the meaning as a generic PROPERTY whose DOMAIN is filled by the event BURY (from *burying*). *In Pittsburgh* was correctly analyzed but incorrectly attached to Leora rather than burying, following a parsing error by CoreNLP. (Reambiguating PP attachments from the CoreNLP parse, so that semantics can weigh in, is on agenda.)

The above presents just a small sampling of linguistic phenomena that the system covers, along with examples of successful analyses. It shows that vision behind the current exercise was not ultimately ill-conceived, and illustrates that the corpus was, in fact, open-domain. But, as we said earlier, we keep this aspect of the report brief in order to focus on the main point: lessons learned.

Lessons Learned

Most of the *types* of outcomes of this exercise were predicable beforehand, but in some cases their *frequency* was rather surprising, thus representing a lesson learned.

1. *It is not possible to automatically detect that a needed multiword expression (idiom, construction, etc.) is missing*

in the lexicon. Multiword expressions are central to a human's knowledge of language and, accordingly, to modeling NLU for LEIAs. When a multiword expression is missing from the lexicon, the system analyzes the components compositionally, which necessarily results in an error. All of the following examples were misanalysed because interpreting the meaning of the underlined portion required a multiword lexical sense that had not yet been acquired. *She is long gone from the club. I got a good look at that shot; The Knicks can live with that. But once Miller gets on a roll, he can make shots from almost 30 feet. I can't say enough about him. This better be good. You miss the point. I should have known better.* The lesson learned involves the frequency with which the system will be overly confident in its analysis, not having recognized that an input component is not semantically compositional.

2. *The methodology of focusing on completely correct TMRs was suboptimal.* Often, the meaning representation of a portion of the input nicely demonstrates a particular functionality, even though some aspect of the overall sentence interpretation is incorrect. Many such mistakes reflect the use of microtheories that are currently underdeveloped, such as those for relative temporal and spatial relations (*in recent weeks, 25 feet right of the hole*, and *for the second time this year*). When, midstream, we decided to revisit partially correct TMRs, we found many interesting correct subanalyses, suggesting that vetting Method #5 described above might be superior to the method we used.

3. *The methodology of focusing exclusively on sentences that resulted in a single TMR was suboptimal*. Outside of context, residual ambiguity is quite common. When we decided to revisit analyses that resulted in two output TMRs – because the analyzer did not have a reason to prefer one over the other – we found examples in which this outcome was actually the correct one. For example, the system correctly detected the ambiguity, and generated multiple correct candidates, for *He stared at the fish,* which could refer to a live fish (FISH) or its meat (FISH-MEAT); and *He glanced at the walls* could refer to parts of a room (WALL) or parts of a person undergoing surgery (WALL-OF-ORGAN).

4. *It can be difficult, even for humans, to describe many intended meanings*. Consider the following sentences: *And he came back from the dead. Training was a way of killing myself without dying. The supporting actor has become the leading man. This is about substance. The roots that are set here grow deep.* Such examples allow for multiple interpretations, at many levels of vagueness and specificity, depending on the specific speech situation. The existence of utterances of this type are among the reasons we believe that, in building agent-oriented NLU capabilities, actionability – not exhaustive understanding – is key. But for this exercise, decision-making about actionability was outside of purview.

5. *The intended meaning can rely more centrally on discourse/pragmatic interpretation than semantic analysis*. In some cases, e.g., for personal pronouns, there is a clear progression from semantic to pragmatic meaning. However, in other cases, semantic meaning is either vague, not directly connected with pragmatic meaning, or even relatively unimportant. Space is too short to flesh out these complex eventualities, but consider the example *It takes two to tango*, which occurred in our corpus. If we were to write a lexical sense for this phrase, how would we describe its meaning? Its propositional meaning – something like "a communication cannot exist without multiple people being agentive" – is much less important than its discourse function. That is, the speaker is saying that the given situation is an example of a generalization about human relations, but the context-specific pragmatic nuances can range from being a barb during a spat (*It's your fault, too, that we're arguing!*) to being advice to a friend (*If you back off, maybe the other person will too*). It seems incorrect to lexically record, and then give a system credit for computing, semantic meanings when it is the pragmatic force that is predictably more important.

6. *Non-literal language is even more prevalent than we had expected – and we had expected a lot.* In fact, we have methods for detecting and recovering from some types of non-literal language, but not the onslaught we encountered in this exercise. For example, *Everyone was saying we won ugly last week* and *He not only hit the ball, he hammered* were imperfectly analyzed because the non-literal meanings were not correctly recovered.

7. *We need to operationalize reasoning about language via affordances*. Just as human vision is well-understood to be largely driven by expectations, so, too, is language understanding. Affordances – i.e., the knowledge of what objects can do and how they can be used – can support reasoning about language inputs, particularly if they involve difficult phenomena, such as non-literal language, unknown words, and indirect modifications. For example, we previously noted that *eyelids rose* resulted in a misinterpretation of 'rise'. It is unlikely that people encode a word sense of 'rise' that covers eyelids; however, we know that eyelids are capable of precious few actions. So a fuzzy matching between words and concepts for moving up and down is sufficient for a person to understand this. A microtheory of applying affordances to reasoning about NLU is on our team's agenda.

8. *It is unclear what credit to give semantics without implicatures*. On the one hand, semantic analysis is hard enough without requiring that NLU systems account for all a speaker's implicatures before claiming any success. On the other hand, in some cases semantics and implicatures cannot be neatly separated. Consider the example, *She's also a woman*. Reading this in isolation, we understand that the context must have been about her in some other social role – as a mother, a co-worker, etc. – and that this utterance focuses attention on her female/sexual side. It is similarly unclear what, if anything, would count as a sufficient semantic (pre-implicature) analysis of the following: *How quickly the city claimed the young. They sat by bloodline. I think he is coming into good years. Fathers were for that.*

9. *Not invoking domain-oriented expectations is more limiting that we had anticipated.* For example, unless you

realize you are in a sports context – and know sports-related lexical and ontological knowledge – the following are not fully interpretable: *The Rangers and the Athletics have yet to make it. He hit his shot to four feet at the 16th. We stole this one. I wanted the shot.*

10. *Although our system is knowledge-based and all processing apart from what is contributed by CoreNLP is fully inspectable, the computational complexity of deep NLU can make it difficult to fully predict, explain, and trouble-shoot results*. Consider the simple example *I almost never talk about it,* whose words have, respectively, 3/5/2/2/2/1 senses. The number of candidate TMRs generated is 50, with their final scores ranging from -22 to 22.9. There was only 1 highest-scoring TMR and it was correct. The CoreNLP parse happened to be correct, but since this is not always the case, our analyzer compensates by considering other syntactic analysis possibilities as well. As a result, the process of mapping syntactic dependencies in inputs to the variables in the syntactic descriptions in lexicon entries can lead to multiple sets of variable assignments for each available sense. At the semantic level, the system needs to select the best sense and set of variable assignments for each word by examining the interactions between the semantic constraints among all the words that interact with it. In the worst case, that can become a computational clique, which has exponential time requirements (we employ various techniques for reducing or sometimes eliminating this computational drain). Scoring functions are also complex – their composition is a research issue in itself. In short, even though we can configure a glassbox evaluation, the analysis process can, in certain cases, still defy complete explanation.

Conclusions

We believe that our original goal – to vet our system's domain-independent microtheories using open text – is achievable. The reason why the focus of this exercise shifted from "vetting" to "investigating lessons learned" is because the methodology for extracting examples and automatically evaluating the quality of output TMRs turned out to be insufficiently developed. The lessons learned will inform the creation of a more sophisticated methodology for future experiments. To give just a few examples of planned enhancements: (a) Including the preceding context for each extracted example to allow for coreference and lateral-constraint heuristics to be leveraged; (b) Automatically excluding excessively short inputs, direct speech, texts from jargon-intensive domains like sports, and inputs containing pronouns whose resolution strongly affects disambiguation decisions (e.g., *it, that* and *they* are more problematic than *he* or *she*); (c) Using an example-extraction methodology that identifies the highest-confidence examples of each word sense, microtheory, etc., from a much larger corpus than was used for this exercise; and (d) Including within purview high-confidence subsentential results.

Apart from lessons learned, this experiment has resulted in promising outcomes. The fact that the system correctly analyzed some inputs from the open domain – even given the shortcomings of the reported methodology and all of the challenges natural language predictably presents – suggests that deep NLU can have near- and mid-term utility, given an appropriate task formulation and improved methods of automatically judging the system's confidence in its analyses.

Lifelong learning has long been understood as a necessary foundation of AI. Even the current capabilities of the reported NLU system can support the learning of lexical units and ontological concepts, with the coverage expected to rise dramatically even with relatively (by industry standards) modest knowledge acquisition efforts.

Our system addresses the open-world problem directly and takes responsibility for all upstream processing errors (currently, from CoreNLP). In some cases, it can successfully learn new meanings and recover from upstream errors, whereas in others it cannot. However, we believe that failures under real-world circumstances are far preferable to the non-real-world experimental set-ups favored by the well-known task-oriented competitions of statistical NLP.

Although the reported exercise focused on stages of NLU that can be, to some degree, computed outside of context, the overall program of work moves toward explainable AI covering integrated agent functionalities.

Acknowledgments

This research was supported in part by Grants #N00014-16-1-2118 and # N00014-17-1-2218 from the U.S. Office of Naval Research. Any opinions or findings expressed in this material are those of the author and do not necessarily reflect the views of the Office of Naval Research.

References

Allen, J. F., Chambers, N. et al. (2007). PLOW: A collaborative task learning agent. *Proceedings of the 22nd National Conference on Artificial intelligence* (AAAI '07), Vol. 2, pp. 1514-1519. AAAI Press.

Davies, M. (2008-). The Corpus of Contemporary American English: 450 million words, 1990-present.

Clark, A., Fox, C., & Lappin, S. (2010). *The Handbook of Computational Linguistics and Natural Language Processing*. Wiley-Blackwell.

Jurafsky, D., & Martin, J. H. (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics*. 2nd edition. Prentice-Hall.

Lindes, P., & Laird, J. E. (2016). Toward integrating cognitive linguistics and cognitive language processing. *Proceedings of the 14th International Conference on Cognitive Modeling* (ICCM). University Park, Pennsylvania.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations,* pp. 55–60. Stroudsburg, PA: The Association for Computational Linguistics.

McShane, M., & Nirenburg, S. (Forthcoming). Context for language understanding by intelligent agents. *Applied Ontology*.

McShane, M., Nirenburg, S., & Beale, S. (2016). Language understanding with Ontological Semantics. *Advances in Cognitive Systems* 4: 35-55.

McShane, M., Nirenburg, S., & English, J. (2018). Multi-stage language understanding and actionability. *Advances in Cognitive Systems* 6: 1-20.

Nirenburg, S., & McShane, M. (2016). Natural language processing. In Chipman, S. (Ed.), *The Oxford Handbook of Cognitive Science, Volume 1*. New York: Oxford University Press. Online publication date: August 2016.

Nirenburg, S., McShane, M., & Beale, S. (2008). A simulated physiological/cognitive "double agent." In Beal, J., Bello, P., Cassimatis, N., Coen, M. & Winston, P. (Eds.), *Papers from the Association for the Advancement of Artificial Intelligence (AAAI) Fall Symposium "Naturally Inspired Cognitive Architectures*". AAAI Technical Report FS-08-06. Menlo Park, CA: AAAI Press.

Nirenburg, S., McShane, M., Beale, S., Wood, P., Scassellati, B., Mangin, O., & Roncone, A. (2018). Toward human-like robot learning. *Natural Language Processing and Information Systems,* Proceedings of the 23rd International Conference on Applications of Natural Language to Information Systems (NLDB 2018), pp. 23-82.

Nirenburg, S., & Raskin, V. (2004). *Ontological Semantics*. The MIT Press.

Nouri, E., Artstein, R., Leuski, A., & Traum, D. (2011). Augmenting conversational characters with generated question-answer pairs. *Proceedings of the AAAI symposium on question generation*.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716. Doi: 10.1126/science.aac4716

Traum, D. R. (1994). A Computational Theory of Grounding in Natural Language Conversation. PhD thesis, Department of Computer Science, University of Rochester.

Generic noun phrases in child speech

Samarth Mehrotra (samarth.1397@gmail.com)
Department of Computer Science, Birla Institute of Technology and Science
Zuarinagar, Goa 403726 India

Amy Perfors (amy.perfors@unimelb.edu.au)
Department of Psychological Sciences, University of Melbourne
Redmond Barry Building, VIC 3010 Australia

Abstract

A wealth of developmental evidence suggests that children essentialise natural kind but not artifact categories, and that both adults and children use generic language less with artifacts as well (Gelman, 2003). Here we further explore the latter result using a novel model for generic identification. We apply our model to a much larger dataset than before, consisting of 26 CHILDES corpora of naturalistic speech involving children at a variety of ages and in a variety of contexts. We found no consistent preference for generic usage in animates over artifacts. Follow-up analyses indicate that this result was probably driven by our inclusion of a wider variety of nouns into our dataset than previous work.
Keywords: essentialism; generics; development; language

Introduction

Psychological essentialism refers to the intuitive belief that many categories have a hidden essence which gives the objects in those categories their identity. Essentialised categories have sharp boundaries, are discovered rather than invented, and have properties that are inherent in some way (e.g., Gelman, 2003). From an early age children behave in ways that are consistent with having essentialist beliefs. This is evident in how they use category information to support induction (Gelman & Markman, 1986) and make predictions about innate potential (Gelman & Wellman, 1991) and identity in the face of transformation (Keil, 1989), among others.

Although there is robust evidence that people essentialise natural kinds, we do not appear to essentialise artifact categories (e.g., Sloman & Malt, 2003). Artifacts do not retain their identity even when transformed (Keil, 1989), often have fuzzy category boundaries (Estes, 2003), and have different insides than animals do (Simons & Keil, 1995).

To what extent is this difference between artifact and natural kinds learned from or supported by environmental differences? One way to answer this question is by investigating one possible source of environmental influence: the use of generic noun phrases (e.g., *Owls sleep during the day* or *Books are heavy*). Generics communicate properties about categories as a whole rather than individuals, and both adults and children appear to make more essentialised inferences when generics are used (Rhodes, Leslie, & Tworek, 2012). Moreover, in a variety of experimental contexts, both children and adults produce generics more often for animals than for artifacts (Gelman & Tardif, 1998; Gelman, Coley, Rosengren, Hartman, & Pappas, 1998; Goldin-Meadow, Gelman, & Mylander, 2005; Brandone & Gelman, 2013). This is highly suggestive that environmental input in the form of generic language usage may play a role in children's early acquisition of essentialised beliefs.

However, the generality of these studies are limited somewhat because they all involved highly structured tasks, often with stimuli specifically created for the experiment. To our knowledge only one study has explored truly *naturalistic* generic language use. Gelman, Sarnecka, and Flukes (2008) hand-coded six corpora for generic language use and found the same bias toward generics in animates over artifacts.

Our work here builds on and extends this research by presenting an automatic model of generic identification. After validating its performance against several external metrics, we apply it to 26 different CHILDES corpora (including the six original ones). Our goal with this larger dataset was to learn more about the range of variation in generic usage in natural speech with children. Are generics used less with artifacts for all corpora, at all ages, and for all speakers? Do the patterns in generic usage support the possibility that psychological essentialism may reflect (or lead to) the statistics of generic speech in the linguistic environment?

Method

The first contribution of our work is the creation of a novel model that can automatically identify generic noun phrases based only on syntactic information. We describe it here.

Model

Although several models for the automatic identification of generic noun phrases exist, they are not ideal for our purposes. For instance, Reiter and Frank (2010) use a Bayesian Network model that relies on a feature set consisting of a large range of both the syntactic and semantic features of the noun itself as well as the clause it is contained in. Example syntactic features include COUNTABILITY, NUMBER, and PART OF SPEECH, while semantic features include SENSE and GRANULARITY. Friedrich and Pinkal (2015) use a conditional random field to label sequences but rely on a similar range of features, both syntactic and semantic.

The reliance on semantic as well as syntactic features is not a problem in general, but does pose an issue for us since our central questions focus on the semantic properties of generic nouns. Do they tend to be animates, artefacts, or something else? We cannot answer this question with a model that identifies generics using semantic features, since any results might emerge due to biases in how the model uses that semantic in-

Word	Part of speech	Dependency label
Elephants	noun	nsubj
do	verb	aux
not	adv	neg
eat	verb	ROOT
birds	noun	dobj
.	punct	punct

Table 1: Example sentence along with the two features used by our model: part of speech and dependency label, which indicates the role each word plays in the syntactic structure.

formation rather than actual distributional properties of the language. We therefore developed a new model of our own which relies only on syntactic features.

Structure Our model is a deep neural network classifier which makes decisions about noun phrases based on their syntactic properties as well as the syntactic properties of other words in the same clause. It therefore incorporates a notion of (local) context: an important consideration when identifying generics because the same word may or may not be a generic depending on how it is used. For instance, the word "dogs" in the sentence *Dogs like to bark* is generic, but the same word in the sentence *Dogs at Pat's house like to bark* is not.

Our classifier was constructed by stacking two different kinds of neural network units together. The first, Long Short-Term Memory (LSTM) units, are especially appropriate to classifying sequence-based data such as words in a sentence, and are widely used in many natural language applications (Hochreiter & Schmidhuber, 1997). We also used Gated Recurrent Units (GRUs) which are similar to LSTMs but often achieve higher performance on smaller datasets like ours. Our model consisted of seven different independently-trained architectures which varied from each other in the dimensionality of the units as well as in how they were stacked.[1]

All of the architectures had a final, fully-connected layer with a softmax activation function which performed the classification task. Each architecture yielded one decision for each noun (generic vs not-generic) and model decisions were made by taking the majority vote among the seven.

Input Our model required two kinds of syntactic information for each of the words in our corpora: the part of speech as well as the dependency label it was associated with in the dependency parse tree. Table 1 illustrates these features for an example sentence. In order to extract this information, we used a number of standard state-of-the art natural language processing tools. We first segmented each of the nouns and their corresponding clauses out of each sentence using the discourse parser SPADE (Soricut & Marcu, 2003). Each word was then assigned a dependency label using the Stanford Dependency Parser (Chen & Manning, 2014) and then tagged with the appropriate part of speech (Toutanova, Klein, Manning, & Singer, 2003).

[1]Our anonymised supplementary materials describe the structure of the architectures: `https://tinyurl.com/ybwg88h5` .

Model	Accuracy	F-score
Reiter and Frank (2010)	71.7	72.3
Friedrich and Pinkal (2015)	79.1	78.8
Our model	76.4	79.3

Table 2: Cross-validation performance on the WikiGenerics dataset. Our model achieves similar performance to the state-of-the-art. Accuracy reflects the total percentage of correct predictions (generics classified as generics, and non-generics as non-generics) while F-score is the harmonic mean of precision and recall, as calculated in Friedrich and Pinkal (2015).

Pronouns posed an interesting dilemma, because they make up a reasonable proportion of all nouns yet cannot be accurately classified for their genericity without determining their referent. For instance, the word "they" in the sentence *Watch out for the piranhas in that fish tank; they bite* is not generic, whereas the word "they" in *I hate mosquitoes; they bite* is generic. We addressed this issue by resolving the coreference of each pronoun using a standard coreference resolution system (Clark & Manning, 2016), and then assigning the genericity of the pronoun to be the same as its referent.

Using these part of speech and dependency features, we created input vectors for our model that corresponded to each noun along with the sequence of words in the clause. This means that for each noun, the model was given not just the noun but also all of the words in the NP it was part of and all of the words in the clause that contained that NP. Each input vector was a concatenation of two vectors consisting of the part-of-speech tag and the dependency label. The model thus used all of the words in the sequence to make a decision about each noun, not just the words that came before it.

Training and validation Each of our seven architectures was trained independently using a weighted categorical cross entropy loss function, which we optimised using the Adam optimiser (Kingma & Ba, 2014). Our loss function weighted the error associated with classifying a non-generic statement as generic (false positive) 1.5 times more than the error associated with classifying a generic statement as non-generic (false negative). By using such a weighted error function, we ensured that the classifier was conservative in its classification of generics, marking a noun as a generic only when it was very confident. This helped to ensure that our model was not overestimating the proportion of generic words.

Before applying our model to CHILDES corpora, we validated its performance in two ways. First we calculated its accuracy and F-score on the WikiGenerics dataset created by Friedrich and Pinkal (2015). This dataset consists of examples from 102 documents from Wikipedia covering a wide variety of topics including animals, games, medicine, music, politics, science, and people, among others. The texts were hand-annotated for genericity by three computational linguists, with contested annotations decided by majority vote. We tested our model using as leave-one-out cross validation strategy. In each cross validation step, examples from 101 of the 102 texts were used for training and the model was tested on the remaining one. The results, shown in Table 2, show

that despite relying on a much smaller range of features our model performed as well as the two best-performing models of generic identification.[2]

Although this level of performance is reassuring, it is not necessarily the case that high performance on a dataset consisting of Wikipedia articles means that the model can accurately identify generics in corpora of child with children. As a second validation of model performance, we thus tested its accuracy against the genericity judgments reported in Gelman et al. (2008).[3] The data we had access to consisted of all of the nouns (in the child speech only) in their six corpora that they coded as generic. Our model had a 88% true positive rate on this data: 88% of the items that they coded as generic were coded as generic by our model. We do not have the list of nouns that they coded as non-generic, but on the assumption that any nouns not coded as generic would have been coded as non-generic, this gives our model an accuracy of 96.8 and an F-score of 81.2 against their gold standard.

CHILDES Datasets We applied our model to 26 different corpora from the CHILDES database (MacWhinney, 2000). The corpora, which are listed in full in the supplemental materials, include the six corpora from Gelman et al. (2008) as well as twenty additional corpora made up of natural conversations between children and adults in English (American or UK). All corpora include both adult and child speech except one (Sawyer) which contained only child speech. Because we were interested in the statistics of language in naturalistic situations, we excluded studies in which children were given a structured task or played with a restricted set of toys.

The supplemental materials list all corpora in detail, but in general, the children ranged in age from less than one year to over five years of age. Given the difficulty in identifying generic usage when grammatical abilities are limited, we excluded all child speech from children less than two years old. However, we do include adult speech to these children because one of our goals with this work is to better understand the distributional properties of the linguistic input they receive at all ages. Our full corpus of child speech contained 1,057,807 utterances total and the corpus of adult speech contained 1,595,305 utterances.

Results

Our first question is about the prevalence of generic speech as a function of age. For the child corpora, we can ask when children begin producing generics. For the adult corpora, we can ask whether adult speech is rich in generics from an early age, and whether there are developmental trends in generic usage. We thus calculated the proportion of generic utterances at different age ranges, coding an utterance as generic if any noun in it was classified as generic. Our results are shown in Figure 1, plotted alongside similar data from Gelman et al. (2008). Although we show more generic usage overall than did Gelman et al. (2008), the patterns are remarkably similar. Children's production increases rapidly over the early years of development, with them producing generics as soon as they have the grammatical capacity. In the early years, adult production is consistently higher than children's, but it then levels off at later ages until they converge. We consider reasons that we estimate more generics in the Discussion.

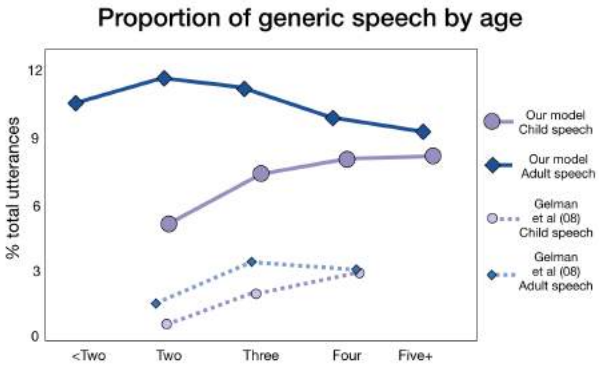


Figure 1: Proportion of generic speech by age. The overall percentage of all utterances coded as generic in our corpora (solid line), broken down by child and adult speech (purple and blue, respectively). For comparison, we plot analogous results from Gelman et al. (2008) with the dotted line. Although we estimated more total generics than they did, the qualitative patterns over development and between child and adult speech are extremely similar.

The primary question motivating this work was how generic usage differs between different kinds of nouns. Do animates, which both children and adults essentialise more, occur more often in generic speech than artefacts, which are essentialised less? In order to investigate this question we had to assign each of the nouns in our corpus to the appropriate category. We accomplished this based on the categories in WordNet, a widely-used lexical database for English. WordNet contains 22 different noun categories, including animals, artifacts, and people as well as feelings, communications, plants, motives, substances, time, and more.

We classified all of our nouns into the four categories used by Gelman et al. (2008): animates, artifacts, food, and other. The artifact and food categories correspond straightforwardly to equivalent categories in WordNet. We constructed our animates category by combining the WordNet animal and person categories, and classified everything else as other. If a word was associated with multiple WordNet categories, we used the Lesk Algorithm to determine which one to assign it to. This algorithm uses the words in the surrounding context to determine the appropriate classification. For instance, the word *fish* would be classified as an animal if it was surrounded by words like *swim* or *water* and as a food if it was surrounded by words like *eat* or *cook*.

What kinds of noun categories do people talk about more, and does this distribution vary between adults and children or by whether generics or non-generics are involved? To answer this question, Figure 2 plots the percentage of each of the four noun categories within generics and non-generics, re-

[2]These numbers are as reported in Friedrich and Pinkal (2015). We did not re-implement their models.

[3]We would like to thank Susan Gelman, who graciously provided this data upon request.

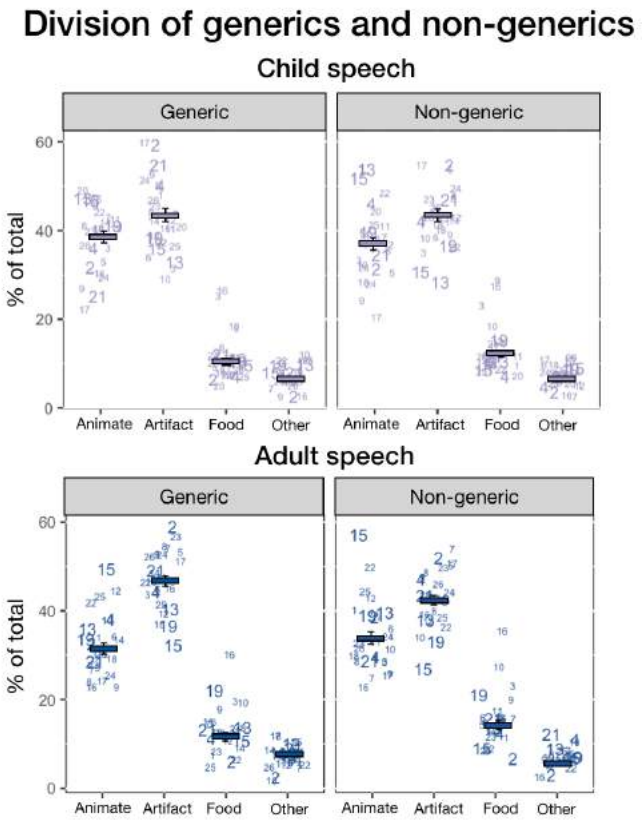


Figure 2: Proportion of generic and non-generic speech across categories. This figure shows the distribution of speech across the four noun categories, for both children (purple) and adults (blue). Lines show the mean when averaged by corpus; error bars indicate standard error. The numbers correspond to the relevant measure for each of the 26 corpora. The corpora from Gelman et al. (2008) are slightly larger and correspond to numbers 2, 4, 13, 15, 19, and 21. It is evident that there is high variability between corpora, but for the most part both children and adults speak about artifacts more often and that there is little difference between generics and non-generics in how they are distributed amongst the four noun categories.

spectively. The left panel thus shows the percentage of all of the generic nouns that are animates, artifacts, foods, or other; the right panel shows the same breakdown out of all of the non-generic nouns. We illustrate the variability in this distribution by plotting the results for each of the corpora individually. It is evident that there is substantial variability overall, and that at least some of that variability is corpus-specific: the correlation between adult and child speech by corpus is $r = 0.95$. This probably largely reflects the fact that children and adults co-create one another's linguistic environment.

This analysis also demonstrates that in general both children and adults talk about artifacts slightly more often than animates.[4] There is also no difference in the distribution of

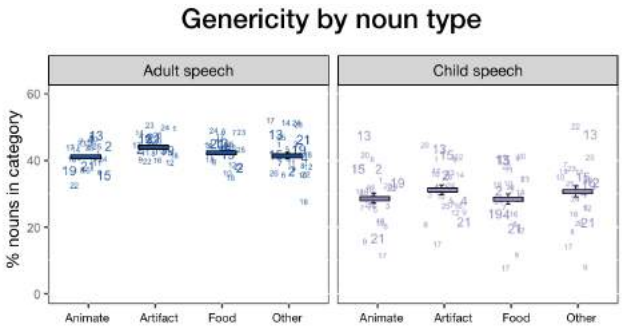


Figure 3: Proportion of generic speech within each noun category. For each of the four categories, this figure shows how often nouns in that category were generic. The large transparent bars indicate the aggregate proportion over all corpora, while the small boxes with error bars show the mean when averaged by corpus. The numbers correspond to the relevant measure for each of the 26 corpora. The corpora from Gelman et al. (2008) are slightly larger and correspond to numbers 2, 4, 13, 15, 19, and 21. There is high variability between corpora (especially for children). However, there is little difference in the pattern of generic usage across noun categories.

speech across noun categories as a function of genericity or speaker.[5] Generics and non-generics have similar distributions across different kinds of nouns, and this holds regardless of whether the speakers are adults or children.

Another way to explore the issue of whether children or adults use generics differently for different categories is to condition on category rather than on genericity. Figure 3 thus shows, for each of the four noun categories, what proportion of time it occurs as a generic in both child and adult speech. Although children are much more variable, we still see little difference in generic usage between noun categories. However, adults were more likely to use generics for artifacts than animates, as well as more overall.[6]

These results are rather surprising, since previous work has suggested that generics tend to be used more often with animate categories. What is going on?

One possibility might be that the six corpora used by Gelman et al. (2008) were outliers in some way relative to our larger set of 26. In order to investigate this possibility, we calculate how many corpora used a higher percentage of animate nouns than artifact nouns as generics. On this measure, the corpora from Gelman et al. (2008) appear to be slight outliers relative to the others. Of the 25 corpora with adult speech, only six used generics more with animates and three of those six were theirs: Bloom (2), Brown (4), and Kuczaj (13). Of the 26 with child speech, six used generics more with animates and four were theirs: 2, 4, 13, and Sachs (19).

[4]Bayesian t-test comparing artifact to animate percentage: For child generics, $BF_{10} = 2.4$ weakly in favour of a model that includes noun type; for child non-generics, $BF_{10} = 7.7$ moderately in favour. For adult generics, $BF_{10} > 10^6$ in favour of a model that includes nountype; for adult non-generics, $BF_{10} = 111$ in favour. All Bayesian analyses used the BayesFactor package in R (version 3.4.4) and compared the model of interest to an intercept-only null model. In also cases we also ran analogous frequentist tests, which always returned qualitatively similar results.

[5]Bayesian ANOVA: $BF_{01} = 10$ for the null model over a model including genericity and $BF_{01} = 10$ for the null over a model including speaker. This indicates strong support for the null model.

[6]Bayesian ANOVA: $BF_{01} = 14.3$ favouring the null model over a model including noun category; $BF_{10} > 10^6$ favoring a model including speaker. Bayesian t-test comparing artifact to animate generic percentage: for child speech, $BF_{01} = 1.9$ favouring the null model; for adult, $BF_{10} = 6.9$ favouring a model including nountype.

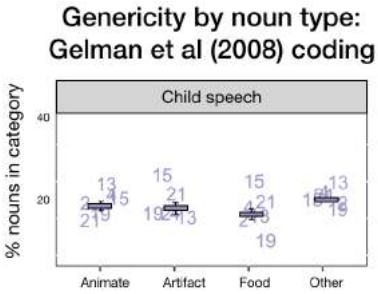


Figure 4: Proportion of generic speech within each noun category using the genericity identifications from Gelman et al. (2008). For each of the four categories, this figure shows how often nouns in that category were generic, using the six corpora and their classifications rather than the classifications from our model. Despite using their classifications, we replicate our previous result, suggesting that the difference between our findings and theirs did not arise due to poor classification performance by our model.

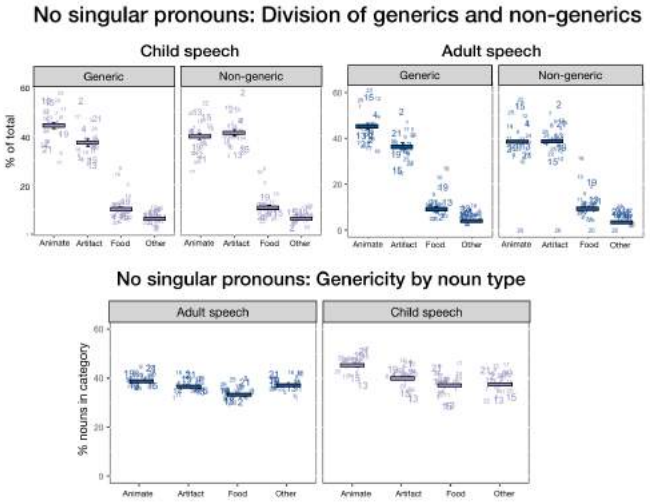


Figure 5: Proportion of generic speech across and within noun categories, on corpora without any singular pronouns. Since Gelman et al. (2008) excluded singular pronouns, we reran our analyses (using our classifications) on our corpora after excluding all singular pronouns. Results are now much more similar to their findings than ours. Generics but not non-generics are used more for animate than artifact categories (top); and for both adults and children, the proportion of generic utterances in animates is higher than in artifacts. This suggests that their exclusion of singular pronouns from the dataset may have driven their results.

These considerations suggest that at least part of the reason our results diverge so markedly from Gelman et al. (2008) is that their corpora were different. However, this cannot be the entire story: the magnitudes of the differences they found are much larger than the magnitudes we found on the same corpora in very similar analyses.

An obvious possibility is that our model is simply classifying many items very differently than they did. Our high accuracy and F-score against their coding scheme suggests that this is not the case, but we were able to test this hypothesis in a much more stringent way as well. For the six corpora in Gelman et al. (2008) that we have their classifications for (child speech only), we took the set of nouns that they identified as generic, assumed that they coded all of the others as non-generic, and applied the same analysis as in Figure 3 to that data. If the difference between our work is because our classifier is coding or identifying items differently than they did, we should find that using their classifications on their corpora replicates their results. However, Figure 4 reveals that we instead replicate our result: there is no difference in generic usage across the four noun categories.[7]

This outcome suggests that the point of divergence between our work and Gelman et al. (2008) must be less due to different decisions about what to code as generic, and more due to different decisions about what nouns to include in the first place. Our analysis included all nouns of any kind, which was straightforward to do since the model could identify them automatically. However, lacking this technology, Gelman et al. (2008) had to process the corpora by hand. They accomplished this by manually identifying potential generics by searching for any bare plurals, plural pronouns, mass nouns, and indefinite singular nouns and then hand-coding that set of nouns as generic (or not). This was justified on the grounds that the vast majority of generics fall into these categories, which is sensible if the goal is to understand the distribution of generics alone. However, if the goal is also to compare to non-generics, it is important to include even those nouns that tend to be non-generic. Their dataset excluded singular pronouns like *he, she, you, I*, and *it*. If singular pronouns tend to "cluster" (for instance, are more likely to be animate and non-generic) then excluding them might result in a mis-estimation of the overall distribution of generics relative to non-generics in different ways for different noun categories.

To test whether the inclusion or exclusion of singular pronouns drove the difference between our results and those of Gelman et al. (2008), we re-ran our original analyses after excluding all singular pronouns from our dataset. As shown in Figure 5, the results now replicate their findings rather than ours. The top panel shows that generics but not non-generics are used more for animate than artifact categories,[8] and the bottom panel shows that for both adults and children, the proportion of generic utterances is higher within animate categories than artifacts.[9] This suggests that Gelman et al. (2008) may have found that animate categories had more generics because they did not count a large number of non-generic animates like *he, she, you,* and *I*. Our other analysis show that once all nouns are included, the proportion of generics across noun categories evens up and if anything favours artifacts.

[7]Bayesian ANOVA: $BF_{01} = 2.6$ for the null over a model including noun category. Bayesian t-test comparing animates to artifacts: $BF_{01} = 2.1$ for the null over a model including noun category.

[8]Bayesian t-test comparing artifact to animate percentage: For child generics, $BF_{10} = 91$ in favour of a model that includes nountype; for child non-generics, $BF_{01} = 2.9$ for the null model. For adult generics, $BF_{10} = 206$ in favour of a model that includes nountype; for adult non-generics, $BF_{01} = 3.5$ for the null model.

[9]Bayesian ANOVA: $BF_{10} > 10^6$ favouring a model including noun category; $BF_{10} = 522888$ favoring a model including speaker. Bayesian t-test comparing artifact to animate generic percentage: for child speech, $BF_{10} = 2529$ favouring a model including nountype; for adult, $BF_{10} = 35$ favouring a model including nountype.

Discussion

This work makes several contributions. First, we present the first fully automatic model for generic identification which uses only synactic features, and demonstrate that it performs well relative to both the state-of-the-art and manual classifications from Gelman et al. (2008). Second, we apply this model to a much larger dataset of child speech than had previously been possible to analyse. Although we replicate the previously-observed developmental trend in generic usage, we find that neither adults nor children use generics more in categories that tend to be essentialised (like animates). Follow-up analyses suggest that our results differ from Gelman et al. (2008) not because of poor classification performance by our model, but primarily because we did not exclude singular pronouns from our dataset (as they did).

A natural question at this point is whether it is better to include singular pronouns or not. Any answer must be conditioned on considerations of what is realistically possible. Given the extreme amount of labour involved in hand-coding corpora, one can reasonably argue that the process for identifying nouns used by Gelman et al. (2008) was a necessary simplification. Other analyses excluded pronouns for other good reasons. For instance, Gelman and Tardif (1998) and Goldin-Meadow et al. (2005) excluded pronouns because of the need to compare English with Mandarin, a pro-drop language. Given these considerations, this too seems reasonable. However, it is possible that this decision is why they as well found a higher proportion of generics for animates.

Overall, we suggest that if the goal is to understand the distribution of generics relative to non-generics in the nouns children hear, it is important to include *all* of the nouns that children hear. Singular pronouns are very common and almost always non-generic; as such, an accurate comparison of generics to non-generics cannot exclude them.

One might also ask why our model identified a larger proportion of generics than previous work did (Figure 1). Part of the reason is probably that a manual identification of generics, as Gelman et al. (2008) had to do, would probably have erred on the side of under-counting them. Another part is that our model appeared to make less conservative choices in some cases. For instance, our model identified many generics that were preceded by the word *the*, as in sentences like *What do bears in the forest do in the day?*. Since our observed developmental trends are very similar and all of our other results hold even when we use the classifications from Gelman et al. (2008), we doubt that our overall higher rate poses a problem.

A final question is what our results mean for our initial question: to what extent does the linguistic environment support the difference in essentialisation of artifact vs animate categories? Our results suggest that this difference is not reflected in differences in generic usage, and thus lends less credence to the possibility that these domin differences in essentialism result from linguistic input. Although this finding is surprising given previous work, one nice aspect of it is that it removes the chicken-and-egg question that otherwise arises: *why* does the linguistic environment have this distribution in the first place? Much remains to be done, but we hope that our model and results offer a useful tool for better understanding how our early biases are shaped by the environment.

References

Brandone, A., & Gelman, S. (2013). Generic language use reveals domain differences in children's expectations about animal and artifact categories. *Cog. Dev.*, *28*, 63–75.

Chen, D., & Manning, C. (2014). A fast and accurate dependency parser using neural networks. In *EMNLP*.

Clark, K., & Manning, C. (2016). Improving coreference resolution by learning entity-level distributed reps. In *ACL*.

Estes, Z. (2003). Domain differences in the structure of artifactual and natural categories. *Mem & Cogn.*, *31*, 199–214.

Friedrich, A., & Pinkal, M. (2015). Discourse-sensitive automatic identification of generic expressions. In *ACL*.

Gelman, S. (2003). *The essential child: Origins of essentialism in everyday thought*. Oxford Univ. Press.

Gelman, S., Coley, J., Rosengren, K., Hartman, E., & Pappas, A. (1998). Beyond labeling: The role of maternal input in the acquisition of richly-structured categories. *SRCD*, *63*.

Gelman, S., & Markman, E. (1986). Categories and induction in young children. *Cognition*, *23*, 183–209.

Gelman, S., Sarnecka, P. G., & Flukes, J. (2008). Generic language in parent-child conversations. *LL&D*, *4*(1), 1–31.

Gelman, S., & Tardif, T. (1998). A cross-linguistic comparison of generic noun phrases in English and Mandarin. *Cognition*, *66*(3), 215–248.

Gelman, S., & Wellman, H. (1991). Insides and essences: early understandings of the nonobvious. *Cogn.*, *38*.

Goldin-Meadow, S., Gelman, S., & Mylander, C. (2005). Expressing generic concepts with and without a language model. *Cognition*, *96*, 109–126.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Keil, F. (1989). *Concepts, kinds, and cogn. devel.* MIT Press.

Kingma, D., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv:1412.6980*.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed.). Lawrence Erlbaum.

Reiter, N., & Frank, A. (2010). Identifying generic noun phrases. In *ACL* (pp. 40–49).

Rhodes, M., Leslie, S., & Tworek, C. (2012). The cultural trans. of social essentialism. *PNAS*, *109*, 13526–13531.

Simons, D., & Keil, F. (1995). An abstract to concrete shift in the development of biological thought: The insides story. *Cognition*, *56*, 129–163.

Sloman, S., & Malt, B. (2003). Artifacts are not ascribed essences, nor are they treated as belonging to kinds. *Language and Cognitive Processes*, *18*, 563–582.

Soricut, R., & Marcu, D. (2003). Sentence level discourse parsing using syntactic and lexical info. In *NAACL*.

Toutanova, K., Klein, D., Manning, C., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *NAACL*.

Online Phonetic Training Improves L2 Word Recognition

Gerda Ana Melnik (gerda.ana.melnik@ens.fr)
Sharon Peperkamp (sharon.peperkamp@ens.fr)

Laboratoire de Sciences Cognitives et Psycholinguistique (ENS, EHESS, CNRS),
Département d'Etudes Cognitives, Ecole normale supérieure - PSL University,
29 rue d'Ulm, 75005 Paris, France

Abstract

High-Variability Phonetic Training (HVPT) has been shown to be effective in improving the perception of even the hardest second-language (L2) contrasts. However, little is known as to whether such training can improve phonological processing at the lexical level. The present study tested whether this type of training also improves word recognition. Adult proficient French late learners of English completed eight online sessions of HVPT on the perception of English word-initial /h/. This sound does not exist in French and has been shown to be difficult to process by French listeners both on the prelexical (Mah, Goad & Steinhauer, 2016) and the lexical level (Melnik & Peperkamp, 2019). In pretest and posttest participants completed an identification task as well as a lexical decision task. The results demonstrated that after training the learners' accuracy had improved in both tasks. The theoretical and applied implications are discussed.

Keywords: second language acquisition; lexical processing; word recognition; speech perception; phonetic training

Introduction

It is well known that producing and perceiving non-native speech sounds can be very challenging (for reviews, see Piske, MacKay & Flege, 2001; Sebastián-Gallés, 2005). In the realm of perception, much research has shown that with auditory training, the difficulty of perceiving even the hardest non-native sounds can be reduced. The most common training paradigm used to improve second language (L2) perception is High-Variability Phonetic Training (HVPT). HVPT uses multiple natural exemplars of the target sounds in a variety of phonetic environments. This variability enhances the process of building novel phonological categories. Importantly, perceptual training involves immediate corrective feedback that provides information to participants about their performance and promotes rapid learning by driving the learner's attention to the relevant phonetic cues of the sounds to be learned (Homa & Cultice, 1984; Logan, Lively & Pisoni, 1991). The effectiveness of this technique has been shown in many studies in a variety of languages, using several target contrasts and structures, including vowels (Carlet & Cebrian, 2014; Lee & Lyster, 2016), consonants (Kim & Hazan, 2010; Shinohara & Iverson, 2018), tones (Wang et al. 1999; Wang, Jongman, & Sereno, 2003), and syllable structure (Huensch & Tremblay, 2015). Moreover, both high- and low-proficiency speakers benefit from HVPT (Iverson, Pinet & Evans, 2012), and HVPT generalizes to new tokens and new speakers (Lively et al., 1994; Okuno & Hardison, 2016). Finally, it gives rise to long-term retention of the new categories (Lively et al., 1994), and it helps to improve L2 production (for a review, see Sakai & Moorman, 2018).

Although the effectiveness of HVPT is well studied, most previous work focused exclusively on prelexical perception, using identification or discrimination tasks. The difficulty with the perception of L2 sounds, though, is paralleled by less efficient lexical processing (e.g., Pallier, Colomé & Sebastián-Gallés, 2001; Weber & Cutler, 2004). Thus, truly successful training should also enhance performance at the lexical level. While prelexical processing only involves a phonetic analysis, lexical processing is more complex as it additionally requires mapping the incoming speech signal onto phonological representations stored in memory, and the performance gap between native and non-native listeners in L2 speech perception increases as the tasks have greater lexical involvement (Díaz et al., 2012).

So far, the only studies on the effect of prelexical auditory training on lexical processing focused on naïve listeners' ability to learn words in a tonal language (Cooper & Wang, 2011; Ingvalson, Barr & Wong, 2013). Both studies found that naïve English listeners' ability to learn words involving difficult tone contrasts improved after auditory training. To our knowledge, no studies have directly assessed the effect of auditory training on enhancing word recognition in L2 learners.

We focused on the perception of the English sound /h/ by intermediate French learners of English. As /h/ does not exist in French, French listeners – even those who are fluent in English – have difficulty perceiving the contrast between the presence vs. absence of /h/ in English stimuli (Mah et al., 2016). At the lexical level, proficient French learners of English tend to accept nonwords such as *usband* (cf. *husband*) and, to a lesser extent, *hofficer* (cf. *officer*), as real words (Melnik & Peperkamp, 2019). Thus, they have difficulty not only in perceiving the contrast between /h/ and silence, but also in distinguishing between words and nonwords that differ only in the presence vs. absence of /h/.

Importantly, there is an almost perfect one-to-one mapping in English of the grapheme <h> onto the phoneme /h/. Most French L2 speakers know how to correctly write /h/-initial words. They are also instructed that <h> is rarely silent in English and that it is pronounced as /h/. If after training learners start better perceiving /h/, they might thus be able to also improve their recognition of /h/-initial English words even if they have imprecise phonological representations of such words, since they can rely on the orthography.

In the current study we trained French learners on the perception of English /h/ in a pretest–training–posttest design. In pretest and posttest, participants performed an identification task aimed at testing their phonetic perception of /h/, and a lexical decision task aimed at testing their processing of /h/ at the lexical level. In the posttest, the identification task also tested for generalization to novel items. In the identification task we used /h/- and vowel-initial nonwords as stimuli. In the lexical decision task we used words and nonwords, where the test nonwords were created from /h/-initial and vowel-initial words by removing or adding /h/, respectively.

Training was administered on-line, and consisted of eight sessions of an identification task using minimal pairs of real words (such as *air*-*hair*), with corrective feedback.[1] We expected the training to enhance performance in the identification task at posttest, thus replicating the findings of previous studies on the effectiveness of HVPT in improving phonetic perception of L2 sounds. Moreover, if the effect of training extends to lexical processing, performance in lexical decision should likewise improve with training.

Method

Pretest-Posttest-Generalization: Identification

Stimuli

For the pre- and posttest we selected 100 pairs of nonwords. The members of each pair differed in the presence or absence of an initial /h/ (e.g. /hasp/ – /asp/). Forty pairs were monosyllabic, 40 dissyllabic and 20 trisyllabic. Ten English vowels (ʌ, ɒ, a, ɪ, ɛ, iː, ʌɪ, əʊ, eɪ, aʊ) were used in the first (or only) syllable, thus creating a large amount of variability in phonetic context.

An additional 30 pairs of nonwords (10 monosyllabic, 10 disyllabic and 10 trisyllabic, containing the 10 vowels mentioned above) were selected to test for generalization at the end of the posttest. Half of the pairs were recorded by a male, and the other half by a female native of American English.

Procedure

Participants were tested individually in a soundproof booth. In each trial they were presented auditorily with a stimulus; their task was to press as quickly as possible the key labelled "h" with their dominant hand if they thought the nonword started with the sound /h/, and to press the key labelled "no h" with their non-dominant hand if they thought it did not start with /h/. There were 194 trials divided over two blocks. Trials were presented in a semi-random order such that no more than four trials of the same type (vowel-initial or /h/-initial) and no more than three trials recorded by the same speaker appeared in a row.

The first block started with a practice phase of six trials, during which participants received feedback. In the case of an incorrect response or no response within 2500 ms, the trial was repeated until the correct response was given. During the test phase, participants received no feedback and if they did not give a response within 2500 ms the next trial was presented. An interval of 1000 ms elapsed between the participant's response or the time-out - whichever came first - and the presentation of the next trial.

At the end of the posttest only, 60 trials with the 30 additional nonword pairs were used to test for generalization.

Pretest-Posttest: Lexical decision

Stimuli

The stimuli were the same as in Melnik & Peperkamp (2019). They consisted of 80 English test words, 40 starting with /h/ (e.g., *husband*) and 40 with a vowel (e.g. *officer*), recorded by the same male American English speaker who recorded stimuli for the identification task. They consisted of nouns, verbs and adjectives, and contained between two and four syllables. The /h/-initial and the vowel-initial words did not differ in mean frequency in the Subtlex database (Brysbaert & New, 2009) or in mean number of syllables (both $t < 1$).[2]

Each word was paired with a nonword, created by deleting or adding /h/ at the beginning (e.g. *husband* → *usband*, *officer* → *hofficer*). In addition, there were 240 English control words (nouns, verbs and adjectives), none of which starting with /h/. They were matched for mean frequency and mean number of syllables with the test words. Each control word was paired with a nonword created by replacing, deleting or inserting one phoneme other than /h/.

The test and control minimal pairs were divided into two equal groups, one for pretest and one for posttest, respecting the matching in terms of frequency and number of syllables. The pretest stimuli were further divided into two counterbalancing lists: list A and list B. Each of them contained only one member of each pretest minimal pair. For instance, if the word *husband* was in list A, its nonword counterpart *usband* was in list B. The posttest stimuli were divided into lists C and D following the same principle. Thus,

[1] Training can be done either with nonwords (e.g., Yamada, 1991) or with real words (e.g., Logan et al., 1991). Here, we chose to use real words because repeated exposure to a large number of nonwords during training might have induced a bias to excessively accepting nonwords in the lexical decision task in pre- and posttest.

[2] The familiarity of these words was evaluated by a separate group of 45 adult French learners of English in an online rating questionnaire. The /h/- and vowel-initial words that were chosen for the experiment did not differ in mean familiarity ($t = 1.0$, $p > 0.1$).

no list contained both members of a given word–nonword pair. Each of the four lists contained 10 /h/-initial and 10 vowel-initial words, 10 /h/-initial and 10 vowel-initial nonwords, as well as 60 control words and 60 control nonwords. Finally, for a practice phase there were two additional words and two additional nonwords, none involving /h/.

Procedure

In pretest half of the participants were randomly assigned to one of the two pretest lists (list A or list B). In posttest, participants who previously heard the list A were given the list C, while participants who previously heard the list B, were now given the list D. Hence, participants heard only one of the members of each word-nonword pair throughout the whole experiment.

The procedure was identical to that in Melnik & Peperkamp (2019): Participants performed a speeded auditory lexical decision task. In each trial they heard a word or a nonword and had to answer if the item was an English word. They were instructed to use their dominant hand for "yes"- and their non-dominant hand for "no"-responses on a button box. There were 160 trials divided over two blocks, each containing the same number of test and control stimuli. Trials were presented in a semi-random order such that between one to three control trials appeared between two experimental ones, and that no more than four trials of the same type (word or nonword) appeared in a row.

The first block started with a practice phase of four trials with control items, during which participants received feedback ('correct' or 'wrong' written on the screen). In the case of an incorrect response or no response within 2500 ms, the trial was repeated until the correct response was given. During the test phase, participants received no feedback and if they did not give a response within 2500 ms the next trial was presented. An interval of 1000 ms elapsed between the participant's response or the time-out and the presentation of the next trial.

Training: Identification

Stimuli

We selected 59 minimal pairs of real words differing in the presence or absence of an initial /h/. Given the limited number of such minimal pairs, we used both frequent words (e.g. *hair-air*) and infrequent ones (e.g. *hosier-osier*) words. However, word frequency was not considered to have an impact, as the task used in training was prelexical.

Four different speakers, two men and two women, recorded the items. One of the male speakers and one of the female speakers were those who recorded the stimuli for the nonword identification task used in pretest and posttest, with the male speaker having also recorded the stimuli for the lexical decision task.

Procedure

The training consisted of eight high-variability phonetic training sessions. In the first four sessions participants heard one speaker per session. In the following four sessions they heard a pair of speakers in each session, such that all four male-female combinations were used.

All training sessions were run at the participants' homes through internet. The online training sessions were designed using the JsPsych library (de Leeuw, 2015) in JavaScript. Before each training session participants received by email a link to the corresponding training session webpage. Stimuli were presented at a comfortable listening level, set individually. The details of each training session (e.g., participant details, day and time of completion, RTs and responses) were automatically sent to the MySql database after the completion of each session. Participants could only do one session per day and there could be no more than one day in between two sessions. Thus, the whole course of training was completed in eight to fifteen days.

In each trial participants first saw the two response alternatives written on the screen (e.g. "hair – air"). The word starting with /h/ was always displayed on the left, and the word without /h/ always on the right. The auditory stimulus was played 800 ms later. The task was to press as quickly as possible the left arrow key if the word started with /h/ and the right arrow key otherwise. When the participant pressed the key, the corresponding word was highlighted in bold. If the response was correct, the word "Correct" written in green appeared in the middle of the screen, in between the two alternatives. If it was incorrect, the word "Wrong" written in red appeared on the screen, followed after 1000 ms by auditory feedback of the form: "*The word was not: XXX. It was: YYY*", spoken by the same speaker as the stimulus itself. For instance, if the stimulus played was the word "hair" but the participant chose instead the word "air", the word "Wrong" was displayed on the screen and the phrase *"The word was not: air. It was: hair"* was played.

If no response was given within 2500 ms, the words "Too slow" appeared on the screen. An interval of 1000 ms elapsed between the participant's response or the time-out - whichever came first - and the presentation of the next trial. There were 118 trials in each session, and trials were presented in a random order. Each session lasted from 15 to 20 min, depending on the accuracy of the participant.

Participants

Participants were French intermediate learners of English, recruited from among university students (about half of which in an English department). In order to avoid ceiling performance or insufficient knowledge of English vocabulary, only participants whose accuracy in pretest was below 80% in the identification task and above 70% on control items in the lexical decision task went through the training and posttest. Of the 51 participants who did the pretest, 25 satisfied these criteria, out of whom a total of 24 completed the study and were included in the data analysis. Among these participants, there were 12 women and 12 men, aged between 19 and 32 (mean: 22.3), who had started

learning English at school. They filled in a questionnaire to self-evaluate their speaking, listening, reading, vocabulary and grammar skills in English and French, on a scale from 1 to 10. The overall mean score was 6.4 (SD = 1.6) for English and 9.4 (SD = 0.9) for French.

None of the participants reported a history of speech or language problems. They received a small payment after the pretest, and those who underwent training received a second, larger, payment when they came back to the laboratory for the posttest.

Results

Pretest-Posttest-Generalization: Identification

Prior to analysis, we discarded responses with a reaction time of 0 ms. Figure 1 displays the identification accuracy of participants in pretest, posttest and generalization. As the identification task is a signal detection task, we used the A' statistic, which provides a non-parametric, unbiased, index of sensitivity (here: to the difference between words and nonwords), with 0.5 indicating chance performance and 1.0 perfect performance. A repeated measures ANOVA by participant with the factor Session (Pretest vs. Posttest vs. Generalization), revealed a main effect of Session ($F(2,46) = 26.75$, $p < .001$), with the accuracy improving from an average A′ score of 0.74 in pretest to 0.86 in posttest and 0.86 in generalization. Bonferroni-adjusted pairwise t-tests revealed that there was a significant difference between pretest and posttest ($p < .01$), as well as between pretest and generalization ($p < .01$). There was no difference between the performance in the posttest and in the generalization ($p = .82$).

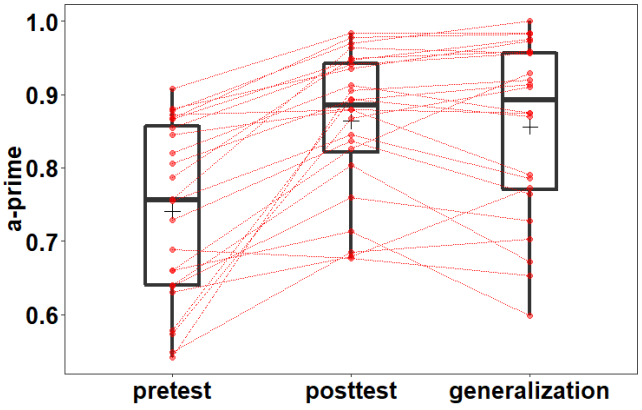


Figure 1. Boxplots of A′ scores in the identification task in pretest, posttest, and generalization. The red dots represent individual participants; the lines link each participant's performance in the three sessions. The black cross marks indicate mean A′ scores in each session.

Pretest-Posttest: Lexical Decision

Prior to analysis, we discarded responses with 0 ms reaction time. Figure 2 displays the accuracy of participants on the test items in pretest and posttest. As the participants had a strong bias for 'yes'-responses (shown by their low accuracy scores on test nonwords), we used the A' statistic as in the analysis of performance in the identification task.

We carried out a repeated measures ANOVA by participant with the factors Session (pretest vs. posttest), Condition (test vs. control) and Lists (AC vs. BD), as well as an interaction between Session and Condition. We found main effects of Session ($F(1, 23) = 39.36$, $p < .001$) and Condition ($F(1, 23) = 73.93$, $p < .001$), and a Session X Condition interaction ($F(1, 23) = 30.87$, $p < .001$). Pairwise t-tests revealed that the interaction was due to the fact that in control items, the effect of Session was not significant, while in test items, there was a significant difference between pretest and posttest ($p < .001$), with the accuracy improving from an average A′ score of 0.62 in pretest to 0.82 in posttest. There was no effect of the counterbalancing factor Lists.

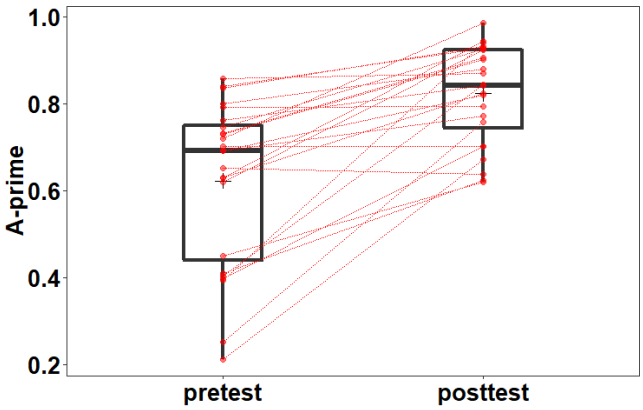


Figure 2. Boxplots of A′ scores in the lexical decision task in pretest and posttest. The red dots represent individual participants; the lines link each participant's performance in both sessions. The black cross marks indicate mean A′ scores in each session.

Discussion

The present study examined if phonetic training can enhance the recognition of words that contain a difficult non-native sound. We tested French learners with intermediate proficiency in English on both their prelexical perception and their lexical processing of stimuli containing /h/. This sound does not exist in French, and French listeners tend to confuse it with silence (Mah et al., 2016). The participants underwent eight sessions of High-Variability Phonetic training, and were tested in pretest and posttest by means of an identification and a lexical decision task.

We found that participants improved in both tasks in posttest compared to pretest. For the identification task, we also observed generalization to new items. The results for this task are in accordance with results from previous studies that used HVPT. Concerning the lexical decision task, this is the first piece of evidence that HVPT can improve not only prelexical but also lexical processing. As mentioned in the

introduction, successful word recognition depends on the correct decoding of the speech signal and the matching of this percept to the phonological representation stored in long-term memory (Pisoni & Luce, 1987). If listeners have difficulty with at least one of those aspects, then word recognition might be less effective. Evidence that this is the case is shown by the fact that in the lexical decision task during pretest, the test items involving the difficult sound /h/ yielded higher error rates than the control items. Note that performance on control items was very good in both pre- and posttest (mean A′ score 0.94). As the test and control items were matched in frequency, this indicates that the difficulty participants encountered with the test items was caused by the presence of /h/ and not by a lack of English vocabulary. Importantly, this difficulty was clearly reduced after training, as in posttest participants made less errors on the test items with /h/ than in pretest, while their performance did not change on control items.

Our findings have both theoretical and practical implications. From a theoretical point of view, they shed light on the relationship between prelexical and lexical processing in L2 learning. It is generally agreed upon that speech processing involves several stages, ranging from auditory processing, phonetic and phonological analysis, to word recognition and lexical access (Pisoni & Luce, 1987). In a study on Dutch L2 learners' processing of the English /æ/-/ɛ/ contrast, Díaz et al. (2012), found that the performance gap between native and non-native listeners increases as the tasks have greater lexical involvement. This is likely due to the fact that different perceptual tasks tap into different processing levels, thus requiring different skills and involving different amounts of cognitive load. Our finding that improvement in prelexical perception is paralleled by an improvement in lexical processing suggests a bottom-up sequential order in learning. Although at a specific time point in learning the proficiency in prelexical perception might be ahead of that in lexical processing, a rapid improvement in the former might give rise to change in the latter. This is in accordance with the Automatic Selective Perception model (Strange, 2011), which proposes that L2 phonological processing is less automatic and therefore requires more attentional resources than phonological processing in L1. Consequently, while the performance of learners might be good on relatively simple prelexical tasks, where they can exclusively focus their attention on crucial phonetic cues, the same performance level might not be obtained in tasks requiring the processing of more complex stimuli and attention to other information, such as word meaning. According to this model, the processing of simple tasks becomes more automatic and nativelike as proficiency grows. Thus, in our study, training possibly rendered the prelexical processing more efficient, thus allowing participants to allocate more cognitive resources to the lexical level of processing.

A similar finding on the benefit of phonetic training for higher processing levels was reported in a study on the perception of L2 speech in noise (Lengeris & Hazan, 2010). Adverse listening conditions such as a high signal-to-noise ratios (SNRs) have been shown to involve increased cognitive load and to have greater negative effects for speech perception in non-native than in native listeners (for a review, see Lecumberri et al., 2010). In this study, it was shown that HVPT in quiet improves the perception of a difficult L2 sound in noise.

On the practical side, the current findings could have implications for language teaching. The above-mentioned aspects of speech processing – lexical perception and perception of speech in noise – are inherent elements of "real life" language processing. The fact that they can be improved by relatively short HVPT is encouraging. Moreover, our training was administered online and not in a well-controlled laboratory setting; it can thus easily complement traditional language teaching methodologies. Finally, we note that participants of our study reported that being trained on real words was very motivating, as they had the occasion not only to enhance their perception but to learn new words as well.

To conclude, we showed that even short online HVPT can improve both prelexical and lexical processing of a difficult L2 sound. Future research should test if these improvements are retained in the long term. Furthermore, although we observed significant improvements, only some participants were at ceiling in posttest. Thus, further studies should look at the effect of training length on learning outcomes. This would help us understand if there is an upper limit of improvement in lexical processing that training can induce.

Acknowledgments

This research was supported by a Ph.D. fellowship from Ecole normale supérieure to G. A. Melnik and by grants from the Agence Nationale de la Recherche (Grant Nos. ANR-17-EURE-0017 and ANR-17-CE28-0007-01).

References

Brysbaert, M., & New, B. (2009). Moving beyond Kucera and Francis: a critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977-90.

Carlet, A. & Cebrian, J. (2014). Training Catalan speakers to identify L2 consonants and vowels: A short-term high variability training study. *Proceedings of the 7th International Symposium on the Acquisition of Second Language Speech* (pp. 85-98). Concordia University Working Papers in Applied Linguistics.

Cooper, A., & Wang, Y. (2011). The influence of tonal awareness and musical experience on tone word learning. *Proceedings of the 17th International Congress of Phonetic Sciences* (pp. 512-515). Hong Kong, SAR, China.

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a web browser. *Behavior Research Methods*, *47*(1), 1-12.

Díaz, B., Mitterer, H., Broersma, M., Sebastián-Gallés, N. (2012). Individual differences in late bilinguals' L2 phonological processes: From acoustic-phonetic analysis

to lexical access. *Learning and Individual Differences*, *22*(6), 680-689.

Homa, D., & Cultice, J. (1984). Role of feedback, category size, and stimulus distortion on the acquisition and utilization of ill-defined categories. *Journal of Experimental Psychology*, *10*, 83-94.

Huensch, A., & Tremblay, A. (2015). Effects of perceptual phonetic training on the perception and production of second language syllable structure. *Journal of Phonetics*, *52*, 105–120.

Ingvalson, E. M., Barr, A. M., & Wong, P. C. (2013). Poorer Phonetic Perceivers Show Greater Benefit in Phonetic-Phonological Speech Learning. *Journal of Speech Language and Hearing Research, 56*(3), 1045.

Iverson, P., Pinet, M., Evans. B. G. (2012). Auditory training for experienced and inexperienced second-language learners: Native French speakers learning English vowels. *Applied Psycholinguistics*, *33*(1), 145-160.

Kim, Y. H. & Hazan, V. (2010). Individual variability in the perceptual learning of L2 speech sounds and its cognitive correlates. *Proceedings of the Sixth International Symposium on the Acquisition of Second Language Speech* (pp. 251-256). Poznań, Poland.

Lecumberri, M. L. G., Cooke, M., Cutler, A. (2010). Non-native speech perception in adverse conditions: a review. *Speech Communication*, *52*(11-12), 864-886.

Lee, A. H. & Lyster, R. (2016). Effects of Different Types of Corrective Feedback on Receptive Skills in a Second Language: A Speech Perception Training Study. *Language Learning*, *66*(4), 809-833.

Lengeris, A. & Hazan, V. (2010). The effect of native vowel processing ability and frequency discrimination acuity on the phonetic training of English vowels for native speakers of Greek. *Journal of the Acoustical Society of America*, *128*(6), 3757-3768.

Lively, S. E., Pisoni, D. B., Yamada, R. A., Tokhura, Y., Yamada, T. (1994). Training Japanese listeners to identify English /r/ and /l/. III. Long-term retention of new phonetic categories. *Journal of the Acoustical Society of America, 96*(4), 2076-2087.

Logan, J.S., Lively, S.E., Pisoni, D.B., (1991). Training Japanese listeners to identify English /r/ and /l/: a first report. *Journal of the Acoustical Society of America*, *89*, 874-886.

Mah, J., Goad, H., Steinhauer, K. (2016). Using event-related brain potentials to assess perceptibility: The case of French speakers and English [h]. *Frontiers in Psychology*, *7*, 1-14.

Melnik, G. A., Peperkamp, S. (2019). Perceptual deletion and asymmetric lexical access in second language learners. *Journal of the Acoustical Society of America, 145*(1), EL13-EL18.

Okuno, T. & Hardison, D. M. (2016). Perception-production link in L2 Japanese vowel duration: Training with technology. *Language Learning & Technology*, *20*(2), 61-80.

Pallier, C., Colomé, A., Sebastián-Gallés, N. (2001). The Influence of Native-Language Phonology on Lexical Access: Exemplar-Based Versus Abstract Lexical Entries. *Psychological Science*, *12*(6), 445-449.

Piske, T., Mackay, I. R., & Flege, J. E. (2001). Factors affecting degree of foreign accent in an L2: A review. *Journal of Phonetics, 29*(2), 191-215.

Pisoni, D. & Luce, P. (1987). Acoustic-phonetic representations in word recognition. *Cognition*, *25*(1-2), 21-52.

Sakai, M. & Moorman, C. (2018). Can perception training improve the production of second language phonemes? A meta-analytic review of 25 years of perception training research. *Applied Psycholinguistics*, *39*(1), 187-224.

Sebastián-Gallés, N. (2005). Cross-language speech perception. In D. B. Pisoni & R. E. Remez (Eds.), *The Handbook of Speech Perception*. Oxford: Blackwell.

Shinohara, Y., Iverson, P. (2018). High variability identification and discrimination training for Japanese speakers learning English /r/–/l/. *Journal of Phonetics*, *66*, 242-251.

Strange, W. (2011). Automatic selective perception (ASP) of first and second language speech: A working model. *Journal of Phonetics*, *39*, 456-466.

Wang, Y., Jongman, A., Sereno, J. A. (2003). Acoustic and perceptual evaluation of Mandarin tone productions before and after perceptual training. *Journal of the Acoustical Society of America*, *113*, 1033-1043.

Wang, Y., Spence, M., Jongman, A., Sereno, J. (1999). Training American listeners to perceive Mandarin tones. *Journal of the Acoustical Society of America, 106*, 3649-3658.

Weber, A., Cutler, A. (2004). Lexical competition in non-native spoken-word recognition. *Journal of Memory and Language*, *50*(1), 1-25.

Yamada, J. (1991). The discrimination learning of the liquids /r/ and /l/ by Japanese speakers. *Journal of Psycholinguistic Research, 20*(1), 31-46.

Explanatory Considerations Guide Pursuit

Patricia Mirabile (patricia.mirabile@sorbonne-universite.fr)
Sciences Normes Décision, Sorbonne Université
Paris, France

Tania Lombrozo (lombrozo@princeton.edu)
Department of Psychology, Princeton University
Princeton, NJ, USA

Abstract

Evidence is typically consistent with more than one hypothesis. How do we decide which hypothesis to pursue (e.g., to subject to further consideration and testing)? Research has shown that *explanatory considerations* play an important role in learning and inference: we tend to seek and favor hypotheses that offer good explanations for the evidence we invoke them to explain. Here we report three studies testing the proposal that explanatory considerations similarly inform decisions concerning pursuit. We find that ratings of explanatory goodness predict pursuit (though to a lesser extent than they predict belief), and that these effects hold after adjusting for subjective probability. These findings contribute to a growing body of work suggesting an important role for explanatory considerations in shaping inquiry.

Keywords: explanation; pursuit; abduction; active learning

From belief to pursuit

"Faced with tracks in the snow of a certain peculiar shape," writes (Lipton, 2003), "I infer that a person on snowshoes has recently passed this way." This form of inference, familiar from both science and everyday life, is known as *inference to the best explanation* (IBE). Harman (1965) explains that in drawing this inference to an explanatory hypothesis, one infers, from the premise that a given hypothesis would provide a better explanation for the evidence than would any other hypothesis, to the conclusion that the given hypothesis is true (Harman, 1965).

Recent work in the cognitive science of explanation has confirmed and helped characterize the role of IBE in human cognition: both children and adults tend to prefer some explanations over others, and these preferences affect which hypotheses they favor (Lombrozo, 2016; for factors that might affect these preferences, see Colombo, Bucher, & Sprenger, 2017). For example, in Douven and Mirabile (2018), participants read about two possible explanations for six realistic events: a target "best" explanation (that had on average received higher quality ratings in a previous experiment) and an alternative explanation (that had received lower ratings). One group of participants was asked to rate the explanatory quality of both explanations (how well each explained the event) and a second group of participants rated the probability of each explanation, additionally indicating whether they accepted the target explanation as the true one. They found that the mean goodness ratings of the first group were better predictors of the acceptance rate of the target explanation by the second group than the probability ratings of that same group.

Findings like these suggest that explanatory considerations play an important role in guiding belief – indeed, in some cases a stronger role than that played by probability (see also Douven & Schupbach, 2015). But they also raise an important puzzle that has been a perennial challenge for advocates of IBE: why treat explanatory considerations as a good guide to what is true? After all, the world may not be simple, elegant, or otherwise conform to a good explanation. This challenge is especially acute when explanatory considerations diverge from probabilistic considerations (see van Fraassen, 1989).

One possibility is that the practice of favoring hypotheses that offer better explanations is a good epistemic policy in the sense that it has positive epistemic consequences, even if it doesnt *directly* result in an inference to a hypothesis that is more likely to be true. Along these lines, Wilkenfeld and Lombrozo (2015) introduce the idea of "Explaining for the Best Inference," whereby the practice of explaining (and of seeking *good* explanations) might improve our epistemic standing through a suite of downstream cognitive effects. Indeed, seeking and evaluating explanations facilitates the discovery of subtle patterns (e.g., Williams & Lombrozo, 2010; Walker & Lombrozo, 2017), even when the generated explanations are inaccurate (Walker, Lombrozo, Legare, & Gopnik, 2014). Explaining also encourages processes such as comparison (Edwards, Williams, Gentner, & Lombrozo, 2019), abstraction (Walker & Lombrozo, 2017), and metacognitive calibration (Rozenblit & Keil, 2002), which can be beneficial even if the agent fails to make an inference to a true explanation.

In the current paper, we turn our attention to the idea of IBE as an effective epistemic policy that could guide learners over time. Rather than focusing exclusively on the role of explanatory considerations in making an inference to (or evaluating the probability of) a given hypothesis at a given time, we consider whether and how explanatory considerations affect the decision to *pursue* one hypothesis over another – that is, to subject a hypothesis to further consideration or testing.

Pursuing explanations

Pursuing hypotheses is an important part of any search for explanations: doctors order medical tests before establishing a diagnosis, detectives interrogate suspects and verify alibis, and scientists collect evidence to assess their theories. Decisions

about pursuit are especially critical in science: not only must we justify to academic peers, funding agencies, and sometimes the general public why a given hypothesis is worthy of pursuit, but this very investigation can also serve as a "criterion of demarcation" between scientific and non-scientific endeavors. According to Popper (2005), a hallmark of science is the generation of theories that can be submitted to a method of critical testing: scientific theories should make predictions that can be falsified by empirical evidence.

However, both in principle and due to time and resource limitations, we are unable to investigate all hypotheses, even all *good* hypotheses: we must instead decide which hypotheses are worth pursuing, and which hypotheses are worth pursuing first. Nyrup (2015) argues that the justification of pursuit is the most legitimate use of IBE – more legitimate even that the justification of belief. His core idea is that a hypothesis that offers a good explanation has higher "epistemic value" if true than its salient competitors, and that this justifies giving it priority when deciding which hypotheses to pursue first.

Formal analyses additionally support the idea that a policy of favoring better explanations could pay off downstream, even if it does not lead to an accurate inference right away. Specifically, Kelly (2007) introduces a formal notion of simplicity, and contends that simple hypotheses should be preferred because adopting simple rather than more complex hypotheses will reduce to a minimum the number of necessary reversals of opinion before arriving at the true hypothesis, and therefore allow us to converge to the truth more quickly. Douven (2016) shows that under certain conditions, artificial agents using update rules that favor better explanations (defined according to a particular measure of "explanatory power") converge faster on the truth than artificial agents with probabilistic (Bayesian) update rules. Evaluating the goodness of an explanation might therefore be a key consideration when deciding whether to pursue it.

In the present research, we report three studies designed to address the following four questions. First (Q1), are people more likely to pursue one hypothesis over another to the extent it offers a good explanation for the data? Second (Q2), is this evaluation partially comparative, such that the explanatory goodness of alternatives will also matter, with a given hypothesis more likely to be pursued to the extent its alternative offers a poor explanation? Third (Q3), does explanatory goodness have an effect on pursuit that is not reducible to the effects of subjective probability on pursuit? Based on the findings from Douven and Mirabile (2018) concerning belief, we expect positive answers to these questions. However, we also expect pursuit and belief to diverge, given their differential costs (in terms of both requisite resources, and the consequences of getting things right vs. wrong). This prompts our final question (Q4): Does explanatory goodness differentially affect pursuit versus belief?

Study 1

In Study 1, we address Q1 - Q4 using materials adapted from Douven and Mirabile (2018). In a within-subjects design, participants were shown six vignettes that each described a disruptive event. They were presented with two possible hypotheses that might explain the event, and asked to rate the goodness and probability of each hypothesis. They also indicated which hypothesis they would recommend investigating first ("pursuit"), and which hypothesis they were more inclined to believe ("belief"). This allowed us to examine the link between perceived explanatory goodness and pursuit, as well as its relationship to probability and belief.

Method

Participants Participants were 72 adults recruited from Amazon Mechanical Turk (33 female, 39 male, ages 20-69, M = 35). Participation was restricted to MTurk workers with unique IP addresses in the United States who had completed at least 1000 HITs with a minimum approval of 99%. An additional 35 participants completed the study, but were excluded from analyses for failing one or more attention checks (described below).

Materials Six vignettes were lightly adapted from the stimuli used by Douven and Mirabile (2018). In these vignettes, experts (scientists, detectives, doctors) are attempting to explain a disruptive event (e.g., the flooding of a village, a murder, or a patient's symptoms), and they have generated two possible explanatory hypotheses, where these hypotheses are independent and are not jointly exhaustive. For instance, in one vignette, participants read about a womans murder, where one hypothesis is that the murder was committed by her jealous husband, and another hypothesis is that the murder was committed by a coworker trying to prevent her from sharing incriminating evidence. Based on the ratings of explanatory goodness provided by participants in Experiment 1 of Douven and Mirabile (2018), one of the hypotheses was classified as offering what we expected to be perceived as the *best explanation*, and the other as offering the *second best explanation*. These designations were used in analyses, but were not presented to participants.

Procedure Each participant received all six vignettes, with the order of the two hypotheses in each vignette randomized across participants. The study consisted of three phases: goodness and probability ratings, belief and pursuit decisions, and distraction questions, which doubled as attention checks. The distraction questions always appeared between the other two phases, which appeared first or last (randomized across participants).

In the goodness and probability ratings phase, participants received all six vignettes, and for each rated the two corresponding hypotheses on *explanatory goodness* and *probability*, with order randomized across participants. For explanatory goodness, participants were asked: "How good do you think each of these hypotheses is as an explanation for why *[the*

event occurred]?", with the corresponding event specified in the stimuli participants saw. Responses were collected on a continuous scale from 0 to 100, where 0 meant that an explanation was very bad, 50 meant that an explanation was neither good nor bad, and 100 meant that an explanation was very good. For probability, participants were asked: "How likely do you think each of these hypotheses is?" Responses were collected on a continuous scale from 0 to 100, where 0 meant that the hypothesis had 0% probability of being true, 50 meant that the hypothesis was equally likely to be true or not true, and 100 meant that the hypothesis had 100% probability of being true. We also included an attention check in which participants were instructed to select zero on the two continuous scales.

In the pursuit and belief decisions phase, participants received all six vignettes, and for each rated the two corresponding hypotheses for pursuit and belief, with order randomized across participants. For the pursuit decisions, participants were told: "The [experts] only have enough resources to investigate and test *one of the two* hypotheses before deciding on an explanation. They could also decide to save their resources and not investigate or test either of the two hypotheses. What do you think they should do?" Participants could select either hypothesis or indicate that they didn't think either of the hypotheses should be investigated. For the belief judgment, participants were asked: "Which of the two hypotheses are you more inclined to believe is the true explanation of why *[the event occurred]*?", (again, the corresponding event was specified in the stimuli participants saw). Participants could select either hypothesis or indicate that they were not inclined to believe either of the hypotheses.

The distraction phase consisted of two questions that doubled as attention checks. Participants read a list of words and, depending on a randomly assigned condition, copied into a text box the first word from that list that referred to an animal, a fruit, or a season. Participants also counted the number of animals in a picture.

After completing these three phases of the study, participants provided demographic information.

Results & Discussion

To examine whether and how explanatory considerations affect pursuit (Q1 and Q2), we fit a logistic binomial mixed-effects model (Q1/Q2 model) predicting participants probability of deciding to pursue the (antecedently defined) best explanation, as opposed to the second best explanation or neither explanation. Our choice of model and dependent variable allowed us to parallel the analyses in Douven and Mirabile (2018), where acceptance of the target "best" explanation was used as a dichotomous dependent variable. Explanatory goodness ratings for the best explanation and for the second best explanation were both centered on 50 and included as fixed effects. Vignette was included as a group-level random effect.

This model found a positive coefficient for the goodness of the best explanation ($p < .001$), with a 5.2% increase in the odds of choosing to pursue an explanation for each one-point increase in goodness. It also found a negative coefficient for the goodness of the second best explanation ($p < .001$), with a 5.2% *decrease* in the odds of choosing to pursue the best explanation for each one-point increase in the goodness of the second best explanation. These results provide a positive answer to Q1: explanatory considerations did predict pursuit. They also provide an answer to Q2: while the goodness of the better hypothesis mattered, the goodness of the alternative mattered as well.

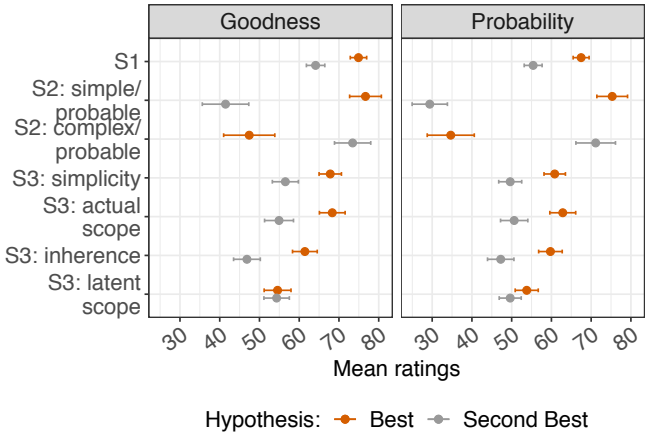


Figure 1: Mean ratings of goodness and of subjective probability for the two hypotheses for Study 1, Study 2 (Simple/Probable and Complex/Probable conditions) and Study 3 (Simplicity, Actual Scope, Latent Scope and Inherence conditions). Error bars represent 95% CI.

We next considered whether there were effects of explanatory considerations on pursuit that were not reducible to the effects of subjective probability on pursuit (Q3). To this end, we fit a logistic binomial mixed-effects model (Q3 model) predicting participants probability of deciding to pursue the best explanation, but in addition to the predictors included above, we also included a fixed effect for the probability assigned to the best explanation, and a fixed effect for the probability assigned to the second best explanation. Vignette was also included as a group-level random effect. There was a positive coefficient for the best explanation ($p < .001$), with a 4.8% increase in the odds of choosing to pursue the best explanation for each one-point increase in probability. There was also a negative coefficient for the second best explanation ($p < .001$), with a 4.9% *decrease* in the odds of choosing to pursue the best explanation for each one-point increase in the probability of its alternative. However, in this model, goodness ratings were not significant predictors. This suggests a negative answer to Q3: there was not evidence of effects of explanatory goodness on pursuit that were not reducible to the effects of subjective probability on pursuit. This result is potentially surprising in light of the findings from Douven and Mirabile (2018), which used essentially the same materials, but could be because goodness and probability were

collected within-subjects and highly correlated: 0.73 for the best explanation, and 0.78 for the second best explanation. One aim of Study 2 is to more successfully tease apart goodness and probability ratings.

Finally, we evaluated whether explanatory goodness differentially predicted pursuit vs. belief (Q4). We fit a logistic regression mixed-effects model (Q4 model) predicting the probability of selecting the best explanation, with goodness rating for the best explanation, goodness rating for the second best explanation, and judgement type (belief vs. pursuit) as fixed effects, as well as interactions between judgement type and each goodness rating. We also included vignette as a group-level random effect. This model found that goodness ratings had a significant effect when predicting pursuit, and that this effect was significantly larger when predicting belief. A one-point increase in the goodness of the best explanation increased the odds of deciding to pursue by 5.1% ($p < .001$), and of deciding to believe by 11.5% ($p = 0.005$). On the other hand, a one-point increase in the goodness of the competing explanation *decreased* the odds of deciding to pursue by 5.0% ($p < .001$), and of deciding to believe by 7.3% ($p < .01$). Explanatory goodness thus had significant and differential effects on pursuit vs. belief, with the impact on belief larger than that on pursuit.

Study 2

Study 2 had two primary aims. First, we sought to revisit Q1-Q4 with materials that induced a weaker correlation between goodness and probability. Second, we sought to vary explanatory quality along a recognizable and objective dimension for which people's explanatory preferences have already been experimentally established: simplicity, defined as the number of unexplained causes invoked in an explanation (Pacer & Lombrozo, 2017). As in Study 1, participants were shown a vignette describing an unusual event with two possible explanatory hypotheses. Each hypothesis was either simple or complex, and described as either probable or improbable. By introducing simple/improbable and complex/probable hypotheses, we hoped to drive apart ratings of goodness and probability.

Method

Participants Participants in Study 2 were 135 adults recruited through Amazon Mechanical Turk as in Study 1 (56 female, 79 male, ages 19-72, M = 37). An additional 25 participants completed the study but were excluded from analyses for failing one or more attention checks.

Materials Two vignettes were created following the same structure as the stimuli used in Study 1. In these vignettes, scientists seek to explain an unusual event (either a change in the reproductive pattern of squirrels, or low crop yields in a given county), and they have generated two possibles hypotheses. One of these hypotheses was simple in the sense that it appealed to a single cause (exposure to one toxin, contamination by one pest), and the other was more complex in that it required the conjunction of two independent causes (two toxins, two pests). In addition, one of the hypotheses was described as being "quite probable" based on the data available to the scientists, and the other hypothesis was described as being "quite improbable."

Procedure The study had a between-subject design (2 vignettes x 2 probability conditions). Each participant received one vignette, with the order of the two presented hypotheses randomized across participants. Participants were randomly assigned to one of two probability conditions. In the simple/probable condition, the simple hypothesis was described as probable and the complex hypothesis as improbable. In the complex/probable condition, this pairing was reversed. In the main part of the study, participants responded to the same questions as in Study 1. They also responded to an attention check and completed one of the distraction tasks from Study 1 midway through the study.

Results & Discussion

First, we verified that Study 2 successfully reduced the high correlations between perceived goodness and probability observed in Study 1. We found correlations of 0.60 and 0.65 between goodness and probability in the simple/probable condition for the simple and complex explanations, respectively, and correlations of 0.58 and 0.76 in the complex/probable condition for the simple and complex explanations respectively. While these correlations remained strong, they were more modest than those in Study 1.

We next conducted the same analyses as those described in Study 1. To examine how explanatory considerations affect pursuit, we fit the Q1/Q2 model, but did not include vignette as a group-level random effect[1]. This model found a positive coefficient for the goodness of the simple explanation ($p < .001$), with a 9.8% increase in the odds of choosing to pursue an explanation for each one-point increase in goodness. It also found a negative coefficient for the goodness of the complex explanation ($p < .001$), with a 8.1% *decrease* in the odds of choosing to pursue the simple explanation for each one-point increase in the goodness of the complex explanation. These results again provide a positive answer to Q1: explanatory considerations did predict pursuit. They also provide an answer to Q2: while the goodness of the better hypothesis mattered, the goodness of the alternative mattered as well.

We next analyzed whether there were effects of explanatory considerations on pursuit that were not reducible to the effects of subjective probability on pursuit (Q3). There was a positive coefficient for the goodness of the simple explanation ($p < .001$), with a 7.9% increase in the odds of choosing to pursue the simple explanation for each one-point increase in goodness. There was also a negative coefficient for the goodness of the

[1] All analyses in Study 2 were first fit using mixed-effects models, with vignette as a group-level random effect. However, the regression analyses indicated a singular fit, so we fit the models again excluding the group-level random effect to ensure that the estimates were stable. Estimated coefficients in the fixed-effects and mixed-effects models were identical.

complex explanation ($p < .035$), with a 5.0% *decrease* in the odds of choosing to pursue the simple explanation for each one-point increase in the goodness of its alternative. However, in this model, subjective probability ratings were not significant predictors. This points to a positive answer to Q3: we found effects of explanatory goodness on pursuit that were not reducible to the effects of subjective probability. These findings could differ from those of Study 1 because goodness and probability were not as highly correlated as in Study 1, or because explanatory goodness was manipulated in the form of simplicity.

Finally, we evaluated whether explanatory goodness differentially predicted pursuit vs. belief by fitting the Q4 model. This model found that goodness ratings had a significant effect when predicting pursuit judgements, and that this effect was not significantly different when predicting belief. A one-point increase in the goodness of the simple explanation increased the odds of a participant deciding to pursue by 9.8% ($p < .001$), and a one-point increase in the goodness of the competing explanation *decreased* the odds of deciding to pursue by 8.1% ($p < .001$). Unlike Study 1, this suggests a negative answer to Q4.

Study 3

Studies 1-2 provided consistent answers to Q1 and Q2: participants were more likely to pursue one hypothesis over another to the extent they judged that hypothesis a good explanation, and its alternative a poor explanation. However, the answers to Q3 and Q4 were more variable across studies. In Study 3, we sought to revisit Q1-Q4 using a larger sample and more varied experimental materials.

Specifically, we varied explanatory quality along four dimensions suggested by prior research to elicit reliable patterns of preferences in people's judgements. The first dimension was simplicity, defined in terms of the number of unexplained causes invoked in each explanation (e.g, explaining an illness with one toxin or the conjunction of two toxins). The second dimension was actual scope, defined as the number of observed effects explained (e.g., explaining all aspects of how a space shuttle had deviated from its trajectory or only some of them). The third dimension was latent scope, defined as the number of *unverified* effects predicted (e.g., one hypothesis predicts that prior to the volcano's irruption, the magma should have been relatively cool and the second predicts that a wider ranger of magma temperatures was possible–however, data on magma temperature prior to the irruption is not available). The fourth dimension was inherence, defined as an appeal to inherent/internal features versus extrinsic features (e.g., a flower's ability to wick off water is explained either by properties of its petals or by properties of the soil where it grows). Prior work has shown that with materials like those used here, people favor explanations that are simpler (Pacer & Lombrozo, 2017), broad in actual scope (Williams & Lombrozo, 2010), narrow in latent scope (Khemlani, Sussman, & Oppenheimer, 2010), and inherent (Cimpian & Salomon, 2014). While simplicity and actual scope are often defended as explanatory virtues, latent scope and inherence are typically assumed to reflect unwarranted biases. The procedure, materials, data collection plan, main predictions, and analyses for Study 3 were preregistered on the Open Science Framework platform prior to data collection and are available at `https://osf.io/6b58k/`.

Method

Participants Participants in Study 3 were 875 adults recruited from Amazon Mechanical Turk as in Studies 1-2 (446 female, 424 male, 2 non-binary/other and 2 who preferred not to respond, ages 18-87, M = 40). Following our preregistration, 1000 participants completed the study, with exclusions (N=125) based on failure to pass one or more attention check(s).

Materials Twelve vignettes were created following the same structure as the stimuli in Studies 1-2. In these vignettes, scientists generate two possible hypotheses to explain an unusual event. The two hypotheses in each vignette differed along a single dimension (simplicity, actual scope, latent scope, or inherence), with three vignettes targeting each dimension. The simplicity vignettes were similar to those in Study 2. In the actual scope vignettes, the best hypothesis explained all aspects of the explanandum, and the second best hypothesis explained only a subset. In the latent scope vignettes, the best hypothesis accounted for the *explanandum* without making unverified predictions, while the second best generated a prediction that it was not possible to verify. In the inherence vignettes, modified from Horne and Khemlani (2018), the best hypothesis invoked an inherent feature of the *explanandum*, and the second best invoked an extrinsic feature.

Procedure Each participant received one vignette, with the order of the two hypotheses randomized across participants. Aside from the fact that this study had a between-subjects design (4 dimensions of explanatory quality x 3 vignettes), the procedure was identical to that of Study 1.

Results & Discussion

To address Q1-Q4, we followed the analyses described in Studies 1-2[2]. We first fit the Q1/Q2 model. This model found a positive coefficient for the goodness of the best explanation ($p < .001$), with a 5.6% increase in the odds of choosing to pursue an explanation for each one-point increase in goodness.

[2]In our preregistered analyses, we planned to fit logistic binomial mixed-effects models that included as predictors the goodness and probability ratings of the best explanation and differences in goodness/probability ratings between the best and the second best explanation. However, upon analyzing the data, we found a high (>0.89) correlation between differences in goodness ratings and differences in probability ratings. Because high correlations between predictors in linear regressions can make the estimated coefficients unreliable, we replaced the difference predictors by the goodness and probability ratings of the second best explanation. The correlation between goodness and probability ratings ranged from >0.8 for the actual and latent scope virtues, to >0.82 for simplicity and >0.94 for inherence.

It also found a negative coefficient for the goodness of the second best explanation ($p < .001$), with a 4.1% *decrease* in the odds of choosing to pursue the best explanation for each one-point increase in the goodness of the second best explanation. These results again provide a positive answer to Q1: explanatory considerations did predict pursuit. They also provide an answer to Q2: while the goodness of the better hypothesis mattered, the goodness of alternatives mattered as well.

We next analyzed whether there were effects of explanatory considerations on pursuit that were not reducible to the effects of subjective probability by fitting the Q3 model. There was a positive coefficient for the goodness of the best explanation ($p < .001$), with a 2.7% increase in the odds of choosing to pursue the best explanation for each one-point increase in goodness, and a positive coefficient for the subjective probability of the best explanation ($p < .001$), with a 3.4% increase in the odds of choosing to pursue the best explanation for each one-point increase in probability. There was also a negative coefficient for the goodness of the second best explanation ($p = .0026$), with a 1.8% *decrease* in the odds of choosing to pursue the best explanation for each one-point increase in the goodness of its alternative, and a negative coefficient for the subjective probability of the second best explanation ($p < .001$), with a 2.9% *decrease* in the odds of choosing to pursue the best explanation for each one-point increase in the goodness of its alternative. Like study 2, this provided a positive answer to Q3: the effect of explanatory considerations held even when the effect of probability judgements was also taken into account.

Next, we evaluated whether explanatory goodness differentially predicted pursuit vs. belief by fitting the Q4 model. This model found that goodness ratings had a significant effect when predicting pursuit judgements, and that this effect was significantly larger when predicting belief: a one-point increase in the goodness of the best explanation increased the odds of a participant deciding to pursue by 5.6% ($p < .001$) and of deciding to believe by 13.6% ($p < .001$). On the other hand, a one-point increase in the goodness of the competing explanation *decreased* the odds of deciding to pursue by 4.1% ($p < .001$), and of deciding to believe by 10.4% ($p < .01$). As in Study 1, explanatory goodness thus had significant and differential effects on pursuit vs. belief, with a larger impact on belief.

Finally, we repeated the three analyses just described for each of the four sets of vignettes corresponding to each virtue. These analyses revealed the same patterns of answers to Q1-Q2 as in the full data set, but some departures for Q3 and Q4. Specifically, we found a negative answer to Q3 for simplicity and actual scope, and a negative answer to Q4 for latent scope.

General Discussion

Across three studies, we find evidence that explanatory considerations affect pursuit: participants were more disposed to pursue a hypothesis to the extent it offered a good explanation, and to the extent its competitor offered a poor explanation. In Studies 2-3, we also found that the effect of explanatory considerations on pursuit were not reducible to the effects of subjective probability on pursuit. Finally, in Studies 1 and 3, we found that explanatory goodness had a larger impact on belief than on pursuit. Discrepancies across the three studies could have resulted from the high correlations between ratings of goodness and of subjective probability, but it is notable that Study 3–which had the largest sample–found positive answers to all four of our guiding questions. However, it is important to note that these results raise open questions about the direction of a potential causal relationship between explanatory considerations and pursuit, and indeed they do not rule out the possibility that pursuit decisions might be causing judgements of explanatory goodness, rather than the reverse.

Why might explanatory considerations affect pursuit? As suggested in the introduction, pursuing good explanations could facilitate learning (Lombrozo, 2016), have higher expected epistemic value (Nyrup, 2015), or provide a more efficient route to the truth (Kelly, 2007; Douven & Schupbach, 2015). The pursuit of good explanations might therefore improve our overall epistemic standing (Wilkenfeld & Lombrozo, 2015), even if the true hypothesis is not the most explanatory. If these ideas are correct, they provide a justification for IBE that side-steps many of the traditional worries concerning its application to belief.

Interestingly, however, the impact of explanatory goodness on pursuit was smaller than that on belief. In a context where unjustified pursuit is more costly (given limited resources) than erroneous belief, participants might be more reluctant to recommend pursuit on the basis of explanatory considerations alone. Moreover, decisions to pursue might be more sensitive to pragmatic considerations that compete with explanatory goodness, or to the goal of reducing uncertainty by maximizing expected information gain.

Several limitations are worth noting. First, participants reasoned about relatively abstract and unfamiliar material. Second, participants did not pursue explanations themselves (e.g., through further consideration or evidence gathering). Future work could investigate decisions to pursue (vs. believe) with more realistic materials, and testing a richer set of pursuit-relevant behaviors. It would also be fruitful to investigate whether the role of explanatory considerations changes as a function of the relevant consideration (as we began to explore in Study 3), in different environments (e.g., with different cost structures), and as a function of the learners goals (e.g., to achieve truth vs. avoid error).

More ambitiously, future research should investigate how pursuit and belief are integrated into a broader model of truth-seeking behavior that involves explanation generation, pursuit, the collection of evidence, hypothesis revision, and ultimately belief. Our findings suggest that explanatory considerations affect this process at two important stages, belief and pursuit, but leave open how they shape everyday and scientific inquiry more broadly.

Acknowledgments

We wish to thank the Concepts and Cognition Lab and Igor Douven for valuable input, as well as the SND research center and Sorbonne Université for funding Patricia Mirabile's stay as a visiting scholar at Princeton University.

References

Cimpian, A., & Salomon, E. (2014, May). The inherence heuristic: An intuitive means of making sense of the world, and a potential precursor to psychological essentialism. *Behavioral and Brain Sciences*, *37*(5), 461–480.

Colombo, M., Bucher, L., & Sprenger, J. (2017, Sep). Determinants of judgments of explanatory power: Credibility, generality, and statistical relevance. *Frontiers in Psychology*, *8*.

Douven, I. (2016). Inference to the best explanation: What is it? and why should we care. In K. McCain & T. Poston (Eds.), *Best explanations: New essays on inference to the best explanation.* Oxford: Oxford University Press.

Douven, I., & Mirabile, P. (2018, Nov). Best, second-best, and good-enough explanations: How they matter to reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(11), 1792–1813.

Douven, I., & Schupbach, J. N. (2015). The role of explanatory considerations in updating. *Cognition*, *142*, 299–311.

Edwards, B. J., Williams, J. J., Gentner, D., & Lombrozo, T. (2019). Explanation recruits comparison in a category-learning task. *Cognition*, *185*, 21–38.

Harman, G. H. (1965). The inference to the best explanation. *The Philosophical Review*, *74*(1), 88–95.

Horne, Z., & Khemlani, S. (2018, July). Conceptual constraints on generating explanations. In C. Kalish, M. Rau, J. Zhu, & T. Rogers (Eds.), *Cogsci 2018* (pp. 1815–1820).

Kelly, K. T. (2007). How simplicity helps you find the truth without pointing at it. *Induction, Algorithmic Learning Theory, and Philosophy*, 111–143.

Khemlani, S. S., Sussman, A. B., & Oppenheimer, D. M. (2010, Nov). Harry potter and the sorcerer's scope: latent scope biases in explanatory reasoning. *Memory & Cognition*, *39*(3), 527–535.

Lipton, P. (2003). *Inference to the best explanation*. Routledge.

Lombrozo, T. (2016). Explanatory preferences shape learning and inference. *TICS*, *20*(10), 748–759.

Nyrup, R. (2015, Dec). How explanatory reasoning justifies pursuit: A peircean view of ibe. *Philosophy of Science*, *82*(5), 749–760.

Pacer, M., & Lombrozo, T. (2017). Ockham's razor cuts to the root: Simplicity in causal explanation. *Journal of Experimental Psychology: General*, *146*(12), 1761.

Popper, K. (2005). *The logic of scientific discovery*. Routledge.

Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive science*, *26*(5), 521–562.

van Fraassen, B. C. (1989). *Laws and symmetry*. Oxford University Press.

Walker, C. M., & Lombrozo, T. (2017). Explaining the moral of the story. *Cognition*, *167*, 266–281. (Moral Learning)

Walker, C. M., Lombrozo, T., Legare, C. H., & Gopnik, A. (2014). Explaining prompts children to privilege inductively rich properties. *Cognition*, *133*(2), 343–357.

Wilkenfeld, D. A., & Lombrozo, T. (2015, Oct). Inference to the best explanation (ibe) versus explaining for the best inference (ebi). *Science & Education*, *24*(9-10), 1059–1077.

Williams, J. J., & Lombrozo, T. (2010). The role of explanation in discovery and generalization: Evidence from category learning. *Cognitive Science*, *34*(5), 776–806.

What information shapes and shifts people's attitudes about capital punishment?

Olivia Miske (omiske@asu.edu)
School of Social and Behavioral Sciences, Arizona State University
Phoenix, AZ, USA

N. J. Schweitzer (njs@asu.edu)
School of Social and Behavioral Sciences, Arizona State University
Phoenix, AZ, USA

Zachary Horne (Zachary.Horne@asu.edu)
School of Social and Behavioral Sciences, Arizona State University
Phoenix, AZ, USA

Abstract

Although most Americans support capital punishment, many people have misconceptions about its efficacy and administration (e.g., that capital punishment deters crime). Can correcting people's inaccurate attitudes change their support for the death penalty? If not, are there other strategies that might shift people's attitudes about the death penalty? Some research suggests that statistical information can correct misconceptions about polarizing topics. Still, statistics might be irrelevant for some people because they may support capital punishment for purely retributive reasons, suggesting other argumentative strategies may be more effective. In Studies 1 and 2, we examined what attitudes shape endorsement of capital punishment and compared how two different interventions shifted these attitudes. Altogether, our findings suggest that attitudes about capital punishment are based on more than just retributive motives, and that correcting misconceptions related to its administration reduces support for capital punishment.

Keywords: capital punishment; coherence; open science

Introduction

In October 2018, Washington state became the 20th state to overturn capital punishment on the grounds that it is unconstitutional, stating that death sentences have been "imposed in an arbitrary and racially biased manner" (Johnson, 2018). Although capital punishment has come under scrutiny at the state-level, a recent poll indicated that 55% of adults in the United States still favor the death penalty for a person convicted of murder (Jones, 2017). However, many people who support the use of capital punishment have misconceptions about its efficacy and administration. For instance, many people believe that capital punishment is an effective deterrent against violent crime, that innocent people are not sentenced with the death penalty, and that it is administered in a fair and unbiased manner (see Manski & Pepper, 2013; DPIC, 2018; Baldus, Woodworth, Zuckerman, & Weiner, 1998). The Death Penalty Information Center (DPIC) and the Innocence Project have publicly impugned these assumptions to better educate the public by releasing informational brochures and short educational videos. Given that people have misinformed attitudes about issues integral to the administration and efficacy of capital punishment, can correcting their misconceptions shift their support for the death penalty, and if not, are there other argumentative tactics that could be used to shift people's attitudes about the death penalty?

Ideally, we could affect attitude change by simply providing people with accurate statistical information—on the basis of this information, people may still support the death penalty, but it would not be based on misconceptions about its efficacy and administration. On the other hand, there is some reason to think that statistics-interventions like these may not be effective at changing people's moral attitudes. In a now classic study, Lord, Ross, and Lepper (1979) found that when people were presented with statistical evidence about capital punishment—especially when that evidence was "mixed" (providing some evidence consistent with and inconsistent with the death penalty)—this led to belief polarization. People who were strongly opposed to or strongly in favor of the death penalty attended to the information that confirmed their position and ignored the information that was inconsistent with their position. These results have led many researchers to conclude that providing statistical information is not an effective tactic for correcting people's misconceptions (e.g., Thaler & Sunstein, 2008; Janis & King, 1954; Gawronski & Bodenhausen, 2006).

More recently, however, some research suggests that statistical information, especially when carefully presented (e.g., using visual aids) can correct misconceptions about polarizing topics like climate change and anti-vaccine attitudes (see Lewandowsky, Ecker, Seifert, Schwarz, & Cook, 2012; Horne, Powell, Hummel, & Holyoak, 2015). These results are some cause for optimism, but they do not establish exactly what interventions are most effective at changing people's attitudes about the death penalty. For example, statistics might be irrelevant to some people's support of capital punishment. For moral reasons alone, people may support the use of capital punishment, not because they believe it deters crime or is more cost effective, but because they think criminals should get what they deserve and that it is the morally right thing to do. Consequently, providing statistics about deterrence, cost, wrongful convictions, and other relevant issues may do little to alter these attitudes about the death penalty.

Current research suggests that even if attitudes about the

death penalty are driven entirely by the desire for retribution, or because it is perceived as the right thing to do, it may still be possible to shift their moral attitudes. A recent line of work has examined how moral attitudes change when related attitudes are manipulated (e.g., Horne, Powell, & Hummel, 2015; Holyoak & Powell, 2016). In the law, coherence is an important theoretical virtue (e.g., Dancy, 1984). Moral theories that are incoherent are generally considered "nonstarters" and inconsistencies in influential moral theories are often the topics of entire books (e.g., Lyons, 1965; Gewirth, 1978; Rawls, 1980; Sen & Williams, 1982; MacIntyre, 2007). These considerations do not appear to only be the concern of academics. For example, Horne and colleagues (2015) found that when people are presented with a situation (e.g., a moral dilemma) that elicits a judgment inconsistent with a general moral principle (e.g., utilitarianism), tension arises due to an internal conflict among participants' attitudes about the dilemma and the general moral principle. This tension induces belief revision because people desire to restore coherence in their network of attitudes (e.g., Festinger, 1962; Holyoak & Powell, 2016). We call this a coherence-based intervention.

Altogether, people may support the use of capital punishment for reasons like deterrence and the cost of execution, which may suggest that presenting accurate statistical information could change people's minds (e.g., Cochran & Chamlin, 2005). On the other hand, the death penalty is a moral issue importantly linked to attitudes about just desserts—this may suggest that coherence-based interventions would be more persuasive than raw statistical information.

In the present studies, we sought to answer two questions: First, what kinds of interventions—statistics or coherence-based—will shift people's support for capital punishment? Second, what might this tell us about what attitudes are most malleable and most central to people's endorsement of the death penalty? In Study 1, we compared the efficacy of two interventions by investigating how statistics versus coherence-based interventions changed people's attitudes about capital punishment. However, because we have reason to think that the effectiveness of these distinct arguments likely depends on the reasons people have for supporting capital punishment, and because of our results in Study 1, we sought to investigate what other related attitudes might be predictive of support for the death penalty in Study 2.

Study 1

Method

Preregistration The data collection plan, predictions, and analysis scripts for our study were preregistered through the Open Science Framework. Data, analyses, and supplemental materials are available at https://osf.io/ek4fh/.

Participants We recruited 504 participants through Amazon Mechanical Turk. Our sample size was determined by conducting a power analysis to detect a Cohen's d of .25 with 80% power. We used an optional stopping procedure by computing a Bayes Factor on the parameter estimating the effect of condition (that is, the parameter of interest). Specifically, we determined that we would continue data collection until the Bayes Factor (BF_{10}) was greater than 100 or less than .01, at which point we would stop data collection (Rouder, 2014). After excluding participants who failed attention checks, 405 participants remained for our final sample (46% female, M_{age} = 36 years old). Each participant was compensated $0.70 for completing the study.

Procedure We developed statistics and coherence-based interventions aimed at countering three common attitudes people have for supporting the death penalty. These attitudes were: (1) People who commit serious crimes, such as murder, deserve to be put to death (retribution), (2) The death penalty discourages people from committing crime (deterrence), and (3) The death penalty is cheaper than life-imprisonment (cost). Participants were randomly assigned to either the statistics or coherence-based intervention, in which they saw either three statistical arguments or three coherence-based arguments in a between-subjects design.

The study proceeded as follows: Participants first were asked to rate how much they agree with three pretest statements (one statement for each commonly-held belief about capital punishment). Then participants received either the statistics or coherence-based intervention, which consisted of statistical or coherence-based arguments designed to counter attitudes about deterrence, cost, and retribution as motivations for supporting the death penalty. After reading these arguments, participants completed the post-intervention measure which captured participants' attitudes about retribution, deterrence, and cost, and their overall attitudes towards capital punishment. Participants then were asked to provide general demographic information. These measures and interventions are described in more detail below. Complete materials for this study can be found in the Supplementary Online Materials (SOM).

Pretest Measure Participants were asked to rate their agreement with three pretest statements about the death penalty. Each of these statements measured three common motivations for supporting the death penalty on a 7-point Likert scale (1 = Strongly disagree, 7 = Strongly agree). For example, the item that measured attitudes about deterrence was, "The death penalty makes criminals think twice before committing murder." These statements were developed based on our post-intervention capital punishment measure.

Interventions As noted, participants were randomly assigned to either the statistics or coherence-based intervention, in which they read three statistical arguments or three coherence-based arguments against each belief for supporting the death penalty.

The statistics intervention was composed of brief summaries of empirical research taken from the Death Penalty Information Center (DPIC). This research contradicts

common misconceptions about capital punishment. For instance, the statistical argument for deterrence summarized information about criminology experts' and researchers' conclusions regarding the efficacy of capital punishment as a deterrent. Excerpts from this argument stated that "88% of these experts rejected the notion that the death penalty acts as a deterrent to murder", and that "studies claiming that the death penalty has a deterrent effect on murder rates are fundamentally flawed."

The coherence-based intervention consisted of brief persuasive arguments adapted from widely-cited law papers. In these papers, authors attempt to persuade readers through coherence-based arguments why the typical reasons taken to support the death penalty are inconsistent with other attitudes they otherwise strongly hold. Therefore, these arguments did not provide information about a belief being objectively false, but rather demonstrated ways in which the reason underlying a belief was incoherent with their other attitudes. For example, the coherence-based argument for cost demonstrated that determining whether someone should live or die based off of financial considerations is not a practice people generally condone and thus, it should not be considered a good reason in the case of capital punishment either. For complete intervention materials, see Table S2 and S3 in the SOM.

Posttest Measure The posttest items measured participants' attitudes about retribution, deterrence, and cost, along with their attitudes towards the death penalty in general. Participants were asked how much they agreed with 13 statements, adapted from the Death Penalty Attitudes Questionnaire (O'Neil, Patry, & Penrod, 2004). An example of a general death penalty item (general items labeled G1 - G4 in Figure 2) was, "I think the death penalty is necessary." Other items concerned attitudes about retribution (labeled R1 - R4 in Figure 2), deterrence (labeled D1 - D3), and the cost of the death penalty (labeled C1 - C2). For example, one item was "Society has a right to get revenge when murder has been committed."

Results

We tested whether statistical or coherence-based arguments would be more effective at changing people's attitudes towards capital punishment. Further, we aimed to understand how the effectiveness of each intervention varied as a function of the specific attitudes, or reasons people have for supporting capital punishment. In order to test this, we performed Bayesian ordinal mixed-effects modeling, predicting post-intervention attitudes towards the death penalty on the basis of condition (1 = statistics, 0 = coherence-based), and participants' pretest attitudes, which we modeled as a monotonic effect. This model treated both participants and scale items as group-level effects, allowing for heterogeneity in the intercept for each participant and question. The model is specified in the syntax of `brms` (Bürkner, 2018):

```
Model 1 <- Response ~ Condition +
mo(PreRetribution) + mo(PreDeterrence) +
mo(PreCost) + (1|Question) + (1|Subject)
```

Bayesian analyses formulate model parameters as probability distributions wherein the posterior distribution for a parameter θ is computed via the prior and the likelihood of θ. To model the joint probability distribution of responses, we specify regularizing priors over the possible effects each parameter could have on the response variable. Model 1 priors are shown below:

$$\beta_{Intercept[1]} \sim \mathcal{N}(2.19, 1)$$
$$\beta_{Intercept[2]} \sim \mathcal{N}(2.94, 1)$$
$$\beta_{Intercept[3]} \sim \mathcal{N}(3.17, 1)$$
$$\beta_{Intercept[4]} \sim \mathcal{N}(3.47, 1)$$
$$\beta_{Intercept[5]} \sim \mathcal{N}(3.89, 1)$$
$$\beta_{Intercept[6]} \sim \mathcal{N}(4.59, 1)$$
$$\beta_{\forall Pretest\ Beliefs} \sim \mathcal{N}(6, 1)$$
$$\beta_{Condition} \sim \mathcal{N}(0, 1)$$
$$\text{Group-level effects} \sim t(3, 0, 10)$$

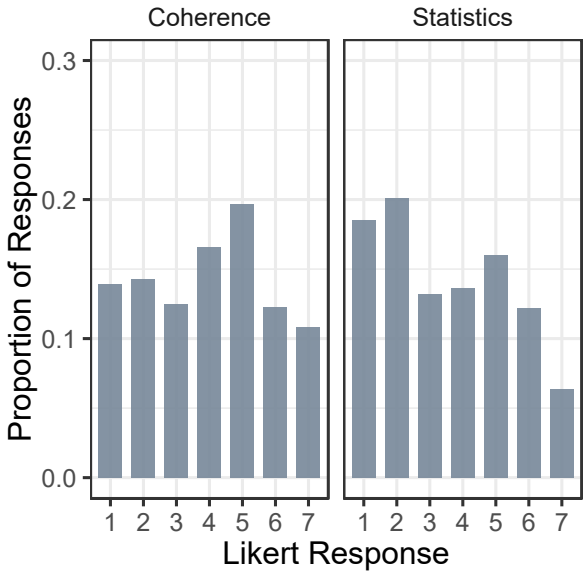


Figure 1: A histogram of the proportion of responses at a given Likert scale point (1 = Strongly disagree, 7 = Strongly agree) in the Coherence and Statistics conditions in Study 1. The figure indicates that participants were less likely to agree with pro-death penalty statements in the Statistics condition than the Coherence condition.

This analysis revealed that the statistics intervention reduced overall support for the death penalty relative to the coherence-based intervention, $b = -0.58$, 95% CI [−0.80, −0.35], $BF_{10} > 100$ (see Figure 1). Models interacting pretest beliefs with condition did not account for additional variance.

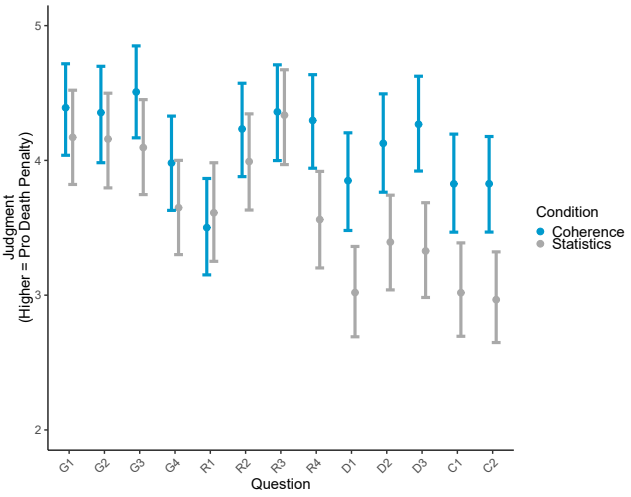


Figure 2: Death penalty attitudes for each scale item in the Coherence and Statistics conditions. Both conditions presented arguments against pro-death penalty beliefs. Error bars represent 95% Credible Intervals. Participants in the Statistical condition were less likely to endorse the death penalty than participants in the Coherence condition, but this effect varied as function of the question under consideration.

Next we investigated how condition interacted with the question to examine whether the statistics intervention affected some reasons for supporting the death penalty more than others. This model is specified below:

```
Model 2 <- Response ~  Condition*Question
+ mo(Deterrence) + mo(Cost) +
mo(Retribution) + (1|Subject)
```

Model 2 Priors:

$$\beta_{Intercept[1]} \sim \mathcal{N}(2.19, 1)$$
$$\beta_{Intercept[2]} \sim \mathcal{N}(2.94, 1)$$
$$\beta_{Intercept[3]} \sim \mathcal{N}(3.17, 1)$$
$$\beta_{Intercept[4]} \sim \mathcal{N}(3.47, 1)$$
$$\beta_{Intercept[5]} \sim \mathcal{N}(3.89, 1)$$
$$\beta_{Intercept[6]} \sim \mathcal{N}(4.59, 1)$$
$$\beta_{\forall_{Pretest\ Beliefs}} \sim \mathcal{N}(6, 1)$$
$$\beta_{Condition} \sim \mathcal{N}(0, 1)$$
$$\beta_{\forall_{Questions}} \sim \mathcal{N}(0, 3)$$
$$\beta_{\forall_{Condition \times Question\ Interactions}} \sim \mathcal{N}(0, 1)$$
$$\text{Group-level effects} \sim t(3, 0, 10)$$

The analysis interacting question with condition indicated that the statistics intervention was more effective at changing people's general death penalty attitudes (i.e., G1 – G4), people's attitudes about the efficacy of capital punishment at deterring crime (D1 – D3), and the cost of capital punishment (C1 – C2) compared to retributive attitudes (R1 – R4), $BF_{10} > 100$, (see Figure 2). This result is consistent with the intuition that for some attitudes, perhaps those that are *particularly* moral in nature, statistical information is irrelevant. When predicting only general attitudes towards the death penalty on

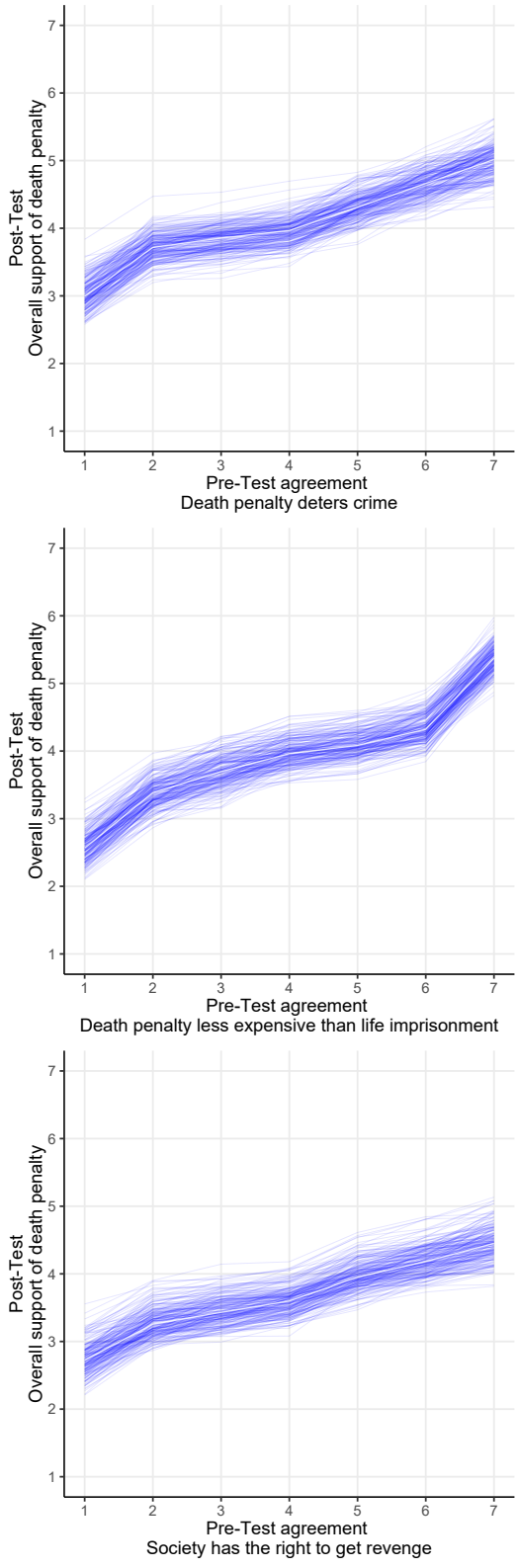


Figure 3: Spaghetti plot of post-test death penalty attitudes predicted by pre-test deterrence (top), cost (middle), and retribution (bottom) attitudes, which were treated as monotonic effects. Each regression line represents a draw from the posterior distribution.

the basis of condition, statistics were still more effective than coherence-based interventions, b = –0.49, 95% CI [–0.88, –0.08].

The nature of our design prohibited us from testing how effective each intervention was at changing every posttest death penalty attitudes because we could not compute a difference score for every item. However, three items were repeated across pretest and posttest. Therefore, we examined whether the coherence-based condition affected these items and how these change scores compared to the statistics condition. This analysis revealed that the statistics and the coherence conditions both decreased posttest endorsement for the three pro-death penalty beliefs, M = –0.68, 95% CI [–0.78, –0.57] and M = 0.25, 95% CI [–0.35, –0.16], respectively. However, for these items alone the statistics intervention was still more effective than the coherence-based intervention.

We followed up on these analyses by conducting a series of exploratory analyses examining how deterrence, cost, and retribution attitudes predicted overall posttest death penalty attitudes. This model regressed posttest death penalty attitudes on each pretest question, allowing us to measure the unique relationship each attitude accounts for in predicting posttest attitudes. Because of the ordinal nature of our predictors, we again treated each as a monotonic effect. These analyses indicated—to our surprise—that attitudes about the cost of the death penalty (b = 3.65, 95% CI [3.11, 4.18]) was more strongly related to people's death penalty attitudes than were beliefs about deterrence (b = 2.52, 95% CI [2.01, 3.03]) and the desire for retribution, b = 2.16, 95% CI [1.60, 2.73] (see Figure 3). Given that cost attitudes are most easily targeted by statistics interventions, and that these interventions proved more effective than a coherence-based intervention, this is further evidence that policy makers interested in shifting attitudes towards the death penalty might focus on the relevant statistics rather than moral imperatives.

Still, the results of Study 1 raised questions about what attitudes, beyond those that have been previously assumed to be relevant, are most strongly related to overall death penalty attitudes. Previous research assessing people's views about the death penalty have predominantly focused on people's retributive and utilitarian motives (i.e., people's desire for retribution and belief in the deterrent effect of capital punishment). Furthermore, some studies have used only a few items or a single dichotomous item to measure death penalty attitudes, even though public opinion polls and other research have shown that people's attitudes about this issue are complex and often dependent on the circumstances of the situation (e.g., Murray, 2003; Roberts & Stalans, 1997). Consequently, relatively simple measures such as these are unlikely to provide substantial insight into why people endorse the death penalty, and what beliefs and motivations underlie their attitudes. This is not to deny that attitudes about retribution and deterrence are central in shaping their attitudes about capital punishment. Rather, in Study 2, we aimed to understand what other understudied factors might also play a significant role in shaping people's attitudes towards capital punishment. For instance, people may not be familiar with the rate at which innocent people are sentenced to death, or they might not know that most other industrialized countries have abolished the death penalty. If these beliefs are related to support for capital punishment, and could also be changed more easily than beliefs about retribution, then researchers could develop more effective interventions using this information (Powell, Weismann, & Markman, 2018).

Study 2

In Study 2, we tested what attitudes are most strongly related to people's general support of the death penalty–what are the most relevant reasons people support capital punishment? We conducted an exploratory correlational study examining the relationship between previously-theorized attitudes (e.g., retribution and deterrence, Finckenauer, 1988; Carlsmith, Darley, & Robinson, 2002) and other understudied attitudes (e.g., the importance of wrongful convictions and perceptions of execution methods) that we hypothesized may be most strongly related to people's general death penalty attitudes.

Method

Preregistration Our sample size and study materials were preregistered through the Open Science Framework at https://osf.io/ek4fh/.

Participants We recruited 249 participants through Amazon Mechanical Turk. After excluding participants who failed attention checks, 184 participants remained for our final sample (45% female, M_{age} = 37 years old). Participants were paid $0.70 for participating in the study.

Procedure Participants were asked to rate how much they agreed with statements which composed 12 scales about capital punishment, the criminal justice system, and other related topics. These attitudes are described in more detail below. After answering these questions, participants provided demographic information.

Death Penalty Attitudes Measure We measured 11 attitudes (54 items total) that we hypothesized would be relevant to people's death penalty attitudes, many of which were suggested by previous studies but not included in most death penalty measures. We again measured attitudes about *retribution*, *deterrence*, and *cost*. The other attitudes we included were: (1) Providing rehabilitation programs for offenders is a good idea (*Rehabilitation*), (2) Innocent people are sometimes sentenced to death and this is a major concern with using the death penalty (*Innocence*), (3) People who are wrongfully convicted of serious crimes must have done something wrong to be in that situation (*Victim Blame*), (4) The death penalty is barbaric (*Barbarity*), (5) The United States has a great deal of crime (*Crime*), (6) America's execution methods are humane (*Humane*), (7) Other countries similar to America have the death penalty (*Common*), and (8)

Torture is acceptable in some cases (*Torture*). Our scales and items were adapted from the Death Penalty Attitudes Questionnaire (O'Neil et al., 2004), the Violence-Related Attitudes and Beliefs Scale (Brand & Anastasio, 2006), and a study by Jiang and colleagues (Jiang, Lambert, Wang, Saito, & Pilot, 2010). Participants rated how much they agreed with each statement on a 7-point Likert scale (1 = Strongly disagree, 7 = Strongly agree; Cronbach's α for all scales were > .70). For the complete list of materials and scales, see the SOM.

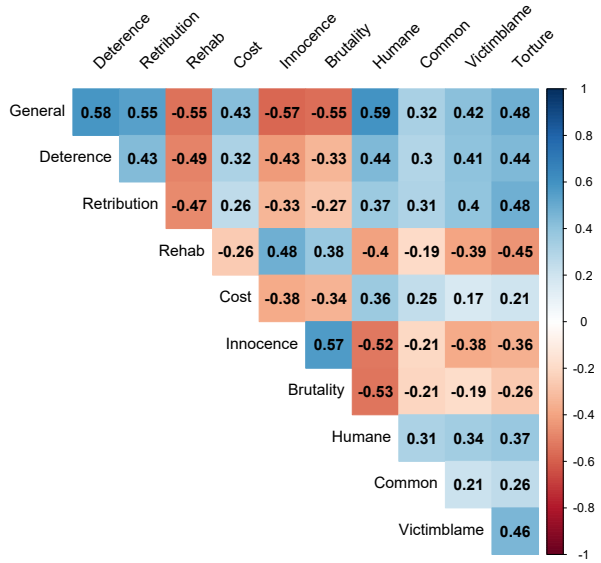


Figure 4: Kendall's tau correlation coefficients of death penalty attitudes in Study 2. Shades of blue indicate a positive correlation and shades of red indicate a negative correlation between two attitudes.

Results

We predicted that each of the 11 attitudes measured would be related to people's overall support for capital punishment, and as expected, all attitudes were correlated with participants' overall death penalty attitudes (see Figure 4). Deterrence, retribution, and the importance of innocence were among the most highly correlated attitudes with general endorsement of the death penalty. However, other attitudes exhibited surprisingly strong relationships with general support of the death penalty as well. For example, participants who endorsed the death penalty were also more likely to think that exonerated people were still nonetheless guilty or partially responsible for them being wrongfully convicted (Victim Blame; τ_b = .42). Strikingly, 28% of participants agreed, at least somewhat, with the idea that wrongfully convicted people on death row were responsible for their conviction (> 4 = Somewhat agree). Taken together, these results suggest that support for capital punishment may be more multidimensional than initially thought and provide further guidance for the development of interventions for correcting misconceptions about the administration of capital punishment.

General Discussion

Over half of the United States supports the use of the death penalty today yet are unaware of the statistics surrounding the deterrent effects and cost of the death penalty. Furthermore, there have been few systematic investigations of the attitudes, both proximal and remote, that may shape people's support for the death penalty. In Study 1, we examined how different types of interventions shift people's attitudes about the death penalty. We found that statistics interventions reduced support for the death penalty, and that these effects were largest for general death penalty attitudes, and attitudes about cost and deterrence. Furthermore, we found that statistics interventions were ineffective at changing attitudes motivated by retribution. Because retribution falls unambiguously within in the moral domain, people likely think statistics are irrelevant to the questions of whether criminals should get what they deserve. Study 1 also revealed that retribution is not the only relevant factor driving people's death penalty attitudes—beliefs about deterrence and cost were also strong predictors of overall endorsement of the death penalty. The results of Study 1 led us to examine what other attitudes, which have perhaps gone unexplored, may shape attitudes towards the death penalty (Powell et al., 2018). Study 2 revealed that many relatively "remote" attitudes were strongly correlated with endorsement of the death penalty. Of note, we observed a relationship between general death penalty attitudes and the belief that people wrongfully sentenced are to some degree responsible for their wrongful imprisonment. From an interventionist perspective, Study 2 also uncovered that many of the attitudes associated with support for the death penalty—for instance, beliefs about innocence and commonality—can be directly addressed by citing statistics. No moral imperative is required. Altogether, these findings highlight new avenues by which researchers can correct and shift people's attitudes about the death penalty.

References

Baldus, D., Woodworth, G., Zuckerman, D., & Weiner, N. (1998). Racial discrimination and the death penalty in the post-Furman era: An empirical and legal overview with recent findings from Philadelphia. *Cornell L. Rev.*, *83*, 1638.

Brand, P. A., & Anastasio, P. A. (2006). Violence-related attitudes and beliefs: Scale construction and psychometrics. *Journal of Interpersonal Violence*, *21*(7), 856–868.

Bürkner, P. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, *10*(1), 395–411.

Carlsmith, K., Darley, J., & Robinson, P. (2002). Why do we punish? Deterrence and just deserts as motives for punishment. *Journal of Personality and Social Psychology*, *83*(2), 284–299.

Cochran, J., & Chamlin, M. (2005). Can information change public opinion? Another test of the Marshall hypotheses. *Journal of Criminal Justice*, *33*(6), 573–584.

Dancy, J. (1984). On coherence theories of justification: Can an Empiricist be a Coherentist? *American Philosophical Quarterly*, *21*(4), 359–365.

Death Penalty Information Center. (2018, February 23). Facts about the death penalty [PDF file]. Retrieved from https://deathpenaltyinfo.org/

Festinger, L. (1962). *A theory of cognitive dissonance.* (Vol. 2). Stanford, CA: Stanford University Press.

Finckenauer, J. (1988). Public support for the death penalty: Retribution as just deserts or retribution as revenge? *Justice Quarterly*, *5*(1), 81–100.

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin*, *132*(5), 692–731.

Gewirth, A. (1978). *Reason and morality.* Chicago, IL: University of Chicago Press.

Holyoak, K., & Powell, D. (2016). Deontological coherence: A framework for commonsense moral reasoning. *Psychological Bulletin*, *142*(11), 1179–1203.

Horne, Z., Powell, D., & Hummel, J. (2015). A single counterexample leads to moral belief revision. *Cognitive Science*, *39*(8), 1950–1964.

Horne, Z., Powell, D., Hummel, J., & Holyoak, K. (2015). Countering antivaccination attitudes. *Proceedings of the National Academy of Sciences*, *112*(33), 10321–10324.

Janis, I., & King, B. (1954). The influence of role playing on opinion change. *The Journal of Abnormal and Social Psychology*, *49*(2), 211–218.

Jiang, S., Lambert, E. G., Wang, J., Saito, T., & Pilot, R. (2010). Death penalty views in China, Japan and the U.S.: An empirical comparison. *Journal of Criminal Justice*, *38*(5), 862–869.

Johnson, K. (2018, October 11). Washington state supreme court deems death penalty unconstitutional. *The New York Times.* Retrieved from https://www.nytimes.com/

Jones, J. M. (2017, October 26). U.S. death penalty support lowest since 1972. *Gallup Poll.* Retrieved from https://www.gallup.com/home.aspx

Lewandowsky, S., Ecker, U., Seifert, C., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, *13*(3), 106–131.

Lord, C., Ross, L., & Lepper, M. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, *37*(11), 2098–2109.

Lyons, D. (1965). *Forms and limits of Utilitarianism.* Oxford, England: Clarendon Press.

MacIntyre, A. (2007). *After virtue: A study in moral theory.* (3rd ed.). Notre Dame, IN: University of Notre Dame Press.

Manski, C., & Pepper, J. (2013). Deterrence and the death penalty: Partial identification analysis using repeated cross sections. *Journal of Quantitative Criminology*, *29*(1), 123–141.

O'Neil, K., Patry, M., & Penrod, S. (2004). Exploring the effects of attitudes toward the death penalty on capital sentencing verdicts. *Psychology, Public Policy, and Law*, *10*(4), 443–470.

Powell, D., Weisman, K., & Markman, E. M. (2018). Articulating lay theories through graphical models: A study of attitudes surrounding vaccination decisions. *Proceedings of the 40th Annual Conference of the Cognitive Science Society*, 906–911.

Rawls, J. (1980). Kantian Constructivism in moral theory. *The Journal of Philosophy*, *77*(9), 515–572.

Roberts, J., & Stalans, L. (1997). *Public opinion, crime, and criminal justice.* Boulder, CO: Westview Press.

Rouder, J. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, *21*(2), 301–308.

Sen, A., & Williams, B. (1982). *Utilitarianism and beyond.* Cambridge, England: Cambridge University Press.

Thaler, R., & Sunstein, C. (2008). *Nudge: Improving decisions about health, wealth, and happiness.* New Haven, CT: Yale University Press.

How much to purchase? - A cognitive adaptive decision making account

Percy K. Mistry (pkmistry@uci.edu)
Department of Cognitive Sciences, University of California Irvine

Abstract

Repeated purchase decisions often violate assumptions of standard economic or rational choice models, such as demonstrating asymmetric or unstable responses to changes in underlying policy, price, or tax variables. I propose a novel framework for how such decisions can be interpreted through the lens of a cognitive process model. This provides psychologically interpretable characterizations of individuals or population groups. It incorporates mental accounting, hedonic adaptation, confirmation bias, and the influence of perceived trust and fairness. It shows how sequential experiences and contextual aspects such as political affiliation, are mediated by this cognitive process to produce evolving consumption patterns. This novel approach can account for empirically observed violations of conventional choice models. The model is quantitatively fit to experimental data for individual purchase decisions and demonstrates improved descriptive, predictive, and inference capabilities. A proof-of-concept analysis using this model to account for real world consumption trends is also demonstrated.

Introduction

Decisions on what quantity (Q) of a particular item to purchase over time depend on individual preferences, expected benefits from purchasing the item, and associated costs (prices, taxes, etc.). Elasticity (ε) is a canonical economic concept that defines the influence of a unit change of an underlying independent variable (e.g. prices or taxes) on the purchase quantity. Typical choice models assume that elasticities are stable at a population level over relatively long periods of time (controlling for income effects, i.e. the influence of higher levels of income on purchasing power), and that elasticities are symmetric (i.e. respond equally to increases versus decreases). They thus assume that observed changes in purchase quantities (∂Q) in response to changes in underlying variables such as prices and taxes (∂p) can be used to estimate empirical elasticities which in turn can be used to accurately forecast future changes.

Standard choice model: Standard economic models of choice assume that decision makers select the optimal quantity (Q) to purchase by maximizing the net benefit stemming from the utility (U_Q) of owning Q units of an item, less the costs of purchase (pQ), where p is the unit price (equation 1).

$$Q = arg\ max_x \{U_x - px\} \tag{1}$$

Without loss of generalizability, the utility function in equation 2 is assumed for the rest of the paper. This is one of a standard set of utility functions used in behavioral and econometric literature (e.g. Chetty, Looney, and Kroft (2009)). Solving for Q using equations 1 and 2, then taking logs, defining $\varepsilon_p = 1/b$, $(\varepsilon_p > 0)$ and $A = -\varepsilon_p\ log(a)$, we obtain equation 3. This is in the form of a log-linear model with a corresponding difference equation 4, with log price elasticity ε_p (a standard economic representation). As per this model $log(Q)$ decreases at a rate of ε_p as $log(p)$ increases and vice versa. Once ε_p is empirically estimated, equation 4 can be used to make forecasts.

$$U_x = \left(\frac{ax^{1-b}}{1-b}\right) \tag{2}$$

$$log(Q) = A - \varepsilon_p\ log(p) \tag{3}$$

$$\partial\ log(Q) = -\ \varepsilon_p\ \partial\ log(p) \tag{4}$$

Evidence against conventional assumptions: There is strong evidence however that elasticities (including, but not limited to, price elasticities such as ε_p described above) may not be stable even in the short run (Hughes, Knittel, & Sperling, 2006; Goodwin, 1992), may not be symmetric (Villas-Boas, Berck, Stevens, & Moe-Lange, 2016; Gately, 1992), may show significant heterogeneity, even directionally, (Chetty, Friedman, Olsen, & Pistaferri, 2009; Ayyagari, Deb, Fletcher, Gallo, & Sindelar, 2009; Fletcher, Frisvold, & Tefft, 2015), and may be easily manipulated by extraneous factors. Whilst these violations are acknowledged, no theory provides a robust and quantitative account of how elasticities evolve over time.

Psychological characterization of dynamic elasticities: In this paper I propose that dynamic characteristics of elasticities can be explained by examining purchase decisions through the lens of a sequential cognitive process. Let $\bar{p}$ define a sequential history of the underlying variable such as prices or taxes, Ψ represent stable cognitive characteristics of an individual or a population (these are elaborated on in subsequent sections), and Δ represent contextual factors (e.g. the measure of political climate or affiliations). Then equation 4 can be replaced with equation 5. Here, $\bar{p}$ and Δ are observable, and cognitive characterization Ψ can be empirically estimated (similar to ε).

$$\partial\ log(Q) = f(\Psi, \bar{p}, \Delta, \partial\ log(p)) \tag{5}$$

Cognitive framework

In this section, I develop a novel cognitive process model to structurally define $f()$ in equation 5, for a sequence of repeat purchase decisions over time. This model is parameterized by psychologically interpretable characteristics Ψ, which interact with sequential history $\bar{p}$ and environmental context Δ to shape continuously evolving patterns of purchases and elasticities. This model is based on bringing together and quantitatively specifying some novel and some previously explored psychological conceptualizations, as elaborated below:

Mental Accounting: Transaction Utility

Thaler (1999, 2008) proposed that consumption quantity (Q) choices were driven by a process of mental accounting that considered a combination of acquisition utility (similar to that defined in equation 2) and transaction utility. Transaction utility reflects the "value" of the deal, typically evaluated against some expectation or reference point, and adds or detracts from the acquisition utility. For this paper, transaction utility is defined by equation 6[1], reflecting the difference between the price p and the expectation or reference price θ.

$$T(Q,p,\theta) = (\theta - p)Q \tag{6}$$

The transaction utility can be positive or negative depending on whether expectations were exceeded. Thus the optimal purchase quantity can now be given by equation 7. This replaces equation 1 of the standard model. The mental accounting theory proposes that the acquisition and transaction utilities are separately evaluated, and may be accorded different weights. Here, δ is a salience weight that emphasizes or reduces the effect of the transaction utility component.

$$Q = arg\ max_x \{[U_x - px] + \delta\,[T(x,p,\theta)]\} \tag{7}$$

Salience Weight: Rational, Hedonic, or Altruistic?

The salience weight δ characterizes the nature of decision making. A rational choice would imply $\delta = 0$, since the transaction utility is driven purely by whether or not internal expectations are exceeded, and should not play any role in objective decision making. Applications of the mental accounting framework typically assume $\delta > 0$. This implies that individuals act hedonically in self-interest to maximize the utility they derive from exceeding their internal expectations. In that sense, $\delta > 0$ implies a reference-point that reflects 'the maximum they should be charging me'. Any price lower than this is treated as a positive utility and vice versa. However, some consumption decisions may involve conflicting considerations, such as those of fairness (Xia & Monroe, 2010). For instance, the decision to purchase goods that damage the environment, the decision to evade taxes, or the decision to purchase mandatory health insurance, may result in a conflict between hedonic utility on one hand, and a moral obligation on the other. Such moral obligations can give rise to utilitarian or altruistic concerns (Greene, 2007, 2009) that are concerned with the fairness of policies and redistribution goals. An altruistic reference point may thus reflect 'the bare minimum I should be paying'. For choices involving such moral obligations, if people do indeed demonstrate altruistic concerns, the salience weight may be $\delta < 0$, for at least a non-trivial subset of the population. A price lower than the reference point would reduce transaction utility and vice versa. While this may seem counterintuitive, an example that makes this comprehensible is the case of purchasing goods that are not environmentally friendly. Paying a price higher than the expected reference point may act as a moral justification, and in fact increase the transaction utility and resulting demand by reducing the associated guilt.

Hedonic reference point adaptation

Transaction utilities may be evaluated positive or negatively against a reference point. However this reference point is not typically constant. I propose that the reference point evolves over time (n), motivated by principles of hedonic adaptation (Frederick & Loewenstein, 1999). At time point n the reference point (θ_n) moves closer to the recently experienced values under consideration (e.g. price p_{n-1}), modulated by a hedonic adaptation rate L_h, as shown in equation 8.

$$\theta_n = \theta_{n-1} + L_h(p_{n-1} - \theta_{n-1}) \tag{8}$$

Hedonic adaptation implies that this mechanism serves to increase satisfaction and reduce dissonance created by any large difference between actual prices and expected reference points. This serves to condition people towards recent levels of p. The hedonic adaptation rate L_h may vary by individual or population - higher values of L_h close to 1 imply smaller transaction utility and rational consumption behavior.

Confirmation Bias: Asymmetric adaptation

Confirmation bias, where people place asymmetrical weights on information that confirm rather than contradict their beliefs and actions has been shown to be pervasive over many cognitive processes (Nickerson, 1998; Jones & Sugden, 2001; Palminteri, Lefebvre, Kilford, & Blakemore, 2016). The rate of hedonic adaptation L_h is proposed to be asymmetric, and depends on whether the prospective movement of the reference point supports or inhibits current behavior. Let m reflect a bias that reduces the rate of adaptation ($0 \leq m \leq 1$) when adaptation would serve to inhibit current behavior. This bias is introduced in equation 9. Here, I is an indicator function, with $I = 1$ if ($p < \theta$ under hedonic salience weight $\delta > 0$), or if ($p > \theta$ under altruistic salience weight $\delta < 0$), and 0 otherwise. These situations reflect a inhibition of current behavior based on prospective adaptation, and hence manifest as a lower adaptation rate mL_h. Confirmation bias will thus manifest as a consumption bias, slowing down adaptation that inhibits consumption.

$$\theta_n = \theta_{n-1} + L_h(p_{n-1} - \theta_{n-1})\ (mI + (1-I)) \tag{9}$$

[1]Note that this assumes a comparison of expectation and realization of the price, however a transaction utility can similarly be expressed based on expectations versus realization for the utility U_x. The framework and model in this paper can be applied without any loss of generalizability to such utility based reference points as well.

Trust based adaptation

Additionally, the reference points are proposed to increase when there is a perception of fairness or trust in the counter-party, and drop otherwise. For example, when considering tax changes and related reference points for tax rates, the government is the counter-party. A reference point for taxes may increase when the government is trusted (e.g. its wealth redistribution goals are considered fair, when political affiliations are in power, and hence higher taxes are more acceptable) than when it is not. Similar considerations may be at work when it comes to prices for goods, and whether people trust a certain brand, or when a brand signals quality, etc. This perception of trust is coded as $\pi = 1$ (trust), $\pi = -1$ (distrust)., or $\pi = 0$ (agnostic). Rate of adaptation in response to these perceptions is governed by L_π, and captured in equation 10. Updating is assumed to occur at every time point n when there is a consumption decision or a change in underlying policy.

$$\theta_n = \theta_{n-1}\,(1+\pi_n L_\pi) + L_h(p_{n-1} - \theta_{n-1})\,(mI + (1-I)) \quad (10)$$

Combined Cognitive-Econometric Model

Here, we replace the standard model in equation 3 by using equations 6, 7, and 10.

Case 1: Purchases Quantity and Price Changes

Equation 7 can now be re-written under the cognitive framework as equation 11, and solved. The log linear demand equation 3 then changes to equation 12. Note that these equations contain the term θ_n given by equation 10[2]. The fully expanded version of equation 12 would thus include the parameters π, L_π, L_h, and m.

$$Q_n = arg\ max_x \left\{ \frac{ax^{1-b}}{1-b} - p_n x + \delta(\theta_n - p_n)x \right\} \quad (11)$$

$$log(Q_n) = A - \varepsilon_p\ log(p_n - \delta(\theta_n - p_n)) \quad (12)$$

Case 2: Purchase Quantity and Tax Changes Next, consider the case where the key variable of interest is how purchase demand may change in response to changes in tax rates t, with the reference point θ being a reference point for what is considered a fair tax rate. There is a lot of evidence to show that there are considerable differences between price and tax elasticity, even when they would have objectively identical impact on consumers (Chetty, Looney, & Kroft, 2009; Chetty, 2015). Following the logic in the previous section, but adding terms for an excise tax rate t that is applied as a percentage on the cost price, so that effective cost would be increased by a value pxt, we obtain equation 13. This considers a situation with constant price p and only changes in the tax rate t and hence a reference point for tax rates only.

$$log(Q_n) = A - \varepsilon_p\ log(p) - \varepsilon_p\ log\big(1 - \delta(\theta_n - t_n)\big) \quad (13)$$

[2]Mathematicaly, extremely high values of the reference point would result in infinite utility, inducing people to spend all resources and maximize the units of consumption. Such reference levels are however *psychologically implausible*, and a mathematical bound for psychological plausibility of θ can be derived in terms of p and δ, such that the $log()$ term in equation 12 never turns negative, implying utility never increases to ∞.

Model Simulation Results

Figure 1 illustrates how different assumptions about the confirmation bias (low or high, governed by m) and mode of processing (hedonic versus altruistic, governed by δ) give rise to systematic deviations from the standard choice model, under different price trend situations (see figure caption for more details). When the predictions from the cognitive model shown in figure 1 are used to infer back what the interpretation of such data would have been under the standard choice model, the resulting inferences reflect highly unstable and variable shifts in conventionally measured elasticity from sub-period to sub-period, as well as asymmetry between elasticity during increasing and decreasing price trends, just as has been reported in literature discussed in the introduction. Such apparent instability and asymmetry is readily explained and generated by stable cognitive characteristics.

Application to Experimental Data (price)

Data: I consider a published experimental dataset from the work reported in Sitzia and Zizzo (2012, 2015). 384 participants made a series of 20 sequential decisions on how many units of a particular lottery to buy. Participants were provided experimental units of currency, and could spend as much of it as they wanted on the lotteries. At the end, the unspent currency, as well as any winnings based on the lotteries were added and converted to real monetary payouts. The lottery remained fixed across all trials, but the purchase price per lottery was varied sequentially. Participants were split into 5 conditions as shown in the left panel in figure 2, where the stimulus (price) patterns for the 5 conditions are represented in different colors. Participants in each of the 5 conditions start with either extremely high (EH), high (H), moderate (M), low (L), or extremely low (EL) levels of prices for the first 10 trials which constitute the "shape" block. All the participants observe the same moderate price levels in the last 10 trials, the "compare" block. The right panel in figure 2 shows the average response (average units bought) for each condition. Key observations made by Sitzia and Zizzo (2012) were that participants with higher initial price purchase more units in the "compare" block than those that have observed a lower initial price. This, as well as the dynamics of many individual patterns of how participants switch purchasing behavior over trials represents a challenge for the standard economic model.

Modeling Results: A standard choice model, as described in the introduction, as well as the cognitive model based on hedonic and asymmetric adaptation (equations 9 and 12) is quantitatively fit to this data, using a Bayesian MCMC framework (JAGS, Plummer et al. (2003)). Since this is an experimental setup, the concept of trust based adaptation is not included in the model. A measure of descriptive fit is evaluated. Additionally, the models are separately tested by providing the first 15 trials for each individual to the models, and obtaining predictions for the last 5 trials. Model comparison using deviance information criteria (DIC, a combined measure of model fit and complexity) was significantly better

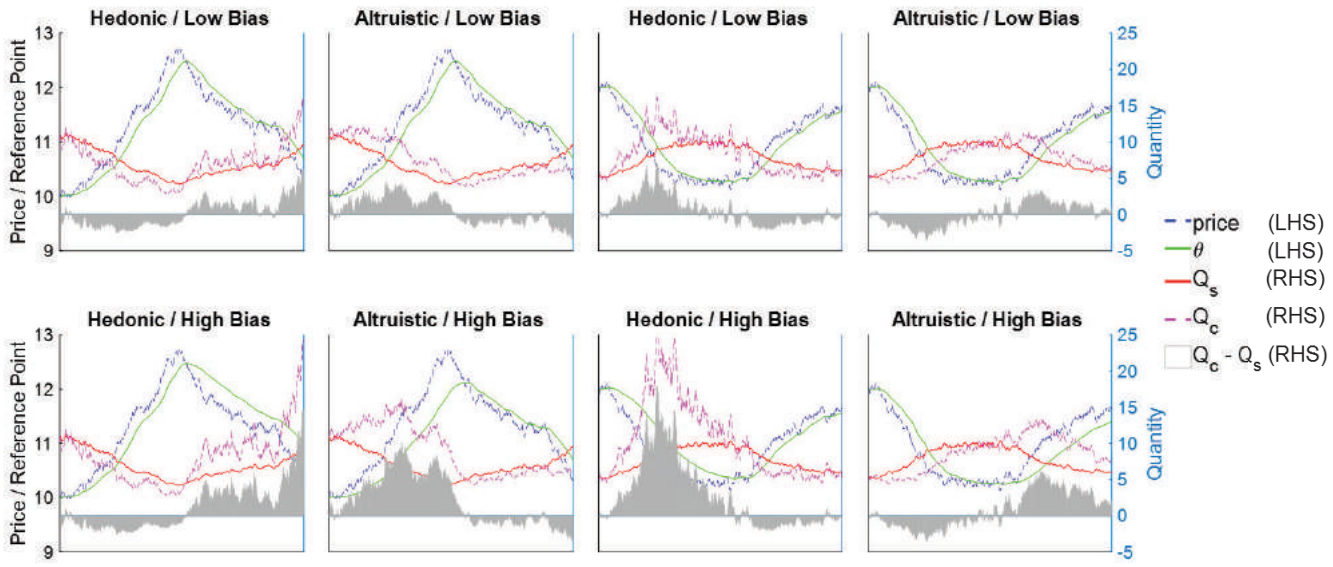


Figure 1: **Model Simulation:** Columns 1 and 2 show price trends that are first increasing and then decreasing. Columns 3 and 4 show price trends that are first decreasing and then increasing. Four parameter combinations reflecting Hedonic ($\delta > 0$) versus Altruistic ($\delta < 0$), and low confirmation bias (high values of m) versus low confirmation bias (high values of m) are compared under each scenario. The blue lines reflect the price changes. The x-axis reflects time, a hypothetical weekly data spread over a 10 year period. The red line gives the purchase quantity based on the standard choice model and typical assumptions about elasticity. The green line reflects the reference point based on the cognitive model assumptions. The pink line reflects the purchase quantity based on the cognitive model that assumes the same base elasticity as the standard model. The gray bars reflect differences between the cognitive model and the standard model quantities predicted. Price and reference points should be read of the left (LHS) axis and quantities off the right (RHS) axis. High bias parameterizations generally produce higher purchase quantity as expected. More interesting is the asymmetry produced by the cognitive model, which is typically seen in real world scenarios, which can be seen in the relative asymmetry between the gray bars in the first and second half of each simulation - reflecting asymmetries involved in responses to increasing versus decreasing prices.

(lower DIC is better) for the cognitive model (DIC = 21,276) compared to the standard model (DIC = 23,871). Figure 3 compares both the fit and prediction errors (RMSE) between the standard and cognitive models. It shows that for a huge majority of individuals, the cognitive model produces better descriptive fits (better for 86%) and better predictions on unseen data (better for 80%).

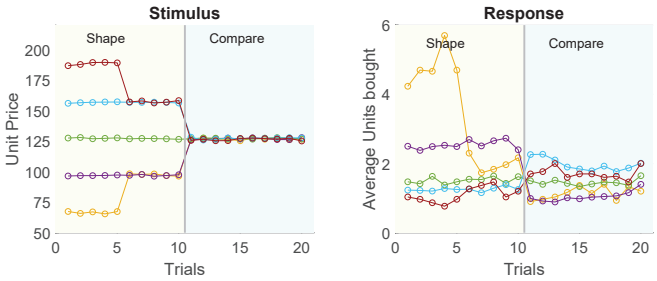


Figure 2: **Stimulus and Responses:** Experimental data from Sitzia and Zizzo (2015)

Illustration of how the model works: Figure 4 shows the latent model inferences about how the reference point evolves, and its relation to consumption patterns for 2 of the participants in the experimental tasks, to illustrate how the model is accounting for behavior. The figure shows the trial by trial stimulus price (red line), the response purchase quantity (black bars) and the latent reference point inferred by the model (green line). For the subject in the left panel, the prices are initially high and then fall. In the second half of the experiment, even though the price stays in the same range, the consumption levels falls as the difference between the latent reference point and the price narrows over time. For the subject in the right panel, the consumption remains almost constant from trials 6-17 even though the price increases after trial 10. This stability when the price is changing significantly is on account of the almost constant difference between the price and the evolving reference point. The standard model finds it difficult to explain these kind of behavioral patterns.

Inferences from the model parameters: Table 1 summarizes the parameter inferences for the cognitive model showing the mean, standard deviation, and the correlation between the parameters and purchase quantities in the "compare" block (trials 11 to 20, where the prices were identical for all 5 conditions). All participants demonstrated con-

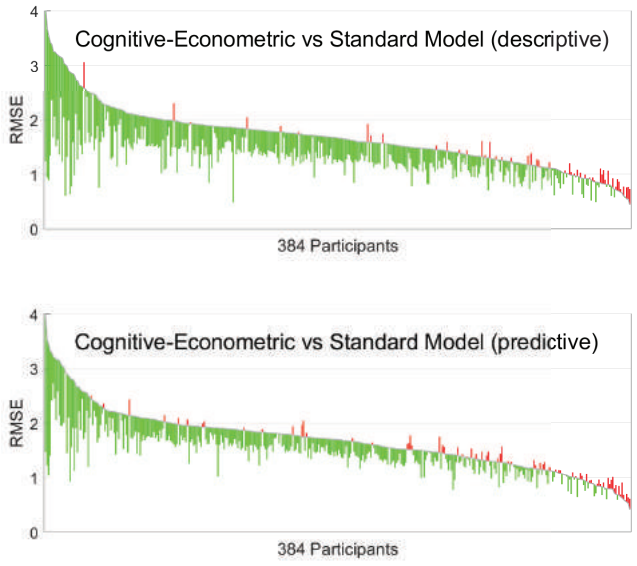


Figure 3: **Comparison of model fit errors:** Significantly better fit (upper panel) and predictions (lower panel) by the cognitive model. Each bar in the figure represents an individual, with the gray line showing the error from the standard model (participants sorted in order of reducing error based on the standard model). The green bars show an improvement (bar going downwards) on account of the cognitive model and red bars show deterioration (bars going upwards).

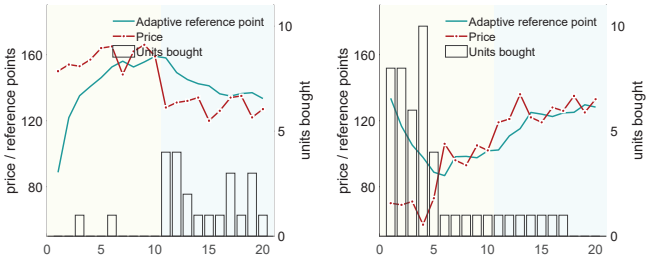


Figure 4: **Illustration of model mechanics:** Examples of the latent model inferences about the reference point trajectory for 2 participants. These behavioral patterns cannot be easily explained by the standard model.

sumption bias ($m < 1$, lower values of m indicate higher consumption bias). The mean value of m of 0.43 indicates that on average, people shifted their reference points twice as much when the price was higher than expectations, than when it was lower than expectations. As intuitively expected, m (lower values = higher consumption bias) is strongly correlated with consumption in the second half of the experiment ($r = -0.75, p < 0.00001$). Most participants show high salience weights δ on the transaction utility, but also individual differences, and higher salience weights are strongly correlated to higher consumption ($r = 0.68, p < 0.00001$). The rate of hedonic adaptation (L_h) did not show strong individual differences, but was consistently less than 1, indicating that adaptation was slower than rationally expected, leading to consistent deviations from the standard model.

Table 1: Key cognitive parameters (Ψ) capturing individual differences in the experimental task.

Characteristics	Mean	std	corr with $Q_{compare}$
m	0.43	0.15	-0.75
δ	1.59	0.71	0.68
L_h	0.48	0.10	0.27

Application to real world data (taxes)

Data: This section provides a brief proof-of-concept for applying this cognitive model to real world population level consumption behavior. Panel data from Chetty, Looney, and Kroft (2009) is used, that includes per capita consumption of beer by state in the US for a period of 34 years, along with the corresponding price and tax changes. As a proof of concept illustration, analysis for 3 states is provided below.

Modeling: A basic standard model[3], and the cognitive model based on equations 10 and 13, that is, including the trust based adaptation, are implemented within a Bayesian inferential framework. The models are fit by providing them with data about consumption changes for 20 years, and then checking model predictions (based on 1000 generated samples for each state for each year) about consumption changes for the last 13 years. The top panels of figure 5 show the changes in tax rates for 3 states over the 34 year period (note the different tax change profiles for the 3 states). The bottom panels show the distribution of prediction errors. The cognitive model produces significantly lower errors ($p < 0.05$ for comparison of error distributions for all 3 states).

Figure 6 shows the influence of the trust based adaptation on reference points (and hence eventually on consumption). The dotted lines show political party regime changes. The three states seem to show graded political affiliations, with the influence of trust switching between high and low (note how the green and blue lines cross over at each regime shift). This is in fact, an inference about the between state differences in trust in existing political regimes, and thus an indicator of state level political affiliation, that was inferred purely from tax rate and beer consumption data. This is an example of the Δ variable suggested in equation 5.

Conclusions

Repeat purchase and consumption decisions are reliant on multiple cognitive processes, and how people respond to

[3]It should be noted that there exist other, more sophisticated and customized econometric models for describing this data. The standard model is used as a baseline comparison to compare the generalizability of standard choice versus cognitive based models.

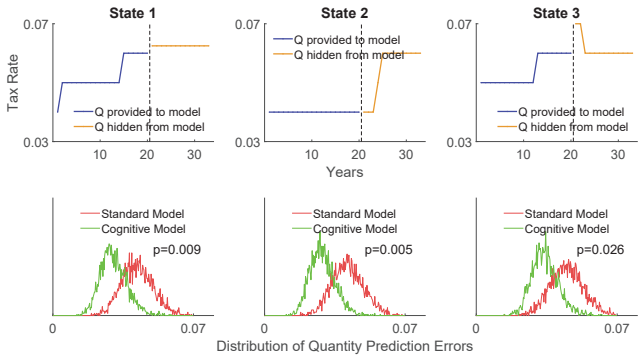


Figure 5: **Application to real world data:** Top panel - Changes in tax rate; Bottom Panel - Prediction error about consumption quantity from standard and cognitive models.

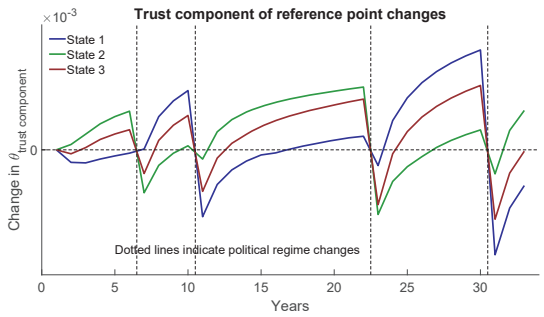


Figure 6: **Effect of Trust based adaptation on reference point:** illustrated are the between state differences based on the shifts in political regime.

changes in prices, taxes, and other policies may deviate significantly from rational models of choice. This paper highlights the importance of a structural model that captures how people's internal expectations may evolve over time, and how capturing this cognitive characterization can help the descriptive and predictive quality of psychological and econometric models. Future work will apply the models to a wider range of experimental and real world data, including identifying heterogeneous sub-population clusters within a larger population (Bell & Lattin, 2000). It will explore the implications for economic predictions, policy implications, and our basic understanding of how adaptive human behavior evolves over time.

References

Ayyagari, P., Deb, P., Fletcher, J., Gallo, W. T., & Sindelar, J. L. (2009). *Sin taxes: do heterogeneous responses undercut their value?* (Tech. Rep.). National Bureau of Economic Research.

Bell, D. R., & Lattin, J. M. (2000). Looking for loss aversion in scanner panel data: The confounding effect of price response heterogeneity. *Marketing Science*, *19*(2), 185–200.

Chetty, R. (2015). Behavioral economics and public policy: A pragmatic perspective. *The American Economic Review*, *105*(5), 1–33.

Chetty, R., Friedman, J. N., Olsen, T., & Pistaferri, L. (2009). *Adjustment costs, firm responses, and micro vs. macro labor supply elasticities: Evidence from danish tax records* (Tech. Rep.). National Bureau of Economic Research.

Chetty, R., Looney, A., & Kroft, K. (2009). Salience and taxation: Theory and evidence. *The American economic review*, *99*(4), 1145–1177.

Fletcher, J. M., Frisvold, D. E., & Tefft, N. (2015). Non-linear effects of soda taxes on consumption and weight outcomes. *Health economics*, *24*(5), 566–582.

Frederick, S., & Loewenstein, G. (1999). 16 hedonic adaptation. *Well-being: Foundations of hedonic psychology*, 302.

Gately, D. (1992). Imperfect price-reversibility of us gasoline demand: asymmetric responses to price increases and declines. *The Energy Journal*, 179–207.

Goodwin, P. B. (1992). A review of new demand elasticities with special reference to short and long run effects of price changes. *Journal of transport economics and policy*, 155–169.

Greene, J. D. (2007). Why are vmpfc patients more utilitarian? a dual-process theory of moral judgment explains. *Trends in cognitive sciences*, *11*(8), 322–323.

Greene, J. D. (2009). The cognitive neuroscience of moral judgment. *The cognitive neurosciences*, *4*, 1–48.

Hughes, J. E., Knittel, C. R., & Sperling, D. (2006). *Evidence of a shift in the short-run price elasticity of gasoline demand* (Tech. Rep.). National Bureau of Economic Research.

Jones, M., & Sugden, R. (2001). Positive confirmation bias in the acquisition of information. *Theory and Decision*, *50*(1), 59–99.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, *2*(2), 175.

Palminteri, S., Lefebvre, G., Kilford, E. J., & Blakemore, S.-J. (2016). Confirmation bias in human reinforcement learning: evidence from counterfactual feedback processing. *bioRxiv*, 090654.

Plummer, M., et al. (2003). Jags: A program for analysis of bayesian graphical models using gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing* (Vol. 124, p. 125).

Sitzia, S., & Zizzo, D. J. (2012). Price lower and then higher or price higher and then lower? *Journal of Economic Psychology*, *33*(6), 1084–1099.

Sitzia, S., & Zizzo, D. J. (2015). Price lower and then higher or price higher and then lower?. [data collection]. uk data service. sn: 851705, http://doi.org/10.5255/ukda-sn-851705.

Thaler, R. H. (1999). Mental accounting matters. *Journal of Behavioral decision making*, *12*(3), 183.

Thaler, R. H. (2008). Mental accounting and consumer choice. *Marketing Science*, *27*(1), 15–25.

Villas-Boas, S. B., Berck, P., Stevens, A., & Moe-Lange, J. (2016). Measuring consumer responses to a bottled water tax policy. *American ournal of Ag Economics*.

Xia, L., & Monroe, K. B. (2010). Is a good deal always fair? examining the concepts of transaction value and price fairness. *Journal of Economic Psychology*, *31*(6), 884–894.

Action prediction during real-time social interactions in infancy

Claire Monroy (claire.monroy@osumc.edu)
Chi-hsin Chen (chi-hsin.chen@osumc.edu)
Derek Houston (derek.houston@osumc.edu)
Department of Otolaryngology—Head and Neck Surgery
The Ohio State University Wexner Medical Center, Columbus, OH 43210 USA

Chen Yu (chenyu@indiana.edu)
Department of Psychological and Brain Sciences
Indiana University, Bloomington, IN 47405 USA

Abstract

Developmental theory considers action prediction as one of several processes involved in determining how infants come to perceive and understand social events (Gredebäck & Daum, 2015). Action prediction is observed from early in life and is considered an important social-cognitive skill. However, knowledge about infant action prediction is limited to evidence from screen-based eye-tracking tasks. Little is known about action prediction in real-life action contexts. Our aim in the current study was to provide new evidence on whether and how infants anticipate actions in free-flowing parent-child interaction. Using dual head-mounted eye-tracking, we analyzed infants' visual anticipations of their parents' reaching actions while they played with objects together. Findings reveal that infants anticipate their parents' actions at a rate higher than would be expected by chance.

Keywords: dual head-mounted eye-tracking; action prediction; parent-child interaction; social-cognitive development

Introduction

Action prediction refers to the ability to anticipate the outcome or endpoint of another person's goal-directed action (Flanagan & Johansson, 2003). This ability serves several important perceptual and cognitive functions: in a noisy and dynamic environment, anticipation allows the observer to direct visual attention to where important events will occur next (Gredebäck, Johnson, & von Hofsten, 2010). Action prediction also facilitates smooth, coordinated interactions. For instance, a simple interaction such as passing an object to another person requires planning a motor response at a precise moment in time and space to grasp the object successfully. Anticipating the other person's action and gazing to the location their hand will go next allows this kind of joint coordination to take place (Knoblich & Sebanz, 2012). For infants, whose developing system is solving the challenge of integrating their motor and visual systems, action prediction is an emerging skill (Falck-Ytter, Gredebäck, & von Hofsten, 2006; Kanakogi & Itakura, 2011; Monroy, Gerson, & Hunnius, 2017). In the current study, we investigated action prediction in 9-month-old infants, who are at the cusp of acquiring new fine motor skills and demonstrating rapid growth in their social-cognitive skills.

Prior research has demonstrated that infants exploit multiple cues to anticipate observed actions. For instance, infants can use kinematic cues from movement trajectories (Rosander & von Hofsten, 2011; Stapel, Hunnius, & Bekkering, 2012), the statistical regularities in familiar action sequences (Monroy, Gerson, & Hunnius, 2017), and knowledge about an actor's goal (Woodward, 1998). This ability develops within the first year of life: at 12 months, but not at 6 months, infants can anticipate unambiguous reaching actions (Falck-Ytter et al., 2006). By 9 months of age, infants can predict the endpoints of simple reaching actions based on motor cues from pincer and palmar grasps (Monroy et al., 2017; Senna et al., 2016).

The research described above is exclusively based on evidence from tightly controlled, yet artificial reaching paradigms. These paradigms have been useful in refining current theories about infants' action perception (Gredebäck & Daum, 2015). However, little is currently known about action prediction abilities 'in the wild', as infants interact with others while freely moving about in the environment. It is unknown whether infants' anticipatory behavior in laboratory contexts would generalize to the messier, more complex action contexts of real life. Here, we aimed to provide new evidence for whether and how frequently infants predict their parents' actions during free-flowing parent-child play.

In real-life contexts such as toy play, infants spend a great deal of time engaged with objects (e.g., almost 90% of the time; Yuan, Xu, Yu, & Smith, 2019) and their visual attention is characterized by long fixations to objects they are holding themselves (Yu & Smith, 2013). Based on these findings from recent head-mounted eye-tracking studies, our first question was whether infants do anticipate others' actions in real life, as they do in controlled laboratory contexts. If so, our second aim was to identify the frequency with which they do so and whether this frequency occurs at a rate higher than would be expected by chance. In the current study, we quantified the proportion of anticipated reaching actions during parent-child play and compared these to chance proportions.

To examine further the contexts in which action prediction can occur during parent-child play, we also analyzed infants' visual attention and manual activity during parents' reaching actions. For instance, to make a successful anticipation, do infants need to be disengaged from other

non-target objects? Do they exploit ostensive cues by attending to their parent's face (Senju & Csibra, 2008)? Given the limitations of infants' visual attention and their tendency to focus on their own manual actions at this age (Yu & Smith, 2017), one possibility is that infants demonstrate anticipations when they are less active themselves (i.e., better opportunity to anticipate) or when they are more socially engaged with their parent (e.g., more face looking).

Method

Participants

The sample consisted of 32 parent-infant dyads (mean age = 9.3 months, range = 9-9.7; 18 females). All children were born full-term and had no developmental diagnoses.

Procedure

Infants and parents were seated at a child-sized table across from one another. Both dyad members were fitted with head-mounted eye-trackers from Positive Science, LLC (Figure 1). Each eye-tracker has an infrared camera that records the right eye and a head camera that records the field of view. Two additional cameras recorded a third-person view of each dyad member. All six cameras recorded at 30Hz and were synchronized offline using custom-written Python scripts.

Figure 1: Experimental setup. A parent and her infant are seated across from one another playing with familiar objects. The crosshair indicates the estimated gaze direction.

To calibrate the eye-trackers, an engaging toy was placed in 15 unique locations on the tabletop to capture the infant's attention. Parents were instructed to attend as well. This phase was use for offline calibration using Yarbus software by marking the locations on the corresponding video frames when the eye was directed at the target.

Following calibration, participants were presented with six familiar, engaging toys (a car, cup, a train, a duck, a plane, and a boat). Toys were grouped into two sets of three, with each set containing one red, one green and one blue toy. Parents were instructed to play with their infants "as they normally would at home". Dyads played with each toy set twice for 90 seconds, yielding six possible minutes of interaction. The order of toy sets was counterbalanced across dyads.

Data processing

After offline calibration, gaze direction was superimposed onto the head camera recording with a crosshair, yielding an additional recording of the calibrated gaze. All camera recordings were then exported into a series of single frames. Each camera contributed a maximum of 10,800 frames per dyad (six minutes of recording at 30 frames per second).

Infants' gaze direction and parents' reaching actions were then manually coded frame-by-frame. For gaze, two independent coders used frames from the calibrated recording to determine whether the crosshair fell within one of four regions of interest (ROIs): the three novel objects and the parent's face. Frames were excluded whenever the eye-tracker failed to capture the eye (e.g., the child knocked the camera out of place), in between trials, or whenever the child was off-task. The second coder annotated a random 10% of the frames. Reliability ranged from 82-95% (Cohen's kappa = .81).

Additional coders annotated parent reaching actions: for each frame, the coder determined whether the parent was reaching for an object and, if so, which one. Reaching was defined as any movement towards an object that ended when contact was made. Right and left hands were coded separately and then merged to yield one data stream.

To identify infants' action prediction—anticipatory looks to the targets of their parents' reaching actions—the two data streams from the infant gaze and the parent reaching were aligned. Action prediction was defined as a gaze to an object that occurred after the onset of a parent reach to that same target, but before the reach was completed (Figure 2). This represents the time window in which the infant had enough information to predict the observed action, but before the hand reached the target. The number of anticipations per interaction was then summed and divided by the total number of valid parent actions to yield the proportion of anticipated actions.

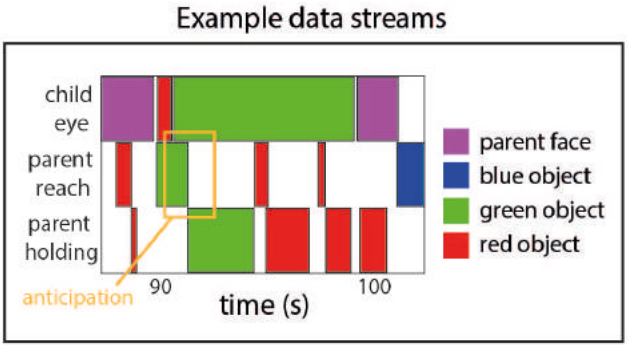


Figure 2: A sample of the aligned gaze and reaching data streams from a representative dyad. The yellow box highlights an example of an anticipation: the infant looks to the green object after the reach onset and prior to the end of the reach. Parent holding is included here for visualization purposes.

Not all parent reaches represented fair opportunities for anticipation. To estimate rates of anticipation out of the child's actual opportunities to anticipate—rather than total

number of reaches—we categorized all parent reaching actions as *valid* or *invalid* opportunities (Table 1).

Table 1: Criteria for categorizing parent reaches as *invalid* opportunities to anticipate.

	Criterion
1.	<200ms (to account for the time needed to program an eye movement)
2.	Subsequent contacts in cases of multiple object contacts (e.g., tapping or switching object from one hand to another)
3.	Infant reaching for the object at the same time
4.	Experimenter was reaching for or touching the object at the same time
5.	Both parent and object were entirely out of the child's view for the entire duration of the reach (e.g., child's eyes were closed, or object was underneath the table)
6.	Child threw or rolled the object to the parent and the parent simply received it[1]

Results

Action prediction

As a group, infants made 78 total anticipations out of 3640 gaze fixations. Per infant, they made an average of 2.44 anticipations throughout the interaction (range = 0-7, *SD* = 1.97). Parents made 1176 total reaching events, an average of 36.75 reaches per parent (range = 17-67, *SD* = 12.14). Of these, 563 represented valid opportunities to anticipate (average = 17.59 per parent, range = 7-33, *SD* = 5.08). The 78 total infant anticipations corresponded to valid reaches; there were no anticipations that corresponded to an invalid reach. Therefore, the mean proportion of anticipated reaches out of all valid reaches across infants was .13 (*SD* = .11).

These results indicate that infants do demonstrate action prediction at 9 months of age during free-flowing interaction, though infrequently (Figure 3). There was a substantial amount of variability among infants: while some infants never anticipated (*n* = 8, or 25% of the sample), others anticipated more than 40% of their parents' actions. After excluding infants who never anticipated, the mean proportion of anticipated reaches was .18 (SD = .09), which is consistent with the findings reported above from all infants.

Out of 563 total reaching events across all parents, in 94 of these events the infant was already looking to the target object when the parent initiated their reach. In these cases, the parent was most likely responding to the child's visual attention by reaching for what the infant is looking at. When these reaching events are removed from the total count—they can also be considered invalid opportunities to anticipate, since the child cannot anticipate a target they are already looking at—the average proportion of anticipated reaches increases to 0.16 (*SD* = 0.13).

Given the low frequency of this behavior, we tested whether infants' anticipations could have been due to chance overlaps between infant gaze and parent reaching behavior. For each infant, we created 1000 randomized time-series by shuffling the sequence of gaze fixations while preserving their overall duration. We then aligned each randomized gaze sequence with the sequence of parents' reaching actions, calculated the number of anticipations that could occur by chance, and averaged over these 1000 values to yield a baseline anticipation rate for each infant. This resulted in a mean of 1.31 baseline anticipations across infants (*range* = 0.31-3.33, *SD* = 0.65). A paired-samples t-test revealed that the average number of anticipations was significantly *higher* than baseline (*mean difference* = 1.13, $t(31) = 3.84$, $p = .001$). This result was the same when comparing the proportion of anticipated reaches with the chance proportion of .07, (*mean difference* = .09, $t(31) = 4.34$, $p < .001$). This finding reveals that infants' action prediction did not simply occur from chance overlaps between looking and parent reaching to the same object.

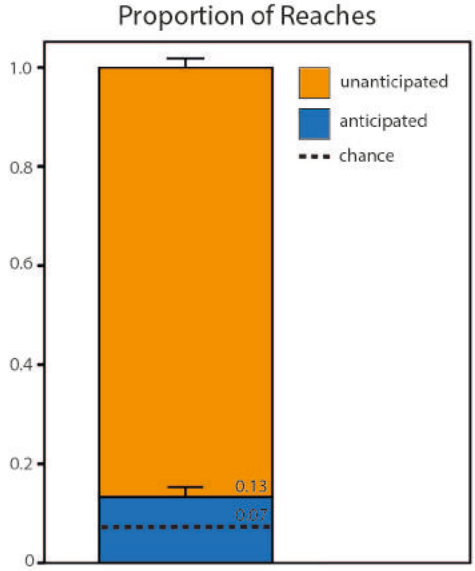


Figure 3: The proportions of reaching actions that were anticipated vs. unanticipated, with the dotted line representing chance. Error bars represent the s.e.m.

Infant visual attention and manual activity during parent reaching

To explore the characteristics of the parent-child interaction during reaching events, we examined the infant and parent behaviors that were occurring during each reaching event. Figure 4 illustrates the proportion of all valid parent reaches in which the infant was attending to or manipulating a different object from the target of the reach. Parents were also holding another object in their other hand during 37.5% of their reaches. In fact, there was not *one single* reaching event across all dyads with no concurrent

[1]Here, the child may be anticipating the causal outcome of their own action or the movement trajectory of the ball rather than their parents' action goal.

visual and/or manual activity to a non-target object from either the child or the parent.

Given this finding, we conducted an additional analysis on the rate of anticipation with the aim of only considering opportunities that did not overlap with infants' own actions. We therefore excluded reaches during which the infant was holding or reaching for a non-target object. In these instances, infants were occupied with planning their own manual actions, which requires vision. After excluding infants' manual activity, this resulted in 251 valid reaches actions across parents (*mean* = 7.84 per parent, *SD* = 3.62). The mean proportion of anticipated actions was 0.36 (*SD* = .31), which remained significantly higher than the chance level of 0.20 (*SD* = .12), $t(31) = 3.61$, $p = .001$. In other words, in a less demanding observation context (i.e., without competing goals from their own manual actions) infants will anticipate close to 40% of their parents' actions.

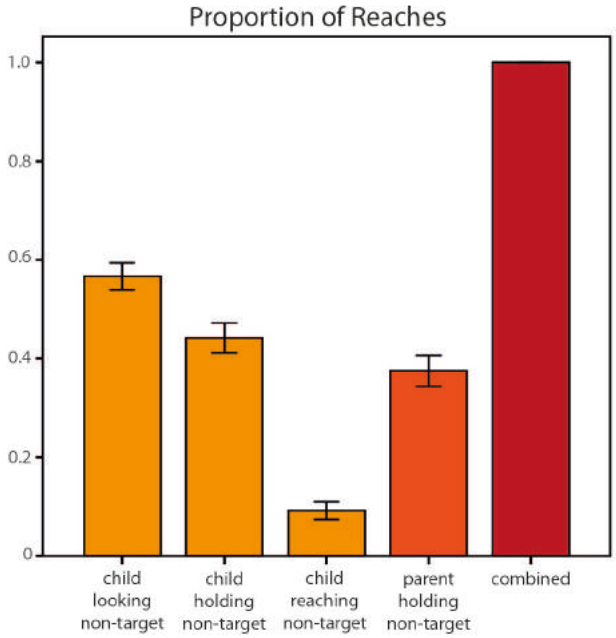


Figure 4: The proportion of reaching events in which infants or parents were concurrently attending to or manipulating a non-target object. From left to right, reaches during which 1) infants were looking at non-target object; 2) infants were holding a non-target object; 3) infants were reaching for a non-target object; 4), parents were holding a non-target object in their other hand, and 5) at least one of the above was occurring. (*Note:1-4 are not mutually exclusive and therefore add up to more than 1.*)

Anticipation latency

The mean duration of parents' anticipated reaches were 611.46ms (*SD* = 314.90, *range* = 200-1770). Figure 5 displays a bar chart of the time-course of infants' anticipations: the latencies between the onset of the reaching actions, infants' looks to the target object, and the contact with the goal (i.e., the end of the reach). The mean latency from the start of the reach to the moment the child looked to the target was 328.88ms (*SD* = 234.67). The mean latency from the gaze onset to the moment the hand reached the target was 282.58ms (*SD* = 282.31). In other words, on average infants required just over 300ms to detect and process their parents' movements and then anticipate. This includes the time required for infants to program an eye movement in response to a visual stimulus (Gredebäck et al., 2010). This finding suggests that movement cues were a strong cue for anticipation.

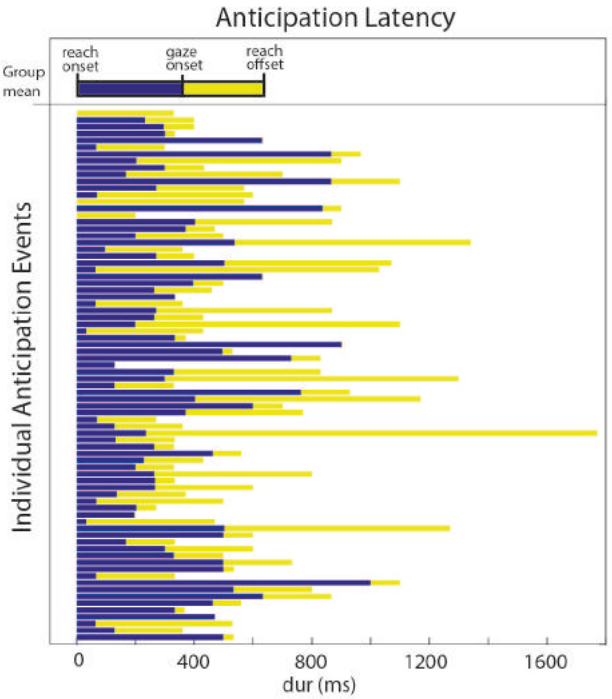


Figure 5: A histogram of the time course of infants' anticipations: blue bars indicate the latency between the start of the reach and the moment the child looked to the target; yellow bars indicate the latency between the child's look and the moment the parent's hand reached the target.

Looks to Parent Faces

If an infant is looking at their parent's face, they could be more likely to perceive the onset of their parent's reach and its trajectory. To determine whether attending to their parents facilitated anticipations, we calculated the proportion of anticipations that were immediately preceded by a fixation to the parent's face. Out of the 78 anticipations performed across all subjects, 18 were immediately preceded by a face look. For the remaining 60 anticipations, 42 of them did not have a face look within a 3-second window before or after the anticipation. This finding suggests that looking to parents' faces was not a strong cue for anticipation.

Discussion

The world of the developing infant is dynamic and constantly changing in both time and space. In the first year of life, infants become increasingly proficient actors and make rapid gains in their abilities to perceive and understand social events. Action prediction reflects an

important part of this social-cognitive process (Gredebäck & Daum, 2015; Hunnius & Bekkering, 2014) and has been widely studied using screen-based eye-tracking methods (Robson & Kuhlmeier, 2016). In the current study, we used dual head-mounted eye-tracking to investigate whether infants' predict their parents' object-directed reaching actions during free-flowing parent-child interaction.

Our primary finding is that, as a group, infants do anticipate their parents' actions at a rate that was significantly higher than what would be expected by chance. This finding demonstrates that infants' action prediction skills are not limited to the unambiguous, controlled action contexts that are typical in laboratory paradigms—they also demonstrate this ability during free-flowing parent-child play while they are also acting themselves.

On the other hand, the low rate of anticipation is also consistent with recent work investigating the real-time dynamics of parent-child interactions. New evidence from head-mounted eye-tracking studies have revealed, for instance, that infants actually rarely look to their parents' faces (Franchak, Kretch, Soska, & Adolph, 2010) and achieve joint attention through their own manual actions rather than through gaze following (Yu & Smith, 2016). The current study adds to this growing literature by revealing that infants also attend less to the goals of their parent's reaching actions. However, it is difficult to evaluate whether proportions of .13-.16 should actually be considered low, as there is no existing data to compare to these values. Future work could, for instance, quantify the rate of parents' anticipations of their infants' actions to provide a reference point.

A second finding was that infants anticipated their parents' actions on a rapid timescale—the mean latency after the reach onset was 328ms. In one recent study that also reported the "disengagement time"—i.e., the $reach_{onset}$ - $gaze_{onset}$ latency—on average infants required 344ms (SD = 209ms) to look to the target (Rosander & von Hofsten, 2011). These authors interpreted this finding as evidence that infants were able to use movement trajectories on a rapid timescale to accurately predict their parents' targets. In that study, infants were seated in a high chair and observed an experimenter move a small ball into a cylinder. Interestingly, infants demonstrated a similar timescale despite the increase in complexity of the toy play context.

A third finding that emerged is that there was not *one single* object-directed reaching action, in the entire sample, in which infants or parents were not simultaneously looking at or holding a different object than the target of the reach. Nevertheless, infants were still able to generate successful anticipations. One recent study may shed some insight into this finding. De Barbaro et al., (2016) investigated the qualitative shift from infant-guided object play to the triadic joint object play that emerges around 9-12 months of age. Their research highlights a "decoupling" between infant gaze and manual activity: in their studies, infants frequently directed their visual and sensorimotor attention to different objects (i.e., they do not look at what they are holding). Likewise, they frequently shifted their gaze from the objects in their own hands and those in their parents' hands. These authors propose that this "sensorimotor decoupling" in hand-eye coordination contributes to the emergence of triadic interactions, by enabling infants to manipulate objects while still attending to the objects in their parents' hands.

This finding also highlights the dissociation between screen-based action contexts and real life: infants rarely experience isolated, unambiguous moments without competing cues or distractors, unlike discrete trials in experimental studies. Although examining what infants can or cannot do in a controlled laboratory setup is an effective approach to understanding infant cognition, it is also critical to examine how behaviors emerge from complex contexts with competing goals. For example, in the toy play context examined here, infants need to efficiently control their visual attention on a rapid timescale to serve multiple tasks—guiding their own manual actions, predicting their partner's actions, and sometimes using gaze to send social signals to their partner. In such real-life contexts, the key question is how the infant cognitive system operates with multiple ongoing tasks and how they distribute cognitive resources (e.g., attention and memory) to coordinate and manage these tasks.

What are the functional consequences of action prediction? Anticipating the actions of their social partners may help infants form associations between other peoples' actions and their goals or intentions. This pathway has been proposed to provide a potential explanation for how infants transition from forming associations between the behaviors they observe and more complex social understanding skills within the first years of life (Hunnius & Bekkering, 2014; Ruffman, Taumoepeau, & Perkins, 2012). In fact, a related study in our lab found correlations between the frequency of infants' action prediction and their vocabulary size, both at the same age and up to 6 months later (Monroy et al., under review). This finding provides preliminary evidence that action prediction may not only reflect infants' current social information-processing skills (Gredebäck & Daum, 2015) but also provide learning opportunities that support their developing social-cognitive system.

Our study represents a first attempt to investigate action prediction in naturalistic parent-child play. In our experimental set-up, parents and infants engaged in unstructured object play limited to three objects. However, this paradigm is also limited in the kinds of information available to infants. For instance, parents' actions did not lead to any meaningful action goal, as our everyday actions do (e.g., making a sandwich or building a Lego tower). In addition, there were no regularities in parents' reaching actions that they could use to anticipate their next target, which is one cue that infants use to generate predictions (Monroy et al., 2017). In future work, we plan to investigate action prediction in a broader range of action contexts—for instance, when infants and parents are engaged in joint activities that feature structure and shared goals.

Acknowledgments

Thank you to Alexis Allard for data coding and to Steven Elmlinger, Melissa Hall, Charlene Tay, Charlotte Wozniak, Melissa Elston, and Seth Foster for data collection. We would also like to thank three anonymous reviewers for helpful comments on an earlier draft of this manuscript. Research reported in this publication was supported by the National Institute on Deafness and Other Communication Disorders under award number F32DC017076 to CM, and National Institutes of Health Grant R01HD074601 and R01HD093792 to CY.

References

de Barbaro, K., Johnson, C. M., Forster, D., & Deák, G. O. (2016). Sensorimotor decoupling contributes to triadic attention: A longitudinal investigation of mother-infant-object interactions. *Child Development*, *87*(2), 494–512. https://doi.org/10.1111/cdev.12464

Falck-Ytter, T., Gredebäck, G., & von Hofsten, C. (2006). Infants predict other people's action goals. *Nature Neuroscience*, *9*(7), 878–879. https://doi.org/10.1038/nn1729

Flanagan, J. R., & Johansson, R. S. (2003). Action plans used in action observation. *Nature*, *424*(6950), 769–771. https://doi.org/10.1038/nature01861

Franchak, J. M., Kretch, K. S., Soska, K. C., & Adolph, K. E. (2010). Head-mounted eye-tracking : A new method to describe infant looking. *Learning*, *82*(6), 1–9. https://doi.org/10.1111/j.1467-8624.2011.01670.x.Head-mounted

Gredebäck, G., & Daum, M. M. (2015). The Microstructure of Action Perception in Infancy: Decomposing the Temporal Structure of Social Information Processing. *Child Development Perspectives*, *9*(2), 79–83. https://doi.org/10.1111/cdep.12109

Gredebäck, G., Johnson, S., & von Hofsten, C. (2010). Eye tracking in infancy research. *Developmental Neuropsychology*, *35*(1), 1–19. https://doi.org/10.1080/87565640903325758

Hunnius, S., & Bekkering, H. (2014). What are you doing? How active and observational experience shape infants' action understanding. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1644), 20130490. https://doi.org/10.1098/rstb.2013.0490

Kanakogi, Y., & Itakura, S. (2011). Developmental correspondence between action prediction and motor ability in early infancy. *Nature Communications*, *2*, 341. https://doi.org/10.1038/ncomms1342

Knoblich, G., & Sebanz, N. (2012). The Social Nature of Perception and Action. *Current Directions in Psychological*, *15*(3), 99–104. Retrieved from http://cdp.sagepub.com/content/15/3/99.short

Monroy, C., Gerson, S., & Hunnius, S. (2017). Infants' motor proficiency and statistical learning for actions. *Frontiers in Psychology*, *8*(DEC). https://doi.org/10.3389/fpsyg.2017.02174

Monroy, C., Gerson, S., & Hunnius, S. (2017). Toddlers' action prediction: Statistical learning of continuous action sequences. *Journal of Experimental Child Psychology*, *157*, 14–28. https://doi.org/10.1016/j.jecp.2016.12.004

Monroy, C., Chen, C., Houston, D., Yu, C. (under review). Action prediction during real-time parent-infant play predicts language development.

Robson, S. J., & Kuhlmeier, V. A. (2016). Infants' Understanding of Object-Directed Action: An Interdisciplinary Synthesis. *Frontiers in Psychology*, *7*(February), 111. https://doi.org/10.3389/fpsyg.2016.00111

Rosander, K., & von Hofsten, C. (2011). Predictive gaze shifts elicited during observed and performed actions in 10-month-old infants and adults. *Neuropsychologia*, *49*(10), 2911–2917. https://doi.org/10.1016/j.neuropsychologia.2011.06.018

Ruffman, T., Taumoepeau, M., & Perkins, C. (2012). Statistical learning as a basis for social understanding in children. *British Journal of Developmental Psychology*, *30*(1), 87–104. https://doi.org/10.1111/j.2044-835X.2011.02045.x

Senju, A., & Csibra, G. (2008). Gaze following in human infants depends on communicative signals. *Current Biology*, *18*, 668–671.

Senna, I., Addabbo, M., Bolognini, N., Longhi, E., Macchi Cassia, V., & Turati, C. (2016). Infants' visual recognition of pincer grip emerges between 9 and 12 months of age. *Infancy*, *22*(3), 389–402. https://doi.org/10.1111/infa.12163

Stapel, J. C., Hunnius, S., & Bekkering, H. (2012). Online prediction of others' actions: The contribution of the target object, action context and movement kinematics. *Psychological Research*, *76*(4), 434–445. https://doi.org/10.1007/s00426-012-0423-2

Woodward, A. L. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition*, *69*(1), 1–34. https://doi.org/10.1016/S0010-0277(98)00058-4

Yu, C., & Smith, L. B. (2013). Joint attention without gaze following: Human infants and their parents coordinate visual attention to objects through eye-hand coordination. *PLoS ONE*, *8*(11), e79659. https://doi.org/10.1371/journal.pone.0079659

Yu, C., & Smith, L. B. (2016). Multiple sensory-motor pathways lead to coordinated visual attention. *Cognitive Science*, *41*, 5–31. https://doi.org/10.1111/cogs.12366

Yu, C., & Smith, L. B. (2017). Hand-eye coordination predicts joint attention. *Child Development*, *88*(6), 2060–2078. https://doi.org/10.1111/cdev.12730

Yu, C., Suanda, S. H., & Smith, L. B. (2018). Infant sustained attention but not joint attention to objects at 9 months predicts vocabulary at 12 and 15 months. *Developmental Science*, e12735. https://doi.org/10.1111/desc.12735

Yuan, L., Xu, T. L., Yu, C., & Smith, L. B. (2019). Sustained visual attention is more than seeing. *Journal of Experimental Child Psychology*, *179*, 324–336. https://doi.org/10.1016/j.jecp.2018.11.020

Eye See What You're Saying: Beat Gesture Facilitates Online Resolution of Contrastive Referring Expressions in Spoken Discourse

Laura M. Morett[1] (lmorett@ua.edu), **Scott H. Fraundorf**[2] (sfraundo@pitt.edu),
James C. McPartland[3] (james.mcpartland@yale.edu)
[1]Department of Educational Studies, University of Alabama, Tuscaloosa, AL 35401
[2]Department of Psychology and LRDC, University of Pittsburgh, Pittsburgh, PA 15260
[3]Yale Child Study Center, New Haven, CT 06520

Abstract

This study investigated how beat gesture and contrastive pitch accenting affect online contrastive reference resolution during spoken discourse comprehension. Evidence from gaze fixations indicated that beat gesture encouraged fixations to target referents of contrastive referring expressions and that contrastive accenting encouraged fixations to competitor referents of non-contrastive referring expressions. Notably, beat gesture and contrastive accenting acted independently, indicating that their effects are additive rather than interactive. Moreover, neither beat gesture nor contrastive accenting affected an observed tendency to anticipate contrastive referring expressions. Together, these results provide the first evidence that beat gesture, like contrastive accenting, is interpreted as a cue to contrast during online reference resolution in spoken discourse comprehension.

Keywords: beat gesture; pitch accent; reference resolution; discourse processing; visual world; eye tracking

Introduction

Successful discourse comprehension entails establishing relations between entities. One such relation is *contrast,* which refers to a contradiction between two themes (Myhill & Xing, 1996). An example can be seen in the distinction between referents in the following discourse: *The report isn't due on Tuesday; it's due on Thursday*. Although contrast can be discerned semantically, cues conveying prominence can be used to highlight it, strengthening the propositional representations of both the speaker and the listener. Two such cues are pitch accent—alterations in in speech fundamental frequency (f0), duration, and intensity (Ladd, 1996)—and beat gesture—simple rhythmic gesture (McNeill, 1992; 2005). Although processing of these cues has been studied in offline discourse comprehension (Kushch & Prieto, 2016; Llanes-Coromina et al., 2018), it is currently unclear how it affects *online* discourse comprehension. The current study uses eyetracking to examine how independently manipulating pitch accent and beat gesture affects online contrast interpretation in spoken discourse. In doing so, it provides insight into the individual and combined contributions of these cues to prediction and resolution of contrast in particular, as well as representation and processing of inter-entity relations more generally, in spoken discourse.

Cues to contrast

Two of the most prominent types of pitch accenting in English discourse are presentational pitch accenting (PPA), which is used to convey new, non-contrastive information, and contrastive pitch accenting (CPA), which is used to convey information contrasting with other mentioned information. These two pitch accents differ acoustically; PPA (H* in the ToBI framework) consists of a high pitch target and f0 high in the talker's range, whereas CPA (L+H* in the ToBI framework) consists of an initial low pitch followed by a sharp rise to a high target on the accented syllable (Beckman & Elam, 1997; K. Silverman et al., 1992). Previous work demonstrates that listeners are sensitive to the distinction between PPA and CPA, and this is reflected in both memory for discourse and real-time discourse comprehension. Referents with CPA are remembered better than referents with PPA, particularly when a salient contrasting item must be rejected (e.g., remembering *Scottish* rather than *British;* Fraundorf et al., 2010; 2012; Lee & Fraundorf, 2016; Lee & Snedeker, 2016; Sanford, Sanford, Molle, & Emmott, 2006). Moreover, CPA facilitates rejection of items contrasting with contrastively-accented referents (e.g., *dish* given *antenna*), but not objects with non-contrastive relations to those referents (e.g., *television* given *antenna*; Braun & Tagliapietra, 2010). Lastly, in eyetracking studies, CPA encourages anticipatory looks to objects contrasting with previously-mentioned referents (e.g., after hearing "red scissors," to purple scissors upon hearing "PURPLE"), even when the referent is subsequently revealed to be non-contrastive (e.g., *book*; Ito, Jincho, Minai, Yamane, & Mazuka, 2012; Ito & Speer, 2008; Kurumada, Brown, Bibyk, Pontillo, & Tanenhaus, 2014; Watson, Tanenhaus, & Gunlogson, 2008; Weber, Braun, & Crocker, 2006).

Like pitch accenting, beat gesture is used to emphasize important information in spoken discourse, such that it serves as a "yellow gestural highlighter" (McNeill, 2006). Indeed, both alone and in combination with pitch accenting, beat gesture enhances memory for information conveyed via discourse (Austin & Sweller, 2014; Igualada, Esteve-Gilbert, & Prieto, 2017; Morett, 2014; Vilà-Giménez, Igualada, & Prieto, in press). Moreover, some work indicates that beat gesture enhances memory for contrastive information in particular, especially when it occurs in conjunction with CPA (Kushch & Prieto, 2016; Llanes-Coromina et al., 2018). These findings suggest that beat gesture strengthens memory traces for information in spoken discourse by increasing its salience visually. In addition to their similarity in function, beat gesture and pitch accenting are closely related in timing. Indeed, beat gesture and pitch accenting are temporally aligned on both the sentential and syllabic levels (Esteve-Gilbert & Prieto, 2013; Leonard & Cummins, 2011), suggesting that the temporal relationship between these two cues to prominence is based on prosody.

Considered as a whole, these findings demonstrate that beat gesture and pitch accenting are closely related in timing and meaning. This suggests that beat gesture—as another cue

to prominence—might facilitate online processing of contrast in spoken discourse, as CPA does, Further, the functional similarity of beat gesture and CPA highlights the need to investigate their effects on online contrast processing not only independently but also *conjointly*.

Cue integration

Despite the similar function and close relationship of pitch accenting and beat gesture, relatively little research has examined how the presence—and absence—of these cues in relation to one another affects interpretation of contrast in spoken discourse. In a focus production task in which participants produced beat gesture and/or pitch accenting on one or both referents of a sentence (*Amanda goes to Malta*), referents produced with beat gesture alone had higher vowel formants and were more likely to be perceived as pitch accented than referents unaccompanied by beat gesture (Krahmer & Swerts, 2007). However, in a similar task that involved producing beat gesture in conjunction with contrastive corrections after hearing sentences (*Baba holds the baby? → Mumu holds the baby*), beat gesture production did not affect the articulatory or acoustic correlates of CPA (Roustan & Dohen, 2010). Moreover, in a focus comprehension task in which pitch accenting and beat gesture were independently manipulated in conjunction with the patients of transitive sentences (e.g., *Yesterday, Anna brought fresh lilies to the room*), pitch accenting elicited a larger N400 response when beat gesture was absent than when beat gesture was present, indicating greater inconsistency with predictions or difficulty of semantic integration in the former case (Wang & Chu, 2013). Taken together, these findings indicate that the co-occurrence patterns of pitch accenting and beat gesture affect their interpretation as cues to contrast in spoken discourse.

The influence of beat gesture on interpretation of pitch accenting is also evident in work indicating that information conveyed via spoken discourse accompanied by both beat gesture and pitch accenting is remembered better than the same information accompanied by pitch accenting alone. This result has been observed for memory of focal, non-contrastive information (Igualada, Esteve-Gibert, & Prieto, 2017; Kushch, Igualada, & Prieto, 2018; Morett, 2014; Vilà-Giménez, Igualada, & Prieto, in press) as well as contrastive information (Kushch & Prieto, 2016; Llanes-Coromina et al., 2018; Morett & Fraundorf, under review). With respect to memory for contrastive information, the authors' previous work indicates that, when both cues are manipulated independently in a within-subjects design, contrastive information with CPA is remembered better than contrastive information with PPA when beat gesture is present, but not when beat gesture is absent. When beat gesture is *never* present, however, contrastive information with CPA is remembered better than contrastive information with PPA, consistent with the findings of previous work demonstrating the same effect using similar paradigms presented only in the auditory modality (Fraundorf et al., 2010; 2012; Lee & Fraundorf, 2016; Lee & Snedeker, 2016; Sanford, Sanford, Molle, & Emmott, 2006). Considered as a whole, these findings suggest that beat gesture and CPA influence one another in offline discourse comprehension and memory. However, it is less clear whether and how these cues interact in online discourse processing.

To elucidate how beat gesture and CPA affect contrastive reference resolution in online spoken discourse, we examined differences in fixations to referents accompanied by beat gesture and/or CPA. To do so, we used a modified version of the visual world paradigm that included video. The visual world paradigm has been used successfully to examine how CPA affects online reference resolution (Ito & Speer, 2008; Kurumada et al., 2014; Watson, Tanenhaus, & Gunlogson, 2008), as well as how representational gesture is integrated with speech online (L. B. Silverman, Bennetto, Campana, & Tanenhaus, 2010). Based on these studies and the related work discussed above, we predicted that beat gesture and CPA would affect online reference resolution. Specifically, we predicted that, when referents contrasted only in color (e.g., *blue triangle* and *red triangle*), the presence of beat gesture alongside the color word would facilitate reference resolution, particularly in conjunction with CPA. By comparison, we predicted that when referents differed in both color and shape (e.g., *blue square* and *red triangle*), the presence of beat gesture alongside the color word would misleadingly suggest a color contrast and hinder reference resolution, particularly in conjunction with CPA.

Methods

Participants

Forty adult native English speakers (age range: 18-35 years; 29 females, 11 males) were recruited to participate in this study on a paid basis. All participants had normal hearing and normal or corrected-to-normal vision and were not colorblind. Additionally, participants were screened for factors affecting eye movements (e.g., psychiatric and neurological disorders, recreational drug use).

Materials

A total of 672 referring expressions conveying simple instructions were audio recorded (see 1a-2b for examples; 32 practice, 640 experimental). In both practice and experimental trials, half of referring expressions provided context, with standard PPA on both color and shape words. The other half of referring expressions provided continuation, consisting of half critical and half filler trials. In critical trials, the color word always differed from that of the preceding context referring expression, and the shape word was either the same (color-contrast; 1a) or different (both-contrast; 1b). In both types of critical trials, pitch accenting was manipulated by splicing color words with CPA or PPA into identical carrier sentences (in which original color and shape words had PPA) to control acoustic realization of the rest of the referring expression. Filler trials were created to represent the other possibilities, in which the color word was the same as that of the preceding context referring expression and the shape word either differed (shape-contrast; 2c) or was the same (no-contrast; 2d). In these trials, pitch accent was always felicitous, such that shape words in shape-contrast referring expressions always had CPA and shape words in no-

contrast referring expressions always had PPA. Sentences in filler trials were recorded wholesale and were not spliced.

1a. *Color-contrast:* Click on the blue triangle → red triangle.
1b. *Both-contrast:* Click on the blue square → red triangle.
2a. *Shape-contrast:* Click on the red square → red triangle.
2b. *No-contrast:* Click on the red triangle → red triangle.

840 videos of a talker producing the sentences described above were recorded to accompany audio recordings. 40 of these videos were used for practice trials, and 800 were used for experimental trials. 336 of these videos, which accompanied context sentences, did not contain beat gestures. In the other 504 videos, which accompanied continuation referring expressions, beat gesture was either present or absent alongside the color word (for critical trials) or shape word (for filler trials). Two videos were recorded to accompany each critical referring expression. In one of these videos, beat gesture was present alongside color words; in the other, beat gesture was absent. Videos recorded to accompany filler trials maintained the association between beat gesture and CPA present in natural speech; in videos accompanying shape-contrast referring expressions), beat gesture occurred alongside CPA-accented shape words, whereas beat gesture was absent from videos accompanying PPA-accented no-contrast referring expressions. A within-participants design was used such that each participant received all combinations of contrast type, beat gesture, and pitch accenting; however, combinations for individual trials were counterbalanced across participants (see Table 1 for experimental design summary). All videos were recorded separately from audio and were aligned temporally with it in post-production. Because beat gestures were produced with one hand, consisting of a single downward stroke with the palm oriented upward, horizontally-flipped duplicates were created for all videos in post-production.

A total of 64 objects (8 colors x 8 shapes) were created for inclusion in arrays accompanying audio and video stimuli. Videos were presented centrally with a circular mask, with objects positioned equidistantly (see Fig. 1). Locations in which objects appeared were counterbalanced to control for contingencies between them and beat gesture orientation.

Table 1: Experimental design (excluding practice trials).

Type	Contrast	Accent	Gesture	Trials
Critical	Color	CPA	Beat	20
Critical	Color	No CPA	Beat	20
Critical	Color	CPA	None	20
Critical	Color	No CPA	None	20
Critical	Both	CPA	Beat	20
Critical	Both	No CPA	Beat	20
Critical	Both	CPA	No Beat	20
Critical	Both	No CPA	No Beat	20
Filler	Shape	CPA	Beat	40
Filler	None	No CPA	None	40

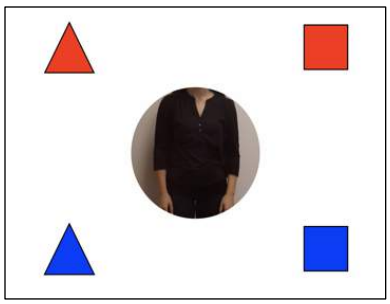


Figure 1: Schematic of screen configuration.

Procedure

Fixation data was collected remotely from the right eye at a 500 Hz sampling rate using an EyeLink 1000 eyetracker. Before beginning the experimental task, participants were seated 55-56 cm from the screen (35° 55' 0.32" visual angle). Gaze was calibrated to within 0.5° of visual angle using 13 points of reference. Drift checks and recalibrations were performed between experimental trial blocks.

At the beginning of the experimental task, participants were told that its objective was to test their ability to follow instructions. Participants were told to respond to all instructions issued in the paradigm by clicking on the appropriate object. The experiment was programmed such that participants who clicked on the wrong object were instructed to click on the correct object to proceed. However, all responses to critical referring expressions were correct. This was not surprising given that the task was simple and straightforward, as is characteristic of visual world tasks (Huettig, Rommers, & Meyer, 2011; Salverda, Brown, & Tanenhaus, 2011), and our intent was to assess the online processing of correctly-understood referring expressions.

To become familiar with the experimental task, participants first completed a practice phase consisting of 8 trials. Participants then proceeded to the experimental phase, which consisted of four blocks of 40 trials each. In both phases, critical and filler trials were randomly interleaved. In each trial, an array of objects appeared and a video began playing, and the context referring expression was presented aurally after a 200 ms delay. This configuration ensured that the apex of the beat gesture occurred 200 ms prior to the onset of the corresponding word, which is consistent with the timing of gesture production relative to speech in natural discourse (Morrel-Samuels & Krauss, 1992) as well as perceptual biases for the timing of beat gesture relative to speech (Leonard & Cummins, 2011). Following a correct response, the video disappeared and was replaced by a gray circular placeholder for 1000 ms while the object array remained on screen. Subsequently, the sequence repeated with the continuation referring expression and corresponding video. Following a correct response, the trial ended and, after a blank screen was displayed for 1000 ms, a new trial began.

Results

We examined fixations during two periods of the critical referring expression: *color word* (color word onset to shape word onset) and *shape word* (shape word onset to response

onset). To account for saccade planning, each period was shifted ahead by 200 ms. Two interest areas relevant to the main research question were defined: the *target object* referred to by the critical referring expression, and the *competitor object* that is temporarily consistent with the unfolding linguistic input. In color-contrast trials, in which the target object contrasted with the referent of the context referring expression only in color, the competitor object differed in both color and shape; in both-contrast trials, in which the target object differed from the referent of the context referring expression in both color and shape, the competitor object contrasted only in color. In addition, the *video* interest area was defined to confirm that participants were watching the video. Participants fixated the video more than target and competitor objects combined during both the color word (77.65% of fixations) and shape word (52.19% of fixations) interest periods.

To account for non-independence of samples, we summed fixations within each interest period in each trial and took the empirical logit (Barr, 2008). Because we were interested in how beat gesture and CPA facilitated and hindered reference resolution, empirical logit values were computed for fixations to target and competitor objects separately. These values were then entered into linear mixed effects models, which were fit using the *lme4* R package (Bates, Mächler, Bolker, & Walker, 2015) and evaluated via null hypothesis statistical testing using the *lmerTest* R package (Kuznetsova, Brockhoff, & Christensen, 2017). Each model implemented the maximal random effect structure permitting convergence, with beat gesture, CPA, contrast, and their interactions as fixed effects and participant and trial as random effects. To account for any effects of spatial orientation of the gesture, we also included gesture orientation and target object side as control variables.

Color Word Interest Period

For target fixations, we observed a main effect of contrast, indicating a higher likelihood of fixating target objects during color-contrast (M = -0.11, SD = 0.55) than both-difference critical referring expressions (M = -0.17, SD = 0.55; t = -3.57, p < .001); however, no interactions between contrast and either accent or gesture were observed. For competitor fixations, no main effect of contrast was observed (color: M = -0.16, SD = 0.54; both: M = -0.09, SD = 0.57; t = 1.62, p = .11). Moreover, there were no main effects or interactions of gesture orientation and target side for target or competitor fixations. Thus, although these results suggest a baseline bias towards contrastive interpretation of critical referring expressions, neither beat gesture nor CPA enhanced resolution of these expressions prior to disambiguation.

Shape Word Interest Period

We observed significant two-way interactions between orientation and target side for both target and competitor fixations (target: B = 0.48, SE = 0.08, t = 6.32, p < .001; competitor: B = -0.10, SE = 0.04, t = -2.58, p = .01), indicating a baseline tendency to fixate target objects and not to fixate competitor objects appearing on the side of the array congruent with the orientation of accompanying beat gesture.

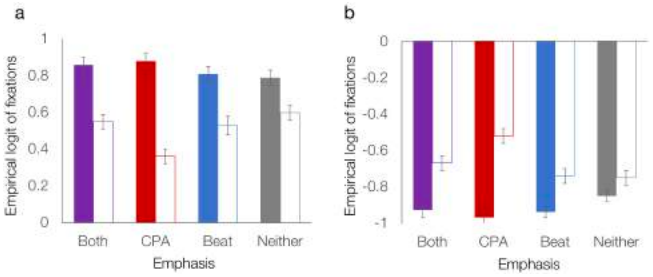


Figure 2: Fixations on (a) target and (b) competitor objects during the shape word period for color-contrast (filled) and both-difference (outlined) critical referring expressions by prominence cue (CPA, Beat, Both, Neither).

We also observed a continuation of the baseline color-contrast preference: There was a higher likelihood of fixating target objects and a lower likelihood of fixating competitor objects during color-contrast than both-difference critical referring expressions. Critically, this preference was qualified by interactions with beat gesture for target fixations (B = 0.13, SE = 0.06, t = 2.18, p = .03) and CPA for competitor fixations (B = 0.26, SE = 0.10, t = 2.51, p = .02; see Fig. 2). A simple-effect analysis revealed a greater likelihood of fixating target objects during color-contrast than during both-difference critical referring expressions when beat gesture was present (color: M = 0.84, SD = 0.80; both: M = 0.54, SD = 0.87; t = -2.01, p = .001) than when it was absent (color: M = 0.83, SD = 0.75; both: M = 0.48, SD = 0.86; t = -3.55, p = .047), indicating that beat gesture facilitated online resolution of contrastive critical referring expressions. Another simple-effect analysis revealed a greater likelihood of fixating competitor objects during both-difference than color-contrast critical referring expressions when CPA was present (both: M = -0.59, SD = 0.76; color: M = -0.95, SD = 0.76; t = 4.49, p < .001) than when it was absent (both: M = -0.75, SD = 0.84; color: M = -0.89, SD = 0.69; t < 1), indicating that CPA contributed to incorrect contrastive interpretation of non-contrastive critical referring expressions. Together, these results indicate that beat gesture and CPA serve as cues to contrast during online reference resolution in spoken discourse. Further, the absence of any significant Gesture x Accent interactions indicates that the effects of these cues are additive rather than interactive.

Discussion

Consistent with our predictions, the results indicate that the effects of beat gesture and CPA vary by contrast type during online reference resolution in spoken discourse. Specifically, beat gesture encouraged fixations on target objects during resolution of color-contrast critical referring expressions, confirming that beat gesture can convey contrast effectively. Moreover, CPA encouraged fixations on competitor objects during resolution of both-contrast critical referring expressions, indicating that it acted as a "garden path" resulting in an incorrect contrastive interpretation. Together, these results indicate that beat gesture and CPA each encourage contrastive resolution of referring expressions during online spoken discourse processing. By providing the

first evidence that beat gesture facilitates online resolution of contrastive referring expressions, the results of the current study build upon previous findings that beat gesture (Kushch & Prieto, 2016; Llanes-Coromina et al., 2018; Morett & Fraundorf, under review; Morett, Roche, Fraundorf, & McPartland, 2018) and CPA (Fraundorf et al., 2010; 2012; Lee & Fraundorf, 2016; Lee & Snedeker, 2016; Sanford, Sanford, Molle, & Emmott, 2006) enhance processing and memory of contrastive information in spoken discourse.

Considered in conjunction with the separate interactions with contrast discussed above, the lack of significant interactions between beat gesture and CPA indicates that these cues exert independent, additive effects on online contrastive reference resolution. This finding differs from work on discourse memory, which has shown interactive effects of beat gesture and CPA (Kushch & Prieto, 2016; Llanes-Coromina et al., 2018; Morett & Fraundorf, under review). Although the reasons for this difference are not entirely clear, one possibility is that separate effects of these cues on contrastive information processing in spoken discourse interact during storage or retrieval, leading to the interactive effects observed in studies of offline processing. Another possibility is that effects of these cues that appear separate in the short-term become interactive in the long-term. Future research should distinguish between these possibilities by introducing a delay during which recollection either is or is not required for discourses containing contrastive information in which these cues are varied.

It is worth noting that beat gesture and CPA affected target and competitor object fixations during the shape word but not the color word period, indicating that these cues affect resolution—but not anticipation—of referents. The timing of the effects of these cues is consistent with some previous work examining the effect of CPA on reference resolution (Ito & Speer, 2008), but is inconsistent with other work examining this same phenomenon (Kurumada et al., 2014; Watson, Tanenhaus, & Gunlogson, 2008) as well as work examining the effect of representational gesture on reference resolution (L. B. Silverman et al., 2010). While the reasons for the absence of effects of beat gesture and CPA on reference anticipation in the current study are not entirely clear, one possibility is that these cues may have elicited a processing cost, increasing reference resolution latency. This possibility is consistent with pupillometry data from the current study (Morett, Roche, Fraundorf, & McPartland, 2018), which indicates that the combination of beat gesture and CPA increases cognitive load during reference resolution. Alternatively, fixations to the video may have persisted during color word processing, reducing fixations to target and competitor objects. This possibility is consistent with the results of an analysis of target fixations during the color word period in which the video was included as an interest area, in which the significant main effect of contrast observed during this interest period in the model excluding the video interest area was absent. Thus, both cognitive load and persistence of fixations on the video may have contributed to the timing of the effects of beat gesture and CPA on online reference resolution in the current study.

In conclusion, the results of the current study indicate that beat gesture and CPA exert independent, additive effects that facilitate online contrastive reference resolution during spoken discourse processing. As such, they provide the first evidence that beat gesture is interpreted as a cue to contrast during online spoken discourse comprehension.

Acknowledgments

This research was funded by a Hilibrand Postdoctoral Fellowship to L.M.M. and NIMH R01 MH107426 to J.C.M. The authors thank Talena Day, Kathryn McNaughton, and Zachary Williams for their help with data collection.

References

Austin, E. E., & Sweller, N. (2014). Presentation and production: The role of gesture in spatial communication. *Journal of Experimental Child Psychology*, *122*, 92-103.

Barr, D. J. (2008). Analyzing 'visual world' eyetracking data using multilevel logistic regression. *Journal of Memory and Language*, *59*, 457-474.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48.

Beckman, M. E., & Elam, G. A. (1997). Guidelines for ToBI labelling. *The OSU Research Foundation*, *3*, 30.

Braun, B., & Tagliapietra, L. (2010). The role of contrastive intonation contours in the retrieval of contextual alternatives. *Language and Cognitive Processes*, *25*, 1024–1043.

Esteve-Gibert, N., & Prieto, P. (2013). Prosodic structure shapes the temporal realization of intonation and manual gesture movements. *Journal of Speech, Language, and Hearing Research*, *56*, 850–864.

Fraundorf, S. H., Watson, D. G., & Benjamin, A. S. (2010). Recognition memory reveals just how CONTRASTIVE contrastive accenting really is. *Journal of Memory and Language*, *63*, 367–386.

Fraundorf, S. H., Watson, D. G., & Benjamin, A. S. (2012). The effects of age on the strategic use of pitch accents in memory for discourse: a processing-resource account. *Psychology and Aging*, *27*, 88–98.

Huettig, F., Rommers, J., & Meyer, A. S. (2011). Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta Psychologica*, *137*, 151-171.

Igualada, A., Esteve-Gibert, N., & Prieto, P. (2017). Beat gestures improve word recall in 3-to 5-year-old children. *Journal of Experimental Child Psychology, 156,* 99–112.

Ito, K., Jincho, N., Minai, U., Yamane, N., & Mazuka, R. (2012). Intonation facilitates contrast resolution: evidence from Japanese adults and 6-year olds. *Journal of Memory and Language*, *66*, 265–284.

Ito, K., & Speer, S. R. (2008). Anticipatory effects of intonation: Eye movements during instructed visual search. *Journal of Memory and Language, 58,* 541-573.

Krahmer, E., & Swerts, M. (2007). The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception. *Journal of Memory and Language*, *57*, 396–414.

Kuznetsova, A. Brockhoff, P.B., & Christensen R.H.B. *lmerTest: Tests for random and fixed effects for linear mixed effect models*. R package version 2.0-0. www.r-project.org (December 2018, date last accessed).

Kurumada, C., Brown, M., Bibyk, S., Pontillo, D. F., & Tanenhaus, M. K. (2014). Is it or isn't it: Listeners make rapid use of prosody to infer speaker meanings. *Cognition*, *133*, 335-342.

Kushch, O., Igualada, A., & Prieto, P. (2018). Prominence in speech and gesture favour second language novel word learning. *Language, Cognition and Neuroscience,* 1–13.

Kushch, O., & Prieto, P. (2016). The effects of pitch accentuation and beat gestures on information recall in contrastive discourse. In *Speech Prosody 2016* (pp. 922–925). Boston: Int'l Speech Communication Association.

Ladd, D. R. (1996). *Intonational phonology*. New York: Cambridge University Press.

Lee, E.-K., & Fraundorf, S. H. (2016). Effects of contrastive accents in memory for L2 discourse. *Bilingualism: Language and Cognition*, 1–17.

Lee, E.-K., & Snedeker, J. (2016). Effects of contrastive accents on children's discourse comprehension. *Psychonomic Bulletin & Review, 23,* 1589-1595.

Leonard, T., & Cummins, F. (2011). The temporal relation between beat gestures and speech. *Language and Cognitive Processes, 26,* 1457–1471.

Llanes-Coromina, J., Vilà-Giménez, I., Kushch, O., Borràs-Comes, J., & Prieto, P. (2018). Beat gestures help preschoolers recall and comprehend discourse information. *Journal of Experimental Child Psychology, 172,* 168–188.

McNeill, D. (1992). *Hand and mind.* Chicago: University of Chicago Press.

McNeill, D. (2005). *Gesture and thought.* Chicago: University of Chicago Press.

McNeill, D. (2006). Gesture: A psycholinguistic approach. In E. Brown & A. Anderson (eds.), *The Encyclopedia of Language and Linguistics* (pp. 58–66). Boston: Elsevier.

Morrel-Samuels, P., & Krauss, R. M. (1992). Word familiarity predicts temporal asynchrony of hand gestures and speech. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 615.

Morett, L. M. (2014). When hands speak louder than words: The role of gesture in the communication, encoding, and recall of words in a novel second language. *The Modern Language Journal*, *98*, 834–853.

Morett, L. M. & Fraundorf, S. H. (under review). Beat gesture alters how pitch accenting affects discourse memory: Evidence from top-down use of talker expectations.

Morett, L. M., Roche, J.M., Fraundorf, S. H., & McPartland, J.C. (2018). Pupillometry and multimodal processing of beat gesture and pitch accent: The eye's hole is greater than the sum of its parts. In C. Kalish, M. Rau, J. Zhu, & T. Rogers (Eds.), *Proceedings of the 40th Annual Meeting of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Myhill, J., & Xing, J. Z. (1996). Towards an operational definition of discourse contrast. *Studies in Language*, *20*, 303-360.

Roustan, B., & Dohen, M. (2010). Co-production of contrastive prosodic focus and manual gestures: Temporal coordination and effects on the acoustic and articulatory correlates of focus. Presented at *5th International Conference on Speech Prosody*, Chicago, IL.

Salverda, A. P., Brown, M., & Tanenhaus, M. K. (2011). A goal-based perspective on eye movements in visual world studies. *Acta Psychologica*, *137*, 172-180.

Sanford, A. J., Sanford, A. J., Molle, J., & Emmott, C. (2006). Shallow processing and attention capture in written and spoken discourse. *Discourse Processes*, *42*, 109–130.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., & Hirschberg, J. (1992). ToBI: A standard for labeling English prosody. In *Proceedings of the Third International Conference on Spoken Language Processing* (pp. 867-870).

Silverman, L. B., Bennetto, L., Campana, E., & Tanenhaus, M. K. (2010). Speech-and-gesture integration in high functioning autism. *Cognition, 115,* 380-393.

Vilà-Giménez, I., Igualada, A., & Prieto, P. (in press). Observing storytellers who use rhythmic beat gestures improves children's narrative discourse performance. *Developmental Psychology.*

Wang, L., & Chu, M. (2013). The role of beat gesture and pitch accent in semantic processing: An ERP study. *Neuropsychologia*, *51*, 2847–2855.

Watson, D. G., Tanenhaus, M. K., & Gunlogson, C. A. (2008). Interpreting pitch accents in online comprehension: H* vs. L+ H. *Cognitive Science, 32,* 1232-1244.

Weber, A., Braun, B., & Crocker, M. W. (2006). Finding referents in time: Eye-tracking evidence for the role of contrastive accents. *Language and Speech*, *49*, 367–392.

A Mechanistic Account of Constraints on Control-Dependent Processing: Shared Representation, Conflict and Persistence

Sebastian Musslick[1,*], **and Jonathan D. Cohen**[1]

[1]Princeton Neuroscience Institute, Princeton University, Princeton, NJ 08544, USA.

[*]Corresponding Author: *musslick@princeton.edu*

Abstract

One of the most fundamental and striking limitations of human cognitive function is the constraint on the number of control-dependent processes that can be executed simultaneously. However, the sources of this capacity constraint remain largely unexplored. Previous work has attributed the constraints on control-dependent processing to the sharing of representations between tasks in neural systems. Here, we examine how shared representations interact with two other factors in producing constraints on control-dependent processing. We first demonstrate that the detrimental effects of shared representations on multitasking performance are contingent on the amount of conflict that is induced by the tasks that share representations. We then examine how the persistence of shared representations between tasks affects processing interference during serial task execution. Finally, we discuss how this set of mechanisms can account for various phenomena in neural architectures, including the psychological refractory period, task switch costs, as well as constraints on cognitive control.

Keywords: cognitive control; capacity constraint; dual-tasking; psychological refractory period; neural networks

Introduction

Despite the powerful abilities that cognitive control affords, and its ubiquitous engagement in daily life (e.g., mentally planning a grocery list, or navigating a new route to work), the capacity for controlled processing appears to be strikingly limited (e.g., the inability to plan and navigate at once). This limitation has been literally paradigmatic in defining cognitive control: it has been used to distinguish it from automatic processing (Posner & Snyder, 1975; Shiffrin & Schneider, 1977), and is used universally to operationalize it in the laboratory (i.e., diagnose it experimentally) in the form of dual task interference (Meyer & Kieras, 1997a; Welford, 1952).

A widely accepted view is that constraints in the capacity for control-dependent processing arise from structural limitations inherent to the control system itself. One of the earliest, and still most influential views, is that cognitive control relies on a centralized, limited capacity mechanism that imposes a seriality constraint on processing (Posner & Snyder, 1975; Shiffrin & Schneider, 1977). However, alternative ("multiple-resource") accounts (Allport, 1980; Meyer & Kieras, 1997a; Navon & Gopher, 1979; Salvucci & Taatgen, 2008) have suggested that the capacity constraints reflect properties of the processes that are being controlled. This proposes that control-demanding tasks, like any others, rely on a constellation of "local" resources; that is, task-specific representations, and that the inability to perform more than one task at a time may reflect the conflict that arises when the tasks involved demand that the same set of representations be used for different purposes, rather than reliance on a single centralized control mechanism. From this perspective, the very purpose of cognitive control is to prevent interference by limiting the number of task processes that make use of shared representations (Cohen, Dunbar, & McClelland, 1990; Botvinick, Braver, Barch, Carter, & Cohen, 2001).

One may argue that the constraints that shared representations between tasks impose on multitasking are negligibly small in a processing system as large as the human brain. However, simulation studies (Feng, Schwemmer, Gershman, & Cohen, 2014), followed by analytic work (Musslick et al., 2016) have studied the multitasking capability of two-layer neural networks as a function of the sharing of representations among tasks and found that the multitasking capability of a network drops precipitously with an increase in shared representations, and is virtually invariant to network size. Moreover, neural architectures appear subject to a tradeoff between learning efficiency and generalization that is promoted through the use of shared task representations, on the one hand, and processing efficiency and multitasking capability that is achieved through the separation of task representations, on the other hand (Musslick et al., 2017). This suggests that limitations in multitasking may reflect a preference of the neural system to learn tasks more quickly (Musslick et al., 2017; Sagiv, Musslick, Niv, & Cohen, 2018).

The studies above were based on the assumption that shared representations between tasks always cause interference. However, the amount of processing interference received by a single task has been shown to depend on the processing strength (automaticity) of the interfering task (Cohen et al., 1990; MacLeod & Dunbar, 1988). Another assumption made by these neural network studies is that multitasking can only be achieved by processing tasks concurrently. However, this assumption does not capture processing interference observed in the sequential execution of multiple tasks (Pashler, 1984; Welford, 1952), task switching effects (Alport, Styles, & Hsieh, 1994), nor multitasking behavior along a continuum from pure parallelism, through rapid task switching, to pure sequential processing (Salvucci, Taatgen, & Borst, 2009) .

In this work, we examine the interactive effect of (a) shared representations between tasks, (b) the conflict induced by

shared representations and (c) the persistence of representations on the constraints on control-dependent processing in two-layered, feed-forward, non-linear networks. Our findings suggest that the detrimental effect of shared representations on multitasking interference is only present if the tasks that share representations induce a sufficient amount of conflict between each other, and that persistence of those representations can lead to delays in the serial execution of two tasks. Finally, we discuss how this set of mechanisms may provide a unifying account of various cognitive phenomena in neural architectures, including the psychological refractory period, task switch costs, as well as constraints on cognitive control.

Neural Network Model

For the simulations described in the paper we focus on a network architecture that has been used to simulate a wide array of empirical findings concerning human performance (e.g. Cohen et al., 1990; Gilbert & Shallice, 2002), including limitations in multitasking (Musslick et al., 2016). In this section we lay out the architecture of this network, its processing, as well as the task environments used to train it.

Network Architecture and Processing

The network consisted of the following layers (Figure 1): an input layer with two partitions, one of which represented the current stimulus (nine units) and projected to an associative layer, and another that encoded the current task (five units) and projected to both the associative and output layers; an associative layer (100 units) that projected to the output layer; and an output layer (nine units) that represented the network's response. Input units were grouped by the stimulus dimensions relevant to performing each task (three units per dimension), and used a one-hot encoding (i.e., a single unit in a stimulus dimension was used to represent the current stimulus feature; the current stimulus feature was clamped to 1 and all others were clamped to 0). The task input units used a similar one-hot encoding, with one unit used to represent each task. Output units were grouped by response dimensions, and trained (see below) using a one-hot encoding for each response within a dimension. Each response dimension of the output layer projected to a leaky competitive accumulator (LCA, Usher & McClelland, 2001) layer (described below), which determined the response for that dimension.

The network was instructed to perform a given task by specifying the current stimulus and task to be performed in the input layer. These stimulus and task input values were multiplied by a matrix of connection weights from each partition of the input layer to a shared associative layer, and then passed through a logistic function to determine the pattern of activity over the units in the associative layer. This pattern was then used (together with the set of direct projections from the task layer) to determine the pattern of activity over the output layer.

The final response within a given response dimension of the network was determined by an LCA (Usher & McClelland, 2001) layer, implementing the assumption that the network could only provide one response per dimension (e.g. the network cannot say RED and GREEN at the same time). One LCA layer was assigned to each response dimension k, which was comprised of a set of units r_i that received as their input the activity of corresponding units in that response dimension. The winning response was determined by the accumulation of activity by each LCA unit, and the competition among them, the dynamics of which were given by

$$dr_i = [y_o - \lambda r_i + \alpha f(r_i) - \beta \sum_{j \neq i} f(r_j)] \frac{dt}{\tau} + \xi_i \sqrt{\frac{dt}{\tau}} \qquad (1)$$

where y_o is the activity of the corresponding response unit in response dimension k, λ is the decay rate of r_i, α is the recurrent excitation weight of r_i, β is the inhibition weight between LCA units, τ is the rate constant, and ξ is noise sampled from a Gaussian distribution with zero mean and standard deviation σ. The activity of each LCA response unit was lower bounded by zero via a threshold such that $f(r_i) = r_i$ for $r_i \geq 0$ and $f(r_i) = 0$ for $r < 0$. The response for response dimension k was determined by the unit within the corresponding LCA layer, the activity $f(r_i)$ of which first reached threshold z. The accuracy for each response dimension k corresponded to the probability of generating the correct response for that dimension $P(\text{correct})_k$ across 100 simulations of the LCA, and the reaction time (RT) for that dimension was the average number of time steps required for the response to reach threshold, scaled by a factor of 0.1. The following parameter values were used for all reported simulations: $\lambda = 0.4$, $\alpha = 0.2$, $\beta = 0.2$, $\sigma = 0.1$, and z for each LCA layer was chosen as the threshold that maximizes reward rate ($P(\text{correct})_k / (\text{ITI} + \text{RT}_k)$) for that dimension, where ITI corresponds to an inter-trial interval of 1s.

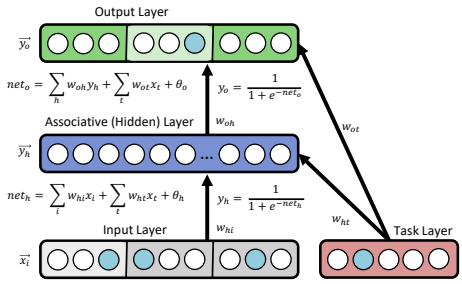


Figure 1: Feedforward neural network used in simulations. The input layer is composed of stimulus vector $\overrightarrow{x_i}$ and task vector $\overrightarrow{x_t}$. The activity of each element in the associative layer $y_h \in \overrightarrow{y_h}$ is determined by all elements x_i and x_t, and their respective weights w_{hi} and w_{ht} to y_h. Similarly, the activity of each output unit $y_o \in \overrightarrow{y_o}$ is determined by all elements y_h and x_t, and their respective weights w_{oh} and w_{ot} to y_o. A bias of $\theta = -2$ is added to the net input of all units y_h and y_o. Blue shades in the input and output units (circles) correspond to unit values of 1 and illustrate an example input pattern with its respective output pattern: The second task requires the network to map the vector of values in the first three feature units to one out of three output units (white shade).

Task Environment

Stimulus input units are structured according to stimulus dimensions (subvectors of the stimulus pattern), each of which was comprised of three feature units with only one feature unit activated per dimension. A task was defined as a mapping from the three stimulus features of a task-relevant stimulus dimension to three output units of a task-specific response dimension, so that only one of the three relevant output units was permitted to be active (see Fig. 1). For each simulation we considered the tasks A-E shown in Figure 2. Tasks A, B and C each map a different stimulus dimension to a different response dimension. Task D shares a stimulus dimension with Task A and shares a response dimension with Task B. Conversely, Task E shares a stimulus dimension with Task B and shares a response dimension with Task A.

Networks were initialized with a set of small random weights and then trained using the backpropagation algorithm (Rumelhart, Hinton, & Williams, 1986) to produce the task-specified response for each stimulus in each task, while suppressing all other responses (both within the task-relevant output dimension, and all task-irrelevant output dimensions). The network was trained in epochs, with each epoch containing all training patterns in random order. The error term used for training was the mean squared error (MSE) of the pattern of activities in the output layer with respect to the correct (task-determined) output pattern. The weights of the network were adjusted with a learning rate of 0.3 after presenting each training pattern within an epoch (online training) until the network reached an MSE of 0.001.

Shared Representation and Conflict

Multitasking limitations have been attributed to shared representations between tasks as they engender interference. However, the amount of interference introduced by shared representation is known to depend on how much conflict they transmit (Cohen et al., 1990; MacLeod & Dunbar, 1988). To illustrate this, consider the simultaneous execution of Tasks A and B depicted in Figure 2. The network can execute a task by limiting processing to the representations involved for that task. For instance, the network can execute Task A by allocating control to the representation that encodes the task-relevant stimulus features for Task A in the associative layer and to the task-relevant response units for Task A in the output layer. Executing Tasks A & B simultaneously would require allocating control to the representations for both tasks in both layers. However, allocating control to Task A would engage Task D if the two tasks share a representation at the associative layer. Once Task D is engaged, it interferes with Task B at the output layer. Similarly, allocating control to a shared associative representation between Tasks B and E would introduce interference with Task A. Shared representations between Tasks A and D, as well as between Tasks B and E therefore introduce a functional dependence between Tasks A and B (Figure 2; Musslick et al., 2016). In contrast, no such interference is expected when the network performs Tasks A & C at the same time.

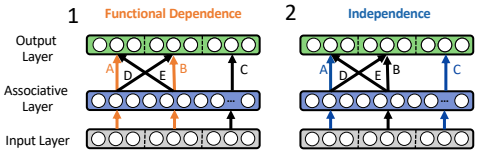


Figure 2: Illustration of dependencies between tasks. (1) Tasks A & B are considered functionally dependent due to shared representations with Tasks D and E, whereas (2) Tasks A & C are considered independent (see text).

In the example above, the amount of conflict introduced by Tasks D and E should decrease if the processing strength of both tasks is weak compared to the processing strength of Tasks A and B. Previous studies have demonstrated that extensive training on a task increases its processing strength which can induce greater conflict with other tasks (Cohen et al., 1990; MacLeod & Dunbar, 1988). This suggests a dilemma: While training on Tasks D and E should improve performance for each individual task, it should also lead to greater interference when dual-tasking seemingly unrelated Tasks A & B. However, dual-tasking performance for the two independent Tasks A & C should be unaffected. Here, we investigated the tradeoff between improvements in single task performance for Tasks D and E, on the one hand, and impairments in dual-task performance for Tasks A & B, as well as Tasks A & C, on the other hand, by varying the amount of training that a network receives for Tasks D and E. We were particularly interested in the amount of training that is required to cause impairments in dual-tasking performance.

We started by initializing 20 networks per training condition. In each condition, we sampled 100 patterns for each of the three Tasks A, B and C per training epoch. However, we varied the number of sampled training patterns for Tasks D and E from 0 (0% task strength) to 150 (150% task strength) across conditions. We then trained every network until it reached performance criterion for Tasks A, B and C. After training, we evaluated whether the network learned shared representations between Tasks A and D, and Tasks B and E in the associative layer of the network. In order to assess the similarity of learned task representations we focus our analysis on the weights from the task units to the associative layer, insofar as these reflect the computations carried out by the network required to perform each task. For a given pair of tasks we compute the learned representational similarity between them as the Pearson correlation of their weight vectors to the associative layer. Finally, we assessed the multitasking accuracy for performing Tasks A & B and the multitasking accuracy for performing Tasks A & C, as well as the single task accuracies of Task D and Task E.

Figure 3.1 shows the correlation between learned task representations in the associative layer of the network, averaged across all networks. As expected, Task A developed a shared representation with Task D in the associative layer since both tasks rely on the same set of stimulus features, as is the case

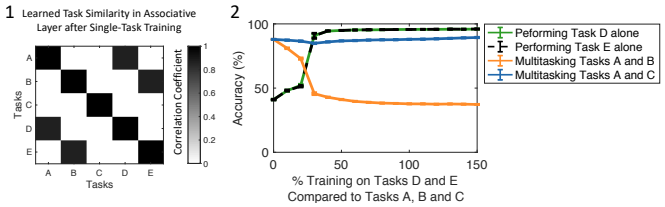


Figure 3: Effects of shared representation and conflict. (1) Average correlations between learned task representations in the associative layer. (2) Multitasking performance for Tasks A & B and Tasks A & C as a function of training on Tasks D and E. Error bars show the standard error of the mean across 20 simulated networks.

for Tasks B and E. Critically, dual-tasking performance for Tasks A & B decreased with the amount of training on Task D and Task E while dual-tasking performance for Tasks A & C was virtually unaffected by the training condition. Even small amounts of training on Tasks D & E (30%) improve performance on these tasks at the expense of impaired multitasking performance of Tasks A & B. Altogether, these results suggest that shared representations alone do not impose constraints on control-dependent processing, but they do so in combination with conflict.

Shared Representation and Persistence

In the network model described above, limitations in multitasking can be circumvented by executing the individual tasks in series. However, a large body of evidence suggests that humans are subject to dual-task interference, even if they execute two tasks one after another (Welford, 1952). To illustrate this, consider the serial execution of two tasks in the psychological response period (PRP) paradigm (Figure 4). A trial in this paradigm begins with the presentation of a stimulus relevant to the first task, followed by a stimulus for the second task. The time between the onset of the first and second stimuli is referred to as stimulus onset asynchrony (SOA) and serves as an independent variable. Participants tend to respond slower to the second stimulus when the SOA is reduced (Welford, 1952). The additional amount of time that it takes to respond to the second task in the presence of a short SOA is referred to as the PRP and serves as the dependent variable.

Symbolic architectures explain the PRP effect in terms of processing bottlenecks that delay execution of the second task while the first task is still being executed (Meyer & Kieras, 1997a; Navon & Gopher, 1979; Salvucci & Taatgen, 2008; Pashler, 1994). While some accounts, such as the EPIC model (Meyer & Kieras, 1997a, 1997b) or the ACT-R/PM model (Byrne & Anderson, 2001) attribute the PRP partly to structural limitations in perceptual processing or motor execution, other accounts claim that the bottleneck is located at a "central" processing stage for response selection (Pashler, 1994) that is preceded by sensory processes and followed by processes for motor execution. However, to date, there is no account of this effect in neural network architectures. For instance, in the feed-forward model considered above, tasks can either be executed concurrently, with the risk of multitasking interference, or in serial, without any risk of interference.

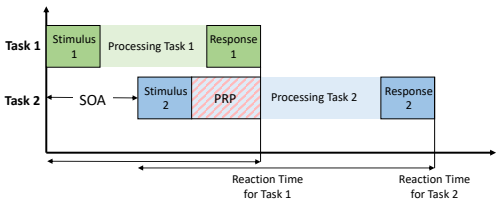


Figure 4: Psychological refractory period paradigm.

A crucial computational feature of neural systems is the integration of information over time, through persisting patterns of activity. Persistence characteristics can account for sequential processing of stimuli (Elman, 1990), working memory (Miyake & Shah, 1999), reconfiguration costs associated with switching tasks (Gilbert & Shallice, 2002) and many other cognitive phenomena. Persistence may also provide a mechanism for how the detrimental effects of shared representation on dual-task interference extend to the sequential execution of two tasks: the more a shared representation of a previously executed task persists in time, the more it may interfere with a subsequent task.

Here, we examine how shared representations interact with the persistence of activity in producing the PRP effect. To examine the PRP effect as a function of both, we first trained 10 networks on Tasks A-E until each network reached the performance criterion across all tasks. After training, we introduced persistence[1] in the computation of the net input of a unit i in the associative and output layers,

$$\overline{net}_i^T = (1-p)\cdot net_i^T + p\cdot \overline{net}_i^{T-1}, \qquad (2)$$

where $\overline{net}_i^{T-1}$ corresponds to the time averaged net input from the previous time step, net_i^T corresponds to the instantaneous net input and p determines how much the time averaged net input of the current time step $\overline{net}_i^T$ depends on the time averaged net input from the previous time step. Thus, the higher p, the longer activity persists over time. For each network, we considered different values for $p \in \{0, 0.5, 0.8, 0.9\}$.

We then simulated the PRP paradigm for two pairs of tasks, A & B, as well as A & C. As demonstrated in the previous section, Tasks A & B are functionally dependent and interfere with each other when executed simultaneously whereas Tasks A & C are independent and interfere less. In both cases, the network was instructed to perform Task A second. Thus, we first presented the network with a feature from the stimulus dimension relevant to Task B or Task C, by activating the corresponding unit in the input layer and by keeping all other input units inactivated. After a number of time steps (determined by the SOA), we presented the network with a feature from the stimulus dimension relevant to the second task (Task A), by activating a unit in the input dimension relevant to that task while the stimulus feature for the first task

[1]Note that persistence in neural networks is typically implemented in the form of recurrent connections between the processing units. Here, we chose, for simplicity, to implement persistence by explicitly integrating processed information over time.

(Task B or Task C) was still present. PRP studies commonly instruct participants to give priority to the first task (Koch, Poljac, Müller, & Kiesel, 2018). We therefore activated the task layer unit for the first task at the beginning of each trial[2] and then determined the optimal onset of the task layer unit for the second task such that the joint reward rate for both tasks is maximized,

$$\text{Reward Rate} = \frac{P(\text{correct})_{\text{first task}} P(\text{correct})_{\text{second task}}}{(\text{ITI} + \text{RT}_{total})} \quad (3)$$

where $P(\text{correct})_{\text{first task}}, P(\text{correct})_{\text{second task}}$ correspond to the accuracies of the first and the second task, respectively, ITI corresponds to an inter-trial interval of 1s , and RT_{total} is the reaction time of the last executed task, measured from the onset of the trial. We then assessed RTs for the first (Task B or Task C) and the second task (Task A) as a function of SOA, by varying the SOA from 1s to 8s in steps of 1s. Finally, we repeated the same analysis for 10 networks that were trained, within each epoch, on 100 patterns of dual-tasking Tasks A & B, as well as 100 patterns for dual-tasking Tasks A & C, in addition to being trained to perform all single tasks as described above. As in the previous section, we also assessed learned representational similarity between tasks as the Pearson correlation of their weight vectors to the associative layer.

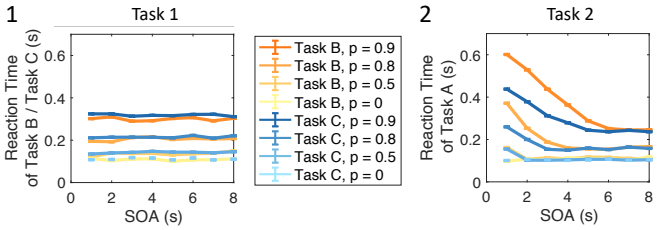


Figure 5: RTs of (1) the first and (2) the second task in the PRP paradigm as a function of persistence p and task. Error bars show the standard error of the mean across 10 simulated networks trained only on single tasks.

Simulation results indicate that higher persistence prolongs the reaction time for both the first and the second task (Figure 5). Moreover, the model replicates the PRP effect, showing a delay of the second task as a function of SOA (Figure 5.2). The delay in RT is overall higher after executing Task B compared to Task C, indicating that Task B interferes more with the subsequently executed Task A. This observation matches simulation results from the previous section, indicating that shared representations between Tasks A & D, as well as Tasks B & E lead to processing interference. However, shorter SOAs still affected RTs for Task A after executing Task C, indicating that there is processing interference between Tasks A & C that is not captured by shared representations in the associative layer alone. Interestingly, higher persistence amplifies the RT difference between Task A followed by a functionally dependent task and Task A followed by an independent task. In line with prior observations (Marill, 1957; Pashler, 1994), the RT of the first task remained unaffected by the SOA, irrespective of whether the first task was functionally dependent or independent of the second task. This observation reflects the embedded strategy of the model to first execute the task associated with the first stimulus. Finally, we observed that dual-task training reduces the amount of shared representation between tasks that rely on a common stimulus dimension (Tasks A and D, as well as Tasks B and E; see Figure 6.1), compared to training the network on single tasks only (cf. Figure 3.1). In addition, training on both dual-task conditions yielded significant reductions in the PRP effect despite high levels of persistence (Figure 6.2). For intermediate levels of persistence ($p \leq 0.5$), dual-task training eliminated the PRP entirely. Such "virtually perfect time sharing" has been observed by Schumacher et al. (2001) after training participants extensively on dual-tasking.

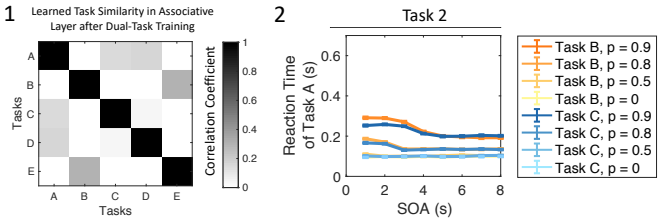


Figure 6: Effects of dual-task training. (1) Average correlations between learned task representations in the associative layer. (2) RT of the second task in the PRP paradigm as a function of persistence p and task. Error bars show the standard error of the mean across 10 simulated networks.

General Discussion and Conclusion

One of the most fundamental limitations of human cognitive behavior is the constraint on the number of control-dependent processes that can be executed simultaneously (Posner & Snyder, 1975; Shiffrin & Schneider, 1977). The multiple-resource hypothesis explains such limitations in terms of shared representations that prevent the interference-free execution of multiple control-demanding tasks (Allport, 1980; Navon & Gopher, 1979). While recent neural network studies provided computational and analytic arguments for the detrimental effects of shared representations on the capacity for control-dependent processing, they were either based on the assumption that shared representations always induce conflict or that tasks can only be executed concurrently (Alon et al., 2017; Feng et al., 2014; Musslick et al., 2016).

In this work, we examined the interactive effect of shared representations and two other factors on limitations associated with control-dependent processing. We first demonstrated that the detrimental effect of shared representations on multitasking interference is present only if the tasks that share representations induce a sufficient amount of conflict. This observation extends previous work, showing that performance of single tasks decreases with the amount of conflict induced by a competing task (Cohen et al., 1990; MacLeod & Dunbar, 1988). In both cases, the conflict induced by the competing task scales with the amount of training on that

[2]We assumed that the task layer unit for the first task becomes deactivated as soon as the model responded to the first stimulus.

task. This suggests that training on a task can improve its performance but may come at the cost of inducing interference with another task that shares a representation.

We also demonstrated that the limitations induced by shared representations can extend to situations in which tasks are executed sequentially. The detrimental effect of shared representation scales with the amount of persistence in the network: the more the representation of a task persists in time, the longer it interferes with other tasks. These observations provide a mechanistic interpretation of the psychological refractory period in neural systems. Symbolic architectures explain this effect in terms of a shared resource that can only be accessed by one task at a time (Anderson, 2013; Navon & Gopher, 1979; Meyer & Kieras, 1997a; Salvucci & Taatgen, 2008). In contrast, the neural network model suggests that tasks may always be processed in parallel but that the outcome of a task process may be strategically delayed to prevent interference from persisting representations of previously executed tasks, yielding a PRP.

The neural network model may also have virtue in explaining findings that central processing bottleneck models struggled to explain. For instance, the second task in the PRP paradigm can be prolonged (relative to single task execution) even if the stimulus for the second task was presented after the participant already responded to the first task (Welford, 1952; Marill, 1957). A central processing bottleneck alone cannot account for a delayed execution of the second task in this situation because a bottleneck should no longer be occupied after executing the first task (Pashler, 1994). The neural network model, however, shows that processing interference induced by shared representation with a previously executed task can persist, irrespective of whether a response for that task has already been generated. Furthermore, modality-specific PRP effects have challenged the notion of a domain-general (amodal) central processing bottleneck: Pairs of tasks with compatible stimulus-response mappings (e.g. a visual-manual task paired with an auditory-vocal task) show greater dual-task interference than two tasks with incompatible stimulus-response mappings (a visual-vocal task paired with an auditory-manual task), lending support to cross-talk models that explain dual-task interference in terms of representational overlap between tasks (Liepelt, Fischer, Frensch, & Schubert, 2011; Hazeltine, Ruthruff, & Remington, 2006). Similarly, our simulation results suggest that functional dependence between tasks induced by representational overlap can lead to higher dual-task interference. Finally, empirical work demonstrated that the PRP can be eliminated with dual-tasking practice, suggesting absence of a central processing bottleneck. (Schumacher et al., 2001). The simulation results presented here suggest that dual-task training may promote the learning of separated, task-dedicated representations that promote interference-free processing.

One of the most robust findings in the cognitive literature is the performance cost associated with the sequential execution of different tasks (Alport et al., 1994). One prominent account of such switch costs is task-set inertia, according to which the task-set of the previously executed task carries over to the next (Alport et al., 1994). Similarly, the findings described here suggest that persistence of task representations lead to a carry over of task interference. The successful sequential execution of two dependent tasks would then afford a temporal switch cost in order to minimize interference-based costs in dual-tasking accuracy. From this perspective, the dependence between tasks induced by shared representation, the amount of conflict, as well as persistence of task representations may all contribute to the performance costs associated with task switches. This suggests that the PRP effect and the costs associated with task switching may originate from the same set of mechanisms in neural systems.

While shared representations may account for limitations in the *number* of control-demanding tasks that can be executed at a time, they do not directly explain limitations in the *amount* of control that can be allocated to a single task (Shenhav et al., 2017). That is, once a commitment has been made to perform a given task (i.e., allocate cognitive control to it), and that precludes the performance of others, then the opportunity cost has already been paid, so why not allocate control maximally to the selected task? Musslick, Jang Jun, Shvartsman, Shenhav, and Cohen (2018) explored the hypothesis that constraints on control intensity (i.e., encoded as cost) reflect, at least in part, an optimal solution to the stability-flexibility dilemma: Allocating more control to a task results in greater activation of its neural representation but also in greater persistence of this activity upon switching to a new task, yielding switch costs. By considering the problem in terms of the parameterization of a nonlinear dynamical system, in which control signals are represented as attractors, Musslick et al. (2018) showed that constraints on the amount of cognitive control allocated to a task can promote cognitive flexibility at the expense of cognitive stability. While this dilemma provides a rationale for why humans should limit the amount of control allocated to a single task it is based on the implicit assumptions that tasks cannot be executed in parallel due to constraints in multitasking capacity and that task representations persist in time. This suggests that both the number of control-demanding tasks that can be executed simultaneously, and the amount of control that can be allocated to a single task may be subject to constraints that arise from (a) the shared use of representation between tasks, (b) the conflict induced by shared representations and (c) persistence of task representations in time.

References

Allport, D. A. (1980). Attention and performance. *Cognitive psychology: New directions*, *1*, 12–153.

Alon, N., Reichman, D., Shinkar, I., Wagner, T., Musslick, S., Cohen, J. D., ... others (2017). A graph-theoretic approach to multitasking. In *NIPS Proceedings* (pp. 2097–2106).

Alport, D., Styles, E., & Hsieh, S. (1994). 17 shifting intentional set: Exploring the dynamic control of tasks.

Anderson, J. R. (2013). *The architecture of cognition*. Psychology Press.

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychol. Rev.*, *108*(3), 624.

Byrne, M. D., & Anderson, J. R. (2001). Serial modules in parallel: The psychological refractory period and perfect time-sharing. *Psychol. Rev.*, *108*(4), 847.

Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: a parallel distributed processing account of the stroop effect. *Psychol. Rev.*, *97*(3), 332–361.

Elman, J. L. (1990). Finding structure in time. *Cognitive science*, *14*(2), 179–211.

Feng, S. F., Schwemmer, M., Gershman, S. J., & Cohen, J. D. (2014). Multitasking vs. multiplexing: toward a normative account of limitations in the simultaneous execution of control-demanding behaviors. *Cogn Affect Behav Neurosci*, *14*(1), 129-146.

Gilbert, S. J., & Shallice, T. (2002). Task switching: A pdp model. *Cognitive Psychology*, *44*(3), 297–337.

Hazeltine, E., Ruthruff, E., & Remington, R. W. (2006). The role of input and output modality pairings in dual-task performance: Evidence for content-dependent central interference. *Cognitive Psychology*, *52*(4), 291–345.

Koch, I., Poljac, E., Müller, H., & Kiesel, A. (2018). Cognitive structure, flexibility, and plasticity in human multitaskingan integrative review of dual-task and task-switching research. *Psychological bulletin*, *144*(6), 557.

Liepelt, R., Fischer, R., Frensch, P. A., & Schubert, T. (2011). Practice-related reduction of dual-task costs under conditions of a manual-pedal response combination. *Journal of Cognitive Psychology*, *23*(1), 29–44.

MacLeod, C. M., & Dunbar, K. (1988). Training and stroop-like interference: Evidence for a continuum of automaticity. *J Exp Psychol Learn Mem Cogn*, *14*(1), 126.

Marill, T. (1957). Psychological refractory phase. *British Journal of Psychology*, *48*(2), 93–97.

Meyer, D. E., & Kieras, D. E. (1997a). A computational theory of executive cognitive processes and multiple-task performance: Part I. Basic mechanisms. *Psychol. Rev.*, *104*(1), 3-65.

Meyer, D. E., & Kieras, D. E. (1997b). A computational theory of executive cognitive processes and multiple-task performance: Part II. accounts of psychological refractory-period phenomena. *Psychol. Rev.*, *104*(4), 749–791.

Miyake, A., & Shah, P. (1999). *Models of working memory: Mechanisms of active maintenance and executive control*. Cambridge University Press.

Musslick, S., Dey, B., Özcimder, K., Patwary, M. M. A., Willke, T. L., & Cohen, J. D. (2016). Controlled vs. automatic processing: A graph-theoretic approach to the analysis of serial vs. parallel processing in neural network architectures. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 1547–1552). Philadelphia, PA.

Musslick, S., Jang Jun, S., Shvartsman, M., Shenhav, A., & Cohen, J. D. (2018). Constraints associated with cognitive control and the stability-flexibility dilemma. In *Proceedings of the 40th Annual Meeting of the Cognitive Science Society* (pp. 806–811). Madison, WI.

Musslick, S., Saxe, A., Özcimder, K., Dey, B., Henselman, G., & Cohen, J. D. (2017). Multitasking capability versus learning efficiency in neural network architectures. In *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 829–834). London, UK.

Navon, D., & Gopher, D. (1979). On the economy of the human-processing system. *Psychol. Rev.*, *86*(3), 214.

Pashler, H. (1984). Processing stages in overlapping tasks: evidence for a central bottleneck. *J Exp Psychol Hum Percept Perform*, *10*(3), 358.

Pashler, H. (1994). Dual-task interference in simple tasks: data and theory. *Psychological bulletin*, *116*(2), 220.

Posner, M., & Snyder, C. (1975). Attention and cognitive control,. In *Information processing and cognition: The loyola symposium* (pp. 55–85).

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, *323*(6088), 533.

Sagiv, Y., Musslick, S., Niv, Y., & Cohen, J. D. (2018). Efficiency of learning vs. processing: Towards a normative theory of multitasking. In *Proceedings of the 40th Annual Meeting of the Cognitive Science Society* (pp. 1004–1009).

Salvucci, D. D., & Taatgen, N. A. (2008). Threaded cognition: an integrated theory of concurrent multitasking. *Psychol. Rev.*, *115*(1), 101.

Salvucci, D. D., Taatgen, N. A., & Borst, J. P. (2009). Toward a unified theory of the multitasking continuum: From concurrent performance to task switching, interruption, and resumption. In *Proceedings of SIGCHI* (pp. 1819–1828).

Schumacher, E. H., Seymour, T. L., Glass, J. M., Fencsik, D. E., Lauber, E. J., Kieras, D. E., & Meyer, D. E. (2001). Virtually perfect time sharing in dual-task performance: Uncorking the central cognitive bottleneck. *Psychological science*, *12*(2), 101–108.

Shenhav, A., Musslick, S., Lieder, F., Kool, W., Griffiths, T. L., Cohen, J. D., & Botvinick, M. M. (2017). Toward a rational and mechanistic account of mental effort. *Annual Review of Neuroscience*, 99–124.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychol. Rev.*, *84*(2), 127.

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: the leaky, competing accumulator model. *Psychol. Rev.*, *108*(3), 550.

Welford, A. T. (1952). The psychological refractory periodand the timing of high-speed performance – a review and a theory. *Br. J. Psychol.*, *43*(1), 2–19.

The effect of stimulus presentation time on bias: A diffusion-model based analysis

Jeremy Ngo (jeremy.ngo@unsw.edu.au),
Christopher Donkin (christopher.donkin@gmail.com)
School of Psychology, UNSW,
Sydney, NSW 2052, Australia

Abstract

There are two main types of bias in simple decision tasks, response bias and stimulus bias. Response bias is a starting level of evidence in favor of a biased response, whereas stimulus bias is the evaluation of stimuli in favor of a biased response. Previous research typically dissociates between these two types of bias. Some studies suggest that it can be difficult to induce response bias without stimulus bias (Ratcliff & McKoon, 2008; van Ravenzwaaij, Mulder, Tuerlinckx, & Wagenmakers, 2012). We used a two-alternative forced-choice brightness discrimination task in which we manipulated the presentation length of the stimuli. We analyzed the data with a hierarchical diffusion model. The results show an overall response bias, as well as stimulus bias that increases as stimulus presentation time decreases. We argue that the results suggest a need to revise how stimulus bias is conceptualized through the drift rate parameter of the diffusion model.

Keywords: diffusion model; response bias; stimulus bias; prior bias; dynamic bias; drift criterion

Introduction

Decision bias is an important area of research because it reveals information about the underlying processes that drive decision making, highlighting how different contexts and goals can influence decision-making behaviour in different ways (White & Poldrack, 2014). Simple decision tasks, where individuals are asked multiple choice questions with only two possible responses, are fairly common in the field of decision making. Research has suggested that there are two distinct types of bias in simple decision tasks: response bias and stimulus bias, also known as prior and dynamic bias, respectively (van Ravenzwaaij et al., 2012; White & Poldrack, 2014). Response bias is a preparedness to make a certain response, whereas stimulus bias is an asymmetry in how two stimuli of equal value/magnitude but opposing valences are processed as evidence for their respective responses.

Decision making can be thought of as sampling information from your environment to build support for a response over time. There have been a number of response time models that have been proposed to formalize this concept. One popular model is the diffusion model (Ratcliff, 1978; Ratcliff & Rouder, 1998; Ratcliff, 2002). To exemplify this model, suppose an observer is tasked with categorizing the stimulus presented in Figure 1a as dark or bright, depending on whether it contains more black or white circles. The diffusion model keeps track of a single quantity of evidence, which reflects the relative amount of accumulated evidence for one choice over the other. This means that in this example, evidence for a 'dark' response counts as evidence against a 'bright' response, illustrated in Figure 1b. Once evidence for one response reaches a boundary, a decision is made.

The basic diffusion model is defined by 4 main parameters that are attributed to different cognitive components that make up the speed and accuracy involved in decision making. These parameters consist of the drift rate, starting point of evidence accumulation, response boundaries, and non-decision time parameters. The non-decision time represents the time taken to perform the processes not directly associated with the evidence accumulation process e.g. motor response to press a button associated with a response. The boundary refers to the amount of evidence required to make a response, and is often characterized as the level of caution the observer has chosen. The starting point of evidence accumulation and drift rate are the two parameters associated with response and stimulus bias respectively.

The starting point parameter is used to represent a baseline level of evidence towards a specific response before stimulus information is accumulated as evidence. Response bias is essentially a shift in the start point parameter, meaning less evidence is required to reach one response boundary compared to the other, as illustrated in Figure 1c. A start point halfway between the two response boundaries indicates no response bias. The drift rate describes the average rate at which evidence is accumulated in favour of one response over the other. Stimulus bias is when one type of stimulus elicits a stronger or weaker drift rate compared to the other type of stimulus. Stimulus bias is illustrated in Figure 1d.

Response bias and stimulus bias both play a large role in decision making, however they are typically presented as independent of each other and dissociable i.e., the preparedness to make a response does not affect the evidence accumulation process. The characterization of these processes as independent confers two main advantages. Firstly, it makes the model more parsimonious. Secondly, it gives a way to account for the different effects that different manipulations have on response times and accuracies.

A number of studies have contributed to this dissociation of response bias and stimulus bias and their associated parameters. The start point can be influenced by the relative frequencies of the presented stimuli and the relative reward rates associated with the stimuli, with limited effects on other parameters (Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006; Diederich & Busemeyer, 2006; Ratcliff & McKoon, 2008; White & Poldrack, 2014).

On the other hand, studies have illustrated that the drift rate is influenced by the quality and discriminability of information presented during a trial (Palmer, Huk, & Shadlen, 2005; Ratcliff & McKoon, 2008; Voss, Rothermund, & Voss, 2004). The standard interpretation for a bias in the drift rate parame-

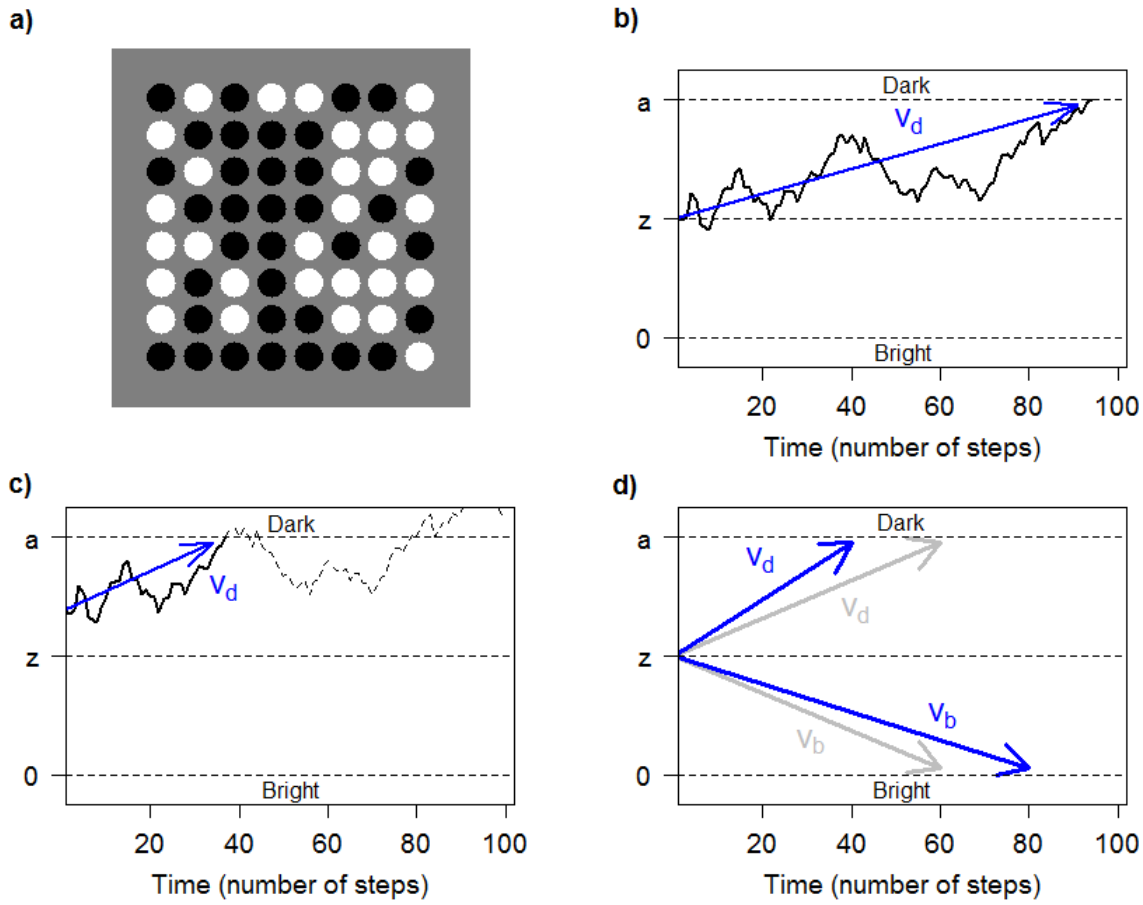


Figure 1: (a) Example of a stimulus where the participant had to decide the stimulus was dark or bright based on the proportions of black and white circles. (b) Diagram of the basic diffusion process. For this example, the top boundary represents the threshold for a 'dark' response, and the bottom boundary is the threshold for a 'bright' response. (c) Effect of a start point bias towards 'dark' responses. There is a shift in the starting point of evidence accumulation such that, given the same course of evidence accumulation observed in Figure 1b, the dark threshold is reached earlier. The dotted line represents the further evidence accumulation if the threshold was not reached. (d) Example of stimulus bias, with V_d and V_b representing drift rates for 'dark' and 'bright' stimuli respectively. Evidence is collected more quickly for the 'dark' response than it is for the 'bright' response. The grey arrows represent drift rates where there is no stimulus bias.

ter is one of criterion setting (Mulder, Wagenmakers, Ratcliff, Boekel, & Forstmann, 2012; Ratcliff, 1985; van Ravenzwaaij et al., 2012). The information observed from a stimulus is compared to a criterion, and the difference between the stimulus information and the criterion yields the evidence value that is to be accumulated in the model. Changing the criterion to permit more evidence for a particular response produces a bias. White, Mumford and Poldrack (2012) demonstrates this in a size discrimination task by showing participants a standard against which upcoming lines should be compared in order to determine what constitutes a 'long' or 'short' response. Their manipulation of this standard selectively influenced a drift criterion parameter in a diffusion model.

There have been studies that used the diffusion model to focus on specifically dissociating these two types of bias. Leite and Ratcliff (2011) examined the effects of stimulus frequency, response payoff, and decision criterion manipulations on start points and drift criterion parameters through a numerosity discrimination task where participants had to decide whether volume of asterisks contained within a 10 by 10 grid could be categorised as a 'low' amount or a 'high' amount, based on some given criteria. They found that changes in the start point parameter alone were able to account for changes in the RT and accuracy data when they manipulated stimulus frequency and payoff. When they manipulated the decision criterion for what was considered 'low' and 'high', they found the data was best fit by shifts in the drift criterion parameter. Similarly, White and Poldrack (2014) used a perceptual discrimination task and a recognition memory task and found that response bias and stimulus bias can be inde-

pendently induced in the diffusion model through the use of stimulus frequency and decision criterion manipulations respectively.

Other research in this field however, has proposed that these biases and the underlying parameters may not be necessarily be independently manipulated. Ratcliff and McKoon (2008) examined the effect of relative frequency and stimulus difficulty manipulations on model parameters using a motion discrimination paradigm. When stimulus difficulty was manipulated, they found that only drift rate varied, however when relative frequency of the stimuli varied, they found a bias in the start point as well as a modest effect on the drift criterion.

Additionally, van Ravenzwaaij et al. (2012) proposed that, theoretically, response bias is sufficient to account for optimal performance in a variable or fixed difficulty task when relative frequency of stimuli is manipulated, but only under certain conditions (cf. Moran, 2015). However, they found that the model fits of empirical data from individuals performing a motion discrimination task show that the relative frequency manipulations had effects on start points and drift criterion in both fixed and variable difficulty tasks.

Rather than being independent of each other, it is possible that base rate information plays a role in moderating how individuals evaluate information under certain circumstances. This provides a potential explanation for why both response bias and stimulus bias were found in studies which manipulated relative frequencies of stimuli. The current experiment aimed to test how response bias and stimulus bias may be expressed under conditions of limited information and differing stimulus frequencies. In doing so, we wanted to observe if this dissociation of these biases and their related parameters holds true. Our experiment empirically evaluated the effect of the relative frequency of stimuli and the duration of stimulus presentation on the parameters of the diffusion model through the use of a hierarchical Bayesian version of the simple diffusion model.

Method

Design

The stimuli used were various combinations of 64 black and white circles in a 8 by 8 grid. In each stimulus, there were 35 circles of one color, and 29 of the other. Participants were instructed to make 'black' or 'white' responses for each stimulus they were presented, indicating which color circle of which there were more. For clarity, stimuli with more black circles will be referred to as 'dark' stimuli and the associated response will be 'dark' responses. Similarly, stimuli with more white circles will be referred to as 'bright' stimuli and the associated response will be 'bright'. An example of a dark stimulus is shown in Figure 1a.

The independent variables manipulated were relative frequencies of stimuli and the presentation length of the stimuli. Relative frequencies of stimuli were manipulated across blocks. This manipulation had three levels, dark biased (two thirds of block were dark stimuli), bright biased (two thirds of block were bright stimuli), and unbiased (even proportions of dark and bright stimuli). There were 13 presentation lengths of stimuli, ranging from 0ms (where no stimulus is shown) to 200ms in 16.7ms (1 frame on a 60 Hz monitor) intervals. The presentation length varied from trial to trial within each block. The experiment consisted of 9 blocks of 80 trials each.

Procedure

At the start of the experiment, participants received instructions on the aim of the task and what they should expect to see on each trial. It was stated that the presentation time of the stimuli will vary within each block. Participants were also told the relative frequencies of each type of stimulus (dark and bright) will differ across blocks and received information about the proportions of dark and bright stimuli at the start of each block. At the end of each block, participants are given an opportunity to take a self-paced break before continuing onto the next block.

At the start of each trial, participants were required to press and hold the spacebar with the index finger of their dominant hand. Once spacebar was held, a fixation cross was presented for 500ms, followed by a mask presented for 100ms, followed by the stimulus. The stimulus is presented for a random duration from 0-200ms, followed by a backward mask of 100ms. There was 16.7ms before the disappearance of the mask and the appearance of the stimulus. Once they were prompted for a response, they had to release the spacebar and indicate a response using the 'F' or 'J' key to indicate whether they thought stimulus was 'dark' or 'bright' using the same finger the held down spacebar with. If they released their finger too early, they received a warning and the experiment would progress to the next trial. These instructions were given to discourage preemptive responses.

After each trial, participants received feedback on screen based on the accuracy of their response. For the 0ms trials where there is no 'correct' response, the feedback for their response was probabilistically determined based on the bias condition for the current block. 'CORRECT' was presented in green if their response was accurate and 'INCORRECT' was presented in red if their response was inaccurate. This feedback was on screen for 750ms before they were allowed to continue to the next trial.

Before the task began, participants were given 12 practice trials consisting of equal numbers of dark and bright stimuli. Each of the presentation lengths, excluding the 0ms presentation length, were used for one of the practice trials.

Model specification

Data were fit using a hierarchical Bayesian version of the simple diffusion model (for more information on hierarchical diffusion models see Vandekerckhove, Tuerlinckx, & Lee, 2011). MCMC estimation was performed through the JAGS Wiener module (Wabersich & Vandekerckhove, 2014) to estimate the parameters by running 3 chains with 5000 iterations each.

Individual-participant level parameters of the diffusion model were assumed to come from Gaussian distributions at the population level. For example, a drift rate for participant i in condition j was modelled as $v_{ij} \sim N(\mu^{v_j}, \lambda^j)$, where μ^{v_j} is the population-level mean drift rate parameter for condition j, and λ^j is the precision of the population-level drift rate parameter for condition j. The priors for the population-level mean parameters were set to be vague and relatively uninformative. For non-decision time, we used a normal distribution of mean 0 and a precision of 100, truncated to be above zero. For the boundary parameter, we used a normal distribution with mean of 3 and a precision of 2, truncated to be above zero. For start-point parameters, we used a uniform distribution from 0 to 1. For drift rate, we used a normal distribution with mean 0 and precision of 1. For the population-level precision parameters, λ, we used gamma distributions with shape and rate parameters of 0.001.

The distance between the boundaries (a), the mean distance of the starting point (z), the average rate of evidence accumulation (v), and the non-decision time parameter (T) were estimated for each individual while also estimated on a population level. The results discussed are the population level parameters estimated by the model. For the purpose of model fitting, dark responses were made when evidence passed the boundary at a and bright responses were made when evidence passed the boundary at 0. This means that higher start points and positive drift rates represent more starting evidence and evidence accumulation for dark responses and lower start points and negative drift rates represent more starting evidence and evidence accumulation for bright responses.

We allowed start points and drift rates to vary freely across all conditions in the experiment, but constrained boundaries and non-decision times to be equal across the three levels of relative frequencies of stimuli conditions. This results in the estimation of 13 boundary parameters and 13 non-decision time parameters (for the each of the trial types), 39 start point parameters (for each trial type across the 3 levels of relative frequencies of stimuli) and 78 drift rate parameters (same as the start point parameters, but estimated separately for the dark stimuli and the bright stimuli). This results in a total of 143 population level parameters. In the following section, we discuss the posterior distributions of the population-level mean parameters.

Results

Figure 2 illustrates posterior distributions of population level start point parameters for each presentation time. A start point closer to 1 and 0 indicates higher starting evidence for dark and bright responses, respectively. When no bias is expected in the start points, a start point of 0.5 is expected. This is what we observed for the unbiased blocks - start points for the unbiased stimulus frequency blocks are distributed around 0.5 across all presentation time conditions, as shown in the green violin plots in Figure 2. From the results of previous experiments that manipulated relative frequencies of stimuli, we expect a bias in the start point in both the dark and bright biased conditions (Leite & Ratcliff, 2011; Ratcliff & McKoon, 2008; van Ravenzwaaij et al., 2012; White & Poldrack, 2014). In the dark biased conditions, we expect start points to be above 0.5 and in bright biased conditions, start points are expected to be below 0.5. Our results are in line with this expectation and are fairly consistent across the different presentation times.

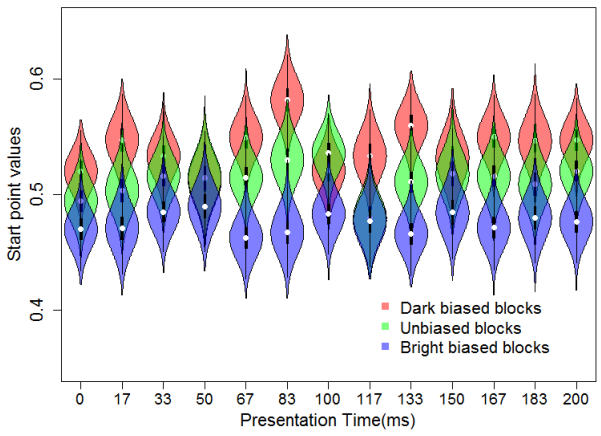


Figure 2: Violin plots of posterior distributions of population level start point parameters for each presentation time.

Regarding the estimates of the drift rate parameters, Since bright and dark stimuli carry the same amount of information (i.e. same proportion of dominant-color circles) we expect the drift rates to have the same magnitude, but in opposite directions. Stimulus bias is calculated as the average drift rate across dark and bright stimuli for a presentation time in a type of block. Since the drift rates for dark and bright stimuli should be equal but with opposite valences, if there is no bias, we expect the average drift rate to be 0.

For the unbiased blocks, the longer a stimulus was presented, the higher the drift rate in the direction of the response associated with that stimulus, with drift rates for short presentation times being distributed around 0, as shown in Figure 3a. This matches our expectations of a higher drift rate when more information (longer presentation time of stimuli) is presented, resulting in limited observed stimulus bias (as shown in the green plots in Figure 3d). For the biased blocks, we observe an overall shift in the drift rates for both bright and dark stimuli, away from 0 and towards the response for the biased stimuli. This is particularly prevalent for the shorter presentation time conditions i.e., for the bright biased blocks, drift rates are in the direction of a bright response when the stimulus is presented for a limited amount of time (0-67ms), regardless of what stimulus was presented (illustrated in Figure 3b). A similar effect is present for the dark biased blocks, illustrated in Figure 3c. Consequently, we found that the stimulus bias observed in the shorter presentation time conditions

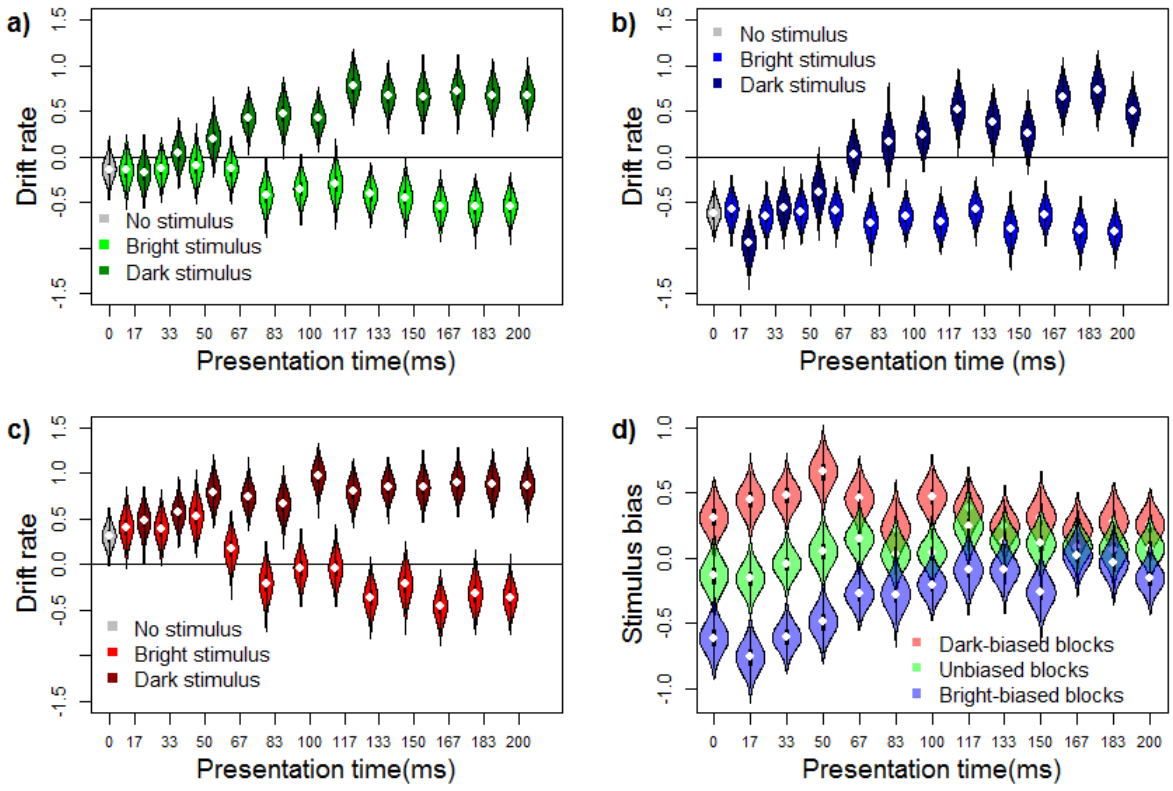


Figure 3: Violin plots of posterior distributions of drift rates for dark and bright stimuli across presentation times for the (a) unbiased, (b) dark biased and (c) bright biased blocks. (d) Average of the dark and bright stimulus drift rates for each presentation time for each block bias type.

was in favor of the biased stimuli for the biased blocks. This also changed as a function of presentation time; we observed a clear trend of stimulus bias decreasing as presentation time increased, summarized in Figure 3d.

Discussion

There are two main findings to take away from the experiment. Firstly, it replicated the response bias effect produced by relative frequencies of stimuli manipulations demonstrated in previous research (Bogacz et al., 2006; van Ravenzwaaij et al., 2012; White & Poldrack, 2014). Secondly, the results from the experiment suggest that stimulus bias has an inverse relationship with presentation time in relative stimulus frequency manipulations. The second finding is particularly interesting because does not coincide with typical interpretations of drift rate and drift rate bias in the diffusion model. If the drift rate reflects the accumulation of information, as stimulus information approaches 0, so too should the drift rate. The results of our current experiment contradict this, continuing to show modest drift rates towards the biased response when there is limited stimulus information.

One possible deviation from this perspective that could explain these results is a model which allows the drift rate to vary across the length of a trial. Its possible that initially, drift rate is driven by biases or sequential effects but is updated as the information from a presented stimulus becomes apparent. Diederich and Busemeyer (2006) discuss a similar concept of a two stage processing model for data from a perceptual discrimination task in which payoffs and deadlines were manipulated. They proposed a model that suggests there are two stages within a trial in which different aspects of the task inform the drift rate. This model suggests that during the first stage, payoff information determined the drift rate but after some period of time, stimulus information takes over. They found that this model was best able to account for the data when compared to two other models, one that allowed boundaries to vary over time, and another that allowed drift rates to vary across time.

On the other hand, a study by Ratcliff and Rouder (2000) manipulated stimulus presentation time in order to examine the concept of non-stationary drift rate in a two choice identification task. They found that a model with a non-stationary drift rate, where the drift rate rose during the onset of a stimulus and then fell to 0 once it was masked, was unable to satisfactorily explain the data. However, a model that used a constant drift rate over time fit the data well, suggesting that there is a constant accumulation of evidence over time even when the stimuli are shown then masked during a trial. In

light of their findings, Ratcliff and Rouder clarify that these findings may not necessarily extend to other domains such as perceptual stimuli (such as the one used in the current experiment) because a cognitive representation may not necessarily be the output of perceptual processing as it is in a letter identification task. Where previous studies focused on purely the onset of a stimulus, none have addressed how response bias may interact with stimulus onset asynchrony.

Another possible explanation is that expectancies or subjective values of responses, more typically reflected in the start point parameter, may moderate how the drift rate is set. The distinction between stimulus and response biases in the diffusion model is analogous to the Bayesian distinction between prior and likelihood. Bogacz, Brown, Moehlis, Holmes and Cohen (2006) argued that the diffusion model is a special case of Wald's (1945) sequential probability ratio test, which is an optimal procedure for deciding between two hypotheses (Wald & Wolfowitz, 1948). Under this equivalence, the start point of evidence accumulation corresponds to the prior probability of the two competing hypotheses (responses). The transformation of information into evidence is carried out by a likelihood function. The posterior probability of the hypotheses are then used as prior probabilities as the next piece of information is to be evaluated. Once the posterior probability of any one hypothesis is large enough, then a response is triggered.

Under the Bayesian framework, a drift rate bias is an adaptation of the likelihood function that is used to transform information from the stimulus into the evidence for competing responses. Our results suggest that the typical interpretation of drift rate bias, the concept of a drift criterion, may not be the whole story. Rather, it seems that the drift rate bias may be also based on what the participant knows about the environment. Usually, such environmental information is assumed to either adjust the prior probability of the different responses, or modify the lens through which stimuli are evaluated. Our data suggest that environmental information may also be accumulated as evidence, at least when the stimulus information is lacking.

Furthermore, some studies have examined how information can be weighted differently in their integration in their response based on their reliability. There has been previous research which show that individuals are able to integrate information from multiple sources, weighing them based on their reliability. (Ernst & Banks, 2002; Fetsch, Pouget, DeAngelis, & Angelaki, 2012; Ohshiro, Angelaki, & DeAngelis, 2011). This has been supported using a modified version of the diffusion model in order to account for the time course of the process (Turner, Gao, Koenig, Palfy, & McClelland, 2017). This further supports the possibility that individuals may be integrating both stimulus information and environmental information when accumulating evidence. When the stimulus is uninformative, individuals may give greater weight to the environmental information which results in the diffusion model showing stimulus bias in the parameter estimates.

When discussing these findings in the context of a diffusion model, it is important to keep in mind that the current set of analyses is a redescription of the observed data through the diffusion model and may not represent the 'true' underlying model. Some alternative models that may be able to account for the results of the current experiment are the leaky, competing accumulator (LCA) model proposed by Usher and McClelland (2001), and Kvam's (2019) theory of bias based on split attention and racing diffusion processes. Usher and McClelland's (2001) LCA model suggests that the observed stimulus bias may be caused by a lateral inhibition between accumulators for the two alternative choices. On the other hand, Kvam (2019) puts forward a model based on a continuous orientation judgement paradigm which suggests that stimulus information and predecision information (such as base rates) compete with each other as separate accumulators and cues can also moderate attention given to a stimulus. Although these models are outside of the scope of this paper, further research in this area should consider these models.

The results of the current experiment highlight that limited information can induce stimulus bias in blocks with uneven stimulus frequencies. Potential avenues for future research include investigating whether this stimulus bias can be induced by other manipulations such as stimulus difficulty or stimulus ambiguity, as well as using other modelling approaches, which may provide alternative explanations for the observed stimulus bias effects. This may help to shed light on the underlying mechanism through which information is processed and how it can result in stimulus bias. Investigating the source of these effects have important implications for understanding how individuals make decisions with different levels of information and may give some deeper insight into the roles of different types of information in decision making.

References

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review*, *113*(4), 700-765.

Diederich, A., & Busemeyer, J. R. (2006). Modeling the effects of payoff on response bias in a perceptual discrimination task: Bound-change, drift-rate-change, or two-stage-processing hypothesis. *Perception & Psychophysics*, *68*(2), 194-207.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*(6870), 429-433.

Fetsch, C. R., Pouget, A., DeAngelis, G. C., & Angelaki, D. E. (2012). Neural correlates of reliability-based cue weighting during multisensory integration. *Nature Neuroscience*, *15*(1), 146-154.

Kvam, P. D. (2019). Modeling accuracy, response time, and bias in continuous orientation judgments. *Journal of Exper-*

imental Psychology: Human Perception and Performance, *45*(3), 301-318.

Leite, F. P., & Ratcliff, R. (2011). What cognitive processes drive response biases? A diffusion model analysis. *Judgment and Decision Making*, *6*(7), 651-687.

Moran, R. (2015). Optimal decision making in heterogeneous and biased environments. *Psychonomic Bulletin & Review*, *22*(1), 38-53.

Mulder, M. J., Wagenmakers, E.-J., Ratcliff, R., Boekel, W., & Forstmann, B. U. (2012). Bias in the brain: A diffusion model analysis of prior probability and potential payoff. *The Journal of Neuroscience*, *32*(7), 2335-2343.

Ohshiro, T., Angelaki, D. E., & DeAngelis, G. C. (2011). A normalization model of multisensory integration. *Nature Neuroscience*, *14*(6), 775-782.

Palmer, J., Huk, A. C., & Shadlen, M. N. (2005). The effect of stimulus strength on the speed and accuracy of a perceptual decision. *Journal of Vision*, *5*(5), 376-404.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59-108.

Ratcliff, R. (1985). Theoretical interpretations of the speed and accuracy of positive and negative responses. *Psychological Review*, *92*(2), 212-215.

Ratcliff, R. (2002). A diffusion model account of response time and accuracy in a brightness discrimination task: Fitting real data and failing to fit fake but plausible data. *Psychonomic Bulletin & Review*, *9*(2), 278-291.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, *20*(4), 873-922.

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science*, *9*(5), 347-356.

Ratcliff, R., & Rouder, J. N. (2000). A diffusion model account of masking in two-choice letter identification. *Journal of Experimental Psychology: Human Perception and Performance*, *26*(1), 127-140.

Turner, B. M., Gao, J., Koenig, S., Palfy, D., & McClelland, J. (2017). The dynamics of multimodal integration: The averaging diffusion model. *Psychonomic Bulletin & Review*, *24*(6), 1819-1843.

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: the leaky, competing accumulator model. *Psychological Review*, *108*(3), 550-592.

Vandekerckhove, J., Tuerlinckx, F., & Lee, M. D. (2011). Hierarchical diffusion models for two-choice response times. *Psychological Methods*, *16*(1), 44-62.

van Ravenzwaaij, D., Mulder, M. J., Tuerlinckx, F., & Wagenmakers, E.-J. (2012). Do the dynamics of prior information depend on task context? an analysis of optimal performance and an empirical test. *Frontiers in Psychology*, *3*(132), 1-15.

Voss, A., Rothermund, K., & Voss, J. (2004). Interpreting the parameters of the diffusion model: An empirical validation. *Memory & Cognition*, *32*(7), 1206-1220.

Wabersich, D., & Vandekerckhove, J. (2014). Extending jags: A tutorial on adding custom distributions to jags (with a diffusion model example). *Behavior Research Methods*, *46*(1), 15-28.

Wald, A. (1945). Sequential tests of statistical hypotheses. *Annals of Mathematical Statistics*, *16*, 117-186.

Wald, A., & Wolfowitz, J. (1948). Optimum character of the sequential probability ratio test. *Annals of Mathematical Statistics*, *19*(3), 326-339.

White, C. N., Mumford, J. A., & Poldrack, R. A. (2012). Perceptual criteria in the human brain. *The Journal of Neuroscience*, *32*(47), 16716-16724.

White, C. N., & Poldrack, R. A. (2014). Decomposing bias in different types of simple decisions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(2), 385-398.

A Resource-Rational Mechanistic Approach to One-shot Non-cooperative Games: The Case of Prisoner's Dilemma

Ardavan S. Nobandegani[1,3], **Kevin da Silva Castanheira**[3], **Thomas R. Shultz**[2,3], **& A. Ross Otto**[3]
{ardavan.salehinobandegani, kevin.dasilvacastanheira}@mail.mcgill.ca
{thomas.shultz, ross.otto}@mcgill.ca
[1]Department of Electrical & Computer Engineering, McGill University
[2]School of Computer Science, McGill University
[3]Department of Psychology, McGill University

Abstract

The concept of Nash equilibrium has played a profound role in economics, and is widely accepted as a normative stance for how people should choose their strategies in competitive environments. However, extensive empirical evidence shows that people often systematically deviate from Nash equilibrium. In this work, we present the first resource-rational mechanistic approach to one-shot, non-cooperative games (ONG), showing that a variant of normative expected-utility maximization acknowledging cognitive limitations can account for important deviations from the prescriptions of Nash equilibrium in ONGs. Concretely, we show that Nobandegani et al.'s (2018) metacognitively-rational model, *sample-based expected utility*, can account for purportedly irrational cooperation rates observed in one-shot, non-cooperative Prisoner's Dilemma, and can accurately explain how cooperation rate varies depending on the parameterization of the game. Additionally, our work provides a resource-rational explanation of why people with higher general intelligence tend to cooperate less in OPDs, and serves as the first (Bayesian) rational, process-level explanation of a well-known violation of the law of total probability in OPDs, documented by Shafir and Tversky (1992), which has resisted explanation by a model governed by classical probability theory for nearly three decades. Surprisingly, our work demonstrates that cooperation can arise from purely selfish, expected-utility maximization subject to cognitive limitations.

Keywords: One-shot non-cooperative games; Nash equilibrium; resource-rational process models; expected utility theory; behavioral game theory; Prisoner's Dilemma; cooperation

1 Introduction

In his seminal work, Nash (1950) introduced a foundational concept of equilibrium, now called "Nash equilibrium," and mathematically proved that any one-shot, non-cooperative, n-player game enjoys (at least) one such equilibrium. In simple terms, Nash equilibrium (NE) is a set of strategies, one for each of the n players of the game, which has the desirable property that each player's strategy is her best response to the strategies adopted by the $n-1$ other players.

Importantly, NE satisfies a number of notable conditions which make it appealing from a normative standpoint. For example, NE passes the key announcement test (Holt & Roth, 2004): If all players publicly announce their strategies, no player would want to reconsider. Furthermore, when the goal is to advise players of a game about which strategies to follow, NE stands out as a rational choice: Any advice that is not an NE would have the unsettling property that there would always be some player(s) who would be better off by deviating from what they are advised (Holt & Roth, 2004). Finally, NE is a self-reinforcing agreement (Holt & Roth, 2004): Once reached by the players, NE does not need any external means of enforcement to endure.

Despite its firm rational grounds, NE has repeatedly failed to provide a descriptively adequate account of human behavior in a variety of important game-theoretic settings (e.g., Mailath, 1998; Goeree & Holt, 2001). By now, extensive empirical evidence shows that people often systematically deviate from Nash equilibrium, thus calling for alternative accounts (e.g., Fehr & Gächter, 2000; Keser & van Winden, 2000; Brandts & Schram, 2001). A prominent example of such violations of NE is the robust empirical finding that people typically cooperate in 2-player, one-shot, non-cooperative, Prisoner's Dilemma (2ONPD) games. Not only does NE prescribe against cooperation in 2ONPD (more precisely, every 2ONPD has only a single NE, and that is for both players to defect), but, more importantly, cooperation is not even *rationalizable* in 2ONPD (Bernheim, 1984; Pearce, 1984) because cooperation is not a best response to any strategy adopted by the other player.

From a purely computational perspective, people's apparent failure to follow the prescriptions of NE is not surprising: Recent theoretical work in computational complexity formally showed that evaluating NE is computationally intractable in general (Daskalakis et al., 2009), and, hence, is generally beyond the capacity of a cognitive system with limited computational power and resources (Simon, 1957).

In this work, we present the first resource-rational mechanistic approach to one-shot, non-cooperative games (ONGs), investigating the extent to which violations of NE could be seen as an optimal response subject to computational and cognitive limitations (Griffiths, Lieder, & Goodman, 2015; Nobandegani, 2017). Concretely, we ask whether these violations can be seen as an optimal behavior with the mind acting as a cognitive miser. To do this, we begin by presenting a general framework allowing us to conceptualize any ONG as a set of risky gambles, thereby reducing the problem of strategy selection in ONGs to a problem of choosing between a set of risky gambles.

To show the efficacy of our framework, we investigate the robust, yet puzzling, experimental finding that people typically cooperate in 2ONPD (e.g., Fehr & Gächter, 2000; Keser & van Winden, 2000; Brandts & Schram, 2001). As we demonstrate, Nobandegani, da Silva Castanheira, Otto, and Shultz's (2018) metacognitively-rational model,

sample-based expected utility (SbEU), not only can provide a resource-rational mechanistic explanation of cooperation behavior in 2ONPD, but also can provide a remarkably accurate quantitative account of how cooperation rate varies depending on the parameterization of 2ONPD (i.e., specific payoffs of the game).

Our paper is organized as follows. After providing a brief overview of SbEU, we present a general framework permitting us to reduce the problem of strategy selection in ONGs to the problem of decision-making under risk. We then turn to modeling cooperation in 2ONPD. Finally, we conclude by discussing the implications of our work for the debate on human rationality.

2 Sample-based Expected Utility Model

Extending the decision-making model of Lieder, Griffiths, and Hsu (2018) to the realm of metacognition, SbEU is a metacognitively-rational process model of risky choice, positing that an agent rationally adapts their strategies depending on the amount of time available for decision-making (Nobandegani et al., 2018). Concretely, SbEU assumes that an agent estimates expected utility

$$\mathbb{E}[u(o)] = \int p(o)u(o)do, \tag{1}$$

using self-normalized importance sampling (Hammersley & Handscomb, 1964; Geweke, 1989), with its importance distribution q^* aiming to optimally minimize mean-squared error (MSE):

$$\hat{E} = \frac{1}{\sum_{j=1}^{s} w_j} \sum_{i=1}^{s} w_i u(o_i), \quad \forall i:\ o_i \sim q^*,\ w_i = \frac{p(o_i)}{q^*(o_i)}, \tag{2}$$

$$q^*(o) \propto p(o)|u(o)|\sqrt{\frac{1+|u(o)|\sqrt{s}}{|u(o)|\sqrt{s}}}. \tag{3}$$

MSE is a standard normative measure of the quality of an estimator, and is widely adopted in machine learning and mathematical statistics (Poor, 2013). In Eqs. (1-3), o denotes an outcome of a risky gamble, $p(o)$ the objective probability of outcome o, $u(o)$ the subjective utility of outcome o, $\hat{E}$ the importance-sampling estimate of expected utility given in Eq. (1), q^* the importance-sampling distribution, o_i an outcome randomly sampled from q^*, and s the number of samples drawn from q^*.

SbEU posits that, when choosing between a pair of risky gambles A, B, people make their choice depending on whether the expected value of the utility difference $\Delta u(o)$ is negative or positive (w.p. stands for "with probability"):

$$A = \begin{cases} o_A & \text{w.p. } P_A \\ 0 & \text{w.p. } 1-P_A \end{cases} \tag{4}$$

$$B = \begin{cases} o_B & \text{w.p. } P_B \\ 0 & \text{w.p. } 1-P_B \end{cases} \tag{5}$$

$$\Delta u(o) = \begin{cases} u(o_A)-u(o_B) & \text{w.p. } P_A P_B \\ u(o_A)-u(0) & \text{w.p. } P_A(1-P_B) \\ u(0)-u(o_B) & \text{w.p. } (1-P_A)P_B \\ 0 & \text{w.p. } (1-P_A)(1-P_B) \end{cases} \tag{6}$$

In Eq. (6), $u(\cdot)$ denotes the subjective utility function of a decision-maker. Following Nobandegani et al. (2018), and consistent with prospect theory (Kahneman & Tversky, 1979; Tversky & Kahneman, 1992), in this paper we assume a standard S-shaped utility function $u(x)$ given by:

$$u(x) = \begin{cases} x^{0.85} & \text{if } x \geq 0, \\ -|x|^{0.95} & \text{if } x < 0. \end{cases} \tag{7}$$

Nobandegani et al. (2018) recently revealed that SbEU provides an account of the availability bias, the tendency to overestimate the probability of events that easily come to mind (Tversky & Kahneman, 1973), and can accurately simulate the well-known fourfold pattern of risk preferences in outcome probability (Tversky & Kahneman, 1992) and in outcome magnitude (Markovitz, 1952; Scholten & Read, 2014). Notably, SbEU is the first rational process model to score near-perfectly in optimality, economical use of limited cognitive resources, and robustness, all at the same time (see Nobandegani et al., 2018; Nobandegani et al., 2019a).

3 From One-shot, Non-cooperative Games to Multi-alternative Risky Choice

In this section, we present a general framework allowing us to conceptualize any ONG as a set of risky gambles $\mathcal{S}$. Importantly, this framework permits us to reduce the problem of strategy selection in ONGs to the problem of risky decision-making. By re-framing the problem this way, the strategy selected by an agent in an ONG corresponds to the risky gamble that the agent would choose among the set of available gambles $\mathcal{S}$.[1]

Without loss of generality, and for ease of exposition, we consider the case of a 2-player, one-shot, non-cooperative game (2ONG) here. Extending the results to the general case of an n-player, one-shot, non-cooperative game is straightforward.

Consider a generic 2ONG whose payoff matrix is given in Fig. 1. The game has two players: Player 1 (Row Player) and Player 2 (Column Player). Player 1 has two pure strategies to choose between: the strategy corresponding to choosing the top row (Top Strategy) and the strategy corresponding to choosing the bottom row (Bottom Strategy). Similarly, Player 2 has two pure strategies to choose from: the strategy corresponding to choosing the left column (Left Strategy) and the strategy corresponding to choosing the right column (Right Strategy). From the perspective of Player 1, Player 2 selects the Left Strategy with probability P_l, and the

[1] We should note that our framework naturally handles "mixed strategies" wherein the agent probabilistically chooses among the set of possible "pure strategies." The validity of this claim follows from the key understanding that the choice between the set of available gambles $\mathcal{S}$ would be also made probabilistically.

		Player 2 (Column Player)	
		Left	Right
Player 1 (Row Player)	Top	a, x	b, v
	Bottom	c, y	d, w

Figure 1: Payoff matrix for a generic 2-player, one-shot, non-cooperative game (2ONG). For example, if Player 1 (Row Player) selects the Top Strategy and Player 2 (Column Player) selects the Left Strategy, Player 1 and Player 2 receive payoffs a and x, respectively.

Right Strategy with probability $P_r = 1 - P_l$. Likewise, from the perspective of Player 2, Player 1 selects the Top Strategy with probability P_t, and the Bottom Strategy with probability $P_b = 1 - P_t$. As such, Player 1 is essentially choosing between the two gambles T and B:

$$T = \begin{cases} a & \text{w.p. } P_l \\ b & \text{w.p. } 1 - P_l \end{cases} \tag{8}$$

$$B = \begin{cases} c & \text{w.p. } P_l \\ d & \text{w.p. } 1 - P_l \end{cases} \tag{9}$$

with gambles T, B corresponding to choosing the Top Strategy and the Bottom Strategy, respectively, and Player 2 is essentially choosing between the two gambles L and R:

$$L = \begin{cases} x & \text{w.p. } P_t \\ y & \text{w.p. } 1 - P_t \end{cases} \tag{10}$$

$$R = \begin{cases} v & \text{w.p. } P_t \\ w & \text{w.p. } 1 - P_t \end{cases} \tag{11}$$

with gambles L, R corresponding to choosing the Left Strategy and the Right Strategy, respectively.

The line of reasoning presented above shows that the problem of strategy selection for a player in 2ONGs can be formally reduced to the problem of deciding between two risky gambles (T, B for Row Player and L, R for Column Player). By the same logic, more generally, the problem of strategy selection for a player in an n-player ONG (with each player having n pure strategies to choose from) can be formally reduced the problem of deciding between n risky gambles.

As evidenced by Eqs. (8-9) depending on the parameter P_l, Player 1's choice between T and B explicitly depends on Player 1's conception of the probability with which Player 2 would select the Left Strategy (i.e., P_l). Likewise, as evidenced by Eqs. (10-11), Player 2's choice between L and R explicitly depends on Player 2's conception of the probability with which Player 1 would select the Top Strategy (i.e., P_t).

As a case-study, in the next section we turn our attention to Prisoner's Dilemma, and we show that, together with the general way of reducing ONGs to risky decision-making discussed above, SbEU can accurately explain cooperation in 2ONPDs, thereby providing a process-level, rational basis for cooperation in 2ONPDs.

4 Cooperation in One-shot, Non-cooperative Prisoner's Dilemma

A wealth of real-life scenarios are modeled as an instance of Prisoner's Dilemma, e.g., conflict of two prisoners independently questioned by the police (Kaminski, 2003), cartel problems (Osborne, 1976), the conflict of two superpowers who engage in a nuclear arms race (Wiesner & York, 1964), doping in sports (Savulescu, Foddy, & Clayton, 2004; Haugen, 2004), and global warming (Milinski et al., 2008).

Although, normatively, one should never cooperate in one-shot, non-cooperative Prisoner's Dilemma games, substantial experimental evidence shows that people typically cooperate in 2ONPDs (e.g., Dawes & Thaler, 1988; Fehr & Gächter, 2000; Keser & van Winden, 2000; Brandts & Schram, 2001).

In a 2ONPD, each player has two strategies to choose from: either to cooperate or to defect. The payoff matrix of a generic 2ONPD is shown in Fig. 2.

		Player 2 (Column Player)	
		Cooperate	Defect
Player 1 (Row Player)	Cooperate	r, r	v, t
	Defect	t, v	p, p

Figure 2: Payoff matrix of a generic 2-player, one-shot, non-cooperative Prisoner's Dilemma (2ONPD), where $t > r > p > v$. For 2ONPDs, the constraint $r > p$ ensures that mutual cooperation is superior to mutual defection, while the constraints $t > r$ and $p > v$ grant that defection is the dominant strategy for both players. Players can either cooperate or defect.

According to the general framework presented in the previous section, assuming that (from the perspective of a player) the other player would cooperate with probability P_c, a player is essentially choosing from the following two risky choices:

$$\text{Cooperate} = \begin{cases} r & \text{w.p. } P_c \\ v & \text{w.p. } 1 - P_c \end{cases}$$

$$\text{Defect} = \begin{cases} t & \text{w.p. } P_c \\ p & \text{w.p. } 1 - P_c \end{cases}$$

According to the normative principle of least-informative priors (i.e., those prior distributions attaining highest entropy), having no priori knowledge of, or any opportunity to learn through interactions about, her opponent—due to the one-shot, non-cooperative nature of the game—it is rationally

justified for a player to assume that $P_c = 0.5$. Accordingly, throughout this paper we make the assumption that $P_c = 0.5$.

Recent work has provided mounting evidence suggesting that people often use very few samples in probabilistic judgments and reasoning (e.g., Vul et al., 2014; Battaglia et al. 2013; Lake et al., 2017; Gershman, Horvitz, & Tenenbaum, 2015; Hertwig & Pleskac, 2010; Griffiths et al., 2012; Gershman, Vul, & Tenenbaum, 2012; Bonawitz et al., 2014; Nobandegani et al., 2018; Lieder, Griffiths, Huys, & Goodman, 2018). Consistent with this finding, throughout this paper we assume that a player draws very few samples ($s = 1$; see Eqs. (2-3)) when deciding between cooperation and defection in 2ONPDs—except for Sec. 4.3 in which we directly investigate the effect of number of samples s on cooperation.

Under these justified assumptions (i.e, $s = 1$ and $P_c = 0.5$), in the following two subsections we show that Nobandegani et al.'s (2018) metacognitively-rational model, SbEU, accurately explains how cooperation rate varies depending on the parameterization of a 2ONPD.

4.1 Manipulation of Cooperation Index

Introduced by Rapoport and Chammah (1965), *cooperation index* (CI) is a concrete measure of cooperativeness in 2ONPDs; CI is a property of the experimental task. For a 2ONPD with a generic payoff matrix shown in Fig. 2, CI is given by (Rapoport & Chammah, 1965):

$$\text{CI} = \frac{r-p}{t-v}. \tag{12}$$

As for any 2ONPD holds $t > r > p > v$ (see Fig. 2), it follows that $0 < \text{CI} < 1$. (The latter result follows from having $t - v > r - p > 0$.)

Rapoport and Chammah (1965) experimentally demonstrated a linear relationship between CI and cooperation rate, with people tending to cooperate more as CI increases. Several studies have replicated this finding (e.g., Steele & Tedeschi, 1967; Vlaev & Chater, 2006).

Next, we show that SbEU can remarkably accurately account for this finding. To test how the cooperation rate predicted by SbEU changes as CI increases, we use nine representative 2ONPD games from Vlaev and Chater (2006, Table 1) which allow us to systematically vary CI equidistantly between 0.1 and 0.9. Recall that $0 < \text{CI} < 1$. We simulate $N = 100,000$ participants, with $s = 1$ and $P_c = 0.5$.

As Fig. 3 demonstrates, there is a significant positive, linear relationship between CI and the cooperation rate predicted by SbEU (Pearson's $r = .9998$, Kendall's $\tau = 1$, Spearman's $\rho = 1$, Ps $< 10^{-5}$).

In the next subsection, we directly compare the cooperation rate predicted by SbEU and human data.

4.2 Manipulation of Defection Payoff p

In a recent experiment investigating the effect of manipulation of defection payoff (i.e., parameter p; see Fig. 2) on cooperation, Engel and Zhurakhovska (2016) presented participants with eleven 2ONPDs; across these stimuli, they systematically varied parameter p while keeping the other parameters fixed (for experimental stimuli see Engel and Zhurakhovska, 2016, Sec. 3).

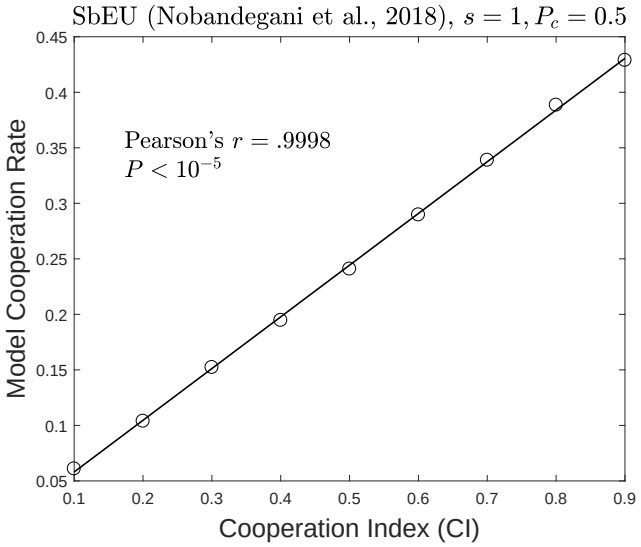


Figure 3: SbEU (Nobandegani et al., 2018) can accurately simulate the linear relationship between CI and cooperation rate, experimentally demonstrated by Rapoport and Chammah (1965).

Fig. 4 shows that SbEU can remarkably accurately account for Engel and Zhurakhovska's (2016) observed cooperation rates, explaining 98% of the variance in the experimental data (Pearson's $r = .9906$, Kendall's $\tau = .9909$, Spearman's $\rho = .9977$, Ps $< .001$). In Fig. 4, we simulate $N = 100,000$ participants, with $s = 1$ and $P_c = 0.5$.

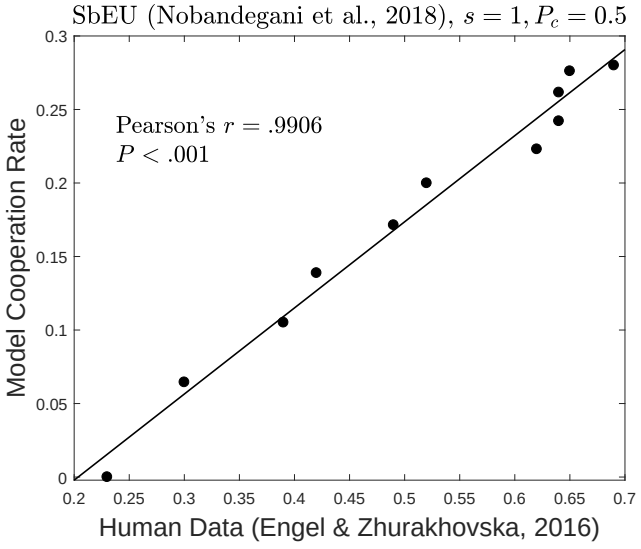


Figure 4: SbEU (Nobandegani et al., 2018) simulates Engel and Zhurakhovska's (2016) experimental data on the effect of manipulation of defection payoff (i.e., parameter p) on cooperation in 2ONPDs.

4.3 The Predictive Relationship between Number of Samples s and Cooperation

In a recent study, Kanazawa and Fontaine (2013) experimentally investigated the effect of general intelligence (measured by a Raven's-type nonverbal test of general intelligence) on cooperation in 2ONPDs, showing that individuals with higher general intelligence are less likely to cooperate.

In this section we investigate the predictive relationship between the number of samples s and cooperation rate in 2ONPDs. In the context of SbEU, we operationalize the well-supported assumption that people with higher general intelligence typically enjoy more cognitive resources, e.g. working memory (e.g., Colom, Jung, & Haier, 2007; Colom et al., 2008, Burgess, Gray, Conway, & Braver, 2011) by positing that these individuals tend to draw more samples.

Consistent with Kanazawa and Fontaine's (2013) finding, SbEU predicts that cooperation rate should decrease as the number of samples s increases; see Fig. 5. In Fig. 5, we adopt the Kanazawa and Fontaine's (2013) specific PD problem given to the subjects (a 2ONPD with $r = 3, v = 0, t = 5, p = 1$), and simulate $N = 100,000$ participants with $P_c = 0.5$.

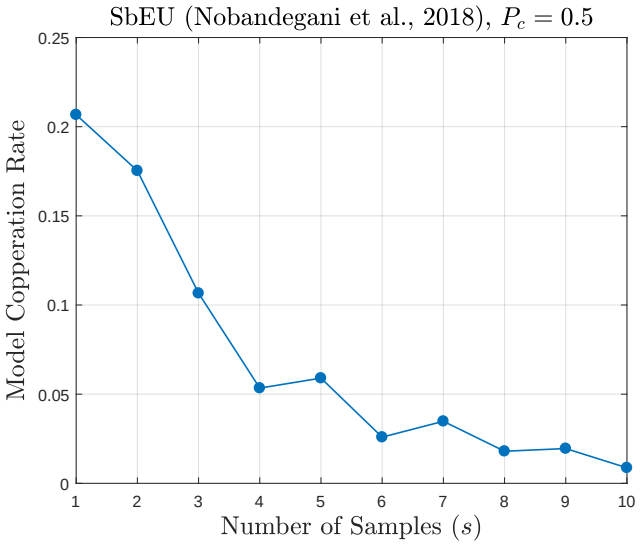


Figure 5: SbEU (Nobandegani et al., 2018) predicts that cooperation rate should decrease as the number of samples s increases, consistent with the experimental findings of Kanazawa and Fontaine (2013).

Importantly, SbEU's prediction depicted in Fig. 5 is supported by substantial evidence revealing that, in the context of 2ONPDs, deliberation (which can be readily operationalized in terms of drawing more samples) leads to increased defection rate, thus bringing behavior closer to the prescriptions of the normative standards of game theory (e.g., Rand, 2016).

4.4 Manipulation of P_c

Shafir and Tversky (1992) examined cooperation rates in a well-known variant of 2ONPD: In some trials, participants were told what the other player was doing. Unsurprisingly, when participants were told that the other person decided to defect, then their probability to cooperate was 0.03; and when they were told that the other person decided to cooperate, then their probability to cooperate was 0.16. However, in trials (within participants design) when participants were not told what the other person did, the probability to cooperate raised to 0.37. This pattern of responding has been independently replicated several times (e.g., Busemeyer, Matthew, & Wang, 2006; Croson, 1999; Li & Taplin, 2002), and has thus far remained a puzzle for optimal decision theorists to explain.

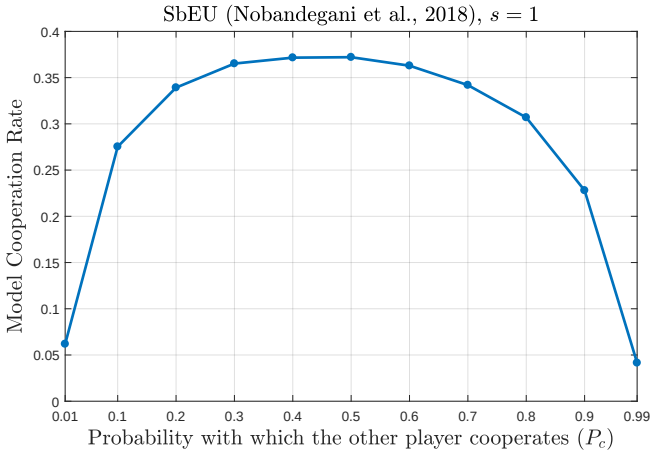


Figure 6: SbEU (Nobandegani et al., 2018) provides a resource-rational, process-level explanation of the puzzling finding of Shafir and Tversky (1992). This finding thus far has defied any (Bayesian) rational explanation. We simulate $N = 100,000$ participants, with $s = 1$. We use a representative 2ONPD game from Shafir and Tversky (1992, Fig. 28.2) with the following parameters: $r = 75, v = 25, t = 85, p = 30$.

The present study offers one, and thus far the only, (Bayesian) rational process-level explanation of this puzzle. As Fig. 6 shows, SbEU predicts that a participant should have only a minuscule tendency to cooperate when the other player is known to either fully cooperate or defect. However, consistent with Shafir and Tversky's (1992) finding, SbEU predicts that participants should have a substantially greater tendency to cooperate when they are maximally uncertain about what the other player would do. As such, SbEU provides a rational explanation of a clear violation of the *law of total probability* in 2ONPDs, as demonstrated by Shafir and Tversky (1992). According to the law of total probability (Durrett, 2010), the cooperation rate under the condition that the opponent's choice is unknown must fall between the cooperation rates observed under the two extreme conditions: full cooperation and full defection.

5 General Discussion

Despite its solid normative ground, NE has failed to provide a satisfying descriptive account of human behavior in many game-theoretic settings. In this work, we focus on a well-

documented, yet puzzling, deviation from NE: cooperation in 2-player, one-shot, non-cooperative, Prisoner's Dilemma games (2ONPDs). By way of introducing a general framework allowing us to conceptualize strategy selection in one-shot, non-cooperative games (ONGs) as the classical problem of decision-making under risk, we investigate whether (seemingly irrational) cooperation in 2ONPDs could be understood as an optimal behavior with the mind acting as a miser.

To our knowledge, our work provides the first (but, admittedly, preliminary) demonstration of how cooperation can arise from *purely selfish*, expected-utility maximization under cognitive limitations. Our findings challenge the widespread view that observed cooperation in 2ONPD games is primarily due to "cooperation bias" in humans, and are supported by recent experimental findings revealing little evidence for such cooperation bias (Pothos et al., 2011). As such, our work refutes the (very intuitive) widely-accepted conclusion that "If players are egoists, cooperation will not be observed in one-shot PD games" (Cooper et al., 1996).

Concretely, in this work we show that the Nobandegani et al.'s (2018) metacognitively-rational process model, SbEU, provides a resource-rational mechanistic explanation of cooperation in 2ONPDs, and offers an accurate quantitative account of how cooperation rate varies depending on the parameterization of a 2ONPD. Furthermore, by operationalizing higher intelligence in terms of drawing a larger number of samples in the available time, our work predicts that more intelligent individuals should tend to cooperate less, fully consistent with recent experimental findings (Kanazawa & Fontaine, 2013).

Shafir and Tversky's (1992) paradoxical finding on the violation of the law of total probability in 2ONPDs has resisted explanation by a model governed by classical probability theory (CPT) for nearly three decades. Interestingly, this paradoxical finding has been recently taken as strong evidence for quantum-probability models of cognition (e.g., Pothos & Busemeyer, 2009; Pothos & Busemeyer, 2013). Our work offers the first, and thus far the only, CPT-based explanation of Shafir and Tversky's (1992) paradoxical finding. As such, our work corroborates the view that decision-making behaviors that appear to be inconsistent with CPT, might after all be reconcilable with CPT when analyzed from an algorithmic perspective acknowledging cognitive limitations.

Being primarily inspired by the experimental finding that deliberation leads to a marked increase in defection rate in 2ONPDs (e.g., Rand, 2016), and applying a dual-process lens to cooperation in 2ONPDs, some researchers have recently argued that intuition favors cooperation while deliberation promotes selfishness (e.g., Rand, Greene, & Nowak, 2012; Rubinstein, 2007; Rand, 2016). Our work offers a completely new way of understanding this experimental finding—both qualitatively and quantitatively.

On the quantitative front, in sharp contrast to a dual-process perspective, our work presents the first, and thus far the only, *single-process* model of cooperation in 2ONPDs, providing a resource-rational mechanistic explanation of why deliberation leads to increased defection. According to our work, it is the optimal use of limited cognitive resources that underlies deliberation promoting selfishness in 2ONPDs. Relatedly, our recent work on modeling fairness in the Ultimatum Game (UG) also supports this view (Nobandegani, Destais, & Shultz, in prep).

On the qualitative side, our work offers a radically different interpretation of cooperation in 2ONPDs than the one provided by the classical dual-process account. From a dual-process perspective, intuition (moderated by System 1) is good and cooperative while deliberation (moderated by System 2) is evil and uncooperative. However, according to our singe-process model (SbEU; Nobandegani et al., 2018), a boundedly-rational agent that selfishly maximizes its expected utility while optimally using its limited cognitive resources should show the highest cooperation rate as an intuitive response, with cooperation rate declining with deliberation. As such, according to our work, humans' intuitive response being to cooperate in 2ONPDs, is still, quite counterintuitively, the effect of selfishly maximizing expected utility while optimally using limited cognitive resources.

Our work contributes to an emerging line of work attempting to explain human cognition as an optimal use of limited cognitive resources (*rational minimalist program*, Nobandegani, 2017; Griffiths, Lieder, & Goodman, 2015), thereby demonstrating that a wide range of human behaviors are rational, provided that the computational and cognitive limitations of the mind are taken into consideration (Simon, 1957).

By demonstrating that SbEU, a recently proposed metacognitively-rational process model of risky choice (Nobandegani et al., 2018), can quantitatively account for ostensibly irrational cooperation rates in 2ONPDs, our work bridges between two related, but distinct, areas of research: game-theoretic decision-making and risky decision-making. As such, the work presented here brings us a step closer to developing a unified, mechanistic account of human decision-making.

Recent work has shown that SbEU can account for the St. Petersburg paradox, a centuries-old paradox in human decision-making (Nobandegani, da Silva Castanheira, Shultz, & Otto, 2019b), and has experimentally confirmed a counterintuitive prediction of SbEU: Deliberation leads people to move from one well-known bias, framing effect, to another well-known bias, the fourfold pattern of risk preferences (da Silva Castanheira; Nobandegani, & Otto, 2019). An important line of future work would be to investigate whether SbEU could also serve as a resource-rational process-level account of contextual effects in multi-attribute decision-making (e.g., the attraction, similarity, and compromise effects), thus bringing us another step closer to developing this unified account of decision-making.

Acknowledgments This work is supported by an operating grant to TRS from the Natural Sciences and Engineering Research Council of Canada. We would like to thank Constance

Destais, Ashley Stendel, Marcel Montrey, and Peter Helfer for helpful comments on an earlier draft of this work.

References

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, *110*(45), 18327–18332.

Bernheim, B. D. (1984). Rationalizable strategic behavior. *Econometrica: Journal of the Econometric Society*, 1007–1028.

Brandts, J., & Schram, A. (2001). Cooperation and noise in public goods experiments: applying the contribution function approach. *Journal of Public Economics*, *79*(2), 399–427.

Burgess, G. C., Gray, J. R., Conway, A. R., & Braver, T. S. (2011). Neural mechanisms of interference control underlie the relationship between fluid intelligence and working memory span. *Journal of Experimental Psychology: General*, *140*(4), 674.

Busemeyer, J. R., Matthew, M. R., & Wang, Z. (2006). A quantum information processing explanation of disjunction effects. In: *Proceedings of the* 28th *Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Colom, R., Abad, F. J., Quiroga, M. Á., Shih, P. C., & Flores-Mendoza, C. (2008). Working memory and intelligence are highly related constructs, but why? *Intelligence*, *36*(6), 584–606.

Colom, R., Jung, R. E., & Haier, R. J. (2007). General intelligence and memory span: evidence for a common neuroanatomic framework. *Cognitive Neuropsychology*, *24*(8), 867–878.

Cooper, R., DeJong, D. V., Forsythe, R., & Ross, T. W. (1996). Cooperation without reputation: experimental evidence from prisoner's dilemma games. *Games and Economic Behavior*, *12*(2), 187–218.

Croson, R. T. (1999). The disjunction effect and reason-based choice in games. *Organizational Behavior and Human Decision Processes*, *80*(2), 118–133.

da Silva Castanheira, K., Nobandegani, A. S., & Otto, A. R. (2019). Sample-based variant of expected utility explains effects of time pressure and individual differences in processing speed on risk preferences. In *Proceedings of the* 41st *Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Daskalakis, C., Goldberg, P. W., & Papadimitriou, C. H. (2009). The complexity of computing a Nash equilibrium. *SIAM Journal on Computing*, *39*(1), 195–259.

Dawes, R. M., & Thaler, R. H. (1988). Anomalies: cooperation. *Journal of Economic Perspectives*, *2*(3), 187–197.

Durrett, R. (2010). *Probability: Theory and Examples*. Cambridge university press.

Engel, C., & Zhurakhovska, L. (2016). When is the risk of cooperation worth taking? The prisoner's dilemma as a game of multiple motives. *Applied Economics Letters*, *23*(16), 1157–1161.

Fehr, E., & Gachter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review*, *90*(4), 980–994.

Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, *349*(6245), 273–278.

Gershman, S. J., Vul, E., & Tenenbaum, J. B. (2012). Multistability and perceptual inference. *Neural Computation*, *24*(1), 1–24.

Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica: Journal of the Econometric Society*, 1317–1339.

Goeree, J. K., & Holt, C. A. (2001). Ten little treasures of game theory and ten intuitive contradictions. *American Economic Review*, *91*(5), 1402–1422.

Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in Cognitive Science*, *7*(2), 217–229.

Griffiths, T. L., Vul, E., & Sanborn, A. N. (2012). Bridging levels of analysis for probabilistic models of cognition. *Current Directions in Psychological Science*, *21*(4), 263–268.

Hammersley, J., & Handscomb, D. (1964). *Monte carlo methods.* London: Methuen & Co Ltd.

Haugen, K. K. (2004). The performance-enhancing drug game. *Journal of Sports Economics*, *5*(1), 67–86.

Hertwig, R., & Pleskac, T. J. (2010). Decisions from experience: Why small samples? *Cognition*, *115*(2), 225–237.

Holt, C. A., & Roth, A. E. (2004). The Nash equilibrium: A perspective. *Proceedings of the National Academy of Sciences*, *101*(12), 3999–4002.

Kahneman, . T. A., D. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263–291.

Kaminski, M. M. (2018). *Games prisoners play: The tragicomic worlds of Polish prison*. Princeton University Press.

Kanazawa, S., & Fontaine, L. (2013). Intelligent people defect more in a one-shot prisoners dilemma game. *Journal of Neuroscience, Psychology, and Economics*, *6*(3), 201.

Keser, C., & Van Winden, F. (2000). Conditional cooperation and voluntary contributions to public goods. *Scandinavian Journal of Economics*, *102*(1), 23–39.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*.

Li, S., & Taplin, J. (2002). Examining whether there is a disjunction effect in prisoner's dilemma games. *Chinese Journal of Psychology*, *44*, 25–46.

Lieder, F., Griffiths, T. L., Huys, Q. J., & Goodman, N. D. (2018). Empirical evidence for resource-rational anchoring and adjustment. *Psychonomic Bulletin & Review*, *25*(2), 775–784.

Mailath, G. J. (1998). Do people play Nash equilibrium? lessons from evolutionary game theory. *Journal of Economic Literature*, *36*(3), 1347–1374.

Markowitz, H. (1952). The utility of wealth. *Journal of political Economy*, *60*(2), 151–158.

Milinski, M., Sommerfeld, R. D., Krambeck, H.-J., Reed, F. A., & Marotzke, J. (2008). The collective-risk social dilemma and the prevention of simulated dangerous climate change. *Proceedings of the National Academy of Sciences*, *105*(7), 2291–2294.

Nash, J. F., et al. (1950). Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences*, *36*(1), 48–49.

Nobandegani, A. S. (2017). The Minimalist Mind: On Mininality in Learning, Reasoning, Action, & Imagination. McGill University, PhD Dissertation.

Nobandegani, A. S., da Silva Castanheira, K., O'Donnell, T. J., & Shultz, T. R. (2019a). On robustness: An undervalued dimension of human rationality. In *Proceedings of the* 17th *International Conference on Cognitive Modeling.* Montreal, QC.

Nobandegani, A. S., da Silva Castanheira, K., Otto, A. R., & Shultz, T. R. (2018). Over-representation of extreme events in decision-making: A rational metacognitive account. In *Proceedings of the* 40th *Annual Conference of the Cognitive Science Society* (pp. 2391-2396). Austin, TX: Cognitive Science Society.

Nobandegani, A. S., da Silva Castanheira, K., Otto, A. R., & Shultz, T. R. (2019b). A resource-rational process-level account the st. petersburg paradox. In *Proceedings of the* 41st *Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Nobandegani, A. S., Destais, C., & Shultz, T. R. (in prep.). An expectation-based model of fairness in the Ultimatum Game.

Osborne, D. K. (1976). Cartel problems. *The American Economic Review*, *66*(5), 835–844.

Pearce, D. G. (1984). Rationalizable strategic behavior and the problem of perfection. *Econometrica: Journal of the Econometric Society*, 1029–1050.

Poor, H. V. (2013). *An Introduction to Signal Detection and Estimation*. Springer Science & Business Media.

Pothos, E. M., & Busemeyer, J. R. (2009). A quantum probability explanation for violations of rationaldecision theory. *Proceedings of the Royal Society B: Biological Sciences*, *276*(1665), 2171–2178.

Pothos, E. M., & Busemeyer, J. R. (2013). Can quantum probability provide a new direction for cognitive modeling? *Behavioral and Brain Sciences*, *36*(3), 255–274.

Pothos, E. M., Perry, G., Corr, P. J., Matthew, M. R., & Busemeyer, J. R. (2011). Understanding cooperation in the prisoners dilemma game. *Personality and Individual Differences*, *51*(3), 210–215.

Rand, D. G. (2016). Cooperation, fast and slow: Meta-analytic evidence for a theory of social heuristics and self-interested deliberation. *Psychological Science*, *27*(9), 1192–1206.

Rand, D. G., Greene, J. D., & Nowak, M. A. (2012). Spontaneous giving and calculated greed. *Nature*, *489*(7416), 427.

Rapoport, A., Chammah, A. M., & Orwant, C. J. (1965). *Prisoner's dilemma: A study in conflict and cooperation* (Vol. 165). University of Michigan press.

Rubinstein, A. (2007). Instinctive and cognitive reasoning: A study of response times. *The Economic Journal*, *117*(523), 1243–1259.

Savulescu, J., Foddy, B., & Clayton, M. (2004). Why we should allow performance enhancing drugs in sport. *British Journal of Sports Medicine*, *38*(6), 666–670.

Scholten, M., & Read, D. (2014). Prospect theory and the forgotten fourfold pattern of risk preferences. *Journal of Risk and Uncertainty*, *48*(1), 67–83.

Shafir, E., & Tversky, A. (1992). Thinking through uncertainty: Nonconsequential reasoning and choice. *Cognitive Psychology*, *24*(4), 449–474.

Simon, H. A. (1957). *Models of Man*. Wiley.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*(2), 207–232.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*(4), 297–323.

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? optimal decisions from very few samples. *Cognitive Science*, *38*(4), 599–637.

Wiesner, J. B., & York, H. F. (1964). National security and the nuclear-test ban. *Scientific American*, *211*(4), 27–35.

A Resource-Rational Process-Level Account of the St. Petersburg Paradox

Ardavan S. Nobandegani[1,3], **Kevin da Silva Castanheira**[3], **Thomas R. Shultz**[2,3], **& A. Ross Otto**[3]
{ardavan.salehinobandegani, kevin.dasilvacastanheira}@mail.mcgill.ca
{thomas.shultz, ross.otto}@mcgill.ca
[1]Department of Electrical & Computer Engineering, McGill University
[2]School of Computer Science, McGill University
[3]Department of Psychology, McGill University

Abstract

The St. Petersburg paradox is a centuries-old philosophical puzzle concerning a lottery with infinite expected payoff, on which people are, nevertheless, willing to place only a small bid. Despite many attempts and several proposals, no generally-accepted resolution is yet at hand. In this work, we present the first resource-rational process-level explanation of this paradox, demonstrating that it can be accounted for by a variant of normative expected-utility-maximization which acknowledges cognitive limitations. Specifically, we show that Nobandegani et al.'s (2018) metacognitively-rational model, *sample-based expected utility* (SbEU), can account for major experimental findings on this paradox. Crucially, our resolution is consistent with two empirically well-supported assumptions: (1) people use only a few samples in probabilistic judgments and decision-making, and (2) people tend to overestimate the probability of extreme events in their judgment.

Keywords: St. Petersburg Paradox; bounded rationality; resource-rational process models; expected utility theory; inference by sampling

1 Introduction

Originally proposed in 1738 by Daniel Bernoulli, the St. Petersburg paradox is a famous economic and philosophical puzzle concerning a risky gamble on which people are asked to place a bid. The gamble goes as follows: The house offers to flip a coin until it comes up heads; the house pays \$1 if heads appears on the first trial (aka initial seed); otherwise the payoff doubles each time tails appears, with this compounding stopping and payment being given at the first heads. The St. Petersburg gamble is outlined in Table 1.

Trial	1	2	3	...	n	...
Event	H	TH	TTH	...	$\underbrace{\text{TT...TH}}_{(n-1)\ \textbf{tails}}$	...
Payoff	\$1	\$2	\$4	...	$\$2^{(n-1)}$	...

Table 1: The St. Petersburg gamble. A fair coin is flipped until the first heads appears. On the n^{th} trial of the gamble, corresponding to the event of having the first heads appear on the n^{th} coin flip, the house pays $\$2^{(n-1)}$ to the bidder and the game ends. The expected value (EV) of this gamble is infinite: EV = $\$1 \times (\frac{1}{2}) + \$2 \times (\frac{1}{4}) + \$4 \times (\frac{1}{8}) + \$8 \times (\frac{1}{16}) + \$16 \times (\frac{1}{32}) + \ldots = \$\frac{1}{2} + \$\frac{1}{2} + \$\frac{1}{2} + \$\frac{1}{2} + \$\frac{1}{2} + \ldots = +\infty$.

Despite the expected value (EV) of the St. Petersburg gamble being infinite (see Table 1), people are typically willing to place only small bids on this gamble (e.g., Bottom, Bontempo, & Holtgrave, 1989; Rivero, Holtgrave, Bontempo, & Bottom, 1990; Kroll & Vogt, 2009; Cox, Sadiraj, & Vogt, 2009; Hayden & Platt, 2009). Under the normative stance that people should prefer gambles with higher EVs, this paradox calls into question human rationality: The EV of the gamble being infinite, people, therefore, should be willing to place *arbitrarily* large bids on this gamble, but this is far from what experimental evidence suggests.

In spite of its innocent appearance, the St. Petersburg paradox occupied the minds of many over the past two centuries, eliciting a variety of reflections and explanations from several notable thinkers, including Daniel and Niklaus Bernoulli, Cramer, de Morgan, Condorcet, Euler, Poisson, and Gibbon, Marschack, Cournot, Arrow, Keynes, Stigler, Samuelson, von Mises, Ramsey and Aumann (see Arrow, 1951; Aumann, 1977; Dutka, 1988; Keynes, 1921; Samuelson, 1960). Nonetheless, no widely accepted explanation of this paradox exists to date.

In this work, we ask whether people's bids on the St. Petersburg paradox could be understood as an optimal behavior with the mind acting as a cognitive miser. Answering this question in the affirmative, we show that the St. Petersburg paradox can be accounted for by a variant of normative expected-utility-maximization which acknowledges computational and cognitive limitations. Specifically, we demonstrate that Nobandegani, da Silva Castanheira, Otto, and Shultz's (2018) metacognitively-rational model, *sample-based expected utility* (SbEU), can account for major experimental findings on the St. Petersburg paradox.

In the present study, our efforts are simultaneously guided by two well-supported observations about human judgment and decision-making under risk: (1) mounting evidence suggests that people often use very few samples in probabilistic judgments and reasoning (e.g., Vul et al., 2014; Battaglia et al. 2013; Lake et al., 2017; Gershman, Horvitz, & Tenenbaum, 2015; Hertwig & Pleskac, 2010; Griffiths et al., 2012; Gershman, Vul, & Tenenbaum, 2012; Bonawitz et al., 2014; Nobandegani et al., 2018; Lieder, Griffiths, Huys, & Goodman, 2018), and (2) people overestimate the probability of extreme events in their judgments (e.g., Tversky & Kahneman, 1972; Ungemach, Chater, & Stewart, 2009; Burns, Chiu, & Wu, 2010; Barberis, 2013; Lieder et al., 2018). As we discuss in the next section, previous explanations of the St. Petersburg paradox fail to respect at least one of these observations.

Our paper is organized as follows. We begin by present-

ing a brief historical overview of major explanations of the St. Petersburg paradox. After providing a brief overview of SbEU, we turn to modeling four major experimental findings on the St. Petersburg paradox. We conclude by discussing the implications of our work for the debate on human rationality.

2 A Brief Historical Overview of the Paradox

In this section, we present a brief overview of major resolutions of the St. Petersburg paradox, followed by notable critiques of them.

It is worth noting that most of the work on the St. Petersburg paradox thus far has been theoretical or philosophical. Comparatively little effort has been directed at providing empirical data on the bids people would be willing to place on the gamble and/or how people's bids would be affected by changing focal characteristics of the gamble, e.g. by varying the initial seed or limiting the number of coin flips in the gamble (e.g., Bottom, Bontempo, and Holtgrave, 1989; Rivero, Holtgrave, Bontempo, and Bottom, 1990; Kroll and Vogt, 2009; Cox, Sadiraj, and Vogt, 2011; Hayden & Platt, 2009; Neugebauer, 2010).

Diminishing marginal utility. Initially presented by Daniel Bernoulli (1738), the diminishing marginal utility explanation of the St. Petersburg paradox argues that, instead of evaluating the expected value (EV) of the gamble (which is infinite, see Table 1), people evaluate the expected utility of the gamble, with the utility function having a concave form (aka diminishing marginal utility).

As this explanation fails to account for super-St. Petersburg paradoxes in which the gamble's payoff increases super-exponentially with every coin flip, recent discussions of this explanation have to make the further assumption that the utility function is bounded from above (e.g., Aumann, 1977; Martin, 2008; Menger, 1934; Samuelson, 1977; Vickrey, 1960).

The diminishing marginal utility explanation has been discredited several times, mainly because it over-predicts bids (Lopes, 1981; Martin, 2008; Menger, 1934; Moritz, 1923; Samuelson, 1960, 1977). (This is not to say that marginal utility does not diminish, just that this factor is insufficient to explain the paradox.) Also, the diminishing marginal utility explanation completely neglects the well-supported observation that people overestimate the probability of extreme events in their judgment (e.g., Tversky & Kahneman, 1972; Ungemach, Chater, & Stewart, 2009; Burns, Chiu, & Wu, 2010; Barberis, 2013; Lieder et al., 2018), mistakenly assuming that the subjective probability of a low-probability extreme event in the St. Petersburg gamble (e.g., to win $\$2^{100}$ with probability $\frac{1}{2^{101}}$) is equal to its objective probability (e.g., $\frac{1}{2^{101}}$). Replacing expected utility with more modern variants which respect the latter observation, e.g. cumulative prospect theory (CPT), does not help either, as empirically fit values strongly over-predict bids in the St. Petersburg paradox (Blavatskyy, 2005; Rieger & Wang, 2006; Camerer, 2005).

Finitude of resources. Another classic explanation is that since the amount of money in the world is finite, the gambler must be skeptical about the ability of the house to pay the large outcomes of the gamble. Relatedly, it has been argued that time is finite, and the gambler, knowing he or she cannot continue playing the game forever, bids less than the expected value of the gamble. This argument has been expressed, in various forms, by several scholars (see Savage, 1954; Tversky & Bar-Hillel, 1983; Vickrey, 1960; Dutka, 1988).

Weaknesses of these arguments have been explicated by several critics. Bertrand argues that, even if the house cannot afford to pay the money, unites of currency can be reasonably replaced by more plentiful stuff, such as grains of sand, inches, or molecules of hydrogen, and the risk aversion still remains (Dutka, 1988). By the same logic, the payment may even be hypothetical or psychological (Martin, 2008; Aumann, 1977).

Ignoring low probabilities. This explanation argues that people consider events whose probability falls below some threshold to be impossible, i.e. they never happen. For example, D'Alembert posited a $1/10,000$ threshold, while Niklaus Bernoulli set the cutoff at a more conservative $1/100,000$ (Dutka, 1988).

However, there is a serious flaw with this argument: According to the well-known availability bias (Tversky & Kahneman, 1972), people over-represent extreme events, i.e., events whose utility is large (Lieder et al., 2018; Nobandegani et al., 2018). As low-probability events have (exponentially) larger payoffs in the St. Petersburg gamble, people should overestimate those low-probability events, putting more weights on those low-probability events in their valuation of the gamble.

A key contribution of our work is to provide a resource-rational process-level explanation of why people are willing to place only a small bid on the gamble *despite* over-representing extreme events in their judgment and decision-making (see Sec. 3). Particularly, past work has shown that SbEU can account for availability bias (Nobandegani et al., 2018).

Computing the median instead of the mean. Recently, Hayde and Platt (2009) proposed that people report the median (and not the mean) of the distribution associated with the St. Petersburg gamble as their bid. The median of the distribution associated with the St. Petersburg gamble is between \$1 and \$2, and is set by convention at \$1.50 (Weissstein, 2008).

The median explanation of Hayde and Platt (2009) is currently the only model which can simultaneously account for all the major experimental findings on the St. Petersburg gamble. We investigate all these major experimental findings in the present study in Sec. 4.

Nevertheless, despite its quantitative coverage, the median explanation remains too limited to explain the St. Petersburg paradox, markedly detached from the extensive literature on human judgment and decision-making. Similar to the diminishing marginal utility explanation, the median explanation completely neglects the well-supported observation that peo-

ple overestimate the probability of extreme events in their judgment (e.g., Tversky & Kahneman, 1972; Lieder et al., 2018), mistakenly assuming that the subjective probability of a low-probability extreme event in the St. Petersburg gamble is equal to its objective probability.

In this work, we seek to provide a resource-rational process model of the St. Petersburg paradox that can additionally account for several well-known effects in decision-making under risk; SbEU meets this criterion (see Sec. 3). As such, we seek to understand the St. Petersburg gamble as a particular risky gamble whose process-level explanation should be consistent with a broader process-level model of decision-making under risk.

3 Sample-based Expected Utility Model

Extending the cognitively-rational decision-making model of Lieder, Griffiths, and Hsu (2018) to the realm of metacognition (Cary & Reder, 2002), SbEU is a metacognitively-rational process model of risky choice that posits that agents rationally adapt their strategies depending on the amount of time available for decision-making (Nobandegani et al., 2018). Concretely, SbEU assumes that an agent estimates expected utility

$$\mathbb{E}[u(o)] = \int p(o)u(o)do, \tag{1}$$

using self-normalized importance sampling (Hammersley & Handscomb, 1964; Geweke, 1989), with its importance distribution q^* aiming to optimally minimize mean-squared error (MSE):

$$\hat{E} = \frac{1}{\sum_{j=1}^{s} w_j} \sum_{i=1}^{s} w_i u(o_i), \quad \forall i:\ o_i \sim q^*,\ w_i = \frac{p(o_i)}{q^*(o_i)}, \tag{2}$$

$$q^*(o) \propto p(o)|u(o)| \sqrt{\frac{1+|u(o)|\sqrt{s}}{|u(o)|\sqrt{s}}}. \tag{3}$$

MSE is a standard normative measure of the quality of an estimator, and is widely adopted in machine learning and mathematical statistics (Poor, 2013). In Eqs. (1-3), o denotes an outcome of a risky gamble, $p(o)$ the objective probability of outcome o, $u(o)$ the subjective utility of outcome o, $\hat{E}$ the importance-sampling estimate of expected utility given in Eq. (1), q^* the importance-sampling distribution, o_i an outcome randomly sampled from q^*, and s the number of samples drawn from q^*.

Recently, Nobandegani et al. (2018) showed that SbEU can account for availability bias, the tendency to overestimate the probability of events that easily come to mind (Tversky & Kahneman, 1972), and can accurately simulate the well-known fourfold pattern of risk preferences in outcome probability (Tversky & Kahneman, 1992) and in outcome magnitude (Markovitz, 1952; Scholten & Read, 2014). Notably, SbEU is the first rational process model to score near-perfectly in optimality, economical use of limited cognitive resources, and robustness, all at the same time (Nobandegani et al., 2018; Nobandegani et al., 2019a).

4 Simulation Results

In this section, we show that SbEU can quantitatively account for four major experimental findings on the St. Petersburg paradox: (1) Bids are only weakly affected by truncating the game (e.g., Cox et al. 2007; Neugebauer, 2010; Hayden & Platt, 2009), (2) Bids are strongly increased by repeating the game (Neugebauer, 2010; Hayden & Platt, 2009), (3) Bids are typically lower than twice the smallest payoff (Hayden & Platt, 2009), and (4) Bids depend linearly on the initial seed of the game (Hayden & Platt, 2009).

Recent work has provided mounting evidence suggesting that people often use very few samples in probabilistic judgments and reasoning (e.g., Vul et al., 2014; Battaglia et al. 2013; Lake et al., 2017; Gershman, Horvitz, & Tenenbaum, 2015; Hertwig & Pleskac, 2010; Griffiths et al., 2012; Gershman, Vul, & Tenenbaum, 2012; Bonawitz et al., 2014; Nobandegani et al., 2018; Lieder, Griffiths, Huys, & Goodman, 2018). Consistent with this finding, in the present study we assume that bidders draw only one sample ($s = 1$; see Eqs. 2-3) when evaluating their (subject) expected utility of the St. Petersburg gamble.

Concretely, we use the Metropolis–Hastings Markov chain Monte Carlo (MCMC) method—a well-known rational process model for sampling from a probability distribution of interest—to generate a single sample ($s = 1$) from the importance distribution q^* given in Eq. 3. MCMC methods have been successful in simulating important aspects of a wide range of cognitive phenomena, e.g., temporal dynamics of multistable perception (Gershman et al., 2012; Moreno-Bote et al., 2011), developmental changes in cognition (Bonawitz, Denison, Griffths, & Gopnik, 2014), category learning (Sanborn et al., 2010), and accounting for many cognitive biases (Nobandegani et al., 2018; Dasgupta et al., 2016).

Also, consistent with prospect theory (Kahneman & Tversky, 1979) and cumulative prospect theory (Kahneman & Tversky, 1992), in this paper we assume a standard S-shaped utility function $u(x)$ given by:

$$u(x) = \begin{cases} x^{0.35} & \text{if } x \geq 0, \\ -|x|^{0.45} & \text{if } x < 0. \end{cases} \tag{4}$$

4.1 Bids are weakly affected by truncating the game

In the original St. Petersburg gamble, there is no a priori upper-bound on number of coin flips; theoretically it can continue indefinitely. In a truncated variant of the St. Petersburg gamble, some a priori upper-bound is placed on the number of coin flips. Several experimental studies have shown that bids that people are willing to offer to play the St. Petersburg gamble are only weakly affected by truncating the game (Cox et eal., 2007; Cox et al., 2008, 2009; Hayden & Platt, 2009). This finding is generally taken as evidence for people ignoring small-probability events in the game (Neugebauer, 2010).

Recently, Hayden and Platt (2009) investigated bids for the St. Petersburg gamble truncated at 3 flips (maximum payoff: \$8, EV: \$2.50), 5 flips (maximum payoff: \$32, EV: \$3.50),

Figure 1: Hayden and Platt's (2009) experimental data on the effect of truncation on bids for the St. Petersburg gamble. Adapted from Hayden and Platt (2009).

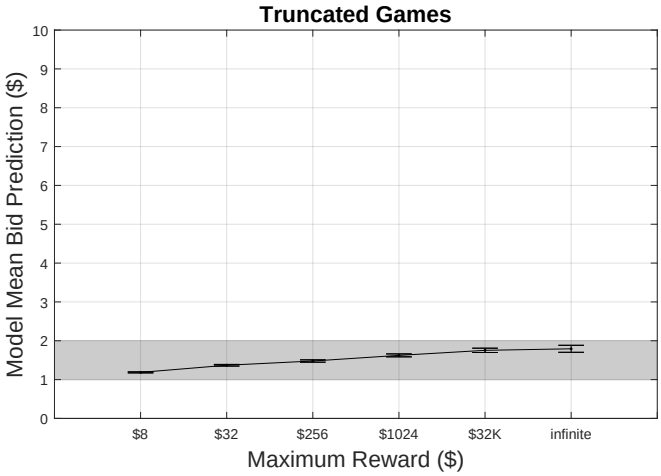


Figure 2: SbEU (Nobandegani et al., 2018) can accurately simulate the experimental data of Hayden and Platt (2009) on the effect of truncation on bids for the St. Petersburg gamble. Error bars indicate $\pm$ 1 SEM.

8 flips (maximum payoff: \$256, EV: \$5), 10 flips (maximum payoff: \$1024, EV: \$6) and 15 flips (maximum payoff: \$32,768, EV: \$8.50); their experimental data are shown in Fig. 1.

Fig. 2 shows that SbEU can account for the experimental data of Hayden and Platt (2009). In Fig. 2, we simulate $N = 1000$ participants, with $s = 1$.

4.2 Bids rise with repetitions of the game

Recently, Hayden and Platt (2009) experimentally showed that bids to play the (un-truncated) St. Petersburg gamble are strongly affected by repeating the game, with people willing to place higher bids with a larger number of game repetitions.

Fig. 3 shows that SbEU can qualitatively simulate people's tendency to place higher bids for a larger number of game repetitions, as experimentally shown by Hayden and Platt (2009). In Fig. 3, we simulate $N = 1000$ participants, with $s = 1$.

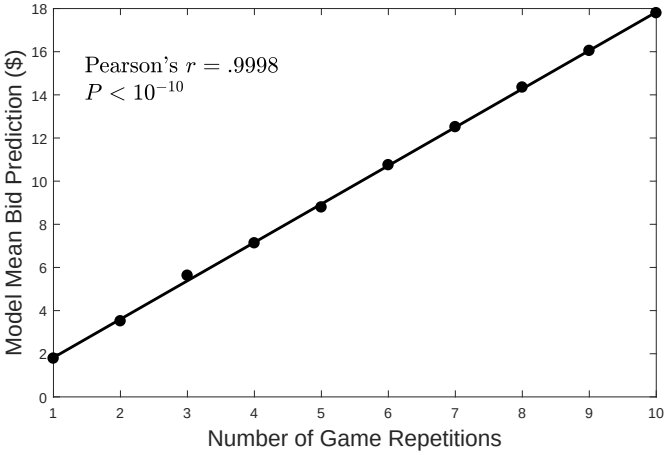


Figure 3: SbEU (Nobandegani et al., 2018) can account for the experimental finding of Hayden and Platt (2009) showing that people willing to place higher bids for a larger number of game repetitions (Pearson's $r = .9998$, Kendall's $\tau = 1$, Spearman's $\rho = 1$, $Ps < .001$).

4.3 Bids are typically lower than twice the smallest payoff

In their recent work, Hayden and Platt (2009) showed that bids to play the (un-truncated) St. Petersburg gamble are typically lower than twice the smallest payoff of the game.

Fig. 4 shows that SbEU can account for this experimental finding of Hayden and Platt (2009). In Fig. 4, we simulate $N = 1000$ participants, with $s = 1$.

4.4 Bids depend linearly on the initial seed

Interestingly, Hayden and Platt (2009) showed that bids to play the (un-truncated) St. Petersburg gamble depend linearly on the initial seed of the game, thus providing a quantitatively well-characterized criterion for evaluating a computational account.

Fig. 5 shows that SbEU can accurately account for this experimental finding of Hayden and Platt (2009) (Pearson's $r = .9758$, Kendall's $\tau = 0.9556$, Spearman's $\rho = .9879$, $Ps < .001$). In Fig. 5, we simulate $N = 1000$ participants, with $s = 1$.

5 General Discussion

The St. Petersburg paradox (Bernoulli, 1738) stands among the oldest philosophical puzzles of human decision-making, and has played a pivotal role in the emergence of the concept of the subjective utility curve, a central concept in economics

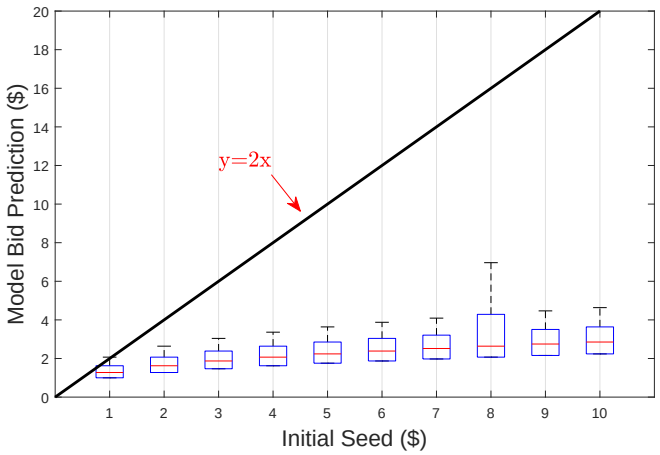


Figure 4: Boxplots of the model's bids. SbEU can account for the experimental finding of Hayden and Platt (2009) showing that people's bids are typically lower than twice the smallest payoff (i.e. initial seed) in the St. Petersburg gamble. On each box, the central red mark indicates the median, and the bottom and top edges of the box indicate the 25$^{\text{th}}$ (denoted by q_1) and 75$^{\text{th}}$ (denoted by q_3) percentiles of the data, respectively. On each box, the whisker extends to the most extreme data points not considered outliers. Outliers are data points that lie outside the interval $[q_1 - 1.5 \times (q_3 - q_1), q_3 + 1.5 \times (q_3 - q_1)]$), and are not shown in this plot. The boldfaced black solid line depicts $y = 2x$.

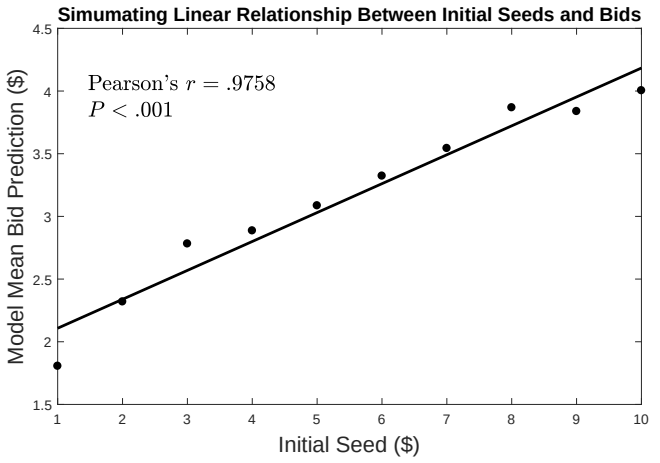


Figure 5: SbEU (Nobandegani et al., 2018) can account for the experimental finding of Hayden and Platt (2009) showing that bids depend linearly on the initial seed of the St. Petersburg gamble (Pearson's $r = .9758$, Kendall's $\tau = 0.9556$, Spearman's $\rho = .9879$, $Ps < .001$).

(Dutka, 1988). Despite occupying the minds of many important thinkers, eliciting many attempts and several proposals, no generally-accepted resolution is yet at hand.

In this work, we provide an algorithmic-level account of major experimental findings on the St. Petersburg paradox. Specifically, we show that a single parameterization of Nobandegani et al.'s (2018) metacognitively-rational model, SbEU, provides a unified, resource-rational, process-level explanation of (1) why bids are only weakly affected by truncating the game, (2) why people are willing to place higher bids for a larger number of game repetitions, (3) why bids are typically lower that twice the smallest payoff of the game (aka initial seed), and (4) why bids depend linearly on the initial seed of the game. As such, Items (1-4) can be understood as optimal behavior subject to cognitive limitations.

As opposed to the competing median explanation of Hayden and Platt (2009) that is too specific to the St. Petersburg paradox, our work provides a resource-rational process model of the St. Petersburg paradox that can additionally account for several well-known effects in decision-making under risk (Nobandegani et al., 2018), and is fully in line with the much broader process-level understanding of human probabilistic judgment and reasoning based on sampling (e.g., Stewart, Chater, & Brown, 2006; Sanborn & Chater, 2016).

Recent work has shown that SbEU provides a resource-rational mechanistic account of (ostensibly irrational) cooperation in one-shot Prisoner's Dilemma games, thus successfully bridging between game-theoretic decision-making and risky decision-making (Nobandegani, da Silva Castanheira, Shultz, & Otto, 2019b). There is also experimental confirmation of a counterintuitive prediction of SbEU: Deliberation leads people to move from one well-known bias, framing effect, to another well-known bias, the fourfold pattern of risk preferences (da Silva Castanheira; Nobandegani, & Otto, 2019).

Crucially, our explanation retains the well-supported assumption that people overestimate the probability of extreme events in their judgment and decision-making (Tversky & Kahneman, 1972; Lieder et al., 2018; Nobandegani et al., 2018), and is fully in line with mounting evidence suggesting that people use only a few samples in probabilistic judgments and reasoning (e.g., Vul et al., 2014; Battaglia et al. 2013; Lake et al., 2017; Gershman, Horvitz, & Tenenbaum, 2015; Hertwig & Pleskac, 2010; Griffiths et al., 2012; Gershman, Vul, & Tenenbaum, 2012; Bonawitz et al., 2014; Nobandegani et al., 2018; Lieder, Griffiths, Huys, & Goodman, 2018).

Recently, Blavatskyy (2005) showed that conventional parameterizations of cumulative prospect theory (CPT; Kahneman & Tversky, 1992) do not explain the St. Petersburg paradox. As we demonstrate in this work, assuming a standard S-shaped utility function, as advocated by CPT, suffices for explaining the St. Petersburg paradox with SbEU (see Eq. 4).

There have been several recent studies (see Lieder & Griffiths, 2018, for a review) attempting to show that many well-known (purportedly irrational) behavioral effects and cognitive biases can be understood as optimal behavior subject to computational and cognitive limitations (*rational minimalist program*, Nobandegani, 2017; Griffiths, Lieder, & Goodman, 2015). The present study contributes to this line of work by providing a resource-rational process-level account of a

centuries-old puzzle concerning human decision-making.

Future work should investigate whether other long-standing paradoxes of human judgment and decision-making, e.g., the Ellsburg paradox (Ellsberg, 1961), could be also understood as optimal behavior subject to cognitive limitations. We see our work as a step in this direction.

Acknowledgments This work is supported by an operating grant to TRS from the Natural Sciences and Engineering Research Council of Canada. We would like to thank Constance Destais, Ashley Stendel, Marcel Montrey, and Peter Helfer for helpful comments on an earlier draft of this work.

References

Arrow, K. J. (1951). Alternative approaches to the theory of choice in risk-taking situations. *Econometrica: Journal of the Econometric Society*, 404–437.

Aumann, R. J. (1977). The st. petersburg paradox: A discussion of some recent comments. *Journal of Economic Theory*, *14*(2), 443–445.

Barberis, N. (2013). The psychology of tail events: Progress and challenges. *American Economic Review*, *103*(3), 611–16.

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, *110*(45), 18327–18332.

Bernoulli, D. (1738). Specimen theoriae novae de mensura sortis (exposition of a new theory on the measurement of risk). *Comentarii Acad Scient Petropolis* (Translated in *Econometrica*), *5*(22), 23–36.

Blavatskyy, P. R. (2005). Back to the st. petersburg paradox? *Management Science*, *51*(4), 677–678.

Bonawitz, E., Denison, S., Griffiths, T. L., & Gopnik, A. (2014). Probabilistic models, learning algorithms, and response variability: sampling in cognitive development. *Trends in cognitive Sciences*, *18*(10), 497–500.

Bottom, W. P., Bontempo, R. N., & Holtgrave, D. R. (1989). Experts, novices, and the st. petersburg paradox: Is one solution enough? *Journal of Behavioral Decision Making*, *2*(3), 139–147.

Burns, Z., Chiu, A., & Wu, G. (2010). Overweighting of small probabilities. *Wiley Encyclopedia of Operations Research and Management Science*.

Camerer, C. (2005). Three cheers—psychological, theoretical, empirical—for loss aversion. *Journal of Marketing Research*, *42*(2), 129–133.

Cary, M., & Reder, L. M. (2002). Metacognition in strategy selection. In *Metacognition: Process, Function and Use* (pp. 63–77). Springer.

Cox, J., Sadiraj, V., & Vogt, B. (2009). On the empirical relevance of st. petersburg lotteries. *Economics Bulletin*, *29*(1), 214–220.

Cox, J., Sadiraj, V., Vogt, B., Dasgupta, U., et al. (2007). Is there a plausible theory for risky decisions? *Georgia State University*.

Cox, J. C., & Sadiraj, V. (2008). Risky decisions in the large and in the small: Theory and experiment. In *Risk aversion in experiments: Research in expermental economics* (Vol. 12, pp. 9–40). Emerald Group Publishing Limited.

Dasgupta, I., Schulz, E., & Gershman, S. J. (2016). Where do hypotheses come from? Center for Brains, Minds and Machines (CBMM) Memo No. 056.

da Silva Castanheira, K., Nobandegani, A. S., & Otto, A. R. (2019). Sample-based variant of expected utility explains effects of time pressure and individual differences in processing speed on risk preferences. In: *Proceedings of the* 41*st Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Dutka, J. (1988). On the st. petersburg paradox. *Archive for History of Exact Sciences*, *39*(1), 13–39.

Ellsberg, D. (1961). Risk, ambiguity, and the savage axioms. *The Quarterly Journal of Economics*, *75*(4), 643–669.

Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, *349*(6245), 273–278.

Gershman, S. J., Vul, E., & Tenenbaum, J. B. (2012). Multistability and perceptual inference. *Neural Computation*, *24*(1), 1–24.

Geweke, J. (1989). Bayesian inference in econometric models using monte carlo integration. *Econometrica: Journal of the Econometric Society*, 1317–1339.

Griffiths, T. L., Vul, E., & Sanborn, A. N. (2012). Bridging levels of analysis for probabilistic models of cognition. *Current Directions in Psychological Science*, *21*(4), 263–268.

Hammersley, J., & Handscomb, D. (1964). *Monte carlo methods.* London: Methuen & Co Ltd.

Hayden, B. Y., & Platt, M. L. (2009). The mean, the median, and the st. petersburg paradox. *Judgment and Decision Making*, *4*(4), 256–272.

Hertwig, R., & Pleskac, T. J. (2010). Decisions from experience: Why small samples? *Cognition*, *115*(2), 225–237.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive psychology*, *3*(3), 430–454.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263–291.

Keynes, J. M. (1921). *A treatise on probability*. Courier Corporation.

Kroll, E. B., & Vogt, B. (2009). The st. petersburg paradox despite risk-seeking preferences. an experimental study.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*.

Lieder, F., & Griffiths, T. L. (2018). Resource-rational analysis: understanding human cognition as the optimal use of limited computatiomal resources. Available on Researchgate.

Lieder, F., Griffiths, T. L., & Hsu, M. (2018). Overrepresentation of extreme events in decision making reflects rational use of cognitive resources. *Psychological Review*, *125*(1).

Lieder, F., Griffiths, T. L., Huys, Q. J., & Goodman, N. D. (2018). Empirical evidence for resource-rational anchoring and adjustment. *Psychonomic Bulletin & Review*, *25*(2), 775–784.

Lopes, L. L. (1981). Decision making in the short run. *Journal of Experimental Psychology: Human Learning and Memory*, *7*(5), 377.

Markowitz, H. (1952). The utility of wealth. *Journal of Political Economy*, *60*(2), 151–158.

Martin, R. (2017). The st. petersburg paradox. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2017 ed.). Stanford University.

Menger, K. (1934). Das unsicherheitsmoment in der wertlehere. *Z Nationaloeken*, *51:*(459–485).

Moreno-Bote, R., Knill, D. C., & Pouget, A. (2011). Bayesian sampling in visual perception. *Proceedings of the National Academy of Sciences*, *108*(30), 12491–12496.

Moritz, R. E. (1923). Some curious fallacies in the study of probability. *The American Mathematical Monthly*, *30*(2), 58–65.

Neugebauer, T., et al. (2010). Moral impossibility in the petersburg paradox: a literature survey and experimental evidence. *Luxembourg School of Finance Research Working Paper Series*(10-14).

Nobandegani, A. S. (2017). The Minimalist Mind: On Mininality in Learning, Reasoning, Action, & Imagination. McGill University, PhD Dissertation.

Nobandegani, A. S., da Silva Castanheira, K., O'Donnell, T. J., & Shultz, T. R. (2019a). On robustness: An undervalued dimension of human rationality. In *Proceedings of the* 17^{th} *International Conference on Cognitive Modeling.* Montreal, QC.

Nobandegani, A. S., da Silva Castanheira, K., Otto, A. R., & Shultz, T. R. (2018). Over-representation of extreme events in decision-making: A rational metacognitive account. In: *Proceedings of the* 40^{th} *Annual Conference of the Cognitive Science Society* (pp. 2391-2396). Austin, TX: Cognitive Science Society.

Nobandegani, A. S., da Silva Castanheira, K., Otto, A. R., & Shultz, T. R. (2019b). A resource-rational mechanistic approach to one-shot non-cooperative games: The case of prisoner's dilemma. In: *Proceedings of the* 41^{st} *Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Poor, H. V. (2013). *An Introduction to Signal Detection and Estimation*. Springer Science & Business Media.

Rieger, M. O., & Wang, M. (2006). Cumulative prospect theory and the st. petersburg paradox. *Economic Theory*, *28*(3), 665–679.

Rivero, J. C., Holtgrave, D. R., Bontempo, R. N., & Bottom, W. P. (1990). The st. petersburg paradox: Data at last. *Commentary, 8 (3–4)*, 46–51.

Samuelson, P. A. (1960). The st. petersburg paradox as a divergent double limit. *International Economic Review*, *1*(1), 31–37.

Samuelson, P. A. (1977). St. petersburg paradoxes: Defanged, dissected, and historically described. *Journal of Economic Literature*, *15*(1), 24–55.

Sanborn, A. N., & Chater, N. (2016). Bayesian brains without probabilities. *Trends in Cognitive Sciences*, *20*(12), 883–893.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: alternative algorithms for category learning. *Psychological Review*, *117*(4), 1144.

Savage, L. J. (1954). *The foundations of statistics*. Courier Corporation.

Scholten, M., & Read, D. (2014). Prospect theory and the "forgotten" fourfold pattern of risk preferences. *Journal of Risk and Uncertainty*, *48*(1), 67–83.

Stewart, N., Chater, N., & Brown, G. D. (2006). Decision by sampling. *Cognitive Psychology*, *53*(1), 1–26.

Tversky, A., & Bar-Hillel, M. (1983). Risk: The long and the short. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *9*, 713–717.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and uncertainty*, *5*(4), 297–323.

Ungemach, C., Chater, N., & Stewart, N. (2009). Are probabilities overweighted or underweighted when rare outcomes are experienced (rarely)? *Psychological Science*, *20*(4), 473–479.

Vickrey, W. (1960). Utility, strategy, and social decision rules. *The Quarterly Journal of Economics*, *74*(4), 507–535.

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? optimal decisions from very few samples. *Cognitive Science*, *38*(4), 599–637.

Weisstein, E. W. (2008). Statistical median. *From MathWorld – A Wolfram Web Resource* http://mathworld.wolfram.com/StatisticalMedian.html.

Toward a Formal Science of Heuristics

Ardavan S. Nobandegani[1,3] **& Thomas R. Shultz**[2,3]
{ardavan.salehinobandegani, william.campoli}@mail.mcgill.ca
{thomas.shultz}@mcgill.ca
[1]Department of Electrical & Computer Engineering, McGill University
[2]School of Computer Science, McGill University
[3]Department of Psychology, McGill University

Abstract

Heuristics are simple, effective cognitive processes that deliberately ignore parts of information relevant to decision-making. Ecological rationality, as an essential part of the Adaptive Toolbox research program on heuristics, investigates the environmental conditions under which simple heuristics would outperform complex models of decision-making, thereby providing support for the surprising less-is-more effect. In this work, we present a new research program, dubbed formal science of heuristics (FSH), that nicely complements the ecological rationality research, developing it into a much richer research program. Concretely, FSH sets to (i) mathematically delineate the broadest class of environmental conditions under which a heuristic is fully optimal, and (ii) formally investigate how deviations from those conditions would lead to degradation of performance, thereby allowing for a mathematically rigorous characterization of their robustness. As an instantiation of the FSH research program, we present several analytical results aiming to delineate the mildest conditions granting the optimality of a well-known heuristic: Take The Best. We conclude by discussing the implications that pursuit of FSH could have on the science of heuristics.

Keywords: Ecological rationality; one-reason heuristics; formal science of heuristics; Take The Best heuristic

1 Introduction

Heuristics—simple, effective cognitive processes that deliberately ignore parts of information relevant to decision-making—are assumed to underpin much of human judgment and decision-making (e.g., Gigerenzer & Selten, 2001; Mousavi, Gigerenzer, & Kheirandish, 2016), and are widely considered to be sub-optimal, attaining higher speed at the expense of lower accuracy (e.g., Payne et al., 1993; Shah & Oppenheimer, 2008; Evans and Over, 2010).

Challenging the latter mindset, the influential *ecological rationality* research program (as part of the Adaptive Toolbox theory) maintains that heuristic are well-matched to the environment they are adopted in (Todd & Gigerenzer, 2007), and seeks to investigates the environmental conditions under which heuristics would outperform complex models of decision-making, giving rise to the surprising less-is-more effect: when less information or computation leads to more accurate judgments than more information or computation (Gigerenzer & Gaissmaier, 2011).

Despite great successes, ecological rationality work has predominantly focused on simulation-based demonstrations of simple heuristics outperforming complex strategies (e.g., Gigerenzer et al., 2008, Gigerenzer & Todd, 1999, Todd & Gigerenzer, 2000, Hoffrage & Reimer, 2004; Gigerenzer & Goldstein, 1996), directing comparatively little effort (but see, e.g., Martignon & Hoffrage, 2002, Hogarth & Karelaia, 2006) toward establishing a mathematically-rigorous characterization of the environmental conditions underpinning the less-is-more effect — Todd and Gigerenzer (2007) explicitly call for developing such deep theoretical accounts.

In this work, we present a new research program, dubbed formal science of heuristics (FSH), that nicely complements the ecological rationality research, developing it into a richer research program, and, additionally, permitting mathematicians and computer scientists to make important contributions to the science of heuristics.

Concretely, FSH pursues the following two objectives. (1) FSH seeks to mathematically delineate the *broadest* class of environmental conditions under which a heuristic is fully optimal (i.e., using the standard terminology of computer science, the environmental conditions under which a heuristic serves as a correct algorithm w.r.t. the objective of interest, or, equivalently, an approximation algorithm with an approximation ratio of one). (2) FSH aims to formally investigate how deviations from optimality conditions would lead to degradation of performance, thereby allowing for a mathematically rigorous characterization of a heuristic's robustness. As such, to provide strongest theoretical support for the robustness of a heuristic, FSH aims to analytically provide the *mildest* technical conditions granting the optimality of a heuristic. According to the Adaptive Toolbox theory, robustness is predominantly responsible for the less-is-more effect, and plays a central role in the success of fast-and-frugal heuristics in everyday life decisions (e.g., Gigerenzer & Todd, 1999; Gigerenzer & Gaissmaier, 2011).

With regard to objective (2) mentioned above, one of the mildest technical conditions worth considering is *distribution-free* performance guarantees, widely studied in statistical learning theory and machine learning (e.g., Valiant, 1984; Kearns, Vazirani, & Vazirani, 1994). As is often the case, a decision-maker lacks (at least partially) the knowledge of the regularities of their environment, and, therefore, is not fully informed as to how information relevant to a decision-making task of interest is distributed. Distribution-free results, as the term suggests, establish performance guarantees that hold true *regardless* of the probability distribution governing a decision-making task (e.g., the distribution of attributes in a multi-alternative decision-making task). As such, distribution-free results provide strong robustness guarantees while demanding *minimal* environmental knowledge on the

part of the decision-maker, thus playing an integral role in the FSH research program.

We should note that establishing distribution-free performance guarantees for a heuristic does *not* imply that: (1) the decision-maker is inattentive to their environment, nor that (2) the decision-maker is not trying to select a heuristic well-matched to the environment — experimental evidence clearly suggests otherwise (e.g., Rieskamp & Otto, 2006; Hoffart, Rieskamp, & Dutilh, 2018; Payne, Bettman, & Johnson, 1988; Bröder, 2003; Pachur, Todd, Gigerenzer, Schooler, & Goldstein, 2011). On the contrary, establishing distribution-free performance guarantees on a heuristic ensures that that heuristic is well-matched to the environment, even when the decision-maker's knowledge of the environment is *imperfect*—a psychologically plausible assumption.

This work is organized as follows. We begin by presenting an overview of a well-known heuristic: Take The Best (TTB). As an instantiation of FSH, we then establish several analytical results, including strong distribution-free performance guarantees, for TTB. Finally, we conclude by discussing the implications that pursuit of FSH could have on the science of heuristics.

2 Take The Best: An Overview

Take The Best (TTB; Tversky, 1969, Gigerenzer, Hoffrage, & Kleinbölting, 1991) belongs to the class of one-reason decision-making heuristics which base decisions on only one attribute value. In its classic form, TTB is concerned with the task of predicting which of two objects, each possessing several binary-valued attributes, has a higher value on a given criterion, e.g., which of two cities has a higher population, or, which of two cookies would be more delicious.

The machinery of TTB is quite simple: Starting with the attribute having the highest validity, make pairwise-comparisons between the attribute values of the two objects; as soon as the first discriminating attribute is encountered (i.e., the attribute on which the two objects differ), announce the object attaining the highest attribute value on the discriminating attribute to be the winning object. TTB visits attributes in a descending order of their validities. If no discriminating attribute is ever encountered, TTB selects the winning object uniformly at random.[1] In TTB, the validity v_i of the i^{th} attribute is given by (Gigerenzer et al., 2008):

$$v_i :\triangleq \frac{R_i}{R_i + W_i},$$

where R_i, W_i are the number of correct and incorrect inferences based on the i^{th} attribute alone, respectively.

The efficacy of TTB receives strong empirical support from a wide range of economic, demographic, environmental, and other prediction tasks (e.g., Gigerenzer et al., 2008; Czerlinski, Gigerenzer, & Goldstein, 1999; Chater, Oaksford, Nakisa, & Redington, 2003). For example, on the task of predicting which of two cities has a higher homeless rate, TTB achieves better prediction accuracy than several competitors, including multiple regression model (Gigerenzer et al., 2008). More strikingly, Czerlinski, Gigerenzer, and Goldstein (1999) empirically showed that, across 20 real-world prediction problems, on average TTB obtains the best prediction accuracy when competing with several prominent alternatives, including multiple regression and tallying heuristic. Relatedly, on the same 20 real-world prediction problems, Gigerenzer et al. (2008) empirically show that the predictive accuracy of TTB came, on average, within three percentage points of a complex Bayesian network model. More broadly, when environments are moderately unpredictable and learning samples are small, as with many social and economic situations, TTB tends to make inferences as accurately as or better than multiple regression and neural networks (Chater, Oaksford, Nakisa, & Redington, 2003).

To provide direct experimental evidence for TTB as a psychological model, Bröder and his colleagues (Bröder 2000; Bröder and Schiffer 2003) conducted 20 studies, concluding that TTB is used under a number of conditions such as when information is costly and the variability of the validity of the attributes is high. Furthermore, Bröder and Gaissmaier (2007) and Nosofsky and Bergert (2007) showed that TTB predicts response times better than weighted additive and exemplar models.

Previous work assessing the prediction accuracy of TTB has mainly focused on computer simulations, with some work establishing analytical results formally supporting the efficacy of TTB (e.g., Martignon & Hoffrage, 2002; Hogarth & Karelaia, 2005, 2006; Baucells, Carrasco, & Hogarth, 2008).

Pursuing the research program proposed by FSH, and contrary to past analytical work, in this work we consider a much broader class of problems involving nonlinear objective functions (Definitions 1-4) with interactions between attributes being also accounted for (Definition 2). For the broad class of problems discussed above, we formally establish conditions granting the optimality of TTB when dealing with both non-binary, discrete attribute values (Propositions 3, 5, and 6) and continuous attribute values (Proposition 4). We also analytically investigate a broad class of prediction problems—involving both structured (Definition 3) and unstructured noise (Definition 4)—for which only *probabilistic* guarantees can be provided. Additionally, and in sharp contrast to past analytical work, we provide strong distribution-free guarantees on TTB for several classes of prediction problems (Propositions 5, 6, and 8).

3 Instantiating FSH: The Case of TTB

As an instantiation of FSH, in this section we establish several analytical results, including strong distribution-free performance guarantees, for TTB.

Before we proceed further, let us formally delineate an objective function which characterizes a broad class of decision-

[1] Without loss of generality, we assume that the decision-maker initially recognizes all the objects which s/he has to choose from. Accordingly, the use of recognition heuristic (Gigerenzer et al., 2008), as the first step of TTB, is implicitly considered in our work.

making problems.

Definition 1. (Objective function) Let $O_1, O_2, \ldots, O_N$ denote the set of N objects a decision-maker should choose from. Let also $O_{ij} \in \{0,1\}$ denote the value of the j^{th} attribute of object O_i where $1 \leq j \leq M$, and w_k denote the weight corresponding to the k^{th} attribute, A_k. Finally, let $\psi(\cdot)$ be an *arbitrary* monotonically-increasing function (i.e., $\forall x\colon \frac{d}{dx}\psi(x) > 0$). Then, the winning object $O_{i^\star}$ is the one whose index $i^\star$ satisfies the following objective function:

$$i^\star :\triangleq \arg\max_i \psi(\sum_{j=1}^{M} w_j O_{ij}). \tag{1}$$

Therefore, in the case of having only two objects O_1, O_2 to choose from, the optimal decision rule is given by:

$$\psi(\sum_{j=1}^{M} w_j O_{1j}) \underset{O_1}{\overset{O_2}{\lessgtr}} \psi(\sum_{j=1}^{M} w_j O_{2j}), \tag{2}$$

where $A \underset{O_1}{\overset{O_2}{\lessgtr}} B$ denotes the following: choose O_1 if $A > B$; choose O_2 if $A < B$; and choose uniformly at random between O_1, O_2 if $A = B$. ■

It is worth noting that in Eqs. (1-2), ψ can be *any* monotonically-increasing function, e.g., $\psi(x) = e^x, \psi(x) = 2^x + \log(x)$.

Proposition 1. (Sufficient condition for optimality) *If there exists a $k \in \mathbb{N}$ such that $\forall p < k$, $\forall i, j \in \{1, \ldots, N\}$ $O_{ip} = O_{jp}$ and $w_k > \sum_{i>k} w_i$, then the following holds true:*

$$\exists j \in \{1, \ldots, N\}\ \forall i \neq j\ O_{jk} > O_{ik} \Rightarrow O_{i^\star} := O_j. \tag{3}$$

Importantly, Proposition 1 establishes a condition granting basing decision on only one attribute while preserving optimality with respect to the objective given in (1). As such, Proposition 1 provides a firm rational basis for the *possibility* of one-reason decision-making for the broad class of decision-making problems characterized in Definition 1.

Next, Proposition 2 establishes a condition granting the optimality of TTB (when choosing between an *arbitrary* number of objects) with respect to the objective given in (1).

Proposition 2. (Generalizing TTB to N-object prediction tasks) *Let $O_1, O_2, \ldots, O_N$ denote the set of N objects a decision-maker is to choose from. Let also w_k denote the weight corresponding to the k^{th} attribute (see Definition 1), and v_k denote the validity of the k^{th} attribute. If $\forall k\ w_k = v_k$ and $\exists r \in \mathbb{R}^{>2}$ s.t. $\forall k\ v_k \leq (\frac{1}{r})v_{k-1}$, then TTB is an optimal strategy for the class of decision-making problems characterized in Definition 1.* ■

In the N-object setting (as in Proposition 2), TTB works as follows: Starting with the attribute having the highest validity, compare attribute values across the N objects; as soon as the first discriminating attribute is encountered (i.e., the attribute on which at least two objects differ), exclude from consideration those objects faring worse on the discriminating attribute; announce the object surviving this elimination process to be the winning object. TTB visits attributes in a descending order of their validities. If no discriminating attribute is ever encountered, TTB selects the winning object uniformly at random.

Proposition 3. (Multi-level attribute values) *Let $O_1, O_2, \ldots, O_N$ denote the set of N objects a decision-maker is to choose from, with each object having M attributes. Let also $O_{ij} \in \{0, 1, \cdots, \theta\}$ denote the value of the j^{th} attribute of object O_i where $1 \leq j \leq M$. Finally, let w_k denote the weight corresponding to the k^{th} attribute (see Definition 1), and v_k denote the validity of the k^{th} attribute. If $\forall k\ w_k = v_k$, and $\exists r > 1 + \theta$ s.t. $\forall k\ v_k \leq (\frac{1}{r})v_{k-1}$, then TTB is an optimal strategy for the class of decision-making problems characterized in Definition 1.* ■

In simple terms, Proposition 3 analytically establishes a conditions granting the optimality of TTB (when generalized to the setting of N objects, each with discrete, multi-level attribute values) with respect to the objective given in (1).

Proposition 4. (Continuous attribute values) *Let $\forall i \in \{0,1\}$, O_i denote the two objects a decision-maker should choose from, with each object having M attributes. Let also $O_{ij} \in \mathbb{R}$ denote the value of the j^{th} attribute of object O_i. Finally, let w_k denote the weight corresponding to the k^{th} attribute (see Definition 1), and v_k denote the validity of the k^{th} attribute. Assuming that k^* denotes the index of the discriminating attribute on which TTB halts, and $|O_{1k^*} - O_{2k^*}| \leq \delta$, the following statement holds true: If $\forall k\ w_k = v_k$, and $\forall i, j \in \{1, \ldots, M\}$ $O_{ij} \leq U$, and $\exists r > 1 + \frac{U}{\delta}$ s.t. $\forall k\ v_k \leq (\frac{1}{r})v_{k-1}$, then TTB is an optimal strategy for the class of decision-making problems characterized in Definition 1.*

Proposition 4 analytically establishes a conditions granting the optimality of TTB (when choosing between two objects, each with continuous attribute values) with respect to the objective given in (1).

Following the line of research proposed by FSH, next we present our first distribution-free guarantee for TTB.

Proposition 5. (Distribution-free guarantee) *Let $O_1, O_2, \ldots, O_N$ denote the set of N objects a decision-maker is to choose from, with each object having M attributes. Let also $O_{ij} \in \{0, 1, \cdots, \theta\}$ denote the value of the j^{th} attribute of object O_i, where $\{O_{ij}\}_{i,j} \overset{d}{\backsim} P$ with P denoting a joint probability distributions over the set of all attribute values $\{O_{ij}\}_{i,j}$. Finally, let w_k denote the weight corresponding to the k^{th} attribute (see Definition 1), and v_k denote the validity of the k^{th} attribute. Then, for* any *joint probability distribution P the following statement holds true: If $\forall k\ w_k = v_k$, and $\exists r > 1 + \theta$ s.t. $\forall k\ v_k \leq (\frac{1}{r})v_{k-1}$, then TTB is an optimal strategy for the class of decision-making problems characterized in Definition 1.* ■

Proposition 5 analytically establishes a condition ensuring the optimality of TTB (when generalized to the setting of N objects, each with discrete, multi-level attribute values) with respect to the objective given in (1), in a strong distribution-

free manner. It is crucial to note that the optimality guarantee given in Proposition 5 holds true for *any* joint distribution P on the set of all attribute values.

Next, in Definition 2, we formally characterize a broad class of prediction problems wherein interactions between attribute values are also accounted for.

Definition 2. (Objective function) Let $O_1, O_2, \ldots, O_N$ denote the set of N objects a decision-maker should choose from, and $O_{ij} \in \{0, 1, \cdots, \theta\}$ denote the value of the j^{th} attribute of object O_i where $1 \leq j \leq M$. Additionally, let w_k denote the weight corresponding to the k^{th} attribute, and r_{pq} denote the weight quantifying the amount of interaction between the p^{th} and the q^{th} attributes. Finally, let $\chi(\cdot)$ be an *arbitrary* monotonically-increasing function (i.e., $\forall x\colon \frac{d}{dx}\chi(x) > 0$). Then, the winning object $O_{i^\star}$ is the one whose index $i^\star$ satisfies the following objective function:

$$i^\star :\triangleq \arg\max_i \chi\Big(\sum_{j=1}^{M} w_j O_{ij} + \sum_{\substack{p,q \\ p\neq q}} r_{pq} O_{ip} O_{iq}\Big). \tag{4}$$

Proposition 6. (Distribution-free guarantee) *Consider the class of prediction problems formally characterized in Definition 2. Let also P denote a joint probability distributions over the set of all attribute values $\{O_{ij}\}_{i,j}$, i.e., $\{O_{ij}\}_{i,j} \overset{d}{\backsim} P$. Then, for* any *joint probability distribution P the following statement holds true: If $\forall k\ w_k = v_k$, and $\exists R \in \mathbb{R}$ s.t. $\forall p, q \in \{1, \cdots, M\}\ r_{pq} < R$, and $\exists r > 1 + \theta$ s.t. $\forall k\ \frac{r-1}{r-(\theta+1)}\binom{M}{2}\theta^2 R \leq v_k \leq (\frac{1}{r})v_{k-1}$, then TTB is an optimal strategy for the class of decision-making problems characterized in Definition 2.* ■

In simple terms, Proposition 6 formally establishes a distribution-free result granting the optimality of TTB (when generalized to the N-object setting, each with discrete, multi-level attribute values) with respect to the broad class of prediction problems characterized in Definition 2 (with interactions between attributes also accounted for).

Next, Definition 3 formally characterizes a broad class of predictions problems under the *noisy-world* setting wherein the noise component contaminating the prediction problem has a particular structured form: Gaussian distribution.

Definition 3. (Objective function) Let $\forall i \in \{0, 1\}$, O_i denote the two objects a decision-maker should choose from, and $O_{ij} \in \{0, 1\}$ denote the value of the j^{th} attribute of object O_i where $1 \leq j \leq M$. Additionally, let w_k denote the weight corresponding to the k^{th} attribute, and C_i denote the score the object O_i attains on the criterion of interest to the prediction task (e.g., the population of a city, if the prediction task is to predict which of two cities has a higher population). Finally, let $\chi(\cdot)$ be an *arbitrary* monotonically-increasing function (i.e., $\forall x\colon \frac{d}{dx}\chi(x) > 0$). Then, consider the class of prediction problems satisfying the following:

$$C_i :\triangleq \chi\Big(\sum_{j=1}^{M} w_j O_{ij}\Big) + \varepsilon, \quad \varepsilon \overset{d}{\sim} \mathcal{N}(0, \sigma^2). \tag{5}$$

Proposition 7. (Noise-level-independent probabilistic guarantee) *Consider the class of prediction problems formally characterized in Definition 3. Then, for any noise variance $\sigma^2 > 0$, the following statement holds true: If $\forall k\ w_k = v_k$, and $\exists r \in \mathbb{R}^{>2}$ s.t. $\forall k\ v_k \leq (\frac{1}{r})v_{k-1}$, then the probability with which TTB correctly selects the superior object is ≥ 0.5.*

Proposition 7 formally establishes the following important result for the inherently-noisy world characterized in Definition 4: For *any* noise level σ^2, TTB dominates the *selection-purely-by-chance* strategy which select the winning object uniformly at random. This result importantly demonstrates that, independent of noise level σ^2, the adoption of TTB (instead of the selection-purely-by-chance strategy) is rationally justified for the inherently-noisy, class of prediction problems formally characterized in Definition 3.

Definition 4 below formally characterizes a broad class of predictions problems, once again, under the noisy-world setting; this time, however, the noise component contaminating the prediction problem has an *un*structured form.

Definition 4. (Probabilistic guarantee) Let $\forall i \in \{0, 1\}$, O_i denote the two objects a decision-maker should choose from. Let also $O_{ij} \in \{0, 1\}$ denote the value of the j^{th} attribute of object O_i, w_k denote the weight corresponding to the k^{th} attribute, and C_i denote the score the object O_i attains on the criterion of interest to the prediction task (e.g., the population of a city, if the prediction task is to predict which of two cities has a higher population). Finally, let $\phi(\cdot)$ be a monotonically-increasing function (i.e., $\forall x\colon \frac{d}{dx}\phi(x) > 0$). Then, consider the class of prediction problems satisfying the following:

$$\mathbb{P}(C_1 > C_2 | \phi(\sum_{j=1}^{M} w_j O_{1j}) > \phi(\sum_{j=1}^{M} w_j O_{2j}) \geq 1 - \eta, \quad 0 < \eta \ll 1.$$

Proposition 8. (Distribution-free guarantee) *Consider the class of prediction problems formally characterized in Definition 4. Let also P denote a joint probability distributions over the set of all attribute values $\{O_{ij}\}_{i,j}$, i.e., $\{O_{ij}\}_{i,j} \overset{d}{\backsim} P$. Then, for* any *joint probability distribution P the following statement holds true: If $\forall k\ w_k = v_k$, and $\exists r \in \mathbb{R}^{>2}$ s.t. $\forall k\ v_k \leq (\frac{1}{r})v_{k-1}$, then the probability with which TTB mistakenly selects the inferior object is less than η, where $0 < \eta \ll 1$.* ■

Proposition 8 establishes a distribution-free condition sufficient to grant that the probability of TTB erring in a prediction task belonging to the class of problems characterized in Definition 4 is minuscule.

4 General Discussion

In this work, we presented a research program, dubbed formal science of heuristics (FSH), that nicely complements the influential ecological rationality research program (Todd & Gigerenzer, 2007), developing it into a much analytically-richer scientific endeavor. By pursuing its two stated goals (see Introduction section), FSH seeks to (i) mathematically

delineate the key premise ecological rationality rests on—that heuristics are well-matched to the environments in which they are adopted (Todd & Gigerenzer, 2007)—and (ii) establish the strongest analytical results supporting this premise. After all, to rigorously and thoroughly answer whether a heuristic is well-matched to its environment, we need to formally characterize the *broadest* class of environments for which that heuristic performs (near) optimally, and experimentally investigate how often people use that heuristic in such environments.

Instantiating FSH with the well-known Take The Best (TTB) heuristic, and contrary to past analytical work, in this work we considered a much broader class of prediction problems involving nonlinear objective functions (Definitions 1-4) with interactions between attributes also being accounted for (Definition 2). For the classes discussed above, we formally established conditions granting the optimality of TTB when dealing with both non-binary, discrete attribute values (Propositions 3, 5, and 6) and continuous attribute values (Proposition 4). We also analytically investigated a broad class of prediction problems—involving both structured (Definition 3) and unstructured noise (Definition 4)—for which only *probabilistic* guarantees can be provided. Additionally, and in sharp contrast to past analytical work, we also provided distribution-free guarantees on TTB for several classes of prediction problems (Propositions 5–6, and 8).

Our work also serves as a potential template for how FSH could be pursued: For a given heuristic, formally characterize the class of decision-making problems with respect to which the performance of the heuristic is to be analytically investigated, followed by analytical results rigorously delineating the extent to which that heuristic is performing (near) optimally for that class. A generic approach would be to start with a narrow class (containing a set of restricted problems) for which a heuristic is performing (near) optimally; and then gradually expand that class into a larger one and see if previously established performance guarantees still hold (or to establish new performance guarantees, in case they fail to hold). A similar approach has been widely and productively used in theoretical computer science and computational complexity theory, e.g., through formally introducing many complexity classes, with one class serving as a relaxation of another.

Our particular focus on TTB in this work was only meant to showcase how the mindset advocated by FSH could be pursued in the case of a given heuristic—in our case, the Take The Best (TTB) heuristic. Ultimately, a serious investigation of FSH should lead to having mathematically rigorous answers to the two stated goals of FSH for *every* experimentally well-documented heuristic that people use, e.g., the Tallying heuristic (Gigerenzer & Gaissmaier, 2011), the Priority heuristic (Katsikopoulos & Gigerenzer, 2008), the Recognition heuristic (Gigerenzer & Gaissmaier, 2011), and the Minimalist heuristic (Gigerenzer et al., 2008). By now, a large number of heuristics are documented in the literature, many of which still lack an adequate characterization of the environmental conditions under which they are (near) optimal and/or how deviations from those conditions would lead to performance degradation. Thus, future work following FSH should address this analytical shortcoming.

Rieskamp and Otto (2006) show that people are sensitive to the distribution of cues in an environment, appropriately applying either TTB or a weighted additive mechanism, depending on which will be more accurate. However, how people are able to determine which type of environment they are in has largely remained an open question. Establishing distribution-free guarantees, as advocated by FSH, sheds new light on this open question, by formally demonstrating that a heuristic may well yield adequate performance despite the decision-maker's possibly incomplete (or, in the worst case, erroneous) assumptions about her environmental conditions, thereby liberating her from having a thorough understanding of her environment—a more psychologically plausible assumption. For example, Proposition 5 establishes a condition granting the optimality of TTB (with respect to the class of problems characterized in Definition 1) that holds true for *any* joint probability distribution $\mathbb{P}$ on the set of attribute values. This result has an important implication: Even if the decision-maker makes wrong assumptions about the true underlying distribution $\mathbb{P}$ governing the set of attribute values, adopting TTB still remains to be the optimal strategy (for the class of problems characterized in Definition 1). Crucially, the latter statement remains valid regardless of how wrong the decision-maker's assumptions about $\mathbb{P}$ are.

We must note, however, that the present work (and pursuit of FSH, in general) does not address the recent conundrum raised by Otworowska et al. (2018) regarding the computational intractability of the Adaptive Toolbox theory. As Otworowska et al. (2018) analytically demonstrate, there exists no efficient (i.e., polynomial-time) process that can adapt toolboxes to be ecologically rational for *all* possible environments. A resolution of this complexity-theoretic conundrum might be attained by restricting the class of environments under consideration, based on the psychologically plausible assumption that the range of environments humans have to deal with is undoubtedly vast, but *not* arbitrary.

Pursuit of FSH would have important implications for experimental wok on heuristics: Every analytical result (however general it may be) is established under a particular set of assumptions the validity of which needs to be experimentally confirmed. Experimental work should therefore investigate the empirical validity of such assumptions. Likewise, an empirical disconfirmation of an assumption on which an analytical result rests should call for the development of new empirically-grounded formal results. Accordingly, pursuit of FSH yields new experimental work, and, conversely, those experimental findings guide the development of new analytical results—a synergetic scientific endeavor.

Finally, pursuit of FSH allows mathematicians and theoretical computer scientists to make important contributions to the science of heuristics by developing a mathematically-

rigorous foundation for the effectiveness of heuristics in everyday life decisions. As such, we hope that FSH paves the way for having a highly interdisciplinary research program on heuristics wherein analytical and experimental studies, hand in hand, deepen our understanding of the effectiveness of the heuristics we live by. We see our work as a step toward that.

Investigations into human judgment and decision-making have led to the discovery of a multitude of cognitive biases and fallacies, with new ones continually emerging, leading to a state of affairs which can be characterized as the cognitive fallacy zoo! Recently, we have formally presented a principled way to bring order to this zoo (Nobandegani, Campoli, & Shultz, 2019). The work presented here, together with recent formal advances on bringing systematic order to the cognitive fallacy zoo (Nobandegani, Campoli, & Shultz, 2019), suggest a fresh formal approach to pursuing the heuristics-and-biases research program: an approach which aims to lay the formal foundations of the "unreasonable" effectiveness of the heuristics we live by, and to bring mathematically-rigorous systematic order to the cognitive biases ensued by those heuristics.

References

Baucells, M., Carrasco, J. A., & Hogarth, R. M. (2008). Cumulative dominance and heuristic performance in binary multiattribute choice. *Operations Research*, *56*(5), 1289–1304.

Bröder, A. (2000). Assessing the empirical validity of the" take-the-best" heuristic as a model of human probabilistic inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(5), 1332.

Bröder, A. (2003). Decision making with the" adaptive toolbox": influence of environmental structure, intelligence, and working memory load. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(4), 611.

Bröder, A., & Gaissmaier, W. (2007). Sequential processing of cues in memory-based multiattribute decisions. *Psychonomic Bulletin & Review*, *14*(5), 895–900.

Bröder, A., & Schiffer, S. (2003). Take the best versus simultaneous feature matching: Probabilistic inferences from memory and effects of reprensentation format. *Journal of Experimental Psychology: General*, *132*(2), 277.

Chater, N., Oaksford, M., Nakisa, R., & Redington, M. (2003). Fast, frugal, and rational: How rational norms explain behavior. *Organizational Behavior and Human Decision Processes*, *90*(1), 63–86.

Czerlinski, J., Gigerenzer, G., & Goldstein, D. G. (1999). How good are simple heuristics? In *Simple Heuristics that Make us Smart* (pp. 97–118). Oxford University Press.

Evans, J. S. B., & Over, D. E. (2010). Heuristic thinking and human intelligence: a commentary on marewski, gaissmaier and gigerenzer. *Cognitive Processing*, *11*(2), 171–175.

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology*, *62*, 451–482.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: models of bounded rationality. *Psychological Review*, *103*(4), 650.

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: a brunswikian theory of confidence. *Psychological Review*, *98*(4), 506.

Gigerenzer, G., Martignon, L., Hoffrage, U., Rieskamp, J., Czerlinski, J., & Goldstein, D. G. (2008). One-reason decision making. In: Handbook of Experimental Economics Results. In C. Plott & V. Smith (Eds.), (1st ed., Vol. 1, pp. 1004–1017). North Holland.

Gigerenzer, G., & Selten, R. (2002). *Bounded rationality: The adaptive toolbox*. MIT press.

Gigerenzer, G., & Todd, P. M. (1999). *Simple Heuristics that Make us Smart*. New York: Oxford University Press.

Hoffart, J. C., Rieskamp, J., & Dutilh, G. (2018). How environmental regularities affect people's information search in probability judgments from experience. *Journal of Experimental Psychology. Learning, Memory, and Cognition*.

Hoffrage, U., & Reimer, T. (2004). Models of bounded rationality: The approach of fast and frugal heuristics. *Management Revue*, 437–459.

Hogarth, R. M., & Karelaia, N. (2005). Simple models for multiattribute choice with many alternatives: When it does and does not pay to face trade-offs with binary attributes. *Management Science*, *51*(12), 1860–1872.

Hogarth, R. M., & Karelaia, N. (2006). take-the-best and other simple strategies: Why and when they work well with binary cues. *Theory and Decision*, *61*(3), 205–249.

Katsikopoulos, K. V., & Gigerenzer, G. (2008). One-reason decision-making: Modeling violations of expected utility theory. *Journal of Risk and Uncertainty*, *37*(1), 35.

Kearns, M. J., Vazirani, U. V., & Vazirani, U. (1994). *An introduction to computational learning theory*. MIT press.

Martignon, L., & Hoffrage, U. (2002). Fast, frugal, and fit: Simple heuristics for paired comparison. *Theory and Decision*, *52*(1), 29–71.

Mousavi, S., Gigerenzer, G., & Kheirandish, R. (2016). Rethinking behavioral economics through fast-and-frugal heuristics. *Routledge Handbook of Behavioral Economics*, 280–296.

Nobandegani, A. S., Campoli, W., & Shultz, T. R. (2019). Bringing order to the cognitive fallacy zoo. In *Proceedings of the* 17^{th} *International Conference on Cognitive Modeling*. Montreal, QC.

Nosofsky, R. M., & Bergert, F. B. (2007). Limitations of exemplar models of multi-attribute probabilistic inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(6), 999.

Otworowska, M., Blokpoel, M., Sweers, M., Wareham, T., & van Rooij, I. (2018). Demons of ecological rationality. *Cognitive Science*, *42*(3), 1057–1066.

Pachur, T., Todd, P. M., Gigerenzer, G., Schooler, L., & Goldstein, D. G. (2011). The recognition heuristic: A review of theory and tests. *Frontiers in Psychology*, *2*, 147.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1988). Adaptive strategy selection in decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(3), 534.

Payne, J. W., R., B. J., & J., J. E. (1994). *The adaptive decision maker*. New York: Cambridge University Press.

Rieskamp, J., & Otto, P. E. (2006). Ssl: a theory of how people learn to select strategies. *Journal of Experimental Psychology: General*, *135*(2), 207.

Shah, A. K., & Oppenheimer, D. M. (2008). Heuristics made easy: An effort-reduction framework. *Psychological Bulletin*, *134*(2), 207.

Todd, P. M., & Gigerenzer, G. (2000). Précis of simple heuristics that make us smart. *Behavioral and Brain Sciences*, *23*(5), 727–741.

Todd, P. M., & Gigerenzer, G. (2007). Environments that make us smart: Ecological rationality. *Current directions in psychological science*, *16*(3), 167–171.

Tversky, A. (1969). Intransitivity of preferences. *Psychological Review*, *76*(1), 31.

Valiant, L. G. (1984). A theory of the learnable. *Communications of the ACM*, *27*(11), 1134–1142.

The Evolutionary Dynamics of Cooperation in Collective Search

Alan N. Tump[1,*] **(tump@mpib-berlin.mpg.de), Charley M. Wu**[1,*] **(cwu@mpib-berlin.mpg.de),**
Imen Bouhlel[2], & **Robert L. Goldstone**[3]
[1]Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin, Germany
[2]Université Côte d'Azur, CNRS, GREDEG, Nice, France
[3]Department of Psychological and Brain Sciences, Indiana University, Bloomington, USA
[*]These authors contributed equally

Abstract

How does cooperation arise in an evolutionary context? We approach this problem using a collective search paradigm where interactions are dynamic and there is competition for rewards. Using evolutionary simulations, we find that the unconditional sharing of information can be an evolutionary advantageous strategy without the need for conditional strategies or explicit reciprocation. Shared information acts as a recruitment signal and facilitates the formation of a self-organized group. Thus, the improved search efficiency of the collective bestows byproduct benefits onto the original sharer. A key mechanism is a visibility radius, where individuals have unconditional access to information about neighbors within a limited distance. Our results show that for a variety of initial conditions—including populations initially devoid of prosocial individuals—and across both static and dynamic fitness landscapes, we find strong selection pressure to evolve unconditional sharing.

Keywords: Collective search; cooperation; evolutionary simulations; pseudo-reciprocity; prosociality; swarm intelligence

Introduction

Social behavior is structured by the dynamics of the environment and how we interact with one another. Strategies that thrive in one context may be poorly suited to others. How do social behaviors arise in an evolutionary context? And can the dynamics of social interactions support the emergence of cooperation without appealing to conditional strategies?

Evolution is often summarized as "survival of the fittest", evoking a notion of fierce competition between individuals. Where is there room for prosociality and cooperation in the midst of evolutionary competition? One of the early challenges for Darwin's theory of evolution (1859) was to explain the origin of prosocial adaptations that improve the welfare of others or one's group as a whole, but at a potential cost to the individual. Darwin's explanation appealed to the notion of *group selection*, where the costs of altruism are ultimately justified by increased fitness for the group (Darwin, 1871). Thus, groups with more prosocial members may outcompete rival groups. Although group selection offers a potential pathway for the emergence of cooperation, it often requires strong assumptions, such as stable group structures and strong competition between groups (Janssen & Goldstone, 2006). Without these assumptions, selection at the individual level can undermine group selection. Thus, a comprehensive understanding of prosociality requires a theory of individual selection (Wilson & Wilson, 2007).

Theories of Cooperation

One traditional explanation for individual selection of prosociality is through the mechanism of *kin selection* (also known as inclusive fitness), where recipients of altruistic acts tend to be genetically related to the donor (Nowak, 2006). Hamilton's law (1964) states that the costs of prosociality C must be justified relative to the benefits of the recipient B by accounting for the relatedness of individuals r such that $\frac{C}{B} < r$. While kin selection explains prosociality between genetically similar individuals, Hamilton's law alone fails to account for all the social behaviors we see in human society (Rand & Nowak, 2013; Fehr & Fischbacher, 2003) and in animals (e.g., Spottiswoode, Begg, & Begg, 2016; Brown, Brown, & Shaffer, 1991). Many mechanisms have been proposed in order to justify the evolution of cooperation towards non-relatives, typically requiring an initial investment of a donor towards a non-related individual with expectations of reciprocity or benefits.

Conditional Cooperation. Theories of conditional cooperation operate on expectations of future reciprocity, where seemingly prosocial behavior is ultimately grounded in self-interest. Often described as impure altruism (Andreoni, 1989), both direct and indirect reciprocity appeal to conditional strategies (e.g., tit for tat; Nowak & Sigmund, 1992), where individuals conditionally cooperate with each other, so long as future reciprocation is expected. Direct reciprocity depends on multiple interactions with the same individual, while indirect reciprocity typically relies on reputation systems, where cooperative behavior is used as a social signal to third-parties (Nowak & Roch, 2007). Conditional cooperation has been widely studied in the context of game theory, yet simple mechanisms of social or spatial dynamics can also explain the origins of cooperation (Nowak & May, 1992).

Unconditional Cooperation. Theories of unconditional cooperation explain the origin of prosocial behavior through changes in the interaction structure for the donor (Perc, Gómez-Gardeñes, Szolnoki, Floría, & Moreno, 2013). Thus, behaving prosocially can make it more likely to interact with other prosocial individuals. *Network reciprocity* operates on similar principles as kin selection, but where the cost-benefit ratio is defined relative to interaction partners (Nowak, 2006). This approach has shown that by situating agents on a network (Ohtsuki, Hauert, Lieberman, & Nowak, 2006) or in a

spatial landscape (Nowak & May, 1992), prosocial individuals tend to interact more with similar partners, thus creating self-organized regions where prosociality proliferates (Perc et al., 2013). It is also possible to replace spatial similarity or network connectivity with some arbitrary feature or tag (Riolo, Cohen, & Axelrod, 2001), such that individuals with similar features are more likely to interact with one another. This provides a useful bridge between individual and group level mechanisms, because it describes how groups can form based on spatial, network, or feature similarity.

Two key assumption are made by these theories. The first is that the initial population already includes multiple prosocial individuals (Nowak & May, 1992; Ohtsuki et al., 2006). Yet this doesn't answer the crucial question of how cooperation emerges *ex nihilo*. Secondly, the interaction structures are more or less static: agents are either embedded in some spatial location (Nowak & May, 1992), as a fixed node in a network (Ohtsuki et al., 2006; Barkoczi, Analytis, & Wu, 2016), or given a fixed feature tag (Riolo et al., 2001). While groups can still emerge through the dynamics of evolution, interaction partners remain relatively stationary (but see Janssen & Goldstone, 2006) and individual dynamics (e.g., search behavior) are largely unaccounted for.

Pseudo-reciprocity is a related theory of unconditional cooperation, where the key difference from network reciprocity is that the fitness of the donor does not depend on the phenotype of the recipient. Thus, prosocial behavior can be beneficial without depending on the presence of other prosocial individuals in a group. Prosociality can alter the social environment for the donor (e.g., by sharing information about resources), such that the donor gains byproduct benefits through self-interested behavior of the recipients (Connor, 1986; Brown et al., 1991). For example, Cliff Swallows (*Hirundo pyrrhonota*) share information about the location of insect swarms through a unique vocal signal (i.e., a food call), which attracts other peers. While it is difficult to track the insect swarms individually, the collective recruited by the information sharer tracks the swarm more efficiently. Hence, even without expectations of reciprocity (i.e., future vocal signals from peers), each individual benefits by behaving prosocially and sharing information (Brown et al., 1991). Thus, pseudo-reciprocity offers a mechanism where individuals can be unconditionally prosocial towards all the members of the group, rather than towards a restricted set of cooperative partners.

Goals and Scope

Here, we analyze the emergence of cooperation through sharing information. We use evolutionary simulations to study how individual selection pressure can give rise to sharing, even from initial populations void of prosocial individuals. We simulate agents searching for rewards on a high dimensional fitness landscape, where the flow of information is dynamically and spatially defined. Agents have a binary phenotype that defines whether or not they share information unconditionally to the rest of the population. We show that this global sharing signal acts as a recruitment mechanism

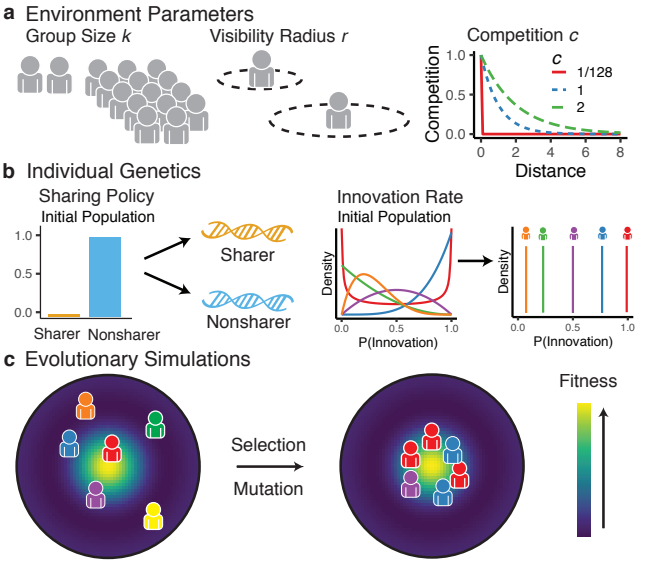


Figure 1: Evolutionary Simulations. **a**) We vary three main environmental parameters: group size, the visibility radius, and competition level. Group size k specifies the number of agents interacting together. Visibility radius r defines the maximum Chebyshev distance between two agents where information can be passively observed. Competition level c defines the decay rate of an exponential competition function that determines how agents split rewards (higher values of c result in splitting over larger distances. **b**) Each agent is defined by a sharing policy (either sharer or non-sharer) and an innovation rate (between 0 and 1). **c**) We use evolutionary simulations over 200 generations to see which individual genes emerge through selection pressure and mutation.

that facilitates the self-organization of dynamic groups. Because groups are more effective at finding rewards than lone individuals, we find that sharing emerges and dominates our evolved populations across a large range of initial conditions and in both static and dynamic fitness landscapes.

Collective Search Simulations

We use a multi-agent framework based on Bouhlel, Wu, Hanaki, and Goldstone (2018), who found that sharing information can be beneficial to the donor, even in competitive contexts and without expectations of reciprocity. The costs of sharing information (through resources lost to competition) can be outweighed by the byproduct benefits of cooperation. A simple coordination mechanism of a local visibility radius (i.e., nearby agents have access to each others' rewards) facilitates the formation of a self-organized collective. Thus, sharing information acts as a recruitment signal, attracting others to the donor, and increasing the likelihood of future social interactions (via the visibility radius). These future interactions are the source of byproducts benefits for the sharer. Here, we use evolutionary simulations and more extreme levels of competition (compared to Bouhlel et al., 2018) in order to study how sharing interacts with innovation, and under which initial conditions there exists individual selection pressure for unconditional sharing, leading to group-level cooperation (Goldstone & Janssen, 2005). Code for reproducing these results is publicly available at https://github.com/alantump/adaptiveSharingEvolution.

Methods

Adopted from Bouhlel et al. (2018), we simulate groups of k agents searching for rewards on a 10-dimensional[1] fitness landscape over $T = 50$ trials. On each trial t, agents can use either individual or social information (see below) to search for rewards on the fitness landscape. Payoffs are proportional to the inverse Manhattan distance of agent i from a global optimum Ω:

$$f(\mathbf{x}_{ti}) = \frac{1}{1 + \|\mathbf{x}_{ti} - \Omega\|_1} \tag{1}$$

where $\mathbf{x}_{ti}$ contains the coordinates for each dimension $m = 1, ..., 10$ of the current location of agent i at trial t. The coordinates of the global optimum Ω are sampled from a uniform distribution $\mathcal{U}(1, 10)$ for each dimension.

Competition. The payoffs $f(\mathbf{x}_{ti})$ are subject to competition, which we implement by having agents split rewards when occupying nearby spaces in the environment. Specifically, we use a competition parameter c that defines an exponentially decaying competition metric $C(\mathbf{x}_{ti}, \mathbf{x}_{tj})$ between each pair of agents i and j:

$$C(\mathbf{x}_{ti}, \mathbf{x}_{tj}) = \exp(-\frac{\left\|\mathbf{x}_{ti} - \mathbf{x}_{tj}\right\|_1}{c}); \tag{2}$$

Larger values of c induce higher competition over larger distances (see Fig. 1a), while in the limit of $c \to 0$, competition only occurs when agents occupy the exact same solution (as in Bouhlel et al., 2018). Splitting of rewards is proportional to the sum of competition values for all other agents. Hence, for location $\mathbf{x}_{ti}$, the acquired reward is:

$$R(\mathbf{x}_{ti}) = \frac{f(\mathbf{x}_{ti})}{1 + \sum_{j \neq i} C(\mathbf{x}_{ti}, \mathbf{x}_{tj})} \tag{3}$$

Individual search. Each agent begins at a random starting location, where each dimension is sampled from a uniform distribution $\mathcal{U}(1, 10)$. On every trial, each agent i stores the location $\mathbf{x}_{tj}$ and reward value $R(\mathbf{x}_{ti})$ of both individually and socially acquired information (see information sharing and visibility radius). We use a local search strategy, where the agent selects the location with the largest observed reward value $\mathbf{x}^*_{ti}$ up until time t, and then has an opportunity to innovate on it by modifying each value in $\mathbf{x}^*_{ti}$ by a discrete value in $\{-1, 0, 1\}$.

We define the *Innovation rate* as the probability that an agent innovates, where otherwise $\mathbf{x}^*_{ti}$ is copied verbatim. If the agent innovates, we modify each dimension of $\mathbf{x}^*_{ti}$ by drawing from a Binomial distribution centered on zero $\sim \mathrm{B}\left(2, \frac{1}{2}\right) - 1$. Intuitively, half of the time there is no change along that dimension, while changes of both -1 or $+1$ are equally likely, each with a probability of 25%.

[1] Bouhlel et al. (2018) studied environments of different dimensionality, while here we use 10-dimensional environments as a prototypical example.

Social information. Depending on their sharing policy, agents are deterministically either sharers or non-sharers. Sharers will unconditionally share information about both reward location $\mathbf{x}_{ti}$ and value $R(\mathbf{x}_{ti})$ to all other agents, while non-sharers will withhold it. Sharing information is associated with an increased cost due to splitting rewards with imitators, but can also confer byproduct benefits by broadcasting high quality solutions, which are subsequently modified by group members and improved upon, before being transmitted back via the visibility radius or by other sharers.

In addition to the global sharing signal, we use a visibility radius as a feature of the environment. At each trial t, agents passively provide information about reward locations and magnitudes to other agents that are within visibility radius r. For any two agents $i \neq j$, agent j is visible to agent i if the maximal distance between the two agents on any dimension (i.e., the Chebyshev distance) is not greater than the visibility radius r:

$$D_{Chebyshev}(\mathbf{x}_{ti}, \mathbf{x}_{tj}) = \max_m |d^t_{mi} - d^t_{mi}| \leq r \tag{4}$$

The visibility radius is a coordination mechanism that allows for localized transmission of information. Whereas the sharing signal is a global mechanism operating at all distances, the local visibility radius allows for dynamic interaction structures to emerge and facilitates the spontaneous formation of spatially coherent groups. Crucially, given the high dimensionality and size of the search space, it is unlikely for any two agents to fall within the same visibility radius without explicit information sharing. For example, there is 0.1% probability of two agent being visible to one another at initialization for a radius of 2.

Evolutionary Simulations

Inspired by biological evolution, we embed the simulation framework in an evolutionary algorithm, which uses selection pressure and mutations over multiple generations to discover which sets of behavioral parameters evolve. The evolutionary algorithm is well suited for our research question because fitness-maximizing behavior (e.g., willingness to share information) depends on the behavior of others in a game theoretical context.

Initial conditions. Beginning with a population of 300 agents, each agent carries genes determining innovation rate and sharing policy (i.e., sharer or non-sharer). We start with an initial population consisting of only non-sharers to address the question of how cooperative behavior can emerge *ex nihilo* through individual selection. We vary the initial mean innovation rate in the populations to ensure that the results of the evolutionary algorithm are not dependent on the starting conditions. The initial values for innovation rate were sampled from a Beta distribution, with the mean of the distribution sampled from a uniform distribution $\mathcal{U}(0, 1)$.

For each generation, we repeatedly sample k agents from the whole population. We simulate these agents performing

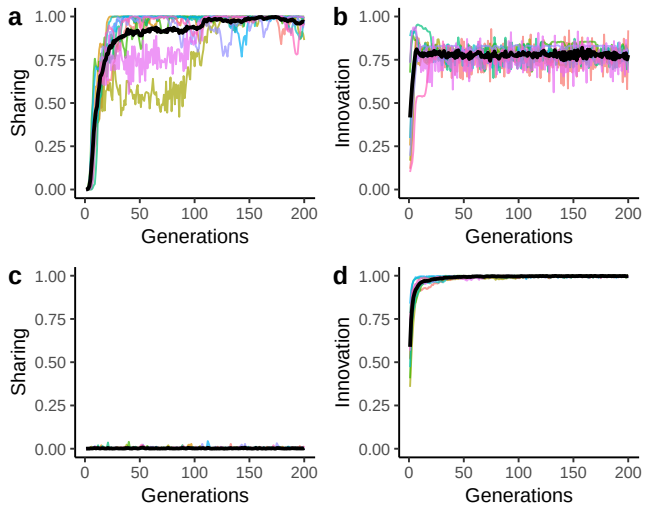


Figure 2: Evolution of sharing and innovation over 200 generations. **a**-**b**) An example where populations evolve high sharing and innovation rates, with group size $k = 6$, visibility radius $r = 4$ and competition level $c = 1/128$. **c**-**d**) An example where individuals adopted high innovation rates but did not evolve sharing, based on group size $k = 6$, visibility radius $r = 0$ and competition level $c = 2$. Each colored line represents the average parameter value within a population, while the black line indicates the average across populations.

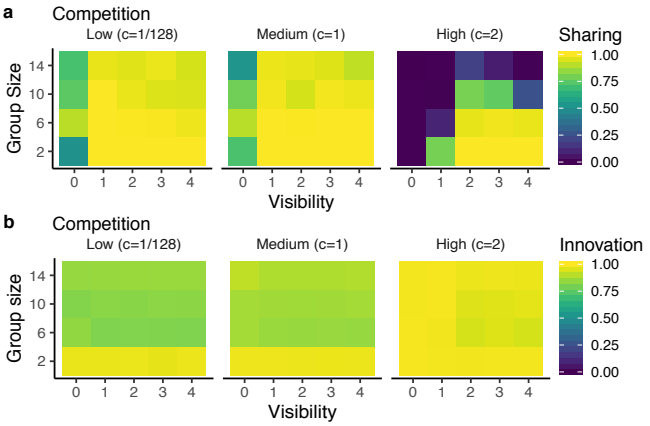


Figure 3: Equilibrium results for different combinations of environmental parameters. **a**) Agents evolved high sharing rates in low and medium competitive environments, although sharing was found in more restricted contexts under high competition (requiring smaller groups and larger visibility radius values). **b**) Overall, we find high levels of innovation, although we also see the trend that larger groups evolve slightly lower innovation rates.

collective search, where behavior is determined by their genetic makeup (innovation rate and sharing policy). We repeat the simulation procedure over $\frac{300}{k} \times 5$ repetitions, resulting in approximately 5 simulations per agent in each generation.

Selection and mutation. We select the agents with the highest fitness to produce genetically similar offspring via tournament selection. In this selection procedure, we repeatedly sample 7 random individuals from the population, whereby the individual with the highest relative performance passes its genes onto the next generation. This selection process is repeated 300 times in order to produce a new generation of 300 agents. The genes of the new generation are exposed to weak mutation to consistently ensure gene variation, where each gene has a probability of mutation. The sharing gene mutates with $p = .002$, whereby a new sharing policy is drawn from a binomial distribution $\sim \mathrm{B}\left(1, \frac{1}{2}\right)$, with the new policy equally likely to be sharer or non-sharer. The innovation gene mutates with $p = .02$, whereby the previous innovation is modified by adding Gaussian noise $\sim \mathrm{N}(0, 0.2)$. The innovation rate was truncated between [0,1]. Note that we chose the mutation probabilities and strengths to be high enough to ensured constant variation in the gen pool.

The genetic algorithm repeats the process of fitness evaluation, selection, and then reproduction with mutation for 200 generations to ensure the population converges to a stable outcome. We ran 10 replications of this procedure and report the average evolved parameters of the last 10 generations (i.e., generations 190 to 200) over each of the 10 replications. We systematically varied group size ($k \in [2, ...14]$), visibility radius ($r \in [0, 1, 2, 3, 4]$), and competition level (low $= \frac{1}{128}$; medium = 1; high = 2) to investigate how the structure of the environment influences the selection of individual characteristics (sharing and innovation).

Results

When exposed to selection pressure via the evolutionary algorithm, the populations evolved different sharing and innovation rates depending on the environmental parameters (see Fig. 2 for examples). Figure 3 shows the proportion of sharers and the innovation rate at equilibrium for different parameter combinations, where yellow tiles indicate high levels of either sharing or innovation.

Sharing evolves ex nihilo. Starting from initial conditions of no sharers in the population, we find that sharing emerges in the overwhelming majority of our simulation parameters, and that sharing often dominates the population at close to ceiling levels (Fig. 3a). However, we also discover the limits of sharing as an adaptive strategy as we increase the level of competition for rewards. Under high levels of competition, only smaller groups with larger visibility radius are able to support sharing.

Sharing and innovation co-evolve. We find that over the entire parameter space, all populations evolved high innovation rates (Fig. 3b), although not at ceiling level (i.e., yellow tiles) compared to sharing behavior. Looking more closely, we find relatively higher innovation rates in small groups compared to large groups, with this effect most pronounced under low or medium levels of competition. Yet, how are sharing and innovation behaviors related to each other?

To further understand the interaction between strategies, we ran additional simulations with innovation rate fixed at low (25%), medium (50%) or high (100%) values. The results are shown in Figure 4, where we replicate the main findings of the previous simulation for high innovation rate (top row). However, we find that sharing becomes substantially

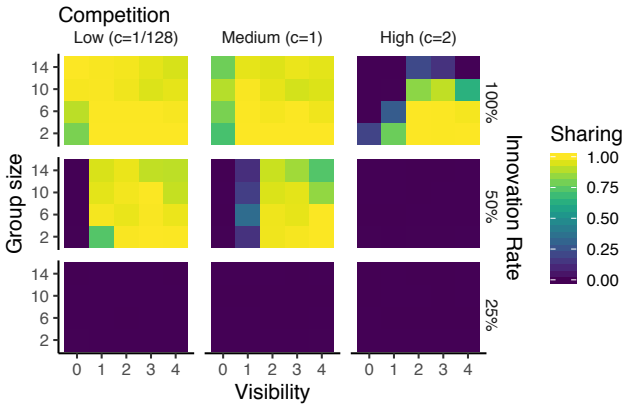


Figure 4: Equilibrium results of sharing for fixed innovation rates (rows) at various parameter combinations. When the innovation rate is fixed at 100%, we largely replicate the results in Figure 3. However, when innovation is fixed at 50%, we find that sharing evolves in a more restricted set of parameters and exclusively with a visibility radius of 1 or larger. When there is an innovation rate of 25%, we find virtually no emergence of sharing.

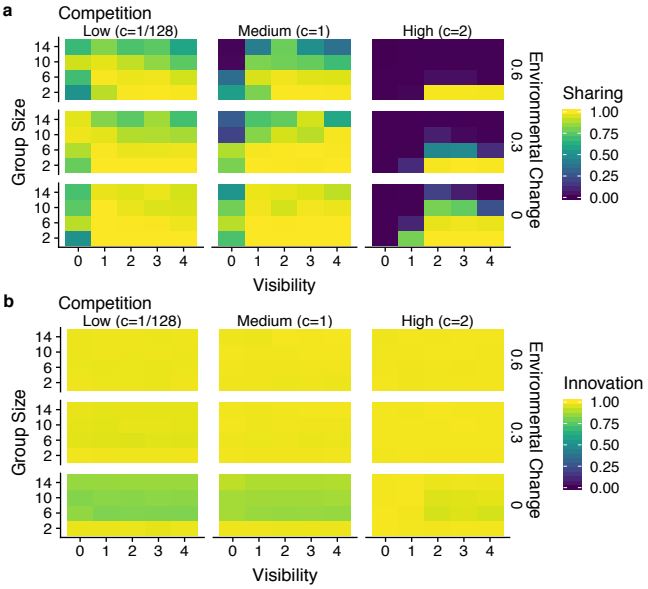


Figure 5: Equilibrium results in a dynamic environment. **a**) Again we find high sharing rates in low and medium competitive environments, but now higher rates of environmental change reduced the levels of sharing in the population. We also see stronger indications of an interaction between group size and visibility radius, where a larger visibility radius is required to coordinate larger groups and support a larger sharing population. **b**) Across all parameters, we find high levels of innovation emerge, although lower competition and larger groups reduces the extent of innovation.

less adaptive for populations with innovation fixed at low or medium levels. Thus, innovation is an essential ingredient for prosocial traits to develop, as has been shown in previous work on cultural transmission through iterative cycles of imitation and innovation (Ehn & Laland, 2012; Wisdom & Goldstone, 2011; Derex, Feron, Godelle, & Raymond, 2015).

Interim conclusion

We show that sharing can evolve across a variety of different environments and in mixed groups with different proportions of sharers and non-sharers. The selection pressure for sharing can lead to it becoming a dominant trait prevalent in the vast majority of the population. The spatial dynamics of this simulation framework facilitated by a visibility radius lead to a setting where selection pressure does not prioritize free-riding and the group does not succumb to a tragedy of the commons.

Dynamic Simulations

We now extend the framework to account for a changing environment, implemented by a wandering global optima. We define the global optima Ω^t and modify it on each time t with a probability determined by the environmental change rate p_e. With probability p_e, the environment's global optima changes, otherwise it stays the same ($\Omega^{t+1} = \Omega^t$). When the environment changes, each coordinate of the global optima $d^t_m \in \Omega^t$ has a 50% probability of being modified by +1 or -1, and a 50% probability of remaining the same. This is the same as the local search rule used by individual agents.

In order to account for the decreasing validity of past observations in a changing environment, we introduce a temporal discount rate γ. Thus, the history of past observations maintained by each agent decays as a function of the elapsed time:

$$\hat{R}(x_{ti}) = \gamma^{\tau} R(x_{ti}) \quad (5)$$

where $\hat{R}(x_{ti})$ is the discounted reward and τ is the elapsed time between the observation and the current time. Thus, agents locally search around the reward location that has the largest discounted reward $\hat{R}(x_{ti})$. Both individually and socially acquired information follow the same decay rate. In our simulations we fixed the Discount rate $\gamma = .99$, which approximately corresponds to a 10% discount after 10 trials.

Dynamic Results

Figure 5 shows the equilibrium results of our dynamic simulations. Again, we find that sharing is a beneficial strategy under many environmental conditions (Fig. 5a). Similar to the static case, there are limits to the conditions under which sharing emerges, particularly in highly competitive environments. The relationship between the visibility radius and group size becomes increasingly important, where a larger radius allows sharing to emerge in larger groups. We also observe that the evolved proportion of sharers decreases in more volatile environments (higher change rates) and in larger groups. This interaction is not observed in the static environment, but may be partially due to the increased difficulty of coordination and because out-of-date information can harm instead of help others (Boyd & Richerson, 1988; Henrich & Boyd, 1998). Additionally, we find that environmental change increases the evolved innovation rates (Fig. 5b). The intermediate levels of innovation found in the static simulations are eclipsed by even higher rates under environmental change.

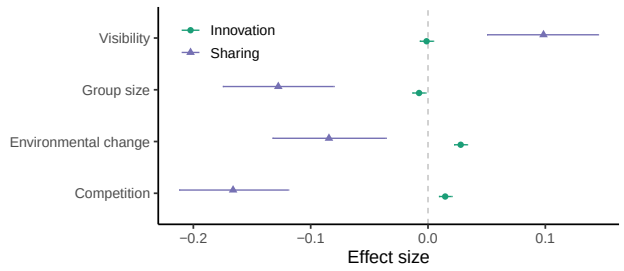


Figure 6: Regression results. The estimated effect sizes of environmental parameters on innovation rate (green) and sharing (purple). Error bars show 95% CI.

General Discussion

We use evolutionary simulations to show that for a variety of initial conditions and across both static and dynamic fitness landscapes, there exists individual selection pressure for the unconditional sharing of information. To summarize the effects of each environmental parameter on the equilibrium characteristics of innovation and sharing, we fit a linear model on the dynamic simulation results (Fig. 6). The size of the visibility radius contributes positively to the rate of sharers in the evolved population, while group size, environmental change, and competition all reduce the rate of sharers. Thus, the evolution of cooperation in the absence of reciprocity operates at a fine balance between coordination (via the visibility radius) and discord (through competition and the communication of out-of-date information).

In comparison, the environmental effects on innovation are relatively small. We find relatively high levels of innovation in all simulations. Environmental change had the strongest influence on innovation, while higher competition also increased innovation. Rather, the more interesting result of our simulations involves the interaction between innovation and sharing, which co-evolve and are dependent on one another for producing the emergent behavior of collective search.

How do the dynamics of cooperation work? To get a deeper understanding of how sharing improves the welfare of the donor, we present a vignette of an agent who is either a sharer or a non-sharer in a population of non-sharers (Fig. 7). The sharer transmits a global signal that recruits peers and gathers them within visible range (Fig. 7a, orange line). This means that a sharer will have access to more social information compared to a non-sharer by being closer to others (Fig. 7a, blue line). Since we find high rates of innovation in all simulations, any imitated information is also tweaked and modified. Some of these modifications will improve upon the originally copied solution. This creates a feedback cycle of solutions that are consistently improved over time, which can benefit the original sharer through local transmissions within the visibility radius (Fig. 7b). Compared to a group of non-sharers (blue line), the sharer is able to explore the reward landscape better and achieve higher rewards despite the stronger local competition (Fig. 7c, orange line).

In summary, as is the case with the Cliff Swallows (Brown et al., 1991), sharers recruit peers within their visibility radius and reap the byproduct benefits of passively acquired modifications to the original solution. Intuitively, larger visibility increases the ability of a group to stay connected with one another. However, the global sharing signal is an essential recruitment device that facilitates the formation of a group in the first place. Group coherency facilitated by the visibility radius provides byproduct benefits to the originator of the sharing signal, creating a feedback loop of imitation with innovation.

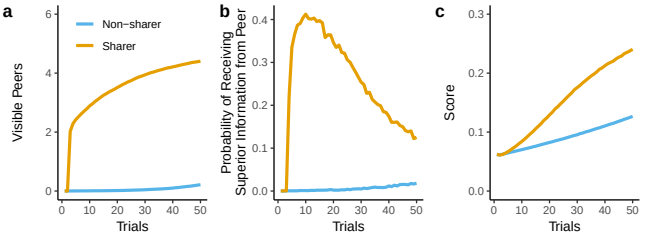


Figure 7: How sharing leads to cooperation. These results are the mean performance over 10,000 replications with a group size of $k = 6$, a visibility radius of $r = 2$, a innovation rate of 1 and a competition level $c = 1$. **a**) The sharer (orange line) attracts other individuals within their visibility radius through the sharing signal, leading to richer informational exchanges than compared to a non-sharer (blue line). **b**) Individuals who imitate the shared information also innovate, and thus passively provide improved information to the sharer through the visibility radius. **c**) As a result, the sharer benefits from passively gained information and acquires an overall higher pay-off compared to individuals in a non-sharing group.

Conclusion

Through the lens of evolution, we show how individual selection pressure can give rise to the unconditional sharing of information. The sharing signal does not require expectations of reciprocity in order to be beneficial, but rather directly benefits the sharer through the byproducts of cooperation. Shared information about a high reward acts as a recruitment signal, which leads to the emergence of a self-organized collective centered on the original donor. A key ingredient is a visibility radius, which allows individuals to observe the rewards of neighbors within a fixed spatial distance. This visibility radius provides a simple yet effective coordination mechanism that is grounded in simple spatial and social dynamics, creating complex patterns of emergent behavior.

More broadly, our results indicate that prosocial behaviour can evolve from initial conditions devoid of other prosocial individuals. While theories explaining the evolution of conditional reciprocity have been very influential (Nowak & May, 1992; Ohtsuki et al., 2006), our results provide an explanation for the initial emergence of prosocial individuals, which is an essential requirement for both conditional cooperation and group selection. Future implementation of conditional strategies in our framework could provide further insight into how various strategies co-evolve.

References

Andreoni, J. (1989). Giving with impure altruism: Applications to charity and ricardian equivalence. *Journal of Political Economy*, *97*, 1447–1458.

Barkoczi, D., Analytis, P. P., & Wu, C. M. (2016). Collective search on rugged landscapes: a crossenvironmental analysis. In A. Papafragou, D. Grodner, D. Mirman, & J. Trueswell (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 918–923). Austin, TX: Cognitive Science Society.

Bouhlel, I., Wu, C. M., Hanaki, N., & Goldstone, R. L. (2018). Sharing is not erring: Pseudo-reciprocity in collective search. In T. T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 156–161). Austin, TX: Cognitive Science Society.

Boyd, R., & Richerson, P. J. (1988). An evolutionary model of social learning: the effects of spatial and temporal variation. *Social learning: psychological and biological perspectives*, 29–48.

Brown, C. R., Brown, M. B., & Shaffer, M. L. (1991). Food-sharing signals among socially foraging cliff swallows. *Animal Behaviour*, *42*, 551–564.

Connor, R. C. (1986). Pseudo-reciprocity: investing in mutualism. *Animal Behaviour*, *34*(5), 1562–1566.

Darwin, C. (1859). *On the Origin of Species by Means of Natural Selection: or the Preservation of Favored Races in the Struggle for Life*. London: Murray.

Darwin, C. (1871). *The Descent of Man and Selection in Relation to Sex*. London: John Murray, Albemarle Street.

Derex, M., Feron, R., Godelle, B., & Raymond, M. (2015). Social learning and the replication process: an experimental investigation. *Proceedings of the Royal Society B: Biological Sciences*, *282*(1808), 20150719.

Ehn, M., & Laland, K. (2012). Adaptive strategies for cumulative cultural learning. *Journal of Theoretical Biology*, *301*, 103–111.

Fehr, E., & Fischbacher, U. (2003). The nature of human altruism. *Nature*, *425*(6960), 785.

Goldstone, R. L., & Janssen, M. A. (2005). Computational models of collective behavior. *Trends in Cognitive Sciences*, *9*(9), 424–430.

Hamilton, W. D. (1964). The genetical evolution of social behaviour. *Journal of Theoretical Biology*, *7*, 1–16.

Henrich, J., & Boyd, R. (1998). The evolution of conformist transmission and the emergence of between-group differences. *Evolution and Human Behavior*, *19*(4), 215–241.

Janssen, M. A., & Goldstone, R. L. (2006). Dynamic-persistence of cooperation in public good games when group size is dynamic. *Journal of Theoretical Biology*, *243*(1), 134–142.

Nowak, M. A. (2006). Five rules for the evolution of cooperation. *Science*, *314*(5805), 1560–1563.

Nowak, M. A., & May, R. M. (1992). Evolutionary games and spatial chaos. *Nature*, *359*, 826–829.

Nowak, M. A., & Roch, S. (2007). Upstream reciprocity and the evolution of gratitude. *Proceedings of the Royal Society of London B: Biological Sciences*, *274*(1610), 605–610.

Nowak, M. A., & Sigmund, K. (1992). Tit for tat in heterogeneous populations. *Nature*, *355*, 250–253.

Ohtsuki, H., Hauert, C., Lieberman, E., & Nowak, M. A. (2006). A simple rule for the evolution of cooperation on graphs and social networks. *Nature*, *441*, 502–505.

Perc, M., Gómez-Gardeñes, J., Szolnoki, A., Floría, L. M., & Moreno, Y. (2013). Evolutionary dynamics of group interactions on structured populations: a review. *Journal of the Royal Society Interface*, *10*, 20120997.

Rand, D. G., & Nowak, M. A. (2013). Human cooperation. *Trends in Cognitive Sciences*, *17*(8), 413–425.

Riolo, R. L., Cohen, M. D., & Axelrod, R. (2001). Evolution of cooperation without reciprocity. *Nature*, *414*, 44–443.

Spottiswoode, C. N., Begg, K. S., & Begg, C. M. (2016). Reciprocal signaling in honeyguide-human mutualism. *Science*, *353*(6297), 387–389.

Wilson, D. S., & Wilson, E. O. (2007). Rethinking the theoretical foundation of sociobiology. *The Quarterly Review of Biology*, *82*(4), 327–348.

Wisdom, T. N., & Goldstone, R. L. (2011). Innovation, imitation, and problem solving in a networked group. *Nonlinear Dynamics-Psychology and Life Sciences*, *15*(2), 229–252.

Absolute Spatial Frames of Reference in Bilingual Speakers of Endangered Ryukyuan Languages: An Assessment via a Novel Gesture Elicitation Paradigm

Rafael E. Núñez (RNUNEZ@Ucsd.Edu)
Department of Cognitive Science, UCSD, 9500 Gilman Drive
La Jolla, CA 92093, USA

Kenan Celik (TAKAMORI.CELIK@Gmail.Com)
National Institute for Japanese Language and Linguistics (NINJAL), 10-2 Midori-cho
Tachikawa City, Tokyo, Japan

Natsuko Nakagawa (NAKAGAWANATUKO@Gmail.Com)
Chiba University, 1-33, Yayoi-cho, Inage Ward
Chiba City, Chiba, Japan

Abstract

We experimentally investigate, by means of a novel gesture-elicitation paradigm, the spontaneous spatial frames of reference (FoRs) used by bilingual individuals who speak Japanese (which has been labeled as a "relative" language) and one of the endangered Ryukyuan languages (Miyako or Shiraho) whose speakers have been reported to routinely use absolute FoRs. How would these last elderly bilingual speakers spontaneously resolve the clashing FoRs the two languages they speak bring forth? We find that despite the fact that Japanese and these Ryukyuan languages have full corresponding grammatical and lexical resources for expressing both, relative and absolute FoR, Ryukyuan speakers tend to markedly prefer the latter gesturally. Methodologically, the results, which are consistent with data obtained with standard FoRs methods, corroborate the reliability of the novel gesture elicitation task, which adds to the battery of techniques for studying FoRs a method that assesses effortless spontaneous real-time cognition with high ecologically validity.

Keywords: spatial construals; gesture; absolute frames of reference; linguistic relativity hypothesis, bilingual; endangered languages; Miyako, Shiraho, Japanese, Japonic languages; Ryukyu islands; elderly participants

Introduction

The study of spatial frames of reference (FoRs) (Levinson, 2003) has received significant attention in the last 20 years. While speakers of some languages have been found to prefer relative FoRs (i.e., egocentric; e.g. 'left' and 'right') to describe or reason about spatial relations among tabletop objects, speakers of others prefer absolute FoRs (i.e., allocentric; e.g. 'north', 'south') (Gumperz & Levinson, 1996; Levinson and Wilkins, 2006). A common interpretation of these results has been that it is language that plays a significant role in structuring the cognition of fundamental domains like space (Majid, Bowerman, Kita, Haun, and Levinson, 2004) —the core of the linguistic relativity hypothesis (Gumperz and Levinson, 1991; Lucy, 1992). This proposal, however, is largely based on the implicit assumption that the human mind is fundamentally monolingual, an assumption that doesn't seem to hold as bi- and multi-linguism have been ubiquitous throughout the history of humanity (Evans, 2011; Pavlenko, 2014). In fact, scholars investigating the linguistic relativity hypothesis have explicitly asked themselves "What are the cognitive consequences of being a bilingual in languages that rely on different frames of reference?" (Majid et al., 2004, p. 113), but no clear answer has been proposed so far, and no significant efforts seem to have been spent in order to address the question properly. Indeed, following the linguistic relativity hypothesis, bilingual individuals who fluently speak languages from the same linguistic family, and which are equipped with exactly the same relevant linguistic resources should not exhibit any marked preference in using relative or absolute FoRs. Here we ask, do fully bilingual individuals who speak such languages spontaneously exhibit any preferences when the linguistic practices of these languages elicit clashing absolute-relative frames of reference?

The last bilingual speakers of endangered languages spoken in the Ryukyus (the chain of islands stretching between Taiwan to Kyushu, Japan) provide a particularly interesting population for addressing this question. The Ryukyuan languages Miyako and Shiraho, are, for instance structurally equivalent to Japanese with respect to lexical spatial encodings. The three languages —all members of the Japonic family— have precise words for left and right, front and back, north and south, etc. (see Table 1), often even sharing cognate words (i.e., sharing the same original root, like "left" and "right" from proto-Japonic **pidari* and **migiri*, respectively). However, while speakers of Japanese have been reported to clearly prefer relative FoRs (Pederson, Danziger, Wilkins, Levinson, Kita, and Senft, 1998), ethnographic descriptions (Suzuki, 1978) as well as empirical psycho-linguistic studies (Celik, Takubo, and Núñez, 2019) have reported that speakers of Ryukyuan languages commonly rely on absolute FoRs. Interestingly, the preference of absolute FoRs of these individuals takes place despite being themselves fluent Japanese-bilinguals, and having been schooled and enculturated into the

mainland Japanese culture for most, if not all of their long lives (Japanese elementary school has been implemented in the Ryukyu islands since the end of 19th century, and there has been an overwhelming presence of mainland Japanese culture in TV and radio for decades).

Table 1: Spatial terms in Japanese, Miyako, and Shiraho

FoR	Translation	Japanese	Miyako	Shiraho
Absolute	east	higashi	agaɪ	anta
	west	nishi	iɪ	inta
	south	minami	pai	penta
	north	kita	nisɪ	niʃanta
Relative	right	migi	ngɪ	neeri
	left	hidari	pɪdaɪ	pitare
Intrinsic/ Relative	side	yoko	juku	jagata/aza
	front	mae	mavkjaa	menta
	behind/back	ushiro	tɕibi, kusɪ	ʃinta
	straight	massugu	massɪgu	meŋga
Intrinsic	be aligned	narab-	narab-	narab-

In this study, we experimentally investigate the spatial FoRs used spontaneously by bilingual Ryukyuan speakers (Miyako-Japanese or Shiraho-Japanese) and by monolingual speakers of standard Japanese from Tokyo. Endangered languages such as the Ryukyuan languages, however, are primarily spoken by elderly people, who are often challenged by problem-solving, reasoning, and memory tasks (Brinley, Jovick, and McLaughlin, 1974) such as the ones used in standard FoRs studies (e.g., "Animals-in-a-row" task (Pederson et al., 1998) or complex arrays of toys (Haun, Rapold, Janzen, and Levinson, 2011)). For this reason, we focus here on spontaneous cognition by means of a novel gesture elicitation paradigm that, while being ecologically valid, adequately assesses effortless spontaneous cognition (McNeill, 1992; Bates and Dick, 2002; Kendon, 2004; Goldin-Meadow, 2005) and can readily be implemented in cross-cultural field studies (Núñez and Sweetser, 2006; Le Guen, 2011; Núñez and Cornejo, 2012; Núñez, Cooperrider, Doan, and Wassmann, 2012; Cooperrider, Slotta, and Núñez, 2018).

Method

Fieldwork locations

The fieldwork took place in the southern part of the Ryukyu islands, where the languages we investigate in this study (Miyako and Shiraho) are spoken. The southern Ryukyus are approximately 1,900 kms southwest from Tokyo, located on the western end of nowadays Okinawa prefecture, Japan. Two sites were involved. The Miyako site was primarily situated in the main island of the Miyako archipelago (total population 54,863 in 2005, Miyakojima shishi hensan iinkai, 2012), which are composed of 7 inhabited islands: Ikema, Irabu, Miyako, Ogami, Kurima, grouped into the city of Miyako, and Tarama, Minna, grouped into the village of Tarama. The Shiraho site was located in the village of Shiraho (population 1,602 in 2010, Ishigakishi kikaku-bu kikaku seisaku-ka, 2013), along the southwestern shores of the Ishigaki island, one of the many islands belonging to the Yaeyama archipelago. This village is the only place where the Shiraho language is spoken. Both Miyako and Yaeyama islands got integrated into the Ryukyu kingdom at the beginning of the 16th century and were ruled and administered by the Ryukyu Kingdom until the Meiji era (1868-1912), before becoming part of Okinawa prefecture (Miyagi, 1968). Both fieldwork sites are located in rural areas that have agriculture (e.g. sugar cane), and, more recently, tourism as main economic activities.

Languages

Ryukyuan languages are traditionally spoken in the Ryukyuan islands and stand in sister relationship with Japanese, with which they form the Japonic family. They all share a common ancestor —proto-Japonic— thought to have been spoken before the 7th century (Pellard, 2015). They are divided into two branches, Northern Ryukyuan spoken in Amami and Okinawa islands, and Southern Ryukyuan, spoken in Miyako and Yaeyama islands. While not mutually intelligible as a consequence of the independent development of each language, both Miyako and Shiraho belong to the Southern Ryukyuan branch. The Ryukyuan languages are virtually entirely unintelligible to speakers of standard Japanese from the mainland, a fact that has been attested empirically (Yamada et al., in press). Ryukyuan speakers, on the other hand, are completely fluent in Japanese due to the fact their entire scholastic, administrative and civic lives are conducted in Japanese. As in the rest of the Ryukyus, there is an on-going language shift to Japanese, resulting in the elderly generations being bilingual in the traditional language and Japanese, and the younger generations monolingual in Japanese, with almost no knowledge of the local language. Miyako and Shiraho are thus reported to be definitely, and severely endangered, respectively (Moseley, 2010). The exact number of fluent speakers is difficult to assess due to the complex sociolinguistic situation induced by the language shift, but coarse estimates have given the figure of 12,000 to 22,000 speakers for Miyako (Jarosz 2015), and only 147 for Shiraho (Nakagawa, Lau, and Takubo, 2016). Miyako, Shiraho and modern Standard Japanese share many broad morpho-syntactic features. Among others, they are characterized by an agglutinative morphology, a SOV word order, the use of postpositions and suffixes, and a nominative/accusative case system.

Participants

Thirty-eight individuals participated in the study, 15 Miyako inhabitants-speakers (5 men, 10 women) tested in the Miyako islands, 8 Shiraho inhabitants-speakers (3 men, 5 women) tested in Shiraho, Ishigaki island, and 15 monolingual speakers of Standard Japanese (8 men, 7

women) all born and raised in Tokyo as the control group, tested in Tokyo. Given that Miyako and Shiraho are endangered languages (like all other Ryukyuan languages), the population of speakers is essentially constituted by senior citizens. As a result, the Ryukyuan samples in this study, as well as the matching control group, were formed by elderly men and women (mean age of the Miyako group was 83.2 and SD 8.13; for Shiraho was 83.63 years and SD 2.62; for the Japanese group was 73.47 years, SD 4.49). Miyako and Tokyo participants were recruited through elderly and rehabilitation centers. Shiraho speakers, being considerably less in number and living in a relatively small village, were recruited through the network of one of the authors (NN). As compensation Ryukyu participants received a small gift, and Tokyo controls a small amount of money. Three Miyako speakers and four Japanese speakers were excluded from the analysis because they failed to produce directional gestures with a clear trajectory in two or more of the four trials. One female Miyako participant was excluded because during the experiment she spoke in Japanese (not in Miyako, as instructed). In total, the data from 30 participants (11, 8, and 11, from the Miyako, Shiraho, and Japanese groups, respectively) were included in the analysis.

Materials

The set of stimuli consisted of eight simple live dynamic events made up of a combination of tabletop objects. These objects were: two ping-pong balls (one white, one black), two book-size boards (one white, one black), and one transparent small box used as a support for reclining the boards. In four of the eight events the experimenter releases a black/white ball to roll down a contrasting reclined white/black board and on the other four events s/he releases the ball (with corresponding color schemes) to bounce down the reclined board (see Figure 1).

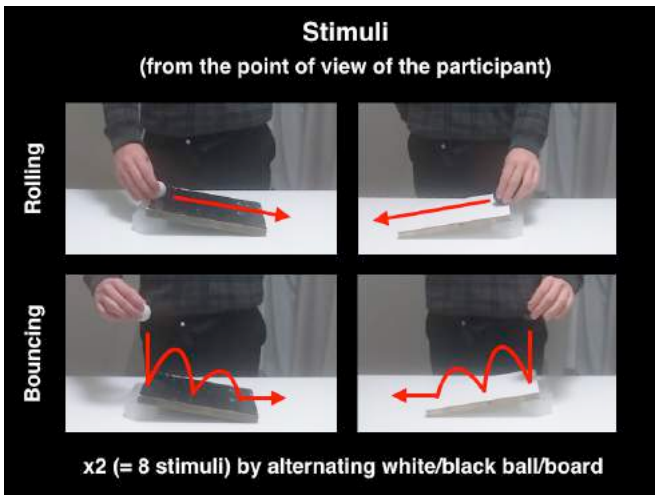


Figure 1: Gesture elicitation stimuli.

From these, four events had the action unfolding towards the left of the participant and four towards the right of him/her. Figure 1 shows four of the eight stimuli (the other four stimuli correspond to the ones shown but with alternated black/white colors of the ball/board).

Procedure

The experimental procedure consisted of four trials, in which participants were tested individually. In each trial the participant was placed in front of a table facing either north or south (the direction was selected randomly).

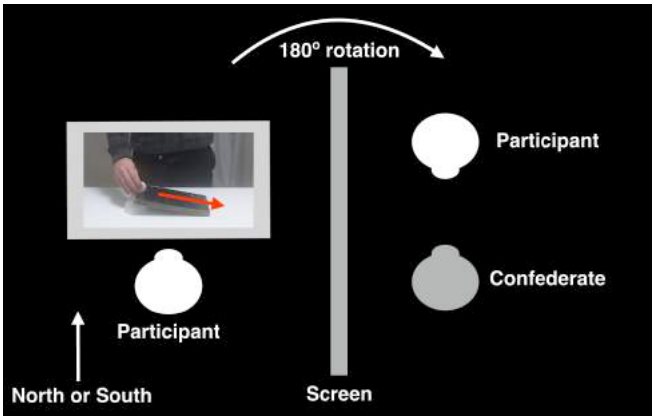


Figure 2: Schema of the gesture elicitation procedure.

Across the table there was an experimenter, who enacted for the participant one of the eight dynamic stimuli (events) described in the "Materials" section. The mode of movement (rolling/bouncing ball) and direction of the movement (eastward/westward) implemented in each of the four trials were selected randomly so that each participant experienced in the 4 trials all possible combinations of mode and direction of movement, namely ball rolling east, ball rolling west, ball bouncing east and ball bouncing west (the color scheme for each trial in the sequence was fixed, alternating the black ball over the white board with the white ball over the black board, beginning with the former configuration[1]). When the brief presentation of the event ended (with the experimenter catching the ball), the participant was invited to go to a section of the room located a few meters away on the other side of a screen that blocked the view of the table where the stimulus had been shown (see Figure 2), and was instructed to sit on an armrest-less chair that placed him/her with a 180° rotation with respect to his/her original orientation (i.e., if s/he observed the dynamic stimulus facing north, s/he was then sat facing south). At this point, sitting in front of the participant there

[1] Since the Ryukyuan languages are endangered, the number of (elderly) speakers available is reduced (especially in Shiraho). This made a full counterbalanced design (manner of movement, direction of movement, and color scheme) not viable.

was a confederate fluent in the participant's language[2] and who, being on this side of the screen, had not observed the dynamic stimulus. The participant was then invited to describe (in the relevant language: Miyako, Shiraho, or Standard Japanese) to the confederate sitting in front of him/her the scene that had been shown to him/her on the table on the other side of the screen. The presence of the confederate was meant to ensure that the participant would have a genuine real-world interlocutor who de facto had not observed the stimulus. The confederate's participation was primarily limited to listening to the descriptions given by the participant (and, occasionally, to nodding to signal comprehension of the participant utterances). Once the description was completed, the participant was invited to go back to the original table on the other side of the screen in order to begin the next trial. There was no mention of the possibility or necessity of gesturing during the instructions. All participants' descriptions were video-recorded.

Results

The directionality of the gestures was easily assessed as they primarily exhibited transversal hand (and sometimes head) movements indexing ball trajectories (often co-produced with utterances that referred to the motion of the ball itself: e.g. "it rolled away", or with vocalizations that characterized the manner of the ball's motion: e.g., "bom, bom, bom" for the bouncing ball). Only gestures that unambiguously unfolded either to the left or to the right of the participant while describing the trajectory of the ball were considered for analysis[3].

Figure 3: Example of a gesture produced by a Miyako speaker exhibiting an absolute frame of reference

Figure 3 shows an example of a Miyako participant describing the trajectory of a westward bouncing ball, which she had observed facing south on the other side of the white screen (i.e., as the ball moved to her right). Her description, as she now faces north, involves a right-handed swiping gesture towards her left (i.e., westwards) reflecting an absolute FoR.

[2] A native speaker for the cases of Miyako and the Standard Japanese controls. For Shiraho, the confederate was the third author (NN), who has done extensive fieldwork on the language.

[3] In this report we ignore gestures co-produced with non-targeted descriptions such as the size of the table or the texture of its surface.

Based on these gestural properties, the gestures depicting the trajectory of the dynamic stimuli of each of the four trials was classified as either "relative" or "absolute", depending on whether it followed an egocentric or an allocentric pattern, respectively[4]. The classification yielded a mean percentage of absolute gestures per participant.

In average both Ryukyuan groups produced a much higher percentage of absolute gestures (Miyako mean = 71.09%, SD = 23.7%; Shiraho mean 81.25%, SD = 34.72%) than the Japanese-only group (Mean = 40.9%, SD = 30.15%). Individual data show that only 1 of the 11 Miyako participants (9.1%), and 1 of the 8 Shiraho participants (12.5%) manifested absolute FoRs in less than 50% of the trials. In contrast, nearly half of the Japanese-only speakers did so (5 of the 11 participants; 45.5%). Besides, more than half of the Miyako participants (6 out of 11; 54.5%), and almost all Shiraho participants (7 out of 8; 87.5%) exhibited absolute FoRs at least in 75% of the trials. In contrast, only 2 of the 11 Japanese speakers (18.2%) did so. A One-Way ANOVA reveals that the difference between the three groups is statistically significant ($F(2,27) = 5.10$, $p = 0.013$), exhibiting a large effect size ($\eta^2 = 0.27$) (see Figure 4). Post-hoc Bonferroni and Holm simultaneous comparisons of the Ryukyu groups versus the Japanese controls yield statistically significant differences for both, Miyako (Bonferroni and Holm T-statistic = 2.42; Bonferroni p-value = 0.045; Holms p-value = 0.023) and Shiraho (Bonferroni and Holm T-statistic = 2.97; Bonferroni p-value = 0.013; Holms p-value = 0.013).

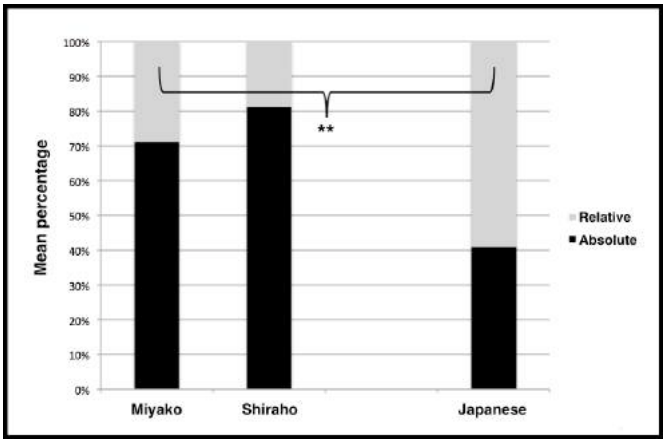


Figure 4: Mean percentage of absolute gestures (black) per participant.

Discussion

The study of FoRs has been an important area for the investigation of the linguistic relativity hypothesis (Gumperz & Levinson, 1996; Li & Gleitman, 2002; Majid et al., 2004; Haun et al. 2011). However, with rare

[4] For the analysis presented here we do not analyze the details and extension of participants' verbal production.

exceptions (e.g., Marghetis, McComsey, and Cooperrider, 2014; Meakins, Jones, and Algy, 2016), many of these studies have focused on monolingual populations. When bilinguism has been present, which is the case of many colonized regions around the world where individuals speak the indigenous language as well as that of the colonizer, more often than not the emphasis has been put only on one of the languages spoken (usually the indigenous one). This situation has left open questions involving the status of the linguistic relativity hypothesis when bilingual (or multilingual) cognition is concerned— an underestimated but ubiquitous condition in human history (Evans, 2011; Pavlenko, 2014). The linguistic relativity hypothesis makes clear predictions of FoRs preferences when the languages under investigation are not structurally equivalent, like when comparing those that do not have terms for "left" and "right" with languages that do, such as Mesoamerican Tzletal and Dutch, respectively (e.g., Brown and Levinson, 1993; Levinson, 2003; Majid et al., 2004). Usually these predictions are about comparisons of speakers of languages that, differing substantially in the deployment of relative vs. absolute FoRs, belong to completely unrelated and radically different linguistic families, such as the one just mentioned (Tzeltal belongs to the Mayan family and Dutch to the Indo-European family) or ≠Akhoe Hai||om form the savannah of Northern Namibia (Central Khoisan language family) and Dutch (Indo-European family) (Haun et al, 2011).

Some studies with bilingual speakers of indigenous/Indo-European languages that exhibit clashing absolute-relative encoding have investigated extra-linguistic factors that appear to delimit the role of language in shaping thought (e.g., Spanish (Indo-European)-Juchitán Zapotec (Otomanguean family) from Oaxaca, Mexico (Marghetis, McComsey, and Cooperrider, 2014); English (Indo-European)-Gurindji (Pama-Nyungan family) from Northern Australia (Meakins, Jones, and Algy, 2016). These studies suggest that environmental and sociocultural factors such as schooling and writing practices affect the choice of FoRs in fundamental ways that are not strictly speaking linguistic.

To help evaluating the precise role that language might play in shaping thought —the essence of the linguistic relativity hypothesis— and further delineate the contribution of extra-linguistic factors, an important step is to investigate cases of bilingual populations of people that speak languages that: (1) in practice tend to elicit clashing FoRs (e.g., absolute vs. relative), (2) that belong to the *same* linguistic family (hopefully sharing cognate words), and, importantly, (3) that are structurally equivalent with respect to the relevant lexicon and grammatical resources. In this study we have done so by investigating the preference of FoRs of Japanese-bilingual speakers of two endangered Ryukyuan languages that belong to the same Japonic linguistic family as Japanese, and that, while being mutually unintelligible, share an important number of grammatical features and cognate words (e.g. Hattori (1959) reports 59% of shared cognates for Miyako and Tokyo Japanese). Importantly, while Japanese has been reported to preferentially elicit the use of relative FoRs for characterizing small-scale spatial relations (Pederson et al., 1998; Kita, 2006) the Ryukyuan languages have been reported to do so primarily via absolute FoRs (Suzuki, 1978; Celik, Takubo, and Núñez, 2019).

The use of FoRs by (elderly) Miyako speakers had been previously studied experimentally via a referential communication task (Celik, Takubo, and Núñez, 2019) adapted from the standard "Man-and-tree" task (Pederson et al. 1998). In this task —of a director-matcher type (e.g., Le Guen, 2011)— speakers are asked to describe to a partner sitting on the other side of a screen simple scenes containing two toy animals that the latter could not see (Cooperrider, Slotta, and Núñez, 2017). The study on Miyako confirmed, as predicted, that (elderly) monolingual Japanese-speaker controls almost exclusively relied on relative terms to describe the stimuli to their partners. In stark contrast, Miyako speakers, when speaking in Miyako, exhibited a marked tendency to describe the spatial configurations of the figurines using absolute terms, such as 'west' and 'east', and this despite being fully bilingual in Japanese and having been massively exposed to mainland Japanese culture for decades. Surprisingly, in addition, speakers from the same Miyako bilingual population when doing the same task in Japanese relied extensively on relative terms, showing no significant difference from Standard Japanese speakers. These results suggest that, beyond the grammatical and lexical resources of languages, the referential communicative practices and conventions brought forth when a speaker is immersed in speaking a particular language may be an important factor influencing thought (and spatial construals, in this case). In order to test whether these results were triggered by the goal-oriented, performance-driven, explicit cooperation dimension that is demanded in this communicative task, in this study we investigated the spatial construals that would be spontaneously enacted gesturally in real-time without the demands of goal-oriented cooperation and performance that are present in the referential communication director-matcher task. Moreover, to gain a richer insight into the Ryukyuan mind, we extended this approach to investigate another Ryukyuan language— Shiraho, from the Ishikagi island.

This study shows that the investigation of spatial construals and FoRs preference can be enriched with the addition of observations of speech-gesture co-production. Spontaneous gesture production is largely effortless, which is an important factor when evaluating elderly speakers, and it is universal, so it is expected to be observed in all human groups (McNeill, 1992). Since, in general, spontaneous gesture production is less monitored than speech, it largely unfolds below the level of awareness, and therefore provides a remarkable backdoor to real-time cognition (McNeill, 1992). And being effortless and largely unconscious, gesture production does not provide the intimidating effect that challenging reasoning and memory tasks might bring for some individuals. This is especially relevant when

studying elderly speakers of endangered languages who often are unschooled (or poorly schooled) and are less exposed to testing and scholastic practices. Importantly, gesture production often provides content that is not observable through speech alone (Kendon, 2004) allowing for the observation of construals that are not accessible via purely linguistic means (Le Guen, 2011; Núñez et al., 2012). Finally, gestures are often co-produced with abstract analogical and metaphorical thinking that reveal important features of the underlying conceptual mappings and inferential affordances (McNeill, 1992), which often rely on spatial construals (e.g., see Núñez and Sweetser (2006) for spatial construals of time).

The novel gesture elicitation paradigm used in this study builds on these important properties of gesture-speech co-production. The results confirm, via a different, but complementary approach into cognition, the preference of absolute FoRs observed in Miyako speakers with a more traditional method such as that of referential communication director-matcher tasks (Celik, Takubo, and Núñez, 2019). These results help establish the reliability of this novel gesture elicitation method for investigating tabletop spatial construals. Morever, the results of this study extend the findings to speakers of another Ryukyuan language — Shiraho— which being consistent with those obtained with Miyako speakers, confirm some of the socio-geographic observations (Suzuki, 1978) that Ryukyuan people exhibit a marked tendency to use absolute FoRs.

The fact that the elderly Ryukyuan individuals exhibit preference for absolute FoRs despite being fluent Japanese-bilinguals and being fully immersed in mainland Japanese culture for most of their extended lives is quite remarkable. In fact, other groups, such as the Gurindji people of Northern Australia who traditionally relied on absolute FoRs, have been shown to have shifted to relative patterns due to exposure to English and the associated cultural and literary practices that go with it (Meakins, Jones, and Algy, 2016). Similary. regarding spatial construals of time, Mandarin-English bilinguals have been reported to exhibit chronic inter-language influences on patterns in thought, and this already by the time they are young adults (Lai and Boroditsky, 2013). But in stark contrast to these cases, the last speakers of the Ryukyuan languages we have studied, despite a life-long immersion in the dominant Japanese culture and language, manifest a preference for an absolute FoR that appears to be robust and long-lasting. How pervasive is this pattern among speakers of other languages of the Ryukyu islands? Why is it so resilient? We don't know. More research is needed for answering this and many other open questions. We are, however, running out of time, as the last speakers of these endangered languages may leave us taking with them their rich cultural heritage along with the precious answers.

Acknowledgments

We thank Yukinori Takubo for comments and insights regarding the Japonic languages and for facilitating access to the fieldwork locations. We also thank Hamakawa Keiko for helping with data collection in Miyako. This research was supported by grants from the Japan Society for the Promotion of Science (17H02333, 25284078, 18K12360).

References

Bates, E., & Dick, F. (2002). Language, gesture, and the developing brain. *Developmental Psychobiology*, *40*, 293-310.

Brinley, J., Jovick, T., & McLaughlin, L. (1974). Age, reasoning, and memory in adults. *Journal of Gerontology*, *29*, 182-189.

Brown, P., & Levinson, S.C. (1993). *Linguistic and nonlinguistic coding of spatial arrays: Explorations in Mayan cognition* (Working Paper No. 24). Nijmegen, The Netherlands: Cognitive Anthropology Research Group, Max Plank Institute.

Celik, K., Takubo, Y., & Núñez, R. (2019). Spatial frames of reference in Miyako: Digging into Whorfian linguistic relativity. In S. Fukuda, M.S. Kim, M-J Park, & H.M. Cook (Eds.), *Japanese/Korean Linguistics Vol. 25*. Chicago: Univ. of Chicago Press (CSLI publications).

Cooperrider, K., Slotta, J., & Núñez, R. (2017). Uphill and Downhill in a Flat World: The Conceptual Topography of the Yupno House. *Cognitive Science*, *41*, 768–799.

Cooperrider, K., Slotta, J., & Núñez, R. (2018). The preference for pointing with the hand is not universal. *Cognitive Science*, *42*, 1375-1390.

Evans, N. (2011). *Dying words: Endangered languages and what they have to tell us*. Chichester UK: John Wiley

Goldin-Meadow, S. (2005). *Hearing gesture: How our hands help us think*. Harvard University Press.

Gumperz, J. J. & Levinson, S. (1991). Rethinking Linguistic Relativity. *Current Anthropology*, *32*, 613-623.

Gumperz, J. J. & Levinson, S. (Eds.) (1996). *Rethinking linguistic relativity. Studies in the social and cultural foundations of language, No. 17*, Cambridge UK: Cambridge University Press.

Hattori, S. (1959). *Nihongo no keito* [Origins of the Japanese language]. Tokyo: Iwanami shoten.

Haun, D., Rapold, C., Janzen, G., Levinson, S. (2011). Plasticity of human spatial cognition: Spatial language and cognition covary across cultures. *Cognition*, *119*, 70–80.

Ishigakishi kikaku-bu kikaku seisaku-ka (Ed.) (2013). *Tokei Ishigaki* [Ishigaki statistics]. Okinawa: Ishigaki city.

Jarosz, A. (2015). *Nikolay Nevskiy's Miyakoan dictionary: reconstruction from the manuscript and its ethnolinguistic analysis*. Doctoral dissertation, Faculty of Modern Languages and Literature, Adam Mickiewicz University.

Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge UK: Cambridge University Press.

Kita, S. (2006). A grammar of space in Japanese. In S.C. Levinson, & D. Wilkins (Eds.), *Grammars of space. Explorations in Cognitive Diversity. Language, Culture and Cognition*. Cambridge: Cambridge University Press

Lai, V. T. & Boroditsky, L. (2013). The immediate and chronic influence of spatio-temporal metaphors on the mental representations of time in English, Mandarin, and Mandarin-English speakers. *Frontiers in Psychology*, *4*, 142.

Le Guen, O. (2011). Speech and gesture in spatial language and cognition among the Yucatec Mayas. *Cognitive Science*, *35*, 905-938.

Levinson, S. (2003). *Space in language and cognition: Explorations in cognitive diversity*. Cambridge, UK: Cambridge University Press.

Levinson, S. C., & Wilkins, D. P. (Eds.). (2006). *Grammars of space: Explorations in cognitive diversity*. Cambridge: Cambridge University Press.

Li, P., & Gleitman, L.R. (2002). Turning the tables: spatial language and spatial reasoning. *Cognition*, *83*, 265–294.

Lucy, J. A. (1992). *Language diversity and thought: A reformulation of the linguistic relativity hypothesis (Vol. 12)*. Cambridge: Cambridge University Press.

Majid, A., Bowerman, M., Sotaro Kita, S., Haun, D., & Levinson, S. (2004). Can language restructure cognition? The case for space. *Trends in Cognitive Sciences*, *8*, 108-114.

Marghetis, T., McComsey, M., & Cooperrider, K. (2014). Spatial reasoning in bilingual Mexico: Delimiting the influence of language. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 36, 940-945.

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. Chicago: University of Chicago press.

Meakins, F., Jones, C., & Algy, C. (2016). Bilingualism, language shift and the corresponding expansion of spatial cognitive systems. *Language Sciences*, (54), 1-13.

Miyagi, E. (1968). Okinawa no rekishi [the History of Okinawa]. Tokyo: Japan Broadcast Publishing.

Miyakojima shishi hensan iinkai (2012). *Miyakojima shishi dai ikkan tsushi-hen* [the History of Miyako city, 1st volume: historical overview]. Miyako: Miyakojima-shi kyoiku iinkai.

Moseley, C. (2010). *Atlas of the World's languages in danger, 3rd edition*. Paris: UNESCO publishing.

Nakagawa, N., Lau, T., & Takubo,Y. (2016). Yaeyama-go shiraho-hogen-no bumpo gaisetsu [A grammar sketch of Shiraho dialect, Yaeyama Ryukyuan]. In Karimata, S. (Ed.), *Ryukyushogo kijutsu bumpo* [Descriptive grammars of Ryukyuan languages]. Okinawa: University of Ryukyus.

Núñez, R & Sweetser, E. (2006) With the future behind them: convergent evidence from Aymara language and gesture in the crosslinguistic comparison of spatial construals of time. *Cognitive Science*, *30*, 401–450.

Núñez, R., & Cornejo, C. (2012). Facing the sunrise: Cultural worldview underlying intrinsic-based encoding of absolute frames of reference in Aymara. *Cognitive Science*, *36*, 965–991.

Núñez, R., Cooperrider, K., Doan, D., & Wassmann, J. (2012). Contours of time: Topographic construals of past, present, and future in the Yupno valley of Papua New Guinea. *Cognition*, *124*, 25-35.

Pavlenko, A. (2014). *The bilingual mind: and what it tells us about language and thought*. Cambridge: Cambridge University Press.

Pederson, E., Danziger, E., Wilkins, D., Levinson, S., Kita, S., & Senft, G. (1998). Semantic typology and spatial conceptualization. *Language*, *74*, 557-589.

Pellard, T. (2015). The linguistic archeology of the Ryukyu islands. In: P. Heinrich, S. Miyara, & M. Shimoji (Eds.), *Handbook of the Ryukyuan languages: history, structure and use*. Berlin: De Gruyter Mouton.

Suzuki, M. (1978). The study of orientation in south-west (Ryukyu) islands. *Japanese Journal of Human Geography*, *30*, 541-554.

Yamada, M., Takubo, Y., Iwasaki, S., Celik, K., Harada, S., Kibe, N., Lau, T., Kakagawa, N., Niinaga, Y., Otsuki, T., Sat, M. Shirata, R., Van der Lubbe, G., & Yokoyama, A. (in press). Experimental study of inter-language and inter-generational intelligibility: Methodology and case studies of Ryukyuan Languages. *Japanese/Korean Linguistics Vol. 26*. Chicago: Univ. of Chicago Press (CSLI publications).

Designing good deception: Recursive theory of mind in lying and lie detection

Lauren A. Oey (loey@ucsd.edu)
Adena Schachner (schachner@ucsd.edu)
Edward Vul (evul@ucsd.edu)
University of California, San Diego, Department of Psychology
9500 Gilman Dr., La Jolla, CA 92093 USA

Abstract

The human ability to deceive others and detect deception has long been tied to theory of mind. We make a stronger argument: in order to be adept liars – to balance gain (i.e. maximizing their own reward) and plausibility (i.e. maintaining a realistic lie) – humans calibrate their lies under the assumption that their partner is a rational, utility-maximizing agent. We develop an adversarial recursive Bayesian model that aims to formalize the behaviors of liars and lie detectors. We compare this model to (1) a model that does not perform theory of mind computations and (2) a model that has perfect knowledge of the opponent's behavior. To test these models, we introduce a novel dyadic, stochastic game, allowing for quantitative measures of lies and lie detection. In a second experiment, we vary the ground truth probability. We find that our rational models qualitatively predict human lying and lie detecting behavior better than the non-rational model. Our findings suggest that humans control for the extremeness of their lies in a manner reflective of rational social inference. These findings provide a new paradigm and formal framework for nuanced quantitative analysis of the role of rationality and theory of mind in lying and lie detecting behavior.

Keywords: deception; Theory of Mind; Bayesian reasoning; non-cooperative games; computational modeling

Introduction

The frank truth is that humans lie frequently, and the abilities to lie and detect lies are practical, but cognitively demanding, tools we develop over time (Vrij, Fisher, Mann, & Leal, 2006). Although much of the research on lying focuses on physical cues that give away lying (like facial expressions), both liars and lie detectors must consider not only the execution of lies (e.g. Vrij, Granhag, & Porter, 2010; Ekman, Friesen, & O'Sullivan, 1988) but also the informational content of lies. In our current era of endemic fake news (Allcot & Gentzkow, 2017), it is ever more critical that we develop an understanding of what cognitive processes contribute to deception and its detection.

Lying *at all* requires believing that the recipient could have a belief different from your own, and thus lying has long been tied to theory of mind (ToM), or the understanding of others' mental states, such as beliefs. Children struggle with the ability to represent false beliefs and second-order beliefs conditioned on false beliefs (Wimmer & Perner, 1983; Talwar, Gordon, & Lee, 2007). This poor ToM in children should also make them terrible liars. Indeed, improvement in children's detection and production of lies appears to be directly related to the development of their ability to use ToM (Ding, Wellman, Wang, Fu, & Lee, 2015). To lie at all, we need to be able to entertain the possibility of a false belief in our interlocutor, however, successful deception requires a far more nuanced process of decision-making interacting with ToM inference.

We usually lie to benefit ourselves. For example, a male date-seeker may want to optimize his chances of attracting potential romantic interests by inflating his height on his online dating profile (Toma, Hancock, & Ellison, 2008). What height should he make up and report to accomplish this goal? A taller height might be more attractive in the eye of potential dates; so perhaps, he could choose the height of his favorite professional basketball player. However, being caught in a lie tends to be costly: he may jeopardize his trustworthiness. An overly tall height is likely to make his date more suspicious, so to decrease the chance of getting caught in a lie, he should not make the height too suspicious. How should he balance these competing pressures on his lie?

On the receiving end of a lie, it is advantageous for humans to be attuned to the detection of lies. Potential dates should want to detect the date-seeker's lie in order to discern whether he is a trustworthy human. But dates cannot haphazardly accuse others of lying, as a false accusation can also result in tarnishing the accuser's reputation. Both liars and lie detectors not only must navigate the constraints placed upon themselves, but they should also consider the other agent's perspective.

In the current study, we argue that good deception not only requires the use of ToM, but we make a stronger claim that good lie detectors evaluate, and good liars conjure, their lies under the assumption that their partner is a rational utility-maximizing agent. We formalize the role of rational and recursive social inference in the production and detection of deception. We argue that it is not only the ability to represent partners' false beliefs that distinguishes good liars from bad liars; rather, good liars balance maximizing reward with maintaining plausibility in their lies, such that liars can avoid having their lies detected by another agent. In order to maximize achievement of these goals, liars consider their partner's prior expectations, the likelihood of observations, and how these expectations shift in response to considering the other agent.

Traditionally psychological studies examining the role of ToM in deception are one-shot experiments. Examples of

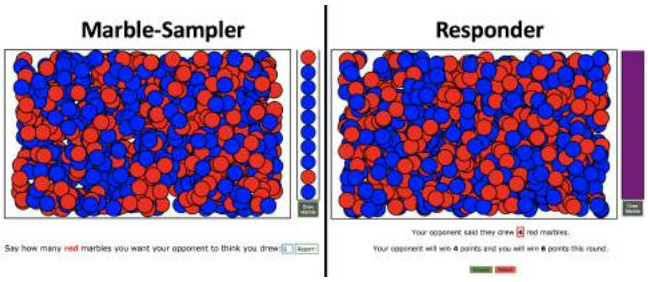


Figure 1: Lying game. In both the marble-sampler and responder roles, participants see the distribution of marbles. (Left) Marble-samplers sample 10 marbles, then either lie or tell the truth about the number of red marbles sampled. (Right) Responders accept or call BS.

such studies are those in which children are instructed to not peek at a toy while the experimenter temporarily leaves the room, and children choose to either lie or tell the truth about peeking (Lewis, Stanger, & Sullivan, 1989; Talwar & Lee, 2002). Alternatively studies of dishonesty in the behavioral economics literature use quantitative measures but emphasize the tendency of people to cheat at an individual level, independent of how other agents affect their deception (e.g. Mazar, Amir, & Ariely, 2008). Taking inspiration from both designs, we developed a novel repeated dyadic stochastic game allowing us to focus on the quantitative, socially-motivated production and evaluation of lies.

Using Bayesian game-theoretic computational modeling and experimental methods, we argue that a well-calibrated, Bayesian ToM supports the production of believable lies and the detection of poorly-formed lies, and introduce a novel ideal observer model of deception.

Lying Game

To study how humans actually behave in lying situations, we developed a novel lying game that rewarded participants for strategic detection and production of lies (Figure 1).

In each round of the game, both players are presented with a box containing red and blue marbles, with some proportion p of red marbles. Players alternate between playing as the marble-sampler or the responder. The marble-sampler randomly samples 10 marbles from the box, of which k are red. However, the sampled marbles are occluded from the responder, so the responder cannot see the true distribution of sampled marbles. The marble-sampler chooses a number k^* to report as the number of red marbles they want their opponent to *think* they sampled. The marble-sampler could choose to (a) tell the truth and report the true number of red marbles sampled, or (b) lie and report a false number of red marbles sampled. The responder then has the opportunity to either (A) accept the reported value or (B) reject it as a lie (i.e. call BS).

Both the marble-sampler's decision to (a) tell the truth or (b) lie about the number of red marbles sampled, and the responder's decision to (A) accept or (B) reject the reported value impact each player's payoff (Table 1). If the reported number of red marbles sampled k^* is accepted, the marble-sampler receives k^* points and the responder received $10-k^*$ points. If the responder rejects the reported number then the payoffs depend on whether or not it was a lie: if the reported number is the truth ($k = k^*$), the marble-sampler gets the k^* points, and the responder pays a penalty of $-k^*$; if the reported number is a lie ($k \neq k^*$) then the responder gains k^* points, while the marble-sampler pays a penalty of $-k^*$. Altogether, this game sets up a reward function that motivates marble-samplers to lie, but not be caught, and motivates the responder to call out egregious lies, but avoid false accusations.[1]

		Marble-Sampler	
		$k = k^*$	$k \neq k^*$
Responder	$BS = 0$	$2k^* - 10$ / $10 - 2k^*$	$2k^* - 10$ / $10 - 2k^*$
	$BS = 1$	$2k^*$ / $-2k^*$	$-2k^*$ / $2k^*$

Table 1: Players' payoff differential (player - opponent points). Utility is determined by reported k^* and whether BS was called. Values to the right of the diagonal in cells indicate points awarded to the marble-sampler, while values to the left are awarded to the responder.

Models

No-Theory-of-Mind Model

As a baseline, let's consider a model that has no model of the opponent, or believes that the opponent is effectively random. In deciding upon what number to report (k^*), such a model does not consider the behavior of the opponent, and would simply lie with probability $1-p$. Moreover, when it lies it would either sample uniformly from values larger than the truth (k), or it would simply pick the largest value (10 – as this is expected-value maximizing response under the assumption that the opponent calls BS at random). This is the best that an agent that has no model of their opponent could do. This no-theory-of-mind model makes a qualitative prediction about lying behavior, such that the expected value of k^* increases linearly as a function of k.

Likewise, a lie-detector that has no model of their opponent, and thus believes them to be random, would only consider the probability of k^* under the true world distribution of $P(k)$. Since this model does not consider the motives and payoffs for their opponent, it would amount to playing the game without knowing the opponent's payoff structure, e.g. whether they would receive points for red or blue marbles. If the marble-sampler were to say that they sampled one red marble when $p = 0.5$, the responder may call BS, simply because such a value is unlikely to occur by chance. This lie detector amounts to conducting a two-tailed hypothesis test. It computes what is statistically significant under a binomial test and calls BS on all k^* that have a p-value $< \alpha$. Regardless

[1] Code available at `github.com/la-oey/Bullshitter`

of α, this lie-detector would call BS on all reports of unlikely k^*, and would thus have a U-shaped lie-detection profile.

Oracle Model

Alternatively, suppose we have a theory-of-mind model that has a *perfect* model of its opponent (i.e. it has an oracle-like omniscience over its opponent's probability to lie and detect lies). This model does not require recursive social inference as a simple first-order inference will suffice, given that they have already perfectly adapted their model of the opponent. It is critical to understand how such a model would behave, as this exemplifies an ideal agent.

To accomplish this, we developed an inferential model of deception, which we term the oracle model, whose opposing agent lies and detects lies using the algorithms from the AI that participants competed against in our lying game in both experiments.

When detecting lies, the AI computes $P_D(BS \mid k^*)$ using the cumulative binomial probability of k^*, $P(X \leq k^*) = \sum_{x=0}^{k^*} Binomial(x \mid p, 10)$ centered at 0.5 when $k^* = 5$. To compute $P_L(k^* \mid k)$, the AI randomly samples a potential k^*, $\hat{k^*}$, from a binomial distribution. If $\hat{k^*}$ is greater than the true k, it lies and uses $\hat{k^*}$ as its reported k^*. Otherwise, it tells the truth by using the true k as its reported k^*.

As participants in our behavioral experiment iteratively competed against this very same non-inferential algorithm over several trials, it seems viable that participants may become perfectly calibrated to the algorithm that their opponent operated upon. In that case, human behavior would rationally match the predictions of the oracle model performing inference over the AI.

Recursive Theory-of-Mind Model

Finally, we consider a model of an ideal observer who does not know *a priori* the behavior of their opponent, but can estimate it from first principles, on the assumption that their opponent is as rational as they are, and is also trying to anticipate their opponent's behavior. This amounts to paired, adversarial ideal observers in which liars L and lie detectors D act as competing rational utility-maximizers. Our model builds on previous Bayesian frameworks of social cognition and communication (Baker, Saxe, & Tenenbaum, 2009; Frank & Goodman, 2012). Both agents perform inference over one another, i.e. L determines what number to report based on his prediction of D's tendency to call BS for different reported numbers, and vice versa. Both agents assume the other agent is acting rationally, namely the other agent is performing optimally given their goal to maximize their own utility. Furthermore, this process of performing inference over the other agent's actions is recursive. In other words, L decides upon his action based on what he believes D will do in light of what she believes L will do, etc. As infinite recursion is memory delimited, our model implements a decay function that breaks the chain of recursion with some degree of probability and implements a base case.

L is constrained by two competing goals: (1) gain (i.e. the agent wants to gain the highest reward possible given that they successfully deceive their partner), and (2) believability (i.e. the probability of having a lie go undetected, which is constrained by the extremeness of their lie). Meanwhile, the competing goals of D include to (1) successfully detect lies, while (2) avoiding falsely accusing their partner of lying when they are in fact telling the truth. The adversarial nature of the agents' goals is captured in the inverse relationship of the utility values for both L and D in our lying game.

Given some true state of the world k, L asserts to D that the state of the world is k^*. If k^* is not equal to k, L is telling a lie, otherwise L is telling the truth. D then sees the reported k^*, and responds by choosing whether to challenge the veracity of k^* by calling $BS = 1$, or accepting k^* as stated ($BS = 0$).

Our formalization of deception is represented as a zero-sum game. We assume that the probability of an action follows a Luce choice rule based on the expected utility of the action relative to alternative actions, with softmax parameter α (Luce, 1959):

$$P(A) = \underset{A}{softmax}(\mathrm{EV}[A]) = \frac{\exp(\alpha \mathrm{EV}[A])}{\sum_{A'} \exp(\alpha \mathrm{EV}[A'])} \quad (1)$$

D chooses to call BS following a Luce choice rule weighting of the expected value of the two options: calling BS, or accepting k^*:

$$P_D(BS \mid k^*) = \underset{BS}{softmax}(\mathrm{EV}_D[BS \mid k^*]) \quad (2)$$

The expected value of calling BS is obtained by marginalizing over the possibilities that $k^* = k$ (here abbreviated as $T = 1$), and $k^* \neq k$ ($T = 0$):

$$\mathrm{EV}_D[BS \mid k^*] = \sum_T \mathrm{u}_D(BS; k^*, T) P(T \mid k^*) \quad (3)$$

where $\mathrm{u}_D(BS; k^*, T)$ is the payoff for D associated with a particular BS response, given k^* and whether or not it corresponds to the true k (T).

The probability of a given k^* being true is given by

$$P(T \mid k^*) = P(k^* = k \mid k^*) = \frac{\sum_k P(k) P_L(k^* \mid k) P(k = k^* \mid k, k^*)}{\sum_k P(k) P_L(k^* \mid k)} \quad (4)$$

relying on the prior probability of k (here: $P(k) = Binomial(k \mid p, 10)$), and the probability that L would produce a given k^* in response to seeing a particular k, $P_L(k^* \mid k)$. Thus, calculating the expected value of calling BS, and choosing whether or not to call the lie requires an estimate of how L is likely to behave.

L, in turn chooses k^* based on a softmax weighting of the expected value of different responses,

$$P_L(k^* \mid k) = \underset{k^*}{softmax}(\mathrm{EV}_L[k^* \mid k]) \quad (5)$$

with the expected values given by:

$$\mathrm{EV}_L[k^* \mid k] = \sum_{BS} \mathrm{u}_L(k^* \mid BS, k^* = k) P_D(BS \mid k^*) \quad (6)$$

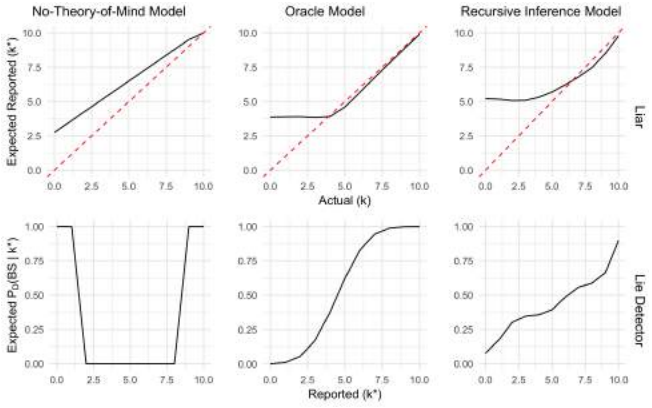


Figure 2: Computed predictions across all models (no-theory-of-mind, oracle, and recursive inference) when $p = 0.5$. Top row displays the liar's predicted performance: reported value as a function of the true value. The red dashed line indicates reported values that would be true. Bottom row displays the lie detector's predicted performance: conditional probability of calling BS by the reported value k^*.

where $u_L(k^* \mid BS, k^* = k)$ is the payoff for the L when reporting k^* given that BS was called, and whether that k^* was a lie. Calculating these expected values requires that L consider $P_D(BS \mid k^*)$ – the probability that D would call BS for a particular reported k^*.

Thus the expected values of various choices for L depends on his beliefs about D, and the expected values of calling BS for D, depends on her beliefs about L. This would yield infinite recursion, so in practice we assume that L's model of D has some probability λ of simply returning a constant $P_D(BS \mid k^*) = c$.

Model Predictions

In Figure 2, we computed predictions from each of the three models about the performance of liars' reported k^* given true k and lie detectors' $P(BS \mid k^*)$ given the reported k^*.

The oracle and recursive inference models make qualitatively similar predictions. For lying, above a certain k, reported values tend to fall on the identity line, indicating that beyond that value, lying is imprudent. For values of k below the average, it is better to lie with a false report of an average outcome (here, $\mathrm{E}[k^*] = 5$ for small values of k). This threshold value in the oracle model is lower than the one in the recursive inference model. This pattern of lying seems to reflect the liar's attempt to balance the gain of lies and the risk of detection from reporting improbable values. For lie detecting, both models predict a sigmoidal pattern, calling BS more often as k^* increases. The fact that both the oracle and recursive inference models appear similar along both the liar and lie detector behaviors indicates that the recursive inference model can emulate the same behavior as the oracle model, despite having no information about the specific behavioral policies of the opposing agent.

In contrast to the theory-of-mind (oracle and recursive inference) models, the no-theory-of-mind model only reduces lying on account of a ceiling effect; thus making k^* a linear function of k. As a lie detector, the no-theory-of-mind model does not consider the reward function of the liar, and thus predicts that both extremely high and extremely low reported values would be called out as lies.

We qualitatively tested these predictions from the theory-of-mind and non-theory-of-mind models in experiment 1. In experiment 2, we tested how manipulating the prior probability of sampling k by varying p would influence human lying and lie detecting behavior. Under the assumption that liars and lie detectors behave rationally, we would expect to see that their behavior would be robust to changes in the probability of the world.

Experiment 1

Participants

We recruited 193 UC San Diego undergraduate students to participate in an online study for course credit.

Procedure

There were a total of 40 trials, with the player acting as the marble-sampler in the initial trial, and then switching roles between each trial, resulting in 20 trials as the marble-sampler and 20 trials as the responder. Participants were instructed to "beat [their] opponent into the ground by winning by the highest point differential possible," in order to motivate participants to successfully lie and detect lies throughout the task. The distribution of marbles was uniform, such that there were 50% red and 50% blue marbles ($p = 0.5$)

Results

When in the marble sampler role, participants showed a non-linear pattern of drift from the truth with lower k values, as shown in the top of Figure 3. We find that this pattern of lying in a positive utility direction for the liar, i.e. above the red line, occurs at lower numbers up until the actual marbles sampled is equal to 5 (i.e. the expected mean).

When in the role of responder, participants' results showed a sigmoidal trend, as shown in the bottom of Figure 3. Both the liar and lie detector pattern of results provide evidence against the no-theory-of-mind model and instead support the oracle and recursive inference models.

It should be noted that due to the nature of binomial distributions, sampling a low number (or high number) of red marbles is rare. As the AI was set up in such a way that the computer tends to lie toward the mean value when the sampled number of marbles is low, this produced a low probability of the computer reporting a low number of red marbles sampled. As a result, there were only a small number of data points available to determine how people detect lies under those conditions. To help offset the wide variance resulting from low counts across k^*, we converted counts to proportion using $(n_{BS=1} + 1)/(n + 2)$ and for all figures, we included points in which there were greater than three observations for a given value along the x-axis.

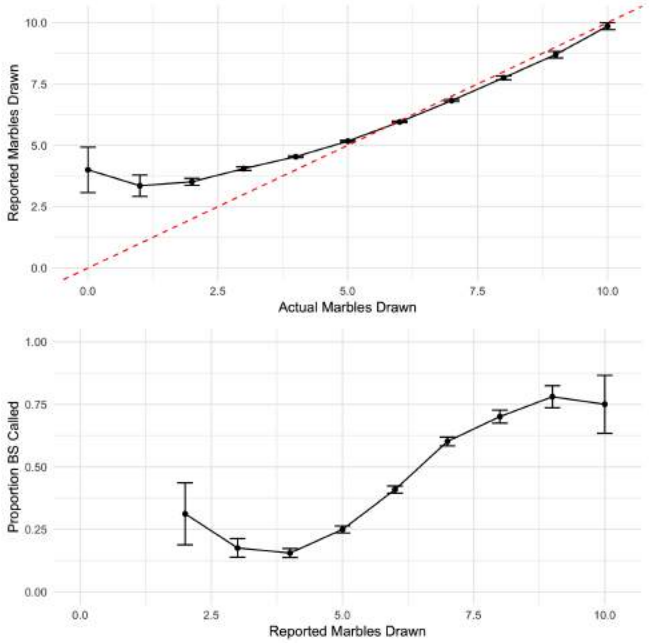


Figure 3: Results from experiment 1: (Top) Marble-sampler's reported number versus actual red marbles sampled. (Bottom) Responder's proportion of calling BS by the reported number of red marbles sampled.

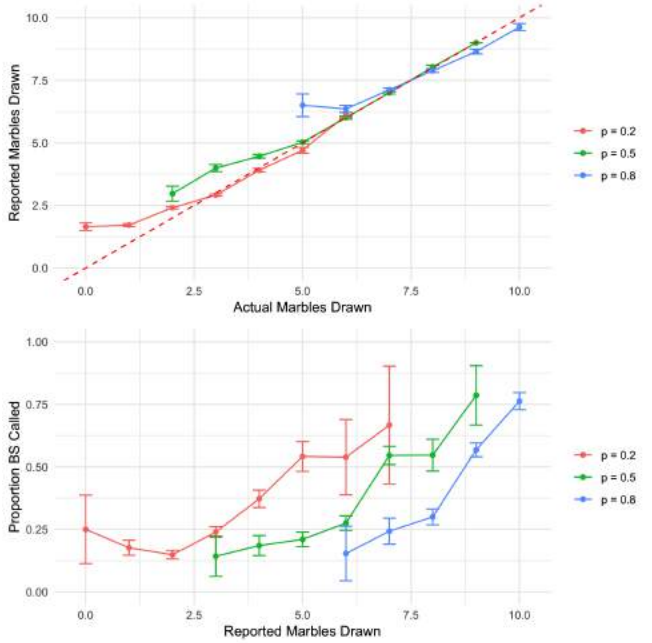


Figure 4: Results from experiment 2: (Top) Marble-sampler's reported versus actual number of red marbles sampled, by changes in the distribution of red-to-blue marbles (indicated by color). (Bottom) Responder's proportion of calling BS by the reported number of marbles sampled.

Experiment 2

We predicted that if people were making flexible, rational inferences, they would be able to flexibly take into account the distribution of marbles in the population, both in the lies they generated and in their detection of others' lies. In lying, we expect a shift in the point on k at which reported values drift from the truth. Similarly, in lie detecting, we expect a change in the tolerance for different k^* values, resulting in a horizontal shift of the BS calling function.

Experiment 2 used a similar design to experiment 1, except that crucially we manipulated the prior probability of k and removed feedback about the other agent's actions. By eliminating feedback, we hoped to distinguish between whether participants are simply adapting to the strategy of the other agent from this feedback, or if participants are performing inference on the other agent's decision process.

Participants

We recruited 86 UCSD undergraduates. Fifteen participants failed to meet the attention check criteria, which entailed accurately answering greater than 75% of the 12 comprehension questions disbursed throughout the experiment. This left 71 participants in our final pool.

Procedure

The procedure for experiment 2 was similar to experiment 1, except we varied between-subject the probability distribution from which the marbles were sampled. There were three conditions: the (red-to-blue) distribution of marbles was either 50-50 ($n = 20$), 20-80 ($n = 32$), or 80-20 ($n = 19$). Participants were not explicitly told about the distribution of marbles; rather they gathered the value of $P(k)$ via visual observation (i.e. the distribution of marbles in the box). In addition, the number of trials increased to 80 trials.

Participants also received no feedback about player decisions between each trial. Thus, the marble-sampler no longer received feedback about whether the responder called BS, and the responder no longer received about whether the marble-sampler lied or not. To ensure that participants understood the payoff structure, participants completed four pre-task practice trials with feedback, i.e. players' decisions, points earned a given trial, and cumulative score, after each trial. These practice trials were included to demonstrate to the participants that the other agent was generating lies or evaluating the participants' lies, and to establish the game's payoff matrix. After the practice trials, participants were told they would no longer receive feedback after each trial, instead only seeing the cumulative score every fifth trial.

Lastly, we used a new payoff structure, in order to contend with a puzzling characteristic of the payoff structure used in experiment 1. In particular in experiment 1, when $k^* \leq 2$, the responder received a lower relative payoff for successfully calling BS on a k^* lie than accepting k^*. Therefore, it would be in the responder's best interest to accept k^* even if they were to believe the marble-sampler was lying. In the $p = 0.2$ condition, the expected value of $k^* = 2$, suggesting that this condition may be affected by the unusual payoff at lower k^* values. To contend with this issue, experiment 2's new payoff structure resulted in a relative gain of 10 for the responder

(and a relative loss of 10 for the marble-sampler) whenever the responder successfully called BS on a lie. Meanwhile, falsely accusing the marble-sampler of lying resulted in a -5 penalty (deducted from the points they would have received had they accepted) for the responder.

Results

Overall we found that participants calibrate their lies, as well as their lie detection, based on the probability structure of the world. Firstly, when examining the lies given by the marble-sampler, we can see a drift in the reported k^* across all conditions, following the pattern of response seen in experiment 1. This point of drift shifts in accordance with the condition, such that the shift in condition $p = 0.2, 0.5, 0.8$ occurs around $k = 3, 5, 7$, respectively.

Secondly, in examining lie detecting behavior, we found that participants shift their judgments of which k^* values they called out as a lie, based on the probability distribution of marbles in the population. We used a mixed-effects logistic regression model to describe the probability that BS is called as a function of the reported number of red marbles sampled k^* and the marble distribution p dummy coded with $p = 0.5$ as the reference group. We found a significant main effect of both reported number of red marbles sampled ($\hat{\beta} = 0.723$, $z = 9.032$, $p < 0.0001$) and marble distribution in the $p = 0.2$ condition ($\hat{\beta} = 2.069$, $z = 3.081$, $p = 0.002$) and in the $p = 0.8$ condition ($\hat{\beta} = -5.823$, $z = -4,714$, $p < 0.0001$). These results not only suggest that people's detection of lies varies as a function of the reported value, but people also calibrate their BS calling depending on the probability of the world.

Using the estimated β values, we computed the number of marbles k^* at which lie detectors would call BS 50% of the time ($P_D(BS \mid k^*) = 0.5$). These thresholds varied systematically across the different marble distributions ($p = 0.2$: 5.236; $p = 0.5$: 7.327; $p = 0.8$: 8.762). The decision boundary shifts to higher k^* values as p increases. This result suggests that lie detectors change their BS calling behavior as a function of their prior expectations about the distribution of the world.

Discussion

In this paper, we report evidence that people lie, and detect lies, in ways that are well-captured by an adversarial recursive Bayesian model. We argue that good liars not only require an ability to represent the idea that others might have mental states different from their own, but they make inferences about the beliefs and actions of their interlocutor to successfully evade detection. In determining what utterances to call out as lies, good lie detectors must rationally consider the goals and utilities of their interlocutor and statistical information about the probability structure of the world.

We introduced the oracle model, in which the model has perfect information about how its opponent behaves. We compared the oracle model to an ideal observer model that does not know the opponent's exact behavioral policies – as is the case in real-world lying – but must instead deduce the opponent's behavior from first principles. This ideal observer assumes that the opponent is rational, and thus, estimates the opponent's behavioral policies by performing recursive social inference. We found that both the oracle and recursive inferential models make qualitatively similar predictions about both lying (i.e. non-linear lies as a function of the true value) and lie detecting (i.e. logistic pattern of calling BS as a function of the lie). This lack of distinguishing predictions across these two models suggests that even though the recursive inferential model lacks the omniscience of the oracle model, it can reproduce qualitatively the same behavior with a far sparser explicit representation of the other agent.

The oracle and recursive theory of mind models are contrasted with an agent that has *no model* of the opposing agent. This agent lies by only considering rewards (and not the opponents' reaction), and detect lies by only considering what is improbable (and not what lies would favor the opponent). This agent makes qualitatively different predictions about both lying and lie detecting behavior. We then tested these model predictions by examining human behavior in a novel lying game. The empirical results suggest that the recursive inferential model of deception capture how human liars choose which lie to tell: they tend to choose lies that are not too implausible. Likewise, we find that lie detecting behavior is consistent with recursive ToM and is calibrated to the probability of the sample under the prior distribution of the world.

To better determine how recursion in these rational models maps onto human behavior, one natural future direction would be to have participants compete against each other in this game. Is it truly the case that liars assume lie detectors are reasoning rationally about the liar, and lie detectors assume liars are reasoning rationally about the lie detector?

In the current experiments, people played against a computer opponent with fixed, non-adaptive behavior. Perhaps over the iterative trials, participants perfectly adapted their model of the other agent, such that they knew how it would behave as a liar and lie detector–essentially acting like the oracle model, with no need for any more than first-order ToM, or ToM over an agent who does not assume rationality about the other agent. Do participants typically perform recursion, or do people only perform first-order ToM? Our first pass at providing evidence against this alternative hypothesis is shown in our second experiment in which the lack of feedback about the opponent's behavior requires players to generate a model of the agent without actual knowledge about the agent's behaviors. Given this lack of feedback, it would be far more difficult to develop an accurate non-inferential generative model of the other agent.

In summary, in the study we present here, we propose and contribute empirical evidence that liars and lie detectors act as rational utility-maximizing agents. Liars and lie detectors choose how to lie and when to call out lies under the assumption that the other agent is also behaving rationally. These findings provide a stepping stone for novel quantitative approaches to studying deception.

Acknowledgments

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1650112 to LAO, and National Science Foundation Grant No. BCS-1749551 to AS.

References

Allcot, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, *31*(2), 211–236.

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*, 329–349.

Ding, X. P., Wellman, H. M., Wang, Y., Fu, G., & Lee, K. (2015). Theory-of-mind training causes honest young children to lie. *Psychological Science*, *26*(11), 1812–1821.

Ekman, P., Friesen, W. V., & O'Sullivan, M. (1988). Smiles when lying. *Journal of Personality and Social Psychology*, *54*(3), 414–420.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*, 998.

Lewis, M., Stanger, C., & Sullivan, M. W. (1989). Deception in 3-year-olds. *Developmental Psychology*, *25*(3), 439–443.

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*. New York: Wiley.

Mazar, N., Amir, O., & Ariely, D. (2008). The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research*, *45*(6), 633–644.

Talwar, V., Gordon, H. M., & Lee, K. (2007). Lying in the elementary school years: Verbal deception and its relation to second-order belief understanding. *Developmental Psychology*, *43*(3), 804–810.

Talwar, V., & Lee, K. (2002). Development of lying to conceal a transgression: Children's control of expressive behaviour during verbal deception. *International Journal of Behavioral Development*, *26*(5), 436–446.

Toma, C. L., Hancock, J. T., & Ellison, N. B. (2008). Separating fact from fiction: An examination of deceptive self-presentation in online dating profiles. *Personality and Social Psychology Bulletin*, *34*(8), 1023–1036.

Vrij, A., Fisher, R., Mann, S., & Leal, S. (2006). Detecting deception by manipulating cognitive load. *Trends in Cognitive Sciences*, *10*(4), 141–142.

Vrij, A., Granhag, P. A., & Porter, S. (2010). Pitfalls and opportunities in nonverbal and verbal lie detection. *Psychological Science in the Public Interest*, *11*(3), 89–121.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, *13*, 103–128.

Imagining the good: An offline tendency to simulate good options even when no decision has to be made

Joan Danielle K. Ongchoco (joan.ongchoco@yale.edu)

Julian Jara-Ettinger (julian.jara-ettinger@yale.edu)

Joshua Knobe (joshua.knobe@yale.edu)

Abstract

Even when we are not faced with any decision, we sometimes engage in offline cognition where we simulate various possible actions we can take. In these instances, which options do we tend to simulate? Computational models have suggested that it is better to focus our limited cognitive resources towards simulating and refining our representations of options that appear, at first blush, to have higher values. Two experimental studies explore whether we use this strategy. Participants went through an 'offline' thinking phase, and an 'online' decision-making phase. Participants first freely viewed various options, which they had to simulate to determine their actual values. They were later asked to decide between good or bad options. Offline simulation produced faster online response times for the options that appeared to have higher values, indicating a pre-computation benefit for these items. These results suggest that people focus their offline cognition on the apparently good.

Keywords: Sampling; simulation; decision-making; mental rotation

Introduction

When people are trying to make decisions, they sometimes proceed by simulating possible options and asking what the outcome would be for each. Existing research has explored the various ways people use such simulations not just in making inferences about the world in general (e.g. Battaglia, Hamrick, & Tenenbaum, 2013; Callaway, Hamrick, & Griffiths, 2017; Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2014), but also in the specific context of decision-making (e.g. Barron, Dolan, & Behrens, 2013; Hamrick, Smith, Griffiths, & Vul, 2015; Hamrick et al., 2016; Lieder, Griffiths, & Hsu, 2018; Wimmer & Shohamy, 2012). Because of the limited time and capacity during online processing, a central question is how to efficiently allocate computational resources. Previous work has investigated how people determine which simulations and how many to run at the moment when they have to make a decision (e.g. Callaway, Gul, Krueger, Griffiths, & Lieder, 2018; Hamrick & Griffiths, 2014; Srivastava, Mller-Trede, Schrater, & Vul, 2016; Vul, Goodman, Griffiths, & Tenenbaum, 2014).

Importantly, however, people are also capable of simulating different possible options offline, i.e., considering possible options when they are not faced with any immediate decision (see, e.g. Gershman, Markman, & Otto, 2014). For example, even when you are not out with someone on a dinner date, you may find yourself simulating various possible ways you might introduce yourself to a (perhaps hypothetical) person. This offline simulation may then prove helpful when you later face an actual online decision-making problem.

Though our capacity for running simulations offline is not quite as limited as our capacity for running simulations online, we still cannot simulate all possible options. Thus, if you are thinking offline about how to introduce yourself on a date, you would inevitably simulate some options (e.g., talking about your background and interests) but not others (e.g., talking in detail about how loudly you snore). This raises the question—which options do people tend to simulate when thinking offline?

What *should* we think about offline?

One way in to this problem is to begin by asking which options it would actually be rational to simulate. Suppose our aim is to select the best action during a subsequent episode of online decision-making. Given this aim, which options would it be best to simulate offline?

Of course, one possible answer would be that it does not matter which specific options we end up simulating. Simulating different possible options for hypothetical situations might simply be helpful in a broad way, for learning the general features of good versus bad options, without having to specifically compute which option is better than another. In other words, running simulations may be a good way of discovering various heuristics about different options that we can then use later, during online decision-making.

An alternative possibility, however, is that simulating offline is not just good for learning various decision-making heuristics, but can also help us get better value estimates for specific options. When we simulate an option, we can improve our representation of the value of that option. This 'pre-computed' value can come in handy when we have to make decisions in the pressure of the moment, when we do not have much time to think.

The problem can then be formulated as follows. At any given point, we have a representation of the value of each option. Some options are represented with high values (i.e. as good options), others with low values (i.e. as bad options), and others as having an intermediate level of value. At first blush, all of these representations will be at least somewhat inaccurate. We may have a sense of what is good or bad, but generally need to think more about which of these is actually the best or the worst. In simulating a specific option, we can then improve our representation of its value. However, we cannot run simulations for all options. Thus, we have to

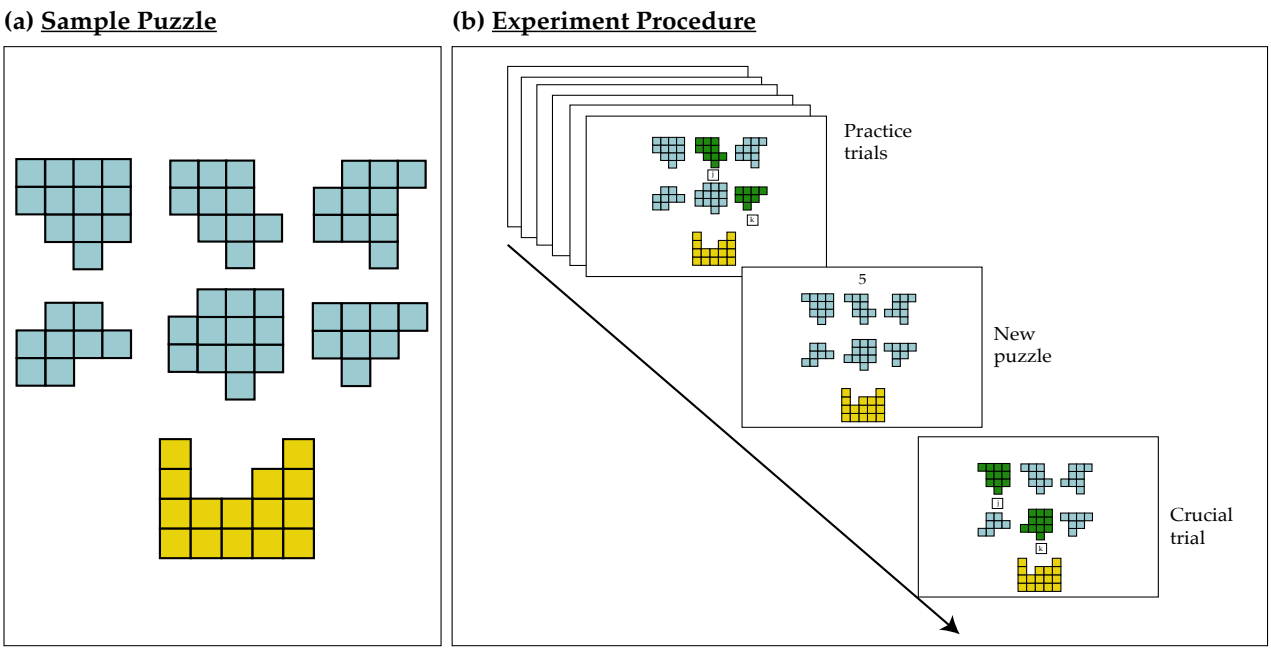


Figure 1: (a) A sample puzzle of the sort used in the experiments. The yellow blocks constitute the puzzle, the blue blocks constitute the pieces. While looking at the puzzle, readers may simulate which pieces would fit or not fit. (b) A caricatured depiction of a sample experimental procedure without the instructions. Participants first go through a series of practice trials, are given a unique puzzle, and are then asked to decide between two options.

decide which of our representations we want to have a more accurate value estimate of, while allowing others to remain inaccurate.

One intriguing finding from existing research is that, other things being equal, it is generally a good idea to run offline simulations of options that we initially think have high values (e.g. Gelly & Silver, 2011; Icard, Cushman, & Knobe, 2018). In other words, if we can only improve our representation of some options, it is better to improve our value estimates of the options we initially represent to have high value than to improve our value estimates of the options we now represent to have low value.

Existing computational work has explored this point within the framework of reinforcement learning (e.g. Icard et al., 2018), but the core intuition is easy to grasp even independently of any formal framework. Suppose that there are now two different inaccuracies in your representations: (a) the option that you mistakenly represent as second-best is actually the best, and (b) the option that you mistakenly represent as second-worst is actually the worst. Now suppose that you are only able to correct one of these inaccuracies, which would you focus on?

The key point is that when you later use these representations in online decision-making, you would ideally want to choose the best option. Thus, it is important to be highly accurate about which of the good options truly is the best, but it is not nearly as important to be accurate about which of the bad options truly is the worst. You should therefore devote your limited offline cognition to the options that you initially think to have high value, and then improve your representations from there.

A question now arises as to whether human cognition actually works in this way. When people only have a limited amount of time to devote to offline cognition, do they tend to run simulations of the options they regard as having high value, even when they do not have to (as there are infinite possibilities one can simulate offline, and there is no immediate specific decision that has to be made)?

The present studies

To address these questions, we conducted two studies in which participants had an opportunity to go through an 'offline' thinking phase before a subsequent 'online' decision-making phase. In the offline phase, participants were given an array of options to freely think about. Crucially, they had to simulate these different options to determine their actual values. In the online phase, participants were asked to decide between two options. The key question was which options would participants think about during the offline phase, when they were not told what decisions they eventually would have to make. To tap into participants' tendencies during offline simulation, we used their response times during the online decision-making phase. We reasoned that if participants were refining their values about specific options and

computing which ones were better than others during the offline phase, then they should respond faster when choosing between those same options during the online phase.

In a novel paradigm, we used incomplete block-puzzles (that look like Tetris) with arrays of different puzzle pieces that would either fit the puzzle or not. In this design, determining whether the puzzle pieces fit would require participants to manipulate these pieces in their minds, akin to classic mental simulation and rotation studies (e.g. Cooper, 1975; Shepard & Metzler, 1971). Moreover, this block-puzzle design allowed us to also specify a 'surface' or apparent value (what people initially think the value of a piece to be) and an actual value (the value of the piece after being simulated) for each puzzle piece, where value could be defined by both the number of blocks the piece had, and whether the piece would actually fit the puzzle.

The idea behind this particular design was that, at first glance, one should be able to immediately 'see' the surface values of the different puzzle pieces, such that some pieces would clearly have higher values (as indicated by the brute number of blocks) than others. From this surface value, one can either simulate the apparently good or the apparently bad. Crucially, it is only by mentally rotating and simulating how these pieces would fit the incomplete block-puzzle that one can get a better sense of the actual values of these pieces. However, one cannot simulate all pieces during the limited window of the offline phase. This time limit allowed us to check which pieces people would simulate over others.

In Experiment 1, we looked at whether people systematically responded faster during online decision-making to some options over others as a function of offline simulation. In Experiment 2, we investigated the mechanism by which offline simulation may lead to benefits in online decision-making, as a function of developing broad heuristics about what options are good and bad in general versus actually precomputing and refining the value of specific options. These experiments altogether explore the principles governing online and offline thinking, and suggest that these may in fact be more closely related than we previously thought: people systematically and actively imagine the good not only when there is an immediate judgement or decision to be made, but also offline, even when they do not have to.

Experiment 1

Participants were given an array of six rotated puzzle pieces per incomplete puzzle during the offline phase. Each piece had a specific value, defined by how many blocks would end up above the puzzle, once the piece fit. In general, the more blocks a piece had, the better. However, to determine the precise value of the piece, participants had to simulate the different pieces. In Figure 1a, the upper leftmost piece would fit the puzzle when rotated counter-clockwise, and would have 7 blocks above the completed puzzle. If participants selected this piece, they would get 7 points. In contrast, the lower middle piece has the same number of blocks, but would not fit the puzzle. If participants chose this piece, they would get 0 points. Thus, we wanted to see whether people would consider the pieces that have a high surface value (like 7) during the offline thinking phase. If participants consider some pieces more than others, they might respond faster when they have to decide between these specific pieces.

Method

All methods and analyses were pre-registered (`http://aspredicted.org/blind.php?x=rd2bd2`). Data and code for all experiments reported here are available on `https://osf.io/npwdq/?view_only=a808e1dd2d594b7992892bfa32fb7e8c`.

Participants. Sixty subjects from the Yale University Library participated (with candy as compensation). The sample size was determined before data collection began.

Apparatus. Stimuli were presented using custom software written in Python with the PsychoPy libraries (Peirce, 2007) and were displayed on a monitor with a 60Hz refresh rate. Participants completed the study on a 13-inch MacBook Air with a 1440 x 900 resolution.

Stimuli. Puzzles were generated randomly through PsychoPy. Puzzles were made of 20 yellow blocks (0.5° black border) stacked in 4 rows of 5 blocks each. Each block was 2° in size. In each puzzle, a number—three or four—of the blocks in the top two rows would be missing. This created an incomplete section at the top of the puzzle.

The puzzle pieces were also generated randomly. Pieces comprised of the specific arrangement of blocks that were determined to be missing, along with additional blocks that made up the value the piece was assigned (e.g. a value of 7 meant that there were 7 blocks on top of the piece). Additional blocks were stacked on top of the piece randomly, for as long as they were always connected to a block in the piece. When the blocks were stacked, the piece was checked on all sides to make sure that only one side would fit in the puzzle. If the piece was not supposed to fit (i.e. have a value of 0), the bottom-most part of the piece was shifted to the left or to the right, in order to ensure that the piece would not fit the puzzle. Puzzle pieces were made out of 2° grey blocks (0.5° black border).

Procedure and design. Throughout the experiment, on the top-left of the screen, there was 'Total Points:' counter. All the text in this experiment was drawn in black Monaco font (0.6° in height). In a single-trial experiment, participants first went through the instructions for the task. They were told that their goal was to earn as many points as they could. They were given sample incomplete block-puzzles and arrays of possible options. They were told that they would be asked questions about these different options afterwards, and would get a number of points corresponding to the particular option they would be asked about. They were told that the value of each piece was defined by the number of blocks

that would end up above the completed puzzle once the piece fit, and that pieces would also be rotated but never flipped. They were also told that speed in responding will be important so the participants would go through a practice section first before the actual trial. After these instructions, participants would then be shown the puzzle, which subtended from 1° above to -3° below the center.

In the offline phase, participants were told that six pieces would now appear above the puzzle, in two rows of 3 pieces each. Participants were told that they did not have to do anything but just look and study the pieces. The six pieces comprised of three pairs with surface values of 3, 5, and 7. In each pair, each piece would have a different actual value: one would fit, and the other was made to not fit the puzzle. During this time, a countdown timer (6° below the center) would start from 5 and decrease per each passing second.

After the offline phase, the online decision-making phase began, where participants now had to decide which of two pieces was the better piece. The blocks of two of the pieces would turn from grey to green to indicate which pieces the participants would have to choose from, and each of these pieces was assigned either a letter j or k. Participants were simply asked, "Which piece is better?", and indicated which piece they preferred by keying in the letter of the piece. In the practice section, participants responded to a total of six practice trials. Throughout the practice trials, if participants responded correctly, the total points counter would increase by the value of the piece they chose (if the piece fit, then this value was determined by the number of blocks above the completed puzzle; if the piece did not fit, the participants would automatically get 0 points).

After participants completed the practice section, they were told that they would be shown a different puzzle and a new set of pieces. Participants were again told that they could be asked about any of these pieces afterwards. To facilitate the pressure of having to decide in the moment, participants were now encouraged to respond as fast as possible, and were told that they would get bonus points for responding quickly. Participants first went through the offline phase, where they were again presented a new puzzle with six puzzle pieces. The countdown timer appeared again. After five seconds, participants began the online decision-making phase, responded to two pieces from the array of six options. In the Good Options condition, participants decided between the two pieces with a value of 7. In the Bad Options condition, participants decided between the two pieces with a value of 3. Participants were randomly assigned to decide either between the good options or the bad options.

Results and discussion

Three participants were excluded because their mean performance in the practice section was 2 standard deviations below the grand population mean (M=29.08 out of 34 total points that could be earned in the practice section; the cut-off was at 18.63). These subjects were replaced, until a total of 60 participants was reached (30 decided between the good options, and 30 decided between the bad options). Response accuracy and response times for the single trial were recorded for each observer. Only response times where participants responded correctly were included in the analysis.

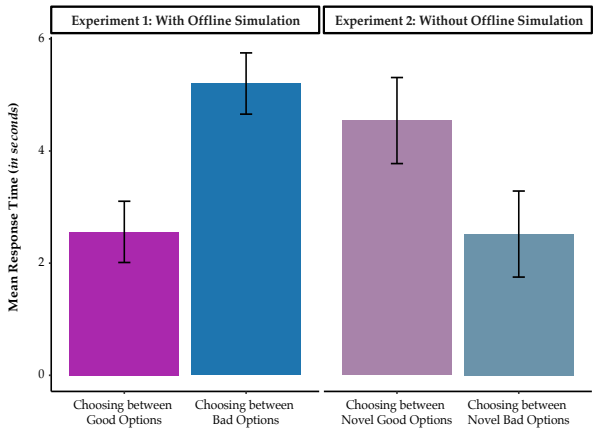


Figure 2: Results from Experiments 1 and 2. The left bar graph presents the mean response times per decision type in Experiment 1, while the right bar graph presents the mean response times per decision type in Experiment 2. The vertical axis represents the mean response time in seconds. The horizontal axis represents the key comparisons in both experiments. The error bars reflect 95% confidence intervals within experiments.

Initial inspection of the left bar graph in Figure 2 shows a lower mean response time for good options than for the bad options. This initial impression was confirmed with statistical tests. Mean reaction time for good options (M=2.56s, MSD=1.77s) was significantly faster than for bad options (M=5.20s, SD=3.82s), $t(25.31)=2.87$, $p=.008$, $d=0.93$. Because the response time distributions violated the normality assumption, we conducted a t-test on the log transformations of the distributions (which now meet the normality assumption), $t(43.37)=3.65$, $p<.001$, $d=1.06$. (We also note that including the incorrect answers did not yield any different results, $t(55.41)=3.51$, $p<.001$, $d=0.91$). There was no significant difference between the percentage of people who responded accurately when choosing between good options (86.67%) vs. bad options (66.67%) (Fisher's exact, $p=.125$).

These results suggest a pre-computation 'imagination' benefit for the good options. In other words, it appears that when given the opportunity to freely think about an array of various options offline, people tend to simulate the options they initially think have a higher value rather than those they initially think to have lower values, even when they do not know what specific decisions they will have to make later on. In simulating the good options, people can determine offline what the actual values of these good options are, such that when it comes to having to make a decision, they re-

spond faster to the good options they had already simulated, computed, and compared beforehand.

Experiment 2

The results from the initial experiment were promising because they suggested that when given an opportunity to think, participants think about and simulate the good options more than the bad options, resulting in a response time benefit at the time of decision-making. But is this a benefit from simulating specific options and refining our representations of their values, and not just from being exposed to and learning what the good or bad options are (for instance, participants could simply have been learning throughout the practice section that the bigger pieces, regardless of their specific configurations, have higher values)? To explore the mechanism underlying the response time benefit observed in Experiment 1, we used the same block-puzzle design. We asked participants to again look at a puzzle and an array of six options. This time, during decision-making, unbeknownst to the participants, instead of presenting them with pieces that were originally in the array of six options, we presented them with a novel pair of good or bad options. Thus, none of their offline thinking strategies should have changed, since they were not told that they would be shown novel pieces. If offline simulation were simply a way of discovering heuristics about which options are good or bad, then participants should still respond faster to the good options than the bad options. However, if offline simulation involves the precomputation of the values of specific pieces, then the precomputation benefit observed in Experiment 1 should disappear when participants are presented with a novel pair.

Method

This experiment was identical to Experiment 1, except as noted. Sixty new participants participated, with this sample size chosen to match Experiment 1. During the decision-making phase, a new pair of pieces with the value of either 7 (i.e. Novel Good Options condition) or 3 (i.e. Novel Bad Options condition) were generated and presented to the participants. All methods and analyses were pre-registered ((`http://aspredicted.org/blind.php?x=bw5n59`).

Results and discussion

One participant was excluded because their mean performance in the practice section was 2 standard deviations below the grand population mean (M=29.08 out of 34 total points that could be earned in the practice section; the cutoff was at 20.65). This subject was replaced, until a total of 60 participants was reached (30 decided between the good options, and 30 decided between the bad options). Response accuracy and response times for the single trial were recorded for each observer. Only response times where participants responded correctly were included in the analysis.

Initial inspection of the right bar graph in Figure 2 shows a lower mean response time for bad options than for the good options. Mean reaction time for good options (M=4.54s, SD=2.51s) was significantly slower than for the bad options (M=2.52s, SD=1.12s), $t(30.52)=3.53$, $p=.001$, $d=1.04$. Again, because the response time distributions violated the normality assumption, we conducted a t-test on the log transformations of the distributions, $t(44.51)=3.45$, $p=.001$, $d=0.99$. (We also note that including the incorrect answers again did not yield any different results, $t(58.22)=2.36$, $p=.021$, $d=0.60$). There was no significant difference between the percentage of people who responded accurately when choosing between good options (76.67%) vs. bad options (76.67%) (Fisher's exact, $p=1$).

To compare these results with those of Experiment 1, we ran a 2 (offline vs. no offline phase) x 2 (good options vs. bad options) ANOVA. There was no main effect of offline thinking, $F(1, 88)=0.12$, $p=.728$, $\eta^2=.002$, or of decision type, $F(1, 88)=0.33$, $p=.570$, $\eta^2=.004$. Crucially, there was a significant interaction, $F(1, 88)=20.99$, $p<.001$, $\eta^2=.193$.

In short, these results show a reversal of the pattern observed in Experiment 1. Since the task is constructed in such a way that the higher value pieces contain more blocks, one might expect at baseline that participants would show longer reaction times for the higher value pieces. In Experiment 1, where participants had an opportunity to engage in offline simulation, we instead found shorter reaction times for the higher value pieces. By contrast, in the present study, we find the expected baseline result: when participants do not have an opportunity to engage in offline simulation, they show longer reaction times for the higher value pieces (perhaps because they had more blocks in general).

General Discussion

There are many instances when we imagine different options without having to immediately make a decision, as when we daydream about which restaurant to go to for dinner or what to say when we are on a date or in an important meeting. In these instances of offline simulation, what do we tend to think about, and why? The present experiments explored this question in terms of the mental simulation of visual stimuli, and asked whether people tend to simulate the apparently good options over the apparently bad options.

The key takeaway from these experiments is simple to summarize: people choosing between two good options responded faster at the point of decision-making than people choosing between two bad options, suggesting that people were thinking more about the good options during the offline thinking phase, when they did not actually have to (and we note, interestingly, even when the good options were more difficult to think about and took longer to process at baseline). Moreover, this does not seem to be just a matter of general practice and exposure to deciding between good versus bad options. When presented a novel pair of good or bad options, participants no longer show this pre-computation benefit, and in fact, perform in the opposite way (responding slower to good options than the bad options). This suggests that thinking in the general does not suffice to produce the

benefit at decision-making. Rather it is thinking and mentally simulating specific possible options offline that proves particularly adaptive when eventually having to choose between these same options.

This result adds to the existing body of work that has explored what people should think about given limited computational resources (e.g. Callaway et al., 2018; Srivastava et al., 2016; Vul et al., 2014). In online decision-making, it generally makes sense to be actively sampling the options with the highest values in order to make the best decision. Our results demonstrate that simulating the best possible options also occurs in offline cognition, when people are allowed to freely think about any option, and do not have to make any decision at all. The tendency to imagine the good options may reflect a more general principle of cognition that is at play while running both online and offline simulations.

This tendency might be interestingly related to recent work on mind-wandering and on memory replay. Research on mind-wandering finds that people tend to spend a good amount of their waking hours just thinking offline (e.g. Mason et al., 2007). Such research indicates that peoples minds are in general more likely to wander to pleasant topics than unpleasant topics (e.g. Killingsworth & Gilbert, 2010). A separate strand of literature, mostly focused on nonhuman animals, has explored the 'replay' of memories. Intriguingly, this literature indicates a similar tendency: animals tend to replay particular memories in proportion to potential gain, and that this process may support future decision-making (see Mattar & Daw, 2018). Future work should explore the potential connection between these two strands of research and the patterns of offline simulation observed here.

These results are also relevant to previous work on people's judgements in moral situations. Existing research suggests that moral judgments can impact people's intuitions about causation, intentional action, and a variety of other apparently non-moral issues (Knobe, 2010). One hypothesis about these effects is that they are explained by a tendency to simulate counterfactuals in which agents perform actions that are morally good, and not to simulate counterfactuals in which agents perform actions that are morally bad (e.g. Icard, Kominsky, & Knobe, 2017; Phillips, Luguri, & Knobe, 2015). Future research could ask whether this tendency is best understood as just another manifestation of the same basic pattern observed in the present studies.

The principal contribution of the present studies is its suggestion that our cognition is particularly attuned to the best possible options, regardless of whether there is an immediate decision that has to be made. One possibility is that our minds are simply wired to default to simulating the good possibilities during offline cognition and that people will therefore show this tendency even when they do not want to be thinking of the good (as when they do not want to get their hopes up), or even when it may not even be beneficial to the task to be thinking of the good (as when they need to be looking out for potential worst-case scenarios).

But another possibility is that our offline tendencies are more flexible depending on the context. Here we explored cases where people can choose which option they want, making it rational to identify the best possible ones. Yet, critical life events are often out of our control, and we can do nothing but prepare for what may come. In cases like these, we may hope for the best and prepare for the worst, making it rational to switch our offline tendencies to focus on bad outcomes to decide what to do in response. We are curious about whether people will show this same tendency when they have less control over which options they end up with, or when there is greater uncertainty about the bad outcomes. Future work can explore the boundaries of this offline tendency to imagine the good.

References

Barron, H. C., Dolan, R. J., & Behrens, T. E. (2013). Online evaluation of novel choices by simultaneous representation of multiple memories. *Nature neuroscience*, *16*, 1492–1498.

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, *110*, 18327–18332.

Callaway, F., Gul, S., Krueger, P. M., Griffiths, T. L., & Lieder, F. (2018). Learning to select computations. In *NIPS workshop on cognitively inspired Artificial Intelligence.*

Callaway, F., Hamrick, J. B., & Griffiths, T. L. (2017). Discovering simple heuristics from mental simulation. In *Proceedings of the Annual Conference of the Cognitive Science Society* (Vol. 39).

Cooper, L. A. (1975). Mental rotation of random two-dimensional shapes. *Cognitive psychology*, *7*, 20–43.

Gelly, S., & Silver, D. (2011). Monte-Carlo tree search and rapid action value estimation in computer Go. *Artificial Intelligence*, *175*, 1856–1875.

Gershman, S. J., Markman, A. B., & Otto, A. R. (2014). Retrospective revaluation in sequential decision making: A tale of two systems. *Journal of Experimental Psychology: General*, *143*, 182–194.

Gerstenberg, T., Goodman, N., Lagnado, D., & Tenenbaum, J. (2014). From counterfactual simulation to causal judgment. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 36).

Hamrick, J. B., & Griffiths, T. L. (2014). What to simulate? Inferring the right direction for mental rotation. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 36).

Hamrick, J. B., Pascanu, R., Vinyals, O., Ballard, A., Heess, N., & Battaglia, P. (2016). Imagination-based decision making with physical models in deep neural networks. In *NIPS 2016 workshop on intuitive physics.*

Hamrick, J. B., Smith, K. A., Griffiths, T. L., & Vul, E. (2015). Think again? The amount of mental simula-

tion tracks uncertainty in the outcome. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 37).

Icard, T. F., Cushman, F., & Knobe, J. (2018). On the instrumental value of hypothetical and counterfactual thought. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 40).

Icard, T. F., Kominsky, J. F., & Knobe, J. (2017). Normality and actual causal strength. *Cognition*, *161*, 80–93.

Killingsworth, M. A., & Gilbert, D. T. (2010). A wandering mind is an unhappy mind. *Science*, *330*, 932–932.

Knobe, J. (2010). Person as scientist, person as moralist. *Behavioral and Brain Sciences*, *33*, 315–329.

Lieder, F., Griffiths, T. L., & Hsu, M. (2018). Overrepresentation of extreme events in decision making reflects rational use of cognitive resources. *Psychological review*, *125*, 1–32.

Mason, M. F., Norton, M. I., Van Horn, J. D., Wegner, D. M., Grafton, S. T., & Macrae, C. N. (2007). Wandering minds: the default network and stimulus-independent thought. *Science*, *315*, 393–395.

Mattar, M. G., & Daw, N. D. (2018). Prioritized memory access explains planning and hippocampal replay. *Nature Neuroscience*, *21*, 1609–1617.

Peirce, J. W. (2007). PsychoPyPsychophysics software in Python. *Journal of neuroscience methods*, *162*, 8–13.

Phillips, J., Luguri, J. B., & Knobe, J. (2015). Unifying moralitys influence on non-moral judgments: The relevance of alternative possibilities. *Cognition*, *145*, 30–42.

Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, *171*, 701–703.

Srivastava, N., Mller-Trede, J., Schrater, P. R., & Vul, E. (2016). Modeling sampling duration in decisions from experience. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 38).

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? Optimal decisions from very few samples. *Cognitive science*, *38*, 599–637.

Wimmer, G. E., & Shohamy, D. (2012). Preference by association: how memory mechanisms in the hippocampus bias decisions. *Science*, *338*, 270–273.

Risk is Preferred at Lower Causal Depth

Jeffrey Parker
University of Illinois at Chicago, Chicago, Illinois, United States

Abstract

Risk and uncertainty are inherent in life, and how people perceive, respond to, and manage both are topics of great academic interest. One critical insight is that people distinguish between types of uncertainty (see, e.g., Fox & lkmen, 2011) and, consequently, may respond to objectively equally probabilistic events differently (e.g., with more polarized predictions of those events outcomes). The current work identifies another way in which risk (a specific form of uncertainty) is differentiated: on the basis of causal depth (Sloman, Love, & Ahn, 1998). Specifically, in contexts where an uncertain outcome (e.g., win/lose) is determined by a causal chain, people tend to prefer for the uncertainty to arise at lower causal depth within the chain (i.e., at later causal stages). This occurs even though the causal depth at which the uncertainty arises makes no difference in the overall probability that the causal chain will generate one outcome or another.

The interactions of rational, pragmatic agents lead to efficient language structure and use

Benjamin N. Peloquin
bpeloqui@stanford.edu
Department of Psychology
Stanford University

Noah D. Goodman
ngoodman@stanford.edu
Department of Computer Science
Stanford University

Michael C. Frank
mcfrank@university.edu
Department of Psychology
Stanford University

Abstract

Despite their diversity, languages around the world share a consistent set of properties and distributional regularities. For example, the distribution of word frequencies, the distribution of syntactic dependency lengths, and the presence of ambiguity are all remarkably consistent across languages. We discuss a framework for studying how these system-level properties emerge from local, in-the-moment interactions of rational, pragmatic speakers and listeners. To do so, we derive a novel objective function for measuring the communicative efficiency of linguistic systems in terms of the interactions of speakers and listeners. We examine the behavior of this objective in a series of simulations focusing on the communicative function of ambiguity in language. These simulations suggest that rational pragmatic agents will produce communicatively efficient systems and that interactions between such agents provide a framework for examining efficient properties of language structure and use more broadly.

Keywords: Communicative efficiency, Rational Speech Act theory, computational modeling, information theory, agent-based simulation

Introduction

Why do languages look the way they do? Zipf (1949) proposed that distributional properties found in natural language were evidence of speaker-listener effort minimization. In his own words, "we are arguing that people do in fact act with a maximum economy of effort, and that therefore in the process of speaking-listening they will automatically minimize the expenditure of effort." Evidence for this claim has been largely derived at the level of the lexicon. Zipf argued that the particular relationship between a word's frequency and its rank, length, and denotation size could be explained as an emergent property of speaker-listener effort minimization.

Zipf articulated what is now considered a *functionalist* approach to language science – analyzing language structure and use in terms of efficiency. Such an approach might reframe our opening question as follows: how does having property x make using language ℓ more or less useful for communication? This efficiency-based framing has produced a rich set of theoretical and empirical targets exploring semantic typology (Regier, Kemp, & Kay, 2015), properties such as ambiguity (Piantadosi, Tily, & Gibson, 2011) and compositionality (Kirby, Griffiths, & Smith, 2014), and the efficient use of reduction and redundancy in production (Genzel & Charniak, 2002; Levy & Jaeger, 2007).

The approaches above typically posit efficiency measures that are motivated by information-theoretic principles, but they typically do not ground out in language use by interacting agents. In this work, we derive a novel objective function from first principles of *rational language use* and show how optimizing this objective can lead to communicatively efficient systems. We also demonstrate that assumptions about interlocutors impact whether language properties are used efficiently. In this way, we integrate questions of language design and language use in a single framework.

Functionalist theories commonly frame language efficiency in terms of a fundamental effort-asymmetry underlying everyday communication: what is "hard" for a speaker is likely different than what is "hard" for a listener. Zipf described this as follows: purely from the standpoint of speaker effort, an optimal language $\ell^*_{speaker}$ would tend toward a vocabulary of a single, low-cost word. Given such a language, the full set of potential meanings would be conveyed using only that word, i.e. $\ell^*_{speaker}$ would be fully ambiguous and all possible meanings would need to be disambiguated by a listener. From the standpoint of listener effort, an optimal language $\ell^*_{listener}$ would map all possible meanings to distinct words, removing a listener's need to disambiguate. In this example, speaker effort is related to *production cost* and listener effort to *understanding or disambiguation cost*. Clearly, natural languages fall between the two extremes of $\ell^*_{speaker}$ and $\ell^*_{listener}$. Zipf proposed that the particular lexicon-level properties he observed were a result of optimization based on these competing forces – the pressure to jointly minimize speaker and listener effort.

But how does this optimization take place? The example given by Zipf (1949) describes local, communicative interactions in terms of a *reference game*. Speakers intend to refer to some object in the world m. They choose some utterance u to transmit this intended meaning, $u \rightarrow m$. The listener attempts to reconstruct this intended meaning given the transmitted utterance, $m \rightarrow u$. Other projects have assumed this basic reference game setting (Piantadosi et al., 2011; Regier et al., 2015) and this simplification of the communicative act has proven productive in theoretical (Ferrer-i-Cancho, 2018), simulation-based (Kirby et al., 2014) and empirical explorations (Hawkins, Franke, Smith, & Goodman, 2018) of efficient language structure and use.

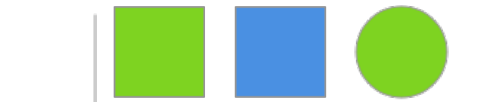
Adopting reference games as a basic unit of analysis suggests that optimization may take place at the level of conversation. Importantly, Zipf's conception of speaker and listener

effort should be connected to how language is used; in particular, whether interlocutors engage in pragmatic reasoning during conversation. Under a Gricean treatment of pragmatics, speakers and listeners follow a set of conversational maxims in which they cooperate to transfer information (Grice, 1975). These maxims appear to emerge from efficiency concerns, however (Horn, 1984). We formalize this connection – showing how system-level efficiencies can emerge from local interaction behavior of pragmatic agents. Our claim is that to understand an "efficient" property of a system it is essential that we consider how that property is *used* efficiently.

We provide a case study for this approach, in which functionalist regularities emerge from the dynamics of pragmatic communication. We choose a property of languages that could, in principle, vary freely, but shows strong regularities across languages. The explanandum is why this regularity holds. We examine ambiguity as our property, extending ideas by Piantadosi et al. (2011). We define a novel measure of efficiency that depends on the interactional behavior of speaker and listener agents. We adopt the reference game as our primary unit of interaction and model language users with the Rational Speech Act (RSA) framework – a computational model of language use, which is supported by experimental data on interaction. Using these ingredients, we show that the property of interest (ambiguity) is prevalent in languages that optimize our measure of efficiency (Simulation 1). Further, we show how ambiguity is used efficiently during local, in-the-moment interactions (Simulation 2). Put differently, these simulations examine efficiency from two angles – in the first we vary languages, fixing agents, and search for efficient language designs. In the second we vary agents, fixing language, and examine efficient use.

The contributions of this work are twofold – we derive a novel measure of linguistic efficiency and also show how the reference game framework, in combination with formal models of communication, can be used to connect ideas about system-level efficiencies to in-the-moment language use.

Exploring efficient language design and use in rational pragmatic agents



square	1	1	0
circle	0	0	1
green	1	0	1
blue	0	1	0

Figure 1: An example reference game with associated literal semantics (in our terminology a "language").

Reference games Zipf's example of optimal speaker- and listener-languages took the form of a reference game. We adopt that formulation here, assuming these communication games as our basic unit of analysis. In this framework, speakers and listeners are aware of a set of objects M (*meanings*) and are knowledgeable about the set of possible signals U (*utterances*) that can be used to refer to a given meaning (see Figure 1). Utterances may have different relative costs, operationalized via a prior over utterances $P(U)$. Similarly, meanings differ in the relative degree to which they need to be talked about, operationalized as a prior over meanings $P(M)$[1]. We consider a set of contexts C with an associated prior $P(C)$. Each context $c \in C$ describes a different distribution over meanings e.g. $p(M|C = c_i) \neq p(M|C = c_j)$. Finally, we consider a set of communicative events $e \in E$ where $< u, m, c >= e$ is an utterance-meaning-context triple.

Languages A language ℓ defines the set of semantic mappings between utterances and meanings. For example, Figure 1 contains four utterances $U = \{$"blue","green","square","circle"$\}$ and three meanings $M = \{\text{green-square}, \text{blue-square}, \text{green-circle}\}$. The boolean matrix describes the literal semantics of the language. We define a language as "ambiguous" if there is some utterance $u \in U$ which can apply to multiple meanings (i.e. $|[[u_i]]| > 1$)[2]. In Figure 1 both the words "square" and "green" are ambiguous so we would say that ℓ contains ambiguity.

Speakers and listeners The Rational Speech Act framework (RSA) is a computational-level theory of pragmatic language use, which has produced good fit to human communication behavior across a range of language phenomena (Frank & Goodman, 2012; Goodman & Frank, 2016). RSA is a formalization of essential Gricean pragmatic principles – agents reason about one another and their shared context (Grice, 1975). We adopt RSA as our representational framework to model Gricean (rational and pragmatic) speakers and listeners in the reference game setting (see SI).

An RSA *speaker agent* defines a conditional distribution over utterances, mapping from intended meanings M to utterances U using ℓ in a given context c. That is, a speaker defines $P_{speaker}(u|m,c;\ell)$. We will use $S(u|m,c;\ell)$ as short-hand throughout. A *listener agent* defines a conditional distribution over meanings, mapping from utterances U to meanings M using ℓ in a given context c (i.e. $L(m|u,c;\ell)$). Speakers and listeners can induce joint distributions over utterance-meaning pairs, although, these distributions may differ:

$$P_{speaker}(u,m|c;\ell) = S(u|m,c;\ell)p(m|c)$$

$$P_{listener}(u,m|c;\ell) = L(m|u,c;\ell)p(u|c)$$

Zipfian objective for linguistic system efficiency

Zipf (1949) proposed that the particular distributional properties found in natural language emerge from competing

[1]The prior over meanings is equivalent to the *need probabilities* assumed in previous work (Regier, Kemp & Kay (2015)).

[2]We use double brackets $[[\ldots]]$ to represent denotation.

speaker and listener pressures. We operationalize this objective in equation (1) – the efficiency of a linguistic system ℓ being used by speaker and listener agents S and L is the sum of the expected speaker and listener effort to communicate over all possible communicative events $e \in E$.

$$\begin{aligned}\text{Efficiency}(S,L,\ell) = \mathbb{E}_{e\sim P(E)}[\text{speaker effort}] \\ +\mathbb{E}_{e\sim P(E)}[\text{listener effort}]\end{aligned} \quad (1)$$

We assume that speaker effort is related to the surprisal of an utterance in a particular context[3] – intuitively, the number of bits needed to encode the utterance u. This particular formalization of speaker-cost is general enough to accommodate a range of cost instantiations, such as production difficulty via articulation effort, cognitive effort related to lexical access, or others (Bennett & Goodman, 2018).

$$\text{speaker effort} = -log_2(p(u|c))$$

We assume listener effort is the semantic surprisal of a meaning given an utterance. This operationalization of listener effort is intuitively related to existing work in sentence processing in which word comprehension difficulty is proportional to surprisal (Hale, 2001; Levy, 2008).

$$\text{listener effort} = -log_2(L(m|u,c;\ell))$$

Importantly, we assume that events $e =< u,m,c >$ are sampled according to the following generative model – some context occurs in the world with probability $P(C = c)$. Within this context, an object m occurs with probability $p(m|c)$. The speaker attempts to refer to that object by sampling from her conditional distribution $S(u|m,c;\ell)$ (i.e. $e \sim p(c)p(m|c)S(u|m,c;\ell)$). From these ingredients it is possible to derive the following objective between the speaker and listener distributions (see SI 2.1 for complete derivation).

$$= \mathbb{E}_{c\sim P(C)}[H_{cross}(P_{speaker}, P_{listener}|c;\ell)] \quad (2)$$

From an information-theoretic perspective this objective is intuitive: H_{cross} denotes the Cross-Entropy (CE), a measure of dissimilarity between two distributions – the average number of bits required to communicate under one distribution, given that the "true" distribution differs. In our case, we have an expectation over this term – the expected difference between the distributions assumed by the speaker $P_{speaker}$ and listener $P_{listener}$ given a set of contexts C[4]. In other words, an "efficient" language ℓ minimizes the distance between what speakers and listeners think.

[3]In the current set of simulations we consider utterance costs as independent from context (i.e.. $p(u|c)p(c) = p(u)p(c)$).

[4]Note that in the single context case $|C| = 1$ this objective is simply the speaker-listener Cross-Entropy.

Simulating the communicative function of ambiguity

The task of understanding language is marked by a frequent need to handle various forms of ambiguity: lexical, syntactic, among others (Wasow, Perfors, & Beaver, 2005). The ubiquity of this property, however, has been argued to provide evidence that languages have not been optimized for communication (Chomsky, 2002).

Piantadosi et al. (2011) argue just the opposite, claiming that ambiguity is an efficient property of any communication system in which *communication is contextualized*. Simply put, it is useful to have a language that re-uses low-cost material (has ambiguity) so long as the cost of disambiguating the material is low. In particular, context (or common ground) can provide useful information for disambiguation.

As an example, say we have two objects (m_1 and m_2), two utterances (u_1 and u_2), differing in cost, and two languages (ℓ_1 and ℓ_2), describing different utterance-meaning mappings. In language ℓ_1, the low-cost u_1 can be used to refer to both m_1 and m_2 ($[[u_1]]_{\ell_1} = \{m_1, m_2\}$), but the high-cost u_2 cannot be used at all ($[[u_2]]_{\ell_1} = \emptyset$). By contrast, in language ℓ_2, u_1 can only refer to m_1 and u_2 can only refer to m_2 ($[[u_1]]_{\ell_2} = \{m_1\}$ and $[[u_2]]_{\ell_2} = \{m_2\}$). While it is cheaper for a speaker to use ℓ_1 (because speaking it is always lower cost), it is more difficult for a listener (because u_1 is ambiguous). Crucially, if context is disambiguating then the speaker can use u_1 to refer to either m_1 or m_2 and ℓ_1 should be preferred to ℓ_2.

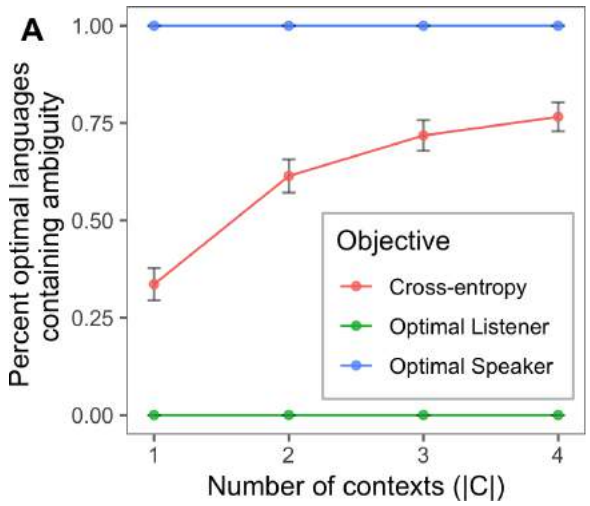
In the following simulations we explore two aspects of Piantadosi's et al.'s claim. In Simulation 1, we examine the efficient language *structure* aspect of their claim, exploring when the optimal linguistic system ℓ^* is most likely to contain ambiguous expressions. In Simulation 2, we explore an efficient language *use* aspect of the claim – under what assumptions will agents use ambiguity efficiently in a conversation?

Simulation 1: Optimal languages contain ambiguity when context is informative

We show that ambiguity is an efficient property under our CE objective in the reference game setting. We proceed by generating languages with different amounts of contextual support (varying the size of $|C|$). We search the space of languages, examining whether ones which minimize our objective contain ambiguity. If context leads to more efficient communication, then optimal languages should be more likely to be ambiguous as the amount of context increases.

Simulation set-up

We conduct $N = 2000$ simulations. For each simulation we enumerate the set of *valid* languages in which $|U| = |M| = 4$ (U is our set of utterances and M our set of meanings). Recall that languages are boolean matrices and a language $\ell \in L$ is "valid" so long as each possible meaning $m \in M$ can be referred to by at least one form $u \in U$ (every column of ℓ has some non-zero assignment) and each form maps to at least one meaning (every row has some non-zero assignment).



B

	m_1	m_2	m_3	m_4
u_1	0	0	**1**	0
u_2	0	**1**	0	0
u_3	**1**	0	0	**1**
u_4	**1**	0	0	0

C

	m_1	m_2	m_3	m_4
u_1	0	0	0	**1**
u_2	0	**1**	0	0
u_3	0	0	**1**	0
u_4	**1**	0	0	0

Figure 2: Panel (A) Vertical axis shows the proportion of optimal languages containing ambiguity. Horizontal axis shows the context-size (1-4) in each condition. Optimal language under CE objective (red). Speaker-optimal (blue). Listener-optimal (green). Error bars represent 95 percent confidence intervals. Panel (B), example CE-optimal language (ambiguous) from a four-context simulation. Panel (C), example CE-optimal language (unambiguous) from a single-context simulation.

For a given simulation, the goal is to find the language ℓ^* which minimizes our objective and then check to see if that language contains ambiguity.

We define language efficiency as a function of the particular semantic mappings induced by that language, the speaker and listener agents (S and L), as well as the utterance ($P(U)$), meaning ($P(M)$), and context priors ($P(C)$). Rather than assume particular structure, for each simulation we generate $P(U) \sim \text{Dir}(1,|U|)$, $P(M|C=c) \sim \text{Dir}(1,|M|)$ (a separate conditional distribution over meanings for each context c), and $P(C) \sim \text{Dir}(1,|C|)$, where $\text{Dir}(1,k)$ specifies the uniform Dirichlet distribution over a k-dimensional probability vector.

Context We want to assess the impact of *context* on the presence of ambiguity in optimal languages. To do so we consider four conditions with $n = 500$ simulations each (that is, 500 unique sets of $\{P(U),P(M|C),P(C)\}$. Our first is a *one-context* condition ($|C| = 1$) – only a single distribution over meanings $P(M)$. In our *two-context* condition ($|C| = 2$), we consider efficiency under both $P(M|C = c_1)$ as well as $P(M|C = c_2)$. *Three-* and *four-context* conditions corresponding accordingly.

Baselines For comparison, we examine properties of optimal languages under two additional objectives. Zipf (1949) proposed that the speaker-optimal language $\ell^*_{speaker}$ would minimize speaker effort and the listener-optimal language $\ell^*_{listener}$ would minimize listener effort. We define these objectives using the first and second half of equation 1 (see SI 2.2.).

Results and Discussion

In Simulation 1 we explored the degree to which ambiguity is an efficient property of languages when communication is contextualized. Figure 2, panel (A) plots the proportion of optimal languages under each objective as a function of number of contexts. The red line shows that as the number of contexts increases, so does the probability that an optimal language ℓ^*_{cross} contains ambiguity (at least one utterance maps to two meanings) under our CE objective. For comparison we also plot the proportion of speaker-optimal $\ell^*_{speaker}$ (blue line) and listener-optimal $\ell^*_{listener}$ (green line) languages that contain ambiguity. In line with Zipf's predictions, if languages are designed only to minimize speaker effort then optimal languages always contain ambiguity. If languages are designed to minimize listener effort then ambiguity is always avoided.

While our results indicate that ambiguity is an efficient property of contextualized language use, these simulations assumed that agents had perfect knowledge of the relevant conditional distributions ($P(M|C)$). This assumption may be too strong for describing much of day-to-day communication – we seldom interact with others with perfect knowledge of the current context (or topic) at the start of a conversation. To explore how ambiguity may be *used* efficiently in our framework, we next examine a case in which the listener has imperfect knowledge of context at the start of the conversation, but may infer it from the discourse history.

Simulation 2: Rational, pragmatic speakers use ambiguity efficiently

In Simulation 1 we showed that efficiency defined in terms of pragmatic agents leads to a preference for languages that contain ambiguity. In Simulation 2 we assume a single fixed language ℓ, which contains ambiguity, and instead vary the types of agents using ℓ. We will show that efficient *use* of ambiguity depends on an agent's ability to use context for disambiguation. More generally, Simulation 2 is intended to

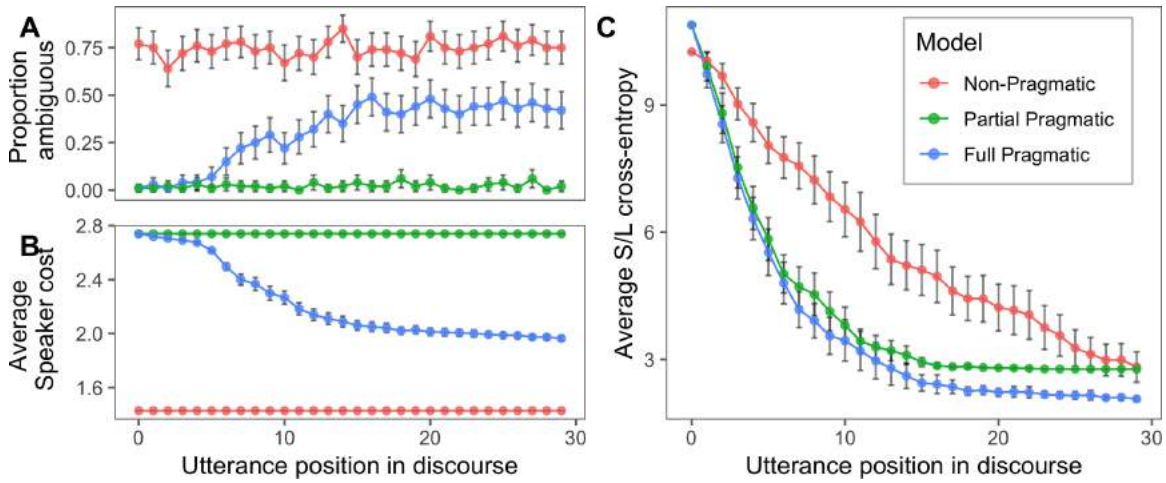


Figure 3: (A) shows the empirical probability that our speaker used an ambiguous utterance as a function of discourse position. (B) shows speaker effort across the three models. (C) shows the Cross-Entropy objective under our three speaker models. Error bars represent 95 percent confidence intervals.

demonstrate how we can assess both questions of efficient use (current simulation) as well as design (previous simulation) in the same framework.

Imagine a scenario in which a reader is beginning a news article. While they may have some knowledge about the article's topic (perhaps from the title), they may not have complete knowledge of its contents, including the persons or events involved. In this setting, using a low-cost, but ambiguous referring expression (say a pronoun like "he") early may lead to misunderstanding if context is not informative. But, if by a later position enough contextual information has accumulated, it may be efficient to use the ambiguous expression. We pursue this general framework in Simulation 2 – examining when in a discourse using ambiguity is efficient. We will consider "context" as analogous to a "topic" of conversation.

Simulation set-up

We consider a single language ℓ, which contains both ambiguous and unambiguous utterances. We assume ambiguous utterances are lower cost. Crucially, we do not assume that the listener knows the particular topic ($c_{current}$) of the conversation *a priori*. Rather, that the listener has knowledge of the set of possible topics $C = \{c_1, \ldots, c_k\}$, but does not know which one is currently being used by the speaker. Formally, this means the listener does not have access to the correct conditional distribution over meanings $P(M|C = c_{current})$ at the start of the discourse.

Over the course of a discourse D, the listener tries to infer both the current topic, $c_{current}$, as well as the particular meaning m of a given utterance u. That is, we consider agents who can track the history of previous utterances D. Importantly, an agent can attempt to infer the current topic of conversation $c_{current}$ using the discourse history D.

We conduct $N = 600$ simulations, generating discourses of length $|D| = 30$ utterances, comparing three speaker models ($n = 200$ each). We consider a single language ℓ[5] with $|U| = 6$ and $|M| = 4$ in which two of the utterances are ambiguous and lower cost than the unambiguous utterances. (Note that use of this particular language is not essential – the results are broadly generalizable to languages that contain ambiguity, but exploring this space is computationally expensive.)

Speaker agents

We vary the degree to which agents can use context for disambiguation. We consider three types of speaker models. Our *Full pragmatics* agent, models a speaker who reasons about her listener and also has complete recall of the set of utterances in the discourse D. This speaker believes that the listener may not know the current topic $c_{current}$ at the start of the discourse, but can infer it over the discourse. We compare two baseline models. The first, a *Partial pragmatics* baseline describes a speaker who reasons about a listener, but assumes they have no access to the discourse history. The second, a *No pragmatics* baseline speaker does not consider a listener at all, but produces utterances according to the underlying language semantics (ℓ) and topic probabilities ($p(M|C = c_{current})$ (see SI 3).

Hypotheses

We are interested in how each speaker-model uses ambiguity over the discourse. A speaker strategy that is mutually efficient for both agents should avoid ambiguity until sufficient contextual information has accumulated. We should expect this to be reflected in our *Full pragmatics* model who reasons about the listener and discourse history. By contrast, a speaker-optimal model who does not consider the listener should greedily use ambiguous utterances (*No pragmatics* model), while a listener-optimal model should avoid ambiguity entirely (*Partial pragmatics* model).

[5] See SI for the matrix notation of this language.

Results and Discussion

Figure 3, panel (A) shows the empirical probability a speaker uses an ambiguous utterance as a function of discourse position. The *No pragmatics* baseline uses ambiguous utterances frequently and at a constant rate over the discourse. The *Partial pragmatics* baseline avoids ambiguous utterances entirely. But, the *Full pragmatics* model avoids ambiguous material only at the start of the discourse, employing it increasingly as the discourse proceeds. Panel (C) tracks our CE objective for each model over the discourse. Note that the objective decreases for all three models, primarily driven by the listeners updating their beliefs about the actual topic ($P(C = c_{current}|D)$). However, the objective declines more quickly under the *Full* and *Partial pragmatics* speakers as listener agents are better able to infer the correct context. Additionally, the difference in CE between the *Full-* and *Partial pragmatics* models at the end of the discourse is driven by the reduction in speaker costs. Panel (B) tracks speaker effort, which remains constant in both *No pragmatics* and *Partial pragmatics* baselines. But, effort declines in the *Full pragmatics* model as speakers increasingly rely on ambiguous material later in the discourse.

General Discussion

How do the competing pressures imposed by speakers and listeners give rise to the distributional regularities found in natural language? Zipf (1949) proposed that the asymmetry between speaker and listener costs gives rise to a range of properties at the level of the lexicon. We explored the interactions of rational pragmatic agents as a framework for understanding efficient language structure and use. We focused on an argument on the communicative function of ambiguity (Piantadosi et al., 2011), deriving a novel speaker-listener Cross-Entropy objective for measuring the efficiency of linguistic systems from first principles of efficient language use. In Simulation 1 we showed that optimal languages are more likely to contain ambiguous material when context is informative. In Simulation 2 we showed how rational pragmatic agents use ambiguous material efficiently in conversation.

A limitation of the current work is an analysis of exactly how the CE objective compares to existing measures. For example, previous work has described competing speaker-listener pressures in terms of a trade-off of simplicity and informativeness (Kemp & Regier, 2012) or expressivity and compressibility (Smith, Tamariz, & Kirby, 2013) to explain linguistic regularities. Future work should assess the degree to which we can derive the same properties as previous studies using our current framework. More generally, we hope that this framework can serve as a domain general tool to assess the range of functionalist theories examining efficient language-structure and use.

SI and simulations: `https://bit.ly/2RBSGcU`, `https://github.com/benpeloquin7/zipf_principles`

References

Bennett, E., & Goodman, N. (2018). Extremely costly intensifiers are stronger than quite costly ones. *Cognition*.

Chomsky, N. (2002). An interview on minimilism. In *N chomsky, on nature and language*.

Ferrer-i-Cancho, R. (2018). Optimization models of natural communication. *Journal of Quantitative Linguistics*, *25 (3)*, 207–237.

Frank, M., & Goodman, N. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*, 998.

Genzel, D., & Charniak, E. (2002). Entropy rate constancy in text. In *Proceedings of the 40th annual meeting on association for computational linguistics*.

Goodman, N., & Frank, M. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, *20(11)*, 818–829.

Grice, P. H. (1975). Logic and conversation.

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. In *Proceedings of the naacl*. 159-166.

Hawkins, R., Franke, M., Smith, K., & Goodman, N. (2018). Emerging abstractions: Lexical conventions are shaped by communicative context. In *Proceedings of the 40th annual conference of the cognitive science society*.

Kemp, C., & Regier, T. (2012). Kinship categories across languages reflect general communicative principles. *Science*, *336*, 1049–1054.

Kirby, S., Griffiths, T., & Smith, K. (2014). Iterated learnign and the evoluation of language. *Current Opinion in Neurobiology*, *28*, 108–114.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106(3)*, 1126–1177.

Levy, R., & Jaeger, T. (2007). Speakers optimize information density through syntactic reduction. In *Proceedings of the twentieth annual conference on neural information processing systems*.

Piantadosi, S., Tily, H., & Gibson, E. (2011). The communicative function of ambiguity in language. *Cognition*, *122*, 280–291.

Regier, T., Kemp, C., & Kay, P. (2015). Word meanings across languages support efficient communication. In B. M. & W. O'Grady (Ed.), *The handbook of language emergence (pp. 237-263)*. Hoboken, NJ: Wiley-Blackwell.

Smith, K., Tamariz, M., & Kirby, S. (2013). Linguistic structure is an evolutionary trade-off between simplicity and expressivity. In *Proceedings of the 36th annual conference of the cognitive science society*.

Wasow, T., Perfors, A., & Beaver, D. (2005). The puzzle of ambiguity. In CSLI (Ed.), *Morphology and the web of grammar: Essays in memory of steven g lapointe*. Stanford, CA: McGraw-Hill.

Zipf, G. (1949). *Human behavior and the principle of least effort*. New York, NY: Prentice-Hall.

Why do echo chambers form? The role of trust, population heterogeneity, and objective truth

Amy Perfors (amy.perfors@unimelb.edu.au)
School of Psychological Sciences, University of Melbourne
Danielle J. Navarro (d.navarro@unsw.edu.au)
School of Psychology, University of New South Wales

Abstract

Many real-world situations involve learning entirely or mostly based on the information provided by other people, which creates a thorny epistemological problem: how does one determine which of those people to trust? Previous work has shown that even populations of rational Bayesian agents, faced with this problem, polarise into "echo chambers" characterised by different beliefs and low levels of between-group trust. In this study we show that this general result holds even when the reasoners have a more complex meaning space and can communicate about their beliefs in a more nuanced way. However, even a tiny amount of exposure to a mutually trusted "ground truth" is sufficient to eliminate polarisation. Societal and psychological implications are discussed.

Keywords: Bayesian reasoning; echo chambers; polarisation; social inference; trust; epistemology

Introduction

The real world is full of situations where the vast majority of what we learn comes from other people. In some, like language learning, the "ground truth" of the matter simply *is* whatever people agree that it is. However, many other situations pose a much more challenging epistemological problem: the ground truth is (at least mostly) inaccessible, and the only way to learn about it is to rely on other people. Regardless of why the truth is often inaccessible – due to spatial or temporal distance, or difficulties in interpreting ambiguous data – people are often faced with questions of this character. Did humans evolve or were we created by a superior being? Did Trump assist the Russians to influence the US 2016 elections? Did Bob have an affair with Mindy? In all of these cases, there *is* a truth of the matter, but it is not a truth that is directly accessible to most people. All of the data is mediated through other agents – scientists studying evolution, politicians receiving confidential documents, journalists deciding what to report on, Bob and Mindy – and few have the access or training necessary to make sense of the data on their own.

What is a rational learner to do in this difficult epistemological situation? One option would be to simply try to communicate fully with everybody and update one's beliefs accordingly. When this happens, groups of Bayesian learners will converge to a shared belief system equivalent to the population prior, at least when organised as chains (Griffiths & Kalish, 2007) or fully interconnected (Whalen & Griffiths, 2017). When data are additionally generated from an external ground truth, the convergent distribution is also shaped by that world (Perfors & Navarro, 2014). However, these results only hold when agents cannot select who to talk to and when all share the same prior. When people have heterogeneous priors, the beliefs of the population are systematically distorted towards the beliefs of the most extreme individuals (Navarro, Perfors, Kary, Brown, & Donkin, 2018).

This amplification of extreme priors is concerning because it suggests that the process of information transmission itself can distort belief – and that this occurs even if all agents are fully rational and can share information fully. But our situation in real world is even more difficult. Limited by temporal and cognitive constraints, people cannot exchange information with everyone else. Moreover, the real world includes people who you might not want to learn from – not just because they have different or more extreme priors, but because they might be completely wrong or actively deceptive.

Intuitively, one solution to this dilemma would be for agents to learn who *not* to trust: to lower the weight given to the data from people who are inaccurate or miscalibrated. This is an appealing idea, but raises an important question: in the absence of any direct access to the ground truth, how should a rational learner determine who is to be trusted? One possibility is that agents might favour those who seem to make sense: those who make claims that are consistent with one's own beliefs. Indeed, there is evidence that people do adopt this strategy (Collins, Hahn, & von Gerber, 2018). Unfortunately, trusting people with similar beliefs more often leads to polarisation (e.g., Axelrod, 1997; Hegselmann & Krause, 2002; Olsson, 2013; Ngampruetikorn & Stephens, 2016; O'Connor & Weatherall, 2018; Madsen, Bailey, & Pilditch, 2018). Instead of converging on a shared set of beliefs, populations split into echo chambers: sub-groups characterised by high trust and shared beliefs within groups, but low trust and shared beliefs between groups.

Although this general result is robust and has been shown in a variety of modelling paradigms, in many cases the reasoners in such paradigms are not meant to be optimal (e.g., Axelrod, 1997; Hegselmann & Krause, 2002; Ngampruetikorn & Stephens, 2016). Some studies that *do* use Bayesian agents have established that polarisation arises even when all of the agents reason rationally (Olsson, 2013; O'Connor & Weatherall, 2018; Madsen et al., 2018); however, these studies generally involve fairly impoverished one-dimensional meaning spaces and agents who can only communicate about those spaces in a limited way. For instance, the agents in Olsson (2013) may believe in a proposition to only some degree (e.g., 70%) but are only capable of communicating binary ("yes" or "no") beliefs about the proposition.

The agents in Madsen et al. (2018) are permitted more nuance, being able to communicate their beliefs about the mean of a one-dimensional Gaussian, but have no way to communicate their level of certainty. Would polarisation still arise in groups of Bayesian agents with a richer space belief and the ability to communicate those beliefs in a more nuanced way? We explore this question here.

In Study 1 we present a new modelling paradigm in which agents must learn and communicate about a two-dimensional meaning space by sampling items from their current beliefs, while simultaneously making inferences about which of the other agents are trustworthy. We show that, as long as the distribution of prior beliefs in the population is sufficiently heterogeneous, echo chambers form even in this circumstance. Study 2 investigates whether polarisation can be eliminated and trust built by selectively communicating about only some topics (dimensions). We find that this is not a solution: doing so does build trust but at the cost of never coming into agreement. In Study 3 we explore another potential solution: access to a mutually trusted ground truth. Reassuringly, when agents have access to such a truth – even if it makes up only a tiny fraction of all of the data – polarisation is eliminated.

Study 1: Baseline

Method

Our simulations involve populations of n optimal Bayesian agents who each learn a hypothesis by receiving data from other agents (we vary $n = 6$ or $n = 18$). Agents perform inference over which other agents are trustworthy t at the same time as inferring which hypothesis h best describes the data x seen so far by calculating the joint posterior $P(t,h|x)$. Performing joint inference over trust and beliefs is somewhat different from the typical approach, in which agents directly prefer others who have similar beliefs (Olsson, 2013; Madsen et al., 2018; O'Connor & Weatherall, 2018). We opted for this approach for two reasons. First, people appear to make inferences about trust at the same time that they evaluate beliefs, and use their perceptions of trust to decide whose data to rely on (Petty & Briñol, 2008; Shafto, Eaves, Navarro, & Perfors, 2012; Perfors, Navarro, & Shafto, 2018). More importantly for our purposes, explicitly differentiating inferences about trust from beliefs allows us to explore what happens if agents can change their communication style (but not their beliefs) in order to build trust, as in Study 2.

Trust is a real value between 0.0 (no trust) to 1.0 (perfect trust) while beliefs consist of 2D Gaussians parameterised by an unknown mean μ and a known symmetric covariance Σ_0, as described in more detail below.

Initialisation. Each agent a is initialised with a different prior belief about the mean $\mu_a \sim N(0,\Sigma)$, where $\Sigma = 0.5\mathbf{I}$. All agents share the same prior about the covariance Σ_0. We manipulate population heterogeneity by changing the size of the prior covariance Σ_0 relative to the initial generating covariance Σ. Populations with high heterogeneity are initialised with means that are more "distant" in belief space relative to their beliefs about how wide the category is. There are three conditions, each defined by their covariance matrix Σ_0: HOMOGENEOUS ($\Sigma_0 = 0.25\mathbf{I}$), NEUTRAL ($\Sigma_0 = 0.15\mathbf{I}$), and HETEROGENEOUS $\Sigma_0 = 0.05\mathbf{I}$).

It would have been mathematically equivalent to manipulate heterogeneity by keeping the agents' covariance priors Σ_0 constant and varying the covariance of the generating distribution Σ; the important thing is the ratio of the two. (We chose to do it this way because one of our dependent variables is the average distance between agents in belief space, and this permits all conditions to be initialised with a similar average distance.) Smaller initial covariance matrices imply more heterogeneity because heterogeneous populations contain more individuals who are more likely to initially disagree (by inferring that the data provided by the other was unlikely). The same intuition is captured in other paradigms via the tendency to seek out those who are distant in belief space; agents with less of this tendency are more likely to polarise (Olsson, 2013; O'Connor & Weatherall, 2018; Madsen et al., 2018).

Agents are also initialised with trust vectors t with one cell for each other agent in the population, such that $t \sim \text{Beta}(1,1)$. This prior means that each agent may initially trust any other to any degree. Because the prior is weak, it is easily changed in response to data.

Iterations. During each iteration we loop through our population of n agents. At each iteration, agent i selects another agent j to learn from, proportional to the relative degree of trust i has in j. Upon being selected, agent j samples a single data point x at random from their hypothesis such that $x \sim N(\mu_j,\Sigma_0)$. Agent i then then updates their beliefs about $\mu_\mathbf{i}$ in the direction of x.[1] Thus, each iteration involves agents learning from others, in all cases revising their beliefs in the direction of the data provided, but weighting the data that was provided by trusted agents more.

At each iteration each agent i also updates their trust in all other agents j, based on the data $\mathbf{X_j}$ provided by each. The intuition is that agents will infer trustworthiness based on the extent that the other says sensible things: in this context, that means that agent j will be trusted proportional to the degree to which the data they provide to i is consistent with i's own beliefs. Agent i accomplishes this by computing the probability that they themselves would have generated that data $P(\mathbf{X}_j|N(\mu_i,\Sigma_0))$ and comparing it to the probability that it was generated by an uninformative and unhelpful other $P(\mathbf{X}_j|N(0,\Sigma_u))$.[2] Agents are thus more likely to trust those who provide data that is consistent with their own beliefs.

[1] Technically, agent i performs $n-1$ Metropolis-Hastings steps, one for each of the other agents j, in which the likelihood is calculated for all of the data points $\mathbf{X}_j$ shared by j, including the new data point x. Likelihood is weighted by trust in that agent, so that agents who are more trusted have more of an affect on belief revision.

[2] The reason for comparing against a baseline is that the raw probability of an agent providing any set of datapoints is low in absolute terms, and without the comparison all simulations tend for all agents to trust nobody. Results are qualitatively similar for a wide range of choices for the covariance of the uninformative baseline, as long as it is larger than Σ_0. All simulations here set $\Sigma_u = \mathbf{I}$.

Our approach is most similar to that of Madsen et al. (2018), but there are a few key differences in addition to those already discussed. First, their agents make inferences about both mean *and* variance, and communicate by providing the mean directly rather than sampling from their posterior. Polarisation occurs in their simulations at least in part because the learned variances approach size zero. This was probably facilitated by the fact that agents could not sample from their distributions when providing data and thus could only give point estimates, leading to a severe underestimation of the variance. Here we test whether polarisation still emerges even with agents with constant variance who can also provide more information about the extent of their distribution.

A second difference is that their agents can revise their beliefs *away* from the data they receive, whereas ours cannot. This sort of belief revision is not necessarily irrational (Jern, Chang, & Kemp, 2014), but it is difficult to determine to what extent it drives polarisation in Madsen et al. (2018). In order to explore whether polarisation arises even when the conditions for it are as unfavourable as possible, our agents disregard data they do not trust rather than move away from it.

Results

For each condition and population size, we ran 50 runs (differing only in the initial random distribution of agents in belief space) for 500 iterations each. All of our simulations were characterised by changes in the beliefs of the agents as well as their mutual trust. We consider each in turn.

Trust. We can visualise the distribution of trust across the population using pairwise mutual trust matrices T in which T_{ij} denotes the trust that agent i has toward j. We are specifically interested in the distribution of trust within the population: does it tend to be uniform, or are there clusters of agents who highly trust in each other but distrust anyone else? As the top right panel of Figure 1 shows, this clustering can be quantified using Gini mean difference (GiniMD): the mean absolute difference between all distinct elements in the pairwise trust matrices. A lower GiniMD indicates a higher shared trust, and GiniMD values over 0.3 correspond to highly polarised populations: the pairwise trust matrices show a "block" structure in which agents are in subgroups characterised by high within-group trust and low between-group trust.

As the top of Figure 1 shows, regardless of the population size, populations with HETEROGENEOUS agents were highly likely to become polarised. An ANOVA found a significant effect of condition on GiniMD ($F(2,296) = 29.34, p < 0.0001$) but not number of agents ($F(1,296) = 0.81, p = 0.369$). Initial random differences in beliefs between agents were exacerbated as they grew to trust those with similar beliefs and minimised data from those with dissimilar beliefs. Heterogeneity was the determining factor because it affected how much weight agent i put on data from j. In heterogeneous populations, more agents had initial beliefs that were far from the covariance of other agents; they were thus more apt to be distrusted. Once distrusted, they could not recover.

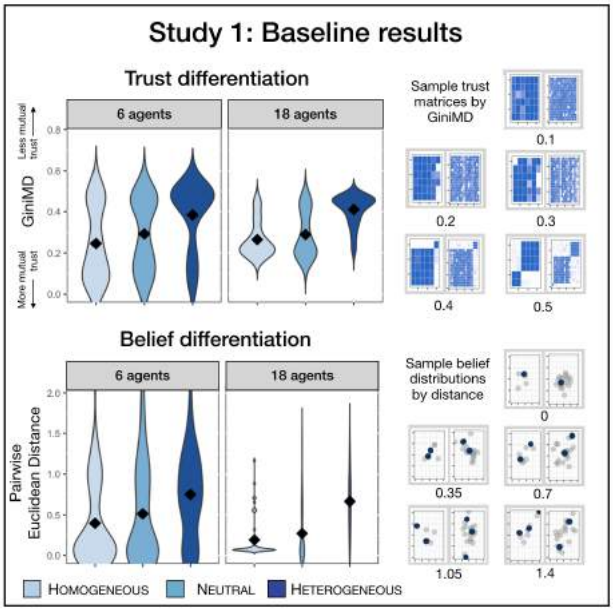


Figure 1: *Study 1: Emergence of polarisation.* Top: Polarisation is evident in the pairwise mutual trust matrices between agents, and quantified using GiniMD (right). Values above 0.3 indicate that agents have formed subgroups characterised by high within-group trust and low between-group trust. Populations of all size become polarised when they are HETEROGENEOUS (left), despite the fact that all agents are optimal Bayesian reasoners. Bottom: More HETEROGENEOUS agents also show a greater divergence in beliefs (left). Sample runs (right) showing the average pairwise Euclidean distance between agents in belief space (grey dots plot the locations of agents' initial hypotheses (μ) and dark blue dots plot the final ones) reveal that larger differences tend to correspond to more than one cluster in belief space.

The bottom of Figure 1 illustrates that these trust-based echo chambers correspond to greater average distance from each other in belief space; agents do not converge on a shared belief. As before, this effect was driven by population heterogeneity ($F(2,296) = 22.11, p < 0.0001$), although population size was also significant ($F(1,296) = 11.24, p = 0.001$). Even though agents in all conditions began the simulations at similar distances in belief space from each other, the HETEROGENEOUS agents tended to form widely-separated clusters while more HOMOGENEOUS agents were more likely to converge on the same belief. Distance in belief space and trust clustering thus both tell the same story: in sufficiently heterogeneous populations, polarisation is highly likely, even when all of the agents involved are optimal Bayesian reasoners. Consistent with this, there is a strong correlation between GiniMD and distance ($r = 0.81, t(298) = 23.7, p < 0.0001$).

How might we disrupt this tendency toward polarisation? Study 2 explores one idea: building trust by communicating tactically. Our agents are always constrained to be honest, but here we make it possible for them to refrain from communicating about topics on which disagreement is likely.

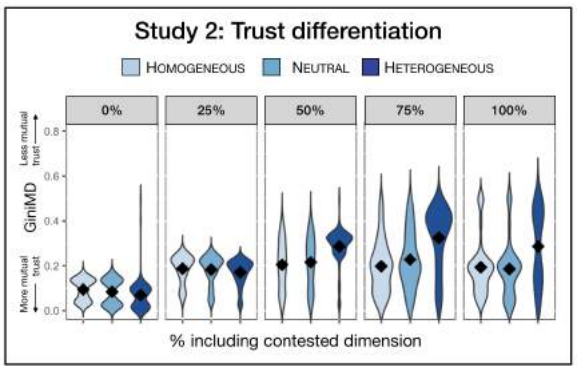


Figure 2: *Emergence of trust when agents can avoid contested subjects.* Average GiniMD as a function of the proportion of time agents included information about the contested dimension. As the dimension is included less, the agents show ever-higher levels of mutual trust. Trust is consistently unpolarised by the time the contested dimension is included 25% of the time, even in the HETEROGENEOUS condition.

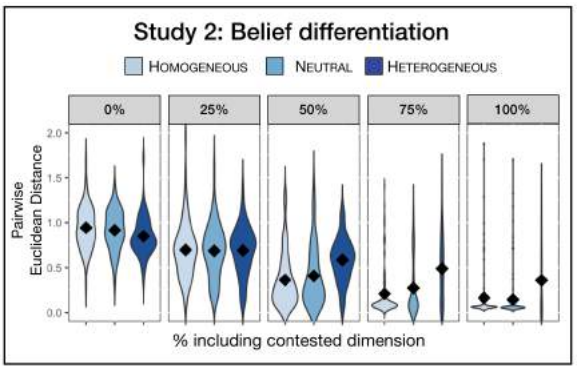


Figure 3: *Evolution of belief when agents can avoid contested subjects.* Average pairwise distance in belief space as a function of the proportion of time agents included information about the contested dimension. As the dimension is included less, the agents show more divergence in beliefs; as trust increases, the divergence in beliefs increases more. Thus, lower polarisation does not reflect more agreement.

Study 2: Tactical topic selection

Method

One of the simplifications we made in Study 1 was to assume that agents were required to communicate fully as well as honestly. In real life, however, people have discretion in what they choose to talk about. If you are visiting an uncle with whom you disagree politically, you might spend the majority of your time talking about something that you agree on, like football. This enables you to grow trust in each other and might give you the space to occasionally talk about politics.

Does adopting this strategy decrease the emergence of polarisation? Key to answering this question is realising that it is important to talk about contested issues at least some of the time: otherwise, you might trust each other, but still have irreconcileable beliefs about the facts of the matter. In these simulations we test whether there are any "sweet spots" in which agents can talk about contested beliefs just enough to come to agreement and maintain trust.

We tested this by initialising the agents differently. Where before the initial means for agents $\mu_{\mathbf{a}}$ were generated by sampling from a Gaussian with symmetric covariance matrix $0.5\mathbf{I}$, in Study 2 we sampled them from an asymmetric matrix with the same covariance as before along one dimension but four times tighter along the other. This meant that agents *a priori* only disagreed on one dimension, rather than two.

We then systematically varied the proportion of time that agents chose to include the contested dimension that they were more likely to disagree on. If an agent received a data point that did not include that dimension, they "filled it in" themselves by sampling it from their own prior. This was done in order to maximise the probability of eliminating polarisation; if it cannot be avoided even when agents are making the most charitable assumptions about what is going unstated, then it would be even harder to avoid if agents are making less charitable assumptions.

Results

The results suggest that enabling agents to only discuss one dimension and avoid contested dimensions *does* increase mutual trust, but the price of this is that agents no longer form a shared set of beliefs. As Figure 2 shows, communicating less about the contested dimension systematically increases trust ($F(4, 1495) = 117.5, p < 0.0001$). If the contested dimension is included only half of the time, GiniMD values are consistently below 0.4, and if it is included 25% of the time or less the level of polarisation is nearly nonexistent.[3]

However, as Figure 3 reveals, that lack of polarisation corresponds to situations where the average distance between beliefs has increased substantially ($F(4, 1495) = 137.7, p < 0.0001$). When the contested dimension is included half of the time, the average distance between beliefs is even higher than in the baseline HETEROGENEOUS case, even though the trust levels are still low. By the time polarisation has been eliminated in the trust matrices (when talking about the contested dimension 25% of the time or less), agents radically differ in their beliefs. What appears to be happening is that, unaffected by external data, evolution along that dimension proceeds in a random walk. Thus, although agents agree with each other on the non-contested dimension, they diverge ever more strongly on the contested one.

Thus, the higher levels of trust have not bought more agreement: they just reflect the fact that some topics are not discussed. Most importantly, we could find no "sweet spots" in our simulations where strategically communicating about contested beliefs only part of the time could allow trust to be maintained *and* beliefs to converge. This finding should be interpreted with caution because it depends to some extent on choices we made about values of Σ_0, Σ_u, and Σ. However, it is not reassuring that the divergence in belief occurs *before*

[3]For ease of presentation, we collapse across population size in the figures and analyses but the qualitative effect is identical whether there are 6 or 18 agents in the population.

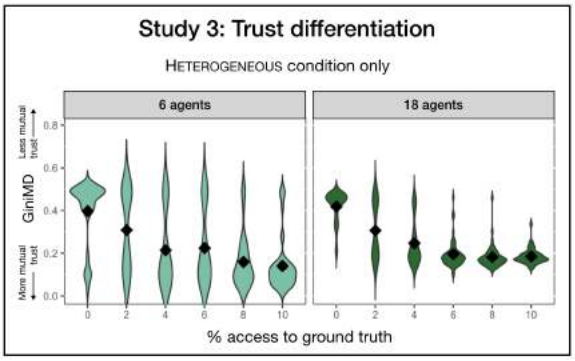


Figure 4: *Evolution of trust when agents have access to the ground truth.* It only takes a little bit of access to a ground truth source that everyone agrees is trustworthy to disrupt the formation of echo chambers, even in a HETEROGENEOUS population. Even 2% of all data points make a big difference, and by 4% or so everyone trusts everyone else.

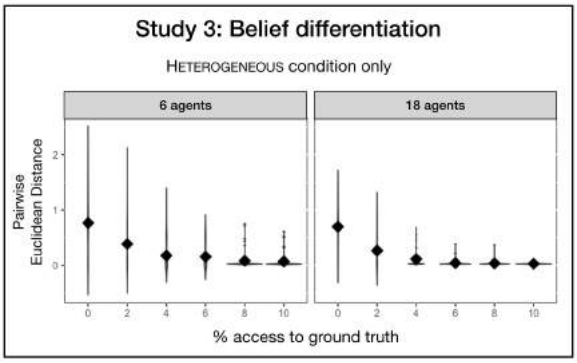


Figure 5: *Evolution of belief when agents have access to the ground truth.* It only takes a little bit of access to a ground truth source that everyone agrees is trustworthy to disrupt the formation of echo chambers, even in a HETEROGENEOUS population. Even 2% of all data points make a big difference, and by 4% or so there are no differences in beliefs.

the emergence of mutual trust, suggesting that even if such a sweet spot exists, it is tiny and highly dependent on a very specific set of parameter choices.

So far we have found that echo chambers persistently form in populations of rational agents, despite making as many charitable assumptions as possible: our agents do not revise beliefs away from those they disagree with and communicate about a rich meaning space in a way that includes their confidence (variance) about the mean rather than the mean alone. Even with these assumptions, as long as the initial beliefs are heterogeneous enough, agents cluster into echo chambers. Allowing them to build trust by communicating more often on less contentious topics does not solve this problem; communicating rarely enough to build trust means not communicating often enough to converge on a set of shared beliefs. Taken together, this appears to support the intuition we began with: this is a very difficult epistemological problem. How can one sensibly learn from others when you have no way to evaluate who to trust aside from the data they provide, and no way to evaluate that data against the state of the world?

These considerations suggest that echo chamber formation might be eliminated by simply giving agents access to some mutually-agreed upon ground truth of the matter. This might be data supplied by the external world directly or information provided by an objective observer; all that is necessary is that everyone has access to it and everyone trusts it. Does access to the ground truth disrupt the formation of echo chambers? If so, how little is required?

Earlier work has investigated these questions and found that access to the ground truth is not sufficient to disrupt echo chamber formation (O'Connor & Weatherall, 2018; Madsen et al., 2018). However, in O'Connor and Weatherall (2018) the agents sought out such evidence in a confirmatory way, testing their current hypothesis only. It is possible that receiving data relevant to all hypotheses might have led to a different result. Furthermore, agents in Madsen et al. (2018) often ended up ignoring the ground truth because it was outside of their inferred variance, which had shrunk to zero. In that sense it was not actually a "ground truth", because although it was available to all, very few people trusted it. In Study 3 we therefore provide a ground truth that all agents have access to and all trust equally.

Study 3: Ground truth

Method

Our method was exactly the same as in Study 1, except that sometimes the agents received a data point x_g sampled from the "ground truth" of the world, $x_g \sim N(0, \Sigma)$. Agents revised their belief based on this data exactly as they did on any other data; the only difference is that they did not perform inference over trust, instead assuming perfect trust in the source. We systematically vary how often agents have access to the ground truth. Because echo chambers only emerged in the HETEROGENEOUS condition in Study 1, we consider only that condition here. As in Study 2, for ease of presentation we combine the runs with 6 and 18 agents.

Results

As Figures 4 and 5 show, even a very small amount of access to ground truth data is sufficient to disrupt the formation of echo chambers. When only 2% of the data comes from the ground truth, a substantial proportion of runs result in high levels of mutual trust and shared beliefs. When 4% of the data is ground truth, polarisation is consistently eliminated: even initially HETEROGENEOUS agents converge on the same set of shared beliefs and trust everybody in the population.

Discussion

This paper is part of a growing literature investigating what happens to populations of rational agents when faced with a difficult epistemological puzzle: how to learn a set of beliefs from other people, without having access to external evidence about those beliefs or knowing *a priori* who to trust.

Consistent with that literature, we find that echo chambers consistently emerge, despite making every effort we could to eliminate them. Even though we provided agents with a richer meaning space and more nuanced communication abilities than other studies, polarisation was still highly likely as long as the population was sufficiently heterogeneous in their initial beliefs. One contribution of our work, therefore, is to further underline the robustness of this effect.

We make several larger contributions as well. First, we show that enabling agents to strategically talk less about topics that they disagree on did not solve the problem. Avoiding those topics did lead to improve trust, but at the expense of *increasing* the distance between beliefs; we found no "sweet spot" where both mutual trust and shared belief were possible. To our knowledge this is the first attempt to simulate the population-level effects that results from agents adopting different communicative tactics. Our framework is rich enough to investigate many other such tactics. What happens if people sample based not just on their own beliefs, but also on their inferences about the beliefs of others? What if people deliberately select more or less extreme beliefs, in an effort to shift the Overton window of acceptable discourse? How vulnerable are these strategies to deceptive or malicious agents?

Our work is also the first, to our knowledge, to show that having access to a trusted "ground truth" is an extremely powerful way to break the echo chamber effect. Previous work found that ground truth did not help that much (Madsen et al., 2018; O'Connor & Weatherall, 2018), but as discussed before, this was probably because of specific modelling choices that resulted in their "ground truth" being neither fully shared nor fully trusted. When it *is* shared and trusted, only a small proportion of data is necessary for even initially heterogeneous populations to develop high trust and converge on shared beliefs. The reason for this is that this common ground breaks the vicious cycle and creates a virtuous one: agents make inferences about their beliefs based in part on the ground truth data, thus trusting agents more who agree with it, and so forth. Our framework is flexible enough to enable further exploration of the robustness of this effect. How important is it that *everyone* have access to it? What if the ground truth is more accessible or less ambiguous to some? Is there any way for agents to identify those people that cannot be "gamed" by malicious agents seeking to mislead?

Our finding about the necessity of the ground truth may have important implications in light of the "post-truth" era that many believe we are now in (Lewandowsky, Ecker, & Cook, 2017). This era is characterised not only by attempts to delegitimise previously trusted sources but, more profoundly, a pervasive denial that a truth exists at all and a persistent belief that no sources are to be trusted (McCright & Dunlap, 2017). Indeed, one of the characteristics of fascism was a denial of the utility of external evidence (Varshizky, 2012), and conspiracy theories are associated with lower levels of trust in external sources (Einstein & Glick, 2015). Our simulations suggest why: shared access to the truth is one of the few things that might rescue agents from an otherwise inescapable epistemic trap. Agents who do not have access or belief in this truth are far easier to confuse, polarise, and manipulate.

Although our work further demonstrates that echo chamber formation is a robust and consistent effect even in populations of perfectly rational learners, it does suggest a key to disrupting them. Perhaps polarisation can be minimised and trust increased not by throwing more evidence toward mistaken beliefs, but by working to persuade people instead that objective truth exists and shoring up their (perceived) capacity to access and evaluate it.

Acknowledgments

This work was supported by the Australian Defence Science Technology Group Strategic Research Initiative scheme. Thanks to Piers Howe, Yoshi Kashima, Nic Fay, Charles Kemp, Alexei Filinkov, and Lucia Falzon for helpful conversations.

References

Axelrod, R. (1997). The dissemination of culture: A model with local convergence and global polarization. *Journal of Conflict Resolution*, *41*(2), 203–226.

Collins, P., Hahn, U., & von Gerber, Y. (2018). The bi-directional relationship between source characteristics and message content. *Frontiers in Psychology*, *9*.

Einstein, K., & Glick, D. (2015). Do I think BLS data are BS? The consequences of conspiracy theories. *Pol. Beh.*, *37*, 679–701.

Griffiths, T., & Kalish, M. (2007). Language evolution by iterated learning with Bayesian agents. *Cogn. Sci.*, *31*(3), 441–480.

Hegselmann, R., & Krause, U. (2002). Opinion dynamics and bounded confidence models, analysis, and simulation. *Jn Artificial Societies and Social Simulation*, *5*(3).

Jern, A., Chang, K., & Kemp, C. (2014). Belief polarization is not always irrational. *Psychological Review*, *121*(2).

Lewandowsky, S., Ecker, U., & Cook, J. (2017). Beyond misinformation: Understanding and coping with the "post-truth" era. *Jn Applied Res. Mem. Cogn.*, *6*, 353–369.

Madsen, J., Bailey, R., & Pilditch, T. (2018). Large networks of rational agents form persistent echo chambers. *Sci. Reports*, *8*.

McCright, A., & Dunlap, R. (2017). Combatting misinformation requires recognizing its types and the factors that facilitate its spread and resonance. *Jn Applied Res. Mem. Cogn.*, *6*, 389–396.

Navarro, D. J., Perfors, A., Kary, A., Brown, S., & Donkin, C. (2018). When extremists win: Cultural transmission via iterated learning when populations are heterogeneous. *Cognitive Science*, *42*(7), 2108–2149.

Ngampruetikorn, V., & Stephens, G. (2016). Bias, belief, and consensus: Collective opinion formation on fluctuating networks. *Physical Review E*, *94*(5).

O'Connor, C., & Weatherall, J. (2018). Scientific polarization. *European Jn for Phil Science*, *8*, 855–875.

Olsson, E. (2013). A Bayesian simulation model of group deliberation and polarization. *Bayesian argumentation*, 113–133.

Perfors, A., & Navarro, D. J. (2014). Language evolution can be shaped by the structure of the world. *Cogn. Sci*, *38*(4), 775–793.

Perfors, A., Navarro, D. J., & Shafto, P. (2018). Stronger evidence isn't always better: A role for social inference in evidence selection and interpretation. In C. Kalish, M. Rau, J. Zhu, & T. Rogers (Eds.), *40th Conf. Cognitive Science Soc.*

Petty, R., & Briñol, P. (2008). Persuasion: from single to multiple to metacognitive processes. *Persp. Psychol. Sci.*, *3*, 137–147.

Shafto, P., Eaves, B., Navarro, D. J., & Perfors, A. (2012). Epistemic trust: Modeling children's reasoning about others' knowledge and intent. *Developmental Science*, *15*, 436–447.

Varshizky, A. (2012). Alfred rosenberg: The nazi weltanschauung as modern gnosis. *Politics, Religion, and Ideology*, *13*, 311–331.

Whalen, D., & Griffiths, T. (2017). Adding population structure to models of language evolution by iterated learning. *Jn of Mathematical Psychology*, *76*, 1–6.

Benefits of active control of study in autistic children

Nicholas Perri (perri@mpib-berlin.mpg.de)
MPRG iSearch, Max Planck Institute for Human Development - Lentzeallee 94, 14195 Berlin, Germany
Mind and Brain, Humboldt-Universität zu Berlin - Unter den Linden 6, 10099 Berlin, Germany

Valentina Fantasia (valentina.fantasia86@gmail.com)
LInC - Interaction & Culture Laboratory - Sapienza, University of Rome, Via dei Marsi 78 00185, Rome

Douglas Markant (dmarkant@uncc.edu)
Department of Psychological Science, University of North Carolina, Charlotte - 9201 University City Blvd
Charlotte, NC 28223-000 USA

Costanza De Simone (desimone@mpib-berlin.mpg.de)
MPRG iSearch, Max Planck Institute for Human Development - Lentzeallee 94, 14195 Berlin, Germany

Gianni Valeri (giovanni.valeri@opbg.net)
Ospedale Pediatrico, Bambino Gesù - Piazza di Sant'Onofrio 4, 00165 Rome, Italy

Azzurra Ruggeri (ruggeri@mpib-berlin.mpg.de)
MPRG iSearch, Max Planck Institute for Human Development - Lentzeallee 94, 14195 Berlin, Germany
School of Education, Technical University Munich - Marsstrasse, 20-22, 8035 Munich, Germany

Abstract

Previous research with typically developing (TD) children and adults show an advantage of active control for episodic memory as compared to conditions lacking this control. The present study attempts to replicate this effect in autistic children. Six- to 12-year-old autistic children (n = 30) were instructed to remember as many of 64 presented objects as possible. For half of the materials presented, participants could decide the order and pacing of study (Active condition). For the other half, they passively observed the study decisions of a previous participant (Yoked condition). We found that recognition memory was more accurate for objects studied in the active as compared to the yoked condition, even after a week-long delay. The magnitude of the effect was comparable to that obtained in previous studies with TD children and adults, suggesting a strong robustness for the benefits of active learning. We discuss how pedagogical approaches may be encouraged to utilize self-directed learning strategies to promote inclusive learning.

Keywords: active learning; Autism Spectrum Disorder; Enactment Effect; recognition memory; pedagogy

Introduction

The opportunity to exert active control over the learning experience, often referred to as active, or self-directed learning, has been shown to lead to improved outcomes as compared to more passive forms of instruction (see Bruner, Jolly, & Sylva, 1976; Gureckis & Markant, 2012; Montessori, 1912; Piaget, 1930). In particular, studies with adults show an advantage of active control for episodic memory of objects (Voss, Galvan & Gonsalves, 2011), faces (Liu, Ward, & Markall, 2007), and in spatial learning tasks (Plancher, Barra, Orriols, & Piolino, 2013; for a review see Markant, Ruggeri, Gureckis, & Xu, 2016), as compared to conditions lacking this control. A more recent study suggests that the benefits of active learning for episodic memory of objects might already emerge during early childhood, and become comparable to adults' by age 8 (Ruggeri, Markant, Gureckis, Bretzke, & Xu, 2019). Difficulties in active selection (and thus control) of the contents of learning may emerge in situations where exploratory behaviours are limited. In this paper we explore the effects of active control of learning on episodic memory in autistic children. The examination of atypically developing children might help to further understand the full spectrum of development (Graham & Madigan, 2016) and support the development of novel pedagogical approaches to promote inclusive learning.

Benefits of active control of study and enactment effect. To investigate the effects of active control for episodic memory, studies have typically employed *yoked* designs, which implicate a pair of learners: An active participant who controls the flow of information during learning (e.g., selecting what to study and for how long), and a yoked participant, who observes the experience generated by the active participant (Markant et al., 2016). By matching the content experienced during study across conditions, yoked designs isolate the effects of active decision making on learning and memory. For example, Ruggeri and colleagues (2019) presented 5- to 11-year-old children with a simple

memory game in which they were tasked to remember and later recognize a set of 64 objects. For half of the materials presented, participants could decide the order and pacing of study (Active condition). For the other half, they passively observed the study decisions of a previous participant (Yoked condition). The authors showed that recognition memory was more accurate for objects studied in the active as compared to the yoked condition, and that this memory advantage persists over a week-long delay. This advantage of active learning has been shown to be fairly robust across different types of tasks, developmental stages, and even populations of learners of different nationalities (Brandstatt & Voss, 2014; Ruggeri et al., 2019).

Self-performed tasks (SPT) (Cohen, 1981) present a similar design: Participants are presented with action phrases (for example, "Clap your hands") that they either have to read/perform (Active condition) or that are read/performed by somebody else (Verbal task/Experimenter performed task; Engelkamp & Zimmer, 1989). Participants are then usually tested through a recall or recognition memory task for the action phrases presented. Results from studies using the SPT have convergently indicated advantages for learning associated with the active condition (Engelkamp, 1998). For example, Baker-Ward and Colleagues (1990) found that children as young as six years old exhibited better recall for actions they performed compared to the observed actions of someone else. This effect, referred to as the e*nactment effect*, is extremely robust and is thought to improve memory mainly through motor actions (Engelkamp & Zimmer, 1998). Along these lines, Engelkamp and Zimmer (1994) found that participants, when they physically performed an action, remembered it better than when they just read a distractor phrase similar to the target action.

The ecological validity of self-performed tasks might be limited though, for instance, as SPTs use stimuli exclusively associated with specific actions. Yet, learning processes, particularly those based on recognition memory, involve interactions of different abilities, functions, semiotics, and experiences with a variety of stimuli. Furthermore, SPT paradigms make it difficult to isolate the sources of enactment effects as the content of the tasks differ across verbal, or experimenter performed conditions. As a participant remembers an action they performed more accurately it becomes challenging to separate motor involvement from other kinds of self-representation, for example. Aside from motor actions, there may be different factors influencing enactment effects like metacognition, attention, motivation, or agency. Further work incorporating different stimuli and target behaviors aimed at isolating motor involvement is needed to expand our knowledge on the function and implications of enactment effects in developmental and learning processes.

Learning strategies in autistic individuals. Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder characterized by persistent social communication difficulties as well as restricted interests, repetitive activities and sensory abnormalities (American Psychiatric Association, 2013). Autistic children seem to explore both space and objects less than others. In autistic children, restricted interaction with objects, or an insistence on exploring only a few features of an object may limit the possibilities for learning (Bjorne, 2007). As a consequence, autistic children might be at risk of missing important opportunities for learning, except for those things that lie within their interests, and this might have important consequences for their development (Pierce & Courchesne, 2001). Bondy and Frost (1994) indicate that 80% of autistic children, aged 5 years and younger, who enter special education are non-verbal, and 30% are minimally verbal at 9 years old (see also Anderson et al., 2007). Verbal tasks may thus not be the most methodologically appropriate to assess active learning in autistic children, considering their well-known communication and other general learning difficulties.

As memory enhancements from active learning paradigms seem to be extremely robust in typically developing individuals, research evidence suggests that the enactment effect may also be intact in autistic individuals (see; Grainger, Williams, & Lind, 2014a; Grainger, Williams, & Lind, 2017; Lind & Bowler, 2009; Summers & Craik, 1994; Williams & Happé, 2009). Summers and Craik (1994) found no significant differences in recognition memory for action-phrases between autistic and typically developing (TD) children from an SPT design. These results were confirmed in a study by Yamamoto and Masumoto (2018), who examined the enactment effect for recall and recognition memory in autistic adults and a TD comparison group through an SPT. They found that although overall recall performance was lower for autistic individuals than for the TD group, there were no differences in the enactment effect between groups. Overall, there seem to be no significant differences in the magnitudes of enactment effects for memory tests in autistic children compared to TD children (see Grainger et al., 2014a for a review), measured through research paradigms adopting self performed tasks.

The present study. The present study aimed to explore the benefits of active learning on episodic memory in autistic children by examining their recognition memory for objects studied in an active compared to a yoked learning condition. The design we have adopted is one step beyond the SPT paradigm used by previous studies to elicit the enactment effect, presenting several advantages: First, we used images of objects that are not explicitly associated with performing an intended action. Second, due to its yoked design, the content experienced during study was carefully matched across conditions, so that we could isolate the effects of active control of study on learning and memory. Third, participants were instructed to perform the same motor actions in both active and yoked conditions. In this way, we could also disentangle the effects of active control from the effects that, in SPTs, have often been attributed to motor engagement. Along these lines, a study by Williams and Happé (2009) designed a task in which autistic children were asked to self-perform an action and to perform the

same action on behalf of a doll that represented a separate agent. The authors found that memory was better for the actions that had been self-performed, suggesting that even the enactment effect cannot be exclusively attributed to motor engagement.

Finally, a number of studies have revealed diminished recall but intact recognition memory in autistic individuals (see Boucher, Mayes, & Bigham, 2012 for a review). For this reason, we thought testing recognition memory would be a sufficient task to isolate the effects of active control. Moreover, evidence has suggested that adopting interactive teaching strategies (i.e. visual-interactive materials paired with music) enhances active engagement and learning of autistic students (Carnahan, Musti-Rao & Bailey, 2009). In this sense, the use of a tablet device, with an interactive interface, to assess autistic children who might have communication impairments might be particularly suitable to deliver the paradigm. Past research has also shown that autistic children seem to be more attentive, and motivated resulting in better performance and enjoyment of intervention sessions implemented through tasks involving technological tools (Moore, & Calvert, 2000). This task can reveal the non-verbal learning strategies adopted by autistic children. Our results will add further information on visual object exploration strategies, and contribute to a broader picture of active learning. Based on the literature reviewed above, we expected that autistic children would show active learning benefits to memory similar to that found in TD children of the same age (Ruggeri et al., 2019). In particular, with this design we can explore whether and how the effects of active control of study depend on how participants explore the objects. These insights would bear relevant implications for future research directions and clinical practice.

Method

Participants

We recruited 30 6- to 12-year-old autistic children (4 female, M_{age} = 113.17 months; *SD* = 19.89 months) from the Neuropsychiatry and Neuroscience Unit, I.R.C.C.S. Bambino Gesù Pediatric Hospital (OPBG), Rome, Italy. Participants had been previously screened for a formal diagnosis of ASD using the Autism Diagnostic Observation Schedule (ADOS; Lord et al., 2000). To minimize differences between participants, we recruited individuals who had scored between 5 to 8 out of 10 on the ADOS (*M* = 6.13; *SD* = 1.04). Once we re-ran the ADOS test with the subjects recruited, we excluded one participant who scored below 5. Participants were also previously screened for IQ (*M* = 109.20; *SD* = 13.43) using the Raven's Coloured Progressive Matrices (Raven, Court, & Raven, 1990). The data from 5 additional autistic children were excluded for reasons due to behavioral issues, symptom severity, and technical difficulties.

Materials

As in Ruggeri and colleagues (2019), the stimuli set consisted of 200 line drawings of the most frequent objects mentioned by 2- to 5-year-old children in their everyday conversations with adults, as recorded by the CHILDES corpus (Child Language Data Exchange System; MacWhinney & Snow, 1985). Eight of the 200 drawings were used as training stimuli for the familiarization trials and 192 drawings were used as stimuli during the first and second experimental sessions. The experimental materials were presented on an Android touchscreen tablet using custom software.

Design and Procedure

The experimental procedure was identical to that implemented by Ruggeri and colleagues (2019). The stimuli were presented as a simple memory game whereby children were tasked with remembering as many of the presented objects as possible.

Familiarization phase. Participants were first presented with two familiarization trials aimed at introducing the goal of the game, the study procedures, and making children comfortable using the touchscreen. During each familiarization trial, children were presented with four objects arranged in a 2x2 grid. The objects were shown on the screen for two seconds before disappearing under occluders (same as for the main experimental session, see Figure 1, top). Participants were instructed that the goal of the game was to remember all the objects presented on the screen. The first familiarization trial introduced the study procedure of the active blocks. Participants were told that in some rounds they could decide which occluder button to touch in order to view the object hidden beneath. After a touch, a red frame appeared for 500 ms, followed by the removal of the occluder that would reveal the hidden object. Children were instructed that, before studying another object, they had to touch the object currently displayed once more to make it disappear behind the occluder. The experimenter modeled the touching actions while explaining the procedure. Children then had the opportunity to practice the active study procedure. If necessary, the experimenter provided feedback and repeated the instructions. Once children were familiar with the active study procedure, they moved on to the second familiarization trial, which introduced children to the study procedure of the yoked blocks. They were told that in other rounds the game would decide what objects they would see and for how long. Children were then presented with a randomly generated study sequence. As in the active blocks, a red frame preceded each object for 500 ms so that children had time to allocate their attention to the new study location before the object appeared. To keep engagement and attention level comparable to the active blocks, during yoked blocks children were asked to touch the objects as soon as they appeared, although this touch had no effect on the display.

There were no time constraints for the familiarization trials.

Study phase. The main experimental session consisted of two active and two yoked study blocks (four blocks total), presented in alternating order (i.e., active, yoked, active, yoked). The active block was always presented first, so that children's initial active study pattern would not be influenced by the study pattern observed in the yoked blocks. Each study block presented children with 16 objects arranged in a 4x4 grid. All 16 objects were visible on the screen for 2 seconds at the beginning of each study block, before disappearing under occluders (see Figure 1, top). Across the four blocks, children were asked to memorize 64 objects. In the active blocks, children had 90 seconds to select and study the objects in order to memorize them. In the yoked blocks, children were presented with the 90-second study sequence (i.e., same objects and pacing) of one of the previous participant's active learning blocks. In between blocks, there was a 20-second break in which children were briefly reminded of the study procedure for the next block.

Test phase. The study phase was immediately followed by a test phase consisting of 8 blocks. In each test block, 16 objects were again presented in a 4x4 grid (see see Figure 1, bottom). Across the 8 test blocks, 64 of the objects had appeared during the study phase (old objects) and 64 were objects that were not presented during study (new objects). The number of old objects in each block was randomly varied between 1 and 15. The number of old objects from active and yoked blocks randomly varied across test blocks (active: M = 4:23, SD = 2:16; yoked: M = 4:3, SD = 2:25).

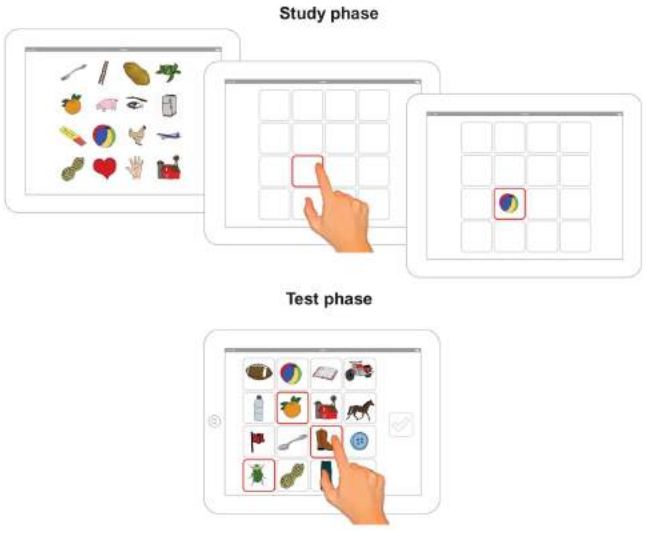


Figure 1. **Top:** Each study round began with all objects displayed for two seconds. After the objects disappeared, participants either selected a location to study (Active condition), causing a red frame to appear, followed by the object, or touched the location where the object appeared (Yoked condition), preceded by a red frame. **Bottom:** During each test block, participants selected the objects that they recognized from the study phase.

All objects were arranged in random locations on the grid. For each block, children were asked to indicate the objects they had studied earlier by touching them on the screen. Selected objects were framed in red to help participants keep track of the objects selected as recognized. Children could deselect any of the previously selected objects by touching them again on the screen and making the red frame disappear. After selecting all the objects they recognized from the study phase, children were prompted to touch a button to proceed to the next test block. Children were not given any feedback about their performance during or after the test phase.

About one week later (range 5 to 8 days; M = 7.04 days; SD = 0.58 days), children revisited the Hospital for a second session in which they were asked to complete 8 new test blocks. The 64 objects studied in the first session were randomly mixed with 64 new objects (i.e., objects that were not used during the first experimental session, neither as study nor as test objects).

Results

We analyzed (1) recognition accuracy (i.e., the number of objects recognized among the ones studied); (2) the correlations between study experience and performance, to test whether certain participants' exploration strategies and patterns lead to better recognition accuracy. In particular, we examined the correlation between the recognition accuracy for a certain object and the time spent studying it, as well as the number of times it had been visited during study. We also examined the correlation between participants' average recognition accuracy and the distance between subsequent study locations (that is, the average distance on the grid between the object currently visible and the one selected next), a basic measure of how systematically a child explored the grid.

Recognition accuracy. We examined recognition accuracy using an ANOVA with study condition (2 levels: active versus yoked) and session (2 levels: test versus one-week-later retest) as within-subject variables. We found a significant main effect of study condition, $F(1, 81) = 16.44$, $p = < .001$. Children recognized more objects studied in the active learning condition (M_{active} = 19.02; SD = 6.70) as compared to the objects studied in the yoked condition (M_{yoked} = 16.26; SD = 6.52), a 9% difference (see Figure 2). We also found a significant effect of session $F(1, 82) = 19.09$, $p = < .001$. Children recognized more objects studied in the first test session (M_{test} = 18.92; SD = 6.72) compared to approximately one week later in the retest session (M_{retest} = 16.22; SD = 6.51). There was no reliable interaction effect between study condition and session (p = .559).

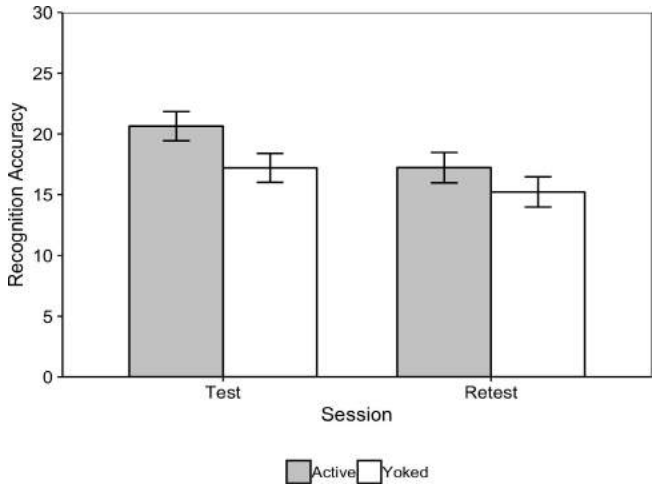


Figure 2: Number of objects correctly recognized in the test trials, displayed by study procedure (active vs. yoked) and session (test vs. retest). Error bars indicate 1 SEM.

Correlations between study experience and performance. Surprisingly, we found that object recognition accuracy was not correlated with the time spent studying an object, nor with the number of times the object had been visited in the active study condition, for both test and retest (see Table 1). However, we found a correlation between recognition memory for objects studied in the active blocks and the distance between the location in which the objects were presented on the study grid and their location on the test grid, $r = .577$, $p = < .01$.

Table 1: Correlations between study measures.

Active study condition				
	Correlations between tests			
Test	1	2	3	4
1. Accuracy in test				
2. Accuracy in retest	.810***			
3. Number of visits	-.093	.087		
4. Study duration	.099	.226	-.340	
7. Distance from study position	.577**	.363	-.242	-.184
Yoked study condition				
	Correlations between tests			
Test	1	2	3	4
1. Accuracy in test				
2. Accuracy in retest	.788***			
3. Number of visits	.067	-.088		
4. Study duration	.479*	.470	-.473*	
5. Distance from study position	-.067	-.356	.096	-.198

* Correlation is significant at the 0.05 level (2-tailed).
** Correlation is significant at the 0.01 level (2-tailed).
*** Correlation is significant at the 0.001 level (2-tailed).

Discussion

The present study investigated whether active control over a learning experience leads to benefits in episodic memory for 6- to 12-year old autistic children. As hypothesized, and similar to previous studies with TD adults and same-aged children, we found that participants' memory was more accurate for objects studied in the active learning compared to the yoked condition. Moreover, the strength of active control for memory encoding is strikingly comparable to the effect found with TD children from an identical recognition task (9% increase over the yoked condition; Ruggeri et al., 2019), and similar to that found in other studies examining the enactment effect in autistic individuals (see Grainger et al., 2014a for a summary). Thus, our results add to the universal robustness of active learning effects. Concurrently, this study complements and adds to the presence of the enactment effect in autistic individuals by employing an alternative paradigm to the commonly used SPTs. Considering our findings with respect to long-term memory, we found that memory improvement for objects studied by active control lasted for at least one week after testing. These findings lend support to established evidence suggesting improved long-term retention as a result of active control of learning (Ruggeri et al., 2019; Yamamoto, & Masumoto, 2018).

As mentioned in the introduction, bearing in mind possible mechanisms responsible for the active learning advantage, motor involvement has often been suggested to play an important role (Engelkamp, 1998; Markant et al., 2016). In experiments that implement SPTs, the participant enacts an action phrase like 'Wave Goodbye' in the active condition, and observes the experimenter performing a different action phrase in the passive condition. This idea (Engelkamp & Zimmer, 1989) supposes that performing an action involves motor components, which add rich contextual properties that help the encoding process of SPTs (Engelkamp, 1998). However, in our study participants are engaged in approximately the same motor actions in both active and yoked conditions. This result suggests that the process of physically performing an action is not necessary to scaffold memory performance.

Rather unexpectedly our results also seem to suggest, that episodic memory is not influenced by autistic children's study patterns, in both conditions. Objects studied for a longer time or visited more often were not recognized more accurately. This differs from all previous adults and children active learning studies that have used this paradigm (see Gureckis & Markant, 2012; Markant et al., 2014; Markant et al., 2016; Ruggeri et al., 2019). This might be related to the deficit in metamemory and metacognition demonstrated in autistic individuals (Grainger, Williams, & Lind, 2014b). That is, due to such deficits, autistic children may not have been strategically devoted to their study effort, allocating the same amount of time and visits to all object images. Therefore, we did not have enough variability to capture a correlational effect. It is extremely interesting to notice that the advantage of active learning for memory encoding does not seem to depend on the efficiency of children's study strategies and metacognitive decision making, and that it persists when such processes do not play a prominent role. Future studies should investigate more thoroughly the role

of metacognition and metamemory, as well as attention and motivation on the active learning benefit for memory encoding.

Again in contrast with results from prior research, we found that recognition accuracy for object studied in the active condition is correlated with the distance between the location in which the object was presented on the study grid and its location on the test grid. Having the objects presented in the same location on the grid across the study and test blocks did help children recognize them more accurately, but only in the active condition. These results might speak, though indirectly, in favor of an active learning advantage for spatial recall in autistic children. However, only a direct test of spatial memory would allow confirmation of this hypothesis.

The natural next step would be to extend this paradigm to include more real-world stimuli and tasks targeted to autistic children as well as other developmental disorders. For example, Ruggeri and colleagues (2019) designed a task to model real learning situations children encounter in school. Using a similar paradigm to our study, children were tasked to learn the French words for images of objects presented in a study space. The experimenters found that French words were remembered more accurately studied in an active as compared to a yoked condition. Based on this research, future studies might explore the role of active learning in learning new actions, words or behaviors. We are currently in the process of collecting a much larger sample, across different age groups and encompassing a wider range of symptom severity and cognitive maturity. On one hand, this would allow us to trace the emergence of the active learning advantage and compare the developmental trajectories of this effect in autistic and TD children. On the other hand, we are keen to explore whether and how general cognitive performance and symptom severity might impact the advantage of active learning and children's active study strategies, although previous research suggests that ASD traits do not impact memory for self-representations (Williams, Nicholson, & Grainger, 2018).

In conclusion, because autistic students often have difficulties participating in classroom activities (Sparapani, Morgan, Reinhardt, Schatschneider, & Wetherby, 2016), it is important to better understand how these children learn to improve and develop current and novel teaching methods. If active control over the learning experience can enhance episodic memory in ASD, then teachers and educators might think of supporting active learning approaches in pedagogical applications. Offering children with developmental disorders opportunities for concrete self-generated, active learning experiences could help promote greater learning outcomes (Haslam, Wagner, Wegener, & Malouf, 2017). Involving the student in their own learning can also be beneficial for reducing problematic behaviors, while at the same time improving skill acquisition (Toussaint, Kodak, & Vladescu, 2016). Alternative modes of teaching based on the use of images and pictures, rather than written words, are encouraging new therapeutic and instructional strategies for autistic children. Consequently, language and communication development devices (e.g. the Picture Exchange Communication System, PECS; Bondy & Frost, 1994) might aim to utilize active learning benefits to ameliorate memory.

Finally, this study tries to bridge atypical, developmental and cognitive research without relying on clinical variations to determine major differences between comparative groups. Rather, our results highlight that autistic individuals share the same memory advantage from active control of learning as TD individuals. This dimensional approach allows for researching *similarities* between typical and atypical groups, and while being as informative as revealing differences (Graham & Madigan, 2016), can support inclusive classrooms. Considering that active learning effects on memory are present in TD as well as autistic children, classrooms could adopt self-directed, active learning methods that would not only benefit both typical and atypical children, but also children who fall somewhere in between these categories.

Acknowledgements

We would like to thank all the children and families who kindly participated, as well as the staff at the Neuroscience and Neuropsychiatric Unit of the Bambino Gesù Pediatric Hospital in Rome for their helpful support in data collection.

References

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA: American Psychiatric Publishing.

Anderson, D. K., Lord, C., Risi, S., DiLavore, P. S., Shulman, C., Thurm, A., . . . Pickles, A. (2007). Patterns of growth in verbal abilities among children with autism spectrum disorder. *J Consult Clin Psychol, 75*(4), 594-604.

Baker-Ward, L., Hess, T. M., & Flannagan, D. A. (1990). The effects of involvement on children's memory for event. *Cognitive Development, 5*, 55-69.

Björne, P. Z. (2007). *A possible world: Autism from practice to theory*. Lund: Cognitive Science.

Boucher, J., Mayes, A., & Bigham, S. (2012). Memory in autistic spectrum disorder. *Psychol Bull, 138*(3), 458-496.

Bondy, A. S., & Frost, L. A. (1994). Picture exchange communication system. *Focus on Autistic Behavior, 9*(3), 1-19.

Brandstatt, K. L., & Voss, J. L. (2014). Age-related impairments in active learning and strategic visual exploration. *Frontiers in Aging Neuroscience*, *6*(FEB), [Article 19].

Bruner, J. S., Jolly, A., & Sylva, K. (1976). *Play: Its role in development and evolution*. New York: Basic Books.

Carnahan, C., Musti-Rao, S., & Bailey, J. (2008). Promoting Active engagement in small group learning experiences for students with autism and significant learning needs.

Education and Treatment of Children, 32, 37-61.

Cohen, R. L. (1981). On the generality of some memory laws. *Scandinavian Journal of Psychology, 22*(1), 267-281.

Engelkamp, J. (1998). *Memory for actions*. Hove: Psychology Press.

Engelkamp, J., & Zimmer, H. D. (1989). Memory for action events: A new field of research. *Psychological Research, 51*, 153-157.

Engelkamp, J., & Zimmer, H. D. (1994). Motor similarity in subject-performed tasks. *Psychological Research, 57*(1), 47-53.

Graham, S. A., & Madigan, S. (2016). Bridging the gaps in the study of typical and atypical cognitive development: A commentary. *Journal of Cognition and Development, 17*(4), 671-681.

Grainger, C., Williams, D. M., & Lind, S. E. (2014a). Online action monitoring and memory for self-performed actions in autism spectrum disorder. *Journal of Autism and Developmental Disorders, 44*(5), 1193-1206.

Grainger, C., Williams, D. M., & Lind, S. E. (2014b). Metacognition, metamemory, and mindreading in high-functioning adults with autism spectrum disorder. *J Abnorm Psychol, 123*(3), 650-659.

Grainger, C., Williams, D. M., & Lind, S. E. (2017). Recognition memory and source memory in autism spectrum disorder: A study of the intention superiority and enactment effects. *Autism, 21*(7), 812-820.

Gureckis, T. M., & Markant, D. B. (2012). Self-Directed Learning: A Cognitive and Computational Perspective. *Perspect Psychol Sci, 7*(5), 464-481.

Haslam, C., Wagner, J., Wegener, S., & Malouf, T. (2017). Elaborative encoding through self-generation enhances outcomes with errorless learning: Findings from the Skypekids memory study. *Neuropsychol Rehabil, 27*(1), 60-79.

Lind, S. E. (2010). Memory and the self in autism: A review and theoretical framework. *Autism, 14*(5), 430-456.

Lind, S. E., & Bowler, D. M. (2009). Recognition memory, self-other source memory, and theory-of-mind in children with autism spectrum disorder. *J Autism Dev Disord, 39*(9), 1231-1239.

Liu, C. H., Ward, J., & Markall, H. (2007). The role of active exploration of 3d face stimuli on recognition memory of facial information. *Journal of Experimental Psychology: Human Perception and Performance, 33*(4):895.

Lord, C., Risi, S., Lambrecht, L., Cook, E. H., Leventhal, B. L., DiLavore, P. C., ... & Rutter, M. (2000). The Autism Diagnostic Observation Schedule-Generic: A standard measure of social and communication deficits associated with the spectrum of autism. *Journal of Autism & Developmental Disorders, 30*, 205–223.

Macwhinney, B., & Snow, C. E. (1985). The child language data exchange system. *Journal of Child Language, 12*, 271-296.

Markant, D., DuBrow, S., Davachi, L., & Gureckis, T. M. (2014). Deconstructing the effect of self-directed study on episodic memory. *Mem Cognit, 42*(8), 1211-1224.

Markant, D. B., Ruggeri, A., Gureckis, T. M., & Xu, F. (2016). Enhanced memory as a common effect of active learning. *Mind, Brain, and Education, 10*(3), 142-152.

Montessori, M. (1964). *The Montessori Method*. Rome 1912. Oxford, England: Bentley, Inc..

Moore, M., & Calvert, S. (2000). Brief report: Vocabulary acquisition for children with autism: Teacher or computer instruction. *Journal of Autism and Developmental Disorders, 30*(4), 359-362.

Piaget, J. (1930). *The child's conception of physical causality*. London: K. Paul, Trench, Trubner & Co.

Pierce, Karen & Courchesne, Eric. (2001). Evidence for a cerebellar role in reduced exploration and stereotyped behavior in autism. *Biological psychiatry, 49*, 655-64.

Plancher, G., Barra, J., Orriols, E., & Piolino, P. (2013). The influence of action on episodic memory: A virtual reality study. *Q J Exp Psychol (Hove), 66*(5), 895-909.

Raven, J. C., Court, J. H., & Raven, J. (1990). *Coloured progressive matrices*. Oxford: Oxford University Press.

Ruggeri, A., Markant, D. B., Gureckis, T. M., Bretzke, M., & Xu, F. (2019). Memory enhancements from active control of learning emerge across development. *Cognition, 186*, 82-94.

Sparapani, N., Morgan, L., Reinhardt, V. P., Schatschneider, C., & Wetherby, A. M. (2016). Evaluation of classroom active engagement in elementary students with autism spectrum disorder. *J Autism Dev Disord, 46*(3), 782-796.

Summers, J. A., & Craik, F. I. (1994). The effects of subject-performed tasks on the memory performance of verbal autistic children. *Journal of Autism and Developmental Disorders, 24*(6), 773-783.

Toussaint, K. A., Kodak, T., & Vladescu, J. C. (2016). An evaluation of choice on instructional efficacy and individual preferences among children with autism. *J Appl Behav Anal, 49*(1), 170-175.

Voss, J. L., Galvan, A., & Gonsalves, B. D. (2011a). Cortical regions recruited for complex active-learning strategies and action planning exhibit rapid reactivation during memory retrieval. *Neuropsychologia, 49*, 3956–3966.

Williams, D., & Happé, F. (2009). Pre-Conceptual aspects of self-awareness in autism spectrum disorder: The case of action-monitoring. *Journal of Autism and Developmental Disorders, 39*, 251-259.

Williams, D. M., Nicholson, T., & Grainger, C. (2018). The self-reference effect on perception: Undiminished in adults with autism and no relation to autism traits. *Autism Res, 11*(2), 331-341.

Yamamoto, K., & Masumoto, K. (2018). Brief report: Memory for self-performed actions in adults with autism spectrum disorder: Why does memory of self decline in ASD? *J Autism Dev Disord*.

Deception in evidential reasoning: Willful deceit or honest mistake?

Toby D. Pilditch[1,2], Alexander Fries[3], and David Lagnado[1]

[1]Department of Experimental Psychology, University College London, 26 Bedford Way, London, WC1H 0AP, UK
[2]University of Oxford, School of Geography and the Environment, South Parks Road, Oxford, OX1 3QY, UK
[3]UCL Institute of Neurology, University College London, Queen Square, London, WC1N 3BG, UK

Abstract

How does one deal with the possibility of deception? Extant literature has mostly focused on identifying deception via cue detection. However, how we reason about the possibility of deception remains under-explored. We use a novel formalism to expose the complexity of this reasoning problem (e.g. separating the uncertainty of an honest mistake, from willful deception), in the process highlighting several reasoning errors regarding deception. Notably, we show reasoners to make substantial errors when reasoning about a (possibly) deceptive source *in isolation* (including base rate neglect errors), but find that reasoning improves when further (independently sourced) corroborative or contradicting reports are introduced.

Keywords: deception; evidential reasoning; probabilistic reasoning; Bayesian Networks; belief updating

Introduction

The question of how to deal with the possibility of deception has long been of interest to police, military and intelligence investigation, among other domains. A potentially deceptive source, more so than a generally unreliable (e.g. incompetent) source, can be particularly deleterious to an investigation, via the wilful sowing of misinformation. Critically, however, investigators seldom have definitive proof of deception, and are therefore placed into the realm of reasoning under uncertainty. In the present paper, we demonstrate a novel Bayesian formalism for capturing the complex uncertainties surrounding (potentially) deceptive sources, such that optimal inferences regarding the likelihood of deception, as well as the hypothesis being informed upon, may be updated with minimised inaccuracy (Pettigrew, 2016). Moreover, we demonstrate that lay reasoners wildly diverge against such a normative expectation.

Deception in Psychology

Deception has typically been researched in terms of lie detection (see Vrij, 2008). Crucially, previous research has noted that individuals struggle with the uncertainties surrounding the possibility of deception (e.g. chance error vs deception) when explaining errors (Schul, Mayo, Burnstein, & Yahalom, 2007).

Research on perceived trustworthiness has shown that it influences attitudes (Cuddy, Glick, & Beninger, 2011; Fiske, Cuddy, & Glick, 2007), persuasive efficacy (Briñol & Petty, 2009), risk perception (Siegrist, Cvetkovich, & Roth, 2000; Earle, Siegrist, & Gutscher, 2010), and advice uptake (Schul & Peri, 2015). But relatively little research has been conducted in regards to not only how people *do reason* about the *possibility of deception*, but also how they *should*. Within evidential reasoning, one can consider deception to be a special case of (dis)trustworthiness. Dual process models in argumentation, like the Heuristic Systematic Model (HSM; Chaiken & Maheswaran, 1994) and Elaboration Likelihood Model (ELM; Petty & Cacioppo, 1984) have argued that cues to the trustworthiness of a source are only attended to in the absence of effortful engagement with the arguments made by that source.

More recently, coherence-based models, such as the Bayesian source credibility model (Hahn, Harris, & Corner, 2009; Harris, Hahn, Madsen, & Hsu, 2015) have provided a framework that moves beyond the directional predictions of earlier models. This has allowed for the integration of a source's trustworthiness (the willingness to impart accurate information) and orthogonally, expertise (the capacity to impart accurate information) into the support provided by a report from a source. Using these models as a normative backdrop, lay reasoners have been shown to take into account the impact of credibility on argument strength (Hahn et al, 2009), and even follow appropriate adjustments in estimations of argument strength and source reliability in light of (shared) compromising reliability information (Madsen, Hahn, & Pilditch, 2018).

Formalising Deception

Taking forward the notion of deception as a special form of (un)reliability, work using Bayesian networks representations to model legal cases has used an idiomatic approach for witness testimony (Fenton, Neil, & Lagnado, 2013). More precisely, when modelling the strength of a witness's testimony, one may consider two possible (non-exclusive) causes of it – the hypothesis being reported on (e.g. guilt of suspect), and the reliability of the witness.

In the same manner, we may model the representation of deception as a possible cause, along with the hypothesis being reported upon (Lagnado, Fenton & Neil, 2013). Fig. 1 below uses an example case of a target hypothesis – "Is the suspect under questioning in fact the mob's hitman?", and a number of informing sources in a police investigation. Two of these, a forensic scientist and an eyewitness (each retaining their own respective reliabilities), and two Inspectors, McGarret and Graham, who are typically accurate in their investigative reports. Critically, each source reports independently of the other, but there is reason to believe McGarret and Graham *may* in fact be in league with the mob, and thus the possibility of deception is introduced (left-most node in Fig. 1).

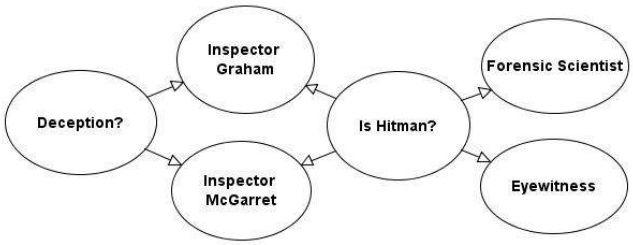


Figure 1. Graphical representation of deception scenario.

To tease apart the levels of uncertainty introduced by the possible deception cause, it is necessary to look at an example conditional probability table (CPT) that represents a (possibly) deceptive agent:

Table 1. Conditional probability table (CPT) representation of a potentially deceptive source, reporting on a hypothesis (Hyp) as either true (T) or false (F).

	Deception = False		Deception = True	
	Hyp = F	Hyp = T	Hyp = F	Hyp = T
Rep = Yes	α	β	γ	δ
Rep = No	(1 -α)	(1 - β)	(1 - γ)	(1 - δ)

In Table 1, α is the probability that the source reports that the suspect is a mob hitman (Rep = Yes) given that he *is honest (Deception = false)* and the suspect is not a mob hitman (Hyp = false). β is the probability that the source reports that the suspect is a mob hitman (Rep = Yes) given that he *is honest (Deception = false)* and the suspect is a mob hitman (Hyp = True). γ is the probability that the source reports that the suspect is a mob hitman (Rep = Yes) given that he *is dishonest (Deception = true)* and the suspect is not a mob hitman (Hyp = false). Finally, δ is the probability that the source reports that the suspect is a mob hitman (Rep = Yes) given that he *is dishonest (Deception = true)* and the suspect is a mob hitman (Hyp=true). δ and (1 - γ) may be due to an imperfect deception, or due to long-run motivations to keep the deception in place. For this initial proof of principle, we simplify the notion of deception by removing this possibility (δ = 0; γ = 1). Put another way, in the scenario we model (and present to participants), deceivers will a) have insider knowledge (i.e. know the true state of "Hyp"), and b) always lie.

Placing Table 1 within the context of the model (and scenario) outlined in Fig. 1, there are a number of important inferences of which to take note.

Firstly, the difference between the probability of a report due to honest error and due to wilful deception (α versus γ), plays a pivotal role in the potential diagnosticity of the report for both P(Deception) and P(Hyp), such that as γ – α increases, the report becomes more diagnostic of deception.

This in turn has a multiplicative effect when considering the second elements: the prior probability of deception (P(Deception)) and – critically – the prior probability of hypothesis being true (P(Hyp)). More precisely, if a report confirms a hypothesis that is *likely* (e.g. P(Hyp) > .5), then P(Deception) should *decrease*, whilst if the report confirms an *unlikely* hypothesis (e.g. P(Hyp) < .5), then P(Deception) should increase. These inferences can best be explained with consideration of how *surprising* a report would be from an honest agent. If unsurprising (e.g. they are saying something *expected*), then an alternative explanation of the report (e.g. deception) is less warranted, and vice versa.

Thirdly, subsequent testimony from independent witnesses will lead to intercausal inferences of P(Deception). For instance, if potentially deceptive agents have their reports *corroborated* by independent testimony, then the increasing probability of P(Hyp) *explains away* the possible deception explanation, lowering P(Deception). Conversely, if independent testimony contradicts the reports of the potentially deceptive sources, then P(Hyp) becomes a less likely explanation of their reports, and again via explaining away, P(Deception) becomes a more likely explanation.

Finally, we seek to provide reasoners with one further clue to deception inferences. The common-cause structure of the deception explanation (left-most node of Fig. 1) – where if deception is true, it explains *both* Inspector Graham and Inspector McGarret's reports, in conjunction with "always liars" (δ = 0; γ = 1) element of their CPTs, allows for an observation-based way of dismissing the possibility of deception. More precisely, given the above, it is not possible for P(Deception) to be true if the two Inspectors contradict each other.

In sum, we use the above formalism to test lay reasoners on 3 different elements of the uncertainty surrounding deception: the prior probabilities of deception (and reported hypothesis) as explanations, the conditional probabilities, and observation-based inference.

The Experiment We present lay reasoners with the above scenario of a police investigation looking into whether a suspect in custody is a mob hitman. The key element to this scenario is to assess how well lay reasoners can integrate the influence of the possibility of deception when integrating testimony from what may otherwise be considered reliable sources.

Of interest is whether reasoners are able to make the following key inferences as more evidence comes in from the available sources:

1. Will reasoners sufficiently account for the likelihood of an honest report when estimating the probability of deception? I.e. If the source is reporting the (a priori) more likely state of the world, then P(Deception) should in fact *decrease*?
2. Will reasoners sufficiently account for the common-cause element of this form of deception? Namely that if the two potentially deceptive sources contradict one another, then they cannot (both) be (all-knowing, perfect liar) deceivers.
3. Will reasoners sufficiently account for the impact of independent sources, when their reports either a) corroborate the deceptive agents (and thus P(Deception) should decrease) or b) contradict the

deceptive agents (and thus P(Deception) should increase)? However, reasoners are likely to get the qualitative direction of these latter inferences.

Method

Participants 180 UK participants were recruited and participated online through the Prolific Academic platform. Participants were native English speakers, with a median age of 28.5 (*SD* = 11.3), and 113 participants identified as female. All participants gave informed consent, and were paid 1.30GBP for their time (*Median* = 8.69 minutes, *SD* = 4.38).

Procedure & Design Participants are provided with a brief background to the scenario, in which they are investigating whether a suspect is in fact a hitman hired by the local mob. They are instructed that they have a number of sources to inform their investigation: two highly reliable inspectors, Graham and McGarret, a Forensic Expert, and an eyewitness – all of whom provide assessments independently of one another. Critically, along with being provided with a prior probability of the suspect being the hitman (P(Hitman) = .1), participants are told there are some logs that suggest the two investigators may be in league with the mob. It is explained to participants that although this is unlikely (P(Deception = .1), if true, the two inspectors will both know the truth (they know the identity of the hitman) and will be motivated to always lie (make sure the innocent suspect takes the fall, or prevent the guilty suspect from going to jail). All the necessary probabilities to populate the underlying model (e.g. error rates of each source) and structures (e.g. common-cause structure of P(Deception)) were provided to participants.[1]

Having had the background explained to them, participants then repeated back the prior probabilities for P(Hitman) and P(Deception):

P(Hitman): *"Until you receive the assessments of other professionals investigating whether the suspect is in fact the hitman, you can safely assume a fairly low (10%) chance of the suspect being the hitman. Please indicate you understood the initial (baseline) probability of the suspect being the hitman:"*

P(Deception): "*... there is only a 10% probability that the two criminal investigators are in fact compromised ...*

Please indicate you understood the initial (baseline) probability of the two criminal investigators being in league with the mob boss:"

Using the gRain package in R (Højsgaard, 2012), these elicited prior probabilities were used to outfit a Bayesian Network (BN) model (Fig. 1) for each participant, creating individually fitted BNs (hereafter termed Behaviorally Informed Bayesian Networks; BIBNs). The remaining structure and parameters were taken from the background information presented to all participants. Thus, a fitted normative comparison could be made for inferences on the participant level.

Following the elicitation of priors, participants then saw three stages of observations, with questions asked at each stage.

(T1) Firstly, participants heard from both the potentially deceptive agents ("DecAgents"). This was manipulated between-subjects, as: Both Report Hitman=True, Both Report Hitman=False, One Contradicts the other.

(T2) Participants then heard from the Forensic Expert, followed by the eyewitness (T3), in separate elicitation stages. These ("OtherAgent") reports were also manipulated between-subjects, as: Both Report Hitman=True, Both Report Hitman=False.

Across these 3 stages, participants were asked two sets of questions:

Probability Estimates (sliders from 0-100%, no default):

- **Hitman Hypothesis:** *"Based on the evidence so far, what do you believe is the current probability of the suspect being the hitman?"*
- **Deception Hypothesis:** "*Based on the evidence so far, what do you believe the current probability is that the criminal investigators are in league with the mob boss?*"

Qualitative Judgments (forced choice; response options: "Increased" / "Decreased" / "Same"; randomized presentation order.):

- **Hitman Hypothesis:** *"Based on the evidence so far, do you believe the probability of the suspect being the hitman has increased, decreased, or remained the same?"*
- **Deception Hypothesis:** "*Based on the evidence so far, do you believe the probability that the criminal investigators are in league with the mob boss has increased, decreased, or remained the same?*"

Thus, this 3 (DecAgents reports) x 2 (OtherAgents reports) x 3 (Elicitation stage) x 2 (Hypothesis) design allows for the testing of the influence of explanation priors, internal (within DecAgents) contradiction, and independent corroboration/contradiction, on estimates (both quantitative and qualitative) of the probability of the hypothesis, and the probability of deception.

Results

Bayesian statistics were employed throughout[2] using the JASP statistical software (JASP Team, 2018). For the sake of brevity, analyses are not reported exhaustively here.

[1] Using the notation of Table 1, participants were given values $\alpha = .05$ (honest false positive); $\beta = .95$ (honest true positive); and $\gamma = 1$, $\delta = 0$ (deception = always lie) for deceptive agents. For full details of the materials used, as well as the collected data, please see https://osf.io/4hvu6/.

[2] Bayes Factors (BF_{10}: likelihood ratio of data given hypothesis, over data given null), may be interpreted as: 1 – 3 = anecdotal support; 3-10 = substantial; 10-30 = strong; 30-100 = very strong; >100 = decisive (Jeffreys, 1961). Conversely, Bayes Factors < .33 can be considered substantial support for the *null* (Dienes, 2014). All analyses used an objective (uninformed) prior. Sample sizes for a given analysis (*N*), and Bayesian Credibility Intervals (95% CI) are indicated wherever appropriate.

Hypothesis 1: Priors and Deception (Base rate neglect)
To understand the impact of priors, we look at estimates and judgments relating to the introduction of the potentially deceptive agents reports (i.e. Baseline to T1), on both P(Hitman) and P(Deception) estimates and judgments, with greater errors predicted for the latter.

P(Hitman) estimates (black lines, Fig. 2). A repeated measures ANOVA was run using elicitation stage (Baseline, T1) and Observed vs Predicted (Data vs BIBN Model) as within-subject factors, and DecAgents condition (restricted to Hitman=True vs Hitman=False reports) as a between-subject factor. This found main effects of elicitation stage (positive trend), $BF_{Inclusion} > 10000$, Observed vs Predicted (data > model), $BF_{Inclusion} = 4.905$, DecAgents condition (Hitman=True > Hitman=False), $BF_{Inclusion} > 10000$, decisive deviations from expectation over time, $BF_{Inclusion} = 5.334$, and opposing trends based on DecAgents condition (increases with Hitman=True, decreases with Hitman=False), $BF_{Inclusion} > 10000$. Crucially, there was no evidence for an interaction of Observed vs Predicted with DecAgents condition, $BF_{Inclusion} = 1.112$, or in conjunction with elicitation stage, $BF_{Inclusion} = 2.178$, indicating no influence of reported base rates on the correctness of P(Hitman) estimates.[3]

P(Deception) estimates (grey lines, Fig. 2). Not only are the same background terms all decisive ($BF_{Inclusion}$'s all > 10000), but there are decisive interactions of Observed vs Predicted and DecAgents condition, $BF_{Inclusion} > 10000$, and the three-way including elicitation stage, $BF_{Inclusion} > 10000$. As can be seen in Fig. 2 by looking at grey solid (participant) vs grey dashed (BIBN model) lines in the middle row (DecAgents reports Hitman=False) vs bottom row (DecAgents reports Hitman=True), estimates *increase* when they should *decrease* in the former, and insufficiently increase in the latter.[4]

Qualitative judgments. Correct responding proportion for the change in P(Hitman) to P(Hitman|DecAgents) did not differ between the DecAgents reports Hitman=True (.39) and DecAgents reports Hitman=False (.25) conditions (N = 121), $BF_{10} = 0.852$. However, in line with probability estimate data, there was substantial evidence for correct responding proportions for the change in P(Deception) to P(Deception|DecAgents) being worse in the DecAgents reports Hitman=False (.1) than DecAgents reports Hitman=True (.28) conditions (N = 121), $BF_{10} = 3.99$.

This latter effect, in conjunction with the P(Deception) estimates, confirms the neglect of the report base rates when considering the possibility of deception, leading to substantial overestimation.

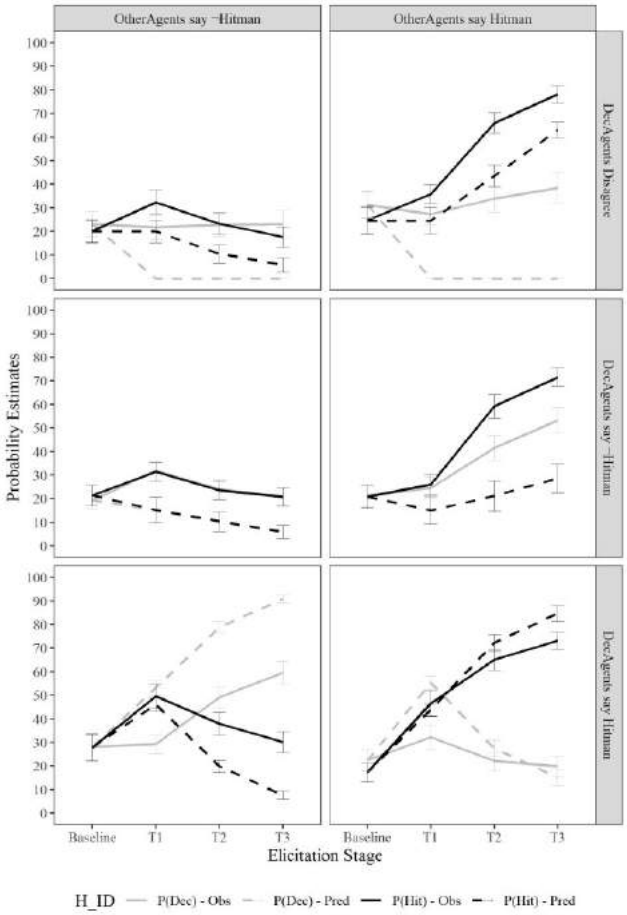


Figure 2. P(Deception) estimates (solid grey lines) and P(Hitman) estimates (solid black lines) across elicitation stages, split by condition. BIBN model predictions are also shown (dashed lines). Error bars reflect standard error.

Hypothesis 2: Common-cause, logic and Deception
To address hypothesis 2, we turn to the DecAgents disagree condition (top row, Fig. 2). Here we focus again on the change in P(Deception) estimates from baseline to T1, as well as the correctness of qualitative judgments. The logic of the structure and conditional probabilities dictate that disagreement between deceptive sources *disproves deception*. However, the repeated measures ANOVA found a decisive deviation from expectation in P(Deception) estimates when moving from baseline to T1, $BF_{Inclusion} > 10000$[5] – an effect corroborated by a t-test showing participants P(Deception) estimates at T1 to be decisively above 0 (N = 59, M = 24.61, SD = 25.44), $BF_{10} > 10000$, δ = 0.937 (95% CI: [0.657, 1.240]).

This error was further confirmed qualitatively, with correct responses (i.e. "Probability decreases") no different from chance level (0.33) responding (N = 59), $BF_{10} = 0.162$, δ = 0.309 (95% CI: [0.203, 0.432]) in a binomial test,

[3] The model with only the above significant terms yielded the best fit, $BF_M = 14.099$, and was significant overall, $BF_{10} = 1.160 * 10^{19}$.

[4] The model with all terms included yielded the best fit, $BF_M = 1.929 * 10^9$, and was significant overall, $BF_{10} = 3.098 * 10^{27}$.

[5] The model including this interaction term yielded the most significant fit, $BF_M = 733042.66$, and was significant overall, $BF_{10} = 3.958 * 10^{15}$.

further confirming an ignorance of the structure and logic based capacity to refute the possibility of deception.
Taken together, these results show that when reasoning about deception, inferences based on structural relations (and logic) alone are highly error prone, once more leading to substantial deception overestimation.

Hypothesis 3: Corroboration, contradiction, and Deception (Explaining Away)

To step through participant estimations of the impact of corroboration / contradiction of possibly deceptive agents, we look first at quantitative estimates (P(Hitman) and P(Deception)) across elicitation stages 1 to 3 – assessing deviation from normative expectation. Second, the correctness of qualitative judgments are assessed over these same stages. This is split by each 2x2 cell (Corroborating Hitman=True, corroborating Hitman=False, contradicting Hitman=True, contradicting Hitman=False).

Corroborating Hitman=True (bottom-right facet, Fig. 2). Repeated measures ANOVAs (elicitation stages T1-T3, and Observed vs Expected) reveal participants do not differ from normative expectation for P(Hitman) estimates, $BF_{Inclusion}$ = 1.777, and track this expectation across elicitation stages, $BF_{Inclusion}$ = 1.436. However, P(Deception) estimates are shown to decisively differ from normative expectation (underestimation), $BF_{Inclusion}$ > 10000, but this deviation decreases across stages, $BF_{Inclusion}$ > 10000.

Table 2 below reveals that whilst qualitative judgments at T1 (when only DecAgents have reported) are correct no better than chance, correct responding at T2 and T3 are greater than chance.

Table 2. Proportion of correct responding in corroborating Hitman=True group. N = 30.

Stage	Hypothesis	Proportion	$\neq .33$ (BF_{10})
T1	Hitman	0.433	0.443
	Deception	0.300	0.218
T2	Hitman	0.900	> 10000
	Deception	0.600	21.16
T3	Hitman	0.733	5313.25
	Deception	0.567	7.527

Corroborating Hitman=False (middle-left facet, Fig. 2). Repeated measures ANOVAs reveal participants decisively differ from normative expectation for P(Hitman) estimates (overestimation), $BF_{Inclusion}$ > 10000, but this deviation does not change across elicitation stages, $BF_{Inclusion}$ = 0.390. Similarly, P(Deception) estimates are shown to decisively differ from normative expectation (overestimation), $BF_{Inclusion}$ > 10000, and this does not change across stages, $BF_{Inclusion}$ = 0.45.

Table 3 below reveals that once again whilst qualitative judgments at T1 (when only DecAgents have reported) are correct no better than chance, correct responding at T2 and T3 are again greater than chance.

Table 3. Proportion of correct responding in corroborating Hitman=False group. N = 32.

Stage	Hypothesis	Proportion	$\neq .33$ (BF_{10})
T1	Hitman	0.25	0.306
	Deception	0.031	707.137†
T2	Hitman	0.781	> 10000
	Deception	0.563	8.124
T3	Hitman	0.656	248.427
	Deception	0.594	22.385

† = Decisively *worse* than chance level.

Contradicting Hitman=True (bottom-left facet, Fig. 2). Repeated measures ANOVAs reveal participants decisively overestimate P(Hitman), $BF_{Inclusion}$ > 10000, and there is strong evidence that this overestimation increases across elicitation stages, $BF_{Inclusion}$ = 17.66. However, participants decisively *underestimate* P(Deception), $BF_{Inclusion}$ > 10000, a trend that does not change across elicitation changes, $BF_{Inclusion}$ = 0.656. Table 4 below reveals that qualitative judgments at T1 (when only DecAgents have reported) are again correct no better than chance, whilst correct responding at T2 and T3 are decisively greater than chance.

Table 4. Proportion of correct responding in contradicting Hitman=True group. N = 31.

Stage	Hypothesis	Proportion	$\neq .33$ (BF_{10})
T1	Hitman	0.355	0.220
	Deception	0.258	0.282
T2	Hitman	0.677	499.34
	Deception	0.774	> 10000
T3	Hitman	0.774	> 10000
	Deception	0.677	499.34

Contradicting Hitman=False (bottom-left facet, Fig. 2). The final repeated measures ANOVAs reveal participants again decisively overestimate P(Hitman), $BF_{Inclusion}$ > 10000, and that this overestimation increases across elicitation stages, $BF_{Inclusion}$ = 334.9. Similarly, participants decisively overestimate P(Deception), $BF_{Inclusion}$ > 10000, but this does not change across elicitation changes, $BF_{Inclusion}$ = 1.965.

Finally, Table 5 below reveals that qualitative judgments at T1 (when only DecAgents have reported) are once again correct no better than chance, whilst correct responding at T2 and T3 are substantially greater than chance.

Table 5. Proportion of correct responding in contradicting Hitman=False group. N = 31.

Stage	Hypothesis	Proportion	$\neq .33$ (BF_{10})
T1	Hitman	0.25	0.307
	Deception	0.179	0.897
T2	Hitman	0.857	> 10000
	Deception	0.714	1164.74
T3	Hitman	0.821	> 10000
	Deception	0.607	20.143

Hypothesis 3 Summary. Taking these 4 sets of analyses together, it is clear that participants can qualitatively appreciate the influence of both corroboration and contradiction from independent sources on potentially deceptive sources, for both P(Hitman), via diagnostic inference, and P(Deception), via an explaining away inference. This is in stark comparison to the substantial qualitative error rates at T1, when only potentially deceptive agents have been observed (see Hypothesis 1). However, estimation data reveals participants consistently overestimate P(Hitman), irrespective of condition (with the exception of corroborating hitman=True). In line with Hypothesis 1, P(Deception) is overestimated when the potentially deceptive agents are reporting the a priori more likely hypothesis (Hitman=False), and underestimated when reporting the less likely hypothesis (Hitman=True). This again suggests a base rate neglect component to assessments of deception.

Conclusions

The issue of how to deal with the possibility of deception when reasoning under uncertainty is as complex as it is potentially deleterious. We present novel findings that lay reasoners are prone to several systematic errors when integrating the possibility of deception, often leading to substantial overestimation.

Using a Bayesian Network formalism, we disentangle the underlying components of deception, including the base rates of deception and the hypothesis the (potentially deceptive) source is reporting on (here, P(Hitman)), structural and logical components, as well as internal (potentially deceptive source reports) and external (corroborative / contradicting reports) observation.

Crucially, we show lay reasoners to be ignorant of the influence of base rates (leading to overestimation of deception, both qualitatively and quantitatively), and structural relations / logic-based negations (again, resulting in deception overestimation). Lay intuitions regarding the impact of corroborative / contradicting testimony on P(Deception) – via explaining away - are (although conservative) shown to qualitatively correspond to normative expectations.

Taken together, this shows erroneous inferences are highest when dealing with potentially deceptive reports alone (where base rates, conditional probabilities, and logical structure are the only active elements to integrate), but accuracy improves when a reference point (other reports / observations) comes into play. This suggests a note of caution for investigative domains in which deception is a possibility (e.g. intelligence analysis), where estimation errors are likely to be substantial until independent evidence (e.g. corroborating testimony) is gathered.

Further work is proposed to incorporate inaccurate / long-run deception motives (i.e. δ), something that we argue may be captured in the present formalism.

Open Practices

All data and materials have been made publicly available via the Open Science Framework at https://osf.io/4hvu6/.

References

Briñol, P., & Petty, R. E. (2009). Source factors in persuasion: A self-validation approach. *European Review of Social Psychology*, *20*(1), 49–96.

Chaiken, S., & Maheswaran, D. (1994). Heuristic Processing Can Bias Systematic Processing: Effects of Source Credibility, Argument Ambiguity, and Task Importance on Attitude Judgement. *Journal of Personality and Social Psychology*, *66*(3), 460–473.

Cuddy, A. J. C., Glick, P., & Beninger, A. (2011). The dynamics of warmth and competence judgments, and their outcomes in organizations. *Research in Organizational Behavior*, *31*, 73–98.

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in psychology*, 5, 1-17.

Earle, T. C., Siegrist, M., & Gutscher, H. (2010). Trust, risk perception and the TCC model of cooperation. In *Trust in risk management: Uncertainty and scepticism in the public mind* (pp. 1–50).

Fenton, N., Neil, M., & Lagnado, D. A. (2013). A general structure for legal arguments about evidence using Bayesian networks. *Cognitive science*, *37*(1), 61-102.

Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: warmth and competence. *Trends in Cognitive Sciences*, *11*(2), 77–83.

Hahn, U., Harris, A. J. L., & Corner, A. (2009). Argument content and argument source: An exploration. *Informal Logic*, *29*(4), 337–367.

Harris, A. J. L., Hahn, U., Madsen, J. K., & Hsu, A. S. (2015). The Appeal to Expert Opinion: Quantitative Support for a Bayesian Network Approach. *Cognitive Science*, *39*(7), 1–38.

Højsgaard, S. (2012). Graphical independence networks with the gRain package for R. *Journal of Statistical Software*, *46*(10), 1-26.

JASP Team (2018). JASP (Version 0.9)[Computer software].

Jeffreys, H. (1961). *Theory of probability* (3rd Ed.). Oxford, UK: Oxford University Press.

Lagnado, D. A., Fenton, N., & Neil, M. (2013). Legal idioms: a framework for evidential reasoning. *Argument & Computation*, *4*(1), 46-63.

Madsen, J. K., Hahn, U., & Pilditch, T. D. (2018). Partial source dependence and reliability revision: the impact of shared backgrounds. In T.T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 722-727). Austin, TX: Cognitive Science Society.

Pettigrew, R. (2016). *Accuracy and the Laws of Credence*. Oxford University Press.

Petty, R. E., & Cacioppo, J. T. (1984). Source Factors and the Elaboration Likelihood Model of Persuasion. *Advances in Consumer Research*, *11*, 668–672.

Schul, Y., Mayo, R., Burnstein, E., & Yahalom, N. (2007). How people cope with uncertainty due to chance or deception. *Journal of Experimental Social Psychology*, *43*(1), 91-103.

Schul, Y., & Peri, N. (2015). Influences of Distrust (and Trust) on Decision Making. *Social Cognition*, *33*(5), 414–435.

Siegrist, M., Gutscher, H., & Earle, T. (2005). Perception of risk: the influence of general trust, and general confidence. *Journal of Risk Research*, *8*(2), 145–156.

Vrij, A. (2008). *Detecting lies and deceit: Pitfalls and opportunities*. John Wiley & Sons.

Zero-sum reasoning in information selection

Toby D. Pilditch[1,2], Alice Liefgreen[1], and David Lagnado[1]

[1]Department of Experimental Psychology, University College London, 26 Bedford Way, London, WC1H 0AP, UK
[2]University of Oxford, School of Geography and the Environment, South Parks Road, Oxford, OX1 3QY, UK

Abstract

Recent research (Pilditch, Fenton, & Lagnado, 2019) shows that people are susceptible to zero-sum thinking in evidence evaluation, where they dismiss or underweight the probative value of evidence that is equally predicted by multiple independent hypotheses. But such an assumption is only valid when explanations are mutually exclusive and exhaustive. The present work extends these findings by looking at the context of information selection, and the decisional consequences of the zero-sum fallacy. It uses an information metric to quantify the cost of the error in terms of overlooked information.

Keywords: zero-sum; evidential reasoning; probabilistic reasoning; Bayesian Networks; belief updating

Introduction

When reasoning under uncertainty, the search and selection of evidence is fundamental to accurate and efficient prediction and diagnosis. Whether in formal investigative domains such as medical diagnosis, forensics, or intelligence gathering, or in everyday reasoning, we often have to search out information to make inferences about a target hypothesis (e.g. which test to conduct? Which source to query? Etc.). To address these questions, reasoners must consider the prospective "value" or *information* provided by new evidence. These estimates are often fraught with biases and errors (e.g. Jones & Sugden, 2001; Nelson, McKenzie, Cottrell & Sejnowski, 2010; Slowiaczek, Klayman, Sherman & Skov, 1992) making accurate choice of what evidence to gather a non-trivial task for lay reasoners.

In the present work, we explore the question of evidence selection in the context of a novel evidential reasoning fallacy, the zero-sum error (Pilditch, Fenton, & Lagnado, 2019), where reasoners assume that evidence which is equally predicted by multiple alternative hypotheses is non-probative. We explore whether this error also drives similar errors in information choice, in particular whether it leads to people overlooking the most useful evidential tests. We explore the mechanisms that might underpin this reasoning fallacy. Furthermore, we highlight the methodological and theoretical value of incorporating information measures into our understanding of how reasoners navigate more complex reasoning structures.

The Zero-sum fallacy

When reasoning about evidence that is equally predicted by two independent explanations, lay reasoners tend to assume that this evidence offers no support to either hypothesis, because it does not discriminate between them (Pilditch, Fenton, & Lagnado, 2019). However, this assumption is only applicable when the explanations are both mutually exclusive and exhaustive (i.e. exactly one of the explanations is true). In fact, given positive evidence, *both* explanations become more probable. Across a number of experiments, reasoners judged such evidence irrelevant to a target hypothesis, even when the inappropriateness of applying the assumptions of exclusivity and exhaustiveness was highlighted.

The posited mechanism behind this error was a fallacy of considering evidential support between hypotheses to be a "zero-sum" situation: one hypothesis may only gain support (i.e. become more probable) at the detriment of another. To elucidate, reasoners were inclined to dismiss a medical test that could not distinguish between 2 diseases – failing to consider that the positive test result could in fact make the patient having *both* diseases more probable.

Work on the zero-sum fallacy has so far looked at qualitative judgments of support. In building on this work, via the incorporation of alternative evidence options and a measure of the amount of overlooked information given a preference, we seek to quantify the *cost* of this error, and further uncover the mechanism underpinning it.

A Bayesian Framework

To further elucidate the nature of the zero-sum fallacy, and outline the foundational formalism upon which information in the context of reasoning under uncertainty may be built, we briefly highlight the role of Bayesian Networks (BNs; Pearl, 1988; 2009) in evidential reasoning.

BNs are directed acyclic graphs (DAGs) that provide a computational framework for modelling the strength of inferential relationships when reasoning under uncertainty. A BN is made up of nodes that represent the variables of interest, and directed arrows capturing probabilistic dependency relations between variables, quantified by conditional probability tables. The probabilities of the unknown nodes are normatively updated given new evidence using Bayes rule (Pearl, 1988). Consequently, BNs are used as a normative comparison against which human reasoning can be compared (e.g. Pilditch, Fenton, & Lagnado, 2019).

To explain in the zero-sum case, two possible hypotheses, each with their own prior probabilities are represented by separate, *independent* nodes (see H1 and H2 in Fig. 1). This reflects the acknowledged assumptions that the two hypotheses are neither mutually exclusive (i.e. both could be true) nor exhaustive (i.e. both could be false), and there are no direct causal links between them. Critical to the fallacy,

however, is the conditional probability table (CPT) of the evidence that depends on both hypotheses (E1 in Fig. 1). Table 1 below provides an example of how likely the evidence is to be observed, given the possible states of the two hypotheses.

Table 1: Example conditional probability table for "common effect" evidence, given two possible causes, H1 and H2.

E	**¬H1, ¬H2**	**H1, ¬H2**	**¬H1, H2**	**H1, H2**
E = T	0.01	0.9	0.9	0.99
E = F	0.99	0.1	0.1	0.01

The two central columns of Table 1 represent the possibilities that participants making the zero-sum fallacy arguably focus on. More precisely, if one (falsely) assumes that only one of the two hypotheses is true (i.e. they are exclusive and exhaustive), then one is only considering two possibilities: the probability of E given H1 being true (P(E|H1,¬H2); center-left column) or given H2 being true (P(E|¬H1,H2); center-right column). Consequently, by adopting this narrow focus, the evidence appears to be *equally predicted* by each possibility (P(E|H1,¬H2) = P(E|¬H1,H2) = 0.9) suggests the evidence is non-probative.

Critically, this reasoning neglects two important possibilities: first, the fact that evidence *could* still occur when neither hypothesis is true (P(E|¬H1,¬H2) > 0) – i.e. the hypotheses are *not* exhaustive explanations of the evidence. Second, that not only is there the possibility that both hypotheses are true i.e. the hypotheses are not exclusive, but that when both *are in fact true*, this results in an even greater probability of observing the evidence (i.e. P(E|H1,H2) > (P(E|H1,¬H2) | P(E|¬H1,H2))). Thus, when making the diagnostic inference from observed evidence to probable hypotheses, *both* H1 and H2 become more probable, given E.

Information Search

In the real world people are habitually required to *actively* seek and acquire information in order to make a decision, causal inference or judgement, and do not merely act as passive observers of their surroundings. Within the psychological literature, measures have been proposed to quantify the informative value of a piece of evidence and the exploration of people's information search behaviour in a variety of contexts (for an overview, see Nelson, 2008). Here we adopt the Kullback-Liebler Divergence (KL-D; Kullback & Liebler, 1951) as a quantitative measure of the expected informative value of different pieces of evidence given a defined probabilistic environment. KL-D is a form of relative entropy and assigns high informative value to evidence that reduces uncertainty the most, entailing the largest divergence between prior and posterior probability distributions (Nelson, 2008). Formally, it quantifies the subjective expected usefulness of evidence before the state of the evidence is known as:

$$KL(E_i) = \sum_{H_j} P(H_j|a_i) \log \frac{P(H_j|a_i)}{P(H_j)}$$

Where E_i is an item of evidence within a set $\{E_1, E_{2...}E_i\}$, H is a set of hypotheses, $\{H_1, H_{2...}H_j\}$ and a_i is a set of possible states of the evidence, $\{a_1, a_2, a_i\}$. This quantification enables not only the evaluation of whether people have a preference for evidence with the highest information value, but also allows for a quantitative measure of the amount of *overlooked* information (as a consequence of sub-optimal search behaviour). This approach directly addresses how violations of normative measures of the value of information relate to known violations of normative models of evidence evaluation such as the zero-sum fallacy. Or more informally, puts an explicit value on the cost of the error.

Present Work

As mentioned above, the goal of the present work is to investigate the zero-sum fallacy further, via the inclusion of information search. To do this we expand the previous zero-sum fallacy model (two hypotheses, H1 and H2, with a single, shared piece of evidence, E1) to include an alternative evidence option (E2) – only explainable by the target hypothesis.

In this way, the reasoning probe shifts from an explicit evaluation of whether E1 provides any support for H1, to a decision-making preference between two evidence items: E1, which has an alternative explanation H2 (and thus invites the zero-sum error), and E2, with no alternative cause represented in the model. To explore the possible influence of zero-sum thinking, and to quantify overlooked information costs, the general structure illustrated in Fig. 1 required populating with several different sets of parameters.

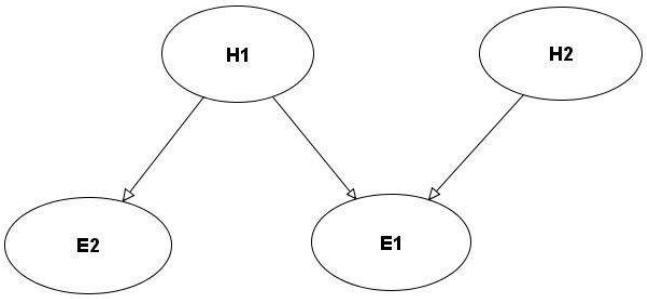


Figure 1. Graphical representation of BN Model.

Four sets of parameters were created (shown in Table 2), each incrementally differing from another, so as to determine the influence of various reasoning components. The prior probabilities of each hypothesis were manipulated as either both rare (P(H1) = P(H2) = .1), both common (P(H1) = P(H2) = .5), or unequal (P(H1) = .5, P(H2) = .1). In this way, the degree to which H2 is providing a "false positive" for E1 (i.e. another explanation for a positive, that is not the hypothesis of interest, H1) is manipulated. This is

of interest to determine whether the zero-sum fallacy is based on the integration of this "false positive" probability (i.e. when H2 is more probable, the zero-sum fallacy is more prevalent), or solely on the *presence* of a possible alternative explanation. Further, the manipulation of 50/50 (or "common") priors can be used to assess whether participants will be more inclined to apply the false assumptions of mutual exclusivity and exhaustiveness that underpin the zero-sum fallacy. Lastly, if the manipulation of unequal priors (P(H1) = .5, P(H2) = .1) resulted in a *reduction* of zero-sum fallacy errors, it would be suggestive of participants using the relative rarity of H2 to *discount it* as an explanation of E1.

In addition, across these three sets, the likelihoods of E1 and E2 were held as unequal, in that E1 was more diagnostic of H1 than E2 (P(E1|H1,¬H2) = .9, vs P(E2|H1) = .6). However, one final parameter set was added in which (along with rare priors) these values were equal across E1 and E2 (P(E1|H1,¬H2) = P(E2|H1) = .8). It should be noted that in all these parameter sets, this results in E1 being the more informative evidence for determining H1, and thus selecting E2 comes at a cost of overlooked information. However, by manipulating the false positive rate of E2 as either high (P(E2|¬H1) = .2/.4), or low (P(E2|¬H1) = .01), we can manipulate between subjects a condition in which E1 is superior (the former), or inferior (the latter), to further determine sensitivity to the parameters underlying the fallacy.

This leads to several predictions: Firstly, there will be a general aversion to selecting E1 (i.e. the decision analogue of a zero-sum fallacy). Secondly, participants will be sensitive to parameter manipulations, such that when E2 is manipulated as more diagnostic (e.g. P(E2|¬H1) = .01 condition), aversion to E1 / preference for E2 will (correctly in this instance) increase. Conversely, when parameter manipulations in fact favour E1 (e.g. equal likelihoods parameter set) participants will (falsely) remain aversive to it.

Method

Participants 180 US participants were recruited and participated online through the Amazon Mechanical Turk platform. Participants were native English speakers (leading to 2 exclusions), with a mean age of 35.88 (*SD* = 10.5), and 90 participants identified as female. All participants gave informed consent, and were paid $1.20 for their time (*Median* = 12.75 minutes, *SD* = 9.62).

Procedure & Design Participants were shown 4 scenarios in a randomized order. These scenarios all originated from the model structure of Fig. 1, to include a target hypothesis (H1), evidence that may inform on the hypothesis (E1), but may also be explainable by an alternative hypothesis (H2), and finally an alternative evidence item only dependent on H1, and not H2 (E2). The scenario contexts were an arson case (identifying an accelerant), a conservation case (tracking a target species), a medical diagnosis case (confirming a brain tumor), and a digital forensics case (identifying a cyberattack culprit).

Crucially, along with the structure of Fig. 1, contexts were also furnished within the text with sufficient parameter details to fully populate a Bayesian Network model of the scenario. These included the priors for each hypothesis (P(H1) and P(H2), the likelihoods for each evidence-hypothesis relationship (P(E1|H1,¬H2), P(E1|¬H1,H2), and P(E2|H1)), and false positives - P(E1|¬H1,¬H2) and P(E2|¬H1). The latter of these parameters (E2 false positive) was manipulated between subjects, as a method of shifting the balance of expected information between E1 and E2. The remaining parameters were deployed as 4 "sets" (see Table 1 below), each designed to test particular parameters trade-offs, and randomly allocated to scenario contexts.[1]

Table 2. Parameter sets, allocated across scenario contexts.

	Parameter Sets			
	RareP. EqL	RareP. UneqL	UneqP. UneqL	ComP. UneqL
P(H1)	.1	.1	.5	.5
P(H2)	.1	.1	.1	.5
P(E1|H1,¬H2)	.8	.9	.9	.9
P(E1|¬H1,H2)	.8	.9	.9	.9
P(E1|¬H1,¬H2)	.01	.01	.01	.01
P(E2|H1)	.8	.6	.6	.6
P(E2|¬H1)	.01 / .2	.01 / .4	.01 / .4	.01 / .4
Information				
KL(E1)*	0.12	0.135	0.27	0.06
KL(E2)	0.22/0.06	0.16/0.005	0.268/0.01	0.268/0.01
KL(E1 – E2)	-0.1/0.06	-0.026/0.13	0.002/0.25	-0.205/0.05

*Only takes into account H1

For each scenario, participants answered the following questions:

Priors: Participants were asked to provide the prior probabilities of H1 and H2 (i.e. *before* observing any evidence). Although participants had already been provided with prior probabilities for H1 and H2, by also eliciting these prior probabilities any participant-based assumptions could be incorporated into the models used for normative comparisons. More precisely, for each participant, elicited priors were used to outfit a Bayesian Network fitting the structure of Fig. 1 (and the remaining parameters drawn from the parameter set being tested), using the gRain package in R (Højsgaard, 2012). These individually fitted BNs (hereafter termed Behaviorally Informed Bayesian Networks; BIBNs) thus provided a fitted normative comparison for participant inferences on the participant by parameter set level. BIBNs were not only then used to generate predicted responses, but also to calculate the informative value (KL-D) of each item of evidence, given

[1] P(E|H1, H2), though not provided explicitly to participants, is based on an assumption of a noisyOR function (see Pearl, 1988), which is based on the reasonable assumption that causes H1 and H2 are independent.

that model – essential for calculating any forgone information.

Preference: Participants were then asked "Which test (evidence item) would you prefer, so as to best determine [H1]?" This qualitative judgment was forced choice [E1 / E2 / "They are the same."]

Confidence in preference: Following the qualitative evidence preference, participants were asked to provide a confidence in that preference ("How confident are you that your response is correct?" 0-100%).

Other DVs: Although posterior probability estimates for each evidence item ("Probability of [H1] ***only*** **given a positive [E1]**" 0 - 100%; "Probability of [H1] ***only*** **given a positive [E2]**" 0 - 100%), and open text reasoning responses were collected, for the sake of brevity, these results are not reported here.

Results

Using the JASP statistical software (JASP Team, 2018), Bayesian statistics were employed throughout[2].

Evidence Preferences

Overall, binomial tests comparing evidence preferences to chance (.33) found the evidence with a single possible cause (E2) to be preferred at a rate decisively greater than chance (.54, $N = 712$), $BF_{10} = 3.06 * 10^{26}$, whilst preferences for the evidence with two potential cause (E1) were no different than chance, (.35, $N = 712$), $BF_{10} = 0.083$, and preferences for "They are the same." occurred decisively less often than expected by chance (.11, $N = 712$), $BF_{10} = 4.59 * 10^{37}$. Further, a contingency table comparing observed to predicted preferences found decisive evidence for these preferences deviating from normative expectation ($N = 1424$), $BF_{10} = 1.196 * 10^{25}$. Importantly, there was a null influence of the potential confounds of scenario order ($N = 712$), $BF_{10} = 0.109$, or scenario context ($N = 712$), $BF_{10} = 5.087 * 10^{-5}$.

In line with expectations, when the false positive rate of E2 was low (.01), and thus sensitivity was higher, then E2 was preferred substantially more often (and E1 less often) than when the false positive of E2 was high (N = 712), $BF_{10} = 4.068$.

Turning next to parameter sets (rows of Fig. 2), we break down the analysis for each set to determine a) the dominant participant preference, and b) whether this deviates from the normative predictions for that set. This split by parameter set is motivated by the potential sensitivity of participants to particular combinations of parameters (e.g. equal likelihoods, or unequal priors).

[2]All analyses assumed an uninformed prior. Bayes Factors (BFs), are interpreted as: 1 – 3 = anecdotal support; 3-10 = substantial; 10-30 = strong; 30-100 = very strong; >100 = decisive (Jeffreys, 1961). Conversely, Bayes Factors < .33 are considered substantial support for the *null* (Dienes, 2014).

Rare Priors, Equal Likelihoods. When both H1 and H2 priors were rare, and evidence likelihoods were equal, participants chose E2 at levels decisively above chance (.612, N = 178), $BF_{10} = 6.729 * 10^{11}$, and E1 significantly less than chance (.23, N = 178), $BF_{10} = 5.565$. This runs contrary to model predictions, where E1 is preferred decisively above chance level (.674, N = 178), $BF_{10} = 1.035 * 10^{18}$, and E2 at no different than chance (.326, N = 178), $BF_{10} = 0.088$. This is further corroborated by a contingency table analysis which finds decisive evidence for a deviation of participant choices from normative expectation (N = 356), $BF_{10} = 6.729 * 10^{11}$.

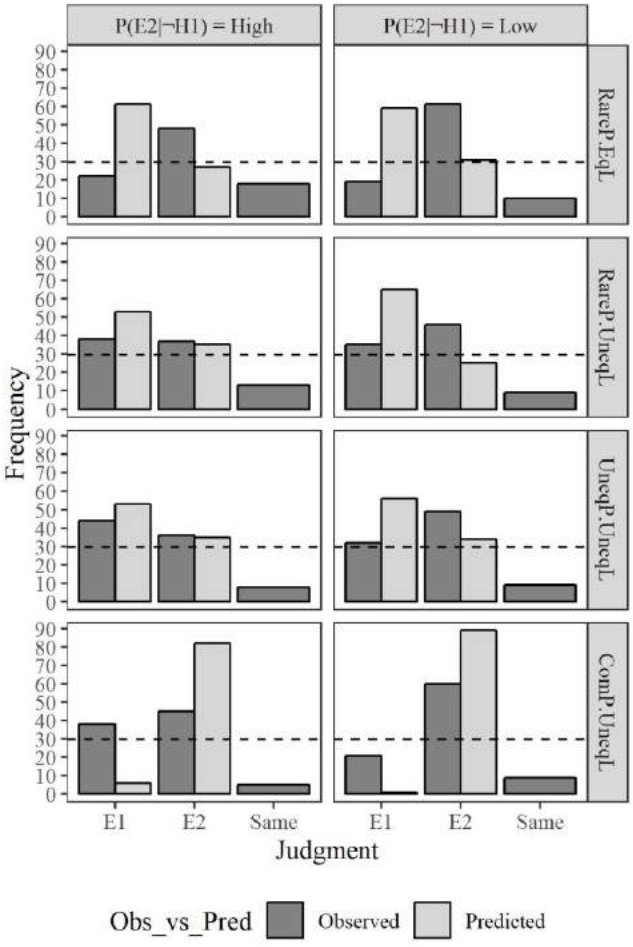


Figure 2. Evidence choice frequencies across parameter sets (rows) and condition (columns).

Rare Priors, Unequal Likelihoods. When priors are rare, and evidence likelihoods are unequal (E2 at .6, and E1 at .9), we again find the same pattern. Participants choose E2 at above chance levels (.466, N = 178), $BF_{10} = 111.88$, and E1 no different than chance (.41, N = 178), $BF_{10} = 1.115$. Once again, however, model predictions show the opposite pattern, with E1 choices above chance level (.663, N = 178), $BF_{10} = 6.223 * 10^{16}$, and E2 choices no different than chance, (.337, N = 178), $BF_{10} = 0.09$. This is again

corroborated by the decisive deviation between participants and their model predictions found by contingency table analysis (N = 356), $BF_{10} = 1.473 * 10^7$.

Unequal Priors, Unequal Likelihoods. When both priors (P(H1) = .5; P(H2) = .1) and likelihoods are unequal, we find the same general trend, albeit to a lesser degree. More precisely, although participant choices for E2 are again greater than chance (.478, N = 178), $BF_{10} = 368.73$, choices for E1 are also just above chance level (.427, N = 178), $BF_{10} = 3.499$. However, model predictions again show a decisive preference for E1 (.612, N = 178), $BF_{10} = 6.729 * 10^{11}$, whilst E2 should be preferred no more often than chance (.388, N = 178), $BF_{10} = 0.335$. This insufficiency of E1 choices is again captured by the decisive difference in judgment proportions when comparing participants and model predictions in a contingency table (N = 356), $BF_{10} = 15735.87$.

Common Priors, Unequal Likelihoods. Turning finally to when priors are both common (.5) and likelihoods are unequal, we see the same behavioral pattern of a preference for E2 above chance level (.59, N = 178), $BF_{10} = 7.748 * 10^9$, and E1 no different than chance (.331, N = 178), $BF_{10} = 0.088$. However, unlike the preceding parameter sets, E2 is also chosen above chance level by model predictions (.961, N = 178), $BF_{10} = 1.996 * 10^{69}$, whilst E1 is in fact chosen decisively less than chance (.039, N = 178), $BF_{10} = 7.246 * 10^{18}$. Further, participant choices for E2 are shown to be insufficient compared to model predictions (N = 356), $BF_{10} = 9.44 * 10^{14}$. This is likely due to the high E1 "false positive" due to marginalization over high H2 probability, making E1 comparatively less diagnostic of H1.

Confidence in evidence preferences. Confidence was generally high across all preferences (M = 66.00, SD = 23.96). Although a Bayesian repeated measures ANOVA revealed confidence to be unaffected by preference, $BF_{Inclusion} = 0.781$, or parameters, $BF_{Inclusion} = 1.064$, but there was strong evidence for confidence being higher in the E2 false positive rate = low condition (M = 68.98, SD = 23.39), rather than high (M = 62.95, SD = 24.18), $BF_{Inclusion} = 11.377$. This finding fits with an easier E2 preference when it is a more sensitive test.

Overlooked information

To elucidate the information cost of the above deviations from normative expectation, for each BIBN model (i.e. each participant-fitted model) the expected informative value (in KL-D) was calculated for E1 and E2. In this way, if a participant selected the evidence with the highest KL-D as predicted by their model, they had not overlooked any information, and thus scored 0. However, if participants selected the less informative evidence, then the overlooked information was the difference (in KL-D) between the optimal (i.e. most informative) evidence and their selected option.[3]

As Table 3 indicates, across all break-downs of evidence choices (overall, by condition, and by parameter set), there was a decisive amount of information overlooked – calculated via Bayesian one sample t-tests (test value = 0). This significant amount of overlooked information can be attributed to the sub-optimal undervaluing of E1 (i.e. the zero-sum fallacy) in all cases barring common priors, unequal likelihoods (bottom row, Table 3). In this latter parameter set, E2 in fact yielded the most information, but was not chosen sufficiently often across participants.

Table 3. Overlooked information; overall, split by condition, and split by parameter sets.

	M	*SD*	*N*	*>0 (BF_{10})*	*δ*	*δ 95% CI*
Overall	.045	.048	712	$5.79*10^{95}$	0.934	.847, 1.021
P(E2|¬H1) = L	.042	.047	360	$1.23*10^{44}$	0.886	.766, 0.999
P(E2|¬H1) = H	.048	.049	352	$2.33*10^{50}$	0.980	.849, 1.113
RareP.EqL	.052	.045	178	$3.12*10^{31}$	1.153	.956, 1.348
RareP.UneqL	.047	.042	178	$2.543*10^{30}$	1.121	.938, 1.318
UneqP.UneqL	.046	.057	178	$6.600*10^{17}$	0.795	.623, 0.958
ComP.UneqL	.034	.045	178	$2.569*10^{16}$	0.753	.59, 0.925

Conclusions

Previous work has shown that evidence equally predicted by multiple explanations is often erroneously dismissed due to the misplaced assumption that support for one hypothesis (of interest) must come at the detriment of another (the zero-sum fallacy; Pilditch, Fenton, & Lagnado, 2019). In the present work, we show that this fallacy results in poor decisions regarding evidence selection, and that such selections come at a quantified cost of overlooked information. Crucially, we also show that participants are sensitive to priors and likelihoods parameters, with different evidence preference patterns as a consequence. However, the general pattern of overlooked information holds despite this sensitivity.

Foremost, the present work confirms the presence of zero-sum reasoning, showing that it is active in people's choice of which evidence to examine. It also highlights the potential costs of the fallacy, via the quantification of (costly) overlooked information. In this way, we argue for the inclusion of different question methods and information measures when investigating reasoning errors – whether across simple or complex structures. This would not only contribute to understanding how violations of normative frameworks of human information acquisition relate to known violations of information evaluation, such as the

[3] If evidence items were equally informative, then participants were pragmatically correct, in terms of information, with any preference (including "They are the same"), and thus scored 0. However, if participants erroneously judged the evidence items the same, the amount of overlooked information was taken from the KL-D of the most informative option.

zero-sum fallacy, but it would allow for the exploration of how consequential sub-optimal evidence selection choices are, in laboratory as well as real-world settings.

Given that information-seeking is a critical aspect of so many areas of decision making – including intelligence analysis, legal reasoning, and medical diagnosis – the use of zero-sum reasoning is a strong concern. Future work will seek ways to alleviate this bias, and shift people towards more normative information gathering.

References

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in psychology*, 5, 1-17.

Højsgaard, S. (2012). Graphical independence networks with the gRain package for R. *Journal of Statistical Software*, *46*(10), 1-26.

JASP Team (2018). JASP (Version 0.9.1)[Computer software].

Jeffreys, H. (1961). *Theory of probability* (3rd Ed.). Oxford, UK: Oxford University Press.

Jones, M., & Sugden, R. (2001). Positive confirmation bias in the acquisition of information. *Theory and Decision*, 50(1), 59-99.

Kullback, S., & Leibler, R. A. (1951). On Information and Sufficiency. *The Annals of Mathematical Statistics*. Advance online publication.

Nelson, J. D. (2008). Towards a rational theory of human information acquisition. In M. Oaksford & N. Chater (Eds.), *The probabilistic mind: Prospects for rational models of cognition* (pp. 143–164). Oxford, United Kingdom: Oxford University Press.

Nelson, J. D., McKenzie, C. R., Cottrell, G. W., & Sejnowski, T. J. (2010). Experience matters: Information acquisition optimizes probability gain. *Psychological science,* 21(7), 960-969.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems*. San Francisco, CA: Morgan Kaufmann.

Pearl, J. (2009). *Causality. Models, reasoning, and inference*. Second edition. New York: Cambridge University Press.

Pilditch, T. D., Fenton, N., & Lagnado, D. (2019). The zero-sum fallacy in evidence evaluation. *Psychological Science*, 1-11. DOI: 10.1177/0956797618818484

Slowiaczek, L. M., Klayman, J., Sherman, S. J., & Skov, R. B. (1992). Information selection and use in hypothesis testing: What is a good question, and what is a good answer? *Memory & Cognition,* 20(4), 392-405.

The effect of semantic relatedness on associative asymmetry in memory

Vencislav Popov
Carnegie Mellon University, Pittsburgh, Pennsylvania, United States

Qiong Zhang
Carnegie Mellon University , Pittsburgh, Pennsylvania, United States

Griffin Koch
University of Pittsburgh, Pittsburgh, Pennsylvania, United States

Regina Calloway
University of Pittsburgh, Pittsrbugh, Pennsylvania, United States

Marc Coutanche
University of Pittsburgh, Pittsburgh, Pennsylvania, United States

Abstract

We provide new evidence concerning two views of episodic associations: The independent associations hypothesis (IAH) posits that associations are unidirectional and separately modifiable links (AB; AB); the associative symmetry hypothesis (ASH) considers the association to be a holistic conjunction of A and B representations. While existing literature focuses on tests that compare the correlation of forward and backward associations and favors ASH over IAH, we provide the first direct evidence of IAH by showing that forward and backward associations are separately modifiable for semantically related pairs. In two experiments, participants studied 30 semantically unrelated and 30 semantically related pairs intermixed in a single list, and then performed a series of up to eight cued-recall test cycles. All pairs were tested in each cycle, and the testing direction (A-? or B-?) alternated between cycles. Consistent with prior research, unrelated pairs exhibited associative symmetry accuracy and response times improved gradually on each test, suggesting that testing in both directions strengthened the same association. In contrast, semantically related pairs exhibited a stair-like pattern, where performance did not change from odd to even tests when the test direction changed; it only improved between tests of the same direction. We conclude that episodic associations can have either a holistic representation (ASH) or separate directional representations (IAH), depending on the semantic relatedness of their constituent items.

Word frequency affects binding probability not memory precision

Vencislav Popov
Carnegie Mellon University, Pittsburgh, Pennsylvania, United States

Matt So
Carnegie Mellon University, Pittsburgh, Pennsylvania, United States

Lynne Reder
Carnegie Mellon University, Pittsburgh, Pennsylvania, United States

Abstract

Normative word frequency has played a key role in the study of human memory, but there is little agreement as to the mechanism responsible for its effects. To determine whether word frequency affects binding probability or memory precision, we examined working memory for spatial positions of words. Each of three experiments included 300 trials in which five words were presented sequentially around an invisible circle followed by one of those words shown in the middle of the circle as a probe to test its location. Participants had to click on the associated location and the degree of error around the circle was the dependent measure. Across experiments we varied word frequency, presentation rate and the proportion of low frequency words on each trial. A mixture model dissociated memory precision, binding failure and guessing rates from the continuous distribution of errors. On trials that contained only low- or high-frequency words, low-frequency words lead to a greater degree of error in recalling the associated location. This was due to a higher word-location binding failure and not due to differences in memory precision or guessing rates. Slowing down the presentation rate eliminated the word frequency effect by reducing binding failures for low-frequency words. Mixing frequencies in a single trial hurt high-frequency and helped low-frequency words, but frequency composition and presentation rate did not interact. These findings support the idea that low-frequency words require more resources for binding and that the binding fails when these resources are insufficient.

Exploring the role that encoding and retrieval play in sampling effects

Keith Ransom (keith.ransom@adelaide.edu.au)
School of Psychology, University of Adelaide
Amy Perfors (amy.perfors@unimelb.edu.au)
School of Psychological Sciences, University of Melbourne

Abstract

A growing body of literature suggests that making different sampling assumptions about how data are generated can lead to qualitatively different patterns of inference based on that data. However, relatively little is known about how sampling assumptions are represented or when they are incorporated. We report the results of a single category generalisation experiment aimed at exploring these issues. By systematically varying both the sampling cover story and whether it is given *before* or *after* the training stimuli we are able to determine whether encoding or retrieval issues drive the impact of sampling assumptions. We find that the sampling cover story affects generalisation when it is presented before the training stimuli, but not after, which we interpret in favour of an encoding account.

Keywords: categorisation; generalisation; memory; sampling assumptions;

Introduction

For most of the reasoning tasks with which we are routinely faced, it is impossible to draw conclusions that are logically entailed by what we know already. Instead, we must by necessity make inductive generalisations on the basis of the limited data we have. In order to make the most of that data, it is important to accurately assess its evidentiary weight – to recognise precisely what kind of generalisations it supports. Doing this assessment accurately depends on understanding the context in which it was observed.

To illustrate why, imagine that you need to buy a present for a colleague as a part of your workplace Secret Santa. You don't know this colleague that well, but while helping them move offices you see a box containing the CDs that they listen to while at work. Sensing an opportunity to re-gift an unwanted copy of *Taylor Swift*, you take a closer look. Upon realising that almost all of their collection consists of 80s Billboard Hits, you conclude that their musical taste is dated[1] and reluctantly decide that Taylor Swift is not for them.

Suppose, instead, that you had seen the exact same data (a box of CDs) but in the context of helping your colleague move their entire music collection – many dozens of boxes worth – and that box just happened to be the only open one. Now the same data is no longer quite so representative: instead of being a carefully culled and chosen set of favourites, it is one of many. Thus, it tells you much less about whether your colleague would like Taylor Swift.

As this example illustrates, knowing something about why one saw the data that one did (and not some other data) enables people to make more valid inferences. Put another way, being able to reason about the generative process behind a set of observations tells people about the weight of evidence that those observations supply. These assumptions about the generative process are often referred to as the *sampling assumptions* that people bring to inference problems. Different sampling assumptions appear to drive qualitatively distinct patterns of generalisation (e.g. Hendrickson, Perfors, Navarro, & Ransom, 2019; Hayes, Navarro, Stephens, Ransom, & Dilevski, 2019), support epistemic trust (Shafto, Eaves, Navarro, & Perfors, 2012) and epistemic vigilance (Landrum, Eaves, & Shafto, 2015; Ransom, Voorspoels, Perfors, & Navarro, 2017), fuel pragmatic implicature (Goodman & Frank, 2016), and promote accelerated learning (Shafto, Goodman, & Griffiths, 2014).

Despite this wealth of empirical support for the utility and importance of sampling assumptions in generalisation, little is known about either how they affect the encoding and retrieval of the data, or how they affect people's mental representations. Is the evidentiary weight of data under a given sampling assumption computed only at the point at which the data is later retrieved? Or is it encoded at the time of learning, thus shaping the underlying representation from the beginning? And how is inference affected as people's memories of the data begin to fade?

Using a single-category learning task, we explore these questions here for the first time. We manipulate both the sampling assumptions people make about the training data (via cover story) as well whether that cover story is available before or after learning. As we explain in the next section, if sampling assumptions affect generalisation at retrieval, we expect no difference in performance regardless of when the cover story was revealed. Conversely, if they affect how the data are encoded, we expect different patterns of generalisation depending on when the cover story was available.

Sampling assumptions and inductive generalisation

The Bayesian generalisation approach of Tenenbaum and Griffiths (2001) provides a useful framework for our research question. In the context of our single category generalisation experiment, we are interested in how the learner decides whether or not to extend the target category c to a novel item y on the basis of previously observed examples x. Within the framework, this decision is assumed to be probabilistic, based on the available evidence. That is:

$$P(y \in c|x,s) = \sum_{h \in \mathcal{H}_c : y \in h} P(h|x,s) \quad (1)$$

where s represents the learner's assumption about the process generating the data x, and $\mathcal{H}_c$ represents the set of alternative hypotheses the learner considers concerning the true extent

[1] The fact that your colleague still uses CDs may have told you this already.

of the category c.[2] In other words, the evidence in favour of category membership is effectively combined across all hypothetical versions of the category containing the novel item. Using a straightforward application of Bayes' rule the term $P(h|x,s)$ may be expressed as:

$$P(h|x,s) \propto P(x|h,s)P(h). \quad (2)$$

This formulation assumes, for simplicity, that the learner entertains a single sampling assumption (i.e. $P(s) = 1$), which we presume was given to them by a cover story describing the generative process.

It is the likelihood function $P(x|h,s)$ that is critical for our current purposes. Substituting different likelihood functions into this system of equations yields different predictions about the way that people generalise from given data. For instance, strong sampling implies a likelihood that embodies the size principle, such that each subsequent datapoint serves as evidence to further tighten one's generalisations around the data; weak sampling uses a different likelihood which implies no such tightening (Tenenbaum & Griffiths, 2001). Thus, the likelihood may be thought of as representing different ways of calculating the weight of evidence that the data provides for the hypothesis under a given sampling assumption.

Our first question here is *when* the likelihood is calculated: when the data is first encoded, or when it is retrieved? If learners do not need to rely on their memories and the sampling cover story is available from the beginning, it is impossible to disentangle these two possibilities. However, if we manipulate when participants are aware of how the data were sampled (i.e., before or after learning), then different possibilities yield different predictions. We consider two main possibilities in detail.

Retrieval. If the likelihood is calculated upon retrieval, then encoding need only involve storing the raw data x in some form. The likelihood calculation would be shaped by whatever sampling assumption was in play during retrieval, regardless of what was assumed during learning. In this sense, the calculation would resemble the conventional or "idealised" interpretation of the Bayesian generalisation model. However, while the conventional interpretation assumes perfect recall of exemplars, a failure to retrieve some data would imply that the likelihood calculation was effectively over a reduced dataset (i.e., smaller sample size). The precise effect that this has will depend on the sampling assumption and on the particular items forgotten. For example, if the diversity of the dataset is largely unaffected by the failure to retrieve certain items, then generalisation under a strong sampling assumption should be wider in this case than under perfect recall. Under weak sampling, in contrast, it is the diversity of the sample and not its size that has an effect on generalisation; thus, a reduction in sample size without a change in diversity would mean that generalisation was unaffected. More generally, as the level of retrieval failure increases, the Bayesian model predicts generalisation increasingly in line with the prior distribution.

[2] In the case that the data x varies over a continuous dimension $\mathcal{H}_c$ will represent a continuum of hypotheses and the sum is replaced with an integral.

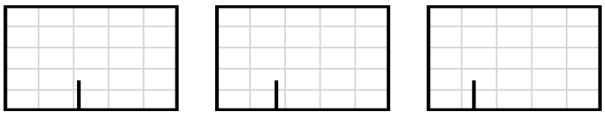


Figure 1: **Example stimuli**. Items varied only in the position of the short black vertical line along the bottom edge of the rectangle.

Encoding. If the likelihood is calculated upon encoding, then the strength of evidence that it represents would have to be stored in some way. In this case, the precise effect of later retrieval failure might vary depending on *how* evidence is encoded. For example, if evidence is stored and retrieved with each exemplar individually then failure to retrieve a given exemplar would mean that subsequent generalisation operates over a smaller dataset, as in the retrieval account (although, unlike the retrieval account, using the sampling assumption that was in play at the time of encoding). If instead, evidence were stored and retrieved in aggregate form (via the hypotheses, for example) then failure to recall any particular exemplar need not imply that the associated evidence was lost. In this way, generalisation might still proceed with all the available evidence (presuming the same hypotheses were accessed). The details of representation notwithstanding, if the likelihood is calculated and stored during encoding, and not at retrieval, then generalisation would be shaped by the sampling assumptions available during learning, even if those assumptions are changed at retrieval.

Method

Our experiment involved a single-category generalisation task modelled on previous work demonstrating that sample size and sampling cover story affect people's willingness to extend category membership to novel examples (Hendrickson et al., 2019; Ransom, Hendrickson, Perfors, & Navarro, 2018). Although we employed stimuli identical to those used in that experiment, we modified the method of presentation so that each stimulus was removed from screen after a (typically brief) period of self-paced study. Using a consistent experimental framework allows us to directly compare our results with the previous findings, and thus to determine if the effect of sampling assumptions on generalisation changes as the memory of training examples decays.

One of our manipulations involved the nature of the cover story people received. Either they were told that the data was given by a HELPFUL teacher (which corresponds to a strong sampling assumption and implies that generalisations should be tighter) or they were given a cover story implying that it was chosen at RANDOM (which corresponds to a weak sampling assumption and implies that generalisations should be looser). Critically, we manipulated whether people were given the sampling story BEFORE or AFTER they saw the training stimuli. If sampling assumptions affect how

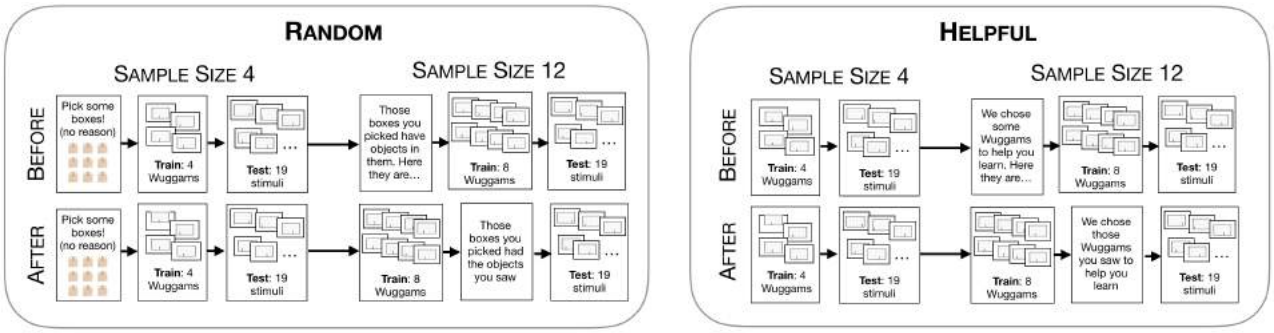


Figure 2: **Experiment design**. Our 2x2x2 design varied Sample Size within-subject and Sampling Explanation and Presentation Sequence between-subjects. All participants began by seeing four individually-presented exemplars followed by a generalisation task to novel stimuli. Those in the RANDOM condition were then given a cover story in which the subsequent eight items were chosen at random from boxes that they themselves had previously selected. Those in the HELPFUL condition were told that the items were selected by a helpful teacher. In the BEFORE condition, the cover story was given before seeing the eight new items; in the AFTER, it came after. In all conditions the experiment ended with a repeat of the generalisation test.

the data are encoded then people should generalise differently depending on when they received the story.

Participants

We recruited 999 people via Amazon Mechanical Turk who were each paid $1.70USD for 5-10 minutes participation. 56% were female, with age varying between 18 and 75 (median: 37 years), drawn predominately from the U.S. population (99%). All participants passed a screening for English language competency prior to participation.

Stimuli

Stimuli were black rectangles containing a vertical black line inside, attached to the bottom edge (see Figure 1). They varied along a single dimension (the *stimulus value*): the horizontal position of the line within the rectangle. Participants were told that this was the way in which stimuli varied. Evenly spaced light grey "guide lines" were drawn within each rectangle in order to improve discriminability. There were 12 training stimuli in total, whose stimulus values ranged from 21% to 43% in increments of 2%. They were divided into two sets corresponding to the two training phases, as described below.

Design and procedure

As shown in Figure 2, our experiment employed a $2 \times 2 \times 2$ mixed factorial design. Two factors (Sampling Explanation and Presentation Sequence) were manipulated between-subjects while another (Sample Size) varied within-subject. People were thus allocated at random to one of four experimental groups.

Across all groups, the experiment involved presenting people with a number of examples of a novel 1D category and then observing whether they generalised category membership to new items based on the examples they had been shown and what they had been told about those examples.

Sample Size To facilitate a baseline against which the effect of additional exemplars could be compared, the experiment involved two rounds of testing. The first (Size 4) occurred after a training phase involving four training examples, and the second (Size 12) after seeing eight more.

Stimuli for the first training phase consisted of the two extreme examples (with values of 21% and 43%) and two others selected at random from the ten whose values lay between the extremes. The eight remaining stimuli formed the second training set and were presented in random order.

Presentation Sequence This between-subjects manipulation varied when the sampling cover story was presented in relation to the second training set. People in the BEFORE condition were told the cover story (RANDOM or HELPFUL, described below) *before* viewing the second set of training items, while people in the AFTER condition were offered the explanation only after all training items had been presented.

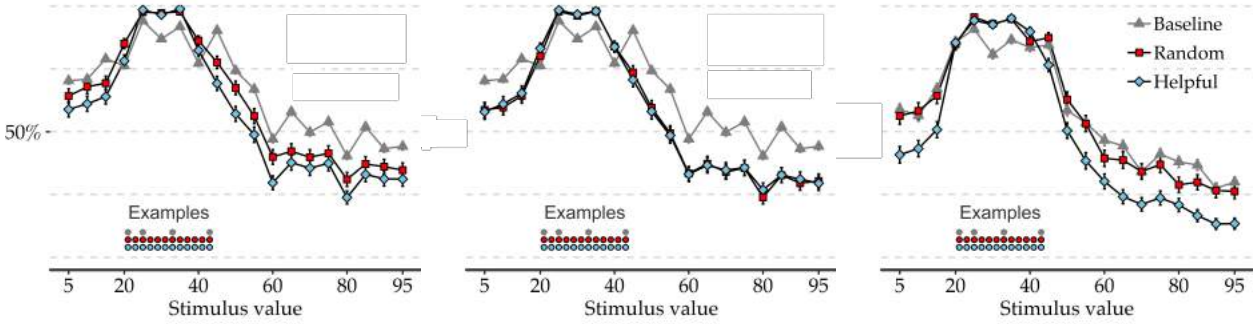
Sampling Explanation The other between-subjects manipulation varied the details of the cover story explaining how the data in the second training phase were generated. The initial training phase, however, was identical for all participants. No explanation was given for how the exemplars were chosen. People were told only that the purpose of the experiment was to see how people judged whether or not unfamiliar objects were in the same category as known examples. In the second training phase people were given one of two different cover stories explaining how the items were selected.

Helpful. People in the HELPFUL condition were told:

> We have a bunch of boxes containing examples of the full variety of «Wuggams». We have chosen 8 of these boxes especially to help you learn the «Wuggam» category, bearing in mind the four training examples we showed you originally.

at which point an array of eight icons resembling open packing boxes were displayed in an adjacent panel. Participants in the BEFORE condition then viewed the eight stimuli one by one. Those in the AFTER condition saw the identical explanation (with verb tenses adjusted) only after all eight stimuli in the second training phase had been shown.

Random. The RANDOM condition was designed to encourage people to believe that each training item was selected



(a) BEFORE. (b) AFTER. (c) Ransom et al. (2018).

Figure 3: Performance on a one category generalisation task as a function of presentation sequence, sampling procedure (manipulated between-subjects) and sample size (manipulated within-subject). The graphs show the proportion of positive responses to the question: "Do you think this object is in the «Wuggam» category?" for each of the test stimuli. People's performance after seeing four examples of the target category with no sampling explanation given (grey line) is contrasted with their performance after seeing all 12 examples and being given an explanation of how the additional examples were selected (black lines). (a) When the sampling explanation was given prior to the presentation of the final 8 examples (BEFORE condition), people tightened their generalisations as more data was observed, but the extent of tightening was affected by the sampling manipulation; those people who actively sampled the additional examples at random (red squares) tightened their generalisation less than those that were told that the items had been selected by a helpful teacher (blue diamonds). (b) In contrast, when the sampling explanation was given only after all training stimuli were presented (AFTER condition), the sampling manipulation had no effect, with people tightening their generalisation equally in both cases. (c) Using the same experimental framework and stimuli, but keeping the training stimuli on-screen during the testing phase, Ransom et al. (2018) demonstrated the effect of sampling manipulation seen only in the BEFORE condition. But when people must rely on their memory of observed examples, their generalisation is wider overall.

at random and that it was at least theoretically possible to see examples not in the target category. To achieve this, people in the RANDOM condition were presented with an additional phase preliminary to the first training round. In this phase, a 6×5 arrangement of packing boxes was displayed on screen, and people were asked to select boxes in any order (but not told why this was necessary). After selecting 11 boxes, people were told that the contents would be revealed later in the experiment. Following this, the first training phase commenced, which was identical for all participants.

During the second training phase, participants in the AFTER condition were immediately shown the eight remaining training items without explanation. Those in the BEFORE condition were told that we had many boxes containing examples from our catalogue, and that these examples included but were not limited to Wuggams. After this, the original array of (closed) boxes was displayed, indicating the ones that the participant had previously selected. People were then told:

> At the start of the experiment we asked you to choose some of these boxes at random. These are the boxes that you selected. We're going to open them now and show you whatever kind of item we find inside.

In order to reinforce the notion that it might have been possible to see items from categories other than Wuggams, the display was updated at this point to reveal eight open boxes and three closed ones. People were told that some of the boxes they had chosen were stuck but that we would show them the contents of the boxes that did open. Participants in the AFTER condition received exactly this cover story (with verb tenses adjusted) only after seeing all eight training examples.

Generalisation test

Immediately after both the first and second training phase, participants in all conditions performed the same generalisation test. In it, they were shown 19 stimuli one at a time in random order; this sequence was repeated four times. The stimuli consisted of 19 items with stimulus values ranging from 5% to 95% in increments of 5%. The test query was a yes or no question: "Do you think this object is in the «Wuggam» category?" Neither training stimuli nor the sampling explanation remained on-screen during testing, requiring people to rely on their memory when making judgements.

Results

Our work is focused on understanding how memory and sampling assumptions interact to affect generalisation. Do we replicate previous findings showing that differences in sampling assumptions lead to differences in generalisation? Does this difference in people's patterns of generalisation change if the sampling manipulation occurs before or after stimulus encoding? We address each question in turn below.

First: do we replicate previous results? Our RANDOM BEFORE and HELPFUL BEFORE conditions are very similar to that of a previous study (Ransom et al., 2018), but are different in one key way. In our version, the training stimuli were removed from the screen after initial presentation; in Ransom et al. (2018) and much of this literature the training stimuli stay visible for the entire experiment. We therefore investigate whether these previously observed effects of sampling manipulation are replicated even when people must rely on their memory of the training stimuli.

To investigate this we first analysed the responses of all participants having seen only the first four exemplars, for which no sampling explanation was given. Against this baseline we separately compared the responses of people in the RANDOM BEFORE and HELPFUL BEFORE conditions. The resulting generalisation curves shown in Figure 3(a) reveal

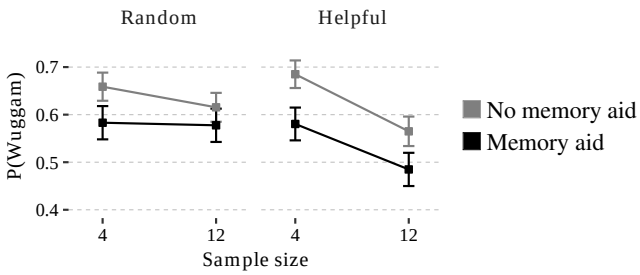
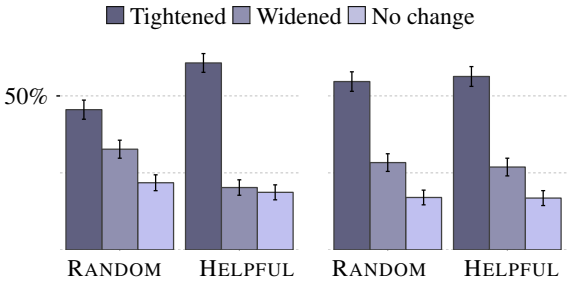


Figure 4: The mean effect of additional exemplars on the marginal probability of generalising the learned category to novel stimuli, as a function of sampling assumption and the presence of a memory aid. When training exemplars remained on-screen throughout the testing phase participants were less willing overall to generalise the target category to novel items than when no memory aid was present. In magnitude, the effect of the memory aid on generalisation was comparable to the effect of observing the eight additional exemplars.



(a) BEFORE (b) AFTER

Figure 5: The proportion of people who either tightened ($\Delta_p < 0$), widened ($\Delta_p > 0$) or showed no change ($\Delta_p = 0$) in their region of generalisation, after seeing additional examples (where Δ_p reflects an individual's change in rates of responding in favour of the learned category). People are grouped according to the explanation they received about the sampling of extra items, and whether it was given before or after the examples themselves. Error bars show standard error of proportion. (a) In the BEFORE condition, where the sampling explanation was given prior to the presentation of the additional examples, the sampling manipulation had an effect. The majority of people who were told that the items had been selected by a helpful teacher tightened their region of generalisation, while the (slight) majority of people in the RANDOM condition, who actively sampled their own additional examples, widened their region of generalisation or showed no change. (b) In contrast, when the sampling explanation was provided after the additional stimuli had been presented (as in the AFTER condition), the majority of people tightened their generalisations regardless of the explanation given.

that the HELPFUL sampling manipulation led to tighter generalisation than the RANDOM manipulation. This replicates a key finding of Ransom et al. (2018), shown in Figure 3(c). To examine the strength of evidence for this finding we analysed generalisation curves for the second test phase (Size 12), calculating the generalisation probability for each person and stimulus separately. A Bayesian ANOVA revealed that a model of generalisation probability including stimulus value and sampling manipulation as predictors is strongly preferred to a model containing stimulus value only ($BF_{10} > 10^6$).

Although we replicated the qualitative difference between sampling conditions, it is evident on visual comparison of Figure 3(a) and (c) that people appeared to generalise further when they had to rely on their memory of the training stimuli. To determine the overall effect that this had on generalisation we calculated the marginal probability of extending category membership to novel items as a function of test phase (4 or 12 items) and sampling manipulation (RANDOM or HELPFUL). We then compared this probability between our experiment (the BEFORE conditions) and Ransom et al. (2018).

The results, shown in Figure 4, demonstrate that the absence of a memory aid had a uniform but significant effect on generalisation overall ($BF_{10} > 10^{100}$).[3] After seeing 12 exemplars, participants in our study (who had no memory aid) showed a willingness to generalise to novel items comparable to participants in Ransom et al. (2018) after seeing only four items that remained on screen throughout. Thus, overall, we find that the *difference* in generalisation according to sampling assumption did replicate, but generalisation was consistently higher when people had to rely on their memory more.

Our second question was whether the effect of sampling manipulation changes when the sampling cover story is given after the training stimuli rather than before. We therefore repeated our analysis for people in the RANDOM AFTER and HELPFUL AFTER conditions, and found that it does: there is no longer a difference in generalisation based on sampling assumption. As Figure 3(b) shows, people tighten their generalisations to a remarkably similar degree across the two conditions, despite the fact that they had opposing sampling cover stories (Bayesian ANOVA now favours the model with stimulus value as the only predictor: $BF_{01} = 42$).

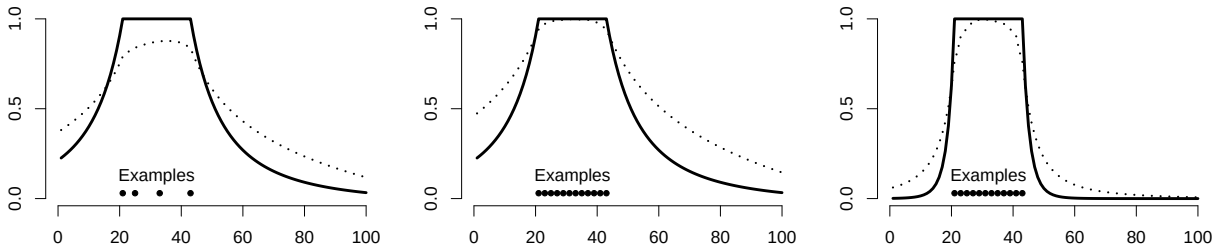
To further assess the effect of our sampling manipulation on the qualitative patterns of responding, we compared each individual's responses between the two test phases, after seeing 4 and 12 exemplars. Figure 5 shows the proportion of people who either tightened, widened or showed no net change in their generalisation (marginalised across test items). Consistent with the patterns at the aggregate level, it is evident that the explanation given to participants regarding the source of the additional exemplars does affect the trajectory of generalisation as more examples are observed. But this explanation only has an effect if it is given before the exemplars are observed ($BF_{10} = 300$) and not after ($BF_{01} = 2.8$).[4]

Discussion

To our knowledge, our work here is the first to explore *when* sampling assumptions affect generalisation, and by extension when the likelihood is calculated. Our results demonstrate that the sampling cover story only had an effect when it was made explicit prior to the presentation of the data. When it was presented at retrieval, then whatever likelihood was the default at the time of encoding (which, in this case, ap-

[3] Based on a Bayesian logistic regression comparing a model of yes/no responses that included stimulus value, sampling manipulation and memory aid as predictors to one without memory aid.

[4] Bayes' factors are based on a multinomial logistic regression comparing a model of qualitative effect (tighten, widen, no net change) with sampling manipulation as a predictor against an intercept only model.



(a) BASELINE. (b) RANDOM. (c) HELPFUL.

Figure 6: Simulated performance on a one category generalisation task as a function of exemplar recall, sampling assumption and sample size. The graphs plot the probability of generalising the learned category as a function of stimulus value. Solid lines represent generalisation performance on the assumption that all exemplars are perfectly recalled at decision time – the default assumption of the Bayesian generalisation model. Dashed lines represent generalisation performance on the basis of imperfect recall. For illustration purposes, the simulation uses an independent probability of recall for each exemplar ($p = 0.5$). Failing to recall exemplars leads to wider generalisation overall. (a) Simulated performance in the BASELINE condition (4 exemplars), assuming the default (strong) sampling. When the sample size is small, the effect of forgetting on generalisation reflects a balance of two forces: the reduction in diversity may reduce generalisation within the range spanned by the exemplars, while the reduced sample size leads to wider generalisation outside the range. (b) Simulated performance in the RANDOM condition (12 exemplars), assuming the BASELINE performance as a prior and that the 8 additional exemplars are weakly sampled. In the case of imperfect recall, the simulation predicts that the 8 additional items, although imperfectly recalled, lead to wider generalisation as a result of increased diversity. (c) Simulated performance in the HELPFUL condition (12 exemplars), assuming the BASELINE performance as a prior and that the 8 additional exemplars are strongly sampled. Under strong sampling, generalisation tightens quickly around the sampled range with each extra exemplar, thus the predicted effect of forgetting is less in this scenario.

peared to have been strong sampling) was the likelihood that shaped generalisation – even though the cover story at retrieval should have contradicted it. While we cannot altogether rule out the influence of sampling assumptions at the point of retrieval, our experiment provides evidence in favour of an encoding account. Under this account, the evidence for different hypotheses is assessed according to the sampling assumption that prevailed at the time that the data were originally presented.

This finding has a variety of interesting implications. First, it suggests that there is no such thing as a "theoryless" learner: at no point do people simply encode the raw data in a veridical fashion. Rather, from the start they are actively engaged in making sense of it for future generalisation even though there is no current need to generalise. The question remains as to how automatic this is: would people be able to inhibit the likelihood calculation if requested to remember each specific data point as precisely as possible, or if they didn't think that a generalisation task would be forthcoming?

This has implications for effective pedagogy as well. It is known that learners benefit from assuming that their teacher is selecting the most informative examples possible given the learner's current beliefs. Such reciprocal assumptions can lead to a highly leveraged form of generalisation in which concepts can quickly be acquired from minimal input (Shafto et al., 2014). Under the idealised account of pedagogical learning, people's inferences should not depend on when the sampling process becomes apparent. However, our results suggest that it is important for the teacher to make the sampling process clear as early as possible.

In a similar way our finding has implications for how people process misinformation and corrections to misinformation. Ransom et al. (2017) found, for example, that people can use truthful but limited data in their efforts to mislead others by attempting to manipulate their counterpart's sampling assumption. Our work suggests that subsequently learning that an information source was biased may not be sufficient to correct the bias. It therefore offers another explanation for the well-established finding that retracting misinformation does not eliminate its influence (Johnson & Seifert, 1994; Ecker, Lewandowsky, Swire, & Chang, 2011). If people are encoding data in such a way that it cannot be disentangled from their theory at the time, interpreting that data under a new theory may be extremely difficult.

Another interesting aspect of this work regards the role of memory. By adopting the experimental procedure of Ransom et al. (2018) but requiring participants to view the simuli one-by-one, we were able to assess how memory decay would interact with sampling assumptions in shaping generalisation. We found that people tightened their generalisations less when they had to rely on their memory more. A simulation of the generalisation task used in our experiment verified our intuition that this should be the case (see Figure 6). Our finding is consistent with previous work using complex linguistic and non-linguistic data rather than a simple one-dimensional category (Perfors, Ransom, & Navarro, 2014), which suggests that the result is reasonably robust.

Our memory manipulation (albeit across two experiments) also provides some basis to distinguish between two possible encoding accounts. One possibility is that evidence is stored and retrieved with each exemplar individually and any failure to retrieve an exemplar would mean that computation occurs over a smaller dataset. A second possibility is that evidence is stored in aggregate (across all data points) and retrieved via the hypotheses. In this case, the contribution of each exemplar would be accounted for at the point of encoding, and so the computation should proceed as if the full dataset were retrieved. The two possibilities suggest contrasting predictions.

In the first case, we would expect generalisation in the present experiment to be wider than in the previous (Ransom et al., 2018, where perfect recall was supported). In the latter case, we should expect the results of the two experiments to be broadly in line with each other. As already noted, we found that manipulating how easy it was to remember exemplars did affect generalisation in a manner consistent with some degree of recall failure. We interpret this as weak evidence favouring the "exemplar encoding" account over the "hypothesis encoding" account: the data is stored in such a way that the strength of evidence is in some way integral to the encoding of the exemplar, at least to the extent that failure to later retrieve the exemplar equates to a failure to incorporate the associated evidence. Our evidence is only weak, however, because it is not entirely clear what "forgetting" in the context of the hypothesis encoding account would amount to. Fleshing out these distinctions more and testing them more systematically is a goal for future work.

While the present experiment should be taken in the spirit of a "proof of concept", our research nonetheless suggests that memory, sampling, and generalisation are intertwined in ways that are still not fully understood. By manipulating when different information is available as well as the cognitive load during learning, it is possible to further illuminate this complex relationship.

Acknowledgements

Thanks to Simon Dennis for helpful comments regarding the initial concept behind this study. Thanks also to the anonymous reviewers for their helpful comments. This work was supported by an Australian Government Research Training Program Scholarship (KR) and ARC Discovery Grant DP180103600

References

Ecker, U., Lewandowsky, S., Swire, B., & Chang, D. (2011). Correcting false information in memory: Manipulating the strength of misinformation encoding and its retraction. *Psychonomic Bulletin & Review*(18), 570–578.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *TiCS*, *20*(11), 818–829.

Hayes, B. K., Navarro, D. J., Stephens, R. G., Ransom, K. J., & Dilevski, N. (2019). The diversity effect in inductive reasoning depends on sampling assumptions. *Psychonomic Bulletin & Review*, 1–8.

Hendrickson, A. T., Perfors, A., Navarro, D. J., & Ransom, K. J. (2019). Sample size, number of categories and sampling assumptions: Exploring some differences between categorization and generalization. *Cognitive Psychology*, *111*, 80–102.

Johnson, H., & Seifert, C. (1994). Sources of the continued influence effect: When misinformation in memory affects later inferences. *Jn Exp Psych: LMC*(20), 1420–1436.

Landrum, A. R., Eaves, B. S., & Shafto, P. (2015). Learning to trust and trusting to learn: a theoretical framework. *Trends in Cognitive Sciences*, *19*(3), 109 - 111.

Perfors, A., Ransom, K. J., & Navarro, D. J. (2014). People ignore token frequency when deciding how widely to generalize. In *36th Annual CogSci Conference* (pp. 2759–2764).

Ransom, K. J., Hendrickson, A., Perfors, A., & Navarro, D. J. (2018). Representational and sampling assumptions drive individual differences in single category generalisation. In *40th Annual CogSci Conference* (pp. 930–935).

Ransom, K. J., Voorspoels, W., Perfors, A., & Navarro, D. J. (2017). A cognitive analysis of deception without lying. In *39th Annual CogSci Conference* (pp. 992–997).

Shafto, P., Eaves, B., Navarro, D. J., & Perfors, A. (2012). Epistemic trust: Modeling children's reasoning about others' knowledge and intent. *Developmental Science*, *15*, 436–447.

Shafto, P., Goodman, N. D., & Griffiths, T. L. (2014). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive Psychology*, *71*, 55–89.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Beh. & Brain Sci.*, *24*(4), 629–640.

Modeling Human Syllogistic Reasoning: The Role of "No Valid Conclusion"

Nicolas Riesterer* (riestern@cs.uni-freiburg.de)

Daniel Brand* (daniel.brand@cognition.uni-freiburg.de)

Hannah Dames (damesh@cs.uni-freiburg.de)

Marco Ragni (ragni@cs.uni-freiburg.de)
Cognitive Computation Lab, Georges-Köhler-Allee 79
79110 Freiburg, Germany

Abstract

"No Valid Conclusion" (NVC) is one of the most frequently selected responses in syllogistic reasoning experiments and corresponds to the logically correct conclusion for 58% of the syllogistic problem domain. Still, NVC is often neglected in computational models or just treated as a byproduct of the underlying inferential mechanisms such as a last resort when the search for alternatives is exhausted. We illustrate that NVC represents a major shortcoming of current models for human syllogistic reasoning. By introducing heuristic rules, we demonstrate that slight extensions of the existing models result in substantial improvements of their predictive performances. Our results illustrate the need for better NVC handling in cognitive modeling and provide directions for modelers on how to integrate it into their approaches.

Keywords: cognitive modeling; heuristics; syllogistic reasoning; no valid conclusion

Introduction

Syllogistic reasoning is one of the core domains in human reasoning research (for a review see Khemlani & Johnson-Laird, 2012). It is concerned with gaining insight into the cognitive processes driving the inference mechanisms for categorical assertions featuring quantifiers ("All", "Some", "Some ... not", and "No") and terms which are inter-related by two premises. The traditional experimental paradigm presents participants with problems of the form "All A are B; All B are C" (substituting A, B, and C with common groups such as gardeners, musicians, etc.) and usually asks "What follows?", i.e., which conclusion can be inferred logically from the premises (generation task; Morley, Evans, & Handley, 2004). Depending on the arrangement of terms, the syllogism is categorized into one of four figures, a property that was found to have a substantial influence on human inferences (Johnson-Laird & Bara, 1984):

	Figure 1	Figure 2	Figure 3	Figure 4
Premise 1	A-B	B-A	A-B	B-A
Premise 2	B-C	C-B	C-B	B-C

For reasons of clarity, syllogistic problems are usually referred to by abbreviating quantifiers with single uppercase letters and the figure number: "All" (A), "Some" (I), "No" (E), "Some ... not" (O). The syllogism "All informative things are useful; Some websites are not informative things" is therefore referred to as AO2. Possible conclusions for syllogistic problems combine the end terms A and C via one of the four quantifiers. Additionally, it is possible to respond with "No Valid Conclusion" (NVC) indicating that the premises have no valid conclusion in accordance to first-order logic. Out of the 64 distinct syllogistic problems, 37 are invalid (58%), i.e., only NVC can be derived.

Experimental investigations have shown that NVC represents one of the most frequently selected conclusions (Khemlani & Johnson-Laird, 2012). Because of this, the role of NVC in syllogistic reasoning is important. However, current models of syllogistic reasoning rarely make explicit statements about NVC. On the extreme, there are heuristic models which do not possess the capability of generating NVC at all. On the other hand, models that do integrate NVC as a conclusion candidate often treat it as a termination criterion when searches for alternatives fail. Currently, there are no strategies to directly infer NVC responses. Additionally, even when going beyond the level of predictions, models are unable to account for statistical phenomena related to NVC responses, such as variations in reaction times (Ragni, Dames, Brand, & Riesterer, 2019).

In this article, we tackle this problem by proposing a set of heuristic rules for generating NVC conclusions based on findings from the syllogistic literature. By attaching these rules to existing models, we show that inadequate NVC handling is indeed one of the core problems of the current state of the art. The following text is split into five sections. After introducing the syllogistic domain of reasoning as well as the current state of the art in modeling (Section 2), we will analyze contemporary models in terms of their capabilities in predicting a human NVC response (Section 3). Section 4 then takes up those results and presents alternative strategies for predicting NVC responses. In Section 5 we evaluate the syllogistic models augmented with the identified strategies for NVC and finally, in Section 6, discuss our results, illustrate the potential with respect to improving models, and give directions for future work in the field of cognitive modeling of human reasoning.

*Both authors contributed equally to this manuscript.

Table 1: Models and their NVC prediction proportions for valid and invalid syllogisms. For models with multiple prediction candidates for a single syllogism, the ratio of NVC is used. Model predictions are taken from Khemlani and Johnson-Laird (2012).

Model	Valid	Invalid
Conversion	44%	86%
Mental Models Theory (MMT)	14%	30%
PSYCOP	0%	100%
Verbal Models	32%	51%
Atmosphere	0%	0%
Matching	0%	0%
Probability Heuristics Model (PHM)	0%	0%

Related Work

Computational modeling is a central part of today's research of human syllogistic reasoning. As of today, at least twelve theories about syllogistic inferences exist. In a meta-analysis, Khemlani and Johnson-Laird (2012) found that the theories have distinct advantages and drawbacks when predicting experimental data obtained by aggregating individual participants' responses. The following paragraphs briefly introduce the different approaches for which the authors were able to provide predictions for the 64 syllogisms. They will be used throughout the following analyses.

The **Conversion Hypothesis** is an attempt at explaining erroneous conclusions resulting from human reasoning processes originally introduced by Chapman and Chapman (1959) and later formalized as a testable model by Revlis (1975). The hypothesis states that while encoding a syllogistic premise, a conversion operation is applied which swaps the direction of the categorical expression (e.g., "All A are B" is interpreted as "All B are A"). As a result, a new syllogism is produced with conclusions that might be inappropriate for the original problem (e.g., Revlin, Leirer, Yopp, & Yopp, 1980). NVC is predicted if the new problem is logically invalid.

The **Mental Model Theory** (MMT; Johnson-Laird, 1975) is a cognitive theory which has successfully been applied to various domains of reasoning (Johnson-Laird & Byrne, 2002; Khemlani & Johnson-Laird, 2012; Ragni & Knauff, 2013). It is based on the assumption that inferential mechanisms operate on mental representations constructed for the given premises. MMT's inference process is composed of a series of phases: model construction, conclusion generation, and the search for counterexamples. First, an initial mental model is constructed integrating the information of the premises, i.e., the relation between the terms of the premises. Second, a candidate conclusion is formulated in accordance to the initial model. Finally, alternative models consistent to the premises are constructed in search of a situation in which the conclusion is false (Ragni, Khemlani, & Johnson-Laird, 2014). If the initial model construction fails, or counterexamples can be found for all models, NVC is returned.

The **Psychology of Proof** model (PSYCOP; Rips, 1994) is a cognitive model of human syllogistic reasoning that claims deduction as a fundamentally human capability (Khemlani & Johnson-Laird, 2012). PSYCOP defines a set of psychologically plausible inference rules approximating the human inferential mechanisms. By applying rules in a deductive forward-inference fashion as well as an inductive backwards-inference fashion, a path between premise information and conclusion is constructed. PSYCOP does not have a guaranteed way to conclude NVC. While it supports exhaustive searches for conclusions and the generation of NVC as fallback option, this behavior is not enforced in its original formulation (Khemlani & Johnson-Laird, 2012).

Verbal Reasoner (Polk & Newell, 1995) is an approach to modeling syllogistic reasoning that assumes that human inferences are fundamentally verbal. It encodes the premise information into a mental model that differentiates between more accessible information (the subject of the premise) and less accessible information (the object of the premise). By defining procedures to extract different degrees of intermediate implicit knowledge about the reasoning problem, the model is able to generate conclusions following more or less complex inferences. The verbal model theory treats NVC as a last-resort option. If no conclusion can be derived from the mental model, the verbal reasoner enters a reencoding loop in search for a solution. NVC is produced when it gives up.

The **Atmosphere Hypothesis** (Woodworth & Sells, 1935) is able to account for a portion of errors in human syllogistic reasoning when compared with formal logics (Revlis, 1975). It is based on a feature extraction step that identifies whether the given premise information is positive/negative ("All", "Some" vs. "Some not", "No") and universal/particular ("All", "No" vs. "Some", "Some not"). By following a combination procedure, the quantifier of the conclusion is determined. Because it only extracts and combines features based on quantifiers, the atmosphere hypothesis is not able to provide information about the direction, i.e., the order of terms in the syllogistic conclusion, and is not able to generate NVC.

The **Matching Hypothesis** (Wetherick & Gilhooly, 1995) reflects a different approach for accounting for errors made in human syllogistic reasoning. It employs a matching strategy which states that the conclusion quantifier is equal to the most conservative quantifier in the premises. Conservativeness in this sense is defined as a preference order of $E > O = I \gg A$ following the estimated number of individuals a quantifier makes a statement about. Similar to Atmosphere, Matching is unable to predict NVC, because it always picks a quantifier from the given premises.

The **Probability Heuristics Model** (PHM; Chater & Oaksford, 1999) is an approach to modeling reasoning that is based on the fundamental idea that reasoning relies on heuristics. PHM defines the inferential process via two phases. First, a conclusion is generated by applying the *min-heuristic* selecting the least informative quantifier from the premises ($A > I > E \gg O$). Second, *probabilistic entailments* can be

applied generating alternative conclusions based on the min-heuristic's result that could probably be true. Next, a third heuristic, *attachment*, is applied to determine the order of terms in the conclusion. Finally, the *max-heuristic* is applied to assess the confidence of the conclusion based on the informativeness of the premises. If confidence is low, the probability of returning NVC instead of the solution candidate rises. Additionally, the *o-heuristic* is applied which states that O-responses should be avoided in favor of NVC.

In Khemlani and Johnson-Laird (2012)'s prediction table, which we use as the source for the models' predictions, PHM is reported without an inclusion of the max- and o-heuristic (Baratgin et al., 2015). While potentially distorting for model comparisons, this does not affect our evaluation of NVC. The max- and o-heuristics are attached to PHM's inference mechanisms (min-heuristic, attachment, and probabilistic entailment) in similar spirit to what we propose as general extensions of cognitive models further below.

The present article investigates the theories based on their NVC prediction capabilities. Table 1 summarizes the models' NVC response proportions in accordance to the prediction data reported by Khemlani and Johnson-Laird (2012) for valid and invalid syllogisms. The table highlights the difference between the cognitive models. While some models are unable to predict NVC at all, the other approaches have a stronger tendency toward responding with NVC for invalid syllogisms. This behavior is expected due to NVC being the logically valid response for invalid syllogisms. PSYCOP reflects formal first order logic in its NVC response behavior. Because all valid and no invalid syllogisms have categorical conclusions, it predicts 0% and 100% NVC, respectively. In the following analyses, we evaluate the models based on their ability to predict the most frequently selected responses.

Analysis State of the Art

Modeling Task

In this article, we aim at uncovering the latent potential of the current state of the art by investigating their prediction capabilities with a special focus on NVC. Hence, we adopt a predictive scenario as the core evaluation setting of the following analyses: Given a dataset of reasoning data, we first compute the most frequent answer (MFA) and assess each model's performance by comparing its predictions with the aggregated response given by the participants.

The dataset used for this article was recorded as an Amazon Mechanical Turk web experiment in 2016 and consists of $N = 139$ participants providing conclusions to all 64 syllogistic problems, each. Participants were asked to select one of the nine syllogistic response candidates following from the premises. After a training phase consisting of four easy syllogisms, the remaining task sequence and order of response options was fully randomized.

The predictions for the model candidates were taken from Khemlani and Johnson-Laird (2012). This prediction data does not feature single explicit conclusions for each model and task. Instead, only sets of possible conclusions can be provided for each model and syllogism. To account for this in our prediction setting, weighted scores were computed for the following analyses via $S(P,T) = |P \cap T|/|P|$, where P and T denote the sets for predicted and true responses, respectively (e.g., Copeland, 2006).

All materials used for the following analyses are openly available via Github[1].

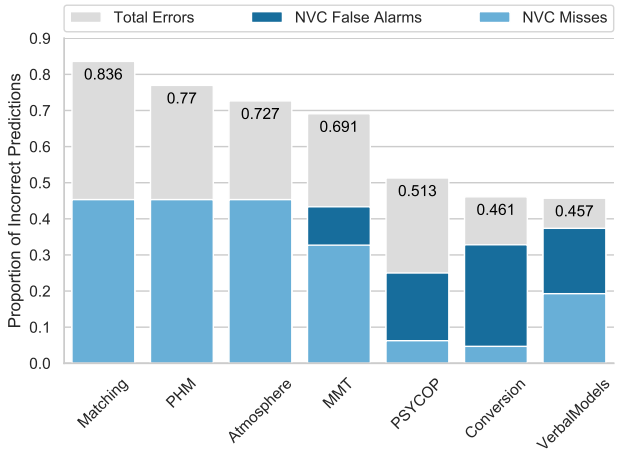


Figure 1: Proportion of model prediction errors (grey) for the 64 syllogisms. False alarms (dark blue), i.e., incorrect, and missed NVC predictions (light blue) are represented as proportions of the prediction error.

State of the Art

Figure 1 illustrates the predictive capabilities of the models in accordance to the prediction table of Khemlani and Johnson-Laird (2012). The grey bars reflect the proportion of incorrect predictions on the 64 syllogisms' MFA responses. Dark blue and light blue bars denote the parts of incorrect responses which can be attributed to unwanted and missed NVC responses, respectively. As an illustrating example, PSYCOP incorrectly predicts 51% of the syllogisms. About 6% of those errors can be attributed to missed NVC responses whereas 19% of the errors were due to false alarms.

The plot highlights the difference between the models in today's state of the art. As expected, the models which are unable to predict NVC responses (Matching, PHM, Atmosphere), perform worst. For the remaining models, the general performance is better. However, NVC-based errors still account for the large parts of the incorrect predictions. As a particularly striking example, more than half of Conversion's errors are due to incorrect NVC predictions.

The depicted results highlight the need for a better understanding of NVC. In the following, we propose strategies for predicting NVC based on results from the literature on human syllogistic reasoning. Since embedding these strategies

[1] https://github.com/nriesterer/syllogistic-nvc

Table 2: Change in predictive accuracy (black, first value), misses (lightblue, middle value), and false alarms (darkblue, right value) of the models' NVC predictions.

Models	PartNeg			EmptyStart			FiguralRule			NegativityRule			ParticularityRule		
Atmosphere	**37.5**%	-28	8	7.8%	-6	2	14.8%	-17	15	25.0%	-16	0	15.6%	-12	4
Conversion	**1.6**%	-2	2	0.8%	-1	1	-3.1%	-2	8	0.0%	0	0	0.0%	0	0
MMT	**28.3**%	-20.1	6	6.2%	-4.4	1.6	11.3%	-12.3	12.2	17.5%	-11.2	0	11.3%	-8.4	2.8
Matching	**42.2**%	-28	8	7.8%	-6	2	19.5%	-17	15	25.0%	-16	0	17.2%	-12	4
PHM	**39.8**%	-28	8	7.8%	-6	2	17.6%	-17	15	25.0%	-16	0	16.4%	-12	4
PSYCOP	**4.2**%	-4	4	3.1%	-2	2	-3.0%	-2	13	0.0%	0	0	0.0%	0	0
VerbalModels	**11.6**%	-11.7	5.2	4.2%	-3.2	1	3.3%	-6.7	8.4	9.1%	-5.8	0	4.9%	-4.7	2

into the assumptions stemming from the high-level theoretical ideas of the models exceeds the scope of this article, we focus on formulating the NVC strategies as rules which can be attached to arbitrary models. If a rule does not predict NVC, the underlying model is queried. This allows us to examine the benefits and assess potential shortcomings of an improved NVC handling in modeling human syllogistic reasoning. Because our rules are purely additive, we expect models with high numbers of NVC misses to benefit most from the proposed strategies. The challenge lies in minimizing the inevitable increase in false alarms.

Towards a Model of NVC

To tackle the problem of missed NVC responses, we introduce a set of heuristic rules detecting NVC which are based on different observations.

The first heuristic, the *Figural Rule* is based on the figural effect, a core result of syllogistic reasoning research. Early studies found that the figure of premises induces a reliable bias on participants' responses: Figure 1 encourages A-C responses while Figure 2 leads to higher proportions of C-A responses (Johnson-Laird, 1975). In a later study it was found that the syllogistic figure also has an effect on the proportion of NVC responses (Johnson-Laird & Bara, 1984): NVC is preferred for syllogisms of Figure 3 and 4. This finding is transformed into a rule generating the NVC response whenever a syllogism of Figure 3 and 4 is encountered. For the remaining figures, the attached model is queried.

The next set of rules draws from the notion of informativeness of quantifiers as a criterion for determining NVC. Informativeness is a driving factor for two models in the current state of the art of syllogistic reasoning. The probability heuristics model (Chater & Oaksford, 1999) assumes an informativeness ordering of $A > I > E \gg O$ based on how unexpected truth about a statement is conceived by humans. Matching, on the other hand, introduces the notion of conservativeness based on the number of individuals a premise makes an assertion about: $E > O = I \gg A$ (Wetherick & Gilhooly, 1995). Both orders assign the least amount of information to the negative quantifiers "Some ... not" (O), and "No" (E). The *negativity rule* integrates both orders by being defined on the assumption that the amount of information encoded by two negative premises does not suffice to license a valid conclusion. This rule relates to PHM's max-heuristic in the sense that it assumes a threshold for insecurity with a generated conclusion candidate that is exceeded for E and O quantifiers. In doing so, it also subsumes PHM's o-heuristic. In analogy to negativity, the *particularity rule* is defined based on the limited information encoded in the particular quantifiers "Some" (I) and "Some ... not" (O). They make assumptions about limited and unspecified sets which might cause the reasoning process to fail. Finally, we define a third rule, *PartNeg* by combining both particularity and negativity: If the syllogism only consists of quantifiers with limited information, i.e., does not contain "All", NVC is predicted.

The last rule, *EmptyStart*, focuses on the syllogisms where information can be propagated transitively through the premises. This is possible for figure 1, i.e., "A-B, B-C", or figure 2, i.e., "B-A, C-B", which can be converted into figure 1 by swapping the premises and substituting C with A and A with C. The heuristic assumes that an information propagation is constructed (A-B-C for figure 1, C-B-A for figure 2). Inferences can only be drawn if the quantifier relating the two terms in the beginning of the chain makes an assertion about a non-empty set of individuals. If this premise features "No", i.e., the most conservative premise (Wetherick & Gilhooly, 1995), no information can be propagated through the chain and NVC is inferred. If we consider syllogism IE1, the chain A-B-C can be extracted starting with quantifier "Some". The reasoner is able to identify a selection of elements from A which can be annotated as B. The information from the second premise can now be integrated easily into the elements from A. If we consider EI1 on the other hand, the reasoner is unable to identify an initial set of elements from A. Therefore, premise 2 cannot be related to elements from A. As a result, there is a higher chance to respond with NVC.

Analysis

Figure 2 depicts the syllogisms for which the introduced heuristics predict NVC along with the syllogisms for which NVC is the most frequent answer (MFA). Comparing the strategies our results show that different parts of the space of syllogisms are covered by different rules. For instance,

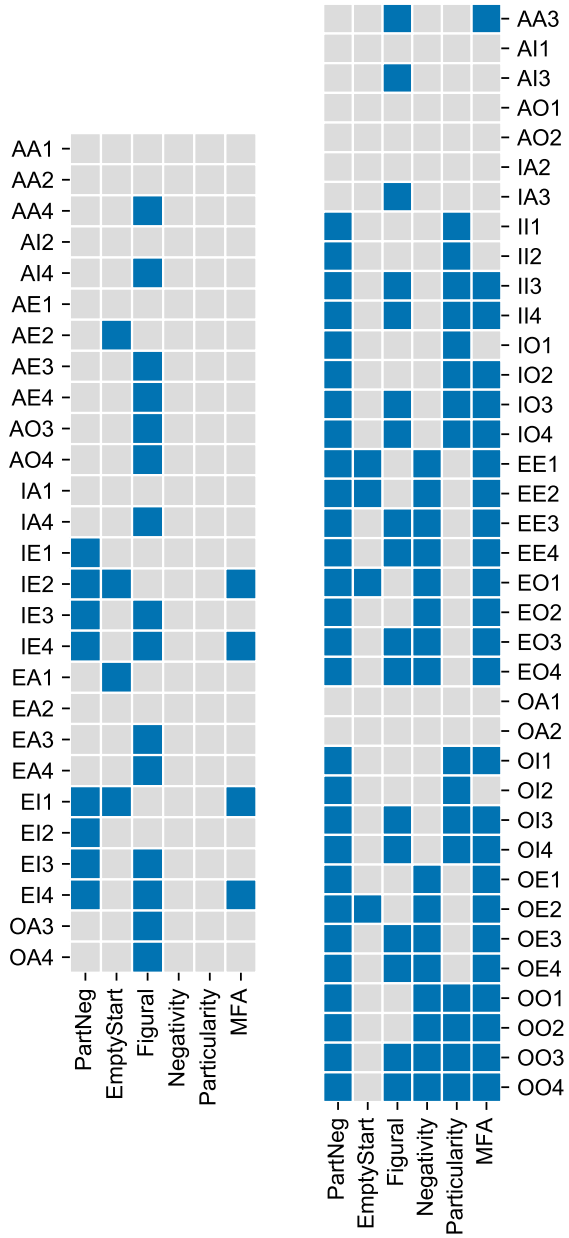


Figure 2: NVC Predictions of the individual rules on valid (left) and invalid syllogisms (right). Syllogisms are abbreviated with the encoded quantifiers of both premises "All" (A), "Some" (I), "No" (E), and "Some ... not" (O), and the figure.

negativity and particularity do not predict NVC for valid syllogisms, because there only exist invalid syllogisms characterized by being fully negative or particular. Figural on the other hand generates NVC for large parts of the syllogistic domain regardless of the validity of the underlying problem. When compared with MFA, the rules vary in predictive performance. PartNeg is capable of covering large parts of the invalid syllogisms correctly and only makes few errors for the valid cases. In contrast, figural's predictions show a more substantial difference in performance between valid and invalid syllogisms.

More generally, the plot also illustrates that most responses were not given by following standard logics. This is especially apparent in the case of the 37 invalid syllogisms where only 25 (68%) of the MFA responses correspond to NVC.

Integrating NVC into Models

To determine the effectiveness of our NVC rules, we attach them to the original state-of-the-art models and evaluate their change in performance. This is depicted in Table 2. It presents the raw improvement of the syllogistic models achieved by attaching the respective NVC rule. Additionally, the decrease in misses (light blue) and increase in false alarms (dark blue) are illustrated. In general, larger improvements (percentages), fewer misses, and fewer false alarms indicate better performance.

Table 2 draws a convincing picture about the qualities of the NVC rules. With the exception of the figural rule, all strategies result in substantial improvements over the standard models. PartNeg achieves the overall peak performance improving up to 42.2% when compared to the base model. EmptyStart has the overall lowest changes in performance but introduces only few additional errors. As expected, models which do not generate NVC at all benefit most from the capability of responding with NVC achieving an improvement of 21.3%, 20.3%, and 20.1% on average across all NVC rules, respectively. PSYCOP (0.9% on average) and Conversion (-1.4% on average) do not benefit from the additional NVC rules with Conversion's performance even decreasing slightly. Surprisingly though, MMT is improved substantially by the additional NVC rules (14.9% on average) even though it already has the capability of generating NVC.

To gain additional insight into the performance of the models, Figure 3 replicates the introductory plot from Figure 1. It depicts the errors in the predictions of the models extended with PartNeg, the overall best NVC rule. Again, the plot depicts the proportion of incorrect predictions (grey) as well as the fractions corresponding to false alarms (dark blue) and misses (light blue).

The figure illustrates that the attached rule, PartNeg, manages to effectively remove NVC misses from the models' predictions. Simultaneously, it achieves this without introducing substantial amounts of false alarms. Consequently, in combination with PartNeg, a heuristic rule was found that is able to nuancedly relate human reasoner's tendencies towards concluding NVC to the syllogistic quantifiers. The fact that the improvement in handling NVC caused a substantial increase in performance for most of the models further strengthens the claim that NVC is one of the core weaknesses of the current state of the art in modeling human syllogistic reasoning.

Figure 4 illuminates the qualities of NVC rules on an individual level. For each model, the values refer to the number of participants for which a certain rule achieves highest

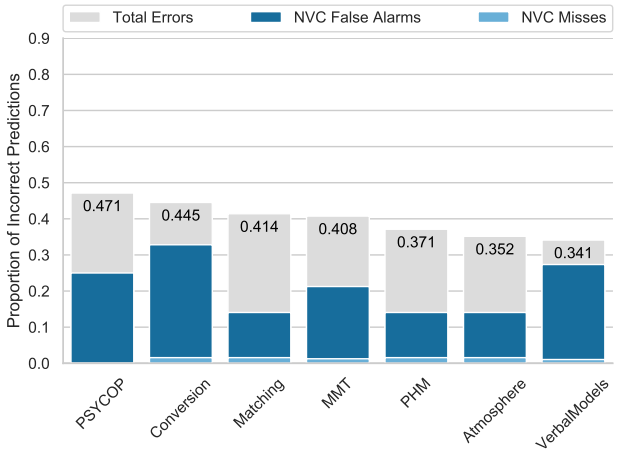


Figure 3: Proportion of the prediction errors (grey) achieved by the models extended with PartNeg, the overall best NVC heuristic, on the 64 syllogisms. False alarms (dark blue), i.e., incorrect, and missed NVC predictions (light blue) are represented as proportions of the prediction error.

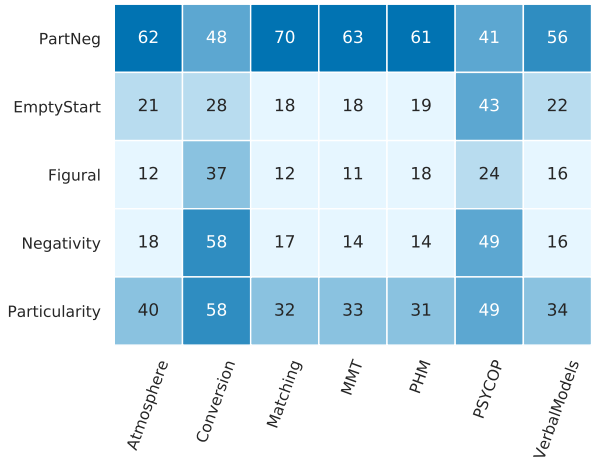


Figure 4: For each model, the values denote the number of participants for which the corresponding NVC rules performs best. In case of ties, the subject is counted for both rules.

performance. The figure illustrates that while PartNeg is the overall best rule, there is quite a substantial number of participants which can be accounted for better by other rules. This suggests that NVC response behavior is dependent on inter-individual differences of reasoning processes.

General Discussion

As the correct response for 58% of syllogisms as well as one of the most frequently given responses by human reasoners (Khemlani & Johnson-Laird, 2012), "No Valid Conclusion" (NVC) is an important response for computational models to capture.

Our results demonstrate that the current state of the art in modeling human syllogistic reasoning is lacking the capabilities for handling NVC correctly. While some other approaches do not feature the ability of producing NVC at all, even the more complex approaches yield false alarm rates of up to 25% (Conversion) and misses of up to 30%. The high miss rates highlight a lack of precision in identifying the problems where NVC responses are adequate.

We combat these shortcomings by introducing five heuristic rules for predicting NVC based on prominent phenomena and properties of syllogistic reasoning (e.g., figural effect; Johnson-Laird, 1975; Johnson-Laird & Bara, 1984, or informativeness of premises; Chater & Oaksford, 1999). By attaching these rules to the cognitive models taken from Khemlani and Johnson-Laird (2012), a substantial improvement can be observed for the majority of models. Models without the capability of predicting NVC could achieve an increase in performance of up to 20% on average across all rules. Combined with PartNeg, the overall best NVC rule, we were able to demonstrate a substantial decrease of misses across the board. Even though these rules introduce low numbers of additional false alarms, this effect is negligible when compared to the substantial reduction of misses.

In conclusion, our work contributes to research in the domain of syllogistic reasoning both on a theoretical and practical level. We isolate NVC as one of the core flaws of the current state of the art in modeling syllogistic reasoning. By demonstrating substantial improvement when attaching NVC predictors, we highlight the remaining potential for modelers to tap into. The next step for cognitive modelers is to integrate these findings into future iterations of their models and derive additional rules from cognitive theories. With PartNeg, we provide a first rule which represents a valuable heuristic candidate for explaining NVC response behavior.

Furthermore, our results show the potential that lies in isolating and improving parts of the problem domain. By highlighting their shortcomings, modelers are given the chance to iteratively improve on their computational models and underlying theories. Apart from NVC, another candidate for improvement is the conclusion direction. Currently, there exist models which completely ignore direction as a predictive factor (e.g., Atmosphere) and others which actively integrate it into their underlying formalisms (e.g., Conversion).

Still, even though PartNeg captures the majority of MFA responses, it is not the optimal choice for each individual. There still is potential left for making better predictions if the relation between individual reasoners' characteristics and their response behavior can be understood. Our results suggest that there is no single rule capable of accounting for all individuals. Therefore, one goal of future models is to determine and use discriminative features enabling the detection of the reasoning strategy most fitting to a specific reasoner.

Acknowledgements

This paper was supported by DFG grants RA 1934/3-1, RA 1934/2-1 and RA 1934/4-1 to MR.

References

Baratgin, J., Douven, I., Evans, J. S. T., Oaksford, M., Over, D., & Politzer, G. (2015). The new paradigm and mental models. *Trends in Cognitive Sciences*, *19*(10), 547–548.

Chapman, L. J., & Chapman, J. P. (1959). Atmosphere effect re-examined. *Journal of Experimental Psychology*, *58*(3), 220–226.

Chater, N., & Oaksford, M. (1999). The probability heuristics model of syllogistic reasoning. *Cognitive Psychology*, *38*(2), 191–258.

Copeland, D. E. (2006). Theories of categorical reasoning and extended syllogisms. *Thinking & Reasoning*, *12*(4), 379–412.

Johnson-Laird, P. N. (1975). Models of deduction. *Reasoning: Representation and process in children and adults*, 7–54.

Johnson-Laird, P. N., & Bara, B. G. (1984). Syllogistic inference. *Cognition*, *16*(1), 1–61.

Johnson-Laird, P. N., & Byrne, R. M. J. (2002). Conditionals: A theory of meaning, pragmatics, and inference. *Psychological Review*, *109*(4), 646–678.

Khemlani, S., & Johnson-Laird, P. N. (2012). Theories of the syllogism: A meta-analysis. *Psychological Bulletin*, *138*(3), 427–457.

Morley, N. J., Evans, J. S. B. T., & Handley, S. J. (2004). Belief bias and figural bias in syllogistic reasoning. *The Quarterly Journal of Experimental Psychology Section A*, *57*(4), 666–692.

Polk, T. A., & Newell, A. (1995). Deduction as verbal reasoning. *Psychological Review*, *102*(3), 533–566.

Ragni, M., Dames, H., Brand, D., & Riesterer, N. (2019). When does a reasoner respond: Nothing follows? In A. Goel, C. Seifert, & C. Freska (Eds.), *Proceedings of the 41st Annual Conference of the Cognitive Science Society.* Montreal: Cognitive Science Society.

Ragni, M., Khemlani, S., & Johnson-Laird, P. (2014). The evaluation of the consistency of quantified assertions. *Memory & Cognition*, *42*(1), 1–14.

Ragni, M., & Knauff, M. (2013). A theory and a computational model of spatial reasoning with preferred mental models. *Psychological Review*, *120*(3), 561–588.

Revlin, R., Leirer, V., Yopp, H., & Yopp, R. (1980). The belief-bias effect in formal reasoning: The influence of knowledge on logic. *Memory & Cognition*, *8*(6), 584–592.

Revlis, R. (1975). Two models of syllogistic reasoning: Feature selection and conversion. *Journal of Verbal Learning and Verbal Behavior*, *14*(2), 180–195.

Rips, L. J. (1994). *The psychology of proof: Deductive reasoning in human thinking*. Mit Press.

Wetherick, N. E., & Gilhooly, K. J. (1995). 'Atmosphere', matching, and logic in syllogistic reasoning. *Current Psychology*, *14*(3), 169–178.

Woodworth, R. S., & Sells, S. B. (1935). An atmosphere effect in formal syllogistic reasoning. *Journal of Experimental Psychology*, *18*(4), 451–460.

Event Participants and Verbal Semantics: Non-Discrete Structure in English, Spanish and Mandarin

Lilia Rissman (l.rissman@let.ru.nl)
Center for Language Studies, Erasmusplein 1
Nijmegen, the Netherlands 6525 HT

Kyle Rawlins (kgr@jhu.edu), Barbara Landau (landau@jhu.edu)
Department of Cognitive Science, Krieger Hall 237, 3400 N. Charles St.
Baltimore, MD 21218 USA

Abstract

Verbs are widely analyzed as functions taking a discrete number of arguments (e.g., *drink* has two arguments but *give* has three). Recent studies, however, suggest that English verbs encode Instruments as more or less salient (e.g., the Instrument is more salient for *slice*, less salient for *eat*). We conducted a judgment task with adult speakers of Spanish and Mandarin and found that verbs in these languages also encode Instruments as having a relative degree of salience, inconsistent with the discrete model of participant encoding.

Keywords: verbal semantics; argument structure; experimental semantics; thematic roles; event representation

Introduction

A fundamental debate in cognitive science concerns whether mental representations have discrete vs. non-discrete structure (Aarts, 2007; Bod, Hay & Jannedy, 2003; Rosch, 1975; Smolensky & Legendre, 2006). This debate arises for theories of verbal semantics. Verbs convey relationships between event participants: *eat*, for example, involves someone who eats and a substance that is eaten. Such relationships have commonly been modeled in logical terms: that a verb is a function taking a discrete number of arguments: *die* has one, *eat* has two, and *lend* has three (Dummett, 1981; Jackendoff, 1972). Theorists have long noted, however, the limits of this logical analogy (Carlson & Tanenhaus, 1989; Parsons, 1990; Williams, 2015). *Eating,* for example, seems to require that the eater have a mouth – is the mouth then one of the arguments of the function *eat*? Although there is broad consensus that verbs encode relations between participants, how precisely these relations are represented is unresolved.

A second unresolved question is whether participant relations are the same for semantically similar verbs across languages. As described by Bowerman and Brown (2008: 10), there is a widespread assumption that "languages will agree on the number of semantic participants there are in events of various types (e.g., one for 'laughing', two for 'pushing', three for 'giving')". There is reason to question this assumption: Wilkins (2008) argues that whereas the English verb *see* has two arguments, in the aboriginal language Arrernte, the translationally equivalent verb *are-* has three: the person who sees, the thing that is seen, and the place where the thing that is seen is located. While differences in argument realization are well-documented across languages (Levin & Rappaport-Hovav, 2005), variability such as described by Wilkins (2008) has received little attention. If variability in how verbs encode participants is widespread, then the mapping from conceptual to linguistic structure is less constrained than previously thought, posing an additional learning challenge to children.

In this study, we address whether discrete argument structures are good models for how verbs encode event participants, as well as whether verbal participant relations are variable across languages. We report the results of a judgment experiment with speakers of Spanish and Mandarin and compare these results with English data previously reported by Rissman, Rawlins and Landau (2015).

Previous Evidence for Semantic Gradience

One of the benefits of the discrete model of verbal participant encoding is that it fits well with syntactic theories of how event participants are overtly expressed: isomorphic mappings can be drawn between a verb's arguments and the surface constituents in a clause. For example, in *Jodi lent a book to her sister*, the arguments <Source, Theme, Recipient> map to the phrases <DP, DP, PP>. The distinction between a verb's arguments and its non-arguments (or "modifiers") is not dichotomous, however (Croft, 2001; Dowty, 2003; Vater, 1978), one reason being that verbal semantics and syntax are sometimes not isomorphic (Haspelmath, 2014; Koenig, Mauner & Bienvenue, 2003). Consider, for example, instrumental participants, as in *Jodi sliced the broccoli with a knife*. Verbs such as *slice* and *chop* activate an Instrument concept during sentence comprehension (Andreu, Sanz-Torrent & Rodríguez-Ferreiro, 2016; Koenig et al., 2003). Nonetheless, instrumental *with*-phrases pattern like modifiers (i.e., not like arguments) given syntactic argument diagnostics (Rissman et al., 2015; Schutze, 1995). For example, *what Jodi did with the knife was slice the broccoli* is acceptable but not **what Jodi did to her sister was lend a book.*

As a result of this mismatch between semantic and syntactic argument diagnostics, researchers cannot rely on syntactic diagnostics to understand how verbs semantically encode event participants. Alternate methods for probing verbal semantics include studies of sentence processing, sentence completion and semantic judgments (Barbu &

Toivonen, 2016; Boland, 2005; Koenig et al., 2003; Rissman et al., 2015; Wittenberg & Snedeker, 2014). In the judgment task in Rissman et al. (2015), English speakers read a paragraph stating that verbs have "arguments," defined as something "essential to the meaning of a verb but not part of the verb itself." This category was elaborated through positive examples, e.g. that *want* has two "arguments" because wanting involves someone who wants and something that is wanted. We distinguish the experimental category "argument" from the theoretical notion of argument.

Following this instruction, subjects judged which of the words in a sentence constituted the "argument" of the verb, for untrained verbs and participant types. Subjects read sentences such as in (1) and had to choose whether either the first or second bracketed phrase was an "argument" of the verb, or whether neither phrase was an "argument":

(1) a. [Last Tuesday] Martha SLICED something [with a steak knife].
b. Tania TAUGHT something [to the students] [in the classroom].

Rissman et al. (2015) hypothesized that if verbs like *slice* and *chop* discretely encode three arguments, an Agent, Patient and Instrument, then they are in an equivalence class with dative verbs such as *teach* and *lend,* which encode a Source, a Theme and a Recipient (Larson, 1988). By prediction, subjects would therefore be equally as likely to choose "with a steak knife" in (1a) as to choose "to the students" in (1b).

Instead, subjects selected Instruments less often than Recipients. In addition, there were differences across the instrumental verbs: an Instrument was selected more often for *slice* and *chop* than for *eat* and *break,* for example. Thus *slice* patterned like neither a 2-argument verb nor a 3-argument verb. Rather, Instruments appeared to have a moderate degree of salience: more salient than a time or a location, but less salient than a Recipient, inconsistent with the discrete model of participant encoding.

A variety of evidence indicates that this judgment task reflects abstract knowledge of verbal meaning. First, on control trials with prototypical arguments and modifiers, subjects almost always chose the Theme in sentences such as "John CARRIED [the books] [in a tote bag]" and almost never chose one of the modifiers in sentences such as "Martha CHOPPED something [on Monday] [in the forest]. Subsequent experiments showed: 1) that the difference between the Recipient and Instrument judgments was likely not driven by the difference in animacy (Recipients were animate whereas Instruments were inanimate), 2) that the Instrument judgments were not correlated with estimates of how often people use tools for these events, and 3) Instrument and Recipient judgments for each verb did correlate with how often people produce Instruments and Recipients in a corpus. Finally, Rissman (2018) found strong positive correlations between Instrument and Recipient judgments for each verb and rates of producing Instrument/Recipient completions for sentence fragments such as *Martha sliced the bread* _____ and *Tania taught the material*____.

Current study

We ask whether Spanish and Mandarin speakers also judge Instruments as having a moderate degree of salience. Such a finding would provide additional evidence against the discrete model of participant encoding. Investigating verbal semantics across multiple languages helps ensure that theoretical developments are not based on English alone.

We also ask whether non-discrete encoding of participants is itself cross-linguistically variable. Although *slice* patterns neither as a 2-argument nor a 3-argument verb, this does not preclude semantically similar verbs in other languages (e.g., Spanish *cortar*) from discretely encoding the Instrument. Languages differ widely as to which semantic role properties are relevant to syntactic argument realization (Bornkessel, Schlesewsky, Comrie & Friederici, 2006; Croft, 2001; Levin & Rappaport-Hovav, 2005). Verbal semantics is also highly variable across languages, with verbs in the same semantic space bundling semantic features in different ways (see Majid, Boster & Bowerman, 2008 for cutting and breaking events and Talmy, 1985 for motion events). In Mandarin, for example, the verb *jie4* encompasses both English *borrow* and *lend*. In the current study, we ask whether instrumental verbs in Spanish and Mandarin encode the Instrument in a discrete way, unlike in English.

For each of the verbs studied by Rissman et al. (2015), we selected semantically similar verbs in Spanish and Mandarin. For these similar verbs, we asked three questions:

1) Do judgments of Instrument salience parallel judgments of Recipient salience, unlike in English?
2) Do some verbs highlight an Instrument more strongly than other verbs, as is true for English?
3) Do verbs with similar meanings across languages give rise to similar judgments of Instrument salience?

In choosing Spanish and Mandarin, we compared one language that is genetically related to English (Spanish) and one language that is genetically distant (Mandarin). These languages both differ from English with respect to argument production: Spanish is a pro-drop language, allowing subject omission, while Mandarin allows both subject and object omission. We can thus test whether in languages that allow pervasive argument omission, subjects are less likely overall to judge that a particular phrase is an "argument."

Experiment 1

Participants

35 native Spanish-speaking adults (F = 22) and 32 native Mandarin-speaking adults (F = 23) participated. Spanish speakers were tested in Chicago and in Baltimore; all Mandarin speakers were tested in Baltimore. All participants reported having some knowledge of English. The Spanish speakers originated from throughout the Spanish-speaking world; Mandarin speakers originated from throughout China and Taiwan. All participants had attended or were currently attending college. Participants received $12 or course credit.

Design and Materials

Native speaker consultants translated the Rissman et al. (2015) materials into Spanish and Mandarin. The prior study tested two types of verbs: 1) verbs compatible with an Instrument ("Instrument verbs"), ranging from strongly to weakly instrumental (e.g., *slice, chop* vs. *eat, drink*), and 2) verbs compatible with a Recipient ("Recipient verbs"), ranging from strongly to weakly Recipient-encoding (e.g., *lend, teach* vs. *bounce, kick*). We selected Spanish and Mandarin verbs by describing to the consultants a set of events that exemplified core uses of each English verb (e.g., *chop* ~ chopping an onion, chopping wood). The consultants then provided the dominant verb in Spanish and Mandarin that would be used to describe these events. If no verb could be found that closely matched the meaning of the English verb and was compatible with the syntactic frames in (2-5), then no verb was tested. Tables 1-2 show the Spanish and Mandarin verbs that were tested, including omissions ("---").[1]

Each sentence in the experiment featured a single verb and two bracketed phrases: participants' task was to choose one of the bracketed phrases as an "argument" of the verb, or to choose that neither phrase was an "argument." Example Instrument and Recipient sentences are shown in (2-3) and (4-5) with English glosses and translations.

(2) Rachel REBANÓ algo [con una hoja de afeitar] [en el puerto].
Rachel slice-3PST something with a razor blade in the port
"Rachel sliced something with a razor blade in the port."

(3) 【在去年復活節那天】小琴 用【一把短柄小斧】⁞砍了⁞一些東西。
in last Easter Sunday Xiaoqin use one hatchet chop-PFV something
"Last Easter Sunday, Xiaoqin used a hatchet to chop something."

(4) [A las 6 am] Ruby le PRESTÓ algo [al nadador].
At 6 AM Ruby 3SG lend-3PST something to the swimmer.
"At 6 AM, Ruby lent something to the swimmer."

(5) 克洛伊【在街上】⁞賣了⁞一樣東西【給演員】。
Chloe in street send-PFV something to actors
"In the street, Chloe sent something to the actors."

The two bracketed phrases constituted several contrasts between two possible participant types. In the main trials of interest, Instruments and Recipients were pitted against prototypical modifiers (location, time and manner phrases). If Instruments and Recipients are arguments, these should be chosen significantly more often than modifiers.

There were two types of control trials. In the first, Themes were pitted against various phrase types including participant locations (e.g. *Layla LLEVÓ [los comestibles] [en una cesta];* "Layla CARRIED [the groceries] [in a basket]") and beneficiaries (e.g. *Jen LEYÓ [el mensaje] [para el detective]*; "Jen READ [the message] [for the detective]"). We predicted that subjects would choose the Theme as an "argument." In the second type of control trial, prototypical modifiers were pitted against each other, as in *Rachel REBANÓ algo [tristemente] [en el puerto]* ("Rachel SLICED something [sadly] [in the port]"). We predicted that in modifier vs. modifier trials, participants would judge that neither phrase was an "argument" of the verb. These control trials assess

Table 1: Instrument verbs

Eng	Span	Mand	Eng	Span	Mand
beat	golpear	qiao1da3	eat	comer	chi1
hit	pegar	da3	drink	beber	he1
touch	tocar	peng4	break	quebrar	da3po4
poke	---	chuo1	open	abrir	da3kai1
stab	apuñalar	ci4	kill	matar	sha1
cut	cortar	qie1	attack	atacar	gong1ji2
chop	picar	kan3	paint	pintar	---
slice	rebanar	---	grow	---	zhong4
write	escribir	xie3	move	mover	yi2dong4
draw	dibujar	hua4	lift	levantar	ju2qi3
dig	---	wa1	clean	limpiar	qing1li3
stir	revolver	jiao3	wash	lavar	xi3

Table 2: Recipient verbs

Eng	Span	Mand	Eng	Span	Mand
serve	servir	duan1	kick	patear	ti1
teach	enseñar	jiao1	throw	tirar	ren1
send	enviar	ji4	toss	---	tou2
tell	decir	---	roll	---	gun3
sell	vender	mai4	push	empujar	tui1
lend	prestar	chu1zu1	slide	---	---
pay	pagar	fu4	take	llevar	na2
offer	ofrecer	ti2gong4	bounce	---	---

[1] In Mandarin, serial verb constructions are common and productive (Li 1990). In Table 1, the verbs *da3po4* ('break'), *da3kai1* ('open'), *yi2dong4* ('move') and *ju2qi3* ('lift') are compound constructions rather than non-compound multi-character verbs. These verbs were included to maintain a close equivalence between the numbers of verbs and the semantic space of the verbs tested in English and Mandarin.

whether subjects distinguish prototypical arguments from prototypical modifiers.

The order of the bracketed phrases was counterbalanced such that some participants saw a trial such as in (2), whereas others saw a structure with a sentence-initial modifier such as *[En el puerto] Rachel REBANÓ algo [con una hoja de afeitar]* ("[In the port] Rachel SLICED something [with a razor blade]"). Each Instrument and Recipient verb appeared six times. In addition, each verb was paired with six unique Instrument/Recipient tokens (e.g., *con una hoja de afeitar* ("with a razor blade"), *con tijeras* ("with scissors")). There were both typical and atypical tokens for each verb. Summing across the experiment, Spanish/Mandarin participants saw a total of 312/318 trials.

Procedure

Participants received a Spanish/Mandarin version of the "argument" instructions from Rissman et al. (2015); the category labels *argumento* and *lun4yuan2* were used in Spanish and Mandarin, respectively. The instruction consisted of two phases: in the first, participants read a prose description about "arguments." Participants were told, for example, that "arguments" are essential to the meaning of a verb but are not part of the verb. Participants were given primarily positive examples, e.g. that *querer/yao4* ('want') has two "arguments," someone who wants and something that is wanted. Participants were also told that "arguments" are not necessarily syntactically required in a sentence. Participants read two negative examples, e.g. in *John ran until he was sick*, the phrase *until he was sick* is not an "argument". In the second phase of the instruction, participants completed practice trials where they read a verb and were asked to indicate the "arguments" of the verb. For example, Spanish participants read the sentence *Jim estaba cocinando* ("Jim was cooking"), were told that *cocinar* has two "arguments," and had to indicate which "argument" of *cocinar* was present and which was absent in the sentence. In another type of practice trial, Mandarin participants were asked to list the "arguments" of *"看"* ('look'); where the correct answer is two "arguments," someone who looks and something that is looked at. Feedback was given on all practice trials in the second phase of training. Across the entire instruction, explicit information was not given about the verbs or participant types that participants would be tested on. The instructions/practice trials were administered by native or near-native speakers of Spanish/Mandarin.

Results and Discussion

Spanish and Mandarin speakers performed as expected on the two types of control trials. For Theme trials (e.g., English ~ *John CARRIED [the books] [in a canvas bag]*), Spanish speakers chose the Theme as an "argument" on 93% of trials (CI_{95} = 1%), and Mandarin speakers chose the Theme on 95% of trials (CI_{95} = 1%). For modifier vs. modifier trials (e.g., English ~ *John CUT something [carefully] [last night]*), Spanish speakers chose the "neither" option on 89% of trials (CI_{95} = 1%) and Mandarin speakers chose "neither" on 96% of trials (CI_{95} = 1%). Thus speakers of Spanish and Mandarin, like the English speakers tested in Rissman et al. (2015), sharply distinguish prototypical arguments from prototypical modifiers in their judgments.

Figure 1 shows the main results, how often Spanish and Mandarin speakers judged Recipients and Instruments to be "arguments" for each verb.[2] The English-*with* data were previously reported in Rissman et al. (2015). These data suggest that Recipients are better examples of "arguments" than Instruments. To test whether Spanish and Mandarin speakers judged Instruments as having the same level of salience as Recipients, and whether these judgments varied across English, Spanish and Mandarin, we modeled the probability of choosing the Instrument or the Recipient (i.e., the *Target*) as an "argument" using mixed-effects logistic regression. Participants almost never selected one of the modifiers as an "argument;" we therefore collapsed the modifier and "neither" responses and modeled these data as a binary choice: whether or not participants chose the Target as an "argument." We fit regression models in R using the *glmer* function in the lme4 package (Bates & Maechler, 2009); models were evaluated through nested model comparison. Possible fixed effects in the model were Language (English vs. Spanish vs. Mandarin), Target type (Instrument vs. Recipient) and Competitor Type (location vs. time vs. manner); Subject was a possible random effect.

The best-fitting model of the data in Figure 1 contained the Subject random effect and the Target fixed effect: participants selected Recipients more often than Instruments (β = 2.53, SE = .05, $p < .001$). None of the following contributed significantly to the model fit: Language, Competitor Type, interaction between Language and Target Type and interaction between Target Type and Competitor Type (p-values for χ^2 tests all > .1). This analysis shows that in both Spanish and Mandarin, Recipients are more prominent for Recipient verbs than Instruments are for Instrument verbs, as in English.

We also observed variation across the individual Instrument verbs, in both Spanish and Mandarin. In Spanish, the rates of selecting the Instrument ranged from 11% (*comer*, 'eat'; CI_{95} = 7%) to 38% (*picar* 'chop'; CI_{95} = 9%). The 95% confidence intervals for these verbs do not overlap, indicating significant variation across verbs. Similarly, for Mandarin, Instrument judgments ranged from 19% (*chi1*, 'eat'; CI_{95} = 9%) to 41% (*ci4*, 'stab'; CI_{95} = 10%). The 95% confidence intervals for these verbs do not overlap.

Finally, we tested the relationship between individual verb

[2] In pilot studies, some Spanish-speakers reported that a Recipient phrase was incompatible with some Recipient verbs. Given this intuition, Spanish participants completed an acceptability questionnaire after the judgment task. If an individual participant judged a verb to be "unnatural" in the Recipient frame, this participant's data for this verb were excluded from analysis. The following percentages of trials were excluded for each verb: *patear* (*'kick'*): 49%; *empujar* (*'push'*): 74%; *llevar* (*'take'*): 11%; *tirar* (*'throw'*): 14%.

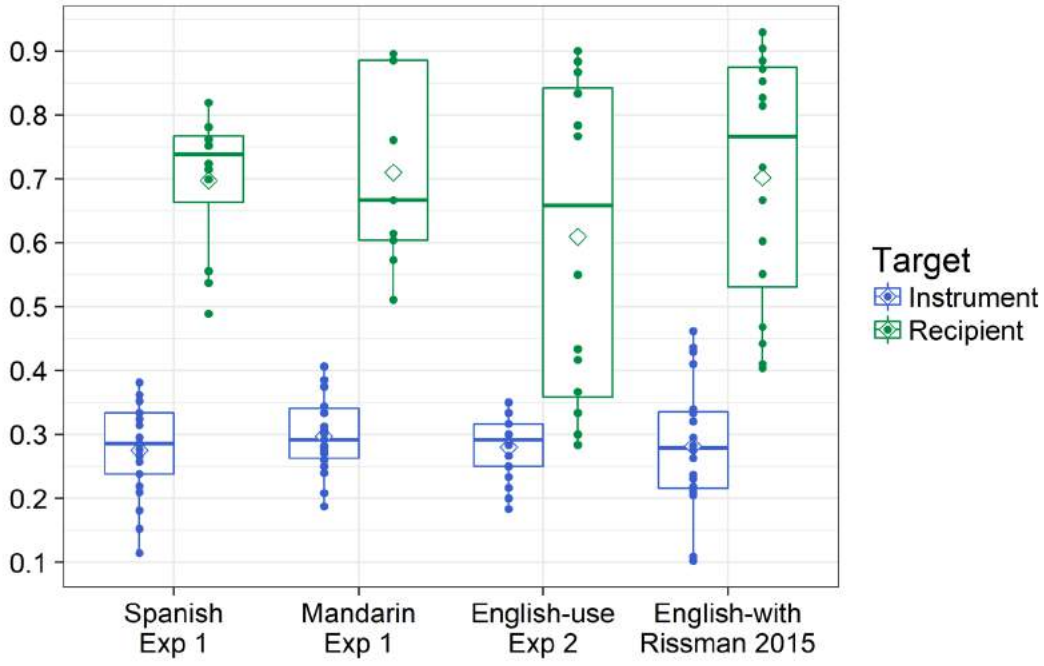


Figure 1. Rates of choosing the Instrument or the Recipient as an "argument" in each experiment. Box plots show median and second and third quartiles; diamonds show the mean; dots represent single verbs

meanings and "argument" judgments across languages, using the verb pairings shown in Tables 1-2. Judgments for individual verbs correlated significantly with each other for each verb category (Spanish-English Instrument verbs: $r(19) = .80$, $p < .001$; Mandarin-English Instrument verbs: $r(22) = .59$, $p < .01$; Spanish-English Recipient verbs: $r(11) = .79$, $p < .01$; Mandarin-English Recipient verbs: $r(11) = .72$, $p < .01$). These correlations show common trends in how verbal semantic features influenced the judgments in each language.

These results provide answers to the three questions raised above: in Spanish and Mandarin, judgments of Instrument salience do not parallel judgments of Recipient salience; some verbs highlight an Instrument more strongly than others; and verbs with similar meanings across languages give rise to similar judgments of Instrument salience. These findings support a gradient theory in which participants can have moderate degrees of salience, and suggest that verbs encode participants in similar ways across languages.

All participants had some knowledge of English. To assess whether English familiarity influenced the judgments, we calculated for each participant the correlation between that participant's judgments for each verb and the mean for the corresponding English verbs, combining Instrument and Recipient verbs. We then calculated correlations between the age at which a participant started learning English and the strength of their correlation with the English data. The correlation with age was non-significant for both Spanish ($r(33) = -.02$, $p > .1$) and Mandarin ($r(30) = -.08$, $p > .1$).

Experiment 2

In the English study, Instruments were introduced by the preposition *with*, whereas Mandarin Instruments were introduced by the verb *yong4*, 'use'. Thus in the English sentences, the Instrument was in the same clause as the main verb, while in the Mandarin sentences, the Instrument was in a separate clause. To assess a possible effect of these different syntactic structures, we collected judgments from English speakers who encountered Instruments in a *use*-frame.

Participants, Design, Materials and Procedure

Twenty English-speaking adults from Baltimore participated ($F = 14$). All subjects reported being native speakers of English. Subjects received $12 or course credit.

Each of the *with*-sentences from Rissman et al. (2015) was converted to a *use*-sentence. As in Experiment 1, the verb *use* was not included in the Instrument bracket. We used two different word orders for each Instrument vs. modifier contrast, e.g., *Jordan used [a shotgun] [in the driveway] to ATTACK someone* and *[In the driveway] Jordan used [a shotgun] to ATTACK someone*. All other trials were the same as in Rissman et al. (2015), as was the instruction.

Results and Discussion

Figure 1 shows the rates of choosing the Instrument/Recipient as an "argument" for Experiment 2. In a mixed-effects logistic regression model of the *with* data from Rissman et al. (2015) and the *use* data from Experiment

2, frame type (*use* vs. *with*) did not significantly affect the likelihood selecting the Instrument ($\chi^2(1) = .04$, $p > .1$). In addition, there was a significant positive correlation between the individual verb means for the English *use* data and the Mandarin data: $r(22) = .76$, $p < .001$. These results show that viewing the Instrument in a *use* frame did not decrease English speakers' likelihood of selecting the Instrument, mitigating the concern that the Mandarin stimuli from Experiment 1 underestimate the extent to which Mandarin verbs highlight Instruments.

General Discussion

Our results suggest that in Spanish and Mandarin, a discrete model of verbal participant encoding does not adequately capture how verbs encode the presence of an Instrument. Some theorists distinguish syntactic arguments from semantic arguments (Jackendoff, 2002). Such a distinction does not help explain our results, however, as both types of argument structures are assumed to be discrete.

It is possible that the gradient judgments we observe reflect probabilistic retrieval of discrete semantic structures (see Hale, 2001; Levy, 2008; among others). This approach, however, does not make an explicit connection between the semantics of a verb and the degree to which an Instrument is salient. Verbal semantics appears to matter: *picar*, 'chop' but not *comer*, 'eat' specifies that the Instrument has a particular physical form (a bladed/pointed shape). If *comer* and *picar* are both associated with 2 and 3-place frames, it is unclear how the semantic difference between the verbs accounts for the different rates of frame retrieval.

An alternate possibility is that the representation wherein verbs encode participant relations is itself gradient. This possibility has been characterized in multiple ways. Langacker (1987) proposes that verbs are conceptually dependent, and dependence is a gradient notion. For example, in *Jim sliced the bread with a knife*, the verb *slice* is dependent on the instrumental phrase *with a knife* because the instrument elaborates a salient substructure within the meaning of *slice,* the bladed-object feature. This salient substructure is not as salient, however, as the substructure indicating the entity that gets sliced, leading to gradient patterns of intuitions.

Similarly, Williams (2015) proposes that the "participant roles" of a verb are given by the "sketch" associated with that verb, a "psychological perspective…engaged by default" (85). Participant roles are entailed, explicit constituents within the sketch. Although Williams does not explicitly describe the sketch as non-discrete, he characterizes the elements of the sketch as psychologically "prominent." The results of Experiments 1-2 could be explained within this framework if: 1) an Instrument is a participant role for verbs such as *picar*, 'chop' and *ci4*, 'stab', and 2) the Instrument is less prominent in the sketch for these verbs than the Recipient is prominent in the sketch of dative verbs.

Rissman et al. (2015) propose a distinction between "primary" and "secondary" participants in event representation: the former are contributed by a discrete argument structure, whereas the latter are generated by the root semantics of the verb. Slicing events, for example, have two primary participants: the agentive causal force and the patient that becomes sliced. Through its root meaning, *slice* encodes that a bladed object comes into contact with the patient, and this bladed object is therefore a secondary participant within the event structure required by the verb. See Rissman and Rawlins (2017) for a proposal for how the instrument-phrase meaning interacts with this event structure.

More recently, Kim et al. (2019a,b) propose that the argument/adjunct distinction is gradient based on an idea from Dowty: certain phrases describing event participants can be *gradient argument/adjunct blends* in the framework of Smolensky et al (2014), i.e. they can be both arguments (to some degree) and adjuncts (to some degree). Kim et al. establish empirically that many prepositional phrases illustrate gradience in terms of whether native speakers categorize them as arguments or adjuncts. Their main aim is to explain variation and gradience in judgments about specific linguistic diagnostics that are supposed to provide evidence for the argument/adjunct distinction, e.g. that adjuncts allow pseudoclefts and that adjuncts are always omissible. Across all of the types of PPs they look at, lexical effects coming from particular verbs are a major factor in determining this gradience, and in particular, verbs vary in the degree to which they prefer for some potential event participant role to be filled; different syntactic frames vary in how much they prefer adjunct phrases. While this proposal makes no specific claims about instrument marking, it does generally predict that verbs will have gradient representations in terms of how they license event participants, something consistent with our results. An open question is whether the very general kinds of verb preferences that Kim et al. show across many PP types can explain the role-specific preferences demonstrated here and in Rissman et al. (2015).

Argument omission is more widespread in Spanish and Mandarin than in English. We did not, however, find a main effect of Language on the judgments. In addition, English speakers' judgments were largely unchanged when they encountered Instruments in a *use*-frame rather than a *with*-frame. These results suggest that the judgments reflect verbal meaning rather than syntactic prominence per se. Given this hypothesized dissociation between syntactic and semantic prominence in this task, we predict that if participants judged whether *the key* is an "argument" in *[The key] OPENED the door*, they would be unlikely to do so.

Across Spanish, Mandarin and English, we observe similarity rather than variability: there are verbs in all three languages where the Instrument has an intermediate level of salience (e.g., *picar*, 'chop,' and *ci4*, 'stab'). We leave future research to explore the interaction of discrete and non-discrete structures that give rise to these gradient judgments.

Acknowledgments

This research was supported by NSF IGERT grant #9972807, NIH RO1 DC000491 and a Radboud Excellence Initiative fellowship awarded to Lilia Rissman. Thank you to Paul Smolensky, Akira Omaki, Colin Wilson, Jenny Culbertson. Thank you also to research assistants Aurora Martinez del Rio, Danny Salevitz, Yijia Hu, Michelle Chu, Lina Montoya, Allison Bellows and Christine Cheseborough.

References

Aarts, B. (2007). *Syntactic gradience: The nature of grammatical indeterminacy*. Oxford: OUP.

Andreu, L., Sanz-Torrent, M., & Rodríguez-Ferreiro, J. (2016). Do Children with SLI Use Verbs to Predict Arguments and Adjuncts: Evidence from Eye Movements During Listening. *Frontiers in Psychology, 6*(1917).

Barbu, R.-M., & Toivonen, I. (2016). Event participants and linguistic arguments. In *Proceedings of the 38th Annual Meeting of the Cognitive Science Society* (pp. 1961-1966).

Bates, D., & Maechler, M. (2009). lme4: linear mixed effects models using S4 classes.

Bod, R., Hay, J., & Jannedy, S. (Eds.). (2003). *Probabilistic linguistics*. Cambridge, MA.: MIT Press.

Boland, J. E. (2005). Visual arguments. *Cognition, 95*(3), 237-274.

Bornkessel, I., Schlesewsky, M., Comrie, B., & Friederici, A. (Eds.). (2006). *Semantic role universals and argument linking: theoretical, typological and psycholinguistic perspectives*. Berlin: Mouton de Gruyter.

Bowerman, M., & Brown, P. (2008). Introduction. In M. Bowerman & P. Brown (Eds.), *Crosslinguistic perspectives on argument structure: Implications for learnability*. New York: Lawrence Erlbaum.

Carlson, G., & Tanenhaus, M. (1989). Thematic roles and language comprehension. In G. Carlson & M. Tanenhaus (Eds.), *Linguistic structure in language processing:* (pp. 413). Dordrecht: Kluwer Academic Publishers.

Croft, W. (2001). *Radical construction grammar: syntactic theory in typological perspective*. Oxford: OUP.

Dowty, D. (2003). The dual analysis of adjuncts/complements in Categorial Grammar. In E. Lang, C. Maienborn, & C. Fabricius-Hansen (Eds.), *Modifying adjuncts* (pp. 33-66). Berlin: Mouton de Gruyter.

Dummett, M. A. (1981). *Frege: Philosophy of language* (Vol. 2): Cambridge University Press.

Haspelmath, M. (2014). Arguments and adjuncts as language-particular syntactic categories and as comparative concepts. *Linguistic Discovery, 12*(2), 3-11.

Jackendoff, R. (1972). *Semantic interpretation in generative grammar*. Cambridge, MA.: MIT Press.

Jackendoff, R. (2002). *Foundations of language: brain, meaning, grammar, evolution*. Oxford: OUP.

Kim, N., Rawlins, K., Van Durme, B., & Smolensky, P. (to appear). Predicting Argumenthood of English Preposition Phrases. *AAAI 2019 Proceedings*.

Kim, N., Rawlins, K., Van Durme, B., & Smolensky, P. (2019). The Complement-Adjunct distinction as Gradient Blends. Manuscript, JHU.

Koenig, J.-P., Mauner, G., & Bienvenue, B. (2003). Arguments for Adjuncts. *Cognition, 89*(2), 67-103.

Langacker, R. W. (1987). *Foundations of cognitive grammar*. Stanford, CA.: Stanford University Press.

Larson, R. K. (1988). On the double object construction. *Linguistic Inquiry, 19*(3), 335-391.

Levin, B., & Rappaport-Hovav, M. (2005). *Argument realization*. Cambridge: Cambridge University Press.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition, 106*(3), 1126-1177.

Li, Y. (1990). On VV compounds in Chinese. *Natural Language & Linguistic Theory, 8*(2), 177-207.

Majid, A., Boster, J. S., & Bowerman, M. (2008). The cross-linguistic categorization of everyday events: A study of cutting and breaking. *Cognition, 109*(2), 235-250.

Parsons, T. (1990). *Events in the semantics of English: a study in subatomic semantics*. Cambridge, MA: MIT Press.

Rissman, L. (2018). "Tools for understanding verb meaning: explicit judgments vs. implicit behavior." Paper presentation at the Linguistic Society of America meeting.

Rissman, L., & Rawlins, K. (2017). Ingredients of Instrumental Meaning. *Journal of Semantics*, 34:3,507-537.

Rissman, L., Rawlins, K., & Landau, B. (2015). Using instruments to understand argument structure: Evidence for gradient representation. *Cognition, 142*(0), 266-290.

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General, 104*(3), 192.

Schutze, C. T. (1995). PP Attachment and Argumenthood. *MIT Working Papers in Linguistics, 26*(Sept), 95-151.

Smolensky, P., & Legendre, G. (2006). *The harmonic mind: From neural computation to optimality-theoretic grammar*. Cambridge, MA: MIT Press.

Smolensky, P., Goldrick, M. and Mathis, D. (2014). Optimization and quantization in gradient symbol systems: a framework for integrating the continuous and the discrete in cognition. *Cognitive Science*, 38, 1102-1138.

Talmy, L. (1985). Lexicalization patterns: Semantic structure in lexical forms. *Language typology and syntactic description, 3*, 57-149.

Vater, H. (1978). On the possibility of distinguishing between complements and adjuncts. In *Valence, semantic case, and grammatical relations*. Amsterdam: John Benjamins.

Wilkins, D. (2008). Same argument structure, Different Meanings: Learning 'Put' and 'Look' in Arrernte. In M. Bowerman & P. Brown (Eds.), *Crosslinguistic perspectives on argument structure: Implications for learnability*. New York: Lawrence Erlbaum.

Williams, A. (2015). *Arguments in syntax and semantics*. Cambridge: Cambridge University Press.

Wittenberg, E., & Snedeker, J. (2014). It takes two to kiss, but does it take three to give a kiss? Categorization based on thematic roles. *Language and Cognitive Processes*, *29*(5), 635-641.

Parametric control of distractor-oriented attention

Harrison Ritz & Amitai Shenhav
Cognitive, Linguistic & Psychological Sciences;
Carney Institute for Brain Science;
Brown University, Providence, RI, 02912

Corresponding Author: *hritz@brown.edu*

Abstract

Traditional models of cognitive control account for a host of classic findings, but these classic tasks have limited our ability to test a broader range of model predictions. In particular, such models predict that control should vary parametrically in response to cognitive demands and that control adjustments should be targeted towards task-relevant stimulus features. We developed a task to probe these predictions across two experiments. Participants responded to one dimension of a stimulus while ignoring the other, and we parametrically varied the conflict between those dimensions and the predictability of this conflict across trials. We found that control adjustments (1) varied parametrically in response to cognitive demands, (2) were sensitive to the predictability of those demands, and (3) were primarily targeted towards task-irrelevant dimensions. These results raise interesting questions about the structure of cognitive control and demonstrate the utility of rich tasks for constraining model predictions.

Keywords: cognitive control; attention; conflict adaptation

Introduction

Cognitive control is vital for adaptive behavior, allowing the brain to balance the consistency of automatic behavior against the flexibility to rapidly perform arbitrary tasks (Miller & Cohen, 2001). Influential models of cognitive control have proposed supervisory processes that parametrically adjust the strength of task-relevant information based on conflict or (dis)utility (Botvinick, Braver, Barch, Carter, & Cohen, 2001; Shenhav, Botvinick, & Cohen, 2013). While these models have successfully explained a host of classic findings in executive control, the evidence from these classic tasks is limited in its ability to constrain models of control. In this experiment, we sought to test several key assumptions of cognitive control using enriched tasks that can better discriminate between different model architectures.

The first feature of control models that has been virtually untested is the parametric nature of adjustments to control. Control adjustments are often examined in reaction to response conflict, but existing paradigms typically vary such conflict in an all-or-none fashion (i.e., stimulus dimensions activate only one response or they activate responses that are fully congruent or fully incongruent). Researchers have studied more granular control adjustments over longer timescales, for instance by varying the overall proportion of incongruent trials at the list level (Logan & Zbrodoff, 1979; Bugg, Jacoby, & Toth, 2008), however parametric manipulations at the single-trial level remain largely unexplored. As a result of this methodological gap in the literature, little is known about how the intensity of control changes when response congruence varies parametrically. A secondary benefit to parametric congruence is that it allows participants to more accurately track changes in congruence over trials, providing clearer evidence for learning-based adjustments (Jiang, Beck, Heller, & Egner, 2015).

The second feature of control models that we sought to test was the assumption that control primarily acts to enhance attention towards targets ('target-oriented' control; Botvinick et al., 2001; Egner, 2007). This assumption is poorly constrained by most studies of response conflict, as they typically only vary the strength of the distractor dimension, and not the target dimension. As a result, existing data cannot distinguish between conflict-related control adjustments that are primarily oriented toward targets, distractors, or both.

To address these gaps in the literature, we developed a novel cognitive control task that varies the strength in the target and/or distractor dimensions of a stimulus, resulting in fine-grained variation in response congruence. We also varied the predictability of this congruence, in order to measure how participants learn to control attention. We found that participant's performance depended on both parametric task demands and parametric control adjustments. In periods when distractor congruence was highly predictable, participants became more sensitive to distractor information. Finally, we found that participants primarily controlled their attention towards distractor dimensions, counter to the predictions of prominent cognitive control models. These experiment demonstrate the need for richer cognitive control tasks that can better distinguish between models of executive functioning.

Experiment 1

Experiment 1 sought to test (1) whether there is a parametric relationship between performance and response congruence; (2) whether participants parametrically adjust control based on recent task demands; and (3) how these control adjustments depend on the learned task demands over longer timescales.

Method

Participants Fifty-eight individuals participated in Experiment 1 for course credit or pay (Mean(SD) age = 20.9(2.6);

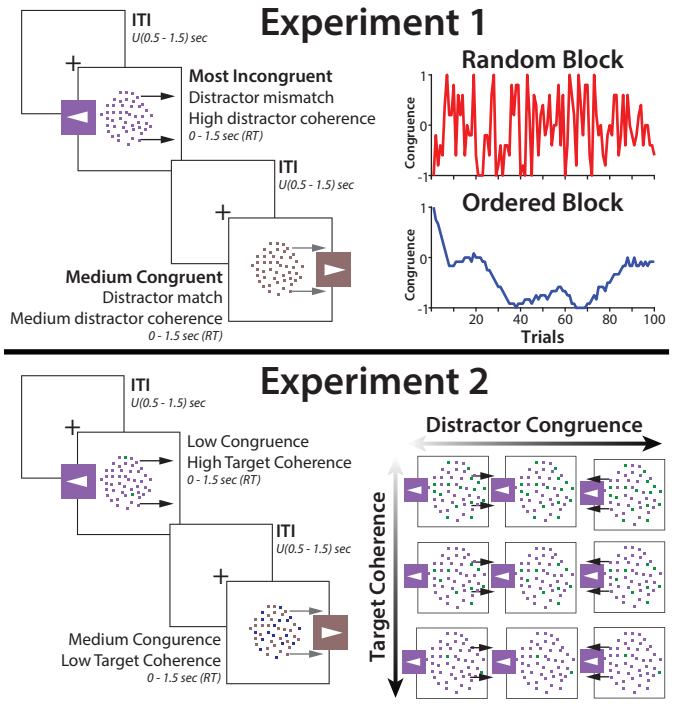


Figure 1: Top Left: In Experiment 1, participants responded to the color and ignored the motion of an array of pseudo-randomly moving dots. Motion coherence induced variable levels of congruence across dimensions. Top Right: Participants performed blocks in which the congruence changed randomly (red) or predictably (blue). Bottom Left: In Experiment 2, both the color and motion dimensions had variable coherence. Bottom Right: These color and motion dimensions were orthogonal

41 females). All participants across all experiments provided informed consent in compliance with our University's Institutional Review Board.

Parametric Conflict Task We developed a parametric version of a previous Simon-like conflict task (Danielmeier, Eichele, Forstmann, Tittgemeyer, & Ullsperger, 2011). On each trial, participants viewed an array of moving dots, presented in one of four colors (see Figure 1). Participants were instructed to press either the left or right key associated with the color of the dots. Each keys was mapped to two possible colors. The direction of the dot motion (leftward or rightward) was task-irrelevant and could be consistent with the response hand for the correct color response (*congruent* trials) or it could be inconsistent with this response hand (*incongruent* trials). To avoid feature priming (Hommel, Proctor, & Vu, 2004), colors did not repeat on adjacent trials.

Uniquely in this experiment, we parametrically varied the *degree* of response congruence on a given trial by varying the coherence of the dot motion (the % of dots moving in a given direction). Congruence was evenly sampled between 100% coherent congruence and 100% coherent incongruence, and was treated as a continuous variable in statistical analyses.

To maintain the salience of the motion dimension throughout the session, participants alternated between blocks of the task above (*color-response* trials) and blocks where participants were instructed to instead indicate the direction of motion (*motion-response* trials; cf. Schneider & Shiffrin, 1977). Mirroring color-response trials, motion coherence was held constant (maximal) during motion-response blocks, while color coherence (the proportion of one color vs. another) was varied across trials.

Procedure Participants first performed 100 motion-only training trials (0% coherent color) and 100 color-only training trials (0% coherent motion) to learn the stimulus-response mappings. During the main experiment, participants performed two types of trial blocks. During *Random* blocks, the distractor congruence varied randomly from trial-to-trial. During *Ordered* blocks, congruence linearly increased and decreased in a predictable manner (see Figure 1).

Variants Data for Experiment 1 incorporate several similar versions of this task. The main differences across versions was the number of congruence levels (mean(range) = 13.5(11-15) levels for Random blocks, mean(range) = 15.4(11-25) levels for Ordered blocks), as well as the number of trials in each block type (mean(range) = 469(300-700) for Random blocks, mean(range) = 643(300-800) for Ordered blocks). We did not find significant differences in performance across versions, nor interactions between task version and our effects of interest, and so our analyses collapse across these versions. Importantly, versions only differed in ways that should produce random rather than systematic error, potentially making our positive findings more conservative.

Results

All analyses were performed using linear mixed effects modelling in MATLAB (lmefit and glmefit). The dependent variables across analyses were log-transformed reaction time (RT) and accuracy. All models included a 'maximal' random effects structure at the participant level and intercept terms (not reported). All analyses excluded trials with RTs faster than 200ms, RT analyses excluded incorrect trials, and adaption analyses required the previous trial to also be accurate and have an RT longer than 200ms. We estimated the effective degrees of freedom with the Satterthwaite approximation for RT models, and used $(n_{Participants} - n_{Predictors})$ for accuracy models. Models were compared on the basis of Akaike Information Criterion (AIC), a goodness-of-fit metric that penalizes model complexity.

Parametric Within-trial Interference Effects Within Random blocks, we found that RT and accuracy varied linearly with our parametric manipulation of congruency (see Figure 2; Table 1). As confirmation that performance varied parametrically across congruence levels, we found that a model that treated congruence as a single continuous variable

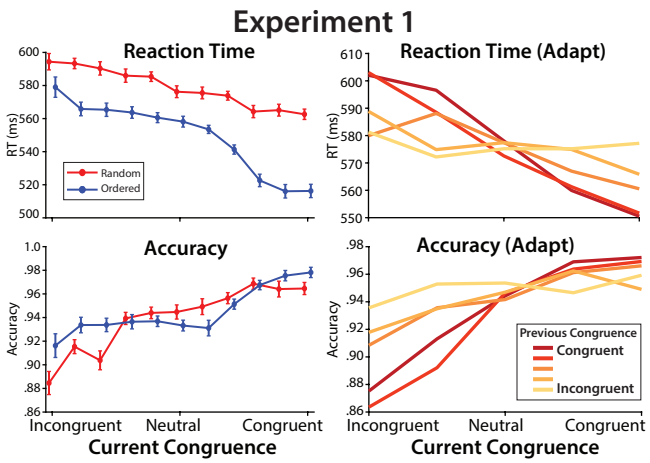


Figure 2: Left: RT (top) and accuracy (bottom) linearly depended on the degree of distractor congruency, moreso in Ordered blocks (blue) than Random blocks (red). Right: The influence of congruence on RT (top) and accuracy (bottom) linearly depended on the previous level of congruence. In all graphs, error bares indicate within-participant SEM.

fit better than a model that treated congruence as a binary variable (congruent vs. incongruent) and a model with separate congruence slopes for trials with target-compatible vs. target-incompatible coherence levels (i.e., levels of congruency vs. levels of incongruency).

Table 1: Parametric Congruence (Exp 1)
performance $\sim$ *congruence*

DV	IV	β	*t*(df)	*p*
logRT	cong	-0.030	-8.0(60)	4.5e-11
accuracy	cong	0.74	9.6(57)	8.2e-14

Parametric Between-trial Adaptation Effects These initial analyses suggest that we were successful in parametrically varying cognitive demands across trials. To test whether this manipulation in turn led to parametric variations in the strength of *control* allocated to the task, we tested how participants' performance changed after a trial that was more or less demanding (i.e., the 'Gratton' or conflict adaptation effect; Gratton, Coles, & Donchin, 1992). Consistent with previous findings of such adaptation effects, we found that RT and accuracy on the current trial was predicted by the interaction between the congruence of the current and previous trials, with stronger congruence on one trial predicting stronger distractor sensitivity on the next trial. Importantly, these adaptation effects were present over and above the effect of current-trial congruency and – like those within-trial effects – *also* varied parametrically (see Figure 2; Table 2).

Interestingly, we found that the previous trial's congruence alone had little influence on current-trial performance, instead modulating the degree to which performance was facilitated by or interfered with by the distractor's current congruence. When the distractor was previously more congruent (i.e., more associated with a correct response), participants incorporated more distractor information into their response; when the previous distractor was more incongruent, participants' performance was virtually independent of the current degree of congruence.

Table 2: Conflict Adaptation (Exp 1)
$performance_t \sim congruence_t * congruence_{t-1}$

DV	IV	β	*t*(df)	*p*
logRT	$cong_t$	-0.030	-8.1(59)	4.2e-11
	$cong_{t-1}$	9.1e-4	0.33(58)	.75
	$cong_t$:$cong_{t-1}$	-0.028	-6.2(56)	7.5e-08
accuracy	$cong_t$	0.69	9.3(57)	5.0e-13
	$cong_{t-1}$	-.099	-1.9(57)	.06
	$cong_t$:$cong_{t-1}$	0.47	6.3(56)	4.9e-8

Influence of Demand Predictability on Control Allocation To determine how control adjustments changed when task difficulty was highly predictable, we compared congruence effects across Random and Ordered blocks (see Figure 2; Table 3). We predicted that participants would match their control allocation to local demands, resulting in weaker congruence effects during Ordered blocks, and better overall performance. While we found that participants were overall faster in Ordered block, they were less accurate, and we found that RTs were in fact more influenced by congruence during Ordered relative to Random blocks.

Table 3: Block Effects (Exp 1)
performance $\sim$ *block*(congruence + coherence)*

DV	IV	β	*t*(df)	*p*
logRT	block	-0.033	-4.0(57)	2.2e-4
	cong	-0.030	-8.0(59)	5.7e-11
	coh	6.1e-4	-0.14(53)	0.89
	block:cong	-0.023	-4.6(59)	2.3e-5
	block:coh	-0.035	-4.7(55)	1.9e-5
accuracy	block	-0.37	-3.9(57)	2.9e-4
	cong	0.69	9.5(57)	2.8e-13
	coh	-0.22	-2.2(57)	.031
	block:cong	-0.016	-0.18(56)	.86
	block:coh	0.91	6.4(56)	3.9e-8

In addition to the block difference in congruence, we found that participant's performance was enhanced when there was greater distractor coherence in Ordered blocks, regardless of whether the distractor was congruent or incongruent with the target. This is consistent with participants learning to use distractor information, i.e., responding in the same or opposite direction of the distractor. In sum, the influence of distractors

on choice was enhanced when the they could be used to make accurate responses, with a stronger bias towards distractor-congruent trials. Interestingly, these effects were present in spite of most participants reporting that they had not noticed the predictability manipulation.

Relative Automaticity of Motion vs. Color Processing We designed our task under the assumption that the response compatibility of the motion dimension would make responding to it more automatic than responding to the color dimension. To validate this assumption, we tested whether these dimensions would interfere with one another asymmetrically (Schneider Shiffrin, 1977). Consistent with this prediction, we found that when participants were instructed to respond based on the motion dimension (rather than color), we did not observe any interference effects associated with the congruency of the color dimension (logRT: b = 5.9e-4, p = .33; accuracy: b = 0.031 p = .088; compare to Table 1), in stark contrast with the results reported above for color-response trials.

Discussion

Experiment 1 sheds new light on how attention is parametrically controlled based on local and long-term task demands. First, we observed that performance depends on the continuous degree of interference, supporting participants' ability to track parametric task demands and control their attention accordingly.

The second major observation from this experiment was that participants parametrically adjust their sensitivity to distracting information based on the degree of interference they previously experienced. Interestingly, we found that participants' performance was not strongly modulated by previous congruence per se, but that the previous congruence influenced distractor sensitivity. This is largely consistent with traditional conflict adaptation effects, which are commonly attributed to a controlled increase in attention towards targets following incongruent trials, which reduce the influence of distractors as a secondary effect (Botvinick et al., 2001; Egner, 2007). In contrast to these models' predictions, the effect of previous congruence was evaluated when the current congruence was neutral (0% coherence), the situation where target enhancement should should be most obvious. Our results are more consistent with changes in distractor processing than target processing.

Finally, we found that under conditions of high predictability, participants increased their attention towards distractors when they were informative (i.e., provided coherent evidence for or against a response), but with a strong bias towards distractors that provided target-congruent evidence. This observation is consistent with the literature on the proportion congruency effect (i.e., weaker congruency effects in blocks of majority-incongruent trials; Logan & Zbrodoff, 1979), originally attributed to participants' learning the predictive value of different stimulus dimensions. This strict learning account does not predict a bias towards distractor-congruent information, making our results more compatible with models that combine learning to weight different cues with adjustments based on the recent history of conflict (Jones, Cho, Nystrom, Cohen, & Braver, 2002).

In sum, these results suggest that participants controlled their attention towards the distracting dimension based on both the learned value of this cue and a bias towards congruent distractors. However, this preliminary evidence for distractor-oriented control is limited by the standard convention of only manipulating distractor congruence. To better isolate control adjustments towards targets and distractors, in Experiment 2 we manipulated the coherence of each dimension to better measure where participants controlled their attention.

Experiment 2

Experiment 2 sought to further characterize the targets of control adjustment in this task. In particular, we examined the degree to which participants adjust their attention towards targets and distractors in response to the demands associated with each dimension. We measured this by independently manipulating the coherence of both the target and distractor dimensions. By 'tagging' these different stimulus dimensions, we sought to determine where participants adjust attention. The traditional target-oriented attention account makes two key predictions: first, if distractor sensitivity is a byproduct of control towards targets (e.g., due to lateral inhibition; Botvinick et al., 2001), then the influence of target and distractor information should strongly interact within a trial. Secondly, we should find that trial-to-trial adjustments to control should primarily influence the sensitivity to the target dimension.

Method

Participants Thirty-three individuals participated in Experiment 2 for course credit or pay (Mean(SD) age = 18.9(0.45); 24 females).

Task & Procedure This task was similar to Experiment 1, except that we varied the coherence of both the distractor (motion) and target (color) dimensions. As in Experiments 1, motion coherence varied from 100% leftward to 100% rightward (11 levels of congruence). Target coherence (i.e., the proportion of dots whose color indicates a leftward or rightward response) varied from 65% to 95% (11 levels of coherence). Participants only performed Random blocks (1200 color-response trials with interleaved motion-response blocks, as in Experiment 1), with the target coherence and distractor congruence independently sampled on every trial.

Results

We used the same linear mixed effects regression approach here as we did in Experiment 1. Target coherence was mean-centered within participants to aid in interpretability.

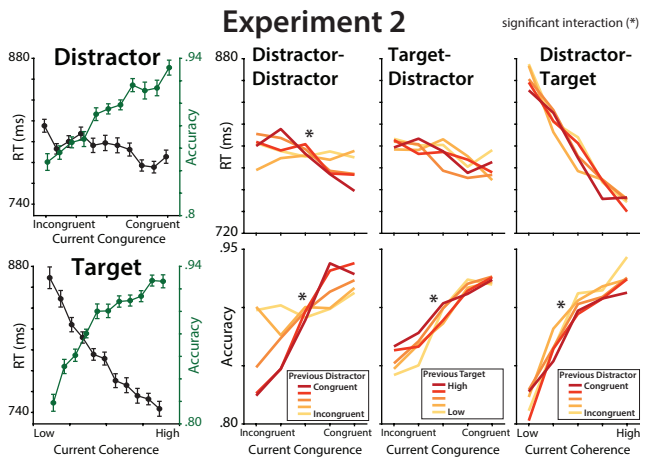


Figure 3: Left: RT (black) and accuracy (green) linearly depended on both distractor congruence (top) and target coherence (bottom). Right: Conflict adaptation was stronger and more consistent for the distractor dimensions. Asterisks indicate significant interactions.

Within-trial Effects of Target and Distractor Information As predicted, we found that performance improved with both greater target coherence and greater distractor congruence within each trial (see Figure 3; Table 4). Interestingly, these dimensions influenced performance largely independently of one another. The interaction between target and distractor information had a significant effect on accuracy ($p<0.01$) but not RT ($p=0.79$), However, relative to models with only main effects, models that included these interaction terms did not improve overall model fit for either RT (Δ AIC = 14) or accuracy (Δ AIC = 0.80).

Table 4: Target & Distractor Within-Trial (Exp 2)
performance $\sim$ *(distractor * target)*

DV	IV	β	*t*(df)	*p*
logRT	dist	0.031	-7.1(32)	4.6e-8
	targ	-0.20	-12(32)	1.2e-13
	dist:targ	-0.0038	-0.26(53)	.79
accuracy	dist	-0.46	-8.6(32)	8.2e-10
	targ	2.3	11(32)	6.1e-12
	dist:targ	-0.44	-2.8(31)	.0094

Target- vs. Distractor-Dependent Adaptation While both target coherence and distractor congruence influenced participants' performance within a given trial, our primary interest was how participants adjust attention from trial to trial. To investigate this, we measured how participants' sensitivity to target and distractor information changed as a function of the previous trial difficulty. The prediction of traditional target-oriented accounts is that previous task demands will most strongly change sensitivity to the target dimension.

We found that the previous distractor congruence strongly influenced participants' sensitivity to the current distractor congruence, replicating our parametric conflict adaption results from Experiment 1 (see Figure 3; Table 5). In contrast to traditional models, we found that the previous distractor had an inconsistent influence over participants' sensitivity to the current target coherence, appearing in the domain of RT but not accuracy. Interestingly, the previous trial's congruence had opposing effects on targets and distractors: more incongruent trials were followed by weaker sensitivity to distractors and stronger sensitivity to targets, albeit with substantively weaker adjustments to target sensitivity.

Table 5: Distractor-Dependent Adaptation (Exp 2)
$performance_t \sim distractor_{t-1} * (distractor_t + target_t)$

DV	IV	β	*t*(df)	*p*
logRT	$dist_{t-1}$	-8.9e-5	-0.03(94)	.97
	$dist_t$	-0.031	-7.2(33)	4.6e-8
	$targ_t$	-0.20	-12(32)	3.4e-13
	$dist_{t-1}$:$dist_t$	-0.028	-6.5(191)	7.3e-10
	$dist_{t-1}$:$targ_t$	-0.012	-0.26(406)	.39
accuracy	$dist_{t-1}$	-0.07	-2.1(32)	.047
	$dist_t$	0.41	8.1(32)	3.4e-9
	$targ_t$	2.3	11(32)	3.2e-12
	$dist_{t-1}$:$dist_t$	0.46	7.6(31)	1.5e-8
	$dist_{t-1}$:$targ_t$	-0.39	-2.1(31)	.043

We also tested a model where the previous target coherence could influence sensitivity towards the current target and distractor (see Figure 3; Table 6). We found, again, no evidence in reaction time that previous target coherence influenced the current target or distractor sensitivity. However, in accuracy we found that weaker previous trial target coherence predicted a stronger reliance on distractor information on the next trial, with no change to the reliance on target information.

Table 6: Target-Dependent Adaptation (Exp 2)
$performance_t \sim target_{t-1} * (distractor_t + target_t)$

DV	IV	β	*t*(df)	*p*
logRT	$dist_t$	-0.031	-7.2(32)	3.0e-8
	$targ_t$	-0.20	-12(32)	3.7e-13
	$targ_{t-1}$	-0.011	-0.98(32)	.34
	$targ_{t-1}$:$dist_t$	-0.0028	-0.18(37)	.86
	$targ_{t-1}$:$targ_t$	0.038	0.73(43)	.47
accuracy	$dist_t$	0.42	8.5(32)	1.2e-9
	$targ_t$	2.3	11(32)	2.9e-12
	$targ_{t-1}$	0.12	1.1(32)	.28
	$targ_{t-1}$:$dist_t$	-0.41	-2.5(31)	.012
	$targ_{t-1}$:$targ_t$	-0.12	-0.22(31)	.83

Overall, target- and distractor-dependent adaptation seem to support a similar mechanism, in which the response-relevance of a dimension modifies the extent to which it is subsequent used for choice, with a strong bias towards modifying the distractor dimension. When the target dimension is incongruent with the response, participants are subsequently more influenced by target and less influenced by distractors. When the target provided weak evidence for the response, participants were subsequently more sensitive to distractors.

Discussion

Experiment 2 replicated the within- and between-trial congruence effects observed in Experiment 1. Critically, Experiment 2 also provided unique evidence in favor of distractor-oriented attentional control in this task.

Within trials, we found that both target and distractor information influenced task performance, there were only weak interactions across dimensions. This runs counter to the predictions from models that posit competitive interactions between the processing of targets and distractors, in which distractor sensitivity changes as a byproduct of target-oriented control (Botvinick et al., 2001; Egner, 2007).

Across trials, we found that adjustments to control primarily acted on distractors, in contrast with traditional models of conflict adaptation. When the previous trial was difficult, participants suppressed distractors if the difficulty was due to incongruent distractors, and enhanced distractors if the difficulty was due to low-coherence targets. It is notable that the latter effect of distractor enhancement appeared to be specific to accuracy, whereas the suppression effect was observed in both speed and accuracy. Whether these reflects different forms of control adjustment (e.g., related to evidence accumulation versus response threshold) demands further investigation with models that can distinguish these processes (e.g., the drift diffusion model).

In addition to these distractor adjustments, we also observed adjustments to target sensitivity in one condition (distractor-dependent adaptation effects in accuracy). However these adjustments to target processing were very subtle, compared to the strong and reliable adaptation effects observed for the distractor dimension, and could plausibly represent a byproduct of these dominant adjustments to distractor processing.

General Discussion

We developed a novel task aimed at examining parametric adjustments of control towards targets and distractors. Across our two experiments, we found consistent evidence that participants parametrically controlled their attention towards distractors based on the recent history of task demands. In Experiment 1, we found that participants adjusted their sensitivity to distractor congruence based on both whether distractors could predict the accurate response, alongside the bias towards congruency predicted by conflict monitoring. In Experiment 2, we narrowed down the sources and targets of this process of control adaptation. We found that participants adjusted attention towards distractors much more than they did towards targets. Together, these results provide strong confirmation for many aspects of existing models of cognitive control, while challenging models that propose unbiased or target-oriented attentional control.

Our experiments leave open the question of why participants would be biased towards distractor-oriented attention. One reason for this asymmetry may due to a primacy for inhibition in cognitive control, exemplified by the well-characterized 'hyperdirect' control of striatal decision-making (Wiecki & Frank, 2013) and the common inhibition factor found across several executive control tasks (Friedman & Miyake, 2017). This may describe why it is easier to (dis)inhibit attention towards distractors, rather than enhance attention to targets, but offers little explanation for *why* there is this preference for inhibition. Another reason for our asymmetry may be that our distracting motion dimension, like many distractors, is easier to control because of its salience. Insofar as attention control requires some form of feature selection, it may be easier to select a distractor's stimulus features to act upon. Finally, this experiment cannot rule out that there is something about motion *per se* that makes it easier to control. Future experiment should test the robustness of these results across multiple stimulus domains and forms of congruence before making more provocative conclusions about the nature of cognitive control.

References

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological review*, *108*(3), 624.

Bugg, J. M., Jacoby, L. L., & Toth, J. P. (2008). Multiple levels of control in the stroop task. *Memory & cognition*, *36*(8), 1484–1494.

Danielmeier, C., Eichele, T., Forstmann, B. U., Tittgemeyer, M., & Ullsperger, M. (2011). Posterior medial frontal cortex activity predicts post-error adaptations in task-related visual and motor areas. *Journal of Neuroscience*, *31*(5), 1780–1789.

Egner, T. (2007). Congruency sequence effects and cognitive control. *Cognitive, Affective, & Behavioral Neuroscience*, *7*(4), 380–390.

Friedman, N. P., & Miyake, A. (2017). Unity and diversity of executive functions: Individual differences as a window on cognitive structure. *Cortex*, *86*, 186–204.

Gratton, G., Coles, M. G., & Donchin, E. (1992). Optimizing the use of information: strategic control of activation of responses. *Journal of Experimental Psychology: General*, *121*(4), 480.

Hommel, B., Proctor, R. W., & Vu, K.-P. L. (2004). A feature-integration account of sequential effects in the simon task. *Psychological research*, *68*(1), 1–17.

Jiang, J., Beck, J., Heller, K., & Egner, T. (2015). An insula-frontostriatal network mediates flexible cognitive control

by adaptively predicting changing control demands. *Nature communications*, *6*, 8165.

Jones, A. D., Cho, R. Y., Nystrom, L. E., Cohen, J. D., & Braver, T. S. (2002). A computational model of anterior cingulate function in speeded response tasks: Effects of frequency, sequence, and conflict. *Cognitive, Affective, & Behavioral Neuroscience*, *2*(4), 300–317.

Logan, G. D., & Zbrodoff, N. J. (1979). When it helps to be misled: Facilitative effects of increasing the frequency of conflicting stimuli in a stroop-like task. *Memory & cognition*, *7*(3), 166–174.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual review of neuroscience*, *24*(1), 167–202.

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. detection, search, and attention. *Psychological review*, *84*(1), 1.

Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2013). The expected value of control: an integrative theory of anterior cingulate cortex function. *Neuron*, *79*(2), 217–240.

Wiecki, T. V., & Frank, M. J. (2013). A computational model of inhibitory control in frontal cortex and basal ganglia. *Psychological review*, *120*(2), 329.

A Definition of Memory for the Cognitive Sciences

Brett A. Ross (brettross96@knights.ucf.edu)
Department of Electrical Engineering and Computer Science, 4328 Scorpius Street
Orlando, FL 32816-2362 USA

Luis H. Favela (luis.favela@ucf.edu)
Department of Philosophy and Cognitive Sciences Program, 4111 Pictor Lane, Suite 220
Orlando, FL 32816-1352 USA

Abstract

We provide a definition of 'memory' that is broad enough to apply to both natural and artificial systems. Inspired by computation and information theory, we define memory as a process that preserves information through time while maintaining its usefulness as an object to be computed. We defend the extensiveness of our definition by explaining how it applies to both brains and modern computers. We then consider potential objections to our definition. Our primary goal is to provide a definition of 'memory' that is broadly applicable across various cognitive sciences subfields.

Keywords: memory; computation; representation; information

Introduction

Memory is a central topic within the cognitive sciences and its various contributing disciplines, such as computer science, neuroscience, and psychology. One likely reason for this is its centrality to various conceptions of cognition. Be it brains or modern computers, memory typically plays a central role. However, it is often unclear if 'memory' is used the same across contexts. What is apparent, however, is the efficacy of computational theory in the cognitive sciences. Given the successes computational theory has provided the study of cognition, and given that memory is central to computation, it follows that a computationally-inspired approach to memory can provide useful insights into the general nature of memory. As such, it is necessary for us to explicate the relevant features of information and computation before discussing our definition of memory.

We begin with Piccinini and Scarantino's (2011) definition of 'computation' as the processing of objects according to rules. Next, we connect that definition to Gallistel and King's (2010) interpretation of Shannon's classic information theory—that is, the reduction in uncertainty regarding the properties of an object—in order to show how computation allows for useful decisions about the world to be made. We then discuss some of the properties necessary for effective information processing and generic computation that describe modern computers, which may also be usefully applied to descriptions of brains as well. We pay special attention to the topic of representation. The ability of the definition to allow for determination of the boundaries of a computational system's memory are examined. Finally, we present and respond to some potential critiques of the definition.

The definition these claims and terms are applied to is as follows: *Memory is a process that carries information forward in time, preserved in a fashion that maintains its usefulness as an object to be computed for the system to which the memory is said to belong*.

Defining Computation and Information

Before presenting our definition of 'memory,' we must first establish definitions for 'information' and 'computation.' We begin with information because, as will be discussed below, computation does not necessarily need to involve information—though it can be more useful when it does. Gallistel and King relate Shannon's definition of 'information' as originating from a source, undergoing a process that 'encodes' the information into a 'signal,' and traveling to a receiver that 'decodes' the signal to derive a 'message' from it (2010, p. 2). The amount of information contained by the signal is determined not only by the signal, but by the receiver as well. The following example will make these points more evident.

Suppose an unseen coin is flipped and you are told, as a hint, that it might be heads or tails. You most likely already knew that and are wondering if this is really a hint at all. This highlights two important criteria for evaluating a signal's informational content: First, a signal must be selected from a possible set of signals. How much information has been transmitted regarding an object depends on how the range of possible object states has been affected. The hint you received does not affect the range of possible outcomes from the coin flip, and thus holds no information. Second, the relative probability of the possible states under consideration plays an important role in evaluating the quantity of information transmitted. A coin is not a truly two-

dimensional object. There is a small possibility that it has landed on its side. The hint you received actually has some informational content, it is just small because the eliminated state is unlikely. Note that Shannon's definition of information does not restrict the types of objects and states that it describes. It may be something as quantitative as numerical data. Likewise, it may be something difficult to quantify numerically, such as the emotions of another. The key point is that there is a spectrum of possible properties and that the signal reduces their domain.

Computation invokes many concepts similar to information. In fact, as Piccinini and Scarantino point out, computation and information processing are often mistakenly held to be synonyms (2011, p. 3). We utilize Piccinini and Scarantino's definition of computation in general: "We use 'generic computation' to designate the processing of vehicles according to rules that are sensitive to *certain vehicle properties* and, specifically, to differences between different portions of the vehicles" (2011, p. 10; italics added). In the case that these vehicles are signals containing information, information processing *is* a form of computation as just defined. However, not all computation involves the processing of information. Informational content is not an intrinsic property of an object. It is relative to an observer and depends on how much the message reduces the observer's uncertainty (Gallistel & King, 2010, p. 7). Consider a computation that outputs 'cuidado' if the input is 'el horno esta encendido' and provides no output if the input is 'el horno esta apagado.' To an English-only-speaking observer, this computation does not process information—the objects have no meaning. But to a Spanish-speaker, this cautions them that the oven has been switched on. The computation performed is the same, regardless of the observer. Even if Spanish is forgotten, and the computation's objects cease to be meaningful to *anyone*, it is still the same computation. Thus, computation does not necessarily process information.

This definition of computation is clearly quite broad. It is so broad, in fact, that some philosophers believe that such an understanding of computation implies that everything performs computation, that is, 'pancomputationalism' (Piccinini & Scarantino, 2011, p. 5; cf. Chalmers, 2011; Copeland, 1996). It may be true that one could pick just about any physical phenomenon and find an arbitrary function that it computes (e.g., a rock; Chalmers, 1996). For "computation" to be a useful concept regarding research on cognition—such as memory—in the cognitive sciences, its scope must be appropriately pared down.

Recall that, based on the above definition, a computation is only sensitive to *certain* properties of objects, not necessarily all of them. A function that determines whether or not a neuron fires may only be sensitive to the firing/pre-firing properties of other neurons. Any additional physical variables are irrelevant to the purposes of the computation at hand, namely, modeling the dynamics of single-neuron activity. Pancomputationalism draws attention to the worry that "computation" may be a meaningless concept in research if it does not refer to some finite range of properties (or messages) that determine the results of the computations carried out by some system. In other words, its properties must have informational content that are relevant to the system.

With these conceptions of information and computation at hand, we can present a way to understand how they are present in the brain. Various brain processes can be usefully understood as computational, for example, the brain's ability to draw conclusions (Gallistel & King, 2010, p. 59). Consider the recognition of an image containing text (Figure 1). The optic nerve transmits visual stimuli to the brain, but it does not interpret the text's meaning. This is the role of a different portion of the brain. In this process there are signals (i.e., visual stimuli) that come from a set of possible messages (i.e., one image is distinguishable from another), which can be understood as processed in accordance to a set of rules that are sensitive to the signal's properties (i.e., the shape of the image is that of a word, and the word has meaning independent from the image). Here we have all the characteristics of computation being used to process information.

An important point to address is that of representation. In our discussion of information, we spoke of it as being encoded. In other words, it is represented within a certain syntactic structure. This encoding is what allows for reliable interpretation of the signal's contents. Modern computers contain a type of software called a "driver." Each driver instructs the computer how to interface with a certain type of peripheral device, such as a mouse or external hard drive. Despite the fact that both of these devices can communicate via a universal serial bus (USB) connection, the computer must use a very different set of rules when interfacing with a mouse than with a hard drive. Similarly, a computational description of the brain must refer to some syntactic structure when describing how the brain processes its signals. However, not any syntax will suffice. Both brains and computers are faced with a tremendous variety of possible objects to represent. A much simpler device than the brain is the TI-84 calculator. The largest number it can represent is approximately 10^{100}. If it was forced to have a unique character for each value, the number of unique characters would exceed the number of atoms in the known universe. The calculator avoids this conundrum by constructing its representations from a small number of symbols (i.e., 0/1 for binary, 0-9 for its decimal display) in a way that is sensitive to their relative positions. Similarly, the English language is represented through the use of twenty-six visual symbols (i.e., the alphabet) and forty-four audible symbols (i.e., phonemes) in a syntax that is sensitive to their relative positions in space and time respectively. All of these methods of representation are "compact," that is, the resources required to construct a representation grow logarithmically as the range of possible messages increases (Gallistel & King, 2010, p. 76). If even the humble TI-84 requires a robust syntax capable of compact representation, it follows that any

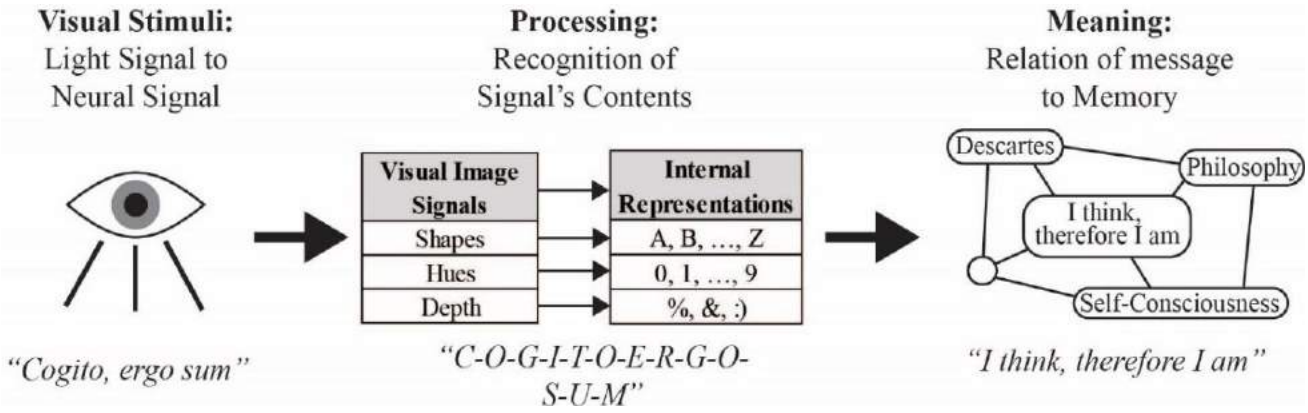


Figure 1: Reading text as the computational processing of information.

syntax that the brain may possess must also be capable of compactly representing the tremendous variety of messages it encounters (Gallistel & King, 2010, p. 82).

Like binary or decimal, the brain's syntax must be capable of constructing representations from a fairly small selection of basic elements. Otherwise, the incredible variety of sensory stimuli and the brain's practically unlimited creative capacity cannot feasibly be represented. These basic elements need not be numerical, and they need not be universal throughout the brain. Different portions of the brain perform different functions, and different syntaxes may be used. There will need to be commonalities between regions that facilitate their interaction, but there may be specialized computations (e.g., facial recognition, speech comprehension, etc.) that take place within regions. These computations might use a specialized syntax internally but use a shared syntax for computations that interact with other portions.

Having discussed computation, information, and representation, as well as what they mean in both the contexts of modern computers and the brain, we now move on to address how these ideas relate to memory and the role memory plays in the aforementioned contexts.

Memory in Computers and Brains

The property that distinguishes memory from other information-carrying signals in a brain or computer is its persistence through time. A fundamental requirement for a signal to be informative is that it is selected from a set of possible messages. There must be some way for this domain of possible messages to be established. When the layman enters an airliner cockpit, the array of dials and knobs are quite mysterious. To a trained pilot, each item denotes a meaningful piece of information. They know if an instrument's reading is alarming or typical. They know this because of their prior experience—information which was presented to them in the past and has persisted. In other words, their memory. Memory is the process that establishes the informational content of new signals.

Memory plays the same role in a modern computer. Without memory, a computer's only information regarding its past is that which is implicitly contained within its current state. As a result, whatever computation it performs must capture every relevant aspect of the computer's current state in order to determine the next state. Computers are often called upon to perform complex tasks that are combinations of a few basic functions (Gallistel & King, 2010, p. 109). Without a compact method for storing and preserving the results of past steps, any practical computation requires an absurdly large number of bits to define its state. Each instruction in a computer's program would have to shepherd hordes of bits. Suppose an instruction were as simple as providing a 1 (ON) or 0 (OFF) for each pixel in a display. To control every pixel of a typical 1080p resolution display, such an instruction would require *two million* bits. For comparison, modern central processing units (CPUs) typically use a humble sixty-four bits for their instructions. For a computational system to deal with these sorts of situations, a system is needed that has accessibility to the information contained in previous states as well as current—memory is needed (Gallistel & King, 2010, p. 131).

How does our definition of memory tackle this issue? In a modern computer, a CPU is the device that actually carries out most of the computations. In order to be able to perform computations quickly, it does not capture all of the information that it needs to perform all of its functionality within itself. If a CPU modifies an image, one cannot look at the CPU a few seconds later and determine how or what modification it performed. Instead, the CPU stores the image in a memory device to be retrieved later as needed. Later, when it needs to use the image in a computation, it calls upon the memory, which loads the image into the CPU. This transfer of bits constitutes a message to the CPU that informs it of the image's contents. The key point is that this is done without repeating the initial computation that resulted in the memory's message. Without the ability to call upon this *persistent information*, a modern computer would be as

cumbersome as a pilot that has to be retrained for dial-reading every time.

Despite the similarity in the role of memory in brains and modern computers, there is a much greater degree of plasticity in the execution of the brain's memory processes. In certain environments, the way the brain remembers events can be highly vulnerable to suggestion (Loftus & Palmer, 1974, p. 588). This supports the idea that memory plays a role in establishing the informational content of new signals and highlights the dynamic nature of the brain's syntax. In Loftus and Palmer's experiment, use of the verb 'hit' versus 'smash' served to alter the subjects' syntax by priming them to think in certain terms. This in turn changed the message they obtained when they referenced their memory regarding the presence of broken glass at the scene of an accident. In this way, our definition addresses the inconsistency of the brain's memory—the syntax being used to interpret the stored information is constantly changing. These changes need not result in an insensible message. Rather, they result in a new interpretation. This differs from modern computers, where even slight changes in syntax can cause total malfunctions.

Explicating the neurobiological processes underlying memory in brains is not necessary for our project. The scope of our definition is readily understood in terms of Marr's three levels of description of information-processing systems (Marr, 1982/2010, p. 24). The first and most abstract level is the *computational theory*, which establishes the general feature of the system being investigated, such as vision, language, or memory. Next, is the *representation and algorithm* level, which describes the procedures for achieving said system feature. The final level is *hardware implementation*, which is concerned with the physical substrate forming the representations and carrying out the algorithms. In terms of explanatory strategies, these levels can be investigated individually. From this perspective, our definition is appropriately understood as working in the first two levels. Although we aim for our definition to be applicable to real systems, we leave work of explicating its physical implementation to others.

For our purposes, we merely note that if our definition of memory is appropriate for the cognitive sciences, then it can guide research that successfully identifies brain regions and processes that facilitate the kind of persistent information seen in modern computers as sketched above (cf. Srimal & Curtis, 2008). If our definition is incorrect, then there will be no empirical evidence of such persistent information. This follows from one consequence of our definition, namely, that in both the brain and in modern computers, memory serves the role of preserving information and establishing the possible set of messages from which new signals arise. In the next section, we explain how our definition of memory provides a way to delineate boundaries around the system in which memory occurs.

The Boundaries of Memory

For a memory to be computationally useful in the system it belongs to, a consistent syntax must be utilized during the encoding process—that is, preservation and representation—of signals. For example, the alphanumeric symbol '6' must always denote the quantity six, and not three or four. This stipulation helps establish who or what a certain memory belongs to. The boundaries of the physical system that consistently realizes a computational system's syntax then defines the boundaries of its memory.

As discussed earlier, a computation is sensitive to some properties of an object but not necessarily all. Specifically, it is reactive to particular forms of content, that is, information. When this computation handles information, the rules of its sensitivity must match up to the syntax in which the information is represented. This feature allows one to determine what contributes to a computational system's memory or not. In order to be memory, a process must not only carry some physical state forward in time, but the state it preserves must be preserved in accordance with the syntactic structure of that to which the memory is said to belong.

Modern computers possess a set of memory addresses, much like a set of street addresses in a neighborhood, that they have access to. Proper usage of these addresses is part of the syntactic structure of the memory process. Searching for an address outside of this range causes the memory process to malfunction. The signal the computer finds with such an address might be encoded using a different syntax, or there might not even be a physical signal present. Either way, if the sought signal is not represented in accordance with the syntax of the memory it is trying to find, it is not a part of the computer's memory. The signals found may inform the computer, but the information will not be accurate.

In many natural and artificial systems, it seems obvious where to draw the boundaries of—at least some of—their memory systems, for example, a human's hippocampus and a laptop's hard drive. In such cases, the syntax used by the computations are consistently applied only within the physical brain and hard drive. Accordingly, such memories are realized within an individual body or casing. With that said, our definition of memory is not a priori confined to brains and hard drives. As long as such features as information preservation and consistent syntax are maintained, the boundaries of memory systems are potentially quite broad. Though we aim here to apply our definition of memory to more traditional work in the cognitive sciences, we leave open the possibility of applying it to cases such as distributed cognition (e.g., shared remembering by couples; Harris et al., 2014) and cultural transmission (e.g., Rowlands, 1993).

The possibility of distributed or extended memory systems should not be controversial. A removable USB storage stick is external memory for any modern computer with a USB port and appropriate software. The memory is only available

when the stick is plugged in, but a large number of different computational systems can all potentially access it. A written grocery list is external memory available to anyone who finds the list and can read the language it is written in (cf. Wagman & Chemero, 2014). The exact message the list presents will vary based on the individual's own internal memory—recall that the informational content of a signal is determined by the observer. Nevertheless, it is information carried forward in time. As long as the list's characters and words represent the intended message in a manner consistent with the reader's understanding of the language, the list can function as contributing to an external memory system.

We have attempted to show that our definition of memory is applicable to both narrow conceptions of cognition (e.g., isolated in brains), as well as more widespread notions (e.g., distributed cognition). Given that we discuss memory in terms of information processing, it is likely that the type of proponent of narrow conceptions who would readily accept our definition are those who think embodied, extended, and distributed cognition are still computational and representational in nature even if cognition is not isolated in brains (e.g., Barsalou, 2008; Hutchins, 1995; Wilson, 1994). On the other hand, it seems far less likely that anti-computational and anti-representationalists regarding cognition would accept our definition. We provide reasons why proponents of more "radical" conceptions of cognition could accept our definition by presenting experimental work involving affordances and memory.

Affordances are opportunities for behavior, and are based on the properties of the organism and environment (Gibson, 1979/1986). A doorway, for example, affords passing through for a human with narrow enough shoulders. Experimental work involving affordances stem from Gibson's ecological psychology (1979/1986). Contrary to representational approaches to perception, Gibson and his proponents argue that perception-action is not properly understood as centering on indirect representations. Visual perception, for example, is not a matter of an organism generating a mental image of the world, but instead is about an organism directly perceiving opportunities the world affords it.

Experimental work on affordances and memory have motivated conceptions of memory that do not appeal to computations or representations of the kind ecological psychologists and their proponents have resisted (e.g., Thomas & Riley, 2014; Vicente & Wang, 1998). Boschker, Bakker, and Michaels (2002), for example, conducted a set of experiments on the visual perception of climbing walls by experts and novices. When asked to recall information concerning the locations and orientations of holds on climbing walls, results suggested that experts can recall more information, clusters of information, and focus on functional aspects of walls (i.e., affordances); whereas novices did not recall clusters and focused on the structure of walls and not their functionals aspects. Boschker et al. argue that their findings show that differences in skill level correspond to differences in visual perception and memory. A central finding is that experts have memory that is better and of a more functional nature because they have more experience of perceived action possibilities than novices. In other words, their increased recall is tied to their increased perception of affordances. Note that this work does not appeal to computations or representations. Yet, our definition still applies: Experts have better task memory because the "information" relevant to action capabilities carries forward in time over the course of experience, and it does so in a manner that maintains its usefulness (i.e., affordance) to be "computed" (i.e., used) by the system (i.e., climber) for which the information belongs. The relationship between our definition and non-computational and anti-representational conceptions of cognition requires further fleshing out. However, we have attempted to demonstrate that the areas are not necessarily mutually exclusive. Having presented our definition of memory and discussed related issues, we now respond to several critiques.

Criticisms of a Computation-Based Definition

The appropriateness of utilizing our definition of memory in the cognitive sciences is contingent on the notion that it is explanatorily fruitful to describe the brain as performing computations. Computational approaches in the cognitive sciences are not without challenge. One source of opposition stems from forceful arguments claiming that phenomena investigated in the cognitive sciences are in no substantial way "computational," that is, "rule-governed manipulations of internal representations" (van Gelder, 1995). Therefore, our understanding of brains and cognition are set back by assuming they are like computers (Barrett, 2012). Another challenge centers on the claim that the prevalence of computationalism results from the prominent role of computers in modern society. Like other metaphors that were popular during their time, so too will the mind-as-computer metaphor pass (e.g., hydraulic pump, steam engine, etc.; Marshall, 1977). A third challenge is that many concepts underlying computational approaches have long and storied histories of imprecision. For example, many definitions of 'memory' now seem outdated in light of further technological advancement (Roediger, 1980). Addressing those challenges is far beyond the scope of the current work. Here, we respond to these criticisms in order to motivate the claim that complete rejection of a "computational" approach in the cognitive sciences is ultimately unwarranted.

First, unlike artifacts such as clay tablets or conveyor belts, computational theory is a set of formalized principles that are independent of any particular physical realization (Gallistel & King, 2010, p. 105). Computational theory becoming obsolete would be more akin to the obsolescence of calculus than that of the cellular phone. Computation is a field of mathematics, not a transient technology. While it is possible that computationally-based theories could be supplanted by

non-computational ones (e.g., Chemero, 2011; Edelman, 1993; Kelso, 2009; van Gelder, 1995) for explaining all forms of cognition, such a shift would likely occur due to conceptual, methodological, and theoretical advances in the cognitive sciences, and not due to a technology's life-cycle.

Second, appealing to computational theory to investigate memory in both brains and computers does not necessitate that both compute digitally or numerically, or that their objects are both numeric or symbolic. Computational theory, in the form we appeal to, is consistent with identifying both modern computers and brains as "computational," even if the objects being computed differ. This is because what matters more than the realizers of particular processes is the syntax. Because modern computers manipulate digital objects (i.e., binary '1s' and '0s'), they are readily able to handle syntax involving computations of large digits, such as multiplication and division. Human brains, on the other hand, may not explicitly manipulate digital objects, which could account for the difference in speed of calculation. Specifically, brains may manipulate analog objects, which may not be as fast at processing syntax involving calculations of discrete numerical values. This would serve to explain why brains and modern computers have a different set of strengths and weaknesses and are better suited for performing different kinds of computation that are computations nonetheless.

A third reason to consider computation in some form is to appeal to the primary motivation for the cognitive revolution, namely, the need to posit "internal" states to more fully account for some kinds of cognition, action, and perception (Gardner, 1985). To be more precise, those cognitive capacities that occur without externally observable processes, for example, predicting and learning. Cognitive systems can make accurate predictions following very complex causal chains. An electrical engineer can look at a wiring diagram and tell what will open a certain contact without interacting with the real circuit. A complete explanation of this capability implies some internal process for simulating events and evaluating them according to a syntax. Cognitive systems can learn to perform behaviors without actually doing them. If a hobbyist reads an article on how to solder a wire before attempting for first time, they will certainly do better than if they had tried with no prior study. From these examples we do not further claim that cognitive systems are not embodied, that learning via action is likely necessary during developmental stages, or that physical practice improves abilities. Yet, such examples motivate the need to appeal to internal processes to fully explain some cognitive phenomena. In some cases, the most parsimonious explanation for these capabilities is the presence of internal representations and rules for consistent execution. All of this suggests that appealing to some form of computation to explain certain cognitive phenomena is well-motivated.

If we are correct that at least some cognitive capabilities (e.g., memory) are appropriately explained via internal processes of some sort, then the nature of how those processes represent must be accounted for as well. Although a tremendous deal of research and effort has gone into mapping and studying brain activity, there is yet to be evidence of a discernible syntax. This could be seen as evidence against computation in the brain. However, this may be a case of a lack of evidence not being evidence of absence. Gallistel and King explain that the more efficient and robust an encoding scheme is (i.e., representation), the less it resembles its message (2010, p. 4). The sheer variety of stimuli the brain is presented with suggests that its syntax would be extraordinarily complex, far more so than binary (e.g., neurons as on-off switches). Additionally, recall that it is not necessary for these representations to be discrete *or* numerical in nature. They might not even be expressible in terms of language. The brain has been produced by natural selection, not a highly-organized team of computer scientists. As such, there is no reason to believe that any criteria other than effectiveness for survival and reproduction has played a role in its development. There has been no force in natural selection pushing the brain's representations to be legible to outside observers. With all this in mind, it is no surprise that the brain's syntax remains a mystery.

The definition of memory posited in this paper proposes a broader definition of computation and representation than are typically applied to the brain. It also does not propose computation as an explanation of brain structure and function. Rather, it appeals to computation to describe memory processes. The aim of this is to enable a discussion that escapes some of the limitations traditionally associated with computationalism. This paves the way for the utilization of computation as a descriptive tool without rejecting other accounts of cognition (e.g., dynamical). While some systems are better explained by either computational or dynamical models (van Gelder, 1995), others benefit from the use of multiple explanatory strategies (Favela & Chemero, 2019). Depending on the goals at hand, one model may be preferable to another, and it is possible that neither can give an all-encompassing account of the system. Here, we are chiefly concerned with defining what memory is, and not the computations or dynamics that explain how it is realized in systems.

Conclusion

We have presented and defended a definition of memory. We began with Gallistel and King's (2010) formulation of Shannon information and highlighted the feature of observer dependence. We then presented Piccinini and Scarantino's (2011) broad conception of computation as the processing of objects according to rules sensitive to certain properties of those objects. Computational systems are distinguished from one another based on what properties they're sensitive to. If the objects being processed are signals with informational content, then the computational system processes information. Both the brain and modern computers can be described as such systems. Casting memory in terms of

computation and information effectively describes memory as playing a role in establishing the meaning of new signals, that is, determining their informational content. The rules according to which these messages are interpreted are their syntax. In order to accommodate the wide variety of messages they represent, both the syntaxes used by brains and computers should be compact, that is, the resources required for representation should grow only with the logarithm of the number of possible signals. Having specific syntax for the purpose of carrying information forward in time allows for delineating boundaries around memory systems. We referred to ecological psychology's concept of affordances in order to illustrate that our definition is not necessarily incompatible with non-computational and anti-representational conceptions of cognition. We defended our use of computation as a tool for describing brain processes. Despite its challenges, computation's status as a set of formalized principles, as well as the ability of representations to serve as a succinct explanation of certain cognitive phenomena, make it well-suited for use as a descriptive tool. By limiting our use of computation to the *description* of memory, we remain nonpartisan as to the methods suitable for explaining its realization. As such, the following definition of memory is broadly applicable across the cognitive sciences: *Memory is a process that carries information forward in time, preserved in a fashion that maintains its usefulness as an object to be computed for the system to which the memory is said to belong*.

References

Barrett, L. (2012). Why behaviorism isn't Satanism. In J. Vonk & T. Shackelford (Eds.), *The Oxford handbook of comparative evolutionary psychology*. Oxford: Oxford University Press.

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology*, *59*, 617-645.

Boschker, M. S., Bakker, F. C., & Michaels, C. F. (2002). Memory for the functional characteristics of climbing walls: Perceiving affordances. *Journal of Motor Behavior*, *34*(1), 25-36.

Chalmers, D. J. (1996). Does a rock implement every finite-state automaton? *Synthese*, *108*, 309-333.

Chalmers, D. J. (2011). A computational foundation for the study of cognition. *Journal of Cognitive Science*, *12*, 323-357.

Chemero, A. (2011). *Radical embodied cognitive science*. Cambridge, MA: MIT press.

Copeland, B. J. (1996). What is computation? *Synthese*, *108*(3), 335-359.

Edelman, G. M. (1993). Neural Darwinism: Selection and reentrant signaling in higher brain function. *Neuron*, *10*(2), 115-125.

Favela, L. H., & Chemero, A. (2019). *Explanatory pluralism: A case study*. Manuscript submitted for publication.

Gallistel, C. R., & King, A. P. (2010). *Memory and the computational brain: Why cognitive science will transform neuroscience.* Chichester, UK: Wiley-Blackwell.

Gardner, H. (1985). *The mind's new science: A history of the cognitive revolution*. New York, NY: Basic Books, Inc.

Gibson, J. J. (1979/1986). *The ecological approach to visual perception*. Boston, MA: Houghton Mifflin.

Harris, C. B., Barnier, A. J., Sutton, J., & Keil, P. G. (2014). Couples as socially distributed cognitive systems: Remembering in everyday social and material contexts. *Memory Studies*, *7*, 285-297.

Hutchins, E. (1995). *Cognition in the wild*. Cambridge, MA: MIT Press.

Kelso, J. A. S. (2009). Coordination dynamics. In R. A. Meyers (Ed.), *Encyclopedia of complexity and systems sciences* (pp. 1537-1564). Berlin: Springer-Verlag.

Loftus, E. F., &, Palmer, J.C. (1974). Reconstruction of automobile destruction: An example of the interaction between language and memory. *Journal of Verbal Learning and Verbal Behavior*, *13*(5), 585-589.

Marr, D. (1982/2010). *Vision: A computational investigation into the human representation and processing of visual information.* Cambridge, MA: The MIT Press.

Marshall, J. C. (1977). Minds, machines and metaphors. *Social Studies of Science*, *7*, 475-488.

Piccinini, G., & Scarantino, A. (2011). Information processing, computation, and cognition. *Journal of Biological Physics*, *37*(1), 1-38.

Roediger, H. L. (1980). Memory metaphors in cognitive psychology. *Memory & Cognition*, *8*(3), 231-246.

Rowlands, M. (1993). The role of memory in the transmission of culture. *World Archaeology*, *25*, 141-151.

Srimal, R., & Curtis, C. E. (2008). Persistent neural activity during the maintenance of spatial position in working memory. *NeuroImage*, *39*, 455-468.

Thomas, B. J., & Riley, M. A. (2014). Remembered affordances reflect the fundamentally action-relevant, context-specific nature of visual perception. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(6), 23612371.

van Gelder, T. (1995). What might cognition be if not computation? *Journal of Philosophy*, *91*, 345-381.

Vicente, K. J., & Wang, J. H. (1998). An ecological theory of expertise effects in memory recall. *Psychological Review*, *105*(1), 33-57.

Wagman, J. B., & Chemero, A. (2014). The end of the debate over extended cognition. In T. Solymosi & J. R. Shook (Eds.), *Neuroscience, neurophilosophy and pragmatism: Brains at work in the world*. New York, NY: Palgrave Macmillan.

Wilson, R. A. (1994). Wide computationalism. *Mind*, *103*, 351-372.

Asking goal-oriented questions and learning from answers

Anselm Rothe[1], **Brenden M. Lake**[1,2], **and Todd M. Gureckis**[1]
[1]Department of Psychology, [2]Center for Data Science, New York University

Abstract

The study of question asking in humans and machines has gained attention in recent years. A key aspect of question asking is the ability to select good (informative) questions from a provided set. Machines—in particular neural networks—generally struggle with two important aspects of question asking, namely to learn from the answer to their selected question and to flexibly adjust their questioning to new goals. In the present paper, we show that people are sensitive to both of these aspects and describe a unified Bayesian account of question asking that is capable of similar ingenuity. In the first experiment, we predict people's judgments when adjusting their question-asking towards a particular goal. In the second experiment, we predict people's judgments when deciding what follow-up question to ask. An alternative model based on superficial features, such as the existence of certain key words in the questions, was not able to capture these judgments to a reasonable degree.

Keywords: Bayesian modeling; active learning; information search; question asking

Introduction

The ability to ask questions is a core quality of human cognition. By asking questions, we can actively seek out information that helps us learn about the world and achieve our goals. Skilled question asking involves the ability to adjust questions towards a particular goal as well as a sensitivity to the context, including what was previously asked.

In contrast, machines have difficulty capturing these aspects of human inquiry. Recent work with neural networks has made progress on generating sensible questions about images, such as "What caused this accident?" for an image displaying a crashed motorbike lying on the street (Mostafazadeh et al., 2016; Jain & Schwing, 2017), or about passages of text (Du, Shao, & Cardie, 2017). Such questions can initiate a conversation between human and computer, however these networks are not able to make sense of any answer they might get to their question. As an intermediate solution, neural networks have been trained to predict the answer to their own questions (Johnson et al., 2017). Another ambitious approach has been to train neural networks end-to-end on entire sequences of questions and answers (Lee, Heo, & Zhang, 2018; Strub et al., 2017). However, the networks still learn a fixed question asking strategy and cannot adapt to new goals that were not included in the training regime.

Unlike neural network approaches, people can flexibly adapt their questions based on their goals and the answers they have received. Previous work has looked at how people ask questions based on specific goals (e.g., Graesser, Langston, & Bagget, 1993), or ask follow-up questions (e.g., Nelson, Divjak, Gudmundsdottir, Martignon, & Meder, 2014; Ruggeri, Lombrozo, Griffiths, & Xu, 2016), but little modeling work has been done to test these aspects directly in naturalistic tasks. Here, we study an intuitive question asking task amenable to formal modeling. By systematically manipulating core components of question asking such as goals and previously asked questions, we can compare people's behavior to an ideal observer in a more naturalistic question asking environment. For this purpose, we extend the computational framework by Rothe, Lake, and Gureckis (2018) to handle these facets of flexible question asking.

In the next section, we will introduce the question asking environment, followed by the computational framework and its extensions. We then report two experiments, in which we test people's ability to identify question quality under changing goals (Experiment 1) and after being provided with answers to previous questions (Experiment 2). Finally, alternative models are discussed.

Battleship game environment

We adopt the Battleship task used by Rothe et al. because it enables intuitive question asking for people while still being amenable to formal modeling. In the Battleship task, participants try to discover geometric shapes (i.e., battleships) on a grid (i.e., game board). These ships have varying shapes, colors, and locations (Figure 1). In our setting, there were always exactly three ships on a 6x6 board and each ship got a unique color from the set {blue, red, purple}. Each ship is a rectangle with a width of 1 and a length sampled from the set {2, 3, 4} and its orientation is sampled from the set {horizontal, vertical}. Each ship is randomly placed on the grid, ensuring they do not overlap.

In our experiments, participants face a partly revealed game board, together with a set of natural-language questions that could reveal more information about the board. Participants rank order these questions by quality taking either a particular goal or an already-answered question into account (Figure 2).

Modeling

We develop a Bayesian ideal-observer model of the task, as used in prior work, and discuss extensions to handle goals and previously answered questions.

Bayesian-ideal observer model

What does the hidden game board look like? The player begins with maximal uncertainty about the game board, modeled as a uniform prior belief distribution $p(h)$ over all possible game boards. Then, the player updates this prior via Bayes rule based on the information d presented by the partly revealed game board,

$$p(h|d) = \frac{p(d|h)p(h)}{\sum_{h' \in H} p(d|h')p(h')}, \quad (1)$$

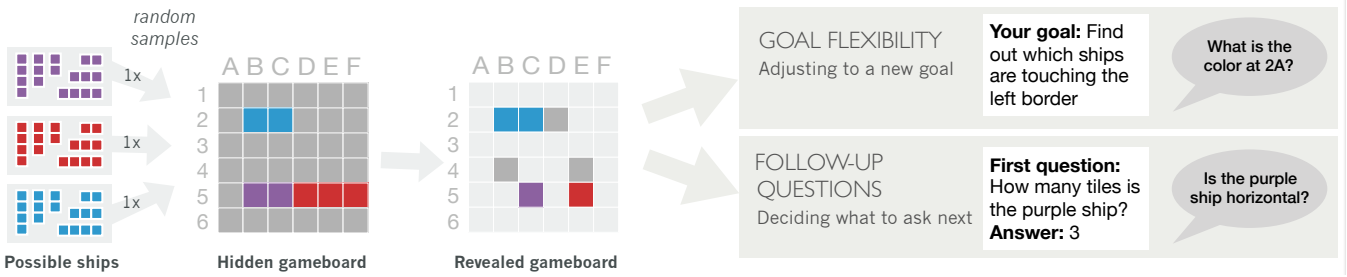


Figure 1: In the Battleship task, three ships are randomly positioned on a game board. The color at each location indicates a ship (blue, red, or purple) or water (dark gray). Participants view a game board that only shows some tiles revealed and many tiles yet unknown (light gray). They can then ask questions to obtain more information. Can people adjust their questions towards specific goals (Experiment 1)? For example for the game board shown above, when the goal is to find out which ships are touching the left border, a question targeting the color of tile 2A would be more useful than the question whether the red ship is horizontal. Do people adjust their questions based on already-answered questions (Experiment 2)? For example, after learning that the red ship is three tiles long, participants might be inclined to keep asking questions about the red ship before addressing the other ships.

where H is the hypothesis space of 1.6 million game boards and $p(d|h)$ the likelihood function, which is 1 if d is consistent with board h, and otherwise 0. The player can now ask a question x to learn more. The answer to the question is assumed to come from an oracle that knows the hidden game board and answers truthfully. We use d again as the label for the answer since it plays the same role as the partly revealed game board before. The likelihood function is again 0 if answer d is inconsistent with board h. Otherwise it is $\frac{1}{n}$, to account for cases where there is more than one valid answer, from which the oracle then chooses uniformly. For instance, for the question "What is the location of one purple tile?" the oracle would indicate the location of one of the n purple tiles on the true game board. Usually though, in our setting there is only one valid answer, $n = 1$ (e.g., yes or no). We now generalize to include the history of previous questions X and their answers D, resulting in

$$p(h|d,D;x,X) = \frac{p(d|h;x)p(h|D;X)}{\sum_{h' \in H} p(d|h';x)p(h'|D;X)}, \qquad (2)$$

where the semi-colon notation indicates that x and X are parameters rather than random variables (for the first question, X and D are empty).

In preparation for the next section, we can compute the posterior predictive probability that d will be the answer to question x via

$$p(d|D;x,X) = \sum_{h \in H} p(d|h;x)p(h|D;X). \qquad (3)$$

Expected Information Gain (EIG)

The player's uncertainty about the hidden game board is measured by the Shannon entropy of the belief distribution (Shannon, 1948; see Crupi, Nelson, Meder, Cevolani, & Tentori, 2018, for a discussion of alternative measures). The Information Gain (IG) of a question x is then defined as the amount by which this uncertainty is reduced when receiving answer d. Since the player does not know the answer at the time of asking, we compute an expected value:

$$\begin{aligned} EIG(x) &= \sum_{d \in A_x} p(d|D;x,X)\Big[I[p(h|D;X)] - I[p(h|d,D;x,X)]\Big] \\ &= \mathbb{E}_{d \in A_x}\Big[I[p(h|D;X)] - I[p(h|d,D;x,X)]\Big], \end{aligned}$$

where $I[\cdot]$ is the Shannon entropy, and A_x are the possible answers to question x. EIG has been used to describe a range of information sampling behavior (see Coenen, Nelson, & Gureckis, 2018, for an overview).

EIG for goal-directed questions. So far, EIG aims to reduce all uncertainty in $p(h)$. In order to only reduce the uncertainty that is relevant for a particular goal, we introduce the goal state space g. To illustrate with an example in the Battleship task, the goal "Find out which ships are touching" has as goal states g the various possibilities of ships that could be touching (i.e., *none*, *blue*|*red*, *blue*|*purple*, etc). Furthermore, g is defined as a projection of the hypothesis space h. Table 1 provides a minimal example of such goal projection. We can now measure the quality of a question x with respect to goal g via

$$EIG_{goal}(x,g) = \mathbb{E}_{d \in A_x}\Big[I[p(g|D;X)] - I[p(g|d,D;x,X)]\Big]. \qquad (4)$$

In detail, we compute the belief distribution over the goal states by marginalizing over h (here shown for the posterior, the equivalent is to be done for the prior)

$$p(g|d,D;x,X) = \sum_h p(g|h)p(h|d,D;x,X),$$

where $p(g|h)$ is 1 if h is goal-projected onto g, and 0 otherwise. More simply stated, for each goal state, we sum the belief values from the hypotheses that are projected onto the goal state. The EIG with respect to this goal is then the expected uncertainty reduction in the belief distribution over

Table 1: Simple example of a goal projection. Four hypotheses in h are projected onto two goal states in g. The projection results in a prior belief $p(g)$ of 0.2 for goal state 1, and 0.8 for goal state 2.

p(h)	h	g
0	1	1
0.2	2	1
0.4	3	2
0.4	4	2

these states. For convenience, we will subsume EIG_{goal} under the label EIG outside of this section.

EIG for follow-up questions. With the setup explained so far, the ability to take an already answered question into account comes out-of-the-box for the EIG model. Observed data D can be the visual information provided by the partly revealed board, as well as the verbal information from the answers to previous questions. The resulting knowledge is encoded in the posterior belief distribution, $p(h|d,D;x,X)$.

Experiment 1 – Asking goal-directed questions

In general, people ask different questions when they have different goals. When their goal changes, people should be able to flexibly adapt the questions they want to ask. In this experiment, we investigate whether people's evaluations of question usefulness are sensitive to specific goals.

Participants

Forty participants recruited on Amazon Mechanical Turk, with restriction to the United States pool, were paid a base of \$2 with a performance based bonus of up to \$4.86.

Method

In order to lead participants into a situation in which they wanted to ask a question, we took a number of steps to make them familiar with the Battleship task. First, participants went through a tutorial that presented the game board and the possible colors, sizes, orientations, and positions of the ships. This key information was shown on the side over the whole experiment and additionally checked in a comprehension quiz after the tutorial. Next, participants went through a warm-up phase, in which they began with a completely unidentified game board and clicked on the grid tiles to turn over their color, revealing more of the game board step by step.

Then, participants started the main phase, which consisted of 18 randomized trials. The schema of a trial is shown in Figure 2. Participants first viewed a partly-revealed game board and received a goal. They then ranked six natural-language questions "such that good questions are at the top and not so good questions are at the bottom" by dragging and dropping each question into a sortable list. To make sure that people paid attention to the questions, we displayed them one by one in a random order and people had to press the correct button

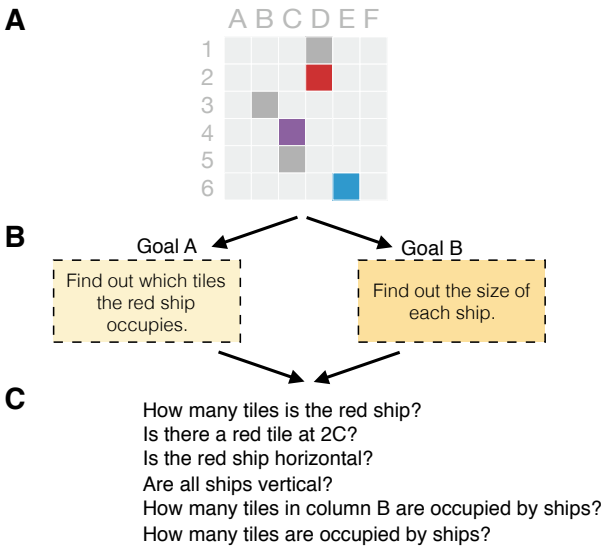


Figure 2: Experimental design of Experiment 1. In a given trial, (A) participants view a partly revealed game board. (B) Participants then receive one of two goals, randomly assigned. (C) Participants rank six questions by quality with respect to the goal. In Experiment 2, Goal A and Goal B are each replaced with an already-answered question.

that described the answer type of the question (either a color, a coordinate on the grid, a number, or yes/no). For each correct response, a bonus of \$0.045 was awarded. Allocating bonuses in this way, rather than basing it on their ranking of questions, discouraged participants from attempting to infer a researcher-preferred ranking of questions.

All participants viewed the same 18 partly-revealed game boards and corresponding question sets. But, as Figure 2 illustrates, the goal they received was randomly chosen from a predefined set of two goals for each context. The 18 game boards and the corresponding questions were the same as in Rothe et al., to ensure maximal comparability across studies (see Rothe et al., 2018, for details on the design of the boards and question sets).

The goals were designed as follows. We created a list of goals that seemed interesting but intuitive, such as "Find out which ships are touching the top border", "Find out which tiles the red ship occupies" which would allow people to ignore the blue and purple ship, or "Find out the size of each ship" which would allow them to ignore the orientation and location of the ships.

For each context, we determined via computer simulation a pair of opposing goals, such that the resulting EIG model scores of the questions were maximally different when evaluated against each goal (as measured by correlation). Examples of these opposing goals are shown as titles of the panels in Figure 3C. The average correlation between model scores within the goal pairs was $r = -0.28$.

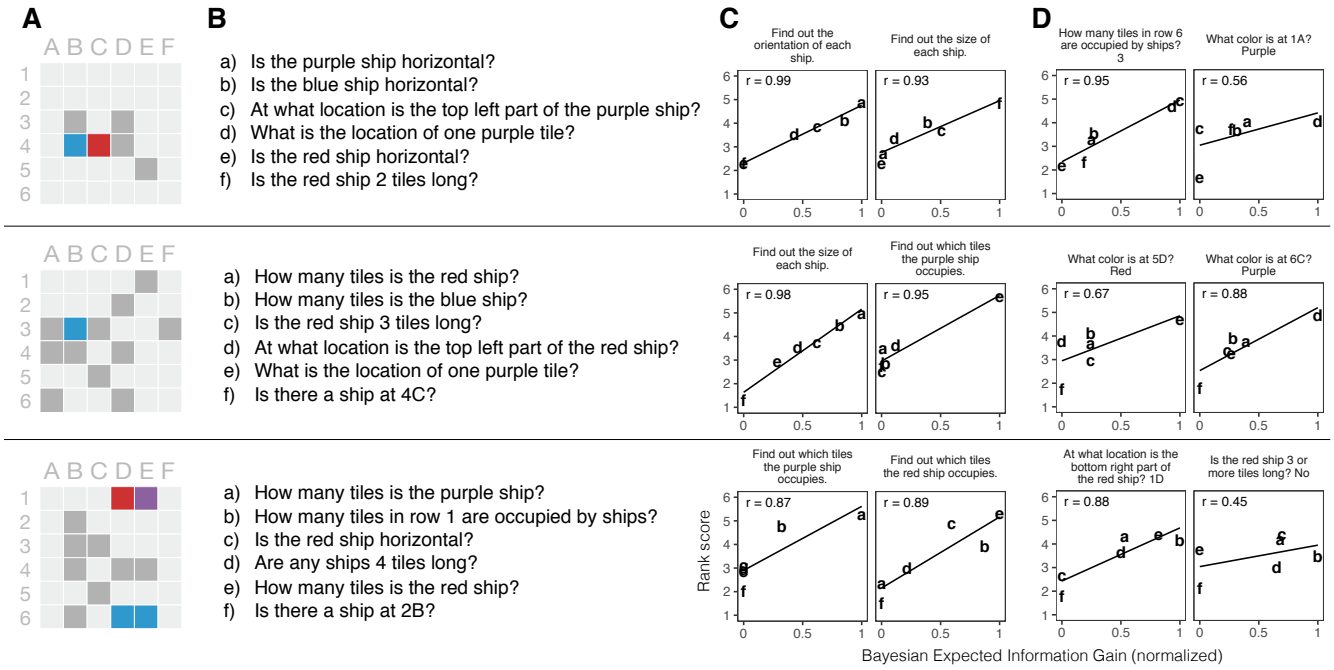


Figure 3: Contexts and questions together with human rankings and model predictions. Three selected trials exemplifying (A) the partly revealed game board, (B) the sets of six questions that were ranked by participants, (C) correlations of human rankings (y-axis; higher is better) and model scores (x-axis) for these questions in Experiment 1, and (D) Experiment 2. The letters a-f in the scatterplots correspond to letters marking the questions to the left. Error bars for $\pm 1SE$ are not plotted as they are only as large as the letters. Model scores are normalized to a maximum of 1.

Results

People's preference for questions were highly sensitive to the specific goals they had. Figure 3C shows how people's rankings of the same questions varied widely depending on the different goals. For example, in the first row in Figure 3, the question "Is the purple ship horizontal?" (marked with the letter **a**) was ranked best for the goal "Find out the orientation of each ship" but very low for the goal "Find out the size of each ship." The different rankings of this and other questions were well captured by the EIG model, which took the respective goal that participants had into account. Figure 3C shows several examples with strong correlations between EIG and human rank scores. Across all contexts, the average Pearson correlation between model scores and human rankings was $r = .84$. In contrast, when we let the EIG model hypothetically take the respective *opposite* goal into account, correlations dropped to an average $r = -.16$.

We also computed an "ignorant" model that ignored the specific goal and instead tried to obtain as much information as possible for the complete game board. A participant whose ratings are well captured by this model is probably ignoring the specific goal and instead plays the original Battleship game. The average correlation for this model was $r = .42$.

Instead of comparing correlation coefficients, we conducted a more sensitive model comparison that takes guessing behavior into account. Model scores were transformed into choice probabilities via the softmax function

$$p(x) = \frac{e^{-\beta M(x)}}{\sum_x e^{-\beta M(x)}},$$

where $M(x)$ is the model score (e.g., EIG(x)) and β is the free temperature parameter, capturing more guessing behavior as $\beta \to 0$. For each model, β was fit per participant to the rankings, and the resulting log-likelihood of the top ranked question computed.

In direct comparison, EIG had higher log-likelihood than EIGopposite, which took the opposite goal into account, for 38 out of 40 participants (95%). EIG also had a higher log-likelihood than EIGignore, which ignored the goal, for 35 out of 40 participants (88%).

We can conclude from this that people are very sensitive towards the specific goals when making question evaluations in our task, and that their evaluations are well predicted by our goal-oriented Bayesian ideal-observer EIG model with zero free parameters.

Experiment 2 – Asking follow-up questions

We test the EIG model further with the very natural task of deciding what to ask next, after a question was already answered.

Participants

A separate set of forty participants recruited on Amazon Mechanical Turk, with restriction to the United States pool, were paid a base of $2 with a performance based bonus of up to $4.86.

Method

The materials and procedure were identical to Experiment 1, except that instead of a goal, a question and its answer were displayed. That is, for each context, there was a predefined set of two already-answered questions from which one was randomly chosen for each participant (cf. Figure 2).

As for the pairs of goals in Experiment 1, we identified via computer simulation pairs of already-answered questions that were as anti-correlated as possible. The following procedure was repeated for each game board context. From a list of 136 unique questions we simulated all possible answers to each question. For example, the question "How many tiles in row 6 are occupied by ships?" (Figure 3D, first panel) has the possible answers {0, 1, ..., 6}. Then, for each question-and-answer combination, we computed what the resulting EIG scores would be for the six questions (Figure 3B) that now served as follow-up question candidates. As before, we created pairs of already-answered questions that had the most different model scores for the follow-up question candidates, as measured by the lowest correlation. The average correlation between model scores within the pairs was $r = 0.02$.

Results

People's rankings of the follow-up questions are generally sensitive to the information provided by the already-answered question. Figure 3D shows the correlations between the EIG model and human rankings. Overall, the average Pearson correlation was $r = 0.71$. When computing EIG by hypothetically taking the opposed already-answered question into account, the average correlation dropped to $r = 0.29$. When computing EIG that ignores the information from the answered question, the average correlation was $r = 0.60$. This suggests that people evaluated the usefulness of the follow-up questions by integrating the verbal information provided by the first question and its answer.

Again, instead of comparing correlation coefficients, we modeled people individually via a softmax function. EIG had a higher log-likelihood than EIGopposite for 28 out of 40 participants (70%). Using a softmax function again and modeling people individually, EIG had a higher log-likelihood than EIGopposite for 28 out of 40 participants (70%). Surprisingly, EIG had a higher log-likelihood than EIGignore for only 21 out of 40 participants (52%). The latter comparison suggests that a fair number of participants were not sensitive to the information from the answered question.

To inspect this result more carefully, we set up a hybrid model that balanced between EIG and EIGignore with a free parameter, $\theta EIG(x) + (1-\theta)EIG_{ignore}$. The balancing parameter θ was fit simultaneously with the softmax guessing parameter β for each person. The resulting distribution suggests

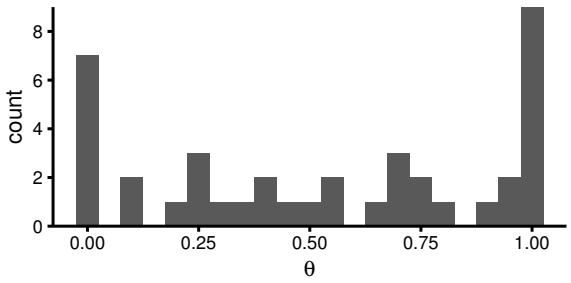


Figure 4: The θ values among participants. A participant's θ could be taken as an indicator for how much she considered the information from the answered question in Experiment 2. The binwidth is .05.

that nine participants (23%) took the information that the answered question provided accurately into account ($\theta > .95$), seven (18%) completely ignored the information ($\theta < .05$), and 24 (60%) exhibited a mixed strategy (Figure 4). Under this analysis it is still possible that the in-between participants used a different strategy, neither captured by EIG nor EIGignore. Yet, the log-likelihoods for these participants were as good as for the others suggesting that they did indeed use a mix of EIG and EIGignore. Thus, a participant's θ could be interpreted as the amount to which she considered the verbal information from the answered question.

Word-based model. We further considered an alternative model that takes a word-based approach. One strategy people might exhibit is to keep focusing on getting information about the ship they already have some details on. For instance, if the answered question provides information that the red ship is horizontal, they might prefer to learn about the size of the red ship before moving on to the next ship. Thus the word-based model looks for signal words that match the already-answered question and the follow-up question. Formally, the Color feature compares the color words {*blue, red, purple, water*} in the answered question with those in the follow-up question candidates. If there exist color words in both questions and they are the same, then the Color feature assigns a 1 to the follow-up question, else a 0. To illustrate, consider the right panel in the third row in Figure 3D. The already-answered question "Is the red ship 3 or more tiles long?" mentions the red ship. Therefore, the model would prefer the follow-up questions **c** and **e** because they also mention the red ship. Indeed, **c** and **e** were both ranked somewhat higher than predicted by the EIG model. Overall, questions that were ranked as best by people had more often matching color words (23%) with the already-answered question than the lower ranked questions (12-21%).

Another strategy that people might employ is to prefer a question of the same type as of the one that was already answered. The Type feature categorizes questions into mutually exclusive groups of *ship orientation, ship size, adjacency, region, location,* and *demonstration* questions. This classifica-

tion follows the one described in Rothe et al. (2018). The Type feature simply assigns a 1 if both questions are classified into the same type, else 0. Illustrating for the same case as above, the feature classifies the answered-question into the *ship size* type, to which also follow-up questions **a**, **d**, and **e** belong, which therefore get a higher score. Overall, questions that people ranked worst were more often of the same question type as the already-answered question (30%) than the higher ranked questions (21-23%).

The word-based model combines both features in a linear combination. We fitted a linear regression using both features as predictors and participants' average rank scores as criterion. However, only little variance in people's rankings could be explained this way, $R^2 = 0.05$.

Discussion

We tested people's preference for questions in two crucial situations: when asking goal-directed questions and when asking follow-up questions. In Experiment 1, we manipulated the goal that people had, while keeping everything else constant. People's rankings of question quality dramatically shifted based on the goal they were assigned. The rankings were well predicted by our Bayesian ideal-observer model of Expected Information Gain (EIG) with zero free parameters. In Experiment 2, we manipulated what already-answered question people received, while keeping everything else constant. Again, people's rankings shifted strongly based on the answered question. However, the picture was less clear than in the first experiment. While generally people's rankings were well predicted by the EIG model, detailed analysis suggested that people varied in the amount to which they integrated the information provided by the already-answered question. An alternative model that approximated question usefulness based on superficial features could not explain human rankings.

So far, neural network approaches to question asking generally struggle with the flexibility that is necessary to take previous answers and goals into account. To reach competitive performance in simple tasks they already need training on large data sets with tens of thousands of questions. In order to add sensitivity towards specific answers and goals would require additional training likely in orders of magnitude more.

One of the strengths of the Bayesian approach is the seamless integration of visual and verbal information. The visual information from the partly revealed game board and the verbal information from the answered question were both integrated into a unified posterior. In our current analysis we only considered varying degrees to which people considered the verbal info from the answered question. It is also possible that people did not perfectly take the visual information from the partly revealed board into account. In future work, we will further explore people's integration of high-level information.

We extended the computational framework to two aspects of question asking—more needs to be done. In our setting, we assumed a reliable, all-knowing oracle that is providing the answers. However, the relationship between question, ground truth, and generated answer is not as deterministic in many real-world settings. For example, in social settings, people need to take into account the knowledge state and goals of their communication partner. This aspect has been elegantly modeled in the Rational Speech Act framework, where a questioner has an internal model of the answered that she simulates recursively before deciding what to ask (Hawkins & Goodman, 2017). We see our approach as complementary to this RSA model. Future work should aim to integrate both.

Acknowledgments

This research was supported by NSF grant BCS-1255538, the John Templeton Foundation "Varieties of Understanding' project, a John S. McDonnell Foundation Scholar Award to TMG, and the Moore-Sloan Data Science Environment at NYU.

References

Coenen, A., Nelson, J. D., & Gureckis, T. M. (2018). Asking the right questions about human inquiry: Nine open challenges. *Psychonomic Bulletin & Review*, 1–41.

Crupi, V., Nelson, J. D., Meder, B., Cevolani, G., & Tentori, K. (2018). Generalized Information Theory Meets Human Cognition: Introducing a Unified Framework to Model Uncertainty and Information Search. *Cognitive Science*, *42*(5), 1410–1456.

Du, X., Shao, J., & Cardie, C. (2017). Learning to Ask: Neural Question Generation for Reading Comprehension. *arXiv:1705.00106v1*.

Graesser, A. C., Langston, M. C., & Bagget, W. B. (1993). Exploring information about concepts by asking questions. *The Psychology of Learning and Motivation*, *29*, 411–436.

Hawkins, R., & Goodman, N. (2017). Questions and answers in dialogue. *PsyArXiv: j2cp6*.

Jain, U., & Schwing, A. (2017). Creativity: Generating Diverse Questions using Variational Autoencoders. *arXiv:1704.03493v1*.

Johnson, J., Hariharan, B., van der Maaten, L., Hoffman, J., Fei-Fei, L., Lawrence Zitnick, C., & Girshick, R. (2017). Inferring and executing programs for visual reasoning. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 2989–2998).

Lee, S., Heo, Y., & Zhang, B. (2018). Answerer in questioner's mind: Information theoretic approach to goal-oriented visual dialog. In *Advances in Neural Information Processing Systems 31* (pp. 2584–2594).

Mostafazadeh, N., Misra, I., Devlin, J., Zitnick, L., Mitchell, M., He, X., & Vanderwende, L. (2016). Generating natural questions about an image. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics.*

Nelson, J. D., Divjak, B., Gudmundsdottir, G., Martignon, L. F., & Meder, B. (2014). Children's sequential information search is sensitive to environmental probabilities. *Cognition*, *130*(1), 74–80.

Rothe, A., Lake, B. M., & Gureckis, T. (2018). Do people ask good questions? *Computational Brain & Behavior*, *1*(1), 69–89.

Ruggeri, A., Lombrozo, T., Griffiths, T. L., & Xu, F. (2016). Sources of Developmental Change in the Efficiency of Information Search. *Developmental Psychology*, *52*(12), 2159–2173.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*, 379–423.

Strub, F., de Vries, H., Mary, J., Piot, B., Courville, A., & Pietquin, O. (2017). End-to-end optimization of goal-driven and visually grounded dialogue systems. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence* (pp. 2765–2771).

Elvis Has Left the Building: Correlational but Not Causal Relationship between Music Skill and Cognitive and Academic Ability

Giovanni Sala (sala.giovanni475@gmail.com)
Graduate School of Human Sciences, Yamada-Oka 1-2 Suita,
Osaka University, 565-0871, Japan

Fernand Gobet (fgobet@liv.ac.uk)
Department of Psychological Sciences, Bedford Street South,
University of Liverpool, L69 7ZA, United Kingdom

Abstract

Music training is commonly thought to have a positive impact on children's cognitive skills and academic achievement. This belief relies on the idea that engaging in an intellectually demanding activity helps to foster overall cognitive function. We here present a meta-analysis of music-intervention studies in children (N = 3,780, k = 204, m = 43). Consistent with the substantial findings in the field of cognitive training, the overall effect size was small ($\bar{g}$ = 0.117, p < .001). Moreover, when active controls were implemented, the effect was practically null ($\bar{g}$ = 0.032, p = .477) and highly homogeneous (ω^2 = 0.000 and τ^2 = 0.000). Finally, we observe that several independent research groups have concluded, via different methodologies, that music skills acquired by training do not generalize to non-music skills. Thorndike and Woodworth's (1901) common elements theory finds thus further support.

Keywords: music; cognitive training; meta-analysis; transfer of skills.

Introduction

Many parents encourage their children to play a musical instrument. Their hopes sometimes go beyond proficiency in playing music: they enroll their children in violin or piano lessons not only to nurture their musical talent but also because they assume that music training will help their children to get better at school or even become more intelligent.

The idea that learning how to play an instrument improves one's cognitive abilities and academic achievement is popular. Music ability is often associated with talent and superior cognitive skills. Blogs and newspapers often report enthusiastically on the benefits of music for the intellect (e.g., Jaušovec & Pahor, 2017). Even the popular TV series *The Simpsons* has echoed this common belief by defining musical instruments as "the way to encourage a gifted child."

The conviction that music training enhances cognitive ability and academic achievement relies on the assumption that music skills acquired by training can generalize to non-music domains. However, what does the scientific research in the field tell us about music training? Is this assumption correct?

Why Should Music Training Enhance Cognition?

As just mentioned, music training has been claimed to improve a broad range of cognitive and academic skills. However, how is music training supposed to provide such diverse benefits? The standard hypothesis relies on the idea that it is possible to train domain-general cognitive abilities by engaging in intellectually demanding activities. Learning how to play a musical instrument engages executive functions such as cognitive control and working memory (Bialystok & Depape, 2009). In addition, music training requires focused attention and learning complex visual patterns. Schellenberg (2006) has thus proposed that the most likely explanation for the presumed broad set of benefits provided by music training is that it enhances individuals' overall cognitive function and general intelligence. These cognitive skills are major predictors of academic achievement (e.g., Deary et al., 2007), and it might be the case that some domain-specific abilities acquired by music training generalize to other non-music skills.

One further theoretical foundation for the hypothesis according to which music training exerts a positive influence on overall cognitive ability is neural plasticity. Neural plasticity is the ability of the neural system to modify and adapt under the pressure of the environment (Strobach & Karbach, 2016). In turn, the changes in the neural system are supposed to account for improvements in cognitive tests. In fact, musicians do exhibit specific anatomical and functional neural patterns. An increased density of gray matter in musicians has been observed in areas involved in cognitive skills such as auditory localization (right Heschl's gyrus; Bermudez et al., 2009) and language production (Broca's area; Sluming et al., 2002).

With regard to functional differences, expert musicians seem to show, for example, enhanced bilateral activation of the Rolandic operculum (for a review, see Neumann, Lotze, & Eickhoff, 2016). This activation probably reflects superior ability in the processing of auditory information (Koelsch et al., 2006). While there is empirical support for the hypothesis that music training induces significant anatomical and functional changes in the brain, which sometimes lead to unexpected behavioral skill differences (e.g., superior memory for randomized music-related material; Sala &

Gobet, 2017a), the evidence that these neural changes lead to increased cognitive function is much weaker, as discussed in the following sections.

Correlational Evidence

There is strong empirical evidence for a link between superior cognitive ability and musical skill. In a study by Ruthsatz et al. (2008), a group of professional musicians outperformed a group of novices in a standardized measure of fluid intelligence (Raven's Progressive Matrices). Lee, Lu, and Ko (2007) found a correlation between music skill and working memory. In the same vein, Saarikivi et al. (2016) found that neural sound discrimination predicted performance on an inhibition task and a set-shifting task in a sample of children and young adolescents. Finally, Schellenberg (2006) reported positive, yet moderate, correlations between engagement in musical activities and IQ in a group of children and undergraduates. Critically, this positive relationship remained even after controlling for parental income and education.

Music ability correlates with academic skills as well. Anvari et al. (2002) found that music perception skills correlated with reading abilities in preschool children. Similarly, Forgeard et al. (2008) reported that music discrimination skill correlated with phonological processing ability in a group of dyslexic and typically-developing children. In line with these studies, Wetter, Koerner, and Schwaninger (2009) reported a positive relationship between engagement in musical activities and overall academic attainment.

Experimental Evidence and Present Study

As just seen, music skill is positively associated with measures of fluid intelligence, memory, and academic achievement. However, while music skill and cognitive ability are correlated, to date there is no clear evidence of a causal relationship from engagement in music training to superior cognitive function.

A meta-analysis of all the available studies (Sala & Gobet, 2017b) has expressed pessimism about the actual possibility of music-training interventions to enhance children's cognitive and academic skills. All the studies included in this meta-analysis are true experiments: individuals with no (or negligible) music experience are allocated to a music-training group and one or more control groups. This meta-analytic review has found modest or null effects of music training on cognitive abilities such as intelligence, memory, spatial ability, and phonological processing (see also Gordon, Fehd, & McCandliss, 2015). Similar modest or null effects have been found with academic skills such as mathematics and literacy. Furthermore, meta-regression analysis has highlighted that the between-study variability is moderated by the type of control group (active or passive) and the type of allocation to the groups (randomized or nonrandomized). While the studies with no random allocation and passive control groups show some positive effects, when the music-trained groups are randomly allocated and compared to an active control group, the effects are null.

Although Sala and Gobet's (2017b) meta-analytic review suggests pessimism, numerous new experimental studies have been carried out in the last three years. Some of these studies have reaffirmed the idea that music training has a positive influence on children's cognitive and academic skills. However, the impact of these new studies on the overall evaluation of the field of music training has not been assessed yet.

The present study intends to update Sala and Gobet's (2017b) meta-analysis and test the recent claims about the presumed cognitive and academic benefits of music training (e.g., Habibi et al., 2018). To achieve this goal, we (a) extend the literature research to the last three years (from January the 1st 2016 to December the 31st 2018), (b) apply a more advanced modeling approach, and (c) provide stricter inclusion criteria to improve, compared to Sala and Gobet (2017b), the average quality of the studies included in the meta-analytic review.

Method

Literature Search

A systematic search strategy was implemented (Moher et al., 2009). Using the following Boolean string search ("music" OR "musical") AND ("training" OR "instruction" OR "education" OR "intervention"), ERIC, Psyc-Info, and ProQuest Dissertation & Theses databases were searched to identify all the potentially relevant studies. In addition, all the studies included in Sala and Gobet (2017b) were reevaluated for inclusion. Also, we e-mailed researchers in the field (n = 8) asking for unpublished studies, clarifications about the study design, and inaccessible data.

Inclusion Criteria

We kept the same inclusion criteria as Sala and Gobet (2017b). The study had to include (a) a cognitively-demanding music-training program (e.g., learning to play instruments, Kodály method,[1] etc.; no correlational studies were included), (b) at least one control group, (c) non-music-related cognitive or academic outcomes,[2] and (d) participants aged between 3 and 16 with no diagnosed clinical condition or previous formal music experience.

In order to improve the overall quality of the reviewed empirical evidence, the present meta-analysis added three more criteria: (e) the article had to report (or the author had to provide) the means and standard deviations in order to calculate the effect size and sampling error variance; (f) the participants had to be allocated by the experimenter to a

[1] The Kodály method is a well-known educational protocol that focuses on singing, ear training, and the creative skills of musicianship. For more details, see http://kodaly.org.uk/.

[2] For a discussion about the potential benefits of music instruction on non-cognitive/academic skills, see Aleman et al. (2017).

group (randomly or nonrandomly); that is, they were not allowed to decide to which group (experimental or control) they would be allocated; and (g) the experimental group and the control group had to include comparable populations (e.g., same grade, comparable baseline IQ, etc.). These additional criteria led to the exclusion of eight studies previously included in Sala and Gobet (2017b).

Moderators

We chose (a priori) five potential moderators that were included in the meta-regression analysis:

1. Allocation (dichotomous variable): Whether the children were randomly allocated to the groups;
2. Type of control group (active or passive; dichotomous variable): Whether the WM training-treated group was compared to an alternative activity (e.g., visual arts) or a do-nothing (passive) control group. This moderator was necessary to check for placebo effects;
3. Baseline difference (continuous variable): The standardized mean difference (Hedges's *g*) between the experimental and control groups at baseline.[3] A negative regression coefficient would suggest the presence of some true heterogeneity due to regression to the mean;
4. Age (continuous variable): The age of the participants in years;
5. Outcome Measure (categorical variable): The effect sizes were clustered into four broad categories: non-verbal ability (e.g., reasoning, mathematical, and spatial skills); verbal ability (e.g., vocabulary and reading skills); memory (e.g., digit-span and working-memory tasks); and speed (e.g., processing speed and inhibition tasks).[4] The interrater agreement was perfect ($\kappa = 1$).

Modeling Approach

We extracted the effect sizes for each relevant dependent variable reported in the studies using the formulas provided by Schmidt and Hunter (2015). Several studies presented multiple-group comparisons – for example, between one experimental group and two control groups (one active and one passive), or between two experimental groups and one control group. In these cases, we calculated as many effect sizes as the number of comparisons.

We grouped all the effect sizes from the same study into the same cluster. Then, we employed *robust variance estimation* (RVE; Hedges, Tipton, & Johnson, 2010) to model statistically dependent effect sizes and calculates adjusted (i.e., increased) overall standard errors. Also, RVE provides estimates of within-cluster true (i.e., not due to random error) heterogeneity and between-cluster true heterogeneity (ω^2 and τ^2, respectively). We ran (a) intercept models to calculate overall effect sizes and (b) meta-regression models to assess the amount of true heterogeneity explained by the moderators.

Publication Bias

To control for publication bias, we first merged the effects from the same study with the method designed by Cheung and Chan (2014; individual-samplewise procedure). The method averages the effect sizes from the same cluster (in this case, the study) and calculates a corrected sampling error variance in order not to miscalculate standard errors and true heterogeneity. Then, we ran a random-effect model with the merged effect sizes and applied the trim-and-fill publication-bias detection method (Duval & Tweedie, 2000; estimators *L0*, *R0*, and *Q0*).

Results

The search yielded 2,462 records, of which 72 studies were thoroughly evaluated for inclusion. Forty-three studies, 13 of which not included in Sala and Gobet (2017b), met the inclusion criteria with a total of 204 effect sizes (Sala & Gobet, 2017b, included 118 effect sizes). The total number of participants was 3,780. Three researchers replied to our emails. The supplemental materials including details about the studies, techniques employed, additional analyses, data, and R codes, can be found at this link: https://osf.io/2gce3/.

Main Model

The intercept model did not include any covariate (i.e., moderator). The overall effect size of the RVE intercept model was $\bar{g} = 0.117$, 95% CI [0.063; 0.170], $m = 43$, $k = 204$, $df = 17.25$, $p < .001$, $\omega^2 = 0.010$, $\tau^2 = 0.005$. The overall impact of music-training interventions was thus small ($\bar{g} = 0.117$, 95% CI [0.063; 0.170]), albeit statistically significant ($p < .001$).

After merging the effects from the same cluster (i.e., the study), the results of the random-effect model were very similar: $\bar{g} = 0.140$, 95% CI [0.064; 0.217], $p < .001$, $k = 43$, $\tau^2 = 0.018$. The trim-and-fill analysis indicated some publication bias (estimates ranging between 0.046 and 0.122).

Meta-Regression Analysis

The meta-regression model included the five moderators described in the Method section. Baseline and Type of control group were the only two significant moderators ($p = .019$ and $p = .003$, respectively). These two moderators explained almost all the observed true heterogeneity ($\omega^2 = 0.000$ and $\tau^2 = 0.005$). We also checked all the pairwise comparisons for the outcome measures with the Holm's method (for details, see the supplemental materials). None of the comparisons yielded significant differences (all ps $\geq .610$).

Finally, we sorted the effect sizes by the moderator Type of control group. The overall effect size of the RVE model including only passive-control comparisons was $\bar{g} = 0.173$,

[3] Five studies implemented an only-post-test design. In those cases, baseline differences were assumed to be null to keep these studies in the moderator analysis.

[4] A more fine-grained categorization was also analyzed (for details, see supplemental materials).

95% CI [0.094; 0.253], $m = 33$, $k = 112$, $df = 19.02$, $p < .001$, $\omega^2 = 0.008$, $\tau^2 = 0.019$. The overall effect size of the RVE model including only active-control comparisons was $\bar{g} = 0.032$, 95% CI [-0.068; 0.132], $m = 19$, $k = 92$, $df = 7.09$, $p = .477$, $\omega^2 = 0.000$, $\tau^2 = 0.000$. Thus, while a small positive and significant effect was observed when passive controls were implemented, no substantial effect occurred when music-treated subjects were compared to controls involved in other activities (Table 1).

Table 1: Summary of the results.

Sample	$\bar{g}$ [95% CI]	p-value
All	0.117 [0.063; 0.170]	.000
Exp. vs passive	0.173 [0.094; 0.253]	.000
Exp. vs active	0.032 [-0.068; 0.132]	.477

Discussion

The present meta-analysis has aimed to update and check the findings of the most recent and comprehensive meta-analysis about the impact of music instruction on children's non-music-related cognitive and academic skills. It has included new studies and nearly doubled the number of effect sizes compared to Sala and Gobet (2017b). Nonetheless, the results of this meta-analysis confirm most of the findings reported in Sala and Gobet (2017b). Most importantly, when only those designs implementing an active control group are considered, the effect of music training is practically null ($\bar{g} = 0.032$, $p = .477$) and highly consistent ($\omega^2 = 0.000$, $\tau^2 = 0.000$). On the other hand, the comparison between music-trained groups and passive controls yields a minimal overall effect ($\bar{g} = 0.173$, $p < .001$) that is easily accounted for by placebo effects. Therefore, the effects of music training on children's cognitive skills and academic achievement are unspecific. Consistent with this explanation, there were no differences between outcome measures, which suggests that the effects of music training (when any) are unspecific.

Finally, beyond supporting Sala and Gobet's (2017b) findings, this meta-analysis highlights new aspects. First, the lack of randomization does not seem to affect the outcomes. On the other hand, compared to Sala and Gobet (2017b) using more rigorous inclusion criteria (e.g., no studies with self-selected participants) lowers the overall effect size (from 0.173 to 0.117) and, most notably, the amount of true heterogeneity (from $\omega^2 = 0.088$ to $\omega^2 = 0.010$, and from $\tau^2 = 0.023$ to $\tau^2 = 0.005$).[5] Second, regression to the mean appears to explain a significant amount of true heterogeneity. This finding does not imply that baseline differences have affected the overall effects. Rather, it means that some of the observed true heterogeneity is spurious (i.e., due to a statistical artifact).

[5] These statistics were obtained by reanalysing Sala and Gobet's (2017b) original dataset with the same multilevel approach used in the current meta-analysis (i.e., RVE).

Triangulation

Beyond meta-analytic evidence, our findings are supported by substantial research into the field of music cognition using different methodologies. Mosing et al. (2016) have shown that music-trained twins do not have a higher IQ than the relative non-music-trained co-twins. The study thus suggests that the level of IQ is determined, to a significant extent, genetically and that engaging in music has no effect on it. Also, Swaminathan, Schellenberg, and Khalil (2017) have recently shown that music aptitude, but not the amount of music training, predicts intelligence in a sample of adults. The association between intelligence (Raven's progressive matrices) and music training is evident until music aptitude is taken into account and added to the regression model.

Strong support for our conclusions is also provided by the fact that the same pattern of results has been found in other domains, including chess training, working-memory training, and brain training. Expertise in chess has been found to correlate with a broad range of cognitive skills such as fluid intelligence, processing speed, short-term memory, and spatial ability (e.g., Burgoyne et al., 2016). Moreover, expert chess players differ from novices and non-players in terms of neural anatomical and functional patterns (e.g., Bilalić et al., 2010; Hänggi et al., 2014). However, chess training does not seem to trigger any genuine improvement in overall cognitive ability or academic achievement (Sala & Gobet, 2016). Analogously, fluid intelligence and working memory capacity are strongly correlated, yet working memory training exerts no effect on fluid intelligence (e.g., Melby-Lervåg et al., 2016). The absence of far-transfer effects is observed even in the presence of functional neural changes (Clark, Lawlor-Savage, & Goghari, 2017). A similar pattern of results has been reported in brain training as well (for a review, see Simons et al., 2016). This outcome upholds the idea that such neural patterns underlie domain-specific skills (e.g., performance in working-memory tasks) rather than overall cognitive function.

These similarities between the results obtained with training studies in different domains induce further pessimism about the concrete possibility of enhancing domain-general cognitive skills through the engagement in intellectually demanding activities. In brief, the idea of enhancing overall cognitive ability through training appears, to date, scientifically implausible (Sala & Gobet, 2019).

Concerning the observed neural patterns in musicians, understanding their actual significance is essential. It is doubtful that functional changes occurring after a music-training intervention represent domain-general improvements in cognitive function. Instead, it is probable that such neural patterns underlie the enhancement of music-related skills such as pitch discrimination (e.g., Nan et al., 2018). It is thus imperative not to erroneously interpret – as sometimes happens (e.g., Habibi et al., 2016) – that

functional neural changes in brain areas involved in domain-general cognitive abilities are evidence of cognitive enhancement. The same applies to anatomical neural changes (e.g., increased density of gray matter). Such patterns frequently observed in professional musicians are most likely neural correlates of their domain-specific expertise rather than superior overall cognitive ability.

Theoretical and Practical Implications

Taken together, the findings of the research into music expertise and music training depict a consistent picture: while a positive relationship between music skill and cognitive ability does exist, the benefits of music training do not go beyond the acquisition of music-related skills. In other words, engaging in music does not make people smarter. Instead, as suggested by the research of Mosing et al. (2016) and Swaminathan et al. (2017), smarter people seem to be more likely to engage and succeed in music.

Two major theoretical implications stem from the above results. First, the lack of generalization of music skills acquired by training provides further corroboration for Thorndike and Woodworth's (1901) common elements theory and theories based on the mechanism of chunking. According to Thorndike and Woodworth's theory, transfer of skills is a function of the extent to which two (or more) domains overlap. Thus, transfer of skill between two (or more) domains only loosely related to each other (i.e., far transfer) hardly occurs. Similarly, chunking and template theories (Chase & Simon, 1973; Gobet & Simon, 1996) predict modest or no transfer across different domains or even subdomains of expertise (for a review, see Gobet, 2016). This is because these theories uphold the idea that skill acquisition is based, to a large extent, on perceptual information (i.e., perceptual chunks and templates), which is hardly transferable across different domains given its highly domain-specific nature. Conversely, theories predicting the generalization of trained skills across different domains (e.g., Strobach & Karbach, 2016) are not supported by these outcomes.

Second, the observed correlation between music skill and cognitive ability, together with the lack of broad cognitive effects following music training, suggests that talent is an essential requisite for achieving expertise in music (Schellenberg, 2015). In line with the conclusions of Macnamara, Hambrick, and Oswald (2014), substantial research into music confirms that the amount of deliberate practice alone cannot account for the individual differences in music expertise.

Beyond theoretical aspects, the obvious practical implication is that music training should not be used as a tool for cognitive enhancement. In fact, music training has failed to offer any specific advantage in terms of both cognitive enhancement and academic achievement. These conclusions are made even stronger if we take into consideration that music training has been found substantially ineffective even at enhancing those skills traditionally believed to be tightly close to music skill, such as phonological processing and literacy (e.g., Kempert et al., 2016).

Recommendations for Future Research and Conclusions

As seen, music training does not affect any non-musical cognitive or academic skills. Importantly, the lack of generalization of music skills acquired by training has been established by different research teams using diverse research methodologies (twin studies, hierarchical multiple regression, and meta-analysis of treatment studies).

We briefly discuss some possible avenues of research. As noted above, the quality of experimental designs is inversely related to the size of the effects of music-training interventions and cognitive-training interventions in general (Moreau, Kirk, & Waldie, 2016). Therefore, future studies should strive for high-quality experimental designs regardless of the particular outcome variables and population under investigation. We thus recommend including both active and passive control groups, random allocation of the participants, pre-, post-, and follow-up assessment, multiple measures of the same constructs, and large samples.

It is worth emphasizing that the findings reported here about the null effects of music training do not imply that music is a worthless activity. Rather, the purpose of this article has been to clarify what are the real effects of music training in order to allow people to make informed decisions. Educators and policymakers should be aware that music training provides no benefits on non-music-related cognitive or academic skills (e.g., Nan et al., 2018). As far as we are concerned, even in the absence of other cognitive or academic benefits, it is worthwhile learning an art present in nearly all the cultures in human history.

Acknowledgments

We thank all the authors who replied to our requests for data and clarification about the primary studies. The support of the Japan Society for the Promotion of Science [17F17313] is gratefully acknowledged.

References

Aleman, X., Duryea, S., Guerra, N. G., Mcewan, P. J., Muñoz, R., Stampini, M., & Williamson, A. A. (2017). The effects of musical training on child development: A randomized trial of El Sistema in Venezuela. *Prevention Science, 18*, 865-878.

Anvari, S. H., Trainor, L. G., Woodside, J., & Levy, B. A. (2002). Relations among musical skills, phonological processing, and early reading ability in preschool children. *Journal of Experimental Child Psychology, 83*, 111-130.

Bermudez, P., Lerch, J. P., Evans, A. C., & Zatorre, R. J. (2009). Neuroanatomical correlates of musicianship as revealed by cortical thickness and voxel-based morphometry. *Cerebral Cortex, 19*, 1583-1596.

Bialystok, E., & Depape, A. M. (2009). Musical expertise, bilingualism, and executive functioning. *Journal of*

Experimental Psychology: Human Perception and Performance, 35, 565-574.

Bilalić, M., Langner, R., Erb, M., & Grodd, W. (2010). Mechanisms and neural basis of object and pattern recognition: A study with chess experts. *Journal of Experimental Psychology: General, 139*, 728-742.

Chase, W. G., & Simon, H. A. (1973). The mind's eye in chess. In W. G. Chase (Ed.), *Visual information processing* (pp. 215-281). New York: Academic Press.

Cheung, S. F., & Chan, D. K. (2014). Meta-analyzing dependent correlations: An SPSS macro and an R script. *Behavioral Research Methods, 46*, 331-345.

Deary, I. J., Strand, S., Smith, P., & Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence, 35*, 13-21.

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel plot based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, *56*, 455-463.

Forgeard, M., Schlaug, G., Norton, A., Rosam, C., Iyengar, U., & Winner, E. (2008). The relation between music and phonological processing in normal-reading children and children with dyslexia. *Music Perception, 25*, 383-390.

Gobet, F. (2016). *Understanding expertise: A multi-disciplinary approach*. London: Palgrave/Macmillan.

Gobet, F., & Simon, H. A. (1996). Templates in chess memory: A mechanism for recalling several boards. *Cognitive Psychology, 31*, 1-40.

Gordon, R. L., Fehd, H. M., & McCandliss, B. D. (2015). Does music training enhance literacy skills? A meta-analysis. *Frontiers in Psychology, 6*:1777.

Habibi, A., Cahn, B. R., Damasio, A., & Damasio, H. (2016). Neural correlates of accelerated auditory processing in children engaged in music training. *Developmental Cognitive Neuroscience, 21*, 1-14.

Habibi, A., Damasio, A., Ilari, B., Sachs, M. E., & Damasio, H. (2018). Music training and child development: A review of recent findings from a longitudinal study. *Annals of the New York Academy of Sciences, 1423*, 73-81.

Hänggi, J., Brütsch, K., Siegel, A. M., & Jäncke, L. (2014). The architecture of the chess player's brain. *Neuropsychologia, 62*, 152-162.

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods, 1*, 39-65.

Jaušovec, N., & Pahor, A. (2017). *Boost your IQ with music*. Retrieved from http://scitechconnect.elsevier.com/boost-your-iq-with-music/

Kempert, S., Götz, R., Blatter, K., Tibken, C., Artelt, C., Schneider, W., & Stanat, P. (2016). Training early literacy related skills: To which degree does a musical training contribute to phonological awareness development? *Frontiers in Psychology, 7*:1803.

Koelsch, S., Fritz, T., von Cramon, D. Y., Müller, K., & Friederici, A. D. (2006). Investigating emotion with music: An fMRI study. *Human Brain Mapping, 27*, 239-250.

Macnamara, B. N., Hambrick, D. Z., & Oswald, F. L. (2014). Deliberate practice and performance in music, games, sports, education, and professions: A meta-analysis. *Psychological Science, 25*, 1608-1618.

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Annals of Internal Medicine*, *151*, 264-269.

Moreau, D., Kirk, I. J., & Waldie, K. E. (2016). Seven pervasive statistical flaws in cognitive training interventions. *Frontiers in Human Neuroscience, 10*:153.

Mosing, M. A., Madison, G., Pedersen, N. L., & Ullén, F. (2016). Investigating cognitive transfer within the framework of music practice: Genetic pleiotropy rather than causality. *Developmental Science, 19*, 504-512.

Nan, Y., Liu, L., Geiser, E., Shu, H., Gong, C. C., Dong, Q., … Desimone, R. (2018). Piano training enhances the neural processing of pitch and improves speech perception in Mandarin-speaking children. *PNAS, 115*, E6630-E6639.

Neumann, N., Lotze, M., & Eickhoff, S. B. (2016). Cognitive expertise: An ALE meta-analysis. *Human Brain Mapping, 37*, 262-272.

Ruthsatz, J., Detterman, D., Griscom, W. S., & Cirullo, B. A. (2008). Becoming an expert in the musical domain: It takes more than just practice. *Intelligence, 36*, 330-338.

Saarikivi, K., Putkinen, V., Tervaniemi, M., & Huotilainen, M. (2016). Cognitive flexibility modulates maturation and music-training-related changes in neural sound discrimination. *European Journal of Neuroscience, 44*, 1815-1825.

Sala, G., & Gobet, F. (2016). Do the benefits of chess instruction transfer to academic and cognitive skills? A meta-analysis. *Educational Research Review, 18*, 46-57.

Sala, G., & Gobet, F. (2017a). Experts' memory superiority for domain-specific random material generalizes across fields of expertise: A meta-analysis. *Memory & Cognition, 45*, 183-193.

Sala, G., & Gobet, F. (2017b). When the music's over. Does music skill transfer to children's and young adolescents' cognitive and academic skills? A meta-analysis. *Educational Research Review, 20*, 55-67.

Sala, G., & Gobet, F. (2019). Cognitive training does not enhance general cognition. *Trends in Cognitive Sciences, 23*, 9-20.

Schellenberg, E. G. (2006). Long-term positive associations between music lessons and IQ. *Journal of Educational Psychology, 98*, 457-468.

Schellenberg, E. G. (2015). Music training and speech perception: A gene-environment interaction. *Annals of the New York Academy of Science, 1337*, 170-177.

Schmidt, F. L., & Hunter, J. E. (2015). *Methods of meta-analysis: Correcting error and bias in research findings (3rd ed.)*. Newbury Park, CA: Sage.

Sluming, V., Barrick, T., Howard, M., Cezayirli, E., Mayes, A., & Roberts, N. (2002). Voxel-based morphometry reveals increased gray matter density in Broca's area in

male symphony orchestra musicians. *NeuroImage, 17*, 1613-1622.

Strobach, T., & Karbach, J. (2016). *Cognitive training: An overview of features and applications*. New York: Springer.

Swaminathan, S., Schellenberg, E. G., & Khalil, S. (2017). Revisiting the association between music lessons and intelligence: Training effects or music aptitude? *Intelligence, 62*, 119-124.

Thorndike, E. L., & Woodworth, R. S. (1901). The influence of improvement in one mental function upon the efficiency of other functions. (I). *Psychological Review, 8*, 247-261.

Wetter, O. E., Koerner, F., & Schwaninger, A. (2009). Does musical training improve music performance? *Instructional Science, 37*, 365-374.

Do cross-linguistic patterns of morpheme order reflect a cognitive bias?

Carmen Saldana (carmen.saldana@ed.ac.uk)
Centre for Language Evolution, The University of Edinburgh,
3 Charles Street, Edinburgh, EH8 9AD, UK

Yohei Oseki (oseki@aoni.waseda.jp)
Faculty of Science and Engineering, Waseda University,
3-4-1 Okubo, Shinjuku, Tokyo 169-8555, Japan

Jennifer Culbertson (jennifer.culbertson@ed.ac.uk)
Centre for Language Evolution, The University of Edinburgh,
3 Charles Street, Edinburgh, EH8 9AD, UK

Abstract

A foundational goal of linguistics is to investigate whether shared features of the human cognitive system can explain how linguistic patterns are distributed across languages. In this study we report a series of artificial language learning experiments to test a hypothesised link between cognition and a persistent regularity of morpheme order: number morphemes (e.g., plural markers) tend to be ordered closer to noun stems than case morphemes (e.g., accusative markers) (Greenberg, 1963). We argue that this typological tendency may be driven by a bias favouring orders that reflect scopal relationships in morphosyntactic composition (Bybee, 1985; Rice, 2000; Culbertson & Adger, 2014). We taught participants an artificial language with noun stems, and case and number morphemes. Crucially, the input language indicated only that each morpheme preceded or followed the noun stem. Examples in which two (overt) morphemes co-occurred were held out—i.e., no instances of plural accusatives. At test, participants were asked to produce utterances, including the held-out examples. As predicted, learners consistently produced number closer to the noun stem than case. We replicate this effect with free and bound morphemes, pre- or post-nominal placement, and with English and Japanese speakers. However, we also find that this tendency can be reversed when the form of the case marker is conditioned on the noun, suggesting an influence of dependency length. Our results provide evidence that universal features of cognition may play a causal role in shaping the relative order of morphemes.

Keywords: linguistic universals; artificial language learning; morpheme order; case; number

Introduction

Human languages are incredibly diverse in the way they combine meaningful units, i.e., morphemes; nevertheless, certain regularities are apparent. For example, some patterns of morpheme order occur more frequently across the languages of the world, while others are rare or even unattested. The typological regularity in morpheme order we target here concerns number and case morphology, specifically, languages in which there is a boundary between these morphemes. For example, in agglutinating languages such as Hungarian or Turkish, there is distinct set of number morphemes (marking plurality) and case morphemes (marking grammatical roles). In such languages, when overt morphemes of both number and case are present on a stem, and both follow or both precede the noun stem, the expression of number is almost always realised closer to the noun stem than the expression of case (Universal 39; Greenberg, 1963). There are a number of candidate explanations for this phenomenon, which intersect with high-level hypotheses about how morpheme (and word) order is determined in language more generally. For example, it has been proposed that semantic or compositional relationships among morphemes, sometimes called *scope*, determine linear order (Bybee, 1985; Wunderlich, 1993; Rice, 2000; Culbertson & Adger, 2014).[1] On one formulation, morphemes which more directly affect or modify the semantic content of the stem have narrower scope (Bybee, 1985; Rice, 2000). Wider-scope morphemes modify the larger semantic constituent which includes any lower scoping morphemes. Perhaps the best-known example of this is the order of derivational and inflectional morphemes (e.g., 'neighbor-hood-s'). On this account, derivational morphemes are ordered closer to the stem because they change its lexical meaning. Inflectional morphemes scope higher, modifying grammatical properties of the stem plus any derivational morphemes. Similarly, it has been claimed that the linear order of nominal modifiers (e.g., adjectives, numerals, demonstratives) reflects semantic scope relations (Culbertson & Adger, 2014; Bouchard, 2002). In the case of Universal 39, the idea would be that case scopes higher than number because number directly modifies the entity referred to by the noun, while the case morpheme signals an external relationship between the entity and some event. Following Culbertson and Adger (2014), we call orders which reflect scope relations *scope-isomorphic*.

A second possible explanation appeals to frequency and its effects on processing. For example, Ryan (2010) shows that in some cases morpheme order reflects the frequency of stem+morpheme bigrams (see also Baayen, 1993; Rice, 2011). Along similar lines, Hay (2001) argues that when a stem is more frequent alone than with a particular affix, then that affix is easier to parse (decompose) from the stem. This in turn determines linear order: more parsable affixes appear farther from the stem than less parsable ones (see also Hay & Plag, 2004; Plag & Baayen, 2009; Manova & Aronoff, 2010). How might this explain Universal 39? It could be that

[1] Related theories argue that universal morphosyntactic hierarchies, potentially reflecting semantics, determine order (Baker, 1985; Grimshaw, 1986; Cinque, 2005).

number tends to be expressed more often than case, or that case morphemes tend to be more parsable than number morphemes. On this account, there is nothing about the semantics of these morphemes that determines their relative order. Indeed, a third possibility is that their relative order reflects patterns of diachronic change: it could be that languages tend to grammaticalise number before case (Givón, 1979).

To date, there is no direct behavioral evidence adjudicating among these potential explanations for Universal 39. In fact, there is no independent evidence beyond the typology to show that placing number closer to the noun stem than case is in fact preferred over the reverse. In a series of three artificial language learning experiments, we test the link between this typological generalisation and a bias towards linear orders that mirror scopal relationships (henceforth scope-isomorphic orders). To summarise, we find support for this hypothesis across two language populations (English, and Japanese) independent of morpheme position (before or after the noun stem), degree of boundedness, and frequency. All things equal, learners therefore prefer scope-isomorphic orders. However, we also find that conditional allomorphy between the stem and the case marker can reverse participants' preferences. We interpret this as a competing bias for local dependencies. This result adds to the growing body of work using these experimental methods to investigate how learning and use shape morphology and word order (Hupp, Sloutsky, & Culicover, 2009; Fedzechkina, Jaeger, & Newport, 2012; Culbertson & Adger, 2014; Culbertson, Smolensky, & Legendre, 2012; Tabullo et al., 2012; Futrell, Mahowald, & Gibson, 2015; Fedzechkina, Chu, & Jaeger, 2018).

Experiment 1

Methods

The artificial language learning experiments described here use an extrapolation paradigm (called 'Poverty-of-the-stimulus' paradigm elsewhere, Wilson, 2003; Culbertson & Adger, 2014). This means learners are trained on input that is designed to be ambiguous between (at least) two patterns of interest: here, two potential ways of ordering case and number morphemes. Learners are exposed to a miniature artificial language with nouns, and case (accusative) and number (plural) morphemes. Crucially, their input indicates whether these morphemes generally precede or follow the noun, but does not include any examples in which the two morphemes co-occur within the same noun phrase. At test, they are asked to produce utterances, including these held out examples. The order they infer will indicate whether they have a preference for placing number closest to the noun (e.g., Noun-Number-Case rather than Noun-Case-Number). All experiment materials and data discussed here are available at osf.io/9fa3v/, and the preregistered design and analysis plan for Experiment 1 is accessible at osf.io/8xuc9.

Participants Forty-one native English speakers were recruited from the University of Edinburgh's Careers Services database. Participants were paid £6 for a 35-min-long experimental session. Participants (N=1) whose vocabulary accuracy was lower than 60% were excluded; testing trials with incomplete sentences were also excluded.

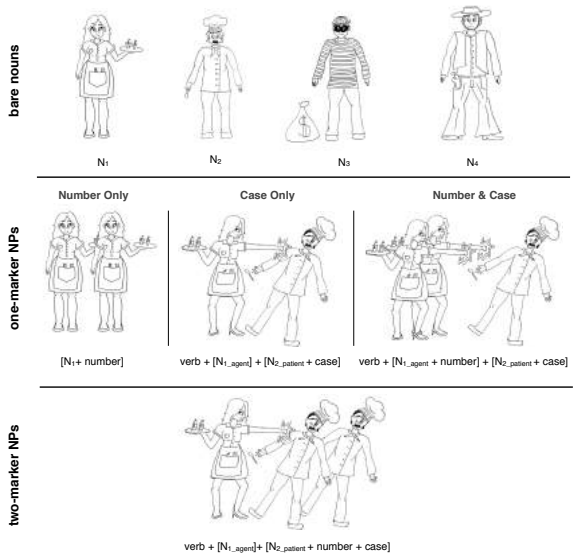


Figure 1: Example visual stimuli and corresponding descriptions. Top to bottom: the four characters in the miniature language in isolation; example events with one marker (either number or case); example event requiring two markers (number and case, testing only).

Input language The lexicon includes three semi-nonce verbs, four nonce nouns, and two nonce markers (one number marker indicating plural; one case marker indicating accusative). All words have initial stress. The three semi-nonce verbs are taken from the English-based creole Tok Pisin: 'kikim'([ˈkʰɪkɪm]), 'poinim'([ˈpʰɔɪnɪm]) and 'straikim'([ˈstraɪkɪm]), which refer to 'kicking', 'pointing' and 'punching' respectively. The (disyllabic) nouns are 'negid'([neʒɪd]), 'nork'([ˈnɔrk]), 'tumbat' ([ˈtʰʌmbət]) , 'vaem' ([ˈvæm]) (based on Fedzechkina et al., 2012), naming four characters: a burglar, a chef, a cowboy, and a waitress. The noun-character mappings are random for each participant. The two markers were randomly mapped to number and case from the set: 'gu' ([ˈguː]), 'sa'([ˈsɑː]), and 'ti'([ˈtʰiː]). Word order in sentences was Verb-Agent-Patient. Half of participants were trained on a language with post-nominal morphemes (case and number morphemes appeared after the noun stem), half with pre-nominal morphemes (case and number morphemes appeared before the noun stem). [2]

Participants are trained on three different NP types: a bare noun, a noun with overt number morphology, and a noun with

[2]We use the terms *pre-* and *post-nominal* instead of *prefixal* and *suffixal* morphology to account for both bound and unbound orthographic representations of case and number morphology.

overt case morphology. Note that singular, and agent case (nominative) are unmarked. During training, participants get descriptions of characters in isolation (singular or plural), or events with a singular patient; plural patients (requiring both number and case morphology) are held-out until testing. See Figure 1 for examples. Crucially, number and case markers appear with the exact same frequency (i.e., absolute, and relative to each given noun) both during training and testing phases, controlling for any potential frequency effects.

The input language is presented both orthographically and auditorily during training. Auditory stimuli were recorded in a sound-attenuated room by a 26yo male speaker of American English. Noun phrases were recorded without a pause between nouns and markers but each marker is orthographically presented surrounded by spaces and thus not bound to the noun.

Experimental procedure The experiment was conducted in a quiet room, with all instructions provided in English, and an English-speaking experimenter. Participants were told that they would be learning part of a foreign language. The session proceeded as follows.

Phase 1, noun training and testing. Participants are first trained on the four nouns in isolation (Figure 1, top row) during a block of 24 trials (6 per noun). In each trial, a single character appears, and its description (a bare noun) is displayed (orthographically and auditorily). Participants are instructed to repeat each description aloud. Participants are then tested on the noun vocabulary using a noun-selection task and an oral production task (12 trial per block, 3 per noun). In noun-selection trials, a character appears, and participants must select the correct noun from 2 choices. The foil noun is randomly selected at each trial. Feedback is provided (an (in)correct-answer sound effect along with the image and correct noun; if incorrect, the audio of the noun is also played). In oral production trials, a character appears, and participants must say the corresponding noun aloud. Feedback is provided (the correct noun is displayed visually and auditorily after participants submit their answer).

Phase 2, one-marker NP training. Participants are next trained on noun phrases with a single marker, either number or case. There are three trial types (Figure 1, middle row): (1) a group of the same characters (2, 3, or 4) in isolation (Number only), (2) an event with (different) singular agent and patient (Case only), or (3) an event with a plural agent, and a singular patient (Number & Case, where crucially each marker belongs to a different noun phrase). On each training trial, participants see an image, and its description is presented (orthographically and auditorily). There are 62 trials total (randomised): 8 bare noun, 18 Number Only (six per character), 18 Case Only (randomly chosen from the 36 possible), and 18 Number & Case images (again randomly chosen).

Phase 3, one-marker NP comprehension test. Participants are then tested on their comprehension of one-marker NPs in a image-selection task. On each trial, they get a description and must select the corresponding image out of an array of two. Feedback is provided (an (in)correct-answer sound effect along with the image and correct orthographic description; if incorrect, the audio description is also played). The foil image is selected according to the trial type. For bare noun and Number Only trials, the foil image is the same character with wrong numerosity (e.g., singular instead of plural). For Case Only and Number & Case trials, the foil is the same event type with agent and patient reversed. There are 34 trials total (randomised): 4 bare noun, 10 each of the three one-marker NP trial types.

Phase 4, one-marker NP written production test. Participants are then tested on their ability to produce one-marker NP descriptions. On each trial, participants see a image and are required to type in the corresponding NP(s). Verb forms are provided for Case Only and Number & Case trials. Feedback is provided (an (in)correct-answer sound is played, along with the image and correct description). There are 16 trials total (randomised): 4 trials for each of the types they have been trained on so far.

Phase 5, two-marker NP production tests. In the two critical testing blocks, participants must provide first written, then oral descriptions which include the held-out phrase type: two marker NPs, with plural patients (Figure 1, bottom row). The written production task is identical to Phase 4, except it only includes the held-out trial types (12 trials, 3×4 events randomly chosen) and no feedback is given. This written task is added with the purpose of familiarising participants with the held-out trial types prior to the final oral production test phase and will not be included in our analyses.

Finally, participants are asked to produce oral descriptions for *all* trial types in the language. On each trial, participants see a image and are asked to provide a description aloud. As in the previous written production trials, participants are provided with the corresponding verb form when necessary. Feedback is provided (as described above) *only* when the target description does not contain a two-marker NP. There are 58 trials total (randomised): 36 two-marker NP trials, 6 trials of each of the three one-marker NP trial types, and four bare noun trials.

Results

Recall that, based on Universal 39 (Greenberg, 1963), participants are predicted to produce number markers closer to the noun stem than case markers. This should hold for both the pre- and post-nominal conditions. Our working hypothesis is that these orders are preferred because they reflect the scopal relations among morphemes. Figure 2 is a stacked histogram, showing the percentage of participants whose oral productions follow scope in 0-100% of trials across both conditions. Experiment 1 results (with English speakers) are on the left-hand side. For critical trials, 95% of participants are (almost) perfectly consistent, producing two-marker NPs in the predicted order 95-100% of the time. We ran a logistic mixed-effects regression model predicting use of scope-isomorphic morpheme orders on two-marker NPs dur-

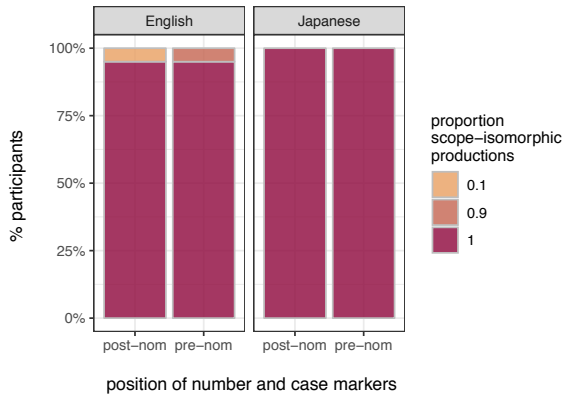


Figure 2: Percentage of participants in Experiments 1 (English) and 2 (Japanese) who produced scope-isomorphic responses a given proportion of the time (rounded to one decimal), ranging from 0% of the time (yellow) to 100% of the time (dark red). Results are split by Marker Position (pre- vs. post-nominal).

Table 1: Model output for Experiment 1.

	β	*SE*	*z*	$Pr(>\|z\|)$
(Intercept)	13.398	3.213	4.169	< 0.001
Marker Position	−0.219	2.428	−0.090	0.928

ing oral production by Marker Position (pre-nominal vs. post-nominal).[3] As shown in Table 1, the intercept (grand mean of scope-isomorphic productions across participants in both conditions) is positive and significant, confirming that the average proportion of scope-isomorphic productions ($P \approx 1$) is above chance. The effect of Marker Position is not significant, confirming that this preference holds regardless of the pre- or post-nominal positioning of the markers.

Experiment 2

The results of Experiment 1 are consistent with the hypothesis that scope relations—here between number and case morphemes—determine proximity to the noun stem. Importantly, we can rule out the effect of raw or bigram frequency in driving our results, since these were held constant in our stimuli. However, an alternative explanation is that our result reflects the fact that English overtly marks (plural) number but it does not have morphological case marking (aside from perhaps the genitive). Exactly how this would lead to a preference for placing number closer than case is not totally clear. Perhaps familiarity with, or accessibility of the number marker leads English speakers to place it closer to the noun. To rule this out, we replicated Experiment 1 with native speakers of Japanese. In contrast to English, Japanese overtly marks cases (including accusative) via suffixation; however, the marking of plurality is exceptional (Nakanishi & Tomioka, 2004). The closest thing to number marking *on nouns* are the associative plural classifiers or collectivising suffixes (*-kata*, *-tachi*, *-ra*, *-domo*). Number is typically expressed instead via plural words (which appear after the case inflected noun), reduplication or numeral words (which precede the noun). Japanese speakers should therefore have no trouble acquiring a novel accusative case marker, and if anything should find the case marker more familiar/accessible than the number marker.

Methods

Experiment 2 is identical to Experiment 1, with one difference: the input lexicon. Rather than using a language with English-like phonotactics, the lexicon for Experiment 2 matched Japanese phonotactics. The preregistered design and analysis plan for Experiment 2 is accessible at osf.io/akcyp.

Participants Forty native Japanese speakers were recruited from Waseda University's student database. Participants were paid ¥1000 for a 35-min-long experimental session. Note that all participants spoke English as an L2.

Input language Lexical items in the language were displayed in Katakana (instead of Latin) script. The three seminonce verbs (which contain the stem of the existing verbs in Japanese) are: ケルラ ([keˈrura]), ナグラ ([naˈgura]) and サスラ ([saˈsura]), which refer to 'kicking', 'punching' and 'pointing' respectively. The (trisyllabic) nonce nouns are: ソギナ ([sogiˈna]), ダクメ ([dakuˈme]), ネチビ ([neˈʧibi]), and タソヌ ([tasoˈnu]), naming four characters (a burglar, a chef, a cowboy, and a waitress). The two nonce markers (one for number, one for case) are randomly chosen from the following set: セヒ ([seˈhi]), ギト ([giˈto]), ヨザ ([yoˈza]). Word order in sentences was Verb-Agent-Patient. Half of the participants were assigned to each of two conditions as per Experiment 1 (i.e, pre-nominal or post-nominal morphology). Auditory stimuli were recorded in a sound-attenuated room by a 28yo female speaker of Japanese.

Procedure The experiment was conducted in a quiet room, with all instructions provided in Japanese, and a Japanese-speaking experimenter. Participants were told that they would be learning part of a foreign language. The session proceeded exactly as outlined for Experiment 1.

Results

The proportion of participants whose oral productions follow scope in 0-100% of trials are shown in Figure 2. The results from Experiment 2 are on the right-hand side. All participants produced number consistently (95-100%) closer to the noun than case. This was true in both the pre-nominal or post-nominal marker conditions. We ran a logistic mixed-effects

[3] In all models, fixed effects were sum coded unless stated otherwise, and random intercepts for both items (noun) and participants were included. The DV consists of a binary variable marking the presence and absence of scope-isomorphism in each oral production trial (1 for a scope-isomorphic pattern, 0 for an anti-scopal pattern).

Table 2: Model output comparing Experiment 1 and 2.

	β	*SE*	*z*	$Pr(>\|z\|)$
(Intercept)	12.112	1.966	6.160	< 0.001
Marker Position	−0.0295	1.302	−0.227	0.821
Experiment	−0.012	1.303	−0.009	0.993
Marker Position × Experiment	0.05	1.302	0.038	0.970

model predicting scope-isomorphic productions by Marker Position (pre- vs. post-) and Experiment (Japanese vs. English). As shown in Table 2, the intercept is positive and significant, confirming that the proportion of scope-isomorphic productions is above chance. The non-significant effects of Marker Position and Experiment confirm that this preference holds regardless of pre- or post-nominal positioning of the markers, and regardless of the native language of participants.

Experiment 3

Experiments 1 and 2 demonstrate that learners have a natural preference to produce number morphology closer to the noun stem than case. These results hold for pre- and post-nominal orders, suggesting that the preference is not driven by linear order: number appears before case in post-nominal orders, but after case in pre-nominal orders. Our results hold for speakers of both English and Japanese, suggesting that they are not driven by L1 knowledge: familiarity with a particular morpheme (number or case respectively) does not mean it is placed closer to the stem. Frequency cannot explain the preference either: markers for case and number occur with equal frequency, as does each stem+morpheme bigram. The parsability of the morphemes is also the same, since frequencies of stem+morpheme forms relative to stems alone is the same for each. We thus conclude that the results obtained so far are consistent with a bias towards scope-isomorphism.

While our results suggest the bias is very strong (almost all participants uniformly preferred scope-isomorphic orders), in natural language, competing pressures may be present. One such pressure, prominent in models of morphological learning comes from the notion of locality. Dependencies between morphemes (e.g., between an allomorph and the stem that triggers it) tend to be local, or adjacent (Embick, 2010; Moskal, 2015; Bobaljik, 2012). In Experiment 3, we test the strength of the scope-isomorphic bias in the face of a competing locality bias. To do this, we use contextual allomorphy: the form of the case marker is dependent on the lexical and phonological identity of the noun. Because this creates a dependency between the noun stem and the case marker, a locality bias would predict that these two elements should be adjacent. The effect of the scope-isomorphism bias uncovered in Experiments 1 and 2 may override the effect of a locality bias. Alternatively, the locality bias may interfere with the placement of number in closer proximity to the noun stem, leading to a higher proportion of anti-scopal order productions (typologically rare) in the presence of stem-dependent case allomorphy.

Methods

Participants Forty-four English speakers were recruited and compensated as for Experiment 1. They were evenly divided between four conditions, as described below. Following our exclusion criteria, the data of four participants were excluded from analysis.

Input languages

This was a 2x2 design, with Marker position (pre- and post-) and Allomorphy (no allomorphy vs. case allomorphy) varying between-subjects. The input language in no allomorphy conditions was as in Experiment 1, except that case and number markers appeared as bound morphemes (i.e., affixes) on the noun when presented in text form (no spaces). The input language in the case allomorphy conditions differed additionally in having *two* accusative case markers, which alternated based on the length of the noun: one marker appeared with bisyllabic nouns ('negid', 'tumbat'), the other with monosyllabic nouns ('vaem', 'nork').

Procedure The procedure was identical to Experiment 1, except that in two-marker written trials, participants could not advance to the next trial until they typed the correct number of characters. This encouraged participants to produce both two markers together.

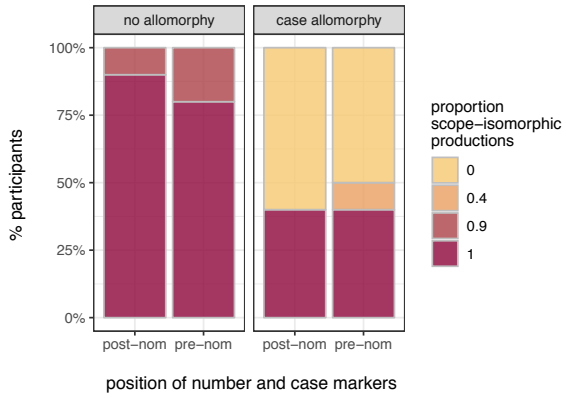


Figure 3: Percentage of participants in Experiment 3 who produced scope-isomorphic responses a given proportion of the time, ranging from 0% of the time (yellow) to 100% of the time (dark red). Results are split by Marker Position (pre- vs. post-nominal) and Allomorphy (no allomorphy vs. case allomorphy).

Results

Figure 3 shows the percentage of participants whose oral productions follow scope in 0-100% of trials across all four conditions. For the no allomorphy conditions, we replicate our previous findings: participants strongly prefer the scope-isomorphic order, with the number marker closer to the noun than case. By contrast, in the case allomorphy conditions,

Table 3: Model output for Experiment 3

	β	*SE*	*z*	$Pr(>\|z\|)$
(Intercept)	15.148	4.557	3.324	< 0.001
Marker Position	0.381	4.355	0.087	0.930
Allomorphy	−27.506	5.386	−5.107	< 0.001
Marker Position × Allomorphy	−0.529	4.691	−0.113	0.910

this pattern is reversed, with most participants producing case closer to the noun. This was confirmed by a logistic mixed-effects regression model predicting use of scope-isomorphic order by Marker Position, and Allomorphy[4]. As shown in Table 3, there is a significant drop in the use of scope-isomorphic orders in the case allomorphy condition.

Discussion

In the experiments reported here, speakers are trained on a language with distinct number and case morphemes, but the relative order of those morphemes is held out. When required to produce both morphemes together during testing, we found that participants' default inference is to place number closer to the noun stem than case (regardless of whether the markers were pre- or post-nominal). This bias provides a potential causal link between human cognition, and a typological generalisation known as Universal 39 (Greenberg, 1963). Importantly, we found strong evidence for this bias across two populations which differ in terms of their prior experience with case and number markers; English marks number but not case, while Japanese marks case but not number. This suggests our results cannot be explained by relative familiarity with these markers. Furthermore, the observed preference is not dependent on distributional information in the input: case and number markers never appear together, and have the same frequency during training. We have suggested that this bias is driven by scope relations among the markers. In particular, case (which marks the grammatical role of the noun in the event) scopes higher than number (which modifies the set properties of the entity), and linear proximity should reflect scope (Bybee, 1985; Rice, 2000; Culbertson & Adger, 2014). While this order is inferred by default, results from Experiment 3 revealed that the presence of stem-dependent contextual allomorphy for case led many participants to place the case morpheme closer to the conditioning noun. This suggests that the default preference may interact with other constraints—i.e., imposed by morphophonological rather than semantic dependency relationships—as predicted by theories of locality (e.g., White et al., 2018; Embick, 2010). Whether such allomorphy patterns are sensitive to locality in natural language points to the need for additional typological research (although see Moskal, 2015).

[4]The fixed effect of Allomorphy was treatment coded (instead of sum coded) so we could directly compare case allomorphy to the baseline no allomorphy.

Conclusion

Our results show that in the absence of explicit evidence, language learners default to a typologically common order of morphemes: with number more proximal to the noun stem than case. This supports a hypothesised link between human cognition and Greenberg's Universal 39. However, this observed bias in principle interacts with constraints on locality driven by morphophonological dependencies.

Data accessibility

The data that support the findings of this study are openly available in the Open Science Foundation repository at https://doi.org/10.17605/OSF.IO/9FA3V.

Acknowledgements

We thank Maki Kubota and Alexander Martin for helping in the recording of the experimental stimuli.

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 757643, held by JC).

References

Baayen, H. (1993). On frequency, transparency and productivity. In *Yearbook of morphology 1992* (pp. 181–208). Springer.

Baker, M. (1985). The Mirror Principle and Morphosyntactic Explanation. *Linguistic Inquiry*, *16*(3), 373–415. Retrieved from www.jstor.org/stable/4178442

Bobaljik, J. D. (2012). *Universals in comparative morphology: Suppletion, superlatives, and the structure of words* (Vol. 50). Cambridge, MA: MIT Press.

Bouchard, D. (2002). *Adjectives, number and interfaces: Why languages vary*. Amsterdam: Elsevier. doi: 10.1353/lan.2006.0187

Bybee, J. L. (1985). *Morphology: A study of the relation between meaning and form* (Vol. 9). Amsterdam: John Benjamins Publishing.

Cinque, G. (2005). Deriving Greenberg's Universal 20 and its exceptions. *Linguistic Inquiry*, *36*(3), 315-332. doi: 10.1162/0024389054396917

Culbertson, J., & Adger, D. (2014). Language learners privilege structured meaning over surface frequency. *Proceedings of the National Academy of Sciences*, *111*(16), 5842–5847. doi: 10.1073/pnas.1320525111

Culbertson, J., Smolensky, P., & Legendre, G. (2012). Learning biases predict a word order universal. *Cognition*, *122*(3), 306–329. doi: 10.1016/j.cognition.2011.10.017

Embick, D. (2010). *Localism versus globalism in morphology and phonology*. Cambridge, MA: MIT Press.

Fedzechkina, M., Chu, B., & Jaeger, T. F. (2018). Human information processing shapes language change. *Psychological science*, *29*(1), 72–82. doi: 10.1177/0956797617728726

Fedzechkina, M., Jaeger, T. F., & Newport, E. L. (2012). Language learners restructure their input to facilitate efficient communication. *Proceedings of the National Academy of Sciences*, *109*(44), 17897–17902. doi: 10.1073/pnas.1215776109

Futrell, R., Mahowald, K., & Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, *112*(33), 10336–10341. doi: 10.1073/pnas.1502134112

Givón, T. (1979). *On understanding grammar*. New York: Academic Press.

Greenberg, J. H. (1963). Some universals of grammar with particular reference to the order of meaningful elements. In J. H. Greenberg (Ed.), *Universals of human language* (pp. 73–113). Cambridge, MA: MIT press.

Grimshaw, J. (1986). A morphosyntactic explanation for the mirror principle. *Linguistic Inquiry*, *17*(4), 745–749. Retrieved from `http://www.jstor.org/stable/4178514`

Hay, J. (2001). Lexical frequency in morphology: Is everything relative? *Linguistics*, *39*(6), 1041–1070. doi: 10.1515/ling.2001.041

Hay, J., & Plag, I. (2004). What constrains possible suffix combinations? on the interaction of grammatical and processing restrictions in derivational morphology. *Natural Language & Linguistic Theory*, *22*(3), 565–596. doi: 10.1023/B:NALA.0000027679.63308.89

Hupp, J. M., Sloutsky, V. M., & Culicover, P. W. (2009). Evidence for a domain-general mechanism underlying the suffixation preference in language. *Language and Cognitive Processes*, *24*(6), 876–909. doi: 10.1080/01690960902719267

Manova, S., & Aronoff, M. (2010). Modeling affix order. *Morphology*, *20*(1), 109–131. doi: 10.1007/s11525-010-9153-6

Moskal, B. (2015). Limits on allomorphy: A case study in nominal suppletion. *Linguistic Inquiry*, *46*(2), 363-376. doi: 10.1162/LING_a_00185

Nakanishi, K., & Tomioka, S. (2004). Japanese plurals are exceptional. *Journal of East Asian Linguistics*, *13*(2), 113–140. doi: 10.1023/B:JEAL.0000019058.46668.c1

Plag, I., & Baayen, H. (2009). Suffix ordering and morphological processing. *Language*, 109–152. doi: 10.1353/lan.0.0087

Rice, K. (2000). *Morpheme order and semantic scope: Word formation in the Athapaskan verb* (Vol. 90). Cambridge, UK: Cambridge University Press.

Rice, K. (2011). Principles of affix ordering: An overview. *Word Structure*, *4*(2), 169–200. doi: 10.3366/word.2011.0009

Ryan, K. (2010). Variable affix order: grammar and learning. *Language*, *86*(4), 758–791. doi: 10.1353/lan.2010.0032

Tabullo, Á., Arismendi, M., Wainselboim, A., Primero, G., Vernis, S., Segura, E., … Yorio, A. (2012). On the learnability of frequent and infrequent word orders: An artificial language learning study. *The Quarterly Journal of Experimental Psychology*, *65*(9), 1848–1863. doi: 10.1080/17470218.2012.677848

White, J., Kager, R., Linzen, T., Markopoulos, G., Martin, A., Nevins, A., … van de Vijver, R. (2018). Preference for locality is affected by the prefix/suffix asymmetry: Evidence from artificial language learning. In *Proceedings of NELS 48* (p. 207-220). Amherst, MA: Graduate Linguistics Student Association.

Wilson, C. (2003). Experimental investigation of phonological naturalness. In *Proceedings of WCCFL 22* (Vol. 22, pp. 533–546).

Wunderlich, D. (1993). Funktionale kategorien im lexikon. In F. Beckmann & G. Heyer (Eds.), *Theorie und praxis des lexikons* (pp. 54–73). Berlin: Walter de Gruyter.

Cumulative cultural evolution in a non-copying task in children and Guinea baboons

Carmen Saldana (carmen.saldana@ed.ac.uk)
Centre for Language Evolution, The University of Edinburgh,
3 Charles Street, Edinburgh, EH8 9AD, UK

Joël Fagot (joel.fagot@univ-amu.fr)
Aix Marseille Université, CNRS, LPC UMR 7290, 13331, Marseille, France
Brain and Language Research Institute, Aix-Marseille University, Aix-en-Provence, France

Simon Kirby (simon.kirby@ed.ac.uk)
Centre for Language Evolution, The University of Edinburgh,
3 Charles Street, Edinburgh, EH8 9AD, UK

Kenny Smith (kenny.smith@ed.ac.uk
Centre for Language Evolution, The University of Edinburgh,
3 Charles Street, Edinburgh, EH8 9AD, UK

Nicolas Claidière (nicolas.claidiere@normalesup.org)
Aix Marseille Université, CNRS, LPC UMR 7290, 13331, Marseille, France
Brain and Language Research Institute, Aix-Marseille University, Aix-en-Provence, France

Abstract

The unique cumulative nature of human culture has often been explained by high-fidelity copying mechanisms found only in human social learning. However, transmission chain experiments in human and non-human primates suggest that cumulative cultural evolution (CCE) might not be dependent on high-fidelity copying after all. In this study we test whether CCE is possible even with a *non-copying* task. We performed transmission chain experiments in Guinea baboons and children where individuals observed and reproduced visual patterns on touch screen devices. In order to be rewarded, participants had to avoid touching squares that were touched by a previous participant. In other words, they were regarded for innovation rather than copying. Results nevertheless exhibited two fundamental properties of CCE: an increase over generations in task performance and the emergence of systematic structure. However, CCE arose from different mechanisms across species: children, unlike baboons, converged in behaviour over generations by copying specific patterns in a different location, thus introducing alternative copying mechanisms into the non-copying task. We conclude that CCE can result from non-copying tasks and that there is a broad spectrum of possible mechanisms that will lead to CCE aside from high-fidelity transmission.

Keywords: social learning; transmission chain; copying; cumulative cultural evolution; Guinea baboons; children;

Introduction

Human culture evolves over time with the gradual accumulation of modifications, from social norms (Nichols, 2002), to art (Morin, 2013), to language (Keller, 2005). In contrast, evidence for cumulative culture has been extremely difficult to find in other animal species (but see, e.g., Grant & Grant, 2010; Garland et al., 2011), and even difficult to induce through experimental manipulations (but see, e.g., Sasaki & Biro, 2017; Fehér, Wang, Saar, Mitra, & Tchernichovski, 2009). It has been proposed that this sharp contrast between human and non-human animal cultures can be explained by the lack of copying fidelity in the social learning of non-human animals (Tomasello, Kruger, & Ratner, 1993; Kempe, Lycett, & Mesoudi, 2014; Lewis & Laland, 2012). Faithful transmission can prevent the loss of cultural modifications and consequently result in cultural accumulation (Tomasello et al., 1993); therefore, the ability to faithfully transmit information through high-fidelity social learning has been taken as a requirement for cumulative culture. However, it is unclear whether there is a critical level of fidelity required to observe cumulative cultural evolution (CCE) and whether that required level of fidelity can ever actually be achieved by social learning mechanisms (Claidière & Sperber, 2009).

Transmission chain experiments have further shown that CCE can occur with learning mechanisms that exist in non-human animals, suggesting that cumulative culture is not after all dependent on special cognitive capacities found only in humans (Caldwell & Millen, 2008; Claidière, Smith, Kirby, & Fagot, 2014; Zwirner & Thornton, 2015). Claidière et al. (2014) for instance, performed a transmission chain study in which baboons observed and reproduced visual patterns on touch screen computers. Transmission led to the emergence of cumulative culture, as indicated by fundamental aspects of human cultural evolution such as (i) a progressive increase in performance and (ii) the emergence of systematic structure. Surprisingly, these results were achieved with an extremely low fidelity of pattern reproduction during the first generations of transmission, suggesting that high-fidelity copying may not always be the cause of cumulative culture and may in fact itself be a product of CCE. Individuals may transform input variants in accordance to their prior biases, and if those biases are shared at the population level, we expect transfor-

mations in the same direction to accumulate at each transmission step. Claidière et al. (2014)'s study therefore shows that cultural transmission may give a misleading impression of high-fidelity transmission when in fact cultural evolution tends to produce variants that become more faithfully transmitted. Similar results have been found in transmission experiments with human participants, for example where the transmission of miniature languages results in the emergence of languages which can be easily learned, even if the initial languages in each chain of transmission are transmitted only with very low fidelity (e.g. Beckner, Pierrehumbert, & Hay, 2017; Kirby, Cornish, & Smith, 2008).

Can we observe CCE in a non-copying task?

Most experiments on social learning and cultural transmission focus on copying tasks in which the individuals goal is to reproduce the input behaviour (for a thorough review, see Mesoudi & Whiten, 2008). However, other mechanisms through which humans and other animals learn, use and transmit information remain under-explored. Encouraged by the results of Claidière et al. (2014) showing that CCE can also result from initially low transmission fidelity, we decided to test whether CCE could occur in a transmission task that did not require direct copying. If high-fidelity copying is essential to CCE, we might not observe it in a task that does not involve copying.

To test this hypothesis, we performed an experiment with baboons and human children using the same protocol as in Claidière et al. (2014) but with a "non-copying" task in which the individuals were trained to avoid directly reproducing the patterns touched by a previous individual. In Claidière et al. (2014)'s original task participants were presented with a grid of 16 squares, four of which were briefly highlighted, and the task was to touch the squares that had been highlighted. The squares touched by one individual then became the highlighted squares for the next individual in the chain of transmission. In our new version of the task, the highlighted squares were instead to be avoided; the squares that one individual touched were the ones the next individual needed to avoid in order to be rewarded.

There are grounds for expecting this change in the pay-off structure of the task would prevent CCE from happening. In every trial, 495 different possible responses lead to a reward (a 27% likelihood of being correct by chance), creating a vast space of "correct" responses in every generation that are all different from the previous individual's response but which all will be rewarded. As well as directly penalising copying behaviour, the fact that the space of possible correct responses is so large and rather unconstrained by the input pattern suggests that any early accumulation of modifications (e.g., incipient structure in the system of patterns produced) could easily be wiped out by any individual in a chain of transmission, preventing cultural accumulation. However, participants could use non-copying alternative strategies that would result in convergent behaviour over generations. For instance, if participants try to minimise the effort of retaining and/or producing non-overlapping patterns, this might progressively cluster the responses on patterns of four connected squares (i.e., tetrominoes, which are easier to retain in memory and produce). This in turn might lead to increased performance over generations because such structured input patterns will be easier to avoid. Thus, if the search in the large evolutionary space is biased and there are alternative strategies which will lead to convergent behavioural output over generations, it might be possible to observe cumulative effects in transmission chains with a non-copying task.

Methods

Guinea baboons

Participants and testing facility Twelve Guinea baboons (Papio papio) belonging to a large social group of 25 from the CNRS Primate Center in Rousset-sur-Arc (France) participated in this study. They were 6 males (median age 8 years, min = 5, max = 11) and 6 females (median age 8 years, min = 5, max = 12).

The study was conducted in a facility developed by J.F. (for further information, see Fagot, Gullstrand, Kemp, Defilles, & Mekaouche, 2014).The baboons live in an outdoor enclosure ($700m^2$) connected to an indoor area which provides shelter when necessary. The outside enclosure is connected to 10 testing booths each equipped with a touchscreen. The key feature of this facility is that baboons have free access to computerised testing booths that are installed in trailers next to their enclosure. Identification of the subjects within each test booth is made possible thanks to two biocompatible 1.2 by 0.2 cm RFID microchips injected into each baboon's forearm. The baboon can thus participate in an experiment whenever they choose, and do not need to be captured to participate. The test program allows an independent test regime for each baboon, irrespective of the test booth it is using. Grains of dry wheat are used as reward. Baboons were neither water- nor food-deprived during the research. Water was provided *ad libitum* within the enclosure. Baboons received their normal ratio of food (fruits, vegetables and monkey chow) every day in the afternoon. The baboons were all born within the primate centre.

This research was carried out in accordance with French and EU standards and received approval from the French Ministère de l'Education Nationale et de la Recherche (approval # APAFIS-2717-2015111708173794-V3). Procedures were also consistent with the guidelines of the Association for the Study of Animal Behaviour.

Computer-based task Each trial began with the display of a grid made of 16 squares, 12 white and four green. Touching this stimulus triggered the immediate abortion of the trial and the display of a green screen for 3 s (time-out). After 400 ms all the green squares became white and, in order to obtain a food reward, the baboon had to select and touch four squares in this matrix which were not previously shown in green colour. Touching these four square could be done in

any order. Squares became black when touched to avoid being touched again and did not respond to subsequent touches. A trial was completed when four different squares had been touched with less than 5 s between touches. If four correct squares were touched, then the trial was considered a success and the computer triggered the delivery of 3-4 wheat grains. If less than four correct squares were touched (i.e. at least one previously green square was touched), then the trial was considered a failure and a green time out screen appeared for 3 s. The stimuli consisted of 80x80 pixel squares (white or green) equally spaced on a 600×600 pixel grid and were displayed on a black background on a 1024×768 pixels screen. The inter-trial interval was at least 3 s, but could be much longer since baboons chose when to initiate a trial.

Training to criterion Twenty-five members of the colony underwent a training procedure to enable them to participate in the transmission chain experiment: only those animals who reached our final criterion (N=12) were admitted to the transmission chain study described below. Training followed a progressive increase in the complexity of the task, starting with one white square and one green square, followed by a stage with an increasing number of white squares (up to six, one green and 2-5 white), then by a progressively increasing number of white and green squares up to 12. Training blocks consisted of 50 non-aborted trials (aborted trials were immediately re-presented, and the abortion rate was very low: mean = 2.2%, min = 0.23% and max = 4.6% for the 25 baboons included in the training). Progress through training was conditioned on performing above a criterion of 80% success on a block of 50 random trials (excluding aborted trials).

Transmission procedure We followed the transmission procedure described in (Claidière et al., 2014) and therefore only report the main elements here. Testing began when all 12 baboons reached the learning criterion with four targets (green squares) and 12 distractors (white squares) randomly placed on the grid. For each transmission chain, a first baboon was randomly selected, and this subject received a first block of 50 transmission trials consisting of randomly-generated patterns. Once the first subject had been tested, its behavioural output (the actual pattern of squares touched) on these 50 transmission trials became the set of target patterns (randomly reordered) shown to the next individual in that chain.

When the individuals were not involved in the transmission chain, they could perform random trials that were generated automatically by the computer and were not part of the transmission process. We ran nine such chains each with 10 generations (i.e., 10 individuals in each chain), each initialised with a different set of randomly-generated trials. We also made sure that each baboon did not appear more than once in each chain and performed at least 500 random trials between sets of transmission trials to avoid interference between chains.

Children

The experimental procedure for children was as similar as possible to the experimental procedure for baboons; in this section we detail the differences.

Participants and materials Participants were 90 English speaking children between the ages of 5 and 7 years old (42 female, mean age = 6 yo), recruited at the Edinburgh Zoo's Budongo Trail. Four further participants were excluded from the study because they failed the pre-established criterion to achieve at least 2/3 successful trials during training. The experiment was carried out in accordance with the research ethics procedures of the Edinburgh Zoo's Bundongo Trail and of the department of Linguistics and English Language at The University of Edinburgh (Ref # 325-1718).

The experiment was conducted on iPads using the iOS application Pythonista 3, in a single session of approximately three minutes. The experiment took place in the hall of the the Edinburgh Zoo's Budongo Trail, with the child seated on a chair and the experimenter beside them throughout, providing all instructions verbally. The experimenter also provided encouragement to the child but no informative feedback during critical trials. All participants were rewarded with stickers at the end of the experiment.

iPad-based task The experiment was divided into two phases, a training phase and a testing phase. The training phase followed a progressive increase in the complexity of the task over three blocks, starting with a grid of two squares (one white, one red)[1], then a grid of four (two red, two white) followed by the final grid of 16 (four red, 12 white). Training blocks consisted of three trials each. We excluded participants who failed to produce a minimum of two successful trials during the last two training blocks (grids of four and 16). During testing, each trial (20 total) began with the display of a grid made of 16 squares as in the baboons' version, 12 white and four red. If four correct squares (any four of those which were not displayed in red) were touched the trial was considered a success and the smiley face of a monkey emoji was displayed along with a reward sound effect. Otherwise, the face of the monkey emoji was displayed with both hands covering the mouth along with a child-friendly incorrect answer sound effect. After the monkey emoji faded away, the screen remained black for 1 s before the next trial began. At the end of the experiment, irrespective of the participants performance, the display filled with animated stars while a reward melody was played.

Transmission procedure

The transmission procedure was exactly as described for the baboons' version, with the only difference being the size of the testing/transmission set, which was 20 trials in the child version instead of 50. We ran nine transmission chains with

[1]We decided to change the colour of the squares in the input patterns to follow the western colour conventions in which red is associated with prohibition.

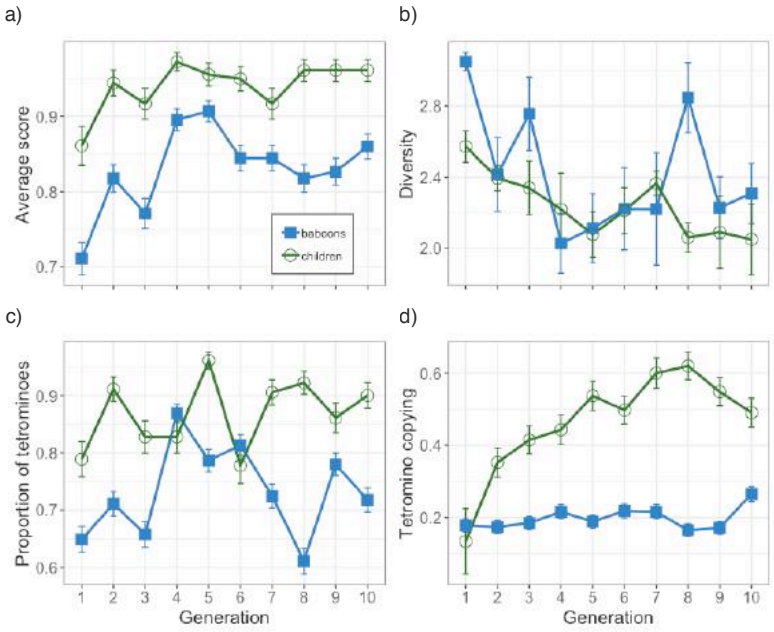


Figure 1: Blue squares and green circles illustrate the results for baboons and children respectively. (a) Average score, defined by the proportion of successful trials. (b) Average Shannon's diversity index within the set of responses. (c) Average proportion of tetrominoes produced. (d) Average proportion of observed tetrominoes which are copied.

a total of 10 generations (i.e., 10 children); each chain was initialised with a different set of randomly-generated trials.

Statistical analyses

The aim of our analyses is to evaluate the strength of the evidence for cumulative culture in baboons and children considering the two criteria highlighted in the introduction, that is, to test (i) a progressive increase in task performance over generations, and (ii) the emergence of systematic structure.

To analyse the results we used mixed-effects Growth Curve Analysis (GCA); the type of model, logistic or linear, will vary according to the dependent variable. All models contain three fixed effects: Generation, a quadratic polynomial for Generation (Generation^2), and Experiment (baboons as the baseline, and children). They also contain two interaction terms, between Generation and Experiment, and between Generation^2 and Experiment. To control for the non-independence within a given transmission chain, all models also contain random intercepts for Chain as well as by-Chain random slopes for the effects of Generation and Generation^2.

Results

Increase in task performance The average score was high across children and baboons, and we found a progressive increase in performance over generations of transmission across children and baboons (see Figure 1a). Using a dependent binary variable (success or failure for each trial) to analyse the evolution of success over generations, the results of the logistic GCA model show a significant effect of Generation ($\beta = 0.307, SE = 0.054, z = 5.700, p < 0.001$) and no significant interaction with Experiment ($z = 0.006, p = 0.995$), suggesting that task performance increases over generations of participants across children and baboons. We also found a significant effect of Generation^2 ($\beta = -0.027, SE = 0.005, z = -4.911, p < 0.001$) and no significant interaction with Experiment ($z = 0.351, p = 0.726$), suggesting that the increase in performance abates as we move along generations of participants. There was a further significant effect of Experiment ($\beta = 1.043, SE = 0.257, z = 4.062, p < 0.001$), suggesting that children generally scored higher in the task than baboons.

Emergence of systematic structure One indicator of the emergence of structure is a progressive decrease in response diversity due to a focus on a subset of responses (Kirby et al., 2008). We observed a reduction of diversity among sets of grids during transmission (see Figure 1b). Using the same model structure as previously specified, and the Shannon's diversity index of the systems of responses (equal to Shannon entropy: Shannon, 1948) as the dependent variable, a linear GCA model reveals a significant effect of Generation ($\beta = -0.237, SE = 0.069, t = -3.535, p < 0.001$) and no significant interaction with Experiment ($t = 1.467, p = 0.144$), suggesting that the diversity of the systems decreased over generations of participants across children and baboons. We also found a significant effect of Generation^2 ($\beta = 0.022, SE = 0.007, t = 2.982, p = 0.003$) and no significant interaction with Experiment ($t = -1.597, p = 0.112$), suggesting that the decrease in diversity deflates as we move along generations of

participants (across species). Moreover, the marginal effect of Experiment ($\beta = -0.356, SE = 0.19, t = -1.853, p = 0.068$) does not provide strong evidence to suggest a difference between children and baboons.

To explore the type of structures that emerged during transmission which might guide the observed decrease in diversity, we looked at the main structures found in Claidière et al. (2014), that is, tetrominoes (grids where all four squares are connected—lines, squares, L-shapes, T-shapes, S-shapes; tetrominoes will be familiar to anyone who has played Tetris). Figure 1c shows the proportion of tetrominoes produced over generations. The results from a logistic GCA model with a binary dependent variable representing the presence or absence of a tetromino suggest that children and baboons have a significant tendency to produce tetrominoes ($\beta = 0.688, SE = 0.225, z = 3.058, p = 0.002$) and that children produced them significantly more than baboons ($\beta = 1.046, SE = 0.353, z = 2.961, p = 0.003$). We also found a weak effect of Generation ($\beta = 0.252, SE = 0.112, z = 2.263, p = 0.024$) and no effect of its interaction with Experiment ($\beta = -0.214, SE = 0.177, z = -1.206, p = 0.228$), suggesting that the proportion of tetrominoes produced slightly increase with generation in baboons as well as in children. However, the significant effect of Generation^2 ($\beta = -0.026, SE = 0.012, z = -2.080, p = 0.038$) and the non-significance of its interaction with Experiment ($\beta = 0.030, SE = 0.020, z = 1.517, p = 0.129$) suggest that such increase in the production of tetrominoes reduces with generation.

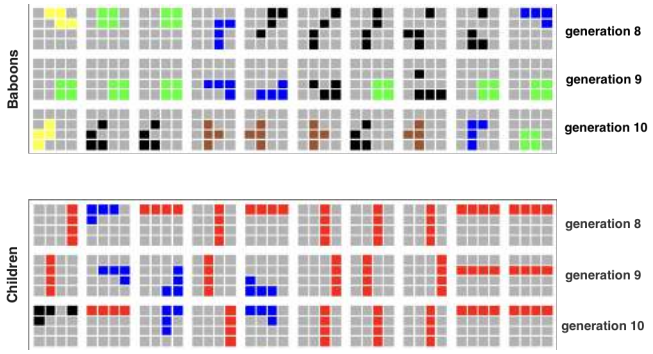


Figure 2: Baboons' and children's example responses, rows correspond to 10 example grids in generations 8-10 of a given chain (from top to bottom). Colouring of each grid reflects the tetromino class each pattern belongs to (green for squares, blue for L-shapes, brown for T-shapes, yellow for S-shapes, and black for non-tetrominoes).

Copying in a non-copying task So far, the general tendencies in the results found in children are very similar to those found in baboons—the only difference so far is that children score higher and produce more tetrominoes than baboons on average. However, an inspection of patterns produced (see e.g. Figure 2) suggested that children tended to copy the overall shape of the response of the previous individual (but shifted its position to avoid direct copying of the observed pattern). Figure 1d indicates that while baboons tend not to copy the overall shape of input tetrominoes in their responses, children seem to do so increasingly over generations. A logistic GCA model confirms that while baboons copy input tetrominoes significantly below chance ($\beta = -1.27, SE = 0.174, z = -7.336, p < 0.001$) constantly across generations (Generation, $z = -1.416, p = 0.157$; Generation^2, $z = 0.012, p = 0.100$) , children increasingly copy input tetrominoes over generations ($\beta = 0.559, SE = 0.121, z = 4.602, p < 0.001$) and more so initially than later on, where the increase abates (as indicated by the interaction between Generation^2 and Experiment, $\beta = -0.048, SE = 0.012, z = -4.620, p < 0.001$).

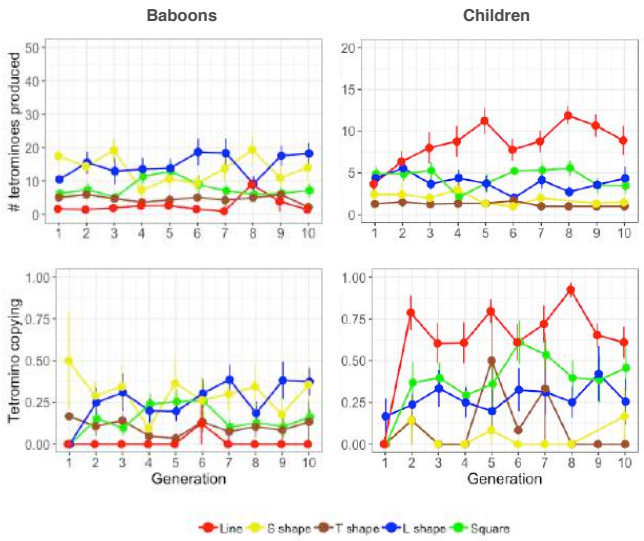


Figure 3: Top row: Average number of tetromino shapes produced by baboons (left) and children (right) for each of the five tetromino classes (over 20 and 50 trials respectively). Bottom row: Average proportion of tetrominoes that are copied from one generation to the next for baboons (left) and children (right).

We further explored the difference in copying in children and baboons by examining specific tetromino shapes because the inspection of the patterns produced (Figure 2) also suggested that children tended to produce many lines and that they copied them more so than any other pattern. Figure 3 shows the average number of tetrominoes produced as well as the proportion of tetromino copying subset by each of the five possible tetromino shapes. A visual inspection of Figure 3 reveals a clear preference for lines over other tetrominoes in children. Moreover, lines are the only pattern that shows an increase in production over time in children. We thus ran a logistic mixed-effects regression model (without the quadratic term, and with an added fixed effect for Tetromino Type with an interaction term) to test whether this observed increase in the production of lines over generations in children could be

accompanied by an increase in tetromino copying specific to lines. Results suggest that children copy lines significantly more than baboons ($\beta = 5.055, SE = 2.075, z = 2.436, p = 0.015$), who in contrast produce lines below chance ($\beta = -4.24, SE = 2.065, z = -2.052, p = 0.040$) equally across generations ($\beta = 0.001, SE = 0.356, z = 0.004, p = 0.997$). Results further suggest that lines are the most copied tetrominoes in children ($\beta = -0.048, SE = 0.012, z = -4.620, p < 0.001$; the smallest difference is shown with square tetrominoes: $\beta = -4.105, SE = 2.096, z = -1.958, p = 0.05$) but that this tendency to copy lines does not change over time ($\beta = -0.017, SE = 0.358, z = -0.047, p = 0.962$). We did not find a single significant interaction of Generation in the model (biggest effect: $z = 0.343, p = 0.732$). Altogether, these results suggest that children have a constant tendency to copy lines (above other tetrominoes), and once lines are introduced in the system, they are maintained. This in turn results in their accumulation and increase in number over time as new lines are introduced.

Discussion

The idea that faithful copying is essential to CCE is both intuitive and appealing: if socially-learned behaviours are not faithfully transmitted, modifications to what is being transmitted will not be passed on to other generations of individuals and will therefore be lost (Tomasello et al., 1993). In a process closely similar to biological replication, faithful copying could guarantee the transmission of modifications and therefore naturally lead to CCE. However, cultural evolution is much broader than biological evolution because it does not fundamentally derive from a process akin to replication and is therefore not constrained to certain modes of transmission (Claidière & André, 2012): several studies illustrate the fact that transmission can be of low fidelity and still lead to CCE (Caldwell & Millen, 2008; Claidière et al., 2014; Kirby et al., 2008; Claidière & Sperber, 2009).

The purpose of this study was to add to this research by examining the possibility of finding CCE with a non-copying task across human and non-human primate species. Results from children and baboons exhibited the two properties of CCE examined: (i) an increase in task performance linked to (ii) the emergence of some type of systematic structure. Despite the presence of a large evolutionary space (1820 possible responses for any single grid) and a very lenient reward function (27% chance of being correct by chance on any trial), we found the emergence of structure. This pattern probably emerged because the participants tended to cluster their responses in tetrominoes.

Although results from children and baboons were strikingly similar we found that, unlike the baboons, children introduced alternative copying mechanisms into the non-copying task by copying the shape of the input pattern in a different location, which was not prevented in the task (the non-copying task only forbid them from copying the exact grid pattern in the input, which included both the shape and location of the stimulus). This strategy adopted by children might in turn potentially explain (at least partially) their higher scores and tetromino production in comparison to baboons.

This observed copying strategy could be in line with children's tendency to high-fidelity copy even when not required in the task (Lyons, Young, & Keil, 2007; Whiten, McGuigan, Marshall-Pescini, & Hopper, 2009). Complementarily, it could also be partly explained by the fact that children, unlike baboons, only saw grids of two and four squares during training before the target grid of 16, and in these grids, the rewarded output is necessarily the mirror image of the input. However, we only observe high-fidelity copying of specific shapes (i.e., tetrominoes), which are potentially already preferred by children. Once these preferred shapes are in the system, they are maintained. Results thus suggest that the observed bias is not a copying bias (at least uniquely), but a bias towards tetromino shapes (stronger than in baboons; on average, almost 80% of responses are tetrominoes in childrens first generations), which results in high-fidelity copying once these patterns are introduced. Further inspection of the results showed that children tended to produce many line tetrominoes as well as to copy them from the input (more so than any other pattern), altogether suggesting that the bias towards tetromino shapes could be particularly strong for line tetrominoes. This bias towards copying and producing lines could be cognitive or task-specific (i.e., lines could potentially be easier and faster to produce altogether or in the context of an iPad game where one finger instead of two is mostly used), or it could simply reflect that lines are particularly salient to children (e.g., because of drawing or colouring).

Conclusion

Our study demonstrates that CCE can be observed in a non-copying task in baboons and children. Results across species exhibited two crucial properties of CCE: (i) an increase in task performance over generations and (ii) the emergence of systematic structure. However, these seemingly similar properties of CCE across species arose from different mechanisms: children, unlike baboons, converged in behaviour across generations by copying specific patterns (i.e., tetrominoes, and in particular lines) in a different location thus introducing biased copying into what was set up as a non-copying task. Together, our results suggest that CCE does not necessarily depend on (at least unbiased) high-fidelity copying and that there is a broad spectrum of possible transmission mechanisms that will lead to CCE; these mechanisms that are not based solely, or even mainly, on high-fidelity copying remain to be further explored.

Data accessibility

The data that support the findings of this study are openly available in the Open Science Foundation repository at https://doi.org/10.17605/OSF.IO/ZA265.

Acknowledgements

The authors thank the staff at the Rousset-sur-Arc Primate Center (CNRS-UPS846, France) and the Edinburgh Zoo's Budongo Trail/Living Links (UK) for technical support, and Julie Gullstrand and Marieke Woensdregt for helping in data collection.

This work was funded by the Agence Nationale de la Recherche ANR-13-PDOC-0004 (ASCE), ANR-16-CONV-0002 (ILCB) and ANR-11-LABX-0036 (BLRI) as well as the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement 681942, held by KS).

References

Beckner, C., Pierrehumbert, J. B., & Hay, J. (2017). The emergence of linguistic structure in an online iterated learning task. *Journal of Language Evolution*, *2*(2), 160–176. doi: 10.1093/jole/lzx001

Caldwell, C. A., & Millen, A. E. (2008). Experimental models for testing hypotheses about cumulative cultural evolution. *Evolution and Human Behavior*, *29*(3), 165–171. doi: 10.1016/j.evolhumbehav.2007.12.001

Claidière, N., & André, J.-B. (2012). The transmission of genes and culture: A questionable analogy. *Evolutionary Biology*, *39*(1), 12–24. doi: 10.1007/s11692-011-9141-8

Claidière, N., Smith, K., Kirby, S., & Fagot, J. (2014). Cultural evolution of systematically structured behaviour in a non-human primate. *Proceedings of the Royal Society of London B: Biological Sciences*, *281*(1797), 20141541. doi: 10.1098/rspb.2014.1541

Claidière, N., & Sperber, D. (2009). Imitation explains the propagation, not the stability of animal culture. *Proceedings of the Royal Society B: Biological Sciences*, *277*(1681), 651–659. doi: 10.1098/rspb.2009.1615

Fagot, J., Gullstrand, J., Kemp, C., Defilles, C., & Mekaouche, M. (2014). Effects of freely accessible computerized test systems on the spontaneous behaviors and stress level of guinea baboons (papio papio). *American Journal of Primatology*, *76*(1), 56–64. doi: 10.1002/ajp.22193

Fehér, O., Wang, H., Saar, S., Mitra, P. P., & Tchernichovski, O. (2009). De novo establishment of wild-type song culture in the zebra finch. *Nature*, *459*(7246), 564. doi: 10.1038/nature07994

Garland, E. C., Goldizen, A. W., Rekdahl, M. L., Constantine, R., Garrigue, C., Hauser, N. D., ... Noad, M. J. (2011). Dynamic horizontal cultural transmission of humpback whale song at the ocean basin scale. *Current biology*, *21*(8), 687–691. doi: 10.1016/j.cub.2011.03.019

Grant, B. R., & Grant, P. R. (2010). Songs of darwin's finches diverge when a new species enters the community. *Proceedings of the National Academy of Sciences*, *107*(47), 20156–20163. doi: 10.1073/pnas.1015115107

Keller, R. (2005). *On language change: The invisible hand in language*. Routledge.

Kempe, M., Lycett, S. J., & Mesoudi, A. (2014). From cultural traditions to cumulative culture: parameterizing the differences between human and nonhuman culture. *Journal of theoretical biology*, *359*, 29–36. doi: 10.1016/j.jtbi.2014.05.046

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*. doi: 10.1073/pnas.0707835105

Lewis, H. M., & Laland, K. N. (2012). Transmission fidelity is the key to the build-up of cumulative culture. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *367*(1599), 2171–2180. doi: 10.1098/rstb.2012.0119

Lyons, D. E., Young, A. G., & Keil, F. C. (2007). The hidden structure of overimitation. *Proceedings of the National Academy of Sciences*, *104*(50), 19751–19756. doi: 10.1073/pnas.0704452104

Mesoudi, A., & Whiten, A. (2008). The multiple roles of cultural transmission experiments in understanding human cultural evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1509), 3489–3501. doi: 10.1098/rspb.2018.0712

Morin, O. (2013). How portraits turned their eyes upon us: visual preferences and demographic change in cultural evolution. *Evolution and Human Behavior*, *34*(3), 222–229. doi: 10.1016/j.evolhumbehav.2013.01.004

Nichols, S. (2002). On the genealogy of norms: A case for the role of emotion in cultural evolution. *Philosophy of Science*, *69*(2), 234–255. doi: 10.1086/341051

Sasaki, T., & Biro, D. (2017). Cumulative culture can emerge from collective intelligence in animal groups. *Nature communications*, *8*, 15049. doi: 10.1038/ncomms15049

Shannon, C. E. (1948). A mathematical theory of communication. *Bell system technical journal*, *27*(3), 379–423. doi: 0.1002/j.1538-7305.1948.tb01338.x

Tomasello, M., Kruger, A. C., & Ratner, H. H. (1993). Cultural learning. *Behavioral and brain sciences*, *16*(3), 495–511. doi: 10.1017/S0140525X0003123X

Whiten, A., McGuigan, N., Marshall-Pescini, S., & Hopper, L. M. (2009). Emulation, imitation, over-imitation and the scope of culture for child and chimpanzee. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1528), 2417–2428. doi: 10.1098/rstb.2009.0069

Zwirner, E., & Thornton, A. (2015). Cognitive requirements of cumulative culture: teaching is useful but not essential. *Scientific reports*, *5*, 16781. doi: 10.1038/srep16781

Bee-ing In the World: Phenomenology, Cognitive Science, and Interactivity in a Novel Insect-Tracking Task

Guilherme Sanches de Oliveira (gui.cogsci@gmail.com)
Department of Philosophy & Center for Cognition, Action and Perception, University of Cincinnati
2700 Campus Way, McMicken Hall, Cincinnati OH 45221

Christopher Riehm (riehmcd@mail.uc.edu)
Department of Psychology & Center for Cognition, Action and Perception, University of Cincinnati
45 W. Corry Blvd, Edwards 1, Cincinnati OH 45221

Colin T. Annand (annandct@mail.uc.edu)
Department of Psychology & Center for Cognition, Action and Perception, University of Cincinnati
45 W. Corry Blvd, Edwards 1, Cincinnati OH 45221

Abstract

Dotov, Nie and Chemero (2010) conducted a set of experiments to demonstrate how phenomenology, particularly the work of Martin Heidegger, interfaces with experimental research in embodied cognitive science. Specifically, they drew a parallel between Heidegger's notion of readiness-to-hand and the concept of an extended cognitive system (Clark 2008) by looking for the presence or absence of interaction-dominant dynamics (Holden, van Orden, and Turvey 2009; Ihlen and Vereijken 2010) in a hand/mouse system. We share Dotov, Nie and Chemero's optimism about the potential for cross-pollination between phenomenology and cognitive science, but we think that it can be better advanced through a shift in focus. First, we argue in favor of using Maurice Merleau-Ponty's phenomenological theory as the philosophical foundation for experimental research in embodied cognitive science. Second, we describe an audio-visual tracking task in virtual reality that we designed and used to empirically investigate human-environment coupling and interactivity. In addition to providing further support for phenomenologically-inspired empirical cognitive science, our research also offers a more generalizable scientific treatment of the interaction between humans and their environments.

Keywords: phenomenology; embodiment; interactivity; agent-environment systems

Introduction

Dotov, Nie and Chemero (2010) illustrated how insights from the philosophical tradition of phenomenology can contribute to experimental research in embodied cognitive science. In a set of experiments, they had participants play a computer game using the mouse cursor to herd a moving target to a designated area of the screen. In the middle of the experiment, the connection between cursor and mouse was briefly "broken," making the cursor move randomly on screen, independently of mouse movements of the participant, until, after a short period of time, normal operation was resumed. The authors recorded time series data of the mouse/hand position and subsequently submitted it to a detrended fluctuation analysis (Kantelhardt et al., 2001), which estimates a measure of temporal correlation within a time domain signal. They found that when the mouse malfunctioned, there was a shift in the fractal scaling of the mouse/hand movements which they took to correspond to the degeneration of interaction-dominant dynamics into component-dominant dynamics (Holden, van Orden, and Turvey 2009). Following Heideggerian phenomenology, they framed this as a transition from the mouse being *ready-to-hand* to being *present-at-hand* for the participant.

For Heidegger (1927), we perceive objects and tools pragmatically as "something in-order-to": for example, you experience the sheet of paper on your desk as *something to write on* and the pen as *something to write with*. In typical circumstances, these objects are "ready-to-hand" in that, while using them, you can focus on the end goal (writing a letter, say) without having to explicitly attend to the tools themselves. But if something goes wrong and the pen runs out of ink, for example, then the pen becomes "present-at-hand": i.e., it suddenly comes to the forefront of your attention, as something that needs to be confronted explicitly and directly before you can resume your work. It is in this sense that Dotov, Nie and Chemero characterize the mouse in their experiment as shifting from being "ready-to-hand" to "present-at-hand" when it becomes unresponsive.

We agree with Dotov, Nie and Chemero about the potential for cross-pollination between phenomenology and cognitive science. In this paper we explore how this interdisciplinary collaboration can be further promoted through a shift in focus. The object of our investigation is perception-action coupling and interactivity in agent-environment systems: as we propose, understanding how agents engage with features of their environment encompasses a broader range of cognitive phenomena that includes, but is not limited to, tool use. In what follows, we first present Maurice Merleau-Ponty's phenomenological theory as providing the philosophical foundation for this shift. Next, to illustrate what this shift looks like experimentally, we describe findings from a novel audio-visual tracking task in virtual reality that we created. We conclude by discussing how this approach offers a more widely-applicable perspective for phenomenologically-inspired empirical research in embodied cognitive science.

Merleau-Ponty, Embodiment and Interactivity

Maurice Merleau-Ponty (1945) introduces the term "bodily schema" to describe the "sensori-motor unity of the body" (p. 114). This unity entails, at once, the integration of each of our senses with one another and the integration of perception with action. Seeing and hearing are "pregnant with one another" and they work together as much as our two eyes complement one another. At the same time, seeing and hearing operate in conjunction with our legs and arms to produce walking and grasping: "my body is, not a collection of adjacent organs, but a synergic system, all the functions of which are exercised and linked together in the general action of being in the world" (p. 272).

Merleau-Ponty's famous example of the blind person navigating the environment with a cane or stick shows how this integrated bodily schema is *fluid* and can *change* over time. If you are adept at getting around using a stick, that is because you no longer perceive the stick itself but you perceive the world "at the end of the stick," which involves an expansion of your integrated sensorimotor bodily schema:

> "To get used to a hat, a car or a stick is to be transplanted into them, or conversely, to incorporate them into the bulk of our own body. Habit expresses our power of dilating our being-in-the-world, or changing our existence by appropriating fresh instruments" (1945, p. 166).

Considered by itself, Merleau-Ponty's example of the blind person's cane is compatible with Heidegger's ideas reviewed above: after all, the cane could be said to be ready-to-hand to the expert user in normal circumstances whereas it would become present-at-hand if it suddenly broke in half, just as it would also be initially present-at-hand to a sighted adult who was trying out the cane for the first time while blindfolded. Yet, Merleau-Ponty's understanding of the bodily schema is much broader than the blind man's cane example suggests and, for this reason, it is also better suited for informing empirical research in embodied cognitive science.

First, although the Heideggerian notions of readiness-to-hand and presence-at-hand help make sense of how we use tools (as in the case studied by Dotov, Nie and Chemero), it is not at all clear how this understanding generalizes to a broader range of cognitive phenomena, such as ordinary instances of perception and action that do *not* involve tool use. In contrast, Merleau-Ponty's richer notion of bodily schema is more versatile, applying to embodied experience no matter the degree of "dilation" and regardless of whether it involves the incorporation of tools. In a telling passage, Merleau-Ponty claims:

> "In the gaze we have at our disposal a natural instrument analogous to the blind man's stick. The gaze gets more or less from things according to the way in which it questions them, ranges over or dwells on them." (1945, p. 177).

As this quote suggests, Merleau-Ponty sees our body and our senses as being tool-like in their instrumental or functional character; yet the bodily schema explicitly applies primarily to our basic embodied activity and only secondarily to literal tool use (such as using a hammer or a mouse) as a particular type of bodily activity.

Second, besides applying to a broader range of cognitive phenomena, Merleau-Ponty's notion of bodily schema is also more theoretically attractive because of how it relates to different views in ongoing debates in cognitive science. Dotov, Nie and Chemero interpreted the ready-to-hand mouse as forming, with the body, an *extended cognitive system*. With this, they explicitly tied their account to the hypothesis that cognition may *sometimes* "leak out" of an individual and into parts of the world that the individual is interacting with (Clark 2008). The extended cognition hypothesis is contentious, to say the least: for many cognitive scientists, cognition just is the name of the processing that goes on within the individual's mind/brain; and for advocates of radical embodied cognitive science (e.g., Chemero 2009), the proper object of study just is the animal-environment system as a whole (Gibson 1979).

Merleau-Ponty's notion of bodily schema does not entail a commitment to the contentious hypothesis of extended cognition, and it thereby circumvents the controversy. In a key passage, Merleau-Ponty explains: "With the notion of the bodily schema we find that not only is the unity of the body described in a new way, but also, through this, the unity of the senses and of the object" (1945, p. 273). Above we saw that the bodily schema entails the sensorimotor unity of the body, that is, the integration of the senses and between perception and action. This quote adds, further, that the bodily schema entails also an integration between subject and the objects of experience. This captures an essential feature of the radical embodied and Gibsonian approaches to studying agent-environment systems, namely the focus on the complex interactivity between agent and environment: in this view, "patterns of an organism's behavior are best understood as the emergent property of the interactions of the organism with its environment" (Kelty-Stephen, Palatinus, Saltzman, and Dixon 2013, p. 2) and "perception and action are best understood in the broader context of the task and environment within which coordination of those biological nuts and bolts takes place" (p. 3). As such, we suggest, embodied agency or "being in the world" is always characterized by an integration of agent and environment through interaction. Interactivity may change qualitatively with changes in task and in the availability of task-relevant information, but it is always present: an agent's perception-action never becomes fully detached from her environment, and understanding this relation is independent of whether some internal feature of the agent "leaks out" into the world or not.

As an illustration of interactivity, imagine an ordinary situation such as trying to track a bumble bee so as to avoid being stung. Although you may initially catch sight of the bee and follow it with your gaze, the bee's erratic movement might cause it to disappear against a cluttered background.

Your desire to avoid being stung persists and you maintain an awareness of the bee's position by listening, trying to regain sight of it. You swivel your head, accommodating for the subtle shifts in interaural sound intensity, allowing your ears to guide your continued search for the bee. Furthermore, the bee may fly along your sagittal plane, momentarily escaping your efforts to track it by sound until, finally, you are able to regain auditory or visual tracking. This dance, between you and the bee, may persist until the bee exits your immediate surroundings. Although it may be true that, at times, the differences in mode and strength of your sensory coupling to the bee change, nothing is ever "broken." There may be differences in how your head or eyes move relative to the bee, but no aspect of this system can be said to transition from readiness-to-hand to presentness-at-hand. Furthermore, the system maintains interactivity throughout. Even though the specific dynamics of a particular aspect of the system may change, the system continues to be unified through the ongoing pursuit of the goals that are implicit to the task (e.g. avoiding a sting). The experiment described below was designed to capture this point.

Method

Undergraduate students (N = 10) at the University of Cincinnati participated in a virtual audio-visual tracking task for class credit. At the start of the experiment, each participant put on an Acer mixed reality headset and a pair of in-ear monitors (IEMs). The virtual scene that they were presented with (depicted in Figure 1) consisted of a white room and a semi-circular line spanning 180 degrees of the participant's visual field on which a black and yellow sphere (henceforth, "the bee") would travel over time, moving in a roughly brownian fashion similar to the moving target from Dotov, Nie and Chemero (2010).

Participants were instructed to track the bee throughout the task by continuously pointing their center of vision, indicated by a small sphere, to its location. They were told that the source would begin emitting a buzzing sound when the experiment started and that it would be necessary to use this sound to continue tracking the bee because the bee would shortly become invisible. The spatial information present in the sound of the bee was imparted by a set of generalized head-related transfer functions (Zhong and Xie 2014). Unbeknownst to the participants, after the bee had been invisible for 12 seconds, the sound spatialization would be removed, making it impossible for the participant to effectively track the bee. After a period of time, the sound spatialization would be added back and then, finally, the bee would reappear. In total, the task consisted of two 12 second periods of audio-visual tracking (at the beginning and at the end), two 12 second periods of audio-only tracking, and one 12 second period of tracking with no spatial information (in the middle). The order of this sequence is illustrated in Figure 2. During the entire experiment, the angular difference between the participant's center of vision and the position of the bee was recorded at 100 hz.

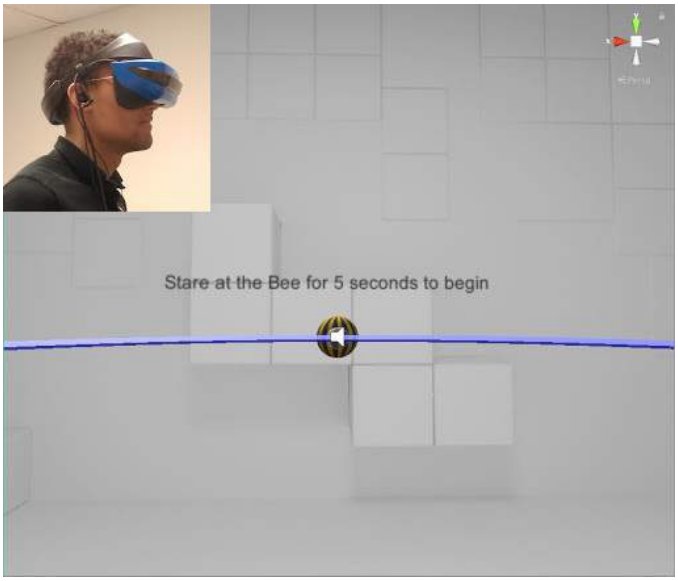


Figure 1: An image of what a participant would see upon starting the experiment, including the bee, the instructions for starting a trial and the line upon which the bee moved.

Fractal Analysis

We submitted the time series data from each trial and condition (Audio-Visual Information (AV), Audio Information Only (AO), and No Spatial Information (NI)) to a detrended fluctuation analysis (DFA) which allows for temporal correlations within a signal, at different scales, to be captured by a single value. Detrended Fluctuation Analysis (DFA) is a form of fractal analysis, which describes a power-law that captures the relationship between the size and occurrence rate of fluctuations for a given time series (Ihlen 2012). Fractal analyses have previously been used to illuminate the nature of embodied cognitive activity by examining continuous measures of agents embedded in environments (Kello, Beltz, Holden & Van Orden 2007).

The DFA measurement of the time series of angular error between gaze and target for each information condition (AV, AO and NI) yields a Hurst exponent and a closely related Alpha value, both of which describe the power-law relationship within the time series. In Dotov, Nie and Chemero (2010), Alpha values were calculated at repeated intervals to identify changes in tool-use behavior that were caused by the perturbation of mouse function. Here, we calculated Hurst exponents for each condition in order to index how gaze activity changes across the information conditions in the bee tracking task, as is visually exemplified in the time series data shown in Figure 2. Our DFA used a minimum window size of 2 samples and a maximum window size of roughly one third of each condition time series, which were each 1200 samples in length.

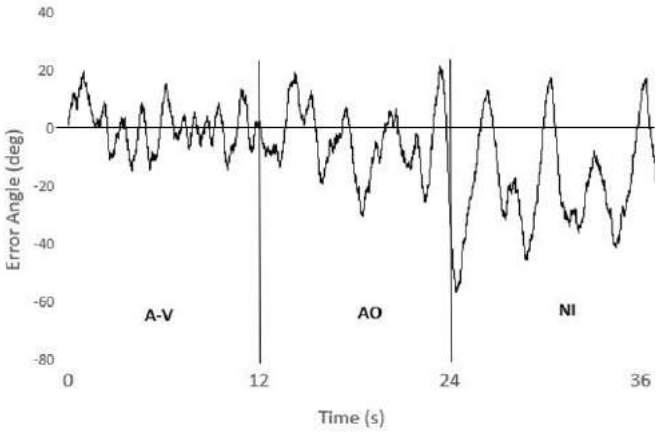


Figure 2: The timeseries above shows the first 36 seconds (x-axis) for the error angle (y-axis) between a participant's center of vision and the bee during the tracking task. The different information conditions are indicated by the lines, Audio and Visual information (AV), Audio information only (AO) and No spatial information (NI). In this particular case, the size and frequency of error clearly changes between the information conditions: error is minimized in the audio-visual condition (AV), increases in the auditory only condition (AO), and displays large, shifting values in the no spatial condition (NI) when the bee target auditory signal switches from stereo to mono.

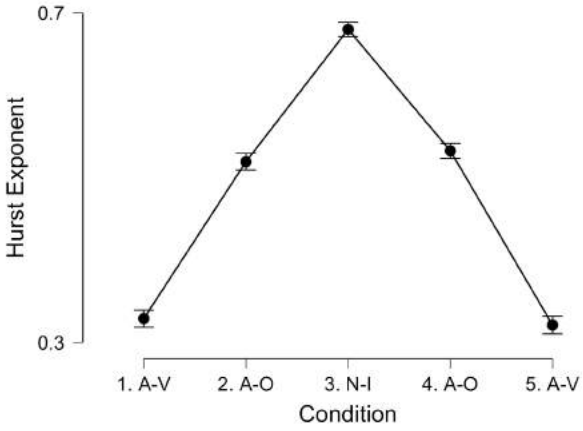


Figure 3: The x-axis indicates the changing conditions between audio-visual information (AV), auditory information only (AO), and no spatial information (NI). The y-axis indicates the value of the mean Hurst exponents and standard error bars for each condition.

Results

The Hurst exponents, calculated for each information condition and trial (AV, AO and NI) were submitted as dependent variables to a repeated measures analysis of variance to examine changes across trials as well as conditional differences.

Neither the within subject main effect of Trial, $F(4,180) = 0.125$, $p = 0.97$, or Trial by Condition interaction effect, $F(16,180)$, $p = 0.1$ were significant. The effect of Condition was significant, $F(4,45) = 86.38$, $p < .001$, $n^2 = 0.88$. It is worth noting that Tukey post-hoc analysis revealed significant differences between the condition types, but not their separate time occurrences: both AV conditions are not significantly different from one another, and both AO conditions are not significantly different from one another. Further details are provided in Table 1 and in Figure 3.

Table 1: Descriptive Statistics

	Mean Hurst Exponents by Condition				
	1. A-V	2. A-O	3. N-I	4. A-O	5. A-V
Mean	0.3290	0.5194	0.6800	0.5326	0.3212
Std. Dev.	0.0641	0.0406	0.0327	0.0459	0.0664

Discussion

Our experimental results reveal differences in the fractal scaling of movement across shifting task conditions. In this way, our results were similar to what was found by Dotov, Nie and Chemero (2010). The key difference between the two experiments lies in our focus. Their investigation is centered on the agent; ours follows radical embodied cognitive science (Chemero 2009) by being primarily concerned with the agent-environment system as a whole. This difference in focus informed both our choice of dependent measure and our interpretation of interactivity.

Because they were trying to find support for the extended cognition hypothesis, Dotov, Nie and Chemero (2010) measured raw hand movement at the tool/hand interface. This meshes well with their goal of demonstrating a shift, from the agent's perspective, between a tool being ready-to-hand to becoming present-at-hand—but this approach misses out on capturing the rest of the agent-environment system. In contrast, we adopted a collective measure at the task performance level. By measuring the error angle between the gaze and the bee's position, we were able to detect shifts in the overall agent-environment dynamics. In this context, specific Hurst exponent values are useful and explicate the nature of the system. For example, in the Audio-Visual Information condition, the low Hurst value indicates that the system corrects for increases in error similarly across timescales, exhibiting anti-persistent dynamics (Riley et al 2012). This makes sense because participants are likely very good at visually orienting to the position of objects. The higher Hurst values from the Auditory Only and No Information conditions show that there is a shift in how error is accommodated for at different scales. In the Auditory Only condition, for example, the

participant may be able only to accommodate for movements of the bee very slowly, but is ineffective at tracking its faster movements. This shift can be characterized as a shift towards persistent system dynamics (Riley et al 2012), which continues in the same direction as information is reduced further in the No Information condition.

A similar interpretation could have been applied to the herding task of Dotov, Nie and Chemero (2010) if the dependent measure had reflected the collective dynamics of the agent-environment system. In their case, it's not that when the tool breaks it is noticed as a tool, external to the system. Rather, the tool appears broken within the context of a task and is used as such. Movements exhibited by participants experiencing a broken mouse are sensible as movements meant to fix or disambiguate the nature of the brokenness of the mouse. Similarly, the movements of our participants who had no information about the bee's position are sensible as exploratory procedures (Riley et al 2002), i.e., movements meant to pick up information. These movements do not reflect a degeneration of interaction, but only a shift in the nature of the ongoing interaction between agent and environment. A participant in either task is never truly decoupled from the specific environment implied by the overarching task.

Dotov, Nie and Chemero characterize the distinction between interaction-dominant dynamics and component-dominant dynamics as follows: "In component-dominant dynamics, behavior is the product of a rigidly delineated architecture of modules, each with predetermined functions; in interaction-dominant dynamics, on the other hand, coordinated processes alter one another's dynamics, with complex interactions extending to the body's periphery and, sometimes, beyond" (2010, p. 3). This characterization works well with their agent-centered approach and their focus on cognition as an internal feature of the agent that can potentially extend out into the world. But when the object of study becomes the agent-environment system, as proposed in radical embodied cognitive science (Chemero 2009), this characterization fails. The dynamic variation in a proper collective measure of a complex agent-environment system will always be governed by the interaction between agent and environment. The system may be redefined across tasks, but can never become broken in the way that Dotov, Nie and Chemero would require. Because interactivity is a universal feature of agent-environment systems, rather than looking for signs of a shift from interaction-dominance to component-dominance, it is more appropriate to inquire into the specific nature of the interactivity. This means focusing on task specific coordination (Turvey, Saltzman and Schmidt 1991), rather than the dynamics that play out at the interface between human and tool.

As seen above, the choice of focus of investigation—whether centered on the agent or on the agent-environment system as a whole—is directly linked to the choice of dependent measure and to the interpretation of interactivity. The focus of investigation is also intimately associated to the phenomenological theory adopted in each case. Heidegger's theory is agent-centric and lends itself to application for investigating the dynamics of tool use and cognitive extension. Merleau-Ponty's theory, on the other hand, motivates thinking in terms of an integration between subject and object, or between agent and environment. This makes it more apt for making sense of a broader range of cognitive phenomena, beyond tool use, where interaction may occur. Merleau-Ponty's approach is thus better suited for conceptually framing research into perceptually driven human-environment interactivity in the ecological and embodied cognitive sciences.

References

Chemero, A. (2009). *Radical embodied cognitive science*. MIT press.

Clark, A. (2008). *Supersizing the mind: Embodiment, action, and cognitive extension*. Oxford University Press.

Dotov, D. G., Nie, L., & Chemero, A. (2010). A demonstration of the transition from ready-to-hand to unready-to-hand. *PLoS One*, *5*(3), e9433.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Houghton Mifflin.

Heidegger, M. (1927/2001). *Being and time* (J. Macquarrie & E. Robins, Eds.). Blackwell Publishers Ltd.

Holden, J. G., Van Orden, G. C., & Turvey, M. T. (2009). Dispersion of response times reveals cognitive dynamics. *Psychological review*, *116*(2), 318.

Ihlen, E. A., & Vereijken, B. (2010). Interaction-dominant dynamics in human cognition: Beyond $1/\alpha$ fluctuation. *Journal of Experimental Psychology: General*, *139*(3), 436.

Kantelhardt, J. W., Koscielny-Bunde, E., Rego, H. H., Havlin, S., & Bunde, A. (2001). Detecting long-range correlations with detrended fluctuation analysis. *Physica A: Statistical Mechanics and its Applications*, *295*(3-4), 441–454.

Kello, C. T., Beltz, B. C., Holden, J. G., & Van Orden, G. C. (2007). The emergent coordination of cognitive function. *Journal of Experimental Psychology: General*, *136*(4), 551568.

Kelty-Stephen, D. G., Palatinus, K., Saltzman, E., & Dixon, J. A. (2013). A tutorial on multifractality, cascades, and interactivity for empirical time series in ecological science. *Ecological Psychology*, *25*(1), 1-62.

Merleau-Ponty, M. (1945). *Phenomenology of perception* (. Translation by Kegan Paul, Ed.). Routledge.

Riley, M. A., Bonnette, S., Kuznetsov, N., Wallot, S., & Gao, J. (2012). A tutorial introduction to adaptive fractal analysis. *Frontiers in physiology*, *3*, 371.

Riley, M. A., Wagman, J. B., Santana, M. V., Carello, C., & Turvey, M. T. (2002). Perceptual behavior: Recurrence analysis of a haptic exploratory procedure. *Perception*.

Turvey, M. T., Saltzman, E., & Schmidt, R. C. (1991). Dynamics and task-specific coordinations. In *Making them*

move: Mechanics, control, and animation of articulated figures (p. 157-170).

Van Orden, G. C., Holden, J. G., & Turvey, M. T. (2005). Human cognition and 1/f scaling. *Journal of Experimental Psychology: General*, *134*(1).

Zhong, X.-l., & Xie, B.-s. (2014). Head-related transfer functions and virtual auditory display. In *Soundscape semiotics-localization and categorization.* IntechOpen.

Sources of knowledge in children's acquisition of the successor function

Rose Schneider
UC San Diego, La Jolla, California, United States

Kaiqi Guo
UC San Diego, La Jolla, California, United States

David Barner
UC San Diego, San Diego, California, United States

Abstract

The successor function a recursive function S which states that for every natural number n, S(n) = n+1 underlies our understanding of the natural numbers as an infinite class. Recent work has found that acquisition of this logical property is surprisingly protracted, completed several years after children master the counting procedure. While such work links successor knowledge with counting mastery, the exact processes underlying this developmental transition remain unclear. Here, we examined two possible mechanisms: (1) recursive counting knowledge, and (2) formal training with the +1 rule in arithmetic. We find that while both recursive counting and arithmetic mastery predict successor knowledge, arithmetic performance is significantly lower than measures of recursive counting for all children. This dissociation suggests children do not generalize the successor function from trained mathematics; rather, we find evidence consistent with the hypothesis that successor knowledge is supported by the extraction of recursive counting rules.

Examining the multimodal effects of parent speech in parent-infant interactions

Sara E. Schroer (seschroe@iu.edu)
Linda B. Smith (smith4@indiana.edu)
Chen Yu (chenyu@indiana.edu)
Department of Psychological & Brain Sciences, Indiana University
1101 East 10th Street, Bloomington, IN 47405 USA

Abstract

Parental input in the form of visual joint attention is hypothesized to serve a critical role in the development of infant attention, acting as a training ground by scaffolding an infant's ability to sustain visual attention in real-time. We extended this hypothesis by studying the effects of parent speech on infant visual and manual attention. Thirty-four toddlers and their parents participated in a free-play study while wearing head-mounted eye trackers. Infant multimodal behaviors were measured in four ways: visual attention, manual action, hand-eye coordination, and joint visual attention with their parent. Overall, we found that longer durations of attention were accompanied by parent speech. Moreover, sustained attention, defined as behaviors lasting 3s or more, almost always occurred with parent speech. Individual differences in parent-infant coordination were also explored. These results suggest that parent-infant interactions create multimodal opportunities for infants to practice sustaining attention.

Keywords: attention, children, cognitive development, eye-tracking, interactive behavior

Introduction

Infants are active learners – they seem to be self-motivated to explore and make predictions about their world. Early development is not solely an individual process, however – it is also embedded in a highly social context as young infants are taught and supported by caregivers. Parents provide scaffolding to their infants in many different ways and in many different contexts, such as recruiting the child's attention, reducing degrees of freedom, and providing demonstrations (Wood, Bruner, & Ross, 1976). Parent scaffolding has been shown to support the development of executive functioning (Bibok, Carpendale, & Müller, 2009) and verbal skills (Smith, Lamdry, & Swank, 2000). In early language learning, parents use infant-directed speech (Thiessen, Hill, & Saffran, 2005), intersensory redundancy (Gogate & Bahrick, 1998), and selective labeling of objects based on infant behaviors (Pereira, Smith, & Yu, 2014) to support word learning. The idea of parental scaffolding has even been adapted by robotics and AI researchers to build a robotic arm that can learn grasp affordances (Ugur, Nagai, Celikkanat, & Oztop, 2015). Understanding how the mature partner influences the sensorimotor experiences and actions of the young infant to support early development and learning is a key question in cognitive development.

Recent work by Yu & Smith (2016) revealed significant effects of parent behaviors on an infant's capacity for sustained attention. In the study, infants and their parents sat at a table while playing with a set of novel toys. Using head-mounted eye tracking, the authors identified moments when parents and infants jointly attended to (or shared attention to) an object and when infants sustained attention on the same object for at least 3s. When the dyad engaged in joint attention, the duration of the infant's sustained attention bout significantly increased, suggesting that 12-month-old infants' ability to sustain attention is scaffolded by parent attention.

Built upon this finding, a recent study (Suarez-Rivera, Smith, & Yu, in press) provided evidence that the social scaffolding effects from parents are not only limited to parent looking behavior. When parent visual attention was accompanied by other types of parent actions – such as talking and manual actions on objects – infants' sustained attention was further improved. Similarly, this redundancy of parent behaviors has been shown to promote joint attention with younger infants (3-to-11-months-old; Deák, Krasno, Jasso, & Triesch, 2018). In parent-infant interactions, both social partners generate various actions moment-by-moment to create multimodal dependencies of looking, talking, and touching, both within the infant's own system and between the two partners. If multimodal behaviors from parents have effects on infants' visual attention, then parent behaviors may also have effects on other, multimodal infant behaviors. The overarching hypothesis in the present study is that parent speech has cascading effects on not only infant visual attention but a suite of multimodal behaviors in parent-infant interactions.

We chose parent speech to study parent scaffolding because it plays a critical role in early communication and early language development. Hart & Risley (1995) famously demonstrated that the amount parents talk to infants is predictive of the varying language abilities of 3-years-olds in different socioeconomic strata. Subsequent studies show both quality and quantity of parent speech is predictive of later language outcomes (Tamis-LeMonda, Bornstein, & Baumwell, 2001; Hirsh-Pasek et al., 2015). While past research has focused on how parent speech and its linguistic properties, such as infant-directed speech and wh-questions

in speech, predict later child vocabulary size (e.g., Rowe, 2012; Weisleder & Fernald, 2013), the present study will examine the non-linguistic effects of parent speech.

Studying the role of parent speech in the micro-level dynamics of parent-infant interactions is a crucial next step in the field. Although joint visual attention facilitates infant sustained attention (Yu & Smith, 2016), we know that joint attention during toy play does not result from infants following the gaze of their caregivers and does not require any overt bid for the partner's attention (Yu & Smith, 2017a; Deák et al., 2018). During play, adult object manipulations (often coupled with other behaviors, such as speech), are the most promotive of joint attention (Deák et al., 2018). However, maternal speech is tightly linked to object manipulation and occurs frequently in an interaction as a response to infants' visual attention to objects, handling of multiple objects, and vocalizations (Chang, de Barbaro, & Deák, 2017). Parents verbally respond to a suite of multimodal infant behaviors, potentially serving as scaffolding for not only joint attention but also other forms of sustained attention.

To test the multimodal effects of parent speech, we chose four infant behaviors from parent-infant interactions that have been shown to be important in early development: 1) visual attention; 2) manual action; 3) hand-eye coordination; and 4) joint attention. Visual attention was chosen because infant sustained visual attention predicts later language learning and cognitive development (Kannass & Oakes, 2008; Lawson & Ruff, 2004; Yu, Suanda, and Smith, 2018). Manual action was chosen because motor skills, including object exploration, are known to play a major role in early language development (Iverson, 2010). Hand-eye coordination was chosen because both infants and parents attend to their own actions and their partner's object manipulations in free play (Yu & Smith, 2017b). Lastly, joint attention between infant and parent was chosen because dyadic differences in the frequency with which parents and children engage in episodes of joint attention predict individual differences in child vocabulary size (Tomasello & Todd, 1983). For all of these behaviors, we will be looking at sustained attention, defined as when infants attend to an object for a long duration (e.g., greater than 3 seconds). While sustained visual attention is known to predict later outcomes (Kannass & Oakes, 2008; Lawson & Ruff, 2004, Yu, Suanda, and Smith, 2018), the ability to sustain attention in other modalities has not been explicitly studied.

The present study had two goals. In Study 1, we examined the multimodal effects of parent speech by measuring the durations of the four types of multimodal behaviors when they were accompanied by parent speech and comparing with when they were not. We hypothesized that parent talk increases infants' ability to sustain their multimodal behaviors. In Study 2, we focused on individual differences in parent speech, given that some parents generated more speech than others did in free play. We examined whether varying amounts of parent speech create different effects on infants' multimodal behaviors.

Methods

Thirty-four toddlers (mean age = 18.67mos [range: 12.3-24.3]; female = 16) and their parents participated in a study on naturalistic parent-infant interactions during free play. An additional 5 dyads were included in the experimental data set but were excluded from the current analyses due to missing parent eye-tracking (n = 2) and non-transcribable speech (n = 3).

Data Collection

Parents and infants played with 24 toys on a carpeted floor in a playroom for an average of 7.15 minutes (range 3.93-11.64). At the beginning of the play session, the toys were randomly spread out across the floor. Parents were instructed to play as they would at home and that they could sit in any orientation (behind, next to, in front of their infant), but were asked to keep their infant sitting on the floor due to the eye tacker's cable.

During the play session, both parent and infant wore a head-mounted eye tracker (Positive Science LLC). The eye tracker system used a scene camera on the participant's forehead to record images from the wearer's perspective with a visual field of 108°. A second, infrared camera pointed to the participant's right eye to record saccades and fixations. Both cameras sampled at a rate of 30Hz. The infant's eye tracker was affixed to a hat and the parent wore their eye tracker like a pair of glasses. Additional cameras were placed in the room to capture traditional third-person views of the dyad (Figure 1).

Figure 1: Experimental set-up (left) and the infant's first-person view, the cross-hair indicates infant gaze (right).

The experiment was run by two researchers. The session began by one researcher placing the eye tracker on the parent and adjusting the scene and eye cameras, while the other researcher engaged with the infant. Afterwards, both researchers worked together to place the eye tracker on the infant. One researcher, and the parent, continued to distract the infant with exciting toys (e.g. a pop-up toy that played music) as the other researcher set up the eye tracker on the infant. After both members of the dyad were wearing their eye trackers, the researchers ran a brief calibration procedure. A large board that had lights and produced sounds was placed in front of the infant (approximately 30 cm away). One of the researchers controlled the board and lit up one of the lights

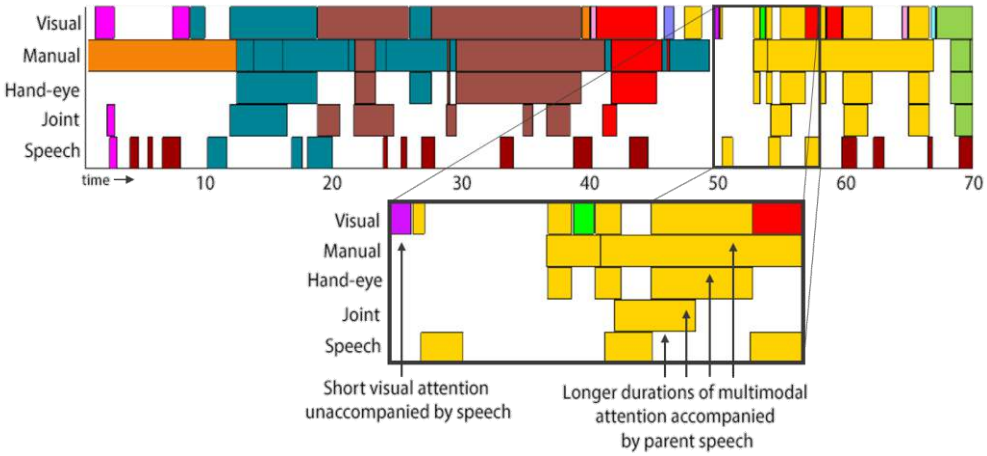


Figure 2: Data streams of infant visual attention, manual action, hand-eye coordination, joint attention, and parent speech over 70s of an interaction. Each block represents a behavioral event and each color represents a different object. The color of parent speech represents the object being named, dark red indicates no naming in that utterance.

until both the parent and infant shifted their gaze to that location. This procedure was repeated for 15 light locations.

The researchers monitored the experiment from an adjoining room. If the infant's eye camera was bumped or moved during play, the researchers reentered, adjusted the camera, and completed an abridged calibration procedure.

Coding and Analyses

Following the experiment, the eye tracking videos from the scene and eye cameras were synchronized and calibrated with a software program to generate a cross-hair that indicated where the participant was looking during each frame of the video (Figure 1). Parent and infant visual gaze were then coded manually using the first-person view (from the scene camera) with the cross-hair overlaid. Using an in-house program, the coder annotated which region of interest (ROI) the cross-hair overlapped with during a fixation. There were 25 ROIs – one for each toy and the social partner's face.

The scene cameras and third-person views were then used to annotate the objects being handled by a participant, frame-by-frame, in an in-house program. If a hand was touching an object, the object was considered "in hand". Participants' left and right hands were coded separately.

Parent speech was transcribed using Audacity at the utterance level. There was no minimum length for an utterance, but separate utterances had to be 400ms or more apart (otherwise they were collapsed together). All parent talk and vocal play (like saying "vroom-vroom" or making a crashing sound) were considered speech. Due to the 400ms criteria, chunks of speech that would be considered sentences could be split apart and separate sentences could be counted as one utterance.

In the current studies, we were interested in five behaviors: infant visual attention, manual action, hand-eye coordination, dyadic joint attention, and parent speech (Figure 2). **Visual attention** was defined as all infant fixations to the 25 ROIs. **Manual action** was similarly defined as all instances of the infant touching an object with either or both hands. **Hand-eye coordination** was defined as moments when the infant looked at and handled the same object, for any duration of time. **Joint attention** between the parent and infant was defined as any moment when the parent's and infant's visual attention fell on the same ROI. All parent utterances were counted as **speech**.

For all four multimodal behaviors (visual attention, manual action, hand-eye coordination, and joint attention) sustained attention was defined as a behavior lasting 3 seconds or longer (to match the previously used definition in Yu & Smith, 2016).

To test the effects of parent speech, we categorized each attention bout as "with speech", if the onset of a parent utterance began after the onset of the attention bout and before the offset of the bout. Other attention bouts, without any overlap with a parent utterance, were categorized as "without speech". With this definition, we can measure the effects of parent speech by comparing attention bouts in the two categories.

Study 1: Multimodal Effects of Parent Support

In Study 1, we tested the relationship between parent speech and the four multimodal measures of infant behavior. Corpus-level analyses were used to compare the durations of all multimodal attentional bouts with and without speech.

Each modality was analyzed separately using mixed effects models to predict the duration of a bout by whether it was accompanied by speech, with subject and attended object as random effects. Each full model was then compared to a null model, with intercept and random effect of object only, using Chi-Square difference tests. All four multimodal behaviors were found to last longer when co-occurring with speech (Figure 3, Table 1).

To specifically test whether parent speech co-occurs with *sustained* attention, an infant behavior known to predict later outcomes (Yu et al., 2018), similar models were used to analyze the subset of sustained attention bouts lasting 3s or more. Bouts of sustained attention of each multimodal

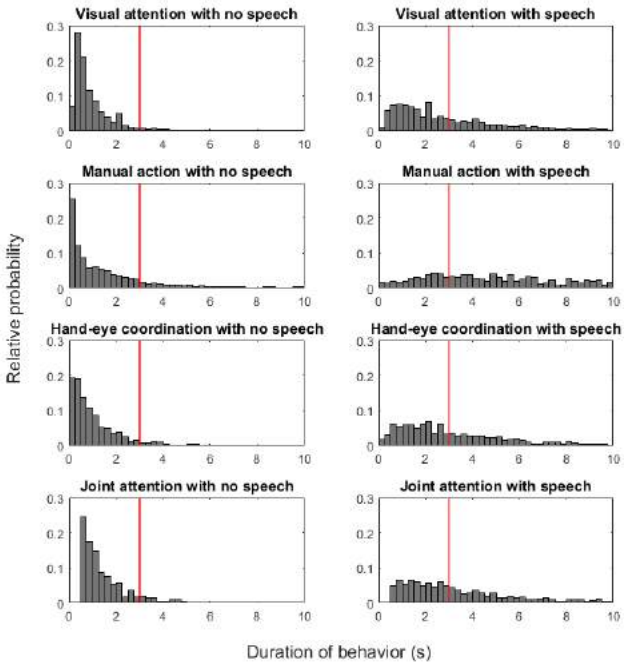


Figure 3: Durations of behaviors without speech (left column) and with speech (right column). Red line indicates the 3-second threshold for sustained attention.

behavior were also longer, and more likely to occur, with speech (Table 1).

The mean duration of **visual attention** bouts with speech was more than 3 times longer than the mean duration of bouts without speech ($M_{with\text{-}speech}$=3.769s, $M_{w/o\text{-}speech}$=1.000s). When we examined the subset of sustained attention bouts that were longer than 3s, sustained attention bouts increased in duration by 50% when accompanied by parent speech ($M_{with\text{-}speech}$=7.196s, $M_{w/o\text{-}speech}$=4.842s).

The mean duration of **manual action** bouts with speech was nearly 8 times longer than the mean duration of bouts without speech ($M_{with\text{-}speech}$=11.187s, $M_{w/o\text{-}speech}$=2.471s). Sustained manual action bouts that co-occurred with parent speech were close to double the duration of bouts without parent speech ($M_{with\text{-}speech}$=13.840s, $M_{w/o\text{-}speech}$=8.629s).

The mean duration of **hand-eye coordination** bouts with speech was nearly 4 times longer than bouts without speech ($M_{with\text{-}speech}$=4.004s, $M_{w/o\text{-}speech}$=1.084s). Sustained hand-eye coordination bouts increased in duration by 50% when accompanied by parent speech ($M_{with\text{-}speech}$=6.961s, $M_{w/o\text{-}speech}$=4.587s).

Lastly, the mean duration of parent-infant **joint attention** with parent speech was more than 2 times longer than joint attention without speech ($M_{with\text{-}speech}$=4.284s, $M_{w/o\text{-}speech}$=1.476s). As with the other behaviors, the duration of sustained attention events with parent speech was longer than sustained attention events without parent speech ($M_{with\text{-}speech}$=6.967s, $M_{w/o\text{-}speech}$=4.071s).

Across all four multimodal behaviors, the duration of infant attention is extended when the bout is accompanied by parent speech. Moreover, when we specifically examined sustained attention bouts, we saw that not only is sustained attention substantially more likely to occur with parent speech, but that bouts of sustained attention with parent speech are significantly longer.

Study 2: Individual Differences

If we view the parent as a coach, training their infant to engage in sustained attention (Yu & Smith, 2016), then we should see differences in how the dyads practice, since different coaches may have different coaching styles. Parents may vary in the "drills", or amount of speech, they use in practice. If so, infants may react differently to parent's coaching which will influence how much they "score" in sustained attention. To understand the individual differences in the coordination of parent speech and infant attention, we examined whether more or less parent talk has different effects on the infant's ability to sustain attention.

Parents varied in how much they spoke to their infants. The average parent produced 16.819 utterances/minute (*SD* =3.844), though the quietest parent only spoke 9.597 times/minute and the most "talkative" parent generated 25.144 spoken utterances per minute. To test the relationship between parent speech and infant sustained attention, we divided the subjects into two groups based on a median split (median = 16.814 utterances/min). Parents in the high frequency speech group produced 19.905 utterances per minute while parents in the low frequency speech group produced on average 13.734 utterances per minute. The low frequency and high frequency groups did not differ in the mean duration of parent utterances (M_{low}=1.309s, M_{high}=1.330s, p = 0.871), suggesting low frequency parents

Table 1: Duration of multimodal behaviors with and without speech

		instances with speech			instances without speech			statistical comparison			
		# of bouts	mean dur	sd	# of bouts	mean dur	sd	beta	p-value	95% CI	null model comparison
visual	overall	2439	3.769	4.938	4053	1.000	1.283	2.741	< 0.001	[2.597 2.885]	$\chi2 = 1262.300$, $p < 0.001$
attention	sustained	986	7.196	5.227	171	4.842	3.971	2.654	< 0.001	[1.839 3.466]	$\chi2 = 40.181$, $p < 0.001$
manual	overall	1736	11.187	15.821	1788	2.471	4.989	8.468	< 0.001	[7.710 9.229]	$\chi2 = 448.450$, $p < 0.001$
action	sustained	1355	13.840	16.984	362	8.629	8.536	4.768	< 0.001	[3.047 6.491]	$\chi2 = 29.238$, $p < 0.001$
hand-eye	overall	920	4.004	4.463	1411	1.084	1.206	2.881	< 0.001	[2.640 3.121]	$\chi2 = 496.190$, $p < 0.001$
coordination	sustained	416	6.961	5.234	88	4.587	1.731	2.708	< 0.001	[1.617 3.791]	$\chi2 = 23.289$, $p < 0.001$
joint	overall	1033	4.284	4.430	865	1.476	1.047	2.796	< 0.001	[2.505 3.087]	$\chi2 = 325.500$, $p < 0.001$
attention	sustained	506	6.967	5.046	76	4.071	1.109	3.681	< 0.001	[2.583 4.767]	$\chi2 = 42.025$, $p < 0.001$

Table 2: Sustained attention in dyads with low frequency and high frequency parent speech

	low frequency group			high frequency group			statistical comparison			
	# bouts	mean dur	sd	# bouts	mean dur	sd	beta	p-value	95% CI	null model comparison
sustained visual attention	401	7.636	6.125	585	6.894	4.490	-0.523	0.121	[-1.185 0.136]	$\chi2 = 2.425$, p = 0.119
sustained manual action	604	14.946	18.215	751	12.950	15.881	-2.424	0.008	[-4.220 -0.619]	$\chi2 = 6.928$, p = 0.008
sustained hand-eye coordination	199	7.404	5.918	217	6.555	4.500	-0.822	0.144	[-1.838 0.195]	$\chi2 = 2.511$, p = 0.113
sustained joint attention	198	7.865	6.571	308	6.389	3.649	-1.165	0.009	[-2.035 -0.298]	$\chi2 = 6.911$, p = 0.009

were truly producing less speech, not just fewer, longer utterances. Therefore, the durations of spoken utterances in the two groups would not be a factor to influence infant's attention.

We then compared the durations of sustained attention bouts produced by infants in the low frequency and high frequency groups. To directly measure the effects of parent speech, we only analyzed sustained attention bouts that were accompanied by parent speech. As in Study 1, each type of multimodal behavior was analyzed separately using mixed effects models, with object attended to as a random effect, and then compared to a null model with intercept and random effect of object only.

For manual actions and joint attention, we found that the duration of attentional bouts was longer for infants in the low frequency group (Table 2). The mean duration of sustained **manual action** bouts was 2 seconds longer in the low frequency group (M_{low}=14.946s, M_{high}=12.950s). The mean duration of sustained **joint attention** bouts was more than a second longer in the low frequency group (M_{low} =7.865s, M_{high} =6.389s). There were no differences between the low frequency and high frequency groups in the durations of sustained visual attention or hand-eye coordination.

We present evidence of two groups of dyads, classified by how much speech parents produced in an interaction. These two groups coordinate their attention in different ways – in the low frequency group, there are less occurrences of speech-attention overlap in all four types of behavior. But, when parent speech co-occurs with manual action or joint attention, infants in the low frequency group had significantly longer durations of sustained attention than infants in the high frequency group. This finding suggests two possible phenomena: 1) parents who talked less may be more selective in when they choose to talk; or 2) infants whose parents talked less are more responsive when their parent does talk.

Discussion

With the current studies, we examined the dynamics of parent-infant interactions, specifically the role of parent behaviors in influencing infant attention. We demonstrated that the duration of infants' visual attention is longer when accompanied by parent speech, extending prior work that focused primarily on parent's visual attention (Yu & Smith, 2016). Furthermore, we measured the relationship between parent speech and multiple infant sensory-motor behaviors beyond visual attention – manual action, hand-eye coordination, and joint attention – and found a similar coordination between parent speech and infant sustained attention. Sustained attention of each of these multimodal behaviors is more likely to occur, and lasts longer, when accompanied by parent speech.

There were, however, individual differences in the observed parent-infant coordination. Parents that spoke less during the interaction had infants with longer durations of sustained manual attention and dyadic joint attention, relative to their talkative peers. This relationship could have two (non-mutually exclusive) causes. One possible explanation is that infants with less talkative parents are more responsive to their parent's speech. Using the coaching analogy, those infants may not get coaching signals very often and therefore they respond to the signals better when they receive them. Another possible explanation is that parents that talk less are more selective in when they choose to talk. Rather than "coach" all the time, irrespective of their infant's attentional state, these parents may find optimal moments to support their infants. Regardless, it suggests that dyads with less talkative parents are still having high-quality practices. Parents that talk more can scaffold their infant's ability to sustain attention more frequently, creating more opportunities for the infant to score. Dyads with less talkative parents, however, appear to employ more effective drills during their practices – even though these infants "score" less, the durations of their sustained manual action and joint attention bouts are longer. Thus, there are two different pathways through which parents can support their infants. Future research needs to examine potential qualitative and quantitative differences between the two pathways used by more and less talkative parents, and how different dyads adjust and adapt to different interaction patterns based on the history of their experiences.

Our results present evidence of a multimodal sustained attention training ground. The coupling of parent speech and infant attention suggests that the more infants sustain their attention, the more parents respond to it, giving the infant even more time to practice. Coaching improves an infant's ability to sustain attention, increasing the time an infant can learn about the object's properties (Ruff, 1986) and creating more opportunities for the parent to talk about and label objects (Yu & Smith 2012; Pereira et al., 2014), fostering a developmental cascade yielding higher language outcomes (Yu et al., 2018). We are also among the first to study sustained attention beyond the visual modality. How

sustained manual attention, hand-eye coordination, and joint attention relate to later outcomes is still an open question to be investigated further, especially given the individual differences seen in manual attention and joint attention.

To better understand the parental scaffolding of sustained attention, we need to study the infant behaviors that elicit parent responses. It is unlikely that parents are randomly speaking during an interaction. Rather, they are responding contingently to certain infant behaviors and following non-linguistic cues like gaze, object manipulation, gesturing, smiling, and more. To create successful object labeling moments, a parent and infant need to couple their behavior so that they are attending to and naming the same object. Infants need to sustain their attention to the object long enough for the parent to provide a label, which requires the infant exhibiting behaviors indicating a readiness to learn (e.g. object-directed vocalizations; Goldstein, Schwade, Briesch, & Syal, 2010) and parents being able to follow these behaviors. One way to address this question is to analyze the temporal dynamics of parent-infant interactions. Measuring parent and infant behaviors seconds before a parent utterance and the subsequent behavioral changes after the utterance will provide further insight into how dyads coordinate their behaviors and influence one another. One possibility is that there are "signatures" that reliably predict whether a parent utterance leads to sustained attention and successful object-label mappings. Studying the temporal dynamics of infant looking and object handling before and after a naming moment revealed developmental changes from 4 to 9 months (Chang et al., 2017), positioning this form of analysis as a pertinent future direction.

Conclusion

Previous work has shown that joint visual attention supports an infant's ability to sustain attention. We extended these findings by measuring the multimodal effect of parent speech on infant visual attention, manual action, hand-eye coordination, and joint attention. When multimodal attention is accompanied by parent speech, the infant sustains their attention for longer periods of time, creating a rich training ground for early development.

Acknowledgements

This research was funded by National Institutes of Health Grant R01HD074601 and R01HD093792 to CY. SES was supported by NSF GRFP 1342962.

References

Bibok, M. B., Carpendale, J. I., & Müller, U. (2009). Parental scaffolding and the development of executive function. *New directions for child and adolescent development*, 2009(123), 17-34.

Chang, L., de Barbaro, K., & Deák, G. (2016). Contingencies between infants' gaze, vocal, and manual actions and mothers' object-naming: Longitudinal changes from 4 to 9 months. *Developmental neuropsychology*, 41(5-8), 342-361.

Deák, G. O., Krasno, A. M., Jasso, H., & Triesch, J. (2018). What leads to shared attention? Maternal cues and infant responses during object play. *Infancy*, 23(1), 4-28.

Gogate, L. J., & Bahrick, L. E. (1998). Intersensory redundancy facilitates learning of arbitrary relations between vowel sounds and objects in seven-month-old infants. *Journal of experimental child psychology*, 69(2), 133-149.

Goldstein, M. H., Schwade, J., Briesch, J., & Syal, S. (2010). Learning while babbling: Prelinguistic object-directed vocalizations indicate a readiness to learn. *Infancy*, 15(4), 362-391.

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday life of America's children.* Baltimore, MD: Paul Brookes.

Hirsh-Pasek, K., Adamson, L. B., Bakeman, R., et al. (2015). The contribution of early communication quality to low-income children's language success. *Psychological science*, 26(7), 1071-1083.

Iverson, J. M. (2010). Developing language in a developing body: The relationship between motor development and language development. *Journal of child language*, 37(2), 229-261.

Kannass, K. N., & Oakes, L. M. (2008). The development of attention and its relations to language in infancy and toddlerhood. *Journal of cognition and development*, 9(2), 222-246.

Lawson, K. R., & Ruff, H. A. (2004). Early focused attention predicts outcome for children born prematurely. *Journal of developmental & behavioral pediatrics*, 25(6), 399-406.

Pereira, A. F., Smith, L. B., & Yu, C. (2014). A bottom-up view of toddler word learning. *Psychonomic bulletin & review*, 21(1), 178-185.

Rowe, M. L. (2012). A longitudinal investigation of the role of quantity and quality of child-directed speech in vocabulary development. *Child development*, 83(5), 1762-1774.

Ruff, H. A. (1986). Components of attention during infants' manipulative exploration. *Child development*, 105-114.

Smith, K. E., Landry, S. H., & Swank, P. R. (2000). Does the content of mothers' verbal stimulation explain differences in children's development of verbal and nonverbal cognitive skills?. *Journal of school psychology*, 38(1), 27-49.

Suarez-Rivera, C., Smith, L. B. & Yu, C. (in press) Multimodal parent behaviors within joint attention support sustained attention in infants. *Developmental psychology.*

Tamis-LeMonda, C. S., Bornstein, M. H., & Baumwell, L. (2001). Maternal responsiveness and children's achievement of language milestones. *Child development*, 72(3), 748-767.

Thiessen, E. D., Hill, E. A., & Saffran, J. R. (2005). Infant-directed speech facilitates word segmentation. *Infancy*, 7(1), 53-71.

Tomasello, M., & Todd, J. (1983). Joint attention and lexical acquisition style. *First language*, 4(12), 197-211.

Ugur, E., Nagai, Y., Celikkanat, H., & Oztop, E. (2015). Parental scaffolding as a bootstrapping mechanism for learning grasp affordances and imitation skills. *Robotica*, 33(5), 1163-1180.

Weisleder, A., & Fernald, A. (2013). Talking to children matters: Early language experience strengthens processing and builds vocabulary. *Psychological science*, 24(11), 2143-2152.

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of child psychology and psychiatry*, 17, 89–100.

Yu, C., & Smith, L. B. (2012). Embodied attention and word learning by toddlers. *Cognition*, 125(2), 244-262.

Yu, C., & Smith, L. B. (2016). The social origins of sustained attention in one-year-old human infants. *Current biology*, 26(9), 1235-1240.

Yu, C., & Smith, L. B. (2017a). Multiple sensory-motor pathways lead to coordinated visual attention. *Cognitive science*, 41, 5-31.

Yu, C., & Smith, L. B. (2017b). Hand–Eye Coordination Predicts Joint Attention. *Child development*, 88(6), 2060-2078.

Yu, C., Suanda, S. H., & Smith, L. B. (2018). Infant sustained attention but not joint attention to objects at 9 months predicts vocabulary at 12 and 15 months. *Developmental science*, e12735.

Spatial Memory of Immediate Environments

Holger Schultheis (schulth@informatik.uni-bremen.de)
University of Bremen
Bremen Spatial Cognition Center
Institute for Artificial Intelligence
Am Fallturm 1, 28359 Bremen, Germany

Abstract

Memorizing and retrieving information about the spatial layout of one's surrounding is of crucial importance for humans. We propose a new theory of spatial memory of immediate environments and develop a corresponding computational realization. We detail how the theory explains key findings on human spatial memory (use) and show that the computational realization accounts well for human behavior from three pertinent experiments. One implication of the theory's success is that enduring spatial memory representations may best be conceptualized as flexible combinations of representation structures and reference frames.

Keywords: spatial memory; spatial reference frames; interference; perspective taking; computational modeling

Introduction

Knowledge about the location of and relation between objects in the immediate environment is crucial for everyday life. Such spatial knowledge can be obtained by perception, but many everyday activities crucially rely on memory representations of spatial information. For example, spatial memory enables avoiding collisions with objects currently outside the perceptual field (e.g., a chair slightly behind oneself when moving around the table), anticipating and planning movements from positions one has not yet reached (e.g., planning the movement to place a plate once one is next to the appropriate spot on the table), and navigating towards objects which are not directly perceivable (e.g., approaching the appropriate cupboard to retrieve the plates contained in it).

In line with its importance, spatial memory of immediate environments has received considerable attention by research in the cognitive sciences and a substantial number of theories of spatial memory has been proposed (Avraamides & Kelly, 2008; Byrne, Becker, & Burgess, 2007; Mou, McNamara, Valiquette, & Rump, 2004; Sholl, 2001; Wang, 2017). A prominent approach to investigating spatial memory of immediate environments have been *perspective taking* (PT) studies, in which people have to judge spatial relations of a previously learned object layout from imaginal perspectives. From these studies a number of main findings have emerged (May, 2007), which can be assumed to characterize key aspects of the structures and processes involved in spatial memory.

We propose a new theory of spatial memory that offers explanations for all main findings. We first describe the PT paradigm and the main findings arising from it. Subsequently, we expound our theory, how it explains the findings, and a computational realization of the theory. After briefly considering related theories, we close with a discussion of the implications of our work.

Main Findings

In typical PT studies on spatial memory of immediate environments people are first asked to memorize the location of objects in their surrounding. After learning the object layout, people are deprived of perceptual access to their surrounding (e.g., by blindfolding) and tested for their knowledge of the spatial relations between objects. Two common forms of testing spatial relations are *judgment of relative direction* and *egocentric pointing*. In a judgment of relative direction task, people are asked to point to obj_1 as if they were standing at obj_2 facing obj_3 (where obj_i are three objects of the previously learned object layout). In an egocentric pointing task, people are asked to point to obj_1 as if they were standing at or facing obj_2. The object to point to is called the *target object*. In particular, the to-be-imagined perspective (e.g., facing obj_2) is usually different from the actual bodily perspective of the participants. The imaginal perspective can differ from the bodily perspective by *rotation* (i.e., the locations of bodily and imaginal perspective coincide, but orientations of the perspectives differ), *translation* (i.e., the orientations of the perspectives coincide, but locations differ), or both (often the case in judgment of relative direction tasks).

By using such a PT approach, existing studies have uncovered many intriguing phenomena of spatial memory organization and access. In the following, we will focus on a set of phenomena, which can be considered the main findings of existing research (May, 2007):

- Taking an imaginal perspective different from the bodily perspective is hard. Indicating the direction to the target object from the imaginal perspective takes more time and is more error prone than from the bodily perspective.
- Imaginal perspectives involving rotations are harder (slower, more error prone) than imaginal perspectives involving only translations.
- The difficulty of pointing to the target object increases with increasing angular disparity between the pointing direction from the imaginal perspective and the pointing direction from the bodily perspective.

- The difficulty of responding from the imaginal perspective can be reduced, if observers are ignorant of the actual spatial relation of their body to the object layout (e.g., when being disoriented).
- Differences between the orientation of the imaginal perspective and salient orientations in the environment (e.g., orientation of axes of symmetry or orientation of learning perspective) may lead to extra processing costs.
- If people are allowed to move their body such that the bodily perspective coincides with the tested perspective, the above mentioned difficulties are reduced notably and sometimes even eliminated.

A further finding that we will consider is the influence of perspective preparation on PT performance. If people are given information about the tested perspective before they are informed about the target object, they may be able to prepare the to-be-taken perspective such that they can respond with less difficulty once the target object is presented. Several studies have investigated the influence of preparation, because preparation effects can help reveal how access to spatial memory is organized. Although preparation has been found to generally reduce processing times associated with PT (Brockmole & Wang, 2003; May, 2004), it seems hard to prepare for certain difficulties (e.g., increase of processing costs with increasing disparity May, 2004; Wang, 2005).

A Theory of Spatial Memory

As virtually all previous theories of spatial memory (Avraamides & Kelly, 2008; Byrne et al., 2007; Mou et al., 2004; Sholl, 2001; Wang, 2017), our theory assumes that one component of spatial memory is what we will call the *sensorimotor representation*. It represents self-to-object relations for (certain) objects in the immediate environment. If any movement of one's body is perceived (through vision, proprioception, etc.) these relations are updated accordingly. In this sense the sensorimotor representation is dynamic and transient. Access to this representation is quick and automatic and it serves as the default basis for motor actions.

In addition, our theory assumes a more enduring representation of the environment as a second component. We will call this component the *LTM representation*. It represents object-to-object relations between the objects in the immediate environment. One's own body can be one of the objects in the LTM representation. The LTM representation is orientation-free and not inextricably linked to some spatial *reference frame* (RF). However, the representation may be associated with a RF in the same sense as items in long-term memory are usually assumed to be associated with each other.

Because the LTM representation is orientation-free, it will be of limited use without further additions. Consider the two object layouts in Fig. 1. Both layouts yield identical representations in an orientation-free representation, but for acting on or within the layout (e.g., approaching object A) it makes a difference which situation is represented. To create the necessary correspondence between the LTM representation and the real world, that is, to anchor the representation in the real world, it has to be oriented. We argue that this is achieved by imposing a spatial RF onto the LTM representation and our theory assumes that any access to the LTM representation involves such an imposition of a RF. A common RF that people will likely employ to access the LTM representation is the *bodily* RF arising from the sensorimotor representation (i.e., a RF that is oriented as the actual body). Other RFs may be RFs associated with the LTM representation (e.g., RFs salient during encoding of the spatial layout) or an *imaginal* RF that allows assessing the spatial layout from a vantage point differing from the current bodily vantage point (e.g., when trying to identify the seats with the best view on the stage in a theater without first walking through the whole theater).

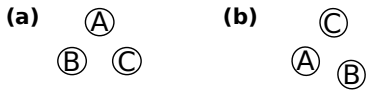


Figure 1: Two spatial layouts (a) and (b) that yield identical orientation-free object-to-object representations.

Notably, differing RFs may concurrently be available for accessing the LTM representation. Accordingly, we propose that accessing the LTM representation requires RF selection and depending on which frames are available this selection may be competitive and effortful. Specifically, our theory assumes that selection probability and effort depend on the conflict between the available frames, where conflict is a function of the salience and the (mis)alignment of the available frames (see further detail below).

According to our theory, taking an imaginal perspective involves the following steps (see also Fig. 2): First, a RF has to be selected and imposed onto the LTM representation. To perform the PT task successfully, the selected RF needs to be the one corresponding to the to-be-taken perspective or a different but aligned RF. Second, once the RF has been imposed, the LTM representation can be accessed to determine the direction towards the target object. Third, the determined target direction is used to activate a pointing movement towards the target. If the imaginal and bodily perspective differ from each other, the determined pointing direction is in conflict with the pointing direction to the target given by the sensorimotor representation. Because access to the sensorimotor representation is automatic, the disagreeing movement directions give rise to motor interference. The strength of this interference is assumed to depend on the dissimilarity of the two movements: The more the two movements' differ in direction, the stronger the interference. Note that in this process, activation of the motor response can only start after LTM access is completed. On the other hand, nothing precludes RF selection and LTM access to happen before the target direction is known. Accordingly, we propose that LTM access may start before the target object is known.

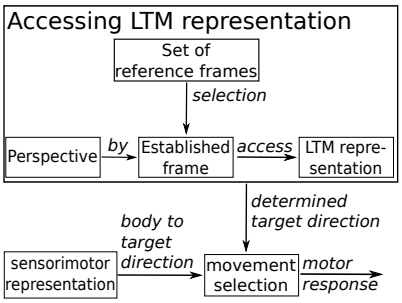


Figure 2: Processing steps in imaginal perspective taking.

Explanation of Main Findings

Our theory explains the main findings mentioned above as follows:

- Imaginal PT is difficult, because it requires RF selection to access the LTM representation. Selecting a frame when (at least) the bodily frame and the imaginal frame are in conflict takes more time than when the frames are aligned. Furthermore, an incorrect frame (e.g., the bodily frame) may be selected, which will result in an erroneous pointing response. Additional difficulty arises from motor interference such that interference leads to slower and more error prone response execution.
- Perspectives involving rotations create a misalignment between the orientation of the bodily and imaginal RF. Perspectives involving only translations do not lead to such a misalignment. Consequently, accessing the LTM representation is harder for rotations than for translations
- With increasing disparity, the difference between movement direction to the target from the bodily and the imaginal perspective increase. This leads to increased motor interference, which results in slower and more error prone responding.
- Lacking a sensorimotor representation has two effects: First, the sensorimotor representation does *not* give rise to a RF, which may otherwise have lead to conflict during LTM access. Second, motor interference is reduced or even eliminated. Accordingly, taking an imaginal perspective different from the bodily perspective can be easier without a sensorimotor representation.
- If a RF is associated with the LTM representation, it may be co-activated with the LTM representation. If this associated RF differs from the imaginal RF, it creates conflict during accessing the LTM representation. As a result, a disagreement between imaginal perspective and, for example, the learning perspective renders PT more difficult.
- Bodily movements towards the to-be-imagined perspective will lead to an accordingly updated sensorimotor representation. This means that the bodily and the imaginal RFs will be aligned and there will be little or no motor interference. Consequently, PT difficulty will be greatly reduced.
- Access to the LTM representation may proceed during preparation and thus reduce the overall processing time. However, motor interference arises only after the target object has been determined and, consequently, cannot be reduced by preparation. This explains why parts but not all of the processing costs of PT can be reduced by preparation.

Formalization

Our theory's ability to provide explanations for the main effects lends support for its assumptions. To allow comparing the behavior predicted by the theory to human behavior in more detail we formalized the theory as a computational model and applied the model to two pertinent PT studies. As a first step, we decided to use a formalization that captures the main assumptions of the theory while remaining as simple as possible. This has the advantage that any successes or failures of the model can be more directly attributed to the theory and its assumptions instead of being a result of implementation-specific detail (see, e.g., Cooper & Guest, 2014). An implementation of the theory that provides more detail on the possible mechanisms is discussed below.

Because establishing a RF and motor interference are the main factors in driving PT difficulty, the model focuses on these two aspects.

RF Selection. Establishing a RF is formalized as follows: Each of the available reference frames RF_i is assumed to have a salience sal_i such that the salience of all available reference frames sums to one. Following Botvinick, Braver, Barch, Carter, and Cohen (2001), we define the strength of conflict (cV) of RF_i with RF_j as

$$cV(i,j) = \delta * sal_i * sal_j * (1 - jConf),$$

where $jConf$ is the conflict of RF_j to all other RF (i.e., $RF_k, k \neq i$) and

$$\delta = \begin{cases} -1, & \text{for } RF_i \text{and } RF_j \text{aligned,} \\ 1, & \text{for } RF_i \text{and } RF_j \text{misaligned.} \end{cases}$$

The overall conflict of RF_i is given as the sum over all pairwise conflict values across all other RF:

$$cV_i = \sum_{j, j \neq i} cV(i,j).$$

The salience and the conflict of each frame are combined to yield an impact score imp_i. Specifically, the frame's salience is scaled based on its conflict value such that higher conflict leads to a lower impact score and $imp_i \in [0.5 * sal_i, 1.5 * sal_i]$. Probability of a frame being selected sp_i and the speed with which it can be selected st_i are proportional to imp_i:

$$sp_i = \frac{imp_i}{\sum_j imp_j},$$
$$st_i = A * (impMax - imp_i),$$

where $impMax$ is the maximum possible impact score and A serves to scale the response time to the order of magnitude of the human data.

Motor Interference. Processing time arising from motor interference is assumed to be directly proportional to the disparity $disp$ between the directions of the two interfering movements: $B * disp$, where B is a scaling factor analogous to A above. Error is determined distinguishing two cases: First, if the imaginal RF (or a frame aligned with it) is used to access the LTM representation, error is also assumed to be proportional to disparity: $C * disp$. If a frame misaligned with the imaginal frame is selected, the error will amount to the angular difference of the selected frame's and the imaginal frame's orientation.

Example. To illustrate the workings of the model, we will consider a situation, in which a person is located in the middle of a previously learned configuration of objects and asked to point to one of the objects as if facing one of the others. Let us assume that the imagined facing direction differs from the actual bodily facing direction and that the learning view coincides with the actual bodily orientation.

In such a situation, the model computes response time and error for each frame individually. The overall response time and error is given as a weighted average of all individual terms: each individual frame's time and error are weighed by the probability of selecting the frame and the resulting values are summed. To obtain the individual frame's values, the model computes the impact score of all three involved frames. Based on the impact scores, the selection probability and selection time of each frame is computed. Given any individual frame RF_i, the model computes the time from motor interference as a linear scaling of the disparity between pointing from the actual bodily perspective and the imaginal perspective given RF_i. If the selected frame is the imaginal frame, error is computed as a linear scaling analogously to the scaling of time. If any of the other two frames is selected, the error equals the orientation difference between the imaginal frame and the bodily/environmental frame. The ultimate output of the model are its prediction of response time and error in the given situation.

Given that humans are generally well able to perform PT tasks, we assumed that the imaginal frame has the strongest salience and set this salience to 0.6. The remaining salience amount of $1-0.6=0.4$ (saliences of all RF sum to 1) was distributed uniformly across all other RF. This left the scaling factors A, B, C as the only free parameters of the model.

Simulations

The first simulation addressed Experiments 2 and 3 of May (2004). These experiments provide a rich dataset of response times and pointing errors across 24 experimental conditions and also exhibit several of the main findings mentioned above.

Experiments 2 & 3 of May (2004). Participants had to perform an egocentric pointing task with to-be-imagined perspectives being either rotations or translations. Across different blocks, participants either had 1 s, 3 s, or 5 s to prepare their perspective before the target object was presented. Furthermore, the disparity between the target direction from the bodily perspective and the target direction from the imaginal perspective was systematically varied to yield levels of increasing disparity: 22.5°, 67.5°, 112.5°, 157.5°. Both experiments revealed that (a) rotations were slower and more error prone than translations, (b) response time and error increased with increasing disparity, and (c) that overall processing time but not the disparity effect decreased with increasing preparation time.

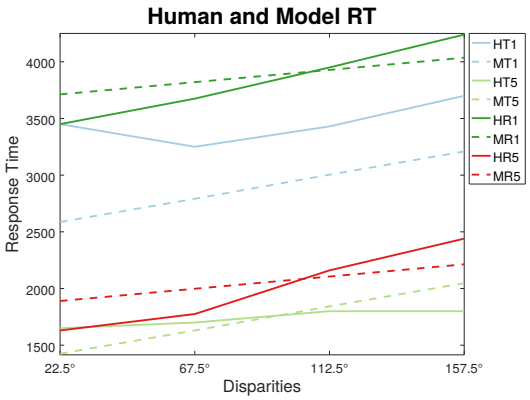


Figure 3: Human (solid lines) and model (dashed lines) response times for Experiment 2 of May (2004). Shown are translation and rotation times for two preparation durations (1 s and 5 s) each. H = Human; M = Model; T = Translation; R = Rotation; 1 and 5 indicate the duration of the preparation interval.

For simulating these experiments, we assumed that the sensorimotor representation, the bodily RF, and the imaginal RF are always present. Given the realization of the learning phase and the spatial layout of the experimental environment (see Fig. 1 in May, 2004), we also assumed the existence of an associated RF, which was—with equal probability—either aligned with the bodily RF or 45° misaligned with the bodily frame. We estimated the 3 free parameters of the model using the Metropolis algorithm (Madras, 2002) by fitting the model to response times and errors of both experiments across all conditions. Since the purpose of the simulation was to investigate the model's ability to account for key effects in the observed behavior, the objective of estimation was to maximize correlations between model and human behavior for response time and error for each of the two experiments (i.e., 4 correlations).

Model response times and errors correlated strongly with human times and errors for both experiments: $\rho = 0.91, \rho = 0.95, \rho = 0.94, \rho = 0.86$ for times and errors of Experiments 2 and 3, respectively. Model behavior is shown alongside human behavior for Experiment 2[1] in Figs. 3 and 4. As can

[1]The fit to Experiment 3 was very similar. For the sake of clarity the data from Experiment 3 are not included in the plot.

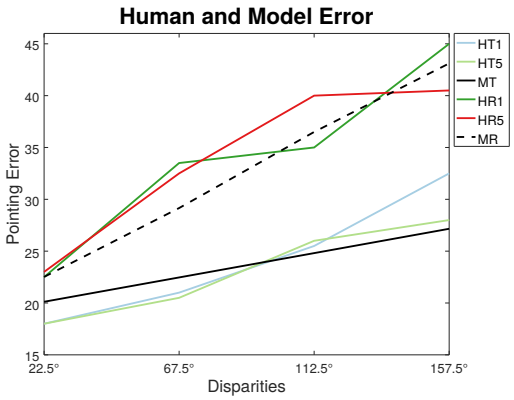


Figure 4: Human (colored lines) and model (black lines) pointing errors for Experiment 2 of May (2004). Shown are human errors for translation and rotation for two preparation durations (1 s and 5 s) each. Because model errors do not differ across different preparation durations, only rotations and translations are distinguished for model errors. H = Human; M = Model; T = Translation; R = Rotation; 1 and 5 indicate the duration of the preparation interval.

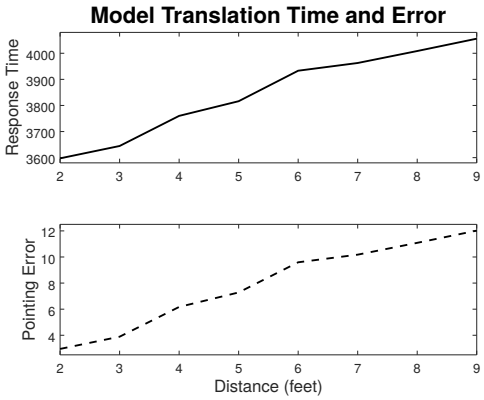


Figure 5: Model response times (solid line) and pointing errors (dashed line) when simulating Experiment 1 of Easton and Sholl (1995)

be seen from the plot, the model mirrors the main effects in the data well: (i) rotations are slower and more error prone than translations; (ii) response times and errors increase with disparity for both rotations and translations; (iii) preparation decreases the overall processing time, but does not substantially impact the disparity effect. Three further aspects of the simulation results seem noteworthy. First, humans show a stronger disparity effect for rotations than predicted by the model. Why humans should exhibit a stronger disparity effect for rotations than translations is currently unclear. Second, the model also captures that translations with short (1 s) preparation are slower than rotations with long (5 s) preparation. Third, the model correctly predicts that preparation time has no impact on error magnitude.

In sum, this first simulation lends further support to the assumptions of our theory in showing that a model based on the theory is able to closely mirror human behavior across a wide range of experimental conditions.

The second simulation will address a potential objection to our theory. Note that the theory makes no reference to mental transformations such as mental rotation or translation (e.g., Sholl, 2001). As a result the theory may seem to be at odds with findings indicating that PT time and error increases with the distance between the bodily and the imaginal perspective in translations (Easton & Sholl, 1995). Because no assumptions of our theory formulate an explicit relation between translation distance and PT performance, it seems an interesting question to what extent our theory can account for such a relationship. To address this question our second simulation models Experiment 1 of Easton and Sholl (1995).

Experiment 1 of Easton and Sholl (1995). In this experiment, participants first memorized an object layout, in which 8 objects were placed with varying distance (2 – 9 ft) from the location of the observer (see Fig. 1B in Easton & Sholl, 1995). After learning the layout, participants were provided with a target object and then a to-be-taken perspective. They were asked to point to the target object as quickly as possible. The main finding was that for translations pointing time and error increased significantly with increasing distance.

We assumed the same RFs (bodily, imaginal, associated) as for the first simulation. Because the bodily orientation was identical for learning and testing, we assumed that all three RFs were aligned for translations. For each object that indicated the position of a translated perspective, we estimated the pointing disparity when pointing to each of the other objects from Fig. 1B of Easton and Sholl (1995). Response times and errors for one translation perspective were computed as the average times / errors across pointing to all other objects from this perspective. Because the purpose of this simulation was to assess the theory's general ability to account for increasing times and errors with increasing distance, we did not fit the model to the human data, but reused the parameters estimated in the first simulation.

As can be seen from Fig. 5, the model nicely accounts for the effects observed by Easton and Sholl (1995): Both times and errors increase with increasing translation distance. Given that our theory makes no reference to mental transformations or distances, it may not be immediately clear why the theory correctly predicts the observed human behavior. It turns out, however, that—at least in Experiment 1 of Easton and Sholl (1995)—the average pointing disparity systematically increases with increasing translation distance. Since our theory assumes increased PT effort with increasing disparity, it predicts the increased effort for increased distance observed in this experiment.

The Algorithmic Level

The model described above was designed to capture the gist of the theory with as little implementational overhead as possible. Consequently, the model remains somewhat abstract and does not provide much detail on the mechanisms and representations underlying the observed behavior. In this section, we propose a more mechanistic realization of our theory.

The object-to-object representation may be realized by the type of representation structure proposed by Schultheis, Bertel, and Barkowsky (2014). This circular representation structure preserves the neighborhood relations between directions, but requires a direction root (i.e., a RF) to ground it in the real world. We further suggest that RF selection may proceed as a leaky, competitive, accumulative process as in RF selection for spatial term use (Schultheis & Carlson, 2017). Finally, we think that activation of a pointing response and interference from the bodily pointing response can appropriately be captured by dynamic field theory (Schöner, Spencer, & DFT Research Group, 2015, with different pointing directions activating different parts of the dynamic field) .

Such a realization has the twofold advantage of promising to capture the main assumptions of the theory while, at the same time, constituting a computational instantiation that is more solidly grounded in previous cognitive theorizing than the above-described model. To what extent such a realization is able to mirror pertinent human behavior will be subject of our future research.

Related Theories

Several theories of spatial memory have previously been proposed and all of them have highlighted important properties of how humans represent and recall spatial information of immediate environments (e.g., Avraamides & Kelly, 2008; Byrne et al., 2007; Mou et al., 2004; Sholl, 2001; Waller & Hodgson, 2006; Wang, 2017). However, none of the existing theories provides an explanation of all of the main findings highlighted above. For some findings the theories do not offer any explanation and for others, the theories' assumptions seem to be in contradiction with the findings. Because space restrictions do not permit a detailed critical appraisal of all theories, we restrict ourselves to a brief exemplary discussion of two of the theories.

Sholl (2001) assumes an orientation-free object-to-object representation that is subject to access by two egocentric RFs: a motor and a cognitive RF. The two frames usually coincide but can be separated. In imaginal PT the cognitive frame is assumed to be mentally rotated / translated to an appropriate place in the object-to-object representation. PT effort is assumed to be driven by the effort to separate the two frames and to mentally transform the cognitive frame. This theory has difficulties, for example, explaining why translation effort increases with pointing disparity and why rotations are generally more effortful than translations.

Mou et al. (2004) also assume an object-to-object representation. In contrast to Sholl (2001) and our theory, however, this representation is assumed to be oriented (i.e., tightly coupled to a RF). If the imaginal frame is aligned with the representation's frame, information can be directly retrieved from memory. If the frames are misaligned, the relation has to be inferred. The mechanisms underlying this inference are sometimes declared outside the scope of the theory (Rump & McNamara, 2013) and sometimes characterized as being some form of mental transformation (Mou et al., 2004, p. 156). In either case, the theory does not offer a satisfactory explanation of some of the key findings.

Conclusion

Our theory constitutes a promising account of spatial memory of immediate environments. As we have shown, the theory provides explanations for a wide range of key findings and a computational realization of the theory accounts well for human behavior in pertinent empirical studies. Moreover, the theory's view on spatial memory of immediate environments also fits well into frameworks of how larger-scale space representations are assembled as networks of more local representations of immediate environments (e.g., Chrastil & Warren, 2014; Meilinger, 2008).

Our theory suggests that enduring spatial memory representations may best be viewed as consisting of two main parts: a representation structure (e.g., the circular structure described above) and a RF. In particular, structure and RF may be flexibly combined such that the same structure / RF can yield different representations when combined with different RFs / structures. Such a view promises a more parsimonious account of spatial representations, because a comparatively small set of structures and frames may be sufficient to explain a wide range of spatial abilities. It also highlights an interesting possible connection between spatial language use and spatial reasoning through sharing RF selection mechanisms.

Future work will focus on refining the computational realization of the theory (see above) and on extending simulations to include further experiments.

Acknowledgments

The research reported in this paper has been partially supported by the German Research Foundation DFG, as part of Collaborative Research Center (Sonderforschungsbereich) 1320 EASE – Everyday Activity Science and Engineering (http://www.ease-crc.org/). The research was conducted in subproject P03 Spatial Reasoning in Everyday Activity.

References

Avraamides, M. N., & Kelly, J. W. (2008). Multiple systems of spatial memory and action. *Cognitive Processing*, *9*(2), 93–106.

Botvinick, M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, *108*(3), 624-652.

Brockmole, J. R., & Wang, R. F. (2003). Changing perspective within and across environments. *Cognition*, *87*, B59–B67.

Byrne, P., Becker, S., & Burgess, N. (2007). Remembering the past and imagining the future: A neural model of spatial memory and imagery. *Psychological Review*, *114*, 340–375.

Chrastil, E. R., & Warren, W. H. (2014). From cognitive maps to cognitive graphs. *PLOS ONE*, *9*(11), 1-8.

Cooper, R. P., & Guest, O. (2014). Implementations are not specifications: Specification, replication and experimentation in computational cognitive modeling. *Cognitive Systems Research*, *27*, 42 - 49.

Easton, R. D., & Sholl, M. J. (1995). Object-array structure, frames of reference, and retrieval of spatial knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(2), 483–500.

Madras, N. (2002). *Lectures on Monte Carlo Methods*. American Mathematical Society.

May, M. (2004). Imaginal perspective switches in remembered environments: Transformation versus interference accounts. *Cognitive Psychology*, *48*(2), 163 - 206.

May, M. (2007). Imaginal repositioning in everyday environments: effects of testing method and setting. *Psychological Research*, *71*(3), 277–287.

Meilinger, T. (2008). The network of reference frames theory: A synthesis of graphs and cognitive maps. In C. Freksa, N. S. Newcombe, P. Gärdenfors, & S. Wölfl (Eds.), *Spatial cognition vi. learning, reasoning, and talking about space* (pp. 344–360). Berlin: Springer.

Mou, W., McNamara, T. P., Valiquette, C. M., & Rump, B. (2004). Allocentric and egocentric updating of spatial memories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(1), 142–157.

Rump, B., & McNamara, T. P. (2013). Representations of interobject spatial relations in long-term memory. *Memory & Cognition*, *41*(2), 201–213.

Schöner, G., Spencer, J. P., & DFT Research Group, T. (2015). *Dynamic Thinking: A Primer on Dynamic Field Theory*. Oxford University Press.

Schultheis, H., Bertel, S., & Barkowsky, T. (2014). Modeling mental spatial reasoning about cardinal directions. *Cognitive Science*, *38*(8), 1521-1561.

Schultheis, H., & Carlson, L. A. (2017). Mechanisms of reference frame selection in spatial term use: Computational and empirical studies. *Cognitive Science*, *41*(2), 276-325.

Sholl, M. J. (2001). The role of a self-reference system in spatial navigation. In D. R. Montello (Ed.), *Spatial information theory* (pp. 217–232). Berlin: Springer.

Waller, D., & Hodgson, E. (2006). Transient and enduring spatial representations under disorientation and self-rotation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(4), 867–882.

Wang, R. F. (2005). Beyond imagination: Perspective change problems revisited. *Psicologica*, *26*, 25–38.

Wang, R. F. (2017). Spatial updating and common misinterpretations of spatial reference frames. *Spatial Cognition & Computation*, *17*(3), 222–249.

An Integrated Trial-Level Performance Measure: Combining Accuracy and RT to Express Performance During Learning

Florian Sense
f.sense@rug.nl
Department of Experimental Psychology & Behavioral and Cognitive Neuroscience
University of Groningen, Groningen, The Netherlands

Tiffany Jastrzembski, Michael Krusmark, Siera Martinez
{tiffany.jastrzembski, michael.krusmark.ctr, siera.martinez.ctr}@us.af.mil
Air Force Research Laboratory, Wright-Patterson Air Force Base, Dayton, OH, USA

Hedderik van Rijn
d.h.van.rijn@rug.nl
Department of Experimental Psychology & Behavioral and Cognitive Neuroscience
University of Groningen, Groningen, The Netherlands

Abstract

Memory researchers have studied learning behavior and extracted regularities describing learning and forgetting over time. Early work revealed forgetting curves and the benefits of temporal spacing and testing for learning. Computational models formally implemented these regularities to capture relevant trends over time. As these models improved, they were applied to adaptive learning contexts, where learning profiles could be identified from responses to past learning events to predict and improve future performance. Often times, past performance is expressed as accuracy alone. Here we explore whether a model's predictions can be improved if past performance is expressed by an integrated measure that combines accuracy and response times (RT). We present a simple, data-driven method to combine accuracy and RT on a trial-by-trial basis. This research demonstrates that predictions made using the Predictive Performance Equation improve when past performance is expressed as an integrated measure rather than accuracy alone.

Keywords: Learning; forgetting; cognitive model; accuracy; response time; integrated measure

Introduction

What data from fact learning trials are needed to predict whether a student will know the correct answer some time in the future? Does it help to know how often (and when) the student has previously answered correctly? Or how long it took them to provide the answer?

These questions are at the heart of models that describe learning and forgetting over time. Computational models are often fit to historical data to demonstrate that they can capture relevant behavioral effects exhibited by human learners (e.g. Pavlik & Anderson, 2008; Walsh, Gluck, Gunzelmann, Jastrzembski, Base, et al., 2018). Yet, the strongest test of a model is accurately predicting future performance—especially if predictions are made for each item studied by each student. The Second Language Acquisition Modeling (SLAM) challenge recently posed by Duolingo required such predictions (Settles, Brust, Gustafson, Hagiwara, & Madnani, 2018). Data from a subset of Duolingo users were made available and users submitted model performance predictions as part of a modeling competition..

As in these challenges, adaptive fact-learning systems must decide which features of the available data are taken into account to detect differences in item difficulty and participants' abilities to make accurate predictions. An obvious candidate is accuracy, since it indicates whether the student knew an answer previously. Forgetting then reduces the probability that responses are correct over time. Systems such as Duolingo (Settles & Meeder, 2016) strive to ensure that study repetitions occur *before* knowledge is forgotten.

Yet, if most responses are correct, there is very little information in the responses if only accuracy is considered, making it difficult to optimally adapt to learning and item difficulty. Response times (RT) can provide an additional source of information to differentiate between otherwise identical responses. The basic assumption is that observed RTs correlate with the difficulty of memory retrieval (e.g., Pavlik & Anderson, 2008; Pyc & Rawson, 2009). Indeed, analyses of the models submitted to the SLAM challenge support the view that RTs provide valuable information for predicting later performance (see Table 4 in Settles et al., 2018).

As accuracy and RT are often correlated (e.g., speed-accuracy trade-offs), methods have been proposed to combine them into a single performance metric. A recent suite of simulation studies discusses the merits of seven such integrated performance measures (Vandierendonck, 2017). All these measures, however, are aggregate measures: For example, the mean RT is combined with the average accuracy to express performance per participant, per condition. As this discards all information pertaining to *when* responses are given, these measures are less suited for parametrizing adaptive learning systems.

To our knowledge, there are at least two adaptive fact-learning systems that use both accuracy and RTs on a trial-by-trial level. Adaptive Response-Time-based Sequencing (ARTS; Mettler & Kellman, 2014; Mettler, Massey, & Kellman, 2016) schedules repetitions adaptively by continuously computing priority scores and presenting

the item with the highest priority. If the previous response was incorrect, that item's priority is increased drastically to ensure timely repetition. If the previous response was correct, however, the priority score is a function of the (log-transformed and scaled) RT associated with that response.

The second system is an extension of ACT-R's declarative memory module and uses the associated equations to approximate an item's memory strength (or "activation") through observed RTs (Pavlik & Anderson, 2008). Instead of using priority scores, items are repeated based on their estimated activation, a value that decreases over time (van Rijn, van Maanen, & van Woudenberg, 2009). Note that the observed RT of incorrect responses is replaced by a fixed, long RT, reflecting that it took "too long" for the correct response to be retrieved. These two examples demonstrate that combining information from accuracy and RTs is feasible in practice. Neither system really uses an *integrated* performance measure, however—they both use a transformation of RT that is conditional on accuracy.

Here, we will present an approach to computing an integrated, trial-level performance measure that combines accuracy and RT. Ideally, such a measure is purely data-driven, easy to interpret, computationally simple, and applicable to existing datasets. We are most interested in situations in which item-level data of the learning history are available and the goal is to validly predict future performance.

In the following, we will outline two datasets that we use as a test bed. Both datasets concern learning of paired associates, which provides a context in which the RT reflects relevant memory processes. We will demonstrate how our trial-level integrated performance measure can be computed for such data. Lastly, we describe how these integrated "Readiness" scores can be used as input to a computational model (the Predictive Performance Equation, Walsh, Gluck, Gunzelmann, Jastrzembski, Base, et al., 2018) to generate predictions based on past performance.

The central focus of this research seeks to explore whether use of this "Readiness score improves model predictions compared to scores that do not integrate accuracy and latency.

Methods

Datasets

We leverage two existing datasets, labelled WSU and TopiCS, to explore the idea of an integrated, trial-level performance measure. Each dataset consists of a study and a test phase. Trial-level information for response accuracy and RT is available for both datasets but they vary drastically in the structuring of the study phase and in the time between study and test. Importantly, accuracy during study is very high in both datasets (85.9% in WSU and 89.8% in TopiCS).

Washington State University (WSU) data WSU data is part of an (as of yet) unpublished multi-day fatigue study. Participants spent four days in a sleep lab and completed a battery of tests throughout that period. Here, we will focus on the paired-associates learning data of 36 participants who were not withheld any sleep (the control group). Fifty-one nonsensical line drawings—e.g.,  —were used as cues and participants learned two-digit numbers—e.g., "79"—as a response. Each paired associated was repeated 20 times according to different presentation schedules.

Table 1: Number of repetitions of an item at each test moment depending on the schedule in the WSU data. RI = retention interval between the last encounter of an item and the test.

	Study phase					Test phase
Schedule	9am	1pm	3pm	7pm	9pm	9am
Spaced	4	4	4	4	4	2 (36h RI)
Massed early	20	·	·	·	·	2 (48h RI)
Massed late	·	·	·	·	20	2 (36h RI)

We will focus on three schedules that distributed the 20 repetitions across a single day. Table 1 shows that the *spaced* schedule distributed the 20 repetitions equally among five study periods throughout the day (four repetitions each), while the *massed* schedules presented each item 20 times either *early* or *late* in the day. The test phase featured two repetitions for each paired associate, and the retention interval (RI; i.e., the temporal space between the last encounter of an item and the test) depended on which study schedule the paired associate was assigned to. Each participant studied with all schedules and encountered three unique paired associate per schedule, resulting in 6,156 observations from the study phase that were used to make predictions for the 648 observations from the test phase.

The recorded RT corresponds to the first key press. If participants did not respond within 6 seconds, the trial was recorded as incorrect (with RT set to 6 sec, hence the spike in the lower right panel of Figure 1A).

TopiCS data The TopiCS data were taken from Sense, Behrens, Meijer, and van Rijn (2016), published in Topics in Cognitive Science. Participants completed three sessions of two blocks each (six total). In each block, material was studied for 20 minutes using an adaptive fact-learning system (van Rijn et al., 2009), followed by a five-minute distractor task (Tetris), followed by a test of the studied material. Here, we will only use the first block of each session. In each of these three blocks, participants studied Swahili-English vocabulary word pairs. Each Swahili block featured 25 unique paired associates.

A total of 50,665 responses are available from 67 participants. Since the introduction and repetition schedules of items during study were governed by an adaptive model, these data do not have the controlled temporal structure of the WSU data: The number of repetitions as well as their timing varied between items and participants. The test was the same for everyone, however. After a five-minute delay, participants

were tested on all 25 potential Swahili cues at the end of each block and accuracy was recorded (4,965 observations).

The study phase was entirely self-paced and RTs correspond to the first key press recorded after a Swahili word appeared on screen. The RT distributions, split by accuracy, are shown in Figure 1A.

Computing integrated "Readiness" scores

The goal is to derive a trial-level, quasi-continuous performance metric from accuracy and RT data. This "Readiness" value can take any value between 0 and 1. Values closer to 1 correspond to a correct, fast response. There are two versions of the metric: For R_0, all incorrect responses are treated equally and set to 0. For R_c, incorrect responses are transformed such that faster incorrect responses are more severely penalized (i.e., closer to 0) than slow incorrect responses (cf. Klinkenberg, Straatemeier, & Van der Maas, 2011). The term "Readiness" is used because values close to the 1 indicate that a response was readily available, resulting in a fast RT. Overall, the higher the "Readiness" value, the better the performance. In the following, we will detail how R_0 and R_c are computed from behavioral data.

Figure 1A depicts the distribution of RTs for correct (top panels) and incorrect (bottom panels) responses from the two datasets. For a more precise depiction of the data, the axes across the panels vary and only RTs faster than 15*s* are shown for the TopiCS data (99% of all observations). Since the vast majority of responses were correct, there are fewer RTs for the incorrect responses in the bottom panels.

Figure 1B makes the mapping from observed RTs to the probability of a correct response explicit: In both datasets, the log-transformed RTs (in ms) are strong predictors of accuracy, such that slower RTs reduce the probability of a correct response. The exact mapping differs in the two datasets: Responses are generally faster in the WSU data and time out after 6 seconds. The mapping is expressed as the two coefficients estimated by a simple logistic regression, which is $\beta_0 + \beta_1 \cdot \log(\text{RT})$. For the WSU data, β_0 is 24.26 and β_1 is -3.09. For the TopiCS data, β_0 is 12.99 and β_1 is -1.39. All four coefficients differ significantly from 0 with $p < 0.001$.

The relationship shown in Figure 1B provides the quantitative basis for the "Readiness" metrics. The mapping provided by the logistic regression allows an unbound performance metric (RT) to be transformed to a continuous metric with range [0, 1].

For the first metric, R_0, all correct responses are transformed using the mapping provided by the logistic regression coefficients. *Incorrect* responses are treated as performance of 0 (as with accuracy; hence the subscript 0). Using this approach, a correct response given quickly is considered "more correct" than a correct response given after longer deliberation, which is in line with behavioral data and theoretical assumptions (Pavlik & Anderson, 2008). Numerically, R_0 is computed by taking the inverse logit (L^{-1}) of the regression formula shown above, using log-transformed RTs (in *ms*) when accuracy (A) is 1:

$$R_0 = \left\{ \begin{array}{ll} A = 0: & 0 \\ A = 1: & L^{-1}(\beta_0 + \beta_1 \cdot \log(\text{RT})) \end{array} \right\} \quad (1)$$

The second "Readiness" metric, R_c, assumes that latencies for incorrect responses are informative too. Specifically, the assumption is that a fast incorrect response is *worse* than an incorrect response given after longer deliberation, a notion also present in other learning systems (e.g., Math Garden—an adaptive, online arithmetic-learning environment used by many schools in the Netherlands—formalized the same idea in the "high speed, high stakes" scoring rule; Klinkenberg et al., 2011, see section 2.3.3. and Fig. 2 specifically). Numerically, this is formalized by using the same approach as for correct responses but then subtracting 1 and taking the absolute value[1]:

$$R_c = \left\{ \begin{array}{ll} A = 0: & |L^{-1}(\beta_0 + \beta_1 \cdot \log(\text{RT})) - 1| \\ A = 1: & L^{-1}(\beta_0 + \beta_1 \cdot \log(\text{RT})) \end{array} \right\} \quad (2)$$

For example, a correct response with an RT of 1,834*ms* would result in the same R_0 and R_c scores but they would depend on the dataset the response was observed in. In the WSU data, the "Readiness" score would be 0.739 ($L^{-1}(24.26 - 3.09 \cdot \log(1,834))$ but in the TopiCS data it would be higher ($L^{-1}(12.99 - 1.39 \cdot \log(1,834) = 0.927)$ because RTs are generally longer, which results in a different mapping (cf. Figure 1A and B). If the RT is the same but associated with an *incorrect* response, the R_0 score is simply 0 (see Eq. 1). The R_c score, on the other hand, would be 0.261 (i.e., $|0.739 - 1|$) in the WSU and 0.073 (i.e., $|0.927 - 1|$) in the TopiCS data (see Eq. 2).

Figure 1C gives an overview of the R_c values computed in the two datasets, split again by accuracy and dataset. Note that the y-axes differ due to the unequal number of observations. For both datasets, the correct responses (top panels) mostly have values between 0.75 and 1. The incorrect responses (bottom panels) are more spread across the range for the R_c metric[2]. For the TopiCS data, most incorrect responses have R_c values between 0 and 0.5 but mostly values are <0.25. In the WSU data, the values are spread more widely. The distributions of correct and incorrect responses barely overlap within a dataset, though (note the small numbers on the y-axis for incorrect reponses from the WSU data). For the R_0 metric, all values corresponding to incorrect responses are simply 0 (cf. Eq. 1).

Taken together, the approach outlined here has multiple advantages. A binary and an unbound performance metric (accuracy and RT) are combined into an integrated, trial-level measure that is continuous and bound between 0 and 1. Importantly, a trial-level performance metric preserves

[1] This could be thought of as flipping the mapping in Figure 1B along the horizontal axis at 0.5.

[2] For the WSU data, timed-out observations with RTs of 6*s* (N = 35) were transformed to the fastest observed RT (391*ms*) to make them "very wrong". If these observations are simply dropped, none of the reported results change qualitatively.

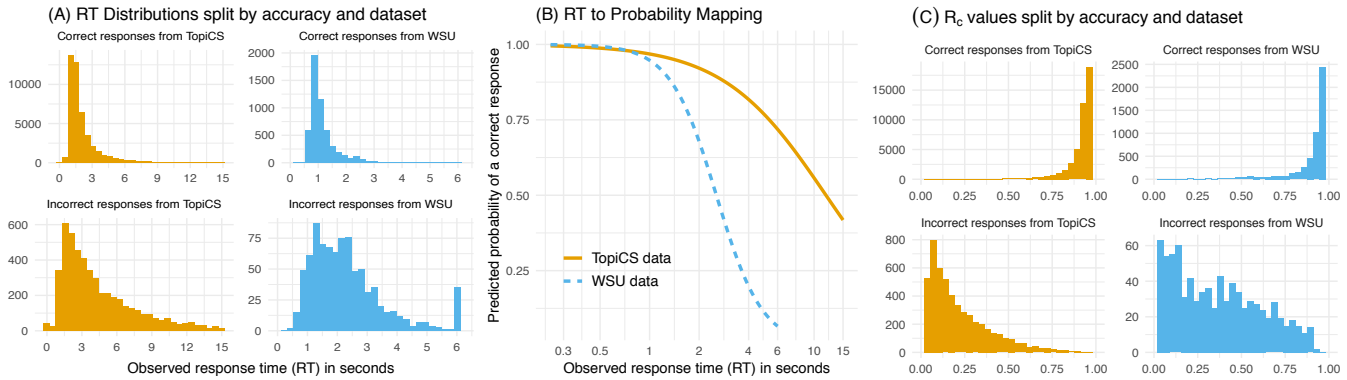


Figure 1: A: The observed RT in seconds split by accuracy and dataset; B: Logistic regression lines showing the mapping from observed RTs in seconds to the probability of a correct response; C: The continuous "Readiness" (R_c) values computed using the mapping in B, split by accuracy and dataset. In all plots, color indicates the dataset. See text for additional details.

information about the timing of individual encounters that would be lost if one simply computed, for example, the mean accuracy during study. The approach is also computationally extremely simple and makes minimal theoretical assumptions that are easily checked: Are longer RTs associated with a lower probability of giving an accurate answer? If the logistic regression's slope coefficient (β_1) does not differ significantly from 0, computing "Readiness" values is probably not sensible. Visual checks akin to Figure 1B and C also provide easy sensibility checks. Finally, the interpretation of "Readiness" values is straightforward: Higher values indicate better performance and values at the boundaries indicate faster responses.

Predictive Performance Equation (PPE)

To explore whether the "Readiness" scores are useful, we will explore their utility as input to the Predictive Performance Equation (PPE), a computational model developed to capture individual differences in learning and forgetting (for an extended description of the model see Walsh, Gluck, Gunzelmann, Jastrzembski, Base, et al., 2018). If the scores expressed meaningful individual differences, a computational model should more closely mimic a participant's learning and forgetting process than when other scores are used. The end result would be more accurate predictions of future performance based on past performance.

In two recent studies, PPE was compared to other models to test "the theoretical adequacy and applied potential of computational models" more generally (Walsh, Gluck, Gunzelmann, Jastrzembski, & Krusmark, 2018) and to shed light on "the mechanisms underlying the spacing effect in learning" specifically (Walsh, Gluck, Gunzelmann, Jastrzembski, Base, et al., 2018). Due to space constraints, we will keep the current description of the model mechanics brief and refer the interested reader to those papers for a detailed overview.

The PPE component we are ultimately interested in is the predicted performance, P, which is a logistic function of activation (M) that has two free parameters, τ and s:

$$P = \frac{1}{1 + \exp(\frac{\tau - M}{s})} \tag{3}$$

The activation M is the product of learning and forgetting, expressed as $N^{0.1} \cdot T^{-d}$. The learning term increases exponentially as a function of the number of repetitions (N) and the forgetting term decreases exponentially. The latter has two components: The elapsed time (T) is the weighted sum of the time since each previous repetition (see Eq. 3 and 4 in Walsh, Gluck, Gunzelmann, Jastrzembski, & Krusmark, 2018) and the decay rate (d), which has free intercept (b) and slope (m) parameters and is a function of the lag between consecutive repetitions:

$$d = b + m \cdot \left(\frac{1}{n-1} \cdot \sum_{j=1}^{n-1} \frac{1}{\ln(\text{lag} + e)} \right) \tag{4}$$

Model fitting In the form outlined above, PPE has four free parameters (b, m, τ, and s) and requires two pieces of information to be fit: The time point of each repetition (to compute T and d) and the observed performance at each time point. The model is agnostic with regards to what the performance metric represents and only requires it to fall in the range of [0, 1], as the "Readiness" measure provides. The best-fitting parameters are found by minimizing the error between the supplied performance metric and the predicted performance P (see Eq. 3) produced by a given combination of the free parameters. The error is defined as the summed squared error between the performance metric and P across the data available for each unique participant-item combination.

Here, we only vary the performance metric that is used during model fitting, using either accuracy, R_0, and R_c. All other factors—free parameters, allowed parameter ranges[3],

[3]The ranges for the free parameters are b = [0, 0.5], m = [0, 0.5],

and timing-related information—are held constant.

Results

For both datasets, we determined the best-fitting PPE parameters for each participant-item combination, and then computed item-level predicted performance, P, on the test. The model fit will be evaluated for the predictions—i.e., comparing predicted P with recorded accuracy—rather than fit to the study data because we consider the ability to predict future performance given historical data most relevant.

For each dataset, PPE was fit three times, using the three performance metrics outlined above: Accuracy, which is binary; R_0, which is continuous for correct responses but all incorrect responses are 0 (see Eq. 1); and R_c, which is continuous for both correct and incorrect responses (see Eq. 2 and Figure 1C). The main results of the comparison are presented in Table 2, which lists two model fit statistics for each performance metric. Also included is a baseline, which simply predicts that all responses during the test are correct.

In both datasets, performance on the test is expressed as accuracy. PPE, on the other hand, predicts the probability of a correct response. To evaluate the model predictions, we use fit statistics commonly used when evaluating performance in binary classification problems. The fit statistics are: (1) The area under the receiver operating characteristic curve (*AUC*), which can be interpreted as the probability of a randomly drawn correct response outranking (i.e., having a higher P value) a randomly drawn incorrect response. Note that the baseline condition, in which all responses are predicted to be correct, would result in an AUC of 50%. (2) *Log loss*, expressing the accuracy of a classifier by penalizing inaccurate classifications. The AUC measure can range from .5 to 1, higher values are better. Log loss is unbound and lower values are better. In Table 2, the best-performing metric is highlighted in bold for each fit statistic.

Table 2: Fit statistics for predictions made in the two datasets. The baseline predicts that all responses on the test are correct.

Dataset	Statistic	Baseline	Accuracy	R_0	R_c
WSU	AUC	0.500	0.687	0.768	**0.769**
	Log loss	19.635	6.170	1.900	**1.141**
TopiCS	AUC	0.500	0.679	0.712	**0.755**
	Log loss	1.298	3.110	1.701	**0.638**

Table 2 shows the fit statistics for the 648 predictions made in the WSU data. All three measures outperform the baseline. Of these, using accuracy as performance measure scores lowest, and there is no clear difference between the two "Readiness" scores. This impression is confirmed by statistical comparison of the AUC values, which tests the null hypothesis that the difference between two AUCs is 0 against the alternative hypothesis that it is not (DeLong, DeLong, & Clarke-Pearson, 1988). The tests yield significant differences between the accuracy- and R_0-based AUCs (z = -4.449; $p < 0.001$) and accuracy- and R_c-based AUCs (z = -3.850; $p < 0.001$) but not between R_0- and R_c-based AUCs (z = -0.063; $p = 0.950$). The log loss is very high for the baseline because the actual accuracy on the test was only 45.5%, resulting in a high penalty.

In the TopiCS data, on the other hand, the observed accuracy on the test was extremely high: 96.4% of the 4,965 responses were correct. Thus, the all-correct baseline gets less than 4% of the predictions wrong, resulting in a relatively low log loss value. Only the R_c score yields predictions that result in a lower log loss value than the baseline. Regarding the AUC values, all predictions derived from the computational model outperform the baseline. The statistical test for the comparison of the accuracy- and R_0-based AUC is inconclusive (z = -1.451; $p = 0.147$), while the R_c-based predictions are significantly better than both the accuracy- (z = -2.809; $p = 0.005$) and R_0-based predictions (z = -2.148; $p = 0.032$).

Discussion

Here, we explored the predictive power of an integrated performance measure that combines accuracy and RT information. Unlike aggregate measures (see Vandierendonck, 2017, for an overview), the "Readiness" scores presented here are computed for each observation individually. Using two datasets, we demonstrated how "Readiness" scores are computed. The practical utility of the resulting integrated performance measures was demonstrated by fitting a computational model to past performance in order to predict future performance. Statistical analyses reveal evidence that predictions are more accurate when past performance was expressed as a "Readiness" score rather than accuracy.

We present two variations of the "Readiness" score that differ in how they treat incorrect responses. The R_0 score regards all incorrect responses equally, setting them to 0 (analogously to accuracy). The continuous score, R_c, scales both correct *and* incorrect responses (see Figure 1C) such that fast incorrect responses are considered worse than slow incorrect responses. Both versions express performance as scores between 0 and 1, with higher values indicating better performance. For R_c, scores closer to either boundary correspond to responses that were given quickly.

Whether R_0 or R_c should be preferred—or whether either should be used—depends on the context and the assumptions the researcher can make, especially regarding incorrect responses. Since the "Readiness" scores are based on empirical data, the data can provide an immediate check. If the slope of the logistic regression model that provides the mapping (cf. Figure 1B) does not significantly differ from 0, the crucial assumption that observed RTs and accuracy are associated is violated and "Readiness" scores are probably

τ = [0, 1], and s = [0, 0.1].

not meaningful. As discussed in the Methods section, visualizations such as those in Figure 1B and C can also inform the researcher's choice. In a very large dataset, for example, the logistic regression model might have a significant but very small slope coefficient, resulting in a mapping (cf. Figure 1B) for which even very slow RTs result in near-ceiling performance, which would in turn yield R_c values that are quasi-equivalent to accuracy.

Exploring to which extent "Readiness" scores could be a useful expression of past performance in different contexts would be a logical extension of the current work, which presents an initial exploration of the idea in two relatively small datasets. This first exploration is promising, however, given that even though both datasets differed in a number of important aspects, the "Readiness" measure outperformed accuracy in both. Most importantly, the retention intervals differed dramatically (five minutes in the TopiCS data and 36–48 hours in the WSU data), which meant that test performance was near-perfect in the TopiCS data and lower than 50% in the WSU data. Another possible extension of the current work would be to investigate the utility of "Readiness" scores in computational models other than PPE.

In conclusion, we present a simple, data-driven way to combine accuracy and response time information into an integrated, trial-level performance measure that we call "Readiness." This approach makes minimal assumptions that are easy to check and resulting performance scores are easy to interpret. This research demonstrates that a single computational model can capture the general learning and forgetting patterns observed across two very diverse sets of paired associate learning data, and that the model's predictive validity is enhanced when past performance is expressed in terms of an integrated "Readiness measure, rather than use of simple accuracy alone.

Acknowledgements

FS was supported by L3 Technologies through the Air Force Research Laboratory at Wright-Patterson Air Force Base.

References

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics*, *44*(3), 837–845. doi: 10.2307/2531595

Klinkenberg, S., Straatemeier, M., & Van der Maas, H. L. J. (2011). Computer adaptive practice of Maths ability using a new item response model for on the fly ability and difficulty estimation. *Computers & Education*, *57*(2), 1813–1824. doi: 10.1016/j.compedu.2011.02.003

Mettler, E., & Kellman, P. J. (2014). Adaptive response-time-based category sequencing in perceptual learning. *Vision Research*, *99*, 111–123. doi: 10.1016/j.visres.2013.12.009

Mettler, E., Massey, C. M., & Kellman, P. J. (2016). A Comparison of Adaptive and Fixed Schedules of Practice. *Journal of Experimental Psychology: General*, *145*(7), 897–917. doi: 10.1037/xge0000170

Pavlik, P. I., & Anderson, J. R. (2008). Using a Model to Compute the Optimal Schedule of Practice. *Journal of experimental psychology. Applied*, *14*(2), 101–117. doi: 10.1037/1076-898X.14.2.101

Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language*, *60*(4), 437–447. doi: 10.1016/j.jml.2009.01.004

Sense, F., Behrens, F., Meijer, R. R., & van Rijn, H. (2016). An Individual's Rate of Forgetting Is Stable Over Time but Differs Across Materials. *Topics in Cognitive Science*, *8*(1), 305–321. doi: 10.1111/tops.12183

Settles, B., Brust, C., Gustafson, E., Hagiwara, M., & Madnani, N. (2018). Second Language Acquisition Modeling. In *Thirteenth workshop on innovative use of nlp for building educational applications* (pp. 56–65).

Settles, B., & Meeder, B. (2016). A Trainable Spaced Repetition Model for Language Learning. *Association for Computational Linguistic (ACL)*, 1848–1858.

Vandierendonck, A. (2017). A comparison of methods to combine speed and accuracy measures of performance: A rejoinder on the binning procedure. *Behavior Research Methods*, *49*, 653–673. doi: 10.3758/s13428-016-0721-5

van Rijn, H., van Maanen, L., & van Woudenberg, M. (2009). Passing the Test: Improving Learning Gains by Balancing Spacing and Testing Effects. In *Proceedings of the 9th international conference on cognitive modeling* (pp. 110–115). Manchester, UK.

Walsh, M. M., Gluck, K. A., Gunzelmann, G., Jastrzembski, T., Base, F., Myung, J. I., … Zhou, R. (2018). Mechanisms Underlying the Spacing Effect in Learning: A Comparison of Three Computational Models. *Journal of Experimental Psychology: General*, *147*(9), 1325–1348. doi: 10.1037/xge0000416

Walsh, M. M., Gluck, K. A., Gunzelmann, G., Jastrzembski, T., & Krusmark, M. (2018). Evaluating the Theoretic Adequacy and Applied Potential of Computational Models of the Spacing Effect. *Cognitive Science*, *42*, 644–691. doi: 10.1111/cogs.12602

Patterns of coordination in simultaneously and sequentially improvising jazz musicians

Matt Setzler (msetzler@iu.edu)
Cognitive Science Program, 1900 E. 10th St.
Bloomington, IN 47406 USA

Rob Goldstone (rgoldsto@indiana.edu)
Department of Psychological and Brain Sciences, 1101 E 10th St
Bloomington, IN 47405 USA

Abstract

In Joint Action (JA) tasks, individuals must coordinate their actions so as to achieve some desirable outcome at the group-level. Group function is an emergent outcome of ongoing, mutually constraining interactions between agents. Here we investigate JA in dyads of improvising jazz pianists. Participants' musical output is recorded in one of two conditions: a *real* condition, in which two pianists improvise together as they typically would, and a *virtual* condition, in which a single pianist improvises along with a "ghost partner" – a recording of another pianist taken from a previous *real* trial. The conditions are identical except for that in *real* trials subjects are mutually coupled to one another, whereas there is only unidirectional influence in *virtual* trials (i.e. recording to musician). We quantify ways in which the rhythmic structures spontaneously produced in these improvisations is shaped by mutual coupling of co-performers. Musical signatures of underlying coordination patterns are also shown to parallel the subjective experience of improvisers, who preferred playing in trials with bidirectional influence despite not explicitly knowing which condition they had played in. These results illuminate how mutual coupling shapes emergent, group-level structure in the creative, open-ended and fundamentally collaborative domain of expert musical improvisation.

Keywords: Joint Action; Music; Improvisation; Complex Dynamical Systems; Situated Cognition

Introduction

Joint action (JA) is a fundamental facet of human life. From the earliest infant-caregiver interactions to the subtle give and take of salsa dancers, we very often coordinate our actions with others (Sebanz, Bekkering, & Knoblich, 2006). In such endeavors, group success has less to do with individual efforts considered in isolation, and more to do with the ability of individuals to successfully coordinate with one another. Understanding behavior in these settings requires shifting the unit of analysis up from the individual to the group level, as collective behavior emerges out of the ongoing interactions among individual agents (Goldstone & Gureckis, 2009).

The past decade has seen a proliferation of research investigating JA in collaborative music performance (Palmer & Zamm, 2017). Music has long been recognized as a rich and meaningful domain for cognitive science. It is a central facet of all human cultures, and music performance demands the simultaneous engagement of a variety of cognitive, emotional and perceptual-motor processes (Pearce & Rohrmeier, 2012). The richness and complexity of music increases still further when we consider collaborative musical performance, where all of these intra-individual processes must be aligned and coordinated amongst an ensemble of interacting musicians in service of a joint musical expression.

JA research has begun to elucidate how musicians meet these collaborative performance demands by examining the role of anticipatory auditory imagery in enabling performers to integrate their actions with one another, and how mutual coupling and leader-follower structures within ensembles facilitate musicians' ability to synchronize and fluidly change tempos (Chang, Livingstone, Bosnyak, & Trainor, 2017; Goebl & Palmer, 2009; Keller & Appel, 2010).

Joint Action in Improvised Music

Most of the work on music JA has taken place in the context of composed music, whereas very little has been done to examine JA in improvised music. JA in improvised music is a relatively neglected topic, and constitutes a uniquely rich and promising domain for examining joint action and complexity which is especially relevant to cognitive science.

When improvising musicians perform together they collectively generate abstract musical structures – rhythm, melody, harmony and sometimes even long-term song structures. In composed music, musicians must coordinate in terms of expressive parameters (like volume, tempo and articulation) but the abstract structure of the music is given *a priori* by the composer. The domain of interpersonal coordination in improvised music extends beyond these expressive parameters and into the formal architecture of the music. Abstract musical structures emerge out of ongoing interactions among improvisers. These interactions are nonlinear, mutually constraining and have the potential to evolve over time.

In many ways JA in improvised music is more closely aligned with other everyday JA situations than is performance of scored music. Improvisation is the norm in our daily life – group problem solving, scientific collaboration, and most of our conversations are improvised. It is actually quite rare that we perform scripted activities with others (composed music, religious ceremonies and theater performances are exceptional in this regard). Given the ubiquity of improvisation in everyday life, we might well expect some aspects of collaborative improvised music to generalize to other areas of cognition.

Despite the paucity of research in this area, some efforts to understand JA in improvised music have begun. In a notable example, coordination in jazz piano duos was analyzed as a

function of musical context (Walton et al., 2018). In the experiment, dyads of jazz pianists were studied improvising together over a swing backing track and a drone (sustained tone with no rhythmic structure). The authors performed Cross Recurrence Quantification Analysis on recordings of musicians' body movements as well as recordings of their musical output. CRQA revealed that pianists spontaneously engaged in different patterns of interpersonal coordination depending on which musical setting they were performing in.

In the current study we directly examine the effects of mutual coupling in improvised music by experimentally manipulating interaction in dyads of professional jazz pianists. Specifically, we recorded pianists improvising in one of two conditions: a *real* condition, in which two pianists improvised together as they typically would, and a *virtual* condition, in which one pianist improvised along with a "ghost partner" – a recording of another pianist taken from a previous *real* trial. In the *real* condition pianists are mutually coupled in the sense that they have the ability to respond to one another in ongoing feedback loops. Such mutual coupling is absent in the *virtual* condition – live pianists have the ability to respond to the recording, but the recording will never respond to the live musician. This feature also makes virtual recordings a nice ground-truth for assessing leader-follower roles. Subjects were blind to which condition they played in, and their musical output was recorded in the form of isolated MIDI tracks[1].

How does the presence of mutual coupling influence the music jointly produced by an ensemble of improvising musicians? This question is addressed by quantitatively comparing rhythmic structures spontaneously generated in *real* performances against *virtual* performances. Notably, these performances were obtained from elite professional pianists from the New York City jazz scene. These are individuals who have dedicated their lives to mastering their instruments and the ability to fluidly interact with others in improvised performance. Our subjects improvised freely, without any specific instructions or musical constraints (other than the implicit constraints imposed by manipulating interaction). The current study thus represents an ecologically valid and scientifically grounded approach to studying JA and mutual coupling in the creative, open-ended and fundamentally collaborative domain of expert musical improvisation.

Methods

Participants

16 professional jazz pianists from the New York City music scene participated in this study. Participant age ranged from 23-35. On average participants had 22 years experience playing piano (sd=4) and 17 years experience improvising (sd=5). All participants received formal training in piano performance and jazz studies at elite conservatories. None of our subjects had prior experience performing with one another.

Apparatus

Two MIDI-enabled keyboards were used: a Roland Juno-Di and Nord Electro 2, both of which had 61 semi-weighted keys. Both keyboards were used on every trial (i.e. virtual trials were arranged such that the live pianist played whatever keyboard their ghost partner did *not* play). Ableton Live 9 Lite (running on a MacBook Air) was used to collect isolated MIDI recordings for each musician. Ableton was also used to synthesize the audio participants heard, which allowed us to ensure time alignment of MIDI recordings, and that participants heard the same exact timbre for themselves and their partner, irrespective of condition. Participants were recorded at a music rehearsal studio in Brooklyn, NY. The studio was divided by a curtain such that participants could not see one another. Participants listened to themselves and their partners through Sony CH700N Noise Cancelling headphones. Thus, from the participants' perspective there was no visual or audible indication of their condition on a given trial.

Procedure

This study employed a within-subjects design, in which each musician played at least 3 trials[2] in both real and virtual conditions (Figure 1). Participants played with the same 'live' partner for each of their real trials and the same 'ghost' partner for each of their virtual trials. Altogether, 32 (128 minutes, 105,766 notes) trials were collected from 9 real pairs and 27 trials (108 minutes, 84,439 notes) were collected from 16 virtual pairs. To control for order effects, conditions were interleaved throughout the course of a session, and sessions were counterbalanced such that the order reversed every other session.

Participants were brought into the studio in pairs, and instructed to improvise a series of short (4-7 minute) duos. These improvisations were 'free', with no accompanying stimuli and no *a priori* musical template or constraints. Other than the suggested timeframe, the only instruction musicians were given was to do their best to improvise a compelling piece of music, as they would in a typical performance setting.

Subjects were told they would be improvising in one of the two conditions (real or virtual), but on any given trial they were not told their condition. At the start of each trial each participant was privately instructed to Play or Don't Play. At the conclusion of each trial (when the musicians had finished improvising) each player was asked to fill out a short questionnaire that had them rate the previous performance in terms of: (1) how easy it was to coordinate with their partner (2) how well coordinated they were with their partner (3) quality of the improvised piece and (4) to what degree they played a supporter or a leader role.

[1] MIDI is a format for representing music on a computer. It symbolically records the pitch, volume and timing (onset and offset) of every note played

[2] Subjects played more trials if time permitted.

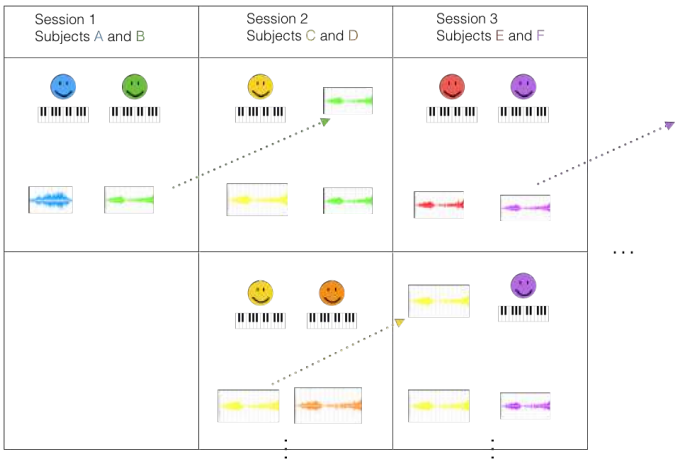


Figure 1: Sequencing of real and virtual trials. Each subject played multiple trials in each condition (repetition of trials not shown in abbreviated figure). Subjects were paired with same partner in all real trials, and a separate ghost partner for all virtual trials. Ghost recordings were taken from real trials of previous sessions, ensuring that the live musician had never heard them before.

Results and Discussion

MIDI data was collected for 32 real trials and 27 virtual trials. Each trial consists of two MIDI recordings, one for each individual (the same MIDI recordings in real trials were used as ghosts in virtual trials). 105,766 improvised note onsets were collected in real trials and 84,439 improvised note onsets were collected in virtual trials. Over 11 hours of music was collected in total. We also collected subjective ratings of participants after every trial they performed.

Figure 2 shows participants' responses to the questionnaire they were given at the conclusion of each trial. Despite the fact that participants were blind to which condition they were in in a given trial, their ratings differed systematically as a function of condition. Overall, subjects rated real trials to be of higher quality than virtual trials (paired $T(df)=15$, $p<.01$). Real trials were also generally rated as being characterized by better inter-musician coordination (paired $T(df)=15$, $p<.01$) and ease of collaboration (paired $T(df)=15$, $p<.01$).

Subjects were also asked to rate the degree to which they felt they played a leader or supporter role, which also revealed a main effect of condition (paired $T(df)=15$, $p<.05$). As expected, participants felt they mostly played a supporter role in virtual trials (in which they were playing with an unresponsive recording), whereas participants neither identified with leader or follower roles in real trials. This last result could indicate multiple things. One possibility is that musicians felt they played an equally leading and supporting role throughout the course of the performance. Alternatively, it could be that leadership roles shifted throughout the course of improvised performances. More data would be needed to differentiate between these possibilities, but in informal conversation with subjects they often alluded to the latter. At the very least, time-evolving leadership dynamics were achievable in real trials characterized by mutual coupling, but not in virtual trials characterized by unidirectional influence.

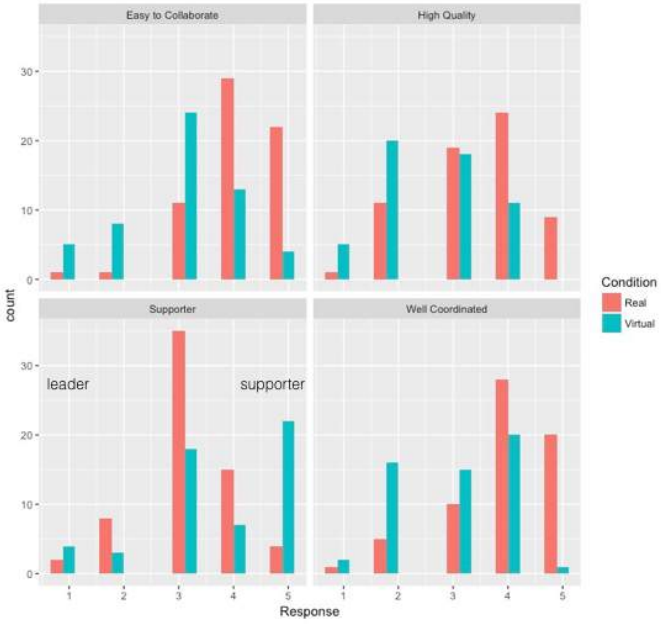


Figure 2: Subjective ratings by condition. Despite being blind to condition, participants generally rated *real* trials to be of higher quality (top right) and characterized by better inter-musician coordination (bottom right) and ease of collaboration (top left) as opposed to *virtual* trials. Participants felt they played more of a supporter role in *virtual* trials, and generally did not identify with either a leader or supporter role on *real* trials (bottom left).

Onset Analysis

MIDI recordings contain a wealth of musical information: rhythmic structure (timing of note onsets and offsets), volume, and tonal structure (sequential pitch information). In expert improvisation, interpersonal coordination occurs in each of these musical dimensions. However, we initially analyzed one clear and unambiguous aspect of the data – timing of note onsets. Timing of note onsets is a good starting point for investigating inter-musician coordination because it is simple to analyze but encapsulates an essential musical component – rhythmic structure.

Synchronization A central challenge in collaborative music making is synchronization. Musicians playing together often need to align their note onsets to occur simultaneously. Previous work has demonstrated that in composed musical settings, piano dyads' synchronize more effectively when they are mutually coupled to another another than in experimental manipulations in which auditory feedback was removed (Demos, Carter, Wanderley, & Palmer, 2017; Goebl & Palmer, 2009). It has also been demonstrated that musical

leaders play onsets of nominally simultaneous notes (notes occurring at the same metrical positions in a written score) slightly before followers (Goebl & Palmer, 2009).

To what degree does mutual coupling facilitate synchronization in improvised music? Without a written score it is difficult to assess this question, because there is no 'ground truth' for when and whether improvisers are trying to synchronize. Nonetheless, we approached the question by identifying near-simultaneous onsets, those occurring within 100ms of one another, played by co-performers. Degree of synchrony can be assessed by looking at the magnitude of asynchronies (henceforth 'asyncs') by which near-simultaneous onsets were displaced from one another. While we cannot be certain whether improvisers were explicitly trying to synchronize, this metric gives us insight into how precisely synchronization occurred spontaneously, as a joint outcome of our subjects' sensibilities and the affordances of inter-musician coupling.

Figure 3 displays the magnitude of onset asyncs colored by experimental condition. Asyncs are more peaked around 0 for real trials compared to virtual trials – indicating that when co-performers did synchronize, they did so more precisely in real conditions compared to virtual conditions. A Kolmogorov-Smirnov test confirmed a significant difference between async distributions in each condition ($p<.01$). This reproduces the result of past work showing that mutual coupling promotes greater synchronization in piano dyads in an improvised context (Demos et al., 2017; Goebl & Palmer, 2009).

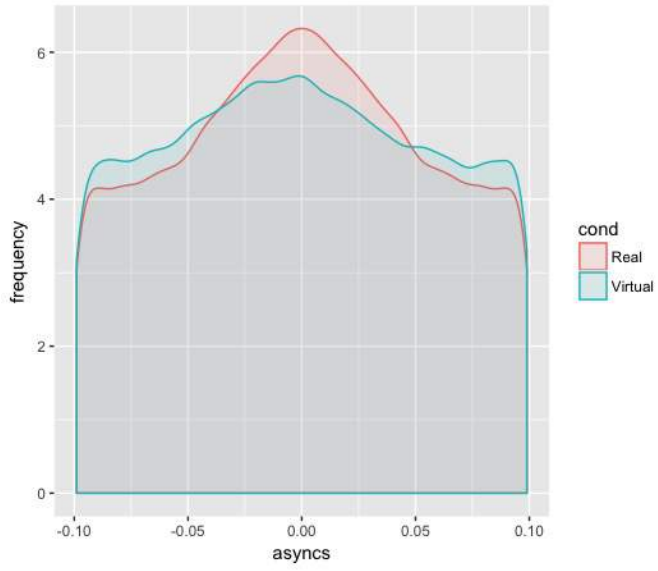


Figure 3: Mutual coupling promotes better synchronization. Density plot of asynchronies between co-performers' near-simultaneous (occurring within 100 ms of one another) note onsets. Asyncs are more tightly clustered around 0 seconds in *real* trials, in which mutual coupling is present.

Async frequency is symmetric around 0 for real trials because the same async was computed once for each partner (and thus represented twice with opposite signs). To assess asymmetries in virtual trials, asyncs were only computed by subtracting onset timestamps of ghost partners from the timestamps of live musicians. Thus positive asyncs indicate that the ghost led the live musician and negative asyncs indicate the reverse. Given past work which demonstrated musical leaders in composed settings play onsets slightly before other ensemble members, we were interested to see if ghost recordings (*de facto* leaders in virtual conditions) would lead the live players (Goebl & Palmer, 2009). However, the mean async across all virtual trials was less than 1 millisecond, indicating a symmetry between how often and how much live players led ghosts and vice versa.

Onset Density Given the lack of musical constraints, the improvised performances in our dataset exhibited high variability in rhythmic structure. Such variability could be found not just between subjects and trials, but even within particular performances. Tempos sped up and slowed down, and dyads moved in and out of "time" – sometimes playing rubato sections that lacked any steady pulse. Even within a given tempo, improvisers had the freedom to play more or fewer notes. To index all of this rhythmic variety, we compute onset density for each performer as the number of note onsets occuring within a given time window. Onset density was computed for each trial using a sliding window of 2 seconds and step size of 0.2 seconds, resulting in one onset density time series per subject-trial (Figure 4A).

How is inter-musician rhythmic coordination influenced by the presence or lack of mutual coupling? This question was approached by looking at cross-correlations between co-performers' onset density throughout the course of each trial. Cross-correlation was computed across a range of lags (+/- 5 seconds) to test for longer-term system memory and directional influence from one musician to another (Figure 4). Overall there was significantly greater cross-correlation in onset density in real trials as opposed to virtual (Figure 4B). This was confirmed with a Mann-Whitney test performed on the distributions collapsing over all time lags for all real trials (mean=.535, sd=.237) and virtual trials (mean=.356, sd=.287); $p < .01$.

Figure 4C displays how cross-correlation varies across lags as a function of condition. At each lag, the mean cross-correlation was obtained for all trials in each condition. In virtual trials we see greater cross-correlation at positive time lags, indicating that onset density of live musicians was more correlated with onset density of ghosts in previous time steps, as opposed to the other way around. This reflects the unidirectional influence inherent in virtual trials, whereby live musicians were influenced by their ghost partners but not the other way around. Onset density is symmetric around 0 for real trials, because data from each co-performer in a given trial was included. But it is interesting to note the dip

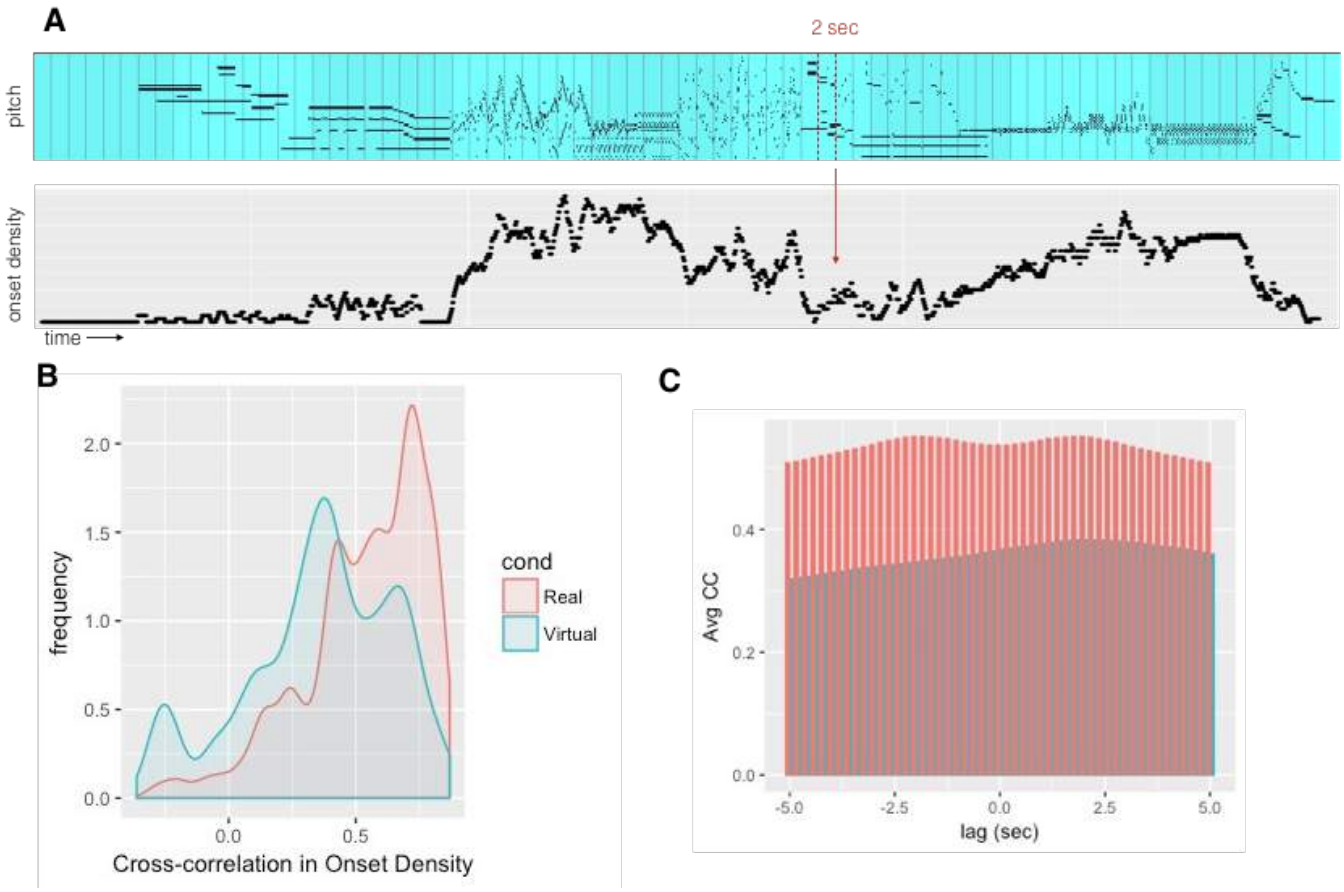


Figure 4: Mutual coupling promotes tighter coordination in onset density. (A) Time series of onset density (lower plot) are obtained by tallying the number of note onsets that occur within a 2 second sliding window of the MIDI recording (upper plot). (B) Cross-correlation in co-performer onset density, collapsing over a range of lags (+/- 5 sec). Overall there is greater cross-correlation in real trials. (C) Cross-correlation in co-performer onset density by lag, averaged across all trials in each condition. Greater cross-correlation in positive lags for virtual conditions reflects ground truth asymmetry inherent in this condition (i.e. live musician can respond to the recording, but not vice versa).

in cross-correlation around lag 0, surrounded by increased cross-correlation around lags +/- 2 sec. In real performances, co-performers' onset density is less correlated right at simultaneous time points, and more correlated at small lags of around 2 seconds. More analysis would be needed to elucidate this pattern, but it could be a result of "call and response" interplay between improvisers, whereby they exchange musical gestures in an interleaved manner.

General Discussion

In this work we have quantitatively demonstrated ways in which inter-musician mutual coupling in improvising jazz ensembles influences the music they spontaneously produce. Specifically, we showed that mutual coupling facilitates more effective coordination in jointly producing rhythmic structure. Musicians synchronized more precisely in performances where they were mutually coupled, and exhibited tighter coupling (greater cross-correlation) in onset density – a metric that captures tempo change and overall rhythmic activity. Subjects were coupled the most not at simultaneous times, but at small lags of about 2 seconds, suggesting a natural timescale of interaction. We also observed a quantitative artifact of musical leadership, as the onset density of improvisers in virtual trials was more correlated more with onset density of ghost partners at previous time points (again at about a 2 second lag). These objective results parallel the subjective intuitions of our performers, who rated trials with mutual coupling to be of higher quality and characterized by better coordination which was easier to achieve than in conditions with unidirectional influence.

When one listens to a great jazz combo, they are not merely listening to the sounds produced, but also to the complex underlying patterns of interaction which give rise to those sounds. This work provides the first controlled investigation of quantitatively measured coordination patterns demonstrated by freely interacting jazz musicians. It builds on prior research that studied the affordances of mutual coupling amongst co-performers by experimentally manipulating in-

teraction to reduce coupling in control conditions (Goebl & Palmer, 2009; Demos et al., 2017). But whereas past work focused on fine-temporal structure and movement dynamics exhibited by pianists, studying improvisers provided the opportunity to examine how mutual coupling influences larger scale musical structures (such as onset density) that are spontaneously generated in joint performance.

In the future we plan to delve deeper into the dynamics of how musicians pick up on and respond to the melodic and rhythmic offerings of their partner. One of the mesmerizing capabilities of expert jazz musicians, such as the pianists recorded in the current work, is their ability to improvise music with coherent and compelling tonal structure (melody and harmony). In ensemble performance, this melodic and harmonic structure emerges out of the ongoing interactions and musical negotiations taking place between ensemble members. In the future we plan to use this same (and expanding) dataset to delve deeper into how mutual coupling affords the emergence of stable tonal structure. Information theory offers a promising framework for inferring synergy and causality in multivariate time series of discretized, non-ordinal data, such as the musical pitches used in our data (Williams & Beer, 2010; Runge, 2018).

We also plan to extend our analyses to investigate the dynamical structure of our performances. Improvised music is essentially dynamic: the ensemble-generated musical structures evolve over time, as do the patterns of interaction between ensemble members. This is immediately evident observing the exemplar MIDI recording in Figure 4A, which appears to transition between regimes of sparse, sustained tones, and pointilistic sections characterized by short punctuated notes. Indeed, such dynamical structure is a central component of what makes improvised music so compelling.

To take another example, it could be the case that cross-correlation in onset density changes throughout performances. Imagine a performance in which there is initially negative cross-correlation between co-performers' onset density. This may transition to a period of positive cross-correlation, which could then be followed by yet another segment exhibiting no cross-correlation, in which performers go off on independent trails of rhythmic exploration. Such time-evolving interpersonal coordination would be lost on our current analysis (in fact it would obscure our results), but could be identified by computing cross-correlation over a sliding-window and analyzing how it varies over the course of a performance.

It is also likely the case that leadership roles shift throughout improvised performances. Without a well-established social structure (as exists in many forms of composed music), the distribution of leadership in ensembles is free to evolve in a self-organized fashion. Given our finding that musical leadership is associated with increased lagged cross-correlation of onset density (where followers are more influenced by the prior rhythmic activity of leaders), a sliding window analysis of onset density cross-correlation may also provide insight into the dynamical patterns governing time-evolving social structures. These kinds of higher-order analysis are a promising avenue towards contributing to the joint (mutually coupled!) efforts of empirical joint action studies and modeling of complex dynamical systems (Richardson, Dale, & Marsh, 2014).

References

Chang, A., Livingstone, S. R., Bosnyak, D. J., & Trainor, L. J. (2017). Body sway reflects leadership in joint music performance. *Proceedings of the National Academy of Sciences*, 201617657.

Demos, A. P., Carter, D. J., Wanderley, M. M., & Palmer, C. (2017). The unresponsive partner: roles of social status, auditory feedback, and animacy in coordination of joint music performance. *Frontiers in psychology*, *8*, 149.

Goebl, W., & Palmer, C. (2009). Synchronization of timing and motion among performing musicians. *Music Perception: An Interdisciplinary Journal*, *26*(5), 427–438.

Goldstone, R. L., & Gureckis, T. M. (2009). Collective behavior. *Topics in cognitive science*, *1*(3), 412–438.

Keller, P. E., & Appel, M. (2010). Individual differences, auditory imagery, and the coordination of body movements and sounds in musical ensembles. *Music Perception: An Interdisciplinary Journal*, *28*(1), 27–46.

Palmer, C., & Zamm, A. (2017). Empirical and mathematical accounts. *The Routledge Companion to Embodied Music Interaction*, 370.

Pearce, M., & Rohrmeier, M. (2012). Music cognition and the cognitive sciences. *Topics in cognitive science*, *4*(4), 468–484.

Richardson, M. J., Dale, R., & Marsh, K. L. (2014). Complex dynamical systems in social and personality psychology. *Handbook of research methods in social and personality psychology*, *253*.

Runge, J. (2018). Causal network reconstruction from time series: From theoretical assumptions to practical estimation. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *28*(7), 075310.

Sebanz, N., Bekkering, H., & Knoblich, G. (2006). Joint action: bodies and minds moving together. *Trends in cognitive sciences*, *10*(2), 70–76.

Walton, A. E., Washburn, A., Langland-Hassan, P., Chemero, A., Kloos, H., & Richardson, M. J. (2018). Creating time: Social collaboration in music improvisation. *Topics in cognitive science*, *10*(1), 95–119.

Williams, P. L., & Beer, R. D. (2010). Nonnegative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515*.

Interaction between Idea-generation and Idea-externalization Processes in Artistic Creation: Study of an Expert Breakdancer

Daichi Shimizu (tothefuture0415@yahoo.co.jp)
Graduate School of Education, University of Tokyo
Tokyo, 113-0033 Japan

Masaya Hirashima (hira@nict.go.jp)
Center for Information and Neural Networks, National Institute of Information and Communication Technology
Osaka, 565-0871 Japan

Takeshi Okada (okadatak@p.u-tokyo.ac.jp)
Graduate School of Education, University of Tokyo
Tokyo, 113-0033 Japan

Abstract

This study develops a cognitive model to explain the process of artistic creation in a dance domain. Many researchers in the field of psychology and cognitive science have investigated the process of creativity and developed various theories that explain this process. Their efforts have mostly focused on higher cognitive functions of artists and scientists. However, in recent years, several studies that have highlighted the importance of the interaction between idea generation and idea externalization processes suggest that people can find and develop new aspects of images and ideas by perceiving and reflecting on the images and ideas they externalize. This study develops a cognitive model that explains this interaction process in dance creation by referring to a famous theory of motor learning, the closed-loop model. We also investigate dance creation of an expert breakdancer and check the validity of our proposed model.

Keywords: creativity, artistic creation, externalization of ideas, closed-loop model, performing arts, breakdance

Introduction

How do professional artists generate their original and fascinating expressions? In the psychology and cognitive science field, many researchers have investigated the process of creativity (e.g., Dunbar, 1993; Okada & Ishibashi, 2017; Wallas, 1926). For example, Finke, Ward, and Smith (1992) proposed the Geneplore model, which explains the process of idea generation, focusing on various cognitive functions. The Geneplore model suggests that people generate and explore their ideas under several task constraints by using cognitive functions such as long-term memory, mental rotation, and concept combination. Additionally, Wallas (1926) proposed the four-stage model based on anecdotal records of several artists. This model explains the creative process in four phases: preparation, incubation, illumination, and verification.

These traditional theories have focused mainly on the cognitive process of the creators. However, in artistic creation, the process of externalizing the creators' images and ideas is also important. Artists externalize their images and ideas in the end or middle of almost all their creations (Glăveanu, 2013). For example, in a dance creation, dancers externalize their images as physical movements, and in paintings, artists externalize their images to the outside as traces, using brushes, paints, and canvases. We propose that this externalization process and the perception or reflection of those externalized images and ideas facilitate the development of the images and ideas. However, previous studies of creation have regarded this externalization process as an implementation phase, and have thus paid it little attention (e.g., Zeng, Proctor, & Salvendy, 2011). In recent years, however, some researchers came to focus on this process of idea-externalization (Glăveanu, 2013). Based on these discussions, we highlight the importance of the interaction between the idea-generation process and the idea-externalization process in artistic creation, and we develop a model that explains the influence of this interaction.

Although these studies have highlighted the importance of the idea-externalization process and its interaction with the idea-generation process, they have not proposed a mechanism as to how this interaction facilitates the creation. Regarding this mechanism, Goldschmidt (1991, 1994) and Kirsh (2009, 2010) offered useful suggestions. Goldschmidt (1991, 1994) investigated the role of sketch in design and claimed that people cannot focus on all features of their images or ideas of expression while they are generating them. For example, when people consider several components that must be included in a design, they cannot focus on the relationships or blank spaces between these components. However, people can find and focus on these features of their images and ideas if they first externalize them as sketches. Furthermore, by focusing on these hidden features, they can develop their images and ideas from different aspects and generate original and fascinating ideas (see Fig. 1). Based on this discussion, Goldschmidt (1991, 1994) emphasized the importance of sketches in design (externalization of images and ideas) and perceiving or reflecting on them. She referred to this perception and/or reflection of sketches and the subsequent development of images and ideas as interactive imagery. Also, Kirsh (2009,

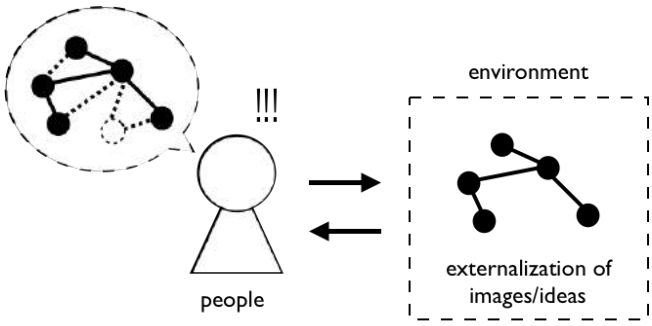


Fig. 1. Interaction between idea-generation process and idea-externalization process

2010) suggested that the idea-externalization process plays various roles in people's cognition. The process of externalization does not only save the internal memory but also facilitate the re-representation of images and the construction of the complex structure of images. He proposed that these processes facilitate the creation of artists.

Based on this discussion, this study explains how the interaction between the idea-generation process and the idea-externalization process facilitates artistic creation. To do so, we develop a cognitive model that explains the process of artistic creation in a dance domain. We also conduct a case study to investigate the creation process of an expert breakdancer and check the validity of our model.

Model development

This study develops a model to explain the influence of that interaction on creation. In particular, this study focuses on dance creation (the creation of breakdance) that existing studies have investigated over the past 10 years. The creation of breakdance is a suitable subject for this study because some fieldwork studies suggested the importance of externalizing images and ideas and perceiving or reflecting on them in breakdance (e.g., Shimizu & Okada, 2013).

To develop our model, we examined models of creation, such as the Geneplore model, reviewed discussions by Goldschmidt (1991, 1994), and referred to a famous theory of motor learning, the closed-loop model, which has been prominent in cognitive science and biomechanics. We consider that this theory provides a clear explanation about the influence of externalization of images and ideas, and the effect of perception and reflection on dance creation.

The closed-loop model emphasizes the importance of a movement-implementation process, especially the feedback error, which refers to a gap between somatosensory feedback derived from the movement and the prediction of that feedback, called efference copy, in the motor-learning process (Schmidt & Lee, 2011). This model explains the mechanism of motor learning (see dotted lines in Fig. 2), which we describe as follows: First, people perceive and identify a stimulus from their surroundings and select their reaction of movements in the cerebellum and primary motor area (movement selection). Then, to implement those

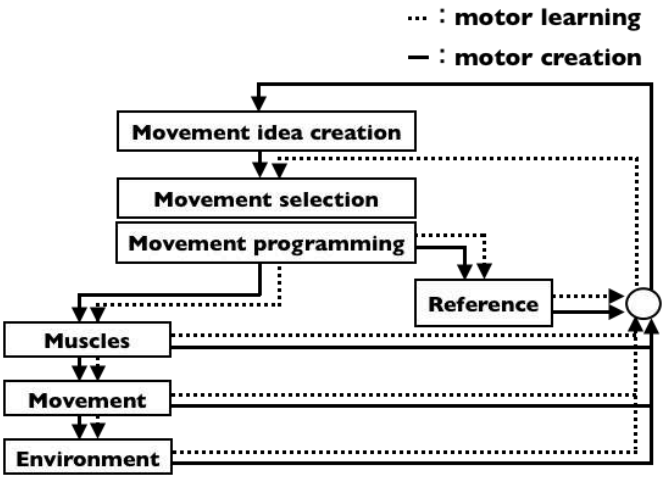


Fig. 2. Motor learning process of closed-loop model and motor creation process of our model. Dotted lines indicate the process of motor learning and solid lines indicate the process of motor creation.

movements, they send signals to the peripheral nerves in their muscles (movement programming) and conduct those movements (muscles, movement). Simultaneously, they send a copy of this motor program, called the efference copy, and generate a prediction of the somatosensory feedback derived from those movements (reference). People receive the gap between the somatosensory feedback and the prediction of that feedback, known as the feedback error, and in the next trial of movements, they refer to this error and correct their motor plan to get their movements as close to their goal (model) movements as possible (movement programming). The closed-loop model explains a mechanism as to how people improve and learn movements by these repetitive processes.

The above-mentioned process explains the mechanism of motor learning and refinement when people have a clear model of movement (i.e., a clear goal). But, how is the mechanism of motor creation achieved when people have no clear goal? In this process, people should first generate their model of movement (i.e., a goal) through cognitive functions proposed by traditional creative theory (such as the Geneplore model), and they should implement this movement plan as a movement. The roles of the feedback error in motor creation also differ from those in motor learning. In motor learning, the feedback error provides information that helps people to approximate their movements to those of the model. However, in motor creation, the feedback error provides information to find and focus on the hidden features (e.g., the relationships between components) of their proposed movement ideas (Goldschmidt, 1991, 1994). As a result, people should develop their images and ideas from various aspects and generate their original movements. In this manner, externalizing images and ideas and perceiving and/or

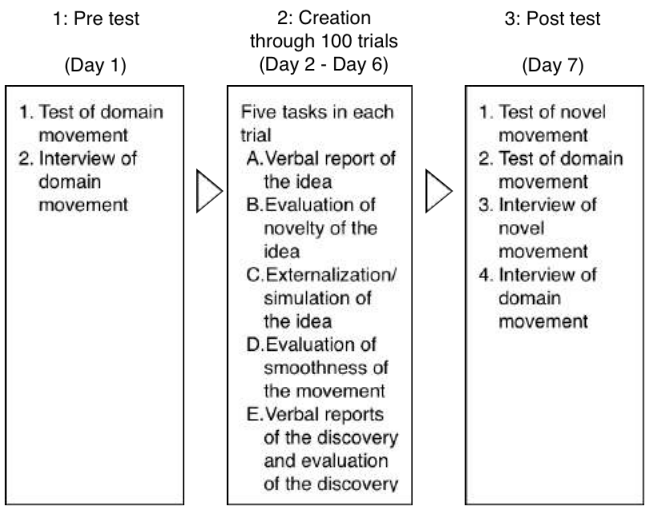


Fig. 3. Procedures of case study

reflecting on them will facilitate the creation of dance movements.

Based on these discussions, we developed a model of dance creation (solid lines in Fig. 2). This model explains the process of motor creation in dance as follows: First, people generate their ideas of movements through cognitive functions such as mental rotation and concept combination (movement idea creation). They generate these ideas by focusing on specific aspects of movements, and they send signals to implement these movements (movement programming) and conduct the movements (muscles, movement). After that, people receive the somatosensory feedback of the movements, compare those with their predictions of them (reference), and calculate the feedback error. Then, they develop and reconstruct their ideas of movements based on this feedback error and by shifting the aspects where they focus. This model thus explains the mechanism of how the interaction between the idea-generation process and the idea-externalization process facilitates dance creation. In particular, the model highlights the importance of the feedback error derived from the idea-externalization process. Notably, previous studies claimed that the feedback error plays an important role in various phenomena such as tickling (Blackmore et al., 1999) and phantom pain (Ramachandran & Ramachandran, 1996), not only in motor learning. We suppose that the feedback error has various functions in human movements.

Case study

Next, we check the validity of our proposed model by conducting a case study of an expert breakdancer's creation, and we verify whether the interaction between the idea-generation process and idea-externalization process facilitated the dance creation. We set two conditions. In the first condition, the dancer generated an original movement in an interactive condition (with the above-mentioned interaction), and in the second condition, the dancer generated an original movement in a non-interactive condition (without the interaction). We compare these two conditions and investigate the differences in the creation process. We also investigate how this interaction facilitates the dance creation by checking the creation process of the first condition in detail. With reference to these two results, we discuss the validity of our model.

Participant

An award-winning Japanese expert breakdancer with nine years' experience in breakdancing participated in our case study. He generated original dance movements over seven days in the two conditions.

Condition

The expert dancer developed original dance movements in the two conditions (interactive and non-interactive conditions). In the interactive condition, he developed his movements by repeating tasks to generate an idea (idea-generation process) and to externalize his idea as movement (idea-externalization process). In the non-interactive condition, he developed his movements by repeating tasks to generate an idea (idea-generation process) and to simulate his idea in his mind, without externalizing it as movement (idea-simulation process).

In breakdancing, dancers generate original movements by focusing on and developing specific movements in the domain (Shimizu & Okada, 2013, 2018). Therefore, in our case study, we asked the dancer to generate an original movement by developing a specific domain movement. We used different domain movements in each condition[1].

Procedure

In each condition, the dancer generated original movements through 100 trials over seven days. Fig. 3 shows the procedures. The dancer followed the same procedures in both conditions, except for task C on days 2–6 (externalization/simulation of the idea).

On days 1 and 7, we tested the domain movement (10 trials each day) using the video camera and motion capture system described in the next section. We also conducted interviews about the domain movement. We conducted further tests (3 trials) and interviews of the dancer's original movements on day 7.

On days 2–6, the dancer generated original dance movements through 20 trials per day. In each trial, he conducted five tasks. First, he generated the idea of movement, and reported its content (task A). Second, he evaluated the novelty of the idea on a one hundred-point scale using a Visual Analog Scale (VAS) (task B). Third, the dancer externalized his idea as a movement in the

[1] We used different domain movements in each condition to exclude the strong influence of the first-time creation on the second-time creation. However, we needed to be careful when interpreting the results of this study because of the different features of the domain movements in each condition.

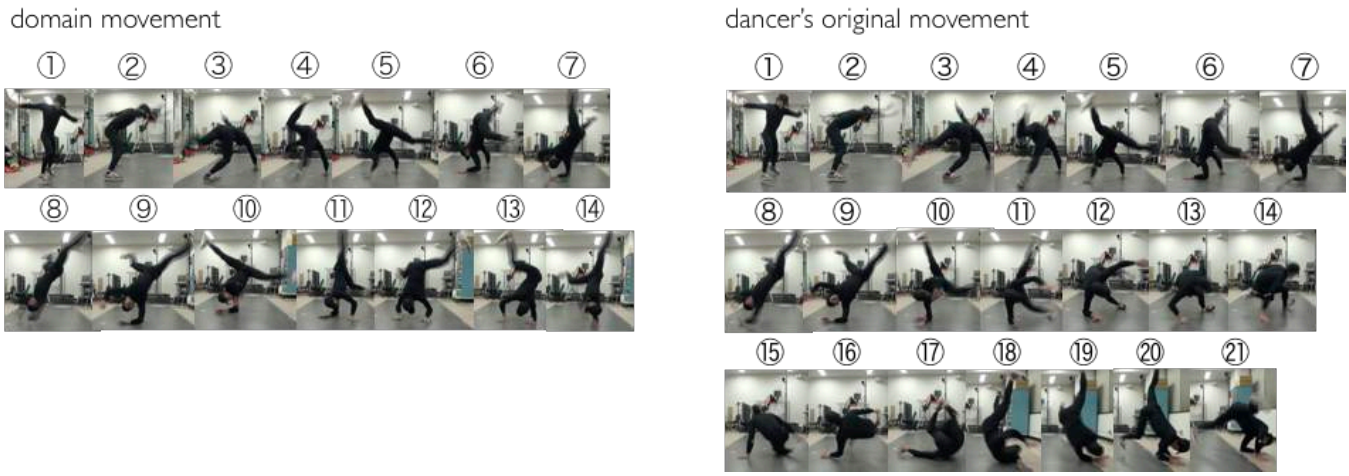


Fig. 4. Domain movement and dancer's original movement in the interactive condition. After the action of picture 14, the dancer goes into the action of picture 7 again, and repeats the rotation in domain movement (left side). In original movement, he goes into the action of picture 7 again after the action of picture 21, and repeats the rotation (right side).

interactive condition and simulated his idea in the non-interactive condition (task C). In the interactive condition, we recorded the movement using a video camera and motion capture system. Fourth, the dancer evaluated the smoothness of the movement using VAS (task D). Fifth, he reported his discovery brought by the idea-externalization and simulation tasks and evaluated the degree of that discovery using VAS (task E). The dancer repeated these five tasks a hundred times to generate his original movement. We set these tasks based on the creation process of expert dancers observed in the fieldwork study (Shimizu & Okada, 2018). We focused on tasks A (verbal report of the idea), C (externalization/simulation of the idea), and E (verbal report of the discovery) in this study because these data include the important information on the idea-generation and idea-externalization processes.

Apparatus

In this study, we used a motion capture system (OQUS 300, Optical motion capture system, QUALISYS co.) to measure the features of the movements in the creation process in the interactive condition. The dancer wore a suit for the system, attached fourteen markers to his body, and worked on the creation. We did not measure the movements in the non-interactive condition because the dancer did not conduct any movements during the creation process; however, for consistency, the dancer wore the suit and attached markers in the non-interactive condition.

Results and Discussion

Outline of the Original Movements First, we explain the outline of the movement that the dancer generated in each condition. In the interactive condition, the dancer generated the movement shown on the right side of Fig. 4 (we also show the domain movement in the left). In this original movement, the dancer stops the rotation of the domain movement by landing on his right leg, and he uses the momentum of rotation for the inverse rotation. He described this action as canceling the rotation, and he generated this original movement by developing this concept. He and another expert dancer confirmed that this was an original movement that they had never seen in the breakdance domain.

In the non-interactive condition, the dancer generated an original movement based on the domain movement called Drill. In this domain movement, the dancer lands on the ground with his head and rotates his whole body (we abbreviate the figure because of space limitations). In the original movement, the dancer lands on the ground on his back after the rotation, and then jumps up and rotates in the different direction (this rotation is similar to another domain movement called Trax). Although mixing the two domain movements seemed interesting, he was not convinced of its originality.偏

Verbal Report of the Idea In the following sections, we compare the creation process in each condition. First, we investigate the verbal report of the idea that the dancer mentioned in each trial. In the analysis, we checked and counted the frequencies of the following three aspects because the dancer mentioned them many times in his reports: (1) specific body parts (e.g., head, right arm, left arm, right leg); (2) abstract concepts of the domain movement (words such as "direction," "speed," and "axis" of the rotation); and (3) other movements in the breakdance domain (e.g., Trax, Baby Windmill, and Ninety). We summed the total frequency of each aspect in each day, divided them by the total frequency of all three aspects, and calculated the relative frequencies of these aspects for each day. These frequencies indicated which aspects the dancer focused on each day.

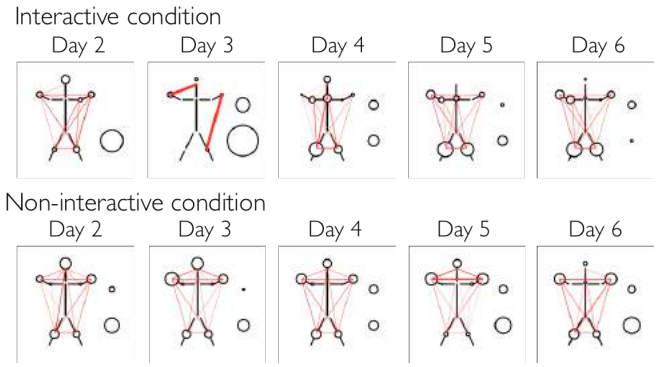


Fig. 5. Results of verbal reports of the idea. Two circles drawn to the right side indicates the frequencies of second aspect (abstract concepts of the domain movement, upper side) and third aspect (other movements in breakdance domain, lower side).

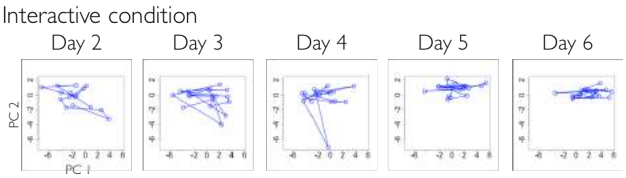


Fig. 6. Scores of PC 1 and PC 2 of the movement

We show these frequencies as sizes of circles in Fig. 5. This figure shows that, in the interactive condition, the dancer changed the aspects of his movement actively between days 2 and 4. Then, on days 5 and 6, he focused on specific aspects such as the right leg when attempting to generate the original movement (the right leg has an important role in his original movement: to stop the rotation of the domain movement). On the other hand, in the non-interactive condition, the dancer did not actively change the aspects of his movement during his creation.

Features of the Movement (Externalized Idea) Next, we investigate the features of the movement (externalized ideas) generated in the interactive condition. We calculated the kinematics data (joint angles and joint angular velocities in each segment) from the time-series position data of 14 markers using the inverse kinematics technique (Hirashima et al., 2008). Then, we conducted principal component analysis and extracted two components that had high contributions (proportions of variance) for explaining these movements (see Kadone & Nakamura, 2007), which we called PC 1 and PC 2. PC 1 and PC 2 are reduced dimensions that explain important features of the movement.

Fig. 6 shows that the scores of PC 1 and PC 2 had various values (PC 1: -7.06~3.84, PC 2: -7.10~1.34) on days 2–4. However, on days 5–6, these scores, especially scores of PC 2 converged at specific values (PC 1: -4.30~5.64, PC 2: -0.61~2.13). These results suggest that, in the interactive

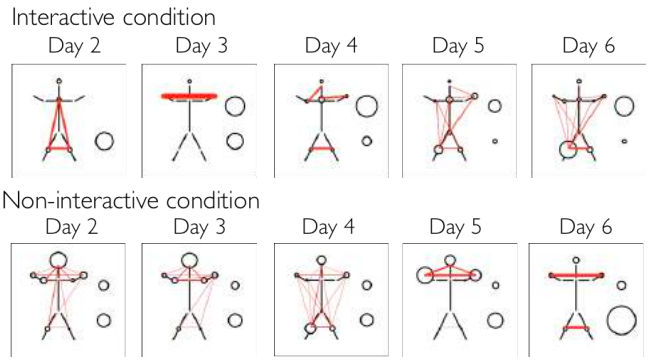


Fig. 7. Results of verbal reports of the discovery. Two circles drawn to the right side indicates the same aspects as those of Fig. 5.

condition, the dancer generated various kinds of movements that had various features in the first half of the creation. The dancer also focused on a particular movement that had a specific feature (the movement which involved stopping the rotation with his right leg) in the second half of the creation.

Verbal Report of the Discovery We investigated the discovery that the dancer mentioned when he externalized and simulated his idea. We examined the verbal report of the discovery and conducted the same analysis of the verbal report of the idea.

Fig. 7 shows that on days 2–4, the dancer actively reported his discovery in various aspects in the interactive condition. On days 5–6, however, he focused on specific aspects such as the right leg and the abstract concept, and he frequently reported his discovery of these aspects. By contrast, in the non-interactive condition, the dancer did not focus on various aspects from days 2–4. He focused on similar aspects to those on days 2–4, and he came to focus on various aspects in his discovery report on day 5.

These results of three analyses indicate that in the interactive condition, the dancer actively changed the aspects on which he focused in the early part of the creation, and in the late part, he focused on the specific aspects and on refining his idea. A retrospective interview conducted on day 7 supports this claim. The dancer mentioned that he found the idea of the original movement around trial 50, and subsequently focused on refining that movement. By externalizing his idea as movement and reconstructing his idea using the feedback error derived from that externalization, he was able to find and focus on the various and hidden aspects of his idea and generate an original movement.

Overall Picture of the Creation in Interactive Condition Finally, we provide an overall picture of the dancer's creation in the interactive condition. Before finding the idea for an original movement (stopping the rotation of the domain movement with his right leg) at around trial 50, the

Table 1. Examples of dancer's discovery reports

"I jumped and rotated in the vertical direction higher than I expected. I was surprised by it." (Discovery report in trial 32)
"This time, I tried to bend both knees when I conducted EAT. This made me jump lower in the horizontal direction than I expected. Though it did not change the final position of EAT, it increased the rotation speed horizontally. I was confused by that." (Discovery report in trial 34)

dancer explored various ideas. Fig. 5 shows that he focused on each body part on day 2, and he said that he found an important aspect of the domain movement (rotation) on day 3, which he then focused on (Fig. 7 also supports this claim). The dancer found this aspect by externalizing his idea as movement and perceiving or reflecting on the feedback error derived from that externalization. In the dancer's verbal reports of the discovery on day 3 (in trials 32 and 34), he mentioned that he was surprised at the gap between the somatosensory feedback derived from the movement and his prediction of that, and became interested in the hidden aspect of his idea: the rotation (see Table 1). After this finding, the dancer focused on this aspect, the rotation, and attempted to generate an original movement by making various changes to it. Finally, the breakdancer developed his idea for an original movement, which involved stopping the rotation with his right leg, at around trial 50. The interaction between the idea-generation process and idea-externalization process led to the findings of the hidden aspect of his idea and facilitated the generation of his original dance movement. These results suggest that the process explained by our model occurred in the expert breakdancer's dance creation.

General discussion

This study developed a model to explain the process of artistic creation in the dance domain. We also conducted a case study that investigated the creation process of an expert breakdancer and verified the validity of our proposed model. Fig. 2 shows that the model developed herein proposes the importance of interactions between the idea-generation process and the idea-externalization process in dance creation based on the closed-loop model (Shmidt & Lee, 2011), the Geneplore model (Finke et al., 1992), and discussions by Goldschmidt (1991, 1994) and Kirsh (2009, 2010). The closed-loop model shows the importance of somatosensory feedback and its error derived from the movement in motor learning. We extended the roles of the feedback error and applied them to the creation of a novel dance movement. By externalizing their idea as a movement and focusing on the feedback error derived from that movement, dancers can find new and hidden aspects of the movement and develop their idea actively. Traditional theories of creation in psychology and cognitive science paid little attention to the importance of interaction between the idea-generation process and the idea-externalization process because the creation of a novel image or idea was considered to be achieved in people's cognitive processes. On the other hand, this study highlights the importance of the interaction between idea-generation and externalization and identified the mechanism of that interaction. We suggest that the processes of idea generation and idea externalization are highly connected, and this connection has a strong influence on creation.

However, we need to consider the generalizability of the influence of this interaction with caution. Based on the hands-on nature of an artistic creation, interactions between the idea-generation process and the idea-externalization process are important in almost all artistic domains. Goldschmidt (1991, 1994) and Glăveanu (2013) proposed the importance of interactions between imagination and externalization in artistic creation. However, there are critical differences between dance creation and other kinds of artistic creation. In particular, media that artists use for externalizing their images and ideas and the feedback they receive from this externalization process are different. In dance creation, dancers externalize their images and ideas as movements through their bodies, and they mainly receive somatosensory feedback from their movements. In paintings, however, artists externalize their images and ideas as traces by using various tools such as brushes, paints, and canvases in addition to their bodies, and they mainly receive visual feedback from their paintings. We thus need to consider these similarities and differences among various artistic domains when discussing the generalizability of our model.

Our model has other limitations. As this study verified the validity of the model by investigating the creation process of only one expert dancer, we should collect data from more expert dancers. Additionally, we should set various domain movements as the base movements and take a counterbalance of those movements between the two conditions. However, to investigate the creation process of experts takes considerable time and effort. We therefore need to develop a method to investigate the creation process of many experts efficiently in more natural field situations.

Acknowledgments

This work was supported by Grant-in-Aid for Scientific Research 26780352 and 16K17306 to the first author and 24243062 to the second author from the Japan Society for the Promotion of Science. We greatly appreciate the cooperation of an expert breakdancer and the comments and support of members of Okada lab at the University of Tokyo.

References

Blakemore, S. J., Frith, C. D., & Wolpert, D. M. (1999). Spatio-temporal prediction modulates the perception of self-produced stimuli. *Journal of cognitive neuroscience*, 11(5), 551-559.

Dunbar, K. (1993). Concept discovery in a scientific domain. *Cognitive Science*, 17(3), 397-434.

Finke, R. A., Ward, T. B., & Smith, S. M. (1992). *Creative cognition: Theory, research, and applications.* Cambridge, MA: MIT Press.

Glăveanu, V. P. (2013). Rewriting the language of creativity: The Five A's framework. *Review of General Psychology*, 17(1), 69.

Goldschmidt, G. (1991). The dialectics of sketching. *Creativity Research Journal*, 4 (2), 123-143.

Goldschmidt, G. (1994). On visual design thinking: the cis kids of architecture. *Design Studies*, 15 (2), 158-174.

Hirashima, M., Yamane, K., Nakamura, Y., & Ohtsuki, T. (2008). Kinetic chain of overarm throwing in terms of joint rotations revealed by induced acceleration analysis. *Journal of biomechanics*, 41(13), 2874-2883.

Kadone, H., & Yoshihiko, N. (2007). Symbolic memory of motion patterns using hierarchical bifurcations of attanctors in an associative memory model. *Journal of the Robotics Society of Japan*, 25 (2), 249-258.

Kirsh, D. (2009). Interaction, External Representations and Sense Making. *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, 1103-1108.

Kirsh, D. (2010). Thinking with external representations. *Ai & Society*, 25(4), 441-454.

Okada, T., & Ishibashi, K. (2017). Imitation, Inspiration, and Creation: Cognitive Process of Creative Drawing by Copying Others' Artworks. *Cognitive science*, 41(7), 1804-1837.

Ramachandran, V. S., & Rogers-Ramachandran, D. (1996). Synaesthesia in phantom limbs induced with mirrors. *Proc. R. Soc. Lond. B*, 263(1369), 377-386.

Schmidt, R. A., Lee, T., Winstein, C., Wulf, G., & Zelaznik, H. (2011). *Motor Control and Learning, 5E*. Human kinetics.

Shimizu, D., & Takeshi, O. (2013). The Process of Creation of New Movements in Street Dance. *Cognitive Studies*, 20(4), 488-492.

Shimizu, D., & Okada, T. (2018). How Do Creative Experts Practice New Skills? Exploratory Practice in Breakdancers. *Cognitive science*, 42(7), 2364-2396.

Wallas, G. (1926). *The art of thought.* New York: Harcourt.

Zeng, L., Proctor, R. W., & Salvendy, G. (2011). Can traditional divergent thinking tests be trusted in measuring and predicting real-world creativity?. *Creativity Research Journal*, 23(1), 24-37.

Partitioning the Perception of Physical and Social Events Within a Unified Psychological Space

Tianmin Shu Yujia Peng Hongjing Lu Song-Chun Zhu
{tianmin.shu, yjpeng, hongjing}@ucla.edu sczhu@stat.ucla.edu
Department of Psychology and Statistics, University of California, Los Angeles, USA

Abstract

Humans demonstrate remarkable abilities to perceive physical and social events based on very limited information (e.g., movements of a few simple geometric shapes). However, the computational mechanisms underlying intuitive physics and social perception remain unclear. In an effort to identify the key computational components, we propose a unified psychological space that reveals the partition between the perception of physical events involving inanimate objects and the perception of social events involving human interactions with other agents. This unified space consists of two prominent dimensions: an intuitive sense of whether physical laws are obeyed or violated; and an impression of whether an agent possesses intentions, as inferred from movements. We adopt a physics engine and a deep reinforcement learning model to synthesize a rich set of motion patterns. In two experiments, human judgments were used to demonstrate that the constructed psychological space successfully partitions human perception of physical versus social events.

Keywords: social perception; intuitive physics; intention; deep reinforcement learning, Heider-Simmel animations

Introduction

Imagine you are playing a multi-player video game with open or free-roaming worlds. You will encounter many physical events, such as blocks collapsing onto the ground, as well as social events, such as avatars constructing buildings or fighting each other. All these physical and social events are depicted by movements of simple geometric shapes, which suffice to generate a vivid perception of rich behavioral, including interactions between physical entities, interpersonal activities between avatars engaged in social interactions, or actions involving both humans and objects.

This type of rich perception elicited by movements within simple visual displays has been extensively studied in psychology. Prior work showed that humans possess a remarkable ability to perceive physical events and to infer physical properties (e.g., masses of objects) (Proffitt & Gilden, 1989), as well as to make causal judgment (Michotte, 1963), based on observations of the movements of two objects. Furthermore, Heider & Simmel (1944) demonstrated that humans also excel in spontaneously reconstructing social events from movements of simple geometric shapes, and describe their observations in terms of agency, goals, and social relations. These classic studies, along with a great deal of subsequent psychological research (e.g., Kassin 1981; Scholl & Tremoulet 2000; Gao et al. 2009, 2010), provide convincing evidence that human inferences about physical and social events are efficient and robust, even given very limited visual inputs.

Although many studies of both intuitive physics and social perception examined dynamic stimuli consisting of moving shapes, these research areas have largely been isolated from one another, with different theoretical approaches and experimental paradigms. In the case of physical events, research has been focused on the perception and interpretation of physical objects and their dynamics, aiming to determine whether humans use heuristics or mental simulation to reason about intuitive physics (see a recent review by Kubricht et al. (2017)). For social perception, some research has aimed to identify critical cues based on motion trajectories that determine the perception of animacy and social interactions (Dittrich & Lea, 1994; Scholl & Tremoulet, 2000; Gao et al., 2009; Shu et al., 2018). Other work focused on inferences about agents' intentions (Baker et al., 2009; Ullman et al., 2010; Pantelis et al., 2014). In contrast to the clear separation between the two research topics, human perception integrates the perception of physical and social events. Hence, it is important to develop a common computational framework applicable to both intuitive physics and social perception to advance our understandings on how humans perceive and reason about physical and social events.

In the present paper, we propose a unified framework to account for the perception of both physical events and of social events based on movements of simple shapes. We aim to construct a unified psychological space that may reveal the partition between the perception of physical events involving inanimate objects and the perception of social events involving human interactions with other agents. Specifically, we hypothesize that this unified space includes two prominent dimensions: an intuitive sense regarding whether physical laws are obeyed or violated; and an impression of whether an agent possesses intentions in the display. Note that the intuitive sense of physical violation may result from observable physical forces that can not be explained by perceived entity properties (such as motion, size, etc.) in a scene. The development of this unified space may shed light on many fundamental problems in both intuitive physics and social perception.

To construct such space, we project a video as a whole onto the space. Hence, a large range of videos can provide a distribution of observed events. We can also project individual entities in one physical or social event onto the same space, and then examine pairwise relations between the projected locations of entities in the space, which could serve as an informative cue for judging social/physical roles of entities (e.g, as an human agent or an inanimate object).

To test the hypothesized psychological space, we report experiments involving many Heider-Simmel animations in which simple moving shapes vary in degrees of physical vi-

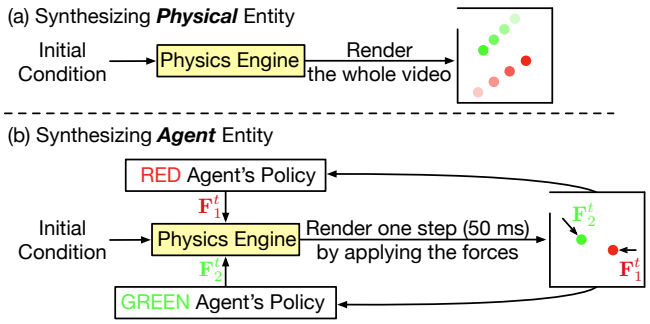


Figure 1: Overview of our joint physical-social simulation engine. For a dot instantiating a physical object, we randomly assign its initial position and velocity and then use physics engine to simulate its movements. For a dot instantiating a human agent, we use policies learned by deep reinforcement learning to guide the forces provided to the physics engine.

olation and the involvement of intention. Prior work usually created Heider-Simmel-type stimuli using manually designed interactions (Gao et al., 2009, 2010; Isik et al., 2017), rule-based behavior simulation (Kerr & Cohen, 2010; Pantelis et al., 2014), and trajectories extracted from human activities in aerial videos (Shu et al., 2018). It is challenging to manually create many motion trajectories, and to generate situations that violate physical constraints. Accordingly, we develop a joint physical-social simulation-based approach built upon a 2D physics engine (Figure 1). A similar idea has been previously instantiated in a 1D environment, Lineland (Ullman, 2015). By generating Heider-Simmel-type animations in a 2D environment with the help of deep reinforcement learning, our simulation approach is able to depict a richer set of motion patterns in animations.

This advanced simulation provides well-controlled Heider-Simmel stimuli enabling the measurement of human perception of physical and social events for hundreds of different motion patterns. We also develop general metrics to measure how well the motion patterns in an animation satisfy physics, and the likelihood that dots are agents showing intentions. These two indices were computed for each stimulus shown to human observers, allowing us to map all videos into a unified space as the two measures providing primary coordinates. In two experiments, we combined model simulations with human responses to validate the proposed psychological space.

Stimulus Synthesis

Overview

Figure 1 gives an overview of our joint physical-social simulation engine. Each video included two dots (red and green) and a box with a small gap indicating a room with a door. The movements of the two dots were rendered by a 2D physics engine (pybox2d[1]). If a dot represents an object, we randomly assigned the initial position and velocity, and then used the physics engine to synthesize its motion. Note that our simulation incorporated the environmental constraints (e.g., a dot can bounce off the wall, the edge of the box), but did not include friction. If a dot represents an agent, it was assigned with a clearly-defined goal (e.g., leaving room) and pursued its goal by exerting self-propelled forces (e.g., pushing itself towards the door). The self-propelled forces were sampled from agent policy learned by deep reinforcement learning (see more details in a later subsection). Specifically, at each step (every 50 ms), the agent observed the current state rendered by the physics engine, and its policy determined the best force to advance the agent's pursuit of its goal. We then programmed the physics engine to apply this force to the dot, and rendered its motion for another step. This process was repeated until the entire video was generated.

[1] https://github.com/pybox2d/pybox2d

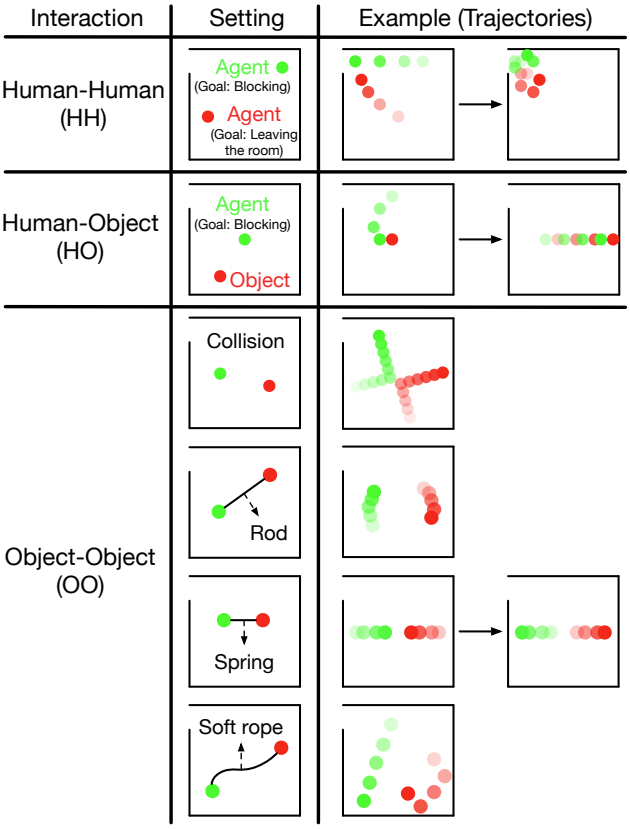


Figure 2: An illustration of three types of synthesized interactions for physical and social events. A few examples are included by showing trajectories of the two entities. The dot intensities change from low to high to denote elapsed time. Note that the connections in OO stimuli (i.e., rod, spring, and soft rope) are drawn only for illustration purpose. Such connections were invisible in the stimuli. Examples of stimuli are available at: https://tshu.io/HeiderSimmel/CogSci19.

Interaction Types

As summarized in Figure 2, we consider three types of interactions, including human-human (HH), human-object (HO)

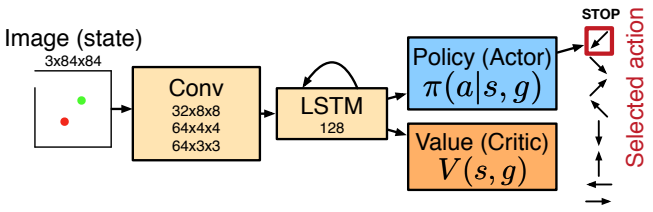


Figure 3: The deep RL network architecture for learning policy for goal-directed movements of an agent. For each goal, we train a separate network with the same architecture.

and object-object (OO) interactions, all of which are generated by the approach depicted in Figure 1. Note that in this paper we treat the terms "human" and "agent" interchangeably. When synthesizing the agents' motion, we set two types of goals for the agents, i.e., "leave the room" (g_1) and "block the other entity" (g_2). Specially, in HH stimuli, one agent has a goal of leaving the room (g_1), and the other agent aims to block it (g_2); in HO stimuli, an agent always attempts to keep a moving object within the room (g_2) and the object has an initial velocity towards the door. By randomly assigning initial position and velocity to an agent, we can simulate rich behaviors that can give the impression such as blocking, chasing, attacking, pushing, etc.

In addition to the three general types of interactions, we have also created sub-categories of interactions to capture a variety of physical and social events. For OO animations, we included four events, as collision, connections with rod, spring and soft rope. Since these connections were invisible in the displays, the hidden physical relations may result in a subjective impression of animacy or social interactions between the entities. In addition, the invisible connections between objects (rod, spring, and soft rope) introduce different degrees of violation of physics in the motion of the corresponding entities if assuming the two entities are independent. For HH animations, we varied the "animacy degree" (AD) of the agents by controlling how often they exerted self-propelled forces in the animation. In general, a higher degree of animacy associates with more frequent observations about violation of physics, thus revealing self-controlled behaviors guided by the intention of an agent. The animacy manipulation introduced five sub-categories of HH stimuli with five degrees of animacy – 7%, 10%, 20%, 50%, and 100%, respectively corresponding to applying force once for every 750, 500, 250, 100, and 50 ms. In an HH animation, we assigned the same level of animacy degree to both dots.

Training Policies

As shown in Figure 1, in order to generate social events, we need sensible policies to infer the self-propelled forces for pursuing goals. However, searching for such policies in a physics engine is extremely difficult. In this study, we use deep reinforcement learning (RL) to acquire such policies, which has been shown to be a powerful tool for learning complex policies in recent studies (Silver et al., 2017).

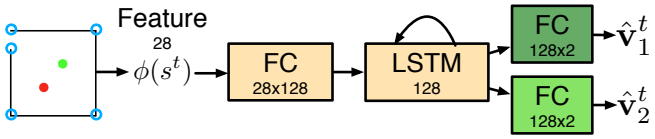


Figure 4: Network for the physical motion prediction model to emulate intuitive physics. Blue circles indicate the corners of the room used for deriving the input features.

Formally, an agent's behavior is defined by an Markov decision process (MDP), $\langle \mathcal{S}, \mathcal{A}, \mathcal{T}, R, \mathcal{G}, \gamma \rangle$, where $\mathcal{S}$ and $\mathcal{A}$ denote the state space (raw pixels as in Figure 3) and action space, $\mathcal{T} : \mathcal{S} \times \mathcal{A} \mapsto \mathcal{S}$ are the transition probabilities of the environment (in our case, deterministic transitions defined by physics), R is the reward function associated with the intended goals $g \in \mathcal{G}$, and $0 < \gamma \leq 1$ is a discount factor. To match to the experimental setup, we define two reward functions for the two goals: i) for "leaving of the room", the agent receives a reward, $r^t = R(s^t, g_1) = \mathbb{1}(\text{out of the room})$, at step t; ii) for "blocking", the reward at step t is $r^t = R(s^t, g_2) = \pm\mathbb{1}(\text{opponent is out of the room})$. To simplify the policy learning, we define a discrete action space, which corresponds to applying forces with the same magnitude in one of the eight directions and "stop" (the agent's speed decreases to zero after applying necessary force).

The objective of the deep RL model is to train the policy network shown in Figure3 to maximize the expected return $E[\sum_{t=0}^{\infty} \gamma^t r^t]$ for each agent. The optimization was implemented using advantage actor critic (A2C) (Mnih et al., 2016) to jointly learn a policy (actor) $\pi : \mathcal{S} \times \mathcal{G} \mapsto \mathcal{A}$ which maps an agent's state and goal to its action, and a value function (critic) $V : \mathcal{S} \mapsto \mathbb{R}$. The two functions were trained as follows (assuming that entity i is an agent):

$$\nabla_{\theta_\pi} J(\theta_\pi) = \nabla_{\theta_\pi} \log \pi(a_i^t | s_i^t, g_i; \theta_\pi) A(s_i^t, g_i), \quad (1)$$

$$\nabla_{\theta_V} J(\theta_V) = \nabla_{\theta_V} \frac{1}{2} \left(\sum_{\tau=0}^{\infty} \gamma^\tau r_i^{t+\tau} \pm V(s_i^t, g_i; \theta_V) \right)^2, \quad (2)$$

where $A(s_i^t, g_i) = \sum_{\tau=0}^{\infty} \gamma^\tau r_i^{t+\tau} \pm V(s_i^t, g_i)$ is an estimate of the advantage of current policy over the baseline $V(s_i^t, g_i)$. We set $\gamma = 0.95$ and limit the maximum number of steps in an episode to be 30 (i.e., 1.5 s). Note that we train a network for each goal with the same architecture. In HH animations, an agent's policy depends on its opponent's policy. To achieve a joint policy optimization for both agents, we adopt an alternating training procedure: at each iteration, we train the policy of one of the agents by fixing its opponent's policy. In practice, we trained the polices by 3 iterations.

Inference of Physical and Social Events

Physics Inference

The first type of inference assesses the degree of violation of physics for each entity. To capture this measure, we used physical events to train a deep recurrent neural network (see

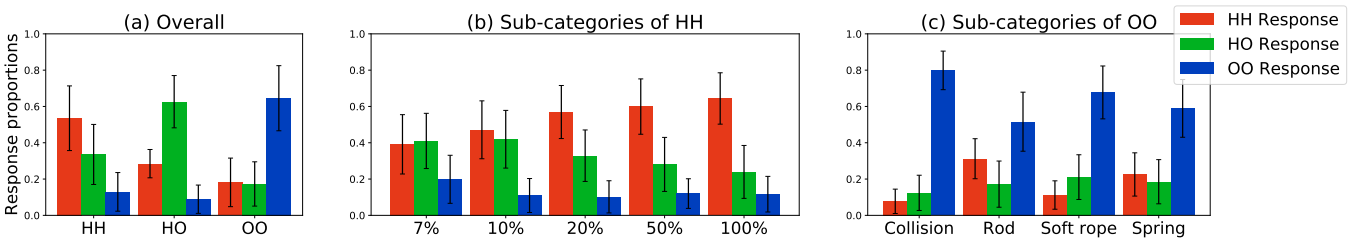


Figure 5: Human response proportions of interaction categories (a) and of the sub-categories (b,c) in Experiment 1. Error bars indicate the standard deviations across stimuli.

Figure 4) as an approximation to emulate intuitive physics. The network can predict the velocities of the two objects $\hat{\mathbf{v}}_i^t$, $i = 1,2$, given their past trajectories $\Gamma_i^t = \{s_i^\tau\}_{\tau=1}^t$. At each step, we feed a 28-dim feature vector to the network by concatenating the two dots' positions in the room, their relative positions to each other and to the five corners highlighted by the blue circles in Figure 4. We generated 2000 collision OO videos and trained the network on these videos with a 4-fold cross-validation. Using the trained network, we then conducted a step-by-step prediction of an entity's movements assuming it is an object. By comparing with the ground truth $\mathbf{v}_i^t$, we can evaluate to what degree an entity's motion is inconsistent with physics predictions:

$$D_i = \frac{1}{T}\sum_{t=1}^{T} ||\mathbf{v}_i^t \pm \hat{\mathbf{v}}_i^t||_2^2, \quad \forall i = 1,2. \tag{3}$$

Intention Inference

To evaluate the impression of whether a dot possesses intentions in the Heider-Simmel display, we estimate a value index (i.e., accumulated reward) from an entity's trajectory w.r.t. each possible goal. We first define a reward function:

$$R(s^t, g) = \frac{(\mathbf{x}_g^t \pm \mathbf{x}^t)^\top \mathbf{v}^t}{||\mathbf{x}_g^t \pm \mathbf{x}^t||_2 || \cdot \mathbf{v}^t||_2}, \tag{4}$$

where $\mathbf{x}^t$ and $\mathbf{v}^t$ are the position and velocity of an entity extracted from its state s^t, and $\mathbf{x}_g^t$ is the position of the goal. For "leaving the room", $\mathbf{x}_g^t$ is the door's position, whereas $\mathbf{x}_g^t$ denotes the position of the other entity for "blocking". Intuitively, this reward function evaluates whether the entity is moving towards certain goal locations. Consequently, we can compute the overall value by selecting the most likely goal:

$$V_i = \left[\max_{g\in\mathcal{G}} \frac{1}{T}\sum_{t=1}^{T} R(s_i^t, g)\right]_+, \quad \forall i = 1,2, \tag{5}$$

where $[x]_+ = \max(x, 0)$. Note that V_i defined here is different from the one in Eq. 2. Ranging from 0 to 1, a higher value of V_i indicates that the entity i shows a clearer intention and is more likely to be an agent. We remove the moments when the denominator in Eq. (4) is too small for the robustness of the value estimate. Considering the complexity of optimal planning in the continuous physical environment, the proposed value index offers a simplified measure of goal inference by inverse planning (Baker et al., 2009; Ullman et al., 2010).

Experiment 1

Participants

30 participants (mean age = 20.9; 19 female) were recruited from UCLA Psychology Department Subject Pool. All participants had normal or corrected-to-normal vision. Participants provided written consent via a preliminary online survey in accordance with the UCLA Institutional Review Board and were compensated with course credit.

Stimuli and Procedure

850 videos of Heider-Simmel animations were generated from our synthesis algorithm described above, with 500 HH videos (100 videos for each AD level), 150 HO videos, and 200 OO videos (50 videos for each sub-category). Videos lasted from 1 s to 1.5 s with a frame rate of 20 fps. By setting appropriate initial velocities, the average speeds of dots in OO videos were controlled to be the same as the average speeds of dots in HH with 100% ADs (44 pixel/s). The dataset was split into two equal sets; each contained 250 HH, 75 HO, and 100 OO videos. 15 participants were presented with set 1 and the other 15 participants were presented with set 2.

Stimuli were presented on a 1024×768 monitor with a 60 Hz refresh rate. Participants were given the following instructions: "In the current experiment, imagine that you are working for a security company. Videos were recorded by bird's-eye view surveillance cameras. In each video, you will see two dots moving around, one in red and one in green. Your task is to 'identify' these two dots based on their movement. There are three possible scenarios: human-human, human-object, or object-object." Videos were presented in random orders. After the display of each video, participants were asked to classify the video into one of the three categories.

Results

Human response proportions are summarized in Figure 5. Response proportion of human-human interaction swas ignificantly greater than the chance level 0.33 ($t(499) = 25.713$, $p < .001$). For HO animations, response proportion of human-object interaction was significantly greater than the other two responses ($p < .001$). Similarly, response proportion of object-object was greater than the other two responses ($p < .001$) for OO animations. These results reveal that human participants identified the main characteristics of different interaction types based on dot movements.

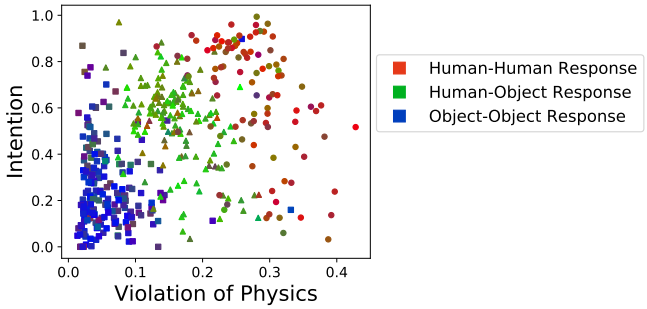


Figure 6: Constructed psychological space including HH animations with 100% animacy degree, HO animations, and OO animations. In this figure, a stimulus is depicted by a data point with coordinates derived by the model, and the colors of data points indicate the average human responses of this stimulus. The two coordinates of the space are the averaged measures between the two entities, as the measure of the degree of violation of physical laws (horizontal) and the measure of values indicating the presence of intention. The mark shapes of data points correspond to the interaction types used in the simulation for generating the corresponding stimuli (circle: HH, triangle: HO, square: OO).

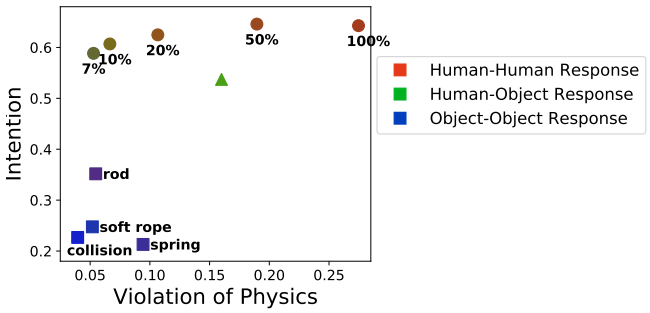


Figure 7: Centers of all types of stimuli.

Next, we examined human responses to the sub-categories within the HH and OO animations. We first used the animacy degree as a continuous variable and tested its effect on human responses in the HH animations. With increases in degree of animacy in HH, the response proportion of human-human interaction increased significantly as revealed by a positive correlation ($r = .42$, $p < .001$). This finding suggests that humans are sensitive to the animacy manipulation in terms of the frequency with which self-propelled forces occurred in the stimuli. For the OO animations, the response proportion for object-object interaction among the four sub-categories yielded significant differences ($F(3,196) = 34.42$, $p < .001$ by an ANOVA), with the most object-object responses in the collision condition, and the least in the rod condition. Pairwise comparisons among the four-categories show significant difference between collision and everything else ($p < .001$), between soft rope and rope ($p < .001$), and also between soft rope and string ($p = .018$); there is a marginally significant difference between rod and string ($p = .079$).

We then combined human responses and the model-derived measures for each animation stimulus to depict the unified psychology space for the perception of physical and social events. Figure 6 presents the distributions of 100 HH videos with 100% animacy degree, 150 HO videos, and 200 OO videos, all in this unified space. In this figure, an animation video is indicated by a data point with coordinates derived by the model, and the colors of data points indicate the average human responses of this stimulus. Specifically, the values of its RGB channels are determined by the average human-human responses in red, human-object responses in green, and object-object responses in blue. The mark shapes of data points correspond to the interaction type used in the simulation for generating the synthesized animations. The coordinates of each data point were calculated as the model-derived measures averaged across the two entities in an animation. The resulting space showed clear separations between the animations that were judged as three different types of interactions. Animations with more human-human interaction responses (red marks) clustered at the top-right corner, corresponding to great values of intention and strong evidence signaling the violation of physics. Animations with high responses for object-object interactions (blue marks), located at the bottom left of the space, show low values of intention index and little evidence of violation of physics. Animations with high responses for human-object interactions (green marks) fell in the middle of the space.

To quantitatively evaluate how well the model-derived space accounts for human judgments, we trained a classifier using the coordinates derived in the space shown in Figure 6 as input features (D and V for the indices of physical violation and intention respectively). For each ground-truth type of interactions $k \in \{\text{HH}, \text{HO}, \text{OO}\}$, we fit a 2D Gaussian distribution $p_k(D,V)$, using half of the stimuli as training data. Then for a given animation with the coordinates of (D,V), the classifier predicts $p(k|D,V) = \frac{p_k(D,V)}{\sum_k p_k(D,V)}$ for animations in the remaining half of the stimuli. The correlation between the model predictions and average human responses was 0.748 ($p < .001$) based on 2-fold cross-validation. Using a split-half reliability method, human participants showed an inter-subject correlation of 0.728 ($p < .001$). Hence, the response correlation between model and humans closely matched inter-subject correlations, suggesting a good fit of the unified space as a generic account of human perception of physical and social events based on movements of simple shapes.

We examined the impact of different degrees of animacy on the perception of social events, and how different subcategories of physical events affect human judgments on interaction types. The unified space provides a platform to compare these fine-grained judgments. Figure 7 shows the centers of the coordinates and the average responses for each of the sub-categories. We first found that, with a decreased degree of animacy, the intention index in HH animations was gradu-

ally reduced towards the level of HO animations. Meanwhile, human judgments of these stimuli varying from low to high degree of animacy transited gradually from human-object responses to human-human responses, consistent with the trend that the data points moved along the physics axis. Among all physical events, the rod and spring conditions showed the highest intention index and the strongest physical violation, respectively, resulting in a greater portion of human-human interaction responses than the other categories.

Experiment 2

In Experiment 1, human participants were asked to classify the three interaction types. But for human-object responses, the assignment of the roles to individual entities was not measured. In Experiment 2, we focused on stimuli that elicited the classification of human-object responses, and asked participants to report which dot was a human agent, and which dot was an inanimate object. Specifically, the role assignment in the human-object responses helps us identify some key characteristics in the psychological space that signal a human-object interaction.

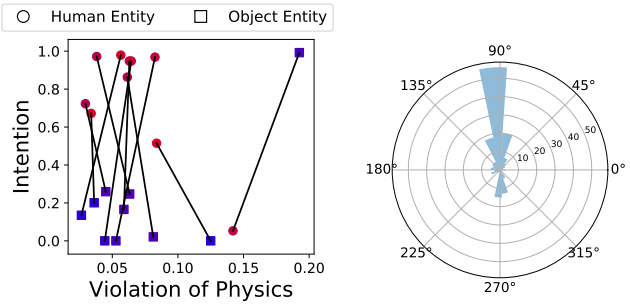
Methods

25 participants (mean age = 21.3; 19 female) were recruited from the UCLA Psychology Department Subject Pool. 216 videos were selected from Experiment 1 based on the criterion that more than 40% of subjects judged the HH videos or OO videos as human-object interaction. 201 videos were HH videos and the other 15 were OO videos.

The procedure was the same as Experiment 1 except that on each trial, subjects were asked to complete two tasks: first to judge the interaction type; then if the judgment was human-object, they were further asked to report which dot represented a human agent and which dot represented an object.

Results

We projected all entities onto the psychological space based on the model-derived measures for each individual entity, and connected a pair of the two entities that appeared in the same video. We visualized 10 animations that yielded high human-object response proportions and the most consistent role judgment among participants as shown in Figure 8a, where circles represent the dots that were frequently identified as humans, and squares represent the dots identified as objects. The resulting segments showed a common feature in that the connection of the two entities in the space depicted a near-vertical orientation, primarily due to high intention value for the human dot, and low intention value for the object dot. To further examine the orientations in the space for the human-object responses, we calculated the histogram of the orientations for animations judged as human-object interactions, which shows a high concentration around 90 degrees (see Figure 8b). This finding suggests that the two dots in the Heider-Simmel animations elicited similar degrees of physical violation, but one of them showed a much clearer intention. Note that this analysis excluded 38 stimuli in which participants did not show consistency in the role judgment (each entity was judged as a human or an object by exactly half of the participants).



(a) (b)

Figure 8: Human and model-simulation results in Experiment 2. (a) Representative cases of animations that elicited the human-object responses, located in the space with model-derived coordinates. The colors reflects average human responses of assigning a dot to the human role (red) and to the object role (blue). (b) Orientation histogram of the segments connected by the concurrent pairs of entities in an animation.

Conclusion

In this study, we propose a unified psychological space to account for human perception of physical and social events from movements of simple shapes in Heider-Simmel animations. The space consists of two primary dimensions: the intuitive sense of violation of physics, and the impression of intentions. We tested the space by measuring human responses when viewing a range of synthesized stimuli depicting human-human, human-object, and object-object interactions in the style of Heider-Simmel animations. We found that the constructed physics-intention space revealed clear separations between social and physical events as judged by humans. Furthermore, we trained a classification model based on the coordinates of each stimulus in this space. The resulting model was able to predict human classification responses at the same level as human inter-subject reliability.

The present paper provides a proof of concept that the perception of physical events and social events can be integrated within a unified space. Such common representation enables the development of a comprehensive computational model of how humans perceive and reason about physical and social scenes. Perhaps the most surprising finding in our work is that the classification result based on just the two measures reflecting the violation of physical laws and the estimate of intention can predict human judgment very well, reaching the same level as inter-subject correlation. The good fit to human responses across a range of Heider-Simmel stimuli demonstrates the great potential of using a unified space to study the transition from intuitive physics to social perception.

The main benefit of constructing this psychological space is to provide an intuitive assessment for general impressions of physical and social events. To build up such representation,

humans or a computation model may use various cues to detect intentions and/or physical violations; such cue-based detection is usually subjected to personal preferences. Instead of discovering a list of cues for distinguishing between physical events and social events, the proposed space offers an abstract framework for gauging how humans' intuitive senses of physics and intentions interplay in their perception of physical and social events.

This work provides a first step toward developing a unified computational theory to connect human perception and reasoning for both physical and social environments. However, the model has limitations. For example, the simulations are limited by a small set of goals, and the model requires predefined goals and good knowledge about the constrained physical environment. Future work should aim to extend the analysis to a variety of goals in social events (Thurman & Lu, 2014), to develop better goal inference, and to support causal perception in human actions (Peng et al., 2017). A more complete model would possess the ability to learn about physical environments based on partial knowledge, and to emulate a theory of mind in order to cope with hierarchical structures in the goal space. In addition, we have only examined human perception of physical and social events on short stimuli with only two entities. Generating longer stimuli with more entities and analyzing human perception on them will further help reveal the mechanisms underlying humans' physical and social perception.

Acknowledgement

We thank Ciaran Zhou, Claire Locke, Huiwen Duan, Suhwan Choi, and Zhibo Zhang for assistance in data collection. We also thank Dr. Tao Gao at UCLA for the helpful discussions. This research was supported by NSF Grant BCS-1655300 to HL, and by DARPA XAI N66001-17-2-4029 and ONR MURI project N00014-16-1-2007 to SZ.

References

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329-349.

Dittrich, W. H., & Lea, S. E. (1994). Visual perception of intentional motion. *Perception*, *23*(3), 253-268.

Gao, T., McCarthy, G., & Scholl, B. J. (2010). The wolfpack effect: Perception of animacy irresistibly influences interactive behavior. *Psychological Science*, *21*, 1845-1853.

Gao, T., Newman, G. E., & Scholl, B. J. (2009). The psychophysics of chasing: A case study in the perception of animacy. *Cognitive Psychology*, *59*(2), 154-179.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *American Journal of Psychology*, *57*(2), 243-259.

Isik, L., Koldewyn, K., Beeler, D., & Kanwisher, N. (2017). Perceiving social interactions in the posterior superior temporal sulcus. *Proceedings of the National Academy of Sciences*, *114*(43).

Kassin, S. (1981). Heider and simmel revisited: Causal attribution and the animated film technique. *Review of Personality and Social Psychology*, *3*, 145-169.

Kerr, W., & Cohen, P. (2010). Recognizing behaviors and the internal state of the participants. In *Proceedings of ieee 9th international conference on development and learning.*

Kubricht, J. R., Holyoak, K. J., & Lu, H. (2017). Intuitive physics: Current research and controversies. *Trends in cognitive sciences*, *21*(10), 749–759.

Michotte, A. E. (1963). *The perception of causality (t. r. miles, trans.)*. London, England: Methuen & Co. (Original work published 1946).

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., ... Kavukcuoglu, K. (2016). Asynchronous methods for deep reinforcement learning. In *International conference on machine learning (icml).*

Pantelis, P. C., Baker, C. L., Cholewiak, S. A., Sanik, K., Weinstein, A., Wu, C.-C., ... Feldman, J. (2014). Inferring the intentional states of autonomous virtual agents. *Cognition*, *130*, 360379.

Peng, Y., Thurman, S., & Lu, H. (2017). Causal action: A fundamental constraint on perception and inference about body movements. *Psychological Science*, 0956797617697739.

Proffitt, D. R., & Gilden, D. L. (1989). Understanding natural dynamics. *Journal of Experimental Psychology: Human Perception and Performance*, *15*(2), 384-393.

Scholl, B. J., & Tremoulet, R. D. (2000). Perceptual causality and animacy. *Trends in Cognitive Sciences*, *4*(8), 299-309.

Shu, T., Peng, Y., Fan, L., Lu, H., & Zhu, S.-C. (2018). Perception of human interaction based on motion trajectories: From aerial videos to decontextualized animations. *Topics in Cognitive Science*, *10*(1), 225–241.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., ... Hassabis, D. (2017). Mastering the game of go without human knowledge. *Nature*, *550*(7676), 354-359.

Thurman, S. M., & Lu, H. (2014). Perception of social interactions for spatially scrambled biological motion. *PloS one*, *9*(11), e112539.

Ullman, T. D. (2015). *On the nature and origin of intuitive theories: Learning, physics and psychology*. Unpublished doctoral dissertation, Massachusetts Institute of Technology.

Ullman, T. D., Baker, C. L., Macindoe, O., Evans, O., Goodman, N., & Tenenbaum, J. B. (2010). Help or hinder: Bayesian models of social goal inference. In *Proceedings of advances in neural information processing systems* (p. 1874-1882).

Seeing the big picture: Do some cultures think more abstractly than others?

Amritpal Singh
Cornell University, Ithaca, New York, United States

Qi Wang
Cornell University, Ithaca, New York, United States

Daniel Casasanto
Cornell University, Ithaca, New York, United States

Abstract

Do some cultures think more abstractly than others? According to tests of formal logic and rule-based reasoning, Westerners tend to think more abstractly than East Asians. Yet, rule-based reasoning is only one type of abstract thinking. More generally, thinking abstractly involves discerning relationships and seeing the big picture. Here we argue that previous tests of attention, perception, and memory can be interpreted as showing that East Asians tend to think more abstractly than Westerners. To test this hypothesis directly we gave a validated measure of abstract thinking (Vallacher & Wegner, 1989) to Chinese and US individuals. Participants chose either abstract or concrete definitions of events. Across six independent national samples (total N=1,798), Chinese participants tended to construe events more abstractly, and US participants more concretely. Within China, more independent (Western-like) groups chose more concrete definitions. Together, these results challenge the generalization that Westerners have a greater propensity for abstract thought.

Measuring Creative Ability in Spoken Bilingual Text: The Role of Language Proficiency and Linguistic Features

Stephen Skalicky (stephen.skalicky@vuw.ac.nz)
School of Linguistics and Applied Language Studies, Victoria University of Wellington
Wellington, New Zealand

Scott A. Crossley (scrossley@gsu.edu)
Department of Applied Linguistics and ESL, Georgia State University
Atlanta, Georgia, USA

Danielle S. McNamara (dsmcnama@asu.edu)
Department of Psychology, Arizona State University
Phoenix, Arizona, USA

Kasia Muldner (kasia.muldner@carleton.ca)
Institute of Cognitive Science, Carleton University
Ottawa, Ontario, Canada

Abstract

Whereas first language (L1) research has demonstrated that perceptions of creative ability are influenced by the complexity and diversity of language used to answer verbal tests of creativity, relatively little is known about the linguistic components of bilingual creative task performance. In this study, we analyze written transcripts of speech produced by 466 Japanese learners of English produced during a creative narrative task for features related to linguistic and cognitive dimensions of creativity. Then, we extract various linguistic features and test whether these features can predict human perceptions of creativity for the transcripts. Unlike L1 data, results suggest text length and L2 proficiency comprise the most parsimonious explanation of creativity scores in this L2 data. At the same time, linguistic features related to positive sentiment explained a significant yet small amount of additional variance in perceptions of creativity, suggesting texts with more positive language were perceived to be more creative.

Keywords: creativity, NLP, language proficiency, bilingualism

Introduction

The relationship between bilingualism and creativity can be approached from a number of perspectives. One is to investigate how learning a second language (L2) impacts creativity. Here, research has shown benefits of language learning, with high-proficiency bilinguals outperforming their monolingual and lower L2 proficiency peers on tests of creative ability (Kharkhurin, 2009; Leikin, 2013; Ricciardelli, 1992). Reasons for this difference have been attributed to the growth of language knowledge that naturally comes with mastering additional languages, suggesting that a specific cognitive ability (i.e., creativity) may be directly associated with language knowledge. Another approach to is to investigate the role of creativity in second language acquisition (SLA). For instance, researchers SLA have highlighted the facilitative role that creativity, play, and humor in an L2 can have on language learning (Cook, 2000; Pomerantz & Bell, 2007).

Yet another approach involves investigating links among creative ability, language use, and language knowledge in order to shed light on how language and cognition (specifically, creative ability) influence one another. One method for doing so, and the one that we adopt in the present work, is by determining whether linguistic features pattern with creativity.

The overarching objective of this study is to better understand how L2 proficiency and linguistic features relate to perceptions of creativity. To do so, this study examines linguistic features in 466 transcribed speech samples produced during an English L2 oral proficiency exam. The speech samples were part of the oral proficiency interviews found in the NICT Japanese Leaners of English (JLE) corpus (Izumi, Uchimoto, & Isahara, 2004; Tono et al., 2001). We trained raters to make creativity judgements for each of the samples. The linguistic features of the samples were then analyzed using automatic text analysis tools and associations between these features and the human judgments of creativity were assessed. This approach allowed us to examine the strength of the relations among L2 language proficiency, linguistic features of L2 speech, and expert raters' perceptions of creativity.

Creativity

Psychologists have defined creativity as a cognitive construct that represents the ability to develop novel and effective solutions to a problem (Kaufman, Plucker, & Baer, 2008; Runco & Jaeger, 2012). One common method for assessing creativity is through the use of divergent thinking tests, where a participant or group of participants generates as many

solutions to a problem that they can in a set amount of time (Runco, 2013). These tests are then most commonly scored for four measures: fluency (total number of ideas), flexibility (range of idea types), elaboration (ability to expand on ideas), and originality (uniqueness of ideas when compared to other participants' answers). In general, participants who score higher on these four features are thought to be more creative than those who score lower. Due to the frequent use of divergent thinking tests in creativity research, these four components have gained widespread acceptance as valid measures of creativity (Kaufman et al., 2008).

Bilingualism and Creativity

One specific application of divergent thinking tests has been to investigate whether bilinguals are more or less creative than monolinguals (Kharkhurin, 2009). A consistent finding from these studies is that the degree of bilingualism or relative proficiency in bilinguals' L2s is strongly related to creative performance (Kharkhurin, 2008). Specifically, language users with more balanced bilingualism (i.e., relatively similar proficiency between a user's two languages) significantly outperform those who report lower L2 proficiency compared to their L1 (Kharkhurin, 2011; Lee & Kim, 2011). These results have been replicated among different language users, including German-English and Dutch-English bilinguals (Hommel, Colzato, Fischer, & Christoffels, 2011) as well as among Hebrew-Russian bilingual children (Leikin, 2013).

Bilingual creative performance has also been identified as an important component of L2 learning. Specifically, language learners who experiment with the sounds, meanings, and forms of a language are a) better equipped to deduce the rules of a language, b) gain more agency over the language they are learning, c) construct more engaging learning environments, and d) enhance interaction with other learners (Bell, 2005; Cook, 2000). Although relatively high L2 proficiency is required to take part in complex forms of language play such as interpersonal humor (Bell, 2005), even lower proficiency L2 learners have demonstrated usage of less complex forms of play (Bell, Skalicky, & Salsbury, 2014).

Linguistic Features, Bilingualism, and Creativity

Learning a second language naturally involves increased knowledge of lexical items and word associations in that language. In English, lexical features such as polysemous word senses, hypernymic categories, and psycholinguistic measures of lexical sophistication have all been shown to change over time as learners increase their L2 English proficiency. Specifically, as L2 English learners become more proficient, they develop more polysemous and less frequent senses for English words (Crossley, Salsbury, & McNamara, 2010), more diverse hypernymic relations among word categories in English (Crossley, Salsbury, & McNamara, 2009), and demonstrate higher levels of lexical sophistication in English (e.g., more abstract lexical items that are less rooted in the immediate context; Salsbury et al., 2011).

Several of these same linguistic features have been associated with higher performance on tests of creativity in English as an L1. For example, words generated by individuals rated higher for creativity have more remote associations among concepts as measured through computationally-derived association strengths such as Latent Semantic Analysis (Acar & Runco, 2014; Beketayev & Runco, 2016; Dumas & Dunbar, 2014). Higher creativity scores are also associated with higher levels of lexical sophistication (i.e., more infrequent, varied, and complex language) and semantic cohesion (Skalicky, Crossley, McNamara, & Muldner, 2017).

Current Study

The current study has two goals. The first is to examine the extent to which L2 English proficiency is associated with perceptions of creativity during an oral picture description task among Japanese-L1 English-L2 bilinguals of eight different L2 proficiency levels. The second is to investigate whether linguistic features of the language produced during the task are predictive of perceptions of creativity. Because SLA research has demonstrated that various aspects of language such as lexical sophistication change over time as one gains proficiency in English as an L2, we examine the extent that differences in creative output based on L2 proficiency are associated with quantifiable features of language. By identifying language features associated with perceptions of creativity, we aim to further define linguistic aspects of L2 creativity, identify associations between creativity and proficiency in a second language, and provide additional explanations for differences in creative performance among bilinguals of differing proficiency levels. The following research questions guide our study:

1. What role does L2 English language proficiency have for human perceptions of creativity during an English L2 oral proficiency exam?
2. Do linguistic features explain differences in creativity scores when controlling for L2 English proficiency?

Method

Corpus

We used a subset of the NICT Japanese Learner English Corpus to collect creativity ratings for L2 speakers of English (Izumi et al., 2004). The JLE comprises over 1200 recorded speech samples of Japanese learners of English who completed an interview activity designed to assess their oral English proficiency. The JLE data also includes the oral proficiency scores for each interviewee assigned by the interviewer at the time of the interview. The scores were derived using the Standard Speaking Test scoring method (Tono et al., 2001), where 2-3 raters used a holistic rubric based on the American Council on the Teaching of Foreign Languages (ACTFL) proficiency guidelines to place

interviewees into one of nine different levels based on their oral proficiency (1 being the lowest and 9 being the highest).

Each interview was conducted between a test taker and a test administrator. The interviews lasted approximately 10-15 minutes and included three interview tasks. In this study, we focus on the final interview task, which was a picture description task where the interviewee was asked to construct a story based on information depicted in a picture or a set of pictures. We focused on this task because it provided the strongest potential for the test takers to produce creative ideas in that they were given the freedom to embellish and elaborate on events in the story as they constructed it. The interviewer provided minimal feedback beyond confirmation checks and backchanneling, ensuring that all the ideas produced during this task belonged to the interviewee. Within the picture sequence description task, there were ten possible picture sets that depicted scenes such as camping, visiting a zoo, eating at a restaurant, and shopping in a grocery store.

We constructed a subset of the JLE corpus by randomly selecting 250 texts from male and female speakers respectively ($N = 500$) while also sampling equally from each proficiency level (levels 2-9 with the exception of 1, which was rare). For each file we manually removed all text not associated with the picture sequence description task and all speech produced by the interviewer, leaving just the text that was on topic and delivered by the test taker. In order to ensure enough coverage for our linguistic measurements, we further removed any text containing less than 50 words (34 texts), resulting in a final JLE subset of 466 texts (237 female, 229 male). The average number of words per text in this final subset was 140.700 ($SD = 62.551$). The resulting distribution of proficiency levels approximated a normal distribution ($M = 5.361$, $SD = 1.744$).

Human Ratings

We developed an analytic rubric to obtain creativity ratings for each text in our dataset. The rubric contained seven different subscales with a range of 1 (does not meet the criterion in any way) to 6 (meets the criterion in every way). The subscales were divided into two larger categories: IDEAS and STYLE. The IDEAS category contained four subscales related to cognitive definitions of creativity: ideation (the speaker produced a large number of different ideas), originality (the speaker's ideas were original when compared to other speakers completing the same task), elaboration (the speaker included additional information elaborating on their ideas) and appropriateness (the speaker's ideas created an effective narrative). The STYLE category contained three subscales related to linguistic creativity: humor (the speaker produced at least one idea intending to provoke humor or amusement), metaphor and simile (the speaker produced ideas which made conceptual comparisons), and word play (the speaker played with the sounds or meanings of words).

Two native English-speaking research assistants were trained on the creativity rubric using a separate subset of 65 JLE texts. Raters were informed that the distance between each number on the rating scale was equal. After calibrating on the initial 65 texts, the raters then independently scored the remaining 466 texts for creativity. The raters were not aware that the samples were from English L2 learners. After scoring, raters were able to adjudicate disagreements greater than two for any of the subscales. Raters reported almost no instances of humor, metaphor and simile, or wordplay in the corpus, and thus these subscales were removed from the study. Table 1 displays the final, adjudicated kappa scores and correlations between the two raters for each of the remaining five subscales. After adjudication, the raters' scores were averaged for each subscale and text.

Table 1: Rater agreement

Subscale	*r*	Kappa
Ideational Fluency	0.830	0.830
Originality	0.825	0.822
Elaboration	0.739	0.738
Appropriateness	0.785	0.781

Linguistic Feature Selection

Based on prior work reporting associations between lexical sophistication, cohesion, and creativity in L1 English research (Acar & Runco, 2014; Beketayev & Runco, 2016; Dumas & Dunbar, 2014; Skalicky et al., 2017), we hand-selected a range of lexical indices representative of these constructs. We also included features related to sentiment in order to explore whether these measures might explain further explain variance in creativity scores. We obtained our measures of lexical sophistication, sentiment, and cohesion using three freely-available automatic text analysis tools, TAALES v2.2, SEANCE, and TAACO, respectively (see Crossley, Kyle, & McNamara, 2016a, 2016b; Kyle, Crossley, & Berger, 2017).

For lexical sophistication, we included linguistic indices of word frequency, word concreteness (i.e., how abstract a word's meaning is), contextual diversity and distinctiveness (i.e., the range of different contexts a word occurs in), word meaningfulness (i.e., number of associations with other words), word polysemy (i.e., the number of different senses a word form has), and word recognition and naming norms (i.e., average time to recognize and name English words). For cohesion, we included features measuring the type-token ratio (i.e., lexical diversity) and number of repeated content words in each text. Finally, for sentiment, we used features measuring the overall valence of a text (i.e., use of positive or negative vocabulary). We used measures calculated for content words (e.g., nouns, verbs, adjectives) only.

Statistical Analysis

We first conducted a principal component analysis using the raters' scores for the four subscales in the IDEAS category from the creativity rubric to develop a single, weighted creativity score to be used as the dependent variable. We then conducted correlations between the creativity score and the oral proficiency scores provided with the JLE corpus, as well as between the creativity score and text length (i.e., number

of content word types in each text). We included text length as a variable because longer texts would include more ideas and thus be biased to higher ideation scores (and therefore higher creativity scores). Then, we controlled the linguistic features based on correlations with the dependent variable and also controlled for multicollinearity using correlations and variance inflation factors.

Next, in order to test whether L2 proficiency and the linguistic features related to lexical sophistication, sentiment, and cohesion were predictive of the creativity scores, we performed comparisons between linear regression models in order to obtain the most parsimonious model (i.e., the model that explained the largest amount of variance with the fewest number of predictor variables).

Results

Principal Component Analysis

A principal component analysis (PCA) was conducted on the averaged ratings of ideation, originality, elaboration, and appropriateness from the analytic rubric for the 466 texts in our subset of the JLE corpus. A Bartlett's test of sphericity was statistically significant ($\chi^2 = 717.179$, $df = 6$, $p < .001$), and the Kaiser-Meyer-Olkin measure of sampling adequacy reported .672, representing acceptable ability for the PCA to yield distinct, reliable factors (Field, 2013). A single component containing all four variables accounted for 59.463% of cumulative variance with an eigenvalue of 2.378. The individual subscale loadings were: ideation = .913, elaboration = .892, appropriateness = .782, originality = .369. In order to calculate a single score reflective of the different strengths of these loadings we multiplied each human score for each subscale for each text by its respective loading and summed these values per text, obtaining a weighted sum component score for each text (DiStefano, Zhu, & Mîndril, 2009), which we refer to as the creativity score (Min = 8.238, Max = 14.780, $M = 12.526$, $SD = 1.545$).

Linguistic Feature Reduction

Using the output from the automatic text analysis programs, we first reduced the number of variables by only including variables of interest that had a significant and meaningful linear relation (i.e., absolute $r > .1$) with the dependent variable (i.e., the creativity score). We then controlled for multicollinearity using variance inflation factors (VIF), removing any variable with a VIF greater than 2. The end result was a selection of seven linguistic indices that demonstrated no strong multicollinearity and possessed a significant linear relation with the dependent variable.

These features were: average Age of Acquisition, which is based on averaged self-reported ratings of the age English users first understood 30,000 different English words collected from over 800,000 English speakers in the United States (Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012), average Spoken Word Frequency calculated from the Corpus of Contemporary American English, The University of South Florida Free Association Norms (i.e., the average number of words a subject voices when presented with a particular word; Nelson, McEvoy, & Schreiber, 1998), LSA Average Top Three Cosine (average LSA cosine values for the top three related words in each text), Vader Positive Sentiment (compound score measuring the overall positive sentiment in a text; Hutto & Gilbert, 2014), Number of Content Word Types (our measure of text length), and Number of Repeated Content Word Lemmas (divided by total text length). Table 2 displays these variables and their correlations with the creativity score, along with the correlation between L2 proficiency level and creativity.

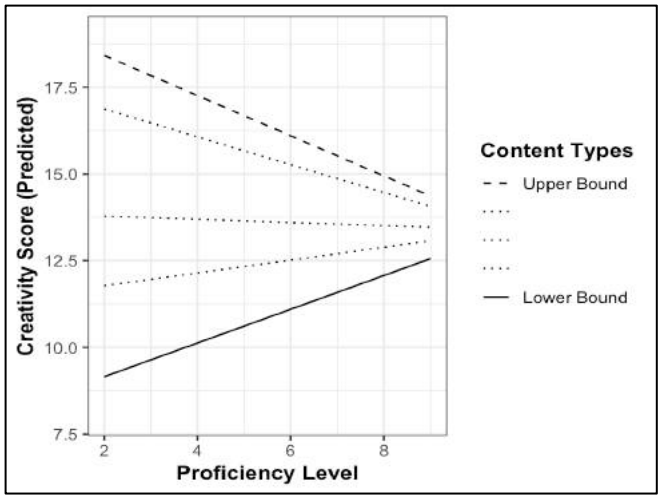


Figure 1: Interaction between English proficiency and text length (number of content word types). Upper and lower bounds represent the minimum and maximum values for number of content word types.

Regression Models

Based on the correlations among creativity, L2 proficiency level, and text length and initial model exploration, the results suggested a large amount of the variance in raters' creativity scores could be captured in a linear regression model fit with L2 proficiency, text length, and an interaction between text length and L2 proficiency. This model explained approximately 48% of the variance in raters' creativity scores ($R^2 = .475$, $F[3, 462] = 141.100$). The significant interaction between text length (i.e., number of content word types) and L2 proficiency indicated that differences in text length at higher L2 proficiency levels had significantly less effect on raters' perceptions of creativity when compared to lower levels of L2 proficiency. Specifically, at lower levels of L2 proficiency, texts with a higher number of content word types were rated significantly higher for creativity, and this effect attenuated significantly at higher levels of L2 proficiency. This interaction is visually plotted in Figure 1, and Table 3 displays the standardized beta coefficients and 95% confidence intervals for the terms in the model.

We then tested whether the separate inclusion of each of the remaining six predictor variables would significantly improve the baseline model based on changes in adjusted R^2 by comparing different linear regression models using the

anova() command in R. Table 4 summarizes the results of these comparisons. As can be seen, only the Vader Positive Sentiment index significantly increased the adjusted R^2 of the baseline model, with an increase of .07% variance explained. The remaining linguistic features did not explain any significant amount of additional variance, further suggesting that perceptions of creativity were strongly associated with L2 proficiency level and the amount of text produced by each participant.

Table 2. Correlations between predictor variables and the creativity score.

Index	*r*
Number of Content Word Types	0.643
L2 English Proficiency Level	0.449
Average Age of Acquisition	0.288
LSA (mean top three cosine)	-0.281
Free Association Norms (USF)	-0.240
Vader Positive Sentiment	0.158
Repeated Content Lemmas	0.134
Spoken Word Frequency (COCA)	0.118

Discussion

Creativity and L2 English Proficiency

Our first research question asked whether L2 English proficiency influenced raters' perceptions of creativity among our speech samples. The moderate correlation between the creativity score and L2 proficiency in Table 2 suggests a positive association between these features. This is further supported by the baseline regression model (Table 3), which included a significant positive effect for L2 proficiency (moderated by text length, see below). Together, these results provide an additional piece of evidence suggesting that a higher L2 proficiency level is associated with greater perceptions of creativity among the creativity raters. This finding aligns well with prior research into bilingual creative performance, which also reported greater creativity levels among bilinguals with higher L2 proficiency (Hommel et al., 2011; Kharkhurin, 2011; Lee & Kim, 2011; Leikin, 2013).

Our findings also suggest that this effect was moderated by text length, in that the overall length of the participants' picture description narratives (i.e., number of content word types) was more strongly associated with raters' perceptions of creativity at lower compared to higher levels of L2 proficiency. Thus, the manifestation of L2 proficiency as the ability to produce more language may be the driving determinant between higher creativity scores and L2 proficiency, as the ability to produce more language allowed for the opportunity to produce more ideas, and therefore receive higher ideation ratings and thus higher creativity scores.

Table 3: Baseline model explaining variance in raters' perceptions of creativity.

Model Term	*Estimate*	*SE*	*t*	*p*	*5% CI*	*95% CI*
(Intercept)	11.977	0.196	61.140	< .001	11.654	12.300
L2 English Proficiency Level	0.132	0.035	3.804	< .001	0.075	0.189
Text Length	1.900	0.176	10.788	< .001	1.609	2.190
L2 English Proficiency Level * Text Length	-0.178	0.028	-6.255	< .001	-0.225	-0.131
Adjusted R^2 = .475, F(3, 462) = 141.100. Estimate represents standardized beta coefficient as all predictor variables were z- scored before being entered into the model.						

Table 4: Comparisons between baseline model and models with different linguistic features.

Model Term	R^2	Adjusted R^2	*F*	*p*	R^2 Difference (Adjusted)
Baseline Model	0.478	0.475	141.099	NA	NA
Vader Positive Sentiment	0.487	0.482	109.275	0.006	0.007
Spoken Word Frequency (COCA)	0.478	0.474	105.597	0.962	0.001
Free Association Norms (USF)	0.480	0.476	106.513	0.167	0.001
LSA (mean top three cosine)	0.478	0.474	105.657	0.721	0.001
Age of Acquisition	0.478	0.474	105.677	0.680	0.001
Repeated Content Lemmas	0.478	0.474	105.684	0.667	0.001
Note: DF for all comparison models = (4, 461). Baseline model R syntax = *creativity ~ L2 English Proficiency Level + Text Length + L2 English Proficiency Level:Text Length*. *F* and *p* values correspond to change in R^2 from baseline model.					

Linguistic Features and English Proficiency

Our second research question asked whether linguistic features explained differences in creativity scores while taking L2 English proficiency into account. Early model exploration as well as a series of hierarchical linear regression comparisons suggested that almost all of the linguistic features selected for this study failed to predict any

meaningful amount of variance beyond the effect of L2 proficiency and text length, which combined to explain nearly 50% (Adjusted R^2 = .475) of the variance in the raters' perceptions of creativity, suggesting a relatively strong effect; see Table 4. The interaction between L2 proficiency and text length demonstrates that while text length was a strong, significant predictor of creativity scores, this effect was significantly stronger at lower L2 proficiency levels. Specifically, while texts with a greater number of content word types predicted increased creativity scores for participants across all eight proficiency levels, this effect was much stronger for participants who received lower L2 proficiency scores by the interviewer.

When comparing the difference in creativity scores between the upper bound and the lower bound of text length in Figure 1 (i.e., the minimum and maximum values for number of content word types), this difference attenuates for participants with higher L2 proficiency scores. This suggests that while differences in creativity scores at lower L2 proficiency levels are strongly predicted by the ability to produce more words (and therefore more ideas), this was not the case at the higher L2 proficiency levels. At higher L2 proficiency levels, variation in creativity scores based on total number of content word types was relatively low, suggesting that other features of the texts may have influenced the raters' creativity scores at higher L2 proficiency levels. However, these additional features, if any, were not captured in any of the linguistic features provided by our automatic text analysis tools. Therefore, unlike results reported in the L1 data, it is difficult at this time to draw concrete connections between specific linguistic features and bilingual performance on tests of creativity. It may be the case that additional linguistic features not included in the current study can explain variance in creativity at higher levels of L2 proficiency, providing ample opportunity for future research.

Aside from text length, one index, Vader Positive Sentiment, did result in a significantly better regression model fit, but only by approximately .07% of variance explained, suggesting that this index had a relatively weak effect. Nonetheless, it is still worth considering why this index may have provided a significant amount of additional variance explained. The Vader Positive Sentiment index is a component score derived from formulas specifically designed to measure sentiment in shorter texts, especially those used in social media (Hutto & Gilbert, 2014). The coefficient for the Vader Index was .148 (intercept = 11.881), suggesting a positive relation between positive sentiment and perceptions of creativity. Thus, narratives with more positive vocabulary may have appeared more creative to the raters in this study. Perhaps narratives with more positive language reflects a greater intent by the speakers in the corpus to create a unique story, as compared to narratives that were more factual descriptions of events. It would thus be worthwhile to further consider the role of sentiment in linguistic investigations of creative performance, as this would help identify links between specific types of linguistic knowledge and the cognitive construct of creativity.

Conclusion and Limitations

Previous investigations of bilingual creativity have reported a tendency for bilinguals with greater L2 proficiency to outperform those with lower L2 proficiency on standardized tests of creativity. The results from the current study support these claims while raising further questions. Specifically, we observed that increased levels of L2 proficiency were associated with higher perceptions of creativity, but this effect was moderated by the length of the speech samples. Moreover, while our results identified Vader Positive Sentiment as a significant linguistic predictor of creativity, this (and our other linguistic features) was overshadowed by the strong effect of text length. As a whole, these results suggest that there may be an L2 proficiency threshold for bilingual creativity, in that raters attended to additional linguistic features beyond text length only for speakers with relatively higher levels of L2 proficiency. In the future, it may be helpful to incorporate diversity-based linguistic information based on the prompts in order to control for potential vocabulary differences among the different prompts, which may influence the raters' perception of creativity (Chiru & Rebedea, 2017).

One final consideration is that the L2 English proficiency measure used in the current study was based solely on oral L2 proficiency at the time of the picture description task. Previous research in bilingual creativity has relied on proficiency assessments based on vocabulary knowledge tests as well as participant self-ratings of L2 proficiency and levels of bilingualism, which captures receptive vocabulary knowledge (i.e., reading and listening ability). The JLE L2 proficiency scores, on the other hand, are a measure of productive vocabulary knowledge (i.e., speaking and writing skill), and productive vocabulary size is typically smaller than receptive size (Schmitt, 2008). However, receptive and productive vocabulary knowledge are inextricably linked, suggesting that the JLE oral proficiency score is also a correlate of receptive L2 vocabulary knowledge (Webb, 2008). In all, these findings further highlight the association between bilingualism and the cognitive ability of creativity while providing avenues for future research.

References

Acar, S., & Runco, M. A. (2014). Assessing associative distance among ideas elicited by tests of divergent thinking. *Creativity Research Journal*, *26*, 229–238.

Beketayev, K., & Runco, M. A. (2016). Scoring divergent thinking tests by computer with a semantics-based algorithm. *Europe's Journal of Psychology*, *12*, 210.

Bell, N. D. (2005). Exploring L2 language play as an aid to SLL: A case study of humour in NS-NNS interaction. *Applied Linguistics*, *26*, 192–218.

Bell, N. D., Skalicky, S., & Salsbury, T. (2014). Multicompetence in L2 language play: A longitudinal case study. *Language Learning*, *64*, 72–102.

Chiru, C.-G., & Rebedea, T. (2017). Profiling of participants in chat conversations using creativity-based heuristics. *Creativity Research Journal*, *29*, 43–55.

Cook, G. (2000). *Language play, language learning*. Oxford University Press.

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016a). Sentiment analysis and social cognition engine (SEANCE): An automatic tool for sentiment, social cognition, and social-order analysis. *Behavior Research Methods*, *49*, 803–821.

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016b). The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods*, *48*, 1227–1237.

Crossley, S. A., Salsbury, T., & McNamara, D. (2009). Measuring L2 lexical growth using hypernymic relationships. *Language Learning*, *59*, 307–334.

Crossley, S. A., Salsbury, T., & McNamara, D. (2010). The development of polysemy and frequency use in English second language speakers. *Language Learning*, *60*, 573–605.

DiStefano, C., Zhu, M., & Mîndril, D. (2009). Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research, and Evaluation*, *14*, 1–11.

Dumas, D., & Dunbar, K. N. (2014). Understanding fluency and originality: A latent variable perspective. *Thinking Skills and Creativity*, *14*, 56–67.

Field, A. (2013). *Discovering statistics using IMB SPSS Statistics: And sex and drugs and rock "n" roll* (4th ed.). Sage.

Hommel, B., Colzato, L. S., Fischer, R., & Christoffels, I. K. (2011). Bilingualism and creativity: Benefits in convergent thinking come with losses in divergent thinking. *Frontiers in Psychology*, *2*, 1–5.

Hutto, C., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Eighth International Conference on Weblogs and Social Media (ICWSM-14)*, 216–225.

Izumi, E., Uchimoto, K., & Isahara, H. (2004). The NICT JLE Corpus: Exploiting the language learner's speech database for research and education. *International Journal of the Computer, the Internet and Management*, *12*, 119–125.

Kaufman, J. C., Plucker, J. A., & Baer, J. (2008). *Essentials of creativity assessment*. Hoboken, NJ: John Wiley & Sons.

Kharkhurin, A. V. (2008). The effect of linguistic proficiency, age of second language acquisition, and length of exposure to a new cultural environment on bilinguals' divergent thinking. *Bilingualism: Language and Cognition*, *11*, 225–243.

Kharkhurin, A. V. (2009). The role of bilingualism in creative performance on divergent thinking and invented alien creatures tests. *The Journal of Creative Behavior*, *43*, 59–71.

Kharkhurin, A. V. (2011). The role of selective attention in bilingual creativity. *Creativity Research Journal*, *23*, 239–254.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, *44*, 978–990.

Kyle, K., Crossley, S., & Berger, C. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): version 2.0. *Behavior Research Methods*, *50*, 1030–1046.

Lee, H., & Kim, K. H. (2011). Can speaking more languages enhance your creativity? Relationship between bilingualism and creative potential among Korean American students with multicultural link. *Personality and Individual Differences*, *50*, 1186–1190.

Leikin, M. (2013). The effect of bilingualism on creativity: Developmental and educational perspectives. *International Journal of Bilingualism*, *17*, 431–447.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). *The University of South Florida word association, rhyme, and word fragment norms [Database]*. Retrieved from w3.usf.edu/FreeAssociation.

Pomerantz, A., & Bell, N. D. (2007). Learning to play, playing to learn: FL learners as multicompetent language users. *Applied Linguistics*, *28*, 556–578.

Ricciardelli, L. A. (1992). Creativity and bilingualism. *The Journal of Creative Behavior*, *26*, 242–254.

Runco, M. A. (Ed.). (2013). *Divergent thinking and creative potential*. New York, NY: Hampton Press.

Runco, M. A., & Jaeger, G. J. (2012). The standard definition of creativity. *Creativity Research Journal*, *24*, 92–96.

Salsbury, T., Crossley, S. A., & McNamara, D. S. (2011). Psycholinguistic word information in second language oral discourse. *Second Language Research*, *27*, 343–360.

Schmitt, N. (2008). Instructed second language vocabulary learning. *Language Teaching Research*, *12*, 329–363.

Skalicky, S., Crossley, S. A., McNamara, D. S., & Muldner, K. (2017). Identifying creativity during problem solving using linguistic features. *Creativity Research Journal*, *29*, 343–353.

Tono, Y., Kaneko, T., Isahara, H., Saiga, T., Izumi, E., Narita, M., & Kaneko, E. (2001). The Standard Speaking Test (SST) Corpus: A 1 million-word spoken corpus of Japanese learners of English and its implications for L2 lexicography. *Second Asialex International Congress, Korea*, 257–262.

Webb, S. (2008). Receptive and productive vocabulary sizes of L2 learners. *Studies in Second Language Acquisition*, *30*, 79–95.

What's Lagging in our Understanding of Interruptions?: Effects of Interruption Lags in Sequential Decision-Making

Jennifer Sloane (j.sloane@unsw.edu.au)
Chris Donkin (c.donkin@unsw.edu.au)
Ben Newell (ben.newell@unsw.edu.au)
Garston Liang (garston.liang@unsw.edu.au)
School of Psychology, University of New South Wales
Sydney, 2052, Australia

Abstract

Interruptions are an inevitable part of every day life. Previous research suggests that interruptions can decrease performance and increase errors and response time. Additionally, there is evidence that providing a lag time prior to an interruption can mitigate some of the interruption costs. The goal of this paper is to investigate the effects of interruptions and interruption lags and explore possible strategies to attenuate interruption costs. A novel sequential decision-making paradigm was used, where the difficulty of the task and type of interruption were the two experimental manipulations. The results indicate that there is a potential benefit to including a lag time when presented with interruptions.

Keywords: interruption; interruption lag; decision making

Introduction

Interruptions are a common occurrence in daily life. From a telephone ringing in the middle of a conversation with a friend, to a nurse handing an X-ray to a surgeon in the midst of a procedure, interruptions can happen at any moment and in any situation. The interruption literature dates back to the early 1900's when Zeigarnik (1938) surprisingly found that interrupted tasks were better recalled compared to tasks that were uninterrupted. This is often referred to as the "Zeigarnik effect". However, research within other fields, such as aviation, suggests interruptions can have negative impacts on behavior. For example, Fitts and Jones (1947) explain, "forgetting may occur when something unusual happens to interrupt or momentarily distract the pilot from his normal routine." Although there has been conflicting results when trying to replicate the Zeigarnik effect and countless of studies on interruptions since the 1920s, Gillie and Broadbent (1989) argue it is even more important to research how easily can people resume a task after being interrupted and what makes interruptions disruptive?

To answer these questions, Gillie and Broadbent (1989) had participants complete a complex computer-based adventure game and manipulated the types and duration of interruptions within the task. They found that similarity to the primary task and the complexity of the task lead to disruptive interruptions, but not the length of an interruption or when it occurred (Gillie & Broadbent, 1989). However, it is worth noting that there were only 10 participants in the experiment and this study was completed 30 years ago. In a more recent review of interruptions, Borst, Taatgen, and van Rijn (2015) conclude that there are three main disruptive factors: duration of the interruption, complexity of the interruption, and the moment of the interruption. Research on the effects of interruptions has dramatically increased in recent years, especially in fields where interruptions can lead to serious and sometimes even fatal consequences, such as in medicine (Westbrook, Raban, Walter, & Douglas, 2018; Walter, Li, Dunsmuir, & Westbrook, 2014; Westbrook et al., 2010), aviation (Gontar, Schneider, Schmidt-Moll, Bollin, & Bengler, 2017), and driving (Klauer et al., 2014; Young, Salmon, & Cornelissen, 2013) just to name a few.

Here, we will define interruptions as a break from one task in order to complete another task, and in our experiment, resuming the primary task can only occur once the secondary task is completed. Within the literature of interruption lags, studies have often used paradigms that are inherently complex and only include one interruption (Gillie & Broadbent, 1989; Trafton, Altmann, Brock, & Mintz, 2003; Cane, Cauchard, & Weger, 2012). Therefore, the main aim of the current experiment is to explore strategies to minimize interruption costs in a decision-making task with varying levels of difficulty so that we can easily manipulate the frequency, type, and location of interruptions. This is a novel sequential decision-making task that will be referred to as "The Mazing Race", which will be explained in greater depth later.

Theoretical Framework

Theories for understanding human cognition have been around for decades. Adaptive Control of Thought-Rational (ACT-R) is one cognitive architecture to model human memory that has been gradually developing for years (J. Anderson, Lebiere, Lovett, & Reder, 1998). Derived from ACT-R, the Altmann and Trafton's Goal Activation Model (GAM) theorizes whichever goal is most *active* will govern behavior. This contrasts to the basic "last-in, first-out" structure to model goal behavior, which assumes the *newest* goal directs behavior. Although this specific model will not be implemented in this study, the model is important to understand as it motivates the research question and design.

GAM predicts that people can take time to prepare before goals are suspended or interrupted. Therefore, the model suggests it may be important to give a cue before an interruption. Specifically, the GAM "predicts that interruption lag is critical to the ability to resume an interrupted goal" (Altmann

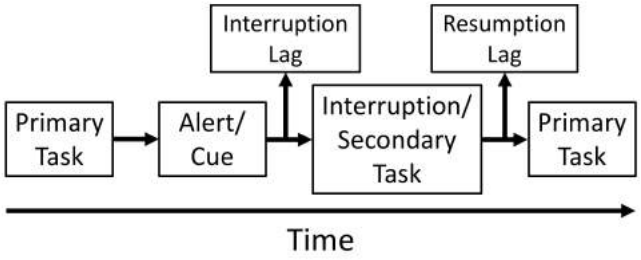


Figure 1: Image modified from Trafton et al. (2003) for a visual representation of the model including interruption and resumption lags.

& Trafton, 2002). Figure 1 illustrates a model to visualize what is happening during an interruption as a function of time. The overall idea of this model is that a person in the middle of completing a primary task is suddenly interrupted with another task, and ultimately has to resume the primary task. Often times there may be an alert, or cue, before the interruption occurs. The interruption lag is the time between the alert and the onset of the interruption. Depending on the context, the duration of the interruption lag may be able to be manipulated. Finally, the resumption lag is the time it takes to resume the primary task once the interruption has ended. This is often the dependent variable in experimental studies investigating interruption lags. One prediction from this model is that the interruption lag gives one time to prepare to resume the primary task after being interrupted.

To further understand this model, we will elaborate on a real-world example alluded to in the introduction. Imagine two people are in the midst of a conversation and suddenly the phone rings. Before the individual goes to answer it, she has the option to quickly end the current conversation, ignore the incoming call, or temporarily pause the conversation. In this scenario, the phone ringing is the alert and choosing to answer the phone would be the interruption to the primary task of the current conversation. If she chooses to pause the conversation, it would be advantageous to take a couple of seconds to remember exactly where the conversation has left off in order to successfully resume the conversation after the call. This is the idea of the interruption lag.

Interruption Lags

Over the past couple of decades, there have been several studies focusing on the effects of interruption lags. However, there are conflicting results with regards to the benefits of interruption lags. On one hand, problem solving tasks (e.g. Tower of London) showed interruption lags lead to faster resumption times compared to no lags (Morgan, Patrick, & Tiley, 2013; Hodgetts & Jones, 2006b, 2006a; Trafton et al., 2003). In fact, Hodgetts and Jones (2006a) found that even a two-second interruption lag can aid resumption on the primary task. Although most research on interruption lags has focused on static contexts, Labonté and colleagues show that a pre-interruption warning can be beneficial in dynamic environments, as well (Labonté, Tremblay, & Vachon, 2019, 2016). On the other hand, there were no benefits to including interruption lags within a reading task (Cane et al., 2012). The authors suggest that the lack of an effect is possibly because interruption lag effects may be dependent on the specific task (e.g. reading task vs. problem solving task).

It is also important to note the complexity of these tasks. For instance, Trafton et al. (2003)'s primary task was a computer game where participants had to keep track of a number of different resources including munitions, fuel, fuel tanks, vehicles, and more. Even the interruption was an involved tactical assessment task lasting 30 seconds. Similarly, the interruption in the reading task was a full minute long. The studies mentioned here investigated the effects of interruption lags in complex primary and secondary tasks. This current study looks to extend the literature by asking what effect, if any, will interruption lags have on a "simpler" task? The "simpler" task will be a novel sequential decision-making task. It is simpler in the sense that participants had to make very quick decisions and the interruptions were relatively short, as well. This paradigm is also novel because the number of interruptions was manipulated, rather than just having one interruption throughout the entire duration of the task. This is arguably a better model of the real world as interruptions are often frequent, unavoidable, and unpredictable.

Method

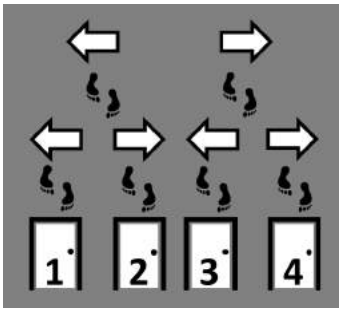
Participants

A total of 64 undergraduate students from the University of New South Wales were recruited to complete the experiment for course credit. Five participants' data were removed from analysis because the programmed crashed, so they were unable to complete the study, leaving 59 participants left for analysis.

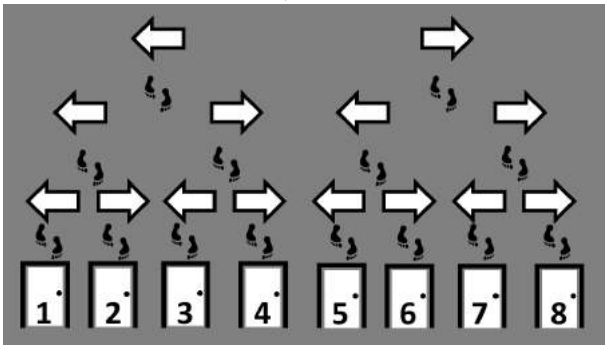
Design

This study was a 3 (difficulty: easy, medium, hard) x 3 (type of interruption: no interruption, interruption, and interruption + lag) fully within-subject design. Participants completed every combination of the conditions once for a total of nine blocks. The Mazing Race was the primary task and the interrupting task was a short-term recognition memory task. The number of interruptions depended on the difficulty level of the block. We were concerned about the difficulty of the task, and so we ensured participants completed the blocks in order of difficulty, from easiest to hardest. Within a set of problems with the same difficulty level, the type of interruption was randomized.

Primary Task: The Mazing Race In The Mazing Race participants had to make a series of decisions to go either "left" or "right" to work their way through a maze to open up doors. Figure 2 shows a visual representation of the underlying structure of the maze. These images were the stimuli used in the experiment and examples of what participants

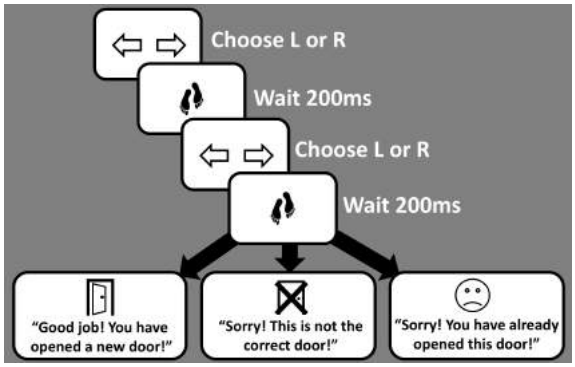


(a) Easy condition

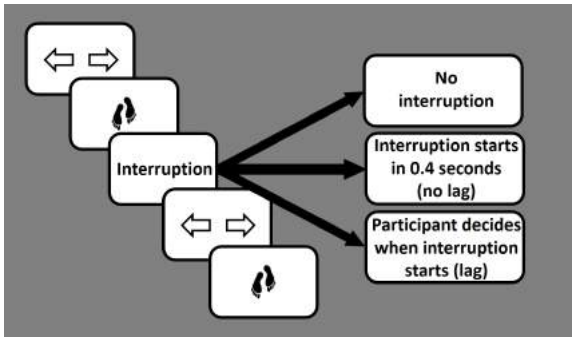


(b) Medium condition

Figure 2: The underlying structure of The Mazing Race for the (a) easy condition and (b) medium condition. The design required participants to open doors in the given order.



(a) Trial outcomes



(b) Interruption types

Figure 3: Schematic representation of the experimental design in the easy condition, depicting the (a) three possible trial outcomes and the (b) interruption types.

saw before each block. This figure illustrates a maze in the easy condition (a) with a total of four doors and a maze in the medium condition (b) with a total of eight doors. Although it is not displayed, the hard condition was of a similar structure, but had one additional decision-making level, resulting in a total of 16 doors. We named it The Mazing Race as it is a race to get to the bottom of every unique path in order to open all of the doors in as few attempts as possible. Once a door was opened, it stayed opened for the remainder of the block. Thus, the main dependent variables were the number of doors successfully opened and the number of trials needed to complete each block. Response times were recorded for further analysis, specifically looking at the response time of every decision (i.e. from when a stimulus is presented until the participant makes a keyboard response).

After participants studied the underlying structure of the maze, they pressed the space bar to start the block. Then, as shown in Figure 3a, two arrows appeared on the screen: one pointing left (L) and one pointing right (R) and participants simply had to choose to go L or R with the respective arrow keys. After every decision, animated footprints appeared for a total of 200ms symbolizing the participant walking down to the next level of the maze, where they made their next decision to go L or R. In the easy condition, for example, after two sequential decisions they reached the bottom of the maze. Every difficulty level had a maximum number of attempts to open all of the doors. In the easy condition it was 8 attempts, medium had 16, and hard had 32. These maximum numbers were included to try to minimize participants' frustration while completing the task.

Participants were required to open the doors in a specific order as shown in Figure 2. The order was always the same: starting by opening the left-most door and systematically working their way to the right-most door. Therefore, on any given attempt to open a door, there was always only one correct response. Feedback was provided every time the participant reached a door (see Figure 3a). If they reached the correct door they received positive feedback saying, "Good job! You have opened the correct door!" If they reached a door that had not already been opened, but was the incorrect door, they received negative feedback saying, "Sorry! This is not the correct door!" Finally, if they opened a door that was already opened, they also received negative feedback saying, "Sorry! You have already opened this door!" To successfully complete an easy block, for example, participants needed to go down the following four paths in this sequence: LL, LR, RL, RR. The block ended when the participant either successfully opened all of the doors or exceeded the maximum number of attempts. The experiment ended when all nine blocks were completed.

Interrupting Task: Recognition Memory Test Past research has shown that similarity and complexity between the

primary and interrupting task are factors that determine if the interruption is disruptive (Borst et al., 2015; Gillie & Broadbent, 1989). Because the main interest was the potential benefits of interruption lags, it was necessary that the interruptions were disruptive. Therefore, a recognition memory task was chosen because we assumed that both the primary and secondary tasks relied on a similar subset of memory-related cognitive processes. Figure 3b illustrates the three types of interruptions: no interruption, interruption, and interruption + lag. Participants were explicitly told what type of interruption to expect before the start of each block.

In our memory task there was a study and a test phase. The stimuli included randomly selected words from a list of 1535 words, where all the words were between three and six letters and one-syllable. This is the same word pool as used in Donkin and Nosofsky (2012). Anticipating that some participants may strategically try to keep count of the number of doors they have opened, numbers (randomly generated between 0-999) were also included in the memory test as a way to interfere with any possible counting. In the test phase, participants were presented with one "old" (i.e. previously studied) item and one "new" (i.e. previously unstudied) item and they were instructed to select the word that they believed to be the old item. During every interruption, there was a total of 10 study items and 10 test pairs. Each study item was randomly selected to be either a word or a number and test pairs could be two words, two numbers, or one of each.

The memory test was programmed to occur once in each set of four trials, where a trial is an attempt of opening a door, in The Mazing Race with a set number of interruptions in each condition. There was only one interruption in the easy condition, up to four in the medium condition, and up to eight in the hard condition. The interruptions were purposefully random and spread out to make it harder for the participant to anticipate when they would be interrupted. Before the memory task began, participants completing an interruption block saw a screen that said: "Start memory test **NOW**" (the task began automatically after 400ms) and in the interruption + lag block they saw a screen that said: "**Think about where you are in the Maze.** Press the space bar to start the memory test". The interruption lag was self-paced, meaning participants decided when to start the memory task. As soon the memory task was completed, participants immediately resumed The Mazing Race at the exact point where they left off and were given no environmental cues about where they were in the maze, which they were told from the start.

Furthermore, after every block, participants were given feedback on their performance for both tasks. For The Mazing Race, they were shown the number of doors they successfully opened and, if there were interruptions, they were shown the percentage of correct answers on the memory test. Lastly, participants were instructed that performance on The Mazing Race and the memory test were equally as important.

Results

We predicted that performance would be best in the no interruption condition and worst in the interruption condition. We expected the interruption + lag condition to fall somewhere between the others, as the lag would provide time to prepare to switch tasks and resume The Mazing Race. As this is a novel paradigm, several different analyses were carried out to try to fully understand the results. We will report the results of both frequentist and Bayesian repeated-measures ANOVAs. The Bayesian anaylses were performed using JASP (JASP Team, 2018), with priors set to their default values within the program. We report Bayes Factors (BF), which express the probability of the data given the alternative hypothesis (H_1) relative to the null hypothesis (H_0). A BF = 1-3 indicates weak evidence for the alternative hypothesis and a $BF > 30$ indicates strong evidence for the alternative hypothesis. Also note that for the purposes of this proceedings paper, due to the large number of comparisons, and exploratory nature of this investigation, we will only present the result of omnibus F-tests as a rough indicator of whether there were differences among conditions as a result of the introduction of interruptions. As such, we will focus on describing the qualitative pattern of the means and attempt to provide a more holistic interpretation of the overall pattern of results.

Before looking at specific dependent variables, Table 1 illustrates results from the interruption task. Participants performed equally well in both the interruption and interruption + lag conditions. Although performance decreased slightly as the primary task got harder, performance was still well above chance in all of the conditions. This suggests that participants were engaged in the secondary task and not using all of their cognitive resources on The Mazing Race.

To measure performance on the task, we first observed the average number of doors participants opened (Figure 4). The dotted lines represent the maximum number of doors in each level of difficulty: four doors in easy, eight doors in medium, and 16 doors in hard. Perfect performance would be to open all the doors in four, eight, and 16 trials, respectively. Looking at the Figure, it doesn't appear that the type of interruption affected the number of doors opened in the easy (BF_{10} = .68; $F(2,116) = 2.85$, $p = 0.06$) or hard (BF_{10} = .12; $F(2,116) = 0.87$, $p = 0.42$) conditions. There may have been an effect of interruptions in the medium condition, but the evidence is

Table 1: Summary Statistics of Interruption Task

	Easy	Medium	Hard
Interruption	0.82 (0.12)	0.79 (0.12)	0.77 (0.13)
Interruption + lag	0.82 (0.14)	0.80 (0.11)	0.77 (0.11)

Average probability of correct responses on the interruption task (memory test) across the different levels of difficulties. Standard deviations are provided in parentheses.

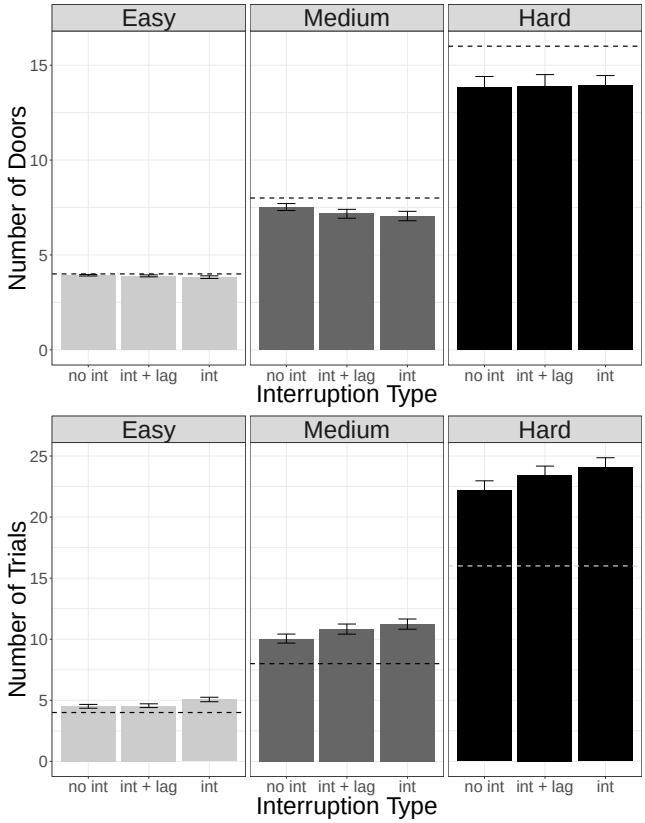


Figure 4: Average number of doors opened (top) and average number of trials needed to successfully complete the maze (bottom) in each condition as a function of interruption type. Dotted lines represent the maximum number of doors. Error bars indicate standard error in this and the subsequent figure.

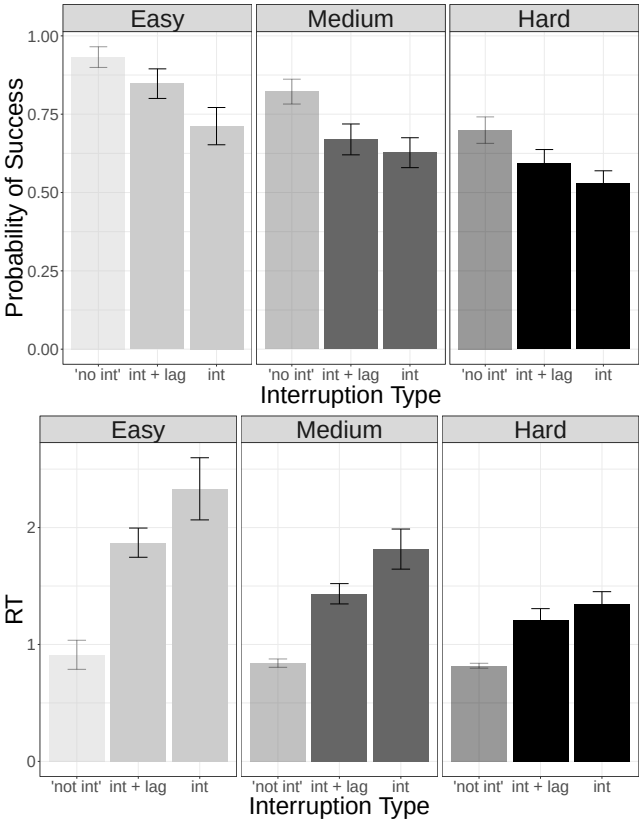


Figure 5: Probability of successfully opening the correct door (top) and median reaction times on trials immediately following interruptions. The "no interruption" condition is a lighter shade to show that it represents a baseline group for comparison even though there was no interruption in these conditions.

inconclusive ($BF_{10} = 1.427$; $F(2,116) = 3.78$, $p = 0.03$). However, it is likely that there were ceiling effects, especially in the easy condition, such that participants were opening all, or close to all, of the doors.

Even if participants successfully opened all of the doors, it is possible that they made more mistakes and needed more trials to open all doors when interrupted. Therefore, we next looked at the average number of trials needed to complete the block (Figure 4). The type of interruption did effect the number of trials in the easy condition ($BF_{10} = 38.22$; $F(2,116) = 7.79$, $p < 0.001$), medium condition ($BF_{10} = 5.93$; $F(2,116) = 5.47$, $p = 0.01$), though the statistical evidence was less clear for the hard condition ($BF_{10} = 1.82$; $F(2,116) = 4.10$, $p = 0.019$). Focusing on the mean scores in all difficulty conditions, we see that performance tends to decrease across interruption type with best performance in no interruption, followed by interruption + lag, with poorest performance in the interruption condition without lag.

Our next analyses examined the probability of successfully opening a door and the average median RT on trials immediately following an interruption (Figure 5). In order to have a baseline condition, we created a no interruption ("no int") condition for these two figures. As a reminder, there was one interruption in the easy condition, two in the medium, and four in the hard. That gives one, two, and four data points per participant in the respective conditions. Therefore, for the no interruption condition, we sampled one, two, and four data points for each respective condition from every participant to represent where an interruption may have occurred. It was predicted that the no interruption condition would have the highest probability of success and the fastest RT, the interruption condition would have the lowest and the slowest, and the interruption and lag would fall somewhere in the middle.

Turning first to the probability of success of opening a door immediately after an interruption, there was an effect of interruption type in all of the conditions: easy ($BF_{10} = 12.60$; $F(2,116) = 5.81$, $p < 0.01$), medium ($BF_{10} = 24.91$; $F(2,116) = 7.13$, $p = 0.001$), and hard ($BF_{10} = 15.11$; $F(2,116) = 6.57$, $p < 0.01$). Looking at the means, we can see there was the biggest difference between "no interruption" and interruption + lag, such that the introduction of the interruption had a relatively large effect on the next trial. In all conditions, however, we do still see a benefit of the lag, with worse performance in the interruption without a lag condition.

Next, we looked at median RTs following interruptions as a way to measure resumption lag. The ANOVA analyses reveal very large effects for easy ($BF_{10} > 100$; $F(2,116) = 22.50$, $p < 0.001$), medium ($BF_{10} > 100$; $F(2, 116) = 34.8$, $p < 0.001$), and hard ($BF_{10} > 100$; $F(2,116) = 28.59$, $p < 0.001$) conditions. Again, we saw a similar pattern when looking at the mean scores. The "no interruption" had the shortest RTs and then a big jump up to interruption + lag, and the interruption conditions had the longest RTs. We will turn to the discussion for further possible interpretations of these results.

Discussion

The aim of this study was to analyze the effects of interruptions and explore the possible benefits of interruption lags in a novel sequential decision making task. While performance was, by no surprise, the best in blocks without interruptions, we did find benefits to having lag time when there was an interruption.

Using the number of doors opened as a dependent variable did not show any large effects of interruption type. Additionally, we did not see the quantitative pattern of data like we saw in the other analyses. However, 49 participants successfully opened every door in all three easy blocks, suggesting performance was at ceiling. This has real world implications, such that if one gets interrupted in the middle of an effortless task, the interruption may not disrupt the primary task at all. Performance, however, did begin to decline when the primary task got harder. In the medium and hard conditions, 39 participants successfully opened all of the doors, with a handful of participants opening less than 50% of the doors.

When looking at the maximum number of trials needed to open all the doors, we did begin to see effects of interruption type. This was the first analysis where a consistent pattern of data emerged. Participants were able to complete the task in the lowest number of trials when there were no interruptions. Performance appeared to decrease in the interruption + lag condition and even more so in the interruption condition. This makes sense as participants were explicitly told to use the interruption time to try to remember their place in the maze.

When interrupted, it often takes time to pick up where you left off. For this reason, we were interested in observing the trials that occurred immediately following interruptions, specifically looking at the probability of success and response time when making the subsequent decision. The probability of success was highest and the average RT was the shortest when examining data from the no interruption condition because participants had nothing from which to be distracted. Additionally, we see a similar trend in the data as previously mentioned, where the interruption lag appears to be improving performance (compared to the interruption condition) in both of these analyses.

Limitations

Observing the effects of interruptions is difficult because it is unreasonable and unrealistic to interrupt participants on every single trial. For that reason, we decided to only include interruptions on $\frac{1}{4}$ of the trials. Therefore, we were left with limited data points for each participant. One solution would be to increase the number of interruptions, but that may be too cumbersome and frustrating for participants. Another solution would be to increase either the number of participants or number of trials per participant. Additionally, participants were required to open the doors in the same order (i.e. left to right) in all the blocks and always completed the blocks in order of difficulty (i.e. easy to hard). Therefore, although the overall RTs are longest in the easy condition and shortest in the hard conditions, this is likely due to practice effects. By the time participants get to the more difficult conditions, they can begin to anticipate their next move resulting in quicker decisions. However, models of volitional action control (Heise, Gerjets, & Westermann, 1997) predict that difficult tasks will protect against distractions. For example, Scheiter, Gerjets, and Heise (2014) and Wirzberger, Bijarsari, and Rey (2017) found that irrelevant interruptions only impaired performance in the easy, and not difficult, conditions of their respective experiments. The competing theories of whether practice effects or volitional control are driving the RT effects can be tested in follow up studies by randomizing the difficulty order of the blocks.

Future Directions

Possible avenues for future research would be to make The Mazing Race more challenging, for example, by randomizing the order of doors to open. Another interesting question is would we see the same pattern of results if the interruption task was different? For example, on one hand, the interruption could be as simple as pushing the space bar every time a cue appears. On the other hand, it is possible that a spatial recognition memory task may be even more disruptive. It is necessary to implement different types of interruptions to see if these results generalize. The relative simplicity and flexibility of The Mazing Race makes it possible to address these questions in follow up studies.

Conclusions

Taken together these findings illustrate the potential benefit to including a lag time when presented with an interruption. Performance increased from interruption < interruption + lag < no interruption across levels of difficulty and across multiple analyses, suggesting there is evidence from this study that interruption lags can reduce some interruption costs. Furthermore, this complements previous research on interruption lags in problem solving tasks (Hodgetts & Jones, 2006a; Trafton et al., 2003). Follow up studies should aim to include more interruptions (if possible) to provide more data points. Additionally, modeling these results could prove invaluable in trying to understand and predict participants' performance and the types of mistakes they make. Interruptions will always be part of our daily lives, so it is not only important to study the effects and costs of interruptions, but also to study possible strategies to minimize those costs.

References

Altmann, E. M., & Trafton, J. G. (2002). Memory for goals: an activation-based model [Journal Article]. *Cognitive Science*, *26*(1), 39-83.

Anderson, J., & Douglass, S. (2001). Tower of hanoi: Evidence for the cost of goal retrieval [Journal Article]. *J. Exp. Psychol.-Learn. Mem. Cogn.*, *27*(6), 1331-1346.

Anderson, J., Lebiere, C., Lovett, M., & Reder, L. (1998). Act-r: A higher-level account of processing capacity [Journal Article]. *Behav. Brain Sci.*, *21*(6), 831-+. doi: 10.1017/S0140525X98221765

Anderson, J. R., & Lebiere, C. (2014). *The atomic components of thought* [Book]. Hoboken: Hoboken : Taylor and Francis.

Borst, J. P., Taatgen, N. A., & van Rijn, H. (2015). What makes interruptions disruptive?: A process-model account of the effects of the problem state bottleneck on task interruption and resumption. In *Proceedings of the 33rd annual acm conference on human factors in computing systems* (pp. 2971–2980).

Cane, J. E., Cauchard, F., & Weger, U. W. (2012). The time-course of recovery from interruption during reading: Eye movement evidence for the role of interruption lag and spatial memory [Journal Article]. *The Quarterly Journal of Experimental Psychology*, *65*(7), 1397-1413. doi: 10.1080/17470218.2012.656666

Donkin, C., & Nosofsky, R. M. (2012). A power-law model of psychological memory strength in short-and long-term recognition. *Psychological Science*, *23*(6), 625–634.

Fitts, P., & Jones, R. (1947). Analysis of factors contributing to 460 "pilot error" experiences in operating aircraft controls 1947. *Report TSEAA*, 694–12.

Gillie, T., & Broadbent, D. (1989). What makes interruptions disruptive? a study of length, similarity, and complexity [Journal Article]. *An International Journal of Perception, Attention, Memory and Action*, *50*(4), 243-250. doi: 10.1007/BF00309260

Gontar, P., Schneider, S. A. E., Schmidt-Moll, C., Bollin, C., & Bengler, K. (2017). Hate to interrupt you, but… analyzing turn-arounds from a cockpit perspective. *Cognition, Technology & Work*, *19*(4), 837–853.

Heise, E., Gerjets, P., & Westermann, R. (1997). The influence of a waiting intention on action performance: Efficiency impairment and volitional protection in tasks of varying difficulty. *Acta Psychologica*, *97*(2), 167–182.

Hodgetts, H. M., & Jones, D. M. (2006a). Contextual cues aid recovery from interruption: The role of associative activation [Journal Article]. *J. Exp. Psychol.-Learn. Mem. Cogn.*, *32*(5), 1120-1132. doi: 10.1037/0278-7393.32.5.1120

Hodgetts, H. M., & Jones, D. M. (2006b). Interruption of the tower of london task: Support for a goal-activation approach [Journal Article]. *J. Exp. Psychol.-Gen.*, *135*(1), 103-115. doi: 10.1037/0096-3445.135.1.103

JASP Team. (2018). *JASP (Version 0.9)[Computer software]*. Retrieved from `https://jasp-stats.org/`

Klauer, S. G., Guo, F., Simons-Morton, B. G., Ouimet, M. C., Lee, S. E., & Dingus, T. A. (2014). Distracted driving and risk of road crashes among novice and experienced drivers. *New England journal of medicine*, *370*(1), 54–59.

Labonté, K., Tremblay, S., & Vachon, F. (2016). Effects of a warning on interruption recovery in dynamic settings. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 60, pp. 1304–1308).

Labonté, K., Tremblay, S., & Vachon, F. (2019). Forewarning interruptions in dynamic settings: Can prevention bolster recovery? *Journal of experimental psychology. Applied*.

Morgan, P. L., Patrick, J., & Tiley, L. (2013). Improving the effectiveness of an interruption lag by inducing a memory-based strategy [Journal Article]. *Acta Psychologica*, *142*(1), 87-95. doi: 10.1016/j.actpsy.2012.09.003

Scheiter, K., Gerjets, P., & Heise, E. (2014). Distraction during learning with hypermedia: difficult tasks help to keep task goals on track. *Frontiers in psychology*, *5*, 268.

Trafton, J., Altmann, E., Brock, D., & Mintz, F. (2003). Preparing to resume an interrupted task: effects of prospective goal encoding and retrospective rehearsal [Journal Article]. *Int. J. Hum.-Comput. Stud.*, *58*(5), 583-603. doi: 10.1016/S1071-5819(03)00023-5

Walter, S. R., Li, L., Dunsmuir, W. T., & Westbrook, J. I. (2014). Managing competing demands through task-switching and multitasking: a multi-setting observational study of 200 clinicians over 1000 hours. *BMJ Qual Saf*, *23*(3), 231–241.

Westbrook, J. I., Coiera, E., Dunsmuir, W. T. M., Brown, B. M., Kelk, N., Paoloni, R., & Tran, C. (2010). The impact of interruptions on clinical task completion [Journal Article]. *Quality & safety in health care*, *19*(4), 284. doi: 10.1136/qshc.2009.039255

Westbrook, J. I., Raban, M. Z., Walter, S. R., & Douglas, H. (2018). Task errors by emergency physicians are associated with interruptions, multitasking, fatigue and working memory capacity: a prospective, direct observation study. *BMJ Qual Saf*, *27*(8), 655–663.

Wirzberger, M., Bijarsari, S. E., & Rey, G. D. (2017). Embedded interruptions and task complexity influence schema-related cognitive load progression in an abstract learning task. *Acta psychologica*, *179*, 30–41.

Young, K. L., Salmon, P. M., & Cornelissen, M. (2013). Distraction-induced driving error: An on-road examination of the errors made by distracted and undistracted drivers. *Accident Analysis & Prevention*, *58*, 218–225.

Zeigarnik, B. (1938). On finished and unfinished tasks. *A source book of Gestalt psychology*, *1*, 300–314.

Asymmetric Switch Costs as a Function of Task Strength

Markus Spitzer[1,2,*], **Sebastian Musslick**[2,*], **Michael Shvartsman**[2], **Amitai Shenhav**[3] **and Jonathan D. Cohen**[2]

[1]Albert Ludwig University of Freiburg, Freiburg, 79085, Germany.
[2]Princeton Neuroscience Institute, Princeton University, Princeton, NJ 08544, USA.
[3]Department of Cognitive, Linguistic, and Psychological Sciences,
Brown Institute for Brain Science, Brown University, Providence, RI 02912, USA.

[*]Equal Contribution, Corresponding Author: *markus.spitzer@psychologie.uni-freiburg.de*

Abstract

Several studies reported that it is harder to switch from a difficult task to an easy task than vice versa. Previous studies explain this paradoxical effect in terms of differences in task strength, by letting participants switch between different types of tasks. However, these studies failed to isolate the effects of task strength from task identity. Here, we present a series of experiments in which we systematically varied the strength of two tasks independent of their identity. We adapted a computational model of task switching by Yeung and Monsell (2003) to derive predictions about the magnitude of asymmetric switch costs (ASC) as a function of task strength, and compared predictions from the model to behavioral data. Our results reveal that ASC depend on the overall and relative task strength across the two tasks. ASC can therefore flip directions if the strength of two tasks is reversed, irrespective of their identities.

Keywords: task switching; paradoxical switch cost; task-set inertia

Introduction

Humans are remarkably flexible in their ability to switch between different tasks. However, a paradoxical finding in task switching experiments is that participants require more time and exhibit more errors when they switch from a harder task to an easier task than vice versa (Monsell, Yeung, & Azuma, 2000; Yeung & Monsell, 2003). For instance, bilinguials take more time to switch from their first language to their second language compared to switching from their second language to their first language (Meuter & Allport, 1999).

Alport, Styles, and Hsieh (1994) explained such asymmetric switch costs (ASC) in terms of the task-set inertia hypothesis. This postulates that the processes needed to execute a task (the task-set) persist in time, causing interference with the next task, and that switch costs reflect the time needed to resolve this interference. Executing a weak[1] task is assumed to require inhibition of automatic processes from a competing dominant task that would otherwise interfere (e.g. speaking a second language would require inhibition of the first language). According to the task-set inertia hypothesis, this inhibition persists when switching back to a dominant task, yielding high switch costs (Allport & Wylie, 2000). In contrast, switching to a weaker task should result in lower switch costs since the weak task would not require to be inhibited when performing the dominant task.

Building on the task-set inertia hypothesis, Yeung and Monsell (2003) devised a formal model that explains ASC as an interaction between task priming and top-down control. In their model, task priming corresponds to a carry over of the previous task-set, resulting in a facilitation of task repetitions (positive priming) but a delay for task switches (negative priming). Top-down control is assumed to vary as a function of task strength, with the weaker task requiring and receiving more control than the more dominant task. Without top-down control, both tasks would be subject to the same switch cost as they would be governed by the same amount of negative task priming. However, higher amounts of top-down control for the weaker task can compensate the effects of negative task priming, yielding lower switch costs for the weaker task relative to the more dominant task.

These and other accounts identify differences in task strength as a necessary condition for ASC (Alport et al., 1994; Allport & Wylie, 2000; Yeung & Monsell, 2003; Gilbert & Shallice, 2002). These accounts predict that the asymmetry in switch costs between two tasks should reverse if their task strengths reverse. In the example above, switching from a second language to a first language should be easier if the task strength of the first language was decreased relative to the second language. However, to date, there is no empirical support for this prediction as previous studies confounded task strength with task identity (e.g. the first language is always easier than the second language, at least for the duration of the experiment). The inability to manipulate task strength independent of task identity has also prevented researchers from testing the precise constellations of task strength under which ASC arise. One may ask if ASC would arise as soon as two tasks differ significantly in task strength, even if both are considered weak, or dominant? Finally, several studies have failed to observe ASC in error rates (ERs) (Meuter & Allport, 1999; Costa & Santesteban, 2004), or reported effects for reaction times (RTs) only (Mayr & Keele, 2000; Philipp, Gade, & Koch, 2007), failing to address whether participants traded off speed against accuracy.

So far, it is unclear (a) whether an asymmetry in switch costs between two tasks reverses if the task strength of the two tasks is reversed, and (b) whether the magnitude of ASC depends on the strength of the dominant task, in addition to

[1]Here, we refer to a task as weak if it requires higher amounts of cognitive control in order to overcome processing interference from more automatic (dominant) tasks.

the difference in strength between the dominant and the weak task. Here, we examine these questions across five experiments in which we manipulate the strength of two tasks independent of their identity. To account for tradeoffs in speed versus accuracy, we fit a hierarchical drift diffusion model (DDM, Ratcliff, 1978; Wiecki, Sofer, & Frank, 2013) to RTs and error rates. Finally, we compare experiment results to predictions derived from the task switching model by Yeung and Monsell (2003)

Experiments

We examined ASC across five experiments in which participants switched between categorizing the motion and categorizing the color of random-dot kinematograms (RDKs) (Kayser, Erickson, Buchsbaum, & D'Esposito, 2010). We manipulated the strength of each task across experiments by varying the signal to noise ratio of the task-relevant stimulus dimension. For each experiment, we then determined the strength of each task, as well as ASC in the drift rate of the fitted DDM.

Participants

All participants were students from Princeton University and received one hour of course credit. The study was approved by the Institutional Review Board of Princeton University. Participants signed a consent form prior to participation and were debriefed about the purpose of the study at the end of testing. We excluded participants whose performance was below 60% accuracy. Table 1 lists participant information for each experiment.

Table 1: Participants across all experiments.

Exp.	Participants	Age	Excluded
1	76 (37 female)	M = 19.5, SD = 0.56	6
2	25 (13 female)	M = 20.5, SD = 0.96	4
3	76 (42 female)	M = 20.4, SD = 0.54	6
4	33 (18 female)	M = 19.8, SD = 0.71	3
5	33 (17 female)	M = 20.1, SD = 0.63	4

Method

Each stimulus was an RDK that consisted of blue and red moving dots. Some of the dots consistently moved in either an upward or downward direction (independent of their color) while the remaining dots moved in a random direction. Participants switched between a color task, in which they had to indicate the color of the majority of the dots (red or blue), using the response buttons 'K' and 'L' respectively, and a motion task in which they had to indicate the direction of coherent motion (up or down), also using the response buttons 'K' and 'L', respectively. Participants performed each task over a mini-block of four to six trials.

Only the first trial of a mini-block was of interest to our analysis. Thus, in each sequence, we counterbalanced seven factors with respect to the first trial of each mini-block: task (color or motion task), task transition (task switch or task repetition), dot motion (upward or downward), dot color (mostly blue or red) and correct response ('K' or 'L' key). Participants were exposed to a total of 256 mini-blocks, divided into four larger experiment blocks.

Each mini-block was preceded by a task cue that instructed participants which task to perform. In some mini-blocks, participants had to repeat the task that they performed in the previous mini-block (task repetition), whereas in other mini-blocks they had to switch to the other task (task switch). The cue was displayed for 700ms before it disappeared for another 500ms. On each trial of a mini-block, the RDK stimulus was shown for 2000ms, followed by an inter-trial interval of 700ms. Participants were asked to respond while the stimulus was on the screen.

Critically, we varied the difficulty for both tasks across experiments, by changing the signal to noise ratio (coherence) for each task (see Tables 2 & 3). Note that variations in the signal to noise ratio of task-relevant stimulus dimensions can mimic the effect of traditional notions of task strength, such as stimulus-response associations in models of cognitive control (Cohen, Dunbar, & McClelland, 1990; Botvinick, Braver, Barch, Carter, & Cohen, 2001) and task switching (Gilbert & Shallice, 2002; Yeung, Nystrom, Aronson, & Cohen, 2006). Thus, differences in the signal to noise ratio between tasks resemble differences in task strength. We defined the color coherence as the percentage of dots that were displayed in the dominant color. For instance, a color coherence of 60% indicated that 60% percent of the dots were colored in blue while the rest of the dots were colored in red. Similarly, we defined motion coherence as the percentage of dots that moved consistently in one direction as opposed to moving in a random direction. In both tasks, coherence was used as a proxy for task difficulty: the higher the coherence, the easier it was to perform the task (Kayser, Buchsbaum, Erickson, & D'Esposito, 2009). It is important to note that equal values for color coherence and motion coherence do not necessarily yield the same level of performance for both tasks.

We adjusted the coherences of both tasks based on results from prior experiments (Tables 2 & 3). The coherence setting for Experiment 1 was determined based on prior pilot studies, with the intention to make the motion task easier than the color task. As expected, we observed that it was easier for participants to perform the motion task relative to the color task in Experiment 1. To test whether this relationship can be inverted, we lowered the coherence of the motion task and increased the coherence of the color task in Experiment 2. In Experiment 3, we tested whether we can invert the relationship observed in Experiment 1 by just lowering the coherence of the motion task while keeping the coherence of the color task the same as in Experiment 1. This manipulation did not yield ASC in terms of RTs, possibly due to a small difference in task strength. We therefore conducted a fourth Experiment in which we decreased the coherence of the motion task even further (relative to Experiment 3) while keeping the coherence of the color task the same as in Experiment 1. Despite significant differences in task strength in terms of both RTs

and error rates, we still failed to observe ASC in Experiment 4, suggesting that ASC may also depend on the strength of the dominant task (in this case, the color task). We therefore decided to increase the coherence of the color task in Experiment 5 while setting the coherence of the motion task to the same value as in Experiment 1.

Data Analysis

We were specifically interested in the performance costs associated with task switches and therefore focused our analyses on the RTs and error rates associated with the first trial of a miniblock (Rogers & Monsell, 1995). We assessed the effects of task (indexing relative task strength), as well as the interactive effect of task and task transition (indexing ASC) on RTs and error rates using a linear mixed model and logistic mixed model, respectively. We then fit a DDM to RTs and error rates, using the HDDM package (Wiecki et al., 2013). The DDM simulates performance on a task as an accumulation process that integrates information about the stimulus until one of two response thresholds is reached. The rate of evidence accumulation, henceforth referred to as drift rate, can be taken as a proxy for task strength, whereas the threshold indicates the degree to which speed is traded against accuracy (Ratcliff, 1978; Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006; Simen et al., 2009).

Fitting performance with the DDM serves two purposes. First, we quantify behavioral differences in task strength, as well as the asymmetry in switch costs in terms of drift rate, thereby isolating tradeoffs in speed versus accuracy. Second, we can compare performance of human participants to performance of the model. In the behavioral experiment, we manipulated the strength of each task in terms of the signal to noise ratio of the stimulus whereas in the computational model described below, we manipulate the strength of a task in terms of how much a corresponding task processing unit is driven by the stimulus. The DDM parameters collapse across RTs and error rates. Thus, quantifying task performance in terms of drift rate allows us to compare the strength of two tasks in comparable terms, and ASC in terms of the interactive effect between task and task transition on drift.

$$Drift \sim TaskTransition * Task + (1|Subj_i). \quad (1)$$

We fit the DDM using a Monte Carlo Markov Chain of 1000 samples of which the first 300 samples were not considered (burned). Other parameters of the DDM (response threshold, starting point, non-decision time and noise) were fit independent of condition.

Results

Table 2 and Table 3 show effects for RTs and error rates, respectively. In Experiment 1, participants were faster and made fewer errors during the motion task. The interaction between task and task transition was significant, with the dominant motion task yielding higher switch costs in RTs and error rates. The results from Experiment 1 were reversed in Experiment 2, with the motion task exhibiting higher RTs and higher error rates, and a significant interaction between task and task transition for error rates, but not RTs. In Experiment 3, participants responded slower to the motion task, but there was no significant difference in error rates between tasks. Moreover, the interaction between task and task transition was only significant for RTs, but not error rates. RTs and error rates were higher in the motion task in Experiment 4, however, we did not observe a significant interaction between task and task transition for RTs and error rates. Finally, in Experiment 5, we observed a speed accuracy tradeoff for the main effect of task, with the motion task showing higher RTs but lower error rates. A significant interaction between task and trask transition suggests that participants exhibited higher switch costs in terms of both RTs and error rates for the motion task relative to the color task.

Table 2: RT results for main effects of task and ASC.

Exp.	Color Coh.	Motion Coh.	Fixed Effects	β	SD	p
1	65%	60%	task***	-51ms	4ms	<0.001
2	80%	30%	task***	132ms	7ms	<0.001
3	65%	30%	task***	37ms	5ms	<0.001
4	65%	24%	task***	57ms	9ms	<0.001
5	80%	60%	task*	16ms	7ms	0.0321
1	65%	60%	ASC***	42ms	9ms	<0.001
2	80%	30%	ASC	21ms	14ms	0.12
3	65%	30%	ASC***	35ms	10ms	<0.001
4	65%	24%	ASC	32ms	18ms	0.0758
5	80%	60%	ASC***	48ms	15ms	0.001

Table 3: Error rate results for main effects of task and ASC.

Exp.	Color Coh.	Motion Coh.	Fixed Effects	β	SD	p
1	65%	60%	task***	-0.37	0.06	<0.001
2	80%	30%	task***	0.41	0.10	<0.001
3	65%	30%	task	0.02	0.05	0.704
4	65%	24%	task***	0.34	0.09	<0.001
5	80%	60%	task***	-0.36	0.11	<0.001
1	65%	60%	ASC***	0.41	0.12	<0.001
2	80%	30%	ASC*	-0.39	0.19	0.045
3	65%	30%	ASC	-0.082	0.11	0.473
4	65%	24%	ASC	0.10	0.19	0.591
5	80%	60%	ASC*	0.52	0.22	0.018

In addition to the RT and error rate analysis, we fitted the data with a hierarchical drift diffusion model (HDDM), to investigate the effects of task strength and ASC in terms of drift rate. In Experiment 1, the drift rate fitted to the motion task was significantly larger than the drift rate for the color task, (M = 0.34, 95% CI = [0.30, 0.40]), suggesting that the motion task was easier than the color task. Furthermore, the drift rate cost of switching to the easier motion task was higher than the drift rate cost of switching to the color task (M = -0.33, 95% CI = [-0.42, -0.24]). We examined the hypothesis that ASC reverse with the relative strength of two tasks, by comparing ASC of Experiment 1 against ASC in Experiment 2 (with lower coherence of the motion task and higher coherence of the color task). Results indicate that the task effect on drift rate reversed in Experiment 2 (M = -0.53, 95%

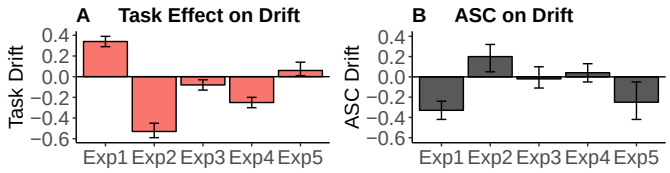


Figure 1: Main effects of task and ASC in terms of drift rate. (A) The main effect of task on drift rate is plotted for each experiment. Positive values indicate that participants accumulated evidence at a higher rate for the motion task relative to the color task (i.e. the strength of the motion task was higher). (B) The ASC effect in terms of drift rates is plotted for each experiment. Positive ASC drift rates indicate that switch costs for the color task were lower than the motion task. Vertical bars indicate 95% confidence intervals.

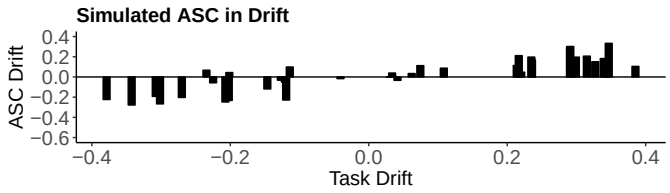


Figure 2: ASC as a function of task effect in drift rate simulated by the model.

CI = [-0.60, -0.46]), suggesting that the color task was easier than the motion task. Moreover, the cost of switching to the motion task was now lower than the cost of switching to the color task, as indicated by a positive interactive effect of task and task transition on drift rate (M = 0.20, 95% *CI* = [0.10, 0.36]). We also compared ASC across Experiments 3-5 to examine whether the magnitude of this effect depends on the task strength of the dominant task. In Experiment 3, we observed no effect of ASC in terms of drift rate (M = -0.02, 95% *CI* = [-0.11, 0.10]), despite significant differences in the task strength of each task (M = -0.08, 95% *CI* = [-0.13, -0.13]). Similarly, in Experiment 4, we observed no drift rate effect of ASC (M = -0.041, 95% *CI* = [-0.05, 0.13]), despite high differences in drift rate between tasks (M = -0.25, 95% *CI* = [-0.30, -0.20]). However, in Experiment 5 where the coherence of both tasks was high, we observed observed ASC (M = -0.25, 95% *CI* = [-0.42, -0.05]) while the difference in drift rate between tasks was relatively small (M = 0.06, 95% *CI* = [0.01, 0.14]).

One concern is that observed differences in drift rates between experiments may arise due to differences in sample size. We therefore performed the same analysis for each experiment on a random sample of 25 participants. Results of this analysis yield the same qualitative effects in drift rates.

Task Switching Model

To test whether our experimental results would be predicted by the model of Yeung and Monsell (2003), we simulated

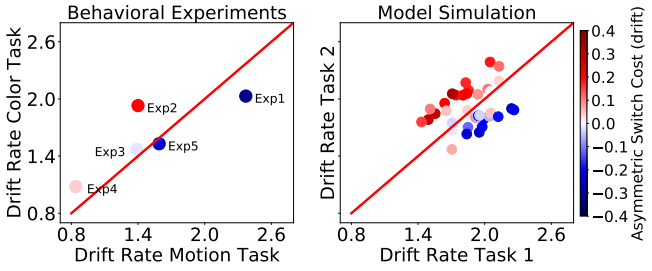


Figure 3: ASC as a function of task drift in the simulated model (left) and 5 behavioral experiments (right). Each colored dot represents a behavioral experiment (left) or a simulation (right) with a different configuration of input strength. Here, we quantified the strength of a task in terms of the drift rate fitted to RTs and error rates for that task. Positive ASC values (red) indicate that switch costs were higher for task 1 relative to task 2, whereas negative values (blue) indicate the opposite. The red line indicates equal strength for both tasks.

ASC as a function of task strength. The model explains the costs of switching between two tasks as a function of their activation. The activation of a task is determined by its strength, the amount of control allocated to a task, as well as task priming. RTs and error rates are generated by two separate equations for response generation and response resolution.

Model Mechanisms

Here, we outline the mechanisms of the model using the notation of Yeung and Monsell (2003). On each trial, the net input for a given task i is determined by a linear combination of four sources of input

$$input_i = strength_i + priming_i + control_i + noise \quad (2)$$

Where $strength_i$ corresponds to the strength of the task[2], $priming_i$ corresponds to inertia from the previously executed task and is set to a constant, $control_i$ is the amount of cognitive control allocated to the task, and *noise* is sampled from a Gaussian distribution with zero mean. The activation of each task is a negatively accelerated function of its net input

$$activation_i = 1 - e^{(-c * input[i])} \quad (3)$$

where c is a scalar that regulates the strength of the net input. Response generation time is computed by first normalizing activation of the two tasks

$$generationrate_i = activation_i / \sum activation \quad (4)$$

and then dividing a threshold by the normalized activation

$$generationtime_i = THRESHOLD / generationrate_i \quad (5)$$

where THRESHOLD is set to 100 in the model. The difference between the generation times for each task determine

[2]Here, the strength of a task is equivalent to the strength of processing weight in a neural network model multiplied by the input signal provided by the stimulus (Cohen et al., 1990; Gilbert & Shallice, 2002)

the time it takes to resolve which task to perform (resolution time). The resolution time depends on the relative time at which response codes for competing tasks are generated, and was computed as follows:

$$resolutiontime = r + f[r - generationtime_j - generationtime_i] \quad (6)$$

where r corresponds to a sample from an ex-Gaussian distribution. Finally, the RT for each task is computed as the sum of generation time, resolution time, and the two constants P and R representing the time taken for perceptual and response-production processes, respectively.

$$ReactionTime_i = P + generationtime_i + resolutiontime_i + R \quad (7)$$

Here, we counted a response as correct if the RT for the currently relevant task was lower than the RT for the currently irrelevant task, and incorrect otherwise [3].

Parameterization

The model parameters were adjusted to yield RT an error rate distributions that matched the behavioral data. The priming factor was set to 0.2 for the current task and set to 0 for the irrelevant task. Noise for the net input was sampled from a Gaussian distribution (μ=0, σ= 0.1) and r was sampled from an ex-Gaussian distribution (μ=200, σ = 240, λ = 150). We fixed c to 0.5 and set perceptual and response-production parameters, P and R, both to 200. In this study, we analyzed switch costs irrespective of stimulus congruency and therefore set f to 0. The priming factor and c were adjusted to receive best ASC fits to the data.

Control for both tasks was first initialized to 0.15, and then adjusted using the stair-casing procedure for each task described by Yeung and Monsell (2003). Each time the model made a correct response the control parameter was decremented by 0.01, and each time the model made an error, the control parameter was incremented by 0.1. The task strength parameter was varied across simulations (see below).

Task Environment

We assessed performance while the model was switching between 64 mini-blocks of two tasks. Each mini-block consisted of four to six trials of the same task. Trial sequences were generated akin to the sequences of the behavioral experiment. We generated the sequence of mini-blocks by counterbalancing which of the two tasks the model is asked to perform, as well as the task transition with respect to the previous mini-block (repetition or switch). Following the analysis procedure of the behavioral experiment, we focused our analyses on RTs and error rates of the first trial of a mini-block (Rogers & Monsell, 1995).

Simulation Procedure

We used the model to generate predictions about ASC as a function of task strength. To do this, we varied the task strength parameters $strength_i$ for both tasks from 0.2 to 0.7 in 0.2 steps across simulations, resulting in 36 parameter configurations. For each parameter configuration, we simulated behavior of the model across 30 task switching sequences. In each trial of a sequence, we set $control_i$ of the relevant task i to the value determined by the adaptation procedure described above while $control_j$ of the irrelevant task $j \neq i$ was set to 0, and fixed all other parameters to their default values. We recorded RTs and error rates for the first trial of each mini-block. We then fitted the drift rate parameter of the DDM separately for each task, as well as for the interaction between task and task transition using the same fitting procedure as described in the Experiment section.

[3] Note, that Yeung and Monsell only analyzed RTs.

Results

Our simulation results indicate that ASC increase with the magnitude of the difference between the strengths of tasks (Fig. 2, 3 & 4). Interestingly, the ASC effect was independent of the absolute magnitude of the task strengths, i.e. the magnitude of the ASC effect remained the same with extremely high or low task strength values measured in terms of main effects on drift rates. However, the range of task drift rates produced by the model did not cover the drift rates obtained from Experiments 1, 3 and 4, preventing a direct comparison. We could only obtain lower task drift rates if the model committed a high amount of errors that did not match human behavior. However, our simulation results suggest that, at least for the range of simulated task strengths, ASC are predicted to depend only on the relative but not the absolute strengths of the two tasks.

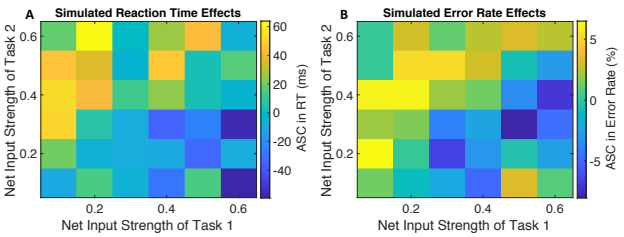


Figure 4: Simulated ASC in RTs and error rates as a function of the input strength for both tasks.

General Discussion and Conclusion

The seemingly paradoxical finding that switching to an easier task is more difficult has been under investigation for the last two decades. Here, we conducted five task switching experiments to systematically investigate ASC as a function of task strength. We manipulated the strength of two tasks, by varying the signal to noise ratio of the corresponding stimulus dimension, and fitted the behavior of human participants with a DDM to quantify (a) the strength of each task and (b) the ASC in terms of changes in the rate of evidence accumulation (drift rate). Our behavioral results indicate that the asymmetry in switch costs between two tasks can indeed reverse if the more dominant task is weakened and the weaker task is strengthened (Experiment 1 and 2). While previous studies

have shown ASC for different tasks, they have typically confounded the difficulty of a task with its identity (Alport et al., 1994; Costa & Santesteban, 2004; Mayr & Keele, 2000; Philipp et al., 2007). Here, we provide empirical support for the hypothesis that ASC can flip if the relative strengths of the two tasks are reversed, even if their identies stay the same. However, our behavioral results suggest that a difference in task strength is not sufficient to yield ASC (Experiment 3-5). That is, we only observed ASC if (a) tasks differed in terms of their strength and (b) the dominant task was relatively easy.

We contrasted human behavior against predictions of a task switching model by Yeung and Monsell (2003) which provides a mechanistic account of this effect in terms of the dynamics of task set priming and top-down control. As in the behavioral experiments, we quantified the strength of each task, as well as the signed magnitude of ASC in terms of drift rate, by fitting the DDM to simulated performance. Our simulation results match the observation that the asymmetry in switch costs should reverse if the strength of two tasks inverses. However, in contrast to participants, the model does not seem to be sensitive to the overall strength of both tasks.

An exaustive analysis of ASC as a function of task strength can help to inform future models of task switching. While most task switching models do not address ASC (Meiran, 1996; Logan & Bundesen, 2003; Brown, Reynolds, & Braver, 2007; Altmann & Gray, 2008), the connectionist model by Gilbert and Shallice (2002) explains ASC, similar to Yeung and Monsell (2003), in terms of differences in top-down input for both tasks: The easier task yields higher switch costs because a stronger top-down input needed to perform a difficult task persists when switching to the easier task. It is worth noting that both models provide a different explanation than Alport et al. (1994). Allport and colleagues suggest that the easier task is associated with higher switch costs because it needed to be suppressed in order to perform the difficult task. In any case, the dependence of ASC on the absolute strength for both tasks presents an interesting challenge for existing and future models of task switching.

Our study provides an important step towards understanding ASC in that it highlights the importance of absolute task strength. However, it does not explain ASC in terms of the factors that contribute to the strength of a task. Here, we operationalized task strength in terms of drift rate that we fitted from RTs and error rates for each task. While this metric allowed us to compare predictions of the model (with respect to task strength) with behavioral performance, it confounds the effects of task automaticity and top-down control on performance. That is, the measured strength of a task may be high, either because it has a high automaticity or because the task receives a high amount of cognitive control. Recent theories of control allocation suggest that the latter can be manipulated by incentivizing accuracy on task (Shenhav, Botvinick, & Cohen, 2013; Musslick, Shenhav, Botvinick, & Cohen, 2015; Botvinick & Braver, 2015). For instance, Umemoto and Holroyd (2015) associated one of two tasks with a higher reward, and observed that participants exhibited lower switch costs when switching to the more rewarded task. Prior modeling work suggests that such incentive-driven differences in switch costs can be attributed differences in allocation of top-down control as opposed to differences in task automaticity (Musslick et al., 2015). Future empirical studies may be able to disentangle the contribution of controlled and automatic processing to ASC, e.g. by manipulating the amount of reward participants receive for a given task.

While task strength appears to play an important role in the explanation of ASC, there are other factors to consider. Yeung and Monsell (2003) found that a delayed onset of the task-irrelevant stimulus (high stimulus onset asynchrony, SOA) could either reduce or reverse the effect of ASC. Moreover, the authors observed no ASC if participants responded with different key presses to each task. These findings identify interference between tasks as a necessary condition for ASC. The results presented here indicate that such interference may not occur when both of the tasks are difficult to perform, even if one of the tasks is much easier than the other.

References

Allport, A., & Wylie, G. (2000). Task switching, stimulus-response bindings, and negative priming. *Control of cognitive processes: Attention and performance XVIII*, 35–70.

Alport, D., Styles, E., & Hsieh, S. (1994). Shifting intentional set: Exploring the dynamic control of tasks.

Altmann, E. M., & Gray, W. D. (2008). An integrated model of cognitive control in task switching. *Psychological review*, *115*(3), 602.

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological review*, *113*(4), 700.

Botvinick, & Braver, T. (2015). Motivation and cognitive control: from behavior to neural mechanism. *Annual Review of Psychology*, *66*.

Botvinick, Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, *108*(3), 624.

Brown, J. W., Reynolds, J. R., & Braver, T. S. (2007). A computational model of fractionated conflict-control mechanisms in task-switching. *Cognitive psychology*, *55*(1), 37–85.

Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: a parallel distributed processing account of the stroop effect. *Psychological Review*, *97*(3), 332–361.

Costa, A., & Santesteban, M. (2004). Lexical access in bilingual speech production: Evidence from language switching in highly proficient bilinguals and l2 learners. *Journal of memory and Language*, *50*(4), 491–511.

Gilbert, S. J., & Shallice, T. (2002). Task switching: A pdp model. *Cognitive Psychology*, *44*(3), 297 - 337.

Kayser, A. S., Buchsbaum, B. R., Erickson, D. T., & D'Esposito, M. (2009). The functional anatomy of a perceptual decision in the human brain. *Journal of Neurophysiology*, *103*(3), 1179–1194.

Kayser, A. S., Erickson, D. T., Buchsbaum, B. R., & D'Esposito, M. (2010). Neural representations of relevant and irrelevant features in perceptual decision making. *Journal of Neuroscience*, *30*(47), 15778-15789.

Logan, G. D., & Bundesen, C. (2003). Clever homunculus: Is there an endogenous act of control in the explicit task-cuing procedure? *Journal of Experimental Psychology: Human Perception and Performance*, *29*(3), 575.

Mayr, U., & Keele, S. W. (2000). Changing internal constraints on action: The role of backward inhibition. *Journal of Experimental Psychology: General*, *129*(1), 4.

Meiran, N. (1996). Reconfiguration of processing mode prior to task performance. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(6), 1423.

Meuter, R. F., & Allport, A. (1999). Bilingual language switching in naming: Asymmetrical costs of language selection. *Journal of memory and language*, *40*(1), 25–40.

Monsell, S., Yeung, N., & Azuma, R. (2000). Reconfiguration of task-set: Is it easier to switch to the weaker task? *Psychological research*, *63*(3-4), 250–264.

Musslick, S., Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2015). A computational model of control allocation based on the expected value of control. In *The 2nd Multidisciplinary Conference on Reinforcement Learning and Decision Making.* Edmonton, Can.

Philipp, A. M., Gade, M., & Koch, I. (2007). Inhibitory processes in language switching: Evidence from switching language-defined response sets. *European Journal of Cognitive Psychology*, *19*(3), 395–416.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological review*, *85*(2), 59.

Rogers, R. D., & Monsell, S. (1995). Costs of a predictible switch between simple cognitive tasks. *Journal of experimental psychology: General*, *124*(2), 207.

Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2013). The expected value of control: an integrative theory of anterior cingulate cortex function. *Neuron*, *79*(2), 217–240.

Simen, P., Contreras, D., Buck, C., Hu, P., Holmes, P., & Cohen, J. D. (2009). Reward rate optimization in two-alternative decision making: empirical tests of theoretical predictions. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(6), 1865.

Umemoto, A., & Holroyd, C. B. (2015). Task-specific effects of reward on task switching. *Psychological Research*, *79*(4), 698707.

Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). Hddm: hierarchical bayesian estimation of the drift-diffusion model in python. *Frontiers in neuroinformatics*, *7*, 14.

Yeung, N., & Monsell, S. (2003). Switching between tasks of unequal familiarity: The role of stimulus-attribute and response-set selection. *Journal of Experimental Psychology: Human perception and performance*, *29*(2), 455.

Yeung, N., Nystrom, L. E., Aronson, J. A., & Cohen, J. D. (2006). Between-task competition and cognitive control in task switching. *The Journal of Neuroscience*, *26*(5), 1429–1438.

Go with Plan A: Backup Plans Help the Powerful but Distract the Powerless

Leila Straub
ETH Zurich, Zurich, Switzerland

Petra C. Schmid
ETH Zurich, Zrich, Switzerland

Abstract

Backup plans represent a safety net that can help ensure goal attainment. However, managing backup plans during goal pursuit can also deflect attention away from the initial plan. We examined how individuals sense of power, which is said to facilitate goal pursuit, affects the extent to which one gets distracted by backup plans. Results from four studies showed that when a backup plan was activated, greater sense of power was associated with lower self-reported distraction and better performance. Studies 2 and 3 further revealed mediating effects of distraction between sense of power and performance. Greater sense of power was associated with less distraction, which in turn was related to better performance. Our findings suggest that when pursuing goals, individuals experiencing high power may be better at allocating their limited cognitive resources to the initial plan.

Ain't that a shame: An exploration into "academic" shame and STEM learning

Anonymous CogSci submission

Abstract

The current study explored the impact that "academic" shame had on learning of the human circulatory system. Participants were randomly assigned to one of two conditions: a shame induction condition or a control condition (no shame induction). Results revealed that the shame induction manipulation was related to higher levels of state shame. Additionally, it was discovered that by and large "in the moment" shame and having a proneness to experiencing shame dampened down any subsequent learning. Implications to education and future research are discussed.

Keywords: shame; cognition; learning; STEM; emotions

Theoretical Framework

Although there are many ways to define shame, for the purposes of this study, shame is an acutely painful affective state that is brought on by a failure to meet internally set rules, ideals, goals, or standards (Turner, Husman, & Schallert, 2002). A gap currently exists in the literature regarding a quantitative exploration of shame. Of the research that has been conducted, much has been qualitative in nature and not focused on "academic" shame (i.e., shame affiliated with learning and education). One possible reason for the underdeveloped exploration of this construct is due to the difficulty in studying it. More specifically, research has shown that individuals may deny their feelings of shame, they tend to self-isolate when they feel shame, and they may be unwilling or unable to express themselves when they feel shame. In fact, one's difficulty in communicating a shameful experience may be a distinctive characteristic of shame (Turner, 2014; Babcock & Sabini, 1990, Lunde, 1958).

Although research has suggested the *difficulties* in studying shame, the difficulty does not detract from the *importance* of studying shame. Tangney and Dearing (2002) suggested that, "Guilt, and especially shame ... are powerful, ubiquitous emotions that come into play across most important areas of life." (p. 8). Contemporary research has shown that experiences of shame can have a "negative impact on interpersonal behavior and functioning" (Tangney & Dearing, 2002, p. 5). Within the context of education, a number of educational psychologists have asserted that feeling shame can interfere with motivation, and negatively impact students' academic goals and achievement (Pekrun, Frenzel, Goetz, & Perry, 2007; Weiner, 1986). Indeed, once students experience shame, their ability to become cognitively engaged may be hindered, they may lose motivation for studying, and, they may feel reluctant to attend class (Turner, Husman, & Schallert, 2002).

Given the importance of gaining a better understanding of this self-conscious emotion, the current study explored the impact that "academic" shame had on learning of the human circulatory system with the hope that we can better understand students' experiences of this emotion.

Current Study

Materials

Test of self-conscious affect The TOSCA-3 (Tangney & Dearing, 2002) was developed as a tool to measure guilt-proneness, shame-proneness, proneness to externalization, and proneness to unconcern. The TOSCA-3 consists of 15 scenario-based situations that test takers may encounter in their day to day lives. Following each scenario, test takers are asked to rate the likelihood of reacting to each of the options on a five-point scale.

Pretest/posttest To assess deep conceptual understanding of the functioning of the human circulatory system, three separate tests were developed in the authors' research laboratory. One test consisted of ten multiple choice questions that were related to the human circulatory system. For example, *"the process of circulation includes which of the following: a) the intake of metabolic materials b) the convergence of metabolic materials throughout the organism c) the return of harmful by products to the environment* ***d) all of the above****".* A second test consisted of 20 matching questions in which the participants had to correctly identify the different components of the human heart. A third and final test consisted of 13 matching questions where the participants had to correctly label the proper functioning of the different parts of the human circulatory system. For example*, "which part of the human circulatory system carries blood away from the heart?" (answer: arteries).*

Self-regulated learning-self report survey (SRL-SRS) The SRL-SRS is intended to measure self-regulation as a relatively stable attribute in multiple learning domains and is based on Zimmerman's self-regulated learning theory. It is comprised of six subscales: planning, self-monitoring, evaluation, reflection, effort, and self-efficacy (Toering, Elferink-Gemser, Jonker, van Heuvelen, & Visscher, 2012).

Casual dimension scale-II The CDS-II consists of 12 closed ended 9-point Likert scale items designed to assess causal attributions related to achievement outcomes. The CDS-II measures attribution across the following four areas: locus of causality (e.g., the cause of your performance reflects an aspect of yourself), external control (e.g., the

cause of your performance is under the power of other people), stability (e.g., the cause of your performance is permanent), and personal control (e.g., the cause of your performance is something you can regulate) (McAuley, Duncan, & Russell, 1992).

Experiential shame scale According to Turner (2014), the Experiential Shame Scale (ESS) is "an opaque measure of physical, emotional, and social markers of shame experiences...developed to address the difficulties of assessing state shame." The ESS consists of eleven questions in which the test taker indicates the number that best describes how they feel right now when comparing two opposite word states. For example, "Physically, I feel [Very Warm 1--2--3--4--5--6--7 Very Cool]".

Participants

Participants consisted of 40 students from a private liberal arts university located in the southern United States. Volunteers fulfilled a course requirement in their general psychology class for their participation.

Procedure

Before entering the lab, participants were randomly assigned to either the experimental (i.e., shame induction) group or the control group. After completing the informed consent, participants were given as much time as needed to complete the TOSCA-3. They then completed the three circulatory system tests. Following completion of the pretests, participants then were asked to fill out the SRL-SRS.

Before beginning the ACT practice problems, participants were read the following instructions: "During this portion of the study you will be asked to complete a series of problems. **These are problems that, as a college student,** should not be extremely challenging for you. In order to recreate a scenario that would match an actual testing environment, you will have 30 minutes to complete the test. After you submit the test, instructions will appear on the screen that will let you know the next steps that you will need to take in this study. Please let the experimenter know if you have any questions at this time. Thank you again for your participation!" The bolded portion in the instructions is the only difference between what is read to participants in the control group and experimental group (i.e., experimental group receives the bolded statement). For the experimental (i.e., shame induction) group, after finishing the ACT, a text box appeared that stated "Your combined score on the test was: 40%. The average (school name; removed for blind reviews) student scored 90%. Please let the experimenter know your score so that it can be catalogued." The control group received the following feedback once they had completed the ACT practice problems: "You have now completed this portion of the study. Please let the experimenter know you are ready to proceed."

Immediately following the completion of the ACT practice problems, participants were asked to complete the Experiential Shame Scale in order to measure state shame (i.e., "in the moment shame"). Participants then filled out the Causal Dimension Scale-II and began interacting with a hypermedia encyclopedia (this served as our instructional delivery to assess the impact of shame on learning). Before interacting with the encyclopedia, they were read a set of instructions by the experimenter which told the learner that their job was to spend 30 minutes learning all they could about the human circulatory system. Participants were required to use the full 30 minutes before moving on from this part of the study. Following completion of the encyclopedia, participants were given the circulatory system posttests, were debriefed, and were then allowed to leave.

Results

Participants in the shame induction condition (M = 4.5) scored significantly higher on the ESS than participants in the control condition (M = 3.6), t (38) = 2.876, p = .007, d = .91. See Figure 1.

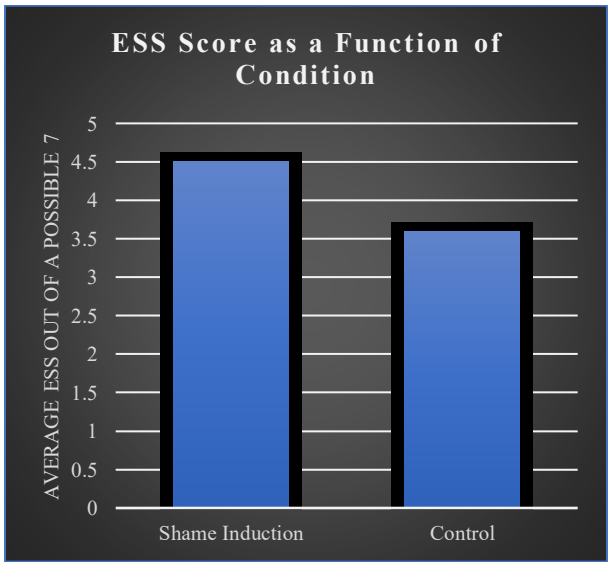


Figure 1: Average shame score as a function of condition.

Initial results revealed that participants in the control condition (M = 1.5) learned significantly more from pretest to posttest compared to participants in the shame induction condition (M = .50), F (1, 38) = 3.188, p = .04 (one-tailed) on the multiple-choice dependent measure. See Figure 2.

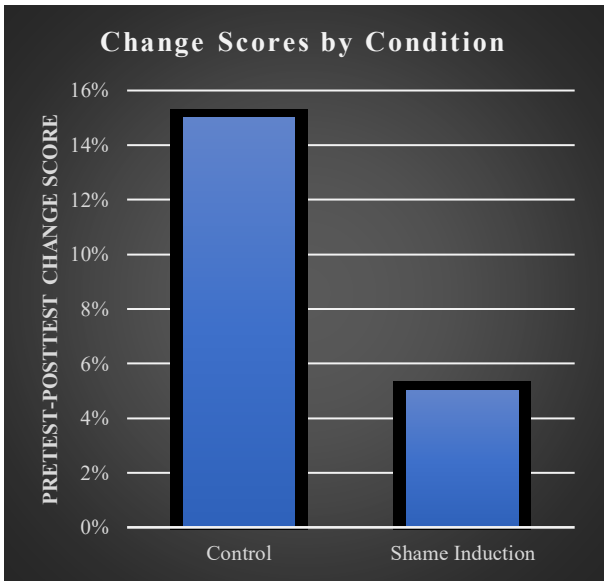


Figure 2: Average learning gain as a function of condition.

A significant main effect was found between the variables "shame proneness" with change scores as the dependent measures. More specifically, change scores on the matching test revealed that participants with a low proneness to shame (M = 5.4) learned significantly more than participants with a high proneness to shame (M = 2.1), p = .000. Additionally, when looking at all tests combined, participant with a low proneness to shame (M = 12.94) learned significantly more than participants with a high proneness to shame (M = 7.34), p =.002. See Figure 3.

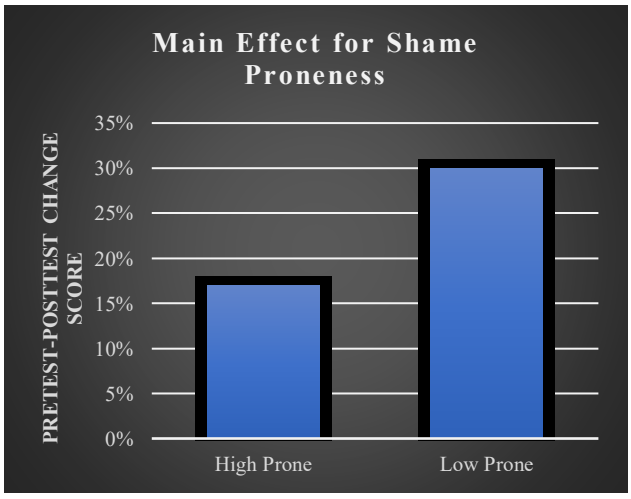


Figure 3: Average learning gain as a function of shame proneness.

Significant interactions were discovered between condition and shame proneness. Participants in the shame induction condition with a high proneness to shame (M = 2.18) learned significantly less than participants in the shame induction condition with a low proneness to shame (M = 6.5), p = .001 (Matching Test).

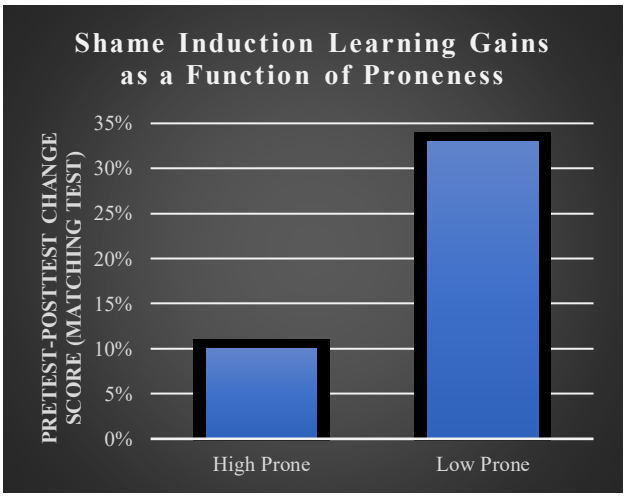


Figure 4: Average matching test learning gain for shame induction condition as a function of proneness.

Similarly, participants in the shame induction condition with a high proneness to shame (M = 3.82) learned significantly less than participants in the control condition with a low proneness to shame (M = 7.8), p = .05 (Labeling Test).

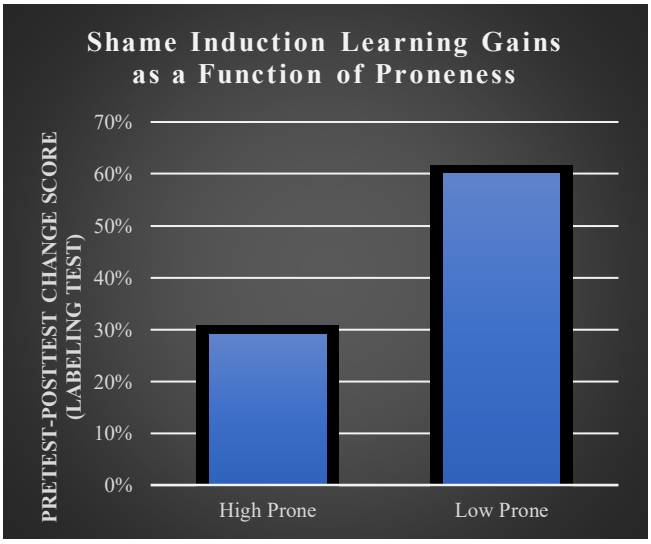


Figure 5: Average labeling test learning gain for shame induction condition as a function of proneness.

Additionally, participants in the control condition with a low proneness to shame (M = 4.3) learned significantly more than participants in the control condition with a high proneness to shame (M = 2.0), p = .036 (Matching Test Only).

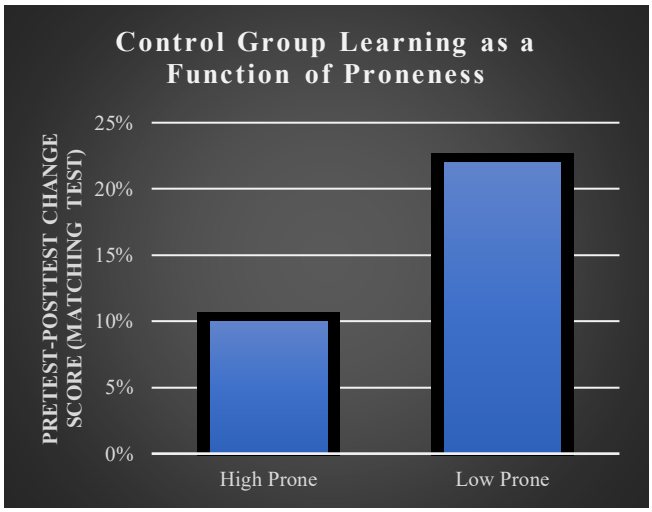


Figure 6: Average matching test learning gain for control condition as a function of proneness.

When looking at the change scores of all tests combined, participants in the shame induction condition with a high proneness to shame (M = 8.2) learned significantly less than participants in the shame induction with a low proneness to shame (M = 10.9), p = .002.

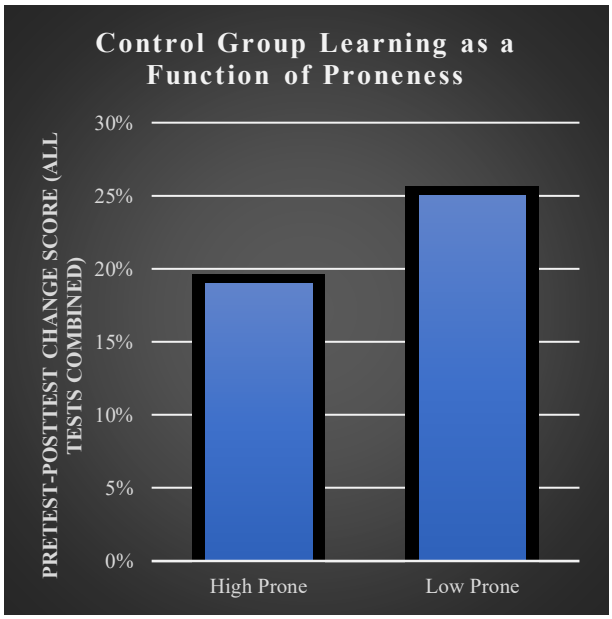


Figure 7: Average learning gains across all tests for control condition as a function of proneness.

Discussion

The results from the current study demonstrated that it is possible to have a systematic quantitative exploration of the self-conscious emotion shame. More specifically, those participants randomly assigned to the shame induction condition had higher instances of "in the moment" shame (as measured by the ESS) compared to those in the control condition. The methodology and findings are consistent with previous research that has found that feelings of shame are significantly positively correlated with feelings of shock (Turner, Husman, & Schallert, 2002).

Furthermore, as can be seen from these preliminary results, by and large, "in the moment" shame and shame proneness appear to be detrimental to the learning of complex science topics (i.e., human circulatory system). Participants randomly assigned to the shame induction condition learned significantly less about the circulatory system compared to participants in the control condition. Furthermore, a main effect was found showing that those with a high proneness to shame learned significantly less about the circulatory system compared to participants with a low proneness to shame. Finally, several significant interactions were discovered that revealed the detrimental impact of shame on learning. As mentioned earlier, this finding is in line with previous findings that have shown that feeling shame can interfere with motivation, and negatively impact students' academic goals and achievement (Pekrun, Frenzel, Goetz, & Perry, 2007; Weiner, 1986). Furthermore, once students experience shame, their ability to become cognitively engaged may be hindered, they may lose motivation for studying, and, they may feel reluctant to attend class (Turner, Husman, & Schallert, 2002).

What if a teacher was able to figure out which subset of students were actually experiencing shame and were able to be proactive to the potential negative consequences? Mitigating shame-consequences by understanding the who- and when-indicators of shame experiences, could facilitate teachers' ability to provide motivational interventions. A better understanding of the when and how of shame may be especially important given that individuals may deny their feelings, and may be unwilling or unable to express themselves, particularly if they self-isolate. In other words, as of now, we have no reliable way (other than perhaps self-report measures) to determine who is experiencing shame. Thus, intervention is near impossible without perceiving reliable indicators.

References

Babcock, M. K., & Sabini, J. (1990). On differentiating embarrassment from shame. *European Journal of Social Psychology, 20*,151-169.

McAuley, E., Duncan, T. E., & Russell, D. W. (1992). Measuring causal attributions: The revised Causal Dimension Scale (CDII). *Personality and Social Psychology Bulletin, 18*, 566-573.

Pekrun, R., Frenzel, A. C., Goetz, T., & Perry, R. P. (2007). The control-value theory of achievement emotions: An integrative approach to emotions in education. In P. A. Schutz & R. Pekrun (Eds.), *Emotions in education*. San Diego: Academic Press.

Tangney, J., & Dearing, R. L. (2002). *Shame and guilt*. New York:Guilford Press.

Toering, T., Elferink-Gemser, M. T., Jonker, L., van Heuvelen, M. J. G., & Visscher, C. (2012). Measuring self-regulation in a learning context: Reliability and validity of the Self-Regulation of Learning Self-Report Scale (SRL-SRS). *International Journal of Sport an Exercise Psychology, 10,* 24–38. doi: 10.1080/1612197X.2012.645132

Turner, J. E. (2014). Researching state shame with the experiential shame scale. *Journal of Psychology,148*,577–601.

Turner, J. E., Husman, J., & Schallert, D. L. (2002). The importance of students' goals in their emotional experience of academic failure: Investigating the precursors and consequences of shame. *Educational Psychologist*, 37, 79 – 89.

Weiner, B. (1986). Cognition, emotion, and action. In R. M. Sorrentino & E. T. Higgins (Eds.), *Handbook of motivation and cognition: Foundations of social behavior* (pp. 281-312). New York: Guilford Press.

Jessie and Gary or *Gary and Jessie*?: Cognitive Accessibility Predicts Order in English and Japanese

Karina Tachihara (tachihara@princeton.edu)
Department of Psychology, Princeton University, Princeton NJ 08544
Miah Pitcher (mpitcher@princeton.edu)
Department of Psychology, Princeton University, Princeton NJ 08544
Adele E. Goldberg (adele@princeton.edu)
Department of Psychology, Princeton University, Princeton NJ 08544

Abstract

Notably, while English tends to prefer shorter before longer complements (*explained to us a very clear effect*), Japanese displays the opposite tendency. Far less cross-linguistic work has investigated possible differences in the ordering of nouns within conjunctions ("binomials'), although a corpus study suggests that the same factors predict binomial ordering in Japanese and English. To investigate the issue experimentally, we report Japanese and English speakers' productions of names of the members of couples that they knew personally. Results confirm that conceptual accessibility is the most important factor in the ordering of familiar name binomials in both languages. That is, both groups tended to name the member they felt closer to first. Length (syllables/mora) was not a significant predictor in either language. Differences in the preferred order of verbs' complements are then attributable to other factors, possibly a very general preference to minimize the average distance between semantically related elements.

Keywords: accessibility; binomials; Japanese; English; word order

Introduction

Accessibility refers to the speed and accuracy with which concepts are activated in memory. When English speakers produce utterances, more accessible and shorter grammatical phrases tend to be produced before less accessible, longer phrases (Bock 1982, 1987; Levelt 1989; Bock & Levelt 1994; McDonald, Bock & Kelly 1993; Bock & Warren 1985; Carroll 1958; Tomlin 1995; Downing & Noonan 1995). This has been argued to allow for more efficient processing insofar as it reduces the need to hold accessible phrases in working memory while less accessible phrases are retrieved and produced first (Ferreira and Dell 2000; Branigan and Feleki 1999; Prat-Sala and Branigan 2000; Ferreira & Yoshita 2003; Kempen & Harbusch 2004).

The factors that have been evoked in discussions of accessibility are quite diverse. They include animacy (McDonald, Bock, & Kelly 1993; Ferreira 1994; Prat-Sala & Branigan 2000; Christianson & Ferreira 2005); givenness in discourse Ferreira & Yoshita 2003; Prat-Sala and Branigan 2000); prototypicality (Onishi, Murphy, and Bock 2008); and basic level status (Lohmann & Takada 2014).

Several of these factors can be quite difficult to tease apart. For example, animacy and discourse-givenness tend to be correlated because people--or agentive entities more generally--are the most common topics of conversation. Moreover, discourse-givenness correlates strongly with length, since previously introduced entities are commonly referred to using pronouns, which are short, and in many languages, discourse-given arguments need not be expressed at all (Ariel 1988; Byrne & Davidson, 1985; McDonald et al., 1993; Narasimhan & Dimroth, 2008).

Yet there is good reason to try to distinguish or control for animacy and discourse-givenness in investigations of length and conceptual accessibility. Importantly, certain factors that result in a shorter-earlier tendency in English (Arnold, Wasow, Losongco, & Grinstrom, 2000; Arnold, 2003; Gries, 1999; Stallings, MacDonald, & O'Seaghdha, 1998; Stallings & MacDonald, 2011; Wasow, 2002) produce the *opposite* order in Japanese (Chang, 2009; Hakuta, 1981; Yamashita & Chang, 2001). For instance, in English, particle placement, the dative-alternation and "heavy NP shift" all prefer particularly long complements to be uttered later in the string, while in Japanese especially long complements tend to be produced earlier (Dryer 2000; Hawkins 1994, 2004; Gibson 1998; Yamashita and Chang, 2001).

The first study to experimentally demonstrate a preference for longer-earlier in Japanese was Yamashita and Chang (2001). They interpreted this finding in a way that attempted to preserve the idea that *all* speakers prefer to express more accessible entities first, by invoking a distinction between formal complexity and cognitive accessibility. They suggested that longer phrases should be considered more semantically or conceptually accessible, even though they are more complex. This raised the following possibility, as described by Jaeger and Norcliffe (2009:876): "Japanese speakers [may be] more sensitive to conveying meaning (putting enriched material earlier), while English speakers prefer to sequence forms (putting easier to produce, e.g., shorter, words earlier, (Yamashita and Chang 2001, 2006)."

The current study tests the possible distinction between conceptual accessibility and length—here, the number of syllables or mora—on how English and Japanese speakers produce the names of familiar couples. If Japanese speakers are influenced more by conceptual accessibility and less affected by length when compared to English speakers, it would provide evidence that a distinction between conceptual accessibility and length underlies the difference between English and Japanese' word order preferences. We refer to

this hypothesis in what follows as the Conceptual Accessibility vs. Formal Accessibility hypothesis (CA v. FA).

Hawkins (1994, 2004) suggested an alternative explanation for the shorter-earlier preference Japanese and the longer-earlier preference in English. He argued that both Japanese and English display a preference to minimize the average distance between a verb and its non-subject complements. His "minimal distance" proposal is satisfied in Japanese and other verb final languages by positioning longer complements before shorter complements (<longer> <short >V). English and other VO languages obey the same preference by expressing short complements *before* longer complements (V<short> <longer>). But if Hawkins' proposal accounts for the shorter-earlier preference in VO and the longer-earlier preference in OV, it raises the question as to whether there *also* exists an accessible-early preference in Japanese, English and other languages.

We address these important issues by considering the preferred word order in both Japanese and English, given a case that clearly involves conceptual accessibility. This allows us to determine whether speakers of both languages prefer to order more conceptually accessible terms earlier (or both prefer to order them later). The idea that the difference between shorter-earlier English and longer-earlier Japanese is due to a Japanese preference for conceptually-accessible-earlier and an English preference for formally-accessible-earlier would predict that Japanese speakers should weigh conceptual accessibility more strongly than length, while English be more strongly affected by length than conceptual accessibility.

We report experimental results which compared the ordering of "binomial" conjunctions (<noun> and <noun>) by speakers of English and speakers of Japanese. Specifically, we investigate the ordering of the names of couples that are personally known to participants (e.g., *Jessie and Gary*). We hypothesized that the person the speaker feels a closer connection to will be named before the other member of the couple in both languages. We recognize that feelings of emotional closeness are hard to decompose, but at the same time, we take it as self-evident that if semantic accessibility is to be a meaningful construct at all, our mental representation of an individual whom we feel closer to should, *ceteris paribus*, be more semantically accessible than our mental representation of someone we feel comparatively less close to. We recognize that if one member of the couple is already under discussion, then all things are not equal. Therefore discourse-givenness is controlled for in the current experiment: participants simply generate the names of couples that they know with no additional context provided. Thus, if, in both Japanese and in English, the name mentioned first tends to be the name of the member of the couple whom the participant feels a greater personal attachment to, it will be evidence that *both* languages prefer to order more cognitively accessible words first.

There already exists a good deal of work on how English speakers order binomial phrases, but with rare exceptions described below, comparative work on the construction is exceedingly rare. Moreover, studies of English binomials have offered a wide range of often quite specific predictors of ordering but have only rarely invoked accessibility explicitly. For instance, Cooper and Ross (1975) suggested 19 factors which included the first element of a binomial being more "Here, Now, Adult, Male, Positive, singular, Living, Friendly, Solid, Agentive, Powerful, at Home, and Patriotic" (pg. 67).

This classic study led to a number of refinements. For example, Benor and Levy (2006) quantified a model that included 20 constraints related to aspects of lexical semantics, phonetics, and frequency. Morgan and Levy (2016) reduced this list to the following seven factors (in order of effect size): iconic sequencing (e.g., early before later), perceptual markedness (which encompassed the majority of factors proposed by Cooper & Ross), formal markedness, power, final stress, length, and frequency. These weighted constraints produced a model that predicted the preferred order in a large corpus of natural speech with 77% accuracy. Notably absent from these discussions was mention of a possible role for accessibility. Onishi et al. (2008), a rare study that did explicitly evoke accessibility as a key factor in English binomial order, introduced yet another predictor: more prototypical members of categories tended to be produced before less-prototypical members.

Importantly, Morgan & Levy (2016) also demonstrated that experience with specific binomial expressions influences the way familiar binomials are expressed. Specifically, they found that the frequency of familiar binomials correlated with reading time when binomials were ordered in the familiar way, and frequency correlated negatively when the two nouns were read in reverse order. Morgan & Levy proposed that the generative factors they proposed influenced the ordering of *novel* combinations of words. While a large number of binomial expressions are familiar, it is equally important to ask how conventional binomials ("freezes") come to be ordered in the particular ways they are (Mollin, 2014). To this end, an early cross-linguistic study of English, Russian and German by Fenk-Oczlon (1989) found that the relative frequency of words determined the ordering of 400 frozen binomial expressions with 84% accuracy; however, Lohmann & Takada (2014) found frequency to be much less influential.

Lohmann & Takada (2014) provides an important precedent for the current work, as they compare results from corpus analyses of binomial expressions in Japanese and English texts. This study included a number of potential predictors including power (including male and "importance"), iconicity (early before later), frequency, discourse-givenness, length (in syllables or mora), and conceptual accessibility. Conceptual accessibility, in this study, was treated as an umbrella category that included animacy, concreteness, prototypicality, basic level, proximal and self before other. In this work, which likely included a number of "frozen" binomials since it was based on corpus data, significant effects were found for length, power,

iconicity, discourse-givenness and accessibility but not frequency in both languages. The Lohmann & Takada work explicitly omitted conjoined proper names from their analyses. But by considering personal names that are known to the participants, the current work is able to index cognitive accessibility with a single factor, closeness. In addition, the ordering of names of familiar couples in our experimental context avoids potential confounds of animacy and givenness, as well as avoiding freezes that are influenced by the language at large. Possibly relevant factors of length and gender are included in the preregistered analyses.

There are two other key precedents for the current study. Like the current study, Wright et al. (2005) also considered the ordering of "Name and Name" phrases. Critically, however, that study differed from the current one in that the experimenters provided names without referents. Therefore, participants had no opportunity to rely on personal experience with the people involved. The study found a bias to order male before female names and shorter before longer names, two factors that have been proposed for English binomials generally, but which are not necessarily related to cognitive accessibility, the key factor of interest in the current work.

A precedent for considering "psychological closeness" to be relevant to binomial order comes from Iliev & Smirnova (2014). This work hypothesized that "psychological closeness of the speaker to one of the poles in the word pair" should predict order with the closer entity positioned earlier (pg. 210). Unlike the current study, all proper names were excluded from analysis. Instead, in one study, websites about cars, politics, religion were analyzed. Results demonstrated that sites sponsored by Honda, for example, were more likely to mention *Honda* before its competitors; liberal leaning websites were more likely to mention *liberal* before *conservative,* and to a lesser extent, websites about Islam showed a tendency to mention *Muslim* before *Christian*. A second study focused on gender and results were more equivocal. The authors hypothesized that male authors should be more likely to order male terms before female, while female authors might show the reverse tendency. Notably, however, male terms were ordered before female terms 93% of the time by male authors and 90% of the time by female authors. The strong skewing toward male-first, also found in previous work, may partially be due to the fact that many relevant phrases are conventionally frozen in English (e.g., *men and women; husband and wife*). A final study was experimental rather than based on corpus data; it elicited various binomials from participants by asking for the top two colleges in Chicago, the two main political parties in the US, the traditional two genders and so on. Participants showed a tendency to name their university first (Northwestern, 67%), and liberal students were more likely to name *Democrat* before *Republican* than were conservative students. Echoing theirs and others' corpus work, an overwhelming majority of respondents produced *male* before *female* (91%), although of the participants who produced *female* first, 80% were women.

An analysis of how participants order the names of familiar couples satisfies several desiderata. It allows us to avoid expressions that are conventional in the language at large, which are recognized to be subject to many general influences as documented in other work. Names are particularly well-suited as an index of cognitive accessibility because a name selects an individual rather than a category: We might know several people named *Gary*, but when we talk about *Gary and Jessie* we have particular individuals in mind, and our representation of Gary, Jessie and their names are dependent our own particular experiences. The experimental context enables us to control for animacy and discourse-givenness, while keeping the generation of names similar to that of natural production. Finally, by comparing Japanese and English, we can determine whether either or both languages tend to order more conceptual accessible names earlier.

Method

Participants

60 native speakers of English living in the US and 60 native speakers of Japanese living in Japan were recruited on Amazon Mechanical Turk as participants and moderately compensated for their time.

Procedure

Participants first answered questions about their gender and native language. They were then asked to name 3 sets of important couples in their life. They entered the name of each member of the couple in blank boxes. For the Japanese survey, participants were also asked to provide the phonetic spelling for each name. The rest of the survey asked whether or not participants were related to either or both of members of each couple, who they felt they were closer to, and the gender of each member of the couple. For these questions, the order of names that had been given were randomized for each participant.

Response coding & model development

To analyze the data, we followed the model of ordering preference for binomial expression introduced in previous work by Levy and colleagues (Benor & Levy 2006; Morgan & Levy 2016). The model predicts the likelihood that the ordering preference for a given pair is consistent with various planned fixed effects. First, each pair was coded in an essentially arbitrary way, specifically whether or not the names were ordered alphabetically. This was used as the outcome variable. Next, for each response, each fixed effect was assigned 1 if the factor predicted the alphabetical order and 0 if it predicted a non-alphabetical order. For example, if the participant indicated that they were closer to Gary than Jessie, closeness would receive a 1 because both alphabetical order and closeness predicted the same order, *Gary and Jessie*. If they had indicated that they were closer to Jessie than Gary, then closeness would receive a 0 because the alphabetical order (*Gary and Jessie*) does not match the closeness preference (*Jessie and Gary*). Note that we are not

testing whether or not there is a preference for alphabetical order. Rather, we use alphabetical order as a basis to get a binary code to compare with the order the participant provided.

To see if the length of names affected their ordering, we counted the number of syllables in each name for English and for Japanese, the number of morae, a more appropriate measure of length in that language (Otake, Hatano, & Mehler, 1993). We then calculated the difference in number of syllables/morae between each pair of names. We assigned this number a positive score if alphabetical order and ordering based on longer-before-short matched (the longer name was earlier in the alphabet) and a negative score when they did not.

Results

Before presenting the results of the model, we present the raw percentages of responses in the pooled data for each coded factor in Table 1. The person whom the participant reported feeling closer to was named first 65% of the time in Japanese and 77% of the time in English.

Cognitive accessibility (closeness) (%)		Gender (%)		Length (%)	
JAPANESE					
1st	65	M-F	56	Long-Short	31
2nd	35	F-M	30	Short-Long	21
		Same	14	Same	48
ENGLISH					
1st	77	M-F	54	Long-Short	33
2nd	23	F-M	43	Short-Long	40
		Same	3	Same	27

Table 1. % of responses for each fixed effect for Japanese (top) and English (bottom). Percentages rounded to the closest integer.

We first created models for each language independently. For this we used a multilevel model with closeness, gender, and length as fixed effects, random intercepts for subject, and alphabetical order as the outcome (Barr, Levy, Scheepers, & Tily 2013), using the lmerTest library (R Development Core Team 2008).

For the English data, the model revealed a significant effect of closeness (β = -0.52, t = -7.31, $p < 0.0001$); the tendency to order males first was not significant, (β = -0.32, t = -1.45, p = 0.15) and neither was a tendency to order shorter before longer names (β = -0.002, t = 0.10, p = 0.9).

The model for the Japanese data also revealed a significant effect of closeness (β = -0.29, t = -4.19, $p < 0.0001$) and no effect of length (β = 0.04, t = 0.90, p = 0.37). Unlike the English data, a marginal effect of gender was found with male names being more likely to appear before female names (β = -0.23, t = 1.98, p = 0.05).

In order to better quantify the importance of each of these effects, we used a leave-one-out method in which we compared a model without each effect to the full model. For both English and Japanese, conceptual accessibility (as operationalized as closeness) significantly improved the model (English, $\chi^2 = 46.02$, $p < 0.0001$; Japanese, $\chi^2 = 16.79$, $p < 0.0001$). Length did not improve either model (English, $\chi^2 = 0.01$, $p = 0.92$; Japanese, $\chi^2 = 0.69$, $p = 0.41$). Gender significantly improved the model only for Japanese (χ^2 = 13.30, $p = 0.001$), and not for English ($\chi^2 = 3.72$, $p = 0.16$). While there seems to be a difference in importance of gender in Japanese and English (or rather Japan and US), all analyses indicate that conceptual accessibility is the most important predictor of binomial expression of proper names.

In order to compare the effect size of conceptual accessibility (closeness) in the two languages, we looked at the interaction of closeness and language using the combined data. For this we used a multilevel model with gender and length as independent fixed effects, closeness and language as interacting fixed effects, random intercepts for subject, and alphabetical order as the outcome. The model found a significant effect of closeness (β = -0.53, t = 7.05, $p <$ 0.0001), and a significant interaction of closeness and language (β = 0.25, t = 2.53, p = 0.01), suggesting that closeness is a larger effect for English than Japanese.

Discussion

The ordering of the names of familiar couples was found to be strongly predicted by which member of the couple the speaker felt closer to. Taking personal closeness as an index of cognitive accessibility, we find that cognitive accessibility was the strongest predictor of name ordering in both English and Japanese, operating in the same direction in both languages: more cognitive accessible names tended to be produced first. This effect was stronger in English than in Japanese, although it is possible that the difference in effect size was due to the fact that a gender effect (male-before-female) was only evident in Japanese. That is, given that gender accounted for some of the variance, it is not surprising that the only other significant effect (cognitive accessibility) accounted for somewhat less in Japanese.

The lack of male-before-female bias in the current English data is intriguing, given that a male-before-female bias has been consistently found in prior corpus work (Cooper & Ross 1975; Lohmann & Takada 2014), and notably, on work involving on non-referential proper names (Wright, Hay, & Bent 2005). The reason a male-first bias exists at all deserves more discussion than we can offer here. Insofar as it is rooted in cultural sexism, it may be relevant that personal contact is recognized to reduce this and other forms of prejudice

(Pettigrew & Tropp, 2006). Japanese society obeys more stereotypical gender norms than the US (Bresnahan, Inoue & Kagawa, 2006; Saito, 2007), which might lead to a weak effect of male-before-female bias in Japanese.

Length was not a significant factor in English or Japanese, nor was there an interaction. And this lack of significance was apparent regardless of whether we treated length as a continuous or binary value. We note that it is possible that the lack of an interaction was due to a lack of power, since the names of each couple were commonly equal in length. That is, 27% of the couple names in English were of equal number of syllables and 48% the two names had the same number of morae in Japanese. Intriguingly, if we consider only the combinations of names that did differ in length, the trends in Japanese and English numerically pattern in opposite directions. Specifically, the ratio of shorter-first in English was roughly 4:3, while in Japanese, the ratio of *Longer*-First was roughly 3:2. Iliev & Smirnova (2014) had found evidence of shorter-first in binomals in both languages, but they had found the effect to be 3x as large in English as Japanese. Thus it is possible that a shorter-first bias only exists in English binomials. Future work with a larger sample may be necessary to confirm this trend.

Let us return to the striking difference in preferred order of especially long complements in English and Japanese. Previous work had appealed to a distinction between conceptual accessibility and lexical accessibility, suggesting that longer phrases are "semantically richer" and that "This semantic richness increases the overall accessibility of the phrase in the conceptual arena" (Chang & Yamashita 2001:B53). Shorter phrases were recognized to be more accessible in the formal (lexical) domain. The difference between Japanese and English then, was that "In English, weight-based shifts [word order variation] seem to be less sensitive to conceptual factors." However, in the current work, we have seen that if anything, English shows a *stronger* conceptually-accessible-early bias than Japanese does.

The current work allows that cognitive and formal/lexical accessibility need not be mutually dissociable. Clearly, certain episodic memories, smells, or images may be more or less cognitively accessible, depending on context and encoding. So clearly conceptual accessibility cannot be reduced to formal or lexical accessibility. But it is reasonable to assume that lexical (or formal) accessibility is simply a type of cognitive accessibility.

Our results are consistent with Hawkin's (1994; 2004) proposal that languages prefer to minimize the distance between the verb and its (non-subject) complements. This ordering is beneficial to listeners since the interpretation of a verb often critically depends on its co-occurring complements. This is clear in English, for instance, in the contrasts between, e.g., *hitting on an idea; hitting on someone; hitting someone up for something*; *hitting a place vs. a person vs. a goal.* See also Chang (2009) for interesting discussion how the minimal distance idea may emerge over the course of learning. In fact, the minimal-distance preference has been generalized to other kinds of semantic dependency relations and validated across a number of languages (Choi, 2007 for Korean; Faghiri & Samvelian 2014 for Farsi[1]; Gildea & Temperley 2010 for English and German; Liu, 2008 for 15 languages; and Futrell et al., 2015 for 20 languages).

The present work finds that both English and Japanese show a preference to produce more conceptually accessible terms first. Prior work has established that languages also generally appear to prefer to minimize the distance between a verb and its arguments. While these types of processing biases may differ in their strength across languages, the present work supports the idea that language processing systems emerge in much the same way in speakers of different languages. This is perhaps to be expected insofar as language processing is shaped by constraints on memory, learning and interpretability.

Conclusion

To conclude, results in both English and Japanese confirmed that the order of names of couples, personally familiar to a participant, were most strongly predicted by which member of the couple the participant felt a closer personal attachment to. By investigating the ordering of the names of familiar couples, animacy and discourse-givenness were controlled for. Investigating the names of couples known to participants was also advantageous because the ordering is not expected to be affected by language-wide conventions. Results did not reveal length to be a significant factor, and gender only played a (relatively small) role in the Japanese data. Therefore, we submit that feelings of personal closeness serve as a useful and relatively direct index of cognitive accessibility.

Thus, the present work provides evidence that cognitive accessibility plays a similar strong role in word order in both Japanese and English. This undermines the possibility that the reverse ordering preferences in Japanese and English clauses is a result of cognitive accessibility influencing the two languages in different ways. Instead, the Japanese

[1] Hawkins had argued for a more specific proposal, namely that the *heads* of dependents should be as close as possible to their external head. This proposal motivates the idea that verb final languages tend to have *post*positions, while verb-medial languages tend to have *pre*positions. However, Faghiri & Samvelian (2014) find that Farsi speakers prefer longer-early, parallel to Japanese. But while Farsi is an SOV language like Japanese, it has prepositions rather than postpositions. Therefore as Farghiri & Samvelian (2014) observe, the longer-early preference in Farsi cannot be explained in terms of a preference to minimize the distance between a verb and the *head* of its complement, since when a PP is is long, the long-early preference actually lengthens the distance between the V and P: <[**P** long>$_{IO}$ <short>$_{DO}$ **V**. Nonetheless, Farsi is consistent with the idea that languages and speakers prefer to reduce the average distance between semantically related units (Gildea & Temperley, 2010).

ordering preference for grammatical phrases (longer-early) must be due to some other factor, quite possibly a preference to indicate a verb's arguments as close to the verb as possible (Hawkins 1994, 2004). The current study demonstrates that, *ceteris paribus*, speakers of both Japanese and English prefer to produce more cognitively accessible words early (Arnold et al. 2000; Ferreira & Dell 2000).

References

Ariel, M. (1988). Referring and accessibility. *Journal of linguistics*, *24*(1), 65-87.

Arnold, J. E., Wasow, T., Losongco, A., & Ginstrom, R. (2000). Heaviness vs. Newness: The effect of structural complexity and discourse status on constituent ordering. *Language*, 76(1), 28–55.

Arnold, J. E. (2003). Multiple Constrains on Reference Form. In *Preferred Argument Structure: Grammar as architecture for function*. John Benjamins Publishing.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Bates, E., & MacWhinney, B. (1987). Competition, variation, and language learning. In B. MacWhinney (Ed.), *Mechanisms of language acquisition* (pp. 157-193). Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Benor, S. B., & Levy, R. (2006). The chicken or the egg? A probabilistic analysis of English binomials. *Language*, 233-278.

Bock, J. K. (1982). Toward a cognitive psychology of syntax: Information processing contributions to sentence formulation. Psychological Review, 89, 1–47.

Bock, J. K. (1987). An effect of the accessibility of word forms on sentence structures. Journal of Memory and Language, 26, 119–137.

Bock, J. K., & Levelt, W. J. M. (1994). Language production: Grammatical encoding. In M. A. Gernsbacher (Ed.), Handbook of psycholinguistics (pp. 945–984). San Diego: Academic Press.

Bock, J. K., & Warren, R. K. (1985). Conceptual accessibility and syntactic structure in sentence formulation. Cognition, 21(1), 47–67.

Branigan, H., & Feleki, E. (1999). Conceptual accessibility and serial order in Greek speech production. In M. Hahn, & SC. Stoness (Eds.), *Proceedings of the 21st Annual Conference of the Cognitive Science Society* (pp. 96-101). Mahwah: Lawrence Erlbaum Associates.

Bresnahan, M. J., Inoue, Y., & Kagawa, N. (2006). Players and Whiners? Perceptions of Sex Stereotyping in Animé in Japan and the US. *Asian Journal of Communication*, *16*(2), 207–217.

Byrne, B., & Davidson, E. (1985). On Putting the Horse before the Cart: Exploring Conceptual Bases of Word Order via Acquisition of a Miniature Artificial Language. *Journal of Memory and Language; New York*, *24*(4), 377–389.

Carroll, J. B. (1958). Communication theory, linguistics, and psycholinguistics. Review of Educational Research, 28, 79–88.

Chang, F. (2009). Learning to order words: A connectionist model of heavy NP shift and accessibility effects in Japanese and English. *Journal of Memory and Language*, *61*(3), 374–397.

Choi, H.-W. (2007). Length and Order: A Corpus Study of Korean Dative-Accusative Construction. *Discourse and Cognition*, *14*(3), 207–227

Christianson, K., & Ferreira, F. (2005). Conceptual accessibility and sentence production in a free word order language (Odawa). *Cognition*, *98*(2), 105-13.

Cooper, W. E., & Ross, J. R. (1975). World order. *Functionalism,* Grossman, RE, James San, L. and Vance, TJ, (Eds.), 63-111.

Downing, & M. Noonan (Eds.), (1995). Word order in discourse (pp. 517–554). Amsterdam: John Benjamins Publishing

Dryer, M. S. (2000). Counting genera vs. counting languages. *Linguistic Typology*, *4*(3), 23.

Faghiri, Pegah & Pollet Samvelian. 2014. Constituent ordering in Persian and the weight factor. In Christopher Pinon (ed.), *Empirical issues in syntax and semantics 10 (EISS10)*, In press.

Fenk-Oczlon, G. (1989). Word frequency and word order in freezes. Linguistics 27, 517– 556

Ferreira, F. (1994). Choice of passive voice is affected by verb type and animacy. *Journal of Memory and Language*, *33*(6), 715–736.

Ferreira, V. S., and G. S. Dell (2000). Effect of ambiguity and lexical availability on syntactic and lexical production. *Cognitive Psychology* 40(4).296–340.

Ferreira, V. S., & Yoshita, H. (2003). Given-new ordering effects on the production of scrambled sentences in Japanese. *Journal of Psycholinguistic Research*, *32*(6), 669–692.

Futrell, R., Mahowald, K., & Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, *112*(33), 10336-10341.

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition,* 68, 1–76

Gildea, D., & Temperley, D. (2010). Do grammars minimize dependency length? *Cognitive Science*, *34*(2), 286-310.

Gries, S. T. (1999). Particle movement: A cognitive and functional approach. *Cognitive Linguistics*, *10*(2).

Hakuta, K. (1981). Grammatical description versus configurational arrangement in language acquisition: The case of relative clauses in Japanese. Cognition, 9, 197–236.

Hawkins, J. A. (1994). A performance theory of order and constituency. Cambridge, UK: Cambridge University Press.

Hawkins, J. A. (2004). Efficiency and complexity in grammars. New York City: Oxford University Press.

Iliev, R., & Smirnova, A. (2016). Revealing Word Order: Using Serial Position in Binomials to Predict Properties of the Speaker. *Journal of Psycholinguistic Research*, *45*(2), 205–235.

Jaeger, T. F., & Norcliffe, E. J. (2009). The cross-linguistic study of sentence production. *Language and Linguistics Compass*, *3*(4), 866-887.

Kempen, G., & Harbusch, K. (2004). Generating Natural Word Orders in a Semi–free Word Order Language: Treebank-Based Linearization Preferences for German. In A. Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Processing* (pp. 350–354).

Levelt, W.J., 1989. Speaking: From Intention to Articulation. MIT Press, Cambridge, MA.

Liu, H.T. (2008) Dependency distance as a metric of language comprehension difficulty. J. *Cognitive Science*. 9 (2):159–191.

Lohmann, A., & Takada, T. (2014). Order in NP conjuncts in spoken English and Japanese. *Lingua*, *152*, 48–64.

McDonald, J. L., Bock, K., & Kelly, M. H. (1993). Word and world order: Semantic, phonological, and metrical determinants of serial position. Cognitive Psychology, 25, 188–230

Mollin, S. (2014). *The (ir) reversibility of English Binomials: Corpus, Constraints, Developments* (Vol. 64). John Benjamins Publishing Company.

Morgan, E., & Levy, R. (2016). Abstract knowledge versus direct experience in processing of binomial expressions. *Cognition*, *157*, 384-402.

Narasimhan, B., & Dimroth, C. (2008). Word order and information status in child language. *Cognition*, *107*(1), 317–329.

Onishi, K. H., Murphy, G. L., & Bock, K. (2008). Prototypicality in sentence production. *Cognitive Psychology*, *56*(2), 103–141.
Otake, T., Hatano, G., & Mehler, J. (1993). Mora or Syllable? Speech Segmentation in Japanese. *Journal of Memory and Language*, *32*(2), 258–278.
Pettigrew, T. F., & Tropp, L. R. (2006). A meta-analytic test of intergroup contact theory. *Journal of personality and social psychology*, *90*(5), 751.
Prat-Sala, Merce & Branigan, Holly. (2000). Discourse Constraints on Syntactic Processing in Language Production: A Cross-Linguistic Study in English and Spanish. *Journal of Memory and Language.* 42. 168-182.
Saito, S. (2007). Television and the cultivation of gender-role attitudes in Japan: Does television contribute to the maintenance of the status quo? *Journal of Communication*, *57*(3), 511–529.
Stallings, L. M., MacDonald, M. C., & O'Seaghdha, P. G. (1998). Phrasal ordering constraints in sentence production: Phrase length and verb disposition in heavy-NP shift. *Journal of Memory and Language*, *39*(3), 392-417.
Stallings, L. M., & MacDonald, M. C. (2011). It's not Just the "Heavy NP": Relative Phrase Length Modulates the Production of Heavy-NP Shift. *Journal of Psycholinguistic Research*, *40*(3), 177–187.
Tanaka, M. (2003). Conceptual accessibility and word-order in Japanese Proceedings of the Postgraduate Conference. Edinburgh: University of Edinburgh.
Tomlin, R. S. (1995). Focal attention, voice, and word order: An experimental, cross-linguistic study. In P. Downing & M. Noonan (eds.), *Word Order in Discourse*. Amsterdam: John Benjamins: 517-552.
Venables, W. N., & Smith, D. M. (2008). the R Development Core Team (2003). *Introduction to R (Version 1.6. 2). http://cran. r-project. org/doc/manuals/R-intro. pdf*.
Wasow, T. (2002). *Postverbal behavior*. (No. 145). CSLI.
Wright, S. K., Hay, J., & Bent, T. (2005). Ladies first? Phonology, frequency, and the naming conspiracy. *Linguistics*, *43*(3), 531-561.
Yamashita, H., & Chang, F. (2001). "Long before short" preference in the production of a head-final language. *Cognition*, *81*(2), B45-B55.

Neural dynamic concepts for intentional systems

Jan Tekülve (jan.tekuelve@ini.rub.de)

Gregor Schöner (gregor.schoener@ini.rub.de)
Institut für Neuroinformatik, Ruhr-Universität Bochum
44780 Bochum, Germany

Abstract

How may intentionality, the capacity of mental states to be about the world, emerge from neural processes? We propose a set of theoretical concepts that enable a simulated agent to have intentional states as it perceives, acts, memorizes, plans, and builds beliefs about a simulated environment. The concepts are framed within Dynamic Field Theory (Schöner et al., 2015), a mathematical language for neural processes models at the level of networks of neural populations. Inspired by Searle's analysis of the two directions of fit of intentional states (Searle, 1980), we recognize that process models of intentional states must detect the match of the world to the mind (for "action" intentions) or the match of the mind to the world (for "perceptual" intentions). Neural representations of Searle's condition of satisfaction implement these detection decisions through dynamic instabilities that are instrumental in enabling autonomous switches among intentional states.

Keywords: Dynamical systems modeling; Mathematical modeling; Neural networks; Intelligent agents; Cognitive Architectures

Introduction

How are neural processes organized to create coherent, complex cognitive function? For instance, how are sequences of actions and processes of active perception generated to orient actions at objects to achieve a desired outcome? How may the nervous system switch between actions and mental states that are driven by current sensory information and actions or mental states that are driven by memory and knowledge?

Philosophers of mind have framed related questions in terms of the notion of *intentionality*: How may an organism with its nervous system generate intentional states that are about objects in the world? How may an organism act to change the world according to its intentional states? The logical structure of this problem has been analyzed in depth by John Searle (Searle, 1980). He postulates that intentional states come in two directions of fit (DoF), the *world-to-mind* direction of fit, in which an intentional state's content represents a desired state of the world, capturing the intuitive "action" flavor of intention. The *mind-to-world* DoF comprises states in which the state's content matches circumstances in the world, a "perceptual" flavor of intention. Each intentional state can be described through its *condition of satisfaction* (CoS), which determines whether the fit between mind and world is achieved. Searle has conjugated these two forms of intentionality through three layers of psychological modes: *intention-in-action* (IiA) and *perception* are intentional states directly linked to the motor or sensory systems. *Prior intention* and *memory* are intentional states with a more indirect form of linkages, in which additional steps are needed to act out or bring about the intentional state. *Beliefs* and *desires* are more abstract forms of intentionality, typically thought to take propositional forms, with an inherent generalization beyond the immediately accessible perceptual or motor experience.

We come to these questions from the theoretical framework of Dynamic Field Theory (DFT) (Schöner et al., 2015), a mathematical language for neural processes models at the level of networks of neural populations. Here, we take inspiration from Searle's concepts to address the neural processes required to autonomously switch between intentional states in these six psychological modes. A key idea has been that there must be neural processes that explicitly represent a CoS and whose activation controls the transitions from one intentional state to another (Sandamirskaya & Schöner, 2010). Specifically, for world-to-mind intentional states, activation of the neural representation of the CoS signals the successful achievement of an intentional state that leads to its deactivation and opens the system to switch to a subsequent intentional state. In mind-to-world intentional states, it is the representation of the content of the intention itself that forms the CoS, which is activated when a detection decision is made and remains activated as long as the intentional state persists.

In this paper we develop this idea into a systematic account of how intentional states can be organized to autonomously generate goal- and object-oriented behavior. We simulate a rudimentary toy scenario, in which an agent explores its simple environment containing colored objects and buckets of paint. The agent may move towards objects and direct an effector to them, either taking up paint (for a bucket) or painting the object (for the colored objects). The agent detects objects, may attentionally select objects, may build scene memories, generate sequences of actions to paint particular objects with a particular paint, and learn and exploit beliefs about which paint applied to which surface generates which outcome. Simple desires (to seek particular outcomes of painting acts) drive the agents goal-oriented and exploratory behaviors. The scenario is chosen such that the amount of time each action or mental operation takes varies, and that during that time the agent is exposed to other perceptions or sensory states that could distract from its current intention. The inherent stability of its intentional states and the capacity to

release these states from stability under the right conditions is thus probed in this scenario.

Dynamic Field Theory

Dynamic Field Theory (DFT) (Schöner et al., 2015) is a theoretical framework for understanding perception, motor behavior, and cognition based on neural principles.The activity in neural populations is modeled by activation fields, $u(x,t)$, spanned across the metric dimensions, x, to which the population is tuned. The neural dynamics of the activation fields,

$$\tau\dot{u}(x,t) = -u(x,t) + h + s(x,t) + \int \omega(x-x')\sigma(u(x',t))dx$$

describes the time-continuous evolution of neural activation on the time scale τ. Activation $u(x)$ below the sigmoidal threshold σ relaxes to the stable solution $h+s(x)$, defined by the field's resting level h and its localized inputs $s(x)$. Field sites, where activation strength surpasses the threshold level, will engage in lateral interaction defined by the field's kernel $\omega(x-x')$, which is locally excitatory and inhibitory over longer distances $x-x'$. This leads to the formation of self-stabilized peaks of supra-threshold activation, which are the unit of representation in DFT (see figure 1).

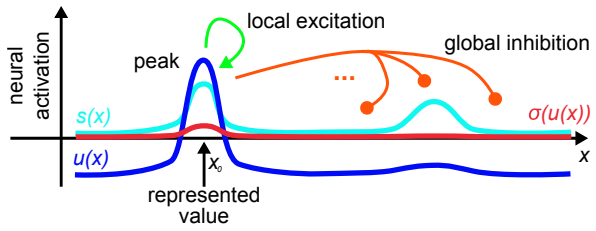


Figure 1: A dynamic neural field spanned across the metric-dimension x representing value x_0 through a supra-threshold activation peak.

Depending on the individual strength of excitatory and inhibitory interaction, fields may allow the formation of multiple peaks (self-stabilized), single peaks (selective) or they may sustain peaks once localized input is removed (self-sustained). Multi-dimensional fields may represent conjunctions of feature dimensions, for example, the conjunction of color and space. Zero-dimensional fields are dynamic neural nodes that represent categorical states.

Two fields u_{src} and u_{tar} may be coupled by adding a field's output $\sigma(u_{\text{src}})$ to the other field's rate of change $\dot{u}_{\text{tar}}$, weighted with a homogeneous connection kernel $\omega_{\text{src,tar}}$. Such projections may preserve the dimensionality of the fields, or may expand or contract the field dimensionality (Zibner & Faubel, 2015). *Dimensionality expansions* may take the form of ridges (or tubes, or slices), in which input along one or several of the receiving field's dimension is constant. *Dimensionality contractions* typically entail integrating along the contracted dimension. Dynamic neural nodes that project homogeneously onto a field by expansion are called *boost nodes*. They may alter the dynamic regime in the target field and induce the formation or vanishing of peaks. Within field architectures such boost nodes may effectively modulate the flow of activation by enabling or disabling particular branches of an architecture to create units of representation. *Concept nodes* project a specific pattern on a higher dimensional field to elicit a peak representing the concept, e.g. a blue-concept node activates neurons tuned to blue hue in a field spanned across the color dimension.

The transition from a stable sub-threshold solution to a new supra-threshold activation pattern marks a discrete event in the presence of time-continuous input variations and is labelled *detection instability*. In the context of intentional states the detection instability is utilized to determine a state's condition of satisfaction, the discrete point in time where a successful match between world and mind representations is achieved.

In the world-to-mind DoF a *matching field* (CoS field) receives sub-threshold input from an *intention-field* representing the desired world-state and sub-threshold input from a *perception-field* representing the current world-state. Due to the resting level h in relation to the strengths of both field inputs, a supra-threshold peak will only form in the matching-field, if both input patterns overlap sufficiently, thus signaling the states CoS through a detection instability. Representation of a world-to-mind CoS is thus independent from the planned timing of the underlying action and signals its termination on a perceptual basis. The formation of a CoS may thus be used to terminate the action and activate the next action in a planned sequence (Richter, Sandamirskaya, & Schöner, 2012).

In the mind-to-world DoF the CoS is determined through the formation of a peak in a field that is connected to sensor or memory substrates. The detection instability may be the result of salient input alone or of the combination of sensor/memory input and top-down attention input from within the neural architecture. Representations of a mind-to-world CoS are made available to the rest of the architecture and may be used in further cognitive processing, e.g. determining a world-to-mind CoS.

Transforming Searles logical analysis of intentional states into a process account has led us to a number of new insights. One is a difference in the time structure of world-to-mind vs. mind-to-world intentional states. World-to-mind intentional states are active before the corresponding state of the world has been achieved and are deactivated once the CoS detects a match between the expected and the sensed state of the world. Mind-to-world intentional states, in contrast, often persist beyond the detection of a match, which is an essential characteristic of memories and beliefs. But what if memories or beliefs (and even percepts) are false? Then they must be deactivated. This is controlled by a condition of dissatisfaction (CoD), which detects a mismatch between current sensory or internal information and an intentional state. Upon activation, a CoD inhibits that intentional state. The CoD responds to evidence against the intentional state, not to the mere absence

of evidence supporting the intentional state.

Model/Scenario

We illustrate how intentional states can be organized to generate autonomous goal- and object-oriented behavior in a minimal scenario requiring Searle's six major psychological modes. The scenario contains a simulated agent engaged in an artificial painting task controlled by a dynamic field architecture connected to the robots sensorimotor surface (see figure 2 for a sketch).

Mind-to-World States

Intentional states of the mind-to-world DoF are the prerequisite to engage in meaningful actions in a given environment as any action at least aims to achieve a perceivable outcome.

Perception The virtual environment contains cuboids of different height and color, which are arranged in an array along a single dimension facing the robotic agent. The agent's visual perception fields are therefore spanned across horizontal retinal space and the two feature dimensions height and color. A selective spatial attention mechanism causes peaks to form in the same spatial location in the *space/color* and *space/height* perception fields, representing a perception of the particular height and color features at that particular location (see Grieben et al. (2018) for details on the attentional selection). To detect successful interaction with the world, the agent perceives changes in the environment through a two-layer transient detector that forms peaks in response to sudden changes in visual input (see Berger et al. (2012) for details).

To monitor its own actions the agent requires self-perception of the task-dependent "body parts", which includes an estimate of the agent's position in the world. A simulated sensor provides input to a one-dimensional *current position* field, as the agent's movement is restricted to driving in parallel to the cuboid array. Arm movement is restricted to two Cartesian dimensions, lateral and forward translation, which leads to a two-dimensional representation of the current end effector position in the *proprioception* field. The painting device is located at the robot's end effector and can either be filled with color or not. This categorical perceptual state is represented through a neural node that is activated if the device is filled.

Attention directed towards particular self-perceptions is modeled through a homogeneous resting level boost, which causes the sub-threshold sensor information to form a peak

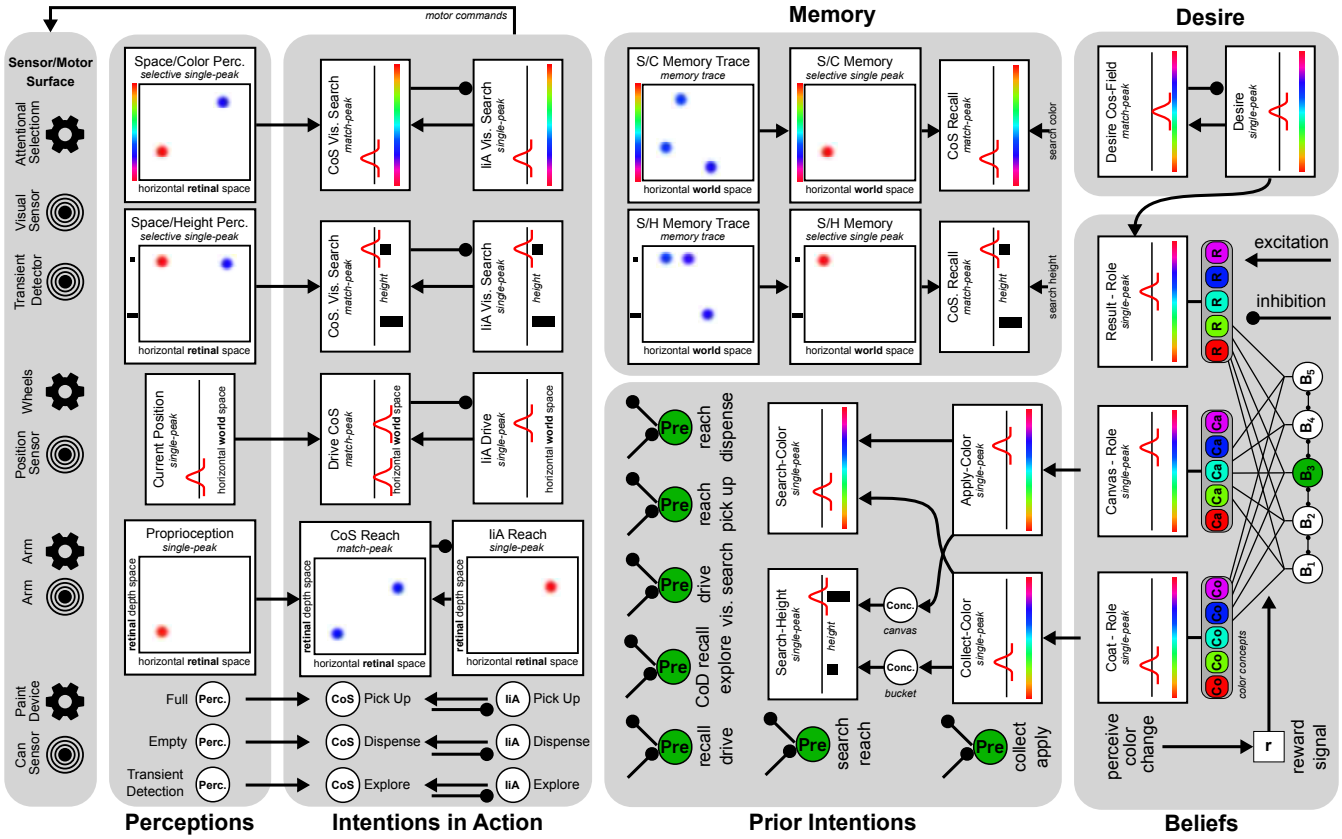


Figure 2: Schematic overview of the dynamic fields and nodes representing the agent's intentional states grouped according to their psychological modes. For clarity's sake only the most relevant connections are shown and parts of the architecture relevant to autonomous learning and exploration are hidden. Prior intentions are depicted as precondition nodes with labels describing the inhibiting CoS followed by the inhibited IiA.

in the respective perception field. Neural interaction in perception fields is strong enough to prevent the destabilization of perceptions through noise, but retains its input coupling such that a continuous change in input induces a drift in peak position.

Memory To allow the agent to engage in more sophisticated actions that are not purely based on current perceptions, the agent stores past perceptions of cuboids in memory. Each visual perception of the agent leaves a slowly decaying two-dimensional memory-trace spanned across world-space and feature, modeling a memory process that is subject to interference (Erlhagen & Schöner, 2002). The trace is forwarded as sub-threshold activation to a space/feature *memory* field analog to the visual perception fields. Memory states represented as peaks in the memory field may emerge through either spatial or feature cues overlapping with the memory trace substrate.

Self-sustained fields retaining task-relevant information, such as the recently collected color, represent working memory, which is functionally closer to the mode of perception than memory, as self-sustained peaks resemble lasting perception representations and do not need an additional detection mechanism to form.

Belief Meaningful interaction with the world also relies on general knowledge or beliefs about the world represented in propositional form. In the toy scenario, beliefs are about relations between the three color concepts: the color of a canvas, the color of the paint, and the color that results from coating the canvas with the paint. Each painting action contributes to the formation of a belief about that relation. The relation is represented through a neural node with reciprocal connections to three color concept nodes, each linked to a different *color role* field. An activated belief state is represented through a supra-threshold belief node that leads to the formation of three peaks, each in one self-sustaining color role field, which provide working memory representations to guide the painting process. The color concept nodes ensure a degree of generalization, as different shades of hue activate the same concept node, while the activation of the concept node activates the mean hue value of the particular color.

A belief is activated when color nodes in either of the three roles become active, to which the belief has learned synaptic connections. For instance, a belief linking the red point on a blue canvas to a yellow result may become activated, if the result color node yellow is activated by a corresponding desire. Inhibitory coupling among belief nodes ensures that only a single belief may be activated at any time. The belief with most matching color role input will typically win the competition and can then be used to guide action. If an active color role does not match the learned projections of any belief, no belief is activated.

The learning of new beliefs is organized by a neural dynamic architecture inspired by Adaptive Resonance Theory (Carpenter & Grossberg, 2016) illustrated in Figure 3. It associates color concepts in the three roles coating, canvas and result color with a single belief node. Such learning steps occur whenever the transient detector registers a change of color during a painting action. This happens under two possible conditions. In one case, a belief has previously been activated that predicts the expected color change. If that prediction is confirmed, a Hebbian learning mechanism consolidates the connectivity. If that prediction is not confirmed, the CoD is activated, and the belief is inhibited. This leaves the system without any activated belief. That second case, no activated belief, may also arise because there was no matching belief to begin with. In this case, a belief note is recruited for learning the new association between coating, canvas, and resulting color. This happens through a homogeneous boost of all belief nodes. Only a previously uncommitted belief node has a chance to become activated, because each belief node is inhibited by a dedicated "commit node" that represents that this belief node is committed to a particular belief it has learned.

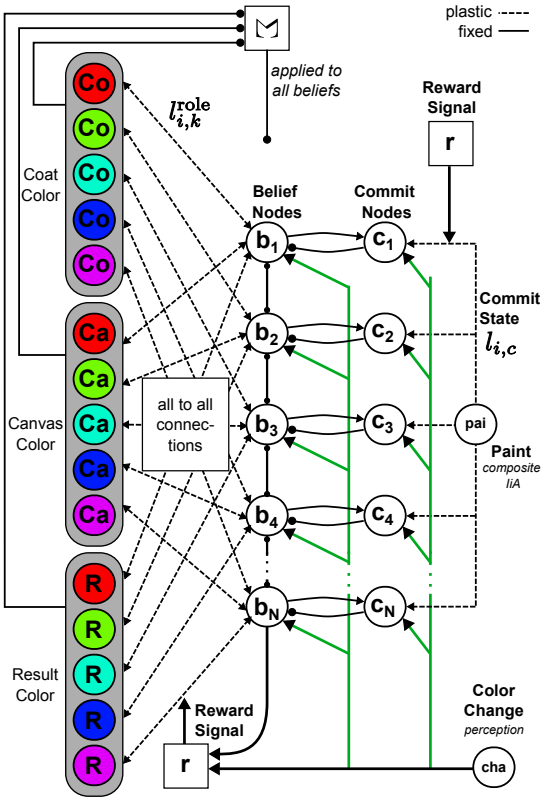


Figure 3: Detailed sketch of the belief learning architecture shown in Figure 2.

The actual learning processes is modulated by a transient reward signal, $r(t)$, that is generated in the presence of an active belief node and a detected color change in the scene. The reward modulated Hebbian learning rule adapts the connections, $l^{\text{role}}_{i,k}$, between belief nodes, b_i, and color-role concept nodes, u^{role}_k (where k is color and role $\in \{$*coat, canvas, result*$\}$):

$$\dot{l}^{\text{role}}_{i,k} = \eta\, r(t)\, \sigma(b_i)\sigma(u^{\text{role}}_k).$$

The learning rate, η, is chosen such that a new belief is learned within a single transient epoch of reward in a form of one-shot learning. For a more detailed analysis of the mechanisms of autonomous learning see (Tekülve & Schöner, 2019).

World-to-Mind States

Intentional states of the world-to-mind DoF are instrumental in bringing about a desired state of affairs in the world, which includes the agent's own body. All world-to-mind states share the representation through the pair of intention and CoS (or match) field. Outgoing connections from the intention field specify the actions driven by a supra-threshold activation peak, while the outgoing inhibitory connection from the CoS field terminates the action once the desired state is detected.

Intention in Action The painting scenario provides several elementary actions that may take a variable amount of time and thus require a representation of a CoS to verify their successful execution. *Reaching* to a particular location in the visual array is realized through a neural field architecture for generating arm movements (see Zibner et al. (2015)). Its duration depends on the relative distance between the agent and the target location. The target location is defined through spatial input from the visual perception fields, which classify reaching as an object oriented action.

Moving to a particular position in the world is motivated through memory instead of perception. The *drive* IiA field thus receives its spatial input from peaks formed in the space/feature memory fields. In absence of a particular target location the agent may also move to either direction until a previously unattended cuboid is perceived. The *explore* IiA realizes this behavior and its CoS is represented through a binary neural node receiving excitatory input from the visual transient detector. The actions explore, *pick up* and *dispense* represent a family of IiAs, where the desired world state is categorical and represented through the activation of a neural node.

Another family of IiAs is represented by the actions *visual search*, *recall* and *activate belief*, which treat the current state of the neural system as part of the world and try to induce particular states of the mind-to-world DoF. Visual search guides the attentional system to achieve a perceptual state matching an intended feature cue, while recall tries to achieve a memory state matching an intended feature cue and activate belief intends to activate a belief node that matches certain color-roles.

Prior Intention Most goal directed actions comprise a sequence of actions such as the painting task in this scenario which requires: Searching for a "color bucket" (high cuboid), collecting color from it, searching for a "canvas" (small cuboid) and applying the collected color on it. Those actions themselves may be described as sequences of more elementary actions, e.g. searching comprises the sequence of recalling a cuboid's position, driving to the position and visually searching for the cuboid, while collecting and applying comprise reaching followed by picking up or dispensing color.

Such a sequence of actions (or composite IiA) is realized through an intention-field that simultaneously activates all IiAs involved in the sequence and an inhibiting precondition node for each IiA. The combination of activating and specifying the input of an IiA, while simultaneously inhibiting it, represents a prior intention. The prior intention turns into an IiA once the precondition node is destabilized by the CoS of a preceding IiA which releases the IiA-field from inhibition (Richter et al. (2012)). These CoS fields may sustain activation in a working memory representation of the current stage within a sequence.

The CoS of a composite IiA is activated through a subset of CoS representations of comprising IiAs determining the successful completion of the composite IiA's goal. This will inhibit the composites IiAs intention field and subsequently destabilize all working memory representations of the comprising IiAs, thus allowing the same sequence of actions to be activated again, which is required in the scenario as searching for a cuboid is part of both collecting and applying color.

Prior intentions may also represent alternative action plans that may occur when a precondition node is destabilized by a CoD, for example, due to failing to recall a specific cuboid.

Desire The agent's desire to observe the change of a cube's color into a desired color is the drive for all actions it executes. The desire specifies the agent's prior intentions of collecting and applying color through the activation of a belief that matches the desired result color. The *desire CoS* is activated through a match between a changing color detected by the visual transient detector and the desired color, which leads to a subsequent inhibition of the desire returning the field architecture to its initial state.

Results

Figure 4 shows activation snapshots of selected fields displaying the formation of CoS peaks during a successful painting sequence. In snapshot $\mathbf{t_1}$ the desire to paint a cube yellow feeds into the result-role field (left column), which triggers a detection instability in belief node B_4 leading to a complete belief representation through the emergence of peaks in the canvas and coat role fields (right column). The coat color leads to an activation of the collect IiA to retrieve blue color and a prior intention to apply the color to a purple canvas, which is represented through a sub-threshold peak in the apply IiA field.

At $\mathbf{t_2}$ the IiA collect activates the "bucket" concept, a high cuboid, which is forwarded as a recall cue to the space/color and space/height memory fields respectively. The collect color also forms the prior-intention to visually search for blue color (left column). The color/height cue leads to the emergence of a single memory peak at the location of the blue/high cuboid, which is read out across space and leads to the formation of a peak in the IiA drive (right column).

The left column of snapshot $\mathbf{t_3}$ shows the IiA drive-field

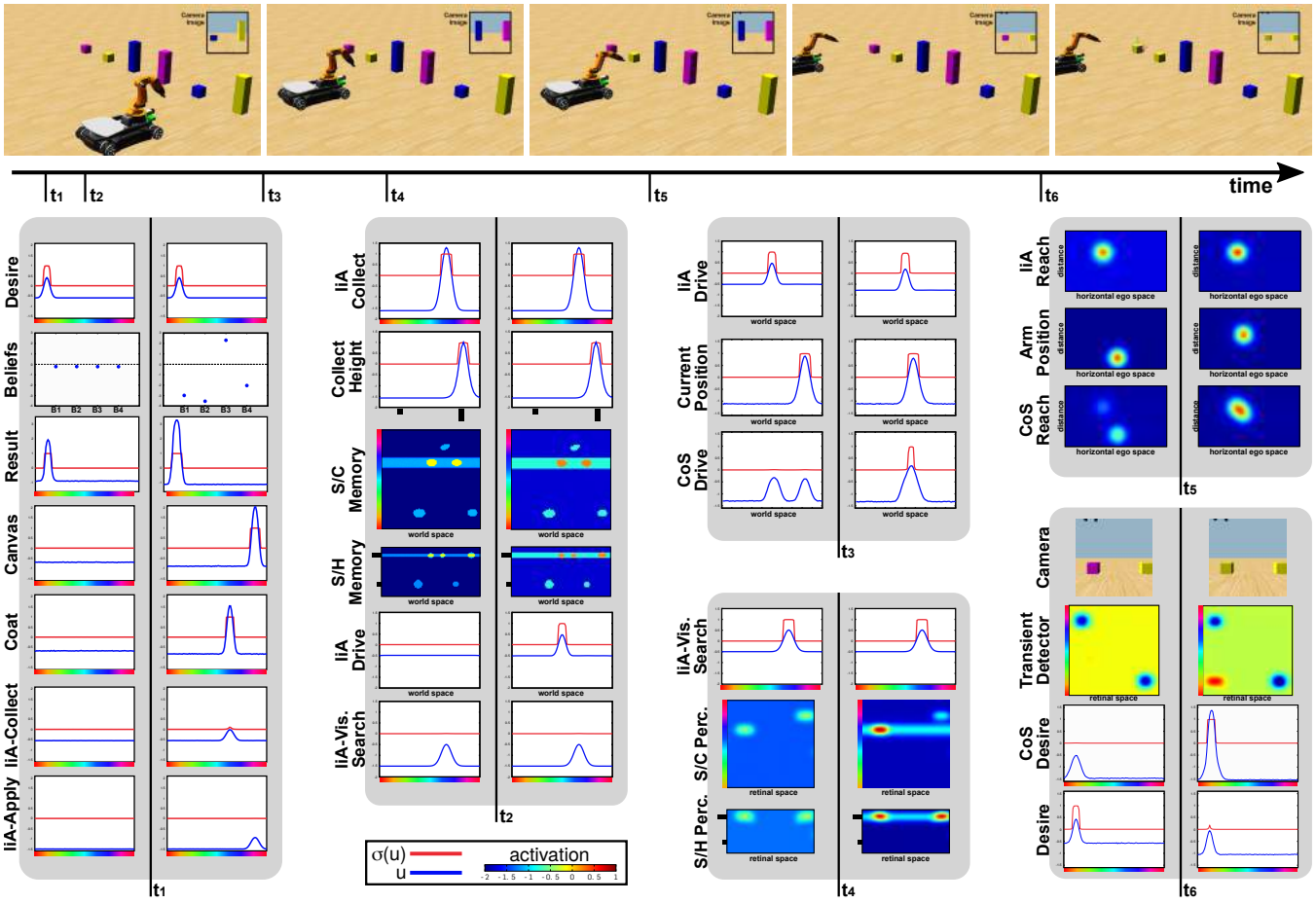


Figure 4: Activation snapshots of selected fields displaying the formation of CoS peaks during a successful painting sequence.

representing the desired goal in the center of the world space, which does not match the currently perceived position. Once movement causes a match between desired and perceived position, a peak emerges in the CoS field causing the termination of the drive IiA (right column).

The detection instability in the drive CoS leads to an autonomous transition to the visual search IiA, which is released from inhibition through the precondition node at $\mathbf{t_4}$ (left column). Ridges of searched color and height are induced in the perception fields causing the emergence of supra-threshold peaks at overlapping positions. (right column).

A match of height and color is detected through the visual search CoS (not shown), which leads to an autonomous transition to the reaching IiA at $\mathbf{t_5}$ by forming a peak at the perceived retinal position in the two dimensional reach IiA field (left column). The reaching IiA is destabilized after a successful arm movement leads to detection instability in the reach CoS-field due to an overlap between proprioception of the eef and the reach goal position (right column).

Snapshot $\mathbf{t_6}$ shows the detection of a color change in the visual scene by the transient detector after the agent successfully collected the color and dispensed it on the purple cube (left column). Activation of the dispense IiA (not shown) changes the color of the purple cube to yellow, which forms a supra-threshold in the transient detector at the conjunction of left and yellow. The color of the perceived change matches with the desired color and leads to the emergence of a peak in the desire CoS which causes a subsequent destabilizing of the desire to paint a cube yellow (right column).

Discussion

We present a neural dynamic architecture that endows a robotic agent with the capability to generate intentional states of the six major psychological modes. Self-stabilized peaks of activation within neural populations determine the content of an intentional state while the state's psychological mode is determined by how the neural population is positioned within a neural dynamic architecture. The CoS of intentional states is modeled through the detection-instability of dynamic neural fields with the DoF determining under which circumstances the instability emerges.

The architecture is demonstrated in a toy scenario, where the agent seeks particular perceptual states (desires), and uses beliefs about contingencies of which paint transforms which color into which new color, to plan sequences of action (prior intentions) based on its current scene representation (mem-

ory). The agent follows these plans by driving towards objects and interacting with them (intention-in-action), if they visually match specified feature combinations (perception). Beliefs about the three different color roles are learned autonomously during each painting sequence.

Similar goals are pursued by Schrodt and colleagues (Schrodt & Butz, 2016; Schrodt et al., 2017), who learn production rules within a cognitive architecture. That work is framed within a probabilistic approach, which is partially embedded in neural networks. Our methods to achieve autonomous sequencing overlap with techniques developed in (Kazerounian & Grossberg, 2014). Globally speaking, we pursue similar aims as the research program of cognitive architectures (Anderson, 1996). Our emphasis is to be pervasively consistent with neural principles, generating the sequence of processing steps autonomously from neural dynamics alone. Although the functions fulfilled by portions of the neural dynamics can be described using concepts of information processing, the system is simply a set of integro-differential equations that generate time courses of activation. These integro-differential equations capture the time-continuous evolution of activation in populations of cortical and subcortical neurons (Erlhagen, Bastian, Jancke, Riehle, & Schöner, 1999). It remains a challenge to provide direct neural support for a complex model like ours (see (Wijeakumar, Ambrose, Spencer, & Curtu, 2017) for an outline of how that may happen). Empirical support for a model like ours may also be sought in the form of behavioral signatures of the neural dynamics, an approach that has been successful for past DFT models. The highly integrative nature of the model makes this difficult, but perhaps not impossible.

Future modeling tasks include scaling the demonstrated principles to more complex task-environments, elaborating the simplistic account for desires, and addressing how believed propositions may be both true and false.

In conclusion, we have explored the requirements on neural processes that arise when embodied cognitive systems are endowed with intentional states of the two directions of fit and the six psychological modes that provide a foundation for intentionality.

Acknowledgments

Support by the Deutsche Forschungsgemeinschaft (SPP Active Self, SCH 336/12-1) and by the Studienstiftung des Deutschen Volkes is gratefully acknowledged.

References

Anderson, J. R. (1996). Act: A simple theory of complex cognition. *American Psychologist*, *51*(4), 355.

Berger, M., Faubel, C., Norman, J., Hock, H., & Schöner, G. (2012). *The counter-change model of motion perception: An account based on dynamic field theory* (Vol. 7552 LNCS).

Carpenter, G. A., & Grossberg, S. (2016). *Adaptive resonance theory*. Springer.

Erlhagen, W., Bastian, A., Jancke, D., Riehle, A., & Schöner, G. (1999). The distribution of neuronal population activation (DPA) as a tool to study interaction and integration in cortical representations. *Journal of Neuroscience Methods*, *94*(1), 53–66.

Erlhagen, W., & Schöner, G. (2002). Dynamic field theory of movement preparation. *Psychological Review*, *109*(3), 545–572.

Grieben, R., Tekülve, J., Zibner, S. K. U., Schneegans, S., & Schöner, G. (2018, July). Sequences of discrete attentional shifts emerge from a neural dynamic architecture for conjunctive visual search that operates in continuous time. In J. Z. Chuck Kalish Martina Rau & T. Rogers (Eds.), *Cogsci 2018* (pp. 427–432).

Kazerounian, S., & Grossberg, S. (2014). Real-time learning of predictive recognition categories that chunk sequences of items stored in working memory. *Frontiers in Psychology*, *5*, 1–28.

Richter, M., Sandamirskaya, Y., & Schöner, G. (2012). A robotic architecture for action selection and behavioral organization inspired by human cognition. In *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on* (pp. 2457–2464).

Sandamirskaya, Y., & Schöner, G. (2010). An embodied account of serial order: How instabilities drive sequence generation. *Neural Networks*, *23*(10), 1164–1179.

Schöner, G., Spencer, J. P., & DFT Research Group, T. (2015). *Dynamic thinking: A primer on dynamic field theory*. Oxford University Press.

Schrodt, F., & Butz, M. V. (2016). Just imagine! learning to emulate and infer actions with a stochastic generative architecture. *Frontiers in Robotics and AI*, *3*, 5.

Schrodt, F., Kneissler, J., Ehrenfeld, S., & Butz, M. V. (2017). Mario becomes cognitive. *Topics in cognitive science*, *9*(2), 343–373.

Searle, J. R. (1980). The intentionality of intention and action*. *Cognitive Science*, *4*(1), 47–70.

Tekülve, J., & Schöner, G. (2019). Autonomously learning beliefs is facilitated by a neural dynamic network driving an intentional agent. In *Ieee conference on development and learning and epigenetic robotics (icdl-epirob 2019)*.

Wijeakumar, S., Ambrose, J. P., Spencer, J. P., & Curtu, R. (2017). Model-based functional neuroimaging using dynamic neural fields: An integrative cognitive neuroscience approach. *Journal of Mathematical Psychology*, *76*, 212–235.

Zibner, S. K. U., & Faubel, C. (2015). Dynamic scene representations and autonomous robotics. In *Dynamic thinking: A primer on dynamic field theory* (pp. 227–246). Oxford University Press.

Zibner, S. K. U., Tekülve, J., & Schöner, G. (2015). The neural dynamics of goal-directed arm movements: a developmental perspective. In *Ieee conference on development and learning and epigenetic robotics (icdl-epirob 2015)* (pp. 154–161).

The Intentional Stance Toward Robots: Conceptual and Methodological Considerations

Sam Thellman (sam.thellman@liu.se)
Department of Computer & Information Science, Linköping University
Linköping, Sweden

Tom Ziemke (tom.ziemke@liu.se)
Department of Computer & Information Science, Linköping University
Linköping, Sweden

Abstract

It is well known that people tend to anthropomorphize in interpretations and explanations of the behavior of robots and other interactive artifacts. Scientific discussions of this phenomenon tend to confuse the overlapping notions of folk psychology, theory of mind, and the intentional stance. We provide a clarification of the terminology, outline different research questions, and propose a methodology for making progress in studying the intentional stance toward robots empirically.

Keywords: human-robot interaction; social cognition; intentional stance; theory of mind; folk psychology; false-belief task

Introduction

The use of folk psychology in interpersonal interactions has been described as "practically indispensable" (Dennett, 1989, p. 342), and its predictive power has been proclaimed to be "beyond rational dispute" (Fodor, 1987, p. 6). The emergence of interactive technologies, such as computers and robots, has sparked interest in the role of folk psychology in human interactions with these systems. For example, John McCarthy stated: "It is perhaps never logically required even for humans, but expressing reasonably briefly what is actually known about the state of a machine in a particular situation may require ascribing mental qualities or qualities isomorphic to them" (McCarthy, 1979, p. 2). The usefulness of folk psychology however does not extend to interaction with all artifacts, and it does not necessarily extend to all kinds of interactions with robots. Although the prevalence of folk-psychological interpretation of robot behavior might be considered as being beyond dispute, its predictive power – i.e., the *usefulness* of taking the intentional stance toward robots – remains largely unassessed.

Researchers from diverse fields have explored people's folk-psychological theories about emerging robotic technologies, such as humanoid robots and autonomous vehicles. For example, Krach et al. (2008) and Chaminade et al. (2012) explored the neural activity of persons engaged in interactive games with robots. Waytz et al. (2014) showed that people's ascriptions of mental states to an autonomous vehicle affected their willingness to trust it. Thellman et al. (2017), Petrovych et al. (2018), and de Graaf and Malle (2018, 2019) investigated whether people judge distinctively human behaviors as intentional when exhibited by robots. Terada et al. (2007) asked people directly about whether they adopted the intentional stance toward a robot. Marchesi et al. (2018) developed a questionnaire-based method specifically for assessing whether people adopt the intentional stance toward robots. These studies all provide insight into people's folk-psychological theories about robots. However, none of them assessed how such theories affect people's predictions of behavior to shed light on the usefulness of taking the intentional stance in interactions with robots.

Moreover, research that has so far explicitly addressed the intentional stance toward robots in many cases conflated the intentional stance with overlapping but different notions, such as folk psychology and theory of mind. In particular, the question whether it is useful for people to predict robot behavior by attributing it to mental states (what we in the present paper will call "the intentional stance question") tends to be confounded with whether robots have minds ("the reality question"), whether people think that robots have minds ("the belief question"), and what kinds of mental states people ascribe to robots ("the attribution question"). For example, Chaminade et al. (2012, p. 8) claimed that participants in their experiments did not adopt the intentional stance when interacting with a robot as opposed to a person based on having "[manipulated] participants' belief about the intentional nature of their opponent" (thereby confounding the attribution question with the belief question). Wykowska et al. (2015, p. 768) stated that "it seems indeed very important to know whether the observed entity is an agent with a mind, and thus, whether the entity's behavior provides some social meaningful content" (confounding the attribution question with the reality question). Wiese et al. (2012, p. 2) stated that "adopting the intentional stance is based on a decision as to whether or not an observed agent is capable of having intentions" (confounding the intentional stance question with the belief question).

In view of these confusions, we aim to provide a clarification of the terminology and different research questions related to the folk psychology about robots in general and the intentional stance toward robots in particular. We also discuss in more detail how (not) to approach research questions specifically targeted at the intentional stance toward robots.

Basic Terminology

We here review Griffin and Baron-Cohen's (2002) distinction between *folk psychology*, *theory of mind*, and *the intentional*

stance and relate these overlapping but different notions to the literature surrounding the role of folk psychology in interactions with robots.

Folk psychology about robots

The notion of folk psychology (also known as belief-desire psychology, naïve or intuitive psychology, or commonsense psychology) broadly encompasses all mind-related theories that people have about themselves and others (Griffin & Baron-Cohen, 2002). This includes views about intentional, content-bearing, representational states (beliefs, desires, intentions, hunches, etc.) as well as phenomenal states (e.g., undirected anxieties, feelings and pain), traits, dispositions, and empirical generalizations such as that *people who are tired are generally irritable*, or – as in the context of folk psychology about robots – the cultural platitude that *robots do not have minds* (Fiala et al., 2014).

Research on people's folk-psychological theories about robots in general (as opposed to specific robots) has been pursued in part because of the societal (e.g., political, legal, or ethical) consequences that such theories might have. For example, European citizens' views on and acceptance of emerging robotic technologies, and their use in different areas of society, have been monitored in extensive surveys by the European Commission (2012, 2015). Ethically motivated research has targeted robot abuse, killer robots, robots in elderly care, child-robot interaction, and sex robots (for an overview, see Lin et al., 2014).

Theory of (robot) mind

Theory of mind refers more narrowly to the ability to attribute the behavior of *specific* others or oneself to underlying mental states, in particular intentional states, such as beliefs and desires, that are perceived to have a causal role in behavior (Griffin & Baron-Cohen, 2002).

People's views about the mental attributes of specific robots are frequently probed for the purpose of evaluating human-robot interactions. Examples of such measures are the Godspeed Questionnaire Series (Bartneck et al., 2009) and the Robotic Social Attributes scale (Carpinella et al., 2017). To the best of our knowledge, these measures have so far not been used in conjunction with measures of people's ability to predict the behavior of specific robots in the context of human-robot interaction research.

The intentional stance toward robots

The intentional stance refers to the *use* of intentional constructs (the beliefs, desires, intentions, etc., that are part of people's folk-psychological theories) as an interpretative strategy or framework to predict the behavior of specific others (Griffin & Baron-Cohen, 2002)[1]. The intentional stance is sometimes mistakenly equated with folk psychology. Dennett (1991) describes the intentional stance as "the craft" of folk psychology and distinguishes it from "the theory" itself. The intentional stance concerns what people *do* with folk psychology (i.e., predict and explain behavior using intentional constructs); folk psychology, in Dennett's view, refers to how we talk about what we do.

Although there seems to be a general consensus in the literature concerning the meaning of "intentionality" as denoting the distinguishing characteristic of certain mental phenomena of being "about" or "directed at" something as an object (Brentano, 1874/2012), some authors have treated it as a biological property (e.g., Searle, 1980; Varela, 1997; Ziemke, 2016) whereas others have refrained from doing so (e.g., Dennett, 1989; McCarthy, 1979). It is also important to recognize that intentionality is a separate notion from having certain intentions. Intentionality is a property of a specific set of mental states, namely intentional mental states. This set includes intentions, but also beliefs, desires, hopes, fears, hunches, and so on. Searle (2008, pp. 85–86) noted that the English translation of the German words for intentionality and intention, "Intentionalität" and "Absicht", are confusingly similar, stating that "we have to keep in mind that in English intending is just one form of intentionality among many".

In some cases, adopting the intentional stance toward an object is a useful strategy for predicting its behavior; in other cases, it is not. Dennett introduced the notion of an *intentional system* to denote objects that are "usefully and voluminously predictable from the intentional stance" (Dennett, 2009, p. 339). Humans are the most obvious example of intentional systems because human behavior is generally successfully predicted from the intentional stance but not from other modes of interpretation. The label "intentional system" is not restricted to humans, but it also does not extend to all non-human objects. Although a person might predict that a thermostat will raise the room temperature in the morning because it *wants* to keep it at 73 degrees and *knows* that it has fallen during the night, the use of such folk-psychological interpretations does not add predictive value *above and beyond* the corresponding non-psychological interpretation. In the words of John McCarthy (1979, p. 11), "ascribing beliefs to simple thermostats is unnecessary for the study of thermostats, because their operation can be well understood without it". In contrast, the moves of a chess-playing computer are, according to Dennett (1971), practically inaccessible to prediction from any other interpretative mode than the intentional stance.

It is reasonable to conjecture, given the complex behavior and social situatedness (Lindblom & Ziemke, 2003) of emerging robotic technologies, that taking the intentional stance might turn out to be crucial in many cases of human-robot interaction (Hellström & Bensch, 2018; Schellen & Wykowska, 2019; Thill & Ziemke, 2017; Vernon et al.,

[1] As noted by Griffin and Baron-Cohen (2002), the intentional stance theory (also known as intentional systems theory; Dennett, 2009) is both Dennett's take on the role of folk psychology in social interactions and on what intentional states really are. These two components can be considered separately (as in this paper); for example, one might agree with Dennett's claims about the role of the intentional stance in social interaction without subscribing to his views about the reality of ascribed mental states.

2016). However, although there is a growing body of evidence that people take the intentional stance toward robots, the usefulness of doing so remains largely unassessed. Hence, the central question in the context of the intentional stance toward robots is the extent to which the behavior of robots is usefully predicted from the intentional stance. The usefulness of the intentional stance toward robots presumably depends on a number of unknown factors, possibly related to the person interacting with the robot, the interaction context, and the robot in question. Answers to the intentional stance question might thus range from "the intentional stance is a practically dispensable mode of interpretation for predicting robot behavior" (cf. thermostat) to "the intentional stance is practically indispensable for predicting robot behavior" (cf. chess-playing computer), depending on these factors. Research into the usefulness of taking the intentional stance toward robots may also reveal unique social cognitive challenges associated with taking the intentional stance specifically toward robots (e.g., compare inferring what a robot vs. a person can perceive in a given situation in order to predict it's behavior), some of which may be universally present in human-robot interactions.

Four Distinct Research Questions

We have attempted to clarify some of the basic terminology surrounding the intentional stance toward robots. We also identified the central question about the intentional stance toward robots as concerning its usefulness for predicting robot behavior. We now move on to distinguish this question from three overlapping but separate research questions that appear frequently in the literature surrounding the intentional stance toward robots.

The reality question: Do robots have minds?

Questions such as "Do robots have minds?" and "Can machines think?" concern the nature or reality of the mental states of robots and other machines. We here collectively refer to such questions as different formulations of *the reality question*. The reality question is clearly independent from people's beliefs about it, and presumably also from people's disposition to predict and explain robot behavior based on mental state ascriptions (and the potential usefulness of doing so). While it seems plausible that ontological "discoveries" about the minds of robots may have a significant impact on how people relate to and interact with robots, there is no apparent reason to believe that they would affect people's predictions of robot behavior in interactions. What matters for the purpose of predicting behavior, it seems, is how people conceptualize behavior, and not the correspondence of those conceptualizations to reality. For example, Heider (1958, p. 5) noted: "If a person believes that the lines in his palm foretell his future, this belief must be taken into account in explaining certain of his actions". Hence, the reality question is conceptually distinct from questions regarding people's attributions and beliefs about the mental states of robots.

The belief question: Do people think that robots have minds?

People's views on the reality of the mental states of robots are part of folk psychology. As stated in the previous section, it is difficult to foresee how (if at all) such considerations affect people's predictions of robot behavior, regardless if they spring from collective scientific discovery or personal belief. Clearly, a person might attribute the behavior of a robot to mental states without necessarily committing to any ontological position about the reality of those mental states. Indeed, people commonly ascribe mental states to cartoon characters and animated geometric figures (Heider & Simmel, 1944). When, for example, we see Donald Duck angrily chasing chipmunks Chip and Dale because they are stealing his popcorn, we know that Donald, Chip, and Dale do not really have mental states, but we attribute their behavior to mental states nevertheless (Ziemke et al., 2015). As stated by Airenti (2018, p. 10), "anthropomorphism is independent of the beliefs that people may have about the nature and features of the entities that are anthropomorphized". There is to our knowledge no evidence that people's beliefs about the reality of the mental states of robots – or of cartoon characters, thermostats, or fellow humans – affect their disposition or ability to predict behavior. It does not seem to matter, for the purpose of predicting the behavior of an agent, whether the person interpreting the behavior of the agent in question believes that the agent *really* has mental states. *The belief question*, therefore, must be treated as distinct from questions concerning people's ascriptions of mental states to robots as well as the reality question.

The attribution question: What kinds of mental states do people ascribe to robots?

There is now an abundance of evidence that people commonly predict and explain the behavior of robots based on attributing it to underlying intentional states. The assumption that they do is arguably even built into many of the methods that are used to evaluate social human-robot interactions, whereby researchers explicitly ask people to evaluate mental properties of robots. The if-question in "Do people take the intentional stance toward robots?" has thus already been answered in the affirmative. Considerably less is known about what we for the present purposes call *the attribution question*, namely what kinds of mental states people ascribe to robots. The lack of knowledge about the attribution question does not stem from a lack of research effort but, at least in part, from issues in the methodology adopted to tackle the attribution question.

There is so far little agreement about what kinds of mental states people ascribe to robots. Gray, Gray and Wegner (2007) found that people tend to attribute the behavior of robots to mental states related to agency (e.g., memory, planning, and thought) but not subjective experience (e.g., fear, pain, and pleasure). Sytsma and Machery (2010) found, in contrast, that people refrain from attributing subjective states

that have hedonic value for the subject, that is, valenced states (e.g., feeling pain and anger) as opposed to unvalenced states (e.g., smelling a banana or seeing red). Buckwalter and Phelan (2013) further showed that people's tendency to attribute (or not) experiential or valenced states depends on the described function of the robot. Fiala et al. (2014) found that respondents in their experiments – when allowed to choose between different ways of describing the capabilities of a robot (e.g., the robot "identified the location of the box" vs. "knew the location of the box") – preferred not to attribute mental states at all. The authors noted that responses to questions about the mental states of robots are influenced by a wide variety of factors, including the apparent function of the robot, the way in which the question is asked, and cultural platitudes about robots.

In sum, it seems problematic to identify what kinds of mental states people ascribe to robots by asking them directly. Part of the problem, we believe, is that such questions are ambiguously open to interpretation as regarding the reality of the mental states of robots. As pointed out previously, people tend to predict and explain robot behavior with reference to mental states without reflecting on the reality of those states. Thus, when asked directly, a person might deny that a robot has a mind, despite having previously attributed mind to it upon being asked to describe its behavior (Fussell et al., 2008).

The intentional stance question: Is it useful for people to predict robot behavior by attributing it to mental states?

The usefulness of predicting robot behavior by attributing it to mental states is not a pre-given. *The intentional stance question* is therefore distinct from the attribution question. The ability to predict behavior based on the intentional stance is also, as evidenced by studies on mental state attribution from Heider and Simmel (1944) and onwards, independent from the reality of the attributed mental states and from people's beliefs about them.

Although the prevalence of people taking the intentional stance toward robots might be considered as beyond dispute, its predictive power – that is, the usefulness of doing so – remains largely unassessed. Hence, the central question in the context of the intentional stance toward robots is to what extent the behavior of robots is usefully predicted from the intentional stance. Other questions of potential interest concern causes of predictive (mis)judgment from the intentional stance toward robots, how misjudgment can be reduced, and potential effects of taking the intentional stance toward robots on human cognition (e.g., cognitive load).

Measures of the Intentional Stance

If one wants to investigate whether the intentional stance is useful as an interpretative framework for predicting robot behavior, then one must, at the very least, measure people's predictions of behavior and ensure that those predictions stem from specific attributed mental states. Very few previous studies concerned with the intentional stance toward robots employed such measures (one exception is Sciutti et al., 2013). In this section, we review established experimental paradigms in interpersonal psychology that accomplish measuring effects of mental state attribution on behavior prediction, namely explicit and implicit false-belief tasks and anticipatory gaze tasks.

Explicit measures

The standard false-belief task (sometimes referred to as the "Sally–Anne test" or the location-change false-belief test) was outlined by Dennett (1978) in a commentary to Premack and Woodruff's seminal paper "Does the chimpanzee have a theory of mind?". This was later turned into an experimental paradigm in which a human study participant must attribute a false belief to an agent in order to predict its behavior (Wimmer & Perner, 1983). In the experiment, the participant is made aware that an agent observes a certain state-of-affairs x. Then, in the absence of the agent the participant witnesses an unexpected change in the state-of-affairs from x to y. The participant now knows that y is the case and also knows that the agent still (falsely) believes that x is the case. After this, the participant is asked to predict how the agent will behave in some circumstance, given its false belief about the state-of-affairs. If the participant fails to predict the behavior of the agent, this can be directly attributed to a failure of the participant to ascribe a false belief to the agent. Frith and Frith (1999, p. 1692) commented on the strength of the false-belief task: "To predict what a person will do on the basis of a true belief is not a sufficiently stringent test [of the ability to take the intentional stance], since here the belief coincides with reality, and it's hard to tell whether the action is governed by physical reality or mental state. In everyday life, beliefs rather than reality determine what people do, and false beliefs play an important role".

False-belief tasks have primarily been used to test for the possession of a theory of mind (e.g., Baron-Cohen, Leslie & Frith, 1985). However, they can also be used to explore the relative difficulty of reasoning about others' beliefs (Bloom & German, 2000). We argue that the false-belief task is a suitable paradigm for assessing the usefulness of the intentional stance toward robots because it enables measuring the extent to which a person's mental state ascriptions to a specific robot are conducive to predicting its behavior. Hence, false-belief tasks would be used in the context of human-robot interaction studies not to test for a person's *possession* of a theory of a specific robot's mind but for the successful or unsuccessful *use* of such theories in interactions with robots.

Concerns have been raised previously in the theory of mind literature about whether the explicit formulation of false-belief questions might impute folk-psychological theory to the task participant or affect his or her disposition to ascribe mental states. In some false-belief experiments, participants were asked questions, such as "Where does the agent *believe/think* that the object is now?", which explicitly suggest that the agent possesses beliefs or thoughts. Other experi-

ments used questions such as "Where will the agent look for the object?" which implicitly suggest the possession of beliefs or thoughts. However, an extensive meta-study of theory of mind research on children showed that the type of question (e.g., explicit vs. implicit statements of belief) provided to participants did not significantly affect participants' success in the false-belief task (Wellman, Cross & Watson, 2001). This finding can be taken as supporting the view expressed by Dennett that "whether one calls what one ascribes to the computer beliefs or belief-analogues or information complexes or Intentional whatnots makes no difference to the nature of the calculation one makes on the basis of the ascription" (Dennett, 1971, p. 91). Regardless of this meta-analytic finding, researchers concerned with the risk of imputing folk-psychological theories about robots to study participants, in the context of false-belief tasks, can employ implicit intentional stance measures. In the following section we review two such measures: implicit false-belief tasks and goal-directed anticipatory gaze tasks.

Implicit measures

Implicit false-belief tasks employ non-verbal measures to assess people's behavior predictions (for an overview, see Schneider & Slaughter, 2015). Using implicit measures, the intentional stance can be investigated by recording anticipatory gaze behavior (Clements & Perner, 1994) or reaction times (Kovács, Téglás & Endress, 2010), even without instructions to predict behavior or providing questions about the mental states of agents (Kovács, Téglás & Endress, 2010; Schneider et al., 2012). Implicit measures also provide an opportunity to investigate the potential effort involved in tracking the beliefs of robots whose sensory perspectives significantly differ from the human case.

Goal-directed anticipatory gaze tasks represent another way to measure the intentional stance toward robots. Using an anticipatory gaze paradigm, Sciutti et al. (2013) showed that people shift their gaze toward perceived "goals states" of robot actions prior to the execution of the actions themselves. One limitation of this paradigm is that it is not always possible to infer which gaze behaviors are anticipatory gazes (and therefore reflect goal ascriptions) and which are not. As such, goal-directed anticipatory gaze measures might not be as strong a measure of the intentional stance as false-belief tasks. Nevertheless, studying goal ascription through anticipatory gaze measures might be suitable as a complement to studying belief ascription using false-belief tasks.

Conclusion

We have attempted to clarify the difference between three overlapping concepts that are used (in many cases confusedly) in the literature surrounding the intentional stance toward robots: folk psychology, theory of mind, and the intentional stance. The central question in research on the intentional stance toward robots was identified as the extent to which the intentional stance is a useful (and potentially even indispensable) interpretative strategy or framework for predicting behavior in interactions with robots. We argued that this question is distinct from questions regarding the reality of the mental states of robots, people's beliefs about the mental states of robots, and what kinds of mental states people ascribe to robots. We also established a "methodological criterion" for investigating the usefulness of the intentional stance toward robots: the measurement of people's predictions of robot behavior and reliable inference that those predictions stem from specific attributed mental states. Last, but not least, we identified explicit and implicit false-belief tasks and anticipatory gaze tasks as fulfilling these criteria, thereby constituting a promising experimental paradigm for future empirical investigations of the intentional stance toward robots.

The ability to infer the intentional states (beliefs, desires, etc.) of robots is presumably in many cases crucial to the successful prediction of robot behavior and, consequently, to well-functioning and socially acceptable human-robot interaction (Hellström & Bensch, 2018; Schellen & Wykowska, 2019; Thill & Ziemke, 2017; Vernon et al., 2016). However, continuously tracking changes in the intentional states of robots as interactions unfold represents a potentially difficult and demanding challenge to humans: robots have different "perspectives" on or sensorimotor couplings with the world than humans. Consider the task of simultaneously navigating interactions with three different types of robots in a crowded environment (e.g., a busy street): the first robot can detect objects behind humanly opaque structures such as walls, vehicles, or humans; the second robot cannot see through glass; and the third robot is sensory-equivalent to most humans. How do humans fare in an interaction scenario like this? We propose that taking the intentional stance toward robots must in some cases be more difficult (in terms of predictive accuracy) and demanding (e.g., in terms of cognitive load; Sweller, 1988) than taking the intentional stance toward humans, and view this as a hypothesis worthwhile exploring in the context of human-robot interaction research. In particular, we speculate that people employ a reasoning heuristic which can be described as "anthropocentric anchoring and adjustment", consistent with the accounts in Epley et al. (2004) and Nickerson (1999) but where people adopt the perspective of specific robots by serially adjusting from their own (human) perspective.

Another question relevant to the intentional stance toward robots is the extent to which its usefulness or predictive power can be improved by providing information about the capabilities and limitations of robots prior to interactions. People base their estimations of the knowledge of robots partly on their assumptions about people (Kiesler, 2005). People's knowledge estimations of robots have been shown to be affected by the physical attributes of robots (Powers & Kiesler, 2006) and information about the robot given beforehand, such as robot gender (Powers et al., 2005) or country of origin (Lee et al., 2005). However, it has to our knowledge not yet been investigated whether providing information or manipulating

social cues can improve the accuracy with which people predict the behavior of a robot. Would prior knowledge about the sensory capabilities of the three types of robots in the example above help people interact with them? This is another question worthwhile exploring in studying the intentional stance toward robots.

We believe that cognitive science has important contributions to make in the continued exploration of the role of folk psychology in human interaction with robots, especially in the development of appropriate methodological approaches to investigating the intentional stance toward robots. As suggested in this paper and elsewhere, the intentional stance can be a confusing concept (Griffin & Baron-Cohen, 2002) and a difficult phenomenon to measure, perhaps especially in the context of interactions with robots (Schellen & Wykowska, 2019). In the folk psychology about robots, robots might not have *real* minds but have *attributed* minds nevertheless, and as for the science of mind, the jury is still out regarding the extent to which mind possession and mind attribution go hand-in-hand in the case of robots (cf. Dennett, 1989; Fodor, 1987; Searle, 1980). We therefore hope that the conceptual clarifications and methodological proposals presented here will pave the way for fruitful research on the intentional stance toward robots.

Acknowledgments

The authors would like to thank Fredrik Stjernberg, Robert Johansson, and members of the Cognition & Interaction Lab at Linköping University for valuable input on the ideas presented in this paper.

References

Airenti, G. (2018). The development of anthropomorphism in interaction: Intersubjectivity, imagination and theory of mind. *Frontiers in Psychology*, *9*, 2136.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a theory of mind? *Cognition*, *21*(1), 37–46.

Bartneck, C., Kulić, D., Croft, E., & Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *Int. J. Soc. Robot.*, *1*(1), 71–81.

Bloom, P., & German, T. P. (2000). Two reasons to abandon the false belief task as a test of theory of mind. *Cognition*, *77*(1), B25–B31.

Brentano, F. (2014). *Psychology from an empirical standpoint*. Routledge. (Original work published 1874).

Buckwalter, W., & Phelan, M. (2013). Function and feeling machines: a defense of the philosophical conception of subjective experience. *Philos. Stud.*, *166*(2), 349–361.

Carpinella, C. M., Wyman, A. B., Perez, M. A., & Stroessner, S. J. (2017). The robotic social attributes scale (rosas): Development and validation. In *Proc. 2017 ACM/IEEE Int. Conf. on Human-Robot Interaction* (pp. 254–262).

Chaminade, T., Rosset, D., Da Fonseca, D., Nazarian, B., Lutscher, E., Cheng, G., & Deruelle, C. (2012). How do we think machines think? An fMRI study of alleged competition with an artificial intelligence. *Frontiers in Human Neuroscience*, *6*, 103.

Clements, W. A., & Perner, J. (1994). Implicit understanding of belief. *Cognitive Development*, *9*(4), 377–395.

de Graaf, M., & Malle, B. F. (2018). People's judgments of human and robot behaviors: A robust set of behaviors and some discrepancies. In *Comp. 2018 ACM/IEEE Int. Conf. on Human-Robot Interaction* (pp. 97–98).

de Graaf, M. M., & Malle, B. F. (2019). People's explanations of robot behavior subtly reveal mental state inferences. In *Proc. 2019 ACM/IEEE Int. Conf. on Human-Robot Interaction* (pp. 239–248).

Dennett, D. C. (1971). Intentional systems. *The Journal of Philosophy*, *68*(4), 87–106.

Dennett, D. C. (1978). Beliefs about beliefs [P&W, SR&B]. *Behavioral and Brain Sciences*, *1*(4), 568–570.

Dennett, D. C. (1989). *The intentional stance*. MIT press.

Dennett, D. C. (1991). Two contrasts: folk craft versus folk science, and belief versus opinion. In J. D. Greenwood (Ed.), *The future of folk psychology: Intentionality and cognitive science* (pp. 135–148). Cambridge University Press.

Dennett, D. C. (2009). Intentional systems theory. *The Oxford handbook of philosophy of mind*, 339–350.

Epley, N., Keysar, B., Van Boven, L., & Gilovich, T. (2004). Perspective taking as egocentric anchoring and adjustment. *J. Pers. Soc. Psychol.*, *87*(3), 327.

European Commission. (2012). *Special eurobarometer 382: Public attitudes towards robots.*

European Commission. (2015). *Special eurobarometer 427: Autonomous systems.*

Fiala, B., Arico, A., & Nichols, S. (2014). You, robot. In E. Machery & E. O'Neill (Eds.), *Current controversies in experimental philosophy.* Routledge.

Fodor, J. A. (1987). *Psychosemantics: The problem of meaning in the philosophy of mind* (Vol. 2). MIT press.

Frith, C. D., & Frith, U. (1999). Interacting minds–a biological basis. *Science*, *286*(5445), 1692–1695.

Fussell, S. R., Kiesler, S., Setlock, L. D., & Yew, V. (2008). How people anthropomorphize robots. In *2008 ACM/IEEE Int. Conf. on Human-Robot Interaction* (pp. 145–152).

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science*, *315*(5812), 619–619.

Griffin, R., & Baron-Cohen, S. (2002). The intentional stance: Developmental and neurocognitive perspectives. In A. Brook & D. Ross (Eds.), *Daniel Dennett* (pp. 83–116). Cambridge University Press.

Heider, F. (1958). *The psychology of interpersonal relations*. New York: Whiley.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *Am. J. Psychol.*, *57*(2), 243–259.

Hellström, T., & Bensch, S. (2018). Understandable robots-what, why, and how. *Paladyn*, *9*(1), 110–123.

Kiesler, S. (2005). Fostering common ground in human-

robot interaction. In *2005 IEEE Int. Workshop on robot and human interactive communication* (pp. 729–734).

Kovács, Á. M., Téglás, E., & Endress, A. D. (2010). The social sense: Susceptibility to others' beliefs in human infants and adults. *Science*, *330*(6012), 1830–1834.

Krach, S., Hegel, F., Wrede, B., Sagerer, G., Binkofski, F., & Kircher, T. (2008). Can machines think? Interaction and perspective taking with robots investigated via fMRI. *PloS one*, *3*(7), e2597.

Lee, S.-l., Lau, I. Y.-m., Kiesler, S., & Chiu, C.-Y. (2005). Human mental models of humanoid robots. In *Proc. 2005 IEEE Int. Conf. Robot. Autom.* (pp. 2767–2772).

Lin, P., Abney, K., & Bekey, G. A. (2014). *Robot ethics: the ethical and social implications of robotics*. The MIT Press.

Lindblom, J., & Ziemke, T. (2003). Social situatedness of natural and artificial intelligence: Vygotsky and beyond. *Adaptive Behavior*, *11*(2), 79–96.

Marchesi, S., Ghiglino, D., Ciardo, F., Perez-Osorio, J., Baykara, E., & Wykowska, A. (2019). Do we adopt the intentional stance toward humanoid robots? *Frontiers in Psychology*, *10*.

McCarthy, J. (1979). Ascribing mental qualities to machines. In M. Ringle (Ed.), *Philosophical perspectives in artificial intelligence.* Humanities Press.

Nickerson, R. S. (1999). How we know – and sometimes misjudge – what others know: Imputing one's own knowledge to others. *Psychol. Bull.*, *125*(6), 737–759.

Petrovych, V., Thellman, S., & Ziemke, T. (2018). Human interpretation of goal-directed autonomous car behavior. In *Proc. 40th Annual Cognitive Science Society Meeting* (pp. 2235–2240). Madison, WI.

Powers, A., & Kiesler, S. (2006). The advisor robot: Tracing people's mental model from a robot's physical attributes. In *Proc. 1st ACM SIGCHI/SIGART Conf. on Human-Robot Interaction* (pp. 218–225).

Powers, A., Kramer, A. D., Lim, S., Kuo, J., Lee, S.-l., & Kiesler, S. (2005). Eliciting information from people with a gendered humanoid robot. In *2005 IEEE Int. Workshop on Robot and Human Interactive Communication* (pp. 158–163).

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, *1*(4), 515–526.

Schellen, E., & Wykowska, A. (2019). Intentional mindset toward robots – open questions and methodological challenges. *Frontiers in Robotics and AI*, *5*, 139.

Schneider, D., Bayliss, A. P., Becker, S. I., & Dux, P. E. (2012). Eye movements reveal sustained implicit processing of others' mental states. *J. Exp. Psychol. Gen.*, *141*(3), 433.

Schneider, D., Slaughter, V. P., & Dux, P. E. (2015). What do we know about implicit false-belief tracking? *Psychonomic Bulletin & Review*, *22*(1), 1–12.

Sciutti, A., Bisio, A., Nori, F., Metta, G., Fadiga, L., & Sandini, G. (2013). Robots can be perceived as goal-oriented agents. *Interaction Studies*, *14*(3), 329–350.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*(3), 417–424.

Searle, J. R. (2008). *Mind, language and society: Philosophy in the real world*. Basic books.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*(2), 257–285.

Sytsma, J., & Machery, E. (2010). Two conceptions of subjective experience. *Philos. Stud.*, *151*(2), 299–327.

Terada, K., Shamoto, T., Mei, H., & Ito, A. (2007). Reactive movements of non-humanoid robots cause intention attribution in humans. In *2007 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems* (pp. 3715–3720).

Thellman, S., Silvervarg, A., & Ziemke, T. (2017). Folk-psychological interpretation of human vs. humanoid robot behavior: Exploring the intentional stance toward robots. *Frontiers in Psychology*, *8*, 1962.

Thill, S., & Ziemke, T. (2017). The role of intentions in human-robot interaction. In *Proc. 2017 ACM/IEEE Int. Conf. on Human-Robot Interaction* (pp. 427–428).

Varela, F. J. (1997). Patterns of life: Intertwining identity and cognition. *Brain and cognition*, *34*(1), 72–87.

Vernon, D., Thill, S., & Ziemke, T. (2016). The role of intention in cognitive robotics. In A. Esposito & L. C. Jain (Eds.), *Toward Robotic Socially Believable Behaving Systems – Volume I* (pp. 15–27). Springer.

Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *J. Exp. Soc. Psychol.*, *52*, 113–117.

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development*, *72*(3), 655–684.

Wiese, E., Wykowska, A., Zwickel, J., & Müller, H. J. (2012). I see what you mean: How attentional selection is shaped by ascribing intentions to others. *PloS one*, *7*(9), e45391.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, *13*(1), 103–128.

Wykowska, A., Kajopoulos, J., Obando-Leitón, M., Chauhan, S. S., Cabibihan, J.-J., & Cheng, G. (2015). Humans are well tuned to detecting agents among non-agents: examining the sensitivity of human perception to behavioral characteristics of intentional systems. *Int. J. Soc. Robot.*, *7*(5), 767–781.

Ziemke, T. (2016). The body of knowledge: on the role of the living body in grounding embodied cognition. *Biosystems*, *148*, 4–11.

Ziemke, T., Thill, S., & Vernon, D. (2015). Embodiment is a double-edged sword in human-robot interaction: Ascribed vs. intrinsic intentionality. In *Proc. 10th ACM/IEEE Human Robot Interaction Conference* (pp. 1–2).

Articulatory features of phonemes pattern to iconic meanings: evidence from cross-linguistic ideophones

Arthur Lewis Thompson[1] (arthurlewisthompson@gmail.com)
Nicolas Collignon[2] (n.collignon@ed.ac.uk) **Youngah Do**[1] (youngah@hku.hk)
[1]Department of Linguistics, the University of Hong Kong, Pokfulam Road, Hong Kong
[2] School of Informatics, University of Edinburgh, 10 Crichton St, Edinburgh

Abstract

Iconic words are known to exhibit an imitative relationship between a word and its referent. Many studies have worked to pinpoint sound-to-meaning correspondences for ideophones from different languages. The correspondence patterns show similarities across languages, but what makes such language-specific correspondences universal, as iconicity claims to be, remains unclear. This could be due to a lack of consensus on how to describe and test the perceptuo-motor affordances that make an iconic word feel imitative to speakers. We created and analyzed a database of 1,888 ideophones across 13 languages, and found that 5 articulatory properties, physiologically accessible to all spoken language users, pattern according to semantic features of ideophones. Our findings pave the way for future research to utilize articulatory properties as a means to test and explain how iconicity is encoded in spoken language.

Keywords: iconicity; ideophones; systematicity; sound symbolism; phonology; semantics

Introduction

Iconicity in spoken language can be summed up as the relation of a linguistic form (or sound) to its meaning (Hinton, Nichols, & Ohala, 1994). One fundamental example is onomatopoeia, as in the English *woof woof* for the sound of a dog bark or *vroom vroom* for the reving a car engine. Sound mapping to meaning in an imitative way is also called sound symbolism. An implicit assumption underlying the term sound symbolism is that phonemes, or clusters of phonemes, map onto meaning below word or morpheme level thus acting as affordances which together allow the sound symbolic word to take on meaning. For example, the /ŋ/ in English /diŋ.doŋ/ seems to be characteristic of the reverberating echo of a bell tolling, while the alternating /i/ and /o/ seems characteristic of movement or a fluctuation in pitch as the bell tolls. While various studies have worked to elicit sub-phonemic sound-to-meaning correspondences (Aryani, 2018; Blasi, Wichmann, Hammarström, Stadler, & Christiansen, 2016; De Carolis, Marsico, & Coupé, 2017; Kawahara, Noto, & Kumagai, 2018; Shih, Ackerman, Hermalin, Inkelas, & Kavitskaya, 2018; Kwon & Round, 2015; Ofori, 2009; Hamano, 1998; Maduka, 1988; Oswalt, 1994; Akita, Imai, Saji, Kantartzis, & Kita, 2013; Ayalew, 2013; McCune, 1985; Assaneo, Nichols, & Trevisan, 2011; Strickland, J, Schlenker, & Geraci, 2017), the underlying mechanisms of such correspondences are unclear. To begin rectifying the issue of what exactly makes a sound symbolic word iconic, and so the field of iconicity can move toward a unified understanding of what affordances in the spoken modality should be classified as iconic, this paper attempts to reveal the gestural affordances underpinning imitative words, i.e., ideophones. Ideophones are marked words which depict sensory meaning and belong to an open lexical class (Dingemanse, 2012, in press). Recent studies have likened ideophones to oral gestures considering that they co-occur with other visual forms of communication so frequently in spontaneous speech (Dingemanse, 2015; Hatton, 2016; Mihas, 2013; Dingemanse, 2013; Nuckolls et al., 2000). This speaks to the importance of analyzing (articulatory) movement in order to understand how ideophones mean what they mean. Ideophones have been shown to be easily learnable by speakers from different language backgrounds, which may also speak to their imitative, gestural nature encoded despite language-specific differences such as phonotactics, phonological inventory, or lexical associations (Lockwood, Hagoort, & Dingemanse, 2016; Dingemanse, Schuerman, Reinisch, Tufvesson, & Mitterer, 2016; Iwasaki, Vinson, & Vigliocco, 2007a, 2007b). Thus, ideophones are an ideal testing ground for how articulatory properties pattern to meaning. Vocal imitations and onomatopoeia created spontaneously by participants in experimental settings have been shown to exhibit sound-meaning correspondences which can be attributed to patterns of oral articulation (Assaneo et al., 2011; Taitz et al., 2018). This leads us to our investigative focus on the articulatory gestures of consonants in imitative words. In a methodological vein similar to Blasi et al. (2016), this study looks at whether articulatory feature (e.g., plosive, fricative, nasal, velar, labial) is more or less found in certain semantic domain (e.g., telic events, human vocal sounds, motion, appearance) following cross-linguistic descriptions of ideophone meaning (Dingemanse, 2012; Hamano, 1998; Van Hoey, 2018; Nuckolls, Swanson, Sun, Rice, & Ludlow, 2017). However, unlike Blasi et al. (2016) who focused on identifying sound-to-meaning mappings in arbitrary words, this study focuses on words which are explicitly iconic in nature. If an oral articulation is more attested in one semantic domain of ideophones than another this could explain why some phonosemantic mappings might be perceived as imitative and therefore iconic of a given percept. Such mappings are therefore explainable as perceptuo-motor affordances grounded in gestural means, e.g., total closure of plosive articulation, affords the semantic category of telic events and their percept coming to an abrupt stop. We created a database of ideophones from 13 languages (in total, 1888 ideophones) to carry out our investigation on how articulatory properties of consonants pattern with ideophone meaning.

Background

Phonosemantics

Sub-phonemic sound-to-meaning mappings have been proposed for a number of languages (Maduka, 1988; Waugh, 1994; Hamano, 1998; Oswalt, 1994; Assaneo et al., 2011; Akita et al., 2013; Ayalew, 2013; Kwon & Round, 2015; Blasi et al., 2016). The general assumption is loosely encapsulated by a broad hypothesis that every phoneme is meaning-bearing, and that this meaning is rooted in its articulation (Diffloth, 1979, 1994; Hamano, 1998; Dingemanse, 2018). Though this study does not assume all phonemes to be meaning bearing in all contexts, we do subscribe to the notion that the meaning of a phonosemantic mapping for iconic word should be rooted in its articulation following previous studies (Diffloth, 1979, 1994; Oda, 2001; Assaneo et al., 2011; Taitz et al., 2018; Strickland et al., 2017).

Ideophone Database

Database

Currently there is no cross-linguistic database dedicated solely to ideophone inventories. We created a database of 13 languages[1] which were selected with the aim of being as typologically diverse as possible despite the limited number of linguistic descriptions for ideophone inventories in the world[2]). The languages are as follows with their number of ideophones in brackets: Manyika Shona (Niger-Congo) [112] , Uyghur (Turkic) [49], Manchu (Tungusic) [91], Chaoyang Southern Min (Sino-Tibetan) [248], Ma'ai Zhuang (Kra-Dai) [232], Kam (Kra-Dai) [223], Akan (Niger-Congo) [190] , Kisi (Niger-Congo) [98], Kuhane (Niger-Congo) [64], Pastaza Quichua (Quechuan) [283] , Upper Necaxa Totonac (Totozoquean) [146], Temne (Niger-Congo) [76] , Yakkha (Sino-Tibetan) [76]. Due to their depictive nature, and the various methods of elicitation, the ideophone inventory numbers reported above are not absolute, but instead reflect a general picture about the semantic "visibility" of ideophones per language. This is in line with a claim recently put forth by Dingemanse (Dingemanse, in press) that ideophones form an open class, speaking to the creative potential for newly coined ideophones.

Total number of ideophones was 1,888. Ideophones were entered into the database with their orthography (if available), International Phonetic Alphabet (IPA) transcription, and reported translation. A phonetically trained transcriber provided IPA transcription of words when original resources do not provide IPA transcriptions. To analyze the phonetic properties of words, the transcriber also provided with place (labial, coronal, dorsal, pharyngeal, laryngeal), manner (sonorant, continuant, nasal, lateral, delayed release), and laryngeal features (voice, spread glottis, constricted glottis) of consonants for each ideophone. An independent transcriber checked the validity of the transcriptions as well as featural descriptions of ideophones. Ideophones were then coded for semantic features following criteria below (Table 1).

Table 1: Semantic features

Semantic Feature	Description of [+] feature
[+/- animal]	vocalization made by animals
[+/- appearance]	depicts visual information, i.e., how something looks or degrees of visibility
[+/- friction]	depicts rubbing together or rough contact of surfaces (not necessarily active movement), i.e., grinding, rustling, sharpening, hacking up phlegm, tearing cloth
[+/- human]	vocalization made by people, i.e., laughter, crying, talking
[+/- loud]	auditory information of inherently high amplitude, i.e., explosion, screaming, shattering
[+/- motion]	depicts active (the act of X) movement, i.e., walking, chopping, splashing, sneaking, flapping, water boiling, bumping, spitting, firecrackers exploding
[+/- sound]	depicts auditory information (the sound of X)
[+/- telic]	depicts an event which reaches completion
[+/- wind]	depicts movement of air, i.e., blowing, coughing, gales

Semantic Features

Feature were created to correspond to Dingemanse (2012) implicational hierarchy of ideophones which lists the following semantic categories: sound <movement <visual patterns <other sensory perceptions <cognitive states. Additional categories were created based on observations of what ideophones depict cross-linguistically (Hamano, 1998; Hinton, Nichols, & Ohala, 2006; Van Hoey, 2018; Nuckolls et al., 2017). It is important to note that semantic features are not mutually exclusive. An ideophone may be coded for multiple. For example, the Chaoyang ideophone /hu.hu/ wind blowing was coded with [+sound] (because this ideophone depicts an auditory percept), [-telic] (because this ideophone does not involve a perceived endpoint of an event), [+wind] (because this ideophone involves a percept created by the movement of air), and [-motion] (because this ideophone is not depictive of an action plus its resulting sound or manner thereof). In total 18 features in Table 1 were considered.

Articulatory Features

We categorized place, manner and laryngeal features in our phonetic transcriptions of ideophones into 7 groups (see Table 2), based on how the articulators (lips, tongue) and airflow are

[1](Franck, 2014; Gerner, 2005; Beck, 2008; Schackow, 2016; Kanu, 2008; Childs, 1988; Ofori, 2009; Nuckolls et al., 2017; Xiao, 2015; Wang & Tang, 2014; Mathangwane & Ndana, 2014; ?, ?)

[2]Ma'ai Zhuang ideophones were collected during ongoing fieldwork.

Table 2: Articulatory features

Articulatory Feature	Description
[+labial]	usage of one or both lips
[+tongue resting]	tongue is in resting/neutral position and not raised
[+tongue root]	usage of back of tongue (dorsum)
[+airflow]	air is forced out through a narrow channel in the mouth, i.e., fricatives
[+velum] (nasal)	velum is lowered and air escapes through the nasal passage
[+oral contact]	lips and/or tongue make constriction
[+vocal folds]	vibration of the vocal folds

involved in their production[3].

While some features can be subsumed by another, such as [+labial] by [+oral contact], which is relevant to both the lips and the tongue, the decision to test a more specific feature is to keep distinctive place features that are frequently contrastive among ideophones; for example, in Chaoyang, we have [+labial] [+oral contact] ideophone /pu.pu/ meaning rapid movement and [-labial] [+oral contact] ideophones /tsu.tsu/ whispering. Likewise, in Pastaza Quichua we have [+labial] [+oral contact] /pɑw/ manner of being turned downward and [-labial] [+oral contact] /kw/ sound of stepping on dry leaves. We also include [+tongue root], which is specific to a particular part of an articulator given their frequent contrastiveness among ideophones. For example, in Akan we have [+tongue root] /kuu/ call of a large bird and [-tongue root] /tuu/ manner of hitting with the fist. The reason we created encompassing features, i.e., [+oral contact] or [+tongue resting], was so that general manner of the consonant is described regardless of place in the oral tract. There are 14 articulatory features when counting the negative counterparts of Table 2. If properties of iconicity are truly universal, then we predict that the universally accessible properties captured by our articulatory features should bear the explanatory power for what perceptuo-motor affordances underpin iconicity and its notions of (analogical) depiction.

We coded semantic and articulatory features in Table 1 and 2 for all consonants found in the ideophones in our database using binary features. Our coding did not consider syllable structures or semantic hierarchy of ideophones, treating all consonants in all ideophones equally. We measured the ratio of each semantic feature given each articulatory feature to see if a semantic feature is more likely to co-occur specifically with certain articulatory features. We used chi-squared tests to see if the distributions of a binary articulatory feature differ between two samples: ideophones that have the semantic feature **X** and those that do not in a given language. The ratios then entered into a Wilcoxon's sign rank test.

[3][+airflow] applies to IPA symbols /ɸ β f v θ ð s z ʃʒ ʂ ʐ ç ʝ x ɣ χ ʁ ħ ʕ h ɦ/. [+tongue root] IPA = /j c ɟ k g ɲ ŋ ç ʝ x ɣ j ɯ ʎ ʟ/. [+labial] IPA = / p b m ʙ ɸ β ɱ v ʋ f ɓ w/. [+velum] IPA = /m ɱ n ɳ ŋ ɴ/. [+oral contact] IPA = /p b m ʙ ɸ β ɱ v ʋ ʋ f θ ð t d n r ɾ z ɮ ɹ l s ɬ ʃ ʒ ɖ ɳ ɽ ʐ ʂ ɻ c ɟ ɲ ç ʝ j ʎ k ŋ x ɣɯ ʟ ɓ ɗ ʄ w/. [+tongue resting] IPA = / t ħ ʕ p b m ʙ β ɸ ɱ v ʋ f ʋ/.

Hypotheses

We have made 4 preliminary hypotheses based on observations from the phonosemantic literature. These observations are grounded in perceptuo-motor analogy but have yet to be tested for ideophone inventories across languages. (1) Stop consonants, characterized by total occlusion of airflow, i.e., [-airflow], have been observed for ideophones indicating complete, i.e., [+telic], events or events with abrupt endings (Diffloth, 1979; Hinton et al., 2006; Hamano, 1998; Alpher, 1994; Taitz et al., 2018; Strickland et al., 2017). (2) Fricatives, i.e., [+airflow], have been associated to wind or friction between two objects (Diffloth, 1979; Oswalt, 1994; Hinton et al., 2006; Ofori, 2009; Taitz et al., 2018). Improvised vocal imitations have shown that (3) labial consonants, i.e., [+labial], are associated with the sounds resulting from motion, i.e., [+motion], (4) while dorsal consonants, i.e., [+tongue root], are associated with movement itself (Taitz et al., 2018) i.e., [+motion] in our feature set.

Analysis

We analysed every semantic feature (32) against every articulatory gesture feature (16), and removed all the pairs when there were two or more languages that did not have ideophones with either the semantic feature, or the articulatory gesture considered (n=213). In total, we analysed 299 pairs of semantic features and articulatory gestures. When evaluating the significance of a mapping between an articulatory gesture and a semantic feature, we compared it against all ideophones that did not have the semantic feature. To do this, we compared on an individual-language basis the distributions of a binary articulatory feature in two samples: ideophones that have the semantic feature **X** (e.g. "human vocal" vs "not human vocal") and those that do not in a given language. We are thus testing for differences in distributions of a binary feature. We first compare these ratios for each individual language. Because the sample sizes were quite low in some comparisons (5-6 samples for e.g. the "human vocal" category in some languages), we use the chi-squared test. To test whether mappings of articulatory gesture to meaning were consistent across languages, we pooled the data from across languages. We decided not to run tests on all the agreggated ideophones across languages, as this would have been biased by the differences in sample sizes across languages. Instead, we compared the ratios from the previous analysis done at the individual language level. There were many maximum ratios (i.e. languages where all words have a given feature). We thus used a wilcoxon's sign rank test, since it is a non-parametric test and does not assume the data to be normal like e.g. a paired Student's t-test.

Table 3: Articulatory feature to semantic feature mappings significant across 13 languages.

#	Articulatory Feature	Semantic Feature	Correlation across all languages (Wilcoxon's test p-values)	Number of individual languages,with significant chi-squared ($p<0.05$)
1	-airflow	+telic	0.0015	7
2	+airflow	+wind	0.0015	6
3	-airflow	+motion	0.0019	4
4	+airflow	+friction	0.0024	4
5	+labial	+motion	0.0107	4
6	-vocal folds	+telic	0.0121	4
7	-tongue resting	+telic	0.0159	4

Results

Out of 299 combinations of articulatory to semantic features, 69 combinations were significant across languages according to Wilcoxon signed rank tests. To be conservative, we used the results of the single-language chi-squared tests as a threshold for reporting Wilcoxon signed rank tests across languages. Specifically, Wilcoxon tests reported here are only those that apply to combinations which were significant ($p < 0.05$) for 4 or more languages on an individual basis. 7 articulatory feature and semantic feature, shown in Table 3, were above this threshold. The correlations in Table 3 are ordered according to the number of languages who had a significant articulatory to semantic feature correlation. The correlation of [-airflow] to [+telic] and [+airflow] to [+wind] are our most robust articulatory feature to semantic feature mapping ($z = 0.00, p = 0.0015$) across all languages, and are significant (χ^2,$p < 0.05$) for 7 and 6 languages on an individual-basis respectively. The correlation of [-airflow] to [+motion] is significant across all languages (Wilcoxon, $p = 0.0019$).

Discussion

Our results overall show that certain articulatory properties map to semantic features of ideophones from 13 languages. More specifically, our results show that phonosemantic mappings as proposed in the ideophone literature (see **Hypotheses** section, hypotheses 1-3) are supported, while [+/-tongue root] was not significant for [+/- motion] as claimed by hypothesis (4). Table 3 shows that five modes of articulation create robust cross-linguistic patterns with regards to imitative meaning. These five modes are: tongue movement, lip movement, airflow, velum lowering (nasal airflow), and vocal fold vibration. This suggests that the imitative nature of ideophones is begotten from perceptuo-motor analogies afforded by such articulatory properties. That is to say, imitative words to an extent derive their imitative meaning through their articulation, implying that articulatory properties of speech are a potential route for explaining the iconic nature of words, such as ideophones. By extension, words of contested iconic nature could thus be deemed more or less iconic depending on whether their articulatory properties support such a claim, For example, if *gl-* of *glisten*, *glimmer*, *glint* was to be proven iconic and therefore imitative, an analogy supported by articulatory features would be required to argue for its purported meaning of luminescence.

If iconicity is imitative due to perceptuo-motor analogy (Dingemanse, Blasi, Lupyan, Christiansen, & Monaghan, 2015) (relations made between sensory percepts and movements), then articulatory properties should likewise map to semantic features for reasons grounded in perceptuo-motor analogy. In Table 4, we propose the perceptuo-motor analogies that allow these articulatory properties to pattern with their semantic features and are in turn embedded in a given ideophone on a sub-phonemic level.

There are few things worth noting regarding the overlap of semantic features. First is that the articulatory feature [+airflow] corresponds to semantic features [+friction] and [+wind] but not motion, i.e., [+air flow] corresponds to [-motion]. This does not imply that [+friction] ideophones are not coded for movement related meaning (as friction must imply some kind of movement). Rather, this implies ideophones which are no to do with motion[4], and are thus beyond motion on Dingemanse (2012) semantic hierarchy for ideophones, involve [+airflow]. With that in mind, the finding that [+labial] corresponds to [+motion] would imply that some (not necessarily complete) occlusion of airflow made by contact with the articulators, is involved in the perceptuo-motor analogy of [+motion]. However, here we would argue that it is the movement of the articulators, not the blockage of air, which affords this perceptuo-motor analogy of movement. This is because [+labial] allows for labio- and labiodental fricatives which of course are consonants coded as [+airflow]. This is further supported by the fact that [-tongue resting], i.e., tongue movement rather than lip movement, also corresponds with [+telic]. Implying that the tongue is used to occlude air in the oral tract to give us the correspondence of [-airflow] to [+telic].

Another observation regarding the overlap of features is that the semantic feature [+telic] is associated with articulatory features [-vocal folds], and [-airflow]. Bear in mind that our feature airflow does not encompass nasal consonants, i.e., air escaping through the nose. We did not find the relation of [+velum] to [+telic] to be significant overall using our Wilcoxon signed rank test (p=0.0869) and thus it is unreported in Table 4. However our chi-squared tests showed it to

[4]There are very few ideophones in our database which are [+motion] but [-sound]. If ideophones are [+motion] they are almost always [+sound], implying that the sound is resultative of the motion and somehow semantically entails it. For example, an ideophone for the sound of footsteps would be [+sound] and [+motion]. The reverse however is not true. For example, the sound of a cow or the sound of wind blowing is [+sound] but [-motion].

Table 4: Analogical justifications for articulatory feature and semantic feature correspondences across 13 languages.

#	Articulation	Corresponds	Justification (≈ analogical to)
1	-airflow	+telic	airflow occlusion ≈ cessation of an event
2	+airflow	+wind	continual airflow ≈ air movement
3	-airflow	+motion	movement of articulators to obstruct airflow ≈ movement depiction
4	+airflow	+friction	airflow sibilance ≈ sibilance of friction and/or ≈ rubbing of two surfaces
5	+labial	+motion	movement of lips ≈ motion depiction
6	-vocal folds	+telic	lack of vocal fold vibration ≈ cessation of an event
7	-tongue resting	+telic	active tongue movement to create [-airflow] articulation ≈ cessation of an event

be significant ($p < 0.05$) in 5 languages on an individual basis (Chaoyang, Akan, Kam, Maai, and Manyika Shona). Though not as robust as other findings, taken together with the other [+telic] associations, this [+velum] to [+telic] pattern would suggest that unvoiced stops are likely associated with telicity, nasal consonants are an exception. This implies that the occlusion created by nasal consonants (air blocked from entering the oral cavity) is just as important as the occlusion of air from escaping the oral cavity for [+telic] ideophones. We can propose that it is the articulatory gesture of blocking of air, an articulatory property common to [+velum] and [-airflow] consonants, that affords the perceptuo-motor analogy of [+telic]. Vocal fold vibration is inherent to nasals. However, as our results show, [-vocal folds] is significantly associated to [+telic] ideophones. This implies that the [-airflow] consonants are those that are unvoiced. Based on the articulatory similarities between [-airflow] and [+velum] consonants, we might also propose that the voicing inherent to nasals is not as important for perceptuo-motor analogy of [+telic] as the occlusion of air. Overall, our results also show for some languages certain articulatory properties pattern with semantic features while others do not. Therefore some perceptuo-motor analogies could be language specific. These language-specific results may have come about for a number of reasons. Firstly, phoneme inventories differ across languages so it is inevitable that some languages make use of certain articulatory features less than others, e.g., voicing. Crucially, we did not take predictable phonotactic processes into account when entering the ideophones into our database. Phonotactic processes could result in the addition or deletion of certain segments in order to satisfy language-specific phonological rules and thus potentially obscuring and/or skewing the articulatory features present for imitative purposes only. Controlling for said phonotactic processes requires in-depth analysis per language (Thompson & Do, in press). We would like to emphasize, however, that our main goal here was to see if there were any cross-linguistic articulatory-semantic patterns despite the presence of language-specific phonotactic patterns. The significance of eight articulatory-semantic feature mappings show that this is possible.

Future directions of research could look into how syllable structure affects the patterning of articulatory features with semantic features. For example, stop consonants, characterized as [-airflow] in our study, might be more attested in codas of ideophones depicting telic events, since the coda is the final segment of a syllable and is thus considered imitative of an events endpoint (Hinton et al., 1994; Strickland et al., 2017). Articulatory properties of vowels are also an obvious direction for future studies, especially given what has been gleaned from the rich literature on *kiki-bouba* studies (Lockwood & Dingemanse, 2015), as well as recent acoustic work on vocal imitations (Perlman & Lupyan, 2018). Given that we only report correlations between individual articulatory features and individual semantic features, future tests could look at how features cluster together, e.g., [+labial] [-airflow] or [+telic] [+motion]. Experimental research could test the results of our study by seeing whether (1) articulatory feature and semantic feature patterns are easily learnable for novel words or ideophones, (2) speakers refer to these articulatory features or perhaps exaggerate them when explaining the meaning of ideophones, as with Dingemanse (2015)'s study on folk definitions of Siwu ideophones. Finally, our study unifies iconicity in the spoken and visual modalities, since both rely on movement to make imitative meanings.

Acknowledgments

We would like to extend our thanks to Ryszard Auksztulewicz, Stephen Matthews, Kirsty Rowan, and to those who attended the *CLS-MPI Iconicity Focus Group Workshop*, Nijmegen (2017) and *the Workshop on Mimetics II: New Approaches to Old Questions*, Nagoya (2017) for the stimulating discussion which inspired this paper.

References

Akita, K., Imai, M., Saji, N., Kantartzis, K., & Kita, S. (2013). Mimetic vowel harmony. *Japanese/Korean Linguistics*, *20*, 115–129.

Alpher, B. (1994). *Yir-yoront ideophones.* Cambridge University Press, Cambridge.

Aryani, A. (2018). *Affective iconicity in language and poetry* (Unpublished doctoral dissertation).

Assaneo, M. F., Nichols, J. I., & Trevisan, M. A. (2011). The anatomy of onomatopoeia. *PloS one*, *6*(12), e28317.

Ayalew, B. (2013). *The submorphemic structure of amharic: toward a phonosemantic analysis* (Unpublished doctoral dissertation). University of Illinois at Urbana-Champaign.

Beck, D. (2008). Ideophones, adverbs, and predicate qualification in upper necaxa totonac. *International Journal of American Linguistics*, *74*(1), 1–46.

Blasi, D. E., Wichmann, S., Hammarström, H., Stadler, P. F., & Christiansen, M. H. (2016). Sound–meaning association biases evidenced across thousands of languages. *Proceedings of the National Academy of Sciences*, *113*(39), 10818–10823.

Childs, G. T. (1988). The phonology of kisi ideophones. *Journal of African languages and linguistics*.

De Carolis, L., Marsico, E., & Coupé, C. (2017). Evolutionary roots of sound symbolism. association tasks of animal properties with phonetic features. *Language & Communication*, *54*, 21–35.

Diffloth, G. (1979). Expressive phonology and prosaic phonology in mon-khmer. *Studies in tai and mon-khmer phonetics and phonology: In honour of Eugenie j. A. Henderson*, 49–59.

Diffloth, G. (1994). I: big, a: small. *Sound symbolism*, 107–114.

Dingemanse, M. (2012). Advances in the cross-linguistic study of ideophones. *Language and Linguistics compass*, *6*(10), 654–672.

Dingemanse, M. (2013). Ideophones and gesture in everyday speech. *Gesture*, *13*(2), 143–165.

Dingemanse, M. (2015). Folk definitions in linguistic fieldwork. *Language Documentation and Endangerment in Africa.*, 215–238.

Dingemanse, M. (2018). Redrawing the margins of language: Lessons from research on ideophones. *Glossa: a journal of general linguistics*, *3*(1).

Dingemanse, M. (in press). Ideophone as a comparative concept. In *Ideophones, mimetics, and expressives* (p. 1333). John Benjamins.

Dingemanse, M., Blasi, D. E., Lupyan, G., Christiansen, M. H., & Monaghan, P. (2015). Arbitrariness, iconicity, and systematicity in language. *Trends in cognitive sciences*, *19*(10), 603–615.

Dingemanse, M., Schuerman, W., Reinisch, E., Tufvesson, S., & Mitterer, H. (2016). What sound symbolism can and cannot do: Testing the iconicity of ideophones from five languages. *Language*, *92*(2), e117–e133.

Franck, G. E. (2014). Ideophones in manyika shona: A descriptive analysis of ideophones and their function in manyika (bantu).

Gerner, M. (2005). Expressives in kam (dong 侗): A study in sign typology (part 11). *Cahiers de Linguistique Asie Orientale*, *34*(1), 25–67.

Hamano, S. (1998). *The sound-symbolic system of japanese.* ERIC.

Hatton, S. A. (2016). The onomatopoeic ideophone-gesture relationship in Pastaza Quichua.

Hinton, L., Nichols, J., & Ohala, J. J. (1994). *Sound symbolism*. Cambridge University Press.

Hinton, L., Nichols, J., & Ohala, J. J. (2006). *Sound symbolism*. Cambridge University Press.

Iwasaki, N., Vinson, D. P., & Vigliocco, G. (2007a). Chapter one how does it hurt, kiri-kiri or siku-sikui: Japanese Mimetic Words oF Pain Perceived By Japanese Speakers and English Speakers. *Applying theory and research to learning Japanese as a foreign language*, 2.

Iwasaki, N., Vinson, D. P., & Vigliocco, G. (2007b). What do English speakers know about gera-gera and yota-yota?: A cross-linguistic investigation of mimetic words for laughing and walking. *Japanese-language education around the globe*, *17*, 53–78.

Kanu, S. M. (2008). Ideophones in temne. *Kansas Working Papers in Linguistics*, 120134. doi: 10.17161/kwpl.1808.3909

Kawahara, S., Noto, A., & Kumagai, G. (2018). Sound symbolic patterns in pokemon names. *Phonetica*.

Kwon, N., & Round, E. R. (2015). Phonaesthemes in morphological theory. *Morphology*, *25*(1), 1–27.

Lockwood, G., & Dingemanse, M. (2015). Iconicity in the lab: A review of behavioral, developmental, and neuroimaging research into sound-symbolism. *Frontiers in psychology*, *6*, 1246.

Lockwood, G., Hagoort, P., & Dingemanse, M. (2016). How iconicity helps people learn new words: Neural correlates and individual differences in sound-symbolic bootstrapping.

Maduka, O. N. (1988). Size and shape ideophones in nembe: A phonosemantic analysis. *Studies in African linguistics*, *19*(1), 93–113.

Mathangwane, J. T., & Ndana, N. (2014). Chiikuhane/chisubiya ideophones: A descriptive study. *South African Journal of African Languages*, *34*(2), 151–157.

McCune, K. M. (1985). *The internal structure of Indonesian roots* (Vol. 23). Badan Penyelenggara Seri Nusa, Universitas Katolik Indonesia Atma Jaya.

Mihas, E. (2013). Composite ideophone-gesture utterances in the ashéninka perené 'community of practice', an amazonian arawak society from central-eastern peru. *Gesture*, *13*(1), 28–62.

Nuckolls, J. B., et al. (2000). Spoken in the spirit of gesture: Translating sound symbolism in a pastaza quechua narrative. *Translating native Latin American verbal art*, 233–251.

Nuckolls, J. B., Swanson, T., Sun, D., Rice, A., & Ludlow, S. (2017). *Quechua realwords: An audiovisual corpus of expressive quechua ideophones.* Retrieved 2019-10-20, from `http://quechuarealwords.byu.edu/`

Oda, H. (2001). An embodied semantic mechanism for mimetic words in Japanese.

Ofori, S. A. (2009). A morphophonological analysis of onomatopoeic ideophones in akan (twi). *IUWPL8: African Linguistics Across the Discipline*, 11-44.

Oswalt, R. L. (1994). Inanimate imitatives in english. *Sound symbolism*, *293306*.

Perlman, M., & Lupyan, G. (2018). People can create iconic

vocalizations to communicate various meanings to naïve listeners. *Scientific reports*, *8*(1), 2634.

Schackow, D. (2016). *A grammar of yakkha*. Language Science Press.

Shih, S. S., Ackerman, J., Hermalin, N., Inkelas, S., & Kavitskaya, D. (2018). Pokémonikers: A study of sound symbolism and Pokémon names. *Proceedings of the Linguistic Society of America*, *3*(1), 42–1.

Strickland, B., J, K., Schlenker, P., & Geraci, C. (2017). *Intuitive iconicity for events and objects: telicity and the count/mass distinction across modalities.* Workshop on Event Representations in Brain and Language Development. Nijmegen: MPI for Psycholinguistics.

Taitz, A., Assaneo, M. F., Elisei, N., Trípodi, M., Cohen, L., Sitt, J. D., & Trevisan, M. A. (2018). The audiovisual structure of onomatopoeias: An intrusion of real-world physics in lexical creation. *PloS one*, *13*(3), e0193466.

Thompson, A. L., & Do, Y. (in press). Defining iconicity: an articulation-based methodology for explaining the phonological structure of ideophones. *Glossa: a journal of general linguistics*.

Van Hoey, T. (2018). Does the thunder roll? Mandarin Chinese meteorological expressions and their iconicity. *Cognitive Semantics*, *4*(2), 230–259.

Wang, S., & Tang, Y. (2014). Hanyu ni sheng ci yu weiwu'er yu moni ci duibi qian xi [A comparative study of onomatopoeia of Chinese and Uyghur]. *Yuyan yu fanyi [Language and translation]*(1), 34–37.

Waugh, L. R. (1994). Degrees of iconicity in the lexicon. *Journal of pragmatics*, *22*(1), 55–70.

Xiao, C. (2015). Manyu nishengci chuyi [Primary Research of Manchu Onomatopoetic Words]. *Manchu Studies*(60: 1), 19–23.

Inductive Biases Constrain Cumulative Cultural Evolution

Bill Thompson (billdthompson@berkeley.edu)
Social Science Matrix, Univeristy of California, Berkeley

Thomas L. Griffiths (tgriffiths@princeton.edu)
Departments of Psychology and Computer Science, Princeton University

Abstract

Cumulative cultural evolution is a distinctively human form of information-processing that endows our societies with improbable and efficient technologies. But how objective is this process? A widely held conjecture is that human cognitive biases can constrain cumulative cultural evolution, and therefore shape our discoveries. We present a Bayesian analysis of a simple form of cumulative cultural evolution. This model allows us to formulate and test the theoretical conjecture in an experimental setting. Across a series of behavioural experiments, we show that people's inductive biases constrain a population's ability to discover counter-intuitive virtual technologies in a simple search problem. Our analysis highlights formal relationships between cumulative cultural evolution, Bayesian inference, and stochastic optimization.

Keywords: cumulative cultural evolution; inductive biases; optimization; computation; Bayes; cultural evolution;

Introduction

We are surrounded by bizarre and complex objects that vastly improve our lives. To our recent ancestors, many of the tools and technologies we rely on today were inconceivable, yet the same innovations will soon seem primitive to our descendants. The capacity for cumulative discovery is a uniquely human form of information processing on a breath-taking scale – but how objective is this process? Is technological evolution an unbiased search for optimal solutions to the problems we face? Or is it shaped by the same representational constraints and biases that limit individuals?

This question has been widely discussed in the context of *cultural evolution* (Mesoudi, 2016). Most theories of cultural evolution agree on the conjecture that in some circumstances, human cognitive biases must constrain cumulative cultural evolution (Morin, 2016; Acerbi & Mesoudi, 2015; Claidière, Scott-Phillips, & Sperber, 2014). This hypothesis has been widely debated and examined in formal models (Claidière & Sperber, 2007; Boyd & Richerson, 1985; Henrich & Boyd, 2002; Griffiths, Kalish, & Lewandowsky, 2008), but it has never been tested experimentally. In part, testing this hypothesis has been challenging because it is difficult to quantify an appropriate set of expectations in an experimental setting (Miton & Charbonneau, 2018).

In this paper, we develop a mathematical model of cumulative cultural evolution that allows us to formulate these expectations precisely. Our model is derived from a Bayesian analysis of individual cognition. The model makes quantitative predictions about the circumstances under which inductive biases are likely to stifle discovery. To test these predictions, we adapt a widely studied experimental paradigm in which participants design and transmit a simple artificial technology: virtual arrowheads. Our strategy is to first characterise participants' inductive biases in this context using serial reproduction chain experiments. On the basis of these estimates, we conduct a series of arrowhead-design experiments which differ only in the extent to which task reward structure contradicts participants' biases. In the process of formalizing our predictions, we identify a formal relationship between cumulative cultural evolution and stochastic optimization.

Background

Cumulative Cultural Evolution

Unlike other species, every generation of humans builds on the insights and actions of their ancestors (Henrich, 2015). When an Apple engineer develops iPhone security updates, she makes use of cognitive resources expended by Turing almost a century before. In this sense, people alive today extend the computations initiated by people who faced similar problems in the past. What kind of process allows us to effectively pool computational resources with strangers over seemingly unbounded timescales? Computation over generations depends on a proclivity to learn from the people around us and the artefacts they create. When this kind of learning is repeated over time, a stochastic process is induced. This process is called *cultural transmission* (Boyd & Richerson, 1985). In some species, cultural transmission leads to cumulative innovation. This special case is known as *cumulative culture* (Mesoudi & Thornton, 2018) and is surprisingly rare (Whiten, Caldwell, & Mesoudi, 2016).

Discovering Technologies

There are many forms of cumulative culture, but one simple example has been heavily studied: refinement of technologies towards consistent functional objectives. Outside of the laboratory, examples of this process are easy to find: motorcycles today are faster, more efficient, more reliable, safer, and longer-lasting than the motorcycles people rode during the second World War. In an experimental setting, small-scale analogues of this process have been studied in several domains. For instance, Caldwell and Millen (2008) showed that micro-societies of experimental participants discover cumulatively more effective ways to design a tall-standing tower

of spaghetti. In these experiments, later participants observed the designs of earlier participants, and created taller and taller spaghetti towers as a result. Similar findings have been reported in lineages of participants designing simple knots (Muthukrishna, Shulman, Vasilescu, & Henrich, 2014), paper aeroplanes (Caldwell & Millen, 2008), rice baskets (Zwirner & Thornton, 2015), and fishing nets (Derex, Beugin, Godelle, & Raymond, 2013), for example.

Inductive Biases in Cumulative Culture

Cumulative cultural evolution can be recreated and manipulated in the laboratory. However, it remains unclear whether the products of these processes are shaped by biases in the way people think, or whether inductive biases and representational constraints are effectively washed out over time. This has been difficult to establish empirically, in part because it is often challenging to quantify the influence of people's inductive biases. Recent reviews have noted that experimental tasks often feature unconstrained or difficult to quantify design spaces (Miton & Charbonneau, 2018), and that there is a need for a better understanding of the information-processing dynamics that link cognition and cultural evolution (Mesoudi & Thornton, 2018; Heyes, 2018). Mathematical analyses have repeatedly identified the potential for human biases to shape cumulative cultural evolution (Claidière & Sperber, 2007; Boyd & Richerson, 1985; Griffiths et al., 2008). However, extending abstract models to an experimental setting remains a challenge. Here, we introduce a formal model that is closely related to these theories of culture, but derived from a Bayesian analysis of cognition, and therefore directly applicable in an experimental context. Our analysis extends prior Bayesian models of cultural evolution (Griffiths & Kalish, 2007; Navarro, Perfors, Kary, Brown, & Donkin, 2018) to the cumulative case. The model we introduce allows us to specify formal predictions about the circumstances in which inductive biases constrain cumulative cultural evolution, and test those predictions experimentally.

Model: Optimization by Cumulative Culture

Our analysis applies to settings in which the design features of a technology can be described in terms of (n) continuous valued parameters $\Theta^t \in \mathbb{R}^n$. This setting offers a natural connection to prior experimental work, in which participants modify design features such as length, height, width, angles, crossing points, mass, or hue.

Induction of a Design

Each new individual estimates these design features from artefacts produced by the previous generation. If this estimation procedure can be given a formulation as Bayesian inference, then an individual's estimate $\hat{\Theta}$ can be decomposed into a trade-off between two quantities: noisy empirical observation of the true design features Θ; and inductive biases imposed by cognition. Inductive biases can expressed as a prior distribution $p(\Theta)$. Using this framework, $\hat{\Theta}^t$ can be treated as a random variable distributed according to the posterior distribution implied by a Bayesian model of learning.

Innovation

After estimating the existing design, each participant attempts an innovation. Assume a $f\colon \Theta \to \mathbb{R}$ is a function that reflects the utility of a technology with respect to it's design features. In the literature on cultural evolution, this quantity would sometimes be referred to as a *fitness* landscape. We will make the assumption that individuals are capable of bounded, local innovation. This is appropriate to scenarios in which innovation is largely driven by an ability to identify similar but improved variants of whatever already exists, through limited experimentation with minor design variations for example. Local information about f can be naturally expressed as its gradient with respect to design features, evaluated at Θ^t. We denote this quantity $\nabla_f = \nabla_\Theta f(\Theta^t)$.

Diffusion Chains

These assumptions formalize a simple theory of cumulative culture as repeated cycles of observation, induction, and local innovation, leading to the expression:

$$\Theta^{t+1} = \Theta^t - \alpha\nabla_f - (\Theta^t - \hat{\Theta}^t)\,, \tag{1}$$

where $t \in 1,\ldots,T$ denotes a specific generation in a transmission chain. This equation describes a single step of a transmission chain in terms of the relationship between an existing technology (Θ^t), its utility (f), its status with respect to human cognitive constraints and the fidelity of transmission ($\Theta^t - \hat{\Theta}^t$), and an innovation rate (α). We examine the properties of this general model under some simplifying assumptions.

Assumption 1: Gaussian Prior & Observation Noise Assume individual learning can be modelled as probabilistic inference in a Gaussian model: observations of an existing design are noisy, and this noise can be approximated by Gaussian corruption of the true design features; inductive biases can be approximated by a Gaussian distribution.

Assumption 2: Independent Features Individual design features $\theta_i \in \Theta$ can be treated independently. This is a limiting assumption, but nonetheless appropriate to many relevant contexts. If μ_i is the prior expectation, the posterior expectation is:

$$\mathbb{E}\left[\hat{\theta}_i^t\right] = \lambda_i\mu_i + (1-\lambda_i)\theta_i^t \tag{2}$$

where $\lambda_i = \sigma_i^2/(\sigma_i^2+\delta_i^2)$ reflects the relative variance of the prior (δ_i^2) and observation noise (σ_i^2) – in other words, the strength of an inductive bias $p(\theta_i)$ relative to the fidelity of transmission.

Chain Dynamics

The expected change at each generation can be written:

$$\mathbb{E}\left[\theta_i^{t+1} - \theta_i^t\right] = \lambda_i(\mu_i - \theta_i^t) - \alpha\nabla_f\,. \tag{3}$$

which implies no further accumulation in expectation when $\nabla_f = \nabla_f^*$, where $\nabla_f^* \equiv (\mu_i - \theta_i^t)(\lambda_i/\alpha)$. This cultural process will halt if the potential for local innovation drops below a threshold determined by: the distance of the current design from the prior expectation ($\mu_i - \theta$), relative to a willingness to explore (α), weighted by the balance of prior and empirical leniencies in learning (δ_i^2/σ_i^2).

Assumption 3: Quadratic Utility Landscape The fate of the process is closely tied to the utility landscape in which it is operating. A simple but broad class of cases can be captured by the assumption that there is an optimal design Θ^*, and that f can be locally approximated by a quadratic surface with the optimum design at its minimum / maximum. In this regime, utility decreases with squared distance from the optimum at a rate proportional to a parameter a. A utility landscape that can be described in this manner has gradients $\nabla_f = a(\theta - \theta_i^*)$. A transmission process acting on a utility landscape of this form will halt in expectation if it reaches $\theta_i^t = \phi_i^*$:

$$\phi_i^* = \lambda_i^* \mu_i + (1 - \lambda_i^*)\theta_i^* \tag{4}$$

which is a linear combination of the prior mean μ_i and the optimum design θ_i^* with mixing proportions:

$$\lambda_i^* = \sigma_i^2/(\sigma_i^2 + \alpha a s) \tag{5}$$

where $s = \delta_i^2 + \sigma_i^2$. Equations (4) and (5) represent our main theoretical result. Our analysis predicts that the outcome of a transmission chain is a compromise between the inductive biases of individuals and the optimal design. When these conflict, the balance of the compromise is quantifiable from the relationships between: transmission fidelity (σ_i^2), strength of inductive bias (δ_i^2), an exploration rate (α), and the slope of the utility landscape (a). The weighting factor $0 \leq \lambda_i \leq 1$ interpolates between cultural evolutionary processes that are constrained by inductive biases ($\lambda \to 1$) and therefore dragged back toward the prior $p(\theta_i)$, and processes that are dominated by information contained in the the utility landscape ($\lambda \to 0$), and therefore destined to discover an objective optimum.

Biased Computation by Cumulative Culture

One way to interpret this finding is as a description of the computation that is being implemented by the *process* we have analysed – the computation implemented by a chain of individuals. Two analogies motivate this interpretation. First, equation (5) has the same form as equation (4). At each generation, an individual person performs a computation that we formalized as a sample from the posterior distribution in a Bayesian model of inference. This computation is biased and *local*: the expectation is a linear combination of the individual's inductive bias and the *currently existing* design θ^t. However, the chain *as a whole* can be understood to implement a biased but *global* computation: equation (5) describes the expectation of a posterior distribution computed by the same kind of learner after observing (a noisy realisation of) the *optimal* design θ^*. Second, in the Gaussian case, equation (1) can be rewritten as:

$$\theta_i^{t+1} = \theta_i^t - \alpha \nabla_f - \lambda_i(\theta_i^t - \mu_i) + \varepsilon_i^t \tag{6}$$

which is a form of stochastic gradient descent with regularisation. Stochastic optimization and Bayesian inference are known to be related (Mandt, Hoffman, & Blei, 2017). This highlights a common interpretation of cognitive and cultural processes – they are both forms of information processing. This cultural process solves an optimization problem subject to regularisation by human inductive biases. In the remainder of this paper, we test this prediction.

Experiment: Discovering Virtual Technologies

We adapted an experimental paradigm that has been widely used to study the influence of social learning on cumulative culture (Mesoudi, Chang, Murray, & Lu, 2015). The experimental task involves designing a virtual arrowhead. The arrowhead has a number of attributes (e.g. length, width) that can be modified and achieves a score when deployed on a virtual hunt. The score reflects the number of calories of food earned by the arrowhead. Participants' goal was simply to test and redesign a single arrowhead they inherited, in an attempt to increase its score. This paradigm allowed us to construct a low-dimensional search problem in which we hypothesised that task-naive participants would display biased expectations. Athough there is signfificant discussion surrounding the definition of cumulative cultural evolution, a central requirement is that over time, it's products must"enhance some measure of performance...[through]...sequential improvements..." (Mesoudi & Thornton, 2018). In our experiment, unconstrained cumulative cultural evolution would correspond to the chains of participants sequentially designing arrowheads that achieve higher scores until the maximum score is achieved.

Method

Stimuli The experiment was presented as a website. Participants designed a virtual arrowhead using two HTML range sliders which modified its width and length. Figure 1 shows the design space. During experimental trials, the screen was split into left (25% screen width) and right (75% screen width) panels. The left panel displayed the participant's estimate of the arrowhead they inherited in reduced proportions that nonetheless preserved the design. Underneath was a depiction of two range-sliders positioned in accordance with the arrowhead's attributes, and text indicating the number of calories that the arrowhead earned. The main panel (Right) displayed an arrowhead in the center, pointing downwards. Two range-sliders were located beneath the arrowhead. Moving a range-slider modified either the length or the width of the arrowhead dynamically. Both dimensions of the arrowhead could take values ranging between 50 and 150 pixels.

Arrowhead scores were determined by a quadratic function of the form $f(\theta_i) = \frac{1}{2}a(\theta_i - \theta_i^*)^2 + c$, where θ_i^* is the

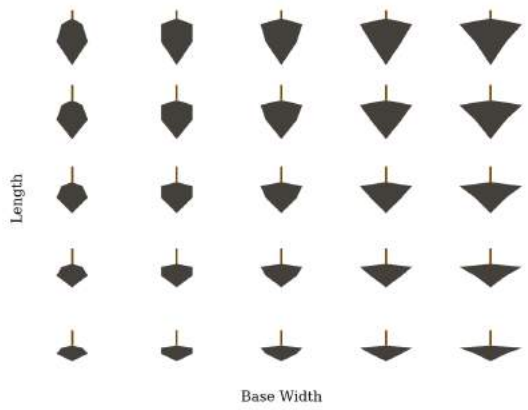


Figure 1: Regularly sampled virtual arrowheads in a design space implied by the ability to modify two features – length and base width.

optimum value for feature θ_i. We chose to use this family of functions because it constructs a smooth utility landscape with a single optimum. Previous work has focused on conditions that allow populations to search effectively through more complex landscapes with both local and global optima (Acerbi, Tennie, & Mesoudi, 2016). In contrast, our analysis focuses on a problem that should be relatively easy to solve optimally if people's inductive biases do no constrain cultural evolution. Quadratic functions are also the class of landscapes we examined in our theoretical analysis.

Within the family of quadratic landscapes, we required a function which was: smooth over the full range; did not return negative scores; had an optimum (maximum) score that lies within a semantically reasonable range given the framing of the task. To meet these requirements, we set $\alpha = -30$ and $c = 10000$ for all experiments and divided the result by 100. Given our settings of θ_i^* (see below), the maximum available score was 1000 calories. Calories decreased away from θ_i^* at a rate given by $\nabla_f = -\frac{3}{10}(\theta_i' - \theta_i^*)$. To award a score, we computed f for both arrowhead features (length and base width) and awarded the mean.

Procedure Participants were informed that: they would go on a virtual hunt; their task was to design a virtual arrowhead that will earn as many calories of food as possible on the hunt; a bonus payment would be made in proportion to the number of calories their arrowhead earned. After consenting to the experiment, participants completed an Information Trial (IT), during which they observed the arrowhead they had inherited (first generation participants inherited an arrowhead with a randomly sampled design). This arrowhead was displayed in the center of the right panel for 3000 milliseconds. Participants then recreated the arrowhead as accurately as possible. Participants then proceeded to the first Modification Trial (MT). The participant's estimate of the arrowhead design was displayed as the arrowhead they inherited in the left panel. Participants completed four MTs. During each MT, $\hat{\theta}_i'$ was displayed in the left panel. At the beginning of an MT, no arrowhead was displayed in the right panel, and the positions of the range-sliders were randomised. Participants could not proceed to the next trial until at least one range-slider had been modified. Upon modifying any of the rage-sliders positions, the arrowhead was redrawn. Modifications were limited to a range of ± 30 pixels around $\hat{\theta}_i'$. This enforced a weak restriction on the innovations participants could make in accordance with our theoretical model. There was no limit to the number of times participants could modify the range-sliders in a given MT. Once satisfied, participants could click Submit to obtain feedback – the number of calories earned by the current arrowhead design. Arrowheads evaluated during previous MTs were displayed (in reduced proportions) in the left panel, in trial order. After completing four MTs, participants were informed that their opportunity to test arrowheads was complete and proceeded to the test trial (TT). Participants were reminded that the arrowhead they designed during this trial would determine a bonus payment. TT was identical to MT in all other respects.

Participants Participants ($n = 1000$) were recruited online using Amazon's Mechanical Turk. The experimental protocol was approved by the University of California, Berkeley's Committee for the Protection of human Subjects. Participants were paid \$0.50 to complete the experiment, and awarded a performance-based bonus of up to \$0.50. Most participants completed the experiment in less than three minutes. Data from any participants who completed the experiment in less than 20 seconds were rejected.

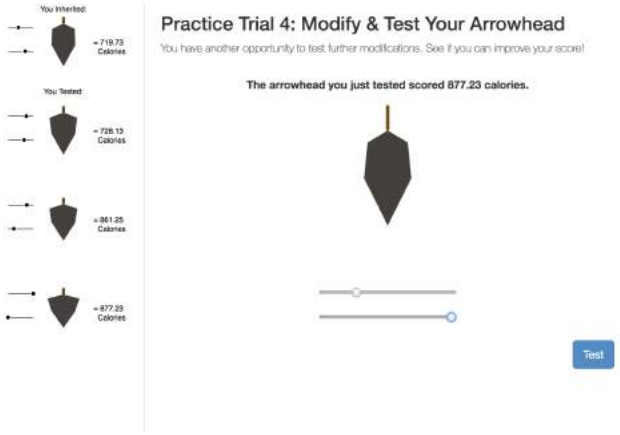


Figure 2: Participant view of the experiment. Screenshot shows the fourth Modification Trial. At this point in the expreriment, the participant has completed the Information trial (and recreated their inherited arrowhead, shown first in the left panel) and three Modification Trials (the arrowheads tested by the particiant so far and their scores are shown in the left panel).

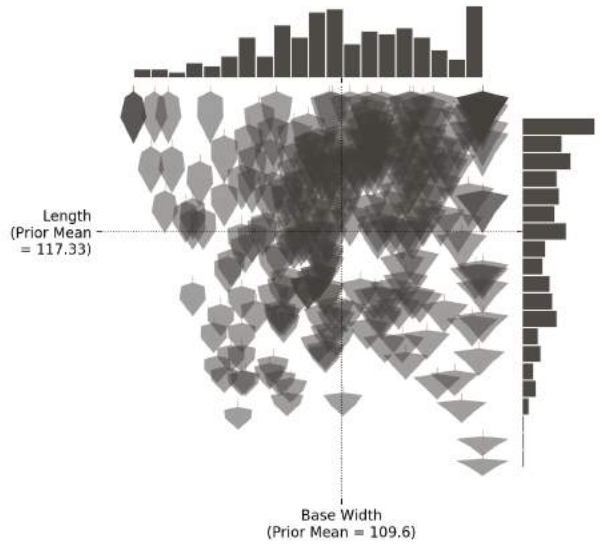


Figure 3: All arrowheads produced by participants at generation 3 or later in 20 serial reproduction chains of 10 generations each. A sample based approximation to participants' inductive biases, $p(\theta)$.

Results

Reproduction Chains We first ran a simpler experiment using the same stimuli. This experiment used *serial reproduction chains* to characterise participants' inductive biases in our stimulus set. In these chains, each participant completed an IT, but did not proceed to MT and TT. Each participant observed the arrowhead designed by the previous participant, and was asked to reproduce it as accurately as possible. Previous mathematical (Griffiths & Kalish, 2007) and experimental (Griffiths et al., 2008; Xu & Griffiths, 2010) research has established that serial reproduction chains characterise participants' inductive biases. Figure 3 shows the distribution of arrowheads produced by all participants at generation 3 or later (in all our analyses, the first two generations of a chain are excluded as *burn-in* generations to minimise the effects of random initial conditions), across 20 serial reproduction chains of 10 generations each. This collection of arrowheads can be understood as a sample-based approximation to the prior distribution $p(\Theta)$. The empirical mean of this distribution is $\hat{\mu}_{width} = 111, \hat{\mu}_{length} = 118$. People favoured arrowheads that are relatively long and relatively wide.

Optimization Chains In light of participants' inductive biases, we conducted four experiments. We treat these as separate experiments rather than experimental conditions because they were carried out sequentially. Each experiment (20 chains of 10 generations) featured a utility landscape with a differently located optimum but the same calorie gradient surface. Our prediction was that differently located optimums would lead to differential discovery of those designs, and differential task success (number of calories). Our mathematical analysis identified the distance between the optimum arrowhead and the mean of the prior distribution $p(\theta)$ as the crucial predictive quantity: optimal arrowheads that are farther from the prior distribution should be harder to find because they are less intuitive. Figure 4 shows our results. Experiment 1 ($\theta^*_{width} = 115, \theta^*_{length} = 115$) was designed to be most consistent with people's inductive biases. In this experiment, the arrowheads people designed scored well (mean calories $M = 948$, $SD = 46$). Experiment 4 was least consistent with people's biases, contradicting people's expectations in both dimensions. Success in the task suffered as a result ($M = 794$, $SD = 183$). Experiments 2 ($\theta^*_{width} = 75, \theta^*_{length} = 115$, $M = 912$, $SD = 98$) and 3 ($\theta^*_{width} = 115, \theta^*_{length} = 75$, $M = 845$, $SD = 132$) were designed to contrast with people's biases in one of the two dimensions – width and length respectively.

We combined data from all four experiments and computed the difference between the optimum arrowhead and the mean of the prior distribution. The main prediction of our formal model (equation 4) can be rearranged into a linear model of the form $\phi^*_i = \hat{\mu}_i + \beta(\theta^*_i - \hat{\mu}_i)$. This allowed us to perform an ordinary least squares regression analysis of this model in our experimental data. The prediction was upheld. Accounting for the mean of the prior distribution ($\hat{\beta} = 1.0, p < .001$) and the difference between the mean of the prior and the optimum design ($\hat{\beta} = 0.42, p < .001$) accounted 96% of the variance in the features of the arrowheads people produced ($R^2 = 0.962$). We also analysed task success, and found significant differences in the distribution of arrowhead scores in all pairwise comparisons of our four experiments (at $\alpha = .05$). Only the comparison between experiments 3 and 4 ($t(197) = 3.2, p = 0.0017$) was not significant at $\alpha = .001$. Figure 4 (b) shows how task success reduced over the four experiments. Finally, we computed the predictions of our mathematical model under the inferred mixing proportions λ explicitly. Figure 4 (c) shows these predictions.

Conclusion

We introduced a simple formal theory of cumulative cultural evolution. We used this theory to predict how the inductive biases of individuals would constrain a cultural process. We tested this prediction in a series behavioural experiments. We found that discovery of an optimal virtual technology in a simple search problem was impeded by people's inductive biases. These results reinforce a theoretical conjecture that had previously not been studied empirically. Our analysis highlighted formal connections between cumulative cultural evolution, Bayesian inference, and stochastic optimization. Our results suggest a more general insight: identifying the algorithm that is implemented by a cultural process can allow us to characterise the computation it performs, yielding a cognitive interpretation of the process in information-processing terms. Our results showed that computation by cumulative culture can be biased. This naturally raises the question: under what circumstances is computation by culture *unbiased*?

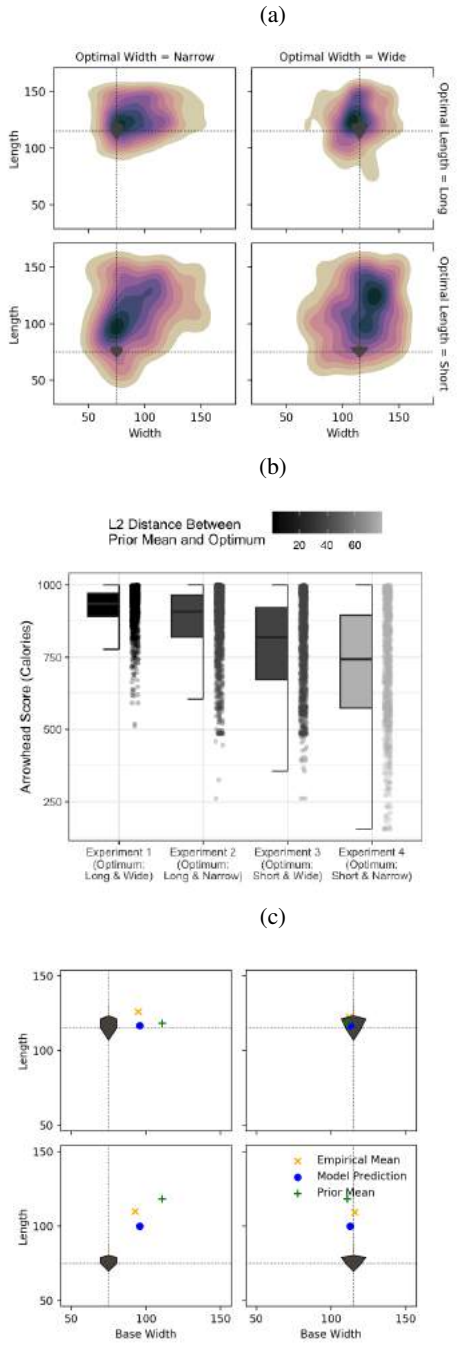


Figure 4: Results of all optimization experiments. Above (a): kernel density estimates of the distribution of arrowheads produced by all participants at generation 3 or later. Dotted lines show the location of the optimum arrowhead in design space. Middle (b): The distribution of scores obtained by the arrowheads produced by participants at generation 3 or later in all optimization chains. The Euclidean distance between the optimum arrowhead and the mean of the prior distribution predicts task success (No. calories). Below (c): Mean arrowhead design produced by all participants at generation 3 or later in all optimization chains (yellow cross), alongside the mean design predicted by our mathematical model (after fitting λ to experimental data, blue circles), the empirical mean of the prior distribution ($\hat{\mu}$, green plus), and the experiment-specific optimum design (black arrowhead).

Acknowledgements

This work was funded in part by NSF grant 1456709 and DARPA Cooperative Agreement D17AC00004.

References

Acerbi, A., & Mesoudi, A. (2015). If we are all cultural Darwinians what's the fuss about? Clarifying recent disagreements in the field of cultural evolution. *Biology & Philosophy*, *30*(4), 481–503. doi: 10.1007/s10539-015-9490-2

Acerbi, A., Tennie, C., & Mesoudi, A. (2016). Social learning solves the problem of narrow-peaked search landscapes: experimental evidence in humans. *Royal Society Open Science*, *3*(9), 160215. doi: 10.1098/rsos.160215

Boyd, R., & Richerson, P. J. (1985). *Culture and the Evolutionary Process*. Chicago, IL: University of Chicago Press.

Caldwell, C. A., & Millen, A. E. (2008). Experimental models for testing hypotheses about cumulative cultural evolution. *Evolution and Human Behavior*, *29*(3), 165–171. doi: 10.1016/J.EVOLHUMBEHAV.2007.12.001

Claidière, N., Scott-Phillips, T. C., & Sperber, D. (2014). How Darwinian is cultural evolution? *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *369*(1642), 20130368. doi: 10.1098/rstb.2013.0368

Claidière, N., & Sperber, D. (2007). The role of attraction in cultural evolution. *Journal of Cognition and Culture*, *7*(1), 89–111.

Derex, M., Beugin, M.-P., Godelle, B., & Raymond, M. (2013). Experimental evidence for the influence of group size on cultural complexity. *Nature*, *503*(7476), 389–391. doi: 10.1038/nature12774

Griffiths, T. L., & Kalish, M. L. (2007). Language evolution by iterated learning with bayesian agents. *Cognitive science*, *31*(3), 441–80. doi: 10.1080/15326900701326576

Griffiths, T. L., Kalish, M. L., & Lewandowsky, S. (2008). Theoretical and empirical evidence for the impact of inductive biases on cultural evolution. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *363*(1509), 3503–14. doi: 10.1098/rstb.2008.0146

Henrich, J. (2015). *The secret of our success: how culture is driving human evolution, domesticating our species, and making us smarter*. Princeton University Press.

Henrich, J., & Boyd, R. (2002). On Modeling Cognition and Culture: Why cultural evolution does not require replication of representations. *Journal of Cognition and Culture*, *2*(2), 87–112. doi: 10.1163/156853702320281836

Heyes, C. (2018). Enquire within: cultural evolution and cognitive science. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *373*(1743), 20170051. doi: 10.1098/rstb.2017.0051

Mandt, S., Hoffman, M. D., & Blei, D. M. (2017). Stochastic Gradient Descent as Approximate Bayesian Inference.

Mesoudi, A. (2016). Cultural evolution: integrating psychology, evolution and culture. *Current Opinion in Psychology*, *7*, 17–22. doi: 10.1016/J.COPSYC.2015.07.001

Mesoudi, A., Chang, L., Murray, K., & Lu, H. J. (2015). Higher frequency of social learning in China than in the West shows cultural variation in the dynamics of cultural evolution. *Proceedings. Biological sciences*, *282*(1798), 20142209. doi: 10.1098/rspb.2014.2209

Mesoudi, A., & Thornton, A. (2018). What is cumulative cultural evolution? *Proceedings. Biological sciences*, *285*(1880), 20180712. doi: 10.1098/rspb.2018.0712

Miton, H., & Charbonneau, M. (2018). Cumulative culture in the laboratory: Methodological and theoretical challenges. *Proceedings of the Royal Society B: Biological Sciences*, *285*(1879), 20180677. doi: 10.1098/rspb.2018.0677

Morin, O. (2016). Reasons to be fussy about cultural evolution. *Biology and Philosophy*, *31*(3). doi: 10.1007/s10539-016-9516-4

Muthukrishna, M., Shulman, B. W., Vasilescu, V., & Henrich, J. (2014). Sociality influences cultural complexity. *Proceedings. Biological sciences*, *281*(1774), 20132511. doi: 10.1098/rspb.2013.2511

Navarro, D. J., Perfors, A., Kary, A., Brown, S. D., & Donkin, C. (2018). When Extremists Win: Cultural Transmission Via Iterated Learning When Populations Are Heterogeneous. *Cognitive Science*. doi: 10.1111/cogs.12667

Whiten, A., Caldwell, C. A., & Mesoudi, A. (2016). Cultural diffusion in humans and other animals. *Current Opinion in Psychology*, *8*, 15–21. doi: 10.1016/J.COPSYC.2015.09.002

Xu, J., & Griffiths, T. L. (2010). A rational analysis of the effects of memory biases on serial reproduction. *Cognitive Psychology*, *60*(2), 107–126. doi: 10.1016/J.COGPSYCH.2009.09.002

Zwirner, E., & Thornton, A. (2015). Cognitive requirements of cumulative culture: teaching is useful but not essential. *Scientific reports*, *5*, 16781. doi: 10.1038/srep16781

Towards a neural-level cognitive architecture: modeling behavior in working memory tasks with neurons

Zoran Tiganj (zorant@bu.edu)
Nathanael Cruzado (nac0005@bu.edu)
Marc W. Howard (marc777@bu.edu)
Center for Memory and Brain
Boston University

Abstract

Constrained by results from classic behavioral experiments we provide a neural-level cognitive architecture for modeling behavior in working memory tasks. We propose a canonical microcircuit that can be used as a building block for working memory, decision making and cognitive control. The controller controls gates to route the flow of information between the working memory and the evidence accumulator and sets parameters of the circuits. We show that this type of cognitive architecture can account for results in behavioral experiments such as judgment of recency, probe recognition and delayed-match-to-sample. In addition, the neural dynamics generated by the cognitive architecture provides a good match with neurophysiological data from rodents and monkeys. For instance, it generates cells tuned to a particular amount of elapsed time (time cells), to a particular position in space (place cells) and to a particular amount of accumulated evidence.

Keywords: Cognitive architecture; Neural-level modeling; Working memory; Cognitive control; Decision making; Judgment of recency; Probe recognition; Delayed-match-to-sample

Introduction

Behavioral experiments provide important insights into human memory and decision making. Building neural systems that can describe these processes is essential for our understanding of cognition.

Here we propose a neural-level architecture that can model behavior in different working memory based cognitive tasks. The proposed architecture is composed of biologically plausible artificial neurons characterized with instantaneous firing rate and with the ability to: 1) gate information from one set of neurons to the other (Hasselmo & Stern, 2018; Bhandari & Badre, 2018; Sherfey, Ardid, Miller, Hasselmo, & Kopell, 2019) and 2) modulate the firing rate of other neurons via gain modulation (Salinias & Sejnowski, 2001). The architecture is based on a canonical microcircuit that represents continuous variables via supported dimensions (Shankar & Howard, 2012; Howard et al., 2014). The microcircuit is implemented as a two-layer neural network. The same microcircuit prototype is used for maintaining a compressed memory timeline, evidence accumulation and for controlling the flow of actions in a behavioral task. Here we demonstrate that this architecture can be used for modeling behavioral responses and neural activity in a variety of working memory tasks.

A neural architecture for cognitive modeling

We sketch a neural cognitive architecture and apply it to three distinct working memory tasks. The architecture is composed of multiple instances of a canonical microcircuit (Figure 1). This microcircuit represents vector-valued functions over variables. These functions can be examined through attentional gain field and then used to produce a vector-valued output. We first discuss the properties of the microcircuit.

Function representation in the Laplace domain

The microcircuit consists of two layers. The first layer approximates the Laplace transform of $\mathbf{f}(t)$ (a vector across the input space) via set of neurons which can be described as leaky integrators $\mathbf{F}(t,s)$, with a spectrum of rate constants s. Each neuron in $\mathbf{F}(t,s)$ receives the input and has a unique rate constant:

$$\frac{d\mathbf{F}(t,s)}{dt} = \alpha(t)\left[\pm s\mathbf{F}(t,s) + \mathbf{f}(t)\right], \qquad (1)$$

where $\alpha(t)$ is an external signal that modulates the dynamics of the leaky integrators. If $\alpha(t)$ is constant, $\mathbf{F}(t,s)$ codes the Laplace transform of $\mathbf{f}(t)$ leading up to the present. It can be shown that if $\alpha(t) = dx/dt$, $\mathbf{F}(t,s)$ is the Laplace transform with respect to x (Howard et al., 2014). We assume that the probability of observing a neuron with rate constant s goes down like $1/s$. This implements a logarithmic compression of the function representation.

The second layer $\tilde{\mathbf{f}}(t,\overset{*}{x})$ computes the inverse of the Laplace transform using the Post approximation. It is implemented as a linear combination of nodes in $\mathbf{F}(t,s)$: $\tilde{\mathbf{f}}(t,\overset{*}{x}) = \mathbf{L}_k^{-1}\mathbf{F}(t,s)$. The operator $\mathbf{L}_k^{-1}$ approximates kth derivative with respect to s. Because $\mathbf{L}_k^{-1}$ approximates the inverse Laplace transform, $\tilde{\mathbf{f}}(t,\overset{*}{x})$ provides an approximation of the transformed function. It turns out (Shankar & Howard, 2012) that the width of the activity of each unit in $\tilde{\mathbf{f}}(t,\overset{*}{x})$ depends linearly on its value of $\overset{*}{x}$ with a Weber fraction that is determined by the value of k.

Accessing the function

The representation described above stores working memory as a vector-valued approximation of a function over an internal variable. We assume that this entire function cannot be accessed all at once, but that one can compute vector-valued integrals weighted by attentional gain over the function. The microcircuit includes an attentional gain function $\mathbf{G}(\overset{*}{x})$ that is

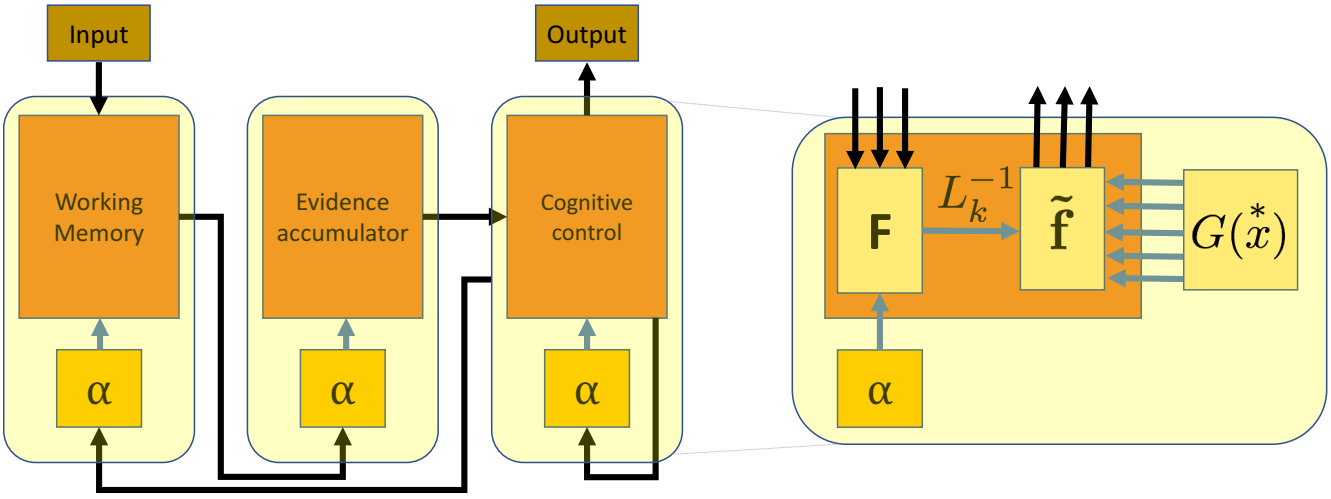


Figure 1: **A schematic of a neural-level circuit that can be used to model different behavioral tasks. This circuit was used to implement all the tasks described here.** The diagram on the left-hand side displays a configuration of the circuit composed of three blocks: working memory, evidence accumulator and cognitive control. The cognitive control block executes a sequence of actions. While a particular action is executed (e.g. waiting for a probe) the sequence is paused by setting its own α to 0. To move to the next action α is set to ± 1. Some actions will access working memory and feed the memory output to the evidence accumulator (e.g. to compare the probe with the content of memory). Output of the evidence accumulator is sent to the cognitive control block where it is used to trigger an appropriate action (e.g. press the left button). Each of the three blocks on the left-hand side is implemented with the microcircuit shown on the right-hand side. The microcircuit takes a vector input (fed into $\mathbf{F}(t,\overset{*}{x})$) and outputs a vector of the same size (through $\tilde{\mathbf{f}}(t,\overset{*}{x})$) selected by the attentional gain field $\mathbf{G}(\overset{*}{x})$ (multiple arrows from $\mathbf{G}(\overset{*}{x})$ represent that it can select different $\overset{*}{x}$ from $\tilde{\mathbf{f}}(t,\overset{*}{x})$). Depending on the initialization and inputs, this multipurpose microcircuit can run a predefined sequence in a self-modulating manner (by modulating its own α), store a compressed memory representation through sequential activation in $\tilde{\mathbf{f}}(t,\overset{*}{x})$ or encode functions of variables (e.g. accumulated evidence) for which a temporal derivative is available.

externally controllable. The output of the microcircuit at any moment is:

$$\mathbf{O}(t) = \sum_{i=1}^{N} \mathbf{G}(\overset{*}{x}_i)\tilde{\mathbf{f}}(t,\overset{*}{x}_i), \tag{2}$$

where N is the number of values of $\overset{*}{x}$ used to implement the function approximation $\tilde{\mathbf{f}}$. In models used here we restrict $\mathbf{G}(\overset{*}{x})$ to be unimodal across $\overset{*}{x}$. Attentional gain field can be made narrow and then activated sequentially, allowing a scan of the function representation or it can be made broad to sum across the $\overset{*}{x}$. This enables one to construct cognitive models based on scanning (e.g., Hacker, 1980) or to construct global matching models (e.g., Donkin & Nosofsky, 2012).

Working memory: Functions of time

When $\alpha(t)$ is constant, $\tilde{\mathbf{f}}$ maintains an estimate of $\mathbf{f}(t)$ as a function of time leading up to the present and we write $\tilde{\mathbf{f}}(t,\overset{*}{\tau})$. If the input stimulus was a delta function at one point in the past, the units in $\tilde{\mathbf{f}}(t,\overset{*}{\tau})$ activate sequentially with temporal tuning curves that are broader and less dense as the stimulus becomes more temporally remote (Figure 2A). Neurons with such properties, called time cells, have been observed in mammalian hippocampus (MacDonald, Lepage, Eden, & Eichenbaum, 2011) and prefrontal cortex (Tiganj, Kim, Jung, & Howard, 2017). Furthermore, different stimuli trigger different sequences of cells (Tiganj et al., 2018), Figure 2B. Taken together at any time t, $\tilde{\mathbf{f}}(t,\overset{*}{\tau})$ can be understood as a compressed memory timeline of the past. The application of the Laplace transform in maintaining working memory in neural and cognitive modeling has been extensively studied (e.g., Shankar & Howard, 2012; Howard, Shankar, Aue, & Criss, 2015).

Evidence accumulation: Functions of net evidence

In simple evidence accumulation models, the decision variable is the sum of instantaneous evidence available during the decision-making process. In these models, a decision is executed when the decision variable reaches a threshold. By setting $\alpha(t)$ to the amount of instantaneous evidence for one alternative, we can construct the Laplace transform of the net amount of decision variable since an initialization signal was sent via the input $f(t)$. If no new evidence has been observed at a particular moment then $\frac{dF(t,s)}{dt} = 0$, thus all the units remain active with sustained firing rate. Large amount of evidence will, on the other hand, mean a fast rate of decay. Inverting the transform results in a set of cells with receptive fields along a "decision axis" (Howard, Luzardo, & Tiganj, 2018) consistent with recent findings from mouse recordings (Morcos & Harvey, 2016).

Cognitive control: Functions of planned actions

The program flow control activates a sequence of actions necessary for completion of a behavioral task. For instance, a typical behavioral task may consist of actions such as attend-

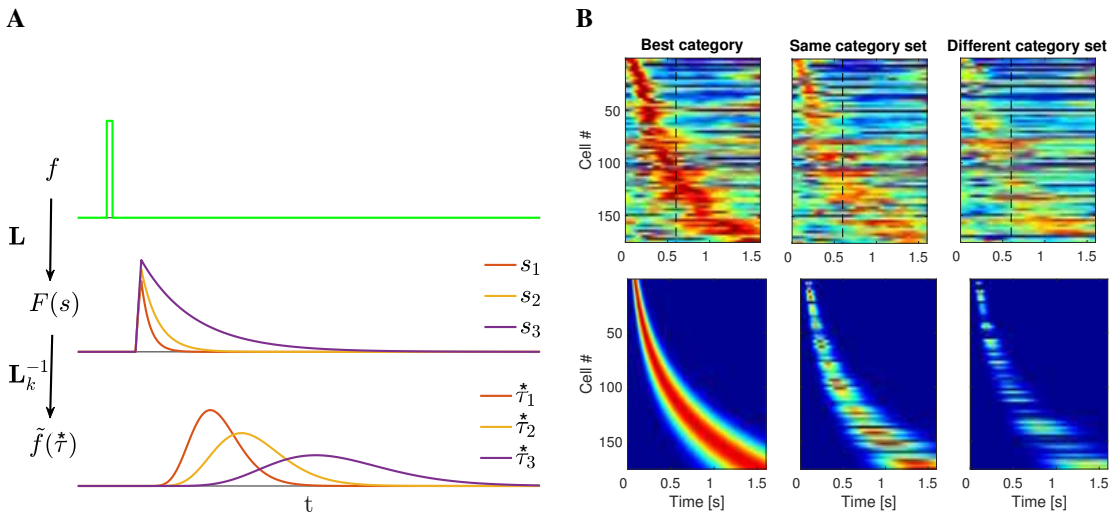


Figure 2: **A scale-invariant compressed memory representation through an integral transform and its inverse: model and neural data**

A. A response of the network to a delta-function input. Activity of only three nodes in each of the two layers is shown. Nodes in $\tilde{f}(\overset{*}{\tau})$ activate sequentially following the presentation of input stimulus f. The width of the activation of each node scales with the peak time determined by the corresponding $\overset{*}{\tau}$, making the memory scale-invariant. Logarithmic spacing of the $\overset{*}{\tau}$ makes the memory representation compressed. **B.** Top: During DMS task sequentially activated cells in monkey lPFC encode time conjunctively with stimulus identity (firing rate encodes visual similarity of the stimuli - stimuli in "Best category" were visually more similar to stimuli in the "Same category set" than to stimuli in the "Different category set"). The three heatmaps show neural activity during the stimulus presentation (first 0.6 s) and the delay period (following 1 s) averaged across trials. (Taken from Tiganj et al. (2018)). Bottom: Activity of the units in the working memory block of the architecture resembles the neural data.

ing to stimuli, detecting the probe, accumulating evidence and taking an appropriate action depending on which of the available choices accumulated more evidence. These operations require the ability to route information to and from the working memory and evidence accumulation modules. For instance, in order to compare a probe to the content of memory, one might route the output of the working memory unit, filtered by a probe stimulus, to the $\alpha(t)$ of an evidence accumulation unit. Because various operations take place in series, we can understand them as a function of future planned actions. Rather than past stimuli, the vectors in $\mathbf{F}(t,s)$ and $\tilde{\mathbf{f}}(t,\overset{*}{\tau})$ can be understood as operations that affect other units (each action has a corresponding two-layer network turning $\mathbf{F}(t,s)$ and $\tilde{\mathbf{f}}(t,\overset{*}{\tau})$ into vectors across the action space).

Different cognitive models correspond to different initial states in $\mathbf{F}(t,s)$ and $\tilde{\mathbf{f}}(t,\overset{*}{\tau})$. The actions will be executed sequentially by setting $\alpha(t) < 0$, winding the planned future closer and closer to the present. For instance, if the first step of a behavioral task is to wait for a probe, then that action will set the controller's $\alpha(t)$ to 0 until the probe is detected. Once the probe is detected, $\alpha(t)$ will be set to a default value of -1 so the neurons in the first layer will grow exponentially and the sequence loaded in $\tilde{\mathbf{f}}(t,\overset{*}{\tau})$ will continue evolving.

Integrating microcircuits into cognitive models

The three blocks described above: working memory, evidence accumulation and cognitive control are all constructed from the same microcircuit (Figure 1 right-hand side). Each circuit has an input, α and output. To demonstrate the utility of this approach, we connected the three blocks such that the program control block gates information from the working memory block to the evidence accumulation block and monitors its output (Figure 1 left-hand side).

Results

We demonstrate performance of the proposed architecture on three classical behavioral tasks: Judgment of Recency (JOR), probe recognition and Delayed-Match-to-Sample (DMS). We compare the results of the model with behavioral data (for JOR and probe recognition) and neural data (for DMS). Critically, even though these three tasks have very different demands, the neural hardware for the models is identical. The only difference is in the initial state of the program block. After initialization, each model runs autonomously and is self-contained.

Judgment of Recency: Sequential scanning of the memory timeline

In JOR subjects are presented with a random list of stimuli (e.g. letters or words) one at a time, and then probed with two stimuli from the list and asked which of the two stimuli was presented more recently. The classical finding is that the time it takes subjects to respond depends on the recency of the more recent probe, but not the recency of the less recent probe (Figure 4A) (Hacker, 1980; Singh & Howard, 2017). This result is consistent with a self-terminating backward scan along a temporally organized memory representation, suggesting

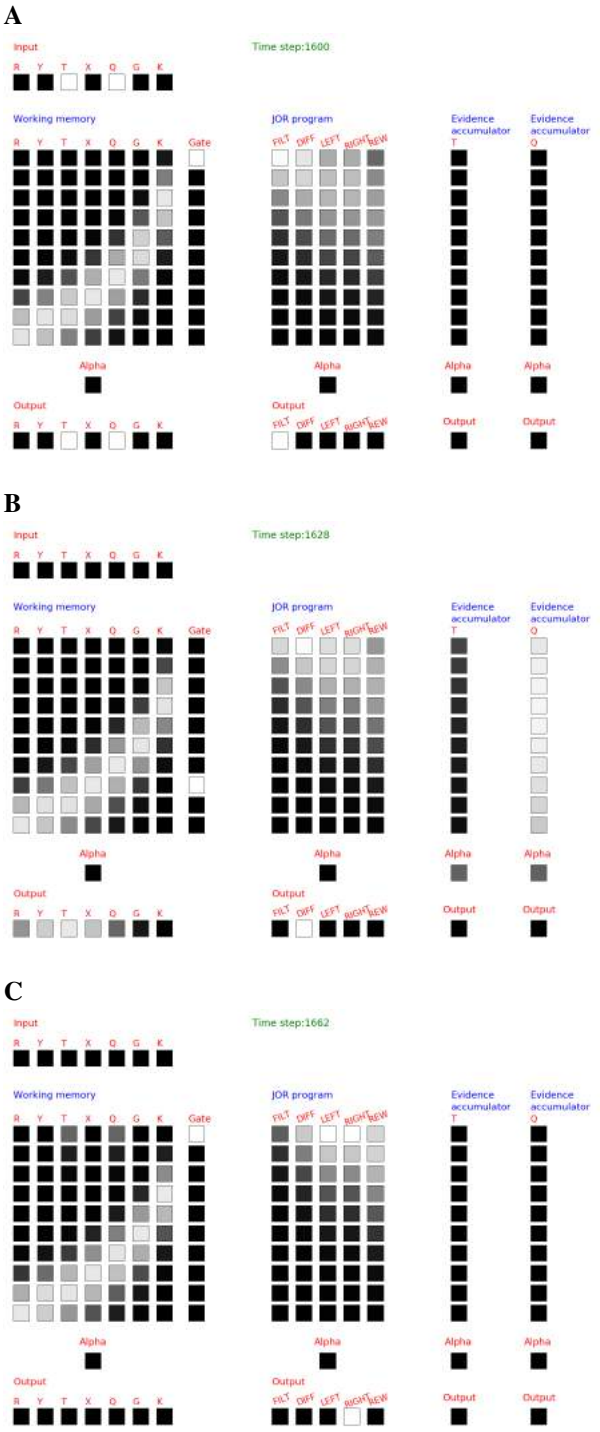


Figure 3: **Example of a JOR task implemented with the proposed architecture.** The implementation is done with microcircuits that correspond to those in Figure 1. Each square corresponds to a single neuron. Squares in the middle layer of each panel correspond to single neurons from $\tilde{\mathbf{f}}(t, \overset{*}{x})$ (neurons from $\mathbf{F}(t,s)$ are not shown). Shading reflects the activity of the neuron at a given time step; darker shading means less activity. **A.** At this time step all the seven items from the test list have been presented and they are stored in the sequentially activated memory. The two probe items T and Q are at the input. **B.** The program (cognitive control) block sequentially gates the information from the working memory into the α neuron of the evidence accumulator (DIFF action in the program block), causing sequential activation in the accumulator. **C.** After the evidence accumulator reaches the threshold, program control continues execution by activating an appropriate action (in this case RIGHT).

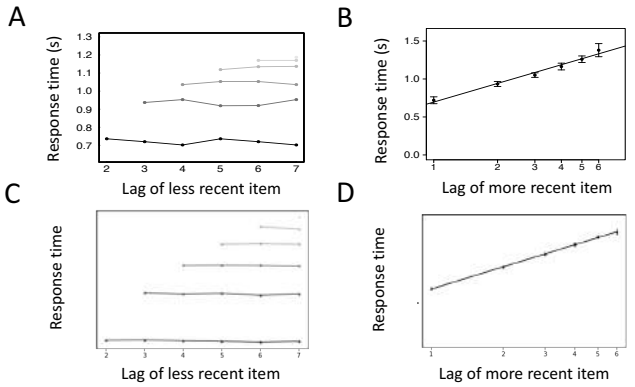


Figure 4: **The model captures behavioral results in the JOR task. A.** In JOR, median response time for correct responses depends strongly on the recency of the more recent probe but not the recency of the less recent probe. Shade of the line denotes lag of the more recent item, with the most recent item shown in black and the most distant item shown in the lightest shade of gray. (From Singh and Howard (2017).) **B.** In JOR, median response time varies sublinearly with recency (x-axis is log-spaced). **C.,D.** Results of the model corresponding to **A** and **B** respectively.

that subjects maintain working memory as a temporally organized, scannable representation. Moreover, the response time is a sublinear function of the lag (Figure 4B) (Singh & Howard, 2017), suggesting that the working memory representation is log-compressed, as proposed by earlier modeling work (Howard et al., 2015; Brown, Neath, & Chater, 2007).

In the model of JOR, the first action was to wait for the probe item to appear (Figure 3A). After that, the gain field over $\overset{*}{\tau}$ was set to scan the memory representation sequentially from more recent towards more distant past. At each step, the value found in the memory was used to drive two evidence accumulators, one independent accumulator for each probe item (Figure 3B). Once one of the two evidence accumulators reached a threshold, the program executed an appropriate action (left or right choice, Figure 3C). Variability in the response times was obtained by adding Gaussian noise to the evidence accumulation process.

Results in Figure 4C indicate that the model captures well the aspect of the data that suggests sequential scanning (Figure 4A): response time depends on the lag of the more recent probe item and does not depend on the lag of the more distant probe item. In addition, the model is consistent with the data regarding compression of the memory representation (Figure 4B - data, Figure 4D - model): the response time grows with the lag of the more recent item.

Old-new probe recognition: Global matching model using the memory timeline

Similarly to JOR, in old-new probe recognition task subjects are presented with a random list of stimuli one at a time. After the list is presented subjects are probed with a single probe that was or was not an item from the list. Subjects choose either *Old* or *New* to indicate their memory. The well-established behavioral results indicate that the response time

increases and accuracy decreases with increasing lag of the probe item (Figure 5A). In other words, if the probe item was further in the past (had larger lag) subjects will take longer to respond and their accuracy will be lower than if the probe was presented less far in the past. Models based on global matching, such as EBRW have managed to capture subjects accuracy and response times (Donkin & Nosofsky, 2012; Nosofsky, Little, Donkin, & Fific, 2011).

Our implementation of probe recognition was similar to JOR, but with several important differences. The main difference between the two tasks was in the way the memory was accessed. Unlike in the implementation of JOR where $G(\overset{*}{x})$ was a delta function resulting in serial scanning, in probe recognition task $G(\overset{*}{x})$ was uniform. This means that the entire memory representation was accessed simultaneously, rather than sequentially scanned. This type of memory access falls under the umbrella of global matching models which includes e.g. EBRW, SAM, Minerva and TODAM (Raaijmakers & Shiffrin, 1980; Murdock, 1982; Hintzman, 1988; Nosofsky et al., 2011).

Figure 5B shows model performance in probe recognition. The two qualitative features observed in the data were captured with the model: response time increased and accuracy decreased as the lag of the probe item increased. Overall, the result of the model resembles the data reported by Donkin and Nosofsky (2012).

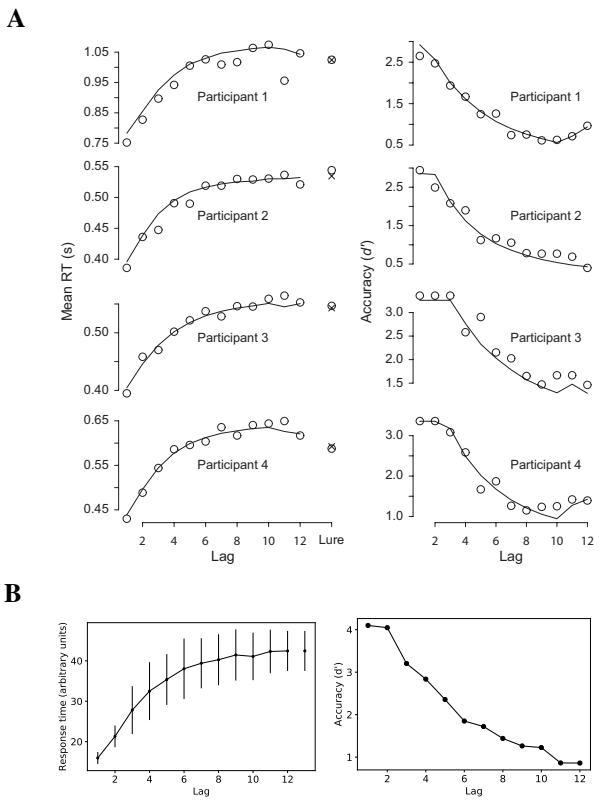


Figure 5: **The model captures behavioral results in the probe recognition task. A.** In probe recognition response times increase and accuracy decreases as the lag of a probe item increases. Circles correspond to data points and solid line is a fit obtained with EBRW model. Taken from (Donkin & Nosofsky, 2012). **B.** Results of the model capture qualitative properties of the data. Response times are shown with standard deviation.

Delayed-Match-to-Sample: Comparing model neurons to empirical evidence for conjunctive coding of what and when

In DMS subjects are presented with a sample stimulus followed by a delay interval, followed by a test stimulus. The action that subjects need to take (e.g. pressing a left or right button) depends on whether the two stimuli were the same or different. We modeled the task with the same components as the JOR task. The only differences were in 1) how the probe item was set (in DMS the second stimulus is by construction the probe, while in JOR the probe is marked by presenting two stimuli at the same time) and 2) what parts of the working memory were gated to the evidence accumulator (in DMS one accumulator accumulated evidence for presence of the probe item in the memory and the other accumulator accumulated evidence that any other item was found in the memory, while in JOR each of the two probe items had its own evidence accumulator). While simple in terms of behavior, DMS task is often done on animals while recording activity of individual neurons. Neural recordings during the delay period of this task show evidence for existence of stimulus-selective sequentially activated cells (Tiganj et al., 2018) that correspond well to the neural activity produced by the sequential memory used here (Figure 2B).

Conclusions

Here we provided an architecture that is based on realistic neural data and that can account for non-trivial behavior. In particular, the behavioral results of JOR task are consistent with the hypothesis that the subjects are scanning along a compressed timeline. The same architecture was used to model DMS task, resulting in neural representation of working memory that closely corresponds to the neural data. Finally, we have also captured qualitative properties observed in probe recognition task by applying an approach analogous to global matching models, but implemented on a neural-level.

Critically, implementation of all three tasks uses the same neural hardware, differing only in the initial condition of the controller. This work is complementary with ongoing efforts of building cognitive architectures such as ACT-R (Anderson, Matessa, & Lebiere, 1997) and SOAR (Laird, 2012). The distinction of the present work is in its attempt to build such architecture with neuron-like units, similar to Spaun (Eliasmith et al., 2012), but with a different type of neural representation. The present work commits to a specific type of representation: variables are represented as supported dimensions via neural tuning curves, tuned to a particular amount of elapsed time, accumulated evidence or a position in a sequence.

Acknowledgments The authors gratefully acknowledge support from ONR MURI N00014-16-1-2832, ONR DURIP

N00014-17-1-2304 and NIBIB R01EB022864.

References

Anderson, J. R., Matessa, M., & Lebiere, C. (1997). Act-r: A theory of higher level cognition and its relation to visual attention. *Human-Computer Interaction*, *12*(4), 439–462.

Bhandari, A., & Badre, D. (2018). Learning and transfer of working memory gating policies. *Cognition*, *172*, 89–100.

Brown, G. D. A., Neath, I., & Chater, N. (2007). A temporal ratio model of memory. *Psychological Review*, *114*(3), 539-76.

Donkin, C., & Nosofsky, R. M. (2012). A power-law model of psychological memory strength in short- and long-term recognition. *Psychological Science*. doi: 10.1177/0956797611430961

Eliasmith, C., Stewart, T. C., Choo, X., Bekolay, T., DeWolf, T., Tang, Y., & Rasmussen, D. (2012). A large-scale model of the functioning brain. *Science*, *338*(6111), 1202–1205.

Hacker, M. J. (1980). Speed and accuracy of recency judgments for events in short-term memory. *Journal of Experimental Psychology: Human Learning and Memory*, *15*, 846-858.

Hasselmo, M. E., & Stern, C. E. (2018). A network model of behavioural performance in a rule learning task. *Phil. Trans. R. Soc. B*, *373*(1744), 20170275.

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in multiple-trace memory model. *Psychological Review*, *95*, 528-551.

Howard, M. W., Luzardo, A., & Tiganj, Z. (2018). Evidence accumulation in a laplace domain decision space. *Computational Brain and Behavior*, *1*, 237–251.

Howard, M. W., MacDonald, C. J., Tiganj, Z., Shankar, K. H., Du, Q., Hasselmo, M. E., & Eichenbaum, H. (2014). A unified mathematical framework for coding time, space, and sequences in the hippocampal region. *Journal of Neuroscience*, *34*(13), 4692-707. doi: 10.1523/JNEUROSCI.5808-12.2014

Howard, M. W., Shankar, K. H., Aue, W., & Criss, A. H. (2015). A distributed representation of internal time. *Psychological Review*, *122*(1), 24-53.

Laird, J. E. (2012). *The Soar cognitive architecture*. MIT press.

MacDonald, C. J., Lepage, K. Q., Eden, U. T., & Eichenbaum, H. (2011). Hippocampal "time cells" bridge the gap in memory for discontiguous events. *Neuron*, *71*(4), 737-749.

Morcos, A. S., & Harvey, C. D. (2016). History-dependent variability in population dynamics during evidence accumulation in cortex. *Nature Neuroscience*, *19*(12), 1672–1681.

Murdock, B. B. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review*, *89*, 609-626.

Nosofsky, R. M., Little, D. R., Donkin, C., & Fific, M. (2011). Short-term memory scanning viewed as exemplar-based categorization. *Psychological Review*, *118*(2), 280-315.

Raaijmakers, J. G. W., & Shiffrin, R. M. (1980). SAM: A theory of probabilistic search of associative memory. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 14, p. 207-262). New York: Academic Press.

Salinias, E., & Sejnowski, T. (2001). Gain modulation in the central nervous system: Where behavior, neurophysiology, and computation meet. *Neuroscientist*, *7*, 430–440.

Shankar, K. H., & Howard, M. W. (2012). A scale-invariant internal representation of time. *Neural Computation*, *24*(1), 134-193.

Sherfey, J. S., Ardid, S., Miller, E. K., Hasselmo, M. E., & Kopell, N. J. (2019). Prefrontal oscillations modulate the propagation of neuronal activity required for working memory. *bioRxiv*. doi: 10.1101/531574

Singh, I., & Howard, M. W. (2017). Recency order judgments in short term memory: Replication and extension of hacker (1980). *bioRxiv*, 144733.

Tiganj, Z., Cromer, J. A., Roy, J. E., Miller, E. K., & Howard, M. W. (2018). Compressed timeline of recent experience in monkey lateral prefrontal cortex. *Journal of cognitive neuroscience*, 1–16.

Tiganj, Z., Kim, J., Jung, M. W., & Howard, M. W. (2017). Sequential firing codes for time in rodent mPFC. *Cerebral Cortex*, *27*, 5663-5671.

Semantic influences on episodic memory distortions

Alexa Tompary (atompary@sas.upenn.edu)
Sharon L. Thompson-Schill (sschill@psych.upenn.edu)
Department of Psychology
University of Pennsylvania, Philadelphia, PA 19104 USA

Abstract

Semantic knowledge can facilitate or distort new memories, depending on their alignment. We aimed to quantify distortions in memory by examining how category membership biases new encoding. Across two experiments, participants encoded and retrieved image-location associations on a 2D grid. The locations of images were manipulated so that most members of a category (e.g. birds) were clustered near each other, but some were in random locations. Memory for an item's location was more precise when it was near members of the same category. Furthermore, typical category members' retrieved locations were more biased towards their semantic neighbors, relative to atypical members. This demonstrates that the organization of semantic knowledge can explain bias in new memories.

Keywords: episodic memory; semantic memory; category membership; typicality; distortion

Introduction

Episodic and semantic memory are commonly studied as distinct cognitive phenomena, the former defined as memory for 'personal experiences and their temporal relations' and the latter as memory for the 'meaning of words, concepts, and classification of concepts' (Tulving, 1972). While this distinction has led to important characterizations of both memory systems, it also oversimplifies the complexity in memories that comprise both episodic and semantic elements. In other words, it neglects the critical notion that new experiences are made up of re-combinations of objects, places, and people for which we already have semantic knowledge. We aimed to probe interactions between the two systems by quantifying how semantic knowledge distorts new episodic learning.

Research on schemas, a type of semantic knowledge defined as a structure of associated information (Bartlett, 1932; Ghosh & Gilboa, 2014), sheds some light on how prior knowledge influences new episodic memory formation. The benefit of prior knowledge for episodic memory is widely documented (Bransford & Johnson, 1972; Alba & Hasher, 1983). Similarly, the presence of prior knowledge accelerates the integration of novel words into existing memory networks (Coutanche & Thompson-Schill, 2014). However, new encoding can also be biased by prior knowledge, resulting in false memories or confabulation (Warren, Jones, Duff, & Tranel, 2014; Webb, Turney, & Dennis, 2016). Taken together, these findings suggest that whether prior knowledge helps or hinders encoding depends on the match between the old and new information.

One weakness of this work is that the operationalization of prior knowledge often ignores its rich, hierarchical structure (Collins & Loftus, 1975). In such a structure, concepts vary in the similarity of their features, giving rise to categories. Typical category members are defined as items that share the greatest number of features with other members, and thus are the best examples of that category (Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976). As a result, typical items are thought to be more strongly associated with category neighbors, relative to atypical items. The consequences of these strong associations are well documented: typical items are more quickly categorized, more efficiently recognized, and less resistant to disruption by brain damage (Patterson, 2007). Furthermore, the features of typical items are more often attributed to category neighbors (Osherson, Smith, Wilkie, López, & Shafir, 1990; Rips, 1975). As an example, if a typical item, rather than an atypical item, is accompanied by a shock, participants are more likely to anticipate shocks with other category members (Dunsmoor & Murphy, 2014). Examining how new memories are formed in the context of this structure may lead to a better understanding of the interactions between episodic and semantic memory.

One promising approach to examining such interactions is by considering retrieval as a construction of different sources of information. According to this view, retrieval is not a veridical recapitulation of past events, but instead an imperfect recombination of event-specific details and other knowledge (Addis, Pan, Vu, Laiser, & Schacter, 2009). Because episodic memories are often noisy and incomplete, successful remembering is thought to combine these partial representations with knowledge from prior experiences (Huttenlocher, Hedges, & Vevea, 2000). Integrating prior knowledge with episodic memories can thus be thought of as a way to improve the 'signal' of a memory. Yet, it also introduces systematic errors if there are discrepancies between a new memory and prior knowledge. For example, exposure to semantically related words (e.g., sour, candy, sugar) often produces a false memory for a non-studied word (sweet; Roediger & McDermott, 1995). Such errors are also captured with continuous measures of bias; for example, memory for the color of shapes is biased towards canonical hues (Persaud & Hemmer, 2014), and estimates of the size of fruits and vegetables are biased by both their superordinate and subordinate mean sizes (Hemmer & Steyvers, 2009). However, it is unknown whether other

properties of semantic knowledge, like category typicality, exert similar distortions on new encoding.

We aimed to quantify distortions in episodic memories due to prior knowledge by examining how differences in category typicality bias new memories for item-location associations. In two experiments conducted on Amazon Mechanical Turk (AMT), participants encoded and retrieved image-location associations on a 2D grid. Critically, the locations associated with each image were determined by semantic relatedness ratings, such that most members of the same category (e.g. birds) were located near each other, but some typical and atypical members were located elsewhere. With this design, participants could learn that items from a certain category tended to be located in a certain area as they encoded item-specific locations.

We used a continuous retrieval measure to disentangle biases driven by semantic knowledge from errors due to forgetting. Critically, these two measures varied independently such that memory for an item could be biased towards or away from category neighbors regardless of its precision. In both experiments, we used these measures to test two predictions. First, we predicted more precise memory for items located near category members, relative to those located farther away, which would replicate past observations that new memories can benefit from prior knowledge if they are aligned. Second, for those typical and atypical items located far from category neighbors, we predicted that their *direction* of error would be different, such that retrieval of typical items would be more biased towards category neighbors relative to atypical items. Such a bias would reflect stronger associations between typical category members and their category neighbors. We did not have strong predictions about precision by typicality, except for the critical notion that any observed differences in bias would be independent of differences in precision.

Experiment 1: Stimulus Development

In the first experiment, we developed a data-driven approach to create item-location associations for the memory task. Specifically, we used semantic relatedness ratings from a separate set of participants to define the images' locations and sort them according to their typicality.

Method

Participants 24 participants (23 – 49 years old, 9 female) completed semantic relatedness judgments. The University of Pennsylvania Institutional Review Board (IRB) approved all consent procedures.

Materials Stimuli comprised 70 100x100-pixel color images on white backgrounds (35 animals, 35 objects). Based on pilot data, we selected images with equivalently high recognition across these two superordinate categories.

Odd-Man-Out Procedure On each trial, participants were presented with three images from a superordinate category and were instructed to click on the image that was least similar to the other two. Once an image was chosen, the images faded away and three new images were displayed after a 200-ms interval. Participants were encouraged to respond in 2 – 4 seconds. They were instructed to make their decisions based on many factors, like whether animals belonged to the same family or shared similar habitats, and whether objects served a similar purpose or tended to be in similar locations. Based on prior piloting, participants completed a random sample of 2,620 combinations per superordinate category, of the 6,545 possible combinations (choose 3 of 35). The trials were divided into 20 separate batches, expected to take 12 - 15 minutes each, and participants were given 1 week to complete them. Of the 35 invited to participate, 24 completed it and 3 were excluded.

The responses were used to create similarity matrices for each participant and superordinate category. Starting with a 35 x 35 matrix of zeros, for every trial on which an odd image was chosen, the value for the other two increased by 1. The summed values across all trials were then divided by the number of times the two images appeared in the same trial. Cells in the matrix thus ranged from 0 to 1, with higher values corresponding to greater similarity between the items. We computed split-half correlations as a test-retest reliability measure for each participant (group mean r = .60, SD = .24). The reliability of the 3 excluded participants was >3 SD lower than the group mean (all r's < .04). Matrices from the 21 remaining participants were averaged into a separate matrix for animals and for objects.

Image-Location Associations Each image was paired with a spatial location on a white 600x1200-pixel rectangle with gray gridlines forming a 50x50-pixel grid. The locations were determined by applying multidimensional scaling (MDS) to the similarity matrices from the odd-man-out procedure. Each matrix was projected into two dimensions, where the x and y coordinates of an item determined its location on the grid. Thus the locations of items represented participants' 2D organization of animals and objects.

We then used k-means clustering of these projections to determine the categories within animals and objects that were captured in the 2D locations. The animal and object locations were separately entered into 10 k-means clustering algorithms with 1 to 10 clusters. The optimal number of clusters was chosen by plotting the sum of within-cluster squared error as a function of the number of clusters used in the algorithm. The 'bend' in this elbow plot signifies the fewest number of clusters that minimize the distance between items in the same cluster. This procedure revealed 3 animal categories (birds, mammals, and sea creatures) and 3 object categories (kitchen, tools/personal care, and office). These clusters were used to identify typical and atypical category members. The center of each cluster was defined as the average x and y coordinate of its constituent items. Then the items were sorted by their distance to its center. The closest 20% were labeled 'typical' and the furthest 20% 'atypical'.

Experiment 1: Memory Task

The item-location associations developed in the prior section were used in an episodic memory task. We probed whether the precision of participants' location memory was related to the consistency between an item's spatial location and the locations of its category neighbors, and whether bias was influenced by its category typicality.

Method

Participants There were 25 participants in the experimental group (21 - 65 years old, 9 female) and 35 in the control group (20 - 61 years old, 16 female). The University of Pennsylvania IRB approved all consent procedures.

Materials See Stimulus Development section.

Image-Location Associations The locations paired from each image were derived from semantic relatedness judgments such that category neighbors were clustered together (see Stimulus Development). The locations of the typical and atypical items were manipulated to be inconsistent with the semantic relatedness ratings. Specifically, they were randomly assigned locations closer to one of the other two cluster centers from the same superordinate category (Figure 1A). In total, 42 images were associated with locations consistent with the ratings ('consistent'), and 28 were associated with a random location ('inconsistent). Of the inconsistent items, 14 were typical and 14 were atypical category members. The projections for animals and objects were arranged side-by-side, randomized for each participant (Figure 3A).

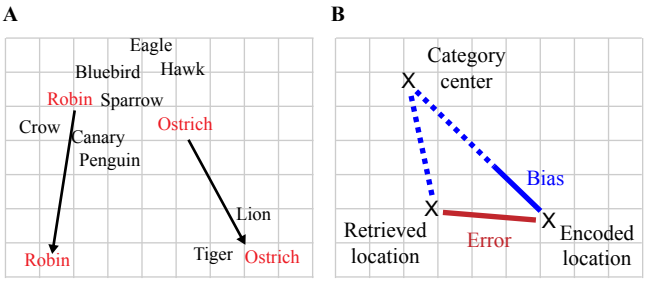


Figure 1 (A) Consistency and typicality for 'birds'. Black indicates 'consistent' and red indicates 'inconsistent' items. Inconsistent items were either typical or atypical category members. (B) Retrieval measures for an item biased towards its category neighbors. Solid red line indicates error. Solid blue line indicates bias.

In the control group, all image-location pairings were randomly shuffled within superordinate category. This group viewed the same locations as the experimental group, but the images assigned to the locations did not cluster by category. In other words, the locations that had originally been associated with (in)consistent or (a)typical images could be associated with any image in that superordinate category, rendering these conditions meaningless.

Memory Procedure The memory experiment comprised an encoding phase and a retrieval phase, separated by a 5-minute break. On each encoding trial, participants viewed an image beneath the grid and a red dot corresponding to that image's location. They were instructed to drag the image onto the dot, click the mouse button once it was positioned over the dot, and memorize its location for a later memory test. Images were presented three times, in three rounds of encoding separated by 1-min breaks. The retrieval task was identical to encoding, but with no dot. Participants were instructed to drag the image to its associated location. The trial order was randomized[1].

Statistical Analyses Two dependent measures were established to quantify error and bias for each image (Figure 1B). Error was defined as the distance between an image's encoded and retrieved location, where greater values indicate less precision. Bias was defined as the relative difference in distance between an item's cluster center and its encoded versus retrieved location: (encoded – center) – (retrieved – center). Thus, values > 0 indicate that retrieval was biased *towards* the cluster center, and < 0 indicate bias *away* from the cluster center. Both measures were averaged across trials by consistency with the relatedness ratings (consistent vs. inconsistent) and by typicality (atypical vs. typical) and entered into two-tailed paired t-tests and repeated measures ANOVAs.

Results

Error We computed a group (experimental, control) x consistency (consistent, inconsistent) ANOVA to examine if memory precision was modulated by the consistency of item locations with those of other category members. This revealed a main effect of group, $F_{(1,58)} = 7.04$, $p = .01$, and consistency, $F_{(1,58)} = 8.46$, $p = .005$. These effects were qualified by an interaction, $F_{(1,58)} = 5.82$, $p = .02$ (Figure 2A), driven by less error for consistent items relative to inconsistent items in the experimental group, $t_{(24)} = 4.11$, $p < .001$, but not the control group, $t_{(34)} = 0.63$, $p = .54$.

We next asked whether, among the inconsistent items, there were differences in precision by typicality. A group x typicality (typical, atypical) ANOVA revealed a main effect of group, $F_{(1,58)} = 4.16$, $p = .046$, but no reliable effect of, or interaction with, typicality (both F's < 2.03, p's > 0.16).

Bias We next asked whether the direction of error differed for typical versus atypical category members. We computed a group x typicality ANOVA amongst the inconsistent items, with bias as the dependent variable (Figure 2B). We found a main effect of group, $F_{(1,58)} = 9.89$, $p = .003$ and typicality, $F_{(1,58)} = 5.46$, $p = .02$, and a group x typicality

[1]Due to a bug, the trial order and locations of the inconsistent items were randomized identically in all participants. Findings from this cohort are reported in this proceeding. After finding the error, we ran a replication experiment (N = 35) where both were randomized individually. All findings were successfully replicated.

interaction, $F_{(1,58)} = 14.12$, $p < .001$. This interaction was driven by greater bias towards category neighbors for typical items relative to atypical items in the experimental group, $t_{(24)} = 6.76$, $p < .001$, but not the control group, $t_{(34)} = 0.55$, $p = .59$.

As predicted, typical items were retrieved as closer to their category neighbors relative to atypical items. It could be the case, however, that this bias was driven by an unrelated difference in how typical and atypical items' locations were retrieved – one possibility is that typical items were retrieved more centrally in the display. To test this possibility, we computed each item's average bias towards the two other clusters in the superordinate category and entered it into a group x typicality ANOVA. There was no main effect of or interaction with typicality (both F's < .34, both p's > .56). This suggests that retrieval of typical items was specifically biased towards category neighbors.

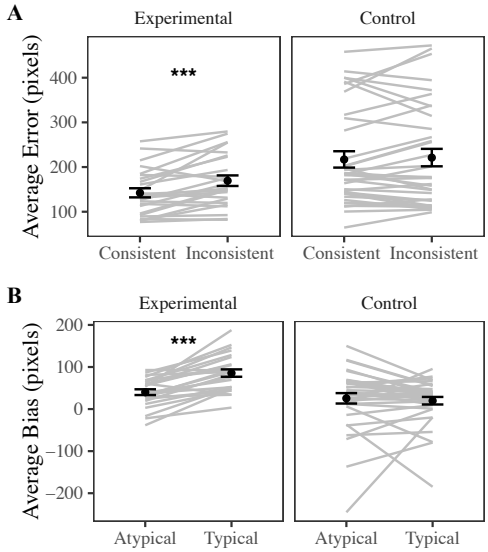


Figure 2 Experiment 1 results. (A) Average error by consistency. (B) Average bias by category typicality. Condition labels in the control group indicate the locations to which (in)consistent and (a)typical items had been assigned in the experimental group; these locations were randomly assigned images in the control group. Lines indicate participants. Error bars signify standard error of the mean (SEM). *** $p < .001$.

Discussion

We found that participants' retrieval was more precise for items located near category neighbors, replicating prior observations of enhancements in memories that are consistent with prior knowledge. Furthermore, of the items that were located far from category neighbors, typical items were more biased towards their category neighbors relative to atypical items, despite no reliable differences in precision. Together, these results suggest that differences in typicality govern the extent of distortion in new memories.

Experiment 2: Stimulus Development

In Experiment 1, we developed data-driven methods to sort items by category typicality and assign them to spatial locations based on their semantic relatedness. We next developed a conceptual replication, using different stimuli, to investigate whether we would observe the same effects with more standard procedures to define category membership and typicality.

Method

Participants 216 participants (27 per category) completed an item ranking procedure. The University of Pennsylvania IRB approved all consent procedures. Demographics were not collected due to experimenter error.

Materials Stimuli comprised 160 100x100-pixel color images on white backgrounds (80 animals, 80 objects). These superordinate categories were divided into 4 categories with 20 images each: birds, insects, sea creatures, mammals, clothes, furniture, kitchen, and office. The categories were selected from prior studies investigating categorization norms (Deyne et al., 2008; Uyeda & Mandler, 1980).

Ranking Procedure We modified a validated item ranking task (Djalal, Ameel, & Storms, 2016) to sort category members by their typicality. Extensive instructions with examples were given to ensure participants understood the sorting procedure. For each category, participants viewed 20 images in a box labeled 'Sort these'. Underneath, there were two empty boxes labeled 'Typical' and 'Atypical'. Participants were instructed to drag 10 images into each box. They were allowed to drag images freely across the three boxes in any order. This resulted in a row of 10 images per box. Then, within each box, participants sorted the 10 images on a scale ranging from most (a)typical to less (a)typical. Arrows and labels in the two boxes indicated the direction that images were to be sorted. The resulting spatial positions in the two boxes were concatenated into a ranked list of category typicality and averaged across participants.

Experiment 2: Memory Task

Results from the ranking task were used in Experiment 2 to define category membership and typicality for a memory task identical to that of Experiment 1. We also aimed to rule out the possibility that memory was more precise for consistent items because they were more densely clustered, increasing the likelihood of guessing the correct location.

Method

Participants 35 participants were in the experimental group (22 - 70 years old, 14 female) and 35 in the control group (24 - 72 years old, 16 female). The University of Pennsylvania IRB approved all consent procedures.

Materials See Stimulus Development section.

Image-Location Associations To generate items' locations, the memory grid was divided into halves with animals on one side and objects on the other, randomized across participants. Each side was divided into four quadrants. Within each quadrant, items from one of the four categories were spaced roughly uniformly apart, resulting in an even distribution of items across the grid (Figure 3B).

As in Experiment 1, some items were assigned locations away from category neighbors. The 15% most typical and 15% most atypical items were swapped across categories, where each quadrant had an equal number of typical and atypical items from the other three quadrants from the same superordinate category. In total, 112 images were located in the quadrant consistent with their category ('consistent'), and 48 in a different quadrant ('inconsistent). Of the 48 inconsistent items, 24 were typical and 24 were atypical category members. In the control group, images were randomly assigned to locations within each superordinate category, identically to the procedure in Experiment 1.

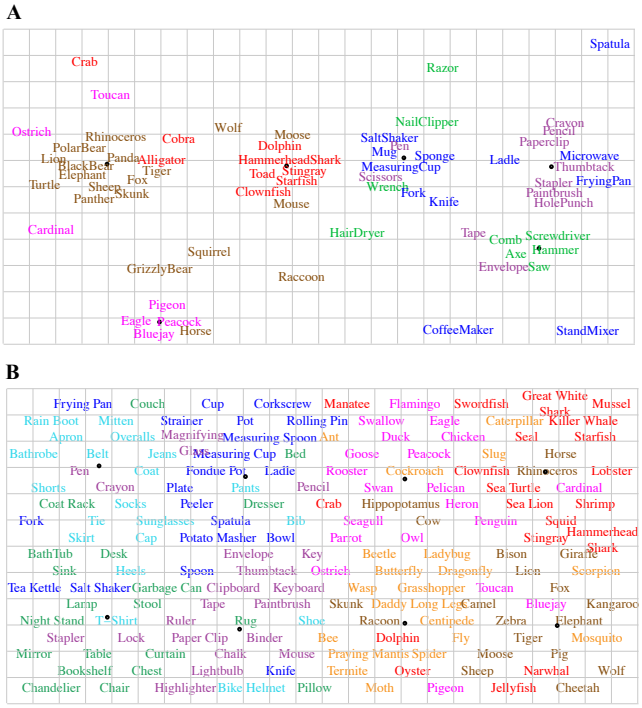


Figure 3 Example memory displays for Experiment 1 (A) and Experiment 2 (B). Each word represents the location of its corresponding image. Colors indicate categories. Black dots indicate a category's cluster center. Each image-location association was presented and tested one at a time.

Memory Procedure The timing and instructions were identical to those of Experiment 1, with an additional confidence measure that will not be discussed.

Statistical Analyses The analyses were identical to those in Experiment 1.

Results

Error We computed a group (experimental, control) x consistency (consistent, inconsistent) ANOVA and found a main effect of consistency, $F_{(1,68)} = 38.63$, $p < .001$, but not of group, $F_{(1,68)} = 0.01$, $p = .91$. Critically, there was also a group x consistency interaction, $F_{(1,68)} = 35.33$, $p < .001$ (Figure 4A). This interaction was driven by less error for consistent items relative to inconsistent items in the experimental group, $t_{(34)} = 7.35$, $p < .001$, but not the control group, $t_{(34)} = 0.24$, $p = .81$. As in Experiment 1, memory precision was modulated by the consistency of an item's location with those of its category neighbors.

We next asked whether, among the inconsistent items, there were differences in error by typicality. A group x typicality (typical, atypical) ANOVA revealed no main effect of group, $F_{(1,68)} = 1.24$, $p = .27$. Interestingly, in contrast to Experiment 1, we found a main effect of typicality, $F_{(1,68)} = 9.65$, $p = .003$, qualified by a group x typicality interaction, $F_{(1,68)} = 5.77$, $p = 0.02$. This interaction was driven by increased error for typical items relative to atypical items in the experimental group, $t_{(34)} = 3.43$, $p = .002$, but not the control group, $t_{(34)} = 0.59$, $p = .56$.

Bias We next focused on differences in bias by typicality and computed a group x typicality ANOVA amongst inconsistent items. We found no reliable main effect of group, $F_{(1,68)} = 1.47$, $p = .23$ and a trending effect of typicality, $F_{(1,68)} = 3.10$, $p = .08$. There was a reliable group x typicality interaction, $F_{(1,68)} = 5.91$, $p = .02$ (Figure 4B). This interaction was driven by greater bias for typical items relative to atypical items in the experimental group, $t_{(34)} = 2.56$, $p = .01$, but not the control group, $t_{(34)} = 0.58$, $p = .56$, replicating the observed difference in bias by typicality in Experiment 1.

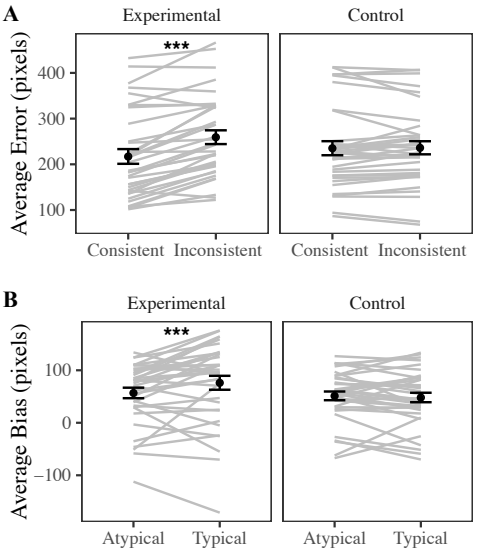


Figure 4 Experiment 2 results. (A) Average error by consistency. (B) Average bias by category typicality. Lines indicate participants. Error bars signify SEM. *** $p < .001$.

General Discussion

Across two experiments, we found that manipulating the match with prior semantic knowledge – by leveraging differences in category typicality – can influence the precision and distortion of new memories. Participants were able to learn associations between an image's category membership and its location on a grid, and this knowledge enhanced their memory of the locations of specific items. Precision of this memory was greater if the items clustered near others from the same category. For items that were located away from category neighbors, participants made systematic errors: typical category members were retrieved closer to category neighbors than atypical category members. These results were observed in two experiments despite differences in the number and type of categories, method of determining typicality, and mapping between category membership and spatial location.

Our findings that consistent items were more precise than inconsistent items (in both experiments), and that precision was greater in the experimental group (in Experiment 1 only), are consistent with a large and diverse body of work showing that prior knowledge facilitates memory for related stimuli (Alba & Hasher, 1983). Our findings extend these results by showing that prior knowledge can improve encoding of new, unrelated features of an item. In our experiment, participants mapped items onto spatial locations on a grid. These locations were not intrinsically related to the items (e.g., nothing about the concept of a 'spatula' implies that it should be located on the top right corner of a grid). However, by associating these locations with the semantic organization of the items, participants treated location as a new 'feature' of items that was explained well by their category membership. Thus, prior knowledge can help to organize the encoding of unrelated contextual details.

When locations did not match expectations, participants' memory was prone to systematic biases. In both experiments, retrieval of typical category members was more biased towards category neighbors relative to retrieval of the atypical category members. While it is well known that memory can be easily distorted (Loftus & Palmer, 1974; Roediger & McDermott, 1995), much of this past work is focused on discrete differences in memory retrieval (e.g. was a word recalled or not). Using continuous reports allows retrieval to be broken down into item-specific error and systematic influences of a particular category or structure (Huttenlocher, Hedges, & Duncan, 1991; Hemmer & Steyvers, 2009; Persaud & Hemmer, 2014). This prior work also demonstrates that new encoding can be biased towards similar stimuli, for example, that memory for the color of an object is biased towards a canonical color. We extended this work by showing that semantic knowledge can exert a stronger or weaker influence on new encoding depending on semantic properties like the typicality of category members, and that such bias can operate independently of memory precision.

What can these biases tell us about how category members are organized; specifically, why is memory for typical members more biased towards neighbors? One possibility is that typical items are more strongly 'pulled' by neighboring items on account of their stronger associations. This interpretation would mirror observations that participants are more likely to cluster the recall of typical items relative to atypical items (Bousfield, Cohen, & Whitmarsh, 1958). Alternatively, because typical items are more similar to other category members, it may be easier to confuse their locations with other item locations that happen to be near the cluster center. This explanation is not specific to category membership but could be applied to any set of memoranda that vary in similarity. Yet another possibility is that because typical items are the closest match to their category, they are more efficiently encoded, but at a cost to in-depth processing of their novel details (Sweegers, Coleman, van Poppel, Cox, & Talamini, 2015) – like their associated location. As we cannot adjudicate between these interpretations with the present design, we have developed follow-up experiments to examine these alternatives.

In summary, we have presented an investigation of the biases that semantic knowledge exerts on episodic encoding. This work demonstrates that semantic knowledge and episodic memory are closely intertwined and offers an opportunity to better understand the interactions between the two systems.

Acknowledgments

This research was funded by an NIH award to Sharon Thompson-Schill (R01 DC009209).

References

Addis, D. R., Pan, L., Vu, M.-A., Laiser, N., & Schacter, D. L. (2009). Constructive episodic simulation of the future and the past: Distinct subsystems of a core brain network mediate imagining and remembering. *Neuropsychologia*, *47*(11), 2222–2238.

Alba, J. W., & Hasher, L. (1983). Is memory schematic? *Psychological Bulletin*, *93*(2), 203–231.

Bartlett, F. C. (1932). A theory of remembering. In *Remembering: A study in experimental and social psychology*. Cambridge University Press.

Bousfield, W. A., Cohen, B. H., & Whitmarsh, G. A. (1958). Associative clustering in the recall of words of different taxonomic frequencies of occurrence. *Psychological Reports*, *4*, 39–44.

Bransford, J. D., & Johnson, M. K. (1972). Contextual prerequisites for understanding: Some investigations of comprehension and recall. *Journal of Verbal Learning and Verbal Behavior*, *11*(6), 717–726.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*(6), 407–428.

Coutanche, M. N., & Thompson-Schill, S. L. (2014). Fast mapping rapidly integrates information into existing memory networks. *Journal of Experimental Psychology: General*, *143*(6), 2296–2303.

Deyne, S. D., Verheyen, S., Ameel, E., Vanpaemel, W., Dry, M. J., Voorspoels, W., & Storms, G. (2008). Exemplar by feature applicability matrices and other Dutch normative data for semantic concepts. *Behavior Research Methods*, *40*(4), 1030–1048.

Djalal, F. M., Ameel, E., & Storms, G. (2016). The typicality ranking task: A new method to derive typicality judgments from children. *PLOS ONE*, *11*(6), e0157936.

Dunsmoor, J. E., & Murphy, G. L. (2014). Stimulus typicality determines how broadly fear is generalized. *Psychological Science*, *25*(9), 1816–1821.

Ghosh, V. E., & Gilboa, A. (2014). What is a memory schema? A historical perspective on current neuroscience literature. *Neuropsychologia*, *53*, 104–114.

Hemmer, P., & Steyvers, M. (2009). Integrating episodic memories and prior knowledge at multiple levels of abstraction. *Psychonomic Bulletin & Review*, *16*(1), 80–87.

Huttenlocher, J., Hedges, L. V., & Duncan, S. (1991). Categories and particulars: Prototype effects in estimating spatial location. *Psychological Review*, *98*(3), 352.

Huttenlocher, J., Hedges, L. V., & Vevea, J. L. (2000). Why do categories affect stimulus judgment? *Journal of Experimental Psychology: General*, *129*(2), 220–241.

Loftus, E. F., & Palmer, J. C. (1974). Reconstruction of automobile destruction: An example of the interaction between language and memory. *Journal of Verbal Learning and Verbal Behavior*, *13*(5), 585–589.

Osherson, D. N., Smith, E. E., Wilkie, O., López, A., & Shafir, E. (1990). Category-based induction. *Psychological Review*, *97*(2), 185–200.

Patterson, K. (2007). The reign of typicality in semantic memory. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1481), 813–821.

Persaud, K., & Hemmer, P. (2014). The influence of knowledge and expectations for color on episodic memory. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society*. Quebec City, Canada.

Rips, L. J. (1975). Inductive judgments about natural categories. *Journal of Verbal Learning and Verbal Behavior*, *14*(6), 665–681.

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(4), 803.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*(3), 382–439.

Sweegers, C. C. G., Coleman, G. A., van Poppel, E. a. M., Cox, R., & Talamini, L. M. (2015). Mental Schemas Hamper Memory Storage of Goal-Irrelevant Information. *Frontiers in Human Neuroscience*, *9*.

Tulving, E. (1972). Episodic and semantic memory. In E. Tulving & W. Donaldson (Eds.), *Organization of Memory*. New York: Academic Press.

Uyeda, K. M., & Mandler, G. (1980). Prototypicality norms for 28 semantic categories. *Behavior Research Methods & Instrumentation*, *12*(6), 587–595.

Warren, D. E., Jones, S. H., Duff, M. C., & Tranel, D. (2014). False recall is reduced by damage to the ventromedial prefrontal cortex: Implications for understanding the neural correlates of schematic memory. *The Journal of Neuroscience*, *34*(22), 7677–7682.

Webb, C. E., Turney, I. C., & Dennis, N. A. (2016). What's the gist? The influence of schemas on the neural correlates underlying true and false memories. *Neuropsychologia*, *93, Part A*, 61–75.

Rapid Presentation Rate Negatively Impacts the Contiguity Effect in Free Recall

Claudio Toro-Serey (ctoro@bu.edu) and **Ian M. Bright (imbright@bu.edu)**
Department of Psychological and Brain Sciences, 64 Cummington Mall
Boston, MA 02215 USA

Brad P. Wyble (bpw10@psu.edu)
Department of Psychology, 140 Moore Building
University Park, PA 16801 USA

Marc W. Howard (marc777@bu.edu)
Department of Psychological and Brain Sciences, 64 Cummington Mall
Boston, MA 02215 USA

Abstract

It is well-known that in free recall participants tend to recall words presented close together in time in sequence, reflecting a form of temporal binding in memory. This contiguity effect is robust, having been observed across many different experimental manipulations. In order to explore a potential boundary on the contiguity effect, participants performed a free recall task in which items were presented at rates ranging from 2 Hz to 8 Hz. Participants were still able to recall items even at the fastest presentation rate, though accuracy decreased. Importantly, the contiguity effect flattened as presentation rates increased. These findings illuminate possible constraints on the temporal encoding of episodic memories.

Keywords: Free recall, lag-CRP, contiguity effect

Introduction

Cognitive neuroscientists have hypothesized that the successful retrieval of an episodic memory is accompanied by a "jump back in time," a recovery of the previous memory's spatiotemporal context (Tulving, 1983). In free recall studies, this recovery manifests as the contiguity effect, wherein following the successful recall of an item, the next item to be recalled is more likely to be a close temporal neighbor than a more distant one (Kahana, 1996). This distance is measured as lag, a directed distance between items in a study list. For example, in the list "absence, hollow, pupil, river, darling", the lag from absence to river is $+3$, while the lag from darling to pupil is -2. In free recall studies the contiguity effect is typically asymmetric, such that forward transitions are more likely to take place than backward transitions of the same distance. This effect is robust, appearing across a variety of methodological manipulations (Kahana, 2012; Healey & Kahana, 2014). For instance, the contiguity effect is observed with more or less the same properties for lists of different modalities (Kahana, 1996), when rehearsal is discouraged (Howard & Kahana, 1999), and when words are widely separated in time (Howard, Youker, & Venkatadass, 2008; Unsworth, 2008). Healey and Kahana (2014) noted that the contiguity effect was observed for every individual participant in a free recall study of 126 subjects. Thus far, dramatic effects on the contiguity effect in free recall have primarily been observed comparing patient populations; older adults and memory disordered individuals show impaired contiguity effects (Kahana, Howard, Zaromb, & Wingfield, 2002; Palombo, Di Lascio, Howard, & Verfaellie, 2019).

Beyond the contiguity effect, free recall contains many other well-explored patterns of behavior. Individuals exhibit a strong recency effect during immediate free recall tests (Glanzer & Cunitz, 1966). In addition, participants exhibit a primacy effect such that items at the beginning of a studied list are more likely to be recalled (B. B. Murdock, 1962). Both primacy and recency effects are observed in the initiation of free recall, and are both also robustly observed in the probability of first recall, a measure of the serial position curve considering only the first recall (Hogan, 1975; Laming, 1999). The relative strength of primacy and recency is not constant however (B. B. Murdock, 1962). For example, Davelaar, Goshen-Gottstein, Ashkenazi, Haarmann, and Usher (2005) found that presentation rates affect the relative strength of primacy and recency, with primacy becoming more prevalent as the presentation rate is increased.

The ubiquity of the contiguity effect in free recall presents something of a challenge for models of memory encoding—if nothing affects the contiguity effect, it makes it more difficult to understand how it comes about. Conversely, if we knew boundary conditions on the contiguity effect it would perhaps shed light on the processes supporting the binding of experiences presented close together in time. In this study we explore the effects of increasing presentation rates on the contiguity effect. If the contiguity effect is disrupted at a particular rate, that suggests the time scale over which the encoding processes necessary for temporal binding take place.

Considerations from the ERP literature and rapid serial visual presentation (RSVP) literature inform the time scale over which contiguity might be disrupted. A to-be-remembered stimulus typically evokes a P300 waveform approximately 500 ms in duration that is thought to represent the updating of memory representations, even when the stimulus duration itself is on the order of 2 seconds (Donchin, 1981). At presentation rates approaching 10 Hz, there is evidence that individual list items are no longer processed as discrete items, and instead are merged into a single extended cognitive event. For example, individual items in 10 Hz lists receive very low hit rates in an immediate recognition test even

when the stimuli are never-before-seen natural images (Potter & Levy, 1969). This poor performance is in stark contrast to the excellent recognition memory for long series of images at slower rates of presentation (Standing, 1973; Brady, Konkle, Alvarez, & Oliva, 2008). However despite this lack of memorability, it is also clear that each item in a 10 Hz stream is processed to some degree, since it is possible to detect specific target items with high probability (Potter, 1976). If the processing of individual items in a list undergoes a qualitative change as the presentation rate is increased to the point at which the representations blend together, then the CRP, primacy effect, and recency effect may be altered. For example, the CRP effect may depend on the ability to place individual items into a discrete temporal representation, and thus it may disappear with faster presentation rates. The probability of first recall could also be altered, since a long-running stream of rapidly presented items imposes a sequential cost on subsequent items due to encoding interference from previous items (Wyble, Bowman, & Nieuwenstein, 2009).

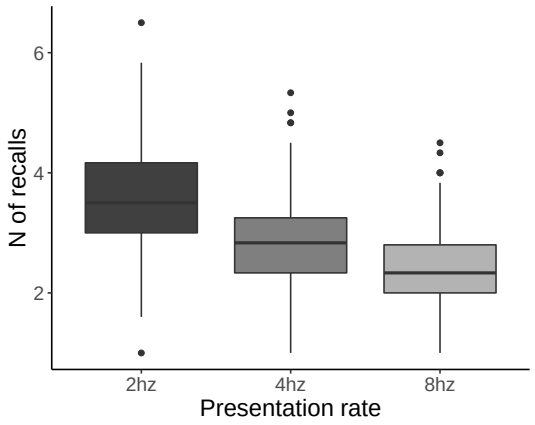


Figure 1: A boxplot of median number of words recalled per trial across participants and presentation rates, with interquartile range, 95% confidence intervals and outliers. Participants recalled fewer words as presentation rate increased.

Methods

Participants

Three hundred and thirty undergraduates from Syracuse University participated in this study. Participants were excluded if they failed to recall a correct word in at least one trial ($n = 15$), and if they did not perform all three conditions ($n = 7$). Data from 308 participants were used in subsequent analyses.

Procedure

Participants took part in 18 trials. Each trial consisted of 20 words from the Toronto Noun Pool (Friendly, Franklin, Hoffman, & Rubin, 1982). Words were visually displayed at three presentations rates: 2 Hz, 4 Hz, and 8 Hz. Participants completed six trials in each condition. Trial order was randomized. Before the start of a trial, participants viewed a bar that discretely rotated at the same rate that words would be presented to help orient them to the upcoming trial (e.g., before a 2 Hz trial the bar would move twice every second). Following the presentation of the list, participants were prompted to verbally recall as many words as possible from the list. Responses were recorded and later parsed using a semi-automatic speech parsing algorithm.

Analysis

We first examined whether presentation rate affected the average number of valid recalls in a trial. This was done with a repeated measures ANOVA. Post-hoc paired permutations (5000 iterations) and Cohen's D effect sizes on mean recalls were then performed to determine significant differences. Serial position curves (SPC) were computed to show the overall probability of a word being recalled based on its position in the list for each participant. We examined whether the recency and primacy effects changed as a function of presentation rate. We performed a paired permutation test in order to predict the difference in the probability of recall for the first and last items in the list (i.e., probability for position 20 minus probability for position 1).

The probability of first recall (PFR) was calculated by dividing the number of times each serial position was recalled first by the total number of first recalls. We then averaged these probabilities across participants per condition. Finally, we calculated the conditional response probability (CRP) for each lag by dividing the number of correct recall transitions at a given lag by the total number of possible correct transitions at that lag. In order to control for serial position effects, which differed across conditions, we restricted the lag-CRP analysis to transitions within the middle of the list where probability of recall was approximately equal across presentation rates. In order to test for differences in the CRP at each lag across conditions, we performed a number of mixed-effects logistic regressions. We estimated the CRP as a function of the interaction between the following fixed-effects predictors: absolute lag, its direction (backwards or forwards from the previously recalled item), and presentation rate. We report Z- and T-scored coefficients for all mixed-effects models.

Results

To anticipate the results, memory performance was reduced at faster presentation rates. We replicated previous findings with respect to changes in the serial position curve at fast presentation rates. Critically, the contiguity effect, even measured at serial positions that avoided contributions from primacy and recency, was severely disrupted at fast presentation rates.

As Presentation Rate Increases, Fewer Words are Recalled

As shown in Figure 1, the total number of words recalled decreased as presentation rates increased (2 Hz: mean = 3.54, SD = 0.85; 4 Hz: mean = 2.86, SD = 0.662; 8 Hz: mean = 2.41, SD = 0.62; ANOVA: $F(2,614) = 309.2, p < 0.001$).

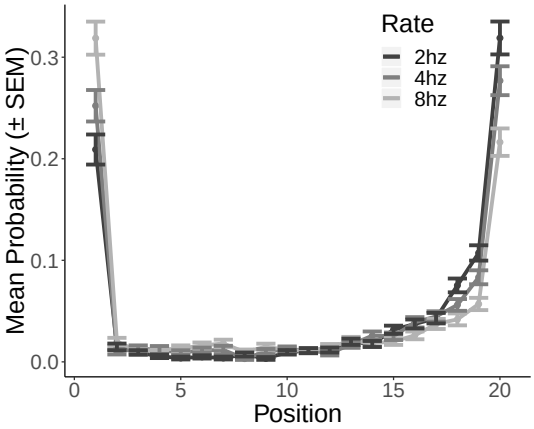


Figure 2: Probability of first recall. Participants tended to begin recall by naming an item from the beginning or end of the list. As presentation rates increased, the probability of initiating recall at the end of the list decreased, and the probability of initiating recall with at the beginning of the list increased.

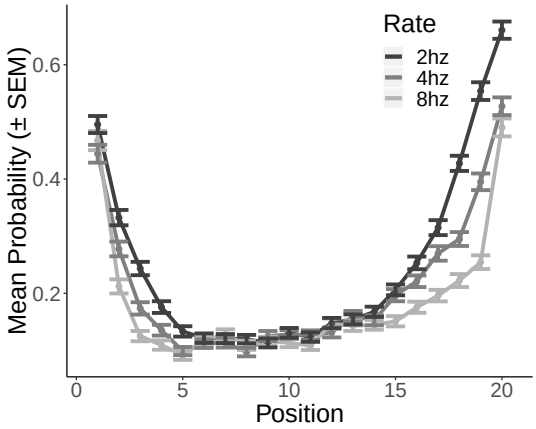


Figure 3: Probability of recall as a function of position in study list. Participants showed the highest level of performance for items at the beginning and end of the list. As presentation rate increased, participants showed a tendency to have a lower recency effect in comparison to the primacy effect.

Post-hoc paired permutations confirmed these results, showing that the presentation rate of 2 Hz yielded significantly higher number of recalls than 4 Hz ($p < 0.001$, Cohen's D $= 0.9$) and 8 Hz ($p < 0.001$, Cohen's D $= 1.5$), and that 4 Hz produced significantly more recalls than 8 Hz ($p < 0.001$, Cohen's D $= 0.69$). This result is consistent with previous findings that faster presentation rates decrease the number of words recalled in a free recall task (B. B. Murdock Jr, 1960).

Increasing Presentation Rates Increases the Primacy Effect and Decreases the Recency Effect

Participants were more likely to begin recall by reporting a word at the beginning or end of the list (Figure 2). As the presentation rate increased, participants initiated recall less frequently at the end of the list and more frequently at the beginning of the list. This was confirmed by paired permutation tests which indicated that the probability of beginning a recall with the first item in a studied list was greater at 8 Hz than both 4 Hz ($p < 0.001$, Cohen's $D = 0.24$) and 2 Hz ($p < 0.001$, Cohen's $D = 0.43$), and greater for 4 Hz than 2 Hz ($p < 0.001$, Cohen's $D = 0.18$). Conversely, the probability of first recalling the last item in a list was greater for 2 Hz than both 4 Hz ($p = 0.002$, Cohen's $D = 0.16$) and 8 Hz ($p < 0.001$, Cohen's $D = 0.40$), and higher for 4 Hz than 8 Hz ($p < 0.001$, Cohen's $D = 0.25$).

As shown in Figure 3, participants showed a higher rate of recalling words from the beginning and end of a list compared to words in the middle (Figure 3). Consistent with previous findings, increasing presentation rates resulted in lower recall for the final item in the list. This was as confirmed by a paired permutation test which found that the probability of recalling the last item in a list was greater for 2 Hz than both 4 Hz ($p < 0.001$, Cohen's $D = 0.50$) and 8 Hz ($p < 0.001$, Cohen's $D = 0.65$), and greater for 4 Hz than 8 Hz ($p = 0.01$, Cohen's $D = 0.15$). In contrast to the PFR, increasing presentation rates did not improve the overall probability of recalling the first item in a list. Rather, it was found that the probability of recalling the first item of a list was greater at 2 Hz than 4 Hz ($p = 0.008$, Cohen's $D = 0.17$), and otherwise there were no significant difference (all $p > 0.05$).

The Contiguity Effect Flattens At Higher Presentation Rates

Figure 4 shows the number of transitions between each serial position in each of the three conditions. The primacy and recency effects can be readily distinguished, as is the tendency to make remote transitions to the beginning of the list. The contiguity effect can be seen as a slightly darker shade along the diagonal; the forward asymmetry appears as a darker shade just above the diagonal. As expected, the contiguity effect appeared to decrease as presentation rate increased. Because primacy and recency effects are a confound in identifying the contiguity effect we calculated the lag-CRP using only transitions that came from items from the middle of the list (serial positions 7-13).

Figure 5 displays the average probability of transitioning from a recalled word to a word at a given lag (with lag 0 corresponding to the diagonal of the matrices in Figure 4), and appears to show a reduction in the temporal contiguity effect as the presentation rate increases. We performed a mixed effect logistic regression to estimate the probability of recall based on absolute lag for each presentation rate separately. This showed that distance from the previously-recalled item significantly decreased the probability of recall for 2 Hz ($z = -9.74, p < 0.001$) and 4 Hz ($z = -4.93, p < 0.001$), but not for 8 Hz ($z = -0.70, p = 0.48$). We then computed another mixed effects logistic regression to test the interaction between absolute lag, its direction (backwards or for-

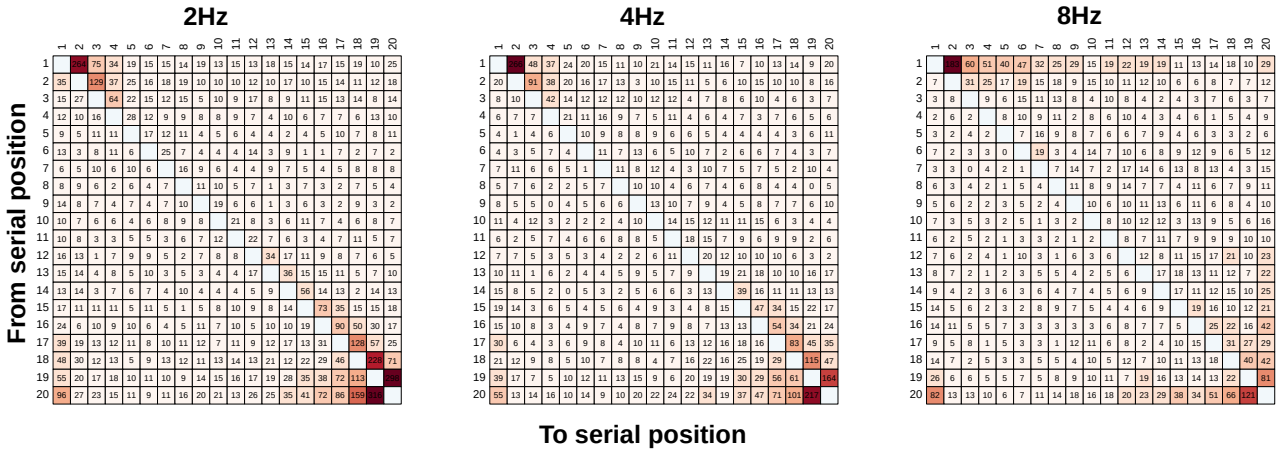


Figure 4: Matrices showing the number of total valid recall transitions between any two study list positions for each presentation rate separately. Colors and numbers correspond to the number of such recalls summed across participants. Transitions between extreme positions in study lists correspond to primacy and recency effects, which persist across rates. In contrast, the likelihood of recalling nearby items (i.e., close to the diagonal) appears to decrease as presentation rate increases.

wards from the previously recalled item), and the presentation rate. This analysis showed that transitions in the forward direction were more probable than backward transitions ($z = 4.18, p < 0.001$); transitions at more distant lags were less probable ($z = -4.83, p < 0.001$); probabilities were higher for 2 Hz compared to 4 Hz ($z = -2.28, p = 0.02$) and 8 Hz ($z = -3.63, p < 0.001$); the effect of absolute lag was stronger for forward transitions than backwards transitions ($z = -2.71, p < 0.01$), and the effect of lag was stronger for 2 Hz compared to 4 Hz ($z = 2.11, p = 0.03$) and 8 Hz ($z = 3.12, p < 0.01$). All other interactions showed no significant effects (all $p > 0.05$). These results show that increasing the presentation rate of studied words decreases the contiguity effect.

Discussion

Remembering past events is associated with a jump back in time, manifesting in a higher probability for temporally contiguous elements to be subsequently recalled. In this study, we investigated whether higher presentation rates would negatively impact the temporal contiguity effect. Many of our results were consistent with previous free recall studies. For instance, the average number of words recalled per list decreased as the presentation rate increased. Also, as the presentation rate increased, the recency effect was diminished. While the primacy effect increased in looking at the probability of first recall, there was not a clear effect on the overall probability of recalling the first item. The novel contribution of this paper is the finding that the temporal contiguity effect was disrupted by fast presentation rates, most notably in the 8 Hz condition. These findings suggest that encoding processes taking place on the order of 125 to 250 ms are im-

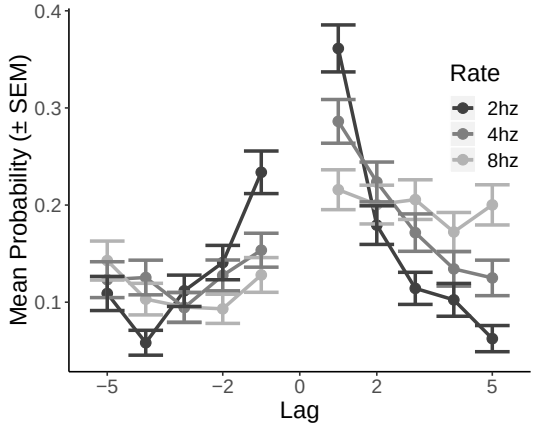


Figure 5: Lag-CRP for transitions restricted to serial positions 7-13 to avoid confounds with primacy and recency effects. As presentation rates increase, the contiguity effect weakens but remains asymmetric. At 8 Hz there is no positive evidence for a contiguity effect.

portant for binding items to their temporal context.

Our results pose questions about the relation of presentation rate and neural coding. Medial temporal lobe theta (3-8 Hz) is related to successful encoding in free recall, particularly when binding elements temporally (Nyhus & Curran, 2010; Sederberg, Kahana, Howard, Donner, & Madsen, 2003). In addition, Guderian, Schott, Richardson-Klavehn, and Düzel (2009) have shown that prediction of successfully-recalled items relies on theta frequency. While presentation rates of 2 Hz and 4 Hz are mostly contained within this frequency band, 8 Hz lies at the upper bound of human theta. It is possible that presening eight words per second outpaces encoding processes that depend on theta (Hasselmo, Bodelón, & Wyble, 2002), thus explaining why lag-CRPs become weaker for this presentation speed. Examination of encoding and retrieval periods using EEG and ECoG could help address this issue in the future.

References

Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences*, *105*(38), 14325–14329.

Davelaar, E. J., Goshen-Gottstein, Y., Ashkenazi, A., Haarmann, H. J., & Usher, M. (2005). The demise of short-term memory revisited: empirical and computational investigations of recency effects. *Psychological Review*, *112*(1), 3-42.

Donchin, E. (1981). Surprise!? surprise? *Psychophysiology*, *18*(5), 493–513.

Friendly, M., Franklin, P. E., Hoffman, D., & Rubin, D. C. (1982). The toronto word pool: Norms for imagery, concreteness, orthographic variables, and grammatical usage for 1,080 words. *Behavior Research Methods & Instrumentation*, *14*(4), 375–399.

Glanzer, M., & Cunitz, A. R. (1966). Two storage mechanisms in free recall. *Journal of Verbal Learning and Verbal Behavior*, *5*, 351-360.

Guderian, S., Schott, B. H., Richardson-Klavehn, A., & Düzel, E. (2009). Medial temporal theta state before an event predicts episodic encoding success in humans. *Proceedings of the National Academy of Sciences*, *106*(13), 5365-5370.

Hasselmo, M. E., Bodelón, C., & Wyble, B. P. (2002). A proposed function for hippocampal theta rhythm: Separate phases of encoding and retrieval enhance reversal of prior learning. *Neural Computation*, *14*, 793-817.

Healey, M. K., & Kahana, M. J. (2014). Is memory search governed by universal principles or idiosyncratic strategies? *Journal of Experimental Psychology: General*, *143*(2), 575.

Hogan, R. M. (1975). Interitem encoding and directed search in free recall. *Memory & Cognition*, *3*(2), 197-209.

Howard, M. W., & Kahana, M. J. (1999). Contextual variability and serial position effects in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 923-941.

Howard, M. W., Youker, T. E., & Venkatadass, V. (2008). The persistence of memory: Contiguity effects across several minutes. *Psychonomic Bulletin & Review*, *15*(PMC2493616), 58-63.

Kahana, M. J. (1996). Associative retrieval processes in free recall. *Memory & Cognition*, *24*, 103-109.

Kahana, M. J. (2012). *Foundations of human memory*. OUP USA.

Kahana, M. J., Howard, M. W., Zaromb, F., & Wingfield, A. (2002). Age dissociates recency and lag-recency effects in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 530-540.

Laming, D. (1999). Testing the idea of distinct storage mechanisms in memory. *International Journal of Psychology*, *34*(5/6), 419-426.

Murdock, B. B. (1962). The serial position effect of free recall. *Journal of Experimental Psychology*, *64*, 482-488.

Murdock, B. B., Jr. (1960). The immediate retention of unrelated words. *Journal of Experimental Psychology*, *60*, 222-34.

Nyhus, E., & Curran, T. (2010). Functional role of gamma and theta oscillations in episodic memory. *Neuroscience and Biobehavioral Reviews*, *34*(7), 1023-35. doi: 10.1016/j.neubiorev.2009.12.014

Palombo, D. J., Di Lascio, J. M., Howard, M. W., & Verfaellie, M. (2019). Medial temporal lobe amnesia is associated with a deficit in recovering temporal context. *Journal of cognitive neuroscience*, *31*(2), 236–248.

Potter, M. C. (1976). Short-term conceptual memory for pictures. *Journal of experimental psychology: human learning and memory*, *2*(5), 509.

Potter, M. C., & Levy, E. I. (1969). Recognition memory for a rapid sequence of pictures. *Journal of experimental psychology*, *81*(1), 10.

Sederberg, P. B., Kahana, M. J., Howard, M. W., Donner, E. J., & Madsen, J. R. (2003). Theta and gamma oscillations during encoding predict subsequent recall. *Journal of Neuroscience*, *23*(34), 10809-14.

Standing, L. (1973). Learning 10000 pictures. *The Quarterly journal of experimental psychology*, *25*(2), 207–222.

Tulving, E. (1983). *Elements of episodic memory*. New York: Oxford.

Unsworth, N. (2008). Exploring the retrieval dynamics of delayed and final free recall: Further evidence for temporal-contextual search. *Journal of Memory and Language*, *59*, 223-236.

Wyble, B., Bowman, H., & Nieuwenstein, M. (2009). The attentional blink provides episodic distinctiveness: sparing at a cost. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(3), 787-807.

The Disappearing "Advantages of Abstract Examples in Learning Math"

Dragan Trninic, Manu Kapur, Tanmay Sinha

ETH Zürich

Abstract

When introducing a novel mathematical idea, should we present learners with abstract or concrete examples of this idea? Considerable efforts have been made over the last decade to settle this question in favor of either abstract or concrete representations. We contribute to this discussion through a critical replication and extension of a well-known study in this area. Whereas the target article argues for the general superiority of abstract representations, we demonstrate that seemingly minor modifications of the study design indicate otherwise. Our results suggest that the previously reported "advantage of abstract examples" manifested not because abstract examples are advantageous in general, but because the earlier studies utilized concrete examples that are pedagogically suboptimal.

Keywords: mathematics education; examples; abstract versus concrete; transfer of learning; replication

Introduction

The use of abstract or concrete representations during mathematics and science instruction has been called a "longstanding controversy" (Fyfe, McNeil, Son, & Goldstone, 2014), and with good reason. Conceptually, we might argue that concrete representations have the advantage of connecting to students' existing knowledge. On the other hand, abstract representations have the advantage of eliminating potentially extraneous perceptual elements. But the elimination of these "extraneous" elements may also reduce the degree to which students can ground a particular representation in their prior knowledge. An advantage of abstract representations is thus at odds with an advantage of concrete representations. Which is better? A review of literature suggests that the answer depends on who asks the question: both pro-concrete and pro-abstract advocates are able to cite research where concrete or abstract representations are more, or less, effective (see, e.g., Koedinger, Alibali, & Nathan, 2008; Schalk, Saalbach, & Stern 2016, for examples).

In this paper, we contribute to the debate through a replication and extension of a well-known and unique study in this area, in particular the central experiment discussed in Kaminski, Sloutsky, and Heckler (2008), "The Advantage of Abstract Examples in Learning Math."

Compared to other papers on the topic, Kaminski et al. is unique in that it makes a universal argument in favor of abstract representations. In particular, the authors argue that "Instantiating an abstract concept in a concrete, contextualized manner… obstructs knowledge transfer. At the same time, learning a generic instantiation allows for transfer" (p. 455). In their study, abstract representations are *in general* superior to concrete representations.

Being a rare mathematics education article published in *Science*, the study caught the attention of not only other scholars, but found recognition in the popular media circuit as well. In a *New York Times* science column, Chang (2008) praised the article, criticized other education researchers for failing to conduct proper research (i.e., "randomized, controlled experiments") and made an even stronger recommendation: "let the apples, oranges and locomotives stay in the real world and… focus on abstract equations." Similar articles appeared in *Le Monde*, *De Standaard*, and elsewhere.

Various elements of the study were criticized over the next few years (see De Bock, Deprez, Dooren, Roelens, & Verschaffel, 2011, for a summary). These criticisms frequently took the form of conceptual disagreements published in math education journals. Despite these conceptual critiques, or perhaps because of them, Kaminski et al. remains steadily cited over the last decade.

The core of the present text is an empirical argument for a more critical re-interpretation of Kaminski et al. In our critical iteration of the experiment, we made relatively minor modifications to the design that nonetheless appear to have had a large impact on the results. We also extended the design to include additional transfer domains, including transfer to a formal mathematical context. The accelerated development of formal knowledge is, after all, a key motivation behind using examples in a math classroom. Our results do not support the hypothesized advantage of abstract examples; on the contrary, they favor the concrete example.

In order to contextualize our own critical replication and extension, we first discuss the central experiment reported by Kaminski et al. (2008), as well as De Bock et al.'s (2011) replication, the first to empirically challenge the original.

Kaminski, Sloutsky, and Heckler (2008)

Kaminski et al. reported a number of experiments drawn from Kaminski's (2006) dissertation. In this paper, we focus on the central experiment, as it forms the foundation of their argument. Here we summarize this experiment, and refer the

reader to Kaminski et al. online supplemental materials for a more extensive description.[1]

The experiment consisted of two phases. In the learning phase, undergraduate students (Ohio, USA) were introduced to a mathematical concept, that of an abstract group of order 3 via rules and examples of these rules. (Briefly, an abstract group of order 3 is a set of three elements and a binary operation that satisfies certain abstract rules—closure, associativity, identity, and inverse. As a consequence of these rules, all mathematical groups of order 3 are isomorphic to each other.) The manipulated variable in the study was whether students were introduced to this mathematical concept via more concrete or more abstract representations.

In the concrete representations condition, participants were provided with three icons of a cup—1/3 full, 2/3 full, and 3/3 full—and rules for combining these cups. In the abstract representations condition, participants were provided with three generic shapes—a flag, a square, and a circle—and rules for combining these shapes. Unbeknownst to the participants, adherence to these rules (in either condition) is mathematically equivalent to operating in an abstract group of order 3.

At the end of the learning phase, a multiple-choice test was administered. The second phase—transfer phase—began immediately after completing this test. There, participants were presented with new, seemingly arbitrary icons of real-world objects (e.g., a vase). Unlike in the learning phase, participants received no explicit training in the transfer domain; they were, however, told that these icons combine in ways structurally identical to the rules they just learned, and provided four examples. Then they answered a series of questions structurally identical to the ones they encountered in the learning phase. The training and the tests were accomplished individually via a computer terminal.

Table 1: Average scores (SD), as a percentage. **A** indicates that the learning phase was conducted with abstract instantiations, **C** with concrete ones.

Condition	Learning	Transfer
A (N = 18)	80 (13.7)	76 (21.6)
C (N = 20)	76 (17.8)	44 (16.0)

See Table 1, above, for a descriptive summary of their results. For now, we note that the "abstract representations" learning condition drastically outperformed the concrete condition on the transfer test: 76% to 44%. This difference is remarkable, all the more so as there were apparently no differences in learning scores or learning times.

[1] The experiments presented are easier to grasp visually. See http://www.sciencemag.org/cgi/content/full/320/5875/454/DC1

De Bock et al. (2011)

In their replication of Kaminski et al., De Bock et al. argued that the transfer domain used by Kaminski et al. is better interpreted as an "abstract transfer" (a terminology we will also use), because it satisfies Kaminski et al.'s own definition of an abstract instantiation. De Bock et al. made the reasonable prediction that, while learning with abstract instantiations may transfer better to an abstract domain, concrete instantiations may transfer better to a concrete domain.

To test this hypothesis, undergraduate students (Belgium) were randomly assigned to one of four conditions:

- AA, abstract learning then abstract transfer
- AC, abstract learning then concrete transfer
- CA, concrete learning then abstract transfer
- CC, concrete learning then concrete transfer.

That is, De Bock et al. kept the two-phase format of the original study, but expanded it to include a transfer to a more concrete domain.

For abstract and concrete learning, and abstract transfer, De Bock et al. used identical materials to Kaminski et al. For concrete transfer, they repurposed one of the alternate concrete learning conditions in the original study—that of a pizza divided in thirds.

Table 2: Average scores (SD), as a percentage. See text, above, for a description of the four conditions.

Condition	Learning	Transfer
AA (N = 23)	71 (16.3)	75 (15.8)
AC (N = 30)	64 (14.6)	73 (17.5)
CA (N = 28)	77 (12.1)	50 (17.9)
CC (N = 24)	76 (14.6)	84 (10.0)

See Table 2, above, for a descriptive summary. In brief, the results confirmed the original findings, as well as De Bock's own hypothesis. In their words: "if transfer to a new abstract domain is targeted, abstract instantiations are indeed more advantageous than concrete instantiations" (p. 120). They continue, "However… the opposite holds as well: Transfer to a new concrete domain is more enhanced by a concrete learning domain than by an abstract one" (p. 120). While not contradicting the original study, De Bock et al. demonstrated that there is more there than meets the eye.

Present Study

We questioned whether the observed "advantage of abstract examples" was due—at least in part—to certain pedagogically suboptimal aspects of the design, which we detail below. To the best of our knowledge, neither De Bock et al., nor anyone else using Kaminski et al.'s materials (e.g., Kaminski, Sloutsky, & Hecker, 2013; McNeil & Fyfe, 2012), attempted to *improve* the materials (as a teacher might). In addition to these pedagogical modification, we also extended the study beyond the original transfer task, as detailed below.

Design modifications and justifications

We identified two aspects of the original materials for improvement. First, the concrete representations used in the main study—those of 1/3, 2/3, and 3/3 liquid-filled cups[2]—caught our attention. The cover story for this instantiation involved combining two or more cups, and trying to determine the "left-over." For instance, 2/3 + 2/3 = 1/3 left-over. Why did Kaminski et al. use the full cup (3/3) as the identity element? The authors presumably used this scheme because it matches our everyday intuition that 1/3 + 2/3 = 3/3. There are at least two issues with this. First, it leads to unintuitive calculations, such as 3/3 + 3/3 = 3/3. More critically, 1/3 + 2/3 = 3/3 is precisely the wrong intuition for mod 3 arithmetic (arithmetic of groups of order 3), because *there* 1 + 2 does not equal 3, but 0.

To us, this suggested that Kaminski's study was not optimized for learning in the concrete condition. An introductory example, we hold, should align not mismatch the superficial concrete elements with the target mathematical structure. Consequently, the concrete representations of cups filled with varying quantities of liquid were modified from 1/3, 2/3, and a full cup (3/3) to 1/3, 2/3 and an empty cup (0/3). This leads to initially surprising but more structurally appropriate 1/3 + 2/3 = 0/3.

Our second concern had to do with the "cover stories" for each of the instantiations. Across these, participants were put into drastically different roles, some believable, others not. These cover stories are as follows (drawn from Kaminski et al. supplementary materials, and Kaminski, 2006):

Abstract instantiation: an archeologist trying to make sense of symbolic combinations left by an ancient civilization.

Concrete (main): an employee at a detergent company calculating the left-over after quantities of liquid are combined.

Concrete (alternative): a pizzeria owner discussing the chef who systemically and persistently burns predetermined portions of every pizza.

Concrete (alternative): an employee at a tennis ball factory dealing with malfunctioning machines producing incorrect quantities of balls.

Transfer: an anthropologist trying to understand a "children's game from another country."

While university students are surely capable of handling nonsense cover stories, such as the one where "the cook systematically burns a portion of each group order," we had concerns about their uneven, varying quality. Specifically, we felt that—pedagogically speaking—the concrete instantiations cover stories were poor in quality, while the generic and transfer narratives impressed us as reasonable. We conjectured that this matters, because a "reasonable" story may be more likely to connect to and activate relevant prior knowledge *without* also being overly distracting. In contrast, a cover story concerning a pizzeria where "the cook systematically burns a portion of each group order" is at odds with any prior knowledge one might have concerning pizzerias, cooking, or business profitability.

A closely related concern has to do with our general sense that the framing of the generic instantiation (an archeological discovery) and the transfer instantiation (a game from another country) had more to do with each other than the concrete instantiations (all of which had to do with odd work).

In response, we made the following modification to the study: every cover story was changed to "a children's game from another country." We generally accept that children play all kinds of games, and recognize that games can involve more concrete instantiations (e.g., combining cups of liquid), or more abstract instantiations (e.g., combining symbols). In other words, this particular cover story was chosen because it naturally accommodates concrete as well as abstract representations.

Because the students in our study would be asked to solve multiple transfer tests rather than one, a compromise was made to remove 4 items from the multiple-choice tests (same 4 from each test); this reduced the number of items on each of the tests from 24 to 20. Specifically, the items removed were 5, 8, 13, and 17 from the original abstract learning instantiation, and all the corresponding items from the other tests. (Of those, items 5 and 8 were chosen for elimination because they were basic and replicated across other questions. Items 13 and 17 were chosen because they used noticeably more text than the other items, a pattern we worried would become apparent across the phases.)

In addition to these modifications to the original study, we extended the study by introducing two additional transfer phases. Similar to De Bock et al.'s study, and for the same reason, we employed a concrete transfer task structurally identical to the original abstract transfer task. While De Bock et al. repurposed the alternative pizza concrete instantiation for this phase, we repurposed the tennis ball factory concrete instantiation.

Finally, we introduced a formal transfer phase, a group of order 5 and consisting of 0, 1, 2, 3, and 4. That is, addition mod 5, formally presented, where 2 + 2 = 4, yet 4 + 2 = 1, 4 + 3 = 2, and so on. We introduced this transfer test to evaluate a particular claim by Kaminski et al., namely that abstract representations lead to superior transfer because they support

[2] Again, we invite the reader to consult the supplementary online materials from the original study. These can be found at: http://www.sciencemag.org/cgi/content/full/320/5875/454/DC1

a deeper understanding of underlying mathematics. But, if students indeed developed a "deep understanding" of groups (or, at least, modular arithmetic), then it stands to reason that they should be able to transfer this knowledge to a formal instantiation of a group of order 5, which shares many similarities with a group of order 3. In this phase, just as in the other transfer phases, participants were not explicitly instructed on the rules, but provided with a few examples and told that the rules of this system are similar to the rules of the previous systems. This phase contained only 11 multiple choice items, focusing on deeper understanding of underlying principles, for example each element having an inverse.

Method

Undergraduate students attending a public university in Switzerland were randomly assigned to one of the (concrete or abstract) learning conditions (38 each; a priori power analysis informed by the original studies indicated that this number was sufficient).[3] Students majoring in mathematics or a computer science field were excluded. Training included explicit training in the rules, with accompanying examples (as in the original study). After this learning phase was completed, participants completed three more transfer phases in the following order: abstract, concrete, and formal. Transfer phases did not include explicit training, but did provide a few examples and inform participants that the rules are "the same" as in the previous tasks (again, as in the original study). To (partially) account for order effects, half the participants in each condition instead completed the phases in the following order: learning, concrete, abstract, formal (formal transfer was always last). No order effects were observed, and the orders are combined for this analysis.

As an illustration, this is what the cover stories and representations looked like for each phase:

Learning, abstract:
In another country, children play a game that involves three symbols: ,  , and  .

Learning, concrete:
In another country, children play a game by combining cups with different quantities of water:  ,  , and  .

Transfer, abstract:
In another country, children play a game that involves these three objects:

 (a ladybug)

 (a vase)

 (and a book).

Transfer, concrete:
In another country, children play a game that involves these three objects:

 - a container with two tennis balls

 - a container with one tennis ball

 - a container with zero tennis balls.

The final phase, formal transfer, did not use a cover story. There, participants were told that they will work with "a number system" and provided with examples of that system.

As in the original study and De Bock's replication, the study was completed individually, on a computer terminal, and there were no breaks during the study. The majority of participants completed the study within an hour, with no one taking more than 75 minutes. The study was conducted by assistants blind to the study expectations.

Reliability analysis for the learning, abstract transfer, concrete transfer, and formal transfer tests yielded McDonald's ω of 0.893, 0.857, 0.888, and 0.856, respectively.

Analysis

No participants were excluded from our analysis. The significance of this is addressed in the Discussion.

For inferential tests (JASP, 2018), Mann-Whitney U test was used as the data were not normally distributed. As is commonplace in education research, we report Cohen's *d*; however, we prioritize the rank-biserial correlation r_B as a more appropriate, unbiased effect size measure.

Results

Table 3, below, provides descriptive statistics for the present study. There were no significant differences in time for completion.

Table 3: Average scores (SD), as a percentage. **A** indicates that the learning phase was conducted with abstract instantiations, **C** with concrete ones. (Note that "Transfer Abstract" in this study corresponds to "Transfer" in previous studies reported in Table 1 and Table 2.)

	Learning	Transfer Abstract	Transfer Concrete	Transfer Formal
A (N = 38)	70 (24.8)	78 (18.3)	90 (14.3)	70 (24.9)
C (N = 38)	95 (12.1)	73 (25.9)	95 (10.7)	78 (27.7)

Comparing concrete to abstract learning conditions, we found a significant difference between the learning scores in

[3] A third condition, a modification of the abstract learning instantiation, was also investigated in the study. As it has no bearing on our current discussion, it is omitted from the analysis.

favor of the concrete learning condition, Mann-Whitney U = 1167.5, $p < .001$, rank-biserial correlation $r_B = 0.617$ with 95% CI [.429, .754] (Cohen's $d = 1.276$). On abstract transfer, we found no difference on performance, U = 701.5, $p = 0.835$, and a very small effect size, $r_B = -0.028$ with 95% CI [-.282, .229] (Cohen's $d = -0.176$). On concrete transfer, we found evidence in favor of the concrete condition, U = 913.5, $p = .033$, and a small-to-moderate effect, $r_B = 0.265$ with 95% CI [.010, .488] (Cohen's $d = 0.396$). Finally, formal transfer favored the concrete condition, but this difference was not significant, U = 877, $p = .103$, $r_B = 0.215$ with 95% CI [-.043, .446] (Cohen's $d = 0.304$).

To check for the influence of outliers, we excluded all participants who scored more than two standard deviations from the mean on any of the tests (same criterion used by Kaminski et al. and de Bock et al.). Three participants were excluded from each condition. Two results were affected. First, the difference on concrete transfer changed from significant to trending in favor of the concrete learning condition, U = 758.5, $p = .064$. Second, the differences on formal transfer reached significance, U = 777, $p = .048$, and a small-to-moderate effect in favor of the concrete learning condition, $r_B = 0.269$ with 95% CI [0.003, 0.499] (Cohen's $d = 0.408$).

Discussion

We aimed to critically replicate and extend an influential study that argued for the advantage of abstract representations in learning mathematics. We made two modification to the original study: (1) using an icon of an empty cup rather than a full cup in the concrete learning condition, and (2) keeping the "cover stories" similar to each other across the tasks. These modifications were made with the intent of removing pedagogically suboptimal elements present in the original design. We also extended the study by included a more concrete transfer task and a formal transfer task. Overall, our results put into question the previously reported advantage of abstract examples.

Whereas Kaminski et al. found no difference in the learning scores, and De Bock's study found a small difference in favor of the concrete instantiation, we found a significant and very large effect in favor of the concrete instantiation. How is it that the concrete instantiation condition in our study performed much higher than participants in the original, and even De Bock's study, on *both* learning and abstract transfer? In the original study, concrete learning to abstract transfer showed 44%, compared to 76% for abstract learning to abstract transfer. In De Bock's study, students fared slightly better, at 50% vs. 75%. In the present study: 73% vs. 78%.

We briefly entertained the (surely self-satisfying) notion that our students are more capable. However, this explanation is unlikely, because our students scored comparatively similar on the other comparable tests, for example across the *abstract* learning condition to abstract transfer (Kaminski: 80%, De Bock: 75%, present study: 78%). This suggests that the concrete instantiation condition performed better because of the changes made to the original materials. But those changes, as detailed earlier, were minor. Of these, we conjecture that using an empty cup rather than a full one may have made the largest difference, as this modification better aligned the concrete representation with the underlying mathematical notion.

As with De Bock et al., we found evidence in favor of the concrete instantiation on the concrete transfer test, although in our case this evidence was not robust.

Furthermore, once outliers were removed, we found evidence in favor of concrete instantiations on the formal transfer test, as well.

An additional point on data analysis may be worth considering. When analyzing our data, we chose to conduct analysis on all the participants, and again after removing those participants scoring more than two standard deviations from the mean. In contrast, the results reported by Kaminski et al. and De Bock et al. (the later following the former), were performed *after* eliminating participants who scored below chance on the learning test, for "failing to learn" (as well as removing the outliers, as we did). This is an unusual method of removing participants in an educational study, and one not conceptually justified in previous articles. Note that it biases the results in favor of students who found the materials useful in the first place. This is an artificial restriction—imagine a mathematics professor evaluating her teaching but refusing to consider those students who "failed to learn" from her lectures, as determined by a learning test immediately following the lecture.

In our data, this "failure to learn" elimination favored the abstract learning condition, because only in that condition did the students score below chance on the learning test. It did not favor it enough to impact the results, but it suggests that this particular elimination introduces bias in favor of the abstract instantiation. This does not explain the drastic differences between our results and those of previous studies, but it raises a question as to why this particular method was employed in the first place. After all, we researchers are unlikely to eliminate data that favors our predictions.

Limitations

Because our design makes not one but multiple modifications to the original study, further work is needed to identify the impact of each modification, as well as to investigate the potential mechanisms through which these modifications influence the learning process.

Summary and Implications

We made a relatively minor change to the concrete learning instantiation in Kaminski et al., in addition to making the various "cover stories" similar to each other. In turn, we observed results that contradict Kaminski et al., and partially support De Bock et al.

Overall, our findings suggest that, if only one instantiation is to be used, and *for these types of tasks*:

Concrete representations facilitate the initial learning of a mathematical concept better than abstract representations of the same idea.

On an abstract transfer, there is no notable advantage between learning via concrete or abstract representations.

On a concrete transfer, learning via concrete representations is preferable, although this difference is relatively small.

When transferring to a formal domain, learning via concrete representations may be preferable, although this difference is, again, relatively small.

Can the current study make any pedagogical recommendations? De Bock et al. (2011) and Jones (2009) caution, and we concur, that brief interventions of this sort should not be applied directly and uncritically to mathematics classrooms. Seen from that perspective, this study claims no more than the following: concrete instantiations may be more or less useful, depending on their quality and context. To be clear, we do not advocate concrete examples as universally advantageous. We agree with Lampinen and McClelland (2018), who argue that it is not the static qualities of "abstractness" or "concreteness" that are likely to impact learning; rather, learning depends on the *interactive* aspects of the learning environment (see Abrahamson & Trninic, 2015). As such, the existence of universally "ideal" learning examples seems unlikely.

The scholarly value of this study lies instead in its contrast to previous work, which found a significant and large effect in favor of the abstract learning instantiation. Our results provide an alternative explanation for those earlier findings. The "advantages of abstract examples" of Kaminski et al. did not manifest because "abstract examples" are better in general. It was because, in that particular design, the concrete learning condition was suboptimal.

References

Abrahamson, D., & Trninic, D. (2015). Bringing forth mathematical concepts: Signifying sensorimotor enactment in fields of promoted action. In D. Reid, L. Brown, A. Coles, & M.-D. Lozano (Eds.), Enactivist methodology in mathematics education research [Special issue]. *ZDM Mathematics Education, 47*(2), 295–306.

Chang, K. (2008, April 25). Study suggests math teachers scrap balls and slices. *The New York Times.* Retrieved January 22, 2019, from http://www.nytimes.com/2008/04/25/science/25math.html

De Bock, D., Deprez, J., Van Dooren, W., Roelens, M., & Verschaffel, L. (2011). Abstract or concrete examples in learning mathematics? A replication and elaboration of Kaminski, Sloutsky, and Heckler's study. *Journal for Research in Mathematics Education, 42*(2), 109–126.

Fyfe, E. R., McNeil, N. M., Son, J. Y., & Goldstone, R. L. (2014). Concreteness fading in mathematics and science instruction: a systematic review. *Educational Psychology Review, 1*(17).

JASP Team (2018). JASP (Version 0.9)[Computer software].

Jones, M. G. (2009). Examining surface features in context. *Journal for Research in Mathematics Education, 40*, 94–96.

Kaminski, J. A. (2006). The effects of concreteness on learning, transfer, and representation of mathematical concepts. (Doctoral dissertation, Ohio State University, 2006). *Dissertation Abstracts International*.

Kaminski, J. A., Sloutsky, V. M., & Heckler, A. F. (2008a). The advantage of abstract examples in learning math. *Science*, 320, 454–455.

Kaminski, J. A., Sloutsky, V. M., & Heckler, A. F. (2008b). Supporting online material for *The advantage of abstract examples in learning math*. Retrieved January 20, 2019, from http://www.sciencemag.org/cgi/content/full/320/5875/454

Kaminski, J. A., & Sloutsky, V. M. (2013). Extraneous perceptual information interferes with children's acquisition of mathematical knowledge. *Journal of Educational Psychology, 105*(2) 351–363.

Koedinger, K. R., Alibali, M. W., & Nathan, M. J. (2008). Trade-offs between grounded and abstract representations: Evidence from algebra problem solving. *Cognitive Science, 32*, 366–397.

Lampinen, A. K., & McClelland, J. L. (2018). Different presentations of a mathematical concept can support learning in complementary ways. *Journal of Educational Psychology, 110*(5), 664–682.

McNeil, N. M., & Fyfe, E. R. (2012). "Concreteness fading" promotes transfer of mathematical knowledge. *Learning and Instruction, 22*(6), 440–448.

Schalk L., Saalbach H., & Stern E. (2016). Approaches to Foster Transfer of Formal Principles: Which Route to Take? *PLoS ONE, 11*(2): e0148787. https://doi.org/10.1371/journal.pone.0148787

Prosodic cues signal the intent of potential indirect requests

Sean Trott (sttrott@ucsd.edu)
Stefanie Reed (sar046@ucsd.edu)
Department of Cognitive Science, 9500 Gilman Dr.
La Jolla, CA

Victor Ferreira (vferreira@ucsd.edu)
Department of Psychology, 9500 Gilman Dr.
La Jolla, CA

Benjamin Bergen (bkbergen@ucsd.edu)
Department of Cognitive Science, 9500 Gilman Dr.
La Jolla, CA

Abstract

Ambiguity pervades language. One prevalent kind of ambiguity is *indirect requests*. For example, "My office is really hot" could be intended not only as a complaint about the temperature, but as a request to turn on the AC. How do comprehenders determine whether a speaker is making a request? We ask whether the *prosody* of an utterance provides information about a speaker's intentions. In a behavioral experiment, we find that human listeners can identify which of two utterances a speaker intended as a request, suggesting that speakers *can* produce discriminable cues. We then show that the acoustic features associated with an utterance allow a classifier to detect the original intent of an utterance (74% accuracy). Finally, we ask which of these features predict listener accuracy on the behavioral experiment.

Keywords: indirect requests; prosody; language production; language comprehension; inference

Introduction

People often make requests indirectly. For example, "Can you open that window?" is literally a question about the hearer's ability to open the window, but is often intended instead as an implied request for the hearer to open the window. Some indirect requests use a highly conventionalized form (in this example, "**Can you** X?"). But other indirect requests are less conventional, such as "My office is really hot." Indirect requests have been a topic of active research for decades in psycholinguistics (Gibbs, 1979), philosophy (Searle, 1990), cognitive psychology (Holtgraves, 1994), and natural language processing (Perrault & Allen, 1980; Williams et al, 2018) for several reasons. First, they're exceedingly frequent. One study eliciting requests from participants found that over 80% were indirect in some way (Gibbs, 1981). Second, successfully comprehending indirect requests requires the hearer to make inferences about the speaker's intent, using linguistic and other contextual knowledge, potentially involving diverse cognitive systems, which can pose challenges to computational implementations of language comprehension (Briggs, Williams, & Scheutz, 2017). But it still remains to be determined what information human comprehenders use to recover the intended interpretation of a potential indirect request.

Previous work suggests that successfully understanding indirect requests requires the integration of extra-linguistic contextual information. For *conventional* indirect requests, comprehenders can use the form of the utterance as a partial cue to its meaning. Consequently, conventional indirect requests are thought to be easier to understand (Gibbs, 1981), and in some cases the *request* interpretation may even be the default (Gibbs, 1986). But even conventional indirect requests can pose a challenge: the conventionality of a particular form is still dependent on context (Gibbs, 1986), and canonical forms can even lead listeners to *misidentify* intended questions as requests (e.g. "Can you play tennis?"), as has been reported for individuals with anterior aphasia and right-hemisphere brain damage (Hirst, LeDoux, & Stein, 1984).

Less conventional indirect requests, such as "My office is really hot", require the hearer to infer both the speech act (e.g. is it a request?) as well as the intended substance of the request, and are thus thought to incur higher processing costs than their literal, non-request counterparts (Tromp, Hagoort, and Meyer, 2016), as well as more conventional indirect requests (Gibbs, 1981). Successful disambiguation of these utterances may benefit from co-speech gesture and eye gaze (Kelly et al, 1999), as well as a representation of what is mutually known across interlocutors (Gibbs, 1987; Trott & Bergen, 2018).

Finally, indirect requests have proven challenging for machine language understanding. Wizard-of-Oz style experiments show that human speakers continue to use indirect requests when speaking to robots (Briggs, Williams, & Scheutz, 2017), even when those robots demonstrably cannot understand them (Williams et al, 2018). Current state-of-the-art solutions (Briggs, Williams, & Scheutz, 2017) use rules relating utterance forms to contexts to probabilistically derive the intended interpretation of ambiguous utterances like "Can you knock down the red tower?" While these solutions work well for established utterance-context mappings, they could still benefit from an increased understanding of precisely which disambiguating

information is available (e.g. paralinguistic or extralinguistic cues), and which is actively exploited by human comprehenders.

Specifically unexplored to date as a candidate source of disambiguating information, is **prosody**: the intonational, rhythmic, and tonal properties of how an utterance is spoken or signed.

Prosodic Cues for Disambiguation

Previous work on other kinds of linguistic ambiguity has already demonstrated that prosodic cues can provide disambiguating information about a speaker's intent.

For one, prosodic features such as *pitch* and *pause duration* can act as "parsing instructions" for listeners. Using speech synthesis, Beach (1991) modified the pitch and duration of critical regions of sentences involving temporary ambiguity (e.g. whether a noun phrase was functioning as a sentential complement or direct object), and found that participants were able to identify the intended parse without listening to the entire utterance. Similarly, Price et al (1991) found that FM radio newscasters, naïve to the purposes of the experiment, produced marked prosodic cues that aided listeners' comprehension of parenthetical statements, apposition, and prepositional phrase attachment ambiguities. This boost in comprehension may even occur before the ambiguity is encountered, as suggested by differences in the visual scan patterns of listeners tasked with determining which object a speaker was referring to (Snedeker & Trueswell, 2003). Nonetheless, there are still substantial debates about the conditions under which *speakers* reliably produce such cues—some studies (Allbritton et al, 1996; Snedeker & Trueswell, 2003) have found that discriminating prosodic cues disappear in the presence of sufficiently disambiguating contextual information, while others (Schafer et al, 2000; Speer et al, 2011) have found that they persist, and have argued that the failure to find such cues is due to limitations on the elicitation paradigms used (e.g. being non-interactive or having low stakes). Regardless, the evidence shows that when such cues *are* available, listeners improve at identifying the intended syntactic parse—pointing to a clear role for prosodic features in syntactic disambiguation.

There is also a growing body of evidence that prosody helps a comprehender decipher a speaker's pragmatic intentions. Early work (Shriberg et al, 1998) found that including prosodic features from conversational speech (including duration, pause, F0, energy, and speech rate) improved a classifier's ability to categorize utterances by Dialogue Act, above and beyond a model equipped with only statistical word-level features. While these results do not indicate that *human* comprehenders infer a speaker's intentions on the basis of prosodic-level features, they do suggest that such features are, in principle, useful. More recently, Hellbernd & Sammler (2016) asked whether trained human speakers could produce cues that identified the intended speech act of one-word utterances—e.g. producing the word "beer" as a Warning, Criticism, or Suggestion. In a behavioral task, human listeners successfully identified the speaker's intended speech act for 82% of words (and 73% of non-words). The authors also trained a machine learning classifier to categorize speech act using prosodic features (duration, mean intensity, harmonics-to-noise ratio, mean fundamental frequency, and pitch rise), obtaining 92% accuracy for words (and 93% for non-words).

Additional evidence that people use prosody to disambiguate comes from research on irony detection. Listeners were able to identify the presence (or absence) of irony in spontaneously-produced speech from radio shows when presented in auditory, but not written, format (Bryant & Fox Tree, 2002), suggesting that success was at least partially dependent on information contained in the speech signal (though see Bryant & Fox Tree (2005) for further discussion of whether these prosodic features are *global* or *local*, and whether they are uniquely characteristic of irony in particular). More recent studies (Deliens et al, 2018) have confirmed that prosodic features aid in the detection of irony; however, listeners appear to exhibit a speed/accuracy trade-off in the integration of prosodic vs. contextual congruity cues, respectively.

Finally, beyond the level of individual speech acts, prosodic features have been shown to improve the detection of a speaker's attitudinal stance (Pell et al, 2018; Ward et al, 2017; Ward et al, 2018). Features such as speech rate and pitch can also influence judgments about the perceived *politeness* of a speech act, including requests (Caballero et al, 2018), though as has been pointed out, the information conveyed by a given prosodic feature is not necessarily independent from the social-interactional context in which that feature is observed (Wichmann, 2000; Culpeper, Bousfield, & Wichmann, 2003).

Together, these findings indicate that speakers are capable of producing signals whose prosodic features provide information about the intended syntactic parse or pragmatic interpretation. Critically, these signals are reliable enough to be detectable—and useful—to both human and machine comprehenders.

However, the role of prosodic features in signaling the intended interpretation of potential indirect requests is currently unexplored. Do speakers and hearers use prosody to overcome the pragmatic ambiguity intrinsic to the most common way to make requests? We addressed this in the current work through three core questions. First, *can* speakers produce reliable cues to indicate to human listeners whether or not they are making a request? Second, *which* cues do speakers actually produce? And third, are these the same cues that listeners seem to use?

Note that all critical data, as well as the code to reproduce the analyses described below, can be found online at: https://github.com/seantrott/prosody_indirect_requests.

Experiment: Listener Judgments of Intent

In a behavioral experiment, we asked whether speakers can produce reliably discriminable prosodic cues. Specifically,

we asked whether these prosodic cues reliably aid human *listeners* in discriminating the speaker's pragmatic intent. On each trial, participants were given two recordings of the same utterance by the same speaker (e.g. "Can you open that window?", or "My soup is cold"), and were asked to select which of the two utterances was intended as a request. If speakers can produce detectable, reliable cues, then participants should be able to identify which utterance was produced as a request; but if speakers cannot produce such cues, or if the cues they produce are not usable by human listeners, then participants should perform at chance.

Methods

Participants 78 participants, all native English speakers, were recruited from Amazon Mechanical Turk. We aimed to recruit 80 participants, but Mechanical Turk under-sampled to 78 participants. The mean age of our participants was 37 (SD=11), ranging from 20 to 69. 30 identified as female, 45 as male, 2 as non-binary, and 1 declined to answer. Each participant was paid $2 for participating, and the experiment took on average 24 minutes to complete.

Materials We recorded five English speakers (2 male, 3 female). Speakers were given 12 utterances to produce (6 conventional indirect requests of the form "Can you X?", and 6 non-conventional indirect requests of the form "My X is Y"), and were instructed to say each utterance twice—once as a request, and once as a literal question or statement. They were allowed to read over the utterances before speaking. The experiment was implemented using JsPsych (de Leeuw, 2015).

Procedure After completing an audio check, participants were instructed that they would listen to a series of paired utterances. They were told that one member of each pair was always intended as a request, and the other member was not. Their task was to indicate which was the request by selecting one of two buttons (either "First" or "Second", corresponding to the first or second utterance presented).

On each trial, participants heard two utterances, containing the same words and produced by the same speaker, with 1 second of silence following each utterance. The order of the utterances (e.g. whether the request or non-request version came first) was counterbalanced within-speaker using a weighted randomization scheme (e.g. for each *speaker-block*, 6 trials contained the request version first, and 6 contained the non-request version first). After listening to both versions, participants indicated which one they thought was intended as a request via button-press.

Each participant performed 60 trials (12 utterance pairs for each of the 5 speakers), blocked by *speaker*. The order of the trials within each *speaker-block* was randomized, as was the order of *speaker-blocks*.

Results

All statistical analyses were performed in R (R Core Team, 2017), using the *lme4* package (Bates et al, 2015). Random effects structure was determined by beginning with the maximal model, then reducing as needed for model convergence (Barr et al, 2013).

Our first question was whether participants could successfully determine which utterance was intended as the request. To test this, we built a generalized linear mixed effects model, with *response* (First or Second) as the dependent variable, and *correct answer* (First or Second) as a fixed effect, as well as random slopes for the effect of *correct answer* for both subjects and items (as well as random intercepts for both). We compared this full model to a reduced model omitting the fixed effect of *correct answer*, and found that the full model explained significantly more variance [$X^2(1)=24.97$, $p=5.8*10^{-7}$]. In other words, participants were able to discriminate request and non-request utterances at a rate above chance.

We were also interested in which characteristics predicted accuracy on particular items—were participants better at identifying pragmatic intent for certain *forms* (conventional vs. non-conventional), or for certain *speakers*? We used nested model comparisons, with *correct* (Yes or No) as a dependent variable, by-item random slopes for *speaker*, by-subject random slopes for *form*, and random intercepts for both items and subjects, to determine whether *form*, *speaker*, and their interaction explained independent sources of variance in participant accuracy. A model with fixed effects for both *form* and *speaker* explained more variance than a model with *form* alone [$X^2(4)=11.5$, $p=.02$], as well as a model with *speaker* alone [$X^2(1)=5.2$, $p=.02$]. Adding an interaction between *form* and *speaker* explained additional variance [$X^2(4)=14.1$, $p=.007$]. In other words, certain speakers produced more discriminable signals overall, and *conventional* requests were generally easier to identify than non-conventional requests, except in the case of one speaker, "S2" (see *Figure 1*).

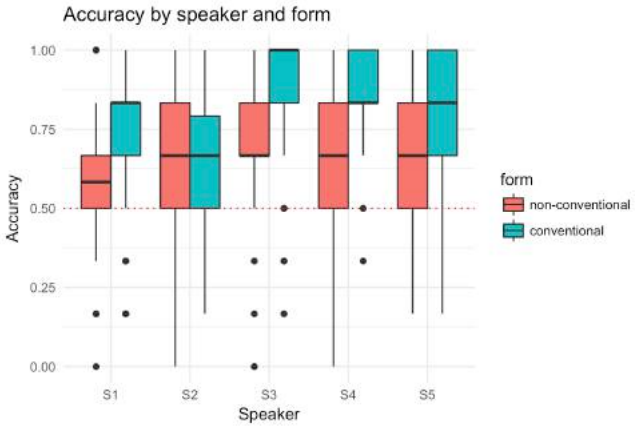


Figure 1: Human accuracy was above chance for all speakers and forms. Accuracy was higher for some speakers (e.g. S3, S4) and some forms (e.g. conventional requests). Dotted red line signifies chance (50%).

One possibility is that participants improved in accuracy over the course of the experiment, perhaps learning which prosodic features signaled intent. We compared a model

with *correct* (Yes/No) as a dependent variable, *Order* (1-60) as a fixed effect, and random intercepts for subjects and items, to a model omitting the fixed effect of *Order*, and found that the full model did not explain significantly more variance [$X^2(1)=.7$, $p=.4$]. Thus, there is no evidence that participants improved over the course of experiment. However, it is also possible that participants improved within each *speaker-block*, but that this adaptation did not carry over across blocks. To test this, we replaced *Order* with *Order-within-block* (1-12) as a fixed effect; a model including *Order-within-block* explained marginally more variance than a model omitting this term [$X^2(1)=3.1$, $p=.08$]. This explanatory power was independent from the variability explained by *speaker*, as determined by comparison of a model including fixed effects of both *speaker* and *Order-within-block* to a model with only *speaker* [$X^2(1)=3.3$, $p=.07$]. Adding an interaction between these factors did not increase explanatory power [$X^2(4)=3.3$, $p=.5$]. This provides weak evidence for within-block adaptation or learning, but requires further analysis and experimentation.

Analysis of Acoustic Features

Listener judgments of pragmatic intent in the behavioral experiment described above demonstrated that speakers produced signals that increased communicative success. However, this analysis does not indicate *which* acoustic features predict a speaker's intended pragmatic interpretation. Here, we asked whether seven acoustic features reliably predicted a speaker's *intent*. Predictive power was assessed in two ways. First, we asked about the explanatory power of each variable in turn using nested model comparisons. Second, we used leave-one-out cross-validation to determine how the combination of *all* features improved the ability of a classifier to identify *intent*.

Data Processing

For each of the 120 recordings (5 speakers producing 12 utterances with two versions each), we used Parselmouth (Jadoul et al, 2018), a Python interface to Praat, to extract the following acoustic features: mean F0, range F0 (max F0 – min F0), standard deviation of F0, duration (number of voiced frames), mean intensity, standard deviation of intensity, and slope of F0 (slope of regressing *F0 ~ time*). We then *z-scored* each of these variables with respect to each speaker's mean and SD, to account for considerable variability in speakers overall.

Results

First, we asked how much independent variance was explained by each feature in turn, comparing a full model (including all seven features) to a model omitting only the feature under consideration. In each case, the full model included *intent* (Request vs. Non-Request) as a dependent variable, fixed effects for each of the seven acoustic features, and random intercepts for each utterance. We adjusted for multiple comparisons using Holm-Bonferroni corrections (Holm, 1979). In each case, a positive coefficient represents a higher likelihood of a *Non-Request*, while a negative coefficient represents a higher likelihood of a *Request*.

For a logistic regression model predicting *intent* of all items (e.g. both *conventional* and *non-conventional* utterances), model fit was improved by including *mean intensity* [$X^2(1)=8.7$, $p=.003$, $p_{adj}=.02$] and *SD intensity* [$X^2(1)=7.8$, $p=.005$, $p_{adj}=.03$]. Higher-intensity utterances were more likely to be Requests [$\beta=-.69$, $SE=.25$, $p=.006$], as were utterances were with greater variation in intensity [$\beta=-1.1$, $SE=.4$, $p=.01$]. No other acoustic features significantly improved model fit after correcting for multiple comparisons.

Because human listener accuracy differed significantly as a function of *form* (see the behavioral experiment), it is possible that distinct prosodic features predict intent for *conventional* and *non-conventional* requests. Thus, we ran the same analysis as above twice: once on only *conventional* and once on only *non-conventional* requests.

For a model predicting *intent* of only *conventional* requests, model fit was improved by including *F0 slope* [$X^2(1)=7.8$, $p=.005$, $p_{adj}=.03$], *SD intensity* [$X^2(1)=7.7$, $p=.005$, $p_{adj}=.03$], and *F0 duration* [$X^2(1)=8.8$, $p=.003$, $p_{adj}=.02$]. More positive slopes were associated with Non-Requests, e.g. literal questions [$\beta=1.1$, $SE=.5$, $p=.01$], as were longer utterances [$\beta=1.1$, $SE=.4$, $p=.01$] and less variation in intensity [$\beta=-1.1$, $SE=.4$, $p=.01$].

For a model predicting *intent* of only *non-conventional* requests, model fit was significantly improved by including *F0 duration* [$X^2(1)=19.6$, $p=9.7*10^{-6}$, $p_{adj}=.00004$], with longer utterances having a higher probability of being Requests [$\beta=-2.5$, $SE=.94$, $p=.008$].

In sum, we identified several acoustic features that predict pragmatic intent. Overall, intent was predicted by *mean intensity* and *SD intensity*. For *conventional* requests in particular, intent was predicted by *F0 slope*, *F0 duration*, and *SD intensity*; for *non-conventional* requests, intent was predicted by *F0 duration*. These results suggest that those features *could*, in principle, be used to identify the intent of an ambiguous utterance.

To determine whether the combination of all seven acoustic features could improve a classifier's ability to detect *intent*, we used leave-one-out cross-validation (LOOCV). A model including all seven acoustic features (as well as their interactions with *form*) accurately predicted *intent* on 74% of the held-out items, a rate substantially above chance (50%).

Predicting Accuracy from Acoustic Features

By regressing pragmatic intent against extracted acoustic features, we isolated multiple features that appear to indicate intent of either conventionally or non-conventionally formatted utterances: F0 slope, F0 duration, mean intensity, and SD intensity. However, this does not entail that listeners actively exploit differences in these features to infer intent. It could be that these features are *statistically* reliable, but

not *psychologically* valid. Which, if any, of these features actually benefit listeners?

One way to test this is to ask: do by-item *differences* in any of the acoustic features explain independent sources of variance in listener *accuracy*, above and beyond the full model specified above in the behavioral experiment (containing an interaction between *form* and *speaker*)? If larger differences from a given dimension (e.g. *F0 slope*) consistently predict accuracy, this suggests that listeners are actively benefitting from those differences, and are thus consistently sampling and deploying information about that particular dimension.

Data Processing

For each utterance pair, we computed the *difference* of each z-scored feature between the Request version and the Non-request version. Thus, a positive value for *F0 slope difference* indicates that the Request version had a larger slope than the Non-request version, while a negative value indicates that the Non-request version had a larger slope. We repeated this procedure for each acoustic feature.

Results

We asked about the informativeness of each acoustic feature (as well as its interaction with *form*) using nested model comparisons. The explanatory power of a given variable was determined by comparing a model including that term to a model without it. We adjusted for multiple comparisons using Holm-Bonferroni corrections (Holm, 1979).

The full model included the terms from the maximal model specified in the behavioral experiment, with *correct* (Yes/No) as a dependent variable, an interaction between *form* and *speaker*, fixed effects for both *form* and *speaker*, and random intercepts for subjects and items. It also included each of the seven acoustic features, as well as their interaction with *form*.

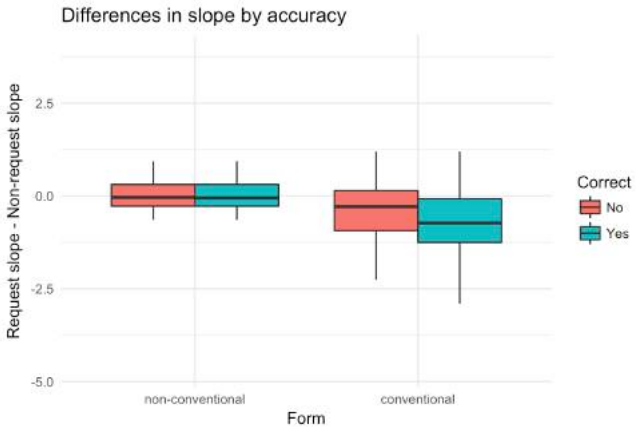


Figure 2: Differences in z-scored F0 slope by *form* and *accuracy*. Conventional items with a larger difference between the Request and Non-Request version (specifically, where the slope on the *Non-Request* version was more positive than the slope on the *Request* version) were more likely to be answered correctly.

Model fit was significantly improved by the interaction between *F0 slope difference* and *form* [$X^2(1)=16.98$, $p=3.78*10^{-4}$, $p_{adj}=.0005$], but was not significantly improved by *F0 slope difference* alone ($p_{adj}>.1$). The direction of this interaction is illustrated in *Figure 2*: accuracy on non-conventional items was not significantly impacted by the difference in *F0 slope* between the Request and Non-Request differences, whereas a larger difference for conventional items predicted more accurate responses. Specifically, conventional items on which the Non-Request version had a more positive slope than the Request version (and thus their difference was more *negative*) were more likely to be answered correctly [$\beta=-.4$, SE=.1, $p=5.5*10^{-5}$].

Model fit was also improved by the interaction between *mean F0* and *form* [$X^2(1)=10.6$, p=.001, $p_{adj}=.01$], as well as the main effect of *mean F0* [$X^2(1)=15.03$, p=.0001, $p_{adj}=.001$]. Specifically, conventional items on which the Request version had a lower *mean F0* than the Non-Request version were more likely to be answered correctly [$\beta=-.47$, SE=.14, p=.001]. Because these comparisons included a term for *F0 slope*, this does not appear to be due simply to conventional Non-Request items exhibiting a sharper final rise (e.g. more positive slope). Differences in *mean F0* explained independent sources of variance from *F0 slope*.

A model including an interaction between *mean intensity* and *form* did not explain more variance than a model omitting that term, but the fixed effect of *mean intensity* did improve model fit [$X^2(1)=9.7$, p=.002, $p_{adj}=.02$]. Specifically, items on which the Request version had a higher overall *mean intensity* than the Non-Request version were marginally more likely to be answered correctly [$\beta=.1$, SE=.05, p=.06].

Model fit was also improved by the interaction between *SD intensity* and *form* [$X^2(1)=7.12$, p=.008, $p_{adj}=02$], though not the fixed effect of *SD intensity* alone ($p_{adj}>.1$). Conventional items on which the Request version exhibited greater variation in intensity than the Non-Request version were more likely to be answered correctly [$\beta=.29$, SE= .11, p=.007].

In summary, four of the acoustic features we extracted predicted listener accuracy—*F0 slope*, *mean F0*, *mean intensity*, and *SD intensity*. *F0 slope* appeared to be useful primarily for conventional requests (with more positive slopes indicating the literal, Non-Request interpretation). *Mean F0* was helpful for both, though again, appeared to be particularly predictive of accuracy on the conventional items (with higher mean F0 on the Non-Request versions predicting higher accuracy). *Mean intensity* was predictive of accuracy on both kinds of items; items on which the Request version exhibited higher overall intensity than the Non-Request version were more likely to be answered correctly. Finally, *SD intensity* was particularly helpful for conventional items—Request versions with more variability in intensity than their Non-Request counterpart were more likely to be correctly identified.

General Discussion

Human listeners were able to discriminate the pragmatic intent of potential indirect requests, indicating that speakers *can* produce discriminable cues, at least when made aware of an utterance's different interpretations. We extracted seven acoustic features from each recorded utterance, and found that four of these features were predictive of listener accuracy in the behavioral experiment: F0 slope, mean F0, mean intensity, and SD intensity. Specifically, larger differences in each of these features were associated with more accurate responses; some were primarily helpful for conventional items (F0 slope, SD intensity, mean F0), while others were helpful for both (mean intensity).

Additionally, using leave-one-out cross-validation, a machine learning classifier trained on these features (and their interaction with utterance *form*) successfully identified the *intent* of potential request utterances 74% of the time (where chance is 50%). Thus, prosodic features are not only useful to human comprehenders attempting to discriminate a speaker's pragmatic intent—they are also informative to machines, suggesting that they could perhaps be integrated into existing natural language understanding architectures (Briggs et al, 2017).

Open questions remain. First, we noted a weak effect of *Order-within-block*, but not *Order* overall, on accuracy. That is, there is no evidence that listeners improved over the course of the entire experiment, but they might have improved while listening to each speaker. If true, this provides weak evidence for *adaptation* to each speaker, which may not successfully carry over across speakers. The effect was marginally significant. Since it arose during exploratory data analysis, it requires further investigation.

Second, a limitation of the behavioral experiment is that participants were asked to explicitly discriminate between two versions of the same utterance (e.g. "which was the request?"), rather than *classifying* an individual utterance (e.g. "is that a request?"). The latter design is clearly more applicable to real-world scenarios, in which comprehenders do not have immediate access to alternative versions of an utterance. We are designing a new set of studies to ask whether comprehenders can identify whether a given utterance was intended as a request, and whether the same acoustic features—e.g. F0 slope, mean intensity, etc.—predict their response. This task design will also allow more direct comparison to the classifier's results, so that we can determine whether the classifier is using similar features (and making similar errors) as human comprehenders.

Third, a long-standing question in the literature on prosody and pragmatic intent is whether particular prosodic features convey direct information about the intended speech act, or whether they function primarily as contrastive markers, which invite the listener to perform additional inference. For example, prosodic features may not directly convey sarcastic intent, but rather prompt listeners to integrate other multimodal, contextual information to recognize irony (Attardo, Eisterhold, Hay, & Poggi, 2003; Bryant & Fox Tree, 2005). Our experiment was not designed to adjudicate between these two possibilities, but our results do suggest that the answer is nuanced, and likely falls somewhere in between. Certain features, such as *F0 slope*, were predictive only of accuracy for conventional forms (E.g. "Can you open that window?"), and thus might be more aptly described as "marking" a deviation from the default interpretation of modal interrogatives as requests (Gibbs, 1986). But other features, such as *mean intensity*, predicted accuracy across forms; in both cases, items with higher intensity on the Request version (vs. the Non-Request version) were more likely to be answered correctly.

Finally, perhaps the most obvious question is whether, or under what conditions, these kinds of prosodic cues would actually be produced. Speakers in our experiment were made aware of the two interpretations of each utterance, and were explicitly asked to produce utterances consistent with those interpretations. While our results indicate that speakers *can* produce discriminable cues, they do not demonstrate that speakers actually *do*. A similar issue arises in the study of prosodic cues for syntactic disambiguation—some (Allbritton, 1996; Snedeker & Trueswell, 2003) have found that these cues are no longer present when the utterance is produced in a disambiguating context, while others (Schafer et al, 2000; Schafer et al, 2005) have argued that the cues are produced regardless of how much information is provided by the context. Thus, the question becomes: are the discriminable prosodic features we observed automatically and conventionally associated with pragmatic intent, or are they deployed strategically for a particular audience in a particular context?

Acknowledgments

We thank Rachel Ostrand for her helpful advice on the modeling of acoustic features. We are also grateful to both the speakers and the participants, and to the reviewers for their suggestions.

References

Allbritton, D. W., McKoon, G., & Ratcliff, R. (1996). Reliability of prosodic cues for resolving syntactic ambiguity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(3), 714.

Attardo, S., Eisterhold, J., Hay, J., & Poggi, I. (2003). Multimodal markers of irony and sarcasm. *Humor*, *16*(2), 243-260.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, *68*(3), 255-278.

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *arXiv preprint arXiv:1506.04967*.

Beach, C. M. (1991). The Interpretation of Prosodic Patterns at Points of Syntactic Structure Ambiguity: Evidence for Cue Trading Relations. *J. of memory and language*, *30*(6), 644.

Briggs, G., Williams, T., & Scheutz, M. (2017). Enabling robots to understand indirect speech acts in task-based interactions. *J. of Human-Robot Interaction*, *6*(1), 64-94.

Bryant, G. A., & Fox Tree, J. E. (2002). Recognizing verbal irony in spontaneous speech. *Metaphor & symbol*, *17*(2), 99-119.

Bryant, G. A., & Fox Tree, J. E. (2005). Is there an ironic tone of voice?. *Language and speech*, *48*(3), 257-277.

Caballero, J. A., Vergis, N., Jiang, X., & Pell, M. D. (2018). The sound of im/politeness. *Speech Communication*, *102*, 39-53.

Culpeper, J., Bousfield, D., & Wichmann, A. (2003). Impoliteness revisited: with special reference to dynamic and prosodic aspects. *Journal of pragmatics*, *35*(10-11), 1545-1579.

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a web browser. Behavior Research Methods, 47(1), 1–12. doi:10.3758/s13428-014-0458-y

Deliens, G., Antoniou, K., Clin, E., Ostashchenko, E., & Kissine, M. (2018). Context, facial expression and prosody in irony processing. *Journal of Memory and Language*, *99*, 35-48.

Gibbs, Jr, R. W. (1979). Contextual effects in understanding indirect requests. *Discourse Processes*, 2(1), 1–10.

Gibbs, R. W. (1981). Your wish is my command: Convention and context in interpreting indirect requests. *J of Verbal Learning and Verbal Behavior*, 20(4), 431–444.

Gibbs, R. W. (1986). What makes some indirect speech acts conventional?. *J. of memory and language*, *25*(2), 181.

Gibbs, R. W. (1987). Mutual knowledge and the psychology of conversational inference. *J. of pragmatics*, *11*(5), 561-588.

Hellbernd, N., & Sammler, D. (2016). Prosody conveys speaker's intentions: Acoustic cues for speech act perception. *Journal of Memory and Language*, *88*, 70-86.

Hirst, W., LeDoux, J., & Stein, S. (1984). Constraints on the processing of indirect speech acts: Evidence from aphasiology. *Brain and language*, *23*(1), 26-33.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, 65-70.

Holtgraves, T. (1994). Communication in context: Effects of speaker status on the comprehension of indirect requests. *J. of Experimental Psychology: Learning, Memory, and Cognition*, *20*(5), 1205.

Kelly, S. D., Barr, D. J., Church, R. B., & Lynch, K. (1999). Offering a hand to pragmatic understanding: The role of speech and gesture in comprehension and memory. *Journal of memory and Language*, *40*(4), 577-592.

Jadoul, Y., Thompson, B., & De Boer, B. (2018). Introducing Parselmouth: a Python interface to Praat. *Journal of Phonetics*, *71*, 1-15.

Pell, M. D., Vergis, N., Caballero, J., Mauchand, M., & Jiang, X. (2018). Prosody as a window into speaker attitudes and interpersonal stance. *The Journal of the Acoustical Society of America*, *144*(3), 1840-1840.

Perrault, C. R., & Allen, J. F. (1980). A plan-based analysis of indirect speech acts. *Computational Linguistics*, *6*(3-4), 167-182.

Price, P. J., Ostendorf, M., Shattuck-Hufnagel, S., & Fong, C. (1991). The use of prosody in syntactic disambiguation. *the Journal of the Acoustical Society of America*, *90*(6), 2956-2970.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Schafer, A. J., Speer, S. R., Warren, P., & White, S. D. (2000). Intonational disambiguation in sentence production and comprehension. *J. of psycholinguistic research*, *29*(2), 169-182.

Searle, J. R. (1990). Indirect Speech Acts 12. *The philosophy of language*, 161.

Shriberg, E., Bates, R., Stolcke, A., Taylor, P.; Jurafsky, D., Ries, K., Coccaro, N., Martin, R., Meteer, M., Van Ess-Dykema, C. (1998). Can Prosody Aid the Automatic Classification of Dialog Acts in Conversational Speech? *Language and Speech*, *341*(4), 443–492.

Snedeker, J., & Trueswell, J. (2003). Using prosody to avoid ambiguity: Effects of speaker awareness and referential context. *Journal of Memory and language*, *48*(1), 103-130.

Speer, S. R., Warren, P., & Schafer, A. J. (2011). Situationally independent prosodic phrasing. *Laboratory Phonology*, *2*(1), 35-98.

Tromp, J., Hagoort, P., & Meyer, A. S. (2016). Pupillometry reveals increased pupil size during indirect request comprehension. The Quarterly Journal of Experimental Psychology, 69(6), 1093–1108.

Trott, S., & Bergen, B. (2018). Individual Differences in Mentalizing Capacity Predict Indirect Request Comprehension. *Discourse Processes*, *00*(00), 1–33.

Ward, N. G., Carlson, J. C., Fuentes, O., Castan, D., Shriberg, E., & Tsiartas, A. (2017). Inferring Stance from Prosody. In *INTERSPEECH* (pp. 1447-1451).

Ward, N. G., Carlson, J. C., & Fuentes, O. (2018). Inferring stance in news broadcasts from prosodic-feature configurations. *Computer Speech & Language*, *50*, 85-104.

Wichmann, A. (2000). The attitudinal effects of prosody, and how they relate to emotion. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*.

Williams, T., Thames, D., Novakoff, J., & Scheutz, M. (2018, February). Thank You for Sharing that Interesting Fact!: Effects of Capability and Context on Indirect Speech Act Use in Task-Based Human-Robot Dialogue. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 298-306).

To Catch a Snitch: Brain potentials reveal knowledge-based variability in the functional organization of (fictional) world knowledge during reading

Abstract

People vary in what they know, yet models of language processing do not take this variability into account. We harnessed the temporal sensitivity of event-related brain potentials alongside individual differences in Harry Potter (HP) knowledge to investigate the extent to which the availability and timing of information relevant for real-time word comprehension are influenced by variation in degree of domain knowledge. We manipulated meaningful (category, event) relationships between sentence contexts about HP stories and critical words (endings), assessed via behavioral ratings and by measuring similarity of word embeddings derived from a high-dimensional semantic model trained on HP texts. Individuals' ratings were sensitive to these relationships according to the degree of their domain knowledge. During reading, N400 amplitudes (neural measures of semantic retrieval) also reflected this variability, suggesting the degree to which information relevant for word understanding is available during real-time sentence processing varies as a function of individuals' domain knowledge.

Keywords: language processing, ERPs, knowledge, individual differences

Introduction

Across cognitive systems, world knowledge allows individuals to organize raw sensation into meaningful experiences. Understanding language is no exception—words cue world knowledge which can be rapidly brought to mind in real time (e.g., Hagoort et al., 2004), incrementally and sometimes even predictively (reviewed in Altmann & Mirković, 2009; Kutas, DeLong, & Smith, 2011). A more precise description of how this occurs—including which types of knowledge, their organization, and the timing of their use—requires a closer look at knowledge availability in real time. It is, however, experimentally challenging to capture the specifics of an individual's world knowledge with standard laboratory procedures.

Troyer and Kutas (2018; Troyer, Urbach, & Kutas, under review) provided a potential solution by focusing on a restricted domain of knowledge with the requisite properties for online language processing studies, including a large, rich set of verbal descriptions, wherein college-aged young adults differed in their degree of knowledge—the fictional world of Harry Potter (HP) by J.K. Rowling. Troyer & Kutas (2018) recorded EEG while participants with varying degrees of knowledge about HP read sentences that described general topics, followed by sentences that described events from the HP stories; sentences ended either in contextually supported or unsupported words. Across participants, and for both sentence types, the effect of contextual support was present on N400 amplitudes—a brain potential sensitive to factors impacting the ease of retrieval from semantic memory, with larger reductions in N400 (i.e., more positive-going potentials) associated with greater ease of retrieval (reviewed in Kutas & Federmeier, 2000). But critically, participants' degree of HP knowledge influenced the size of this effect only for the sentences about HP. More specifically, individuals' HP knowledge was correlated with N400 amplitudes to contextually supported, but not to unsupported, words. These results empirically demonstrate that the rapid influence of written sentence context, known to modulate N400 brain potentials, is a function of each individual's knowledge.

These findings are not surprising given the vast literature showing that people rapidly make use of a variety of word and world knowledge as they understand words in real time, such as orthographic neighborhood density (Laszlo & Federmeier, 2009), word frequency (Van Petten & Kutas, 1990), and non-linguistic knowledge including the organization of categories in semantic memory (Federmeier & Kutas, 1999), facts about the world (Hagoort et al., 2004), generalized event knowledge (Metusalem et al., 2012), personal preferences (Coronel & Federmeier, 2016), and fictional characters (Filik & Leuthold, 2013). It stands to reason that the structure and organization of individuals' knowledge would have consequences for the availability, contents, and timecourse of bringing to mind these varied sources of knowledge in real time.

One way to ask whether and, if so, when people bring different types of information to mind as they read sentences is to probe them with words that are linguistically anomalous, yet systematically related to the sentence context and/or a likely upcoming, linguistically licensed word. This related anomaly paradigm has been fruitfully employed to investigate the influence of the functional organization of semantic memory on sentence processing. For example, in sentence contexts setting up an expectation for the word *pines*, categorically related words (e.g., another type of tree, *palms*) elicited reduced N400 amplitudes compared to words from a different category (e.g., *tulips*), but which were larger than those to the expected word (Federmeier & Kutas, 1999, 2002). In a different study, where individuals read short paragraphs about common events (e.g., playing football) that set up linguistic expectations for a word (e.g., *touchdown*), unexpected and linguistically unlicensed words related to the event being described (e.g., *helmet*) also elicited reduced N400 amplitudes compared to unrelated words (e.g., *license*) (Metusalem et al., 2012). It is worth noting that the "related anomaly" in Federmeier & Kutas shared many features with an expected word whereas in Metusalem et al. the related anomaly was related in one or more of several ways to the generalized event being described in the context, but not did not share features with the linguistically expected word. Nonetheless, the related anomaly ERP effects in both studies had a similar timecourse and scalp topography, maximal around 400 ms over centro-parietal recording sites,

suggesting that people made quick use of both types of related information during real-time sentence processing.

The availability of related/relevant information stored in semantic memory during real-time language processing must, at least to some degree, be modulated by each individual's degree of domain knowledge. Indeed, the literature on expert knowledge proposes that the functional organization of information around themes, events, and categories is likely to depend on individuals' degree of expertise (reviewed in Ericsson et al., 2006). To investigate the extent to which variation in domain knowledge influences the nature and timing of the availability of knowledge stored in long-term memory during real-time sentence processing, we probed semantic memory using a related anomaly paradigm incorporating sentences describing the narrative world of Harry Potter.

Using freely available materials (including Wikipedia and HP fan sites) along with the text of the HP book series by J.K. Rowling, the first author created a set of 156 sentence pairs that accurately described events and entities from the series. Each sentence context ended either in (a) a contextually Supported (and linguistically expected) word; (b) a word which was factually incorrect and Unrelated to the context and to the supported word; and (c) a word which was factually incorrect but which was Related in one of two ways to the context and/or contextually supported word. For half of the materials, the related words were taken from the same category as the linguistically expected word, as in Federmeier & Kutas (1999). For the remaining materials, the related word was related in some way to the episode/event being described by the preceding sentence context, as in Metusalem et al. (2012). Based on the previous findings, we expected that both types of relationships would lead to N400 related anomaly effects which might be similarly influenced by the degree of individuals' domain knowledge. Three lists were constructed such that each sentence frame and each critical word appeared only once per list (examples provided in Table 1).

In order to verify that the words we deemed related via category or event to contextually supported words were indeed more closely related than the unrelated ending, we conducted a series of experiments to examine these relationships. First, we trained a high-dimensional semantics/language model directly on the text of the HP book series; we then asked whether the word embeddings learned by the model reflected the manipulation in our materials (e.g., with Supported-Related word embeddings being closer in semantic space than Supported-Unrelated word embeddings). Next, we conducted two experiments asking participants of varying degrees of HP knowledge to rate critical words from our materials for their similarity and relatedness, respectively. Finally, with these measures in hand, we conducted an EEG/ERP study to ask to what extent and when domain knowledge impacts the availability of contextually supported as well as contextually unsupported yet functionally (categorically, event-based) related knowledge during written sentence comprehension.

Experiment 1: Word embeddings

We trained a word2vec model (Mikolov et al., 2013a,b) on the text from the HP book series. This model uses a neural net to learn word embeddings (vectors) in high-dimensional semantic space from word co-occurrences in the input. The semantic "contents" of such embeddings can reflect various aspects of meaning, including category and event-based relationships (reviewed in Lenci, 2018). We could then use these embeddings to quantify relative similarities/differences between word pairs (or average vectors computed over sequences of words).

Methods

Word2vec model. We trained a word2vec model (distribution by D. Yaginuma, https://github.com/dav/word2vec) on the text from the seven books of the HP series, taken from the official electronic publication (https://usd.shop.pottermore.com)—a total of 1,125,854 words, with a vocabulary size of 8,046 words (subject to the constraint of each word appearing at least 5 times in the HP books). We used the continuous bag-of-words (CBOW) architecture, which learns to predict a word based on its context—in our case, a window of 10 words on either side. Each word from the HP books was modeled as a point (i.e., vector) in a 200-dimensional space.

Word-word similarity Using this model, we extracted word embeddings for critical words from each of our experimental conditions (Supported, Related, and Unrelated). For each

Table 1. Sample HP sentence materials.

Sentence frame	Supported	Related	Unrelated	Related Anomaly Type
Sybill Trelawney is a Hogwarts Professor. She teaches	*Divination*	*Transfiguration*	*basilisk*	**Category**
In Quidditch, games are usually won in one way. This is when the seeker catches the	*Snitch*	*Bludger*	*dragon*	**Category**
Harry has a patronus. It takes the form of a	*stag*	*dementor*	*Sectumsempra*	**Event**
When Harry is one year old, Hagrid brings him to the Dursleys'. For transportation, he uses a borrowed	*motorcycle*	*Sirius*	*Vow*	**Event**

item (156 total), we then computed the cosine similarity (angular distance) between the word embeddings for Supported-Related and Supported-Unrelated pairs of critical words. We expected that the similarity for the Supported word-Related word pair would be greater Supported word-Unrelated word pair.

Sentence context-word similarity We also extracted word embeddings for each word (where possible) of our sentence pair frames/contexts. To create a single embedding (i.e., vector) for each item's sentential context, we took the average of all its words' vectors. We then computed the cosine distance between this aggregate context vector and the vector for each ending type (Supported, Related, Unrelated). We expected this distance would be greatest for the Context-Supported pair, followed by Context-Related and finally Context-Unrelated.

Results

As expected, we found that word embeddings—derived from a corpus of the HP novels' text—for Supported words were more similar to Related words than to Unrelated words (Fig. 1a). This pattern held for both category- and event-related item subsets, though it was somewhat larger within category-related items ($p < .01$). Also as expected, average word embeddings for sentential contexts were most similar to Supported words, followed by Related and Unrelated words (Fig. 1b). This pattern held both for category- and event-related item subsets. These findings show that the high-dimensional semantic space learned by the word2vec model captured systematic, meaningful differences in the relationships between the sentence context and the Supported, Related, and Unrelated endings.

Experiments 2a-b: Ratings studies

To further assess the manipulation in the HP sentences, and to examine the extent to which the manipulation was dependent on HP knowledge, we conducted two behavioral studies, asking participants to rate critical word-pairs (Supported-Related and Supported-Unrelated) on similarity (Exp. 2a) or relatedness (Exp. 2b). These criteria were chosen specifically to examine the two types of relationships we targeted in our HP sentence materials, namely categorical relationships (words share many similar features) and event relationships (words are related via an event/episode from the HP books). In addition, these experiments allowed us to assess the ratings of similarity/relatedness as a function of individuals' degree of HP knowledge. We expected that individuals with greater knowledge would be more sensitive to our experimental manipulations—i.e., that more knowledgeable individuals would indicate relatively greater similarity/relatedness for Supported-Related than for Supported-Unrelated word pairs.

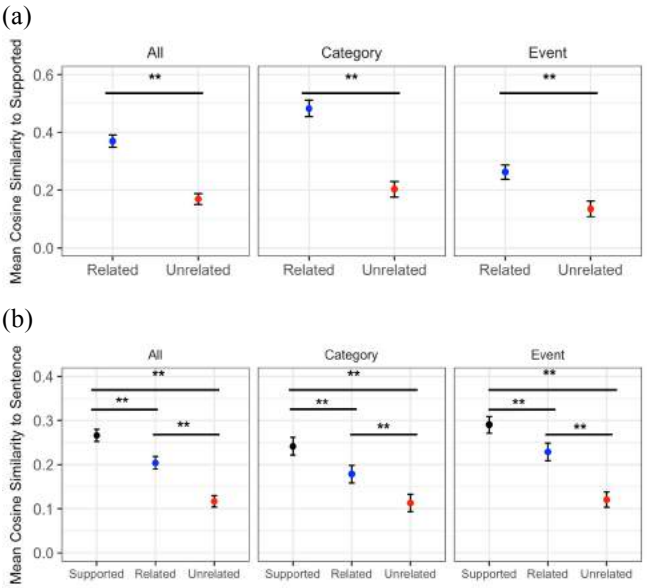


Fig. 1. (a) Across items and within each subset (Category-related, Event-related), mean cosine similarity for Supported & Related endings is greater than for Supported & Unrelated (all ps < .01). (b) For all 156 items (and within each subset) there was a significant three-way difference between cosine similarity of averaged word embeddings for sentences and Supported < Related < Unrelated endings.

Methods

Participants 24 participants completed similarity ratings; 25 different participants completed relatedness ratings. All participants were UCSD students; they received partial course credit as compensation.

Procedure For the similarity ratings experiment, participants were asked to consider word-pairs in the context of the Harry Potter stories and to judge their similarity in meaning using a scale ranging from 1 ("not similar at all") to 7 ("nearly the same meaning"). They were given the following guide to judging similarity of word meanings:

(1) Do the two word meanings behave similarly (e.g., do they perform the same actions)?
(2) Do the two word meanings share physical / sensory properties (e.g., do they look, taste, smell, sound or feel similarly)?
(3) Do the two word meanings share many functional properties (e.g., are they used in similar ways, or do they serve a similar purpose)?
(4) Do the two word meanings share any other properties and/or features in common?

For the relatedness ratings experiment, instructions were similar, except participants were asked to judge words on how *related* they were using a scale ranging from 1 ("not related at all") to 7 ("very closely related)". They were given

the following guide to judge whether the pairs were meaningfully related:

(1) How likely are the words to show up within the same context (that is, in/around the same part of the HP stories)?
(2) How important does one word seem to be for understanding the meaning of the other?
(3) Are the two words related via some theme, topic, event, or episode/scenario in the HP stories?
(4) Are the two words related via any other relationship?

Participants in Exp. 2a-b also completed a 10-question trivia quiz assessing their HP knowledge and a questionnaire about their HP experience.

Results

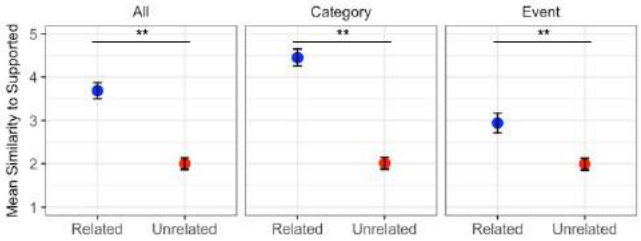
As expected, mean similarity ratings for Supported-Related word pairs were greater than those for Supported-Unrelated word pairs (Fig. 2a). This pattern was similar for both the category- and event-related item subsets, but was larger for the category-related subset, which might be expected based on greater similarity due to feature overlap between members of the same category (compared to words related via an event or episode).. Also as expected, HP knowledge was positively correlated with the size of the effect (i.e., similarity for Supported-Related word pairs minus similarity for Supported-Unrelated word pairs) at $r = .51$, $p < .05$.

In addition, mean relatedness ratings for Supported-Related word pairs were greater than those for Supported-Unrelated word pairs (Fig. 2b). This pattern was similar for both the category- and event-related item subsets. Also as expected, HP knowledge was positively correlated with the size of the effect (i.e., relatedness for Supported-Related word pairs minus relatedness for Supported-Unrelated word pairs) at $r = .68$, $p < .001$.

We also examined the correlation between the word2vec cosine similarity measures (Exp. 1) and the similarity and relatedness ratings for Supported-Related and Supported-Unrelated word pairs, respectively (Exp. 2). Cosine similarity was positively correlated with both similarity ($r = .43$, $p < .0001$) and relatedness ($r = .26$, $p < .01$) ratings for the Supported-Related word pairs, but not for the Supported-Unrelated word pairs (n.s.).

These results empirically indicate that our Supported sentence endings were indeed more similar/related to our Related, compared to Unrelated, endings. Moreover, that the size of these effects was positively correlated with HP knowledge further supports the notion that sensitivity to the relatedness manipulation depends on knowledge specific to the HP book series. Next we describe an ERP/EEG study designed to investigate the extent to which individual differences in domain knowledge influence the availability of information relevant for word processing—i.e., information cued by categorically / event related words—during real-time sentence processing.

(a)



(b)

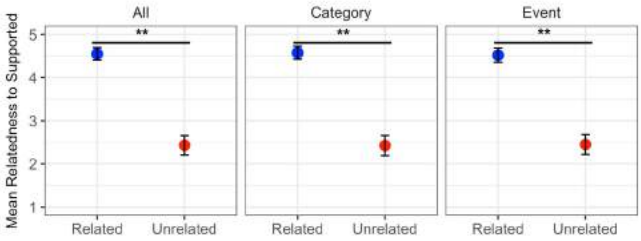


Fig. 2. (a) Across items and within each subset (Category-related, Event-related), similarity ratings for Supported & Related endings are greater than Supported & Unrelated (all ps < .01); this effect was larger for the category-related subset of items compared to the event-related subset ($p < .0001$). (b) Across items and within each subset, relatedness ratings for Supported & Related endings are greater than Supported & Unrelated (all ps < .001).

Experiments 3a-b: ERP studies

In Experiment 3, we asked whether certain aspects of the functional organization of semantic memory—namely organization of words/concepts via categories (wherein members share many similarities/features) and event/episode relationships—would be available to comprehenders as they read sentences about a fictional domain. We were particularly interested in whether the availability of not just the contextually supported information, but also the contextually unsupported, related information would be a function of the degree of individuals' domain knowledge.

To this end, participants of varying degrees of HP knowledge read 156 sentence pairs about the fictional world of HP while we recorded EEG (Experiment 3b). Sentences ended in a critical word that was either Supported, Related, or Unrelated (HP sentence materials described above). Three lists were then constructed so that every participant read each sentence frame and each critical word only once. That is, even though the same critical word appeared in other conditions on other lists, it never appeared in the critical position more than once in the same list. All but three words appeared as critical words in two or all three conditions. We expected that for individuals knowledgeable about HP, we would see a three-way difference in the amplitude of N400 potentials to the critical words, with the largest amplitude for Unrelated, the most reduced amplitude for Supported, and an intermediate amplitude for Related. We also expected that

N400 amplitude to Supported words would be positively correlated with HP knowledge (replicating Troyer & Kutas, 2018 and Troyer, Urbach, & Kutas, 2018). Moreover, we expected that N400 amplitude to Related words would be positively correlated with HP knowledge, consistent with the real-time use of differential functional organization of long-term memory as a function of degree of domain knowledge. For visualization of ERPs, we present subgroups of participants based on a median split on HP knowledge.

To demonstrate (a) that individuals spanning a range of HP knowledge scores could elicit N400 effects, more generally, and (b) that the relationship between HP knowledge and N400 effects was specific to sentences about HP, we also recorded EEG while the same participants read sentences about general topics (Experiment 3a). Due to time constraints, we included only 40 such sentences, half of which ended in a contextually Supported word / the best completion (determined by an offline cloze norming task in which participants provided completions to sentence frames) and the other half of which ended in a contextually Unsupported word (a plausible word that was low-cloze).

Methods

Participants 48 students from the UCSD community participated in the EEG study (Experiments 3a-b).

Experimental procedures Participants were instructed to silently read sentences for comprehension, first about general topics (Experiment 3a) and then about Harry Potter (Experiment 3b). In each experiment, the whole first sentence appeared in the center of the screen. When ready, participants pressed a button to move on to the second sentence, which was presented one word at a time in the center of the screen with a 500 ms SOA (200 ms on, 300 ms off). Following the ERP study, participants completed a 10-question HP trivia quiz and a questionnaire about their HP experience.

In addition, we collected several other measures of individual differences to better understand group differences among participants (see Troyer & Kutas, 2018, for more details). We combined measures of general print/reading experience (media and reading habits questionnaire, author and magazine recognition tests; Stanovich & West, 1989) for an aggregate reading experience score, and we also collected a measure of general knowledge (trivia quiz developed from freely available materials), and verbal working memory (sentence span, Daneman & Carpenter 1980). Finally, we administered a debriefing questionnaire.

ERP recording and data analysis The electro-encephalogram (EEG) was recorded from 26 tin electrodes geodesically arranged in an ElectroCap, with impedances kept below 5 KΩ. Recordings were referenced online to the left mastoid and re-referenced offline to an average of the left and right mastoids. EEG was recorded by Grass bio-amplifiers with a bandpass of .01-100 Hz at a sampling rate of 250 Hz. Trials contaminated by artifacts (e.g., eye movements or blinks) were not included in analyses.

Grand average ERPs to sentence-final words were computed across all 26 recording sites for each experiment and by Ending Type (3a: Supported / Unsupported; 3b: Supported / Related / Unrelated). For statistical analyses, we used linear mixed effects models and focused on a region of interest (ROI) where N400 effects are typically largest, including an average of 8 centro-parietally distributed channels (MiCe, LMCe, RMCe, MiPa, LDPa, RDPa, LMOc, and RMOc) in a canonical N400 time period (250-500 ms) relative to a 200 ms pre-stimulus baseline.

Results

ERPs from our centro-parietal ROI are shown in Fig. 3. ERPs to critical words are characterized by N1 and P2 sensory components. Across all participants, the P2 is followed by a relative negativity (N400), which is most reduced for Supported words compared to Unsupported (Control) and Unrelated/Related (HP) words.

Experiment 3a: Control sentences. Our primary aim in analyzing the control experiment was to ask (a) whether individuals, irrespective of their degree (or depth) of HP knowledge, would elicit standard N400 effects to contextually supported vs. unsupported words in sentences about general topics and (b) to determine whether HP knowledge influenced the size of this effect. We predicted that HP knowledge would have a specific influence on sentences about HP, but not on the size of the effect for control sentences. Our results confirmed this prediction. We observed main effects of ending type (Supported < Unsupported; $p < .0001$) and HP knowledge (individuals with greater HP knowledge tended to yield overall somewhat more positive-going N400 potentials; $p < .01$), but critically, no interaction between ending type and HP knowledge. That is, the size of individuals' N400 reduction to contextually supported, compared to unsupported, words did not differ as a function of individuals' degree of HP knowledge.

Experiment 3b: HP sentences. We expected that HP knowledge would modulate N400 amplitude to contextually supported words in sentences about HP, as in previous studies. In addition, we asked whether HP knowledge would also modulate N400 amplitude to contextually unsupported, but related, words in HP sentences. We observed a reliable interaction between HP knowledge and ending type ($p < .05$) and followed up using planned comparisons examining (a) the Unrelated vs. Supported endings and (b) the Unrelated vs. Related endings, finding that HP knowledge interacted with ending type in both cases (both ps < .05). Follow-up analyses revealed that HP knowledge was correlated with N400 amplitudes to Supported words (Pearson's $r = .57$, $p < .0001$) and Related words (Pearson's $r = .47$, $p < .001$), but not to Unrelated words (n.s.).

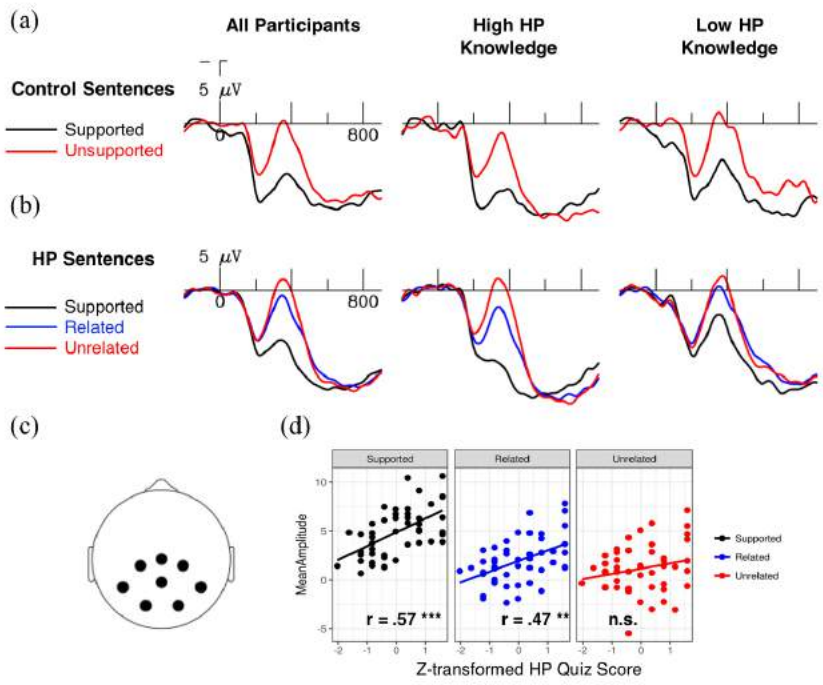


Fig 3. ERPs are plotted to critical words from Control (a) and HP (b) sentences for an ROI based on an average of 8 centro-parietal channels (c). For Control sentences, across all participants, Supported words elicited a reduced N400 compared to Unsupported words; there was no interaction between ending type and degree of HP knowledge. For HP sentences, across all participants, Supported words showed reduced N400 amplitude compared to Unrelated words, with Related words eliciting an intermediate N400. Reductions in N400 amplitude for both Supported and Related words were largest for individuals with high HP knowledge and smallest for individuals with low HP knowledge. (d) N400 amplitude is plotted against HP knowledge; the correlation between the two was strong for Supported and moderate for Related words, but was not significant for Unrelated words.

Due to some differences present in similarity metrics for category- vs. related-anomaly subgroups of items (Exp. 1-2), we also tested whether there were systematic differences in the N400 response between subgroups of materials. However, linear mixed effects models revealed no interaction of ending and related anomaly type on N400 potentials nor any interaction between these predictors and HP knowledge.

To rule out the possibility that other existing individual differences (namely reading experience, general knowledge, and verbal working memory scores) could better account for the observed variability in N400 ERPs, we tested a model that incorporated fixed effects of ending type, HP domain knowledge, general knowledge scores, reading span scores, and aggregate reading experience scores along with interaction terms for each individual differences measure with ending type. We compared this model and a similar model that did not incorporate interaction terms with any individual differences measures (except for the HP domain knowledge-by-ending type interaction term), and found that the more complex model did not explain additional variance.

Discussion

We asked whether, when, and to what extent individuals' degree of domain knowledge of the fictional world of HP would reliably influence the availability of meaningfully relevant information during written language comprehension, even when it was linguistically unexpected. To that end, we assessed a set of materials in which sentence contexts set up expectations for contextually supported words, along with sentence endings that were contextually unsupported, but were meaningfully related or unrelated to the sentence contexts and/or to the supported endings. In a word-by-word reading ERP study, we probed the extent to which real-time access to the same sentence endings was modulated by domain knowledge.

Importantly, individuals' degree of HP knowledge did not influence the size of the contextual support effect for Control sentences about general topics. Replicating Troyer & Kutas (2018), we found that N400 reduction to supported words was strongly predicted by degree of domain knowledge. Moreover, we observed a similar pattern for critical words that were contextually unsupported, yet related to the sentence context and/or supported word.

These results suggest that variation in knowledge—even of a fictional narrative world—influences what knowledge is retrieved in real time, which we believe is likely to reflect the way that knowledge is functionally organized. Our results further suggest that having relatively more knowledge, and thereby more organization around categories and/or events, allows for quick availability of this relevant organization during real-time reading. That is, knowledgeable individuals can quickly (pre)-activate relevant featural and/or thematic (as in the event-related subset of items) information—the very knowledge that is needed to make sense of words in real time. Moreover, individuals' degree of knowledge seems to predict the likelihood with which and/or extent to which such information becomes available for use. These methods and findings invite new research using knowledge-based individual differences to better understand how language processing interfaces with knowledge in real time. For example, future work could combine subject-level domain knowledge and sentence-and-word-level similarity and/or relatedness measures (e.g., based on computational models (Exp. 1) or human judgments (Exp 2.)) to investigate their joint influences at the individual trial level.

References

Altmann, G.T.M. & Mirkovic, J. (2009). Incrementality and prediction in sentence processing. *Cognitive Science*, *33*, 583-609.

Coronel, J.S. & Federmeier, K.D. (2016). The N400 reveals how personal semantics is processed: Insights into the nature and organization of self-knowledge. *Neuropsychologia*, *84*, 36-43.

Daneman, M. & Carpenter, P.A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, *19*, 450-466.

Ericsson, K.A., Charness, N., Feltovich, P.J., Hoffman, R.R. (Eds). (2006). *The Cambridge Handbook of Expertise and Expert Performance*. Cambridge, UK: Cambridge University Press.

Federmeier, K.D. & Kutas, M. (1999). A rose by any other name: Long-term memory structure and sentence processing. *Journal of Memory and Language*, *41*, 469-495.

Federmeier, K.D., McLennan, D.B., de Ochoa, E., & Kutas, M. (2002). The impact of semantic memory organization and sentence context information on spoken language processing by younger and older adults: An ERP study. *Psychophysiology*, *39*, 133-146.

Filik, R. & Leuthold, H. (2013). The role of character-based knowledge in online narrative comprehension: Evidence from eye movements and ERPs. *Brain Research*, *1506*, 94-104.

Hagoort, P., Hald, L., Bastiaansen, M., Petersson, K.M. (2004). Integration of word meaning and world knowledge in language comprehension. *Science*, *304*, 438-441.

Kutas, M., DeLong, K.A., & Smith, N.J. (2011). A look around at what lies ahead: Prediction and predictability in language processing. In M. Bar (Ed.), *Predictions in the Brain*. Oxford, UK: Oxford University Press, pp. 190-207.

Kutas, M. & Federmeier, K.D. (2000). Electrophysiology reveals semantic memory use in language comprehension. *Trends in Cognitive Sciences*, *4*(12), 463-470. doi: 10.1016/S1364-6613(00)01560-6

Laszlo, S. & Federmeier, K.D. (2009). A beautiful day in the neighborhood: An event-related potential study of lexical relationships and prediction in context. *Journal of Memory and Language*, *61*, 326-338.

Lenci, A. (2018). Distributional models of word meaning. *Annual Review of Linguistics*, *4*, 151-171.

Metusalem, R., Kutas, M., Urbach, T.P., Hare, M., McRae, K., & Elman, J.L. (2012). Generalized event knowledge activation during online sentence comprehension. *Journal of Memory and Language*, *66*(4), 545-567.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In *Proceedings of Workshop at ICLR*.

Mikolov, T., Sutskever, I, Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances on Neural Information Processing Systems*.

Stanovich, K.E. & West, R.F. (1983). On priming by a sentence context. *Journal of Experimental Psychology: General*, *112*(1), 1-36.

Troyer, M. & Kutas, M. (2018). Harry Potter and the Chamber of *What*?: The impact of what individuals know on word processing during reading. *Language, Cognition, and Neuroscience*, 1-17.

Van Petten, C. & Kutas, M. (1990). Interactions between sentence context and word frequency in event-related brain potentials. *Memory and Cognition*, *18*(4), 380-393.

Environmental Regularities Shape Semantic Organization throughout Development

Abstract

Our knowledge of the world is an organized lexico-semantic network in which concepts can be linked by relations, such as "taxonomic" relations between members of the same stable category (e.g., *cat* and *sheep*), or association between entities that occur together or in the same context (e.g., *sock* and *foot*). Prior research has focused on the emergence of knowledge about taxonomic relations, whereas association has received little attention. The goal of the present research was to investigate how semantic organization development is shaped by both taxonomic relatedness and associations based on co-occurrence between labels for concepts in language. Using a Cued Recall paradigm, we found a substantial influence of co-occurrence in both 4-5-year-olds and adults, whereas taxonomic relatedness only influenced adults. These results demonstrate a critical and persistent influence of co-occurrence associations on semantic organization. We discuss these findings in relation to theories of semantic development.

Keywords: semantic development; semantic organization; categories

Introduction

Our knowledge about the world is fundamental to many of the cognitive feats we accomplish on an everyday basis, including applying what we know to new situations, retrieving knowledge from memory, and incorporating new information into existing knowledge (Bower, Clark, Lesgold, & Winzenz, 1969; Heit, 2000; Tse, Langston, Kakeyama et al., 2007). These feats are possible due to the organization of our knowledge into an interconnected lexico-semantic network of related concepts (Cree & Armstrong, 2012; McClelland & Rogers, 2003). For example, our knowledge of dogs is often connected to our knowledge of other similar animals (e.g., cats), as well as to our knowledge about the contexts in which dogs appear, such as with leashes and doghouses.

Although the fact that our concepts are organized is hardly controversial (e.g., McClelland & Rogers, 2003), the processes that drive the development of semantic organization are a topic of considerable debate. To date, this debate has focused on how connections between concepts from the same stable, "taxonomic" category (e.g., *animals*, *foods*) are formed, in spite of the fact that they may be difficult to observe: Members of the same (especially superordinate) taxonomic category do not necessarily look similar, or occur together. Some have proposed that semantic development begins with easy to observe relations that are then used to bootstrap taxonomic knowledge (Lucariello, Kyratzis, & Nelson, 1992). Alternately, others have proposed that we are endowed with early-emerging biases towards learning taxonomic relations (e.g., Gelman & Markman, 1986).

The goal of this research is to investigate another possibility: That easy to observe relations – specifically, co-occurrence – play a fundamental role in shaping knowledge organization from early in development through adulthood. In this paper, we first review traditional theoretical accounts that have focused on taxonomic relations, then highlight key findings suggestive of a role for co-occurrence that these accounts fail to capture, and an alternate perspective that we test in the present experiment.

Traditional Accounts of Semantic Development

Most extant accounts of the development of semantic organization have focused on how semantic knowledge becomes organized according to membership in taxonomic categories, such as *foods*. According to some accounts, referred to here as *restructuring* accounts, taxonomic relations are the endpoint of development. Critical to these accounts is the idea that the order in which relations between concepts are acquired is dictated by how observable they are. For example, it is easy to observe that cups have the same shape, or reliably co-occur with juice or milk, whereas membership in the same superordinate taxonomic category is more difficult (if not impossible) to observe. Restructuring accounts propose that early organization is shaped by information readily available in the environment, and that taxonomic knowledge comes to replace this (more rudimentary) organization.

An early restructuring account was proposed by Inhelder and Piaget (1964), in which the transition to taxonomic organization is driven by experiences that highlight the inadequacy of earlier modes of organization (although the mechanisms by which this transition occurs are not clear). Another, more specified restructuring account is Nelson and Lucariello's (1992) slot-filler account, which highlights environmental input in which some members of the same taxonomic category play the same role in the same context, such as some members of the taxonomic category of *foods* (e.g., eggs and bacon) reliably *being eaten* in a *breakfast* context. According to this account, young children are sensitive to these regularities, such that semantic knowledge is first organized into contextually-constrained taxonomic groups, which are gradually integrated together as children recognize when entities play the same role in different contexts (e.g., foods *being eaten* in different meal contexts).

According to another set of accounts, referred to here as *taxonomic bias* accounts, taxonomic relations predominate semantic organization from early in development due to early-emerging (possibly innate) biases towards learning which entities are members of the same taxonomic category. These biases include beliefs that entities in the world belong to taxonomic categories, and that labels are indicative of category membership (e.g., Gelman & Coley, 1990). A role for other types of environmental input, such as the regularity with which entities co-occur, is not specified.

A final type of account reviewed here, which we refer to as *featural learning*, posits that the development of semantic organization is driven by detecting clusters of features

whose appearance in entities is reliably correlated, and which are often associated with taxonomic category membership (Rosch, 1975). For example, membership in the category of *birds* is associated with possessing *wings*, *feathers*, and a *beak*. Featural learning accounts propose that sensitivity to these correlations yields taxonomic organization (e.g., McClelland & Rogers, 2003). In contrast with taxonomic bias accounts, featural learning accounts argue in favor of the gradual emergence of taxonomic organization over the course of development. However, featural learning accounts do not consider spatial or temporal co-occurrence of items in the world (or language) as contributors to semantic organization.

Environmental Regularities Overlooked by Traditional Theoretical Accounts

Of the influential accounts reviewed in the previous section, only some restructuring accounts posit any role in semantic development for environmental regularities with which entities and their labels co-occur. Even in these accounts, these regularities are ultimately overwritten. However, several findings highlight a potential importance of co-occurrence regularities *throughout* development.

First, statistical learning studies suggest that sensitivity to the regularity with which different entities co-occur is apparent from very early in development (Bulf, Johnson, & Valenza, 2011). Moreover, numerous findings attest to the influence on children's reasoning of semantic relations that may be derived from co-occurrence, such as *schematic* relations between entities that occur in the same context (e.g., *cow* and *barn*) and *thematic* relations between entities that play complementary roles (e.g., *nail* and *hammer*) (Blaye, Bernard-Peyron, Paour, & Bonthoux, 2006; Fenson, Vella, & Kennedy, 1989; Lucariello et al., 1992; Walsh, Richardson, & Faulkner, 1993). Additionally, a handful of studies conducted by Fisher, Godwin and Matlen (Fisher, Matlen, & Godwin, 2011; Matlen, Fisher, & Godwin, 2015) point more directly towards an influence of co-occurrence on children's semantic reasoning. In these studies, participants were asked to infer whether a property (e.g., "has blicket inside") attributed to a target (e.g., glove) was shared by either a strongly taxonomically related item (e.g., mitten) or a more weakly taxonomically related item (e.g., sweater). These studies revealed that four year old children only reliably chose the strongly taxonomically related item when its label co-occurred with the target either in corpora of children's speech input (e.g., bunny-rabbit, Fisher et al., 2011) or an empirically manipulated speech stream (Matlen et al., 2015). These findings suggest that accounts of semantic development that do not posit any role for co-occurrence are at best incomplete.

Second, a handful of findings suggest that semantic relations that may be derived from co-occurrence continue to shape semantic organization into adulthood. For example, Lin and Murphy (2001) found that relations between entities that adult raters judged as associated in scenes or events (which likely co-occur) had a pervasive influence on adults' categorization and reasoning that was frequently greater than the influence of taxonomic relations. This evidence is inconsistent with restructuring accounts, in which an early influence of co-occurrence is eventually overwritten.

Finally, the potential contributions of co-occurrence regularities are highlighted by a mechanistic account and corroborating behavioral evidence presented by Sloutsky, Yim, Yao, and Dennis (2017). According to this account, exposure to co-occurrence regularities in language fosters both the learning of associations between concepts whose labels directly co-occur in sentences (e.g., *fork* and *spaghetti*), and between taxonomically related concepts whose labels share patterns of co-occurrence (e.g., *spaghetti* and *pie*). However, whereas co-occurrence in a sentence can be directly gleaned from input and therefore rapidly learned, shared patterns of co-occurrence that often link members of the same taxonomic category are learned more slowly because they can only be derived from multiple instances of direct co-occurrence. This account predicts both that (1) direct co-occurrence should contribute to semantic organization throughout development, and (2) the contributions of direct co-occurrence to semantic organization should be evident earlier in development than the contributions of taxonomic relatedness. Initial evidence for this account comes from a series of experiments presented in Sloutsky et al. (2017) in which children and adults were asked to infer the category membership of a novel word (e.g., whether it was an animal or a machine) that was presented within a list of familiar words. Both children and adults readily inferred the category membership of the novel word when it appeared in a list of words that are associated (and therefore likely to co-occur) with the same category. For example, participants inferred that the novel word referred to an animal when it appeared in a list of words including "furry" and "zoo". However, only adults inferred this meaning when the novel word appeared in a list of words referring to *members* of the category, such as "lion" and "bunny".

Together, these prior findings suggest that co-occurrence regularities may shape semantic development. However, in addition to being overlooked in traditional theoretical accounts of the development of semantic organization, this possibility has received only limited empirical investigation to date, and the way in which it has been investigated has not been designed to assess relational knowledge for items that *actually* co-occur in the environment. Critically, this research has instead investigated knowledge for relations between items either judged by researchers or participants as co-occurring according to researcher-specified criteria, or produced in free association tasks. Neither ratings nor free associations are inputs from the environment from which semantic relations can be learned: They are *outcomes* of relations already learned and present in semantic knowledge (Hofmann, Biemann, Westbury et al., 2018). A more direct investigation of the role of co-occurrence in shaping semantic development could be accomplished by assessing

the contributions of co-occurrence regularities present in actual environmental input.

Current Study

The overall purpose of the current study was to investigate the contributions of co-occurrence regularities and taxonomic relatedness to the organization of lexico-semantic knowledge from early childhood to adulthood. This investigation was designed to arbitrate between competing theoretical accounts of the development of knowledge organization. Specifically, restructuring accounts predict that co-occurrence should contribute to knowledge organization in childhood, but be replaced by taxonomic relations in adulthood. Both taxonomic bias and featural learning accounts are agnostic about the contributions of co-occurrence, but whereas the former predict that taxonomic relations should contribute from childhood through to adulthood, the latter predict that the contributions of taxonomic relations should substantially increase with age.

A different developmental pattern is predicted by recent proposals that highlight a key role throughout development for co-occurrence in which it both directly fosters relations between concepts, and indirectly fosters relations between concepts that share patterns of co-occurrence and are often taxonomically related (e.g., Sloutsky et al., 2017). Specifically, such proposals predict that the contributions of co-occurrence should be evident in both children and adults, whereas contributions of taxonomic relatedness should be evident only later in development.

We accomplished this investigation by measuring the degree to which familiar concepts were related in young children (4-year-olds) and adults' semantic knowledge when either the concepts' labels reliably co-occur in linguistic input, or when they are members of the same taxonomic category. To target actual experienced co-occurrence, we identified pairs of words familiar to young children that co-occurred more reliably with each other than with other words in corpora of child-directed speech.

To measure the contributions of co-occurrence and taxonomic relations to children and adults' lexico-semantic knowledge, we used a Cued Recall paradigm to measure the effects of co-occurrence and taxonomic relatedness on memory retrieval. We selected this paradigm for two reasons. First, the sensitivity of this task to semantic relatedness is attested by numerous findings that semantic relatedness influences the accuracy with which people (including children) recall word pairs and lists (Bjorklund & Jacobs, 1985; Blewitt & Toppino, 1991). Second, this task facilitates a comparison between children and adults because it measures contributions to lexico-semantic knowledge without requiring participants to reason about relations, which adults may more easily.

Table 1: Pairs of words used in the Co-Occur, Taxonomic, and Unrelated conditions

Co-Occur		Taxonomic		Unrelated	
bottle	baby	ball	puzzle	crayon	frog
foot	shoe	pig	bear	towel	bread
brush	hair	horse	bunny	blocks	cereal
cup	juice	carrot	banana	balloon	tree
cheese	mouse	fork	bowl	sheep	pancake
car	street	popcorn	fries	pizza	lion
soup	spoon	airplane	boat	fish	bed
milk	cow	sock	pajamas	duck	swing

Method

Participants

The sample included 30 4-5 year old children (M_{age}=4.50 years, *SD*=1.62 years), and 29 Adults (M_{age}=20.16 years, *SD*=3.66 years). The child age group was selected because the 4-5 year period is one during which the nature of relations that organize lexico-semantic knowledge has been the subject of active debate (Lucariello et al., 1992; Nguyen & Murphy, 2003; Waxman & Namy, 1997). Children were recruited from families, daycares, and preschools in a metropolitan area in a Midwestern US city. Adults were undergraduates from a public university in the same city and participated in exchange for partial course credit.

Stimuli and Design

The primary stimuli used in this experiment were word pairs that belonged to one of three Semantic Relatedness conditions: Co-Occur (pairs that reliably co-occurred with each other more often than with other words in child speech input), Taxonomic (words close in meaning from the same taxonomic category) or Unrelated. (words that neither reliably co-occur nor are similar in meaning).

Co-Occurrence Criteria. The first step taken to select pairs in each condition was to identify a set of words for which lexical norms collected using the MacArthur-Bates Communicative Development Inventory (MB-CDI) were available from WordBank (an open database of children's vocabulary development, Frank, Braginsky, Yurovsky, & Marchman, 2016), and measure their rates of co-occurrence in 25 child speech input corpora from the CHILDES database (MacWhinney, 2000). To reduce the computational expense of measuring word co-occurrence rates, some classes of words that would *a priori* not be used as stimuli were removed, such as sounds (e.g., "moo"), leaving a list of 538 words. Additionally, to ensure that co-occurrences were measured from speech *input*, CHILDES corpora were pre-processed to remove speech produced by children. Co-occurrences between these words were then calculated by taking all possible pairs of words in this set, and calculating how frequently they co-occurred with each other within a 7-word window across 25 CHILDES corpora. Finally, to account for the fact that more frequent words co-occur with other words simply by chance, t-scores (Evert, 2008) were calculated for each word pair using the formula below based on their measured co-occurrence frequencies (O), adjusted for the frequency of co-occurrence expected by chance based on their respective frequencies across the corpora and the size of the corpora (E):

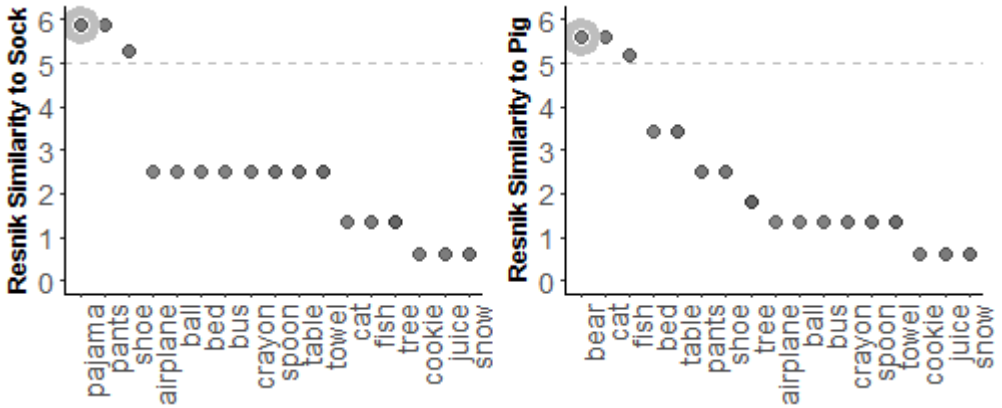


Figure 1: Graphs depicting Resnik similarity between one item from a Taxonomic pair and: (1) The other item from the pair (highlighted), (2) Other items from the same taxonomic category, and (3) Items from other categories.

$$t.score = \frac{O - E}{\sqrt{O}}$$

Word pairs for use in the Co-Occur condition were then selected as pairs of nouns with t-scores > 2.5 (following Baayen, Davidson, & Bates, 2008) in which, according to lexical norms accessed from WordBank, both words were produced by >80% of 36-month-old children (one year younger than children in our sample).

Taxonomic Criteria. Taxonomic relatedness was determined based on both the membership of concepts in the same taxonomic category (e.g., clothing, foods, animals) and similarity in *meaning* between their labels. Similarity in meaning was measured as similarity between the definitions of candidate words from WordNet (a database of word definitions composed by lexicographers). This measure captures the essence of taxonomic relatedness – i.e., close similarity in meaning – without relying on participant judgments that may be influenced by non-taxonomic relations (Wisniewski & Bassok, 1999). In WordNet, nouns are first grouped into sets of synonyms, which are in turn linked into a hierarchy according to "IS A" and part-whole relations. Similarity in meaning between word pairs was measured using Resnik similarity, i.e., the information content (specificity) of the word lowest in the WordNet hierarchy within which the pair of words is subsumed. For example, *dog* and *cat* are subsumed within *carnivore*, whereas *dog* and *kangaroo* are subsumed within mammal; because the information content of carnivore is greater than the information content of mammal, Resnik similarity is higher between dog and cat versus dog and kangaroo.

Candidate Taxonomic pairs nouns with Resnik similarities of > 5 and t-scores < 1.5 in which both were produced by at least 80% of 36-month-old children according to WordBank norms. The rationale of the Resnik similarity criterion of > 5 is illustrated in Fig. 1, which shows that this value distinguished between same- vs. different-category items.

Unrelated Criteria. Candidate Unrelated word pairs were noun pairs that met the WordBank production norm criterion with t-scores and Resnik similarities of < 1.5.

Composition of Full Set. From the sets of candidate pairs, eight pairs were selected for each of the Relation conditions (Co-Occur, Taxonomic, and Unrelated) such that: 1) The mean percentage of 36-month-olds who produced the words in the pairs according to Wordbank norms was equated across conditions, and 2) No words appeared in more than one condition (Table 1). An additional 4 nouns that met the WordBank production norm criterion were selected to construct pairs used for demonstration and practice (see Procedure below). All words were recorded by both a male and a female speaker using an engaging, child-friendly intonation.

The eight pairs in each Relation condition were divided into two Stimulus Sets, each with four pairs in each condition, because pilot testing indicated that 12 pairs was the maximum number that could be presented to children without producing floor effects. Within each Stimulus Set, each word in a pair was randomly assigned to be either the Cue or Target. In the experiment, Cue words were presented using the male speaker's voice, and Targets using the female's voice. Additionally, the 12 word pairs were pseudorandomized into three blocks, such that each block contained 1-2 pairs from each condition. The order of these blocks was counterbalanced across participants.

Procedure. Adult participants were tested in a quiet space in the lab, and children were tested either in a quiet space in the lab, or at their preschool or daycare. The procedure was identical for adults and children (including the auditory presentation of the same recorded Cue-Target pairs), with the exceptions that: 1) Instructions were conveyed by an experimenter for children, and as text on a computer screen for adults, and 2) Children made verbal responses recorded by the experimenter, whereas adults typed responses.

To start, participants were informed that they were going to play a game with two sock puppets depicted on the computer, Izzy and Ozzy, in which Izzy and Ozzy would say pairs of words. The two demonstration/practice unrelated Cue-Target spoken word pairs were then played, while animations depicted one puppet "saying" the Cue word, and the other saying the Target word. Participants

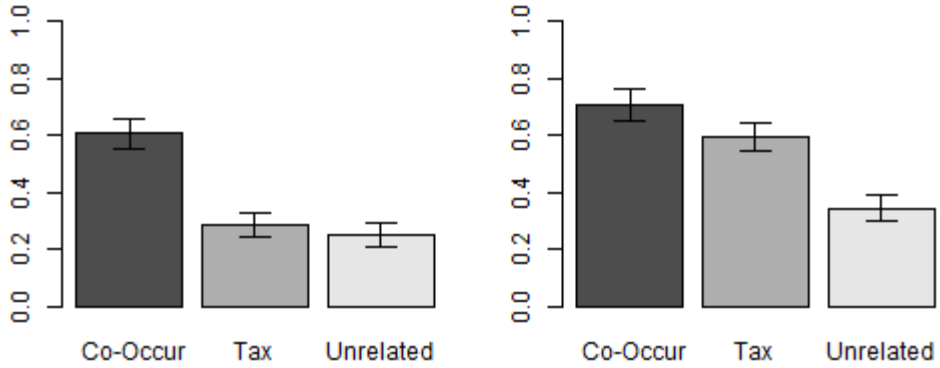


Figure 2: Accuracy in children (left) and adults (right) in the Relation Conditions. Error bars represent standard errors.

then completed two practice rounds with the same Cue-Target pairs consisting of a Study Phase, in which participants were instructed to remember the words that went together in pairs, and a Test phase, in which only the Cue in each pair was presented and participants were prompted to either say or type the Target that had been spoken by Ozzy. Participants received corrective feedback after each practice trial, and completed up to three practice rounds until they either responded with the correct Target for both Cues within around, or the experiment was terminated.

Participants then proceeded to complete the three blocks of Cue-Target pairs in the Stimulus Set to which they had been randomly assigned. Each block followed the same Study and Test phase format as the practice rounds, with the exception that participants did not receive feedback.

Results

The primary outcome measure of interest for this study was the accuracy with which participants recalled Target words paired with Cues in each of the three Relation conditions: Co-Occurrence, Taxonomic, and Unrelated[1]. Responses were scored as accurate when participants made responses identical to the Target, morphological variants of the Target (e.g., "spoons" instead of "spoon"), or close synonyms to the Target (e.g., "road" instead of "street").

All analyses were conducted in the R environment. Mixed effects models were generated using the lme4 (Bates, Maechler, Bolker, & Walker, 2015) package, and corresponding χ^2 or F-statistics for main effects and interactions were generated using the car package (Fox & Weisberg, 2011).

Preliminary Analyses: Stimulus Set Comparison

We first tested whether any effect of condition varied across the two Stimulus Sets in children and adults. For data from each age group, we generated a binomial generalized linear mixed effects model with Accuracy (0 or 1) as the outcome variable, Relation condition (Co-Occurrence, Taxonomic, and Unrelated) and Stimulus Set (1 vs. 2) as fixed effects, and participant and item as random effects. This analysis revealed no significant interaction between Relation condition and Stimulus Set (ps > .23). For all subsequent analyses, we therefore collapsed across Stimulus Sets.

Primary Analyses

Accuracy by age and condition is presented in Figure 2. To test the relative influences of Relatedness conditions (Co-Occurrence, Taxonomic, and Unrelated) on accuracy, we generated an omnibus binomial generalized linear mixed effects model with Accuracy (0 or 1) as the outcome variable, Relatedness condition and Age group (children and adults) as fixed effects, and participant and item as random effects. This analysis yielded main effects of Relatedness condition ($\chi^2(2)=25.26$, $p<.001$) and Age group ($\chi^2(1)=10.36$, $p=.001$) that were qualified by an interaction ($\chi^2(2)=7.87$, $p=.02$).

To investigate the interaction between Relatedness condition and Age group, we conducted two sets of analyses: A first set in which we compared the effects of the different Relatedness conditions in each Age group, and a second set in which we compared the effects of each Relatedness condition in children versus adults.

Relation Conditions in Each Age Group. In these analyses, we generated for each age group a binomial generalized linear mixed effects model with Accuracy as the outcome variable, Relatedness condition as a fixed effect, and participant and item as random effects. These models

[1] We also analyzed participants' errors to test the frequency with which the incorrect responses participants in each age group produced either co-occurred with or were taxonomically related to the Cue. However, these analyses did not contribute meaningfully to our results. The majority of incorrect responses in both age groups were other words from the set of word pairs the participant heard (64% in children, 82% in adults). Of these responses, only a small minority (7-14%) were either co-occurring with or taxonomically related to the Cue, which was likely the result of the random chance with which some words from the list, when randomly recombined with Cues, happen to be related to them in some way. Of responses not drawn from the list of word pairs, the only detectable pattern was a tendency for children to respond with incorrect words that co-occurred with the Cue (52%) more often than words that were taxonomically related to the Cue (6%). This pattern mirrors the results of analyses of children's accuracy.

revealed significant effects of Relatedness condition in each age group (ps < .001) (Figure 3). To conduct pairwise comparisons of the Relatedness conditions in each age group, we re-generated the model for each age with each of the Relatedness conditions as the reference level, and applied Bonferroni-adjustments to the resulting p-values. In children, these analyses revealed significant differences between the Co-Occurrence (M=0.60, SD=0.49) and both Unrelated (M=0.25, SD=0.43) and Taxonomic conditions (M=0.29, SD=0.45) (ps < .001), but no difference between the Taxonomic and Unrelated conditions (p > .99). In adults, these analyses revealed a significant difference between the Co-Occurrence (M=0.71, SD=0.46) and Unrelated conditions (M=0.34, SD=0.48) (p < .0001), the Taxonomic (M=0.59, SD=0.49) and Unrelated conditions (p=.033), and no significant difference between Co-Occurrence and Taxonomic conditions (p=.237).

Comparison of Children and Adults. To compare the accuracy of children versus adults in each Relatedness condition, we generated a binomial generalized linear mixed effect model for each Relatedness condition, each with Age Group as a fixed effect, and participant and item as random effects. Additionally, we applied Bonferroni-adjustments to all p-values to correct for multiple comparisons. These analyses revealed only a significant difference between children and adults in accuracy in the Taxonomic condition (p<.001). In comparison, there was no significant difference in accuracy between children and adults in either the Co-Occur or Unrelated conditions (ps>.2).

General Discussion

The purpose of the present experiment was twofold: (1) To investigate how semantic development is shaped by co-occurrence regularities and taxonomic relatedness, and (2) More broadly, to investigate whether the development of semantic organization involves the maintenance of early-emerging taxonomic organization throughout development (as in taxonomic bias accounts), the restructuring of semantic organization (as in restructuring accounts), or the addition of new semantic knowledge that does not replace earlier-emerging knowledge.

In this experiment, we observed substantial effects of co-occurrence in both young children and adults. In contrast, an influence of taxonomic relatedness was only apparent in adults. Importantly, due to our use of an implicit measure of semantic knowledge, this developmental pattern is unlikely to be attributable to developmental improvements in reasoning. These findings therefore support a key role for co-occurrence in semantic development, and are consistent with an overall developmental trajectory in which some types of semantic knowledge (such as taxonomic) tend to supplement rather than supplant earlier-emerging knowledge.

Generalizability of Findings

In order to evaluate the support for a key role for co-occurrence in lexico-semantic development, it is important to consider the possibility that the cued recall paradigm used in this experiment biased the results in favor of this outcome. Specifically, accurately recalling pairs of words may better evoke participants' prior knowledge of word pairs that they have experienced occurring together than their knowledge of taxonomically related words.

However, this possibility is undermined by corroborating evidence from very different paradigms that do not involve recalling word pairs. First, as described in the introduction, findings from studies conducted by Fisher, Godwin, and Matlen (Fisher et al., 2011; Matlen et al., 2015) have provided evidence for the contribution of co-occurrence to semantic reasoning. Specifically, these studies found that young children only reliably infer that an item shares a property with another, strongly taxonomically related item when their labels co-occur (e.g., bunny-rabbit). Moreover, the pattern of results in adults and children has recently been replicated using another, very different paradigm in which the contribution of a given form of relatedness is measured based on the degree to which it interferes with participants' ability to identify when a picture (e.g., of a baby) does *not* depict the same thing as a preceding word (e.g., "bottle") (Unger & Sloutsky, Under Review). Taken together, these findings suggest a general contribution of co-occurrence to lexico-semantic knowledge that is not dependent upon the use of a cued recall-based assessment.

References

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 390-412.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*, 1-48.

Bjorklund, D. F., & Jacobs, J. W. (1985). Associative and categorical processes in children's memory: The role of automaticity in the development of organization in free recall. *Journal of Experimental Child Psychology, 39*, 599-617.

Blaye, A., Bernard-Peyron, V., Paour, J.-L., & Bonthoux, F. (2006). Categorical flexibility in children: Distinguishing response flexibility from conceptual flexibility. *European Journal of Developmental Psychology, 3*, 163-188.

Blewitt, P., & Toppino, T. C. (1991). The development of taxonomic structure in lexical memory. *Journal of Experimental Child Psychology, 51*, 296-319.

Bower, G. H., Clark, M. C., Lesgold, A. M., & Winzenz, D. (1969). Hierarchical retrieval schemes in recall of categorized word lists. *Journal of Verbal Learning and Verbal Behavior, 8*, 323-343.

Bulf, H., Johnson, S. P., & Valenza, E. (2011). Visual statistical learning in the newborn infant. *Cognition, 121*, 127-132.

Cree, G. S., & Armstrong, B. C. (2012). Computational models of semantic memory *The Cambridge Handbook of*

Psycholinguistics (pp. 259-282). Cambridge: Cambridge University Press.

Evert, S. (2008). Corpora and collocations. *Corpus linguistics. An international handbook, 2*, 1212-1248.

Fenson, L., Vella, D., & Kennedy, M. (1989). Children's knowledge of thematic and taxonomic relations at two years of age. *Child Development, 60*, 911-919.

Fisher, A. V., Matlen, B. J., & Godwin, K. E. (2011). Semantic similarity of labels and inductive generalization: Taking a second look. *Cognition, 118*, 432-438.

Fox, J., & Weisberg, S. (2011). *An R Companion to Applied Regression* (Second Edition ed.). Thousand Oaks, CA: Sage.

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2016). Wordbank: An open repository for developmental vocabulary data. *Journal of Child Language*.

Gelman, S. A., & Coley, J. D. (1990). The importance of knowing a dodo is a bird: Categories and inferences in 2-year-old children. *Developmental Psychology, 26*, 796.

Gelman, S. A., & Markman, E. M. (1986). Categories and induction in young children. *Cognition, 23*, 183-209.

Heit, E. (2000). Properties of inductive reasoning. *Psychonomic Bulletin & Review, 7*, 569-592.

Hofmann, M. J., Biemann, C., Westbury, C., Murusidze, M., Conrad, M., & Jacobs, A. M. (2018). Simple Co-Occurrence Statistics Reproducibly Predict Association Ratings. *Cognitive Science, 42*, 2287-2312.

Inhelder, B., & Piaget, J. (1964). *The early growth of logic in the child*. New York: Norton.

Lin, E. L., & Murphy, G. L. (2001). Thematic relations in adults' concepts. *Journal of Experimental Psychology: General, 130*, 3-28.

Lucariello, J., Kyratzis, A., & Nelson, K. (1992). Taxonomic knowledge: What kind and when? *Child development, 63*, 978-998.

MacWhinney, B. (2000). *The CHILDES project: The database* (Vol. 2): Psychology Press.

Matlen, B. J., Fisher, A. V., & Godwin, K. E. (2015). The influence of label co-occurrence and semantic similarity on children's inductive generalization. *Frontiers in Psychology, 6*, 1146.

McClelland, J. L., & Rogers, T. T. (2003). The parallel distributed processing approach to semantic cognition. *Nature Reviews Neuroscience, 4*, 310-322.

Nguyen, S. P., & Murphy, G. L. (2003). An Apple is More Than Just a Fruit: Cross-Classification in Children's Concepts. *Child development, 74*, 1783-1806.

Rosch, E. (1975). *Basic objects in natural categories*: Language Behavior Research Laboratory, University of California.

Sloutsky, V. M., Yim, H., Yao, X., & Dennis, S. (2017). An associative account of the development of word learning. *Cognitive Psychology, 97*, 1-30.

Tse, D., Langston, R. F., Kakeyama, M., Bethus, I., Spooner, P. A., Wood, E. R., . . . Morris, R. G. (2007). Schemas and memory consolidation. *Science, 316*, 76-82.

Unger, L., & Sloutsky, V. M. (Under Review). Environmental Regularities Influence Semantic Organization throughout Development.

Walsh, M., Richardson, K., & Faulkner, D. (1993). Perceptual, thematic and taxonomic relations in children's mental representations: Responses to triads. *European Journal of Psychology of Education, 8*, 85-102.

Waxman, S. R., & Namy, L. L. (1997). Challenging the notion of a thematic preference in young children. *Developmental Psychology, 33*, 555-567.

Wisniewski, E. J., & Bassok, M. (1999). What makes a man similar to a tie? Stimulus compatibility with comparison and integration. *Cognitive Psychology, 39*, 208-238.

Impatient to Receive or Impatient to Achieve: Goal Gradients and Time Discounting

Oleg Urminsky
University of Chicago, Chicago, Illinois, United States

Indranil Goswami
University of Buffalo, Buffalo, New York, United States

Abstract

When people behave impatiently, prioritizing sooner outcomes at the expense of latter ones, is it because they value achieve their goal sooner, or because they value receiving the benefits sooner? Prior research has often confounded goal gradient (the stronger motivational effect of more proximal goals) and time discounting effects on decision-making. We first establish a preference to invest in the earlier of two equally difficult goals (e.g, a first-goal preference) that could be explained either by relative goal gradients or by differences in time discounted value. We then experimentally separate the timing of goal completion and reward receipt. We find separate and disassociated large goal gradient and somewhat smaller time discounting effects. Our results suggest that goal gradient effects may provide a partial, but substantial, explanation of time discounting and, consequently, can inflate estimated discount rates when not accounted for.

Structural Thinking about Social Categories: Evidence from Formal Explanations, Generics, and Generalization

Nadya Vasilyeva (nadezdav@princeton.edu)

Tania Lombrozo (lombrozo@princeton.edu)

Department of Psychology, Princeton University, Princeton, NJ, 08540, USA

Abstract

Most theories of kind representation suggest that people posit internal, essence-like factors believed to underlie kind membership and the observable properties of members. Across two studies (N = 234), we show that adults can construe properties of social kinds as products of both internal and *structural* (stable external) factors. Internalist and structural construals are similar in that both support formal explanations (i.e., "category member has property P due to category membership C"), generic claims ("Cs have P"), and a particular pattern of generalization to individuals when the individuals' category membership and structural position are preserved. Our findings thus challenge these phenomena as signatures of essentialist thinking. However, once category membership and structural position are unconfounded, different patterns of generalization emerge across internalist and structural construals, as do different judgments concerning category definitions and property mutability. These findings have important implications for reasoning about social kinds.

Keywords: structural explanation; kind representation; generalization; essentialism; inherence; social categorization

Introduction

Kind representations allow people to organize, store, and use conceptual information efficiently and productively. We rely on our representations of social groups and natural kinds to make sense of the world, generate explanations, and make predictions about the individual category members we encounter. Most theories of kind representation, especially for natural and social kinds, emphasize an internalist bias, a tendency to look "within" the kind for deep, causally active, and explanatorily powerful factors that hold categories together, and that shape and maintain the properties of their members. This internalist bias can take the form of assumptions about internal causal structure (psychological essentialism; Gelman, 2003), or a preference for explanations citing factors that are inherent, as opposed to contextual or extrinsic (inherence heuristic; Cimpian & Salomon, 2014). Different manifestations of internalist bias have been widely documented (Haslam et al., 2010; Gelman, 2003; Rangel & Keller, 2011), and it has been proposed as a conceptual default, with profound – and often negative - consequences for the way we think about and behave towards members of social categories. For example, explaining a dearth of women in mathematics by appeal to their "essential" or inherent nature can discourage girls from pursuing careers in this field (Leslie, Cimpian, Meyer, & Freeland, 2015).

Some linguistic forms have been argued to promote internalist construals, in particular of social kinds. Generic expressions, which attribute a property to a category in general (e.g., "women fail math tests") have received particular attention (e.g., Cimpian, 2010; Cimpian & Markman, 2010; Rhodes, Leslie, & Tworek, 2012). There is also evidence that formal explanations, which appeal to category membership to explain a property (e.g., "Priya doesn't like math because she's a girl"), reflect internalist beliefs (Gelman, Cimpian, & Roberts, 2018; Prasada & Dillingham, 2006).

While internalist modes of thinking have been extensively explored in the psychological literature, alternative ways of representing kinds have received much less attention. One such alternative is *structural thinking*, and in particular, a structural construal of category-property connections (Haslanger, 2015; Vasilyeva, Gopnik, & Lombrozo, 2018). On a structural construal, stable associations between categories and their properties arise from stable external constraints acting on category members. For example, the categories "women," "men," "Blacks," and "Latin@s" occupy relatively stable social positions within a given social structure. These positions can differ across cultures and possess their own properties. To illustrate, the generics "women don't drive," or "women are bad a math," can be true in one social system but false in another. Such culture-dependence is one cue that a property-category association should be attributed to a *social position* rather than to *the category occupying that position.*

Because a social position and the category that occupies it can share the same label (e.g., "women"), we contend that generics and formal explanations can be interpreted in either internalist or structural terms. For example, a person could endorse a formal explanation ("He ended up in prison because he's Black") or a generic ("Black men end up in prison") for different reasons: under an internalist construal, attributing the property ("being in prison") to the category itself (e.g., presumed criminal inclinations), or under a structural construal, attributing the same property to the social position, constituted by a conglomeration of stable constraints acting on members of the category in virtue of occupying that position (e.g., unequal opportunities for Black youth, biased hiring and other barriers to wealth, racial profiling by the police, etc. – all the factors that together constitute the social position "Black" in the US).

In the current research, we test the prediction that adults can construe property-category associations in either

internalist or structural terms, and that both construals support formal explanations (Study 1) and generics (Study 2). However, we also investigate important ways in which internalist and structural construals are expected to differ. Because internalist and structural construals allocate different roles to category membership (vs. a category's social position) in explaining an associated property, we expect internalist and structural construals to result in different intuitions about using the property in category definitions (Study 1), different "mutability" judgments about true category membership when the property is removed (Study 1), and different patterns of property generalization as category membership and/or social position change (Study 2).

Documenting these predicted patterns of similarity and difference across internalist and structural construals is important for a number of reasons. First, alternatives to internalist thinking have rarely been articulated and tested. Documenting a psychologically real alternative can thus enrich our understanding of the mental representations that support our thinking about social (and potentially non-social) kinds. Second, given that internalist construals have been linked with the perpetuation of stereotypes and other negative social effects (Bastian & Haslam, 2004; Cimpian, 2010), an alternative form of construal could identify changes in mindset that would mitigate these effects. A structural construal is especially promising in that it explains (rather than ignores) property-category associations that in fact obtain (such as a low proportion of women in math) while also pointing to structural factors that could be targets of intervention.

While structural explanation has received attention within the philosophy of social science (Ayala, 2018; Ayala & Vasilyeva, 2015; Haslanger, 2015; Garfinkel, 1981), there has been little empirical work on the topic to date. In a recent paper, Vasilyeva, Gopnik, and Lombrozo (2018) reported a study investigating structural thinking in adults and children aged 3-6. Using open-ended explanations, category definition tasks, mutability judgments, and measures of formal explanation, they found that even 3-year-olds showed signs of early structural thinking, with greater differentiation between internalist and structural construals in older children and adults. The present work goes beyond Vasilyeva, Gopnik, and Lombrozo (2018) in five important ways: in using more realistic social categories (a group of immigrants); in exploring structural reasoning about novel social groups; in using a wide range of properties matched in terms of property/cue validity (Study 1) or content (Study 2); in the introduction of a control condition (Study 1 and Study 2), and in exploring judgments concerning generics and generalizations under different conditions (Study 2).

Study 1

In Study 1, we introduce participants to a novel social category ("Borunians," an immigrant group in the fictional country of Kemi), along with a suite of associated properties (e.g., holding low-paying jobs). Across properties, we vary whether the category-property connections are explained in a way that is internalist (e.g., appealing to group identity), structural (appealing to social position), or incidental (the associations just happen to be true). To test whether this manipulation is successful in inducing different construals, we adapt measures originally developed in Prasada and Dillingham (2006, 2009) to differentiate "principled" and "statistical" connections, and used also in Vasilyeva, Gopnik, and Lombrozo (2018). These measures include partial definition evaluation (i.e., whether the category can be defined in terms of the property), category mutability ratings (i.e., whether an individual missing the property is a true category member), and formal explanation evaluation (i.e., whether the presence of the property can be explained by appeal to category membership).

First, as explained in the introduction, we expected both internalist and structural construals to support formal explanations (e.g., "He holds a poorly paid job because he's a Borunian"). Second, we expected the internalist and structural conditions to differ with respect to partial definitions and mutability. A definition of a category in terms of an "essential"/inherent feature should be more appropriate than a definition citing a feature that holds only in virtue of a category's position in a social structure. Likewise, removing an internal feature should produce more damage to category membership than removing a feature acquired through a social position, and therefore contingent on external structure. These predictions found support in Vasilyeva, Gopnik, and Lombrozo (2018); we test them here with a more realistic social kind, a broader range of features, and a modified mutability measure.

Finally, and going beyond Vasilyeva, Gopnik, and Lombrozo (2018), we included features with an "incidental" explanation for which we predicted a profile of effects different from either internalist or structural thinking, based on Prasada and Dillingham (2006, 2009). We expected that incidental features would not support definitions and would be seen as easily mutable (like structural features), but that they would not support formal explanations (in contrast to both internalist and structural features).

Method

Participants Seventy-seven participants (38 women, 39 men; mean age 33) were recruited on Amazon MTurk in exchange for $1.50; in this and subsequent studies participation was restricted to workers with an IP address within the United States and with a HIT approval rating of 95% or higher from at least 50 previous HITs. An additional 33 participants were excluded for failing a memory check.

Materials, Design, and Procedure Participants read a short vignette introducing the novel social category of "Borunians" - a group of immigrants settled in a fictional country, *Kemi*, who originally immigrated from Bo-Aaruna. Borunians were characterized by 18 unique features, with 6 of each type: *Internalist* (tying the feature to Borunians' tradition and identity), *incidental* (roughly equivalent to Prasada and Dillingham's (2006) "statistical"), and

Table 1. Examples of features used in Study 1.

Internalist: Borunian traditions are extremely important to them, and form part of their identity: Borunians have a special tattoo on one arm.	*Incidental:* Here are some statements about Borunians that are true, but there's nothing about these features that ties them to Borunian culture, tradition, personality or anything about their place in Kemi society: Borunians barbeque in their back-yards all year round, so they buy a lot of barbequing coal all year round.	*Structural:* Here are a few characteristics that Borunians have due to their position in the Kemi society and governmental policies applying to Borunians: Borunians are *not* allowed to take any job with an income over 20,000 Kemi dollars per year (approximately 20,000 USD) if other appli-cants for the same job include Kemi citizens who are equally or more qualified. Due to this regu-lation, Borunians hold mostly poorly paid jobs.

structural (tying the feature to the structural constraints acting on Borunians due to their position within Kemi society). Sample features are shown in Table 1. All features were presented in generic form. A norming study with a separate group of 23 participants verified that the three feature types did not differ in mean cue and category validity.

After learning the features, each participant performed one of three judgments – formal explanation (e.g., Question: Why does he hold a poorly paid job? Answer: Because he is a Borunian. How good is this explanation? 1 not good at all - 7 very good), *partial definition* (e.g., Question: What is a Borunian? Answer: A Borunian is a person who holds a poorly paid job. How good is this answer? 1 not good at all – 7 very good), or *mutability* (e.g., Imagine an alternative world where people we call Borunians do not hold mostly poorly paid jobs. From your perspective, would you call them really and truly Borunians? (1 definitely no - 7 definitely yes). Each judgment involved 18 ratings, one about each feature. Prior to the main set of ratings, participants practiced the judgment type they were assigned on two practice trials that involved rating a feature of a dog ("has four legs") and of a barn ("is red").

In sum, the study implemented a 3 (judgment type: formal explanation, partial definition, mutability; between subjects) by 3 (feature type: internalist, structural, incidental; within subjects) design.

Results and Discussion

Participants' ratings were analyzed in an ANOVA with feature type as a within-subjects factor and judgment as a between-subjects factor, followed by planned *t*-tests. The main effect of judgment was significant, $F(2,74) = 5.70$, $p = .005$, $\eta_p^2 = .133$, and the main effect of feature type was marginal, $F(2,148)=2.99$, $p=.053$, $\eta_p^2=.039$. However, of most theoretical importance was the significant interaction between judgment and feature type, $F(4,148)=31.54$, $p<.001$, $\eta_p^2=.460$. As shown in Figure 1, each feature type had a unique "profile" across the three judgments. As predicted, and replicating Prasada and Dillingham's (2006, 2009) findings, internalist features (relative to incidental features) better supported formal explanations ($p=.003$) and definitions ($p<.001$), and were judged less mutable ($p< .001$). Also as predicted, structural features (relative to internalist features) supported definitions less strongly, and mutability judgments more strongly ($ps<.001$). However, they supported formal explanations to the same extent ($p=.327$), and more strongly than incidental features did ($p<.001$).

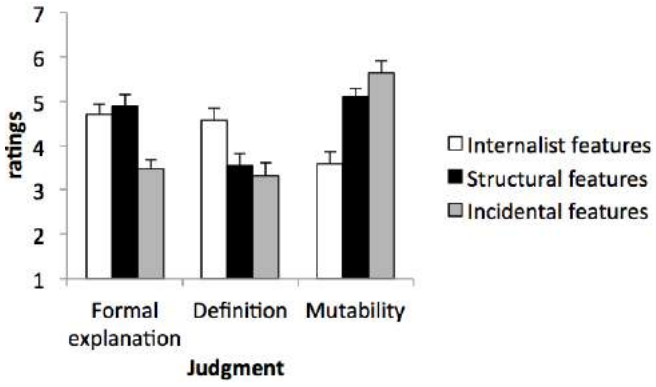


Figure 1: Participants' ratings as a function of feature type and judgment in Study 1. Error bars represent 1 SEM.

In sum, we find the predicted profile of effects for category-property associations introduced with a structural explanation. These associations behaved like internalist features in supporting formal explanations, but like incidental features in terms of partial definitions and mutability. It is worth noting that this structural pattern of responses was elicited by offering appropriate cues in the feature description, but did not require any explicit guidance or training in structural reasoning. This suggests that this mode of thinking may occur naturally in adults' cognitive lives when appropriate cues are present. It's also notable that the cues took the form of explanations, which presumably fed into causal-explanatory models that supported a representation that attached the property to the category versus the social position it occupied.

Study 2

In Study 2, participants received information about the prevalence of a property in the Borunian population, as well as an internalist, structural, or no explanation for the category-property association. Participants then rated their endorsement of a corresponding generic claim ("Borunians [have property]"), and generalized the properties in question to individual targets that varied both in category membership (same or different) and in social position (same or different).

This design allowed us to test two predictions. First, we predicted that internalist and structural construals would similarly support generic claims, with higher endorsement the greater the prevalence. Prior work has already shown that an internalist construal is not *necessary* for a generic to be endorsed; even statistical connections can support generics (Prasada & Dillingham, 2006, 2009; Tessler & Goodman, 2016). However, a structural construal additionally supports an interpretation of a generic claim whereby the category label refers to the social position that the category occupies.

Second, we predicted different patterns of generalization for properties explained internally vs. structurally as the generalization target varied in category membership and social position. In most real-life contexts, social category and social position are confounded, meaning that both internalist and structural explanations support the extension of category properties to individual category members (albeit for different reasons). De-confounding the category and position in our task, however, allows a predicted divergence between internalist and structural consturals to emerge. We expected a structural explanation to support greater generalization on the basis of shared position (relative to internalist), and an internalist explanation to support greater generalization on the basis of shared category (relative to structural). Additionally, we expected the effect of prevalence on generalization to be moderated by explanation, such that for participants who received a structural explanation (vs. internalist), prevalence effects would be weaker when social position was not preserved, and for participants who received an internalist explanation (vs. structural), prevalence effects would be weaker when category membership was not preserved. Comparisons to the control condition allowed us to assess the extent to which these effects were driven by a structural construal, an internalist construal, or both.

Method

Participants One-hundred-and-fifty-seven adults (76 women, 80 men, 1 agender; mean age 37) participated online in exchange for $1.50. An additional 30 participants were excluded for failing memory and attention checks.

Materials, Design, and Procedure We developed a new set of twelve features describing a fictional immigrant category, Borunians, introduced as in Study 1, and an internalist explanation and a structural explanation for each feature (see Table 2 for sample features and explanations). For the internalist condition, we intentionally chose a range of explanations spanning from more biological to those citing group preferences, values, and traditions (see further comments on this in the General Discussion). We also took care to keep the internalist and structural explanations of similar average length.

Each participant was assigned to one explanation condition (internalist, structural, or control), and completed two blocks of measures: generic truth ratings, and individual generalizations. In the generic truth rating block, participants saw the 12 features of Borunians, one at a time, in a random order, each accompanied by prevalence information (e.g., "Percentage of Borunians who hold poorly paid jobs: 48%"). For participants in the internalist or structural conditions, this was also accompanied by an explanation of the corresponding type (e.g., "Reason: in order to hire a Borunian for a well-paid job, employers in Kemi are required to file complicated government paperwork"). The feature prevalence (i.e., the percentage of Borunians with the feature) was drawn from a pool of 12 unique values, binned into Low (M=25%, range 20-29), Medium (M=50%, range 46-55), and High (M=75%, range 71-80). Below the prevalence information and explanation (if presented), participants read a generic statement attributing the feature to the category (e.g., "Borunians hold poorly paid jobs"), and were asked to classify it as "True" or "False."

In the individual generalization block, participants were asked to generalize a property from the kind (Borunians) to an individual. Participants were asked to rate their confidence that one of the properties previously attributed to Borunians (e.g., "holds a poorly paid job") held for that individual on a 9-point scale ranging from -4 (I'm confident it's false) to +4 (I'm confident it's true). Crucially, we manipulated both the category membership and the social position of the target individual: same vs. different category membership, and same vs. different social position. The resulting four scenarios are described in Table 3.

To ensure that participants still remembered the prevalence level and the explanation of each feature, the generalization rating block was split into three sets of four questions each. Each set of four questions was preceded by a reminder display with four features along with their prevalence levels and explanations (repeating the information from the first block). Further, to reduce memory load for prevalence levels, all four features in a set were pulled from the same prevalence bin (e.g., all had High prevalence). Following the reminder, participants saw

Table 2. Sample features and explanations used in Study 2. Each explanation was presented within the frame "Reason: [explanation]."

Feature	**Internalist Explanation** [Reason: ….]	**Structural explanation** [Reason: ….]
Follow a largely vegetarian diet	… a deficiency in digestive enzymes required for digesting meat	…special access to municipal subsidies to purchase vegetables directly from local farmers
Sell artisan souvenirs	…a natural affinity for design and great facility with fine-motor tasks	....special subsidies from the Kemi government to Borunians to obtain vendor permits for artisan booths
Get sunburn easily	…a genetic variation which makes Borunian skin very vulnerable to the effects of sunlight	…a high proportion of contaminants and skin irritants in the neighborhoods where Borunians live; these substances make their skin vulnerable to the effects of sunlight
Participate in donkey races	…agility and inherent skill with animals	…not allowed to participate in horse or car races
Live with their parents through adulthood	…a special value attached to family and elders, as well as living in tight-knit communities	… inability to afford the cost of maintaining independent residences
Hold poorly paid jobs	… strong preference to work regular hours; avoidance of demanding jobs that may require over-time	…in order to hire a Borunian for a well-paid job, employers in Kemi are required to file complicated government paperwork
Have poor credit ratings	…Borunians' reliance on a peculiar calendar with a different month length results in frequent late payments	….government banks imposed an additional step to verify every transaction for new immigrants, resulting in frequent late payments

Table 3. Descriptions of generalization targets produced by crossing same/different social category with same/different social position (note: social position is not the same as geographic location; non-Borunians in Kemi occupy a different social position from Borunians).

Scenario	Category	Position	Description
ALL SAME	Same	Same	Azz is a Borunian, and lives in Kemi.
MOVED	Same	Different	Nuvo is a Borunian who moved from Kemi a long time ago, and now lives in a completely different country, with an entirely different social system and regulations.
ADOPTED	Different	Same	Pau is a NON-Borunian by birth, who was adopted into a Borunian family in Kemi at a very young age, a long time ago, in a secret adoption (meaning the fact of adoption was never revealed, nobody except the parents knew that the child was adopted, and the child was brought up as a Borunian).
ALL DIFFERENT	Different	Different	Eken is a NON-Borunian who lives in Kemi.

the four generalization questions (one from each row of Table 3), in random order. The assignment of features to prevalence levels and question types, as well as the order of question sets, were counterbalanced across participants.

At the end of the survey, participants responded to a series of memory and comprehension checks (e.g., asking them to classify a list of characteristics and explanations as mentioned vs. not mentioned in the survey), as well as individual difference measures that are not analyzed here.

Results

Generic Truth Ratings Data were analyzed in a mixed effects logistic regression, predicting generic truth ratings from numerical prevalence, explanation, and their interaction (allowing for random intercepts for participants and items). To compare all three explanation conditions in this and the following regression models, the model was fit with the control condition as the reference group, and then re-fit with the structural condition as the reference group. Prevalence was the only significant predictor ($p<.001$): the odds of a "true" judgment increased 1.10 times per unit of increase in prevalence. Binning the prevalence predictor into three levels, the mean proportions of "true" responses were .25 (Low), .74 (Medium), and .94 (High). All other predictors were not significant, $ps \geq .244$, indicating that explanation condition did not affect overall generic endorsement, nor moderate the effect of feature prevalence.

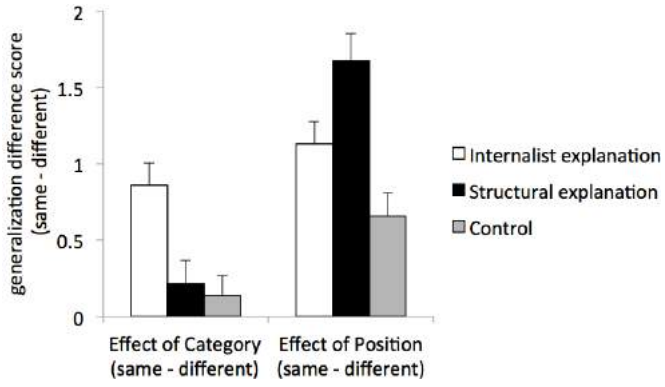


Figure 2: To represent the interactions between explanation condition, shared category membership, and shared social position, we created "generalization difference scores" (mean difference in generalization to individual in same vs. different category, and same vs. different social position). Error bars represent 1 SEM.

Individual Generalization A hierarchical linear model predicting individual generalization from centered numerical prevalence, explanation condition, shared category (yes or no), and shared social position (yes or no), with random intercepts across participants, revealed a four-way interaction, $p=.001$. To investigate this interaction further we ran additional analyses. First, to evaluate the prediction that an internalist explanation elevates the importance of shared category membership as a basis for generalization (relative to structural or control), we dropped prevalence and shared social position from the model, and predicted individual generalization from condition and shared social category. As expected, we observed significant interactions between regressors. The effect of shared category membership was stronger in the internalist condition relative to structural, $p=.006$, and to control, $p=.002$, which did not differ from each other, $p=.734$ (Figure 2). Second, to evaluate the prediction that a structural explanation elevates the importance of shared social position as a basis for generalization (relative to internalist or control), we predicted individual generalization from condition and shared social position. Again, we observed the expected interactions between regressors (see Figure 2), revealing a stronger effect of shared social position in the structural condition than either the internalist, $p=.014$, or control condition, $p<.001$. The internalist condition also heightened the relevance of social position relative to the control condition, $p=.036$, which suggests that our internalist explanations (perhaps by appealing to culture) also involved some social / structural elements.

Next, we addressed the prediction that the effect of prevalence on generalization would be moderated by explanation type. Given that the prevalence estimates that were offered corresponded to Borunians in Kemi (and not necessarily to non-Borunians or Borunians in other social positions), we expected the effect of prevalence to weaken with distance from the "ALL SAME" generalization target. However, we also expected that a change in category membership would attenuate the effect of prevalence more strongly in the internalist than in the structural condition, and that a change in social position would attenuate the effect of prevalence more strongly in the structural than in the internalist condition. To address this prediction, we considered the two cells that crossed category membership and social position ("ADOPTED" and "MOVED"; see Table 3), and ran separate models predicting individual generalization from prevalence and explanation condition (see Figure 3).

In the "MOVED" scenario, prevalence positively predicted generalization, β=.41, p<.001. Mirroring the results presented in Figure 2, participants were also *less* likely to generalize in the structural condition than in the internalist condition, B=-.45, p<.001, or in the control, B=-30, p=.008 (the latter two did not differ, B=.15, p=.181). Most crucially, however, we also observed interactions, such that the effect of feature prevalence was *weakened* in the structural condition relative to the internalist condition (B=-.31, p=.002) and control (B=-.33, p=.001); the effect of prevalence did not vary across the latter two explanation conditions (B=.018, p=.865).

In the "ADOPTED" scenario, prevalence positively predicted generalization, β=.67, p<.001. However, the predicted interaction between prevalence and explanation, with an attenuated effect of prevalence in the internalist condition (relative to control) was only marginal, B=-.19, p=.0502. Moreover, contrary to our expectations, the effect of prevalence was significantly attenuated in the structural condition (relative to control), B=-.33, p=.004. The extent to which the prevalence effect was attenuated, relative to control, did not differ across the two explanation condition, p=.117.

Finally, we considered the remaining two generalization targets, "ALL SAME" and "ALL DIFFERENT," for which we did not predict differential effects of explanation type. In

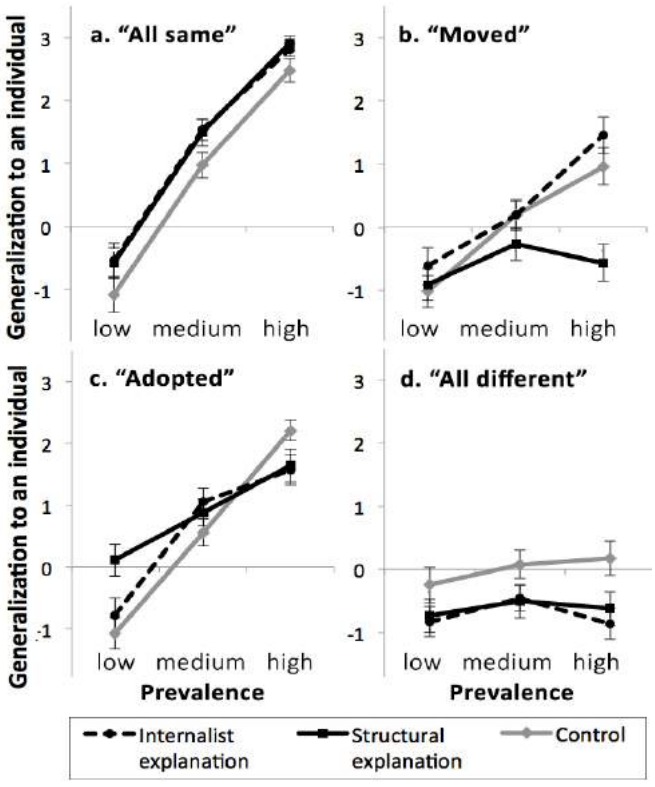


Figure 3: Mean individual generalization ratings as a function of within-category feature prevalence (binned into low, medium, and high ranges for presentation) and explanation type, split by the scenario (same or different category, and same or different social position). Error bars represent 1 SEM.

predictor of generalization, β = 70, p<.001, and both the "ALL SAME" scenario, prevalence was a positive internalist and structural explanations boosted generalization relative to control (B_{Int}=.20, p=.031; B_{Str}=.19, p=.036); the internalist and structural explanations did not differ, p=.945. As predicted, there were no significant interactions, ps ≥.780.

In the "ALL DIFFERENT" scenario, feature prevalence did not predict generalization, β=.10, p=.162. Participants were less likely to generalize in either explanation condition, relative to control ($B_{Int\ vs.\ control}$ =-.39, p=.005; $B_{Str\ vs.\ control}$ = -.34, p=.013); the two explanations did not differ, p=.708. As predicted, there were no significant interactions, ps >.238.

Discussion

Study 2 identified important respects in which internalist and structural construals overlap: both support generics, and both are equally sensitive to within-category/position statistics when it comes to endorsing generics or drawing generalizations to individuals within the same category/position. On the other hand, internalist and structural construals diverge when it comes to generalizations that break the typical confounds between categories and social positions: an internalist construal favors generalization (and reliance on within category/position statistics) across changes in social position; a structural construal is less sensitive to the preservation of category membership. These patterns emerged clearly in the "MOVED" scenario; the "ADOPTED" scenario (which was also the most unusual) was less clear.

In real life, the divergence between internalist and structural construals might be even more pronounced than that observed here. For experimental purposes, we used the same features across explanation conditions; as a result, many invoked culture and group identity, possibly downplaying more internalist factors. Indeed, shared social position was more influential overall than shared category, and shared position boosted the generalization of internalist features relative to control (Figure 2). Plausibly, the internalist condition could have been made "more internalist" by using different feature sets across conditions and citing exclusively biological factors in internalist explanations, as is common within the abundant literature documenting essentialist (or more broadly internalist) reasoning. Given that our goal was instead to document the reality of a structural construal as distinct from an internalist construal, we opted for greater experimental control over maximally representative features.

General Discussion

Across two studies we document underappreciated flexibility in people's construal of social kinds: in addition to adopting an internalist construal (familiar from prior research), people are capable of adopting a structural construal, which makes sense of observed correlations between properties and categories without tying them to the inherent nature of the category. Given the dangers of internalist construals in the social domain, an alternative that

makes sense of observed correlations without perpetuating them could be of no small social value.

Contrary to the dominant view, generic language does not necessarily convey or induce essentialist beliefs: in Study 1, the generic language that introduced a category-property association did not prevent an alternative construal, and in Study 2, both construals supported the endorsement of generic claims. This calls for refining numerous claims about generics and formal explanations as ways of inducing essentialism or signaling which kinds should be essentialized (e.g., Rhodes, Leslie, & Tworek, 2012). At the same time, it remains a possibility that internalist construals are the default, or cognitively less demanding.

Our generalization results also have important implications: from a theoretical standpoint, they offer yet another illustration of how explanation shapes generalization (Lombrozo & Gwynne, 2014; Sloman, 1994; Vasilyeva & Coley, 2013; Vasilyeva, Ruggeri, & Lombrozo, 2018), directing it along the dimensions of shared category and/or position. From a methodological standpoint, they offer a cautionary note about interpreting generalization measures as indices of essentialism; willingness to generalize a category's property can signal either an internalist *or* a structural construal. Finally, from a practical standpoint, getting a fuller picture of how people generalize from categories to individual has important real-life implications (e.g., a manager deciding whether to hire a woman, based in part on an inference about whether she'll take a parental leave).

One interesting question that deserves future attention is how people represent and reason about the "cultural" properties of social groups, such as food or religious customs. Where do such features fall on the internalist-structural continuum? One possibility is to identify "cultural" with "structural." However, many aspects of culture, including preferencves, values, and attitudes, can be understood in internalist terms, where cultural properties reflect shared internal characteristics. Consistent with these dual interpretations, the very same cultural properties in Study 2 were treated as internalist (through explanations citing "a special value attached to family and elders" or "a strong preference to work regular hours") or as structural (by attributing them to stable external constraints). However, the question of how people reason about "cultural" features in more naturalistic contexts remains open.

In sum, across two studies, we show that internalist and structural construals elicit different representations of categories: in the former case a property is attached to the category, in the latter case to its social position. While both kinds of representations can effectively track environmental statistics and support inferences, they work differently, in ways that could have tangible social consequences.

References

Bastian, B. & Haslam, N. (2005). Psychological essentialism and stereotype endorsement. *Journal of Experimental Social Psychology*, *42*, 228–235

Cimpian, A. (2010). The impact of generic language about ability on children's achievement motivation. *Developmental Psychology*, *46*(5), 1333–1340.

Cimpian, A., Brandone, A.C., & Gelman, S. A. (2010). Generic statements require little evidence for acceptance but have powerful implications. *Cognitive science*, *34*(8), 1452-1482.

Cimpian, A., & Markman, E.M. (2011). The generic/nongeneric distinction influences how children interpret new information about social others. *Child Development*, *82*(2), 471–492.

Cimpian, A., & Salomon, E. (2014). The inherence heuristic: An intuitive means of making sense of the world, and a potential precursor to psychological essentialism. *Behavioral and Brain Sciences, 37*(5), 461–480.

Garfinkel, A. (1981). *Forms of explanation: Rethinking the questions in social theory*. New Haven: Yale University Press.

Gelman, S. A. (2003). *The essential child: Origins of essentialism in everyday thought*. Oxford University Press.

Gelman, S. A., Cimpian, A., & Roberts, S. O. (2018). How deep do we dig? Formal explanations as placeholders for inherent explanations. *Cognitive psychology*, *106*, 43-59.

Haslam, N., Rothschild, L., & Ernst, D., (2000). Essentialist beliefs about social categories. *British Journal of Social Psychology, 39*, 113–27.偏

Haslanger, S. (2015). What is a (social) structural explanation? *Philosophical Studies*, *173*(1), 113-130.

Leslie, S.-J., Cimpian, A., Meyer, M., & Freeland, E. (2015). Expectations of brilliance underlie gender distributions across academic disciplines. *Science, 347(6219)*, 262-265.

Lombrozo, T., & Gwynne, N.Z. (2014). Explanation and inference: mechanistic and functional explanations guide property generalization. *Frontiers in Human Neuroscience, 8*(September), 700. doi; 10.3389/fnhum.2014.00700

Prasada, S., & Dillingham, E. M. (2006). Principled and statistical connections in common sense conception. *Cognition*, *99*(1), 73–112.

Rangel, U., & Keller, J. (2011). Essentialism goes social: Belief in social determinism as a component of psychological essentialism. *Journal of Personality and Social Psychology, 100*(6), 1056-1079.

Rhodes, M., Leslie, S.-J., & Tworek, C. M. (2012). Cultural transmission of social essentialism. *PNAS*, *109*, 13526–13531.

Sloman, S. (1994). When explanations compete: The role of explanatory coherence on judgments of likelihood. *Cognition, 52*, 1-21.

Vasilyeva, N., & Coley, J.C. (2013). Evaluating two mechanisms of flexible induction: Selective memory retrieval and evidence explanation. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Vasilyeva, N., Gopnik, A., & Lombrozo, T. (2018). The development of structural thinking about social categories. *Developmental Psychology, 54*(9), 1735-1744.

Vasilyeva, N., Ruggeri, A., & Lombrozo, T. (2018). When and how children use explanations to guide generalizations. In T.T. Rogers, M. Rau, J. Zhu, & C.W. Kalish, (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society*. (pp. 2609-2614). Austin, TX: Cognitive Science Society.

Onomatopoeia, gestures, actions and words: How do caregivers use multimodal cues in their communication to children?

Gabriella Vigliocco (g.vigliocco@ucl.ac.uk)
Yasamin Motamedi, Margherita Murgiano, Elizabeth Wonnacott, Chloe Marshall, Iris Milan Maillo, Pamela Perniss

Abstract

Most research on how children learn the mapping between words and world has assumed that language is arbitrary, and has investigated language learning in contexts in which objects referred to are present in the environment. Here, we report analyses of a semi-naturalistic corpus of caregivers talking to their 2-3 year-old. We focus on caregivers' use of non-arbitrary cues across different expressive channels: both iconic (onomatopoeia and representational gestures) and indexical (points and actions with objects). We ask if these cues are used differently when talking about objects known or unknown to the child, and when the referred objects are present or absent. We hypothesize that caregivers would use these cues more often with objects novel to the child. Moreover, they would use the iconic cues especially when objects are absent because iconic cues bring to the mind's eye properties of referents. We find that cue distribution differs: all cues except points are more common for unknown objects indicating their potential role in learning; onomatopoeia and representational gestures are more common for displaced contexts whereas indexical cues are more common when objects are present. Thus, caregivers provide multimodal non-arbitrary cues to support children's vocabulary learning and iconicity – specifically – can support linking mental representations for objects and labels.

Keywords: language development; word learning; iconicity; onomatopoeia; co-speech gestures; child directed speech; naturalistic observation.

Introduction

Understanding how children acquire language, its onset and the developmental path thereafter - is one of the great challenges for the social sciences, with critical implications for education and for intervention in atypically developing children. Vocabulary learning is a central part of language development and is characterized as a hard problem: How do children know that the sounds people produce are 'words' for objects, actions and properties? At the core of most existing proposals is the long-held assumption that language is purely arbitrary: there is no recognizable link between a label and the corresponding referent in the world (e.g. between the English word *dog* and the furry, four-legged animal; de Saussure, 1916). Arbitrariness makes the task of learning words especially hard: how can children learn the correct referent in a visually cluttered world (where multiple objects, actions and properties are all possible candidates for a given label), or even worse, when the objects, actions and properties talked about are absent from the immediate environment?

However, in addition to being arbitrary, language presents also other types of form-meaning mapping characterized by a more transparent and motivated link (Dingemanse et al. 2015). For example, iconicity, across languages, can be found in the phonology of words, e.g. in onomatopoeia such as *meow* or *drip*. This expressive richness is particularly prominent once we look at the multimodal communicative context in which language is learnt: prosodic modulations (e.g. prolonging a vowel to indicate prolonged extension, *loooong*), iconic, representational gestures (e.g., tracing an up and down movement with the index finger while talking about a bouncing object), points and hand actions with objects (e.g., showing a toy hammer to a child or showing how to use the toy hammer) also contribute to the meaning of the message. In vocabulary learning, these iconic and indexical communicative cues may scaffold the mapping between words and world (Perniss et al., 2010; Perniss & Vigliocco, 2014).

Such cues have been previously documented. Onomatopoeia are over-represented early on in children's language development, both in children's vocabularies (Laing, 2014) and in the input they receive (Perry et al., 2017), though this prevalence declines as children age. Points have been reported as the most common gestures used by caregivers especially with very young children (under the age of 2, Iverson et al. 1999; Özçalıskan & Goldin-Meadow 2005), helping to isolate the referent from a complex scene and to link it to the provided label. Though points are common, iconic gestures are also present in parental input from early on in child development (Rowe et al., 2008), and present in the gestural repertoires of children (Acredolo & Goodwyn, 1988). Furthermore, Rowe et al. (2008) showed that parents' gesture use (including points and iconic gestures) predicts children's gesture use, which in turn predicts later vocabulary development, suggesting the importance of such cues for overall language development. Lastly, research has shown a link between direct manipulation of objects (i.e., hand actions) in caregiver-child interaction and children's learning (see Rohlfing 2011 for a review). However, most previous studies focus on a single cue (e.g. gestures or hand actions), rather than considering how the different cues are used together (and together with speech). Cartmill et al. (2013) find that the quality of parental communication, operationalised as how predictable certain words are given the surrounding context (e.g. speech, gesture, surrounding

objects), predicts child vocabulary size at 54 months. Though this suggests that the multiplex nature of child-directed communication might scaffold language learning, they do not analyse the information provided by different cues. Moreover, most studies have focused on learning contexts where label and referent co-occur spatially and temporally (e.g., when objects are present in the visual scene, or words are uttered while actions are ongoing). However, displaced contexts (i.e., when objects are absent) can also provide learning opportunities and previous research indicates that children do learn in these contexts (e.g., Tomasello, Stroberg & Akhtar 1996).

Here, we provide a first investigation that comprehensively assesses the distribution of iconic and indexical cues both in learning contexts in which objects are present, and contexts in which they are absent. We expect to find that iconic cues (onomatopoeia and representational gestures) will be especially important in displaced contexts because iconicity can evoke perceptual or auditory features of the object, in this way providing an imagistic link with the referent and help in bringing it to the 'mind's eye. Both iconic and indexical (points and actions with objects) cues can single out referents, when present, in complex and messy visual scenes, and thus provide cues to solve the referential ambiguity problem.

We use a semi-naturalistic method in which we video-recorded caregivers interacting with their child talking about objects (provided by the experiments) which were either known and unknown to the child. We introduced this manipulation as cases in which the child is unfamiliar with the object and its label are more clearly learning episodes. Moreover, we manipulated whether the objects talked about are either present or absent. We focus on children aged 2 to 3 years old as this is a time of remarkable vocabulary growth in which all critical elements of child-directed language are present, communication about displaced referents is present, and finally, at which children are assumed to be able to understand and produce iconic gestures (Özcaliskan & Goldin-Meadow, 2005).

Method

Participants. Thirty-four caregiver-child dyads participated in the study. The language used between the caregiver and the child was British English. All children included in our sample were aged between 24 and 42 months.

Materials. We used toys from four categories: foods, musical instruments, animals and tools. We chose these categories because they are very common for children of this age and because they offer opportunities for vocal and manual iconicity. We created sets of 6 toys from each of the four categories, such that each set contained 3 toys known and 3 toys unknown to the child (based on parental reports). Toys were selected for each child from a larger set of about 20 toy items per category, each of which were used for a roughly equal number of participants.

Procedure. Caregiver-child interactions took place in the families' homes. Before the session, caregivers were given a list of toy names from our full list and they were asked to indicate whether their child knew those objects and those words. They were also asked to fill in the Oxford Communicative Development Inventory (OCDI)[1]. During the session, two experimenters visited the family, and recorded interactions with two videocameras (one focusing on the caregiver, one focusing on the child and the interaction space). One experimenter checked the correct working of the videocameras while the other carried out the manipulations. The interactions were carried out at a table with the caregiver and the child sitting at 90 degrees from each other. Caregivers were asked to interact with their child in a natural way, as they usually did, but to try to talk about each of the objects provided. Drawings of the set of toys was given to the caregiver to help them remember which toys were in the set. The order of object present vs absent was counterbalanced across participants. When the interaction started with objects present, the experimenter brought to the table 6 toys from one category (e.g., animals) and left the room. The dyad talked about these toys for 3-5mins, then the experimenter re-entered the room, asked the child to help in tidying up the toys and then left the room for the displaced condition asking the caregiver and child to continue to talk (again for 3-5mins) about the toys they just played with. The experimenter then reappeared with a new set of toys until all toy categories had been used. When the toy absent condition came first, the caregiver was asked to begin talking about the toys that were about to come while she was going to get them from another room (caregivers were first familiarised with the toys). After 3-5mins, the experimenter brought in the set of toys, repeating this process for all four categories. The whole recording session lasted approximately 45-60 mins.

Coding of caregiver communication. The caregiver communicative behaviour was coded in the following manner.

(1) Speech. Data was transcribed by *utterance*, which is our unit of analysis (Berman & Slobin 1994). Lexical elements were transcribed further for *onomatopoeias* (including lexical onomatopoeia as well as sound effects) and for explicit mention of the referent (the toys in our sets) label. For each utterance, we coded the *topic*, as the specific toy (or multiple toys) that each utterance referred to, regardless whether labels were produced or not. Utterances were assigned to the known/unknown condition on the basis of their topic. Utterances not about the toy referents were coded as "other" for topic and were not included in any analysis.

(2) Points: gestures (using the index finger or the whole hand) that single out a referent by pointing to it*;*

[1] Performance on the CDI was at ceiling and therefore no analyses including this measure are reported.

(3) Iconic/representational gestures: gestures that represent referents by e.g. depicting aspects of their shape or manipulation.

(3) Hand actions: We coded hand actions and movements performed while holding or manipulating an object. These were divided into (i) *deictic* (i.e., showing) and (ii) *depicting* (e.g., demonstrating the use of a tool). Hand actions were only coded for the toys we provided, and thus can only occur in the toy present condition.

Thus, we distinguish in our coding between iconic cues (onomatopoeia and representational gestures) and indexical cues such as points. Another category is hand actions. Hand actions are indexical in that they direct attention to the referent. They can however, differ and we coded separately those that depicted some properties of the referent (depicting hand actions) from hand actions that showed the object to the child (deictic hand actions). We consider the distinction between iconic and indexical cues to reflect two different manners in which cues can be non-arbitrary: iconic cues *stand for* the object; indexical cues provide a visual link to the object but they do not stand for it. Fig. 1 shows screenshots of the different categories.

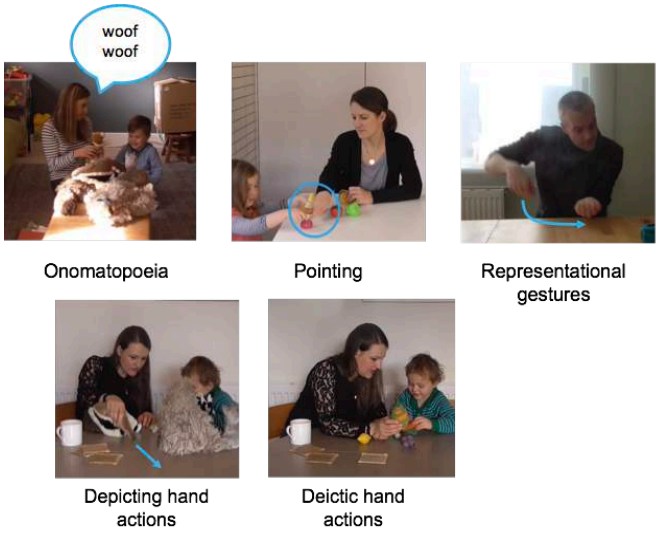


Figure 1. Examples of communicative behaviors coded in the different categories.

Results

Before looking at the distribution of the multimodal cues, we examined the distribution, across our four conditions, of caregiver utterances. Figure 2 illustrates how often parents talk about objects across conditions. Parents talk more when toys are present, and talk more about items unfamiliar to the child than those that are familiar. The larger number of utterances with objects present may indicate that it is easier to maintain the child's attention, or greater ease of production about present objects than about objects that need to be recalled.

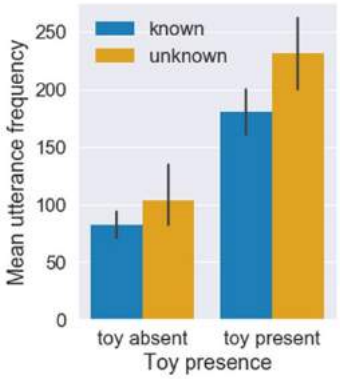


Figure 2. Mean utterance frequency

Iconic and indexical cue use. The primary aim of this study is to understand whether, and how, parents use onomatopoeia, representational gestures, points and hand actions (deictic and depicting) in their interactions with children. As such, we analysed whether age, familiarity (known vs. unknown object) and presence (present vs. absent) affect the use of each cue type.

Analyses use logistic mixed effects models to asses which factors affect the presence of absence of different cues. Age of the child (in months), presence or absence of the object, and familiarity of the label (known/unknown) were included as centered fixed effects, as well as their interaction, and the centered fixed effect of category (category is a control variable). We included a random intercept for participant with random slopes of presence/absence and label familiarity, plus their interaction. Dependent variables are presence/absence of each cue – referent label, onomatopoeia, representational gesture, point, hand action – in an utterance in a given condition). This model structure is used in all models throughout this section, unless otherwise specified. In the interest of space, only the effects of interest are reported here. Full results from the models can be found at https://osf.io/yegxh/.

First, we find that parents make use of all of these cues: approximately 39% of all utterances (11, 755 out of 30,283 utterances) in the dataset are modified by at least one iconic or indexical cue. Figure 3 shows the proportion of each cue across conditions in the study. Second, the proportion of points is low in comparison to the other, especially manual, cues. We attributed this to the affordances of the interaction context: toys were in close proximity to the caregiver and the child, therefore hand actions in this context can take the place of points (indeed, deictic hand actions represent 57% of all hand actions). Deictic and depicting hand actions are by definition only present when objects are present. These are more common for unknown than known objects and their frequency is not modulated by the children's age. *Points* are more common when objects are present, but we find no modulation regarding the familiarity of the label, or based on the age of the child.

Crucially, iconic cues that can be used across all our four conditions (onomatopoeia and representational gestures) show a clear effect of both toy presence and familiarity. In particular, for *onomatopoeia* we find that caregivers use them more often when toys are absent. We also find an interaction between familiarity and toy presence. When toys are present, caregivers use onomatopoeia more with known items. However, when toys are absent, we see the reverse, such that onomatopoeia occur more for unknown items. Interestingly, onomatopoeia decrease as the child's age increases; parents use fewer onomatopoeia with older children. For *representational gestures*, we see that these are overwhelmingly used when toys are absent and for unknown objects. No effect of age is observed. Table 1 summarises the model results.

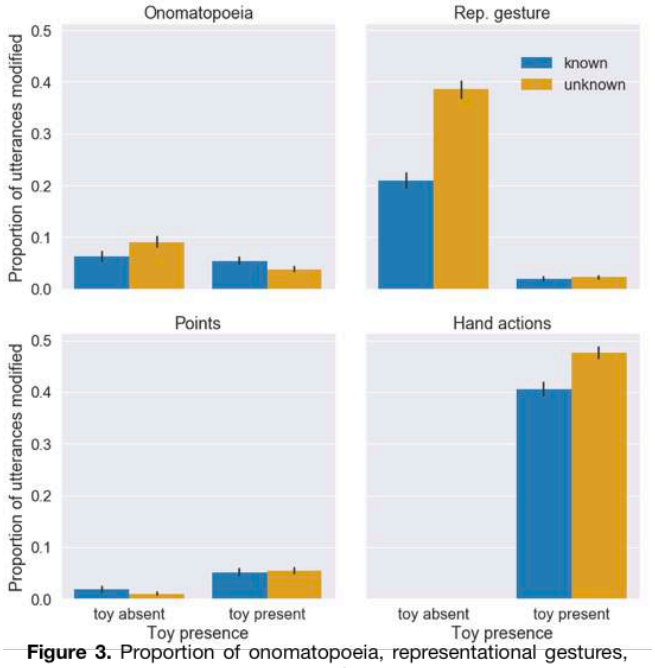


Figure 3. Proportion of onomatopoeia, representational gestures, and points across conditions (toy presence – situated vs displaced, and status – known vs. unknown). Hand actions can only be found in the situated condition. Error bars represent 95% confidence intervals.

Within utterances, cues can co-occur with other cues within and across modalities (e.g., in an utterance we may have a hand action and a representational gesture; or we may find an onomatopoeia and a hand action). Co-occurrences between cues (e.g., between onomatopoeia and hand actions) were remarkably rare in the dataset, occurring in only approximately 3% of parent utterances. Although we have a high proportion of cue modifications, we do not see a high proportion of cases where multiple cues co-occur.

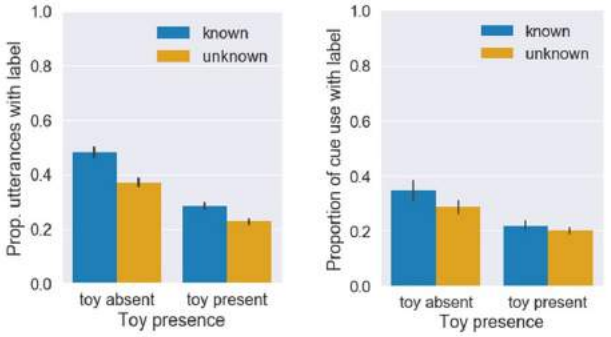


Figure 4. a) Mean proportion of label use, and **b)** Mean proportion of multimodal cues that co-occur with a label.

Label use. Finally, we looked how often parents use explicit labels for the objects (e.g. saying the word 'cat'). We find that parents use explicit referent labels more when the objects are not present, and tend to use the label more for familiar objects than unfamiliar objects. Analysis of a model predicting referent use confirms this: a decrease in referent use for unknown labels, compared with known ones (β=-0.43, SE=0.06, z=-7.60, p<0.001), and a decrease in referent use in the toy present condition, compared to the toy absent (β=-0.90, SE=0.08, z=-11.22, p<0.001). We also address the question of whether and when any of our multimodal cues co-occur with explicit naming of referents. If cues specifically help to link label and referent, then we might expect that use of multimodal cues occur in close proximity to the referent.

Onomatopoeia	β	SE	z	p
Age	-0.05	0.02	-2.39	0.02
Label familiarity	-0.04	0.13	-0.33	0.74
Presence	-0.53	0.14	-3.77	<0.001
Familiarity*Pres.	-0.85	0.21	-4.10	<0.001
Points				
Age	-0.004	0.02	-0.19	0.85
Label familiarity	-0.14	015	-0.95	0.34
Presence	1.62	0.25	6.46	<0.001
Familiarity*Pres.	0.54	0.34	1.60	0.11
Gesture				
Age	0.006	0.03	0.18	0.86
Label familiarity	0.14	0.20	0.70	0.48
Presence	-3.42	0.19	-18.18	<0.001
Familiarity*Pres.	-0.86	0.33	-2.60	0.009
Hand actions				
Age	-0.02	0.02	-1.10	0.27
Label familiarity	0.25	0.06	4.14	<0.001
Hand action type (deictic-depicting)				
Age	0.03	0.03	0.86	0.39
Label familiarity	0.07	0.14	0.45	0.65

Table 1. Summary of model results from logistic mixed effects models. Output variable given in bold. Note that the model for hand action does not include toy presence, as hand actions are not possible in cases where toys are absent.

All of the cues we coded for can co-occur with explicit labelling of the referent (e.g., naming the referent while producing a hand action or representational gesture; naming the referent in the same utterance in which an onomatopoeia is produced). We found that, overall, the multimodal cues occur with explicit naming of a referent approximately 35% of the time. Figure 4 illustrates referent-cue co-occurrence across conditions. We subsetted rows in the dataset where

any of our four cues were produced, to analyse how co-occurrence between cues and labels differed across conditions. The model revealed both an effect of familiarity (β=-0.32, SE=0.08, z=-4.30, p<0.001), and of toy presence (β=-0.58, SE=0.12, z=-4.98, p<0.001). Label-cue co-occurrence occurs more when toys are absent, and when the label is known to the child.

Discussion

The work reported here aimed to characterize the distribution of iconic and indexical cues in the input to 2-3 year-old children. We see that approximately 40% of the clauses produced by caregivers contains at least one of these multimodal cues which often co-occur with explicit labelling of objects especially when toys were present and unknown to the child.

Iconicity as a bridge between words and world

One main goal was to establish if multimodal cues are differentially distributed across contexts: (i) whether the child knows the object and its label, and (ii) whether the objects being talked about are present in the communicative context vs. absent. The latter manipulation has been introduced in order to assess the extent to which the multimodal communicative strategy of caregivers is responsive to the presence vs. absence of object, and whether they modify their language based on the physical setting in which the communication takes place.

The hypotheses we have test is one in which non-arbitrary cues in learning provide a stepping stone to the child to bridge between words and world (Perniss & Vigliocco, 2014). Both iconic and indexical cues can single our referents when these are present in the environment. Moreover, iconic cues can be used when the objects are absent to bring to the mind's eye properties of referents. Thus, indexical and iconic cues may play an important role in learning. We see that this is the case in our data. With the exception of points, which are equally likely for familiar and unfamiliar objects (and labels), all other cues are more commonly used for unknown objects (learning contexts). Crucially, the iconic cues (onomatopoeia and representational gestures) are also used more often when the objects are not present in the physical environment. The results for iconic cues are in line with previous work using a similar paradigm, where it was found that deaf caregivers modify iconic signs in British Sign Language (BSL) to highlight iconic properties of signs (e.g., enlarging the up-and-down movement path of the arm in the sign HAMMER) far more often when objects where absent than present (Perniss et al., 2017), suggesting that this tendency holds across language modalities. In contexts where the label is known to the child, we see that while representational gestures are still overwhelmingly most common for displaced contexts, this is not the case for onomatopoeia. Another interesting difference between onomatopoeia and representational gestures (as well as all other cues) is that onomatopoeia show a decrease with age, in the age range we considered (24-42 months). This finding is in line with previous work showing that use of onomatopoeia in caregivers and children's speech decreases from 0.8 to 2 years (Kauschke & Hofmeister, 2002; Kauschke & Klann-Delius, 2007; Laing, 2014). It has been suggested that for spoken languages, iconicity embedded in wordforms as onomatopoeia may act as a bootstrapping mechanism, a sort of protolanguage, guiding infants' attention to the fact that what comes out of the mouth is linked to what happens in the world (see Imai & Kita, 2014; Laing, 2014).

Iconic vs Indexical cues

We have distinguished points from representational gestures: points don't stand for an object - like representational gestures do - but they direct attention to the object via direct deixis. Just like representational gestures, they are non-arbitrary. Although, in principle iconic cues could be found both when objects are present as well when they are absent, they are far more common in displaced contexts (note however that we observe depicting hand actions in situated contexts); points also, in principle could be found in both contexts, but they are overwhelmingly more common in situated contexts (and deictic hand actions can only be present in situated contexts). This is in line with what was observed in a previous study in BSL where pointing was also much more common in situated than displaced contexts, though the use of indexing to abstract locations in space is common in signed language (Perniss et al., 2017). In the introduction, we mentioned two ways in which non-arbitrary cues can support language development. First, they can help singling out referents when these are present. Both iconic and indexical cues can do this, however, indexical cues may be better placed, as they can be used from earlier age and they provide an unambiguous visual link to the referent. Second, they can help evoking - via imagery - properties of referents that are not present. Iconic cues are best suited to support this type of learning scenario. Note that our study might have called for the use of iconic cues also linked to learning about properties of novel objects. When children were presented with unfamiliar objects, caregivers also used iconicity (especially depicting hand actions) to show the child how the object is used, or how it moves.

How are cues orchestrated?

Previous work suggests that some cues co-occur. For example, Laing et al. (2017) showed that onomatopoeia are usually prosodically marked and Kita (1997) reports that representational gestures tend to co-occur with onomatopoeic and other sound-symbolic words in Japanese. However, we did not observe any tendency for cues to co-occur (although we did not code for prosody in our dataset at this point and therefore we acknowledge that things might be different when considering prosody). For the cues we have considered, it is clear that caregivers choose one cue, presumably on the basis of affordances of the objects (e.g., onomatopoeia for toy animals, representational gesture for tools) to associate to

each utterance. We found that explicit label productions are more likely to co-occur with a multimodal cue when objects are absent and when objects are known to the child. Precisely why this may be the case is unclear; given that labels themselves appear more frequently when toys are absent, it may be that parents use the label in conjunction with cues when toys are absent to make reference to a given object more salient. When the toys are present, it is possible to interact with or point to the toys, making direct reference less necessary.

Conclusions

This study provides a first snapshot of the distribution of multimodal cues in child-directed language. We found a clear indication that iconic as well as indexical cues are well represented in caregivers' input and crucially, they are especially used in those contexts where they may be most useful to children: namely in learning contexts, where the objects and labels talked about are unfamiliar to the child and when the learning occurs in displaced contexts where the objects are not available. It is important to note that the work reported here only provides a partial picture, however. First, the interactions in this study are focussed on contexts of play, which may not be representative of other interactional contexts. Secondly, missing from the current picture is prosodic modulation, which is a key feature of child-directed speech (e.g., Fernald & Simon, 1984; Fernald, 1989; Fernand et al., 1989) and which has been shown to be associated to onomatopoeia (Laing, 2017). Finally, and most important, is the fact that the present work focuses on the communication by the caregiver only, without considering the child's communication, thus giving the impression that the child is a passive receiver of input from caregivers. There is clear evidence this is not the case (e.g. Pereira, Smith & Yu 2014), however, while we plan to code the children's productions, we nonetheless believe that considering the distribution of multimodal cues in caregivers' communication can already provide insight into important questions that has received little attention so far such as which and how cues are used in displaced contexts.

The work reported here was supported by a ESRC grant (ES/P00024X/1) and ERC Advanced Grant (743035) to GV.

References

Acredolo, L., & Goodwyn, S. (1988). Symbolic Gesturing in Normal Infants. *Child Development*, *59*(2), 450–466.

Bates, D., Bolker, B., Machler, M., & Walker, S. C. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Cartmill, E. a, Armstrong, B. F., Gleitman, L. R., Goldin Meadow, S., Medina, T. N., & Trueswell, J. C. (2013). Quality of early parent input predicts child vocabulary 3 years later. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(28), 11278–83.

de Saussure, (1916) *Course in general linguistics*. *New York, NY: McGraw-Hill*

Dingemanse, M., Blasi, D. E., Lupyan, G., Christiansen, M. H., & Monaghan, P. (2015). Arbitrariness, iconicity and systematicity in language. *Trends in Cognitive Sciences*, *19*(10), 603–615.

Fernald, A., Taeschner, T., Dunn, J., Papousek, M., de Boysson-Bardies, B., Fukui, I. (1989). A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *Journal of Child Language* 16, 477–501.

Fernald, A. (1989). Intonation and communicative intent in mothers' speech to infants: Is the melody the message? Child Development, 60, 1497-510.

Fernald, A. & Simon, T. (1984). Expanded intonation contours in mothers' speech to newborns. Developmental Psychology, 20, 104-113.

Imai, M., & Kita, S. (2014). The sound symbolism bootstrapping hypothesis for language acquisition and language evolution. *Phil. Trans. R. Soc. B* 369, 20130298.

Iverson, J.M., Capirci, O., Longobardi, E., & Caselli, M. (1999). Gesturing in mother–child interactions. *Cogn. Dev.* 14, 57–75.

Kauschke, C., & Hofmeister, C. (2002). Early lexical development in German: A study on vocabulary growth and vocabulary composition during the second and third year of life. *Journal of Child Language* 29, 735–757.

Kauschke, C. & Klann-Delius, G. (2007). Characteristics of maternal input in relation to vocabulary development in children learning German. In Gülzow, Insa and Gagarina, Natalia (eds.), *Frequency Effects in Language Acquisition. Defining the Limits of Frequency as an Explanatory Concept*, 181-204. Berlin, New York: De Gruyter Mouton.

Kita, S. (1997). Two-dimensional semantic analysis of Japanese mimetics. Linguistics, 35, 379-415.

Laing, C.E. (2014). A phonological analysis of onomatopoeia in early word production. *First Language* 34(5), 387-405.

Laing, C.E., Viham, M. & Portnoy, T.K. (2017). How salient are onomatopoeia in the early input? A prosodic analysis of infant-directed speech. *J. Child Lang., 44,* 1117-1139.

Özçalışkan, Ş. & Goldin-Meadow, S. (2005). Do parents lead their children by the hand? *Journal of Child Language* 32(3), 481-505.

Özçalışkan, Ş, Levine, S.C. & Goldin-Meadow S (2013). Gesturing with and injured brain: how gesture helps children with early brain injury learn linguistic constructions. *Journal of Child Language, 40*: 69-105.

Pereira, A.F., Smith, L.B. & Yu. C. (2014). A bottom-up view of toddler word learning. *Psychon Bull Rev, 21,* 178-185.

Perniss, P., Lu, J.C., Morgan, G. & Vigliocco, G. (2017). Mapping language to the world: The role of iconicity in the sign language input. *Developmental Science* DOI: 10.1111/desc.12551.

Perniss P, Thompson T, Vigliocco G. (2010). Iconicity as a

general property of language: evidence from spoken and signed languages. *Front. Psychol.* 1, 1–15.

Perniss, P., & Vigliocco, G. (2014). The bridge of iconicity: from a world of experience to the experience of language. *Phil. Trans. R. Soc. B* 369, 20130300.

Perry, L. K., Perlman, M., Winter, B., Massaro, D. W., & Lupyan, G. (2017). Iconicity in the speech of children and adults. *Developmental Science*, *21(3)*, e12572.

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rohlfing, K. J. (2011). Meaning in the objects. *Experimental Pragmatics/Semantics*, 151–176, Amsterdam: John Benjamins.

Rowe, M. L., Özçalışkan, Ş., & Goldin-Meadow, S. (2008). Learning words by hand: Gesture's role in predicting vocabulary development. *First Language*, *28*(2), 182–199.

Tomasello, M. & Kruger, A. (1992). Acquiring verbs in ostensive and non-ostensive contexts. *Journal of Child Language* 19, 311—33.

Tomasello, M. & Barton, M. (1994). Learning words in non-ostensive contexts. *Developmental Psychology* 30, 639-50

Tomasello, M., Strosberg, R., & Akhtar, N. (1996). Eighteen-month-old children learn words in non-ostensive contexts. *Journal of Child Language* 23, 157-176.

Yu, C. & Smith, L. (2012). Embodied attention and word learning by toddlers. Cognition 125, 244-262.

Modeling Ungrammaticality: A Self-Organizing Model of Islands

Sandra Villata (sandra.villata@uconn.edu)
Department of Linguistics, 365 Fairfield Way
Storrs, CT 06269-1145 USA

Jon Sprouse (jon.sprouse@uconn.edu)
Department of Linguistics, 365 Fairfield Way
Storrs, CT 06269-1145 USA

Whitney Tabor (whitney.tabor@uconn.edu)
Department of Psychology and Haskins Laboratories, 406 Babbidge Road
Storrs, CT 06269-1020 USA

Abstract

Formal theories of grammar and traditional parsing models, insofar as they presuppose a categorical notion of grammar, face the challenge of accounting for gradient judgments of acceptability. This challenge is traditionally met by explaining gradient effects in terms of extra-grammatical factors, positing a purely categorical core for the language system. We present a new way of accounting for gradience in a self-organized sentence processing (SOSP) model, which generates structures with a continuous range of grammaticality values. We focus on islands, a family of syntactic domains out of which movement is generally prohibited. Islands are interesting because, although most linguistic theories treat them as fully ungrammatical and uninterpretable, experimental studies have revealed gradient patterns of acceptability and evidence for their interpretability. We report simulations in which SOSP largely respects island constraints, but in certain cases, consistent with empirical data, coerces elements that block dependencies into elements that allow them.

Keywords: whether islands; subject islands; D-linking; acceptability; ungrammaticality; gradient effects; self-organized sentence processing model; SOSP

Introduction[1]

Acceptability judgments are gradient: sentences' acceptability spans from full acceptability to full unacceptability passing through a range of intermediate values which can be statistically distinguished. Grammaticality, on the contrary, is traditionally conceived of as categorical: sentences are either grammatical or ungrammatical but cannot be "partially" (un)grammatical. Degrees of acceptability have been attributed to extra-grammatical factors, such as memory limitations, plausibility etc. It is commonly assumed that this view comes with the advantage of simplicity: a grammar admitting only two states is claimed to be simpler than a grammar involving a continuous, infinite, number of states. We argue that, despite its apparent simplicity, this position is actually less parsimonious than one that accounts for graded acceptability judgments as deriving from the grammar itself. We present a self-organized sentence processing (SOSP) framework, which accounts for gradient effects through a single mechanism of structure building (e.g. Tabor & Hutchins, 2004; Smith & Tabor, 2018; Villata, Tabor, & Franck, 2018). Unlike most classical parsing and grammatical models, SOSP conceives of grammar as residing in a continuous space where fully grammaticality and fully ungrammaticality are two endpoints of a continuum (e.g. Kempen & Vosse, 1989; Cho, Goldrick, & Smolensky, 2017). As a result, gradient effects are understood as generated by the grammar itself, rather than deriving from extra-grammatical factors. To test this theory, we focus on what is arguably one of the most prototypical, and yet also most theoretically challenging syntactic phenomena: islands. Islands are encapsulated syntactic environments out of which almost nothing can be extracted (Ross, 1967). Islands come in two flavors: strong and weak. Strong islands are claimed to block all kinds of extraction. In particular, non D(iscourse)-linked (e.g. *what*, *who*) and D-linked elements (e.g. *which NP*) are equally unextractable from strong islands. This is illustrated in (1) and (2) for subject islands, where the NP (*what* or *which dissertation*) is extracted from a NP subject (*the first chapter of*)[2]:

(1) *__What__ do you think [the first chapter of _] is full of errors?
(2) *__Which dissertation__ do you think [the first chapter of_] is full of errors?

In contrast, weak islands are traditionally claimed to be selective: they prohibit the extraction of non D-linked wh-elements, but allow the extraction of D-linked wh-elements (e.g. Cinque, 1990; Rizzi, 1990). This is illustrated in (3) and (4) for whether islands, where the extraction of the NP is from a whether-clause:

(3) *__What__ do you wonder [whether the student read _]?
(4) __Which book__ do you wonder [whether the student read _]?

The sharp distinction between the examples in (1), (2) and (3) on the one hand, which are standardly deemed ungrammatical, and (4) on the other, which is typically considered grammatical, is very much in line with the traditional, categorical view of grammar, which only admits binary outcomes. However, with the development

[1]This work was supported, in part, by a grant from the Marica de Vincenzi Foundation.

[2]The island domain is in brackets, and the asterisk indicates ungrammaticality.

of finer-grained techniques for gathering acceptability judgments, experimental studies have revealed gradient patterns of acceptability for island effects. Here we focus on three empirical facts indicating gradient island effects.

First, acceptability judgment studies have revealed that weak island acceptability is gradient (e.g. Sprouse, Wagers, & Phillips, 2012; Sprouse & Messick, 2015). In particular, D-linked whether islands (4) are more acceptable than non D-linked ones (3), and yet still ungrammatical, contra the traditional wisdom that conceives of D-linked whether islands as grammatical (see Villata, Rizzi, & Franck 2016 for similar evidence for wh-islands). These studies used a 2x2 factorial design that isolates the island effect from two processing factors that are known to interact with the effect: (i) STRUCTURE TYPE (island vs. non-island),[3] and (ii) DEPENDENCY LENGTH (long vs. short) (5). The contrast between (5a) and (5c) isolates the cost of structure, while the contrast between (5a) and (5b) isolates the dependency length effect. We define the island effect as a statistical interaction between the two factors: it is what remains after the linear sum of the two processing factors is taken into account.

Sprouse & Messick (2015) found a significant interaction for both non D-linked and D-linked whether islands, indicating an island effect in both cases. However, the island effect was stronger in the non D-linked condition as compared to the D-linked condition, providing evidence that D-linking reduces the island effect in weak islands (see Figure 5; empirical data are in black).

(5) Factorial design measuring the whether island effect

a. NON-ISLAND, SHORT
Who/Which woman _ thinks that John bought a car?
b. NON-ISLAND, LONG
What/Which car do you think that John bought _?
c. ISLAND, SHORT
Who/Which woman _ wonders whether John bought a car?
d. ISLAND, LONG
What/Which car do you wonder whether John bought _?

The second empirical fact is that D-linking interacts with island types: while D-linking ameliorates the acceptability of weak islands, it does not help strong islands — e.g., subject islands. Example (6) shows a corresponding factorial design for subject islands, and Figure 6 (black lines) shows the empirical data from Sprouse & Messick (2015).

(6) Factorial design for measuring the subject island effect

a. NON-ISLAND, SHORT
Who/Which leader _ thinks the speech interrupted the TV show?
b. NON-ISLAND, LONG
What/Which speech does the leader think _ interrupted the TV show?
c. ISLAND, SHORT
Who/Which leader _ thinks the speech by the president interrupted the TV show?
d. ISLAND, LONG
Who/Which politician does the leader think the speech by _ interrupted the TV show?

Third, D-linked whether islands with an intransitive embedded verb (e.g., *Which joke does the comedian wonder whether the audience* ***laughed****?*) are less acceptable than those with a transitive embedded verb (e.g.,*Which necklace does the detective wonder whether the thief* ***stole****?*), an effect that was significant for both D-linked and non D-linked whether islands, although it was greater for the former (Villata, Sprouse, & Tabor, 2018) (Figure 1). We take this result as evidence that whether islands, though ungrammatical, are interpreted. This suggests that the dependency between the extracted wh-phrase and the gap inside the island can, at least sometimes, be established.

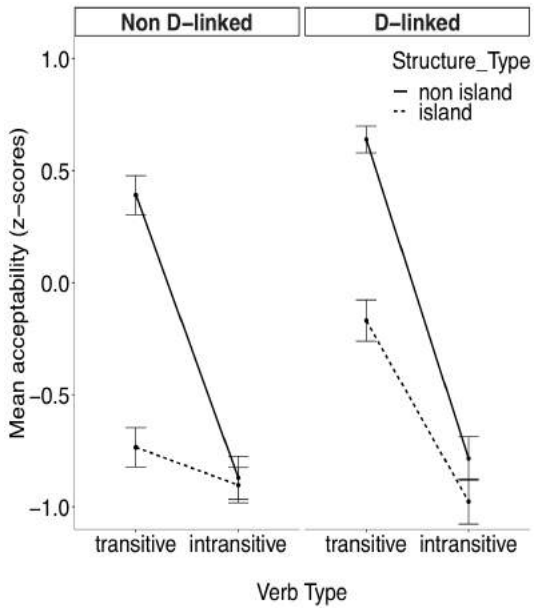


Figure 1: Acceptability proportions for weak islands (data from Villata, Sprouse, & Tabor (2018)).

Summarizing: First, weak islands are ungrammatical. Though D-linking improves their acceptability, it does not cancel the island effect. Hence, their acceptability is gradient. Second, D-linking does not improve the acceptability of strong islands. Hence, gradience is not evident in all cases. Third, the evidence suggests that weak islands are interpreted. Hence, the dependency between an extracted wh-phrase and a gap inside the island can sometimes be established. The last fact seems to point to an account of islands not couched in terms of perfectly impenetrable syntactic domains.

In the next section we introduce the SOSP model. In the section *Model Implementation* we describe SOSP's implementation and, in the section *Simulations*, we describe

[3] "Island" here does not refer to an island-violating structure, but to the mere presence of a structural domain that does not tolerate extractions, like a whether embedded clause or a complex subject.

how the model accounts for the data at hand.

The SOSP Model

In SOSP, structures are formed through continuous dynamical interaction among their constituent elements. Building on several linguistic foundations (Fillmore et al., 1988; Fodor, 1998; Gazdar, 1981) and following the psycholinguistic formulation of Kempen & Vosse (1989), we take the constituent elements to be treelets. Treelets are subtrees formed by a mother node and a finite number of daughter nodes that become active when a word is encountered. Each treelet is associated with a vector of syntactic and semantic features that specifies the properties of the word and its expected dependents. Treelets interact in all possible ways to form structure, creating competition for attachment. Attachments between treelets with a good feature match generally outcompete attachments with a poor feature match, which leads the system to stabilize, most of the time, on a grammatical structure. Structures in which all attachments perfectly satisfy the requirements of the feature vectors of the treelets receive the maximum *harmony value* of 1. Harmony is a formal measure of the degree of coherence in a set of interacting treelets — details below (e.g. Smolensky, 1986).

Importantly, SOSP also allows the generation of intermediate structures, i.e. structures with a harmony value strictly between 0 and 1 (0 harmony = no structure). This happens when an attachment is made between treelets whose features only partially match. This can happen in two ways. First, due to noise in the system, attachments between treelets with a poor feature match can sometimes outcompete attachments between treelets with a good feature match. However, this will happen in a small proportion of the cases, for attachments with a good feature match tend to win competitions. Second, when no optimal bond is available, as in ungrammatical sentences and difficult garden paths, the system forces the attachments to form anyway, generating (sub-optimal) structures. This leads to a variety of differently-valued outcomes which are *internally* generated (i.e. generated by the functioning of the system itself), rather being the result of factors that are external to the system.

Model Implementation

The implemented model (Smith & Tabor, 2018) consists of sets of differential equations that converge on fixed points corresponding to locally optimal structures. Treelets are encoded as banks of feature vectors (all of the same dimensionality, $nfeat$) with one bank for each attachment site (mother/daughters). The general implementation is achieved by first determining all the possible structures (both fully and partially grammatical) that can be formed from the vocabulary of the language, treating the concatenated banks of features and link values as forming a single vector space, and identifying the location of each locally optimal structure in this space. The local harmony, h_i, associated with such a point in the feature space is given by (1):

$$h_i = \prod_{l \in links} \left(1 - \frac{dist(\mathbf{f}_{l,daughter}, \mathbf{f}_{l,mother})}{nfeat}\right) \quad (1)$$

where $dist(\vec{x} \cdot \vec{y})$ is Hamming distance between $\vec{x}$ and $\vec{y}$. In other words, the local harmony is a product, across links, of a measure of similarity between the daughter feature vector $\mathbf{f}_{l,daughter}$ and the mother feature vector $\mathbf{f}_{l,mother}$ on the end of each link. Thus, if every link has a perfect match, then h_i is maximal and equals 1. The minimum possible value is 0, and various degrees of mismatch give intermediate values.

For each such structural locus, we specify a radial basis function (RBF), ϕ_i (Muezzinoglu & Zurada, 2006):

$$\phi_i(\mathbf{x}) = \exp\left(-\frac{(\mathbf{x}-\mathbf{c}_i)\ (\mathbf{x}-\mathbf{c}_i)}{\gamma}\right)$$

Here, $\mathbf{x}$ is the state of the system encoding the values of all features on all activated treelets and all possible links between them, $\mathbf{c}_i$ is the location of the ith (partial) parse, denotes the vector transpose, and γ (a free parameter) specifies the width of the RBFs. We define the harmony function $H(\mathbf{x})$ as the height of that RBF among n RBFs that is maximal at x, where n is the number of optimal and partially-optimal structures (harmony peaks) that can be formed with the currently activated elements, and h_i is the height of the i'th mode:

$$H(\mathbf{x}) = \max_{i \in 1...n} h_i \phi_i(\mathbf{x}) \quad (2)$$

This equation interpolates a harmony landscape between the structural loci, c_i, associated with the local harmony peaks.[4]

Parsing starts with all features equal to 0. The perception of the first word of a sentence causes features of a lexical treelet associated with that word to be turned on. This, in turn, causes links and additional treelet feature banks corresponding to the most viable parse of just that word to be turned on. In SOSP, treelets are interacting subsystems that attempt to assemble themselves through local interactions that locally maximize harmony. This is implemented as noisy gradient ascent on the harmony surface, $H(\mathbf{x})$:

$$\frac{d\mathbf{x}}{dt} = \nabla_{\mathbf{x}} H(\mathbf{x}) = -\frac{2}{\gamma} h_{i_{max}} (\mathbf{x} - \mathbf{c}_{i_{max}}) \phi_{i_{max}}(\mathbf{x}) + \sqrt{2D}\, dW \quad (3)$$

where $D > 0$ scales the magnitude of the Gaussian noise process dW. In other words, the system moves approximately uphill on the harmony landscape as it processes each word. Moving uphill is equivalent to growing the link structures and adjusting values of unspecified or conflicting features to

[4] This definition differs from the form specified in Smith & Tabor (2018) who summed the RBFs to form H. We have found that, in systems with many harmony peaks, if a summation is used, there are often ganging effects that influence the structure of the gradient and flummox effective parsing: many proximal ungrammatical structures gang together to pull the state toward their mean and away from a lone worthy grammatical candidate. Humans seem to be strongly influenced by the presence of a good candidate even if there are also many bad ones around, so the max method yields more plausible parsing than does the summing method when the language model is realistically rich.

reach a locally optimal parse state. After a local optimum is reached, new features specified by the next word are turned on (moving the system off its current hilltop and into a nearby valley). The gradient ascent process then begins anew and a new harmony maximum is reached, corresponding to the next step of the parse. Across multiple trials, the noise produces a distribution over the harmony maxima, generally favoring those that correspond to plausible parses of the input seen up to the present moment. At the end of parsing a sentence, the system will be at a particular harmony peak that has a value between 0 and 1. We take this harmony value to correspond to the model's assessment of the acceptability of the sentence.

Simulations

In the terms of the model, based on the empirical results reviewed above, we identify the following desiderata: (i) non D-linked whether islands should receive a low harmony value, and D-linked whether islands should receive a higher, but not maximal, harmony value; (ii) the high-but-not-maximal harmony for D-linked whether islands should be generated by linking the gap to the filler inside the island with some strain, in line with experimental findings suggesting that these structures are interpreted (Villata, Sprouse, & Tabor, 2018); (iii) subject islands should receive a low harmony value irrespective of the presence of D-linking (comparable to non D-linked whether islands) .

Figure 2 portrays the model's processing of a non-D-linked whether island. The model considers, in parallel, all conceivable parses of the input string. However, since many of these parses have extremely low harmony and do not have much influence on the processing, the figure only shows those that play a significant role in the parsing dynamics. One reads the figure from left to right and bottom-up.[5] Typically, when a word is perceived, bonds between treelets form. For example, when "you" is perceived, a bond between "NP_{you}", the mother of "you", and "NP_{you}", the daughter of "S/NP_{what}", typically forms. Bonds between treelets that are formed by the system are illustrated with dashed lines, while straight lines indicate the treelet's structure as it is defined in the lexicon based on phrase structure rules (e.g. S $\rightarrow$ NP VP). Crucially, the treelet feature vectors are mutable within a range of values corresponding to the syntactic/semantic range that the treelet affords. For example, in the present case, the mother of the "S $\rightarrow$ NP VP" treelet has mutated to acquire a slash buffer that specifies the syntax and semantics of the fronted element "what". Due to this mutation, links can often achieve a perfect feature match, causing the relevant term in Equation (1) to take on the value, 1.

The crucial developments in the case of the non-D-linked whether sentences occur when the words "wonder" and "whether" are perceived. As shown in Figure 2, the system is deciding between two possible, not-fully-grammatical structures at "wonder". The first, shown by the left

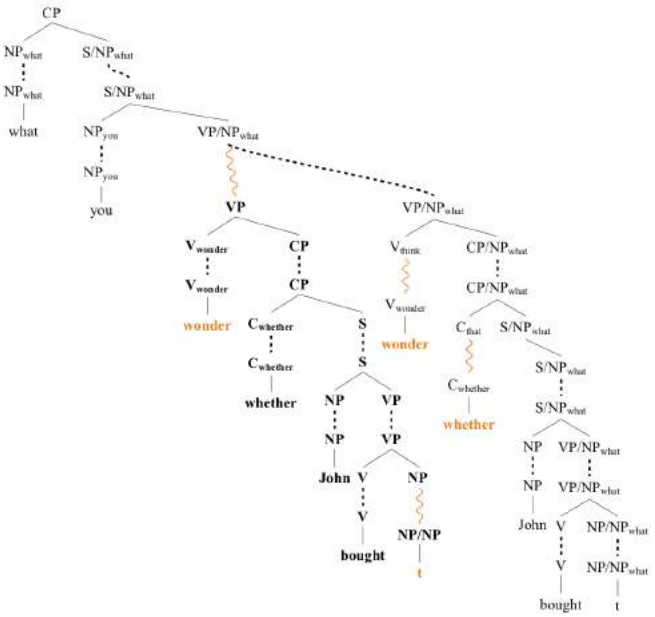


Figure 2: Simplified tree for non D-linked whether island. Subscripts indicate which feature has been transmitted to the node (e.g. S/NP_{what} means that *what* has been propagated to the S node). Words in orange are those that trigger coercion. Wavy orange lines illustrate the dynamic of the coercion. Here the parsing that ultimately wins out in most trials is the one on the left branch (bold font).

VP-branch, respects the constraint imposed by the verb "wonder", which cannot take as a complement an element with a slash feature. This implements the "islandhood" of the CP-complement of "wonder". This parse makes it possible for "$V_{wonder} \rightarrow$ wonder" to attach with perfect harmony to its CP-complement, but at the cost of failing to propagate the slash buffer ("$/NP_{what}$") onto the VP node below. We assume this failure has a cost, but not a severe cost because "what" is a very abstract element, so its encoding plausibly contains only a few features — that is to say, the difference between the "NP_{what}" slash buffer and an empty slash buffer is a small difference. This mild penalty is indicated by the orange color of the link between "VP/NP_{what}" and "VP". The second parse in Figure 2, shown by the right-branch, takes an opposite approach: it propagates the slash buffer, "$/NP_{what}$", onto the VP node below, but it can only do this by coercing "wonder" into a verb that licenses slash propagation. We have taken the verb "think" as a canonical example of such a verb. The orange squiggly line from "V_{think}" to "V_{wonder}" indicates this penalty. Again, this penalty is not extreme because the semantics and syntax of "wonder" and "think" are fairly similar. At the next word, "whether", the system undergoes a second coercion, of "whether" into "that". This coercion, which is triggered by the requirements of the verb "think", allows the slash buffer to propagate down the tree, for "that", unlike "whether", does not act as a slash propagation blocker. As before, although this coercion comes at a cost, the cost is mild, because of the similarity of the two complementizers.

[5]The model employs slash-propagation (Gazdar, 1981) to implement long-distance dependencies.

We now consider the case of the D-linked whether-island, illustrated in Figure 3. In this case, not propagating the slash feature onto the VP node (the parse on the left-branch) comes with a strong penalty (illustrated by the red squiggly line in the figure). This is because, D-linked words, unlike non D-linked ones, are associated with a rich bundle of semantic and syntactic features. As a result, failing to propagate the features associated with D-linked NPs incurs a strong penalty. The system therefore tends to prefer the second parse (right-branch): the close analogy between "think" and "wonder", and "that" and "whether" makes the mild coercions option better than any other parsing option, leading the parser to stabilize on the option that propagates the slash buffer inside the whether island.

In subject islands, on the contrary, there is no such close analogy between the words in the sentences and alternatives in the lexicon. As a result, the parser systematically fails to consider the possibility of propagating the slash feature down the subject branch and then is caught up short when a gap appears in the subject, and no gap appears in the main verb phrase (see Figure 4).

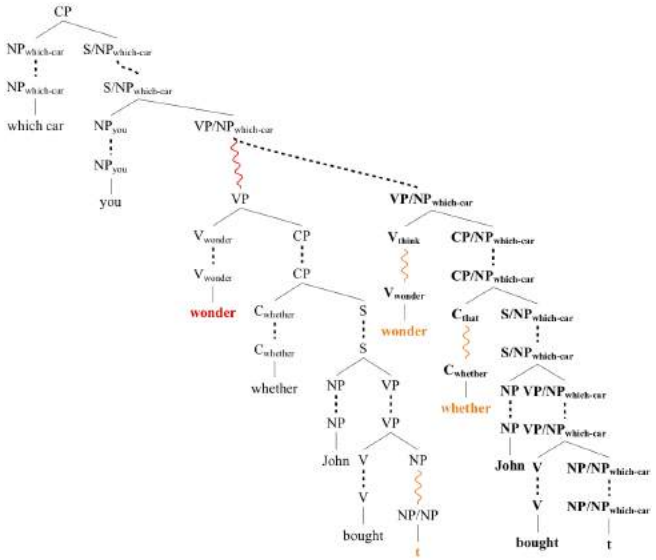


Figure 3: Simplified tree for D-linked whether island. Words in orange trigger a mild coercion, while words in red trigger a severe coercion. Here the parsing that ultimately wins out in most trials is the right branch (bold font).

We ran 20 runs of the model on simplified versions of each sentence in examples 5 and 6 (no determiners, ignoring English do-support). The model, somewhat revised from the one described in the first, reviewed version of this paper, is both an elaboration and a simplification of the model described in Smith & Tabor (2018).[6] It used 45

[6]We describe the revised model here rather than the original one because its assumptions are more plausible and easier to describe, as requested by several anonymous reviewers, and the causal dynamics by which it produces the data points reported here—specified in the analyses above—are the same as those previously described.

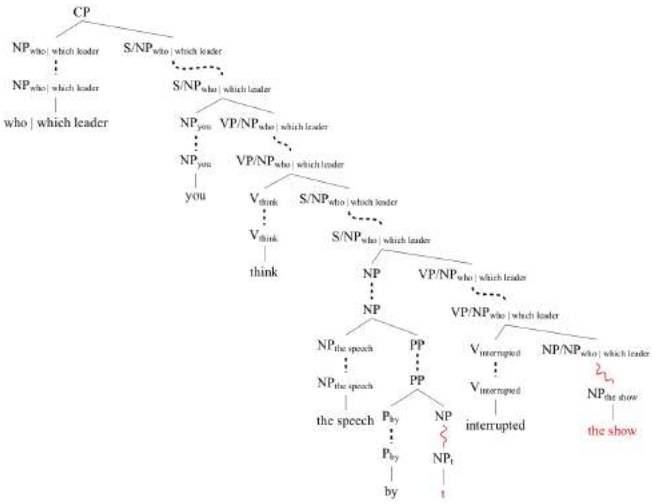


Figure 4: The low-harmony structure that stabilizes when the model is presented with a subject island.

distinct feature vectors for coding the lexical and syntactic nodes needed for the tree configurations described in the analyses above (as well as variants needed for all the stimulus sentences listed in (5) and (6)). Whereas previous versions of the model were hand-coded with roughly plausible linguistic features, the current version started by generating random bit vectors in 20 space for each feature vector (which was either a mother or a daughter of a treelet). This made all the feature vectors relatively distant from one another. Then, in keeping with the hypotheses described above, the vector for "wonder" was made to be equal to the vector for "think" except in two dimensions where it had contrasting bits; the vector for "whether" was analogously made similar to the vector for "that"; and the vector for "CP/What" was analogously made similar to the vector for "CP" (i.e., to CP with an empty slash buffer). SOSP entertains a plethora of possible ways of combing the treelets, most of which give rise to very low harmony structures. In the simulations reported here, motivated by the assumption that many of the low-harmony variants have little effect on the parse trajectory and to simplify implementation, we only considered the variants that we have mentioned as alternatives in the analyses above. An earlier version of the model had trouble telling sentences apart if many were included in the stimulus set. Here, we introduced a two-fold magnification of the dimension coding the features for the lexical elements. This effectively moved the harmony peaks for sentences with different word forms farther apart from one another, causing the system to prefer parses that are faithful to the input, though not rigidly—see Levy (2008). To allow the model to detect harmony maximization upon processing of each word, we allowed the dynamics to settle through a quadratic velocity profile: the model had to speed up (associated with reaching the steep section of one of the RBF humps) and then slow down (indicating that it was topping out on a harmony maximum) before moving on to the next word.

In addition to the number of feature dimensions (20) and the degree of lexical isolation (2 x) mentioned above, important free parameters are γ, specifying the width of the RBFs, D, specifying the magnitude of the noise, and ρ which takes its values in [0, 1] and determines how far over the velocity "hump" the model must travel before moving to the next word ($\rho = 1$ implies immediate transition, $\rho = 0$ implies infinite processing time per word), and Δt which specifies the step size in the Euler Integration that we used to approximate the dynamics. We explored these parameters by hand finding a way to roughly optimize behavior in a test grammatical sentence and the D-linked whether island extractions (D-linked, whether, island, long) to establish the settings $\gamma = 4$, $D = 7 \times 10^{-1}$, $\rho = 0.4$, $\Delta t = 0.5$ and then examined the results in the other fourteen conditions.

One other point about the implementation is particularly important. The current versions of SOSP add dimensions to the state space with every new word (these dimensions correspond to the feature banks in treelets that the word introduces, and to the links this treelet can potentially form with other activated treelets). The behavior of the dynamical equations is sensitive to the dimensionality, so to achieve reasonable parsing, such an implementation needs to change the dynamical parameters (γ, D, ρ, Δt) as the sentence grows. We do not think this is very plausible. Instead, we think the dimensionality of human processing is kept roughly constant via a focusing mechanism (possibly related to what is called "Working Memory" in other work). We suspect that the form of this focusing involves fractal scaling as has been proposed in work on neural encoding of arbitrary dependency languages (Plate, 2003; Tabor, 2000). However, we do not know how to apply such scaling techniques to the SOSP encodings, so we have used a kind of Poor Man's focusing method: run the dynamics on just the vectors associated with the current word and the previous word. Coupled with slash propagation, this technique is capable of tracking of all the dependencies needed for the current stimuli.

Figures 5 and 6 present a comparison of the predicted island effects by the SOSP model (in red) and the observed island effects (in black)—the model exhibited very little variance within trials so no model error bars are shown.[7] Indeed the qualitative behavior of the model matched the desiderata we have mentioned, often succeeding in linking the gap to the fronted element in D-linked whether islands, and in extractions from non-islands, but not in the subject islands, and rarely in the non-D-linked whether islands.

[7]For subject islands, Sprouse & Messick (2015) report a reverse effect of D-linking, with D-linked subject islands showing a stronger interaction than non D-linked ones. However, this reverse D-linking effect appears to be driven by the non-island/long condition, which exhibited lower ratings in the non D-linked than the D-linked condition. Although it is unclear what might have driven this effect, for current purposes it is sufficient to observe that the reverse D-linking effect is not driven by the island condition itself. Moreover, the ratings for the island condition are comparable with those obtained by the non D-linked whether island, and also by the Complex NP and adjunct islands tested by Sprouse & Messick (not reported here), for which no reverse D-linking effect was observed.

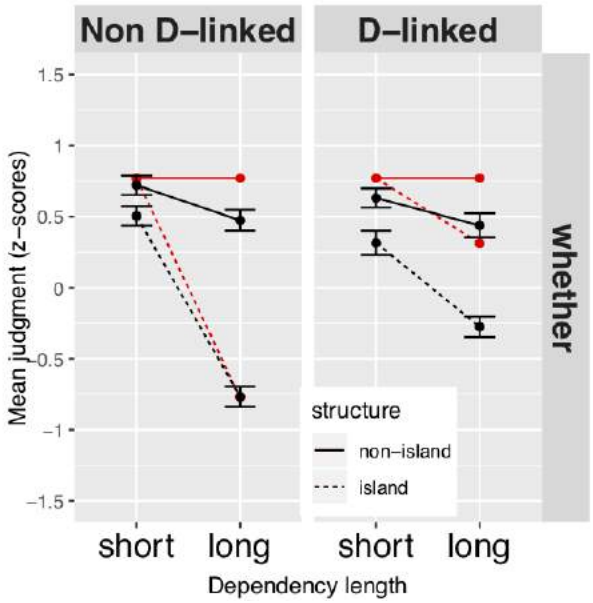


Figure 5: Interaction plots for whether island. The points correspond to the 4 conditions in (5). Empirical results are in black (data from Sprouse & Messick, 2015) and results from the model's simulation are in red.

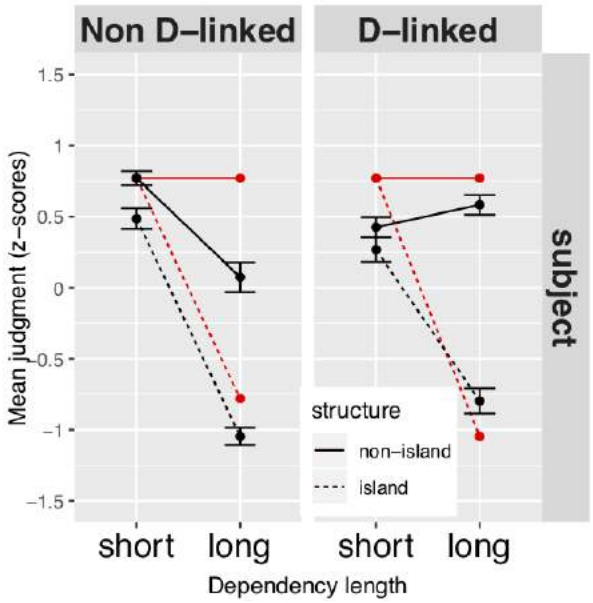


Figure 6: Interaction plots for subject island. The points correspond to the 4 conditions in (6). Empirical results are in black (data from Sprouse & Messick, 2015) and results from the model's simulation are in red.

Discussion

We reported three empirical findings from the literature pointing to gradient effects in island acceptability. We presented a new way to account for islands' ungrammaticality in a self-organized sentence processing (SOSP) framework.

SOSP's key novelty lies in its conception of grammatical states as lying in a continuum of grammaticality values. As a consequence, and unlike traditional theories of grammar, SOSP treats degrees of acceptability as deriving from the grammar itself, rather from extra-grammatical factors. This occurs because self-organizing treelets, not being under the control of a central coordinator, build whatever structure they can, sometimes achieving only partial coherence. This is the case of D-linked whether islands: the system succeeds in coercing them into a non-island structure, leading to the propagation of the slash feature inside the island, thus rendering these structures interpretable, in line with empirical findings. However, coercion comes at a cost, which is what causes the sentence to be given a suboptimal harmony value by the model, thus accounting for the fact that D-linked whether islands, although improved as compared to non D-linked ones, are still degraded. Importantly, the system is also able to generate extreme grammaticality values, in line with classical models. On the ungrammatical side, this happens when no grammatical parse is available and no coercion can take place, either because no grammatical structure is similar-enough to the to-be-parsed structure or because the system is not sufficiently prompted in undergoing the coercion. The first case is illustrated by subject islands, where no alternative (coerced) parse is available. The second case is illustrated by non D-linked whether islands: here the non D-linked wh-phrase is not powerful enough to cause the system to discover the coercion, resulting in failure to propagatied the slash feature inside the island, and very low harmony value.

Shortcomings of the current model are that the treelet forms are based on linguistic theorizing, not on a machine-learning method. A machine learning approach would make the method more completely formalized. Also, the feature vector composition, which ends up determining the harmony values, was mainly random. It will be valuable to explore more realistic feature analyses motivated by linguistic theory. Finally, as noted above, it is desirable to find a more principled method of keeping the state space finite.

All in all, we argue that SOSP offers a valuable new way of approaching the relationship between grammar and processing. It is closely related to generative linguistic theory. Nevertheless, it differs in non-trivial ways from traditional assumptions, notably continuity, and a central role for processing in grammatical explanation. We hope our results will spur new discussion on these topics.

References

Cho, P. W., Goldrick, M. A., & Smolensky, P. (2017). Incremental parsing in a continuous dynamical system: Sentence processing in dradient symbolic computation. Linguistic Vanguard, 3, 76-96.

Cinque, G. (1990). Types of Ā-dependencies. MIT press.

Fillmore, C., Kay, P., & O'Conner, M. C. (1988). Regularity and idiomaticity in grammatical constructions: the case of let alone. Language, 64, 501–538.

Fodor, J. D. (1998). Unambiguous triggers. Linguistic Inquiry, 29, 1–36.

Gazdar, G. (1981). On syntactic categories. Philosophical Transactions (Series B) of the Royal Society, 295, 267-83.

Kempen, G., & Vosse, T. (1989). Incremental syntactic tree formation in human sentence processing: a cognitive architecture based on activation decay and simulated annealing. Connection Science, 1, 273–290.

Levy, R. (2008). A noisy-channel model of rational human sentence comprehension under uncertain input. In Proceedings of the 13th conference on empirical methods in natural language processing (p. 234-243).

Muezzinoglu, M. K., & Zurada, J. M. (2006). RBF-based neurodynamic nearest neighbor classification in real pattern space. Pattern Recognition, 39, 747–760.

Plate, T. A. (2003). Holographic reduced representation: Distributed representation for cognitive structures. Stanford, CA: CSLI Publications.

Rizzi, L. (1990). Relativized minimality. MIT press.

Ross, J. R. (1967). Constraints on variables in syntax. (Ph.D. thesis, MIT)

Smith, G., & Tabor, W. (2018). Toward a theory of timing effects in self organized sentence processing. In Proceedings of the 16th international conference on cognitive modeling (p. 138-143).

Smolensky, P. (1986). Foundations of harmony theory. In D. E. Rumelhart, J. L. McClelland, & the PDP Group (Eds.), Parallel distributed processing, volume I (pp. 194–281). MIT Press.

Sprouse, J., & Messick, T. (2015). How gradient are island effects?. (Poster presented at NELS 46)

Sprouse, J., Wagers, M., & Phillips, C. (2012). A test of the relation between working-memory capacity and syntactic island effects. Language, 82-123.

Tabor, W. (2000). Fractal encoding of context-free grammars in connectionist networks. Expert Systems: The International Journal of Knowledge Engineering and Neural Networks, 17(1), 41-56.

Tabor, W., & Hutchins, S. (2004). Evidence for self-organized sentence processing: Digging in effects. Journal of Experimental Psychology: Learning, Memory, and Cognition, 30(2), 431-450.

Villata, S., Rizzi, L., & Franck, J. (2016). Intervention effects and relativized minimality: New experimental evidence from graded judgments. Lingua, 179, 76-96.

Villata, S., Sprouse, J., & Tabor, W. (2018, March). Modeling ungrammaticality: A self-organizing model of islands. (Poster presented at the CUNY conference on sentence processing)

Villata, S., Tabor, W., & Franck, J. (2018). Encoding and retrieval interference in sentence comprehension: Evidence from agreement. Frontiers in psychology, 9(2).

The End's in Plain Sight: Implicit Association of Visual and Conceptual Boundedness

Jonathan Wehry (jonwehry14@gmail.com)
Department of Psychology, University of California, Berkeley
2121 Berkeley Way, Berkeley, CA 94720

Alon Hafri (ahafri@gmail.com)
Departments of Cognitive Science and Psychological & Brain Sciences, Johns Hopkins University
237 Krieger Hall, 3400 N. Charles Street, Baltimore, MD 21218

John Trueswell (trueswel@psych.upenn.edu)
Department of Psychology, University of Pennsylvania
425 S University Ave, Philadelphia, PA 19104

Abstract

What are the categorical distinctions shared between conceptual and visual representations? One distinction may be between bounded and unbounded entities. Previous research in sign language has shown that even non-signers associate signs with repetitive motion with atelic verbs, such as "run", and signs with sudden motion with telic verbs, such as "arrive". In our first study, we show this distinction holds even when the visual stimuli depicted bear no intrinsic linguistic reference: we used non-linguistic random dot motions. In our second study, we demonstrate this association occurs spontaneously, even when subjects are not making explicit semantic judgments about verbs. We use a cross-modal lexical decision task in which verbs and non-words appear superimposed on bounded or unbounded dot stimuli. We find congruency when the motion boundedness matches the conceptual boundedness of the verb. Together, these studies provide evidence for an automatic link between visual and conceptual boundedness in the mind.

Keywords: telicity; motion perception; visual boundedness

Introduction

Language allows us to describe our own perceptual experience and understand the experiences of others. As social creatures, this is critical. Thus, an understanding of how and what information is common between language and perception is both interesting and important. One category of information that might be shared across these two systems is boundedness. In the current work, we investigated whether the visual perceptual system makes a distinction between bounded and unbounded stimuli, and whether this distinction is common across visual and linguistic experience (i.e., through verbs and verb phrases).

To understand what we mean by visual boundedness, imagine you're observing an event (e.g., something moving across your field of vision), but everything is blurred so that you cannot make out objects or what category of event is taking place. Because the low-level motion properties of the scene are preserved, you could still perceive the motion properties of the event, such as whether it started and stopped. In other words, there are perceptual correlates of boundedness even when you do not have access to high-level information about objects, goals, or events.

A second form of boundedness is telicity, or conceptual boundedness. Telicity is a similar concept to visual boundedness but in the linguistic domain. Telicity refers to whether an event described by a verb or verb phrase is construed as having an intrinsic endpoint (telic) or an undefined one (atelic; Vendler, 1957). For instance, "run" is an atelic verb. While a person could not run forever, the verb itself does not entail an endpoint. This is as opposed to a verb such as "arrive." There is a definite endpoint entailed by the verb such that the event has only occurred when someone arrives at their destination. A simple test for this distinction is to probe the felicity or grammaticality of a sentence when adding the phrases "for an hour" (atelic) and "in/within an hour" (telic; Todorova, Straub, Badecker, & Frank, 2000). For example, one could say someone *ran for an hour* but could not say someone *ran in an hour*; conversely, it is infelicitous to say someone *arrived for an hour* but fine to say someone *arrived within an hour*.[1]

One way to investigate the link between visual and conceptual boundedness is through sign language, as sign language is inherently both linguistic and visual. In Malaia & Wilbur (2012), signers were instructed to produce signs for verbs, and motion capture technology was used to record the maximum deceleration, maximum velocity, and duration of the signs. It was found that the motion properties of the signs for atelic verbs, e.g. "run", are consistent with one another, and visually distinct from telic verbs, e.g. "arrive". Such findings suggest that signs carry information about verb telicity iconically in the form of the sign itself.

However, although this study showed that a difference in telicity may be visually distinct, it does not indicate whether humans have access to this boundedness distinction (whether implicitly or explicitly). Strickland et al. (2015) addressed this issue: they demonstrated that even among non-signers, there is an implicit bias to map atelic signs (i.e.

[1] However, this rule is not absolute. For example, the telic verb die can be used with both "in an hour", an instantaneous event, and "for an hour", an extended process with an undefined endpoint.

signs for atelic verbs) onto atelic verbs and telic signs (i.e. signs for telic verbs) onto telic verbs. In that study, English-speaking individuals without sign language experience were shown an atelic or a telic sign and were forced to choose one of two verbs that they believed matched the meaning of that sign. For example, participants viewed the sign for "float" (an atelic verb) and were asked to choose between two words that differed in telicity (e.g., "float" vs. "leave"). Participants significantly preferred the verb that matched the sign in telicity, even when neither verb referred to the true meaning of the sign (e.g., "talk" vs. "buy"), and even for verbs with no visual correlate (e.g., "think" vs. "decide"). This shows that the human mind has access to boundedness information in visual input and can associate it implicitly with word meanings that are conceptually bounded, even though iconicity for telicity does not exist in their own language.

Although the Strickland et al. (2015) results are compelling, the scope of their conclusions is limited to perception within linguistic communication. Sign language is inherently linguistic and referential. Thus, participants can presume that these visual cues have specific linguistic meanings. This raises the question of whether these results only hold when people are performing a task where they must map from one language to another (even if one of the languages is a visual one that they have no knowledge of).

We address this question here via a new set of experiments. We used visual stimuli that were not overtly referential. Participants were shown non-linguistic motion composed of scrambled dots (extracted from biological motion stimuli) that could not be recognized as interpretable events, but nevertheless contained motion information consistent with bounded or unbounded events.

In the first experiment, participants were asked to make atelic vs. telic verb choices after viewing the visual stimulus, just as in Strickland et al. (2015). As these random dot motions are not linguistic or referential, positive findings would offer strong support for a connection between visual and conceptual boundedness. In the second experiment, we test whether such an association is automatic. In a cross-modal lexical decision task, we observed a congruency effect, such that participants were faster to confirm that a stimulus was a word when the background motion matched the boundedness of the displayed verb.

Experiment 1: Verb-Motion Matching

In this experiment, each trial consisted of the participant viewing a 3-second video clip of scrambled moving dots derived from biological motion stimuli, after which the participant had to indicate which of two visually presented verbs (one atelic and one telic) best described the clip. The video clip was designed to depict an unbounded or a bounded event as determined by separate ratings of the repetitiveness of the motion. It was predicted that participants would be more likely to select telic verbs for bounded events and atelic verbs for unbounded events. Following Strickland et al. (2015), effects of motion boundedness were tested within three different semantic domains of verbs: Physical (e.g., *fly* vs. *hit*), Social (e.g., *argue* vs. *give*) and Mental Verbs (e.g., *think* vs. *decide*). If effects hold for all three types of verbs it suggests that motion boundedness is linked to the abstract notion of telicity rather than, for example, spatial-motion aspects of the events denoted by these verbs.

Method

Participants Twenty-four participants were recruited from the University of Pennsylvania undergraduate body and participated for course credit. This was the same number as in the Strickland et al. (2015) study as their effect size was not available for a power analysis. All participants were fluent speakers of English with normal or corrected to normal vision.

Visual Materials A personal computer running the Psychophysics Toolbox Version 3 for MATLAB was used to run this experiment (Kleiner et al., 2007). Sixty biological motion, or biomotion, videos of three seconds from the CMU Graphics Lab Motion Capture Database were used with the BioMotion Toolbox (van Boxtel & Lu, 2013).

Biomotion videos are produced via motion capture, whereby each joint on a person's body is attached to a sensor. The positions of these sensors are then recorded during movement. This produces a video, composed of dots, in which the overall shape, size, and movement of an individual is maintained but the fine details and body form are removed.

Crucially, in our versions, body structure information was removed from these videos while preserving the overall motion signal. This was done by randomizing the start point of each individual dot, but then preserving its relative motion path from that start point. For example, the dot that corresponded to the person's right elbow may, at the start of the animation, be located to the left of where their left ankle was and the dot corresponding to their left ankle may now start right above where their right knee was. This removes the benefits of being able to tell what action is occurring (because the intact structure of the body is removed) and ensures that participants only get information about the motion properties of the dots, e.g., velocity and acceleration.

Selection and Norming of Video Materials Videos were initially selected by JW, and then their boundedness was confirmed using a norming procedure. JW rated a random set of 574 scrambled videos from the CMU database on perceived boundedness. Subsequently, we presented 79 unbounded and 61 bounded candidate videos to twenty-two undergraduates in a norming study. Although the CMU database includes descriptions for each video, they were ignored for video selection. Using these ratings, we chose a set of 60 videos to use in the subsequent experiments.

Participants were asked to rate each video for repetitiveness or deceleration (between subjects) on a scale

of 1 to 7, e.g "Rate the video based upon how repetitive you think the motion is" or "based upon how fast you think the motion decelerates." These properties were used as they were found to be indicators of boundedness in previous studies (Malaia & Wilbur, 2012; Strickland et al., 2015).

Although our intention was to define bounded videos as those with the highest deceleration and least repetition in the motion, deceleration ratings proved inconclusive as across all videos there was little deviation from the average of 4. That is, across all 140 videos participants tended to choose a middle value (*SD* = 0.85). This was perhaps due to the difficulty of the task and the nature of the stimuli (independently moving dots). In contrast, repetition ratings had high variety across items and participants (*SD* = 1.51).

As a result, only the repetition ratings were used to select the sixty videos for the main study. The sixty videos were selected by taking the 40 videos with the highest average ratings and the 40 videos with the lowest average ratings and then sorting these videos by lowest standard deviation across ratings. The 30 videos from each group with the lowest standard deviation were then selected. We considered the videos with high repetition ratings as Unbounded and videos with low repetition ratings as Bounded. The mean repetition rating for unbounded videos was 5.70 and for bounded videos was 1.91, and the two groups differed reliably (p = <0.0001).

Verbal Materials Five atelic and five telic verbs were chosen from each of three separate conceptual domains: physical, social, and mental. This resulted in fifteen telic and fifteen atelic verb pairs. Each pair consisted of one telic and one atelic verb, approximately matched for log frequency. Fourteen of the 18 Strickland et al. verbs were used (4 not used due to low frequency), with an additional sixteen verbs (seven telic and nine atelic) generated by author JW to maintain approximate match in log frequency. The verbs were the following. Atelic: *run, fly, play, paint, sing, think, consider, imagine, dream, study, talk, discuss, fight, love, argue*; Telic: *enter, die, leave, hit, grab, decide, accept, forget, choose, remember, marry, sell, buy, give,* and *take*. To create each participant's list of paired verbs for each trial, the atelic and telic verbs for each domain were shuffled and then paired (within domain) to produce fifteen total pairs (five for each domain). This shuffling was performed four times (60 trials in total). Verb pairs (trials) were shuffled and paired randomly with videos.

Procedure During the experiment, participants were instructed that they would be shown a short clip of moving dots and asked to choose which of two verbs better fit the clip. They were told to use their intuition to make their verb choice and that there was no right answer. After the instructions, two practice trials were given.[2] On each trial, participants were shown the video twice before making their selection, to ensure they could adequately perceive the motion. The video slowly faded in over the first half second and then faded out during the last half second to diminish influences of motion onset. After the video disappeared, the two verbs appeared on the screen, one on each side. Participants then made their selection (f for left or j for right). See Figure 1 for a schematic of trial types.

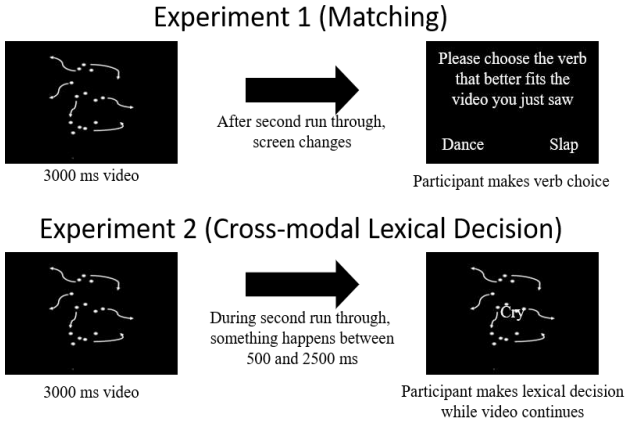


Figure 1: Trial schematics for Experiment 1 (explicit matching) and 2 (lexical decision). In Exp. 1, participants watched a 3 second video twice, then made a choice of which of two verbs better fit the video. In Exp. 2, during the second viewing, they were instead given a lexical decision (word/nonword) or attention task. Lines with arrows illustrate motion paths and were not seen by subjects.

Results

Figure 2 presents the mean proportion of telic verbs selected (subject means) as a function of whether the motion event was unbounded or bounded, overall and separately for each verb domain. As predicted, telic verbs were selected less often for Unbounded motion events (M=0.39, SE=0.02) than for Bounded motion events (M=0.55, SE=0.02). Moreover, the effect of motion Boundedness was very similar for each verb domain.

To test for reliability, we used a multilevel logistic regression to model binary trial-level choices for the telic verb. Fixed effects consisted of motion Boundedness (bounded vs. unbounded), Verb type (physical, social or mental), and the interaction. The maximal random effects structure was used for each subject and a random intercept was used for each item (each video). The significance of factors was performed by comparing likelihood-ratio values for nested models that included main effects and interactions of factors to models without them.

The best-fitting model showed a reliable effect of motion Boundedness (β=0.339, SE=0.096, z=3.531, p=0.0004), but no reliable effects or interactions with Verb type. Removing Verb type and the interaction from the model did not decrease the fit of the model ($\chi^2(4)$=4.38, p=0.346), but further removing the effect of Boundedness did ($\chi^2(1)$= 27.92, p<0.0001) – indicating that the motion Boundedness of the videos reliably predicted telic choices.

[2] Thus, there were 2 practice videos and 58 trial videos. Later "test" trials using these videos were discarded from analysis.

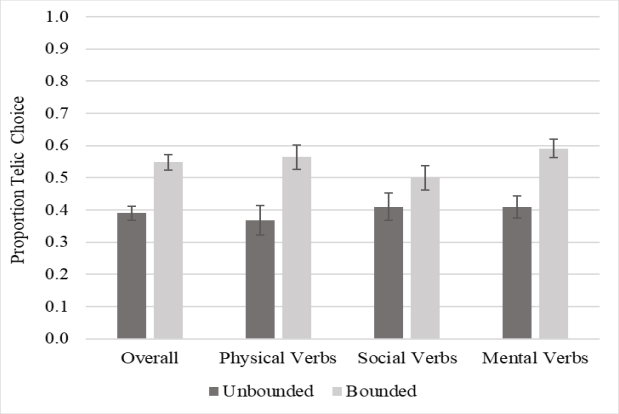


Figure 2: Mean proportion of telic verb choices as a function of Motion Boundedness (Unbounded, Bounded), overall, and by verb domain. Average of Subject Means. Error bars indicate ±1 Standard Error.

One concern is that people may have developed a strategy over time or discovered the purpose of the experiment and acted accordingly. Although unlikely, as no participant revealed an explicit awareness of the hypothesis or purpose when questioned, we examined the time course of the effect. If it was an explicit strategy, we might expect a difference between the beginning and the end of the experiment. We tested whether there was a statistical difference between the first twenty trials and last twenty trials regarding whether participants showed a differing proportion of telic verb preference for bounded and unbounded videos. This test showed no significance (paired t (23) = -0.635, p = 0.532), suggesting a strategy did not emerge during the experiment.

Discussion

The current results show that the visual system can distinguish between "bounded" and "unbounded" motion stimuli, as defined by repetitiveness. Furthermore, people are biased to associate visual boundedness with conceptual boundedness in verbs. Surprisingly, this occurs even when these verbs have no visual manifestation, as is the case for the domain of mental verbs. This implies that people were implicitly encoding boundedness and telicity when observing these biomotion stimuli. Put another way, these results show that the distinction of boundedness that is present in both the visual and linguistic systems is shared or otherwise accessible by the two systems. This extends Strickland et al. (2015) by demonstrating that this association is not just due to the referential task (associating two linguistic items), but rather it exists even when using visual stimuli that have no inherent referential properties (i.e. scrambled biomotion stimuli). Thus, it appears that conceptual boundedness is a basic property of the visual and linguistic systems.

Although we find these results to be compelling, it would be of interest to understand if similar results can be obtained when participants are not attempting to link the video of the motion event to the linguistic stimuli. That is, could such a connection between motion boundedness and linguistic telicity arise spontaneously, even when linguistic judgements do not involve connecting the verb to the motion video? If so, it would suggest that the perception of motion boundedness automatically activates linguistic telicity. We explore this issue in Experiment 2.

Experiment 2: Cross-Modal Lexical Decision

In this experiment, we present preliminary results from two versions of a cross-modal lexical decision task. Each trial involved the participant viewing the same clips used in Exp. 1. However, participants were not presented with a forced choice task. Instead, for target trials, a single telic or atelic verb appeared centrally over the video, and the participant's task was to make a lexical decision (word / nonword). The core prediction of this experiment is the following: *If the perception of motion boundedness spontaneously activates linguistic telicity, then we would expect a congruency effect, such that Bounded motion events would speed judgments of telic verbs whereas Unbounded motion events would speed judgments of atelic verbs.*

The results are preliminary because each version of the experiment suffered from some issues related to response time measurement precision, such that we may have been underpowered to detect robust results in either alone. Nevertheless, across versions, the patterns are significant and consistent with our hypothesis. Thus, we present both sets of results together, noting differences between versions below as needed. Further versions of the experiment with more precise RT measurements are planned.

Differences between Versions of the Experiment Each version of the experiment suffered from precision issues with RT collection. In version 1 of the experiment, participants used a keyboard to make their responses (f for word, j for nonword, or spacebar if the dots changed color). However, due to a coding error, responses were only recorded at each screen refresh (every 16.67 ms); thus, these measurements were imprecise. In version 2, we used an E-prime button box instead of keyboard, since it is known for its measurement precision. Mean accuracy and mean RTs were nearly on par across experiments. However, the buttons on the button box used in version 2 were differentially sensitive; 4-5% of trials were timeouts, while those in version 1 were nearly 0%, suggesting that sometimes the buttons were not responsive.

Other differences were the following. In version 1, unbounded videos were randomly assigned one of the onset times for the bounded video. In version 2, this assignment was fixed for each list. That is, in version 2, unbounded video A would always have onset time of bounded video 1, video B would have onset time of bounded video 2, etc.

Method

Participants Experiment 2 consisted of two separate versions. A total of 116 subjects were recruited from a

university's undergraduate student body. Participants were excluded based on pre-determined criteria: accuracy below 80% on any of the three tasks (described below). After exclusion, there were 54 in the first version and 47 in the second. For both versions, the goal sample size was determined by doubling the sample size of experiment 1. The data from participants 49-54 from version 1 were included due to this experiment being underpowered (described below). Students signed up to participate in exchange for study credit. All participants were fluent English speakers with normal or corrected to normal vision.

Materials The same materials from Experiment 1 were used (i.e., the same fifteen telic and atelic verbs across three domains and the same sixty test videos).

An additional sixty videos from the CMU Graphics Lab database were used for filler attention trials (described below). An additional twelve videos and four verbs were used for practice trials. Sixty-four non-words were created using Wuggy, a word generation tool that creates nonwords matched with inputted real words on phonotactics and word length (Keuleers & Brysbaert, 2010). These non-words were used for the lexical decision task (described below).

Although our videos were rated for overall boundedness in the previous norming study, this does not indicate *when* in the video a boundary occurred. To ensure that the onset time of the word stimulus coincide with a motion boundary (for bounded videos), study authors JW and AH and research assistants chose the boundary point by watching each video frame by frame. In version 1, JW and AH individually watched and chose the points. If the differences between the two values was more than thirty frames, JW re-watched the video and made the final decision. If the difference was less than thirty frames, the average of the two was taken. In version 2, to get a more reliable value for boundedness point, median frame values for each video across an additional four research assistants, in addition to JW and AH, were taken. The average difference between the first and second experiment version was three frames. Since each video was only 3 seconds long, boundaries closer than 0.5 seconds towards the beginning or end of the video were constrained to be at 0.5 or 2.5 seconds, respectively. Since they do not have a motion boundary, the distribution of onset times for unbounded videos and filler (attention) videos were matched to the bounded videos.

Procedure During the experiment, there were three trial types: two were lexical decision (word or non-word). In these trials, participants simply had to press one button if the string of letters that appeared was a word, and another if it was not. The third trial type was an attention catch task, designed to ensure participants attended to the visual dot stimulus. In this catch task, dots would briefly change color from white to blue (0.5 sec); thus, this task made no reference to visual motion.

Participants were instructed to press a different key for each of these three trial types. On each trial, one of the three visual changes would appear: a word superimposed on the dot stimuli, a non-word on the dot stimuli, or the color change type. Participants were not made aware of what the current or next trial would be ahead of time. After the instructions were given, twelve practice trials (four of each type) were given. On each trial, participants were shown the 3-second video twice to ensure they could adequately perceive the full motion. The visual stimulus appeared at the pre-determined onset time during the second viewing. The videos faded in and out over the first and last half second to avoid sudden visual transients.

The visual stimuli (word, non-word, color change) appeared either *at* the pre-selected onset frame, or 0.25 sec *after* (counterbalanced with each condition). However, this timing factor was collapsed over in analyses.

Each verb and nonword was paired once with a bounded video and once with an unbounded video. There were 192 total trials (60 word, 60 non-word, 60 catch, 12 practice).

Results

The results for the first and second version of the experiment are being presented together. For an explanation and discussion of this decision, see below. Reaction times +/- 2.5 *SDs* from each subject mean were excluded (2.9% of trials), as well as timeout trials. Word trials: Accuracy 94.8% (*SD* 3.5%), mean RTs 676 ms (*SD* 119ms); Non-word trials: Accuracy 92.3% (*SD* 5.1%), mean RTs 808ms (*SD* 155ms); Attention trials: Accuracy 94.1% (*SD* 4.0%), mean RTs 629ms (*SD* 139 ms). All statistical analyses were performed on inverse RTs (-1000/RT), on accurate word trials only. Inverse RTs were used to improve normality of RT distributions for model fitting (Baayen & Milin, 2010).

Figure 3 presents the inverse reaction time (subject means) as a function both of verb telicity (telic vs. atelic), and motion boundedness (whether the visual stimulus was unbounded or bounded). As predicted, we observe an interaction between verb telicity and visual boundedness on reaction times for lexical decision: reaction times to atelic verbs were faster when the visual motion was unbounded, and reaction times to telic verbs were faster when the visual motion was bounded. However, the effect here was subtler than in Exp. 1, as should be expected: participants were not performing an explicit matching task but deciding whether the word that appeared was a real word or a non-word (or were performing a color change detection task, for filler trials).

To test for reliability, we used a multilevel linear regression to model inverse reaction time. Fixed effects consisted of Verb Telicity (telic vs. atelic), Motion Boundedness (bounded vs. unbounded), Verb Type (physical, social or mental), and all relevant interactions. To account for mean RT differences in experiment versions, a main effect of experiment Version was also included. The maximal random effects structure that converged was used for each subject (random intercept and random slopes for telicity and motion boundedness), and a random intercept was used for each verb. We compared nested models with and without these factors and interactions.

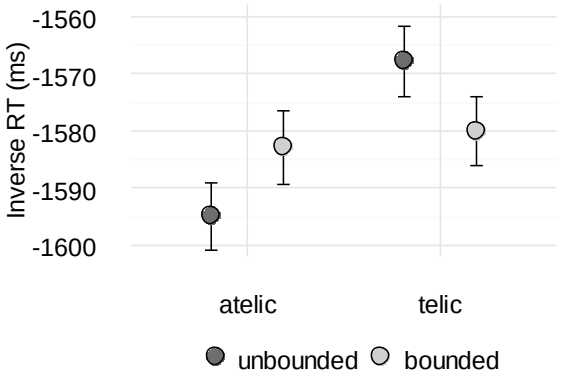


Figure 3: Mean inverse response times (-1000/RT) on lexical decision trials (word only), as a function of verb telicity (telic, atelic) and motion boundedness (unbounded, bounded). Average of Subject Means. Error bars indicate ±1 Standard Error. Up indicates slower RTs, down faster RTs, as is standard for raw RT plots.

The best fitting model was one that included main effects and interactions of Verb Telicity and Motion Boundedness, and Verb Type and Motion Boundedness. This model produced a reliable interaction between Verb Telicity and Motion Boundedness (β=6.56, *SE*= 2.91, t(5555)=2.26, p=0.024). Telicity and Motion Boundedness were contrast coded in the following way: telic=1, atelic=-1; unbounded=1, bounded=-1. Thus, the positive β indicates greater RTs for the mismatched conditions (e.g. telic+unbounded or atelic+bounded). Adding a triple interaction of Telicity, Boundedness, and Verb Type only marginally improved the fit (χ^2(4)=8.34, p=0.08). Further, removing the interaction of Telicity and Boundedness *did* decrease the fit of the model (χ^2(1)=5.09, p=0.024). This confirms the interaction of Verb Telicity and Motion Boundedness that we expected.

Discussion

Experiment 2 results produced a pattern of congruency that would be expected if the perception of motion boundedness activated linguistic telicity. Reaction times to atelic verbs were faster when preceded by Unbounded as opposed to Bounded motion events whereas the opposite was found for telic verbs. That this effect arose even though participants were not trying to relate the verbs to the motion videos suggests that the relation between boundedness and telicity occurs spontaneously, without conscious effort.

There are a few important issues that force us to see these results as preliminary and necessary of replication. Even though the timing issues mentioned in the Method section merely decrease RT precision, it is also the case that the results of both versions were pooled post-hoc, after discovering these RT precision issues. Although these are not small issues that should be overlooked, we believed it was still worthwhile to report on these data as they are promising, consistent with the judgment data of Experiment 1, and spur the need for replication.

Although we did not find strong evidence for differences among verb domains in the congruency effect, future replications with higher power will allow us to determine whether this difference is real, and if so, whether only certain of the domains (e.g. physical) demonstrate the effect. Nevertheless, results of Experiment 1 (matching) and Experiment 2 (lexical decision) do suggest that the congruency effect is general, regardless of domain.

General Discussion

We have shown here that the motion properties of what can only be characterized as scrambled moving points of light yield systematic and expected interpretive responses from observers, concerning the detection of motion boundedness, and its relation to conceptual and linguistic telicity.

In Experiment 1, we observed that participants were more likely to choose a telic verb over an atelic verb to describe a bounded non-repetitive motion, even when the meanings of these verb pairs denoted abstract mental events (e.g., *think* vs. *decide*). This extends the Strickland et al. (2015) findings to visual stimuli more generally, even stimuli without linguistic and referential properties (signs). Experiment 2 offered preliminary results that activation of conceptual telicity from motion signals arises automatically, such that telic verbs show a congruency effect with bounded motion and atelic verbs show an effect with unbounded motion. Activation of concepts from motion signals has been observed before in similar tasks, e.g., that upward motion will speed reaction time to *rise* as opposed to *fall* (Meteyard et al., 2008). Additionally, previous research has shown that the ability to judge an action verb on a lexical decision task is correlated with the ability to judge a non-scrambled point-light action on an action decision task, e.g. is this a valid human action (Bidet-Ildei & Toussaint, 2015). What is surprising in the present study is that activation is for a highly abstract categorical feature (boundedness) that arises from seemingly continuous motion signals, and that the activation affects judgments even for verbs labeling events in the social and mental domains, which have no overt visual boundedness cues.

Observing an implicit association between visual and linguistic boundedness suggests there is an underlying amodal conceptual distinction that both systems have access to: a distinct categorical representation of boundedness, which may indeed be a conceptual primitive similar to that proposed for causation (see, e.g., Jackendoff, 1996; Rolfs, Dambacher, & Cavanagh, 2013). Of course, a representation of boundedness is not limited to the event domain; things may be conceived of as objects or substances, which have perceptual consequences of their own (vanMarle & Scholl, 2003). In future work we plan to refine our experimental tasks to replicate our lexical decision effects in the event domain, and to determine if they extend to boundedness across conceptual domains (events and objects).

References

Baayen, R. H., & Milin, P. (2010). Analyzing Reaction Times. *Int. Journal of Psychology Research, 3,* 12–28.

Bidet-Ildei, Christel & Toussaint, Lucette. (2014). Are judgments for action verbs and point-light human actions equivalent?. Cognitive processing. 10.1007/s10339-014-0634-0.

Jackendoff, R. (1996). Conceptual semantics and cognitive linguistics. *Cognitive Linguistics, 7(1),* 93-129.

Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior Research Methods 42(3),* 627-633.

Kleiner M, Brainard D, Pelli D, 2007, "What's new in Psychtoolbox-3?" *Perception 36 ECVP*.

Malaia, E., & Wilbur, R. B. (2012a). Kinematic signatures of telic and atelic events in ASL predicates. *Language and Speech*, *55*(3), 407–421.

Meteyard, L., Zokaei, N., Bahrami, B., & Vigliocco, G. (2008). Visual motion interferes with lexical decision on motion words. *Current Biology, 18*(17).

Rolfs, M., Dambacher, M., & Cavanagh, P. (2013). Visual adaptation of the perception of causality. *Current Biology*, *23*(3), 250-254.

Strickland, B., Geraci, C., Chemla, E., Schlenker, P., Kelepir, M., & Pfau, R. (2015). Event representations constrain the structure of language: Sign language as a window into universally accessible linguistic biases. *Proceedings of the National Academy of Sciences*, *112*(19), 5968–5973.

Todorova, M., Straub, K., Badecker, W., & Frank, R. (2000). Aspectual coercion and the online computation of sentential aspect. In *Proceedings of the Cognitive Science Society* (Vol. 22).

van Boxtel, J. J. A., & Lu, H. (2013). A biological motion toolbox for reading, displaying, and manipulating motion capture data in research settings. *J. of Vision, 13(12),* 7–7.

vanMarle, K., & Scholl, B. J. (2003). Attentive Tracking of Objects Versus Substances. *Psychological Science, 14(5),* 498–504.

Vendler, Z. (1957). Verbs and Times. *The Philosophical Review*, *66*(2), 143–160.

Individual differences in bodily attention: Variability in anticipatory mu rhythm power is associated with executive function abilities and processing speed

Staci Meredith Weiss (sweiss@temple.edu)
Department of Psychology, 1701 N. 13th Street
Philadelphia, PA 19122 USA

Rebecca L. Laconi (devlab@temple.edu)
Department of Psychology, 1701 N. 13th Street
Philadelphia, PA 19122 USA

Peter J. Marshall (pjmarsh@temple.edu)
Department of Psychology, 1701 N. 13th Street
Philadelphia, PA 19122 USA

Abstract

The ability to anticipate, attend and respond appropriately to specific stimuli is involved in the execution of everyday tasks. The current investigation examined the relations between cognitive skills measured by the NIH Toolbox and changes in the power of mu oscillations during anticipation of and in response to a tactile stimulus. Electroencephalographic (EEG) activity was measured after a visuospatial cue directed adults (n=40) to monitor their right or left hand for upcoming tactile stimulation. In the 500 ms prior to the onset of the tactile stimulus, a desynchronization was apparent 8 – 14 Hz at contralateral central sites, consistent with prior investigations of mu rhythm; a widespread synchronization was apparent in the 250 ms proceeding delivery of the tactile stimulus. The extent of contralateral reduction in mu power was associated with speed processing ability, while ipsilateral mu power was associated with flanker performance and marginally correlated with card sort performance. Regression further probe the significance and specificity of these effects. Increases in mu power following onset of the tactile stimulus were not associated with any behavioral measures. Mu modulation during attention to a specific bodily location appears related to variability in the broader ability to regulate behavior in a goal-directed manner, and perhaps to speed of stimulus processing.

Keywords: tactile; mu; EEG; executive function; sensorimotor; oscillations; anticipation;

Introduction

Anticipation of an impending event or sensation can guide perception and action. In experimental settings, when the presentation of a visual, auditory or tactile stimulus is preceded by a stimulus-relevant cue, participants report higher rates of accurate stimulus perception and demonstrate more rapid reaction time than when a stimulus is presented without a preparatory cue (Posner, 1980; Frey et al., 2015). These behaviors suggest prior to stimulus presentation, deployment of attention in a selective, focused manner is conducive to stimulus processing (van Ede & Nobre, 2017). Exploiting the temporal precision of electroencephalogram (EEG), we can eavesdrop on the changes in neural oscillations which occur before and after the presentation of a stimulus (Cheyne et al., 2003; Engel, Fries, & Singer, 2001), with the goal of identifying how these changes facilitate perception and the regulation of behavior.

In this study we assessed individual differences in oscillatory neural responses during anticipation of a tactile stimulus and in response to that stimulus. We investigate the association of subject-specific changes in oscillatory activity with variation in 1) reaction time in responding to the tactile stimulus, 2) general processing speed and receptive language abilities, as well as 3) executive function abilities, or the constellation of skills involved in the regulation of behavior.

The Active Role of Alpha Oscillations in Perception

Oscillatory activity in the alpha band of the EEG signal, broadly defined as activity within the 8-14 Hz frequency range in adults, has been identified as a correlate, gate and predictor of behavioral responses and cognitive functioning (Zanto & Gazzaley, 2009; van Ede & Nobre, 2017). As the most prominent oscillation in the EEG, alpha-range signals were originally associated with an 'idling' state but are now seen as more active in perceptual and cognitive processes (Klimesch et al., 1998). The oscillations apparent in the EEG signal arise from fluctuations in the polarity of cortical tissue, which reflect the shifting, homeostatic balance of postsynaptic potentials released by assemblies of excitatory pyramidal cells and inhibitory interneurons (Lopes da Silva, 2013; Cohen, 2016). The presence (or mere expectation) of a stimulus disrupts the default synchronized firing rate of post-synaptic potential which generated the rhythmic alpha activity, eliciting an event-related desynchronization (ERD) in the oscillatory signal (Haegens et al. 2011; Lopes da Silva, 2013). Changes in amplitude, phase and frequency of oscillations evoked by a discrete event can be computed using event-related spectral perturbation (ERSP), in which sinusoidal wavelets are used to estimate the shift in amplitude and phase of EEG oscillations in each successive, overlapping time window (Pfurtscheller & Da Silva, 1999; Makeig & Delorme, 2004). Thus, ERSP can quantify the changes in power of a given frequency range (relative to a baseline period), tracking the temporal sequence of postsynaptic potentials discharged synchronously from a particular neuronal population (Klimesch et al., 1998; Lopes da Silva, 2013).

To study changes in alpha power in anticipation of or in response to an upcoming event or stimulus, participants are presented with a cue that orients them to a feature of the forthcoming stimulus. In the widely-used Posner paradigm, a spatial cue indicates whether a visual stimulus will be presented to the participant's right or left visual field (Posner, 1980). During the interval following the cue but prior to the predicted onset of a visual stimulus, anticipatory ERD of

rhythmic alpha activity is observed over contralateral visual cortex, measured as a decrease in alpha power relative to baseline (Thut et al., 2006; Nobre & van Ede, 2017).

Contemporary accounts of 'top-down' or attention-related modulation of alpha-range activity rest upon the inhibition-timing hypothesis (Klimesch, Sauseng, & Hanslmayr, 2007), which explains that during rest, oscillatory EEG activity arises from the synchronized cortical firing of neurons that may limit the sampling of sensory events (Schroeder & Lakatos, 2009). When a stimulus disrupts the default state of rest or inattention, there is a reallocation of resources diverted to the local processing of the new or expected stimulus, which is facilitated by the suppression or inhibition of global neural activity. As such, widespread *increases* in alpha power from baseline reflect inhibited sampling of irrelevant sensory events, which permit concentrated cortical firing by neurons in the sensory cortex relevant to the stimulus. Focused attention and perceptual awareness of a stimulus is thus facilitated by concomitant *global increases* and *local decreases* in alpha power, indicating an adjustment in the sampling of sensory events adaptive to the expected temporal and spatial presentation of an upcoming stimulus (Frey et al., 2015; Schroeder & Lakatos, 2009; Thut et al., 2006). During anticipation, it is thought that sensory-specific alpha responses initiate the coordination of multisensory attentional control networks, enabling dynamic prediction of events across modalities and preparation for action (Engel, Fries & Singer, 2001; Sadaghiani & Kleinschmidt, 2016). In reaction to a stimulus, during rest and under most other conditions, these modality-specific alpha rhythms exhibit dissociable properties and operate independently (Mazaheri et al. 2009). Thus, the state of stimulus anticipation enables a unique opportunity for studying variability in oscillatory neural activity and centrality to behavior (Weiss et al., 2018).

Although much of the extant work on alpha power fluctuations has focused on the visual alpha rhythm at posterior occipital sites, another prominent alpha-range oscillation is the sensorimotor mu rhythm observed at central electrode sites (Jones et al., 2010; Pfurtscheller, 1989). Expectation of tactile stimulation in adults elicits changes in the mu rhythm which exhibit a somatotopic pattern (Anderson & Ding, 2011; Jones et al., 2010), in accord with the organization of the homunculus (Penfield & Boldrey, 1937). Jones et al. (2010) demonstrated reductions of mu power in anticipation of tactile stimulus, with responses lateralized according to the direction of a spatial cue (pointing left or right) as participants monitor their hands in expectation of sensation. Particularly when a distracting tactile sensation is presented simultaneous to the uncued hand, ipsilateral increases in mu power have also been demonstrated during the suppression of tactile attention (Haegens, Luther, & Jensen, 2012; van Ede, de Lange, & Maris, 2014). The utility of mu oscillatory power as an index of individual difference in behavior, beyond tactile stimulus processing to more general control of voluntary attention and action (executive functioning), has yet to be fully explored.

Anticipatory Mu Power and Tactile Processing

Across auditory, visual and tactile modalities, both contralateral alpha ERD and increases in ipsilateral alpha power during stimulus anticipation and response have been correlated with behavioral responses to stimuli (Thut et al., 2006; van Ede et al., 2014; Frey et al., 2015). In the tactile modality, the relation between mu power and behavioral indicators of tactile processing appears to differ depending on the strength and salience of the expected tactile stimulation, as well as the load on tactile attention (Haegens et al., 2012; Gomez-Ramirez, Hysaj, & Niebur, 2016). When a reliable spatial cue directs participants to expect tactile stimulation at the cued location, the magnitude of anticipatory mu ERD in electrode sites over the contralateral somatosensory cortices is linearly, inversely associated with rate of stimulus detection (Anderson & Ding, 2011; Haegens et al., 2011; Jones et al., 2010). Van Ede et al. (2012) examined anticipatory and post-stimulus mu power to parse their relative contributions to behavioral indicators of tactile processing. The authors reported that anticipatory mu ERD significantly accounted for the accuracy of participant's tactile judgements, while both the magnitude of anticipatory mu ERD and post-stimulus mu increases in mu power accounted for participant's reaction time to the stimulus. Reductions in anticipatory contralateral mu power have also been linearly associated with higher hit rates on tactile feature detection and temporal judgement tasks (Gomez-Ramirez et al., 2016).

Haegens, Handel and Jensen (2011) employed magnetoencephalography to investigate whether the lateralization of anticipatory mu oscillations varied according to how accurately a visual arrow cue relayed the location (right or left thumb) of an upcoming tactile stimulus. The authors reported that anticipatory contralateral mu power significantly distinguished between trials with above- and below-average reaction times, but not in accurate identification of the tactile stimulus (Haegens et al., 2011). This relation depended on the validity of the visual cue in predicting the location of the tactile stimulus. The authors found that the extent of oscillatory mu modulation reflects the predictability of the environment, such that differences in ipsilateral and contralateral mu power decreased under conditions with increasing uncertainty.

When tactile stimulation is expected simultaneously to a target location and another body part, it appears that variance in ipsilateral mu may index the suppression of tactile attention, partially accounting for behavioral responses to a tactile stimulus. In a subsequent MEG study, Haegens, Luther and Jensen (2012) reported that when tactile stimulation is presented simultaneously to the cued and uncued hand, both ipsilateral and contralateral mu power significantly distinguish between correct and incorrect trials. Thus, similar to the importance of increases in ipsilateral anticipatory alpha power in the visual modality in accounting for variability in stimulus response (Thut et al., 2006; Frey et al., 2015), anticipatory ipsilateral mu power may facilitate focus when tactile attention is under load.

To address inconsistencies in the literature associating oscillatory mu activity with task-specific indicators of tactile processing, we note the potential importance of subtle differences in task demands (Gomez-Ramirez et al., 2016). The dynamic adjustment of lateralized mu modulation to anticipated features of a tactile stimulus may be indicative of its sensitivity to the load on tactile attention, divided by managing competing expectancies, allocating tactile attention according to goals and bracing for potential distraction (Haegens et al., 2012).

One suggestion arising from work linking anticipatory neural responses to basic sensory responses is the proposition that 'low-level' indicators of attentional processing reciprocally influence, gate and cascade into individual level differences in the 'higher-order' ability to control behavioral responses (Engel, Fries & Singer, 2001; Gazzaley and Nobre, 2012; Sadaghiani & Kleinschmidt, 2016). We further suggest that executive function, defined by the planning, regulating and monitoring of goal-directed behavior, may partially be a manifestation of individual differences in how adults use information in their environment to anticipate upcoming sensory events and adjust their behavior to such expectancies.

The Present Study

The goal of the current investigation is to utilize an individual differences approach to the analysis of sensorimotor mu oscillatory activity during anticipation of and in response to a tactile stimulus. Our objectives were (i) to develop a subject-specific approach to identifying sensorimotor mu rhythm reactivity (ii) to examine whether mu reactivity is associated with variance in participant's reaction time in stimulus detection (iii) to test if mu reactivity is associated with variance in a battery of cognitive skills, which includes measures of receptive language, processing speed and executive function. We employed a task in which a visual cue directed adults to focus their attention on a specific bodily location (the left or right hand) in anticipation of a tactile stimulus to that location. Using a foot pedal, participants responded to the tactile stimulus to indicate whether they detected one or two stimuli. We expected neural indicators of heightened attention (greater mu desynchronization or ERD in the contralateral hemisphere, and greater mu synchronization or ERS in the ipsilateral hemisphere) to relate to higher-order cognitive abilities (i.e., the executive function measures) and response time to target stimuli.

The logic of presenting a preparatory cue in a *different* modality from the target stimulus allows temporal and spatial differentiation of anticipatory activity (over sensory cortex relevant to the target) from neural responses elicited by the cue. There are also several strengths of employing somatosensory rather than visual targets: (i) Compared with the visual modality, tactile attention is not complicated by factors such as ocular shifts or visual preference; (ii) Neural indices of anticipation of touch are readily measurable through EEG recordings from electrodes overlying somatosensory cortex (Anderson and Ding, 2011; Haegens et al., 2011; Jones et al., 2010); (iii) The ability to focus attention to a body part in expectation of touch may be amenable to change and enhancement via specific interventions (Jones et al., 2010).

Methods

Fifty undergraduate students received course credit in return for participation. Data from six participants were excluded from analyses due to technical issues. Four additional participants were excluded due to excessive artifact that contaminated more than 25% of trials. The final analyzed sample comprised 40 participants (mean age = 21.24 years; SD = 3.85; 37 females). All participants were right-handed according to the Oldfield Handedness questionnaire, neurologically healthy, and had normal or corrected vision. Once consented, participants were fitted with an EEG cap and tactile stimulators, seated at a table facing a computer screen, and instructed to rest their hands on their lap, out of sight.

Procedure

Participants were instructed to prepare for tactile stimulation to the index finger of the hand indicated by the direction of the arrow, and to indicate how many stimuli they detected (one or two) by pressing a foot pedal once or twice. The foot used to report stimulus detection was counterbalanced across participants. The specific sequence of visual stimuli in each trial comprised a fixation cross for 500 ms, followed by the arrow cue for 2250 ms, followed by a response screen that read "Copy with Your Foot!" (Figure 1). The tactile stimulation was delivered 1500 ms after the onset of the arrow cue, which remained on the screen for the 750 ms following tactile stimulation. The direction of the arrow was randomized, with an equal number (100) of left and right trials. Individual participant's reaction time was retrieved from the onset of the response screen to the foot pedal press. Two tactile stimuli were delivered in rapid succession ("double stimuli") on 20 out of the 200 trials, and 80 single-pulse trials were delivered to the right or left hands of participants. Prior to the experimental trials, 5 practice trials were presented to ensure that participants distinguished between the single and double tactile stimuli.

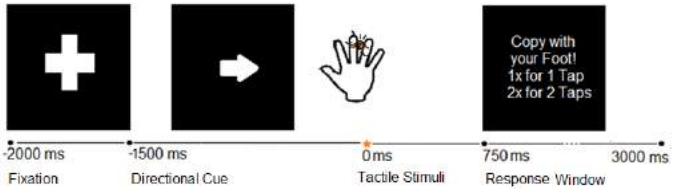


Figure 1. Trial structure: A fixation point was displayed for 500 ms, followed by an arrow spatial cue displayed continuously for 2250 ms, and the onset of the tactile stimulus occurred 1500 ms later (at 0 ms). The response prompt was displayed 750 ms after the tactile stimulus.

Tactile stimuli were delivered to the distal tip of the left and right index fingers using an inflatable membrane (10 mm diameter; MEG Services International, Coquitlam) mounted

in a plastic casing and secured with a finger clip. The membrane was inflated by a short burst of compressed air delivered via flexible polyurethane tubing (3 m length, 3.2 mm outer diameter). The compressed air delivery was controlled by STIM stimulus presentation software in combination with a pneumatic stimulator unit (both from James Long Company, Caroga Lake) and an adjustable regulator that restricted the airflow to 60 psi. To generate each tactile stimulus, the STIM software delivered a 10 ms trigger that served to open and close a solenoid in the pneumatic stimulator. Expansion of the membrane started 15 ms after trigger onset and peaked 35 ms later, with a total duration of membrane movement of around 100 ms.

EEG Recording and Processing EEG was recorded at a 512 Hz sampling rate using a stretch cap (ANT Neuro, Berlin) with electrodes placed at Fp1, Fpz, Fp2, F3, Fz, F4, F7, FC6, FC1, FC2, FC5, F8, Fz, C3, Cz, C4, CP1, CP2, CP5, CP6, T7, T8, P3, Pz, POz, P4, P7, P8, O1, Oz, O2, GND, and the left and right mastoids. Vertical EOG was recorded above and below the orbital rim of the left eye. Conducting gel was used and scalp electrode impedances were kept under 25 kΩ (values were typically lower). EEG channels were collected referenced to the vertex (Cz) and were re-referenced offline to an average mastoids reference prior to further analysis. The signal was amplified using optically isolated, high input impedance (> 1 GΩ) custom bioamplifiers (SA Instrumentation) and digitized using a 16-bit A/D converter (+/- 2.5 V input range). Bioamplifier gain was 4000 and filter (12 dB/octave rolloff) was set to .1 Hz (high-pass) and 100 Hz (low-pass).

Initial processing of the data utilized the EEG Analysis System (James Long Company) followed by analysis using the EEGLAB toolbox (Makeig et al., 2004) implemented in MATLAB. Independent component analysis was used to clear the EEG data of ocular and muscle artifact (Hoffmann and Falkenstein, 2008). Visual inspection of the EEG signal rejected epochs containing excessive remaining artifact. There was no difference in the number of usable trials between the left and right cued conditions ($p = 0.81$). Out of 80 trials, the mean number of artifact-free trials per condition was 69 ($SD = 5.62$).

For each single-pulse trial with a correct behavioral response, an epoch of 2500 ms was extracted (beginning 2000 ms prior to onset of the tactile stimulus and extending 500 ms after tactile stimulus onset). To avoid contamination of the anticipatory and response window by stimulus delivery, we set the initial membrane expansion as the onset of the tactile stimulus (0 ms) and the post-stimulus window to 20ms following the peak of membrane expansion. Spectral power over this epoch was estimated using Gaussian-tapered Morlet wavelets (Makeig & Delorme, 2004). Changes in power were computed as event-related spectral perturbation (ERSP) from initial visual cue presentation until after tactile stimulus presentation (i.e., -1500 to 300 ms) relative to a 500 ms baseline preceding the visual cue (i.e., -2000 to -1500 ms prior to tactile stimulation onset). For statistical analyses, a key variable was anticipatory mu ERSP, which was extracted from mean ERSP value at C3 or C4 from 8 – 14 Hz in the 500 ms prior to onset of the tactile stimulus to the onset of the tactile stimulus (0 ms). We extracted post-stimulus mean mu ERSP by extracting the mean mu ERSP for the period from the delivery of the tactile stimulation at 20 ms to the following 270 ms.

Behavioral Measures Following the tactile task and removal of the EEG cap, four tasks from the NIH Cognition Toolbox were administered (for details, see Zelazo et al., 2013): the Flanker task, the Card Sort task, a Processing Speed task, and a picture vocabulary test that measured Receptive Language. On the Card Sort task, participants selected one of two test stimuli which matched either the shape or color of the target stimuli. In the Flanker task, participants indicated the direction of a central arrow that was presented between distracting 'flanker' arrows. Processing Speed was measured by the average reaction time to detecting if two images were identical. Participant's scores on the Card Sort and Flanker tasks were calculated to reflect both accuracy and reaction time for participants who correctly identified targets on 80% of trials; accuracy alone was considered when this threshold was not met. For all four measures, we used t-standardized test scores (standardized around $\mu=100$) provided by the NIH Cognitive Toolbox.

Results

Behavioral Responses to Tactile Stimuli

Aggregated across the sample ($N = 40$), participants correctly identified the single or double tactile stimuli on 96.7% of trials. Reaction time was calculated as the duration from response screen until the initiation of the foot pedal press. Only single-stimulus trials were included in analyses.

Identifying Mu ERSP

Time-frequency plots (Figure 2) show a clear mu rhythm (8-14 Hz) ERD at the central electrode site (C3 or C4) contralateral to cue direction. In contrast, there was minimal change in mu power at the central electrode ipsilateral to the cue direction. Significant differences between contralateral and ipsilateral central sites (Figure 2) are driven by mu ERD during anticipation of tactile stimulation (-500 ms to 0 ms) at the site contralateral to the cue direction. At the left central electrode site (C3), mu ERD was apparent as participants attended to their right hand. At the right central electrode site (C4), mu ERD was present during attention to the left hand.

Quantifying Anticipatory and Post-Stimulus Mu ERSP The envelope of the amplitude-modulated signal was computed via the Hilbert transform ("hilbert" function in Matlab), which discards phase information and reveals oscillatory power fluctuations over time. A subject-specific approach to identifying peak mu activity was used, with a peak quantified in R as the largest local maximum within the 7-14 Hz range (Goljahani et al., 2014). This value was extracted from individual participant power spectra for C3 and C4, for each condition (right/left).

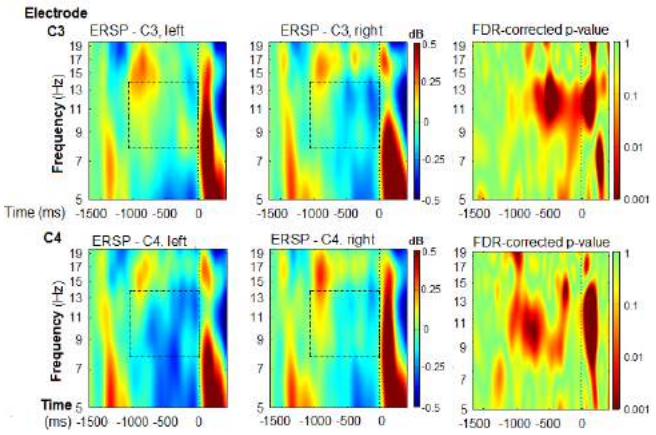


Figure 2. Time-frequency plots showing ERSP (event-related spectral perturbation) at left and right central sites (C3/C4) across a range of 5-20 Hz for the period from 1500 ms before the tactile stimulus (0 ms) to 300 ms after.

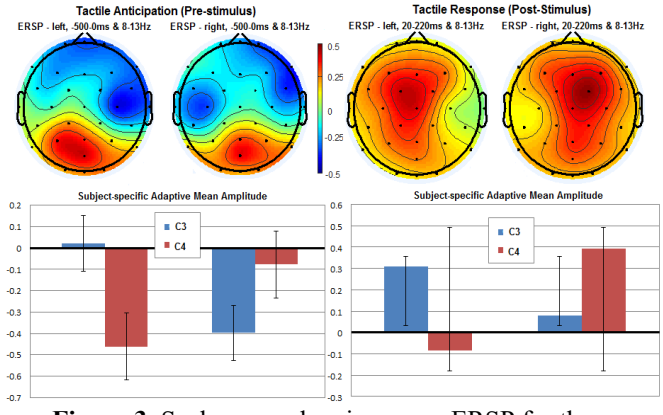


Figure 3. Scalp maps showing mean ERSP for the anticipatory period (-500 to 0 ms) and stimulus response period (20-270 ms) at each of 30 electrodes. Mu power for each participant was calculated for the subject-specific frequency band (+/-2 Hz) at C3 and C4.

The bin with the highest number of observations was centered at 10.5 Hz. The mean alpha peak frequency across subjects was 10.1 Hz with a between-subject SD of 2.1 Hz and the median was 10.4 Hz.

A repeated-measures ANOVA was conducted, comparing anticipatory mean 8-14 Hz ERSP in the -500 to 0 ms window prior to tactile stimulation, by electrode (C3/C4) and cue direction (left/right). No main effects were observed. A significant interaction was observed between cue direction and electrode, $F(1, 39) = 25.757$, $p < .001$, $\eta^{2p} = 0.398$. As suggested by the ERSP scalp maps (Figure 3), this interaction was driven by greater mu ERD at the contralateral site than at the ipsilateral site. When stimulation was expected to the left hand, greater mu ERD was observed at C4 ($M = -0.461$, $SD = 0.988$) than at C3 ($M = -0.022$, $SD = 0.984$, $t = 3.246$, $p < .001$, $d = .588$). When stimulation was expected to the right hand, greater mu ERD was observed at C3 ($M = -0.398$, $SD = 1.026$) than at C4 ($M = -0.077$, $SD = 0.844$, $t = -3.246$, $p < .002$, $d = -0.513$).

A repeated-measures ANOVA was conducted, comparing mean 8-14 Hz ERSP in the 20 to 270 ms window by electrode (C3/C4) and cue direction (left/right). No main effects were observed. A significant interaction was observed between cue direction and electrode, $F(1, 39) = 11.823$, $p < .001$, $\eta^{2p} = 0.233$. Following stimulation of the left hand, mu ERSP was greater at the ipsilateral site C3 ($M = 0.308$, $SD = 1.337$) compared to the contralateral site C4 ($M = -0.083$, $SD = 1.555$, $t = -3.506$, $p = .015$, $d = .403$). Following stimulation to the right hand, mu ERSP was greater at the contralateral site C3 ($M = 0.393$, $SD = 1.545$) compared to ipsilateral site C4 ($M = 0.079$, $SD = 1.686$, $t = -2.240$, $p = .031$, $d = -.354$).

To examine the relations between mu ERSP and scores on the behavioral tasks, the dependent variables used in the previous ANOVA were collapsed into *contralateral* (mu ERSP at C3 for the right hand cue and at C4 for the left hand cue) and *ipsilateral* (mu ERSP at C3 for the left hand cue and at C4 for the right hand cue) mean *mu ERSP* values.

Correlation of Mu ERSP with Behavior

Pearson correlations were computed among ipsilateral and contralateral mu ERSP in anticipation of (*anticipatory*) and in response to (*post-stimulus*) tactile stimulation, and the measures from the NIH Cognitive Toolbox. Contralateral anticipatory (CL TA) mu ERSP was inversely associated with Processing Speed (PS) ($r = -.321$, $p = .02$), while Flanker score was significantly associated with ipsilateral anticipatory (IP TA) mu ERSP ($r = .293$, $p = .03$). Similarly, Card Sort was marginally associated with ipsilateral anticipatory mu ERSP, ($r = .230$, $p = .06$). Processing Speed ability and task-specific reaction time were significantly correlated ($r = .245$, $p = .02$). Language (PVT) was not significantly correlated with other measures; contralateral tactile response (CL TR) mu ERSP was marginally associated with processing speed, ($r = -.219$, $p = .07$), but ipsilateral tactile response (IP TR) did not relate with other behavioral measures.

Table 1. Correlation Matrix of Study Variables.

	CL TA Mu ERSP	IP TA Mu ERSP	CL TR Mu ERSP	IP TR Mu ERSP	Flan-ker (EF)	Card Sort (EF)	PS	PVT
CL TA Mu ERSP	—	.017	**.718**	**.667**	—	—		
IP TA Mu ERSP	—	—	**.421**	**.511**	—	—		
Flanker	.006	**.293**	.021	.067	—	—		
Card Sort	-.001	*.230*	.018	.110	**.598**	—		
PS	**-.254**	-.043	*-.219*	.016	**.333**	**.421**		
PVT	-.100	-.047	-.132	-.137	-.203	.079	.047	
Reaction Time	.026	.117	.052	.008	.047	.115	**.245**	-.092

Regressions of Anticipatory Mu ERSP with Behavior

To address our hypotheses on the relations between cognitive skills and neural indicators of anticipation, multiple regressions were conducted predicting scores on the Flanker, Card Sort, Receptive Language, and Processing Speed tasks from contralateral and ipsilateral mu ERSP. For both Flanker and Card Sort tasks, greater ipsilateral mu ERSP was associated with better EF task performance. Flanker performance was related to ipsilateral mu ERSP, $t(39) = 2.026$, $\beta = 0.531$, $p = 0.046$, but not with contralateral mu ERSP. Card Sort performance was also related with ipsilateral mu ERSP, $t(39) = 2.219$, $\beta = 0.576$, $p = 0.033$, but was not significantly associated with contralateral mu ERSP. Contralateral mu ERSP was related to Processing Speed, $t(39) = -2.418$, $\beta = -0.621$, $p = 0.021$, and marginally associated with ipsilateral mu ERSP. Receptive Language scores were not related to anticipatory mu ERSP, nor were there further significant relations detected among regressions of behavioral measures and contralateral and ipsilateral mu tactile responses. Further, variance accounted for in Card Sort and Flanker by anticipatory ipsilateral mu ERSP remained significant when the extent of mu ERSP during the response to the tactile stimulus was used as a covariate. Similarly, variance in Processing Speed accounted for by anticipatory contralateral mu ERSP remained significant when controlling for variance in post-stimulus mu ERSP.

Discussion

We were interested in whether individuals differed in their neural activity during anticipation of and in response to a tactile stimulus, and whether such differences had meaningful relations with behavior, including indicators of tactile attention relevant to the task and measures of other attentional and cognitive skills. Consistent with previous investigations, sensorimotor mu ERD was observed in the hemisphere contralateral to the expected location of a tactile stimulus, indicating that participants indeed directed their attention to the relevant hand during the anticipatory epoch (Haegens et al., 2011; Anderson & Ding, 2011; Van Ede et al., 2014). The magnitude of contralateral mu ERD was associated with how quickly and accurately participants compared the similarity of two stimuli on a separate task assessing processing speed. In turn, performance on the processing speed task was found to be related to how quickly participants pressed a foot pedal to indicate how many tactile stimuli they perceived in the EEG task. Individual differences in executive function were also associated with variation in the magnitude of anticipatory mu oscillations, but only at central electrodes sites ipsilateral to the cued hand.

Mu activity in the ipsilateral somatosensory cortices is relevant to the coordination of behavioral responses, with animal and human research indicating that somatosensory processing is distributed across bilateral primary sensory cortices (van Ede et al., 2014; Tamè et al., 2016). The dynamic adjustment of bilateral mu modulation to anticipated features of a tactile stimulus indicates that oscillations originating in the somatosensory cortices are acutely sensitive to the load on tactile attention (Gomez-Ramirez et al., 2016; Haegens et al., 2012). In primates and humans, neural responses in bilateral somatosensory cortices may serve to simultaneously managing competing expectancies, reflecting allocation of tactile attention according to goals and bracing for potential distraction (Haegens et al., 2012; Tame et al., 2016)

In interpreting the relation of ipsilateral mu activity (rather than contralateral mu ERD) to executive function, we look to two possible explanations for the generation of alpha oscillations. Global alpha oscillations have been ascribed an inhibitory function (Klimesh et al., 1998; Mahzeri and Jensen, 2010). In past examinations of anticipation in the visual and auditory modalities, the 'gating' function of increases in alpha power has offered an account for the association between anticipatory ipsilateral alpha power with task-relevant stimulus detection rate and speed of behavioral responses across sensory modalities (Frey et al., 2015). Alternatively, and supported by previous investigations of anticipation in visual and tactile modalities (van Ede et al., 2014; Thut et al., 2006), the ipsilateral mu power over the relevant sensory cortices might increase or hover at baseline to suppress sampling of events at the unattended location (Shroeder and Latkos, 2009). These complementary accounts of anticipatory alpha oscillations may provide insight into how variability of neural responses contributes to individual differences in measures of cognitive ability.

The association of processing speed ability and reductions of mu power expands the existing literature focused on relations of mu modulation with task-specific reaction time. A previous investigation of children aged 6-8 found a significant association between executive function abilities and contralateral reductions of mu power (Weiss et al., 2018). There may be developmental differences in how attention is allocated in expectation of a tactile stimulus: speculatively, younger children may deliberately focus on monitoring sensation at the cued location while adults deploy effort into inhibiting sensation at the uncued bodily location. Such task-specific strategies could explain the observed patterns of lateralized mu oscillations and the difference in which hemisphere accounted for a greater share of variance in executive function skills. It is possible that attention to bodily sensations and variability in perceived boundaries between the body, peripersonal space and extrapersonal space contributed to these developmental and individual differences (Bremner & Spence, 2017), or that mu oscillations may have greater inter-individual variability than other alpha-range rhythms (Coll et al., 2017). Regardless, our findings indicate that neural responses during anticipation of a tactile stimulus index variation in stimulus processing speed, which could cascade into meaningful individual differences captured by measures of executive function (Willougby et al., 2018).

Further studies can address the potential utility of mu oscillations as an indicator of individual differences in how attention is deployed to the body. Neural responses during anticipation of a stimulus may offer a potential source of

variation in behavioral responses and stimulus processing speed, which could cascade into individual differences in measures of executive function.

Acknowledgments

The authors thank Yuheiry Rodriguez and Jebediah Taylor for their help with data collection. The education of SMW is supported by an NSF Graduate Research Fellowship. This research was supported by grants from the NSF (BCS-1460889 to PJM) and Pennsylvania Department of Health.

References

Anderson, K. L., & Ding, M. (2011). Attentional modulation of the somatosensory mu rhythm. Neurosci., 180, 165–180.

Bremner, A. J., & Spence, C. (2017). The Development of tactile perception. Adv. in Child Develop., 52, 227–268.

Cheyne, D., Gaetz, W., Garnero, L., Lachaux, J.-P., Ducorps, A., Schwartz, D., & Varela, F. J. (2003). Neuromagnetic imaging of cortical oscillations accompanying tactile stimulation. Cogn. Brain Res., 17, 599–611.

Cohen, M. X. (2018). Analyzing Neural Time Series Data: Theory and Practice. The MIT Press.

Coll, M.-P., Press, C., Hobson, H., Catmur, C., & Bird, G. (2017). Crossmodal Classification of Mu Rhythm Activity during Action Observation and Execution Suggests Specificity to Somatosensory Features of Actions. J. of Neuro., *37*, 5936–5947.

Engel, A. K., Fries, P., & Singer, W. (2001). Dynamic predictions: oscillations and synchrony in top-down processing. Nature Reviews, 2, 704–716.

Gomez-Ramirez, M., Hysaj, K., & Niebur, E. (2016). Neural mechanisms of selective attention in the somatosensory system. J. Neurophys., 116, 1218–1231.

Goljahani, A., Bisiacchi, P., & Sparacino, G. (2014). An Analyzing individual EEG alpha rhythm. *Comp. Methods Biomed.*, *113*, 853–861.

Haegens, S., Händel, B. F., & Jensen, O. (2011). Top-down controlled alpha band activity in somatosensory areas determines behavioral performance in a discrimination task. J. of Neuro., 31, 5197–5204.

Haegens, S., Luther, L., & Jensen, O. (2012). Somatosensory anticipatory alpha activity increases to suppress distracting input. J. of Cog. Neuro., 24, 677–685.

Hoffmann, S., & Falkenstein, M. (2008). The correction of eye blink artefacts in the EEG: A comparison of two prominent methods. PLoS ONE, 3, e3004.

Jensen, O., & Mazaheri, A. (2010). Shaping functional architecture by oscillatory alpha activity: gating by inhibition. Front. Human Neuro., 4, 186.

Jones, S. R., Kerr, C. E., Wan, Q., Pritchett, D. L., Hämäläinen, M., & Moore, C. I. (2010). Cued spatial attention drives functionally relevant modulation of the mu rhythm in primary somatosensory cortex. *J. of Neuro., 30*, 13760–13765

Klimesch, W. (1999). EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. Brain Res., 29, 169–195.

Klimesch, W., Sauseng, P., & Hanslmayr, S. (2007). EEG alpha oscillations: The inhibition–timing hypothesis. Brain Res., 53, 63–88.

Lopes da Silva, F. (2013). EEG and MEG: relevance to neuroscience. Neuron, 80, 1112–1128.

Makeig, S. & Delorme, A. (2004). Mining event-related brain dynamics. TICS, 8, 204–210.

Mazaheri, A., & Jensen, O. (2009). Prestimulus alpha and mu activity predicts failure to inhibit motor responses. Human Brain Map., 30, 1791–1800.

Nobre, A. C., & van Ede, F. (2017). Anticipated moments: temporal structure in attention. Nature Reviews, *19*, 34–48.

Penfield, W., & Boldrey, E. (1937). Somatic motor and sensory representation in the cerebral cortex of man as studied by electrical stimulation. Brain, 60, 389–443.

Pfurtscheller, G., & Da Silva, F. H. L. (1999). Event-related desynchronization. Elsevier BV.

Posner, M. I. (1980). Orienting of attention. The Quarterly J.of Exper. Psych., 32, 3–25.

Sadaghiani, S., & Kleinschmidt, A. (2016). Brain Networks and α-Oscillations: Structural and Functional Foundations of Cognitive Control. TICS, 20, 805–817.

Schroeder, C. E., & Lakatos, P. (2009). Low-frequency neuronal oscillations as instruments of sensory selection. Trends in Neuro., 32, 9–18.

Tamè, L., Braun, C., Holmes, N. P., Farnè, A., & Pavani, F. (2016). Bilateral representations of touch in the primary somatosensory cortex. Cog. Neuropsych., 33, 48–66.

Thut, G., Nietzel, A., Brandt, S. A., & Pascual-Leone, A. (2006). α-Band EEG activity over Occipital Cortex indexes visuospatial attention bias and predicts visual target detection. J. of Neuro., 26, 9494–9502.

van Ede, F., de Lange, F. P., & Maris, E. (2012). Attentional cues affect accuracy and reaction time via different cognitive and neural processes. J. of Neuro., 32, 10408–10412.

van Ede, F., de Lange, F. P., & Maris, E. (2014). Anticipation increases tactile stimulus processing in the ipsilateral primary somatosensory cortex. Cerebral Cortex, 24, 2562–2571.

Weiss, S. M., Meltzoff, A. N., & Marshall, P. J. (2018). Neural measures of bodily attention in children: Relations with executive function. Dev. Cog. Neuro, *34*, 148-158.

Willoughby, M. T., Blair, C. B., Kuhn, L. J., & Magnus, B. E. (2018). The benefits of adding a brief measure of simple reaction time to the assessment of executive function skills in early childhood. *J. of Exp. Child Psyc.*, *170*, 30–44.

Zanto, T. P., & Gazzaley, A. (2009). Neural suppression of irrelevant information underlies optimal working memory performance. J. of Neuro., 29, 3059–3066.

Zelazo, P. D., Anderson, J. E., Richler, J., Wallner-Allen, K., Beaumont, J. L., & Weintraub, S. (2013). II. NIH Cognitive Battery (CB): Measuring executive function and attention. *Monographs of the Society for Research in Child Development*, *78*, 16–33.

What Syntactic Structures block Dependencies in RNN Language Models?

Ethan Wilcox[1], **Roger Levy**[2], and **Richard Futrell**[3]

[1]Department of Linguistics, Harvard University, wilcoxeg@g.harvard.edu
[2]Department of Brain and Cognitive Sciences, MIT, rplevy@mit.edu
[3]Department of Language Science, UC Irvine, rfutrell@uci.edu

Abstract

Recurrent Neural Networks (RNNs) trained on a language modeling task have been shown to acquire a number of non-local grammatical dependencies with some success (Linzen, Dupoux, & Goldberg, 2016). Here, we provide new evidence that RNN language models are sensitive to hierarchical syntactic structure by investigating the **filler–gap dependency** and constraints on it, known as **syntactic islands**. Previous work is inconclusive about whether RNNs learn to attenuate their expectations for gaps in island constructions in particular or in *any* sufficiently complex syntactic environment. This paper gives new evidence for the former by providing control studies that have been lacking so far. We demonstrate that two state-of-the-art RNN models are are able to maintain the filler–gap dependency through unbounded sentential embeddings and are also sensitive to the hierarchical relationship between the filler and the gap. Next, we demonstrate that the models are able to maintain **possessive pronoun gender expectations** through island constructions—this control case rules out the possibility that island constructions block all information flow in these networks. We also evaluate three untested islands constraints: coordination islands, left branch islands, and sentential subject islands. Models are able to learn left branch islands and learn coordination islands gradiently, but fail to learn sentential subject islands. Through these controls and new tests, we provide evidence that model behavior is due to finer-grained expectations than gross syntactic complexity, but also that the models are conspicuously un-humanlike in some of their performance characteristics.

Keywords: Syntactic Islands, Recurrent Neural Networks, Blocking Effects, Acquisition of Syntax

Introduction

Recurrent Neural Networks (RNNs) with Long Short-Term Memory architecture (LSTMs) have achieved state-of-the-art scores at a number of natural language processing tasks, including language modeling and parsing (Hochreiter & Schmidhuber, 1997; Jozefowicz, Vinyals, Schuster, Shazeer, & Wu, 2016). In addition, they have begun to be used as a plausible sub-symbolic model for a variety of cognitive functions, including visual perception and language processing and comprehension (J. Elman, 1990). However, the distributed representations learned by RNNs and neural networks in general are notoriously opaque, posing a challenge for their interpretability as models of human sentence processing and for their controllability as NLP systems.

One recent line of work aims to uncover what these 'black boxes' learn about language by treating them like human psycholinguistic subjects. In this **psycholinguistic paradigm** RNNs trained on the language modeling task are fed handcrafted sentences, designed to expose their underlying syntactic knowledge (Linzen et al., 2016; McCoy, Frank, & Linzen, 2018). Much of this work has investigated what RNNs trained on a language modeling objective are capable of learning about natural syntactic dependencies. For the purposes of this investigation, we define **dependency** as any systematic co-variation between two words. For example, in one experiment networks were tested as to whether they had learned the number agreement dependency between a subject and a verb. They were fed with the prefix *The key to the cabinet...* and correctly gave a higher probability to the grammatical *is* over the ungrammatical *are*. Networks were shown to successfully complete this task for a number of languages, as well as for sentences whose content words were replaced with random alternatives of the same syntactic category rendering them syntactically licit but semantically implausible (Gulordava, Bojanowski, Grave, Linzen, & Baroni, 2018).

But learning that covariance exists between certain words or word forms, without reference to their relative positions, is not enough to say that the RNN models have fully learned a dependency. Natural language dependencies consist of covariation between two elements in *certain syntactic positions*. Agents must both attend to the structural relationship between the two elements bound by the dependency and filter out intervening material in syntactically irrelevant positions. The subject–verb number agreement task above provides compelling evidence that RNNs are capable of the latter: they were able to maintain correct predictions despite a number of *distractors* that mismatched the subject in number, such as *cabinet* in the example provided (Marvin & Linzen, 2018).

Evidence suggesting that RNN language models are also sensitive to the structural relationship between the two bound elements has emerged from the study of **filler–gap dependencies** (Wilcox, Levy, Morita, & Futrell, 2018; Chowdhury & Zamparelli, 2018). The filler–gap dependency is the dependency between a *filler*—such as *who* or *what*—and and a gap, which is an empty syntactic position. Crucially, filler–gap dependencies are subject to a number of constraints, known as *island constraints*, which are a set of structural positions that prevent the filler and the gap from entering into a dependency with each other (Ross, 1967). (1-b) gives one example island, in which the dependency is blocked by a wh-complementizer.

(1) a. I know what the guide said that the lion devoured __ yesterday. NO VIOLATION
b. *I know what the guide said whether the lion devoured __ yesterday. WH-ISLAND ISLAND VIOLATION

While it has been shown that both simple Elman RNNs and more contemporary LSTMs are able to represent the basic covariance between fillers and gaps, as well as other non-structural aspects of dependency, it is still uncertain whether

the models are sensitive to island constraints (J. L. Elman, 1991). Previous work has demonstrated that two state-of-the-art models are sensitive to three of the most-studied island constraints (wh-islands, complex NP islands and adjunct islands) but insensitive to a fourth (subject islands) (Wilcox et al., 2018). Others have concluded that the models are merely sensitive to syntactic complexity plus order. Chowdhury and Zamparelli (2018) compared sentence-level perplexity scores obtained by RNN LMs for wh-questions that violate island constraints, and yes-no questions and statements that violate no grammatical rules but contain the same syntactic structures. While the models obtained better perplexity scores on the statements compared to the island-violation questions, they performed similarly on the island-violations and non-violating yes/no questions. These results may indicate that RNNs are not learning to attenuate their expectations for gaps in island constructions in particular, but in *any* sufficiently complex syntactic environment.

This paper adjudicates between these two accounts of model behavior by providing control studies that have been lacking so far. In the first section, we demonstrate that two state-of-the-art LSTM models are sensitive to some forms of syntactic complexity, but not to others. Models are able to maintain the filler–gap dependency through **unbounded sentential embeddings** and yet are sensitive to the **hierarchical relationship** between the filler and the gap, suggesting that only specific types of syntactic complexity block gap expectations. In the second section, we turn to **possessive pronoun gender dependencies**, demonstrating that the models are able to maintain general expectations through island constructions—it is not the case that island constructions block all information flow in these networks. In this section we also evaluate three untested islands constraints: **coordination islands**, **left branch islands**, and **sentential subject islands**. Models are able to learn left branch islands and coordination islands gradiently, but fail to learn sentential subject islands. Through these controls and new tests, we provide evidence that model behavior is due to finer-grained expectations than gross syntactic complexity, but also that the models are conspicuously un-humanlike in some of their performance characteristics.

Methods

Language Models

We assess two state-of-the-art pre-existing LSTM models trained on English text for a language modeling objective. The first model, which we refer to as the **Google Model**, was trained on the One Billion Word Benchmark and has two hidden layers with 8196 units each. It uses the output of a character-level convolutional neural network (CNN) as input to the LSTM (and was originally presented as the *BIG LSTM+CNN Inputs*) (Jozefowicz et al., 2016). The second model, which we refer to as the *Gulordava Model* was selected for its previous success at learning the subject-verb number agreement task. It was trained on 90 Million tokens of English Wikipedia, and has two hidden layers of 650 units each (Gulordava et al., 2018).

Dependent Measure: Surprisal

In this work we take a grammatical dependency to be the covariance between an upstream *licensor* and a downstream *licensee*. We assess the model's knowledge of the dependency by measuring the effect that the licensor has on the **surprisal** of the licensee, or on material immediately following the licensee when it is a gap. Surprisal, or negative log-conditional probability , $S(x_i)$ of a sentence's i^{th} word x_i, tells us how strongly x_i is expected under the language model's probability distribution. For sentences out of context, the surprisal is: $S(x_i) = -\log p(x_i|x_1 \dots x_{i-1})$. Surprisal is known to correlate directly with processing difficulty in humans (Smith & Levy, 2013; Hale, 2001; Levy, 2008). In this work, we expect that grammatical licensors set up expectations for licensee, reducing its surprisal compared to minimal pairs in which the licensor is absent. We derive the word surprisal from the LSTM langauge model by directly computing the negative log of the predicted conditional probability $p(x_i|x_1 \dots x_{i-1})$ from the softmax layer.

Experimental Design: Wh-Licensing Interaction

The filler–gap dependency is biconditional: Fillers set up expectations for gaps and gaps require fillers to be licensed. To measure this bi-directionality we employ the 2x2 interaction design proposed in Wilcox et al.. There, the authors measure the **wh-licensing interaction**, which they compute from four sentence variants, given in (2), that contain the four possible combinations of fillers and gaps for a specific syntactic position. Note that the underscores are for presentational purposes only, and were not included in test items. Subsequent examples will be given via the (2-d) example, but all four variants were created in order to compute the licensing interaction.

(2) a. I know that you insulted your aunt yesterday. [-FILLER -GAP]
b. *I know who you insulted your aunt yesterday. [+FILLER -GAP]
c. *I know that you insulted __ yesterday. [-FILLER +GAP]
d. I know who you insulted __ yesterday. [+FILLER +GAP]

If the filler sets up an expectation for a gap, then the filled syntactic position where a gap would typically occur should be more surprising in contexts that contain an upstream filler. That is $S(b) - S(a)$ should be a large positive number. If the gap requires a filler to be licensed, then the transition from the embedded verb to the S-modifying PP 'yesterday' that skips over the otherwise-required grammatical object should be more surprising in contexts without an upstream filler. That is, $S(d) - S(c)$ should also be a large negative number. We can assess how well the model has learned both expectations by measuring the difference of differences: $[S(b) - S(a)] - [S(d) - S(c)]$. This is the wh-licensing interaction. If the models are learning the filler–gap dependency, we expect this to be a large positive number, with typical models showing about 4 bits of licensing interaction in simple object extracted clauses such as (2). Although we might expect the

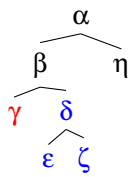


Figure 1: C-Command in a binary-branching tree structure. γ *c*-commands all the nodes in blue, but does not *c*-command the black nodes.

strongest difference in surprisal between (2-a) and (2-b) to be on the filled-gap position, *your aunt*, this material is elided in two of the conditions. Therefore, in order to keep the measurement site the same across all four conditions, we measure wh-licensing interaction in the post-gap prepositional phrase ('yesterday' in (2)).

In previous work using this methodology, RNN knowledge of island constraints was assessed by comparing the licensing interaction in island configurations to that in non-island minimal pairs. Strong evidence for an island constraint would be if the wh-licensing interaction dips to zero for a gap in island position, indicating that the model has decoupled expectations for fillers from gaps in this position. In practice we look for a significant decrease in wh-licensing interaction as indication that the models have learned to attenuate their expectations for gaps within islands. We derive the statistical significance of the interaction from a mixed-effects linear regression model, using some-coded conditions (Baayen, Davidson, & Bates, 2008). We include random intercepts by item but omit random slopes as we do not have repeated observations within items and conditions (Barr, Levy, Scheepers, & Tily, 2013). In our figures, error bars represent 95% confidence intervals of the contrasts between conditions, computed by subtracting out the by-item means before calculating the intervals as advocated in (Masson & Loftus, 2003). [1]

Syntactic Complexity

Unboundedness

The filler–gap dependency can span through a potentially unbounded number of sentential embeddings. To test whether models' expectations were attenuated with greater embedding depth, we created 23 items in five experimental conditions with between 0 and 4 layers of embedding and gaps in either object or indirect object (goal) position, following the examples in (3), and measured the licensing interaction in the post-gap material. (In this and subsequent examples, the material in which the interaction is measured will be highlighted in bold.)

(3) a. I know who you insulted __ **at the party**. [OBJECT GAP, 0 LAYERS]

b. I know who the gardener reported the butler said the hostess believed her aunt suspected you insulted __ **at the party**. [OBJECT GAP, 4 LAYERS]

c. I know who you delivered a challenge to __ **at the party**. [GOAL GAP, 0 LAYERS]

d. I know who the gardener reported the butler said the hostess believed her aunt suspected you delivered a challenge to __ **at the party**. [GOAL GAP, 4 LAYERS]

The results for this experiment can be seen in figure 2, with the object gap results on the top and goal gap results on the bottom. First, we find a significant interaction between fillers and gaps resulting in supperaditive reduction of surprisal ($p < 0.001$ for all conditions) indicating that both models have learned the filler–gap dependency. Starting with the object gap conditions: For the google model, we find no effect of embedding depth on the wh-licensing interaction ($p > 0.85$ in all cases); for the gulordava model, we find a significant decrease in wh-licensing interaction only between the *no embedding* conditions and conditions with 3 or 4 additional layers of embedding ($p < 0.001$ in both). When the gap occurs in the goal position, for the google model, we find no significant effect of embedding depth of the wh-licensing interaction. For the gulordava model, we find a generally smaller wh-licensing interaciton, as well as a significant effect of embedding between the *no embedding* condition and conditions with two or more additional embedding layers ($p < 0.05, p < 0.05, p < 0.01$ for 2 ,3 and 4 layers). We take these results to indicate that the google model has learned the unboundedness of the filler–gap dependency whereas the gulordava model has learned only relative unboundedness and shows behavior that reflects human performance more than human competence. However, these results indicate that both models can, in principle, thread their expectations for gaps through complex syntactic structures, if we take the number of syntactic nodes as a proxy measure for syntactic complexity.

Syntactic Hierarchy

Although the filler–gap dependency is unbounded, it is subject to a number of hierarchical constraints, the most basic of which is that the filler must be "above" the gap, structurally. Here, we take this to mean that the filler must *c*-command the gap, although the precise relationship is more complex (Pollard & Sag, 1994). Structurally-speaking node γ *c*-commands node δ if neither node directly dominates the other and every node X that dominates γ also dominates δ. Figure 1 demonstrates this relationship, with the noes *c*-commanded by γ highlighted in blue.

To assess whether the models had learned this constraint on the structural relationship we created 24 variants following the examples in (4) and measured the wh-licensing interaction in the post-gap PP. If the model has learned the structural constraints on the filler–gap dependency, an undischarged filler in the matrix clause should not make a gap in subsequent parts of the sentence more or less likely, leading to near-zero licensing interaction in the *Matrix Clause* condition.

[1] Our studies were preregistered on `aspredicted.org`: To see the preregistrations go to `aspredicted.org/blind.php?=X` where $X \in \{$`sz8f5d,2r2eu7,zt73qt,es8rx7,f9pk9f,se6i2e`$\}$.

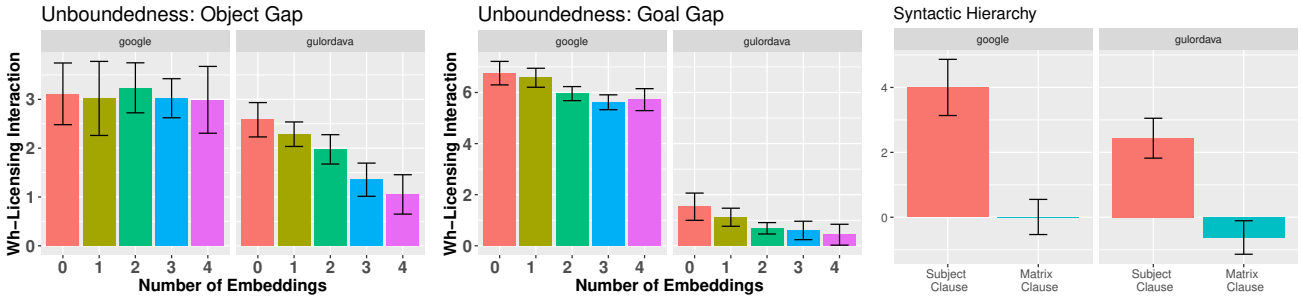


Figure 2: Effect of sentential embedding and syntactic hierarchy on wh-licensing interaction.

(4) a. The fact that the mayor knows who the criminal shot __ **shocked the jury** during the trial. [SUBJECT]
b. *The fact that the mayor knows who the criminal shot the teller shocked __ **during the trial**. [MATRIX]

The results from this experiment can be seen in Figure 2, on the far right panel. We find strong licensing interaction for the grammatical *Subject Clause* conditions (in red), but a striking reduction in licensing interaction for the *Matrix Clause* conditions (in blue), which is significant for both models ($p < 0.001$). As the results in (2) and Wilcox et al. have shown that RNN models are insensitive to linear distance between the filler and the gap, we take these results suggest that it is the relevant structural properties which block the models' expectations for gaps inside the matrix clause.

Island Effects: Gender Expectation vs. Filler–Gap Dependency

Island constraints are specific syntactic configurations that block the filler–gap dependency. One way to show that the RNN models are learning island conditions as constraints on the filler–gap dependency is to demonstrate that they are capable of threading other expectations into island configurations. To do this, we used **pronoun gender expectation** between a gendered noun, such as 'actress' or 'husband', and a possessive pronoun such as 'his' or 'her.'. Nouns that carry overt gender marking or culturally-imbued gender bias set up expectations that subsequent pronominals match them in gender. Previous work has shown that humans thread expectations set up by *cataphoric pronouns* into syntactic islands (Yoshida, Kazanina, Pablos, & Sturt, 2014). Cataphoric pronouns are pronouns that precede the nominal element to which they refer, as in (5).

(5) **Her** manager revealed that the studio notified **Judy Dench** about the new film.

Because cataphoric pronouns are relatively less frequent than anaphoric pronouns, which follow the nominal to which they refer, we use sentences such as those in (6) to assess whether RNN LMs can thread expectations into island environments. We measure the strength of the gender expectation by calculating the difference in surprisal between the matching condition and the mismatching condition, or S((6-b))-S((6-a)). If the models attenuate their expectation for gender agreement in island positions, then we expect an interaction between MISMATCH and ISLAND resulting in supperaditivally lower surprisal.

(6) a. The actress said that they insulted **her** friends. [MATCH, CONTROL]
b. #The actress said that they insulted **his** friends. [MISMATCH, CONTROL]
c. The actress said whether they insulted **her** friends. [MATCH, ISLAND]
d. #The actress said whether they insulted **his** friends. [MISMATCH, ISLAND]

In order to test whether the models maintained their gender expectations through island constructions, we created six suites of experiments following the pattern of (6) for six of the most frequently studied islands constructions. For each of the gender expectation experiments, we created 30 variants, 15 with masculine subjects and 15 with feminine subjects and measured the surprisal at the possessive pronoun. The results are presented on the bottom row in Figure 3 alongside model performance on the filler–gap dependency for the same syntactic constructions (top row). For the filler–gap dependency, results for four islands had already been tested in Wilcox et al. (2018), which we present alongside novel results for *Coordination Islands*, *Sentential Subject Islands* and *Left-Branch Islands*, the latter separately without a gender expectation control. For these experiments, we created between 20-24 experimental items and measured the wh-licensing interaction in the post-gap material. We take a reduction in wh-licensing interaction in island constructions and no such reduction in the gender expectation as evidence that the model has both learned the island constraint, and has applied that constraint uniquely to the filler–gap dependency.

Wh-Islands The wh-constraint states that the filler–gap dependency is blocked by S-nodes introduced by a wh-complimentizer, as demonstrated in the unacceptability of (7-b) compared to (7-a). We created experimental items following the examples in (7) and measured their gender expectation and filler–gap dependency (filler–gap dependency materials were taken from Wilcox et al.).

(7) a. I know who Alex said your friend insulted __ **yesterday**. [CONTROL, FILLER–GAP]
b. *I know who Alex said whether your friend insulted __ **yesterday**. [ISLAND, FILLER–GAP]
c. The actress said they insulted {**his/her**} friends. [CONTROL, GENDER EXP.]
d. The actress said whether they insulted {**his/her**}

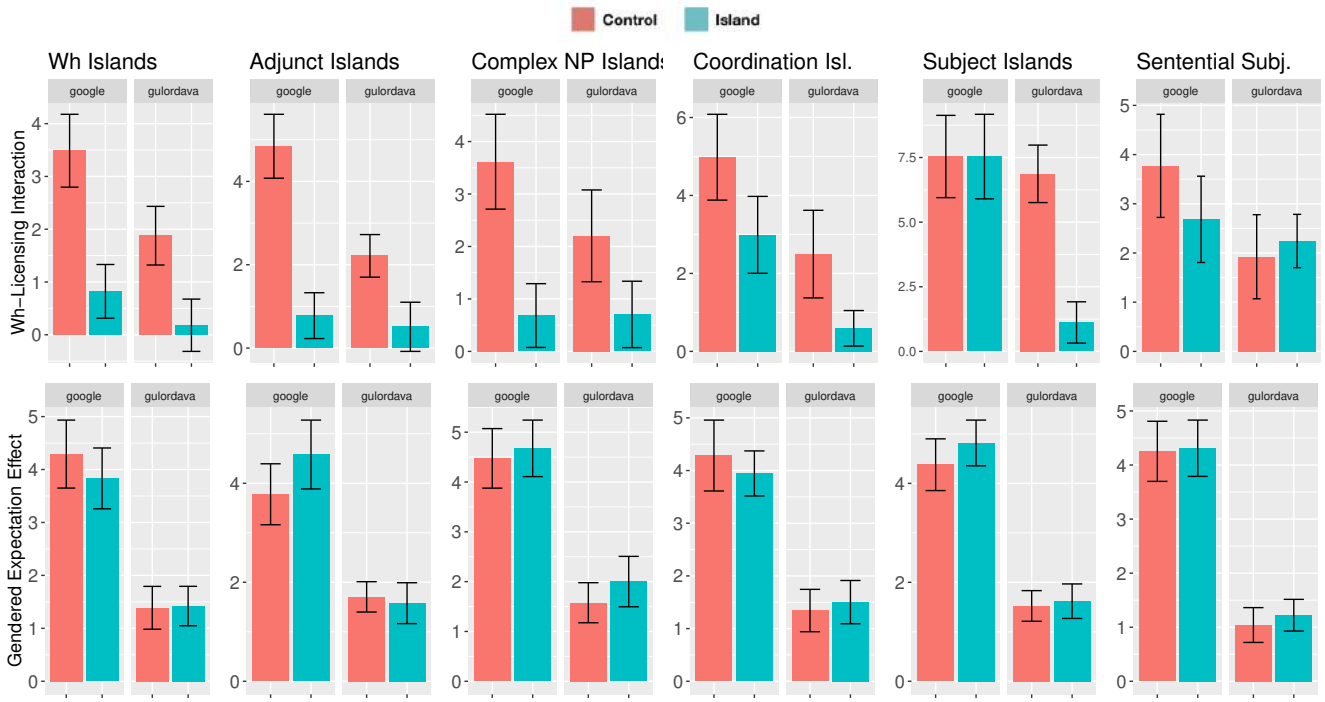


Figure 3: Effect of island construction on gender dependency.

friends. [ISLAND, GENDER EXP.]

The results for this experiment can be seen in the far left panel of Figure 3, with island structures graphed in blue and non-island controls in red. We find a significant difference in licensing interaction between the island and non-island conditions for both the google and gulordava models ($p < 0.001$ for both models), but no such difference in gender expectation.

Adjunct Islands Gaps cannot be licensed inside an adjunct clause, as demonstrated by the relative unacceptability of (8-a) over (8-b).

(8) a. I know what the librarian placed __ **on the wrong shelf**. [CONTROL, FILLER–GAP]
b. *what the patrong got mad after the librarian placed __ **on the wrong shelf**. [ISLAND, FILLER–GAP]
c. The actress thinks they insulted {**his/her**} performance [CONTROL, GENDER EXP.]
d. The actress got mad after they insulted {**his/her**} performance. [ISLAND, GENDER EXP.]

The results for this experiment can be seen in Figure 3, second panel from the left. We find a significant reduction of wh-licensing interaction between the control and island conditions in the case of the filler–gap dependency for both models ($p < 0.001$ google; $p < 0.01$ gulordava; materials taken from]Wilcox et al.). However, we find no effect of syntactic structure on the gender effect.

Complex NP Islands Gaps are not licensed inside S-nodes that are dominated by a lexical head noun, as demonstrated by the relative badness of (9-b) compareid to (9-a).

(9) a. I know what the actress bought __ **yesterday**. [CONTROL, FILLER–GAP]
b. *I know what the actress bought the painting that depicted __ **yesterday**. [ISLAND, FILLER–GAP]
c. The actress said they saw her {**his/her**} performance. [CONTROL, GENDER EXP.]
d. The actress said they saw the exhibit that featured {**his/her**} performance. [ISLAND, GENDER EXP.]

We created items follwing the examples in (9), with filler–gap items adopted from (Wilcox et al., 2018). The results from this experiment can be found in the middle-left panel of Figure 3. We found an effect of syntactic location on wh-licensing interaction for both models ($p < 0.001$ google; $p < 0.01$ gulordava) but no such interaction for gender expectations.

Coordination Islands The coordination constraint states that a gap cannot occur in one half of a coordinate structure as demonstrated by the difference between (10-b) and (10-a), in which a whole conjunct has been gapped.

(10) a. I know what the man bought __ **at the antique shop**. [CONTROL, FILLER–GAP]
b. *I know what the man bought the painting and __ **at the antique shop**. [ISLAND, FILLER–GAP]
c. The fireman knows they talked about {**his/her**} performance. [CONTROL, GENDER EXP.]
d. The fireman knows they talked about the football game and {**his/her**} performance. [ISLAND, GENDER EXP.]

We created experimental items following the examples in (10). Results can be seen in 3 center-right panel. For the filler–gap dependency, in both models there is a significant difference between the *control* condition and *island* conditions ($p < 0.05$ for both models). These results indicate that the models have somewhat attenuated expectations for gaps when they occur in the second half of a coordinate struc-

ture. However, note that, at least for the google model, the wh-licensing interaction is significantly greater than zero, indicating that this model still maintains *some* expectation for gaps in this syntactic location. For both models there is no difference in gender expectation between the *control* and *island* conditions).

Subject Islands Gaps are generally licensed in prepositional phrases, except when they occur attached to sentential subjects. We created experimental items following the examples in (11), with filler–gap materials adapted from Wilcox et al..

(11)a. I know what _ **fetched** a high price. [CONTROL, FILLER-GAP]

b. *I know who the painting that depicted _ **fetched** a high price. [ISLAND, FILLER–GAP]

c. The actress said they sold the painting by {**his/her**} friend. [CONTROL, GENDER EXP.]

d. The actress said the painting by {**his/her**} friend sold for a lot of money. [ISLAND, GENDER EXP.]

The results from this experiment can be seen in Figure 3, second panel from the right. For the filler–gap dependency, we found a significant difference between the *control* and *island* condition in the case of the gulordava model ($p < 0.01$), but no such reduction in the case of the google model. For gender expectation, we found no significant difference between the two conditions.

Sentential Subject Islands The sentential subject constraint states that gaps are not licensed within an S-node that plays the role of a sentential subject. To assess whether the RNN models had learned this constraint we created items following the variants in (12).

(12)a. I know who the seniors defeated _ **last week**. [CONTROL, FILLER–GAP]

b. I know who for the seniors to defeat _ **will be trivial**. [ISLAND, FILLER–GAP]

c. The fireman knows they will save {**his/her**} friend. [CONTROL, GENDER EXP.]

d. The fireman knows for them to save {**his/her**} friend will be difficult. [ISLAND, GENDER EXP.]

The results for this experiment can be seen in Figure 3, in the far right panel. We found no decrease in gender expectation between the *control* and *island* conditions for either model. Likewise, for the filler–gap dependency we found no significant decrease in wh-licesning interaction between the island and non island conditions in either model. These results indicate that neither model suspends its expectations for gaps within sentential subjects.

Left Branch Islands The left-branch constraint states that modifiers which appear on the left branch under an NP cannot be gapped, which accounts for the relative ungrammaticality of (13-b) compared to (13-a). Because possessive pronouns cannot grammatically occur in left-branches under an NP, this experiment examines only the filler–gap dependency. We created 20 items following the examples in (13) and measured the wh-licensing interaction in the post-gap material.

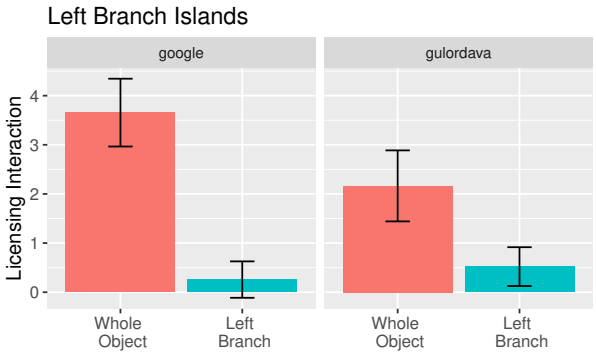


Figure 4: Left Branch Islands.

(13)a. I know what color car you bought _ **last week**. [WHOLE OBJECT]

b. I know what color you bought _ **car last week**. [LEFT BRANCH]

The results from this experiment can be seen in Figure 4 with experimental conditions on the x-axis and wh-licensing interaction on the y-axis. We see strong wh-licensing interaction in the two *whole object* conditions, but a significant reduction in licensing interaction when the gap consists of the Adjective Phrase modifier ($p < 0.001$ for the google model; $p < 0.05$ for the gulordava model). This results indicate that the models have learned the left branch islands, insofar as they do not expect left-branching modifiers to be extracted without the NP to which they are attached.

For every condition tested we found that the expectation set up by gendered subjects for possessive pronouns is not affected by the pronoun's location inside island constructions. For the three novel structures, we found that the two models tested are sensitive to left branch islands and gradiently to coordination islands, but not to sentential subject islands.

Discussion

The filler–gap dependency has been the focus of intense research for over fifty years because it is both far reaching and tightly constrained. It can be threaded through a potentially unbounded number of sentential embeddings; yet the filler must syntactically dominate the gap and the dependency is subject to a number of highly-specific blocking 'island' conditions. In this work we have shown that RNNs trained on a language modeling objective have learned both the power and the constraints imposed on this dependency. First, we provided evidence that they are able to thread the dependency through an unbounded number of sentential embeddings, and have also learned the constraints that govern the syntactic hierarchy of the filler relative to the gap.

Second, using gender expectation effects, we have demonstrated that the models are able to thread some contextually-dependent expectations into island constructions, providing evidence that previously-observed island effects have been learned for the filler–gap dependency *in particular*, and are not due to the model's inability to thread *any* information into syntactic islands. In addition, we have increased the experimental coverage of island effects, demonstrating that the models were able to learn left-branch islands and gradiently

learn coordination islands, but failed to learn sentential subject islands. This brings the total number of islands learned to 5/7 for the google model and 6/7 for the gulordava model. Although some of the model behavior remains strikingly unlike human acceptability judgements (in e.g. coordination islands), these experiments demonstrate that sequence models trained on a language modeling objective are able to separate natural language dependencies from each other and learn different fine-grained syntactic rules for each.

References

Baayen, R. H., Davidson, D., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Chowdhury, S. A., & Zamparelli, R. (2018). Rnn simulations of grammaticality judgments on long-distance dependencies. In *Proceedings of the 27th international conference on computational linguistics* (pp. 133–144).

Elman, J. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211.

Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine learning*, *7*(2-3).

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., & Baroni, M. (2018). Colorless green recurrent networks dream hierarchically. In *Proceedings of naacl.*

Hale, J. T. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the north american chapter of the association for computational linguistics and language technologies* (pp. 1–8).

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Jozefowicz, R., Vinyals, O., Schuster, M., Shazeer, N., & Wu, Y. (2016). Exploring the limits of language modeling. *arXiv*, *1602.02410*.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.

Linzen, T., Dupoux, E., & Goldberg, Y. (2016). Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, *4*, 521–535.

Marvin, R., & Linzen, T. (2018). Targeted syntactic evaluation of language models. *arXiv preprint arXiv:1808.09031*.

Masson, M. E., & Loftus, G. R. (2003). Using confidence intervals for graphically based data interpretation. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *57*(3), 203.

McCoy, R. T., Frank, R., & Linzen, T. (2018). Revisiting the poverty of the stimulus: hierarchical generalization without a hierarchical bias in recurrent neural networks. *arXiv preprint arXiv:1802.09091*.

Pollard, C., & Sag, I. A. (1994). *Head-driven phrase structure grammar*. Stanford, CA: Center for the Study of Language and Information.

Ross, J. R. (1967). Constraints on variables in syntax.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302–319.

Wilcox, E., Levy, R., Morita, T., & Futrell, R. (2018). What do rnn language models learn about filler-gap dependencies? *arXiv preprint arXiv:1809.00042*.

Yoshida, M., Kazanina, N., Pablos, L., & Sturt, P. (2014). On the origin of islands. *Language, Cognition and Neuroscience*, *29*(7), 761–770.

An ACT-R approach to investigating mechanisms of performance-related changes in an interrupted learning task

Maria Wirzberger (maria.wirzberger@tuebingen.mpg.de)
Max Planck Research Group "Rationality Enhancement",
Max Planck Institute for Intelligent Systems, Tübingen, Germany

Jelmer P. Borst (j.p.borst@rug.nl)
Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence,
University of Groningen, Groningen, Netherlands

Josef F. Krems (josef.krems@psychologie.tu-chemnitz.de)
Cognitive and Engineering Psychology, Institute for Psychology,
Faculty of Human and Social Sciences, TU Chemnitz, Germany

Günter Daniel Rey (guenter-daniel.rey@phil.tu-chemnitz.de)
Psychology of Learning with Digital Media, Institute for Media Research,
Faculty of Humanities, TU Chemnitz, Germany

Abstract

Learning constitutes an essential part of human experience over the life course. Independent of the domain, it is characterized by changes in performance. But what cognitive mechanisms are responsible for these changes and how do situational features affect the dynamics? To inspect that in more detail, this paper introduces a cognitive modeling approach that investigates performance-related changes in learning situations. It leverages the cognitive architecture ACT-R to model learner behavior in an interrupted learning task in two conditions of task complexity. Comparisons with the original human dataset indicate a good fit in terms of both accuracy and reaction times. Although interruption effects are more obvious in the human data, they are prevalent as well in the model. Furthermore, the model can map the learning effects, particularly in the easy task condition. Based on the existing mapping of ACT-R module activity with fMRI data, simulated neural activity is computed to investigate underlying cognitive mechanisms in more detail. The resulting evidence connects learning and interruption effects in both task conditions with activation-related patterns to explain changes in performance.

Keywords: Learning; Interruption; Cognitive performance; ACT-R; Simulated neural activity

Introduction

As an omnipresent requirement, learning happens throughout the entire life. From speaking the first words as a child to operating new technical devices as an elderly, the establishment of knowledge structures constitutes a core outcome of learning processes of all kind. Previous research indicated benefits in terms of performance, once already existing knowledge structures can be applied automatically (e.g., Wirzberger, Herms, Esmaeili Bijarsari, Eibl, & Rey, 2018). Besides these internally occurring process-related changes, externally induced situational characteristics such as interruptions also effect cognitive performance. Interruptions are highly prevalent across various contexts in daily life, including learning situations (e.g., Scheiter, Gerjets, & Heise, 2014). They can be described as usually neither planned nor expected cognitive breaks in the task performed up to that time (Brixey et al., 2007). To avoid or at least minimize resulting impairments, the interplay of interruptions and learning effects needs to be inspected in more detail on a cognitive level. On this account, computational cognitive modeling approaches offer a promising way to gain insights into underlying dynamics.

Based on that, the current paper introduces an ACT-R model that performs an interrupted learning task and is inspected in terms of behavioral parameters and underlying neural processes. After briefly describing the modeled experimental task and core results from human data, the paper outlines characteristics of the cognitive architecture ACT-R (Anderson, 2007). Following an explanation of the underlying model concept, the behavioral results obtained from the model runs are presented and compared with the described human data. The subsequent chapter addresses model performance on a neural level by reporting results from simulated fMRI analyses.

In summary, the obtained evidence highlights the connection of observable changes in cognitive performance due to learning and interruption effects with the mechanism of activation.

Task outline

The task setting underneath the cognitive model is reported in more detail in Wirzberger, Esmaeili Bijarsari, and Rey (2017). Participants in this study had to learn four abstract geometric symbol combinations via trial and error by verifying feedback (Shute, 2008) over a total of 64 trials. As outlined in Figure 1, they were shown the first part of the combination at the beginning of a trial and had to select the appropriate response by mouse click. Afterward, they were informed about the correctness of their response as well as

the correct response in terms of errors. Task complexity was represented by the number of symbols in a defined order that formed a combination. In the easy task condition, symbol combinations consisted of two symbols (input-response), whereas in the difficult task condition three symbols (input-input-response) formed a combination.

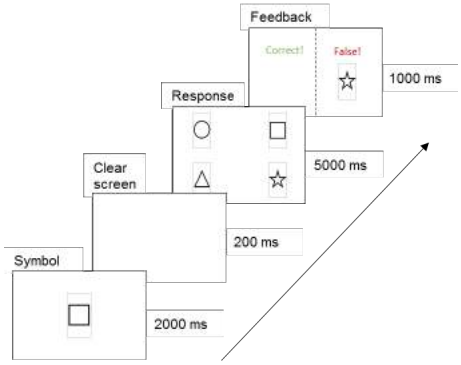


Figure 1: Sample learning trial in the easy task condition (adapted from Wirzberger et al., 2017).

At five pre-defined stages over the task (i.e., after trials 8, 24, 32, 40, and 56), an interrupting visual search task was induced. As displayed in Figure 2, it required participants to count the number of two out of four types of geometric symbols on a visual search screen and provide their responses afterward. The screens were accompanied by an instruction on the target symbols. A high similarity to the symbols used in the learning task (Gillie & Broadbent, 1989; Trick, 2008) and an appropriate task duration (Monk, Trafton, & Boehm-Davis, 2008) should ensure its interrupting potential.

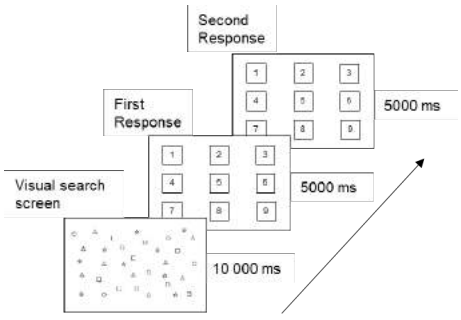


Figure 2: Sample interruption trial (adapted from Wirzberger et al., 2017).

Experimental results

In terms of reaction times in correctly solved trials, Figure 3 shows that participants speed up with increasing task progress in both conditions. Standard errors decrease over trials due to the increasing number of correct reactions. In addition, resumption effects are observable in both conditions, but are more distinctive in the easy task condition. Approaching accuracy, Figure 4 indicates that participants in the difficult task condition learn slower, but in the end both conditions reach a comparable level. Again, resumption effects are more prevalent in the easy task condition. These effects raise the question which cognitive mechanisms underlie the observed learning- and interruption-related patterns.

Computational cognitive modeling

Building on vested psychological evidence on human information processing, computational cognitive modeling approaches offer the opportunity to derive well-founded explanations of behavioral phenomena. The idea of building models to explain cognitive phenomena has already been discussed by Wegener (1967), who outlined the indicative value of an electronic simulation of mental processes for deriving and validating the related hypotheses.

Constituting a prevalent and broadly used production-based approach, ACT-R (Anderson, 2007) is particularly characterized by its modular brain-inspired structure. The included *modules* represent goal planning (goal module), declarative memory (declarative module), intermediate problem states (imaginal module), action coordination (procedural module), the handling of visual and auditory inputs (visual and aural module), and motor and vocal outputs (motor and vocal module). The mapping of these modules on corresponding regions-of-interest (ROIs) in the human brain has been validated with fMRI data by Borst, Nijboer, Taatgen, van Rijn, and Anderson (2015). For instance, when a model retrieves content from declarative memory, increased activity in the declarative module corresponds to activity in the prefrontal cortex, which has proven to be sensitive to both retrieval and storage operations. Activity in the goal module corresponds to activity in the anterior cingulate cortex, which is involved in higher-level control functions such as attentional allocation or performance monitoring. *Buffers* with limited capacity serve as interface between modules and enable their communication. They can hold one information element at the same time, representing existing limitations in information processing resources.

ACT-R uses a hybrid approach of both symbolic and subsymbolic mechanisms: *chunks* of information from declarative memory are retrieved not only on the match of content but also based on their level of activation. Activation is calculated from the history and context of use of a chunk and has to exceed a defined threshold to be eligible for selection. The full equation for each chunk *i* involves the components displayed in the subsequent equation:

$$A_i = B_i + \sum_k \sum_j W_{kj} S_{ji} + \sum_l PM_{li} + \varepsilon. \qquad (1)$$

The recency and frequency of use of the chunk *i* is reflected by the *base-level activation* B_i. Each time a chunk is presented, its base-level activation is increased, which decays as a power function of the time since that presentation. These decay effects are summed up and then transformed

logarithmically. With the *spreading activation* mechanism (Anderson, 2007), ACT-R accounts for the fact that activation is distributed across related chunks that share information elements. It is represented in the equation by W_{kj}, the amount of activation from source j in buffer k, and S_{ji}, the strength of association from source j to chunk i. W_{kj} and S_{ji} are summed over all buffers that provide spreading activation and all chunks in the slot of the chunks in buffer k. As humans sometimes retrieve related but ultimately wrong information from memory, ACT-R further includes a partial matching mechanism. Based on initially defined similarities between chunks, a mismatch between request and actual retrieval is calculated. The higher the mismatch, the more the activity of the chunk is penalized (Lebiere, 1999). In the equation, P reflects the amount of weighting given to the similarity in slot l and M_{li} represents the similarity between the value l in the retrieval specification and the value in the corresponding slot of chunk i. M_{li} is summed over the slot values of the retrieval specification. The value of ε represents noise, which is computed at the time of a retrieval request for each chunk.

Model concept

Each model run starts with an initial set of the task goal to the symbol learning task, which is assumed to result from the previously read instruction. In the following, each learning trial builds upon three task-related steps: at first, the presented symbol is encoded, which is repeated for the second symbol in the case of the difficult condition. This procedure stores an intermediate representation of all encoded visual content in the problem state (Borst, Taatgen, & van Rijn, 2010, 2015; Nijboer, Borst, van Rijn, & Taatgen, 2016), for instance, the input symbols 'square – circle' in the difficult condition. Next, the model attempts to retrieve the associated response symbol from declarative memory. In the second step, a response is selected from the provided opportunities on the screen, either according to the retrieved chunk or by random choice in case of no successful retrieval. In the final step, the model searches for visual feedback on the given response and, in the case of a false response, an update of the existing intermediate representation. The final information contains both the input and the correct response parts of the symbol combinations, such as 'square – circle – square' in case of the previous example.

In the first trials, there is no sufficiently matching content or no content at all to retrieve, resulting in slower and less accurate responses. After being presented the input symbols several times and retrieving related content from declarative memory, the model performance gets increasingly faster and more accurate due to increasing chunk activation. In the current task, the above outlined spreading activiation mechanism particularly effects the difficult task condition. In more detail, symbol combinations including the same input symbols, such as 'square – circle' and 'circle – square', obtain equal activation, independent of the correct symbol order. Following the concept of element interactivity in instructional research (Sweller, 2010), task demands increase with more logically interrelated information elements that have to be processed simultaneously. In the current task, the symbols that form a combination can be regarded as information elements that are related to each other by order. Without considering the order information, a wrong input-response association is more likely to be retrieved, which is then penalized by the partial matching mechanism. In consequence, due to more potentially mismatching information, the chunks in the difficult condition receive less activation and are harder to retrieve.

Following a goal change due to the bottom-up triggered saliency of the interrupting task, the task procedure involves the steps of searching, counting, and responding to the indicated target symbols. Using a color to indicate the task switch followed the model implemented by Wirzberger and Russwinkel (2015) and represents the immediate attention to the related screen change. Tying in with evidence on pre-attentive and attentive processes in the visual module of ACT-R (Nyamsuren & Taatgen, 2013), the second visual-location request in the visual search is enhanced by additional information on stimulus color that relates to distinct characteristics of the presented symbols. After finishing the counting part that also employs the problem state (Borst et al., 2010, 2015; Nijboer et al., 2016), on each of the two response screens the model encodes the requested symbol and attempts to retrieves the potential answer. Again, the possibility to retrieve a wrong answer persists due to the partial matching mechanism. When resuming the learning task, in line with Altmann and Trafton (2002) the model attempts to retrieve the previous task goal and thus restores its representation. Emerging interruption effects can be attributed to a decay in the activation of chunks related to the learning task that slows down subsequent retrieval requests (Borst et al., 2010, 2015; Trafton, Altmann, Brock, & Minz, 2003).

Model comparison

The inspected model data based on $n = 100$ model runs in each condition to obtain robust conclusions from the average model performance. A further goal was to achieve a balanced fit pattern across both accuracy and reaction time in either condition. Compared to a base model[1] that includes neither spreading activation nor partial matching, the overall root mean squared scaled deviation (RMSSD) decreased by almost one standard error and fit indices were quite aligned.

Besides the shared prevalence of interruption effects, in both conditions the model speeds up in reaction time over trials. The visual inspection in Figure 3 indicates that it can map the decreasing progression particularly in the difficult task condition. However, the model performs slightly slower than human participants during most of the trials. On the level of numerical goodness-of-fit indices, the model achieved

[1] In addition to the base model and the reported model, models including either only spreading activation or partial matching were inspected. Due to the superior fit, only the final model that applies both mechanisms is reported.

RMSSD = 2.16 and R^2 = 0.58 in the difficult task condition. Apart from a subtler decrease in the beginning, the mapping also fits quite well for later trials in the easy task condition. On a numerical level, RMSSD = 1.67 and R^2 = 0.52 resulted in this condition.

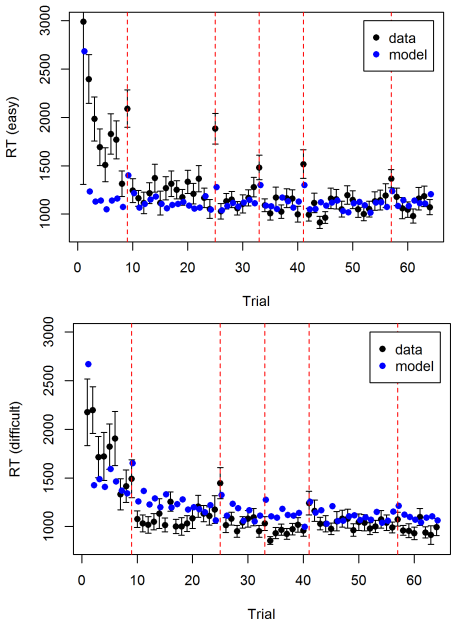


Figure 3: Comparison of human and model behavior in terms of reaction time. Red dashed lines indicate the first trial that immediately follows an interruption.

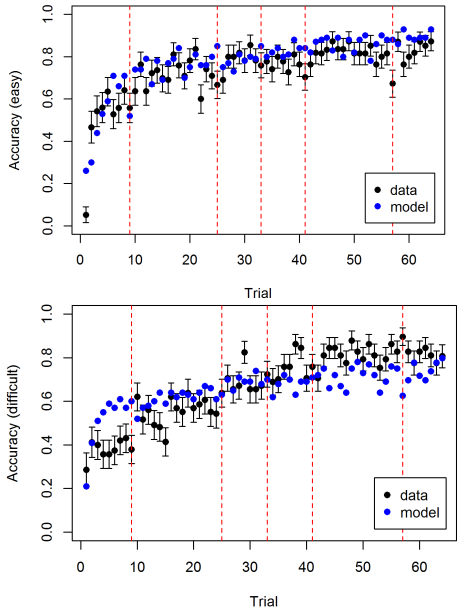


Figure 4: Comparison of human and model behavior in terms of accuracy. Red dashed lines indicate the first trial that immediately follows an interruption.

For accuracy, Figure 4 indicates that the model can map the progression in human behavior quite well in the easy task condition, although it achieves a higher performance in the end and shows a subtler reflection of interruption effects. On a numerical level, RMSSD = 1.51 and R^2 = 0.69 were achieved in this condition. The model in the difficult task condition learns slower compared to the easy task condition, but still faster than the human participants. However, apart from the nearly perfect location match in the last data points, it cannot fully map the final increase in the human data. The goodness-of-fit indices for the difficult task condition resulted in RMSSD = 2.07 and R^2 = 0.57.

Simulated fMRI data

Based upon the already mentioned mapping of activity in ACT-R modules on defined brain regions, simulated neural activity in predefined ROIs is computed to investigate underlying cognitive mechanisms in more detail (Borst & Anderson, 2017). This approach uses the recorded start and end times of module activity to simulate a signal comparable to the blood oxygenation level obtainable via fMRI, which shows peaks about 4-6 s after the occurrence of neural activity. In the first step, the activity of each inspected module is represented as 0-1 demand function and convolved afterward with the hemodynamic response function displayed in Figure 5. As an example, related to the task of the current model, longer retrieval times due to lower levels of chunk activation would result in increased activity in the declarative module. Such patterns are expectable in early stages of the task, with increased task difficulty, or caused by decay during an interruption, and would be observable by higher peaks in the resulting simulated signal.

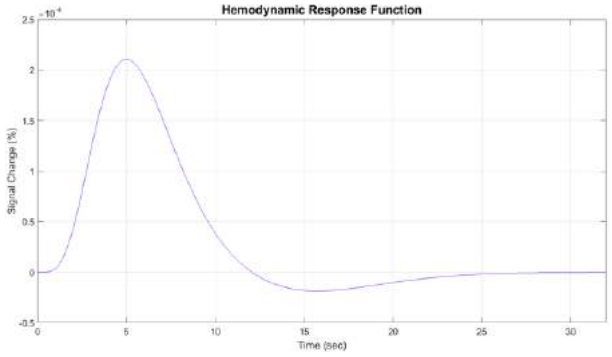


Figure 5: Hemodynamic response function (adapted from Borst & Anderson, 2017).

Prevalent changes in the declarative module activity across the learning task, which simulates activity in the prefrontal cortex, are displayed in Figure 6. Whereas blue lines represent the first third of the trials in the task, the red lines indicate the middle third of the trials, and the black lines refer to the last third of the trials. The curves predict a decrease in cognitive activity in later task stages in both conditions in the prefrontal cortex due to task-inherent learning processes. In the difficult task condition, represented by the dashed lines, a

higher level of activity is prevalent across all stages, with a particularly distinctive peak across early task stages. As already outlined, this relates to increased retrieval demands from lower levels of chunk activation.

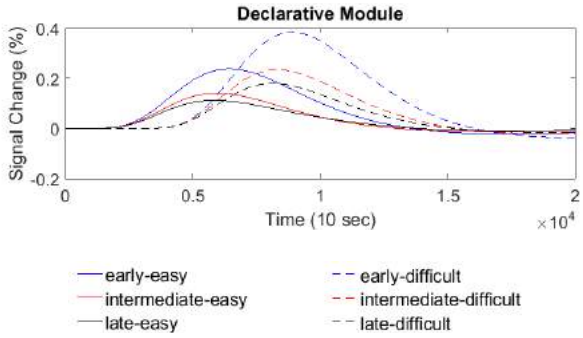


Figure 6: Simulated neural activity in the declarative module (corresponding to activity in the prefrontal cortex) across stages of the learning task.

Comparisons between the interrupting task and the learning task are depicted in Figure 7 and Figure 8. These include a separate visualization of the resumption phase (red lines), defined as the first trial that immediately follows the interrupting task. Across all inspected modules, activity levels in the interrupting task do not differ between both task conditions, since the solid and dashed blue lines overlap almost all the time. For both the declarative module, relating to the prefrontal cortex, and the goal module, relating to the anterior cingulate cortex, a higher activity across resumption trials compared to the remainder of trials in the learning task (black lines) is predicted for both conditions. In addition, differences between task conditions during the resumption phase are predicted for the anterior cingulate cortex and indicate higher levels of activity in the easy task condition. Even if these effects are less obvious in the behavioral model data, this also corresponds to the higher prevalence of resumption effects in the easy condition in the human data.

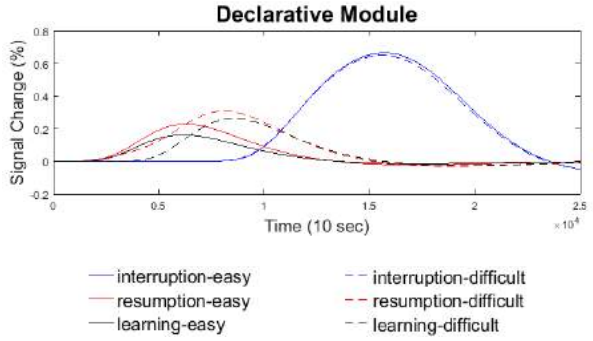


Figure 7: Simulated neural activity in the declarative module (corresponding to the prefrontal cortex) across interruption, resumption, and learning stages.

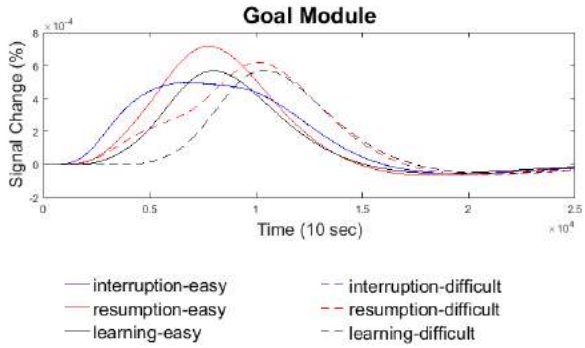


Figure 8: Simulated neural activity in the goal module (corresponding to the anterior cingulate cortex) across interruption, resumption, and learning stages.

Discussion

The current model explores cognitive mechanisms that underlie changes in performance due to the inserted interruptions and task-related learning processes. Comparing model performance across both conditions on a behavioral level, the obtained results indicate a good fit in terms of reaction times and accuracy. The model can map both the learning-related increase in performance and the decrease in performance due to experiencing an interruption. A potential improvement to increase the visibility of interruption effects in the model might involve adjusting when the model starts to retrieve information related to the correct response symbol. For the difficult task condition, the accuracy result pattern hints on a shift in task-related strategies. Due to the small number of learned symbol combinations, over time people in the difficult condition might have applied a more heuristic encoding strategy with focus on the first symbol, directly mapping task execution in the easy task condition. Taking this into account, the current modeling approach offers potential for future work by explaining such strategy shift with a more complex model on both the level of production rules and corresponding selection mechanisms.

The pattern observed in the simulated neural activity relates to the fact that the model needs to invest a higher amount of declarative memory resources upon each retrieval request in the early task stage due to the lack of suitable chunks and lower levels of chunk activation. The smaller level of cognitive activity with increasing task progress emphasizes the prevalence of learning effects in both conditions, as existing content in the declarative memory receives increasingly higher activation and thus can be retrieved faster and more accurately. In the difficult task condition, invested declarative resources are constantly higher across all stages, which by closer inspection relates to effects of spreading activation and the increased influence of partial matching that penalizes chunk activation and extends retrieval times. Increased levels of resumption-related activity in the declarative module arise from the activation decay in chunks related to the acquired symbol combinations. Observable differences in goal activity during the resumption stage align well with predictions stated by the memory-for-

goals model (Altmann & Trafton, 2002). They relate to the demand to rebuild the goal-representation of the learning task after each interruption. The obtained fMRI predictions will be compared with human data sets in the next step.

Conclusion

Taken together, the obtained results emphasize the importance of considering activation-related dynamics when approaching changes in performance in learning situations. The outlined cognitive modeling approach inspects the influence of both internal and external factors in these contexts and can be taken as promising step to investigating related patterns of cognitive resource investment. Since it extends beyond human experiments and model-based behavior on a neural level, it provides a more detailed understanding, which is crucial for developing adequate support and minimizing harmful effects.

Acknowledgements

The reported research was funded by the German Research Foundation (DFG), GRK 1780/1 and a travel grant by the Saxon State Ministry for Higher Education, Research and the Arts (SMWK).

References

Altmann, E. M., & Trafton, J. G. (2002). Memory for goals: an activation-based model, *Cognitive Science, 26*, 39–83. doi:10.1207/s15516709cog2601_2

Anderson, J. R. (2007). *How can the human mind occur in the physical universe?* New York, NY, USA: Oxford University Press.

Borst, J. P., & Anderson, J. R. (2017). A step-by-step tutorial on using the cognitive architecture ACT-R in combination with fMRI data. *Journal of Mathematical Psychology, 76*, 94-103. doi:10.1016/j.jmp.2016.05.005

Borst, J. P., Nijboer, M., Taatgen, N. A., van Rijn, H., & Anderson, J. (2015). Using data-driven model-brain mappings to constrain formal models of cognition. *PLoS ONE, 10*, e0119673. doi: 10.1371/journal.pone.0119673

Borst, J. P., Taatgen, N. A., & van Rijn, H. (2015). What makes interruptions disruptive? A process-model account of the effects of the problem state bottleneck on task interruption and resumption. In *CHI '15. Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 2971-2980). New York, NY, USA: ACM Press. doi:10.1145/2702123.2702156

Borst, J. P., Taatgen, N. A., & van Rijn, H. (2010). The problem state: A cognitive bottleneck in multitasking. *Journal of Experimental Psychology: Learning, Memory and Cognition, 36*, 363-382. doi:10.1037/a0018106

Brixey, J. J., Robinson, D. J., Johnson, C. W., Johnson, T. R., Turley, J. P., & Zhang, J. (2007). A concept analysis of the phenomenon interruption. *Advances in Nursing Science, 30*(1), E26–E42.

Gillie, T., & Broadbent, D. (1989). What makes interruptions disruptive?: A study of length, similarity, and complexity. *Psychological Research, 50*, 243-250. doi:10.1007/BF00309260

Lebiere, C. (1999). The dynamics of cognition: An ACT-R model of cognitive arithmetic. *Kognitionswissenschaft, 8*, 5–19. doi: 10.1007/BF03354932

Monk, C. A., Trafton, J. G., & Boehm-Davis, D. A. (2008). The effect of interruption duration and demand on resuming suspended goals. *Journal of Experimental Psychology: Applied, 14*, 299-313. doi:10.1037/a0014402

Nijboer, M., Borst, J., van Rijn, H., & Taatgen, N. (2016). Contrasting single and multi-component working-memory systems in dual tasking. *Cognitive Psychology, 86*, 1-26. doi: 10.1016/j.cogpsych.2016.01.003

Nyamsuren, E., & Taatgen, N. A. (2013). Pre-attentive and attentive vision module. *Cognitive Systems Research, 24*, 62-71. doi:10.1016/j.cogsys.2012.12.010

Scheiter, K., Gerjets, P., & Heise, E. (2014). Distraction during learning with hypermedia: Difficult tasks help to keep task goals on track. *Frontiers in Psychology, 5*, 268. doi:10.3389/fpsyg.2014.00268

Shute, V. (2008). Focus on formative feedback. *Review of Educational Research, 78*, 153-189. doi:10.3102/0034654307313795

Sweller, J. (2010). Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educational Psychology Review, 22*, 123-138. doi:10.1007/s10648-010-9128-5

Trafton, J. G., Altmann, E. M., Brock, D. P., & Mintz, F. E. (2003). Preparing to resume an interrupted task: Effects of prospective goal encoding and retrospective rehearsal. *International Journal of Human-Computer Studies, 58*, 583–603. doi:10.1016/S1071-5819(03)00023-5

Trick, L. M. (2008). More than superstition: Differential effects of featural heterogeneity and change on subitizing and counting. *Perception & Psychophysics, 70*, 743-760. doi:10.3758/PP.70.5.743

Wegener, H. (1967). Die Simulation menschlicher Intelligenz – Leistungen. In W. Kroebel (Ed.) *Fortschritte der Kybernetik* (pp. 375-385). Munich-Vienna, Germany-Austria: R. Oldenbourg Verlag.

Wirzberger, M., Esmaeili Bijarsari, S., & Rey, G. D. (2017). Embedded interruptions and task complexity influence schema-related cognitive load progression in an abstract learning task. *Acta Psychologica, 179*, 30-41. doi: 10.1016/j.actpsy.2017.07.001

Wirzberger, M., Herms, R., Esmaeili Bijarsari, S., Eibl, M., & Rey, G. D. (2018). Schema-related cognitive load influences performance, speech, and physiology in a dual-task setting: A continuous multi-measure approach. *Cognitive Research: Principles and Implications*, 3:46. doi: 10.1186/s441235-018-0138-z

Wirzberger, M., & Russwinkel, N. (2015). Modeling interruption and resumption in a smartphone task: An ACT-R approach. *i-com, 14*, 147-154. doi:10.1515/icom-2015-0033

Modality Effects in Vocabulary Acquisition

Merel C. Wolf (merel.wolf@mpi.nl)[1]
Alastair C. Smith (alastair.smith@mpi.nl)[1]
Antje S. Meyer (antje.meyer@mpi.nl)[1]
Caroline F. Rowland (caroline.rowland@mpi.nl)[2]

[1] Max Planck Institute for Psycholinguistics, Wundtlaan 1, 6525 XD Nijmegen, The Netherlands
[2] ESRC LuCiD Centre & Department of Psychological Sciences,
University of Liverpool, Bedford Street South, L69 7ZA, Liverpool, United Kingdom

Abstract

It is unknown whether modality affects the efficiency with which humans learn novel word forms and their meanings, with previous studies reporting both written and auditory advantages. The current study implements controls whose absence in previous work likely offers explanation for such contradictory findings. In two novel word learning experiments, participants were trained and tested on pseudoword - novel object pairs, with controls on: modality of test, modality of meaning, duration of exposure and transparency of word form. In both experiments word forms were presented in either their written or spoken form, each paired with a pictorial meaning (novel object). Following a 20-minute filler task, participants were tested on their ability to identify the picture-word form pairs on which they were trained. A between subjects design generated four participant groups per experiment 1) written training, written test; 2) written training, spoken test; 3) spoken training, written test; 4) spoken training, spoken test. In Experiment 1 the written stimulus was presented for a time period equal to the duration of the spoken form. Results showed that when the duration of exposure was equal, participants displayed a written training benefit. Given words can be read faster than the time taken for the spoken form to unfold, in Experiment 2 the written form was presented for 300 ms, sufficient time to read the word yet 65% shorter than the duration of the spoken form. No modality effect was observed under these conditions, when exposure to the word form was equivalent. These results demonstrate, at least for proficient readers, that when exposure to the word form is controlled across modalities the efficiency with which word form-meaning associations are learnt does not differ. Our results therefore suggest that, although we typically begin as aural-only word learners, we ultimately converge on developing learning mechanisms that learn equally efficiently from both written and spoken materials.

Keywords: modality effects; word learning; vocabulary acquisition; reading

Introduction

Novel words can be encountered through listening to speech or through reading text. Inherent properties of each modality will have specific processing demands and will pose specific constraints on the learning mechanisms that enable learning in these modalities. It is, however, not yet understood whether these modality-specific demands influence the efficiency of learning in these modalities. The present study aimed at investigating to what extent the modality in which information is presented affects the efficiency of learning novel word form – meaning associations.

The existing literature shows conflicting findings regarding the effect of modality on novel word learning. Concerning word form learning only, benefits have been found in favour of the spoken modality (Bakker, Takashima, Van Hell, Janzen, & McQueen, 2014; Van der Elst, Van Boxtel, Van Breukelen, & Jolles, 2005). Multiple theoretical explanations have been proposed for these observed auditory learning benefits. Firstly, it has been argued that learning from spoken input is more efficient as a result of such mechanisms being developmentally and/or evolutionarily older than those operating on written stimuli (Bakker et al., 2014).

Further, evidence suggests that, relative to the visual modality, in the auditory modality stronger associations develop between sequential events (Penney, 1989) and/or that temporal events are more accurately stored (Glenberg & Jona, 1991). Auditory cortices have been suggested to be more sensitive to sequencing information, due to the sequential nature of auditory information (Frost, Armstrong, Siegelman & Christiansen, 2015).

Cognitive load theory (Sweller, Van Merrienboer & Paas, 1998) also predicts a spoken learning benefit when learning word forms and visual meanings (e.g. a picture or graph) in combination. It argues that cognitive overload is less likely under conditions in which information processing can be divide between the visuo-spatial sketchpad and phonological loop (Baddeley, 1992), compared to conditions in which all information must be processed within the same modality and thus by the same cognitive resources.

In contrast to the above, a written advantage has also been observed particularly when word forms are learned in conjunction with their meanings, (Balass, Nelson, & Perfetti, 2010; Nelson, McEvoy, & Schreiber, 2004; Van der Ven, Takashima, Segers, & Verhoeven, 2015). Multiple theories have also been proposed in explanation for these findings. It is argued that when reading (novel words) phonological representations are automatically activated alongside orthographical representations, therefore, two separate representations of the word form are stored. However, on exposure to the spoken word form, automatic activation of its

orthographic form is less likely (Perfetti, Bell & Delaney, 1988; Paivio, 1991). Further, the spoken modality is fleeting by nature, posing additional demands on attention and working memory capacity. Reading allows rereading and processing at one's own pace and this flexibility leads to greater availability of memory and attentional cognitive resources for learning (Van der Ven, 2015).

Alternatively, in contrast to the above findings it remains possible that learning mechanisms operating on written and spoken stimuli are equally efficient and instead observed contradictory effects result from modality specific biases in the experimental design. Although typically, prior to literacy, word learning is only possible via the auditory modality, it is feasible that proficient readers develop learning mechanisms that overcome modality specific constraints such that learning occurs equally effectively in both modalities.

Previous studies, that have reported modality effects, have potentially generated contradictory findings due to an absence of one or more of the following controls. First, exposure duration was not controlled in studies that found a written learning advantage (Balass et al., 2010; Nelson et al., 2006; Van der Ven et al., 2015). People were given unlimited time with the spoken and written materials, but the exact exposure time was not measured. Participants thus might have exposed themselves more to materials in one modality, evoking a learning effect that does not result from a more efficient modality specific learning mechanism but simply due to a mechanism having greater exposure to the stimulus.

Second, in all studies that found a written benefit the test was presented in a written form (Balass et al., 2010; Nelson et al., 2006; Van der Ven et al., 2015); likewise, some studies that found a spoken benefit performed only a spoken test (Van der Elst et al., 2005). According to Tulving and Thomas's (1973) encoding specificity principle, recall is enhanced if the conditions during retrieval match the conditions during learning. Thus, such modality effects observed in these studies might be evoked by encoding specificity rather than by differences in the efficiency of the spoken and written learning mechanisms. Similarly, studies examining learning of word form-meaning associations only used written meanings. Thus, the congruency of the format between written word forms and written word meanings potentially benefits learning in the written modality.

Fourth, many previous studies have used explicit learning tasks (Balass et al., 2010; Nelson et al., 2006; Van der Ven et al., 2015). Therefore, in such studies, it is difficult to exclude the possibility that observed modality effects do not result from modality-specific conscious learning strategies, such as repeating heard words or rereading written words, rather than differences in the efficiency of modality specific cognitive mechanisms.

Finally, many previous studies (Bakker et al., 2014; Balass et al., 2010; Nelson et al., 2006; Van der Ven et al., 2015; Van der Elst et al., 2005) do not control for cross-modal orthographical and phonological transparency. Therefore, any learning benefit observed may not result from differences in the efficiency of learning mechanisms but instead may result from it being easier to accurately transform the phonological form to the orthographic or vice versa.

In order to gain an understanding of modality effects on word learning it is first necessary to control for each of these potential confounds. The present study aims to do precisely this, controlling for the many confounds that have potentially generated observed modality effects that do not result from difference in efficiency of the spoken and written learning mechanisms.

In two experiments, participants learned 24 Dutch-like, fully transparent pseudowords and pictorial meanings. After a short period of consolidation, participants were tested on their knowledge of the learned word forms and meanings. A between-subjects design generated four participant groups per experiment 1) written training, written test; 2) written training, spoken test; 3) spoken training, written test; 4) spoken training, spoken test. In addition, non-verbal IQ, vocabulary and reading tasks were administered to control for differences across groups. In Experiment 1 written word forms were presented for a time period equal to the duration of the spoken form. In Experiment 2, to control for the fact that a written word can be read quicker than its spoken form takes to unfold, the written stimulus was presented only for the period necessary to read the written stimulus.

Experiment 1

Methods

Participants 60 participants (M = 22.96 years, SD = 2.53; 46 female) were recruited. All participants were right-handed, with no language, sight or hearing disorders. Participants earned €10 for participating.

Design The two between-subjects factors were modality during training and modality during testing. Words could be learned in either modality and also testing could occur in two modalities. There were therefore four between-subjects conditions. Participants were randomly assigned to a condition.

Materials Twenty-four orthographically and phonologically transparent Dutch pseudowords were created using Wuggy (Keuleers & Brysbaert, 2010). The words had a Levenshtein's Distance (Levenshtein, 1966) of above three to avoid confusability. Pilot studies ensured the words were not reminiscent of existing Dutch words. The words varied between five and nine letters and four and eight phonemes and graphemes. Speech duration of the words varied between 664 and 993 ms.

In addition twenty-four pictures of unknown objects from The Novel Object and Unusual Name (NOUN) Database were used (Horst & Hout, 2016). The pictures were not visually similar to each other. To limit item-specific effects, for each group of four participants the pictures were randomly assigned to one of the word forms.

Procedure Participants were trained and tested on the same day. First the training phase was administered. The experiment was designed to minimize opportunities for participants to utilize explicit learning strategies. For this reason no explicit instruction to learn the picture–word form pairs or indication of a later test was provided, images and word forms were presented briefly and in rapid succession, and both auditory and visual masks immediately followed presentation of the word form. In each trial (Figure 1), participants saw a fixation cross (250 ms), a picture (1000 ms), then again saw a fixation cross (250 ms), either heard the word or read the word depending on the condition, and then heard a auditory mask in the form of a continuous tone and saw a visual mask in the form of a grey diamond (500 ms). The exposure to the word form varied for each word: the written word was presented for the speech duration of that specific word (M = 863 ms, SD = 97 ms). The next word in the training sequence always had a Levenshtein's distance above three and a different onset. Each training trial was repeated seven times in a blocked, semi-randomized order. To ensure attention during the training phase, eight pictures of familiar known objects (e.g. a bus) were shown in-between the trials and participants had to press a button as soon as they saw one of these familiar objects. Participants were instructed to pay attention to the pictures and words and press a button if they saw one of the eight familiar objects, but critically were not explicitly told to learn the word form – picture pairs.

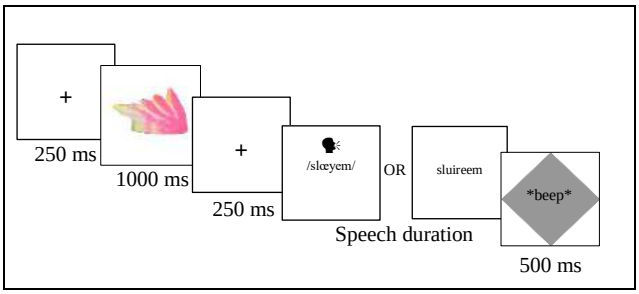


Figure 1: Experimental procedure of a training trial

The training phase was followed by a filler task. This purely visual, nonverbal IQ-task lasted for 20 minutes (Raven's progressive matrices, 1965). Then, in the test phase, participants performed a subsequent matching task. Participants saw a fixation cross (250 ms), a picture (1000 ms), then again a fixation cross (250 ms), heard or saw the word depending on the condition, and had to decide within 2 seconds whether the picture and word matched what they had learned by using a button box. The written words were again presented for a time period equal to the speech duration of that particular word. Each word was presented twice: once with the correct picture and once with a foil picture (i.e., a different picture presented in the training phase). There were several constraints regarding the relationship between the foil picture and the target word form. The corresponding learned word form of the foil picture did not share the onset of the target word form and possessed a Levenshtein's distance of above four. Regarding the order of the trials, the corresponding word form of the next (foil) picture could not be one of the previous ten word forms. Also, half of the target words were first shown with the correct picture before they appeared with a foil picture and vice versa. Participant's ability to identify both matching and mismatch picture-word form pairs was recorded. Then, several individual difference tests were administered, including word reading, pseudoword reading (Van den Bos, Spelberg, Scheepsma & de Vries, 1994; Brus & Voeten, 1973) and vocabulary (Dunn, Dunn, & Schlichting, 2005).

Results

Violin plots depicting, per condition, the proportion of picture–word form pairs that were correctly identified as a match or mismatch can be found in Figure 2.

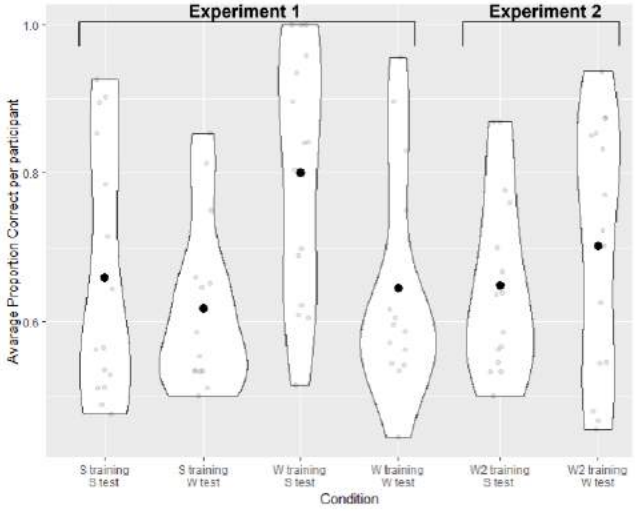


Figure 2: Proportion of correctly identified matching and mismatching picture-word form pairings per participant

A mixed effects logistic regression model (lme4 package: Bates, Maechler, Bolker, & Walker, 2015) using R (R Development Core Team, 2008) was constructed with response on test (match or mismatch) as the dependent variable, i.e. whether a participant recorded the corresponding image and word form pair as matching or mismatching. Model structure was compatible with the conventions of standard signal detection analysis and was consistent with current best practice (e.g. Jacobs, Dell, Benjamin, & Bannard, 2016; Zormpa, Brehm, Hoedemaker & Meyer, 2019). The model included fixed effects of trial type (whether the trial was a match or mismatch), training modality (written or spoken) and test modality (written or spoken), in addition to their interactions. The full random effect structure was also included in the model with random intercepts and slopes by item for trial type, training modality and test modality, and random intercepts and slopes by participant for trial type.

Model results revealed a main effect of trial type, showing participants displayed sensitivity to trained versus untrained picture–word pairs, providing a match more frequently when presented with the picture – word pairs on which they were trained (estimate = -1.03, *SE* = 0.16, *z* = -6.35, *p* < .001). The interaction between trial type and training modality was also significant (estimate = 0.29, *SE* = 0.13, *z* = 2.20, *p* = .03) with participants in trained on written word forms displaying greater sensitivity in identification of trained vs. untrained picture – word for pairs. Finally, a significant interaction between trial type and test modality was also observed (estimate = -0.37, *SE* = 0.13, *z* = -2.78, *p* = 0.006) with participants displaying greater sensitivity when tested on spoken word forms.

Conclusion

Experiment 1 results show that when controlling for exposure time by providing equal exposure duration in both modalities, learning from written materials is greater. One explanation for this might be differences between modalities in the speed with which the full word form can be accessed from the stimulus. The speech duration of the word forms was between 664 and 993 ms, and thus, the written words were presented for a duration of between 664 and 993 ms, depending on the word. First pass single word reading is however much faster than the reading time provided in Experiment 1. Literature using lexical decision or naming tasks show that bisyllabic word can be read at between 525-610 ms, and pseudowords between 575 and 650 ms (Brunswick, McCrory, Price, Frith, & Frith, 1999; De Groot & Nas, 1991; Schilling, Rayner, & Chumbley, 1998; Weekes, 1997). However, these estimates include time necessary to make a decision and speech planning. Studies using ERP and eye-gaze measures, which give a more accurate estimate of reading times, show that frequent, known words can be read around 150 ms and infrequent words within 200-250 ms (Rayner, Pollatsek, Ashby, & Clifton Jr, 2012; Schilling et al., 1998; Sereno, Rayner, & Posner, 1998). This means that, although exposure time to the written and spoken stimuli was equal in Experiment 1, people had more time with the full word form in the written condition.

Experiment 2 tested whether the modality effects found in Experiment 1 would hold if exposure to written and spoken materials was equivalent, taking into account that written information is presented instantaneously and that reading is faster than listening to speech. Literature has shown that people need slightly longer to read infrequent words (200-250 ms) than frequent words (150 ms). Pseudowords are thus likely to be read slightly slower. Therefore, in Experiment 2, the written exposure time was set at 300 ms for all 24 words, which is a written exposure time reduction of 65% on average relative to Experiment 1.

Experiment 2

Methods

Participants 30 participants (*M* = 23.02 years, *SD* = 2.40; 26 female), all right-handed, with no language, sight or hearing disorders participated in this experiment. Participants earned €10,- for participating.

Design Experiment 2 only concerned written modality learning. Testing occurred in both modalities, creating two conditions. Participants were randomly assigned to a condition.

Materials The materials were the same as in Experiment 1.

Procedure The procedure was similar to that of Experiment 1, except for the training phase. In the training phase, the written word was now presented for 300 ms rather than the speech duration of that specific word. This reduced the total duration of the training phase by 560 ms. To ensure that this shortening of the trial did not affect learning, in each trial the first fixation cross was elongated from 250 to 530 ms and the mask at the end of a trial was elongated from 500 to 780 ms. After training participants again performed a non-verbal IQ test, followed by the picture-word form matching task and the individual difference measures.

In addition, to test that 300 ms was sufficient time for participants to read the word-forms, a simple retyping task was added to test whether participants could read 120 additional Dutch pseudowords equally well when presented for either 300 ms or 860 ms (the mean written exposure time of Experiment 1). This retyping task was only administered to the participants in the written training condition.

Results

One participant from Experiment 2 had to be removed, because no buttons were pressed during the matching task. Violin plots of the accuracy data can be found in Figure 2. Four one-way ANOVA's indicated that the six groups (four from Experiment 1 and two from Experiment 2) did not differ regarding average general IQ ($F(5,83) = 0.46$, $p = .81$), vocabulary ($F(5,83) = 0.64$, $p = .67$), word reading ($F(5,83) = 0.69$, $p = .63$) or pseudoword reading ability ($F(5,83) = 0.67$, $p = .65$).

To analyse performance on the retyping task, a frequentist mixed-effect logistic regression model was applied using R package lmer (lme4 package: Bates, Maechler, Bolker, & Walker, 2015) with retyping accuracy as dependent variable, and word length and exposure time (300 or 860 ms) as independent variables, plus a random intercept by participant and word. This analysis showed no difference in accuracy of retyping after a 300 or 860 ms exposure (estimate = 0.54 *SE* = 1.81, *z* = 0.29, *p* = .77).

The mixed-effects logistic regression model used to analyse results in Experiment 1, was extended to analyse results of both experiments, with modality at training now possessing three levels: spoken training in Experiment 1, written training

in Experiment 1 where written exposure time was equal to spoken exposure time, and written training in Experiment 2 where written exposure time was reduced to 300 ms. The bias effects of modality at training and test on hits and false alarms are illustrated in Figure 2.

Analyses revealed a significant main effect of trial type with participants more likely to produce a match response when trials included the picture – word form pairs on which they were trained (estimate = -0.73, *SE* = 0.19, *z* = -3.89, *p* < 0.001). The interaction between trial type and training modality was not significant when comparing the reduced written training condition (Experiment 2) to that of the spoken training condition (estimate = -0.24, *SE* = 0.25, *z* = -0.96, *p* = 0.34) indicating that sensitivity of participants did not differ significantly between groups. Similarly, the interaction between trial type and training modality was not significant when comparing the reduced written training condition (Experiment 2) to the longer written training condition (Experiment 1) (estimate = -0.38, *SE* = 0.25, *z* = -1.52, *p* = 0.13). The three-way interaction between training modality, test modality and trial type was significant when comparing the two written conditions (estimate = -0.73, *SE* = 0.25, *z* = -2.90, *p* = 0.004). The three-way interaction was not significant (estimate = -0.43, *SE* = 0.25, *z* = -1.69, *p* = 0.09) when comparing the longer written training condition (Experiment 1) to the spoken training condition or the reduced written training condition (Experiment 2) to the spoken training condition (estimate = 0.31, *SE* = 0.25, *z* = 1.25, *p* = 0.21). Thus, participants trained in the longer written condition (Experiment 1) displayed greater sensitivity during the spoken test than participants trained in the shorter written training condition (Experiment 2) or spoken training condition.

Conclusion

Experiment 2 aimed to investigate whether the written modality benefit found in Experiment 1 resulted from participants having more time with the word form in the written condition, due to the fact that it takes longer for a spoken word to unfold than to read its written from. By reducing written word exposure to 300 ms per word, we controlled for this inherent advantage of the written modality. Results showed that when the exposure time to the written materials was reduced, learning in the written condition did not differ from that in the spoken condition. Further, this was not a result of participants having insufficient time to read the written form as participants did not differ in their ability to retype written pseudowords when they were presented for 300 ms or 860 ms.

General Discussion

This study aimed to test whether modality specific learning mechanisms, engaged when learning novel picture–pseudoword form pairings, are more effective when words are presented in their written or spoken form. This study is the first to test for such effects of modality while controlling for the following factors, which potentially give rise to modality effects independent of differences in the efficiency of modality specific learning mechanisms: 1) differences in orthographic and phonological transparency, 2) congruence in modality of word form and word meaning, 3) duration of exposure, 4) engagement of explicit learning strategies, 5) congruence in modality of training and modality of test.

Our results showed that when the duration of written and spoken exposure is equal (the written stimulus is presented for a time period equal to the duration of the spoken word), participants' accuracy in identifying picture-word form pairs is greater when trained on written word-forms. This finding replicates earlier findings of a written learning benefit when learning word forms and their meanings (Balass et al., 2010, Nelson et al., 2004, Van der Ven et al., 2015).

However, Experiment 2 shows that the written learning benefit disappears when controlling for the fact that the time required to read a word in its written form is shorter than the time required for its spoken form to unfold. Our results demonstrate that once controlling for this property of reading there is no additional advantage in learning word form – picture associations when words are presented in their written rather than spoken form.

Our conclusions are therefore at odds with previous studies that argue for differences in the efficiency of modality specific learning mechanisms. Based on the results produced by this study we believe such findings are likely driven by an absence of one or more of the confounds listed above (see list 1-5), which alone may generate such observed modality effects.

Bakker et al, (2014), one of few studies to train and test participants in both modalities, provides evidence that auditory benefits of learning novel word forms emerge only at longer periods of consolidation. Within their study phoneme and letter monitoring tasks were used to probe lexical integration of novel word forms after 24 hrs and 8 days. It is feasible therefore that the findings within the current study are limited to short-term episodic memory. This can be tested in a follow up study by extending the current paradigm to include tests of lexical integration at longer periods of consolidation.

Unexpectedly, our results did not produce a modality congruency effect as predicted by Tulving and Thomas's (1973) encoding specificity principle, in that the experimental groups for which the test modality was the same as the training modality, did not show superior performance. Paradoxically, the written benefit observed in Experiment 1 was mainly driven by the written learning spoken test group. However, we believe this to be caused by the perceived erratic response window in the written test condition. Participants were required to respond within 2 seconds plus the speech duration of the written word. Because they did not hear the word, the response time was therefore difficult to predict. This conclusion is supported by participant's performance on the same task in Experiment 2, when participants were habituated in the written training phase to a fixed exposure time which did not appear to result in a decrease in performance on the written test.

Still, our experiments do not provide evidence for Tulving and Thomas's (1973) encoding specificity principle, since participants in all cross-modal conditions were consistently able to recognize words in a modality in which they had not seen the word form before. Further, no interaction was observed between the written reduced training condition of Experiment 2 and the spoken test condition of Experiment 1, indicating that when participants have equivalent exposure to either the written or spoken word form in training, their ability to recognise the novel word form in the alternative, unseen modality does not differ.

Within the current study, attempts were made to limit strategic cross-modal encoding: no explicit instruction to learn the materials was provided, participants were trained in a single modality, stimuli were presented rapidly, and visual and auditory masks immediately followed the presentation of the word form. Thus, our results suggest that proficient readers, such as those tested in our study, automatically rapidly recode novel word forms into both their phonological form when presented with written stimulus (Perfetti et al., 1988) and their orthographic form when presented with an auditory stimulus.

Our findings also do not support a developmental and/or evolutionary advantage for learning from spoken materials. It appears that even though the ability to learn from written materials has developed later in human's lives and their evolution as a species, this ability is sufficiently developed in adult proficient readers to perform equally effectively.

This study set out to test for modality effects on novel word learning. Specifically it tested for differences in the efficiency of modality specific mechanisms engaged when learning novel object - pseudoword pairs, from either spoken or written stimuli. Results showed a written benefit when equal exposure time was provided. However, once we controlled for the fact that reading allows faster access to the full word form than listening to speech, no modality effect was observed. This suggests that modality specific learning mechanisms operating on spoken or written stimuli were equally efficient. Given that we typically begin learning words from auditory input only, the findings of the present study indicate that once we become proficient readers, the cognitive system converges on learning equally efficiently from both modalities.

References

Bakker, I., Takashima, A., Van Hell, J. G., Janzen, G., & McQueen, J. M. (2014). Competition from unseen or unheard novel words: Lexical consolidation across modalities. *Journal of Memory and Language, 73*, 116-130.

Balass, M., Nelson, J. R., & Perfetti, C. A. (2010). Word learning: An ERP investigation of word experience effects on recognition and word processing. Contemporary *Educational Psychology, 35*(2), 126-140.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1-48.

Brunswick, N., McCrory, E., Price, C. J., Frith, C. D., & Frith, U. (1999). Explicit and implicit processing of words and pseudowords by adult developmental dyslexics: A search for Wernicke's Wortschatz? *Brain, 122*(10), 1901-1917.

Brus, B. T., & Voeten, M. (1973). Een-minuut test. Vorm A en B. *Verantwoording en handleiding*.

De Groot, A. M. B., & Nas, G. L. J. (1991). Lexical representation of cognates and noncognates in compound bilinguals. *Journal of Memory and Language, 30*(1), 90-123.

Dunn, L. M., Dunn, L. M., & Schlichting, J. E. P. T. (2005). *Peabody picture vocabulary test-III-NL*. Amsterdam, Netherlands: Harcourt Test Publishers.

Frost, R., Armstrong, B. C., Siegelman, N., & Christiansen, M. H. (2015). Domain generality versus modality specificity: the paradox of statistical learning. *Trends Cognitive Sciences, 19*(3), 117-125.

Glenberg, A. M., & Jona, M. (1991). Temporal coding in rhythm tasks revealed by modality effects. *Memory & Cognition, 19*(5), 514-522.

Horst, J. S., & Hout, M. C. (2016). The Novel Object and Unusual Name (NOUN) Database: A collection of novel images for use in experimental research. *Behavior Research Methods, 48*(4), 1393-1409.

Jacobs, C. L., Dell, G. S., Benjamin, A. S., & Bannard, C. (2016). Part and whole linguistic experience affect recognition memory for multiword sequences. *Journal of Memory and Language, 87, 38-58.*

Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior Research Methods, 42*(3), 627-633.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady, 10*(8), 707-710.

Nelson, J. R., Balass, M., & Perfetti, C. A. (2006). Differences between written and spoken input in learning new words. *Written Language & Literacy, 8*(2), 25-44.

Paivio, A. (1991). Dual Coding Theory - Retrospect and Current Status. *Canadian Journal of Psychology-Revue Canadienne De Psychologie, 45*(3), 255-287.

Penney, C. G. (1989). Modality effects and the structure of short-term verbal memory. *Memory & Cognition, 17*(4), 398-422.

Perfetti, C. A., Bell, L. C., & Delaney, S. M. (1988). Automatic (prelexical) phonetic activation in silent word reading: Evidence from backward masking. *Journal of Memory and Language, 27*(1), 59-70.

R Development Core Team. (2008). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing.

Raven, J. (1965). *Advanced Progressive Matrices. Sets I and II.* London: HK Lewis & Co.

Rayner, K., Pollatsek, A., Ashby, J., & Clifton Jr, C. (2012). *Psychology of Reading, 2nd ed.* New York, NY, US: Psychology Press.

Schilling, H. H., Rayner, K., & Chumbley, J. I. (1998). Comparing naming, lexical decision, and eye fixation times: word frequency effects and individual differences. *Memory & Cognition, 26*(6), 1270-1281.

Sereno, S. C., Rayner, K., & Posner, M. I. (1998). Establishing a time-line of word recognition: evidence from eye movements and event-related potentials. *Neuroreport, 9*(10), 2195-2200.

Sweller, J., van Merrienboer, J. J. G., & Paas, F. G. W. C. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*(3), 251-296.

Tulving, E., & Thomson, D. M. (1973). Encoding Specificity and Retrieval Processes in Episodic Memory. *Psychological Review, 80*(5), 352-373.

Van den Bos, K., Spelberg, H., Scheepsma, A., & De Vries, J. (1994). *De Klepel. Vorm A en B. Een test voor de leesvaardigheid van pseudowoorden. Verantwoording, handleiding, diagnostiek en behandeling.* Nijmegen: Berkhout.

Van der Elst, W., Van Boxtel, M. P., Van Breukelen, G. J., & Jolles, J. (2005). Rey's verbal learning test: normative data for 1855 healthy participants aged 24-81 years and the influence of age, sex, education, and mode of presentation. *Journal of the International Neuropsychological Society, 11*(3), 290-302.

Van der Ven, F., Takashima, A., Segers, E., & Verhoeven, L. (2015). Learning word meanings: overnight integration and study modality effects. *PLoS One, 10*(5), e0124926.

Weekes, B. S. (1997). Differential Effects of Number of Letters on Word and Nonword Naming Latency. *The Quarterly Journal of Experimental Psychology Section A, 50*(2), 439-456.

Zormpa, E., Brehm, L. E., Hoedemaker, R. S., & Meyer, A. S. (2019). The production effect and the generation effect improve memory in picture naming. *Memory, 27*(3), 340-352.

Under pressure: The influence of time limits on human exploration

Charley M. Wu[1,*] **(cwu@mpib-berlin.mpg.de), Eric Schulz**[2,*],
Kimberly Gerbaulet[1,3,*], **Timothy J. Pleskac**[1,4], & **Maarten Speekenbrink**[5]
[1]Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin, Germany
[2]Department of Psychology, Harvard University, Cambridge, Massachusetts, USA
[3]Institute of Cognitive Science, University of Osnabrück, Osnabrück, Germany
[4]Department of Psychology, University of Kansas, Lawrence, Kansas
[5]Department of Experiment Psychology, University College London, London, UK
[*]Contributed equally to this work.

Abstract

How does time pressure influence attitudes towards uncertainty? When time is limited, do people engage in different exploration strategies? We study human exploration in a range of four-armed bandit tasks with different reward distributions and manipulate the available time for each decision (limited vs. unlimited). Through multiple behavioral and model-based analyses, we show that reactions towards uncertainty are influenced by time pressure. Specifically, participants seek out uncertain options when time is unlimited, but avoid uncertainty under time pressure. Moreover, larger relative differences in uncertainty between options slowed down reaction times and dampened the drift rate of a linear ballistic accumulator model. These results shed new light on the differential effect of uncertainty and time pressure on human exploration.

Keywords: Exploration-exploitation; Uncertainty; Time Pressure; Directed Exploration; Multi-armed Bandits

Introduction

Searching for rewards requires navigating the exploration-exploitation dilemma: Should one exploit options known to produce high rewards, or explore lesser known options to gain information that could potentially lead to even higher rewards? Because optimal solutions (Gittins, 1979) are generally intractable in realistic settings, practical solutions usually rely on heuristics (Auer, Cesa-Bianchi, & Fischer, 2002), which can be classified as directed exploration, random exploration, or both.

Directed exploration is often implemented using an exploration bonus that inflates the expected value of an option proportional to the estimated uncertainty, to encourage the exploration of uncertain options. Whereas earlier studies produced mixed evidence for the use of exploration bonuses in human reinforcement learning (Daw, O'doherty, Dayan, Seymour, & Dolan, 2006), there is now an increasing amount of evidence for directed exploration in vast problem spaces (Wu, Schulz, Speekenbrink, Nelson, & Meder, 2018), planning (Wilson, Geana, White, Ludvig, & Cohen, 2014), dynamic decision making (Knox, Otto, Stone, & Love, 2012), and simple two-armed bandit tasks (Gershman, 2018).

Unlike directed exploration, *random exploration* increases choice stochasticity in accordance to the agent's uncertainty about the value of available actions (Speekenbrink & Konstantinidis, 2015). One recent theory proposed that random and directed exploration can be dissociated, where the balance is influenced by the total and relative uncertainty of available options (Gershman, in press). If there are multiple options with similar expected rewards, directed exploration makes an option more likely to be sampled when its uncertainty is higher relative to the other options (Schulz & Gershman, 2019). We make use of this effect by studying how patterns of decision making and exploration are affected by both uncertainty and expected reward in a four-armed bandit task. Compared to previously studied two-armed bandit tasks, the richer set of options makes exploration more pertinent and observable over more trials. Crucially, we manipulate the presence or absence of time pressure to gain insights into the cognitive processes underlying exploration. If directed exploration is a reasoned and controlled process, which requires taking the uncertainties of each options into account, then time pressure may limit the capacity for directed exploration.

As predicted, we find that participants are more likely to sample options with high relative uncertainty in the absence of time pressure. However, when we impose time pressure by limiting the allowed decision time to under 400 milliseconds, we find that relative uncertainty reduces the probability that an option is chosen. Additionally, relative uncertainty slows down reaction times more strongly and dampens the evidence accumulation process more heavily under time pressure. In other words, time pressure moderates the effect of environmental uncertainty, such that risk-seeking behavior arising through directed exploration transforms into risk-aversion under time pressure. These results enrich our understanding of human exploration strategies under changing task demands.

Experiment

Participants and Design. We recruited 99 participants (36 female, aged between 21 and 69 years; M=34.82; SD=10.1) on Amazon Mechanical Turk (requiring 95% approval rate and 100 previously approved HITs). Participants were paid \$3.00 for taking part in the experiment and a performance contingent bonus of up to \$4.00 (calculated based on the performance of one randomly selected round). Participants spent 13.0 ± 5.6 minutes on the task and earned $\$5.87 \pm \$0.91$ in total. The study was approved by the Ethics Committee of the Max Planck Institute for Human Development.

We used a 2×4 within-subject design to examine how the presence or absence of time pressure and the payoff structure of the task (see Fig. 1b and Tab. 1) influenced choices and reaction times. In total, the experiment consisted of 40 rounds

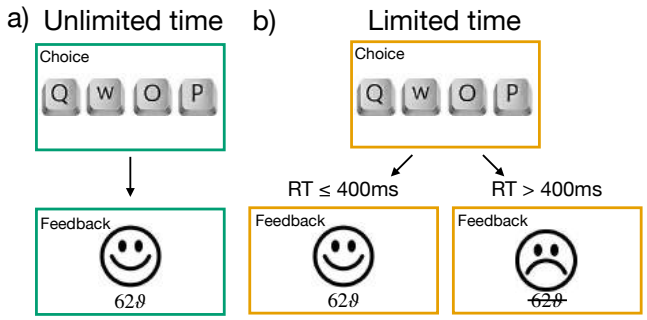


Figure 1: **Experimental design.** We used a four-armed bandit task where each option was randomly mapped to the Q, W, O, and P keys on the keyboard. **a**) In unlimited time rounds, participants could take as long as they wanted to make each selection and received positive feedback (happy face) and were shown the value of the acquired payoff. **b**) In limited time rounds, participants were only given 400 ms to make each selection. If they exceeded the time limit, they would forgo earning any rewards, and received negative feedback (sad face) along with the value of the payoff they could have earned (crossed out).

with 20 trials each. In each round, a condition was sampled (without replacement) from a pre-randomized list, such that each combination of time pressure and payoff structure was repeated five times, with a total of 100 trials in each.

Materials and Procedure. Participants were required to complete three comprehension questions and two practice rounds (one with unlimited time and one with limited time) consisting of 5 trials each before starting the experiment. Each of the 40 rounds was presented as a four-armed bandit task, where the four options were randomly mapped to the $[Q,W,O,P]$ keys on the keyboard (Fig. 1). Selecting an option by pressing the corresponding key yielded a reward sampled from a normal distribution, where the mean and variance was defined by the round's payoff structure (Fig. 2a and Tab. 1). Participants completed 20 trials in each round and were told to acquire as many points as possible.

Before starting a round, participants were informed whether it was an unlimited or a limited time round. In unlimited time rounds, participants could spend as much time as they needed to reach a decision, upon which they were given feedback about the obtained reward (displayed for 400 ms) before continuing to the next trial (Fig. 1a). In limited time rounds, participants were instructed to decide as fast as possible. If a decision took longer than 400 ms, they forfeited the reward they would have earned (presented to them as a crossed-out number with an additional sad smiley; Fig. 1b). We used the same inter-trial period of 400 ms to display feedback about obtained rewards in both limited and unlimited time rounds.

We applied a random shifting of rewards across rounds (i.e., different maximum reward) to prevent participants from immediately recognizing when they had chosen the optimal option. For each round, we sampled a value from a uniform distribution $\mathcal{U}(30,60)$, which was then added to the rewards. Together with random shifting, we also truncated rewards such that they were always larger than zero. In order to convey intuitions about the random shift of rewards, payoffs were presented using a different fictional currency in each round (e.g., ß, Þ, ϑ), such that the absolute value was unknown, but higher were always better.

Table 1: Payoff Conditions

Payoff Conds	Means (μ)	Variances (σ^2)
IGT	$[-10,-10,10,10]$	$[10,100,10,100]$
Low Var	$[-10,-\frac{1}{3},\frac{1}{3},10]$	$[10,10,10,10]$
High Var	$[-10,-\frac{1}{3},\frac{1}{3},10]$	$[100,100,100,100]$
Equal Means	$[0,0,0,0]$	$[10,40,70,100]$

At the end of each round, participants were given feedback about their performance in terms of the bonus they would gain (in USD) if this was the round selected for determining the bonus. The bonus was calculated as a percentage of the total possible performance, raised to the power of 4 to accentuate differences in the upper range of performance:

$$\text{Bonus} = \left(\frac{\text{total reward gained}}{\text{mean reward of best option} \times 20 \text{ trials}}\right)^4 \times \$4.00$$

Payoff conditions We used four different payoff conditions as a within-participant manipulation (Tab. 1 and Fig. 2a). Each payoff condition specified the mean μ_i and variance σ_i^2 of the reward distribution $R_i \sim \mathcal{N}(\mu_i,\sigma_i^2)$ for each option i. Each distribution was randomly mapped to one of the four $[Q,W,O,P]$ keys of the keyboard in each round. The Iowa Gambling Task (IGT) is a classic design that has been related to a variety of clinical and neurological factors affecting decision-making (Yechiam, Busemeyer, Stout, & Bechara, 2005; Bechara, Damasio, Damasio, & Anderson, 1994). We implemented a reward condition inspired by the IGT such that there are two high and two low reward options, with a low and high variance version of each. We also constructed two conditions with equally spaced means, but with either uniformly low variance or uniformly high variance. Lastly, the equal means condition had identical means and gradually increasing variance, such that we can observe the influence of uncertainty independent of mean reward.

Behavioral Results

Participants acquired higher rewards in the unlimited than in the limited time condition (Fig. 2b; $t(98) = 3.1$, $p = .002$, $d = 0.3$, $BF = 10$). Participants also improved over trials, signified by an average correlation between trial and rewards (Spearman's $\rho(98) = 0.16$, $p < .001$, $BF > 100$). This correlation did not differ between limited and unlimited time rounds ($t(98) = -1.3$, $p = .196$, $d = 0.1$, $BF = .25$).

We also compared performance across payoff conditions. This is possible, since all games had the same expected reward under the assumption of a random sampling strategy. We found that participants performed better in the IGT-like condition than in the low variance condition ($t(98) = 3.2$, $p = .002$, $d = 0.3$, $BF = 14$). We see an even larger difference when comparing the low variance and high variance conditions, which had the same means but different levels of

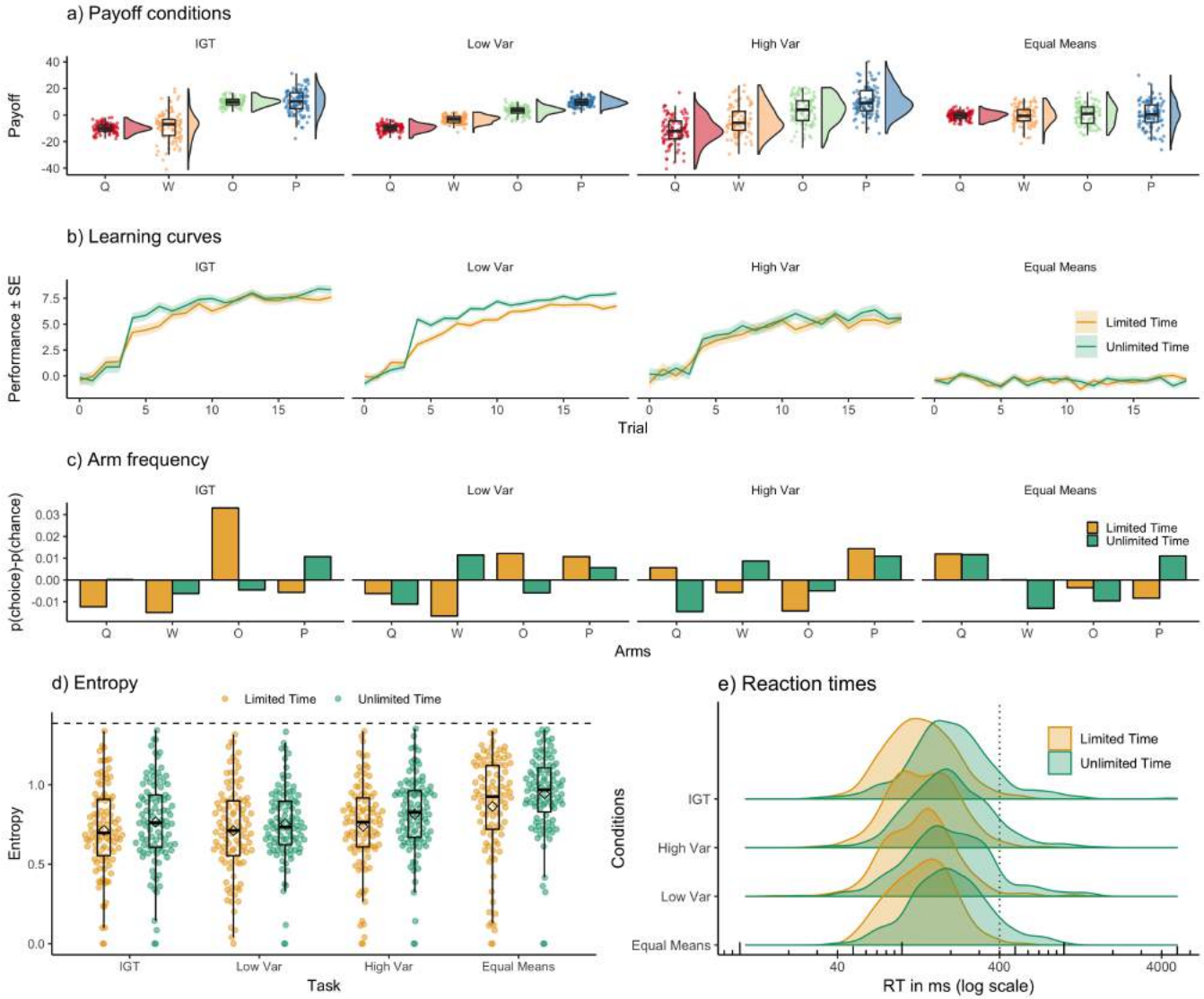


Figure 2: **Payoff conditions and behavioral results**. **a**) Four different payoff conditions were combined with either limited or unlimited time rounds to create 8 different scenarios. Each condition specifies a normal payoff distribution for each option; the means and variances are shown in Table 1. Each dot represents a randomly drawn payoff, while the Tukey boxplots and half violin plots show the distribution for 100 simulated draws. Note that rewards were randomly shifted in each round by adding a constant $\sim \mathcal{U}(30,60)$ to all payoffs. **b**) Learning curves of average participant performance (using unshifted rewards) over trials by payoff condition. Ribbons indicate standard error. **c**) Choice proportions (normalized for chance) for each option, mapped to the canonical ordering shown in panel a). **d**) The entropy of choices in each round, where higher entropy corresponds to more diverse choices and the dotted line indicates random chance (i.e., playing each arm with equal probability). Each dot represents a participant, and overlaid are Tukey boxplots with the diamond indicating the group mean. **e**) Distributions of reaction times in milliseconds (ms) and shown on a log scale. The vertical dotted line indicates the time limit (400 ms) of the limited time condition

risk and uncertainty. Participants performed substantially better in the low variance condition than the high variance condition ($t(98) = 6.2$, $p < .001$, $d = 0.6$, $BF > 100$). Thus, higher variance increased the difficulty of the task. Lastly, participants performed better in the high variance than in the equal means task ($t(98) = 25.5$, $p < .001$, $d = 2.6$, $BF > 100$), which is intuitive since improvement is not possible if all arms have the same mean reward.

Choice proportions. Figure 2c shows the proportion of choices, which illustrates differences across time conditions. We used a Bayesian mixed-effects logistic regression and found that in the IGT condition, participants chose the high reward-low variance option (indicated as 'O' in Fig. 2c) less frequently in the unlimited time than in the limited time condition ($\hat{\beta} = -.22$, 95% Highest Posterior Density (HPD) interval: $[-.28, -.15]$, $BF > 100$)[1].

Additionally, we also find differences across time-pressure conditions in the Equal Means task, where participants selected the highest variance option ('P') more frequently in the unlimited time condition $\hat{\beta} = .11$, 95% HPD: $[.05, .17]$,

[1]We use Bridge sampling (Gronau, Singmann, & Wagenmakers, 2017) to approximate the Bayes Factor by comparing against an intercept-only null model (i.e., without time pressure as a predictor).

$BF = 15$). This illustrates a shift in preferences away from uncertain options when time pressure is introduced. Whereas participants tend to be risk-seeking and choose highly uncertain options under unlimited time, they become more risk-averse and choose them less often under time pressure.

We also calculated the Shannon entropy of participants' choices in each round (Fig. 2d), where higher entropy corresponds to higher diversity of choices and the maximal entropy strategy would be to choose each option an equal number of times (indicated by the dotted line). Averaged across participants, we find higher choice entropy (i.e. more diversity in choice) under unlimited time than limited time ($t(98) = 4.1$, $p < .001$, $d = 0.4$, $BF > 100$). This further strengthens the evidence for reduced exploration under time pressure, since we find a lower diversity of choices.

Reaction times. Figure 2d shows reaction times. Unsurprisingly, participants responded faster in the limited time than in the unlimited time conditions (comparing RTs in logms: $t(98) = 9.7$, $p < .001$, $d = 1.0$, $BF > 100$). There were no differences across payoff conditions ($F(3,95) = 0.12$, $p = .951$, $BF = 0.01$).

Model-Based Analyses

In order to model learning and decision making in our task, we use a *Bayesian mean tracker* (BMT) as a reinforcement learning model for estimating rewards and uncertainties, which are then updated based on prediction error. The BMT is a variant of a Kalman filter, but assumes a time-invariant reward distribution (as is the case in our experiment) instead of a dynamically changing one. Both models use an updating rule based on prediction error, and have been described as a Bayesian extension of the classic Rescorla-Wagner model of associative learning (Gershman, 2015). Variants of the BMT have been used to describe human behavior in a variety of multi-armed bandit and decision-making tasks (Gershman, 2018, in press; Yu & Dayan, 2003; Schulz, Konstantinidis, & Speekenbrink, 2015; Dayan, Kakade, & Montague, 2000; Speekenbrink & Konstantinidis, 2015).

The BMT learns a posterior distribution over the mean reward μ_j for each option j. Rewards are assumed to be normally distributed with a known variance but unknown mean. The prior distribution of the mean is also a normal distribution. This implies that the posterior distribution for each mean is also a normal distribution:

$$p(\mu_{j,t}|\mathcal{D}_{t-1}) = \mathcal{N}(m_{j,t}, v_{j,t}) \tag{1}$$

where $\mathcal{D}_{t-1}$ denotes the previously observed rewards for all options. For a given option j, the posterior mean $m_{j,t}$ and variance $v_{j,t}$ are only updated when it has been selected at trial t:

$$m_{j,t} = m_{j,t-1} + \delta_{j,t} G_{j,t} [y_t - m_{j,t-1}] \tag{2}$$

$$v_{j,t} = [1 - \delta_{j,t} G_{j,t}] v_{j,t-1} \tag{3}$$

where $\delta_{j,t} = 1$ if option j is chosen on trial t, and 0 otherwise. Additionally, y_t is the observed reward at trial t, and $G_{j,t}$ is defined as:

$$G_{j,t} = \frac{v_{j,t-1}}{v_{j,t-1} + \theta_\varepsilon^2} \tag{4}$$

where θ_ε^2, referred to as the error variance, is the variance of the rewards around the mean. For our model-based analysis, we set the error variance to 1 (which led to competitive task performance in prior simulations).

Intuitively, the estimated mean of the chosen option $m_{j,t}$ is updated based on prediction error, which is the difference between the observed reward y_t and the prior expectation $m_{j,t-1}$, multiplied by learning rate $G_{j,t} \in [0,1]$. At the same time, the estimated variance $v_{j,t}$ of the chosen option is reduced by a factor $1 - G_{j,t}$. The error variance (θ_ε^2) can be interpreted as an inverse sensitivity, where smaller values result in more substantial updates to the mean $m_{j,t}$, and larger reductions of uncertainty $v_{j,t}$. We set the prior mean to $m_{j,0} = 45$ and the prior variance to $v_{j,0} = 55$ based on the expectation across payoff conditions.[2]

Results

We followed Gershman (in press) and generated predictions from the BMT by feeding in a participant's observations on a particular round until time t, and then predicting the mean and standard deviation for each option at time point $t+1$. We used the resulting predictions of rewards and uncertainties to conduct three model-based analyses of choices, reaction times, and evidence accumulation.

Choices. In our first analysis, we assessed how the predicted mean and uncertainty of an option affected the likelihood of it being chosen on each trial (estimated separately for limited and unlimited time conditions). We applied hierarchical Bayesian inference to estimate the parameters of a softmax policy, under the assumption that a participant's choice on each trial is influenced by both the predicted mean and uncertainty of an option, where each participant's parameters are assumed to be jointly normally distributed. The probability of choosing option j on trial t is a softmax function of its decision value $Q_{j,t}$:

$$P(C_t = j) = \frac{\exp(Q_{j,t})}{\sum_{k=1}^{4} \exp(Q_{k,t})}. \tag{5}$$

The decision value $Q_{j,t}$ is a linear function of the estimated mean $m_{j,t}$ and uncertainty $\sqrt{v_{j,t}}$ (estimated as a standard deviation) of each option according to the BMT:

$$Q_{j,t} = \beta_1 m_{j,t} + \beta_2 \sqrt{v_{j,t}}. \tag{6}$$

Formally, we assume that the β-coefficients for each participant $\beta_i = (\beta_{1,i}, \beta_{2,i})$ are drawn from a normal distribution

$$\beta_i \sim \mathcal{N}(\mu_\beta, \sigma_\beta^2), \tag{7}$$

[2] We use the shifted reward values that were observed by participants, where the means in each condition were centered on 0 and shifted by $\mathcal{U}(30,60)$.

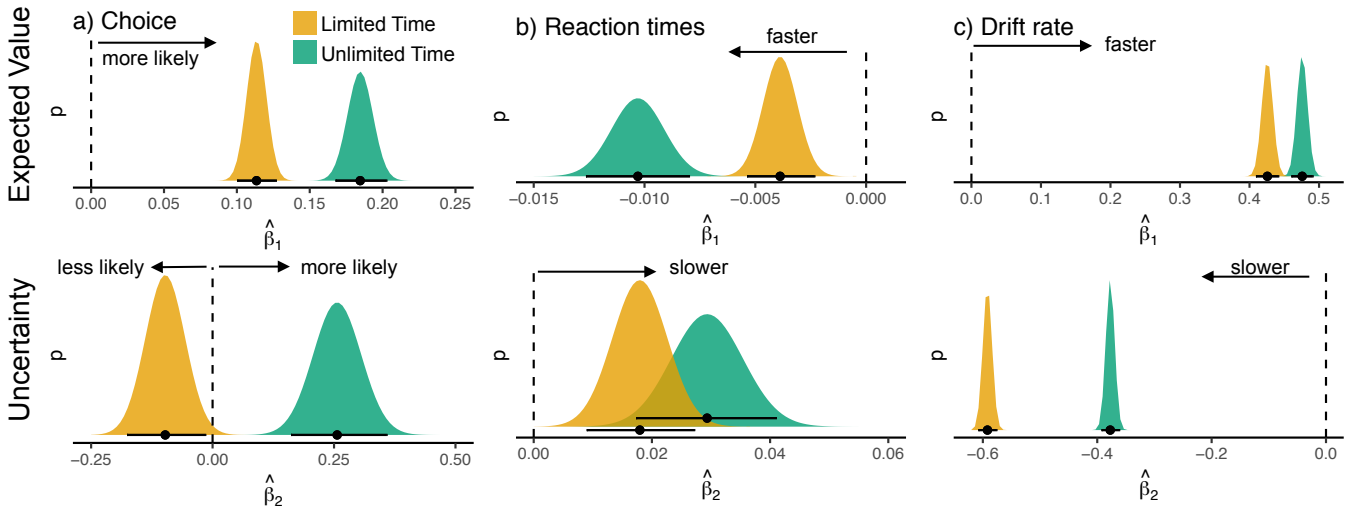


Figure 3: **Posterior parameter estimates. a**) Effects of BMT predicted mean rewards ($\hat{\beta}_1$) and uncertainties ($\hat{\beta}_2$) on an option's probability of being chosen, estimated by a hierarchical Bayesian softmax regression. **b**) Influence of BMT means ($\hat{\beta}_1$) and uncertainties ($\hat{\beta}_2$) on participant response times, estimated by a hierarchical Bayesian linear regression. **c**) Influence of BMT means ($\hat{\beta}_1$) and uncertainties ($\hat{\beta}_2$) on drift rates in a Bayesian Linear Ballistic Accumulator model. In all plots, the vertical dashed line indicates an effect of 0, while the black dot indicates the mean effect and confidence intervals show the 95% highest posterior density (HPD).

and we estimate the group-level mean μ_β and variance over participants σ^2_β. We used the following priors on the group-level parameters:

$$\mu_\beta \sim \mathcal{N}(0, 100) \tag{8}$$

$$\sigma_\beta \sim \text{Half-Cauchy}(0, 100) \tag{9}$$

In each time condition, we arrive at group-level parameter estimates describing how expected rewards (β_1) and uncertainty (β_2) influence choice probability under the softmax policy.

We estimated the hierarchical model using Hamiltonian Markov chain Monte Carlo sampling with `PyMC3` (Salvatier, Wiecki, & Fonnesbeck, 2016). The results (Fig 3a) show that the expected value of an option increased choice probability for both the limited time ($\hat{\beta}_1 = .11$, 95% HPD: $[.10, .13]$) and the unlimited time conditions ($\hat{\beta}_1 = .19$, 95% HPD: $[.17, .2]$). Options estimated to have higher expected rewards were more likely to be chosen in both conditions, with a stronger effect in the unlimited time conditions.

Notably, we found contrasting effects of uncertainty on choice probability. In the unlimited time conditions, uncertainty had a positive effect on choice probability ($\hat{\beta}_2 = .26$, 95% HPD: $[.16, .36]$). This replicates previous findings reported in two-armed bandit tasks without time pressure (Gershman, 2018, in press). However, uncertainty had a negative effect on choice probability in the limited time condition ($\hat{\beta}_2 = -.10$, 95% HPD: $[-.18, -.02]$). Thus, whereas participants sought out uncertain options in the unlimited time condition, they shunned uncertain options in the limited time condition.

Reaction Time. Our second analysis looked at how the estimated means and uncertainties of options influenced reaction times. We normalized the BMT predictions of mean reward and uncertainty by calculating the difference between the chosen option and the average of the unchosen options on each trial. Thus, positive values indicate that expected reward/uncertainty are relatively larger than those of the unchosen options. We regressed these normalized means and uncertainties onto participant log reaction times[3] in a hierarchical Bayesian linear regression, using the same priors over the β-coefficients as before (Eq. 9).

The resulting posterior parameter estimates (Fig. 3b) show that participants were faster at choosing options with relatively higher expected reward in both conditions, but with a stronger effect in the unlimited ($\hat{\beta} = -.01$, 95% HPD: $[-.013, -.008]$) than in the the limited time condition ($\hat{\beta} = -.004$, 95% HPD: $[-.005, -.002]$). Furthermore, participants were slower at choosing options with higher relative uncertainty in both the limited ($\hat{\beta} = .02$, 95% HPD: $[.01, .03]$) and the unlimited conditions ($\hat{\beta} = .03$, 95% HPD: $[.02, .04]$). Thus, whereas higher relative value made participants act faster, higher relative uncertainty slowed them down. This differs from previous findings using two-armed bandits (Gershman, in press), which showed higher relative uncertainty makes participants choose faster.

Evidence Accumulation. In our third analysis, we used the Linear Ballistic Accumulator (LBA; Brown & Heathcote, 2008) to model choices and reaction times simultaneously. This model assumes that choices are the result of a process in which evidence for each option is accumulated continuously over time, and that option is chosen for which the accumulated evidence first exceeds a set decision threshold.

Formally, the LBA assumes that, after an initial period of non-decision time τ, evidence for an option j on trial t accumulates at a rate of $v_{j,t}$, starting from an initial evidence

[3] 1 ms was added to each RT to avoid $\log(0)$. Additionally, RTs were truncated at 5000 ms.

level $p_{j,t} \sim \mathcal{U}(0, A)$. Evidence accumulates for each option j until a threshold b is reached. We follow the Bayesian implementation proposed by Annis, Miller, and Palmeri (2017) and assume that the priors for the drift rates stem from truncated normal distributions

$$v_{j,t} \sim \mathcal{N}(2,1) \in (0,\infty). \tag{10}$$

Additionally, we assume a uniform prior on non-decision time

$$\tau \sim \text{Uniform}(0,1), \tag{11}$$

and a truncated normal prior on the maximum starting evidence

$$A \sim \mathcal{N}(0.5,1) \in (0,\infty). \tag{12}$$

Finally, we reparameterized the model by shifting b by k units away from A, and put a truncated normal distribution as the prior on the resulting relative threshold k:

$$k \sim \mathcal{N}(0.5,1) \in (0,\infty). \tag{13}$$

We estimated the LBA parameters for each participant in every round using No-U-Turn Hamiltonian MCMC (Hoffman & Gelman, 2014), with reaction times truncated at 5000 ms. Participants had higher mean drift rates under limited time compared to unlimited time ($t(98) = 7.1$, $p < .001$, $d = 0.7$, $BF > 100$), consistent with the need to arrive at decisions more quickly. Participants in the limited time conditions also had shorter non-decision times τ ($t(98) = -4.6$, $p < .001$, $d = 0.5$, $BF > 100$), less maximum starting evidence A ($t(98) = -7.8$, $p < .001$, $d = 0.8$, $BF > 100$), and lower relative thresholds k ($t(98) = -5.2$, $p < .001$, $d = 0.5$, $BF > 100$), compared to participants in the unlimited time conditions. Thus, our LBA results confirm the intuition that participants thought more carefully about different options given unlimited time.

We then regressed the BMT predictions of relative expected reward and relative uncertainty for each option onto its estimated drift rate using a Bayesian linear regression. The result of this analysis revealed that the relative expected value of an option had a positive effect on drift rate for both the limited ($\hat{\beta} = .43$, 95% HPD: $[.41, .44]$; see Fig. 3c) and unlimited time conditions ($\hat{\beta} = .48$, 95% HPD: $[.46, .49]$), with a stronger effect in the latter. Conversely, relative uncertainty had a negative effect on drift rate, which was larger in magnitude for the limited ($\hat{\beta} = -.59$, 95% HPD: $[-.61, -.58]$) than for the unlimited time conditions ($\hat{\beta} = -.38$, 95% HPD: $[-.39, -.36]$). Thus, the behavioral patterns in Figure 2b suggest that uncertainty reduced the rate of evidence accumulation, with a stronger effect under time pressure than in the unlimited time conditions.

Discussion and Conclusion

How do people explore uncertain options under time pressure? We investigated this question using several variants of a four-armed bandit task with continuous rewards, while manipulating the available decision time to be either unlimited or limited to less than 400 ms.

Our models showed that higher relative uncertainty made an option more likely to be chosen in the absence of time pressure. This matches previous findings showing evidence for an exploration bonus consistent with directed exploration (Gershman, in press). However, putting participants under time pressure inverted this relationship, and caused uncertainty to reduce the probability that an option was chosen. Thus, the uncertainty bonus found in standard multi-armed bandit tasks can turn into an uncertainty penalty when people are under time pressure. This is similar to findings from description-based gambles, where time pressure increased risk aversion (Nursimulu & Bossaerts, 2013).

We also found that relative uncertainty slowed down choices and dampened evidence accumulation. These results suggest that uncertainty can have reversible effects on preference: sometimes people seek out uncertainty, and sometimes they actively avoid it. Both of these cases suggest people track uncertainty in their expectations, and that uncertainty feeds into the decision-making process. This is similar to what has been observed in tasks that directly elicit confidence judgments (Boldt, Blundell, & De Martino, 2017; Stojic, Schulz, Analytis, & Speekenbrink, 2018; Schulz, Wu, Ruggeri, & Meder, 2018; Wu, Schulz, Garvert, Meder, & Schuck, 2018), while previous work has shown that changing the context from only gains to adding risky options can also cause a shift from actively seeking uncertainty to avoiding it (Schulz, Wu, Huys, Krause, & Speekenbrink, 2018).

Our results provide a richer understanding of the cognitive processes underlying human learning and exploration. While we found evidence that time pressure reduces directed exploration—consistent with directed exploration being a controlled and reasoned process—we did not predict uncertainty avoidance under time pressure. Together with the finding that relative uncertainty slowed down reaction times and dampened evidence accumulation, our results suggest that time pressure does not eliminate the ability to track uncertainty. Rather, it alters attitudes towards it, from seeking out uncertainty to avoiding it. Future studies should therefore investigate the conditions that cause uncertainty-seeking or uncertainty-avoidance and test whether uncertainty-avoidance is a deliberate behavior (Schulz, Klenske, Bramley, & Speekenbrink, 2017).

Acknowledgments

CMW is supported by the International Max Planck Research School on Adapting Behavior in a Fundamentally Uncertain World; ES is supported by the Harvard Data Science Initiative

References

Annis, J., Miller, B. J., & Palmeri, T. J. (2017). Bayesian inference with Stan: A tutorial on adding custom distributions. *Behavior Research Methods*, *49*, 863–886.

Auer, P., Cesa-Bianchi, N., & Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, *47*, 235–256.

Bechara, A., Damasio, A. R., Damasio, H., & Anderson, S. W. (1994). Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition*, *50*, 7–15.

Boldt, A., Blundell, C., & De Martino, B. (2017). Confidence modulates exploration and exploitation in value-based learning. *bioRxiv*, 236026.

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology*, *57*, 153–178.

Daw, N. D., O'doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*, 876.

Dayan, P., Kakade, S., & Montague, P. R. (2000). Learning and selective attention. *Nature neuroscience*, *3*(11s), 1218–1223.

Gershman, S. J. (2015). A unifying probabilistic view of associative learning. *PLoS Computational Biology*, *11*, e1004567.

Gershman, S. J. (2018). Deconstructing the human algorithms for exploration. *Cognition*, *173*, 34–42.

Gershman, S. J. (in press). Uncertainty and exploration. *Decision*.

Gittins, J. C. (1979). Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society. Series B (Methodological)*, 148–177.

Gronau, Q. F., Singmann, H., & Wagenmakers, E.-J. (2017). Bridgesampling: an r package for estimating normalizing constants. *arXiv preprint arXiv:1710.08162*.

Hoffman, M. D., & Gelman, A. (2014). The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, *15*, 1593–1623.

Knox, W. B., Otto, A. R., Stone, P., & Love, B. (2012). The nature of belief-directed exploratory choice in human decision-making. *Frontiers in Psychology*, *2*, 398.

Nursimulu, A. D., & Bossaerts, P. (2013). Risk and reward preferences under time pressure. *Review of Finance*, *18*, 999–1022.

Salvatier, J., Wiecki, T. V., & Fonnesbeck, C. (2016). Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, *2*, e55.

Schulz, E., & Gershman, S. J. (2019). The algorithmic architecture of exploration in the human brain. *Current Opinion in Neurobiology*, *55*, 7–14.

Schulz, E., Klenske, E., Bramley, N., & Speekenbrink, M. (2017). Strategic exploration in human adaptive control. *bioRxiv*, 110486.

Schulz, E., Konstantinidis, E., & Speekenbrink, M. (2015). Learning and decisions in contextual multi-armed bandit tasks. In *Thirty-Seventh Annual Conference of the Cognitive Science Society.*

Schulz, E., Wu, C. M., Huys, Q. J., Krause, A., & Speekenbrink, M. (2018). Generalization and search in risky environments. *Cognitive science*, *42*, 2592–2620.

Schulz, E., Wu, C. M., Ruggeri, A., & Meder, B. (2018). Searching for rewards like a child means less generalization and more directed exploration. *bioRxiv preprint*.

Speekenbrink, M., & Konstantinidis, E. (2015). Uncertainty and exploration in a restless bandit problem. *Topics in Cognitive Science*, *7*, 351–367.

Stojic, H., Schulz, E., Analytis, P. P., & Speekenbrink, M. (2018). It's new, but is it good? How generalization and uncertainty guide the exploration of novel options. *PsyArXiv preprint*.

Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A., & Cohen, J. D. (2014). Humans use directed and random exploration to solve the explore–exploit dilemma. *Journal of Experimental Psychology: General*, *143*, 155-164.

Wu, C. M., Schulz, E., Garvert, M. M., Meder, B., & Schuck, N. W. (2018). Connecting conceptual and spatial search via a model of generalization. In T. T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.), *Proceedings of the 40th annual conference of the cognitive science society* (pp. 1183–1188). Austin, TX: Cognitive Science Society.

Wu, C. M., Schulz, E., Speekenbrink, M., Nelson, J. D., & Meder, B. (2018). Generalization guides human exploration in vast decision spaces. *Nature Human Behaviour*, *2*, 915—924.

Yechiam, E., Busemeyer, J. R., Stout, J. C., & Bechara, A. (2005). Using cognitive models to map relations between neuropsychological disorders and human decision-making deficits. *Psychological Science*, *16*, 973–978.

Yu, A. J., & Dayan, P. (2003). Expected and unexpected uncertainty: ACh and NE in the neocortex. In *Advances in Neural Information Processing Systems* (pp. 173–180).

Preschoolers jointly consider others expressions of surprise and common ground to decide when to explore

Yang Wu
Stanford University, Stanford, California, United States

Hyowon Gweon
Stanford University, Stanford, California, United States

Abstract

Prior work on early social learning suggests that children are sensitive to adults pedagogical demonstrations and verbal instructions. Yet, people also display various emotional expressions when interacting with children. Here we show that young children draw rich causal inferences and guide their own exploration based on others expressions of surprise. Preschoolers (age:3.0-4.9) saw an experimenter discover a function of a novel causal toy. Then, either the same experimenter or a nave confederate expressed surprise while playing with the toy behind an occluder. Children explored the toy more broadly to search for a hidden function following the experimenters surprise than following the confederates surprise, suggesting that children integrated others expressions of surprise and others epistemic states to infer the presence of hidden functions and explore accordingly. This study synthesizes perspectives from literature on social learning, exploration, and affective cognition towards a more comprehensive science of learning. Preprint:https://psyarxiv.com/ckh6j

A predictability-distinctiveness trade-off in the historical emergence of word forms

Aotao Xu (a26xu@uwaterloo.ca)
Computer Science Program, University of Waterloo

Christian Ramiro (chrisram@berkeley.edu)
Cognitive Science Program, University of California, Berkeley

Yang Xu (yangxu@cs.toronto.edu)
Department of Computer Science, Cognitive Science Program, University of Toronto

Abstract

It has been proposed that language evolves under the joint constraints of communicative expressivity and cognitive ease. We explore this idea in the historical emergence of word forms. We hypothesize that new word forms that enter the lexicon should reflect a trade-off between predictability and distinctiveness. An emergent word form can be highly predictable if it efficiently reuses elements from the existing word forms, resulting in low cognitive load. An emergent word form should also be sufficiently distinctive from the existing lexicon, facilitating communicative expressivity. We test our hypothesis by examining the properties of 34,478 emergent word forms over the past 200 years of Modern English. We show how word forms at future time $t+1$ are bounded statistically between n-gram generated word forms (highly predictable) and slang words that are outside the standard lexicon (highly distinctive) at time t. Our work supports the view of cognitive economy in lexical emergence.

Keywords: word form; lexicon; lexical emergence; language evolution; cognitive economy

Introduction

The lexicon is a central locus of human thought, but it undergoes constant change over time. In particular, new words may emerge due to changing sociocultural needs, resulting in growth of the lexicon. Taking the English lexicon as an example, it has grown by approximately tenfold over the past millennium, with more than 150,000 word forms having emerged from the period of Old English to the present day (Figure 1a). Here we ask what principles might underlie the historical emergence of word forms above and beyond the external sociocultural factors that could influence lexical emergence.

Our starting point is the idea that language evolves under the dual considerations of communicative function and cognitive effort (Labov, 2011; Jespersen, 1959; Otto, 1956; Kirby, Tamariz, Cornish, & Smith, 2015), a prominent proposal that has been framed similarly in linguistics as the principle of least effort (Zipf, 1949) and in cognitive psychology as the principle of cognitive economy (Rosch, 1978). This proposal also relates to a growing line of research that explores design principles of language through the lens of efficient communication (Piantadosi, Tily, & Gibson, 2012; Kemp & Regier, 2012; Kemp, Xu, & Regier, 2018). Most relevant to the current study is work by Labov who suggests that words may be selected under the joint constraints of least effort (cf. Zipf, 1949)—a drive for cognitive ease of production, and the competing force of communicative informativeness (Labov, 2011). There is evidence for each of these constraints in the design of word forms. For example, it has been shown that word forms that conform to well-formed phonotactic properties can facilitate production (Edwards, Beckman, & Munson, 2004), and words that sound similar to many existing words, or having dense lexical neighbourhoods, tend to reduce speech error (Stemberger, 2004). On the other hand, separate work has suggested that perceptual distinctiveness matters in the lexicon because it minimizes confusion and facilitates clarity in communication (Flemming, 2004; Meylan & Griffiths, 2017).

We extend previous work by exploring principles in the historical emergence of novel word forms. We believe that the same proposal of language evolution should apply to explaining how new word forms enter the lexicon over time. In particular, we hypothesize that the emergence of word forms should trade off *predictability* against *distinctiveness*. An emergent word form is highly predictable if it efficiently recombines elements from existing word forms, resulting in low cognitive effort in production and memory. Our notion of predictability is rooted in classic work by Shannon (1951) on the information analysis of English text. However, this criterion of predictability is likely in competition with distinctiveness: An emergent word form should be sufficiently distinctive from words in the existing lexicon, hence generating minimal confusion and facilitating communicative expressivity. Predictability and distinctiveness trade off against each other because a highly predictable word form is necessarily similar in form to existing words, so it is unlikely to be distinctive. Similarly, a highly distinctive word form is necessarily novel in its composition, so it is unlikely to be very predictable. Here we examine the possibility that the emergent word forms in history are shaped under these two joint forces, such that they appear sandwiched between (plausible) word forms that are highly predictive and those that are highly distinctive (see Figure 1b for illustration).

We test our hypothesis by examining new word forms that entered the Modern English lexicon over the past 200 years. At each future decade $t+1$, we compare the actual emergent words against a control set of computer-generated words and slang words that did not enter the standard lexicon up to the previous decade t. We show how the actual word forms are interleaved between the highly predictable and distinctive control words in terms of their statistical properties.

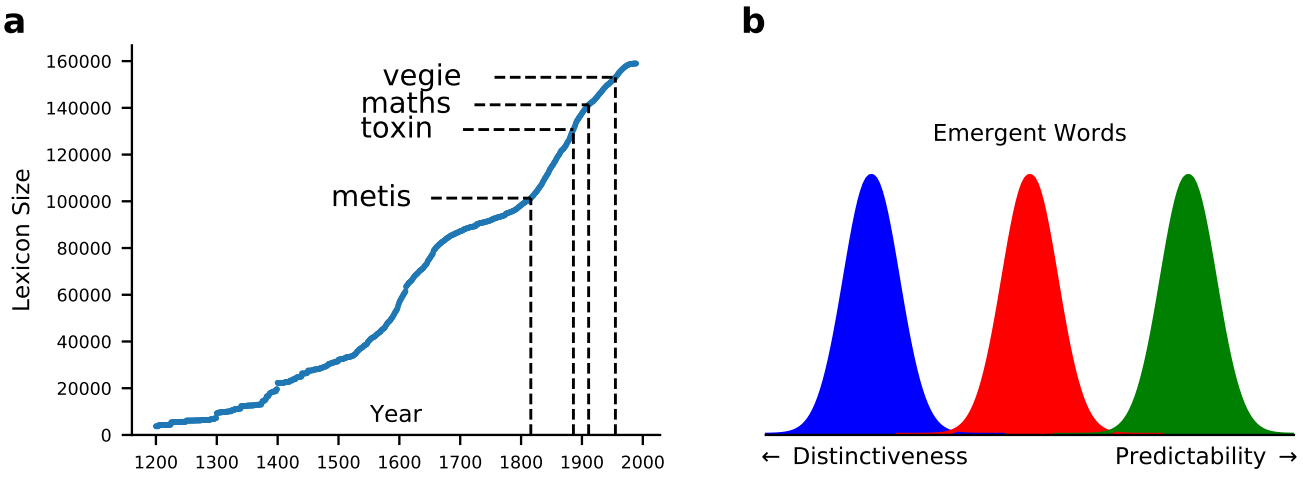


Figure 1: Illustration of the phenomenon of lexical emergence and our hypothesis. a) Growth of the English lexicon over the past 800 years based on data from the Historical Thesauraus of English (HTE). Lexicon size represents the number of unique word forms that exist during this period, and examples of emergent words are shown at respective years of emergence. b) The hypothesis that emergent words (middle) at future time $t+1$ should reflect a trade-off between predictability and distinctiveness among the space of plausible word forms (on the two sides) given the current lexicon at time t, or effectively sandwiched.

Materials and Methods

We analyzed 34,478 word forms during each decade from 1800 to 1980 as recorded in the Historical Thesauraus of English (HTE) (Kay, Roberts, Samuels, & Wotherspoon, 2017) (`https://ht.ac.uk/`) which is based on the Oxford English Dictionary. We considered single-word lexemes that are composed of 26 English letters (a-z) and took the first year recorded in HTE for a given word form as its emerging date. We focused our analysis on the Modern English period due to both orthographic and phonetic changes in words during the remote periods of Old English and Middle English (Baugh & Cable, 1993, p279), and the lack of control words for the same periods that is critical for our analyses. We grouped word forms according to the lengths of their orthographic forms, and we considered lengths ranging from 4 to 9 because lengthier words are more likely formed due to rule-based compositional strategies (Krott, 1996). The grouping by word length is necessary because longer words are by chance more distinctive in form than shorter words, so a principled analysis of our trade-off hypothesis should be independent of length. We focus on reporting results based on orthographic forms, although we observed similar results with phonological forms that we do not include here due to space limit.

We used two standard measures to quantify the statistical properties of word forms along the predictability–distinctiveness dimension: letter n-gram probability and lexical neighbourhood density. We quantify the two measures for word forms at a future decade $t+1$ based on statistical properties of the existing lexicon at a decade earlier at t. Formally, we define the probability of an emergent word form w of length $|w|$ by using the n-gram probabilities of its constituent letters (or phonemes), extending the classic work by Shannon on information analyses of English words (Shannon, 1951):

$$p^{t+1}(w) = \prod_{i=1}^{|w|} p^t(l_i|l_{<i}) \quad (1)$$

$$= p^t(l_1|\cdot) \times p^t(l_2|l_1) \times p^t(l_3|l_2,l_1) \times \quad (2)$$
$$\ldots \times p^t(l_{|w|}|l_{|w|-1},\ldots,l_1)$$

Equations 1-2 effectively estimate how probable a novel word form w would be at decade $t+1$ given the n-gram statistics at the current decade t. We considered n-gram of up to order 5 because statistics of higher orders are sparse and prohibitively expensive to compute. Under this measure, a highly predictable word form at $t+1$ for a given length should be one that maximizes the n-gram probability based on the lexical statistics at t. On the contrary, a highly distinctive word form should have low predictability that minimizes the same probability measure.

To ensure the robustness of our approach, we considered lexical neighbourhood density as an alternative measure. We define the neighbourhood density of an emergent word form w based on how similar it is to existing word forms v in the lexicon at time t (L^t), grounded in the psycholinguistic study of English word forms by Bailey and Hahn (2001):

$$ND^{t+1}(w) = \sum_{v \in L^t} e^{-d(w,v)} \quad (3)$$

Equation 3 effectively estimates how crowded a novel word form w would be at decade $t+1$ given existing word forms at the current decade t. We used the standard Levenshtein

edit distance for calculating $d(\cdot,\cdot)$ that considers if two word forms are similar or distant based on the number of edits required to match the forms via insertion, deletion, or substitution (Yarkoni, Balota, & Yap, 2008; Bailey & Hahn, 2001). For example, the edit distance between "cat" and "maths" is 3 since the edit involves one substitution and two insertions. Similar to the case of the *n*-gram measure, a highly predictable word form at a given length should be one that maximizes neighbourhood density at $t+1$. On the contrary, a highly distinctive word form should not be crowded and hence minimizes its lexical neighbourhood density.

To evaluate the hypothesis that emergent word forms trade off predictability against distinctiveness, we considered a set of control word forms that are representative of the extremities of this dimension yet did not formally enter the English lexicon. Our goal is to assess whether the trade-off hypothesis might explain why certain word forms have entered the lexicon over time, whereas other plausible forms have not appeared. Because the set of all possible word forms is enormous (e.g., there are over 10 million possible word forms of length 5 that did not appear in English up to 1980), we chose control words by focusing on word forms that are either likely to be very predictable or distinctive.

We first obtained the *predictable control set* by generating word forms according to the *n*-gram probability measure in Equations 1-2. At each yet-to-emerge decade, we sampled these word forms from the *n*-gram statistics obtained from the previous decade in a sample size that matches the number of the emergent words. The sample does not intersect with the lexicon, but it can intersect with the set of actual emerging words. We then partitioned these control words by length and calculated their *n*-gram probabilities and neighbourhood densities according to Equations 1-3. This control word set approximates the extremity of predictability because the candidates are directly generated from the distribution of the existing lexicon, so they should be statistically equivalent to the existing word forms in the lexicon. Because *n*-gram probability correlates with neighbourhood density (Sanders & Chin, 2009), we also expect this word set to have high (but not necessarily the maximal) neighbourhood density. If the trade-off hypothesis is correct, the emergent word forms should generally have lower but not near-identical *n*-gram probability and neighbourhood density to this control set.

We next obtained the *distinctive control set* by sampling word forms from slang that did not enter the standard English lexicon. Slang is likely to represent the extremity of distinctiveness because slang words are known to differ from the standard lexicon (Mattiello, 2008, 2013), and 2) they serve as a more conservative measure for plausible word forms (plausible because a subset of slang can eventually become actual words (Baugh & Cable, 1993, p293)) than random samples of non-existent word forms that can be distinctive but not permissible, e.g., "jxyzh" is very distinctive from existing words in English but it is not permissible based on the knowledge of English. We drew data from a large online resource, the Urban Dictionary (`https://www.urbandictionary.com/`), for this control set. We used word forms containing only the letters a-z conforming to the same selection standard with the emergent words. During each decade of interest, we excluded homographs of word forms or words that have overlapping lemma in the lexicon via the lemmatizer from the Natural Language Toolkit (NLTK) (Bird, Klein, & Loper, 2009). We then sampled from the rest of the 317,403 unique word forms in matching size to the emergent lexicon per length, and calculated the *n*-gram and neighbourhood statistics for these word forms. If the trade-off hypothesis is correct, the emergent word forms should generally have higher but not near-identical *n*-gram probability and neighbourhood density to this control set.

Results

We evaluated our hypothesis by first examining whether newly emerging word forms tend to fall between predictable control words and distinctive (slang) control words in terms of *n*-gram probability and lexical neighbourhood density. At each decade, we compared the actual emergent word forms to the two sets of control words of the same length under the two measures separately. We took the average values of the two measures for each word group and every length that we examined.

Figure 2 summarizes the results for these comparisons for every decade from 1800 to 1980 and word forms of lengths 4 to 9. In most cases, we observed that the emergent word group is situated in the middle between the predictable and distinctive control word sets, and the rank order of these three groups based on *n*-gram probability and neighbourhood density conforms to our prediction. Specifically, the predictable control words exhibit the highest mean predictability, manifested in the highest overall *n*-gram probability (or equivalently, the lowest overall negative logarithmic *n*-gram probability) and lexical neighbourhood density among the three groups. In comparison, the slang/distinctive control words exhibit the highest mean distinctiveness, manifested in the lowest overall *n*-gram probability and neighbourhood density. The emergent word group tends to fall in between the two control groups.

To evaluate the significance of these trade-offs, we tested a null hypothesis for each comparison between the emergent group and each of the control groups. The null hypothesis is that the mean estimate of the emergent word set does not differ in *n*-gram probability or lexical neighbourhood density from each of the control sets. We tested this by performing a two-tailed *t*-test for every comparison. Across different word lengths and time periods, we observed consistent evidence for rejecting the null hypothesis (see Figure 3; the variations in the magnitude of p values correlate with time and changing sample sizes, as the number of actual emergent words are different in every decade). These results show that the emergent words are significantly different from the control words.

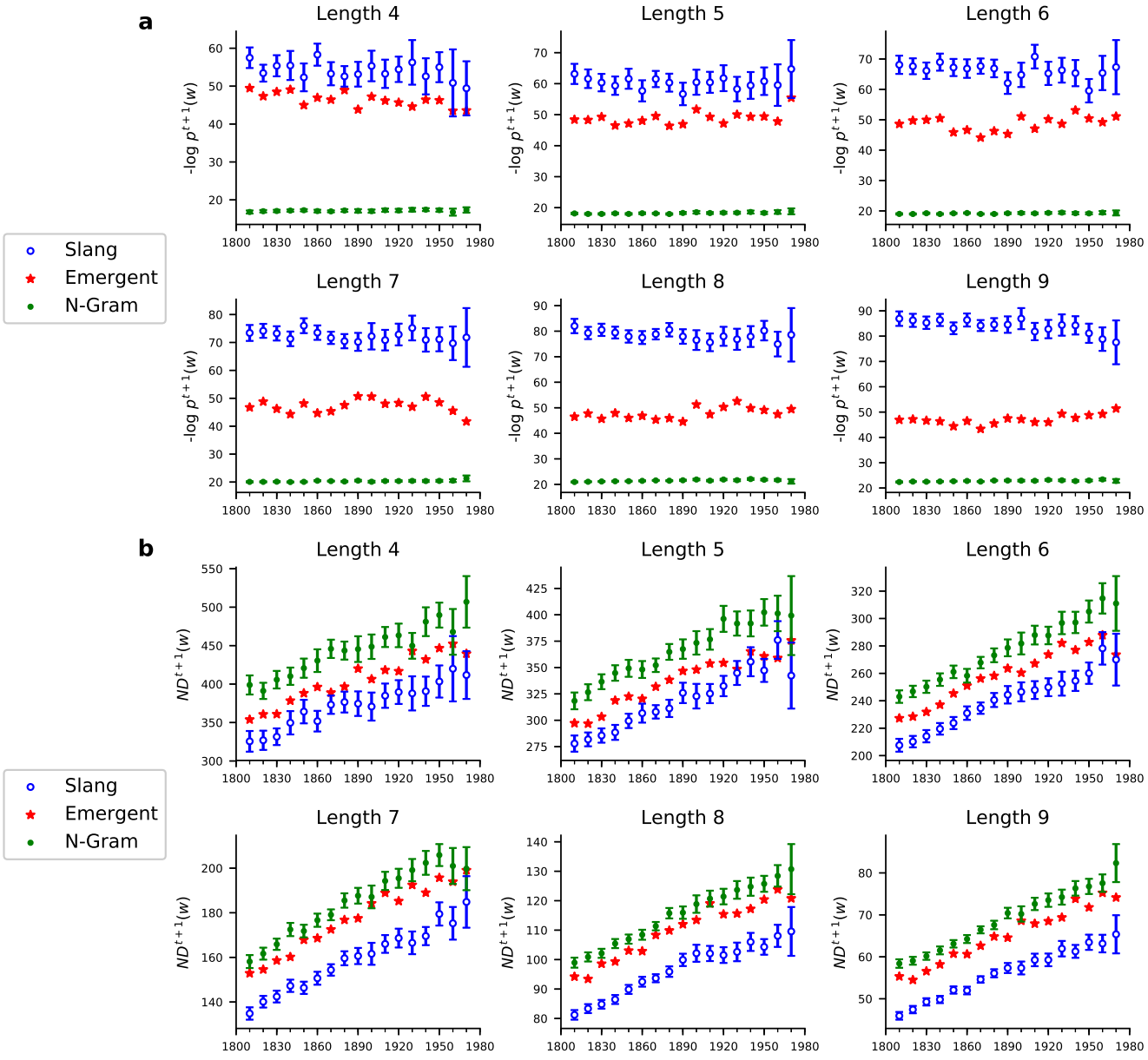


Figure 2: Summary of main results on the trade-off hypothesis of lexical emergence. Each of the panels summarizes the results computed under one of the two measures: a) *n*-gram probability (negative logarithm) and b) neighbourhood density. The vertical axes represent magnitudes under these two measures, and the horizontal axis represents the temporal dimension where each tick corresponds to one decade over the period between 1800 and 1980. Each subplot corresponds to the results for word forms of a different length as specified. Dots (green), stars (red), and circles (blue) correspond to the *n*-gram (predictable) control words, the actual emergent words, and the slang (distinctive) control words, respectively. Each error bar indicates a 95% confidence interval (constructed from the *t*-distribution) for the estimated mean value of the control group. This confidence level is uncorrected for multiple comparisons, and we expect 5% of all intervals to exclude emergent word groups by chance.

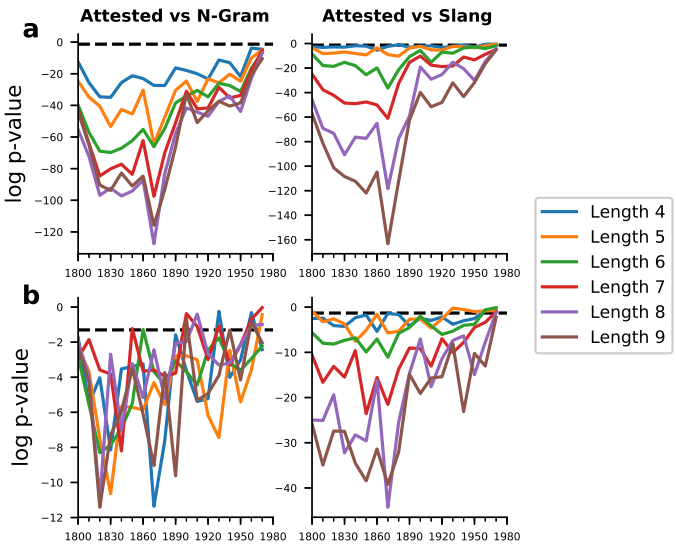


Figure 3: Time courses of p-values. Panels on the left summarize the comparison of the emergent word sets against the corresponding predictable control sets. Panels on the right summarize similar comparisons against the distinctive control sets. The comparisons were based on the measures from a) n-gram probability and b) neighbourhood density. The vertical axis indicates p-values from the t-tests in logarithmic scale, and the horizontal axis represents the time dimension in decades. The black dashed line represents the significance level $p = 0.05$. For each measure, we made 216 simultaneous uncorrected tests, so we expect 11 rejections by chance.

To assess the robustness of these findings, we performed similar analyses based on 1) word forms defined in phonological space as opposed to orthographical space; 2) alternative lexicons obtained by excluding morphologically derived words from the HTE data; 3) an alternative control set based on slang words from a historical resource as opposed to a modern resource. We found that the effects are robust to this variation in design choices, and we omit the details of these analyses due to space constraints. In sum, this set of results provide empirical evidence for our proposal that emerging word forms reflect a trade-off between predictability and distinctiveness and suggest why certain words have entered the lexicon over time, but others have not.

As a follow-up analysis, we assessed whether we can reliably predict emergent word forms from possible words that did not formally enter the lexicon. In particular, we performed a simple logistic regression analysis to predict the identity of each word form from the three groups: emergent words, predictable control words, and distinctive control words. We applied a logistic classifier with $L2$ penalty and the multinomial loss function using the `scikit-learn` package (Pedregosa et al., 2011). For each future decade, we trained the classifier using data from the previous decade t and used the same classifier to make predictions for data from $t+1$. We used three feature sets for classification: 1) n-gram probabilities of words from the three groups; 2) lexical neighbourhood densities of the same words; 3) a combination of their n-gram probabilities and neighbourhood densities.

In general, we observed that predictive accuracies of the three word groups are above chance (33.3% for a three-way classification) under all three feature choices for each decade and length that we considered (predictive accuracy when using neighbourhood density, mean = 43.0%, and standard deviation across word length groups and time periods, SD = 4.2%; using n-gram probability, mean= 61.0%, SD = 3.6%; using the combined features, mean= 61.0%, SD = 3.6%). We noted that the above-chance predictive accuracies are sustained over time, suggesting the trade-off holds generally and not just for certain periods in the history of Modern English. We also noted that the n-gram model performed generally better than the neighbourhood density model, partly because one of the control word groups was directly simulated using n-gram statistics.

Overall, these findings suggest that there are predictable differences in the compositional structure of emergent word forms and that of n-gram generated and slang word forms from the control groups.

Figure 4 further demonstrates the trade-off idea with three example word forms chosen from the three word groups in the 1930s, along with their nearest-neighbour word forms measured by edit distance from the same period. The emergent word form "macro" reflects a trade-off in neighbourhood density: It has fewer 1-edit lexical neighbours (6) than the highly predictable n-gram generatd word "codet" (9 neighbours), but it has more neighbours than the highly distinctive slang word "porph" that has the fewest neighbours (3).

Figure 4: Demonstration of the predictability-distinctiveness trade-off. Panels a), b), and c) show an example word form and its lexical neighbours from the lexicon in the 1930s under the n-gram control set, emergent word set, and slang control set, respectively. The examples are placed in the center, surrounded by their neighbours. Each example word is exactly one edit distance away from each of its neighbours.

Conclusion

We have shown that the historical emergence of English word forms follows a trade-off between predictability and distinctiveness. This trade-off is manifested in the properties of emergent words that straddle between 1) highly predictable computer-generated word forms that conform to statistical properties of the existing lexicon, and 2) highly distinctive word forms originated from slang that had not yet enter into the standard lexicon. We have suggested that such a trade-off may reflect the general principles of language evolution discussed in prior work, under the joint functional pressures for communicative expressivity and cognitive ease (Labov, 2011; Jespersen, 1959; Otto, 1956). Future research should explore whether the same set of principles holds in the emergence of word forms in languages other than English and how word forms interact with meaning (cf. Ramiro, Srinivasan, Malt, & Xu, 2018) in lexical evolution.

Acknowledgments

We thank the University of Glasgow for licensing of the HTE data. We also thank Barbara C. Malt, Peter Turney, Suzanne Stevenson, and Barend Beekhuizen for their constructive comments on this work. This research is supported by an NSERC DG grant and a Connaught New Researcher Award to YX.

References

Bailey, T. M., & Hahn, U. (2001). Determinants of wordlikeness: Phonotactics or lexical neighborhoods? *Journal of Memory and Language*, *44*(4), 568–591.

Baugh, A. C., & Cable, T. (1993). *A history of the english language*. London, UK: Routledge.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with python: analyzing text with the natural language toolkit*. Sebastopol, CA: O'Reilly Media, Inc.

Edwards, J., Beckman, M. E., & Munson, B. (2004). The interaction between vocabulary size and phonotactic probability effects on children's production accuracy and fluency in nonword repetition. *Journal of speech, language, and hearing research*, *47*(2), 421–436.

Flemming, E. (2004). Contrast and perceptual distinctiveness. In B. Hayes, R. Kirchner, & D. Steriade (Eds.), *The phonetic bases of markedness.* (pp. 232–276). Cambridge, UK: Cambridge University Press.

Jespersen, O. (1959). *Language: Its nature, development and origin*. London: Allen & Unwin.

Kay, C., Roberts, J., Samuels, M., & Wotherspoon, I. (Eds.). (2017). *The historical thesaurus of english, version 4.21*. Glasgow, UK: University of Glasgow. Retrieved from `http://historicalthesaurus.arts.gla.ac.uk/`

Kemp, C., & Regier, T. (2012). Kinship categories across languages reflect general communicative principles. *Science*, *336*, 1049–1054.

Kemp, C., Xu, Y., & Regier, T. (2018). Semantic typology and efficient communication. *Annual Review of Linguistics*, *4*, 109–128.

Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, *141*, 87–102.

Krott, A. (1996). Some remarks on the relation between word length and morpheme length. *Journal of Quantitative Linguistics*, *3*(1), 29–37.

Labov, W. (2011). *Principles of linguistic change, volume 3: Cognitive and cultural factors* (Vol. 36). Hoboken, NJ: John Wiley & Sons.

Mattiello, E. (2008). *An introduction to english slang: A description of its morphology, semantics and sociology* (Vol. 2). Monza, Italy: Polimetrica-International Scientific Publisher.

Mattiello, E. (2013). *Extra-grammatical morphology in English: Abbreviations, blends, reduplicatives, and related phenomena* (Vol. 82). Berlin, Germany: Walter de Gruyter.

Meylan, S. C., & Griffiths, T. L. (2017). Word forms–not just their lengths–are optimized for efficient communication. *arXiv preprint arXiv:1703.01694*.

Otto, J. (1956). *Language: Its nature development and origin.* Crows Nest, Australia: George Allen & Unwin Limited.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Piantadosi, S. T., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, *122*, 280–291.

Ramiro, C., Srinivasan, M., Malt, B. C., & Xu, Y. (2018). Algorithms in the historical emergence of word senses. *Proceedings of the National Academy of Sciences*, *115*(10), 2323–2328.

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (Eds.), *Cognitionandcategorization* (pp. 27–48). Hillsdale: Lawrence Erlbaum.

Sanders, N. C., & Chin, S. B. (2009). Phonological distance measures. *Journal of Quantitative Linguistics*, *16*(1), 96–114.

Shannon, C. E. (1951). Prediction and entropy of printed english. *Bell System Technical Journal*, *30*(1), 50–64.

Stemberger, J. P. (2004). Neighbourhood effects on error rates in speech production. *Brain and Language*, *90*(1-3), 413–422.

Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond coltheart's n: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, *15*(5), 971–979.

Zipf, G. K. (1949). *Human behavior and the principle of least effort: An introduction to human ecology*. Cambridge: Addison-Wesley.

Explaining intuitive difficulty judgments by modeling physical effort and risk

Ilker Yildirim
Massachusetts Institute of Technology, Cambridge, Massachusetts, United States

Basil Saeed
Massachusetts Institute of Technology, Cambridge, Massachusetts, United States

Grace Bennett-Pierre
Stanford University, Stanford, California, United States

Tobias Gerstenberg
Stanford University, Stanford, California, United States

Josh Tenenbaum
MIT, Cambridge, Massachusetts, United States

Hyowon Gweon
Stanford University, Stanford, California, United States

Abstract

The ability to estimate task difficulty is critical for many real-world decisions such as setting appropriate goals for ourselves or appreciating others' accomplishments. Here we give a computational account of how humans judge the difficulty of a range of physical construction tasks (e.g., moving 10 loose blocks from their initial configuration to their target configuration, such as a vertical tower) by quantifying two key factors that influence construction difficulty: physical effort and physical risk. Physical effort captures the minimal work needed to transport all objects to their final positions, and is computed using a hybrid task-and-motion planner. Physical risk corresponds to stability of the structure, and is computed using noisy physics simulations to capture the costs for precision (e.g., attention, coordination, fine motor movements) required for success. We show that the full effort-risk model captures human estimates of difficulty and construction time better than either component alone. Preprint link https://arxiv.org/abs/1905.04445.

Tensions Between Science and Intuition in School-Age Children

Andrew G. Young (ayoung2@oxy.edu), Isabel Geddes, Claire Weider, & Andrew Shtulman (shtulman@oxy.edu)
Department of Psychology, Occidental College
1600 Campus Road, Los Angeles, CA 90041

Abstract

Adults with extensive science education exhibit cognitive conflict when reasoning about counterintuitive scientific ideas, such as whether clouds have weight or whether bacteria need nutrients. Here, we investigated whether elementary-school-aged children show the same conflict and whether that conflict can be reduced by targeted instruction. Seventy-eight 5- to 12-year-olds verified, as quickly as possible, statements about life and matter before and after a tutorial on the scientific properties of life or matter. Half the statements were consistent with intuitive theories of the domain (e.g., "frogs reproduce") and half were inconsistent (e.g., "cactuses reproduce"). Participants verified the latter less accurately and more slowly than the former, both before instruction and after. Instruction increased the accuracy of participants' verifications for counterintuitive statements within the domain of instruction but not their speed. These results indicate that children experience conflict between scientific and intuitive conceptions of a domain in the earliest stages of acquiring scientific knowledge but can learn to resolve that conflict in favor of scientific conceptions.

Keywords: conceptual development, scientific reasoning, explanatory coexistence, intuitive theories

Introduction

Our first theories of natural phenomena are often incompatible with the scientific theories we learn later in life. We first conceive of heat as an invisible substance that flows in and out of objects rather than kinetic energy at the molecular level (Reiner, Slotta, Chi, & Resnick, 2000). We conceive of forces as properties imparted to objects, propelling them forward, rather than as interactions between objects, changing their velocity (McCloskey, 1983). Colds and flus are thought to be caused by cold air rather than a virus (Au et al., 2008). And lunar phases are thought to be caused by the earth's shadow on the moon rather than our changing view of the moon's illuminated surface (Trundle, Atwood, & Christopher, 2002).

Our first theories are known as folk theories, naïve theories, or intuitive theories. They are developed by children from a variety of inputs, including innate biases, firsthand experience, cultural artifacts, and cultural beliefs (Carey, 2009; Shtulman, 2017; Vosniadou, 1994). Intuitive theories play the same inferential role as scientific theories, helping us explain past events, predict future events, and intervene on present events (Gopnik & Wellman, 2012). They differ from scientific theories, however, in that they carve up the world into entities and processes that do not align with the true causes of natural phenomena.

One well-studied example of intuitive theories are children's theories of life (Hatano & Inagaki, 1994; Slaughter & Lyons, 2003; Stavy & Wax, 1989). Life is a metabolic state—the consumption of energy to further an organism's survival and reproduction—but young children do not know of the internal components of organisms that make metabolism possible. In the absence of such knowledge, they interpret "life," "living," and "alive" as descriptions of motion. Entities that move on their own are deemed alive, regardless of their metabolic status. Thus, preschoolers mistakenly classify moving but nonliving entities, like the sun and the clouds, as alive, and they mistakenly classify living but nonmoving objects, like plants and trees, as not alive. These mistakes persist until children conceive of life as supported by the interrelated functions of internal organs, typically by age ten.

Another well-studied example are children's theories of matter (Carey, 1991; Nakhleh, Samarapungavan, & Saglam, 2005; Smith, 2007). Matter is anything composed of atoms, but many such substances betray no perceptible sign of their underlying composition. Gases, vapors, and microscopic objects are all composed of atoms, but children can neither see them nor hold them, so they classify them as nonmaterial. They also deny that such entities have weight or take up space. Children also make the converse mistake of classifying nonmaterial entities that they can see or feel as matter, including echoes, shadows, and heat. This pattern persists until early adolescence, when children learn a particulate theory of matter in introductory physical science.

Learning to reinterpret phenomena covered by an intuitive theory through the lens of a scientific theory requires conceptual change, or knowledge revision at the level of individual concepts. Conceptual change has traditionally been viewed as a process of restructuring and replacement (Carey, 1985; Chi, 1992; Nersessian, 1989; Vosniadou, 1994). Intuitive theories are restructured to accommodate counterintuitive scientific information and are thus replaced in the process, in the same way that remodeling a house erases the footprint of its original layout.

This view has been challenged by recent research revealing that intuitive theories are not entirely erased by scientific theories and will, in fact, influence domain-relevant reasoning under cognitive load or cognitive impairment. In the domain of life, for instance, college undergraduates instructed to classify entities as "alive" or "not alive" as quickly as possible are prone to make the kinds of mistakes preschoolers make, classifying moving but nonliving things as alive and living but nonmoving things as not alive (Goldberg & Thompson-Schill, 2009). That is, undergraduates are less accurate at classifying plants as alive relative to animals, and they are less accurate at classifying dynamic objects (like clocks, geysers, comets,

and rivers) as not alive relative to static ones. They are also slower to do so. Similar results have been found for Alzheimer's patients with moderate dementia, who not only misclassify moving but nonliving entities as alive but also explicitly define life in terms of motion rather than metabolic activity (Zaitchik & Solomon, 2008). Even elderly adults without Alzheimer's Disease are inclined to make these errors (Tardiff, Bascandziev, Sandor, Carey, & Zaitchik, 2017), indicating that motion-based conceptions of life are pervasive across the lifespan and must be inhibited to reason about life as a metabolic process.

Early intuitions about matter also reemerge under cognitive load. Adults instructed to decide whether something is material or nonmaterial as quickly as possible will mistakenly classify gases and heft-less objects, like dust and snowflakes, as nonmaterial and mistakenly classify perceptible forms of energy, like rainbows and lightning, as material (Shtulman & Legare, 2019). Adults also make systematic mistakes in deciding whether an object will sink or float. An object's buoyancy is related to its density—a property that makes sense only if matter is composed of smaller particles. When adults are shown two balls of equal size, one made of wood and one made of lead, they judge that the wood ball is more likely to float than the lead one. But shown a large ball of wood and a small ball of lead, they take reliably longer to make the same judgment (Potvin & Cyr, 2017; Potvin, Masson, Lafortune, & Cyr, 2015).

Research over the past decade has revealed that this pattern is widespread (Shtulman & Lombrozo, 2016). Adults verify counterintuitive scientific ideas more slowly and less accurately than closely-matched intuitive ones in several domains, including astronomy, genetics, mechanics, thermodynamics, and evolution (Shtulman & Harrington, 2016; Shtulman & Valcarcel, 2012). And these effects been observed in several populations, including high schoolers (Babai, Sekal, & Stavy, 2010), undergraduate science majors (Foisy, Potvin, Riopel, & Masson, 2015), high school science teachers (Potvin & Cyr, 2017), and elderly adults (Barlev, Mermelstein, & German, 2018). Even professional physicists (Kelemen, Rottman, & Seston, 2013) and professional biologists (Goldberg & Thompson-Schill, 2009) exhibit cognitive conflict when reasoning about counterintuitive scientific ideas. Such conflict indicates that early intuitions about natural phenomena survives the acquisition of scientific knowledge in some form or another.

In previous research (Young, Laca, Dieffenbach, Hossain, Mann, & Shtulman, 2018), we sought to determine whether participants could be trained to verify counterintuitive scientific ideas more quickly and more accurately. We focused our investigation on statements about life and statements about matter. Some statements were intuitive (e.g., "bricks have weight," "goats need nutrients"), and others were counterintuitive (e.g., "dust has weight," "yeast needs nutrients"). Participants completed this task before and after a tutorial on the scientific properties of life or matter. The tutorials helped participants close the gap in accuracy between intuitive and counterintuitive statements within the domain of instruction but not the gap in latency. In other words, the tutorials were ineffective at reducing the immediate conflict elicited by counterintuitive statements (as indexed by response times), but they did help participants favor scientific responses over intuitive ones.

In the present study, we extended this line of research to elementary-school-aged children. Our motivation was threefold. First, children are in the earliest stages of learning science, and it's unclear whether their nascent scientific theories would pose a measurable challenge to their well-worn intuitive theories of the same phenomena. Second, any conflict that children experience between science and intuition may be more malleable than that experienced by adults, either because children's scientific theories are less developed (and thus more easily bolstered) or because their intuitive theories are less entrenched. Third, adapting our task for use with children may have pedagogical value if it proves to be an informative measure of early science learning or early scientific reasoning.

Our study followed the same protocol as Young et al. (2018), which included a pretest, a tutorial, and a posttest. At pretest, we expected children to show conflict between science and intuition, given that the children in our age range were beginning to learn about life and matter in school, but it was an open question whether that conflict would manifest itself in both response accuracy and response latency. Children might, for instance, verify counterintuitive statements less accurately than intuitive ones but show no difference in speed. At posttest, we expected children to verify counterintuitive statements more accurately within the domain of instruction, but it was an open question whether they would also verify those statements more quickly.

Method

Participants

Seventy-eight children in kindergarten through 6th grade participated. Their mean age was 8 years and 7 months, and they were approximately balanced for gender (37 female, 41 male). Children were recruited from public playgrounds and a children's museum, and they completed the study onsite.

Materials

Statement-Verification Task. We measured the conflict between science and intuition using a child-modified version of Shtulman and colleagues' statement-verification task. Children were presented with four types of scientific statements and asked to judge those statements as "true" or "false" as quickly as possible. Some statements were true from both a scientific perspective and an intuitive perspective ("tigers need nutrients"); some were false from both perspectives ("forks need nutrients"); some were true from a scientific perspective but false from an intuitive perspective ("bacteria need nutrients"), and some were false from a scientific perspective but true from an intuitive perspective ("fire needs nutrients"). The first two types of

statements will be referred to as *intuitive* and the latter two types as *counterintuitive*.

For each domain, statements were generated by pairing three predicates with 32 entities. In the domain of life, the predicates were "reproduces," "needs nutrients," and "grows and develops." In the domain of matter, the predicates were "has weight," "takes up space," and "is made of atoms." The biological predicates apply to all living things, but we predicted that children would be more inclined to apply them to entities that appear to move on their own. Likewise, the physical predicates apply to all material things, but we predicted that children would be more inclined to apply them to entities that can be seen or felt. These predictions were derived from prior work with adults (Young et al., 2018), as well as the extensive literatures on intuitive theories of life and matter referenced above.

We created the four types of statements by pairing predicates with four types of entities, as shown in Table 1. In the domain of life, those entities were animals (deemed alive by both science and intuition), inanimate artifacts and inanimate natural kinds (deemed alive by neither science nor intuition), plants and microorganisms (deemed alive by science but not by intuition), and animate natural kinds (deemed alive by intuition but not science). In the domain of matter, those entities were physical objects (deemed material by both science and intuition), abstract ideas (deemed material by neither science nor intuition), gases and other bulk-less or heft-less objects (deemed material by science but not by intuition), and the visible or tangible components of energy transfer (deemed material by intuition but not science).

Children completed the task on an iPad. Statements were displayed on the screen and children responded via touch screen. Twenty-two children opted into a version of the task that played audio recordings of the statements as they were displayed on the screen, thus supporting children who had difficulty reading independently. Audio recordings of each statement were generated via Apple's macOS text-to-speech engine. Children who listened to the audio-recorded stimuli received only four of the six predicates (randomly selected), due to the additional time required to play the recordings.

Tutorials. Children completed a tutorial on life or matter midway through the experiment. The tutorial on life emphasized that all living things need energy and nutrients, grow and develop, react to stimuli in their environment, and reproduce. It also addressed the misconception that life is synonymous with self-directed motion, providing examples of entities that do not appear to move on their own but are alive (e.g., moss) and entities that move on their own but are not alive (e.g., comets). The tutorial on matter emphasized that all matter occupies space, has weight, is made of atoms, and can undergo phase transitions. It also addressed the misconception that matter is synonymous with visibility or tangibility, providing examples of entities that cannot be seen or felt but are material (e.g., gases) and entities that can be seen or felt but are not material (e.g., lightning). Both tutorials contained a mixture of text, images, and videos and took approximately seven minutes to complete.

Table 1: Sample items used in the biological statements (top) and physical statements (bottom), organized by their role in scientific and intuitive views of the domain.

Is it alive?	Intuition: Yes	Intuition: No
Science: Yes	Rabbits Turtles Snails	Mushrooms Grass Bacteria
Science: No	Sun Wind Fire	Hammers Caves Shells

Is it matter?	Intuition: Yes	Intuition: No
Science: Yes	Bricks Ice Logs	Smoke Snowflakes Air
Science: No	Rainbows Shadows Heat	Dreams Songs Numbers

Procedure

Each study session proceeded in three phases. First, children verified 48 statements about life and 48 statements about matter (pretest). Next, they completed a tutorial on life or matter. Last, they verified 48 additional statements about life and 48 additional statements about matter (posttest). Children were randomized to tutorial condition—41 received the tutorial on life and 37 received the tutorial on matter.

Children completed the pretest and posttest in blocks. They saw a screen introducing a particular predicate (e.g., "Does it grow and develop?"), followed by 16 statements with that predicate (e.g., "Seaweed grows and develops"). Four of the statements were scientifically and intuitively true; four were scientifically and intuitively false; four were scientifically true but intuitively false; and four were scientifically false but intuitively true. The statements were randomly ordered within a block, and the blocks were randomly ordered within the testing phase, meaning that biological and physical predicates were intermixed. Children saw the same predicates at pretest and posttest, but those predicates were paired with 16 new entities. The entities presented at pretest for half the children were presented at posttest for the other half and vice versa. This variable was crossed with whether children received the tutorial on life or the tutorial on matter to ensure that the effects of the tutorial were not confounded with the effects of particular pretest or posttest items.

Results

The statement-verification task yielded two outcome measures: response accuracy and response latency. We analyzed each outcome with a linear mixed model (LMM), with statement type (intuitive or counterintuitive), test

(pretest or posttest), instruction (instructed or uninstructed), and their interactions as fixed effects and by-participant and by-predicate random effects. The response latency model additionally adjusted for whether children read or listened to the statements. Models with maximal random effects structures had convergence issues, and thus we followed the procedure recommended by Bates, Kliegl, Vasishth, and Baayen (2015) to guide removal of random effects that were not supported by the data. Inference for fixed effects was carried out via Type 3 likelihood ratio test (LRT) model comparison.

The present analyses collapse across tutorial domain (life or matter) for lack of space and focus instead on whether the statements were targeted by instruction or not. Children did verify biological statements more accurately than physical statements (87% vs. 75%). However, mean response latencies were similar across domains, as were the effects of the tutorial.

Finally, we did not consider age effects in the following analyses. In general, older children were more accurate, faster, and learned more from instruction. However, the overall pattern of reported results was similar across the age distribution of our sample.

Response Accuracy

As seen in Figure 1, there was an effect of statement type, such that children verified intuitive statements more accurately than counterintuitive statements, LRT $\chi^2(1) = 12.18$, $p < .001$. Overall, accuracy for intuitive statements was 18.4% greater than accuracy for counterintuitive statements, 95% CI [12.1, 24.7].

Additionally, there was a three-way interaction between statement type, test period, and instruction, LRT $\chi^2(1) = 25.17$, $p < .001$. We were specifically interested in children's response to instruction. In the instructed domain, children's posttest accuracy for counterintuitive statements was 11.9% greater than their pretest performance, 95% CI [9.4, 14.4]. However, pretest and posttest scores were similar for intuitive statements in the instructed domain and similar for both statements types in the uninstructed domain. Thus, instruction was effective at improving children's accuracy at verifying counterintuitive scientific ideas within the targeted domain.

Response Latency

Following prior research, we analyzed response latencies for correctly verified statements only. Before doing so, we first removed latencies shorter than 250 ms, as responses produced that quickly were unlikely to have been deliberate. Second, we calculated the mean response latency across participants and statements (M = 2743 ms) and removed latencies more than two standard deviations above the mean (i.e., latencies greater than 7565 ms). We then calculated the mean latency for each predicate, separating intuitive statements from counterintuitive statements and pretest statements from posttest statements.

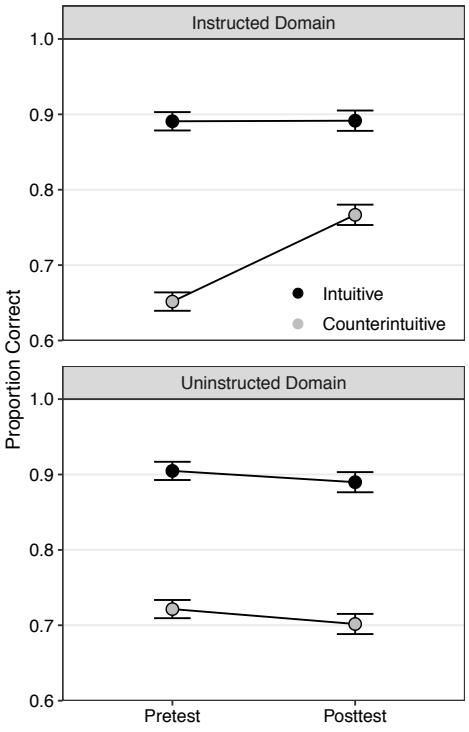


Figure 1: Estimated proportion of correct verifications by statement type, test, and instruction. Error bars represent standard errors.

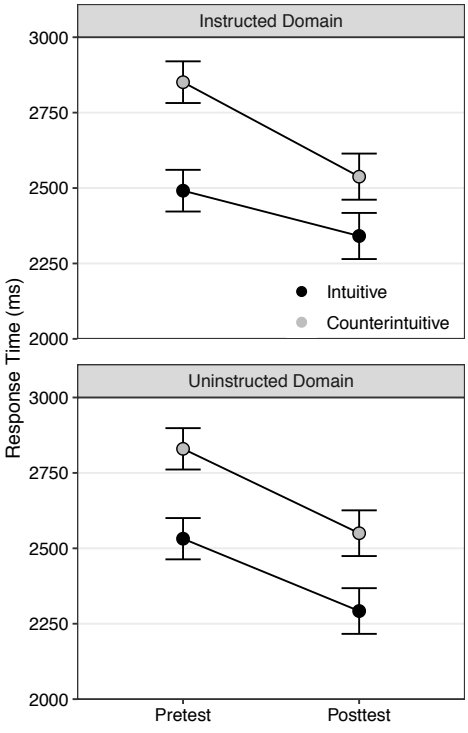


Figure 2: Estimated response latency for correct verifications by statement type, test, and instruction. Error bars represent standard errors.

As seen in Figure 2, there was an effect of statement type, such that children correctly verified counterintuitive statements more slowly than intuitive statements, LRT $\chi^2(1) = 90.16$, $p < .001$. Response latencies for counterintuitive statements were 278 ms slower than response latencies for intuitive statements, 95% CI [222, 335].

Additionally, there was an effect of test, such that children correctly verified statements faster at posttest than pretest, LRT $\chi^2(1) = 6.22$, $p = .013$. Response latencies at posttest were 246 ms faster than response latencies at pretest, 95% CI [70, 421]. We suspect this effect was due to increased familiarity with the task, as it did not vary by instruction and statement type (three-way interaction: LRT $\chi^2(1) = 1.14$, $p = .285$).

Discussion

Do elementary schoolers exhibit cognitive conflict when reasoning about counterintuitive scientific ideas? Our findings suggest they do. Children between ages five and twelve were slower and less accurate at verifying scientific statements that conflict with their intuitive theories of life or matter (e.g., "bacteria grow and develop," "steam is made of atoms") relative to closely-matched statements that accord with those theories (e.g., "tigers grow and develop," "rocks are made of atoms"). Instructing children on the scientific properties of life or matter increased their accuracy for counterintuitive statements in the instructed domain but not in the uninstructed domain. However, instruction did not reduce the gap in response latency between counterintuitive and intuitive statements, at least in comparison to the uninstructed domain. These findings indicate that children experience conflict between scientific ideas and intuitive ideas, despite limited exposure to science, but this conflict can be resolved in favor of scientific ideas with targeted instruction.

Our findings parallel those of Young et al. (2018), who administered the same task to adults. Adults were faster and more accurate overall, but both children and adults verified counterintuitive statements more slowly and less accurately than closely-matched intuitive statements. The effect of instruction was also similar across age groups, increasing participants' accuracy at verifying counterintuitive statements but not their speed. Thus, the same signatures of cognitive conflict observed in adults were observed in children ten to fifteen years younger.

Our findings accord with other findings on the speed and accuracy of children's scientific reasoning, documented by Vosniadou et al. (2018). Vosniadou and colleagues asked third- and fifth-graders to sort physical and biological items into one of two categories: a category that emphasized the item's intuitive features or a category that emphasized its scientific features. The categories were characterized by exemplars rather than by labels. For instance, on one trial participants decided whether water should be grouped with other liquids (coke, lemonade, milk) or with other forms of H_2O (ice, vapor, snow). Children of all ages preferred intuitive categories over scientific categories, and they took longer to make their judgments when they opted for the scientific category instead. Vosniadou et al.'s findings, like our findings, suggest that children must suppress an intuitive conception of the target item in order to endorse a competing scientific conception.

Vosniadou and colleagues did not administer a tutorial to their participants, and it's open question whether instructing participants on the scientific properties of the target items would change the nature of their categorizations. They did, however, measure executive function skills—namely, set-shifting ability and inhibitory control—and they found that children with higher executive function were more likely to categorize the target items by their scientific properties and were also faster to do so. Children with higher executive function have also been found to learn more from science instruction in the domain of vitalist biology (Bascandziev, Tardiff, Zaitchik, & Carey, 2018). Future research is needed to determine whether executive function plays a role in children's statement verifications as well. If it does, executive function tasks could be administered alongside our statement-verification task as a diagnostic for assessing young children's understanding of science and their receptiveness to science instruction.

One limitation of the current study is that we sampled children who had already begun learning the scientific properties of life and matter in school. Younger children (i.e., preschoolers) would likely show a different pattern of results. Without any scientific knowledge of life or matter, they should view statements like "bubbles have weight" and "dandelions need nutrients" as demonstrably false. Their accuracy for such statements would be lower, but their responses should be faster. Thus, in comparison to older children, younger children should show a larger gap in response accuracy between intuitive and counterintuitive statements but a smaller gap in response latency. And teaching preschoolers about the scientific properties of life or matter should increase the gap in latency, not reduce it. There are challenges, however, to adapting the task for use with preschoolers. Preschoolers are unlikely to know the meaning of terms like "atoms," "nutrients," and "reproduces," and the alternative terms they do know may not carry the same meaning. "Has babies," for example, may not be a substitute for "reproduces" because the offspring of plants, fungi, and bacteria are rarely referred to as "babies."

In conclusion, we have shown that tensions between science and intuition emerge early in the acquisition of scientific knowledge. While children can be taught to privilege scientific ideas over intuitive ones, the conflict between them—as manifested in slower response times for statements that elicit both ideas—appears to be immediate and robust.

Acknowledgements

We would like to thank the James S. McDonnell Foundation for supporting this research with an Understanding Human Cognition Scholar Award to Andrew Shtulman. We would

also like to thank Kidspace Children's Museum, Robin Pounders, and Claudia Lechner for their assistance with data collection.

References

Au, T. K. F., Chan, C. K., Chan, T. K., Cheung, M. W., Ho, J. Y., & Ip, G. W. (2008). Folkbiology meets microbiology: A study of conceptual and behavioral change. *Cognitive Psychology, 57*, 1-19.

Babai, R., Sekal, R., & Stavy, R. (2010). Persistence of the intuitive conception of living things in adolescence. *Journal of Science Education and Technology*, *19*, 20-26.

Barlev, M., Mermelstein, S., & German, T. C. (2018). Representational coexistence in the God concept: Core knowledge intuitions of God as a person are not revised by Christian theology despite lifelong experience. *Psychonomic Bulletin & Review, 25*, 2330-2338.

Bascandziev, I., Tardiff, N., Zaitchik, D., & Carey, S. (2018). The role of domain-general cognitive resources in children's construction of a vitalist theory of biology. *Cognitive Psychology, 104*, 1-28.

Bates, D., Kliegl, R., Vasishth, S., & Baayen, R. H. (2015). Parsimonious mixed models. arXiv:1506.04967.

Carey, S. (1985). *Conceptual change in childhood*. Cambridge, MA: MIT Press.

Carey, S. (1991). Knowledge acquisition: Enrichment or conceptual change. In S. Carey & R. Gelman (Eds.), *The epigenesis of mind: Essays in biology and cognition* (pp. 257–291). Hillsdale, NJ: Lawrence Erlbaum.

Carey, S. (2009). *The origin of concepts*. New York: Oxford University Press.

Chi, M. (1992). Conceptual change within and across ontological categories: Examples from learning and discovery in science. In R. Giere (Ed.), *Cognitive models of science* (pp. 129-186). Minneapolis, MN: University of Minnesota Press.

Foisy, L. M. B., Potvin, P., Riopel, M., & Masson, S. (2015). Is inhibition involved in overcoming a common physics misconception in mechanics? *Trends in Neuroscience and Education, 4*, 26-36.

Goldberg, R. F., & Thompson-Schill, S. L. (2009). Developmental "roots" in mature biological knowledge. *Psychological Science, 20*, 480-487.

Gopnik, A., & Wellman, H. M. (2012). Reconstructing constructivism: Causal models, Bayesian learning mechanisms, and the theory theory. *Psychological Bulletin, 138*, 1085-1108.

Hatano, G., & Inagaki, K. (1994). Young children's naive theory of biology. *Cognition*, *50*, 171-188.

Kelemen, D., Rottman, J., & Seston, R. (2013). Professional physical scientists display tenacious teleological tendencies: Purpose-based reasoning as a cognitive default. *Journal of Experimental Psychology: General*, *142*, 1074-1083.

McCloskey, M. (1983). Naïve theories of motion. In D. Gentner & A. L. Stevens (Eds.), *Mental models* (pp. 299-324). Hillsdale, NJ: Erlbaum.

Nakhleh, M. B., Samarapungavan, A., & Saglam, Y. (2005). Middle school students' beliefs about matter. *Journal of Research in Science Teaching, 42*, 581–612.

Nersessian, N. J. (1989). Conceptual change in science and in science education. *Synthese, 80*, 163-183.

Potvin, P., & Cyr, G. (2017). Toward a durable prevalence of scientific conceptions: Tracking the effects of two interfering misconceptions about buoyancy from preschoolers to science teachers. *Journal of Research in Science Teaching, 54*, 1121-1142.

Potvin, P., Masson, S., Lafortune, S., & Cyr, G. (2015). Persistence of the intuitive conception that heavier objects sink more: A reaction time study with different levels of interference. *International Journal of Science and Mathematics Education, 13*, 21-43.

Reiner, M., Slotta, J. D., Chi, M. T. H., & Resnick, L. B. (2000). Naïve physics reasoning: A commitment to substance-based conceptions. *Cognition and Instruction, 18*, 1-34.

Shtulman, A. (2017). *Scienceblind: Why our intuitive theories about the world are so often wrong*. New York: Basic Books.

Shtulman, A., & Harrington, K. (2016). Tensions between science and intuition across the lifespan. *Topics in Cognitive Science, 8*, 118-137.

Shtulman, A., & Legare, C. H. (2019). Competing explanations of competing explanations. *Manuscript under review*.

Shtulman, A., & Valcarcel, J. (2012). Scientific knowledge suppresses but does not supplant earlier intuitions. *Cognition, 124*, 209-215.

Shtulman, A., & Lombrozo, T. (2016). Bundles of contradiction: A coexistence view of conceptual change. In D. Barner & A. Baron (Eds.), *Core knowledge and conceptual change* (pp. 49-67). Oxford, UK: Oxford University Press.

Slaughter, V., & Lyons, M. (2003). Learning about life and death in early childhood. *Cognitive Psychology, 46*, 1-30.

Smith, C. L. (2007). Bootstrapping processes in the development of students' commonsense matter theories: Using analogical mappings, thought experiments, and learning to measure to promote conceptual restructuring. *Cognition and Instruction, 25*, 337-398.

Stavy, R., & Wax, N. (1989). Children's conceptions of plants as living things. *Human Development, 32*, 88-94.

Tardiff, N., Bascandziev, I., Sandor, K., Carey, S., & Zaitchik, D. (2017). Some consequences of normal aging for generating conceptual explanations: A case study of vitalist biology. *Cognitive Psychology, 95*, 145-163.

Trundle, K. C., Atwood, R. K., & Christopher, J. E. (2002). Preservice elementary teachers' conceptions of moon phases before and after instruction. *Journal of Research in Science Teaching, 39*, 633-658.

Vosniadou, S. (1994). Capturing and modeling the process of conceptual change. *Learning and Instruction, 4*, 45-69.

Vosniadou, S., Pnevmatikos, D., Makris, N., Lepenioti, D., Eikospentaki, K., Chountala, A., & Kyrianakis, G. (2018, in press). The recruitment of shifting and inhibition in online science and mathematics tasks. *Cognitive Science*.

Young, A., Laca, J., Dieffenbach, G., Hossain, E., Mann, D., & Shtulman, A. (2018). Can science beat out intuition? Increasing the accessibility of counterintuitive scientific ideas. *Proceedings of the 40th Annual Conference of the Cognitive Science Society*, 1238-1243.

Zaitchik, D., & Solomon, G. E. A. (2008). Animist thinking in the elderly and in patients with Alzheimer's disease. *Cognitive Neuropsychology, 25*, 27-37.

Perceived area plays a dominant role in visual quantity estimation

Sami R. Yousif (sami.yousif@yale.edu)[1]
Emma Alexandrov (ealexandrov@vassar.edu) [2]
Elizabeth Bennette (ehbennet@ucsd.edu) [3]
Frank C. Keil (frank.keil@yale.edu) [1]

[1]Department of Psychology, Yale University
New Haven, CT, 06520 USA
[2]Department of Cognitive Science, Vassar College
Poughkeepsie, NY, 12604 USA
[3]Department of Psychology, UCSD
La Jolla, CA, 92093 USA

Abstract

Many studies have investigated the roles that area and number play in visual quantity estimation. Yet, recent work has shown that *perceived* area is not equal to true, mathematical area. This simple fact calls into question many findings in numerical cognition and suggests a new theoretical perspective: that area estimation plays a dominant role in visual quantity estimation. We examine two 'case studies': (1) a 'general magnitude' account of visual quantity estimation, which posits bi-directional influences between area and number. In contrast with prior work, controlling for perceived area reveals a unidirectional relation between area and number (Experiments 1 and 2), and (2) acuity of area and number estimation (Experiment 3). We show how an understanding of the perception of area forces a reevaluation of several findings concerning the relative acuity of number and area estimation. Combined, and in contrast to many prior studies, our findings suggest a dominant role of area in visual quantity estimation.

Keywords: approximate number, number, area, perception

Introduction

The ability of human adults, infants, and nonhuman animals to rapidly approximate large numbers is a cornerstone of research on numerical cognition. This propensity supposedly relies on an evolutionary ancient system -- the Approximate Number System -- which serves as a foundation for downstream numerical and mathematical ability (Cantlon & Brannon, 2007; Dehaene, 1997; Feigenson et al., 2004; Xu & Spelke, 2000).

Yet this widely accepted notion also raises questions: in our evolutionary environment, how often would number have been the most relevant cue for approximating quantity? Area perception rather than number perception would seem to have been prioritized evolutionarily: if foraging for food, for example, would you prefer to have 100 berries, or 50 berries four times in volume? Nevertheless, approximate area has been vastly understudied relative to approximate number (but see Brannon et al., 2006; Lourenco et al., 2012; Odic et al., 2013). In hundreds of studies, numerosity is assumed to be perceived independently of area (and other continuous dimensions; e.g., average size, density, or convex hull), thereby relegating area manipulations to little more than pesky control conditions in 'bigger' questions about number.

However, visual area approximation has recently emerged as an ability in its own right. Recent work has revealed that the visual approximation of area is guided by a cue other than area (Yousif & Keil, 2019). Instead, visual approximations of area are roughly equivalent to the sum of objects' dimensions rather than their product, resulting in potentially large distortions of perceived space. This continues to be true after accounting for confounds such as numerosity and perimeter. This phenomenon is known as the 'Additive Area Heuristic' (AAH).

An area estimation heuristic raises questions about the relation between area and number. While numerous papers have documented bidirectional 'congruity effects' between area and number (e.g., Hurewitz et al., 2006, Walsh, 2003), perceived area (per the AAH) may not be influenced by numerosity; these past results may arise because of a confound between *perceived* area and numerosity (Yousif & Keil, 2019). Only when unconfounded is it possible to understand the relation between number and area in visual quantity estimation.

The AAH calls into question many other findings in the field of numerical cognition, raising the possibility that many of them can also be explained by a failure to account for perceived area. For example: if numerosity does not influence the perception of area, does the perception of area influence the perception of numerosity? Though this question has been asked before (e.g., Hurewitz et al., 2006), it has operated under a false premise: that true, mathematical area accurately reflects the percept of area. Thus, to the extent that area perception is best captured not by mathematical area but by some other means (e.g., the AAH), this question ought to be revisited.

If perceived area is dissociable from mathematical area, it suggests a reinterpretation – and, in some cases, a reexamination – of many prior findings. The present work explores the relation between number and *perceived* area in the context of two 'case studies': (1) a 'general magnitude' account of number and area, and (2) relative area and

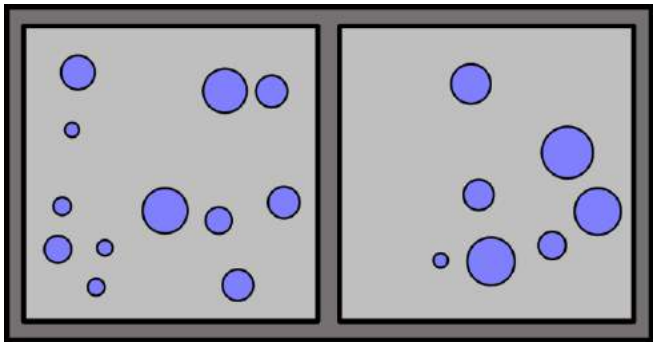


Figure 1. An example display for Experiments 1-3. Most observers report that the left panel is greater in area, despite the fact that the two are equal in true area. However, the left panel is greater in 'Additive Area' (which causes the illusion).

number estimation acuity. In both cases, we demonstrate that accounting for perceived area reveals a qualitatively different pattern from what has been previously observed.

The current study

In a first experiment, we assess the 'general magnitude' account of number and area approximation by examining how increased 'Additive Area' (AA) affects numerosity estimation. To do so, we manipulate AA while number is held constant. Most work has suggested bidirectional interactions between area and number (e.g., Hurewitz et al., 2006), but recent work has shown that manipulating number *does not* influence perceived area (Yousif & Keil, 2019). Here, we show that this relation is in fact unidirectional in that perceived area influences number judgments to a large extent. In a second experiment, we follow up on this by pitting AA and number against each other in a maximally implicit design, by having one group of observers make area judgments and another group of observers make number judgments on the exact same stimuli. Again, we demonstrate influences of area on number perception. In a third experiment, we assess number estimation acuity under different conditions (e.g., controlling AA vs. true, mathematical area). Number acuity appears to differ dramatically depending on how area is controlled.

Experiment 1: Area influences number

Mimicking a design in prior work (Yousif & Keil, 2019), we created stimuli for which additive area, mathematical area (MA), and number could be manipulated independently. AA is used as a proxy for perceived area, given the prior work showing that AA captures perceived area more accurately than MA. Observers viewed two stimuli side-by-side and were simply asked to indicate which was greater in number.

Method

Participants 100 observers were recruited via Amazon Mechanical Turk. Observers were excluded if and only if they began but did not complete the task (5 observers). All observers consented prior to participation, and these studies were approved by the IRB at Yale University.

Stimuli All of the stimuli were generated via custom software written in Python with the PsychoPy libraries (Peirce, 2007). The aim was to create pairs of stimuli that varied in either AA, MA, or number while the other values were equated. For each stimulus pair, we randomly generated an initial set of discs (ranging from 20 pixels to 100 pixels in diameter, with a buffer of at least 10 pixels between any two discs), then pseudo-randomly generated a second set of objects based on a given AA/MA/Number ratio (specific values varied for each experiment; see, e.g., Table 1). The displays always had between 20 and 26 discs (the initial set always having 20). Stimulus pairs were generated randomly until a pair met both the AA, MA, and number criteria, at which point that pair would be rendered another time and saved. The second stimulus always had more area (whether AA or MA) than the initial stimulus. For the details of how AA, MA, and number covaried, see Table 1. All discs were rendered with a thin, black border (4-pixel stroke width). The images depicted in Figure 1 are representative of those used in the experiment.

Procedure The task itself was administered online via Amazon Mechanical Turk, using custom software. On each

Number Ratio	Type	AA Ratio	MA Ratio
1.00	**AA Constant**	1.00	1.00
		1.00	1.05
		1.00	1.10
	MA Constant	1.00	1.10
		1.05	1.10
		1.10	1.10
1.15	**AA Constant**	1.15	1.05
		1.15	1.10
		1.15	1.15
	MA Constant	1.10	1.10
		1.15	1.10
		1.20	1.10
1.30	**AA Constant**	1.20	1.10
		1.20	1.15
		1.20	1.20
	MA Constant	1.15	1.15
		1.20	1.15
		1.25	1.15

Table 1. The number, AA, and MA ratios for Experiment 1.

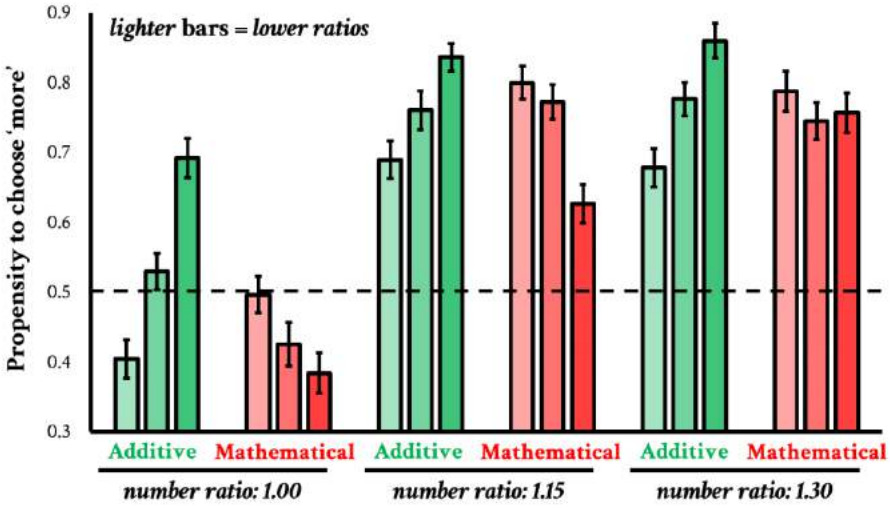


Figure 2. Results from Experiment 1. Three number ratios are represented along the x-axis. Green bars represent MA-controlled sets, where AA varied in three steps. Red bars represent AA-controlled sets, where MA varied in three steps. Lighter bars represent lower ratios. E.g., for the leftmost set of green bars, the lightest bar represents the lowest AA ratio and the darkest bar represents the highest AA ratio. Error bars represent +/- 1 SE. The dashed line represents chance performance.

trial, observers saw two spatially separated sets of lavender-colored dots, presented side-by-side in the center of the screen, with 50 pixels of space in between (see Figure 1). Each stimulus was 400 pixels by 400 pixels. The stimuli were always counterbalanced so that an equal number containing more AA, MA, or number appeared on each side of the screen. Observers were instructed to press 'q' if the image on the left had more cumulative number, and 'p' if the image on the right had more cumulative number. They were also given an additional, explicit warning to respond according to number regardless of area. The stimuli stayed on the screen for 700ms, but there was no time limit on responses. Between each trial, there was a 1000ms ITI. Observers completed 72 trials. All trials were presented in a unique random order for each participant. Observers completed two representative practice trials before beginning the actual task.

Results and Discussion

The results of Experiment 1 are shown in Figure 2. An ANOVA revealed a main effect of numerosity, confirming that observers were able to discriminate on the basis of numerosity, $F(2,93)=149.65$, $p<.001$, $\eta_p^2=.61$. Further, increased MA generally *decreased* the probability that an observer would select a stimulus as more numerous $F(2,93)=12.78$, $p<.001$, $\eta_p^2=.12$. Yet, critically, increased AA *did* increase the likelihood that observers would indicate a stimulus was more numerous, $F(2,93)=49.08$, $p<.001$, $\eta_p^2=.34$ (and this pattern was observed across all ratios, as can be seen in Figure 2). Note that this is in stark contrast to other results showing that changes in numerosity *do not* influence area judgments (Yousif & Keil, 2019).

These results (in combination with prior results) suggest a relation between perceived area and perceived number – but one that is unidirectional (i.e., perceived area influences number, but not vice versa). In contrast to a 'general magnitude' account, which predicts positive relations between various magnitudes, the present results suggest area may play a dominant role in quantity estimation. However, these results do not reveal the extent to which number is perceived independently of AA. The following experiments aim to address that question.

Experiment 2: Number versus area

To understand whether the results of Experiment 1 could be explained by a General Magnitude account (e.g., a 'more-is-more' heuristic), we directly pitted AA and number against each other in a between-subjects experiment. In this way, we can directly assess the effect of increased area on number perception and vice versa. Borrowing from previous work which dissociated AA and MA (Yousif & Keil, 2019we manipulated both AA and number while holding the other constant. In one condition, observers made area judgments; in another condition, a separate group of observers made number judgments.

Method

Participants 200 observers were recruited via Amazon Mechanical Turk (100 for each condition). Observers were excluded if and only if they began but did not complete the task (3 observers, all in the area condition). All observers consented prior to participation, and these studies were approved by the IRB at Yale University.

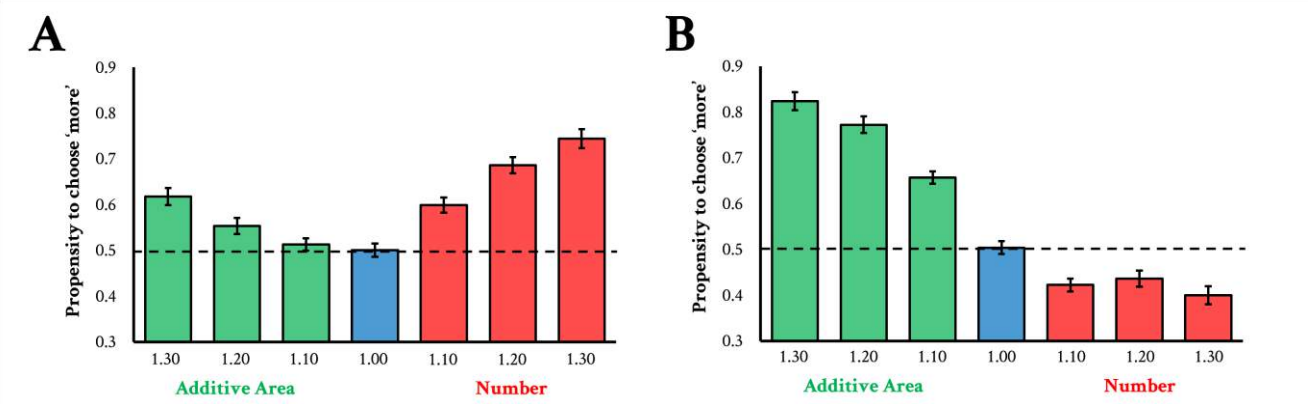


Figure 3. Results from number discriminations (A) and area discriminations (B) in Experiment 2. The green bars represent trials where AA varied (in a 1.1, 1.2, or 1.3 ratio) but number was held constant, while the red bars represent trials where number varied (in a 1.1, 1.2, or 1.3 ratio) while AA was held constant. The y-axis represents the propensity to choose 'more', whether that be more number or more area. Error bars represent +/- 1 SE. The dashed line represents chance performance.

Stimuli The stimuli for this experiment were generated in the same way as those of the prior experiment. The same stimuli were used for each condition. There were seven ratios: three in which number varied (in a 1.1, 1.2, and 1.3 ratio) while AA was held constant, three in which AA varied (in a 1.1, 1.2, and 1.3 ratio) while number was held constant, and one in which both were held constant (to serve as a baseline).

Procedure The procedure is identical to Experiment 1 except that observers completed 84 trials instead of 72. For the number judgment condition, the instructions were the same. For the area judgment condition, observers were told the following: "Your task is simply to indicate which set of circles has **more cumulative area**. In other words: if you printed the images out on a sheet of paper, which would require more total ink?" Later, they were told: "The sets of dots will sometimes vary in number, but the number of dots does not matter. Instead, you should answer only which has more area, regardless of number."

Results and Discussion

The results of the number discrimination condition are shown in Figure 3a. Observers indicated that images containing more discs were more numerous ($t(96)=11.85$, $p<.001$, $d=1.20$). However, observers also indicated that images with greater perceived area (but were equal in number) were more numerous ($t(96)=5.35$, $p<.001$, $d=.54$). In other words, it appears that the perception of area affects the perception of numerosity.

The results of the area discrimination condition are shown in Figure 3b. Observers indicated that images greater in AA were greater in perceived area ($t(96)=17.60$, $p<.001$, $d=1.76$). However, observers were slightly below chance when selecting between displays equal in AA but differing in numerosity ($t(96)=5.81$, $p<.001$, $d=.58$). Thus, all else equal, observers judged displays with more number to have *less* area – replicating the findings of recent work (Yousif & Keil, 2019) but in stark contrast to many existing studies (e.g., Hurewitz et al., 2006).

These results suggest three primary conclusions. First, the results of the number discrimination condition cannot be explained by a response bias to simply pick the image with 'more' on some dimension. Indeed, observers indicated that displays with more number appeared to have less cumulative area. Second, this experiment provides converging evidence with Experiment 1 that perceived area influences perceived numerosity (i.e., people confuse 'more' perceived area for 'more' number). Third, and critically, this experiment shows that number does *not* influence perceived area. This indicates a unidirectional relation between perceived area and number (in contrast to views that posit bidirectional interactions between these domains of magnitude; e.g., Walsh, 2003). There *is* an effect of number on area (such that more number is related to less perceived area) – but our findings challenge a general magnitude account, and are contrary to prior work (e.g., Hurewitz et al., 2006).

Experiment 3: Number and area acuity

A third experiment assessed number discrimination acuity (i.e., the level of precision with which observers can discriminate two non-symbolic numerosities) in a more traditional number acuity task, while controlling for either AA or MA. We predicted that performance will be lower when AA is controlled. The goal of this study is to ascertain whether there is a 'true' number discrimination acuity (or area discrimination acuity, for that matter), as this would bear on studies that have tried to interpret relative acuity in each domain (e.g., Lourenco et al., 2012; Odic et al., 2013).

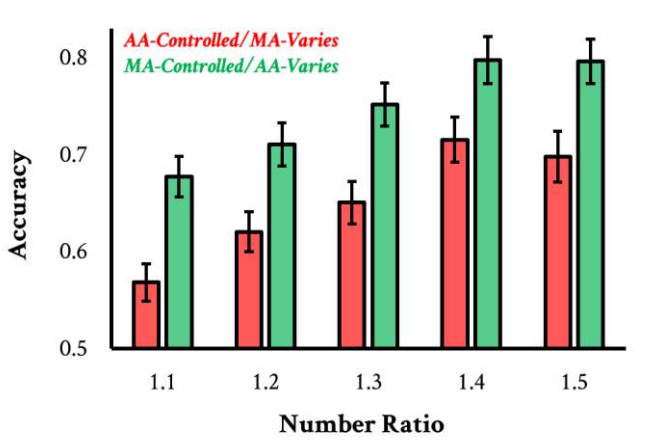


Figure 4. Results from Experiment 3. Five number ratios are represented along the x-axis. Green bars represent MA-controlled sets, where AA varied. Red bars represent AA-controlled sets, where MA varied in three steps. The y-axis represents accuracy for number discriminations, i.e., the proportion of time observers chose the display that was more numerous. Error bars represent +/- 1 SE. The x-axis corresponds to chance performance.

Method

All elements of the experimental design were identical to those of Experiment 1, except as stated below. 80 new observers were tested via Amazon Mechanical Turk. One observer was excluded for failing to complete the task. Observers completed number discriminations at five distinct ratios: 1.10, 1.20, 1.30, 1.40, and 1.50. Half the trials were controlled for AA, and the other half of trials were controlled for MA (while allowing the other dimension to vary). The displays always had between 10 and 30 discs (the initial set having 10 half the time, and 20 the other half of the time). Observers completed 80 trials.

Results and Discussion

The results of Experiment 3 are displayed in Figure 4. Accuracy was indeed lower for the AA-controlled trials, $t(79)=6.97$, $p<.001$, $d=.79$, and this was independently true for each number ratio ($ps<.002$). Of the 80 observers tested, 66 were as good or better at discriminating number in the MA-controlled condition (where AA varied; $p<.001$). Critically, performance across the two different area controls was highly correlated $r(78)=.69$, $p<.001$ – about as highly as performance in each condition was to itself (MA-control: $r=.66$; AA-control: $r=.65$).

Once again, differences in perceived area strongly influenced perceived number. While prior work has made conclusions on the basis of relative acuity (e.g., Lourenco et al., 2012; Odic et al., 2013), these results suggest that comparing acuity across dimensions should be interpreted with caution. In other words: what is 'true' number acuity, if number acuity varies so greatly across different area controls? This is especially relevant for developmental studies which make claims about relative acuity across development (e.g., Odic et al., 2013).

General Discussion

Our first two experiments demonstrate that accounting for perceived area challenges our understanding of the relation between area and number. In particular, we have shown an apparent unidirectional relation between area and number such that area influences number judgments but not the other way around. This contrasts with work documenting a bidirectional relation and forces a reconsideration of the roles of area and number in quantity estimation.

In addition, we have shown how accounting for perceived area challenges our understanding of area and number *acuity*. In particular, number discrimination acuity appears to vary dramatically depending on whether AA or MA is controlled (as revealed explicitly in Experiment 3, but also evident in the results of Experiment 1). This raises questions about prior studies that have interpreted the relative acuity of area and number discriminations (e.g., Lourenco et al., 2012; Odic et al., 2013).

Conclusion: is number special?

Is number special in visual processing? The answer to this question seems obvious: the field of numerical cognition is perhaps one of the largest and most prominent in all of cognitive science, and the ability to discriminate visual number is often thought to be the foundation of our 'core' mathematical competency (Feigenson et al., 2004). Yet, this seemingly obvious conclusion is not evident from first principles. In what evolutionary context would an approximate number system have been more critical for survival than approximate area or volume? Few plausible examples come to mind.

Our studies do not ask whether number is special *somewhere* in the mind. Instead, the question is whether number is special *visually* – or even whether, as more extreme views have suggested, it is a visual feature (like color or orientation; e.g., Anobile et al., 2016; Burr & Ross, 2008). This question has been heavily discussed (e.g., Durgin, 2008; Leibovich et al., 2017). Yet this debate, here and elsewhere, has been plagued by the use of artificial stimuli with a seemingly unbounded number of possible confounds. How can one hope to isolate numbers amidst the continuous dimensions of area, perimeter, convex hull, density, average element size, variance in element size, variability in inter-dot distance, etc. (some of which are often negatively correlated with one another)? This list is only a small subset of all the continuous cues that *may* be related to the perception of number.

The present work is not immune to such confounds. However, our studies do provide clear predictions about a particular cue, AA, (rather than a collection of them) and its relation to numerosity. This prediction is borne out of the theoretical position that visual number estimation is unlikely to have been prioritized in evolution. More consequentially, we find clear influences of area on number, but not the other way around.

What should be said, then, about the perception of number? We have presented evidence for area playing a

dominant role in quantity estimation, automatically and irresistibly influencing the estimation of number. Yet, number discrimination ability across very different displays (i.e., displays controlled for either AA or MA), is highly correlated – suggesting that number estimation cannot be explained by perceived area (or by some superficial strategy that operates differently over different sets of stimuli). Thus, while the human visual system is clearly able to extract number, it does not seem to be wired to do so first and foremost. Indeed, area may play the leading role in quantity estimation. This also suggests that number may not be a true visual feature as has been claimed (see Burr & Ross, 2008).

Across several paradigms and stimuli configurations, one salient pattern consistently emerges: area influences number approximation but not the other way around. This is a fundamentally different pattern from what has been observed in tasks that do not control for AA, and these findings offer a new theoretical perspective on the relation between number and area in vision: that number may not be so special after all.

Acknowledgments

For helpful comments and conversation, we thank Brynn Sherman and all the members of the Yale Cognition and Development Lab.

References

Anobile, G., Cicchini, G. M., & Burr, D. C. (2016). Number as a primary perceptual attribute: A review. *Perception*, *45*, 5-31.

Burr, D., & Ross, J. (2008). A visual sense of number. *Current Biology*, *18*, 425-428.

Brannon, E. M., Lutz, D., & Cordes, S. (2006). The development of area discrimination and its implications for number representation in infancy. *Developmental Science*, *9*, F59-F64.

Cantlon, J. F., & Brannon, E. M. (2007). How much does number matter to a monkey (Macaca mulatta)? *Journal of Experimental Psychology: Animal Behavior Processes*, *33*, 32.

Dehaene, S. (1997). The number sense: how the mind creates mathematics. New York: Oxford University Press.

Durgin, F. H. (2008). Texture density adaptation and visual number revisited. *Current Biology*, *18*, R855-R856.

Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in Cognitive Sciences*, *8*, 307-314.

Hurewitz, F., Gelman, R., & Schnitzer, B. (2006). Sometimes area counts more than number. *Proceedings of the National Academy of Sciences*, *103*, 19599-19604.

Leibovich, T., Katzin, N., Harel, M., & Henik, A. (2017). From "sense of number" to "sense of magnitude": The role of continuous magnitudes in numerical cognition. *Behavioral and Brain Sciences*, *e164*, 1-62.

Lourenco, S. F., Bonny, J. W., Fernandez, E. P., & Rao, S. (2012). Nonsymbolic number and cumulative area representations contribute shared and unique variance to symbolic math competence. *Proceedings of the National Academy of Sciences*, *109*, 18737-18742.

Odic, D., Libertus, M. E., Feigenson, L., & Halberda, J. (2013). Developmental change in the acuity of approximate number and area representations. *Developmental Psychology*, *49*, 1103-1112.

Peirce, J. W. (2007). PsychoPy—psychophysics software in Python. *Journal of Neuroscience Methods*, *162*, 8-13.

Xu, F., & Spelke, E. S. (2000). Large number discrimination in 6-month-old infants. *Cognition*, *74*, B1-B11.

Yousif, S. R., & Keil, F. C. (2019). The 'Additive Area Heuristic': An efficient but illusory means of visual area estimation. *Psychological Science*, *30*, 495–503.

Statistical learning generates implicit conjunctive predictions

Ru Qi Yu (ruqiyu@psych.ubc.ca)
Department of Psychology, University of British Columbia

Jiaying Zhao (jiayingz@psych.ubc.ca)
Department of Psychology, and Institute for Resources, Environment and Sustainability, University of British Columbia

Abstract

The cognitive system readily detects statistical relationships where the presence of an object predicts a specific outcome. What is less known is how the mind generates predictions when multiple objects predicting different outcomes are present simultaneously. Here we examine the rules with which predictions are made in the presence of two objects that are associated with two distinct outcomes. In three experiments, participants first implicitly learned that an object predicted a specific target location in a visual search task. When two objects predicting two different target locations were present simultaneously, participants were reliably faster to find the target when it appeared in the conjunctive location than in disjunctive locations. This was true even if participants were not consciously aware of the association between the objects and target locations. The results suggest that in the presence of multiple predictors, statistical learning generates implicit expectations about the outcomes in a conjunctive fashion.

Keywords: Implicit learning, regularities, conjunctive inference, visual search, attention

Introduction

The visual environment contains widespread regularities in terms of co-occurrences between individual objects or events over time. For example, the red light turns on after the yellow light at traffic intersections, and thunder follows lightening in a thunderstorm. The mind can detect such regularities effortlessly, automatically, or even outside of conscious awareness. One form of extracting these regularities (i.e., A predicts B) is statistical learning, which involves the detection of statistical relationships among individual objects over space or time (Fiser & Aslin, 2001; Saffran, Aslin, & Newport, 1996; Turk-Browne, Jungé, & Scholl, 2005). Statistical learning occurs incidentally to ongoing tasks and quickly after a few exposures to the regularities (Turk-Browne, Scholl, Johnson, & Chun, 2010), and proceeds without explicit awareness or conscious intent (Baker, Olson, & Behrmann, 2004).

The implicit extraction of regularities has a number of consequences on the representations of the individual objects that comprise the regularities. Recent studies suggest that statistical learning spontaneously biases attention to the co-occurring objects in a persistent manner (Zhao, Al-Aidroos, & Turk-Browne, 2013; Yu & Zhao, 2015), interferes with summary perception (Hall, Mattingley, & Dux, 2015; Zhao, Ngo, McKendrick, & Turk-Browne, 2011), updates object representations (Yu & Zhao, 2018a, Yu & Zhao, 2018b), facilitates the compression of information in working memory (Brady, Konkle, & Alvarez, 2009; Zhao & Yu, 2016), and leads to automatic transitive inferences (Luo & Zhao, 2018).

To date, research on statistical learning has predominately focused on the relationship between individual objects or events. However, in the broader visual environment different objects or events are often present at the same time where each predicts a specific outcome. For example, excessive smoking can lead to cardiovascular problems as well as lung complications, while excessive alcohol consumption can lead to similar cardiovascular problems and also potential brain damage. When excessive smoking occurs with excessive drinking, what consequences would follow? In this example, a conjunctive inference would generate an expectation that satisfies both predictors (i.e., cardiovascular problems), whereas a disjunctive inference would generate an expectation that satisfies either one of the two predictors (i.e., cardiovascular problems, lung complications, and potential brain damage). When people are presented with both predictors at the same time, what kind of inference do they make automatically (Mendelson, 2009)? Understanding automatic conjunctive or disjunctive inferences can help illuminate reasoning biases such as the conjunction fallacy where people mistakenly judge a conjunctive statement to be more probable than a disjunctive statement (Tversky & Kahneman, 1983).

In the current study, we examine the rules with which predictions are made in the presence of two objects that are associated with two distinct outcomes. In a visual search paradigm, participants first viewed one color circle and then searched for a target (a rotated T) in an array during the exposure phase. Each color predicted a specific location of the target in the array. For example, after a blue circle the target would always appear in the top half of the array; and after a red circle the target would always appear in the left half of the array. The question is: Where was the target expected to appear when both the blue circle and the red circle were present at the same time? A conjunctive prediction would suggest that the target was expected to appear in the top left quadrant of the array, whereas a disjunctive prediction would suggest that the target was expected to appear in the top half or the left half of the array. Importantly, at the inference phase when both color circles were present, the target was equally likely to appear in any quadrant of the array. We used response time of target search during the inference phase to gauge in which location the target was expected to appear.

In Experiments 1 and 2, we found that participants were reliably faster to find the target when it appeared in the conjunctive quadrant than in the disjunctive quadrant. This was true even if participants were not consciously aware of the association between the color circles and target locations during debriefing. We further replicated the finding in Experiment 3 where the two predictors were two feature dimensions in one object. This effect was equally strong whether participants implicitly learned the association or were explicitly told about the association.

Experiment 1

This experiment examined which type of inference participants would make when they saw a pair of colors, each predicting a different half of the array.

Participants

A total of 120 students (81 female, mean age=20.0 years, SD=2.3) from the University of British Columbia (UBC) participated for course credit.

Stimuli

For each trial in the experiment, participants saw one colored circle first, followed by a search array (Figure 1). The color circle could appear in one of four colors (R/G/B): red (255/0/0), yellow (255/255/0), blue (0/0/255), or grey (192/192/192). Each circle subtended 2.2° of visual angle. For each search array following the circle, 16 objects were presented in an invisible 8-by-8 grid. Each cell in the grid subtended 1.7° of visual angle. The 8-by-8 grid was divided into four 4-by-4 quadrants, where each quadrant was separated from the adjacent two quadrants by 2.2° of visual angle. Each quadrant contained four objects, where no row or column in the quadrant could be empty.

Out the 16 objects in each array, 15 were distractors in "L" shapes, randomly pointing to the left or right. There was only one target in each array, which was a rotated "T", randomly determined to be pointed to the left or right. Participants were asked to find the target "T" and indicate which direction the "T" was pointing (left or right) by pressing a key on the keyboard, as quickly and accurately as possible.

For each trial, the color circle was presented on the screen for 1000ms. Followed by a 1000ms blank screen, the search array appeared on the screen until response. There was a 1000ms blank screen interval between trials.

Procedure

Participants first completed the exposure phase (Figure 1). During exposure, one color circle appeared on the screen at a time followed by a visual search array. Each of the four colors was presented for 40 times during exposure, resulting in a total of 160 trials (the order of the trials was random). Each color predicted that the target "T" in the search array always appeared in a unique half of the array (the top, left, bottom, or right half). For example, after the blue circle, the target always appeared in the top half of the array. After the red circle, the target always appeared in the left half of the array. The target location within each half of the array was counter-balanced between the two quadrants (e.g., counterbalanced between top-left and top-right quadrants for the top half of the array), and the target location within each quadrant was randomly determined. The color-location associations were randomly determined for each participant but remained fixed throughout the experiment for the participant.

We wanted to examine whether there were differences in conjunctive inferences made from explicit knowledge versus incidentally learned predictions. Therefore, half of the participants (N=60) were randomly selected to be explicitly told about the associations between colors and target locations before exposure (explicit condition), and the other half were told to only pay attention to the color circle and search for the target (implicit condition).

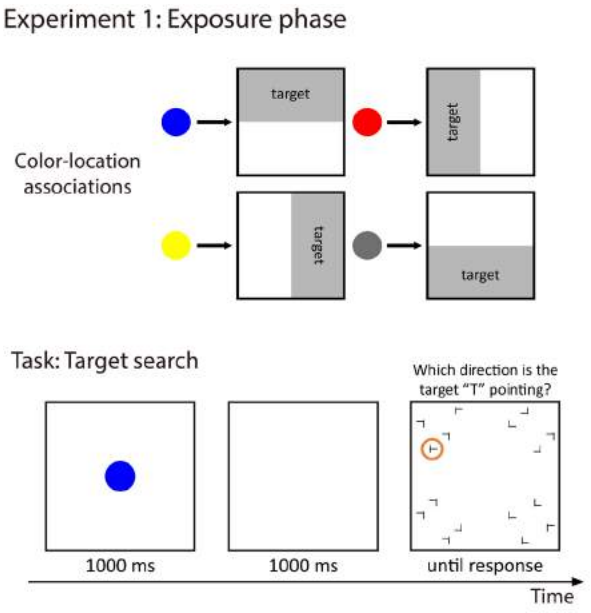


Figure 1. Experiment 1 exposure phase. Each color circle predicted the location of the target in the subsequent search array. In the visual search task, participants saw the color circle first, and then searched for a target (the rotated "T") and judged the direction of target as quickly and accurately as possible.

After exposure, participants completed the inference phase (Figure 2). During this phase, two color circles were presented at the same time in each trial, followed by a search array. There were six unique color pairs. Each color pair and the following search array were presented four times in the inference phase in a random order, resulting in 24 trials in total. In each trial, the target appeared in any of the four quadrants with equal probability (the top-left, top-right, bottom-left, and bottom right quadrant). The location of the target within the quadrant was randomly determined.

Since the target now appeared in the four quadrants with equal probability, faster response time in target search in a given quadrant would indicate that the participant prioritized that quadrant for target search. This would mean that the

participant expected that the target would appear in that quadrant, suggesting a prediction of where the target would appear after seeing the two color circles.

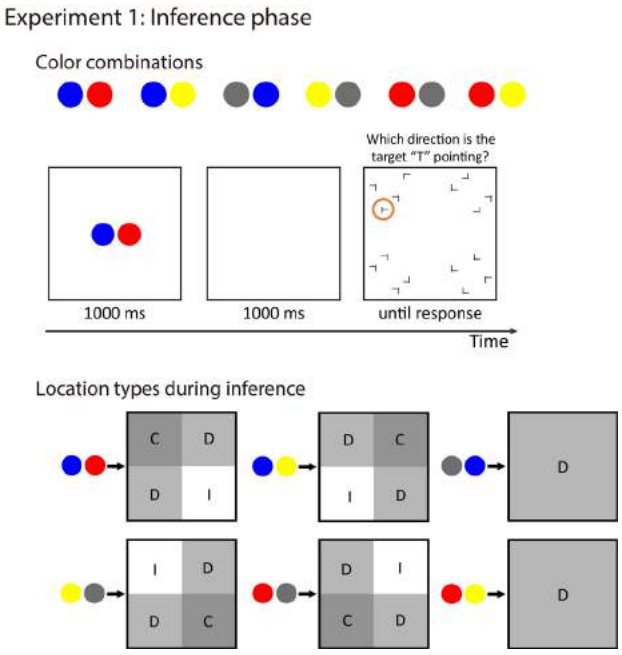


Figure 2. Experiment 1 inference phase. The four colors were combined into six color pairs. The pairs were presented first, followed by a search array. The target appeared in all four quadrants with equal probability following each pair. Based on the color-location associations during exposure, there were four types of target location following each color pair. These include the locations consistent with a conjunctive inference (C), locations consistent with a disjunctive inference (D), and the impossible locations (I).

During the inference phase, the two color circles were presented next to each other horizontally or vertically (randomly determined), and the order of the two colors for each pair was counter-balanced. Based on the color-location associations during exposure, there were four types of target location following each pair: locations consistent with a conjunctive inference (C), locations consistent with a disjunctive inference (D), and the impossible locations where the target would never appear based on the prior color-location associations (I). In both the explicit and implicit conditions, participants were only told that they would now see two color circles appearing simultaneously on the screen before each search array, and they were asked to search for the target as in the exposure phase.

After the inference phase, participants in the implicit condition also completed a test phase to probe their awareness of the color-location associations. They were asked where the target would appear (the top, left, bottom, and right half of the array) after seeing each of the four colors, so guessing would result in an accuracy of 0.25 in the test phase.

Results and Discussion

The test phase accuracy for participants in the implicit condition was 0.51, reliably above chance [chance=0.25, p<.001], indicating that participants in the implicit condition have successfully learned the color-location associations.

We then analyzed the responses time (RT) of correct trials in the inference phase to see what type of inferences participants made when they saw the color pairs. We grouped the trials in the inference phase into four types: conjunction, disjunction (2 quadrants vs. 4 quadrants), and impossible. Take the blue and red pair, the blue circle previously predicted that the target would appear in the top half of the array and the red circle previously predicted that the target would appear in the left half of the array. This means that the top left quadrant was the conjunctive quadrant, the top right and the bottom left quadrants were the disjunctive quadrants, and the bottom right quadrant was the impossible quadrant. For example, faster RT in the conjunctive quadrant would indicate that participants expected the target would appear in that quadrant, suggesting a conjunctive prediction. We plotted the RT in each type of quadrant in the inference phase (Figure 3).

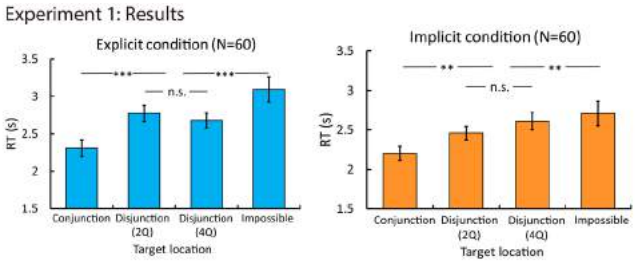


Figure 3: Experiment 1 results. The response time (RT) for each type of trials was graphed separately for the explicit and implicit conditions (Error bar reflect ± 1 SE; **p<.01, ***p<.001).

A 2 (condition: explicit vs. implicit, between-subjects) × 4 (trial type: conjunctive, 2-quadrant disjunctive, 4-quadrant disjunctive, and impossible quadrant, within-subjects) mixed-design ANOVA revealed a significant main effect of trial type [F(3,354)=16.04, p<.001, η_p^2=0.12], but no main effect of condition [F(1,118)=3.27, p=.07, η_p^2=0.03], or interaction [F(3,354)=1.29, p=.28, η_p^2=0.01]. This suggests that participants attended to the four quadrants differently during the inference phase, suggesting that they made specific predictions about where the target would appear. There was no significant difference in RT across different trial types when the knowledge was explicitly told vs. when the knowledge was implicitly learned. Post-hoc Tukey HSD tests showed that RT in the impossible trials was reliably slower than that in the other three types of trials [p's<.03], the RT in the 2-quadrant disjunction trials was not reliably different from that in the 4-quadrant disjunction trials [p=.99], and the RT in the conjunction trials was reliably faster than both the 2-quadrant and 4-quadrant disjunction trials [p's<.01]. We then performed planned contrast analysis separately for the implicit and explicit conditions. The 2-quadrant and 4-quadrant disjunction trials were combined as

one category in the analysis. For both conditions, RT in conjunction trials was significantly faster than that in disjunction trials, which in turn was faster than that in the impossible trials [p's<.01].

Additionally, we examined RT performance separately for learners (whose test phase accuracy>0.25, N=42) and non-learners (whose test phase accuracy≤0.25, N=18). For learners, RT in conjunction trials was significantly faster than that in disjunction trials [p=.014], which in turn was faster than that in the impossible trials [p<.001]. For non-learners, RT in conjunction trials was marginally faster than that in disjunction trials [p=.09], but there was no difference in RT for the disjunction and impossible trials [p=.97]. This suggests that participants with higher test phase accuracy showed the effect more robustly than participants with lower test phase accuracy did.

These results suggest that when two objects each predicting a different outcome were presented at the same time, participants automatically made a conjunctive prediction which contained the shared property of the different outcomes.

Experiment 2

One explanation for faster RT in the conjunction trials in Experiment 1 was that the conjunctive quadrant was smaller in terms of spatial scope than the disjunctive quadrants. The smaller spatial scope might have facilitated visual search, leading participants to prioritize the conjunctive quadrant over the other quadrants. To examine this possibility, in Experiment 2, we aimed to equate the spatial scope of conjunctive and disjunctive quadrants in the inference phase.

Participants

A new group of 120 students (95 female, mean age=20.2 years, SD=1.9) from UBC participated for course credit.

Stimuli and Procedure

The stimuli and procedure in the experiment were the same as those in Experiment 1, except for one critical difference: During the exposure phase, after a color circle, the target could appear in three of the four quadrants. This means that in the inference phase, for each pair of color circles, two of the quadrants on the array would be consistent with a conjunctive inference, and the other two quadrants would be consistent with a disjunctive inference (Figure 4).

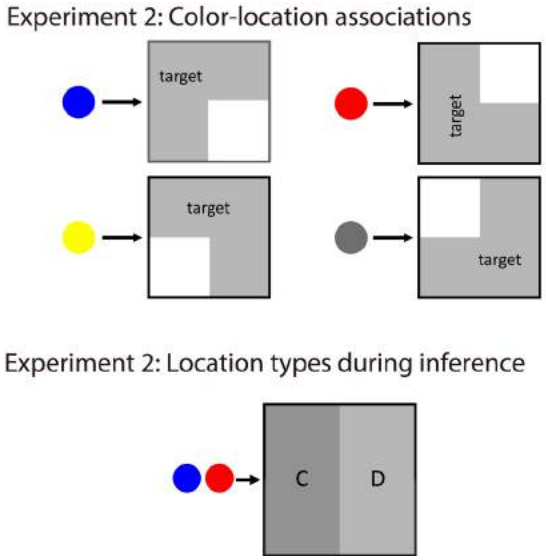


Figure 4: Experiment 2 paradigm. The stimuli and procedure were the same as those in Experiment 1, except that each color predicted the target would appear in three quadrants in the array during exposure. Consequently, two quadrants during the inference phase were consistent with a conjunctive inference (C), and the other two were consistent with a disjunctive inference (D).

Results and Discussion

The test phase accuracy for participants in the implicit condition was 0.31, which was not reliably above chance [p=0.11], suggesting that participants in the implicit condition did not successfully learn the color-location associations during exposure. This may be due to the difficulty of learning that the target could appear in three quadrants instead of two.

A 2 (condition: explicit vs. implicit, between-subjects) × 2 (trial type: conjunctive vs. disjunctive, within-subjects) mixed-design ANOVA revealed a marginal interaction between condition and trial type [$F(1,118)=3.865$, p=.05, η_p^2=0.03], but no main effect of condition [$F(1,118)=0.40$, p=.53, η_p^2=0.00], or trial type [$F(1,118)=1.22$, p=.27, η_p^2=0.01]. We then compared the RT in conjunction and disjunction trials separately for the implicit and explicit conditions. In the explicit condition, RT in conjunction trials was reliably faster than that in disjunction trials [$t(1,59)=2.03$, p<.05, d=0.24], but in the implicit condition, the RT in conjunction trials was not reliably different from that in disjunction trials [$t(1,59)=0.66$, p=.51, d=0.07] (see Figure 5).

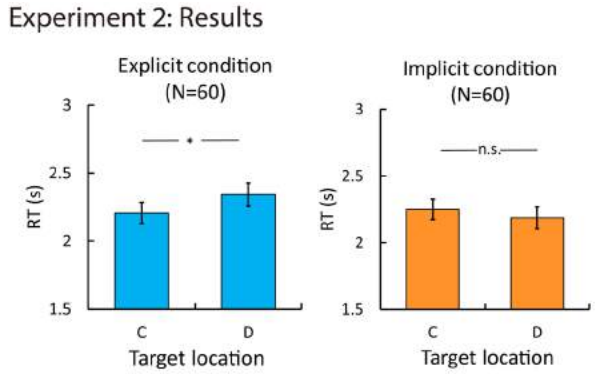


Figure 5: Experiment 2 results. The RT for each type of trials (C for conjunction, D for disjunction) was graphed separately for the explicit and implicit conditions (Error bar reflect ± 1 SE; *p<.05).

These results suggested that when participants learned the color-location associations, they automatically made conjunctive inferences when they saw two color circles, even when the conjunctive quadrants were of the same spatial scope as the disjunctive quadrants. On the other hand, if participants did not successfully learn the color-location associations, they failed to make such conjunctive inferences.

Experiment 3

In Experiments 1 and 2, the two color circles were presented simultaneously side by side during the inference phase to elicit conjunctive predictions. An alternative method to represent conjunctions is to combine two features into one object, such as combing the color red and the shape square into a red square (Treisman & Gelade, 1980; Singer & Gray, 1995). Therefore, in this experiment, we tested this alternative presentation where the two predictors were combined into a new object, rather than manifesting them as two different objects, to elicit conjunctive predictions.

Participants

A new group of 60 students (47 female, mean age=19.6 years, SD=2.6) from UBC participated for course credit. In the current experiment, only the implicit condition was examined (we did not examine the explicit condition due to time constraints in participant recruitment).

Stimuli and Procedure

The stimuli and procedure in the experiment were the same as those in Experiment 1, except for two critical differences.

First, during the exposure phase there were two color circles (red and blue, as described in Experiment 1) and two textured circles (dotted and stripy circles, see Figure 6). The two color circles were always presented with a filled texture, and the two textured circles were always presented in a black color (R/G/B: 0/0/0). The two color circles always predicted two parallel halves of the array (e.g., the top and bottom halves), and the two textured circles predicted the other two halves of the array (e.g., the left and right halves). The assignment of a color or texture to a given half was randomized across participants, but remained constant for a given participant throughout the experiment.

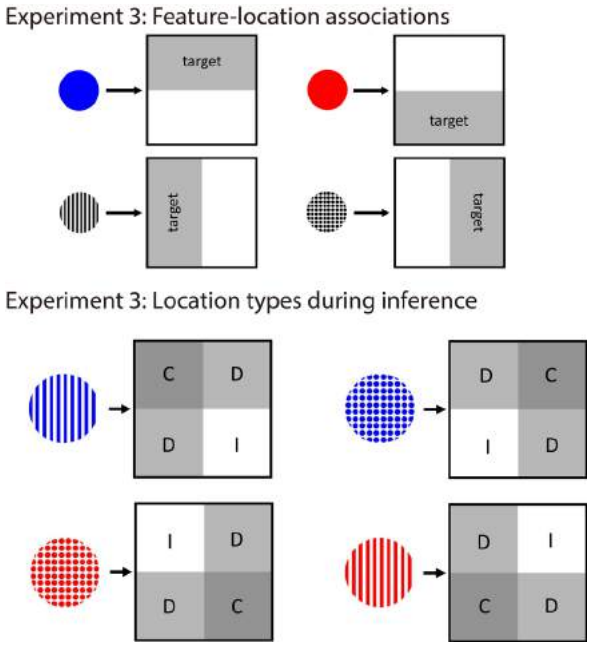


Figure 6. Experiment 3 paradigm. Two unique colors and two unique textures each predicted the target location in the following array during exposure. Circles with both a unique color and a unique texture were presented one at a time during inference. The trial types during the inference phase were the same as those in Experiment 1.

Second, during the inference phase participants saw one circle at a time on the screen. Each circle contained one of the two colors and one of the two textures presented in the exposure phase (i.e., a blue stripy circle, a blue dotted circle, a red stripy circle, or a red dotted circle). There were four trials for each unique colored textured circle. Since a color and a texture never predicted two parallel halves during exposure, there were three types of trials in the inference phase as in Experiment 1: conjunctive trials where the target could appear in a conjunctive quadrant (C), disjunctive trials where the target could appear in a disjunctive quadrant (D), and impossible trials where the target never appeared in a quadrant based on exposure (I).

Results and Discussion

The test phase accuracy in this experiment was 0.33, which was marginally above chance [p=.07], suggesting that learning was weak.

As before, we analyzed RT of correct trials in the inference phase (Figure 7). A one-way repeated-measures ANOVA revealed a main effect of trial type [$F(2,118)=5.32$, $p<.001$, $\eta_p^2=0.24$]. Post-hoc Tukey HSD tests showed that there was reliable RT difference in the conjunction trials and impossible trials [$p<.01$]. Other pair-wise comparisons were numerically similar to those in Experiment 1, but not statistically reliable [p's>.11]. These results suggest that the participants made conjunctive predictions when the two

features were presented in a new object. However, the effect was not as strong as in previous experiments.

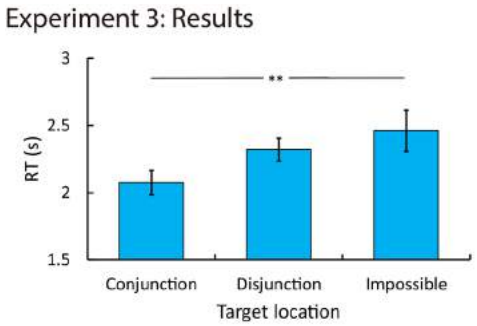


Figure 7: Experiment 3 results. The RT for each type of trials was graphed (Error bar reflect ± 1 SE; **p<.01).

General Discussion

In this study, we examined how predictions were made in the presence of two objects that were associated with two different outcomes. Using a visual search paradigm, unique colors (all three experiments) or textures (Experiment 3) predicted a specific location of the target in the search array in the exposure phase. In the inference phase, we examined where the target was expected to appear when two color circles (Experiments 1 and 2) or a circle with a unique color and a unique texture (Experiment 3) were presented at the same time. Importantly in the inference phase, the target appeared in any location with equal probability.

Based on the speed of visual search (RT), we found that participants were faster to find the target when it appeared in a conjunctive quadrant than in disjunctive or impossible quadrants. This was surprising because the simultaneous presentation of the two circles or features did not necessarily dictate a conjunctive or disjunctive inference. For example, just because the blue circle previously predicted the top half and the red circle previously predicted the left half, the blue and red circles together, in principle, could predict either the top left quadrant (conjunctive inference), or the top left, top right, and bottom left quadrants (disjunctive inference). What we found was that participants automatically prioritized the conjunctive quadrant over the disjunctive quadrant in the visual search task, at the presence of the two predictors. This conjunctive preference occurred without prior instructions, or even explicit awareness of the color-location associations.

Across all three experiments, participants were not told anything about where to look when two color circles or two different features were presented together. Therefore, the differential RT in the conjunctive quadrant indicated an automatic expectation resulting from the previously learned color- or feature-location associations during exposure.

In Experiment 1, the expectation to find the target in a location consistent with a conjunctive prediction was equally strong whether participants implicitly learned the associations or were explicitly told about the associations. However, in Experiment 2 when there was no successful learning of the associations in the implicit condition, this conjunctive prediction was absent. In fact, the conjunctive prediction was only present when participants were explicitly told about the color-location associations in the explicit condition. This suggests that the conjunctive predictions were only made when participants have successfully learned the color-location associations, either after implicit statistical learning, or after explicit instructions of these associations.

It is important to note that the disjunctive quadrants in the current study were exclusively disjunctive, not containing the conjunctive quadrant. The fact that the RT in the disjunction trials was faster than that in the impossible trials but slower than that in the conjunction trials suggests that the impossible quadrant may be inhibited and the conjunctive quadrant may be prioritized during visual search.

We think that both the learning process and the prediction process were implicit. In all three experiments, participants were not told anything about the object-location associations before the exposure phase in the implicit condition. That is, participants were only told to find the target in the search array and were not told that the object before each search array predicted the location of the target. Therefore, learning of the associations in the implicit condition was automatic and implicit. In the inference phase, there was no explicit instruction as to what to do with the two objects. Again, participants were only told to find the target in the search array. Moreover, the target in the inference phase could appear in any quadrant with equal probability, so the two objects were completely task-irrelevant. Finally, the RT was relatively fast so any explicit reasoning process may not occur in the period between object presentation and target search. For these reasons, we think that the conjunctive predictions were implicit.

There are several limitations of the current study. First, we only presented two objects side by side, or two features in a single object as cues. There might be other ways to represent such joint cues using semantic categories (e.g., if object A is associated with the "dog" category and object B is associated with the "small" category, will people automatically predict Chihuahuas and Pomeranians upon seeing A and B?). Second, we only used RT as a measure to probe whether participants made conjunctive or disjunctive predictions. A richer method can involve eye tracking to see the timecourse of attention to the different quadrants in the inference phase. Finally, there was a confound of proximity in the current study, where the conjunctive quadrant was spatially closer to the disjunctive quadrant than to the impossible quadrant. This could explain the RT advantage of the disjunction trials over the impossible trials.

In conclusion, the current results suggest that in the presence of multiple predictors, statistical learning generates automatic expectations about the outcomes in a conjunctive fashion.

Acknowledgements

We would like to thank Chaz Firestone, Justin Halberda, Chris Mole, Yu Luo, Brandon Tomm, and five anonymous reviewers for their helpful comments. This work was

supported by NSERC Discovery Grant (RGPIN-2014-05617 to JZ), Canada Research Chairs program (to JZ), Leaders Opportunity Fund from the Canadian Foundation for Innovation (F14-05370 to JZ), and Alexander Graham Bell Canada Graduate Scholarships-Doctoral Program (to RY).

References

Baker, C. I., Olson, C. R., & Behrmann, M. (2004). Role of attention and perceptual grouping in visual statistical learning. *Psychological Science, 15*, 460-466.

Brady, T. F., Konkle, T., & Alvarez, G. A. (2009). Compression in visual working memory: using statistical regularities to form more efficient memory representations. *Journal of Experimental Psychology: General, 138*, 487-502.

Fiser, J., & Aslin, R. N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological Science, 12*, 499-504.

Hall, M., Mattingley, J., & Dux, P. (2015). Distinct contributions of attention and working memory to visual statistical learning and ensemble processing. *Journal of Experimental Psychology: Human Perception and Performance, 41*, 1112-1123.

Luo, Y. & Zhao, J. (2018). Statistical learning creates novel object associations via transitive relations. *Psychological Science, 29*, 1207-1220.

Mendelson, Elliott. *Introduction to mathematical logic.* Chapman and Hall/CRC, 2009.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science, 274*, 1926-1928.

Singer, W., & Gray, C. M. (1995). Visual feature integration and the temporal correlation hypothesis. *Annual review of neuroscience, 18*, 555-586.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive psychology, 12,* 97-136.

Turk-Browne, N. B., Jungé, J. A., & Scholl, B. J. (2005). The automaticity of visual statistical learning. *Journal of Experimental Psychology: General, 134*, 552–564.

Turk-Browne, N. B., Scholl, B. J., Johnson, M. K., & Chun, M. M. (2010). Implicit perceptual anticipation triggered by statistical learning. *The Journal of Neuroscience, 30*, 11177-11187.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological review, 90*, 293.

Yu, R., & Zhao, J. (2015). The persistence of attentional bias to regularities in a changing environment. *Attention, Perception, & Psychophysics, 77*, 2217-2228.

Yu, R. & Zhao, J. (2018a). Object representations are biased toward each other through statistical learning. *Visual Cognition, 26*, 253-267.

Yu, R. & Zhao, J. (2018b). Implicit updating of object representation via temporal associations. *Cognition, 181*, 127-134.

Zhao, J., Al-Aidroos, N., & Turk-Browne, N. B. (2013). Attention is spontaneously biased toward regularities. *Psychological science, 24*, 667-677.

Zhao, J., Ngo, N., McKendrick, R., & Turk-Browne, N. B. (2011). Mutual interference between statistical summary perception and statistical learning. *Psychological Science, 22*, 1212-1219.

Zhao, J., & Yu, R. (2016). Statistical regularities reduce perceived numerosity. *Cognition, 146*, 217-222.

Semantic categories of artifacts and animals reflect efficient coding

Noga Zaslavsky[1,2] **(noga.zaslavsky@mail.huji.ac.il)**
Terry Regier[2,3] **(terry.regier@berkeley.edu)**
Naftali Tishby[1,4] **(tishby@cs.huji.ac.il)**
Charles Kemp[5] **(c.kemp@unimelb.edu.au)**

[1]Edmond and Lily Safra Center for Brain Sciences, Hebrew University, Jerusalem 9190401, Israel
[2]Department of Linguistics, University of California, Berkeley, CA 94720 USA
[3]Cognitive Science Program, University of California, Berkeley, CA 94720 USA
[4]Benin School of Computer Science and Engineering, Hebrew University, Jerusalem 9190401, Israel
[5]School of Psychological Sciences, University of Melbourne, Parkville, Victoria 3010, Australia

Abstract

It has been argued that semantic categories across languages reflect pressure for efficient communication. Recently, this idea has been cast in terms of a general information-theoretic principle of efficiency, the Information Bottleneck (IB) principle, and it has been shown that this principle accounts for the emergence and evolution of named color categories across languages, including soft structure and patterns of inconsistent naming. However, it is not yet clear to what extent this account generalizes to semantic domains other than color. Here we show that it generalizes to two qualitatively different semantic domains: names for containers, and for animals. First, we show that container naming in Dutch and French is near-optimal in the IB sense, and that IB broadly accounts for soft categories and inconsistent naming patterns in both languages. Second, we show that a hierarchy of animal categories derived from IB captures cross-linguistic tendencies in the growth of animal taxonomies. Taken together, these findings suggest that fundamental information-theoretic principles of efficient coding may shape semantic categories across languages and across domains.

Keywords: information theory; language evolution; semantic typology; categories

Introduction

Cross-linguistic studies in several semantic domains, such as kinship, color, and numeral systems, suggest that word meanings are adapted for efficient communication (see Kemp, Xu, & Regier, 2018 for a review). However, until recently it had remained largely unknown to what extent this proposal can account for soft semantic categories and inconsistent naming, that could appear to pose a challenge to the notion of efficiency, and how pressure for efficiency may relate to language evolution. Recently Zaslavsky, Kemp, Regier, and Tishby (2018; henceforth ZKRT) addressed these open questions by grounding the notion of efficiency in a general information-theoretic principle, the Information Bottleneck (IB; Tishby, Pereira, & Bialek, 1999). ZKRT tested this formal approach in the domain of color naming and showed that the IB principle: (1) accounts to a large extent for cross-language variation in color naming; (2) provides a theoretical explanation for why observed patterns of inconsistent naming and soft semantic categories may be efficient; and (3) suggests a possible evolutionary process that roughly recapitulates Berlin and Kay's (1969) discrete implicational hierarchy while also accounting for continuous aspects of color category evolution. However, it is not yet clear to what extent these results may generalize to other semantic domains, especially those that are fundamentally unlike color.

Here we test the generality of this theoretical account by considering two additional semantic domains: artifacts and animals. These domains are of particular interest in this context because they are qualitatively different from color, they have not previously been comprehensively addressed in terms of efficient communication, and at the same time it is possible to apply to them the same communication model that has previously been used to account for color naming.

First, we consider naming patterns for household containers. This is a semantic domain in which categories are known to overlap and generate inconsistent naming patterns (Ameel, Storms, Malt, & Sloman, 2005; Ameel, Malt, Storms, & Assche, 2009). Although it has previously been shown that container naming in English, Spanish, and Chinese is efficient compared to a large set of hypothetical naming systems (Xu, Regier, & Malt, 2016), that demonstration did not consider the full probability distribution of names produced by different speakers, did not explicitly contrast monolingual and bilingual speakers, and was based on a smaller set of stimuli than we consider here. In this work we show that the full container-naming distribution in Dutch and French, including overlapping and inconsistent naming patterns, across a large set of stimuli, both in monolinguals and bilinguals, is near-optimally efficient in the IB sense.

Second, we test the evolutionary account of ZKRT in the case of animal categories. By analogy with Berlin and Kay's implicational hierarchy of color terms, Brown (1984) proposed an implicational hierarchy for the evolution of animal taxonomies based on cross-language comparison. We show that aspects of this hierarchy are captured by a sequence of efficient animal-naming systems along the IB theoretical limit. Our results also support the view that both perceptual and functional features shape animal categories across languages (Malt, 1995; Kemp et al., 2018).

The remainder of this paper proceeds as follows. First, we review the theoretical framework and formal predictions on which we build. We then present two studies that apply this approach to the aforementioned semantic domains.

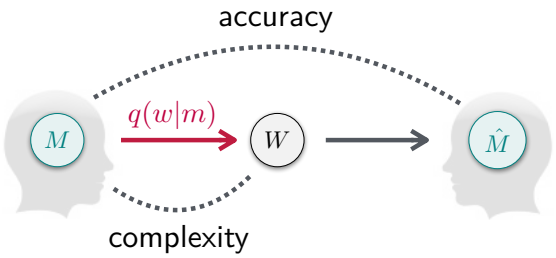


Figure 1: Communication model adapted from ZKRT. A speaker communicates a meaning M by encoding it into a word W according to a naming distribution $q(w|m)$. This word is then interpreted by the listener as $\hat{M}$. Complexity is a property of the mapping from meanings to words, and accuracy is determined by the similarity between M and $\hat{M}$.

Theoretical framework and predictions

We consider here the theoretical framework proposed by ZKRT, which is based on a simplified interaction between a speaker and a listener (Figure 1), formulated in terms of Shannon's (1948) communication model. The speaker communicates a meaning m, sampled from $p(m)$, by encoding it into a word w, generated from a naming (or encoder) distribution $q(w|m)$. The listener then tries to reconstruct from w the speaker's intended meaning. We denote the reconstruction by $\hat{m}_w$, and assume it is obtained by a Bayesian listener.[1] These meanings, m and $\hat{m}_w$, are taken to be mental representations of the environment, defined by distributions over a set $\mathcal{U}$ of relevant features. For example, if communication is about colors, then $\mathcal{U}$ may be grounded in a perceptual color space, and each color would be mentally represented as a distribution over this space.

Under these assumptions, efficient communication systems are those naming distributions that optimize the Information Bottleneck (IB; Tishby et al., 1999) tradeoff between the complexity and accuracy of the lexicon. Formally, complexity is measured by the mutual information between meanings and words, i.e.:

$$I_q(M;W) = \sum_{m,w} p(m)q(w|m)\log\frac{q(w|m)}{q(w)}, \qquad (1)$$

which roughly corresponds to the number of bits used to encode meanings into words. Accuracy is inversely related to the discrepancy between m and $\hat{m}_w$, measured by the expected Kullback–Leibler (KL) divergence between them:

$$\mathbb{E}_q[D[m\|\hat{m}_w]] = \mathop{\mathbb{E}}_{\substack{m\sim p(m)\\ w\sim q(w|m)}}\left[\sum_{u\in\mathcal{U}} m(u)\log\frac{m(u)}{\hat{m}_w(u)}\right]. \qquad (2)$$

Accuracy is defined by $I_q(W;U) = \mathbb{E}_q[D[\hat{m}_w\|m_0]]$, where m_0 is the prior representation before knowing w, and maximizing accuracy amounts to minimizing equation (2).[2]

[1] The reconstruction of a Bayesian listener with respect to a given naming distribution is defined by $\hat{m}_w = \sum_m q(m|w)m$.

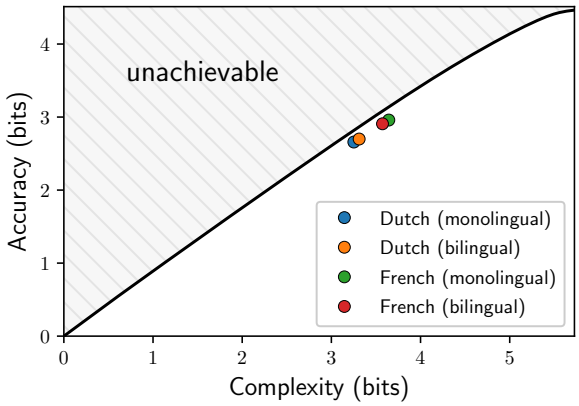


Figure 2: The black curve is the IB theoretical limit of efficiency for container naming, obtained by varying β. Points above this curve cannot be achieved. Complexity and accuracy tradeoffs in the four naming conditions are near-optimal.

Achieving maximal accuracy may require a highly complex system, while minimizing complexity will result in a non-informative system. Efficient systems are thus pressured to balance these two competing goals by minimizing the IB objective function,

$$\mathcal{F}_\beta[q] = I_q(M;W) - \beta I_q(W;U), \qquad (3)$$

where $\beta \geq 0$ controls the efficiency tradeoff. The optimal systems, $q_\beta(w|m)$, achieve the minimal value of equation (3) given β, denoted by $\mathcal{F}^*_\beta$, and evolve as β gradually shifts from 0 to ∞. Along this trajectory they become more fine-grained and complex, while attaining the maximal achievable accuracy for their level of complexity. This set of optimal systems defines the theoretical limit of efficiency (see Figure 2).

If languages are pressured to be efficient in the IB sense, then for a given language l with naming system $q_l(w|m)$, two predictions are made. (1) Deviation from optimality, or *inefficiency*, should be small. This is measured by $\varepsilon_l = \frac{1}{\beta_l}(\mathcal{F}_{\beta_l}[q_l] - \mathcal{F}^*_{\beta_l})$, where β_l is estimated such that ε_l is minimized. (2) The *dissimilarity* between q_l and the corresponding IB system, q_{β_l}, should be small. This is evaluated by a dissimilarity measure (gNID) proposed by ZKRT. In addition, ZKRT suggested that languages evolve along a trajectory that is pressured to remain near the theoretical limit.

These predictions were previously supported by evidence from the domain of color naming. To apply this approach to other domains, i.e. to instantiate the general communication model, two components must be specified: a *meaning space*, which is the set of meanings the speaker may communicate; and a prior, $p(m)$, also referred to as a *need distribution* (Regier, Kemp, & Kay, 2015), since it determines the frequency with which each meaning needs to be communicated. In the following sections we present two studies that

[2] See (Zaslavsky et al., 2018) for detailed explanation.

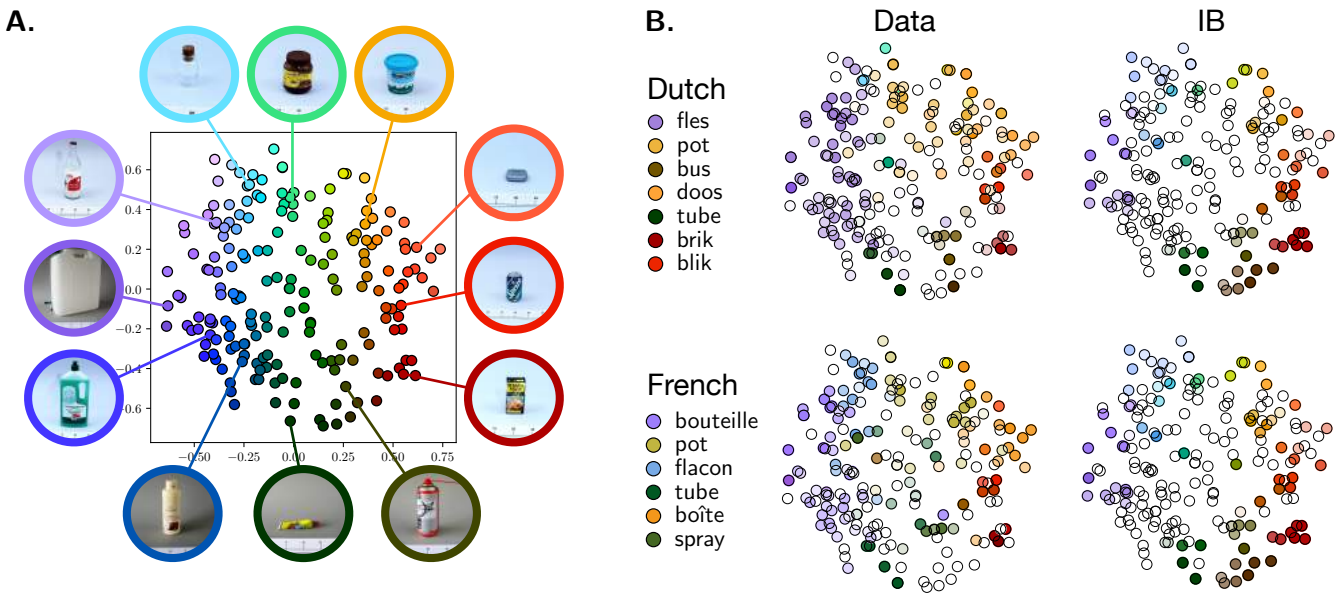


Figure 3: **A**. Two dimensional nMDS embedding and color coding of the containers stimulus set used by White et al. (2017). Images show a few examples. **B**. Monolingual naming distributions for Dutch (upper left) and French (lower left), together with their corresponding IB systems (right column), are visualized over the 2D embedding shown in (A). Each color corresponds to the color centroid of a container category, w, based on the color map in (A). Colors show category probabilities above 0.4, and color intensities reflect the values between 0.4 and 1. White dots correspond to containers for which no category is used with probability above 0.4. Legend for each language shows only major terms.

follow this approach and test its predictions in qualitatively different semantic domains.

Study I: Container names

The goal of this experiment is to test the theoretical predictions derived from IB in the case of container naming. It is not clear whether previous findings for color would generalize to this case for several reasons. First, the representation of artifacts is likely to involve more than just a few basic perceptual features, unlike color. Second, categories in this domain are believed to be strongly shaped by adaptation to changes in the environment (Malt, Sloman, Gennari, Shi, & Wang, 1999). At the same time, container categories tend to overlap, as in the case of color categories, posing a similar theoretical challenge to explain this observation in terms of communicative efficiency. Finally, the bilingual lexicon in this domain has been extensively studied, and it has been shown that bilingual naming patterns tend to converge (Ameel et al., 2005, 2009). However, it is not yet clear whether this convergence, or compromise, comes at a cost in communicative efficiency, or whether it may actually be formalized and explained in terms of efficiency.

Data. To address these open questions, we consider sorting and naming data collected by White et al. (2017), relative to a stimulus set of 192 images of household containers (see Figure 3A for examples). This set is substantially larger than those used in previous container-naming studies (e.g. Malt et al., 1999; Ameel et al., 2005), thus providing a better representation of this semantic domain. In the naming task, 32 Dutch and 30 French monolingual speakers, as well as 30 bilingual speakers, were asked to provide names for the containers in the stimulus set. Bilingual participants performed the task once in each language. The container-naming distribution in each of the four conditions (language $\times$ linguistic status) is defined by the proportion of participants in that condition that used the word w to describe a container c. A separate sorting task was performed by 65 Dutch speakers, who were asked to organize all containers into piles based on their overall qualities. Participants were also allowed to form higher-level clusters by grouping piles together. White et al. (2017) evaluated the similarity between two containers, denoted here by $\mathrm{sim}(c, c')$, based on the number of participants that placed them in the same pile or cluster (see White et al., 2017 for detail). In both tasks, participants were instructed not to take into account the content of the object (e.g., water).

Model. We ground the meaning space in the similarity data, following a related approach proposed by Regier et al. (2015) and Xu et al. (2016). While these data are from Dutch speakers, there are only minor differences in perceived similarities among speakers of different languages (Ameel et al., 2005). Therefore, we assume that these similarity judgments reflect a shared underlying perceptual representation of this domain. We take $\mathcal{U}$ to be the set of containers in the stimulus set, and define the mental representation of each container c by the similarity-based distribution it induces over the domain, $m_c(u) \propto \exp(\gamma \cdot \mathrm{sim}(c, u))$, where γ^{-1} is taken to be the

empirical standard deviation of $\mathrm{sim}(c, u)$. In contrast with the case of color, in which these mental representations were grounded in a standard perceptual space, here there is no standard perceptual space for containers, and so our assumed underlying perceptual representation requires further validation, which we leave for future work. We define the need distribution, $p(m_c)$, by averaging together the least informative (LI) priors for the different languages, as proposed by ZKRT. We used only the monolingual data for this purpose, and regularized the resulting prior by adding $\epsilon = 0.001$ to it and renormalizing.

Results

We estimated the theoretical limit of efficiency for container naming by applying the IB method (Tishby et al., 1999), as ZKRT did in the case of color naming, here with 1500 values of $\beta \in [0, 1024]$. We evaluated the empirical complexity and accuracy in the four naming conditions by entering the corresponding naming distributions in the equations for $I_q(M;W)$ and $I_q(W;U)$. The results are shown in Figure 2 and Table 1. It can be seen that container naming in Dutch and French lie near theoretical limit, both for monolinguals and bilinguals, and that bilinguals achieve similar levels of efficiency as monolinguals (Table 1). In all four cases, the corresponding IB solution is at $\beta_l \approx 1.2$, suggesting that there is only a weak preference for accuracy over complexity in this domain, as also found for color naming.

Consistent with the empirical observations of convergence in the bilingual lexicon, the complexity-accuracy tradeoffs in bilinguals are closer to each other (Figure 2, orange and red dots) compared to the monolingual tradeoffs (Figure 2, blue and green dots). This may be explained by a need to reduce the complexity of maintaining two naming systems simultaneously, while achieving monolingual-like levels of efficiency in each language. To test this possibility, we compared two joint French-Dutch systems that bilinguals may employ: one that randomly selects one of the two monolingual systems to name objects, and another that randomly selects one of the two bilingual systems. We found a 0.16% reduction in the complexity of the joint bilingual system compared to the joint monolingual system. Although this is a small effect, it may accumulate across domains to have a substantial impact. In addition, our simple calculation did not take into account similar word forms, which may also reduce complexity (Ameel et al., 2005). Thus, this finding suggests that the convergence in the bilingual lexicon may be shaped, at least in part, by pressure for efficiency.

The remainder of our analysis focuses on the monolingual systems, as they are more distinct and presumably more representative of each language. To get a precise sense of how challenging it may be to reach the observed levels of efficiency, we compared the actual naming systems to a set of hypothetical systems that preserve some of their statistical structure. This set was constructed by fixing the conditional distributions of words, while shifting how they are used by applying a random permutation of the containers. For each language we constructed $10,000$ such hypothetical systems. Table 1 shows that these hypothetical systems are substantially less efficient than the actual systems, and are also less similar to the IB systems. In fact, both languages achieve better (lower) scores than all of their hypothetical variants, providing a precise sense in which they are near-optimal according to IB. One possible concern is that this outcome may be a result of the LI prior, which was fitted to the naming data. To address this, we repeated this analysis with a uniform need distribution. The results in that case are similar (not shown), although as expected the fit to the actual systems is not as good compared to the LI prior.

Table 1: Evaluation of the IB container-naming model. Lower values indicate a better fit of the model. Values for hypothetical systems are averages $\pm$SD over $10,000$ systems.

		Inefficiency	Dissimilarity
Dutch	monolingual	0.16	0.11
	bilingual	0.17	0.12
	hypothetical	0.29 ($\pm$0.02)	0.59 ($\pm$0.05)
French	monolingual	0.18	0.11
	bilingual	0.17	0.09
	hypothetical	0.31 ($\pm$0.01)	0.56 ($\pm$0.06)

The low dissimilarity scores for the actual languages, shown in Table 1, suggest that the observed soft category structure in this domain may also be accounted for by the IB systems. This is indeed supported by a fine-grained comparison between the naming distribution in both languages and their corresponding IB systems. To see this, we embedded the 192 containers in a 2-dimensional space by applying non-metric multidimensional scaling (nMDS) with respect to the similarity data, similar to Ameel et al. (2009). This was done using the scikit-learn package in Python. We initialized the nMDS procedure with a solution for the standard metric MDS that achieved the best fit to the similarity data out of 50 solutions generated with random initial conditions. For visualization purposes, we assigned a unique color to each container. The resulting 2D embedding and color coding of the containers stimulus set are shown in Figure 3A.

The monolingual systems in Dutch and French are shown in Figure 3B, together with their corresponding IB systems. These two IB systems are very similar, although not identical, which is not surprising given that the naming patterns in Dutch and French are fairly similar. Both the actual systems and the IB systems exhibit soft category structure and similar patters of inconsistent naming, as shown by the white dots. In addition, since each category is colored according to its centroid, similarity between the category colors together with their spatial distribution reflect the similarity between the full naming distributions. For example, the IB systems have a category that is similar to *fles* and *bouteille*, as well as a category that is similar to *doos* and *boîte* in Dutch and French respectively, although these categories in the IB systems are a bit narrower. The IB systems also capture the

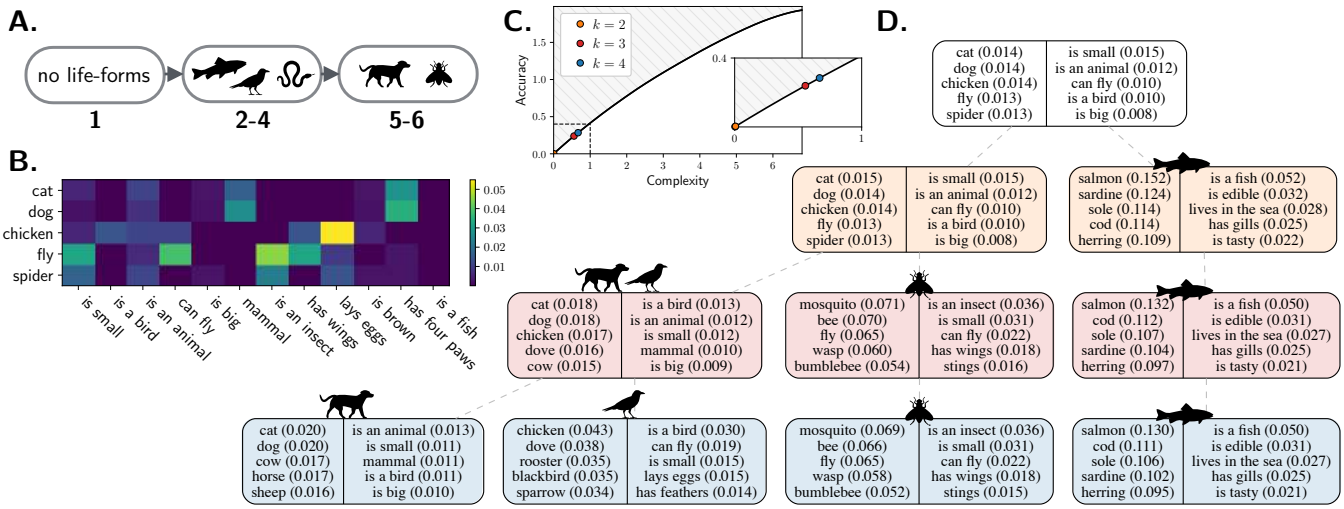


Figure 4: **A**. Brown's (1984) proposed hierarchy for animal categories. **B**. Subset of the conditional probabilities of features (columns) given animal classes (rows), for the 5 most familiar classes and 12 most frequently generated features. **C**. Theoretical limit for animal naming. Colored dots along the curve correspond to the systems shown in (D), with $k = 2, 3, 4$ categories. **D**. Animal category hierarchy derived from IB. Each level corresponds to an IB system. Each box corresponds to a category, which is represented by its top five classes (left) and features (right) and their probabilities given the category.

category *tube* quite well in both languages. However, there are also some apparent discrepancies. For example, the distinction between *bouteille* and *flacon* in French is reflected in both IB systems, although Dutch does not have the same pattern in this case (Ameel et al., 2005).

This analysis shows that efficiency constraints may to a substantial extent explain the container-naming distribution in Dutch and French, including soft category boundaries and inconsistent naming observed empirically, both in monolinguals and bilinguals. It thus supports the hypothesis that a drive for information-theoretic efficiency shapes word meanings across languages and across semantic domains. However, since this analysis is based only on two closely related languages, we were not able to test how well the results for this domain generalize across languages. Important directions for future research include testing whether these results generalize to other, preferably unrelated, languages, and further testing the extent to which the convergence in the bilingual lexicon is influenced by pressure for efficiency. The next section focuses on another semantic domain for which we are able to obtain broader cross-linguistic evidence.

Study II: Folk biology

Cross-language variation and universal patterns in animal taxonomies have been extensively documented and studied (Berlin, 1992), however this domain has not yet been approached in terms of efficient communication. By analogy with Berlin and Kay's theory, Brown (1984) proposed an implicational hierarchy for animal terms, based on data from 144 languages. Brown identified six stages for animal taxonomies, as illustrated in Figure 4A. Languages at the first stage do not have any lexical representation for life-forms. Languages at stages 2-4 add terms for *fish*, *bird* and *snake*, but Brown does not argue for any particular order for these categories. Terms for *mammal* and *wug* ("worm-bug", referring in addition to small insects) are added in stages 5 and 6, again with no implied order. Much of the data analyzed in this domain is not fine-grained, and Brown's proposal has been criticized (Randall & Hunn, 1984) mainly due to lack of sufficiently accurate data. Nonetheless, his observations can be considered as a rough approximation of cross-linguistic tendencies in this semantic domain. Therefore, in this work we aim at testing whether broad cross-linguistic patterns, as summarized by Brown's proposal, can be accounted for in terms of pressure for efficiency. More specifically, our goal is to derive from the IB principle a trajectory of efficient animal-naming systems, analogous to ZKRT's trajectory for color, and to compare this trajectory to the naming patterns reported by Brown. However, unlike previous comparisons to IB optima, due to the nature of available data, here we only attempt to make coarse comparisons.

To derive a trajectory of efficient animal-naming systems, we first need to specify the communication model in this domain. We ground the representations of animals in high-level, human-generated features. Specifically, we consider the Leuven Natural Concept Database (De Deyne et al., 2008), which contains feature data and familiarity ratings for animal classes (e.g., "cat", "chicken", etc.). These data were collected from Dutch speakers, and then translated to English. We follow Kemp, Chang, and Lombardi (2010), who considered 113 animal classes and 757 features from this database, and for each feature u and class c estimated the conditional probability $p(u|c)$ based on the number of participants who generated this feature for that class (see Figure 4B for exam-

ples). We take $\mathcal{U}$ to be the set of animal features, and assume each animal class is mentally represented by the distribution it induces over features, i.e. $m_c(u) = p(u|c)$, as estimated by Kemp et al. (2010). In addition, we follow Kemp et al. (2010) in using a familiarity-based prior over animal classes, in which the probability of a class is proportional to its familiarity score. We define the need distribution to be this prior.

Given these components, we estimated the theoretical limit for animal naming (Figure 4C) using the same method as before, this time with 3000 values of $\beta \in [0, 2^{13}]$. We then selected the most informative systems with $k = 2, 3, 4$ categories. The number of categories, k, was determined by considering categories w with probability mass $q_\beta(w) > 0.00001$. These systems are shown in Figure 4D, where each layer of the hierarchy corresponds to a system and each box corresponds to a category within that system. The top layer, with a single category, corresponds to a non-informative system that does not distinguish between different animal classes. This can be considered as a stage 1 system in Brown's sequence. The second layer (shown in orange) roughly corresponds to a stage 2 system. It consists of a *fish* category, as can be inferred from the distribution it induces over features and animals, and another category for all other animals. It lies very close to the origin in Figure 4C, as it maintains little information about most animals. The third layer (shown in red) corresponds to a system with categories for *fish* and *wug*, as well as a category that is dominated by birds and mammals. The *bird-mammal* category has greater probability mass (0.8) than the *wug* category (0.14), suggesting that it is more prominent even though these two categories appear together. This transition deviates from Brown's sequence in the early appearance of *wug* (although not strongly weighted here), and in lacking a *snake* category (although animals from that category do appear in the Leuven database). One possible explanation for this deviation is that the feature data on which we relied were obtained from Dutch participants, and are thus strongly biased toward Western societies. In the next layer (shown in blue), the 3-category system evolved to a 4-category system by refining the *bird-mammal* category, resulting in a system that roughly corresponds to a Brown stage 6 system, with the exception of *snake*.

These results suggest that animal naming systems may evolve under efficiency pressure much as color appears to, despite the qualitative difference between these domains. However, in order to test this proposal more comprehensively, fine-grained cross-linguistic animal naming data is required, comparable to the naming data for colors and containers. The fact that systems along the theoretical limit capture some cross-linguistic tendencies in animal taxonomies is notable, given that our characterization of the domain, in terms of features, was necessarily strongly biased toward animal representations in Western societies. This finding supports the idea that to some extent at least there is a shared underlying representation of animals across cultures (Mayr, 1969), while also raising the interesting possibility of some cross-language and cross-cultural differences in underlying representations. It is also worth noting that the salient features in the IB systems tend to be both perceptual (e.g., "is big") and functional (e.g., "is edible"), suggesting that both types of features may shape animal categories across languages, and that this may be consistent with pressure for efficiency (Kemp et al., 2018).

Although we introduced the hierarchy in Figure 4D as an account of category structure across languages, the same hierarchy could potentially serve as a model of hierarchical structure within a single language. This within-language interpretation resembles previous applications of the IB principle to language (Pereira, Tishby, & Lee, 1993), although these applications were based on corpus statistics. The within-language interpretation seems useful in the case of animal taxonomies, a semantic domain with strong hierarchical structure, as opposed to containers and even colors. A possible, yet speculative, reconciliation of the within-language and cross-language interpretations is that speakers may internally represent a hierarchy induced by an evolutionary sequence. For example, Boster (1986) showed that English speakers can recapitulate Berlin and Kay's implicational color hierarchy in a sequential pile-sorting task. Thus, it seems at least possible that a similar phenomenon may also hold for animal categories.

General discussion

Artifacts, animals, and colors are qualitatively different elements of human experience, yet our findings suggest that their semantic representations across languages is governed by the same general information-theoretic principle: efficient coding of meanings into words, as defined by the IB principle. We have shown that this theoretical account, which was previously tested only in the domain of color naming (ZKRT), generalizes to container names and animal taxonomies. This finding resonates with the proposal that word meanings may be shaped by pressure for efficient communication (Kemp et al., 2018). However, it goes beyond that proposal by explaining how pressure for efficiency may account for soft categories and inconsistent naming, both in monolinguals and bilinguals, and how it may relate to language evolution.

An important direction for future research is to test to what extent our results extend to other semantic domains, and ideally, to the lexicon as a whole. While it may not be possible to apply this approach to every aspect of the lexicon, we believe that the theoretical formulation considered here may be broadly applicable across semantic domains.

Acknowledgments

We thank Anne White, Gert Storms, and Barbara Malt for making their container naming and sorting data publicly available. The animal features and familiarity data we used were preprocessed by Kemp et al. (2010). We thank Simon De Deyne for initially sharing these data, and for useful discussions. This study was partially supported by the Gatsby Charitable Foundation (N.Z. and N.T.), and by the Defense

Threat Reduction Agency (N.Z. and T.R.); the content of the study does not necessarily reflect the position or policy of the U.S. government, and no official endorsement should be inferred.

References

Ameel, E., Malt, B. C., Storms, G., & Assche, F. V. (2009). Semantic convergence in the bilingual lexicon. *Journal of Memory and Language*, *60*(2), 270–290.

Ameel, E., Storms, G., Malt, B. C., & Sloman, S. A. (2005). How bilinguals solve the naming problem. *Journal of Memory and Language*, *53*(1), 60–80.

Berlin, B. (1992). *Ethnobiological classification: Principles of categorization of plants and animals in traditional societies*. Princeton University Press.

Berlin, B., & Kay, P. (1969). *Basic color terms: Their universality and evolution*. Berkeley and Los Angeles: University of California Press.

Boster, J. (1986). Can individuals recapitulate the evolutionary development of color lexicons? *Ethnology*, *25*(1), 61–74.

Brown, C. H. (1984). *Language and living things: Uniformities in folk classification and naming*. Rutgers University Press.

De Deyne, S., Verheyen, S., Ameel, E., Vanpaemel, W., Dry, M. J., Voorspoels, W., & Storms, G. (2008). Exemplar by feature applicability matrices and other Dutch normative data for semantic concepts. *Behavior Research Methods*, *40*(4), 1030–1048.

Kemp, C., Chang, K. K., & Lombardi, L. (2010). Category and feature identification. *Acta Psychologica*, *133*(3), 216–233.

Kemp, C., Xu, Y., & Regier, T. (2018). Semantic typology and efficient communication. *Annual Review of Linguistics*, *4*(1).

Malt, B. C. (1995). Category coherence in cross-cultural perspective. *Cognitive Psychology*, *29*(2), 85–148.

Malt, B. C., Sloman, S. A., Gennari, S., Shi, M., & Wang, Y. (1999). Knowing versus naming: Similarity and the linguistic categorization of artifacts. *Journal of Memory and Language*, *40*(2), 230 - 262.

Mayr, E. (1969). The biological meaning of species. *Biological Journal of the Linnean Society*, *1*(3), 311-320.

Pereira, F., Tishby, N., & Lee, L. (1993). Distributional clustering of English words. In *Proceedings of the 31st annual meeting of the Association for Computational Linguistics* (pp. 183–190).

Randall, R. A., & Hunn, E. S. (1984). Do life-forms evolve or do uses for life? Some doubts about Brown's universals hypotheses. *American Ethnologist*, *11*(2), 329-349.

Regier, T., Kemp, C., & Kay, P. (2015). Word meanings across languages support efficient communication. In B. MacWhinney & W. O'Grady (Eds.), *The handbook of language emergence* (pp. 237–263). Hoboken, NJ: Wiley-Blackwell.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*.

Tishby, N., Pereira, F. C., & Bialek, W. (1999). The Information Bottleneck method. In *37th annual Allerton conference on communication, control and computing.*

White, A., Malt, B. C., & Storms, G. (2017). Convergence in the bilingual lexicon: A pre-registered replication of previous studies. *Frontiers in Psychology*, *7*.

Xu, Y., Regier, T., & Malt, B. C. (2016). Historical semantic chaining and efficient communication: The case of container names. *Cognitive Science*, *40*(8), 2081–2094.

Zaslavsky, N., Kemp, C., Regier, T., & Tishby, N. (2018). Efficient compression in color naming and its evolution. *PNAS*, *115*(31), 7937–7942.

Sampling to learn words: Adults and children sample words that reduce referential ambiguity

Martin Zettersten (zettersten@wisc.edu) & Jenny Saffran (jsaffran@wisc.edu)
Department of Psychology, 1202 W. Johnson Street
Madison, WI 53706 USA

Abstract

How do learners gather new information during word learning? We present evidence that adult learners will choose to receive additional training on object-label associations that reduce ambiguity about reference during cross-situational word learning. This ambiguity-reduction strategy is related to improved test performance. We find mixed evidence that children (4-8 years of age) show a similar preference to seek information about words experienced in ambiguous word learning situations. In an initial experiment, children did not preferentially select object-label associations that remained ambiguous during cross-situational word learning. However, this may be explained by some children having relatively high certainty about object-label associations for which they did not see evidence disconfirming their initial hypothesis. In a second experiment that increased the relative ambiguity of two sets of novel object-label associations, we found evidence that children preferentially make selections that reduce ambiguity about novel word meanings.

Keywords: cross-situational word learning; mutual exclusivity; active learning; self-directed learning; sampling

Introduction

What makes us seek out new information during learning? One proposal is that information-seeking behavior is driven by uncertainty reduction (e.g., Kidd & Hayden, 2015). A variety of studies have demonstrated that – at least in some contexts - children are motivated to gather information to reduce the uncertainty after ambiguous or surprising events (Schulz & Bonawitz, 2007; Stahl & Feigenson, 2015).

To what extent does ambiguity-reduction play a role in word learning? A classic problem is how learners disambiguate the meaning of words in potentially ambiguous situations (Quine, 1960). One solution is that children can disambiguate word meanings by tracking co-occurrences of object-label pairs across multiple ambiguous situations (Yu & Smith, 2007). This proposal would be particularly powerful if learners are naturally drawn to isolating object-label associations that have remained ambiguous over the course of past learning (Hidaka, Torii, & Kachergis, 2017). A previous study of cross-situational word learning has shown that being able to actively select sets of object-label pairs to learn about increases participants' accuracy compared to a passive condition in which random sets of objects are presented (Kachergis, Yu, & Shiffrin, 2013). However, we still know little about what sampling strategies adult and child learners display when given the opportunity to control their learning input.

In the current work, we investigated whether adult and child learners seek information that aids in reducing ambiguity about the meaning of novel words. We manipulated the ambiguity of novel word mappings by varying the degree to which object-label pairs co-occurred with one another during cross-situational word learning (Experiments 1A, 1B, 2A) or whether children could use mutual exclusivity to disambiguate the referents of novel words (Experiment 2B). The central question was whether adults and children would choose to learn more about those items that most strongly reduce referential ambiguity.

Experiments 1A & 1B

We tested whether adult learners would seek information that aided in disambiguating reference. Participants completed a cross-situational learning task in which their goal was to learn a set of object-label associations by determining the referent of each label across training. Participants were then given the opportunity to select which object-label association they would hear on the next learning trial. The central question was whether adult learners would make selections that reduce referential ambiguity about the novel object-label associations. We collected data in an online experiment (Experiment 1A) and in an in-lab experiment (Experiment 1B) that we discuss together due to their similarity in design and results.

Method

Participants. For Experiment 1A, we recruited 31 participants through Amazon Mechanical Turk. Three participants were excluded for not passing an initial auditory attention check (2) or for restarting the experiment (1). All participants were assigned to the Fully Ambiguous Condition (n = 28) and paid $0.75 for completing the study.

For Experiment 1B, 62 University of Wisconsin-Madison's undergraduates participated for course credit and were randomly assigned to the Fully Ambiguous Condition (n = 28) or the Partially Ambiguous Condition (n = 34).

Stimuli. The object stimuli were 8 images of novel 'alien' creatures used in previous word learning studies (Partridge, Mcgovern, Yung, & Kidd, 2015). 8 novel word stimuli (*beppo*, *finna*, *guffi*, *kita*, *noopy*, *manu*, *sibu*, *tesser*) were recorded by a female native speaker of English and normalized in duration and average loudness. The association between each label and its target referent and the roles of the stimuli within a condition were randomized across participants. The stimuli were presented using a web-based experiment created using jsPsych (de Leeuw, 2014).

Design & Procedure. The experiment was split into a *Training Phase*, a *Sampling Phase*, and a *Test Phase*.

Training Phase. Participants completed 24 cross-situational learning trials (2 blocks of 12 trials), presented in random order. The goal was to learn the association between eight novel labels and their referents. On each training trial, participants were presented with two referents and two labels. The labels appeared sequentially in random order, both visually and auditorily. Consequently, the association between a particular label and its referent remained ambiguous on any single trial, but could be disambiguated by aggregating information across trials. Each object and its label occurred 6 times across the 24 training trials.

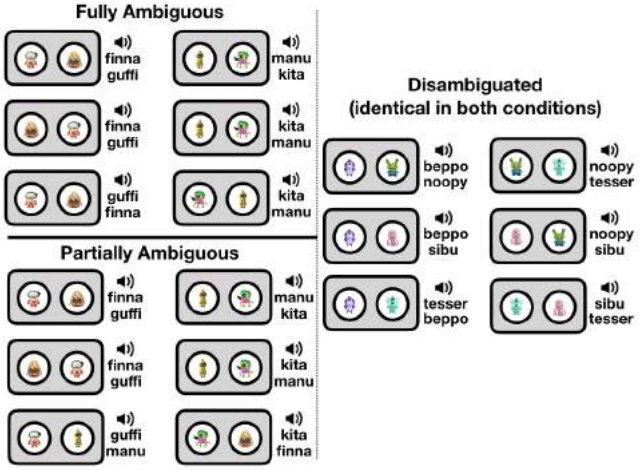


Figure 1. Overview over one block of the Training Phase for the Fully Ambiguous Condition and the Partially Ambiguous Condition

We manipulated whether the object-label associations became disambiguated across trials during training, and therefore, how uncertain participants were at the onset of the *Sampling Phase* about the specific object-label pairs. Across Experiments 1A and 1B, participants were assigned to one of two conditions: the Fully Ambiguous condition or the Partially Ambiguous condition. In the Fully Ambiguous condition, half of the object-label pairs remained ambiguous: two sets of two items were yoked together such that they were never disambiguated across training (ambiguous items; Figure 1, top left). The remaining items in the Fully Ambiguous condition were disambiguated across trials, occurring with three different object-label pairs (disambiguated items; Figure 1, right panel). In the Partially Ambiguous condition, two sets of two objects were grouped such that two specific objects co-occurred on 4 out of their 6 occurrences, but each occurred with one other object from the ambiguous object set on the remaining 2 trials (partially ambiguous items; Figure 1, bottom left). The other four objects were disambiguated as in the Fully Ambiguous condition. Note that across both conditions, participants saw each individual object and label equally frequently.

Sampling Phase. Participants next completed four sampling trials. On each trial, all 8 objects appeared in randomized locations. Participants were instructed to select which of the 8 items they wanted to hear in the next cross-situational learning trial. After participants' selection, a second object was chosen at random from the remaining objects. The two objects and their labels then appeared together in a cross-situational word learning trial with the same structure as in the training phase.

Test Phase. Participants' knowledge of the object-label associations was probed in an 8-AFC recognition test. On each test trial, all 8 objects appeared in randomized locations on the screen, along with one of the 8 labels. Participants were then asked to select the object that went with the label. No feedback was provided after a choice. Participants were tested on each label in random order, for a total of 8 recognition test trials.

Predictions. We predicted that participants would be more likely to choose to learn more about the ambiguous items than about the disambiguated items in the sampling phase. For the Partially Ambiguous condition, we expected participants to have a weaker preference for ambiguous items over the disambiguated items, since adults accurately tracking the co-occurrence evidence could successfully learn all word-referent pairs. We did not predict large differences in test accuracy between items. One possible outcome was that test accuracy would be higher for items that were disambiguated during training. However, another possibility was that ambiguous items could be learned at comparable levels to disambiguated items if participants preferentially sampled ambiguous items.

Results

Sampling choices. We report the results combining the data from Experiments 1A and 1B for convenience – however, qualitatively similar results are obtained when considering the data from Experiment 1A or Experiment 1B separately. We used the lme4 package version 1.1-18-1 in R (version 3.5.1) to fit a logistic mixed-effects model testing participants' likelihood of making an ambiguous selection against a chance level of 0.5 (Bates & Maechler, 2009; R Development Core Team, 2018), including by-subject and by-item random intercepts and a fixed effect for condition. In the Fully Ambiguous condition, participants were more likely to choose ambiguous items than disambiguated items, $b = .59$, $z = 3.61$, $p < .001$. Participants chose an object from the ambiguous set on 63.4% of trials (95% CI = [55.7%, 71.0%] (Figure 2A). Participants in the Partially Ambiguous condition selected the partially ambiguous items on 47.8% of trials (95% CI = [39.1%, 56.5%]), thus showing no sampling preference between the two item types ($p = .64$). Participants were in the Fully Ambiguous condition were more likely than participants in the Partially Ambiguous condition to select the more ambiguous object-label

associations, b = .68, z = 2.64, p = .008. Non-parametric analyses yielded equivalent results.

Test performance. Overall, participants showed learning of the label-object pairs, accurately selecting the correct referent in both the Fully Ambiguous condition (M = 69.2%, 95% CI = [60.8%, 77.5%], chance = 12.5%) and in the Partially Ambiguous condition (M = 77.6%, 95% CI = [67.1%, 88.0%]) (Figure 2B). Notably, within the Fully Ambiguous condition, test accuracy was lower for the ambiguous items (M = 61.6%) than for the disambiguated items (M =76.8%; logistic mixed-effects model with by-subject and by-item random intercepts and a by-subject random slope for item type, z = 3.25, p = .001).

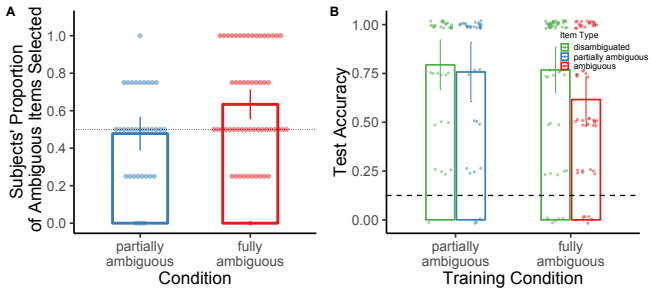


Figure 2. (A) Proportion of more ambiguous items selected in each condition and (B) test accuracy by condition and item type. Error bars in represent within-subject 95% CIs.

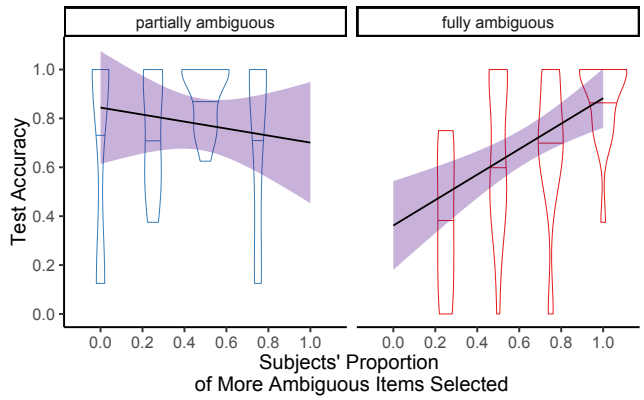


Figure 3. Relationship between choosing more ambiguous items and test accuracy for each condition. Error bands represent +/-1 SE.

Relationship between sampling and test performance. In the Fully Ambiguous condition, participants who chose more objects from the ambiguous set during the sampling phase accurately identified more words at test, r(54) = .48, 95% CI = [0.25, 0.66], p < .001 (Figure 3). There was no significant relationship between participants' tendency to select the partially ambiguous items and their accuracy at test (r(32) = -.11, p = .50).

Discussion

In a cross-situational learning task, adult learners chose to learn more about those object-label pairs that remained ambiguous throughout training. Adults showed this tendency when the object-label pairings were truly ambiguous based on the training evidence (Fully Ambiguous condition), but not when the object-label pairs became disambiguated at any point during training (Partially Ambiguous condition). While participants showed poorer overall learning of the (more difficult) ambiguous object-label pairs, their success at test correlated strongly with the degree to which they chose more ambiguous items during the sampling phase. This experiment provides 'proof-of-concept' evidence that adult learners will seek to reduce ambiguity about object-label associations when given the opportunity to control which items they will learn about.

Experiment 2A

Next, we asked whether children would demonstrate a similar tendency to seek new words that reduce ambiguity during cross-situational learning. As in Experiment 1A, children (4-8 years of age) first completed a cross-situational word learning task. Across training, one set of novel object-label associations could be inferred based on the object-label associations they co-occurred with, while another set of words remained ambiguous. Then, participants were given the opportunity to sample object-label associations presented in isolation, i.e. in unambiguous learning trials. The central question was whether children would prefer to select object-label associations with ambiguous evidence during training, suggesting that children sample words that reduce referential ambiguity.

Method

Participants. We recruited 38 participants (M = 5.9 years, range = 4.1 – 8.1 years, 19 female) at a local children's museum. Two additional participants were excluded due to inattention during experiment.

Stimuli. The object stimuli were 8 images of novel 'alien' creatures used in previous word learning studies (Partridge et al., 2015) and 2 cartoon images of familiar animals (penguin, dog). 8 novel word stimuli (*biffer, deela, guffi, sibu, tibble, leemu, zeevo, pahvy*) and two familiar word stimuli (*penguin*, *dog*) were recorded by a female native speaker of English and normalized in duration and average loudness. The association between each novel label and its novel target referent, as well as the particular roles of the novel word-referent stimuli, were randomized across participants. The stimuli were presented using in a web-based experiment created in jsPsych (de Leeuw, 2014).

Design & Procedure. Children were tested in a quiet room in the children's museum on a 10.1" Samsung Galaxy Note tablet. An experimenter guided children through the experiment by giving instructions at the beginning of each

new phase. The experiment was presented as a game in which a cartoon bear named Teddy would first teach children the names of new alien friends, and then ask children to help her find her friends. The experimenter began with the following introduction:

> In this game, Teddy went up to space and met a bunch of new alien friends. Teddy is going to tell you the names of aliens, and your job is to try to remember which name goes with which alien. Later, you're going to help Teddy find them.

The experiment then proceeded to a Practice Phase, followed by the main experiment consisting of three phases: the Training Phase, the Sampling Phase, and the Test Phase.

Practice Phase. Participants first completed a practice phase in which they encountered the two familiar word object stimuli and two novel object-label associations. We introduced this short practice phase to give children experience with the overall structure of the main experiment under less demanding circumstances, using a smaller set of items and mixing familiar and novel items. First, children were exposed to 4 training practice trials similar in structure to the training trials in the main experiment. On each trial, two referents appeared on the screen on either side of the Teddy character and children heard two labels, one for each object, in random order. On the first trial, children always saw the two familiar items (i.e., the penguin and the dog), followed by a second trial in which children saw two novel object-label associations (i.e., an ambiguous labeling event). On the final two training practice trials, children saw each of the familiar items occur with one of the two novel items (permitting the disambiguation of the novel object-label associations). Next, children saw two sampling practice trials, in which children had the opportunity to select which of the four items they wanted to learn about next, followed by four practice test trials, in which participants' knowledge of the items was tested in a 4-AFC recognition test. The procedure for each of these practice trial types mirrored the procedure for the Sampling Phase and the Test Phase described in more detail below.

Training Phase. Participants completed 9 cross-situational learning trials (3 blocks of 3 trials each). On each training trial, participants saw two referents appear on the screen on either side of the Teddy character and heard the labels of the two objects presented sequentially in random order. Next, the objects switched locations in a brief animation, and participants heard the same two labels presented in the same order. We introduced this trial repetition with flipped locations in order to reduce children's tendency to interpret the labeling event as moving from left to right on the screen, i.e. assuming that the first label went with the object on the left and the second label went with the object on the right.

As in the Fully Ambiguous condition of Experiment 1A, we manipulated whether the object-label associations could be disambiguated across trials during training (Figure 4). Every object-label pair occurred on three cross-situational training trials. Four of the objects occurred with three different object label pairs (disambiguated items). The remaining two object-label associations always occurred with one another (ambiguous items), such that children never saw evidence allowing them to link the two words unambiguously with their respective referent.

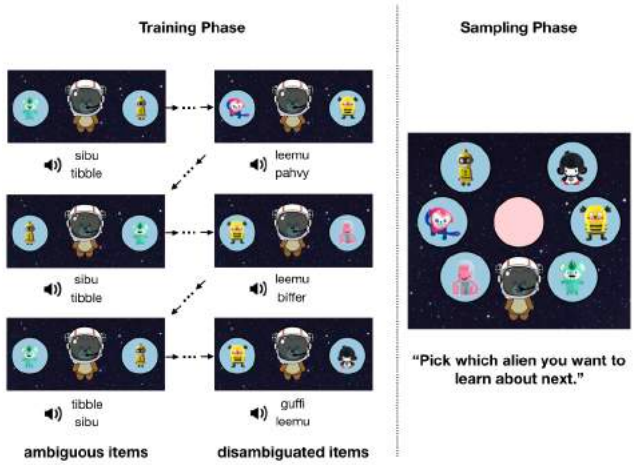


Figure 4. Overview over the design of the Training and Sampling Phase in Experiment 2A

Sampling Phase. After completing the training phase, participants completed four sampling trials. On each sampling trial, all 6 referents appeared in randomized locations on the screen. Participants were instructed to select which of the 6 items they wanted to learn about next (Figure 4). When participants tapped one of the 6 referents, a brief animation moved the item to the center of the screen while the remaining items disappeared, and the referent was subsequently labeled in isolation.

Test Phase. Participants' knowledge of the object-label associations was probed in a 6-AFC recognition test. On each test trial, all 6 referents appeared in randomized locations on the screen surrounding the Teddy character. When participants tapped Teddy in the center of the screen, they heard one of the 6 labels. Participants were instructed to help Teddy by selecting the friend she was looking for. No feedback was provided after a choice. Participants were tested on each label in random order, for a total of 6 recognition test trials.

Predictions. As in Experiment 1A, our main prediction was that children would preferentially select object-label associations that remained ambiguous during the cross-situational word learning trials of the training phase.

Results

Sampling choices. Contrary to our prediction, children did not preferentially select ambiguous object-label associations during the Sampling Phase, b = -0.01, z = -.11, p = .91.

Participants chose an object from the ambiguous set on 32.9% of trials (95% CI = [27.1%, 38.7%]).

Test performance. Overall, participants showed significant learning of the label-object pairs, choosing the correct object to go with a label at above-chance levels (chance = 0.167), M = 38.6%, 95% CI = [30.7%, 46.5%], $t(37) = 5.65$, $p < .001$. However, surprisingly, children performed more accurately on the ambiguous items (M = 48.6%, 95% CI = [36.9%, 60.4%]) than on the disambiguated items (M = 33.6%, 95% CI = [24.9%, 42.2%]), $b = .68$, $z = 2.23$, $p = .028$. When tested on ambiguous items, children had a strong preference to select one of the two ambiguous objects (61.8% of trials, 95% CI = [50.7%, 72.9%]) over the four disambiguated objects (chance = 0.33). When tested on disambiguated items, children tended not to choose the two ambiguous objects, selecting them on only 18.4% of trials (95% CI = [12.8%, 24.1%]).

Discussion

Unlike adult learners, children did not show a preference for selecting object-label associations for which they had experienced ambiguous evidence during training. Interestingly, children performed better at test for ambiguous object-label associations than for object-label associations that were disambiguated across training trials. There are likely two reasons why children showed higher accuracy on the ambiguous items. First, since the two ambiguous items always co-occurred with one another, the training could help learners constrain the set of possible competitors for a given ambiguous label to two objects (compared to four possible objects for the disambiguated items). Indeed, children appeared to constrain their choices to the two objects that co-occurred on ambiguous trials when tested on their respective labels and rarely chose these objects when tested on the labels that occurred with the disambiguated objects

Second, anecdotally, we observed that many children explicitly pointed to specific objects during training while listening to each label and even repeated the respective label for each object. This behavior may indicate that some children were making an explicit hypothesis about each word mapping (Trueswell, Medina, Hafri, & Gleitman, 2013). If a child formed a specific hypothesis about the mapping between the two labels and objects on the first ambiguous trial, they would subsequently hear evidence that would appear to confirm their hypothesis: the two labels and the two objects would occur together again on the subsequent two training trials. "Hypothesis-testers" would never experience evidence disconfirming their initial hypotheses and thus have a 50% chance of responding correctly at test for these items (note that our child participants' test accuracy was 48.6% on average). Crucially, one consequence of learners approaching the task in this manner is that the two object-label associations deemed "ambiguous" according to the experimental design may have actually appeared *less* ambiguous to children performing the task than the putatively disambiguated items. Thus, in our next step, we adapted the task to create a learning situation in which one set of object-label associations would be more clearly ambiguous from the standpoint of the child learner.

Experiment 2B

In Experiment 2B, we sought to increase the likelihood that children would perceive some novel object-label associations as more ambiguous than others. We used mutual exclusivity to increase the ease with which children could infer word-referent pairs for one set of novel objects (Markman & Wachtel, 1988) while maintaining the ambiguity of a second set of novel word-referent pairs as in the previous experiments. By giving children the opportunity to infer the referents for novel objects occurring in mutual exclusivity trials, we aimed to make it easier for children to recognize the referential ambiguity of novel object-label associations that always co-occurred.

Method

Participants. We recruited 53 participants (M = 5.7 years, range = 4.1 – 7.9 years, 32 female) at a local children's museum. One additional participant was excluded due to experimenter error.

Stimuli. The novel object and word stimuli were six images and recordings composed of a subset of the items used in Experiment 2A. In addition, 4 cartoon images of familiar animals (cow, dog, monkey, pig) along with audio recordings of their respective labels were used. All word stimuli were recorded by the same female native speaker of English and normalized in duration and average loudness.

Design & Procedure. The procedure and testing conditions were identical to Experiment 2A. The experiment followed the same structure as Experiment 2A, beginning with a *Practice Phase* and then proceeding through three phases: *Training Phase*, *Sampling Phase*, and *Test Phase*.

Training Phase. Participants completed 9 cross-situational learning trials (3 blocks of 3 trials each) with 6 object-label pairs, two familiar object-label pairs (e.g., pig and dog) and four novel object-label pairs chosen randomly from the set of novel stimuli. As in Experiment 2A, on each trial, participants saw two referents appear on the screen and heard two labels presented in random order. Two novel object-label associations always occurred with one another (ambiguous items), mirroring the ambiguity manipulation from Experiments 1A/B and 2A. The two remaining novel object-label associations were each yoked to one of the two familiar object-label pairs (i.e., one alien always occurred with the dog image, while the other always occurred with the pig image; mutual exclusivity items). We reasoned that children would successfully disambiguate reference for mutual exclusivity items (i.e., when seeing an image of a dog and a novel "alien", on hearing the words *leemu* and *dog*, children would successfully infer that *leemu* referred to

the novel alien). This would make it more likely that the ambiguous items would be perceived by child learners as having high referential uncertainty. As in previous experiments, all novel objects and their labels occurred equally frequently across the training phase.

Sampling Phase. Participants next completed two sampling trials. On each trial, the four novel objects appeared on the screen and children were instructed to choose which object they wanted learn more about. The procedure was otherwise identical to Experiment 2A.

Test Phase. Participants' knowledge of the six words from the training phase (4 novel, 2 familiar words) was tested in a 6-AFC recognition task as in Experiment 2A.

Results

Sampling choices. Children preferentially selected ambiguous object-label associations during the Sampling Phase, b = .58, z = 2.87, p = .004. Participants chose an object from the ambiguous set on 64.2% of trials (95% CI = [55.0%, 73.3%]) (chance level = 0.5; Figure 5A). The likelihood of children making ambiguous selections increased with age, b = .49, z = 2.42, p = .016 (logistic mixed-effects models; Figure 5B).

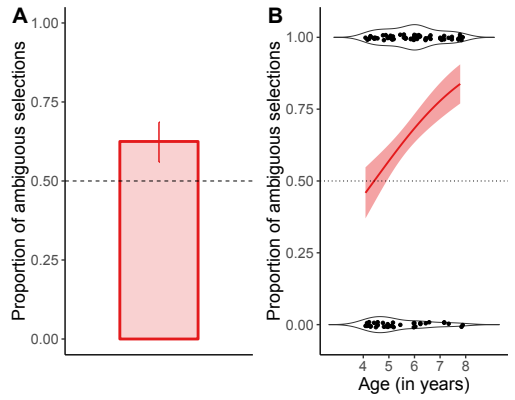


Figure 5. Proportion of ambiguous item selections in Experiment 2B overall (A) and across age (B). Error bars represent 95% CIs and error bands are +1/-1 SEs based on model estimates.

Test performance. Overall, participants showed significant learning of the label-object pairs, choosing the correct object to go with a label at above-chance levels (chance selection of novel object = 0.25), M = 59.9%, 95% CI = [50.4%, 69.4%], $t(52)$ = 7.38, p < .001. Accuracy for mutual exclusivity items (M = 64.2%, 95% CI = [53.2%, 75.1%]) and for the ambiguous items (M = 55.7%, 95% CI = [44.0%, 67.3%]) was similar, b =.45, z = 1.20, p = .23.

Discussion

When given the opportunity to select which object-label pairs they wanted to learn more about, 4-8-year-olds preferentially selected object-label pairs that remained ambiguous during training over object-label pairs that could be disambiguated through mutual exclusivity. These findings demonstrate that – at least in some ambiguous word learning situations – children prefer to select learning events that aid in reducing referential uncertainty. The tendency to make ambiguity-reducing selections began to emerge around 5 years of age in our sample.

General Discussion

When learning the referents of novel labels in ambiguous contexts, adult learners chose to learn more about object-label associations that remained more ambiguous at the end of training. These choices appear to help learners improve performance: participants' learning performance at test was higher if they had selected more ambiguous items during the Sampling Phase. It is interesting to note the modest magnitude of adults' preference on the task: ambiguous items were selected on slightly less than two-thirds of adults' sampling trials. This may be partly related to the design of the sampling phase, which allowed for a number of potentially successful sampling strategies (e.g., selecting a known word on each sampling trial in order to hear that known word in combination with other words). However, another intriguing possibility is that there are individual differences in how adults organize their learning, and that these differences may lead to distinct learning outcomes.

We find mixed evidence that children spontaneously sample object-label associations that reduce ambiguity. When presented with a similar task, 4-8-year-olds did not choose to learn about object-label associations that remained ambiguous during training. However, we think this result may be partially explained by the fact that word-referent pairs occurring in ambiguous contexts also never provided children with evidence that could disconfirm an existing hypothesis about word reference. In a simplified design that highlighted the ambiguous nature of the trials in which two referents always occurred together, older children in our sample chose to learn about items that reduced uncertainty about the words' referents.

Children have substantial control over their "curriculum" as they learn new words in the world (Smith, Jayaraman, Clerkin, & Yu, 2018), with potentially immense consequences for the difficulty of the learning problem they face (Hidaka et al., 2017). The results from Experiment 2B are consistent with results from domains such as causal learning that suggest that children are motivated to explore novel objects when presented with confounded evidence (e.g., Schulz & Bonawitz, 2007). However, the limits on the extent to which children spontaneously make ambiguity-reducing selections also raise important questions about what sampling strategies children employ when in control of what they learn next. A key question for future research will be investigating how children's sampling strategies and the structure of their environment interact to support learning.

Acknowledgements

This research was supported by NSF-GRFP DGE-1256259 awarded to MZ and grants from the NICHD awarded to JS (R37HD037466) and the Waisman Center (U54 HD090256). We would like to thank Joseph Austerweil for advice on the project and Tess Gapinski, Louisa Forrest, and Andrew Kressin for aiding in data collection.

References

Bates, D., & Maechler, M. (2009). *lme4: Linear mixed-effects models using S4 classes.*

de Leeuw, J. R. (2014). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, 1–12.

Hidaka, S., Torii, T., & Kachergis, G. (2017). Quantifying the impact of active choice in word learning. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.), *Proceedings of the 39th Annual Meeting of the Cognitive Science Society* (pp. 519–525). London, UK: Cognitive Science Society.

Kachergis, G., Yu, C., & Shiffrin, R. M. (2013). Actively learning object names across ambiguous situations. *Topics in Cognitive Science*, *5*, 200–213.

Kidd, C., & Hayden, B. Y. (2015). The psychology and neuroscience of curiosity. *Neuron*, *88*, 449–460.

Markman, E. M., & Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology*, *20*, 121–157.

Partridge, E., McGovern, M. G., Yung, A., & Kidd, C. (2015). Young children's self-directed information gathering on touchscreens. In R. Dale, C. Jennings, P. Maglio, T. Matlock, D. Noelle, A. Warlaumont, & J. Yoshimi (Eds.), *Proceedings of the Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Quine, W. V. O. (1960). *Word and object*. Cambridge, MA: MIT Press.

R Development Core Team. (2018). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Schulz, L. E., & Bonawitz, E. B. (2007). Serious fun: preschoolers engage in more exploratory play when evidence is confounded. *Developmental Psychology*, *43*, 1045–1050.

Smith, L. B., Jayaraman, S., Clerkin, E., & Yu, C. (2018). The developing infant creates a curriculum for statistical learning. *Trends in Cognitive Sciences*, *22*, 325–336.

Stahl, A. E., & Feigenson, L. (2015). Observing the unexpected enhances infants' learning and exploration. *Science*, *348*(6230), 91–94.

Trueswell, J. C., Medina, T. N., Hafri, A., & Gleitman, L. R. (2013). Propose but verify: fast mapping meets cross-situational word learning. *Cognitive Psychology*, *66*, 126–156.

Yu, C., & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, *18*, 414–420.

Availability-Based Production Predicts Speakers' Real-time Choices of Mandarin Classifiers

Meilin Zhan (meilinz@mit.edu) Roger Levy (rplevy@mit.edu)
Department of Brain & Cognitive Sciences, Massachusetts Institute of Technology
43 Vassar Street, Cambridge, MA 02139 USA

Abstract

Speakers often face choices as to how to structure their intended message into an utterance. Here we investigate the influence of contextual predictability on the encoding of linguistic content manifested by speaker choice in a classifier language, Mandarin Chinese. In Mandarin, modifying a noun with a numeral obligatorily requires the use of a classifier. While different nouns are compatible with different SPECIFIC classifiers, there is a GENERAL classifier that can be used with most nouns. When the upcoming noun is less predictable, using a more specific classifier would reduce the noun's surprisal, potentially facilitating comprehension (predicted to be preferred under Uniform Information Density, Levy & Jaeger, 2007), but the specific classifier may be dispreferred from a production standpoint if the general classifier is more easily available (predicted by Availability-Based Production; Bock, 1987; Ferreira & Dell, 2000). Here we report a picture-naming experiment confirming two distinctive predictions made by Availability-Based Production.

Keywords: Language production; speaker choice; Chinese classifiers; noun predictability

Introduction

The simple act of speaking may typically seem effortless, but it is extraordinarily complex. Speakers must plan the message they wish to convey, choose words and constructions that accurately encode that message, organize those words and constructions into linearly-sequenced utterances, keep track of what has been said, and execute each part of their speaking plans at the correct time. Throughout this process, speakers face choices in structuring their intended message into an utterance. One central question for a computationally precise theory of language production is thus: When multiple options are available to express more or less the same meaning, what general principles govern a speaker's choice? To what extent do speakers make choices that potentially facilitate comprehenders, and to what extent do they make choices that are preferable from a production standpoint? Here we approach these questions from the standpoint of contextual predictability, which is known to affect a wide range of speaker choices. Specifically, we investigate the influence of contextual predictability on the encoding of linguistic content manifested by speaker choice in a classifier language. Two major theories of sentence production, Availability-Based Production (ABP; Bock, 1987; Ferreira & Dell, 2000) and Uniform Information Density (UID; Levy & Jaeger, 2007; Jaeger, 2010), make conflicting predictions about the distribution of speaker choices when more than one classifier could be used in a given context. We report a language production experiment on classifier choice that adjudicates between these theories.

In languages with a grammaticalized count–mass distinction, such as English, count nouns such as *table* can be used with a numeral directly and typically exhibit a singular–plural morphological marking (e.g., *one table, three tables*), whereas mass nouns such as *sand* cannot co-occur with numerals directly without some kind of measure word (e.g., *three cups of sand*) and do not have a plural morphology on the noun (e.g., **three sands*). In classifier languages such Mandarin, in contrast, nouns lack obligatory singular–plural morphological marking and cannot directly co-occur with numerals. Instead, a numeral classifier is required when a noun is modified by a numeral or a demonstrative. Linguists generally agree that there is a distinction between two types of Chinese classifiers: count classifiers, which we focus on here, and mass classifiers (Tai, 1994; Cheng, Sybesma, et al., 1998; Li, Barner, & Huang, 2008). [1] Among count classifiers, which are used with nouns that denote individuals or groups of individuals, different SPECIFIC classifiers are compatible with different nouns, but the GENERAL classifier *ge* (个) can be used with almost any noun. Often, the choice of general versus specific classifier for a given noun carries little to no meaning distinction for the utterance, as illustrated in (1) and (2) below.

(1) 我 卖了 三 **台** 电脑
wo mai-le san **tai** diannao
I sold three CL.machinery computer ("I sold three computers")

(2) 我 卖了 三 **个** 电脑
wo mai-le san **ge** diannao
I sold three CL.general computer ("I sold three computers")

In this study, we focus on speaker choice between general and specific count classifiers for nouns where both options

[1] A count classifier (e.g., two CL.top hat ("two hats")) is used to categorize a class of noun entities in reference to their salient perceptual properties, which are often permanently associated with the entities named by the class of nouns. A mass classifier (e.g., two box (of) hat ("two boxes of hats")) creates a unit and form a temporary relationship with the noun. Because using different mass classifiers often change the semantics of the noun phrase, here we only focus on count classifiers (henceforth, classifiers).

convey more or less the same meaning. When the upcoming noun is unpredictable, a specific classifier would constrain the range of possible nouns more than the general classifier, thus increasing the predictability of the upcoming noun and potentially benefiting comprehension. The Uniform Information Density account thus predicts that speakers will prefer specific classifiers for unpredictable nouns. However, Availability-Based Production predicts that the specific classifier may be dispreferred from a production standpoint if the general classifier is more easily available. Which of these two accounts better predict classifier choice in real-time production? In other words, does noun predictability affect classifier choice, and if so, in which direction? Here we use a picture-naming experiment to address this question.

Before diving into the experiment, we first briefly introduce why we focus on predictability effects and how the two accounts predict speaker choices with regard to optional reduction in language.

Predictability Effects on Optional Reduction

It has been shown that contextual predictability plays a role in optional reduction in language, where more predictable content tend to yield a greater rate of reduction in the linguistic form. At the lexical level, predictable words are phonetically reduced (Jurafsky, Bell, Gregory, & Raymond, 2001; Bell, Brenier, Gregory, Girand, & Jurafsky, 2009; Seyfarth, 2014) and tend to have shorter forms (Piantadosi, Tily, & Gibson, 2011; Mahowald, Fedorenko, Piantadosi, & Gibson, 2013). At the syntactic level, optional function words are more likely to be omitted when the phrase they introduce is predictable (Levy & Jaeger, 2007; Jaeger, 2010). For example, in English relative clauses (henceforth RCs) such as (3), speakers can but do not have to produce the relativizer *that*. We refer to the omission of *that* as OPTIONAL REDUCTION.

(3) I created a mobile app dancers like.

(4) I created a mobile app that dancers like.

For optional function word omission, predictability effects have been argued to be consistent with both the speaker-oriented account of Availability-Based Production, where the speaker mentions material that is readily available first, and the potentially audience-oriented account of Uniform Information Density, where the speaker aims to convey information at a relatively constant rate. These two accounts have proven difficult to disentangle empirically. For different reasons, both accounts predict that the less predictable the clause introduced by the function word, the more likely the speaker would be to produce the function word *that*.

Uniform Information Density

Uniform Information Density proposes that within boundaries defined by grammar, when multiple options are available to express the message, speakers prefer the variant that distributes information density more uniformly throughout the utterance, to lower the chance of information loss or miscommunication (Levy & Jaeger, 2007; Jaeger, 2010). Multiple formalizations are possible under this account (Genzel & Charniak, 2002; Aylett & Turk, 2004; Maurits, Navarro, & Perfors, 2010; Levy, 2018).

In (3), where the relativizer *that* is omitted, the first word of the relative clause w_1 (*dancers* in this case) is highly unpredictable and would convey two pieces of information: both the onset of the relative clause and part of the content of the relative clause itself. These both contribute to the information content of w_1, which can be measured using SURPRISAL, the negative log-probability of the word in context: $-\log P(w|\text{Context})$ (Hale, 2001; Levy, 2008; Demberg & Keller, 2008; Smith & Levy, 2013). In (4), having *that* at the onset of the RC splits these two pieces of information apart, offloading the relative clause's onset onto *that* so that *dancers* only conveys relative clause-internal content and thus has lower information content, potentially avoiding a peak in information density and thus facilitating comprehension.

Availability-Based Production

Availability-Based Production proposes that production is more efficient if speaker mentions material that is readily available first. According to ABP, speaker choice is governed by: 1) when a part of a message needs to be expressed within an utterance; 2) when the linguistic material to encode that part of the message becomes available (Bock, 1987; Ferreira & Dell, 2000). Specifically, if material that encodes a part of the message becomes available when it comes time to convey that part of the message, it will be used. However, if that material is not yet available, then other available material will be used, as long as it is compatible with the grammatical context produced thus far and it does not cut off the speaker's future path to expressing the desired content. This is also referred to as THE PRINCIPLE OF IMMEDIATE MENTION (Ferreira & Dell, 2000).

Suppose a speaker has just uttered the word *app* in (3) and has in mind to convey the remainder of the utterance meaning as a relative clause. If the word *dancers* becomes available quickly, then according to the principle of immediate mention, a sentence without *that* should be produced (see (3)). If *dancers* does not become available quickly, however, ABP predicts that the speaker will utter *that* to buy more time for *dancers* to become available. (Note that this account relies on an implicit auxiliary assumption that that *that* will generally become available quite quickly; this assumption is rendered plausible by the fact that it is a high-frequency word used in a wide variety of contexts.) If the first word of the RC takes longer to become available the lower its contextual predictability—an assumption consistent with previous work on picture naming (Oldfield & Wingfield, 1965) and word naming (Balota & Chumbley, 1985)—then the less predictable the relative clause, the lower the probability that its first word, *dancers*, will be available at when the speaker reaches the RC, and the higher the probability that the speaker will use *that*. Since an RC is required after *app* in order for it to be followed by the word *dancers*, the lower the contextual

probability of an RC the lower the contextual probability of its first word, predicting the empirically observed relationship between phrasal onset probability and optional function word omission rate.

Distinguishing theories of predictability-driven speaker choice

Although UID and ABP are substantially different theories of what drives speaker choice, they make the same prediction for the effect of contextual predictability on optional reduction of function words for cases such as (3). It is thus intrinsically difficult to use optional reduction phenomena to tease these accounts apart. Prior work (Jaeger, 2010) acknowledged this entanglement of the predictions and attempted to tease these accounts apart via joint modeling using logistic regression. There are other phenomena for which the accounts make similar predictions, as well. Consider the case of ordering choices for words or phrases, such as subject–object versus object–subject word order for languages in which both options are available, such as Russian. Availability-Based Production predicts that whichever becomes available earlier will be uttered first (Levelt & Maasen, 1981); if the lexical encodings of more contextually predictable references tend to become available more quickly, then more predictable arguments will tend to be uttered first. This prediction is indeed likely to be true: a given-before-new word order preference is widely recognized to influence many languages (Behaghel, 1930; Prince, 1981; Gundel, 1988), and discourse-given entities are generally more contextually predictable than discourse-new entities. But UID turns out to make the same prediction. Two arguments of the same verb generally carry mutual information about each other, so any argument will typically be less surprising if it is the latter of the two. Thus, putting the argument that is more predictable from sentence-external context before the less-predictable argument will lead to a more uniform information density profile and will be preferred.

In the case of speaker choice for Mandarin classifiers, however, UID and ABP turn out to make different predictions as we describe in the next section. The empirical facts regarding speaker choice for classifiers are thus of considerable theoretical interest.

Predictions on Mandarin Classifiers

Zhan and Levy (2018) have argued that UID and ABP make different predictions on Mandarin Classifier use with regard to noun predictability. As regards UID, the choice between a specific classifier and a general classifier will typically affect the contextual predictability of the noun modified by the classifier. In particular, a specific classifier constrains the space of possible upcoming nouns more tightly than the general classifier (Klein, Carlson, Li, Jaeger, & Tanenhaus, 2012), thus generally reducing the actual noun's surprisal. The UID hypothesis thus predicts that speakers choose a **specific** classifier more often when the noun predictability would otherwise be low than when the noun is more predictable. This is because the use of a specific classifier makes the distribution of information density more even between the noun and the classifier.

Availability-Based Production, on the other hand, makes different predictions than UID. The fundamental prediction of ABP is that the harder the noun lemma is to access, the less often the speaker will use a specific classifier, provided two plausible assumptions. First, the general classifier *ge* is always available, regardless of the identity of the upcoming noun, as it is the most commonly used classifier and is compatible with practically every noun. Second, in order to access and produce an appropriate specific classifier, a speaker must complete at least some part of the planning process for the production of the nominal reference: accessing the noun lemma, or minimally accessing the key semantic properties of the referent that determine its match with the specific classifier. On these two assumptions, any feature of the language production context that makes the noun lemma less accessible or that more generally makes noun planning more difficult will favor the general classifier. In out-of-linguistic-context picture naming, for example, noun lemma accessibility is known to be driven by noun frequency (Oldfield & Wingfield, 1965). The lower the noun frequency, the less accessible the noun lemma, thus the less likely a specific classifier will be used. To make predictions about the effect of noun predictability on classifier choice in linguistic contexts, we must add a third, theoretically plausible assumption: that less predictable noun lemmas are harder and/or slower to access than more predictable noun lemmas. On these three assumptions, in corpus data the link between noun lemma accessibility and classifier choice will show up as an effect of noun predictability, which by hypothesis is determining noun lemma accessibility. For less predictable nouns, their specific classifiers will less likely be available to the speaker when the time comes to initiate classifier production. Because noun lemmas need to be accessed in order to produce specific classifiers, and the less predictable the noun, the harder the noun lemma is to access and hence the specific classifier associated with the noun becomes available by the time a classifier needs to be produced.

In other words, the link between noun lemma accessibility and classifier choice will manifest in different predictions depending on whether one we are looking at usage in linguistic context versus picture-naming. Under our assumptions about ABP, we can identify three predictions. First, in out-of-context picture naming, speakers should as described above choose the **general** classifier more often the more frequent the noun (provided there is high naming agreement for the picture, so that there is not competition among nouns that affects the production process). Second, in corpus data, speakers should as described above choose a **general** classifier more often the less predictable the noun. Finally, we can add a third prediction based on the temporal dependence of specific classifier production on noun planning: when speakers are under greater time pressure, they should produce the

general classifier more often, as it can be used even when noun planning has not proceeded far enough for a specific classifier to be available.

Zhan and Levy (2018) tested the second prediction in an investigation of naturally occurring texts, using language models to estimate noun predictability and mixed logistic regression to infer its relationship with classifier choice. They found that the less predictable the noun, the lower the rate of using a specific classifier. While these results lend support for the Availability-Based Production account, the study has some limitations. One limitation is that the corpus being used was a collection of online news texts. Written language may serve as a first approximation of testing theories of language production, but it would be ideal to use real-time language production task to further test the hypotheses. Another limitation is that there was no experimental control of context, so it is possible that predictability was confounded with some other contextual factor that was not included in their regression analysis but that is actually responsible for speaker choice.

In the present study, we use a real-time language production task involving picture naming varying noun frequency and whether speakers are put under time pressure, allowing us to further investigate the two models of language production by testing the first and third predictions described above.

Methods

We used a picture-naming experiment to test the predictions of Uniform Information Density and Availability-Based Production by manipulating noun frequency and whether or not the speaker is under time pressure. This picture-naming experiment offers a simple yet effective way to elicit real-time language production.

Participants

Thirty-six self-reported native speakers of Mandarin Chinese were recruited via Witmart, a China-based online crowdsourcing platform. Participants received compensation for their time.

Materials

We adapted images from the Pool of Pairs of Related Objects (POPORO) (Kovalenko, Chaumon, & Busch, 2012) image set to create our visual stimuli. We selected images from the image set based on the following criteria: 1) the image can be described by a count noun; 2) the preferred count noun is compatible with the general classifier and at least one specific classifier. We developed a web-based version of the experiment using `jsPsych` (de Leeuw, 2015), a JavaScript library for creating behavioral experiments in a web browser. We estimated the frequencies of occurrence of the preferred count nouns from `SogouW` (Sogou, 2006), a word frequency dictionary for online texts in Chinese.

Procedure

Participants were presented with scenes of various countable object kinds such as cabbages and tables. Figure 1 shows a sample display. In each scene, there were several instances of the same object kind. The number of objects in each trial varied from two to four. Participants were asked to describe the number and the name of the object in Mandarin, eliciting utterances such as “three CL chairs” which we recorded.

Figure 1: Sample visual display for the picture-naming experiment. The red dot below the picture is the recording light. Below it is the text indicating the status of the recorder, in this case, it is ”recording stopped”. The English translation for the sentence in the bottom is: “Please describe the number and the name of the objects in the picture.”

Participants were assigned to one of the two conditions. In the **Quick** condition, recording started 50 ms after the picture was shown, indicated by a recording light at the bottom of the picture. Each trial ended after 5 seconds of recording, and the next trial began automatically. In the **Slow** condition, recording stared 3 seconds after the picture was shown, and participants clicked on the screen to move toward the next trial.

Predictions

Availability-Based Production predicts that (1) the rate of specific classifier use will be lower in the **Quick** condition, when speakers are under time pressure, than in the **Slow** condition; and (2) the rate of specific classifier use will be lower for less frequent nouns. This latter prediction derives from evidence that lexical access, as manifested by response latencies, takes longer for lower frequency words in language production experiments requiring word production outside of sentence context; this holds not only for picture naming (Oldfield & Wingfield, 1965), as we require of participants here, but also of visually-presented word naming (Balota & Chumbley, 1985). If lower-frequency nouns are slower to ac-

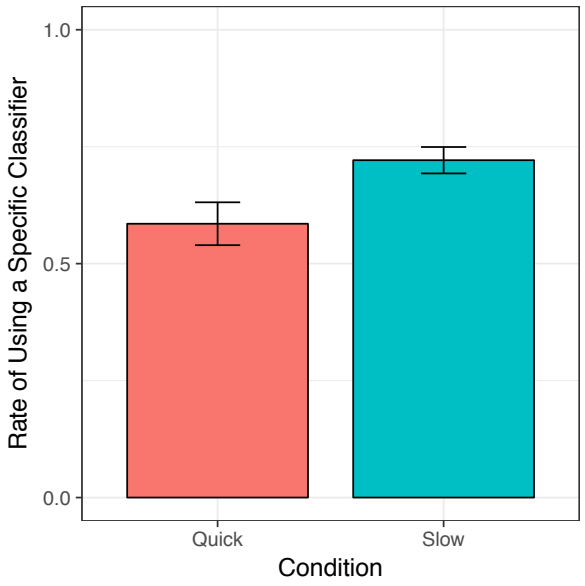


Figure 2: Quick vs. Slow manipulation and rate of using a specific classifier as opposed to the general classifier *ge*. Error bars are standard errors over by-participant means.

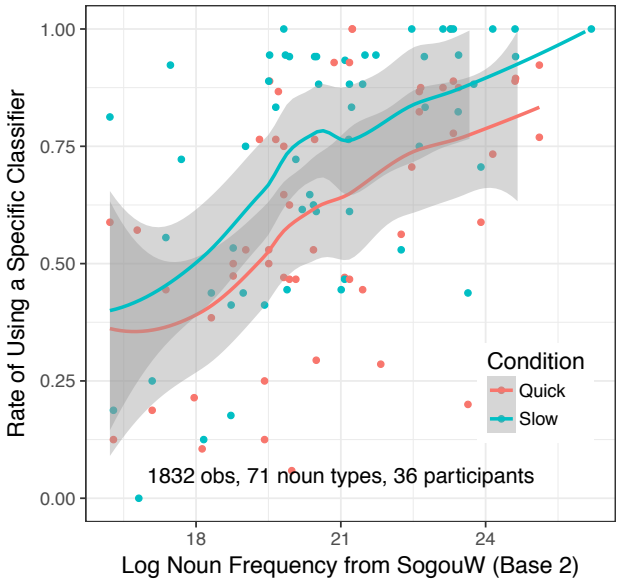


Figure 3: The relationship between noun frequency and rate of specific vs. general classifier use in picture naming.

cess, their specific classifiers may also be slower to access and thus less often used than the general classifier, which is available for all nouns.

The predictions of Uniform Information Density for the effect of the **Quick/Slow** manipulation are unclear. As regards noun frequency, UID predicts that if anything low-frequency nouns should have a *higher* rate of specific classifier usage, as a noun's frequency may effectively serve as its predictability in this experimental setting without broader linguistic context.

Analysis

Audio responses were first transcribed to texts using Google's speech-to-text application programming interface (API), and then checked manually to correct transcription errors. We excluded trials when the participant did not produce a classifier or a noun. For each item, we used the nouns that were most frequently produced as the noun for that item. We also compiled a list of acceptable nouns for each items, and excluded nouns that were not on the list.

We used a mixed-effect logit model to investigate whether noun frequency and time pressure affect classifier choice. The dependent variable was the binary outcome of whether the general classifier or a specific classifier was produced. For each noun type, we also identified its preferred specific classifier (using native speaker introspective judgment and predominant responses by experimental participant, which were concordant). We included two predictors in the analysis: log noun frequency and condition. We included noun, preferred specific classifier, and participant as random factors. We used the maximal random-effects structure with respect to these two predictors (Barr, Levy, Scheepers, & Tily, 2013). For condition, this entailed random slopes by noun and by preferred specific classifier, but not participant because the condition manipulation was between subject. For log noun frequency, this entailed a random slope by participant. The full formula in the style of `R`'s `lme4` is:

```
response ~ log_noun_freq + condition
+(1+condition|noun)
+(1+condition|preferred_spec_cl)
+(1+log_noun_freq|participant)
```

Statistical significance was determined using Markov chain Monte Carlo (MCMC) methods in the R package `MCMCglmm` (Hadfield, 2010) with p-values based on the posterior distribution of regression model parameters with an uninformative prior, as is common for MCMC-based mixed model fitting (Baayen, Davidson, & Bates, 2008).

Results

Looking just at the **Quick/Slow** contrast, we find (Figure 2) that speakers produced more instances of the general classifier when they are under time pressure than when they are not ($p < 0.05$), suggesting that specific classifiers are slower than the general classifier to access and thus supporting the Availability-Based Production account.

Further breaking out our results by noun log-frequency, we find (Figure 3) that the lower frequency the noun, the more likely a **general** classifier is to be produced ($p < 0.001$). This pattern holds within both experimental conditions and is consistent with previous results from the corpus analysis (Zhan & Levy, 2018), and also supports the Availability-Based Production account.

One potential concern arises in the frequencies of the different specific classifiers. One could argue that it was not the noun's frequency that determined the use of the general classifier, rather it was the frequency of the preferred specific classifier that affected the choice of which classifier was used. In the mixed-effect logit model presented above, we included a by-specific-classifier random intercept, which largely rules out the possibility that specific classifier frequency were confounding the effect of noun frequency. To further investigate this issue, we tried a version of our regression model that also includes a fixed effect for the log frequency of preferred specific classifier as a control factor. We did not find any qualitative change to the results. The effects of noun frequency ($p < 0.001$) and condition ($p < 0.05$) on classifier choice remain qualitatively similar to the results of the original model. Furthermore, in this new analysis, there is no effect of specific classifier frequency on classifier choice ($p = 0.483$). This additional analysis suggests that it is unlikely that specific classifier frequency to be driving the effect of noun frequency.

Conclusion

Using a picture-naming experiment, we show that Availability-Based Production predicts speakers' real-time choices of Mandarin Chinese. The lower a noun's frequency, the more likely a general classifier is to be used. We also found that the use of classifier is moderated by whether the speaker is under time pressure when speaking, where the speaker tends to produce more instances of the general classifier if they are under greater time pressure to speak. This real-time effect confirms that the general classifier is easily accessible when the speaker is about to produce a noun phrase with numeral.

Taken together, the present study and previous corpus work on Mandarin classifier (Zhan & Levy, 2018) offer converging evidence regarding the relationship between noun frequency, predictability, and classifier choice, and thus shed light on the mechanisms influencing speaker choice. While the corpus work provides ecological validity through naturalistic data, the experimental work helps us to eliminate potential correlation-based confounds with a clean setup, and enables us to get dense data that are theoretically important but naturalistically sparse. When combined together, this work is complementary with previous corpus work and together paint a more comprehensive picture of language production.

These studies also underscore the importance of investigating a wide variety of speaker choice phenomena, taking advantage of the many types of phenomena offered by the languages of the world. Optional reduction and word order choice are perhaps the best-studied types of such alternations, but they have proven ill-suited to teasing apart the predictions of Uniform Information Density and Availability-Based Production. The approach taken here could be extended to the many types of classifier systems in languages around the world, and might inspire investigation of yet different speaker choice configurations that shed new insights into the mechanisms of language production.

In future work on classifier choice, we plan to investigate other potentially relevant factors such as mutual information. It is possible that some classifier-noun pairs are especially prominent and accessible in memory. If the mutual information between the noun and classifier is high, speakers might be more likely to use that classifier for the noun selected. Although we have not found direct evidence supporting the UID hypothesis, it is possible that this particular experimental setting is not very communicative in nature. In future work, we plan to do an real-time language production experiment in a more communicative setting, with virtual or real listeners in the experiment to further test speaker choice in language production. We also plan to add additional production measures such as phonetic reduction of classifiers, pause durations, and disfluencies to enrich our understanding of language production.

Viewed most broadly, using speaker choice in classifier production as a test case has helped us investigate computationally explicit theories of language production, and advance our understanding of the psychological processes involved in converting our thoughts to speech.

Acknowledgements

We gratefully acknowledge valuable feedback from members of MIT Computational Psycholinguistics Laboratory and three anonymous reviewers, technical advice on web-based experiment development from Jon Gauthier and Wenzhe Qiu. This research was funded by NSF grants BCS-1551866 to Roger Levy, and BCS-1844723 to Roger Levy and Meilin Zhan.

References

Aylett, M., & Turk, A. (2004). The Smooth Signal Redundancy Hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. , *47*(1), 31–56.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412.

Balota, D. A., & Chumbley, J. I. (1985). The locus of word-frequency effects in the pronunciation task: Lexical access and/or production. *Journal of Memory and Language*, *24*(1), 89–106.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, *68*(3), 255–278.

Behaghel, O. (1930). Von deutscher wortstellung. *Zeitschrift für Deutschkunde*, *44*(1930), 81–89.

Bell, A., Brenier, J. M., Gregory, M., Girand, C., & Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*, *60*(1), 92–111.

Bock, K. (1987). An effect of the accessibility of word forms on sentence structures. *Journal of Memory and Language*, *26*(2), 119–137.

Cheng, L. L.-S., Sybesma, R., et al. (1998). Yi-wan tang, yi-ge tang: Classifiers and massifiers. *Tsing Hua journal of Chinese studies*, *28*(3), 385–412.

de Leeuw, J. R. (2015). jspsych: A javascript library for creating behavioral experiments in a web browser. *Behavior research methods*, *47*(1), 1–12.

Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, *109*(2), 193–210.

Ferreira, V. S., & Dell, G. S. (2000). Effect of ambiguity and lexical availability on syntactic and lexical production. *Cognitive Psychology*, *40*(4), 296–340.

Genzel, D., & Charniak, E. (2002). Entropy rate constancy in text. In (pp. 199–206).

Gundel, J. K. (1988). Universals of topic-comment structure. *Studies in syntactic typology*, *17*(1), 209–239.

Hadfield, J. D. (2010). Mcmc methods for multi-response generalized linear mixed models: the MCMCglmm R package. *Journal of Statistical Software*, *33*(2), 1–22.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the north american chapter of the association for computational linguistics on language technologies* (pp. 1–8). Retrieved from `http://aclweb.org/anthology/N/N01/N01-1021.pdf`

Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, *61*(1), 23–62.

Jurafsky, D., Bell, A., Gregory, M., & Raymond, W. D. (2001). Probabilistic relations between words: Evidence from reduction in lexical production. *Typological Studies in Language*, *45*, 229–254.

Klein, N. M., Carlson, G. N., Li, R., Jaeger, T. F., & Tanenhaus, M. K. (2012). Classifying and massifying incrementally in chinese language comprehension. *Count and mass across languages*, 261–282.

Kovalenko, L. Y., Chaumon, M., & Busch, N. A. (2012). A pool of pairs of related objects (poporo) for investigating visual semantic integration: behavioral and electrophysiological validation. *Brain Topography*, *25*(3), 272–284.

Levelt, W., & Maasen, B. (1981). Lexical search and order of mention in sentence production. In W. Klein & W. J. M. Levelt (Eds.), *Crossing the boundaries in linguistics* (pp. 221–252). Springer.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.

Levy, R. (2018). Communicative efficiency, uniform information density, and the rational speech act theory. In *Proceedings of the 40th annual meeting of the cognitive science society* (p. 684–689).

Levy, R., & Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. In *Advances in neural information processing systems* (pp. 849–856).

Li, P., Barner, D., & Huang, B. H. (2008). Classifiers as count syntax: Individuation and measurement in the acquisition of mandarin chinese. *Language Learning and Development*, *4*(4), 249–290.

Mahowald, K., Fedorenko, E., Piantadosi, S. T., & Gibson, E. (2013). Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition*, *126*(2), 313–318.

Maurits, L., Navarro, D., & Perfors, A. (2010). Why are some word orders more common than others? a uniform information density account. In *Advances in neural information processing systems* (pp. 1585–1593).

Oldfield, R. C., & Wingfield, A. (1965). Response latencies in naming objects. *Quarterly Journal of Experimental Psychology*, *17*(4), 273–281.

Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, *108*(9), 3526–3529.

Prince, E. F. (1981). Towards a taxonomy of given-new information. In P. Cole (Ed.), *Radical pragmatics* (pp. 223–256). New York: Academic Press.

Seyfarth, S. (2014). Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation. *Cognition*, *133*(1), 140–155.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302–319.

Sogou. (2006). *Sogou lab data: Internet lexicon 2006 version.* `https://www.sogou.com/labs/resource/w.php`. (Accessed: 2018-11-30)

Tai, J. H. (1994). Chinese classifier systems and human categorization. *In honor of William S.-Y. Wang: Interdisciplinary studies on language and language change*, 479–494.

Zhan, M., & Levy, R. (2018). Comparing theories of speaker choice using a model of classifier production in Mandarin Chinese. In *Proceedings of the 17th annual conference of the north american chapter of the association for computational linguistics: Human Language Technologies* (p. 1997–2005).

Why do people reject mixed gambles?

Joyce Wenjia Zhao (zhaowenj@sas.upenn.edu)
Department of Psychology, University of Pennsylvania
Philadelphia, PA

Lukasz Walasek (L.Walasek@warwick.ac.uk)
Department of Psychology, University of Warwick
Coventry, UK

Sudeep Bhatia (bhatiasu@sas.upenn.edu)
Department of Psychology, University of Pennsylvania
Philadelphia, PA

Abstract

Decision makers often reject mixed gambles offering equal probabilities of a larger gain and a smaller loss. This important behavioral pattern is generally seen as evidence for loss aversion, a psychological mechanism according to which losses are given higher utility weights than gains. In this paper we consider an alternate mechanism capable of generating high rejection rates: A predecisional bias towards rejection without the calculation of utility. We use a drift diffusion model of decision making to simultaneously specify and test for the effects of these two psychological mechanisms in a gambling task. Our results indicate that high rejection rates for mixed gambles result from multiple different psychological mechanisms, and that a predecisional bias applied prior to the computation of utility (rather than loss aversion) is the primary determinant of this important behavioral tendency.

Keywords: drift diffusion model; risky choice; predecisional bias; loss aversion

Introduction

Consider a gamble that offers you a gain of $11 if a coin toss lands heads, and a loss of $10 if it lands tails. Would you accept or reject this gamble? Most people choose to reject similar positive expected value mixed gambles (gambles that offer both a possibility of a gain and a possibility of a loss; Kahneman & Tversky, 1979; Samuelson, 1960), suggesting an aversion to risk. Yet risk aversion for such small monetary payoffs cannot be easily explained by conventional applications of expected utility theory. Such models predict that anyone who rejects a 50-50 gamble between a gain of $11 and a loss of $10, displays such a strong degree of risk aversion, so as to also reject a 50-50 gamble involving a loss of $100 (regardless of the magnitude of the corresponding gain; Rabin, 2000).

This (clearly unreasonable) prediction presents compelling evidence against expected utility theory, and indicates that additional psychological mechanisms need to be incorporated into models of risky choice in order to account for high rejection rates in mixed gambles (Rabin, 2000). The psychological mechanism that is widely considered to be responsible for these high rejection rates is loss aversion, which states that losses have a greater impact on utility than gains (Kahneman & Tversky, 1979; Kőszegi & Rabin, 2007; Rabin & Thaler, 2001). For example, in the mixed gamble presented at the start of this paper, loss aversion predicts that individuals experience more negative utility from the $10 loss than positive utility from the $11 gain. Thus the gamble, despite having a positive expected value, appears unattractive, and is rejected.

If loss aversion is the only mechanism responsible for the rejection of mixed gambles, an individual's degree of loss aversion can be estimated by observing how likely he or she is to accept or reject such gambles. This measure can then be used to relate loss aversion to various psychological, clinical, and neurobiological variables. Following this logic, researchers have argued that loss aversion plays an important role in irrational financial decision making, problem gambling, suicidal decision making, and incorrect affective forecasting (Hadlaczky et al., 2018; Kermer, Driver-Linn, Wilson, & Gilbert, 2006; Lorains et al., 2014; Takeuchi et al., 2015); in explaining differences in risky decision making between decision contexts (Polman, 2012; Schulreich, Gerhardt, & Heekeren, 2016; Vermeer, Boksem, & Sanfey, 2014) and between individuals with varying psychological traits, demographic profiles, and life experiences (Barkley-Levenson & Galvan, 2014; Bibby & Ferguson, 2011; Pighin, Bonini, Savadori, Hadjichristidis, & Schena, 2014; Sokol-Hessner, Hartley, Hamilton, & Phelps, 2015a); and in determining physiological and neural responses to risky prospects (Canessa et al., 2017; De Martino, Camerer, & Adolphs, 2010; Gelskov, Henningsson, Madsen, Siebner, & Ramsøy, 2015; Lazzaro, Rutledge, Burghart, & Glimcher, 2016; Markett, Heeren, Montag, Weber, & Reuter, 2016; Sokol-Hessner, Lackovic, Tobe, Camerer, Leventhal, et al., 2015b; Tom, Fox, Trepel, & Poldrack, 2007). An influential example of this approach is presented in Tom et al. (2007): In this paper, neural activity is correlated with loss aversion, measured using gamble rejection rates, and is used to identify brain regions that encode loss aversion in risky choices involving mixed gambles.

However, loss aversion may not be the only mechanism responsible for the rejection of mixed gambles. Another possibility, one which we explore in the present paper, is that

individuals exhibit a predecisional bias towards rejecting such gambles. Psychologically, this form of behavior may reflect a general preference for the status quo, whereby a decision to accept a lottery is regarded as a departure from one's status quo (Gal, 2006; W. Samuelson & Zeckhauser, 1988). We refer to this tendency as a *predecisional* bias to capture the intuition that individuals may be predisposed towards maintaining the status quo in mixed gamble tasks even *before* they have inspected and learnt about the monetary amounts that could be gained or lost. Although such a tendency could be overridden after monetary amounts are evaluated, we would nonetheless expect the predecisional bias to influence people's decisions and, in many settings, lead to a higher probability of rejection than acceptance. An important prediction of this account is that the effect of such a bias would be greatest early on in the decision, and would diminish as the decision maker deliberates about the money that could be gained or lost.

Although the predecisional bias mechanism provides a fairly intuitive explanation for high rejection rates in mixed gambles, it hasn't yet been formally compared against loss aversion, which remains the dominant explanation for this important behavioral phenomenon. The reason for this is that predecisional biases cannot be accommodated within the types of economic models used to specify loss aversion and predict risky choice. Typically, these models assume that choices depend entirely on *utility*, which itself is a product of the gains and losses offered by the gamble in consideration (e.g., Kahneman & Tversky, 1979). Thus there is no place for a mechanism that influences choice prior to the formation of utility.

There are, however, neurocomputational models of decision making that permit a more nuanced understanding of the deliberation process underpinning people's choices. One such model is the drift diffusion model (DDM), which assumes that individuals gradually accumulate evidence over the time course of the decision, with the decision being made when evidence reaches a threshold value (e.g., Bhatia, 2014; Dai & Busemeyer, 2014; Krajbich, Armel, & Rangel, 2010; Ratcliff, 1978). The evidence being accumulated depends on features of the choice alternatives, such as gains and losses, and subsequently on relative utilities. However, the start of this accumulation process can be biased towards a response (such as rejection), even before these utilities have been evaluated by the decision maker.

Mathematically, DDM implements a sequential probability ratio test, and with this interpretation, its predecisional bias can be seen as a biased prior. The DDM has also been shown to capture aspects of neural information processing, for which a predecisional bias corresponds to a bias in baseline firing rates (Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006; Gold & Shadlen, 2007). In either case, a predecisional bias in the DDM generates unique patterns in response times, and can be quantitatively estimated and differentiated from other DDM parameters (including those that govern the use of decision features like gains and losses) with a combination of choice and response time data (White & Poldrack, 2014). In prior work, psychologists and neuroscientists have used these estimates to compare predecisional biases against alternate decision mechanisms in a variety of perceptual, lexical, and motor choice tasks (Leite & Ratcliff, 2011; Mulder, Wagenmakers, Ratcliff, Boekel, & Forstmann, 2012; A. Voss, Rothermund, & Voss, 2004; White & Poldrack, 2014). The goal of this paper is to use a similar methodology to establish the extent to which a predecisional bias can account for choices in the popular mixed gamble task.

As an example of this task, consider the decision to accept or reject a gamble i, offering a 50% chance of gaining G_i and a 50% chance of losing L_i. The utility for accepting the gamble in the presence of loss aversion is given by $U_i = G_i - \lambda \cdot L_i$ (as the probabilities of the gains and losses are identical, they can be ignored without any effect on model predictions). Here λ is the loss aversion parameter, where $\lambda > 1$ indicates the larger impact of loss than gains. Assuming that the utility for rejecting the gamble is 0, the decision maker will accept gamble i when $U_i > 0$, and reject the gamble when $U_i < 0$. Stochasticity in choice can be modelled with a logistic response function. With such specification, the magnitude of λ (the loss aversion parameter) can be estimated using a logistic regression: $A_i \sim \beta_G \cdot G_i - \beta_L \cdot L_i$. Here A_i is the participant's binary response to the i^{th} gamble (1 if Accept, 0 if Reject), and β_G and β_L are regression coefficients that yield $\lambda = \beta_L/\beta_G$. In practice, researchers often include an additive intercept (α) in the logistic regression: $A_i \sim \alpha + \beta_G \cdot G_i - \beta_L \cdot L_i$. Here the additive intercept corresponding to a fixed impact on utility favoring acceptance or rejection.

Although commonly used to make inferences regarding the psychological and neural underpinnings of risky choice (Tom et al., 2007), the logistic model outlined above neglects the possibility that decision makers may be predisposed towards one of the choice options (acceptance or rejection) prior to evaluating the underlying utilities. To permit this possibility, we model the decision using a drift diffusion process, which is illustrated in Figure 1A. This model assumes that decision makers accumulate evidence in favor of accepting vs. rejecting the gamble over time, with a *drift rate* that relates the utility of the gamble to the accumulation process. To keep model specifications consistent with the static logistic model outlined above, we write the drift rate for a trial involving gamble i, as $v_i = \alpha + \beta_G \cdot G_i - \beta_L \cdot L_i$. Choices are made when the accumulated evidence reaches a positive threshold $+\theta$ (corresponding to acceptance) or a negative threshold $-\theta$ (corresponding to rejection). The magnitude of θ quantifies the amount of evidence required for reaching a decision. Mechanistically, this threshold captures the speed-accuracy tradeoff in decision making, with higher value of θ generating slower but more accurate choices.

In the DDM, the predecisional bias takes the form of a starting point $\gamma > 0$, that is closer to $+\theta$ (predisposing the decision maker towards accepting the gamble), or $\gamma < 0$, that is closer to $-\theta$ (predisposing the decision maker

towards rejecting the gamble). When $\gamma = 0$, the preference accumulation process starts from a neutral state, and the choice probabilities generated by the DDM are identical to those predicted by the static logistic model introduced above. Allowing for the gradual accumulation of evidence prior to the decision enables the DDM to predict response times (RTs). The response time in a trial is assumed to be the time taken for the accumulating evidence to reach a decision threshold added to a fixed non-decisional time τ (which captures the time taken to perceive the stimuli, execute motor responses after the decision has been made, and so on).

The response times predicted by the DDM depend critically on the gamble that is offered on a given trial. Responses times on trials with extremely desirable or undesirable gambles (which generate large positive or negative drift rates) will be shorter, capturing the fact that easier decisions are made relatively quickly compared to more difficult decisions. Besides the influence of the specific gamble at hand, response times also depend on the predecisional bias. If there is a predecisional bias in favor of rejection ($\gamma < 0$), response times associated with rejection will tend to be shorter than those associated with acceptance, and correspondingly, the rejection rates in quicker choices will be higher than those in slower choices, controlling for the difficulty of the choice in consideration (see Figure 1A). Intuitively, the effects of the drift rate (i.e. the utilities used in evaluation) persist throughout the preference accumulation process; whereas the impact of a non-neutral starting point (predecisional bias) gets gradually washed out over time. Crucially, such a prediction cannot be made by the DDM in the absence of the predecisional bias (i.e. when $\gamma = 0$, and DDM choice probabilities mimic the standard logistic specification), indicating that the choice-RTs patterns can be used as a behavioral marker to infer the existence of a predecisional bias (White & Poldrack, 2014).

Methods

Our main experimental task incentivized accept-reject decisions for mixed gambles with a 50% chance of a gain and a 50% chance of a loss. We preregistered our study at OSF(https://osf.io/varx6/?view_only=b9b9f84bd9fc4a56b8df19ea02998fec). In addition to our preregistered study, we also conducted three additional non-incentivized studies (Experiments 1A-1C), which we do not report in the paper due to space limit. The main conclusions of Experiment 2 were replicated in those studies.

Experimental design

Participants. 49 participants were recruited from a paid participant pool at the University of Pennsylvania.

Procedures. Participants were instructed to accept or reject a sequence of 200 gambles, presented in four blocks of 50 gambles. Each gamble had two possible outcomes: A gain of some amount of tokens occurring with a 50% chance and a loss of some amount of tokens occurring with a 50% chance. The outcomes were displayed side by side, with positive/negative values indicating gains and losses (see Figure 1B). Participants pressed up or down arrow keys on a keyboard to indicate acceptance or rejection, with the specific key-response associations alternating across blocks to control for response biases favoring one of the keys. Choices and reaction times were recorded.

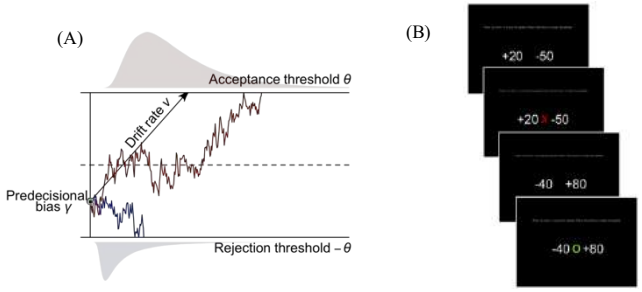


Figure 1 A: The drift diffusion model. B: Task presentation.

Each token was worth US$0.10, and participants began the experiment with an endowment of 100 tokens (US$10). Participants were informed that their choices in the experiment would determine their bonus payment, which they would receive on top of a fixed show-up fee of US$8. This was accomplished by selecting one of the gambles at random. If the participant rejected the gamble, the bonus payment would be 100 tokens (US$10). If the participant accepted the gamble, then they would flip a coin in front of the experimenter to play out the gamble. Their received token amount would be their initial endowment (100 tokens = US$10) plus or minus the gain or loss associated with the coin flip. Average total payments in the experiment were US$ 10.43 per participant.

Stimuli. The possible gain and loss values were taken from the set of {10, 20, 30, 40, 50, 60, 70, 80, 90, 100} tokens, or equivalently US$ {1, 2, 3, 4, 5, 6, 7, 8, 9, 10}. With this stimuli set we were able to generate a total of 100 unique gambles. We counterbalanced the positions of the gain/loss outcomes across blocks, resulting in 200 total trials.

Model Fitting

The models were fit to choice and RT data using HDDM (Wiecki, Sofer, & Frank, 2013), a Python package for hierarchical Bayesian estimation of drift-diffusion models, using its default priors. To fit the models, 4 chains of 50,000 samples were generated, where the first 25,000 were burn-ins, and a thinning of 2 was applied.

Results

Overall, the average rejection probability across participants was 71.5%, with 79.6% of participants being more likely to reject than accept the gambles. These probabilities are significantly different to 50% which is the rate we would expect if choices were made by chance or if individuals did not display loss aversion or predecisional biases ($p < 0.001$ when compared to 50% using t-tests). On average, participants accepted the gambles only when the size of the gain exceed 1.75 times the size of the loss. This pattern of behavior can be explained by both the loss aversion and the

predecisional bias mechanisms. According to a model with loss aversion but no predecisional bias, the probability of acceptance is greater than the probability of rejection only when the utility for the gamble exceeds 0, which happens only when the size of the gain exceeds the size of the loss by a large enough margin to counteract loss aversion. According to a model with a predecisional bias but no loss aversion, the probability of acceptance is greater than the probability of rejection only when the utility of the gamble is large enough to override the starting point bias favoring rejection. This happens only when the size of the gain exceeds the size of the loss by a large enough margin, giving a sufficiently positive utility.

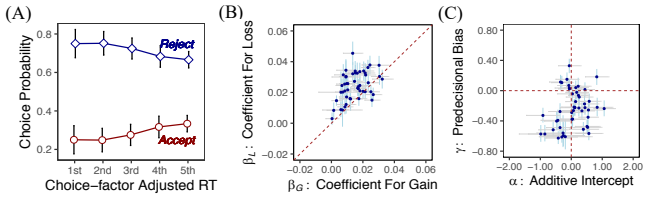


Figure 2. A: Choice-RT relationships. Error bars indicated 95% CI. B: Loss aversion in the DDM. C: Predecisional bias and additive intercept in the drift rate. Most participants have negative posterior means for predecisional bias (i.e., bias towards rejecting gambles). In panel B and C each dot represents a participant and the error bars indicate 95% posterior credible intervals for the parameters in the two figures.

We also found that rejections were quicker than acceptances. Overall, the average rejection decision took 1.30 seconds, whereas the average acceptance decision took 1.72 seconds (the difference is significant: $t(46) = 4.04, p < 0.001$). Additionally, 74.5% of participants took less time to reject than to accept. The RT distributions for acceptance and rejections are different from each other (Wilcoxon signed rank test: $V = 935, p < 0.001$).

Although the observed response time pattern appears consistent with those generated by a predecisional bias favoring rejection, they do not control for choice factors (gains and losses) of the gamble, and thus can also be generated by a DDM model without this bias. More specifically, it is possible that trials on which gambles are rejected involve highly undesirable gambles (and therefore quicker response times), whereas trials on which gambles are accepted involve only moderately desirable gambles (and thus slower response times). To address this issue, Figure 2A shows these choice-RTs patterns, with RTs adjusted for choice factors. These adjusted RTs are residuals from participant-level regressions, in which log RTs are regressed on gain values and loss values of the mixed gambles for each participant. With choice factors controlled for, we observe a negative relationship between choice probability and response time for rejection decisions, and a positive relationship between choice probability and response time for acceptance decisions, showing that decision makers are quicker to reject and slower to accept. This is a novel behavioral pattern that suggests that our participants displayed a predecisional bias favoring rejection. Importantly, this pattern cannot be generated by a DDM model with only loss aversion and no predecisional bias (or by the standard logistic specification of the loss aversion mechanism).

A more rigorous comparison of the loss aversion and predecisional bias mechanisms requires quantitative model fitting. We did so using hierarchical Bayesian techniques applied to choice and RT data. This approach allows for three flexible parameters for the drift rate (α, β_L and β_G) as well as a flexible starting point bias (γ), threshold (θ) and non-response time (τ). Thus this model can simultaneously display both loss aversion and a predecisional bias. We also allowed the threshold (θ) to be dependent on the monetary loss, in order to capture the effect of losses on attention (as specified in our preregistration plan; Yechiam & Hochman, 2013).

Overall, we observe best-fit parameter values such that $\beta_L > \beta_G$ for 85.7% participants, with 57.1% of participants having a 95% credible interval for $\beta_L - \beta_G$ that is strictly positive. The posterior mean of $\lambda = \frac{\beta_L}{\beta_G}$ averaged across our participants is 2.11 ($SD = 1.35$). We also observe a negative posterior mean of γ for 77.6% participants (significant for 69.4% of participants as indicated by 95% credible intervals). The averaged participant-level posterior mean of γ is -0.24 ($SD = 0.25$) across all participants. Finally, we observe a negative posterior mean of α for only 40.8% participants (significant for 12.2% of participants as indicated by 95% credible intervals), with a mean value of $\alpha = 0.05$ ($SD = 0.45$) across our participants. This analysis indicates that most participants display loss aversion and predecisional biases favoring rejection, but do not display any systematic additive intercepts in the drift rate. The posterior means for participant-level parameters are shown in Figures 2B and 2C.

To better understand the descriptive power of the predecisional bias, and to compare it against the descriptive power of loss aversion, we also fit three restricted variants of the DDM. The first constrained model set $\beta_L = \beta_G$ (eliminating loss aversion while permitting flexible values of γ, as well as other DDM parameters). The second set $\gamma = 0$ (eliminating the predecisional bias while permitting flexible values of β_L and β_G, as well as other DDM parameters). The third constrained model is a baseline model that set both $\beta_L = \beta_G$ and $\gamma = 0$ (but permitted flexible values for the remaining DDM parameters). We compared the relative fits of these three constrained models against each other, and against the full model. The model comparisons were performed using the deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & van der Linde, 2002), which measures model fits while penalizing model complexity to avoid over-fitting. Smaller DICs indicate better model performance. This measure revealed that despite having more parameters than the remaining models, the full model ($DIC = 16{,}871$) generated the best fit to the observed data (indicated by DIC differences between this

model and the remaining models, which we denote as ΔDIC). Conversely, despite having fewer parameter than the other models, the baseline model generated the worst fit to the observed data ($DIC = 18{,}456$, $\Delta\text{DIC} = 1{,}586$). This indicates that loss aversion and predecisional biases are useful for describing behavior in our experiment. However, out of the two constrained models, the one that set $\beta_L = \beta_G$ ($DIC = 17{,}332, \Delta\text{DIC} = 461$) yielded much better fits than the one that set $\gamma = 0$ ($DIC = 17{,}979, \Delta\text{DIC} = 1{,}108$), indicating that the predecisional bias plays a more important role than loss aversion.

Although our quantitative fits do provide strong evidence in favor of the predecisional bias mechanism, using such fits as a single piece of evidence for theory testing is problematic (Roberts & Pashler, 2000). Ideally, we should also compare our models in terms of their ability to account for a qualitative behavioral marker, in this case, the finding that rejection rates are higher for trials with shorter RTs compared to trials with longer RTs (Figure 2A, and solid blue lines in Figures 3A-D). As discussed above, this pattern is consistent with the effect of a predecisional bias towards rejecting mixed gambles. A model without such a bias cannot account for RT differences between acceptance and rejection, controlling for choice factors. To establish this more rigorously, we used simulated data from the best-fitting full and constrained models. In line with our intuition, we found that the choice-RT relationship can be captured by the best-fit full model (Figure 3A), as well as by the best-fit constrained model with flexible predecisional bias but no loss aversion (Figure 3B). However, both the best-fit model with loss aversion but no predecisional bias (Figure 3C) and the best-fit baseline model (Figure 3D) fail to capture this relationship. This finding provides one explanation for why the predecisional bias plays a more important role than loss aversion in our quantitative model fits.

In our final analysis we tested the relationship between individual-level model parameters and observed heterogeneity in participant behavior. For this purpose, we correlated best-fitting participant-level estimates of loss aversion ($\lambda = \beta_L/\beta_G$) and predecisional bias (γ) with average participant-level rejection rates. The Pearson correlation between acceptance rates and the predecisional bias is 0.91 ($t(47) = 14.79, p < 0.001$; $Spearman\ Corr = 0.92, p < 0.001$); whereas the correlation between acceptance rates and loss aversion is -0.25 ($t(47) = 1.78, p = 0.08$; $Spearman\ Corr = -0.43, p = 0.002$). These correlations are displayed in Figures 3E and 3F. From the perspective of describing participant heterogeneity, the predecisional bias is clearly the more important psychological mechanism.

Did the participants develop the predecisional bias over the course of the experiment, or did they already have a predecisional bias for gamble choices based on previous life experiences? To test this, we examined the choice-RT relationship (the behavioral marker for predecisional biases) in the first 25 trials of the experiment (first half of the first block). As Figure 3G shows participants were quicker to reject gambles than accept gambles when choice factors are controlled for. In other words, participant already had a predecisional tendency to reject gambles, even when they had limited knowledge regarding the gain and loss value distributions involved in the experiment.

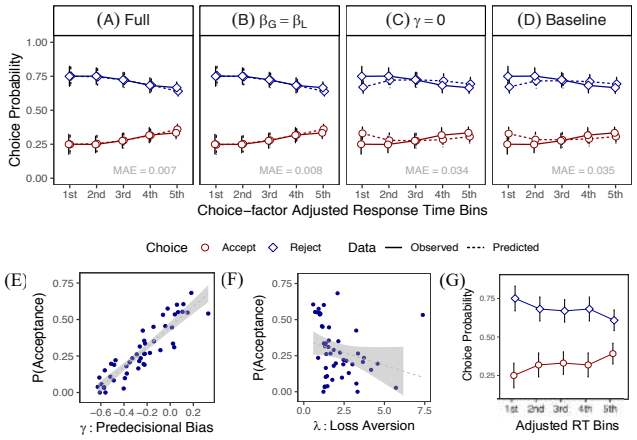


Figure 3. A-D: Choice-RT relationships for observed data (solid lines) and model simulated data (dashed lines). Rejection rates are higher in quicker trials compared to slower trials, controlling for choice factors (gain and loss values). This pattern can only be generated by models that permit a predecisional bias (panels A and B). MAE: Mean absolute error. E-F: Relationships between the DDM mechanisms and acceptance rates. Each dot represents a participant. The predecisional bias is more strongly correlated with the observed choice outcomes, compared to loss aversion. G: Choice-RT relationship for observed data in the first 25 trials of the experiment.

Discussion

The results presented above have a number of important implications for the study of risk preference. First, these results shed light on the psychological underpinnings of one of the most important behavioral findings pertaining to risk: The rejection of small scale 50-50 mixed gambles with positive expected values (Kahneman & Tversky, 1979; Samuelson, 1960). They show that this phenomenon is not just a product of loss aversion (i.e., higher weights attached to losses relative to gains), but is also due to a predecisional bias favoring the status quo. This bias generates a tendency to reject the gamble even before the gamble's payoffs are evaluated, and the effect of this bias is the strongest early on in the decision process. For this reason, the predecisional bias makes unique predictions regarding the relationship between response time and rejection probability. Our experiments provide novel evidence in support of these predictions, indicating that a model equipped with a predecisional bias is necessary to account for behavioral patterns in mixed gamble tasks.

We also used model fitting to evaluate the relative contributions of the loss aversion and predecisional bias mechanisms. Although both loss aversion and predecisional bias play a valuable quantitative role, a model with the predecisional bias but without loss aversion fits better than a model with loss aversion but without predecisional bias. A

second test evaluating the predictive power of best fit model parameters shows that individual-level predecisional bias parameters correlate more strongly with individual-level rejection rates than do individual-level loss aversion parameters. These findings provide strong quantitative evidence that predecisional biases are the primary determinant of high rejection rates in mixed gamble tasks. In doing so they complement recent experimental results showing that loss aversion is not as good of a descriptor of choice behavior as has been previously assumed (Bhatia, 2017; Birnbaum, 2008; Erev, Ert, & Yechiam, 2008; Ert & Erev, 2013; Walasek & Stewart, 2015).

Our findings have important implications for how we interpret people's tendency to reject mixed gambles. A lot of prior work in psychology, economics, and neuroscience infers loss aversion through mixed gamble rejection rates, and subsequently uses this measure of loss aversion to explain the effect of social, cognitive, emotional, developmental, demographic, clinical, physiological, and neural variables on risky choice (e.g., Bibby & Ferguson, 2011; Canessa et al., 2017; Engelmann et al., 2015; Gelskov et al., 2015; Hadlaczky et al., 2018; Kermer et al., 2006; Lazzaro et al., 2016; Lorains et al., 2014; Markett et al., 2016; Pighin et al., 2014; Polman, 2012; Tom et al., 2007; Vermeer et al., 2014). Yet our results indicate that these explanations may be incorrect, and that these variables may be better understood in terms of predecisional bias tendencies. Thus, for example, the well-known finding that ventral striatum activity correlates with mixed gamble rejection rates (Tom et al., 2007) could be due to the relationship between brain activity and predecisional bias rather than the relationship between brain activity and loss aversion, as is commonly assumed. Additional research is needed to untangle these relationships, and future work should consider the possibility that gamble rejection rates, as well as the psychological and neurobiological correlates of high rejection rates, can be understood in terms of multiple different psychological mechanisms.

The tests presented in this paper rely critically on response time data: without this type of data, it would be impossible to identify and measure the predecisional bias. Our analysis uses the drift diffusion model to account for trends in response time data, and by doing so, illustrates the descriptive power of this popular neurocomputational theory (Ratcliff, 1978). The DDM has been previously used to model perceptual, lexical, motor phenomena, and the predecisional bias has been shown to be an important parameter in these low-level tasks (e.g., Forstmann, Ratcliff, & Wagenmakers, 2016; Mulder et al., 2012; Ratcliff et al., 2004; Ratcliff, Smith, Brown, & McKoon, 2016; White & Poldrack, 2014). Additionally, this bias has a theoretically compelling interpretation in terms of baseline firing rates in neural models and statistical priors in optimal sequential evaluation tasks (Bogacz et al., 2006; Gold & Shadlen, 2007). Recent work applying DDM and related models to preferential choice data has also shown that these models provide a powerful account of a variety of choice anomalies. We recommend that future research utilizes the DDM, alongside response time data, to obtain a more comprehensive understanding of the psychological and neurobiological determinants of risky choice.

Reference

Barkley-Levenson, E., & Galvan, A. (2014). Neural representation of expected value in the adolescent brain. *Proceedings of the National Academy of Sciences*, *111*(4), 1646–1651.

Bhatia, S. (2014). Sequential sampling and paradoxes of risky choice. *Psychonomic Bulletin & Review*, *21*(5), 1095–1111.

Bhatia, S. (2017). Comparing theories of reference-dependent choice. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(9), 1490–1507.

Bibby, P. A., & Ferguson, E. (2011). The ability to process emotional information predicts loss aversion. *Personality and Individual Differences*, *51*(3), 263–266.

Birnbaum, M. H. (2008). New paradoxes of risky decision making. *Psychological Review*, *115*(2), 463–501.

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review*, *113*(4), 700–765.

Canessa, N., Crespi, C., Baud-Bovy, G., Dodich, A., Falini, A., Antonellis, G., & Cappa, S. F. (2017). Neural markers of loss aversion in resting-state brain activity. *NeuroImage*, *146*(C), 257–265.

Dai, J., & Busemeyer, J. R. (2014). A probabilistic, dynamic, and attribute-wise model of intertemporal choice. *Journal of Experimental Psychology: General*, *143*(4), 1489–1514.

De Martino, B., Camerer, C. F., & Adolphs, R. (2010). Amygdala damage eliminates monetary loss aversion. *Proceedings of the National Academy of Sciences*, *107*(8), 3788–3792.

Erev, I., Ert, E., & Yechiam, E. (2008). Loss aversion, diminishing sensitivity, and the effect of experience on repeated decisions. *Journal of Behavioral Decision Making*, *21*(5), 575–597.

Ert, E., & Erev, I. (2013). On the Descriptive Value of Loss Aversion in Decisions under Risk. *Judgment and Decision Making*, *8*(3), 214–235.

Forstmann, B. U., Ratcliff, R., & Wagenmakers, E. J. (2016). Sequential Sampling Models in Cognitive Neuroscience: Advantages, Applications, and Extensions. *Annual Review of Psychology*, *67*(1), 641–666.

Gal, D. (2006). A psychological law of inertia and the illusion of loss aversion. *Judgment and Decision Making*, *1*(1), 23–32.

Gelskov, S. V., Henningsson, S., Madsen, K. H., Siebner, H. R., & Ramsøy, T. Z. (2015). Amygdala signals subjective appetitiveness and aversiveness of mixed gambles. *Cortex*, *66*(C), 81–90.

Gold, J. I., & Shadlen, M. N. (2007). The Neural Basis of Decision Making. *Annual Review of Neuroscience*, *30*(1), 535–574.

Hadlaczky, G., Hökby, S., Mkrtchian, A., Wasserman, D., Balazs, J., Machín, N., et al. (2018). Decision-Making in Suicidal Behavior: The Protective Role of Loss Aversion. *Frontiers in Psychiatry*, *9*, e1000123–9.

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, *47*(2).

Kermer, D. A., Driver-Linn, E., Wilson, T. D., & Gilbert, D. T. (2006). Loss Aversion Is an Affective Forecasting Error. *Psychological Science*, *17*(8), 649–653.

Kőszegi, B., & Rabin, M. (2007). Reference-Dependent Risk Attitudes. *American Economic Review*, *97*(4), 1047–1073.

Krajbich, I., Armel, C., & Rangel, A. (2010). Visual fixations and the computation and comparison of value in simple choice. *Nature Neuroscience*, *13*(10), 1292–1298.

Lazzaro, S. C., Rutledge, R. B., Burghart, D. R., & Glimcher, P. W. (2016). The Impact of Menstrual Cycle Phase on Economic Choice and Rationality. *PloS One*, *11*(1), e0144080–15.

Leite, F. P., & Ratcliff, R. (2011). What cognitive processes drive response biases? A diffusion model analysis. *Judgment and Decision Making*, *6*(7), 651–687.

Lorains, F. K., Dowling, N. A., Enticott, P. G., Bradshaw, J. L., Trueblood, J. S., & Stout, J. C. (2014). Strategic and non-strategic problem gamblers differ on decision-making under risk and ambiguity. *Addiction*, *109*(7), 1128–1137.

Markett, S., Heeren, G., Montag, C., Weber, B., & Reuter, M. (2016). Loss aversion is associated with bilateral insula volume. A voxel based morphometry study. *Neuroscience Letters*, *619*, 172–176.

Mulder, M. J., Wagenmakers, E.-J., Ratcliff, R., Boekel, W., & Forstmann, B. U. (2012). Bias in the brain: a diffusion model analysis of prior probability and potential payoff. *Journal of Neuroscience. 32*(7), 2335–2343.

Pighin, S., Bonini, N., Savadori, L., Hadjichristidis, C., & Schena, F. (2014). Loss aversion and hypoxia: less loss aversion in oxygen-depleted environment. *Stress*, *17*(2), 204–210.

Polman, E. (2012). Self–other decision making and loss aversion. *Organizational Behavior and Human Decision Processes*, *119*(2), 141–150.

Rabin, M. (2000). Diminishing marginal utility of wealth cannot explain risk aversion. In D. Kahneman & A. Tversky (Eds.), *Choices, Values and Frames*. New York.

Rabin, M., & Thaler, R. H. (2001). Anomalies: Risk Aversion. *Journal of Economic Perspectives*, *15*(1), 219–232.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108.

Ratcliff, R., & Childers, R. (2015). Individual Differences and Fitting Methods for the Two-Choice Diffusion Model of Decision Making. *Decision*, *2015*.

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion Decision Model: Current Issues and History. *Trends in Cognitive Sciences*, *20*(4), 260–281.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*(2), 358–367.

Samuelson, P. A. (1960). The St. Petersburg Paradox as a Divergent Double Limit. *International Economic Review*, *1*(1), 31–37.

Samuelson, W., & Zeckhauser, R. (1988). Status quo bias in decision making. *Journal of Risk and Uncertainty*, *1*(1), 7–59.

Schulreich, S., Gerhardt, H., & Heekeren, H. R. (2016). Incidental fear cues increase monetary loss aversion. *Emotion*, *16*(3), 402–412.

Sokol-Hessner, P., Hartley, C. A., Hamilton, J. R., & Phelps, E. A. (2015a). Interoceptive ability predicts aversion to losses. *Cognition & Emotion*, *29*(4), 695–701.

Sokol-Hessner, P., Lackovic, S. F., Tobe, R. H., Camerer, C. F., Leventhal, B. L., & Phelps, E. A. (2015b). Determinants of Propranolol's Selective Effect on Loss Aversion. *Psychological Science*, *26*(7), 1123–1130.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B*, *64*(4), 583–639.

Takeuchi, H., Kawada, R., Tsurumi, K., Yokoyama, N., Takemura, A., Murao, T., et al. (2015). Heterogeneity of Loss Aversion in Pathological Gambling. *Journal of Gambling Studies*, *32*(4), 1143–1154.

Tom, S. M., Fox, C. R., Trepel, C., & Poldrack, R. A. (2007). The neural basis of loss aversion in decision-making under risk. *Science*, *315*(5811), 515–518.

Vermeer, A. B. L., Boksem, M. A. S., & Sanfey, A. G. (2014). Neural mechanisms underlying context-dependent shifts in risk preferences. *NeuroImage*, *103*(C), 355–363.

Voss, A., Rothermund, K., & Voss, J. (2004). Interpreting the parameters of the diffusion model: an empirical validation. *Memory & Cognition*, *32*(7), 1206–1220.

Walasek, L., & Stewart, N. (2015). How to Make Loss Aversion Disappear and Reverse: Tests of the Decision by Sampling Origin of Loss Aversion. *Journal of Experimental Psychology: General*, *144*(1), 7–11.

White, C. N., & Poldrack, R. A. (2014). Decomposing bias in different types of simple decisions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(2), 385–398.

Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). HDDM: Hierarchical Bayesian estimation of the Drift-Diffusion Model in Python. *Frontiers in Neuroinformatics*, *7*, 14.

Yechiam, E., & Hochman, G. (2013). Loss-aversion or loss-attention: The impact of losses on cognitive performance. *Cognitive Psychology*, *66*(2), 212–231.

Towards a space of contextual effects on choice behavior: Insights from the drift diffusion model

Wenjia Joyce Zhao
University of Pennsylvania, Philadelphia, Pennsylvania, United States

Aoife Coady
University of Pennsylvania, Philadelphia, Pennsylvania, United States

Sudeep Bhatia
University of Pennsylvania, Philadelphia, Pennsylvania, United States

Abstract

Choice behavior can be influenced by many different types of incidental contextual effects, including those pertaining to presentation format, emotion, social belief, and cognitive capacity. Many of these contextual effects form the basis of nudges, used by academics and practitioners to shape choice. In this paper, we use data from a very large-scale choice experiment to uncover a space of contextual effects. We construct this space by analyzing fifteen contextual effects using the parameters of the drift diffusion model (DDM). DDM is a quantitative theory of decision making whose parameters offer a theoretically compelling characterization of the cognitive underpinnings of choice behavior. By representing a large number of contextual effects in terms of how they influence the parameters of the DDM, our space is able to precisely measure, quantify, and compare the contextual effects, and interpret these effects in terms of their behavioral, mechanistic, and statistical implications.

Does Video Content Facilitate or Impair Comprehension of Documentaries? The Effect of Cognitive Abilities and Eye Movement Strategy

Yueyuan Zheng (u3514160@connect.hku.hk)
Department of Psychology, University of Hong Kong,
Pokfulam Road, Hong Kong

Xinchen Ye (yexc@connect.hku.hk)
Department of Psychology, University of Hong Kong,
Pokfulam Road, Hong Kong

Janet H. Hsiao (jhsiao@hku.hk)
Department of Psychology, University of Hong Kong,
Pokfulam Road, Hong Kong

Abstract

It remains unclear whether multimedia facilitates or impairs knowledge acquisition. Here we examined whether subtitles and video content facilitate comprehension of documentaries consisting of statements of facts and whether the comprehension depends on participants' cognitive abilities and eye movement strategies during video watching. We found that subtitles facilitated comprehension regardless of participants' cognitive abilities or eye movement strategies for video watching. In contrast, with video content but not subtitles, comprehension depended on participants' auditory working memory, task switching ability, and eye movement strategy. Through the Eye Movement analysis with Hidden Markov Models (EMHMM) method, we found that a centralized (looking mainly at the screen center) eye movement strategy predicted better comprehension as opposed to a distributed strategy (with distributed regions of interest) after contributions from cognitive abilities were controlled. Thus, whether video content facilitates comprehension of documentaries depends on the viewers' eye movement strategy in addition to cognitive abilities.

Keywords: multimedia; eye-movement; hidden Markov model

Introduction

Multimedia learning refers to knowledge construction from both verbal and pictorial information, with the verbal form including spoken words or printed texts, and the pictorial form including illustrations, graphs, pictures, photos, animations, and videos (Mayer, 2014a). The modality principle suggests that it is generally beneficial to receive both visual and audio information in the learning process (Low & Sweller, 2014). Similarly, subtitles have been shown to enhance learning: same-language subtitles in video advertisements were shown to enhance the viewers' memory of the brand and slogan (Brasel & Gips, 2014), and watching recorded lectures with subtitles was associated with better comprehension performance (Kruger & Steyn, 2014).

The cognitive theory of multimedia learning further posits that learning would be undermined if multiple sources of information are received from the same perceptual channel, and would be facilitated if different sources of information are received from independent channels (Mayer, 2014b). For example, people displayed worse learning outcome when watching animations with on-screen texts than without because both text and animation information came from the visual modality (Mayer, Heiser & Lonn, 2001). However, some more recent research reported no trade-off between image and text processing (Perego, Del Missier, Porta & Mosconi, 2010; Kruger, Soto-Sanfiel & Doherty, 2017), and that participants learning through multimedia displayed better knowledge acquisition and improved content comprehension as compared with those who learned through text reading or traditional lectures in academic learning (Starbek, Erjavec, Starcic & Peklaj, 2010).

The inconsistent findings in the literature may be due to differences in the type of learning materials used and learners' language proficiency across studies. The effect of multimedia learning may depend on how the most important information for comprehension was delivered. For example, for documentaries containing mainly statements of facts, the auditory narratives may contain most of the information, and thus video content may be distracting and impair comprehension whereas same-language subtitles may be helpful especially for second-language learners. Indeed, unsubtitled videos were reported to create a higher cognitive load (as indicated by pupil diameter change) and frustration levels (as measured by EEG) than subtitled versions for students learning English as a second language (Kruger, Hefer & Matthew, 2013). In contrast, for learning involving graphic demonstrations, video content may contain additional information, which may compete with on-screen texts for cognitive resources.

In addition, individual differences in cognitive abilities and strategies may also influence whether multimedia helps or impairs learning. For example, the ability to flexibly

switch attention among various sources of information and to focus on the relevant information while inhibiting irrelevant information may be important for successful multimedia learning (Miyake et al., 2000). Indeed, research has reported that readers with high working memory capacity were more effective in selecting and integrating information and achieved better comprehension in multimedia learning (Schnotz, 2005; Fenesi, Kramer & Kim, 2016). Similar results were found for those with better task switching ability (Baadte, Rasch & Honstein, 2015).

Another possible factor is individual differences in cognitive strategy or perception style, which may be better revealed through eye tracking (Hyona, 2010; van Gog & Scheiter, 2010). Previous research using eye tracking to understand multimedia processing typically only focused on group level comparisons, such as comparing adults' and children's learning (D'Ydewalle & De Bruycker, 2007). However, recent studies have reported significant individual differences in eye movement patterns that can reflect differences in cognitive strategy and perception style (e.g., Chan, Chan, Lee & Hsiao, 2018). It is possible that participants adopting different eye movement strategies during multimedia learning differed in whether multimedia facilitates or impairs learning.

Here we aimed to examine how individual differences in cognitive abilities and eye movement strategies modulate multimedia learning of documentaries consisting of mainly statements of facts, and thus the important information would be mainly in the auditory narratives. Specifically, we examined how adding subtitles and video content would influence participants' comprehension of auditory science documentaries, and whether participants' working memory capacity, switching ability, and eye movement strategy for video watching could predict the comprehension of the documentaries. We recruited native speakers of the language used in the documentaries to control for language experience. To discover common eye movement strategies for video watching from the participants and quantitively measure individual differences in eye movement pattern, we used the Eye Movement analysis with Hidden Markov Models (EMHMM, Chuk, Chan, & Hsiao, 2014) method to analyze eye movement data. In this method, each participant's eye movement pattern during a visual task is summarized using an HMM, including personalized regions of interest (ROIs) and transition probabilities among the ROIs. Individual HMMs then can be clustered according to similarities to discover common strategies in the participants (Coviello, Chan & Lanckriet, 2014). The similarity between a participant's eye movement pattern to a common strategy can be assessed as the log-likelihood of the participant's eye movement data being generated by the HMM of the common strategy. This quantitative measure of eye movement pattern similarity then can be used to examine the associations between eye movement patterns and other cognitive measures. We hypothesized that (1) subtitles are helpful in the comprehension of the documentaries because of the exact match to the content of the documentaries, which may facilitate retrieving the meanings of the auditory narratives, (2) people with higher working memory capacity or task switching ability may show better comprehension when learning from documentaries with video content, and (3) people with more explorative eye movement patterns during video watching may be more distracted by video content, leading to worse comprehension performance.

Methods

Participants

Sixty native Mandarin speakers (40 females, 18 to 30 years old, M = 21.07, SD = 3.32) were recruited from the University of Hong Kong. Participants were from different majors except for ecology, astronomy, geography and chemistry, which were the topics of the documentary clips used here. All participants reported normal or corrected-to-normal vision.

Materials

The materials consisted of 16 documentary video clips in ecology, astronomy, geography, and chemistry, with 4 clips in each topic. The length of each clip was 75 s, and the resolution was 1920 x 1280 pixels. All clips used were produced by China Central Television (CCTV) and Shanghai Education Television (SETV) and were accessible to the public. The clips were selected based on the following criteria: a) Mandarin narratives; b) with simplified Chinese subtitles; c) not translated from foreign languages; d) produced as statements of facts, as to ensure the understandability of the clips to native speakers and to avoid possible linguistic biases.

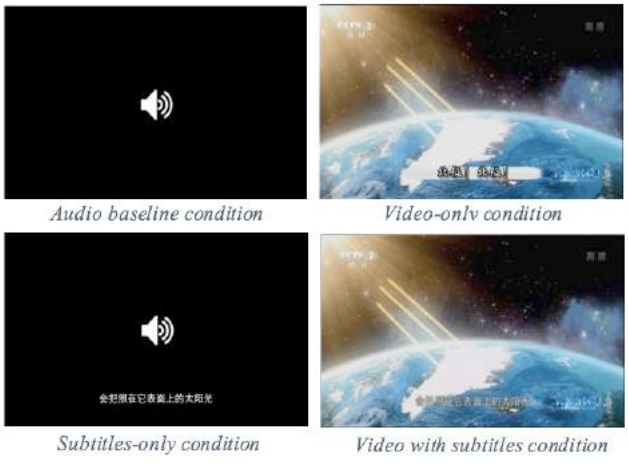


Figure 1: Four conditions of documentary presentation.

Each documentary could be presented in 4 different conditions: a) *Audio baseline* condition (i.e., narrated sound only without video content or subtitle, with a static icon presented at the center of a black screen); b) *Subtitles-only* condition (i.e., Audio baseline condition with original subtitles); c) *Video-only* condition (i.e., narrated sound with full-screen video content, with a fixed title masking the original subtitles); and d) *Video with subtitles* condition (i.e., narrated sound with both video content and original subtitles; Figure 1). Among the 16 original video clips, half of them had subtitles located on the bottom left of the screen, while the other half set the subtitles on the bottom center.

Design

Here we examined how video content and subtitle affected comprehension of documentaries. The independent variables were video content (with vs. without) and subtitle (with or without), resulting in four experimental conditions: audio baseline, subtitles only, video only, and video with subtitles. Each participant viewed 16 clips in total with 4 clips in each condition (one from each topic); the clips used in the 4 conditions were counterbalanced across participants. The dependent variable was accuracy in answering comprehension questions related to the clips. Repeated measures ANOVA was used for the data analysis.

In a separate analysis, we examined whether eye movement strategies used in the video only and video with subtitles conditions could predict comprehension of documentaries. The EMHMM approach was used (Chuk et al., 2014). More specifically, for each of the two video conditions separately, we used one HMM to summarize a participant's eye movement pattern across all clips. We then clustered the individual HMMs into two groups according to similarities to discover two representative eye movement strategies among the participants. Participants' eye movement pattern similarity to the two strategies then could be quantitatively assessed by calculating the log-likelihood of their data being generated by the HMM of the representative strategies. We examined whether this eye movement pattern similarity measure could predict participants' comprehension.

We also examined whether participants' cognitive abilities could predict comprehension of documentaries. Participants completed a n-back task for testing working memory capacity (Owen, McMillan, Laird & Bullmore, 2005), a Tower of London task for assessing executive function and planning abilities (Shallice, 1982), and a multitasking task for testing task switching abilities (Pashler, 2000). We then performed a hierarchical analysis to examine whether eye movement pattern could still predict comprehension after variation due to cognitive abilities was controlled.

Procedure

Comprehension of Documentaries During the task, participants' eye movements were recorded by an eye tracker, Eyelink1000. The tracking mode was pupil and corneal reflection with a sampling rate of 1000 Hz. Stimuli were displayed on a 22" CRT monitor with a resolution of 1920 by 1440 pixels and 150 Hz frame rate. The viewing distance was 60 cm. The standard nine-point calibration procedure was carried out before the experiment and whenever the drift correction error was larger than 1° of visual angle. Each trial started with a white solid dot appearing at the center of the screen. Participants were asked to look at the dot whenever it appeared for drift correction. Afterwards, a documentary clip was played in full screen. After each clip, participants were asked to answer 6 aurally-presented multiple-choice questions (MCQs) according to the content. The MCQs were presented one at a time binaurally in Mandarin, and the voice was synthesized by the online Baidu voice producer. Participants could replay each question unlimited times before their response. Participants performed the task in the 4 different documentary presentation conditions in separate blocks, with the block order counterbalanced across participants. They proceeded to the cognitive tasks described below after the comprehension task.

Cognitive Tasks

1. N-back Test: Two-back tests with 3 types of stimuli, including visual English letters, spoken numbers, and irregular shapes (Figure 2A) were used to test visual and verbal working memory. For each type of stimuli, participants were presented with 30 items one at a time, each for 2.5 s with a 0.5 s interval (Lau et al., 2010), and asked to judge whether the item presented in a trial matched the one that appeared 2 trials back.

2. Tower of London Test: Participants were asked to move 3 color discs one at a time from an initial position to match a goal position with the minimum number of moves, and to plan the moves in mind before execution (Figure 2B; Phillips, Wynn, McPherson & Gilhooly, 2001). Participants completed 10 trials. The total number of moves, total execution time, and total preplanning time before executing the first move were measured. Five practice trials were provided.

3. Multitasking Test: Four types of figures with different combinations of shapes and fillings were used as the stimuli (Figure 2C). The stimuli were presented one at a time in either the top or the bottom half of a box at the center of the screen (Figure 2C, left). Participants were asked to perform a dual task where they judged the shape of the figure (the shape task) as fast and correctly as possible when the figure was presented in the top half of the box, and judged the number of dots in the filling of the figure (the filling task) when the figure was presented in the bottom half of the box (Stoet, O'Connor, Conner & Laws, 2013). The figure was

presented for 2500ms, followed by a 500ms blank screen. Participants were asked to respond by the end of the 3-second trial. A shape-only and a filling-only task were tested sequentially before the dual-task to measure participants' baseline behavior where no task switching was involved. Their task switching ability was measured as the response time (RT) in the dual task minus the average RT during the two no-switching tasks.

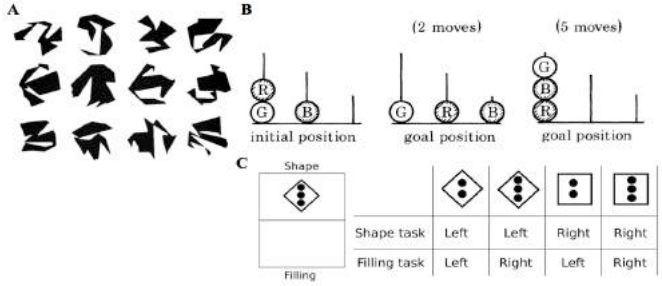


Figure 2: (A) 12 pictorial stimuli (Attneave and Arnoult structures) in the 2-back test, (B) Example of Tower of London test, (C) Stimuli used in the multitasking test.

Results

Effect of Video Content and Subtitle

The results showed a significant main effect of subtitle, F (1, 59) = 13.359, p = .001 (Figure 3). There was no main effect of video content or interaction between video content and subtitle. This result suggested that subtitles, but not video content, facilitated comprehension of documentaries.

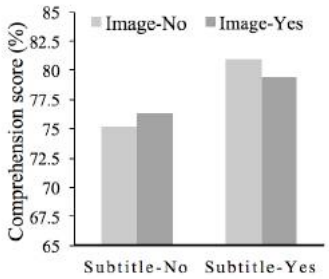
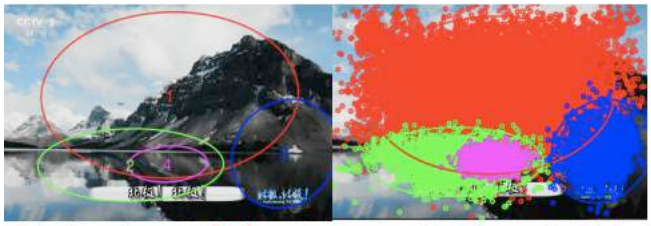


Figure 3. Effect of video content and subtitle on the comprehension of documentaries.

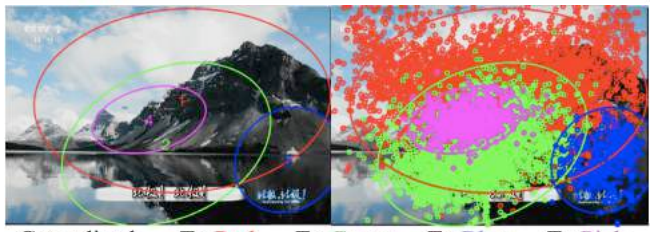
Eye Movement Strategies for Viewing Videos

Using the EMHMM method, for the video-only and video with subtitles conditions separately, we summarized each participant's eye movement pattern using an HMM. For each HMM, a variational Bayesian method was used to determine the optimal number of ROIs. More specifically, we ran each HMM with a different number of ROIs (ranging from 1 to 6) 300 times with a random initialization each time and selected the model with the largest log-likelihood given the data. We then clustered all individual HMMs into two groups and generated a representative HMM for each group with the number of ROIs set to 4.

Figure 4 shows the results of the *Video-only* condition. In the strategy on the top, after an initial fixation at the center of the video, participants had 8% of probability to look at either the blue ROI containing a logo on the bottom right or the green and the pink ROIs containing the fixed title on the bottom center of the screen. Afterwards, they tended to stay in the same ROI or switch back to the red ROI, or occasionally switched among the green, blue, and pink ROIs. We referred to this strategy as the distributed strategy. 46 participants were classified in this group. In contrast, in the strategy on the bottom, there were overlapping ROIs around the screen center, and participants mainly focused at the center of the screen. We referred to this strategy as the centralized strategy. There were 14 participants in this group.



Distributed	To Red	To Green	To Blue	To Pink
Priors	1.0	.00	.00	.00
From Red	.92	.05	.02	.01
From Green	.37	.56	.07	.00
From Blue	.38	.12	.47	.04
From Pink	.27	.00	.08	.64



Centralized	To Red	To Green	To Blue	To Pink
Priors	.46	.44	.01	.00
From Red	.91	.08	.01	.00
From Green	.18	.78	.04	.01
From Blue	.17	.39	.44	.00
From Pink	.03	.04	.00	.93

Figure 4. Distributed (top) and centralized strategies in the video only condition. Ellipses show ROIs as 2-D Gaussian emissions. The table shows transition probabilities among the ROIs. Priors show the probabilities that a fixation sequence starts from the ellipse. The image on the right shows raw eye fixation data and their ROI assignments.

To better understand the relationship between eye movement pattern and comprehension performance, following previous studies (e.g., Chan et al., 2018), we defined a Distributed-Centralized scale (D-C scale) for each participant as

$$Distributed\text{-}Centralized\ scale = \frac{D - C}{|D| + |C|}$$

Where D is the log-likelihood of the participant's eye movement data being generated by the representative HMM of the distributed strategy, and F is the log-likelihood of the participant's data being generated by the representative HMM of the centralized strategy. This log-likelihood measure reflects the similarity of the participant's eye movement pattern to the representative strategies. A more positive value in the D-C scale indicated higher similarity to the distributed strategy, whereas a more negative value indicated higher similarity to the centralized strategy. We found that participants' D-C scale was negatively correlated with comprehension performance in the video-only condition, r = -.291, p = .024: the more distributed the pattern, the lower the performance in comprehension (Figure 5A).

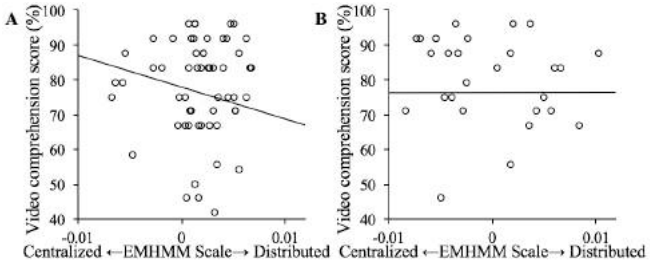
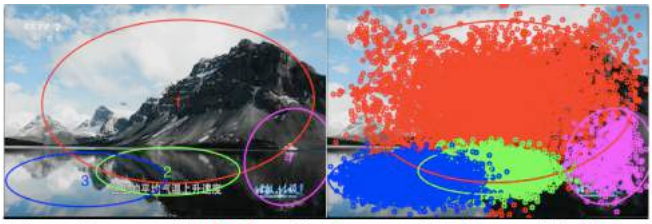


Figure 5. Correlation between eye-movement pattern and comprehension performance in the (A) video-only, and (B) video with subtitles condition.

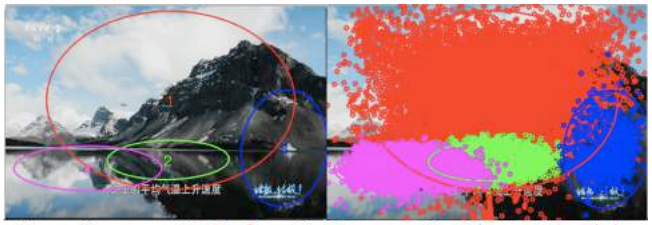
A similar analysis was conducted with eye movement data in the video with subtitles condition. Figure 6 shows the results of clustering participants' eye movement patterns into 2 groups. The 2 groups showed similar concentrations on the ROIs on the bottom left and bottom center of the screen, where the subtitles were located, in addition to the screen center. Group 1 strategy showed a higher probability to look at the subtitle regions after looking at the screen center. One-third of the participants (20 out of 60) adopted Group 1 strategy (one participant's eye movement data was invalid due to technical problems). We also measured participants' eye movement pattern similarity using the Group 1-2 scale in the same way as the D-C scale, and found that it did not correlate significantly with comprehension performance (Figure 5B).

Effect of Cognitive Abilities and Eye Movement Strategy on Comprehension

The above results suggested that in the video-only condition, participants' online eye movement pattern (D-C scale) predicted participants' comprehension. In addition to eye movement strategy, we found that comprehension performance was also significantly correlated with auditory working memory ability as measured in the n-back task, r = .337, p = .008, and task switching ability as measured in the multitasking test r = .262, p = .043. To examine whether eye movement pattern significantly contributed to comprehension after variation due to cognitive abilities was controlled, a three-stage hierarchical multiple regression was conducted to predict comprehension score. At stage one, auditory working memory capacity (N-back test) contributed significantly to the regression model, F(1,58) = 7.432, p = .008, and accounted for 11.4% of the variation. Adding task switching ability (Multitasking test) to the regression model explained an additional 5.5% of the variation in comprehension and the change in R^2 was significant, F(2,56) = 5.813, p = .005. Finally, introducing eye movement pattern (D-C scale) explained an additional 7.9% of the variation in comprehension score and this change in R^2 was significant, F (3,57) = 6.172, p = .001. Thus, when watching video documentaries without subtitles, participants' online eye movement behavior played an important role in comprehension in addition to cognitive abilities. A similar regression analysis was conducted for predicting comprehension performance in the video with subtitles condition, and no significant predicting variable was found.



Group1	To Red	To Green	To Blue	To Pink
Priors	0.98	.00	.02	.00
From Red	.74	.14	.11	.02
From Green	.22	.76	.01	.01
From Blue	.24	.01	.74	.01
From Pink	.23	.13	.13	.51



Group2	To Red	To Green	To Blue	To Pink
Priors	.99	.00	.00	.01
From Red	.79	.10	.01	.09
From Green	.22	.76	.01	.00
From Blue	.22	.13	.53	.11
From Pink	.24	.00	.01	.75

Figure 6. The two strategies observed in the video-with-subtitles condition.

Discussion

The present study aimed to investigate the role of video and subtitle in knowledge-based multimedia learning, with the effect of individual differences in multimedia processing including cognitive abilities and eye movement strategies considered. We hypothesized that for documentaries consisted of statements of facts, where auditory narratives provide most of the information, subtitles would facilitate comprehension due to the exact match to the content. In contrast, with video content, the comprehension may depend on participants' working memory, planning, and task switching abilities, as well as their eye movement strategy. Consistent with our hypothesis, we found that subtitles facilitated comprehension whereas video content did not. Through the EMHMM method, we discovered the distributed and centralized eye movement strategies in watching videos without subtitles. Interestingly, the more similar participants' eye movement pattern to the distributed strategy, the worse their comprehension in the video-only condition. Hierarchical regression analysis further showed that, while both auditory working memory and task switching abilities were significant predictors for comprehension, participants' eye movement pattern contributed significantly to comprehension after variation due to these cognitive abilities was controlled. This result showed that the facilitation of video content in the comprehension of documentaries depended on participants' online eye movement strategy in addition to working memory and task switching abilities. In contrast, participants' comprehension in the video with subtitles condition did not depend on either cognitive abilities or eye movement strategy.

Our results showed that adding subtitles is beneficial to knowledge acquisition of documentaries consisting of statements of facts. Previous studies on the effect of subtitles have reported inconsistent findings, with some showing that on-screen text is distracting (Mayer, Heiser & Lonn, 2001) whereas others suggesting facilitating effects (Starbek et al., 2010). This inconsistency may be due to differences in the amount of information carried in each medium during multimedia learning. In cases where pictorial stimuli contain important content for knowledge acquisition, simultaneous on-screen texts may be distracting (e.g., Mayer, Heiser & Lonn, 2001). In contrast, for materials where auditory narratives already provide most of the information for learning, such as the documentaries used in the current study, subtitles that match well with the auditory narratives may help maintaining participants' attention to the content of the knowledge and consequently facilitate comprehension (Kruger, Hefer & Matthew, 2013).

The finding that the distributed eye movement strategy, as opposed to the centralized strategy, was correlated with worse comprehension may be related to how attention was allocated in these two cases. According to the cognitive theory of multimedia learning (e.g., Mayer, 2014), engaging in active eye movement planning as demonstrated in the distributed strategy where there were specific ROIs located at different regions of the video may increase the cognitive load, resulting in decreased attentional resources for listening comprehension. The EMHMM method allows discovery of representative eye movement strategies from individual patterns in a data-driven fashion and provides a quantitative assessment of eye movement pattern similarities, leading this novel finding.

We also observed that in the video with subtitles condition, most participants focused on the subtitle locations in addition to the screen center, and participants' comprehension could not be predicted by either eye movement strategies discovered in this condition, working memory ability, or task switching ability, in contrast to the video-only condition. This result suggests that subtitles with video may help maintain participants' attention to a specific location of the video and reduce the possibility of active eye movement planning to other regions of the video, resulting in more attentional resources to listening comprehension and reduced task switching requirements. The redundancy effect from subtitles may also decrease the demands on working memory.

Previous research has suggested that cognitive abilities such as working memory capacity could modulate multimedia learning effects (Fenesi, Kramer & Kim, 2016). Here we further showed that in addition to working memory and task switching abilities, comprehension performance in the video-only condition could be predicted by online eye movement strategy: people who adopted a more centralized eye movement pattern had better comprehension. Future work will examine whether an explicit instruction of using a centralized eye movement strategy during video documentary viewing will facilitate comprehension.

To conclude, here we showed that for knowledge acquisition from auditory narratives, subtitles facilitated comprehension, whereas with video content, comprehension depended on participants' working memory and task switching abilities, as well as online eye movement strategy. When watching videos without subtitles, participants' comprehension could be facilitated by better auditory working memory and task switch abilities, and a more centralized eye movement pattern. In contrast, when watching videos with subtitles, subtitles seemed to have attracted and stabilized eye movements as well as reduced the demands on working memory and task switching, and thus neither cognitive abilities nor eye movement strategy could predict comprehension performance. These findings demonstrated the importance of taking individual differences into account in the research on instructional design and science of learning, and eye tracking with EMHMM provides a useful tool for revealing and quantitatively assessing these individual differences.

Acknowledgements

We are grateful to RGC of Hong Kong (Project # 17609117 to Hsiao).

References

Baadte, C., Rasch, T., & Honstein, H. (2015). Attention switching and multimedia learning: The impact of executive resources on the integrative comprehension of texts and pictures. *Scandinavian Journal of Educational Research*, *59*(4), 478-498.

Brasel, S. A., & Gips, J. (2014). Enhancing television advertising: Same-language subtitles can improve brand recall, verbal memory, and behavioral intent. *Journal of the Academy of Marketing Science*, *42*, 322-336.

Chan, C. Y. H., Chan, A. B., Lee, T. M. C., & Hsiao, J. H. (2018). Eye-movement patterns in face recognition are associated with cognitive decline in older adults. *Psychonomic Bulletin & Review*.

Chuk, T., Chan, A. B., & Hsiao, J. H. (2014). Understanding eye movements in face recognition using hidden Markov models. *Journal of Vision*, *14*(11), 8-8.

Coviello, E., Chan, A. B., & Lanckriet, G. R. (2014). Clustering hidden Markov models with variational HEM. *The Journal of Machine Learning Research*, *15*(1), 697-747.

D'Ydewalle, G., & De Bruycker, W. (2007). Eye movements of children and adults while reading television subtitles. *European Psychologist*, *12*(3), 196-205.

Fenesi, B., Kramer, E., & Kim, J. (2016). Split-attention and coherence principles in multimedia instruction can rescue performance for learners with lower working memory capacity. *Applied Cognitive Psychology, 30*(5), 691-699.

Hyona, J. (2010). The use of eye movements in the study of multimedia learning. *Learning and Instruction, 20*(2), 172-176.

Kruger, J. L., Doherty, S., & Soto-Sanfiel, M. T. (2017). Original language subtitles: their effects on the native and foreign viewer. *Comunicar: Media Education Research Journal*, *25*(50), 23-32.

Kruger, J. L., Hefer, E., & Matthew, G. (2013). Measuring the impact of subtitles on cognitive load: Eye tracking and dynamic audiovisual texts. In *Proceedings of the 2013 Conference on Eye Tracking South Africa* (pp. 62-66). ACM.

Kruger, J. L., & Steyn, F. (2014). Subtitles and eye tracking: reading and performance. *Reading Research Quarterly*, *49*, 105-120.偏

Lau, E. Y. Y., Eskes G. A., Morrison, D. L., Rajda, M., Spurr, K. F. (2010). Executive function in patients with obstructive sleep apnea treated with continuous positive airway pressure. *J. Int. Neuropsych. Soc., 16,* 1077- 1088.

Low, R., & Sweller, J. (2014). The modality principle in multimedia learning. In *The Cambridge Handbook of Multimedia Learning, Second Edition*, 227-246.

Mayer, R. E. (2014). Introduction to multimedia learning. In R. Mayer (Ed.), *The Cambridge Handbook of Multimedia Learning, Second Edition*, 1-24. Cambridge: Cambridge University Press.

Mayer, R. E. (2014). Cognitive theory of multimedia learning. In *The Cambridge Handbook of Multimedia Learning, Second Edition*, 43-71. Cambridge University Press.

Mayer, R. E., Heiser, J., & Lonn, S. (2001). Cognitive constraints on multimedia learning: When presenting more material results in less understanding. *Journal of Educational Psychology*, *93*(1), 187-198.偏

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology*, *41*(1), 49-100.

Perego, E., Del Missier, F., Porta, M., & Mosconi, M. (2010). The cognitive effectiveness of subtitle processing. *Media Psychology*, *13*(3), 243-272.

Phillips, L. H., Wynn, V. E., McPherson, S., & Gilhooly, K. J. (2001). Mental planning and the Tower of London task. *Q. J. Exp. Psychol. - A*, 54, 579–597

Schnotz, W. (2005). Integrated model of text and picture comprehension. In R. Mayer (Ed.), *The Cambridge Handbook of Multimedia Learning* (2nd ed.), 72–103. Cambridge: Cambridge University Press.

Starbek, P., Erjavec, M. Starcic, & Peklaj, C. (2010). Teaching genetics with multimedia results in better acquisition of knowledge and improvement in comprehension. *Journal of Computer Assisted Learning, 26*(3), 214-224.

Stoet, G., O'connor, D., Conner, M., & Laws, K. (2013). Are women better than men at multi-tasking? *BMC Psychology, 1*(1), BMC Psychology, 12/2013, Vol.1(1).

Van Gog, T., & Scheiter, K. (2010). Eye tracking as a tool to study and enhance multimedia learning. *Learning and Instruction*, *20*(2), 95-99.

Conceptualization of Cultural Diversity for Efficient and Flexible Manufacturing Systems of the Future

Kashif Zia (kzia@su.edu.om)
Faculty of Computing and Information Technology, Sohar University,
Al Jamia Street, Sohar, Oman

Alois Ferscha (ferscha@soft.uni-linz.ac.at)
Institute of Pervasive Computing, Johannes Kepler University Linz
Altenberger Strae 69, 4040 Linz, Austria

Dari Trendafilov (dari.trendafilov@pervasive.jku.at)
Institute of Pervasive Computing, Johannes Kepler University Linz
Altenberger Strae 69, 4040 Linz, Austria

Abstract

Manufacturing systems of the future need to have flexible resources and flexible routing to produce extremely personalized products, even of lot size equal to one. In this paper we have proposed a framework, which is designed to achieve this goal. Towards this we have integrated an established cultural evolution model to achieve desirable flexibility of resources and acceptable routing time. Promising results are evidenced through a simple proof-of-concept agent-based simulation. The simulation results reveal that the products need to move less in more diversified cultural groups when looking for suitable resources. It was also observed that the more time we provide for cultural dissemination, the cultural groups become increasingly coherent due to homophily. For scenarios, which require diversification of resources, we need to find a balance between coherence and diversification. This paper provides first insights into these aspects for a production shop floor.

Keywords: Industry 4.0; resource flexibility; routing flexibility; personalized production; cultural dissemination; group coherence.

Introduction

The industrial manufacturing paradigm has already evolved from mass production to mass customization. Fueled by initiatives like Industry 4.0 (Lee, Bagheri, & Kao, 2015), we foresee a further improvement in coming years, namely the paradigm of personalized production. Personalized production targets an extremely flexible manufacturing system which could respond to predicted and unpredicted changes in the production environment and allows customers to create and design themselves (Hu, 2013; Mourtzis & Doukas, 2014). Manufacturing systems supporting personalized production should exhibit the following features (Ogunsakin, Mehandjiev, & Marín, 2018):

- Resource Flexibility: flexibility of processing stations (or machines) to make multiple parts, which means that one processing station is not designated for one task and can perform different tasks as required.
- Routing Flexibility: flexibility to execute same operation (or function related to a task) using multiple processing stations, which means that a single task can be performed by many processing stations.
- Lot Size Flexibility: ability to produce a very small customized and/or personalized lot size in a non-batch mode, which is a direct consequence of at least (if not any other dimension) the above two features.

The progress towards a truly flexible manufacturing system (FMS) is naturally driven by technological needs from industrial process management viewpoint, which falls into general knowledge areas of scheduling (Wang, Zhong, Dai, & Huang, 2016; Marichelvam, Prabaharan, & Yang, 2014), resource optimization (Ogunsakin et al., 2018; Beruvides, 2017), constraint satisfaction (Ezpeleta, Colom, & Martinez, 1995), and related.

Still, the body of work considering the aspect of "personalization" is quite lean and requires further attention. Realizing this, several projects and activities are already under progress. One significant effort endeavours to develop cognitive products and production systems incorporating *human-like* capabilities like "perception, understanding, interpretation, memorizing and learning, reasoning, planning and hence cognition-based acting" (Pro2Future, n.d.). The project is about complex cognitive modalities of humans, products and machines and their interrelationships. In this paper, we argue that one does not need to have high-level cognitive capabilities to be effective. At a scale of a population or a group, a very basic level cognition of interacting agents may result in a desirable global situation. We just need to find the conditions in which this may happen.

Agent-based modeling (ABM) (Bonabeau, 2002) is a method used for modeling such inquiries. One particular area of interest of a production unit is the layout of shop floor which should be optimized for maximum gain in productivity, particularly in case of FMSs. This case study is adopted in our paper. At a conceptual level, a group of agents comprising an interactive social network is augmented with the notion of culture to ground them with the physical world.

Most optimization mechanisms (as stated above) either consider a mathematical abstraction or imitate a real-world situation as their manufacturing environment (which is

mostly *static*) while modeling, and then proposing a solution within these presumptions. A more recent work (Ogunsakin et al., 2018) also considers mobile processing stations as a mean to achieve flexibility of shop floors. The idea is to make resources available as and where these are required. Although, this approach addresses the challenge of routing flexibility to an extent, the capabilities of resources still remain static.

In our research, we are mostly focusing on resource flexibility, which means that the processing units are able to dynamically change their *capabilities* and therefore a resource is able to perform several tasks. The goal is to keep resources stationary (and avoid expensive process of mobility) and arrange resources in groups of *complementing* capabilities. Ideally, a resource would designate itself for a capability that would optimize the manufacturing process in several dimensions, such as production rate, lead-time per order and reactivity index (Ogunsakin et al., 2018). However, we only focus on resource availability and mobility of products.

We postulate that flexibility in resources, routing and personalizing relate to evolution of culture as it emerges at the physical level due to local interactions of *mostly* stationary individuals. In the context of resources (processing units) of a production shop floor, we seek for groups of *complementing* capabilities, self-organizing to produce an approximately optimized layout for the products, which ensures availability of resources and reduces products' mobility. This novel idea would provide an entirely new perspective for the future research in this domain.

In the next section, formal definition of culture and cultural diversification is presented; followed by detailed description of the methods. Next, we present details about our model and simulation; followed by discussion on initial findings. We end the paper with an elaborate outlook of future work.

Culture, Diversification and FMS

A culture is a multi feature system evolving in time. One characteristic of culture is its coherence when seen from outside. Definitely, this coherence results due to a majority of people trying to acquire a similar behavior (often termed as a trait) in a certain context (often termed as a feature).

Relating these concepts to FMS, we need to conceptualize features and traits of resources and products, where a resource is a processing unit in the production line, whereas a product is obviously a product under production. Although a product can also be considered as a cultural entity, it is not the case for the purpose of this paper. Only a resource is a cultural entity.

Resources are flexible, initially having some randomly chosen features and a randomly chosen trait against a feature. For example, a processing unit may have ability to perform one, two or more tasks T_1, T_2, ... with certain levels of precision P_1, P_2, ... Here, a tuple consisting of n values is a set describing capabilities of a resource. For example, the set $\{P_2, P_1, P_3\}$ can be interpreted as: this resource can perform task 1 with precision 2, task 2 with precision 1 and task 3 with precision 3. Furthermore, it cannot perform any other task.

Further, all products have a *sequential* list of capability requirements. For example, a product with set $\{P_1, P_1, P_2\}$ requires task 1 with precision 1, followed by task 2 with precision 1, finally followed by task 2 with precision 2. The question is: would cultural diversification be able to generate a physical layout that would ensure availability of capable resources with minimal mobility for all the products in the system? Technically, what are conditions which lead us to an acceptable (and approximate) solution of the problem?

Such a scheme is naturally compatible with the requirement of a flexible manufacturing system stated above, namely, flexibility in resources, routing and personalizing. Axelrod provides evidence in his seminal work (Axelrod, 1997) for such a simple configuration of cultural descriptions which can result into a locally coherent, but globally polarized culture as a consequence of localized interactions of participating entities.

Our intuition is that unbounded coherence between cultural groups would not help in this scenario. The reason is that limitless coherence has no control over where the boundaries of the global polarization would occur, which is not compatible with a system which seeks for economy of resources and optimizations in several dimensions. That is the reason, we try to find conditions which end up in approximately acceptable structuring in terms of coherence (termed as limited coherence) vs. polarization. To achieve this, we have taken motivation from Centola et. al's work (Centola, Gonzalez-Avella, Eguiluz, & San Miguel, 2007) in which a random drift is used to deviate a highly coherent environment. This drift is achieved through change in the neighborhood of an agent. Theoretically it is possible to do it, however in scenarios like FMS it is not practical as we cannot move processing units so frequently after deployment. Hence, we have fine tuned Axelrod's model of cultural dissemination (Axelrod, 1997) with focus on limited coherence between cultural groups and tried to find out how much we can achieve and in which conditions. Definitely, at run time, the dynamics of requirements and products may change and make a particular layout extremely inefficient. To address it, a further investigation is required, which is planned for the future.

Methods

Axelrod's Model of Cultural Dissemination

Axelrod's model (Axelrod, 1997) thrived for cultural homogeneity (Bednar, Bramson, Jones-Rooy, & Page, 2010), where adjacent cultures get influence from each other. The model is based on cultural components defined by three factors: features, traits and persons. A culture has many features, such as habits of eating, recreation and leisure. These features may not be identical across different cultures. Each of these features have several traits, which may differ across cultures. A person is a placeholder of a culture described by one of f features and t traits. Axelrod proposed a model seeking for cultural homogeneity proclaiming that different cultures are

destined to cohere together so that they appear as a cultural unity, but at the same time, there exists a clear-cut differentiation between cultures.

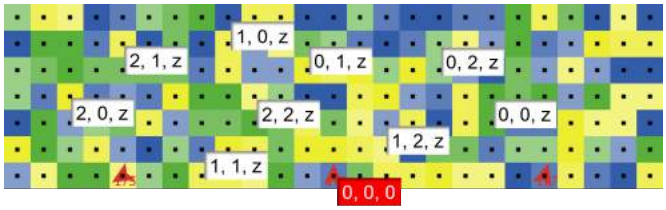


Figure 1: Initial distribution of a 25×7 grid constituted by blocks of culture (anchored on central black persons); each block is a tuple of 3, representing three features (green, blue, yellow) of three traits each (3 shades of a color). Each cell's color has a meaning; for example, all green cells have capability to perform task 1 with precision value 0, which is followed by precision values of task 2 (0, 1 or 2); last value is not path dependent and represented by z. Possible combinations of colors are shown with values; each tuple relating to a person on the top-left corner. A product has a unique sequence of task to perform represented with an arrow shape (at the bottom center of the space).

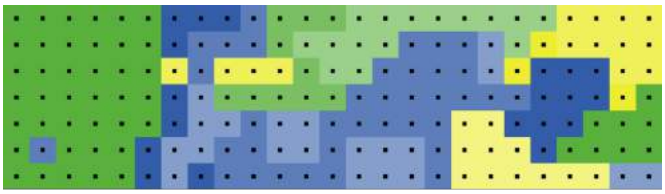
Axelrod model was able to demonstrate that the above two (rather contradictory) goals can be achieved by a simple interaction model (realized through N coordination games) between neighboring persons. Axelrod showed that N coordination games are necessary for a broader scale evolution of a culture. Furthermore, groups' consistency across different aspects of societal norms makes a group culturally coherent and different from others.

We developed a simple simulation model for demonstration purposes using NetLogo (Tisue & Wilensky, 2004). Figure 1 presents a grid of 25×7 cells. Each cell is represented by a person (in black) and the corresponding culture acquired, depicted by cell color of the cell the person is occupying.

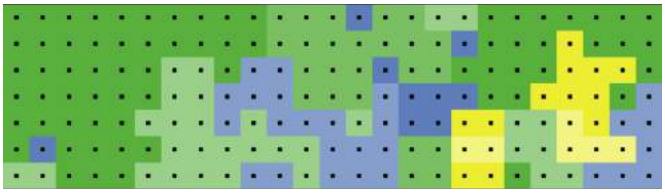
Axelrod's model calculated similarity s between neighboring cultures (based on Von Neumann's neighborhood). If s is not 1 (100%), with a probability p, the value of a $different$ column of a person is replaced by the corresponding value of the neighboring person. This simple mechanism is able to generate clusters of coherent cultures as shown in Figure 2. If we define **diversity index** as the mean diversification of cultures of all persons when compared to their neighbors, the Axelrod model would converge into a single culture most of the time with diversity index equal to 0. This is not desirable in the context in which we want to use this model. Therefore, the model was extended as detailed in the following.

Model Motivation: Constrained, N-Coordination Games for Cultural Diversity

Before describing the model, we will emphasize the scenario given in Figure 1. Given that a processing unit is able to perform three possible tasks with three possible precision values,



(a) Axelrod Model: diversity index = 0.34 at 10000th iteration



(b) Axelrod Model: diversity index = 0.25 at 20000th iteration

Figure 2: Axelrod's Model: Evolution of cultures shown in Figure 1. (a) at simulation iteration 10000 showing clusters of cultures starting to form. (b) at simulation iteration 20000 showing further consolidation of clusters of cultures. The evolution is destined to end up in very few cultures (1 or 2).

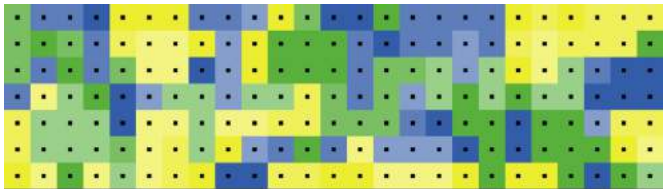
we can see a clear capability matching through colors. Furthermore, a product is introduced which need to complete a sequence of three tasks offered by different resources. We hypothesize that using the constraint, N coordination games, we can achieve cultural diversity closer to what is desirable. This would directly impact products' traversing efforts in a positive way.

The Proposed Diversity Dissemination Mechanism

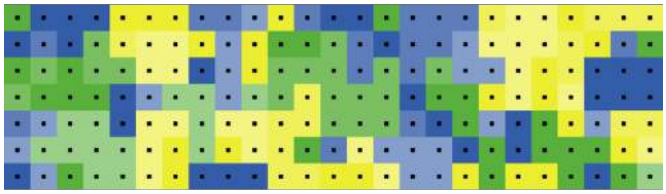
The Axelrod model is too skewed towards coherence and would end up in too few cultures. Hence we propose to refine the Axelrod model in the following way. Axelrod model seeked for similarity s between neighboring cultures. If s is not 1 (100%), with a probability p, the value of a $different$ column of a culture is replaced by corresponding value of the neighboring culture. We extend this model by applying an extra constraint. That is, the replacement is only possible if s is also less than a threshold th, which is for now given a static value of 0.5. This obviously increases the overall *diversity_index* of the system as shown in Figure 3. Before analyzing the impact of this refinement we explain the mechanism of product traversing.

Traversing Mechanism

All products have a sequence of tasks to perform in the form [x, y, z]. A product first gets the value x, and maps it onto resources with identical capability and residing close to its position. Let's denote the resource with r. After visiting r, the product seeks for the next nearest resource corresponding to y. It is assumed that y has a relationship with x. This means that, in terms of colors, this cell (and the resource residing on top of it) should have the same color. The last task z is independent and just show the range of flexibility that the system



(a) Proposed Model: diversity index = 0.50 at 10000th iteration



(b) Proposed Model: diversity index = 0.44 at 20000th iteration

Figure 3: Proposed Model: cultural diversity at iteration (a) iteration 10000 and (b) 20000.

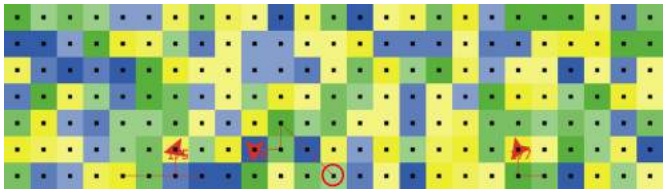


Figure 4: Traversing behavior in random configuration of resource capability.

may have.

An Example Walk-through on Random Configuration (without applying diversification mechanism): Referring to Figure 4, each resource (black agent at the center of a cell) is randomly populated with vector [x y z], where x, y and z may have three possible values 0, 1 and 2. Products have to perform three tasks in a sequence. One product (at the center) has to perform task 1 with precision 0, task 2 with precision 1 and task 3 with precision 1. It starts at the position marked with red circle. First it performs task 1 with precision 0. That takes it 2 steps to the top left cell, which has the nearest resource with this capability. Next, it has to perform task 2 with precision 1. The nearest resource, which has first column equal to 0 (assuming a connection between task 1 and 2) and second column equal to 1 is the resource at the bottom; hence the product would move there. Next task is task 3 with precision 1. Assuming that it is an independent task, the product would try to find the nearest resource that has the third column equal to 1 (any color). This can be any resource (cell at the left is selected). Hence, the **mobility index** of this product is 4, the total number of hops traversed. The other two products also traverse to complete their tasks. The average mobility index turns out to be 3.94.

It seems that random configurations would be the best, but this cannot be the case in a structured environment, particularly in case of an assembly line type of manufacturing.

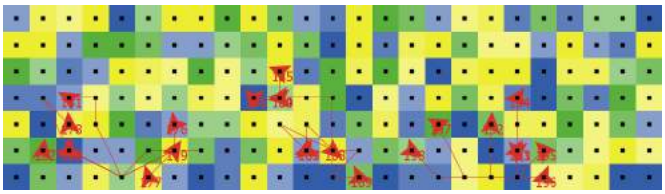
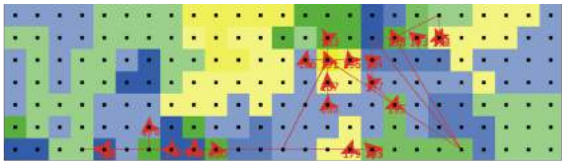
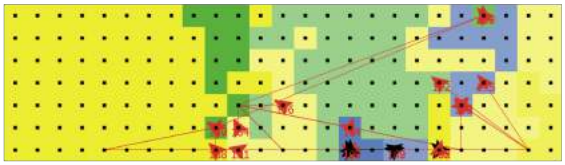


Figure 5: Traversing behavior in random Layout without diversification applied.



(a) Axelrod Model (10000th iteration): diversity index = 0.34, mobility index = 6.78



(b) Axelrod Model (20000th iteration): diversity index = 0.24, mobility index = 8.66

Figure 6: Traversing behavior in Axelrod's Model.

Analysis of Initial Findings

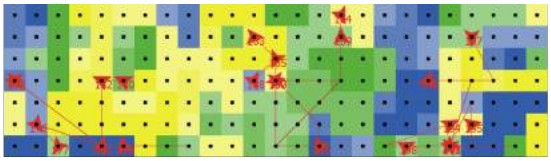
Definitely, the introduction of *th* retains diversity index in case of extension of Axelrod's model. This helps in task completion capability of the system. This claim can be verified by analyzing the mobility of products and the diversity index in three cases. We have used 25 products distributed at three places. In each case, the simulation was run for 100 times and the results were averaged. In the following, we present a sample visualization for each case which is close to average values, at two sampling points (iteration 10000 and iteration 20000) if applicable.

Random Layout

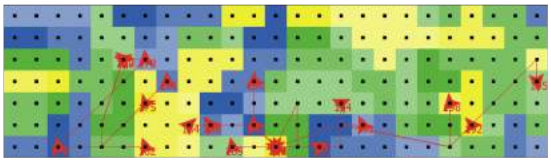
In Figure 5, the system has a diversity index equal to 0.70 and a mobility index equal to 3.4. This is also confirmed by the graphs shown in Figure 8 (diversity index) and Figure 9 (mobility index). As we mentioned already, random configuration is most flexible and would always be best in its task completion capability. However, this configuration is unrealistic. In reality, we need to plan placement of resources and deploy them accordingly.

Axelrod's Model

In case of Axelrod's model, we have analyzed the results for diversification period of 10000 and 20000 iterations. These two situations are represented in Figure 6. With increasing polarization and decreasing diversity index, the average mobility index drops. After running the simulation for 100 runs and averaging, it was observed (see Figure 8 (diversity index) and Figure 9 (mobility index)) that mobility index is just less



(a) Proposed Model (10000th iteration): diversity index = 0.47, mobility index = 5.00



(b) Proposed Model (20000th iteration): diversity index = 0.24, mobility index = 8.66

Figure 7: Traversing behavior in proposed Model.

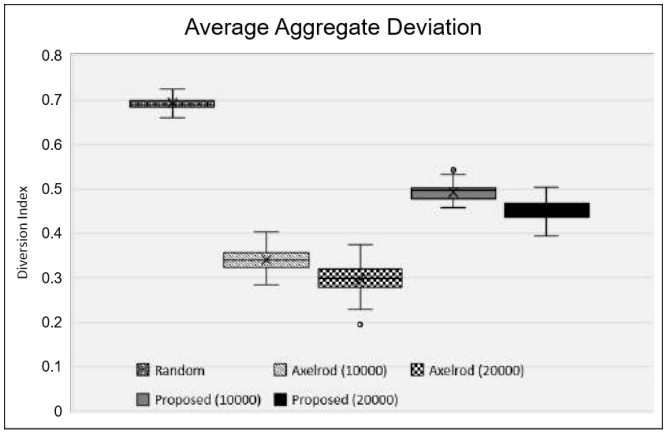


Figure 8: Graph showing diversity index of 100 simulations runs.

than 8 (diversity index = 0.35) in case the diversification happens for 10000 iteration, whereas, mobility index is slightly higher than 8 (diversity index = 0.30) in case the diversification happens for 20000 iteration. As shown in Figure 6, this decrease is due to nonavailability of resources indicated by products turning into black color.

Proposed Model

Lastly, the proposed model solves the above issue. We can see a smooth performance of tasks for all the products, which is evident from Figure 7. Again, we have analyzed the results for diversification period of 10000 and 20000 iterations. These two situations are represented in Figure 7. After running the simulation for 100 runs and averaging, it was observed (see Figure 8 (diversity index) and Figure 9 (mobility index)) that mobility index is equal to 4.57 (diversity index = 0.49) in case the diversification happens for 10000 iteration, whereas, mobility index is about 5 (diversity index = 0.45) in case the diversification happens for 20000 iteration.

Comparative Analysis

As diversity decreases, the availability of resources becomes more difficult. In this particular scenario, the products need

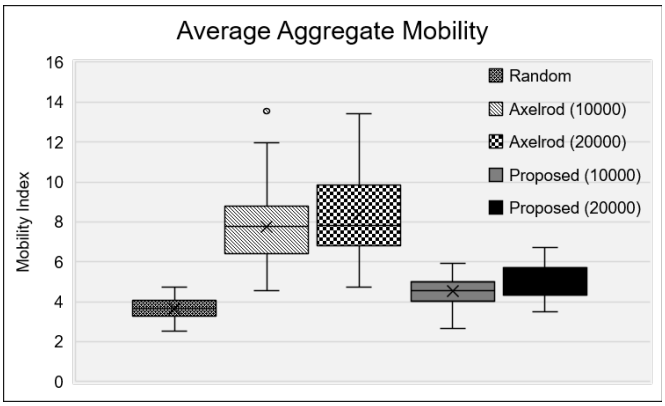


Figure 9: Graph showing mobility index of 100 simulations runs.

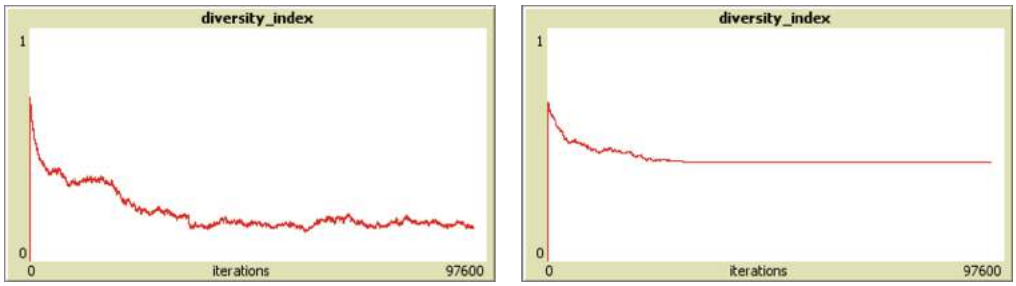
to move less in more diversified cultural groups. The ideal case is random layout in which the products need to move the least. As diversity decreases from random layout to Axelrod's model, the mobility increases substantially. In case of Axelrod's model, it was also observed that the more time we provide for cultural dissemination, the cultural groups become increasingly coherent. In the simulation world's geometry used, the number of culture clusters goes down to a few if the number of iterations is increased to 100000. Obviously, this is not an interesting case to show. However, in the case of the proposed model, this does not happen with such high intensity. In fact, the diversity index never drops below 0.4 and interestingly it reaches an equilibrium in most runs. Hence, it is possible to provide a drift against unbounded homophily effect resulting into an extremely low diversification by using a simple threshold based control mechanism. The graphs shown in Figure 10 validate our claim.

Conclusion and Outlook

Manufacturing systems of the future need to have flexible resources and routing to produce extremely personalized product, even of lot size equal to one. What we have seen is that flexible manufacturing systems can be realized without moving the resources (processing units) by enabling reconfiguration of capabilities of resources based on dissemination of culture concept proposed by Axelrod. However, the Axelrod model has a focus on coherence of cultural groups, which most of the times ends up in one or very few cultures. If we equate such an instance of a culture with a single capability of a resource, we are left with extremely limited resources and products cannot complete their production life cycle.

Hence, we proposed to have a constrained cultural coherence mechanism by introducing a threshold. This tiny development has a significant impact on the increase in diversity of the culture along with related resources being in close vicinity to each other on average. This not only ensured an increase in resource availability as a whole, but also managed to decrease the mobility of products in search of suitable resources.

However, the real contribution of the paper is integration of



(a) Axelrod's Model (b) Proposed Model

Figure 10: Comparative Analysis of diversity index: Axelrod's Model vs. Proposed Model.

manufacturing processes with cultural considerations, which naturally fits into the problem. In our view this is a novel approach of real significance. However, the work reported in this paper is just a proof-of-concept. We need to have more thorough experiments to measure the efficiency of the model in challenging environments such as environments having inflow and outflow points, more in-depth capabilities and richer relationships between tasks.

In the next phase of the project, we will induct models of dynamics, which include timing of tasks, conflict and deadlock resolution between products seeking for identical resources, and more realistic analytics such as production rate, lead-time per order and reactivity index. Lastly, we would also include an autonomous learning system, which would help resources learn and change their configurations on the fly based on product types, requirements and trajectories.

Acknowledgment

The authors would like to acknowledge support by FFG funded Pro^{2}Future under contract No. 6112792.

References

Axelrod, R. (1997). The dissemination of culture: A model with local convergence and global polarization. *Journal of conflict resolution*, *41*(2), 203–226.

Bednar, J., Bramson, A., Jones-Rooy, A., & Page, S. (2010). Emergent cultural signatures and persistent diversity: A model of conformity and consistency. *Rationality and Society*, *22*(4), 407–444.

Beruvides, G. (2017). Artificial cognitive architecture with self-learning and self-optimization capabilities. case studies in micromachining processes.

Bonabeau, E. (2002). Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences*, *99*(suppl 3), 7280–7287.

Centola, D., Gonzalez-Avella, J. C., Eguiluz, V. M., & San Miguel, M. (2007). Homophily, cultural drift, and the co-evolution of cultural groups. *Journal of Conflict Resolution*, *51*(6), 905–929.

Ezpeleta, J., Colom, J. M., & Martinez, J. (1995). A petri net based deadlock prevention policy for flexible manufacturing systems. *IEEE transactions on robotics and automation*, *11*(2), 173–184.

Hu, S. J. (2013). Evolving paradigms of manufacturing: from mass production to mass customization and personalization. *Procedia CIRP*, *7*, 3–8.

Lee, J., Bagheri, B., & Kao, H.-A. (2015). A cyber-physical systems architecture for industry 4.0-based manufacturing systems. *Manufacturing Letters*, *3*, 18–23.

Marichelvam, M. K., Prabaharan, T., & Yang, X. S. (2014). A discrete firefly algorithm for the multi-objective hybrid flowshop scheduling problems. *IEEE transactions on evolutionary computation*, *18*(2), 301–305.

Mourtzis, D., & Doukas, M. (2014). Design and planning of manufacturing networks for mass customisation and personalisation: challenges and outlook. *Procedia CIRP*, *19*, 1–13.

Ogunsakin, R., Mehandjiev, N., & Marín, C. A. (2018). Bee-inspired self-organizing flexible manufacturing system for mass personalization. In *International conference on simulation of adaptive behavior* (pp. 250–264).

Pro2Future. (n.d.). *Area 4.1 cognitive products.* Retrieved from `http://www.pro2future.at/research-en/areas-en/area-41-en/`

Tisue, S., & Wilensky, U. (2004). Netlogo: A simple environment for modeling complexity. In *International conference on complex systems* (Vol. 21, pp. 16–21).

Wang, M., Zhong, R. Y., Dai, Q., & Huang, G. Q. (2016). A mpn-based scheduling model for iot-enabled hybrid flow shop manufacturing. *Advanced Engineering Informatics*, *30*(4), 728–736.

The price of knowledge: Children infer epistemic states and desires from explorations cost

Rosie Aboody
Yale University, New Haven, Connecticut, United States

Caiqin Zhou
Wellesley College, Wellesley, Massachusetts, United States

Julian Jara-Ettinger
Yale University, New Haven, Connecticut, United States

Abstract

When deciding whether to explore, people must consider both their need for information, and the cost of obtaining it. Thus, to judge why someone explores (or decides not to), we must consider not only their actions, but also the cost of information. Do children attend to the cost of agents exploratory choices when inferring what others know or desire to know? In Experiment 1, four- and five-year-olds judged that an agent who rejected an opportunity to gain low-cost information must have already known it. In Experiment 2, four- and five-year-olds judged that an agent who incurred a greater cost to gain information had a greater epistemic desire. In two control experiments, we show that these results cannot be explained by a low-level heuristic linking competence with knowledge. Our results suggest that childrens Theory of Mind includes expectations about how costs interact with epistemic desires to produce action.

Ignorance = doing what is reasonable: Children expect ignorant agents to act based on prior knowledge

Rosie Aboody[1] (rosie.aboody@yale.edu), Madison Flowers[1] (madison.flowers@yale.edu), Caiqin Zhou[2] (czhou@wellesley.edu), & Julian Jara-Ettinger[1] (julian.jara-ettinger@yale.edu)

[1] Department of Psychology, Yale University. New Haven, CT 06520 USA.
[2] Department of Psychology, Wellesley College. Wellesley, MA 02481 USA.

Abstract

When deciding how to act in new situations, we expect agents to draw on relevant prior experiences. This expectation underlies many of our mental-state inferences, allowing us to infer agents' prior knowledge from their current actions. Do children share this expectation, and use it to infer others' epistemic states? In Experiment 1, we find that five- and six-year-olds (but not four-year-olds) attribute additional knowledge to agents whose prior experiences cannot explain their success. In Experiment 2, we find that six-year-olds (but not younger children) also attribute greater knowledge to agents whose prior experience cannot explain their failure. We show that by age five or six, children expect ignorant agents' beliefs (and therefore their actions) to be guided by their prior knowledge. This work adds to a growing body of research suggesting that, while infants can represent mental states, the ability to infer mental states continues to develop throughout early childhood.

Keywords: Ignorance; Knowledge; Social Cognition; Theory of Mind

Introduction

To discuss someone's ambitions, frustrations, or disappointments is to talk about a mind that works much like our own, except that we cannot see it or know what it knows. Yet, we make surprisingly accurate inferences about what others think or want, just by watching how they act. For example, if your friend gives you her keys but later rummages in her bag upon reaching the car, you might infer that she forgot you have them. If she doesn't slow for a pedestrian at a crosswalk, you'd probably assume she didn't see them. And if she suddenly takes a detour, you might suspect she knows something you don't (perhaps the usual route is under construction).

The ability to infer other people's thoughts and desires from their behavior involves building a working model of how their mental states relate to their actions. The foundations of this capacity, called a Theory of Mind (Dennett, 1987; Gopnik & Wellman, 1992), are in place and at work early in infancy (Woodward, 1998; Liu, Ullman, Tenenbaum, & Spelke, 2017) but continue to mature throughout early childhood (Wellman, Cross, & Watson, 2001), and well into adolescence (Richardson, Lisandrelli, Riobueno-Naylor & Saxe, 2017).

Within Theory of Mind, our ability to reason about other people's beliefs—what they know, what they don't, and what they think they know—is particularly slow to develop. While infants can represent other people's beliefs (Onishi & Baillargeon, 2005), knowledge (Surian, Caldi & Sperber, 2007), and ignorance (O'Neill, 1996), children do not use these representations explicitly until several years later (Bartsch & Wellman, 1995; Wellman, et al., 2001).

As adults, we understand that other people's past experiences shape their current beliefs, and that these beliefs guide their actions. If, for example, your friend starts their car by inserting and turning a key, you can reasonably predict they will try the same the first time they drive yours. And you'd expect this even if you know your car works differently (for example, starting when a button is pushed in proximity to the key fob).

This expectation not only allows us to predict how others will act: it also allows us to infer what they know by observing how they act. In the example above, if your friend defied your expectations by immediately locating the button that starts your car, you might wonder if they had some prior experience you didn't know about (perhaps they've driven other cars like yours before). Such reasoning may seem intuitive, but how exactly do we predict what actions agents are likely to take in new situations? Prior research suggests that adults solve this problem by integrating over agents' uncertainty (Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017). For instance, when we reason about an agent who does not know whether a car starts via a key or a button, we consider what they would do in each situation, and we expect them to choose a plan weighted by their confidence.

While effective, these types of inferences are computationally complex. They require considering multiple possible worlds (at least implicitly), and deciding what an agent would do in each. Perhaps unsurprisingly, children's expectations for how ignorant agents are likely to act appear to rely on simpler strategies. Children sometimes equate being ignorant with getting things wrong (Ruffman, 1996; Saxe, 2005); although, in other contexts, their intuitions reverse (Friedman & Petrashek, 2009; German & Leslie, 2001).

While expecting ignorant agents to fail may support accurate inferences and useful predictions, such strategies are limited. Even ignorant agents can make reasonable guesses based on past experience. For instance, even if you've only used PC's, you probably have some idea of what you'd try if you had to turn on a Mac. And ignorant agents can always get lucky, succeeding by chance.

Do children understand how previous experiences affect agents' future actions? And do they leverage this expectation to infer what an agent knows based on what she does? In the current work, we investigate these questions with four- to six-year-olds. The ability to explicitly and flexibly represent beliefs emerges in the mid-preschool years (e.g., Rubio-Fernández, 2019; Wellman et al., 2001). Therefore, if children have expectations about the relation between ignorance and action, we might expect them to emerge in this age range.

In two experiments, participants watched two puppets learn how to activate a novel toy. Each puppet later attempted to activate a different (but outwardly identical) toy. One agent's actions were consistent with their prior experience, while the other agent's actions were inconsistent with their prior experience. In Experiment 1, both agents succeeded in activating a toy. If children expect agents to act based on their prior knowledge, they should judge that the inconsistent agent (whose actions cannot be explained by their experience with the initial toy) must have had additional knowledge. We find that five- and six-year-olds (but not four-year-olds) attribute additional prior knowledge to this agent.

To control for the possibility that children attribute knowledge to agents who teach them something new, in Experiment 2, children learned how a toy worked, and then watched two agents fail to activate this toy. Children again judged that the inconsistent agent (whose action couldn't be explained by his experience with the initial toy) had *greater* additional knowledge. These results suggest that by age five, children expect ignorant agents to act according to their prior knowledge, and further, that children leverage this expectation to infer what others know from what they do. All experiments' procedures, predictions, exclusion criteria, and analyses were pre-registered.

Experiment 1

In Experiment 1, children watched two puppets learn how to activate a novel toy. Next, each puppet was given the chance to activate a different toy (always outwardly identical to the original). One puppet stated that his chosen toy worked the same as the original, and pressed the same button he had seen activate the original toy. The other puppet stated that his chosen toy worked differently to the original, and pressed a different button. Both puppets succeeded in activating their chosen toy. Children were then asked which of the two agents already knew how the toys worked.

If children expect ignorant agents to behave in accordance with their prior beliefs, then they should judge that the agent who acted inconsistently with their prior experience is more likely to be knowledgeable. But if children attribute epistemic states by relying on a rule of thumb (e.g., expecting ignorant agents to be wrong), or have no representation of what it means to be ignorant, then children should have no preference for either agent.

Method

Participants 72 four-, five- and six-year-olds (mean age: 5.46 years, range: 4.05 – 6.99 years; n = 24 participants per age group) were recruited at a local children's museum. 22 participants were excluded from the analyses and replaced because: they did not pass the pre-registered inclusion questions (n = 9), due to experimenter error (n = 5), interruptions from other children (n = 3), because the participant did not answer the test question within 30s (n = 2), distraction (n = 1), interference with the procedure (n = 1), or due to developmental delays (n = 1).

Stimuli Stimuli consisted of two male puppets, and three novel toys. These toys were externally identical machines, each covered in black construction paper and measuring approximately 5 x 3 x 2.75 in. Toys had three buttons on top: a red button in the middle, and two black buttons flanking the red one (see Figure 1).

Although they all looked the same, the toys worked in different ways. The first toy (called the "training" toy) activated and played music only when the central red button was pressed. Of the remaining toys, the "consistent" one worked the same way. However, the "inconsistent" toy worked differently: only pressing the black button to the participant's far left made it activate. For clarity, we refer to this button as the "correct" black button, and the other as the "incorrect" black button (since it did not activate the toy).

Procedure First, participants were familiarized with the training toy (which turned on when the central red button was pressed). Participants learned that the red button made the toy go, but that the black buttons did nothing. They were then given a chance to press all of the buttons themselves. Next, participants were introduced to two puppets. The experimenter explained that she was going to show the puppets how the toy worked, and told the puppets that while the red button made the toy go, the black buttons did not do anything. Upon the experimenter's request, the puppets pressed the red button together.

Next, the remaining toys were placed on the table (one on either side of the training toy). The experimenter explained that one of the puppets had snuck out from under the table and played with all the toys, and discovered which buttons made the toys play music. The other puppet had stayed underneath the table, and hadn't seen anything. The child's task was to help figure out which puppet had snuck out and played with all the toys.

Each puppet was questioned individually, while the other agent was placed under the table. During his turn, each puppet was asked: "Can you show us how to make one of these toys go?" To make the relation between agents' actions and their experience with the initial toy more explicit, each agent explained himself as he acted. One puppet chose the consistent toy, saying, "Hmm. Well, the red button made this [original] toy go, so the red button makes this toy go too," pointing to the two relevant buttons as he spoke. Finally he pressed the red button, successfully

activating the toy. The other puppet chose the inconsistent toy, saying, "Hmm. Well, the red button made this [original] toy go, but this black button makes this toy go," pointing to the two relevant buttons as he spoke. Finally he pressed the correct black button, successfully activating the toy.

After each puppet demonstrated one of the toys, the experimenter asked the test question: "[Child name], remember how I told you at the beginning of the game that only one of my friends snuck out from underneath the table, and played with all the toys? Can you tell me, which one of my friends snuck out and played with all the toys?" Participants were then asked to explain their answer. The memory check questions (pre-registered as inclusion questions) were asked last, with subjects asked to match each puppet to the toy he had demonstrated: "[Child name], can you remind me, which friend showed us how to make this toy go [both puppets point to a toy]? And which friend showed us how to make this toy go [both puppets point to the other toy]?"

Puppets always demonstrated the toy they were standing closest to. This was to avoid pragmatic concerns that could arise if puppets undertook a cost to demonstrate a particular toy. Therefore, the puppet on the experimenter's left hand demonstrated the leftmost toy, and vice versa. The identity of the puppet whose turn was first, and the toy this agent acted on was always counterbalanced. Additionally, the side each puppet was presented on (left/right) was randomized.

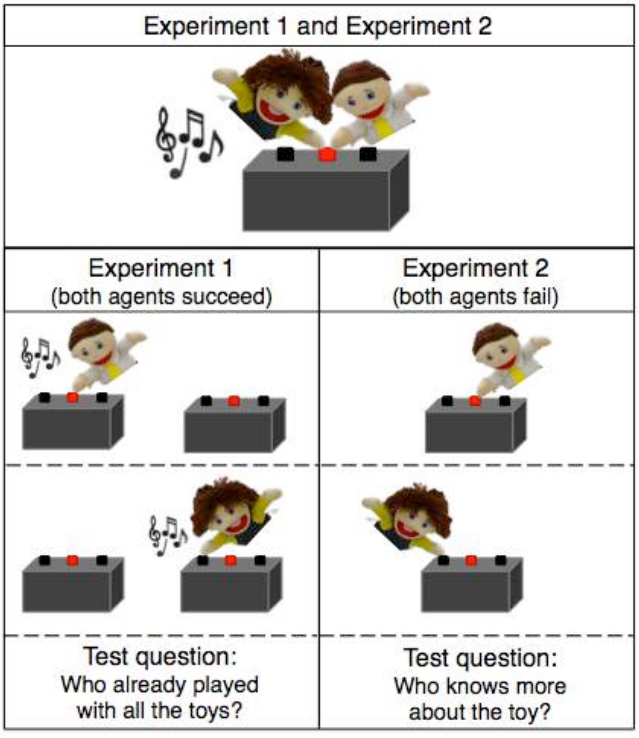


Figure 1: Procedure of both experiments. In Experiment 1 both puppets succeeded in activating the toy. In Experiment 2, both failed. Crucially, one agent's actions were always consistent with his prior experience (pressing the red button); the other agent's were not (he pressed one of the black buttons).

Results and Discussion

Two coders who were not involved in data collection determined exclusions. The first coder determined whether the experiment had been run correctly, blind to children's final answers. The second coder coded only children's answers, unaware of each puppet's role (that is, whether he demonstrated the consistent or inconsistent toy). 22 participants were excluded and replaced (see Participants).

Overall, of 58.3% of children judged that the agent who pressed the black button (and acted inconsistently with his prior experience) was more likely to have had additional knowledge. This proportion is not reliably different from chance (42 of 72; 95% CI: 47.2 – 69.4). However, a logistic regression predicting performance based on age revealed a significant age difference ($\beta = 0.87$, $p = .006$). While only 37.5% of four-year-olds judged that the agent who activated the inconsistent toy had prior knowledge (9 of 24; 95% CI: 16.67 - 58.33), 66.6% of five-year-olds (16 of 24; 95% CI: 50 – 87.5) and 70.8% of six-year-olds (17 of 24; 95% CI: 54.17 - 87.5) selected this agent. And consistent with five- and six-year-olds' success, a logistic regression predicting performance based on age also predicts that children will be more likely to answer the test question correctly (as opposed to incorrectly) by 5.04 years of age.

These results suggest that children do not simply expect ignorant agents to act successfully or unsuccessfully. Rather, by age five, children seem to expect ignorant agents to act reasonably, applying their prior knowledge in novel situations. This is consistent with prior findings that children do not think ignorance means having a false belief (Friedman & Petrashek, 2009; Jara-Ettinger, Floyd, Tenenbaum, & Schulz, 2017). If children assumed that ignorant agents should fail due to a false belief, then participants should have judged that both agents were equally knowledgeable (since both were successful). Our results suggest that by age five, children make principled belief inferences from agents' behavior. Specifically, children expect both knowledgeable and ignorant agents to act consistently with their prior knowledge, and they use these expectations to infer what other people know.

Note, however, that children were only ever taught how the training toy worked. If children (reasonably) assumed all the toys worked in the same way, they may have been surprised to see a puppet activate the inconsistent toy. Perhaps children attributed greater knowledge to this agent not because his actions were inconsistent with his prior knowledge, but because the actions (and their outcome) were inconsistent with children's own beliefs. In other words, children might simply attribute knowledge to agents who teach them something new, or show them something unexpected. We test this possibility in Experiment 2.

Experiment 2

Participants in Experiment 1 learned only how the first (training) toy worked. If participants attributed greater knowledge to the inconsistent actor because he taught them something new or unexpected, teaching children how all the toys work should cause performance to fall to chance because, now, neither agent can provide any novel information.

To address this, Experiment 2 differs in three substantial ways. First, we taught participants how all the toys worked. To reduce concerns about memory load, we used only two machines in this task: the training toy, and the inconsistent toy. Second, when trying to activate the novel toy, both puppets failed. One puppet pressed the red button (consistent with his prior experience), and one pressed the incorrect black button (inconsistent with his prior experience). Finally, we emphasized throughout that one of the puppets knew more, but not all, about the toy, making it plausible that both puppets could fail. Together, these changes allow us to test whether children attribute greater prior knowledge to agents whose actions are not explained by their prior experience, even when the agent fails to achieve their goal.

Method

Participants 72 four-, five- and six-year-olds (mean age: 5.56 years, range: 3.99 – 6.92 years; n = 24 participants per age group) were recruited at a local children's museum. 26 participants were excluded from analyses and replaced because: they did not pass the pre-registered inclusion questions (n = 13), due to experimenter error (n = 5), interruptions or interference with the procedure (n = 3), because the participant did not answer the test question within 30s (n = 3), because the participant had already participated in the past (n = 1), or due to developmental delays (n = 1).

Stimuli Materials were identical to those of Experiment 1, except that now only two machines were used: the training toy, and the inconsistent toy.

Procedure Experiment 2 began identically to Experiment 1. Participants and then puppets were familiarized with the training toy. Next, after placing the puppets underneath the table, the experimenter produced the additional (inconsistent) toy. In contrast to Experiment 1, the experimenter told participants that this toy was "a little bit different." She explained that the red button did not activate this toy, and that only one of the black buttons (the correct black button) made the toy play music. She demonstrated all of the buttons, and then allowed the participant to press each button. Thus, participants were explicitly taught how the toys worked, and experienced for themselves that the toys worked differently.

Next, both puppets returned. The experimenter explained that one of the puppets had seen the toy before, and knew a little bit about it. And she explained that the other puppet had never seen the toy before. The experimenter noted that one of the puppets knew more about the toy, but she didn't know which one. The participant's task was to help the experimenter identify which puppet knew more about the toy.

Each puppet was asked to make the toy go in turn. During each puppet's turn, the other agent was placed underneath the table. One puppet's actions were consistent with his prior knowledge, saying, "Hmm. Well, the red button made this [original] toy go, so the red button makes this toy go too," pointing to the two relevant buttons as he spoke. He pressed the red button. The button did not activate the toy, and the puppet exclaimed "oh!" in surprise when nothing happened. The other puppet's actions were inconsistent with his prior knowledge, saying, "Hmm. Well, the red button made this [original] toy go, but this black button makes this toy go," pointing to the two relevant buttons as he spoke. He pressed the incorrect black button. The button also did not activate the toy, and the puppet exclaimed "oh!" in surprise when nothing happened.

After each puppet pressed a button, the experimenter asked the test question: "[Child name], remember how I told you that one of my friends knows more about this toy? Can you tell me, which friend knows more?" Participants were asked to explain their answer. The inclusion questions were asked last, with children asked to match each puppet to the button he had pressed on the novel (inconsistent) toy: "[Child name], can you remind me, which one of my friends pressed this button [both puppets point to one button]? And which one of my friends pressed this button [both puppets point to the other button]?"

The identity of the puppet whose turn was first, and the button this agent pressed was always counterbalanced. Additionally, the side each puppet was presented on (left/right) was randomized.

Results and Discussion

Results were coded as in Experiment 1, with 26 participants excluded and replaced (see Participants). Overall, 61.1% of participants attributed knowledge to the puppet who pressed the black button, a proportion reliably higher than chance (44 of 72; 95% CI: 50 - 72.2). A logistic regression predicting performance based on age did not reveal a significant age difference ($\beta = 0.42$, $p = .14$). But while participants in all age groups preferred to attribute knowledge to the agent whose actions were inconsistent with his prior experience, only six-year-olds' preferences were robust. While 70.8% of six-year-olds judged that the agent who pressed the black button was more knowledgeable (17 of 24; 95% CI: 54.17 - 87.5), only 54% of four-year-olds (13 of 24; 95% CI: 33.33 - 75) and 58% of five-year-olds (14 of 24; 95% CI: 37.5 - 79.17) also made this judgment. In sum, although no age difference was obtained, only six-year-olds reliably judged that the agent whose failure was inconsistent with his prior experience had greater knowledge.

These findings suggest that children do not simply attribute knowledge to agents who show them something new. If they did, they should have performed at chance, as neither puppet taught children anything new. Instead, our results suggest that, by age six, children not only expect ignorant agents to act based on their prior knowledge, but also understand that knowledge runs along a continuum: agents can know more or less about any given topic. Thus, by age six, children attribute more knowledge to agents whose prior experience cannot explain their actions, even when these actions fail to fulfill their goal.

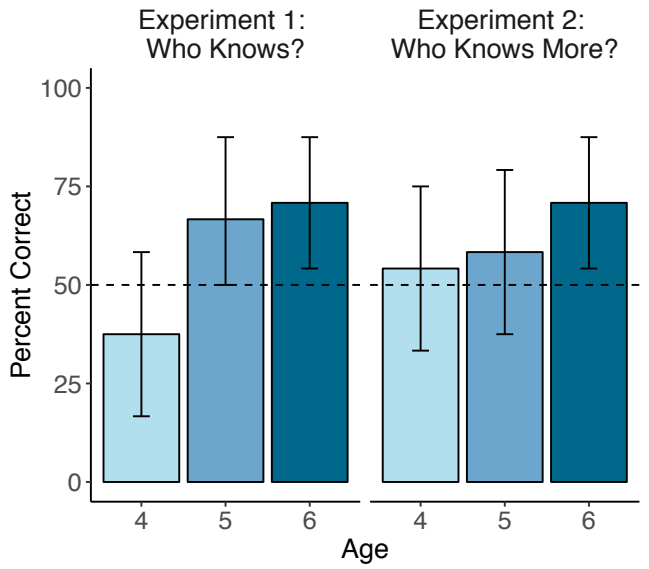


Figure 2: Results from both experiments. The error bars are bootstrapped 95% confidence intervals, and the dotted line indicates chance performance (50%). In Experiment 1, five- and six-year-olds judged that an agent whose success could not be explained by his prior experience had additional knowledge. In Experiment 2, six-year-olds judged that an agent whose failure could not be explained by his prior experience had additional (albeit incomplete) knowledge.

General Discussion

To successfully interact with others, we must understand what they know and believe, what they feel, and what they want. Children understand the link between mind and behavior early in life, inferring goals (Csibra, Gergely, Bíró, Koós & Brockbank, 1999; Jara-Ettinger, Gweon, Schulz & Tenebaum, 2016), beliefs (Onishi & Baillargeon, 2005; Rubio-Fernandez & Geurts, 2013) and desires (Doan, Denison, Lucas & Gopnik, 2015; Repacholi & Gopnik, 1997) from others' actions. Yet, while much work has shown that even young children have expectation about how knowledgeable agents should behave (Surian, Caldi & Sperber, 2007), less work has investigated whether children understand how ignorant agents might apply their prior knowledge to new situations.

Here we found that preschoolers expect ignorant agents to act based on their prior knowledge. When agents' past experience cannot explain their actions, children infer that these agents must have additional knowledge. In Experiment 1, five- and six-year-olds (but not four-year-olds) judged that an agent whose observable past experience could not explain his successful actions must've had additional knowledge. In Experiment 2, four- to six-year-olds (but only six-year-olds reliably) judged that an agent whose observable past experience could not explain his failure must've had some (incomplete) additional knowledge.

Our results show that, by age five, children expect past experiences to shape agents' beliefs and guide their actions in new situations. These results are consistent with related work, which suggests that children do not reliably link ignorance to specific outcomes (Friedman & Petrashek, 2009; German & Leslie, 2001; Ruffman, 1996).

These findings also suggest several broader implications. First, while we often talk about "knowing" or "not knowing," knowledge is not binary. People are rarely completely ignorant or completely knowledgeable. More frequently, knowledge lies along a continuum. In Experiment 2, six-year-olds succeeded in identifying which of two agents knew more, even when both agents were wrong. If children believe that agents can only be fully knowledgeable or fully ignorant, they may not have attributed even partial knowledge in this case (perhaps judging that any agent who is wrong is equally ignorant). The results of this experiment suggest that, by age six, children represent knowledge and ignorance as two poles of the epistemic continuum, leveraging their expectations about how prior experience should affect agents' actions to infer the extent of their knowledge.

Second, these findings provide insight into the development of children's epistemic inferences. While prior work has thoroughly investigated young children's ability to represent others' beliefs (e.g., Onishi & Baillargeon, 2005; Wellman et al., 2001), less research has investigated how children infer belief from action. In our tasks, children had to infer agents' beliefs from their actions. This required understanding that each agent pressed the button they believed would make the toy go, and considering what role their past experiences played in shaping these beliefs. Past work suggests that children infer knowledge from action via a naïve theory of knowledge: a set of expectations about how ignorant and/or knowledgeable agents should act (Aboody, Huey, & Jara-Ettinger, 2018). Our results are consistent with this account, demonstrating that across varied contexts, children can infer what others know or believe by observing their actions.

Our results also open avenues for future work. First, Experiment 2 shows that children do not simply attribute knowledge to agents who show them something new or surprising. However, other simple rules may explain participants' performance. For example, children may expect ignorant agents to act the same way they've acted in the past, without representing their knowledge or beliefs. In our studies, specifically, children may have solved the task by matching agents' current actions to their prior acts,

licensing knowledge any time these acts were inconsistent. Future work can address this possibility by providing agents with knowledge, but not experience (e.g., by telling the puppets in Experiment 1 how the toy works but not allowing them to try it for themselves).

A second possibility is that children expect ignorant agents to try whatever is most reasonable, not in the context of agents' knowledge, but in the context of what children themselves think is reasonable. For example, children in our task could have assumed that the red button was the most obvious thing to try (regardless of agents' past experiences), and attributed prior knowledge to any agent who rejected this obvious solution. While it is unclear whether children in fact find the red button to be the obvious solution in this task, future work can address this possibility by reversing Experiment 1, and introducing children to a training toy that works the same as the inconsistent toy. If children now attribute greater knowledge to the agent who presses the (more visually salient) red button, this would show that children do not just think that ignorance means trying the most perceptually obvious answer.

Third, in both experiments, puppets' actions differed, but so did their explanations of their actions. Namely, one agent said: "Hmm. Well, the red button made this [original] toy go, **so** the red button makes this toy go **too**," and the other said, "Hmm. Well, the red button made this [original] toy go, **but** this black button makes this toy go." Although only two words differed between explanations, it is possible that this could explain children's epistemic attributions in our task. Note, however, that this would be consistent with our account, showing that children attribute knowledge to those who explicitly reject past experience. In addition, if the linguistic cue guides children's inferences, this would be interesting in its own right—the difference between "so too" and "but" is subtle, and to our knowledge, little work has investigated how such words affect children's belief inferences. To identify whether these explanations were critical to children's inferences, future work will leave them out. If children make the same judgments, this would provide evidence that performance in this task did not hinge upon puppets' explanations.

Fourth, in Experiment 2, it is possible that children did not think both puppets were equally wrong. Conceptually, the puppet who pressed the black button may have been closer to being right (since he knew that one of the black buttons made the toy go). It is possible that children didn't consider whether agents' prior knowledge explained their actions, and instead simply attributed greater knowledge to the agent who was closer to being correct. While possible, this account does not explain children's success in Experiment 1. Furthermore, it is unclear how to operationalize what it means to be "closer" to being right in Experiment 2: while one agent was conceptually closer (pressing a black button), the other was physically closer (pressing the red button, which was right next to the correct black button). It is unclear how the magnitude of agents' errors may have guided children's inferences in the current task, but future work should investigate how this factor affects children's epistemic judgments.

Last, across both experiments, children's preferences strengthened with age (significantly in Experiment 1, and non-significantly in Experiment 2). Four-year-olds' failures in both experiments are consistent with prior work, which suggests that the ability to infer knowledge from behavior continues to develop between the ages of four and five (Aboody, Huey & Jara-Ettinger, 2018). But while five-year-olds succeeded in Experiment 1, they were not reliably above chance in Experiment 2. Why might this be?

One possibility is that identifying a completely knowledgeable agent (Experiment 1) is easier than judging which agent has greater (but still incomplete) knowledge (Experiment 2). Furthermore, given that children may equate accuracy with knowledge (Brosseau-Liard & Birch, 2010; Ronfard & Corriveau, 2016), it might be harder for them to attribute knowledge in the face of a failure.

It is also possible that five-year-olds do attribute knowledge based on a rule (for example, attributing knowledge to agents who act in a surprising way). This could explain their weaker performance in Experiment 2, although it is unclear why four-year-olds would not have followed the same rule (which would have led to success in Experiment 1). It is possible that four-year-olds have no rule for inferring belief from knowledge, five-year-olds depend on a rule (e.g., knowledge = rejecting the obvious), and six-year-olds have a deeper understanding of how prior knowledge shapes beliefs. Finally, it is always possible that task demands affected children's performance, although this would fail to explain the difference in five-year-olds' performance across the two studies. Future work will address these possibilities to further clarify how children's epistemic intuitions emerge and develop.

In sum, across two experiments, we find evidence that young children have expectations for how prior knowledge is likely to shape people's beliefs and guide their behavior. We find that children use these expectations to infer what others know (or don't know) from their actions and that, by age five, children do not expect ignorant agents to act as blank slates; rather, they expect ignorant agents to leverage relevant prior knowledge when planning their actions. Altogether, our findings suggest that even young children may understand how ignorance begets belief and action.

Acknowledgments

We thank the Boston Children's Museum, the Peabody Natural History Museum, and the families who participated in this research. We thank Sarah Wong and Ivana Bozic for help with coding, and Lindsay Stoner for help with coding and data collection. This material is based upon work supported by the Center for Brains, Minds, and Machines (CBMM), funded by NSF-STC award CCF-1231216.

References

Aboody, R., Huey, H., & Jara-Ettinger, J., (2018). Success does not imply knowledge: Preschoolers believe that

accurate predictions reveal prior knowledge, but accurate observations do not. *Proceedings of the 40th Annual Conference of the Cognitive Science Society*.

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, *1*(4), 0064.

Bartsch, K., & Wellman, H. M. (1995). *Children talk about the mind*. Oxford university press.

Brosseau-Liard, P. E., & Birch, S. A. (2010). 'I bet you know more and are nicer too!': what children infer from others' accuracy. *Developmental Science*, *13*(5), 772-778.

Csibra, G., Gergely, G., Bíró, S., Koós, O., & Brockbank, M. (1999). Goal attribution without agency cues: the perception of 'pure reason' in infancy. *Cognition*, *72*(3), 237-267.

Dennett, D. C. (1987). *The intentional stance.* Cambridge, MA: MIT Press

Doan, T., Denison, S., Lucas, C., & Gopnik, A. (2015). Learning to reason about desires: An infant training study. *Proceedings of the 37th Annual Conference of the Cognitive Science Society*.

Friedman, O., & Petrashek, A. R. (2009). Children do not follow the rule "ignorance means getting it wrong". *Journal of Experimental Child Psychology*, *102*(1), 114-121.

German, T. P., & Leslie, A. M. (2001). Children's inferences from 'knowing' to 'pretending 'and 'believing'. *British Journal of Developmental Psychology*, *19*(1), 59-83.

Gopnik, A., & Wellman, H. M. (1992). Why the child's theory of mind really is a theory. *Mind & Language*, *7*(1‐2), 145-171.

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in Cognitive Sciences*, *20*(8), 589-604.

Jara-Ettinger, J., Floyd, S., Tenenbaum, J. B., & Schulz, L. E. (2017). Children understand that agents maximize expected utilities. *Journal of Experimental Psychology: General*, *146*(11), 1574.

Liu, S., Ullman, T. D., Tenenbaum, J. B., & Spelke, E. S. (2017). Ten-month-old infants infer the value of goals from the costs of actions. *Science*, *358*(6366), 1038-1041.

O'Neill, D. K. (1996). Two‐year‐old children's sensitivity to a parent's knowledge state when making requests. *Child Development*, *67*(2), 659-677.

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, *308*(5719), 255-258.

Repacholi, B. M., & Gopnik, A. (1997). Early reasoning about desires: evidence from 14-and 18-month-olds. *Developmental Psychology*, *33*(1), 12.

Ronfard, S., & Corriveau, K. H. (2016). Teaching and preschoolers' ability to infer knowledge from mistakes. *Journal of Experimental Child Psychology*, *150*, 87-98.

Richardson, H., Lisandrelli, G., Riobueno-Naylor, A., & Saxe, R. (2018). Development of the social brain from age three to twelve years. *Nature Communications*, *9*(1), 1027.

Rubio-Fernández, P. (2019). Memory and inferential processes in false-belief tasks: An investigation of the unexpected-contents paradigm. *Journal of Experimental Child Psychology*, *177*, 297-312.

Rubio-Fernández, P., & Geurts, B. (2013). How to pass the false-belief task before your fourth birthday. *Psychological Science*, *24*(1), 27-33.

Ruffman, T. (1996). Do children understand the mind by means of simulation or a theory? Evidence from their understanding of inference. *Mind & Language*, *11*(4), 388-414.

Saxe, R. (2005). Against simulation: the argument from error. *Trends in Cognitive Sciences*, *9*(4), 174-179.

Surian, L., Caldi, S., & Sperber, D. (2007). Attribution of beliefs by 13-month-old infants. *Psychological Science*, *18*(7), 580-586.

Wellman, H. M., Cross, D., & Watson, J. (2001) Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development*, *72*(3), 655-684.

Woodward, A. L. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition*, *69*(1), 1-34.

Mathematics Skills and Executive Functions Following Preterm Birth: A Longitudinal Study of 5- to 7-Year Old Children

Julia Anna Adrian[1], Frank Haist[2], & Natacha Akshoomoff[2]
Department of Cognitive Science[1], Psychiatry[2], UC San Diego
9500 Gilman Drive, La Jolla, CA 92093

Abstract

Early mathematics skills are an important predictor of later academic, economic and personal success. Children born preterm, about 10% of the US population, have an increased risk of deficits in mathematics. These deficits may be related to lower levels of executive functions and processing speed. We investigated the development of mathematics skills, working memory, inhibitory control and processing speed of healthy children born very preterm (between 25 and 32 weeks gestational age, n=51) and full-term (n=29). Children were tested annually from ages 5 to 7 years. We found persistent lower overall mathematics skills in the preterm group, driven by differences in more informal skills (e.g. counting) at earlier time points, and by differences in more formal skills (e.g. calculation) at later time points. We did not find significant differences between preterm and full-term born children in spatial working memory capacity or processing speed. However, these cognitive measures were significant predictors of mathematics skills in the preterm but not the full-term group, hinting towards the use of different strategies when solving problems.

Keywords: Early Mathematics; Executive Functions; Cognitive Development; Preterm Birth; longitudinal;

Introduction

Mathematics skills are beneficial for success in life. Early mathematics skills at school-entry are predictive of later academic achievement (Duncan et al., 2007; Geary, Hoard, Nugent, & Bailey, 2013), and socioeconomic status (Ritchie & Bates, 2013).

Very preterm birth (before 33 weeks of gestation) has a negative effect on academic achievement in general (Johnson, Wolke, Hennessy, & Marlow, 2011), and mathematical ability in particular (Akshoomoff et al., 2017; Taylor, Espy, & Anderson, 2009). Every year about 15 million children are born preterm, with preterm birth rates ranging from 5-18% (Liu et al., 2016). Recent studies on the effects of preterm birth often examine children born extremely preterm (<28 weeks) and/or with very low birth weight (<1500g). However, these individuals make up a small proportion of the preterm born population. Furthermore, with medical advances, severe complications of preterm birth can be treated, and the rates of preterm born children without severe neurodevelopmental disorders have increased. Yet even in otherwise healthy children, preterm birth is associated with long-term cognitive consequences such as developmental and learning problems (Anderson, 2014).

Mathematical ability is related to executive functions and processing speed in typically developing children (Geary, 2011; Purpura, Schmitt, & Ganley, 2017). The core executive functions are working memory, inhibitory control and shifting (Miyake, Friedman, Emerson, Witzki, & Howerter, 2000). These cognitive skills are affected by preterm birth (Aarnoudse-Moens, Duivenvoorden, Weisglas-Kuperus, Van Goudoever, & Oosterlaan, 2012) and are likely to be related to mathematics deficits. Rose, Feldman, and Jankowski (2011) showed that differences in math and reading skills of 11 year old full-term and preterm children can be explained through preterm deficits in executive function and processing speed. They argue for a cascade of effects: preterm birth leading to lower processing speed, leading to lower executive functions, leading to lower math and reading scores. It is unknown when this cascade of effects begins and how processing speed, executive functions, and mathematics achievement are connected during early childhood.

While we know that preterm birth affects processing speed and executive functions, and that children born preterm exhibit deficits in mathematics achievement, to our knowledge no study has investigated these components longitudinally in preterm born children during childhood.

The Present Study

The present study examines how mathematics ability develops and how it is related to other neuropsychological functions following preterm birth. The children in this study were born between 25 and 32 weeks gestational age. They are considered healthy and do not suffer from any severe medical conditions or neurodevelopmental disorders. However, they make up about 2% of the general population in the US and are thus an important group to study. This longitudinal comprehensive study allows controlling for individual differences and will give insight into the development of the interplay of processing speed, executive functions, motor skills and mathematics ability.

Methods

Participants

Participants were preterm and full-term children who were tested at three time points, each about a year apart. First testing was performed within six months of starting kindergarten, at a mean age of 5.3 years (SD: 0.38). Mean age for the following two time points was 6.4 (SD: 0.37) and 7.3 (SD: 0.35), respectively. A total of 51 preterm and 29 full-term children completed the mathematics, working memory, and inhibition tests at all three time points. Not all of these children completed the other cognitive and behavioral measures; sample size for each of the subtests is stated below.

The preterm sample was primarily recruited from the UC San Diego High-Risk Infant Follow Up Clinic. Inclusion criteria was gestational age at birth of <33 weeks. Out of the 51 children in the preterm group, 10 children were born <28 weeks gestational age, and 41 children between 28 and 32 weeks gestational age. We did not find a correlation between gestational age at birth and mathematics performance within the preterm group, therefore the children were not further divided into subgroups. In the following the term preterm includes both the extremely and very preterm born children of this study.

Exclusion criteria from the preterm sample were a history of severe brain injury (e.g., cystic periventricular leukomalacia), disability (e.g., bilateral deafness or blindness), genetic abnormalities likely to affect development, and acquired neurological disorder unrelated to preterm birth.

Inclusion criteria for the full-term sample was gestational age at birth of >38 weeks and no history of neurological, psychiatric, or developmental disorders. All participants had a score of 80 or higher on the Verbal Comprehension Index of the Wechsler Preschool and Primary Scale of Intelligence (WPPSI-IV) to insure comprehension of tasks (Wechsler, 2012).

Participant characteristics are summarized in Table 1. A social economic status (SES) score was calculated as the sum of rank in maternal education and household income. Maternal education was ranked as 1: high school, 2: 1-3 years of college, 3: four year college, 4: professional/ post-graduate degree. Household income was ranked as 1: less than $50,000, 2: $50,000 - $99,999, 3: $100,000 - $199,999, 4: $200,000 and above.

By definition, the preterm and full-term group differed in gestational age at birth ($F=693.86$, $p<0.0005$) and birth weight ($F=338.96$, $p<0.0005$). They were not significantly different in terms of gender composition, household income or socioeconomic status (SES) composite. Maternal education was significantly higher in full-term compared to preterm children ($F=4.74$, $p=0.033$). They did not differ in age at any testing time.

Table 1: Participants characteristics. GA: gestational age, SES: socioeconomic status

	Preterm (n=51)	Full-term (n=29)
GA at birth (weeks): mean (min-max)	29.5 (25-32)	39.7 (38-41)
Birth weight (g): mean (min-max)	1328 (680-2410)	3411 (2353-4422)
Gender (% female)	47.1	48.3
SES composite: mean (min-max)	5.0 (2-8)	5.43 (2-8)
Maternal education	2.47 (1-4)	2.93 (1-4)
High school	13.7%	10.3%
1-3 years college	35.3%	20.7%
College graduate	41.2%	34.5%
Professional	9.8%	34.5%
Household income	2.52 (1-4)	2.50 (1-4)
Less than $50,000	15.7%	3.4%
$ 50,000 - $99,999	31.4%	44.8%
$100,000 - $199,999	35.3%	44.8%
$200,000 and above	15.7%	3.4%

Cognitive and Behavioral Measures

Measures of mathematics ability, working memory, inhibitory control, processing speed, and motor skills, were examined. A number of tasks were drawn from the Cambridge Neuropsychological Testing Automated Battery (CANTAB). These tasks are well established and standardized computerized non-verbal tasks, administered on a touch screen. They are suitable for children 4 years of age and older. In addition, a more challenging task was administered that can be thought of as a composite measure of executive functions and motor skills, the Head Toes Knees Shoulders (HTKS) Task. To be able to parse apart the effect of motor function on HTKS performance, the Movement Assessment Battery for Children (MABC-2) was administered.

Mathematics Ability was assessed via the Test for Early Mathematics Ability, Third Edition (TEMA-3, Ginsburg & Baroody, 2003). It is designed for children between 3:0-8:11 years of age. It comprises up to 72 items. The items can also be broadly categorized into informal and formal mathematics, and more specifically into seven subcategories: Verbal Counting, Counting Objects, Numerical Comparison, Numeral Literacy, Set Construction, Calculation, and Number Facts (Ryoo, et al. 2015). Measure of overall performance is the TEMA-3 total (raw) score.

Spatial Working Memory was assessed via the CANTAB Spatial Working Memory (SWM) task. The participant's task is to find a token that is hidden under one of several colored boxes. Once found, the token is hidden again, but not under the same box twice. Thus the participant has to remember where the tokens were previously found. The number of trials is gradually increased up to eight boxes. Working memory is measured inversely based on the number of errors made (searching under the same box multiple times).

Inhibitory Control was assessed via the CANTAB Stop Signal Task (SST). The participant has to choose between pressing one of two buttons depending on where an arrow points. If they hear an auditory signal when the arrow appears, the participant has to withhold their response and not press the button. Performance is measured via the stop signal reaction time in the second half of the task, were poor performance is reflected in longer reaction times.

Processing Speed was assessed via the CANTAB Reaction Time Task (RTI). The participant holds a button at the bottom of the touch screen. Above the button are five circles. Once a yellow dot appears in one of the circles, the participant releases the button at the bottom and taps the circle with the dot. Performance is measured via median response time of pressing the circle in which the yellow dot appeared. 10 preterm children from the full sample did not complete the RTI, resulting in a sample size of 41 preterm and 29 full-term children.

The Head Toes Knees Shoulders (HTKS) Task can be seen as a composite measure of executive function and motor skills (Ponitz et al., 2008). It has three rounds: In the first round the participant has to touch their toes when the examiner says to touch their head and vice versa. In the second round the participant has to touch their knees whenever the examiner says to touch their shoulders and vice versa. The third round includes all four body parts and requires remapping of the previously learned instructions. This task requires working memory, as the participant has to keep in mind which body part to touch instead of the one the examiner said; inhibitory control, as the participant has to keep themselves from plainly following the instruction, shifting, as the instructions change; and motor control. The task was added while the study was already ongoing, leading to a relatively small sample of 33 preterm and 13 full-term children who were assessed at all three time points.

Motor Skills were assessed using the Movement Assessment Battery for Children-2 (MABC-2; Henderson, Sugden, & Barnett, 2007). It tests manual dexterity, aiming & catching, and balance. The total test scores were used to compare motor skills. In this analysis, the MABC-2 is used to disentangle wheather performance differences in the HTKS are due to differences in motor or executive function. Thus the MABC-2 was examined on the same sample as the HTKS. It was administered at time point 1 and 3 only.

Results

Executive Functions, Processing Speed & Motor Skills

Group differences in cognitive and behavioral measures other than mathematics ability are summarized in Table 2. Spatial working memory scores did not differ significantly between preterm and full-term children at any time point. The stop signal reaction time in the SST was significantly longer in the preterm group at time 1 and 2, and approached significance at time 3. A longer reaction time indicates more difficulty with inhibitory control. Processing speed as measured via reaction time in the RTI did not differ significantly between groups in the 5 choice version of the RTI. Performance on HTKS was significantly lower in preterm children at all three time points. Preterm children also scored significantly lower on MABC-2 at the two time points in which motor function was measured.

Correlation analysis between performance on these cognitive and behavioral tasks controlled for time point, SES, and group revealed that performance on HTKS was correlated with errors on the SWM ($r=-0.204$, $p=0.035$), SST reaction time ($r=-0.202$, $p=0.036$), and reaction time on RTI ($r=-0.313$, $p=0.001$). Importantly, HTKS and MABC-2 total scores were not correlated ($r=0.097$, $p=0.317$). Thus differences in HTKS reflect differences between participants other than motor function.

Table 2: Group differences in performance on the spatial working memory (SWM), the stop signal task (SST), the reaction time task (RTI), the Head Toes Knees Shoulders (HTKS) task, and the Movement Assessment Battery for Children-2 (MABC-2) as measured via ANOVA. ° no significant difference, ↑ sign. longer in the preterm group, ↓ sign. lower in preterm group. n.d.: no data.

Task	Time point 1		Time point 2		Time point 3	
		F (p)		F (p)		F (p)
SWM	°	0.507 (0.478)	°	3.374 (0.070)	°	0.218 (0.642)
SST	↑	4.627 (0.035)	↑	6.503 (0.013)	°	3.328 (0.072)
RTI	°	0.302 (0.585)	°	2.128 (0.149)	°	1.205 (0.276)
HTKS	↓	7.648 (0.008)	↓	8.628 (0.005)	↓	4.685 (0.008)
MABC-2	↓	14.303 (<0.0005)		n.d.	↓	24.092 (<0.0005)

Mathematics Skills

Preterm born children had lower TEMA-3 total scores compared to full-term born children on all three time points. The difference decreased from time 1 to 2, and increased again at time 3 (figure 1, table 3). Categorization of TEMA-3 test items into subcategories according to Ryoo et al. (2015) revealed that differences between preterm and full-term children did not share a common developmental pattern. Verbal Counting, a more informal skill showed group differences at age 5 and 6 that diminished at age 7 (figure 2). Similarly, other skills such as Numerical Comparison and Counting Objects showed a narrowing of the performance gap between preterm and full-term born children between time point 1 and 3. In contrast, at age 5, preterm children did not score significantly lower on items testing for Set Construction, a more formal, skill. Over time, the a deficit emerged such that preterm born children scored about 15 percent points below full-term children at age 7. Differences in Calculation skills were present at all time points. Notably, the differences show a high increase over time, with about twice the effect size at time point 3, compared to 1 and 2. Performance of both groups on Number Facts was a floor at time point 1, and at time point 2, preterm and full-term children did not score significantly differently. However, by time point 3, large differences emerged, with lower scores in the preterm sample.

In the following Verbal Counting and Calculation skills, as well as the overall TEMA-3 performance are analyzed in more detail. These skills were chosen as they exemplify the differential trajectories of informal and formal skills.

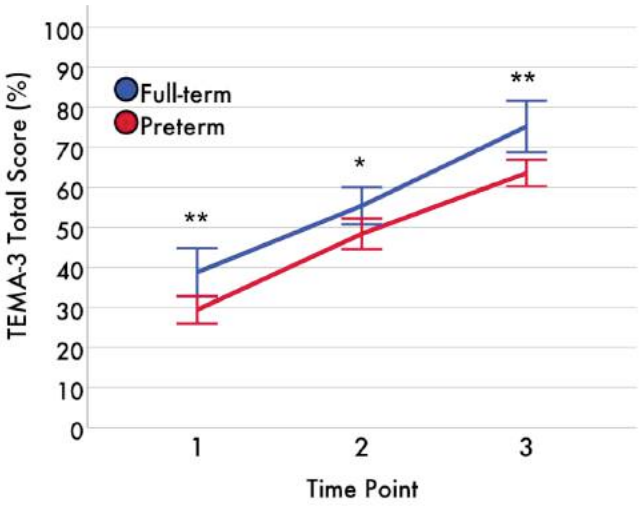


Figure 1: Standardized mathematics ability score of preterm and full-term group. The preterm group has a lower score at all three time points. * $p < 0.05$, ** $p < 0.01$.

Repeated measures ANOVAs showed a significant interaction of time and group for Verbal Counting ($F=4.234$, $p=0.016$, $\eta^2=0.051$) and Calculation ($F=7.079$, $p=0.001$, $\eta^2=0.083$). TEMA-3 total score showed a significant effect of time and group, but not their interaction ($F=1.272$, $p=0.283$, $\eta^2=0.016$). A three-way ANOVA with time and skill (Verbal Counting/Calculation) as within subject repeated measures and group as between subjects factor revealed a significant interaction between time, skill, and group ($F=11.191$, $p<0.005$, $\eta^2=0.125$).

Table 3: Summary of TEMA-3 overall performance (total score) and score of distinct skills as defined by Ryoo, et al. (2015). Scores presented as percentage of total possible score. M: mean, SD: standard deviation, group comparisons via ANOVA.

	Time point 1			Time point 2			Time point 3		
	Full-term	Preterm		Full-term	Preterm		Full-term	Preterm	
Skill	M (SD)	M (SD)	F (p)	M (SD)	M (SD)	F (p)	M (SD)	M (SD)	F (p)
Total score	38.75 (15.89)	29.36 (12.30)	8.689 (0.004)	55.36 (12.26)	48.37 (13.57)	5.265 (0.024)	75.19 (16.92)	63.56 (11.76)	13.065 (0.001)
Verbal Counting	50.49 (26.72)	34.87 (20.24)	8.693 (0.004)	75.86 (18.05)	63.03 (19.68)	8.343 (0.005)	86.95 (13.51)	83.19 (11.32)	1.763 (0.188)
Counting Objects	77.34 (19.66)	64.99 (14.37)	10.403 (0.002)	89.66 (12.01)	85.99 (12.61)	1.611 (0.208)	97.04 (5.89)	92.16 (10.81)	5.056 (0.027)
Numerical Comparison	37.55 (16.38)	27.67 (15.46)	7.234 (0.009)	51.72 (8.54)	43.33 (13.34)	9.186 (0.003)	60.92 (12.46)	54.68 (10.15)	5.904 (0.017)
Numeral Literacy	40.95 (24.07)	24.26 (16.09)	13.765 (<0.0005)	66.81 (19.27)	58.82 (18.42)	3.361 (0.071)	86.64 (14.15)	75.98 (13.66)	10.971 (0.001)
Set Construction	50.57 (13.80)	46.62 (11.76)	1.838 (0.179)	72.03 (16.96)	59.48 (13.67)	13.065 (0.001)	87.74 (14.95)	72.11 (14.80)	20.455 (<0.0005)
Calculation	23.75 (13.19)	15.90 (13.01)	6.668 (0.012)	49.04 (18.43)	37.04 (22.07)	6.136 (0.015)	88.89 (44.64)	59.48 (27.83)	13.197 (0.001)
Number Facts	Performance at floor			10.73 (15.57)	5.66 (16.24)	1.852 (0.177)	49.43 (31.23)	23.53 (23.59)	17.544 (<0.0005)

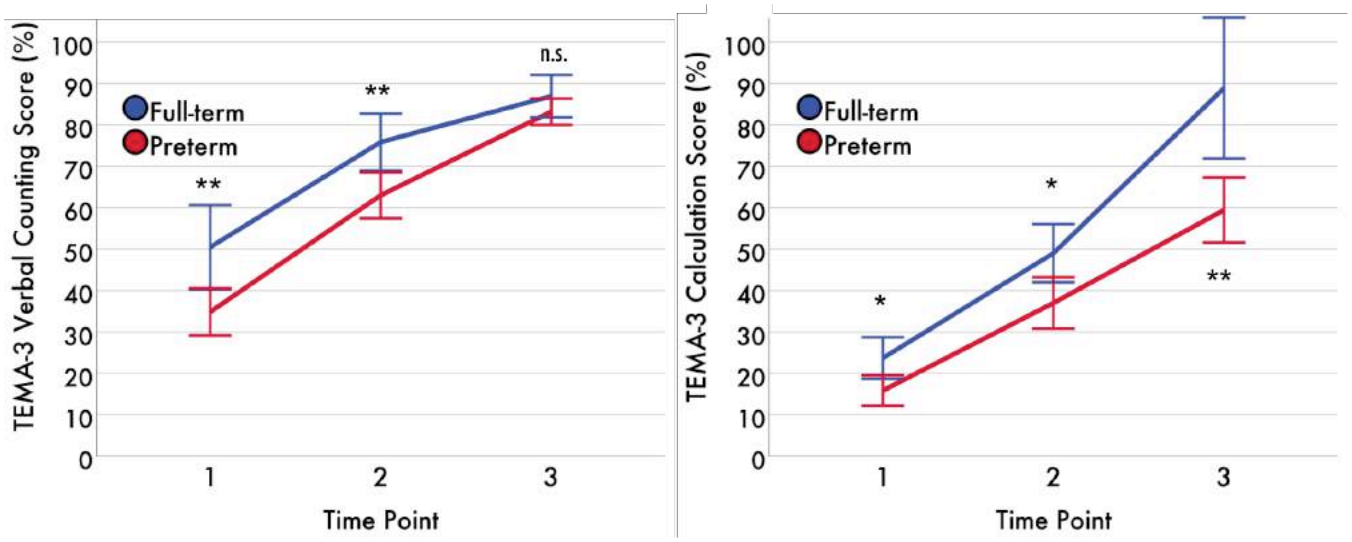


Figure 2: Development of Verbal Counting (left) and Calculation skills (right) as measured via the Test of Early Mathematics Ability – Third Edition (TEMA-3). Error bars: 95% confidence interval. * $p < 0.05$, ** $p < 0.01$, n.s.: $p > 0.05$

Effect of Executive Functions and Processing Speed on Mathematics Skills

Linear models were used to determine which of the cognitive and behavioral measures are predictive of mathematics skills. As the sample size for HTKS is considerably smaller, two separate models were evaluated: Model 1 used SWM, SST and RTI as predictors, while model 2 used HTKS score as predictor (table 4). Time point, SES and gender were included in all models. Time point and SES were significant predictors for both groups, and gender was a significant predictor in the full-term but not the preterm group.

In model 1, SWM and SST were predictive of Verbal Counting and Calculation in the preterm group. RTI was predictive of Verbal Counting but not Calculation. In contrast, SWM was not predictive in the full-term group for either skill, and SST was a significant predictor only for Verbal Counting. RTI did not significantly predict either skill.

Model 1 was a better fit when predicting Verbal Counting (r^2=0.6293) compared to Calculation skills (r^2=0.4580) in the preterm group. Conversely, the model performed slightly better when predicting Calculation (r^2=0.5743) compared to Verbal Counting (r^2=0.5261) skills in full-term children.

Model 2 showed that in both groups performance on HTKS was predictive of Verbal Counting, but not Calculation. Consequently, it was a better fit when predicting Verbal Counting compared to Calculation skills.

Table 4: Summary of linear models predicting Verbal Counting and Calculation Skills in full-term and preterm born children, respectively. Time point, SES and gender are included in all models. Model 1 additionally uses SWM, SST and RTI as predictors, model 2 uses HTKS.

	Verbal Counting		Calculation	
	Full-term	Preterm	Full-term	Preterm
Base Model – predictors: Time point, SES, gender				
Adj. r^2	0.4974	0.5852	0.5275	0.4166
Model 1 – n(preterm)=41, n(full-term)=29				
Adj. r^2	0.5261	0.6293	0.5743	0.4580
Predictors: std. β coeff. (p)				
SWM	-0.0259 (0.2084)	-0.0309 (0.0137)	-0.0376 (0.0522)	-0.0245 (0.0190)
SST	-0.0080 (0.0129)	-0.0037 (0.0309)	-0.0043 (0.1426)	-0.0035 (0.0155)
RTI	0.0003 (0.9025)	-0.0036 (0.0299)	-0.0025 (0.0844)	-0.0009 (0.5056)
Model 2 – n(preterm)=33, n(full-term)=13				
Adj. r^2	0.5549	0.6528	0.4942	0.4061
Predictors: std. β coeff. (p)				
HTKS	0.0570 (0.0316)	0.0559 (0.0008)	0.0473 (0.1607)	0.0233 (0.1351)

Discussion

The present study examined the trajectories of specific mathematics skills and overall mathematics ability in preterm and full-term born children from before starting kindergarten to the end of first grade (age 5 to 7). Consistent with previous studies, we found lower overall mathematics score in the preterm born group at all time points. This observation by itself masks the fact that the differences between preterm and full-term group are not consistent over time. Importantly, one has to distinguish between the developmental trajectory of informal skills (e.g. Verbal Counting) and formal skills (e.g. Calculation).

We found a deficit in Verbal Counting skills in preterm compared to full-term children at time point 1 and 2. However, by time point 3, there were no significant group differences. This type of developmental pattern signals an initial delay of Verbal Counting skills in preterm children, followed by catch up in skill level.

In contrast, the difference in Calculation skills between preterm and full-term group increases over time. This is likely because more formal mathematics skills are commonly introduced in kindergarten, and rapidly develop with schooling. We predict that this deficit in the preterm group will persist over time, and possibly increase further. In line with this, mathematics deficits have been found in pre-teens (Akshoomoff et al., 2017; Rose et al., 2011), and teenagers (Litt et al., 2012) born preterm. The same applies for other aspects of formal mathematics, such as Set Construction and Number Facts.

Differences in mathematics skills might be mediated to some extent by differences in other cognitive functions (Mulder, Pitchford, & Marlow, 2010). We examined measures of spatial working memory, inhibitory control, processing speed, and motor function. Interestingly, while there were no group differences in SWM, we found that it is predictive of both Verbal Counting and Calculation skill in preterm but not full-term children. Similarly, we did not find significant differences in processing speed between preterm and full-term children, and performance on RTI was predictive of Verbal Counting in the preterm group only. This may indicate that the two groups are employing different problem-solving strategies. Children born preterm may rely on different cognitive processes as they develop mathematics skills.

Preterm compared to full-term children show significantly longer reaction times at the SST, an inverse measure of inhibitory control, at time point 1 and 2. While SST is predictive of Verbal Counting in both groups, it is predictive of Calculation in the preterm group only. This may be another hint towards more effortful task completion in the preterm born children, and potentially having to recruit inhibition skills to a greater extent than full-term children.

The HTKS, a task requiring working memory, inhibitory control, shifting, and motor skills, revealed deficits of the preterm compared to the full term group at all time points. Since there was no correlation between performance on the HTKS and the MABC-2, it appears to reflect group differences in composite executive function ability that is not driven by the group differences in motor control.

Performance on the HTKS has previously been shown to be predictive of academic achievement of typically developing children in kindergarten (McClelland et al., 2014). Consistent with this, we found that HTKS scores were a significant predictor for Verbal Counting, but not Calculation for both groups. However, the number of participants who were administered the HTKS was smaller, and it remains to be examined if the results hold up with a larger sample size.

Our study is an important contribution to the existing body of literature as it examines the crucial transition from preschool through the end of first grade, capturing the first formal instruction of mathematics in school. Further follow-up of these children, and their formal mathematics skills in particular, would give valuable insight into the potential differences in developmental trajectory of preterm and full-term born children from childhood through adolescence.

It should be noted that this preterm group had no significant neonatal complications and was considered healthy. Nevertheless, we found deficits in mathematics skills, particularly in formal skills at age 7, that may be heralds of important inequalities later in life (Basten, Jaekel, Johnson, Gilmore, & Wolke, 2015). These differences are important to consider for parents and teachers of preterm born children, and for our society at large.

Acknowledgments

This study was funded by a grant from the National Institutes of Health (R01HD075765).

We are grateful to the children and families who participated in this study. Thank you to Holly Hasler, M.S., Stephanie Torres, Kelly McPherson, Akshita Taneja, and Rubaina Dang; and our collaborators: Martha Fuller, Ph.D., RN, PNP, Yvonne Vaucher, M.D., Terry Jernigan, Ph.D., Don Hagler, Ph.D., Anders Dale, Ph.D., John Hesselink, M.D., and Joan Stiles, Ph.D.

References

Aarmoudse-Moens, C. S., Duivenvoorden, H. J., Weisglas-Kuperus, N. Y. N. K. E., Van Goudoever, J. B., & Oosterlaan, J. (2012). The profile of executive function in very preterm children at 4 to 12 years. *Developmental Medicine & Child Neurology*, *54*(3), 247-253.

Anderson, P. J. (2014). Neuropsychological outcomes of children born very preterm. In *Seminars in Fetal and Neonatal Medicine* (Vol. 19, No. 2, pp. 90-96). WB Saunders.

Akshoomoff, N., Joseph, R. M., Taylor, H. G., Allred, E. N., Heeren, T., O'shea, T. M., & Kuban, K. C. (2017). Academic achievement deficits and their

neuropsychological correlates in children born extremely preterm. *Journal of Developmental & Behavioral Pediatrics*, *38*(8), 627-637.

Basten, M., Jaekel, J., Johnson, S., Gilmore, C., & Wolke, D. (2015). Preterm birth and adult wealth: mathematics skills count. *Psychological science*, *26*(10), 1608-1619.

CANTAB® [Cognitive assessment software]. Cambridge Cognition (2019). All rights reserved. www.cantab.com

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., ... & Sexton, H. (2007). School readiness and later achievement. *Developmental psychology*, *43*(6), 1428.

Geary, D. C. (2011). Cognitive predictors of achievement growth in mathematics: a 5-year longitudinal study. *Developmental psychology*, *47*(6), 1539.

Geary, D. C., Hoard, M. K., Nugent, L., & Bailey, D. H. (2013). Adolescents' functional numeracy is predicted by their school entry number system knowledge. *PloS one*, *8*(1), e54651.

Ginsburg, H., & Baroody, A. J. (2003). *TEMA-3: Test of early mathematics ability*. Pro-ed.

Henderson, S. E., Sugden, D. A., & Barnett, A. L. (2007). *Movement assessment battery for children-2*. Harcourt Assessment.

Johnson, S., Wolke, D., Hennessy, E., & Marlow, N. (2011). Educational outcomes in extremely preterm children: neuropsychological correlates and predictors of attainment. Developmental neuropsychology, 36(1), 74-95.

Litt, J. S., Gerry Taylor, H., Margevicius, S., Schluchter, M., Andreias, L., & Hack, M. (2012). Academic achievement of adolescents born with extremely low birth weight. *Acta Paediatrica*, *101*(12), 1240-1245.

Liu, L., Oza, S., Hogan, D., Chu, Y., Perin, J., Zhu, J., ... & Black, R. E. (2016). Global, regional, and national causes of under-5 mortality in 2000–15: an updated systematic analysis with implications for the Sustainable Development Goals. *The Lancet*, *388*(10063), 3027-3035.

McClelland, M. M., Cameron, C. E., Duncan, R., Bowles, R. P., Acock, A. C., Miao, A., & Pratt, M. E. (2014). Predictors of early growth in academic achievement: The head-toes-knees-shoulders task. *Frontiers in psychology*, *5*, 599.

Mulder, H., Pitchford, N. J., & Marlow, N. (2010). Processing speed and working memory underlie academic attainment in very preterm children. *Archives of Disease in Childhood-Fetal and Neonatal Edition*, *95*(4), F267-F272.

Ponitz, C. E. C., McClelland, M. M., Jewkes, A. M., Connor, C. M., Farris, C. L., & Morrison, F. J. (2008). Touch your toes! Developing a direct measure of behavioral regulation in early childhood. *Early Childhood Research Quarterly*, *23*(2), 141-158.

Purpura, D. J., Schmitt, S. A., & Ganley, C. M. (2017). Foundations of mathematics and literacy: The role of executive functioning components. *Journal of Experimental Child Psychology*, *153*, 15-34.

Ritchie, S. J., & Bates, T. C. (2013). Enduring links from childhood mathematics and reading achievement to adult socioeconomic status. *Psychological science*, *24*(7), 1301-1308.

Rose, S. A., Feldman, J. F., & Jankowski, J. J. (2011). Modeling a cascade of effects: The role of speed and executive functioning in preterm/full-term differences in academic achievement. *Developmental science*, *14*(5), 1161-1175.

Ryoo, J. H., Molfese, V. J., Brown, E. T., Karp, K. S., Welch, G. W., & Bovaird, J. A. (2015). Examining factor structures on the Test of Early Mathematics Ability—3: A longitudinal approach. *Learning and Individual Differences*, *41*, 21-29.

Taylor, H. G., Espy, K. A., & Anderson, P. J. (2009). Mathematics deficiencies in children with very low birth weight or very preterm birth. Developmental disabilities research reviews, 15(1), 52-59.

Wechsler, D. (2012). *Wechsler preschool and primary scale of intelligence—fourth edition*. San Antonio, TX: The Psychological Corporation.

Decision-Making in a Social Multi-Armed Bandit Task: Behavior, Electrophysiology and Pupillometry

Julia Anna Adrian[1], Siddharth Siddharth[2], Syed Zain Ali Baquar[1], Tzyy-Ping Jung[3], & Gedeon Deák[1]

Department of Cognitive Science[1], Electrical Engineering[2], Bioengineering[3], UC San Diego
9500 Gilman Drive, La Jolla, CA 92093

Abstract

Understanding, predicting, and learning from other people's actions are fundamental human social-cognitive skills. Little is known about how and when we consider other's actions and outcomes when making our own decisions. We developed a novel task to study social influence in decision-making: the social multi-armed bandit task. This task assesses how people learn policies for optimal choices based on their own outcomes and another player's (observed) outcomes. The majority of participants integrated information gained through observation of their partner similarly as information gained through their own actions. This lead to a suboptimal decision-making strategy. Interestingly, event-related potentials time-locked to stimulus onset qualitatively similar but the amplitudes are attenuated in the solo compared to the dyadic version. This might indicate that arousal and attention after receiving a reward are sustained when a second agent is present but not when playing alone.

Keywords: Decision-Making; Uncertainty; Multi-Armed Bandit; Social Interaction; Dyadic EEG

Introduction

For successful social interaction it is useful to represent and predict other people's actions and the consequences of those actions. Joint action is defined as the ability to coordinate one's actions with others to achieve a goal (Vesper et al., 2016). Although it occurs in many sorts of human activities, it can be conveniently studied using the social or two-player versions of standardized cognitive tasks. Such tasks modified for social interactions can reveal complex dynamic in people's use of social information for judgments and action-planning.

For example, the *joint Simon task* (Sebanz, Knoblich, & Prinz, 2003) is a modified, two-player version of the standard Simon task that measures stimulus-response compatibility. Participants learn to respond with left or right button press for visual or auditory cues and show a longer reaction times are shorter when the cue location is compatible with the response hand, than when the cue occurs contralaterally. This *Simon effect* (Simon, 1990), interestingly, remains in the two-player version in which each participant is only responsible for one stimulus-response pair (Sebanz et al., 2003). This can be interpreted as evidence that human action planning is automatic and is elicited by processing another person's actions as well as planning and executing our own actions. The propensity to develop this ability might have evolved to enable efficient social learning (Kilner, Friston, & Frith, 2007; Liao, Acar, Makeig, & Deák, 2015). Particularly under conditions of uncertainty, the capacity to observe, encode, and imitate others' actions can be beneficial (Laland, 2004), permitting a sort of vicarious embodied modeling.

However, it is not always adaptive to generalize from other's actions and outcomes to one's own. The findings from joint Simon and other tasks have shown that representation of other's internal states occurs even when it is unnecessary, or disadvantageous, for optimal task completion. To study the extent to which people use observation of other's actions and outcomes to influence their own choices, even when it is unfavorable, we developed a novel task: the *social multi-armed bandit* task. The standard multi-armed bandit is a single-player paradigm to study decision-making under uncertainty. Named after the 'one-armed bandit' slot machines of casinos that have a fixed reward probability, multi-armed bandit tasks present several different options ('arms') of different, unknown reward probabilities. They manifest a classic exploitation/ exploration problem (Cohen, McClure, & Yu; Gittins, Glazebrook, & Weber, 2011). Commonly, after an initial phase of exploration players employ one of two strategies: maximizing or matching. *Maximizing*, or consistently choosing the most-rewarding arm (based on prior observations), is the optimal strategy for problems with static reward probabilities. By contrast, *matching,* or choosing each arm in proportion to its relative reward probability, is suboptimal but nevertheless seen in humans and other animals (Sugrue, Corrado, & Newsome, 2004).

Notably, although a great deal of problem solving and prediction updating occurs in social or joint tasks, only a few studies have included multiple decision-makers in social versions of prediction tasks such as multi-armed bandit, and even these have not investigated effects of social interaction or observation on decisions (Liu & Zhao, 2010).

In addition to studying behavior, we recorded participants' electroencephalogram (EEG) and pupil size as physiological metrics of cortical and neuromodulatory concomitants of social decision-making. These bio-sensing methods may provide insights into the underlying neural dynamics of decision-making with high temporal resolution. Both of these physiological measures are common in affective computing (Partala, 2003) to measure valence and arousal, and cortical changes (Fink, 2009).

The present study

The present study aims to address a *"key question of today's cognitive science: how and to what extent do individuals mentally represent their own and others' actions, and how do these representations influence, shape, and constrain an individual's own behavior when interacting with others?"* (Dolk et al., 2014).

To do so, we converted a classic three-armed bandit paradigm into a turn taking game. Reward probabilities for the three arms were different for each of two participants, allowing us to estimate the distinct effects of their own and their partner's action and outcome history on their ongoing decision-making. We studied three outcome measures: (1) decision-making behavior, (2) event-related EEG potentials, and (3) pupil dilation. Details are described below.

In the multi-player version, the probabilities remain constant for each player, however, they differ between the two players (see Table 1). This allows us to examine to what extent each participant takes into account their own and their partner's choices and outcomes. We expected to observe two different core strategies:

Egocentric strategy: Participants might make their decisions only based on their own outcome history and ignore information from their partner's outcomes. Players using this strategy should converge on choosing their own highest gaining arm (90% reward probability) most of the time.

Joint strategy: Participants might take into account information from their partner's outcomes to the same extent as information from their own outcomes. Players using this strategy should not converge on choosing one arm, because all arms average the same reward probability if both participants' outcomes are encoded equally. Alternately, in an intermediate strategy, participants might take into account their partner's outcomes but weigh them less than their own outcomes, and then more slowly converge on their own optimal choice.

Table 1: Reward probabilities for the different arms of the social multi-armed bandit

	Arm 1	Arm 2	Arm 3
Player 1	30%	60%	90%
Player 2	90%	60%	30%

Experiment 1

Participants

Participants were 28 female undergraduate students (14 dyads) recruited through the university's SONA system. They received course credit for participation in addition to a small monetary reward based on performance in the social multi-armed bandit (0.05 USD per reward).

One pair was excluded from EEG and eye-tracking analysis due to recording failure; another pair was excluded because one player chose the same arm on every trial. This left behavioral data from 26 participants, EEG data from 12 participants, and pupillometry data from 12 participants.

Experimental Design

The (social) multi-armed bandit was described to participants as 'the ice-fishing game' and presented on a touchscreen. They were shown three 'ice holes' (arms) distinguished by shape, at approximately equal distances from each other (see Figure 2). The arms were associated with discrete and constant reward probabilities (30%, 60%, and 90%) unknown to the players. Upon choosing and touching a hole, participants heard and saw differential reward feedback. Participants had 100 trials each (200 total) to catch as many fish as possible, choosing one ice hole per trial. Each participant played the game once on their own (solo version) and once as a turn-taking game (dyadic version). For each dyad of participants, EEG and pupillometry data were recorded from one player, and behavioral data were recorded from both.

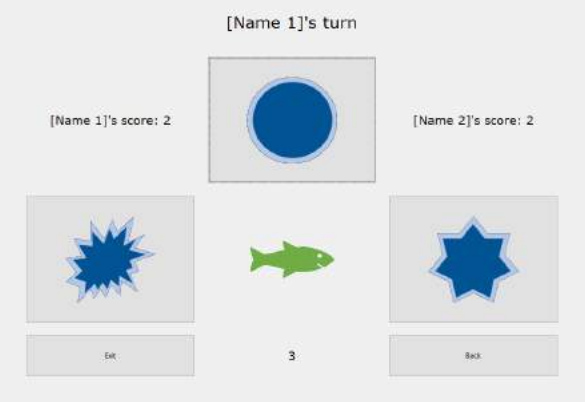


Figure 1: The game screen when a player has won a reward (green fish). The display shows the two players' accumulated rewards as well as which player's turn it is.

Figure 2: Two participants playing the social multi-armed bandit. The player on the left is wearing an Emotiv EEG headset and PupilLabs eye-tracker.

Data Acquisition

The game was presented on a table-mounted capacitive touch screen monitor (diagonal: 66cm). During the dyadic turn-taking game the participants sat facing each other (figure 2). An Emotiv headset (www.emotiv.com) recorded 14-channel EEG data, and a PupilLabs headset (pupil-labs.com/pupil/) captured pupillometry data. These sensors were chosen for participants' comfort and natural movement during a social interaction. EEG was sampled at 128 Hz, and the eye-facing camera sampled at 120 Hz. PupilLabs Software was used to detect the pupil in each frame and calculate its diameter. Lab Streaming Layer (LSL) (Kothe, 2015) was used to synchronize all of the data streams (i.e. EEG, eye-gaze video, and game events) by time stamping each event and each sample.

Synchronized EEG and pupillometry data were locked to participants' game choices in LSL-created XDF files so that behavioral and physiological data were epoched to trials. On each turn the 2 sec of data following the outcome stimulus presentation (win/loss) was used for further analysis.

Data Analysis

The first 20 trials of each game for each player were considered training trials, to teach the participant the game. These trials were not considered in the current analyses.

Decision-Making Behavior Participants' ice hole choice patterns were analyzed via Kullback-Leibler divergence (KLD), a measure of relative entropy.

$$D_{KL}(p,q) = \sum_{i=1}^{N} p(x_i) * log \frac{p(x_i)}{q(x_i)}$$

This quantifies the divergence of one probability distribution to another one. In our experiment we use the KLD to measure the difference between a participant's observed choices, and expected choices according to potential strategies. We hypothesize the employment of four different strategies with the expected choice probabilities as summarized in Table 2. For the egocentric-maximizing strategy, the player chooses the highest gaining arm (arm 3) at every trial. In the egocentric-matching strategy, each arm is chosen in proportion to their reward probability. The joint-equal strategy assumes that the outcomes of both players are weighted equally, resulting in an apparent reward probability of 60% for each arm. In that case each arm is chosen ⅓ of the trials. The joint-social strategy assumes a social value of the arm that has an equal, relatively high reward probability for both of the players (arm 2) and is thus chosen most often.

KLD of observed vs. expected probability distribution for each of the hypothesized strategies was calculated and compared to classify each participant's preferred strategy. As the joint strategy is not applicable in the single-player version, only the two egocentric strategies were compared.

Table 2: Expected choice probabilities for player 1 for each of the four hypothesized strategies. Reward probabilities for

Strategy of player 1		Arm 1	Arm 2	Arm 3
Ego-centric	maximizing	0%	0%	100%
	matching	16.7%	33.3%	50%
Joint	equal	33.3%	33.3%	33.3%
	social	0%	100%	0%

Game Data The game data was an 8 x 200 matrix which included the turn number, player number, reward state, choice, time taken, player 1 reward and player 2 reward. In the single player case the last value was set as -1 and disregarded.

EEG Data The EEG data was cleaned using EEGLAB's Artifact Subspace Reconstruction (ASR) noise removal pipeline (Delorme, 2004; Mullen, 2013). Region of interest was the occipital cortex (channel O1).

Eye-Tracking Data The current analyses only consider a single channel containing pupil diameter information. Samples with abnormally high or zero pupil diameter values (due to detection errors or eye blinks) were ignored and data was interpolated by adjacent values.

After interpolating, the data was normalized to range between 0 and 1 to account for discrepancies in pupil diameter across subjects.

Further Analysis Epochs for each trial containing the response and 2 seconds of subsequent data (including the reward outcome). Our goal was to illustrate the pupil dilation (indicating autonomic response) and cortical dynamics (focusing on updating responses) upon perceiving a reward stimulus after choosing a specific action.

Pupil and EEG data for each type of choice and reward combination were then averaged across all subjects. For EEG data, each channel was averaged independently, to facilitate Event Related Potential (ERP) analyses. The 0.2 sec of data before the event were used as baseline for the normed succeeding EEG data, to control variance in EEG amplitude across subjects.

Results

Decision-Making Behavior Over the course of the game, participants received information through trial and error and could learn that different arms were associated with different probabilities of receiving a reward. In the solo version, participants chose the highest gaining arm more often than the other two. Table 3 summarizes the decision-making behavior via mean total scores and mean number of choices for each arm. Participants distributed their choices more equally and scored lower during the dyadic game.

Table 3: Means (SD) of each decision type in the single- and multi-player games. All measures differ significantly between single and multi-player version ($p<0.001$).

	Reward probability	30%	60%	90%
No. of choices by game version	single	7 (5.0)	17 (10.3)	56 (13.4)
	multi	18 (9.1)	32 (15.1)	30 (19.5)

Mean total score in the single-player game was 76 (SD: 7.0) compared to 63 (SD: 10.1) in the multi-player game ($p<0.001$).

For each version, participants were categorized based on the strategy employed. For each strategy, KLD was calculated between the observed and the expected choices. 70% (18/26) of participants employed a maximizing strategy in the solo version. In the dyadic version, strategies were more varied (see Figure 3). Most common was the joint-social strategy, used by 32% (9/28). There was no correlation between individuals' strategies in the solo and dyadic versions.

Strategy use affected overall scores in the dyadic version ($F=6.083$, $p=0.004$) and had a marginally significant effect in the solo version ($p=0.073$).

Brain Dynamics ERP locked to outcome stimuli for each of the differentially rewarding arms were compared between three conditions: (1) *solo*: a player's responses in the solo version of the game, (2) *dyad (self)*: a player's responses to an outcome of their own action, and (3) *dyad (other)*: a player's responses to an outcome of the partner's action.

Figure 4 illustrates the findings. The ERP displayed a prominent positive potential around 300ms (P3) after dyadic self-reward events in the dyadic version, but not for partner's reward or for reward in the solo game. We also note that the ERP response for partner's rewards as well as for own-reward in the solo game is attenuated but follows the same profile as that of the self-reward condition.

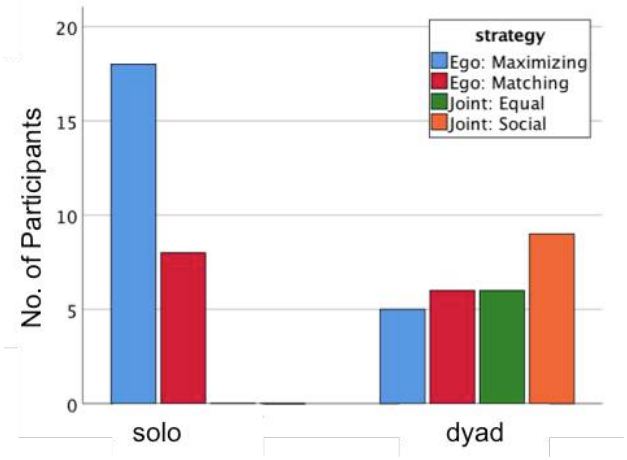


Figure 3: Distribution of strategies being employed by the participants in the single- and multi-player game, as determined via KLD.

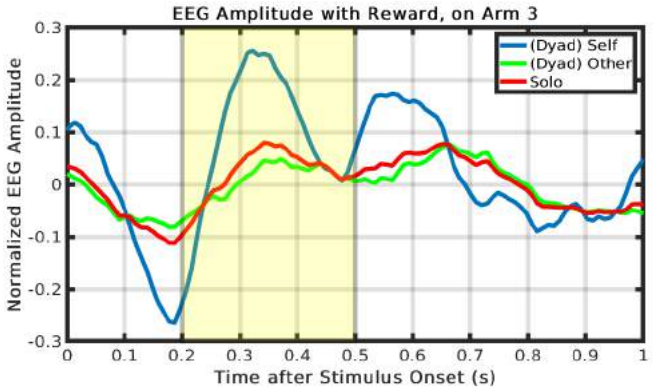


Figure 4: ERP after reward at arm 3 (90%/30%) averaged across participants from channel. Dyad (Self): player receives reward in dyadic version, Dyad (Other): player observes partner receive a reward in dyadic version, Solo: player receives a reward in solo version.

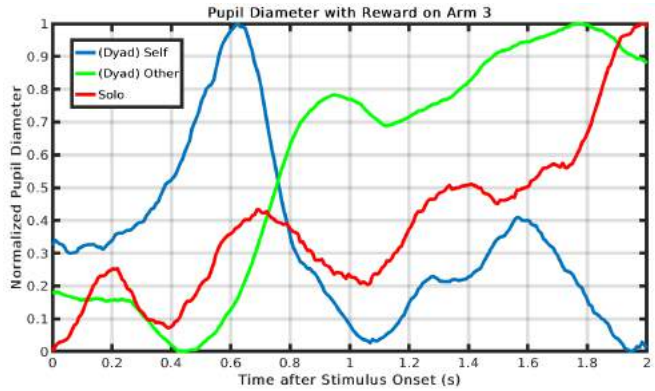


Figure 5: Pupillometry data after reward at arm 3 (90%/30%) averaged across participants from channel. Dyad (Self): player receives reward in dyadic version, Dyad (Other): player observes partner receive a reward in dyadic version, Solo: player receives a reward in solo version.

Pupillometry Pupil dilation significantly increases initially (Figure 5) when a reward is obtained for the player themselves when playing the social multi-armed bandit, but not in the single-player version and not when observing the partner receive a reward. After this initial response, we see that after 0.8 seconds of the reward onset, pupil diameter increases for the Dyad (Other) condition whereas it decreases for the self-reward one.

Discussion

The Social Multi-Armed Bandit revealed that adults take into account information from others' when making decisions. In consistency with previous studies, the majority of participants employed maximizing strategy in the solo version of the task. With this novel paradigm we found that this is not the case when there is a second player present, and when the reward probabilities for the arms are not the same for the two players. Instead, more than half of the participants employed a 'joint'-strategy, in which the actions and outcomes from the other player are integrated with their own when making decisions. One explanation for this phenomenon is a high prior belief of the same underlying probability structure for both players. This is likely the case because the visual representation remains constant throughout the game aside from updating the score count and the display whose turn it currently is. This issue is addressed in Experiment 2.

The ERP time-locked to stimulus onset showed a qualitatively similar pattern after receiving a reward at arm 3 (90%/30% reward probability) for the three conditions analyzed. However, the amplitude is highest when receiving a reward in the dyadic version, and attenuated when observing the partner receive a reward. When receiving a reward in the solo version of the game, the amplitude is also attenuated in comparison to receive a reward in the presence of a second player. We believe that this is due to the higher stakes and reward scenario attached with the dyadic version of the game.

Interestingly, pupil dilation increases drastically at about 0.6 seconds after stimulus onset when receiving a reward in the dyadic version of the game. In contrast, pupil dilation after observing the partner receive a reward has a longer latency of about 0.8 second. This likely reflects differential activation of the parasympathetic nervous system (PNS) for self/other reward scenarios.

Experiment 2

Participants

Participants were 32 undergraduate students (16 dyads, 10 female-female, 4female-male, 2 male-male) recruited through the university's SONA system. They received course credit for participation in addition to a small monetary reward based on performance in the social multi-armed bandit (0.05 USD per reward).

Experimental Design

The experimental design of Experiment 2, is very similar to experiment 1. The modifications of the experiment are:

(1) Whereas in experiment 1, one person in each dyad played the solo version of the game before the dyadic version and the other person played in the reverse order, in experiment 2 both played the solo version either before or after the dyadic version. In other words, game order was randomly assigned by dyad. This ensured that the game process was not driven by prior knowledge of only one of the players.

(2) To reduce the prior belief of a constant underlying reward structure of the game, we changed the background color of the game after each turn, such that there was a distinct visual cue to signal each player's turns.

(3) EEG data was recorded from both participants in Experiment 2 (vs. only one participant per dyad in Exp. 1). Pupillometry data was not recorded.

Data Acquisition

See Experiment 1.

Data Analysis

As in Experiment 1, the first 20 trials were excluded from analysis.

Decision-making behavior

The analysis performed in Experiment 1 is based upon the assumption that participants make use of particular strategies. In this experiment, a different type of analysis was performed, considering choice behavior 'bottom-up' without assumptions of specific strategies.

Participants' choices were analyzed via the Jensen-Shannon Divergence (JSD). The JSD is a distance metric between two probability distributions and based on the KLD:

$$D_{JS}(p,q) = \frac{1}{2}D_{KL}(p,x) + \frac{1}{2}D_{KL}(q,x)$$

with $$x = (p+q)/2$$

The JSD between the relative choice distribution of the last 80 trials of the participants' empirical behavior and the relative choice distribution if all choices were made towards the highest gaining arms (=(0,0,1), maximizing strategy) was used to analyze the data. Hence, the decision-making behavior for each participant could be characterized by their JS divergence in the solo and the dyadic version. As reference, the JS divergence of relative choice distribution between matching behavior (0.17, 0.33, 0.5) and maximizing (0,0,1) is 0.31. This value was used to further cluster participants into 'learners' (JSD < 0.31) and 'non-learners' (JSD ≥ 0.31).

Results

Decision-Making Behavior Participants could be categorized into four groups, depending on if they learned which option was the highest gaining in the solo and/or dyadic version of the game. 50% (16/32) of participants were grouped into 'learners' in the solo version, but into 'non-learners' in the dyadic version. As shown in Figure 6, in the solo version, they choose the 90% reward arm significantly more often than the other two arms ($F = 179.1$, $p < 0.0005$) and significantly more often than in the dyadic version ($F = 59.7$, $p < 0.0005$). 22% (7/32) of participants were clustered into 'learners' in both versions of the game, and 19% (6/32) were clustered into 'non-learners' in both versions of the game. 9% (3/32) were clustered into 'learners' in the dyadic version of the game, but not in the solo version.

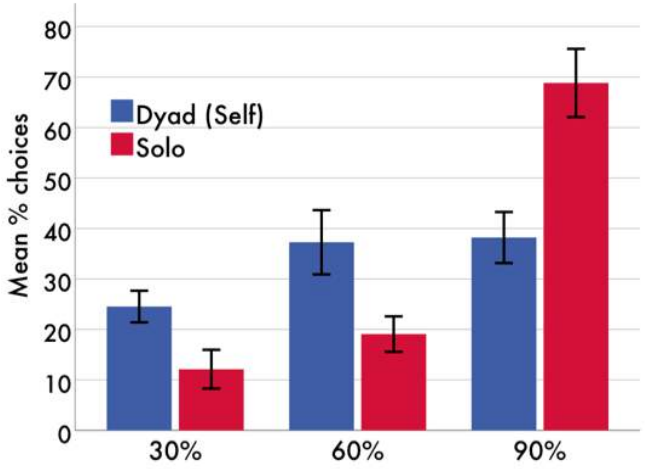


Figure 6: Choice behavior of 50% of participants who learned which arm has the highest reward probability in the solo but not the dyadic version of the multi-armed bandit.

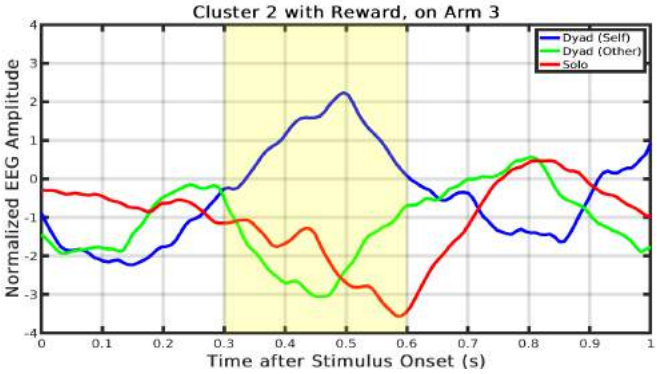


Figure 7: ERP after reward at arm 3 (90%/30%) averaged across participants from channel. Dyad (Self): player receives reward in dyadic version, Dyad (Other): player observes partner receive a reward in dyadic version, Solo: player receives a reward in solo version.

Brain Dynamics ERPs were examined at channel O1 of participants who chose the highest gaining arm most of the time in the solo version, but not in the dyadic version (Figure 7). We observed a high increase in amplitude for the Dyad (Self) condition. In comparison, there was a negative deflection in ERP for the Dyad (Other) and Solo condition.

Discussion

Even when the prior belief of a common underlying reward structure was decreased, half of the participants integrated information gained through observation of their partner similarly as information gained through their own actions.

In Experiment 2, our goal was to combine the subjects' decision-making behavior with their physiology. Similarly as in Experiment 1, the ERP time locked to stimulus onset when receiving a reward at arm 3 showed a high increase in amplitude for the Dyad (Self) condition. In comparison, there was a negative deflection in ERP for the Dyad (Other) and Solo condition. We consider this a good starting point to move towards extracting more high-level features such as EEG power spectrum density, mutual information and pupil diameter-based fixations and saccades in the future.

General Discussion

We developed a Social Multi-Armed Bandit task to examine the influence of social interaction on decision-making. We found that while some individuals do figure out that the other player's information does not apply to them, the majority of participants converged to a suboptimal decision-making strategy. We termed this strategy 'joint' as it most likely emerges through averaging the reward probabilities for both players. Measurement of electrophysiology showed a distinct P3 when the player receives a reward in the dyadic version of the multi-armed bandit but not the solo version. P3 is thought to emerge through stimulus-driven 'top-down' processes when the participant pays focused attention to a task. The distinct presence of the P3 in the dyadic task might thus hint towards heightened attention, particularly towards own rewards, in the presence of another player. Interestingly, the pupillometry data revealed a similar pattern as the ERP.

This task has considerable possibilities for further studies of social interaction. Next steps include a similar experiment with participants are previously acquainted with each other, e.g. friends, and children with their parents. It is likely that having a prior relationship with the other partner will alter the joint strategy for one or both partners. It would also be interesting to test how an asymmetric relationship (e.g., parent-child) would influence decision-making strategies, compared to a more symmetric relationship. It is possible that less-experienced participants (e.g., children) are more likely to follow, or match, the behavior of a 'reliable' person, as is the case in imitation (Poulin-Dubois, Brooker, & Polonia, 2011). Lastly, this task might give interesting insights into decision-making processes of neuro-divergent people, particularly those with potential differences in social behaviors (Montague, 2018).

Acknowledgements

This study was funded with an Innovative Research Grant from the Kavli Institute for Brain and Mind.

References

Cohen, J. D., McClure, S. M., & Yu, A. Y. (2007). Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *362*(1481), 933-942.

Delorme, A., & Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of neuroscience methods*, *134*(1), 9-21.

Dolk, T., Hommel, B., Colzato, L. S., Schütz-Bosbach, S., Prinz, W., & Liepelt, R. (2014). The joint Simon effect: a review and theoretical integration. Frontiers in Psychology, 5, 974.

Fink, A., Grabner, R. H., Benedek, M., Reishofer, G., Hauswirth, V., Fally, M., ... & Neubauer, A. C. (2009). The creative brain: Investigation of brain activity during creative problem solving by means of EEG and fMRI. *Human brain mapping*, *30*(3), 734-748.

Gittins, J., Glazebrook, K., & Weber, R. (2011). *Multi-armed bandit allocation indices*. John Wiley & Sons.

Kilner, J. M., Friston, K. J., & Frith, C. D. (2007). Predictive coding: an account of the mirror neuron system. *Cognitive Processing*, 8(3), 159-166.

Kothe, C. (2014). Lab streaming layer (LSL). *https://github.com/sccn/labstreaminglayer. Accessed on February*, *1*, 2019.

Liao, Y., Acar, Z. A., Makeig, S., & Deak, G. (2015). EEG imaging of toddlers during dyadic turn-taking: Mu-rhythm modulation while producing or observing social actions. *NeuroImage, 112*, 52-60.

Liu, K., & Zhao, Q. (2010). Distributed learning in multi-armed bandit with multiple players. *IEEE Transactions on Signal Processing*, *58*(11), 5667-5681.

Montague, P. R. (2018). Computational Phenotypes Revealed by Interactive Economic Games. In *Computational Psychiatry*.

Mullen, T., Kothe, C., Chi, Y. M., Ojeda, A., Kerth, T., Makeig, S., ... & Jung, T. P. (2013, July). Real-time modeling and 3D visualization of source dynamics and connectivity using wearable EEG. In *2013 35th annual international conference of the IEEE engineering in medicine and biology society (EMBC)* (pp. 2184-2187).

Partala, T., & Surakka, V. (2003). Pupil size variation as an indication of affective processing. *International journal of human-computer studies*, *59*(1-2), 185-198.

Polich, J. (2007). Updating P300: an integrative theory of P3a and P3b. *Clinical neurophysiology*, *118*(10), 2128-2148.

Poulin-Dubois, D., Brooker, I., & Polonia, A. (2011). Infants prefer to imitate a reliable person. Infant Behavior and Development, 34(2), 303-309.

Sebanz, N., Knoblich, G., & Prinz, W. (2003). Representing others' actions: just like one's own?. *Cognition*, *88*(3), B11-B21.

Sugrue, L. P., Corrado, G. S., & Newsome, W. T. (2004). Matching behavior and the representation of value in the parietal cortex. *science*, *304*(5678), 1782-1787.

Vesper, C., Abramova, E., Bütepage, J., Ciardo, F., Crossey, B., Effenberg, A., ... & Schmitz, L. (2017). Joint action: mental representations, shared information and general mechanisms for coordinating with others. *Frontiers in psychology*, *7*, 2039.

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta‐analysis of theory‐of‐mind development: The truth about false belief. *Child development*, *72*(3), 655-684.

Using Machine Learning to Guide Cognitive Modeling: A Case Study in Moral Reasoning

Mayank Agrawal (mayank.agrawal@princeton.edu)
Department of Psychology, Princeton University

Joshua C. Peterson (peterson.c.joshua@gmail.com)
Department of Computer Science, Princeton University

Thomas L. Griffiths (tomg@princeton.edu)
Departments of Psychology and Computer Science, Princeton University

Abstract

Large-scale behavioral datasets enable researchers to use complex machine learning algorithms to better predict human behavior, yet this increased predictive power does not always lead to a better understanding of the behavior in question. In this paper, we outline a data-driven, iterative procedure that allows cognitive scientists to use machine learning to generate models that are both interpretable and accurate. We demonstrate this method in the domain of moral decision-making, where standard experimental approaches often identify relevant principles that influence human judgments, but fail to generalize these findings to "real world" situations that place these principles in conflict. The recently released Moral Machine dataset allows us to build a powerful model that can predict the outcomes of these conflicts while remaining simple enough to explain the basis behind human decisions.

Keywords: machine learning; moral psychology

Introduction

Explanatory and predictive power are hallmarks of any useful scientific theory. However, in practice, psychology tends to focus more on explanation (Yarkoni & Westfall, 2017), whereas machine learning is almost exclusively aimed at prediction. The necessarily restrictive nature of laboratory experiments often leads psychologists to test competing hypotheses by running highly-controlled studies on tens or hundreds of subjects. Although this procedure gives a better understanding of the specific phenomenon, it can be difficult to generalize the findings and predict behavior in the "real world," where multiple factors are interacting with one another. Conversely, machine learning takes full advantage of complex, nonlinear models that excel in tasks ranging from image classification (Krizhevsky et al., 2012) to video game playing (Mnih et al., 2015). The performance of these models scales with their level of expressiveness (Huang et al., 2018), which results in millions of parameters that are difficult to interpret.

Interestingly, machine learning has long utilized insight from cognitive psychology and neuroscience (Rosenblatt, 1958; Sutton & Barto, 1981; Ackley et al., 1985; Elman, 1990), a trend that continues to this day (Banino et al., 2018; Lzaro-Gredilla et al., 2019). We believe that the reverse direction has been underutilized, but could be just as fruitful. In particular, psychology could leverage machine learning to improve both the predictive and explanatory power of cognitive models. We propose a method (summarized in Figure 1) that enables cognitive scientists to use large-scale behavioral datasets to construct interpretable models that rival the performance of complex, black-box algorithms.

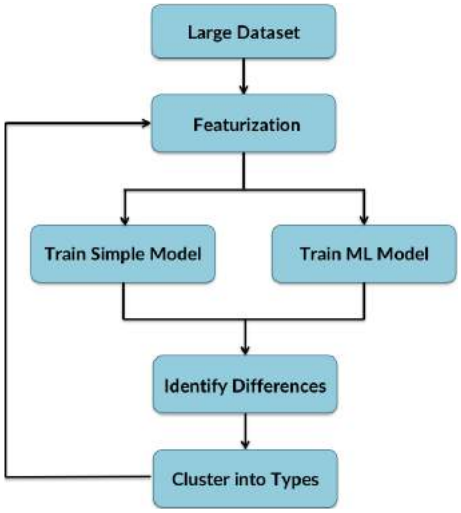


Figure 1: A systematic, data-driven procedure for building interpretable models that rival the predictive power of complex machine learning models.

This methodology is inspired by Box's loop (Box & Hunter, 1962; Blei, 2014; Linderman & Gershman, 2017), a systematic process of integrating the scientific method with exploratory data analysis. Our key insight is that training a black-box algorithm gives a sense of how much variance in a certain type of behavior can be predicted. This predictive power provides a standard for improvement in explicit cognitive models (Khajah et al., 2016). By continuously critiquing an interpretable cognitive model with respect to these black-box algorithms, we can identify and incorporate new features until its performance converges, thereby jointly maximizing our two objectives of explanatory and predictive power.

In this paper, we demonstrate this methodology by building a statistical model of moral decision-making. Philosophers and psychologists have historically conducted thought experiments and laboratory studies isolating individual principles responsible for human moral judgment (e.g. consequentialist ones such as harm aversion or deontological ones such as not using others as a means to an end). However, it can be difficult to predict the outcomes of situations in which these principles conflict (Cushman et al., 2010). The recently released

Moral Machine dataset (Awad et al., 2018) allows us to build a predictive model of how humans navigate these conflicts over a large problem space. We start with a basic rational choice model and iteratively add features until its accuracy rivals that of a neural network, resulting in a model that is both predictive and interpretable.

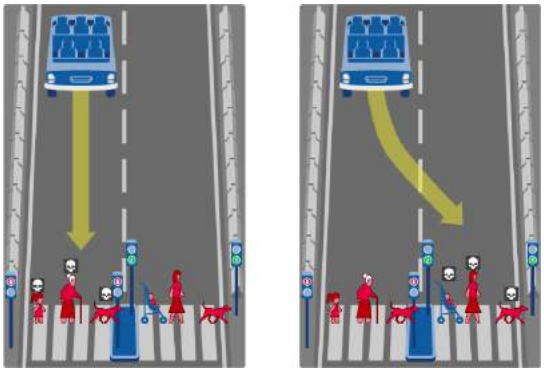
Background

Theories of Moral Decision-Making The two main families of moral philosophy often used to describe human behavior are *consequentialism* and *deontology*. Consequentialist theories posit that moral permissibility is evaluated solely with respect to the outcomes, and that one should choose the outcome with the highest value (Greene, 2007). On the other hand, deontological theories evaluate moral permissibility with respect to actions and whether they correspond to specific rules or rights.

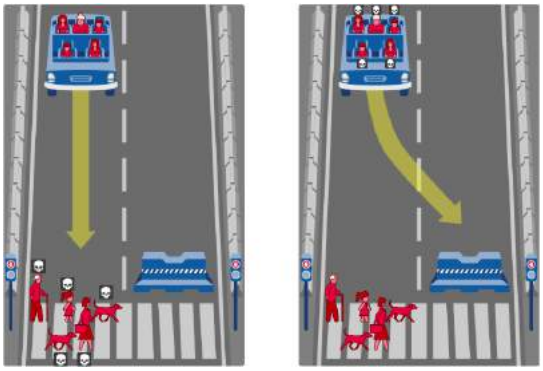
The trolley car dilemma (Foot, 2002; Thomson, 1984) highlights how these two families differ when making moral judgments. Here, participants must determine whether it is morally permissible to sacrifice an innocent bystander in order to prevent a trolley car from killing five railway workers. The "switch" scenario gives the participant the option to redirect the car to a track with one railway worker, whereas the "push" scenario requires the participant to push a large man directly in front of the car to stop it, killing the large man in the process. Given that the outcomes are the same for the "switch" and "push" scenarios (i.e., intervening results in one death, while not intervening results in five deaths), consequentialism prescribes intervention in both scenarios. Deontological theories allow for intervening in the "switch" scenario but not the "push" scenario because pushing a man to his death violates a moral principle, but switching the direction of a train does not.

Empirical studies have found that people are much more willing to "switch" than to "push" (Greene et al., 2001; Cushman et al., 2006), suggesting deontological principles factor heavily in human moral decision-making. Yet, a deontological theory's lack of systematicity makes it difficult to evaluate as a model of moral judgment (Greene, 2017). What are the rules that people invoke, and how do they interact with one another when in conflict? Furthermore, how do they interact with consequentialist concerns? Would people that refuse to push a man to his death to save five railway workers still make the same decision and with the same level of confidence when there are a million railway workers? Any theory of human moral cognition needs to be able to model how participants trade off different consequentialist and deontological factors.

Moral Machine Paradigm As society anticipates autonomous cars roaming its streets in the near future, the trolley car dilemma has left the moral philosophy classroom and entered into national policy conversations. A group of researchers aiming to gauge public opinion created "Moral Machine," an online game that presents users with moral dilemmas (see Figure 2) centered around autonomous cars (Awad et al., 2018). Comprising roughly forty million decisions from users in over two hundred countries, the Moral Machine experiment is the largest public dataset collection on human moral judgment.



(a) An autonomous car is headed towards a group of three pedestrians who are illegally crossing the street. The car can either stay and kill these pedestrians or swerve and kill three other pedestrians crossing legally.



(b) An autonomous car with five human passengers is headed towards a group of pedestrians who are illegally crossing the street. Staying on course will kill the pedestrians but save the passengers, while swerving will kill the passengers but save the pedestrians.

Figure 2: Two sample dilemmas in the Moral Machine dataset. In every scenario, the participant is asked to choose whether to *stay* or *swerve* (Awad et al., 2018).

In addition to the large number of decisions, the experiment operated over a rich problem space. Twenty unique agent types (e.g. man, girl, dog) along with contextual information (e.g. crossing signals) enabled researchers to measure the outcomes of nine manipulations: action versus inaction, passengers versus pedestrians, males versus females, fat versus fit, low status versus high status, lawful versus unlawful, elderly versus young, more lives saved versus less, and humans versus pets. The coverage and density of this problem space provides the opportunity to build a model that predicts how humans make moral judgments when a variety of different principles are at play.

Predicting Moral Decisions

As described earlier, the iterative refinement method we propose begins with both an initial, interpretable model and a more predictive black-box algorithm. In this section, we do exactly this by contrasting rational choice models derived from moral philosophy with multilayer feedforward neural networks.

Model Descriptions

We restricted our analysis to a subset of the dataset ($N = 12,478,340$) where an empty autonomous vehicle must decide between saving the pedestrians on the left or right side of the road (see Figure 2a for an example). The models we consider below are tasked to predict the probability of choosing to save the left side.

Interpretable Models Choice models (CM) are ubiquitous in both psychology and economics, and they form the basis of our interpretable model in this paper (Luce, 1959; McFadden et al., 1973). In particular, we assume that participants construct the values for both sides, i.e., v_{left} and v_{right}, and choose to save the left side when $v_{\text{left}} > v_{\text{right}}$, and vice versa. The value of each side is determined by aggregating the utilities of all its agents:

$$v_{\text{side}} = \sum_i u_i l_i \tag{1}$$

where u_i is the utility given to agent i and l_i is a binary indicator of agent i's presence on the given side.

McFadden et al. (1973) proved that if individual variation around this aggregate utility follows a Weibull distribution, the probability that v_{left} is optimal is consistent with the exponentiated Luce choice rule used in psychology, i.e.,

$$P(v_{\text{left}} > v_{\text{right}}) = P(c = \text{left} | v_{\text{left}}, v_{\text{right}}) = \frac{e^{v_{\text{left}}}}{e^{v_{\text{left}}} + e^{v_{\text{right}}}} \tag{2}$$

In practice, we can implement this formalization by using logistic regression to infer the utility vector $\mathbf{u}$. We built three models, each of which provided top-down different constraints on the utility vector. Our first model, "Equal Weight," required each agent to be equally weighted. At the other extreme, our "Utilitarian" model had no restriction. A third model, "Animals vs. People," was a hybrid: all humans were were weighted equally and all animals were weighted equally, but humans and animals could be weighted differently.

Research in moral psychology and philosophy has found that humans use moral principles in addition to standard utilitarian reasoning when choosing between options (Quinn, 1989; Spranca et al., 1991; Mikhail, 2002; Royzman & Baron, 2002; Baron & Ritov, 2004; Cushman et al., 2006). For example, one principle may be that allowing harm is more permissible than doing harm (Woollard & Howard-Snyder, 2016). In order to incorporate these principles, we moved beyond utilitarian-based choice models by expanding the definition of a side's value:

$$v_{\text{side}} = \sum_i u_i l_i + \sum_m \lambda_m f_m \tag{3}$$

where f_m is an indicator variable of whether principle m is present on the side and λ_m represents the importance of principle m. We built an "Expanded" model that introduces two principles potentially relevant in the Moral Machine dataset. The first is a preference for allowing harm over doing harm, thus penalizing sides that require the car to swerve in order to save them. Another potentially relevant principle is that it is more justified to punish unlawful pedestrians than lawful ones because they knowingly waived their rights when crossing illegally (Nino, 1983). This model was trained on the dataset to infer the values of $\mathbf{u}$ and λ.

Neural Networks We use relatively expressive multilayer feedforward neural networks (NN) to provide an estimate of the level of performance that statistical models can achieve in this domain. These networks were given as inputs the forty-two variables that uniquely defined a dilemma to each participant: twenty for the characters on the left side, twenty for the characters on the right side, one for the side of the car, and one for the crossing signal status. These are the same inputs for the "Expanded" choice model. However, the "Expanded" model had the added restriction that the side did not change an agent's utility (e.g., a girl on the left side has the same utility as a girl on the right side), while the neural network had no such restriction.

The networks were trained to minimize the crossentropy between the model's output and human binary decisions. The final layer of the neural networks is similar to the choice model in that it is constructing the value of each side by weighting different features. However, in these networks, the principles are learned from the nonlinear interactions of multiple layers and the indicators are probabilistic rather than deterministic.

To find the optimal hyperparameters, we conducted a grid search, varying the number of hidden layers, the number of hidden neurons, and the batch size. All networks used the same ReLU activation function and and no dropout. Given that most of these models both performed similarly and showed a clear improvement over simple choice models, we did not conduct a more extensive hyperparameter search. A neural network with three 32-unit hidden layers was used for all the analyses in this paper.

Model Comparisons

Standard Metrics Table 1 displays the results of the four rational choice models and the best performing neural network. All models were trained on eighty percent of the dataset, and the reported results reflect the performance on the held-out twenty percent. We report accuracy and area under the curve (AUC), two standard metrics for evaluating classification models. We also calculate the normalized Akaike information criterion (AIC), a metric for model comparison that integrates a model's predictive power and simplicity. All metrics resulted in the same expected ranking of models: Neural Network, Expanded, Utilitarian, Animals vs. People, Equal Weight.

Table 1: Comparison of Standard Metrics

Model Type	Accuracy	AUC	AIC
Equal Weight	0.571	0.616	1.301
Animals vs. People	0.630	0.702	1.234
Utilitarian	0.732	0.780	1.146
Expanded	0.763	0.826	1.046
Neural Network	0.774	0.845	0.983

Performance as a Function of Dataset Size Table 1 demonstrates that our cognitive models aren't as predictive as a powerful learning algorithm. This result, however, is only observable with larger datasets. Figure 3 plots each metric for each model over a large range of dataset sizes. Choice models performed very well at dataset sizes comparable to that of a large laboratory experiment. Conversely, neural networks improved with larger dataset sizes until reaching an asymptote where $N > 100,000$, at which point they outperform rational choice models. These results suggest that while psychological models are robust in the face of small datasets, they need to be evaluated on much larger ones.

Identifying Explanatory Principles

The neural network gives us an aspirational standard of how our simpler model should perform. Next, our task is to identify the emergent features it constructs and incorporate them into our simple choice model.

Calculating Residuals in Problem Aggregates By aggregating decisions for each dilemma, we can determine the empirical "difficulty" of each dilemma and whether our models predict this difficulty. For example, assume dilemmas A and B have been proposed to one hundred participants. If ninety participants exposed to dilemma A chose to save the left side and sixty participants exposed to dilemma B did, the empirical percentages for A and B would be 0.90 and 0.60, respectively. An accurate model of moral judgment should not only reflect the binary responses but also the confidence behind those responses.

We identified the specific problems where the neural network excelled compared to the "Expanded" rational choice model. Manually inspecting these problems and clustering them into groups revealed useful features beyond those employed in the choice model that the neural network is constructing. We formalized these features as principles and incorporated them into the choice model to improve prediction. Two examples are represented in Table 2.

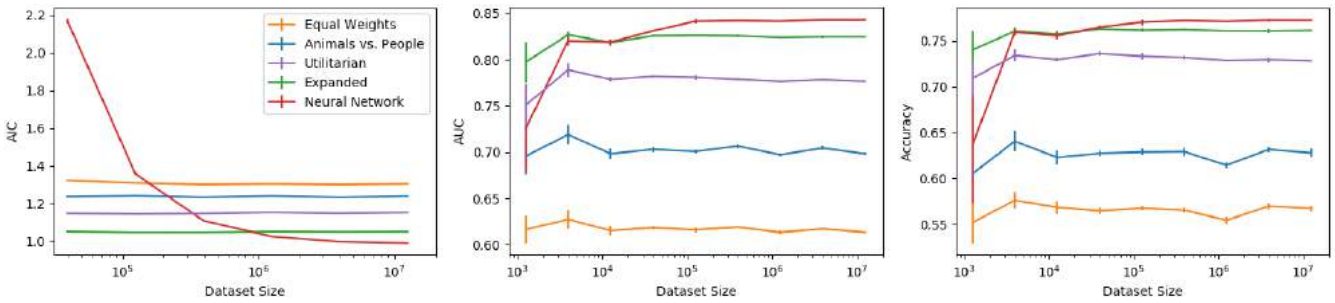
Table 2a describes a set of scenarios where one human is crossing illegally and one pet is crossing legally. Empirically, users tend to overwhelmingly prefer saving the human, while the choice model predicts the opposite. Our choice model's inferred utilities and importance values reveal a strong penalty (i.e., a large negative coefficient) for (1) humans crossing illegally and (2) requiring the car to swerve. However, the empirical data suggests that these principles are outweighed by the fact that this is a humans-versus-animals dilemma, and that humans should be preferred despite the crossing or intervention status. Thus, the next iteration of our model should incorporate a binary variable signifying whether this is an explicit humans-versus-animals dilemma.

We can conduct a similar analysis for the set of scenarios in Table 2b. Both models output significantly different decision probabilities, the neural network being the more accurate of the two. Most salient to us was an effect of age. Specifically, when the principal difference between the two sides is age, both boys and girls should be saved at a much higher rate, and information about their crossing and intervention status is less relevant. To capture this fact, we can incorporate another binary variable signifying whether the only difference between the agents on each side is age.

Incorporating New Features The two features we identified are a subset of six "problem types" the Moral Machine researchers used in their experiment: humans versus animals, old versus young, more versus less, fat versus fit, male versus female, and high status versus low status. These types were not revealed to the participants, but the residuals we inspected suggest that participants were constructing them from the raw features and then factoring them into their decisions.

Incorporating these six new features as principles resulted in 77.1% accuracy, nearly closing the gap entirely between our choice model and neural network performance reported in Table 1. Figure 4 illustrates the effects of incorporating the problem types into both the choice model and the neural network in details. Importantly, we observe that "Neural Network + Types" outperforms "Neural Network" at smaller dataset sizes, but performs identically at larger dataset sizes. This result suggests that the regular "Neural Network" is constructing the problem types we identified as emergent features given sufficient data to learn them from. More importantly, our augmented choice model now rivals the neural network's predictive power. And yet, by virtue of it being a rational choice model with only a few more parameters than our "Expanded" (and even the "Utilitarian") model, it remains conceptually simple. Thus, we have arrived at an interpretable statistical model that can both quantify the effects of utilitarian calculations and moral principles and predict human moral judgment over a large problem space.

Figure 4b still displays a gap between the AUC curves, suggesting there is more to be gained by repeating the process and potentially identifying new even more principles. For example, the last iteration found that when there was a humans-versus-animals problem, humans should be strongly favored. However, residuals suggest that participants don't honor this principle when all the humans are criminals. Rather, in these cases, participants may favor the animals or prefer the criminal by only a small margin. Thus, our next iteration will include a feature corresponding to whether all the humans are criminals. Our model also underperforms by overweighting



(a) Dataset Size vs. AIC (b) Dataset Size vs. AUC (c) Dataset Size vs. Accuracy

Figure 3: Test-set performance metrics of choice models and neural network[1] as a function of dataset size. Models were trained on five 80/20 training/test splits. Error bars indicate ± 1 SEM.

Table 2: Problem Aggregate Comparisons (Left Side Save Percentage)

Left Side Agents	**Right Side Agents**	**Car Side**	**Empirical**	**CM**	**NN**
Pregnant Woman Crossing Illegally	Cat Crossing Legally	Left	0.779	0.411	0.797
Stroller Crossing Illegally	Cat Crossing Legally	Left	0.826	0.425	0.801
Dog Crossing Legally	Male Doctor Crossing Illegally	Right	0.312	0.693	0.293
Cat Crossing Legally	Man Crossing Illegally	Right	0.308	0.692	0.266
Old Woman Crossing Illegally	Cat Crossing Legally	Left	0.670	0.306	0.622

(a) Problems indicating Human vs. Animals Principle

Left Side Agents	**Right Side Agents**	**Car Side**	**Empirical**	**CM**	**NN**
Old Man Crossing Legally	Boy Crossing Illegally	Right	0.350	0.647	0.341
Old Woman Crossing Legally	Girl Crossing Illegally	Right	0.337	0.642	0.321
Man	Boy	Left	0.113	0.417	0.097
Old Woman Crossing Legally	Girl Crossing Illegally	Left	0.268	0.570	0.269
Old Woman	Woman	Right	0.256	0.475	0.269

(b) Problems indicating Old vs. Young Principle

the effects of intervention. In problem types such as male versus female and fat versus fit, the intervention variable is weighted much differently than in young-versus-old dilemmas. The next iteration of the model should also include this interaction. Thus, this methodology allows us to continuously build on top of the new features we identify.

Conclusion

Large-scale behavioral datasets have the potential to revolutionize cognitive science (Griffiths, 2015), and while data science approaches have traditionally used them to predict behavior, they can additionally help cognitive scientists construct explanations of the given behavior.

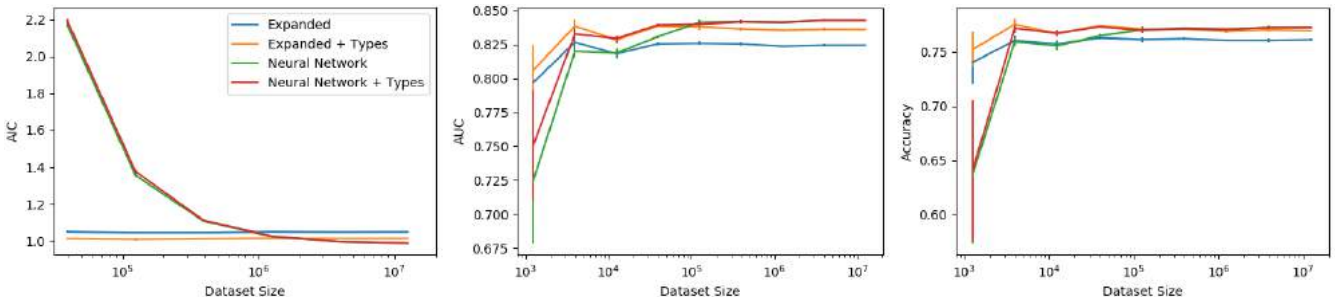
Black-box machine learning algorithms give us a sense of the predictive capabilities of our scientific theories, and we outline a methodology that uses them to help cognitive models reach these capabilities:

1. Amass a large-scale behavioral dataset that encompasses a large problem space
2. Formalize interpretable theories into parameterizable psychological models whose predictions can be evaluated
3. Compare these models to more accurate, but less interpretable black-box models (e.g., deep neural networks, random forests, etc.)
4. Identify types of problems where the black-box models outperform the simpler models
5. Formalize these problem types into features and incorporate them into both the simple and complex models
6. Return to Step 4 and repeat

We applied this procedure to moral decision-making, starting off with a rational choice model and iteratively adding principles until it had a comparable predictive power with black-box algorithms. This model allowed us to quantitatively predict the interactions between different utilitarian concerns and moral principles. Furthermore, our results regarding problem types suggest that moral judgment can be better predicted by incorporating alignable differences in similarity judgments (Tversky & Simonson, 1993), such as whether the dilemma is humans-versus-animals or old-versus-young.

The present case study, while successful, is only a limited application of the methodology we espouse, and further demonstrations are required to illustrate its utility. It will be

[1]While a batch size of 8,192 was used for Table 1, a batch size of 512 was used here because of the smaller dataset sizes.



(a) Dataset Size vs. AIC (b) Dataset Size vs. AUC (c) Dataset Size vs. Accuracy

Figure 4: Test-set performance metrics before and after incorporating new principles. Models were trained on five 80/20 training/test splits. Error bars indicate ± 1 SEM.

particularly interesting to apply our method to problems with even larger gaps between classic theories and data-driven predictive models. It is also likely that transferring insights from data-driven models will require moving beyond the sorts of featurization we consider here (i.e., problem clustering). In any case, we hope the microcosm presented here will inspire similarly synergistic approaches in other areas of psychology.

Acknowledgments. We thank Edmond Awad for providing guidance on navigating the Moral Machine dataset.

References

Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985). A learning algorithm for boltzmann machines. *Cognitive science*, *9*(1), 147–169.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., ... Rahwan, I. (2018). The moral machine experiment. *Nature*, *563*(7729), 59.

Banino, A., Barry, C., Uria, B., Blundell, C., Lillicrap, T., Mirowski, P., ... others (2018). Vector-based navigation using grid-like representations in artificial agents. *Nature*, *557*(7705), 429.

Baron, J., & Ritov, I. (2004). Omission bias, individual differences, and normality. *Organizational Behavior and Human Decision Processes*, *94*(2), 74–85.

Blei, D. M. (2014). Build, compute, critique, repeat: Data analysis with latent variable models. *Annual Review of Statistics and Its Application*, *1*, 203–232.

Box, G. E., & Hunter, W. G. (1962). A useful method for model-building. *Technometrics*, *4*(3), 301–318.

Cushman, F., Young, L., & Greene, J. D. (2010). Our multi-system moral psychology: Towards a consensus view. *The Oxford handbook of moral psychology*, 47–71.

Cushman, F., Young, L., & Hauser, M. (2006). The role of conscious reasoning and intuition in moral judgment: Testing three principles of harm. *Psychological science*, *17*(12), 1082–1089.

Elman, J. L. (1990). Finding structure in time. *Cognitive science*, *14*(2), 179–211.

Foot, P. (2002). The problem of abortion and the doctrine of the double effect. *Virtues and Vices and Other Essays in Moral Philosophy*, 1932.

Greene, J. D. (2007). The secret joke of kant's soul. In W. Sinnott-Armstrong (Ed.), *Moral psychology: The neuroscience of morality: Emotion, brain disorders, and development* (Vol. 3, chap. 2). MIT Press.

Greene, J. D. (2017). The rat-a-gorical imperative: Moral intuition and the limits of affective learning. *Cognition*, *167*, 66–77.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fmri investigation of emotional engagement in moral judgment. *Science*, *293*(5537), 2105–2108.

Griffiths, T. L. (2015). Manifesto for a new (computational) cognitive revolution. *Cognition*, *135*, 21–23.

Huang, Y., Cheng, Y., Chen, D., Lee, H., Ngiam, J., Le, Q. V., & Chen, Z. (2018). Gpipe: Efficient training of giant neural networks using pipeline parallelism. *arXiv preprint arXiv:1811.06965*.

Khajah, M., Lindsey, R. V., & Mozer, M. C. (2016). How deep is knowledge tracing? *arXiv preprint arXiv:1604.02416*.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

Linderman, S. W., & Gershman, S. J. (2017). Using computational theory to constrain statistical models of neural data. *Current opinion in neurobiology*, *46*, 14–24.

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*.

Lzaro-Gredilla, M., Lin, D., Guntupalli, J. S., & George, D. (2019). Beyond imitation: Zero-shot task transfer on robots by learning concepts as cognitive programs. *Science Robotics*, *4*(26).

McFadden, D., et al. (1973). Conditional logit analysis of qualitative choice behavior.

Mikhail, J. (2002). Aspects of the theory of moral cognition: Investigating intuitive knowledge of the prohibition of intentional battery and the principle of double effect.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... others (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529.

Nino, C. S. (1983). A consensual theory of punishment. *Philosophy & Public Affairs*, 289–306.

Quinn, W. S. (1989). Actions, intentions, and consequences: The doctrine of doing and allowing. *The Philosophical Review*, *98*(3), 287–312.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, *65*(6), 386.

Royzman, E. B., & Baron, J. (2002). The preference for indirect harm. *Social Justice Research*, *15*(2), 165–184.

Spranca, M., Minsk, E., & Baron, J. (1991). Omission and commission in judgment and choice. *Journal of experimental social psychology*, *27*(1), 76–105.

Sutton, R. S., & Barto, A. G. (1981). Toward a modern theory of adaptive networks: expectation and prediction. *Psychological review*, *88*(2), 135.

Thomson, J. J. (1984). The trolley problem. *Yale Law Journal*, *94*, 1395.

Tversky, A., & Simonson, I. (1993). Context-dependent preferences. *Management science*, *39*(10), 1179–1189.

Woollard, F., & Howard-Snyder, F. (2016). Doing vs. allowing harm. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Winter 2016 ed.). Metaphysics Research Lab, Stanford University.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122.

Quantifying the Conceptual Combination Effect on Word Meanings

Nora Aguirre-Celis (naguirre@cs.utexas.edu)
ITESM, Monterrey, Mexico & The University of Texas at Austin
Department of Computer Science, 2317 Speedway
Austin, TX 78712 USA

Risto Miikkulainen (risto@ cs.utexas.edu)
The University of Texas at Austin
Department of Computer Science, 2317 Speedway
Austin, TX 78712 USA

Abstract

How do people understand concepts such as *dog*, *aggressive dog*, *dog house* or *house dog*? The meaning of a concept depends crucially on the concepts around it. While this hypothesis has existed for a long time, only recently it has become possible to test it based on neuroimaging and quantify it using computational modeling. In this paper, a neural network is trained with backpropagation to map attribute-based semantic representations to fMRI images of subjects reading everyday sentences. Backpropagation is then extended to the attributes, demonstrating how word meanings change in different contexts. Across a large corpus of sentences, the new attributes are more similar to the attributes of other words in the sentence than they are to the original attributes, demonstrating that the meaning of the context is transferred to a degree to each word in the sentence. Such dynamic conceptual combination effects could be included in natural language processing systems to encode rich contextual embeddings to mirror human performance more accurately.

Keywords: Context Effect; Concept Representations; Conceptual Combination; fMRI Data Analysis; Neural Networks; Embodied Cognition

Introduction

In the embodied cognition approach. (Barsalou, 2008, Binder et al., 2009), the meaning of a concept is not a set of verbal features that people associate with the concept, but rather a set of neural processing modalities that are involved while experiencing instances of the concept. This approach provides a direct correspondence between conceptual content and neural representations, and suggests that concepts can be represented through a number of weighted semantic dimensions that correspond to different brain areas. Recently it has become possible to ground this approach to brain imaging. In particular, Binder et al. (2009) identified a distributed large-scale brain network linked to the storage and retrieval of words. This brain network was used as the foundation for the Concept Attributes Representation (CAR) theory (a.k.a. the experiential attribute representation model). CAR theory proposes that words are represented as a set of weighted attributes stimulated by context.

People weigh concept features differently based on context, i.e., they construct a meaning dynamically according to the combination of concepts that occur in the sentence. Such conceptual combination either uses an attribute of one concept to describe another (in attribute combination) or forms some relation between two concepts to create a new one (in relational combination). In case of attribute combination, the modifier features adapt other concepts in the combination to some degree, and as a result, the words involved are alike (Wisniewski, 1998). For example, listeners must realize that *red apple* could mean just a fruit having a certain color by selecting salient features that dominate in the combination. The noun *apple* is defined by color, size, shape, taste, etc. and one or more of those dimensions will be modified during the attribute combination. In relational combination, the modifier features have nothing to do with the combination. For example *apple basket* or *apple pie* contain a variety of relations that often do not include *apple's* features as in *apple baskets* are not edible, red or a fruit. To help understand that *apple pie* is made of apples but *apple baskets* are not, a thematic relation needs to be built based on world knowledge about plausible combinations. Both attribute and relational combinations play an important role in the construction of new or complex concepts (Gagné & Shoben, 1997; Murphy 1990; Pecher, Zeelenberg, & Barsalou, 2004).

This paper focuses on the attribute combination process. It describes how such a dynamic construction of concepts in the brain can be quantified. This question has been studied in previous work anecdotally, by analyzing a few example cases of how the meaning attributes are weighted differently in various contexts for individual concepts, combinations of concepts, and for sentences (Aguirre-Celis & Miikkulainen, 2017, 2018). The current study expands on this prior work by evaluating the robustness and generality of these conclusions across an entire corpus of sentences and semantic roles. A neural network is trained to map brain-based semantic representations of words (CARs) into fMRI data of subjects reading everyday sentences. Backpropagation is then repeated separately for each sentence, reducing the remaining error by modifying only the CARs at the input of the network. As a result, the strengths of the attributes in the CARs change according to how important each attribute is for that sentence context.

The CAR theory is first reviewed, and the sentence collection, fMRI data, and word representation data described. The computational model is presented, followed by the experiments: an example individual case of how

conceptual combinations affect word meanings, and an aggregate study across a corpus of sentences.

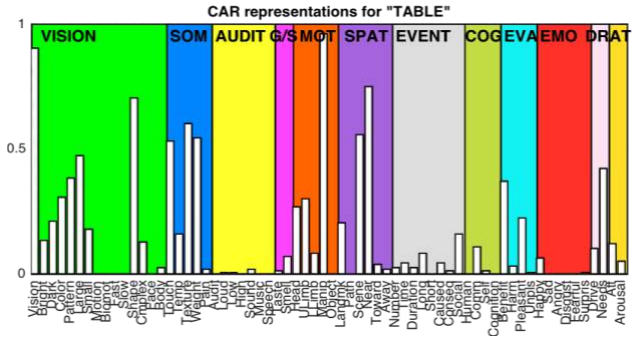


Figure 1: Bar plot of the 66 semantic features for the concept *table*. The values represent average human ratings for each feature. Given that *table* is an object, it gets low weightings on human-related attributes such as Face, Speech, Head, and emotions including Happy, Sad, and Angry, and high weightings on attributes like Vision, Shape, Touch, and Manipulation.

Concept Attribute Representation Theory

CAR theory represents the basic components of meaning defined in terms of observed neural processes and brain systems thereby relating semantic content to systematic modulation of neuroimaging activity (Anderson, et al., 2016; Binder, et al., 2009). They are composed of a list of modalities that correspond to specialized sensory, motor and affective functions, and are therefore not limited to the classical sensory-motor dimensions of most embodied theories.

CARs capture aspects of experience central to the acquisition of event and object concepts, both abstract and concrete. For example, concept ratings on visual and sensory components include brightness, color, size, shape, temperature, weight, pain, etc. These aspects of mental experience model each word as a collection of a 66-dimensional feature vector that captures the strength of association between each neural attribute and the word meaning. For instance, Figure 1 shows the CAR for the concept *table*.

The attributes in CAR theory were selected after an extensive body of physiological evidence based on two assumptions: (1) All aspects of mental experience can contribute to concept acquisition and consequently concept composition; (2) experiential phenomena are grounded on neural processors representing a particular aspect of experience. For a more detailed account of the attribute selection and definition see Binder, et al., (2009, 2011, 2016a, and 2016b). The next section describes how the CAR theory is instantiated by acquiring attribute ratings from human subjects.

Data Preparation

Three data collections were used in this study: A sentence collection prepared by Glasgow et al., (2016), the fMRI images for these sentence by the Medical College of Wisconsin (Anderson, et al., 2016; Binder, et al., 2016), and semantic Vectors (CAR ratings) for words obtained via Mechanical Turk (Anderson, et al., 2016; Binder, et al., 2009). In addition, fMRI representations were synthesized for individual words from the sentence fMRI. Each of these data collections is described in more detail below.

Sentence Collection

The sentence set was prepared for the fMRI study as part of the Knowledge Representation in Neural Systems Program (KRNS). A total of 240 sentences were composed from two to five content words from a set of 242 words (141 nouns, 39 adjectives and 62 verbs). The words were selected toward imaginable and concrete objects, actions, settings, roles, state and emotions, and events. Examples include *couple*, *author*, *boy*, *theatre*, *hospital*, *desk*, *red*, *flood*, *damaged*, *drank*, *gave*, *happy*, *old*, *summer*, *chicken*, *dog*.

The sentence collection is not fully balanced and systematic, but instead aims to be a natural sample. In order to investigate the effect of context, pairs of contrasting sentences were identified in this collection in an early study. This pairs include differences and similarities such as live mouse vs. dead mouse, family celebrated vs. happy family, and playing soccer vs. watching soccer. The resulting collection of 77 such sentences, with different shades of meaning for verbs, nouns and adjectives, as well as different contexts for nouns and adjectives was used to identify anecdotal examples (Table 1). However, the entire colection of sentences was used in the aggregate study described below.

Table 1: Contrasting Sentences. Sentence examples with differences and similarities in meaning. For instance, the role of the verb *flew* is used in two different contexts, bird and duck flying (animate) vs. plane flying (inanimate). Such sentence pairs illustrate the idea of conceptual combination well. However, the entire set of sentences was used in the aggregate study described in this paper.

SEMANTIC CONTRAST		SENTENCES
GOOD	94	*The soldier delivered the medicine.*
AGGRESSIVE	112	*The soldier kicked the door.*
ANIMAL	203	*The yellow bird flew over the field.*
	207	*The duck flew.*
OBJECT	210	*The red plane flew through the cloud.*
BAD PEOPLE	119	*The dangerous criminal stole the television.*
	152	*The mob was dangerous.*
NATURE	99	*The flood was dangerous.*

Neural fMRI Representation of Sentences

To obtain the neural correlates of the 240 sentences, subjects viewed each sentence on a computer screen while in the fMRI scanner. The sentences were presented word-by-word using a rapid serial visual presentation paradigm, with each content word exposed for 400ms followed by a 200ms inter-stimulus interval. Participants were instructed to read the sentences and think about their overall meaning.

Eleven subjects took part in this experiment producing 12 repetitions each. The fMRI data were preprocessed using standard methods, including slice timing and motion correction (AFNI software, Cox 1996). The most stable,

active and discriminative voxels were then selected and Principal Component Analysis and zero mean normalization were performed on them. These transformed brain activation patterns were converted into a single-sentence fMRI representation per participant by taking the voxel-wise mean of all repetitions (Anderson, et al., 2016; Binder, et al., 2016, 2016b). To form the target for the neural network, the most significant 396 voxels per sentence were then chosen (to match six case-role slots of the content words consisting of 66 attributes each) and scaled to [0.2..0.8].

Synthetic fMRI Word Representations

The neural data set did not include fMRI images for words in isolation. Therefore a technique developed by Anderson et al. (2016) was adopted to approximate them. The voxel values for a word were obtained by averaging all fMRI images for the sentences where the word occurs. These vectors, called SynthWords, encode a combination of examples of that word along with other words that appear in the same sentence. Thus, the SynthWord representation for *mouse* contains aspects of running, forest, man, seeing, and dead, from sentence 56:*The mouse ran into the forest* and sentence 60:*The man saw the dead mouse*.

This process of combining contextual information is similar to many semantic models in computational linguistics (Baroni et. al., 2010; Burgess, 1998; Landauer et al., 1997; Mitchell & Lapata, 2010). In other studies, this approach has been used successfully to predict brain activation (Anderson, et al., 2016; Binder, et al., 2016a, 2016b; Just, et al., 2017).

Due to the limited number of combinations, some of SynthWords became identical and were excluded from the dataset. The final collection includes 237 sentences and 236 words (138 nouns, 38 adjectives and 60 verbs).

Semantic CAR Representations for Words

CAR ratings were collected for the original set of 242 words (Glasgow et al., 2016) through Amazon Mechanical Turk. In a scale of 0-6, the participants were asked to assign the degree to which a given concept is associated to a specific type of neural component of experience (e.g., "To what degree do you think of a *table* as having a fixed location, as on a map?"). Approximately 30 ratings were collected for each word. After averaging all ratings and removing outliers, the final attributes were transformed to unit length yielding a 66-dimensional feature vector (Figure 1).

Note that this approach build its representations by directly mapping the conceptual content of a word (expressed in the questions) to the corresponding neural processes and systems for which the CAR dimensions stand. This approach thus contrasts with systems where the features are extracted from text corpora and word co-occurrence (Baroni et. al., 2010; Burgess, 1998; Harris, 1970; Landauer & Dumais, 1997).

Computational Approach

The approach for quantifying the effect of context in the fMRI data is based on the FGREP neural network (Forming Global Representations with Extended BP, Miikkulainen & Dyer, 1991). The idea is to train a neural network to predict what the sentence fMRI should be, based on the CAR representations, and then use FGREP to modify the CARs so that that prediction becomes correct.

Therefore, a simple three-layer neural network is first trained to map the CAR representations to word fMRI (in the left side of Figure 2, the mapping from CARWords, or word attribute ratings, to SynthWords, i.e., fMRI synthetic words).

After training, this network is used to predict what the sentence fMRI would be without the context effects. The SynthWords in the sentence are averaged to form this prediction called SynthSent. The SynthSent is then compared to fMRISent (the original fMRI data) to form an error signal.

That signal is backpropagated through the network (right side of figure 2), but the neural network weights are no longer changed. Instead, the error is used to change the CARWords (which is the FGREP method). This modification can be carried out through multiple iterations until the error goes to zero, or no additional change is possible (because the CAR attributes are already at their max or min limits). Eventually, the revised CARWord represents the word meaning for the current sentence such that when combined with other CARWords in the sentence, the prediction of sentence fMRI is correct.

For the experiments, the FGREP model was trained 20 times with different random seeds for each of the eleven fMRI subjects. A total of 20 different sets of 786 context-based word representations (one word representation for each sentence where the word appear) were thus produced for each subject. Afterwards, the mean of the 20 representations was used to represent each word.

Results

Previous work showed (1) that words in different contexts have different representations, and (2) these differences are determined by context (Aguirre-Celis & Miikkulainen 2017, 2018). These effects were demonstrated by analyzing individual sentence cases across multiple fMRI subjects. This paper verifies these same conclusions in the aggregate through a statistical analysis across an entire corpus of sentences. It measures how the CAR representation of a word changes in different sentences, and correlates these changes to the CAR representations of the other words in the sentence. In other words, it quantifies the conceptual combination effect statistically across sentences and subjects.

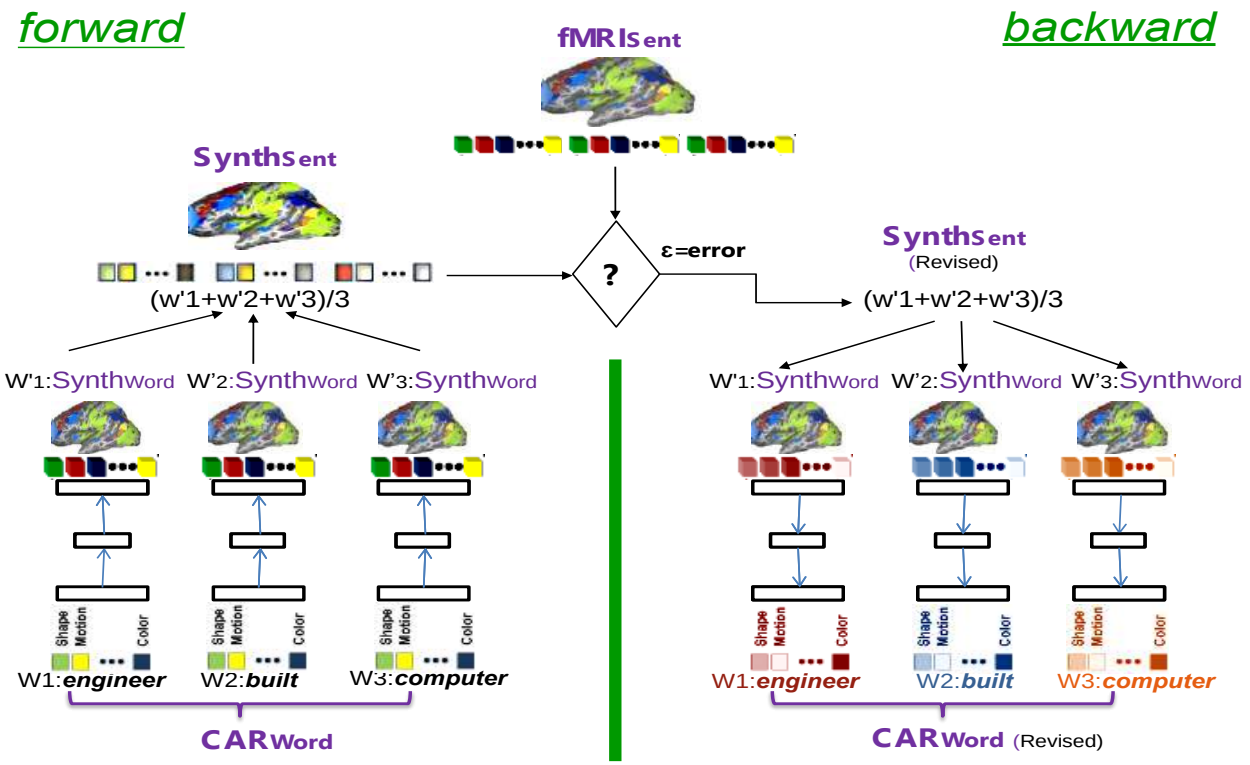


Figure 2: Using FGREP to quantify the effect of context on word meaning. Brain-based semantic representations (CARs) are changed based on a difference of predicted and actual fMRI image for a sentence. (1) Propagate CARWord to SynthWord. (2) Construct SynthSent by averaging the words into a prediction of the sentence. (3) Compare SynthSent against Observed fMRISent. (4) Backpropagate the error with FGREP for each sentence, freezing network weights and changing only CARWord. (5) Repeat until error reaches zero or CARs reach their upper or lower limits. As a result, the changes in CARs illustrate the effect of context on word meaning.

A detailed individual example of the conceptual combination effect is first presented, followed by the aggregate analysis.

The Conceptual Combination Effect

As discussed above, in CAR theory, concepts' interactions arise within multiple brain networks, activating similar brain zones for both concepts. These interactions determine the meaning of the concept combination (Binder, 2016a, 2016b).

As an example, consider the noun-verb interactions in Sentence 200: *The yellow bird flew over the field*, and Sentence 207: *The red plane flew through the cloud*. Since *bird* is a living thing, animate dimensions related to agency such as sensory, gustative, motor, affective, and cognitive experiences are expected to be activated, including potentially attributes like Speech, Taste, and Smell. In contrast, *plane flew* is expected to activate inanimate dimensions related to perceiving an object, as well as possibly Emotion, Cognition, and Attention.

Figure 3 shows the CARs for the word *flew* in the two sentences after they were modified by FGREP as described in Figure 2 and averaged across all 11 subjects. In Sentence 200 there were indeed high activations on animate attributes like Small, Pain, Smell and Taste, Audition, Music, Speech, as well as Communication and Cognition. In contrast, Sentence 207 emphasizes perceptual features like Color, Size, and Shape, Weight, Audition, Loud, Duration, Social, Benefit, and Attention.

These results illustrates the effect of conceptual combination on word meaning. As the context varies, the overlap on neural representations create a mutual enhancement, producing a clear difference between animate and inanimate contexts. The FGREP method then encodes this effect into the CAR representations where it can be measured. In other experiments, a similar effect was observed for several other noun-verb pairs, as well as several adjective-noun pairs. In the next section the effect is quantified statistically across the entire corpus of sentences.

Aggregation Analysis

So far, the conceptual combination effect has been demonstrated in a number of example cases, like the one above, and others in earlier work (Aguirre-Celis & Miikkulainen 2017, 2018). The goal of the aggregation study in this paper is to demonstrate that the effect is robust and general across the entire corpus of sentences and case roles. The hypothesis is that similar sentences have a similar effect, and this effect is consistent across all words in the sentence.

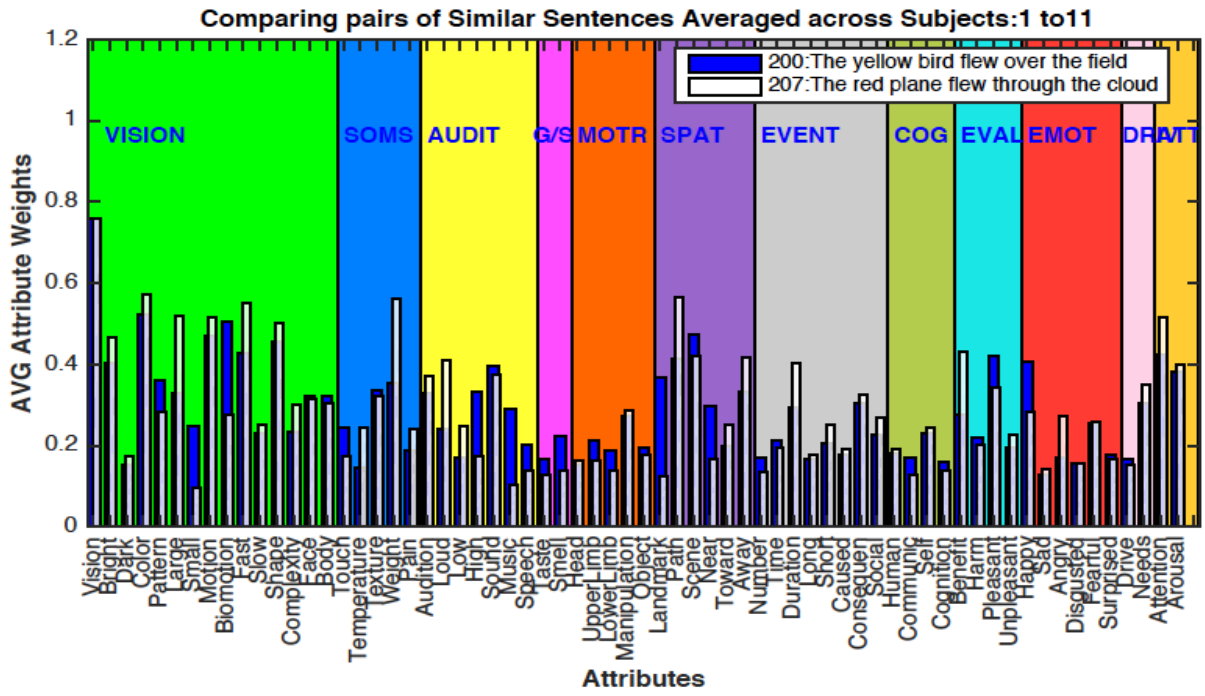


Figure 3: Contrasting the conceptual combination effect in two different sentences. In Sentence 200 (blue bars), the CAR representation modified by FGREP for the word *flew* has salient activations on animate features, presumably denoting *bird* properties like Pain, Small, Smell and Taste, and Communication. In Sentence 207 (white bars), it has high activations on inanimate object features, describing a Loud, Large, and Heavy object such as a *plane*.

This hypothesis was verified in the following process:

1. For each subject, modified CARs for each word in each sentence were formed through FGREP as described in Figure 2.
2. A representation for each sentence, SynthSent, was assembled by averaging the modified CARs.
3. Clusters of sentences were formed by running the Matlab function ***linkage*** on the set of SynthSents. Linkage measure the distance between clusters using the Ward method and the distance between elements with Euclidean distance. It treats each sentence as a single cluster at the beginning and then successively merges pairs of clusters. The process was stopped at 30 clusters, i.e., at the point where the granularity appeared most meaningful (e.g., sentences describing open locations vs. closed locations).
4. For each cluster, CAR representations with similar roles (agent, verb patient) were identified.
5. For each word in each such role, the differences between the modified CAR representations and the original CARs were calculated and averaged, and statistical significance of the difference measured using t-test across the entire set for each CAR dimensions.
6. The CARs of the other words in the sentence were averaged.
7. Pearson's Correlations were then calculated between the modified CARs and the averages CARs of other words across all the dimensions.
8. Similarly, correlations were calculated for the original CARs.
9. These two correlations were then compared. If the modified CARs correlate with the CARs of other words in the sentence better than the original CARs, there is evidence of context effect based on conceptual combination

In other words, this process aims to demonstrate that changes in a word CAR originate from the other words in the sentence. As in the example presented in the previous subsection, the noun-verb combination of *bird flew* and *plane flew* showed how some of the noun properties (animate/inanimate) were transferred to the verb, adapting the combination to the extent that the words share similar features. For example, if the other words in the sentence have high values in the CAR dimension for *Small*, then that dimension in the modified CAR should be higher than in the original CAR for that word. The correlation analysis measures this effect across the entire CAR representation. It measures whether the word meaning changes towards the context meaning.

The results are shown in detail in Table 2. The correlations are significantly higher for new CARs than for the original CARs across all subjects and all roles. As a summary, the average correlation was 0.3201 (STDEV

0.020) for original CAR representations and 0.3918 (STDEV 0.034) for new CAR representations. The results indeed confirm that the conceptual combination effect occurs reliably across subjects and sentences, and it is possible to quantify it by analyzing the fMRI images using the FGREP method on CAR representations.

Table 2: Correlation results. Average correlations analyzed by word class for 11 subjects comparing the original and new CARs vs. the average of the other words in the sentence. A moderate to strong positive correlation was found between new CARs and the other words in the sentence suggesting that features on one word are transferred to other words in the sentence during conceptual combination.

AVERAGE CORRELATIONS PER SUBJECT (3 ROLES)						
	ORIGINAL			NEW		
SUBJECTS	AGENT	VERB	PATIENT	AGENT	VERB	PATIENT
5051	0.2956	0.2884	0.3138	0.3908	0.3760	0.4147
5146	0.3272	0.3103	0.3476	0.3854	0.3585	0.4096
9322	0.3097	0.3049	0.3209	0.3746	0.3661	0.3905
9324	0.3264	0.3021	0.3456	0.3613	0.3373	0.3800
9362	0.3595	0.3029	0.3252	0.3918	0.3621	0.3959
9637	0.3195	0.3076	0.3391	0.3585	0.3319	0.3755
9655	0.3306	0.3045	0.3176	0.3627	0.3435	0.3835
9701	0.3839	0.3046	0.3074	0.4360	0.3992	0.4383
9726	0.3311	0.3064	0.3075	0.4185	0.3989	0.4258
9742	0.3410	0.3119	0.3250	0.3941	0.3682	0.4203
9780	0.3377	0.3023	0.3046	0.4706	0.4483	0.4610
AVERAGE	0.33293	0.30417	0.32312	0.39494	0.37182	0.40865
STDEV	0.02364	0.00611	0.01525	0.03464	0.03355	0.02670

Discussion and Future Work

This study aimed to verify the hypothesis that during sentence comprehension, people adjust the word meanings according to the combination of the concepts that occur in the sentence. This effect had been demonstrated in individual cases before, and the goal was to demonstrate it more broadly across many subjects, and entire corpus of sentences, and different semantic case roles in the sentence. The correlation results indeed demonstrated that the effect is robust, and can be quantified by analyzing fMRI images through the FGREP mechanism.

These findings are significant considering that the dataset was limited and was not designed to answer the question of dynamic effects in meaning. In the future, it may be possible to extend the data with identical contexts and contrasting contexts, and such fully balanced stimuli could be used to test the hypothesis more systematically.

Similarly, it would be desirable to extend the data with fMRI images of individual words. The current approach of synthetic words (SynthWords) is an approximation often used in computational linguistic (Baroni et. al., 2010; Burgess, 1998; Landauer et al., 1997; Mitchell & Lapata, 2010) and neural activity prediction research (Anderson, et al., 2016; Binder, et al., 2016a, 2016b; Just, et al., 2017). The FGREP process of mapping semantic CARs to SynthWords and further to sentence fMRI, refines the synthetic representations by removing noise. Still, such representations blend the meanings of many words in many sentences, therefore including word fMRI should lead to stronger and clearer results.

One important advantage of CAR theory is that it is grounded on brain representations, and therefore a good choice when mapping semantic representations to fMRI. In the future, it would be interesting to compare whether similar effects can be observed with semantic representations based on co-occurrence in text corpora, or perhaps even a combination of the two. Another important direction of future work is to take advantage of this effect in an artificial natural language processing system. The vector representations for words can be modified dynamically based on context. Such a process should match human behavior better, and result in a more effective and robust system.

Conclusion

This paper shows how word meanings change dynamically depending on context. Using FGREP as a mechanism it was possible to show that the difference between the expected and observed fMRI images can indeed be explained by a change in CARs. Across an entire corpus of sentences, the new CARs are more similar to the other words in the sentence than to the original CARs, demonstrating how features of the context are transferred to each word in the sentence. In the future it may be possible to utilize such dynamic representations in an artificial natural language processing system, by making the word embeddings more sensitive to the semantic meanings that humans actually perceive.

Acknowledgments

We would like to thank Jeffery Binder (Medical College of Wisconsin), Rajeev Raizada and Andrew Anderson (University of Rochester), Mario Aguilar and Patrick Connolly (Teledyne Scientific Company) for their work and valuable help regarding this research. This work was supported in part by IARPA-FA8650-14-C-7357 and by NIH 1U01DC014922 grants.

References

Aguirre-Celis, N., Miikkulainen R. (2017). From Words to Sentences & Back: Characterizing Context-dependent Meaning Rep in the Brain. Proc.39th Annual Meeting of the Cognitive Science Society, London, UK. 1513-1518.

Aguirre-Celis N., Miikkulainen R. (2018) Combining fMRI Data and Neural Networks to Quantify Contextual Effects in the Brain. In: Wang S. et al. (Eds.). Brain Informatics. BI 2018. Lecture Notes in Computer Science. 11309, 129-140. Springer, Cham.

Anderson, A. J., Binder, J. R., Fernandino, L., Humpries C. J., Conant L. L., Aguilar M., Wang X., Doko, S., Raizada, R. D. (2016). Perdicting Neural activity patterns associated with sentences using neurobiologically

motivated model of semantic representation. *Cer. Cortex*. 1-17. DOI:10.1093/cercor/bhw240.

Baroni, M., Murphi, B., Barbu, E., Poesio, M. (2010). Strudel: A Corpus-Based Semantic Model Based on Properties and Types. *Cognitive Science*, 34(2), 222-254.

Barsalou, L. W. (2008). Grounded Cognition. *Annual Review of Psychology*, 59, 617-845.

Binder, J. R., Desai, R. H., Graves, W. W., Conant L. L. (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex*, 19:2767-2769.

Binder, J. R., Desai, R. H. (2011). The neurobiology of semantic memory. *Trends Cognitive Sci*, 15(11):527-536.

Binder, J. R., Conant L. L., Humpries C. J., Fernandino L., Simons S., Aguilar M., Desai R. (2016a). Toward a brain-based componential semantic representation. *Cog. Neuropsychology*, 33:(3-4), 130-174.

Binder, J. R. (2016b). In defense of abstract conceptual representations. *Psychonomic Bulletin & Review*, 23.

Burgess, C. (1998). From simple associations to the building blocks of language: Modeling meaning in memory with HAL model. *Behavior Research Methods, Inst. & Com.*, 30, 188-198.

Cox, R.W. (1996). AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Comput Biomed Res.*, 29:162-173.

Gagné, R., Shoben, E. (1997). Influence of Thematic Relations on the Comprehension of Modifier-Noun Combinations. *Journal of Experimental Psychology Learning Memory and Cognition*, 23, 71-87 DOI: 10.1037/0278-7393.23.1.71

Glasgow, K., Roos, M., Haufler, A. J., Chevillet, M., A., Wolmetz, M. (2016). *Evaluating semantic models with word-sentence relatedness*. arXiv:1603.07253.

Harris, Z. (1970). Distributional Structure. *In Papers in Structure and Transformational Linguistics*, 775-794.

Just, M.A., Wang, J., & Cherkassky, V. (2017). Neural representations of the concepts in simple sentences: Concept activation prediction and context effects. *NeuroImage*, *157*, 511-520.

Kiefer M, Pulvermüller F. (2012). Conceptual representations in mind and brain: theoretical developments, current evidence and future directions. *Cortex*. 48, 805–825.

Mitchell J, Lapata M. (2010). Composition in distributional models of semantics. *Cogn Sci*. 38(8), 1388–1439. DOI: 10.1111/j.1551-6709.2010.01106.x

Landauer, T.K., Dumais, S. T. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104, 211-240.

Miikkulainen, R., Dyer, M., G. (1991). Natural Language Processing with Modular PDP Networks and Distributed Lexicon. *Cognitive Science*, 15, 343-399.

Murphy, G. (1990). Noun Phrase Interpretation and Conceptual Combination. *Journal of Memory and Language*, 29, 259-288.

Pecher, D., Zeelenberg, R., Barsalou, L. W. (2004). Sensorimotors simulations underlie conceptual representations: Modality-specific effects of prior activation. *Psychonomic Bulletin & Review, 11*, 164-167.

Wisniewski, E., Property Instantiation in Conceptual Combination. (1998). *Memory & Cognition*, 26, 1330-1347. https://doi.org/10.3758/BF03201205

Numerosity capture of attention

Santiago Alonso-Díaz[a], **Jessica F. Cantlon**[b]
[a]Department of Economics, Universidad Javeriana, COL, 110231 alonsosantiago@javeriana.edu.co
[b]Department of Psychology, Carnegie Mellon, US, 15213, jcantlon@andrew.cmu.edu

Abstract

Numerosity is informative for living organisms. It can transmit, among many things, amount of food available, heading direction of the troop, which group could win a territorial dispute, the decision of were to build a beehive. Given its ecological importance, we test the hypothesis that numerosity captures visual selection. In five experiments we confirmed that an irrelevant visual stimulus that was numerically large slowed down participants in detecting a task-relevant visual target (Exp. 1 and 2). This capture was not driven by sensory variables that could correlate with numerosity: cumulative area (Exp. 3) and element size (Exp. 4). We also confirmed that the underlying numerosity representations were analogue, not set-based (Exp. 5). In a crowded visual scene numerosity is a relevant cue for visual selection, but represented only in approximate/coarse fashion.

Keywords: Attention; Attention capture; Numerosity

Introduction

Numbers can guide visual selection (Hamilton, Mirkin, & Polk, 2006; Reijnen, Wolfe, & Krummenacher, 2013; Sobel, Puri, & Faulkenberry, 2016; Utochkin, 2013). Imagine going to a crowded town fair for the first time, with different novel attractions. Your decision on where to look will be affected by the number of people around each attraction. Number is a natural and intuitive cue for behavior in uncertain contexts (Arganda, Pérez-Escudero, & de Polavieja, 2012).

A recent review proposed a list of features that could guide attention in visual search and placed them in a scale with five levels of certainty (Wolfe & Horowitz, 2017). The "undoubted guiding attributes" were color, motion, orientation, and size. On the lower side of the scale, the "probably not guiding attributes" were, among others, material type, blur, optic flow, and 3D objects. Importantly, our feature of interest, namely numerosity, was on the third level of certainty: "Possible guiding attributes". This means that even though there are some indications in the literature that it is a guiding feature, more research is required.

A classic task to study attention capture is the additional singleton search task (Theeuwes, 1992). This is a visual search task in which participants have to locate a distinct shape, say a diamond, among many other homogenous shapes present in the visual field, say circles. All the shapes have a line segment inside and subjects must report the orientation of the line in the distinct shape. The main experimental manipulation is that in a set of trials one of the homogenous shapes is turned into a distractor, usually by coloring it differently (e.g. all the shapes are green, including the target, but one is red, the distractor). The notable result is that response times are slower when there is a distractor, suggesting interference in the visual selection of the target. Moreover, the singleton search task is a compound task: participants perceive shape but report line orientation thus the effect is due to perceptual interference not response difficulty.

In a series of experiments we modified the singleton search task and created a distractor by placing more lines inside one of the non-target circles (Fig. 1; Exp. 1) or making the target more numerous while displaying a shape distractor (Exp. 2). A slower response time in the former and no distraction in the latter would indicate spontaneous capture of attention by numerosity.

We further explored whether equating total whiteness inside each of the shapes (Exp 3) (lines were white against a black background) or reducing element size/width could modify the effect (Exp. 4). The overall results indicate that the presence of number capture is robust to those perceptual features and they are consistent with the idea that number is a perceptual dimension guiding visual selection on its own terms (Anobile, Cicchini, & Burr, 2016).

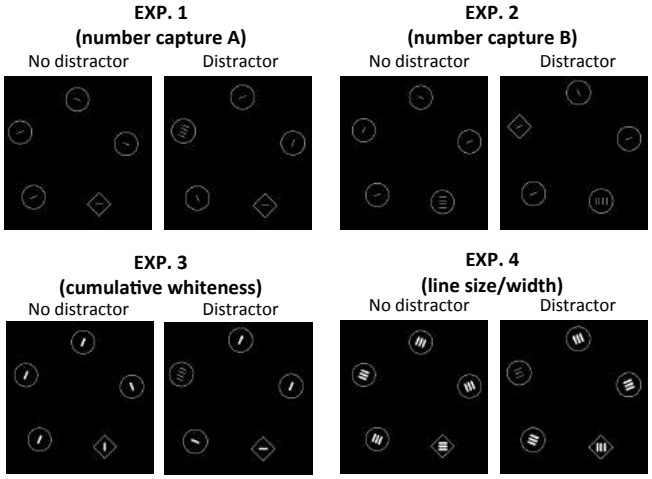


Figure 1: Tasks. In experiments 1,3, and 4 participants had to report the orientation of the line(s) inside the diamond shape (vertical or horizontal). In experiment 2 they reported the orientation of the more numerous one i.e. shape is the distractor. In half of the trials there was a distractor (counterbalanced blocked design). There were 3 different set sizes: 3, 5 (presented here), and 7.

Experiments 1-4: Numerosity Guides Visual Selection

Methods

All experimental procedures adhered to university standards, as approved by the Research Subjects Review Board. For each experiment we aimed to recruit 10 subjects, based on sample sizes of similar attention capture studies (Theeuwes, 2010).

Participants. 42 university students participated in four experiments (26 females, mean age: 21.21 years, s.d.: 3.43. We assigned 10 to each experiment; 2 were dropped due to lack of task enhancement (sleepiness and high error rate). They received $10 as compensation. The task took approximately 60 minutes, including instructions.

Stimuli. Display elements were equally spaced around a fixation point of an imaginary circle (3.4° in radius). Each display element was either a circle (1.4° in diameter) or a diamond (1.4° on each side). Inside each shape there was one or four line segments (0.42° in length) randomly oriented. The orientation inside the target was not random; it could be either vertical or horizontal. Shapes and lines were white on a black background. Participants saw three different set sizes: 3, 5, or 7 shapes equally distributed across trials (Fig. 1 has examples of set size 5).

Procedure. Subjects sat 50 cm from screen and placed their head on a chin rest. Each trial began with a fixation cross and eyes were monitored with an EyeLink 1000 desktop mount system. Images only appeared if fixation was confirmed. After a random fixation time (700 ms – 1700 ms), the fixation-cross disappeared and the shapes became visible. Set size changed randomly on each trial, as well as the position of the target and distractor. The task was to report the orientation of the lines in the target using 'z' and '/' in a qwerty keyboard to indicate vertical or horizontal, respectively. In experiment 1,3, and 4 the target was the diamond shape, and in experiment 2 the shape with more lines inside. Distractors were number (Exp. 1 and 3), a diamond shape (Exp. 2), or line width (Exp 4) (Fig. 1). Instructions emphasized a quick but accurate response. If a response was not detected after 1200 ms., the display images disappeared, the trial aborted, and a reminder text indicated that the response was too slow.

There were 240 training trials and 300 test trials with four resting breaks. Training and test trials were identical but we only analyzed test trials. The objective of training was to make subjects as fast as possible. Trials were blocked. One half had no distractor and the other did. Half of the subjects started with no distractor. Before starting, participants received an explanation of the blocked design and saw example images of each block with the main elements (target and distractor) pointed out. When a new block started, an on-screen instruction reminded participants whether there was going to be a distractor or not.

Data analysis. We analyzed each experiment individually using repeated measures ANOVAs on response times. To statistically compare effect sizes across experiments, we bootstrapped the distribution of effect size differences and compute a 95% confidence interval (samples = 1000) (Kirby & Gerlanc, 2013). For effect size we used the generalized Eta squared of the ANOVAs, suited for repeated measures analysis (Bakeman, 2005). No response time outlier detection was implemented as all trials were forced to last less than 1200 ms (see Procedures above). We report correct trials in the main text (error rates were low). All analysis were done in R.

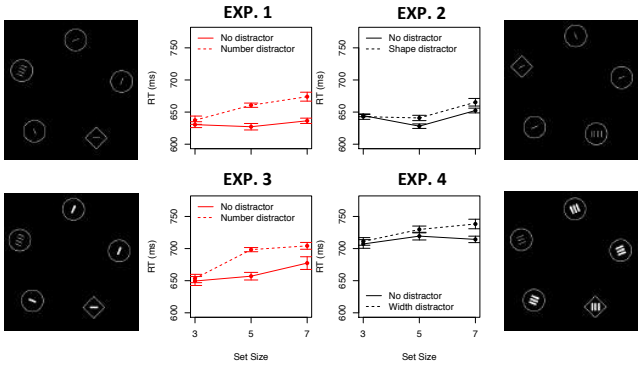


Figure 2: Experiment results. Alongside each plot there is an example image of the corresponding distracting condition (set size 5). Plots are colored red for significant distractor effects ($p < 0.05$). Bars are within subject standard errors (Cousineau et al., 2005).

Results

The presence of a number distractor increased response times in participants of Exp. 1 (Fig. 2). A repeated measures ANOVA on response times found main effects of distractor ($F(1, 9) = 41.138$, $p < 0.001$, $\eta_g^2 = 0.099$), set size ($F(2, 18) = 4.166$, $p = 0.032$, $\eta_g^2 = 0.046$), and their interaction ($F(2, 18) = 4.998$, $p = 0.018$, $\eta_g^2 = 0.029$). The slope of response time in no distractor trials is indistinguishable from zero (1.73 ms per shape; Table 1) and when there is a distractor it increases (9.31 ms/shape) causing the interaction effect. These slopes are really shallow suggesting that the diamond shape can be located in parallel when there is no distractor and even when there is a distractor the detection is much faster than a traditional serial process (Bacon & Egeth, 1994).

In the next experiment we aimed to check if capture occurred due to the generic presence of structured, but irrelevant, information in the visual field. With the same stimulus a different set of participants did the mirror task of Experiment 1: report the line orientations of the circle with more lines and be distracted by the diamond shape (Fig. 1). This time there was no significant attention capture (Fig. 2; distractor: $F(1, 9) = 0.651$, $p =$

0.440, η_g^2 = 0.002, set size: F(2, 18) = 18.914, p > 0.001, η_g^2 = 0.017, interaction: F(2, 18) = 1.038, p = 0.374, η_g^2 = 0.002). Even though not significant, there was still a minimal distraction in Exp. 2 in the same direction as Exp. 1 (Fig. 2) In such cases is important to statistically compare effect sizes (Nieuwenhuis, Forstmann, & Wagenmakers, 2011). We bootstrapped the difference of the effect sizes (η_g^2) of the distractor in both experiments (Exp. 1 minus Exp. 2). The obtained 95% confidence interval is positive [0.037, 0.142], meaning that the effect of distractor is highly unlikely to be larger in Exp. 2., confirming that subjects were at most weakly distracted by shape.

This is not saying that number is uniquely special. In a supplemental experiment we found that a square shape can also capture attention and previous work has established that forms are attractive (Theeuwes, 1992). The unique finding of Exp. 1 and 2 is that sensory stimulation was identical but when human observers are asked to find shape they are distracted by number but not vice versa. This asymmetry is not self-evident as in both versions number and shape are irrelevant for orientation detection.

An alternative explanation for the asymmetry is that in Exp. 2 the distractor was a shape which has nothing to do with the target (lines) and so is less distracting. In Exp. 1, on the other hand, the distractor were lines and the task was to detect orientation of lines, and so is more distracting. However, we selected the Theeuwes task precisely to avoid such confounds. Participants need to detect the relevant feature, shape or number, and then report the orientation. The alternative strategy of trying to directly detect line orientations in this type of task has been shown to be too inefficient (Theeuwes, 2010). That being said, if the alternative explanation holds, our result would implicate that numerosity breaks the strategy of detecting the feature and reporting the orientation; an interesting finding on its own terms that does not invalidate Exp. 1 findings.

In our stimulus capture seems to be driven by a parser that detects more lines. During training and between blocks participants were reminded that the distractor had more segments. And, prefacing the next set of experiments, attention capture was not detectable when numerosity was equal (Exp. 4). It only appeared when there was an increase in the number of lines (Exp. 3).

The next pair of experiments probe with more detail the sensory aspects of the more numerous lines that could have mobilized attention. In Exp. 3 we equated total amount of whiteness in all shapes by making single segments four times thicker (Fig. 1). If the observed number capture in Exp. 1 is due to an overall integration of whiteness (cumulative area/brightness) then distraction should disappear. This was not observed. There were detectable interferences of the irrelevant more numerous location (Fig. 2; distractor: F(1, 9) = 4.137, p = 0.072, η_g^2 = 0.040, set size: F(2, 18) = 8.702, p = 0.002, η_g^2 = 0.071, interaction: F(2, 18) = 4.618, p = 0.024, η_g^2 = 0.016). A direct comparison of effect sizes in Exp. 1 and Exp. 3 actually includes the possibility that the distractor effect is larger when cumulative area is controlled for (95% CI of Exp. 1 minus Exp. 3: [-0.006, 0.116]). The slopes relating set size and RT were again really low (Table 1), lower than a steorotypical serial search (Bacon & Egeth, 1994; Treisman & Gelade, 1980), indicating that the task was done in partially parallel fashion. Number capture is not related in a simple manner to an attraction to overall whiteness.

Table 1. RT slopes

	Exp. 1	Exp. 2	Exp. 3	Exp. 4
No distractor	1.73	1.31	6.79*	1.21
Distractor	9.31*	5.58*	11.56*	6.98*

* p<0.05

It is possible that what drove number capture in Exp. 3 was the width of the lines (Fig. 1). In Exp. 4 we fixed the number of lines inside each of the shapes and made their line width three times bigger than the one in the distractor. If line width is the critical distracting aspect in Exp. 3, then Exp. 4 should reveal attention capture. This was not observed (Fig. 2; distractor: F(1, 9) = 1.332, p = 0.278, η_g^2 = 0.007, set size: F(2, 18) = 2.767, p = 0.089, η_g^2 = 0.011, interaction: F(2, 18) = 1.079, p = 0.360, η_g^2 = 0.003). A comparison of the effect sizes of Exp. 3 and 4 indicates that distraction was more notable in the latter (95% CI of Exp. 3 minus Exp. 4: [0.004, 0.066]). Again, the slopes were really shallow suggesting an efficient search process, close to parallel (Table 1). Line width draws little attention in our visual stimulus.

Discussion

Attention is captured by numerosity, beyond basic perceptual features that could correlate with number: cumulative area/whiteness and element size/width. This was obtained with a compound visual search task that differentiates perception from response difficulty. This is important because distractor effects can be traced back to perceptual interference and not to response interference (Theeuwes, 2010). The overall results are consistent with the idea that numerosity is a basic perceptual feature that guides attention (Anobile et al., 2016; Wolfe & Horowitz, 2017).

Previous reports have demonstrated the importance of number for attentional process. Reijnen et al., 2013 used a task where the target and distractors were numerical. However they used large numerosities and the task of participants actually required numerical estimation. Here we confirmed attentional effects with a much simpler compound visual task with small numerosities.

Utochkin, 2013 found that numerosity guides attention as an aide to find perceptual features, in their case color. Thus, numerosity was actually useful in their task. Our Exp. 1 - 4, numerosity was irrelevant and as such is closer to the notion of attention capture.

Attention capture is usually framed around the conceptual dichotomy of bottom-up or top down sources of the observed distraction (Theeuwes, 2010). However, the notion of priority maps, a working space that integrates current goals, selection history, physical salience, is perhaps more relevant (Awh, Belopolsky, & Theeuwes, 2012). For our purposes, number must induce a priority signal and be a relevant source of information for the nervous system to be able to capture attention. Visual selection would emulate other decision contexts in which numerosity is routinely used, mostly as an heuristic to solve complex uncertain choices (Gigerenzer & Brighton, 2009; Reyna & Brainerd, 2008).

There is great deal of debate on the abstract or sensory nature of number (Anobile et al., 2016; Gebuis, Kadosh, & Gevers, 2016; Leibovich, Katzin, Harel, & Henik, 2017) . We argue that the number capture observed here is consistent with the proposal that number is abstract and a basic perceptual feature. First, the sensory aspects evaluated (cumulative whiteness and element size/width) failed to capture attention. Second, the shallow slopes relating set sizes and response times were not so different from previous attention capture studies using other basic perceptual stimulation (e.g. color) ((Bacon & Egeth, 1994; Theeuwes, 1992). They were not necessarily different from zero to claim any preattentive mechanism, but they are certainly really close to those previous works that demonstrated attention capture from basic features.

There are at least three limitations of our study. First, we did not control for line separation, which may be a feature driving attention in our task. If line separation means frequency then we are not sure how to distinguish frequency from number as they would correlate perfectly. Also, even though we cannot rule out that possibility, a recent review on features that have been found to guide attention did not report line separation (Wolfe & Horowitz, 2017).

The second limitation is that we did not control for overall contrast. We manipulated line width to control for cumulative area effects (Exp. 3 and 4) and the number distractor ended up looking more dim (Fig. 1). We would argue that this actually made our results more robust because it is not about higher contrast. Still, it would have been interesting to determine how much of the effect changes with different contrast levels.

The third and final limitation is that attention may have been driven by the presence of a texture formed by the patch with more lines. However, we would argue that texture is a vague term and we narrowed down on an aspect, namely numerosity. Also, texture is obtained preattentively (Julesz, 1981) and search slopes in Exp. 1-4 were different from zero.

In general, as with most studies of numerosity, it is almost impossible to discard 100% that our results are not influenced by a preattentive sensory features. They may indeed have a role in the underlying effect but we think that there is sufficient evidence in the literature to believe that number is a basic sensory aspect (Anobile et al., 2016); and we think our results add to that line of research.

In many behavioral contexts numerosity is a basic heuristic that hinders or facilitates learning and decision-making (Gigerenzer & Brighton, 2009; Reyna & Brainerd, 2008). Also, the approximate number system seems to influence higher order behavior such as risk attitudes and math scores (Halberda, Mazzocco, & Feigenson, 2008; Schley & Peters, 2014). Our study furthers the link between numerosity and attention which may provide clues on why raw numerosity is such a strong driver of learning and behavior.

References

Anobile, G., Cicchini, G. M., & Burr, D. C. (2016). Number as a primary perceptual attribute: A review. *Perception*, *45*(1-2), 5–31.

Arganda, S., Pérez-Escudero, A., & de Polavieja, G. G. (2012). A common rule for decision making in animal collectives across species. *Proceedings of the National Academy of Sciences*, *109*(50), 20508–20513.

Awh, E., Belopolsky, A. V., & Theeuwes, J. (2012). Top-down versus bottom-up attentional control: A failed theoretical dichotomy. *Trends in cognitive sciences*, *16*(8), 437–443.

Bacon, W. F., & Egeth, H. E. (1994). Overriding stimulus-driven attentional capture. *Perception & psychophysics*, *55*(5), 485–496.

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior Research Methods*, *37*(3), 379–84.

Burr, D. C., Anobile, G., & Turi, M. (2011). Adaptation affects both high and low (subitized) numbers under conditions of high attentional load. *Seeing and Perceiving*, *24*(2), 141–150.

Cousineau, D. et al. (2005). Confidence intervals in within-subject designs: A simpler solution to loftus and masson's method. *Tutorials in quantitative methods for psychology*, *1*(1), 42–45.

Gebuis, T., Kadosh, R. C., & Gevers, W. (2016). Sensory-integration system rather than approximate number system underlies numerosity processing: A critical review. *Acta psychologica*, *171*, 17–35.

Gigerenzer, G., & Brighton, H. (2009). Homo Heuristicus: Why Biased Minds Make Better Inferences. *Topics*

in Cognitive Science, *1*(1), 107–143. doi:10.1111/j.1756-8765.2008.01006.x

Halberda, J., Mazzocco, M. M. M., & Feigenson, L. (2008). Individual differences in non-verbal number acuity correlate with maths achievement. *Nature*, *455*(7213), 665–8.

Hamilton, J. P., Mirkin, M., & Polk, T. A. (2006). Category-level contributions to the alphanumeric category effect in visual search. *Psychonomic Bulletin & Review*, *13*(6), 1074–1077.

Julesz, B. (1981). A theory of preattentive texture discrimination based on first-order statistics of textons. *Biological Cybernetics*, *41*(2), 131–138.

Kirby, K. N., & Gerlanc, D. (2013). Bootes: An r package for bootstrap confidence intervals on effect sizes. *Behavior research methods*, *45*(4), 905–927.

Leibovich, T., Katzin, N., Harel, M., & Henik, A. (2017). From "sense of number" to "sense of magnitude": The role of continuous magnitudes in numerical cognition. *Behavioral and Brain Sciences*, *40*.

Nieuwenhuis, S., Forstmann, B. U., & Wagenmakers, E.-J. (2011). Erroneous analyses of interactions in neuroscience: A problem of significance. *Nature neuroscience*, *14*(9), 1105.

Reijnen, E., Wolfe, J. M., & Krummenacher, J. (2013). Coarse guidance by numerosity in visual search. *Attention, Perception, & Psychophysics*, *75*(1), 16–28.

Reyna, V. F., & Brainerd, C. J. (2008). Numeracy, ratio bias, and denominator neglect in judgments of risk and probability. *Learning and Individual Differences*, *18*(1), 89–107. doi:10.1016/j.lindif.2007.03.011

Schley, D. R., & Peters, E. (2014). Assessing "economic value": symbolic-number mappings predict risky and riskless valuations. *Psychological science*, *25*(3), 753–61. doi:10.1177/0956797613515485

Sobel, K. V., Puri, A. M., & Faulkenberry, T. J. (2016). Bottom-up and top-down attentional contributions to the size congruity effect. *Attention, Perception, & Psychophysics*, *78*(5), 1324–1336.

Theeuwes, J. (1992). Perceptual selectivity for color and form. *Perception & psychophysics*, *51*(6), 599–606.

Theeuwes, J. (2010). Top–down and bottom–up control of visual selection. *Acta psychologica*, *135*(2), 77–99.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive psychology*, *12*(1), 97–136.

Utochkin, I. S. (2013). Visual search with negative slopes: The statistical power of numerosity guides attention. *Journal of vision*, *13*(3), 18–18.

Wolfe, J. M., & Horowitz, T. S. (2017). Five factors that guide attention in visual search. *Nature Human Behaviour*, *1*(3), 0058.

Intrinsic whole number bias in an indigenous population

Santiago Alonso-Díaz[a], Jessica F. Cantlon[b], & Steven T. Piantadosi[c]
[a]Department of Economics, Universidad Javeriana, COL, 110231 alonsosantiago@javeriana.edu.co
[b]Department of Psychology, Carnegie Mellon, US, 15213, jcantlon@andrew.cmu.edu
[c]Department of Psychology, Berkeley University, US, 94720, stp@berkeley.edu

Abstract

Probabilities can be described by a numerator and a denominator and students and decision-makers are not indifferent to numerical values of the components. For instance, when people compare two equal ratios their choices gravitate to the option with larger number, even if they know both ratios are equal. To the date, however, it is unclear if whole number biases are present in other cultures. We tested a farming-foraging group living in the Bolivian rain forest in a simple 2AFC ratio comparison task. After appropriate training, the Tsimane were highly accurate in this task, confirming that visual proportional reasoning is present across cultures. Importantly, they had a strong tendency to favor large numbers in equal ratio comparisons, similar to what is found in educated populations. Even though our sample size is moderate (n=76), the whole number bias we found occurred under good proportional reasoning. The bias may be a general feature of cognition, rather than a cultural or education artifact, that may help humans solve ambiguous situations.

Keywords: Tsimane; Numerical cognition; Fraction; Probability; Whole number bias

Introduction

Detecting differences in discrete visual ratios is useful. They convey a variety of critical information, like how much units of food there is available per competitor or heading direction of a troop by a majority rule (Real, 1993; Strandburg-Peshkin, Farine, Couzin, & Crofoot, 2015). Infants, indigenous population without formal education, and non-human primates can act upon probabilities expressed by visual proportions (Denison & Xu, 2010; Fontanari, Gonzalez, Vallortigara, & Girotto, 2014; Rakoczy et al., 2014). The spontaneous mapping of visual ratios to probabilities in the context of no formal education suggests that this is a core cognitive feature akin to detectors of abstract numerosity and geometry relations found across cultures and species (Carey & Spelke, 1994; Spelke & Lee, 2012)

Discrete probability comparisons, however, suffer from numerosity interferences (Reyna & Brainerd, 2008). It is much easier to compare ratios when the largest one happens to have the larger numerosity. It is unclear if this is caused by cultural characteristics shared by Western, Educated, Industrialized, Rich, and Democratic people (WEIRD) (Henrich, Heine, & Norenzayan, 2010) or if sticking to numerosity is a general feature of how quotients are compared in the mind (Alonso-Diaz, Piantadosi, Hayden, & Cantlon, 2018). The latter option is what we call an intrinsic whole number bias: a pull towards numerical magnitude even though ratio estimates are available. The presence of ratio estimates is critical because it distinguishes it from denominator-neglect or any other strategy used to cover up the inability to compute the value of the fraction.

Previous work probing proportional reasoning in non-WEIRD people, found that the Kaqchikel and K'iche', two indigenous Mayan groups in Guatemala, had refined probabilistic abilities in the absence of formal probability education (Fontanari et al., 2014). Of importance, one of the experiments (Exp. 2) revealed that proportional reasoning was not affected by the numerosity of the options. Participants excelled in comparing 0.25 against 0.75, both when the larger probability had more or fewer number of winners.

Experiment 2 of Fontanari et al., 2014 established probabilistic cognition with no formal education but there were no indications in their study of a whole-number bias, and their analyses nor experimental design tried to uncover one. In fact, to the best of our knowledge, there is no evidence of the whole number bias outside WEIRD populations (perhaps in other species, but not across the WEIRD-NON WEIRD divide). There are at least three hypothesis. Our hypothesis is that it should be similar in NON-WEIRD humans because is a reflection of the inner workings of basic perceptual proportional choice (Alonso-Diaz et al., 2018). A second hypothesis is that the whole-number bias is a mistake caused by deficient education (Reyna & Brainerd, 2008). Under this hypothesis, the whole-number bias should be notably stronger in the NON-WEIRD humans because they lack formal education on probability principles. The third and final hypothesis is that the bias only appears in WEIRD humans because of specific cultural practices (e.g how they learn probabilities and fractions).

We tested a 2AFC ratio comparison task in the Tsimane', a farming-foraging group living in the Bolivian rain forest (Huanca, 2008). A wealth of studies have been done on the Tsimane's cognitive and decision making processes (Apaza et al., 2003; Apaza et al., 2002; Godoy & Jacobson, 1999; Godoy, Jacobson, & Wilkie, 1998; Henrich et al., 2010; Kirby et al., 2002; McDermott, Schultz, Undurraga, & Godoy, 2016; Piantadosi, Kidd, & Aslin, 2014; Reyes-Garcıa et al., 2003). Their aptitude to probabilistic cognition, however, has not been properly researched. The Tsimane are fairly isolated, with low literacy, and no formal instruction on probability principles. We hypothesized the existence of probabilistic reasoning in the Tsimane. Perhaps more important, a detectable bias towards more numerous options in equal ratio trials.

To detect a whole-number bias, we will exploit the fact that when ratios are equal participants should be indifferent to the numerosity of the options and pick randomly; but this is not observed empirically (Denes-Raj, Epstein, & Cole, 1995) even when the proportions are known to be equal (Alonso-Diaz et al., 2018). To be clear, the bias is not exclusive to equal ratio trials. Also, we are not suggesting that only on them probabilistic reasoning fails. The bias towards larger numerosities is intricate and with many explanations (Alonso-Diaz et al., 2018). We are using equal ratio trials as a methodological tool to detect the bias in an indigenous population.

The originality of our work is that we seek an intrinsic whole number bias, one that is detected under appropriate probabilistic reasoning (Alonso-Diaz et al., 2018). To prove good reasoning we will use congruent (the larger probability has larger numerosity), incongruent (the larger probability has smaller numerosity) and equal ratio trials. Congruent and incongruent trials will help us discard illusory Stroop effects by which the irrelevant dimension of numerosity could affect ratio estimates by changing the subjective psychophysical properties of the alternatives (Barth, 2008). In simple words, if the Tsimane are successful in both congruent and incongruent trials we can be sure that they tried to pick the best ratio, not the one with more numerosity.

Their choice on equal ratio trials will be a metric on how intense the bias is. If it is considerably larger or smaller, then we can conclude that cultural practices (e.g. formal education) affect the bias. If it is similar, then it is consistent with being a generic human adaptation (Alonso-Diaz & Penagos, under review).

Methods

The study procedures were approved by the Gran Consejo Tsimane' (Tsimane' grand council), as well as institutional IRBs. Tomás Huanca and the Centro Boliviano de Investigación y de Desarrollo Socio Integral (CBIDSI) provided logistic support (translators, transportation, and general expertise about the Tsimane community).

Participants. We evaluated two groups of Tsimane. This was not an explicit design strategy but rather reflects the dynamics of field-work (details below). The first group received verbal instructions in their native language (n=86, 60 females, M age = 34.13 years, s.d. = 15.09, Education M = 3.18 years, s.d. = 3.28). The second received non-verbal training version (n=78, 53 Females, Age M = 31.884 years; s.d. = 14.528; Education M = 4.012 years, s.d. = 4.037). 76 Tsimane succeed non-verbal training (two subjects failed the training stage). We only present the results for the Tsimane who did non-verbal training (see Alonso-Diaz, 2017 for the verbal-training sample). Each Tsimane did many cognitive tasks sequentially including language, numerosity, color perception, and the probability task reported here.

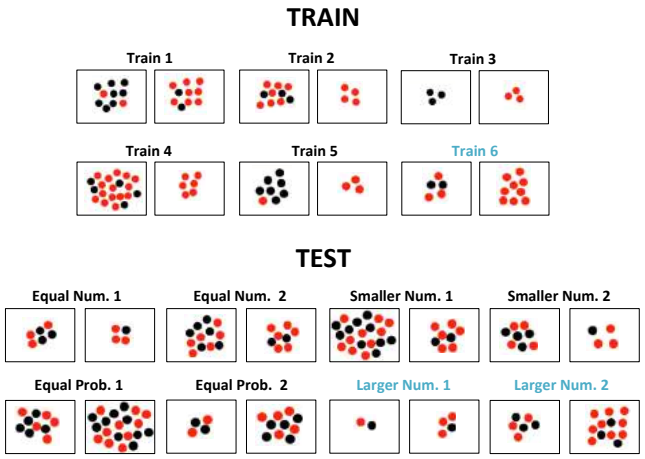


Figure 1: Training and test trials. The ratios of winners (red) to losers (black) in Test trials were 0.5 and 0.75. The titles in the Test images indicate whether the largest ratio had equal, smaller, or larger number of winners or if both options had equal probability. Cyan titles indicate stimulus that just a subsample of Tsimane observed (n=24). If a participant is successful in all control trials (Equal, Smaller, and Larger Num.), we would be more confident that the bias in Equal Prob. trials is intrinsic i.e. it can appear under proper proportional reasoning.

Materials and Procedure. Participants saw two images, one to the left and another to the right side. Each image was presented in individual laminated sheets (legal size) that contained a mix of red and black dots (Fig. 1). Participants had to select the option with best chances of winning (red). The best option was randomly placed on either side. The Tsimane heard a verbal instruction in their native language.

The behavior of the initial 86 Tsimane (those who only received verbal instructions) was hard to classify as either following ratio or numerosity (an analysis of this subsample is provided in Alonso-Diaz, 2017). To make sure it was not related to translation issues, halfway during field research we included non-verbal training with feedback. After verbal instructions we randomly presented six pair of training images until all were correct i.e. we cycled through them until all responses were correct and most Tsimane were quick dispatching training. Training trials were mostly trivial (5 out of 6) in that one side only had winners (Fig. 1), randomly placed to the left or right side. The intention of trivial trials was to deter number-based strategies: the correct option had the same, fewer, or more winners than the wrong side. We presented test trials (Fig. 1) in pseudo-random order with no feedback.

Of the 76 participants with non-verbal training, 52 did three types of test trials: 1) both ratios had equal num-

ber of winners, 2) the best ratio had smaller number of winners, and 3) ratios were identical but one had more winners (Figure 1). To further discard a strategy of low-number of losers (the confound present in Fontanari et al., 2014 Maya's study), the last 24 Tsimane saw the same images as the 52 but also new Test images with identical number of losers (blacks) (and also one more training image, Fig. 1 cyan color).

Data analysis. We will use the following acronyms: EN = both ratios had equal numerator; SN = larger ratio had smaller numerator; LN = larger ratio had larger numerator; EP = both images had equal probability. In EP accuracy reflects the proportion of choices favoring the option with larger numerosity

Binomial tests evaluated if performance was greater than chance. In control trials (EN, SN, and LN), chance means picking the larger ratio more than 50% of the times. In test trials (EP), chance means picking the option with larger numerosity more than 50% of the times. In the binomial tests we used the total number of choices. Because each Tsimane made two choices on each trial type (Fig. 1), n = 2 times sample size.

We classified each Tsimane's behavior according to one of the following potential strategies: consistently picked A) More winners; B) Fewer winners; C) More total number of balls; D) Fewer total number of balls; E) More losers; F) Fewer losers; G) Larger ratio; H) Other. Some behaviors were ambiguous as they could be consistent with more than one strategy. For instance, in the stimuli presented to the subsample of 52 Tsimane (the one similar to Fontanari et al., 2014), being correct in all trials and selecting the option with fewer losers when both ratios were equal will necessarily occur if the agent decides based on fewer losers or the larger ratio. When such coding conflicts occurred, we used the unambiguous behavior. In the example, we would code the Tsimane as following a strategy that picks fewer losers because a strategy of only ratios will be random when both bags have equal ratio.

Results

Tsimane's accuracy in test trials was high (Fig. 2A; EN: 149/152 = 0.98 trials correct, $p < 0.001$; SN: 143/152 = 0.94, $p < 0.001$; LN: 40/48 = 0.83, $p < 0.001$; EP: 66/152 = 0.43, $p = 0.12$). Fig. 2A seems to suggest that in equal probability trials (EP) the Tsimane did not tend to pick the bag with larger numerosity. A closer look reveals that the majority of Tsimane behave in accordance to a ratio-based strategy (n = 36), followed by strategies that follow small number of balls (n = 26), other unidentifiable strategy (n = 8), low number of losers (n = 4), and large number of winners (n = 2) (Fig. 2B). The diversity of strategies is only normal in such unnatural task. Interestingly, we detected a large number bias in equal probability trials in Tsimane' whose performance was flawless in EN, SN, and LN test trials (52/72 = 0.72 trials favored the option with more winners, $p < 0.001$). As a reminder, EN, SN, and LN were control trials for a simple numerosity-based behavior. For instance, a Tsimane who had only followed large numerosities would have failed in both SN trials because in those trials the larger probability had smaller numerosity. This means that the manifestation of the whole number bias in equal ratio trials is hardly explained by a straightforward behavior based on numerical cardinalities in those who did not fail in EN, SN, and LN trials.

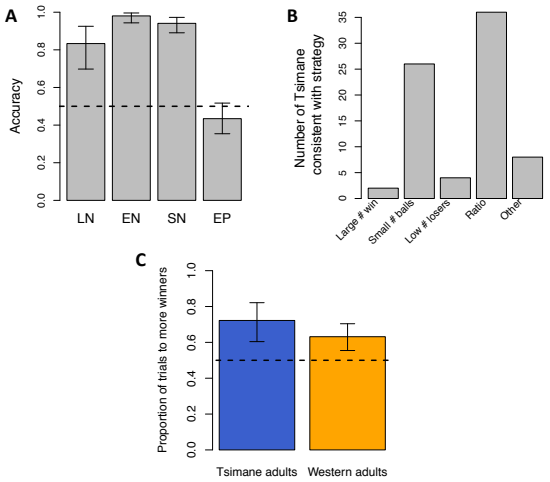


Figure 2: Accuracy (A), distribution of strategies (B), and whole number bias in the Tsimane who were reliably coded as following a ratio-based strategy (C). In A: LN: the best ratio had the large numerator (winners); EN: both ratios had equal numerator; SN: the best ratio had small numerator; EP: both options had equal probability. In EP accuracy reflects the proportion of choices favoring the option with larger numerosity. Dashed line is chance. Error bars are 95% confidence intervals of a binomial test (all $p<0.05$, except EP).

The rate of the bias is comparable to the one found in American adults doing a similar one-shot task. (Fig. 2C; χ^2 (1) = 1.469, p = 0.225; American data in (Alonso-Diaz et al., 2018)). The Tsimane whose behavior is consistent with ratio use (n=36) exhibit the same intuition to favor large numbers in ambiguous contexts. We emphasize that in the other strategies is hard to classify the bias as such because it is baked in the actual definition e.g. in a "large # win" strategy the task is solved following winners. We argue that the theoretically relevant bias is when appropriate proportional reasoning is present.

As it was mentioned in the methods, some participants did additional trials in which both images had identical number of losers. The reason for this was to discard a losers-based behavior. An analysis of these two subsamples reveals a similar pattern. In the subsample that did not see images with identical number of losers

(n=52) they had a behavior above chance in all control trials (chance means picking the larger probability) and in equal probability trials (chance means picking the the image with larger numerosity) (i.e. all binomial tests $p<0.05$) confirming the presence of a whole number bias. Of those 52, 48 had perfect performance in control trials and also revealed a whole number bias in equal probability trials albeit not significant (30/48 = 0.625, p = 0.11). This first subsample was clearly trying to solve the task through ratio-based strategies because they succeeded in SN, LN, and EN trials. However, when faced with equal ratio trials they showed a whole-number bias. This does not necessarily mean that in equal ratio trials their proportional reasoning shuts down (Alonso-Diaz et al., 2018); our experimental design cannot solve that question.

In the second subsample, we can definitely discard a loser-based strategy (n = 24). Their behavior in control trials was different than chance (i.e. all binomial tests $p<0.05$), and revealed a whole number bias in equal probability trials (22/24 = 0.91 trials favored the option with more winners, $p < 0.001$). Thus, the effect was particularly present in those who we can discard any form of number-based strategy in SN, LN, and EN trials.

A caveat of our results is that even though significant, we had a small sample size (n=76), specially the ones that we can confidently discard a number-based strategy (n=24). Future work could increase sample size, but two things make us confident of the results. First, the whole number bias is not a controversial finding (e.g. Alonso-Diaz et al., 2018; Reyna and Brainerd, 2008). Second, the bias we reported was stringent, making sure that it was present under proper proportional reasoning.

Discussion

The Tsimane', similar to other populations (Mayas, human infants, non-human primates), are capable of visual proportional reasoning. Even though the task used was artificial, based on laminated sheets, it was possible to elicit ratio-based responses. Perhaps more relevant, in ambiguous trials, in which both options had equal probability, the intuition of adult Tsimane was in line with that of adult Americans: pick the option with larger numerosity. This was not a number-based strategy induced by lack of proportional abilities as they were very capable of solving congruent (larger prob. has more winners), incongruent (larger prob. has fewer winners), and trials where the large probability had the same number of winners as the wrong alternative.

Perhaps more insightful is that the bias was comparable in size between WEIRD and NON-WEIRD samples (Fig. 2C). Because we obtained the bias under good proportional reasoning and with equal ratio trials, it suggests that numerosity could be a generic cognitive tool to solve ambiguity, not merely a quick heuristic to substitute an inability to compute ratios as previously proposed. In fact, the bias could be a sign of adaptive agents (Alonso-Diaz & Penagos, under review).

The automatic activation and use of numerical values, despite appropriate visual proportional reasoning, is confirmable through more rigorous psychophysics tasks and computational models (Alonso-Diaz et al., 2018). What is suggestive of the Tsimane results is that number intrusions may not be a WEIRD phenomenon of developed economies (Henrich et al., 2010) but the outcome of some generic computation, perhaps influenced by the fact that larger numerosities elicit a greater sense of confidence and capture attention (Alonso-Diaz, 2017; Alonso-Diaz & Cantlon, 2018).

The intuition of relying in numerosities is usually observed during learning and manipulation of symbolic fractions (Ni & Zhou, 2005; Siegler, Fazio, Bailey, & Zhou, 2013). At the same time, there is growing evidence that perceptual and symbolic systems are not independent (Melnick, Harrison, Park, Bennetto, & Tadin, 2013), for instance the approximate number system correlates with formal math tests (Halberda, Mazzocco, & Feigenson, 2008). It is possible, then, that the effects of number in perceptual proportional reasoning transpire to symbolic education and decision-making settings where numerosity should not be employed.

An alternative explanation of our results is that the Tsimane tested were not fully illiterate (mean years of education 4.012) and some negative pedagogical influence in those years may have impacted behavior in our task. The main problem with this interpretation is that the Tsimane succeeded in ratio comparisons with different numerosity manipulations. If anything, the contra argument is also plausible: education might have helped them in solving the task. Rather, we argue that the intuition of relying in larger numerosities is a generic feature of cognition. The human mind is endowed with probabilistic knowledge. However, the mechanisms that lead to overt probabilistic behavior do not necessarily drop the numerical values, even when holistic ratio computations are available. Number intrusions seem to be present across cultures.

Another interesting result is that the Tsimane required non-verbal training to succeed in our task. The first subsample only received verbal instructions in their native language but their performance was lower than those who received non-verbal training (see Methods). It is hard to narrow down the reasons for such difference between verbal and non-verbal instructions but it is relevant for future studies on non-WEIRD populations.

Acknowledgments

We thank the Tsimane people for their gracious help. Also, Tomás Huanca, from the Centro Boliviano de Investigación y de Desarrollo Socio Integral (CBIDSI), and

the translators who were essential to the project success. We want to thank Ted Gibson, Richard Futrell, Julian Jara-Ettinger, and Steve Ferrigno for support and advice in the field.

References

Alonso-Diaz, S. (2017). *Number representation in perceptual decisions* (Doctoral dissertation, University of Rochester, Department of Brain and Cognitive Sciences).

Alonso-Diaz, S., & Cantlon, J. F. (2018). Confidence judgments during ratio comparisons reveal a bayesian bias. *Cognition*, *177*, 98–106.

Alonso-Diaz, S., & Penagos, G. (under review). Human Adaptation to the Empirical Distribution of Relative Quantities. *Cognitive Science*.

Alonso-Diaz, S., Piantadosi, S. T., Hayden, B. Y., & Cantlon, J. F. (2018). Intrinsic whole number bias in humans. *Journal of Experimental Psychology: Human Perception and Performance*, *44*(9), 1472.

Apaza, L., Godoy, R., Wilkie, D., Byron, E., Huanca, T., Leonard, W., ... Vadez, V. T. (2003). Markets and the use of wild animals for traditional medicine: a case study among the tsimane amerindians of the Bolivian rain forest. *Journal of Ethnobiology*, *23*(1), 47–64.

Apaza, L., Wilkie, D., Byron, E., Huanca, T., Leonard, W., Perez, E., ... Godoy, R. (2002). Meat prices influence the consumption of wildlife by the Tsimane ' Amerindians of Bolivia. *Oryx*, *36*(4), 1–7. doi:10.1017/S0030605302000000

Barth, H. C. (2008). Judgments of discrete and continuous quantity: An illusory stroop effect. *Cognition*, *109*(2), 251–266.

Carey, S., & Spelke, E. (1994). Domain-specific knowledge and conceptual change. In L. A. Hirschfeld & S. A. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture* (Chap. 7, pp. 169–200). New York: Cambridge University Press.

Denes-Raj, V., Epstein, S., & Cole, J. (1995). The generality of the ratio-bias phenomenon. *Personality and Social Psychology Bulletin*, *21*(10), 1083–1092.

Denison, S., & Xu, F. (2010). Twelve- to 14-month-old infants can predict single-event probability with large set sizes. *Developmental science*, *13*(5), 798–803. doi:10.1111/j.1467-7687.2009.00943.x

Fontanari, L., Gonzalez, M., Vallortigara, G., & Girotto, V. (2014). Probabilistic cognition in two indigenous mayan groups. *Proceedings of the National Academy of Sciences*, *111*(48), 17075–17080.

Godoy, R., & Jacobson, M. (1999). Covariates of Private Time Preference: A Pilot Study Among the Tsimane' Indians of the Bolivian Rain Forest. *Evolution and Human Behavior*, *20*, 249–256.

Godoy, R., Jacobson, M., & Wilkie, D. (1998). Strategies of Rain-Forest Dwellers against Misfortunes : The Tsimane ' Indians of Bolivia. *Ethnology*, *37*(1), 55–69.

Halberda, J., Mazzocco, M. M. M., & Feigenson, L. (2008). Individual differences in non-verbal number acuity correlate with maths achievement. *Nature*, *455*(7213), 665–8.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and brain sciences*, *33*(2-3), 61–135. doi:10.1017/S0140525X0999152X

Huanca, T. (2008). *Tsimane' Oral Tradition, Landscape, and Identity in Tropical Forest*. SEPHIS, South-South Exchange Programme for Research on the History of Development.

Kirby, K., Godoy, R., Reyes-Garcia, V., Byron, E., Apaza, L., Leonard, W., ... Wilkie, D. (2002). Correlates of delay-discount rates : Evidence from Tsimane ' Amerindians of the Bolivian rain forest. *Journal of Economic Psychology*, *23*, 291–316.

McDermott, J. H., Schultz, A. F., Undurraga, E. A., & Godoy, R. (2016). Indifference to dissonance in native Amazonians reveals cultural variation in music perception. *Nature*, *535*, 547–550. doi:10.1038/nature18635

Melnick, M. D., Harrison, B. R., Park, S., Bennetto, L., & Tadin, D. (2013). A strong interactive link between sensory discriminations and intelligence. *Current Biology*, *23*(11), 1013–1017.

Ni, Y., & Zhou, Y.-D. (2005). Teaching and learning fraction and rational numbers: The origins and implications of whole number bias. *Educational Psychologist*, *40*(1), 27–52.

Piantadosi, S., Kidd, C., & Aslin, R. (2014). Rich analysis and rational models: Inferring individual behavior from infant looking data. *Developmental Science*, 1–16.

Rakoczy, H., Clüver, A., Saucke, L., Stoffregen, N., Gräbener, A., Migura, J., & Call, J. (2014). Apes are intuitive statisticians. *Cognition*, *131*(1), 60–8. doi:10.1016/j.cognition.2013.12.011

Real, L. A. (1993). Toward a cognitive ecology. *Trends in Ecology & Evolution*, *8*(11), 413–417.

Reyes-Garcıa, V., Godoy, R., Vadez, V., Apaza, L., Byron, E., Huanca, T., ... Wilkie, D. (2003). Ethnobotanical Knowledge Shared Widely Among Tsimane' Amerindians, Bolivia. *Science*, *299*, 1707.

Reyna, V. F., & Brainerd, C. J. (2008). Numeracy, ratio bias, and denominator neglect in judgments of risk and probability. *Learning and Individual Differences*, *18*(1), 89–107. doi:10.1016/j.lindif.2007.03.011

Siegler, R. S., Fazio, L. K., Bailey, D. H., & Zhou, X. (2013). Fractions: The new frontier for theories of numeri-

cal development. *Trends in cognitive sciences*, *17*(1), 13–19.

Spelke, E. S., & Lee, S. A. (2012). Core systems of geometry in animal minds. *Philosophical Transactions of the Royal Society B*, *367*, 2784–2793. doi:10.1098/rstb.2012.0210

Strandburg-Peshkin, A., Farine, D. R., Couzin, I. D., & Crofoot, M. C. (2015). Shared decision-making drives collective movement in wild baboons. *Science*, *348*(6241), 1358–1361.

Distinguishing learned categorical perception from selective attention to a dimension: Preliminary evidence from a new method

Janet Andrews (andrewsj@vassar.edu)
Department of Cognitive Science, 124 Raymond Ave.
Poughkeepsie, NY 12604 USA

Joshua de Leeuw (jdeleeuw@vassar.edu)
Department of Cognitive Science, 124 Raymond Ave.
Poughkeepsie, NY 12604 USA

Rebecca Andrews (reandrews@vassar.edu)
Department of Cognitive Science, 124 Raymond Ave.
Poughkeepsie, NY 12604 USA

Cole Landolt (clandolt@vassar.edu)
Department of Cognitive Science, 124 Raymond Ave.
Poughkeepsie, NY 12604 USA

Chrissy Griesmer (chgriesmer@vassar.edu)
Department of Cognitive Science, 124 Raymond Ave.
Poughkeepsie, NY 12604 USA

Abstract

A novel experimental method is motivated and applied in an effort to test for effects of category learning on perceptual discrimination so as to clearly distinguish category boundary effects of expansion and compression from changes in sensitivity to stimulus dimensions. The method includes a control group performing a task that, like category learning, requires attention to one systematically varying stimulus dimension rather than another. Discrimination accuracy is tracked over time and measured using a psychophysical staircase procedure tailored to individual participants that doesn't rely on memory. Initial results suggest improvement in discrimination accuracy over time, particularly on the dimension relevant to the categorization or control task, but no evidence of category boundary effects or effects of category learning on dimension perception stronger than those of the control task. Possible reasons for this and directions for further research are briefly discussed.

Keywords: categorical perception; categorization; learning; expansion; compression; dimensional modulation; selective attention

Introduction

It is well known that various kinds of experience can produce perceptual learning, i.e., improved ability to distinguish objects, features, or values on a dimension (Goldstone, 1998). One of the processes that is claimed to have special effects on the perceptual judgment of stimuli is learning to categorize the items, the phenomenon known as learned categorical perception (CP) (Goldstone & Hendrickson, 2009). Learned CP effects reported in the literature include boundary effects whereby items placed in different categories become more distinguishable, sometimes called expansion, and/or items placed in the same category become less distinguishable, sometimes called compression. However, these are not always clearly distinguished from dimension-wide effects where there is sensitization to the category-relevant dimension(s) and/or desensitization to the category-irrelevant dimension(s).

There are potentially many tasks besides category learning that require or benefit from greater attention to one dimension rather than another whereas only category learning would be expected to produce the boundary effects of expansion and/or compression. It is therefore very important that measures of learned CP carefully distinguish dimensional effects from boundary effects, something that previous research has not necessarily done. An important goal of the work reported here is to develop a method that distinguishes boundary effects of category learning from dimension-wide effects and, if category learning does cause dimension-wide effects, to determine if it does so to a greater extent than a task that doesn't involve category learning.

One reason that learned CP effects are of theoretical interest is that they may provide key evidence of genuine top-down effects on perception, an issue of considerable current controversy (Firestone & Scholl, 2016). But since the vast majority of learned CP evidence is based on measures that rely on memory (e.g., successive judgments of pairs of stimuli for same-different or similarity judgments), it is hard to argue that they are genuinely perceptual effects rather than reflecting higher level cognitive processes. Another purpose of the method adopted here is to eliminate the role of memory and determine if learned CP effects still occur. (Of course, even

if they do, other challenges raised by Firestone and Scholl might still need to be addressed.)

An examination of the existing body of learned CP research also reveals a bewildering pattern of effects and non-effects (compression vs. expansion vs. both, boundary effects with or without accompanying dimensional effects and vice versa, etc.). Researchers rarely have specific predictions regarding which effects will or won't occur and often don't distinguish clearly between them or test for all of them. As noted above, our study will clearly distinguish boundary effects from dimension-wide effects of category learning.

A recent p-curve meta-analysis of this body of research (Andrews, de Leeuw, Larson, & Xu, 2017) found a low level of statistical power, suggesting that it may be unproductive to try to interpret the patterns of effects and non-effects in the existing literature, since low statistical power is likely to produce both false positive and false negative results. Without a firm grasp on which learned CP effects do and don't occur under what conditions, it will be very difficult to make progress understanding the theoretical basis of learned CP or modeling the relevant mechanism(s). In addition to simply running better powered studies, another strategy to increase the informativeness of the data that are collected is to use analysis techniques such as Bayesian statistics that indicate the relative support for different hypotheses regarding learned CP effects, including the null hypothesis of no effects.

Another important methodological feature that renders previous results difficult to interpret is the fact that learned CP experiments almost always use a before-after comparison, a control group that only performs the final task performed by the learning group after category training, or at most, a control group that receives passive exposure to the category training stimuli. The goal of the research reported here is to address this and the other features of learned CP research that render its results ambiguous. Our approach relies on the use of a new method for tracking the effects of learning to categorize a set of patterns *over time* and in *comparison to the effects of performing an appropriate non-category-based control task*. Tracking over time is important for addressing another ambiguity when effects are only measured after training: expansion effects cannot be distinguished from a combination of compression and sensitization to the category-relevant dimension. These could potentially be distinguished if they emerge at different rates or times over the course of training. In order to track effects of learning over time, we test for changes in discrimination ability using a psychophysical staircase procedure throughout the entire experiment, alternating with classification or control task trials.

Because we use *simultaneous* stimulus presentation to avoid memory effects, a standard same-different or XAB task would allow successful performance based on the comparison of meaningless pixel-level features. We therefore developed a stimulus set where the potentially category-relevant dimensions vary both systematically in one respect (e.g., number/density of dots inside a circle) and also randomly (e.g., the exact location of the dots). This means that two stimuli with the same values on the two systematically varying dimensions will not be identical, much in the same way that individual instances of real world categories are usually unique. This allows us to use a variation on same-different judgments that highlights the role of the dimensions and works with simultaneous presentation, as explained in the method section.

The above features of our method make it different from the usual learned CP experiment in a number of ways, but we think it is essential to determine whether learned CP will occur under these more controlled conditions. If it does not, we can systematically re-introduce more traditional methodological features, such as successive presentation on the discrimination test, to determine which are necessary to produce the effects in order to better understand them. While we only report one experiment and acknowledge that our method likely needs adjustment to be fully successful in achieving its goals, our hope is that by sharing our work at this stage we can obtain useful feedback to inform and guide our next steps.

Method

All study materials, data, and analysis scripts are available at this OSF site: https://osf.io/msq57/.

Participants

A total of 101 participants (52 women; mean age 34.8; age range 18-72) were recruited using the online crowdsourcing platform Prolific and paid $4 for participating. Data from 8 participants were missing or incomplete leaving a final total sample size of 93.

Stimuli

Stimuli for this experiment were sunbursts. The number/density of dots and lines was systematically varied across stimuli but the exact placement of the dots and lines and the length of the lines were random (see Figure 1). For each participant in the experimental group (see below), category membership was randomly assigned to be based on either line or dot density. The density of dots or lines in a particular stimulus ranges from 300-2000 dots and 30-550 lines. (Each range is treated as 0.0-1.0 here.)

Procedure

The software jsPsych was used to create the experiment (de Leeuw, 2015). **Phase 1** used a same-different task variant we call the odd-one-out task. Four sunbursts appeared simultaneously: three had the same dot and line densities and one differed on one of those dimensions. Participants had 4 seconds to press a number key (1-4) to indicate the odd one out and receive feedback (see Figure 2).

At the beginning of Phase 1, the dimension that differed in the odd one differed by a large amount from the others. This distance was subsequently adjusted through a staircase

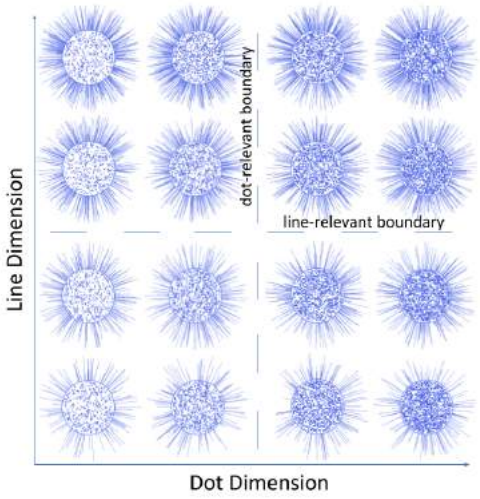


Figure 1. The stimulus space illustrating the two dimensions and the two possible sets of categories.

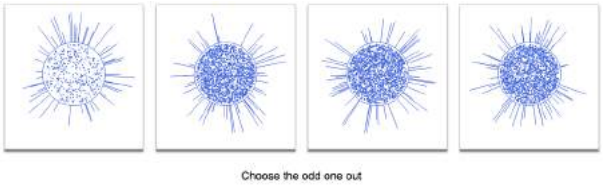


Figure 2. An odd-one-out trial display in Phase 1.

procedure, decreasing or increasing by 15% depending on whether the response was correct or incorrect. Trials continued until at least eight reversals occurred on each dimension. The goal of Phase 1 was to identify an approximation of each individual participant's just noticeable difference (JND) on each dimension, defined as the average of the distances of the last four reversals.

In **Phase 2**, odd-one-out trials alternated with one of two other tasks, classification or number judgment, to which participants were randomly assigned. For both tasks, a single sunburst appears with a question and participants press a key to answer and receive feedback (see Figure 3).

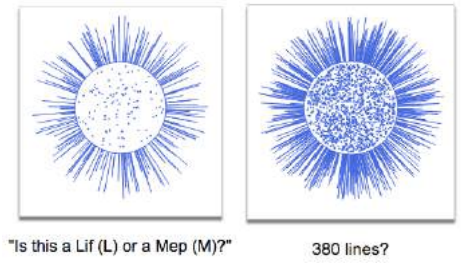


Figure 3. A classification task trial display (left) and a number judgment task trial display (right) in Phase 2.

The number judgment (control) task is to say "more" (M) or "less" (L) in response to a question about the number of dots or lines, where the number varied from trial to trial. For a given participant, the number judgment questions are always about just one of the two dimensions, randomly assigned, so that the control task matches the category learning task in relying on attention to one "relevant" dimension to answer correctly. For the classification ("experimental") group, the randomly assigned relevant dimension defined the category boundary as shown in Figure 1.

The specific stimuli used in Phase 2 odd-one-out trials were initially based on each JND value from Phase 1 for each participant and dimension. The sets of four stimuli (see Figure 2) were of three types as shown below in Figure 4. For both BE (between category) and WI (within category) comparisons, the odd one out differed from the other three only on the relevant dimension while for IRR comparisons, it differed only on the irrelevant dimension. All 48 possible adjacent stimulus pairs were used as the basis for the odd-one-out trials and drawn from the participant's JND-based dimensional space at a given moment.

Phase 2 trials proceeded in 40 blocks each containing six odd-one-out trials (one BE, two WI, and three IRR trials to sample the stimulus space evenly) and six classification or number judgment trials in a random order. The staircase procedure on the odd-one-out task was continued individually for each participant throughout Phase 2 just as in Phase 1, but separately for these six comparison subtypes. This controls for discriminability differences due to stimulus magnitude (e.g., Weber's law). Since adjacent dimensional values were already near JND level, the proportion change from one trial to the next of that subtype was reduced from 15% to 5% and the maximum distance allowed between dimension values was .33.

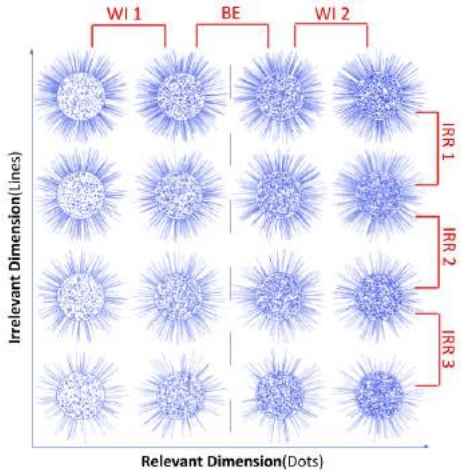


Figure 4. Illustration of the six comparison subtypes for the odd-one-out trials with dots as the relevant dimension.

Analysis Plan

A traditional learned CP analysis takes a behavioral measure such as similarity rating or same-different accuracy and compares the experimental (category learning) and control

groups on that measure for between-category vs. within-category pairs. Our experiment tracked changes in the size of the distance between the two dimensional values used in odd-one-out task trials. Therefore, our learned CP measure was the change in this value for a given dimension from the beginning to the end of Phase 2. If participants improved on the odd-one-out task, their scores will be negative since they will become able to accurately judge smaller differences, and a larger negative score represents more improvement. Because differences in speed of discriminating between-category vs. within-category pairs are sometimes taken as evidence for CP, we also used mean correct reaction time over the last four blocks on odd-one-out trials as an alternate measure. We standardized RTs within subject by converting them to z scores. Note that for both of these measures, a smaller score reflects better performance.

It is traditional for the above types of analysis to adopt some criterion of successful category learning and exclude participants who don't meet it. However, the choice of the criterion is arbitrary, may well influence the results, and is not explicitly motivated in learned CP research. In addition, because our continuous staircasing procedure kept dimensional differences between adjacent stimuli near JND, we expected category learning to be relatively difficult and produce a wide range of performance levels. Since it seems reasonable to predict that learned CP measures should positively correlate with category learning success (see Gureckis & Goldstone, 2008 for a similar approach and positive evidence), we only reported that type of analysis.

Results

Figure 5 shows an example of a result of the Phase 1 staircase procedure for illustrative purposes. Participants whose Phase 1 JND on either dimension exceeded the maximum of .33 allowed in Phase 2 by more than .05 were excluded from subsequent analysis since the Phase 2 staircasing procedure would not apply correctly to them. This produced a final n of 72 (35 control, 37 experimental).

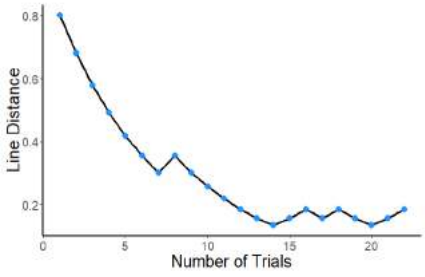


Figure 5. Example outcome of Phase 1 staircase procedure.

The mean proportion correct over Phase 2 on the classification task was .678 (*SD* = .145) and on the number judgment task it was .807 (*SD* = .13).

Phase 2 began with dimensional differences based on each individual participant's Phase 1 JND. Did the staircase procedure continuing throughout Phase 2 (in alternation with the classification or number judgment task) produce further perceptual learning? Figure 6 shows that in general, averaging across all participants, it did, particularly on the relevant dimension comparisons, as one might expect. Using the mean distance change for each participant averaging over the three odd-one-out trials differing on the relevant dimension (BE, WI1, and WI2) in the final block, the mean of the entire sample (M = -2.57) was significantly less than zero ($t(71)$ = -3.547, $p < .0001$). This was not the case for the irrelevant dimension (averaging over IRR1, IRR2, and IRR3 trials) (M = -0.72, $t(71)$ = -0.996, p = .16). A one-tailed paired samples t-test yielded a significant difference between relevant and irrelevant mean distance change ($t(71)$ = -2.014, p = .024).

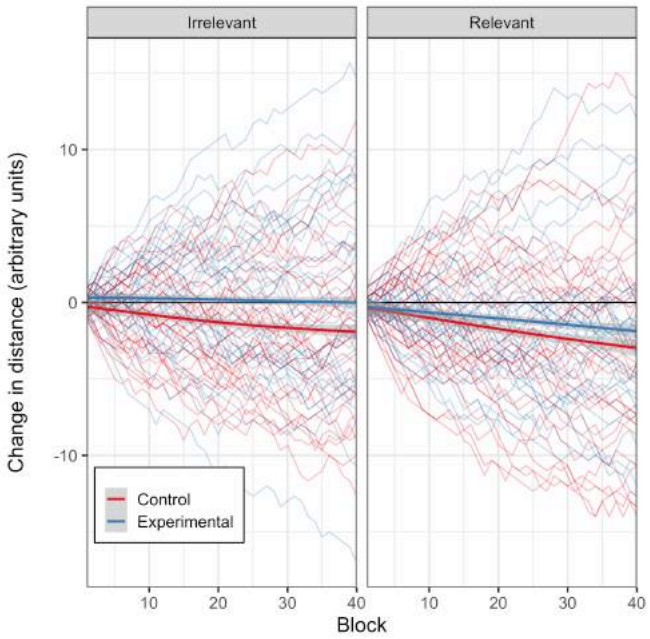


Figure 6. Overall perceptual learning in the experiment; the y axis represents number of staircase steps, e.g., a change in distance of -10 means the staircase has gotten 10 steps more difficult, indicating improved discrimination accuracy.

The left panel in Figure 7 illustrates the pattern that would be expected to hold for the control group, with better performance on the number judgment task coinciding with better performance on the odd-one-out task only (or to a greater degree) for the dimension relevant to the number judgment task, and no difference in the patterns for between and within category comparisons. The right panel shows what the pattern would be if the experimental group showed learned CP boundary effects, with better classification performance associated with better odd-one-out performance on between-category comparisons (expansion) and/or worse odd-one-out performance on within-category comparisons (compression) relative to the control group. If the experimental group were to show stronger sensitization to the relevant dimension or desensitization to the irrelevant

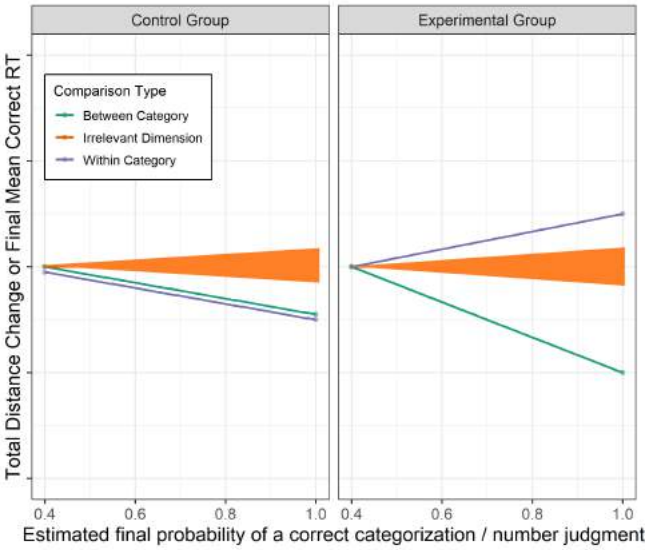


Figure 7. Relationship between classification or number judgment task performance (x axis) and either dependent variable (y axis) predicted by learned CP boundary effects.

dimension relative to the control group, the patterns would be slightly different, but we will focus on the boundary effects that are typically what is meant by learned CP.

Figure 8 shows the actual relationship in our data between total distance change over Phase 2 (y axis) and an estimate of the probability of a correct response at the end of Phase 2 on the number judgment (left) or classification (right) task (x axis) obtained by fitting a logistic regression model for each individual subject.

Overall there is a weak negative relationship such that discrimination performance tended to be better when classification or number judgment was more accurate, perhaps reflecting a general effect of effort. These data were analyzed using a Bayesian linear model to predict total distance change from three variables: comparison type (between, within, or irrelevant), group (control or experimental), and estimated final performance on the number judgment or classification task. The model also included the three-way interaction between these three variables since, as shown in Figure 7, this would have to be present if learned CP effects occurred. The analysis produced a BF_{10} of 119 for the estimated final performance variable, supporting the effort effect mentioned previously. To assess evidence for the critical three-way interaction, we determined the ratio of the BF_{10} for the full model containing the three predictor variables and the three-way interaction (.94) to the BF_{10} for the model containing just the three predictor variables (7.18). This yielded a BF_{10} of .131 indicating moderate support for H0 and therefore no evidence for learned CP.

The same analysis was performed for the RT measure (see Figure 9) and showed only one result favoring the alternative hypothesis and that was for comparison type

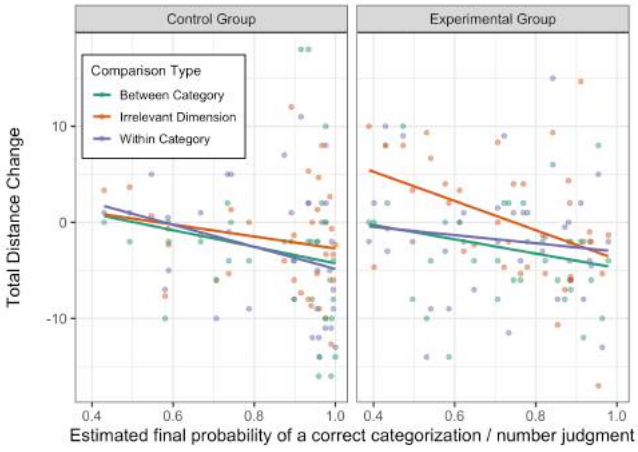


Figure 8. Relationship between estimated final performance on the classification or number judgment task and actual discrimination accuracy improvement over Phase 2 for the three comparison types.

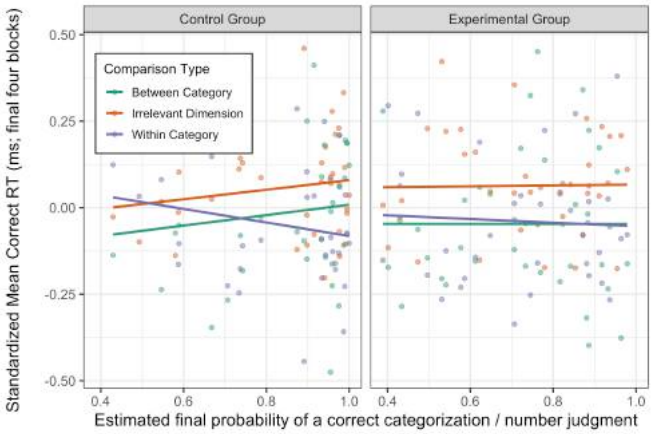


Figure 9. Relationship between estimated final performance on the classification or number judgment task and standardized mean correct RT on the last four blocks of Phase 2 for the three comparison types.

(BF_{10} = 41). The graph shows this to be due to irrelevant dimension odd-one-out trial responses being slower in general than relevant dimension trials of either type. The ratio of the BF_{10} for the full model containing the three predictor variables and the three-way interaction (.095) to the BF_{10} for the model containing just the three predictor variables (.942) yielded a BF_{10} of .101. This constitutes fairly strong support for H0 and thus no learned CP effects.

Discussion

This experiment employed a novel methodology designed to rigorously test for learned CP effects. Stimuli varied systematically on two dimensions, only one of which was relevant for either category learning or a control task. The stimuli also varied in random low-level features to allow for

simultaneous presentation in the discrimination ("odd-one-out") task to eliminate reliance on memory. A staircasing procedure was used to initially determine the JND for each participant on each dimension and this staircasing continued with discrimination trials alternating with classification or control task trials to allow for continuous measurement of discrimination ability on each dimension.

The results provide evidence of sensitization to the relevant dimension for both the classification task and a control task that was comparable in requiring attention to one of the two dimensions. This was seen in significant discrimination performance improvement from the beginning to the end of Phase 2 for the sample as a whole on the relevant but not the irrelevant dimension. The interesting question then is whether there were differences in discrimination performance between the two groups that fit any of the patterns consistent with learned CP.

We did not report traditional analyses of learned CP effects, comparing successful category learners to the control group on our odd-one-out performance measures as a function of comparison type, due to the arbitrariness of setting a criterion for successful learning and the fact that our continuous staircasing procedure kept discrimination across the category boundary difficult. Instead, we examined whether learned CP effects appeared in the form of different *relationships* between category learning performance and discrimination performance as a function of comparison type and in relation to the control group.

The only effects we found were a positive correlation between success on the classification or number judgment task on the one hand and the odd-one-out task on the other, and slower response times by the end of the experiment for odd-one-out trials that required distinguishing stimuli differing on the irrelevant dimension. The critical three-way interaction between group, comparison type, and level of classification or number judgment performance that would be required in order to demonstrate any variety of learned CP effects was lacking for both dependent measures, and the analyses showed more than anecdotal support for its absence.

Note that if learned CP effects had occurred in this experiment, our continuous measurement of discrimination ability on the three types of comparisons would have been valuable for tracking the emergence of different types of effects (e.g., expansion vs. compression) and would have potentially allowed us to distinguish otherwise similar end results (i.e., expansion vs. a combination of compression and relevant dimension sensitization). However, since we did not obtain any learned CP effects overall, we were not able to take advantage of this capability.

There are many possible reasons for these negative results, due to the ways in which our methodology deviated from typical learned CP experiments. Perhaps the constantly changing stimulus set and its randomly varying sub-features below the dimensional level prevented learned CP from occurring. Or it may be that constantly alternating between a classification task and the odd-one-out task interfered with learned CP. If learned CP effects depend on memory and thus require tasks with a delay between stimuli in order to occur, our simultaneous stimulus presentation would be the cause. Or it could be that, previous evidence of boundary effects notwithstanding, so-called learned CP effects are really due to paying attention selectively to one dimension rather than another, and thus also occur as a result of other tasks besides category learning such as the number judgment task used by our control group.

We believe it is very important to determine the conditions under which learned CP effects do and do not occur, which has not been addressed sufficiently in the literature. Our negative results can provide a useful initial reference point. One strategy for building on this would be to next conduct a traditional version of the experiment utilizing a fixed set of the same stimuli and and a successive presentation version of our discrimination task to establish whether learned CP effects do occur under those conditions. If they do, methodological changes can then be incorporated one at a time, such as simultaneous rather than successive discrimination testing and comparison to a control group that performs a task requiring attention to one dimension, to determine which manipulations change and/or eliminate learned CP effects. This would allow us to make real progress in understanding the phenomenon of learned CP and its scope and limits.

Acknowledgments

This project was supported by the Vassar College Undergraduate Research Summer Institute.

References

Andrews, J. K., de Leeuw, J. R., Larson, C., & Xu, Xiaoqing. (2017). A preliminary p-curve analysis of learned categorical perception research. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.), *Proceedings of the 39th Annual Meeting of the Cognitive Science Society* (pp. 1550-1555). Austin, TX: Cognitive Science Society.

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, *47*(1), 1–12.

Firestone, C., & Scholl, B. J. (2016). Cognition does not affect perception: Evaluating the evidence for "top-down" effects. *Behavioral and Brain Sciences, 39*, 1-72.

Goldstone, R. L. (1998). Perceptual learning. *Annual Review of Psychology, 49*, 585-612.

Goldstone, R. L., & Hendrickson, A. T. (2009). Categorical perception. *Interdisciplinary Reviews: Cognitive Science*, *1*, 69–78.

Gureckis, T. M., & Goldstone, R. L. (2008). The effect of the internal structure of categories on perception. In B. Love, K. McRae, & V. Sloutsky (Eds.), *Proceedings of the 30thh Annual Conference of the Cognitive Science Society* (pp. 1876-1881). Austin, TX.: Cognitive Science Society.

Distant Concept Connectivity in Network-Based and Spatial Word Representations

Abhilasha A. Kumar (abhilasha.kumar@wustl.edu)
Department of Psychological & Brain Sciences, Washington University in St Louis, MO 63130 USA
David A. Balota (dbalota@wustl.edu)
Department of Psychological & Brain Sciences, Washington University in St Louis, MO 63130 USA
Mark Steyvers (mark.steyvers@uci.edu)
Department of Cognitive Sciences, University of California, Irvine, Irvine, CA 92697 USA

Abstract

It is presently unclear how localized, word association network representations compare to distributed, spatial representations in representing distant concepts and accounting for priming effects. We compared and contrasted 4 models of representing semantic knowledge (5018-word directed and undirected step distance networks, an association-correlation network and *word2vec* spatial representations) to predict semantic priming performance for distant concepts. In Experiment 1, response latencies for relatedness judgments for word-pairs followed a quadratic relationship with network path lengths and spatial cosines, replicating and extending a pattern recently reported by Kenett, Levi, Anaki, and Faust (2017) for an 800-word Hebrew network. In Experiment 2, response latencies to identify a word through progressive demasking showed a linear trend for path lengths and cosines, suggesting that simple association networks can capture distant semantic relationships. Further analyses indicated that spatial models and correlation networks are less sensitive to direct associations and likely represent more higher-level relationships between words.

Keywords: neural networks; *word2vec*; semantic priming; semantic space model; word association; network science.

Introduction

Understanding language requires the retrieval of meaning from underlying semantic representations of words. A class of models of semantic memory represent words as nodes in a large memory network, where words with similar meanings are connected to each other via edges (see Kenett, Kenett, Ben-Jacob & Faust, 2011; Steyvers & Tenenbaum, 2005). Semantic network models propose localized word representations, in contrast to feature-based or distributed space models (Smith, Shoben & Rips, 1974; Landauer & Dumais, 1997).

Spatial models of semantic memory represent words in a multi-dimensional space, where words are an aggregate of the individual dimensions of the space. The spatial dimensions are derived from statistical co-occurrences in natural language. For example, Latent Semantic Analysis (LSA; Landauer & Dumais, 1997) is a distributional model that measures semantic similarity by calculating co-occurrences of words in a text corpus. LSA successfully simulates complex human behavior in a variety of cognitive tasks but has had difficulty accounting for semantic priming effects (Hutchison et al., 2008) and power laws (Steyvers & Tenenbaum, 2005), suggesting that spatial models may have some limitations.

A more recent spatial model, *word2vec* (Mikolov, Chen, Corrado & Dean, 2013) has received considerable attention in the fields of computer science and natural language processing. *word2vec* uses neural networks to compute continuous vector representations of words. These semantic representations can then be used to compute an index of semantic similarity between words via vector cosines (higher cosines indicate greater semantic similarity). Interestingly, *word2vec* is able to solve verbal analogy problems (e.g., king: queen::man:?) using simple vector arithmetic, although recent research suggests that *word2vec* successfully captures only certain, simpler types of semantic relationships and not others (Chen, Peterson & Griffiths, 2017). The question of whether individuals use an association-based representation or represent meaning in a high-dimensional space is currently controversial (Griffiths, Steyvers & Tenenbaum, 2007; Jones, Gruenenfelder & Recchia, 2011). Thus, direct comparisons among different types of meaning representations and how they account for more distant semantic relationships is an important next step for the field.

Recently, Kenett, Levi, Anaki and Faust (2017) used a semantic relatedness task to explore the impact of network path length derived from an 800-word Hebrew semantic network. The Hebrew network was created using correlations from continuous free association responses of 60 participants to 800 target words (for complete methodology, see Kenett et al., 2011). The results from the semantic relatedness task indicated that as network path length between word pairs (i.e., shortest distance between two words in the network) increased, fewer word pairs were judged as related. They also reported a quadratic relationship between path length and response latencies to make relatedness judgments, such that response times (RTs) increased for word pairs at shorter path lengths (e.g., BUS-CAR), but after path length 3, RTs systematically decreased for word pairs at longer path lengths (e.g., CHEATER-CARPET). They also showed that this network outperformed LSA and another measure of semantic distance, Positive Pointwise Mutual Information (PPMI) in explaining task performance. However, given that Kenett et

al. used a novel association-correlation methodology based on a Hebrew network, it remains unknown how simpler association networks (e.g., Steyvers & Tenenbaum, 2005) and more recent spatial models (e.g., Mikolov et al., 2013) capture such distant semantic relationships. Moreover, it is important to extend the Kenett et al. network structure to a larger English-based network analysis to examine the generalizability of their findings.

The present set of experiments were designed to compare and contrast the structural differences between three different network-based models and the *word2vec* model, across two behavioral tasks. It is important to note here that we do not claim that association-based networks are a complete account of semantic memory, but the issue we are interested in whether networks created from simple associations can indeed capture distant semantic priming effects, and how they compare to other models of semantic memory, such as spatial models and the association-correlation network. There is a rich tradition of using network-based models to accommodate priming effects (Anderson, 2000; Collins & Loftus, 1975), and we were mainly interested in comparing different types of network-based approaches to each other in accounting for this well-studied task, and also to other spatial representations. In Experiment 1, we extended and replicated the patterns reported by Kenett et al. in the Hebrew semantic relatedness task in three large semantic networks in English along with cosines from the *word2vec* model. We created these networks from a 5018-word database of free association norms collected by Nelson, McEvoy & Schreiber (2004) to examine the extent to which network path lengths would predict performance in the relatedness judgment task.

A potential concern regarding the performance of network models created through human association norms in Experiment 1 is both relatedness judgments and word associations direct attention to the meaning dimension, and thus the patterns observed may just be due to overlap in the type of task. Further, the quadratic pattern observed may just reflect how the semantic "distance" between two words might influence the related/unrelated decision and how a particular individual partitions items into these arbitrary categories. We attempted to address this concern by employing a task that does not require accessing meaning-related information to make the response. Thus, in Experiment 2, participants first viewed a briefly presented prime (120 ms) and then identified targets through a visual demasking task. Hence, we were able to directly compare the different network configurations and spatial representations in accounting for performance in two behavioral tasks.

Semantic Network Construction

To construct the semantic networks, we used a 5018-word database of free-association norms collected by Nelson et al. (2004), in which 150 participants on average wrote down the first word that came to mind in response to approximately 120 word-cues. The cues were selected by Nelson et al. after multiple rounds of data collection, and typically, the most

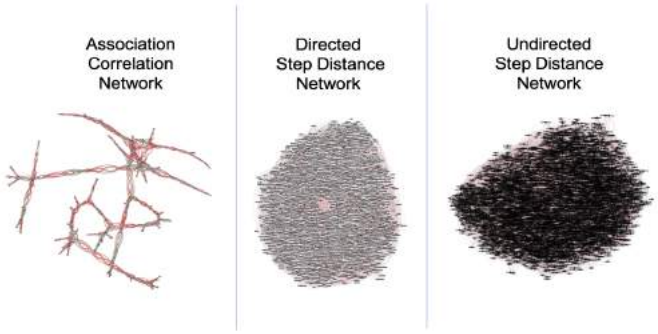


Figure 1: Large-scale visualization of the Association-Correlation Network, Directed and Undirected Step Distance Networks.

frequent responses for each of the cues were contained within the 5018 cues themselves. Responses were included only if at least two participants produced the same response, thus excluding idiosyncratic responses from the database. Responses that were not within the 5018 cues were also excluded during network construction. We constructed three networks from this database: an Association-Correlation Network (ACN), an Undirected Step Distance Network (Undirected SDN) and a Directed Step Distance Network (Directed SDN).

Association-Correlation Network

The ACN was created based on the methodology described by Kenett et al. (2011). Associative responses to 5018 cue words were first converted into a matrix, in which each column represented a cue word, and each row indicated unique associative responses for the target word. This matrix was converted to an association-correlation matrix, where the correlations between two target word profiles (i.e., the words produced to the two targets) was calculated based on the Pearson's formula. This correlation matrix was converted into a weighted, undirected network, such that each target word was a node in the network, and the correlation between two target words represented the weight of the edge between them. This fully-connected network was then reduced to a planar maximally filtered graph, resulting in a smaller planar network (a network in which no edges cross each other) with the same target nodes, but only edges that represent the most relevant associations between target words. Path length between word-pairs was then calculated as the shortest path from one word to another in this smaller network. Figure 1 (Left panel) displays a large-scale visualization of the ACN, and Figure 2 (Left panel) displays the 6-step shortest path from RELEASE to ANCHOR.

Undirected and Directed Step Distance Networks

Following Steyvers and Tenenbaum (2005), in the Directed SDN, two words (*a* and *b*) were connected by an edge if the word *a* evoked the word *b* as an associative response for at least two participants in the Nelson database. In the Undirected SDN, words were connected if *a* evoked *b* or *b*

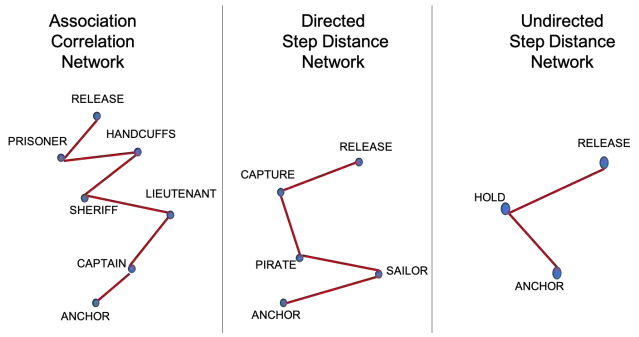


Figure 2: Shortest path from RELEASE to ANCHOR in the Association-Correlation Network, Undirected and Directed Step Distance Networks.

evoked *a*, independent of the associative direction. Path length for each word pair in the network was calculated as the shortest path from one word to another. Figures 1 and 2 (Middle and right panels) display visualizations of the two SDNs, and the shortest path from RELEASE to ANCHOR.

Network Comparisons

Table 1 displays the network parameters for the three networks. As is evident from the large-scale visualizations, ACN is sparser than the SDNs, with a greater clustering coefficient (an index of network connectivity, i.e., the extent to which neighborhoods of neighboring nodes overlap) and longer average path lengths, indicating more distant associations compared to the direct associations captured by SDNs with shorter path lengths overall. Table 2 displays the correlation among the path lengths derived from each of the networks for the sets of words used in our experiments. As is clear, there were considerable differences across the different types of network configurations. As shown in Figure 1, the ACN is a sparsely connected network, in which obscure,

Table 1: Network parameters for the semantic networks

	Simple Step Distance Networks		Association-Correlation Networks	
	Undirected	Directed	English	Hebrew
n	5018	5018	5018	800
<*k*>	22	12.7	5.85	5.94
L	3.04	4.27	23	10
D	5	10	61	25
C	.186	.186	.69	.68
L_{random}	3.03	4.26	1.95	3.94
C_{random}	.004	.004	.05	.005

Note. *n* = the number of nodes; <*k*> = average number of connections; *L* = average shortest path length; *D* = diameter of network; *C* = clustering coefficient; L_{random} = average shortest path length with random graph of same size and density; C_{random} = the clustering coefficient for a random graph of same size and density.

Table 2: Correlation matrix for network path lengths and *word2vec* cosines for word-pairs in Experiments 1 and 2

	ACN	Undirected SDN	Directed SDN	*word2vec* Cosines
ACN	1	-	-	-
Undirected SDN	.49	1	-	-
Directed SDN	.35	.58	1	-
word2vec Cosines	-.42	-.55	-.45	1

Note: All correlations were significant at the $p < .05$ level

higher-level associations are closely represented (e.g., TRAGEDY-REMORSE is 1 step away), whereas several direct (e.g., VOLCANO-ASH is 15 steps away) and mediated associations (e.g., LION-STRIPES is 38 steps away) are exaggerated. Overall, path lengths derived from the two SDNs were very highly correlated, suggesting that the simple associative networks largely overlap in their network structure, and differ from the ACN.

Vector Cosines via word2vec

The *word2vec* model (Mikolov et al., 2013) trains neural networks based on words that naturally co-occur in a text corpus and uses this contextual information to predict a word's immediate contextual neighborhood. Typically, these contextual words have probabilities associated with them, which indicate the likelihood of words co-occurring together in natural language. If two words occur in similar contexts, the model learns similar vector representations for those words. Cosines between these vector representations thus serve as indices of semantic similarity. For all the word pairs used in the current experiments, we obtained *word2vec* cosines from a pre-trained model trained on 100 billion words from a Google News dataset (Mikolov et al., 2013). Table 2 displays the correlations between *word2vec* cosines and path lengths derived from the three networks described above. Note that *word2vec* cosines were negatively correlated with the path lengths, due to the direct cosine similarity measure used. Further, there were considerable differences across the models in the extent to which they captured "semantic similarity", given that the average correlation among all the measures was only .46.

Experiment 1

Methods

Participants Forty Amazon Mechanical Turk users (M_{age}= 37 years, SD = 10.4) and 40 undergraduate students (M_{age}= 20 years, SD = 0.8) recruited from Washington University in St Louis participated in the study. All participants were self-reported native English speakers.

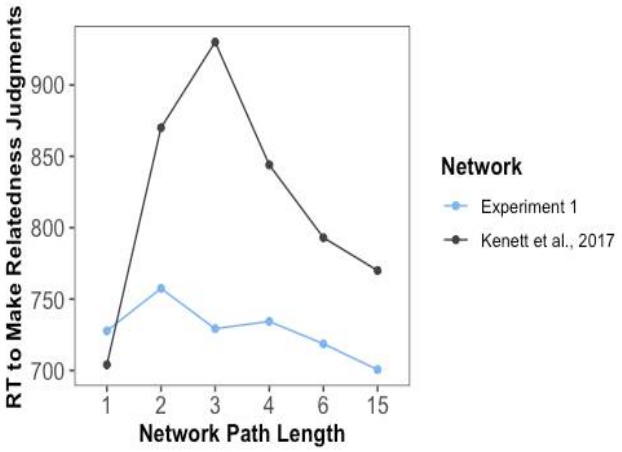


Figure 3: Response times for relatedness judgments in Experiment 1 and Kenett et al. (2017)

Materials In order to extend and replicate the Kenett et al. study, we randomly sampled 40 word-pairs from path lengths 1, 2, 3, 4, 6 and 15 from the ACN. The stimuli consisted of 1200 distinct word-pairs across 5 lists. For each word-pair sampled from the ACN, we also obtained path lengths in the Undirected and Directed SDN and obtained vector cosines from the *word2vec* model. We also obtained lexical characteristics (word length, frequency, lexical decision times and concreteness) for all the words from the English Lexicon Project (ELP; Balota et al., 2007) and used these as covariates in our analyses. All items used in the current study are available at https://github.com/abhilasha-kumar/Distant-Semantic-Connectivity.

Procedure

The relatedness task was developed in JSPsych, an online software for conducting psychological experiments. Each participant completed the experiment online. Following Kennett et al., on each trial, participants saw a fixation cross for 200 ms, followed by a blank screen for 100 ms. Then, the prime was briefly presented for 120 ms, followed by the target for 120 ms. Participants decided whether the prime and target were related or unrelated and responsed by pressing a key (K or L, counterbalanced). After a response, participants saw a blank screen for 500 ms before the next trial.

Results

There were no differences in the overall patterns between the five lists, or the Amazon Mechanical Turk or Washington University sample, thus all analyses included the full sample.

Effect of ACN Path Length on RTs To replicate the analysis procedures reported in Kenett et al. (2017), each path length was first classified as related or unrelated, based on the percentage of related and unrelated responses to specific word pairs. The following were the percentages of "related" responses for the path lengths: 1 (66%), 2 (47%), 3 (29%), 4 (27%), 6 (16%) and 15 (13%). Based on these percentages and the criterion of at least 50% of words producing a related response, only path length 1 was considered related, and the remaining path lengths were considered unrelated. To minimize any effects of slowing and individual differences, all RTs faster than 250 ms and slower than 2000 ms were removed. Second, a mean and standard deviation were calculated from the remaining trials for each participant and any RTs that exceeded 3 standard deviations (SDs) from the participant mean were also removed. This process excluded 5.4% of the total trials. After this trimming procedure, we standardized the remaining trials within each participant and conducted all primary analyses using trial-level standardized RTs. A repeated measures Analysis of Variance (ANOVA) on mean RT revealed a significant main effect of path length, $F_1(5, 395) = 7.42$, $p < .001$, $\eta_p^2 = .09$. RTs significantly increased from path length 1 to 2 ($p = .006$), decreased from path lengths 2 to 3 ($p = .001$) and 4 to 15 ($p = .015$). As shown in Figures 3 and 4, we successfully replicated the pattern reported by Kenett et al. for RTs as a function of path length in the ACN. Importantly, this pattern persisted after including degree of relatedness as a predictor in our analyses, standardizing the RTs and controlling for lexical variables such as word frequency, length, concreteness and standardized lexical decision times, as well as mean degree (i.e., number of direct neighbors of the words) using linear mixed effects models.

Effect of SDN Path Length on RTs In addition to the ACN based on Kenett et al., as noted, we also examined the effect of path lengths derived from two SDNs (Undirected and Directed) based on the method used in Steyvers and Tenenbaum (2005) on standardized RTs in the relatedness task. As shown in Figure 4, both the Undirected and Directed networks also showed a quadratic trend for standardized RTs as a function of path length, with RTs significantly rising from path lengths 1 to 2 ($p<.001$) and then reliably decreasing

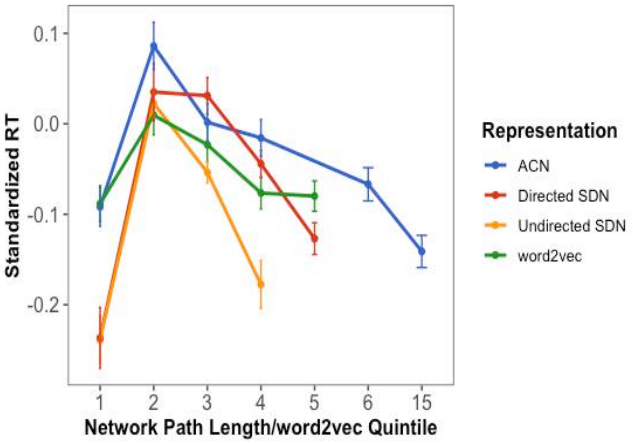


Figure 4: Standardized RTs for relatedness judgments in Experiment 1 as a function of network path lengths and *word2vec* cosine quintiles (reverse-scored)

from path length 2 onwards. We observed a significant decline in RTs from path lengths 3 to 4 in the Undirected ($p < .001$), and from 2 to 5 in the Directed network ($p < .001$).

Effect of word2vec Cosines on RTs We also computed vector cosines derived via *word2vec* for each of the word pairs in Experiment 1. As shown in Figure 4, continuous *word2vec* cosines successfully predicted standardized RTs to make relatedness judgments ($b = -.22$, $t = -3.54$, $p < .001$), and reproduced the quadratic pattern previously observed.

Discussion

The results from Experiment 1 provide strong evidence for multiple-step priming in the relatedness judgment task, and also replicate and extend the quadratic pattern observed by Kenett et al. (2017) for the Hebrew network. In addition, simple directional and nondirectional SDNs also captured distant semantic relationships between concepts. This is noteworthy, as it indicates that the number of "steps" in the ACN do not necessarily reflect *direct* associative strength, at least based on distances captured by simple SDNs. Of course, this does not imply that the ACN distances are unimportant, as the ACN shows comparable performance in the current task. We also found that the *word2vec* model successfully captured the quadratic trend, although there do seem to be differences in the semantic information captured by all the models, based on the relatively low correlations across the networks.

It is important to note that the nature of the relatedness decisions is likely driving the quadratic trend. Specifically, RTs are slowed to make "unrelated" decisions for the more ambiguous items e.g., at path lengths 2 and 3. Interestingly, the RTs for only the "related" decisions continued to increase with greater path lengths, a finding that is more consistent with a spreading-activation account. In addition, the networks in this study were explicitly created from free association norms, and their explanatory power may reflect the high degree of overlap between the base task (free association) and the relatedness judgment task. Thus, in Experiment 2, we explored whether network path length and vector cosines can account for semantic priming in a primed progressive demasking task, which does not explicitly involve explicit semantic retrieval to make a response.

Experiment 2

Methods

Participants Thirty-nine young adults (M_{age} = 20.9 years, SD= 2.8) were recruited from undergraduate courses at Washington University in St Louis. All participants were Native English speakers.

Materials One list of 240 items was randomly chosen from one of the five lists used in Experiment 1. As before, the list contained 40 word-pairs from path lengths 1, 2, 3, 4, 6 and 15 from ACN. Each word pair also had corresponding path lengths in the undirected and Directed SDN, as well as *word2vec* cosines. This list was then used to create two lists counterbalanced across participants, so that each word was a prime as well as a target in the study.

Procedure

The primed progressive demasking task was developed using E-Prime 2.2. Participants saw a black fixation cross on the screen for 500 ms. Next, a blank screen was displayed for 200 ms, followed by the prime word, displayed for 120 ms. Immediately after, the target word was progressively demasked on the screen. During progressive demasking, the display alternated between the target (e.g., XXXX) and a mask (a row of pound signs matching the length of the word, e.g., ######). The total duration of target-mask pair was held constant at 500 ms but the ratio of target display time to target display time progressively increased. The duration of the target increased at each cycle (0, 16, 32,...,500 ms) and the duration of the mask decreased (500, 484, 468,…0 ms). The demasking procedure continued until the target was fully revealed for 500 ms, or until the target was identified by the participants by pressing the spacebar and typing in the target word. The next trial began immediately after typing the target and pressing spacebar.

Results

Effect of ACN Path Length on RTs All trials in which the correct target was not identified were excluded from analyses (2.7%). Next, we standardized the RTs to identify the target as in Experiment 1. A repeated measures ANOVA revealed a significant effect of path length, $F_1(5,190) = 53.85$, $p < 0.001$, $\eta_p^2 = .586$. As shown in Figure 5, we observed a significant increase in RTs from path lengths 1 to 2 ($p <.001$), and 2 to 3 ($p <.001$). Differences between RTs at path length 3 and higher ACN path lengths were not reliable. These effects persisted after controlling for lexical variables & mean degree (i.e., number of direct neighbors of the words).

Effect of SDN Path Length on RTs. We also examined the effect of path lengths from the Undirected and Directed SDNs on standardized RTs. As shown in Figure 5, path lengths from the Undirected SDN significantly predicted RTs to identify the target. RTs increased from path length 1 to 2 ($p = .001$), from path lengths 2 to 3 ($p < .001$), and then marginally from 3 to 4 ($p = .058$). Path lengths from the Directed SDN also predicted RTs to identify the target. RTs increased from path lengths 2 to 3 ($p = .015$) and 4 to 5 ($p = .038$).

Effect of word2vec cosines on RTs We also obtained vector cosines derived via word2vec for each of the word pairs, as in Experiment 1. As shown in Figure 5, continuous word2vec cosines also successfully predicted standardized RTs to identify the target (($b = -1.34$, $t = -9.18$, $p < .001$).

Model Comparisons Because the results from this task were not complicated by the relatedness decision as in Experiment 1 (i.e., RTs should be linearly related to demasking performance), we were able to directly compare the model

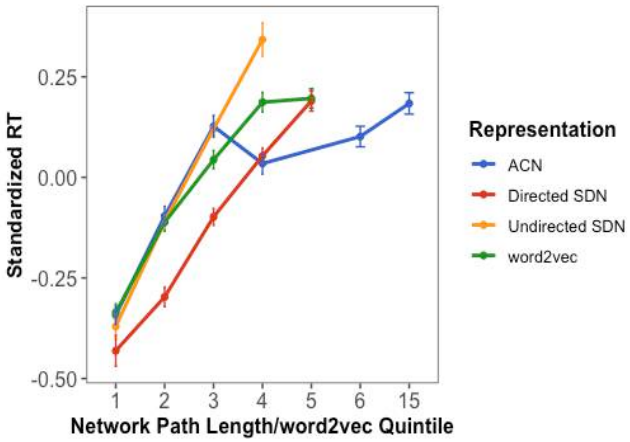


Figure 5: Standardized RTs to identify target word in demasking in Experiment 2 as a function of network path lengths and *word2vec* cosine quintiles (reverse-scored)

estimates. To estimate the unique variance accounted for by each type of network configuration at the item level, we calculated the individual R^2 for each model, as well as estimates of AIC and BIC, after controlling for covariates. As shown in Table 3, the models had overall comparable fits, and explained a significant amount of variance over and above the model with just the covariates, although as discussed before, these models seem to capture somewhat different semantic information.

Discussion

Results from Experiment 2 indicated that network path lengths can indeed account for performance in a progressive demasking task. RTs linearly increased as a function of SDN path lengths and *word2vec* cosines. This is especially interesting as the demasking task does not require any direct retrieval of semantic association to make the response, and yet, we see that path lengths derived from word associations directly predict demasking response latencies. Further, we found reliable differences at relatively distant path lengths in the simple association networks, suggesting that simple association networks are able to capture distant semantic relationships in the memory network, even in tasks that do not necessarily direct attention to semantics. Interestingly, we find that path lengths from the ACN increase linearly only up to 3 steps, after which the network seems to no longer be sensitive to priming effects in this task, suggesting differences in the network structures.

General Discussion

A primary goal of the present study was to compare the extent to which measures of semantic similarity derived from different types of network-based models explained distant semantic priming. In Experiment 1, we replicated and extended a pattern previously reported by Kenett et al. (2017) to a larger 5018-word association network in English and also

Table 3: Model comparison metrics for Experiment 2

Model	R^2 (%)	AIC	BIC	Likelihood ratio test
Covariates	13.33	561.9	586.1	---
ACN	26.99	500.8	545.1	p <.001
U-SDN	22.16	523.3	559.6	p < .001
D-SDN	25.98	506.4	550.8	p <.001
word2vec	28.03	486.8	515	p < .001

compared their graph-theoretical approach of capturing semantic similarity with simpler Undirected and Directed Step Distance Networks (Steyvers & Tenenbaum, 2005). Our results indicated that simple association networks can also capture similar distant relationships between words in the lexicon. Experiment 2 indicated that network models also successfully capture performance in tasks that do not directly rely on word association.

As described earlier, the ACN uses correlations between association responses and a planarity criterion to construct the network, and possibly captures more higher-level associations. This leads to several direct word associations (e.g., TIGER-STRIPES is 37 steps away in the ACN and 1 step away in the SDNs) being dropped, giving rise to more high-level associations (e.g., SUEDE-SERPENT is only 2 steps away in the ACN but farthest, i.e., 4 steps away in the SDNs). The SDNs, on the other hand, capture *direct* associations between words. Importantly, given that all networks had comparable fits, it seems that each network captured different sources of variance in the task.

It is possible that the ACN may be differentially sensitive to semantic relationships if a different criterion for network construction was used, or possibly in a conceptually driven semantic task, which would suggest that different types of stimuli/tasks emphasize different properties of the lexicon. Indeed, Gruenenfelder, Recchia, Rubin and Jones (2015) recently argued for a hybrid representation of semantic memory and suggested that individuals switch between a contextual representation and associative networks when generating free associations. Our results suggest that there may also be differences in how individuals use semantic representations in tasks that do not explicitly involve word association but are still sensitive to semantic relationships.

Another important goal of the current study was to investigate how network-based models of semantic representation compare to a distributed model, *word2vec*, which has been shown to explain human performance in some semantic tasks. Our results indicate that *word2vec* successfully captures similar patterns of behavior as the semantic networks. However, we also observed important differences in the semantic relationships captured by each of the models. For example, the word BOXING is 2 steps away from the word SPLINTER in the Undirected SDN but is very weakly associated in the *word2vec* space with a cosine of -0.022. Thus, there appear to be differences in the type of semantic information the models capture, e.g., the path from

BOXING to SPLINTER is mediated by the word PAIN in the association networks, but it is possible that this particular usage of SPLINTER does not co-occur in the same contexts as BOXING very often, which is the mechanism underlying *word2vec* model. Thus, these findings indicate that the nature of the task as well as the underlying representation are both critical variables that determine the extent to which semantic models explain human performance. Importantly, the tasks in the current study focused on semantic priming, and it is possible that spatial models and correlation networks are most useful in conceptual tasks like verbal analogies.

There were some limitations to the current study. First, the Hebrew network used in Kenett et al. (2017) was based on responses from a continuous free association task, whereas the Nelson et al. norms are based on a discrete free association task. The validity of both continuous and discrete responses has been debated (Hahn, 2008; Nelson, McEvoy & Dennis, 2000) and our use of discrete responses may have produced a different network structure than one based on continuous responses. However, given that the English ACN and SDNs were created from the same norms, we believe that the differences observed between the ACN and the SDNs were not critically influenced by the nature of associative responses per se, although this issue deserves further exploration. Further, the *word2vec* model was trained on a Google News corpus, which is very different from the Nelson et al. database, and the type of corpus can impact how well semantic models account for performance (Recchia & Jones, 2009). Thus, the nature of the task, stimuli and training corpora are all likely to influence the extent to which semantic models explain cognitive task performance.

In conclusion, the current set of experiments investigated the predictive power of path lengths derived from three large semantic networks and cosines derived from a neural network model in two behavioral tasks and provided strong evidence for multiple-step priming. We also demonstrated important structural differences between correlation-based networks and simple association networks and showed that simple association networks are also able to capture relatively distant semantic relationships. Finally, we showed that *word2vec* successfully captures similar behavioral patterns across two tasks. However, based on preliminary analyses, it appears that *word2vec* and the ACN are more likely to capture higher-level semantic representations, whereas simple step networks are more likely to capture direct associations. Clearly, further work is needed to substantiate these observations.

References

Anderson, J. R. (2000). *Learning and memory: An integrated approach.* (2nd ed.). New York: Wiley.

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., ... & Treiman, R. (2007). The English lexicon project. *Behavior Research Methods*, *39*(3), 445-459.

Chen, D., Peterson, J. C., & Griffiths, T. L. (2017). Evaluating vector-space models of analogy. *arXiv preprint arXiv:1705.04416*.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*(6), 407.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*(2), 211.

Gruenenfelder, T. M., Recchia, G., Rubin, T., & Jones, M. N. (2016). Graph-theoretic properties of networks based on word association norms: implications for models of lexical semantic memory. *Cognitive Science*, *40*(6), 1460-1495.

Hahn, L. W. (2008). Overcoming the limitations of single-response free associations. Electronic Journal of Integrative Biosciences, 5(1), 25-36.

Hutchison, K. A., Balota, D. A., Cortese, M. J., & Watson, J. M. (2008). Predicting semantic priming at the item level. *The Quarterly Journal of Experimental Psychology*, *61*(7), 1036-1066.

Jones, M., Gruenfelder, T., & Recchia, G. (2011, January). In defense of spatial models of lexical semantics. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 33, No. 33).

Kenett, Y. N., Levi, E., Anaki, D., & Faust, M. (2017). The semantic distance task: Quantifying semantic distance with semantic network path length. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(9), 1470.

Kenett, Y. N., Kenett, D. Y., Ben-Jacob, E., & Faust, M. (2011). Global and local features of semantic networks: Evidence from the Hebrew mental lexicon. *PloS one*, *6*(8), e23912.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Nelson, D. L., McEvoy, C. L., & Dennis, S. (2000). What is free association and what does it measure?. *Memory & Cognition*, 28(6), 887-899.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 402-407.

Recchia, G., & Jones, M. N. (2009). More data trumps smarter algorithms: Comparing pointwise mutual information with latent semantic analysis. *Behavior Research Methods*, *41*(3), 647-656.

Smith, E. E., Shoben, E. J., & Rips, L. J. (1974). Structure and process in semantic memory: A featural model for semantic decisions. *Psychological Review*, *81*(3), 214.

Steyvers, M., & Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science*, *29*(1), 41-78.

Garnering Support for Number and Area as Integral Dimensions

Lauren Aulet
Emory University, Atlanta, Georgia, United States

Stella Lourenco
Emory, Atlanta, Georgia, United States

Abstract

Non-numerical magnitudes such as cumulative area, element size, and density influence the perception of number. However, it is unclear whether interactions between number and non-numerical magnitudes reflect independent representations that interface vis--vis other systems (e.g., language) or, conversely, reflect holistic perception of number and other magnitudes. In the present work, we found converging evidence that number and cumulative area are perceptually integral dimensions. Whether assessed explicitly (Experiment 1) or implicitly (Experiment 2), perceived similarity for dot arrays that varied parametrically in number and area was best modeled by Euclidean, as opposed to city-block, distance. Critically, we also found that the integrality of number and area is comparable to other integral dimensions (Exp. 1: brightness/saturation; Exp. 2: radial frequency components), but different from separable dimensions (Exp. 1: shape/color; Exp. 2: thickness/curvature). In summary, these findings suggest that non-symbolic number perception is holistic, such that the processing of non-numerical magnitudes is obligatory.

A computational model of feature formation, event prediction, and attention switching

Eman Awad and **Fintan Costello**
School of Computer Science,
University College Dublin,
Belfield, Dublin 6, (eman.awad@ucdconnect.ie,fintan.costello@ucd.ie)

Abstract

In this paper we present a model of three central aspects of probabilistic cognition: event prediction, feature formation, and attention allocation. While most models of probabilistic reasoning take a parameter estimation and error minimisation approach (sometimes referred to as 'predictive coding', and often described in terms of Bayesian updating), our model takes a contrasting frequentist hypothesis-testing approach. This choice is motivated by a series of recent results suggesting that people's probabilistic reasoning follows frequentist probability theory. In simulation tests we demonstrate that this frequentist model, in which predictive features are formed by a process of null hypothesis significance testing, can give a successful account of event prediction and attentional switching behaviour.

Introduction

There are, broadly speaking, two approaches to statistical reasoning: a 'parameter estimation' approach (associated primarily with Bayesian statistics), where some form of generative model is used to predict data and the estimation process involves adjusting parameters of this model so as to reduce errors in prediction; and a 'hypothesis testing' approach (associated primarily with frequentist statistics) where a decision is made to reject a hypothesis (that is, to reject a possible generative model) when the probability of the observed data under that model is less than some significance level. Most current models of probabilistic cognition, learning and attention take the parameter estimation and error minimisation approach, sometimes referred to as 'predictive coding'; this approach is naturally described in terms of Bayesian priors (values of generative model parameters) which are 'updated' by experience, to produce more accurate posterior estimates of those parameters (see e.g. Clark, 2013; Griffiths and Tenenbaum, 2006; Tenenbaum et al., 2011; Miller et al., 1995).

In this paper we present a model of probabilistic cognition based on frequentist hypothesis-testing rather than parameter estimation and error minimisation. We apply this model to the processes of probabilistic learning, feature formation, event prediction, and attention. There are three motivations for this frequentist hypothesis-testing approach to probabilistic cognition. First, the contrasting parameter estimation and hypothesis-testing approaches to statistical reasoning are known to have different strengths and weaknesses: modelling probabilistic cognition via frequentist hypothesis-testing is worthwhile because it allows us to see this type of cognition in a new light.

Second, the hypothesis-testing approach applies very naturally to one core aspect of probabilistic cognition; that of decision making. Decision making is central to feature formation (given observed pattern of co-occurrence between events, how do we decide whether to treat that pattern as representing a single complex event, and so form a feature representing that event?), event prediction (given estimated probabilities of various future events or outcomes, how do we decide which event to predict?) and attention (given multiple sources of information, how do we decide whether to direct our attention to one source rather than another?) The frequentist hypothesis-testing approach was specifically developed to guide decision-making on the basis of data (see e.g. Fisher, 1937), and so provides a natural normative framework for modelling decision making in prediction, feature formation and attention allocation.

Finally, this model is motivated by recent evidence suggesting that people's probabilistic reasoning processes follow the requirements of frequentist probability theory (Costello and Watts, 2018a, 2016), and that a range of well-known biases in probabilistic reasoning can be explained as a consequence of regression produced by random variation or noise in normatively correct frequentist reasoning (Costello and Watts, 2014, 2018b). The model described here represents a computational implementation of this account; in this model random variation arises simply as a consequence of sampling.

We present this model incrementally, focusing first on prediction, feature formation and probabilistic learning for a single 'stream' of input (that is, with fixed attention). We then generalise to learning, feature formation and prediction across multiple simultaneous streams of input (where attention moves from stream to stream). We test the frequentist approach by comparing the effectiveness of an attention switching mechanism derived from frequentist hypothesis testing against the effectiveness of a switching mechanism based on error minimisation (as used in predictive coding), and against a random switching baseline.

The model

At an abstract level, temporal prediction involves taking a temporally ordered stream of categorical events or labels, such as

$$A,B,S,A,B,S,A,-,S,A,-,S,-,A,B,S,A,B,A,B \qquad (1)$$

and predicting the next event in the stream. Our model predicts future events in such sequences by constructing features from observation of a given stream, identifying features which allow statistically reliable predictions, and then combining these 'predictive features' to give an overall predicted probability for the next event in the stream.

Each feature in our model consists of an antecedent event A, a consequent event S, and a time interval t between them. Each feature also holds two counts: k, a count of the number of times A has been followed, after time t, by S; and n, a count of the number of times A has been followed, after time t, by any event. Finally, each feature holds a conditional probability $P(S|A) = k/n$, representing the probability of seeing the consequent S at time t after the occurrence of the antecdent A.

The antecedent A in a given feature may be a single event (e.g. the label A as a predictor of the next event, in our example in (1)), or may be a combination of events occurring over time (e.g the consecutive labels A,B as a predictor of the next event). Our model stores, in 'Short Term Memory' (*STM*), the N most recent events in the stream. Our model stores, in 'Long Term Memory' (*LTM*), a large number of simple or complex features that have been observed, with some features marked as 'reliable', meaning that there is statistically significant evidence supporting the relationship between antecedent A and consequent S in that feature. These reliable features are used to make predictions about the next event in the stream. The model has two free parameters: N, the size of short term memory, which we set by default at 4, and c, the significance criterion, which we set by default at $c = 0.05$.

Reliable features and prediction

To decide whether a given feature describes a statistically reliable relationship between antecedent A and consequent S, our model follows the hypothesis-testing approach of standard frequentist probability theory. We consider two possible cases: one where the antecedent event is a single 'atomic' event; and one where the antecedent is a complex or composite event, made up of multiple subevents.

In the case of a single event as antecedent A, we have two hypotheses: a null hypothesis (that there is no relationship between A and S; under this hypothesis the probability getting S after A is simply the base probability of event S, $P(S)$) and an alternative hypothesis (that there is a reliable relationship between A and S; under this hypothesis the probability seeing S after A is given by $P(S|A)$). The probability of obtaining k instances of S after A in a sample of n occurrences of A, assuming that $P(A) = p$, is given by the binomial

$$Bin(k,x,p) = \binom{n}{k} p^k (1-p)^{n-k} \tag{2}$$

This means that if $Bin(k,n,P(S)) < c$, for some critical significance level c, then we can reject the null hypothesis that $P(A) = P(S)$, and can instead accept the alternative hypothesis, that there is a reliable predictive relationship between A and S. When $Bin(k,n,P(S)) < c$ the model thus marks the feature linking A with S after time t as a statistically reliable predictive feature.

We now consider the situation where we have a complex event made up of sub-events A and B (each of which may itself be made up of further subevents), and where this complex event *AthenB* is itself an antecedent of our consequent S. Here we take k to represent the number of times consequent S has occurred at time t after antecedent *AthenB* in the observed time series, and n to represent the number of times any event at all has occurred at time t after antecedent *AthenB* in the series. In this situation we test against three possible 'null hypotheses': that $P(S|AthenB) = P(S)$, as before; that $P(S) = P(S|A)$ (that the probability getting S after *AthenB* is simply the probability of getting S after A, $P(S|A)$); and that $P(S) = P(S|B)$ (that the probability getting S after *AthenB* is simply the probability of getting S after A, $P(S|B)$). These three 'null hypotheses' are tested using the binomial as before. If all three tests are significant, the model concludes that the complex feature *AthenB* is itself a distinct, statistically reliable predictor of the occurrence of S: given that *AthenB* has occurred, the probability of S is given by $P(S|AthenB)$ (rather than by $P(S), P(S|A)$ or $P(S|B)$). By contrast, if there is no additional relationship between the feature *AthenB* and S beyond that given by A, B, and the base rate $P(S)$ then we can say that A and B (or some combination of their subevents) are independent predictors of the consequence S.

Using this hypothesis-testing procedure our model identifies, for a given sequence of events, the set of independent, statistically reliable, features occurring in that sequence which predict a given event S at the next timestep. Following standard frequentist probability theory the overall predicted probability of S is calculated by 'ORing' these independent predictions

$$Pr(S|A_1,\ldots,A_n) = 1 - \prod_{i=1..n} (1 - P(S|A_i)) \tag{3}$$

to give an overall predicted probability for S occurring next in the sequence.

An example

We can give an example of the model's operation for our example sequence in (1). In that series there are 20 events in total, of which 5 are S (so $P(S) = 0.25$). There are 3 occurrences of B followed one step later by S, and 1 occurrence of B followed one step later by a different event. To test the hypothesis that the occurrence of B predicts the occurrence of S at the next time step, we calculate the probability of seeing 4 occurrences of B, 3 of which are followed by S, if S was occurring at its base rate probability of 0.25: $Bin(3,4,0.25) = 0.0469$. Since this probability is less than our significance criterion of $c = 0.05$ we conclude that there is a statistically reliable relationship between B and S. Similarly, there are 5 occurrences of A followed two steps later by S, and 1 occurrence of A followed two steps later by a different event. To test the hypothesis that the occurrence of A predicts the occurrence of S

two steps later, we calculate the probability of seeing 6 occurrences of A, 5 of which are followed by S, if S was occurring at its base rate probability of 0.25: $Bin(5,6,0.25) = 0.0004$. This probability is also less than our significance criterion of 0.05 and we conclude that there is a statistically reliable relationship between A and S.

We also consider the complex predictive feature *AthenB*. There are 3 occurrences of *AthenB* followed one step later by S, and 1 occurrence of *AthenB* followed one step later by a different event. Since we've seen 4 occurrences of B, 3 of which were immediately followed by S, we estimate $P(S|B) = 3/4 = 0.75$. Testing against the hypothesis that the observed relationship between *AthenB* and S is explained simply be the presence of B we calculate the binomial probability $Bin(3,4,0.75) = 0.42$, and we see that there is no evidence for an additional relationship between the *AthenB* and S beyond that given by B. Since we've seen 5 occurrences of A, 4 of which were immediately followed by S, we estimate $P(S|A) = 4/5 = 0.8$, and since $Bin(3,4,0.8) = 0.41$ we similarly see that there is no evidence for an additional relationship between the *AthenB* and S beyond that given by A. We can thus conclude that the complex event *AthenB* is not a reliable predictor of S; instead, the two simple events A and B are independent predictors of the occurrence of S, with A predicting S after 2 steps with $P(S|A) = 0.8$, and B predicting S after 1 step with $P(S|B) = 0.75$. If events A, B have just occurred, the probability of S occurring next is obtained by ORing the predictions of these two statistically reliable features, giving

$$P(S) = 1 - (1 - P(S|A))(1 - P(S|B)) = 1 - 0.2 \times 0.25 = 0.95$$

as the predicted probability of S occurring at the next timestep in the stream shown in (1).

Switching between multiple streams

In this section we apply this model to feature formation and prediction across multiple different streams of input. We assume a single Long Term Memory (*LTM*), as before. To deal with multiple streams of input we assume multiple separate *STM* stores, one for each stream, and with each *STM* storing the last N events that have occurred in that stream. The model uses the statistically reliable features in *LTM* to calculate predicted probabilities for the next event in each input stream. The predicted next event for input stream i is calculated by finding statistically reliable predictive features in *LTM* whose antecedent event has occurred in STM_i (that is, in the store of recent events from stream i), and then combining the predictions from those features as described above.

Prediction, for each stream, happens in parallel and is computationally cheap (there are typically very few statistically reliable predictive features whose antecedents are present in a given *STM*). Learning and feature formation, however, are computationally 'expensive' and so take place only for one particular stream: the stream that is the current focus of attention. The model forms new features, updates antecedent and consequent occurrence counts, and identifies statistically reliable predictive features just as before, but only for this focal stream.

As the overall goal of the model is to accurately predict its environment (to accurately predict event occurrence in all streams), the model must occasionally switch its focus of attention from one stream to another. Attentional switching is a form of decision making: the model must decide to switch attention away from one stream of input (and so cease any predictive learning from events in that stream) and towards another stream of input (so beginning the process of learning from events in that new stream). The overall goal of the model is to form statistically reliable predictive features; satisfaction of this goal requires a decision process where attention is switched towards streams where statistically reliable predictive features are more likely to be formed, and away from streams where such reliable features are less likely to be formed. As before, frequentist hypothesis testing gives a natural and normatively correct way to make such switching decisions.

Suppose we are considering forming a reliable feature *ApredictsS*, and have observed k instances of A followed by S, out of n occurrences of A overall. This feature will be judged reliable when the observed pattern of co-occurrence between A and S has a low probability of occurrence under the null hypothesis that A does not predict S (when $bin(k,n,P(S)) < c$). This means that the lower the value of this binomial expression $bin(k,n,P(S))$, the more likely it is that the null hypothesis is false and the there is some reliable relationship between A and S. In other words, if we have we have some feature for which $bin(k,n,p) > c$ (some feature which is not yet reliable), then the probability that this feature will become reliable is proportional to $1 - bin(k,n,p)$. More generally, if a given stream i contains a number of not-yet-reliable features with counts k_j and n_j and null hypothesis values p_j, then the overall probability of forming a reliable feature in that stream is obtained by ORing the individual probabilities of each of these features becoming reliable, as given in the expression

$$F(i) = 1 - \sum_{\substack{\text{j= not yet} \\ \text{reliable} \\ \text{feature in} \\ \text{stream i}}} bin(k_j, n_j, p_j) \qquad (4)$$

The greater the value of this expression $F(i)$ for a given stream i, the greater the probability that switching attention to that stream will lead to the formation of a new reliable predictive feature. The model uses these values $F(i)$ to guide switching; at each timestep the model calculates $F(i)$, the probability of constructing a new reliable feature, for all streams i including the current stream x, and identifies the stream *max* with the highest value. If $F(max) - F(x) > s$, where s is a switching decision criterion (if the chance of forming a new feature in stream *max* is s greater than the chance of forming a new feature in the current stream) then the model switches attention to stream *max*; otherwise attention remains in the current stream.

Testing the Model

We test our model via Markov Chain Monte Carlo simulation, as follows.

Any process producing a series of events can be represented by an nth order Markov chain (for some value of n). Such chains thus represent realistic generative models of sequential event occurrence. For each stream of information in our simulation, we construct a generative n-th Markov chain with $m = 4$ (4 distinct categorical events) and $n = 4$ (the current state consists of the last 4 events). There are $4^4 = 256$ distinct states in this chain, each with 4 transition probabilities. For each state these 4 transition probabilities are assigned random values, normalised so their sum for that state equals 1. Each stream thus represents a (randomly constructed) Markov chain process.

We use the Markov chain to generate a large sequence of categorical events from that stream. We first pick an 4 initial events at random, representing the initial state of the Markov Chain. We identify the 4 transition probabilities associated with that state, and choose one transition (one new event) at random, proportional to its transition probability from the current state in the Markov Chain. The selected event is added to our sequence of generated events. The new state of our Markov model now consists of the 4 most recent events (three previous events and the event that was just added to the series), and the cycle repeats.

The events for each stream are fed in parallel to our model, which forms statistically reliable predictive features, makes predictions for the next event to occur at each time step in each stream, and switches attention between streams as described above.

After an initial training phase we continue running the model and the Markov Chain generators for an additional test phase. We gather, at each time step of this phase, the model's predicted probability for each event in each stream, and whether or not that event actually occurred. We assess the model by gathering together all cases where the model predicts that an event will occur with probability in some range R. If the model is accurate, the proportion of those predicted events that did actually occur should be in or near the range R. For example: we gather together all cases where the model predicted some event with probability in the range $0.1 - 0.15$. If the model's predictions are accurate, then the predicted event should have actually occurred around $10\% - 15\%$ of the time; the probability (or proportion) of actual occurrence of the predicted event should be close to the range $0.1 - 0.15$.

Test 1: Learning from a single stream

Table 1 shows results obtained when running the model with a single stream of input (no switching between streams), for a 5000 timestep training phase (during which the model formed predictive features) followed by a 5000 timestep test phase (during which we gathered the model's predicted probabilities for the next event, at each timestep). This table gives the proportion of times the model's predicted event occurred, for prediction ranges from $0.05 - 0.10$ to $0.55 - 0.60$. As the table demonstrates, the model's predicted probabilities corresponded closely with the actual probability (or proportion) of occurrence of the predicted event.

Table 1: This table shows the number of times our model predicted that an event would occur with a probability that fell into a given range **R** (column 2), and of those predictions, the number of times when the predicted event actually occurred (column 3). If the model is making accurate predictions, the proportion of occurrence of the predicted event (the observed probability, column 4) should follow the range value R. The two values are highly correlated ($r = 0.99$) indicating that the model is predicting event probabilities accurately.

Predicted probability range **R**	Number of predictions in **R**	Predicted event occurred	Observed probability of predicted event
0.05 - 0.10	1393	193	0.14
0.10 - 0.15	2268	401	0.18
0.15 - 0.20	2600	516	0.19
0.20 - 0.25	3016	6463	0.21
0.25 - 0.30	3901	1094	0.28
0.30 - 0.35	3341	1005	0.3
0.35 - 0.40	1806	597	0.33
0.40 - 0.45	788	272	0.35
0.45 - 0.50	307	117	0.38
0.50 - 0.55	171	64	0.47
correlation with probability range **R**			0.99

As Table 1 also demonstrates, there is regression in the model's predictions: for low predicted probability ranges, observed event probabilities tend to be significantly higher than the probability range, while for high predicted probability ranges, observed event probabilities tend to be significantly lower than the probability range. This pattern of regression in turn implies that the models predicted probabilities are regressive towards the center of the probability scale, relative to true event probabilities Erev et al. (1994). This pattern of regression is just as assumed in Costello & Watts frequentist account of probabilistic reasoning (Costello and Watts, 2014, 2016, 2018a,b). This model thus provides a mechanistic implementation of that account, in which regression arises as a consequence of random sampling variation.

Test 2: Learning from a multiple streams

To test the hypothesis-testing model of switching given above, we test the model in the same Random Markov chain regime, but with 5 parallel streams of input, each with its own randomly initialised Markov Chain generator. Specifically, we compare learning under this model against learning under random switching, and learning under an alternative 'predic-

Table 2: Average observed probability of occurrence of predicted event for prediction range *R*, across all streams. Observed probabilities are calculated as in Table 1. Data is given for each switching mechanism, running the model for 5000 times/steps in each case. Both the 'Switch to max error' and the 'switch to form reliable features' switching methods gave predictions closely correlated with observed probability.

Range **R**	Observed probability of predicted event for predictions in range **R** (by Switching method)		
	Random switching	Switch to max. error	Form reliable features
0.05 - 0.10	0.17	0.00	0.14
0.10 - 0.15	0.18	0.16	0.17
0.15 - 0.20	0.17	0.17	0.19
0.20 - 0.25	0.17	0.19	0.21
0.25 - 0.30	0.2	0.22	0.25
0.30 - 0.35	0.34	0.32	0.3
0.35 - 0.40	0.41	0.39	0.34
0.40 - 0.45	0.48	0.45	0.39
0.45 - 0.50	0.56	0.49	0.39
0.50 - 0.55	0.49	0.61	0.42
correlation	0.90	0.98	0.99

tive coding' model of attentional switching, where we switch attention to the stream where errors in event prediction are highest.

Random switching As a baseline for comparison, we run the model with a fixed-length random-choice method for switching between streams. Under this switching method, the model will remain in a certain stream for 100 timesteps at a stretch; after each sequence of 100 timesteps has passed, the model will switch to a randomly selected other stream. Given that the learning model performs well in learning to predict events in a single stream, we expect that the model will perform relatively well in predicting events across multiple streams under this random switching regime.

Switching to minimise predictive error As an alternative for comparison, we run the model with a switching method designed to minimise predictive error. In the predictive coding view, a learning model makes predictions which are compared with outcomes: attention is driven towards locations where those predictions are incorrect (and so more learning is required) and away from locations where predictions are accurate (and so less learning is needed). In our model, predictive error in a given stream at a given time is simply equal to the predicted probability of the event that actually occurred in that time: if S is the event that actually occurs and the model's prediction probability for S was high, then there is little predictive error; if S occurs and the model's prediction probability of S was low, there is significant predictive error.

Table 3: Average correlation between observed and predicted event probability, obtained after learning with each switching method, for runs of different length (500,1000,5000,10000 and 50000 times/steps). Both the 'Switch to max error' and the 'switch to form reliable features' switching methods gave predictions that were more closely correlated with observed even occurrence rates, with the 'switch to form reliable features' approach giving the highest average correlation between observed and predicted probability.

Run size	Correlation between observed and predicted probability (by Switching method)		
	Random switching	Switch to max. error	Form reliable features
500	0.85	0.96	0.88
1000	0.89	0.97	0.96
5000	0.9	0.98	0.99
10000	0.9	0.98	0.99
50000	0.9	0.96	0.99

To implement a switching mechanism based on predictive error, we give a method which calculates, for each stream i, the average predicted probability of the last N events that occurred in this stream. The lower this average, the more prediction error in stream i. Letting $G(i)$ be equal to 1 minus the average prediction probability for stream i, the model uses a decision criterion s to guide switching; at each timestep the model calculates $G(i)$ for all streams i including the current stream x, and identifies the stream *max* with the highest value. If $G_{max} - G(x) > s$ (if prediction error in stream *max* is s greater than that in the current stream) then the model switches attention to stream *max*; otherwise attention remains in the current stream.

Results

We ran the model separately with each of the three different switching mechanisms described above, and with different number of training and test timesteps (500, 1000, 5000, 10000 and 50000 timesteps: in each case the training phase was the same length as the test phase). As before, we grouped the model's predictive probabilities into a series of 'buckets' or ranges **R** (so one bucket would hold all cases where the model predicted some event with a probability between 0.05 and 0.1, another would hold all cases where the model predicted some event with a probability between 0.1 and 0.15, and so on). For each bucket we counted the number of times the predicted event actually occurred. If the model is accurate, the proportion of those predicted events that did actually occur (the observed probability of occurrence of predicted events), for a given range R should be in or near the range R (the predicted probability for those events).

Table 2 shows the results of this analysis of model prediction for the 5000 timestep run with each of the three switching methods. This table gives the observed probability of events whose predicted probability fell in range R, in runs of the model with each of the 3 possible switching methods. The observed probabilities shown here are averages across predictions made in all 5 parallel streams of input, in a single run of the model (observed probabilities in each stream closely follow the pattern seen here, and closely follow the pattern seen in Table 1). As this table shows, there was a reliable correlation between the range in which the model predicted an event will occur and the actual rate or ratio of occurrence of that event, for all switching methods. This is expected: as we saw in the earlier 'single-stream' simulation tests, the model does well in learning to predict events accurately from observed event sequences, and so we would expect the model to learn to predict all streams relatively well no matter what attentional switching mechanism was being used.

The table also shows that both the 'switch to form reliable features' and the 'switch to maximum error' methods give results that matched observed probabilities much more closely than those given by the 'random switching' mechanism. These results demonstrate the contribution that effective attentional switching can make to prediction accuracy.

Table 3 shows the correlation between model predicted probabilities and observed event probabilities for each switching method, across increasing training and test time. As this table shows, correlation between predicted and observed probabilities increased with learning to some degree for all switching methods, but increased to very high correlation values for both the 'switch to max error' and the 'switch to form reliable features' methods. Taken together, these results demonstrate that this hypothesis-testing model forms features which reliably predict future events, and switches attention in a way that maximises formation of such features.

Conclusions

We have described a computational model of prediction, feature formation, and attentional switching. This model is interesting because it is based on the frequentist, hypothesis-testing approach to statistical reasoning, as opposed to the parameter-estimation approach currently popular in models of these cognitive processes. This model represents a computational implementation of a general account of probabilistic reasoning, also based on frequentist probability (Costello and Watts, 2014, 2016, 2018a). That account sees human probabilistic reasoning as being based on normatively correct processes, but subject to random variation or 'noise': that noise has systematic regressive effects, producing a range of biases in people's probabilistic judgement. The computational model implemented here demonstrates just the pattern of regression assumed in that more general account, and so inherits its account for those biases.

The frequentist, hypothesis-testing approach described here may usefully address two problems with the standard parameter-estimation approach to probabilistic prediction. These problems arise because the Bayesian approach to learning and prediction often require the specification of initial priors, in two separate ways. First, such models require initial assumptions as to the form of the generative model being used to predict data (assumptions which specify which features are 'available' for use in prediction, for example, and which features are not). The hypothesis-testing approach described here in some ways avoids this requirement, by providing a mechanism whereby predictive features are 'built' out of observed event. Second, such models require assumptions as to the initial values of parameters in that generative model (assumptions about the initial, prior, probability distribution associated with features in the generative model). The hypothesis-testing approach described here avoids this requirement also, because features are not identified as 'reliable' (and so do not contribute to predictions) until the events making up those features have been repeatedly observed (that is, until any initial 'prior' has been made irrelevant by repeated experience with the events in question). These points suggest that an integrated approach, combining the parameter-estimation and the hypothesis-testing perspectives, may prove insightful. Understanding the interplay between hypothesis testing and parameter estimation in human probabilistic reasoning is an important aim for future work.

References

Clark, A. (2013). Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences*, 36(3):181–204.

Costello, F. and Watts, P. (2014). Surprisingly rational: probability theory plus noise explains biases in judgment. *Psychological review*, 121(3):463.

Costello, F. and Watts, P. (2016). Peoples conditional probability judgments follow probability theory (plus noise). *Cognitive Psychology*, 89:106–133.

Costello, F. and Watts, P. (2018a). Invariants in probabilistic reasoning. *Cognitive psychology*, 100:1–16.

Costello, F. and Watts, P. (2018b). Probability theory plus noise: Descriptive estimation and inferential judgment. *Topics in cognitive science*, 10(1):192–208.

Erev, I., Wallsten, T. S., and Budescu, D. V. (1994). Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological Review*, 101(3):519–527.

Fisher, R. A. (1937). *The design of experiments*. Oliver And Boyd; Edinburgh; London.

Griffiths, T. L. and Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological science*, 17(9):767–773.

Miller, R. R., Barnet, R. C., and Grahame, N. J. (1995). Assessment of the rescorla-wagner model. *Psychological bulletin*, 117(3):363.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., and Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *science*, 331(6022):1279–1285.

Transferability of calibration training between knowledge domains

Christopher Babadimas, Christopher Boras, Nicholas Rendoulis, Matthew Welsh & Steve Begg
([Christopher.babadimas; Christopher.boras; Nicholas.Rendoulis] @student.adelaide.edu.au; [matthew.welsh; steve.begg] @adelaide.edu.au)
Australian School of Petroleum, University of Adelaide, Adelaide, SA 5005, Australia

Abstract

Many industry professionals are poorly calibrated, overestimating their ability to make accurate forecasts. Previous research has demonstrated that an individual's calibration in a specific domain can be improved through calibration training in that domain; however devising a training program for each specific domain within a field is laborious. A more efficient method would be if individuals from different disciplines could undertake the same general training and transfer the skills learnt to their respective, specific domains. This study investigated whether calibration training in a general domain was transferable to the specific domain of petroleum engineering. The results showed that, whilst the feedback training was effective within the general domain, there was only limited transfer to the specific domain. This is argued to be due to recognition failure, where the participants failed to recognise that the skill learnt through training in the general domain could be transferred to the specific domain.

Keywords: calibration; overconfidence; training; skill transfer.

Introduction

In technical disciplines and industries, individuals are required to provide range estimates, such as 80 percent confidence intervals, for uncertain parameters used in modeling and decision making (see, e.g., Capen, 1976). The accuracy of the individual's estimates can greatly influence decisions, with significant impacts on company bottom lines (see, e.g., Welsh, Begg & Bratvold, 2007). Calibration is the measure of how well individuals' estimates match real world outcomes (Lichtenstein, Fischhoff & Phillips, 1982). For example, if a weather forecaster makes multiple predictions of an 80% chance of rain, and on 80% of those occasions it does rain, they are well calibrated (for 80%), meaning they have a higher likelihood of providing more accurate estimates, which lead to more informed decisions.

Poor calibration in range estimation tasks can result from cognitive biases (e.g., biases from the anchoring and availability heuristics; Tversky & Kahneman, 1974) and is described as overprecision - one type of overconfidence bias (Moore & Healy, 2008). Overprecision describes the observation that individuals provide overly narrow ranges that do not represent their true degree of knowledge (Moore, 2008). The tendency for individuals to be over-precise in estimation has been demonstrated repeatedly (e.g., Soll & Klayman, 2004; Lichtenstein *et al.,* 1982) and seems to affect experts similarly to novices (McKenzie, Lierch & Yaniv, 2008). (NB, many studies use the term 'overconfidence' rather than overprecision and, in order to stay consistent with past literature, this will be done hereafter.)

Calibration Training

Past research has shown calibration can be improved through debiasing techniques, the most effective being domain-specific performance feedback training, wherein a subject receives timely feedback on the accuracy of their estimates within a particular area of knowledge (e.g., a field like petroleum engineering or meteorology; see, e.g.: Adams & Adams, 1958; Fischhoff, 1981) or learns this over an extended period in an amenable environment (see, e.g., Tetlock & Gardner, 2016). Whilst domain-specific training may be effective, devising training programs for numerous specific domains within a wider field or industry is laborious. For example, oil industry personnel include engineers and geoscientists across various specialties and a generalised training program, with calibration training learnt in a general domain and learnings transferred to specific domains, would be a more efficient method of improving calibration for a company employing these people.

Despite this previous research on domain-specific performance feedback training, it has seldom extended to the idea of creating generalised performance feedback training. Adams and Adams (1961) showed that training a subject's calibration in a series of tasks lead to an improvement in calibration in a separate task, an idea termed "generalisation"; although the degree of improvement in the untrained task was lower than in the trained task. Lichtenstein & Fischhoff (1980) showed that calibration training in a base task improved calibration in other, similar, tasks but not on dissimilar tasks, which was attributed to the subjects' inability to spontaneously relate the new task to the base task.

Similarly, Bornstein & Zickafoose (1999) demonstrated that individuals' confidence and accuracy were stable across domains of general knowledge and eyewitness memory, and that training using general knowledge questions reduced overconfidence in eyewitness memory. Improvements in calibration and resolution, however, were not observed, implying no improvement in accuracy. Thus, the above studies suggest that generalised training could be effective but, given inconsistent results and the fact that this was not their primary focus, the question of whether generalised training transfers to specific domains remains open.

An argument supporting the plausibility of the generalizable calibration training is analogical transfer,

where studies have shown transfer of knowledge (although not calibration training) across domains. Analogical transfer involves the use of a familiar problem to solve a novel problem of the same structure (Reeves, 1994). By identifying similarities in the structure of base and target problems, a subject can transfer the principles of the base problem to solve the target problem (Glick & Holyoak, 1983). Analogical transfer is argued to be the main method used to solve novel problems in all domains (Rumelhart, 1989). Given this, if the process of improving calibration training can be stripped down to its base structure, analogical transfer may facilitate transfer of calibration training across domains.

Whilst the structural similarities of the base and analogue problems are essential to facilitate transfer, they do not guarantee *recognition* of the relationship, which would prevent spontaneous transfer from one problem to the next without instructions or help (Day & Goldstone, 2012). Recognition failure is argued to occur largely as a result of dissimilar surface elements in the respective problems (Day, 2012) – for example, questions drawn from different domains - but may be improved by providing multiple base problems, as this will allow the subject to derive a more general analogy (Glick & Holyoak, 1983). Recognition failure may provide an explanation for the limitations in generalisation seen in Lichtenstein & Fischoff (1980), and Bornstein & Zickafoose (1999). Conversely, Adams and Adams (1958) achieved moderate generalization - using training in multiple, different tasks.

Given the paucity of research into the generalization of calibration training and the apparent absence of research connecting transferability of calibration training to analogical transfer, this paper has the opportunity to fill a distinct research gap. The research is further warranted by the paper's focus on the practical issue of how best to provide training. That is, seeing whether analogical transfer facilitates calibration training transferring to a new domain is both practically and theoretically interesting.

Aims

Given the unclear evidence in the literature, this paper's primary aim is to see whether generalised training in calibration can be developed to enable transfer of improved calibration to problems in a different, specific knowledge domain – specifically, petroleum engineering. This leads to two main hypotheses, as shown below:

H1: Calibration training will improve calibration within the domain in which the training is given.
H2: Improvements in calibration training will transfer to a new, specific domain.

It is important to highlight that the term "general domain" is used to describe a domain, unrelated to the specific domain, in which training will be given. The term generalized training thus refers to training applied in the general domain. In the context of a real world application it makes sense for the general domain selected to be general knowledge, as this domain is accessible to all, and is clearly separate from a subject's specific domain of expertise. A general domain in this context could, however, be any domain other than the participant's specific domain.

Methodology

Participants

Participants were 54 (15F and 39M) recent (n=7) and current (n=47) students of the Australian School of Petroleum, University of Adelaide, ranging in age from 18 to 35 (M=22, SD=3.0). Previous experience with calibration varied amongst the participants, with 31 participants having previously undertaken a course that taught calibration, and 15 who had not undertaken the course but who indicated (prior to the study) that they understood what calibration was and how it affects decision making. Participants entered a draw (1 in 6 chance) to win one of several $200 gift cards.

Materials

Testing materials consisted of three questionnaires - two general knowledge, and one in the domain of petroleum engineering - and a feedback/training package (described below). In this scenario, petroleum engineering is the specific domain, and general knowledge the general domain. Petroleum engineering was chosen to be specific domain due to the higher level of expertise in this field (compared to the general populace) shared by all participants. This higher level of experience is expected to elevate their knowledge of this domain above the participant's understanding of more general knowledge; separating it from the general domain. In terms of knowledge transfer, the assumption is that participants may think differently about their area of specialty than general knowledge questions and, thus, that recognition failure across the two domains may be more likely.

The first general knowledge test – designated "Pre-Training" - contained 30 questions; however, the number of questions in the remaining tests (designated "Post-Training" and "Domain Specific") were reduced to 20 each following participant feedback. The tests consisted of questions that had definite numerical answers and were sufficiently difficulty for participants not to simply know the true answer. An example of a question used in the general knowledge domain (i.e., the Pre-Training and Post-Training questionnaires) was "How many countries does the Nile River cross over?" For comparison, an example question used in the Domain Specific questionnaire was "How many times greater is the Young's Modulus of a stiff sandstone compared to the Young's Modulus of coal?"

In all cases, participants were asked to provide a low and a high value such that they were 80% confident their range would contain the true value. (The initial page of each test provided information about how to answer the questions, including an example question.)

While the second test is designated "Post-Training",

feedback materials were provided only to the Experimental Group. This was a pdf document, consisting of: information about calibration and overconfidence, a calibration curve illustrating the subject's under/overconfidence, a histogram showing the subject's calibration score relative to other participants, and a graphical depiction of the subject's confidence intervals, plotted against the corresponding true answers. These figures were intended to help participants understand the degree of overconfidence they had shown in the Pre-Training test. Each figure was accompanied by a short explanation, and information on methods for improving calibration on the remaining tests – including recognition of their current calibration in order to prompt them to give wider ranges.

Procedure

After registering their interest, participants were provided with links to access the Pre-Training questionnaire online (on SurveyMonkey) with instructions to complete each question by providing 80% confidence interval estimates. Based on the results of the Pre-Training questionnaire, participants were divided into two groups with similar levels of calibration. Feedback training was then distributed to the Experimental Group via email, at most two weeks after completing the test, with instructions to read and understand the material completely before continuing to the general knowledge Post-Training questionnaire). To test whether participants understood the feedback, a four-question quiz was given on the material covered in the training package. Participants who scored less than 3 out of 4 (2 participants) were moved from the Experimental Group to the control group, as it was adjudged they had not read the material and hence not received the feedback (NB – while recognizing that removing the participants may have been a more appropriate, this choice was made in light of the already small sample). Links to the Post-Training and the Domain Specific questionnaires were then provided to participants straight after the feedback training was distributed.

Improvements in calibration due to the feedback training were measured by comparing the Experimental Group's Post-Training and Domain Specific questionnaires to baselines of the Experimental Group's Pre-Training questionnaire and the Control Group's Post-Training and Domain Specific questionnaires. Comparisons were made under the assumption that the tests were of equal hardness and both groups were equally well calibrated. This yields measures of both the effectiveness of the feedback training and the transferability of the training across domains.

Results

Descriptive Statistics

Table 1 shows the demographics for the experimental and control groups while Figure 1 shows the mean calibration achieved by each group under each condition (with 95% CIs - recalling that questions asked for 80% ranges meaning numbers under 80% reflect overconfidence). Prior experience refers to the knowledge the participants had acquired regarding calibration and overconfidence prior to this experiment's start. Participants who answered 'Yes' indicated they had received prior training or learning regarding calibration and overconfidence. 'Partial' referred to participants who believed they understood the concepts at least vaguely. 'No' referred to the participants believing they had no understanding of calibration or overconfidence.

Table 1. Descriptive statistics for demographic variables including prior knowledge of calibration by group

	Overall	Control	Experimental
N	54	29	25
Gender (%)	M: 72 F:28	M: 72 F: 28	M: 72 F: 28
Age (SD)	22.0 (3.0)	22.7 (1.3)	22.3(4.2)
Prior Experience (%)	Yes: 57 Partial:28 No:15	Yes: 72 Partial:24 No: 4	Yes: 40 Partial:28 No:32

Looking at the figure, ones sees clear evidence of overconfidence across both groups and tests with none of the 95% CIs containing the 'expected' 0.8 proportion correct. The two groups seem to show similar levels of calibration on the Pre-Training questionnaire and Domain Specific Test but differ on the Post-Training questionnaire.

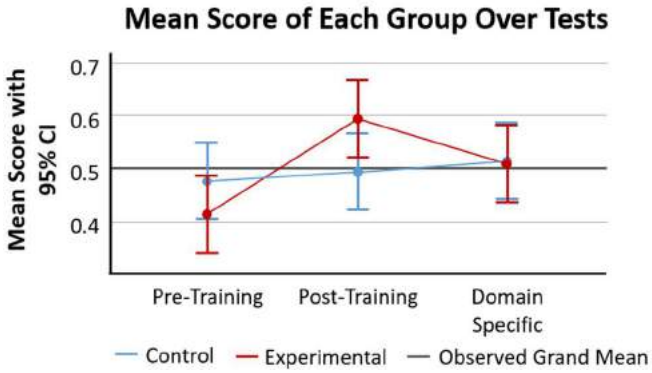


Figure 1. Calibration by group and condition.

Repeated Measures ANOVA

A Repeated Measures ANOVA was used in SPSS to test the two hypotheses simultaneously. Table 2 summarises the significant results from this.

As shown in Table 2, participant's calibration scores differ across the three tests and there is also an interaction between test and group – supporting the observations made above. Independent samples t-tests were used, post-hoc, to compare the mean calibration scores of the Control Group and Experimental Group for each of the three tests as shown in Table 3. The tests indicated that, for both the Pre-Training questionnaire and Domain Specific questionnaire, the difference observed in mean calibration score between the Experimental and Control Group was not significant. However, there was a significant difference in the means of the Experimental and Control Groups on the Post-Training

questionnaire with the mean calibration score of the Experimental Group noticeably higher. This supports Hypothesis 1 – that the feedback training improved the experimental group's Post-Training questionnaire results.

Table 2: Significant results of RM ANOVA

Comparison	F(df)	F-value	P-value
Questionnaire	F(2,104)	10.4	<0.001
Questionnaire*Group	F(2,104)	6.1	0.003

Table 3: Independent t-tests between Experimental and Control Group for each questionnaire.

Questionnaire	t(52)	p
Pre-Training	0.925	.359
Post-Training	-2.189	.033
Domain Specific	0.164	.871

Paired samples t-tests were used, post-hoc, to compare the relative difficulty of the tests, and to verify improvements in calibration observed in the independent samples t-tests. The tests, shown in Table 4, indicated that differences in the means between all the Control Group's tests were non-significant. That is, the tests were equally difficult for the Control group. Conversely, the tests indicated that differences in the means between all Experimental Group tests were significant – as shown in Table 5.

Table 4: Paired t-tests between each questionnaire for the Control Group.

Comparison	t(28)	P
Pre-Training– Post-Training	0.852	.402
Post-Training– Domain Specific	0.627	.536
Domain Specific– Pre-Training	1.315	.199

Table 5: Paired t-tests between each questionnaire for the Experimental Group.

Comparison	t(24)	p
Pre-Training– Post-Training	-5.368	0.000
Post-Training– Domain Specific	3.633	0.001
Domain Specific– Pre-Training	-2.439	0.023

The results of the ANOVA and t-tests, along with observation of Figure 1, suggest no significant difference in calibration scores in the Control Group – as would be expected. However, the figure and analyses show that calibration score for the Post-Training questionnaire of the Experimental Group is significantly higher than both the Experimental Group's Pre-Training questionnaire, and the Control Group's Post-Training questionnaire, which is taken as evidence that feedback training improved calibration.

The near-identical scores of the Control and Experimental groups on the Domain Specific questionnaire, however, suggests this benefit did not transfer to the new domain. That is, despite all tests using the same question format (80% confidence intervals), the change in domain was seemingly sufficient to prevent the training transferring, meaning Hypothesis 2 was not supported. A caution to this interpretation, however, is the observation that the Experimental Group's calibration in the Domain-Specific questionnaire was significantly higher than in the Pre-Training questionnaire, but statistically no different to Control Group's calibration in the Domain-Specific questionnaire. This discrepancy is explored further, below.

Discussion

Experimental Findings

Baseline Measure

The performance on the Pre-Training questionnaire between the Experimental Group and the Control Group suggested both groups were similarly calibrated, indicating that the method for dividing participants into two groups was successful and that the control group can, justifiably, be compared to the experimental group as a baseline.

The consistent results of the Control Group across all tests similarly showed that each test was of similar difficulty, justifying comparisons between tests within a group.

Feedback Effectiveness

The comparison between the mean calibration scores of the Pre-Training questionnaire and Post-Training questionnaire of the Experimental Group shows that the feedback was effective - to a degree. This was reinforced through the comparison of the Experimental Group and the Control Group for the Post-Training questionnaire, which also found a significant result. Between the Pre-Training questionnaire and the Post-Training questionnaire for the Experimental Group, calibration scores improved by 17% (from 42% to 59%). This improvement in calibration was expected, as a wealth of previous research has shown that performance feedback training improves a subject's calibration (Adams & Adams, 1961; Lichtenstein & Fischhoff, 1980; Moore et al., 2017; Stone & Opel, 2000).

Transfer

As noted above, the comparison of the Experimental and the Control Groups on the Domain-Specific Test showed no significant difference, suggesting the Experimental Group was *not* able to transfer their knowledge of calibration to a different domain and thus arguing for recognition failure.

Comparing the Experimental Group's Pre-Training and Domain Specific results, however, showed a significant result driven by an approximately 8% increase (from 42% to 50%) in the Experimental group's mean calibration score. Considering the two tests were of similar difficulty – according to the baseline measure from the Control group - the significant increase in calibration suggests participants *were* able to, at least partially, transfer their skills from the general knowledge domain to the specific domain of petroleum engineering and that this is being obscured in the analyses above by the Experimental Group's slightly lower scores on the Pre-Training questionnaire.

To quantify the extent of the transfer, the Post-Training

questionnaire was compared to the Domain-Specific for the Experimental Group. This showed an ~10% decrease in mean calibration score from the Post-Training questionnaire to the Domain-Specific Test (60% compared to 50%, respectively). This significant difference suggests participants were not able to transfer all of what they had learnt about improving calibration to the new domain. Looking solely at the Experimental group's results in Figure 1 suggests that about half of the improvement seen following training transferred to the Domain Specific Test.

This, of course, contradicts the previous results and the discrepancy between these means that no strong conclusion can be drawn regarding whether the transfer of knowledge between domains did or did not occur. However, one conclusion that can be drawn is that, if the transfer occurred, it is well under 100%, in agreement with Adams & Adams (1961). This is also reminiscent of Glick & Holyoak's (1983) work on analogical transfer, where they argue that incomplete transfer may be due to recognition failure; that is, a failure to recognise the similarities in the problem structure and, hence, to recognise that the skills used successfully in one problem are applicable to the other.

While the question formats used in the three tests herein were identical – asking for 80% confidence intervals - it is possible that having experience in a domain evokes a different thought process to that which may be used to solve general knowledge type questions - reminiscent of knowledge partitioning (Lewandowsky & Kirsner, 2000) - and suggesting that individuals' domain specific knowledge could be separated from their general knowledge and thus processes used to access one may not work for another.

This may have caused the participants to not recognise the similar structure between the general knowledge and specific domain type questions. That is, participants may have simply not recognised that their calibration training should also be applied to the specific-domain questions.

External Factors

Initial Calibration and prior knowledge

Simple comparisons showed participants, regardless of their stated prior experience with calibration (trained, aware or unaware) had similar calibration, and similar improvement after feedback. This is likely due to participants with prior knowledge not being able to apply the knowledge they learnt previously when setting confidence intervals. These results suggest that participants with previous experience with calibration were unsuccessful in reducing their overconfidence long-term, likely due to the fact they did not receive frequent calibration training or regularly practice calibration – as has been observed in previous research (see, e.g., Welsh, Bratvold & Begg, 2005).

Caveats

Sample Size

The sample size was smaller than hoped, as a result of strict time constraints for the project, meaning that statistical power is low. A larger sample might, for example, have helped determine to what extent transfer was actually occurring or whether the effect is an artefact of differences between groups and tests aligning coincidentally. As noted above, the low sample size also resulted in the decision to move participants from the experimental group and control group.

Expertise

The type of questions asked throughout the Domain-Specific Test were designed to relate to the expertise of the participants. As petroleum engineering students, participants have increased knowledge about the petroleum engineering field, but would not be classified as 'experts'. This is doubly true, as the sample includes student participants from different year levels and thus with differing amounts of learning within the field. This concern is somewhat alleviated by the fact that the majority of participants were final year students or recent graduates, who could be expected to have similar levels of understanding of the field (which might, in fact, be less true of professionals further into their careers who tend to specialise into a sub-field). The selection of students of all year levels as the sample, however, meant that, despite all of the questions being related to the oil and gas industry, they had to be kept general enough that all participants could reasonably understand what they referred to – rather than being specific, technical questions that only a fully trained petroleum engineer could understand. That is, while the questions were *about* petroleum engineering, they did not truly test fundamental skills learnt by the participants. Questions more central to the petroleum engineering domain would provide a more accurate measure of knowledge transfer across domains but would require an expert sample.

Testing Conditions

As noted, all tests were online, meaning participants were unable to ask clarifying questions if they did not understand the point of the test - or may have approached the test in unanticipated ways. Although instructions indicated that questions should take no longer than 30 seconds, many participants spent much longer than that on some questions, which may have repercussions on the consistency of the answers. Future work could, therefore, be conducted face to face, or more time be spent explaining the purpose of the test, possibly with the aid of a video. This would make it easier to see if participants are engaged in the test and answering the questions as expected. Conducting training feedback sessions in person could also be beneficial in ensuring that the main points of the training session are highlighted to the participant, so that they can better learn how to improve their calibration for future tests.

Questions

Another concern related to the amount of time available to pilot the general knowledge questions with people similar to the expected participants. As noted elsewhere, the study was conducted as part of the student authors' coursework and, as

such, had to be completed within semester, meaning that time for piloting was limited. While efforts were made to include a wide variety of questions participant feedback indicated there were several questions where some participants did not understand what the question was asking, and hence had no point of reference.

Future Research

As noted in the acknowledgements, this experiment was conducted as part of the first three authors' final-year, undergraduate, research project. As such, there were strict time and budgetary constraints which dictated the approach taken and resulted in unavoidable limitations. Given this, and the equivocal evidence observed herein, larger, more rigorous follow-ups are warranted.

Analogical Transfer

The results from this paper could be extended to determine the true extent at which analogical transfer of calibration training can occur. As shown by Glick & Holyoak(1983), one method to overcome recognition failure and improve transfer is to provide hints about applying the solution of the base problem to solve the analogue problem. In terms of this study, providing hints could simply entail telling participants to apply the training to the specific domain. The purpose of these hints is to remind the participants to use the knowledge and skills learnt from the training on the Post-Training questionnaires, in order to improve calibration. Directly reminding them to incorporate these skills when providing their ranges, would show how much of the training could be transferred in optimal conditions.

Individual Differences

An interesting approach would be to examine responses to a larger study of this type at an individual level – in order to determine whether the group-level improvements are driven by the majority of people improving a small amount or a smaller number of people showing a large improvement in calibration. Which of these better represents the true state of nature has implications for how to improve training processes. If the first, one might consider that better, or more intensive training is required to get participants closer to optimal calibration. If some participants are reaching optimal calibration with the current training, by comparison, the characteristics of or explanations provided by those participants might help improve current training to assist others in achieving similar benefits.

Initial Calibration

The results from this experiment suggested that having prior knowledge of calibration did not influence the participants calibration estimates at any point during the test (in line with previous research from Welsh et al, 2005). An extension to the research could thus conduct a second, Post-Training questionnaire at a later date to determine if the effectiveness of the feedback training remained over time for participants who either had or had not been provided continuing feedback aimed at maintaining better calibration. This could assist in determining how durable any benefits of training are and, thus, how often they need to be reinforced.

Conclusion

Participants in this experiment showed levels of miscalibration in the form of overconfidence (overprecision) consistent with previous literature. The Control group, who received no feedback on their performance, showed very similar levels of overconfidence across the three tests with around half of their (theoretically) 80% interval estimates containing the true value on each test – suggesting that they were appropriately matched for difficulty and that the participants degree of expertise within a specific domain did not alter their degree of calibration relative to the general knowledge domain. Additionally, no benefit was seen for participants who reported having prior experience or knowledge of calibration and overconfidence.

The feedback training provided to participants in the Experimental group proved effective, increasing the number of their ranges containing the true value from 42% to 60%. Whether this benefit transferred to the Domain-Specific Test, however, was less clear, with different analyses pointing in different directions. The Experimental group did not significantly outperform the control group on the Domain Specific Test (in fact, they performed very slightly but not significantly worse). This may, however, reflect their having started from a somewhat lower base – as their Domain Specific Test results *were* significantly better than their own Pre-Training questionnaire results.

Given this conflict, the strongest conclusion that can be drawn is that, while it seems that transfer may have occurred, it was less than complete and that future research is needed to more accurately determine the bounds on the efficiency of transfer of expertise in calibration across domains.

Acknowledgements

This study was conducted as part of the first three named authors' final year Petroleum Engineering Project at the Australian School of Petroleum. These authors contributed equally to the project and this paper and are named alphabetically. MBW is supported by ARC Linkage Grant LP160101460, which was awarded to SHB and includes support from Santos and Woodside. Please address correspondence regarding this article to MBW.

References

Adams, P. A., & Adams, J. K. (1958). Training in confidence-judgments. *The American Journal of Psychology*, *71*(4), 747-751.

Adams, J. K., & Adams, P. A. (1961). Realism of confidence judgments. *Psychological review*, *68*(1), 33.

Bornstein, B. H., & Zickafoose, D. J. (1999). " I know I know it, I know I saw it": The stability of the confidence–accuracy relationship across

domains. *Journal of experimental psychology: Applied*, *5*(1), 76.

Capen, E. C. (1976). The Difficulty of Assessing Uncertainty (includes associated papers 6422 and 6423 and 6424 and 6425). *Journal of Petroleum Technology*, *28*(08), 843-850.

Day, S. B., & Goldstone, R. L. (2012). The import of knowledge export: Connecting findings and theories of transfer of learning. *Educational Psychologist*, *47*(3), 153-176.

Fischhoff, B. (1981). *Debiasing* (No. PTR-1092-81-3). DECISION RESEARCH EUGENE OR.

Glick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive psychology*, *15*(1), 1-38.

Lewandowsky, S., & Kirsner, K. (2000). Knowledge partitioning: Context-dependent use of expertise. *Memory & cognition*, *28*(2), 295-305.

Lichtenstein, S., & Fischhoff, B. (1980). Training for calibration. *Organizational behavior and human performance*, *26*(2), 149-171.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1977). Calibration of probabilities: The state of the art. In *Decision making and change in human affairs* (pp. 275-324). Springer, Dordrecht.

McKenzie, C. R., Liersch, M. J., & Yaniv, I. (2008). Overconfidence in interval estimates: What does expertise buy you?. *Organizational Behavior and Human Decision Processes*, *107*(2), 179-191.

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological review*, *115*(2), 502.

Moore, D. A., Swift, S. A., Minster, A., Mellers, B., Ungar, L., Tetlock, P., ... & Tenney, E. R. (2016). Confidence calibration in a multiyear geopolitical forecasting competition. *Management Science*, *63*(11), 3552-3565.

Reeves, L., & Weisberg, R. W. (1994). The role of content and abstract information in analogical transfer. *Psychological bulletin*, *115*(3), 381.

Rumelhart, D. E. (1989). of human reasoning. *Similarity and analogical reasoning*, 298.

Soll, J. B., & Klayman, J. (2004). Overconfidence in interval estimates. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(2), 299.

Tetlock, P. E., & Gardner, D. (2016). *Superforecasting: The art and science of prediction*. Random House.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *science*, *185*(4157), 1124-1131.

Welsh, M. B., Begg, S. H., & Bratvold, R. B. (2007, January). Modelling the economic impact of common biases on oil and gas decisions. In *SPE Annual Technical Conference and Exhibition*. Society of Petroleum Engineers.

Welsh, M. B., Bratvold, R. B., & Begg, S. H. (2005, January). Cognitive biases in the petroleum industry: Impact and remediation. In *SPE Annual Technical Conference and Exhibition*. Society of Petroleum Engineers.

Efficient Data Compression Leads to Categorical Bias in Perception and Perceptual Memory

Christopher J. Bates (cjbates@ur.rochester.edu)

Robert A. Jacobs (rjacobs@ur.rochester.edu)
Department of Brain and Cognitive Sciences, University of Rochester
Rochester, NY

Abstract

Efficient data compression is essential for capacity-limited systems, such as biological memory. We hypothesize that the need for efficient data compression shapes biological perception and perceptual memory in many of the same ways that it shapes engineered systems. If true, then the tools that engineers use to analyze and design systems, namely rate-distortion theory (RDT), can profitably be used to understand perception and memory. To date, researchers have used deep neural networks to approximately implement RDT in high-dimensional spaces, but these implementations have been limited to tasks in which the sole goal is compression with respect to reconstruction error. Here, we introduce a new deep neural network architecture that approximately implements RDT in a task-general manner. An important property of our architecture is that it is trained "end-to-end", operating on raw perceptual input (e.g., pixels) rather than an intermediate level of abstraction, as is the case with most psychological models. We demonstrate that our framework can mimick categorical biases in perception and perceptual memory in several ways, and thus generates specific hypotheses that can be tested empirically in future work.

Keywords: Perception; memory; deep neural networks; rate-distortion theory; categorical bias

Introduction

Biological cognitive systems are not infinite. For instance, it is commonly hypothesized that people have finite attentional and memory resources, and that these constraints limit what people can process and remember. In this regard, biological systems resemble engineered systems which are also capacity-limited. For any capacity-limited system, biological or engineered, efficient data compression is paramount. After all, a capacity-limited system attempting to achieve its goals should maximize the amount of information that it processes and stores, and this can be accomplished through efficient data compression. Of course, this raises the question of what one means by "efficient".

In engineered systems, resources (e.g., bandwidth, finite memory) are limited, and thus system designers allocate these resources so as to maximize a system's performance, a process referred to as "bit allocation" (Gersho & Gray, 1992). Consider the design of digital compression algorithms. For example, file sizes can be reduced by a substantial factor using JPEG (image) or MP3 (audio) compression while still maintaining enough fidelity for most applications. When thinking about how to best perform bit-allocation, engineers must consider several questions. Which data items are frequent, and thus should be encoded with short digital codes, and which data items are infrequent, and thus can be assigned longer codes? Which aspects of data items are important to task performance, and thus should be encoded with high fidelity via long codes, and which aspects are less task relevant, and thus can be encoded with lower fidelity via short codes? For example, frequencies beyond the range of the human ear are less important when compressing audio waveforms with MP3, and can be stored with less fidelity. To address these questions, engineers have developed rate-distortion theory (RDT), a sophisticated mathematical formalism based on information theory (Cover & Thomas, 1991).

Our goal in this paper is two-fold. First, although exact methods already exist for RDT analysis in low-dimensional spaces, approximate methods are needed for high-dimensional spaces. To date, researchers have used deep neural networks to approximately implement RDT in high-dimensional spaces, but these implementations have been limited to tasks in which the sole goal is data compression with respect to reconstruction error (e.g. Ballé, Laparra, & Simoncelli, 2016). An innovation of the research presented here is that we introduce a new deep neural network architecture that approximately implements RDT in a task-general manner. That is, our architecture discovers good data compressions even when the data will be used for regression, classification, recognition, or other tasks. An important property of our model is that it is trained "end-to-end", operating on raw perceptual input (e.g., pixels) rather than intermediate levels of abstraction (e.g., orientation, texture, shape), as is the case with most psychological models. In this way, our framework represents an early step toward scaling up models of perception and perceptual memory toward levels of complexity faced in real-world situations.

Our second goal is to present one important and previously uninvestigated implication of efficient data compression which can be compared against empirical phenomena in perception and perceptual memory. While in this paper we present only a qualitative comparison, future work can focus on more rigorous, empirical evaluations of the hypotheses that our modeling framework generates. Specifically, we examine the phenomenon of categorical bias, which we explain in more detail below.

Principles of Efficient Data Compression and their Implications for Perception and Memory

This section examines important principles and implications of efficient data compression. We focus on one implication in particular, categorical bias, and draw a connection between

categorical bias in efficient compression and that found in perceptual memory.

All physically-realized systems are finite, and thus have finite limits on processing and storage capacities. For people, this implies that faulty perception and memory—what engineers refer to as "lossy compression"—is inevitable. If perception and memory cannot be perfect, can they at least be as good as possible given their capacity limits? This question has been explored in the context of low-level perception ("efficient coding"; see Barlow, 1961; Simoncelli & Olshausen, 2001), and researchers have found that low-level perceptual representations tend to be highly efficient with respect to the statistics of the environment.

Here, we focus on explaining higher-level sensory perception from the standpoint of efficient data compression. As we show in our results and analyses below, abstraction and categorization may be data-efficient strategies in many capacity-limited situations. There is strong empirical evidence that people employ these strategies in memory. For instance, research suggests visual working memory (VWM) avails of a wide array of summary statistics (e.g. Brady & Tenenbaum, 2013; Brady, Konkle, & Alvarez, 2009; Sims, 2016; Mathy & Feldman, 2012). In addition, various forms of abstract conceptual structures have been studied extensively in the context of long-term memory (LTM), such as schemas and scripts (Bartlett & Burt, 1933; Schank & Abelson, 1977).

A central assumption for our analysis below on categorical bias is that memory traces decay. Evidence for decay can be found in many experiments, including iconic visual memory and VWM (e.g. Sperling, 1960; Luck, 2008). We account for the decay of individual memory traces by hypothesizing that memory is biased toward representing recent information because recent information tends to be more task-relevant (Anderson, 1991). Consequently, memory engages in a form of adaptive bit-allocation in which fewer resources are devoted to older perceptual traces (suggesting that these traces are recoded in more compact and abstract ways over time) until so few resources are devoted to a trace that, effectively, the trace has fully decayed. This process frees up resources that can then be used to encode new information.

We propose that this reallocation happens both across and within memory subsystems. Within a subsystem (e.g. visual short-term memory), an individual trace tends to lose information over time to decay. Across systems, decay rates for individual traces vary. First, at stimulus offset, highly-detailed sensory information decays very rapidly. Next, sensory (e.g. iconic) memory representations are less detailed (more categorical) and decay more slowly. Short-term or working memory representations contain still less detail about the stimulus, are even more categorical and abstract, and decay more slowly than those of sensory memory. Finally, LTM contains the least amount of detail about the originally-observed stimulus, is the most categorical and abstract, and decays slowest.

For a well-designed system with limited storage, making decay rates proportional to information content is an efficient strategy—abstract representations (e.g those found in LTM) have low information content, and therefore can be retained "cheaply". As an analogy, imagine you are trying to make room on a full hard drive. It would be efficient to first remove large video files, before worrying about much smaller text files. Because highly abstract traces can be retained cheaply, LTM can accrue and store a large amount of traces over time. By contrast, working or sensory memory subsystems contain more detailed representations, and therefore cannot keep as many traces concurrently.

Consistent with our theory, experimental findings indicate that nearly all subsystems are influenced by a mix of perceptual and conceptual factors, but that the balance tilts more in favor of the conceptual the longer something is held in memory. Irwin (1991, 1992) demonstrated that iconic memory maintained more visual detail about an array of dots than VWM, whereas VWM representations seemed to be more abstract, coding information in a way that was robust to spatial translations. Brady and Alvarez (2011) found that observers' memories for the size of an object are systematically biased toward the mean of the object's category (see also Hemmer & Steyvers, 2009). Several experiments also indicate that memories for spatial location are biased toward spatial "prototypes" (Huttenlocher, Hedges, Corrigan, & Crawford, 2004; Huttenlocher, Hedges, & Duncan, 1991; Huttenlocher, Newcombe, & Sandberg, 1994). VWM representations not only encode "gist" or summary statistics (Oliva, 2005) over low-level visual features and textures, they also summarize high-level constructs such as the emotion of a face (Haberman & Whitney, 2007, 2009).

Visual LTM representations appear to be even more abstract. Konkle, Brady, Alvarez, and Oliva (2010) performed a visual LTM experiment in which subjects studied images of real-world objects drawn from different categories. Subjects studied between one and 16 exemplars per category, and later performed memory recognition test trials. It was found that as the number of exemplars from a category increased during study, memory performance decreased. Further analysis revealed that the conceptual distinctiveness of a category—low when category exemplars belong to the same subcategories and high when exemplars belong to different subcategories—is correlated with visual LTM performance but perceptual distinctiveness is not. The authors concluded that "observers' capacity to remember visual information in long-term memory depends more on conceptual structure than perceptual distinctiveness" (Konkle et al., 2010, p. 558).

To understand how abstraction results from efficient compression, it is important to understand the two central principles of RDT, which we name the "Prior Knowledge Principle" and the "Task-Dependency Principle". Now, we will briefly explain each principle and intuitively how each one can give rise independently to categorical representations.

Prior Knowledge Principle: Prior or domain knowledge is crucial to designing information-efficient systems. Accurate knowledge of stimulus statistics allows an agent to form

efficient representational codes given a limited capacity. To code a stimulus efficiently, a code must be designed using knowledge of the statistics of the to-be-coded items. Consider Morse code which is an algorithm for encoding letters of the alphabet as binary signals ("dots" and "dashes"). The designers of this code realized that they could increase its efficiency (i.e., decrease average code length) using knowledge of letter frequencies by assigning the shortest binary sequences to the most frequently transmitted letters. The more "peaky" the frequency of letters, the less information messages convey, and the shorter codes can be on average. For example, if 90% of the English language consisted of the letter 'e', then messages could be coded much more compactly on average than with real English in which e's are not nearly so frequent.

In many domains, the stimulus prior (i.e. distribution over stimuli) is highly peaked around several values. For example, if the set of stimuli consists of many photographs of various apples and bananas, this would constitute two different peaks (or modes) in the space of images around apples and bananas respectively. Efficient data compression predicts that these types of "modal" stimulus distributions will result in categorical bias. Specifically, as memory capacity is decreased (e.g. when decaying from short-term to LTM), representations should be attracted to one of the two modes, resulting in categorical bias.

Task-Dependency Principle: In addition to prior knowledge, for a code to be optimal, it must also take into account the current behavioral goals (or task) of an agent. Codes should allocate resources according to how an agent will use the encoded information. In particular, if it is costly to an agent to confuse stimulus values x and y, then codes should be designed so that these values are easily discriminated, even if this means a loss of precision for other discriminations.

As was the case with prior knowledge, efficient data compression predicts that certain behavioral goals will result in categorical bias. Namely, if effective behavior depends on making category distinctions, then when capacity is decreased, efficient codes should become more biased toward category prototypes, even when the stimulus prior is uniform. Thus, efficient data compression produces two distinct hypotheses for the existence of categorical bias. Either it results from modalities in the stimulus prior or from behavioral goals. These hypotheses may be evaluated in future work.

In the next section, we present the RDT formalism in order to make the prior knowledge and task-dependency principles mathematically precise. Then, we will demonstrate in simulation that each principle can indeed give rise to categorical bias.

Overview of Rate-Distortion Theory

Information theory addresses the problem of how to send a message over a noisy channel (e.g., a telephone wire) as quickly as possible without losing too much information. How much information can be sent per unit time (or per symbol) is the information 'rate' of a channel. Rate-distortion theory focuses on the case when the capacity (or rate) is too low to send the signal perfectly for a particular application (e.g., trying to hold a video conference with a slow internet connection). In this situation, one's goal is to design a channel that minimizes the average cost-weighted error (or distortion) in transmission, subject to the capacity limitation. Crucially, the optimization depends on two factors: (i) the prior distribution over inputs to the channel, and (ii) how the transmitted signal will be used after transmission. The first factor is important because common inputs should be transmitted with greater fidelity than uncommon inputs. The second factor is important because, depending on the application, some kinds of errors may be more costly than others.

Whereas much of the cognitive science literature uses the number of remembered "items" as a measure of memory capacity, information theory defines channel capacity as the mutual information between the input distribution and the output distribution. That is, if you know what comes out of a channel, how much information does that give you about what was inserted into the channel? If mutual information is high (high capacity), then the outputs tell you a lot about the inputs, but if it is low (low capacity), then the channel does not transmit as much information. The mutual information $I(x;y)$ for discrete random variables x and y is given by:

$$I(x;y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)\ p(y)}. \tag{1}$$

In the case of memory, sensory stimuli (e.g., pixel values) can be regarded as inputs to a channel, and neural codes are the channel's outputs (e.g., firing rates, changes to synaptic weights). The capacity of memory is the mutual information between the stimulus distribution and the neural code.

RDT seeks to find the conditional probability distribution of channel outputs (neural codes, denoted $\hat{x}$) given inputs (sensory stimuli, denoted x) that minimizes an error or distortion function $d(x,\hat{x})$ without exceeding an upper limit C on mutual information. For example, the distortion could be defined as the squared difference between the channel input and output, $(x-\hat{x})^2$. Mathematically, this minimization is the following constrained optimization problem:

$$\begin{aligned} Q^* = \underset{p(\hat{x}|x)}{\arg\min} & \sum_{x,\hat{x}} p(x)\ p(\hat{x}|x)\ d(x,\hat{x}) \\ & \text{subject to } I(x;\hat{x}) \leq C \end{aligned} \tag{2}$$

where Q^* is the optimal channel distribution.

Rate-Distortion Theory and Categorical Bias

Above, we described abstract or categorical representations as being an efficient strategy for compression, and pointed to evidence that human cognition makes use of this strategy. Furthermore, we noted that as the average information-content of memory traces decreases, the degree of categorical bias increases. We suggested that LTM might be viewed as using highly-compressed and categorical compressions, whereas perception uses less-compressed, less-categorical

compressions. For example, suppose you view an image of an apple. At short delays, you may remember that it was a red apple, at a longer delay, you may only remember that it was an apple, and perhaps at still longer delays, you may only remember that you saw a fruit. At long delays, categorical bias is large, because your memory for one apple is very similar to your memory for a different apple. Here, we demonstrate this phenomenon in simulation. We use a toy, one-dimensional domain in which it is possible to find the optimal lossy compression. In experiments below, we use approximate methods to extend this result to high-dimensional spaces, closer to the level of complexity that real brains must cope with.

As mentioned above, lossy compression can produce categorical bias when the stimulus prior is modal or when the loss function penalizes miscategorizations. Figure 1 demonstrates categorical bias effects in each case for unidimensional stimuli. The top panel (A) shows the case of a modal prior and squared error loss for d, while the bottom panel (B) shows the case of a uniform prior and categorical loss for d. According to the categorical loss, there is high cost to misremembering a stimulus that belongs to category A as one that belongs to category B, but low cost to misremembering a stimulus as another member of the same category. For example, consider plants that can be grouped as edible or poisonous. Misremembering a poisonous plant as an edible plant has high cost, whereas misremembering an edible plant as a different edible plant has low cost.

Figure 1A and B illustrate that channels optimized for a modal prior or a categorical loss, respectively, yield strong categorical bias at low capacity, but little at higher capacity. In the top rows of each (low capacity), $p(\hat{x}|x)$ is nearly identical for all values of $x = x_0$ within a category, but differs for two x_0 from different categories. In both A and B, categorical bias arises because values closer to the modes are "safer" when capacity is low and transmission errors are likely. On the other hand, at high capacity (bottom rows), $p(\hat{x}|x)$ is tightly peaked around the true input x_0 in both cases. In experiments below, for brevity we only elicit categorical bias via the distortion function (panel B).

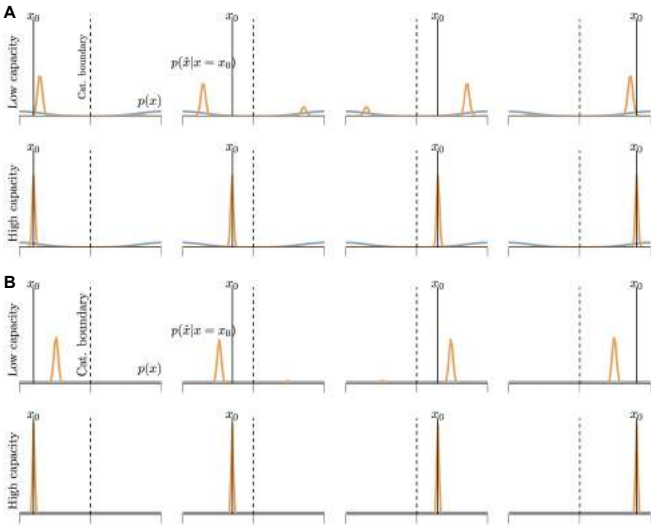


Figure 1: Illustration of how categorical bias can be explained via the prior (A) or the distortion function d (B). Horizontal axes plot stimulus space, vertical axes plot probability, dotted vertical line is the category boundary, solid vertical line marks the true stimulus value ($x = x_0$), and orange line plots output distribution $p(\hat{x}|x)$. Input distribution $p(x)$ is given by the blue line. Top and bottom rows in A and B show results for low and high capacity channels, respectively. In A, distortion function was squared error and $p(x)$ was bimodal. In B, distortion function was a weighted sum between a pure categorical loss and a square-error loss with weights of 1 and 0.001, respectively, and $p(x)$ was uniform.

RDT Neural Networks

Although RDT can be implemented exactly to find optimal compressions for problems using low-dimensional stimuli, it is too computationally expensive to be used with high-dimensional stimuli. Therefore, researchers have considered approximate implementations based on deep neural networks. To date, however, these implementations have been limited to tasks in which the sole goal is data (e.g., image) compression (e.g. Ballé et al., 2016). In this section, we introduce a new deep neural network architecture that approximately implements RDT in a task-general manner. In other words, our architecture discovers good data compressions even when the data will be used for regression, classification, recognition, or other tasks. Like previous RDT neural network implementations, our architecture is trained "end-to-end", meaning that it operates on raw sensory input (e.g., pixel values) rather than intermediate levels of abstraction (e.g., orientation, texture, shape), as is the case with most psychological models. The combination of end-to-end operation and task generality represents an important step toward scaling up models of perception and perceptual memory toward levels of complexity faced in real-world situations.

Rate-distortion (RD) Autoencoders: A key component of our models is the "autoencoder", parameterized models (e.g., neural networks) that map inputs to themselves subject to an information bottleneck. This bottleneck "forces" a model to find a more abstract, latent representation of the data. These abstract representations can then be used in subsequent tasks. Conventional neural network autoencoders consist of one or more 'encoder' layers, a middle 'latent' layer, and one or more 'decoder' layers. The latent layer typically has many fewer units than there are input dimensions, effectively reducing the dimensionality of the representation.

RD autoencoders differ from traditional autoencoders in that (i) they have a stochastic latent layer, and therefore a clear probabilistic interpretation, and (ii) a regularization term is added to the training objective function which acts to constrain how much information is represented in the latent units. If the coefficient on this term is high, then the network will seek a highly compressed latent representation. In our experiments, the latent unit activations are our models' "memory" of an input. Several variants of the rate-distortion autoencoder have been proposed, but here we choose the β-

Variational Autoencoder (β-VAE; Alemi et al., 2018).

Architecture: The models for all experiments presented here are defined by deep feedforward neural networks. Our general architecture (see Figure 2) consists of two modules: a β-VAE autoencoder and a decision module. The decision module takes as input the memory code (i.e., the activations of the latent units in the autoencoder) and optionally a task-related "probe" image, and outputs a decision variable. For example, in a change-detection task, the input to the autoencoder would be a target image, the input to the decision module would be a probe image and memory representation of the target, and the output of the decision module would be the probability that the probe is different than the target. Correspondingly, the training objective function has three terms, which can all be weighted differently to achieve different tradeoffs, corresponding to: (1) the distortion (or error) of the autoencoder's image reconstruction, (2) the information capacity of the memory representation, and (3) the decision error. Crucially, we can manipulate what kind of information is encoded in memory by varying how much reconstruction error is weighted relative to decision error during training, as well as how one kind of decision error is weighted relative to others (e.g., up-weighting errors along one stimulus dimension relative to other dimensions).

Implementation Details: Specific architectural choices for both experiments discussed below were standard within the neural network literature, and no specific fine-tuning was required to produce our results. In Experiment 1, we chose standard fully-connected layers with 'tanh' activation functions. The encoder and decoder both had two hidden layers, and the decision module had one. The latent layer and all hidden layers had 500 units. However, results were relatively insensitive to the choices of number of hidden units and layers, as long as the number of units was large. In Experiment 2, the encoder was composed of four 3×3 convolutional layers (32, 64, 64, and 64 filters for each layer, respectively), followed by a fully-connected layer with 1000 units. There were 1000 latent (memory) units. The decoder mirrored the encoder, except that convolutional layers were replaced with standard convolution-transpose layers. All hidden units used rectified-linear activations (ReLU). Again, a range of architectural choices can produce similar results. Finally, the decision module output was a single sigmoidal unit in Experiment 1, while in Experiment 2, the output was a softmax layer with one output unit for each of the three categories. All networks were trained with the "Adam" optimization algorithm.

Training sets: For Experiment 1, the dataset consists of images of an artificial plant-like object which we varied along two dimensions: leaf width and leaf angle. Images were converted to gray scale, down-sampled, and cropped to a size of 120×120 pixels. The space was discretized to 100 values along each dimension, for a total of 10,000 unique stimuli.

For Experiment 2, we used the Fruits-360 database[1]. We chose a subset of the classes to train on, specifically apples, tomatoes, and bananas. We augmented the dataset during training by randomly zooming and cropping inputs, as well as flipping the inputs horizontally at random. All images were resized to 112×112 pixels.

[1] https://github.com/Horea94/Fruit-Images-Dataset

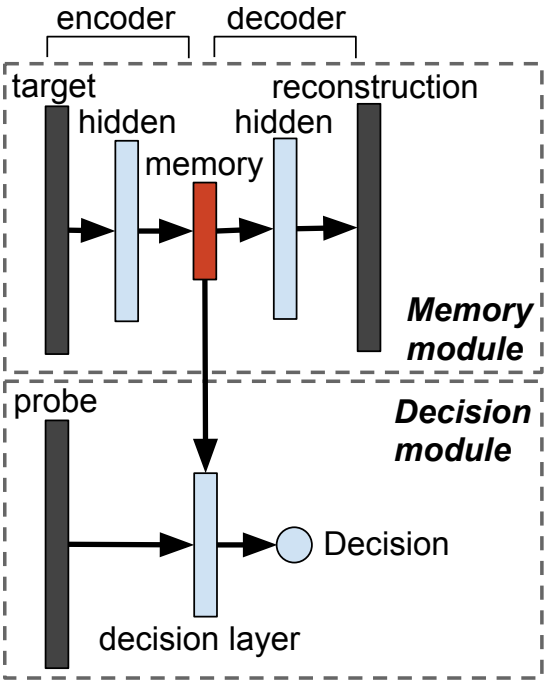


Figure 2: Schematic of the general model architecture. Dark gray boxes represent a vector of pixel values, while other boxes represent layers (or a set of layers) in the network. Layer that represents the memory code is in red.

Experiment 1: Artificial Images: Experiment 1 used the artificial plants dataset to demonstrate that the categorical bias effect depicted in Figure 1 extends to models operating in high-dimensional pixel-space. We show that, as expected, when a limited-capacity network is highly penalized for miscategorizing a stimulus, its memories exhibit categorical bias.

We trained the architecture on the full plants dataset. Following panel B in Figure 1, the training objective function was a mixture of pixel reconstruction error and categorical error, with a high relative coefficient on the latter. Specifically, the decision module was tasked with deciding whether the target image (input into the autoencoder) was the same category as a subsequent randomly-chosen probe image (input to the decision module). Given the high penalty for miscategorization, the optimal strategy for a model with very little capacity is to store little more than the category label. Figure 3 demonstrates this outcome by plotting target image reconstructions (outputs from the decoder) corresponding to a range of possible inputs. At low capacity (top panel), reconstructions of exemplars to the left of the category boundary are all nearly identical, and reconstructions of exemplars to the right of the boundary are also nearly identical. However, reconstructions on one side of the boundary are quite different from those on the other. In other words, there is a strong bias in the reconstructions to the appropriate category means, and thus a

sharp discontinuity at the category boundary. These results imply that at low capacity, the memory representation is a code that simply indicates which category the input belonged to. The best the autoecoder can do in this case is to produce the mean or prototype of that category. At higher capacities, the memory code contains more perceptual details beyond the category membership.

Experiment 2: Natural Images Experiment 2 used the Fruits-360 dataset to show that our approach scales to natural images. Again, we show that our models have increasing categorical bias as capacity decreases. However, our analyses in this experiment differ in a few ways. First, because natural image datasets do not contain a clear set of dimensions along which stimuli vary (like leaf width and leaf angle in Experiment 1), we indirectly measure the categorical bias in the trained models using autoencoder reconstructions and principle components analysis (PCA). An additional difference is that the decision module was trained to categorize each image, rather than to detect a change between target and probe.

Figure 4 (top panel) shows image reconstructions from the autoencoder at high, medium, and low capacity. These images demonstrate that the amount of detail that is retained in memory decreases as capacity decreases. At low capacity, the reconstructions are clearly categorical: each type of fruit corresponds to a unique output, which is the average of all images in that category. At medium capacity, different varieties within each species of fruit can begin to be distinguished. The figure's bottom panel demonstrates that the model's memory codes become more categorical at lower capacities. We performed PCA on memory vector activations and plotted stimuli in the space defined by the first two principle components. At medium or low capacity, memory codes for stimuli that belong to the same class are very similar, whereas at high capacity, memories of stimuli within a category are quite distinguishable from each other, and thus more perceptual details may be recovered[2].

Conclusion

We have argued, from both theoretical and empirical standpoints, that efficient data compression may be a central goal of perceptual and memory subsystems. In future work, we will discuss the extensive empirical evidence that efficient data compression is implemented in biological perception and memory, beyond the limited examples given here. In the current work, we highlighted one interesting piece of evidence that neural systems follow these principles, specifically that categorical representations are prevalent in memory. In simulation, we showed how categorical representations can be a natural outgrowth of efficient compression. These mechanisms for categorical bias generate hypotheses that can be tested in future empirical work. Because our modeling framework operates in an end-to-end and task-general manner, we believe that it shows promise for being scalable in ways that most psychological models are not.

[2]Note that even though the principle-components space appears to scale with capacity, this does not imply that the degree of categorical bias stays constant. For example, if the magnitude of noise that is added to the latent activations is fixed, more separation between two points in principle-components space implies that the decoder can more easily distinguish between them despite the noisiness. In fact, as network capacity is increased, the magnitude of noise added to the latents tends to *decrease* (because this allow more information to be stored), and thus two points that are a distance d apart in principle-components space are at least as distinguishable at high capacity compared to low capacity.

Acknowledgments

We thank Chris Sims for many useful discussions. The first author was supported by an NSF NRT graduate training grant (NRT-1449828) and an NSF Graduate Research Fellowship (DGE-1419118). This work was also supported by an NSF research grant (DRL-1561335).

References

Alemi, A. A., Poole, B., Fischer, I., Dillon, J. V., Saurous, R. A., & Murphy, K. (2018). Fixing a broken elbo. *arXiv preprint arXiv:1711.00464v3*.

Anderson, J. R. (1991). Is human cognition adaptive? *Behavioral and Brain Sciences*, *14*, 471–485.

Ballé, J., Laparra, V., & Simoncelli, E. P. (2016). End-to-end optimized image compression. *arXiv preprint arXiv:1611.01704*.

Barlow, H. B. (1961). Possible principles underlying the transformations of sensory messages.

Bartlett, F. C., & Burt, C. (1933). Remembering: A study in experimental and social psychology. *British Journal of Educational Psychology*, *3*(2), 187–192.

Brady, T. F., & Alvarez, G. A. (2011). Hierarchical encoding in visual working memory: Ensemble statistics bias memory for individual items. *Psychological Science*, *22*(3), 384–392.

Brady, T. F., Konkle, T., & Alvarez, G. A. (2009). Compression in visual working memory: Using statistical regularities to form more efficient memory representations. *Journal of Experimental Psychology: General*, *138*(4), 487.

Brady, T. F., & Tenenbaum, J. B. (2013). A probabilistic model of visual working memory: Incorporating higher order regularities into working memory capacity estimates. *Psychological review*, *120*(1), 85.

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory*. New York: Wiley.

Gersho, A., & Gray, R. M. (1992). *Vector quantization and signal compression*. Norwell, MA: Kluwer Academic Publishers.

Haberman, J., & Whitney, D. (2007). Rapid extraction of mean emotion and gender from sets of faces. *Current Biology*, *17*(17), R751–R753.

Haberman, J., & Whitney, D. (2009). Seeing the mean: ensemble coding for sets of faces. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(3), 718.

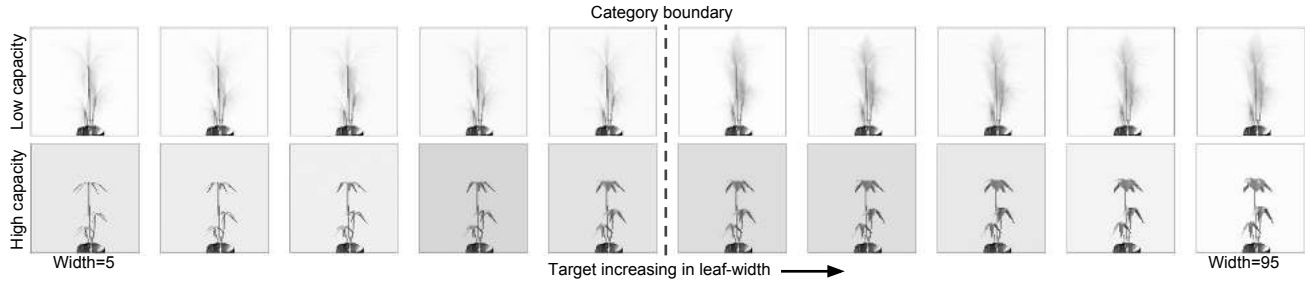


Figure 3: Visualizing categorical bias. At low capacity, all images on either side of the category boundary are highly similar to each other. Training-set stimuli included all combinations of leaf width and leaf angle. The category boundary divided skinny leaves from wide leaves, but was agnostic to leaf angle.

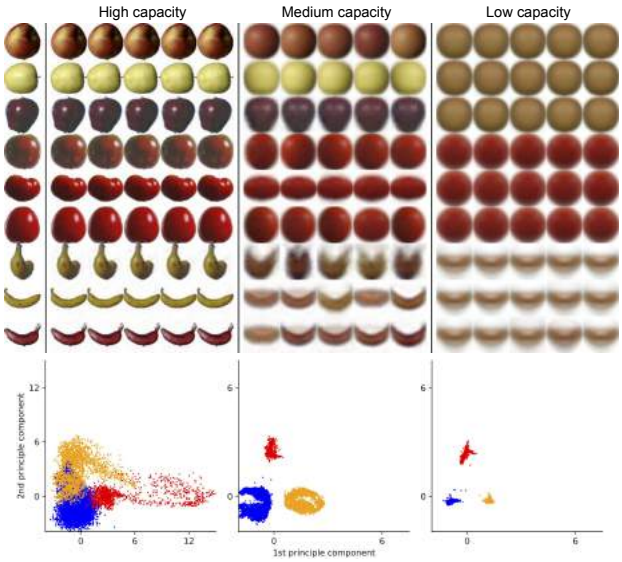


Figure 4: Top: Reconstructions from models trained on Fruits-360 dataset at three different capacities. Bottom: Corresponding PCA analysis of latent (memory) unit activations. At high capacity, latent activations have high variability within a class, whereas at lower capacities, same-class memories are highly similar.

Hemmer, P., & Steyvers, M. (2009). A bayesian account of reconstructive memory. *Topics in Cognitive Science*, *1*(1), 189–202.

Huttenlocher, J., Hedges, L. V., Corrigan, B., & Crawford, L. E. (2004). Spatial categories and the estimation of location. *Cognition*, *93*(2), 75–97.

Huttenlocher, J., Hedges, L. V., & Duncan, S. (1991). Categories and particulars: prototype effects in estimating spatial location. *Psychological review*, *98*(3), 352.

Huttenlocher, J., Newcombe, N., & Sandberg, E. H. (1994). The coding of spatial location in young children. *Cognitive psychology*, *27*(2), 115–147.

Irwin, D. E. (1991). Information integration across saccadic eye movements. *Cognitive psychology*, *23*(3), 420–456.

Irwin, D. E. (1992). Memory for position and identity across eye movements. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(2), 307.

Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010). Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. *Journal of Experimental Psychology: General*, *139*(3), 558.

Luck, S. J. (2008). Visual short-term memory. In *Visual memory* (pp. 43–85). New York: Oxford University Press.

Mathy, F., & Feldman, J. (2012). Whats magic about magic numbers? chunking and data compression in short-term memory. *Cognition*, *122*(3), 346–362.

Oliva, A. (2005). Gist of the scene. In *Neurobiology of attention* (pp. 251–256). Elsevier.

Schank, R. C., & Abelson, R. P. (1977). *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Lawrence Erlbaum Associates.

Simoncelli, E. P., & Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual review of neuroscience*, *24*(1), 1193–1216.

Sims, C. R. (2016). Rate–distortion theory and human perception. *Cognition*, *152*, 181–198.

Sperling, G. (1960). The information available in brief visual presentations. *Psychological Monographs: General and Applied*, *74*, 1–29.

Representing lexical ambiguity in prototype models of lexical semantics

Barend Beekhuizen
Department of Language Studies
University of Toronto, Mississauga
Depts. of Linguistics and Computer Science
University of Toronto
barend@cs.toronto.edu

Chen Xuan Cui
Department of Computer Science
University of Toronto
bobcui@cs.toronto.edu

Suzanne Stevenson
Department of Computer Science
University of Toronto
suzanne@cs.toronto.edu

Abstract

We show, contrary to some recent claims in the literature, that prototype distributional semantic models (DSMs) are capable of representing multiple senses of ambiguous words, including infrequent meanings. We propose that word2vec contains a natural, model-internal way of operationalizing the disambiguation process by leveraging the two sets of representations word2vec learns, instead of just one as most work on this model does. We evaluate our approach on artificial language simulations where other prototype DSMs have been shown to fail. We furthermore assess whether these results scale to the disambiguation of naturalistic corpus examples. We do so by replacing all instances of sampled pairs of words in a corpus with pseudo-homonym tokens, and testing whether models, after being trained on one half of the corpus, were able to disambiguate pseudo-homonyms on the basis of their linguistic contexts in the second half of the corpus. We observe that word2vec well surpasses the baseline of always guessing the most frequent meaning to be the right one. Moreover, it degrades gracefully: As words are more unbalanced, the baseline is higher, and it is harder to surpass it; nonetheless, Word2vec succeeds at surpassing the baseline, even for pseudo-homonyms whose most frequent meaning is much more frequent than the other.

Keywords: distributed semantic models; word meaning; ambiguity; prototype models; exemplar models; word2vec

Introduction

A central question for the cognitive science of language is how word meanings are represented in the minds of language users. Distributional semantic models (DSMs) represent word meanings as vectors in a high-dimensional space (Landauer & Dumais, 1997; Erk, 2012). The location of these points is based on the words in the neighbouring linguistic context (e.g., a window of words around the target word, or the document the word occurs in). DSMs have been successful in simulating diverse facets of human cognition, such as similarity judgments and analogy completion (e.g., McNamara, 2011; Pereira, Gershman, Ritter, & Botvinick, 2016).

Given that a vast majority of the words in English (and presumably most languages) are ambiguous (Klein & Murphy, 2001), the question arises whether a single vector, which functions as a 'prototype' of the word's meaning, can adequately represent the multiple meanings of an ambiguous word. Several researchers have argued that this is indeed the case. Schütze (1998), Burgess (2001), and Kintsch (2001) each show, using different models and set-ups, how aggregate representations of the context words can disambiguate ambiguous words. Arora, Li, Liang, Ma, and Risteski (2018) propose that word vectors are combinations of the vectors of the component meanings, and that these meaning vectors can be recovered from the 'compact' representation. Further circumstantial evidence for the adequacy of prototype representations comes from the fact that the DSMs successfully model various aspects of cognition even when representing a massively ambiguous vocabulary (Pereira et al., 2016).

Other work, however, suggests that single vector representations are inadequate for the representation of word meaning ambiguity. In the computational linguistics literature, this consideration has led to approaches in which multiple vector representations are learned for a word, each serving as the prototype of *one* of its senses (Reisinger & Mooney, 2010; Li & Jurafsky, 2015). In cognitive science, this assumption has led to the proposal of exemplar-based models, in which a word meaning is represented not as one or more prototype vectors, but as a weighted trace of the memorized contexts that a word occurred in. Jamieson, Avery, Johns, and Jones (2018), for instance, demonstrate that their exemplar-based model of word meaning representation succeeds where two widely-used DSMs (LSA; Landauer & Dumais, 1997 and BEAGLE; Jones & Mewhort, 2007) fail: While the prototype DSMs are able to represent the dominant (most frequent) meaning of a word, subordinate meanings are poorly captured by a single vector, suggesting that these models cannot reliably identify the intended meaning of an ambiguous word in context.

Given the general success of prototype DSMs, such a failure to simulate a key cognitive behaviour would indeed be worrisome if it applied to the entire class of approaches. However, Beekhuizen, Milić, Armstrong, and Stevenson (2018) show in a series of corpus experiments that not all prototype DSMs behave alike in representing ambiguous meanings. In this paper, we will argue that claims concerning the inadequacy of prototype DSMs are not justified. We will do so by showing that another prototype DSM, the CBOW algorithm of word2vec, has model-internal properties that enable it to disambiguate word meaning, and to succeed at accurately representing the infrequent meaning of ambiguous words. Crucially, we believe that this success in disambiguating infrequent meanings is driven by the fact that word meaning interpretation is distributed over two sets of representations in word2vec.

Our Approach

Word2vec (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013) is a word embedding model that learns a distributed semantic space that enables it to best predict words from their contexts. Figure 1 illustrates the process graphically, when using the continuous bag-of-words (CBOW) algorithm of word2vec. Each word in the vocabulary is represented as a row vector in the context matrix C and as a column vector in the target word matrix T. At every training step, the model is given a target word t as well as a window of k context words on either side of the target word. The model then learns to best predict the target word from the context words.

To determine its prediction, word2vec first takes the vector representations in C of the k context words and averages them, forming the aggregate context vector a. The context vector a is then compared with the current word representations in T to predict which word is most likely the target in that context. Intuitively, the more similar the context a is to the current vector representation of a word in T, the higher the predicted probability of observing that word. In training, after making a prediction for an example context, the model checks how far it is off from the desired probability distribution – that is, a probability of 1 for observing the given target word t and 0 for all other words – and proportionally updates the vectors in both C and T to minimize this error.

Although word2vec trains both a context matrix C and target matrix T, researchers typically just use one set of the trained representations (those of the context matrix C) as the resulting DSM of word meaning. Then, for disambiguating a word, a natural approach is to combine the vector representations (from that matrix) for the ambiguous word and its (presumably disambiguating) context words, and then to compare the resulting vector to other representations – for instance, synonyms of the two possible meaning of the ambiguous word – from the same matrix, under the assumption that the aggregate vector will be closest to the appropriate synonym (i.e., the one corresponding to the intended meaning of the ambiguous word). This approach has been explored in computational linguistics by Iacobacci, Pilehvar, and Navigli (2016).

In contrast, we propose a novel approach to using word2vec representations in modeling the disambiguation process, by drawing on its *training procedure* to derive the contextual interpretation of a word. Our insight is that both the context matrix C and the target matrix T contain learned knowledge that is important in disambiguation, just as they work together in the training process to form compatible representations of the context and target words (cf. Mitra, Nalisnick, Craswell, & Caruana, 2016). Rather than throwing away this important information and using representations from just one of the matrices, we use both the C and T matrices: We form an aggregate context vector a using C as a representation of the context of a word to be disambiguated, and compare that aggregate vector to representations of synonyms of its possible meanings embedded in the target word matrix T.

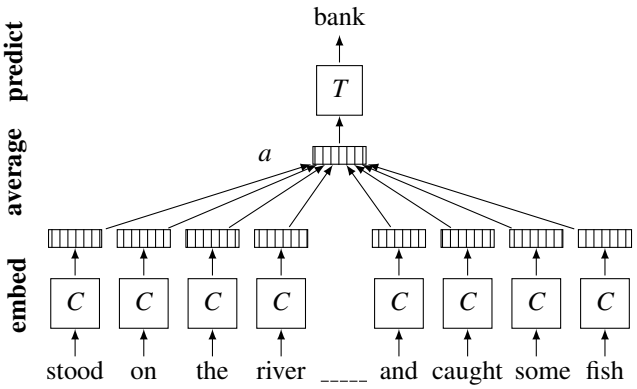


Figure 1: Word2vec model, using the CBOW algorithm

The use of a part of the training procedure is desirable, as it addresses an issue Jamieson et al. (2018) raise, namely that the prototype DSMs that have been shown to work use ad-hoc patches that are added to the models in order to represent word meanings in context. For word2vec, the aggregate context vector a is a representation of the context that is native to the model, as is the process of comparing a to representations in the T embedding space.

On a conceptual level, we believe that word2vec reflects an important property of word meaning interpretations, namely that they are not completely represented 'in' the word itself (cf. Elman, 2009). The word can be thought to provide a 'sketch' of the meaning (Levinson, 2000) that is completed through inferential processes by the linguistic and extralinguistic context in which it is embedded (e.g. Sperber & Wilson, 1986). This consideration is in fact one of the motivations of an exemplar-based approach. However, in word2vec too, ambiguous meanings are similarly not fully 'represented' in the word vectors of C or T. Rather, C and T, along with the algorithm that compares them, share the responsibility for predicting the target words from the context.

With regard to interpretation of infrequent meanings of a word, this approach gives word2vec an advantage. Given that word2vec's objective is to predict the target word, it suffices to optimize the representations in T so that the vector of the ambiguous target word represents just enough of the infrequent meaning to enable the appropriate context words to predict it (cf. the notion of 'good enough semantic processing' in Ferreira, Bailey, & Ferraro, 2002; Frisson, 2009). In the experiments below, we will illustrate how using the context and target matrix together allows word2vec to represent infrequent word meanings and identitfy them in context.

Artificial Language Simulations

As a first proof of concept, we replicate the artificial language simulation of Jamieson et al. (2018), which compared disambiguation in an exemplar-based model of word meaning to two prototype DSMs, and found the latter less successful.

An artificial corpus was generated in which the homophone sound form /breɪk/ (i.e., the sound of *break* or *brake*) was used in three contexts corresponding to three different meanings (to brake a car, to break the news, or to break a plate; henceforth all referred to as *break*). The models were tested to see whether they could identify each of the three meanings of *break* used in various disambiguating contexts (e.g., *man break car*, *woman break news*, *woman break plate*). Aside from sentences containing *break*, sentences with verbs that are synonymous to one of the three meanings were generated as well (e.g., *woman stop car*, *man report news*, *man smash glass*). These unambiguous verbs enabled evaluation of whether disambiguation models were able to identify the correct meaning of *break* in the context. Crucially, the corpus was generated either so that all meanings of *break* were equally frequent (balanced), or so that one meaning was 4 times as frequent as the other two meanings (unbalanced). For further details, see Jamieson et al. (2018).

We replicate this experiment for word2vec by generating the corpus in the same way as outlined above and training word2vec on it.[1] We then apply our approach using word2vec (described in the previous section) to see if it can correctly disambiguate the different meanings of *break*. To do so, we see whether the prediction of *break* in a sample context (e.g., *woman+car*) is as strong as the prediction of the appropriate unambiguous word (in this case, *stop*), and much stronger than the inappropriate unambiguous words (those corresponding to the other meanings of *break*). Importantly, the approach follows the flow of the learning procedure of word2vec: we average the representations of the context words in C to create an aggregate context vector a, and then compare a to the representation in T of each of the four different words (*break*, *stop*, *smash*, *report*) to determine the strength of prediction.

As Figure 2 shows, word2vec successfully predicts both *break* and its contextually appropriate synonym, both for the balanced corpus (where the three meanings of *break* are equally frequent) and the unbalanced corpus (where one meaning is more frequent than the others). Note that in all cases, the aggregate context vector is about as similar to the correct unambiguous verb as it is to *break*. For example, the model has learned that in the context of *woman* and *news*, both *report* and *break* are similarly predicted, and thus are similar to each other *in this context*.

Interestingly, we found that this behaviour is only present when both C and T are used; when aggregating context word vectors in C and then comparing them to the vectors of the unambiguous words in C again, the appropriate disambiguation behaviour was not achieved.[2] This means that word2vec is able to represent the contextually disambiguated meaning of a verb through the interaction of its context matrix C with its target matrix T. This behaviour can be expected, as the training algorithm of word2vec optimizes the similarity of the aggregate representations in C (i.e., the vector a) to that of the target word in T. That is: a and the vector of the target word in T are (by design) embedded in the same space, whereas an *aggregate* representation of the context words in T (as opposed to the individual words' representations in T) and the vector of the target word in T are not.

Our successful results contrast with those in Jamieson et al. (2018), who found that, while their exemplar-based word meaning model (Instance Theory of Semantics, henceforth ITS) performed well in this task, the two prototype DSMs – LSA and BEAGLE – were not as successful. In particular, in the balanced condition, all three models show the desirable disambiguation behaviour, but in the unbalanced condition, ITS can successfully disambiguate, but LSA and BEAGLE cannot. For these prototype models, only the most frequent meaning (the *stop* sense of *break*) is activated correctly, whereas the contexts of infrequent meanings (the *report* and *smash* senses) also activate (incorrectly) the most frequent meaning.

While our approach using word2vec demonstrates that a prototype DSM *can* successfully disambiguate infrequent meanings, a potential point of criticism is that our approach may work in an artificial setting like this, but not when the model is trained on a corpus with a realistic vocabulary size and many more unique contexts. After all, a realistic set-up necessitates a far greater degree of compression to allow for a maximally accurate prediction given only 200 dimensions to store all information in — and thereby a greater chance of having infrequent meanings being pushed out by the more frequent ones. Furthermore, the artificial language set-up tests the disambiguation on the data it was trained on, and so we are not directly addressing whether the model can carry out disambiguation in a generalizable way. These issues led to the design of the next experiment.

Disambiguation in a Naturalistic Setting

While the artificial language experiment provides a proof-of-concept of contextual disambiguation, it cannot test whether models have *generalizable* knowledge that scales to *naturalistic* contexts. The obstacle to larger-scale, more realistic scenarios is that testing disambiguation requires knowing the "correct" answer – that is, for any given instance of an ambiguous word in context, we need to know which meaning was intended in order to judge whether a model is performing appropriately. This requires a natural corpus that has the instances of homonyms annotated with the correct meaning in each case.

Since no such corpora of substantial size exist, we follow Arora et al. (2018) in adopting a method of using "pseudo-

[1] In all experiments reported, we used the implementation of word2vec in gensim (Řehůřek & Sojka, 2010), using CBOW with 200 vector dimensions, a window size of 5, a minimum frequency of 1, and otherwise default parameter settings. All software used is available as supplementary material at `https://tinyurl.com/w2vcogsci`.

[2] We also tried other ways to use word2vec, including its Skip-Gram variant, but CBOW with both C and T matrices was the most robust with unbalanced homonyms.

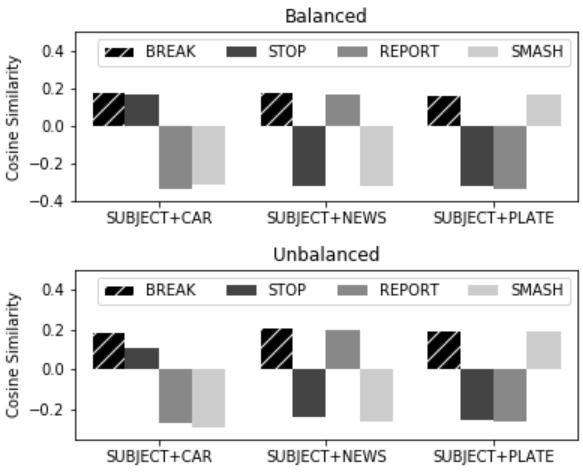


Figure 2: Cosine similarities between the word2vec representations of the context words (on x-axis) and the representations of the target words (in legend), for the balanced corpus and the unbalanced corpus.

homonyms" – pairs of words that are considered as if they were a single word. For example, if we consider the set of usages of *pizza* and *water* as if they were a single word with meanings PIZZA and WATER, then we would have a corpus in which all the instances in context of `pizza_water` are known to be disambiguated as either PIZZA or WATER (corresponding to the original word in that instance).

This set-up allows us to present our word2vec-based disambiguation approach with test cases (contexts containing a pseudo-homonym), and see whether it can identify which component meaning of the pseudo-homonym was intended in that context. We similarly evaluate the performance of ITS (Jamieson et al., 2018) on the same data, to see how our approach, based on a prototype DSM, compares to an exemplar-based approach to word meaning.

We use the TASA corpus of Landauer, Foltz, and Laham (1998), with the first half of the corpus as training data, and the second half as test data. Using a training-test split of the data, we made sure the models were actually tested on their capacity to disambiguate target words in novel, unseen contexts. We sampled 100 pairs of non-homonymous words that were similar to one of the real homonyms listed in Armstrong, Tokowicz, and Plaut (2012) in their length, frequency, and relative frequency of the two component meanings. This was done to make sure the pseudo-homonyms displayed similar relevant properties as real homonyms (Piantadosi, Tily, & Gibson, 2012).[3] We next explain how we can test each model under this approach.

Pseudo-homonym set-up for word2vec. For word2vec, we need to modify the corpus to enable training on a set of pseudo-homonyms, which were created by merging two non-homonymous words – e.g., replacing all instances of the words *pizza* and *water* with the single token `pizza_water`. The context and target matrices of word2vec were trained once on the original version of the training data, yielding C and T, and again on the version with pseudo-homonyms, yielding C' and T'. In this way, we have representations both for the pseudo-homonyms and for their component words individually. Then, for each instance of a pseudo-homonym in the test data, say `pizza_water`, we tested whether its aggregate context vector a from C' (based on the pseudo-homonym version of the corpus) was more similar to the correct or incorrect component meaning representation in T – *pizza* or *water* – whichever occurred in the original corpus).[4]

Pseudo-homonym set-up for ITS. ITS (Jamieson et al., 2018) follows the intuition that an accurate representation of word meaning is derived from all previously encountered instances of the word. Starting with words represented as high-dimensional random vectors, ITS represents the *memory trace* of each document in a corpus as the sum of the random vectors of all the words in that document. Word meanings in context are then derived from the matrix of memory traces by presenting the model with a *probe* in the form of a set of words, and retrieving its *echo*: an aggregate of all memory traces, weighted by how similar they are to the probe. Figure 3 presents a graphical representation of the echo retrieval process.

In our ITS set-up, we constructed a matrix of 20K-dimensional memory traces for the training portion of the original TASA corpus. Then, for each instance of either of the component words of a pseudo-homonym in the test data, a context probe was constructed out of the five words to the left and to the right of the word (excluding stopwords and punctuation), plus the two component words of the pseudo-homonym themselves. The echo of this aggregate probe was retrieved and compared to the echo of each component word individually. The component word whose echo had the highest cosine similarity to the echo of the aggregate context probe was selected as the disambiguated meaning.[5]

Results This approach gives us 91,703 ambiguous pseudo-homonym tokens in the test data, aggregated over the 3 simulations (on average 306 per pseudo-homonym). We find that word2vec scores an overall accuracy (proportion of correctly disambiguated test items) of .85 versus .69 for ITS. This means that overall, word2vec is better able to disambiguate words in their naturalistic contexts.

It is important to also consider how these accuracies compare to a chance baseline – is either model doing better than random guessing? Assuming there is some way to know which is the most frequent (dominant) meaning, a model that always guessed the dominant meaning would achieve a score

[3]Due to the random sampling, we ran three simulations, each with a new set of 100 pseudo-homonyms, and report aggregate findings of the three simulations.

[4]To compare vectors from C' to those from T, we use Orthogonal Procrustes, a standard method, to rotate T to T' so the vectors are all in a compatible vector space.

[5]This set-up was found to yield the best results for ITS compared to other set-ups we tried.

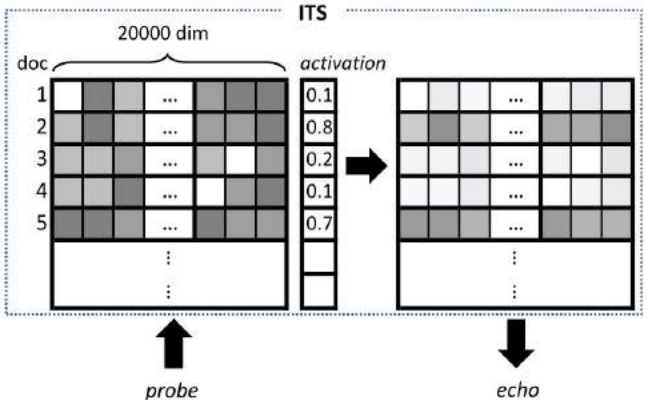


Figure 3: A visual example of the retrieval of an echo in ITS through the selective activation of memory traces when presented with a probe.

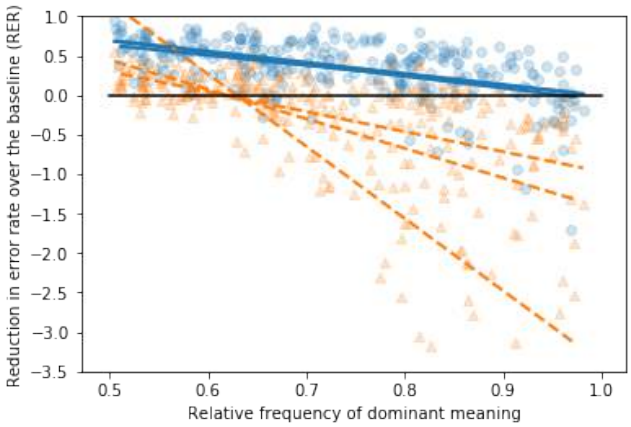


Figure 4: Reduction in error rate over the baseline (RER), aggregated over the three simulations. Dots (orange triangles for ITS, blue circles for word2vec) represent pseudo-homonyms. Regression lines are given for each simulation (orange dashed lines for ITS, blue solid lines (all overlapping) for word2vec). The black line represents zero error rate reduction; values below 0 are error rate increase, above 0 error rate reduction.

of .73 for our pseudo-homonyms – i.e., the average relative frequency of the dominant meaning. This seems like a reasonable baseline to assume, since we are interested in whether a model can learn the non-dominant meanings of ambiguous words. For each simulation, a two-tailed paired-samples t-test compared the accuracy per pseudohomonym for the model predictions and the dominant meaning baseline. In all simulations, word2vec did significantly better than the baseline (Sim. 1: $T = 9.71, p < 0.001$ / Sim. 2: $T = 11.96, p < 0.001$ / Sim. 3: $T = 10.01, p < 0.001$). ITS, however, performed significantly worse than the baseline (Sim. 1: $T = 3.32, p < 0.01$ / Sim. 2: $T = 2.34, p < 0.05$ / Sim. 3: $T = 2.74, p < 0.01$).[6]

Critical for our purposes is whether each model not just performed accurately for words with balanced meanings, but also was able to accurately disambiguate cases where one of the meanings is much more dominant. To assess this, we look at each pseudo-homonym individually. To compare fairly across pseudo-homonyms with different baselines (different degrees of dominance of meanings), we need a measure which looks at the amount by which each model surpasses (or falls short of) that baseline. A common measure to do so is the so-called reduction in error rate over the baseline (RER), defined as the amount by which the model improves over the baseline, divided by the error rate of the baseline.[7]

Figure 4 plots the RER for each pseudo-homonym as a function of its baseline (the relative frequency of its dominant meaning). The lines indicate the best linear fit between the two per simulation (all linear fits with Pearson's r are significant at $p < .001$). Both models display a downward slope across all simulations. This is unsurprising, since we would expect for any model that it is more difficult to disambiguate a very unbalanced homonym toward the infrequent meaning.

However, as can be gleaned from Figure 4, the slopes for word2vec are less negative than those of ITS, a difference that is significant across all three simulations (Sim. 1: $T = 3.03, p < .01$ / Sim. 2: $T = 4.83, p < .001$ / Sim. 3: $T = 2.90, p < .01$). This means that word2vec degrades more gracefully as homonyms become more unbalanced than ITS. Indeed, ITS only surpasses the baseline for relatively balanced items, and is unable to do better than the baseline for items whose most frequent meaning has a relative meaning frequency of around .66 or more. By contrast, the regression lines for word2vec only touch the null line (meaning always guessing the most frequent meaning) for the most unbalanced pseudo-homonyms (right end of the x-axis).

This means that, contrary to the predictions of Jamieson et al. (2018), and arguments raised in the computational linguistics literature (Reisinger & Mooney, 2010; Li & Jurafsky, 2015), not all prototype DSMs are unable to represent a contextually-resolved meaning of an unbalanced ambiguous word: word2vec performs adequately on such disambiguation tasks. Scaling up the disambiguation experiment to a more naturalistic corpus size and set of contexts, our approach using word2vec consistently surpasses the most-frequent sense baseline, and can thus be said to robustly resolve lexical ambiguities on the basis of the context words. Furthermore, word2vec degrades gracefully: it is harder to do better than chance for very unbalanced items than it is for balanced ones, but word2vec nonetheless on average surpasses the baseline even for very unbalanced pseudo-homonyms.

General Discussion

In this paper, we set out to show that, contrary to claims in the literature (Griffiths, Steyvers, & Tenenbaum, 2007; Reisinger & Mooney, 2010; Jamieson et al., 2018), proto-

[6] By virtue of transitivity, this also means that word2vec performs better than ITS (Sim. 1: $T = 12.13, p < 0.001$ / Sim. 2: $T = 11.82, p < 0.001$ / Sim. 3: $T = 11.23, p < 0.001$).

[7] That is, $RER = (model\_acc - baseline\_acc)/(1 - baseline\_acc)$

type distributed semantic models are capable of representing infrequent meanings of ambiguous words. We proposed that word2vec contains a natural, model-internal way of operationalizing the disambiguation process, and tested this approach successfully on the artificial language simulations for which Jamieson et al. (2018) showed that other prototype DSMs failed.

Importantly, we further assessed whether these results scaled to the disambiguation of naturalistic corpus examples. We generated a pseudo-homonym corpus by replacing all instances of sampled pairs of words in a corpus with pseudo-homonym tokens. We then trained word2vec on one half of the corpus, and assessed if the model was able to disambiguate pseudo-homonyms on the basis of their linguistic contexts in the second half of the corpus. We observed that our disambiguation approach using word2vec well surpasses the baseline of always guessing the most frequent meaning to be the right one, in contrast to an exemplar-based model (Jamieson et al., 2018). Word2vec moreover degrades gracefully: as words are more unbalanced (i.e., as the most frequent meaning has a higher relative frequency), the baseline is higher, and it is harder to surpass it. Word2vec nonetheless succeeds at surpassing the baseline, even for very unbalanced pseudo-homonyms.

A follow-up question is why Word2vec can represent infrequent meanings while LSA and BEAGLE cannot. It is tempting to speculate that this is due to the fact that word2vec vectors are trained to *predict* words, whereas LSA and BEAGLE vectors reflect the *counting* of words, and prediction-based DSMs have been found to outperform count-based DSMs (Baroni, Dinu, & Kruszewski, 2014). However, Levy and Goldberg (2014) argue the skipgram variant of word2vec performs implicit factorization of a count-based matrix in its objective function, so the actual differences between count-based and prediction-based models are not completely clear. This is an open area of research to which our findings contribute an important data point – i.e., that our approach to using the prediction mechanism of word2vec in semantic disambiguation outperforms a non-predictive approach using count-based DSMs (BEAGLE and LSA, as shown in Jamieson et al., 2018). A relevant future step is the comparison of our approach using the CBOW algorithm of word2vec to other prediction-based models or variants such as skipgram (Mikolov et al., 2013), as well as other contemporary approaches such as GloVe (Pennington, Socher, & Manning, 2014) and ELMo (Peters et al., 2018).

Another option is that it is the use of both the context word and target word matrices that allows us to achieve these results. Whereas off-the-shelf vectors have been used extensively in cognitive modeling experiments, our paper proposes to use a model-internal approach that leverages the fact that word2vec represents meaning as context word vectors *and* as target word vectors. This approach addresses the concern of Jamieson et al. (2018) that many prototype models only have ad hoc ways of carrying out the disambiguation procedure. It furthermore instantiates two critical points of the perspective on lexical semantics put forward by Elman (2009), namely: (1) that the drive to predict upcoming (linguistic) behaviour has sizable impact on the kinds of representations learned, and (2) that the interpretation of a word is always a function of some prior knowledge of the word as well as its context. It is effectively this idea that, combined with high-parametric representations and an abundance of data to train on, has led to the success of contemporary NLP word-meaning models such as ELMo (Peters et al., 2018).

We would like to argue that because of this distributed way in which word2vec learns to predict words, its representations reflect the important point that not all of a word meaning representation needs to be stored 'inside of' the word itself, but also by how word meanings relate to other word meanings (i.e., the 'oppositions' with other lexical items they have; Trubetzkoy, 1969 (1939)), as well as by rich pragmatic interpretive processes (Sperber & Wilson, 1986; Levinson, 2000). An important goal for the cognitive sciences of word meaning is to develop computationally precise models of how these processes work and interact. The present paper constitutes a stepping stone towards that goal.

Acknowledgments

SS and CXC are supported by an NSERC Discovery Grant RGPIN-2017-06506 to SS. We would like to thank the anonymous reviewers for their thoughtful comments.

References

Armstrong, B. C., Tokowicz, N., & Plaut, D. C. (2012). eDom: Norming software and relative meaning frequencies for 544 English homonyms. *Behavior Research Methods*, *44*(4), 1015–1027.

Arora, S., Li, Y., Liang, Y., Ma, T., & Risteski, A. (2018). Linear algebraic structure of word senses, with applications to polysemy. *TACL*, *6*, 483–495.

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings ACL*.

Beekhuizen, B., Milić, S., Armstrong, B., & Stevenson, S. (2018). What company do semantically ambiguous words keep? Insights from distributional word vectors. In *Proceedings CogSci.*

Burgess, C. (2001). Representing and resolving semantic ambiguity: A contribution from high-dimensional memory modeling. In D. S. Gorfein (Ed.), *On the consequences of meaning selection: Perspectives on resolving lexical ambiguity* (pp. 233–261). American Psychological Association.

Elman, J. L. (2009). On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive science*, *33*(4), 547–582.

Erk, K. (2012). Vector space models of word meaning and phrase meaning: A survey. *Language and Linguistics Compass*, *6*(10), 635-653.

Ferreira, F., Bailey, K. G., & Ferraro, V. (2002). Good-enough representations in language comprehension. *Current directions in psychological science*, *11*(1), 11–15.

Frisson, S. (2009). Semantic underspecification in language processing. *Language and Linguistics Compass*, *3*(1), 111–127.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*(2), 211.

Iacobacci, I., Pilehvar, M. T., & Navigli, R. (2016). Embeddings for word sense disambiguation: An evaluation study. In *Proceedings ACL*.

Jamieson, R. K., Avery, J. E., Johns, B. T., & Jones, M. N. (2018). An instance theory of semantic memory. *Computational Brain & Behavior*, *1*(2), 119–136.

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, *114*(1), 1–37.

Kintsch, W. (2001). Predication. *Cognitive science*, *25*(2), 173–202.

Klein, D. E., & Murphy, G. L. (2001). The representation of polysemous words. *Journal of Memory and Language*, *45*(2), 259–282.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, *104*(2), 211.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes*, *25*(2-3), 259–284.

Levinson, S. C. (2000). *Presumptive meanings: The theory of generalized conversational implicature*. MIT Press.

Levy, O., & Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. In *Proceedings NeurIPS*.

Li, J., & Jurafsky, D. (2015). Do multi-sense embeddings improve natural language understanding? In *Proceedings EMNLP*.

McNamara, D. S. (2011). Computational methods to extract meaning from text and advance theories of human cognition. *Topics in Cognitive Science*, *3*(1), 3–17.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings NeurIPS*.

Mitra, B., Nalisnick, E. T., Craswell, N., & Caruana, R. (2016). A dual embedding space model for document ranking. *CoRR*, *abs/1602.01137*.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings EMNLP*.

Pereira, F., Gershman, S., Ritter, S., & Botvinick, M. (2016). A comparative evaluation of off-the-shelf distributed semantic representations for modelling behavioural data. *Cognitive Neuropsychology*, *33*(3-4), 175–190.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings NAACL* (pp. 2227–2237).

Piantadosi, S. T., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, *122*(3), 280–291.

Řehůřek, R., & Sojka, P. (2010, May 22). Software Framework for Topic Modelling with Large Corpora. In *Proceedings LREC*.

Reisinger, J., & Mooney, R. J. (2010). Multi-prototype vector-space models of word meaning. In *Human Language Technologies: NAACL* (pp. 109–117).

Schütze, H. (1998). Automatic word sense discrimination. *Computational Linguistics*, *24*(1), 97–123.

Sperber, D., & Wilson, D. (1986). *Relevance: communication and cognition*. Harvard University Press.

Trubetzkoy, N. S. (1969 (1939)). *Principles of phonology*. University of California Press.

Are all Remote Associates Test equal?
An overview and comparison of the Remote Associates Test in different languages

Jan Philipp Behrens (jan.behrens@fu-berlin.de)
Cognitive Systems Group, Human-Centered Computation
Freie Universität Berlin, Germany

Ana-Maria Olteţeanu (ana-maria.olteteanu@fu-berlin.de)
Cognitive Systems Group, Human-Centered Computation
Freie Universität Berlin, Germany

Abstract

The Remote Associates Test (RAT, CRA) is a classical creativity test used to measure creativity as a function of associative ability. The RAT has been administered in different languages. Nonetheless, because of how embedded in the language the test is, only a few items are directly translatable, and most of the time the RAT is created anew in each language. This process of manual (and in two cases computational) creation of RAT items is guided by the researchers' understanding of the task. However, are the RAT items in different languages comparable? In this paper, different RAT stimuli datasets are analyzed qualitatively and quantitatively. Significant differences are observed between certain datasets in terms of solver performance. The potential sources of these differences are discussed, together with what this means for creativity psychometrics and computational vs. manual creation of stimuli.

Keywords: Remote Associates Test; RAT; CRA; Creativity; Creativity evaluation and metrics; Creativity Test

Introduction

The Remote Associates Test is a creativity test often used in the literature (Ansburg & Hill, 2003; Cunningham, MacGregor, Gibb, & Haar, 2009; Mednick & Mednick, 1971; Cai, Mednick, Harrison, Kanady, & Mednick, 2009; Ward, Thompson-Lake, Ely, & Kaminski, 2008).
A RAT problem given to a participant contains three words, for example FISH, MINE, RUSH; the participant has to come up with a fourth word related to all of the three given words. In this case, GOLD is an answer, because the compounds GOLD FISH, GOLD MINE, GOLD RUSH can be built with it. For a human or a machine (Olteţeanu & Falomir, 2015) to solve the RAT, knowledge about the compound words of a language is needed.

Because solving the RAT relies on knowing various expressions and compound words from a language, native speakers have an advantage and are generally the target population when deploying the RAT. This raises the need for various RAT stimuli sets in different languages.

As the RAT relies on knowledge and expressions which are language specific, the RAT is, in most part, not translatable between languages. An exception to this are the rare cases in which all compounds required as knowledge by a RAT item in a specific language also exist in another language - for example GOLDFISCH, GOLDMINE, GOLDRAUSCH as the German counterpart of the above mentioned query.

As only a few items are translatable, RAT sets of items are created anew by researchers in each language. This entails that RAT queries are probably impacted by the language itself, and quite likely by the preferences and knowledge of compound words of the stimuli dataset authors. The Remote Associates Test (RAT) is administered in many creativity studies, in the native language of the participants. Results reported in these studies are therefore impacted by the quality and difficulty of RAT items in each language. How can this impact be assessed?

No overview exists of the human performance in the RAT / CRA in the different languages. Such an overview would help us understand whether significant differences exist between performance on different RAT problem sets in the various languages in which it is employed. If no significant differences exist, this may indicate that results reported on creativity studies which use the RAT in different languages are, indeed, cross-comparable. If a significant difference however does exist, the comparability of the RAT across languages may require more nuance, and the development of an understanding of the sources of this difference.

This paper sets out to construct an overview of the RAT across eight languages and two types of the RAT (compound and functional), and provide an initial analysis between RAT sets across all these languages.

The RAT and languages

Sets of RAT / CRA problems of the following languages were analyzed - please note that some languages come with multiple datasets (D):

- German (Landmann et al., 2014)
- Chinese (Shen, Yuan, Liu, Yi, & Dou, 2016)
- Italian (Salvi, Costantini, Bricolo, Perugini, & Beeman, 2016)
- Romanian (Olteţeanu, Taranu, & Ionescu, n.d.)
- Polish (Sobków, Połeć, & Nosal, 2016)
- English D1 (Bowden & Jung-Beeman, 2003)
- English D2 (Olteţeanu, Schultheis, & Dyer, 2017)
- English D3 (Olteţeanu, Schöttner, & Schuberth, 2019)

– Finnish (Toivainen, Olteţeanu, Repeykova, Likhanov, & Kovas, 2019)

– Russian (Toivainen et al., 2019)

RAT comparison

A qualitative and quantitative comparison of the RAT datasets above is provided in the next sections.

Qualitative comparison

English datasets D2 and D3 contain different types of items: *compound* versus *functional*. For compound items, the relationship between the three given words and the answer word is a relationship manifested in language – for example, GOLD FISH, GOLD MINE and GOLD RUSH are compounds which all appear in language. By contrast, the relationship between functional query words and the answer reflects a functional relationship between the two, but may or may not be a compound linguistic relationship. For example, the relationship between CLOCKWISE and RIGHT or WRONG and RIGHT is a functional relationship. Of the above datasets, English D3 is functional.

Independent of the compound/functional classification, RAT problems have also been divided into two types based on the order of the words: homogeneous and heterogeneous items. RAT items are homogeneous if the solution word is either a prefix or a suffix to all the three words of the problem (like in the query FISH, MINE, RUSH, where GOLD acts as a prefix to each of the query items). Problems are heterogeneous, if the solution word is the prefix for some of the words and a suffix to other words of the problem (e.g. in the query RIVER, NOTE, ACCOUNT, the answer BANK is a suffix for the first word, and a prefix for the other two).

Of the above datasets, the German, Italian and English D1 ones distinguished between the heterogenous and homogeneous type of the queries. ANOVA with task type as a factor were run by the authors on these sets. The task type factor showed no significant effect on Accuracy (the number of queries solved by the participants). In the German version, a significant effect of the task type factor was observed on reaction times.

Finally, of the dataset items above, most are manually created. An exception to this are items from the English D2 and English D3 datasets. English D2 (Olteţeanu et al., 2017) successfully attempts the computational creation of RAT items, and compares results with an existing (English D1) normative dataset. English D3 (Olteţeanu et al., 2019) applies the computational approach using a new type of language knowledge to the creation of functional items, thus resurrecting an older idea of Worthen and Clark (Worthen & Clark, 1971) regarding the existence of such items, and their differences to compound items. These items are compared to compound items in the paper.

Quantitative comparison

In the following, a descriptive statistics overview of the different datasets is provided. To answer the question whether differences exist between RAT datasets in the various languages, Welch's unequal variances t-test is used on each two language pairs to determine the effect of language on the Remote Associates Test.

Descriptive data

The various RAT datasets contained varying numbers of items, between 17 (Polish) and 144 (English D1). Furthermore, the various items were deployed either (a) giving participants different timeframes to solve each query, between 2s and 60s, or (b) without setting a time limit. Since 2s, 5s and 7s timeframes were only used once across these datasets, only items between 15s and 60s are analysed in this paper. The stimuli were deployed on populations of various sizes, with n ranging between 26 and 317 participants. The Accuracy (number of correct answers given by the participants) fluctuated between .31 and .58. The response times ranged between $7.26s$ and $37.34s$. Please note that means and standard deviations were calculated for this paper from the given data, where they were not provided by the initial dataset. Table 4 gives an overview of all the datasets and various descriptive metrics across all languages.

Cronbach's alpha

Cronbach's alpha is the most commonly used method for estimating the reliability of a test, as reflected by its internal consistency between items. Scores below 0.5 indicate an unacceptable internal consistency, whereas higher scores indicate a better one. Generally scores above 0.7 are considered to reflect an acceptable amount of reliability, and an α above 0.9 is excellent. The Cronbach α scores were calculated by authors for some of the intial papers (see Table 4) and vary between .73 and .99.

Differences between languages

In order to measure differences between languages, heterogenous and homogeneous items were combined and Welch's unequal variances t-test was conducted to measure the difference between means on two existing performance metrics: Accuracy and Response Times.

Accuracy in 15s timeframe

As shown in Table 1, there were significant differences of means between:

- Italian ($M = .39$; $SD = .23$) and German ($M = .30$; $SD = .27$); $t(250) = 2.86$, $p = .0046$
- Italian and English D1 ($M = .31$; $SD = .22$); $t(253.88) = 2.95$, $p = .0035$

Table 1: Welch test results for accuracy in a 15s timeframe

accuracy	GER			ITA		
15s	t	df	p	t	df	p
ITA	2.86	249.99	.005**	-	-	-
ENG D1	0.13	260.92	.89	2.95	253.88	.004**

Accuracy in 30s timeframe

Like displayed in Table 5, there were significant differences of means between:

- Chinese ($M = .58$; $SD = .25$) and Polish ($M = .41$; $SD = .23$); $t(38.29) = 4.92$, $p < .0001$
- Chinese and German ($M = .30$; $SD = .27$); $t(254.28) = 5.92$, $p < .0001$
- Chinese and English D1 ($M = .31$; $SD = .22$); $t(265.86) = 3.47$, $p = .0006$
- English D1 and German; $t(262.27) = 2.72$, $p = .007$

Accuracy without timeframe

As reported in Table 6, there were significant differences of means between:

- English D2 ($M = .52$; $SD = .14$) and Finnish ($M = .46$; $SD = .11$); $t(93.95) = 2.1$, $p.038$
- English D3 ($M = .33$; $SD = .16$) and Romanian ($M = .54$; $SD = .43$); $t(83.26) = 3.98$, $p = .0002$
- English D3 and Russian ($M = .55$; $SD = .14$); $t(92.87) = 3.73$, $p = .0003$
- English D3 and English D2; $t(93.46) = 3.83$, $p = .0002$

RT in 15s timeframe

As presented in Table 2, there was a significant difference of means between:

- English D1 ($M = 7.26$; $SD = 1.65$) and Italian ($M = 6.52$; $SD = 1.46$); $t(258.86) = 3.87$, $p = .0001$.

RT in 30s timeframe

As shown in Table 3, there were significant differences of means between:

- English D1 ($M = 10.45$; $SD = 3.47$) and Polish ($M = 14.03$; $SD = 3.06$); $t(21.38) = 4.48$, $p = .0002$
- Chinese ($M = 9.74$; $SD = 3.13$) and Polish; $t(20.7) = 5.42$, $p < .0001$

RT without timeframe

As stated in Table 7, there were significant differences of means between:

- Finnish ($M = 37.34$; $SD = 17.36$) and Romanian ($M = 15.37$; $SD = 10.53$); $t(52.72) = 6.67$, $p < .0001$
- Finnish and Russian ($M = 23.53$; $SD = 10.38$); $t(58.18) = 5.05$, $p < .0001$
- Finnish and English D2 ($M = 14.52$; $SD = 9.89$); $t(76.07) = 4.79$, $p < .0001$
- Finnish and English D3 ($M = 11.68$; $SD = 10.96$); $t(67.26) = 6.48$, $p < .0001$
- Russian and English D3; $t(83.71) = 2.99$, $p = .004$
- Russian and Romanian; $t(91.38) = 3.37$, $p = .001$
- English D2 and English D3; $t(91.92) = 2.09$, $p = .04$

Discussion and further work

The hardest sets to solve seem to be the English D3 set of items from Study 2, with a an average accuracy of .30, and the Finnish dataset in terms of response times, with a mean 37.34 seconds. The response times of the Russian RAT were also noticeably higher (23.53s).

This paper set out to compare the RAT in different languages, and across different datasets. Significant differences were observed between multiple languages and datasets, on both the Accuracy and Response Times performance metrics.

The significant difference observed between the English D2 and English D3 sets may have as a source the difference between types of items (compound versus functional).

In the cases in which a significant difference exists between different language datasets, various causes could act as the source:

(a) different population samples are more creative (or at least better at the associative factor in creativity);

(b) the RAT is more difficult in some languages, because of the language itself and the cognitive factors resulting from encoding linguistic knowledge and solving the RAT in that language and/or

(c) sets of RAT queries vary in difficulty, because they are created without using standardized methods, thus depend on the inspiration and knowledge base of the researchers creating them.

This initial investigation shows that differences between the RAT in various languages need to be addressed in more detail. Before cross-comparison of creativity results can be declared, the source of these differences needs to be found. Experimental or analytical setups need to be designed in order to establish which one of (a), (b) and (c), or combination thereof, is the source of the differences.

An initial thought on establishing comparability could be to attempt to find translatable items across the various languages. By keeping stimuli items constant, differences of creativity pertaining to the population or use of language could be established.

However, even if translatable, the same RAT items may not be the same difficulty in different languages. Some light on this is shed by computational models like comRAT-C (Olteţeanu & Falomir, 2015), essentially models of memory search, which can solve the RAT by organizing their knowledge in a semantic net-like structure and propagating activation through word associations. The comRAT-C's probability

of solving a query correlates with human performance. This model entails that, even if different RAT queries can be translated in different languages, equivalence does not necessarily exist between them: the number of word associates and the strength of association may not be the same in different languages. Different tools may thus need to be used to try to establish query equivalence.

A potential solution may be to establish a stronger item equivalence in computational terms: for example by using computational RAT query generators like comRAT-G (Olteţeanu et al., 2017), to create sets of items where a high degree of control can be maintained over the number of associates and the association strength of the query words. Such approaches have already proven fruitful in the deployment of more precise empirical designs (Olteţeanu & Schultheis, 2017), and in the creation of other types of items (Olteţeanu et al., 2019).

Another direction of future work would be to establish a creative association measure which transcends the constraints of language like a visual Remote Associates Test - some work in this direction has already been done by (Olteţeanu, Gautam, & Falomir, 2015; Toivainen et al., 2019).

This paper gives an overview of RAT datasets in multiple languages, and shows that cross-linguistic comparability should not be taken for granted in the case of this broadly used creativity test.

Acknowledgements

The support of the Deutsche Forschungsgemeinschaft (DFG) for the project CreaCogs via grant OL 518/1-1 is gratefully acknowledged.

References

Ansburg, P. I., & Hill, K. (2003). Creative and analytic thinkers differ in their use of attentional resources. *Personality and Individual Differences*, *34*(7), 1141 - 1152.

Bowden, E. M., & Jung-Beeman, M. (2003). Normative data for 144 compound remote associate problems. *Behavior Research Methods*, *35(4)*, 634–639.

Cai, D. J., Mednick, S. A., Harrison, E. M., Kanady, J. C., & Mednick, S. C. (2009). Rem, not incubation, improves creativity by priming associative networks. , *106*(25), 10130–10134.

Cunningham, J. B., MacGregor, J., Gibb, J., & Haar, J. (2009). Categories of insight and their correlates: An exploration of relationships among classic-type insight problems, rebus puzzles, remote associates and esoteric analogies. *The Journal of Creative Behavior*, *43*(4), 262-280.

Landmann, N., Kuhn, M., Piosczyk, H., Feige, B., Riemann, D., & Nissen, C. (2014). Entwicklung von 130 deutsch sprachigen Compound Remote Associate (CRA)-Wortraetseln zur Untersuchung kreativer Prozesse im deutschen Sprachraum. *Psychologische Rundschau*, *65*, 200–211.

Mednick, S. A., & Mednick, M. (1971). Remote associates test: Examiner's manual.

Olteţeanu, A.-M., & Falomir, Z. (2015). comrat-c : A computational compound remote associate test solver based on language data and its comparison to human performance. *Pattern Recognition Letters*, *67*, 81-90.

Olteţeanu, A.-M., Gautam, B., & Falomir, Z. (2015). Towards a visual remote associates test and its computational solver..

Olteţeanu, A.-M., Schöttner, M., & Schuberth, S. (2019). Computationally resurrecting the functional remote associates test using cognitive word associates and principles from a computational solver. *Knowledge-Based Systems*.

Olteţeanu, A.-M., & Schultheis, H. (2017). What determines creative association? Revealing two factors which separately influence the creative process when solving the remote associates test. *The Journal of creative behavior*. doi: 10.1002/jocb.177

Olteţeanu, A.-M., Schultheis, H., & Dyer, J. B. (2017). Computationally constructing a repository of compound Remote Associates Test items in American English with comRAT-G. *Behavior Research Methods*, *7*.

Olteţeanu, A.-M., Taranu, M., & Ionescu, T. (n.d.). Normative data for 111 compound Remote Associates Test problems in Romanian. *Frontiers*.

Salvi, C., Costantini, G., Bricolo, E., Perugini, M., & Beeman, M. (2016). Validation of Italian rebus puzzles and compound remote associate problems. *Behavior Research Methods*, *48*, 664–685.

Shen, W., Yuan, Y., Liu, C., Yi, B., & Dou, K. (2016). The development and validity of a chinese version of the compound remote associates test. *American Journal of Psychology*, *129*, 245–258.

Sobków, A., Połeć, A., & Nosal, C. (2016). Rat-pl – construction and validation of polish version of remote associates test. *Studia Psychologiczne*, *54*, 1–13.

Toivainen, T., Olteţeanu, A.-M., Repeykova, V., Likhanov, M., & Kovas, Y. (2019). Visual and linguistic stimuli in the remote associates test: A cross-cultural investigation. *Frontiers in Psychology*, *10*.

Ward, J., Thompson-Lake, D., Ely, R., & Kaminski, F. (2008). Synaesthesia, creativity and art: What is the link? *British Journal of Psychology*, *99*(1), 127-141.

Worthen, B. R., & Clark, P. M. (1971). Toward an improved measure of remote associational ability. *Journal of Educational Measurement*, *8*(2), 113–123.

Appendix

Table 2: Welch test results for RT in a 15s timeframe

RT	ITA		
15s	t	df	p
ENG D1	3.87	258.86	.0001***

Table 3: Welch test results for RT in a 30s timeframe

RT	ENG D1			CHI D1		
30s	t	df	p	t	df	p
CHI D1	1.77	265.91	.08	-	-	-
POL	4.48	21.38	.0002***	5.42	20.7	2e-5****

Table 4: Number of elements($|x|$), sample size(n), mean($\bar{x}$) and standard deviation(s) of accuracy and response time and Cronbach's α for the RAT in the different languages. S1 and S2 reflect different studies using the same set of stimuli.

	Timeframe			Accuracy $\bar{x}$ (s)		RT $\bar{x}$ (s)	Cronbach's α	
Language	in sec	$\|x\|$	n	sum	per item	per item [sec]	Accu	RT
German both	60	130	80	54.99 (34.97)	.44 (.27)	16.97 (7.12)	-	-
heterogeneous	60	56	80	26.10 (15.79)	.47 (.28)	15.80 (6.70)	-	-
homogeneous	60	74	80	30.19 (19.17)	.41 (.26)	18.50 (7.50)	-	-
German both	30	130	80	-	.39 (.27)	-	-	-
German both	15	130	80	-	.30 (.27)	-	-	-
Chinese	30	128	123	74.46	.58 (.25)	9.74 (3.13)	.92	
Italian both	15	122	317	47.58 (28.06)	.39 (.23)	6.52 (1.46)	-	-
heterogeneous	15	66	317	25.48 (14.72)	.39 (.22)	-	-	-
homogeneous	15	56	317	22.12 (13.44)	.40 (.24)	-	-	-
Romanian	none	111	63	59.94 (47.73)	.54 (.43)	15.37 (10.53)	.93	.97
Polish	30	17	206	6.90 (3.90)	.41 (.23)	14.02 (3.06)	.79	-
English D1 both	30	144	289	72.72	.51 (.25)	10.45 (3.47)	-	-
heterogeneous	30	59	289	29.74	.50	-	-	-
homogeneous	30	85	289	42.93	.51	-	-	-
English D1 both	15	144	289	-	.31 (.22)	7.26 (1.65)	-	-
English D2 both	none	100	113	52.64 (16.16)	.53 (.16)	-	.94	.99
comRAT-G	none	50	113	26.20 (7.03)	.52 (.14)	14.52 (9.89)	.85	.99
Bowden, J.-B.	none	50	113	26.41 (11.24)	.53 (.23)	16.56 (12.84)	.93	.99
English D3 S1 fRAT	none	75	26	35.27 (7.99)	.47 (.11)	13.91 (8.42)	-	-
comRAT	none	50	26	25.02 (7.26)	.50 (.15)	12.38 (6.23)	-	-
English D3 S2 fRAT	none	48	61	17.10 (5.77)	.36 (.12)	14.14 (13.39)	.79	.90
Compound both	none	48	61	15.85 (7.60)	.33 (.16)	11.68 (10.96)	.87	.96
comRAT-G	none	24	61	7.25 (3.72)	.30 (.16)	11.00 (10.62)	.75	.93
Bowden, J.-B.	none	24	61	8.61 (5.06)	.36 (.21)	11.64 (0.65)	.85	.92
Finnish	none	47	67	21.60 (5.30)	.46 (.11)	37.34 (17.36)	.73	-
Russian	none	48	67	26.60 (6.90)	.55 (.14)	23.53 (10.38)	.83	-

Table 5: Welch test results for accuracy in a 30s timeframe

accuracy	GER			CHI D1			POL		
30s	t	df	p	t	df	p	t	df	p
CHI D1	5.92	254.28	1e-8****	-	-	-	-	-	-
POL	0.39	43.32	.7	4.92	38.29	2e-5****	-	-	-
ENG D1	2.72	262.27	.007**	3.47	265.86	.0006***	2.03	36.62	.05

Table 6: Welch test results for accuracy without timeframe

accuracy	ROM			FIN			RUS			ENG D2		
no tf	t	df	p	t	df	p	t	df	p	t	df	p
FIN	1.66	78.25	.10	-	-	-	-	-	-	-	-	-
RUS	0.32	91.93	.75	1.71	90.40	.09	-	-	-	-	-	-
ENG D2	1.00	74.86	.32	2.10	93.95	.038*	0.66	89.52	.51	-	-	-
ENG D3	3.98	83.26	.0002***	1.82	92.66	.072	3.73	92.87	.0003***	3.83	93.46	.0002***

Table 7: Welch test results for response time without a timeframe

RT	ROM			FIN			RUS			ENG D2		
no tf	t	df	p	t	df	p	t	df	p	t	df	p
FIN	6.67	52.72	2e-8****	-	-	-	-	-	-	-	-	-
RUS	3.37	91.38	.001**	5.05	58.18	5e-6****	-	-	-	-	-	-
ENG D2	1.96	66.05	.054	4.79	76.07	8e-6****	0.30	80.50	.76	-	-	-
ENG D3	0.64	68.42	.52	6.48	67.26	1e-8****	2.99	83.71	.004**	2.09	91.92	.04*

Investigating the Use of Word Embeddings to Estimate Cognitive Interest in Stories

Morteza Behrooz
morteza@ucsc.edu
University of California Santa Cruz
1156 High St. Santa Cruz, California 95064

Justus Robertson
jjrobert@ncsu.edu
North Carolina State University
Raleigh, NC 27695

Arnav Jhala
ahjhala@ncsu.edu
North Carolina State University
Raleigh, NC 27695

Abstract

Narrative processing is an important skill to model both from a cognitive science perspective and a computational modeling perspective which applies to intelligent agents. Communication between humans often involves storytelling patterns that make the mundane exchange of information more interesting and with proper emphasis on important communicative goals. Current narrative generation models evaluate their generations based on either a priori domain semantics (e.g. game state for an in-game conversation with player agents) or generic text quality measures (e.g. coherence). However, in utilizing storytelling as a communicative tool for real-world interactions, domain-specific approaches fail to generalize and text quality measures fail to ensure that the narrative is perceived as *interesting*. Hence, such generation needs to consider the cognitive processes involved in the perception of narrative. Using theories of cognitive interest, we present results of an investigation of whether word embeddings (e.g. GloVe (Pennington, Socher, & Manning, 2014)) could be used to model and estimate cognitive interestingness in stories.

Introduction and Background

In computational narrative generation, the communication context for which the narratives are generated plays an integral role in determining both the method constraints during the generation and the evaluation metrics for the resulting narratives. Not all approaches to narrative generation are compatible with all narrative communication paradigms, because they result in vastly different qualities in the generated narratives and also differ in their assumptions and constraints.

Moreover, no single set of evaluation or optimization metrics can ensure the success of a narrative generator across multiple paradigms. Such "success" is usually dependent upon being received positively by the audience and achieving any potential communicative or social goals. In simpler terms, a "good" narrative has to be interesting to the audience.

Entertainment, and games in particular, have been a prominent context for narrative generation and communication. Many games change the events that are not (at least directly) in control of the player, or affect what the players say (in voice or text), in order to create the "best" storyline possible with a goal of maximal immersion and character believability (Mateas & Stern, 2003; McCoy, Treanor, Samuel, Mateas, & Wardrip-Fruin, 2011; Ryan, Mateas, & Wardrip-Fruin, 2016). Other games can involve an interactive settings, where the player can influence the progression of the story through making choices (Riedl & Bulitko, 2012).

In such game-related use cases, it is often possible to infer the quality or interestingness of the generated story using known domain semantics. For instance, if a simple generator is making a story about chess, it is easy to know which sequence of events or moves are worthy of being recited as a story, since we know the significance of every move, or sequence of moves, to the game progression or to the winning chances of each side. Similar inferences about event sequences can be made about more complex games as well, given that some game semantics are available. Moreover, even when games are not involved, many story modeling and narrative generation approaches rely on a semantic model of a particular domain (e.g. characters, goals, entity relationships, etc.) which allows the derivation of a sequence of events and ultimately a narrative, such as in (Elson, 2012a). The same is true about classic story generation systems that while inspiring, rely on a bank of previous stories and their assumed structures to generate new ones with a measure of interestingness or success, such as Minstrel (Turner, 1994) and Mexica (PÉrez & Sharples, 2001).

Other narrative generation approaches are less dependent on a particular context of communication and use case, and consequently, do not depend on a priori semantic models. Instead, they attempt to generate narrative of stories that make general sense (as a sequence of events) and contain correct sentences (if presented in text). Thus, in order to assess the quality of the generated story, such approaches often focus on the general properties and qualities of the generated text, such as coherence or the causal plausibility of the sentence ordering (Papineni, Roukos, Ward, & Zhu, 2002). This way of generating narrative is sometimes referred to as *open story generation* (Martin et al., 2017; Swanson & Gordon, 2008).

Improving on generic text-based evaluation metrics, in (Purdy, Wang, He, & Riedl, 2018), a set of proxy measures are introduced to assess the "story quality" in an open story generation task. These measures are shown in (Purdy et al., 2018) to correlate with human judgment of story quality; hence, they can be used towards a better evaluation of the generated narrative and an easier and faster fine-tuning of many generative models, such as Recurrent Neural Networks (RNNs). They include:

- Correct grammar use ("grammaticality"),
- Complexity of used language ("narrative productivity"),
- Similarity of adjacent sentences ("local contextuality"),
- Level of adherence to the usual ordering of events in most

stories, e.g. "eat" comes after "order" ("temporal ordering").

Humans possess an intuitive evaluation metric for stories, one that goes beyond linguistic measures. Expert human storytellers are not considered experts merely because of the quality of their use of language (however sophisticated it may be), but also because of their ability to tell stories that seem interesting to a large number of audience. Such experts master narrative authorship techniques and can recognize the processes involved in human's cognitive perception of narrative. In other words, they tell stories in ways that are informed by their understanding of how human perception of narrative works.

To that end, proxy measures introduced above are a useful start to assessing narrative quality when it is not tied to a specific domain of semantics. However, an important aspect of story quality, i.e. "how good a story is", depends on more complex evaluations metrics than language use, local contextuality, or the normality of the event orderings. While those measures are relevant, they do not inform the generation process about the perception of narrative. Ideally, a generator should also optimize for its generated narratives to be perceived as interesting. Moreover, as mentioned above, a computational generation of narrative heavily depends on the communication context in which it operates. A particular reason why a focus on narrative perception is imperative is the rapid evolution of such contexts, which will increasingly include interactive and sociable agents (e.g. embodied or virtual agents (Goodrich, Schultz, et al., 2008; Fong, Nourbakhsh, & Dautenhahn, 2003) or conversational agents (NPR, 2017)).

Story Interestingness

Storytelling, as an intuitive, natural and commonplace human behavior, seems deceptively simple to judge in terms of "interestingness". However, similar to some other intuitive and natural behaviors, such as nodding and gazing, it is extremely complicated to predict or reconstruct a story's interestingness. This perceived interest can be subjective, is often cultural and it can also change over time (e.g. a popular movie's narrative becomes less popular among a new generation). Moreover, the subtleties and arts of authorship makes the ways in which a narrative can seem interesting incredibly diverse, subtle and nuanced. Despite such difficulties, there are ways in which we can start understanding this phenomenon and begin developing proxy measures for perceived story interestingness, to be used in generative models. To this end, the related work in the field of cognitive science is a great resource to draw from.

While various types of interest can be established in a story, many researchers have broadly categorized these interests in two main groups. Under various names, such as *individual* and *situational* (Hidi & Baird, 1986), or *cognitive* and *emotional* (Kintsch, 1980), researchers have focused on the source of interest to make such categorization. "Cognitive" interests are largely the properties of the narrative (or authorship techniques) and "emotional" interests are largely rooted in an audience's predispositions. The latter group is more subjective, and can consist of instinctive "absolute" (Schank, 1979) interests (e.g. danger, power, sex), or "topic interests" (Campion, Martins, & Wilhelm, 2009).

While it is plausible to assume that all kinds of interest affect each other when it comes to perception, cognitive interests are categorized as the less subjective factors, ones that have a larger focus on the stimuli: the properties of the narrative. Many researchers have developed theories of the mechanisms that lead to the establishment of cognitive interest in stories. Notable theories include: unexpectedness (Schank, 1979), the interaction between background knowledge, uncertainty and postdictability (Kintsch, 1980), incongruity (Mandler, 1982), change in one's belief (Frick, 1992), generation of inference (Kim, 1999), and the generation of predictive inference (Campion et al., 2009).

Many of these theories above are conceptually close to and can overlap with each other. In this paper, we focus on two of these theories that represent familiar notions: **unexpectedness** (closely related to surprise) and **predictive inference** (closely related to foreshadowing).

A detailed overview of the theories of story interestingness is provided in (Behrooz, Mobramaein, Jhala, & Whitehead, 2018).

Search for Specificities

Another reason for creating proxy measures for story interestingness is the potential roles of such measures in choosing an appropriate set of specificities in a narrative.

Picking the Right Specificity in a Situated Context If a narrative generation system, for instance one used by an agent operating in the real world, attempts to build a narrative from events that have previously happened, there would be a search problem involved to choose which observations, details or specificities (if any) should be included in the story. At a minimum, a sequence of events can be described as a mundane narrative that minimally describes the story's events. However, the inclusion of certain specificities about the elements in the story is usually what allows for authorship skills.

The "Chekhov's Gun" principle says: "every element in a story must be necessary, and irrelevant elements should be removed." On the other hand, many seemingly unnecessary parts of a telling of a story serve the particular purpose of making the narrative more interesting (e.g. through foreshadowing or red herring techniques). For instance, specifying that "the moon was shining bright" a few events before two characters (that the audience may suspect are in love) kiss for the first time, asserts a property of the moon that is (most likely) inconsequential to what happens in the story, but is nonetheless a part of what makes the telling of it interesting.

Thus, while completely irrelevant details and specificities

can violate Chekhov's Gun principle, some details and specificities, when chosen and employed in an informed and artistic way, can contribute to the interestingness of narrative when perceived by an audience.

Complementing Approaches That Involve Generalization of Concepts This search problem can also arise when generative neural networks (such as RNNs) are used to generate stories. In order to increase the chances of convergence in such models, researchers sometimes replace verbs and words in a story corpus that is used to train the model with generalized concepts (Martin et al., 2017) using semantic word networks such as VerbNet and WordNet (Schuler, 2005; Miller, 1995). This would result in the replacement of both of the words "car" and "automobile"with the semantic label "self-propelled vehicle.n.01", and consequently, it becomes easier for the model to find event patterns involving either of these words. However, the narratives generated using such models would then also include the generalized concepts, and hence, they can be more mundane and less specific as a result. Having proxy measures to find the more interesting specificities may offer a solution to this problem. In particular, word vectors can help with choosing a specific instance of a semantic label. This lack of specificity can occur in any generative method for open story generation that involves generalization of concepts or events, and consequently results in mundane generated stories, such as in (Li, Lee-Urban, Johnston, & Riedl, 2013).

Cognitive Interest as a Proxy Measure

In the absence of a domain's semantic model (as explained in previous sections), we explore the idea of using word embedding vectors with the goal of developing proxy measures for story interestingness. Word vectors introduce a way to estimate the semantic similarity and relationships between words, largely based on co-occurrence. The rapid improvements in deep learning have greatly contributed to the quality of word embeddings and they have seen much success in many computational linguistic tasks. In this paper, we investigate the use of word embeddings to estimate the cognitive interest in stories.

Foreshadowing

As briefly reviewed before, one of the main causes of the establishment of cognitive interest in stories is predictive inference by the audience (Campion et al., 2009). Among the diverse set of reasons why and ways in which a reader may try to infer what will occur in the continuation of a story, we focus on a common way in which authors attempt to intentionally cause such inference in the reader. Commonly known as *foreshadowing* (Chatman, 1980), this authorship techniques involves giving readers implicit hints that can, in various ways, provide clues about the upcoming noteworthy events in the story. Foreshadowing can have various degrees of subtly. In some cases, it can create a vivid question mark in user's mind about why a particular point is mentioned in the story (e.g. "the road seemed scary and dark, with no barriers in the middle of it"). In such cases, foreshadowing is more likely to lead to predictive inference. At other times, what is also recognized as foreshadowing may be too subtle of a hint to drive predictive inference and may not pose a question mark to the user until a later event reveals a rather cryptic connection. In both cases, the goal is for the reader to realize this connection and make sense of a "coherent macrostructure" of the story in retrospect; a notion called postdictability by Kintsch (Kintsch, 1980).

There have been a few notable attempts to generate foreshadowing in stories. Minstrel (Turner, 1994), relying on a bank of stories that it has seen before and knows about, attempts to foreshadow those upcoming events that are uncommon and hence unexpected. In (Bae & Young, 2008), another planning-based system provides solutions for generating foreshadowing and flashbacks for events that are found to be surprising. Suspenser (Cheong & Young, 2006) uses similar approaches to generate suspense in a planning-based story generation system. While our focus on cognitive interest and foreshadowing is not part of a story generation system, it can be used in one and the aforementioned system are a great source of inspiration for our work. However, as explained earlier, our focus is on systems that cannot assume the levels of semantics needed for use in planning-based systems.

Using Word Vectors to Find Foreshadowing

Estimating the presence of foreshadowing, without a semantic model of the domain, is a complicated task. Foreshadowing can take many different shapes, be causal or non-causal, and can depend on domain-specific clues. However, certain cases of foreshadowing involve usage of words that co-occur in many contexts and hence, are likely to have similar word vectors in an embedding space. This is the main intuition behind our approach.

Obtaining the Story Keywords Consider the example story in Table. 1. It contains a case of foreshadowing with a potential to cause predictive inference in the reader: event 5 (waiter is distracted and tired) foreshadows event 7 (food is wrong, waiter apologizes). Treating all the words in the story as a bag-of-words, we first remove stop words (e.g. "the", "is"), and then further narrow down our selection of words using part-of-speech tags. In order to focus on the words that capture most of the events and descriptions in the story, we select verbs, nouns and adjectives. Specifically, for verbs we use verb roots extracted via VerbNet (Schuler, 2005) and for nouns we exclude named-entities such as "Sam". It is worth noting that the current target for state-of-the-art open story generation approaches is short stories that are 6-10 sentences (Purdy et al., 2018).

Table 1: An example story which contains a case of foreshadowing. The numbers on the left are story event indexes.

1	Sam and Judy went out for dinner at their favorite restaurant.
2	While driving to the restaurant, Judy's favorite song played on the radio.
3	Sam found a parking space at the very front of the restaurant.
4	Sam and Judy were seated immediately and ordered their favorite food to the waiter.
5	The waiter looked distracted and tired but was polite while taking their order.
6	Sam's favorite song played on the radio while they waited for their food.
7	When the waiter returned with their food it was all wrong! The waiter apologized and returned a few minutes later with the correct order.
8	Sam and Judy enjoyed their meal.
9	They paid their tab, left a tip for the waiter, and drove back home.

Table 2 shows the keywords extracted as above for the story in Table. 1 (using Stanford CoreNLP (Manning et al., 2014) for part-of-speech tags).

Table 2: Extracted keywords from the story in Table. 1.

waiter, return, pay, song, seat, order, radio, look, go, apologize, dinner, take, home, wrong, favorite, find, space, leave, minutes, restaurant, food, enjoy, parking, tired, drive, distracted, front, correct, meal, tip, tab, play, wait

Vectorizing and Visualizing the Story Keywords We used GloVe embeddings, pre-trained on Wikipedia articles, in order to obtain a set of vectors that represent the words in Table 2. Hence, this set of vectors represent the major occurrences and descriptions in the story, as they map onto the embedding space at use. Moreover, by extension, these vectors can also represent major groups of concepts that are perceived by the audience when reading the story.

The original embedding space used is 300-dimensional. In order to visualize the word vectors, we used the T-SNE algorithm (Maaten & Hinton, 2008) to yield a 2-D representation of them. The results can be seen in Fig. 1.

Interpreting the Vector Space The T-SNE visualization shows to us that certain clusters of words can be distinguishable from others. These clusters can semantically categorize the contents of the story without any semantic models of the domain, such that each focus on a particular aspect of the story, involving its own events, objects and specificities. Attempting to extract some of the clusters seen in Fig. 1, we notice the following by grouping the words that are reasonably close to each other:

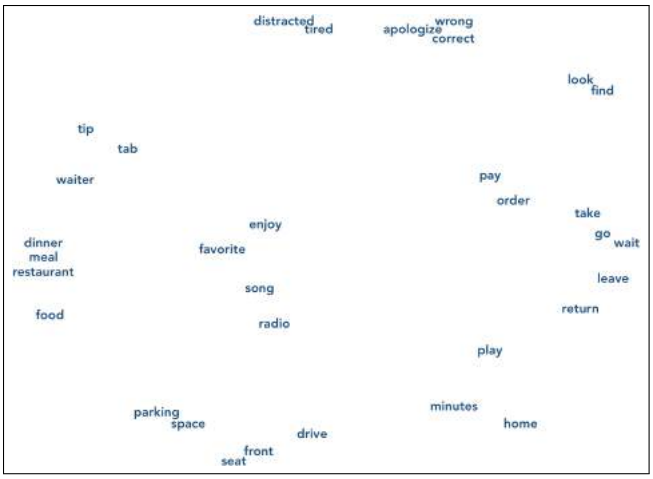


Figure 1: 2-dimensional T-SNE visualization of the GloVe vectors representing the keywords in Table 2.

- **dining:** waiter, restaurant, dinner, meal, food, tip, tab
- **car:** parking, front, seat, space, drive
- **logistics:** return, take, wait, go, leave
- **music:** song, radio, enjoy, favorite
- **cashier:** pay, order
- **searching:** look, find
- **mistake-recovery:** distracted, tired, apologize, wrong, correct
- **play-minute-home:** play, minute, home

It is also worth noting that other unsupervised clustering approaches, such as K-means, would lead to very similar clusters. We used T-SNE for this analysis because K-means proved less deterministic and could yield less predictable results depending on its starting state; however, the distance between two given pairs of word vectors is constant, hence, T-SNE depicts an appropriate representation of those constant distances.

Finding the "Key Event" Usually, a key event in a short story (or a segment of a long one) is the target of foreshadowing. In classic dramatic structures, such event can play the role of the story "climax" (Elson, 2012b). Alternatively, an "inciting incident" in the story (McKee, 1997) can become the subject of foreshadowing. Such events are often followed by a resolution (e.g. the correct food order is then brought, in our example story). Usually, this key event is unexpected,

surprising, or otherwise interesting to the audience, such that it would justify the telling of the story in the first place. Finding this key event without semantic models of the story's domain is not an easy task. Most techniques employed for this purpose depend on irregularities and unexpectedness in a story. In order to find irregularities, one would need to first develop an understanding of regular progressions of the story first (without relying on a priori semantics about them). In (Behrooz, Swanson, & Jhala, 2015), for instance, sequence modeling is employed to build a model of regular event sequence in a domain, and subsequently, irregular progressions of the story and the events that cause them are identified.

In this paper, we use the *cosine similarity* of vectors representing all of the verbs in the story in order to find the most anomalous verb. Based on the above, this verb has the highest chance of being part of the key event. In Table 3, all of the roots of the verbs in the story in Table 1 are listed along with the cosine similarity metric between *each verb root vector* and the *mean of all verb root vector* in the story. This measure can indicate how close or far each verb vector is from the rest of the verbs in the story, and hence, how semantically related or unrelated.

Table 3: Verb roots of all of the verbs in the story in Table 1 (excluding stop words), along with a cosine similarity distance between each verb root vector and the mean of all verb root vectors. Verb roots are obtained using VerbNet (Schuler, 2005), and word vectors using a pre-trained GloVe model (Pennington et al., 2014).

Verb root	Cosine similarity
go	0.838
drive	0.517
play	0.609
find	0.734
seat	0.416
order	0.566
look	0.698
take	0.839
favorite	0.458
wait	0.697
return	0.697
apologize	**0.335**
enjoy	0.57
pay	0.631
leave	0.744

As we can see in the Table 3, the verb `apologize` is the most anomalous verb in our example story, since it has the lowest cosine similarity score with the mean of all verb root vectors. We identify this verb as the *key verb* in the story, and since the key verb is mentioned in event 7 (in Table 1), we also identify that event as the *key event* in the story.

Finding the Foreshadowing Cluster Given the key event and key verb, as described above, we can use the keyword clustering of the story, seen in 1, to find out if there exists a cluster whose constituent keywords:

1. play a role in the key event and include the key verb, and,
2. play a role in one other preceding event (or sentence) in the story.

With such constraints considered, we can see that the **mistake-recovery** can be the *foreshadowing cluster*; a cluster that includes the words involving the foreshadowing in the story.

Finding the Foreshadowing The preceding event or sentence in the story in Table 1, in which the foreshadowing cluster plays a role, is event 5. Hence, we can guess that event 5 has a chance of foreshadowing our key event, 7. Moreover, as a whole, these steps can result in an estimate of the presence of foreshadowing in the story.

Unexpectedness

As mentioned before, many approaches to open story generation focus on finding the usual progressions of events in the story. Among such approaches are story scripts (Schank & Abelson, 2013) which argue that plots about many domains of storytelling usually follow a similar general pattern. Another example are Plot Graphs (Li et al., 2013), which use crowdsourcing to build networks of usual progressions and precedence rules of events (e.g. a graph covering many of the usual paths that a "dining at a restaurant" story would cover). In (Purdy et al., 2018), using a corpus of movie plot summaries, a temporal ordering network is created to capture the common ordering of verbs in stories. The resulting proxy measure, introduced earlier as "temporal ordering", is then used to find the extent to which a new sequence of events adheres to the common ordering of events in stories.

While such adherence would help estimate a correct causal chain of events or logical precedence between them, it is noteworthy that one of main reasons for cognitive interest in stories is the *unexpectedness* of events (Schank, 1979). Hence, as a story generator would benefit from a proxy measure for correct temporal ordering of events, it may also benefit from one that rewards it for having some unexpected event.

"The Inverted-U Function" Kintsch (Kintsch, 1980) argues that cognitive interest can be an "inverted-U" function of knowledge and uncertainty about the story. Simply described, this view argues that if a story creates too many or too few question marks in user's mind, it is less likely to be perceived as interesting. This guides us towards a proxy measure that can have a higher value if a story deviates in small amounts from the usual ordering of events, and a lower

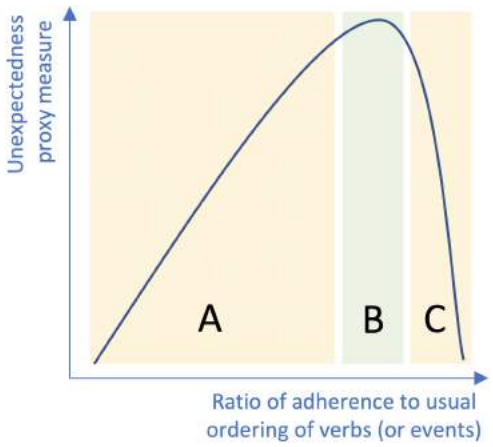


Figure 2: An illustration of a cognitive interest proxy measure based on unexpectedness and inspired by Kintsch arguments (Kintsch, 1980). The area marked as *A* denotes a story that does not sufficiently adhere to the usual ordering of verbs (or events). *C* shows an area where there is no or too little deviation from the usual ordering for the story to cause cognitive interest. *B* shows an area indicating that the story generally adheres to the usual ordering, but contains enough deviations and hence may cause cognitive interest.

value if it deviates too much from (or does adheres at all to) the usual ordering. Using a temporal ordering network, for instance, an unexpectedness proxy measure can have its highest value if most but not all (e.g. 90%) of the pairs of verbs in the story adhere to the network's order. The proxy measure would sharply decrease if this adherence ratio is much less, or approaches 1. An illustration of such proxy measure function can be seen in Fig. 2.

Unexpectedness and Word Vectors Using a vector space that represents verbs (or sentences (Pagliardini, Gupta, & Jaggi, 2017)) in a story, the distance between each vector and the average of all vectors belonging to a story (similar to Table 3) can estimate how unexpectedly each verb is perceived compared to the rest of the story. Hence, in order to follow an inverted-U pattern, a proxy measure of unexpectedness can have the highest value when most entries in Table 3 have large values, but at least one entry has a much lower value than others.

Conclusion

Communication context is a consequential factor in narrative generation, in terms of approach, constraints, and evaluation criteria. Certain narrative generation approaches are tied to a specific communication context (e.g. games) and depend on that context's a priori semantics to evaluate how good a generated story is. Other approaches are not bound to a specific context (called *open story generation*) and often use generic text quality measures to assess the quality of the story. Given the importance of narrative perception in real-world use cases of such story generation (e.g. by an intelligent agent), we draw from theories of cognitive interest and investigate the use of word embeddings vectors to find how interesting a generated narrative is. Specifically, we assess the existence of predictive inference (through foreshadowing) and unexpectedness in stories, using GloVe word vectors (Pennington et al., 2014). We plan to evaluate this approach in a situated scenario and seek to find correlations between proxy measures of cognitive interest and judgments of human subjects.

References

Bae, B.-C., & Young, R. M. (2008). A use of flashback and foreshadowing for surprise arousal in narrative using a plan-based approach. In *Joint international conference on interactive digital storytelling* (pp. 156–167).

Behrooz, M., Mobramaein, A., Jhala, A., & Whitehead, J. (2018). Cognitive and experiential interestingness in abstract visual narrative. *Cognitive Science Society (CogSci)*.

Behrooz, M., Swanson, R., & Jhala, A. (2015). Remember that time? telling interesting stories from past interactions. In *Interactive storytelling*. Springer.

Campion, N., Martins, D., & Wilhelm, A. (2009). Contradictions and predictions: Two sources of uncertainty that raise the cognitive interest of readers. *Discourse Processes*, *46*(4), 341–368.

Chatman, S. B. (1980). *Story and discourse: Narrative structure in fiction and film*. Cornell University Press.

Cheong, Y.-G., & Young, R. M. (2006). A computational model of narrative generation for suspense. In *Aaai* (pp. 1906–1907).

Elson, D. K. (2012a). Detecting story analogies from annotations of time, action and agency. *Proceedings of the Third Workshop on Computational Models of Narrative*(1981), 91-99.

Elson, D. K. (2012b). *Modeling narrative discourse*.

Fong, T., Nourbakhsh, I., & Dautenhahn, K. (2003). A survey of socially interactive robots. *Robotics and autonomous systems*, *42*(3-4), 143–166.

Frick, R. W. (1992). Interestingness. *British Journal of Psychology*, *83*(1), 113–128.

Goodrich, M. A., Schultz, A. C., et al. (2008). Human–robot interaction: a survey. *Foundations and Trends® in Human–Computer Interaction*, *1*(3), 203–275.

Hidi, S., & Baird, W. (1986). Interestingnessa neglected variable in discourse processing. *Cognitive Science*, *10*(2), 179–194.

Kim, S.-i. (1999). Causal bridging inference: A cause of story interestingness. *British Journal of Psychology*, *90*(1), 57–71.

Kintsch, W. (1980). Learning from text, levels of comprehension, or: Why anyone would read a story anyway. *Poetics*,

9(1-3), 87–98.

Li, B., Lee-Urban, S., Johnston, G., & Riedl, M. (2013). Story generation with crowdsourced plot graphs. In *Aaai.*

Maaten, L. v. d., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, *9*(Nov), 2579–2605.

Mandler, G. (1982). The structure of value: Accounting for taste. *Center for Human Information Processing Report*, *101*.

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., & McClosky, D. (2014). The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations* (pp. 55–60).

Martin, L. J., Ammanabrolu, P., Wang, X., Hancock, W., Singh, S., Harrison, B., & Riedl, M. O. (2017). Event representations for automated story generation with deep neural nets. *arXiv preprint arXiv:1706.01331*.

Mateas, M., & Stern, A. (2003). Façade: An experiment in building a fully-realized interactive drama. In *Game developers conference* (Vol. 2).

McCoy, J., Treanor, M., Samuel, B., Mateas, M., & Wardrip-Fruin, N. (2011). Prom week: social physics as gameplay. In *Proceedings of the 6th international conference on foundations of digital games* (pp. 319–321).

McKee, R. (1997). *Substance, structure, style, and the principles of screenwriting*. New York: HarperCollins.

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, *38*(11), 39–41.

NPR, E. R. (2017). The smart audio report [Computer software manual]. Retrieved from `https://www.nationalpublicmedia.com/smart-audio-report/` (Accessed: 2018-07-20)

Pagliardini, M., Gupta, P., & Jaggi, M. (2017). Unsupervised learning of sentence embeddings using compositional n-gram features. *arXiv preprint arXiv:1703.02507*.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics* (pp. 311–318).

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp)* (pp. 1532–1543).

PÉrez, R. P. Ý., & Sharples, M. (2001). Mexica: A computer model of a cognitive account of creative writing. *Journal of Experimental & Theoretical Artificial Intelligence*, *13*(2), 119–139.

Purdy, C., Wang, X., He, L., & Riedl, M. (2018). Predicting generated story quality with quantitative measures.

Riedl, M. O., & Bulitko, V. (2012). Interactive narrative: An intelligent systems approach. *AI Magazine*, *34*(1), 67.

Ryan, J., Mateas, M., & Wardrip-Fruin, N. (2016). Characters who speak their minds: Dialogue generation in talk of the town. *Proc. AIIDE*.

Schank, R. C. (1979). Interestingness: controlling inferences. *Artificial intelligence*, *12*(3), 273–297.

Schank, R. C., & Abelson, R. P. (2013). *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Psychology Press.

Schuler, K. K. (2005). Verbnet: A broad-coverage, comprehensive verb lexicon.

Swanson, R., & Gordon, A. S. (2008). Say anything: A massively collaborative open domain story writing companion. In *Interactive storytelling* (pp. 32–40). Springer.

Turner, S. R. (1994). Minstrel: A computer model of creativity and storytelling.

Multimodal Event Knowledge in Online Sentence Comprehension: The Influence of Visual Context on Anticipatory Eye Movements

Valentina Benedettini (valentina.benedettini@sns.it)
Scuola Normale Superiore, p.za dei Cavalieri 7
I-56126 PISA, Italy

Pier Marco Bertinetto (piermarco.bertinetto@sns.it)
Linguistica Generale, Scuola Normale Superiore, p.za dei Cavalieri 7
I-56126 PISA, Italy

Alessandro Lenci (alessandro.lenci@unipi.it)
Dipartimento di Filologia, Letteratura e Linguistica, Università di Pisa, Via Santa Maria 36
I-56126 PISA, Italy

Ken McRae (kenm@uwo.ca)
Department of Psychology, Social Science Centre, University of Western Ontario
1151 Richmond St, London, ON N6A 3K7, Canada

Abstract

People predict incoming words during online sentence comprehension based on their knowledge of real-world events that is cued by preceding linguistic contexts. We used the visual world paradigm to investigate how event knowledge activated by an agent-verb pair is integrated with perceptual information about the referent that fits the patient role. During the verb time window participants looked significantly more at the referents that are expected given the agent-verb pair. Results are consistent with the assumption that event-based knowledge involves perceptual properties of typical participants. The knowledge activated by the agent is compositionally integrated with knowledge cued by the verb to drive anticipatory eye movements during sentence comprehension based on the expectations associated not only with the incoming word, but also with the visual features of its referent.

Keywords: event knowledge; anticipatory eye movements; visual perception; prediction.

Introduction

People use their experiences of events in the world to organize their semantic knowledge about objects and actions (Radvansky & Zacks, 2014). For example, the event of "going to a restaurant" implies the presence of waiters, tables, food, and money as well as actions of cooking, serving, and eating. Several studies have illustrated the central role of knowledge about events in online sentence comprehension. Event knowledge is cued by lexical items, integrated to form a coherent representation of the situation being described, and used to generate expectations about incoming input. (Tanenhaus et al., 1995; Altmann, 1999; Altmann & Kamide, 1999, 2004, 2007; Kamide et al., 2003; Knoeferle, Crocker, Scheepers, & Pickering, 2005; Knoeferle & Crocker, 2006, 2007; Bicknell et al. 2010; Matsuki et al. 2011; Metusalem et al., 2012). In this paper, we present an eye-tracking experiment that investigates the hypothesis that event knowledge activated during sentence comprehension is inherently multimodal, because it derives from people's sensori-motor (i.e., watching and performing events) and linguistic experiences (i.e. talking and reading about events), and allows people to generate expectations not only about the most likely noun filler of a verb's thematic role (e.g., *ball* as a typical patient of *throw*), but also about the visual properties of the noun referent (e.g., oval ball vs. round ball).

We used the visual world paradigm to investigate how event knowledge activated by an agent-verb pair is integrated with perceptual information about the referent that fits the patient role. For instance, the noun *ball* can refer to a small white baseball, to a large orange basketball, or to a large oval (American) football. We call these nouns *perceptually underspecified*, because the noun in isolation does not entail a specific type of perceptual referent. This affects the kind of predictions that people will generate. Compare for instance the following sentences:

(1) a. *The man threw the ball*.
b. *The quarterback threw the ball.*

In (1a), we cannot anticipate which type of ball was thrown, without further contextual information. Conversely, in (1b) we can predict that the ball is likely to be an oval football. Our hypothesis is that this prediction about the patient in (1b) depends on the integration of event-based knowledge cued by the agent and the verb. In particular, *quarterback* activates knowledge about football, including

that the ball is oval. Once this information is integrated with *throw*, predictions are generated that make *ball* a highly expected patient noun and allow comprehenders to anticipate the specific object to which it refers.

In the present experiment, participants read sentences such as *The doctor/bartender uncaps the bottle*, in which agent-verb pairs denote events that activate knowledge about plausible noun fillers of the patient role. The visual scenes contained two objects that may fit the event expressed by the verb (a pill bottle and a beer bottle). The patient role was filled by a perceptually-underspecified noun that can denote both objects (*bottle*). Anticipatory eye movements on the predicted object mirror the integration of the event-based knowledge activated by the agent-verb pair and perceptual information coming from the visual input during online sentence comprehension.

Related Studies

Words encode mutual expectations between events and their typical participants (McRae et al., 1998; Ferrettti et al., 2001; McRae et al., 2005; Hare et al., 2009). McRae, et al. (2005) found that agents, patients and instruments prime verbs that describe events in which they typically are involved (*waiter*, *chainsaw* and *guitar* prime verbs like *serving*, *cutting* and *strummed*). Bicknell et al. (2010) conducted an Event Related Potential (ERP) experiment to investigate whether an already filled role affects how another role can be filled. They found that typical agent-patient pairs such as *journalist-spelling* and *mechanic-brakes* in *The journalist checks the spelling* and *The mechanic checks brakes* elicited reduced N400s as compared to *The journalist checked the brakes* and *The mechanic checked the spelling*. The effects on N400 amplitudes show both generalization across input modalities and regularity between N400 properties and sensory, conceptual and linguistic factors, suggesting that the effects are modality sensitive but not modality specific (Kutas & Federmier, 2011). According to Kuperberg and Jaeger (2016), "prediction" concerns a change in the state of the language processing system based on the context prior to the availability of new input. The context involves both linguistic and extralinguistic information, that can facilitate the processing of new information at multiple levels of representation, which interact and communicate during language processing. Contextual information includes semantic knowledge about specific events, event structures, event sequences, and general schemas (Altmann & Mirković, 2009; Radvansky & Zacks, 2014). According to Knoeferle and Guerra (2016), during sentence comprehension visual perceptual information interacts with word knowledge. Some eye tracking studies have manipulated argument-verb combinations to investigate anticipatory eye movements (Altmann & Kamide, 1999; Kamide, et al., 2003; Knoeferle & Crocker, 2006, 2007).

Altmann and Kamide (1999) investigated the hypothesis that people tend to predict which object will fit the patient role after hearing the verb. They used sentences like *The boy will eat the cake* in combination with pictures of a boy, a birthday cake, a toy car, a toy train and a ball. Subjects fixated the single edible object in the scene (birthday cake) more often than the other depicted objects before hearing *cake*. By contrast, when subjects heard *The boy will move the cake* with the same visual scene they looked equiprobably at all of the movable objects. This shows that verb selectional preferences constrain the set of possible objects that follow the verb. Kamide, Altmann and Haywood (2003) investigated whether agent-verb pairs elicit anticipatory eye movements toward entities that fit the patient role. Sentences such *as The man will ride the motor bike* and *The girl will ride the carousel* were combined with pictures of a motorbike, a carousel, a beer and a sweet. The same visual scene was presented while participants listened to *The man will taste the beer* and *The girl will taste the sweet*. Anticipatory eye movements on the predicted objects (motorbike and carousel; beer and sweet) were triggered by the verb. The results are consistent with the assumption that expectations associated with agent-verb pairs help people to predict which entity fills the incoming patient role.

Knoeferle and Crocker (2006, 2007) performed an eye tracking experiment to investigate the interplay between current visual context and event knowledge during sentence comprehension. Sentences such as *The detective will soon spy on the pilot* and *The wizard will soon spy on the pilot* (in German) were combined with pictures of a wizard looking a pilot through the telescope, a detective serving the pilot some food, a pilot and a tree. In the verb time window (*spy*) when listening to *The wizard will soon spy on the pilot* (which corresponds to the event occurring in the visual scene) participants often looked more at the wizard, though spying is a detective's typical action. Since the visual scenes provided information that conflicts with typical event knowledge (wizard spies vs. detective spies), the outcomes are consistent with the assumption that listeners exploit information coming from current visual context during online comprehension. These studies suggest that contextual information includes multiple types of knowledge such as event structures and sensory input. Predictions are strongly associated with the interplay among words, event contingencies and conceptually combined knowledge (Altmann & Mirković, 2009; Altmann & Kamide, 2004, 2007; Barsalou, 2008; Hagoort et al., 2004).

Experiment

We investigated how event knowledge activated by an agent-verb pair influences pre-activation of multimodal information about the referent that fits the patient role. Sentences like *The doctor uncaps the bottle* were combined with four pictures such as a pill bottle (target), a beer bottle

(action related object), a syringes (agent related object) and a comb (unrelated object), as shown in Figure 1:

1. **target objects** fit the patient role given the agent-verb combination. Since doctors prescribe and sometimes administer medication, typically they open pill bottles rather than beer bottles;
2. **action related objects** fit the verb (a beer bottle can be uncapped), but not the agent-verb combination
3. **agent related objects** corresponded to objects that commonly occur in situations together with the agents, such as doctors and syringes;
4. **unrelated objects** were not congruent with the agent, verb, or agent-verb combination.

Figure 1. Combination of visual and linguistic stimuli.

The sentence stimuli were divided into two lists and the targets of the first list became the action related objects in the second list, which contained the same verb but a different agent. In *The bartender uncaps the bottle*, for example, the beer bottle was the predicted object (target), and the pill bottle was the action related object. Since the verb-patient pairs co-occur with different agents in the two lists, the agent related objects changed as well. The noun *bartender* cues situations that involve objects such as taps and mug, while *doctor* triggers situations involving surgical scalpels and stethoscopes. The agents activate knowledge about objects that commonly occur in the events performed by them (targets and agent related objects).

Method

Norming

We measured the strength of the association between the agents and the predicted object (target) images. We used the Figure Eight crowdsourcing platform[1] to create a task in which participants evaluated how likely it was that the agent and the object appeared in the same situation, using a scale that ranged from 1 (not very likely) to 7 (very likely). Participants read the name of the agent, such as *doctor*, opened the link for the object picture (pill bottle), and rated "How likely is it that the person and the object appear in the same situation?". The mean ratings were 6.3 and the 95% confidence interval was 0.1. Thus, the agents and the objects were judged to co-occur strongly in the same real-world situations.

Participants

Twenty-four University of Western Ontario undergraduate students were compensated $10 for their participation. They ranged in age from 19 to 28 years. All participants had normal or corrected to normal visual acuity and self-reported English as their native language. Self-reportedly, participants had never endured a traumatic brain injury or illness and were not currently diagnosed with any major psychiatric illness.

Sentences

There were 60 trials consisting of 30 experimental and 30 filler trials. In the experimental trials, participants heard sentences in which the agent performs an action that could be associated with two pictures in the visual scene, the target and the action related object. The patient role was filled by a perceptually underspecified noun that could refer to both objects. The sentences were split into two lists to present only one type of verb-patient pair to each participant. Fifteen filler trials consisted of two pictures of objects that could be denoted by the same word but the sentence did not refer to either of them. It referred instead to a third object. For example, *The man does not like candies* was combined with pictures of a candy, a fishing hook, a coat hook and a candelabra. An additional 15 filler sentences had various syntactic structures and one word referred to one of the pictures (e.g., *Karen made the tea with her new pot* with pictures of a teapot, a marble, a picture frame, a mitten). We used four practice trials to familiarize participants with the experiment.

Auditory Stimuli

A female native English speaker recorded all sentences. They were recorded using Audacity Cross-Platform Sound Editor 2.2.2 (released February 20 2018), and annotated by marking relevant points of the sentence using a customized script in Praat 6.0.37 (retrieved February 3 2018). For each sentence we set a pointer at: agent onset, agent offset/verb onset, verb offset/second article onset, second article offset/patient onset and patient offset as well as the start and end of the sentence. The agent offset/verb onset was normalized in all auditory files (1200 ms).

[1] https://www.figure-eight.com/

Visual Stimuli
All images were presented at 300x300 pixels in colour. Each picture was placed in a different quadrant of the screen at a 45-degree angle from the center. The location of the four images was randomized across trials and participants. The pictures were selected from BOSS[2], KONKLAB[3] and COGPSY Image Corpora.

Eye Tracker
We used a desktop mounted Eyelink 1000 and Experiment Builder, Version 1.10.1241 software (SR Research Ltd.). The camera lens was positioned approximately 60 cm from the participant's head at an approximately 35-degree angle to the participant's eyes. Participants were positioned 70 cm away from a 16-inch monitor displaying the visual stimuli (resolution set to 1024 x 768 dpi). Calibration was performed prior to the start of the experiment, as well as at any time the equipment registered significant head movement.

Procedure
During the first ten seconds of each trial a fixation cross was presented. The participant was then redirected to calibration. After three seconds during which the participant fixated the cross, this was replaced by the four trial images. Participants had one second to become familiar with the images before the auditory stimulus began. A series of red circles were flashed in the center of the screen to bring the participant's attention back to the fixation cross. The sentence began when participants fixated the cross. The four pictures remained on the screen while the sentence was presented and participants' eye movements were recorded. An additional 300 ms of silence followed the end of the sentence. When the images disappeared, the next trial began. Before starting the session, participants were assigned to a list. Each list contained three trial blocks. At the start of the experiment, participants received the following instructions: "You will see a display with four pictures while hearing a sentence. There is no task involved; just look at the pictures and listen to the sentences. We'll start with some practice trials to see how it works." The first block contained four practice trials. Thereafter, participants saw: "This is the end of the practice sessions for part one. Do you have any questions before the experiment begins?" The other two trial blocks contained the experimental and filler trials randomly presented for each participant. Instructions were repeated at the start of each block. An equal number of experimental and filler items were presented in each list. Participants were given a short break between blocks to rest their eyes.

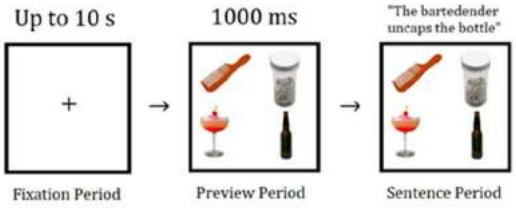


Figure 2. Example of the procedure for one trial.

Results

We recorded the proportion of fixations on the target pictures and compared them to the proportions of fixations on the other pictures (agent related, action related and unrelated) in specific time windows (agent, verb and patient). We analyzed three time windows: the agent (*bartender*); the verb + article (*uncaps the*), which is the anticipatory time window, and the patient (*bottle*). The Area Of Interest (AOI) for each picture consisted of each screen quadrant. The analyses were conducted with RStudio Version 1.1.463 (2009-2018). We fit one Linear Effects Mixed Model (LMER) for each time window using the lmer() function from the linear mixed effects package lme4 (Bates et al., 2015; Baayen et al., 2008; Barr et al., 2013). The four AOIs and the two lists are the fixed effects. We calculated two random slopes accounting for random effects (subjects and trials). Fixed and random effects remain stable for each model and during all the analyses conducted on the dataset. For each time window, we calculated estimated means of proportions, Standard Errors, *t*-values, and *p*-values of AOIs comparisons (Table 1).

Agent window
The agent time window extended from agent onset (610 ms) to verb onset (1200 ms). The duration was 590 ms. The onsets of the spoken sentences were preceded by a silence to normalize the verb onset (457 ms). There were no significant differences in proportions of fixations. Moreover, there were no significant differences in proportions of fixations between the action, agent related and unrelated objects (Table 1).

[2] https://sites.google.com/site/bosstimuli/
[2] http://konklab.fas.harvard.edu/#

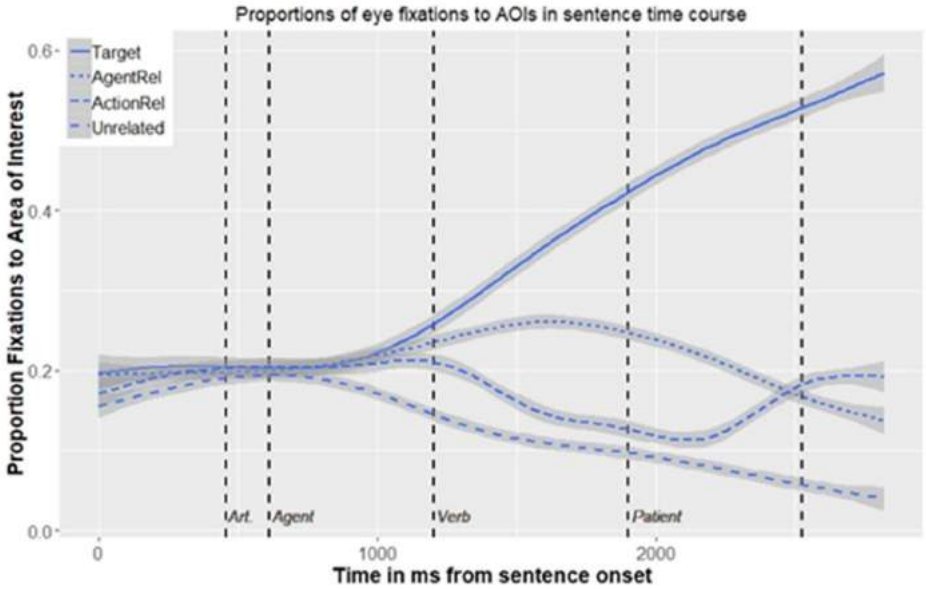


Figure 3. Proportions of fixations on AOIs across the sentence time course. "Art", "Agent", "Verb" and "Patient" correspond to the mean onset of the first article (456 ms), agent (610 ms), verb (1200 ms) and patient (1899 ms).

Table 1. Results of comparisons between pairs of AOIs in each time window (* = p < 05).

Time Window	Comparison	Estimate	SE	t value	p-value
Agent	Target-ActionRel	0.01	0.02	0.51	0.62
	Target-AgentRel	0.00	0.02	0.21	0.83
	Target-Unrelated	0.04	0.02	1.68	0.11
	ActionRel-AgentRel	-0.01	0.02	-0.31	0.76
	ActionRel-Unrelated	0.03	0.02	1.26	0.22
	AgentRel-Unrelated	0.03	0.02	1.56	0.13
	List1-List2	0.03	0.03	1.07	0.30
Verb	Target-ActionRel	0.17	0.03	5.91	**3.84e-06***
	Target-AgentRel	0.08	0.02	4.74	**8e-05***
	Target-Unrelated	0.22	0.03	8.05	**2.19e-08***
	ActionRel-AgentRel	-0.09	0.02	-3.72	**0.001***
	ActionRel-Unrelated	0.05	0.02	3.18	**0.002***
	AgentRel-Unrelated	0.14	0.02	6.19	**1.30e-06***
	List1-List2	0.05	0.03	1.85	0.08
Patient	Target-ActionRel	0.35	0.04	8.02	**2.97e-08***
	Target-AgentRel	0.31	0.04	7.42	**1.16e-07***
	Target-Unrelated	0.43	0.04	11.48	**2.80e-11***
	ActionRel-AgentRel	-0.04	0.02	-1.79	0.09
	ActionRel-Unrelated	0.81	0.02	4.88	**1.83e-05***
	AgentRel-Unrelated	0.12	0.02	6.71	**1.78e-07***
	List1-List2	0.03	0.02	1.73	0.1

Verb window

The verb time window extended from verb onset (1200 ms) to the second article offset/patient onset (1899 ms). The duration was 699 ms. Participants fixated the object that fit the agent-verb combination more often than the objects that were associated with the verb only, the agent only or the unrelated object. Furthermore, the agent-related and action-related objects were fixated significantly more often than the

unrelated object. Finally, participants fixated the agent-related object more often than the action related object.

Patient window

The patient time window extended from the patient onset (1899 ms) and to end of sentence (2524 ms). Again, participants fixated the object that fit the agent-verb combination more often than each of the other objects. Both the agent-related and action-related objects were fixated more often than the unrelated object.

Discussion

Our results support the hypothesis that the knowledge activated by the agent concerning events in which it typically appears is compositionally integrated with knowledge cued by the verb, so as to drive anticipatory eye movements during online sentence comprehension. This is consistent with the assumption that during language comprehension people generate expectations using their multimodal knowledge about experienced situations in the world (Zwaan & Radvansky 1998; Barsalou 2008; Radvansky & Zacks 2014). Such integrated multimodal event knowledge allows comprehenders to resolve the perceptual underspecification of the patient noun and to anticipate the appropriate type of referent in the situation triggered by the agent-verb combination. According to Huettig and McQueen (2007), there is an interplay during the comprehension between the stored knowledge of visual properties of referents elicited by the spoken words and perceptual information in the current visual input. Our results suggest that the information in the current visual context was integrated with event knowledge cued by agent-verb pairs, eliciting the knowledge of the correct referent of the unfolding patient role. This is also consistent with Altmann and Kamide (1999), Kamide, Altmann and Haywood (2003), and Knoeferle and Crocker (2006, 2007), who demonstrated that word meaning combines with visual perceptual information to contribute to predictive processes involving event-based knowledge. This supports the hypothesis that the stored event knowledge is associated with perceptually based information that can be elicited by the current visual context and by specific agents. These cue information about particular referents that could fit the unfolding patient. What distinguishes this study from Kamide et al. (2003) is the use of very specific agents (doctor/bartender vs. girl/man) and referents (pill bottle/beer bottle vs. sweet/beer) in linguistic and visual stimuli respectively. Their combinations allowed us to investigate the hypothesis that comprehenders make extremely fine-grained predictions about referents of patient roles exploiting the event knowledge cued by agent-verb combinations and the visual context.

From a computational linguistic perspective, predicate-argument expectations have been modeled using distributional semantics (Erk, Padò and Padò 2010; Erk & Padò 2008; Lenci 2011; Santus et al. 2017). Distributional Semantic Models collect corpus-based co-occurrence statistics and encode them in vectors (also known as *word embedding*) that represent word meaning according to the so-called Distributional Hypothesis (Lenci 2018). Since these models represent the meaning exclusively in terms of connections between words, several recent studies have focused their attention on the combination of textual and visual information extracted from pictures, yielding Multimodal Distributional Semantic Models (Bruni, Tran, Baroni 2014; Lazaridou, Pham & Baroni 2015; Kiela 2016).

We plan to use multimodal distributional semantics to model the behavioral data we have collected in our experiment. We expect this computational model should be able to predict that a quarterback throws an oval ball while a pitcher throws a small white ball based on the integration of multimodal distributional information cued by lexical items.

References

Altmann, G. T. M. (1999). Thematic role assignment in context. *Journal of Memory and Language*, 41, 124-145.

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, 73, 247-264.

Altmann, G. T. M., & Kamide, Y. (2004). Now you see it, now you don't: Mediating the mapping between language and the visual world. In J. Henderson & F. Ferreira (Eds.), *The integration of language, vision and action*. Hove: Psychology Press.

Altmann, G. T. M., & Kamide, Y. (2007). The real-time mediation of visual attention by language and world knowledge: Linking anticipatory (and other) eye movements to linguistic processing. *Journal of Memory and Language*, 57, 502-518.

Altmann, G. T. M., & Mirkovic, J. (2009). Incrementality and prediction in human sentence processing. *Cognitive Science*, 33, 583-609.

Baayen, R. H., Davidson, D. J., and Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language*, *59*(4), 390-412.

Barr, D. J. (2013). Random effects structure for testing interactions in linear mixed effects models. *Frontiers in psychology,* 4.

Barsalou, L. (2008). Grounded cognition. *Annual Review of Psychology*, 59, 617-645.

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2015). Fitting Linear Mixed-Effects Models using lme4. *Journal of Statistical Software*, 67(1), 1-48.

Bicknell, K., Elman, J. L., Hare, M., McRae, K., & Kutas, M. (2010). Effects of event knowledge in processing verbal arguments. *Journal of Memory and Language*, 63(4), 489-505.

Bruni, E., Tran, N. K., Baroni, M. (2014) Multimodal Distributional Semantics. *Journal of Artificial Intelligence Research,* 49, 1-47.

Elman, J. L. (2009). On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive Science*, 33(4), 547-582.

Erk, K. & Padò, S. (2008) A structured vector space model for word meaning in context. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, 897-906.

Erk, K., Padò, S., Padò, U. (2010) A flexible, corpus-driven model of regular and inverse selectional preferences. *Computational Linguistics*, 36(4), 723-764.

Ferretti, T. R., McRae, K., & Hatherell, A. (2001). Integrating verbs, situation schemas, and thematic role concepts. *Journal of Memory and Language*, 44.

Hagoort, P., Hald L., Bastiaansen M., Petersson K. M. (2004). Integration of word meaning and world knowledge in sentence comprehension. *Science*, 304, 438-441.

Hare, M., Jones, M., Thomson, C., Kelly, S., & McRae, K. (2009). Activating event knowledge. *Cognition*, 111(2), 151-167.

Huettig, J., & McQueen, J. M. (2007) The tug of war between phonological, semantic and shape information in language-mediated visual search. *Journal of memory and language*, 57, 460-482.

Kamide, Y. (2008). Anticipatory Processes in Sentence Processing. *Language and Linguistics Compass*, 2/4 (10), 647-670.

Kamide, Y., Altmann, G. T. M., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language*, 49, 133-159.

Kiela, D. (2016) MMFEAT: a toolkit for extracting multi-modal features. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics—System Demonstrations*, 55-60.

Knoeferle, P., Crocker, M. W., Scheepers, C., & Pickering, M. J. (2005). The influence of the immediate visual context on incremental thematic role-assignment: Evidence from eye movements in depicted events. *Cognition*, 95(1), 95-127.

Knoeferle, P., & Crocker, M. (2006). The coordinated interplay of scene, utterance, and world knowledge: Evidence from eye tracking. *Cognitive Science*, 30, 481-529.

Knoeferle, P., & Crocker, M. W. (2007). The influence of recent scene events on spoken comprehension: Evidence from eye movements. *Journal of Memory and Language*, 57, 519-543.

Knoeferle, P., & Guerra, E. (2016). Visually situated language comprehension. *Language and Linguistics Compass*, 10(2), 66-82.

Kuperberg, G. R.., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, 31(1), 32-59.

Kutas, M., & Federmeier K. D. (2011). Thirty Years and Counting: Finding Meaning in the N400 Component of the Event-Related Brain Potential (ERP). *Annual Review of Psychology*, 62, 621-647.

Lazaridou, A., Pham, T. N., Baroni, M. (2015). Combining language and vision with a multimodal Skip-gram model. In *Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL*, 153-163.

Lenci, A. (2011) Composing and updating verb argument expectations: a distributional semantic model. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, 58-66.

Lenci, A. (2018). Distributional Models of Word Meaning. Annual Review of Linguistics, 4, 151-171.

Matsuki, K., Chow, T., Hare, M., Elman, J. L., Scheepers, C., & McRae K. (2011). Event-based plausibility immediately influences on-line language comprehension. J*ournal of Experimental Psychology: Learning, Memory, and Cognition*. 37(4), 913-934.

McRae, K., Spivey-Knowlton, M. J., & Tanenhaus, M. K. (1998). Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language*, 38.

McRae, K., Hare, M., Elman, J. L., & Ferretti, T. R. (2005). A basis for generating expectancies for verbs from nouns. *Memory & Cognition*, 33(7).

Metusalem, R., Kutas, M., Urbach, T. P., Hare, M., & McRae, K. (2012). Generalized event knowledge activation during online sentence comprehension. *Journal of Memory and Language*, (66), 545-567.

Radvansky, G. A., & Zacks, J. M. (2014). Event Cognition. Oxford University Press.

Santus, E., Chersoni, E., Lenci, A., & Blache, P. (2017). Measuring Thematic Fit with Distributional Feature Overlap. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 659-669.Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268(5217), 1632-1634.

Zwaan, R. A., & Radvansky, G. A. (1998) Situation models in language comprehension and memory. *Psychological Bulletin*, 123, 162-185.

Where Do Heuristics Come From?

Marcel Binz (binz@staff.uni-marburg.de)
Department of Psychology, Theoretical Neuroscience Group
Philipps-Universität Marburg

Dominik Endres (dominik.endres@uni-marburg.de)
Department of Psychology, Theoretical Neuroscience Group
Philipps-Universität Marburg

Abstract

Human decision-making deviates from the optimal solution, i.e. the one maximizing cumulative rewards, in many situations. Here we approach this discrepancy from the perspective of computational rationality and our goal is to provide justification for such seemingly sub-optimal strategies. More specifically we investigate the hypothesis, that humans do not know optimal decision-making algorithms in advance, but instead employ a learned, resource-constrained approximation. The idea is formalized through combining a recently proposed meta-learning model based on Recurrent Neural Networks with a resource-rational objective. The resulting approach is closely connected to variational inference and the Minimum Description Length principle. Empirical evidence is obtained from a two-armed bandit task. Here we observe patterns in our family of models that resemble differences between individual human participants.

Keywords: Bounded rationality; computational rationality; variational inference; reinforcement learning; meta-learning; individual differences; multi-armed bandit

Introduction

In this work we study human decision-making strategies on a stationary multi-armed bandit task. These are among the simplest sequential decision-making problems, that require reasoning about trade-offs between exploration and exploitation. In the special case of an infinite horizon and geometric discounting their Bayes-optimal solution is the Gittins index strategy (Gittins, 1979), while in general it is defined as the result of a planning process in an augmented Markov Decision Process (Duff & Barto, 2002). Prior work however suggests, that several heuristics appear to be favourable as a model of human decision-making, when compared to the Bayes-optimal solution (Steyvers, Lee, & Wagenmakers, 2009; Zhang & Angela, 2013).

Understanding human cognition in terms of heuristics has been a major theme in cognitive science over the past decades (Tversky & Kahneman, 1974; Simon, 1990; Gigerenzer & Todd, 1999). They can be viewed as crude, but realizable, approximations of optimal behavior. Heuristics are thus connected to the idea of rationality under resource constraints, which is commonly referred to as bounded rationality (Simon, 1972), computational rationality (Gershman, Horvitz, & Tenenbaum, 2015), or resource-rationality (Griffiths, Lieder, & Goodman, 2015). Examples for resource constraints include related prior experience on a given task, limited capacity of our brain or restricted deliberation times. For a more general overview of computational rationality we refer the reader to Gershman et al. (2015). Here we are interested in the hypothesis, that humans employ a learned, resourced-constrained approximation of an optimal decision-making strategy. More specifically we show, that different, potentially sub-optimal, human strategies emerge naturally in artificial learning systems when varying the strength of the constraints placed upon them. For a realization of this principle, we rely on information-theoretic concepts, similar to the approach of Ortega and Braun (2013).

We instantiate a particular kind of such resource-rational agents using recent advances from the meta-learning literature (Wang et al., 2016; Duan et al., 2016). In this framework the algorithm to be learned is parametrized by a Recurrent Neural Network (RNN). RNNs are known to be Turing-complete and hence are in theory able to realize any algorithm (Siegelmann & Sontag, 1991). The RNN is trained on a set of related tasks to act as an independent Reinforcement Learning algorithm for solving the original problem. We treat all parameters of the RNN as random variables and infer approximate posterior distributions by solving a regularized optimization problem. Varying the regularization factor leads to a spectrum of resource-rational algorithms, each possessing different properties. Models with large constraints need to rely more on prior assumptions and thus prefer simple strategies, while models with weaker constraints will approach the optimal solution (up to the representational capabilities of the RNN and the limitations of the meta-learning procedure).

The resulting approach is closely related to the Minimum Description Length (MDL) principle (Hinton & Van Camp, 1993; Grunwald, 2004), which asserts that the best model is the one, that leads to the best compression of the data, including a cost for describing the model. The bits-back argument establishes a link between the MDL principle and Bayesian learning (Honkela & Valpola, 2004), opening up connections to Bayesian theories of cognition (Griffiths, Kemp, & Tenenbaum, 2008). Indeed several heuristics have been recently interpreted as Bayesian models under strong priors (Parpart, Jones, & Love, 2018).

Our hypothesis is validated on a classical two-armed bandit task. However we view multi-armed bandits merely as

the first step towards investigating more complex tasks and the proposed algorithm is not limited to any specific problem class. The following section first introduces the framework in more general terms, before considering multi-armed bandits as a special case. We then identify different strategies of human participants and subsequently show how the proposed class of models captures important characteristics of human behavior on both a qualitative and quantitative level. Our results indicate, that the seemingly sub-optimal decision strategies used by humans might be a consequence of the constraints under which these very strategies are learned.

Methods

Reinforcement Learning

Let $M = (\mathcal{S}, \mathcal{A}, p, \gamma)$ be a Markov Decision Process (MDP), with a set of states $\mathcal{S}$, a set of actions $\mathcal{A}$, a joint distribution over the next state and a scalar reward signal, describing the dynamics of the environment, $p(s_{t+1}, r_t | s_t, a_t)$ and a discount factor $\gamma \in [0,1]$. The objective of a Reinforcement Learning (RL) agent is to find a policy $\pi(a_t|\cdot)$, that maximizes the discounted, expected return $\mathbb{E}_{p,\pi}\left[\sum_{t=0}^{\infty} \gamma^t r_t\right]$ without having direct access to the true underlying dynamics p.

Learning Reinforcement Learning Algorithms

Following the approach of Wang et al. (2016); Duan et al. (2016) we want to *learn* a RL algorithm for solving a MDP sampled from a distribution over MDPs. We parametrize the algorithm to be learned with a Recurrent Neural Network (RNN), in form of a Gated Recurrent Unit (Cho et al., 2014), followed by a linear layer. The set of all model parameters is denoted with θ in the following. The RNN takes previous actions and rewards as inputs in addition to the current state, making the output a function of the entire history $X_t = (s_0, a_0, r_0, s_1 \ldots, a_{t-1}, r_{t-1}, s_t)$. A good algorithm has to integrate information from the history in order to identify the currently active MDP, based on which it subsequently has to select the appropriate strategy. The RNN is trained to accomplish this using standard model-free RL techniques. In this work we utilize n-step Q-Learning (Mnih et al., 2016), although in theory any other algorithm could be applied as well. The RNN implements a freestanding RL algorithm through its recurrent activations after training is completed (the parameters of the RNN are held constant during evaluation). Throughout this work we use the abbreviation LRLA – for learned Reinforcement Learning algorithm – to refer to this kind of model. Alternatively we can view this procedure as a model-free algorithm for partially observable MDPs, where the hidden information consists of the currently active task.

Resource-Rational Decision-Making

We consider maximizing the following regularized objective for inferring a distribution q_ϕ over parameters θ of LRLAs:

$$\mathcal{L}(\phi, \mathbf{X}, \mathbf{y}) = \mathbb{E}_{q_\phi(\theta)}\left[\log p(\mathbf{y}|\mathbf{X}, \theta)\right] - \beta \mathrm{KL}(q_\phi(\theta)||p(\theta)) \quad (1)$$

where the hyperparameter β controls how much the posterior is allowed to deviate from the prior in terms of the Kullback-Leibler (KL) divergence. We assume a likelihood $p(\mathbf{y}|\mathbf{X}, \theta)$, that factorizes over data points $\prod_{i=1}^{N} p(y_i|X_i, \theta)$ and we approximate each factor with a normal distribution of fixed scale σ_y: $\mathcal{N}(y_t; Q_\theta(X_t, a), \sigma_y)$. In our setting $Q_\theta(X_t, a)$ corresponds to the RNN output after seeing history X_t and y_t corresponds to the n-step return $\sum_{k=0}^{n-1} \gamma^k r_{t+k} + \gamma^n \max_a Q_\theta(X_{t+n}, a)$. The corresponding policy is derived as follows:

$$\pi(a_t|X_t) = \begin{cases} 1 & \text{if } a_t = \arg\max_{a \in \mathcal{A}} Q_\theta(X_t, a) \\ 0 & \text{else} \end{cases} \quad (2)$$

Setting β to a specific value can be interpreted as implicitly defining a constraint on $\mathrm{KL}(q_\phi(\theta)||p(\theta))$. Importantly the KL term determines how much the model parameters can be compressed in theory (Hinton & Van Camp, 1993). Hence our models are resource-constrained with regard to a hypothetical lower bound on their storage capacity. Intuitively, if the regularization factor β is large, parameters are forced to match the prior closely. In this work we employ priors favoring simple functions, hence models are only allowed to realize more complex functions as $\beta \to 0$.

Bayesian Interpretation

If we set $\beta = 1$, we recover the evidence lower bound (ELBO) as an objective for performing variational inference. In the setting of large data-sets subsampling techniques are often employed to approximate Equation 1 using mini-batches $\mathcal{B}$ of size M with an appropriately scaled log-likelihood term:

$$\log p(\mathbf{y}|\mathbf{X}, \theta) \approx \frac{N}{M} \sum_{i \in \mathcal{B}} \log p(y_i|X_i, \theta) \quad (3)$$

If data arrives in sequential fashion, as it does in the RL setting, the data-set size N is not known in advance and has to be treated as an additional hyperparameter. This leads to a Bayesian interpretation of Equation 1 even for $\beta \neq 1$. For any values of β and N maximizing Equation 1 is equivalent to performing stochastic variational inference with an assumed data-set size of $\hat{N} = \frac{N}{\beta}$. In practice we optimize a by N^{-1} scaled version of Equation 1, which leads to $\hat{N}^{-1}$ as a factor for the KL term.

In the following section we investigate whether we can understand individual differences in human decision-making in terms of optimal solutions to Equation 1 for varying values of β. It is worth clarifying, that we are only interested in the computational aspects of this hypothesis, i.e. we want to test, whether human decision-making can be characterized through resource-rational strategies. We do not attempt to answer how this objective is realized on an algorithmic or implementational level.

Technical Details

We maximize Equation 1 using standard gradient-based optimization techniques. For this we simulate k environments in

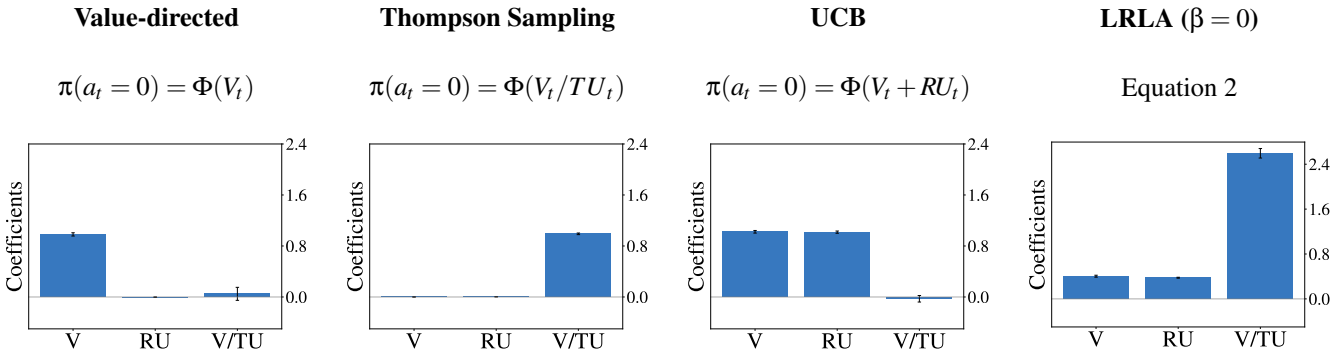


Figure 1: Illustration of different algorithms for two-armed bandits. **Middle**: Definitions of the respective policy. **Bottom**: Coefficients obtained from fitting the probit regression (Equation 5) to corresponding trajectories. Error bars indicate the uncertainty (one standard deviation) in the coefficients estimated through a Laplace approximation. Note, that for LRLAs the coefficients are task-dependent. For this plot we use the set of two-armed bandits described in the later sections to compute the coefficients. Φ denotes the cumulative distribution function of a standard normal distribution.

parallel and update the model at the end of each episode. All models in this work employ a group horseshoe prior, which can be viewed as a continuous relaxation of a spike-and-slab prior (Mitchell & Beauchamp, 1988), over their weights:

$$s \sim \mathcal{C}^+(0, \tau_0); \quad \tilde{z}_i \sim \mathcal{C}^+(0,1);$$
$$\tilde{\theta}_{ij} \sim \mathcal{N}(0,1); \quad \theta_{ij} = \tilde{\theta}_{ij}\tilde{z}_i s$$

and we represent the approximate posterior $q_\phi(\theta)$ through a fully factorized distribution as proposed in (Louizos, Ullrich, & Welling, 2017). The hyperparameter of the horseshoe prior is fixed to $\tau_0 = 10^{-5}$. The horseshoe prior is a sparsity-inducing prior, which causes our models to implement simple functions in absence of any experience. During training we approximate the expectation of the log-likelihood term with a single sample from $q_\phi(\theta)$ and make use of the reparametrization trick (Kingma & Welling, 2013). Resampling of weight matrices is done only at the beginning of an episode as proposed in Gal and Ghahramani (2016); Fortunato, Blundell, and Vinyals (2017). Target values y_t are computed using the maximum a posteriori estimate of a separate target network (Mnih et al., 2013; Lipton et al., 2017). For additional details we refer the reader to the publicly available implementation[1].

Multi-Armed Bandits

Experiments in the following section involve a multi-armed bandit task. These are MDPs consisting of a single state. At each step t an agent selects one out of multiple actions and is rewarded according to an unknown, stationary distribution based on its choice. This interaction is repeated T times.

The trade-off between exploiting good options and exploring yet unknown ones is the central theme in multi-armed bandits (and in RL in general). Methods for resolving this exploration-exploitation dilemma can be categorized in two major groups: directed and random exploration strategies. Directed exploration attempts to gather information about uncertain, but learnable, parts of the environment, while random exploration injects stochasticity of some form into the policy. Gershman (2018) showed, that these two principles can be distinguished exactly under certain conditions. For this we consider a two-armed bandit task with normal distributions over both the mean of rewards for each arm and their reward noise at each time-step. Let $\mathcal{N}(r_a; \mu_{0,a}, \sigma_{0,a})$ be an independent normal prior over expected rewards for each action a and $\mathcal{N}(r_a; \mu_{t,a}, \sigma_{t,a})$ be the posterior after t interactions. Many popular strategies can be formulated using the parameters of these distributions. Define:

$$\begin{aligned} V_t &= \mu_{t,0} - \mu_{t,1} \\ RU_t &= \sigma_{t,0} - \sigma_{t,1} \\ TU_t &= \sqrt{\sigma_{t,0}^2 + \sigma_{t,1}^2} \end{aligned} \tag{4}$$

V_t constitutes the estimated difference in value, while RU_t and TU_t describe relative and total uncertainty respectively. Choice probability in Thompson sampling (an example for random exploration) is only a function of V_t and TU_t, while it is a function of V_t and RU_t for the UCB algorithm (an example of directed exploration). Figure 1 (middle row) shows definitions of all strategies under consideration. For a given set of observed trajectories $\mathcal{D}$ one can fit a probit regression model to infer the importance of factors from Equation 4:

$$p(a_t = 0 | \mathcal{D}, \mathbf{w}) = \Phi(w_1 V_t + w_2 RU_t + w_3 V_t/TU_t) \tag{5}$$

Analyzing the resulting coefficients $\mathbf{w}$ can reveal, which exploration strategy generated the observations, as shown in Figure 1 (bottom row). We utilize this form of analysis throughout the following sections.

Empirical Analysis

Human Participants

We initially inspect human exploration strategies on a two-armed bandit task with episode length $T = 10$. The mean

[1] https://github.com/marcelbinz/MDLDQN

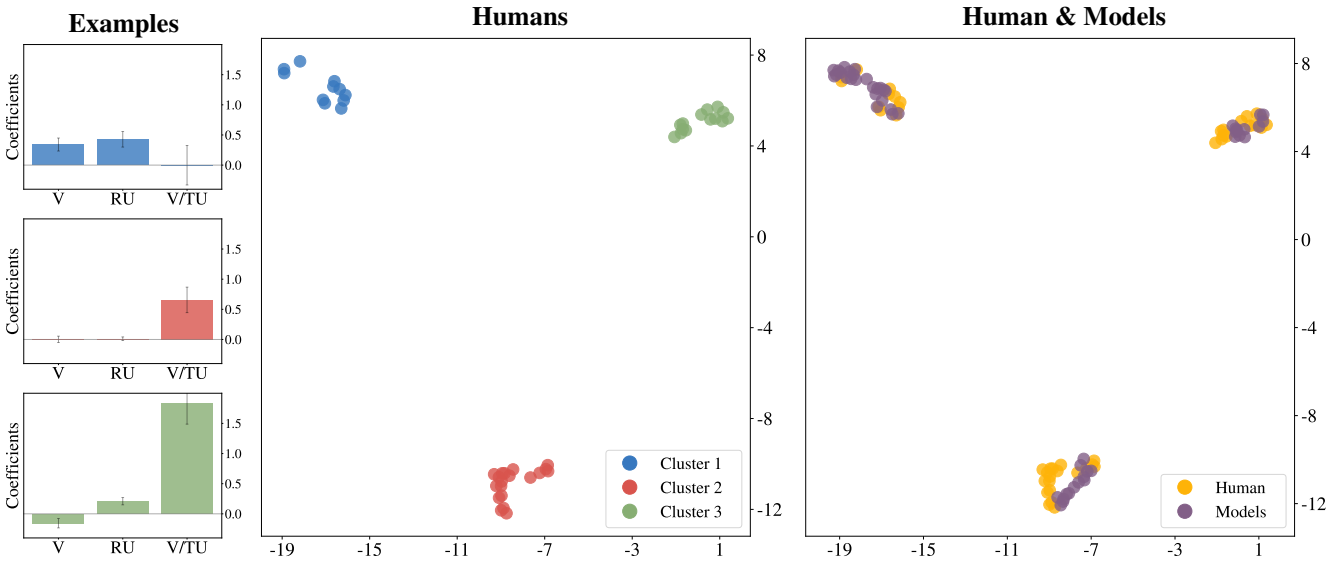


Figure 2: Visualization of human policies alongside resource-constrained LRLAs. **Left**: Probit regression coefficients of prototype participants. Prototypes were obtained from a mean-shift clustering, shown in the middle plot. Colors correspond to clusters. Error bars indicate the uncertainty (one standard deviation) in the coefficients estimated through a Laplace approximation. **Middle**: UMAP (McInnes & Healy, 2018) embedding of coefficients for all participants. **Right**: Joint UMAP embedding of coefficients for human participants and LRLAs $\in \mathcal{H}_{\text{LRLA}}$.

reward for each action is drawn from $\mathcal{N}(\mu_a; 0, \sqrt{100})$ at the beginning of an episode and the reward in each step from $\mathcal{N}(r_t; \mu_{a_t}, \sqrt{10})$. Intuitively we expect some participants to be more proficient at the task, for example because they have more experience at related problems (higher $\hat{N}$), while the opposite is true for others. We rely on data gathered by Gershman (2018), which contains records of 44 participants, each playing 20 of the aforementioned two-armed bandit problems. Figure 2 (middle) shows the result of fitted probit regression coefficients for individual participants. This analysis reveals three major subgroups within the population, each using a different set of strategies. We visualize coefficients of three example participants (Figure 2, left) and observe, that a large fraction is well-described through Thompson sampling (clusters 2 and 3), while other participants have tendencies towards a mixture of strategies (cluster 1).

Learned Reinforcement Learning Algorithms

Next we show, that optimizing LRLAs with different regularization factors leads to the emergence of diverse exploration pattern. We train otherwise identical models for $\hat{N} \in \mathcal{H}_{\text{LRLA}} = \{256, 512, 1024, 2048, 4096, 8192\}$ on the same two-armed bandit task until convergence and report average results over 10 random seeds unless otherwise noted. Equation 1 is approximated with a batch of samples from complete episodes of 16 parallel simulations and gradient-based optimization is performed using Adam (Kingma & Ba, 2014). Figure 3 (left) shows, that performances continuously improves as $\hat{N}$ increases, confirming our expectation that models become more sophisticated for large $\hat{N}$. Fitting the aforementioned probit regression model to the resulting policies (Figure 3, right) reveals value-based characteristics at one end of the spectrum. Towards the other end we observe coefficients, that slowly transition to those of the unconstrained ($\beta = 0$) model.

Modelling Human Behavior

We are mainly interested in whether the set of resource-constrained LRLAs can help us to understand human behavior on an individual level. To answer this question, we compare the optimized models to human decision-making strategies in terms of the probit regression analysis. We visualize the regression coefficients for 50 models (10 for each value of $\hat{N} \in \mathcal{H}_{\text{LRLA}}$, excluding $\hat{N} = 256$) alongside those of the human participants in Figure 2 (right). Although some parts of the low-dimensional embedding are over- and underrepresented, the overall variation of human exploration strategies is captured by the resource-constrained LRLAs.

Model Comparison

The regression analysis performed so far provides only qualitative indicators for our hypothesis. In order to obtain a quantitative measure for the explanatory power of the proposed hypothesis, we performed a Bayesian model comparison. Figure 4 (left) shows log-likelihoods for each participant and model. We observe, that different participants are modelled best with different values of $\hat{N}$.

To verify that the class of resource-constrained LRLAs $\mathcal{H}_{\text{LRLA}}$ contains a good model, we compute Bayes factors (BF) between the marginal probability of the resource-

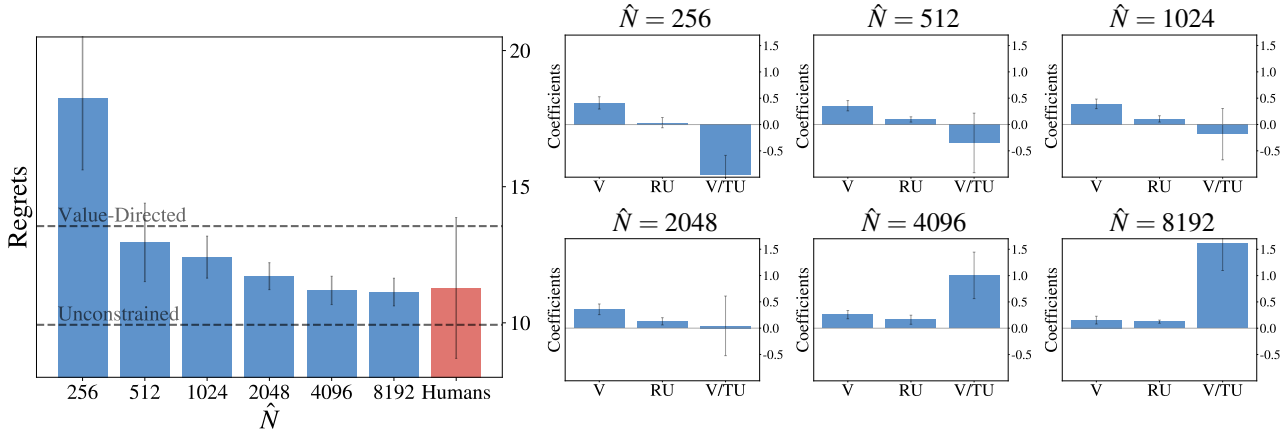


Figure 3: Results for optimized LRLAs with different $\hat{N}$. **Left**: Visualization of per episode regret averaged over 10 models and 1000 episodes. Horizontal lines correspond to the performance of a value-directed policy and an unconstrained LRLA. **Right**: Coefficients of the probit regression from Equation 5. Error bars indicate standard deviations across the 10 models.

constrained LRLAs and a value-directed policy:

$$\log BF_i = \log p(\mathcal{D}_i|\mathcal{H}_{\text{LRLA}}) - \log p(\mathcal{D}_i|H_{\text{value-directed}})$$
$$p(\mathcal{D}_i|\mathcal{H}_{\text{LRLA}}) = \frac{1}{|\mathcal{H}_{\text{LRLA}}|} \sum_{H \in \mathcal{H}_{\text{LRLA}}} p(\mathcal{D}_i|H) \tag{6}$$

where $\mathcal{D}_i$ refers to all actions taken by a specific participant and $\frac{1}{|\mathcal{H}_{\text{LRLA}}|}$ is a prior that corrects for multiple comparisons across different values of $\hat{N}$. The resulting $\log BF$s (see Figure 4, right) reveal strong evidence for 42 of the 44 participants in favor of the class of resource-constrained LRLAs, when compared with the baseline. This indicates, that one of the models in $\mathcal{H}_{\text{LRLA}}$ explains the participant's behavior much better than the value-directed policy. There are nine participants best described by letting $\hat{N} = 512$, nine by $\hat{N} = 1024$, 20 by $\hat{N} = 4096$ and six by $\hat{N} = 8192$. This heterogeneity highlights, that the model class is able to accomodate individual differences between human participants.

Finally we want to show, that the proposed class of models captures exploration strategies across all participants better than any standard exploration strategy alone. To verify this, we computed Bayes factors between $\prod_i p(\mathcal{D}_i|\mathcal{H}_{\text{LRLA}})$ and two baseline exploration strategies: $\prod_i p(\mathcal{D}_i|H_{\text{Thompson}})$ and $\prod_i p(\mathcal{D}_i|H_{\text{UCB}})$. We find $2\log BF = 72.8$ against Thompson sampling and 5391.4 against UCB, indicating that our class of models is overall better at representing exploration strategies for all participants in comparison to any single, fixed strategy.

Discussion

In this work we proposed a justification for seemingly sub-optimal human strategies in sequential decision-making problems based on the idea of computational rationality. We view human decision-making as an instance of a learned, resource-constrained RL algorithm. This is formalized through learning distributions over parameters of a meta-learning model with a regularized, resource-rational objective. The emerging spectrum of strategies resembles characteristics of human decision-making without being explicitly trained to do so. Additional model comparison suggests, that the resulting resource-constrained LRLAs describe human policies well on a quantitative level. However, the correspondence between human behavior and the LRLA model class is not perfect. Looking at Figure 2 (right) we observe, that some clusters are not represented exactly. Furthermore it remains open, why none of the participants is best described through the model with $\hat{N} = 2048$. Accounting for these observations remains a question for future work.

The analysis on the two-armed bandit task presented in this work can be extended in several ways. Relating deliberation times to regularization factors could, for example, provide additional evidence for our hypothesis. It also remains to be seen whether our conclusions transfer to other sequential decision-making problems beyond the bandit setting. In this context we are especially interested in tasks, where descriptive models of individual human behavior consist of a set of different heuristics. We are also interested in methods, that allow us to disentangle resource-rational behavior from the Bayesian interpretation.

Recent work on model-free meta-learning methods, similar to the one employed in this work, indicates an emergence of model-based behavior (Wang et al., 2016) and causal reasoning (Dasgupta et al., 2019) as well as the ability for few-shot learning (Santoro, Bartunov, Botvinick, Wierstra, & Lillicrap, 2016), properties supposedly absent in artificial systems. Having systems capable of such feats, opens the possibility for interesting studies on human cognition.

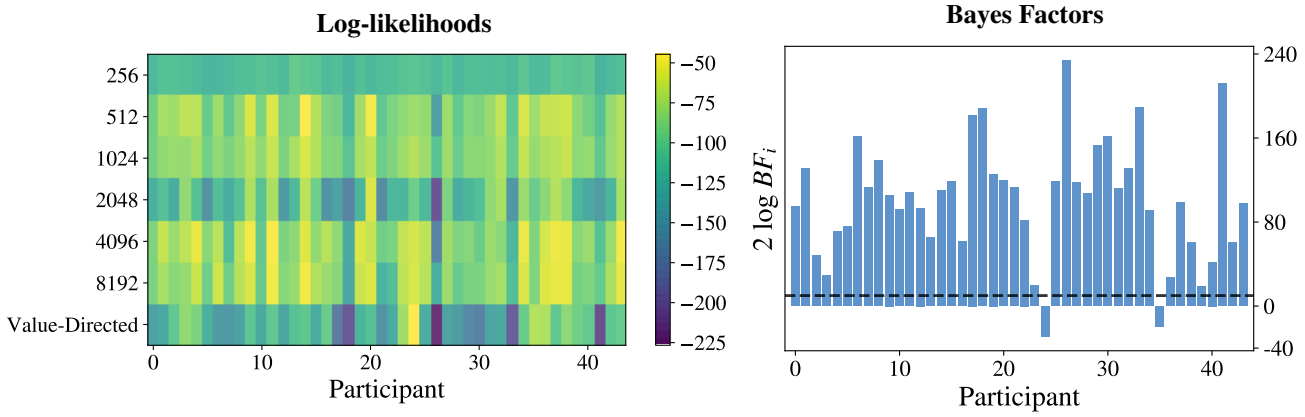


Figure 4: Model comparison of the set of resource-constrained LRLAs with a value-directed baseline. **Left**: Log-likelihoods for each participant and model. Higher values indicate a better fit. **Right**: Bayes factors (see Equation 6) for each participant i. The dotted horizontal line (equal to 10) corresponds to the threshold for very strong evidence (Kass & Raftery, 1995) in favour of $\mathcal{H}_{\text{LRLA}}$.

Acknowledgments

This work was supported by the DFG GRK-RTG 2271 'Breaking Expectations'.

References

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Dasgupta, I., Wang, J., Chiappa, S., Mitrovic, J., Ortega, P., Raposo, D., ... Kurth-Nelson, Z. (2019). *Causal reasoning from meta-reinforcement learning.* Retrieved from `https://openreview.net/forum?id=H1ltQ3R9KQ`

Duan, Y., Schulman, J., Chen, X., Bartlett, P. L., Sutskever, I., & Abbeel, P. (2016). Rl ^{2}: Fast reinforcement learning via slow reinforcement learning. *arXiv preprint arXiv:1611.02779*.

Duff, M. O., & Barto, A. (2002). *Optimal learning: Computational procedures for bayes-adaptive markov decision processes*. Unpublished doctoral dissertation, University of Massachusetts at Amherst.

Fortunato, M., Blundell, C., & Vinyals, O. (2017). Bayesian recurrent neural networks. *arXiv preprint arXiv:1704.02798*.

Gal, Y., & Ghahramani, Z. (2016). A theoretically grounded application of dropout in recurrent neural networks. In *Advances in neural information processing systems* (pp. 1019–1027).

Gershman, S. J. (2018). Deconstructing the human algorithms for exploration. *Cognition*, *173*, 34–42.

Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, *349*(6245), 273–278.

Gigerenzer, G., & Todd, P. M. (1999). *Simple heuristics that make us smart*. Evolution and Cognition (Paper.

Gittins, J. C. (1979). Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society. Series B (Methodological)*, 148–177.

Griffiths, T. L., Kemp, C., & Tenenbaum, J. B. (2008). Bayesian models of cognition.

Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in cognitive science*, *7*(2), 217–229.

Grunwald, P. (2004). A tutorial introduction to the minimum description length principle. *arXiv preprint math/0406077*.

Hinton, G. E., & Van Camp, D. (1993). Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the sixth annual conference on computational learning theory* (pp. 5–13).

Honkela, A., & Valpola, H. (2004). Variational learning and bits-back coding: an information-theoretic view to bayesian learning. *IEEE Transactions on Neural Networks*, *15*(4), 800–810.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the american statistical association*, *90*(430), 773–795.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Lipton, Z., Li, X., Gao, J., Li, L., Ahmed, F., & Deng, L. (2017). Bbq-networks: Efficient exploration in deep reinforcement learning for task-oriented dialogue systems. *arXiv preprint arXiv:1711.05715*.

Louizos, C., Ullrich, K., & Welling, M. (2017). Bayesian compression for deep learning. In *Advances in neural information processing systems* (pp. 3288–3298).

McInnes, L., & Healy, J. (2018, February). UMAP: Uniform Manifold Approximation and Projection for Dimen-

sion Reduction. *ArXiv e-prints*.

Mitchell, T. J., & Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, *83*(404), 1023–1032.

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., ... Kavukcuoglu, K. (2016). Asynchronous methods for deep reinforcement learning. In *International conference on machine learning* (pp. 1928–1937).

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., & Riedmiller, M. (2013). Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*.

Ortega, P. A., & Braun, D. A. (2013). Thermodynamics as a theory of decision-making with information-processing costs. *Proc. R. Soc. A*, *469*(2153), 20120683.

Parpart, P., Jones, M., & Love, B. C. (2018). Heuristics as bayesian inference under extreme priors. *Cognitive psychology*, *102*, 127–144.

Santoro, A., Bartunov, S., Botvinick, M., Wierstra, D., & Lillicrap, T. (2016). Meta-learning with memory-augmented neural networks. In *International conference on machine learning* (pp. 1842–1850).

Siegelmann, H. T., & Sontag, E. D. (1991). Turing computability with neural nets. *Applied Mathematics Letters*, *4*(6), 77–80.

Simon, H. A. (1972). Theories of bounded rationality. *Decision and organization*, *1*(1), 161–176.

Simon, H. A. (1990). Invariants of human behavior. *Annual review of psychology*, *41*(1), 1–20.

Steyvers, M., Lee, M. D., & Wagenmakers, E.-J. (2009). A bayesian analysis of human decision-making on bandit problems. *Journal of Mathematical Psychology*, *53*(3), 168–179.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *science*, *185*(4157), 1124–1131.

Wang, J. X., Kurth-Nelson, Z., Tirumala, D., Soyer, H., Leibo, J. Z., Munos, R., ... Botvinick, M. (2016). Learning to reinforcement learn. *arXiv preprint arXiv:1611.05763*.

Zhang, S., & Angela, J. Y. (2013). Forgetful bayes and myopic planning: Human learning and decision-making in a bandit setting. In *Advances in neural information processing systems* (pp. 2607–2615).

Predicting Learned Inattention from Attentional Selectivity and Optimization

Abstract

Although selective attention is useful in many situations, it also has costs. In addition to ignoring information that may become useful later, it can have long term costs, such as learned inattention – difficulty in learning from formerly irrelevant sources of information in novel situations. In the current study we tracked participants' gaze while they completed a category learning task designed to elicit learned inattention. During learning an unannounced shift occurred such that information that was most relevant became irrelevant, whereas formerly irrelevant information became relevant. We assessed looking patterns during initial learning to understand how different aspects of attention allocation contribute to learned inattention. Our results indicate that learned inattention depends on both the overall level of selectivity (measured as entropy of proportion of looking to each feature) and the extent to which participants optimized attention (becoming more selective over time).

Keywords: selective attention; categorization; learning; attention

Introduction

Category learning is a critical cognitive process that enables abstract thought and allows for generalization of knowledge to novel situations. Since the early work by Shepard, Hovland, and Jenkings (1961) selective attention has been considered a critical component of categorization and category learning. Selective attention refers to the ability to prioritize task-relevant information and filter out task irrelevant information (Desimone & Duncan, 1995; Hanania & Smith, 2010; Plude, Enns, & Brodeur, 1994; Pashler, Johnston, & Ruthruff, 2001; Yantis, 2000).

Most models of categorization and category learning adopt the Shepard et al. (1961) view and consider selective attention a critical contributor to human categorization. Exemplar models (Hampton, 1995; Medin & Schaffer, 1978; Nosofsky, 1986), prototype models (Smith & Minda, 1998), clustering models (Love, Medin, & Gureckis, 2004), and dual process models (Ashby, Alfonso-Reese, Turken, & Waldron, 1998) all include some form of selective attention as a factor determining the influence (or weight) of stimulus dimensions on categorization. According to some of these models, as participants learn categories, they tend to shift attention to features that are more likely to predict category membership, while attending less to less predictive features, the idea known as attention optimization. This idea has been confirmed empirically: there is eye-tracking evidence indicating that when learning categories, people indeed tend to optimize their attention over time, increasing fixating at the features most predictive of a category and decreasing fixating at irrelevant features (Rehder & Hoffman, 2005; Blair, Watson & Meier, 2009).

Costs and Benefits of Selectivity

Although selective attention is often beneficial in many learning scenarios (e.g. faster, efficient processing of attended information), selective attention also has costs (Best, Yim, & Sloutsky, 2013; Hoffman & Rehder, 2010; Plebanek & Sloutsky, 2017; Rich & Gureckis, 2018). Some of the costs are relatively short-term: people miss non-selected information. Other costs are longer term in that they affect future learning. One such short-term cost is that non-selected information is filtered out. Focusing attention is a tradeoff in that it results in efficient learning and performance, but it also results in missing information that could be used later.

While short-term costs of selectivity affect only the task at hand, longer term costs also affect performance on future tasks. One type of long-term cost has been recently discussed by Rich & Gureckis (2018), who referred to it as a "learning trap". The authors demonstrate that under certain circumstances selective attention can be a trap in that it can result in getting stuck on inaccurate representations of the to-be-learned structure and preventing the exploration needed to discover the correct structure. This happens particularly when there are possible negative outcomes to exploring, in which case selective attention results in overgeneralizing which things should be avoided.

A more general long-term cost is that selective attention may result in learned inattention (see Hoffman & Rehder, 2010, for a review) – difficulty in learning from sources of information that were uninformative in a previous situation. Optimizing one's attention to the currently most relevant sources of information can result in not only learning to ignore currently irrelevant sources of information, but continuing to ignore these sources in novel situations in the future (Kruschke & Blair, 2000). As a result, if those sources of information become relevant later, learning is more difficult than it would have been if one had not first learned to ignore them. Learned inattention can be detrimental when task demands change, or when a new classification contrast depends on previously irrelevant features.

For example, Hoffman & Rehder (2010) had participants learn to classify stimuli consisting of three dimensions. Learning occurred in either a classification condition or a feature inference condition. In classification, participants predict the category label from all of the stimulus' features. In inference they predict one missing feature from the label and the remaining features.

In the first phase of the experiment only dimension 1 was relevant to distinguish two categories, while the other two were irrelevant. In a second phase, only dimension 2 was relevant for distinguishing two new categories, while in a third phase dimension 3 was relevant for a novel contrast between categories. Classification encourages selective

attention since only a single feature predicts category membership. Inference, on the other, encourages distributed attention, since participants may need to predict any of the three features.

Participants performing the classification task were impaired at learning the new contrast when the relevant dimension changed compared to participants doing the feature inference task. Additionally, eye-tracking showed that the classification participants were much less likely to fixate the relevant dimension after learning it was irrelevant in a prior phase of the experiment compared to baseline levels at the start of the experiment. These costs occurred because learners selectively attended the relevant feature while classifying the stimuli—optimizing their attention to ignore (or inhibit) the other features.

The Current Study

The goal of this study is to further investigate learned inattention during category learning by examining how different aspects of attention allocation contribute to learned inattention. This study also serves as a first step toward a larger investigation of developmental differences in attention allocation and optimization and their consequences. The current study investigates only adults, but developmental implications and predictions for children are touched on in the Discussion.

In our experiment, we presented participants with a category learning task while tracking their gaze. The to-be-learned categories had a rule-plus-similarity structure, with one deterministic feature perfectly predicting category membership and multiple probabilistic features, providing good, but imperfect prediction (see Deng & Sloutsky, 2016, for a similar structure). In addition, one feature was completely irrelevant to categorization.

Given the structure, participants could either form a rule-based representation (by relying on the deterministic feature) or a similarity-based representation (by relying on all features). So, either selective or distributed attention could lead to effective learning. Once participants mastered the categories in this initial phase, there was an unannounced shift in the category learning task. After the shift, feature dimensions that had been deterministically predictive in became irrelevant, and features that had been irrelevant in became deterministically predictive.

Learned inattention was expected to produce costs on learning in Phase 2, making participants less likely to attend to and learn to use the new deterministic feature. We examined what aspects of attention during initial learning were most likely to manifest these costs. In particular we investigated the effects of overall selectivity and of attention optimization (increase in selectivity over time).

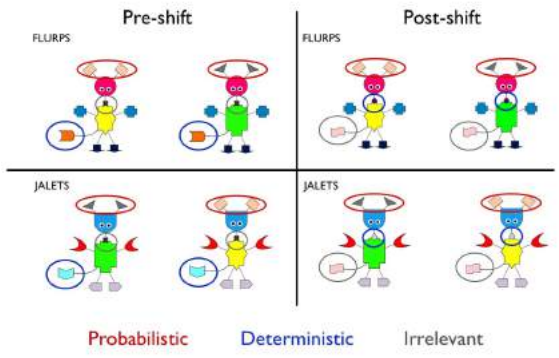


Figure 1. Examples of stimuli. The stimuli were creatures composed of seven discrete-valued dimensions differing in shape and color. One Deterministic feature perfectly predicted category membership (the tail in this example). One Irrelevant feature was the same across both categories (the button on the neck here). Five Probabilistic features predicted category membership with 80% accuracy. After an unexpected shift occurred, the Deterministic became Irrelevant, and the Irrelevant feature became Deterministic Probabilistic features were unchanged by the shift.

Method

Participants

A total of 38 adults (26 women) participated in the experiment. Participants were undergraduate students participating for course credit.

Materials and Design

Stimuli were colorful images of creatures composed of seven discrete-valued dimensions (see Figure 1). The creatures were divided into two categories, referred to as Flurps and Jalets. Of the seven features (antenna, head, body, button, hands, feet, and tail), one feature deterministically predicted category membership (henceforth the Deterministic feature), five features were probabilistically predictive with 80% accuracy (the Probabilistic features), and one feature was non-predictive—having the same value across all exemplars of both categories and therefore irrelevant to classification (i.e., the Irrelevant feature). Table 1 shows the stimulus structure used in the task.

Stimuli were organized into pairs of complementary sets. Each set in a pair was identical to its counterpart except that the Deterministic and Irrelevant features swapped roles. Probabilistic feature values and the category labels remained the same. As discussed below, participants learned one set in Phase 1 of the experiment (i.e., before the shift), and then the stimuli were unexpectedly replaced with the complimentary set for Phase 2 (i.e., after the shift). There were two pairs: one where feet and hands were the Deterministic/Irrelevant features, one where tail and neck button were the Deterministic/Irrelevant features. Which pair was presented and which set in that pair was learned in Phase 1 were counterbalanced between participants.

Stimulus sets also contained Ambiguous items. These were hybrid items having the Probabilistic features from one category and the Deterministic feature from the other category. These items were presented only during the testing sessions and were designed to test which features controlled categorization. There were 10 Ambiguous items total per set—one corresponding to each exemplar seen during training (i.e. with identical Probabilistic features but with the Deterministic feature from the opposite category). These items allow us to determine whether participants' category judgments were based more on the single Deterministic feature or on one or more of the Probabilistic features.

Table 1. Stimuli structure during training

	Pre-shift							Post-shift							
Feature	D/I	I/D	P1	P2	P3	P4	P5		D/I	I/D	P1	P2	P3	P4	P5
Category A															
A1	1	2	0	1	1	1	1	A6	2	1	0	1	1	1	1
A2	1	2	1	0	1	1	1	A7	2	1	1	0	1	1	1
A3	1	2	1	1	0	1	1	A8	2	1	1	1	0	1	1
A4	1	2	1	1	1	0	1	A9	2	1	1	1	1	0	1
A5	1	2	1	1	1	1	0	A10	2	1	1	1	1	1	0
Category B															
B1	0	2	1	0	0	0	0	B6	2	0	1	0	0	0	0
B2	0	2	0	1	0	0	0	B7	2	0	0	1	0	0	0
B3	0	2	0	0	1	0	0	B8	2	0	0	0	1	0	0
B4	0	2	0	0	0	1	0	B9	2	0	0	0	0	1	0
B5	0	2	0	0	0	0	1	B10	2	0	0	0	0	0	1

Note: D/I is the feature that is Deterministic prior to the shift and Irrelevant after the shift. I/D is the feature that is Irrelevant pre-shift and Deterministic post-shift. P1-P5 are the probabilistic features.

Procedure

Adult participants conducted a classification procedure while their gaze was monitored with an EyeLink 1000 hydraulic-arm eyetracker at 500Hz (SR research, Ontario, Canada). The experiment was divided into two phases. Both phases contained a training section (with feedback) followed by a testing section (no feedback). In Phase 1 participants learned to classify two categories of creatures (and then were tested), and then in Phase 2 an unannounced shift occurred wherein the previously Deterministic feature and the previously Irrelevant feature swapped roles.

The formerly Deterministic dimension took on a new, previously unseen, value that was fixed across all stimuli of both categories, while the formerly Irrelevant dimension now had two new potential values that perfectly predicted category membership. Participants were given no warning that this shift would occur at any point. Like Phase 1, Phase 2 consisted of training followed by testing.

At the beginning of the experiment participants were given information about the Deterministic and Probabilistic features. For Probabilistic features they were told that most of the members of the category had that particular feature value. For the Deterministic feature they were told that all members of category A have one value while those of category B have another value, while being shown both. The Irrelevant feature was never mentioned in the instructions. These informative instructions were included to ensure good learning in Phase 1, since any expected costs due to the shift rely on the categories initially being learned well. Additionally, as noted above we plan to eventually investigate developmental differences, and young children require this type of informative instructions and feedback in order to learn well within the timeframe of the experiment.

Training Training in each phase consisted of 30 trials (in 3 blocks of 10 trials). In each block of 10 trials the ten training exemplars, five from each category, were presented in random order, so participants saw each exemplar three times throughout training (see Table 1 and Figure 1).

On each training trial, one stimulus was presented in the middle of the screen and participants indicated which category they thought it belonged to. Corrective feedback was then given which tried to equally encourage attention to general appearance (similarity-based responding) and to the Deterministic feature (rule-based responding). For example feedback would be "Correct this is a Flurp. It looks like a Flurp and has the Flurp hands.", or "Oops this is actually a Jalet. It looks like a Jalet and has the Jalet hands.", in the case where hands were the Deterministic feature.

In Phase 2, after the unannounced shift, feedback was simplified to mention only the correct category without drawing attention to the features (e.g. "Correct this is a Flurp."). While this change in feedback may have given participants some indication that a change had occurred, it was necessary so that participants would need to figure out on their own the new contingencies between features and categories. That learning process in Phase 2, discovering what is informative, is the critical area of interest, while parity between Phase 1 and 2 is not critical.

Testing Testing in each phase consisted of 20 trials. Again, participants saw the stimuli one at a time and classified each one, but no feedback was provided during the test. The 10 items seen during training (henceforth the High Match items) and 10 Ambiguous items were each presented once, in random order. These two types of items, respectively, were the basis for the behavioral measures of general shift costs and shift costs due to learned inattention.

Accuracy on High Match items indicates how well each participant learned during the immediately prior training session. A decrease in accuracy from Phase 1 to Phase 2 on these items would indicate a general cost (in terms of poorer learning) due to the unexpected shift, which may occur for a number of reasons.

Responses to the Ambiguous items, in contrast, provide the cornerstone of our behavioral analyses related to learned inattention. Prior to the shift they provide the baseline level that each participant tended to categorize based on the single Deterministic feature. After the shift responses to the Ambiguous items tell us whether participants learned and used the rule on the new Deterministic (formerly Irrelevant) feature. Low deterministic responses indicate learned inattention since it suggests that the participant had difficulty finding the new rule on the feature that was previously irrelevant.

Results

Behavioral Results

Initial learning in Phase 1 was good overall (see Figure 2). Mean accuracy was 92.1% correct. A repeated measures ANOVA found a main effect of block [F(1,75) =. 14.74, p = 0.0003, partial η^2 = 0.164] suggesting that participants learned well during training. Categorization accuracy on High Match items during the test was also high (M = 93.9% correct).

Responses to the Ambiguous items provide insight into which features controlled participants' categorization: Responding based on the Deterministic feature points to rule-based categorization, whereas responding based on the Probabilistic features suggests similarity-based (or at least non-rule-based) categorization. Distributing attention during training could lead to either type of responding (since all features were attended), but selective attention to the Deterministic feature should only result in rule-based categorization. Participants were overwhelmingly deterministic in their categorization of the Ambiguous items: 88.4% deterministic responses, which was well above chance, $t(37) = 9.69$, $p < 0.001$, $d = 1.57$.

Post-shift Learning Participants learned well in Phase 2 after the shift (77.0% correct), which was above chance, $t(37) = 19.09$, $p < 0.001$, $d = 3.10$, but accuracy was lower than prior to the shift, $t(37) = 6.91$, $p < 0.001$, $d = 1.12$, which is expected—even in the absence of learned inattention—due to less informative feedback in Phase 2 compared to Phase 1 and any general costs of adapting to the shift. A repeated measures ANOVA on Phase 2 training accuracy found a main effect of block [F(1,75) = 7.192, p = 0.009, partial η^2 = 0.087], suggesting accuracy increased over time. Accuracy on High Match items during the test was significantly higher than chance (M = 81.3% correct), $t(37) = 10.06$, $p < 0.001$, d = 1.63, but was also lower than in Phase 1, $t(37) = 3.52$, p = 0.001, d = 0.571, suggesting a learning cost due to the unexpected shift. This represents a general shift cost which may be due to a variety of factors including learned inattention. We assess costs specific to learned inattention in the next section.

Shift Costs Due to Learned Inattention We assessed effects of learned inattention by examining responses to the Ambiguous items in the Phase 2 test (Figure 3). Learned inattention would result in relatively low levels of deterministic responses on the post-shift test, since participants would be less likely to attend to the previously Irrelevant (now Deterministic) feature, and thus would have difficulty learning the new rule on that feature. This would result in participants relying primarily on the Probabilistic features instead when making category judgments after the shift. Participants were significantly below chance, M =36.6%, $t(37) = 2.53$, $p = 0.016$, $d = 0.41$, suggesting that they primarily relied on Probabilistic features to categorize after the shift, in contrast to their behavior prior to the shift.

Figure 3 shows individual participants proportion of classifying Ambiguous items based on the Deterministic feature in the post-shift test.

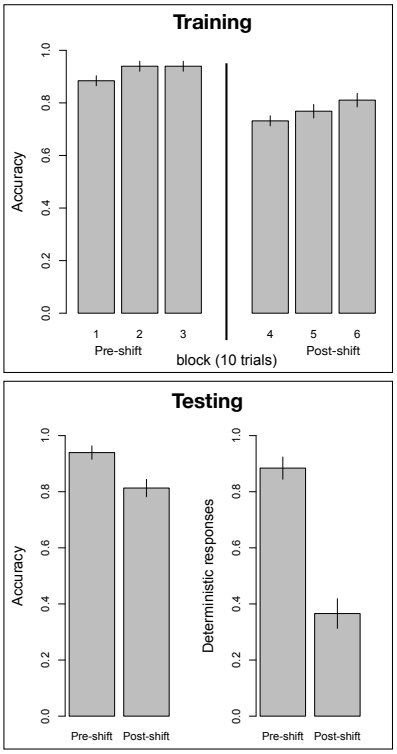


Figure 2. Behavioral results. During initial training participants learned well and achieved high accuracy. A substantial drop occurred after the shift. Accuracy during test was high for old (High Match) items, and was lower after the shift, displaying a general cost of the shift. Responding based on the Deterministic feature was very high prior to the shift, but dropped substantially in the post-shift test, suggesting effects of learned inattention. Error bars represent standard error of the mean.

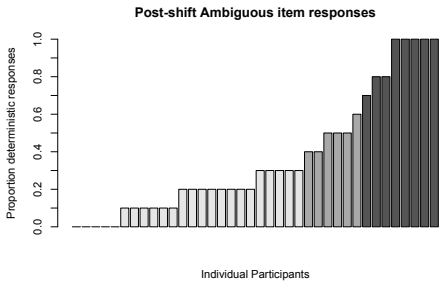


Figure 3. Individual participants' post-shift deterministic responding. Responses to Ambiguous items on the post-shift test varied widely between individuals. The majority classified based on the Probabilistic features, with some intermediate, and less than one-quarter of participants classifying based on the new Deterministic feature.

Eye-Tracking

Regions of interest isolating each feature were used to calculate the proportion of each trial spent looking at each feature for each subject. Timepoints where participants were not looking at any of the features were removed. Figure 4 shows the average proportion of looking at each feature type during training. Prior to the shift the proportion of the trial spent looking at the Deterministic feature increased over trials, while time spent looking at Probabilistic features decreased. Looking at the Irrelevant feature was extremely low throughout all of training (M = 2.82% of total looking time), but did decrease over the course of training—as measured by comparing block 1 (M = 3.60%) to block 3 (M = 1.52%), $t(37) = 3.39$, $p = 0.002$, $d = 0.161$.

After the shift, looking at the previously Deterministic (now Irrelevant) feature dropped off rapidly, while looking to the Probabilistic features shot up. Critically, looking at the newly Deterministic feature (that was previously Irrelevant) did not increase from block 1 to block 3, $t(37) = 0.53$, p = 0.600, demonstrating the effects of learned inattention. They continued to ignore this feature despite the high level of information it now contained.

We assessed attentional patterns during initial learning by calculating the entropy (Shannon, 1948) for each trial. Entropy was defined as,

$$S = -\sum_{i=1}^{n} p_i * \log(p_i)$$

where p_i is the proportion of the trial spent looking at feature i. Higher entropy indicates more distributed attention, where maximum entropy is produced by looking at all seven features equally, and lower entropy indicates more selective attention—focusing on a smaller number of features. We use entropy as a measure of selectivity rather proportion of looking at the Deterministic feature since some participants may have optimized their attention to one of the Probabilistic features, and that selectivity should still produce learned inattention despite being suboptimal. Entropy for each trial was normalized by dividing by maximum possible entropy, such that all values were between 0 and 1.

Drop in entropy was calculated as average entropy per trial in block 1 minus average entropy in block 3. This served as a measure of attention optimization, since greater drops indicate an increase in selectivity of attention.

We performed a logistic regression predicting classification of Ambiguous items on the post-shift test from average entropy during pre-shift training, the drop in entropy over training, and their interaction. This analysis revealed a significant interaction, $z = 3.318$, $p = 0.001$. To better understand the interaction, we divided participants into low and high entropy groups based on a median split. We then performed a logistic regression on Ambiguous item responses predicted from the drop in entropy for each group. For the low entropy group, there was a significant negative relationship between entropy drop and Ambiguous item responses, $z = -3.153$, $p = 0.002$, indicating that those who optimized their attention in the initial training were less likely to use the new Deterministic feature to categorize items after the shift (see Figure 5). In contrast, participants in the high entropy group did not show a relationship between the drop in entropy and responses to Ambiguous items, $z = 1.899$, p = 0.0575.

In other words, attention optimization was associated with greater learned inattention in the low entropy (selective attention) group, but not in the high entropy (distributed attention) group. This interesting interaction has important implications. It implies, not surprisingly, that a certain level of selectivity is necessary to produce learned inattention. But more importantly, this high level of selectivity is not enough. Learned inattention seems to also require that attention allocation be incrementally learned over time. This suggests that it occurs when people initially consider multiple features, but learn through experience to ignore them (in contrast to a top-down strategy implemented from the beginning to focus on one or few features).

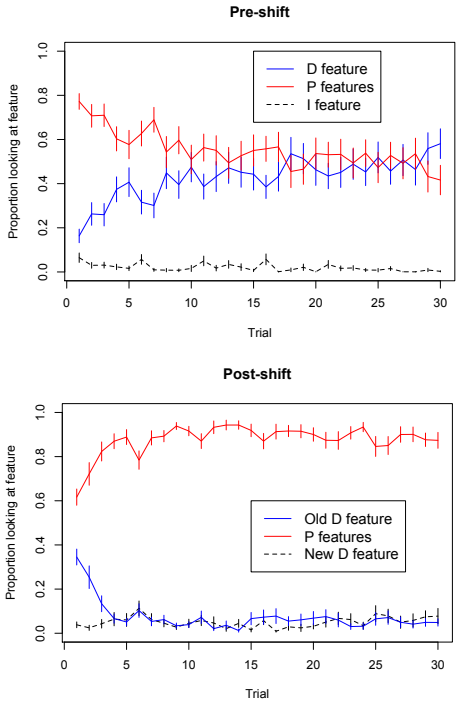


Figure 4. Proportion looking at each feature type during training. Note that the proportion for Probabilistic features is summed across all five Probabilistic features. During initial training looking to the Deterministic feature increased, while looking to the other features decreased (i.e. attention optimization occurred). Post-shift looking to the previously Deterministic feature quickly dropped and was replaced by increased looking to the Probabilistic features, while the previously Irrelevant (now Deterministic feature) remained low. Error bars represent standard error of the mean.

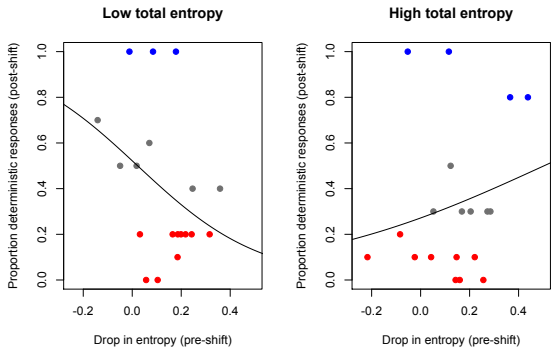


Figure 5. The role of selectivity and attention optimization. Overall entropy (how distributed attention was) interacted with the change in entropy (attention optimization) to produce costs consistent with learned inattention. Participants who were both highly selective (low entropy) and who optimized (large drop in entropy) were the least likely to categorize based on the new deterministic feature after the shift. Line shown the fit line of logistic regression. Dot color indicates deterministic responders (blue), probabilistic responders (red), intermediate (grey).

Discussion

Although selective attention is often effective and efficient, there are potential costs. One particular (longer-term) cost is that selective attention can result in learned inattention to non-selected information, which in turn affects future learning. In the current study, participants performed a category learning task designed to induce learned inattention while we tracked their gaze. Both behavioral and eye-tracking measures showed evidence of learned inattention. When making category judgments after the unexpected shift, participants were less likely to use a stimulus dimension that was previously irrelevant, but was now highly informative. Eye-tracking showed that after the shift occurred, participants quickly shifted attention away from the previously informative (and now irrelevant) feature. Their attention instead shifted to probabilistically predictive features, but they continued to ignore the previously irrelevant, now perfectly predictive, feature—with looking to that feature remaining low and not increasing over the course post-shift training. Participants simply ignored this feature despite its potential usefulness in their task.

The level of learned inattention that was exhibited varied across individuals, though. We used measures of selectivity and attention optimization for each individual to determine what aspects of initial learning best predicted the level of learned inattention that occurred. Our results suggest an interaction between these two measures, such that learned inattention was most likely for participants who were overall highly selective, but importantly, who optimized their attention over time. Participants who were highly selective from the beginning, and so did not optimize attention over time, did not show high levels of learned inattention (see Figure 5). These results suggest that learned inattention crucially depends on incremental learning over time, and is not simply an effect of ignoring sources of information, but of *learning* to inhibit them after initially considering them.

Participants who did optimize attention over time, but whose attention was overall relatively distributed (having high entropy), also did not show high levels of learned inattention. One possibility is that these participants needed more time to optimize attention before reaching the level required for substantial learned inattention to occur, and that with more training trials they would reach that level.

That both the level of selectivity and attention optimization predict individual differences in learned attention has important implications for cognitive development. Young children tend to distribute their attention broadly and do not optimize attention as much as adults do (Best, Yim, & Sloutsky, 2013; Deng & Sloutsky, 2016), so they may be largely protected against the adverse effects of learned inattention.

Allocating attention is always a tradeoff: selective attention results in more efficient processing of attended information, but has several potential pitfalls, including learned inattention. In contrast, if attention is distributed, processing is less efficient, but these traps are avoided. Therefore, to allocate attention effectively estimations must be made about information's potential future relevance. With less general knowledge, children have less basis to make solid conclusions about what might and might not be useful to know later. Additionally, these types of costs could be particularly damaging early in the learning process, and so perhaps children's tendency to distribute attention may be not only a result of immature control, but also adaptive for their particular situation. Understanding the developmental differences in this process is an important direction for future research.

Acknowledgements

Acknowledgments to be completed after review to facilitate double-blind review process.

References

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological model of multiple systems in category learning. *Psychological Review, 105*, 442–481.

Best, C. A., Yim, H., & Sloutsky, V. M. (2013). The cost of selective attention in category learning: Developmental differences between adults and infants. *Journal of Experimental Child Psychology, 116*(2), 105-119.

Blair, M. R., Watson, M. R., & Meier, K. M. (2009). Errors, efficiency, and the interplay between attention and category learning. *Cognition*, *112*(2), 330-336.

Deng, W. S., & Sloutsky, V. M. (2016). Selective attention, diffused attention, and the development of categorization. *Cognitive Psychology*, *91*, 24-62.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, *18*, 193-222.

Hampton, J. A. (1995). Testing the prototype theory of concepts. *Journal of Memory and Language*, *34*(5), 686-708.

Hanania, R., & Smith, L. B. (2010). Selective attention and attention switching: towards a unified developmental approach. *Developmental Science, 13*(4), 622–635.

Hoffman, A. B., & Rehder, B. (2010). The costs of supervised classification: The effect of learning task on conceptual flexibility. *Journal of Experimental Psychology: General*, *139*, 319–340.

Kruschke, J. K., & Blair, N. J. (2000). Blocking and backward blocking involve learned inattention. *Psychonomic Bulletin & Review*, *7*(4), 636-645.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review, 111*, 309–332.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological review*, *85*(3), 207.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57.

Pashler, H., Johnston, J. C., & Ruthruff, E. (2001). Attention and performance. *Annual Review of Psychology*, *52*(1), 629-651.

Plebanek, D. J., & Sloutsky, V. M. (2017). Costs of selective attention: when children notice what adults miss. *Psychological Science*, *28*(6), 723-732.

Plude, D. J., Enns, J. T., & Brodeur, D. (1994). The development of selective attention: a life-span overview. *Acta Psychologica*, *86*(2-3), 227-272.

Rehder, B., & Hoffman, A. B. (2005). Eyetracking and selective attention in category learning. *Cognitive Psychology*, *51*(1), 1-41.

Rich, A., & Gureckis, T. (2018). The limits of learning: Exploration, generalization, and the development of learning traps. *Journal of Experimental Psychology: General, 147*(11), 1553-1570.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell system technical journal*, 27(3), 379-423.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs: General and Applied*, *75*(13), 1-41.

Smith, J. D., & Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*, 1411–1430.

Yantis, S. (2000). Goal-directed and stimulus-driven determinants of attentional control. *Attention and Performance*, *18*, 73–103.

Translation Tolerance in Vision

Anonymous CogSci submission

Abstract

A fundamental challenge in object recognition is to recognize an image when it is projected across different retinal locations, an ability known as translation tolerance. Although the human visual system can overcome this challenge, the mechanisms responsible remain largely unexplained. The 'trained-tolerance' approach holds that an object must be experienced across different retinal locations to achieve translation tolerance. Previous studies have supported this approach by showing that the visual system struggles to generalize recognition of novel objects to translations as small as 2° of visual angle. The present paper outlines a series of eyetracking studies that show novel objects can be recognized at translations as far as 18° from the trained retinal location, challenging the standard account of translation tolerance in neuroscience and psychology.

Keywords: Translation Tolerance; Translation Invariance; Object Recognition; Vision

Introduction.

We can identify familiar objects despite the variable images they project on our retina, including variation in image size, orientation, illumination, and position on retina. How the visual system succeeds under these conditions is still poorly understood. Here we focus on our ability to identify objects despite variations in retinal location and consider the extent to which the visual system relies on "on-line" vs. "trained" translation tolerance. In the case of on-line tolerance, learning to identify an object at one location immediately affords the capacity to identify that object at multiple retinal locations. At one extreme, the visual system can immediately generalize to all locations (to the limit of visual acuity), what might be called on-line translation invariance; at the other extreme, generalization is limited to a few degrees of visual angle. Trained tolerance, by contrast, refers to the hypothesis that we learn to identify familiar objects across locations by explicitly training the visual system to identify each object across a broad range of retinal locations. On this view, one of the functions of eye-movements is to ensure that objects are projected to multiple locations. These two accounts trade-off on one another: the more restricted on-line translation tolerance is the more trained tolerance is required to support the ability to identify objects across a wide range of retinal locations.

As detailed below, most of the empirical research in psychology and neuroscience suggests that on-line tolerance is restricted to a few degrees of visual angle, and to date, all neural network models of object identification show the same restriction. As a consequence, most theories of vision assume that trained tolerance plays an important role in our ability to identify objects across a range of retinal locations.

Early long-term priming studies by Biederman and colleagues (Biederman & Cooper, 1991; Cooper, Biederman, & Hummel, 1992; Fiser & Biederman, 2001) provided evidence for extensive on-line translation tolerance, and indeed, in some cases, translation invariance. For example, Fiser and Biederman (2001) asked participants to name line-drawings of objects as fast and accurately as possible in a study phase. In a later test block, participants were faster and more accurate to name repeated images compared to a set of control objects (same name, different exemplar) regardless of whether retinal position was the same at study and test or displaced by 10 degrees (°) of viewing angle. A limitation of all these studies, however, is that they assessed priming for familiar objects, and accordingly, participants had seen the same type of objects in a wide variety of retinal locations prior to the experiment. This leaves open the possibility that the findings reflected trained rather than on-line translation tolerance within the visual system, or indeed, trained tolerance outside the visual system with priming effects occurring within semantic or verbal systems (Kravitz, Vinson, & Baker, 2008).

In contrast with the Biederman studies, a number of authors have failed to observe robust translation tolerance for novel objects that participants had not seen prior to the experiment (e.g., Afraz & Cavanagh, 2008; Cox & DiCarlo, 2008; Dill & Fahle, 1997; Dill & Fahle, 1998; Newell, Sheppard, Edelman, & Shapiro, 2005). *Figure 1* outlines a selection of studies that used different experimental paradigms and found highly limited (in one case no) translation tolerance (adapted from Kravitz, Vinson, & Baker, 2008). Based on the outcome of such studies, Chen et al. (2017) state that "the translation-invariance of the human visual system is limited to shifts on the order of a few degrees - almost certainly less than 8°" (p.5). In line with this, Kravitz et al.'s (2008) review of behavioral studies found that "most of the training and matching studies found a significant decrement in discrimination performance with translations varying from 0.5° to 2°" (p. 118).

Neural data are also consistent with the idea that on-line translation tolerance is highly limited. Researchers have identified neurons in inferior-temporal cortex (IT) with a range of receptive fields (ranging from 2.8° to 26°; for review see Kravitz et al., 2008). The larger receptive fields are thought to provide the neural underpinning of translation tolerance, but it is important to note that these receptive fields have only been observed for familiar or newly-trained stimuli that have been seen at multiple

retinal locations (e.g., Gross et al., 1972; Ito, Tamura, Fujita, & Tanaka, 1995; Tovee, Rolls, & Azzopardi, 1994). Accordingly, these large receptive fields could reflect trained or on-line translation tolerance. Consistent with the former hypothesis, Cox and DiCarlo (2008) only observed small receptive fields for their novel stimuli that were trained in one location, as shown in *Figure 1C*. That is the neural data appear to mirror the behavioral data: robust translation tolerance and large receptive fields are found for familiar stimuli, limited generalization and small receptive fields are observed for unfamiliar stimuli. Indeed, Cox and DiCarlo reach a similar conclusion to Kravitz et al. (2008), writing "...the computational machinery of the ventral visual stream is not constructed in a manner that automatically produces position tolerance in IT, even across relatively small changes in retinal position. Instead, the creation and/or maintenance of IT position tolerance might require experience".

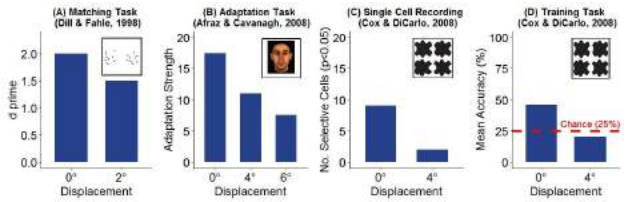


Figure 1. Behavioral studies of translation tolerance.

Similarly, work with artificial neural network models has reported robust translation tolerance for words and objects at trained locations, but highly limited generalization to untrained locations. For example, Dandarund et al. (2013) and DiBono & Zorzi (2013) showed that that their models of visual word identification supported translation invariance, but the models were trained with each word at each location. Elliffe, Rolls and Stringer (2002) showed that a biologically inspired neural network model called VisNet supported on-line translation tolerance to untrained locations for simple stimuli, but each stimuli had to be trained at multiple spatial locations (after training in 7 locations the model generalized to an 8th and 9th location), and the authors tested small translations (translations of 8 pixels in a 128x128 retina). The above behavioral, neural, and computational above findings have led most theorists in psychology and neuroscience to endorse the 'trained' account of translation tolerance.

Despite empirical and computational results, there are still reasons to question the trained tolerance hypothesis. Behavioural studies that failed to observe on-line translation (e.g., *Figure 1*) suffer from a number of limitations. For example, stimuli are typically unlike real objects (e.g., Dill & Fahle, 1997; see *Figure 1a*), and/or are very similar to each other (e.g., Cox & DiCarlo, 2008; see *Figure 1c*). Differentiating between highly similar items may rely on low-level visual representations that are retinotopically constrained (Kravitz et al., 2008). Additionally, stimuli in these experiments were typically trained at a given location for just 100ms (contrary to everyday visual experiences in which stimuli can be encoded for longer intervals). More extended studying time may be required for robust online translation tolerance. Consistent with the first possibility, Dill and Edelman (2001) observed much more extensive translation tolerance for unfamiliar objects that were more object-like and discriminable from one another. Indeed, they reported no significant reduction in performance in five of six experiments when images were displaced by 8 degrees. And consistent with both possibilities, Bowers, Vankov and Ludwig (2016) reported more robust translation tolerance still when participants trained to identify more discriminable stimuli novel stimuli that were studied at one retinal location for longer periods of time. Indeed, in their Experiment 3, participants were ~80 % accurate in naming novel objects following a shift of 13° (when chance was 16.7 %).

In this article we explore on-line translation tolerance in humans given the conflicting empirical evidence regarding on-line translation tolerance and the theoretical implications for theories of vision in psychology and neuroscience. Two gaze-contingent eye-tracking studies are reported and include the following critical design features. First, in all studies, 24 novel objects were used, each of which was a member of a pair of objects built from similar parts but in a different global configuration (see Method). Using a large set of stimuli of this sort should encourage participants to learn the complete objects rather than just the parts. Previous studies have rarely matched items on the basis of their parts (but see Dill & Edelman, 2001), and have typically included far fewer stimuli. Second, the novel three-dimensional objects we included were designed to be more naturalistic compared to the novel stimuli used in previous experiments, such as those depicted in *Figure 1a* and *1c*. This makes it more likely that the visual system will process these new stimuli in a manner more similar to everyday recognition. Third, we included study conditions in which stimuli were presented for unlimited time at study as opposed to the brief display conditions in previous studies that may have artificially reduced on-line translation tolerance. Fourth, we included test conditions in which objects were presented for 100ms durations, reducing the likelihood that participants adopted artificial strategies at test. Note that the Bowers et al. (2016) experiments reporting robust on-line translation invariance included a smaller number of less realistic objects that were displayed for an extended time at test. Accordingly, the current studies provide a much stronger test of the on-line translation tolerance hypothesis.

Experiment 1.

Method

Participants and Equipment. Fifteen participants (Experiment 1a=6, 1b=9) were recruited from the University of Bristol's course credit scheme for Psychology students. Eye-movements were monitored using the Eyelink 1000 plus system (SR Research). Stimuli were presented using Psychopy v1.85.3 (platform: Linux-Ubuntu), and on a 120Hz monitor with a spatial resolution of 1920 x 1080 pixels (screen width = 53cm), at a distance of 70cm.

Stimuli. Twenty-four novel objects were taken from Leek, Roberts, Oliver, Cristino, and Pegna (2016). Each object was part of a pair that had similar local features but a different global configuration (see *Figure 2*). For each participant, one member of each pair was randomly assigned the label 'A' and the other was assigned 'B'.

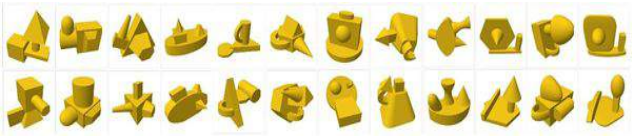


Figure 2. Twenty-four novel objects. Each column contains a pair of objects that are matched for similar local features, but which differ in global configuration.

Procedure. In the learning phase of the experiment participants were trained to categorize the 24 objects as 'A' or 'B'. Each object was presented one-by-one in the centre of the screen and occupied 5°x5° of visual angle. Participants were required to maintain their gaze on a centrally located fixation-cross for 1000ms for an object to appear. If gaze moved 1.5° beyond the fixation-cross, a mask replaced the object. The learning task was split into two phases: (i) *Familiarization*. The familiarization phase consisted of two presentations of each object. Each object was presented for an unlimited time and was accompanied by a sound-file indicating its category (A or B). (ii) *Training*. Each object was displayed at the same location without the sound file and for unlimited time until the participant pressed a button to indicate the image's category. Audio feedback was then provided. The training phase continued until the participant correctly identified each object consecutively (i.e., 24/24 consecutive correct answers). After completing the first training phase (Block 1), participants completed two additional training phases - Block 2 and Block 3 - which were identical except for their shorter presentation times of 500ms and 100ms respectively.

After the learning phase, participants completed seven test-blocks, each consisting of 24 presentations (one of each object); each test-block differed in terms of horizontal eccentricity from the centre of the object to the central fixation cross (i.e., displacement from training location). Test blocks 1, 2, 3, 4, 5, 6 and 7 presented images at 0° (i.e, trained-position), 3°, 3°, 6°, 6°, 9°, and 9° displacement from the centrally trained position, respectively. Test-blocks with the same displacement (e.g., block 2 and 3 were both 3°) differed in terms of presentation side (left or right). Within each test-block, order of presentation was randomised and no feedback was provided. In Experiment 1a images remained on the screen until participants responded. Experiment 1b was the same as Experiment 1a except that images were presented for 100ms at test in order to reduce possible response strategies. These final presentation durations are similar to previous studies that have failed to find online translation tolerance (see *Figure 1*).

Results.

Table 1. Mean (+/- 95% CI range) Accuracy in Experiment 1. *Columns show displacement of the test presentation from the trained location.*

	Mean (+/- 95% CI range) Accuracy			
	0°	**3°**	**6°**	**9°**
Exp 1a (N = 6)	98% (5%)	98% (3%)	95% (9%)	94% (9%)
Exp 1b (N = 9)	98% (2%)	95% (3%)	91% (4%)	84% (7%)

As shown in *Table 1*, novel objects were recognised with high accuracy at untrained retinal-positions (chance is 50%). Even at the most distal untrained position (9°), objects were recognised with a mean accuracy of 94% when unlimited time was afforded at test (Experiment 1a), and although translation tolerance was reduced when stimuli were presented for 100ms at test (Experiments 1b), accuracy was still 84% when at 9° displacement.

Experiment 2.

Experiment 2 served two purposes. (i) Although we observe near complete translation invariance for newly acquired objects displaced by 9° when objects were presented for an unlimited amount of time, there was a significant reduction when stimuli were presented for 100 ms at test. In an attempt to reduce any effects of retinal specificity, Experiment 2a adopted a learning condition known as 'location-training' (i.e., training at the test location - see below). (ii) Experiment 2b also used location training to examine whether the robust on-line translation reported in our experiments (1a to 2a) could be extended to an even more distal untrained location, 18° from the trained location.

Location training has been used by previous studies to show that participants can overcome retinal-specificity for low-level visual discrimination tasks. Xiao et al. (2008) demonstrated that participants who had been trained to discriminate contrasts at location 1 showed complete transfer of this ability to location 2 only when they had also been trained to discriminate different stimuli on a different dimension (orientation) at location 2 (otherwise, learned contrast discrimination was location specific). Xiao et al. concluded that training at location 2 trained participants to overcome stimulus-nonspecific factors like local noise at the stimulus location, and this enabled complete location transfer. Our Experiment 2 investigated whether location training may also reduce retinal specificity in high-level visual recognition tasks (Experiments 2a and 2b investigated this at displacements of 9° and 18°, respectively).

Participants, Equipment and Stimuli. Experiment 2 used identical equipment, stimuli and recruitment methods as Experiment 1. Ten participants were used in Experiment 2a, and 10 different participants in Experiment 2b.

Experiment 2a: Learning and Test Phase. During learning blocks 1 and 2, objects were trained for unlimited time and 100ms duration respectively, at the centre of the screen (at fixation). In block 3, 'location training' took place: twelve objects were trained at one peripheral location, 9° from the central fixation-cross and the remaining 12 objects were trained at a contralateral peripheral location, 9° to the other side of the fixation-cross (all presentations were 100ms). Participants were trained until they got 12/12 consecutive correct answers at each peripheral location. In the test phase, objects were tested at three test locations: 9° left, 9° right, and centre of fixation, giving three test conditions: "trained-central" (0° displacement from central training location), "trained-peripheral" (0° displacement from peripheral training location) and "novel-peripheral" locations (9° displacement from central training location, on the opposite side to the trained peripheral location). To control for possible order effects, the three test locations were randomly interleaved within each test-block.

Experiment 2b: Learning and Test Phase. Experiment 2b examined whether the robust on-line translation reported in Experiment 2a could be extended to an even more distal untrained location, 18° from the trained location. In order to displace presentations by 18° at test, images were presented at one peripheral location only (and never at central fixation): 12 images were presented 9° to the right, and the remaining 12 were presented 9° to the left of central fixation (images were presented for unlimited time in block 1, and for 100ms duration in blocks 2 and 3). In an attempt to boost performance compared to Experiment 2a, participants were required to get 24/24 consecutive correct answers in each block. At test, objects were tested at two test locations: 9° left, and 9° right of fixation, giving two test conditions: "trained-peripheral" (0° displacement from peripheral trained location) and "novel-peripheral" locations (9° displacement from central fixation, and thus 18° displacement from the opposite peripheral location).

Results.

Table 2. **Mean (+/- 95% CI range) Accuracy scores in Experiment 2a and 2b.** Columns show degrees by which the test presentation was displaced from the nearest training location and the screen position of that test presentation.

	Mean (+/- 95% CI range) Accuracy			
Displacement	**0°**	**0°**	**9°**	**18°**
Screen Position	**Centre (trained)**	**Peripheral (trained)**	**Peripheral (novel)**	**Peripheral (novel)**
Exp 2a (N=10)	93% (5%)	83% (5%)	81% (6%)	not tested
Exp 2b (N=10)	not tested	97% (3%)	not tested	89% (7%)

The results of Experiment 2a and 2b are summarised in *Table 2*. In Experiment 2a, mean accuracy scores at the *novel-peripheral* position (9°) were significantly above chance and were nearly equivalent to those yielded in the *trained-peripheral* position (0° Peripheral). Thus, Experiment 2a provided strong evidence for robust online translation tolerance over 9° displacement even when objects are presented for just 100ms at test.

In Experiment 2b, objects were recognised with a very high degree of accuracy even when the trained location was displaced by 18°. Moreover, 5 of the 10 participants scored above 90% on mean accuracy at 18°. As illustrated in *Figure 3*, the findings from Experiment 2b, show on-line translation tolerance for novel stimuli over much more distal displacements compared to all previous work.

Discussion

The present paper has provided evidence of robust on-line translation tolerance in the human visual system. Participants trained to recognise novel objects at one retinal position could recognise the same objects at untrained distal retinal-locations (up to 18°) with high accuracy.

The findings are contrary to previous studies that demonstrate much more limited generalization over translations as small as 2 and 4° (e.g., Cox & DiCarlo, 2008; Dill & Fahle, 1998) and that have been used to support trained theories of translation tolerance. Indeed,

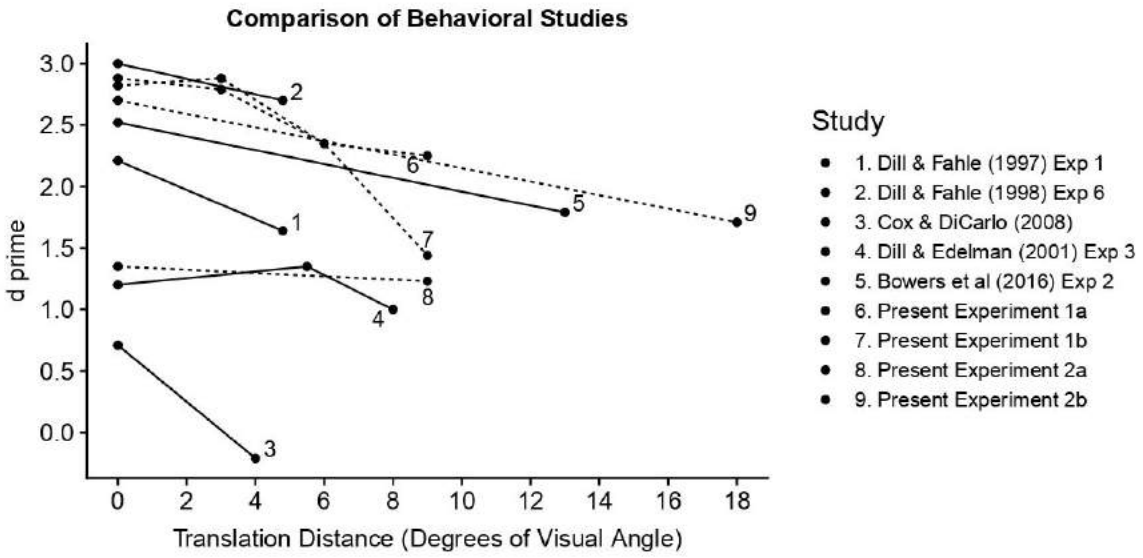


Figure 3. Comparison of d-prime scores for previous and present experiments as a function of translation distance. For each study, the experiment with the best performance at the largest displacement is illustrated. The present experiments (dashed lines) show more robust on-line translation tolerance than the majority of previous experiments. Experiment 2b showed robust on-line translation tolerance over a larger translation distance (18°) than any previous experiment. d-prime scores were calculated using the psyphy package (Knoblauch, 2014) for R (R Development Core Team, 2018).

advocates of the approach have recently claimed "the translation-invariance of the human visual system is limited to shifts on the order of a few degrees - almost certainly less than 8°" (Chen et al., 2017; p. 5). Rather, the present findings indicate that novel object recognition can be generalized on-line to more distal untrained retinal positions than previously demonstrated (see *Figure 3*): objects were recognised over translations as large as 18° with performance near 90%.

Why was robust on-line translation tolerance demonstrated in the present experiments whereas most previous experiments demonstrated highly limited generalization? As described above, previous studies typically used stimuli that are unlike real objects and/or are very similar to each other. Differentiating such stimuli may rely on low-level visual processes that are highly retinotopically constrained. High-level visual processes may also be more retinotopically constrained under these conditions. Indeed, there is some physiological evidence that receptive field (RF) sizes of neurons in the infero-temporal cortex (IT) are a function of stimuli size (DiCarlo & Maunsell, 2003). The present study used more naturalistic stimuli and included a number of variations in the procedure used by most psychophysical studies, including extended sampling times and, in Experiment 2, 'location training'. The more naturalistic conditions may have encouraged recruitment of IT neurons with larger RFs. Other studies that have also observed robust translation tolerance have also used more naturalistic, easily discriminable stimuli (Dill & Edelman, 2001) and extended sampling times (Bowers et al., 2016), but our findings go beyond this work by showing that robust translation tolerance extends to 18° under conditions in which strategic processing is minimized (by flashing items at test for 100 ms and by including a large set of novel objects that differed in the configuration of similar parts).

The findings are also relevant to computational modelling of the visual system. As noted in the Introduction, previous attempts to achieve on-line translation tolerance with artificial neural networks have reported highly limited tolerance. Such demonstrations may have been considered a strength given similarly limited tolerance reported in humans (e.g., Dill & Fahle, 1997; 1998). The present results show the need for these models to capture the robust on-line translation tolerance we have reported in humans. There is reason to believe that at least one class of artificial neural network can achieve this. Deep convolutional neural networks (CNNs) are designed to support translation tolerance by including convolutional layers and global pooling layers. Convolutions involve copying learning that occurs at one retinal location to other locations (the premise that information learned at one location is relevant at others), whilst pooling layers aggregate information from multiple

spatially organized units to a single unit in order to down-size the image. Both of these inbuilt ("innate") mechanisms are widely claimed to support translation tolerance, but there is surprisingly little evidence as to whether these mechanisms can support robust on-line translation tolerance as we have observed. We are in the process of running simulations to assess whether CNNs can support our empirical results.

Overall, the evidence outlined above is a clear demonstration that the human visual system can support recognition of novel objects at untrained distal retinal positions. Since the standard approach within psychology and neuroscience is to deny such robust generalization, there is a need for the field to more widely acknowledge an on-line generalization mechanism that can account for these results.

Acknowledgments

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme.

References

Afraz, S.R., & Cavanagh, P. (2008). Retinotopy of the face aftereffect. *Vision Research, 48,* 42–54.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review, 94*, 115–147.

Biederman, I., & Cooper, E. E. (1991). Evidence for complete translational and reflectional. *Perception, 20*, 585–593.

Biederman, I., Cooper, E.E., Kourtzi, Z., Sinha, P., & Wagemans, J. (2009). Biederman and Cooper's 1991 Paper. *Perception*, *38*, 809–825.

Bowers, J. Vankov, I. & Ludwig, C. (2016). The visual system supports online translation invariance for object identification. *Psychonomic Bulletin Review, 23,* 432-438

Chen, F. X., Roig, G., Isik, L., Boix, X, & Poggio, T. (2017). "Eccentricity dependent deep neural networks: Modeling invariance in human vision," in AAAI Spring Symposium Series, Science of Intelligence, 2017.

Cooper E. E., Biederman I., & Hummel J. E. (1992). Metric invariance in object recognition: A review and further evidence. *Can. Psychol. 46*, 191-214.

Cox, D. D. & DiCarlo, J.J. (2008). Does Learned Shape Selectivity in Inferior Temporal Cortex Automatically Generalize Across Retinal Position? *Journal of Neuroscience*, *28*, 10045-10055.

Dandurand, F., Hannagan, T., & Grainger, J. (2013). Computational models of location-invariant orthographic processing. *Connection Science, 25*, 1–26.

Di Bono, M. G., & Zorzi, M. (2013). Deep generative learning of location-invariant visual word recognition. *Frontiers in Psychology, 4*, 635.

DiCarlo, J. J., & Maunsell, J. H. (2003). Anterior inferotemporal neurons of monkeys engaged in object recognition can be highly sensitive to object retinal position. *Journal of Neurophysiology, 89,* 3264–3278.

Dill, M., & Fahle, M. (1997). The role of visual field position in pattern discrimination learning. *Proceedings of the Royal Society B, 264,* 1031–1036.

Dill, M. & Fahle, M. (1998) Limited translation invariance of human visual pattern recognition. *Perception & Psychophysics, 60*, 65–81

Dill, M., & Edelman, S. (2001). Imperfect Invariance to Object Translation in the Discrimination of Complex Shapes. *Perception, 30*, 707–724.

Elliffe M.C.M., Rolls E.T., & Stringer S.M. (2002). Invariant recognition of feature combinations in the visual system. *Biological Cybernetics, 86,* 59– 71.

Fiser, J., & Biederman, I. (2001). Invariance of long-term visual priming to scale, reflection, translation, and hemisphere. *Vision Research, 41*, 221–234.

Gross, C.G. et al. (1972). Visual properties of neurons in inferotemporal cortex of the macaque. *J. Neurophysiol., 35*, 96–111

Ito, M. et al. (1995) Size and position invariance of neuronal responses in monkey inferotemporal cortex. *J. Neurophysiol., 73*, 218–226

Kravitz, D. J., Vinson, L. D., & Baker, C. I. (2008). How position dependent is visual object recognition? *Trends in Cognitive Sciences, 12,* 114–122.

Leek, E. C., Roberts, M. V., Oliver, Z. J., Cristino, F., & Pegna, A. (2016). Early differential sensitivity of evoked-potentials to local and global shape during the perception of three-dimensional objects. *Neuropsychologia, 89,* 495–509.

Nazir T, & O'Regan J.K. (1990). Some results on translation invariance in the human visual system. *Spatial Vision, 5,* 81-100.

Newell, F. N., Sheppard, D. M., Edelman, S., & Shapiro, K. L. (2005). The interaction of shape- and location-based priming in object categorisation: Evidence for a hybrid "what + where" representation stage. *Vision Research, 45*, 2065–2080.

Tovee, M. J., Rolls, E. T., Azzopardi, P., (1994). Translation invariance in the responses to faces of single neurons in the temporal visual cortical areas of the alert monkey. *J. Neurophysiol., 72,* 1049-1060.

Ullman, S. (2007). Object recognition and segmentation by a fragment based hierarchy. *Trends Cogn Sci, 11,* 58–64.

Xiao, L.Q., Zhang, J.-Y., Wang, R., Klein, S. A., Levi, D. M., & Yu, C. (2008). Complete transfer of perceptual learning across retinal locations enabled by double training. *Current Biology, 18*, 1922–1926.

Is It Better to Be in Shape or on Top of It? The Impact of Control, Valence, and Expectedness on Non-Spatial Uses of *in* and *on*

Brooke O. Breaux (brookebreaux@louisiana.edu)
Jessi Lynne LaSalle (jessi.lasalle@yahoo.com)
Peyton Lute (peytonlute@gmail.com)
Catherine Brousse (cmb10795@gmail.com)
Claudia Mijares (cdmijares@gmail.com)
Department of Psychology, University of Louisiana at Lafayette
Lafayette, LA 70504 USA

Abstract

Using the prepositions *in* and *on*, Jamrozik and Gentner (2015; 2014; 2011) explored a particular factor of meaning that was hypothesized to serve as a metaphorical link between spatial and abstract concepts. Across several studies, these researchers have provided evidence for the idea that there is a "continuum of control" that exists for both spatial and abstract uses of *in* and *on*. Our research explores other potential meaning factors that might play a role in non-spatial uses of *in* and *on*. Our results replicate and extend Jamrozik and Gentner's (2011) findings. We advocate using a multi-componential approach as research involving indirect metaphors continues moving forward.

Keywords: prepositions; spatial language; abstract language; metaphor; language understanding; semantics

Introduction

A popular assumption in cognitive linguistics is that metaphors are extremely common in both language and thought (e.g., Lakoff & Johnson, 1980). Historically, the evidence provided for this assumption has been primarily linguistic in nature. For example, a conceptual metaphor such as LOVE IS A CONTAINER is proposed to exist in the minds of speakers because it is natural to talk about people being in love or people falling out of love regardless of containment being a spatial concept and love being an abstract concept. One question, which arises from this assumption, is the nature of the connection between the spatial and the abstract. In other words, how might the spatial and abstract concepts activated by a conceptual metaphor be connected?

The potential for this type of metaphorical connection is most apparent when people talk about abstract concepts, such as time, thoughts, and emotions, using terms drawn from physically-based domains, such as space, force, and motion. Jamrozik and Gentner (2015; 2014; 2011) have explored the possibility of such connections using the prepositions *in* and *on*. Steen (2010) refers to these as indirect metaphors. Across several studies, these researchers present evidence suggesting that control is an important concept not only for spatial uses of *in* and *on* but also for abstract uses of *in* and *on*. For example, consider the scenario of a marble in a jar that has been secured with a lid. The marble is considered the figure and the jar is the ground. Even if you were to shake the jar or turn it upside down, the marble has very little control over where it can be at any given moment in time. In other words, the ground has more control than the figure in this situation. Alternatively, consider the scenario of a marble on a plate. The marble is still the figure, but this time the plate is the ground. If you were to move the plate, the marble might easily roll right off. In this example, the figure has more control than the figure. Borrowing from Beitel, Gibbs, and Sanders' (2001) terminology, the plate does not constrain the movement of the marble as well as the closed jar.

Jamrozik and Gentner's (2015) research suggests that this difference in the amount of control associated with *in* as compared to *on*—what they refer to as a "continuum of control"—also exists for abstract uses of *in* and *on*. Again, consider an example. If someone was described as being *in trouble*, participants in Jamrozik and Gentner's (2015) studies thought that the figure had a low degree of control over their situation. If someone was described as being *on a roll*, participants in Jamrozik and Gentner's (2013) studies thought that the figure had a higher degree of control over their situation.

Jamrozik and Gentner (2015) selected control as their factor of interest because previous researchers (Coventry, Carmichael, & Garrod, 1994) have considered control to be the most likely candidate for extension to abstract contexts. Our goal is to extend Jamrozik and Gentner's (2015) research by considering other factors that might be useful in differentiating between abstract uses of *in* and *on*. If control is not the only candidate for extension, then one place to look for other potential candidates is in literature involving the spatial semantics of the locative prepositions *in* and *on*.

Early research on the spatial meaning of *in* and *on* tended to focus on the role of geometric constructs such as inclusion and contact (e.g., Bennett, 1975; Herskovits, 1986; Leech, 1969; Miller & Johnson-Laird, 1976); however, these approaches cannot account for cases in which the necessary geometric constructs are present but the lexical item in question is dispreferred (e.g., a pear on a counter that is being covered by an overturned bowl is not considered *in* the bowl) as well as cases in which the necessary geometric constructs are not present but the lexical item in question is preferred (e.g., describing a book that is on top of another book as being *on* a table even though it is not directly in contact with or being supported by the table). In response to these inadequacies, researchers have developed more multi-

componential approaches in which they propose that a variety of different factors feed into the meanings of spatial relational terms (Coventry & Garrod, 2004; Feist, 2000, 2010). These factors include geometric contact and geometric inclusion, which are factors less likely to extend to abstract concepts, as well as factors such as location control that are more likely to extend to abstract concepts. One example of a factor that may extend to abstract concepts is object association. More specifically, research by Coventry and Prat-Sala (2001) revealed a complex interaction between the factors of control and object association such that when figure control was high, acceptability of *on* was higher when the figure-ground combination was unusual (e.g., a brick on a plate) and lower when the figure-ground combination was expected (e.g., a fish on a plate). Whether or not this spatial factor of expectedness plays a significant semantic role in non-spatial uses of *in* and *on* and whether it has the same complex interaction with figure control has yet to be explored.

Interestingly, Jamrozik and Gentner's (2015) research on the abstract uses of *in* and *on* suggests another factor that should be taken into consideration. It is a factor that has not previously been considered as a spatial meaning component: valence. Throughout their studies, Jamrozik and Gentner (2011; 2014; 2015) provide a variety of stimulus examples. More often than not, these examples are indicative of a particular relationship between control and valence: Statements associated with higher figure control are often more positive (e.g., *Jordan is on a roll*) than statements associated with lower figure control (e.g., *Casey is in a depression*). Jamrozik and Gentner (2015) explain that this is evidence of a natural correlation between control and positive valence. They point out that the correlation is not perfect, and there are certainly examples for which the relationship does not hold (e.g., *on thin ice*; *in shape*).

Given all of this, we first set out to explore the potential relationships between control, valence, and the comprehension of *in* and *on* in non-spatial contexts. We then set out to explore the potential relationships between control, valence, and expectedness and how these factors might influence the production of *in* and *on* in non-spatial contexts.

Experiment 1

In line with previous results (cf. Jamrozik and Gentner, 2015), we predicted that participants would rate the figures of conventional *on* phrases as having more control than the figures of conventional *in* phrases. We also predicted that participants would rate conventional *on* phrases as more positive than conventional *in* phrases.

Method

Participants A total of 47 college students participated in exchange for course credit: 24 in Version 1 and 23 in Version 2. The mean age of participants was approximately 19 (M = 18.8, SD = 1.7). Forty were female (85%) and seven were male (15%). Only one participant reported being a non-native English speaker.

Materials and Design Participants in this experiment were presented with 160 sentences. Each sentence consisted of a human figure and either a prepositional phrase or verb phrase. Of these sentences 89 were the target stimuli used by Jamrozik and Gentner (2015): 44 *in* sentences and 45 *on* sentences. We developed the remaining 71 filler sentences by following a procedure similar to the one described by Jamrozik and Gentner (2015) in which we selected conventional abstract uses of prepositions (other than *in* and *on*) and verbs from online idiom dictionaries and randomly assigned them common gender ambiguous names. Using a procedure similar to that described by Jamrozik and Gentner (2015), names were selected for use if they appeared in the top 1,000 names given to both males and females in social security records of American children born between 1990 and 2000. Of the 71 filler sentences, 21 were prepositional phrases (e.g., *Peyton is at ease*; *Bailey is under the weather*), 26 were verb + preposition phrases (e.g., *Quinn is letting the cat out of the bag*; *Alex is beating around the bush*), and 24 were verb phrases (e.g., *Noel is taking it easy*; *Taylor is jumping the gun*).

Procedure Participants took part in this study via SurveyMonkey.com. After consenting, participants were asked to answer a standard set of demographic questions: their age, gender, native language, and other languages they are able to speak, write, read, or understand.

Participants then read and rated the 160 sentences described previously. Following from Jamrozik and Gentner (2015), the sentences were presented one at a time on the screen and in a pseudo-randomized order such that there were never two *in* sentences, *on* sentences, or target sentences presented back-to-back. Participants were instructed to "imagine the scenario each sentence describes and think about how much the person controls or is controlled by the situation" as well as "the degree to which the situation being described is likely to be a positive or negative event in that person's life." For each sentence, participants were presented with a figure-control question (i.e., "To what degree does the person have control of the situation?") to which they would respond on a scale ranging from "1-extremely low control of the situation by the person" to "5-extremely high control of the situation by the person." On the same page participants were presented with a valence question (i.e., "To what degree is the situation being described likely to be a positive or negative event in their life?)" to which they would respond on a scale from "1-extremely negative" to "5-extremely positive." We also developed 19[1] catchtrials that were presented in a pseudo-randomized order such that there was only ever one catchtrial per sentence and never more than two

[1]There were originally 20 catchtrials; however, due to experimenter error, one catchtrial in which the response was five was not included in the final version of the study.

catchtrials appearing back-to-back. Catchtrials involved asking participants to provide a specific numerical rating from one to five on the scale provided.

At the end of the experiment, participants were then presented with an electronic debriefing form. Two versions of the experiment were developed due to concerns that participants might simply provide the same response to both the figure control and valence questions. Participants were assigned to one version, and the only difference between them being that the valence scale response options were flipped such that "extremely positive outcome" was associated with one and "extremely negative outcome" was associated with five. These responses were reversed scored before analysis.

Results

Of the 47 participants described previously, 12 were removed from further data analysis: one for reporting a native language other than English, two for responding incorrectly to three or more of the catchtrials, and nine for not finishing the study. Thus, data for 35 participants were used in the following analyses.

To determine whether the mean figure-control ratings produced by participants in the current study were aligned with the mean figure-control ratings produced by participants in Jamrozik and Gentner's (2015) Experiment 1a, we conducted a pairwise correlation, $r(88) = .94$, $p < .001$.

It should not be surprising, then, that when we conducted a mixed-model ANOVA with preposition as a within-subjects variable and version as a between-subjects variable that we found the same significant effect of preposition on ratings of figure-control reported by Jamrozik and Gentner (2015): Participants in the current study rated figures *on* ground as having more control ($M = 3.6$, $SD = .48$) than figures *in* ground ($M = 3.1$, $SD = .47$), $F(1, 33) = 132.81$, $p < .001$, $\eta p^2 = .801$. As expected, there was no significant main effect or interaction involving version.

To determine the proportion of stimuli that fit with our predictions, we conducted a median-split analysis using the control ratings. The results showed that our predictions held for 64% of the *on* sentences (ratings falling above the median) and 66% of the *in* sentences (ratings falling below the median).

When we conducted a mixed-model ANOVA with preposition as a within-subjects variable and version as a between-subjects variable using valence scores, we found that participants rated figures *on* ground as more positive ($M = 3.26$, $SD = .28$) than figures *in* ground ($M = 2.92$, $SD = .24$), $F(1, 33) = 103.97$, $p < .001$, $\eta p^2 = .759$. We also expected that if participants were simply providing the same response to the control and valence questions that we would see significant difference across versions when the anchoring of the valence question was reverse. No significant main effect or interaction involving version was observed.[2]

To determine the proportion of stimuli that fit with our predictions, we conducted a median-split analysis using the valence ratings. The results showed that our predictions held for 42% of the *on* sentences (ratings falling above the median) and 57% of the *in* sentences (ratings falling below the median).

Discussion

As predicted, we were able to replicate the results of Jamrozik and Gentner's (2015) Experiment 1a using their set of conventional phrases. In addition to finding that *on* is associated with more figure-control than *in*, we also observed that *on* phrases were rated as more positive than *in* phrases. This evidence suggests that, in addition to control, valence may serve as a meaning component that can be used to differentiate between abstract uses of *in* and *on*. That being said, our median-split analyses revealed that control might play a more predictable role than valence across a variety of different contexts. We argue that these patterns are consistent with a multi-componential approach, meaning that consideration of a variety of different semantic factors might be useful when investigating the non-spatial semantics of *in* and *on*.

A significant limitation of Experiment 1 is that these findings might also be due to a lack of counterbalancing because participants were always presented with the same order of ratings: control followed by valence. It could be that a significant effect of valence was found only because valence ratings were always considered through the lens of control. It is also possible that these findings might simply be the result of characteristics particular to the sentences that were used as stimuli. Furthermore, even though using conventional phrases served to enhance the ecological validity of Experiment 1, it also limited the degree to which we were able to explore the potential interaction between control and valence.

Experiment 2

With mounting evidence that control is an important factor in distinguishing between the meanings of *in* and *on* in non-spatial contexts, we wanted to take a multi-componential perspective and explore the potential relationships between control, valence, and expectedness and how these factors

[2]Due to experimenter error, the *in* sentence *Noel is in the know* was absent from the stimuli presented to the participants. Instead the filler phrase *out of practice* was inadvertently presented twice, once with the name Avery and once with the name Noel. Because the missing phrase would have likely been rated by our participants as positive and a situation in which there was low control of the situation by the person, this could be viewed as a confound since we may have inadvertently eliminated data that could have served to refute our hypothesis. To address this possibility, we eliminated from analysis the *on* phrase with the most positive valence rating (*Adrian is on task*; $M = 4.71$) and conducted the same set of analyses described in the main text. Even without this sentence, the pattern of results and the significance tests outcomes were the same: A significant main effect of preposition was found for both control and valence ratings.

might influence the production of *in* and *on* in non-spatial contexts.

Method

Participants A total of 122 college students participated in exchange for course credit. The mean age of participants was approximately 18 (M = 18.1, SD = 1.8). Nineteen were male (16%) and 103 were female (84%). Only three participants reported being non-native English speakers.

Materials and Design We developed 64 stories for this experiment. Eight base stories were created. The stories averaged 3 sentences (or 45 words) in length. Each base story was associated with a general theme and a fictitious person name that was gender ambiguous. Eight story types were developed from each base story by fully crossing the following factors: control of the figure (high vs. low), valence of the outcome (positive vs. negative), and expectedness of the scenario (expected vs. unexpected). Eight versions of the experiment were then created, each containing only one story from each base and only one of each story type. Each participant was randomly assigned to a particular version of the experiment, and the order in which the stories were presented to participants was completely randomized.

We also selected eight nonwords to serve as novel prepositional objects for the production portion of this experiment. Using a procedure similar to the one described by Jamrozik and Gentner (2015), we generated these eight nonwords using the ARC Nonword Database (Rastle, Harrington, & Coltheart, 2002: https://www.cogsci.mq.edu.au/research/resources/nwdb/nwdb.html): *vight*, *slief*, *thwom*, *yease*, *prach*, *gwinn*, *malse*, and *zaiff*. The eight nonwords were assigned in a pseudo-randomized fashion such that each participant saw each nonword only once during the experiment and that each nonword appeared with a different story type across each of the eight versions.

Procedure Participants took part in this study via SurveyMonkey.com. After consenting, participants were asked to answer a standard set of demographic questions: their age, gender, native language, and other languages they are able to speak, write, read, or understand.

We then asked participants to read and respond to eight short stories (see Table 1 for examples). Four catch trials were also included in each version of the experiment. Catchtrials involved asking participants to provide a specific numerical rating from one to five on the scale provided.

After reading each story, participants were presented with three questions and their respective scales: a figure-control question (i.e., "To what degree does the person have control of the situation?") to which they would respond on a scale ranging from "1-extremely low control of the situation by the person" to "5-extremely high control of the situation by the person;" a scenario expectedness question (i.e., "To what degree is the situation natural and expected?") to which they would respond on a scale ranging from "1-extremely unnatural and unexpected" to "5-extremely natural and expected;" and an outcome valence question (i.e., "To what degree is the situation being described likely to be a positive or negative event in their life?)" to which they would respond on a scale from "1-extremely negative" to "5-extremely positive." Participants were then asked to imagine that they overheard someone talking about the fictitious person they just read about in the story and to decide whether *in* or *on* was more likely to appear in the novel statement that they overheard (e.g., "Adrian is ____ a gwinn").

After reading eight stories and responding to the four questions following each story, participants were then presented with an electronic debriefing form.

Table 1: Example of two story types developed from one base story.

Figure-control	Valence outcome	Scenario expectedness
High	Positive	Expected
Lee works at a local pizza place and has spent lots of time developing their customer service skills. Because of what they learned, Lee always wears a clean uniform to work. Lee just found out that they are going to be promoted.		
Low	Negative	Unexpected
Lee's parents forced them to get a job at the local pizza place owned by their family. Lee's manager makes every employee wear a clown suit to work. Lee just found out that they are going to be fired.		

Results

Of the 122 participants described previously, four indicated that they were fluent in another language other than English but failed to specify which language despite explicit instructions to do so. It was determined that these failures to respond were likely due to participants not paying close enough attention to the questions being presented to them; therefore, their data was excluded from further analysis. Of the remaining 118 participants, 11 were removed from further data analysis: three for reporting a native language other than English, four for responding incorrectly to one or more of our four catchtrials, three for not finishing the study, and one for attempting to complete the study a second time. Thus, the data for 111 participants was used in the following analyses.

Ratings We first needed to determine whether participants were sensitive to the ways in which we manipulated the three variables across the different story types; therefore, we conducted a one-way by items ANOVA for each factor. Table 2 shows that, on average, stories were rated in accordance with our manipulations.

In order to explore the roles that outcome valence and scenario expectedness might play in decisions of figure-control, we then conducted a repeated measures ANOVA on control ratings using figure-control, outcome valence, and scenario expectedness as within-subjects variables and version as a between-subjects variable. Version did have a significant effect overall, $F(7, 103) = 4.75$, $p = .004$, $\eta p^2 = .180$, suggesting that participants' control ratings differed depending on the sets of stories they received. Moreover,

every time version participated in an interaction, the interaction was statistically significant. Even though these complex interactions were not analyzed further, this pattern suggests that the following pattern of effects might be driven by only a subset of the stories presented to participants.

Table 2: Mean ratings (and standard deviations) across factor levels for figure-control, outcome valence, and scenario expectedness.

	Factor Levels		*p*-value
Figure-Control	Low	High	
	2.59(1.38)	3.53(1.45)	< .001
Outcome Valence	Negative	Positive	
	1.65(0.83)	4.18(1.06)	< .001
Scenario Expectedness	Unexpected	Expected	
	2.70(1.40)	3.26(1.26)	< .001

Of the three within-subjects variables, two had significant effects on participants' ratings of control: figure-control and valence. The significant effect of valence was such that participants' ratings of control were higher in response to high figure-control stories (M = 3.53, SD = 1.45) as compared to low figure-control stories (M = 2.59, SD = 1.38), $F(1, 103) = 143.42$, $p < .001$, $\eta p^2 = 0.533$. Such an effect is not surprising. What is more interesting is that ratings of control were also influenced by valence, $F(1, 103) = 87.17$, $p < .001$, $\eta p^2 = .458$. More specifically, ratings of control were higher in response to positive valence stories (M = 3.39, SD = 1.44) as compared to negative valence stories (M = 2.72, SD = 1.47). Unlike valence, expectedness did not have a significant impact on participants' control ratings: Control ratings made in response to stories with expected outcomes (M = 3.06, SD = 1.48) were not significantly different from stories with unexpected outcomes (M = 3.05, SD = 1.50).

Interestingly, there were two significant interactions that did not involve the between-subjects variable of version. One was an interaction of figure-control and outcome valence, $F(1, 103) = 41.02$, $p < .001$, $\eta p^2 = .285$ (see Figure 1). This interaction was such that stories with positive outcomes received higher control ratings when they described high control situations (M = 4.10, SD = 1.16) as compared to low control situations (M = 2.69, SD = 1.34), $F(1, 103) = 172.85$, $p < .001$, $\eta p^2 = 0.627$; however, these differences in control ratings were observed to a lesser degree when stories with negative outcomes described high control situations (M = 2.96, SD = 1.49) as compared to low control situations (M = 2.49, SD =1.41), $F(1, 103) = 15.85$, $p < .001$, $\eta p^2 = .133$.

The other interaction of interest involved expectedness, which alone did not have a significant impact on control ratings. The interaction of expectedness and valence, $F(1, 103) = 46.17$, $p < .001$, $\eta p^2 = .310$, was such that participants rated the figures of stories with positive outcomes as having more control when the scenario was expected than when it was unexpected, $F(1, 103) = 23.01$, $p < .001$, $\eta p^2 = .183$; however, the opposite pattern was observed for stories with negative outcomes: Participants rated figures as having less control when the scenario was expected than when it was unexpected, $F(1, 103) = 18.31$, $p = .001$, $\eta p^2 = .151$ (see Figure 2).

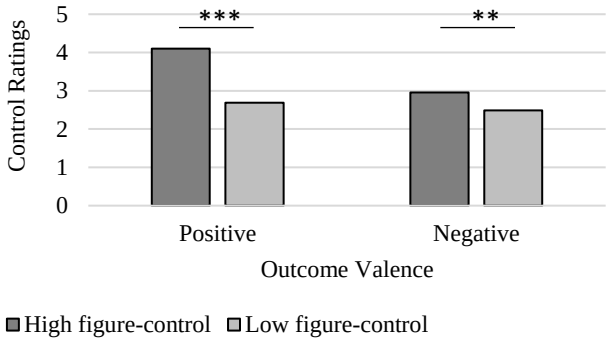


Figure 1: Mean control ratings across levels of figure-control and valence.

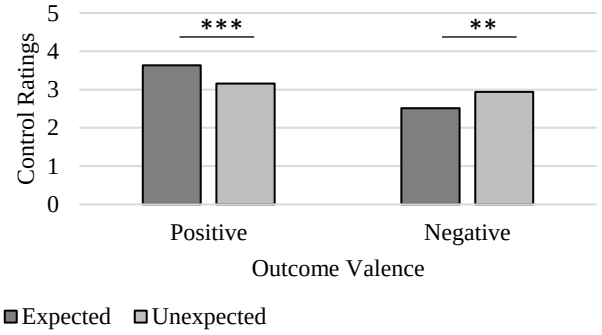


Figure 2: Mean control ratings across levels of expectedness and valence.

In addition to the pervasive influence of version, another concern arose during data analysis. Due to experimenter error 34% of the stories presented to participants actually contained the lexical items *in* and *on*; therefore, it is possible that ratings of control might have had more to do with lexical priming than the factors of interest. This lexical items confound was distributed across versions such that of the eight stories read by each participant at least one but no more than three contained a lexical item confound.

To analyze the impact that this lexical items confound may have had, we coded stories containing the word *in* as one and stories containing the word *on* as three. All "neutral" stories were coded as two, which included all stories that did not contain the lexical items *in* and *on* as well as two stories that contained both *in* an *on*. These codes were constructed such that high scores on the lexical items confound would be associated with high figure-control. The concern, then, was that high lexical confound ratings might predict high figure-control ratings; however, the results of a linear regression

actually showed the opposite to be significant: Low lexical confound ratings were associated with higher figure-control ratings, $F(1, 886) = 4.74$, $p = .030$, with an R^2 of .01. Participants' predicted control ratings decreased -.19 points for each increase in lexical confound, suggesting that seeing the lexical item *on* (as opposed to *in*) caused participants to produce lower ratings of control. This is a surprising finding that cannot be readily explained on the basis of lexical priming and does not align with the findings of Jamrozik and Gentner (2015).

Production We hypothesized that the factors of figure-control, outcome valence, and scenario expectedness would have an influence on production. In particular, we were interested in the likelihood that participants would choose either *in* or *on* to complete a novel phrase about the figure described in the story. A mixed effect logistic regression analysis looking at the choice between *in* and *on* as a function of all three within-subjects factors, the between-subjects factor of version, and all of their possible interactions revealed only significant effects involving figure-control, $F(1, 103) = 4.69$, $p = .033$, and outcome valence, $F(1, 103) = 24.13$, $p < .001$ (see Table 3). The figure-control effect was such that *on* was more preferred when figure-control was high (43%) and less preferred when figure-control was low (35%). The outcome valence effect was such that *on* was more preferred when the outcome was positive (47%) and less preferred when the outcome was negative (32%). Despite the pervasive influence of version in the ratings data, version did not play a significant role in production.

Table 3: Percentage of on responses and number of biased version across the eight story types.

Figure-Control	Outcome Valence	Scenario Expectedness	% *on* Responses	*on* vs. *in* Biased Versions
High	Positive	Expected	57	1 vs. 1
High	Positive	Unexpected	49	2 vs. 1
High	Negative	Expected	30	1 vs. 1
High	Negative	Unexpected	34	2 vs. 1
Low	Positive	Expected	41	1 vs. 2
Low	Positive	Unexpected	40	0 vs. 2
Low	Negative	Expected	31	2 vs. 2
Low	Negative	Unexpected	32	2 vs. 2

Discussion

The severe limitations related to this experiment do not allow us to make any strong claims regarding the observed findings; however, we think that a discussion of these findings is useful for generating hypotheses that can be addressed in future research. The most severe limitations of Experiment 2 involve the unaccounted for variation across the story sets that resulted in significant effects tied to version and the lexical confounds present in particular stories. Another significant limitation of Experiment 2 is a lack of counterbalancing due to the fact that participants were always presented with the same ordering of ratings: control, expectedness, valence, and production.

Very generally, the results of this experiment suggest that the relationships between control, valence, and expectedness may be more complex than we originally anticipated. For example, when stories had positive outcomes, participants' ratings of control may have been influenced more by figure-control than when stories had negative outcomes. It is possible that when situations have a negative outcome, we would prefer to think that the person did not have as much control over the events. Another potential example of these complex relationships is the significant interaction of expectedness and valence. What this interaction may suggest is that when ordinary events are involved, positive outcomes are associated with more figure control; however, when strange occurrences result in positive outcomes, people may be more likely to think that the person has less control over the situation. Interestingly, the pattern seems to switch when going from positive outcomes to negative outcomes. When something bad happens and it is an ordinary event, people may tend to sympathize and not attribute control to the person involved. When things go awry, the abnormality of a situation may cause us to think that the person had more control over the events that took place.

As for the production of *in* and *on*, our results were consistent with Jamrozik and Gentner's (2015) findings: *On* was more preferred when the person was described as having more control and less preferred when the person was described as having less control. We also found evidence to suggest that *on* may be more preferred when outcomes are positive and less preferred when outcomes are negative.

Conclusion

In the domain of spatial semantics the success of multi-componential approaches is clear. Our data is consistent with the hypothesis that non-spatial uses of prepositions have a multi-componential structure like their spatial counterparts. As a case in point, consider the title of this paper. Despite the connections *in* has with negative valence, being in shape is not perceived as significantly worse ($M = 4.57$) than being on top of it ($M = 4.63$). Even though multiple factors of meaning will likely be needed to account for the types of complex patterns we observed in our data, what is less clear is the origin of these factors and the relative impact each of them might have during either comprehension or production. It may be that the spatial and abstract meanings of spatial prepositions share features due to happenstance; however, many cognitive linguists propose that the abstract meanings associated with indirect metaphors are derived via metaphorical connections from their spatially-based meanings (e.g., Brugman & Lakoff, 1988/2006; Tyler & Evans, 2001, 2003). We argue that future research should focus on exploring the implications this type of research has for theories of indirect metaphors, specifically, and lexical semantics, more generally.

Acknowledgments

We are extremely grateful to Michele I. Feist for sharing her insights and recommendations with us during the early stages of this project. We would also like to extend our thanks to members of the Cognition and Psycholinguistics (CaP) Research Lab for their continued support. A special thanks goes out to Mateja Pavlic for her role in helping develop the stories used in Experiment 2. Finally, we would like to thank the reviewers of our original submission. Your feedback was instrumental in shaping this final product.

References

Beitel, D. A., Gibbs, R. W., & Sanders, P. (2001). The embodied approach to the polysemy of the spatial preposition on. In H. Cuyckens & B. Zawada (Eds.), *Polysemy in cognitive linguistics* (pp. 241-260). Philadelphia, PA: John Benjamins.

Bennett, D. (1975). *Spatial and temporal uses of English prepositions.* London: Longmans.

Brugman, C., & Lakoff, G. (2006). Cognitive topology and lexical networks. In D. Geeraerts, R. Dirven, J. R. Taylor, & R. Langacker (Eds.), *Cognitive linguistics: Basic readings* (pp. 109-139). New York, NY: Mouton de Gruyter. (Reprinted from *Lexical ambiguity resolution: Perspectives from psycholinguistics, neuropsychology, and artificial intelligence*, pp. 477-508, by S. L. Small, G. W. Cottrell, & M. K. Tanenhaus, Eds., 1988, San Mateo, CA: Morgan Kaufmann)

Coventry, K. R., Carmichael, R., & Garrod, S. C. (1994). Spatial prepositions, object-specific function and task requirements. *Journal of Semantics, 11*, 289-309. Retrieved from http://www.kennycoventry.org/pdfs/CoventryCarmichaelGarrod1994.pdf

Coventry, K. R., & Prat-Sala, M. (2001). Object-specific function, geometry, and the comprehension of *in* and *on*. *European Journal of Cognitive Psychology*, *13*, 509-528. http://doi.org/10.1080/713752404

Herskovits, A. (1986). *Language and spatial cognition: An interdisciplinary study of the prepositions in English*. Cambridge, UK: Cambridge University Press.

Jamrozik, A., & Gentner, D. (2015). Well-hidden regularities: Abstract uses of *in* and *on* retain an aspect of their spatial meaning. *Cognitive Science*, 39, 1881-1911. doi:10.1111/cogs.12218

Jamrozik, A., & Gentner, D. (2014). Making sense of the abstract uses of the prepositions *in* and *on*. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 2411-2416). Austin, TX: Cognitive Science Society.

Jamrozik, A., & Gentner, D. (2011). Prepositions *in* and *on* retain aspects of spatial meaning in abstract contexts. *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 1589-1594). Austin, TX: Cognitive Science Society.

Leech, G. N. (1969). *Towards a semantic description of English.* London: Longman.

Miller, G., & Johnson-Laird, P. (1976). *Language and perception*. Belknap: Harvard University Press

Rastle, K., Harrington, J., & Coltheart, M. (2002). 358,534 nonwords: The ARC Nonword Database. *Quarterly Journal of Experimental Psychology*, *55A*, 1339-1362. doi:10.1080/02724980244000099

Tyler, A., & Evans, V. (2001). Reconsidering prepositional polysemy networks: The case of *over*. *Language*, *77*(4): 724-765. doi:10.1353/lan.2001.0250

Tyler, A., & Evans, V. (2003). *The Semantics of English Prepositions: Spatial Sciences, Embodied Meaning, and Cognition.* New York: Cambridge University Press.

Children's exploration as a window into their causal learning

Sophie Bridgers (sbridge@stanford.edu)
Department of Psychology, Stanford University

Yvonne Wang (yvonnewjy.wang@mail.utoronto.ca)
Daphna Buchsbaum (buchsbaum@psych.utoronto.ca)
Department of Psychology, University of Toronto

Abstract

How do children's beliefs about a causal system influence their exploration of that system? Children watched an experimenter try to make a machine play music by placing blocks on top; one block always activated the machine and the other block never did (Deterministic condition), or one block activated the machine a higher proportion of times than the other (Probabilistic condition). Subsequently, we measured children's exploratory behaviors without feedback (the machine never activated). We predicted that children in the two conditions would differ in their beliefs about how the system should work, leading to different hypotheses about why the machine was no longer working, and to differential exploration. Compared to the Probabilistic condition, children in the Deterministic condition intervened more often with the previously more effective block, experimented more with how to activate the machine, and explored for less time. Children's exploration provides a rich, nuanced view of their causal reasoning.

Keywords: cognitive development; causal learning; causal uncertainty; statistical learning; exploration

Introduction

Children continually learn about the causal relationships in their environment: Switches turn on lights; germs make people sick; unkind words make people sad. Learning such relationships is not trivial because causal links are not directly observable but rather inferred from statistical contingencies between events. Furthermore, causal relationships are graded in strength – you not only need to learn that a relationship exists but also the probability of the cause generating the effect, which in turn can lead to more or less certainty about the underlying causal structure (Griffiths & Tenenbaum, 2005).

From as young as 24 months, children can infer both the existence of causal links and the relative strength of different causes from deterministic and probabilistic data (Waismeyer, Meltzoff, & Gopnik, 2015; for older children see, Gopnik et al., 2004; Kushnir & Gopnik, 2007; Sobel, Tenenbaum, & Gopnik, 2004). Further, children are more likely to trust testimony that conflicts with probabilistic data (e.g., a block only sometimes makes a machine play music) than testimony conflicting with deterministic data (e.g., a block always makes a machine play music), suggesting sensitivity to the relative strength implied by these different patterns of evidence (Bridgers, Buchsbaum, Seiver, Griffiths, & Gopnik, 2016).

The majority of prior research on children's causal reasoning, however, has examined children's judgments of causal structure. That is, children are prompted to identify which objects are causes and which are not or to produce a single intervention on the system to bring about an effect (e.g., Bridgers et al., 2016; Buchanan & Sobel, 2011; Gopnik et al., 2004; Kushnir & Gopnik, 2005; Kushnir, Wellman, & Gelman, 2008; Sobel et al., 2004). These forced-choice dependent measures provide insight into children's inferences about what is and what is not a cause, but more graded measures could provide additional insight into children's sensitivity to causal strength, especially since causal strength itself is inherently graded. Such measures could also reveal children's certainty or confidence in their inferences. Here, we look to children's exploratory behavior as a window into their causal learning in the hope of gaining a more nuanced view.

Children are sophisticated active learners and acquire knowledge from their own direct exploration of the world (Schulz, 2012; Xu & Kushnir, 2013). Children's exploratory play is not just driven by their enjoyment but also by their inductive inferences and expectations about how the world works. For instance, if children learn that members of a category have an unobservable property, they expect other category members to also have that property and attempt to elicit the property from them (Baldwin, Markman, & Melartin, 1993; Butler & Markman, 2012). The stronger the cues provided to category membership are (e.g., the objects are similar vs. different in appearance; have the same vs. different labels), the longer children persist in their attempts to elicit the property, revealing sensitivity to gradations in the inductive strength of these cues (Baldwin et al., 1993; Schulz, Standing, & Bonawitz, 2008). Children also explore more when evidence is ambiguous or an event challenges their prior beliefs. They opt to play with causally confounded or belief-violating toys over novel toys (Schulz & Bonawitz, 2007; Stahl & Feigenson, 2015). In this play, they even spontaneously design novel interventions or experiments to test their beliefs, and generate sufficient evidence to disambiguate the causal system (E. B. Bonawitz, van Schijndel, Friel, & Schulz, 2012; Gweon & Schulz, 2008; Legare, 2011; Schulz & Bonawitz, 2007). Taken together, childrens certainty and surprise appear to influence how much and how long they explore, making exploration a good dependent measure of their underlying beliefs and confidence in those beliefs.

Here, we take up the hypothesis that children's free exploration of a causal system is supported by some of the same rational principles of inductive inference that inform their explicit causal judgments (e.g., Schulz, Standing, & Bonawitz, 2008). We examine children's exploration in the context of causal uncertainty, and in particular investigate how their exploration might differ after observing the deterministic vs. probabilistic patterns of evidence often used in classic ex-

periments on causal judgments. We predicted that children would be sensitive to differences in causal strength implied by deterministic vs. probabilistic data, and that this would be reflected in their free play with a causal system. Using exploration as a dependent measure has the additional benefit that it is non-verbal and so could be particularly useful in measuring young children's certainty in their inferences; explicit meta-cognition is still developing in early childhood, making it difficult to elicit explicit certainty judgments from young children (Ghetti, Hembacher, & Coughlin, 2013; Hembacher & Ghetti, 2014).

A feature common to experiments using exploration as a dependent measure is that children are presented with objects that lack the previously observed property. This design decision prevents children from eliciting confirmatory evidence and eliminates the distraction of the interesting causal property, to more easily isolate how children's expectations affect their play (see Baldwin et al., 1993; Schulz, Standing, & Bonawitz, 2008). However, it also raises the question of what conclusions children are drawing about their failed attempts to elicit the property. This failure could be due to the causal system (i.e., it does not work as expected or has stopped working) or due to one's own actions (i.e., I am doing something wrong) (Karmiloff-Smith & Inhelder, 1974), a distinction to which even 16-month-old infants are sensitive when responding to their own failed actions (Gweon & Schulz, 2011). Thus, children's exploration may not only reflect uncertainty in their prior inferences about the causal system but also uncertainty about why the system is no longer working, making it a useful measure of both their initial inductive predictions and how such predictions inform later inferences about unexpected outcomes. Next, we present our specific experimental hypotheses about how children's sensitivity to these sources of uncertainty might manifest in their exploration of an inert causal system.

Overview

In the current experiment, four- and five-year-olds were introduced to a machine that could be activated by placing blocks on top. Children observed either deterministic or probabilistic evidence that one of two blocks was better than the other at making the machine go. Children were then given the opportunity to explore the blocks and the machine on their own, but during this time, the machine never activated no matter what interventions children performed.

We predicted that children's exploration would reflect both their initial inferences from the demonstrated data about how the system works and their subsequent inferences about why it was no longer working (Legare, 2011; Schulz, Hooppell, & Jenkins, 2008). In both conditions, we predicted children would first attempt to activate the machine with the block that was demonstrated to be more causally efficacious. However, when faced with evidence that this block was no longer working, children would explore differently across conditions.

Children in the Deterministic condition will likely develop a strong expectation that the previously more effective block should work (it always did before) and that the the previously less effective block should not (it never did before), while children in the Probabilistic condition should have less certainty about the causal strength of each block (both blocks previously succeeded and failed in activating the machine). We thus predicted that children in the Deterministic condition would persist in trying to activate the machine with the previously better block more than children in the Probabilistic condition, who would be more likely to explore the previously worse block.

The stronger belief in the previously better block's efficacy might also lead children in the Deterministic condition to infer that the block is no longer working because they are doing something wrong. If that is the case, children might not only persist in trying this block but also be more likely to experiment with different ways of activating the machine (e.g., placing the block in different locations on the machine) to try and find the right way to use the block (Legare, 2011). However, this stronger initial belief in the better block's causal strength might also lead children to give up more quickly, because of a belief that they are still doing something wrong (e.g., not placing the block in the right location) or that the system has somehow changed (e.g., it is out of batteries). In contrast, children in the Probabilistic condition might explore longer overall, but produce less variable interventions. Since the system is stochastic, if it is not activating it is not necessarily because the system has stopped working or that they are doing something wrong, so it is worth continuing to test out the blocks.

Prior work has actually suggested that children engage in more variable exploration when presented with ambiguous or probabilistic evidence (see E. Bonawitz et al., 2011; Schulz, Hooppell, & Jenkins, 2008). Here, we predict that the probabilistic evidence will result in more even exploration of both blocks and longer exploration overall. In contrast to previous work, we anticipate that the deterministic evidence will render the inert system more surprising and so will lead to more novel interventions to try to figure out why this is the case.

Given these predictions, we not only look at children's overall persistence and exploration time, but also at which blocks children place on the machine and what else they try with the blocks and the machine to see if these behaviors also reflect different causal inferences and sources of uncertainty.

Methods

Participants

Seventy-seven children (41 4-year-olds and 36 5-year-olds; 38 females) were recruited from local museums in Toronto, Ontario. An additional 10 children were tested but excluded from analysis due to experimenter error ($n = 5$), missing date of birth ($n = 2$), or ending the experiment early ($n = 3$). Children were randomly assigned to the Deterministic condition ($n = 39$, $M(SD) = 59.43(6.30)$ months, 19 females) or the Probabilistic condition ($n = 38$, $M(SD) = 58.96(7.56)$ months, 19 females). The diversity of the sample was repre-

sentative of the diversity of the local population.

Materials

The causal system was presented to children as a machine that could play music when blocks were placed on top. The "machine" was a decorated cardboard box as shown in Figure 2A. There were four wooden blocks, differing in shape and color, but similar in size: the red oval, yellow square, blue triangle, and purple star blocks. The machine appeared to play music when blocks were placed on top, but in reality, it contained a wireless door bell that could be activated surreptitiously by the experimenter via a hidden remote control. For some children, a bell they could ring to indicate that they were done exploring ($n = 36$) or a distractor toy ($n = 2$) were also placed on the table.[1]

Procedure

Participants were tested individually in a quiet off-exhibit location at the museum. Children's behavior was video recorded and coded offline by the second author.

Demonstration Phase The experimenter first introduced the child to the novel machine, explained that you could make the machine play music by putting blocks on top and that some blocks made it play music while others did not. The experimenter then brought out a pair of blocks, either the yellow square and red oval blocks, or the blue triangle and purple star blocks (counterbalanced across participants) and told children that she had never played with these blocks before.

Next, the experimenter demonstrated each block on the machine. In the Deterministic condition, each block was placed on the machine six times; one block deterministically activated the machine on all six trials (*better* block), while the other block failed to activate the machine on all six trials (*worse* block). In the Probabilistic condition, one block was placed on the machine three times and activated the machine twice, always on the first and third trial (*better* block); the other block was placed on the machine six times and also activated it two times, on the second and fifth trial (*worse* block). As in previous work, this pattern of data controls for the frequency with which each block activates the toy, providing stronger evidence that children are reasoning about the probability of each cause generating the effect and not just the number of times the effect was associated with the cause (Bridgers et al., 2016; Kushnir & Gopnik, 2007). We counterbalanced which block was the better block and whether the better block was demonstrated first.

The experimenter then asked the child which block was better at making the machine play music. If the child did not answer or chose both blocks, the experimenter asked them to select one. The block children pointed to, touched, or named was coded as their answer. If children answered incorrectly, the experimenter corrected them (e.g., "Oh remember, the red oval block was better at making the machine play music.")

Exploration Phase The experimenter then informed the child that she had to leave for a bit, and that the child could play freely with the blocks and the machine while she was gone. The experimenter left the table and pretended to be busy in another part of the room. During this time, the child could explore the blocks and machine but did not receive any feedback, i.e., neither block activated the machine. If the child asked the experimenter questions, she explained that she was still working but the child could keep playing with the blocks and machine, and let the experimenter know when they were done. The experimenter returned when the child indicated they were done or after two minutes.[2] Lastly, the child was given an opportunity to activate the machine with a new pair of blocks to ensure they left in good spirits.

Exploration Phase Coding There were three main variables of interest. First, we measured the proportion of interventions children performed with the better block, out of the total interventions they made with just a single block (*single trials*). A single trial was when children placed a block that was off of the machine onto the machine, or lifted and then put back down part or all of a block that was already on the machine. Each single trial was coded according to whether the *better* or *worse* block was used. Only children who placed a block on the machine were included in this analysis (Deterministic condition: $n = 33$; Probabilistic condition: $n = 35$). A small number of trials (3.4%) in which children placed both blocks on the machine simultaneously were not included, but were considered a strategy as described below.

Second, we measured the total time in seconds children spent exploring the blocks and machine. Exploration began when children first touched a block or the machine and ended when they met any of the following criteria: They (1) explored for two minutes; (2) indicated that they were done; (3) did not interact with the blocks or machine for 15 seconds (end time was coded as the last second they touched the block and/or the machine); (4) only played with the blocks off of the machine for 15 seconds (end time was coded as the last second when they removed the block(s) from the machine). Children who never put the blocks on the machine and only played with the blocks off of the machine, were coded as having an exploration time of one second. Children who did not interact with the blocks or machine at all even after additional

[1] We initially experimented with how children could indicate that they were done. Children were either instructed to verbally alert the experimenter ($n = 39$), given a fun distractor toy they could switch to ($n = 2$), or told that the experimenter would bring out new blocks when she returned and given a bell to ring when they were done exploring ($n = 36$; we kept this latter version for our preregistered replication of this pilot experiment). These approaches were evenly distributed across conditions.

[2] We were initially concerned children might feel they had to explore until the experimenter returned and so would continue due to normative pressures rather than interest. Thus, for a subset of children (n = 23), evenly distributed across conditions, the experimenter checked-in prior to 2 minutes if they stopped exploring for 5-10 seconds, i.e., before their exploration had otherwise ended. Most children, however, did stop exploring before 2 minutes had passed, so we decided to remove these check-ins for the remaining children in this experiment and the children in our preregistered replication.

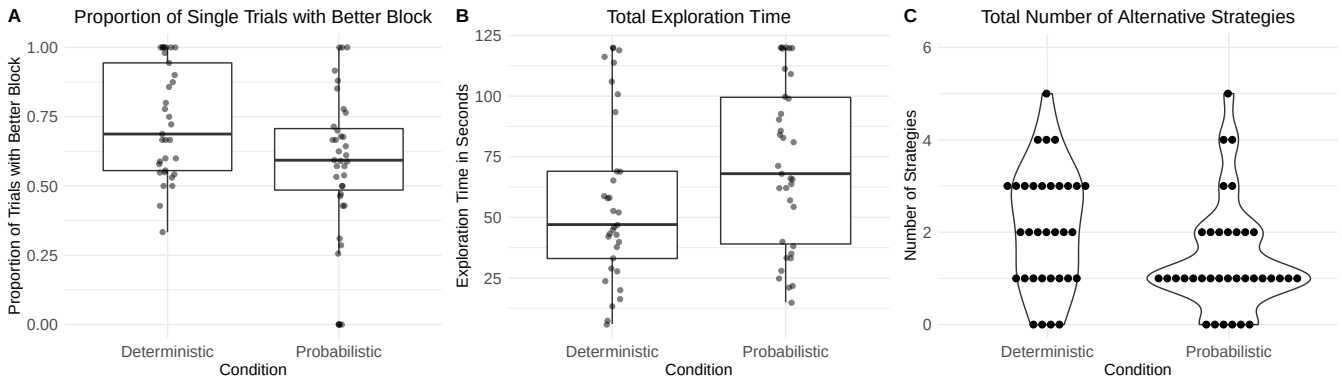


Figure 1: (A) Proportion of single trials children performed with the better block across conditions. (B) The amount of time children explored the blocks and/or the machine across conditions. Median and 1st and 3rd quartiles are displayed. (C) The total number of alternative strategies children performed across conditions. For all plots, dots are individual children.

encouragement from the experimenter were recorded as having an exploration time of zero seconds. These children did not indicate that they wanted to end the experiment; the experimenter also made clear through prompting that they could place blocks on the machine, so we were confident they understood the instructions and interpreted their lack of exploration as a choice not to explore, rather than confusion about the task. Nine children did not place a block on the machine, and this tendency did not differ by condition (Deterministic: $n = 6$; Probabilistic: $n = 3$; two-tailed Fisher's Exact test, $p = 0.481$). The end of exploration was determined offline according to the above criteria.

Third, we noticed that during children's exploration, they indeed performed alternative actions that were not demonstrated by the experimenter. We considered these different actions experimentation or alternative strategies children were employing to try to activate the machine. We identified seven different types of strategies: (1) exploring the machine alone (e.g. knocking on or poking it; flipping it over), (2) exploring the blocks alone (e.g., tapping the blocks together off of the machine or on the table), (3) flipping a block over to try a different side, (4) placing the block in a different location on the machine, (5) placing both blocks on the machine, (6) dropping the blocks onto the machine from above, and (7) applying force when placing the blocks on the machine (See Figure 2A). Children received a score of 1 for each strategy type they produced and a 0 otherwise (i.e., children could be coded as exhibiting 0 to 7 different strategies). Note that this is not a measure of how many times children produced a strategy but rather a count of how many different kinds of strategies children exhibited. Only children who interacted with the machine at some point were additionally coded for strategies (Deterministic condition: $n = 33$; Probabilistic condition: $n = 35$; same children as those included in analysis of the proportion of single trials with the better block).

Results

Consistent with previous work, in response to the explicit question about which block was better at activating the machine, the majority of children correctly selected the *better* block (two-tailed Binomial test, $p < 0.001$; Deterministic: $37/39$; Probabilistic: $33/38$; two-tailed Fisher's Exact test comparing across conditions: $p = 0.263$). This was also true when looking only at children who later placed a block on the machine during exploration (two-tailed Binomial test, $p < 0.001$; Deterministic: $31/33$; Probabilistic: $32/35$; two-tailed Fisher's Exact test across conditions: $p = 1$). These children were similarly more likely to first intervene with the better block, rather than the worse block (two-tailed Binomial test, $p < 0.001$; Deterministic: $22/33$; Probabilistic: $24/35$; two-tailed Fisher's exact test across conditions: $p = 1$). Intriguingly, these children were more likely to identify the better block as the better cause ($63/68$) than to select it first to place on the machine ($46/68$; two-tailed Fisher's Exact test, $p < 0.001$), suggesting they may have had motivations other than maximizing the probability of activating the machine when they first intervened.

To compare the proportion of single trials on which children used the better block, the total time children explored, and the total number of strategies children exhibited across conditions, we conducted three one-way ANCOVAs with condition as a factor and age in months as a covariate.

Children in the Deterministic condition intervened with the better block on a higher proportion of single trials than children in the Probabilistic condition ($M \pm$SE = 0.73 ± 0.035 v. 0.59 ± 0.041, respectively; $F(1,65) = 6.17$, $p = .016$), and this tendency to intervene with the better block did not differ by age ($F(1,65) = 0.23$, $p = .635$). Children in the Deterministic condition, however, explored for a shorter amount of time than children in the Probabilistic condition ($M \pm$SE = 48.60 ± 6.14 seconds v. 66.40 ± 6.14 seconds, respectively). This difference was significant ($F(1,74) = 4.37, p = .040$), and the length of time children explored did not differ by age ($F(1,74) = 1.26, p = .264$). If we only consider the children who placed a block on the machine, the difference in exploration time is trending (Deterministic: $M \pm$SE = 57.33 ± 6.12 seconds; Probabilistic: 71.89 ± 5.76 seconds; $F(1,65) = 3.07$, $p = .084$). (See Figure 1A-B.)

Most children employed at least one alternative strategy in their attempts to make the machine play music (Determin-

istic: 29/33; Probabilistic: 29/35), and this tendency did not differ across conditions (two-tailed Fisher's exact test, $p = .735$). However, children in the Deterministic condition employed more strategies overall than children in the Probabilistic condition ($M \pm$SE = 2.09 ± 0.23 v. 1.43 ± 0.20, respectively; $F(1,65) = 4.74, p = .033$); children's overall tendency to perform these alternative actions did not differ by age ($F(1,65) = 0.18, p = .670$). (See Figure 1C.)

Looking at the different strategies separately reveals that the modal strategy in the Deterministic condition was flipping the blocks over and in the Probabilistic condition, placing the blocks in different locations on the machine. Roughly the same number of children in both conditions changed the location of blocks (16 in Deterministic and 15 in Probabilistic) but about twice as many children flipped blocks over in the Deterministic than in the Probabilistic condition (21 v. 10 respectively), and this difference was significant (two-tailed Fisher's Exact test, $p = .007$). We are cautious to draw conclusions about the less common strategies since so few children exhibited them overall, but they do provide additional suggestive evidence that the conditions not only differed in the total number of strategies children exhibited but also in which strategies children employed. (See Figure 2B.)

Discussion

We provide evidence that children's exploratory behaviors can serve as a graded and detailed window into their causal reasoning. We find that presenting children with covariation information that deterministically or probabilistically supports a particular causal system leads them to explore this system differently, reflecting different inferences about causal strength and the uncertainty inherent in these inferences.

Children in both conditions drew rational inferences from the evidence they observed about the relative causal strength of the more effective (better) block compared to the less effective (worse) block. Children who previously observed deterministic evidence for the blocks' effectiveness attempted to activate the machine with the better block, rather than the worse block, more often than children who observed probabilistic evidence, suggesting a stronger inference that this block should work. Children's differential exploration across conditions suggests they did not simply draw a binary inference about which cause was better but rather were sensitive to the relative magnitude of causal strength.

Children's exploration also provided a richer picture of their causal inferences in this task than their causal judgments prior to the exploration phase. The causal judgments revealed that children had correctly inferred which block was more effective. If we had only considered this measure, however, we would not have seen that children were differentiating between the evidence presented in each condition and correctly retaining more uncertainty in the more ambiguous, probabilistic case. Similarly, if we had only looked at children's first intervention on the machine, we would have lacked the sensitivity to pick up on differences across conditions.

Interestingly, only about two-thirds (68%) of children across conditions first intervened with the better block, though over 90% explicitly identified this block as more effective. This is particularly surprising in the Deterministic condition; in prior work using similar or even weaker patterns of deterministic evidence, when explicitly asked to select a block to make the machine go, children overwhelmingly tended to intervene with the more effective block (e.g., Sobel et al., 2004; Sobel, Sommerville, Travers, Blumenthal, & Stoddard, 2009). Children's first intervention in our task suggests they were not simply motivated to generate the effect but rather to explore from the get go (e.g., perhaps some children wanted to understand why the worse block did not work or see if they could make it work). Direct questions likely place pressure on children to respond correctly, while self-directed exploratory play is more open-ended, removing such pressures and potentially revealing different behavior.

Prior work shows that children explore more when presented with events that violate their expectations. In many of these studies, including the present, children are not just faced with evidence that the causal system works differently than how they predicted but that it does not work at all. What inferences do children draw about the source of their own failed actions and how might their exploration reflect these inferences? Children in the Deterministic condition may have thought the problem lay with them since the evidence demonstrated by the experimenter strongly suggested that one block made the machine go and the other did not. These children tended to exhibit a wider variety of alternative actions, suggesting they may indeed have interpreted their failures to elicit music from the machine as the fault of their own actions rather than the causal system. Children in the Probabilistic condition, however, could explain away the lack of activation due to the system's stochasticity. Many did try at least one alternative strategy but overall did not experiment as much as children in the Deterministic condition, suggesting they did not necessarily think the problem lay with them but rather with the system itself. Indeed the one alternative strategy more children in the Probabilistic condition produced was exploring the machine on its own.

Though children in the Deterministic condition tried more strategies, they actually appeared to give up more quickly, exploring for a shorter amount of time overall than children in the Probabilistic condition. One possible explanation for this difference, consistent with why children in the Deterministic condition may have experimented more, is that the lack of machine activation led them to conclude more quickly that they were incapable of activating the machine either because they just could not figure out how or because it had stopped working. In contrast, children in the Probabilistic condition may have explored longer because they continued believing the machine might activate.

Taken together, these different aspects of children's exploration – their tendency to explore the better block, to experiment with different ways of activating the machine, and total

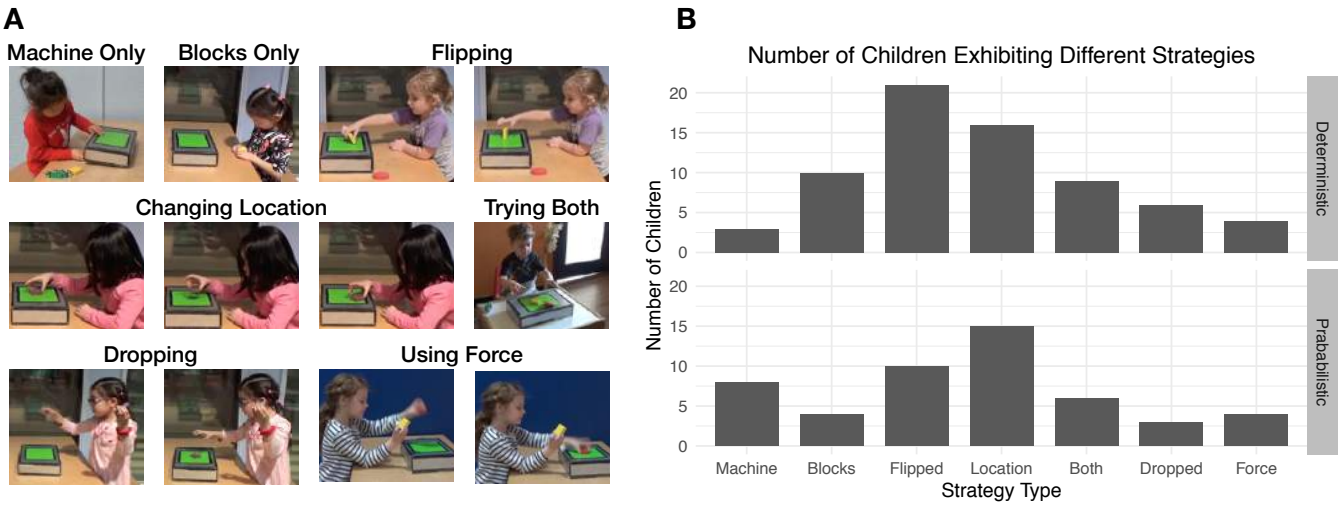


Figure 2: (A) Examples of alternative actions or strategies children performed to try to activate the machine. (B) The total number of children exhibiting each of the different strategies across conditions.

length of exploration – are suggestive of children integrating different sources of uncertainty. This includes uncertainty about the causal system itself (which block is better, and how much better) and uncertainty about the source of failure ("Is it me or the world?"; Gweon & Schulz, 2011).

In the deterministic case, children appear to persist in believing that the better block should work, so something must be wrong either with their own actions or with the blocks and the machine (e.g. it ran out of batteries). This behavior suggests continued certainty about how the causal system should work (even after receiving negative evidence about the previously better block, they still do not think it is likely that the previously worse block will work) but higher uncertainty about what has gone wrong (resulting in them trying more strategies to see if they can fix it).

In the probabilistic case, children appear less certain about the causal relationships. The better block continuing to not work is less surprising, because the system is stochastic. Children therefore demonstrate high uncertainty about the strength of the blocks and probability of activation (as evidenced by trying both blocks more evenly, rather than favoring the previously better one) but lower uncertainty about the source of failure (as evidenced by fewer strategies and overall more persistence).

Decades ago, Karmiloff-Smith and Inhelder (1974) argued that children's exploration is not just driven by their prior implicit theories but also by the evidence they generate as they explore, and that their failures to bring about an expected outcome would be interpreted as either relevant to their theory or to their action. Moving forward, looking at the time course of exploration (e.g., how early exploration differs from later exploration) could provide more compelling evidence for how children's inferences evolve as they accumulate evidence of the system's failure. For example, do the type of strategies children employ change over time? Prior work indicates children this age can design informative interventions to disambiguate causal evidence (Cook, Goodman, & Schulz, 2011). Are children's alternative strategies indeed targeting different hypotheses across conditions about why the machine is not working? In the present study, there are too few children who perform any particular strategy to probe these questions in more detail, but we are currently running a larger scale replication with over 100 children to address these questions (see `osf.io/sc54w` for preregistration).

Children's experimentation also raises interesting questions about conditions that lead to innovation. Along certain dimensions, and consistent with prior work (e.g., E. Bonawitz et al., 2011; Schulz, Hooppell, & Jenkins, 2008), deterministic evidence seemed to constrain children's exploration: they were less likely to explore the block another person had demonstrated to be less effective and explored for a shorter total amount of time than children in the Probabilistic condition. Along other dimensions however, they appeared to be more exploratory and innovative. They were more likely to experiment with how they placed blocks onto the machine. Children's ability to generate alternative means for achieving a goal and flexible problem solving in the face of failed actions is an interesting avenue for future work.

Exploration is a powerful, ecologically valid dependent measure that is more sensitive than binary questions and does not rely on children's language skills or explicit introspection. It does come with limitations; it is indirect and influenced by other factors besides children's beliefs. The use of open-ended, dynamic measures such as exploration, however, in conjunction with direct questions will allows us to paint a richer, more graded picture of children's inferences, as well as offer the potential of investigating how these inferences might change across time and affect behavior. Just as children harness the power of their exploratory play to learn about the world, we, as scientists, can harness this same play to learn more about what and how children learn.

Acknowledgments

We would like to thank Katrina Palad, Maureen Huang, Sandy Chao, and Aaron Philipp-Muller for their help with data collection and coding. We also thank the meta-reviewer and three reviewers of this paper for their thoughtful and constructive feedback. This research was supported by the Social Sciences and Humanities Research Council of Canada [435-2018-0890], and the Canada Foundation for Innovation.

References

Baldwin, D., Markman, E., & Melartin, R. (1993). Infants' ability to draw inferences about nonobvious object properties: evidence from exploratory play. *Child Development*, *64*(3), 711–728.

Bonawitz, E., Shafto, P., Gweon, H., Goodman, N. D., Spelke, E., & Schulz, L. (2011). The double-edged sword of pedagogy: Instruction limits spontaneous exploration and discovery. *Cognition*, *120*(3), 322–330.

Bonawitz, E. B., van Schijndel, T. J. P., Friel, D., & Schulz, L. (2012). Children balance theories and evidence in exploration, explanation, and learning. *Cognitive Psychology*, *64*(4), 215–234.

Bridgers, S., Buchsbaum, D., Seiver, E., Griffiths, T. L., & Gopnik, A. (2016). Children's causal inferences from conflicting testimony and observations. *Developmental Psychology*, *52*(1), 9–18.

Buchanan, D. W., & Sobel, D. M. (2011). Mechanism-based causal reasoning in young children. *Child Development*, *82*(6), 2053-2066.

Butler, L. P., & Markman, E. M. (2012). Preschoolers Use Intentional and Pedagogical Cues to Guide Inductive Inferences and Exploration. *Child Development*, *83*(4), 1416–1428.

Cook, C., Goodman, N. D., & Schulz, L. E. (2011). Where science starts: spontaneous experiments in preschoolers' exploratory play. *Cognition*, *120*(3), 341–349.

Ghetti, S., Hembacher, E., & Coughlin, C. A. (2013). Feeling uncertain and acting on it during the preschool years: A metacognitive approach. *Child Development Perspectives*, *7*(3), 160–165.

Gopnik, A., Glymour, C., Sobel, D., Schulz, L., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: causal maps and Bayes nets. *Psychological Review*, *111*(1), 3–32.

Griffiths, T., & Tenenbaum, J. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*(4), 334–384.

Gweon, H., & Schulz, L. (2008). Stretching to learn: Ambiguous evidence and variability in preschoolers' exploratory play. *Proceedings of the 30th annual meeting of the Cognitive Science Society*, 570–574.

Gweon, H., & Schulz, L. (2011). 16-Month-Olds Rationally Infer Causes of Failed Actions. *Science*, *332*(6037), 1524–1524.

Hembacher, E., & Ghetti, S. (2014). Dont look at my answer: Subjective uncertainty underlies preschoolers exclusion of their least accurate memories. *Psychological Science*, *25*(9), 1768–1776.

Karmiloff-Smith, A., & Inhelder, B. (1974). If you want to get ahead, get a theory. *Cognition*, *3*(3), 195–212.

Kushnir, T., & Gopnik, A. (2005). Young children infer causal strength from probabilities and interventions. *Psychological Science*.

Kushnir, T., & Gopnik, A. (2007). Conditional probability versus spatial contiguity in causal learning: Preschoolers use new contingency evidence to overcome prior spatial assumptions. *Developmental Psychology*, *43*(1), 186–196.

Kushnir, T., Wellman, H., & Gelman, S. (2008). The role of preschoolers' social understanding in evaluating the informativeness of causal interventions. *Cognition*, *107*(3), 1084–1092.

Legare, C. H. (2011). Exploring Explanation: Explaining Inconsistent Evidence Informs Exploratory, Hypothesis-Testing Behavior in Young Children. *Child Development*, *83*(1), 173–185.

Schulz, L. (2012). The origins of inquiry: inductive inference and exploration in early childhood. *Trends in Cognitive Sciences*, *16*(7), 382–389.

Schulz, L., & Bonawitz, E. (2007). Serious fun: Preschoolers engage in more exploratory play when evidence is confounded. *Developmental Psychology*, *43*(4), 1045–1050.

Schulz, L., Hooppell, C., & Jenkins, A. (2008). Judicious imitation: children differentially imitate deterministically and probabilistically effective actions. *Child Development*, *79*(2), 395–410.

Schulz, L., Standing, H., & Bonawitz, E. (2008). Word, thought, and deed: the role of object categories in children's inductive inferences and exploratory play. *Developmental Psychology*, *44*(5), 1266–1276.

Sobel, D., Sommerville, J. A., Travers, L. V., Blumenthal, E. J., & Stoddard, E. (2009). The role of probability and intentionality in preschoolers' causal generalizations. *Journal of Cognition and Development*, *10*(4), 262-284.

Sobel, D., Tenenbaum, J. B., & Gopnik, A. (2004). Children's causal inferences from indirect evidence: Backwards blocking and Bayesian reasoning in preschoolers. *Cognitive Science*, *28*(3), 303–333.

Stahl, A. E., & Feigenson, L. (2015). Observing the unexpected enhances infants' learning and exploration. *Science*, *348*(6230), 91–94.

Waismeyer, A., Meltzoff, A. N., & Gopnik, A. (2015). Causal learning from probabilistic events in 24-month-olds: an action measure. *Developmental Science*, *18*(1), 175–182.

Xu, F., & Kushnir, T. (2013). Infants Are Rational Constructivist Learners. *Current Directions in Psychological Science*, *22*(1), 28–32.

Elicitation of Quantified Description Under Time Constraints

Gordon Briggs (gordon.briggs@nrl.navy.mil)
Christina Wasylyshyn (christina.wasylyshyn@nrl.navy.mil)
Paul F. Bello (paul.bello@nrl.navy.mil)
Navy Center for Applied Research in Artificial Intelligence
U.S. Naval Research Laboratory
4555 Overlook Avenue SW
Washington, DC USA

Abstract

Quantity can be expressed in a variety of ways and at different levels of precision. One factor that influences numerical description of elements in a visual scene is how long the scene is observed. We extend a previous incremental model of numerical perception to model quantified description under time constraints. Our extended model predicts that as presentation duration decreases and as the quantity of items to be enumerated increases, the frequency of inexact quantifiers will increase. We conducted two human subject elicitation studies to test these predictions. Our findings were consistent with our model's predictions. Additionally, we demonstrate that our novel model of incremental numerical perception and quantified description closely predicts the precise proportion of exact numerical responses generated by in these experiments.

Keywords: numerical language; numerical perception; quantifiers; subitizing; counting; estimation; computational model

Introduction

Quantity can be expressed in a variety of ways and at different levels of precision. Speakers can use exact numbers to describe quantities (e.g., "there are *twenty-two* guests at the party") or they can use more vague language (e.g., "there are *many* guests at the party"). Many factors influence the form and degree of precision of quantified language a speaker uses, including pragmatic considerations (Cummins, 2015). For example, speakers can express quantities in strategically vague ways for the purpose of influencing behavior (Hesse & Benz, 2018).

In the context of visual scene description, another factor influences quantified language: how long the scene is observed. Research in numerical perception suggests that mental representation of visual quantity is incrementally acquired through temporally extended and attentionally-dependent perceptual processes (Trick & Pylyshyn, 1994; Railo, Koivisto, Revonsuo, & Hannula, 2008). Glancing at a scene allows one to form a less precise representation of quantity, while taking the time to count each relevant item gives one a precise numerical representation of quantity. Using exact numbers when insufficient time has been devoted to complete enumeration often results in incorrect numerical guesses, and psychologists studying numerical perception often rely on analysis of error patterns during exact number elicitation tasks to make inferences about underlying processes and representations (e.g., Mandler & Shebo, 1982).

However, outside psychophysics experiments, people are rarely forced to express themselves using only exact numerical expressions. Some recent work has begun to examine patterns of quantified language usage in more unconstrained situations. In particular, Barr, Deemter, and Fernández (2013) found that when individuals produce quantified reference expressions (QREs), the form of the QRE was dependent on the numerosity of the sets under consideration. When the quantities in each set were large, people tended to produce relational expressions (e.g., "my set is the largest one"). However, when the target set of objects consisted of a small quantity, people tended to produce QREs with exact numerical descriptors despite inexact QREs being sufficient to disambiguate the expression. This result suggests that people balance pragmatic concerns of informativity with minimization of perceptual effort or cost.

The contribution of this paper is two-fold. First, we extend a previous incremental model of numerical perception to model quantified description under time constraints. This provides an explicit model of perceptual cost that was lacking in prior literature. The second contribution of this paper is two novel human subject elicitation studies designed to test the predictions generated by this model. Our extended model predicts that as presentation duration decreases and as the quantity of items to be enumerated increases, the frequency of inexact quantifiers will increase. The findings from the experiments were consistent with our model's predictions. Additionally, we demonstrate that our novel model of incremental numerical perception and quantified description closely predicts the precise proportion of exact numerical responses generated by in these experiments.

Computational Models of Numerical Perception

The perception of numerosity consists of multiple processes, each occurring at different rates and resulting in mental representations of varying precision. Explicit counting provides a slow, but precise, determination of number (Gelman & Gallistel, 1986) rooted in linguistic representation in a phonological buffer (Whalen, Gallistel, & Gelman, 1999). Estimation provides a rapid but less precise judgment of the quantity of a group of objects (Barth, Kanwisher, & Spelke, 2003) rooted in what has become known as the approximate number system (ANS) (Dehaene, 2011). Between these two procedures, a third process, called subitizing, provides both rapid and pre-

cise judgments of numerosity, but only for small quantities, from one to typically around four objects (Kaufman, Lord, Reese, & Volkmann, 1949). Consequently, the range of numerosities between one and four has become known as the subitizing range. While debate still continues about the representations and processes underlying subitizing, there are converging lines of evidence that suggest that the object-tracking system (OTS) plays a central role (Feigenson, Dehaene, & Spelke, 2004).

Recently, there has been renewed interest in developing neural models of numerical perception. These models typically focus on accounting for only a single process and form of representation, such as estimation and the ANS (Chen, Zhou, Fang, & McClelland, 2018) or counting and exact number (Fang, Zhou, Chen, & McClelland, 2018). Some researchers have begun to examine the generation of quantified descriptions of visual scenes with varying levels of precision (Pezzelle, Marelli, & Bernardi, 2017). However, these models also generally do not attempt to model the time course of enumeration. The psychophysical literature on numerical perception has shown that within the subitizing range, each additional object requires only 40–100 ms to accurately enumerate, while outside the subitizing range, each additional object requires 250–350 ms to enumerate (Trick & Pylyshyn, 1994). Most existing computational models of numerical perception do not attempt to capture this aspect of enumeration, nor do they provide accounts for how estimates can be refined with additional time.

An Incremental Model

In contrast with these previous models, Briggs, Bridewell, and Bello (2017) developed a computational model, implemented in the ARCADIA cognitive system (Bridewell & Bello, 2016), that models temporally extended numerical perception and integrates various forms of numerical representation. The model contains components that implement three distinct numerical representation systems: the ANS, the OTS, and the phonological buffer. We will denote this model as `INP-Guess` (incremental numerical perception and guessing).

`INP-Guess` operates by first ascertaining an approximate, noisy estimate of quantity by deploying visual attention toward the entire group of items to be enumerated.[1] Subsequently, serial attention is deployed to each individual item in the group. This process of serial attention first fills up the visual short-term memory (vSTM) slots in ARCADIA's object-tracking system. If there are no more items to be enumerated or no more available slots in vSTM, then a lexical representation of the quantity of relevant items in vSTM is subvocalized within the system's phonological buffer. After this point, subsequent serial focus to new items is accompanied by subvocalization of the next count word in the counting sequence.

[1] We refer the reader to the original model paper for details about how visual attention is realized in the ARCADIA system.

Numerical Guessing. If the visual scene ends, or enumeration otherwise ends, the model merges both the results from the ANS and the lexicalized count into a single numerosity judgment. If time allows for an explicit count to be fully generated (i.e., all items had received individual attentional focus), the explicit count is recorded. Otherwise, an educated guess is made:

$$Guess(n_c, n_e, w) = n_c + sample(\mathcal{N}(n_e - n_c, \sqrt{w \cdot (n_e - n_c)}))$$

where w denotes the Weber fraction of the ANS, n_e denotes the number of items that collectively received attentional focus during estimation, and n_c denotes the number of items that received individual focus during subitizing and counting.

As more items are individually attended to, an exact representation of a lower bound on the number of visual items in the scene increases. The equation above reflects the variance of the noisy numerosity representation upon which linguistic description is based decreasing as this lower bound increases. Note, we are not proposing that serial deployment of attention directly affects the variance of the representation produced by the ANS. Rather, what we are proposing is that the partial exact, lexical representation of number and the noisy ANS representation are merged (Briggs, Bridewell, & Bello, 2017), such that the resulting merged representation of numerosity will have lower variance when more items have received serial focus of attention. If there is enough time to devote attention to each item individually, then $n_c = n_e$ and guess is equal to the lexicalized count n_c.

While the precise time T_{attend} it takes to fully attend to n items individually within the `INP-Guess` model depends on multiple task-related factors, we can formulate a mathematical approximation of the time required in a simple case (i.e., a single-task involving enumeration of all items in a visual scene):

$$T_{attend}(n) \approx T_f \cdot min(r_s, n) + \prod_{max(r_s,n) \leq i \leq n} T_{subvocal}(i) + T_f$$

where T_f denotes the time necessary to attend to encode a single item into vSTM, r_s denotes the subitizing limit, and $T_{subvocal}(i)$ denotes the time necessary to subvocalize the i-th count word. Based on the original parameter values used by Briggs and colleagues (2017), we set the following values: $T_f = 50$ms, $r_s = 4$. Subvocalization time $T_{subvocal}(i)$, varies by number and is based on the formula from Huss and Byrne (2003).

Therefore, the number of items n_c that can be individually attended to in `INP-Guess` within a time window of T can be approximated as:

$$n_c(T) \approx \underset{i \geq 0}{\text{argmax}} \{i | T_{attend}(i) \leq T - T_{estimate}\}$$

where $T_{estimate}$ denotes the time necessary for the initial estimation of quantity within the visual scene.

Briggs and colleagues (2017) demonstrated that the `INP-Guess` model could account for the bilinear reaction time

curve in enumeration (Trick & Pylyshyn, 1994). Additionally, the INP-Guess model could account for the pattern of error in studies of subitizing during conditions of divided attention (Railo et al., 2008).

However, while being able to capture the pattern of error in numerical guessing during tasks with time and attentional constraints is desirable in a model of numerical perception, it is an incomplete account of quantified language use. When speaking with one another, people are faced with a variety of communicative norms. Communicating exact numbers in cases of noisy numerical representations would likely violate these norms, including prohibitions against communicating without adequate evidence (Grice's Maxim of Quality) and failure to be informative about one's own certainty (Grice's Maxim of Quantity) (Grice, 1975). The use of inexact quantified description (e.g., "there are between 4 to 7 items", "there are about 6 items", etc.) is one way to satisfy these communicate norms. While the wide range of quantified language provides ample opportunity for investigation by researchers in semantics and pragmatics (Cummins, 2015), an even more basic question arises: how can we model when people decide to use inexact quantified description?

Extension to Inexact Language

To model when people use inexact quantified descriptions, we propose a simple extension to the INP-Guess model, which we will denote as INP-Hedge. Instead of sampling a single guess value, INP-Hedge obtains ν distinct guesses, which we will denote as the multiset $G = \{g_1, ..., g_\nu\}$. This corresponds to a collection of values an individual may find plausible. If all the guesses in G are the same ($g_1 = g_2 = ... = g_\nu$), then we would predict an exact numerical description is generated equivalent to the value of these guesses. Otherwise, we would predict an inexact numerical description is generated, which can be derived from the set of guesses. For instance, consider a set of guesses $G = \{6, 8, 8\}$. Potential ways to linguistically describe this set are "about eight" or "six to eight."

How particular forms of inexact quantified description are generated is a question beyond the scope of this paper. Here, we do not attempt to model the distribution of specific forms of inexact description. In the INP-Hedge model we currently generate two forms of inexact expression: hedged numbers (e.g., "about eight") and intervals (e.g., "six to eight"). If the majority of the sampled guesses are equal to a value X, then INP-Hedge produces a hedged number expression anchored in this majority guess (i.e., "about X"). Otherwise, the model produces an interval response ("between X and Y"), where X corresponds to the minimum guess and Y corresponds to the maximum guess.

This quantified description mechanism is still preliminary, and we will discuss how the INP-Hedge model can direct future work on inexact quantifier realization in the general discussion below. Overall, INP-Hedge assumes that speakers would detect the uncertainty of their mental representation of quantity by considering multiple plausible exact number responses, and then elect to hedge their quantified descriptions. Thus, the INP-Hedge model predicts that the limits of human perceptual performance would influence language usage, because there may be insufficient time to completely eliminate uncertainty about quantity through serial attention. Specifically, the INP-Hedge model predicts that enumeration duration and numerosity have the following effects on quantified language:

(P1) In the subitizing range (quantity 1-4), participants will predominately use exact quantifiers.

(P2) For indefinitely long presentation durations, participants will use exact quantifiers.

(P3) As presentation duration decreases, the frequency of inexact quantifiers will increase.

If people elect to use inexact quantified language to avoid being incorrect, as we hypothesized above, we can propose one additional prediction:

(P4) In difficult duration/quantity pairings, participants that responded by describing quantity using inexact quantifiers will indicate higher confidence in the correctness of their response vs. participants that responded in the same duration/quantity condition with exact numbers.

Experiment 1

To test our model's predictions, we conducted an online numerical perception and quantified language elicitation experiment. Participants viewed videos in which varying quantities of black dots were presented for varying durations.

Method

Participants. Thirty-nine participants (mean age = 36.3; 19 females, 19 males, and 1 other) volunteered through the Amazon Mechanical Turk online platform (Paolacci, Chandler, & Ipeirotis, 2010). All but one participant reported being native English speakers.

Design, procedure, and materials. We manipulated the duration of stimulus presentation of dot clusters and the quantity of elements (dots) presented. Three possible presentation durations were used: 200 ms, 1000 ms, and an indefinite amount of time (dots remained on the screen while participants responded). Three possible quantity ranges were used: $[1-4]$, $[5-8]$, and $[9-12]$. The dot clusters in each video were randomly arranged, and four videos were produced for each specific numerosity, yielding 16 unique videos for each stimulus duration and numerosity condition (4 videos per number with 4 possible numbers per quantity range). Participants were presented with one video from each of these duration/quantity conditions in a random order, viewing nine videos in total. Videos were 512x512 pixels in dimension with a light grey

Num. Range	Stimulus Duration		
	0.2s	1.0s	∞ s
$[1-4]$	94.9%	97.4%	100.%
$[5-8]$	61.5%	87.2%	100.%
$[9-12]$	46.2%	53.8%	100.%

Table 1: Percentage of responses categorized as EXACT-NUM by stimulus duration and numerosity range conditions in Experiment 1.

background. A dark grey fixation cross appeared for one second before the cluster of dots. A masker grid was displayed following the stimulus interval (except in the indefinite enumeration time condition). After a video had concluded, participants were asked to complete the following sentence, being as accurate and precise as possible:

"In the above video, there are _____ black dot(s)."

Additionally, participants were asked to report their confidence in their completed description (1 = very unsure to 5 = very confident). Because we were primarily interested in investigating the precision of quantified description of a visual scene, we used a free-response, sentence-completion task instead of a completely free-response task. This was done to encourage quantified description and avoid descriptions of dot clusters based on other attributes, such as spatial arrangement (e.g., "I see a group of dots shaped like a constellation of stars").

Results

Analysis. The expressions used to complete the description were categorized into five types: (1) exact numbers, which we will denote as EXACT-NUM (e.g., "In the video above, there are *four* black dot(s)"); (2) hedged numbers, denoted as HEDGED-NUM (e.g., "In the video above, there are *about ten* black dot(s)"); (3) intervals, denoted as INTERVAL (e.g., "In the video above, there are *five to seven* black dot(s)"); (4) vague quantifiers, denoted as VAGUE-Q (e.g., "In the video above, there are *several* black dots(s)"); and (5) other miscellaneous expression, denoted as OTHER (e.g., "In the video above, there are *groups of* black dot(s)"). Two annotators classified each response. High inter-annotator agreement was found (Cohen's κ = .945). The proportions of exact numerical responses (EXACT-NUM) for each duration and numerosity condition are reported in Table 1. All four predictions were supported by the data.

Consistent with P1, 114 of 117 (97.4%) of responses in conditions within the subitizing range were EXACT-NUM responses. Additionally, consistent with P2, only exact descriptions were used when participants had an unlimited amount of time to enumerate. Consistent with P3, the number of exact responses decreased from 117 out of 117 in the indefinite duration condition to 93 out of 117 with one second of duration, yielding a significant difference (Fischer exact test, $p < .001$). The number of exact responses further decreased from 93 out of 117 to 79 out of 117 in the 200ms duration condition, though this difference was only marginally significant (Fischer exact test, $p = .054$).

Finally, Wilcoxon-signed rank tests indicated confidence ratings were significantly lower for exact responses than inexact responses for quantity ranges $[5-8]$ ($p = .039$) and $[9-12]$ ($p < .001$) at 200 ms presentation time and quantity range $[9-12]$ ($p = .007$) at 1000 ms presentation time, supporting our final prediction (P4).

Intriguingly, only 23 out of the 39 participants (59%) used inexact quantifiers. The other 16 out of 39 (41%) only used exact number responses, guessing in cases of uncertainty. As previously discussed, we would expect a participant who uses inexact quantified descriptions to generate these inexact descriptions to express uncertainty and avoid incorrectness when exact enumeration is difficult. Therefore, we would predict the EXACT-NUM responses from participants who used only EXACT-NUM responses to be less accurate than those generated by participants who switch between exact and inexact expressions. The data supported this prediction. The accuracy of EXACT-NUM responses from participants who used only exact expressions (69.4%) was found to be lower than those from participants who sometimes used inexact expressions (91.8%), yielding a significant difference (Mann-Whitney U test, $p < .001$).

Model Fit

While our four main predictions were supported by the data, we also sought to test how well the INP-Hedge model predicts the precise frequency of EXACT-NUM responses in each stimulus duration and numerosity condition. We ran the INP-Hedge model[2] ten times on each video from Experiment 1. The proportion of exact quantifier responses produced by INP-Hedge, compared with the human data from Experiment 1, is found in Figure 1. The results of the human data were highly correlated with the proportion of exact/inexact quantifiers selected by the INP-Hedge model (Spearman's ρ = 0.938). Though the correlational fit is high, we can see that INP-Hedge underestimates the amount of EXACT-NUM responses in the hardest duration and numerosity conditions. Specifically, INP-Hedge underestimates exact responses in numerosity ranges $[5-8]$ and $[9-12]$ by 43% and 28%, respectively. Exact responses during medium durations (1000 ms) for larger numerosities ($[9-12]$) are also underestimated by 33%.

What could explain this underestimation? Recall that a sizable portion (41%) of participants only gave EXACT-NUM descriptions. That is to say, about 4 out of every 10 participants guessed an exact number response when they were uncertain about the quantity of dots. To account for this, we revisited the Experiment 1 videos, running the INP-Guess model four times on each video and the INP-Hedge model six times (replicating the mixed set of strategies found in human participants). Not only was the correlational fit of this

[2]Setting $\nu = 3$.

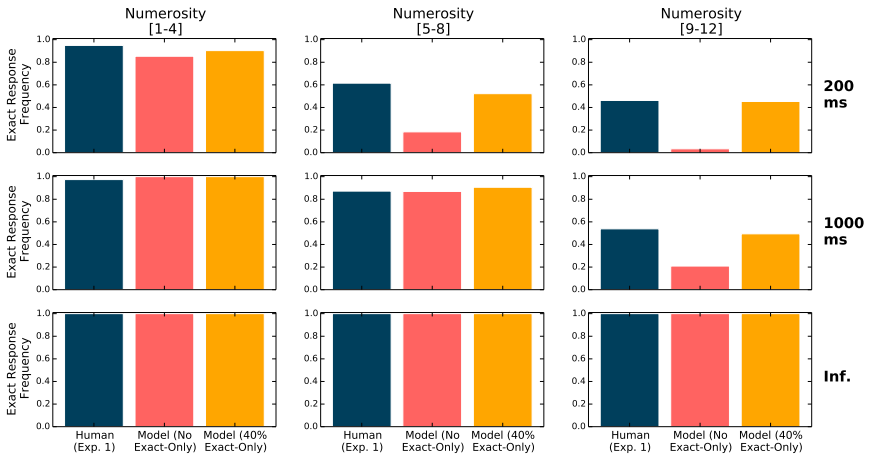


Figure 1: Predicted proportions of EXACT-NUM responses based on our model (pink) and our model adjusted for the number of participants that give only exact-number responses (yellow), compared with proportion of EXACT-NUM responses from human data in Experiment 1 (blue).

mixed-model improved (Spearman's ρ=.956), but the precise predictions about exact response frequency are much closer to the observed frequencies from the human data. Proportion differences are reduced during short stimulus durations (200ms) to 9% and 1%, for numerosity ranges $[5-8]$ and $[9-12]$, respectively. Finally, the difference in exact response proportion is reduced to approximately 4% for larger quantities ($[9-12]$) during medium durations (1000ms).

Experiment 2

Roughly 40% of participants in Experiment 1 only gave exact numerical responses. This is in line with other studies, where a subset of participants use only exact numerical expressions for all items. For instance, in a QRE elicitation task, about 20% of participants used only exact numerical expressions (Barr, Deemter, & Fernández, 2013). However, unlike in Barr and colleagues (2013), participants in Experiment 1 did not have unlimited amounts of time to view stimuli. Therefore, it seems likely that participants are limiting their set of potential quantified response forms *a priori*. One possible explanation is that because each trial in Experiment 1 involved a question asking the participant to rate the confidence of their numerical expression, participants may have felt less pressure to hedge uncertainty about the observed quantity in the language of the numerical expression itself. Rather, participants may have been more inclined to guess an exact number, then hedge their uncertainty in the confidence question. In Experiment 2, we sought to eliminate this possibility.

Method

Participants. Forty participants (mean age = 35.4; 19 females and 21 males) volunteered through the Amazon Mechanical Turk online platform. All participants reported being native English speakers.

Design, procedure, and materials. The experimental design, procedures, and materials were identical to Experiment 1, except in two respects. First, the confidence question was eliminated. Second, the number of trials each participant completed was increased to 18 (two trials per numerosity range and stimulus duration condition). Videos were randomly sampled from each stimulus duration and numerosity category without replacement.

Results

As with Experiment 1, two annotators used the same labels to categorize all the expressions participants produced. Inter-annotator agreement was again high (Cohen's $\kappa = .938$). Table 2 lists not only the proportion of exact quantified descriptions, but the precise counts of each type of expression found. We found that 14 out of 40 participants (35%) in Experiment 2 used only EXACT-NUM expressions, compared with the 16 out of 39 (41%) in Experiment 1. While the proportion of participants using only exact expressions slightly decreased in Experiment 2, this difference is not statistically significant (Fischer's exact test, $p = .647$) given the number of participants in each study. Consistent with Experiment 1, we also found that EXACT-NUM responses from participants who used only exact descriptions were less accurate (61.9%) than those from participants who sometimes used inexact expressions (88.9%), yielding a significant difference (Mann-Whitney U test, $p < .001$).

Predictions. Aside from P4, which could not be tested as there was no confidence data in this experiment, the main predictions were also still supported. Consistent with P1, 233 out of 240 (97.1%) responses in the subitizing range were exact. Also, 235 out of 240 (97.9%) responses in the

Response Type	$[1-4]$			$[5-8]$			$[9-12]$		
	0.2s	1.0s	∞	0.2s	1.0s	∞	0.2s	1.0s	∞
EXACT-NUM	75	79	79	45	67	80	35	36	76
HEDGED-NUM	2	0	0	16	8	0	13	15	1
INTERVAL	1	0	0	9	4	0	14	13	0
VAGUE-Q	2	1	1	10	0	0	17	14	3
OTHER	0	0	0	0	1	0	1	2	0
Exp. 2: Exact %	93.8	98.8	98.8	56.3	83.8	100.	43.8	45.0	95.0
Model (INP-Hedge): Exact %	85.3	100.	100.	18.4	86.9	100.	3.4	20.9	100.
Model (40% INP-Guess): Exact %	90.3	100.	100.	52.2	90.6	100.	45.3	49.4	100.

Table 2: Counts of response types by stimulus duration and numerosity conditions for Experiment 2.

indefinite duration condition were exact, consistent with P2. Consistent with P3, the number of exact responses decreased from 235 out of 240 in the indefinite duration condition to 182 out of 240 with one second of duration, yielding a significant difference (Fischer exact test, $p < .001$). The number of exact responses further decreased from 182 out of 240 to 155 out of 240 in the 200ms duration condition, yielding a significant difference (Fischer exact test, $p = .009$).

Model Fit. The pattern of exact/inexact response was similar to Experiment 1. Correlation of the proportion of exact is high (Spearman's ρ = 0.926). Underestimation of usage of exact numerical expressions remains in the post-subitizing range for short stimulus durations (200ms) and larger numerosity ranges. Specifically, INP-Hedge underestimates exact responses in numerosity ranges $[5-8]$ and $[9-12]$ by 38% and 40%, respectively. Exact responses during medium durations (1000 ms) for larger numerosities ($[9-12]$) are also underestimated by 24%. Our mixed model, (40% INP-Guess, 60% INP-Hedge) increases model correlation (Spearman's ρ = 0.944), while reducing this observed underestimation. Proportion differences are reduced during short stimulus durations (200ms) to 4% and 3%, for numerosity ranges $[5-8]$ and $[9-12]$, respectively. Finally, the difference in exact response proportion is reduced to approximately 4% for larger quantities ($[9-12]$) during medium durations (1000ms).

General Discussion

The results of our two quantified language elicitation experiments demonstrate that the use of precise quantified language to describe visual scenes decreases with decreased viewing time or increased stimulus quantity. Future computational models of quantified description of visual scenes, regardless of how they are implemented, need to account for this phenomenon to fully capture human quantified language use. We contend that to make sense of these results, one must view numerical perception as a temporally extended process in which uncertainty is reduced by additional perception of the visual scene. Both computational models we presented above, INP-Guess and INP-Hedge, account for this reduction of uncertainty by a proposed model of serial deployment of attention to individual items in the visual scene.

While we have demonstrated that a combination of these psychologically grounded and attention-driven models of numerical perception and quantifier use is able to closely fit human patterns of quantified language use under time constraints, many open questions still remain. One question raised by our quantifier elicitation experiments (and results from Barr and colleagues, 2013) is how do people decide to constrain the set of quantifiers they elect to even consider generating? Because our elicitation experiment contained relatively small quantities of visual items (1-12), participants may have felt that the degree of potential error in exact number guessing to be acceptable. With this explanation, increasing the number of potential visual items (e.g., 50-120) may reduce the proportion of participants giving only exact responses. Likewise, task motivation and context would potentially affect quantifier use. Situations were precision is critical and error may lead to highly negative consequences are likely to elicit more exact quantified language.

This work raises another series of questions regarding the realization of inexact quantified language. If people do consider multiple forms of inexact quantified language, how do people choose the precise language to use in a particular context? The mechanism for selecting different forms of quantified description in our proposed model is still rudimentary. However, it does begin to make some predictions. For instance, the current model would predict that the bounds of interval expressions would increase proportionally with reduced enumeration time or increased numerosity. In future work, we hope to address a variety of these open questions and hypotheses.

Acknowledgments

We would like to thank Hillary Harner, Andrew Lovett, Will Bridewell, Derek Brock, Brian McClimens and Sangeet Khemlani for their feedback and discussions pertaining to this project. We would also like to thank Kevin Zish, Kalyan Gupta, and the Knexus Research Corporation for their assistance in supporting these studies. Additionally, we would like to thank the reviewers for helpful feedback

in improving the presentation of this work. This work was supported by a NRC Research Associateship award to GB, and AFOSR MIPR grant F4FGA07074G001 and ONR grant N0001416WX00762, both awarded to PB. The views expressed in this paper are solely those of the authors and should not be taken to reflect any official policy or position of the United States Government or the Department of Defense.

References

Barr, D., Deemter, K., & Fernández, R. (2013). Generation of quantified referring expressions: evidence from experimental data. In *Proceedings of the 14th European Workshop on Natural Language Generation* (pp. 157–161).

Barth, H., Kanwisher, N., & Spelke, E. (2003). The construction of large number representations in adults. *Cognition*, *86*(3), 201–221.

Bridewell, W., & Bello, P. F. (2016). A Theory of Attention for Cognitive Systems. In *Proceedings of the Fourth Annual Conference on Advances in Cognitive Systems* (pp. 1–16). Evanston, USA.

Briggs, G., Bridewell, W., & Bello, P. F. (2017). A computational model of the role of attention in subitizing and enumeration. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society* (pp. 1672–1677). London, UK.

Chen, S. Y., Zhou, Z., Fang, M., & McClelland, J. L. (2018). Can Generic Neural Networks Estimate Numerosity Like Humans? In *Proceedings of the 40th Annual Meeting of the Cognitive Science Society* (pp. 202–207). Madison, WI.

Cummins, C. (2015). *Constraints on Numerical Expressions* (Vol. 5). Oxford University Press.

Dehaene, S. (2011). *The number sense: How the mind creates mathematics*. New York, NY: Oxford University Press.

Fang, M., Zhou, Z., Chen, S. Y., & McClelland, J. L. (2018). Can a Recurrent Neural Network Learn to Count Things? In *Proceedings of the 40th Annual Meeting of the Cognitive Science Society* (pp. 360–365). Madison, WI.

Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in Cognitive Sciences*, *8*, 307–314.

Gelman, R., & Gallistel, C. R. (1986). *The child's understanding of number*. Harvard University Press.

Grice, H. P. (1975). Logic and conversation. *1975*, 41–58.

Hesse, C., & Benz, A. (2018). Giving the wrong impression: Strategic use of comparatively modified numerals in a question answering system. In *Proceedings of The Conference on Natural Language Processing (KONVENS)* (pp. 148–157). Vienna, Austria.

Huss, D., & Byrne, M. (2003). An ACT-R/PM model of the articulatory loop. In *Proceedings of the Fifth International Conference on Cognitive Modeling* (pp. 135–140).

Kaufman, E. L., Lord, M. W., Reese, T. W., & Volkmann, J. (1949). The discrimination of visual number. *The American Journal of Psychology*, *62*, 498–525.

Mandler, G., & Shebo, B. J. (1982). Subitizing: An analysis of its component processes. *Journal of Experimental Psychology: General*, *111*, 1–22.

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running Experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, *5*, 411–419.

Pezzelle, S., Marelli, M., & Bernardi, R. (2017). Be precise or fuzzy: Learning the meaning of cardinals and quantifiers from vision. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers* (Vol. 2, pp. 337–342).

Railo, H., Koivisto, M., Revonsuo, A., & Hannula, M. M. (2008). The role of attention in subitizing. *Cognition*, *107*, 82–104.

Trick, L. M., & Pylyshyn, Z. W. (1994). Why are small and large numbers enumerated differently? a limited-capacity preattentive stage in vision. *Psychological Review*, *101*, 80–102.

Whalen, J., Gallistel, C. R., & Gelman, R. (1999). Nonverbal counting in humans: The psychophysics of number representation. *Psychological Science*, *10*, 130–137.

Mapping visual features onto numbers

Erik Brockbank (ebrockbank@ucsd.edu)
Department of Psychology, 9500 Gilman Dr.
La Jolla, CA 92093-109 USA

Edward Vul (evul@ucsd.edu)
Department of Psychology, 9500 Gilman Dr. #109
La Jolla, CA 92093-109 USA

Abstract

Modern society frequently requires that we express our subjective senses in objective, shared formal systems; this entails mapping multiple internal variables onto a common scale. Here we ask whether we accomplish this feat in the case of estimating number by learning a single mapping between explicit numbers and one integrated subjective estimate of numerosity, or if we separately map different perceptual features onto numbers. We present people with arrays of dots and ask them to report how many dots there are; we rely on the systematic under/overestimation of number at higher quantities to estimate error in the mapping function. By comparing how this error changes over time, as the mapping fluctuates for different visual cues to numerosity, we can evaluate whether these cues share a single mapping, or are mapped onto number individually. We find that area, size, and density all share a common mapping, indicating that people obtain a unified subjective estimate of numerosity before mapping it onto the formal number line.

Keywords: numerosity; number; estimation

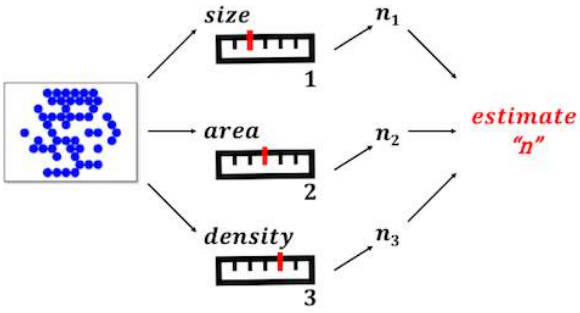
Introduction

Human reasoning and planning frequently involves mapping internal estimates onto formal systems: we can compare the weights of two rocks using our subjective sense of weight, but to provide an estimate of one rock's weight in kilograms requires mapping that subjective sense of weight onto a formal metric system. This task of expressing our internal subjective senses in objective, standard systems is commonplace, from making time estimates to evaluating prices. To accomplish this we somehow learn to map from perceptual and internal states onto formal systems like weight in kilograms. The task is often complicated by the fact that we might have many subjective variables that must map on to the same formal system: weight might be estimated by the pressure a rock exerts on our hand, or by its inertia as we try to move it. How do we deal with multiple subjective cues: do we map each one individually onto the formal system, or do we combine them to come up with a single subjective estimate, and then map that estimate onto the formal system? In this paper, we approach this question for people's ability to estimate numerosity.

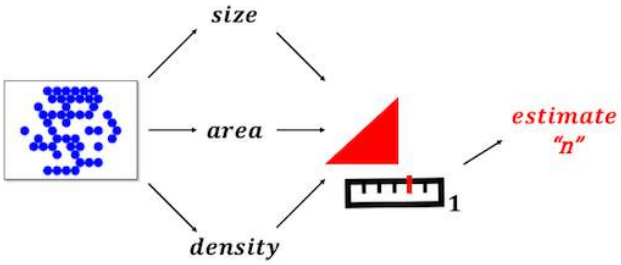
Based on a quick glance at a display of many objects, people can estimate the number of objects present in the display. Even if there is insufficient time to explicitly count the objects, there are enough visual features that correlate with number, that a number estimate may be obtained just based on these internal, analog signals which together give us a sense of "Approximate Magnitude". Imagine for example stepping into a room full of people: as you look around, you can get a rough sense of how many are present just based on the density of the crowd and the size of the room faster than you would be able to count each person individually. In general, displays with higher numerosity tend to have objects distributed over a larger portion of the visual field (area) and the number of objects in a constant area tends to be higher, either because the objects themselves tend to be smaller (size), or because the inter-object distances are smaller (spacing/density). These separate cues to numerosity may be treated in different ways by the visual system. They may be combined into one internal representation of numerosity which forms the basis of estimation. Or, because visual cues to numerosity all tend to correlate together in the real world, it may not be necessary to undertake the extra calculation of integrating them to form an internal sense of number. Instead, these features may be mapped onto numbers directly. Both explanations posit internal representations which must be mapped onto formal numbers when making estimates, but differ as to how this mapping occurs.

A large body of research has examined the representations that support our internal sense of number but comparatively little work examines how we might map from that internal sense to number estimates. The degree to which we directly perceive and represent number is an area of active debate (for a recent review see Leibovich, Katzin, Harel, and Henik (2017)). Researchers have proposed that an internal sense of number, the "Approximate Number System" or ANS, exists in many animals and is developed by infants at a young age (Feigenson, Dehaene, & Spelke, 2004). However, competing accounts emphasize that perceptual features of a quantity such as size, area, and density are highly correlated with number: this has led some to argue that our ability to estimate numerical quantities can be served directly by these continuous magnitudes without any internal number sense (Gebuis & Reynvoet, 2012) or that insofar as we have an internal representation of number, it is assembled directly from our sense of continuous magnitudes (Leibovich et al., 2017).

The present experiment is agnostic about the precise mechanisms for visual processing and internal representation of numerosity. We are interested in understanding how people map from various internal representations to the formal number line during numerical estimation. One hypothesis is that



(a) A relationship between visual properties and number estimates that relies on multiple independent mappings (numbered 1, 2, and 3) from distinct visual features to an estimate "n"



(b) A relationship between visual properties and number that specifies a *single mapping* (numbered 1) between some internal quantity representation (the red incline) and an estimate "n"

Figure 1: Two different ways of thinking about how visual cues to magnitude map onto numbers

people have multiple mappings which take as their inputs features associated with numerosity such as size, density, and area. Another hypothesis is that people instead have a single mapping from some internal representation—whether a number sense or a broader integrated magnitude—to an estimated quantity. Both mapping hypotheses are plausible a priori and might inform the broader debate about how people perceive or represent number. In what follows, we discuss in greater detail the research on number representation, which supports the availability of various possible *inputs* to this mapping function.

ANS and Continuous Magnitudes

The predominant theory in number processing holds that people have an internal approximate number system which they map onto the formal number system for purposes of estimation and other related tasks (Izard & Dehaene, 2008). Work in this space has sought to model the characteristics of this number system, including how it is represented internally (Izard & Dehaene, 2008) and how it develops in infants and young children (Carey, 2009). Research on development of the approximate number system has found that ability to distinguish between distinct numbers—the *acuity* of the approximate number system—develops independently of ability to discriminate area, density, length, and time (Odic, 2018) and that acuity of number sense in children is correlated with mathematical ability later in life (Halberda, Mazzocco, & Feigenson, 2008). In line with the idea that numerosity is a core part of how we represent the world around us, some have argued that numerosity is even available as a primary feature of perception and not reducible to related properties like texture density (Burr & Ross, 2008). In support of this, it has been shown that numerosity estimates are subject to visual adaptation effects, much like other visual properties such as color and motion (Burr & Ross, 2008).

In contrast to proposals that humans have an internal approximate number system, some have argued that number estimation is inferred directly from visual properties that correlate with number (Dakin, Tibber, Greenwood, & Morgan, 2011). Evidence that people's ability to estimate quantities stems directly from their processing of visual cues comes primarily from work showing that people struggle to infer numerosity independently of the information they receive from visual cues (Leibovich et al., 2017). For example, Gebuis and Reynvoet (2012) presented participants with a series of dot arrays which manipulated the convex hull, aggregate surface, and density of the dots such that none of these visual properties correlated with true quantity across all the trials. They found that participants' estimates of the number of dots in the arrays were largely explained by each of these features even though these features provided no information about the true number of dots. They argue that people are therefore unable to estimate numerosity independently of the visual cues which tend to provide certain signals about numerosity. More recent work has argued that the basis for our sense of number is our ability to process *continuous magnitudes* (density, area, size, etc.): Leibovich et al. (2017) challenge the degree to which research on the approximate number system is able to isolate numerosity from visual cues and argue for a more general magnitude system from which number is inferred.

Regardless of whether people have an internal sense devoted specifically to numerosity or assemble their sense of quantity from continuous magnitudes that correlate with number, it's clear from the existing research that a.) numerosity and visual features such as area, size, and density are closely tied and b.) that a mapping from internal quantity estimates onto the formal number system could plausibly take as its inputs any combination of visual features and numerical representation. In light of this, we propose two hypotheses about how such a mapping might work. One holds that we have multiple mappings from size, area, and density features to number estimates. These mappings could independently serve our estimation needs. The other holds that we have a single mapping from some internal quantity representation onto the number line. This internal representation could be our approximate number sense or a quantity estimate assembled by combining information from size, area, and density. In what follows, we summarize research which has examined people's performance on number estimation tasks and describe a novel method of investigating mappings from internal number to formal number.

Individual differences and drift in mental number-line calibration

Work investigating the approximate number system has sought to understand how we map from our internal sense of number to the verbal number system. Several key findings have informed this line of inquiry. First, people's mapping from internal quantity representations to formal numbers is often miscalibrated (Izard & Dehaene, 2008). Specifically, individuals asked to estimate quantities outside the *subitizing* range tend to systematically over- or underestimate those quantities. This relationship follows a power law: the higher the true quantity, the more people reliably over- or underestimate (Izard & Dehaene, 2008). Second, the amount that people are miscalibrated in their estimations varies considerably across individuals (Vul, Barner, & Sullivan, 2013). Some people reliably overestimate while others reliably underestimate, suggesting that whatever mapping we use onto formal number varies from person to person. Finally, the amount that people are miscalibrated in their estimations varies *within* individuals. In other words, the degree to which people over- or underestimate has been shown to *drift* over many successive estimations (Vul et al., 2013).

In this study, we use the slow drift of individuals' number estimates to investigate the mapping between internal number and quantity estimates across various visual conditions. If people rely on multiple mappings from independent visual features to numerosity estimates, then we would expect these mappings to drift independently over many estimates based on different visual features. However, if people utilize a singular mapping function from some internal quantity estimate onto number, then we would expect the drift in their estimates to be invariant to changes in the visual cues used to form each estimate. This difference is illustrated in Figure 1.

Experiment

We tested the degree to which the numerosity of a display is estimated through independent mappings from correlated visual features (e.g., size, area, and density) to number, or if these features map onto a single internal numerosity estimate, which is then mapped onto a symbolic number. We presented participants with an estimation task in which the stimuli varied along one dimension (size, area, or density) as magnitude changed, while holding the other two dimensions constant. We compared the drift in participants' magnitude estimations across the three conditions to assess whether number estimates may be independently obtained from size, area, and density cues.

Participants

Participants were 57 undergraduates at the University of California, San Diego who received course credit for their participation.

Methods

Participants were shown a series of dot arrays on a white background like those in Figure 2. The dots appeared on the screen for 500ms and then disappeared. Participants were then prompted to guess the number of dots on the screen. We did not use a mask between trials, as any aid that participants received in estimating due to an after image would have been consistent across all trials. For the first 25 trials, participants were given feedback after each round about the true number of dots they had just seen. Participants were awarded points after each round on a logarithmic scale based on the difference between their guess and the true number of dots. Participants performed 1,000 estimation trials or 50 minutes on the task, whichever came first.

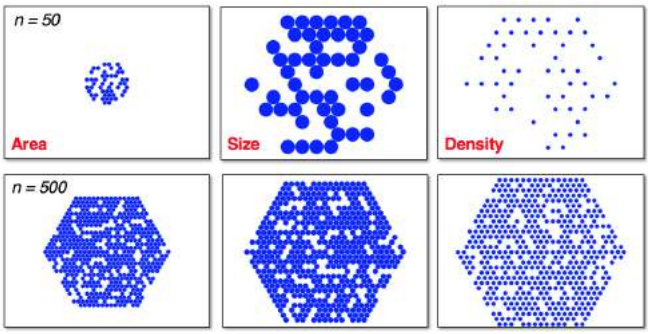


Figure 2: Sample stimuli for n = 50 and n = 500 across *area*, *size*, and *density* conditions.

Stimuli

The number of dots on each trial was selected by sampling a number between 10 and 750 from an exponential distribution with a mean of 100. Each trial of the experiment was randomly selected to vary one of the *area*, *density*, or *size* of the dots while keeping the other two constant.

Trials that varied the *area* of the dots used a predetermined constant for spacing between dots and dot size so that the number of dots on the screen was indicated by how much area the dots occupied. For trials that varied the *density* of dots, dots were populated in a constant area on the screen with a constant size: when there were more dots, the spacing between them was lower and when there were fewer dots, there was greater spacing between them. Finally, trials that varied the *size* of the dots used a consistent area of the screen and a consistent spacing between dots, generating larger dots when there were fewer in a given trial and smaller dots when the magnitude was greater. See Figure 2 for examples of dot arrays that varied each visual feature. In each trial, a random selection determined whether the dot display would vary size, area, or density so that over the course of the experiment, all three features would present cues to numerosity but on any given trial, only one would be informative.

Results

To understand how well participants estimate visual quantities based on size, area, and density inputs, we can compare their estimates for trials in which each of these features were informative to the actual numbers presented. For

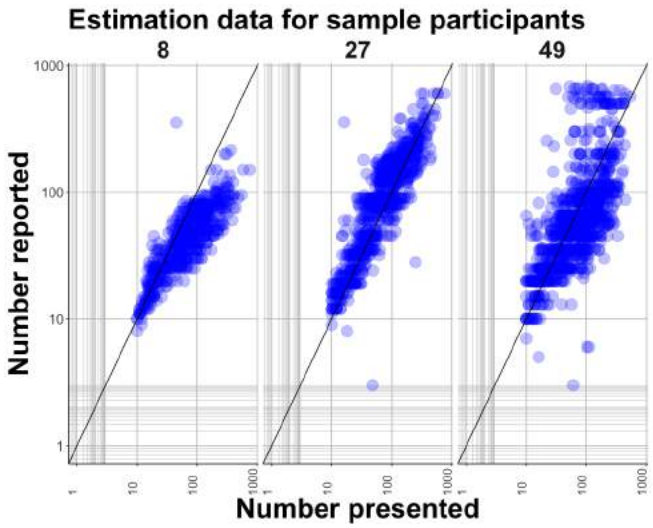


Figure 3: Individual data for *Number presented* and *number reported* from three sample participants. The degree to which each particpant underestimates, overestimates, or is accurate reflects individual differences in this task

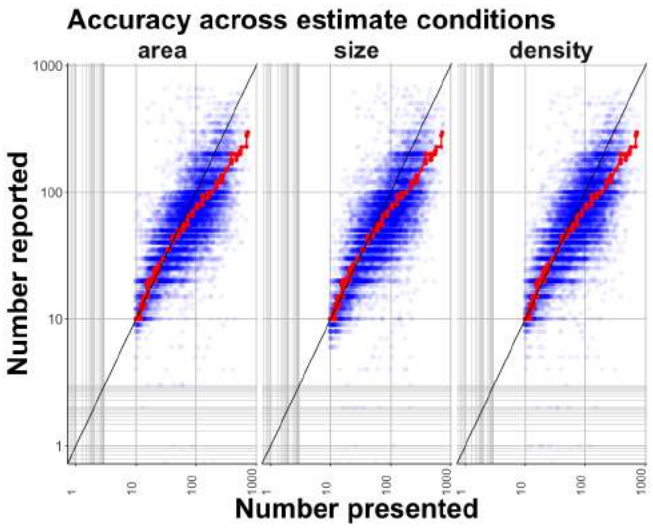


Figure 4: *Number presented* and *number reported* across estimate conditions. The red line is the median response for each number presented. Participants show similar underestimations across estimate conditions.

perfect estimators, each estimate plotted this way would lie along the identity line. Estimates show a high degreee of variance across individuals: Figure 3 shows data from three sample participants which illustrate this. Combining this data across all participants, Figure 4 shows participants' accuracy by plotting their estimates alongside the true number presented for trials varying size, area, and density. Consistent with earlier findings, people are accurate up to numerosities of about 20-30, but they reliably underestimate larger numbers on average (even setting aside the individual variance: see Vul et al., 2013). The underestimation pattern in a given set of trials can be described as a bilinear function which follows the identity line up to a threshold, and deviates from the identity line with some slope thereafter. This slope amounts to the "calibration" of the mental number line (Izard & Dehaene, 2008), and was precisely shown to (a) vary across subjects, and (b) within subjects, slowly drift over the course of an experiment (Vul et al., 2013). Figure 4 shows that the calibration of the mapping to the formal number line is similar regardless of whether area, size, or density is the numerosity-informative variable.

Previous research has shown that perception of structure and groupings can lead to systematic underestimation. For example, objects connected by lines are underestimated relative to disconnected objects (Franconeri, Bemis, & Alvarez, 2009). When dot arrays are seen as grouped, the degree to which they're clustered increases underestimation (Im, Zhong, & Halberda, 2016). Even sub-conscious processes like statistical learning of co-occurrence in colored dot arrays can lead to underestimation (Zhao & Yu, 2016). In the stimuli presented here, it is possible that perception of grouping among dots on various trials led to underestimation. However, given the similarity in patterns of underestimation in Figure 4 across modalities, it's unlikely that underestimation due to perceptual grouping affected the size, area, or density informative trials more than any other.

Consistent individual differences across modalities

Dividing each participant's estimates into blocks of size 50 (after the initial 25 "calibration" trials), we can evaluate the best fitting slope estimates for each block in each condition. For example, block 11 for each participant will contain trials 476–525. Of these, some number will be area trials, some will be size trials, and some will be density trials. We can extract the trials belonging to each estimate condition (size, area, density) for a given block of trials for a given participant, and compute a best fitting slope for that subset of number estimates for that participant. This gives us a slope for each estimate condition, for each participant, in each block of 50 trials over the course of the experiment.

To the degree that processes of estimating quantity based on changes in size, area, and density are independently calibrated to the data participants have seen, individual differences in slopes should not be consistent across these different modalities. However, if different variables are mapped onto a subjective quantity estimate, and the uncertain, idiosyncratic mapping lies between approximate number and reported number, then an individual's slope for area-determined numerosity will be consistent with their density- and size-determined numerosity as well.

Figure 5 shows each of the possible correlations between best fitting slopes for size, area, and density estimates across participants in each block: size-area slope correlations, size-density slope correlations, and density-area slope correlations. Block zero (trials 1-25) includes the trials in which participants received feedback after each guess. In these tri-

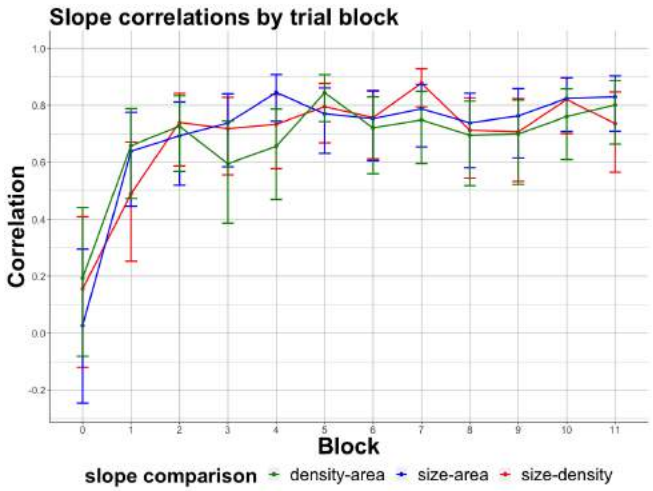


Figure 5: Slope correlations of size-area, size-density, and density-area for the first 11 blocks of the experiment. Participants received feedback in block 0. After that, correlations across estimate conditions are relatively high and are tightly coupled in remaining blocks (error bars indicate 95% confidence intervals on the correlation coefficient).

als, participants were fairly accurate in their estimates. Correlation between slope estimates in block zero is therefore low due to low variance in slope estimates across all conditions as a result of the feedback. However, from block one onwards, correlation of slope estimates between size and density trials, size and area trials, and density and area trials increases to 0.6–0.8. Critically, each of these three slope estimate correlations (size-area, size-density, and density-area) remain high and tightly in tandem from block one onwards. Such closely aligned correlations would be unusual if variations in stimulus size, density, and area across trials each directly and independently enabled an estimate of quantity.

Consistent within-individual drift across modalities

The correlation between different slope estimates for each block, described above, indicates how similar size, area, and density estimates were to each other across participants in each set of 50 trials throughout the experiment. In other words, this shows whether individual differences in numerosity estimations are consistent across modalities. Another feature of numerosity estimates is their drift in calibration over time *within individuals*. Here, we examine the data from estimation across modalities in light of this pattern: if size, area, and density each independently map to a formal number estimate, the drift in calibration for each of these modalities should be independent. However, if each modality maps to an internal estimate and the drift reflects changes in the mapping of internal estimates to formal number, then we will not detect any difference in drift across modalities. Figure 6 shows the correlation between slopes in blocks 1 – 11 for each possible comparison of estimate conditions: autocorrelation of e.g. density slopes across blocks and correlations across modalities of e.g. size to area slopes between each block. The overall pattern of correlations across blocks looks very similar for each of these comparisons, further reinforcing the idea that these features do not map separately onto number estimates. Individual drift in over- and underestimations can be seen in the lower correlation between blocks that are farther apart: this pattern is also similar across comparisons.

To better compare the slope correlations within and across modalities, we group pairs of blocks based on their temporal separation: their *trial distance*. For example, the correlation between blocks 1 and 4, and 2 and 5, and 3 and 6, all have a trial distance of 3 blocks (150 trials). The decline in correlations over longer trial distances indicates the drift of mapping over time. We can thus compare these cross-correlation functions for different modality-modality comparisons. Comparing slope estimates for each block of a given condition to those that are all an equal distance away in the same or alternate estimate conditions gives us a correlation between trials across a range of trial distances. Figure 7 shows these correlations by distance for the same combinations of estimate conditions shown in Figure 6. Across all comparisons, the correlations decrease as distance between trials increases. This drift in estimate slopes—the slopes of trial blocks farther from each other are less similar to the slopes of closer blocks—reflects the drift in calibration of the mapping function onto precise quantity over time (Vul et al., 2013).

To ensure that the "drift" shown in Figure 7 is not attributable to differential distributions of each trial type across blocks of increasing distances, we shuffled trial order for each participant and re-calculated the correlation of slopes by trial distance. For the shuffled data, the correlations within and between modalities were very stable across *all* trial distances: in other words, there was no sign of systematic drift. We fit linear models to the mean correlations at each trial distance for each modality comparison to ensure that indeed there was no drift in the shuffled data: for the six comparisons shown in Figure 7 (with shuffled trial order), none had a slope significantly different from 0[1].

Figure 7 illustrates that the drift in calibration occurs not only within each estimate condition but also across them: size-density, size-area, and density-area slope comparisons show similar decreasing correlations at greater trial distances. Most importantly, these correlations over trial distances are indistinguishable whether we consider within-modality correlations (e.g., area-area) or across-modality correlations (e.g., area-size). The correlation of slopes over varying trial distances is indistinguishable within and across modalities. If visual cues to density, size, and area each map to a subjective numerosity which *then* maps to precise quantity estimates, the similarity of slope correlations within and across esti-

[1] Shuffled trial order correlation slopes (per block): size-size 95% CI = [-0.013, 0.001] p = 0.07, density-density 95% CI = [-0.010, 0.002] p = 0.16, area-area 95% CI = [-0.001, 0.015] p = 0.08, size-density 95% CI = [-0.019, 0.003] p = 0.12, size-area 95% CI = [-0.014, 0.003] p = 0.19, density-area 95% CI = [-0.006, 0.013] p = 0.39

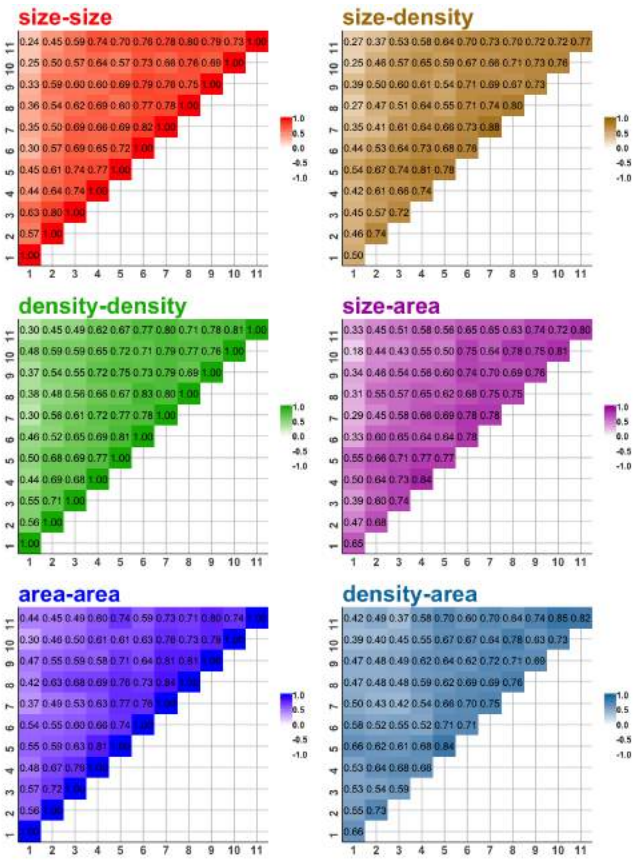


Figure 6: Slope comparisons between each block across all estimate conditions. The correlation pattern is similar across all these comparisons, a pattern seen when plotting correlation by trial distance as well.

mate conditions could be accounted for by separate mapping functions from visual properties to approximate number being similarly calibrated. The drift would then be attributed to the mapping from this internal numerosity to the number reported. If the estimation of reported quantities was accomplished by separate mapping functions from each of the visual modalities, it would be improbable for these mapping functions to change in lock-step, thus rendering the within- and across-modality correlations identical.

Conclusion

We asked how people map their analog, perceptual features onto explicit numbers. Specifically, we investigated whether (a) people have one mapping from a cohesive, internal estimate of numerosity/magnitude onto the number line, or if (b) people have multiple mappings onto number from different visual features that tend to correlate with number. In this experiment, we asked participants to estimate the number of dots present in a display, while we varied which visual feature varied with number. The numerosity of the dot arrays in a given trial could only ever be inferred from a single perceptual modality (one of size, area, or density) of the dots.

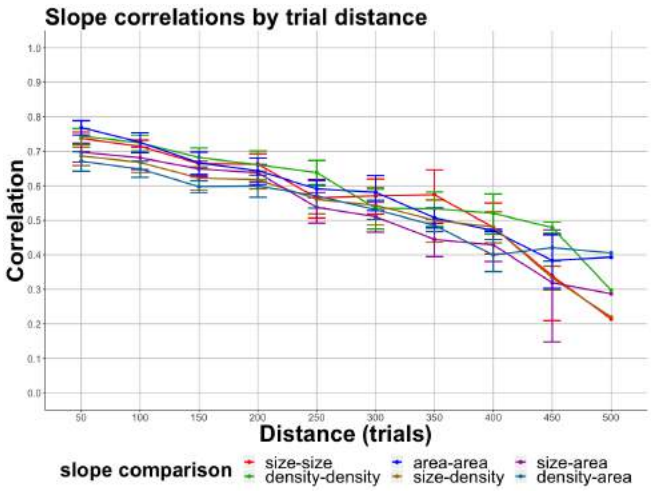


Figure 7: Slope comparisons *within* and *across* estimate conditions by trial distance. The correlation of slopes over varying trial distances is indistinguishable within and across estimate conditions (error bars represent standard error across measurements of each distance).

Using results from prior research indicating that people drift in their mapping of numerosity to number over the course of many trials (Vul et al., 2013), we analyzed whether such drift occurs independently across estimates based on size, density, and area. We find that the (mis)calibration of numerosity onto number is indistinguishable between size, area, and density trials. Moreover, as this mapping drifts over the course of many trials, it changes in lock-step for all the modalities, indicating that there is only one mapping function that drifts, which is shared across all modalities. From this we conclude that size, area, and density all share a common mapping onto formal number. In other words, perceptual features that are cues to numerosity must be combined into an internal representation of numerosity which is then mapped onto the formal number line when reporting an exact number.

It would be rash to generalize these results to all cases in which we might map subjective senses onto objective, external standards (e.g, estimating weight in kilograms, or our willingness-to-pay in dollars). It seems likely that some formal systems do not have a corresponding unified internal representation, and instead have an assortment of independent mapping functions which may be inconsistent and incommensurate. However, these results are encouraging that for at least some formal systems, we have unified, coherent internal representations that serve as their substrate. We postulate that the methods we develop here—of relying on fluctuations in the mapping of subjective states onto formal, objective systems—might be used to identify whether notions like "subjective value" are indeed unified monolithic entities, or if they are an ensemble of related, but independent, internal senses.

Acknowledgments

We thank our reviewers for their thoughtful suggestions.

References

Burr, D., & Ross, J. (2008). A visual sense of number. *Current biology*, *18*, 425–428.

Carey, S. (2009). Where our number concepts come from. *The Journal of philosophy*, *106*.

Dakin, S., Tibber, M., Greenwood, J., & Morgan, M. (2011). A common visual metric for approximate number and density. *Proceedings of the National Academy of Science*, *108*, 19552–19557.

Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in cognitive sciences*, *8*, 307–314.

Franconeri, S., Bemis, D., & Alvarez, G. (2009). Number estimation relies on a set of segmented objects. *Cognition*, *113*(1), 1–13.

Gebuis, T., & Reynvoet, B. (2012). The role of visual information in numerosity estimation. *PLoS One*, *7*.

Halberda, J., Mazzocco, M., & Feigenson, L. (2008). Individual differences in non-verbal number acuity correlate with maths achievement. *Nature*, *455*.

Im, H., Zhong, S., & Halberda, J. (2016). Grouping by proximity and the visual impression of approximate number in random dot arrays. *Vision research*, *126*, 291–307.

Izard, V., & Dehaene, S. (2008). Calibrating the mental number line. *Cognition*, *106*, 1221–1247.

Leibovich, T., Katzin, N., Harel, M., & Henik, A. (2017). From sense of number to sense of magnitude: The role of continuous magnitudes in numerical cognition. *Behavioral and Brain Sciences*, *40*.

Odic, D. (2018). Children's intuitive sense of number develops independently of their perception of area, density, length, and time. *Developmental Science*, *21*.

Vul, E., Barner, D., & Sullivan, J. (2013). Slow drift of individuals' magnitude-to-number mapping. In *Proceedings of the 35th annual meeting of the cognitive science society.*

Zhao, J., & Yu, R. (2016). Statistical regularities reduce perceived numerosity. *Cognition*, *146*, 217–222.

When do people use containment heuristics for physical predictions?

Erik Brockbank [1] (ebrockbank@ucsd.edu)
Edward Vul [1] (evul@ucsd.edu)
Kevin Smith [2] (k2smith@mit.edu)

[1] Department of Psychology, 9500 Gilman Dr. La Jolla, CA 92093-109 USA
[2] Department of Brain and Cognitive Sciences, MIT Building 46-4053, 77 Massachusetts Avenue Cambridge, MA 02139

Abstract

Accounts of human physical reasoning based on simulation from a noisy physics engine have enjoyed considerable success in recent years. However, simulating complex physical dynamics can be a computationally expensive process, and it is possible that people use faster, cheaper shortcuts to make predictions and inferences in complicated physical scenarios. Here we asked people to predict the eventual destination of a ball on a 2D bumper table (in the style of Smith, de Peres, Vul, and Tenenbaum (2017)). We designed scenarios that we expected would modulate the use of heuristics and simulation: the bumper table provided varying degrees of containment to constrain future outcomes and to make a containment heuristic more useful, and could have more or less internal structure to vary the reliability of noisy simulation. As the containment heuristic becomes more useful, and as simulation becomes more expensive, we expected that people would switch from using simulation to rely more on rapid heuristic-based predictions and therefore respond faster. Instead, we found that even when containment was very predictive, people were progressively slower and less accurate as simulation complexity increased, indicating that they persisted in using simulation rather than containment heuristics.

Keywords: simulation; heuristics; physics

Introduction

In everyday life we are constantly tasked with making predictions about how physical objects will behave and interact, whether changing lanes in traffic or stacking dishes in the sink. Such inferences are so commonplace that we rarely think twice about them. However, the mechanisms by which we are able to make these inferences are far from obvious: at a minimum, they require a rich understanding of how things in the world tend to move and the ability to make rapid predictions based on this knowledge, both non-trivial achievements from a computational perspective.

Prior research has shown that a range of human physical inferences can be captured by *Intuitive Physics Engine* models that rely on simulations of physical outcomes performed with a probabilistic physics engine similar to those used in computer games (Battaglia, Hamrick, & Tenenbaum, 2013). By sampling from these simulations, probabilistic models can generate a reasonable representation of the physical world and make predictions accordingly (Ullman, Spelke, Battaglia, & Tenenbaum, 2017). Such models have been successful in reproducing human judgments across a range of tasks and domains, from predictions about object balance (Battaglia et al., 2013), mass (Hamrick, Battaglia, Griffiths, & Tenenbaum, 2016), and velocity (Smith & Vul, 2013) to liquid dynamics (Bates, Yildirim, Tenenbaum, & Battaglia, 2018) and causal attribution (Gerstenberg, Peterson, Goodman, Lagnado, & Tenenbaum, 2017).

While simulation allows us to reproduce many features of human physical reasoning, there are also situations where people's behavior is inconsistent with the use of an intuitive physics engine (Smith, Battaglia, & Vul, 2018). Empirically, human behavior sometimes differs significantly from predictions made by simulation-based models, suggesting that we have sophisticated strategies for avoiding simulations when other forms of inference will suffice (Smith, Dechter, Tenenbaum, & Vul, 2013). In particular, research on errors in physical judgment have shown that people often hold a number of systematic biases which are inconsistent with even basic physical simulations (see Davis & Marcus, 2015 for an overview of some of these). Underlying this difference is a criticism of simulation as a computational account of all human physical reasoning: simulation of almost any sort, but particularly of complex physical phenomena, may require considering the interactions between a large number of objects over time. Because interactions between objects add uncertainty to predictions (Smith et al., 2013), in complex scenarios these simulations might therefore require keeping a large number of objects in mind and yet still produce very uncertain predictions. These sorts of considerations have led some to argue for a limited role of simulation in human physical reasoning (Davis & Marcus, 2016).

In light of the challenges posed to a simulation-based account of human physical reasoning, what alternatives can account for people's ability to make diverse predictions about physical interactions in the world around them? A large body of research supports the idea that humans are adept in their use of heuristics and other simplified qualitative prediction strategies, including in the domain of physical predictions (Gigerenzer & Todd, 1999). Prior work has shown that people can represent certain topological relationships like containment using only first-order logic (Davis, Marcus, & Frazier-Logue, 2017). Given the large number of strategies available to reasoners and the flexibility with which we navigate the physical world, it has been proposed that humans selectively utilize a toolbox of prediction techniques, including simulation, qualitative reasoning, and logical inference, as well as analogical and rule-based strategies (Davis & Marcus, 2015).

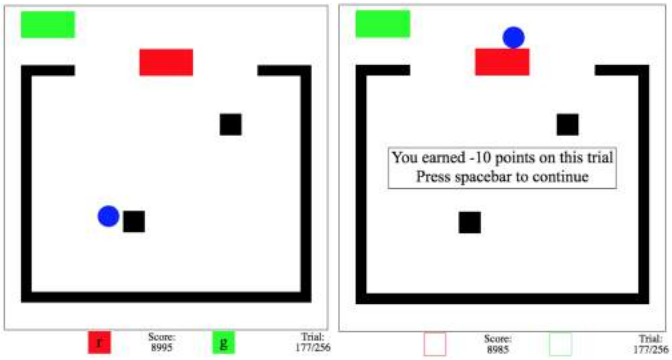


Figure 1: A simple trial with partial containment and two obstacles. At left is what participants see when prompted to guess a target after 2.5s of animated ball movement. At right is feedback after guessing "green" and seeing the ball animated on the remainder of its path.

The idea that humans are able to balance simulation-based prediction with alternative prediction strategies is intuitively appealing because it offers a way to unify simulation-based accounts with complementary accounts of physical inference based on e.g. topological and visual features. However, it raises a number of additional questions. If humans are able to flexibly recruit different strategies for making physical predictions, what determines the choice of one strategy over the other? How and when do we switch between fine-grained simulation methods and more coarse qualitative analyses? The exact mechanisms for such decisions remain poorly understood. For example, when novel but reliable and visually salient heuristics are available, people often fail to use them unless the existence of such heuristics are made explicit (Callaway, Hamrick, & Griffiths, 2017). A simple hypothesis is that compared to simulations, topological predictions are faster, lower fidelity, and less generally applicable; consequently, topology ought to be used when the scenario makes topological cues particularly useful, and renders simulations particularly imprecise and costly by complex scenarios. In other words, if computationally expensive simulations are unlikely or unable to produce a confident prediction, while topology can, a rational agent should make a guess based on simpler heuristics or visual features rather than waste resources on repeated simulations.

In the present study, we tested the hypothesis that people balance the precision and cost of simulation against the applicability of topological analysis when making physical predictions. Our experiment builds on prior research in several important ways. First, we examine people's reasoning about containment scenarios because prior research has shown that containment relationships can be expressed propositionally and that intuitive inferences about containment can be made with such knowledge-based reasoning even with very little information (Davis et al., 2017). As such, it is an ideal simplified model for physical inference. Second, containment relationships can in some cases be visually processed rapidly and automatically (Strickland & Scholl, 2015). Finally, prior research has used a similar paradigm to explore the degree to which people simulate or use topological inference when making physical predictions in scenarios involving containment relationships. Smith et al. (2013) modeled inference on a prediction task using noisy simulation but found that people's predictions were more rapid than the model predicted in scenarios involving containment. Building on these results, Smith et al. (2017) presented participants with similar tasks in which a containment heuristic was available but found evidence for simulation across all the tasks. However, in the tasks presented to participants, the simulation required was fairly straightforward and temporally limited. Therefore, insofar as simulation and topological processing happened in parallel or participants reasoned that simulation was a consistently viable strategy, they may have failed to leverage a more coarse containment-based judgment out of habit or convenience (Smith et al., 2017). We hypothesize that when topological predictions are available *and* simulation proves intractable or uncertain, participants will be more likely to make their predictions based on topology. In line with this hypothesis, Davis & Marcus (2015) argue that simulation is most effective on relatively short time scales and small spatial scales such that simulation is straightforward and reliable. Here we violate this condition by including trials in which the number of obstacles (complexity level) makes simulation both more uncertain and potentially longer. We expect that participants, faced with predictions involving unreliable simulations, will pursue alternative strategies for prediction: an agent that rationally trades off the advantages of simulated inference with the computational costs should select more favorable knowledge-based inference strategies when conditions support them.

Experiment

In the present study, we tested the hypothesis that people would switch from using slower simulation to faster heuristics when simulation becomes less efficient. Specifically, we presented participants with a task which required them to make predictions about the path of a ball in a series of two-dimensional environments. We manipulated (a) how much the topography of the environment allowed a simple topological "containment" heuristic to identify the answer (degree of containment), (b) the complexity and uncertainty of simulations in the environment (degree of complexity). The core prediction is that participants would favor using simulation to obtain an answer when simulations were easy and topology was uninformative, but would switch to relying on containment, or other coarse topological cues when they were particularly effective, and simulation was particularly ineffective. Specifically, we rely on the assumption that using a fast containment heuristic would be more efficient than simulation, thus we predict that for high-containment scenarios, increasing complexity would decrease response times.

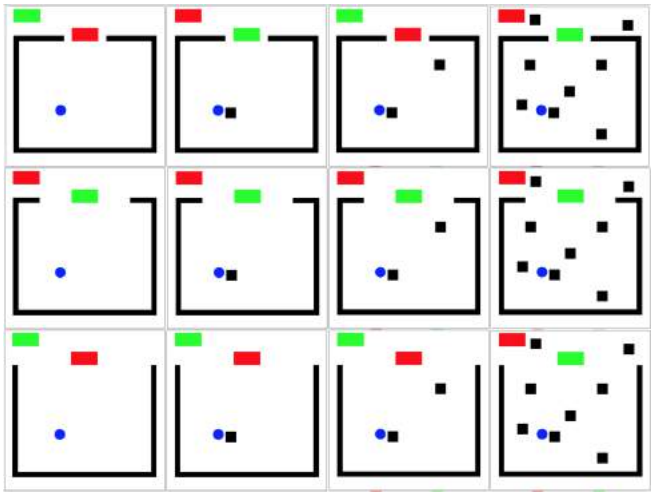


Figure 2: Twelve trials for Scenario 1, increasing in *simulation complexity* in the horizontal direction and *containment* in the vertical direction (highest containment at top). The high containment, high complexity trials offer a simple topological prediction without needing to simulate the ball's interaction with the walls and obstacles.

Participants

Participants were 81 undergraduates from the University of California, San Diego who received course credit for participation. The experiment lasted approximately 25 minutes.

Methods

We used a task that is very similar to Smith et al. (2017). During the experiment, participants were shown a series of trials depicting a blue ball on a flat surface (600 pixels by 600 pixels). The ball was surrounded by walls and square obstacles that the ball could bounce off. Each trial contained a red and a green target and the goal of the task was to determine whether the ball would hit the red target or the green target first (see Figure 1). Before making a guess, participants were shown 2.5s of the ball's movement, after which the ball paused in its trajectory and participants pressed either the "R" or "G" key to indicate their guess for the *red* or *green* target. After participants made their guess (or 10s elapsed), the ball would resume its movement until it hit one of the targets. At the end of each trial, participants received points based on their accuracy and their response time: -10 for an incorrect guess, 0 for no guess, variable points for a correct guess based on time to respond (see Figure 1). The points for a correct guess were allotted based on an exponential decay function of response time so participants were rewarded for guessing quickly if they could generate an accurate guess rapidly, but the penalty for longer response times quickly diminished. To illustrate, participants received 100 points for responses at 250ms, 71 points at 1000ms, and 45 points at 2000ms.

Participants read a brief set of instructions and completed three practice trials before doing the experimental trials. Each participant completed all trials in the experiment: 48 trials representing each complexity and containment level across four scenarios, with 64 "distractor" trials (discussed below) for a total of 256 trials. The order of the trials was randomized for each participant, as was the selection of the red and green target for reach trial.

Stimuli

The trials were grouped into four qualitative scenarios, and within each scenario they were parametrically manipulated along two dimensions that modified the uncertainty of simulations and the availability of topological predictions.

Scenario: Each trial belonged to one of four possible scenarios corresponding to the containment structure that the targets were placed in. For example, one scenario placed the ball inside variants of a box where one of the targets was placed in the opening, while another had the ball traveling down a right-angled tunnel with a target at one end. (see Figure 5).

Containment: Each scenario had three distinct *containment* levels that varied how much the ball and one of the targets were contained by the set of walls in the scenario. In the high containment trials for each scenario, the ball was virtually guaranteed to hit one of the targets because the ball and that particular target were almost entirely contained by the walls. In the low containment trials, the walls provided only minimal containment for the ball and one of the targets, rendering topology and containment fairly uninformative.

Simulation Complexity: For each scenario and containment level, there were four *complexity* levels which varied the degree of uncertainty involved in simulating the ball's path. This was accomplished by placing an increasing number of square obstacles throughout the scene: simulation therefore required accommodating the growing possibility of the ball bouncing off one or more obstacles before hitting one of the targets, making simulation results less certain. The lowest complexity levels for each scenario and containment level had no such obstacles, while the highest complexity levels had eight obstacles spread throughout the scene (see Figure 2).

Each unique scenario, containment, complexity combination was rotated 90, 180, and 270 degrees to allow for more trials and to prevent the scenarios from being too predictable. In addition, there were 64 *distractor* trials that were identical to the high containment trials in each scenario, except that both targets were placed inside or outside the containing structure. These were added to prevent participants from adopting a strategy of assuming that every trial would have a containment structure or other topological best guess once they had seen a number of trials in which that was the case.

For each trial, we captured participants' accuracy (correct or incorrect) and response time. Previous results using the same target task have provided evidence that participants are likely to make simulated inferences for this task across a range of scenarios and further that response time is correlated with time required to simulate the outcome (Hamrick, Smith, Griffiths, & Vul, 2015; Smith et al., 2017). We ex-

pected response time to be a reasonable measure of participants' reliance on simulation for the inference in the task: as the complexity of the simulation required to make a prediction increased, so too should the response time. In contrast, predictions made via topological inference should show little change in response time as complexity of the scene increased. When one of the targets was clearly contained in the same space as the ball, the uncertainty or duration of the ball's simulated path should not have had any bearing on judging which target the ball would hit first if participants were taking advantage of this containment information. Therefore, we expected to see a relationship between simulation complexity and response time which held for trials in which participants made a prediction by simulation but failed to hold for trials where participants were instead using visual cues which facilitated more coarse topological predictions.

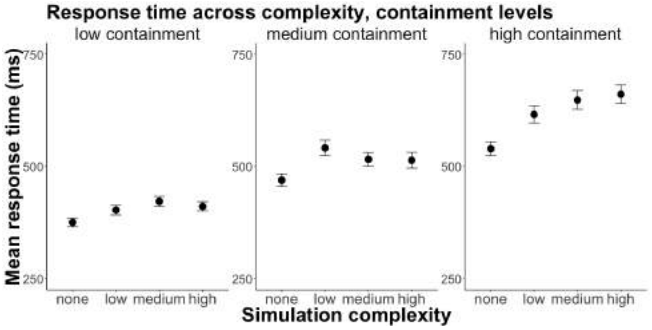
Results

Two of the 81 participants were excluded from analysis due to technical difficulties logging their data. For each participant, we excluded data from the 64 distractor trials. These were included in the experiment to prevent the inference that all trials would have a more and a less contained target. However, the data from these trials is not relevant to the present analyses. All subsequent analyses were therefore conducted with data from 79 participants over 192 trials (twelve trials for each of the four scenarios, rotated each of 0, 90, 180, and 270 degrees). For all analyses, response times were log-transformed to account for their skewed distribution (Whelan, 2008) but transformed back for reporting and display.

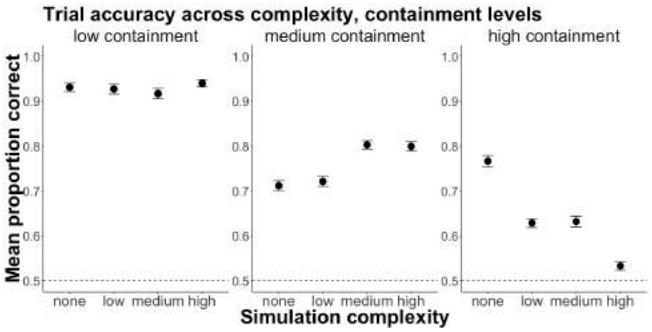
Response times

To assess whether participants were avoiding costly simulations when simulations were particularly uncertain and topological conditions supported more efficient predictions, we examined average response times across each level of complexity and containment. The results are illustrated in Figure 3a. We were interested in comparing response times in low-containment trials to high-containment trials, where an efficient topological prediction about which target the ball would hit was available. Rather than a stabilization or even a decrease in response time as complexity increased in high containment trials (signaling a switch to topological prediction), Figure 3a shows that response times increased progressively as containment increased from low to high and within each containment level as complexity increased from none to high. Moreover, the high-containment trials were slower, and less accurate (Figure 3b), than low-containment trials.

In a repeated measures ANOVA, response times vary with containment and complexity, ($F(2, 156) = 55.63$, $p < 0.001$ and $F(3, 234) = 8.87$, $p < 0.001$, respectively). However, consistent with the fact that participants are not treating complexity differently in high containment trials, there is no containment-complexity interaction ($F(6, 468) = 0.487$). Participants relying on topological information to infer which target the ball would hit in high containment trials would have been able to do so quickly. As complexity increased, so too would the time required to simulate the ball's possible outcomes. Therefore, predictions made via topological analysis in high containment, high complexity trials could potentially be done in less time than required for prediction by simulation in trials with the same degree of complexity but lower containment. Even with complexity levels which make simulation difficult and topological information which makes prediction simple, participants showed no sign of using a containment heuristic.



(a) Response times across conditions



(b) Accuracy across conditions

Figure 3: (a) Mean response time across containment and complexity levels. Despite the availability of simple topological predictions in the high containment, high complexity trials, response time is highest. (b) Mean accuracy across containment and complexity levels. Accuracy steadily decreases at higher containment levels, even though more contained trials would seem to make prediction more certain.

Accuracy

In light of our findings that response times both increased as complexity increased within each containment level and also increased across containment levels, one interpretation is that this pattern was a result of a speed-accuracy tradeoff. Insofar as additional complexity in a given scenario made simulation more difficult and uncertain, participants may have spent more time confirming their predictions without any other change in their simulations or prediction strategies. To test this, we looked at each participant's accuracy in a given containment and complexity level (there are 16 trials in a given containment and complexity level for each participant). The

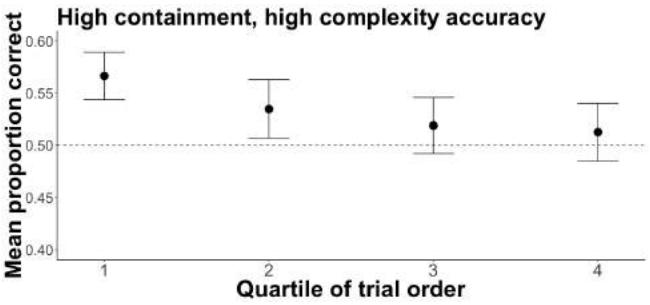


Figure 4: Performance of participants in *high containment, high complexity trials only* by trial order quartile. Accuracy remains close to chance and does not improve over the course of the experiment, suggesting that participants likely did not switch at any point to topological inference cues or other strategies that would have improved their accuracy.

mean accuracy proportions across all participants for each complexity and containment level are shown in Figure 3b.

In contrast to what would be predicted by a speed-accuracy tradeoff in which the containment and complexity levels that participants spent the most time on also have the highest accuracy, mean accuracy steadily decreases from low to high containment scenarios. In low containment trials, mean accuracy was above 90% across all complexity levels, while in high containment and high complexity trials, where participant response times were the largest, accuracy was only nominally above chance (95% CI 51.3% - 55.4%). In a repeated measures ANOVA, both containment and complexity accounted for a significant portion of the variance in accuracy ($F(2, 156) = 829.5$, $p < 0.001$ and $F(3, 234) = 15.7$, $p < 0.001$, respectively), as well as the interaction between them ($F(6, 468) = 61.79$, $p < 0.001$). As the containment and complexity of trials increased, participants spent more time making judgments and their accuracy decreased: these data are inconsistent with an account of prediction in which people process topological features to make the judgment as efficiently as possible. One alternative is that people persist in simulating outcomes in such trials even when alternatives are readily available. Under this account, participants would be expected to simulate more as complexity increased in order to overcome the uncertainty imposed by increases in complexity. They might do this even when increasing levels of containment made topological predictions simple.

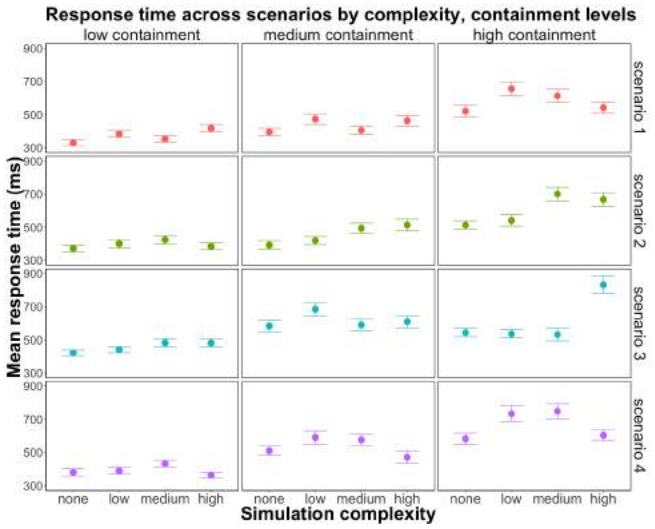
Strategy changes

Another interpretation of the current results is that people may have *eventually* switched to heuristic-based strategies in the more complex trials, but not right away. We predicted that the difficulty of simulation on high complexity trials would encourage participants to employ alternative inference strategies where available. But it may be that the complexity of a trial in and of itself is insufficient to induce strategy change. For example, participants might need to see several complex trials and infer that high complexity trials are likely to recur and are not "one offs". Or, participants might overestimate the accuracy of simulation-based inferences: only after getting wrong answers on complex trials would they pursue alternative inference strategies.

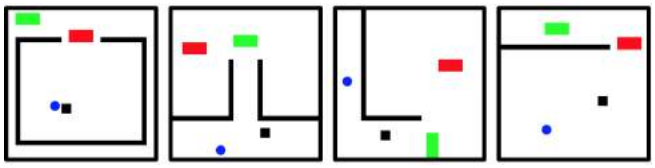
If participants were switching to heuristic-based strategies as a result of familiarity with the task or low accuracy on complex trials, we might expect a difference between high complexity, high containment trials encountered *earlier* versus *later* in the experiment. This difference would be revealed in changes in accuracy over the course of the experiment: if participants eventually ended up using a topological inference strategy for these high containment, high complexity trials, we would expect near perfect accuracy for any such trials. Figure 4 shows accuracy on *high containment, high complexity trials only*, arranged by the trial order quartile in which participants saw them. Participants performed relatively poorly on the high containment, high complexity trials at the outset. Critically, there is no sign of improvement over the course of the experiment: in an ANOVA with participants' mean accuracy by quartile, accuracy did not vary significantly across quartiles ($F(1, 78) = 2.359$, $p = 0.129$). If participants had switched to a topologically based inference strategy, we would expect an increase in accuracy since high complexity, high containment trials enable a very confident containment-based solution. Figure 4 suggests that directionally, participants appeared to get worse on the high accuracy, high containment trials and remain fairly inaccurate throughout the experiment.

Scenario and rotation differences

A third account for the higher response times and lower accuracy as containment and complexity increased is that this overall pattern reflects a great deal of variance across scenarios. In a repeated measures ANOVA of response time that adds scenario on top of containment and complexity, there are significant main effects of containment and complexity (as described before) as well as scenario ($F(3, 234) = 64.89$, $p < 0.001$), reflecting the fact that participants' response times seemed to vary across scenarios. In Figure 5, we show mean response times across containment and complexity levels but further broken down by scenario. The pattern of response times is fairly consistent for low containment trials but the directionality of response times as complexity increases in high containment trials varies across scenarios. In scenarios 1, 2, and 4, response times in high containment trials stabilize or diminish at higher complexity levels, which is qualitatively consistent with our hypothesis that participants would make faster predictions when topological conditions supported a coarse analysis and made simulation highly uncertain. Indeed, the effects of containment and complexity are not homogeneous across scenarios, revealed by significant interactions between scenario and containment ($F(6, 468) = 6.935$, $p < 0.001$), and scenario and complexity ($F(9, 702) = 4.662$, $p < 0.001$). The three-way interaction between scenario, containment, and complexity is weaker, but also significant ($F(18, 1404) = 1.626$, $p = 0.047$), indicating that the



(a) Response time broken down by scenario



(b) High containment example of each scenario (1–4)

Figure 5: (a) Response times are consistent across scenarios in lower containment and complexity levels but diverge considerably at higher containment and complexity levels. (b) A high containment (low complexity) trial for each scenario. Complexity was increased by adding more square obstacles.

pattern in scenario 3 is quite unusual. However, we cannot confidently conclude that any of the scenarios would reliably produce the sort of two-way interaction between containment and complexity that our hypothesis predicts.

Finally, it is worth noting that even though rotated versions of the trials were identical in configuration and ball movements, simply turned 90, 180, or 270 degrees, participants may have treated rotated versions of the trials differently. A repeated measures ANOVA of response time as a function of scenario and rotation found that rotation accounted for a significant amount of the variance ($F(3, 234) = 6.995$, $p < 0.001$), scenario was significant (as outlined above) and that there was a significant interaction between scenario and rotation ($F(9, 702) = 2.939$, $p = 0.002$). Whether this reflects some sort of bias towards e.g. the targets being at the top of the screen is unclear.

Conclusion

In this study we presented participants with physical prediction tasks that simultaneously varied the degree to which a simple containment heuristic could be used to make effective predictions and the complexity required to simulate outcomes instead. Our hypothesis was that as increasing complexity made simulations more and more uncertain and effortful, participants would pursue less costly topological prediction strategies. When conditions permitted such knowledge-based predictions, response times would reflect the rapid and efficient use of containment heuristics. We found no evidence of participants flexibly using heuristics when simulation was complex. In fact, participants spent the longest on trials that had the highest degree of containment; meanwhile, their accuracy was lowest on these same trials.

Why might participants have spent more time and been less accurate on trials where a simple containment-based prediction was available? First, it's possible that the structure of the task at the outset biased participants towards a simulation-based strategy in a way that might have been difficult to overcome, even when complexity of trials made simulation difficult. Earlier work that used static control stimuli in a similar task found evidence that people used simulation even with static stimuli (Smith et al., 2017). Therefore, it's possible that participants had a high "fixedness" when confronted with complex trials. Additionally, it has been shown that when explicitly instructed to apply distinct simulation strategies, participants show notable performance differences on mental rotation tasks (Flusberg & Boroditsky, 2011). In the present study, participants were not instructed to simulate or make a containment-based inference and were solving the problems as they naturally would, but future work might look at how instructions play a role in guiding more efficient strategies.

Alternatively, Smith et al. (2017) suggest that if simulation and alternative prediction strategies are running in parallel, detecting scenarios in which participants switch from a default simulation-based prediction to a more qualitative one that is quicker but more coarse might require enough time for simulation to *run out*. In the present study, average response times in the slowest high containment, high complexity trials were still less than one second on average (see Figure 3a). Perhaps participants, upon finding that they were not able to make an accurate simulation-based prediction on these trials, still did not spend long enough attempting an accurate answer to detect the containment relationship or make a prediction based on such a holistic topological feature. Insofar as the higher containment and complexity trials simply required longer to visually process the full scene, participants may have resorted to an even quicker and more general heuristic in order to respond quickly, such as the target that was the shortest Euclidean distance or seemed most directly along the ball's initial path irrespective of obstacles. Alternatively, participants may have simply persisted in slower and less efficient simulations on high containment, high complexity trials rather than pursue alternate strategies (Hamrick et al., 2015). Future research will need to carefully design stimuli in order to control for the many ways participants might make predictions and consider other hypotheses that allow for manipulation of the uncertainty of simulations during prediction.

Acknowledgments

We thank our reviewers for their thoughtful suggestions.

KAS is supported by CBMM funded by NSF STC award CCF-1231216, and ONR grant N00014-13-1-0333.

References

Bates, C. J., Yildirim, I., Tenenbaum, J. B., & Battaglia, P. (2018). Modeling human intuitions about liquid flow with particle-based simulation. *arXiv:1809.01524 [cs, q-bio]*.

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. In *Proceedings of the national academy of sciences.*

Callaway, F., Hamrick, J. B., & Griffiths, T. (2017). Discovering simple heuristics from mental simulation. In *Proceedings of the 39th annual meeting of the cognitive science society.*

Davis, E., & Marcus, G. (2015). The scope and limits of simulation in cognitive models. *arXiv preprint arXiv:1506.04956*.

Davis, E., & Marcus, G. (2016). The scope and limits of simulation in automated reasoning. *Artificial Intelligence*, *233*, 60–72.

Davis, E., Marcus, G., & Frazier-Logue, N. (2017). Commonsense reasoning about containers using radically incomplete information. *Artificial Intelligence*, *248*, 46-84.

Flusberg, S., & Boroditsky, L. (2011). Are things that are hard to physically move also hard to imagine moving? *Psychonomic bulletin & review*, *18*(1), 158-164.

Gerstenberg, T., Peterson, M. F., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2017). Eye-Tracking Causality. *Psychological Science*, *28*(12), 1731-1744. doi: 10.1177/0956797617713053

Gigerenzer, G., & Todd, P. (1999). *Simple heuristics that make us smart*. USA: Oxford University Press.

Hamrick, J. B., Battaglia, P. W., Griffiths, T. L., & Tenenbaum, J. B. (2016). Inferring mass in complex scenes by mental simulation. *Cognition*, *157*, 61–76.

Hamrick, J. B., Smith, K. A., Griffiths, T. L., & Vul, E. (2015). Think again? optimal mental simulation tracks problem difficulty. In *Proceedings of the 37th annual meeting of the cognitive science society.*

Smith, K. A., Battaglia, P. W., & Vul, E. (2018). Different Physical Intuitions Exist Between Tasks, Not Domains. *Computational Brain & Behavior*, *1*(2), 101-118. doi: 10.1007/s42113-018-0007-3

Smith, K. A., Dechter, E., Tenenbaum, J. B., & Vul, E. (2013). Physical predictions over time. In *Proceedings of the 35th annual meeting of the cognitive science society.*

Smith, K. A., de Peres, F., Vul, E., & Tenenbaum, J. B. (2017). Thinking inside the box: Motion prediction in contained spaces uses simulation. In *Proceedings of the 39th annual meeting of the cognitive science society.*

Smith, K. A., & Vul, E. (2013). Sources of uncertainty in intuitive physics. *Topics in cognitive science*, *5*, 185–199.

Strickland, B., & Scholl, B. (2015). Visual perception involves event-type representations: The case of containment versus occlusion. *Journal of Experimental Psychology: General*, *144*, 570–580.

Ullman, T. D., Spelke, E., Battaglia, P., & Tenenbaum, J. B. (2017). Mind games: Game engines as an architecture for intuitive physics. *Trends in Cognitive Sciences*, *21*, 649–665.

Whelan, R. (2008). Effective analysis of reaction time data. *The Psychological Record*, *58*, 475–482.

Simplicity and Probability in Human Judgment

Tyler Brooke-Wilson
Massachusetts Institute of Technology, Cambridge, Massachusetts, United States

Jonathan S. Rosenfeld
Massachusetts Institute of Technology, Cambridge, Massachusetts, United States

Matthias Hofer
Massachusetts Institute of Technology, Cambridge, Massachusetts, United States

Junyi Chu
MIT, Cambridge, Massachusetts, United States

Josh Tenenbaum
MIT, Cambridge, Massachusetts, United States

Abstract

Children and adults prefer simpler to more complex explanations, a penchant they share with scientists and philosophers. While the preference has been widely remarked, its mechanisms and justification remain contested (Kitcher1987, Lombrozo 2007, Lombrozo2015). Explanations for the simplicity preference have included over-hypotheses, resource rationality, pragmatic justifications, and quirks of the hypothesis generation process. We present a model of key results from Pacer and Lombrozo (Pacer2017) and show that one form of the simplicity bias can be explained on probabilistic grounds alone. This modeling work provides an explanation for one manifestation of the simplicity bias, and allows us to formalize questions within the 'Explanation for Best Inference' Framework (Lombrozo2015), asking explicitly what makes the best explanation 'best.'

Memory maintenance of gradient speech representations is mediated by their expected utility

Wednesday Bushong (wbushong@ur.rochester.edu) and T. Florian Jaeger (fjaeger@ur.rochester.edu)
Department of Brain and Cognitive Sciences, University of Rochester

Abstract

Language understanding requires listeners to quickly compress large amounts of perceptual information into abstract linguistic categories. Critical cues to those categories are distributed across the speech signal, with some cues appearing substantially later. Speech perception would thus be facilitated if *gradient sub-categorical* representations of the input are maintained in memory, allowing optimal cue integration. However, indiscriminate maintenance of the high-dimensional signal would tax memory systems. We hypothesize that speech perception balances these pressures by maintaining gradient representations that are expected to facilitate category recognition. Two perception experiments test this hypothesis. Between participants, an initial exposure phase manipulated the utility of information maintenance: in the *High-Informativity* group, following context always was informative; in the *Low-Informativity group*, following context always was uninformative. A subsequent test phase measured the extent to which participants maintained gradient representations. The Low-Informativity group showed less maintenance, compared to the High-Informativity group (Experiment 1). We then increased the task demands and made the targets of the manipulation less obvious to participants (Experiment 2). We found a qualitatively similar pattern. Together, these results suggest that listeners are capable of allocating memory to gradient representations of the speech input based on the expected utility of those representations.

Keywords: speech perception; cue integration; memory; expected utility

Introduction

Spoken (and signed) language is a temporally unfolding signal. In order to comprehend language, humans must quickly compress kilobits of information per second into abstract linguistic representations and meanings that contain more manageable amounts of information. At the same time, cues to linguistic categories often do not temporallly co-occur in neatly delimited segments of the speech signal, but rather are distributed across the signal. For example, one of the primary cues to stop voicing in English is the duration of the *preceding* vowel (Klatt, 1976). To make optimal categorization judgments, listeners must retain some sub-categorical information about the preceding vowel in memory in order to integrate it with later-arriving information (i.e., the stop itself). This kind of information distribution is typical across languages and can occur at several timescales: cues to sound categories can come not only from proximate acoustic properties, but also from, e.g., later lexical and semantic context that can occur anywhere. But maintaining rich representations of all incoming input would seemingly overload working memory. Thus, many theories of language processing claim that listeners simply do not maintain gradient representations of the input on any significant timescale, but instead immediately compress input into abstract representations (Just & Carpenter, 1980; Christiansen & Chater, 2016). According to these accounts, listeners throw away rich representations of the input as soon as a categorical perceptual judgment has been made.

However, a growing body of literature has suggested that listeners can and do maintain sub-categorical representations of prior input (McMurray, Tanenhaus, & Aslin, 2009), even at quite long perceptual timescales (Connine, Blasko, & Hall, 1991; Brown-Schmidt & Toscano, 2017; Gwilliams, Linzen, Poeppel, & Marantz, 2018). For example, Connine et al. (1991) exposed participants to sentences like "When the ?ent in the [fender/campground]...", where the ?-segment ranged between /d/ and /t/ (by manipulating one of the primary cues to voicing perception, the voice-onset time or VOT). The context following the ?-segment contained additional semantic context toward the identity of the original word. Participants had to categorize whether they had heard the word "tent" or "dent" in the sentence. Connine and colleagues found that participants' categorizations were influenced *both* by the VOT of the sound *and* by subsequent context, suggesting that listeners maintained a gradient representation of the initial sound for later use in cue integration and categorization. Subsequent studies have confirmed that listeners can maintain sub-categorical representations well beyond word boundaries (Szostak & Pitt, 2013; Bushong & Jaeger, 2017; Bicknell, Bushong, Tanenhaus, & Jaeger, under review).

How is this possible when language contains too much information to be held in memory indefinitely? We hypothesize that listeners use a memory strategy based on *expected utility*: the more important a piece of input is deemed to be, the more likely a detailed gradient representation should be maintained in memory; the less important the input, the more likely a categorical, less detailed representation will be maintained. In the case of the VOT and subsequent context example above, we can operationalize utility as the likelihood that *subsequent* context will be informative for categorization of the current input—if there is likely to be later information relevant to categorization, listeners should maintain a gradient representation of the speech input in order to be able to use it during cue integration (when the relevant subsequent context arrives).

In order to test this proposal, we conduct two experiments where we manipulate the probability that subsequent content in the sentence is relevant to the target word that participants have to categorize. Participants listen to sentences where a critical target word is acoustically manipulated to range between *tent* and *dent*. Like in the experiments by Connine and colleagues, these words are embedded in sentences. Unlike in earlier work, one group of participants hears sentences

that always contain subsequent contextual information that is *informative* for categorization, whereas another group of participants hears sentences with *uninformative* subsequent context. Following exposure, we then test how much participants in the two groups maintain sub-categorical representations about the target word. Here, we operationalize whether participants are maintaining sub-categorical representations of the initial target as the extent to which each group integrates *both* acoustics of the target word *and* subsequent context into their categorization responses.

General Methods

Participants

We recruited 128 native English-speaking participants each for Experiments 1 and 2. Participants were recruited from Amazon Mechanical Turk and rewarded $3.00 for their participation. Participants could only participate in in either Experiment 1 or Experiment 2. The average age of our participants was comparable across experiments suggesting they had similar amounts of language experience (Experiment 1: 37.55 ± 12.04; Experiment 2: 34.14 ± 8.11).

Materials

We take the paradigm from Bushong and Jaeger (2017) as a starting point for our experiments. We constructed 12 sentence triplets like the following:

(1) After the ?ent Sue had found in the campgrounds collapsed, we went to a hotel. **(*tent*-biasing context)**

(2) After the ?ent Sue had found in the teapot was noticed, we threw it away. **(*dent*-biasing context)**

(3) After the ?ent was noticed, we continued on our way. **(neutral context)**

We manipulated two aspects of the sentence stimuli. First, we acoustically manipulated the "?" to range between /d/ and /t/ by changing the value of its voice-onset time (VOT), the primary cue distinguishing voiced from voiceless syllable-initial stop consonants in English. Based on norming and previous experiments, we chose to test VOT values of 10, 40, 50, 60, 70, and 85ms to cover a perceptual range from /d/ to /t/ with ambiguous points in between. Each VOT step occurred equally often. Second, we manipulated whether later context biased toward a /t/-interpretation (1) , /d/-interpretation (2), or neither (3). Informative words in the subsequent context—if present—occurred between 6-9 syllables after the target word, as in (1) and (2) above.

Procedure

Both experiments consisted of two phases (participants were unaware of this implicit structure): Exposure (72 trials) and Test (48 trials). Participants were randomly assigned to one of two groups: Low-Informativity exposure and High-Informativity exposure. During exposure, the Low-Informativity exposure group only heard sentences with neutral subsequent context (e.g., sentence (3) above), such that the only relevant information to sound categorization was VOT. The High-Informativity exposure group, by contrast, always heard sentences that contained informative later context (split evenly between /t/-biasing and /d/-biasing contexts), as in previous studies. In the test phase, both groups heard sentences that contained informative later context (split evenly between /t/-biasing and /d/-biasing contexts). This allowed us to assess context effects during the test phase, following Connine et al. (1991). Figure 1 illustrates the design of both experiments.

Both during exposure and test, participants' task was simply to categorize whether they heard one of two alternative words after they heard the full sentence. In Experiment 1, participants always made judgments about our critical target words of interest—i.e., they were asked whether they heard "tent" or "dent" on every trial. In Experiment 2, on half of all trials, participants instead had to categorize another word in the sentence (e.g., for sentence (3) above they were asked whether they heard "way" or "day"). We motivate this difference in design after presenting Experiment 1.

Predictions

We analyze responses from the test phase. Specifically, we analyze the influence of VOT and subsequent context on categorization responses to assess whether listeners maintained gradient representations of VOT. If participants maintain sub-categorical information about the /t/ and /d/ in the target word *?ent* until the end of the sentence, we should see effects of both VOT and context. Critically, if listeners can monitor the utility of subsequent context for the target word, and if expectations about this utility affect the degree to which listeners maintain sub-categorical representations, we should see that the main effect of context is smaller in the Low-Informativity exposure group, compared to the High-Informativity exposure group. Note that observing a continuous effect of VOT is not sufficient to establish that listeners maintain gradient representations of VOT since this could still reflect initial deterministic categorizations; it is the ability for listeners to integrate continuous VOT information with later-arriving context that is critical (for more discussion of this point, see Bicknell et al., under review).

Experiment 1

The goal of Experiment 1 was to test the simplest version of our proposal, whether listeners can adapt their expectations about the utility of subsequent context and use them to guide whether to maintain gradient representations of initial acoustic input.

Analysis

Following previous work (Bicknell et al., under review; Bushong & Jaeger, 2017), we excluded participants whose

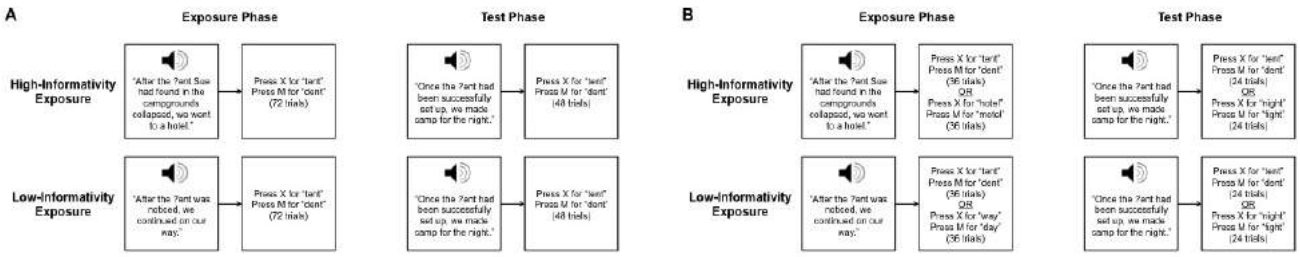


Figure 1: Design of the two experiments. **A**: Experiment 1. **B**: Experiment 2.

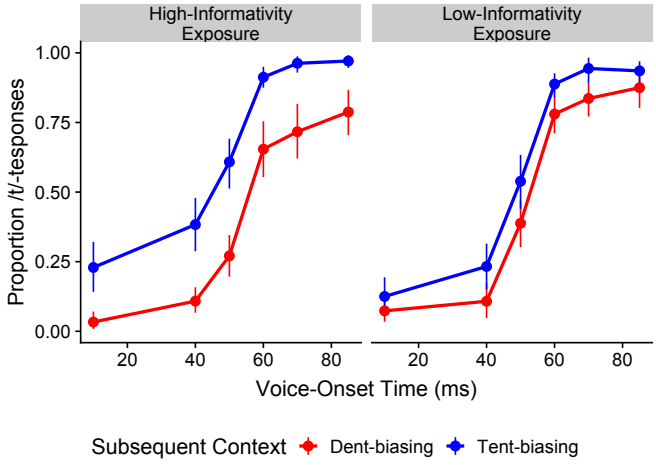


Figure 2: Experiment 1 test phase categorization results by VOT, subsequent context, and exposure group. Error bars are bootstrapped 95% confidence intervals over subject means.

categorization responses were not modulated by VOT, suggesting that they did not understand the task. This resulted in the exclusion of 11 participants for Experiment 1 (8.6%).

We fit mixed-effects logistic regression (Jaeger, 2008) predicting the proportion of /t/ responses in the test phase from VOT (z-scored to help with model convergence), squared VOT (z-scored), subsequent context (sum-coded: 1 = *tent*-biasing vs. -1 = *dent*-biasing), group (sum-coded: 1 = High-Informativity vs. -1 = Low-Informativity exposure), the two-way interaction of group with subsequent context, and the two-way interaction of subsequent context and squared VOT.[1] We included the interaction between squared VOT and context to test whether listeners' behavior was ideal observer-like (for a longer discussion on why this is important, see Bicknell et al., under review). The analysis also contained the full random effects structure that allowed model convergence. Analyses were conducted in the `lme4` package in `R` (Bates, Maechler, Bolker, Walker, et al., 2014).

Results

Figure 2 shows /t/-responses by group, VOT, and subsequent context over the test phase.

We found main effects of VOT ($\hat{\beta} = 2.87, p < 0.001$), squared VOT ($\hat{\beta} = 1.15, p = 0.01$) and subsequent context ($\hat{\beta} = 0.88, p < 0.001$) such that participants were more likely to respond /t/ when VOT increased and when subsequent context was /t/-biasing. The interaction between squared VOT and context was not significant ($\hat{\beta} = 0.06, p = 0.47$), replicating previous findings that suggest rational information integration.

There was no main effect of group on /t/-responses ($\hat{\beta} = 0.01, p = 0.89$). Crucially, there was a significant interaction between group and context ($\hat{\beta} = 0.41, p = 0.001$) such that the High-Informativity group showed a larger context effect than the Low-Informativity group. A simple effects analysis[2] revealed that both groups showed a significant context effect in the same direction (High-Informativity group: $\hat{\beta} = 1.25, p < 0.001$; Low-Informativity group: $\hat{\beta} = 0.47, p = 0.007$).

Discussion

We found that both exposure groups showed effects of VOT and subsequent context on their categorization responses, suggesting that they maintained gradient representations of speech input (VOT) in memory. As predicted, however, the effect of context was much smaller for the Low-Informativity group as compared to the High-Informativity group. These results suggest that the average informativity of later context influences whether listeners maintain gradient information about VOT in memory.

Experiment 1 shares with most previous work on the maintenance of sub-categorical representations that our paradigm involved a large degree of repetition (see Connine et al., 1991; Szostak & Pitt, 2013; Bicknell et al., under review; Bushong & Jaeger, 2017). This raises questions about the extent to which the results of Experiment 1 generalize to scenarios that more closely resemble the task demands of everyday language processing. In particular, participants in Experiment 1 were asked to make categorization judgments about the same critical target words of interest (*tent* and *dent*) throughout the entire experiment. This target word always occurred in

[1] Fixed effects R formula: `/t/-response ~ VOT + VOT^2 Context + Group + Group:Context + VOT^2:Context`.

[2] R formula: `/t/-response ~ VOT + Group / Context` plus the same random effects as the main analysis.

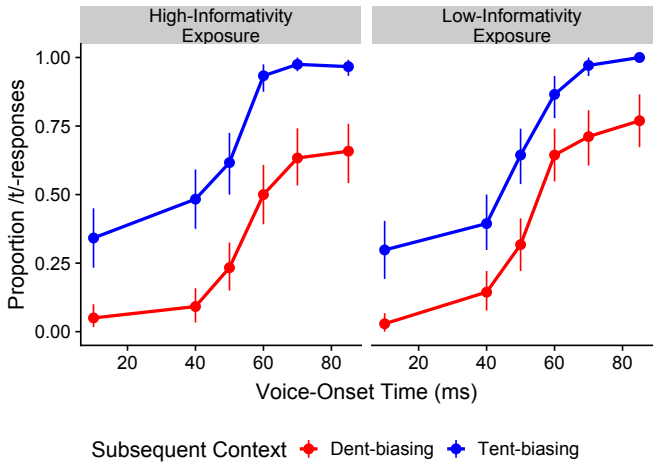


Figure 3: Experiment 2 Test Phase categorization results by VOT, subsequent context, and group. Error bars are bootstrapped 95% confidence intervals over subject means.

the same position within the 12 different sentence contexts. Additionally, subsequent information context (in the High-Informativity exposure group) always occurred about 6-9 syllables after the target. All of these factors likely directed participants' attention towards the target words.

One possibility is thus that the large context effects in the High-Informativity group is due to participants limiting their attention solely on the target word and the informative context. Similarly, the small context effect in the Low-Informativity group might be due to participants recognizing that they can do the task just as well while tuning out after hearing the target word, as they always get asked about the same target word. Experiment 2 presents a first step towards addressing this question.

Experiment 2

Experiment 2 is identical to Experiment 1, except that participants in both exposure groups of Experiment 2 only made judgments about the target words of interest (*tent* and *dent*) on half of the trials. On the other half of trials, participants were instead asked to categorize another word in the sentence; these alternate target words were never the informative subsequent context words for our critical words of interest—i.e., participants were never asked about the word "campgrounds" in sentence (1) above. This change from Experiment 1 had two purposes: (i) it directed participants' attention away from the target words, thus allowing us to test how participants behave when targets are not perfectly predictable, as in natural speech; and (ii) made it more likely that participants in both exposure groups remained attentive throughout the entire sentence. One of our concerns about Experiment 1 is that participants in the low-informativity group may have just 'tuned out' the rest of the sentence, and thus subsequent biasing context in the test phase, after hearing the target word. We reasoned that asking participants about words near the end of the sentence would generally increase attention toward those areas of the sentence, thus making it more likely that they heard and processed the later context. Participants were not told before the experiment what types of words they would be making judgments about or how often, so it is unlikely that they had a priori expectations about the distribution of target words (beyond general expectations about what kinds of words are usually tested in experiments, e.g. content words).

Analysis

Analyses were identical to Experiment 1. Our exclusion criteria resulted in the removal of 16 participants (12.5%) from analysis.

Results

Figure 3 shows /t/-responses by group, VOT, and subsequent context over the test phase.

We found main effects of VOT ($\hat{\beta} = 0.71, p = 0.01$), squared VOT ($\hat{\beta} = 1.68, p < 0.001$) and subsequent context ($\hat{\beta} = 1.26, p < 0.001$) such that participants were more likely to respond /t/ when VOT increased and when subsequent context was /t/-biasing. The interaction between squared VOT and context was significant ($\hat{\beta} = 0.22, p = 0.01$), in contrast to Experiment 1. In Experiment 2, we did not find an interaction between group and context ($\hat{\beta} = 0.18, p = 0.14$), although the numerical difference was in the same direction. The simple effects analysis revealed that both groups showed a significant context effect in the same direction (High-Informativity group: $\hat{\beta} = 1.39, p < 0.001$; Low-Informativity group: $\hat{\beta} = 1.01, p < 0.001$).

Discussion

The results of Experiment 2 are qualitatively similar to Experiment 1: participants overall showed evidence of integration of VOT and context into their responses, but the effect of context was numerically smaller in the Low-Informativity group than in the High-Informativity group. In contrast to Experiment 1, this difference between groups was not significant. This may suggest that shifting participants' attention away from our main manipulation made it harder to track how informative subsequent context was for our target words. However, it is hard to draw firm conclusions from a null result. To investigate whether the context effect difference is different between the two experiments, we directly compare them.

Interestingly, we also found a significant interaction between squared VOT and context, suggesting that participants' integration of VOT and context was non-optimal—unintuitively, participants seemed to use context more for *unambiguous* stimuli. This seems to contradict previous proposals that context is either integrated as a constant regardless of ambiguity (optimal integration, Bicknell et al., under review), or is used more for ambiguous than unambiguous stimuli (Connine et al., 1991). Since this aspect was not the focus of this experiment we will not discuss it further here, but further work should investigate why we observe suboptimal integration behavior in this more naturalistic setting

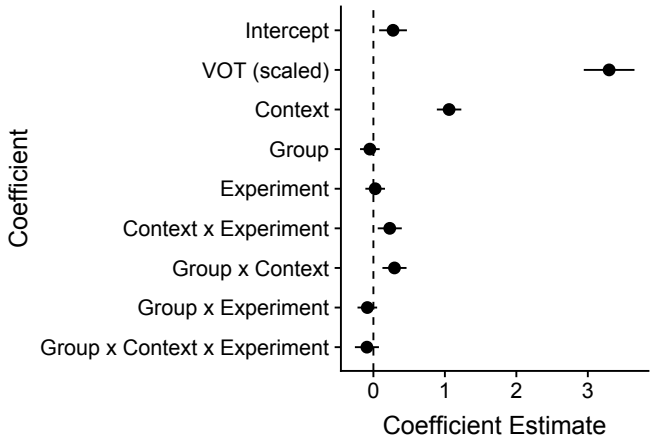


Figure 4: Model coefficients from combined analysis of Experiments 1 and 2. Error bars are 95% confidence intervals.

(for a first step in modeling sub-optimal integration strategies in this paradigm, see Bushong & Jaeger, 2019).

Comparison of Experiment 1 & 2

Since we observed a different pattern of results in Experiments 1 and 2, we conducted a post-hoc combined analysis in an attempt to assess overall evidence for the group by context interaction across experiments. In order to test the relationship between the group and context interaction and experiment, we fit a combined regression model to both of the datasets, allowing experiment (sum coded such that Experiment 1 = -1, Experiment 2 = 1) to interact with context and group. The other fixed effects remained the same as the above analyses for both experiments.

Results

Figure 4 shows the results of the combined analysis of Experiments 1 and 2.

In the combined analysis, we found main effects of VOT ($\hat{\beta} = 3.3, p < 0.001$) and subsequent context ($\hat{\beta} = 1.06, p < 0.001$). There was also a significant interaction between group and context ($\hat{\beta} = 0.29, p < 0.001$) such that the context effect was larger for the High-Informativity group than the Low-Informativity group. Critically, there was no three-way interaction between experiment, group, and context ($\hat{\beta} = -0.09, p = 0.29$), suggesting that the two-way interaction between group and context did *not* differ significantly across experiments.

We also found an unexpected two-way interaction between experiment and context ($\hat{\beta} = 0.23, p = 0.008$), such that the context effect was larger in Experiment 2 than Experiment 1. We return to this difference below. None of the other effects reached significance ($p > .2$).

General Discussion

When comprehending language, listeners must process thousands of bits of incoming information per second, compressing it into more manageable abstract representations. However, sub-categorical information about input can be useful to maintain in memory for later integration with relevant cues. Previous work has shown that listeners seem to be able to maintain such gradient representations in memory for up to several seconds. Here, we asked how this is possible when maintaining gradient representations of all incoming input would presumably overload short-term memory. We proposed that these effects may be driven by the *expected utility* of maintaining such information. In the case of these experiments, we tested this by manipulating the informativity of subsequent context: we reasoned that if subsequent context is likely to be informative for phonemic categorization, then the utility of maintaining gradient representations of VOT is higher, since it will be available for cue integration.

In Experiment 1, we found that both experimental groups maintained gradient representations of VOT over the timescale tested in this experiment (6-9 syllables). In line with our predictions, this effect was significantly smaller in the Low-Informativity group compared to the High-Informativity group. This provides support for our hypothesis that expected utility mediates maintenance of gradient representations of speech input in memory.

In order to make our experiments more naturalistic, we added an additional manipulation in Experiment 2. Participants were not always making judgments about our critical target words of interest. Instead, on half of all trials (during exposure and test), they made categorization judgments about non-critical words in the sentence. This change in the design takes a (small) step towards the task demands of natural language use: listeners don't necessarily *a priori* know which parts of the speech input they will need to comprehend and respond to. When we made this change, we observed the same numerical trend toward a smaller context effect in the Low-Informativity group, but this difference was not significant ($p = 0.14$). It is possible our manipulation in Experiment 2 successfully directed participants' attention away from the *tent*- and *dent*-biasing context, so that participants had a harder time estimating the informativity of subsequent context.[3] A follow-up combined analysis found no evidence that there was a difference in the interaction between the two experiments, though such an interaction might have been difficult to detect. We tentatively conclude that both experiments support the hypothesis that listeners maintain gradient representations according to their expected utility, but further experimentation with similar design is needed.

Of note, our follow-up analysis also found that participants in Experiment 2 exhibited an even *larger* context effect as compared to participants in Experiment 1. This might be seen as surprising: if anything Experiment 2 directed atten-

[3] As suggested by Figures 2 and 3, the difference in the context effects between experiments was driven by the Low-Informativity group: post-hoc analyses revealed that the context effect was larger in the Low-Informativity group in Experiment 2 compared to Experiment 1, but there were no differences in the High-Informativity group.

tion *away* from the critical word, compared to Experiment 1. In Experiment 2, participants had to make categorization judgments about words that occurred in different parts of the sentence rather than only the *tent* and *dent*. Critically, participants were never asked to make judgments about the *tent*- and *dent*-biasing context. The large context effect in Experiment 2 would thus seem to suggest that maintenance of gradient representations in memory is the default during speech perception, rather than the exception. While this interpretation stands in stark contrast with received wisdom (Just & Carpenter, 1980; Christiansen & Chater, 2016), it is line with a number of other recent findings that have found maintenance, for example, on the first trial of experiments (Bushong & Jaeger, 2017) or for lexically heterogeneous stimuli in naturalistic task-based language use (Burchill, Liu, & Jaeger, 2018; Brown-Schmidt & Toscano, 2017). Thus the present results may suggest that 'turning off' maintenance, rather than continued maintenance of gradient representations, requires sustained attention to specific, known targets.

Why would listeners show such a robust maintenance effect? One possible explanation consistent with our expected utility proposal is that natural language is *typically* informative: not only are low-level features like acoustic cues highly correlated even at long distances (providing helpful redundancy in light of perceptual inferences over noisy input, Hermansky, 2018), speakers talk about coherent topics that naturally provide semantic context that adds categorization-relevant information about the speech signal. Given these long-distance informational dependencies, maintaining gradient representations will typically be beneficial since it allows for optimal integration of these cues (Bicknell et al., under review). This would explain why we observe robust maintenance effects in paradigms that test sentences which follow these general natural constraints.

The present paradigm shares some caveats with previous work (Bicknell et al., under review; Connine et al., 1991; Szostak & Pitt, 2013): Experiments 1 and 2 involve a high degree of repetition; very much unlike in everyday language use, participants had to categorize the same word dozens of times. While the results of Experiment 2 show that listeners *do* maintain gradient representations about the speech input even when the target word is not *perfectly* predictable, Experiment 2 still allowed listeners to limit their attention—and maintenance of gradient representations: the critical target words (*tent* and *dent*) were the target of categorization on half of all trials. It is thus an open question whether equally strong maintenance of gradient representations is observed when it is less clear which aspects of the speech signal will turn out to be particularly relevant later (for preliminary evidence, see Burchill et al., 2018; Brown-Schmidt & Toscano, 2017).

Acknowledgments

This work was partially funded by NIHCD R01 HD075797 (to T.F.J.) and an NSF NRT #1449828 fellowship (to W.B.). The authors would like to thank Evan Hamaguchi, Nicole Vieyto, and Chelsea Marsh for assistance with stimulus sentence creation and recording.

References

Bates, D., Maechler, M., Bolker, B., Walker, S., et al. (2014). lme4: Linear mixed-effects models using eigen and s4. *R package version*, *1*(7), 1–23.

Bicknell, K., Bushong, W., Tanenhaus, M. K., & Jaeger, T. F. (under review). Listeners can maintain and rationally update uncertainty about prior words.

Brown-Schmidt, S., & Toscano, J. C. (2017). Gradient acoustic information induces long-lasting referential uncertainty in short discourses. *Language, Cognition and Neuroscience*, *32*(10), 1211–1228.

Burchill, Z., Liu, L., & Jaeger, T. F. (2018). Maintaining information about speech input during accent adaptation. *PloS one*, *13*(8), e0199358.

Bushong, W., & Jaeger, T. F. (2017). Maintenance of perceptual information in speech perception. In *Proceedings of the thirty-ninth annual conference of the cognitive science society.*

Bushong, W., & Jaeger, T. F. (2019). Modeling long-distance cue integration in spoken word recognition. In *Proceedings of naacl 2019 cognitive modeling and computational linguistics workshop.*

Christiansen, M. H., & Chater, N. (2016). The now-or-never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, *39*.

Connine, C. M., Blasko, D. G., & Hall, M. (1991). Effects of subsequent sentence context in auditory word recognition: Temporal and linguistic constraints. *Journal of Memory and Language*, *30*(1), 234–250.

Gwilliams, L., Linzen, T., Poeppel, D., & Marantz, A. (2018). In spoken word recognition, the future predicts the past. *Journal of Neuroscience*, *38*(35), 7585–7599.

Hermansky, H. (2018). Coding and decoding of messages in human speech communication: Implications for machine recognition of speech. *Speech Communication*, 112–117.

Jaeger, T. F. (2008). Categorical data analysis: Away from anovas (transformation or not) and towards logit mixed models. *Journal of memory and language*, *59*(4), 434–446.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological review*, *87*(4), 329–354.

Klatt, D. H. (1976). Linguistic uses of segmental duration in english: Acoustic and perceptual evidence. *The Journal of the Acoustical Society of America*, *59*(5), 1208–1221.

McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2009). Within-category vot affects recovery from lexical garden-paths: Evidence against phoneme-level inhibition. *Journal of memory and language*, *60*(1), 65–91.

Szostak, C. M., & Pitt, M. A. (2013). The prolonged influence of subsequent context on spoken word recognition. *Attention, Perception, & Psychophysics*, *75*(7), 1533–1546.

Executive Functions in Aging: An Experimental and Computational Study of the Wisconsin Card Sorting Task

Andrea Caso
Birkbeck, University of London, London, United Kingdom

Richard Cooper
Birkbeck, University of London, London, United Kingdom

Abstract

In this paper we explore the effect of normal aging on executive function and present a computational account of the effect of aging in a standard executive task. We tested 25 younger adults and 25 older adults (both with no known neurological condition) on the Wisconsin Card Sorting Task (WCST), a classic test of executive function. The test produces multiple measures related to the types of error made by participants, the rate of learning, and so on. As hypothesised, results show no difference between the groups in the number of perseverative errors (i.e., in continuing with a previously successful rule in the presence of negative feedback), but a significantly increased tendency for older adults relative to younger adults to commit set loss errors (i.e., to switch away from a rule despite positive feedback). We fit an existing neurocomputational model of the task to the experimental data by searching through the models parameter space in order to find the best set of parameter values for the two different age groups. This leads to a proposition regarding the effect of aging on the value of the epsilon_ctx parameter, which we argue elsewhere reflects cortical dopamine concentration. We further reanalyse the data by clustering participants by performance (rather than by age) and show that there are multiple points in parameter space that fit each cluster of participants. We argue on the basis of this and the behavioural data, that different parameter values reflect different solutions to optimizing task performance, and that older participants may compensate for changes in epsilon_ctx (reflecting dopamine concentration) by effortful changes in other parameters (specifically, by increasing attentional focus).

Taxonomic and Whole Object Constraints: A Deep Architecture

Mattia Cerrato (mattia.cerrato@unito.it)
Dipartimento di Informatica, University of Torino
Corso Svizzera 185, 10149, Torino, Italy

Edoardo Arnaudo (edoardo.arnaudo@edu.unito.it)
Dipartimento di Informatica, University of Torino
Corso Svizzera 185, 10149, Torino, Italy

Roberto Esposito (roberto.esposito@unito.it)
Dipartimento di Informatica, University of Torino
Corso Svizzera 185, 10149, Torino, Italy

Valentina Gliozzi (valentina.gliozzi@unito.it)
Center for Logic, Language, Cognition & Dipartimento di Informatica, University of Torino
Corso Svizzera 185, 10149, Torino, Italy

Abstract

We propose a neural network model that accounts for the emergence of the taxonomic constraint and for the whole object constraint in early word learning. Our proposal is based on Mayor and Plunkett (2010)'s neurocomputational model of the taxonomic constraint and extends it in two directions. Firstly, we deal with realistic visual and acoustic stimuli. Secondly, we model the well-known whole object constraint in the visual component. We show that, despite the augmented input complexity, the proposed model compares favorably with respect to previous systems.

Keywords: Neural Networks; Children; Language acquisition.

Introduction

How do infants learn the referent of words? As Quine (1960) famously pointed out, for every word heard in a given circumstance, there are several possible referents: in order to infer the appropriate one, infants have to rule out several possible alternatives. But how? Markman (1989) proposed that infants rule out inappropriate referents by means of three constraints. By the *taxonomic constraint* children extend words to taxonomically-related objects: when a child hears the word "dog" pronounced by a caregiver while pointing at a specific dog, she generalizes the referent of "dog" to all dogs, not just to the one in front of her. By the *whole object constraint* children assume that novel words refer to objects as a whole, rather than to their parts, substance, color, or the visual context in which it appears. Lastly, by the *mutual exclusivity constraint* children assume that two labels usually do not refer to the same object.

This paper concerns the first two constraints, namely the taxonomic and the whole object constraint.

Our starting point is Mayor and Plunkett (2010)'s neurocomputational model of the taxonomic constraint. Their model provides an account of how the taxonomic constraint may emerge from infant experience, as the result of the interplay between (i) taxonomic organization of visual inputs in visual areas, (ii) phonetic organization of the acoustic inputs in acoustic areas, (iii) Hebbian learning developing connections between the two organizing areas. The model uses self-organizing maps (Kohonen, 2001) and Hebbian learning (Hebb, 1949), which are considered cognitively plausible mechanisms, describing at an abstract level realistic forms of information organization in the brain (Hebb, 1949; Miikkulainen, Bednar, Choe, & Sirosh, 2005). The powerful interplay between these structures allows word-object associations to taxonomically generalize after a single (*one-shot*) joint word object presentation[1].

Here we extend Mayor and Plunkett (2010)'s seminal model in two directions:

1. We intend to investigate whether the taxonomic constraint can emerge from experience if we consider *realistic visual and acoustic* stimuli (photographic images with different size, color, location in the picture, point of view, etc. and audio excerpts embodying spoken words synthesized via software) instead of the very simple, artificially built stimuli examined in the original model. A first effort in this direction was undertaken by (Fenoglio, Esposito, & Gliozzi, 2017), in which, however, only realistic visual stimuli were considered. Here we enrich that proposal by considering visual *and acoustic* realistic stimuli (as well as the whole object constraint, see below). To this purpose, we insert in the model two deep architectures, one convolutional to process visual stimuli and the other recurrent to process realistic acoustic stimuli.

2. We insert the *whole object constraint* in the model. Whether early learned or innate, the capacity of picking up the objects in a scene is present in early infancy (see e.g. Spelke, 1990). However, this primacy of the object concept

[1] For a critical discussion of the breadth of one shot learning and fast mapping see for instance (Yurovsky, Fricker, Yu, & Smith, 2014) or (McMurray, Horst, & Samuelson, 2012).

in visual scene analysis is not present in most recent convolutional neural network (CNN) models, that are the state of the art in vision tasks. In fact, these models usually process visual images as a whole (object and background context together), see e.g., (Zhu, Xie, & Yuille, 2017). Here we overcome this limitation of CNNs by inserting a segmentation module that extracts the object from the visual scene before feeding it to the CNN for feature extraction. In the experimental section we show that the whole object constraint improves the performance of the model.

Remarkably, our model replicates Mayor and Plunkett (2010)'s performance with realistic visual and acoustic stimuli, albeit requiring *very-few* joint presentations of image and spoken word pairs.

It is worth mentioning here that we do not try to maintain that CNNs or LSTMs are cognitively plausible models of how realistic stimuli are processed in biological brains – more details about this point in the "New Model" Section. We are just validating the hypothesis that the (Mayor & Plunkett, 2010) model could generalize to more complex stimuli and that the whole-object constraint can be helpful in this model.

Mayor and Plunkett (2010)'s model

Mayor and Plunkett (2010) neurocomputational model of taxonomic constraint (Figure 1) is based on (i) a visual self-organizing map (SOM) that processes visual inputs, (ii) an acoustic SOM that processes acoustic inputs, (iii) Hebbian connections between the two maps. Both self-organizing maps and Hebbian learning are considered cognitively plausible mechanisms (Hebb, 1949; Miikkulainen et al., 2005)

Firstly, the two maps are independently trained (using the standard learning algorithm for self-organizing maps, see Kohonen, 2001) to categorize the visual and the acoustic stimuli. This first learning phase is preliminary to word learning, and unsupervised, proper word learning starting to occur once infants have already started to learn to organize visual and acoustic information in isolation.

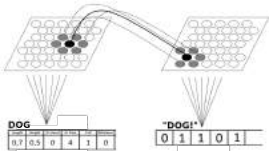


Figure 1: Mayor and Plunkett (2010)'s model

In this way, the two maps learn to represent the stimuli of their training set in a topologically significant way: close units respond (activate) similarly to similar stimuli. The *neural activation* a_j of a neuron j in response to a stimulus x is defined as: $a_j = e^{-\frac{q_j}{\tau}}$, where q_j is the *quantization error* (i.e., the distance between the input vector x and j's weight vector: $q_j = \|\mathbf{x} - \mathbf{w}_j\|$)), and τ is a parameter that modulates the neural activation. The neuron having the strongest activation is the stimulus' Best Matching Unit (BMU).

Once this first phase of learning is complete, the actual word learning can start. This is the Hebbian Learning phase, in which visual and acoustic stimulus are presented to their respective maps and the synapses between the two maps are strengthened. In particular, for each neuron v on the visual map and neuron p on the acoustic map, the Hebbian connection $u_{v,p}$ is strengthened proportionally to the resulting neural activations a_v and a_p, as follows:

$$u'_{v,p} = u_{v,p} + 1 - e^{-\lambda a_v a_p}$$

where λ is the Hebbian training learning rate, and $u'_{v,p}$ is the Hebbian connection after the update.

A single Hebbian learning event, combined with the previously acquired categorization capabilities of the visual and acoustic SOMs, allows the model to generalize the association to other stimuli belonging to the same category.

Comprehension is assessed by considering what visual category is retrieved when a word is presented to the auditory map and its activation is propagated via Hebbian connections. *Production* is assessed by considering what word is produced by the auditory map when a visual stimulus is presented to the visual map and the resulting activation is propagated through Hebbian connections.

The ability of the model to extend the learned word-object associations to other words and objects belonging to the same category is measured by the *Taxonomic Factor*, which is the percentage of correct word-object associations generated by the model (i.e., the average of the Production and Comprehension statistics). Results show that when the SOMs are adequately trained the Taxonomic Factor reaches 80% after a single joint word-object presentation.

New Model

We have enriched the original Mayor & Plunkett model (2010) (and Fenoglio et al. (2017)) so that (i) it can deal with realistic visual *and acoustic stimuli*, and (ii) it captures *the whole object constraint*. A graphical representation of the overall model is contained in Figure 2. For the visual and the acoustic stimuli we propose to use Deep Neural Networks to act as powerful feature extractors: we use two deep convolutional neural networks (Mask R-CNN and Inception V3) to process visual information and a deep recurrent neural network (Deep Speech) to process acoustic information. These models have been widely adopted for this purpose by the Machine Learning community, as they are able to output highly discriminative features (Razavian, Azizpour, Sullivan, & Carlsson, 2014; Graves, Mohamed, & Hinton, 2013; Hannun et al., 2014). Even if it is not the main focus of this paper, it is worth mentioning that these models have also been proposed as realistic models of visual and acoustic processing. Several studies establish a parallel between the representations of the visual input created by the different levels of CNNs and the way in which visual stimuli are processed by the visual cortex (Serre, 2016; Kriegeskorte, 2015; Khaligh-Razavi & Kriegeskorte, 2014). Furthermore, comparisons have been drawn between Recurrent Networks units (specifically the LSTM cell (Hochreiter & Schmidhuber, 1997)) and biologically plausible models of working memory such as the

PBWM model (prefrontal cortex and basal ganglia) (O'Reilly & Frank, 2006).

Visual Component

The visual stimuli that we consider are images taken from the Common Objects in COntext (COCO) dataset (Lin et al., 2014). In this dataset images are labelled pixel-wise, meaning that it is possible to extract the foreground objects from the background scene (i.e. performing image *segmentation*). As a first component of the visual module, we included a Mask R-CNN segmentation model (He, Gkioxari, Dollár, & Girshick, 2017), which separates foreground objects from the background content. Then the foreground object is cut from the background, the background erased and the new image so obtained is fed into an InceptionV3 Deep Convolutional Network (Szegedy, Vanhoucke, Ioffe, Shlens, & Wojna, 2016), an architecture which displays human-like performance on Object Recognition tasks. The deep network processes the object and builds a representation for it. We extract the representations (i.e. the neural activations) that are found in the deepest layer before the fully connected neural classifier (as they contain the most abstract features which have the best chance to depict the abstract concept the object instance refers to), and feed these representations to the visual self-organizing map.

To summarise, we employ a stack of two Deep Neural Networks in our visual module: the first one segmenting the object from the context; the second one analysing that output by means of a standard convolutional deep network. This architecture allows the model to overcome one limitation of standard deep convolutional models that, differently from humans (Spelke, 1990), do not use the notion of object when processing an image, and, on the contrary, rely very much on background information in object recognition tasks (Zhu et al., 2017).

Acoustic Component

We process spoken words using a Deep Speech Recurrent Network (Hannun et al., 2014) which is close to the state of the art in the Speech Recognition (i.e. parsing speech into text) task. This network is able to extract highly discriminative representations from our input stimuli, which have been generated using a realistic voice synthesizer that can be set up to use both male and female voices as well as different regional English accents. In our experiments, we had the generator pronounce labels from the COCO dataset. Similarly to the visual module, we extract features by concatenating the hidden state of the recurrent units after each time step. The resulting vector representations are then truncated to the same length and reduced in dimensionality by means of principal component analysis.

Overall Model

The upper component of the model, comprising the visual and acoustic self-organizing maps and their Hebbian connections, is trained as in the original (Mayor & Plunkett, 2010) model:

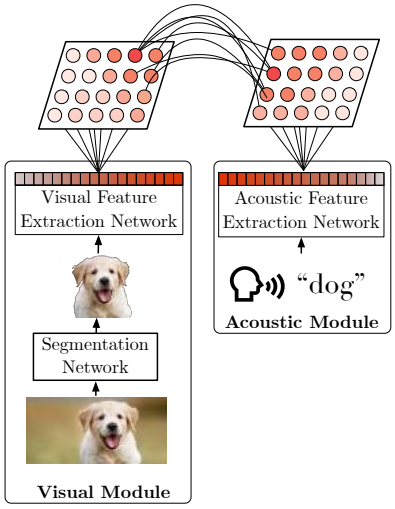


Figure 2: Our model

at first the visual and acoustic self-organizing maps are separately trained to organize their stimuli, then Hebbian learning starts.

Similarly to the original model, in order to assess the quality of the word/object association (the Comprehension and Production ability), we proceed as follows. For Production, we present to the visual map a visual stimulus, individuate a BMU, propagate its activity through Hebbian connections, and then evaluate the induced activation on the acoustic map with a cascading mechanism where neurons are interrogated in order of activation intensity. The first neuron corresponding to a single spoken word (i.e. a neuron that is the BMU for a single acoustic stimulus category) indicates which word is produced. We say that a Production task is successful if the category of the word matches the category of the visual stimulus. We proceed in a similar way for Comprehension.

Experiments

In our experimental phase, we set out to answer the following two questions:

1. Does our extension to the original word learning model by (Mayor & Plunkett, 2010) still account for the taxonomic constraint? In other words, is it possible to use realistic auditory and visual stimuli and achieve good word learning performance?

2. Is the whole object constraint beneficial to the word learning process?

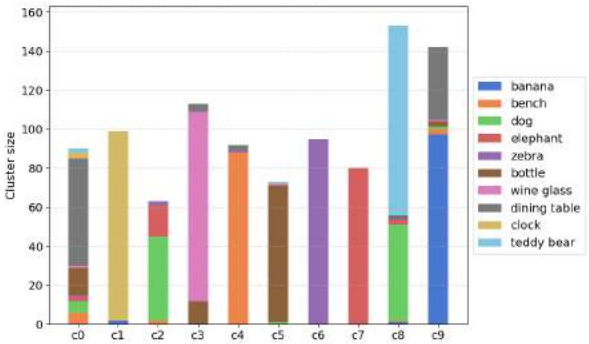
In order to extract whole object and non whole object representations (the inputs of the visual SOM), we trained two separate InceptionV3 networks for the same amount of time (i.e. epochs, full passes of the COCO dataset). However, one network was trained on images where the main object was cut out using the Mask R-CNN model, while the other one em-

ployed images that include a portion of the full visual scene[2]. We refer to these models respectively as the "whole object" and "non whole object" networks. Therefore, we explored the impact of using one network or the other in our visual module as a way to quantify the impact of including the whole object constraint in the overall model. An early evidence of the importance of the whole object constraint is provided by the performances (in terms of Object Recognition accuracy) of the two networks: the whole object network reaches a higher accuracy (93%) than the non-whole object network (77%) after the same amount of training time and similar learning rate schedules[3]. For the experiments that follow, however, we decided to only use as visual stimuli those images that have been correctly classified by both networks; thus, both convolutional models have perfect accuracy on the final visual dataset and can be compared on a fair ground. As far as the acoustic stimuli are concerned, we used a voice synthesizer to generate realistic voice recordings of both male and female voices pronouncing the object categories which appear in the visual dataset. To augment the size of the auditory dataset, we also varied the synthesizer's pronunciation speed. The representations were then extracted using a pre-trained Deep Speech network[4]. We truncated the representations to a length of 25 and kept the 20 most informative factors of variation using PCA. In the following sections, we report representation quality, SOM quality and taxonomic factor measurements for a dataset composed by 1000 visual stimuli and 390 acoustic stimuli belonging to 10 different word-object categories.

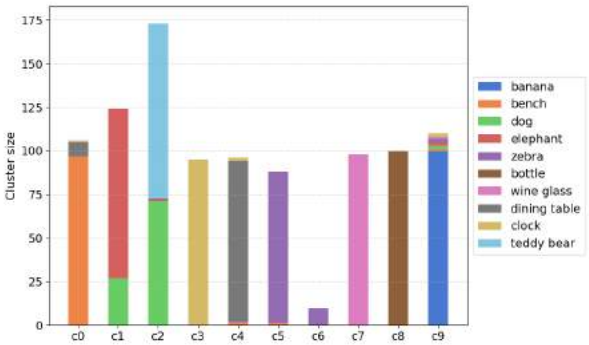
Representation Quality

First off, we set out to understand whether the representations extracted from the realistic stimuli are well-behaved. To this end, we performed an experiment in which the representations are used as input for the k-Means clustering algorithm with k, the number of clusters, set to 10. After fitting the clustering model, we visualize the resulting clusters (see Figure 3) using a histogram plot.

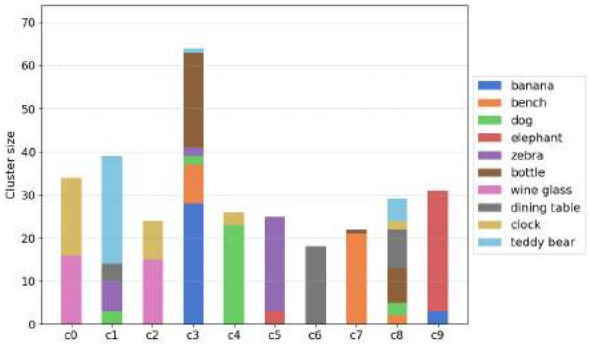
We also assess the trained SOMs' topological organization by visualizing them. In Figure 4, we see that representations belonging to the same category are mapped on neurons that are topologically close. Moreover, we evaluate the organization quality by using the *class compactness* measure; this is computed by averaging the Euclidean distances between neurons that are BMUs for stimuli belonging to the same class and dividing by the average distance between BMUs for any stimulus. Lower values indicate better topological structure.



(a) Non whole object representations.



(b) Whole object representations.



(c) Audio representations.

Figure 3: K-means clusters. Colors show how representations of each category contribute to the clusters.

[2]More specifically, we used the bounding box information for each object in COCO and expanded it by 40% so to preserve a significant amount of visual context.

[3]We trained the whole object network for 60 epochs. We used a learning rate of 10^{-3}, decreasing it to 10^{-4} after 40 epochs. This schedule, however, appeared to be very sub-optimal when training the non whole learning network, as the object recognition accuracy progressed very slowly. Therefore, the second network was trained with a learning rate of 10^{-2} and decreased it to 10^{-3} after 40 epochs.

[4]https://github.com/mozilla/DeepSpeech/

Averaging this formula over the categories in the dataset results in the overall *SOM compactness* value. We report (Table 1) lower compactness values for the SOM we trained on the whole object representations, and robust compactness for the non whole object and acoustic SOMs.

Word Learning

As an experimental evaluation of the overall model, we compare the word learning capabilities of our model with and without the whole object constraint. After training with a number of joint word-object presentations, the model has to be able to produce an appropriate acoustic stimulus when presented with an image (*understanding*) and viceversa (*comprehension*). The algorithm used to obtain the final word-object association is described in the "New Model" Section.

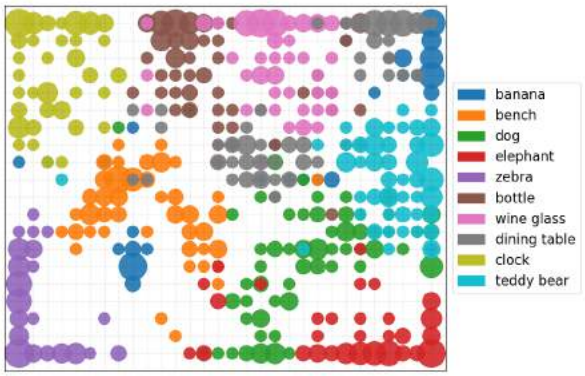
In Figure 5 we report the understanding and comprehension performances alongside the Taxonomic Factor (their average). A set of stimuli (20% of all the visual and acoustic representations) was reserved for testing and was excluded from the training sets.

Discussion

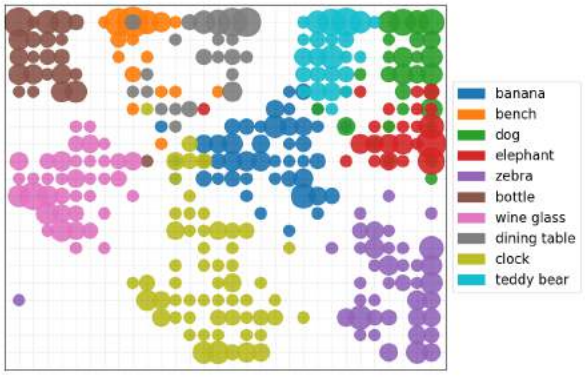
Coming back to the questions we set out to answer at the start of the section, Figure 5 makes apparent that our model manages to perform word learning with appropriate performance (set at a Taxonomic Factor of over 80% in (Mayor & Plunkett, 2010)) after very few word-object presentations. The performance is also in line with previous work on this model (Fenoglio et al., 2017), in which, however, very simplified acoustic stimuli were considered. Therefore, this computational model can still account for the taxonomic constraint even in the face of realistic visual and acoustic stimuli. As for the contribution of the whole object constraint to the model, we first observe that the comparison of the clusters is favorable to the whole object model (Figure 3); furthermore, the self-organizing maps that were trained using the aforementioned representations display good topological organization (Figure 4) and solid compactness values. In addition, as implied by Table 1, the SOM trained with the whole object representations displays stronger topological organization. As for the word learning performance, we obtain a significantly higher Taxonomic Factor when using the whole object representations and conclude that including the whole object constraint in this model is highly beneficial.

Related Work

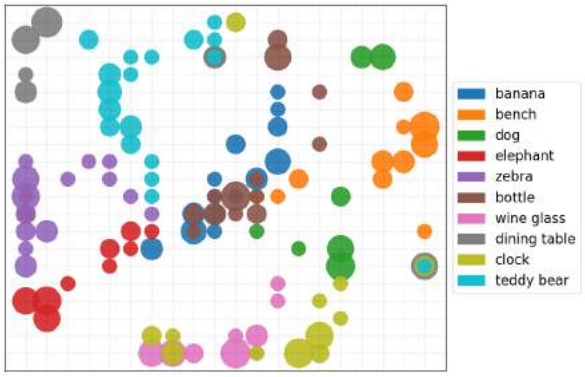
Our model bears some family resemblance to recent models of image-speech association learning (Synnaeve, Versteegh, & Dupoux, 2014; Harwath & Glass, 2015; Harwath, Torralba, & Glass, 2016; Chrupala, Gelderloos, & Alishahi, 2017), which, at least in part, have been proposed as cognitive models of spoken words referent acquisition. Similarly to Synnaeve et al. (2014), here we consider associations between images and single spoken words, whereas Harwath and Glass (2015); Harwath et al. (2016); Chrupala et al. (2017) consider associations between images and



(a) SOM over non whole object representations.



(b) SOM over whole object representations.



(c) SOM over audio representations.

Figure 4: SOM representations. Colors represent different categories, larger circles are for neurons that activate more often.

Table 1: Compactness values for the three SOMs.

Visual Whole Object	Visual Non Whole Object	Acoustic
0.228	0.372	0.429

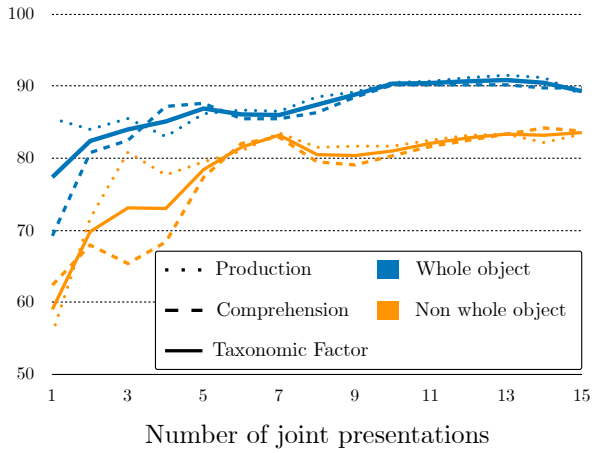


Figure 5: *Taxonomic factor* of the model, using an increasing number of pairs of stimuli per class during the training of the Hebbian Connections (on the x-axis).

more complex acoustic stimuli, as whole spoken sentences (Harwath & Glass, 2015; Harwath et al., 2016; Chrupala et al., 2017). With respect to all these models, the specificity of our model is that it learns to generalize image-speech associations to whole visual categories and all phonetic variants of a corresponding word, out of few *positive* joint image-speech presentations, without any need of explicit counterexamples. This parallels the training schedule by which humans usually learn to associate words (or sentences) to visual stimuli.

Vinyals, Blundell, Lillicrap, Kavukcuoglu, and Wierstra (2016) address the problem of *One Shot Learning*: how to build models that reproduce the crucial ability of humans, infants and adults, of learning out of *few examples*, as opposed to the massive training currently used for many neural network models? The proposed model is trained to integrate in one-shot new observations into pre-existent knowledge, represented by a support set. Similarly to our work, representations extracted from pre-trained neural networks are also employed. The authors test their model on classification tasks in which the training dataset is composed by 1 or 5 examples for each category; while a direct comparison would not be proper, as the experimental setups and datasets are fundamentally different, it is worth mentioning that word learning in the present approach does not rely on the supervised, gradient-based optimization of a training objective (i.e. a loss function). On the contrary, in our model word learning emerges after the unsupervised training of the SOMs and a few joint, positive presentations of word-object pairs.

Conclusions

In this paper we expand on the the model originally introduced by (Mayor & Plunkett, 2010) and extended by (Fenoglio et al., 2017). Our work focused on two objectives: allowing the model to process realistic acoustic stimuli, and injecting the whole object constraint into it. We also introduce experiments allowing one to assess the effects of these two changes to the model.

In summary, the empirical evidence shows that the realistic stimuli are not hindering the ability of the model to learn the association between objects and word. In fact, even though the greater complexity of the stimuli representation makes the task harder, the system only requires a few joint presentations to reach the 80% taxonomic accuracy performance shown in the original work by Mayor and Plunkett (2010).

For what concerns the whole object constraint, the evidence demonstrates the remarkable impact of this constraint on the performances of the system. In practice the whole object constraint allows for better performances with respect to the model by Fenoglio et al. (2017) even considering that the latter is dealing with simpler acoustic stimuli. It is worth debating whether the whole-object representations extracted by the visual module contain all the parts of the original objects. Indeed, given the discriminative nature of the CNN training process, the representations may only contain few, very specialized features which suffice for the classification task. As a future work, one may investigate this problem by designing experiments in which one studies whether the activation of the visual SOM, elicited by an acoustic stimulus (a word), allows one to reconstruct a prototypical version of the object referenced by the word. Furthermore, we intend to investigate how to cope with the uttering of whole sentences instead of single words.

References

Chrupala, G., Gelderloos, L., & Alishahi, A. (2017). Representations of language in a model of visually grounded speech signal. *ACL 2017*.

Fenoglio, G., Esposito, R., & Gliozzi, V. (2017). A neural network model for taxonomic responding with realistic visual inputs. *COGSCI 2017*, 1–6.

Graves, A., Mohamed, A.-r., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. *ICASSP 2013*, 6645–6649.

Hannun, A. Y., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., et al. (2014). Deep speech: Scaling up end-to-end speech recognition. *CoRR*, *abs/1412.5567*.

Harwath, D., & Glass, J. R. (2015). Deep multimodal semantic embeddings for speech and images. *ASRU 2015*.

Harwath, D., Torralba, A., & Glass, J. R. (2016). Unsupervised learning of spoken language with visual context. *NIPS 2016*.

He, K., Gkioxari, G., Dollár, P., & Girshick, R. B. (2017). Mask R-CNN. *CoRR*, *abs/1703.06870*.

Hebb, D. (1949). *The organization of behavior: A neuropsychological theory*. Weley & Sons.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014, 6). Deep Supervised, but Not Unsupervised, Models May Explain

IT Cortical Representation. *PLOS Computational Biology*, *10*(11).

Kohonen, T. (2001). *Self-organizing maps*. Springer Berlin.

Kriegeskorte, N. (2015). Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, *1*, 417–446.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., et al. (2014). Microsoft coco: Common objects in context. *ECCV 2014*, 740–755.

Markman, E. M. (1989). *Categorization and naming in children: Problems of induction*. MIT Press.

Mayor, J., & Plunkett, K. (2010). A neurocomputational account of taxonomic responding and fast mapping in early word learning. *Psychological review*, *117 1*, 1–31.

McMurray, B., Horst, J., & Samuelson, L. (2012). Word learning emerges from the interaction of online referent selection and slow associative learning. *Psychological Review*, *119*, 831–877.

Miikkulainen, R., Bednar, J., Choe, Y., & Sirosh, J. (2005). *Computational maps in the visual cortex*. Springer.

O'Reilly, R. C., & Frank, M. J. (2006). Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. *Neural computation*, *18*(2), 283–328.

Quine, W. V. O. (1960). Word and object. *MIT Press*.

Razavian, A. S., Azizpour, H., Sullivan, J., & Carlsson, S. (2014). Cnn features off-the-shelf: an astounding baseline for recognition. *CVPRW 2014*, 806–813.

Serre, T. (2016). Models of visual categorization. *Cognitive Science*, *7*(3), 197–213.

Spelke, E. (1990). Principles of object perception. *Cognitive Science*, *14*(1), 29-56.

Synnaeve, G., Versteegh, M., & Dupoux, E. (2014). Learning words from images and speech. *NIPS Workshop on Learning Semantics 2014*.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. *CVPRW 2016*, 2818–2826.

Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K., & Wierstra, D. (2016). Matching networks for one shot learning. *NIPS 2016*, 3630–3638.

Yurovsky, D., Fricker, D., Yu, C., & Smith, L. B. (2014). The role of partial knowledge in statistical word learning. *Psychonomic Bullettin Review*, *21(1)*, 1–22.

Zhu, Z., Xie, L., & Yuille, A. L. (2017). Object recognition with and without objects. *IJCAI 2017*, 3609–3615.

Simulating Bilingual Word Learning: Monolingual and Bilingual Adults' Use of Cross-Situational Statistics

Kin Chung Jacky Chan (k.c.chan@lancaster.ac.uk)
Lancaster University, Department of Psychology, Lancaster, LA1 4YF, UK

Padraic Monaghan (p.j.monaghan@uva.nl)
University of Amsterdam, Faculty of Humanities, Amsterdam, 1000 BP, NL
Lancaster University, Department of Psychology, Lancaster, LA1 4YF, UK

Abstract

Children learning language in multilingual settings have to learn that objects take different labels within each different language to which they are exposed. Previous research has shown that adults can learn one-to-one and two-to-one word-object mappings via cross-situational statistical learning (CSSL), and that socio-pragmatic cues may differentially influence monolingual and bilingual adults' learning of such mappings. However, the extent to which monolingual and bilingual learners can keep track of multiple labels from multiple speakers has not yet been investigated. We manipulated the number of speakers in a CSSL task that involved learning both mapping types. We successfully replicated previous studies that found that both monolinguals and bilinguals could learn both types of mappings via CSSL. In addition, we found that bilinguals showed a steeper learning rate for two-to-one mappings than monolinguals, and bilinguals were more likely to accept two words for the same object than monolinguals. These results show that the effect of speaker identity on tracking word-object mappings varies according to language experience.

Keywords: statistical learning; bilingualism; mutual exclusivity; cross-situational learning; word learning

The learning of the mapping between a word and its referent is difficult, as there are infinitely many potential referents for a word. This uncertainty is referred to as the "Gavagai" problem (Quine, 1960). The uncertainty is increased still further when children grow up in multilingual environments, as this means there are multiple words for a particular referent. The present study aims to investigate how speaker identity, as a socio-pragmatic cue, impacts on language learning under such conditions of referential and reference uncertainty.

A prominent suggestion as to how language learners overcome the "Gavagai" problem has been that language learners make use of constraints on which mappings can be formed. For instance, the ME constraint suggests that language learners tend to assign only one word to a referent (Markman & Wachtel, 1988). When language learners hear a novel word and see a familiar object, of which they already know the name, and an unfamiliar object, they would, based on ME, pair the novel word with the unfamiliar object. Other constraints include the whole-object assumption and the taxonomic assumption (Markman, 1991; Markman & Hutchinson, 1984; Markman & Wachtel, 1988). Another account of word learning is the socio-pragmatic account, which suggests that language learners' word learning rely on their socio-cognitive skills and the social cues available in communicative contexts (Tomasello, 2000). This account explains word learning in terms of language learners' ability to actively monitor others' attention (Akhtar & Tomasello, 1996) and intention (Tomasello & Barton, 1994) to discover intended referents of novel words. In general, both of these accounts posit that language learners make use of certain strategies to limit the number of potential referents for a word to help solve the "Gavagai" problem. Yet, constraining the problem space is not the only way to solve the word-learning problem.

Recently, cross-situational statistical learning (CSSL) ability has been proposed as a valuable contributor to word learning. Though the referent of a novel word might be ambiguous within the context of a single learning instance, across multiple learning instances, learners would be able to track the co-occurrences of the novel word and its referent, with which it reliably appears. This statistical information can then help learners to disambiguate which words refer to which referents. Yu and Smith (2007) presented adults with a series of trials containing two to four unfamiliar objects and novel words. Within each trial, the word-object pairings were ambiguous (i.e., novel words were presented in a random order in all trials and there was no correspondence between the order of words and the location of objects on the computer screen), but across trials, with the presentation of different combinations of novel words and their referring objects, the word-object pairings could become apparent. Yu and Smith found that adults could learn the meanings of words via CSSL. This finding has been replicated in various similar studies (e.g., Fitneva & Christiansen, 2011; Hamrick & Rebuschat, 2012; Monaghan & Mattock, 2012; Vouloumanos, 2008).

In these studies, only one-to-one word-object pairs were used. Yet, although learners favour ME (i.e., one-to-one word-referent mappings) when learning the meaning of words, overcoming ME is important for learning categories, homonyms, and synonyms (e.g., Markman & Wachtel, 1988). It is also particularly important for bilinguals as they have to learn translation equivalents (forming many-to-one word-referent mappings; e.g., both "apple" and "manzana" refer to a particular fruit) and interlingual homographs (forming one-to-many word-referent mappings; e.g., "tuna" refers to a kind of fish in English but prickly pear in Spanish).

Ichinco, Frank and Saxe (2009) familiarised and then tested adults on a set of one-to-one word-object pairs. Then, the participants were familiarised to a second set of one-to-one word-object pairs. Some of the pairs in the second set required the remapping of objects or words. Thus, although each set consisted of one-to-one word-object pairs, across the two sets, there was a combination of one-to-one, two-to-one, and one-to-two word-object pairs. The two-to-one and one-to-two word-object pairs were critical for testing whether adults could relax ME during a CSSL task. It was found that the participants were successful in learning the one-to-one word-object pairs and the first mapping of the two-to-one and one-to-two word-object pairs. By contrast, they failed to learn the second mapping of the two-to-one and one-to-two word-object pairs. Ichinco et al. took the results of their study as evidence against a simple associative learning account of word learning.

Yet, Kachergis, Yu and Shiffrin (2009) argued that the results of Ichinco et al.'s (2009) study could be due to a blocking effect, giving rise to the participants favouring the first mapping learnt. Using a similar paradigm to that in Ichinco et al.'s study, Kachergis et al. manipulated the number of occurrences of the second mapping of the word-object pairs. It was found that the extent to which the participants relaxed ME – successful at learning the second mapping of the word-object pairs – was associated with the number of times they had been exposed to the pairs, such that the participants were more likely to relax ME when there was more evidence (i.e., exposure) in the input for the second mapping.

These CSSL studies examined CSSL in a monolingual population. Only a few studies have looked at CSSL in a bilingual population. A study similar to that of Yu and Smith's (2007) by Escudero, Mulak, Fu and Singh (2016) showed that bilingual adults could learn one-to-one word-object pairs via CSSL, outperforming their monolingual counterparts. Another study by Poepsel and Weiss (2016) investigated whether bilingual adults would learn one-to-two word-object pairs better than monolingual adults do, owing to them encountering more instances where they have to relax ME in order to learn new words. They tested the participants' learning of the first and second word-object mappings of the one-to-two word-object pairs in separate testing blocks after the first and second block of learning trials respectively, and tested all word-object mappings in the final testing block after the third learning block. Consistent with Poepsel and Weiss' prediction, it was found that the bilingual adults were quicker than the monolingual adults at learning and showed higher proficiency in learning the one-to-two word-object pairs.

Further, Benitez, Yurovsky and Smith (2016) familiarised monolingual and bilingual adults with a set of one-to-one and two-to-one word-object pairs and tested their learning of the word-object mappings. They found that the monolingual and bilingual adults performed similarly on the task. Both groups showed learning of both the one-to-one and two-to-one word-object pairs, but both groups were better at learning the one-to-one pairs. This is surprising, but not unreasonable, as monolinguals, who have to learn synonyms, are also experienced in learning two-to-one word-object mappings. An interesting finding of their study was that when a phonological cue distinguished sets of labels, the bilingual adults were more likely to learn both words of the two-to-one pairs. This suggests that bilingual adults are more sensitive to the linguistic cues that hint at different languages present in the linguistic input. Taken together, there is evidence that bilingual adults are better than their monolingual counterparts when it comes to learning word-object pairs that violate ME via CSSL.

Other studies have investigated whether socio-pragmatic cues in the linguistic input would affect learners' cross-situational word learning (e.g., Metzing & Brennan, 2003; Trude & Brown-Schmidt, 2012). Poepsel and Weiss (2014) manipulated the socio-pragmatic information available to participants. In one condition, the participants were told that there were two languages involved in the task. In the other two conditions, the participants were not told anything explicitly, but in one of these conditions, the participants were provided with information on speaker identity – they heard a male and a female voice. In the two-voice condition, the two speakers used the same word to refer to a different object, which could be seen as an implicit cue that there could be two different linguistic structures involved in the task. It was found that the manipulation of socio-pragmatic information did not affect the monolingual adults' performance on learning one-to-two word-object pairs. Yet, in multilingual environments it is more usual for one object to be labelled differently by distinct speakers. Whether varying speaker identity would affect bilingual adults' cross-situational word learning, and whether speaker identity can influence learning of two-to-one mappings is as yet unknown.

The aim of the present study was to examine whether speaker identity would differentially affect monolingual and bilingual adults' performance on a CSSL task that involved the learning of one-to-one and two-to-one word-object pairs. We included two conditions – one where there was a single speaker labelling objects in two ways, and one where different speakers labeleld objects in two ways. The present study employed a CSSL paradigm similar to that in Monaghan and Mattock's (2012) study, which is slightly different from many of the CSSL paradigms used in other studies. The crucial difference was that the CSSL paradigm used in the present study did not distinguish between familiarisation and test trials – participants were required to make a forced choice response, without feedback, between two objects in all trials. This allowed an online measure of how quickly and reliably participants form one-to-one and two-to-one word-object mappings across trials.

Another unique feature of the present study was that an additional ME block was administered at the end of the CSSL paradigm to determine whether successful learning of two-to-one word-object pairs was due to successful tracking of two structures in the linguistic input or a general tendency to relax ME.

It was predicted that bilingual adults would be quicker and more accurate at learning two-to-one word-object pairs than monolingual adults. Also, it was predicted that the presence of speaker identity would further benefit bilingual adults' learning of two-to-one word-object pairs due to them being more experienced than monolingual adults in using socio-pragmatic information to track multiple structures in their linguistic input.

Method

Participants

Forty monolingual (M_{age} = 22.80, SD = 4.56, 4 male) and forty bilingual (M_{age} = 23.58, SD = 3.71, 10 male) participants were recruited through SONA (the departmental online recruitment system) and advertisements on social networking websites. Half of the participants in each language group were randomly assigned to the one-speaker condition, and the other half the two-speaker condition. Nine additional participants were tested but excluded due to technical difficulties (n = 8) and experimenter error (n = 1).

Participants rated their language proficiency on a 10-point Likert scale from 1 (limited knowledge) to 10 (highly proficient). Monolinguals rated their English proficiency at an average of 9.95 (SD = 0.22). Ten monolingual participants indicated exposure to additional languages, but were considered functionally monolingual, as all such proficiency ratings were below 4 (M = 2.23, SD = 0.93), a similar cut-off to that used in Poepsel and Weiss (2016). The bilingual group rated the proficiency of their first language at an average of 9.85 (SD = 0.43) and that of their additional languages at an average of 7.36 (SD = 2.01).

Materials and apparatus

Fourteen images of unfamiliar objects and 20 novel words were selected from the Novel Object and Unusual Name (NOUN) Database (Horst & Hout, 2016). Sound files of the novel words were generated using the system voices Kate (female voice) and Daniel (male voice) on Macintosh computers. Pictures were randomly paired with the novel words for each participant, such that there were eight one-to-one word-object pairs and six two-to-one word-object pairs. In the one-speaker condition, all words were uttered by the same speaker. The gender of the speaker was counterbalanced across participants assigned to the one-speaker condition. In the two-speaker condition, half of the words were uttered by a male, and the other half a female. For words in the two-to-one word-object pairs, the two words referring to the same object were uttered by voices of different gender. The gender of speaker of each word was counterbalanced across participants assigned to the two-speaker condition. In addition, eight images of familiar objects were selected from the TarrLab Object Databank (1996) for use in the familiarisation trials. Sound files of the familiar words were generated using the system voice Allison (female voice) on Macintosh computers. Note that this was a different voice from those used in the test trials, so that the participants did not have any reliable information on what language(s) the speakers in the test trials spoke. The pictures and audio files of words were presented on a Macintosh computer using PsychoPy (Peirce, 2009).

Procedure

The experiment took place in a quiet room. Participants were tested in groups of less than five people. After receiving an information sheet and signing informed consent, each participant was asked to complete the experiment on a Macintosh computer. Participants were asked to put on headphones for the experiment.

For each trial, the participants saw two pictures presented on the screen. After 500 ms, they heard a word. The target and foil were randomised for screen position (left vs. right) across trials. The participants were instructed to press the right arrow key if they thought the word presented refers to the object on the right and press the left arrow key if they thought the word presented refers to the object on the left. The participants were also instructed to make a guess if they did not know the answer to any of the test trials.

The participants first took part in a familiarisation block, in which they were presented with four trials containing known words and objects. This was to familiarise the participants with the experimental procedure. For the main experiment, the participants first took part in eight CSSL blocks of 40 test trials each. Within each of the CSSL blocks, each object occurred four times as the target and four times as the foil. The screen position of the target and foil were pseudo-randomised, such that the target appeared an equal number of times as the left and as the right object. Words in the one-to-one word-object pairs occurred four times within a block, whereas those in the two-to-one word-object pairs occurred only two times within a block. The order of trials within each block was pseudo-randomised, such that none of the objects appeared in two consecutive trials. An important point to note is that the participants were not provided with any information on the number of languages involved in the main experiment – the only socio-pragmatic cue available to them was the number of speakers in the task. The participants were allowed to take a short break after every two blocks. After all eight blocks, the participants were exposed to each one-to-one word-object pair 32 times and each two-to-one word-object pairs 16 times.

Immediately after the eighth CSSL block, the participants took part in an ME block containing eight test trials. Each trial featured one of the objects from the one-to-one pairs from the CSSL blocks and a new unfamiliar object. Each object occurred one time as the target and one time as the foil. As in the CSSL blocks, the screen positions of the target and foil were pseudo-randomised. For each of the first four trials, the participants heard a word that they had just had the opportunity to learn during the CSSL blocks. These four trials served the purpose of familiarising the participants with the new unfamiliar objects and to control for a possible novelty bias during later trials, where the new unfamiliar objects were the target. For each of the final four trials, the participants

heard a new novel word, which was spoken by the speaker who spoke the word for the foil in the same trial. These final trials were critical for determining the extent to which the participants relied on ME when learning new words. If a participant was relying on ME, they would be more likely to choose the familiar object in the first four trials and the less familiar objects in the last four trials. However, if a participant was relaxing ME, their performance would be at chance level – choosing either object as the answer in any given trial.

Upon completing the ME block, all participants were given a full debrief and received £3.50 for taking part in the experiment. Each testing session lasted less than 30 minutes.

Results

Learning over the training blocks

Data from six participants, one from the monolingual group and five from the bilingual group were excluded from analysis, due to them not demonstrating learning across testing blocks (i.e., average proportion correct across first two blocks > average proportion correct across final two blocks).

To compare whether number of speakers had influenced the monolingual and bilingual adults' learning of the two types of mappings, generalised linear mixed-effects (GLM) modelling was used to predict the adults' response accuracy. The data for GLM modelling consisted of the response accuracy from each participant on each trial, giving a total of 23680 observations.

A series of GLM models were fitted using the glmer function (family = binomial) in the lme 4 package in R. A backwards elimination approach was used, entering as fixed factors: language group, speaker number, block, mapping type of the target (whether it had one or two labels), and mapping type of the foil. Extraneous variables, including participant gender and speaker gender, did not influence the participants' performance. For training accuracy, the best model (*AIC* = 19977.7, *BIC* = 20131.1, *logLik* = -9969.9, deviance = 19939.7) given the data is the model with the following fixed effects: the three-way interaction, all two-way interactions, and main effects of block, language group, and target mapping and the main effect of foil mapping; the following random intercepts: subject, word, target, and foil; and the following random slopes: block on subject and language group on word and target.

As expected, there was a significant effect of block (β = 0.26655, 95% CI [0.2204, 0.3127]), suggesting that, in general, performance improved across testing blocks. The main effect of target mapping was also significant (β = 0.74309, 95% CI [0.5191, 0.9670]). To our surprise, and contrary to Benitez et al. (2016), the participants were *better* at learning the two-to-one than one-to-one mappings. There was also a significant main effect of foil mapping (β = 0.26248, 95% CI [0.1826, 0.3424]), suggesting that performance was better if the foil in a given trial was a two-to-one mapping. In addition, the interaction between block and target mapping was also significant (β = 0.06914, 95% CI [0.0194, 0.1189]), showing a convergence of the participants' performance in learning the two mapping types across blocks, such that although their learning of the two-to-one mappings was better than that of the one-to-one mappings across blocks, their learning rate for the one-to-one mappings was steeper.

Though there was no significant main effect of language group, the interaction between language group and target mapping was significant (β = -0.49190, 95% CI [-0.8068, -0.1770]), indicating that although both language groups were better at learning the two-to-one mappings, the monolingual group's performance difference between the two mapping types was greater than that of the bilingual group.

Finally, there was a significant three-way interaction between block, language group, and target mapping (β = 0.08267, 95% CI [0.0103, 0.1551]; see Figure 1). The three-way interaction suggests that, for the monolingual group, the learning rate of the two-to-one mappings was more gradual than that of the one-to-one mappings, whereas for the bilingual group, there was faster learning of the two-to-one mappings over the blocks.

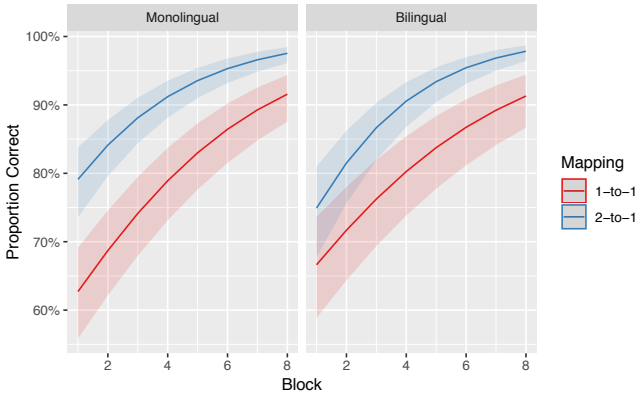


Figure 1: Three-way interaction of block, language group and target mapping.

Performance on the ME task

Though there were no significant main effect or interactions with number of speakers in the task, it was possible that monolingual and bilingual speakers relied on different strategies – either relaxing ME or successfully tracking two labels in the linguistic input would produce a similar pattern of results. In order to determine whether the two language groups relied on similar strategies, their performance in the ME block was analysed. In any given trial, if a participant picked the object that was in line with the application of ME, they scored 1, otherwise they scored 0. Similar to the treatment of the data from the CSSL blocks, GLM models were fitted to participants' scores on each trial (592 observations). Predictor variables of the GLM models were language group, speaker number, and word type (familiar vs. new), and a backwards elimination approach was used.

Extraneous variables, including participant gender and speaker gender, did not influence the participants'

performance. The best model (*AIC* = 180.6, *BIC* = 202.6, *logLik* = -85.3, deviance = 170.6) given the data is the model with the following fixed effects: the two-way interaction and the main effects of language group and speaker number; and the random intercept of subject.

Of particular note, the significant interaction between language group and speaker number (β = 3.3748, 95% CI [0.0054, 6.7442]; see Figure 2) suggests that although both language groups were able to systematically apply ME, there was a tendency for the bilingual group to relax ME when there was only one speaker in the task.

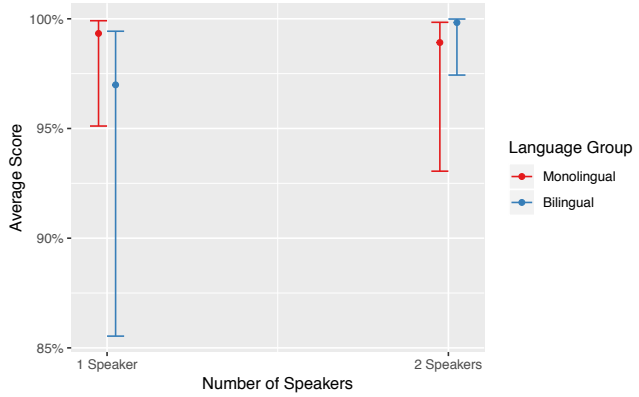


Figure 2: Interaction between speaker number and language group.

Discussion

Using a CSSL paradigm, the influence of speaker identity on monolingual and bilingual adults' learning of one-to-one and two-to-one word-object pairs was examined. In line with previous research (e.g., Benitez et al., 2016), our results showed that both monolingual and bilingual adults are capable of learning one-to-one and two-to-one word-object mappings through CSSL. Yet, inconsistent with Benitez et al.'s main finding, both groups of participants in the present study were better at learning the two-to-one than one-to-one word-object mappings. This could be due to the imbalanced number of objects for each type of mapping – there were six objects that mapped onto two words, whereas only four objects mapped onto one word. As the majority of the objects had two names, it was possible that the participants had formed an expectation that each object could take on two names and used this as a learning strategy for the task.

This explanation is given further weight by the finding that the participants were more likely to accurately map a label to the target object if the foil object was a referent of a two-to-one mapping. The indication here is that, in any given trial, if a participant had already learnt two labels for the foil, they would map the different word presented to the target, due to the foil already taking on the expected maximum number of words. Yet, in order to detect that some objects were named with one, and others with two, labels, participants had to gain this knowledge from tracking implicitly the association between particular words and objects over multiple scenes. That participants were adept at acquiring both one and two labels for objects so early in training demonstrates the power of this learning mechanism. Yet, it should also be noted in Benitez et al.'s study, there were instances where two-to-one mappings were better learnt than one-to-one mappings in that the presence of a second label seemed to have improved learning of the first label. Our task could be showing a similar advantage.

However, there were subtle differences in the learning trajectory of the monolingual and bilingual speakers in our study. The significant three-way interaction between block, language group, and target mapping shows that the learning of the two-to-one mappings was different for the two language groups. The performance of the monolingual group showed less improvement in learning of the two-to-one mappings, whereas the bilingual group had a steeper learning rate for the two-to-one compared to one-to-one mappings. This steeper learning rate could be due to their experience with language. In bilingual adults' linguistic environment, two-to-one word-object mappings would be more dominant than one-to-one word-object mappings. This experience could have benefited them in learning the two-to-one mappings in the task, which is in line with the finding of Kalashnikova, Mattock and Monaghan's (2015) study that bilingual experience would lead to more flexible use of ME, exhibited by higher tendency to accept lexical overlap.

Alternatively, the observed difference between the two language groups could be due to the monolingual adults displaying an early advantage in learning the two-to-one mappings from the first testing block, whereas the bilingual adults' learning of the two mapping types did not differ until the third testing block. A possible explanation to this initial difference of the learning of the two-to-one mappings could, again, relate to the imbalanced number of objects pertaining to each type of mapping. The imbalanced number may have served as a cue for the monolingual adults to more readily learn two words for one object, which could have been salient because this was inconsistent with their usual experience (i.e., one-to-one word-object mappings being the norm). For the bilingual adults, as they frequently confront two-to-one word-object mappings, the imbalance may be less salient and thus a less effective cue to influence their cross-situational learning early on in the experiment. These significant interactions suggest that language experience plays a role in the application of different word-learning strategies.

However, our results for the bilingual participants do not tally with those of Benitez et al.'s (2016) finding, which showed that bilingual adults' learning of two-to-one word-object mappings in a CSSL task was worse than their learning of one-to-one mappings. In Benitez et al.'s study, participants were presented with four objects and four words at a time during training, whereas the participants in the present study were only presented with two objects and one word at a time. The complexity of Benitez et al.'s task could have favoured the learning of one-to-one mappings. In their study, although the number of co-occurrences of each corresponding word-object pair was the same for both mapping types, the spurious

co-occurrences of unpaired word-object mappings was higher for the two-to-one mappings, making the learning of the two-to-one mappings more difficult than that of the one-to-one mappings. In the present study, although the participants were presented with fewer tokens of the two-to-one mappings, the reduced number of objects and words in a given trial were likely to have more closely mimicked actual word-learning experiences than Benitez et al.'s task, making the learning of both types of mapping relatively easy to the participants in the present study.

In addition, the design of the present study required participants to make a decision about a pairing on every trial, unlike in previous studies where participants went through a familiarization phase and then a test phase. This could have made the participants' learning of the word-object mappings more explicit and highlighted to the participants that the majority of the mappings were two-to-one, giving rise to the observed better learning of the two-to-one mappings. Determining the extent of referential ambiguity and the relative occurrence of two-to-one versus one-to-one mappings in the language learner's experience will enable us to determine more closely which experimental task better resembles natural language learning.

In contrast to our prediction, manipulating speaker identity did not influence CSSL of either language group. It was perhaps less surprising for the monolingual group, as Poepsel and Weiss (2014) found that manipulating speaker identity did not affect monolingual adults' learning of word-object mappings that violate ME. Taking into account Benitez et al.'s (2016) finding that linguistic cue could affect bilingual adults' learning of two-to-one word-object mappings and the non-significant effect of speaker identity in the present study, it is likely that information about the languages involved in the linguistic input per se is more important than speaker identity as a cue in influencing bilingual adults' word learning. In reality, information about languages in the input is a more reliable cue than speaker identity, as one speaker could speak multiple languages and different speakers could speak the same language.

Nevertheless, speaker identity did seem to have an effect on the strategy used by the two language groups. In the ME block, both language groups demonstrated majority use of ME. Yet, when there was only one speaker involved in naming objects, the bilingual group showed a greater tendency to relax ME (Kalashnikova et al., 2015). This suggests that although speaker identity did not have an effect on the observed responses of the participants in the CSSL task, it may have altered the strategies that they use. The bilingual speakers were more likely than the monolingual speakers to relax ME when more than one language structure was used by the same speaker. This may have been due to greater familiarity by bilingual speakers that individuals may speak more than one language.

In a broader sense, the results of the present study have demonstrated that language learners can flexibly use multiple word-learning strategies to learn different language structures in solving the "Gavagai" problem. In an environment with multiple language structures, learners have to quickly discriminate the different structures (Gebhart, Aslin & Newport, 2009). Previous studies (e.g., Qian, Jaeger & Aslin, 2012) have shown that socio-pragmatic cues, such as a voice change, can help learners focus on the syntactic structures available in the input. The lack of overall influence of speaker identity on the CSSL task in the present study should, therefore, not be taken as evidence that socio-pragmatic cues do not contribute to word learning, as it could instead be that word learning across multiple situations does not rely so heavily on this particular socio-pragmatic cue. Other socio-pragmatic cues, for example information on the languages that the speakers in the CSSL task speak or more information on the speakers' linguistic identities, might be more effective in influencing learners' reliance of word-learning strategies. Nevertheless, the results of the present study, in terms of trajectory of learning on the CSSL task and performance in the ME task, suggest that the extent to which a word learning strategy is relied upon depends in part on an individual learner's previous experience with languages and the learning context. These results also begin to give us some insights into how language experience, contextual cues and task design contribute to shaping learners' use of different word-learning strategies.

In summary, we replicated previous studies that found that language learners are adept at accepting multiple labels for the same object. Curiously, when only one word is heard, and two possible objects viewed, both monolingual and bilingual speakers were better at learning two labels for an object than one label for an object. The effects of participants' linguistic background exerted subtle effects on this ability, with a steeper learning rate of two-to-one mappings for bilinguals compared to monolinguals, and a greater ability for bilinguals to be flexible in the application of the ME constraint. These results show that the parameters determining how word-object mappings are acquired and the role of language experience in driving this learning are complex and varied.

Acknowledgements

This work was supported by a Leverhulme Trust Doctoral Scholarship to KCJC and the International Centre for Language and Communicative Development (LuCiD) at Lancaster University, funded by the Economic and Social Research Council (United Kingdom; ES/L008955/1).

References

Akhtar, N., & Tomasello, M. (1996). Two-year-olds learn words for absent objects and actions. *Developmental Psychology*, *14*, 79-93.

Benitez, V. L., Yurovsky, D., & Smith, L. B. (2016). Competition between multiple words for a referent in corss-situational word learning. *Journal of Memory and Language*, *90*, 31-48.

Escudero, P., Mulak, K. E., Fu, C. S. L., & Singh, L. (2016). More limitations to monolingualism: Bilinguals outperform monolinguals in implicit word learning. *Frontiers in Psychology*, *7*, 1218.

Fitneva, S. A., & Christiansen, M. H. (2011). Looking in the wrong direction correlates with more accurate word learning. *Cognitive Science*, *35*, 367-380.

Gebhart, A. L., Aslin, R. N., & Newport, E. L. (2009). Changing structures in midstream: Learning along the statistical garden path. *Cognitive Science*, *33*, 1087-1116.

Hamrick, P., & Rebuschat, P. (2012). How implicit is statistical learning?. In P. Rebuschat & J. N. Williams (Eds.), *Statistical learning and language acquisition* (pp. 365-382). Berlin: de Gruyter Mouton.

Horst, J. S., & Hout, M. C. (2016). The Novel Object and Unusual Name (NOUN) Database: A collection of novel images for use in experimental research. *Unpublished manuscript*.

Ichinco, D., Frank, M., & Saxe, R. (2009). Cross-situational word learning respects mutual exclusivity. In N. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st Annual Meeting of the Cognitive Science Society* (pp. 2214-2219). Austin, TX: Cognitive Science Society.

Kachergis, G., Yu, C., & Shiffrin, R. M. (2009). Frequency and contextual diversity effects in cross-situational word learning. In N. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st Annual Meeting of the Cognitive Science Society* (pp. 2220-2225). Austin, TX: Cognitive Science Society.

Kalashnikova, M., Mattock, K., & Monaghan, P. (2015). The effects of linguistic experience on the flexible use of mutual exclusivity in word learning. *Bilingualism: Language and Cognition*, *18*, 626-638.

Markman, E. M. (1991). The whole-object, taxonomic, and mutual exclusivity assumptions as initial constraints on word meanings. In S. A. Gelman & J. P. Brynes, *Perspectives on language and thought: Interrelations in development* (pp.72-106). New York, NY: Cambridge University Press.

Markman, E. M., & Hutchinson, J. E. (1984). Children's sensitivity to constraints on word meaning: Taxonomic versus thematic relations. *Cognitive Psychology*, *16*, 1-27.

Markman, E. M., & Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology*, *20*, 121-157.

Metzing, C., & Brennan, S. E. (2003). When conceptual pacts are broken: Partner-specific effects on the comprehension of referring expressions. *Journal of Memory and Language*, *49*, 201-213.

Monaghan, P., & Mattock, K. (2012). Integrating constraints for learning word-referent mappings. *Cognition*, *123*, 133-143.

Pierce, J. W. (2009). Generating stimuli for neuroscience using PsychoPy. *Frontiers in Neuroscience Methods*, *162*, 8-13.

Poepsel, T. J., & Weiss, D. J. (2014). Context influences conscious appraisal of cross situational statistical learning. *Frontiers in Psychology*, *5*, 691.

Poepsel, T. J., & Weiss, D. J. (2016). The influence of bilingualism on statistical word learning. *Cognition*, *152*, 9-19.

Qian, T., Jaeger, T. F., & Aslin, R. N. (2012). Learning to represent a multi-context environment: More than detecting changes. *Frontiers in Psychology*, *3*, 228.

Quine, W. V. O. (1960). *Word and object*. Cambridge, MA: MIT Press.

The Object Databank [Computer software]. (1996). Carnegie Mellon University. Retrieved from: http://wiki.cnbc.cmu.edu/Objects

Tomasello, M. (2000). The social-pragmatic theory of word learning. *Pragmatics*, *10*, 401-413.

Tomasello, M, & Barton, M. E. (1994). Learning words in nonostensive contexts. *Developmental Psychology*, *30*, 639.

Trude, A. M., & Brown-Schmidt, S. (2012). Talker-specific perceptual adaptation during online speech perception. *Language and Cognitive Processes*, *27*, 979-1001.

Vouloumanos, A. (2008). Fine-grained sensitivity to statistical information in adult word learning. *Cognition*, *107*, 729-742.

Yu, C., & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, *18*, 414-420.

Modeling Delay Discounting using Gaussian Process with Active Learning

Jorge Chang[1], Jiseob Kim[2], Byoung-Tak Zhang[2], Mark A. Pitt[1], Jay I. Myung[1]
[1]Department of Psychology, The Ohio State University, Columbus, OH 43210, USA
[2]School of Computer Science and Engineering, Seoul National University, Seoul 151-742, KOREA

{*changcheng.1,pitt.2,myung.1*}*@osu.edu,* {*jkim,btzhang*}*@bi.snu.ac.kr*

Abstract

We explore a nonparametric approach to cognitive modeling. Traditionally, models in cognitive science have been parametric. As such, the model relies on the assumption that the data distribution can be defined by a finite set of parameters. However, there is no guarantee that such an assumption will hold, and it may introduce undesirable biases. For these reasons, a nonparametric approach to model building is appealing. We propose a novel framework that combines Gaussian Processes with active learning (GPAL), and evaluate it in the context of delay discounting (DD), a well-studied task in decision making. We evaluate GPAL in a simulation and a behavioral experiment, and compare it against a traditional parametric model. The results show that GPAL is a suitable modeling framework that is robust, reliable, and efficient, exhibiting high sensitivity to individual differences.

Keywords: Gaussian processes; optimal experimental design; delay discounting; nonparametric modeling; Bayesian inference

Introduction

Models of human cognition are built by designing an explanatory or descriptive model that fits data generated in a behavioral experiment. Although a model's parameterization is motivated by assumptions about the cognitive process under study, the empirical data strongly influence model design. Because of this, the design of the behavioral experiment from which the data were generated (e.g., which stimuli were presented to participants) can introduce bias into the model. This can occur, for example, by not sampling the stimulus space adequately, which can then lead to an incomplete or imprecise model. Two ways to reduce such bias are to not commit in advance to which stimuli should be sampled and to make as few assumptions about the cognitive model as possible, such as parameterization and functional form. In other words, make model-building and data collection as data-driven as possible, at least in the initial stage of model development. Gaussian processes (GP) provide a means of achieving these two goals, functioning as a nonparametric framework for experimentation. We evaluated the viability of a GP-based approach for cognitive modeling in humans.

Researchers in psychology have explored the use of GP to model human behavior (e.g., Cox, Kachergis, & Shiffrin, 2012; Griffiths, Lucas, Williams, & Kalish, 2009; Schulz, Speekenbrink, & Krause, 2018; Song, Sukesan, & Barbour, 2018) but it has yet to be a wide-spread approach. Here, we propose a flexible framework for cognitive modeling by combining GP with active learning (GPAL). GPAL extends traditional GP regression by including appealing features for cognitive science tasks. GPAL is capable of simultaneously modeling the data with minimal assumptions and optimizing the experimental design to find the underlying function efficiently. By virtue of being nonparametric, GPAL shows high sensitivity to individual differences and is able to capture a wider array of patterns compared to parametric approaches. This sensitivity should provide high-fidelity models. Optimization is desirable to minimize the length of a testing session, such as when experimentation is expensive (neuroimaging research) or time-constrained (clinical or special populations). While models produced by the nonparametric framework may not provide interpretable parameters, inferences about cognitive functioning can still be made by examining mathematical properties of the function, such as gradient, curvature or area under the curve.

In our study here, we examined the efficiency, reliability, robustness, and sensitivity of GPAL in the context of modeling delay discounting (DD). Data were collected from 30 participants in a delayed discounting task (e.g., "Do you prefer $10 today or $40 dollars in two weeks?"), which measures an individual's ability to delay gratification. This is a common task in decision-making research, and performance shows a strong correlation with other psychological phenomena, including impulsivity and addiction (Green & Myerson, 2004). The one-parameter hyperbolic model is a popular model that assumes future rewards decline in value hyperbolically with the length of the delay. Recent work from Cavagnaro, Aranovich, McClure, Pitt, and Myung (2016) used adaptive design optimization (ADO) to estimate the parameters of the function in an active learning fashion. One of the conclusions from that study is that none of the six models tested were able to capture the full range of behavioral patterns participants displayed in the task. Thus, DD provides a good test-bed in which to evaluate GPAL. The present investigation represents the first step toward validating GPAL as a premier modeling tool for cognitive science research.

Gaussian Process with Active Learning (GPAL)

This section provides background on each component of the proposed GPAL framework. Figure 1 shows a schematic representation of it.

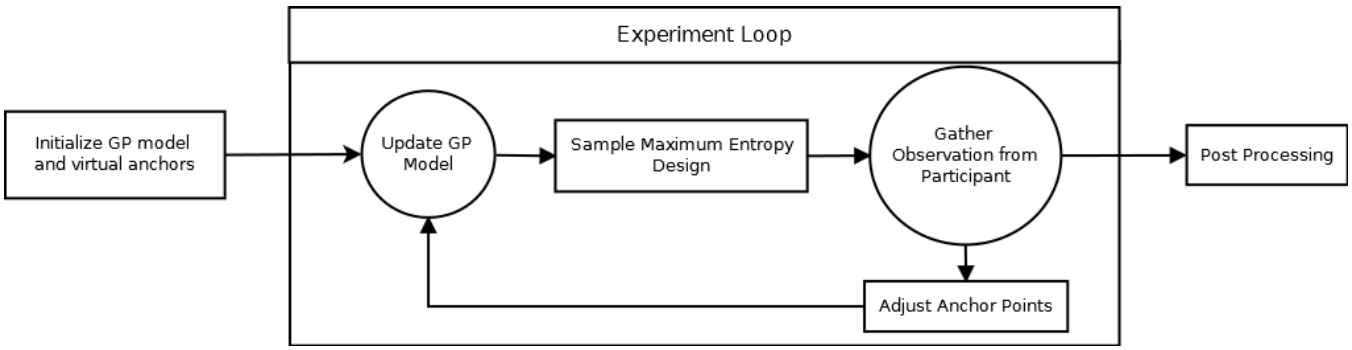


Figure 1: Schematic representation of the GPAL framework. The task is formulated as an active learning based classification task. Virtual anchors are used to restrict the sampling of the design space. On each trial in the experiment loop, an optimal design is picked from the restricted design space according to the maximum entropy criterion, an observation is made, and the GP model and virtual anchors are updated. After the looping, a post processing step may be used to refine the final GP model.

Gaussian Processes

Gaussian processes (GP) are tools for nonparametric Bayesian modeling that establish priors over functions and are a popular approach in machine learning for regression and classification tasks (Rasmussen & Williams, 2006). Formally, GP is a stochastic (random) process where any subset of random variables forms a Gaussian distribution. For a set of observed value pairs (X, f) and a set of unobserved pairs $(\widetilde{X}, \widetilde{f})$, the joint posterior distribution under GP is given by

$$\begin{bmatrix} f \\ \widetilde{f} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu \\ \widetilde{\mu} \end{bmatrix}, \begin{bmatrix} K_{f,f} & K_{f,\widetilde{f}} \\ K_{\widetilde{f},f} & K_{\widetilde{f},\widetilde{f}} \end{bmatrix} \right) \tag{1}$$

where K is a kernel function that defines the covariance between two function values. The kernel function used in this study is the squared exponential kernel which is defined by a length scale parameter l that controls the smoothness and the variance parameter σ^2, which is a measure of the average distance to the mean. This kernel function is a popular choice that has several desirable properties and is known to work well with very smooth functions. The posterior in Eq. 1 can then be used to model $\widetilde{F}$ using the conditional of the multivariate normal distribution.

Many tasks in cognitive science such as DD are not able to observe f directly due to the nature of human experiments. Instead, it is common to give participant choices resulting in multinomial observations. In the case of DD, participants are given two choices on each trial, thus resulting in binomial observations which can then be modeled as a GP binary classification task. This is commonly done by applying a sigmoid transformation function (e.g., probit in our case) to restrict the predicted values to a unit interval. As a consequence of this transformation, the likelihood is no longer Gaussian and requires the use of approximate methods to be estimated, such as expectation propagation as done here. We direct the readers to Rasmussen and Williams (2006), and Vanhatalo et al. (2012) for a comprehensive tutorial on GP and related techniques.

Active Learning

Behavioral experiments can be expensive in terms of both money and time, and the more time an experiment takes, the greater the chance that data quality will suffer due to fatigue or boredom. Systems that incorporate active learning are appealing because they mitigate these problems by optimizing efficiency through identifying highly informative design points based on previous observations (e.g., Cohn, Ghahramani, & Jordan, 1996). It is possible to incorporate active learning in GP based system by deriving a measure of information from the GP and then finding the design point that maximizes this objective function. For our experiment, we used entropy as an information theoretic objective function. We use the derivation of entropy in Houlsby, Huszr, Ghahramani, and Lengyel (2011) which approximates the entropy for GP classification.

Like many tasks in cognitive science, design points in DD are sampled from a discrete space. This space needs to be sparse enough to allow human subjects to make meaningful and differentiable choices. Thus, the search space for optimizing experiments is significantly smaller than in other areas that would use this kind of approach, thereby making grid search a better choice to maximize the entropy function than the proposed method in Houlsby et al. (2011).

Constrained Gaussian Process

Models of natural phenomena are often constrained by prior knowledge or experimental design. For example, when studying natural organisms, the range of the model can be constrained by the physical limitations of such organism. Similarly, a model can be constrained by the experimental design. For example, researchers often design experiments such that some of the outcomes are trivial and well-anticipated. Traditionally, these factors are incorporated in the model design and the range of the parameters and design space. This is a bit more difficult to do in GPAL since it is built to be a fully general modeling tool. It is, however, possible to impose some "reasonable" constraints on the bounds, derivatives, and convexity of a GP. For a more detailed explanation of these

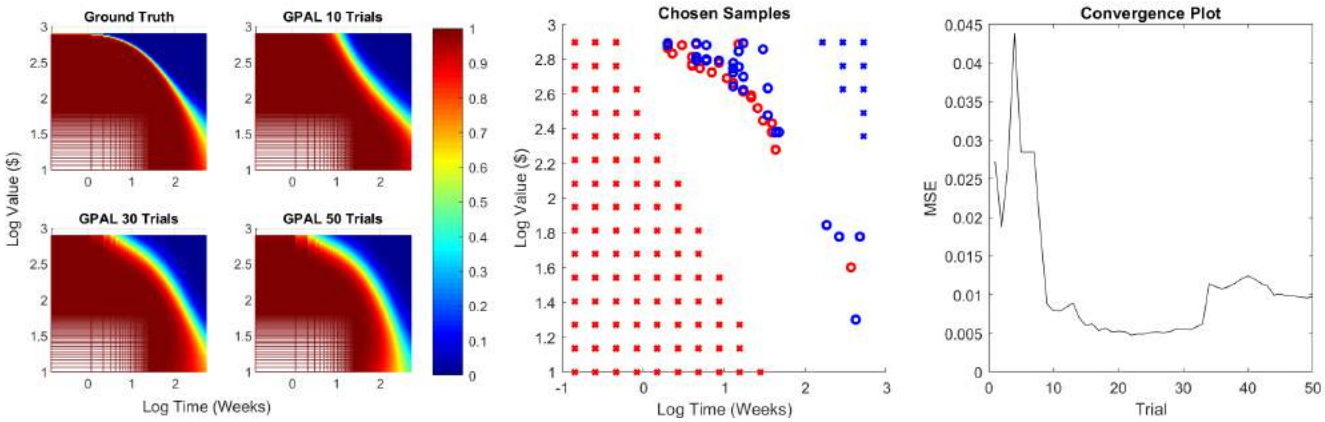


Figure 2: Example of a GPAL simulation using a hyperbolic model as the ground truth. The left panel show GPAL estimated models at four different trials in a top-down view where the z-axis represents the probability of choosing the delayed reward as a function of the two-dimensional, logarithmically scaled design space of (t_{LL}, A_{SS}). The x-shaped and circle-shaped samples in the middle panel represent virtual anchor points and design points selected by GPAL, respectively. The blue and red circles represent choosing the immediate and delayed reward, respectively. The convergence plot in the right panel represents the mean squared error (MSE) between the model at each trial and the ground truth.

methods, see Da Veiga and Marrel (2012). For our delay discounting experiment, we focus on two properties that are often desired in psychological tasks: (a) monotonicity; and (b) local constraints.

Tasks based on subjective preference such as DD often assume that humans follow the axioms of rationality. This leads to researchers building monotonic models that predict a preference for a choice with a higher reward value (or a shorter delay) over another choice with a lower reward value (or a longer delay), if all other things are equal. One way to force monotonicity in GP models is by systematically adding virtual observations in areas where the constraint is violated (Riihimki & Vehtari, 2010). Specifically, this is done by building a joint model of the GP and its derivative. The derivative domain is then used to inject virtual observations into the model when the derivative of the GP violates the monotonicity constraint.

Regarding local constraints, experiments are often designed in a way such that the outcomes are trivial for some design points. For instance, the probability of preferring a choice of $790 now to another choice of $800 in 10 years should be virtually equally to 1. Ideally, we would like GPAL to avoid sampling such trivial design points. Again, inspired by Riihimki and Vehtari (2010), GPAL implements local constraints as follows: We can extend their idea of virtual noiseless observations to local constraints by placing them in trivial regions. We refer to these virtual observations as "virtual anchors". Virtual anchors act as a prior to the function being estimated by reducing its variance so as to avoid sampling in this region. They also have the benefit of removing the need for initial random sampling to start the active learning process. Further expanding on this idea, we can cover large areas of the design space with virtual anchors and systematically remove them using a moving margin that recedes when a design point is sampled nearby. In our simulations and experiment with human participants described below, we used a linear receding margin, though other schemes could also be used, depending upon the problem at hand. We would like to note that these are preliminary results and future work will focus on increasing the robustness of the model.

Models of Delay Discounting

Delay discounting (DD) is a preferential choice task that is often employed to measure impulsivity by quantifying the preference of an sooner-smaller reward (*SS*) against a later-larger reward (*LL*). This measure of impulsivity has been linked to various mental illnesses such as addiction, gambling, and ADHD (Koffarnus, Jarmolowicz, Mueller, & Bickel, 2013; Sharp et al., 2012; Reynolds, 2007). Models of DD typically start by defining the relation between the value of a reward A at time t as:

$$V = AD_t \tag{2}$$

where V represents the discounted value of A, and D_t the discounting factor. Under this framework, DD behavior is modeled by fitting choice data to a discounting curve that models D_t as a function of t. A popular model of choices is the 1-parameter hyperbolic model (Mazur [1987]):

$$D_t = \frac{1}{1+kt} \tag{3}$$

where $k(>0)$ is the parameter related to impulsivity in that high values of k are associated with high levels of impulsivity. Participant choices are fitted to this model by defining a sigmoid choice function for the probability of choosing the LL option over the SS option:

$$P(LL|k,\varepsilon) = \frac{1}{1+e^{\varepsilon(V_{SS}-V_{LL})}} \tag{4}$$

where V_{SS} and V_{LL} are the discounted values of the SS and LL choice options, respectively, and ε (>0) is a free parameter

reflecting consistency of choice behavior. To aid participants in making more meaningful choices and ease visualization, we fix A_{LL} to \$800 and t_{SS} to 0 weeks (i.e., an immediate reward). Thus, the design space becomes a two-dimensional space of (t_{LL}, A_{SS}).

Cavagnaro et al. (2016) extended the hyperbolic model to include active learning by using an adaptive design optimization (ADO) framework. ADO is a parametric framework for Bayesian optimal adaptive experimentation that can be used to select the most informative design for parameter estimation as well as model discrimination (Cavagnaro, Myung, Pitt, & Kujala, 2010; Myung, Cavagnaro, & Pitt, 2013). For our experiments, we use ADO as baseline to compare the performance of GPAL against a parametric model.

Simulations of GPAL

We first tested the feasibility of GPAL in a simulation study in which GPAL was to recover a hyperbolic function used in Cavagnaro et al. (2010) with 10% noise in the observations. Virtual anchor points were incorporated by adding noiseless design points with a value of one or zero at the extreme values of the design space. Figure 2 (left panel) shows an example of the performance with a top-down view of the GPAL estimated DD models at four different trials. Each plot in the left panel represents the probability of choosing the delayed reward as a function of the "later-larger" time t_{LL} and the "sooner-smaller" value A_{SS}. Both dimensions are plotted in the log domain to highlight the difference between functions in our experiments. The decision boundary refers to the regions of interest where the probabilities are closer to 0.5.

The results suggest that GPAL can achieve reasonable convergence within the first 20 trials. Afterwards, as shown in the right panel, we see a decrease in performance (rise in MSE), likely due to Gaussian noise. The design points (red and blue circles) sampled by GPAL at the end of 50 trials, as shown in the middle panel of Figure 2, are reasonable choices that lie close to the decision boundary. The virtual anchors (red and blue x-shaped symbols) provide a reasonable starting point and leave enough distance in case the decision boundary needs to be pushed in one direction or the other. That is, the example shown on the left panel shows that by trial 10 the curvature hasn't been captured yet and the model is still heavily influenced by the anchors. By trial 30, we see a generally close match in shape, with a slight difference in very steep regions as expected due to Gaussian noise.

Modeling Delay Discounting Using GPAL

Experiment

We recruited 30 participants from a pool of undergraduate students at Ohio State University. Participants were asked to perform a DD task over two sessions, ADO and GPAL. The ADO session used ADO to fit the participant choices to the hyperbolic model. The GPAL session fit the data using the GPAL framework with virtual anchors. Both sessions were further divided into two identical and independent sub-sessions to test for reliability.

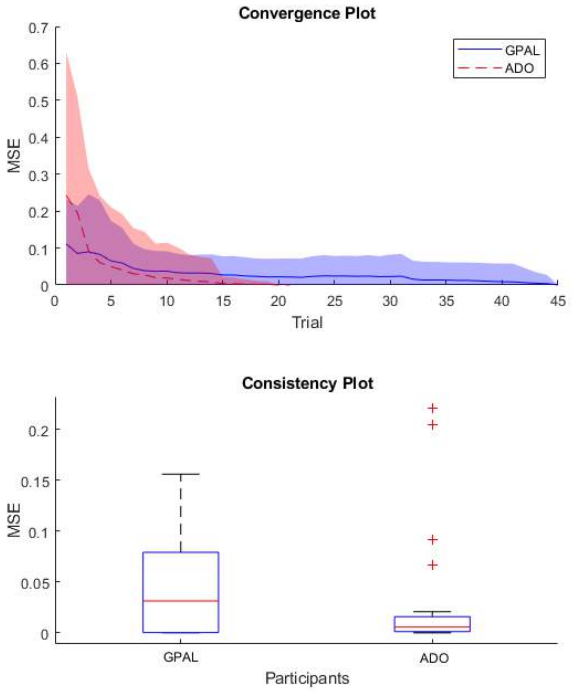


Figure 3: Aggregated results from the experiment. The top panel shows the MSE between the model at each trial and the last. The bottom panel shows the MSE between the first and second session for both experimental conditions.

The sessions were presented in random order and participants were unaware of the identity of the session they were in. Each trial consists of a preference choice presented in the format "\$$X$ now or \$800 in Y time in the future". The value of X (i.e., A_{SS}) ranged from \$10 to \$790 in multiples of 10 whereas the value of Y (i.e., t_{LL}) took on 48 values ranging from 1 day to 10 years spaced on a logarithmic scale. Each session started with 5 practice trials to familiarize participants with the task. This was followed by 20 trials for ADO and 50 trials for each of the two GPAL sub-sessions, for a total of 120 experimental trials. The two GPAL sessions were presented as a single testing block with no break between them. The number of trials was chosen based on previous ADO experiments and GPAL simulations. All the GPAL software was developed and implemented in MATLAB with the aid of *GPStuff* library for Gaussian processes (Vanhatalo et al., 2012).

Results and Analysis

We tested GPAL on its efficiency, reliability, robustness and sensitivity. Efficiency was assessed by comparing the convergence speed between the two frameworks, ADO and GPAL. We expect efficient models to converge quickly to a final solution. We measured this by the speed at which they approach their final solution. Figure 3 (top panel) shows the MSE at

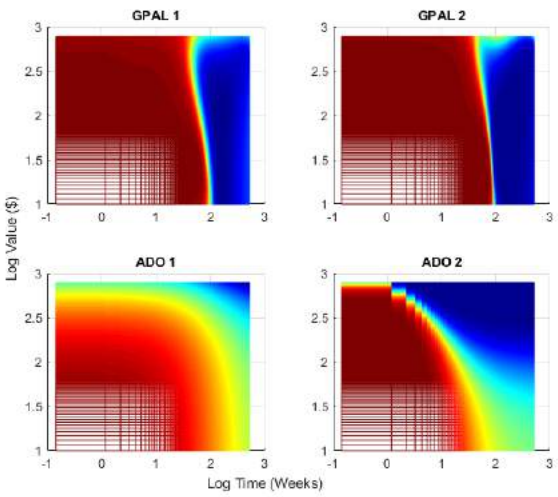


Figure 4: An example participant with consistent results between sessions in the GPAL condition but with inconsistent results in the ADO condition.

each trial between the current estimated model and the last model of the experiment. Both models achieve reasonable convergence quickly with ADO flattening out at around 20 trials and GPAL at 30. GPAL starts with the advantage of having access to the virtual anchors, which act as priors. This means that the initial estimate is much closer to the its final solution compared to ADO which is reflected in the initial values of the results. However, ADO shows faster convergence, which can be seen in the rate and consistency at which the MSE decreases. This is an expected result since ADO assumes a hyperbolic model which allows it to make stronger inferences. We also expect efficient models to pick design points that lie close to the decision boundary of each participant, as these represent the most informative points.

One way to assess reliability is by comparing the GPAL function across the two testing sessions. GPAL, if reliable, should produce consistent results across the sessions, and this is what we find. Figure 3 (bottom panel) shows the MSE between sessions for all 30 participants in both conditions. Overall GPAL performance was good, with an average difference of 0.047, which is deemed quite small. For comparison the average MSE between ADO sessions was 0.026. Again, this is expected due to the added flexibility of GPAL.

As seen in Figure 3, the ADO condition had several outliers that were very inconsistent across sessions. Interestingly, these participants were much more consistent in the GPAL condition. Figure 4 shows an example in which the results for the GPAL condition are significantly more consistent than their ADO counterpart. This was the case for all the outliers in the ADO condition. We find that this phenomena tends to happen when GPAL predicts a function shape that is hard for the hyperbolic model to fit in ADO.

Further inspecting the GPAL results, we find that inconsistent samples are largely produced by a shift in the decision

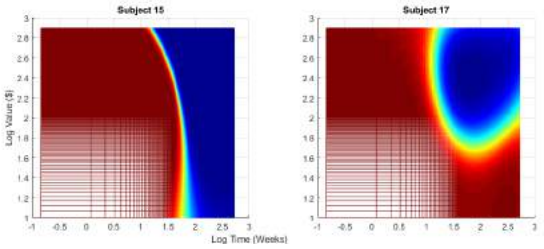


Figure 5: Two selected examples of non-monotonic patterns that consistently emerged during our experiments. The left panel is an example of a non-monotonic pattern with respect to the monetary reward dimension (i.e., y-axis). The right is an example of a non-monotonic pattern with respect to both dimensions.

boundary in the extremes of the design space. Points in this region are also more influential for GPAL because they determine the concavity of the function whereas for ADO, this is determined by the form of the hyperbolic function. These results also suggest that participants tend to be less consistent for designs in this region. Methods to address this problem will be discussed in the next section.

Regarding robustness, we assessed this property by examining a model's ability to predict unseen data. Operationally, robustness was measured by turning an estimated model, whether ADO or GPAL, into a classifier by setting a decision threshold for the predicted probability to generate predicted outcomes. We then tested classification performance by performing cross validation between the observations of each session. In other words, the GPAL-estimated model was used to predict the designs picked by ADO and the ADO-estimated model was used to predict the designs picked by GPAL. Note that both datasets are comprised of data points that are considered to be hard by their respective framework, making them significantly harder to predict than a random sample. We found that ADO performed literally at the chance level of 49.99% accuracy whereas GPAL achieved a 56.53% accuracy. While this result is not particularly impressive, we take this result as evidence that GPAL is able to produce a better classifier or learn better from noisier data than ADO. This result can also be taken as evidence of higher sensitivity to individual differences, since we expect a sensitive model to produce a better and more robust classifier.

Discussion and Conclusion

How does one build a model of human cognition? We introduced GPAL as a data-driven (bottom-up), nonparametric approach with the aim of overcoming biases in parametric modeling approaches for model development and inference. The diversity of data patterns in our experiments illustrates these features of GPAL. GPAL can uncover concave, convex and approximately linear shapes, and do so quickly, providing the modeler with a higher fidelity description of performance. We envision researchers using this information in one of two ways. The straightforward way is to use GPAL as an exploratory tool for providing an unbiased picture of the raw

data to aid the formulation of a parametric model. This gives traditional models a stronger justification in which to ground their assumptions. A second way to utilize GPAL is to replace parametric models altogether. While this second approach requires a paradigm shift in the way models are interpreted, it comes with the potential benefit of providing more accurate measurements. Below we illustrate these ideas by discussing the benefits of GPAL in the context of DD.

Previous models of DD have assumed a monotonic function in both dimensions of money and time. This is a reasonable assumption to make since participants are expected to prefer larger sums of money and shorter time spans. However, Figure 5 shows a few instances that violate this "rationality" assumption. It might be enticing to think that non-monotonic functions in GPAL are a product of noise or model biases. If this is the case, GPAL can be adapted to produce monotonic function using the approach in Riihimki and Vehtari (2010). However, we believe that these non-monotonic patterns are not caused by artifacts. To show this, we focus on the two patterns exemplified in Figure 5. These patterns can be seen across several different participants which we do not show in the interest of space. Since these patterns are repeatable and present across several participants, we find it is unlikely that they are a product of random noise. To support this hypothesis, participants that showed non-monotonic behavior were given five additional trials in which GPAL picked designs using gradient information to identify key regions of non-monotonocity. This process gave a chance for participants to correct their choices to reflect a more conventional function. However, only about a fourth of the participants corrected themselves with the rest confirming their previous behavior. We hypothesize that this non-monotonic, irrational behavior is caused by the interaction between non-linearities in the perceived value of money and delay time. Using the two non-monotonic patterns shown in Figure 5, we provide two possible explanations that would produce this outcome. The first pattern on the left could be caused by a "soft" threshold at which the value of money rapidly decreases, making the time component less relevant. Similarly, the pattern on the right could be caused by a threshold in time at which the delayed reward becomes significantly less appealing. In short, we think that being able to observe these kinds of patterns using GPAL can be a powerful tool to justify choices in parametric models.

A more radical idea is to use GPAL as the primary modeling tool. One of the main benefits of a parametric approach is the ability to formulate theories based on a small set of parameter values. In the case of DD, the k parameter of the hyperbolic model is of particular interest because it is thought to be related to an individual's impulsivity. An analysis such as this is not possible when using a nonparametric model like GPAL since the number of parameters is not constant. However, one could still extract meaningfully information from a GPAL-estimated model. One approach that has been explored in the literature is to interpret the hyperparameters of the kernel function (e.g. Wu, Schulz, Speekenbrink, Nelson, & Meder, 2018). In our case however, the hyperparameters of the square exponential kernel do not relate well to the k parameter. A different alternative could be to use the (parameter-free) estimate of the area under the curve (AUC) of the GPAL function across the input space as an alternative to the k parameter. When this was done for the data from our experiment, the AUC shows a positive correlation to the k parameter. This suggest that the AUC could be used as a measure of impulsivity in a fully nonparametric model but more work needs to be done in this regard. More generally, it is possible to attribute meaning to mathematical properties of GPAL models, which would allow GPAL to function as a primary modeling tool. One benefit of this approach is that the increased sensitivity of the framework might produce more accurate measurements compared to their more constrained counterpart. Additionally, these measurements come from mathematical properties which can be applied to other types of models allowing for easier comparison between models.

Future work will also focus on evaluating the performance of GPAL in a wider array of behavioral tasks. This will allow us to show additional techniques that were not applicable to the DD task. We must also address issues that come from combining GP with active learning. We found that GPAL can be overly sensitive when observations were sparse. While our data suggest that the model is likely to converge within 30 trials, we need to develop the means of ensuring model fidelity while not sacrificing efficiency. The source of this problem is likely due to the greedy nature of active learning. One way to address this problem is to extend active learning to include a bias towards region that are hard for human subjects.

In conclusion, the work in this paper represents a first step towards the development of a novel modeling framework in cognitive science. We propose the use of a nonparametric, model-free approach for cognitive modeling based on GP. This framework serves as a middle ground between raw-data, which are hard to visualize, and parametric models, which rely on strong assumptions. The experiments in the DD task showed that GPAL is a practical framework that yields consistent results efficiently. GPAL showed a high degree of sensitivity to individual differences that were able to uncover non-trivial patterns. This is exemplified by the presence of non-monotonic discounting functions that are present in several participants. These characteristics make GPAL a promising tool for constructing unbiased and sensitive models of cognition.

Acknowledgments

The work was supported by grant FA9550-16-1-0053 from the Air Force Office of Scientific Research.

References

Cavagnaro, D. R., Aranovich, G. J., McClure, S. M., Pitt, M. A., & Myung, J. I. (2016, Jun 01). On the functional form of temporal discounting: An optimized adaptive test. *Journal of Risk and Uncertainty*, *52*(3), 233–254.

Cavagnaro, D. R., Myung, J. I., Pitt, M. A., & Kujala, J. V. (2010). Adaptive design optimization: A mutual information-based approach to model discrimination in cognitive science. *Neural Computation*, *22*(4), 887-905.

Cohn, D. A., Ghahramani, Z., & Jordan, M. I. (1996). Active learning with statistical models. *Journal of Artificial Intelligence Research*, *4*, 129-145.

Cox, G. E., Kachergis, G., & Shiffrin, R. M. (2012). Gaussian process regression for trajectory analysis. *Proceedings of the 34th annual conference of the Cognitive Science Society*, 1440-1445.

Da Veiga, S., & Marrel, A. (2012). Gaussian process modeling with inequality constraints. *Annales de la Faculté des sciences de Toulouse : Mathématiques*, *Ser. 6, 21*(3), 529-555.

Green, L., & Myerson, J. (2004). A discounting framework for choice with delayed and probabilistic rewards. *Psychological Bulletin*, *130*(5), 769–792.

Griffiths, T. L., Lucas, C., Williams, J., & Kalish, M. L. (2009). Modeling human function learning with gaussian processes. In D. Koller, D. Schuurmans, Y. Bengio, & L. Bottou (Eds.), *Advances in neural information processing systems 21* (p. 553-560). Curran Associates, Inc.

Houlsby, N., Huszr, F., Ghahramani, Z., & Lengyel, M. (2011, 12). Bayesian active learning for classification and preference learning. *arXiv preprint:1112.5745*.

Koffarnus, M. N., Jarmolowicz, D. P., Mueller, E. T., & Bickel, W. K. (2013). Changing delay discounting in the light of the competing neurobehavioral decision systems theory: a review. *Journal of the Experimental Analysis of Behavior*, *99*(1), 32-57.

Myung, J. I., Cavagnaro, D., & Pitt, M. A. (2013, 06). A tutorial on adaptive design optimization. *Journal of Mathematical Psychology*, *57*, 53-67.

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes for machine learning (adaptive computation and machine learning)*. The MIT Press.

Reynolds, B. (2007, 01). A review of delay-discounting research with humans: Relations to drug use and gambling. *Behavioural pharmacology*, *17*, 651-67.

Riihimki, J., & Vehtari, A. (2010). Gaussian processes with monotonicity information. In Y. W. Teh & M. Titterington (Eds.), *Proceedings of the thirteenth international conference on artificial intelligence and statistics* (Vol. 9, p. 645-652). PMLR.

Schulz, E., Speekenbrink, M., & Krause, A. (2018). A tutorial on gaussian process regression: Modelling, exploring, and exploiting functions. *Journal of Mathematical Psychology*, *85*, 1 - 16.

Sharp, C., Barr, G., Ross, D., Bhimani, R., Ha, C., & Vuchinich, R. (2012). Social discounting and externalizing behavior problems in boys. *Journal of Behavioral Decision Making*, *25*(3), 239-247.

Song, X. D., Sukesan, K. A., & Barbour, D. L. (2018, Apr 01). Bayesian active probabilistic classification for psychometric field estimation. *Attention, Perception, & Psychophysics*, *80*(3), 798-812.

Vanhatalo, J., Riihimki, J., Hartikainen, J., Jylanki, P., Tolvanen, V., & Vehtari, A. (2012). Bayesian modeling with gaussian processes using the gpstuff toolbox. *arXiv preprint:1206.5754*.

Wu, C. M., Schulz, E., Speekenbrink, M., Nelson, J. D., & Meder, B. (2018). Generalization guides human exploration in vast decision spaces. *Nature Human Behaviour*, *2*, 915-924.

Influence of linguistic tense marking on temporal discounting:
From the perspective of asymmetric tense marking in Japanese

Qixiang Chen (keisyou1993@gmail.com)
Graduate School of Arts and Sciences, The University of Tokyo
3-8-1, Komaba, Meguro-ku, Tokyo 153-8902, Japan

Hidehito Honda (hitohonda.02@gmail.com)
Department of Psychology, Yasuda Women's University
6-13-1, Yasuhigashi, Asaminami-ku, Hiroshima 731-0153, Japan

Kazuhiro Ueda (ueda@gregorio.c.u-tokyo.ac.jp)
Graduate School of Arts and Sciences, The University of Tokyo
3-8-1, Komaba, Meguro-ku, Tokyo 153-8902, Japan

Abstract

There has been much discussion around the Linguistic-Savings Hypothesis (LSH), which postulates that language can affect intertemporal choices of its speakers; the validity of this claim has remained controversial. To test the LSH independent from the possible influencing factors, such as cultural differences, we focused on the Japanese language, which features asymmetric tense marking, in that past tense is grammatically marked but future tense is not. Adopting a within-participant design, we compared the discounting behavior between past and future gains in native Japanese participants. Our results revealed that Japanese speakers tended to discount the values placed on rewards in an asymmetry way: to discount the value of past gains more heavily than that of future gains. We believed our results corroborated the LSH and linguistic relativity.

Keywords: Intertemporal discounting; Intertemporal choice; Linguistic-Savings Hypothesis; Tense; Linguistic relativity

Introduction

Intertemporal choices, regarding trade-offs between time and benefits, are very common in everyday life, such as the decision on whether to spend the salary on a trip to Kyoto immediately after receiving it, or whether to save up for years to buy an apartment in Tokyo. Economical behaviors such as investment and insurance purchasing, as well as health issues such as nicotine addiction and drug abuse, are also manifestations of intertemporal decisions (see Thaler, 1981; Frederick, Loewenstein & O'Donoghue, 2002 for reviews). On a macro level, it may even play a part in very important economic issues such as national saving rates (Springstead & Wilson, 2000). Because of this ubiquity and significance, intertemporal choices remain a topic of lasting research interest.

Numerous studies dealing with intertemporal preference and temporal discounting behavior have shown that people psychologically discount future gain or loss, and tend to discount more for longer temporal distance (Thaler, 1981; Kirby& Marakovic, 1996).

Previous studies have also found that intertemporal discounting behaviors vary individually and culturally (Gell, 1992; Hofstede, 1997). One of the most intriguing hypotheses holds that people's native language may exert an influence on their intertemporal choices (Chen, 2013).

Does language matter in intertemporal discounting?

Whether the language people speak will influence their intertemporal choices has recently been under hot debate. Linguistic-Savings Hypothesis (LSH), proposed by Chen (2013), has been one of the most intriguing hypotheses on this topic. According to Chen (2013), speakers of languages which grammatically distinguish between present and future, such as English, and speakers of languages with no such distinction, such as Mandarin, tend to have different feelings about temporal information, leading to different discounting behaviors. Specifically, for futureless language speakers, the distinction between present and future is vaguely construed, while speakers of languages with separate tense marking for present and future tend to perceive the distinction more clearly. As a result, speakers of a futured language tend to discount future rewards more heavily than those of a futureless language. Chen (2013) substantiated the hypothesis by results from analysis of massive databases of savings rates, health behaviors and retirement assets across many countries. This simple yet intriguing hypothesis has attracted major attention (e.g., by 2018, over 1,790 thousand views on TED talk, 2012).

Meanwhile, the hypothesis has been also criticized and challenged from multiple perspectives. First of all, it has been pointed out that the analysis of massive database is basically indirect (i.e., focusing on correlational relationships). Thus, the validity of causal inferences may be doubtful (Roberts &Winters, 2013). Secondly, it may be difficult to eliminate the influence of cultural differences. Previous cross-cultural analyses of temporal discounting behavior have generated contradictory results (Thoma & Tytus, 2018), thus rendering the results indefensible when taking cultural differences into account. Lastly, empirical evidence on the hypothesis is

mixed. While there is evidence from behavioral experiments in support of the LSH (e.g., Lergetporer et al., 2014), opposite results have also been obtained (e.g., Thoma & Tytus, 2018).

Against this background, our research started with the same point of view with the LSH, but tried to eliminate the influence of alternative explanations, such as the influence of cultural differences, by conducting the experiment using within-participant design.

Asymmetric tense marking in Japanese

In the present study, we examined how Japanese people discounted the value of past and that of future.

Even though the measurement of the intertemporal discount in past is not as familiar as that of future in the field of economics and psychology, it is widely applied in the field of health behavior. The discounting rate of past can be a valid indicator of patience as well as an effective predictor of cigarette and drug abuse.

Previous research found native speakers of English in the U.S. tend to discount the value of future and past gains in a symmetrical way, with no significant difference between the discounting rate for past and future gains (Bickel et al., 2008; Yi et al., 2006). This finding can be explained by the LSH, which predicts that the tense encoding in a language can influence its speakers' time perception and intertemporal discounting behavior. Therefore, native speakers of English are predicted to discount the future and past values in the same way since both past tense and future tense exist in the English language. This symmetric tense marking in English can lead to the symmetric discounting behavior towards past and future gains.

However, not all languages encode tense in the same way. Japanese, for example, has asymmetric encoding in marking only the past tense. The grammaticalization of tense in English, Japanese, and Mandarin is summarized in Table 1.

The LSH can be tested in the Japanese language which features asymmetric tense marking. According to the LSH, the grammaticalization of tense in a language will influence its speakers' discounting behavior towards past and future gains. Therefore, Japanese speakers are predicted to showcase discounting behaviors also in an asymmetric way. To be specific, since there is past tense and no future tense in Japanese, speakers may feel the distinction between past and present more clearly (i.e., larger) than that between present and future, leading to higher discounting rate for past gains than that for future gains.

In the current research, we recruited native speakers of Japanese as participants, and compared the discounting rate of past gains with that of future gains of each participant. Since only Japanese speakers were targeted, the effect of culture was controlled for. Furthermore, the within-participant design also excluded the influence of other potent factors such as individual characteristics, educational level, and economic status between different groups.

To our knowledge, this is the first study to examine LSH directly by comparing discounting rates of past and future gains while excluding the influence of cultural differences as well as other factors.

Table 1: Tense marking in English, Japanese, and Mandarin.

	Past Tense	Future Tense
English	+	+
Japanese	+	-
Mandarin	-	-

Statistical analyses

Indicators and models of temporal discounting

A brief account of the analytical procedure is given in this sub-section. The first step is to estimate the indifference point between two intertemporal choices (e.g., Kirby & Marakovic, 1996; Toubia et al, 2013). An indifference point is reached where the amount available now (following Yi et al., (2006), we regarded 'one hour ago' and 'in one hour' as 'now') is equivalent to the delayed amount in the future. For instance, if a participant preferred to 'receive ¥80,000 (¥10,000 is approximately $100) in one hour' rather than 'receive ¥100,000 in seven days,' but meanwhile chose to 'receive ¥100,000 in seven days' rather than 'receive ¥70,000 in one hour,' we thus assume that the indifference point between now and a delay of seven days lies between ¥70,000 and ¥80,000 and we determined the indifference point at a delay of seven days of ¥100,000 to be the average of two amounts (in the above example, ¥75,000 in one hour). Likewise, the indifference points can be located for the past scenarios. For example, if a participant preferred to 'receive ¥60,000 one hour ago' rather than 'receive ¥100,000 seven days ago,' but at the same time chose to 'receive ¥100,000 seven days ago' rather than 'receive ¥50,000 one hour ago,' we assumed ¥55,000 as the indifference point.

Based on the estimated indifference points, we conducted both statistical and model-based analyses. For statistical analysis, we compared discounting rate and the Area Under the Curve (AUC) between values in the past versus future scenarios. Analysis was based on four models: linear model, exponential model, hyperbolic model, and q-exponential model, among which the fitted models serves as a basis for discussion on participants discounting behavior.

In the following section, we shall explain these methods in detail.

Discounting rate

Discounting rate of specific temporal distance (r_d) can be calculated with the following equation:

$$r_d = \frac{V_0 - \mathrm{V}'}{V_0}$$

where V_0 is the original value and V' the discounted value.

This equation estimates the discounting rate for each specific temporal distance. To reveal the general tendency in

individual participant's discounting behavior, we also adopted AUC and model-based approaches.

AUC

AUC is a very common model-free approach to estimate the discounting behavior (Myerson et al., 2001). To calculate AUC, each indifferent value point should be plotted in the same figure and then lined up to form a curve. The area under the curve is then calculated to be AUC. In general, as a participant discounts the value more heavily, the AUC value become lower.

In the current study, we used the standardized AUC (i.e., ranging from 0 to 1) as an indicator of a general tendency of discounting.

Model-based analyses

In the following explanation of the four models, V', V_0, and d denote discounted value, original value, and the temporal distance, respectively. k and q are discounting and adjusting parameter, respectively.

Linear Model

Linear model is the simplest model to predict how value is discounted with the span of time.

$$V' = V_0 - \mathrm{kd}$$

Exponential Model

Exponential model is the standard model adopted in related empirical works, with an advantage in explaining drastic discounting behaviors.

$$V' = V_0 \cdot e^{-\mathrm{kd}}$$

Hyperbolic Model

Overall, the hyperbolic model (Mazur, 1987) shows a better fit than the exponential model for its strength in predicting a more decelerated rate of value depreciation over time, which resembles discounting behavior.

$$V' = \frac{V_0}{1 + kd}$$

Q-exponential Model

Apart from the most popular models (i.e., the exponential model and the hyperbolic model) in intertemporal behavior study, recent research suggests that q-exponential model could be a better fit since it can be seen as the generalized style of the above models (Cajueiro, 2006; Takahashi et al., 2014).

$$V' = \frac{V_0}{(1 + k(1-q)d)^{\frac{1}{1-q}}}$$

In this model, q is the adjusting parameter and determines the form of fitting model. When q reaches 1, the model equals the exponential model. In contrast, when q reaches 0, the model equals the hyperbolic model.

Behavioral experiment

Participants Five hundred and five Japanese people (M_{age} = 45.08, SD_{age} = 14.55) participated in this experiment, with balanced age groups, i.e., 98, 102, 102, 102, 101 participants respectively in their 20s, 30s, 40s, 50s and over 50s. There were 255 males and 250 females. They were recruited online and enrolled the study via the Qualtrics system (https://www.qualtrics.com). As a reward, each participant received a coupon which could be redeemed for online shopping in Japan.

Tasks Participants were instructed to perform altogether the following three tasks.

Task 1: Binary choice task Participants were instructed to make a series of binary choices in two hypothetical scenarios, i.e., past (Figure 1) and future (Figure 2).

In both scenarios, instructions were given (i.e., 'Which option would you prefer?') and participants were required to make binary choices between an ¥100,000 reward with temporal distance, and an immediate reward with 10 monetary amounts evenly divided between ¥100,000 and ¥10,000 (i.e. ¥10,000, ¥20,000, ¥30,000, …¥100,000). In the example, the temporal distance is 30 days and the choices are presented in a descending order (from ¥100,000 to ¥10,000). Six temporal distances (i.e., 1, 7, 30, 90, 180, and 365 days) were involved and the amounts were presented in two possible orders (ascending or descending).

Altogether, each participant was required to make 120 binary choices (two tense scenarios × 10 monetary amounts × six temporal distances) in this task. Presentation was counterbalanced for tense scenario (past or future) and order of amount (ascending or descending) and randomized for temporal distances.

Task 2: Impulsiveness measurement Participants were then asked to answer the questionnaire of Barratt Impulsiveness Scale 11 (BIS11), a widely-used measure of individual impulsivity (Patton et al., 1995) containing 30 questions. The Japanese version of the scale was used in the present study (Someya et al., 2001).

Task 3: Demographic information collection Participants were requested to report age, sex, nationality and language skills. The language skills reported included four languages, i.e., Japanese, English, Mandarin and French, and participants' self-evaluation was anchored on a scale of 101 points, from 0 (Unable to Understand), 40 (Conversational Level), 70 (Business Level) to 100 (Native Speaker Level).

Procedure All the participants were presented with the same questions, and with the order of task 1, task 2 and task 3. The questions in the task 2 and task 3 were presented in the same order for all participants and the questions in task 1 were kept counting balanced (as described above). All the questions were presented in Japanese.

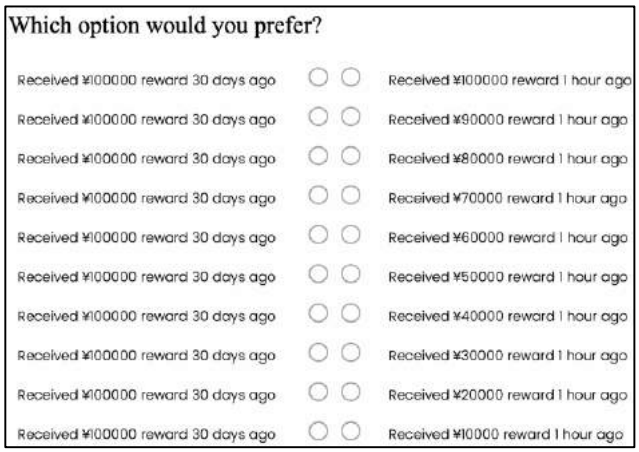


Figure 1: Binary choice task: past scenario

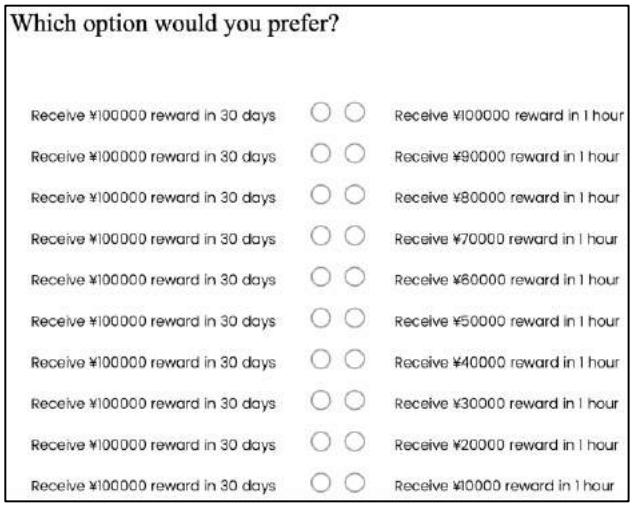


Figure 2: Binary choice task: future scenario

Results

General tendency

Based on participants' answers in the discounting task, we identified the points where the immediate reward of ¥100,000 was equivalent to the amount at temporal distances of 1, 7, 30, 90, 180, 365 days in the past as well as in the future scenario for each participant.

Then, we calculated the indifference points by averaging the equivalent amounts for each temporal distance and plotted them. We fitted the four models to data and chose the best one based on Akaike information criterion (AIC). AIC for each model is summarized in Table 2. It was found that q-exponential was the best model for both past and future discounting. Thus, we analyzed the data based on the q-exponential model.

Figure 3 showed the indifference points plot and q-exponential model for past and future discounting. Overall, as predicted, Japanese speakers discounted past gains ($k_{past}=0.480$) more heavily than future gains ($k_{future}=0.229$). We also used the indicator of (Area Under the Curve) to evaluate the temporal discount. We standardize the area to restrict the value from 0 to 1. The average AUC of past gains ($M_{AUCpast}=0.547$) is significantly smaller than that of future gains ($M_{AUCfuture}=0.624$, $t(504) = 6.843$, $p < .001$, $d = 0.305$), suggesting that the value of past is more sensitive to time transition.

Figure 4 shows the discounting rates for each temporal distance (1, 7, 30, 90, 180, and 365 days) in past and future scenarios, with significant differences between the two scenarios for all temporal distances: 1d ($t(504) = 4.770$, $p < .001$, $d = 0.212$), 7d ($t(504)= 3.980$, $p < .001$, $d = 0.177$), 30d ($t(504) = 5.602$, $p < .001$, $d = 0.249$), 90d ($t(504) = 4.780$, $p < .001$, $d = 0.213$), 180d ($t(504) = 5.503$, $p < .001$, $d = 0.245$) and 365d ($t(504) = 5.710$, $p < .001$, $d = 0.254$).

In line with our prediction, results suggest that Japanese speakers tended to discount past gains more drastically than future gains, as indicated by data of discounting rate, AUC and q-exponential model. This finding also supported the LSH.

Table 2: AIC for the models.

	Past	Future
Linear	15.092	13.789
Exponential	28.231	25.576
Hyperbolic	26.249	23.906
Q-exponential	9.472	8.031

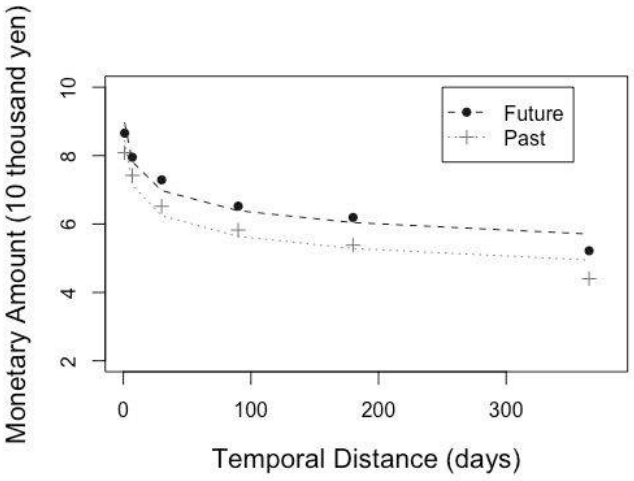
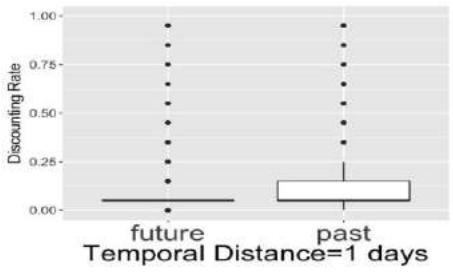


Figure 3: Q-exponential model fits for discounting results.



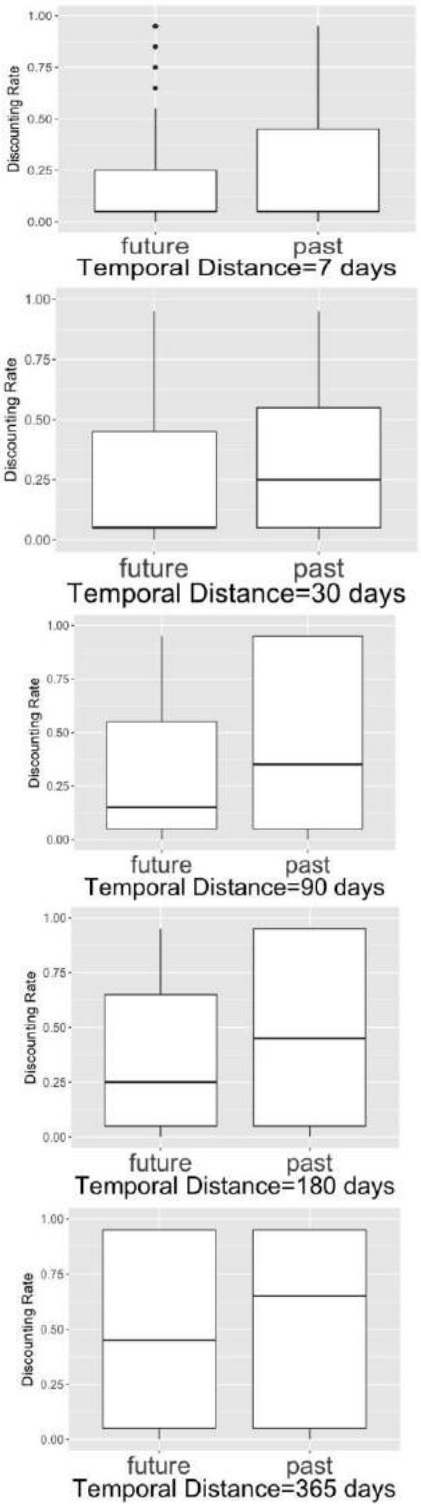


Figure 4: Discounting rate for each temporal distance.

Individual differences and personal characteristics

To further assess individual differences and the effect of demographic factors, we fitted individual participants' data with each of the four models to decide on the best-fitting model for individuals (evaluated by AIC), as summarized in Table 3. Although the q-exponential model was found to be the best-fitting for the overall data, the linear model explained the individual data the best. However, other models were also selected with non-negligible proportions, making it difficult to directly compare individual behaviors. Therefore, we used AUC to evaluate the discounting behavior at the individual level.

To identify the tendency of individuals, we first executed k-means clustering to categorize the AUC values obtained in past and future scenarios.

The first step is to determine the optimal number of clusters. We applied 30 indices in the R package 'NbClust' (Charrad et al.,2014) and experimented with the optimal cluster number from two to ten. Among the 30 indices, 27 returned valid results. Although two clusters were suggested by the largest number of indices (*8/27*), it gave much less information than three clusters, suggested by the second largest number of indices (*6/27*). Balance between parsimony and informativeness, we decided on three as the optimal number of clusters as shown in Figure 5. Each dot displays data for one individual, and the triangles represent the center of each cluster.

Figure 5 showcases the plausible clustering of participants' discounting behavior into three groups, i.e. high discounting group (green dots, n=136), middle discounting group (blue dots, n=183) and low discounting group (red dots, n=186). The diagonal line represents identical discounting rate of past and future gains. The dots above the line are individuals who discounted the value of future gains more, while those below the line denote individuals who discounted the value of past gains more.

As the figure illustrates, individuals of the three clusters show very different tendencies. On average, the discounting rate for past and future gains is very close in both high ($M_{AUCpast}$=*0.205*, $M_{AUCfuture}$=*0.200*) and low ($M_{AUCpast}$=*0.905*, $M_{AUCfuture}$=*0.883*) discounting group. However, it is obvious that participants in the middle discounting group discount the value of past gains much more heavily than that of future gains ($M_{AUCpast}$=*0.436*, $M_{AUCfuture}$=*0.676*). Thus, although the LSH well predicted the general tendency of the participants' discounting behaviors, it failed to capture the specificity at the individual level as it was found to have explained individuals with middle level discounting behaviors better than on average.

We then conducted multiple regression analysis to identify individual characteristics that have influenced discounting behavior. To reveal a full picture, we included age, impulsiveness (measured by BIS-11 questionnaire), language ability (in English, Mandarin and French) as numerical independent variables, and sex (male = 1, female = 0) and tense (past = 1, future = 0) as dummy independent variables, to predict AUC. Since we have confirmed that all participants reported that Japanese is their native language, we excluded the variable of Japanese skill. We also confirmed that correlations between every two variables were low (*cor*<.2).

As multiple regression results (Table 4) show, among all variables, tense (past or future) and impulsiveness had

significant influences on discounting behaviors ($p < .01$). The significant effect of tense is consistent with our major finding that people have the tendency to discount the value of past gains more strongly than that of future gains. With regard to impulsiveness, participants who scored high in the BIS-11 questionnaire tended to have stronger discounting behaviors than those with lower scores in the measurement. This result is consistent with that in the previous study, suggesting that in general impulsive individuals tend to showcase more drastic discounting behaviors (Bickel et al., 2008). No significant influence was found for the other variables. These results indicated that the difference of discounting behavior was explained more in terms of impulsiveness than other demographic factors such as age and sex. Besides, we tried to include the interaction factor of tense and impulsiveness in the model and found there was no significant interaction between tense and impulsiveness ($p > .1$). The result implied that impulsiveness and tense functions on temporal discounting separately.

Table 3: Individual best model percentage.

	Past	Future
Linear	264 (52.3%)	286 (56.6%)
Exponential	90 (17.8%)	78 (15.5%)
Hyperbolic	80 (15.8%)	59 (11.7%)
Q-exponential	71 (14.1%)	82 (16.2%)

Table 4: Multiple regression results for AUC values.

	Est.	SE	p
Tense	**-0.077**	**0.020**	***p<.001***
Age	0.000	0.001	*p=.966*
Sex	0.008	0.020	*p=.683*
English Skill	-0.001	0.001	*p=.375*
Mandarin Skill	-0.001	0.001	*p=.320*
French Skill	0.002	0.002	*p=.373*
Impulsiveness	**-0.005**	**0.001**	***p<.001***

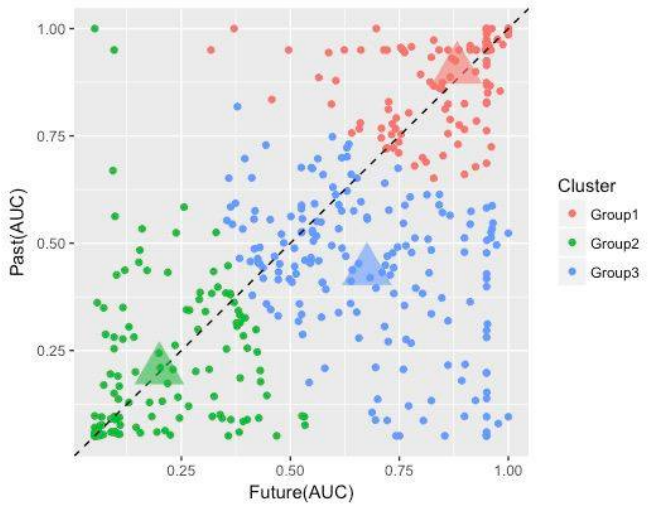


Figure 5: Cluster analysis result of AUC values.

Discussion

The present study examined the LSH by comparing past and future discounting behaviors of individual Japanese speakers to eliminate the influence of potent factors such as culture. We found that Japanese speakers discounted the value of past gains more than that of future gains. This pattern was consistent with the prediction based on the asymmetric grammatical marking of tense in Japanese as there is grammatically marked past tense but no future tense. Thus, our results supported the LSH.

Moreover, detailed analysis of individual characteristics revealed that although the theory could explain the general tendency of discounting behavior, remarkable individual differences remained unexplained. Furthermore, our results suggested that the difference in discounting behavior was explained more in terms of impulsiveness than in terms of demographic characteristics.

Finally, we need to acknowledge that our study focused only on Japanese speakers. Even though there were several previous studies on native English speakers in the U.S. and found they discounted the value of future and past gains in a symmetrical way, we haven't replicated this result and executed the direct comparative analysis by far. This may cause some doubt and alternative explanations here. Our next step is to collect data from native speakers of English and Mandarin for comparative studies to strengthen and broaden our conclusion.

Acknowledgments

This study was supported by JSPS KAKENHI Grant (No. 18H03501) for the second author and JSPS KAKENHI Grant (No. 16H01725) for the last author.

References

Bickel, W.K., Yi, R., Kowal B.P. & Gatchalian K.M. (2008). Cigarette smokers discount past and future rewards symmetrically and more than controls: is discounting a measure of impulsivity? *Drug and Alcohol Dependence*, **96**, 256-262.

Cajueiro, D.O (2006). A note on the relevance of the q-exponential function in the context of intertemporal choices. *Physica A*, **364**, 385–388.

Charrad, M., Ghazzail, N., Boiteau V. &Niknafs A. (2014). NbClust: an r package for determining the relevant number of clusters in a data set. *Journal of Statistical Software*, **61**, issue 6.

Chen, M. K. (2013). The effect of language on economic behavior: Evidence from savings rates, health behaviors, and retirement assets. *The American Economic Review*, **103**(2), 690–731.

Frederick, S., Loewenstein, G. & O'Donoghue, T. (2002). Time discounting and time preference: a critical review. *Journal of Economic Literature*, **40**(2), 351-401.

Gell, A. (1992). The anthropology of time. Oxford, UK: Berg Publisher.

Hofstede, G. (1997). Culture and organizations: software of mind. London, UK: McGraw-Hill

Kirby, K.N. & Marakovic, N.N. (1996) Delay-discounting probabilistic rewards: rates decrease as amounts increase. *Psychonomic Bulletin & Review*, **3**(1), 100-104.

Lergetporer, P., Sutter, M., Angerer, S., & Glatzle-Rutzler, D. (2014). The effects of language on children's intertemporal choices. Beitrage zur Jahrestagung des Vereins fur Socialpolitik 2014: Evidenzbasierte Wirtschaftspolitik—Session: Culture and Social Background.

Mazur, J. (1987). An adjusting procedure for studying delayed reinforcement. In M. Commons, J. Mazur, J. Nevin, & H. Rachlin (Eds.), *The effect of delay and of intervening events on reinforcement value*, 55-73. Hillsdale, NJ: Erlbaum.

Myerson, J., Green, L. & Warusawitharana, M. (2001). Area under the curve as a measure of discounting. *Journal of Experimental Analysis of Behavior*, **76**, 235-243.

Patton, J.H., Stanford, M.S. & Barratt, E.S. (1995). Factor structure of the Barratt Impulsiveness Scale. *Journal of Clinical Psychology*, **51** (6), 768–74.

Roberts, S., & Winters, J. (2013). Linguistic diversity and traffic accidents: Lessons from statistical studies of cultural traits. *PLoS ONE*, **8**(8), e70902.

Someya, T., Sakado, K., Seki, T., Kojima, M., Reist, C., Tang, S. W. & Takahashi, S. (2001). The Japanese version of the Barratt Impulsiveness Scale, 11th version (BIS-11): its reliability and validity. *Psychiatry and Clinical Neurosciences*, **55**, 111-114.

Springstead, G.R., & Wilson, T.M. (2000). Participation in voluntary individual savings accounts: an analysis of IRAs, 401(k)s, and TSP. *Social Security Bulletin*, **63**, 34-39.

Takahashi, T., Tokuda, S., Nishimura, M. & Kimura, R. (2014). The q-exponential decay of subjective probability for future reward: a psychophysical time approach. *Entropy*, **16**, 5537-5545

TED talk. (2012). Could your language affect your ability to save money? Available at: https://www.ted.com/talks/keith_chen_could_your_language_affect_your_ability_to_save_money?language=en Accessed December 30, 2018.

Thaler, R.H. (1981) Some empirical evidence on dynamic inconsistency. *Economic Letters*, **8**(3), 201-207.

Thoma, D. & Tytus, A. E. (2018). How cross-linguistic differences in the grammaticalization of future time reference influence intertemporal choices. *Cognitive Science,* **42**, 974-1000.

Toubia, O., Johnson, E., Evgeniou, T. & Delquie, P. (2013). Dynamic experiments for estimating preferences: an adaptive method of eliciting time and risk parameters. *Management Sciences*, **59**(3), 613-640.

Yi, R., Gatchalian, K.M., Bickel, W.K. (2006). Discounting of past outcomes. *Experimental and Clinical Psychopharmaco*logy, **14** (3), 311–317.

The Goal-Dependent Nature of Automatic Semantic Priming

Lin Khern A. Chia (lachia2@illinois.edu)

Jon A. Willits (jwillits@illinois.edu)

Department of Psychology, University of Illinois at Urbana-Champaign, IL 61820 USA

Abstract

Despite the fact that priming is one of the most studied phenomena in cognitive psychology, many questions remain about exactly when, why and under what task conditions we ought to observe priming in the lab, and what types of relationships between words or concepts reliably lead to priming. This project contrasted two priming experiments where the primary manipulation was the decision the subjects were making about words (as well as manipulating other factors, like relatedness proportion, known to affect priming). We found evidence that: 1) automatic priming for semantically related words does happen under some conditions, but 2) semantic priming, and whether it happens independent of association, is dependent on the task in which participants are engaged. These results provide evidence for the context sensitive nature of the activation of semantic memory.

Keywords: Semantic memory**;** Semantic Priming; Associative Priming; Goals; Explicit Awareness

Introduction

Priming, or the improvement in performance in a perceptual or cognitive task relative to some baseline, is one of the most studied effects in cognition (McNamara, 2005). Much of this interest is because of priming's potential for giving us a window into our representations and how we access them. For example, if the word or concept *dog* is responded to more quickly when it is preceded by the word or concept *cat* then when it is preceded by *shoe*, it suggests that our representation of *dog* and *cat* share some relation, association, or overlap that *dog* and *shoe* do not (Collins & Loftus, 1975).

By discovering systematicities about what kinds of words or concepts prime each other, cognitive scientists hope to unravel the nature and structure of how knowledge is represented. For example, McRae, de Sa, & Seidenberg (1997) found different patterns of priming for human-made artifact and natural kind words, leading to claims about differences in the nature of the representations of those words' meanings. Statistically significant priming was only found between natural kind words if those words had high correlated feature overlap. In other words, priming was observed between words like *canary* and *robin*, which share a set of intercorrelated features that co-occur across a broad range of words, like "*has wings*", "*has feathers*", and "*can fly*". But no priming was observed between words like *raspberry* and *ruby*, despite the fact that they superficially share many features in common (like "*is red*", "*is small*" and "*is round*"). Unlike "*has wings*", "*has feathers*", and "*can fly*", these features are not correlated across a broad set of items. McRae et al found that, in contrast to natural kind words, priming occurred for human-made artifact words that had high feature overlap, regardless of whether those features were correlated or uncorrelated. Based on these results, McRae et al. argued that correlated features are in some way important to the representational structure of natural kind concepts but not artifact concepts.

Decades of research has investigated a wide range of relationships between words, and whether those relationships lead to priming, including: normative association strength, co-occurrence in language, synonymy, antonymy, perceptual similarity, feature overlap, shared category membership, shared script/schema membership, functional relations, and others (for review, see Hutchison, 2003; & McNamara, 2005). However, conclusions about what types of relations systematically lead to priming are made difficult by the fact that many factors unrelated to the target-prime relationship also influence the extent to which semantic priming occurs. One such factor found by Moss et al. (1995) is that words belonging to the same category prime when presented auditorily, but not as text.

A second moderating factor is the type of task subjects are asked to perform during presentation of a target, can influence priming results. Examples of tasks are naming the target word aloud, or deciding whether the target is a legal string in the English language. The latter is called a lexical decision task, which we will abbreviate LDT.

A third moderating factor is the time duration between the prime and target (the stimulus onset asynchrony, or SOA). Hutchison (2003) reviewed 36 experiments (shown in Figure 1 below) examining priming for words belonging to the same category (and which were not normatively associated). He found that in experiments where the task was lexical decision, priming almost always occurred regardless of SOA, whereas in naming studies, priming effects were much less consistent.

A final moderating factor is the relatedness proportion (RP) of words in the study. In a typical study this can range from as high as 50% of the items being related, to sometimes being as low as 5% (Hutchison, 2003). Like SOA, RP effects can dramatically alter whether words of certain types prime each other. RP and SOA are often seen as working in a similar mechanistic fashion, by altering the extent to which the subject is explicitly aware of the potential for a relationship between the prime and target.

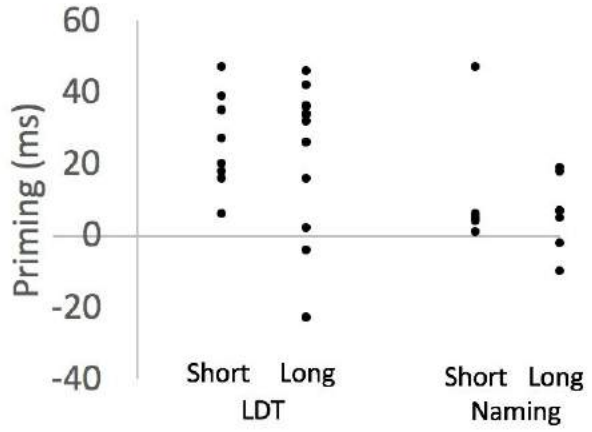


Figure 1. Priming effect sizes for same category words (in ms) in 36 experiments, as a function of task (lexical decision vs. naming) and short (<300 ms) or long (>300 ms) SOA. Data taken from Hutchison (2003) Table 4.

When subjects have a long time to process the prime, and a high proportion of items are related, there is a high chance they are making active, explicit, or strategic predictions about the items (Neely, 1991). In contrast, when the speed is very quick, or a small proportion of items are related, the chance of this is much reduced, and priming effects are often attributed to unconscious or automatic effects like spreading activation (Collins & Loftus, 1975).

With so many factors moderating or eliminating priming effects, we do not yet understand priming well enough to use it as a tool for probing which words' or concepts' representations are related. In this paper, we hope to bring some clarity to these issues. We do so by controlling and contrasting task, SOA, and RP within the same experiment and using the same items. This will allow us to see if semantic priming can be consistently obtained and the factors that affect these priming effects. In doing so, we hope to answer three primary questions.

First, can we reliably obtain priming for items that are semantically related (in terms of being from the same category), in the absence of other types of relations? For example, one other factor moderating the studies shown in Figure 1 was the extent to which the words were normatively associated (e.g. the prime reliably elicited the target in a free association task). Of the studies shown that failed to find semantic priming, the overwhelming majority used same category words that were not normatively associated, whereas the studies that did find priming used words that were both associated and from the same category. This had led some to argue that most priming is "associative" priming, and that purely semantic (e.g. category or feature-based) priming are rare, weak, or nonexistent. This explanation is somewhat dissatisfying, however, because the fact that two words are associated in a normative task does not tell us much about the nature or cause of that association.

One possible explanation for the lack of priming without association deals with the strength of the similarity of the items. In some of the studies testing same-category priming for unassociated words, the category-based relationship was rather weak. For example, Shelton & Martin (1992)'s experiments 2 and 3 found priming times of 2 ms and -23 ms (in a long SOA LDT task), but many of their "related" items were of questionable relatedness, including *dirt* and *cement*, *soup* and *juice*, *barn* and *home*, and *duck* and *cow*. Thus, in this experiment the lack of associativity was confounded with the lack of strong semantic similarity (in terms of feature overlap or any other definition). Other studies, such as McRae and Boisvert (1998) that used more strongly related words, found evidence for semantic priming in the absence of association. In order to address the question of whether semantic priming exists independent of association, in our experiments we choose items that are from the same set of eight categories (mammals, birds, fruits, vegetables, vehicles, clothing, weapons, and tools), and were as highly similar as possible, but varied the degree of association so that its effect could be investigated statistically as a covariate in our analyses.

Second, is semantic category-based priming consistent across different relatedness proportions and stimulus onset asynchronies? As noted, there has been inconsistency in whether semantic priming is found with short SOAs or low RPs, leading some to suggest that semantic priming (as opposed to associative priming) is only an explicit or strategic phenomenon that occurs when subjects might be aware of the fact that words in the study are related, and that therefore automatic unconscious semantic priming does not occur. But again, many of these studies have problems, ranging from small sample sizes to relatively dissimilar "semantically related" words, to not fully crossing RP and SOA. In the experiments described below we ran different sets of subjects in a 2x2x2 design crossing extremely short SOAs (50ms) and moderately short SOAs (250ms), two RP conditions (0.25 and 0.50 related), in addition to whether the words' meanings are related or unrelated (priming would or would not be expected).

A **third** question being tested in this paper is whether automatic priming is dependent on the task-related goals of the subject. Contrary to the depiction of semantic priming as a static phenomenon by a large majority of the literature, Willits et al. (2015) found that what types of

verb-instrument relations led to priming could be manipulated by changing the task. In tasks that had a linguistic bias (such as naming words aloud), priming was observed for words that that have strong linguistic co-occurrence relationships, but not for words that were semantically related that do not co-occur frequently. In contrast, in tasks that were heavily semantic (such as making a category decision about the target word), priming occurred for those words that shared a semantic relationship, regardless of their linguistic co-occurrence probability. Across the current two experiments, we manipulated the task in a similar fashion, observing whether semantic priming is independent of the tasks-specific goals. In Experiment 1, subjects' task was to decide whether the target was a concrete (vs. abstract) entity. Unlike other tasks often used in semantic priming (like naming and lexical decision), this task is one that involves activation of semantic information, and thus may make semantic priming more likely. In Experiment 2, the subjects performed a semantic categorization decision, deciding if words belonged to a particular category (selected from the same eight categories from which the stimulus words were drawn). Critically, sometimes the related pairs were aligned with the category decision being made (e.g. *eagle-hawk* for "is the second word a bird" vs. "is the second word a vehicle). Thus, the contrast between Experiments 1 and 2 allows us to investigate the extent to which semantic priming is consistent across tasks, and whether or not it matters that the kind of semantic relationship being primed is consistent with the subject's current goal.

Experiment 1: Priming in a Concrete/Abstract Decision Task

In Experiment 1, subjects saw a sequence of 128 prime-target pairs. They were asked to judge whether the target was a concrete real object (like a *rock*, *bird*, or *cloud*) or an abstract concept (like *truth*, *beauty*, or *honesty*). Half the items were concrete, and half the items were abstract. Among the concrete items, either 50% or 25% of the items were semantically related. Subjects were randomly assigned to each RP condition, and to either a 50 ms or 250 ms SOA condition. This resulted in a 2x2x2 mixed design, with RP and SOA as between-subject factors, and prime-target relatedness as a within-subject factor.

Method

Subjects. There were 339 undergraduate students who participated in the experiment for course credit. All subjects were fluent speakers of English.

Procedure. Subjects were seated in front of a computer screen. Each of 128 trials consisted of the presentation of the following sequence of events. First, a fixation cross for 50 ms. Second, the prime word (for either 25 ms or 225 ms, depending on SOA condition). Third, a pattern mask ("&&&&&&") for 25 ms (with the duration of the prime word plus the pattern mask constituting the SOA). Fourth, the target word, which stayed on the screen until a response. The inter trial interval was one second. Subjects were required to answer yes or no as to whether the target word was a concrete real object. The trials were randomly divided into eight blocks of 16 words, allowing the subjects a brief resting period between each block.

Materials. Each subject saw 128 noun-noun trials which consisted of 64 concrete-abstract pairs and 64 concrete-concrete pairs. The specific 64 concrete-concrete trials varied across subjects depending on their RP condition. The 128 words making up the 64 concrete-concrete pairs were chosen according the following parameters. First, 16 words from eight semantic categories (mammals, birds, fruits, vegetables, vehicles, clothing, weapons, and tools) were chosen resulting in 64 pairs that were from the same category, maximized semantic similarity, while varying normative association strength.

The experiment's 64 related pairs were then arranged into counterbalanced lists that re-paired 50% or 75% of the targets with unrelated primes (depending on the RP condition). These lists also ensured that each prime and target occurred only once in each list, and that each word occurred as a related prime and target, and as an unrelated prime and target across different lists. For example, the RP=.50 condition had four lists, so that *dog* could occur as a related prime and target (*dog-cat*, and *cat-dog*) and an unrelated prime and target (*dog-shoe* and *shoe-dog*) across the four lists. Each subject saw only one list.

The 128 words making up the concrete-abstract trials were chosen by selecting 64 abstract words and then pairing each one with an unrelated concrete prime word (chosen equally distributed from the same eight categories). These same 64 concrete-abstract pairs were added to each of the lists described above. Note that this means that the RP conditions could be considered .25 and .125 rather than .5 and .25, depending on whether you are considering the relatedness of all trials, or of just the concrete-concrete *yes* trials which constituted our analyses.

Results and Discussion

As per standard convention in priming experiments, we first inspected accuracy scores to check to make sure there were no speed accuracy tradeoffs. Then we analysed the reaction times in the *yes* trials, after removing outlier trials that were shorter than 400 ms or greater than three standard deviations of the mean leaving 20,560 trials (out of 21,234 total *yes* trials) left for analyses. The resulting mean reaction times for related and unrelated trials in our four RP-by-SOA conditions are shown in Figure 2.

Next, we used relatedness, SOA, and RP as fixed factors predicting RT in a mixed-effects regression model, with subject and target word as random factors (Bates, Maechler, Bolker & Walker, 2015). The results of this model are shown in Table 1. We found significant main

effects of relatedness which did not interact with either SOA or RP. Thus we found evidence for a priming effect independent of SOA or RP, and no evidence that our subjects were generating strategic expectations about prime-target relationships, even in SOA/RP conditions that encouraged such expectations.

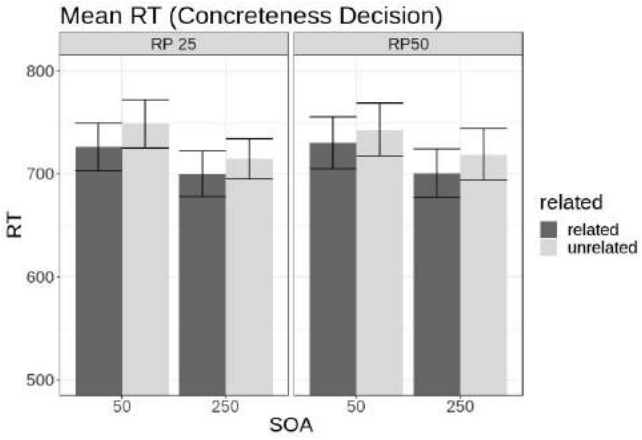


Figure 2. Mean reaction times (and standard deviations computed across subjects) in an abstract/concrete judgement task for related and unrelated trials as a function of relatedness proportion and stimulus onset asynchrony.

Table 1. Fixed effects of mixed effect model analyzing reaction time on *yes* trials in Experiment 1. As per convention, t scores of greater than 2 are typically considered statistically "significant" (Baayen, 2008).

Fixed Effect	b	t
Relatedness	-15.6	-5.39*
SOA	-26.7	-2.34*
RP	-1.20	-0.10
Relatedness x SOA	-2.92	-0.51
Relatedness x RP	5.74	0.99
SOA x RP	7.58	0.33
Relatedness x SOA x RP	-21.6	-1.87

Table 2. Fixed effects predicting residual variance in RT after removing variance in RT predictable by normative association strength in Experiment 1.

Fixed Effect	b	t
Relatedness	-8.22	-1.48
SOA	-21.9	-1.73
RP	-2.76	-0.22
Relatedness x SOA	3.97	0.36
Relatedness x RP	12.4	1.11
SOA x RP	-1.06	-0.04
Relatedness x SOA x RP	-38.0	-1.71

We also fit a second model to the RT data after removing the variance in RT that could be predicted by association strength. This removal of variance was done by excluding the 15,175 trials that: 1) involved normatively associated prime-target pairs, and 2) included targets shared by the normatively associated prime-target pairs so as to ensure equal treatment, leaving 5385 trials for analysis. The effect of relatedness disappeared after removing variance in RT predictable by normative association strength. Thus, in an abstractness judgement task, although we found evidence for an RP and SOA-independent priming effect, we did not find evidence for priming due to "semantic" relatedness (i.e. high similarity items belonging to the same category), when the effect of normative association was removed. This is true even though our items were picked to maximize strength of the relationship between the related prime and target items.

Experiment 2: Priming in a Category Decision Task

Experiment 2 was designed to investigate the extent to which the priming results found in Experiment 1 were dependent on the task in which the subject was engaged. In Experiment 2, subjects' performed a category decision task, deciding if the target word belonged to a specific category (one of the same eight from which the stimuli were drawn).

Method

Subjects. There were 253 undergraduate students who participated in the experiment for course credit. All subjects were fluent speakers of English.

Materials. Items here were identical to those of experiment 1 but for one difference: the *no* trials were, like the *yes* trials, concrete-concrete pairs drawn from the same eight categories. These *no* trials were chosen such that their relatedness proportion matched that of the subject's condition (either 0.25 or 0.50). Thus, each block consisted of related pairs that aligned with the category decision relevant for that block, unrelated pairs with either prime or targets (but not both) aligned with the category, as well as related trials that were misaligned with the category. As an example, consider Table 3, with a sample showing two counterbalanced lists of eight items.

Table 3. Sample items demonstrating related and unrelated pairings in Experiment 2 when task was to decide "Is the second word a mammal?" and RP=0.50. For this sample of words, other counterbalancing lists would have been created allowing all words to serve as both primes and targets in both related and unrelated trials, across different subjects.

Prime	Target	Condition	Correct Response
dog	cat	Related	Yes
rat	mouse	Related	Yes
eagle	deer	Unrelated	Yes
hammer	cow	Unrelated	Yes
sword	knife	Related	No
subway	train	Related	No
moose	shirt	Unrelated	No
zebra	blueberry	Unrelated	No

Procedure. The procedure in Experiment 2 was identical to that of Experiment 1 except for the nature of the yes-no decision, now a category decision. The 128 trials were divided into eight blocks, such that the specific category about which the subject was evaluating the word changed every 16 trials. These eight categories were the same from which the items were drawn. The order of the eight blocks was randomized across subjects.

Results and Discussion

Data in Experiment 2 were analyzed the same way as we analyzed the data in Experiment 1. 14,977 trials (out of 17,092 total *yes* trials) were left after our trimming process. The mean reaction times in each condition are shown in Figure 3.

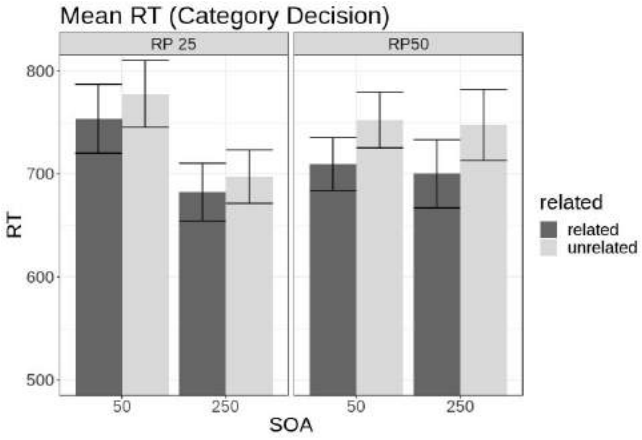


Figure 3. Mean reaction times (and standard deviations computed across subjects) in a category judgement task.

Again, we used relatedness, SOA, and RP as fixed factors predicting RT in a mixed-effects regression model, with subject and target word as random factors. The results of this model are shown in Table 3. We found a significant main effect of relatedness, and also a significant interaction of relatedness with RP (but not SOA).

Table 4. Fixed effects of mixed effect model analyzing reaction time on *yes* trials in Experiment 2.

Fixed Effect	b	t
Relatedness	-22.1	-4.95
SOA	-42.6	-2.95
RP	-3.15	-0.22
Relatedness x SOA	4.76	0.69
Relatedness x RP	-19.3	-2.77
SOA x RP	67.8	2.35
Relatedness x SOA x RP	-4.35	-0.31

As per Experiment 1, we also fit a second model to the RT data after removing the variance in RT that could be predicted by association strength. This removal of variance was done by excluding the 10,563 trials that 1) involved normatively associated prime-target pairs, and 2) included targets shared by the normatively associated prime-target pairs so as to ensure equal treatment, leaving 4414 trials for analysis.

The results of Experiment 2 turned out interestingly different than those of Experiment 1. Here, the main effect

Table 5. Fixed effects predicting residual variance in RT after removing variance in RT predictable by normative association strength in Experiment 2.

Fixed Effect	b	t
Relatedness	-33.4	-3.05
SOA	-40.2	-2.40
RP	9.36	0.56
Relatedness x SOA	-2.23	-0.17
Relatedness x RP	-2.10	-0.16
SOA x RP	80.8	2.41
Relatedness x SOA x RP	0.63	0.02

of relatedness survived the removal of normatively associated word pairs. Thus, the priming observed in Experiment 2 was at least partly due to semantic relatedness, independent of association. This stands in sharp contrast to the way that the priming observed in Experiment 1 can be attributable to effects of normative association.

Why the difference between Experiments 1 and 2? In comparison to an abstract/concrete judgement decision, semantic information (in particular, semantic similarity, overlapping semantic features, or the category to which a word belongs) is clearly more relevant when the task in which the subject is engaged is a category judgment decision. The results of Experiment 2, and their contrast with Experiment 1, strongly suggested that the manifestation of semantic priming depends on the goals or task in which a person is engaged. If their goals beg heavy use of knowledge about semantic features, empirical phenomena of cognitive access like priming should be organized semantically.

There was a significant Relatedness x RP interaction found in the model that included normatively associated pairs. However, this interaction disappeared in the model that excluded normatively associated pairs. This is a curious finding. Alone, these results might suggest that RP effects are selective and only relevant to associative priming. Unfortunately, this conclusion is untenable. Experiment 1, where priming was associative in nature, showed no RP effect at all. We are still left with some uncertainty about the exact role that RP plays in priming.

General Discussion

"Priming is an improvement in performance in a perceptual or cognitive task, relative to an appropriate baseline, produced by context or prior experience." (McNamara, 2005). Priming is typically called semantic when the improvement in performance is brought about by prior experience with semantically related concepts. Due to the fleeting nature of semantic priming, some have expressed doubts that it reflects the true organization of

concepts in our mind. If semantic priming only manifested itself in conditions that encouraged strategic processing, no researcher would be able to use it as evidence that semantic memory is organized semantically. Given further evidence that semantic priming is predictable by word association norms, it might even be reasonable to say that semantic memory is instead organized associatively. However, this is not the conclusion warranted by our data. In Experiment 2, we found evidence for automatic priming (i.e. priming even with very short SOAs and low RPs) for words with no associative relationship.

Despite these findings, it should be stressed that automatic semantic priming is nonetheless a fleeting phenomenon. Consistent with other work about the task and/or goal dependent nature of semantic activation (Willits et al., 2015), we found that priming does not occur independent of task, or with words of low relatedness (Shelton & Martin, 1992). Our results indicate that semantic priming should only manifest reliably when subjects' goals involve heavy usage of semantic information. Subjects' task in Experiment 2 was heavily reliant on semantic information, where they were required to make decisions about the words' membership to categories that were directly aligned to the related vs. unrelated contrasts in stimuli. Experiment 1 on the other hand, required a lighter use of semantic information where nuanced distinctions between and matches of sets of features were not needed. Instead, all that was needed was whether or not the word referred to something that is tangible.

Given the ubiquity of SOA effects in the priming literature, it might be remarkable that our experiments showed no effects of SOA on priming at all. It is worth noting, though, that one of our limitations lie in our SOA manipulations: they were relatively small (50 ms vs. 250 ms) compared to some previous work (which has investigated SOAs as long as 1000 ms). While many have argued that 250 ms is where strategic effects begin to appear, it lies too close to the borderline for us to be certain of any conclusions drawn about pure-SOA effects. Future work could extend these studies to using much longer SOAs, resolving this uncertainty about strategic effects.

Other future directions include investigating the true nature of associative priming, a phenomenon that, because it is defined by a word norm task, is unsatisfying as a mechanistic explanation for priming. An alternative would be to ground associative priming in something more tangible as a mechanistic explanation. Willits et al. (2015), for example, used language co-occurrence statistics to fruitfully predict priming results. Corpora analyses therefore offers a step away from defining the phenomenon by a word norm task. Finally, given our promising results of the existence of semantic priming under at least some circumstances, it is natural that another next steps would be a computational model that is able to predict priming at the level of individual words and/or subjects. Such a model would be a major step towards a truly mechanistic and unambiguous account of the source of lexical priming effects.

References

Baayen, R. H. (2008) *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R*. Cambridge, UK: Cambridge University Press.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*, 1–48.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*, 407-428.

Hutchison, K. A. (2003). Is semantic priming due to association strength or feature overlap? A microanalytic review. P*sychonomic Bulletin & Review*, *10*, 785-813.

McNamara, T. P. (2005). *Semantic Priming: Perspectives from Memory and Word Recognition.* New York: Psychology Press.

McRae, K., & Boisvert, S. (1998). Automatic semantic similarity priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*, 558-572.

McRae, K., De Sa, V. R., & Seidenberg, M.S. (1997). On the nature and scope of featural representations of word meaning. *Journal of Experimental Psychology: General*, *126*, 99-130.

Moss, H. E., Ostrin, R. K., Tyler, L. K., & Marslen-Wilson, W. D. (1995). Accessing different types of lexical semantic information: Evidence from priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 863-883.

Neely, J. H. (1991). Semantic priming effects in visual word recognition: A selective review of current findings and the, ones. In D. Besner & G. Humphreys (Eds.), *Basic processes in reading: Visual word recognition* (pp. 264-337). Hillsdale, NJ: Erlbaum.

Shelton, J. R., & Martin, R. C. (1992). How semantic is automatic semantic priming?. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 1191-1210.

Willits, J. A., Amato, M. S., & MacDonald, M. C. (2015). Language knowledge and event knowledge in language use. *Cognitive Psychology*, *78*, 1-27.

The Explanatory Value of Mathematical Information in Everyday Explanations

Seth Chin-Parker (chinparkers@denison.edu), Department of Psychology
Sam Cowling (cowlings@denison.edu), Department of Philosophy
May Mei (meim@denison.edu), Department of Mathematics & Computer Science
Denison University, 100 West College Street, Granville, OH, 43023 USA

Abstract

With two experiments, we begin an inquiry into the perceived explanatory value of mathematical entities in everyday explanations. This work is motivated by a philosophical debate about the role mathematical entities play in explanation. Simply put, are the mathematical entities themselves explanatory, or is mathematical talk elliptical or shorthand for talk about the physical entities we are concerned with? Across the two experiments, we found clear evidence that situational factors affected how the mathematical entities were considered. However, when those situational factors are accounted for, participants tended to see more explanatory value for mathematical entities that point to other objects involved in the explanation as opposed to mathematical entities that assume the explanatory role themselves.

Keywords: explanation; mathematical explanation; indispensability argument; nominalism; platonism

As scientists, we often appeal to mathematical entities within the explanatory frameworks we adopt. These entities can take a variety of forms, from simple numerals (e.g., '7' and 'thirty') and functions (e.g., '*f(x)*') to complex computational models. Most cognitive scientists, but by no means all, recognize the usefulness of this mathematical information, and there has been extensive commentary on its role and how it should be interpreted. In recent years, philosophers have taken up this question with an increased focus on historical and contemporary case studies in the natural sciences (e.g. Lange, 2016; Pincock, 2011). However, outside of these formal, scientific frameworks, there is arguably a less well developed sense of what role mathematical entities play in explanations.

In this paper, we consider how a live philosophical debate about the explanatory role of mathematical entities relates to everyday explanations. Do mathematical entities contribute to the explanatory work themselves or are they "merely" drawing out the structure necessary for the explanation, identifying the relevant conceptual entities that are actually doing the explanatory work?

Within philosophy of mathematics, platonists affirm the existence of mind-independent and abstract mathematical objects, while nominalists deny that there are any such entities (see Cowling, 2017, for a general discussion of the platonist-nominalist debate). An influential line of argument in defense of platonism is the "Indispensability Argument", which posits that an ontological commitment to mathematical entities of the sort held by platonists is warranted because mathematical entities like numbers and functions play an indispensable explanatory role (Colyvan, 1998). Put differently, platonists make a claim about what exists—namely, that along with concrete entities like electrons and tables, there are also imperceptible, non-spatiotemporal mathematical entities. In contrast, nominalists deny that mathematical entities exist while acknowledging that we must nevertheless explain their usefulness in explanations. We examine whether this distinction that has motivated philosophical debate plays a role in everyday explanation.

In most scientific frameworks, mathematical entities are used to provide formal descriptions of processes and components theorized within conceptual frameworks. For instance, in the categorization literature numerous mathematical models have been proposed to account for how individuals organize items into coherent classes. These models vary from rather simple computations of feature overlap among the items to complex systems of probabilistic computation. They employ mathematical entities in a variety of ways, but there is no assumption that the explanatory value of the models rests on a commitment to the existence of those mathematical entities. Instead, the mathematical entities reference the things, e.g. the features, that are doing the explanatory work. We describe this approach as a *nominalist friendly* (NF) position. On nominalism and the various accounts that have been developed to account for mathematical explanation, see Burgess and Rosen (1997).

One can also accept an ontological commitment to the mathematical entities and allow them to assume explanatory relevance. In this case, the mathematical entities themselves ground the explanation as opposed to simply representing the physical-causal entities and their relations. For instance, consider the explanation for why certain species of cicadas emerge from their nymph state in either 13 or 17 year cycles. The explanation for these life cycles can be understood in terms of avoiding predation

(the cicadas would evolve to have a life cycle that minimizes overlap with the life cycle of predators), but that explanation ultimately rests on the fact that 13 and 17 are prime numbers. The mathematical reality of prime numbers is that they cannot be factored. The explanation for the life cycle of these species of cicada thus relies on a commitment to the mathematical entities as having particular qualities and would therefore be no less real or existent than familiar objects like chairs and racecars (Baker, 2005). We describe this stance as a *platonist friendly* (PF) position.

We use this philosophical debate to background an initial inquiry into how lay people use and evaluate mathematical entities in everyday explanations. We want to be clear that we do not think that people ponder the ontological commitments they are making as they produce or evaluate these kinds of explanations. However, there may be an effect tied to whether the explanations induce genuine ontological commitments to mathematical entities. Indeed, platonists who endorse the indispensability argument standardly assert that, without PF-friendly claims, certain proposed explanations will seem non-explanatory and that, generally, PF explanations are superior to NF explanations (Colyvan, 2018). As we consider below, whether the mathematical entities are represented with regard to their number theoretic value or are merely non-referring placeholders for information about the items they reference will, according to platonists, impact explanatory processes.

Psychologists have examined why people engage in explanation, what implications explaining has for other cognitive activities, and what cognitive structures underlie explanation. There is evidence that people value explanations that are simple and provide coverage in terms of how widely the explanation can be applied (Lombrozo, 2012). There is also evidence that explanatory processes rely on structured internal representations (Chin-Parker & Bradner, 2017; Johnson, Johnston, Koven, & Keil, 2018). These two aspects of explanation suggest that the ontological commitment could indeed play a role in how people regard explanations. For instance, if an explanatory relationship is represented in terms of the number theoretical values (e.g. $5 < 6$), it might be considered simple and widely applicable. If the mathematical entities facilitate the kind of structured representations implicated in explanatory processes, there could be a preference for PF explanations.

However, insights from the psychological study of mathematical reasoning complicate this simplistic rendering of the situation. This literature is vast, so we focus here on two issues. First, there is variability in the ability of people to use and understand mathematical information (Rittle-Johnson, 2017). This variability in mathematical reasoning would likely impact whether an individual is able to easily use the mathematical information to instantiate the requisite representations that the explanatory processes operate over. Second, how the information is presented also impacts mathematical reasoning (Koedinger, Alibali, & Nathan, 2008). In a simple problem, people tend to be more successful when the relevant information is grounded, when it has a clear relationship to concrete referents. When the information is presented in a more abstract manner, e.g. algebraic notation, people are less able to solve the problem. At the same time, the more abstract mathematical entities can facilitate more complex mathematical reasoning. Given these patterns, we expect the type of explanation may interact with the content of the explanation.

We use the logical form of the sentence to determine the ontological commitment of a mathematical statement. For example, 'Thirteen is prime' is PF because it entails that there is something that is prime, which is logically equivalent to the claim that thirteen—a mathematical entity—exists. A NF stance would, consequently, be one in which mathematical terms only appear in non-subject positions—e.g., 'There are thirteen dogs'. Here, 'thirteen' merely modifies the subject, dogs, and the sentence directly entails that there are dogs, but does not, without auxiliary logical assumptions, entail that there is a number thirteen. We note, however, that this assumption is a familiar point of controversy among philosophers and linguists and it is far from clear that lay persons are sensitive to the complex relationship between syntactic position and ontological commitment even if such a view is defensible upon sustained philosophical analysis (see Hofweber, 2016, for a recent discussion). In taking on this account of ontological commitment for the present study, we are, in part, investigating whether certain factors that philosophers of mathematics to take to be of paramount importance are represented in everyday explanatory practices.

To begin our inquiry (Experiments 1a and 1b), we asked participants to generate, and subsequently evaluate, explanations for a series of scenarios. The scenarios varied in terms of their content so that we could assess the generalizability of the participants' ontological commitments across situations. By asking the participants to both generate and evaluate explanations, we were also able to assess whether those commitments vary across different explanatory processes. Thus, the first experiments allowed us to examine whether there is a consistent preference for one type of explanation over the other, or whether the explanation and, in turn, commitment to mathematical entities varies between individuals, situations, and how the information is used. Because of the exploratory nature of this inquiry, we focus on describing the patterns of participant responses relevant to these topics as opposed to testing a priori hypotheses.

Experiment 2 presents a more controlled examination of the issue. We used modified versions of the cicada life-

cycle scenario (Baker, 2005) and asked participants to rate the explanatory value and complexity of various explanations the cicada life cycle. As prior, the explanations varied in terms of whether the mathematical entities made a NF or PF commitment. Also, we varied whether the mathematical terms were developed using a more or less specific example. This manipulation was intended to affect the ease with which participants could represent the information provided in the explanation.

When the explanation was developed using a more specific example, we expected that the participants should find the explanation to be less complex and they should give higher explanatory ratings for PF explanations. This prediction rests on the idea that the grounded mathematical terms could be more easily incorporated into an internal representation and the PF explanation would provide better coverage because it reflects the existence of those entities. However, when the explanation was developed with a less specific example, we expected that the participants would rate it as more complex and they should give higher explanatory ratings for the NF explanations. If the participant has more difficulty representing the situation due to the development of the mathematical terms, an explanation that does not rely on the existence of those mathematical terms should be seen as more explanatory.

In sum, we predict a main effect of the information in the explanation (specific vs. non-specific) on the complexity ratings, and an interaction between the development and the ontological commitment (PF vs. NF) for the explanatory ratings. These predictions rest on the assumption that the explanatory value will reflect both the generalizability and simplicity of the explanation. In order to get a better understanding of the individual differences in play, we also asked participants to report their comfort with mathematics and belief about the existence of mathematical entities.

Experiments 1a and 1b

Methods

Participants Undergraduate students participated as partial fulfillment of a requirement for an introductory psychology course. Thirty participants completed Exp. 1a. Forty participants completed Exp. 1b. Two participants in Exp. 1b failed to complete the explanation generation task, but they did provide ratings of the explanations.

Materials and Procedure The two studies used the same materials, but the method of data collection differed. In Exp. 1a, participants completed the study in small groups. Materials were projected onto a screen, and participants wrote out their responses in prepared packets. In Exp. 1b, participants completed the study on-line by completing a questionnaire created using the Qualtrics platform. See the *Appendix* for the full set of materials used in Exp. 1a and Exp. 1b.

Four scenarios were developed for this experiment. Each scenario presented a set of initial conditions that included mathematical entities (e.g. "The editor of the Daily News has 127 remaining newspapers to deliver and only three paperboys to deliver them.") and then a specific why-question related to those conditions (e.g. "Why can't the editor distribute the papers equally to each of the paperboys?"). The scenarios were designed such that it was possible to answer the question by positing the existence of the mathematical entities (a PF explanation), but a suitable explanation could be made without such a commitment (a NF explanation). In Exp. 1a, the order of the scenarios was balanced, and in Exp. 1b, the order of the scenarios was randomized. In both cases, participants were presented with the scenario and why-question and asked to generate a response.

After responding to all of the scenarios, the participants were told that other students had also generated explanations and those explanations needed to be evaluated. The participants were presented with the same four scenarios – the order of the situations was again balanced (Exp. 1a) or randomized (Exp. 1b). Each scenario was accompanied by two short explanations. One of the explanations reflected PF commitment (e.g. "Because 127 is not divisible by three") and the other reflected NF commitment (e.g. "Because if he gives each paperboy 42 papers, there will be one paper remaining"). In Exp. 1a, participants were asked to select which of the two explanations they considered to be the better explanation. In Exp. 1b, the participants were asked to rate how explanatory each explanation was. Along with each explanation was a slider that could be adjusted from 0 ("not explanatory") to 10 ("ideally explanatory").

Results

Explanation Generation The explanations generated by the participants were coded as to whether they rested on a PF claim, a NF claim, or whether the claim was ambiguous. The explanations were independently coded by two of the study authors, and disagreements were resolved through discussion including the third author. The inter-rater agreement was 87% for the responses from Exp. 1a and 84% for responses from Exp. 1b. Disagreements were easily resolved.

The distribution of the explanatory claims was similar across the two studies. In Exp. 1a, 62% of the explanations were NF, 31% PF, and 8% ambiguous. In Exp. 1b, 65% were NF, 31% PF, and 4% ambiguous. The results indicated that people tend to rely more on NF claims, but that they also will invoke PF claims when deemed appropriate. None of the participants in either study

generated PF explanations for all four scenarios, 8% generated three PF explanations, 22% generated two PF explanations, 49% generated a single PF explanation, and 16% generated no PF explanations. The scenario being explained had an effect on the type of explanation generated. In the "paperboy" scenario, participants readily generated PF explanations (85% of the explanations), while in the other three scenarios, they tended to rely on NF explanations (over 75% of the explanations for each scenario).

Explanation Selection, Exp. 1a The variability between participants and among scenarios is also evident in Exp. 1a when the participants were asked to select one explanation, NF or PF, as more explanatory. No participant consistently selected the PF explanation for every scenario while only six participants consistently selected the NF.

Table 1: Explanation Selection in Exp. 1a

Scenario	PF Selection	NF Selection
Championship	13/30	17/30
Fishing	5/30	25/30
Paperboy	22/30	8/30
Wheat	9/30	21/30

In order to assess explanatory preference, the selection data for each situation were compared to an assumed equal distribution of explanation types using a one-sample binomial test. In both the "fishing" and "wheat" scenarios, the participants showed a consistent preference for the NF explanations (both $ps < .05$). In the "paperboy" scenario, the participants showed a clear preference for the PF explanation ($p < .05$). Only the "championship" scenario had a distribution that indicated that the participants had no preference for the type of explanation.

We did not find evidence that participants made a consistent ontological commitment across the generation and selection tasks. When the participant generated a PF (or NF) explanation for a particular scenario, they subsequently selected the same type of explanation for that scenario only 52% of the time.

Explanation Rating, Exp. 1b The participant ratings for the PF and NF explanations for each scenario were analyzed using a 2 (type of explanation) X 4 (scenario) repeated measures ANOVA. There was no overall effect of the type of explanation, $F(1, 37) = 0.37, p = .55, \eta_p^2 = .01$, a significant effect of the scenario, $F(3, 111) = 6.70, p < .001, \eta_p^2 = .15$, and a significant interaction between the type of explanation and the scenario, $F(3, 111) = 8.54, p < .001, \eta_p^2 = .18$. As can be seen in *Figure 1*, the explanations for the "championship" scenario were significantly lower than the ratings for the other three scenarios (all $ps < .01$).

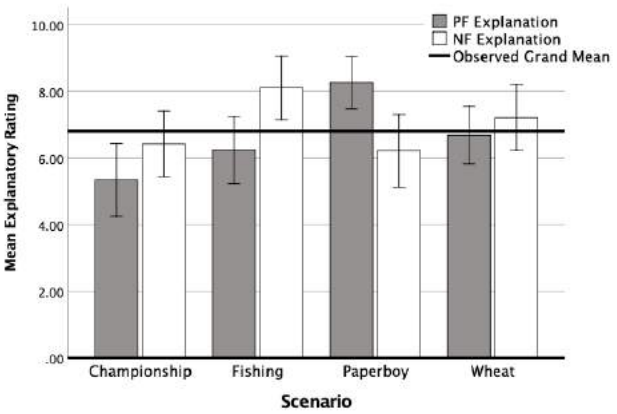


Figure 1: Mean Rating for PF and NF Explanations by Scenario from Exp. 1b

Note. Error bars represent 95% confidence intervals.

The other scenarios did not differ significantly from one another. Paired sample *t*-tests revealed that the ratings for the PF and NF explanations in the "championship" and "wheat" scenarios were not different; $t(39) = -1.25, p = .22$, and $t(39) = -0.64, p = .53$, respectively. However, the type of explanation did affect the ratings in the "paperboy" scenario, $t(39) = 3.14, p < .01$, and "fishing" scenario, $t(39) = -3.03, p < .01$, although in opposite directions.

In Exp. 1b, the participants showed a more consistent pattern of commitment to a particular type of explanation than in Exp. 1a. Overall, the participants gave a higher rating to the explanation that matched the type of explanation they had generated 66% of the time. However, this was primarily driven by participants that generated NF explanations and subsequently rated the NF explanations as better. The participants that initially generated PF explanations rated the PF explanations as better only 50% of the time.

Experiment 2

Methods

Participants Participants ($n = 173$) were obtained using the Mechanical Turk platform. They had to have above a 98% positive approval rating and successfully completed at least 100 tasks within the system. Eight participants were removed for not following directions. The questions included in this study were embedded within an unrelated memory study. Participants were paid for their participation.

Design Explanations varied in terms of commitment of the mathematical entities (either NF or PF) and how the mathematical terms were developed in the explanation (whether they contained a specific or non-specific example). Combining these factors created four conditions, and participants were randomly assigned to one condition.

Materials and Procedure Participants completed the study online using the Qualtrics platform. Each participant read a short passage about cicadas that provided some basic information about their appearance and diet. Importantly, they were informed about the cicada life-cycle being either thirteen or seventeen years. The description ended with the statement, "A question that has interested scientists is why cicadas have this particular life-cycle." Four different explanations were created to address that question.

All of the explanations consisted of six sentences, the first and last sentences were identical across all of the explanations. The second sentence reintroduced the idea that the life cycle of the cicadas was either thirteen or seventeen years. In PF explanations that point was connected to the notion that these numbers are prime:

Interestingly, 13 and 17 are prime numbers – this means that no smaller value (such as 2 or 3) can be divided into these numbers.

In the NF explanations, the number of years was connected to the notion that those numbers could not be evenly segmented:

Interestingly, cicadas' life-cycles are 13 or 17 years long – this means that these periods cannot be segmented evenly into durations of two years, durations of three years, and so on.

In all explanations, the next (third) sentence noted that the length of the life cycle minimized overlap with potential predators. In the PF explanations, this point was explicitly tied to the fact that the length of the life-cycle was a prime number. In the NF explanations, the point was tied to the length of the life-cycle generally. The fourth and fifth sentences developed the idea raised in the third sentence with either a specific or non-specific example. For instance, in the specific PF explanation, the example described how a predator with a three year life cycle would overlap with cicadas with a thirteen year life cycle only once every 39 years, but it would overlap every life cycle with a cicada that had a twelve year life cycle. In the non-specific PF explanation, the development of the explanation relied on algebraic notation:

*If a predator of the cicada had a life-cycle of x years (where x is equal to 2, 3, or 4), it would threaten cicadas with a 13 year life-cycle only once every 13*x years because that number would be the first number that can be divided by both x and 13.*

In the NF explanations, the specific and non-specific examples differed similarly except the examples referred to how the life-cycle could be segmented as opposed to the characteristics of prime numbers.

Immediately following the explanation, two rating scales were presented. The first scale asked participants, "How well does the above account explain the cicada life-cycle?". The participants could move a slider along a scale from 1 ("Not at all explanatory") to 9 ("Very explanatory"). The second scale asked, "How complex would you consider the explanation provided above?". The scale went from 1 ("Extremely simple") to 9 ("Extremely complex").

Following the critical questions, the participant was asked, "Would you consider yourself to be a scientist?" and "Are numbers real?" (Yes/No options for both measures). There was also a measure where the participant was asked to report their comfort with math from 1 ("Not comfortable") to 9 ("Very comfortable").

Results

The participants in the study predominately reported that they did not consider themselves scientists, (7.5% responded "yes" and 92.5% responded "no") and that they considered numbers to be real (95.4% responded "yes" and 4.6% responded "no"). Overall, they reported that they were "reasonably comfortable" with math (m = 5.62, s = 2.23), but there was some variability in those responses. Importantly, the reported comfort with math did not meaningfully vary by condition, $F(3, 169) = 1.71$, $p = .16$, $\eta^2 = .03$.

The explanatory ratings were analyzed using a 2 (specificity) X 2 (commitment) ANOVA (see Figure 2). There was no effect of specificity on the explanatory rating, $F(1, 169) = 0.12$, $p = .73$, $\eta_p^2 = .001$. The mean for the specific explanations ($m = 6.41$, $s = 2.03$) was similar to the mean for the non-specific explanations ($m = 6.32$, $s = 2.05$). There was a significant effect of the mathematical commitment on the ratings, $F(1, 169) = 3.85$, $p = .05$, $\eta_p^2 = .02$. The mean for the NF explanations ($m = 6.67$, $s = 1.86$) was significantly higher than the mean for the PF explanations ($m = 6.06$, $s = 2.17$). There was no interaction between the specificity and mathematical commitment of the explanations, $F(1, 169) = 1.19$, $p = .28$, $\eta_p^2 = .01$.

The complexity ratings were similarly analyzed (see Figure 3). There was no effect of specificity on the complexity rating, $F(1, 169) = 1.88$, $p = .17$, $\eta_p^2 = .01$. The

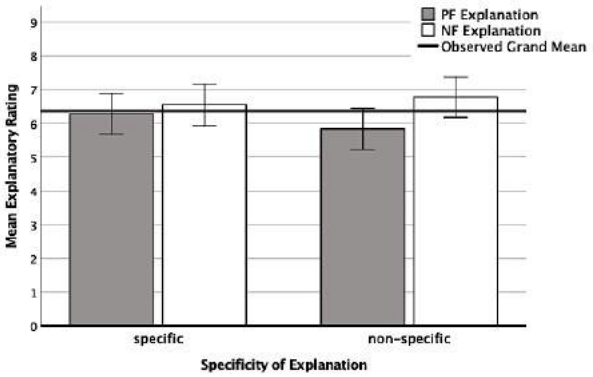


Figure 2: Mean Explanatory Ratings from Exp. 2
Note. Error bars represent 95% confidence intervals.

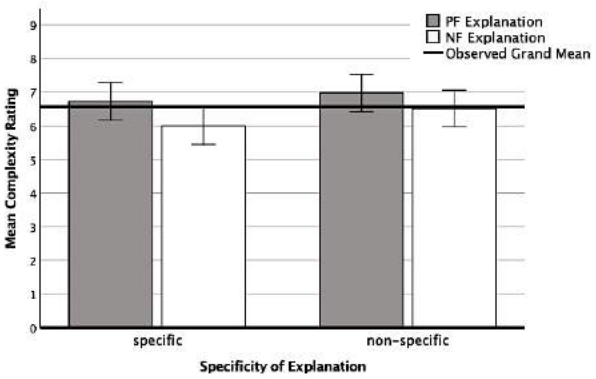


Figure 3: Mean Complexity Ratings from Exp. 2
Note. Error bars represent 95% confidence intervals.

mean for the specific explanations ($m = 6.36$, $s = 1.90$) was similar to the mean for the non-specific explanations ($m = 6.74$, $s = 1.82$). There was a significant effect of the mathematical commitment on the ratings, $F(1, 169) = 4.49$, $p = .04$, $\eta_p^2 = .03$. The mean for the NF explanations ($m = 6.26$, $s = 1.88$) was significantly lower than the mean for the PF explanations ($m = 6.85$, $s = 1.81$). There was no interaction between the specificity and the mathematical commitment, $F(1, 169) = 0.21$, $p = .65$, $\eta_p^2 = .001$.

The participant ratings of how explanatory and complex the explanations were had a weak, negative relationship, $r = -0.12$, $p = .10$. There was no relationship between the participants' reported comfort with math and their explanatory ($r = -0.02$, $p = .80$) or complexity ($r = -0.01$, $p = .88$) ratings. There were too few people that reported themselves to be scientists (or to not believe numbers to be real) to assess how those factors might have impacted their ratings of the explanations.

Discussion

Experiments 1a and 1b showed that there is variability in the mathematical commitments people are willing to make when generating or evaluating explanations for relatively simple situations. It was clear that the variability was not simply an individual difference issue – i.e. there was no evidence that some people always use and value PF entities and other people do not. This result suggests, contrary to some philosophical discourse, that there are not distinctively nominalist or platonist reasoners.

The variation in Exp. 1a and 1b appeared to be largely driven by differences among the scenarios. Across both samples and all measures, participants readily committed to PF explanations for the "paperboy" scenario. It involves the simplest mathematical relations as the explanatory value rests on a single mathematical operation. The "fishing" scenario tended to be the one where explanations ontologically committed to mathematical entities were least valued, and that scenario involves multiple operations across several potential numerical values. The other two scenarios tended to show less consistent patterns of response. This suggests to us that the complexity of the structure of putative explanations might drive much of the variation seen in the participants' preferences for the different types of mathematical explanations. Further study, and more careful control, of the various factors that differentiate these kinds of everyday explanations should provide more clarity as to why people shift in the ontological commitments.

In Experiment 2, we did not find the predicted effect of the specificity on the rated complexity of the explanations. We also did not find the predicted interaction between the specificity and the type of mathematical commitment on the explanatory value. However, among the non-specific explanations, the explanatory ratings differed between the PF and NF conditions ($p = .03$). Even though our manipulation of specificity did not have the expected effect on the complexity ratings, the participants responded to the PF and NF explanations differently when the information was non-specific.

The main results of Experiment 2 were that participants considered the PF explanations to be less explanatory and more complex than the NF explanations. It is possible that the mathematical relations underlying prime numbers are more difficult for people to grasp than we had anticipated. If that is the case, the results across the two experiments align; with more difficult mathematical relations, people perceive the explanations are being less explanatory. This assessment fits with recent work by Johnson, Johnston, Koven, and Keil (2017) and aligns with findings that there is a negative relationship between complexity and explanatoriness in non-mathematical explanations (Lombrozo, 2012). However, that relationship may not always hold (Johnson, Valenti, & Keil, 2017). Alternatively, it is possible that the participants were receptive to the more verbal depictions of the mathematical relations found in the NF explanations (see Koedinger & Nathan, 2004). This would suggest that it could relate more generally to how easily participants are able to represent the relations that underlie the explanation.

The present results suggest that lay people often find NF explanations satisfactory and, in certain instances, preferable to PF explanations. So, if platonists seek to defend the existence of mathematical entities because of their explanatory value or because of the manifest superiority of platonist over nominalist explanations, the present study provides preliminary evidence that such claims cannot be substantiated by our everyday explanatory practices, which are often quite friendly to would-be nominalists. We fully recognize we are not able to resolve the philosophical debate that backgrounds this study, but it does provide an interesting glimpse into how people use mathematical information in explanations.

Acknowledgments

The authors would like to recognize the Lisska Center at Denison University for its support of the pursuit of interdisciplinary research. This research was supported in part by a grant from the Denison University Research Foundation to the first author.

References

Baker, A. (2005). Are there genuine mathematical explanations of physical phenomena?, *Mind*, *114,* 223–238.

Burgess, J., & Rosen, G. (1997). *A Subject with no Object: Strategies for Nominalistic Interpretation of Mathematics*. Oxford University Press.

Cowling, S. (2017) *Abstract Entities*. Routledge.

Chin-Parker, S., & Bradner, A. (2017). A contrastive account of explanation generation. *Psychonomic Bulletin & Review*, *24*, 1387-1397.

Colyvan, M. (1998). In defense of indispensability, *Philosophia Mathematica, 3, 39–62*.

Colyvan, M. (2018). The Ins and Outs of Mathematical Explanation. *The Mathematical Intelligencer*, *40*, 26-29.

Hofweber, T. (2016). *Ontology and the Ambitions of Metaphysics*. Oxford University Press.

Johnson, S. G. B., Johnston, A. M., Koven, L. K., & Keil, F. C. (2018). Principles used to evaluate mathematical explanations. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. J. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 612-617). Austin, TX: Cognitive Science Society.

Johnson, S.G.B., Valenti, J.J., & Keil, F.C. (2017). Opponent uses of simplicity and complexity in causal explanation. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. J. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 612-617). Austin, TX: Cognitive Science Society.

Koedinger, K. R., Alibali, M. W., & Nathan, M. J. (2008). Trade-Offs between grounded and abstract representations: Evidence from algebra problem Solving. *Cognitive Science*, *32*, 366-397.

Koedinger, K. R., & Nathan, M. J. (2004). The real story behind story problems: Effects of representations on quantitative reasoning. *The journal of the learning sciences*, *13*, 129-164.

Lange, Marc. (2016). *Because without Cause*. Oxford University Press.

Lombrozo, T. (2012). Explanation and abductive inference. *Oxford handbook of thinking and reasoning*, 260-276.

Pincock, C. (2011). *Mathematics and scientific representation*. Oxford University Press.

Rittle-Johnson, B. (2017). Developing mathematics knowledge. *Child Development Perspectives*, *11*, 184-190.

Appendix: Exp. 1a and 1b Materials

Championship Scenario

Central High School hosts the league basketball championship game every three years and they host the league volleyball championship every four years. Eighteen years ago, they won both championships on their home court. Why can't they duplicate that feat this year?

(PF) Because the 18 is a multiple of three, but not four.

(NF) Because it will be another six years until they host both championship games again.

Fishing Scenario

Lana has $30 and wants to buy a fishing rod, fishing reel, and fishing line. There are two rods priced at $21 and $22. There are three reels priced at $7, $8, and $9. Fishing line is $2. Lana wants to spend exactly $30. Why should Lana buy the $21 rod?

(PF) Because the sum of 22, 7, and 2 is greater than 30.

(NF) Because any way of combining the $22 rod purchase with the purchase of a fishing reel and fishing line requires spending more than $30.

Paperboy Scenario

The editor of the Daily News has 127 remaining newspapers to deliver and only three paperboys to deliver them. Why can't the editor distribute the papers equally to each of the paperboys?

(PF) Because 127 is not divisible by three.

(NF) Because, if he gives each paperboy 42 papers, there will be one paper remaining.

Wheat Scenario

Fred needs 86 lbs of wheat for winter and he can't afford to waste any money on unused wheat. Wheat comes in bags of 8 lbs. He has 54 lbs of wheat already. Why can Fred avoid buying any unnecessary bags of wheat?

(PF) Because Fred must buy 32 lbs of wheat, and thirty-two divided by eight is four.

(NF) Because Fred must buy 32 lbs of wheat, and if Fred buys four 8 lb bags, he will have 32 lbs.

Problem Difficulty in Arithmetic Cognition: Humans and Connectionist Models

Sungjae Cho[1] (sj.cho@snu.ac.kr), Jaeseo Lim[1] (jaeseolim@snu.ac.kr),
Chris Hickey[1] (chris.hickey@ucdconnect.ie), Byoung-Tak Zhang[12] (btzhang@bi.snu.ac.kr)
[1] Interdisciplinary Program in Cognitive Science, Seoul National University,
[2] Department of Computer Science and Engineering, Seoul National University,
1, Gwanak-ro, Gwanak-gu, Seoul, 08826, Republic of Korea

Abstract

In mathematical cognition, problem difficulty is a central variable. In the present study, problem difficulty was operationalized through five arithmetic operators — addition, subtraction, multiplication, division, and modulo — and through the number of carries required to correctly solve a problem. The present study collected data from human participants solving arithmetic problems, and from multilayer perceptrons (MLPs) that learn arithmetic problems. Binary numeral problems were chosen in order to minimize other criteria that may affect problem difficulty, such as problem familiarity and the problem size effect. In both humans and MLPs, problem difficulty was highest for multiplication, followed by modulo and then subtraction. The human study found that problem difficulty was monotonically increasing with respect to the number of carries, across all five operators. Furthermore, a strict increase was also observed for addition in the MLP study.

Keywords: problem difficulty; arithmetic cognition; binary numeral system; connectionist model; multilayer perceptron

Introduction

Mathematical cognition is the field of research concerned with the cognitive processes that underlie mathematical abilities (Campbell, 2005). Mathematical cognition involves complex mental activities, such as identification of relevant quantities, encoding those quantities into an internal representation, mental comparisons, and cognitive arithmetic. Most notably, cognitive arithmetic is concerned with the mental representation of numbers and arithmetic, and the processes that access and use this knowledge (Ashcraft, 1992).

In cognitive arithmetic, problem difficulty is a central variable (Ashcraft, 1992, 1995). There are at least three criteria for operationalizing problem difficulty: (a) operand magnitude (e.g., 1 + 1 vs. 8 + 8); (b) number of digits in the operands (e.g., 3 + 7 vs. 34 + 78); and (c) the presence or absence of carry[1] operations (e.g., 15 + 31 vs. 19 + 37). In particular, criterion (c) has been further investigated with regard to the number of carries required to correctly solve a problem (Fürst & Hitch, 2000; Imbo, Vandierendonck, & Vergauwe, 2007). In the present study, we investigated how the number of carries affected problem difficulty. *Response time* (RT) from the time a participant sees a problem to the time the participant answers the problem was used in the present study to measure problem difficulty.

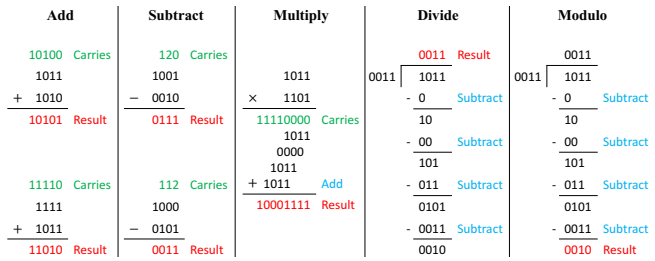


Figure 1: Guiding examples of the five operators with carries.

Previous studies that examine the ways humans process numbers are mostly based on the highly familiar decimal numeral system. Instead, the present study used the binary numeral system, which may offer a novel way to mitigate against the effect of previous experience with conventional mathematical operations. Moreover, since the binary system uses only 0 or 1 digits, it may reduce the *problem size effect*; criterion (a): problems with smaller operands (e.g., $5+2$, $4-1$) are solved more quickly and accurately than problems with larger operands (e.g., $7+6$, $9-6$) (Campbell, 1994; LeFevre et al., 1996; Miller, Perlmutter, & Keating, 1984). Therefore, to observe the effect of carries on problem difficulty, the present study employed the binary system to control for familiarity with the decimal system and criterion (a).

Extending the connectionist approach (Rumelhart & Mcclelland, 1986) to address problems of mathematical cognition may help us understand in detail why mathematics is hard (McClelland, Mickey, Hansen, Yuan, & Lu, 2016). This approach is effective because connectionist models are able to learn many aspects of mathematical cognition. Also, these models offer the possibility to provide concrete instantiations of the mechanisms that grasp the nature of human knowledge and learning within the domain of mathematics.

Previous research has demonstrated how the connectionist model can simulate arithmetic operations. For instance, McCloskey and Lindemann (1992) simulated associative-memory neural networks that learn single-digit multiplication operations. However, these networks were unable to learn all the given arithmetic operations. Utilizing recent advances in

[1] A *carry* in binary addition is the leading digit 1 shifted from one column to a more significant column when the sum of the less significant column exceeds a single digit. A *borrow* in binary subtraction is the digit $10_{(2)} = 2$ shifted to a less significant column in order to obtain a positive difference in that column. This paper refers to borrows as carries.

deep learning, Kaiser and Sutskever (2016) implemented a recurrent network capable of learning either addition or multiplication of two long binary numbers. This model achieved 100% test accuracy. Franco and Cannas (1998) designed multilayer perceptrons (MLPs) that computed either the addition or multiplication of two binary numbers. The MLPs were constructed with at least one hidden layer and binary step functions as activations. Instead of being learned from data, the weights of the MLPs above were analytically designed. Hoshen and Peleg (2016) made MLPs that learned arithmetic addition, subtraction and multiplication from images of two 7-digit decimal integers through a numerical method.

The MLP was chosen as the connectionist model for the present study due to its strong learning ability, owing to the universal approximation theorem. According to the theorem, an MLP can learn any function if its hidden layers are large enough and its activation functions are squashing functions like sigmoid (Hornik, Stinchcombe, & White, 1989). This implies that MLPs should be capable of learning arithmetic/modulo operations including addition, subtraction, multiplication, division, and modulo[2]. Also, MLPs are a general type of neural networks capable of learning through backpropagation (Rumelhart, Hinton, & Williams, 1986). Based on these properties, we applied the MLP model to help better understand problem difficulty in arithmetic/modulo operations. In order to measure MLP's problem difficulty, we used *conquest epoch*, which is the number of epochs taken by a model to correctly learn a given problem set. We propose this empirical measure since complex nonlinear mappings from harder problems to their correct answers tend to require more epochs than easier problems. In this regard, the conquest epoch can be used to measure the difficulty of learning and solving a particular problem set by MLPs.

Previous studies used one or two arithmetic operators to study problem difficulty. In contrast, the present study investigated problem difficulty across five arithmetic operators[3] — addition, subtraction, multiplication, division, and modulo. The present study also examined problem difficulty across the number of carries for each operator. This provides a more complete view of the impact of both arithmetic operators and carries on problem difficulty. Furthermore, as far as we know, the present study is the first to investigate the impact of operators and carries on problem difficulty in the context of both humans and connectionist models.

Datasets

Operation Datasets For each operator, we constructed an *operation dataset*, containing all possible operations between two 4-digit binary nonnegative integers (ranging $[0, 2^4 - 1]$) that generate nonnegative results. The dataset consists of $(\mathbf{x}, \mathbf{y})$ where $\mathbf{x}$ is an 8-dimensional input vector that is a concatenation of the two operands, and $\mathbf{y}$ is an 8-dimensional output vector that is the result of computing the operands. In Table 1, 'Total' is the number of pairs in each operation dataset. Let us simply refer to, for example, the operation dataset of subtraction as the *subtraction dataset*, and problems from the subtraction dataset as *subtraction problems*. The subtraction dataset is nearly half the size of any other dataset because only problems satisfying $a - b \geq 0$ were included. In the case of division and modulo, the dataset size is $240 = 2^8 - 2^4$ because $a \div b$ where $b = 0$ were excluded. The entirety of these operation datasets was used to train MLPs.

Table 1: Operation and carry datasets. One operation dataset exists for each operator, and this dataset is subdivided into carry datasets.

	Operation datasets			
# Carries (n)	$+$	$-$	$\times$	$\div$, mod
0	81	81	161	214
1	54	27	11	13
2	52	19	17	9
3	42	9	20	4
4	27		29	
5			5	
6			4	
8			8	
12			1	
Total	256	136	256	240
Carry datasets	5	4	9	4

Carry Datasets Operation datasets were further subdivided into carry datasets. A *carry dataset* refers to the total set of problems requiring a specific number of carries to solve correctly, for a given operator. With n denoting the number of carries required to correctly solve a problem, multiplication has 9 possible n. Hence, multiplication has 9 carry datatsets. The number of carry datasets for the other operators are shown in Table 1. Let us simply refer to the carry dataset involving n carries as the *n-carry dataset*, and problems from the n-carry dataset as *n-carry problems*.

Experiment 1: Humans

Participants

153 undergraduate students (89 men, 64 women) from various departments completed the experiment for course credit. The average age of participants was 21.3 ($SD = 1.8$).

Materials

Problem Sets A *problem set* for a specific operator was given to participants. Problems in a problem set were evenly distributed across carry datasets so that participants answered equal numbers of questions from each carry dataset. Question distributions per problem set were as follows: addition – 50 problems across 5 carry datasets; subtraction – 40 problems across 4 carry datasets; multiplication – 45 problems across 9 carry datasets; division – 40 problems across 4 carry datasets;

[2]The present study refers to the modulo operation as modulo.

[3]Strictly speaking, modulo is not an arithmetic operator; however, for simplicity, the present study assumes there are five arithmetic operators including modulo.

modulo – 40 problems across 4 carry datasets. Arithmetic problems were randomly sampled from each carry dataset without replacement; let us refer to a set of problems sampled from an n-carry dataset as an *n-carry problem set*. Sampling without replacement prevented participants from answering previously seen problems. However, in rare cases where the number of problems in a specific carry problem set were insufficient[4], participants were presented with the same problem multiple times. Each participant was given a unique randomly sampled problem set. In a given problem, two operands were given in a fixed 4-digit format (Figure 2). This was done in order to control for the extraneous influence of the number of operand digits on problem difficulty, as outlined by criterion (b) in the introduction.

Calculation Guidelines Calculation guidelines were prepared for participants because of their unfamiliarity with the binary system. The guidelines first explained the concept of binary numbers, followed by guiding examples with detailed step-by-step calculations, based on the right-to-left standard algorithm (Wu, 2011). Guiding examples (Figure 1) for each operator were organized as follows: addition – 2 addition problems; subtraction – 2 subtraction problems; multiplication – 1 multiplication problem with 2 addition problems; division – 1 division problems with 2 subtraction problems; modulo – 1 modulo problem with 2 subtraction problems. More than one carry was involved in all guiding examples so that participants grasped the mechanism of carry operations.

Procedure

Participants were randomly assigned to a subset of problems pertaining to one of the five operators; 30 students were assigned to addition, 30 to subtraction, 33 to multiplication, 30 to division, and 30 to modulo. Participants studied detailed calculation guidelines containing one or two guiding examples (Figure 1) for a given operator until they fully understood the given operator. Participants then began the experiment, solving problems through the command line interface (Figure 2). The use of pen and paper was permitted to assist in problem solving. After solving each problem, the true answer was displayed (Figure 2) in order to help participants understand their mistakes and perform more accurately for subsequent problems.

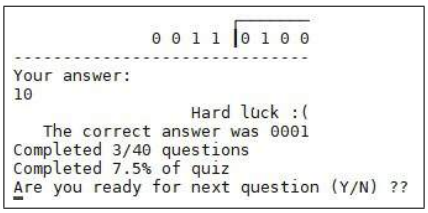


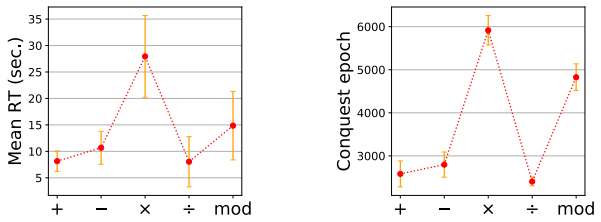
Figure 2: Sample program output.

[4]This was the case for the multiplication 6-carry and 12-carry problem sets, the subtraction 3-carry problem set, and the division/modulo 2-carry and 3-carry problem sets.

Results

If a participant provided a correct answer for a problem, it is reasonable to assume that this participant performed the correct number of carries to arrive at that answer. As such, only RTs for correct answers were used in Experiment 1. Data and detailed analytical results are available in the footnoted repository[5].

Response Time by Operator Each participant's mean RTs across all five operators were analyzed. Let us denote the mean RT for a problem set of operator $*$ as RT^*. Analysis of Variance (ANOVA) was used to investigate differences in RT^* across the five operators $* \in \{+, -, \times, \div, mod\}$. ANOVA showed significant differences between all RT^* [$F(4,148) = 78.65$, $p < .001$, $\eta^2 = .68$]. Further, a post hoc analysis was performed to analyze comparisons between all RT^*. The results of the Games-Howell post hoc test can be denoted by using the following notation[6]: $RT^\times > RT^+$, $RT^\times > RT^-$, $RT^\times > RT^\div$, $RT^\times > RT^{mod}$, $RT^{mod} > RT^+$, $RT^{mod} > RT^\div$ [$p < .001$], $RT^{mod} > RT^-$ [$p < .05$], $RT^+ < RT^-$ [$p < .01$], but $RT^+ \approx RT^\div$, $RT^- \approx RT^\div$ [$p > .05$]. These results can be summarized as: $RT^+ \lessapprox RT^\div \lessapprox RT^- < RT^{mod} < RT^\times$ (Figure 3a).



(a) Humans: Mean RT. (b) MLPs: Conquest epoch.

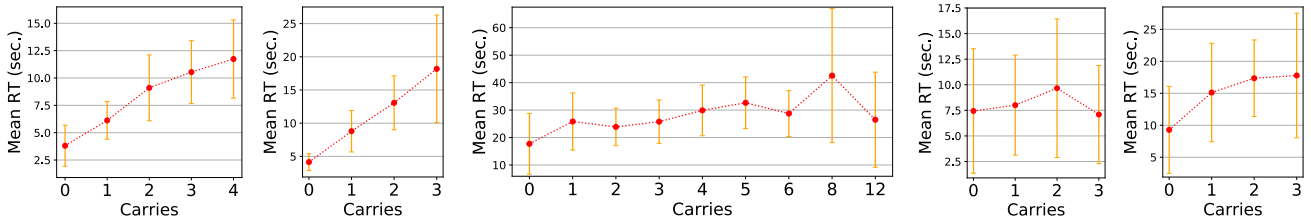
Figure 3: Problem difficulty by operator. The error bars are $\pm 1SD$.

Response Time by Carries Each participant's mean RTs across carry problem sets were analyzed. Let us denote the mean RT for an n-carry problem set of operator $*$ as RT^*_n.

Addition had 5 types of n-carry problems, $n \in \{0,1,2,3,4\}$. ANOVA showed significant differences between all RT^+_n [$F(4,145) = 43.45$, $p < .001$, $\eta^2 = .55$]. The Games-Howell post hoc test revealed that $RT^+_0 < RT^+_1 < RT^+_2$, $RT^+_0 < RT^+_3$, $RT^+_0 < RT^+_4$, $RT^+_1 < RT^+_3$, $RT^+_1 < RT^+_4$ [$p < .001$], $RT^+_2 < RT^+_4$ [$p < .05$], but $RT^+_2 \approx RT^+_3$, $RT^+_3 \approx RT^+_4$ [$p > .05$]. These results can be summarized as: $RT^+_0 < RT^+_1 < RT^+_2 \lessapprox RT^+_3 \lessapprox RT^+_4$ (Figure 4a). As such, for

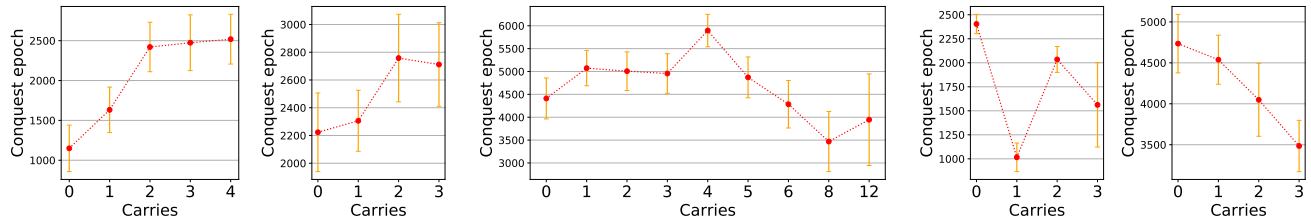
[5]https://github.com/sungjae-cho/cogsci2019-appendix/tree/master/human

[6]$A \approx B$ denotes $E[A]$ and $E[B]$ are not significantly different [$p > .05$]. $A < B$ and $B > A$ denote A and B are significantly different [$p < .05$] and their expectations hold $E[A] < E[B]$. $A \lessapprox B \lessapprox C \lessapprox D$ represents $A \approx B$ but A is less than any other right-hand operand (C, D); namely, $A < C$ and $A < D$. Likewise, concerning D, it indicates $C \approx D$, $A < D$ and $B < D$.



(a) Addition. (b) Subtraction. (c) Multiplication. (d) Division. (e) Modulo.

Figure 4: Humans: Mean RT by carries. The error bars are $\pm 1SD$.



(a) Addition. (b) Subtraction. (c) Multiplication. (d) Division. (e) Modulo.

Figure 5: MLPs: Conquest epoch by carries. The error bars are $\pm 1SD$.

$n \in [0,2]$, RT_n^+ was strictly increasing[7], but for all n, RT_n^+ was monotonically increasing[8].

Subtraction had 4 types of n-carry problems, $n \in \{0,1,2,3\}$. ANOVA showed significant differences between all RT_n^- [$F(3,116) = 46.07$, $p < .001$, $\eta^2 = .54$]. The Games-Howell post hoc test revealed all pairs of RT_n^- had significant differences. More specifically, $RT_0^- < RT_1^- < RT_2^-$, $RT_0^- < RT_1^- < RT_3^-$ [$p < .001$], $RT_2^- < RT_3^-$ [$p < .05$]. The results can be summarized as follows: $RT_0^- < RT_1^- < RT_2^- < RT_3^-$ (Figure 4b). Therefore, RT_n^- was strictly increasing with respect to the number of carries n.

Multiplication had 9 types of n-carry problems, $n \in \{0,1,2,3,4,5,6,8,12\}$. ANOVA showed significant differences between all $RT_n^\times$ [$F(8,284) = 9.24$, $p < .001$, $\eta^2 = .21$]. The results of the Games-Howell post hoc test can be summarized as follows: $RT_0^\times < RT_3^\times$ [$p < .05$], $RT_0^\times < RT_6^\times$ [$p < .01$], $RT_0^\times < RT_4^\times$, $RT_0^\times < RT_5^\times$, $RT_0^\times < RT_8^\times$ [$p < .001$], $RT_1^\times < RT_8^\times$ [$p < .05$], $RT_2^\times < RT_5^\times$, $RT_2^\times < RT_8^\times$ [$p < .01$], $RT_3^\times < RT_8^\times$ [$p < .05$]; there were no significant difference between the remaining pairs $RT_n^\times$, and only $RT_{12}^\times$ was not significantly different from any other $RT_n^\times$. These results can be summarized as: for $n \in [0,8]$, $RT_n^\times$ was monotonically increasing (Figure 4c).

Division had 4 types of n-carry problems, $n \in \{0,1,2,3\}$. ANOVA showed no significant differences between all $RT_n^\div$ [$F(3,116) = 1.20$, $p > .05$, $\eta^2 = .03$]. These results can be summarized as: $RT_0^\div \approx RT_1^\div \approx RT_2^\div \approx RT_3^\div$. Despite no significant difference between any $RT_n^\div$, a weak monotonically increasing trend in mean RT was observable (Figure 4d).

Modulo had 4 types of n-carry problems, $n \in \{0,1,2,3\}$. ANOVA showed significant differences between all RT_n^{mod} [$F(3,116) = 7.78$, $p < .001$, $\eta^2 = .17$]. The Tukey HSD post hoc test revealed that only RT_0^{mod} had significant differences from any other RT_n^{mod}. Specifically, $RT_0^{mod} < RT_1^{mod}$ [$p < .05$], $RT_0^{mod} < RT_2^{mod}$, $RT_0^{mod} < RT_3^{mod}$ [$p < .001$]. These results can be summarized as: $RT_0^{mod} < RT_1^{mod} \approx RT_2^{mod} \approx RT_3^{mod}$ (Figure 4e). RT_n^{mod} was monotonically increasing.

Experiment 2: Connectionist Models

Model

3000 MLPs (Figure 6) were trained for each operator. An 8-dimensional input vector comprised of two concatenated 4-digit operands was fed to the MLP. The MLPs had only one 2^6-unit hidden layer with sigmoid. An 8-dimensional output

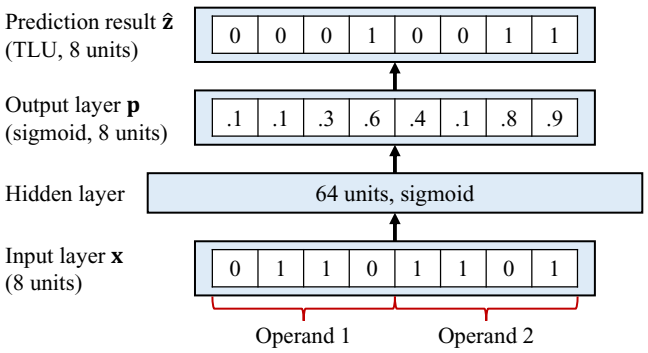


Figure 6: The structure of the multilayer perceptron. The model above predicts that $110 + 1101$ is equal to 10011.

[7]For every x and x' such that $x < x'$, if $f(x) < f(x')$, then we say f is *strictly increasing*.

[8]For every x and x' such that $x < x'$, if $f(x) \leq f(x')$, then we say f is *monotonically increasing*.

vector with sigmoid was used in order to match the maximum digit output of the arithmetic results. The predicted result is acquired by processing the output layer through the threshold logic unit (TLU), which transforms output numbers to 1 if they are greater than 0.5, and to 0 otherwise.

Training Settings

MLPs learned arithmetic operations by using backpropagation (Rumelhart et al., 1986) and a stochastic gradient method (Bottou, 1998) called Adam optimization (Kingma & Ba, 2015) with settings: $\alpha = .001$, $\beta_1 = .9$, $\beta_2 = .999$, $\varepsilon = 10^{-8}$. An entire operation dataset was utilized as a training set to train an MLP. For each epoch, a 32-sized mini-batch was randomly sampled without replacement (Shamir, 2016) from the training set. The weight matrix $W^{[l]}$ in layer l was initialized to samples from the truncated normal distribution ranging $[-1/\sqrt{n^{[l-1]}}, 1/\sqrt{n^{[l-1]}}]$ where $n^{[l]}$ is the number of units in the l-th layer; all bias vectors $b^{[l]}$ were initialized to 0. The objective function was the sum of the cross-entropy H between the true result $\mathbf{z}(\mathbf{x})$ and output activation vector $\mathbf{p}(\mathbf{x})$ where $\mathbf{x}$ is an input vector from a mini-batch: $\sum_{\mathbf{x}} H(\mathbf{z},\mathbf{p}) = \sum_{\mathbf{x}} [-\mathbf{z}(\mathbf{x}) \cdot \log(\mathbf{p}(\mathbf{x})) - (1-\mathbf{z}(\mathbf{x})) \cdot \{1 - \log(\mathbf{p}(\mathbf{x}))\}]$.

For every epoch, accuracy was evaluated on the total operation dataset and each carry dataset (Table 1). When 100% accuracy for a carry or operation dataset was attained, the current number of epochs was recorded as the conquest epoch of the dataset. Training was stopped when 100% accuracy for the operation dataset was reached.

Results

Data and detailed analytical results are available in the footnoted repository[9].

Conquest Epoch by Operator The conquest epochs of MLPs across operation datasets were analyzed. Let us denote the conquest epoch for the operation dataset of operator $* \in \{+,-,\times,\div,mod\}$ as e^*. ANOVA showed significant differences between all e^* [$F(4,14995) = 92838.78$, $p < .001$, $\eta^2 = .96$]. The Games-Howell post hoc test revealed that the differences between all pairs of e^* were significant [$p < .001$]. More specifically, these results can be summarized as: $e^{\div} < e^{+} < e^{-} < e^{mod} < e^{\times}$ (Figure 3b). This mirrors the ordering of the three highest mean RT in humans, as seen in Experiment 1.

Conquest Epoch by Carries The conquest epochs of MLPs across carry datasets were analyzed for each operator. Let us denote the conquest epoch of the n-carry dataset for operator $*$ as e^*_n.

Addition had 5 carry datasets, $n \in \{0,1,2,3,4\}$. ANOVA showed significant differences between all e^+_n [$F(4,14995) = 11835.66$, $p < .001$, $\eta^2 = .76$]. The Games-Howell post hoc test revealed that all pairs of e^+_n were significantly different [$p < .001$]. These results can be summarized as: $e^+_0 < e^+_1 < e^+_2 < e^+_3 < e^+_4$ (Figure 5a). As such, the conquest epoch e^+_n was strictly increasing with respect to n. Again, this mirrors results of RT^+_n from Experiment 1.

Subtraction had 4 carry datasets, $n \in \{0,1,2,3\}$. ANOVA showed significant differences among all e^-_n [$F(3,11996) = 2831.77$, $p < .001$, $\eta^2 = .41$]. The Games-Howell post hoc test revealed that all pairs of e^-_n were significantly different [$p < .001$]. These results can be summarized as: $e^-_0 < e^-_1 < e^-_3 < e^-_2$ (Figure 5b). Therefore, the conquest epoch e^-_n was strictly increasing for both $n \in \{0,1,2\}$ and $n \in \{0,1,3\}$.

Multiplication had 9 carry datasets, $n \in \{0,1,2,3,4,5,6,8,12\}$. ANOVA showed significant differences between all $e^\times_n$ [$F(8,26991) = 5024.17$, $p < .001$, $\eta^2 = .60$]. The Games-Howell post hoc test revealed that differences between all pairs of $e^\times_n$ were significant [$p < .001$]. Specifically, the results can be summarized as: $e^\times_8 < e^\times_{12} < e^\times_6 < e^\times_0 < e^\times_5 < e^\times_3 < e^\times_2 < e^\times_1 < e^\times_4$ (Figure 5c).

Division had 4 carry datasets, $n \in \{0,1,2,3\}$. ANOVA showed significant differences between all $e^\div_n$ [$F(3,11996) = 17788.62$, $p < .001$, $\eta^2 = .82$]. The Games-Howell post hoc test revealed that all pairs of $e^\div_n$ had significant differences [$p < .001$]. More specifically, the results can be summarized as follows: $e^\div_1 < e^\div_3 < e^\div_2 < e^\div_0$ (Figure 5d). Thus, the conquest epoch $e^\div_n$ was not increasing with respect to n.

Modulo had 4 carry datasets, $n \in \{0,1,2,3\}$. ANOVA showed significant differences between all e^{mod}_n [$F(3,11996) = 7281.45$, $p < .001$, $\eta^2 = .65$]. The Games-Howell post hoc test revealed that all pairs of $e^\div_n$ had significant differences [$p < .001$]. The results can be summarized as follows: $e^{mod}_0 > e^{mod}_1 > e^{mod}_2 > e^{mod}_3$ (Figure 5e). Hence, the conquest epoch e^{mod}_n was strictly decreasing with respect to n.

Discussion and Conclusion

Experiment 1 Results of the present study demonstrate how problem difficulty varies depending on the five arithmetic operators and the number of carries. In Experiment 1, results showed that for the five operators, problem difficulty was monotonically increasing with respect to the number of carries (Figure 4). Notably, for subtraction, RT was strictly increasing (Figure 4b). Another notable result was that RT for multiplication was the highest among the five operators (Figure 3a). In order to successfully perform multiplication, several sub-multiplication steps must first be completed (e.g. $1011 \times 1 = 1011$, $1011 \times 0 = 0$, see Figure 1). A participant may have to complete as many as 4-operand addition steps in order to correctly solve a single multiplication problem (Figure 1). It has been shown that the number of steps (DeStefano & LeFevre, 2004) and operands (Seitz & Schumann-Hengsteler, 2000, 2002) involved in arithmetic problems increases working memory demands. As such, the additional arithmetic steps involved in multiplication problems may have led to multiplication having the highest RT among the five operators. It is worth highlighting that participants solved the same 12-carry problem five times, due to

[9] https://github.com/sungjae-cho/cogsci2019-appendix/tree/master/mlp

there being only one problem in the 12-carry dataset (Table 1). This problem repetition may be responsible for the decreased RTs seen in the 12-carry problem set, relative to the 8-carry problem set (Figure 4c). As such, it is not valid to compare RT of the 12-carry problem set to other carry problem sets. Like multiplication, modulo problems also require many sub-operations to solve correctly. This may explain why modulo had substantially higher RT than addition, subtraction, and division (Figure 3a). Within the modulo problem set, RT for the 0-carry problems was significantly less than RT for problems involving carries (Figure 4e). However, no significant difference was found in RT between any pair of problem sets involving carries. Modulo involves the use of arithmetic sub-operations in order to correctly answer problems (Figure 1). However, unlike in multiplication, the subtraction sub-operations involved in modulo problems showed consistent patterns. The second operand of the sub-operations was always equivalent to either 0 or the denominator (e.g. 11, see Figure 1). These patterns may have lowered RT for higher n-carry datasets (Figure 4e). For division, even if a given problem was an n-carry problem, it did not necessarily involve n carries, as the final subtraction sub-operation may have been unnecessary in solving the problem (Figure 1). This may have meant that the number of carries in a division problem did not always impact on RT (Figure 4d).

Experiment 2 Experiment 2 found that problem difficulty (conquest epoch) for addition, subtraction, and division was substantially less than problem difficulty for multiplication and modulo. Addition, subtraction, and division may have been easier for MLPs to learn than the other two operators, due to the repeated occurrence of digit patterns in these problems. This implies MLPs learned multiplication and modulo problems by memorizing each problem, rather than by finding digit patterns. This experiment also found that addition problem difficulty for MLPs was strictly increasing with respect to the number of carries involved in a problem (Figure 5a). However, no such increase was seen in the other operators (Figure 5b, 5c, 5d, 5e). Generally, MLPs are sensitive to statistical properties of experience, such as the frequency and typicality of patterns they meet while they learn (Rumelhart & McClelland, 1986). In this regard, MLPs appear to require more epochs to conquer datasets that contain lots of infrequent and atypical patterns. However, the frequency and typicality of patterns in our datasets does not offer a satisfactory explanation as to why an increasing relationship between problem difficulty and carries was observed in addition, but not for the other operators.

Experiments 1 & 2 Comparing Experiment 1 with Experiment 2, humans and the MLPs showed partial similarities in their solving of binary arithmetic problems. For both humans and the MLPs, problem difficulty was highest for multiplication, followed by modulo and then subtraction. (Figure 3). Addition problem difficulty for both humans and MLPs showed increasing trends as a function of the number of carries (Figure 4a, 5a). However, the trajectories of these increases followed notably different paths (Figure 4a, 5a).

Contributions The present study makes four notable contributions to the current literature on mathematical cognition and cognitive science: Firstly, the present study compares problem difficulty across the five operators. This contrasts with preceding work, which has generally dealt with three or fewer operators. Furthermore, to the best of our knowledge, the present study is the first to investigate problem difficulty with regards to the modulo operation. Secondly, the present study for humans showed that the number of carries had a discernible effect on problem difficulty across four of the five arithmetic operators: addition, subtraction, multiplication, and modulo. Thirdly, the use of the binary numeral system allowed the present study to somewhat control for other criteria that may have impacted problem difficulty. These criteria include the problem size effect and over-familiarity with the decimal numeral system. This allowed for a targeted investigation into the effect of carries on problem difficulty. Finally, the present study found that MLPs experienced problem difficulty for some operators similarly to humans: For both humans and MLPs, problem difficulty was highest for multiplication, followed by modulo and then subtraction (Figure 3a, 3b). Also, the effect of carries on problem difficulty in addition problems showed increasing trajectories for both agents (Figure 4a, 5a). This supports previous research (McClelland et al., 2016) suggesting that there may be some similar cognitive processes underlying mathematical cognition in both humans and connectionist models.

Future Study Future studies should aim to uncover what underlying mechanisms caused the MLPs to experience relative problem difficulty similarly to humans across the five operators. Also, the internal representations MLPs use to perform arithmetic operations could be investigated. However, MLPs do not have the innate ability to dynamically process information as humans do. MLPs always take a fixed number of computational steps to produce answers, while humans take a variable amount of time to produce answers. One direction for future work could introduce a new dynamic connectionist model to learn arithmetic, namely, a recurrent network such as the Elman network (Elman, 1990) or the Jordan network (Jordan, 1997). Such recurrent networks can produce answers through variable computational steps depending on the problem. These variable steps can be directly compared to humans' RT, providing a more valid comparison to human arithmetic cognition.

Acknowledgments

We thank Prof. Sungryong Koh for comments on the psychological experiment; Kyoung-Woon On, Gi-Cheon Kang, Taehyeong Kim, Chung-Yeon Lee and Joonho Kim for discussion; Paula Higgins, Seung Hee Yang, and Woosuk Choi for writing comments. This work was partly supported by the Institute for Information & Communications Technology Pro-

motion (R0126-16-1072-SW.StarLab, 2017-0-01772-VTT, 2018-0-00622-RMI, 2019-0-01367-BabyMind) and Korea Evaluation Institute of Industrial Technology (10060086-RISF) grant funded by the Korea government (MSIP, DAPA).

References

Ashcraft, M. H. (1992). Cognitive arithmetic: A review of data and theory. *Cognition*, *44*(1-2), 75–106.

Ashcraft, M. H. (1995). Cognitive psychology and simple arithmetic: A review and summary of new directions. *Mathematical cognition*, *1*(1), 3–34.

Bottou, L. (1998). *Online algorithms and stochastic approximations.* Cambridge University Press.

Campbell, J. I. (1994). Architectures for numerical cognition. *Cognition*, *53*(1), 1–44.

Campbell, J. I. (2005). *Handbook of mathematical cognition*. Psychology Press.

DeStefano, D., & LeFevre, J.-A. (2004). The role of working memory in mental arithmetic. *European Journal of Cognitive Psychology*, *16*(3), 353–386.

Elman, J. L. (1990). Finding structure in time. *Cognitive science*, *14*(2), 179–211.

Franco, L., & Cannas, S. A. (1998). Solving arithmetic problems using feed-forward neural networks. *Neurocomputing*, *18*(1), 61–79.

Fürst, A. J., & Hitch, G. J. (2000). Separate roles for executive and phonological components of working memory in mental arithmetic. *Memory & cognition*, *28*(5), 774–782.

Hornik, K., Stinchcombe, M. B., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*(5), 359–366.

Hoshen, Y., & Peleg, S. (2016). Visual learning of arithmetic operation. In *Proceedings of the thirtieth AAAI conference on artificial intelligence* (pp. 3733–3739).

Imbo, I., Vandierendonck, A., & Vergauwe, E. (2007). The role of working memory in carrying and borrowing. *Psychological Research*, *71*(4), 467–483.

Jordan, M. I. (1997). Serial order: A parallel distributed processing approach. In *Advances in psychology* (Vol. 121, pp. 471–495).

Kaiser, L., & Sutskever, I. (2016). Neural GPUs learn algorithms. In *4th international conference on learning representations, conference track proceedings.*

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *3rd international conference on learning representations, conference track proceedings.*

LeFevre, J.-A., Bisanz, J., Daley, K. E., Buffone, L., Greenham, S. L., & Sadesky, G. S. (1996). Multiple routes to solution of single-digit multiplication problems. *Journal of Experimental Psychology: General*, *125*(3), 284.

McClelland, J. L., Mickey, K., Hansen, S., Yuan, A., & Lu, Q. (2016). A parallel-distributed processing approach to mathematical cognition. *Manuscript, Stanford University*.

McCloskey, M., & Lindemann, A. M. (1992). MATHNET: Preliminary results from a distributed model of arithmetic fact retrieval. In *Advances in psychology* (Vol. 91, pp. 365–409). Elsevier.

Miller, K., Perlmutter, M., & Keating, D. (1984). Cognitive arithmetic: Comparison of operations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(1), 46.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*, 533-536.

Rumelhart, D. E., & McClelland, J. L. (1986). On learning the past tense of English verbs. In *Parallel distributed processing* (Vol. 2, pp. 216–271). MIT Press.

Rumelhart, D. E., & Mcclelland, J. L. (1986). *Parallel distributed processing* (Vol. 1). MIT Press.

Seitz, K., & Schumann-Hengsteler, R. (2000). Mental multiplication and working memory. *European Journal of Cognitive Psychology*, *12*(4), 552–570.

Seitz, K., & Schumann-Hengsteler, R. (2002). Phonological loop and central executive processes in mental addition and multiplication. *Psychological Test and Assessment Modeling*, *44*(2), 275.

Shamir, O. (2016). Without-replacement sampling for stochastic gradient methods. In *Advances in neural information processing systems 29: NIPS 2016* (pp. 46–54).

Wu, H. (2011). The standard algorithms. In H. Wu (Ed.), *Understanding numbers in elementary school mathematics* (pp. 57–60). American Mathematical Society.

Observing child-led exploration improves parents' causal inferences

Koeun Choi[1,2](koeun@vt.edu)
Milagros Grados[2] (milagrosgrados95@gmail.com)
Elizabeth Bonawitz[2](lbaraff@gmail.com)
[1] Department of Human Development and Family Science, Virginia Tech, VA 24061
[2] Department of Psychology, Rutgers University, Newark, NJ 07102

Abstract

Do children's flexible causal inferences promote more creative causal discovery for observing adults? Inspired by a task in which children are more likely to consider unconventional causal forms (Lucas, Bridgers, Griffiths, & Gopnik, 2014; Wente et al., 2019), we designed a new method in which child-adult pairs work together to solve a causal task and assessed the relative influence of each member of the pair on the other's causal inference. Consistent with previous research, children were better than parents at learning the unusual conjunctive relationship, suggesting that children make more flexible causal inferences than adults. Our research also revealed a surprising and new result – that observing a child explore broadly helped parents to be more flexible and open-minded in their causal learning. In contrast, a child observing an adult's exploratory interventions had no negative consequence on the child's ability to infer the correct relation. Follow-up experiments explored the degree to which this child-led bootstrapping for adults was due to the particular exploratory evidence generated by the child during play, or merely the presence of a child. Results suggest that both factors may play a role in shaping adult's causal inferences.

Keywords: causality, cognitive development, parent-child interaction

Introduction

Like scientists, children explore, discover, and learn. Those of us with the good fortune to spend ample time with these little scientists can't help but be inspired by their curiosity and reminded of our own creative and inquisitive pasts. As Gopnik (2016) has suggested, childhood may be a unique time for greater exploration, cognitive flexibility, and creativity, leading to innovation for our species driven by our youngest. Of course, much research has focused on how children learn from adults, but perhaps there are cases when adults can learn from these innovative explorers. Perhaps there are cases when the flexible minds of children lead to knowledge and learning when adults lack.

Indeed, evidence from several research studies indicated that children learn specific and abstract causal structure and sometimes do so more readily than adults (Gopnik et al., 2017; Lucas et al., 2014; Wente et al., 2019). These findings suggest that children may be more open to new possibilities and willing to consider different hypotheses than adults (Gopnik et al., 2017). Often times, children encounter and explore new information in the presence of adults who may hold contrasting ideas about the world. Here we ask, how do children and adults interact with each other to explore and come to understand the world around them? In this study, using a new method in which child-adult pairs work together to solve a causal task, we look at whether exploratory patterns differ between children and adults and the extent to which these differences have consequences for causal inference in the observers.

Young children's ability to infer abstract causal principles has been studied using the forms of overhypotheses including conjunctive and disjunctive causal relationships. An overhypothesis is a broad framework that constrains the range of hypothesis learners consider (Goodman, 1955; Griffiths & Tenenbaum, 2007). A conjunctive causal relation is a functional form in which multiple causes jointly produced an effect; a disjunctive relation is a functional form in which a single cause can bring out an outcome independently (e.g., see Cheng, 1997). These overhypotheses are not bounded to a particular context but are applicable to many other scenarios, and having these assumptions shape future learning by limiting the number of possible hypotheses that are considered.

Prior research has revealed developmental differences in inferring a certain form of overhypotheses (Lucas et al., 2014). After having the same amount of exposure to evidence that is statistically best explained by (the unconventional) conjunctive causal form, children outperformed adults by correctly generalizing the conjunctive causal relationship to new objects. While both adults and children were successful at inferring a disjunctive form, the ability to infer conjunctive forms appears to be decreased with age. When given evidence that supported a conjunctive form, adults instead maintained a disjunctive relationship (Lucas et al., 2014). The developmental differences in learning of the conjunctive (but not disjunctive) causal form suggest that young children are more flexible than adults in incorporating evidence to guide future learning (See also Gopnik et al., 2017; Wente et al., 2019; Gopnik, Griffiths, & Lucas, 2015).

In these past studies, participant's ability to infer abstract causal forms was tested by asking for judgments about the causal efficacy of each cause or to use potential causes to produce an outcome. However, in these studies, participants were not given the opportunity to explore and generate their own evidence. Thus, it remains an open question whether adult and child participants will generate different patterns of exploration when given the opportunity to test out possible causal forms.

One concern is whether children will be able to generate meaningful play at all. However, recent findings are supporting the claim that young children may be more competent and capable explorers than previously believed. For example, children shape their explorations to conduct inter-

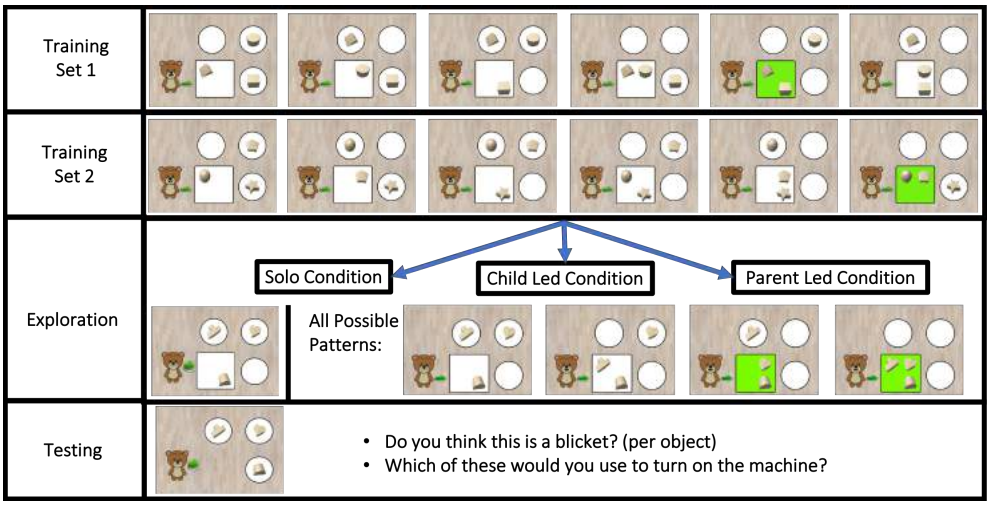


Figure 1: Stimuli and task procedure; During two sets of training, parent-child pairs watched experimenter demonstrating evidence in favor of a conjunctive causal form. Following the training, parent-child pairs were randomly assigned to solo, child-led, and parent-led conditions. In the solo condition, the parent and child each explored the set of testing objects. In the child-led group, the child explored the testing objects while parent watched. In the parent-led group, the parent explored the testing objects while the child watched. Finally, the experimenter asked the parent and child individually to judge whether each object was a blicket and to turn on the machine.

ventions to deconfound variables (Schulz & Bonawitz, 2007; Schulz, Kushnir, & Gopnik, 2007; Schulz, 2012; Schulz, Gopnik, & Glymour, 2007). Further, children plan their future exploration based on the inference about pedagogical goals of teachers based on available information (Bonawitz et al., 2011; Eaves & Shafto, 2012; Gweon, Pelton, Konopka, & Schulz, 2014). These studies provide evidence for the claim that children's exploration is guided by the evidence. However, it remains an open question whether the evidence generated during children's and adults' explorations may differently reflect beliefs going into the task.

Prior studies revealed developmental differences in the conjunctive causal inference by examining child and adult groups individually. Despite the importance of caregiver-child interaction on play and development (Weisberg, Hirsh-Pasek, & Golinkoff, 2013; Fisher, Hirsh-Pasek, Newcombe, & Golinkoff, 2013; Honomichl & Chen, 2012), little is known about the impact of observing either child's or parent's patterns of exploration on one another. Thus, we were also interested in whether observing children's broad hypothesis search would promote more creative and flexible thinking in causal learning for observing adults. Of course, observing adults' exploration may also influence children's conjunctive causal learning. For example, instructions constrain children's explorations indicating that children are sensitive to inductive biases in their explorations (Bonawitz et al., 2011). Similarly, a body of literature on guided play highlights that it is critical for adults to scaffold learning goals as well as let children direct their exploration and discovery (Weisberg et al., 2013; Fisher et al., 2013). This balance between scaffolding and letting the child take the lead could be particularly important when children and adults hold different assumptions about the world.

In the present work, we examined the extent to which exploration patterns differed between children and adults and whether observing another's exploration shaped consequent learning. In Experiment 1, we examined children and their parents' exploration and learning following the exposure to evidence consistent with a conjunctive causal relationship. Critically, we manipulated who the actor was generating the evidence including a child-led condition, parent-led condition, and solo conditions for each group as controls. Consistent with prior research, we hypothesized that children would be more likely than adults to generalize a conjunctive causal relationship. However, we also predicted differential exploratory patterns for children and adults. We looked at whether these differences have consequences for learning in the observers. Specifically, observing parent-led exploration may restrict children's causal inference, thus resulting in more adult-like responses. On the other hand, observing child-led exploration may result in flexible learning in parents.

Experiment 1

Methods

Participants Seventy-two parent-child dyads were recruited from various settings (i.e., museum, home, and community event; Children: n = 72, 53% Female, M = 5.03, SD = 0.84, Range = 4.0-6.9 years; Parents: n = 72, 56% Mothers). The dyads were randomly assigned to the Solo (n = 24), Child-Led (n = 24), and Parent-Led (n = 24) conditions.

Additional two dyads were recruited but excluded due to not finishing the study ($n = 1$) or experimental errors ($n = 1$).

Stimuli and Apparatus Our procedure involved both real objects and interactive video stimuli. The interactive video stimuli were developed using jsPsych (De Leeuw, 2015) and displayed on a touch-screen tablet computer (10.1-in. Galaxy Tab; Samsung America, San Jose, CA). The video included images of three circles, a square, a green button, and a cartoon bear (see Figure 1). Images of objects were presented on the three circles, and each object moved to the square when tapped. The button was designed to test objects once they were placed on the square. In addition, to provide a way for participants to respond without influencing the listening other, two identical yes-no response sheets were created so participants could silently point to their response behind a barrier. The sheets included two rectangles (green and red), each includes a smiley or frowny face with "yes" or "no" written at the bottom, respectively.

Procedure Participants were tested in a quite place. The yes-no response sheets were placed in front of the participants. The experimenter asked a simple question about color (i.e., Is this white?) to both the child and the parent. If participants pointed to the wrong answer or responded verbally, the experimenter asked additional questions until children successfully responded using the yes-no response sheet.

Next, a backward blocking task was conducted (e.g. see Sobel, Tenenbaum, & Gopnik, 2004). This task was designed to acclimate the participants to an ambiguous causal reasoning task as well as familiarize participants with the instructions. The participants were introduced to a machine that detects "wugness". The experimenter explicitly stated that wugs are very rare and can not be judged solely by looking, but they possess wugness inside them. First, the experimenter placed two potential causes (Objects A and B) on the machine, which produced an outcome. Then the participants observed that the outcome occurred with the presence of only one of the causes (A). After observing these two events, the participants were asked to judge whether each object (A, B) was a wug, respectively.

Upon the completion of the backward blocking task, the experimenter introduced the tablet and stated that they would now play a completely different game (see Figure 1). The experimenter also mentioned that blickets are very rare and can not be judged just by looking, but have blicketness inside them. After a brief introduction to the features of the tablet game, the experimenter introduced an object (C) and ask if the participant thought that object as a blicket without any evidence; this allowed us to test participants' priors for the probability of an object being a blicket. Then the experimenter presented the first set of three training objects (D, E, F) which activated the machine according to a conjunctive causal rule. This was followed by another set of different training objects (D', E', F') that also provided evidence for a conjunctive rule, as in Lucas et al. (2014).

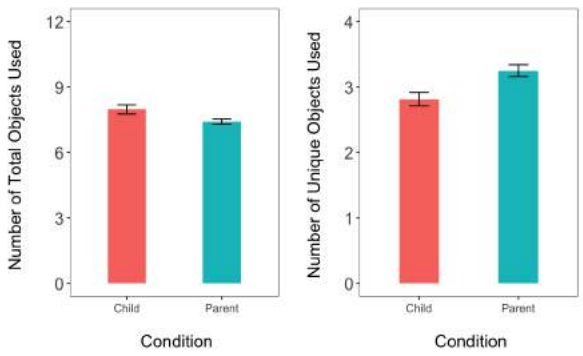


Figure 2: Differing exploratory patterns as measured by total object used (left) and unique actions (right) during the Exploration Phase. Compared to children, parents used more unique combinations of objects, but also fewer objects were tried on each trial. Error bars denote *SE*.

Following the two training trials, the experimenter introduced the new set of three testing objects (G, H, I) that the participants would have the opportunity to explore and test themselves. During this phase, one object (I) was permanently attached to the square. This was designed so that the evidence generated during the free intervention phase would maintain ambiguity. At the beginning of the exploration phase, the participants were told that the object (I) was stuck on the machine and that they can test the object (I) by itself or with the other objects. There were four exploration trials, and participants could choose one of the four possible options (I, GI, HI, GHI) in each trial. Our critical between-subjects design varied who controlled the interventions. In the solo condition, both parent and child had their own tablets, and could not see the screens or exploratory choices of the other. In the Child-Led condition, the child made all intervention choices while the parent watched. In the Parent-Led condition, the parent made all intervention choices while the child watched. Participants were given four intervention trials (a trial was counted once the participant depressed the test button); the intervention choices were recorded automatically with the tablet software. Next, the experimenter asked the parent and child to judge whether each object was a blicket. Lastly, as our critical test measure, the experimenter asked the parent and child individually to generate the effect using the objects ("Which of these objects would you use to turn on the machine?"). We coded whether two or more objects were used.

Results

We first assessed what kinds of interventions children and adults performed during the exploration phase. Results revealed that overall, children were more likely than adults to explore objects jointly, Welch $t(90.5) = 3.19$, $p = .001$ (Figure 2, left), suggesting that children may have been more amenable to the conjunctive rule as early as the exploration

phase. However, the quality of children's interventions was not strictly better than adults: parents tried more deconfounding causal explorations by testing more unique combinations of objects, $t(72.9)$= -2.36, p = .021 (Figure 2, right). Within each age group (child and adult, respectively), there was no significant difference between solo and joint groups, $p >$ .250.

Critically, we explored whether any particular group was more successful at generating the correct response in the final test phase. Overall, and replicating previous findings, children performed better on average than the adults, $\chi^2(1)$ = 20.39, $p < .001$. We conducted a logistic regression to predict the probability of selecting one or more objects to activate the machine as a function of condition. As shown in Figure 3, the parents in the Child-Led group were more likely to use multiple objects to activate the machine than those in the Parent-Led group, b = 1.44, p = .022, suggesting that observing evidence generated by children helped parents to be more flexible and exploratory in their own causal inferences. The probability of choosing multiple objects as blickets in the Parent-Solo condition did not differ from that in the Child-Led condition b = 0.84, p = .152, or the Parent-Led condition, b = 0.59, p = .353.

Consistent with previous research, children were better than parents at learning the unusual conjunctive relationship, suggesting that children make more flexible causal inferences than adults. Our research also revealed a surprising result – that merely observing a child's exploratory behavior may suffice to help parents to be more flexible and open-minded in their causal learning. Two possible explanations exist for this result. One possibility is simply that watching a child interact with the toy (regardless of the patterns of exploration) was sufficient to get adults in a childlike frame of reference, opening their mind to a broader set of hypotheses. Another possibility is that the particular evidence generated by children (which differed from adults) was critical in helping adults infer the conjunctive form. In Experiment 2, we explored these possibilities by having adults view a child actor perform interventions for all conditions. Critically we varied the particular interventions presented, yolking the evidence to the specific interventions attempted in the Child-Solo, Child-Led, and Parent-Led conditions of Experiment 1.

Experiment 2

Methods

Adults participants were recruited from Amazon Mechanical Turk. The final sample of 72 participants were randomly assigned to the Child-Solo-Exploration-Data (n = 24), Child-Led-Exploration-Data (n = 24), and Parent-Led-Exploration-Data (n = 24) conditions. An additional 10 participants were excluded from analysis because of the failure to pass an attention check question. Participants were paid $.75 for completing the 6-8 minute survey.

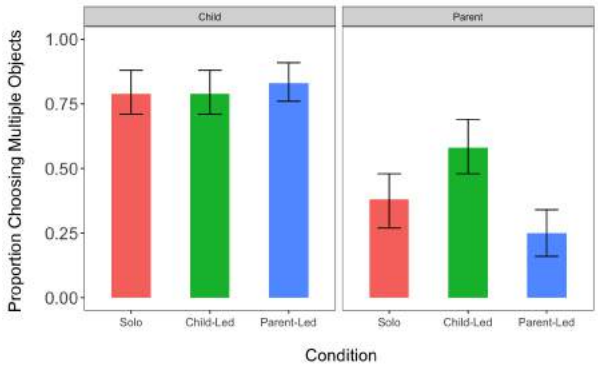


Figure 3: Proportion of participants who selected multiple objects on the final test trial by condition (Solo, Child-Led, Parent-Led). Children correctly attempted multiple objects to turn on the machine regardless of condition. However, adults were only more likely to test multiple objects in the Child-Led condition. Error bars denote *SE*.

Procedure

The stimuli and procedure was the same as that of Experiment 1 except the following differences. First, the data collection was conducted online; thus, the training phase was introduced with a series of screenshots of tablet games, and the participants were required to click a button to proceed. Second, for the exploration phase, the participants saw a video of a preschooler trying to figure out which objects are blickets using the tablet game. The same child actor generated different exploratory patterns, which was organized to match the patterns of exploration data generated from the three conditions (Child-Solo, Child-Led, Parent-Led) in Experiment 1. The preschooler was described to the participants as being a randomly selected example of a child exploring the toy.

Results

We used a logistic regression model as a function of condition (Child-Solo-Exploration-Data, Child-Led-Exploration-Data, Parent-Led-Exploration-Data) to predict the probability of using two or more objects to turn on the machine. There was a significant difference between the Child-Solo-Exploration-Data (67%) condition and the Parent-Led-Exploration-Data (38%) condition such that the group of participants who observed child-solo-exploration data showed a higher probability of using multiple objects to activate the machine compared to those who observed parent-led-exploration data, b = 1.20, p = .046. Unexpectedly, there was also a marginally significant difference between the Child-Solo-Exploration-Data (67%) and Child-Led-Exploration-Data condition (42%), b = 1.02, p = .085.

These results revealed that observing children's exploratory patterns based on broad, exploratory hypotheses supported adults' learning of an unconventional abstract causal form (at least in cases when data from the Child-Solo-Exploration-Data condition were observed). Experi-

ment 2 provides additional support for the idea that observing child-generated exploratory patterns increases the flexibility of adult's causal reasoning. It remains unclear whether the child-generated exploratory evidence alone would be sufficient to promote adults' causal reasoning, or whether it is this evidence *in conjunction* with a child-directed play that helps adults. Further, Experiment 2 was conducted via an online survey platform; thus, the findings may be limited in their generalization to adults in a live setting. To explore this further, in Experiment 3, we used an in-lab setting to conduct child-yoked interventions, but performed by adults. We focus our attention on the two critically different yolked-data conditions: Child-Led-Exploration-Data and Parent-Led-Exploration-Data.

Experiment 3

Methods

Forty-eight undergraduate students (M = 20.60, SD = 3.13, range: 18-31 years) were randomly assigned to the Child-Led-Exploration-Data (n = 24) or Parent-Led-Exploration-Data (n = 24) conditions. An additional 5 participants were excluded from analysis because of experimental errors in generating the data from the yolked trials.

Procedure

The stimuli and procedure was the same as to that of Experiment 1 except the following differences. First, the data collection was conducted in the lab. Second, each participant was paired with a confederate who secretly worked with the lab but who was introduced as another naive participant. During the exploration phase, the participant observed the confederate exploring the testing objects, as if choices were "in the moment" decisions. Instead, however, the confederate completed the four exploration trials as yolked to the data generated from the two conditions (Child-Led, Parent-Led) in Experiment 1.

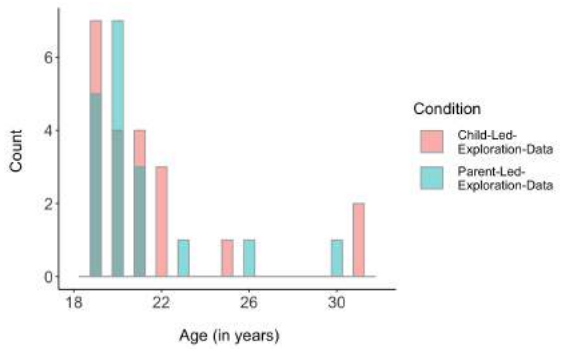


Figure 4: Histogram of age of participants for each condition (Child-Led-Exploration-Data, Parent-Led-Exploration-Data) in the sample. Age was ranged from 18 to 31 years.

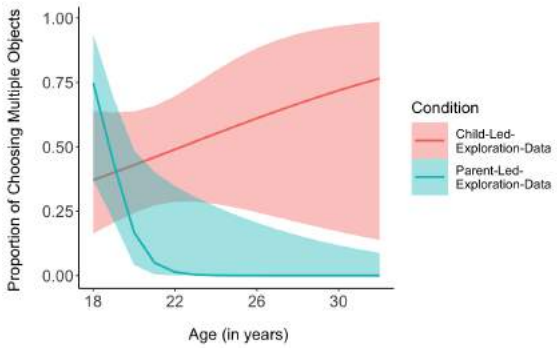


Figure 5: Proportion of participants who selected multiple objects for each condition (Child-Led-Exploration-Data, Parent-Led-Exploration-Data). Adults in the Child-Led-Exploration-Data condition were more likely to use two or more objects to turn on the machine than those in the Parent-Led-Exploration-Data condition. Shading indicates 95% confidence intervals.

Results

Comparing overall performance in terms of endorsement of two or more casual blocks between Child-Led-Exploration-Data and Parent-Led-Exploration-Data conditions revealed no overall differences, $\chi^2(1) = 0.35$, $p = .555$. This result is surprising, given the condition differences observed in Experiment 2. One possible explanation for this difference is that indeed the presence of the child generating the particular interventions was required to help adults consider the unlikely conjunctive form.

However, we also noticed that the age of the participants in our lab sample (ranging from 18-31 years; see Figure 4) significantly differed from the parents in Experiment 1 (ranging in mid-thirties to forties), and so we performed an unplanned exploratory analysis using a logistic regression model with age as a continuous variable and condition (Child-Led-Exploration-Data, Parent-Led-Exploration-Data) to predict the probability of using two or more objects to turn on the machine. In fact, there was a significant interaction between age and condition such that the group of participants who observed child-led-exploration data showed a higher probability of using multiple objects to activate the machine compared to those who observed parent-led-exploration data with increasing age, $b = 1.46$, $p = .018$ (see Figure 5). The pattern stayed the same for a narrower age range (18-24 years).

These results suggest that while child-yoked interventions may assist adults with causal form inferences, age may moderate this effect. Specifically, observing child-generated evidence was particularly helpful for the older participants of our sample.

Discussion

Consistent with previous research showing that adults are less likely to generate a conjunctive causal form than children, children were better than parents at generalizing the unusual

conjunctive form to their exploration and learning (Gopnik et al., 2017; Lucas et al., 2014; Wente et al., 2019). By examining exploration, we revealed that parents tried more deconfounding explorations than children. In contrast, children performed more interventions that involved multiple blocks, suggesting that children were engaging in hypothesis confirmation consistent with having inferred the conjunctive form from the previous training trails.

Strikingly, parents in the Child-Led group were more likely to generalize the conjunctive relationship than those in the Parent-Led group. Child-yoked interventions performed by either a child or adult similarly improved causal form inferences, suggesting that observing evidence generated by children may help adults to be more flexible in their own causal inferences.

In our study, young children generalized the unconventional conjunctive relationship to their exploration and learning regardless of whether the free play period was led by an adult or not. Of course, if children had already inferred the correct causal form from the initial training trials, than any intervention observed would continue to confirm children's overhypothesis because we designed the toy to produce outcomes consistent with the conjunctive form. In contrast, if adults had not yet inferred the correct form prior to the exploration phase, then observing their children repeatedly use multiple blocks to activate the machine may have been sufficient to raise the salience of this alternative hypothesis and facilitate learning.

The results from Experiments 2 and 3 suggest that adults' causal inferences can benefit from observing child-yoked explorations, especially when those exploratory patterns were generated by a child than an adult. However, as these two experiments were conducted in different environments (in-person vs. online), the contextual factor may have contributed to the differences. Thus, an important next step would be to test the effect of the age of the model who demonstrates child-led exploratory patterns in the same setting. Future work will examine whether watching a video of an adult demonstrator performing child-yoked interventions similarly improves adults' causal form inferences, controlling for the familiarity of the adults to the demonstrator. Further, future studies could explore the characteristics of adult observers such as age and experience working with young children.

The current findings show the importance of observing other's exploration when beliefs are in conflict with each other. Adults at least may be able to recognize the relationship between attempted interventions and considered hypotheses, raising awareness of hypotheses that were not previously considered. Such an account is consistent with the Wisdom of the Crowds (Vul & Pashler, 2008) or the adage that two heads are better than one. However, our results go one step further, suggesting that even observing the exploratory actions of another may help bootstrap inference to the best explanation.

More broadly, these findings support the importance of adult-child play, but with a surprising twist. Adults may benefit from play with children (rather than the other way around, as is often considered in the literature). Such work suggests the importance of giving children opportunities to lead their own exploration and discovery.

Parent-child joint play occurs in numerous settings involving both concrete and digital materials. In light of the pervasive interactive technology in young childrens everyday lives, it is important to understand how these tools can be used not only to transmit information but also support active exploration and discovery. This line of research can help us understand the ways in which parents and children conjunctively learn about the world.

Acknowledgments

We thank staff, parents, and children of Newark Museum for their participation and Cierra Clark, Kiabeth Guazhco, and Micah Connolly for data collection. This research has been supported by NSF-CNS-1623486 (EB), Jacobs Foundation Fellowship (EB), and NSF-SES 1627971 (EB).

References

Bonawitz, E. B., Shafto, P., Gweon, H., Goodman, N. D., Spelke, E., & Schulz, L. E. (2011). The double-edged sword of pedagogy: Instruction limits spontaneous exploration and discovery. *Cognition*, *120*(3), 322–330. doi: 10.1016/j.cognition.2010.10.001

Cheng, P. W. (1997). From covariation to causation: a causal power theory. *Psychological Review*, *104*(2), 367. doi: 0033-295 X/97/S3.00

De Leeuw, J. R. (2015). jspsych: A javascript library for creating behavioral experiments in a web browser. *Behavior Research Methods*, *47*(1), 1–12. doi: 10.3758/s13428-014-0458-y

Eaves, B. S., & Shafto, P. (2012). Unifying pedagogical reasoning and epistemic trust. *Advances in Child Development and Behavior*, *43*, 295–319. doi: 10.1016/B978-0-12-397919-3.00011-3

Fisher, K. R., Hirsh-Pasek, K., Newcombe, N., & Golinkoff, R. M. (2013). Taking shape: Supporting preschoolers' acquisition of geometric knowledge through guided play. *Child Development*, *84*(6), 1872–1878. doi: 10.1111/cdev.12091

Goodman, N. (1955). *Fact, fiction, & forecast*. Cambridge, MA, US: Harvard University Press.

Gopnik, A. (2016). *The gardener and the carpenter: What the new science of child development tells us about the relationship between parents and children*. New York, NY: Farrar, Straus, and Giroux.

Gopnik, A., Griffiths, T. L., & Lucas, C. G. (2015). When younger learners can be better (or at least more open-minded) than older ones. *Current Directions in Psychological Science*, *24*(2), 87–92. doi: 10.1177/0963721414556653

Gopnik, A., O'Grady, S., Lucas, C. G., Griffiths, T. L., Wente, A., Bridgers, S., ... Dahl, R. E. (2017). Changes

in cognitive flexibility and hypothesis search across human life history from childhood to adolescence to adulthood. *Proceedings of the National Academy of Sciences*, *114*(30), 7892–7899. doi: 10.1073/pnas.1700811114

Griffiths, T. L., & Tenenbaum, J. B. (2007). From mere coincidences to meaningful discoveries. *Cognition*, *103*(2), 180–226. doi: 0.1016/j.cognition.2006.03.004

Gweon, H., Pelton, H., Konopka, J. A., & Schulz, L. E. (2014). Sins of omission: Children selectively explore when teachers are under-informative. *Cognition*, *132*(3), 335–341. doi: 10.1016/j.cognition.2014.04.013

Honomichl, R. D., & Chen, Z. (2012). The role of guidance in children's discovery learning. *Wiley Interdisciplinary Reviews: Cognitive Science*, *3*(6), 615–622. doi: 10.1002/wcs.1199

Lucas, C. G., Bridgers, S., Griffiths, T. L., & Gopnik, A. (2014). When children are better (or at least more open-minded) learners than adults: Developmental differences in learning the forms of causal relationships. *Cognition*, *131*(2), 284–299. doi: 10.1177/0963721414556653

Schulz, L. E. (2012). The origins of inquiry: Inductive inference and exploration in early childhood. *Trends in Cognitive Sciences*, *16*(7), 382–389. doi: 10.1016/j.tics.2012.06.004

Schulz, L. E., & Bonawitz, E. B. (2007). Serious fun: preschoolers engage in more exploratory play when evidence is confounded. *Developmental Psychology*, *43*(4), 1045. doi: 10.1037/0012-1649.43.4.1045

Schulz, L. E., Gopnik, A., & Glymour, C. (2007). Preschool children learn about causal structure from conditional interventions. *Developmental Science*, *10*(3), 322–332. doi: 10.1111/j.1467-7687.2007.00587.x

Schulz, L. E., Kushnir, T., & Gopnik, A. (2007). Learning from doing: Intervention and causal inference. *Causal learning: Psychology, Philosophy, and Computation*, 67–85. doi: 10.1093/acprof:oso/9780195176803.003.0006

Sobel, D. M., Tenenbaum, J. B., & Gopnik, A. (2004). Children's causal inferences from indirect evidence: Backwards blocking and bayesian reasoning in preschoolers. *Cognitive Science*, *28*(3), 303–333. doi: $10.1207/s15516709cog2803_1$

Vul, E., & Pashler, H. (2008). Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science*, *19*(7), 645–647. doi: 10.1111/j.1467-9280.2008.02136.x

Weisberg, D. S., Hirsh-Pasek, K., & Golinkoff, R. M. (2013). Guided play: Where curricular goals meet a playful pedagogy. *Mind, Brain, and Education*, *7*(2), 104–112. doi: 10.1111/mbe.12015

Wente, A. O., Kimura, K., Walker, C. M., Banerjee, N., Fernández Flecha, M., MacDonald, B., … Gopnik, A. (2019). Causal learning across culture and socioeconomic status. *Child Development*, *90*(3), 859-875. doi: 10.1111/cdev.12943

Query-guided visual search

Junyi Chu
MIT, Cambridge, Massachusetts, United States

Jon Gauthier
Massachusetts Institute of Technology, Cambridge, Massachusetts, United States

Roger Levy
Massachusetts Institute of Technology, Cambridge, Massachusetts, United States

Josh Tenenbaum
MIT, Cambridge, Massachusetts, United States

Laura Schulz
Massachusetts Institute of Technology, Cambridge, Massachusetts, United States

Abstract

How do we seek information from our environment to find solutions to the questions facing us? We pose an open-ended visual search problem to adult participants, asking them to identify targets of questions in scenes guided by only an incomplete question prefix (e.g. Why is..., Where will...). Participants converged on visual targets and question completions given just these function words, but the preferred targets and completions for a given scene varied dramatically depending on the query. We account for this systematic query-guided behavior with a model linking conventions of linguistic reference to abstract representations of scene events. The ability to predict and find probable targets of incomplete queries may be just one example of a more general ability to pay attention to what problems require of their solutions, and to use those requirements as a helpful guide in searching for solutions.

Female advantage in visual working memory capacity for familiar shapes but not for abstract symbols

Adam Chuderski
Institute of Philosophy, Jagiellonian University in Krakow
Adam.Chuderski at uj.edu.pl

Jan Jastrzębski
Institute of Psychology, Jagiellonian University in Krakow
Jastrz.Jan at gmail.com

Abstract

Both behavioral studies and the neurophysiological data modelling suggested female advantage in memory for objects, however, most research pertained to long-term memory, whereas data from visual working memory (VWM) are scanty. In a large sample of 2044 people, the number of objects supposedly encoded in VWM was measured during the change detection task. The stimuli were either relatively familiar geometric shapes or less familiar Greek symbols. Controlling for the general ability level, a small but significant advantage for memorizing shapes in VWM was found in females over males, but no effect was observed for memorizing abstract symbols. The present results support neuroimaging models of human cognitive architecture, suggesting that female VWM relies on a more complex network of domain-specific brain modules, as compared to males. Consequently, formal models of VWM and related cognitive processes should account for sex and material type.

Keywords: visual working memory, sex differences, change detection task, neural architecture of memory

Introduction

Notable sex differences are observed in human memory, especially in long-term memory (LTM) and episodic memory (Halpern, 2013; Kimura, 1999). Research suggested that female brains are more effective in encoding and retrieving information pertaining to objects, episodes, faces, and verbal material, whereas males seem to better memorize spatial information (for reviews see Cahill, 2006; Herlitz & Rehnman, 2008). However, a relatively smaller number of studies were devoted to sex differences in working memory (WM), making this topic worth of closer examination.

WM is defined as a key neurocognitive mechanism responsible for active maintenance, effective updating, and controlled retrieval of task-relevant information during short periods of time (Cowan, 2001). At the same time, WM is believed to block task-irrelevant and distracting information (Kane & Engle, 2002). WM operation relies on short-term memory (STM), but most likely also involves memory processes beyond the sheer passive storage in STM. The key feature of WM is its limited capacity comprising the simultaneous representation of only several "chunks of information," being objects, their features, and their bindings. Given that WM is a key construct in psychology as well as a strong predictor of complex cognitive abilities, such as problem solving, fluid reasoning, and education (Kane, Hambrick, & Conway, 2005), vast research has been devoted to the neurocognitive underpinnings of WM. Establishing whether sex differences, commonly observed in other types of memory, do exist also in the case of WM performance, may contribute to our understanding of WM mechanisms. Moreover, if potential sex differences in WM are driven by the type of to-be-memorized content, in either a similar or a different way, in comparison to the content effects found for the other memory systems, such an observation may provide additional evidence for theoretical models that assume either close links between WM and LTM (Crowder, 1993; Nairne, 2002; Neath & Suprenant, 2003) or their relative separateness (e.g., Cowan, 2001; Kane & Engle, 2002; Vogel, Woodman, & Luck, 2001). The data can also guide design of the future studies on WM.

Sex differences in working memory

Early studies reported sex differences in WM that matched those found for LTM, with verbal tasks such as the reading span favoring females (e.g., Cochran & Davis, 1987), whereas spatial tasks such as the Corsi blocks favoring males (e.g., Grossi, Matarese, & Orsini, 1980). However, later studies reported differences that either were negligible, or highly variable from task to task (e.g., Duff & Hampson, 2001; Lejbak, Crossley, & Vrbancic, 2011; Postma, Jager, Kessels, Koppeschaar, & van Honk, 2004; Reed, Gallagher, Sullivan, Callicott, & Green, 2017; Robert & Savoie, 2006; for meta-analyses see Halpern, 2013; Voyer, Postma, Brake, & Imperato-McGinley, 2007; Voyer, Voyer, & Saint-Aubin, 2017; Wang & Carr, 2014).

Crucially, Speck et al. (2000) suggested that most of these cognitive studies were underpowered to detect robust sex differences in WM, whereas functional brain connectivity patterns differentiating females and males may be more informative than comparing WM capacity between females and males. Indeed, such patterns have been found (Filippi et al., 2013; Grabowski, Damasio, Eichhorn, & Tranel, 2003; Piefke et al., 2005). A comprehensive meta-analysis of 44

neuroimaging papers that studied males and another 15 that studies females, which used activation likelihood estimation (ALE) method, identified the brain regions most likely associated with sex differences. It showed that besides large overlapping structures in the frontal and parietal cortices, which are commonly attributed to WM, the female performance seems to depend also on additional prefrontal sites as well as hippocampus and anterior cingulate, whereas the male performance relies on additional parietal sites as well as insula (Hill, Laird, & Robinson, 2014).

However, it is still interesting to see how the observed sex differences in the brain activations might translate into behavioral differences in coping with tasks tapping WM. Only, a larger power is most likely needed to detect such behavioral consequences of neural differences. The present study aimed to investigate sex differences in visual working memory (VWM) using the varied types of to-be-memorized content. On the basis of previous research on sex differences in LTM, as well as interpreting the Hill et al.'s findings, it was assumed that the involvement of hippocampal regions in female WM may result in a better encoding of more concrete and familiar objects by females, as compared to males, because neurons in and around hippocampus are known to encode episodic information that can be linked to existing memory traces (see Eichenbaum, Yonelinas, & Ranganath, 2007; Wixted et al., 2014). The involvement in male WM of parietal and insular regions, commonly associated with awareness and attention (see Cowan et al., 2011; Eckert et al., 2009), can in turn yield the male advantage in encoding of less concrete and unfamiliar objects, which cannot be easily memorized using episodic and semantic traces, and thus require increased attentional effort. Indeed, some studies on episodic memory suggested the female advantage in recall and recognition of concrete pictures (Herlitz, Airaksinen, & Nordström, 1999) and familiar odors (Lehrner, 1993), but no advantage for more abstract images (ink blots and snow crystals; Goldstein & Chance, 1970) and unfamiliar odors (Öberg et al., 2002). Unfortunately, such a prediction has never been tested with regard to WM nor in large samples.

The present study analyzed data of 2044 people, collected over several published studies, conducted in the authors' laboratory between year 2007 and 2018. All these studies assessed VWM capacity using a simple recognition paradigm, called the change detection task, with stimuli being either geometric shapes or Greek symbols. The systematic use of the shape and the symbol variant of the change detection task gave an unique opportunity to check the sex × material interaction in a sample size never examined to date, which might allow to overpass the Speck et al. objections regarding the behavioral studies of sex differences in WM. In line with Hill et al. (2014), the female advantage in VWM capacity for more concrete, more familiar geometric figures was expected, as such figures could be encoded via episodic/semantic traces in and around hippocampus, which was a brain structure identified as a more specific to females. The male advantage in VMW capacity for more abstract, more unfamiliar Greek symbols was predicted, as such symbols could not be easily associated with episodic/semantic information, and might require increased attentional effort supported by the parietal sites and insula, found to be more active in males. Although the classification of shapes as more concrete, while Greek symbols as more abstract is not univocal (see Discussion), these two kinds of material were clearly different, and the examination of sex × material interaction was worthwhile.

The study

Participants

The total sample encompassed 1310 females (aged 17 to 46 years, M = 23.2, SD = 4.6) and 734 males (aged 18 to 46 years, M = 23.7, SD = 4.7). All participants were recruited from general population via internet adverts, in a Central-European city. The prevalence of females in the sample unfortunately resulted from the robust tendency for female enrolment in the psychological study recruitment. All participants signed a written consent to participate, were screened for normal or corrected-to-normal vision and no history of neurological problems, and were informed that they could stop the experiment and leave the lab at will. All data were anonymized. All other procedural aspects of the study conformed to the WMA's Declaration of Helsinki.

Materials

The stimuli in each trial of the symbol variant of the change-detection task were randomly drawn from the set of 16 small Greek letters (α, β, δ, θ, λ, μ, π, etc.), whereas in the shape variant they were drawn from the set of 16 simple shapes (circle, square, rhombus, etc.). Each stimulus was approximately 2×2 cm in size and was presented in black on a grey background. Each variant included either 60 or 90 trials, depending on a study (preceded by several training trials). Each trial consisted of a virtual, 4 × 4 array filled with several stimuli (see Fig. 1). From four to nine stimuli were used across studies. The array was visible between 1 s and 4 s (depending on set size), and was followed by a 1-s black square mask. On random, either the second array was identical to the first or both differed by exactly one item at one location. Either the new or random item, respectively, was highlighted by a square border. The task was to press one of two response keys (Z, M) depending on whether the highlighted item differed or not. The order of task variants was random between the studies, and they were preceded and followed by other tasks. The task score was the estimated average number of objects that were effectively maintained in VWM (k; Rouder et al., 2011), calculated as the participant' difference between the proportion of correct responses for arrays with the item change and the proportion of incorrect responses for unchanged arrays, multiplied by

the set size. For example, if in a six-object condition 80% correct was scored in the former trials, and 70% correct was scored in the latter trials, formula yielded $k = 6 \times (.80 - .30) = 3$ objects supposedly maintained in VWM. Thus, the k value is relatively insensitive to the actual set size.

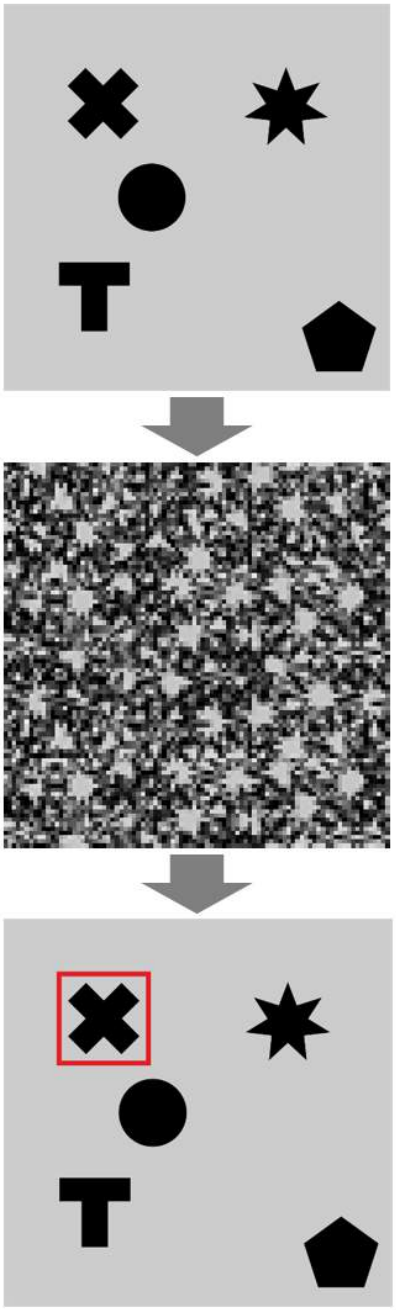


Fig. 1: Example stimuli and the sequence of events in the change detection task used in the study. The familiar shapes condition is shown; in the abstract symbols condition the task was identical except for Greek symbols were displayed. The "no-change" trial is shown; the "change trial was identical except for the shape (or symbol) surrounded by the rim in the bottom screen differed from the respective shape (or symbol) in the top screen.

As a short presentation time (2.5 s on average) and the graphical nature of stimuli practically eliminated their verbalization, it was assumed that the main difference between the tasks pertained to the concreteness and familiarity of stimuli, predicted to be larger for geometric shapes that commonly appear in the environment, whereas expected to be smaller for foreign Greek symbols that are not taught in schools and are rarely encountered in daily life and media in Poland, where the studies were held.

Additionally, general fluid intelligence (*gf*) was screened with two reasoning tests, Raven's APM (Raven et al., 1983) and (depending on a study) either Figural analogies (Chuderski & Nęcka, 2012) or Culture Fair Test Version 3 (Cattell & Cattell, 1961). The test results were converted to *Z* scores, separately for each study, and then averaged to yield the *gf* factor value. As VWM capacity strongly correlates with fluid intelligence (see Kane et al., 2005), this *gf* factor was used as a covariate in comparisons between females' and males' VWM scores, in order to make sure that any sex difference in fluid intelligence does not account for the expected sex differences in VWM.

Results

Males displayed the mean *gf* value of 0.047 ($SD = 0.960$), whereas females scored $gf = -0.019$ ($SD = 0.883$), and this difference was not statistically significant, $t(2042) = 1.59$, $p = .112$. Fig. 2 presents the female and male distribution of k values, separately for each material variant. All four distributions were normal, and yielded comparable standard deviations both for the shape variant, $SD_{female} = 1.49$, $SD_{male} = 1.52$, and the symbol variant, $SD_{female} = 1.49$, $SD_{male} = 1.52$. Visual inspection of Fig. 2 suggests that in the shape variant the female distribution was shifted right, relative to the male distribution, while in the symbol variant the distributions closely matched.

To formally test this observation, the k values were submitted to ANCOVA, with sex and material as two factors, and the *gf* factor as a covariate. Fig. 2 shows the respective means and 95% CIs. The shape variant yielded a comparable performance ($k = 3.05$) to the symbol variant ($k = 2.98$), $F(1, 4080) = 2.17$, $p = .141$, suggesting that overall both materials were equally demanding. The key analysis pertained to sex differences. There was a marginal effect of sex ($k_{female} = 3.04$, $k_{male} = 2.99$), $F(1, 4080) = 4.09$, $p = .043$, but the sex effect was qualified by its significant interaction with material, $F(1, 4080) = 8.15$, $p = .004$, $\eta^2 = .002$. In the shape variant, females performed significantly better than males ($\Delta k = 0.18$), $F(1, 4080) = 11.94$, $p = .0005$, $d = .20$, but no significant sex difference was noted for the symbol variant, ($\Delta k = -0.08$), $F(1, 4080) = 0.34$, $p = .559$. For females, the difference between variants was statistically significant in favor of the shape variant ($\Delta k = 0.20$), $F(1, 4080) = 8.15$, $p = .004$, Cohen's $d = .13$, while for males there was no significant difference between the two variants ($\Delta k = -0.06$), $F(1, 4080) = 0.75$, $p = .385$.

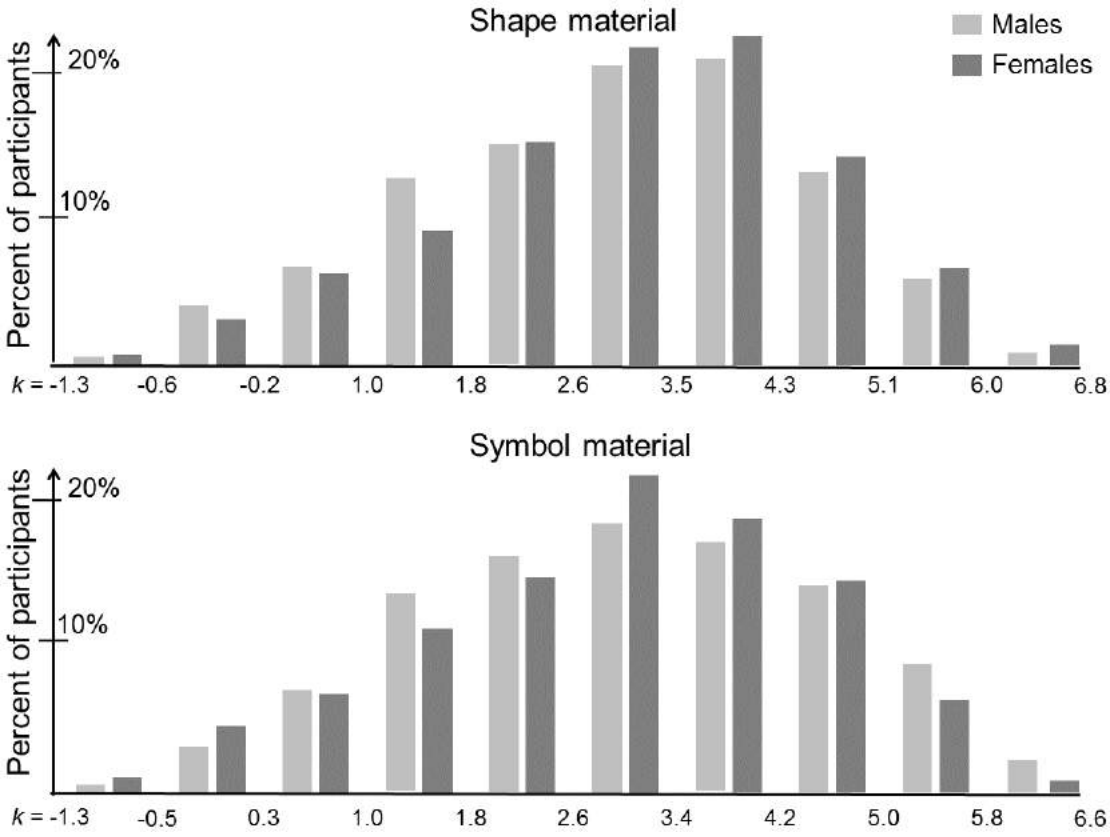


Figure 2: The distribution of *k* values for females and males.

The above analysis was also run without *gf* as a covariate, using rmANOVA with sex as a between-subjects factor and material as a within-subjects factor. The effect of sex was no longer statistically significant, $F(1, 2042) = 0.73$, $p = .392$, but the interaction of this factor with material was fully comparable with the preceding analysis, $F(1, 2042) = 15.21$, $p = .0001$, $\eta^2 = .001$. Female advantage over males for the shape material was highly significant, $F(1, 2042) = 7.03$, $p = .008$, while for the symbol material again there was no significant sex difference, $F(1, 2042) = 1.35$, $p = .246$. The shape material, as compared to symbols, yielded larger *k* values in females, $F(1, 2042) = 23.67$, $p < .0001$, but no significant difference related to material was noted for males, $F(1, 2042) = 1.51$, $p = .215$

In order to validate the null sex effect for the symbol material, ANCOVA was applied to another sample of 1486 people (aged 15 to 46, $M = 22.76$, $SD = 4.06$, $N_{female} = 938$), who performed only the symbol variant. They were also screened with two reasoning tests, which this time more visibly differentiated the two sexes, $t(1484) = 2.41$, $p = .016$ ($gf_{female} = -0.052$, $gf_{male} = 0.075$, $SD_{female} = 0.89$, $SD_{male} = 0.96$). However, also in this sample ANCOVA showed no significant difference for the symbol material between females ($k = 3.21$) and males ($k = 3.31$), $F(1, 1483) = 0.36$, $p = .546$. This difference was not significant even when the two samples were combined, $F(1, 2527) = 0.74$, $p = .389$.

Finally, no differences in correlation between the *k* and *gf* values was observed. For the shape material, the respective correlation coefficient was numerically the same both in the female and the male sub-sample, $r = 40$, $p < .0001$. It was quite comparable to the respective coefficients for the symbol material, $r_{female} = .37$, $r_{male} = .34$, ps $< .0001$.

Discussion

Neuroimaging data (Hill et al., 2014) suggested that WM tasks, besides the common prefrontal and parietal sites, activate additional prefrontal and hippocampal regions in the female brains, whereas additional parietal and insular regions in the male brains. The present analysis of the large set of scores in the change detection task tested behavioral consequences of this female/male neuronal specificity. Results indicated that one potential consequence of the sex differences in brain networks underlying WM is the female advantage in VWM capacity for more concrete, more familiar stimuli (possibly encoded by episodic/semantic traces in and around hippocampus), which is absent for more abstract, less familiar stimuli (possibly requiring more attention rooted in parietal sites and insula). As the sample size was particularly large as for the WM research, the data were *gf*-corrected, and the female advantage was specific for one type of material but not for the other, the reported effect very likely reflects factual difference between the sexes, and not just a sample-dependent variation.

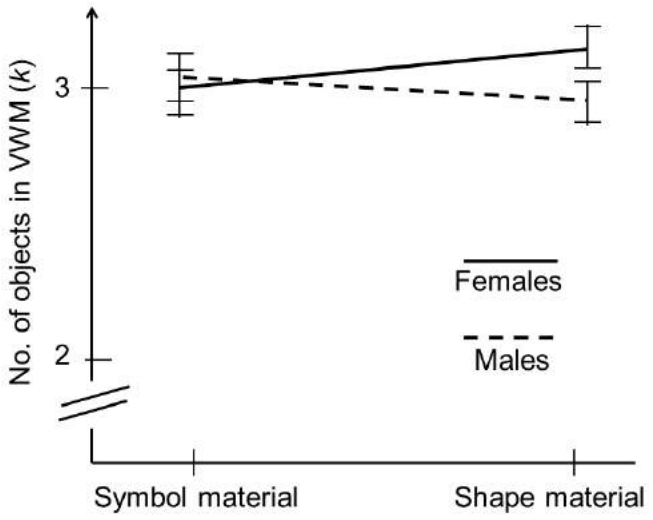


Fig. 3: Mean *k* values for females and males, depending on material type. Bars = 95% confidence intervals.

However, the female advantage for the shape material was quite small (Δk = .18). As the average VWM capacity was about three objects, females surpassed males by around 6%. This amount is in line with the Hill et al. conclusion that the lion's share of WM processing in both females and males relies on the shared prefrontal and parietal regions. However, given a strongly limited nature of WM, even such 6% can count, and females' potential reliance on specialized memory mechanisms may boost memory performance when a memorized content is compatible with those mechanisms. The present effect of stimulus familiarity was also much smaller than differences in episodic and semantic memory that were reported in the literature (see Cahill, 2006; Halpern, 2013; Herlitz & Rehnman, 2008; Kimura, 1999).

In contrast, the initially predicted male advantage for the symbol material was not observed in the data. After consideration, it seems that this prediction might be premature. Male advantage has been reported primarily for spatial material (Cahill, 2006; Grossi et al., 1980; Herlitz & Rehnman, 2008; Lewin, Wolgers, & Herlitz, 2001; Voyer et al., 2017), whereas more abstract material in fact did not differentiate the sexes (Goldstein & Chance, 1970). A more plausible interpretation of the null effect for the abstract symbols is that the additional involvement of attention might just have eliminated the female advantage rooted in more effective specialized memory processes (Herlitz & Rehnman, 2008; Voyer et al., 2007), but its contribution was too weak to yield the performance advantage of males.

One limitation of the study was the material used. Using other material than geometric shapes and Greek symbols would broaden the scope of conclusions that could be drawn from the present study. However, this study relied on the already existing data set, which included only two types of material. Moreover, the assumption that only shapes were familar to participants, and could be encoded in episodic memory, but the symbols could not, might be objected. Obviously, Greek symbols are also a kind of shapes, and at least some of them (e.g., α, β) could be verbalized, what helps in episodic encoding. So, we agree with all those objections. However, we think that the attenuated variant of this assumption, stating that shapes are *relatively* more familiar than Greek symbols (at least in the population with minimal exposure to Greek alphabet), and can be *relatively* more easily encoded in episodic memory, can be valid.

Summary

The present analysis of the existing large-size data set revealed the statistically significant difference between female and male WM performance on the relatively familiar, graphical material (but not on the more abstract material), which, on the one hand, most likely would be overlooked by a single study, whereas, on the other hand, might not be easily identifiable in meta-analyses of multiple studies applying diverse and not easily comparable methods. Overall, this kind of neuroimaging-driven psychometric analyses of sex differences in memory performance can shed light on the mechanisms underlying various memory systems as well as human cognitive architecture. The present study suggests that formal models of memory and related processes should account for sex and material type.

Acknowledgments

Raw data can be obtained from osf.io/cas4q/. This work was supported by the National Science Centre of Poland under grant number 2015/17/B/HS6/04152 to Adam Chuderski.

References

Cahill, L. (2006). Why sex matters for neuroscience. *Nature Reviews Neuroscience, 7,* 477-484.

Cattell, R. B., & Cattell, A. K. S. (1961). *Culture Free Intelligence Test, Version 3, Handbook.* Champaign, IL: Institute of Personality and Ability Testing.

Chuderski, A., & Necka, E. (2012). The contribution of working memory to fluid intelligence: Capacity, control, or both? *Journal of Experimental Psychology: Learning, Memory, & Cognition, 38,* 1689-1710.

Cochran, K. F., & Davis, J. K. (1987). Individual differences in inference processes. *Journal of Research in Personality, 21,* 197-210.

Crowder, R. G. (1993). Short-term memory: where do we stand? *Memory and Cognition, 21,* 142–145.

Duff, S. J., & Hampson, E. (2001). A sex difference on a novel spatial working memory task in humans. *Brain and Cognition, 47,* 470-493.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioural and Brain Sciences, 24,* 87-114.

Cowan, N., Li, D., Moffitt, A., Becker, T. M., Martin, E. A., Saults, J. S., & Christ, S. E. (2011). A neural region of abstract working memory. *Journal of Cognitive Neuroscience, 23,* 2852-2863.

de Bourbon-Teles, J., Bentley, P., Koshino, S., Shah, K., Dutta, A., Malhotra, P., Egner, T. Husain, M., & Soto, D. (2014). Thalamic control of human attention driven by memory and learning. *Current Biology, 24,* 993–999.

Eckert, M. A., Menon, V., Walczak, A., Ahlstrom. J., Denslow, S., Horwitz, A., Dubno, J. R. (2009). At the heart of the ventral attention system: the right anterior insula. *Human Brain Mapping, 30,* 2530–2541.

Eichenbaum, H., Yonelinas, A. P., & Ranganath, C. (2007). The medial temporal lobe and recognition memory. *Annual Review Neuroscience, 30,* 123–152.

Filippi, M., Valsasina, P., Misci, P., Falini, A., Comi, G., & Rocca, M. A. (2013). The organization of intrinsic brain activity differs between genders: A resting-state fMRI study in a large cohort of young healthy subjects. *Human Brain Mapping, 34,* 1330–1343.

Goldstein, A. G., & Chance, J. E. (1970). Visual recognition memory for complex configurations. *Perception & Psychophysics, 9,* 237-241.

Grabowski, T. J., Damasio, H., Eichhorn, G. R., & Tranel, D. (2003). Effects of gender on blood flow correlates of naming concrete entities. *Neuroimage, 20,* 940–954.

Grossi, D., Matarese, V., & Orsini, A. (1980). Sex differences in adults' spatial and verbal memory span. *Cortex, 16,* 339-340.

Halpern, D. F. (2000). *Sex differences in cognitive abilities,* (3rd ed). Mahwah, NJ: Erlbaum.

Herlitz, A., Airaksinen, E., & Nordström, E. (1999). Sex differences in episodic memory: The impact of verbal and visuospatial ability. *Neuropsychology, 13,* 590-597.

Herlitz, A., & Rehnman, J. (2008). Sex differences in episodic memory. *Current Directions in Psychological Science, 17,* 52-56.

Hill, A. C., Laird, A. R., & Robinson, J. L. (2014). Gender differences in working memory networks: A BrainMap meta-analysis. *Biological Psychology, 102,* 18-29.

Kane, M. J., Hambrick, D. Z., & Conway, A. R. A. (2005). Working memory capacity and fluid intelligence are strongly related constructs: Comment on Ackerman, Beier & Boyle (2005). *Psychological Bulletin, 131,* 66-71.

Kane, M. J., & Engle, M. J. (2002). The role of prefrontal cortex in working memory capacity, executive attention, and general fluid intelligence: An individual-differences perspective. *Psychonomic Bulletin & Review, 9,* 637-671.

Kimura, D. (1999). *Sex and cognition*. Cambridge, MA: MIT Press.

Lehrner, J. P. (1993). Gender differences in long-term odor recognition memory: Verbal versus sensory influences and the consistency of label use. *Chemical Senses, 18,* 17-26.

Lejbak, L., Crossley, M., & Vrbancic, M. (2011). A male advantage for spatial and object but not verbal working memory using the n-back task. *Brain and Cognition, 76,* 191–196.

Lewin, C., Wolgers, G., & Herlitz, A. (2001). Sex differences favouring women in verbal but not visuospatial episodic memory. *Neuropsychology, 15,* 165-173.

Nairne, J. S. (2002). Remembering over the short-term: the case against the standard model. *Annual Review of Psychology, 53,* 53–81.

Neath, I., & Surprenant, A. (2003). *Human memory (2nd ed.).* Belmont, CA: Wadsworth.

Öberg, C., Larsson, M., & Bäckman, L. (2002). Differential sex effects in olfactory functioning: The role of verbal processing. *Journal of International Neuropsychological Society, 8,* 691-698.

Piefke, M., Weiss, P. H., Markowitsch, H. J., & Fink, G. R. (2005). Gender differences in the functional neuroanatomy of emotional episodic autobiographical memory. *Human Brain Mapping, 24,* 313–324.

Postma, A., Jager, G., Kessels, R. P. C., Koppeschaar, H. P. F., & van Honk, J. (2004). Sex differences for selective forms of spatial memory. *Brain and Cognition, 54,* 24-34.

Raven, J. C., Court, J. H., & Raven, J. (1983). *Manual for Raven's Progressive Matrices and vocabulary scales (Section 4: Advanced Progressive Matrices)*. London: H. K. Lewis.

Reed, J. L., Gallagher, N. M., Sullivan, M., Callicott, J. H., & Green, A. E. (2017). Sex differences in verbal working memory performance emerge at very high loads of common neuroimaging tasks. *Brain and Cognition, 113,* 56-64.

Robert, M., & Savoie, N. (2006). Are there gender differences in verbal and visuospatial working-memory resources? *European Journal of Cognitive Psychology, 18,* 378-397.

Rouder, J. N., Morey, R. D., Morey, C. C., & Cowan, N. (2011). How to measure working memory capacity in the change detection paradigm. *Psychonomic Bulletin & Review, 18,* 324-330.

Speck, O., Ernst, T., Braun, J., Koch, C., Miller, E., & Chang, L. (2000). Gender differences in the functional organization of the brain for working memory. *NeuroReport, 11,* 2581–2585.

Vogel, E. K., Woodman, G. F., & Luck, S. J. (2001). Storage of features, conjunctions, and objects in visual working memory. *Journal of Experimental Psychology: Human Perception & Performance, 27,* 92–114.

Voyer, D., Postma, A., Brake, B., & Imperato-McGinley, J. (2007). Gender differences in object location memory: A meta-analysis. *Psychonomic Bulletin & Review, 14,* 23-38.

Voyer, D., Voyer, S. D., & Saint-Aubin, J. (2017). Sex differences in visual-spatial working memory: A meta-analysis. *Psychonomic Bulletin & Review, 24,* 307-334.

Wang, L., & Carr, M. (2014). Working memory and strategy use contribute to gender differences in spatial ability. *Educational Psychologist, 49,* 261–282.

Using transcranial Direct Current Stimulation (tDCS) to modulate the face inversion effect on the N170 ERP component.

Ciro Civile (c.civile@exeter.ac.uk)
School of Psychology, College of Life and Environmental Sciences, University of Exeter, UK.
Brad Wooster (bmw211@exeter.ac.uk)
School of Psychology, College of Life and Environmental Sciences, University of Exeter, UK.
Adam Curtis (a.curtis3@exeter.ac.uk)
School of Psychology, College of Life and Environmental Sciences, University of Exeter, UK.
R. McLaren (r.p.mclaren@exeter.ac.uk)
School of Psychology, College of Life and Environmental Sciences, University of Exeter, UK.
I.P.L. McLaren (i.p.l.mclaren@exeter.ac.uk)
School of Psychology, College of Life and Environmental Sciences, University of Exeter, UK.
Aureliu Lavric (a.lavric@exeter.ac.uk)
School of Psychology, College of Life and Environmental Sciences, University of Exeter, UK.

Abstract

In the present study, we combined tDCS and EEG to examine the electrophysiological responses to the tDCS-induced effects on the face inversion effect showed in recent studies. A double-blind procedure with a between-subjects design (n=48) was used with the subjects, recruited from the student population, being randomly assigned to either tDCS anodal or sham condition. The tDCS stimulation was delivered over the DLPFC at Fp3 site for 10 min at an intensity of 1.5mA while subjects engaged in an old/new recognition task traditionally used to obtain the inversion effect. The behavioural results generally confirmed previous findings. Critically, the results from the N170 show an effect of tDCS. Specifically, the tDCS procedure was able to modulate the N170 peak component by reducing the inversion effect on the latencies (i.e. less delay between upright and inverted faces) and by increasing the inversion effect on the amplitudes (i.e. larger N170 for inverted vs upright faces). We interpret the results based on the previous literature in regard to the inversion effect on the N170 component.

Keywords: Inversion effect; tDCS; N170, perceptual learning

Introduction

Several researchers have studied the nature of face recognition skills by investigating the causes of a robust phenomenon known as the face inversion effect. This refers to reduced performance when we try to recognize familiar faces turned upside down (Yin, 1969). When it was first discovered this phenomenon was used as a marker for "specificity" of face processing. This was because the inversion effect was found to be larger for faces than for other visual stimuli such as houses or planes (Valentine & Bruce, 1986; Yovel & Kanwisher, 2005). However, Diamond and Carey's (1986) finding of a large inversion effect for dog images when participants were dog breeders (vs that exhibited by novices), and Gauthier's work on perceptual expertise and the inversion effect for novel categories of objects named Greebles (Gauthier & Tarr, 1997) challenged the idea that faces are special and introduced "expertise" as a contributing factor to the inversion effect. Importantly, in 1997, McLaren using a set of artificial stimuli, checkerboards (so that expertise can be fully controlled), reported the first evidence of an inversion effect for novel stimuli that was predicted based on a specific model of perceptual learning, the MKM model (McLaren, Kaye & Mackintosh, 1989; McLaren & Mackintosh, 2000). Following this, Civile et al. (2014) extended McLaren's findings to the type of *old/new recognition task* originally used to investigate the face inversion effect (e.g. Yin, 1969). Taken together, Gauthier and Tarr's (1997), McLaren's (1997), and Civile et al's (2014) studies provide support for the Diamond and Carey's (1986) expertise account of face recognition; they have also served as a basis for further investigations of face and object recognition using Electroencephalogram (EEG) derived event-related potentials (ERPs).

Early studies on face recognition claimed the N170 ERP component to be the neural signature for face stimuli (Bentin et al., 1996). The N170 is a negative-polarity ERP deflection (peak) maximal at 140-200ms usually found at occipital-temporal electrodes after a face stimulus is presented (Bentin et al., 1996; George et al., 1996). The N170 has been found to be larger in amplitude and delayed in latency for inverted faces compared to upright faces. This is what has been commonly defined as the inversion effect on the N170 (Eimer, 2000). Rossion et al (2002) directly compared the N170 for faces and Greebles demonstrating how after the training phase with upright Greebles, the inversion effect (i.e. delayed and larger amplitude for inverted stimuli) was of a similar magnitude for both faces and Greebles. In a similar vein, Busey and Vanderkolk (2005) showed that

fingerprint experts exhibited an inversion effect on the N170 (similar to that for faces) in response to images of fingerprints. Furthermore, Civile et al. (2014a, Exp. 4), found an inversion effect on the N170 for checkerboards drawn from a familiar prototype-defined category (a larger and delayed N170 for inverted checkerboards compared to upright ones). The results from these studies provided motivation for a departure from the original account of the N170 component as being specific to faces, toward a position where the inversion-induced enhancement and delay of the N170 can be obtained for non-face categories of stimuli if they are made sufficiently familiar.

In recent years, Civile et al (2016) first, and then Civile, McLaren, and McLaren (2018a) (for a pilot see also Civile, Obhi & McLaren, 2018b) strengthened the analogy between the inversion effect for checkerboards (Civile et al., 2014), and that for faces, through demonstrating that they both share the same causal mechanism. Using a specific tDCS paradigm, the authors were able to modulate perceptual learning and selectively affect the robust inversion effect that otherwise would have been obtained for checkerboards and face stimuli. Anodal tDCS delivered over the DLPFC at Fp3 site (see Ambrus et al., 2011 for an example of previous studies targeting the same brain area to modulate categorization for prototype-defined stimuli) for 10 mins at an intensity of 1.5mA eliminated the inversion effect found for checkerboards by reducing performance for upright checkerboards taken from a familiar category (compared to controls) (Civile et al., 2016). Critically, the same tDCS paradigm is also able to reduce the robust face inversion effect by affecting recognition performance for upright faces (Exp.1 and the replication Exp.2 in Civile et al., 2018a). Furthermore, through an *active control* study the authors showed that applying the same tDCS anodal stimulation on a different brain area did not result in any difference between the face inversion effect compared to the sham group (Exp.3, Civile et al., 2018a). Overall the results from these studies using tDCS show how a particular tDCS procedure can modulate perceptual learning and so reduce the robust inversion effect that would otherwise be obtained with checkerboards (after participants have gained enough expertise with them) or faces.

In the present study, we extended the tDCS procedure adopted by Civile et al (2016) and Civile et al (2018a) to the face inversion effect on the N170 ERP component. To our knowledge, this is the first study that attempts to examine the behavioural tDCS-induced effects on the inversion effect to electrophysiological responses on the N170. Showing that the tDCS procedure used to affect the inversion effect for checkerboards and for faces can also modulate the N170, would strengthen the link between perceptual learning (and in general the expertise account) and face reocognition.

Method

Subjects

Overall, 48 naïve (right-handed) subjects (18 male, 30 Female; Mean age = 21.3 years, age range= 18-27, *SD*= 2.25) took part in the study. Subjects were randomly assigned to either sham or anodal tDCS groups (24 in each group). All the subjects were students from the University of Exeter and were selected according to the safety screening criteria approved by the Research Ethics Committee at the University of Exeter. The sample size was determined from earlier studies that used the same tDCS paradigm, EEG paradigm, face stimuli, and counterbalancing (Civile et al., 2018a, b, c).

Materials

The study used a set of 256 face images standardized to grayscale on a black background (Civile et al., 2018a, b, c). All stimuli images were cropped removing distracting features such as hairline, and adjusted for extreme differences in image luminance. The stimuli, whose dimensions were 5.63 cm x 7.84 cm, were presented at resolution of 1280 x 960 pixels. The experiment was run using. Examples of the stimuli used are given in Figure 1. The experiment was run using E-prime software Version 1.1 installed on a PC computer.

The Behavioural Task

The experiment consisted of a 'study phase' and an 'old/new recognition phase' (Civile et al., 2018a,b,c).

Study Phase. Once subjects gave their consent, the instructions for the Study Phase were presented on the screen. The aim of the task was for the subjects to try to memorize the faces presented on the screen. The trial started with a fixation cross (500ms) in the center of the screen, immediately followed by a blank screen (500ms), and then by a facial stimulus (3000ms). Then the fixation cross and the black screen were repeated, and another face presented, until all stimuli had been presented. Overall, 128 face stimuli were presented, 64 in their upright orientation and 64 were presented inverted. After all the 128 face stimuli had been presented, the program displayed another set of instructions, explaining the recognition task.

Recognition Task. In this task, subjects were asked to press the 'z' key if they recognized the face stimulus as having been shown in the study phase on any given trial, or press 'm' if they did not (the keys were counterbalanced). All the stimuli previously seen in the study phase were presented again, "old", intermixed with 128 "new" faces split by the two conditions (upright and inverted). All the faces were presented one at a time at random order. The trial structure was as that in the study phase however this time the stimuli were presented for a longer period (4000ms).

The tDCS Paradigm

Stimulation was delivered by a battery driven constant current stimulator (neuroConn DC-Stimulator Plus) using a pair of surface sponge electrodes (7cm x 5cm i.e.35 cm^2) soaked in saline solution and applied to the scalp at the target area of stimulation. We adopted the same tDCS montage used in Civile et al (2018a)'s study (Exp. 1 & 2). Hence, one of the electrodes (anode) was placed over the target

stimulation area (Fp3) and the other (cathode) on the forehead over the reference area (right eyebrow). The study was conducted using a double-blind procedure reliant on the neuroConn study mode in which the experimenter inputs numerical codes (provided by another experimenter otherwise unconnected with running the experiment), that switch the stimulation mode between "normal" (i.e. anodal) and "sham" stimulation. In the anodal condition, a direct current stimulation of 1.5mA was delivered for 10 mins (5 s fade-in and 5 s fade-out) starting as soon as the subjects began the behavioral task and continuing throughout the study. In the sham group, the identical stimulation mode was displayed on the stimulator and subjects experienced the same 5 s fade-in and 5 s fade-out, but with the stimulation intensity of 1.5mA delivered for just 30 s, following which a small current pulse (3 ms) was delivered every 550 ms (0.1mA over 15 ms) for the remainder of the 10 mins to check impedance levels. Subjects were randomly assigned to one of the tDCS groups (Sham or Anodal). For every subject the stimulation started at the beginning of the Study Phase and finished before the Old/New Recognition Task started.

Given the novelty inherent in combining tDCS and EEG techniques, especially with using two separate pieces of equipment, it is worth noting some of the practical challenges faced during the implementation of the study. Specifically, we realised the tDCS stimulation (both sham and anodal) induced strong artefacts on the EEG data. Thus, we made sure that the tDCS stimulation ended by the end of the study phase before we started recording the EEG for the recognition phase. Hence, our analysis of the EEG data will be entirely for the recognition phase.

EEG Recordings

The EEG was sampled at 1000 Hz, with a band-pass of 0.016-100 Hz, the reference at Cz and the ground at AFz using 32 Ag/AgCl active electrodes and BrainAmp amplifiers. The electrodes were placed on the scalp in an extended 10-20 configuration plus one on each earlobe (references during online recording). Their impedances were kept below 10 kΩ.

Data Processing and Analysis

As mentioned above in the *tDCS Paradigm* section the ERP analysis was limited to the recognition phase. Data processing was performed in BrainVision Analyzer. The data was first filtered offline using a Butterworth Zero Phase filter with a low cutoff of 0.5 Hz and a high cutoff of 30 Hz, each with a 24 dB/oct slope. Individual channels were manually inspected and removed from further analysis where physical interference from a tDCS electrode was noted during set-up, or where data otherwise showed signs of significant artefacts throughout. Electrodes retained the online reference to Cz. Peak amplitudes of the N170 were examined for differences between the experimental conditions. To improve the estimates of the amplitude and latency the N170 extraction was aided by linear decomposition of the EEG using Independent Component Analysis (ICA, Bell & Sejnowski, 1995). The ICA was run separately for each subject using all scalp channels and the entire dataset. The EEG segments were then averaged for every participant and experimental condition. For each subject, we identified ICA components that: (1) showed a deflection (peak) in the N170 time-range (at 160-220 ms following stimulus onset), and (2) had a scalp distribution containing an occipital-temporal negativity characteristic of N170 (the scalp distributions of components are the columns of the inverted unmixing matrix). This resulted in 1-4 ICA components corresponding to the N170 identified in most subjects - these were back-transformed into the EEG electrode space (by multiplying the components with the inverted unmixing matrix that had the columns corresponding to other components set to zero) and submitted to statistical analysis of N170 peak amplitude and latency. N170 latency and amplitude analyses were restricted to electrode PO8, (over the right temporal hemisphere) which often in the literature has shown bigger effects on the N170 (Civile et al., 2018c; Civile et al., 2014; Civile et al., 2012; Rossion & Jacques, 2008).

Results

Behavioral Results

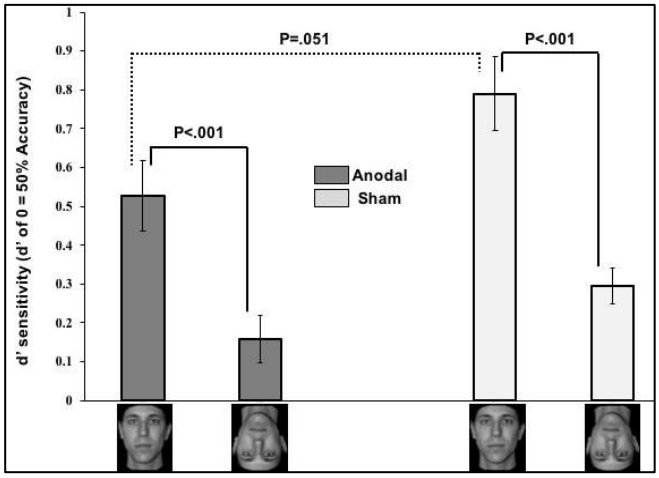
Following Civile et al (2018a,b,c) the data from all the participants were used in the signal detection d' sensitivity analysis of the recognition task (seen and not seen stimuli for each stimulus type) where a *d'* = of 0.00 indicates chance-level performance (Stanislaw & Todorov, 1999). We assessed performance against chance to show that both upright and inverted face stimuli in both the tDCS sham and anodal groups across the three experiments were recognized significantly above chance (for Sham Inverted, Sham Upright and Anodal Upright we found $p < .001$ for this analysis, for Anodal Inverted we found $p = .016$). Each p-value reported for the comparisons between conditions is *two-tailed*, and we also report the F or t value along with effect size. We also analyzed the reaction times (RTs) data to check for any speed-accuracy trade-off. We do not report this analysis here because it does not add anything to the interpretation of our results. For completeness, we give mean RTs for each of the stimulus' conditions: Sham Upright = 1240 ms; Sham Inverted = 1277 ms; Anodal Upright = 1263 ms; Anodal Inverted = 1267 ms.

d-Prime Analysis

We computed a 2 x 2 mixed model design using, as a within-subjects factor, *Face Orientation* (upright or inverted), and the between-subjects factor *tDCS Stimulation* (sham or anodal). Based on previous studies (Civile et al., 2018a,b) we expected the inversion effect for the anodal group to be smaller than that in the sham group. Analysis of Variance (ANOVA) revealed that numerically this was case but this time the interaction was not statistically significant, $F(1, 46) = .947$, $p = .33$, $\eta^2_p = .02$. There was a significant main effect of *Orientation* $F(1, 46) = 43.95$, $p < .001$, $\eta^2_p = .48$, which confirmed that upright faces were better responded to than inverted ones. A main effect of *tDCS Stimulation* was found with performance in the anodal

stimulation (M=.343, SE=.06) being significantly reduced compared to that in sham group (M=.542, SE=.05), $F(1, 46) = 5.39$, $p = .025$, $\eta^2_p = .10$. Paired t test analyses were conducted to compare performance on upright and inverted face stimuli (the inversion effect) in each tDCS group (sham, anodal). Based on previous studies that used the same stimuli and tDCS paradigm (Civile et al., 2018a,b) our primary measure was the face inversion effect given by comparing performance on upright and inverted faces in each tDCS group. We also directly compared the performance for upright faces in the sham vs tDCS group. This is particularly appropriate because the same stimulus sets are rotated across participants in a counterbalanced manner; so that each upright face seen in the anodal group for a given participant will equally often serve as an upright face for the participants in the sham group. A significant inversion effect was found in the sham group (M=.495, SE=.10), $t(23) = 4.97$, $p < .001$, $\eta^2_p = .38$, and a numerically reduced inversion effect was found in the tDCS anodal group (M=.368, SE=.07), $t(23) = 4.62$, $p < .001$, $\eta^2_p = .25$ (see Figure 1). Recognition for upright face stimuli in the anodal group was lower compared to that in the sham group, $t(46) = 2.05$, $p = .051$, $\eta^2_p = .19$. We also found a trend towards performance for inverted faces being reduced in the anodal relative to the sham group, $t(46) = 1.81$, $p = .083$, $\eta^2_p = .16$.

Figure 1. Results for the old/new recognition task. The *x*-axis shows the stimulus conditions. The *y*-axis shows sensitivity



d' measure. Error bars represent s.e.m.

Bayes Factor Analysis

Because we did not find a significant interaction in this experiment, as we had expected, we performed Bayesian analyses to check that our results fell within the usual parameters of our previous work. Using the procedure outlined by Dienes (2011), we first conducted a Bayes analysis on the Face Orientation by Stimulation interaction. Thus, we used the interaction effect averaged over Experiments 1 and 2 (0.30) from Civile et al. (2018a; same tDCS procedure, behavioural paradigm, stimuli, and sample size as in the study here reported)'s work as the prior (standard deviation of p). Then we used the standard error (0.03) and mean difference (0.13) for the interaction in our study, assumed a one-tailed distribution for our theory, and gave it a mean of 0. This resulted in a Bayes factor of 2162.84, which is strong evidence (greater than 10, for the conventional cut-offs see Jeffrey, 1961 and Dienes, 2011) indeed for the theory, in this case that the interaction will be positive and non-zero. Next, because in Civile et al. (2018a) both Experiments 1 & 2 had performance for the upright faces significantly better in the sham group compared to that in the anodal group, we calculated the Bayes factor for this effect in our study using as the prior the difference between sham minus anodal upright faces averaged over Civile et al. (2018a)'s Experiments 1 & 2 (0.28). We then used the standard error (0.11) and mean difference (0.26) between sham upright faces minus anodal upright faces in our study and assumed a one-tailed distribution for our theory with a mean of 0. This gave a Bayes factor of 8.10, which provides good evidence (as greater than 3) that sham performance on upright faces is higher than that under tDCS.

N170 ERP Results

In analyzing the N170 peak component we computed the same statistical analyses as for the behavioral data.

N170 Peak Latency Analysis

A 2 x 2 repeated measure ANOVA revealed a trend towards a significant interaction for peak latency, $F(1,46) = 3.26$, $p = .077$, $\eta^2_p = .06$. A significant main effect of *Orientation* was found, $F(1, 46) = 51.19$, $p < .001$, $\eta^2_p = .52$. No main effect of *tDCS Stimulation* was found, $F(1, 46) = .077$, $p = .783$, $\eta^2_p = .00$. A significant inversion effect (i.e. a delayed N170 peak for inverted vs upright faces) was found in the sham group (M=7.95ms, SE=1.28), $t(23) = 6.20$, $p < .001$, $\eta^2_p = .62$, and a numerically reduced inversion effect was found in the tDCS anodal group (M=4.70ms, SE=1.22), $t(23) = 3.86$, $p < .001$, $\eta^2_p = .39$. No difference was found between the N170 latencies for upright stimuli in the anodal vs sham group, $t(46) = .235$, $p = .815$, $\eta^2_p = .00$. No significant difference was found between inverted faces in the anodal vs sham group, $t(46) = .903$, $p = .375$, $\eta^2_p = .04$.

N170 Peak Amplitude Analysis

A 2 x 2 ANOVA revealed a significant *Orientation* by *Stimulation* interaction for peak amplitude, $F(1,46) = 4.06$, $p = .049$, $\eta^2_p = .09$, and a main effect of *Orientation*, $F(1, 46) = 45.47$, $p < .001$, $\eta^2_p = .49$. No main effect of *tDCS Stimulation* was found, $F(1, 46) = .178$, $p = .679$, $\eta^2_p = .00$. Contrarily to what we found for N170 latencies, the inversion effect (larger N170 for inverted vs upright faces) was found to be larger in amplitude in the anodal group (M=3.32μV, SE=.63) $t(23) = 5.22$, $p < .001$, $\eta^2_p = .54$, compared to that found in the sham group (M=2.41, SE=.52) $t(23) = 4.32$, $p < .001$, $\eta^2_p = .44$ (see Figure 2). No difference was found between the N170 amplitude for upright stimuli in the anodal vs sham group, $t(46) = .033$, $p = .975$, $\eta^2_p = .00$. Despite a numerically larger N170 for the inverted faces in the anodal vs sham group, no significant difference was found, $t(46) = .882$, $p = .386$, $\eta^2_p = .03$.

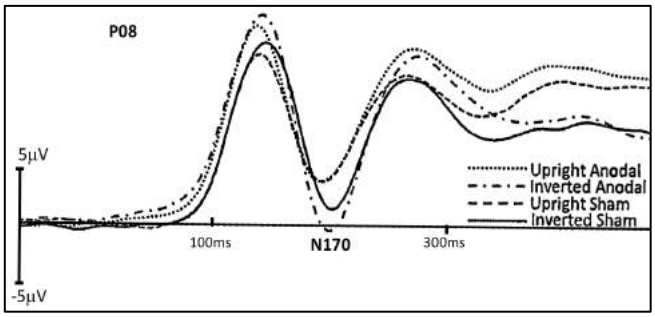


Figure 2. Waveforms at electrode P08 for the four conditions. The X axis shows the elapsed time after a stimulus was presented. The Y axis gives the amplitudes (μV) of the ERPs in the recognition phase of the experiment.

General Discussion

In the study reported here we examined the effects of tDCS on electrophysiological correlates of the face inversion effect. Specifically, we adopted the same tDCS procedure used by Civile et al (2018a,b) and Civile et al (2016) to modulate perceptual learning and affect the inversion effect for newly acquired stimuli (i.e. checkerboards) and long-term learnt stimuli i.e. faces. Our behavioral results are in line with previous work. Despite the inversion effect in the anodal group being only numerically reduced compared to sham, the additional Bayes Factor analysis gives us confidence that our effects are in line with previous work (Civile et al., 2018a). Importantly, as in previous studies, we find that anodal tDCS is particularly effective in reducing the recognition performance for upright faces, a result also supported by the Bayes Factor analysis. Our behavioral results also hint at a tendency (not significant) for anodal tDCS to reduce performance for the inverted faces. This is a new trend that previous studies (Civile et al., 2018a,b) did not show.

The most novel aspect of the present study involves the ERP results. To our knowledge, the current study provides the first evidence for tDCS being able to modulate a robust ERP component such as the N170 often associated with faces as well as sets of prototype-defined artificial stimuli that participants have become familiar with (Rossion et al., 2002; Civile et al., 2014). Intriguingly, our results suggest a dissociation of the effects that tDCS has on the N170. Specifically, in the latencies we find tDCS reduces the inversion effect compared to sham (less delay between the peaks for inverted vs upright faces). At the same time, tDCS increases the inversion effect on the N170 amplitudes compared to sham (a larger difference between the peak amplitude of the N170 for inverted vs upright faces). The effects of tDCS on the N170 latencies are more easily interpreted. Specifically, on the expertise account we can argue that a delayed N170 is recorded for a target face or familiar stimulus as a consequence of the familiarity lost when the target stimulus is turned upside down. We know from the behavioral results that anodal tDCS affects perceptual learning (by reducing expertise) for upright faces, making them more similar to stimuli drawn from an unfamiliar category, and thus this would result in a latency more similar to that for inverted faces.

Remarkably, the results from the N170 amplitude analysis provide some evidence for a dissociation from the tDCS-induced effects on the ERP latencies. Here anodal tDCS increased the inversion effect seen in the N170 amplitudes, and the inverted faces in this condition were found to elicit the largest N170 (i.e. more negative) compared to all the other stimulus conditions. But we should beware of attributing this effect to the impact of tDCS on the inverted faces, as this wasn't independently significant. All we can be sure of is that the inversion effect (difference between peak amplitudes) increased as a result of tDCS. In line with our explanation for the N170 latencies, if we assume that anodal tDCS is affecting participant's expertise for faces, then why would this have any impact on inverted faces when we have already argued that it will affect upright ones? Instead, it may be better to just focus on the significant effect (i.e. the difference between upright and inverted), and speculate that there may be some shift in baseline effects in our tDCS condition (not unlikely, we are, after all, activating a substantial region of frontal cortex using anodal stimulation) that results in the inverted face ERP apparently showing the greatest effect.

That still leaves us with the effect on peak amplitude to explain, and here it may be that we have to appeal to the difference between upright faces, for which we have expertise, and two different types of stimuli for which we do not. We assume that inverted faces do not benefit from our expertise with upright faces, whilst still acknowledging that they are readily recognized as faces. Another type of stimulus that would not benefit from expertise would be an entirely novel stimulus (a Greeble, a checkerboard). But this stimulus is not an inverted face. Now, if we postulate that tDCS makes the upright faces more like a novel stimulus, and that novel stimuli, other things being equal, do not show such a pronounced N170, then the greater difference from the inverted face N170 could be explained. Essentially, we would argue that tDCS shifts the upright face N170 towards that of a novel stimulus, which has a smaller amplitude and a greater latency, and that this is why we get our apparently "opposite" effects.

Interestingly, something like this pattern of results has previously been found in EEG studies where the level of familiarity for the stimuli presented was manipulated directly by means of training to the stimuli or by altering the typical familiar stimulus configuration (e.g. rearrange the locations of the features within a face). In Civile et al. (2014)'s study, the N170 peak amplitudes for inverted checkerboards taken from a familiar category were larger compared to the other conditions (upright checkerboards from a familiar category and upright and inverted novel checkerboards). Furthermore, Civile et al (2018c) found normal inverted faces elicited a larger N170 amplitude compared to normal upright faces and scrambled (i.e. the facial features were shuffled) upright/inverted faces (see also Civile et al., 2012 for similar results using Thatcherised faces). Finally, also Rossion et al. (2002) showed (in the pre-training phase) the N170 peak

amplitude being larger for normal inverted faces compared to normal upright faces, and upright/inverted Greebles. Civile et al (2018c) suggested that this effect is due to the fact that the normal inverted faces possess all the configural information (spatial relations) of a normal upright face, but presented in an orientation that not only makes it difficult to make use of them but imposes an additional cost. Thus the idea here would be that the differences in the N170 caused by inversion only partly index the effect of perceptual learning (in the latencies), the amplitude difference reflects something else (perceptual effort perhaps).

In conclusion, in the study reported here, we have provided some evidence in support of a tDCS procedure able to modulate the face inversion of the N170 component. Importantly, the tDCS-induced effects on the N170 seem to dissociate between latencies and amplitudes of the N170. Further studies will be needed to establish these effects.

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Curie grant agreement No.743702 awarded to Ciro Civile; from the Economic and Social Research Council *New Investigator Grant (Ref.ES/R005532)* awarded to Ciro Civile (PI) and I.P.L. McLaren (Co-I).

References

Ambrus G. G., Zimmer M., Kincses Z. T., Harza I., Kovacs G., Paulus W., et al. (2011). The enhancement of cortical excitability over the DLPFC before and during training impairs categorization in the prototype distortion task. *Neuropsychologia 49*, 1974–1980.

Bell, A. J, & Sejnowski, T. J. (1995). An information–maximization approach to blind separation and blind deconvolution. *Neural Computation, 7*, 1129-59.

Bentin, S., Allison, T., Puce, A., Perez, E., & McCarthy, G. (1996). Electrophysiological studies of face perception in humans. *Journal of Cognitive Neurosci*ence, 8, 551-565.

Busey, T., & Vanderkolk, J. (2005). Behavioural and electrophysiological evidence for configural processing in fingerprint experts. *Vision Research*, 45, 431-448.

Civile, C., Elchlepp, H., McLaren, R., Lavric, A & McLaren, I.P.L. (2012). Face recognition and brain potentials: Disruption of configural information reduces the face inversion effect. *Proceedings of the 34th Annual Conference of the Cognitive Science Society,* (pp. 1422-27). Austin, TX: Cognitive Science Society.

Civile, C., Zhao, D., Ku, Y., Elchlepp, H., Lavric, A., & McLaren, I.P.L. (2014). Perceptual learning and inversion effects: Recognition of prototype-defined familiar checkerboards. *Journal of Experimental Psychology: Animal Behavior Processes*, 40, 144-61.

Civile, C., Verbruggen, F., McLaren, R., Zhao, D., Ku, Y., & McLaren, I.P.L. (2016). Switching off perceptual learning: Anodal transcranial direct current stimulation (tDCS) at Fp3 eliminates perceptual learning in humans. *Journal of Experimental Psychology: Animal Learning and Cognition*, 42, 290-296.

Civile, C., McLaren, R., and McLaren, I.P.L. (2018a). How we can change your mind: Anodal tDCS to Fp3 alters human stimulus representation and learning. *Neuropsychologia*, *119, 241-246*.

Civile, C., Obhi, S.S., & McLaren, I.P.L. (2018b). The Role of Experience Based Perceptual Learning in the Face Inversion Effect. *Vision Research*, doi.org/10.1016/j.visres.2018.02.010.

Civile, C., Elchlepp, H., McLaren, R., Galang, C.M., Lavric, A., & McLaren, I.P.L. (2018c). The effect of scrambling upright and inverted faces on the N170. *Quarterly Journal of Experimental Psychology, 71, 2464-2476*.

Diamond, R. & Carey, S. (1986). Why faces are and are not special: An effect of expertise. *Journal of Experimental Psychology*: *General*, 115, 107-117.

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, 6, 274–290.

Eimer, M. (2000). The face-specific N170 component reflects late stages in the structural encoding of faces. *NeuroReport, 11, 2319-2324.*

Gauthier, I., & Tarr, M. (1997). Becoming a "Greeble" expert: exploring mechanisms for face recognition. *Vision Research,* 37, 1673-1682.

McLaren, I.P.L (1997). Categorization and perceptual learning: An analogue of the face inversion effect. *The Quarterly Journal of ExperimentalPsychology 50A (2),. 257-273.*

George, N., Evans, J., Fiori, N., Davidoff, J., & Renault, B. (1996). Brain events related to normal and moderately scrambled faces. *Cognitive Brain Research*, 4, 65-76.

McLaren, I.P.L., Kaye, H. & Mackintosh, N.J. (1989). An associative theory of the representation of stimuli: Applications to perceptual learning and latent inhibition. In R.G.M. Morris (Ed.) *Parallel Distributed Processing - Implications for Psychology and Neurobiology*. Oxford, Oxford University Press.

McLaren, I.P.L. & Mackintosh, N.J. (2000). An elemental model of associative learning: Latent inhibition and perceptual learning. *Animal Learning and Behavior*, 38, 211-246.

Rossion, B., Gauthier, I., Goffaux, V., Tarr, M.-J., Crommelinck, M. (2002). Expertise training with novel objects leads to face-like electrophysiological responses. *Psychological Science, 13, 250-257*.

Rossion B. & Jacques C. (2008). Does physical interstimulus variance account for early electrophysiological face sensitive responses in the human brain Ten lessons on the N170. *Neuroimage*, 39, 1959–1979.

Valentine, T., & Bruce, V. (1986). Recognizing familiar faces : The role of distinctiveness and familiarity. *Canadian Journal of Psychology,* 40, 300-305.

Yin, R. K. (1969). Looking at upside-down faces. *Journal of Experimental Psychology*, 81, 141-145.

Yovel G., & Kanwisher N. (2005) The neural basis of the behavioral face-inversion effect *Current Biology,* 15, 2256-62.

Reinforcement Learning and Insight in the Artificial Pigeon

Thomas R. Colin (thomas.colin@plymouth.ac.uk)
School of Mathematics and Computing, University of Plymouth
Plymouth, U.K.

Tony Belpaeme (tony.belpaeme@plymouth.ac.uk)
School of Mathematics and Computing, University of Plymouth
Plymouth, U.K.

Abstract

The phenomenon of insight (also called "Aha!" or "Eureka!" moments) is considered a core component of creative cognition. It is also a puzzle and a challenge for statistics-based approaches to behavior such as associative learning and reinforcement learning. We simulate a classic experiment on insight in pigeons using deep Reinforcement Learning. We show that prior experience may produce large and rapid performance improvements reminiscent of insights, and we suggest theoretical connections between concepts from machine learning (such as the value function or overfitting) and concepts from psychology (such as feelings-of-warmth and the einstellung effect). However, the simulated pigeons were slower than the real pigeons at solving the test problem, requiring a greater amount of trial and error: their "insightful" behavior was sudden by comparison with learning from scratch, but slow by comparison with real pigeons. This leaves open the question of whether incremental improvements to reinforcement learning algorithms will be sufficient to produce insightful behavior.

Keywords: reinforcement learning; insight; creativity

Introduction

Insight moments are one of the most spectacular manifestations of human creativity. Revolutionary insights are paradigmatic examples of creativity, whether historically suspicious (Aristotles' "Eureka!", Newton's apple), or better documented such as those described by the mathematician Poincaré (1909) or the chemist Kekulé (Rothenberg, 1995). In this article, however, we focus on the insights which occur in everyday human and animal problem-solving.

Over a century of research in psychology underlies our knowledge of insightful problem-solving. In contrast, to our knowledge there has been relatively little work considering insight from an artificial intelligence perspective, especially since the momentous advent of deep learning techniques in AI. We seek to remedy this omission. The objective is not to build a precise model of biological neural processes, but to uncover analogies between the two domains of deep Reinforcement Learning (RL) and biological insight. We do this by simulating a classic experiment on insight (Epstein, Kirshnit, Lanza, & Rubins, 1984), dealing with insight in the pigeon.

We will first discuss established results from insight research in psychology on humans and animals, and the difficulties associated with modeling insight problems from a machine learning perspective. We will then describe the original experiment and its simulation, and the results obtained using a simple deep RL approach (a deep actor-critic). Finally, we discuss the analogies between insight and various sub-disciplines within reinforcement learning, suggesting directions for future research.

Background: insight

Psychological research on insight begins with studies on chimpanzees by Köhler (1921). These studies sought to demonstrate that animals, far from being Cartesian automatas as suggested in the work of Thorndike (1898), are capable of human-like intelligence. One of Köhler's experiments involved attaching a banana to the ceiling of the chimpanzee enclosure, and placing a box within the enclosure. The chimpanzees had to carry the box underneath the banana and climb onto it in order to reach the fruit. When solving the problem, the chimpanzees displayed behavior that more closely resembled Aristotle's "Eureka!" than the trial-and-error learning of cats locked in puzzle-boxes by Thorndike (1898). In Köhler's "gestalt" perspective, it was understood that chimpanzees had to interpret the situation from scratch in order to discover the "roundabout" way of reaching for the objective.

Later work by Birch (1945) showed that chimpanzee insight was not achieved from scratch, but was instead made possible by relevant prior experiences. Epstein et al. (1984) showed that with adequate training, "even" pigeons could display the kind of insight observed in chimpanzees. Epstein's findings are robust: several variations of this experiment were performed by Epstein and colleagues, and the original was recently replicated by Cook and Fowler (2014). For Epstein, who was a student of Skinner, this made the argument that seemingly complex mental processes could be explained from behaviorist principles.

There has been continued interest in insight since the cognitive turn in psychology. This body of work has established several key behavioral, cognitive, and metacognitive characteristics of insight:

1. The insight sequence: search – (impasse) – restructuring – verification (Ohlsson, 2011; Weisberg, 2015).
2. Insights are sudden and surprising to the problem-solver, as evidenced by "feeling-of-warmth" ratings measuring subjective closeness to the solution (Metcalfe & Wiebe, 1987).
3. The "restructuring" which accompany insight involves

changes in problem representation (Knoblich, Ohlsson, & Raney, 2001), in the heuristics used (Kaplan & Simon, 1990), and in the constraints on operators (MacGregor, Ormerod, & Chronicle, 2001).

4. Insight depends on previous experience (Wiley, 1998) and is facilitated by sleep (Wagner, Gais, Haider, Verleger, & Born, 2004).

Recent research on insight has used imaging techniques such as fMRI[1]. Much of this work has focused on associative cortices (notably middle and temporal gyri) and on hemispheric differences (Kounios & Beeman, 2015); however the involvement of structures associated with executive control is a robust finding (prefrontal cortex, especially anterior cingulate cortex), and recent ultra high-field work (Tik et al., 2018) suggests the involvement of deeper brain structures during insight, including those underlying biological reinforcement learning (subcortical dopaminergic structures including the striatum, thalamus, nucleus accumbens and ventral tegmental area).

Summarizing: a rich body of research has investigated insight according to different psychological research paradigms, establishing the key characteristics of insight enumerated above. However, the precise nature of the cognitive mechanisms that enable insight remains unclear.

This is not to say that there have not been models, or theories, of the cognitive basis of insight (computational, mathematical, or otherwise); those of Hélie and Sun (2010), Friston et al. (2017), Schilling (2005), and Stephen, Boncoddo, Magnuson, and Dixon (2009) are among the most influential. A review of and comparison with these variegated models is beyond the scope of this paper, if only due to their great diversity, which ranges from bayesian inference (Friston et al., 2017) to dynamical systems (Stephen et al., 2009) and graph theory (Schilling, 2005). We note in passing that the model presented later in this article may be compatible with several of these other models: for instance phase transitions such as those described by Stephen et al. (2009) are conjectured to occur in neural networks.

None of the four models mentioned above aim to give rise to artificial agents capable of solving problems through insight[2]. In contrast, we seek to produce a model of insight problem-solving which, when implemented, not only predicts the behavior of a biological insightful problem-solver, but also solves the problem.

AI: which insight problems to model?

Most of the contemporary insight literature focuses on humans, using a wide array of experimental designs (for instance, the nine-dots problem (MacGregor et al., 2001), the mutilated checkerboard problem (Kaplan & Simon, 1990), or the Compound Remote Associates (Bowden & Jung-Beeman, 2003)). Despite their apparent variety, virtually all insight studies involving humans make use of verbal instructions which define the objective for the problem-solver in their language.

Consider the nine-dot problem: the instructions specify the number of segments, with constraints over their properties (four segments, drawn in a sequence "without lifting the pen"; every dot should end up on one of the segments). Language thus allows for a description of the desired "goal-state" which is abstract enough to specify the solution without giving it away. Simulating such a problem using AI would require either very task-specific algorithms (which seems to defeat the point of replicating human insight), or the algorithmic mastery of language as a prerequisite for understanding instructions.

A "roundabout" solution is to focus instead on insight experiments which feature animals solving problems that are not specified by instructions, but instead by some intrinsic need, typically for food, and by the situation in which the experimenter puts the animal[3]. This is the approach taken in this article.

Insightful (real) pigeons

The experiment by Epstein et al. (1984) is a reproduction of Köhler's banana-and-box experiment, adapted for pigeons. Chimpanzees would naturally want to acquire a banana; but pigeons might not be interested in that fruit. Therefore Epstein et al. first reinforced pecking a facsimile banana (hereafter just "the banana") by providing a suitable food reward upon pecks. In the "test" situation, the banana is suspended from the ceiling of the room, such that pigeons cannot reach it by stretching towards it (they do not attempt to fly towards it (Cook & Fowler, 2014)). However, a small cardboard cube ("the box") has been placed in the pigeon's Skinner box. The problem is solved when the animal pushes/pecks the box underneath the banana and, standing on the box, reaches for/pecks at the banana; see figure 1.

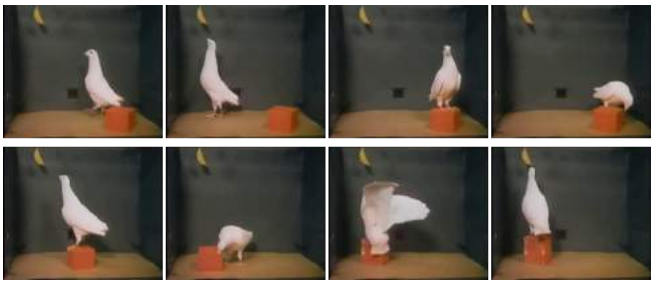


Figure 1: Left-to-right, then top-to-bottom: a pigeon solves the banana-and-box test (snapshots from https://youtu.be/mDntbGRPeEU, with permission from Dr. Epstein).

[1] See Sprugnoli et al. (2017) for a review of brain imaging studies.

[2] A notable exception is the model of MacLellan (2011), who investigates insight as a change of heuristics in a search process, and tests this on the nine-dot problem.

[3] See Shettleworth (2012) for a judicious review of insight research on animals.

Prior to this apparent display of ingenuity, the behavior of Epstein's pigeons was carefully *shaped*. Shaping is a technique used in animal training (with closely related applications in certain behavioural therapies for humans), consisting of reinforcing successive approximations of a desired behavior. Two skills ("behavioural repertoires") are taught to the pigeons by reinforcing the corresponding behaviors:

- In the absence of the banana: push a box to a green spot.
- With the box nailed underneath the banana, and in the absence of a spot: jump on the box and peck the banana.

Teaching pigeons to push a box towards an objective is considerably more difficult than getting them to hop onto the pre-placed box. To achieve this, Epstein et al. proceeded gradually, the shaping sequence including teaching the pigeons to move the box, then progressively placing the box at an increased distance from the spot. Additionally, the pigeons were sometimes put in the presence of the box and in the absence of both banana and spot, in order to extinguish aimless pushing behavior (which eventually would result, via a random walk within the Skinner box, in reaching the correct position and thereby triggering the food reward).

It is of special importance that the two behaviors are not exactly applicable to the final test: the pigeons are trained to push the box towards a green spot, but in the test situation they must spontaneously generalize this behavior to a slightly different problem: pushing towards the yellow banana. It is by combining two behaviors, and generalizing one behavior to a novel situation, that the pigeons solve the test task.

Epstein's pigeons proved remarkably adept in the test - all of them succeeding in minutes, save for one, and presenting behavior that seemed insightful: after a period of hesitation and some trial and error, the pigeons began acting in a seemingly directed, intentional manner, moving the box towards the banana and jumping on top of it. The lone laggard failed in a manner reminiscent of AI failures: during the test, a projector had been used to illuminate the (filmed) performance. When the additional lighting was turned off, the pigeon succeeded quickly.

Simulation

Admittedly, the displays of insight by Epstein's pigeons are less impressive than those of Köhler's chimpanzees: they received substantial training in the form of shaping. However, just as pigeons could not solve the test without having first acquired relevant skills, so chimpanzees were not able to solve insight problems without having first engaged in spontaneous play with the relevant objects (Birch, 1945). This suggests that similar cognitive mechanisms may be at play, and that it may be wise to begin by modeling the version of the task completed by pigeons.

In addition to requiring no instructions or verbal skill, the task used by Epstein et al. (1984) allows for a simulation which preserves much of what makes the task difficult: the pigeons had to combine pre-existing skills (pushing the box, and jumping on top of it to peck at the banana) while also generalizing to a new stimulus (pushing is shaped using a green dot, but in the test situation the pigeons must aim instead for a banana).

Thus, in simulating this task, we seek to preserve the difficulty inasmuch as it is relevant to problem solving, as opposed to the complete difficulty of the task including subjective perception and full physical coordination.

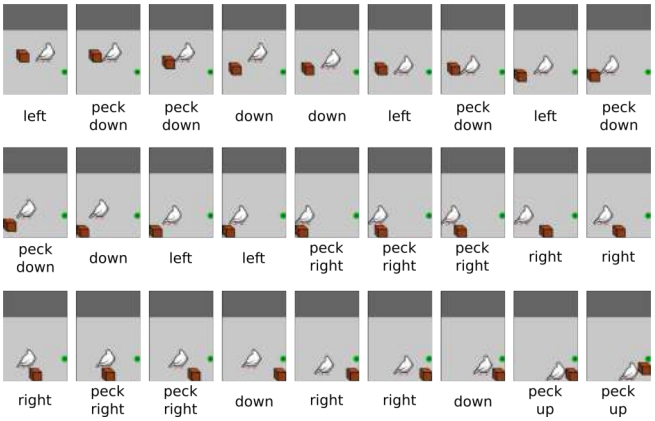


Figure 2: Successive frames of an artificial pigeon solving the "push box to spot" shaping task. The pigeon succeeded despite a sub-optimal policy (first pushing the box in the wrong direction, then pecking it out of corners). Also note the stochasticity of the "peck" actions: pecking actions have up to 4 different outcomes.

Task and shaping model

We model the task as an RGB image, such that the complete situation is perceived at each time-step. The pigeon, box, banana and spot consist of squares identifiable by size and color. For visualization, an interpretable representation is also provided (see figure 2). The dimensions of the various elements, and the dynamics of the actions are chosen to match those observed in the experiment. In particular, the size of the various elements (Skinner box, pigeon, box, banana, spot), the effects of the actions (walking, directional pecking, and jumping) and the consequences of interactions (box movement) closely match those of the initial experiment.

Specifically, the pigeon has 9 actions: walking in either cardinal direction, pecking towards either cardinal direction, and jumping on/off the box. Walking is deterministic and moves the pigeon by 1 square in the corresponding cardinal direction unless an obstacle is present. Pecking the box will result in its stochastic displacement in the general direction opposite to that from which it was pecked: assuming there are no obstacles and the box is not fixed in place, the box moves with equal probability (0.25) by 1 or 2 squares forward, or by 1 square forward and 1 in either perpendicular direction. With respect to direction, the pigeon can push the box south if the northern edge of the pigeon is at least as far north as the northern edge of the box, and if the pigeon is adjacent to the box;

likewise (mutadis mutandis) for the other directions (refer to figure 2 for some examples of stochastic box movement and pigeon positioning). The white pigeon is 3×3, the orange box 2×2, the green spot and yellow banana are 1 square each, and the background environment 10×10. Assuming squares approximately 4cm across, this roughly matches the size of the real objects (10x10cm box, 7x2cm facsimile banana, 4x4cm spot, approx. 25x8cm pigeons), Skinner box (45x45cm for the square box), and the effects of recognizable discrete actions in the original. In Skinner boxes, pigeons are rewarded by receiving food through a little window; in this simulation, a reward of 10 is provided instantaneously upon success.

The artificial pigeons undergo shaping similar to that used by Epstein et al.: artificial pigeons perform the **push-to-spot**, **jump-and-peck**, or **push-extinction** tasks. In the push-to-spot task, the box is initially placed immediately next to the spot. The distance between the box and the spot is sampled uniformly between 0 and X, where X increases progressively as the artificial pigeons become more adept at solving the task: pigeons "graduate" to the next distance once they achieve good performance (100 successive successes in a maximum duration 50 timesteps each) on the task. Other than box-spot distance, the position of the various elements of the task is randomized for each shaping and test instance. For jump-and-peck, the box is fixed in place underneath a banana, and for push-extinction the box is present with no reward is available. The three shaping tasks are interleaved.

Artificial pigeons trained in this way did not succeed at the test on their first try in an "insightful" manner, unlike real pigeons. Instead, we present results for repeated tests, in which, after training, the pigeons face a succession of randomized test problems (with the box and banana placed randomly).

"Pigeon Insight" Model

Learning is modeled using deep Reinforcement Learning (Sutton & Barto, 2018), specifically an actor-critic algorithm. Reinforcement Learning is learning what to do in order to maximize a reward signal, where obtaining a reward often requires multiple successive actions. To know whether an action was good, it is therefore useful to evaluate the resulting situation, without waiting for the reward itself: if the new situation is promising (as opposed to dire), the tendency to repeat that action in similar contexts should be reinforced (as opposed to weakened). Many reinforcement learning algorithms exploit these ideas by making use of an actor which selects actions, and a critic which evaluates situations and generates a learning signal.

A technical description of these ideas and their implementation is given below in order to make the present work reproducible. Readers who wish to familiarize themselves further with Reinforcement Learning are encouraged to consult the article by Kaelbling, Littman, and Moore (1996) or the more expansive book by Sutton and Barto (2018). For a discussion of the connections between Reinforcement Learning approaches in AI and in psychology, see chapters 14 and 15 of Sutton and Barto (2018, accessible online).

The simulated environment is a Markov Decision Process, where images count as states s from a set $\mathcal{S}$ ($s \in \mathcal{S}$), pigeon behavior as actions $a \in \mathcal{A}$, with rewards $r \in \mathcal{R}$ (10 on successful completion, 0 otherwise), and a transition function $p : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0,1]$ defining the dynamics of the environment. In an actor-critic algorithm, the agent, with no prior knowledge of the environment dynamics, learns from experience a policy $\pi_\theta : \mathcal{S} \times \mathcal{A} \to [0,1]$ (mapping states to a probability of selecting each action, based on the parameters θ of the *actor*) and a value function $v_w : \mathcal{S} \to \mathbb{R}$ (which denotes the agent's future prospects, or return, assuming it follows its policy from the current state; it is approximated as $\hat{v}_w$ based on parameters w of the *critic*). Actor-critic systems are considered more plausible models for biological agents (Sutton & Barto, 2018, pp395-402).

Two convolutional neural networks are used to approximate the value function v as $\hat{v}_w$ (critic network) and to implement the policy (actor network). The architecture is shown for the actor network in figure 3; the critic network is identical save for the last layer, which has only one output and no nonlinearity. Learning proceeds online by gradient descent, according to the update rules:

$$w \leftarrow w + \alpha_w \delta \nabla \hat{v}_w(S')$$
$$\theta \leftarrow \theta + \alpha_\theta \delta \nabla \log \pi_\theta(A|S)$$

Where S is the state, A is the action chosen (according to the policy π), R is the reward, S' the following state, and $\delta = R + \gamma \hat{v}_w(S') - \hat{v}_w(S)$ is the one-step time-difference error. We use a discount γ (0.9) and learning rates α_w and α_θ (0.003 and 0.0003). Thus, by way of the time-difference error, the critic adjusts its estimate of the value of a state based on that of the next state. Meanwhile, the actor learns to preferentially select actions which lead to surprisingly high-valued states (states with positive time-difference errors). The interleaved processes of estimating the value of states and improving the policy leads (demonstrably under certain conditions) to a locally optimal policy. In our implementation, the actor was regularized based on the entropy of its output to ensure continued adequate exploration (as in Mnih et al. (2016)), and learning and acting was parallelized (16 concurrent agents) to accelerate computation time.

A first cohort of 20 agents was given shaping training up to a performance of 90% completion within 50 time-steps, and then continued learning in the test condition; we call this *condition 1*. A second cohort of 20 was given more extensive training (150,000 additional timesteps after meeting the criteria for condition 1); we call this *condition 2*. The expectation was that additional training would result in overfitting and render transfer more difficult (as observed for human insight in the work of Wiley (1998)). A third cohort was directly given the test without any prior training; we call this *condition 3*. In all cases, the primary measure is the rate of success: how likely each simulated pigeon is to succeed at its task within 50 time-steps. This is measured as a running average (cf. figure 4).

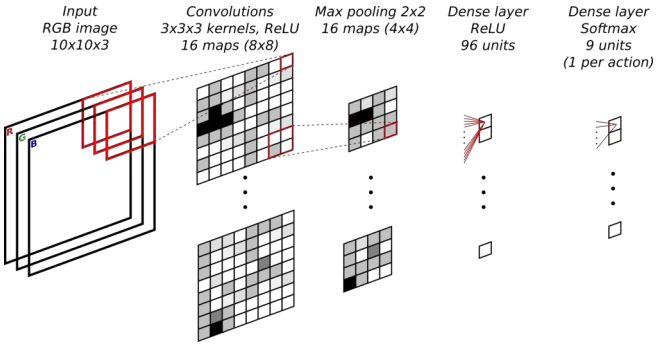


Figure 3: The neural network architecture used for the actor. For illustrative purposes, example activations are given in shades of grey, and example connections in red.

Results

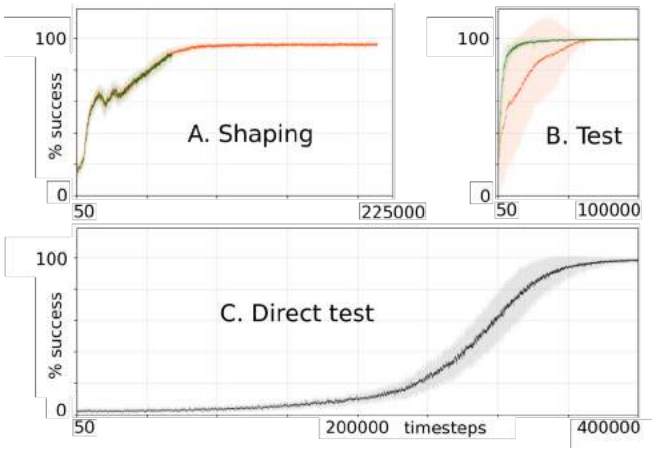


Figure 4: Performance of the actor-critic model. All graphs show the success rate for 20 runs, smoothed over 100 timesteps; color bands show the standard deviation. Note that the success rate is initially high during shaping because the shaping tasks are easy in the beginning, and progressively made more difficult as performance increases. Graphs A and B show the performance for conditions 1 (dark green) and 2 (lighter orange) for the shaping and test, whereas Graph C shows the performance for condition 3 (naive agents). Condition 2 had worse average performance on the test, with greatly increased variance.

The shaping program was successful in improving performance. Agents in conditions 1 and 2 transferred successfully to the final task, rapidly learning the new task in condition 1, although there was often a delay for those of condition 2 who had been given more extensive training. Agents in condition 2 showed considerable variance in the transfer - some of them necessitating a much longer time than others. Condition 1 and 2 both showed substantially better performance than condition 3 on the test. These results are shown in figure 4.

In condition 1, agents adapted rapidly to the new task. However, in condition 2 there often was a period of "impasse" during which the agents displayed low performance; individual curves are shown in figure 5. These impasses remained short compared to condition 3, but were substantial compared to condition 1 (see figure 5b); impasse was followed by a rapid performance increase, which was accompanied by an increase in expected value as estimated by the critic components of the agents. There was also an increase in positive time-difference errors, which correspond to unexpected progress, from the agent's perspective.

Discussion

Did the simulated pigeons experience "insight"? Unlike the real pigeons, few solved the test situation on their first try, suggesting that out-of-the-box RL is not sufficient for insight. However, especially for condition 2, they displayed patterns that are reminiscent of findings on the insight process. Recall the characteristics of insight enumerated in the background section. Many of them are reflected in the behavior of the deep RL agents:

1. The insight sequence: in condition 2 especially, one can distinguish a fruitless search/impasse phase from a sudden resolution.

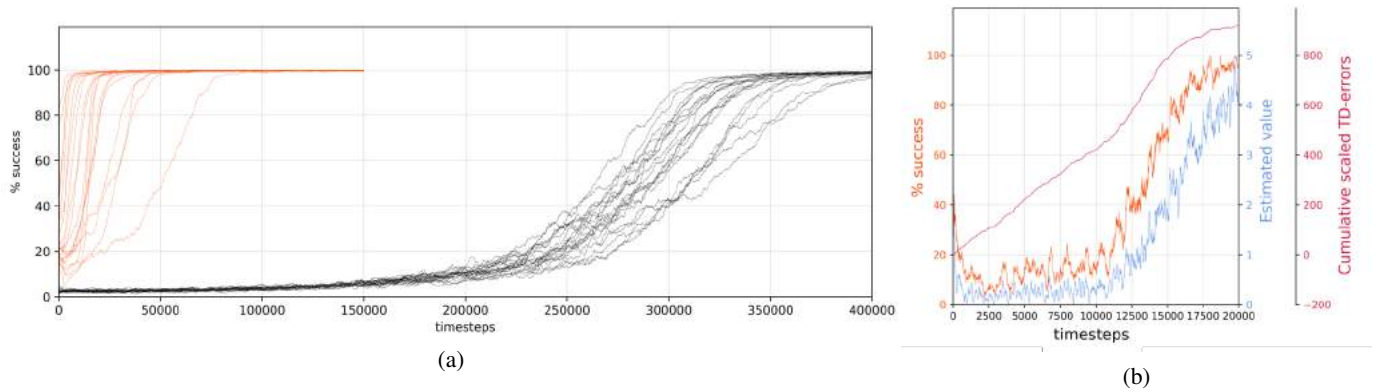
2. Sudden and surprising solution: the sudden increase of "feelings of warmth" in humans Metcalfe and Wiebe (1987), i.e. their subjective appreciation of how close they are to solving the problem, resembles the sudden increase of the estimated value function in the agents. (Recall that the value function, estimated by the critic component of the agents, measures their expectation of acquiring reward; it is thereby analogous to the "feelings of warmth" measure.) The steepness of the learning curve for shaped agents (conditions 1 and 2) is sudden by comparison to naive agents (condition 3).

3. Restructuring: the agents ought to behave "as if" the yellow objective is the green spot with which they trained. We conjecture that when the agent learns this, the rest of the correct solution "falls into place" rapidly due to prior learning[4].

4. Role of experience: "insight" is made possible by prior experience, with extensive experience having an ambiguous role – too much experience being detrimental to performance, as in Wiley (1998).

Additionally, we note several associations between the concepts of reinforcement learning and those of psychology, which are known in RL and cognitive psychology, but have received little attention in the insight literature. Readers familiar with RL may have recognized transfer and curriculum learning techniques used for instance in robotics; those well-read in psychology noticed that the overfitting of condition 2

[4]The distributed nature of neural networks makes this difficult to verify; we reserve such investigations to future work.



(a) (b)

Figure 5: **(a)** All "overfitting" transfer curves (orange, left), compared to learning from scratch (black, right), over 20 runs. (All curves have been smoothed for readability, showing the average over 4000 timesteps.).
(b)A single learning curve on the test (one of the 20 shown in orange in subfigure (a)). The top curve is the cumulative TD-error, the middle curve is the % of success, the bottom curve is the estimated value.

is reminiscent of the Einstellung effect, by which prior experience gets in the way of finding an optimal solution to a new problem (Luchins, 1942).

Thus although the artificial pigeons needed a considerable amount of interaction with the test by trial and error (note that both pigeons (Epstein et al., 1984) and chimpanzees (Köhler, 1921; Birch, 1945) also showed some amount of trial and error even during the test), they also presented learning patterns resembling those of insight: namely (1) a comparatively sudden increase of performance, accompanied by (2) an increase in expected return, which (3) was made possible by a "just-right" amount of prior experience.

The proposed model thus displays some characteristics of insight while being limited in other respects. The most notable of these limitations is the time needed to discover the full solution during the test. This might be a matter of learning quickly from limited data during the test (this is the solution favored by Epstein (2014)), or of making use of more profound regularities in the shaping tasks, e.g. via temporal abstraction as suggested by Colin, Belpaeme, Cangelosi, and Hemion (2016). Alternatively, they might identify new regularities between old and new tasks on the fly (Friston et al., 2017), or use off-policy learning to make use of prior experience (as suggested by Richard Sutton in personal communication; cf. Tolman and Honzik (1930)). Finally, perhaps the use of model-based reinforcement learning allows for trial and error to occur in subconscious simulation "in the agent's mind" (Hamrick et al., 2016; Hélie & Sun, 2010). These various approaches are not mutually exclusive - indeed, all of them are compatible, and perhaps only some (yet-to-be-realized) combination of all of these methods can produce behavior truly comparable to animal and human insight.

Conclusion

Insight problem-solving was historically presented by Köhler as a challenge for Thorndike's concepts of animal learning. Nowadays Aha!-moments, due to the sheer speed of the phenomenon in human beings and animals, remain puzzling for modeling approaches that rely on statistical trial-and-error. However, their apparent reliance on learning and thereby generalization, and their representational component, has made them equally challenging for traditional cognitive models. Both symbolic and statistical approaches have difficulty explaining insight.

We suggest that the statistical approaches offer, after all, a promising avenue of research for explaining insight. The established importance of learning for insight (Birch, 1945; Wiley, 1998) suggests a model based on learning. Our results show how transfer learning can accelerate the resolution of a new problem to the point of making it seem, in contrast to solving it "from scratch", rather sudden. This and the focus of contemporary machine learning techniques on representation designates them as clear candidates for modeling insight.

We have presented a simulation of a psychological experiment on insight, with the aim of proposing a model of the cognitive processes underlying animal behavior in the experiment. Our artificial pigeons were not a match for the real pigeons performance-wise: they required more experience to solve a simplified version of the task; their "insights" were slower and clumsier. However the proposed model showed qualitative properties reminiscent of those seen in pigeons. It is a long way to recreating the insights of chimpanzees, let alone humans; we have given some directions for future research, and we hope that the methodology presented here (replicating insight studies on non-human animals) can serve as a basis for future investigations of the creativity of great apes - such as ourselves.

Acknowledgments

This work was completed as part of Marie Curie Initial Training Network FP7-PEOPLE-2013-ITN, CogNovo, grant number 604764. We would like to thank Dr. Robert Epstein for

his helpful comments, and the reviewers whose constructive criticism helped make this a better article.

References

Birch, H. G. (1945). The relation of previous experience to insightful problem-solving. *Journal of Comparative Psychology*, *38*(6), 367–383.

Bowden, E. M., & Jung-Beeman, M. (2003). Normative data for 144 compound remote associate problems. *Behavior Research Methods*, *35*(4), 634–639.

Colin, T. R., Belpaeme, T., Cangelosi, A., & Hemion, N. (2016). Hierarchical reinforcement learning as creative problem solving. *Robotics and Autonomous Systems*, *86*, 196–206.

Cook, R. G., & Fowler, C. (2014). "Insight" in pigeons: absence of means–end processing in displacement tests. *Animal cognition*, *17*(2), 207–220.

Epstein, R. (2014). On the orderliness of behavioral variability: Insights from generativity theory. *Journal of Contextual Behavioral Science*, *3*(4), 279–290.

Epstein, R., Kirshnit, C., Lanza, R., & Rubins, L. (1984). "insight" in the pigeon: antecedents and determinants of an intelligent performance. *Nature*, *308*, 61–62.

Friston, K. J., Lin, M., Frith, C. D., Pezzulo, G., Hobson, J. A., & Ondobaka, S. (2017). Active inference, curiosity and insight. *Neural computation*, *29*(10), 2633–2683.

Hamrick, J. B., Pascanu, R., Vinyals, O., Ballard, A., Heess, N., & Battaglia, P. (2016). Imagination-based decision making with physical models in deep neural networks. In *Proceedings of the NIPS 2016 workshop on intuitive physics.*

Hélie, S., & Sun, R. (2010). Incubation, insight, and creative problem solving: a unified theory and a connectionist model. *Psychological review*, *117*(3), 994–1024.

Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of artificial intelligence research*, *4*, 237–285.

Kaplan, C. A., & Simon, H. A. (1990). In search of insight. *Cognitive psychology*, *22*(3), 374–419.

Knoblich, G., Ohlsson, S., & Raney, G. E. (2001). An eye movement study of insight problem solving. *Memory & Cognition*, *29*(7), 1000–1009.

Köhler, W. (1921). *Intelligenzprüfungen an menschenaffen [the mentality of apes]*. Berlin: Springer-Verlag.

Kounios, J., & Beeman, M. (2015). *The eureka factor: Creative insights and the brain*. Random House.

Luchins, A. S. (1942). Mechanization in problem solving: The effect of einstellung. *Psychological monographs*, *54*(6), i–95.

MacGregor, J. N., Ormerod, T. C., & Chronicle, E. P. (2001). Information processing and insight: A process model of performance on the nine-dot and related problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(1), 176–201.

MacLellan, C. J. (2011). An elaboration account of insight. In *AAAI fall symposium: Advances in cognitive systems* (pp. 194–201).

Metcalfe, J., & Wiebe, D. (1987). Intuition in insight and noninsight problem solving. *Memory & cognition*, *15*(3), 238–246.

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., ... Kavukcuoglu, K. (2016). Asynchronous methods for deep reinforcement learning. In *Proceedings of the 33rd international conference on machine learning* (Vol. 48, pp. 1928–1937).

Ohlsson, S. (2011). *Deep learning: How the mind overrides experience*. Cambridge University Press.

Poincaré, H. (1909). *Science et méthode*. Flammarion.

Rothenberg, A. (1995). Creative cognitive processes in Kekulé's discovery of the structure of the benzene molecule. *The American Journal of Psychology*, *108*(3), 419–438.

Schilling, M. A. (2005). A" small-world" network model of cognitive insight. *Creativity Research Journal*, *17*(2-3), 131–154.

Shettleworth, S. J. (2012). Do animals have insight, and what is insight anyway? *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, *66*(4), 217-226.

Sprugnoli, G., Rossi, S., Emmerdorfer, A., Rossi, A., Liew, S.-L., Tatti, E., ... Santarnecchi, E. (2017). Neural correlates of eureka moment. *Intelligence*, *62*, 99–118.

Stephen, D. G., Boncoddo, R. A., Magnuson, J. S., & Dixon, J. A. (2009). The dynamics of insight: Mathematical discovery as a phase transition. *Memory & Cognition*, *37*(8), 1132–1149.

Sutton, R., & Barto, A. (2018). *Reinforcement Learning: An Introduction*. MIT Press. (Accessible at `http://incompleteideas.net/book/the-book-2nd.html`)

Thorndike, E. L. (1898). Animal intelligence: an experimental study of the associative processes in animals. *The Psychological Review: Monograph Supplements*, *2*(4), i–109.

Tik, M., Sladky, R., Luft, C. D. B., Willinger, D., Hoffmann, A., Banissy, M. J., ... Windischberger, C. (2018). Ultra-high-field fmri insights on insight: Neural correlates of the aha!-moment. *Human brain mapping*, *39*(8), 3241–3252.

Tolman, E. C., & Honzik, C. H. (1930). Introduction and removal of reward, and maze performance in rats. *University of California Publications in Psychology*, *4*, 257–275.

Wagner, U., Gais, S., Haider, H., Verleger, R., & Born, J. (2004). Sleep inspires insight. *Nature*, *427*(6972), 352–355.

Weisberg, R. W. (2015). Toward an integrated theory of insight in problem solving. *Thinking & Reasoning*, *21*(1), 5–39.

Wiley, J. (1998). Expertise as mental set: The effects of domain knowledge in creative problem solving. *Memory & cognition*, *26*(4), 716–730.

Epistemic drive and memory manipulations in explore-exploit problems

Nicolas Collignon n.collignon@ed.ac.uk
Christopher Lucas clucas2@inf.ed.ac.uk
School of Informatics, University of Edinburgh

Abstract

People often navigate new environments and must learn about how actions map to outcomes to achieve their goals. In this paper, we are concerned with how people direct their search and trade off between selecting informative actions and actions that will be most immediately rewarding when they are faced with new tasks. We find that some people selected globally informative actions and were able to generalize from few observations in order learn new reward structures efficiently. These participants also displayed the ability to transfer knowledge across similar tasks. However, a consistent proportion of participants behaved sub-optimally, caring more about observing novel information instead of maximizing reward. Across four experiments, we present evidence that participants' motivation to explore was influenced by 1) how much they already knew about the underlying task structure and 2) whether their observations remained available. We discuss possible explanations behind people's exploratory drive.

Keywords: active learning; generalization; exploration-exploitation; transfer learning; data-availability;

Introduction

In order to act, plan, and achieve goals, people must learn about their environment and the outcome of possible actions. One reason for human successes in developing new theories and strategies when confronted with new problems is that people are not passive observers. Indeed, children ask informative questions and can adapt their strategies when inquiring about things they don't know (Ruggeri & Lombrozo, 2014), and play with new toys in ways that help them disambiguate uncertain causal relationships and gather information (L. Schulz & Bonawitz, 2007; Cook et al., 2011). The idea that humans learn and interact with their environment by performing intuitive experiments, maximizing information gain, is a popular one (Coenen et al., 2017; Gureckis & Markant, 2012; Nelson, 2005; Gopnik et al., 2004).

In this paper, we are interested in how people learn to select actions that are most rewarding when faced with a sequence of novel but potentially related tasks. We designed experiments to better understand people's exploration and reward maximizing strategies across a sequence of tasks. Do those strategies evolve over time, as they encounter related tasks? Can people transfer structural knowledge and improve their performance by leveraging similarities between tasks? What is the relationship between people's search strategies, their ability to learn and generalize from observations, and how well they do?

When encountering new situations, people are often faced with the decision of either gathering more information about the task to improve the quality of their decision, or choosing an action that has been shown to be rewarding (Hills et al., 2015). A doctor might, for example, want to run more tests to have a better diagnosis for their patient or give them the treatment they believe will best relieve them from their symptoms. To better understand human decision strategies when dealing with the explore-exploit trade-off, Multi-armed Bandits (MAB) have been used extensively. In these experiments, participants have to select between different possible actions (e.g. the arms of a bandit) yielding stochastic rewards, so as to maximize their rewards. In the real world, an essential part of solving problems lies in discovering the underlying structure of the problem, where each action can be represented as a set of continuous and discrete features. In a Contextual MAB (CMAB), there are observable feature that provide information about the arms' reward distributions. Learning how features relate to rewards allows for an efficient representation of the environment, and enables the learner to generalize to new events. Previous studies of human behavior in CMAB problems have shown that people are able to generalize across observations when faced with a large number of options, and make use of uncertainty to direct their search (E. Schulz et al., 2017; Wu et al., 2018; Borji & Itti, 2013). These experiments have assumed the basic structure of the underlying problems to be static, or known in advance. When confronted with unknown task structures, Teodorescu and Erev (2014) showed that people were able to adaptively learn purely exploratory or purely exploitation-oriented policies. However, in their experiment there was no systematic relationship between an option's features and its reward, aside from whether it had been previously explored.

Unlike a CMAB-type task, the tasks we presented to participants were deterministic, meaning that re-selecting an option would always yield the same reward. This was done to ensure a clear distinction between exploration and exploitation in participant decisions. To examine people's ability to use generalization to guide their search we presented them with tasks that contained a large number of choices and a relatively limited number of actions, meaning that generalizing over previous observations is necessary for optimal performance. We chose a simple structure to ensure it would be possible for participants to learn and exploit it when maximizing rewards.

Our first two experiments focus on sequential tasks where participants had no prior information about the underlying reward structure, and where a combination of exploration – to discover task structure and discover optima – and exploitation is necessary to do well. The next two experiments provided participants with training about the reward structures before the task itself. In all of these experiments, we found that some participants selected actions that resolved uncertainty about the underlying structure of the task, and traded off between exploration and exploitation in order to maximize

reward. These participants were also able to transfer knowledge across tasks and gradually improved their performance. We also found a significant number of participants engaged in purely exploratory behavior, consistently preferring to choose novel actions, even when these actions were relatively unrewarding. These results highlight the importance of studying individual differences to better identify the multiple factors that influence human behavior, and of accommodating these differences in models of learning and exploration.

Experiment 1

Across our four experiments participants were given a sequence of grids composed of 9-by-9 arrays of tiles (see Figure 1), with each tile corresponding to a possible action. In this paper, we limit our analysis to the first three grids presented to participants (out of nine), as the latent task structure changed after that point. The grids studied here shared a similar underlying task structure: they had the same kind of relationship between features and rewards, but details of those relationships varied. In our experiment an action consists of selecting an individual tile, which has two features: its horizontal (x), and vertical position (y). Participants had to select tiles to maximize their cumulative rewards over 20 choices in each grid. The task presents a classical explore-exploit tradeoff: Succeeding requires carefully balancing between choosing new tiles to learn about the underlying reward structure or re-selecting tiles that were observed to be rewarding. In Experiment 1, participants received no prior knowledge about the reward structure of the tasks, nor about whether the tasks were related to one another in any way.

We predicted participants would be able to generalize from previous observations and improve by using their growing knowledge of the underlying task structure to select better actions. We measure this by looking at whether participants were able to select more rewarding tiles as they collected more information, and whether they demonstrated confidence in their knowledge by repeatedly selecting (i.e., exploiting) optimal actions. Our second hypothesis was that participants would be able to re-use knowledge across grids, since they shared the same structure, and thus improve their performance from one grid to the next.

We also studied the distance between participants' selections throughout the task to better understand their behavior. Distance between selections is a useful marker of different exploration strategies. For example, participants who seek to reduce uncertainty about the task structure are likely to select tiles that are far apart from each other, as these tend to yield more information about the broad shape of the reward function, in addition to having more uncertain rewards themselves. We call these selections *globally informative* actions. In contrast, participants might sample tiles adjacent to their previous observations, e.g., because they believe they are close to a maximum or because they want to observe local gradients. We call this kind of selection *local search*.

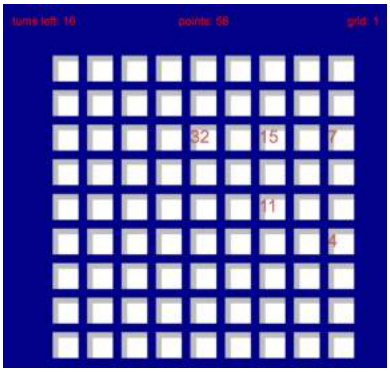


Figure 1: Screenshot of grid presented to participants after 5 observations. Note that in Experiment 1, the rewards disappear shortly after a tile has been selected.

Methods We recruited 79 participants using Amazon's Mechanical Turk service. They received \$0.75-\$1, which was doubled for participants whose final scores were in the top 10 percent. Following the instructions given to participants, we excluded participants whose performance was worse than chance ($n = 3$). We also excluded participants who failed to select more than 2 different tiles on the majority of grids ($n = 5$), as it showed a lack of engagement with the task.

The three grids analysed here used a reward structure where one location (x_m, y_m) was sampled uniformly at random in each grid, and the grid's maximum reward m was sampled from $(\mathcal{N}(\mu = 200, \sigma^2 = 50))$. The reward r for a given tile location (x, y) was exponentially decreasing with its Euclidean distance d from that maximum-reward tile: $r(x,y) = C \cdot e^{-k \cdot d((x,y),(x_m,y_m))}$, rounded to the nearest integer. We chose an exponential relationship between features and rewards to ensure there would be a clear advantage for participants who discovered the maximum-reward tile. We chose a constant $(k = 0.4)$ that led to large differences between the maximum and its closest neighbors while making it unlikely that any tiles would have rewards of zero or one. We used a random maximum reward in order to make it difficult for participants to know they had found the most rewarding tile without knowing the reward structure of the task.

When a tile was selected, the reward was displayed on the tile for 1.5 seconds and added to the cumulative score on the current grid. Participants could re-select tiles they had previously chosen. Participants were given no information about the underlying structure of the grid prior to the task, and were not informed that the tasks were related in any way, apart from a note that there could be patterns behind the rewards.

Results and Discussion For this and all subsequent experiments, we report the normalized scores (between 0 and 1), by dividing each reward by the maximum reward in that grid. We were first interested in seeing whether participants were able to recognize similarities between tasks. We use a general linear model (GLM), with the reward as outcome variable. The turn and grid index were used as predictor variables. Both the turn ($b = 0.02, se = 0.001, p < 0.001$) and the grid ($b = 0.05, se = 0.005, p < 0.001$) were significant factors. Following our hypothesis, participants selected better

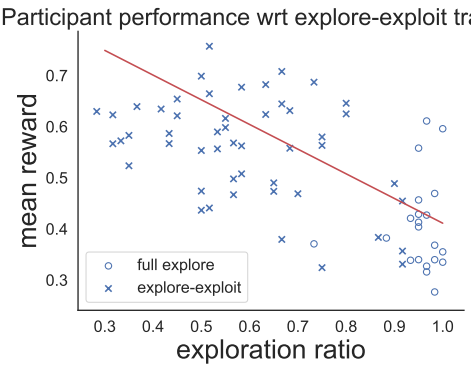


Figure 2: Each point represents a participant. The y-axis is the average reward across all three grids. The x-axis is the proportion of novel selections across all three grids. A value of 1 would mean only selecting new tiles, 0 only selecting the previously-selected tiles.

tiles over time, suggesting that they were able to exploit the underlying reward structure. Participants also improved their performance across grids, suggesting they were able to transfer structural knowledge across tasks (see Figure 3).

As a simple measure of a participant's propensity to explore, we used the proportion of actions that selected a previously-unseen tile ("exploration") versus re-selecting a previously-seen tile ("exploitation"). This distinction is more natural in our tasks than in a traditional stochastic bandit task, as in the latter it can be informative to re-select previously-seen tiles to learn about their reward distributions. There were significant behavioral differences indicated by how people traded off between exploration and exploitation among participants, and in the cumulative rewards they collected ($M = 0.49, SD = 0.30$) (see Figure 2).

Twenty-two participants (31 percent) never re-selected tiles more than twice in the majority of grids. We call these participants *full explore* (FE) participants. We call the other participants (n=49), that traded off exploration and exploitation, *Explore-Exploit* (EE) participants.

EE participants improved across tasks ($b = 0.07, se = 0.006, p < 0.001$) (see Figure 3), supporting our hypothesis that participants who used the underlying task structure to direct their search and maximize reward were able to re-use what they had learned to a new task.

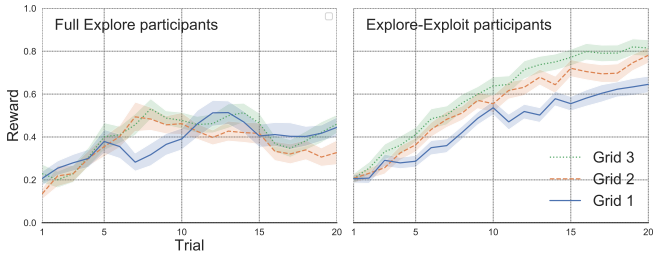


Figure 3: Performance of FE participants (n=22) and EE participants (n=49) in Experiment 1 across all three grids. Error bars in this and all subsequent plots reflect standard errors of the mean.

Across all participants, the proportion of exploratory selections correlated negatively with score ($r(140) = -0.71, p < 0.001$), and FE participants earned lower scores than EE participants ($t(69) = 5.77, p < 0.001, d = 0.15$). Their average scores barely improved from one grid to the next (Figure 3; $b = 0.02, se = 0.008, p = 0.06$).

We used a logistic regression model to evaluate participants' ability to find the maximum across grids. More participants found the maximum as they went on with the grids, hinting that they were better at utilising the underlying task structure ($b = 0.64, se = 0.11, p < 0.001$). Whether participants were engaging in *full exploratory* or *explore-exploit* strategies did not predict if they found the maximum in the tasks ($b < 0.001$). Participants were significantly better than chance at finding the maximua (0.65 of grids, vs. upper bound chance proportion of 0.25; $\chi^2(1, N = 1174) = 188.1, p < 0.001$). Furthermore, participants had overall a strong 'local bias' in their sampling, where they choose tiles close to their last choice more often than chance given the distribution of inter-tile distances ($t(151) = -50.8, p < 0.001, d = -2.34$) (see Figure 4). This suggests that participants engaged in local search strategies, rather than globally informative actions. Both EE and FE groups showed this bias, with adjacent tiles selected in 49% of FE participants' exploratory choices ($SD = 0.17$) and 39% for EE participants ($SD = 0.17$).

In conclusion, Experiment 1 showed that some participants were able to learn the underlying task structure when it was new and traded off between exploration and exploitation to maximize their rewards. These participants transferred knowledge across tasks that shared similarities in their underlying structure. However, a large proportion of participants had a strong tendency to explore in circumstances where exploitation would have yielded much higher scores, preferring unobserved tiles over known tiles with a high reward value. FE participants presented some evidence for learning the underlying structure, but this was not reflected in their score. Why did so many participants adopt such an extreme exploratory policy? One possibility is that they were motivated to learn more about the reward structure, or ensure they had found the maximum possible reward, in line with the inherent curiosity bias observed in people (Kidd & Hayden, 2015; Gottlieb et al., 2013).

We also observed a locality bias in participants' choices. This may have been due to the memory demands of the task. Wu et al. (2018) presented evidence that participants displayed an ability to use generalization to direct their search. Unlike the task used in their study, our task had the rewards disappear after participants selected a tile. Remembering past observations when generalizing might be difficult, and could have led participants to adopt policies that alleviated the complexity of the task. For example, if participants tracked local gradients in rewards and followed increasing rewards, this would only require tracking 2-3 past observations while being less demanding than computing a surrogate model over the general task structure. This would be consistent with the local search strategies exhibited in other domains such as causal learning (Bramley et al., 2015) and category learning (Markant et al., 2016), and the idea that people adapt their high-level strategies to make the most of limited resources (Lieder et al.,

2014). For FE participants, the local bias during exploration could reflect a systematic and memory-efficient policy for exhaustively searching a subset of the tiles for a maximum.

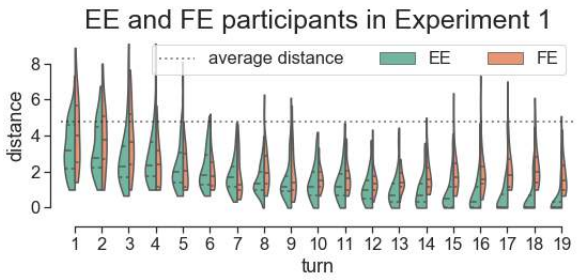


Figure 4: Average distance between selections of EE and FE participants in Experiment 1 presented with quantiles and kernel density estimations. We use Euclidian distance between selections, with 0 counting for a re-selection of the previous click. The dotted line represents the average distance between all tiles in a grid. The shape of the distribution is drawn using a (normal) Gaussian Kernel Density Estimate cut at 0.

In Experiment 2, we presented participants with the same task structure as in Experiment 1, but with changes designed to understand and potentially reduce their strong tendency to explore new tiles. These included persistent indicators of explored tiles' rewards, checks of participants' understanding of the instructions, and different incentives.

Experiment 2

In this experiment, the reward associated with a given tile is displayed continuously once it has been observed. We hypothesized that with participants observations remaining visible, the overall reward pattern would be more evident. We predicted that participants would be able to make more globally informative actions (i.e. exploratory selections would be more distant from each other). Because the underlying structure is made more evident, we also assumed fewer participants would engage in *full exploratory* behavior.

Methods We recruited 72 participants using Amazon's Mechanical Turk service identically to Experiment 1. Participants all received a base payment of $0.75. The reward scheme differed from that in Experiment 1: rather than granting bonuses to the top 10 percent, we gave all participants a bonus proportional to their cumulative score, up a maximum of $0.75. We excluded participants who failed to select more than 2 different tiles on the majority of grids ($n = 4$). In Experiment 2 when a tile is selected by a participant the reward is continuously displayed on the tile and is added to the current cumulative score on the current grid.

In another change from Experiment 1, tiles' rewards were persistently visible after they had been selected, under the logic that it might improve participants' ability to learn the underlying reward structure and increase their ability to find and exploit the maximum. We also added explicit instructions that participants could re-select tiles, and added a pre-task questionnaire to make sure participants understood these instructions. The questionnaire also required participants to understand that their goal was to maximize reward (as opposed to discovering the underlying pattern, or finding the maximum. Participants were not allowed to proceed with the task until they answered all questions correctly.

Results and Discussion Contrary to our predictions that participants would be less prone to *full exploratory* behavior, a significantly larger proportion of participants showed FE behavior in Experiment 2 as compared with Experiment 1 (.47, $n = 32$ vs. .31, $n = 22$; $\chi^2(1, N = 139) = 18.6, p < 0.001$). As in Experiment 1, the proportion of exploratory selections correlated negatively with performance ($r(134) = -0.75, p < 0.0001$). In Experiment 2, EE participants also performed significantly better than FE participants ($t(66) = 9.31, p < 0.0001, d = 0.23$) and improved significantly across tasks ($b = 0.04, se = 0.007, p < 0.0001$), whereas FE participants did not ($b = 0.01, se = 0.006, p = 0.14$).

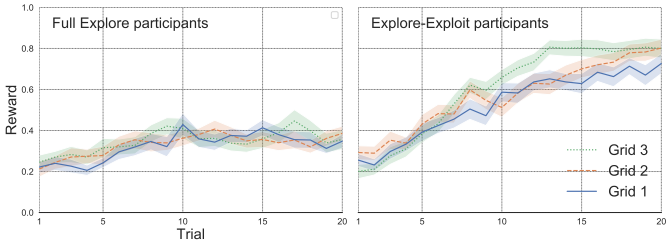


Figure 5: Performance of FE participants ($n = 49$.) vs EE participants ($n = 36$).

To understand the effect of having observations available throughout the task, we compare the performance of EE participants in Experiment 2 (n=36) to the performance of EE participants in Experiment 1 (n=49). Overall, EE participants in Experiment 2 (M=0.58) did slightly better than EE participants in Experiment 1 (M=0.56) ($b = 0.04, se = 0.008, p < 0.001$). This was most pronounced in the first grid ($t(84) = 2.18, p = 0.03, d = 0.08$). We conjecture that EE participants in Experiment 2 learned the reward pattern faster, and EE participants caught up in subsequent grids. This supports the hypothesis that visible observations allowed participants to generalize better, by supporting more global strategies. To test this idea, we looked at the inter-selection distances between the first 5 selections of participants. EE participants' choices in Experiment 2 were more global, with greater distances than EE participants' choices in Experiment 1 ($t(84) = -2.25, p = 0.03, d = 0.66$) (see Figure 6).

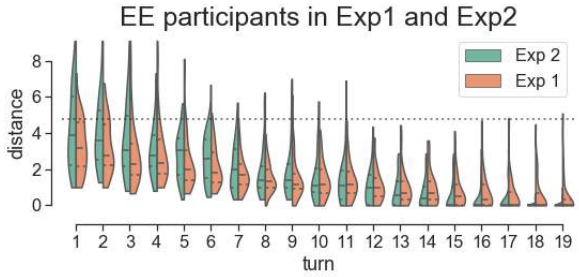


Figure 6: Comparison of distances between selections of EE participants in Experiment 1 and Experiment 2 (see Figure 4 for details). EE participants in Experiment 2 selected more "global" actions (longer distances between selections) during their first actions.

Why did more participants engage in FE behavior in Ex-

periment 2? We conjectured that participants were more motivated to observe rewards for new tiles when previous rewards remained visible, because the overall pattern – and the possibility of better understanding it – might have been more salient to them.

In Experiment 3, we sought to understand why some participants might want to select new tiles almost exclusively, rather than occasionally exploiting what they had learned to earn greater rewards. After Experiment 1, we hypothesized that this might have been due to an intrinsic epistemic drive in participants. Experiment 2 showed that for EE participants were able to leverage visible observations to conduct more global exploration, and led to a better overall performance. However, the observable rewards also seemed to add an incentive for many participants to exclusively choose novel actions, rather than maximising rewards. We hypothesized that this would only be the case for new tasks when participants had no prior knowledge about the underlying reward structure of the tasks, since new observations would not be very informative if participants had a prior about the underlying reward structure.

Experiment 3

We designed Experiment 3 to control explicitly for the potential epistemic drive of FE participants by familiarizing them with the underlying reward structures prior to the task. By making the structure clear to participants prior to the tasks, our primary prediction for Experiment 3 was that fewer participants would engage in FE behavior. We presumed the intrinsic motivation of observing new observations would be attenuated when participants did not gain new information about the task from those observations.

We also hypothesized there would be weaker or no progress across grids since participants would already be familiar with the reward structure when they engage with the first grid. Because of the training, we predicted participants would be more efficient at finding and re-selecting tiles with high values, and would thus perform better overall than in Experiment 1 and 2. Experiment 3 was set up identically to Experiment 2. Participants were told about the underlying pattern and given three practice grids so they could learn the reward structure prior to the task.

Methods We recruited 43 participants using Amazon's Mechanical Turk service, identically to Experiment 2, with the following changes: Participants were only recruited for three grids rather than nine, following the same reward pattern discussed in Experiment 1 and Experiment 2. Because of the shorter duration, participants were paid a base reward of \$0.2. We used a proportionally larger bonus of \$0.6 under the logic that this would further reduce the effects of epistemic drive. Apart from the training grids presented prior to the task, instructions were identical to Experiment 2. During the training, participants were told that each grid had one maximum tile, and the closer a tile is to the maximum the higher the reward. The first training grid had all rewards displayed and participants were instructed to familiarize themselves with the nature of the task. The next two grids were similar to the grids in the actual task (i.e. only observed tiles display reward values) but participants were encouraged to learn the pattern as well as they could. Throughout the task, instructions regarding reward maximisation and the possibility of reselecting tiles were also displayed. We excluded one participant who failed to select more than two different tiles on the majority of grids and one participant who reported not following the instructions upon completing the experiment.

Results and Discussion Surprisingly, 37 percent (15 out of 41) of participants engaged in *Full Exploration* (FE) in Experiment 3. The proportion of FE participants in Experiment 3 was significantly less than the 47 percent we observed in Experiment 2 ($\chi^2(1, N = 109) = 8.82, p = 0.003$), but was nonetheless a higher proportion than anticipated.

As expected, EE participants in Experiment 3 did not improve significantly across grids, since they had been trained extensively on the rule before the assessed task started ($b = -0.01, se = 0.008, p = 0.112$). The average performance of EE participants was significantly better than EE participants in Experiment 2 ($t(61) = 2.29, p = 0.03, d = 0.07$) and EE participants in Experiment 1 ($t(74) = 3.11, p = 0.003, d = 0.09$, suggesting that participants were able to learn the rule during the training and relied on this knowledge when faced with new grids in the task.

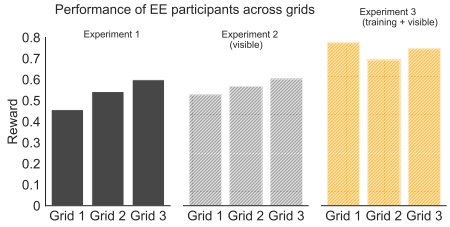


Figure 7: Average performance of EE participants (participants that traded off between exploration and exploitation) across all three grids in Experiment 1, 2 and 3.

To understand the effect of prior knowledge on participants' exploratory patterns, we compared how EE participants explored compared to EE participants in Experiment 2. Participants in Experiment 3 were significantly more locally biased in their initial five selections ($t(359), p < 0.001, d = 1.19$). Participants in Experiment 3 were already familiar with the *Location rule*, and it is probable that they were able to find the maximum by ascending towards the maximum through small incremental steps. EE participants in Experiment 3 had a significantly lower proportion of reselections (0.19 in Experiment 3 vs 0.28 in Experiment 2) ($\chi^2(1, N = 1367) = 17.16, p < 0.001$). Given their higher performance scores, EE participants in Experiment 3 were likely to have a strategy more adapted to the task than in Experiment 2, where participants were still learning the reward structure. Indeed, EE participants in Experiment 2 had a tendency to settle on a sub-optimal tile, finding the maximum tile in 0.62 of grids. EE participants in Experiment 3 took smaller exploratory steps but found the maximum in 0.81 of the grids

($\chi^2(1, N = 185) = 6.69, p = 0.01$).

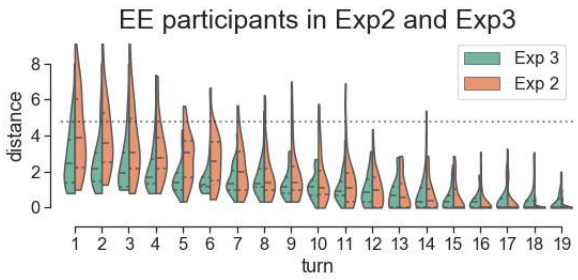


Figure 8: Distance between selections of participants (see Figure 4 for details). EE participants in Experiment 2 had more global observations than EE participants in Experiment 3. This can be explained by that fact that they had no prior knowledge about the task structure.

Contrary to our hypothesis, many participants still engaged in full exploratory behavior. Given this result, we hypothesized that participants might be motivated by observing new rewards rather than learning the underlying reward structure *per se* and that this effect might be emphasized when rewards remain visible after having been observed. Indeed, in Experiment 2, where rewards remained visible, significantly more participants engaged in full-exploratory behavior than in Experiment 1. We designed Experiment 4 to account for these two factors of epistemic motivation: 1) wanting to learn about the underlying reward structure and 2) wanting to attend novel information.

Experiment 4

Experiment 4 followed the design details of Experiment 3, except that rewards were not displayed continuously after they had been selected - they are displayed on the tile and disappear shortly after, like in Experiment 1.

Our main hypothesis for Experiment 4 was that fewer participants would engage in *full exploratory* behavior, since the epistemic reward is attenuated by not having the tiles visible after they have been selected and having training grids prior to the task. We predicted EE participants would perform similarly or slightly worse than in Experiment 3, because of the constraints of not having previous observations visible, but better than in Experiment 1 and 2. We also predicted we would observe little or no transfer effect across grids.

Methods 39 participants were recruited using Amazon Mechanical Turk. One participant was excluded for failing to select more than two different tiles, and one was excluded because their performance was worse than chance.

Results In agreement with our hypothesis, only one participant out of 37 engaged in *Full Exploration*. This was significantly less than in any other experiment. This supports the idea that participants' strategies were driven by an epistemic drive which was twofold:

First, participants were motivated to reveal the underlying reward structure, e.g., reducing the entropy about the structure of the task, or about the location of the maximum. Indeed, participants were less likely to engage in FE behavior in Experiment 4 (known structure and disappearing observations) than Experiment 1 (unknown structure and disappearing observations), and significantly less in Experiment 3 (known structure and visible observations) than Experiment 2 (unknown structure and visible observations).

Second, participants were motivated to observe the outcomes of individual actions. In Experiment 1,2 and 3 a significant proportion of FE participants selected the maximum but consistently opted for selecting novel options rather than re-selecting a previous maximum observation, with a preference for actions that were local to their last one. Participants' drive to select novel actions was enhanced by the fact that information did not need to be kept in working memory. They were less engaged in FE behavior in Experiment 1 (non-visible observations) than Experiment 2 (visible observations), and, similarly, less in Experiment 4 (non-visible observations) than Experiment 3 (visible observations). Though EE participants in Experiment 3 performed slightly better than in Experiment 4, this was not significant ($t(61) = 0.93, p = 0.35, d = 0.04$). Participants in Experiment 4 improved their average performance slightly across tasks ($b = 0.02, se = 0.007, p = 0.02$).

The average distance between the initial five exploratory selections of EE participants was not significantly different in Experiment 3 and Experiment 4 ($t(309) = -0.90, p = 0.37, d = -0.15$). EE participants in Experiment 4 explored significantly more locally than EE participants in Experiment 1 ($t(374) = -2.73, p = 0.007, d = 0.47$). Like in Experiment 3, this supports the hypothesis that participants who were familiar with the underlying structure of the grid were able to find the maximum by taking local exploratory steps until they eventually found the maximum.

Conclusion

In this paper, we focused on the behavioural analysis of participants across four experiments to study how people learn to select rewarding actions in a sequence of novel tasks. We found that some participants were able to learn the underlying structure while balancing exploration and exploitation to maximize their rewards across tasks. They improved their performance from one task to the next by transferring abstract knowledge about their environment. However, consistently across tasks, we observed that a significant proportion of participants engaged in purely exploratory behavior, largely ignoring the reward incentive. We showed that this behavior could be manipulated by controlling the availability of information as the learner selected actions, and by giving prior knowledge before participants engaged with the task. We suggest that people are motivated by two types of epistemic drives: 1) to reduce uncertainty and learn about the structure of the task and 2) to observe new evidence, regardless of its informativeness about the global task structure. The latter was evident when participants continued valuing new actions over maximising rewards, even when they were familiar with the task structure.

Different mechanisms for curiosity have been discussed in the literature, and could be connected to how people learn

in new environments when combined with trying to achieve goals or maximising utility. One such strategy is to entirely dismiss reward feedback, giving rise to a strong novelty drive. This *novelty search* mechanism has been been shown to be very successful in the context of Evolutionary Strategies for tasks with tricky reward functions (Lehman & Stanley, 2011). Some studies have shown that people are biased towards surprise (Gottlieb et al., 2013; Itti & Baldi, 2006). Selecting new actions would make sense under the assumption of possible change, or if one believes that the environment is trying to fool us. Third, the idea of *epistemic actions* could explain part of people's strong drive to select new actions, especially under the constraint of cognitive load, when storing observations is expensive or unrealistic. Epistemic actions refer to actions *in the world* that help solve problems by changing the mental state of the agent, as opposed to performing computations in the head (Kirsh & Maglio, 1994). An example of this behavior is the use of sticky-notes, or of arranging documents in a way that makes it easier to retrieve them rather than by memory alone. In the case of our experiment, observing new information might have been perceived as much cheaper than the possibility of generalizing from few observations.

In our study, we highlight that studying individual differences amongst participants can help us better understand the complex mechanisms at play during active learning in new environments. We hope that by pointing out surprising facets of human behavior, this empirical study can guide the design of better computational models of human learning and exploration. We are currently investigating how computational models of memory, generalization and search (León-Villagrá et al., 2018; Wu et al., 2018; Lucas et al., 2015) can give us further insight into people's representations and strategies when learning in new environments.

References

Borji, A., & Itti, L. (2013). Bayesian optimization explains human active search. In *Advances in neural information processing systems* (pp. 55–63).

Bramley, N. R., Lagnado, D. A., & Speekenbrink, M. (2015). Conservative forgetful scholars: How people learn causal structure through sequences of interventions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(3), 708.

Coenen, A., Nelson, J. D., & Gureckis, T. (2017). Asking the right questions about human inquiry.

Cook, C., Goodman, N. D., & Schulz, L. E. (2011). Where science starts: Spontaneous experiments in preschoolers? exploratory play. *Cognition*, *120*(3), 341–349.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: causal maps and bayes nets. *Psychological review*, *111*(1), 3.

Gottlieb, J., Oudeyer, P.-Y., Lopes, M., & Baranes, A. (2013). Information-seeking, curiosity, and attention: computational and neural mechanisms. *Trends in cognitive sciences*, *17*(11), 585–593.

Gureckis, T. M., & Markant, D. B. (2012). Self-directed learning: A cognitive and computational perspective. *Perspectives on Psychological Science*, *7*(5), 464–481.

Hills, T. T., Todd, P. M., Lazer, D., Redish, A. D., Couzin, I. D., Group, C. S. R., et al. (2015). Exploration versus exploitation in space, mind, and society. *Trends in cognitive sciences*, *19*(1), 46–54.

Itti, L., & Baldi, P. F. (2006). Bayesian surprise attracts human attention. In *Advances in neural information processing systems* (pp. 547–554).

Kidd, C., & Hayden, B. Y. (2015). The psychology and neuroscience of curiosity. *Neuron*, *88*(3), 449–460.

Kirsh, D., & Maglio, P. (1994). On distinguishing epistemic from pragmatic action. *Cognitive science*, *18*(4), 513–549.

Lehman, J., & Stanley, K. O. (2011). Abandoning objectives: Evolution through the search for novelty alone. *Evolutionary computation*, *19*(2), 189–223.

León-Villagrá, P., Preda, I., & Lucas, C. G. (2018). Data availability and function extrapolation. In *Proceedings of the Annual Meeting of the Cognitive Science Society.*

Lieder, F., Plunkett, D., Hamrick, J. B., Russell, S. J., Hay, N., & Griffiths, T. (2014). Algorithm selection by rational metareasoning as a model of human strategy selection. In *Advances in neural information processing systems* (pp. 2870–2878).

Lucas, C. G., Griffiths, T. L., Williams, J. J., & Kalish, M. L. (2015). A rational model of function learning. *Psychonomic bulletin & review*, *22*(5), 1193–1215.

Markant, D. B., Settles, B., & Gureckis, T. M. (2016). Self-directed learning favors local, rather than global, uncertainty. *Cognitive science*, *40*(1), 100–120.

Nelson, J. D. (2005). Finding useful questions: On bayesian diagnosticity, probability, impact, and information gain. *Psychological review*, *112*(4).

Ruggeri, A., & Lombrozo, T. (2014). Learning by asking: how children ask questions to achieve efficient search. In *Proceedings of the 36th annual conference of the cognitive science society* (pp. 1335–1340).

Schulz, E., Konstantinidis, E., & Speekenbrink, M. (2017). Putting bandits into context: How function learning supports decision making.

Schulz, L., & Bonawitz, E. B. (2007). Serious fun: preschoolers engage in more exploratory play when evidence is confounded. *Developmental psychology*, *43*(4), 1045.

Teodorescu, K., & Erev, I. (2014). On the decision to explore new alternatives: The coexistence of under-and over-exploration. *Journal of Behavioral Decision Making*, *27*(2), 109–123.

Wu, C. M., Schulz, E., Speekenbrink, M., Nelson, J. D., & Meder, B. (2018). Generalization guides human exploration in vast decision spaces. *Nature Human Behaviour*, *2*(12), 915.

Kinematic Specification of Intention in Full-body Motion

Sierra F. Corbin[1], Charles H. Moore[1], Gaurav Patil [3], Lillian Rigoli [4], Tehran Davis [1], Kevin Shockley [1], Tamara Lorenz [1,2,3]

[1]Center for Cognition, Action, & Perception, Department of Psychology, University of Cincinnati, Cincinnati, USA
[2]Department of Electrical Engineering and Computation Science, University of Cincinnati, Cincinnati, USA
[3]Department of Mechanical and Materials Engineering, University of Cincinnati, Cincinnati, USA
[4]Department of Psychology, Macquarie University, Sydney, Australia

Abstract

Kinematic specification of dynamics (KSD) states that full-body kinematic patterns of daily activities are reflective of a person's plans, goals, and intentions. Furthermore, it has been shown that observers of those activities are well attuned to differences between those kinematic patterns. However, despite a substantial body of research on the identification of intentional motion, it is not yet clear what the essential kinematic information is required to perceive the intention from the kinematic pattern. Therefore, we analyzed four different intentional full body motions (sit-to-stand transitions: stand, press-stand, press-sit, and reach-up), to determine the essential kinematic information that differentiates them. We utilized principal component analysis (PCA), linear mixed models, and hierarchical multinomial logistic regression to create two predictive regression models that allow us to successfully identify and distinguish the four intentional motions.

Keywords: Intention Recognition; Kinematic Specification of Dynamics; Sit-to-Stand Transition; Point-Light Displays;

Introduction

Activities that people perform in their daily lives are reflected in a person's full body kinematic patterns (Johansson, 1973). Moreover, human observers can easily perceive even small differences in the patterns of a person's motion profile (Ansuini, 2005; Becchio, Manera, Sartori, Cavallo, & Castiello, 2012; Richardson & Johnston, 2005). It has therefore been argued that the information humans derive from another person's biological motion profile can be used to establish successful coordination with others (Pezzulo, Donnarumma, & Dindo, 2013; Sartori, Becchio, Bara, & Castiello, 2009) and with intelligent machines, such as robots (Vernon, Thill, & Ziemke, 2016).

Kinematic Specification of Dynamics

In order to study biological motion, Johansson (1973) created the first point-light displays by attaching small lights to his participants and limiting the exposure of his camera recording to capture only those lights. Johansson called these recordings point-light displays and discovered that when he placed the lights on key joint centers, observers of the displays could identify that the moving points represented a person performing a specific action.

Runeson (1994) framed the findings behind point-light displays in his principle of kinematic specification of dynamics (KSD). The KSD principle postulates that because movement kinematics are lawfully related to the dynamics that produce a movement, the movement specifies the dynamics from which it arose. In other words, the relations among a person's joint centers and joint angles specify the action that they are performing.

Specification of Action Capabilities

Overall, kinematic information has been shown to be remarkably rich. For example, point-light displays of an actor pretending to lift a heavy box are noticeably different from displays of the actor actually lifting a heavy box (Runeson, 1994). Furthermore, observing the kinematics of a person can not only specify the action that is being performed, it can also carry rich information about the action capabilities of the observed person. For example, Ramenzoni, Riley, Davis, Shockley, & Armstrong (2008) have shown that after observers view another person walking, they become more accurate at estimating the walker's maximum reach-with-jump height. Additionally, after watching another person walk while wearing (unseen) ankle weights, observers are sensitive to reductions in the walker's maximum reach-with-jump height caused by the additional weight. These findings indicate that a person's general movement pattern provides sufficient information to make an educated judgment of a person's action capabilities.

Specification of Intention

Although it is evident that people can distinguish another person's activities based on the kinematic structure of the displayed motion, there has been some debate about the richness of KSD in terms of social interaction and

intentions. In general, it is hypothesized that the intentionality as reflected in human motion can be used to understand another person's action plans. Thus, one's own actions can subsequently be adjusted in response to this understanding and smooth action coordination can be executed (Becchio, Sartori, Bulgheroni, & Castiello, 2008). However, Jacob & Jeannerod (2005) argued that the kinematics involved at the start of a chain of movements might not reflect the end goal of that chain of movements, meaning the kinematics might not accurately reflect the intention. They proposed a thought experiment involving the story of Dr. Jekyll and Mr. Hyde; the two identities belong to the same person, but the former is a renowned surgeon who performs surgeries on anesthetized patients. The latter is a dangerous sadist who performs the same hand movements on his non-anesthetized victims. Jacob & Jeannerod argued that if someone were to witness one of the two identities reaching and grasping for a scalpel, then it would be impossible to specify the social nature and intention through the grasping motion.

Several studies were performed in response to Jacob and Jeannerod's thought experiment and found evidence against their claim. Ansuini, Giosa, Turella, Altoè, & Castiello (2008) showed that prior intention shapes kinematics by measuring prior-to-contact grasping kinematics for reach-to-grasp movements performed toward a bottle filled with water. By comparing hand shaping across tasks involving different subsequent actions such as pouring the water into a container, throwing the bottle, and moving the bottle from one spatial location to another, the authors demonstrated how prior intention in grasping an object strongly affected the positioning of the fingers, the duration of the reaching, and the contact phase of the movement. Becchio et al. (2008) performed a similar experiment investigating differences between grasping an object to move it to another location and grasping an object to hand it to another person. The velocities and shapes of participants' hands for both the opening and closing phases of the grasping movement were significantly different between the two conditions, as well as the trajectory of the movement during the passing phase. While Becchio and colleagues demonstrated that movement kinematics differ based on the social or operational intention, Manera, Becchio, Cavallo, Sartori, & Castiello (2011) showed that observers can also differentiate between distinct reaching intentions. They presented point-light displays of different grasping movements including a slow grasping movement, a fast grasping movement, a grasping movement with the cooperative intention of passing an object to another person, and a grasping movement with the competitive intention of grabbing an object before another person. With only access to the kinematic information of the initial forward movement, observers were able to accurately classify which of the four actions was being presented 72% of the time, indicating that the observed kinematic patterns may be used for action coordination during joint activities (see also Sartori et al., 2009).

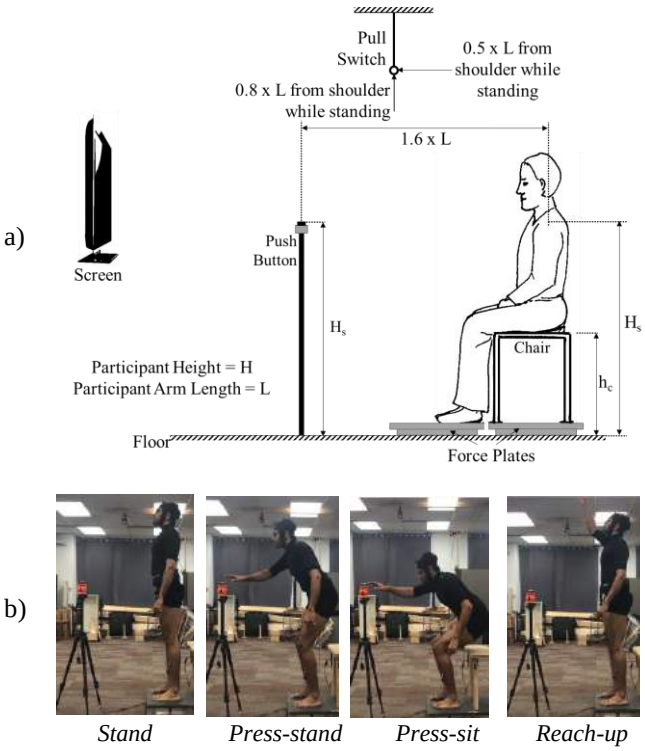


Figure 1: a) Experimental setup for data collection as established and published in Patil et al. (2018). b) Intentional sit-to-stand transitions.

Kinematic Specification of Action and Intention

Although it has been shown that one's movement kinematics provide the information necessary for another person to identify one's action capabilities and intentions, the informational basis for this ability has not been identified (though see Ansuini, Cavallo, Bertone, & Becchio, 2015). Therefore, in the current study, we adopted a similar approach to Weast-Knapp et al. (2019) who used Principal Components Analysis (PCA) to decompose the kinematic data of walking movements to isolate the informational basis for an observer trying to perceive a person's action capabilities. However, rather than focusing on the informational basis to estimate a person's action capabilities, here we explore a different type of full-body movement (sit-to-stand transition, STS) that was executed with different intentions that altered the basic STS motion (see Figure 1b) in order to identify the essential kinematic information of a full-body motion with varying intentions. To gain insight on how people can perceive intention from motion, we must first confirm if the essential kinematic information is different between the motions. If the differences exist, the next goal will be to confirm that humans can extract the same information. Therefore, this paper tackles the first goal of clarifying the essential kinematic structures for STS intentional motion.

Method

In order to enable the analysis of intentional motion, we utilized one subset of a larger data set that was originally collected in context of understanding joint angle variations for exoskeleton control (Patil et al., 2018). The data subset was taken from a healthy, 28-year old male participant.

Setup and Procedure

In order to induce four different intentional STS transitions, a setup was created by Patil et al. (2018) as shown in Figure 1a. A button was placed in front of participants at the shoulder height while sitting and 1.6-times the arm length from the shoulder. A pull switch was positioned above the participant at a height of 0.8 times the arm length and at a distance of 0.5 times the arm length from the shoulder while standing. Motion data was recorded using a 20-camera 3D motion capture system (Motion Analysis Corporation, Santa Rosa, CA). A 29-marker set based on the Helen Hayes body marker placement protocol (Kadaba, Ramakrishnan, & Wootten, 1990) was used to track the motion. A screen was positioned at eye level in front of the participant to provide instructions for the specific trial. Every trial started with the participant sitting at a stool (height 45.72 cm) without any hand or back rests.

The participants were shown a "ready" signal on the screen and, after 3 seconds, the instruction to perform any of four randomized tasks marked the go-signal. The participants performed 100 trials of four intentional STS transition tasks (25 trials per intention): *stand* - the participants were asked to stand up at a comfortable speed without any intention of subsequent activity, *press-stand* - the participants were asked to stand up from the chair while pushing the button in front of them and finish standing up; *press-sit* - the participants were asked to stand up from the chair while pushing the button in front of them and to then immediately sit back down; *reach-up* - the participants were asked to stand up to pull on the switch above their head and after pulling the switch finish standing up. For all trials, the participants were instructed not to use their hands to push down on the chair or their thighs during STS and not to lift their feet from the heel or toes during the trial. The participants were allowed to take breaks whenever they felt fatigued. For the purpose of the current analysis, we included the first four trials of each performed intentional sit-to-stand transitions within the data subset.

Data Analysis and Results

Determining Essential Principal Components

Seven of the original 29 markers (corresponding to the head, right shoulder, right elbow, right hip, right knee, right ankle, and right hand) were selected to form a simple sideview configuration of each motion with the right side of the body represented. We cut off each time series using the furthest point forward in the motion of the hip marker as shown in Figure 1b. This served to truncate the movement to the initial intention-expressing forward portion of rising from a seated position and excluded the stand-to-sit backwards transition. In the future, we plan to use this motion data to explore how human observer respond when viewing it. The Y and Z coordinates of the recorded 3-dimentional motion data were used to perform a Euclidean transformation which provided one value for each marker for each frame of the data set (cf. Weast et al., 2019). The data was then submitted to PCA via R (*base package: prcomp*). PCA is a statistical tool that allows for the reduction of high-dimensional data with the goal of revealing hidden structure in the underlying relationship between variables. For example, previous research has utilized PCA to uncover the most important factors contributing to variation in movement kinematics in gait (Vallery & Buss, 2006), juggling (Post, Peper, & Beek, 2003) and even the movements of cooperating actors (Ramenzoni, Riley, Shockley, & Baker, 2012). Here, we used PCA to reduce the seven-marker data set to a subset of principal components (PC) that captured the dynamics of the kinematic movements. We decided to use PCA rather than machine learning techniques, as we are interested in which joint centers hold the essential kinematic information to differentiate the motions. Though machine learning can help determine the presence of differences and classify each motion, it will not offer insight as to which body segments participate in providing the structure that differentiates movements. Subsequent analyses were performed on this subset to identify the activity of key markers for discriminating between intended movements.

We completed 16 PCA analyses, one per STS motion file (4 intentions × 4 repetitions). Each PCA analysis yielded a **7** (markers) × **7** (PC dimensions) matrix of coefficients, as well as a vector of the amount of total variance accounted for by each PC. The variance vectors were used as criteria to reduce the original data to those PCs that (1) accounted for at least 10% of the total variance in the motion pattern, and (2) provided sufficient variation between intentional movement profiles to be a useful candidate for future discrimination. PC1 and PC2 reliably met criteria (1), suggesting that the data could be reduced to the first two PCs without much loss in information. To determine (2) we submitted the percent explained variance of each PC to a linear mixed effects model (R package: lme4) with intention as a fixed effect and instance (each intentional motion was performed four times) as a random effect. For the sake of brevity, we only report the *F*-tests (Satterthwaite's degrees of freedom method) for overall significance of the models. Only models for PC1, $F(3,12) = 47.49$, $p < .001$, and PC2, $F(3, 9) = 54.51$, $p < .001$, were significant, suggesting that the amount of explained variance for both PC1 and PC2 differed by intentional movement. For the remaining PCs this relationship was non-significant, supporting our choice to further analyze PC1 and PC2.

Elimination of Non-significant STS Markers in PC1 & PC2

Having reduced the data to the first two PCs, two additional sets of linear mixed model analyses (7 per PC1 and PC2; 14 total) were completed to establish whether the PCA coefficients for each marker systematically varied as a function of the intentional motion.

Table 1. Results of linear mixed model for STS markers on PCs 1 and 2

Marker	Model PC1	Model PC2
M1 (Head)	$F = 86.10$, $p < .001$ *	$F = 4.82$, $p = .03$ *
M2 (Shoulder)	$F = 100.44$, $p < .001$ *	$F = 9.37$, $p = .004$ *
M3 (Elbow)	$F = 8.52$, $p = .003$ *	$F = .57$, $p = .65$
M4 (Hip)	$F = 132.40$, $p < .001$ *	$F = 19.64$, $p < .001$ *
M5 (Ankle)	$F = 3.10$, $p = .08$	$F = .09$, $p = .96$
M6 (Knee)	$F = 1.07$, $p = .41$	$F = .08$, $p = .97$
M7 (Hand)	$F = 3{,}132$, $p < .001$ *	$F = 2.86$, $p = .10$

** candidate markers*

Again, intentional motion was entered into the model as a fixed effect with instance as a random effect. The purpose of this series of analysis was to further reduce the dimensionality of the data by identifying candidate markers, whose activity might be used to build a parsimonious model for predicting the intended motion. In short, we sought to determine *which markers* might qualify for submission to a predictive model for intention, as well as *how few* may be used to build a model that reliably discriminates between the intentional movements.

As can be seen in Table 1, the analysis on PC1 revealed that the coefficients corresponding to markers M1, M2, M3, M4, and M7 varied systematically as a function of intentional motion; repeating this analysis for the coefficients in PC2, we found significant systematic variability for markers M1, M2, and M4. These sets of markers provided a list of variables to enter into subsequent regression analysis for PC1 and PC2.Table 2. Results of linear mixed model for STS markers on PCs 1 and 2

Regressing Intention Categories onto Candidate Markers

Using our candidate markers, we performed two hierarchical multinomial logistic regressions (one for each PC) to determine which combination of markers was most parsimonious in reliably discriminating between the intentional movements. For the analysis along PC1, a single marker per hierarchical step was loaded into the regression model in order from largest to smallest PCA coefficient mean. This resulted in a statistically significant model containing markers M7 (hand) and M3 (elbow), which improved the likelihood of determining the corresponding intentions, above and beyond the null (chance) model, as well as all models formed by prior steps in the analysis (see Table 2).

Table 2. Summary of model results for hierarchical multinomial regression for STS markers in PC 1. Only models that yielded a significant improvement are reported.

Step	Variables Entered	df	Likelihood Ratio	*p*
1	**M7**	42	35.81	< .001
2 *final*	**M7 + M3**	39	7.85	.049
	M7 = hand, M3 = elbow			

We followed an identical method for PC2, hierarchically entering each marker into the regression model beginning with the marker possessing the largest PCA coefficient mean. The resulting model was statistically significant, containing M1, M2, and M4, and improved the likelihood of determining the corresponding intentions, above and beyond the null (chance) model, as well as all models formed by prior steps in the analysis (see Table 3).

Table 3. Summary of model results for hierarchical multinomial regression for STS markers in PC 2.

Step	Variables Entered	df	Likelihood Ratio	*p*
1	**M1**	42	6.37	< .001
2	**M1 + M2**	39	13.10	.004
3 *final*	**M1 + M2 + M4**	36	14.88	.002
			M1 = head, M2 = shoulder, M4 = hip	

Examining Improved Accuracy from Model 1 to Final Model

Finally, we compared the accuracy in intention categorization for each of the regression results by calculating the predicted probabilities derived from the fitted values of the Step 1 and final models. For brevity, we report the predicted probability of the true (correct) intentional movement given the PC coefficients for each marker. As expected, we observed significant improvement in predictive probabilities from the first to the final models. For both PC1 and PC2, the predicted probabilities of the final model that corresponded to the correct intention was greater than 95%. Moreover, this was achieved using relatively few markers (PC1: 2 out of 7, PC2: 3 out of 7). Our results suggest that, for PC1, hand and elbow marker activity appear to provide the essential kinematic information to differentiate movement categories (see Table 4).

Table 4. Correct predicted probabilities of multinomial regression models using PC1.

	PC 1: Predicted Probabilities	
Intention	Model 1	Final Model
Stand	97.99%	100%
Press-stand	70.20%	97.03%
Press-sit	86.50%	98.70%
Reach-up	58.22%	95.78%

For PC2, the reduction to head, shoulder, and hip suggests that these markers may contain the essential kinematic information to further differentiates movement categories (see Table 5).

Table 5. Correct predicted probabilities of multinomial regression models using PC2

	PC 2: Predicted Probabilities	
Intention	Model 1	Final Model
Stand	44.50%	100%
Press-stand	62.07%	100%
Press-sit	52.35%	99.99%
Reach-up	22.58%	100%

Discussion

We aimed to identify the essential kinematic information available to observers for distinguishing intentional STS transitions.

Overall, the results suggest that, while the four intentional STS transitions (stand, press-stand, press-sit, reach-up) are built upon similar motion profiles, there are distinct differences regarding the essential kinematic information along the first two PCs, which allows for the accurate differentiation of each intention by means of a specific subsets of markers. Analyses of the coefficients in PC1 and PC2 sufficiently capture the majority of the variance attributed to differentiating the four intentional STS transitions. Additionally, the stratification of specific markers within the two PCs allows us to specify (and differentiate) the essential kinematic structure of each intentional motion. This provided the opportunity to formulate regression models that were capable of accurately predicting intention, above and beyond chance level. Both predictive models allowed for the classification of each intentional STS transition with 95-100% accuracy.

Thus, within each PC, there exists essential kinematic information that can be extracted from the time series of similar, yet distinct, intentional motions. Each marker in the final model, can then be understood as *one of the essential communicators of intention* for each STS transition. In turn, the variation in coefficients indicates *how* each marker contributes to the overall movement pattern of each intention.

For example, analyzing PC1 showed that the majority of variance in the motion data can be explained by the markers reflecting the arm motion (i.e.: elbow and hand marker). Considering that the arm motion differed significantly across intentions (e.g. reaching up vs. reaching forwards), this result is consistent with expectations. Subsequently, analyzing PC2 indicated the presence of additional essential kinematic information in the head, shoulder, and hip markers, which distinguishes suprapostural differences in the kinematics of the full-body motion.

Ultimately, our results reinforce empirical findings showing that humans are capable of visually distinguishing different intentional motion patterns (c.f. Ansuini, 2005; Becchio, Manera, Sartori, Cavallo, & Castiello, 2012; Pezzulo, Donnarumma, & Dindo, 2013; Sartori, Becchio, Bara, & Castiello, 2009) by revealing the essential information that defines and differentiates the kinematic structure of intentional motion.

References

Ansuini, C. (2005). Effects of End-Goal on Hand Shaping. *Journal of Neurophysiology*, *95*(4), 2456–2465. http://doi.org/10.1016/j.tet.2015.04.102

Ansuini, C., Cavallo, A., Bertone, C., & Becchio, C. (2015). Intentions in the brain: The unveiling of Mister Hyde. *Neuroscientist*, *21*(2), 126–135. http://doi.org/10.1177/1073858414533827

Ansuini, C., Giosa, L., Turella, L., Altoè, G., & Castiello, U. (2008). An object for an action, the same object for other actions: Effects on hand shaping. *Experimental Brain Research*, *185*(1), 111–119. http://doi.org/10.1007/s00221-007-1136-4

Becchio, C., Manera, V., Sartori, L., Cavallo, A., & Castiello, U. (2012). Grasping intentions: from thought experiments to empirical evidence. *Frontiers in Human Neuroscience*, *6*(May), 1–6. http://doi.org/10.3389/fnhum.2012.00117

Becchio, C., Sartori, L., Bulgheroni, M., & Castiello, U. (2008). The case of Dr. Jekyll and Mr. Hyde: a kinematic study on social intention. *Consciousness and Cognition*, *17*(3), 557–64. http://doi.org/10.1016/j.concog.2007.03.003

Jacob, P., & Jeannerod, M. (2005). The motor theory of social cognition: A critique. *Trends in Cognitive Sciences*, *9*(1), 21–25. http://doi.org/10.1016/j.tics.2004.11.003

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics*, *14*(2), 201–211. http://doi.org/10.3758/BF03212378

Kadaba, M. P., Ramakrishnan, H. K., & Wootten, M. E. (1990). Measurement of lower extremity kinematics during level walking. *Journal of Orthopaedic Research*,

8(3), 383–392. http://doi.org/10.1002/jor.1100080310

Manera, V., Becchio, C., Cavallo, A., Sartori, L., & Castiello, U. (2011). Cooperation or competition? Discriminating between social intentions by observing prehensile movements. *Experimental Brain Research. Experimentelle Hirnforschung. Expérimentation Cérébrale*, *211*(3–4), 547–56. http://doi.org/10.1007/s00221-011-2649-4

Patil, G., Rigoli, L., Richardson, M. J., Kumar, M., Kiefer, A. W., & Lorenz, T. (2018). Joint Angle Variation in Intentional Sit-to-Stand Transitions. In *Proceedings of the 2nd IFAC Conference in Cyber-Physical and Human Systems*. Miami, FL.

Pezzulo, G., Donnarumma, F., & Dindo, H. (2013). Human sensorimotor communication: A theory of signaling in online social interactions. *PLoS ONE*. http://doi.org/10.1371/journal.pone.0079876

Post, A. A., Peper, C. E., & Beek, P. J. (2003). Effects of Visual Information and Task Constraints on Intersegmental Coordination in Playground Swinging. *Journal of Motor Behavior*, *35*(1), 64–78. http://doi.org/10.1080/00222890309602122

Ramenzoni, V., Riley, M. A., Davis, T., Shockley, K., & Armstrong, R. (2008). Tuning in to another person's action capabilities: Perceiving maximal jumping-reach height from walking kinematics. *Journal of Experimental Psychology: Human Perception and Performance*, *34*, 919–928. http://doi.org/10.1037/0096-1523.34.4.919

Ramenzoni, V. C., Riley, M. A., Shockley, K., & Baker, A. A. (2012). Interpersonal and intrapersonal coordinative modes for joint and single task performance. *Human Movement Science*, *31*, 1253–1267. http://doi.org/10.1016/j.humov.2011.12.004

Richardson, M. J., & Johnston, L. (2005). Person Recognition from Dynamic Events: The Kinematic Specification of Individual Identity in Walking Style. *Journal of Nonverbal Behavior*, *29*(1), 25–44. http://doi.org/10.1007/s10919-004-0888-9

Runeson, S. (1994). Perception of biological motion: The KSD-principle and the implications of a distal versus proximal approach. In G. Jansson, S. S. Bergström, & W. Epstein (Eds.), *Resources for ecological psychology. Perceiving events and objects* (pp. 383–405). Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Sartori, L., Becchio, C., Bara, B. G., & Castiello, U. (2009). Does the intention to communicate affect action kinematics? *Consciousness and Cognition*, *18*(3), 766–72. http://doi.org/10.1016/j.concog.2009.06.004

Vallery, H., & Buss, M. (2006). Complementary limb motion estimation based on interjoint coordination using principal components analysis. *Proceedings of the IEEE International Conference on Control Applications*, 933–938. http://doi.org/10.1109/CACSD-CCA-ISIC.2006.4776770

Vernon, D., Thill, S., & Ziemke, T. (2016). The Role of Intention in Cognitive Robotics. In *Intelligent Systems Reference Library* (Vol. 105, pp. 15–27). http://doi.org/10.1007/978-3-319-31056-5_3

Weast-Knapp, J. A., Shockley, K., Riley, M. A., Cummins-Sebree, S., Richardson, M. J., Wirth, T. D., & Haibach, P. C. (2019). Perception of another person's maximum reach-with-jump height from walking kinematics. *Quarterly Journal of Experimental Psychology*, 174702181882193. http://doi.org/10.1177/1747021818821935

Working Memory and Co-Speech Iconic Gestures

Seana Coulson (scoulson@ucsd.edu) and Ying Choon Wu (yingchoon@gmail.com)

Department of Cognitive Science

9500 Gilman Drive, La Jolla 92093 USA

Abstract

The importance of verbal and visuospatial working memory (WM) for co-speech gesture comprehension was tested in two experiments using the dual task paradigm. Healthy, college-aged participants encoded either a dot locations in a grid (Experiment 1), or a series of digits (Experiment 2), and rehearsed them as they performed a discourse comprehension task. The discourse comprehension task involved watching a video of a man describing household objects, and judging which of two words probes was most related to the video. Following the discourse comprehension task, participants recalled either the verbally or visuo-spatially encoded information. In both experiments, performance on the discourse comprehension task was faster when gestural information was congruent with the speech than when it was incongruent. Moreover, performance on the discourse comprehension task was impacted both by increasing the load on the visuospatial WM system (Experiment 1) and the verbal WM system (Experiment 2). However, in both studies effects of WM load and gesture congruency were additive, suggesting they were independent.

Keywords: depictive gesture; discourse comprehension; iconic gesture; multimodal meaning; representational gesture; verbal working memory; visuospatial working memory

Introduction

Co-speech gestures, which are produced spontaneously in co-ordination with speaking, offer an exciting opportunity to explore the relationship between body movement and higher order cognitive functions, such as language comprehension and conceptualization (for review, see Goldin-Meadow, 2003). To date, little research has addressed the cognitive resources that allow us to understand these gestures and to relate their meaning to that conveyed by the accompanying speech. Because gestures relate to linguistic information at varying levels of granularity, including the word-, phrase, and sentence- levels (Kendon, 2004), one fairly straightforward possibility is that working memory (WM) plays an important role in these processes, allowing listeners to maintain information conveyed in the gestural stream until it can be integrated with relevant information presented in the speech.

Previous research has contrasted the *verbal resources hypothesis*, that speech gesture integration primarily recruits verbal WM, with the *visuo-spatial resources hypothesis*, that speech gesture integration recruits the visuo-spatial WM system. That work employed a discourse comprehension task in which participants viewed a multi-modal discourse prime of a speaker describing everyday objects, followed by a picture that participants judged as either related or unrelated to the prime (Wu &Coulson, 2014). Reaction times for related picture probes are typically faster following discourse primes with congruent gestures that match the concurrent speech, relative to incongruent gestures that do not, suggesting congruent iconic gestures help convey information about the discourse referents (Wu & Coulson, 2014).

Consistent with the visuo-spatial resources hypothesis, the magnitude of these congruity effects has been shown to be larger in participants with greater visuo-spatial WM capacity (Wu & Coulson, 2014). Moreover, imposing a concurrent verbal load during this task yielded additive effects of gesture congruity and WM load, while a concurrent visuo-spatial load yielded interactive effects, as gesture congruity effects were greatly attenuated under conditions of high visuo-spatial load (Wu & Coulson, 2014). Prior research thus suggests that speech-gesture integration recruits cognitive resources shared by visuo-spatial WM load tasks.

One shortcoming of research by Wu and Coulson (2014) is that their measure of speech-gesture integration involved participants' responses to picture probes that followed videos of multimodal discourse. Given that responding to pictorial stimuli presumably imposes a load on participants' visuospatial processing resources, this task may overestimate the importance of visuospatial WM for the comprehension of co-speech gestures.

The present study explored the role of verbal versus visuospatial WM in speech-gesture integration by utilizing a dual task paradigm similar to that in Wu & Coulson (2014). However, rather than using performance on a picture probe task to index comprehension of the gestures, we asked participants to choose which of two words was most related to the preceding discourse video. Experiment 1 paired this

discourse comprehension task with a visuospatial WM task, and Experiment 2 paired it with a verbal WM task.

Experiment 1

Experiment 1 tests how increasing the load on participants' visuospatial WM system impacts their sensitivity to the meaning of co-speech gestures in multimodal discourse. The logic of the dual task paradigm is that if the two tasks recruit shared cognitive resources, engagement in the secondary task will impair performance on the primary one. In Experiment 1, the primary task is that of discourse comprehension, as indexed by a word probe task, while the secondary task involved memory for a sequence of dot locations in a grid. We manipulated the difficulty of multi-modal discourse comprehension by varying the semantic congruity of the gestures and the speech in our discourse videos. The difficulty of the visuospatial recall task was varied by asking participants to remember a sequence of either four locations (high load), or to remember a single location (low load). Consequently, if the recall task diverts cognitive resources from speech-gesture integration, it would be reflected in a change in the congruency effects as a function of visuospatial load – that is, either the amplification of congruency effects, the reduction of congruency effects, or their elimination altogether.

Methods

Participants Participants were 51 healthy undergraduates who, in exchange for participation, received extra credit for a course in cognitive science, linguistics, or psychology.

Materials A total of 84 discourse primes were kindly provided by Dr. Wu. These primes were derived from continuous video footage of spontaneous discourse centered on everyday activities, events, and objects. The speaker in the video was naïve to the experimenters' purpose and received no explicit instructions to gesture. Short segments (2-8s) were extracted in which the speaker produced both speech and gesture during his utterance. Topics varied widely, ranging from the height of a child, the angle of a spotlight, the shape of furniture, swinging a golf club, and so forth. For congruent primes, the original association between the speech and gesture was preserved. To create incongruent counterparts, audio and video portions of congruent clips were swapped such that across all items, all of the same speech and gesture files were presented; however, they no longer matched in meaning.

In an independent norming study using a five point Likert scale, the degree of semantic match between speech and gesture in the congruent trials was rated on average as 1.6 points higher than in the incongruent trials (congruent = 3.8, sd=0.8 versus incongruent = 2.2, sd = 0.7). Because of the discontinuity between oro-facial movements and verbal output in incongruent items, the speaker's face was blurred in all discourse primes (i.e. both the congruent and incongruent version of each).

Each discourse prime was followed by the presentation of two word probes arrayed vertically in the center of the monitor. The *related* probe was a word related to the audio content of the video, and was intended to specifically highlight the semantic content of the congruent gesture. The *unrelated* probe was intended to be unrelated to any aspect of the audio or video. The same two word probes followed the congruent and the incongruent version of each audio file. The location of the related probe (i.e. at the top or the bottom of the array) was chosen randomly on each trial.

Half of the trials (n=42) were accompanied by a low load version of the visuo-spatial recall task, and half with a high load version of the same task. The visuospatial recall task was similar to the dot movement task employed by Wu & Coulson (2014), in which participants were asked to remember a single location in a 4 x 4 grid on low load trials, and an ordered sequence of four locations on high load trials. The gesture congruity and memory load manipulations were fully counterbalanced.

Procedure Each trial began with a fixation cross (1s), followed by the encoding phase of the secondary task (visuospatial WM). Secondary encoding involved the visual presentation of a sequence of dots in a 4x4 grid. High load trials involved a sequence of four distinct locations, while low load trials involved the presentation of a single dot. Each dot remained visible on the grid for one second. A 500ms pause concluded the encoding phase.

The discourse comprehension portion of each trial began with a discourse video, presented at a rate of 30ms per frame in the center of a computer monitor. Immediately following the video offset, the probes appeared above and below the fixation cross. The mouse cursor was initialized to a location equidistant between the two. Participants were asked to respond by clicking the mouse in the square that contained the word that best matched the scenario described by the speaker. No feedback was given.

After a 250ms pause, participants were prompted to recall the location of dots in the grid in the order that they had been presented. Written feedback ("correct" versus "incorrect") was provided following each trial for 500ms. Between trials, the screen was blank for half a second and the mouse cursor was reset to a neutral hidden position.

After completion of the dual-task portion of the experiment, verbal and visuo-spatial WM capacity were assessed through two short tests – an auditory version of the Sentence Span task (Daneman and Carpenter, 1980) and a computerized version of the Corsi Block task (Milner, 1971). The Listening Span task involved listening to sequences of unrelated sentences and remembering the sentence final word in each. All trials contained between two and five items, and were presented in blocks of three. An individual's span was the highest consecutive level at which all sentence final words were accurately recalled (in any order) on at least two of the three trials in a block.

In the Corsi Block task, an asymmetric array of nine squares was presented on a computer monitor. On each trial, between three and nine of the squares flashed in sequence, with no square flashing more than once. Participants reproduced patterns of flashes immediately

afterwards by clicking their mouse in the correct sequence of squares. An individual's Corsi span was the highest level at which at least one sequence out of five was correctly replicated (Conway et al., 2005). The entire experimental session lasted approximately two hours.

Results

Visuospatial Recall Task Performance on the visuo-spatial recall task was indexed by the number of trials in which the participant correctly recalled all of the to be remembered locations. These values were subjected to repeated measures ANOVA with factors memory load (High, Low) and gesture congruity (Congruent, Incongruent). This analysis revealed only an effect of Load, $F(1, 67) = 130.2$, $p < 0.05$, $ges = .24$. Figure 1 shows the average number of correct trials in each condition and clearly indicates better performance in trials with a low load (1 dot location) than in the high load trials (4 dot locations). These data suggest the memory load task was more difficult in the high than the low load condition.

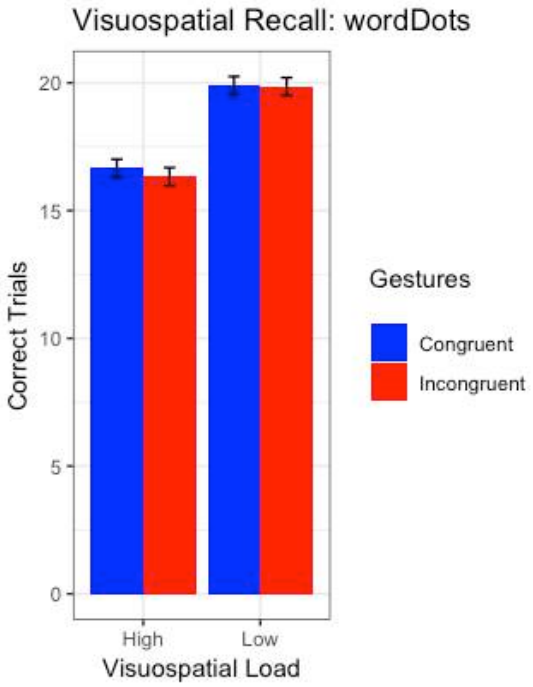


Figure 1: Mean number of correct trials on the recall task in each condition of Experiment 1.

To test the importance of WM capacity for sensitivity to our memory load manipulation, we computed the difference between each participant's accuracy on the high and low load trials. We then constructed a linear model to predict this difference due to the load manipulation as a function of participants' scores on the Corsi Block and Listening Span tasks. This model significantly predicted accuracy on the recall task, F (2, 64) = 5.18, p < 0.01, accounting for 13.95% of the variance. ANOVA on the output of the model suggested scores on the Corsi Block Task served as significant predictors, F (1, 64) = 10.3, p < 0.01, while scores on the Listening Span did not, F (1, 64) = 0.03, n.s. The systematic relationship between scores on the Corsi Block Task with the visuo-spatial load effect supports our contention that the dots task recruits visuo-spatial WM.

Discourse Comprehension Task

Accuracy on the discourse comprehension task was scored by counting the number of correct trials in each condition for each participant. Figure 2 shows the mean scores in each condition. These values were subjected to repeated measures ANOVA with factors gesture congruity (congruent/incongruent) and memory load (high/low). This analysis revealed a main effect of gesture congruity, F(1, 66) = 3.4, p < 0.05, ges 0.08, as participants were more accurate when speech was accompanied by congruent than incongruent gestures. Memory load was not significant, either as a main effect, F(1, 66) = 2.5, n.s., or as an interaction with gesture congruity, F(1, 66) = 1.08, n.s.

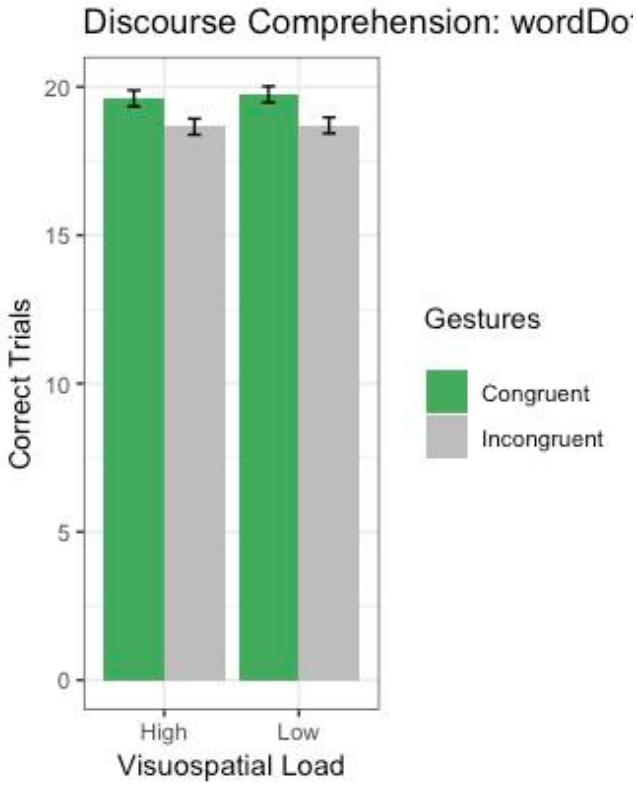


Figure 2: Mean number of correct trials in the discourse comprehension task in Experiment 1. Error bars represent 95% confidence intervals.

To explore the relationship between sensitivity to gestures and our individual difference measures, we computed the difference between the total number of trials each participant responded to correctly in the congruent gesture condition and the incongruent gesture condition. A linear model was constructed to predict this difference score from the Corsi Span score and the Listening Span score. This model accounted for 10.4% of the variance in difference scores, F(2, 64) = 3.72, p < 0.05. ANOVA on the output of

the model suggested only Corsi Span scores served as a significant predictor, $F(1, 64) = 5.3$, $p < 0.05$.

Response times for correct trials on the discourse comprehension task were analyzed with linear mixed effects models with fixed effects for gesture congruity and visuospatial load, and random effects for subject and item (viz., the audio file held constant across congruent and incongruent gesture versions of each stimulus). Random effect structure was determined via backwards model comparison using the step function in lmerTest, beginning with the 'maximal' structure allowed by the design.

Mean response times in each condition are shown in Figure 3. Performance on this task was an additive function of gesture congruity, $t = -6.84$, $p < 0.001$, with responses that were on average 383ms faster following congruent than incongruent gestures, and memory load, $t = -3.95$, $p < 0.001$, with responses an average of 170ms faster in high load trials than low load trials. The latter presumably results because participants desire to rush through the discourse comprehension task in order to 'unload' memory items in the recall task that immediately followed.

Discussion

Experiment 1 suggests a relationship between visuospatial WM capacity and sensitivity to speech-gesture congruity, but fails to support a causal link between visuospatial WM and the comprehension of gestures.

First, did the visuospatial recall task (viz. the dot task) serve to divert visuospatial resources from the primary task? Indeed, recall performance was worse under conditions of high than low load. Moreover, participants' performance on the dot task was systematically related to their visuospatial WM capacity as indexed by their scores on the Corsi block task. These data suggest that the dot task did indeed recruit our participants' visuospatial processing resources, thereby making them less available for primary task performance.

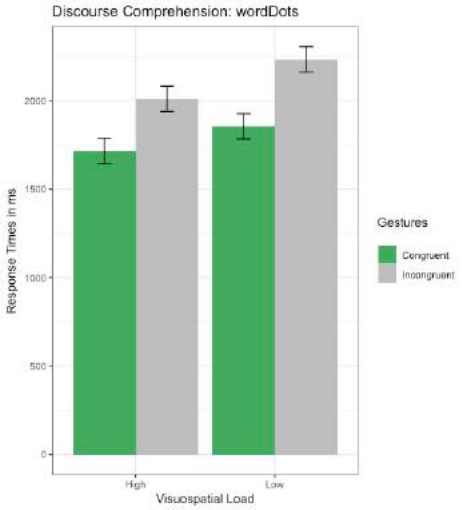


Figure 3: Mean response times in the discourse comprehension task in Experiment 1. Error bars depict 95% confidence intervals.

Second, was the discourse comprehension task employed here sensitive to the relationship between the gestures and the speech? Participants responded more quickly and more accurately on trials preceded by discourse with congruent gestures than incongruent ones. Experiment 1 thus extends results reported in Wu & Coulson (2014), showing that the facilitative impact of congruent gestures can be detected with the word probe paradigm employed in the present study. Moreover, as in the report by Wu and Coulson (2014), the participants who scored the highest on our independent assessment of visuospatial WM capacity were those who showed the largest gesture congruity effects.

Finally, how was performance of the discourse comprehension task impacted by the diversion of visuospatial processing resources? Apart from the gesture congruity effect noted above, the discourse comprehension task was also impacted by visuospatial load. Load had a somewhat paradoxical impact on responses as participants responded faster but less accurately on high load trials. Importantly, though, these two effects were additive, suggesting the discourse comprehension task proceeded somewhat independently of the visuo-spatial recall task.

Experiment 2

Experiment 2 paired the discourse comprehension task with a verbal WM task to explore how reducing the availability of verbal resources impacted participants' sensitivity to iconic co-speech gestures.

Methods

Audio and video materials were identical to those used in Experiment 1, as were the word probes. As in Experiment 1, half of the trials were accompanied by a low load recall task, and half with a high load recall task. The secondary recall task was similar to the digit recall task employed by Wu & Coulson, in which participants were asked to remember a single digit on low load trials, and an ordered series of four digits on high load trials. As in Experiment 1, the gesture congruity and memory load manipulations were fully counterbalanced.

During the encoding phase of the verbal task, a series of four numbers (each ranging between one and nine) were selected pseudo-randomly, and presented via digitized audio files while a central fixation cross remained on the computer screen. As for the visuospatial WM task in Experiment 1, the stimulus onset asynchrony for to-be-remembered items was 1 second.

During the recall phase of the task, an array of randomly ordered digits from 1-9 appeared in a row in the center of the screen, and participants clicked the mouse on the numbers that they remembered hearing in the order that they were presented. Written feedback (either "Correct" or "Incorrect") on the recall task was shown on the monitor for half a second after the final mouse click.

Results and Discussion

Verbal Recall

Performance on the verbal recall task was indexed by the number of trials in which the participant correctly recalled all of the to be remembered digits (see Figure 4). These values were subjected to repeated measures ANOVA with factors memory load (High/Low) and gesture congruity (Congruent/Incongruent). This analysis revealed only an effect of memory load, $F(1, 47) = 35.9$, $p < 0.05$, $ges = .13$. Figure 4 shows the average number of correct trials in each condition and clearly indicates better performance in the low load (1 digit) trials than in the high load trials (4 digits). These data suggest the task worked as intended to occupy verbal WM.

To test the importance of WM capacity for sensitivity to our verbal memory load manipulation, we computed the difference between each participant's accuracy on the high and low load trials. We then constructed a linear model to predict this memory load effect as a function of scores on the Corsi Block and Listening Span tasks. This initial model only approached significance, F (2, 45) = 2.92, p = 0.06. Backwards model selection via the step function in the MASS package in R indicated that the best model of memory load effects was one that included a single factor, participants' Listening Span scores.

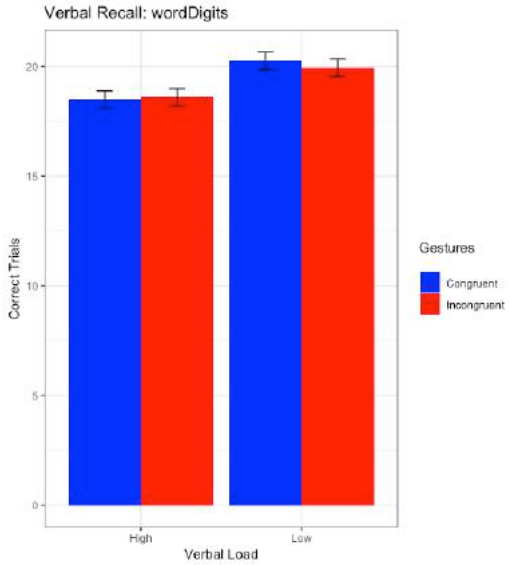


Figure 4: Mean number of correct trials in the verbal recall task for each condition in Experiment 2. Error bars represent 95% confidence intervals.

Accordingly, Corsi Span scores were dropped. The reduced model was significant, F(1, 46) = 5.81, p < 0.05, predicting 11.2% of the variance. The coefficient on the Listening Span factor was -1.2, indicating the load effect was most pronounced in participants with the lowest Listening Span scores. These data indicate a relationship between sensitivity to the digit load manipulation with our independent assessments of participants' verbal WM capacity, consistent with our assumption that the digit recall task diverted verbal WM resources. The systematic relationship between scores on the Listening Span Task with the verbal load effect supports our contention that the digit recall task recruits verbal WM resources.

Discourse Comprehension

Accuracy on the discourse comprehension task was scored by counting the number of correct trials in each condition for each participant. Figure 5 shows the mean scores in each condition. These values were subjected to repeated measures ANOVA with factors gesture congruity (congruent/incongruent) and memory load (high/low). This analysis revealed a main effect of gesture congruity, F (1, 47) = 3.4, $p < 0.05$, ges 0.12, as participants were more accurate when speech was accompanied by congruent than incongruent gestures. Memory load was not significant, either as a main effect, F (1, 47) = 0.03, $ges < 0.01$ or as an interaction with gesture congruity, F (1, 47) = 1.34, n.s, $ges < 0.01$.

To explore the relationship between sensitivity to gestures and our individual difference measures, we computed the difference between the total number of trials each participant responded to correctly in the congruent gesture condition and the incongruent gesture condition. A linear model was constructed to predict this difference measure from the Corsi Span score and the Listening Span score. However, neither this model nor any of the models explored with backwards model selection provided a significant account of these effects, indicating the absence of a systematic relationship between working memory capacity and this measure of sensitivity to gesture congruity.

Response times were analyzed in the same manner as in Experiment 1. Analysis involved the construction of linear mixed effects models with fixed effects of memory load and gesture congruity, and random effects of subject and item. As in Experiment 1, backwards model selection was used to simplify the random effects structure and choose the best model.

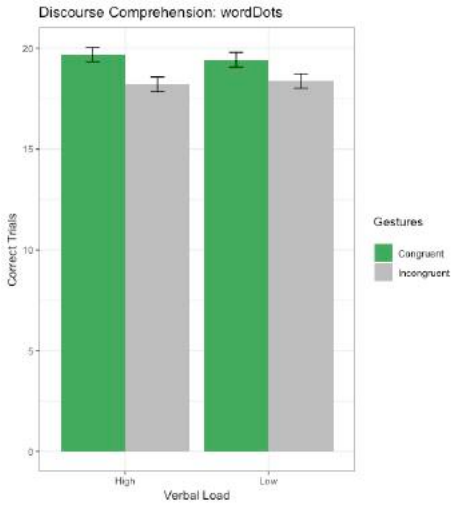


Figure 5. Number of correct trials on the discourse comprehension task in each condition of Experiment 2. Error bars depict 95% confidence intervals.

As in Experiment 1, the memory load effect results due to a 122ms faster responses in the high load trials than in the low load ones, $t = -2.55$, $p < 0.05$. Further, responses were 327ms faster in the congruent trials than the incongruent ones, $t = -5.21$, $p < 0.001$. Figure shows mean response times in each condition.

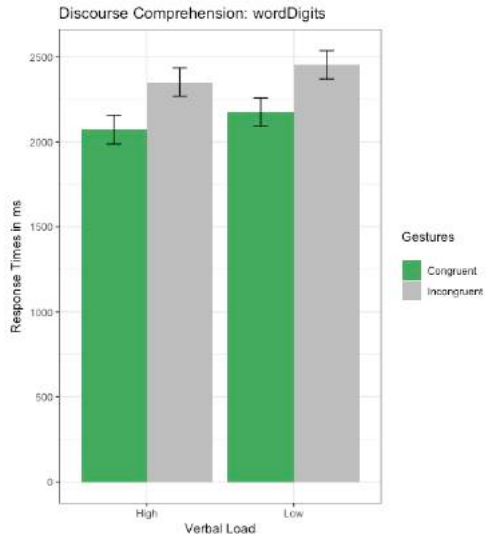


Figure 5: Mean response times for the discourse comprehension task in each condition of Experiment 2. Error bars represent 95% confidence intervals.

General Discussion

Results of the present study provide only modest support for the visuospatial resources hypothesis, and no support for the verbal resources hypothesis. Reducing the availability of visuospatial resources impacted multimodal discourse comprehension, but did not modulate participants' sensitivity to the semantic congruity of co-speech gestures. Likewise, reducing the availability of verbal resources impacted the discourse comprehension task, but did not modulate participants' sensitivity to the semantic congruity of co-speech gestures. In Experiment 1, however, sensitivity to gesture congruity was systematically greater among participants with the greatest visuospatial WM capacity. Thus, while we find no support for a direct causal role of visuospatial WM and speech-gesture integration, visuospatial resources may be relevant to some aspect of gestural processing.

Results of the present study stand in stark contrast to those reported in Coulson & Wu (2014) using the same discourse materials, the same secondary memory task, but that utilized a picture probe to test gesture comprehension rather than the word probes employed here. In tests with picture probes, Coulson & Wu (2014) found that participants were less sensitive to gesture congruity when visuospatial resources were taxed. In the present study, responses to word probes were significantly impacted by gesture congruity, but sensitivity to gestural information was similar under conditions of high and low visuospatial load. This discrepancy might result because the discourse comprehension task in Wu & Coulson (2014) was more taxing than that in the present study. Alternatively, it might be more related to the extent that the picture probe task draws more on the visuospatial resources shared with gesture processing than does the word probe task.

Indeed, the latter interpretation is consistent with the similarity between the impact of verbal memory load in Experiment 2 of the present study with that in the parallel study in Wu & Coulson (2014). Using a picture probe to assess discourse comprehension, they found that performance was impacted both by gesture congruity and by verbal memory load, although the two factors did not interact. Similarly, here we find that performance on the word probe task was independently influenced by gesture congruity and by verbal memory load. The similar impact of verbal versus visuospatial memory load on discourse comprehension as assessed with the word probes employed here also mitigates the concern raised by Wu & Coulson (2014) that the two secondary tasks differ in their demands on central processing resources.

We suggest that the greater impact of the dots task on the processing of picture probes than word probes may be indicative of the role that iconic co-speech gestures play in communication. Congruency effects on the word probe task suggest that speakers readily exploit the information in gestures to detect semantic relationships between novel words and the extant discourse context. However, perhaps because gestures are habitually used to interpret words, this process exerted minimal enough cognitive demands as to resist interference from concurrent demands on either verbal or visuospatial memory systems. By contrast, the picture probe task used by Wu & Coulson (2014) suggested that visuospatial resources were particularly important for detecting a relationship between the pictures and multimodal discourse about concrete topics.

Future research should increase the demands of either the discourse comprehension task or those of the secondary memory tasks in order to elucidate the reason for our failure to observe a differential impact of memory load on sensitivity to gestures. Perhaps titrating memory load demands individually (as in Frank, et al., 2012) will allow us to better estimate its impact on discourse comprehension.

References

Frank, M. C., Fedorenko, E., Lai, P., Saxe, R., & Gibson, E. (2012). Verbal interference suppresses exact numerical representation. *Cognitive psychology*, *64*(1-2), 74-92.

Goldin-Meadow, S., Nusbaum, H., Kelly, S. D., & Wagner, S. (2001). Explaining math: Gesturing lightens the load. *Psychological Science*, *12*(6), 516-522.

Kendon, Adam. (2004). *Gesture: Visible Action as Utterance*. Cambridge University Press.

Wu, Y. C., & Coulson, S. (2014). Co-speech iconic gestures and visuospatial working memory. *Acta Psychologica*, *153*, 39-50.

Subtle differences in language experience moderate performance on language-based cognitive tests

Maury Courtland[†], Aida Davani[‡], Melissa Reyes[*], Leigh Yeh[‡], Jun Leung[*], Brendan Kennedy[‡], Morteza Dehghani[*‡], and Jason Zevin[*†]

[†] Department of Linguistics
[‡] Department of Computer Science
[*] Department of Psychology
University of Southern California
{landerpo, mostafaz, reyesmel, leighyeh, junyenle, btkenned, mdehghan, zevin}@usc.edu

Abstract

Cognitive tests used to measure individual differences are generally designed with *equality* in mind: the same "broadly acceptable" items are used for all participants. This has unknown consequences for *equity*, particularly when a single set of linguistic stimuli are used for a diverse population of language users. We hypothesized that differences in language variety would result in disparities in psycholinguistically meaningful properties of test items in two widely-used cognitive tasks, resulting in large differences in performance. As a proxy for individuals' language use, we administered a self-report survey of media consumption. We identified two substantial clusters from the survey data, roughly orthogonal to *a priori* groups recruited into the study (university students and members of the surrounding community). We found effects of both population and cluster membership. Comparing item-wise differences between the clusters' language models did not identify specific items driving performance differences.

Introduction

Cognitive tests are increasingly used in research on individual differences. For example, a number of recent studies reported correlations between speech perception in noise and working memory (for meta-analysis, see: Dryden, Allen, Henshaw, and Heinrich 2017). Widely used tests for both (Daneman & Carpenter, 1980; Kalikow, Stevens, & Elliott, 1977) were developed without much regard for potential individual differences in language experience, however. This raises the possibility that at least some of the variability in these tasks is related to differences in participants' language experience, as demonstrated in studies of higher-level language processing (Moore & Gordon, 2015; Wells, Christiansen, Race, Acheson, & MacDonald, 2009). Currently, it remains unclear how much this robust correlation between the two tasks – found in 26 of the 30 studies surveyed by Akeroyd (2008) – reveals a correlation between the target constructs or a latent variable of language experience.

Linguists have long considered the communicative capacities of every language to be equal and equally expressive (Joseph & Newmeyer, 2012; Pellegrino, Coup, & Marsico, 2011). Guidelines from the American Speech-Language-Hearing association on cultural competence encourage clinicians to take cultural variables into account in assessing and treating language disorders and differences (American Speech-Language-Hearing Association (ASHA), n.d.). Despite these commitments in allied fields, and the demonstrable existence of multiple American Englishes (e.g. see for review: Labov, Ash, and Boberg 2006; Schneider and Kortmann 2004), most cognitive tests assume "Mainstream" American English (MAE) as a default in the construction of stimuli, potentially confounding cognitive test performance with experience and fluency in MAE. Conversely, language experience is not deterministically related to the usual features that define distinct "dialects" – region, ethnicity, class, etc. People are cosmopolitan and idiosyncratic in the language experiences they seek out, and as a result, may be familiar with multiple language varieties, with potential consequences for their performance on cognitive tests.

Statistical learning, hypothesized to underlie much of language development (Elman, 2001; Seidenberg & MacDonald, 1999), is driven by patterns in language input. Given different input, then, language learners will necessarily construct different distributional models to generate and process speech and language. Online speech and language processing relies heavily on learned statistical regularities to facilitate top-down anticipatory processes. This is evidenced by the effects of surprisal observed when these anticipations are violated (Federmeier, Mai, & Kutas, 2005; Kutas & Hillyard, 1980, 1984). Given the highly demanding nature of online speech and language processing, anticipatory mechanisms help lessen the cognitive effort needed to accomplish the task. The greater the difference between the listener or reader's language model and the statistics of the language material they are processing, the greater the cognitive burden on the listener. For example, intelligibility levels in noise are better for one's own dialect than for a familiar, but less commonly encountered dialect (Clopper & Bradlow, 2008). In children who prefer a non-mainstream English, familiarity with "school English" is associated with performance on literacy tests (Charity, Scarborough, & Griffin, 2004).

The current research examines the effect of variability in language experience on cognitive tests. We hypothesized that measuring people's language experience indirectly, by having them complete a "media diet" survey, would allow us to identify distinct clusters of individuals based on their viewing, listening, and reading habits. We expect these clusters to only loosely covary with the demographic factors that commonly define distinct "dialect" groups. This new measure

of language differences between participants thus provides a novel aspect of individual variability that we expect to moderate performance on language-based cognitive tasks. As this measure probes the role of language directly, it may be more informative in predicting task performance variability than standard demographic information. To test this we recruit from two populations that differ along traditional demographic lines: USC undergraduates – typically high-SES students pursuing higher education (*Facts and Figures | About USC*, n.d.) – and members of the downtown Los Angeles community – mostly African American and Latinx lower-SES individuals, many of whom not pursuing education beyond high school (e.g. the zip code 90062: US Census Bureau n.d.). We administer the aforementioned functional hearing and working memory tasks and expect survey responses to at least partly predict variability in task performance. As we expect this effect to be linguistic, we also predict that language models trained on the media sources will predict participants' behavioral performances.

Methods

Participants

We recruited participants from the USC undergraduate population (N=70) and on a local community college campus (Los Angeles Trade-Technical College, N=25). USC students participated in exchange for course credit and community participants were compensated for their time at $15 per hour, prorated at 20 minute intervals. No requirements were placed on age, but due to recruitment populations, 80% of participants were between the ages of 19 and 26 (mean=22, std=6.25).

Cognitive Tests

To test participants' language abilities, we used the reading span task (Daneman & Carpenter, 1980) that was developed to assess verbal working memory and the speech perception in noise task (SPiN, Kalikow et al. 1977) that was developed to assess functional hearing. The reading span task presents sets of sentences to be read aloud while participants maintain the last word of each sentence in memory. At the end of a set, participants are tested on how many sentence-final words they can recall, and set length is increased until they cannot complete the task. Testing is terminated when participants cannot completely recall any of the three sets of sentences at a particular set length. The SPiN consists of short sentences presented over headphones in 12 talker babble (6 female, 6 male). Participants must identify the final word of the target sentence. We used recordings from the Nationwide Speech Project (Clopper & Pisoni, 2006) to create the stimuli and present trials at +6dB SNR which produced large individual differences in accuracy in pilot results. We choose these tests due to their importance as widely used individual difference measures in clinical populations to diagnose age-related decline (Byrne, 1998), aphasia (Caspari, Parkinson, LaPointe, & Katz, 1998), Alzheimer's (Kempler, Almor, Tyler, Andersen, & MacDonald, 1998), and schizophrenia (Stone, Gabrieli, Stebbins, & Sullivan, 1998). We also choose these tests for the important – but often unacknowledged – role language processing is likely to play in both.

Survey

We constructed an online survey (approx. 20 minutes long) that probes participants' current and formative media consumption habits, elicits short language production passages, and collects basic demographic information. We use this tool to glean each participant's media diet, which forms the basis for later linguistic grouping and analysis. We use the language obtained from the sources participants report as a model for participants' actual language input and a proxy of language experience.

Equipment

Subjects sat in a noise attenuating booth and participated in the survey and behavioral tasks on a desktop PC computer. USC participants were allowed to complete the survey online prior to their lab session. Participants first completed the reading span task, followed by the SPiN, and finally the survey. The reading span task was administered and scored by a researcher to ensure subjects read aloud continuously. Upon completion of each sentence, the researcher advanced the display to the next sentence in the set and solicited verbal responses at the end of each set. After a brief training phase, participants were not given feedback on their performance and were not told their failure had caused the end of the test, simply that it had ended. The SPiN test was administered using Paradigm experiment software; participants typed their responses into a free-response text box. Trials began after a 500ms delay once participants had submitted their response. Stimuli were presented at a comfortable level, standard across participants.

Clustering

We create a media source space in which each dimension represents a reported source (e.g. movie) collected in our survey. Each participant is thus represented as a binary vector in this space, with 1s in dimensions corresponding to sources they consume, and 0s in those they do not. To ensure each dimension is informative (and reduce the dimensionality), we only represent sources reported 10 or more times – thus avoiding dimensions that would only differentiate a few participants (i.e. the rest would all receive 0s in that dimension). This leaves 314 dimensions along which participants were clustered using the k-means algorithm (Lloyd, 1982). Figure 1 shows the distortion values for different numbers of clusters, revealing 3 clusters to be the inflection point at which more clusters provide only marginal returns. The algorithm takes this point as the true number of clusters because increasing the number of clusters beyond this simply subdivides the true clusters, thus over-fitting.

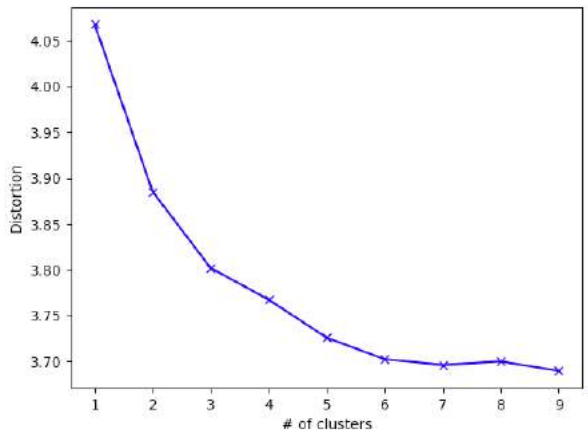


Figure 1: K-means clustering reveals 3 clusters of participants in our media consumption space. This is evidenced by the inflection in distortion decrease that occurs at $k = 3$.

Corpora Construction

We aggregate language data from the sources participants reported in our survey for further linguistic analysis. This produces two corpora (one for each cluster) that allow us to model their language differences. We fully acknowledge the difference between consuming sources as text, as our models do, and speech, as our participants do. Despite this, however, text fully captures the regularities of lexical and supra-lexical features we expect to influence performance on our behavioral tasks.

We collected each corpus by scraping repositories of television scripts (*Springfield! Springfield!*), movie subtitles (*YIFY Subtitles*), and song lyrics (the *Genius API*). For TV and music, we collected all the content for one show (e.g. all the scripts from *Law & Order*) or one artist (e.g. all the songs by *Bruno Mars*). We then pre-process these sources by removing anything the viewer would not hear (e.g. stage directions) and anything non-linguistic (e.g. non-alphanumeric characters or non-verbal noises).

Language and Surprisal Modeling

To model the language statistics of each cluster's corpus, we use 5-gram language models with backoff (Katz, 1987). These models estimate the likelihood of a sentence as the product of the conditional probabilities of its words given the words that precede them. Thus for a sentence of length L, the likelihood is:

$$\prod_{l=1}^{L} P(w_l | w_{l-(n-1)} \ldots w_{l-1}) \tag{1}$$

where n is a hyperparameter set to control the number of preceding words considered for context ($n = 1$ is simply the marginal probability). Because the probability of encountering the preceding string of words in training decreases as the length of the string increases, backoff allows the algorithm to decrease n until the preceding string *has* been seen in training (thus allowing the conditional probability to be estimated). Therefore, while we initially set our $n = 5$, probabilities may be calculated given less prior context.

In addition to the 5-gram model which proceeds from the beginning of a sentence seeking to model its probability, we also model the surprisal associated with encountering the final word of the sentence. This is a particularly important quantity considering both our behavioral tasks use sentence final words as their testing target. While in theory the model aligns with the concept of cloze probability – the probability of the sentence-final word given every preceding word: $P(w_L | w_1 \ldots w_{L-1})$ – this rarely occurs in practice given the sparseness of a training corpus. To model this, we adopt a similar method to n-gram models with backoff. We calculate the conditional probability of the last word given the $n-1$ preceding terms:

$$P(w_L | w_{L-(n-1)} \ldots w_{L-1}) \tag{2}$$

where we initialize $n = 5$ and reduce its value until the preceding string has been encountered in the training corpus ($n = 1$ is simply the marginal probability of the word occurring sentence-finally).

Results

Clustering

The clustering included all reported media sources and revealed three clusters based on participants' consumption habits. Despite a substantial drop in distortion from 2 clusters to 3 (see Fig. 1 for distortions), cluster 0 proved too small to analyze: it contains just 2 participants. Its size precludes both behavorial analysis, which requires an adequate number of samples to be statistically feasible, and computational modeling, which requires a corpus built from an adequate number of reported sources (aggregated across a cluster). Given these limitations, the following analyses will only use clusters 1 and 2 as the sample population (still 98% of the original sample). This clustering, far from an artifact of random seed, proved stable across random restarts. Over 1000 iterations, on average 75% of participants were re-clustered in the same groups (see **Behavioral Data** for the effects on statistical tests).

Regarding cluster membership, we expected USC students and community members to be unevenly distributed between clusters, and this was true, although not categorically. As seen in Table 1, the two are relatively balanced across clusters. Thus, cluster membership and *a priori* group membership are treated as orthogonal in the following analyses.

In addition to the a priori population, we examined the distribution of traditionally considered covariates across the clusters. We wanted to test whether self-reported media consumption provided new information beyond existing measures (i.e. we were not just capturing an existing highly correlated dimension of variance). As seen in Table 1, typical demographic variables were fairly evenly distributed across the clusters. One-way chi-square tests revealed that none of the demographic variables significantly differed from an even

split across clusters (i.e. the expected values if cluster and variable were independent).

Variable	Level	Cluster Ns		Cluster %	
		1	2	1	2
Population	USC	34	24	59%	41%
	LATTC	7	13	35%	65%
Gender[1]	Female	33	21	61%	39%
	Male	7	16	30%	70%
Schooling	High School	10	10	50%	50%
	Associate	4	4	50%	50%
	Some College	19	15	56%	44%
	Bachelor's	7	6	54%	46%
	Master's	1	2	33%	66%
Mono-lingual	True	10	11	48%	52%
	False	31	26	54%	46%
SES Self-Report[2]	High	13	13	50%	50%
	Medium	16	10	62%	38%
	Low	12	14	46%	54%

Table 1: The distribution of traditionally considered covariates across clusters is fairly even. We observe no obvious imbalance between clusters along any demographic dimensions our survey measured. One-way chi-square tests confirm this.

Given the orthogonality of self-reported media consumption to traditional demographic variables, we hereafter focus on the observed dimension of variance: media diet. We probe how the clusters differ in their media habits in order to delineate their makeup. We examine the clusters' centroids to calculate which dimensions (i.e. sources) they differ maximally along. This provides a measure of which media sources are most distinct between clusters. We find the following sources to be the 5 most different between clusters 1 and 2 and provide the difference in mean consumption between the two (i.e. $\bar{x}_1 - \bar{x}_2$) in parentheses: Star Wars (.64, specific films reported in the series were less powerful, on the order of .11-.17), Yes! (-.47), CNN (-.26), People (-.12), and Harry Potter (-.12). We hesitate to draw any conclusive generalities on the two clusters' media diets, but at a glance it appears that cluster 1 consumes lots of high fantasy (Star Wars, Lord of the Rings, The Chronicles of Narnia, etc.) while cluster 2 consumes more nonfiction (Yes!, CNN, People, etc.).

Behavioral Data

As seen in Figure 2, the SPiN task revealed a main effect of cluster, $F(1,74) = 7.30$, $p < .01$, but no main effect of population and no interaction between the two. In the reading span data, we again find a main effect of cluster, $F(1,74) = 4.05$, $p < .05$, and a main effect of population $F(1,74) = 13.57$, $p < .001$, and no interaction between the two. In our 1000 clustering iterations, 63% of iterations revealed statistically significant effects of cluster on the SPiN task (at $\alpha = .05$). This was not replicated with the span task, however: only 4% of our iterations found statistically significant effects.

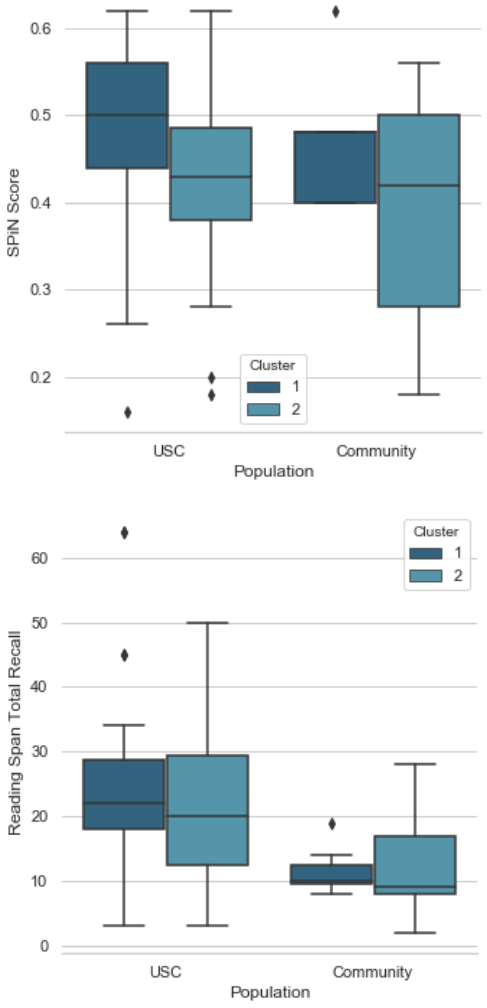


Figure 2: Results from the SPiN test reveal a significant difference between clusters but not populations. Reading span also shows an effect of cluster, but a larger effect of population. We observe no significant interactions.

The span test will play a minor role in further analyses, due to the difficulty in handling test result data and its scoring. Because the span task is terminated whenever participants fail to recall a set, participants provide unequal numbers of observations. The analyses are additionally constrained by the small number of items a typical participant completes. While observations exist for items later in the test, they are for a few extraordinary participants. This presents a problem not only in the paucity of observations, but also in the fact that these participants are unrepresentative of the general sample in their task abilities. As such, both item-level statistics and graphical representations are challenging.

Our survey obtains several pieces of demographic infor-

[1] One participant in cluster 1 chose not to report gender.

[2] Participants reported their SES on a continuous scale. Here, we bin responses into 3 quantiles to report distribution across clusters.

mation that are traditionally considered relevant covariates of performance on our cognitive tasks, such as socioeconomic status (SES, self-reported), age, education level, and monolingual status. None of these correlated significantly with performance on either task.

Language Media Input Modality

The above findings of differences between cluster performances motivated us to explore differences between clusters' survey behavior (other than the categorical responses which were used in clustering) to explain their performance data. In particular, we wondered whether the stronger task performances of cluster 1 might be due to increased experience with the tasks of speech perception and reading.

To probe this, we tested whether cluster 1 reported significantly more speech sources (TV, Movies, Music, and News shows) and significantly more text sources (Books, Newspapers, Magazines, Online News, and other online reading) than cluster 2. Indeed, we find that cluster 1 participants report significantly more listening on average than cluster 2: $t(42.82) = 3.09, p =< .005, d = 0.67$ (a medium effect). We also find that cluster 1 participants report significantly more reading on average than cluster 2: $t(72.6) = 5.10, p < .001, d = 1.13$ (a large effect). This may indicate an effect of modality-specific training on task performance. To probe this, we test the correlation between the number of speech sources a participant reports and their SPiN task performance. We test rank correlation rather than linear correlation as we are unsure of the linearity of the relationship between number of sources and modality-specific benefit, as well as to control for the effect of outliers in both performance and reporting volume. We observe a significant correlation between the two: $\rho(76) = .31, p = .005$. We do not, however, observe a significant correlation between number of text sources and span performance.

We also tested whether past modality preference (solicited with "when you were growing up...") would relate to current modality preference. We find a strong correlation between the amount of spoken language items reported growing up and amount of current items reported: $r(76) = .91, p < .001$. This correlation extends to the number of written language items, although not as strongly: $r(76) = .49, p < .001$.

Language Models

To evaluate the claim that our language models were capturing meaningful statistical regularities in the language of each cluster's corpus, we tested whether the log-likelihood produced by a model for each of the test items would correlate with mean performance on those items for the cluster. We do not observe a significant correlation between cluster 1's 5-gram model and performance on either the SPiN ($r(48) < 0.01$) or span ($r(25) = -0.01$). We also observe no significant correlation between cluster 2's 5-gram model and its performances on SPiN ($r(48) = -0.02$) or span ($r(25) < 0.01$). Additionally, we tested the correlation between cluster 1's 5th-order surprisal model and its performance and found no correlations with SPiN ($r(48) = 0.07$) or span ($r(25) = -0.09$). Similar results were obtained for cluster 2 (SPiN: $r(48) = 0.21$, span: $r(25) = 0.02$).

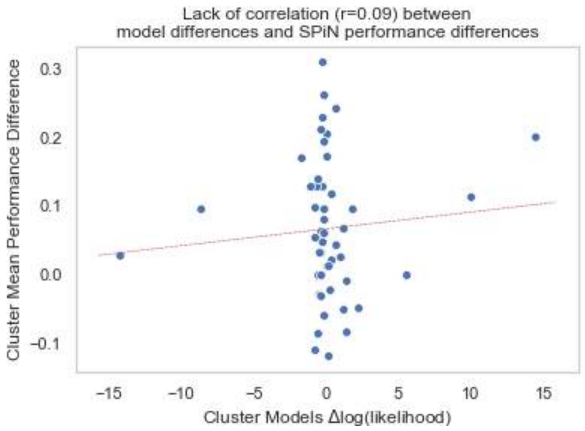


Figure 3: The non-correlation of cluster performance differences with model likelihood differences indicates the statistical information captured by the models is a poor predictor of behavioral performance. The significant cluster performance difference can be seen here by the majority of items occurring above 0-difference on the y-axis. A LMS-Regression line is drawn in red for reference.

In addition to modeling statistical properties of particular items, we also tested whether the difference between the language and surprisal models might capture the significant differences we see on our behavioral tasks. This method avoids any idiosyncrasies of particular items (as comparisons are within item) and instead captures any language differences of media sources. We again find a lack of significant correlation between 5-gram likelihood differences and task performance for both the SPiN ($r(48) = 0.09$) and span ($r(25) = .25$). Similar results are observed for the 5th-order surprisal model (SPiN: $r(48) < 0.01$, span: $r(25) = 0.08$). As shown in Fig. 3, differences between the model likelihoods are close to zero for most items, with a few outliers.

To examine the non-correlations and clustering around 0 on Fig. 3's x-axis, we tested the correlation between models and found strong correlations for both the SPiN ($r(48) = 0.94, p < .001$) and span ($r(86) = 0.97, p < .001$) test. These strong correlations, coupled with linear regression slopes of $\beta_1 = 0.96$ (SPiN) and $\beta_1 = 0.94$ (span) imply nearly identical log-likelihood scores between models despite training on categorically different sources. While the results reported here are specific to 5-gram language models and 5th-order surprisal models, other lower-ordered models of both yielded similar results.

Discussion

We observed significant performance differences on a speech perception in noise task and a working memory task between

clusters of participants derived from self-reported media consumption. These differences were above and beyond differences driven by *a priori* participant groups – students at a university vs. participants from the surrounding community. This clustering was robust to randomness and orthogonal to any traditionally considered demographic variables. As we have no reason to believe that the tests' target constructs systematically vary between our clusters, we conclude that media diet represents an uncorrelated latent variable moderating task performance. To our knowledge, our identifying media consumption as a significant orthogonal predictor of cognitive task performance is a novel contribution of this work.

This novel predictor is surprisingly powerful at explaining language test performance considering its complete lack of explicit linguistic information. In pursuing a linguistic explanation for our finding, we used statistical language models trained on sources participants reported consuming to analyze test items. These models did not identify particular stimuli driving performance differences, and we found no obvious differences in how well stimuli fit our models. However, a follow-up study we performed with more complex recurrent neural models did in fact reveal a correlation between models trained on our media corpora and behavioral performance (Courtland et al., 2019). This implies the statistics used here are not sophisticated enough. Cloze probability, for example, is computed as a simple ratio of the tokens of a word in context to all tokens in that exact context in the corpus.

Also of note is the highly significant difference in the number of sources reported by cluster 1 compared with cluster 2. It is possible that the greater number of sources indicates that cluster 1 contains more voracious consumers of media than cluster 2. This increased media consumption in the modalities of our tests may be providing cluster 1 members with modality-specific training they leverage at test time. Indeed, the correlation we observed between number of speech sources reported and SPiN performance supports this explanation. This is especially plausible given that watching a TV show or movie involves perceiving character dialog often obscured by various sources of noise (soundtrack, sound effects, etc.). It is also possible, however, that the increased responses and performance from cluster 1 is indicative not of their increased modality-specific training but rather a latent variable such as attentiveness or enthusiasm at participating in all aspects of the study.

It should be mentioned that participants' responses may reflect a (possibly implicit) choice to make specific habits known in the context of the survey. Given the importance of shared experience in forming relationships, what pieces and types of information people share and what they keep private often acts as a type of signaling that forms the basis of social cohesion. Thus, media diet survey responses may be more appropriately interpreted as signalling membership in a language community than literally reflecting the language practices of that community. Indeed, the vast majority of items in the corpora are professionally produced texts, which are likely to differ less than spontaneous spoken and written communication. In future work, we plan to obtain rich, naturalistic language samples in addition to the media corpora included so far to strengthen the evidence found here.

The identification of a dimension (other than the target construct) that test performance differs significantly along brings into question not only specific test validity probed here, but also the validity of the entire practice of test item standardization. This is true whether this dimension is categorical media consumption, shown here, or the linguistic content of the media, shown in Courtland et al. (2019). Tests that use language to probe target constructs must take the language of their test into account – not as a static entity to be standardized, but as the diverse and dynamic communication medium that it is. Test validity relies on the ability to generalize a test's result to participants' everyday behavior. This is only valid if the test is representative of the language they encounter in their daily lives (Coleman, 1964). Thus, tests employing standardized language not only contain inherent inequity for those less familiar with the test language, they are also less valid.

Here we aim to show that participants' diverse language experiences must be taken into account when diagnostic tools like those tested here are designed. Ideally, given the unique nature of language experience, test creators should strive to create tests that present equal difficulty to each participant by using personalized test language. This step to ensure equity is especially important given that test scores cannot simply be adjusted for using traditionally defined dialectal boundaries – as demonstrated here by the uninformativeness of the demographic variables that define these boundaries.

Generating equitable stimuli is a difficult or possibly infeasible task for human researchers, but could potentially be automated using generative models. If such models were driven by statistics that are highly representative of participants' language experience, they may do a better job of capturing cognitive constructs without smuggling in variability resulting from differences in language experience. Perhaps the most exciting future direction of this research will be to facilitate using more representative language statistics in designing stimuli for cognitive tests. The study of how language experience influences test performances that we take here represents a first step to understanding and mitigating this test inequity.

Acknowledgments

This work was supported by the NIH (5R21DC017018-02) and data were collected under USC IRB #UP-18-00006.

References

Akeroyd, M. A. (2008, January). Are individual differences in speech reception related to individual differences in cognitive ability? A survey of twenty experimental studies with normal and hearing-impaired adults. *International Journal of Audiology*, *47*(sup2), S53–S71. doi: 10.1080/14992020802301142

American Speech-Language-Hearing Association (ASHA). (n.d.). *Cultural competence.*

Byrne, M. D. (1998, June). Taking a computational approach to aging: The SPAN theory of working memory. *Psychology and Aging*, *13*(2), 309–322. doi: 10.1037/0882-7974.13.2.309

Caspari, I., Parkinson, S. R., LaPointe, L. L., & Katz, R. C. (1998, July). Working Memory and Aphasia. *Brain and Cognition*, *37*(2), 205–223. doi: 10.1006/brcg.1997.0970

Charity, A. H., Scarborough, H. S., & Griffin, D. M. (2004). Familiarity with school english in african american children and its relation to early reading achievement. *Child development*, *75*(5), 1340–1356.

Clopper, C. G., & Bradlow, A. R. (2008, September). Perception of Dialect Variation in Noise: Intelligibility and Classification. *Language and Speech*, *51*(3), 175–198. doi: 10.1177/0023830908098539

Clopper, C. G., & Pisoni, D. B. (2006). The Nationwide Speech Project: A new corpus of American English dialects. *Speech Communication*, *48*, 633–644.

Coleman, E. B. (1964, February). Generalizing to a Language Population. *Psychological Reports*, *14*(1), 219–226. doi: 10.2466/pr0.1964.14.1.219

Courtland, M., Davani, A., Reyes, M., Yeh, L., Leung, J., Kennedy, B., Dehghani, M., & Zevin, J. (2019). Modeling performance differences on cognitive tests using LSTMs and skip-thought vectors trained on reported media consumption. In *Proceedings of NLP+CSS: Workshops on Natural Language Processing and Computational Social Science,*. Minneapolis, MN.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of verbal learning and verbal behavior*, *19*(4), 450–466.

Dryden, A., Allen, H. A., Henshaw, H., & Heinrich, A. (2017, December). The Association Between Cognitive Performance and Speech-in-Noise Perception for Adult Listeners: A Systematic Literature Review and Meta-Analysis. *Trends in Hearing*. doi: 10.1177/2331216517744675

Elman, J. L. (Ed.). (2001). *Rethinking innateness: a connectionist perspective on development* (1. MIT Press paperback ed., 5. print ed.). Cambridge, Mass.: MIT Press.

Facts and Figures | About USC. (n.d.). Retrieved 2018-10-22, from `https://about.usc.edu/facts/`

Federmeier, K. D., Mai, H., & Kutas, M. (2005, July). Both sides get the point: Hemispheric sensitivities to sentential constraint. *Memory & Cognition*, *33*(5), 871–886. doi: 10.3758/BF03193082

Joseph, J. E., & Newmeyer, F. J. (2012). 'All languages are equally complex': The rise and fall of a consensus. *Historiographia Linguistica*, *39*(2-3), 341–368. doi: 10.1075/hl.39.2-3.08jos

Kalikow, D., Stevens, K., & Elliott, L. (1977). Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *The Journal of the Acoustical Society of America*, *61*(5), 1337–1351.

Katz, S. M. (1987). Estimation of probabilities from sparse data for the language model component of a speech recognizer. In *IEEE Transactions on Acoustics, Speech and Signal Processing* (pp. 400–401).

Kempler, D., Almor, A., Tyler, L. K., Andersen, E. S., & MacDonald, M. C. (1998, October). Sentence Comprehension Deficits in Alzheimer's Disease: A Comparison of Off-Line vs. On-Line Sentence Processing. *Brain and Language*, *64*(3), 297–316. doi: 10.1006/brln.1998.1980

Kutas, M., & Hillyard, S. A. (1980, January). Reading senseless sentences: brain potentials reflect semantic incongruity. *Science*, *207*(4427), 203–205. doi: 10.1126/science.7350657

Kutas, M., & Hillyard, S. A. (1984, January). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, *307*(5947), 161–163.

Labov, W., Ash, S., & Boberg, C. (2006). *The atlas of North American English: phonetics, phonology, and sound change: a multimedia reference tool*. Berlin ; New York: Mouton de Gruyter.

Lloyd, S. (1982, March). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, *28*(2), 129–137. doi: 10.1109/TIT.1982.1056489

Moore, M., & Gordon, P. C. (2015). Reading ability and print exposure: item response theory analysis of the author recognition test. *Behavior research methods*, *47*(4), 1095–1109.

Pellegrino, F., Coup, C., & Marsico, E. (2011). Across-Language Perspective on Speech Information Rate. *Language*, *87*(3), 539–558. doi: 10.1353/lan.2011.0057

Schneider, E. W., & Kortmann, B. (Eds.). (2004). *A handbook of varieties of English: a multimedia reference tool*. Berlin ; New York: Mouton de Gruyter.

Seidenberg, M. S., & MacDonald, M. C. (1999, October). A probabilistic constraints approach to language acquisition and processing. *Cognitive Science*, *23*(4), 569–588. doi: 10.1016/S0364-0213(99)00016-6

Stone, M., Gabrieli, J. D. E., Stebbins, G. T., & Sullivan, E. V. (1998, April). Working and strategic memory deficits in schizophrenia. *Neuropsychology*, *12*(2), 278–288. doi: 10.1037/0894-4105.12.2.278

US Census Bureau. (n.d.). *American Community Survey Data.* Retrieved 2018-10-22, from `https://www.census.gov/programs-surveys/acs/data.html`

Wells, J. B., Christiansen, M. H., Race, D. S., Acheson, D. J., & MacDonald, M. C. (2009). Experience and sentence processing: Statistical learning and relative clause comprehension. *Cognitive Psychology*, *58*(2), 250–271.

Efficiency of Learning in Experience-Limited Domains: Generalization Beyond the WUG Test

Christopher R. Cox (chriscox@lsu.edu)
Department of Psychology, Louisiana State University
1005 Field House Dr, Baton Rouge, LA 70802 USA

Matthew Cooper Borkenhagen and **Mark S. Seidenberg**
Department of Psychology, University of Wisconsin-Madison
1202 W. Johnson Street, Madison, WI 53706 USA

Abstract

Learning to read English requires learning the complex statistical dependencies between orthography and phonology. Previous research has focused on how these statistics are learned in neural network models provided with as much training as needed. Children, however, are expected to acquire this knowledge in a few years of school with only limited instruction. We examined how these mappings can be learned efficiently, defined by tradeoffs between the number of words that are explicitly trained and the number that are correct by generalization. A million models were trained, varying the sizes of randomly-selected training sets. For a target corpus of about 3000 words, training sets of 200–300 words were most efficient, producing generalization to as many as 1800 untrained words. Composition of the 300 word training sets also greatly affected generalization. The results suggest directions for designing curricula that promote efficient learning of complex material.

Keywords: reading; efficient learning; generalization; computational modeling; human and machine learning

Introduction

Generalization—the ability to apply existing knowledge to novel cases—is an important capacity observed, with varying complexity, in many species (Santolin & Saffran, 2018). Human generalization encompasses a broad range of behaviors, ranging from generalizations about the properties of three dimensional space to ones based on physical appearance. The behavioral and neurobiological bases of generalization are a focus of much research (e.g., Goldberg, 2009; Zhang, Bengio, Hardt, Recht, & Vinyals, 2016).

Generalization is especially important in language acquisition and learning to read. Children rapidly acquire knowledge that allows them to generalize beyond the limited sample of utterances they experience (Chomsky, 1965). The classic demonstration is the WUG Test (Berko, 1958). A child who has learned about plural formation can generalize to novel cases: one wug, two wugs. Similarly, a beginning reader who has learned correspondences between spelling and pronunciation can read aloud nonce words such as NUST and GLORP (Seidenberg & McClelland, 1989). Generalization has traditionally been taken as evidence for symbolic rules, but it is also observed in neural networks of varying complexity (Seidenberg & Plaut, 2014; LeCun, Bengio, & Hinton, 2015).

Our research examined generalization from a different perspective, efficiency of learning. Efficiency is a concern in real-world contexts in which, unlike most machine learning applications, learning opportunities are constrained. For example, children's vocabulary development depends on their time- and context-limited exposure to spoken language, which varies considerably (Hart & Risley, 1995; Gilkerson et al., 2017). The resulting differences in vocabulary size and quality have an enormous impact on learning to read and other aspects of schooling (Seidenberg, 2017). Knowledge gaps cannot be closed solely through explicit instruction because there isn't sufficient classroom time. The same holds for learning mappings between written and spoken language. Instruction ("phonics") is helpful, but only a small subset of patterns can be taught. In these and other knowledge domains, children learn from relatively limited data and generalization is paramount.

In the classic WUG test generalization is assessed by performance on nonce forms or, in machine learning, withheld words. The exact composition of the examples that support generalization is not the focus of attention, but is critical in experience-limited domains. We therefore re-formulated the generalization question as follows, using spelling-sound knowledge as a test case:

- Children need to acquire the ability to generate pronunciations for many written words (the target set);
- They are explicitly taught the correspondences between orthography and phonology for a much smaller subset of words (the training set);
- Generalization is assessed in terms of correct performance on untrained items from the target set, rather than nonce forms. This shifts the focus of generalization to acquiring real-world knowledge.

The research question is then how the size and composition of the training set affects generalization to untrained items. Learning is efficient if the ratio between the number of trained items and the number of generalization items is low. We examined efficiency of learning as a function of the size of the training set using simple, well-studied models of learning orthography-phonology correspondences (Seidenberg & McClelland, 1989; Harm & Seidenberg, 1999). We also examined how efficiency was affected by the composition of a training set of a given size. The results suggest that it may be possible to structure children's reading experiences in ways that promote more efficient learning.

Materials and Methods

Words

The simulations used a set of 2881 monosyllabic English words employed in previous research (Harm & Seidenberg, 1999). Word length ranged from 2–8 letters and 1–7 phonemes.

Model architecture

The model was a simple feedforward network with an input orthographic layer (102 units), an output phonological layer (66 units) and a single hidden layer (100 units). It was structured and trained in standard ways, with weights updated with gradient descent and backpropagation after accumulating cross-entropy error over all words in the training set.

Orthographic representations were generated as follows. Words were centered on the vowel (or the first vowel in a digraph), adding empty letters to the onset as necessary. If the first vowel was followed immediately by a consonant, an empty letter was also added between them, except in cases where the consonant is voiced as part of the vowel (e.g., the letter *w* in SAW). The letter *y* was treated as a consonant when it began a word and a vowel otherwise. Finally, empty letters were added to the end of each word, resulting in orthographic codes of uniform length (14 letters including empty ones).

Each letter was represented by one unit in a 26 element vector, with no units activated for the empty letter. The 14 vectors were concatenated to represent each word. To make these representations more concise, they were stacked to create a 2881×364 matrix, and all-zero columns were dropped, leaving 102 units.

Phonological word forms were represented using 41 phonemes (26 consonants, 15 vowels). They were aligned on the first vowel, adding empty phonemes at the beginning or end to produce phonological representations of equal length (10 phonemes including empty phonemes). Each phoneme was defined by 25 phonetic features (Harm & Seidenberg, 1999). The 10 phoneme by 25 feature vectors were condensed by eliminating nodes for unused features, resulting in an output layer with 66 features.

The model was implemented using scikit-learn in Python 3.6 using a multilayer perceptron, and training was executed in parallel using HTCondor (Thain, Tannenbaum, & Livny, 2005) and computational resources maintained by the Center for High Throughput Computing at UW Madison.

Model training

One million models were run, each using a set of words sampled randomly without replacement from the 2881 word target set. Training sets ranged from 100 to 1000 words in increments of 100, with an equal number of each size.

Each model was trained for 3000 weight updates with a constant learning rate (0.1). The model was exposed to the whole training set before each update. Each model was then tested on the untrained remainder of the target corpus to evaluate generalization. An output pattern was scored as correct if all unit activations were within 0.5 of their target state.

Model evaluation

Using all untrained words as the holdout set to evaluate generalization performance for each model means that the holdout set is not held constant. This is a deliberate design decision: when a word is explicitly trained, it no longer needs to be generalized to. Training on exceptional, irregular words may be the only way to accurately produce them—that explicit training not only develops the model to encode that orth-phon relationship, but also removes that exceptional word from the generalization set. On the other hand, this exceptional word may not teach the model anything generally useful. The give and take between what is in the training set or test set is central to the research question.

An alternative approach is possible, where a single test set is constructed a priori and used for all generalization. This has the advantage of serving as a true benchmark, but poses a critical challenge. It requires composing a representative test set that expresses all relational othographic and phonological structure. Our attempts at dimensionality reduction on the model representations that map between orthography and phonology for the full corpus indicate that 50 dimensions are necessary to express 80% of the variance in that structure. Sampling representatively from that high dimensional space would be necessary for constructing a useful benchmark test set. The problem of constructing this test set is the same as the problem of constructing a representative and efficient training set, and does not have a simple solution.

Results

Training set size and generalization

Figure 1A shows generalization to untrained items as a function of training set size. Smaller training sets afford more opportunities for generalization, but are less able to provide representative coverage of the corpus. Increasing the size of the training set produced diminishing generalization returns. Increasing training sets beyond 500 words did not yield greater

Size	Mean	(Ratio)	Max	(Ratio)
100	333	(3.33)	590	(5.90)
200	889	(4.45)	1252	(6.26)
300	1240	(4.13)	1546	(5.15)
400	1404	(3.51)	1634	(4.08)
500	1469	(2.94)	1668	(3.34)
600	1484	(2.47)	1654	(2.76)
700	1470	(2.10)	1618	(2.31)
800	1438	(1.80)	1566	(1.96)
900	1395	(1.55)	1510	(1.68)
1000	1344	(1.34)	1444	(1.44)

Table 1: Mean and maximum generalization performance over 100k models fit with each training set size. Ratios divide the previous descriptive statistic by the training set size.

generalization.

Figure 1B shows total number of words correct (trained and generalized). No model produced correct performance for all words. Some words were only learned if they were included in the training set; they were never produced correctly by generalization. These include words with highly atypical spellings and pronunciations such as SIXTH, DRAUGHT, SCHEME, COUPS, and JINX.

Figure 1C shows an index of *training set efficiency*, defined as the number of words correct by generalization divided by the number of words trained. Training sets with 100 words are less efficient than those with 200 words on average and in the limit, indicating that the larger set captures more of the structure relevant to untrained words. Training sets of 300 words are somewhat less efficient than those with 200, but after 300 words efficiency drops rapidly. Taking all three metrics into account, 300 words appears to be a sweet spot (see also Table 1).

Analyses of training environments containing 300 words show that they yielded reading vocabularies of 1540 words on average ($SD = 76.62$) and 1846 words at best (failing to decode 1035). Given that efficiency is a primary concern for early reading curricula, it is noteworthy that this is 75.5% of the largest reading vocabulary achieved with any training set (2444 words, achieved after training on 1000 words). Note that this 598 word increase required growing the training set by 700 words. If we subtract the training set from all reading vocabularies and just focus on words that were generalized to, the best model trained on 300 words (1546) achieves 92.7% of the maximum amount of generalization achieved with any training set (1668, achieved after training on 500 words).

These results indicate that nearly all systematic structure relating English orthography and phonology within our corpus of 2881 monosyllabic words can be learned from an appropriately constructed 300 word subset. It is possible to establish a reading vocabulary of over 1800 words based on explicit training on only 300 words, a 6-fold return on instructional investment. However, achieving this level of performance is highly dependent on the composition of the training set: the best and worst models trained with 300 words are separated in performance by over 600 accurate generalizations (min: 906; max: 1546). Thus, in future work it will be important to understand how properties of training sets are related to generalization.

What makes a word likely to be correct by generalization?

The rates at which individual words were correct by generalization across training sets varied greatly, forming a roughly bimodal distribution (Figure 2).

At one extreme are words that are correct by generalization with almost any random selection of training words; at the other are words that for which generalization is highly sensitive to training set composition. The former contain spelling patterns and orthography-phonology mappings that occur more often in this corpus; the latter words have less common patterns and more idiosyncratic mappings.

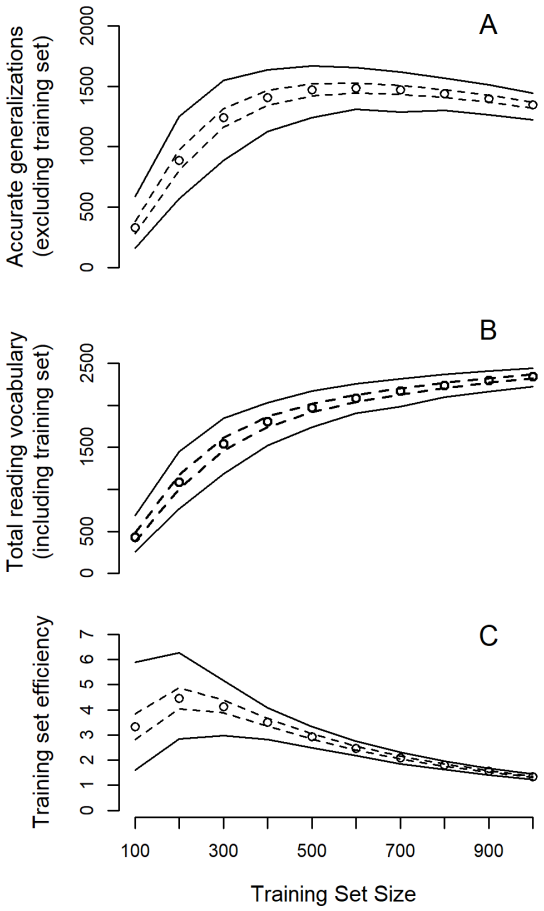


Figure 1: Reading vocabulary size and generalization ability for increasing training set sizes. A) The number of accurate generalization peaks at lower training set sizes and B) the rate of reading vocabulary growth slows. No model trained on a subset of words is capable of reading all words. C) The ratio of generalization performance and training set size, efficiency, is highest with training sets with 200–300 words. Dots indicate the mean; dotted lines are $\pm 1SD$; solid lines are minimum and maximum values.

Whether a word was likely to be generalized to was related to quantifiable measures of orthographic, phonological, and relational (mapping) typicality. We examined several lexical factors that have been employed in previous research:

- Word length: number of letters
- Orthographic neighborhood: number of words whose spelling differs from a word by a one letter substitution, deletion, or addition ($D_{Levenshtein} < 1$).
- Phonological neighborhood: number of words with the same rime (e.g., for "must", the "ust" words like "dust" and "lust").

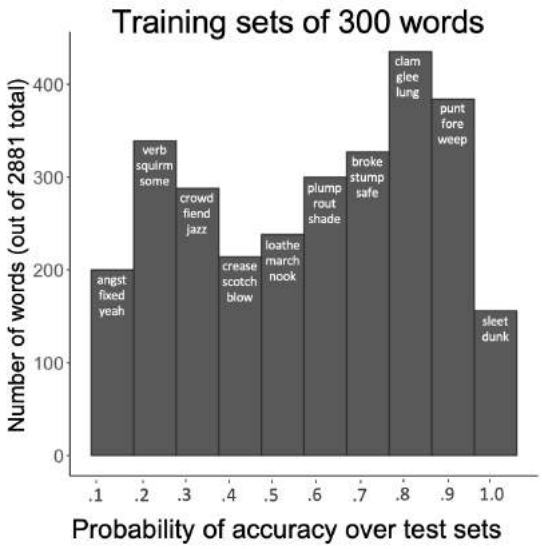


Figure 2: When aggregating over the 100k 300-word model training environments, each word occurs in many test sets. The proportion of times a word occurs in the test set and is accurately generalized to corresponds to how difficult that word is to learn. Representative words belonging to each bin are displayed.

- Consistency: the proportion of words with a given word body (the orthographic equivalent of the rime) and the same phonological rime (e.g., for GAVE, the proportion of -AVE words pronounced "ave"; (Plaut, McClelland, Seidenberg, & Patterson, 1996).

The correlations among these variables, and between these variables and the probability of accurate generalization, are reported in Table 2. The number of orthographic neighbors tends to decrease as word length increases ($r = -0.65$); a similar but weaker trend applies to the size of phonological neighborhoods ($r = -0.28$). This is representative of the English language in general. There is also a moderate relationship between neighborhood size across modalities, such that words that belong to large othographic neighborhoods are expected to belong to large phonological neighborhoods ($r = 0.35$). That this correlation is not higher demonstrates the asymmetry of structure across the modalities. The consistency of a word's pronunciation given its orthography, however, is uncorrelated with the modality-specific metrics. Words are more likely to be generalized to if they are short, belong to large phonological and (especially) orthographic neighborhoods, and have consistent pronunciation given their spelling (Table 2, bottom row).

Given the high correlations among variables, and to gain perspective on how jointly-predictive these factors are of the probability of accurate generalization, we regressed the probability of accuracy over test sets on all four variables in an additive linear model (no interaction terms). This simple model accounts for 39% of the variance in generalization accuracy. Of the variables we considered, the consistency metric accounted for the most unique variance ($\Delta R^2 = 0.15$), but orthographic neighborhood size was a close second ($\Delta R^2 = 0.13$). Once accounting for other variables, phonological neighborhood size and word length did not appreciably improve the model.

These results are broadly consistent with previous research. Effects of spelling-sound consistency have been observed in many behavioral studies of skilled and beginning readers (Jared, McRae, & Seidenberg, 1990), and simulated in earlier models that examined performance over the course of learning many words (Seidenberg & McClelland, 1989; Plaut et al., 1996). Our results suggest that factors that affected ease of learning in the earlier models also affect probability of generalization as studied in the present work.

Out of the variables we considered, phonological neighborhood size is the most studied in the context of word acquisition, where it is understood to influence the order in which words are acquired (Storkel, 2003). Orthographic neighborhood size is often studied in terms of performance, specifically visual word recognition and lexical access (Andrews, 1997). It is also negatively correlated with age of acquisition norms, which indicates that words with more dense orthographic neighborhoods tend to be learned earlier (Cameirão & Vicente, 2010). Words with consistent orthographic to phonology relationships are also processed more efficiently (Ziegler, Ferrand, & Montant, 2004).

	WL	ON	PN	Con.
Word Length	1.00			
Orth. Neighbors	-0.65	1.00		
Phon. Neighbors	-0.28	0.35	1.00	
Consistency	-0.03	-0.02	-0.02	1.00
P(accuracy)	-0.27	0.47	0.28	0.38

Table 2: Correlation among lexical measures. The bottom row reports the pairwise correlation of each variable with the probability of generalization accuracy for each word, defined as the number of times accurately generalized to divided by the number of test sets a word appeared in.

regressor	η_p^2	$\Delta\mathbf{R^2}$
Word length	0.01	0.00
Orth. Neighbors	0.17	0.13
Phon. Neighbors	0.03	0.02
Consistency	0.20	0.15

Table 3: Effect sizes for the regressors that account for variance in the probability of accurately generalizing each word. These effect size metrics are perspectives on the unique variance explained by each variable. Because of collinearity among the regressors, the sum of the ΔR^2 values will be less than total $R^2 = 0.39$.

What makes a good training set?

The word-level features reviewed above give some insight into which words will tend to be generalized to, and which will not, in the context of any given training set. The deeper question pertains to the qualities of the training set foster the most efficient generalization to untrained words in the language. One angle on this question is to consider that the word-level features are in fact reflective of how the word is situated relative to the broader linguistic environment. While we did not test this directly, it is plausible to assume that neighborhood size predicts how likely a word is to be generalized to. Good training sets are *representative* of the broader environment. If a neighborhood is split across training and test sets, the consequence is that the neighbors in the test set have representation within the training set. Given that we randomly split our corpus into training and test sets, there is no guarantee that neighborhoods are efficiently split in this way. However, words that belong to larger neighborhoods are more likely to be split across training and test sets by chance, so we might expect that training sets with larger orthographic and phonological neighborhoods on average will foster more generalization. It is clear that words with no orthographic neighbors ($n = 271$) are generalized to far less often (median probability 0.10) than words with at least one neighbor (median probability 0.56).

Such a crude metric, however, would be largely insensitive to the relative composition of the two sets. For instance, training sets that contain many words with large neighborhoods may simply contain all the words belonging to those large neighborhoods. Such a training set would be unrepresentative of the test set, and unlikely to foster generalization. What we would rather know is each word's neighborhood size relative to the number of its neighbors that also belong to the training set.

On the other hand, orthographic and phonological neighborhood structure is only helpful to the extent that they are aligned. An orthographic neighborhood populated with words with irregular and idiosyncratic pronunciations is not likely to foster generalization on a reading-aloud task. Thus, training sets that have a large and varied collection of words with consistent pronunciations may be expected to generalize well. While it is easy to determine the mean consistency of a training set, it is less clear how to account for the variability across consistent relationships and determine the representativeness of such relationships to the target environment.

We regressed the generalization performance of the 100,000 models trained on 300 word training sets on the mean word length, orthographic and phonological neighborhood sizes, and consistency over all 300 words in each set. The effect sizes are reported in Table 4. We see that, despite being a very crude measure, mean orth-phon consistency accounts for about 13.6% of the variance unexplained by the other variables, indicating that item level characteristics may provide insight on how to construct efficient training sets. However, the vast majority of variance remains unexplained and provides fertile ground for continued research.

regressor	η_p^2	$\Delta\mathbf{R^2}$
Word length	0.002	0.001
Orth. Neighbors	0.006	0.005
Phon. Neighbors	0.000	0.000
Consistency	0.137	0.136

Table 4: Effect sizes for the regressors used to account for variance in generalization accuracy over the 100,000 models fit to random 300 word training sets. Generalization was to all untrained words in the corpus. Because of collinearity among the regressors, the sum of the ΔR^2 values will be less than total $R^2 = 0.14$.

Discussion

We have established a computational procedure for investigating two aspects of generalization in learning basic reading skills: how many words need to be learned to generalize to real English words yet to be learned, and what aspects of reading vocabulary promote this transfer. Our findings indicate that while printed vocabulary continues to grow along with the number of words taught, the efficiency of learning does not grow along with it.

These findings are relevant to real-world learning conditions. As a human teacher grows the number of words they would like to teach, the amount of learning time needed grows along with it. Our findings suggest a trade-off where a smaller number of words could be taught, increasing efficiency of learning and teaching for sake of near-optimal generalization capacity. This has potentially important implications for reading education where there is a need to teach spelling-sound patterns (phonics) but only enough time to sample from the large set of patterns. Many educators oppose teaching phonics because it is seen as requiring "drill and kill" amounts of instruction and practice. This may be less of a concern if, as our results suggest, patterns can be selected in a way that maximizes generalization.

The problem of maximizing generalization with the smallest possible training set can be formalized as a *machine teaching* optimization problem (Zhu, 2015). We have drawn on this literature by manipulating the learning environment while holding the abilities of the learner constant, and then performing careful analyses of the outcomes to identify the factors that contribute to training the most proficient models. In doing so we have demonstrated systematic relationships between the composition of the training set and generalization performance that machine teachers may be able to discover and exploit.

These results are empirical; our next step will be to identify properties of words and word-sets responsible for better generalization both at the word- and set-level. As indicated in our regression model reported, item-wise measures of phonology, orthography, and especially orth-phon consistency account

for non-trivial amounts of generalization error. Next steps will be oriented towards accounting for more of the variance in generalization accuracy, and to scale up analyses to model-wise characteristics that promote generalization. It may also be possible to improve efficiency even further by using training sets attuned to children's vocabulary development, and by optimizing the sequence of learning experiences. Ultimately the aim is to discover the principle axes of the orth-phon mapping space, and exploit that structure in a theory-driven way to construct idealized training environments.

The reported models were trained on representations of the orthography with 14 "slots" for letters and tested on phonology with 10 "slots" for phonemes. This has consequences for learning that are artificial relative to how a child learns to decode orthography. Most salient is that each slot has an independent set of weights that project to the hidden layer. This means that what is learned about letters in one slot is not necessarily transferred to other slots—once the model has learned to pronounce the consonant *K* in the third slot, it will fail to generalize that knowledge when presented with a *K* in the fifth slot. This and other limitations of the slot based representation scheme contribute to our focus (and the focus of the modeling literature, generally) on monosyllabic words. Monosyllabic words are short and fairly consistent in length with a single vowel phoneme. After vowel-centering, the limits of using slots are effectively attenuated in the monosyllabic context, but it is not a solution that scales up. Models of reading that attempt to reflect more plausible visual processes and accommodate disyllabic words are needed. The slot-based approach may add some complexity to the decoding problem while simplifying the "visual" experience of our models.

Though preliminary, these simulations demonstrate that it is possible to be more efficient with curricula that attend to the number of words taught and the words that are prioritized in teaching.

References

Andrews, S. (1997). The effect of orthographic similarity on lexical retrieval: Resolving neighborhood conflicts. *Psychonomic Bulletin & Review*, *4*(4), 439–461. doi: 10.3758/bf03214334

Berko, J. (1958). The child's learning of english morphology. *Word*, *14*, 150 - 177.

Cameirão, M. L., & Vicente, S. G. (2010). Age-of-acquisition norms for a set of 1,749 portuguese words. *Behavior Research Methods*, *42*(2), 474–480. doi: 10.3758/brm.42.2.474

Chomsky, N. (1965). *Aspects of the theory of syntax.* Cambridge, M.I.T. Press.

Gilkerson, J., Richards, J. A., Warren, S. F., Montgomery, J. K., Greenwood, C. R., Oller, D. K., ... Paul, T. D. (2017). Mapping the early language environment using all-day recordings and automated analysis. *American Journal of Speech-Language Pathology*, *26*, 248–265.

Goldberg, A. E. (2009). The nature of generalization in language. *Cognitive Linguistics*, *20*(1), 93 - 127.

Harm, M. W., & Seidenberg, M. S. (1999). Phonology, reading acquisition, and dyslexia: insights from connectionist models. *Psychological review*, *106*, 491–528.

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young american children.* Baltimore, MD: Paul H. Brookes.

Jared, D., McRae, K., & Seidenberg, M. S. (1990). The basis of consistency effects in word naming. *Journal of Memory and Language*, *29*(6), 687–715. doi: 10.1016/0749-596x(90)90044-z

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. doi: 10.1038/nature14539

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review*, *103*(1), 56 - 115.

Santolin, C., & Saffran, J. R. (2018). Constraints on statistical learning across species. *Trends in Cognitive Sciences*, *22*(1), 52 - 63.

Seidenberg, M. S. (2017). *Language at the speed of sight: How we read, why so many can't, and what can be done about it.* New York : Basic Books.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, *96*(4), 523 - 568.

Seidenberg, M. S., & Plaut, D. C. (2014). Quasiregularity and its discontents: The legacy of the past tense debate. *Cognitive Science*, *38*(6), 1190 - 1228.

Storkel, H. L. (2003). Learning new words II. *Journal of Speech, Language, and Hearing Research*, *46*(6), 1312–1323. doi: 10.1044/1092-4388(2003/102)

Thain, D., Tannenbaum, T., & Livny, M. (2005). Distributed computing in practice: the condor experience. *Concurrency and Computation-Practice and Experience*, *17*(2-4), 323 - 356.

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2016). Understanding deep learning requires rethinking generalization. *arXiv*.

Zhu, X. (2015). Machine teaching: An inverse problem to machine learning and an approach toward optimal education. In *Proceedings of the twenty-ninth AAAI conference on artificial intelligence* (pp. 4083–4087). AAAI Press. Retrieved from `http://dl.acm.org/citation.cfm?id=2888116.2888288`

Ziegler, J. C., Ferrand, L., & Montant, M. (2004). Visual phonology: The effects of orthographic consistency on different auditory word recognition tasks. *Memory & Cognition*, *32*(5), 732–741. doi: 10.3758/bf03195863

Iconic Prosody is Rooted in Sensori-Motor Properties: Fundamental Frequency and the Vertical Space

Aleksandra Ćwiek (cwiek@leibniz-zas.de)
Leibniz-Centre General Linguistics
Berlin, Germany

Susanne Fuchs (fuchs@leibniz-zas.de)
Leibniz-Centre General Linguistics
Berlin, Germany

Abstract

The iconic cross-modal correspondence between fundamental frequency and location in vertical space ("high is up") has long been described in the literature. However, an explanation for this relationship has not been proposed. We conducted an experiment in which participants shot at cans projected on the wall in different vertical positions. We found that mean fundamental frequency was significantly influenced by vertical head position. Moving the head upwards changes the position of the larynx, which pulls on the cricothyroid muscle and changes the fundamental frequency. We thus propose that the iconic relationship between fundamental frequency and vertical space is grounded in the body.

Keywords: iconicity; prosody; fundamental frequency; vertical space; sensori-motor properties; embodied cognition

Introduction

Iconicity refers to the resemblance between the linguistic *form* and the intended *meaning* (Blasi, Wichmann, Hammarstrm, Stadler, & Christiansen, 2016; Dingemanse, Blasi, Lupyan, Christiansen, & Monaghan, 2015). The evidence brought forward in the last decade suggests that iconicity is an essential part of language (Perniss, Thompson, & Vigliocco, 2010; Perniss & Vigliocco, 2014). The origin of iconicity, however, is far from clear (cf. e.g., Imai & Kita, 2014, p. 9). This article brings together earlier work in phonetics and cognitive sciences by discussing the relationship between iconicity and prosody in speech.

The term prosody has often been used interchangeably with intonation, though these terms are used in different ways by different authors (for a discussion, cf. Hirst & Di Cristo, 1998). Here, we use prosody as an umbrella term for the suprasegmental stress, rhythm, and intonation properties in speech (Bussman, 1996; Trask, 2004). We investigate a single basic aspect of prosody, fundamental frequency, which is an acoustic correlate of intonation. Fundamental frequency expresses the rate at which the vocal folds vibrate during speech. It is most frequently measured in hertz. Pitch expresses how fundamental frequency is perceived. It is quantified by listeners' judgments. In this paper we use f0 in reference to production studies and pitch in reference to perception studies.

Previous research has shown that there is a relationship between location in vertical space and pitch (in speech perception) or f0 (in speech production), though little in the way of explanation has been proposed. In the literature, iconicity is mainly considered to be both a finding and the explanation for the finding: there is a form–meaning mapping, because the relationship is iconic. However, no explanation is given for why iconicity is present in natural language. Therefore, the aim of this study is to investigate a potential origin for iconicity. Using an experimental approach, we demonstrate that there is a link between vertical head movement and f0. We propose that the iconic correspondence between location in vertical space and f0 is rooted in head movement required to look at objects that are higher up. In other words, the origin for the relationship between f0 and location in vertical space lies in bodily constraints, namely vertical head movement.

Evidence for Iconic Pitch in Speech Perception

Researchers have long been interested in how people localize sounds in vertical space and what kind of cognitive processes are involved (e.g., Seashore, 1899). Pioneering work by Pratt (1930) and Trimble (1934) and later experiments by Mudd (1963) and Roffler and Butler (1968) show that, regardless of the actual vertical position of the sound source, participants tend to locate high-pitched sounds higher in the vertical plane and low-pitched sounds lower in the vertical plane (i.e., the Pratt effect, cf. above). Recent work by Parise, Knorre, and Ernst (2014) provides an insight into the possible source of this "frequency–elevation mapping." The authors recorded sounds in different natural environments with directional microphones mounted at various heights. They found a strong correlation between the frequencies of the noises in the environment, especially in the 1–6 kHz range, and the sound location in the vertical space. After accounting for the filtering properties of the outer ear, Parise et al. conclude that the ear is fine-tuned to the statistics of the noises in the environment. Thus, the relationship between a given sound's frequency and its vertical position is not language-specific. Apart from that, their results suggests that the listeners' expectations of an object's location in vertical space may be grounded in the statistical probabilities in the natural environment. It has to be noted that the noises in the environment come from both animate, like birds, and inanimate sources, like wind in the trees.

In 1994, Ohala proposed the *frequency code* as a possible cause of prosodic iconicity (Ohala, 1994). He argued that in various species, low-pitched vocalizations are associated with a large-sized animal, since the mass of the vocal folds correlates with body mass and, thus, size. Both body mass

and the size of an animal are crucial to estimate a potential threat, which in the animal kingdom is a matter of life and death. The lower the perceived pitch of a given animal, the more threatening, dominant, or aggressive the animal is assumed to be. And conversely, the higher the perceived pitch emitted by the sound source, the smaller its size is estimated to be, thus the source itself is interpreted as less threatening, dominant, and aggressive.

Keeping in mind the studies mentioned above and Ohala's frequency code, when we consider humans, it is apparent that there are two main, possibly contradictory, factors that affect pitch estimation. The first is vertical position – higher pitch is located higher in vertical space. The second factor is body size – higher pitch is associated with smaller body size. If two people of different size (height) stand beside each other, the larger person's mouth will always be higher in the vertical plane. On the one hand, according to Ohala's frequency code, we would expect the larger person to emit lower-pitched sounds. On the other hand, according to, e.g., Parise et al. (2014), the larger person, because their mouth is higher up in the vertical plane, should sound higher-pitched. This contradictory predictions have, to our knowledge, only been addressed by one study (Pisanski, Isenstein, Montano, O'Connor, & Feinberg, 2017).

Pisanski et al. (2017) explicitly investigated body size estimation based on pitch vs. vertical location. Their results show that low pitch was associated with a large body size even when it was played from a low vertical position. This suggests that pitch cues override spatial cues in the body size estimation. However, this finding may be due to the task, which was to estimate the body size of an animate being. Hence, animacy and the experimental task may also have influence on the perception of different frequencies.

Evidence for Iconic F0 in Speech Production

There is considerably less work on iconic f0 in speech production. Although the effects found in the studies mentioned below are mostly subtle, nevertheless they provide evidence for the use of iconic pitch with regard to location in vertical space. Items that are located higher up in vertical space (whether they are actual objects or mental concepts located in a metaphorical plane) are marked with higher fundamental frequency than item that are located lower in space.

In a series of experiments, Clark, Perlman, and Johansson Falck (2014) asked the participants to read stories related to vertical motion (up vs. down), emotions (positive vs. negative), and perceived sound (high-pitched vs. low-pitched). The authors expected that the participants would produce higher fundamental frequency in stories with higher elevation in the physical space, positive emotions, and high auditory pitch in contrast to stories reporting lower vertical space, negative emotions, and lower auditory pitch. A significant effect was found only for stories describing a vertical motion in which the f0 was on average 5 Hz higher in the "up" condition than in the "down" condition.

Nygaard, Herold, and Namy (2009) investigated prosody of adjective antonym pairs (e.g., *happy/sad*, *big/small*, *tall/short*). The adjectives were embedded in carrier sentences next to novel non-words and the participants were asked to use infant-directed speech. Their analyses suggest that the fundamental frequency was higher in the adjectives *happy*, *big*, *hot*, *tall*, *yummy* and *strong*, compared to their antonym counterparts. Depending on the item, f0 differences were between 20 and 90 Hz. However they might have been affected by how engaged the participant was in the infant-directed speech task.

Another study that tackles the problem of iconic f0 with regard to vertical space is that of Shintel, Nusbaum, and Okrent (2006). The authors analyzed the fundamental frequency of participants saying if an animated dot was moving up or down. The data revealed a significantly higher f0 for the "up" condition. The f0 differences were, however, relatively small, similarly to those reported by Clark et al. (2014).

The reviewed previous literature, both in perception and production, documents the relationship between f0 and location in vertical space that is proposed to be iconic. But the only explanation given for this relationship is iconicity itself. The purpose of this study is to test one potential origin of iconicity and thus also the origin of the relationship between f0 and location in vertical space.

Potential Anatomical Explanations for the Relationship between F0 and Vertical Space

The control of fundamental frequency is anatomically complex. It involves a difference in the subglottal pressure between the lungs and the oral cavity, in addition to the tensing of the vocalis muscles (the vocal folds). Moreover, different extrinsic muscles can indirectly influence the vocal folds. For example, activation of the cricothyroid muscle (CT) leads to the rotation of the crycoid cartilage, which in turn tenses the vocal folds, and thus increases f0 (Honda, 1996). Apart from that, fundamental frequency can be lowered by the actions of the external strap muscles (Erickson, Baer, & Harris, 1982).

But additional factors may come into play in f0 control by causing a change within the other parameters affecting f0, e.g., by varying the muscle tension around the larynx. Anatomically, muscle tension around the larynx can be changed by head movement. To the best of our knowledge, only one empirical study exists showing the influence of head motion on f0 in speech production (Munhall, Jones, Callan, Kuratate, & Vatikiotis-Bateson, 2004). Their sentence-by-sentence multiple regression analysis revealed that 63% of the variation in the fundamental frequency could be explained by speaker's head movement during speech production. The upward head motion raises the larynx, thereby pulling on the CT muscle, which elongates the vocal folds and thereby increases f0. Still, it has to be noted that the analyses were carried out on recordings of only one speaker.

The aim of this study is to investigate the potential anatomical origin of the iconic relationship between f0 and location in vertical space. Our first hypothesis is that the fundamental frequency is affected by the position of the head. Changes in

the position of the head influence the placement of the larynx, and simultaneously affect the fundamental frequency. We assume that people look at an object, the head follows the gaze. When an object is located higher up in space, people move their head upwards and when it is located lower, they move their head down. Thus, if an object is located at the higher position in space, we expect a higher f0 within the utterance produced at this head position. Our second hypothesis is that the size of an object has an impact on the fundamental frequency when referring to that object. We expect a higher f0 for smaller objects and lower f0 for larger objects.

Methods

Experimental Design

In the experimental task, participants were asked to "shoot" cans, which were projected onto the wall in front of them, with a laser pointer. Additionally, the participants had to say the word written on the can. One of two words was written on each can: *piff* [pɪf] or *paff* [paf], both of which are German onomatopoeic words imitating the sound of shooting, like English 'bang' or 'pow'. To measure the hypothesized effect of size of the object referred to, we used two sizes of cans in the experiment – a small and a large one, which was approximately twice as large. The cans appeared on five equidistant positions on the vertical and horizontal axes, resulting in 25 possible positions. The varying vertical position of the can enabled us to elicit the head movement, expected to have an influence on fundamental frequency – according to the first hypothesis. All conditions sum up to a total of 100 tokens per participant: two words x two can sizes x five vertical positions x five horizontal positions.

During the task, the participants stood at a landmark on the ground, at approximately 1 m distance from the wall. The projection surface measured 130 x 130 cm, with the lowest edge starting at 145 cm above the ground. When a can appeared, the participants were instructed to (1) point at the can with a laser pointer, and (2) say the word written on the can. After the participant successfully pointed at the can and uttered the word written on it, an animation of the can falling down was played and, after a short blank screen, a new can in a different position appeared. The presentation of cans and their animation was manually controlled to prevent the participants from predicting when the next can would appear.

Five datasets with prerandomized order of presentation were created to avoid order effects. For technical reasons, the presentation of the stimuli was divided in two sections, both of which consisted of 50 items. The whole experiment took 15–20 minutes, though the experimental task itself lasted no longer than 5–6 minutes. This was highly relevant in order to avoid boredom and its potential effect on fundamental frequency. The main task was preceded by a short familiarization trial, which consisted of five items – cans with different words written on them, than those that were used in the main task. The participants were given no specific instructions regarding the alignment of their movements. If one of them asked about it, they were told to act naturally, in a way similar to pointing a laser pointer at a particular word or image while giving an oral presentation.

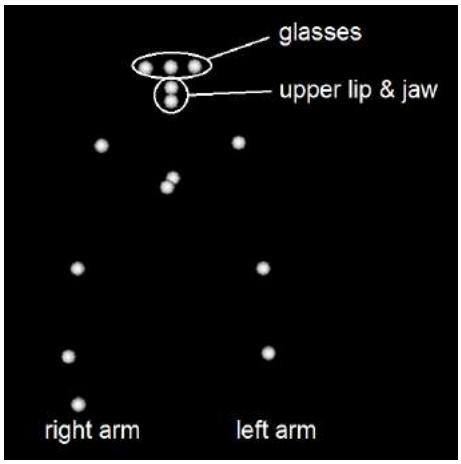


Figure 1: The placement of markers for the motion capture.

Acoustic and motion data were recorded simultaneously using a Sennheiser ME 64 cardioid microphone and an Optitrack motion capture system (Motive, version 1.9.0) with 12 cameras (Prime 13). Motion data was captured at 120 Hz sampling frequency and acoustic data at 44.1 kHz. In total, 15 markers were placed on different body parts of the participants: three on the glasses (center, left, and right); one each on the upper lip and the lower lip (jaw); one at the position of the sternum; one approximately at the location of the fourth thoracic spine vertebra; one on the laser pointer; and symmetrically two on the shoulders, elbows, and wrists. The placement of all markers is illustrated in Figure 1.

Due to technical problems with the recording equipment, the data of one participant had to be excluded.

Participants

Since males have an on average lower fundamental frequency than females, including different sexes would have yielded an additional factor in the experimental design. To avoid this, we focused on females as a participant group. A total of 31 German native speakers took part in the study (mean age = 27.84; mean height = 167.7 cm, with min = 152 cm and max = 183 cm). Twenty-six participants were monolingual and five reported being bilingual; all apart from one were right-handed. The participants were all recruited using a participant database. At the beginning of the session, they were given information about the experiment and signed a consent form. They received monetary compensation after the completion of the task. The project was approved by the ethical board of the DFGS and preregistered at Open Science Framework[1].

[1]The OSF repository can be visited under the following address: https://osf.io/ysr75/

Data Preprocessing and Annotation

The acoustic data were automatically labeled at the phoneme level using WebMAUS (Kisler, Reichel, & Schiel, 2017) and subsequently manually corrected using Praat (Boersma & Weenink, 2018). The on- and offset of the vowel were defined as the onset and offset of vocal fold oscillations, respectively. So far, we have annotated and corrected the data for 20 participants (mean age = 29.45; mean height = 165.5 cm, with min = 152 cm and max = 177 cm). Mean fundamental frequency was calculated for the whole vowel interval. The fundamental frequency range was set to 150–400 Hz to avoid octave jumps of the pitch tracker due to creaky voice.

The motion capture data were first extracted and processed with Mokka version 0.6.2 (Barré & Armand, 2014), and subsequently converted to be further processed with MATLAB (version R2017b). The maximal vertical position of the pointing wrist and of the center of the glasses were calculated within the vowel interval, provided by the annotated acoustic data.

Statistical Analyses

All statistical analyses were carried out within the R environment, version 3.5.1 (R Core Team, 2018), using the following packages: `plyr` (Wickham, 2011) for data wrangling, `car` (Fox & Weisberg, 2011) and `lme4` (Bates, Mächler, Bolker, & Walker, 2015) for statistical modelling, `RePsychLing` (Baayen, Bates, Kliegl, & Vasishth, 2015) for model evaluation, and `stargazer` (Hlavac, 2015) for the output table.

After the initial data exploration, we were forced to exclude the data of one participant (id7) from subsequent analysis. Her behavior during the experiment was atypical; she shrieked and laughed a lot during the experiment. This later had a negative effect on the reliable parameter extraction. Thus, all results refer to the group of 19 female participants.

Results

General Remarks

Though we did not explicitly investigate the coordination between the head movement (gaze), articulation (lip and jaw movement), and the pointing gesture, their temporal organization is shown in Figure 2 for reference. The speaker first visually locates the target and elevates the head (in the example shown in Figure 2, the target is situated higher in the vertical space). Then she starts the pointing gesture by visibly lifting the wrist. Finally, the speech itself begins. It can be observed in the acoustic signal itself, but also in the larger maximal distance of the upper lip and jaw.

Statistical Hypotheses Testing

The parsimonious mixed model approach was used to more reliably analyze data that are highly variable between subjects (Bates, Kliegl, Vasishth, & Baayen, 2015, p. 2). This approach starts by calculating a maximal model, as suggested by Barr (2013). Subsequently, a principal component analysis of the random effects' structure is run using the `RePsychLing`

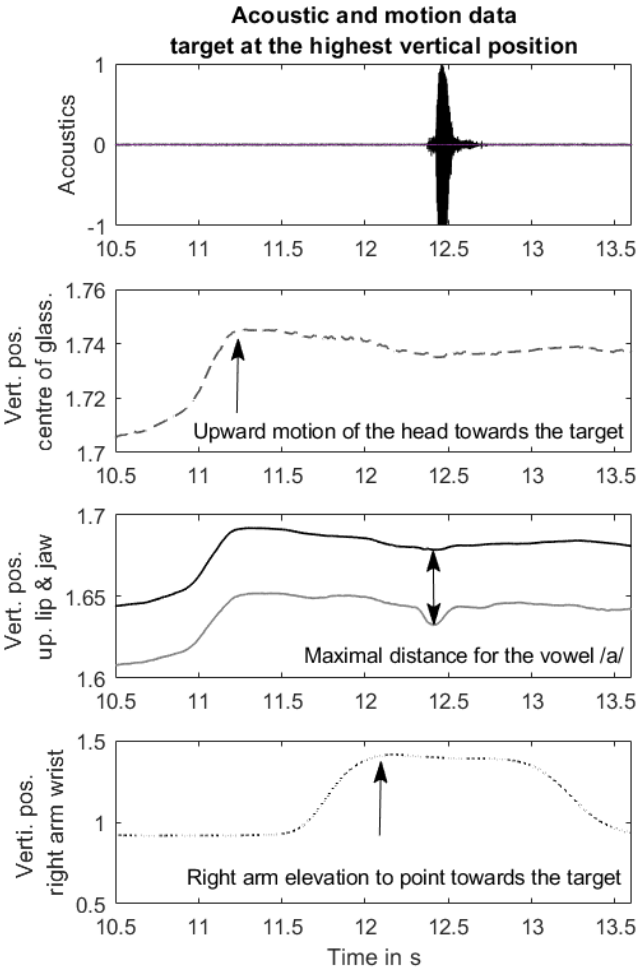


Figure 2: An example of acoustic and motion data for a trial with the stimulus located high in the vertical plane. Plots from top to bottom: (1) acoustic signal; (2) vertical head position (dashed line), determined by the marker in the middle of the glasses; (3) upper lip (black line) and jaw (gray line) markers; (4) marker on the pointing arm wrist (dashed line). All motion data are in meters. Note that the arrows in charts 2–4 only illustrate the coordination among different body parts, but all positions are calculated within the vowel interval.

package (Bates, Kliegl, et al., 2015). This procedure allows to reduce the random effects structure to the necessary components, according to the variance in the data. It is done by an iterative reduction of the model's structure and step-by-step comparison of subsequent models. In the current analysis, the parsimonious best-fit model was in fact not significantly different from the maximal model ($p = 0.54$).

All computed models consisted of the same set of fixed factors: vowel in the uttered word, subject's height, subject's head position, can's vertical position, can size, and the interaction of vowel and can size. Random intercept for subject was included in the random effects structure to account for intersubject variability. The random slopes after the iterative model reduction consisted of vowel in the uttered word, and the interaction of vowel and can size.

The results of the the linear mixed effects model presented in Table 1 reveal four main effects on the mean fundamental frequency, namely that of: the vowel segment, participant's body height, participant's head position, and the size of the can. The first two effects have been described in the past. No effect on the mean f0 has been found for either the can's vertical position or the interaction of vowel segment and can size.

The effect of the vowel on the mean f0 is in line with previous reports on the intrinsic f0 in vowels (Whalen & Levitt, 1995; Whalen, Gick, Kumada, & Honda, 1999). This is the most robust effect found in our data and it shows that higher mean f0 values were found for the high vowel /ɪ/ and lower values for the low vowel /a/. Furthermore, the data show that the taller the participant is, the lower her fundamental frequency is. This negative effect found for participant's height corroborates with the frequency code (Ohala, 1994) and the work by Pisanski and Rendall (2011).

Most importantly, both hypotheses put forward earlier gain support from our analysis. The model shows that the mean f0 increases with the elevation of the head, which is consistent with the first hypothesis. In addition, the size of the can had a significant impact on the mean f0 – participants produced lower mean f0 when referring to larger cans. This result is in line with the second hypothesis. However, it has to be pointed out that the mean f0 differences found between smaller and larger cans are rather small at only 2–3 Hz. Figure 3 illustrates that the effect found for size is the strongest in the highest can positions (1 and 2) and it diminishes or disappears completely at the lower positions (3–5).

Discussion

The analysis presented above demonstrates that in our data the mean fundamental frequency is influenced by the vertical head position rather than the vertical position of the object referred to. In the computed best-fit model with mean fundamental frequency as a dependent variable, we found that the tested anatomical factor – head position – plays a crucial role for the mean f0. In contrast, a factor depicting a purely iconic relationship in the location on the vertical plane – can's ver-

Table 1: The results of the linear mixed model analysis. The table shows the effect of fixed factors, listed on the left, on the dependent variable: mean f0. The estimated effect size is given for each factor, with the standard error given in brackets in the line below.

	Dependent variable:
	Mean f0
Intercept	225.455***
	(5.358)
Vowel	15.978***
	(2.347)
Participant's height	−13.481**
	(5.986)
Head position	9.371***
	(3.041)
Can's vertical position	−0.441
	(0.572)
Can size	2.225**
	(1.005)
Vowel * can size	1.326
	(1.894)
Observations	1,872
Log Likelihood	−7,841.494
Akaike Inf. Crit.	15,704.990
Bayesian Inf. Crit.	15,765.870
Note:	*p<0.1; **p<0.05; ***p<0.01

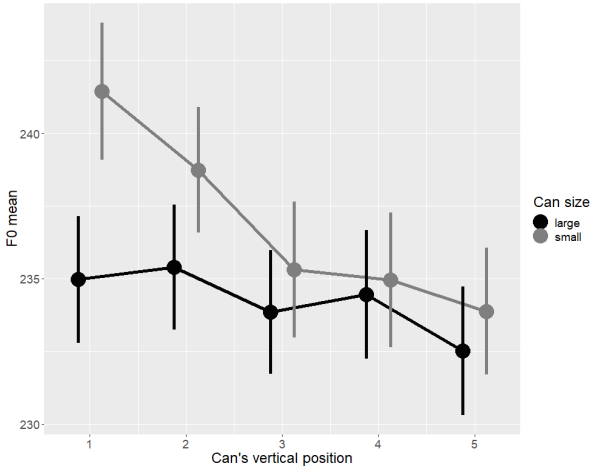


Figure 3: Mean f0 values for the can in different sizes and vertical positions. On the horizontal axis, 1 means high and 5 means low position. The darker line depicts a large-sized can and the lighter one a small-sized can.

tical position – did not significantly account for the variance in the mean f0. Thus, we propose that the origin for the relationship between f0 and location in vertical space in speech production is rooted in the body and its sensori-motor properties. The cross-modal relationship between f0 and vertical space, as discussed in the framework of iconicity, is not simply a result of an iconic form and meaning mapping ("high is up"). It is a result of embodiment, driven by the sensorimotor properties, which are influenced by the changes in the head position.

Furthermore, we found that the size of an object had a significant impact on mean f0 values in the current experiment (cf. Figure 3). In speech production, a correspondence between fundamental frequency and size of an object has been previously found in a story reading task by Perlman, Clark, and Falck (2015), apart from other studies mentioned in the introduction. In this study, participants were asked to read stories with concepts of *fast/slow* and *big/small*. Reading pace and f0 were measured and it was found that (1) the stories with "fast" concepts were read faster than those with "slow" ones, and (2) the stories with "big" concepts were read with lower pitch than those with "small" ones. Our study supplements previous findings on f0–size symbolism by showing that participants reliably signal a difference in size of an inanimate object by adapting their fundamental frequency. There is a large body of work on the iconic f0–size relationship on the segmental level (cf. e.g., Shinohara & Kawahara, 2010; Tsur, 2006; Ultan, 1978). It has been shown in various cross-linguistic analyses that high vowels are more frequently used in words depicting smaller objects, and low vowels in words depicting larger objects. Therefore it was highly relevant to control for the interaction of vowel and can size in the current analysis, though no significant effect was found.

We would like to point out that even though vertical head movement does affect the f0 in our data, it is not the only thing that affects f0 in speech production. Speakers have a high degree of control over f0 and it is often employed to signal prominence in speech (Terken, 1991), such as word stress and sentence accent. Speakers can manipulate their pitch according to the needs of the communicative situation. We found that speakers varied greatly in how they completed the task. Some participants barely moved their head, while others did, even if it did not seem necessary. Even though a small number of participants showed very little head movement, we still found that the head position was one of the strongest predictors of mean f0 variance in our data.

The degree of the vertical head movement varied not only between the participants, but also between the vertical positions of the cans. It can be seen in Figure 3 – lower vertical positions (3–5) yield smaller or no differences in mean f0 between one another. This can be a side effect of a sufficient body size to visually process lower targets. A thorough analysis of speaker behavior is yet to be conducted.

The present study proposes a potential origin for the iconic relationship between f0 and object location in vertical space. Our data show that head movement influences f0: upward head movement leads to higher f0, which is most likely a result of the larynx pulling on the cricothyroid muscle. We thus propose that the iconic relationship between f0 and vertical space is rooted in the body: when looking at an object located higher on the vertical plane, head movement generally reflects the location of the object. In this case, the head movement itself could be considered iconic, because the form is aligned with the meaning – the head moves upwards toward an upward target. Previous literature has already established an iconic relationship between f0 and object location. Our study supplements previous interpretations of this iconic relationship with evidence that the correspondence stems from bodily constraints. The upward movement of the head causes physiological changes that influence f0. In this way, we argue that the correspondence between f0 and vertical space is both embodied – because it is rooted in the body – and iconic – because the form and the meaning correspond. "High is up", because the head moves upwards and the effect is thus an embodied form–meaning correspondence.

Acknowledgments

This work was funded by the German Research Council as a part of the XPrag.de project PSIMS: The Pragmatic Status of Iconic Meaning in Spoken Communication (FU 791/6-1).

References

Baayen, H., Bates, D., Kliegl, R., & Vasishth, S. (2015). Repsychling: Data sets from Psychology and Linguistics experiments [Computer software manual]. (R package version 0.0.4)

Barr, D. J. (2013). Random effects structure for testing interactions in linear mixed-effects models. *Frontiers in Psychology*, *4*, 328.

Barré, A., & Armand, S. (2014). Biomechanical ToolKit: Open-source framework to visualize and process biomechanical data. *Computer Methods and Programs in Biomedicine*, *114*, 80-87.

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *arXiv preprint arXiv:1506.04967*.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Blasi, D. E., Wichmann, S., Hammarstrm, H., Stadler, P. F., & Christiansen, M. H. (2016, September). Sound-meaning association biases evidenced across thousands of languages. *Proceedings of the National Academy of Sciences*, *113*(39), 10818–10823.

Boersma, P., & Weenink, D. (2018). Praat: doing phonetics by computer [Computer program]. Version 6.0.43, retrieved 2018-03-20. `http://www. praat. org/`.

Bussman, H. (1996). *Routledge Dictionary of Language and Linguistics*. New York: Routledge. (Translated and edited by Gregory P. Trauth and Kerstin Kazzasi)

Clark, N., Perlman, M., & Johansson Falck, M. (2014). Iconic pitch expresses vertical space. In M. Borkent, B. Dancygier, & J. Hinnell (Eds.), (pp. 393–410). Stanford: SCLI Publications.

Dingemanse, M., Blasi, D. E., Lupyan, G., Christiansen, M. H., & Monaghan, P. (2015, October). Arbitrariness, Iconicity, and Systematicity in Language. *Trends in Cognitive Sciences*, *19*(10), 603–615.

Erickson, D., Baer, T., & Harris, K. S. (1982). The role of the strap muscles in pitch lowering. *Haskins Laboratories: Status Report on Speech Research*, *70*, 275–284.

Fox, J., & Weisberg, S. (2011). *An R Companion to Applied Regression* (Second ed.). Thousand Oaks CA: Sage.

Hirst, D., & Di Cristo, A. (1998). *Intonation systems: a survey of twenty languages*. Cambridge University Press.

Hlavac, M. (2015). stargazer: Well-formatted regression and summary statistics tables [Computer software manual]. Cambridge, USA. (R package version 5.2)

Honda, K. (1996). Biological Mechanisms for Tuning Voice Fundamental Frequency. *Koutou (THE LARYNX JAPAN)*, *8*(2), 109–115.

Imai, M., & Kita, S. (2014, August). The sound symbolism bootstrapping hypothesis for language acquisition and language evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1651), 20130298.

Kisler, T., Reichel, U., & Schiel, F. (2017). Multilingual processing of speech via web services. *Computer Speech & Language*, *45*, 326 - 347.

Mudd, S. A. (1963). Spatial stereotypes of four dimensions of pure tone. *Journal of Experimental Psychology*, *66*(4), 347–352.

Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., & Vatikiotis-Bateson, E. (2004). Visual prosody and speech intelligibility: Head movement improves auditory speech perception. *Psychological Science*, *15*(2), 133–137.

Nygaard, L. C., Herold, D. S., & Namy, L. L. (2009, January). The Semantics of Prosody: Acoustic and Perceptual Evidence of Prosodic Correlates to Word Meaning. *Cognitive Science*, *33*(1), 127–146.

Ohala, J. J. (1994). The frequency code underlies the sound-symbolic use of voice pitch. In L. Hinton, J. Nichols, & J. J. Ohala (Eds.), *Sound symbolism* (pp. 325–347). Cambridge University Press.

Parise, C. V., Knorre, K., & Ernst, M. O. (2014, April). Natural auditory scene statistics shapes human spatial hearing. *Proceedings of the National Academy of Sciences*, *111*(16), 6104–6108.

Perlman, M., Clark, N., & Falck, M. J. (2015, October). Iconic Prosody in Story Reading. *Cognitive Science*, *39*(6), 1348–1368. doi: 10.1111/cogs.12190

Perniss, P., Thompson, R. L., & Vigliocco, G. (2010). Iconicity as a General Property of Language: Evidence from Spoken and Signed Languages. *Frontiers in Psychology*, *1*.

Perniss, P., & Vigliocco, G. (2014, August). The bridge of iconicity: from a world of experience to the experience of language. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1651), 1–14.

Pisanski, K., Isenstein, S. G. E., Montano, K. J., O'Connor, J. J. M., & Feinberg, D. R. (2017, feb). Low is large: spatial location and pitch interact in voice-based body size estimation. *Attention, Perception, & Psychophysics*, *79*(4), 1239–1251.

Pisanski, K., & Rendall, D. (2011). The prioritization of voice fundamental frequency or formants in listeners assessments of speaker size, masculinity, and attractiveness. *The Journal of the Acoustical Society of America*, *129*(4), 2201-2212.

Pratt, C. C. (1930). The spatial character of high and low tones. *Journal of Experimental Psychology*, *13*(3), 278–285.

R Core Team. (2018). R: A Language and Environment for Statistical Computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org/`

Roffler, S. K., & Butler, R. A. (1968, June). Localization of Tonal Stimuli in the Vertical Plane. *The Journal of the Acoustical Society of America*, *43*(6), 1260–1266.

Seashore, C. (1899). Localization of sound in the median plane. *Univ. Iowa Stud. Psychol*, *2*, 46–54.

Shinohara, K., & Kawahara, S. (2010, August). A Cross-linguistic Study of Sound Symbolism: The Images of Size. *Annual Meeting of the Berkeley Linguistics Society*, *36*(1), 396–410.

Shintel, H., Nusbaum, H. C., & Okrent, A. (2006). Analog acoustic expression in speech communication. *Journal of Memory and Language*, *55*, 167-177.

Terken, J. (1991). Fundamental frequency and perceived prominence of accented syllables. *The Journal of the Acoustical Society of America*, *89*(4), 1768–1776.

Trask, R. L. (2004). *A Dictionary of Phonetics and Phonology*. New York: Routledge.

Trimble, O. C. (1934, January). Localization of Sound in the Anterior, Posterior and Vertical Dimensions of Auditory Space. *British Journal of Psychology*, *24*(3), 320–334.

Tsur, R. (2006, June). Sizesound symbolism revisited. *Journal of Pragmatics*, *38*(6), 905–924.

Ultan, R. (1978). Size-sound symbolism. In J. H. Greenberg (Ed.), *Universals of human language* (Vol. 2, pp. 525–568). Stanford University Press Stanford, CA.

Whalen, D. H., Gick, B., Kumada, M., & Honda, K. (1999, April). Cricothyroid activity in high and low vowels: exploring the automaticity of intrinsic F0. *Journal of Phonetics*, *27*(2), 125–142. doi: 10.1006/jpho.1999.0091

Whalen, D. H., & Levitt, A. G. (1995). The universality of intrinsic f0 of vowels. *Journal of Phonetics*, *23*(3), 349–366.

Wickham, H. (2011). The Split-Apply-Combine Strategy for Data Analysis. *Journal of Statistical Software*, *40*(1), 1–29. Retrieved from `http://www.jstatsoft.org/v40/i01/`

Sample-based Variant of Expected Utility Explains Effects of Time Pressure and Individual Differences in Processing Speed on Risk Preferences

Kevin da Silva Castanheira[1], **Ardavan S. Nobandegani**[1,2], **& A. Ross Otto**[1]
{kevin.dasilvacastanheira, ardavan.salehinobandegani}@mail.mcgill.ca
ross.otto@mcgill.ca
[1]Department of Psychology, McGill University
[2]Department of Electrical & Computer Engineering, McGill University

Abstract

While previous models of economic decision-making offer descriptive accounts of behavior, they often overlook the computational complexity of estimating expected utility. Here, we seek to understand how both environmental and individual constraints on cognition shape our daily decision. Informed by the predictions of a recently-proposed resource-rational process model of risky choice, *sample-based expected utility* (SbEU; Nobandegani, da Silva Castanheira, Otto, & Shultz, 2018), we reveal that both time pressure and individual differences in processing speed have a convergent effect on risk preferences during a risky decision-making task. Under severe time constraints, participants' risk preferences manifested a strong framing effect compared to little time pressure in which choice adhered to the classic fourfold pattern of risk preferences. Similarly, individual differences in processing speed, measured using an established task, predicted similar effects upon risk attitudes as extrinsic time pressure. These findings reveal a converging contribution of environmental and individual limitations on risky choice, and provide empirical support for SbEU as a resource-rational process model of risky decision making. Notably, SbEU serves as a single-process model of two well-established biases, and the transition between the two, in risky choice.

Keywords: Behavioral economics; Risky decision-making; Time pressure; Processing speed; resource-rational process models

1 Introduction

Our capacity to adapt our decision-making strategies—financial or otherwise—to environmental demands such as time pressure is an invaluable asset for successful behavior. From an online sale which expires in a few minutes, to the rapid trading of stocks in volatile financial markets, our decisions are inevitably constrained by time pressure. Furthermore, internal limitations in processing speed—that is, the speed with which an individual can perform any cognitive operation—should interact with these environmentally imposed limitations (Gigerenzer & Selten, 2002; Salthouse, 1985) as making a choice is widely thought to require a computation of the relative values of the options under consideration (Kahneman and Tversky, 1979). In light of these constraints, one might wonder if our apparent failures to abide by rational decision-making frameworks (e.g., expected utility theory) could reflect a strategic use of limited cognitive resources. To this end, a number of recent theories have proposed that human cognition, with all its apparent biases, can in fact be understood as optimal response—subject to computational and cognitive limitations (*rational minimalist program*, Nobandegani, 2018; Griffiths, Lieder, & Goodman, 2015; Icard, 2014).

Thus, it is of interest to better understand both how we have adapted our decision-making processes to meet these demands and to what extent our ostensibly irrational choices are shaped by these limitations. While previous work has investigated the effects of environmental constraints like time pressure on irrational choice (Guo, Trueblood, & Diederich, 2017), here we seek to corroborate the contributions of both environmental and individual limitations on risky decision-making.

Perhaps one of the most studied departures from rational theories of decision-making is the violation of *description invariance*, which posits that preferences should remain consistent across choices, regardless of the context in which available options are presented. For example, according to expected utility theory (von Neumann & Morgenstern, 2007), whether a decision is made to avoid a loss or seek gains, it should not change one's choice. However, this assumptions is challenged by a wealth of data supporting the framing effect: people tend to be risk seeking for losses and risk averse for gains (Tversky & Kahneman, 1981). Sensitivity to choice framing has been documented in a variety of real-world circumstances including consumer (e.g., Levin & Gaeth, 1988; Loke & Lau, 1992), and medical decisions (e.g., McNeill, Pauker, Sox, & Tversky, 1982; Moxey, O'Connell, McGettigan, & Henry, 2003). This classic pattern of choice—risk-seeking in the domain of losses and risk-aversion in the domain of gains—is perhaps most famously explained by the S-shaped utility function posited by prospect theory (Kahneman & Tversky, 1979; Tversky & Kahneman, 1992)—a well-known descriptive model of choice behavior.

Prospect theory also explains another choice phenomenon: a decision-maker's risk preference depends not only on the framing of the problem (gains vs. losses), but also the probability of the outcome (small vs. large) associated with the risky option. For example, people buy lottery tickets for which winning in unlikely (low probability gain) but prefer to pay to insure their houses against unlikely disasters (low probability loss). On the other hand, when faced with highly probable outcomes, people prefer to select a sure gain over a probabilistic one—"something is better than nothing"—but prefer to risk it all when faced with two unfavorable options—"Ive got nothing to lose" (Di Mauro & Maffioletti, 2004; Fehr-Duda et al., 2010; Kahneman & Tversky, 1979; Markowitz, 1952; Scholten & Read, 2014; Tversky & Kahneman, 1992). According to prospect theory, the fourfold

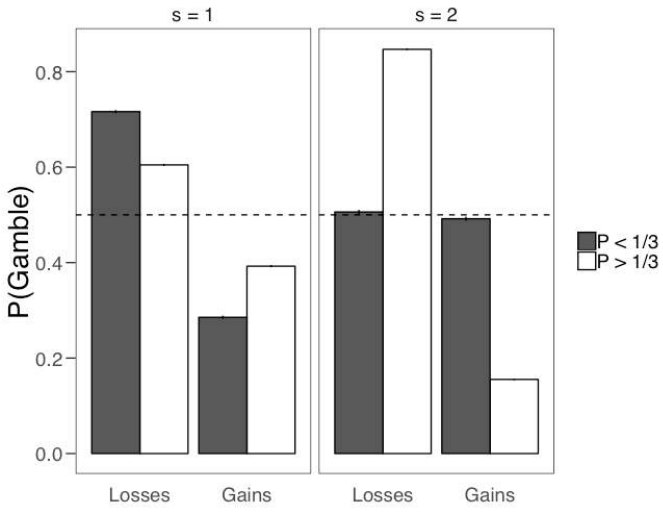


Figure 1: Sample-based expected utility (SbEU; Nobandegani et al., 2018) model predictions for the differential effect of the number of samples on choice. With limited samples (Left) model predicts a framing effect whereas, with more samples, the model predicts more of a fourfold pattern.

pattern of choice arises from the interplay between the S-shaped utility function and the subjective over-weighting of small probabilities (below 1/3) and underweighting of large probabilities (above 1/3) (Tversky & Kahneman, 1992).

While prospect theory offers a descriptive account for the framing effects, it fails to explain either how the decision-making process evolves over time, or how time constraints might bear upon the decision-making process. In order to answer questions about the role of time in these risky choices one must turn to dynamic models of choice. Sequential sampling models are a class of models which assume that choice preferences are estimated by the simulation of an action's potential consequences and where samples are simulated outcomes (Shadlen & Shohamy, 2016). In such models, each simulation takes a non-negligible amount of time and cannot be run in parallel, making time a valuable resource for the decision-maker (Lieder, Griffiths, & Hsu, 2018; Nobandegani, da Silva Castanheira, Otto, & Shultz, 2018). Thus, both total available time and the speed at which these simulations (i.e., samples) are run are directly proportional to the total number of potential outcomes considered (i.e., samples).

If sampling is costly in terms of elementary mental processes, then the number of effective 'samples' an individual is able to draw in a fixed amount of time should also vary in accordance with individual differences in the speed at which an individual processes information—a well-documented capacity limitation termed "processing speed"—which varies considerably across individuals (Kail & Salthouse, 1994). Accordingly, we leverage time pressure manipulations and these individual differences in processing speed to investigate the effect of limiting the number of samples on risky decision-making. Using these two manipulations, will test the effect of varying the number of samples on risky decision-making. Our hypotheses on the directionality of the effect of the time pressure are chiefly informed by a recently-proposed resource-rational process-level model of risky decision-making, *sample-based expected utility* (SbEU; Nobandegani et al., 2018). Extending an earlier model by Lieder et al. (2018), SbEU posits that an agent rationally adapts their strategies depending on the amount of time available for deciding.

Recently, Lieder et al. (2018) proposed a rational process model of risky choice. This model estimates the difference in expected utility of two prospect by using importance sampling, whereby outcomes are sampled in proportion to both its objective probability and its utility (e.g., important outcomes are overrepresented). Lieder et al.'s model, however, was developed under restrictive technical assumptions, making it only optimal when a large number of samples can be drawn. Fortunately, recent developments have determined an optimal sampling distribution which holds for both small and large number of samples (Nobandegani et al., 2018). This is of particular importance as mounting empirical evidence suggests that decision-makers draw very few samples (e.g., Vul, Goodman, Griffiths, & Tenenbaum, 2014); thus, providing an opportunity to explore the effect of limiting cognitive resources (i.e., available samples) on risk preferences.

Accordingly, we used SbEU to generate predictions of people's behavior for a mixture of gambles (i.e., both gains and losses and large and small outcome probabilities) under both conditions of time pressure—in which they can draw very few samples ($s = 1$)—and less constrained conditions—in which they can draw more samples ($s = 2$). Both prospects and time conditions modeled are conceptually identical to those experienced by participants during the task. As depicted in Fig. 1, drawing more samples to estimate the expected utility results in moving from a 'pure' framing effect (Fig. 1a) to the classic fourfold pattern of risk preferences (Fig. 1b). This prediction is in line with the empirical work which suggests that time pressure reduces the amount of information one can process (Miller, 1960; Zur & Breznitz, 1981), as the fourfold pattern requires integrating both outcome and outcome probability information (Kahneman & Tversky, 1979, 1979). Thus, informed by the SbEU's predictions, we sought out to test whether the effects of time pressure on economic choice would conform to the hypothesized pattern. Furthermore, as these predictions are not specific to external time pressure, but any internal constraint on the amount of information that can be processed per unit time, we simultaneously test if differences in cognitive capacity (i.e., processing speed) can also predict a similar pattern of results.

Method

Participants

Data were collected online using Amazon's Mechanical Turk; 100 (41 Female) US-based adult volunteers (mean age =

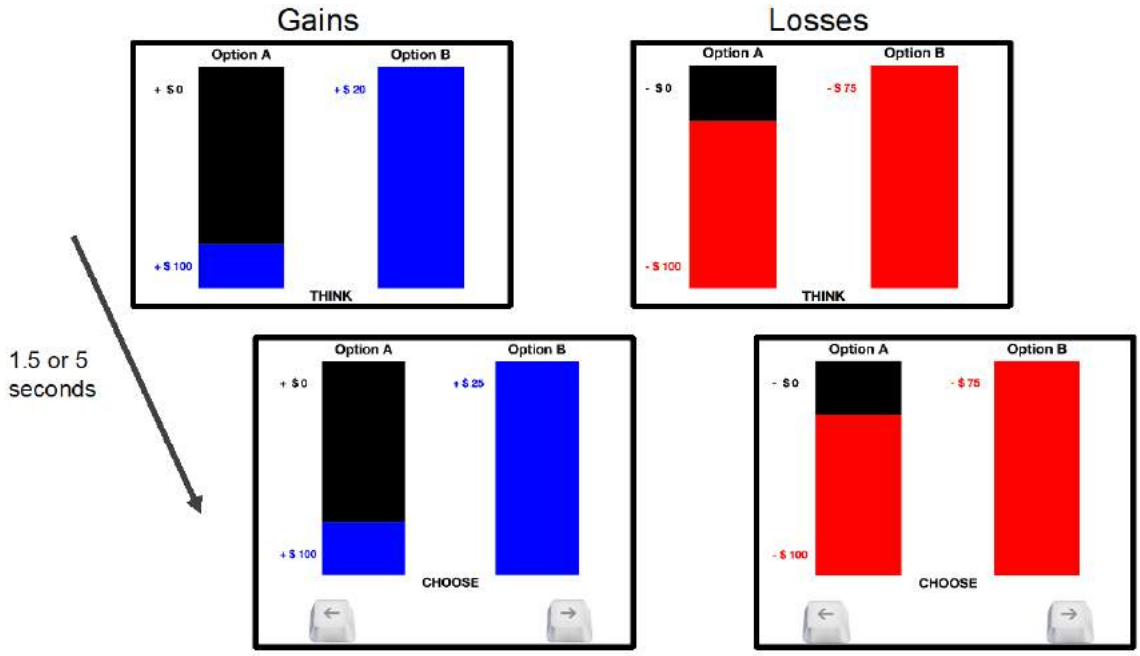


Figure 2: Screenshots of the Gambling task. Participants were given instructed to think about the gamble presented to them before being prompted to respond. The time allotted to think about the problem varied between time pressure conditions: under severe time pressure participants were given 1.5 second to think whereas under light time pressure participants had 5 seconds. Gambles were represented as bar charts where the probability of an outcome was depicted as proportional to the size of the colored portion. Color was used to emphasize the frame of the problem: red represented losses and blue represented gains.

34.77, SD = 9.88), recruited via Amazon's Mechanical Turk (Crump, McDonnell, & Gureckis, 2013), participated in the experiment for a base remuneration of $3.00 USD and a cash bonus—computed in proportion to the outcomes of all trials, with a mean overall payment of $5.85 USD. This study was approved by the McGill University Research Ethics Board.

Processing Speed Measurement

Individual differences in processing speed were assessed using a computerized Digit-Symbol Coding task (Mathias et al., 2017; Salthouse, 1985) which we adapted for use online. Participants were asked to indicate whether or not the digit-symbol pair presented in the center of the screen matched according to the key of associations presented to them. In order to asses processing speed, participants were given 90 seconds to respond to as many trials as correctly and quickly as possible. To ensure participants were taking the task seriously and to minimize exclusions due to random responding, participants were asked to complete the task a second time if their accuracy was below 70%. We subsequently only analyzed the data from a participant's final attempt at the task.

Risky Decision-Making Task

Participants were presented with 120 pairs of binary choices, 60 of which were presented during the light-time pressure (LTP) block and the remaining 60 were presented in the severe time pressure block (STP). Time pressure was manipulated by allowing the participants either 1.5 seconds (STP blocks) or 5 seconds (LTP blocks) to think about their choice. After this lock-out period, participants had a 1 second window to respond in both time pressure conditions; this response window was implemented to minimize the variability in response times and isolate the effects of processing speed on decision-making. Participants were prompted to think about their choice before the response window opened which was signaled by a switch in the cue—from "think" to "choose"—and the image of two arrow keys. The order of presentation of the two time pressure blocks was counterbalanced across participants.

Each pair of options involved a certain option and a risky option with probability p of winning the indicated amount and probability $1 - p$ of winning nothing; all gambles were of equal expected value except for 12 "catch" trials in which the expected value greatly favored an option (expected value = ± 90). Half of the stimuli were framed as losses and half of the stimuli were framed as gains. In both frames, the outcome probability of the risky options varied between extremely likely (0.90, 0.95 or 0.99) or extremely unlikely (0.10, 0.05, 0.01).

Information about each pair of options were presented in a manner similar to that used by Tymula et al. (2012): at the start of each trial participants were presented with two stacked bar-graphs where framing was depicted by the color of the bars (red for losses, and blue for gain) and the outcome probability was depicted by the proportion of the bar which

was colored (either red or blue), while the amounts ranged from $1 to $200 (see Fig. 2). The outcomes of gain trials were added to total earnings while the outcomes of loss trials were subtracted from total earnings—making the task incentives compatible. Participants were paid a bonus in proportion to their total earnings.

Data Analysis

In order to ensure that participants' choices were not made randomly but were based on the information presented, participants with less than 75% accuracy on catch trials across both conditions (operationalized as the proportion of choices which maximize expected value) were excluded from the sample, resulting in the exclusion of 21 participants. Participants who also failed to score above 70% accuracy during the last run of the digit-symbol task were also excluded from the sample—one in total. Finally, six participants were excluded for failing to meet the specified deadline resulting in a total exclusion of 28 participants of the 100 collected.

We used a mixed-effects logistic regression to predict risky versus certain choice on the basis of 1) the framing of the problem (losses or gains) 2) the outcome probability (coded as >0.5 or <0.5), and 3) time pressure condition (light or severe), and all two- and three-way interactions between these predictors. This regression model then gives us two terms of interest: the two-way interaction between probability and framing—an estimate of the fourfold pattern of choice effect since it represents the extent to which mean differences between gain and loss frames depend on outcome probability (large or small), and the three-way interaction between probability, framing and time pressure, which indicates the extent to which the presence of fourfold pattern is modulated by time pressure. Similarly, two additional regression models were run to test the effects of individual differences in processing speed on choice within each time pressure condition. Specifically, to assess the influence of individual differences in processing speed on choice, a similar regression was run for each time pressure condition except with normalized processing speed score added as an independent variable instead of time pressure condition. For all regressions, all categorical independent variables were effect coded and entered as both fixed and random effects. These regressions were estimated using the lme4 package (Pinheiro & Bates, 2002) for the R programming language.

Results

As predicted, under little time constraints participants exhibited a fourfold pattern of choice: they were both sensitive to the framing of the problem ($\beta = 1.44$, SE = 0.10, $p < .001$), and the interaction between the outcome probability and the framing of the problem ($\beta = -0.37$, SE = 0.18, $p = 0.04$).

However, under strong time pressure participants exhibited a marked framing effect, becoming less sensitive to outcome probability ($\beta = 0.28$, SE = 0.08, $p = 0.001$). Thus, the effect of time pressure on risky choice, surprisingly, changed participant's preferences from one ostensibly irrational pattern to another (see Fig. 3).

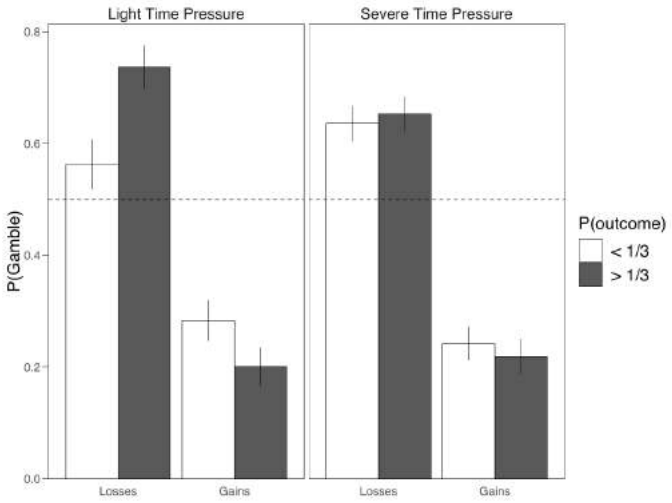


Figure 3: The effect of time pressure on risky decision-making. Under the light time pressure condition (LTP; 5 sec) participants showed more of a fourfold pattern compared to when under the severe time pressure condition (STP; 1.5 sec).

However, it remains unclear if this change in preference is a result of a reduction in the participant's ability to comprehend the gambles offered and correctly respond. It is possible that time limitations would lead to a nonspecific increase in choice randomness, as opposed to the proposed reduction in cognitive resources used. To test this alternative account, we compared the percentage of correct responses to the catch trials in the strong time pressure condition to test if it was significantly higher than chance. Using an Exact Binomial test, we were able to confirm that participants were capable of responding to the catch trials well above chance (Accuracy = 0.91, $p \leq 2.2 \times 10^{-16}$). This is to be expected as those participants who did not respond accurately in general—either due to lack of attention or understanding—were excluded from the analyses.

Finally, individual differences in processing speed were found to be related to risk preference in the predicted direction. Under light time pressure (LTP condition), individual differences in processing speed interacted with both framing of the problem ($\beta = 0.29$, SE = 0.13, $p = 0.02$) and the interaction between outcome probability and the framing of the problem ($\beta = -0.51$, SE = 024, $p = 0.03$). As processing speed increased, the extent to which participants exhibited a fourfold pattern also increased. Put another way, as processing speed decreased they were less likely to endorse a fourfold pattern (see Fig. 4). Moreover, these changes in risk preferences were not likely due to random performance on the task as processing speed and catch trials accuracy was not correlated ($r = -0.0081$, $p = 0.94$). Similarly, under severe time pressure (STP condition), both the two-way interaction between processing speed and the framing of the problem ($\beta = 0.24$, SE = 0.09, $p = 0.01$) and three-way interaction

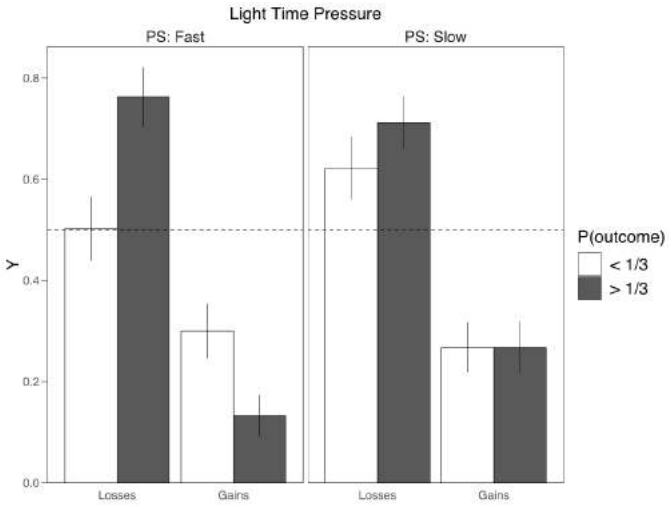


Figure 4: Under the light time pressure condition individual differences in processing speed (PS) predicted the extent to which participants endorsed a fourfold pattern. Processing speed conditions were assigned based on a median split.

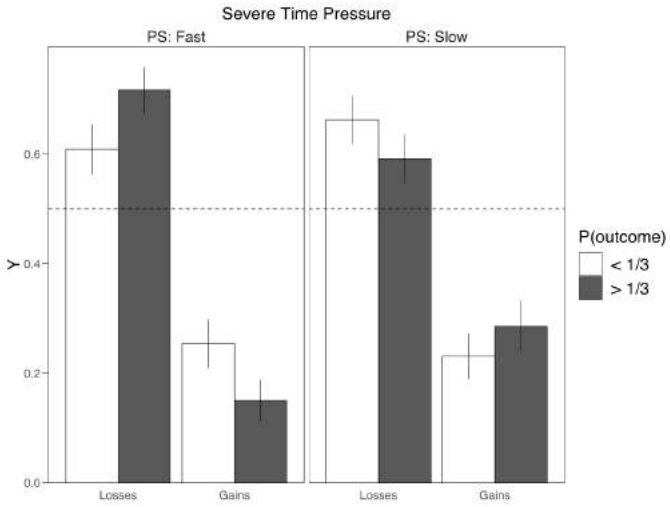


Figure 5: Under the severe time pressure condition individual differences in processing speed predicted the extent to which participants endorsed a fourfold pattern. For ease of exposition, processing speed scores (PS) were split based on the median.

between outcome probability, framing, and processing speed ($\beta = -0.30$, SE = 0.12, $p = 0.01$) were statistically significant (see Fig. 5).

General Discussion

The results presented here show that both situational and personal factors which limit cognitive resources contribute to changes in participants' risk preferences. Under little time constraints, participants produced a fourfold pattern of risk preferences—consistent with prospect theory (Kahneman & Tversky, 1979; Tversky & Kahneman, 1992). However, as predicted by a recently-proposed sample-based variant of expected utility theory, SbEU (Nobandegani et al., 2018), limitations in available cognitive resources—induced either through time pressure or measured by individual differences in processing speed—lead participants to go from showing a fourfold pattern to a framing effect.

While, descriptive models like prospect theory describe the risk preferences when selecting between gambles, this provide no account for how these preferences evolve over time or how limiting cognitive resources affects preferences. Thus, our results surprisingly reveal that the ostensibly irrational framing effect, fourfold pattern, and the demonstrated transition between the two, can all be explained as resulting from rational use of limited cognitive resources.

Interestingly, Stanovich and West (1998) demonstrated that performance on classic reasoning and judgment tasks and relationships to measures of academic achievement, correlates within individuals. Taken together with the results presented here, there is mounting evidence that the use of heuristics and biases may reflect the rational use of limited processing resources, thus suggesting that future models of choice should take into consideration individuals' cognitive abilities (or limitations). However, more work needs to be done to identify which specific cognitive capacities (e.g., processing speed or working memory) significantly contribute to the use of certain heuristics, thus providing an opportunity to better understand the cognitive mechanisms required for the performance of a given task.

Recent empirical work has also found time pressure to produce a similar pattern of results—severe time pressure lead to stronger framing effects—but failed to observe a fourfold pattern; instead they found individual preferences to reflect a weaker framing effect under light time pressure (Guo et al., 2017). However, the results were interpreted to arise from using a fast, intuitive systemas opposed to a slow, deliberative system.

Some have suggested that heuristics and biases are more than merely a result of flaws in human reasoning but are adaptive strategies to deal with conditions of limited time, knowledge or computational capacities (Simon, 1956; Todd & Gigerenzer, 2012) or take advantage of the structure of information in the environment (Todd & Gigerenzer, 2012). In fact, both experimental work (Goldstein & Gigerenzer, 2002), and theoretical work (e.g., Nobandegani & Shultz, 2019) has shown that fast and frugal algorithms can outperform standard integrative algorithms when knowledge is limited. Our results are in accordance with this compromise between normative and heuristic views of cognition as we show that biases like the framing effect can be explained as a strategic use of limited cognitive resources.

While previous work has also interpreted the framing effect as being a result of quick and intuitive thinking, these explanations make appeal to dual-process theory (De Martino, Kumaran, Seymour, & Dolan, 2006; Guo et al., 2017; Kah-

neman & Frederick, 2005; Sloman, 1996; Stanovich & West, 1998). Surprisingly, here we show that a rational *single-process* model can account for the observed results: an apparent framing effect can arise from limiting the number of samples in a resource-rational, sample-based expected utility model, SbEU (Nobandegani et al., 2018). A single-process framework is favorable over dual-process models as it provides a more parsimonious account of the observed effect.

Interestingly, unlike dual process theory would suggest, our results reveal that even when using a slow, deliberative system one can produce ostensibly irrational behavior. Concretely, according to our findings, deliberation takes us from one ostensibly irrational bias (framing effect) to another (the fourfold pattern of risk preferences)—and, as our work suggests, all of this can be understood as the optimal use of limited cognitive resources.

References

Crump, M. J., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating amazon's mechanical turk as a tool for experimental behavioral research. *PloS One*, *8*(3), e57410.

De Martino, B., Kumaran, D., Seymour, B., & Dolan, R. J. (2006). Frames, biases, and rational decision-making in the human brain. *Science*, *313*(5787), 684–687.

Di Mauro, C., & Maffioletti, A. (2004). Attitudes to risk and attitudes to uncertainty: experimental evidence. *Applied Economics*, *36*(4), 357–372.

Fehr-Duda, H., Bruhin, A., Epper, T., & Schubert, R. (2010). Rationality on the rise: Why relative risk aversion increases with stake size. *Journal of Risk and Uncertainty*, *40*(2), 147–180.

Gigerenzer, G., & Selten, R. (2002). *Bounded rationality: The adaptive toolbox*. MIT press.

Goldstein, D. G., & Gigerenzer, G. (2002). Models of ecological rationality: the recognition heuristic. *Psychological Review*, *109*(1), 75.

Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in Cognitive Science*, *7*(2), 217–229.

Guo, L., Trueblood, J. S., & Diederich, A. (2017). Thinking fast increases framing effects in risky decision making. *Psychological Science*, *28*(4), 530–543.

Icard, T. (2014). Toward boundedly rational analysis. In *Proceedings of the* 36^{th} *Annual Conference of the Cognitive Science Society* (pp. 2391-2396). Austin, TX: Cognitive Science Society.

Kahneman, D., & Frederick, S. (2014). A model of heuristic judgment. *The Cambridge Handbook of Thinking and Reasoning*.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decisions under risk. *Econometrica*, *47*, 263–291.

Levin, I. P., & Gaeth, G. J. (1988). How consumers are affected by the framing of attribute information before and after consuming the product. *Journal of Consumer Research*, *15*(3), 374–378.

Lieder, F., Griffiths, T. L., & Hsu, M. (2018). Overrepresentation of extreme events in decision making reflects rational use of cognitive resources. *Psychological Review*, *125*(1), 1.

Loke, W. H., & Lau, S. L. L. (1992). Effects of framing and mathematical experience on judgments. *Bulletin of the Psychonomic Society*, *30*(5), 393–395.

Markowitz, H. (1952). The utility of wealth. *Journal of political Economy*, *60*(2), 151–158.

Mathias, S. R., Knowles, E. E., Barrett, J., Leach, O., Buccheri, S., Beetham, T., ... Glahn, D. C. (2017). The processing-speed impairment in psychosis is more than just accelerated aging. *Schizophrenia Bulletin*, *43*(4), 814–823.

McNeil, B. J., Pauker, S. G., Sox Jr, H. C., & Tversky, A. (1982). On the elicitation of preferences for alternative therapies. *Preference, Belief, and Similarity*, *583*.

Miller, J. G. (1960). Information input overload and psychopathology. *American Journal of Psychiatry*, *116*(8), 695–704.

Moxey, A., OConnell, D., McGettigan, P., & Henry, D. (2003). Describing treatment effects to patients. *Journal of General Internal Medicine*, *18*(11), 948–959.

Nobandegani, A. S. (2017). The Minimalist Mind: On Mininality in Learning, Reasoning, Action, & Imagination. McGill University, PhD Dissertation.

Nobandegani, A. S., da Silva Castanheira, K., Otto, A. R., & Shultz, T. R. (2018). Over-representation of extreme events in decision-making: A rational metacognitive account. In *Proceedings of the* 40^{th} *Annual Conference of the Cognitive Science Society* (pp. 2391-2396). Austin, TX: Cognitive Science Society.

Nobandegani, A. S., & Shultz, T. R. (2019). Toward a formal science of heuristics. In *Proceedings of the* 41^{st} *Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Salthouse, T. A. (1985). Speed of behavior and its implications for cognition. In *Handbook of the Psychology of Aging,* 2^{nd} *Ed. (pp. 400-426).* New York, NY.

Scholten, M., & Read, D. (2014). Prospect theory and the forgotten fourfold pattern of risk preferences. *Journal of Risk and Uncertainty*, *48*(1), 67–83.

Shadlen, M. N., & Shohamy, D. (2016). Decision making and sequential sampling from memory. *Neuron*, *90*(5), 927–939.

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review*, *63*(2), 129.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*(1), 3.

Stanovich, K. E., & West, R. F. (1998). Individual differences in rational thought. *Journal of Experimental Psychology: General*, *127*(2), 161.

Starmer, C. (2000). Developments in non-expected utility theory: The hunt for a descriptive theory of choice under risk. *Journal of Economic Literature*, *38*(2), 332–382.

Todd, P. M., & Gigerenzer, G. (2012). *Ecological rationality: Intelligence in the world*. OUP USA.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*(4481), 453–458.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*(4), 297–323.

Tymula, A., Belmaker, L. A. R., Roy, A. K., Ruderman, L., Manson, K., Glimcher, P. W., & Levy, I. (2012). Adolescents risk-taking behavior is driven by tolerance to ambiguity. *Proceedings of the National Academy of Sciences*, *109*(42), 17135–17140.

von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior (commemorative edition)*. Princeton University Press.

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? optimal decisions from very few samples. *Cognitive Science*, *38*(4), 599–637.

Zur, H. B., & Breznitz, S. J. (1981). The effect of time pressure on risky choice behavior. *Acta Psychologica*, *47*(2), 89–104.

Lifting the Curse of Knowing: How Feedback Improves Readers' Perspective-Taking

Debby Damen (d.j.damen@uvt.nl), Marije van Amelsvoort (m.a.a.vanamelsvoort@uvt.nl), Per van der Wijst (per.vanderwijst@uvt.nl), Emiel Krahmer (e.j.krahmer@uvt.nl)
Tilburg center for Cognition and Communication, Tilburg University, PO Box 90153, 5000 LE, The Netherlands

Abstract

Previous studies have shown that readers often overestimate the similarity between their perspective and the perspective of protagonists in a story. This egocentric projection is argued to originate from readers' tendency to use their own knowledge as a frame of reference from which they (insufficiently) adjust away to account for protagonists' less informed perspective. This experimental study demonstrated that readers use feedback about protagonists' knowledge status to draw inferences that are more accurate on future perspective-taking trials. Readers who were given the opportunity to learn through feedback not only *adjusted* their perspective-judgment more than those who did not receive feedback, these readers also showed less egocentric projection on future assessments.

Keywords: perspective-taking; egocentricity bias; anchoring and adjustment; privileged information; feedback

Introduction

Communication processes rely on our ability to successfully reason about others' mental states. Research examining this perspective-taking, however, paints a contradictory picture with regard to communicators' *tendency* to be accurate perspective-takers. On the one hand, a large body of research suggests that communicators rapidly and accurately assess others' perspective (Brown-schmidt, Gunlogson, & Tanenhaus, 2008; Nadig & Sedivy, 2002). In contrast to this view, studies have shown that rapid (and automatic) judgments of others' mental state are often influenced by communicators' own knowledge and attentional status (Apperly et al., 2010; Keysar, Barr, Balin, & Brauner, 2000). These studies argue that perspective-taking activities follow an *egocentric anchoring and adjustment* process (Epley, Keysar, Van Boven, & Gilovich, 2004). During this perspective-taking process, communicators adopt another's perspective by using their own perceptions as a frame of reference and adjust this frame to take into account possible informational differences between their own and others' perceptions. These *perspective-adjustments*, however, are often insufficient due to the immediate accessibility or saliency of one's own perceptions. The accessibility and, hence, saliency of one's own knowledge in contrast to the seemingly impermeable nature of the other's mind makes it hard for perceivers to ignore or suppress their own perception as a possible estimate of others' perspective. The failures to inhibit one's own perspective during perspective-taking may result in egocentric projection (Ames, 2004), during which perceivers wrongly assume that their private perspective is shared by others.

Studies have shown that egocentric projection might also occur during reading when readers try to take story characters' perspective (e.g., Keysar, 1994; Weingartner & Klin, 2005, 2009). In these studies, readers overestimated the extent to which their knowledge was accessible to uninformed protagonists. That is, readers read stories in which a speaker protagonist sent an ambiguous message (e.g., "About that dancing class: I can't think of better ways to spend my Tuesday evenings") to a friend. Readers learned how to interpret the speaker's message by the clarifying event information they received beforehand. When this disambiguating information suggested counterfactual (e.g., "The dance class had been dull") rather than factual (e.g., "The dance class had been interesting) information, readers interpreted the speaker's message to be sarcastic. This disambiguating information was not accessible to the recipient of the speaker's message and, for each story, this addressee protagonist had no reason to believe that the speaker was being sarcastic. Studies showed, however, that readers were very likely to use their own interpretation of the speaker's communicative intention to judge that the uninformed addressees would perceive the speaker's message in a similar way. That is, when privileged information suggested that the speaker was being sarcastic, readers assumed addressees would also perceive this sarcasm. In these instances, readers' own knowledge about the speaker's experience "cursed" (Birch & Bloom, 2007; Keysar, 1994) their ability to suppress their own interpretation of the speaker's communicative intention while imaging the perspective of the uninformed protagonists.

Epley and his colleagues (2004) showed that this "curse of knowledge" (Keysar, 1994) effect on perspective-taking originates from an egocentric anchoring and insufficient adjustment process. In their "Sarcastic Messages" experiment, Epley et al. (2004) asked readers to read similar stories in which a speaker protagonist left ambiguous voicemail messages on the answering machine of his friends. Subsequently, readers indicated either the speaker's intention with his voicemail or how they thought the recipient of the voicemail would interpret the message. Following egocentric anchoring, Epley et al. (2004) expected readers to interpret the addressee's perception of the voicemail based on information that was accessible to themselves. Findings indeed showed that readers were more likely to indicate that addressees would perceive the speaker's sarcasm when readers' privileged information suggested the speaker was being sarcastic rather than sincere. Epley et al. (2004) further

showed that this perception of sarcasm was more moderate when readers only judged addressees' interpretation of the message rather than only their own perception of sarcasm. The more moderate perception of the speaker's sarcasm in the perspective-taking condition showed that readers acknowledged that the messages sounded more ambiguous to the uninformed addressees than to themselves. However, since readers still believed that addressees perceived the speaker's sarcasm, readers' perspective-judgments still reflected their own knowledge about the speaker's communicative intention. Even though readers adjusted their egocentric interpretation into a more moderate judgment, these adjustments were not sufficient in order to reflect addressees' true perspective.

Inhibiting Egocentric Information

Perceivers learning to inhibit their own cognitions during mental state reasoning can perhaps counter insufficient perspective-adjustments. For instance, recent perspective-switching research (Samuel, Roehr-Brackin, Jelbert, & Clayton, 2018) showed that communicators found it difficult to switch back to an egocentric judgment once they had learned to adopt another frame of reference. In addition to this, it is argued that the more cues perceivers receive about the knowledge status of others, the less likely they are expected to engage in egocentric projection (Eyal, Steffel, & Epley, 2018; Keysar, Barr, & Horton, 1998; West, 1996). However, studies have shown that directing perceivers' attention to focus on other people's knowledge and attentional status does not always improve perspective-taking accuracy (Damen, Van der Wijst, Van Amelsvoort, & Krahmer, 2018; Eyal et al., 2018). For instance, in a direct replication and extension of Epley et al.'s (2004) "Sarcastic Messages" study, Damen et al. (2018c)[1] examined whether explicit and repeated instructions to focus on addressee protagonists' *uninformed* perspective helps readers to acknowledge that their privileged information was not accessible to these addressees. However, not only did Damen et al. (2018c) replicate readers' egocentric anchoring and insufficient adjustment during perspective-taking, their findings also showed that explicit perspective-focus instructions did not stimulate the adjustment phase. Regardless of an explicit focus on addressees' uninformed perspective, readers still overestimated the extent to which the uninformed protagonists shared their interpretation of the voicemail.

Gaining Interpersonal Insight

Readers in Epley et al. (2004) and Damen et al. (2018c) were more likely to rely on privileged rather than common-ground information while interpreting protagonists' perspective. Interesting to note here is that readers' perspective-taking appertained to a "top-down process" (Eyal et al., 2018), whereby readers *selected* perspective-information that, according to them, was the most relevant to use. In turn, highlighting or enhancing the accessibility of more reliable information (i.e., protagonists' perspective) did not make readers more likely to *use* this information during mental state reasoning. This finding raises the question whether, during this top-down inferencing, readers did not see the need to *adjust* their judgment because they were unaware of its inaccuracy. In this case, increasing readers' awareness of the inaccuracy of their judgments might make them better future perspective-takers.

West (1996) found some support for this line of reasoning by showing that an awareness of inaccurate (egocentric) predictions allowed perceivers to learn from their mistake and to improve their perspective-taking skills. In West (1996), participants learned to predict a target's preference for quilt patterns through the feedback they received from the target. In each trial, agents made a prediction of the target's preference for the pattern (rated from "1 = dislike very much" to "7 = like very much"). Subsequently, the target responded by showing his actual preference (rating) for the pattern, after which agents rated their own preference. Findings showed that the agents' first predictions of the target's preferences showed egocentric projection. That is, if agents liked the pattern, they assumed the target did too. Interestingly, this egocentric projection decreased on subsequent trials due to the target's feedback. The more agents learned about the target's preferences, the less likely they were to project their egocentric preferences onto the target on subsequent perspective-taking trials. Apparently, feedback about their perspective-judgments allowed agents to disregard their own preferences and to select perspective-information that more reliably predicted the target's true perspective.

In addition, recent research by Eyal and colleagues (2018) showed that *receiving* accurate perspective-information rather than relying on existing knowledge improved communicators' perspective-taking accuracy. In Eyal et al. (2018), romantic partners who had the opportunity to discuss each other's preferences on a range of topics were able to use this gained insight on future assessments of their partner's preferences. This in contrast to the partners in the perspective-taking conditions who were not given this discourse opportunity, but who had to rely solely on their *imagination* of their partner's preferences. According to Eyal et al. (2018), the act of trying to *take* others' perspective does not necessarily lead to a more accurate insight into these imagined mental states, because perceivers are very likely to select the wrong information to base their inferences on. In this sense, providing communicators with the opportunity to *gain* reliable perspective-information of which they are also *aware* of its appropriateness should improve perspective-taking accuracy.

[1] Damen et al.'s (2018c) preregistration, materials and data are available in the Open Science Framework (doi: 10.17605/osf.io/kv5mu).

Current Study

This study investigates the role of feedback as a strategy to gain accurate insight into others' perspective. In particular, we examine whether confronting readers with the accuracy of their perspective-judgment (i.e., feedback) allows them to accurately assess protagonists' perspective on subsequent perspective-taking trials. Additionally, we aim to explore whether readers adjust their perspective differently depending on how they gain this perspective-insight. In this study, we contrast two approaches. For the first approach, we rely on perceivers' "bottom-up inferencing" (e.g., Eyal et al., 2018), through which perceives gain interpersonal insight by perceiving others' thoughts and actions. Since this strategy indirectly communicates to perceivers whether their first assessment had been correct, we will term this approach as *indirect feedback*. We contrast this approach against a strategy through which perceivers gain insight by receiving explicit feedback about the accuracy of their assessment (e.g., West, 1996). We will term this type of information as *direct feedback* and we will use this term to refer to the situation in which readers are made explicitly aware that they have made an error and why their judgment was inaccurate (e.g., Ellis, Loewen, & Erlam, 2019).

This study replicates Damen et al.'s (2018c) study in which readers judge addressees' interpretation of voicemails sent by a speaker protagonist. We extend the experimental design by adding a feedback manipulation and a subsequent second measurement of readers' judgment of addressees' interpretation of the voicemail. In line with previous egocentric anchoring findings (Epley et al., 2004; Damen et al., 2018c), we expect readers to overestimate the extent to which uninformed addressee protagonists will also perceive a speaker's sarcasm. We expect that this egocentric projection occurs more at readers' first than at their second prediction of addressees' perspective. In addition, we expect that this relationship is qualified by whether readers receive feedback about the accuracy of their first prediction. In particular, compared to a baseline in which readers do not receive feedback, we expect that both feedback types will help readers to *adjust* their first prediction into a perspective-judgment that more accurately reflects addressees' sincere interpretation of the message. Finally, we expect that readers' second predictions will be more accurate after they had been explicitly told their judgment had been wrong (direct feedback), than when readers need to infer the accuracy of their judgment from a description of addressees' response to the message (indirect feedback). This study is preregistered in the Open Science Framework (doi 10.17605/osf.io/kpw6u).

Method

Participants

A total of 149 undergraduates were invited to participate in the study. Seven participants were excluded because they recognized the voice-actor ($N = 5$) or because they were non-native speakers of the language of the experiment ($N = 2$). The remaining participants were randomly allocated to the control ($N = 48$), direct feedback ($N = 47$), and indirect feedback ($N = 47$) conditions (105 women, 37 men, $M_{age} = 21.57$, age-range 18-38).

Design

In each condition, participants read 12 scenarios in which a speaker protagonist (Tom) left a voicemail-message on the answering machine of an addressee protagonist. After hearing this voicemail, participants judged the addressee's perception of the speaker's sarcasm both before (time 1) and after (time 2) they received feedback about their first perspective-judgment. This resulted in a 3 (*Condition*: control, direct feedback, indirect feedback) x 2 (*Time*: time 1, time 2) design in which *Condition* was treated as a between-subjects factor and *Time* as a within-subjects factor.

Procedure and Materials

We replicated and extended the experimental materials and procedure of Damen et al.'s (2018c) "interpretation" condition. On a computer, participants read 12 stories describing an event in the life of Tom. For instance, in the story "The Dance Class", participants read the following:

> Tom was on his way to the first night of his ballroom dancing class when he saw Eileen, an old friend from his dorm last year. When he told her that he was on his way to a ballroom dancing class, she excitedly replied, "I'm thinking of taking that class, but I can't make it to tonight's class--I am having dinner with friends. Could you call me when you get back and tell me how it is?"

Subsequently, participants learned that Tom's experience had been either negative (e.g., "(…) the instructor spent the entire time taking attendance and filling out lengthy forms and questionnaires.") or positive (e.g., "(…) the instructor spent the entire time teaching the class fun, new dances."). Both experiences followed with Tom leaving a voicemail on the answering machine of his friend. In "The Dance Class" story, Tom left the following message:

> Eileen, this is Tom. Hope you enjoyed your dinner. About that ballroom dancing class: Judging from tonight's class, I can't think of better ways to spend my Tuesday evenings. Anyways, give me back a call and I'll fill you in on the details. Bye.

We re-used the 12 voicemails from Damen et al. (2019b) who demonstrated the validity of the voicemails. In a separate rating experiment, Damen et al. (2019b) asked listeners to rate the voicemails in the absence of clarifying (positive, negative) event information (1 = as very sincere, 7 = as very sarcastic). This rating experiment showed that the voicemails sounded truly ambiguous to the uninformed listeners. That is, participants rated the voicemails to sound neither as very sarcastic or as very sincere (M = 3.73, SD = 0.83).

We followed the experimental procedure described in Damen et al. (2019b), and asked participants to indicate – immediately after listening to Tom's voicemail – how the addressee protagonist (Tom's friend) would perceive the voicemail message (1 = definitely as sincere, 7 = definitely as sarcastic). For this study, this constituted the first measurement of participants' judgment of the addressee's perception of sarcasm (time 1). All stories were presented to participants in digital booklets, and half the stories in these booklets described a positive event, whereas the other half described a negative event. We created four versions of these booklets: The first booklet contained a random order of negative versus positive events (booklet 1), and another one contained its mirror image (booklet 2). Additionally, for each booklet, we created a version that contained a reversed order of the events. In contrast to Damen et al. (2019b), we chose to focus on participants' judgments of the addressee protagonist's perspective only for those stories in which participants' privileged information suggested that Tom was being sarcastic (negative events). We thereby treated the stories that suggested Tom was being sincere (positive event) as fillers.

Additionally to our replication procedure, we manipulated the extent to which participants received feedback about their first judgment of the addressee's perception of sarcasm. This feedback was automated in the sense that the computer provided participants with either direct or indirect feedback. In the direct feedback condition, participants' received explicit feedback about the accuracy (i.e., ranging from "You are completely right!" to "You are completely wrong!") of their judgment based on the answer they provided on the 7-point scale (see Table 1).

Table 1: Example of the direct feedback participants received after judging Eileen's perception of sarcasm (1 = definitely as sincere, 7 = definitely as sarcastic)

Answer	Direct Feedback
1	**"You are completely right!** Eileen thinks that Tom liked the class."
2 / 3	**"You are almost right!** Eileen thinks that Tom liked the class."
4	**"You are not right!** Eileen thinks that Tom liked the class."
5 / 6	**"You are wrong!** Eileen thinks that Tom liked the dance class."
7	**"You are completely wrong!** Eileen thinks that Tom liked the class."

Participants in the indirect feedback condition received feedback about the accuracy of their perspective-judgement regardless of their choice on the 7-point scale. This feedback constituted a follow-up text that described addressees' sincere interpretation of Tom's voicemail. For instance in "The Dance Class" story, participants could derive from Eileen's thoughts and actions in response to Tom's voicemail that she thought that Tom had enjoyed attending the class:

> After saying goodbye to her friends, Eileen cycled home. She decided she was going to search for her dancing shoes the minute she would arrive at home. She could hardly wait to join Tom in the dance class. If Tom had liked the dance class, she definitely would like it too.

In contrast to the two feedback conditions, participants in the control condition did not receive feedback about their first assessment of addressees' perception of sarcasm. Subsequently to their first judgment, these participants read a follow-up text that described the addressee's thoughts and actions that did not target her interpretation of the voicemail:

> After saying goodbye to her friends, Eileen cycled home. She and her friends had enjoyed dinner. They had known each other since high school and had built up a close friendship. Although they only saw each other a few times a year, it was always like they never had been apart.

In all three conditions, participants subsequently re-judged addressees' interpretation of the voicemail (1 = definitely as sincere, 7 = definitely as sarcastic). After this second assessment, participants answered a comprehension question that encouraged participants to attend to the materials. These 12 questions did not target participants' privileged information. When participants answered the comprehension question incorrectly, they were informed to attend to the materials more carefully. Participants answered almost all questions correctly (M = 10.52, SD = 1.07), but the number of correct responses differed between conditions, $H(3)$ = 9.73, p <.01. Pairwise comparisons with adjusted p-values showed that participants answered more comprehension questions correctly in the indirect feedback condition (M = 10.81, SD = 0.95) than in the direct feedback condition (M = 10.13, SD = 1.15), (p < .01). The accuracy scores did not differ between the control and the two feedback conditions (p > .05). After reading 12 stories, participants filled out their demographics and were debriefed about the purpose of the experiment.

Results

We computed a mean sarcasm score of participants' first (time 1) and second (time 2) judgment of addressees' perception of the speaker's sarcasm for the scenarios in which participants' privileged information suggested that the speaker was being sarcastic (negative events). We submitted these mean scores to a mixed analysis of variance in which *Condition* (control, direct feedback, indirect feedback) was treated as a between-subjects factor and participants' judgment of addressees' perception of sarcasm (*Time*; time 1, time 2) as a within-subjects factor. The means of participants' judgment of addressees' perception of sarcasm as a function of *Time* and *Condition* are presented in Figure 1.

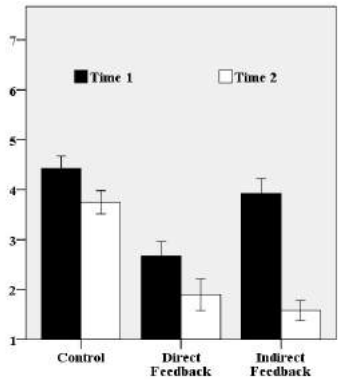


Figure 1: Mean scores of participants' judgment of addressees' perception of sarcasm (1 = definitely as sincere, 7 = definitely as sarcastic) as a function of *Time* (time 1, time 2) and *Condition* (control, direct feedback, indirect feedback).

In line with our first hypothesis, participants thought addressees would perceive the speaker's sarcasm more at their first ($M_{time\ 1}$ = 3.68, *SD* = 1.19) than at their second perspective-judgment ($M_{time\ 2}$ = 2.42, *SD* = 1.29), $F(1, 139) = 198.96$, $p < .001$, $\eta_p^2 = .59$[2].

We expected that feedback (direct, indirect) would help participants to adjust their first prediction of addressees' perspective into a judgment that more accurately reflected addressees' sincere interpretation of the voicemail than when this feedback was absent (control). Results indeed showed that the main effect of *Time* was qualified by a significant interaction with *Condition*, $F(2, 139) = 35.93$, $p < .001$, $\eta_p^2 = .34$[2]. Pairwise comparisons that compared participants' perspective-taking accuracy of their second perspective-judgment showed that participants had more successfully adjusted their first prediction after they had received both direct (*M* = 1.89, *SD* = 0.13) and indirect (*M* = 1.58, *SD* = 0.13) feedback, compared to the control condition in which this feedback was absent (*M* = 3.75, *SD* = 0.12), $p < .001$. The accuracy of participants' second prediction did not differ between the two feedback types ($p = .245$).

Interestingly, results also showed that participants' perspective-taking accuracy of their first prediction differed as a function of *Condition*. Pairwise comparison revealed that participants in the control condition (*M* = 4.43, *SD* = 0.14) thought addressees would perceive sarcasm more at time 1 than the participants in both the direct (*M* = 2.67, *SD* = 0.14, $p < .001$) and indirect (*M* = 3.93, *SD* = 0.14, $p < .05$) feedback conditions. For their first prediction, participants in the indirect feedback condition also thought addressees would perceive sarcasm more than the participants in the direct feedback condition ($p < .001$).

To examine whether the degree to which participants adjusted their perspective differed as a function of *Condition*, we computed a mean difference score between participants' first and second judgment of addressees' perception of the speaker's sarcasm and submitted this difference-score to an one-way analysis of variance. This follow-up analysis showed that participants' perspective-adjustments differed between conditions, $F(2, 139) = 35.93$, $p < .001$. Simple contrasts revealed that participants had adjusted their perspective more in both the direct ($M_{difference}$ = 0.78, *SE* = 0.16) and indirect ($M_{difference}$ = 2.34, *SE* = 0.16) feedback conditions compared to the control condition ($M_{difference}$ = 0.68, *SE* = 0.15), $t(139) = -4.63$, $p < .001$. In addition, participants who had received indirect feedback had adjusted their perspective more than those who had received direct feedback, $t(139) = 7.10$, $p < .001$.

Discussion

This study examined the influence of feedback on readers' perspective-taking. In an extension study of Damen et al. (2018c), we have shown that readers learned from the feedback they received to make better perspective-taking judgments immediately after the feedback (within the same trial) and on subsequent trials. The extent to which readers improved their perspective-taking accuracy depended on the type of feedback they received. In contrast to our expectation, we found that readers' predictions were more accurate immediately after indirect rather than direct feedback. This could have been due to the benefit these readers had from having to exert more cognitive effort to calculate addressees' interpretation. That is, readers who received the feedback indirectly not only had to infer addressees' interpretation of the voicemail from the description of addressees' actions and thoughts, these readers also had translate this information to a reliable score (i.e., 1 = definitely as sincere, 7 = definitely as sarcastic). This in contrast to the readers who were explicitly informed about the extent to which their judgment deviated from addressees' actual interpretation (direct feedback) and who, therefore, could have converted this feedback to a rating more easily.

Interestingly, the accuracy of readers' *first* predictions also differed due to the type of feedback they had received on previous trials. Although readers receiving indirect feedback made better adjustment *within* the same perspective-taking trial, their first predictions on new trials showed more egocentric projection errors than those who received direct feedback[3]. This finding needs to be interpreted with caution, because it could have been the result of task characteristics. That is, for each trial, readers receiving direct feedback could have learned that a sincere interpretation (i.e., a score of 1) was the correct response for all experimental trials, reducing egocentric projection on first predictions. This in contrast to the indirect feedback condition in which readers could have been more cautious to assume the addressees' sincere interpretation until they had actually received addressees' reaction to the voicemail. However, in all experimental conditions and for each experimental trial, the correct

[2] The findings remained unchanged when we controlled for the presentation order of the scenarios.

[3] This finding could also be an explanation as to why we see bigger adjustments in the indirect than the direct feedback condition.

response always reflected addressees' sincere interpretation of the messages. Therefore, this possible confound cannot explain why there are still significant differences within experimental trials and adjustment differences across conditions.

Although readers in the control condition did not receive feedback about the accuracy of their interpretation, these readers also adjusted their first prediction to a more accurate second prediction of the addressees' perspective. This 'positive' adjustment could have been the result of readers reflecting on their earlier assessment and subsequently coming to a more accurate conclusion (e.g., Epley et al., 2004). However, important to note is that these adjustments were still less accurate than when readers were provided with reliable information (feedback) to base their re-assessment on.

In line with findings of both West (1996) and Eyal and colleagues (2018), this study showed that providing readers with reliable perspective-information ("perspective-getting") allows them to disregard their own knowledge and to use this new information to more accurately predict others' perspective. It should be noted that readers in this experiment paid attention to the feedback they received and, therefore, could have been more aware that they could or *should* use this information to adjust their predictions appropriately. In addition, in Eyal et al. (2018), the discourse through which partners gained relevant perspective-information was demarcated with regard to the topics partners had to discuss. Therefore, an interesting question for future research is whether this "perspective-getting" effect generalizes to situations in which reliable perspective-information (and its appropriateness) is not been made explicit.

References

Ames, D. R. (2004). Inside the mind reader's tool kit: Projection and stereotyping in mental state inference. *Journal of Personality and Social Psychology*, *87*(3), 340–353.

Apperly, I. A., Carroll, D. J., Samson, D., Humphreys, G. W., Qureshi, A., & Moffitt, G. (2010). Why are there limits on theory of mind use? Evidence from adults' ability to follow instructions from an ignorant speaker. *Quarterly Journal of Experimental Psychology*, *63*(6), 1201–1217.

Birch, S. A. J., & Bloom, P. (2007). The curse fo knowledge in reasoning about false beliefs. *Psychological Science*, *18*(5), 382–386.

Brown-Schmidt, S., Gunlogson, C., & Tanenhaus, M. K. (2008). Interpreting questions during interactive conversation. *Cognition*, *107*(3), 1122–1134.

Damen, D., Van der Wijst, P., Van Amelsvoort, M., & Krahmer, E. (2018a). Perspective-taking in referential communication: Does stimulated attention to addressees' perspective influence speakers' reference production? *Journal of Psycholinguistic Research*, 1–32.

Damen, D., Van der Wijst, P., Van Amelsvoort, M., & Krahmer, E. (2018b). The curse of knowing: The influence of explicit perspective-awareness instructions on perceivers' perspective-taking. In T. T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 1578–1583). Austin, TX: Cognitive Science Society.

Damen, D., Van der Wijst, P., Van Amelsvoort, M., & Krahmer, E. (2018c). *Can the curse of knowing be lifted? The influence of explicit perspective-focus instructions on readers' perspective-taking*. Manuscript submitted for publication.

Ellis, R., Loewen, S., & Erlam, R. (2019). Implicit and explicit corrective feedback and the acquisition of L2 grammar. *Studies in Second Language Acquisition*, *28*(2), 339–368.

Epley, N., Keysar, B., Van Boven, L., & Gilovich, T. (2004). Perspective taking as egocentric anchoring and adjustment. *Journal of Personality and Social Psychology*, *87*(3), 327–339.

Eyal, T., Steffel, M., & Epley, N. (2018). Perspective mistaking: Accurately understanding the mind of another requires getting perspective, not taking perspective. *Journal of Personality and Social Psychology*, *114*(4), 547–571.

Keysar, B. (1994). The illusory transparency of intention: Linguistic perspective taking in text. *Cognitive Psychology*, *26*, 165–208.

Keysar, B., Barr, D. J., Balin, J. A., & Brauner, J. S. (2000). Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science*, *11*(1), 32–38.

Keysar, B., Barr, D. J., & Horton, W. S. (1998). The egocentric basis of language use: Insights from a processing approach. *Current Directions in Psychological Science*, *7*(2), 46–49.

Nadig, S. A., & Sedivy, J. C. (2002). Evidence of perspective-taking constraints in children's on-line reference resolution. *Psychological Science*, *13*(4), 329–336.

Samuel, S., Roehr-Brackin, K., Jelbert, S., & Clayton, N. S. (2018). Flexible egocentricity: Asymmetric switch costs on a perspective-taking task. *Journal of Experimental Psychology Learning Memory and Cognition*, *45*(2), 213-218.

Weingartner, K. M., & Klin, C. M. (2005). Perspective taking during reading: An on-line investigation of the illusory transparency of intention. *Memory & Cognition*, *33*(1), 48-58.

Weingartner, K. M., & Klin, C. M. (2009). Who knows what? Maintaining multiple perspectives during reading. *Scientific Studies of Reading*, *13*(4), 275-294.

West, P. M. (1996). Predicting Preferences: An examination of agent learning. *Journal of Consumer Research*, *23*(1), 68–80.

Abstract concepts and the suppression of arbitrary episodic context

Abstract

Context is important for abstract concept processing, but a mechanism by which it is encoded and re-instantiated with concepts is unclear. We used a source-memory paradigm to determine whether *episodic* context is attended more when processing abstract concepts. Experiment 1 presented abstract and concrete words in colored boxes at encoding. At test, memory for the frame color was worse for abstract concepts, counter to our predictions. Experiment 2 showed the same pattern when colored boxes were replaced with male and female voices. Experiment 3 presented words from encoding in the same or different box color to determine whether a greater advantage is conferred by context retention in memory for abstract concepts. There was instead a *disadvantage*: abstract concepts were *less* likely to be identified when the encoding color was retained at test. Concrete concepts are more sensitive to simple episodic detail, and in abstract concepts, *arbitrary* context may be suppressed.

Keywords: concepts, semantic memory, episodic memory, abstract concepts, concreteness

Introduction

Abstract concepts like *decision* are central to the human experience, yet little is understood about how they are processed. Contextual information is thought to be important to abstract concepts—the specific meaning of *decision* varies more depending on context than does the meaning of *river*. While a river in New England shares many properties with a river in Papua New Guinea, consider the case of *decision*: your decision on which beverage to buy at a café late at night differs greatly from the decision a judge might make in determining sentencing for a felon. It is the context which determines the antecedents, outcomes, and consequences in these two instantiations of *decision*. While it seems that context should be important in processing abstract concepts, the mechanism by which context is encoded and re-instantiated with the concept remains unclear. One possibility is that the episodic memory system, which supports encoding and recall of contextually detailed memories, is critical in understanding abstract concepts. Thus, here we probed a potential mechanism underpinning abstract concepts' sensitivity to context by using a source memory paradigm to test whether *episodic* context is better bound to abstract than concrete concepts.

Episodic memory is classically defined as explicit memory for unique events (Tulving, 1983, 2002), where episodic context is the detail that colors an episode. There are circumstances under which we are more likely to *encode*, and therefore, *recall* the arbitrary contents of a particular episode (e.g., the color of a frame or the identity of a speaker). A standard paradigm for assessing this ability is the source memory task (see Davachi, 2006; Yonelinas, 2001, 2002). In this task, participants are asked at test to determine whether an item (e.g., a word) was previously presented in an exposure phase, and then probed as to whether they can recognize some prior contextual detail. Greater confidence in having seen a word at exposure is associated with greater likelihood of having encoded the contextual detail (e.g., Kirwan, Wixted, & Squire, 2008; Yu, Johnson, & Rugg, 2012). Therefore, we would predict that greater confidence in having seen or heard a word during an encoding phase is associated with better memory for an arbitrary context, such as a box color or voice, at a test phase.

In addition to confidence in recollection or strength of the memory, emotionality in words, including both valence and arousal (Kensinger & Corkin, 2003), influences the likelihood of recalling the context in which something was presented, suggesting that the *content* of the stimuli at exposure can influence the likelihood that the context is identified at test. More specifically, *conceptual* or *semantic* content might affect likelihood of context encoding. In this set of experiments, we investigated whether this is true for concreteness: are we better at encoding contextual detail for abstract than for concrete concepts?

The notion that episodic context is more important for interpreting abstract concepts suggests that we should be more sensitive to the episodic context in which abstract concepts are placed and, in turn, be more accurate at retrieving even non-systematically related elements in the context. We opted to test this hypothesis by examining whether *arbitrary* contexts are better recognized when paired with abstract as compared to concrete concepts. Because memory is generally better for concrete than for abstract words (e.g., Paivio, Walsh, & Bons, 1994), we expected that although overall memory for concrete concepts would be better, *when* abstract concepts are correctly recognized, the context would be better encoded. To foreshadow the results, we find evidence *against* this hypothesis, suggesting that arbitrary episodic context may be *inhibited* in abstract concepts. In the General Discussion, we propose an alternative framework in which these results might be accommodated.

In the studies below, context is operationally defined as an aspect of a stimulus that is irrelevant to the central stimulus, such as whether a target word is presented within a red or green frame or whether stimuli are presented in a male or female voice.

Experiment 1

Methods

Participants Forty-two University of Connecticut (UConn) students with normal or corrected-to-normal vision and hearing provided informed consent and received course credit for participating. One participant was excluded for non-compliance, leaving $N = 41$. The study was approved by the UConn IRB.

Stimuli In the encoding phase, 100 (60 target, 40 non-target) abstract (e.g., *decision*) and 100 (60 target, 40 non-target) concrete (e.g., *chair*) noun concepts were used. (Non-targets were synonym words which functioned as positive responses for the synonym-judgment task described below. Targets were non-synonyms.) Stimuli were matched across all stimulus subsets on word length and word frequency based on English Lexicon Project data (Balota et al., 2007), and were sorted into abstract and concrete conditions based on Brysbaert, Warriner, and Kuperman's (2014) concreteness norms (Table 1). Half of the words were enclosed in red boxes, and the other half in green, and this was balanced across concrete and abstract words. In the test phase, an additional 50 abstract and 50 concrete words were added to the target and non-target items.

Table 1
Stimulus Characteristics

	Targets			Synonyms		
	n_{letter}	log^F	conc	n_{letter}	log^F	conc
Abs	6.7	5.0	1.8	7.3	5.7	2.1
Conc	7.0	5.1	4.9	6.2	5.7	4.8

Procedure Participants performed a two-phase source memory task. Stimuli were presented visually one at a time, in pseudorandomized order, with an arbitrary box context (either a red or a green box). On each word, participants performed a synonym-judgment 1-back task. To ensure that they did not ignore the boxes, the hand they used to make their response was determined by box color (left hand for words in green boxes and right for red). Stimuli were presented for 2000 ms with a 1000-ms interstimulus interval. Participants were told there would be a later memory test on the words, but not that source (i.e., box color) memory would be tested.

In the test phase, participants performed two tasks for each word. First, they responded whether they had seen the word at encoding, indicating their degree of confidence in the decision (high, medium, and low confidence for either "old" or "new"). Second, for old words, they indicated the color of the box on initial encoding. The task was the same for new words, except that they were asked simply to select the color they thought the box would have been had it been presented at encoding. Participants were given 6000 ms each for the old/new and the box color judgment.

Data analysis Data were analyzed using R. Memory for items (i.e., words) and their contexts (i.e., box color) was first analyzed using descriptive statistics, calculating accuracy, hit rate, miss rate, correct rejections, false alarms, and d' (calculated as z(Hit) – z(FA)) for all words, and accuracy was also assessed by level of confidence. Source (i.e., box) memory accuracy was calculated only for target hits, and was assessed across confidence levels. Source memory accuracy was analyzed as a function of word type and confidence in having seen the word at encoding. Logistic mixed effects models (lme4 package; Bates et al., 2017) were used to analyze the data, with subject and word as random intercepts, and word type (abstract or concrete), level of confidence (low, medium, high), and their interaction as treatment-coded fixed effects. Each predictor was entered in a successive model, and statistical significance was assessed by comparing the models using likelihood ratio tests. Here, p-values < .05 were considered statistically significant.

Results

Item recognition First, to provide a baseline measure of memory for concrete and abstract words, we report the accuracy and hit, miss, correct rejection, and false alarm rates across all words (Table 2). Hit rates were higher and false alarms lower in concrete words, demonstrating the mirror effect (Glanzer & Adams, 1985), which has previously been observed for concreteness (Glanzer & Adams, 1990). For overall accuracy, there were main effects of both word type and confidence. Concrete words were better recognized than abstract ($\chi^2(1) = 10.21$, $p = .001$), and accuracy increased with greater confidence ($\chi^2(2) = 571.37$, $p < .001$). The interaction was non-significant. Among targets only (i.e., non-synonym words presented at encoding), there was no main effect of word type ($\chi^2(1) = 0.29$, $p = .59$), but a main effect of confidence level ($\chi^2(2) = 675.22$, $p < .001$), with words recognized better with higher confidence. The interaction was non-significant. Means and 95% CIs for word and source (i.e., box) memory are shown in Figure 1. Finally, d' analysis showed that when considering response sensitivity, accuracy was better for concrete concepts, $t(39) = -5.37$, $p < .001$.

Table 2
Mean Item Recognition Accuracy

Word type	Acc	Hit	Miss	CR	FA	d'
Abstract	.73	.77	.23	.65	.35	1.21
Concrete	.78	.81	.19	.71	.29	1.57

Note. CR = correct rejection; FA = false alarm.

Source memory Here, we included only trials for which the word had been correctly identified. There was a main effect of word, where the box was less likely to be remembered for abstract words ($\chi^2(1) = 5.45$, $p = .02$), but not of confidence level. The interaction was non-significant. Participants were less likely to correctly remember the box color for abstract words (Figure 1).

According to d' scores, there was a baseline advantage for recognizing concrete words, which would then bias the source memory models. Correct memory trials for abstract words may have been less likely to reflect true hits where the word was in fact encoded. Accordingly, we also constructed models with d' as a predictor. A likelihood ratio test comparing the model with both d' and word type versus the model with only d' was significant, $\chi^2(1) = 5.27$, $p = .02$, suggesting that the effect of word type, where box recognition was worse in abstract than it was in concrete concepts, was significant even after accounting for the d' concreteness advantage.

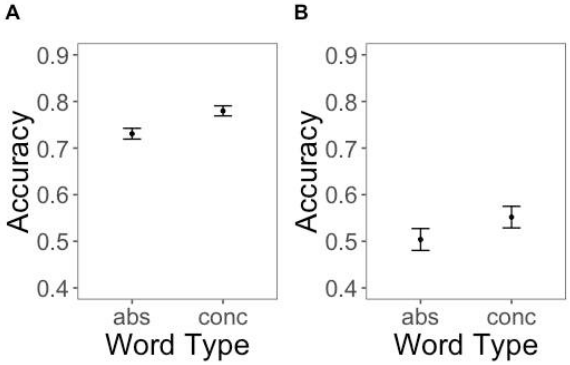


Figure 1. Effects of concreteness on (a) overall item recognition accuracy and (b) source (i.e., box) memory.

Discussion

Source memory was worse for abstract concepts, and this was true even after controlling for a concreteness advantage detected in d'. Thus, the results of Experiment 1 ran counter to our hypothesis: source memory was *worse* for abstract than for concrete words, even when participants were highly confident in having seen the word at encoding. Why did this unexpected difference emerge? It may be that concrete concepts are more amenable to a mnemonic strategy wherein a color adjective (i.e., "red" or "green") could readily be bounded to concrete objects (e.g., "table"), making source memory better for concrete words. Thus, it may be that contextual detail is better encoded in abstract concepts, but only when not systematically related to concrete objects (as may be the case for colored boxes). A second explanation is that counter to our main hypothesis, the concreteness advantage extends to memory for arbitrary contextual details. Experiment 2 evaluated these competing explanations.

Experiment 2

In Experiment 2, we utilized a variant of the source memory paradigm, where instead of the box, the context to be encoded was a male or female voice. Concepts were presented auditorily, and memory was assessed on visually presented words (e.g., Wilding & Rugg, 1996). In line with the original prediction that contextual detail is encoded to a greater extent in abstract concepts, it was predicted that source memory (i.e., male or female voice) would be better for abstract concepts. This prediction is further buoyed by the finding that person-related social properties may be more important for abstract concepts (Barsalou & Wiemer-Hastings, 2005).

Methods

Participants Forty-two UConn undergraduates with normal or corrected-to-normal vision who had not participated in Experiment 1 provided informed consent and were given course credit for their participation.

Stimuli The words were the same as those used in Experiment 1, but rather than being presented visually they were instead recorded by a male and a female speaker, with half the words presented by the male speaker and half by the female speaker. As with box color, this list was held constant across participants. There were no differences in the length of the sound files between the two speakers, and all files were normalized to a peak amplitude.

Procedure In the encoding phase, the procedure was the same as in Experiment 1. In the memory phase, the first judgment—whether the word was in the initial set (old) or not (new)—was the same. For the second judgment, participants were asked to indicate whether the person who said the word in the initial set was "Jane" or "Sid." The test phase was conducted with visually presented words, as in Experiment 1 (for a similar paradigm, see Wilding & Rugg, 1996).

Data analysis Data were analyzed in the same way as in Experiment 1.

Results

Item recognition Accuracy and hit, miss, correct rejection, and false alarm rates across all words are shown in Table 3. Among all words, there was a

significant main effect of both word type, with concrete words showing better recognition ($\chi^2(1)$ = 6.77, p = .009), and confidence level, with both medium and high showing greater accuracy than low confidence ($\chi^2(2)$ = 610.85, p < .001). The word type × confidence interaction was non-significant. Among targets, there was a main effect of confidence ($\chi^2(2)$ = 961.49, p < .001), but not of word type. The interaction was significant ($\chi^2(2)$ = 9.18, p = .01) at high confidence, suggesting that at greater memory strength, item recognition was worse for abstract words. Means and 95% CIs for the main effects of word type on word and source (i.e., voice) memory are visualized in Figure 2. Finally, d' analysis revealed that after considering response sensitivity, accuracy was better for concrete concepts, $t(40)$ = -3.49, p = .001.

Table 3
Mean Word Recognition Accuracy

Word type	Acc	Hit	Miss	CR	FA	d'
Abs	.70	.72	.28	.64	.36	1.04
Conc	.73	.77	.23	.67	.33	1.28

Note. CR = correct rejection; FA = false alarm.

Source memory Here, we again included only trials for which the word had been correctly recognized. There was a main effect of word type, with source memory for the voice context worse for abstract words ($\chi^2(1)$ = 5.70, p = .02), as well as a main effect of confidence ($\chi^2(2)$ = 25.22, p < .001). The interaction was non-significant. Thus, participants were again less likely to recognize the context correctly for abstract as compared to concrete words. Means and 95% CIs are shown in Figure 2.

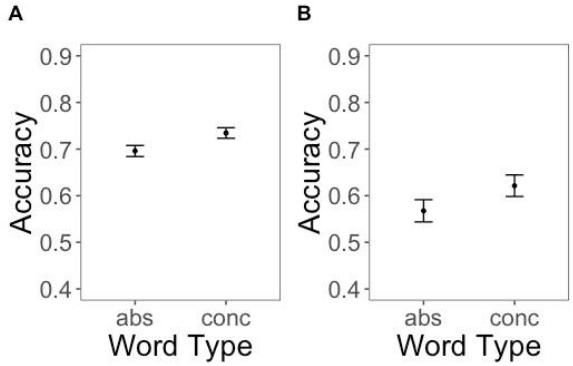


Figure 2. Effects of concreteness on (a) overall item recognition accuracy and (b) source (i.e., voice) memory.

As in Experiment 1, d' was greater for concrete than it was for abstract concepts (Table 3), and so we constructed models with d' as a predictor. A likelihood ratio test comparing the model with both d' and word type versus the model with only d' was significant, $\chi^2(1)$ = 5.75, p = .02, suggesting that the effect of word type, where source memory was worse for abstract than it was for concrete concepts, was significant even after accounting for the d' concreteness advantage.

Discussion

There was again a concrete word advantage in overall item recognition. Moreover, source memory for the voice context was worse for abstract concepts. This provides support for the interpretation that the concreteness advantage also extends to episodic memory, at least for memory for simple episodic detail. However, Experiments 1 and 2 showed a baseline memory advantage for concrete words, and thus they may have been more strongly encoded, and the strength with which the words were encoded, not concreteness, may have facilitated source memory. Accordingly, we conducted a third experiment.

Experiment 3

In Experiment 3, we simplified the memory phase by instead *only* probing recognition memory: half of the words were presented in the same box color as they were at encoding, while half of the words were presented in a different box color. The aim here was to investigate whether there is a selective advantage in recognition memory when the context is retained in abstract concepts—that is, is recognition memory facilitated to a greater extent in abstract concepts by context preservation? This would suggest that while the memory trace left by abstract concepts may be weaker, it can be strengthened when context is consistent across exposures. On the other hand, if recognition memory accuracy for abstract concepts is *worse* when the box color at encoding is preserved at test, it would suggest that arbitrary episodic detail may be *inhibited* in abstract concepts.

Methods

Participants Forty UConn undergraduates with normal or corrected-to-normal vision who had not participated in Experiment 1 or 2 provided written informed consent and received course credit.

Stimuli The stimuli were the same as those in Experiments 1 and 2, and box color assignment was counterbalanced across participants.

Procedure The encoding procedure was the same as in Experiment 1. At test, participants were asked to identify as many old words as possible, ignoring the color of the box. Words were presented in the red and green boxes. Half of the words retained the box color from encoding, and half changed color.

Data analysis Item recognition data were analyzed in the same way as in Experiments 1 and 2. However, box retention (old vs. new) was used as a second fixed effect in the mixed logit model, and the interaction was word type × box retention.

Results and discussion

Accuracy and hit, miss, correct rejection, and false alarm rates across all words are shown in Table 4. In overall old/new item recognition memory, there was a main effect of word type ($\chi^2(1) = 12.29$, $p < .001$), where memory was better for concrete words. Among targets only, there was no main effect of word type, nor was there a main effect of box retention. There was, however, an interaction between word type and box retention ($\chi^2(1) = 4.92$, $p = .03$; Figure 3). Accuracy was *worse* when the box color was retained in abstract concepts, again operating counter to the original hypothesis, and leading to the perhaps surprising conclusion that arbitrary episodic context may even be *suppressed* in abstract concepts.

Table 4
Mean Item Recognition Accuracy

Word type	Acc	Hit	Miss	CR	FA	d'
Abstract	.76	.76	.24	.78	.22	1.59
Concrete	.81	.80	.20	.84	.16	2.04

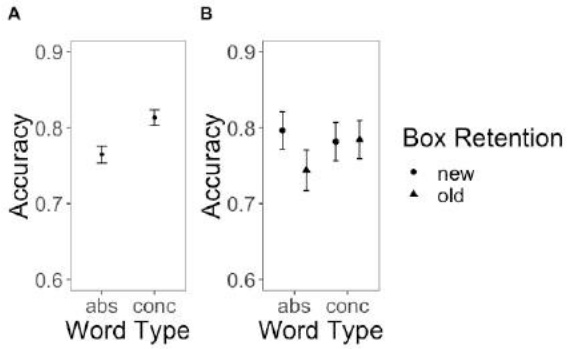


Figure 3. Plots showing (a) the main effect of concreteness on item recognition memory for all words and (b) the interaction between word type and box retention on target recognition memory accuracy (means and 95% CIs).

General Discussion

Abstract concepts are sensitive to context, but what is the mechanism by which this sensitivity emerges? The episodic memory system was identified as a potential candidate for encoding contextual information when processing abstract concepts. In Experiments 1 and 2, however, there was a concreteness advantage for recognizing episodic contexts. In Experiment 3, context preservation conferred a *disadvantage* for recognizing abstract concepts, suggesting the presence of a mechanism whereby arbitrary association are *inhibited* in the episodic experience(s) of the situations that activate abstract concepts.

In spite of these findings, across several literatures it is agreed that context is critical for understanding abstract concepts. However, there are differences across frameworks in terms of the *type* of context specified as being critical to processing abstract concepts, ranging from semantically constraining linguistic context in context-availability theory (Schwanenflugel & Shoben, 1983), to thematic associations in the qualitatively different representations framework (Crutch & Warrington, 2005), to meaningful situational and internal factors in grounded cognition (Barsalou & Wiemer-Hastings, 2005). While this study sought to uncover a basic mechanism that might unify these approaches (i.e., sensitivity to episodic information), the results unequivocally ran counter to our hypothesis: there is a concreteness advantage for encoding simple episodic detail.

Concreteness, context, and episodic memory

Concreteness is a powerful organizing factor in semantic memory (e.g., De Deyne, 2017; Hollis & Westbury, 2016), and concreteness effects are near ubiquitous in recognition memory studies. The present results suggest that such effects extend beyond stronger memory for concrete concepts to better associative. *relational memory* for concrete concepts, at least when the relation is a simple, arbitrary context. One important consideration here is the way in which we might expect context to be differentially recruited for processing concrete and abstract concepts, as this has implications for the relation between context sensitivity and concreteness.

In a review of the pervasiveness of context effects in cognition and perception, Yeh and Barsalou (2006) present two primary theses for how context affects concept processing: (1) contexts and concepts mutually activate each other, such that when processing a context, associated concepts are activated, and vice versa; and (2) when processing a concept in a particular context, properties of the concept which are relevant to that context become active. These two theses have different implications for the relation between context sensitivity and concreteness.

The first thesis resonates strongly with context availability theory, and likely suggests a concrete word advantage: concrete concepts activate contexts more strongly because they have stronger implicit ties to specific contexts. Thus, building implicit, direct associations between context and concept may have been facilitated by a similar mechanism to that which underpins context availability effects—if concrete concepts are typically associated with these sorts of contexts, then such contexts (such as boxes and

voices) might be more likely to be encoded with concrete concepts.

The second thesis may be more pertinent to abstract concept processing: when processing *decision* in the context of your choice of beverage at 9pm in the local café, the activated properties will be different from when processing decision in the context of a judge determining the appropriate sentence for a felon convicted of battery. That is, the schema-based knowledge necessary in these two situations differs considerably. *Decision* has a number of possible interpretations, and its precise meaning—and thus, the properties activated—depends on the situation and (systematically) associated schema-based knowledge. Research on the neural dynamics underpinning schema processing (e.g., van Kesteren et al., 2013) suggests that activating these systematic associations may in fact suppress the formation of associations with arbitrary elements of an episode. This dynamic is rooted in the interplay between neural systems in medial frontal and medial temporal lobe, where medial frontal activation when processing systematic associations may dampen activation of medial temporal lobe, thereby suppressing the formation of arbitrary bindings. Exploring these neural dynamics in this paradigm is an important direction for future work.

In summary, we contend that abstract concepts activate systematic—or *schema*-based—contextual information, and when processing *decision*, the activation of schema-based information may in fact *inhibit* formation of arbitrary associations. This would explain why our arbitrary episodic contexts were not well remembered for abstract concepts (Experiments 1 and 2) and why context retention may have even *inhibited* word recognition (Experiment 3). That is, abstract concepts may be particularly sensitive to *systematic* or *schema*-based contextual constraints, implicitly activating these associations when they are absent, and thus simultaneously inhibiting *arbitrary* contextual associations.

Limitations

The synonym judgment task used at encoding may have worked to a disadvantage: as abstract concepts tend to have more diverse meanings, synonym judgments may be more difficult for abstract concepts, as it must be determined whether *any particular sense* of the word is a synonym to the target (Hoffman et al., 2013). Thus, an abstract concept like *decision* when paired with *judgment* might leave fewer resources available to process immediately available relational information (i.e., in the present study, the box color or the voice) because we must search for a context in which *decision* and *judgment* are in fact synonyms (a recent computational model makes this prediction; Popov & Reder, 2018). Relatedly, if abstract concepts are simply more difficult to process, and the context does not help with accessing the meaning of the word, it could render the immediate context less salient. Thus, future research on context encoding in abstract and concrete concepts might benefit from departing from low-level episodic contexts. While we focused on arbitrary episodic detail, it might be fruitful to instead explore *systematic* contextual relations. For example, abstract concepts are thought to be represented in thematic or associative networks (e.g., *faith–church*), and so we might expect to see an abstract advantage in such contexts (for related evidence showing precisely this in relational vs. entity concepts, see Asmuth & Gentner, 2017). Finally, our finding that context reinstatement did not improve item recognition even for concrete words was perplexing. This may be because we only used two contexts—context-preservation advantages may not be observed when the context is shared across too many items (Park et al., 2006). That said, with just two contexts, reinstatement still *impaired* item recognition for abstract words, implying that a context preservation *disadvantage* can be detected with only two contexts. Nevertheless, further research is necessary to better understand the interaction between abstract concepts and arbitrary episodic contexts.

Conclusions

This research suggests that arbitrary episodic detail is better bound with concrete than abstract concepts. Abstract concepts rely on situational context for interpretation, and given that activation of situational information is known to inhibit formation of arbitrary associations (van Kesteren et al., 2013), formation of arbitrary associations may be *inhibited* in abstract concepts. More broadly, the way in which the episodic memory system is recruited appears to differ as a function of concreteness, suggesting that engagement of the episodic memory system is modulated by semantic content.

References

Asmuth, J., & Gentner, D. (2017). Relational categories are more mutable than entity categories. *The Quarterly Journal of Experimental Psychology*, *70*(10), 2007–2025.

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., … & Treiman, R. (2007). The English lexicon project. *Behavior Research Methods*, *39*(3), 445–459.

Barsalou, L. W., & Wiemer-Hastings, K. (2005). Situating abstract concepts. In D. Pecher & R. A. Zwaan (Eds.), *Grounding cognition: The role of perception and action in memory, language, and*

thought (pp. 129–163). Cambridge, UK: Cambridge University Press.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904–911.

Crutch, S. J., & Warrington, E. K. (2005). Abstract and concrete concepts have structurally different representational frameworks. *Brain*, *128*(3), 615–627.

Davachi, L. (2006). Item, context and relational episodic encoding in humans. *Current Opinion in Neurobiology*, *16*(6), 693–700.

De Deyne, S. (2017). Mapping the lexicon using large-scale empirical semantic networks. Talk presented at the Annual Meeting of the Psychonomic Society, Vancouver, BC.

Glanzer, M., & Adams, J. K. (1985). The mirror effect in recognition memory. *Memory & Cognition*, *13*(1), 8–20.

Glanzer, M., & Adams, J. K. (1990). The mirror effect in recognition memory: data and theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16(1), 5–16.

Hoffman, P., Lambon Ralph, M. A., & Rogers, T. T. (2013). Semantic diversity: A measure of semantic ambiguity based on variability in the contextual usage of words. *Behavior Research Methods*, *45*(3), 718–730.

Hollis, G., & Westbury, C. (2016). The principals of meaning: Extracting semantic dimensions from co-occurrence models of semantics. *Psychonomic Bulletin & Review*, *23*(6), 1744–1756.

Kirwan, C. B., Wixted, J. T., & Squire, L. R. (2008). Activity in the medial temporal lobe predicts memory strength, whereas activity in the prefrontal cortex predicts recollection. *Journal of Neuroscience*, *28*(42), 10541–10548.

Paivio, A., Walsh, M., & Bons, T. (1994). Concreteness effects on memory: When and why? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(5), 1196–1204.

Park, H., Arndt, J., & Reder, L. M. (2006). A contextual interference account of distinctiveness effects in recognition. *Memory & Cognition*, *34*(4), 743-751.

Popov, V., & Reder, L. (2018; preprint). Frequency effects on memory: A resource-limited theory.

Rugg, M. D., Vilberg, K. L., Mattson, J. T., Sarah, S. Y., Johnson, J. D., & Suzuki, M. (2012). Item memory, context memory and the hippocampus: fMRI evidence. *Neuropsychologia*, *50*(13), 3070–3079.

Schwanenflugel, P. J., & Shoben, E. J. (1983). Differential context effects in the comprehension of abstract and concrete verbal materials. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *9*(1), 82–102.

Tulving, E. (1983). Elements of episodic memory. Oxford, UK: Clarendon.

Tulving, E. (2002). Episodic memory: From mind to brain. *Annual Review of Psychology*, *53*(1), 1–25.

Wilding, E. L., & Rugg, M. D. (1996). An event-related potential study of recognition memory with and without retrieval of source. *Brain*, *119*(3), 889–905.

Yeh, W., & Barsalou, L. W. (2006). The situated nature of concepts. *The American Journal of Psychology*, *119*(3), 349–384.

Yonelinas, A. P. (2001). Components of episodic memory: The contribution of recollection and familiarity. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *356*(1413), 1363–1374.

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language*, *46*(3), 441–517.

Yu, S. S., Johnson, J. D., & Rugg, M. D. (2012). Hippocampal activity during recognition memory co-varies with the accuracy and confidence of source memory judgments. *Hippocampus*, *22*(6), 1429–1437.

Rapid learning of word meanings from distributional and morpho-syntactic cues

Margherita De Luca (margherita.deluca@uniroma1.it)
University of Rome "La Sapienza", Department of Document studies, Linguistics, Philology and Geography, Piazzale Aldo Moro, 5
Rome, 00185 Italy

Gary Lupyan (lupyan@wisc.edu)
University of Wisconsin-Madison, Department of Psychology, 1202 W. Johnson Street
Madison, WI 53706-1611 USA

Abstract

What does it take to learn a new word? Many of the words we learn, we have learned from language itself – by encountering them in various informative contexts. Here, we investigate the limits of learning from context by studying how people learn new words from very sparse contexts, at the extreme, a context in which all content words are replaced by nonsense words. We find that participants exposed to even such extremely *sparse* contexts nevertheless learn something about the meaning of words embedded in those contexts. Performance tended to be better when knowledge was assessed by first directing people's attention to the part of speech of the target words.

Keywords: language; word learning; distributional semantics; syntactic bootstrapping.

Introduction

How do we learn the meanings of words? One way is by associating words with external referents. This is of particular importance in early word-learning when children's vocabulary is dominated by concrete nouns ("cup", "nose"), action verbs ("jump", "bark") and words used by children as imperatives ("more", "up"). As we become more mature language users, our vocabulary expands to contain a large proportion of items that do not have concrete referents (e.g., "terrible", "hope" and "fun"). These words must be learned through language itself.[1]

One way to learn word meanings through language is through explicitly provided definitions. For example, on encountering the sentence "The Celtics Coach was livid over the call, hurling an expletive at the officials.", one might look up "livid" to learn that a common meaning is "furiously angry". But we can also learn something about the possible meaning of livid through its context. Indeed, school-age children learn, on average, between 600 and over 3,000 new words per year (Nagy & Herman, 1984; Nagy, Herman, & Anderson, 1985). That these words are learned via direct instruction or through dictionaries is highly unlikely (Nagy & Anderson, 1984).

In this work we investigate the limits of learning from linguistic context. We already know that school-age children and adults are adept at learning from informative contexts such as the "livid" example above. We were curious whether learning something about a word's meaning is possible even when contexts are highly impoverished. At the extreme, we investigate learning meanings of nonce words from contexts in which all content words are replaced with nonce terms. How can people learn word meanings from contexts that themselves contain only nonce words? One way is through the use of morpho-syntactic cues.

The idea that people can learn words through morpho-syntax is generally known as "syntactic bootstrapping". Preschool children can use both morphology and syntax to infer the meanings of novel words (Gleitman, 1990; A. E. Goldberg, 2003; Naigles & Swensen, 2007). In a classic demonstration, Brown (1957) showed 3-5 year old children images such as a pair of hands emptying an odd container of a novel slushy material. The children were then told that "in this picture you can see sibbing/a sib/some sib", and were asked to point to another instance of sibbing/a sib/some sib. The finding that children were more likely to point to a substance when presented with "some sib" and an object when presented with "a sib" suggests that they are learning (coarse) aspects of meaning from morpho-syntactic cues. In a more recent study, Naigles presented 2-year old children with the novel word "gorping" and two videos of novel actions, one causative and one non-causative. The sentence frame in which the verb was presented influenced which action children chose for the verb (Naigles, 1990). Although these experiments show that even young children can make use of morpho-syntactic cues to learn something about what the word means, they do not tell us about the limits of our ability to learn word meanings in this way. In these classic studies, the to-be-learned are marked, e.g., by being the only unfamiliar word in the utterance. The contexts that are used tend to be quite informative and the meanings being learned are constrained by the accompanying pictures or videos. Lastly, the assessment of what is being learned in such studies tend to be quite limited. The children in Brown's study could answer "correctly" simply by inferring that "sibbing" is an action, without learning anything more specific about its meaning.

But we know that people can do more. For example, even in a language with relatively simple morphology – English – a sentence like "The gostak distims the doshes" is surprisingly meaningful. We can infer that a "gostak" is doing something ("distimming") and it is doing it to the "doshes" (Ingraham, 1903). We may further infer that a gostak may be an animate

[1] Indeed, in the largest set of concreteness/abstractness norms (Brysbaert, Warriner, & Kuperman, 2014), words were defined as "abstract" if their meanings could not be communicated via direct reference or through action, but rather would have to be explained through language.

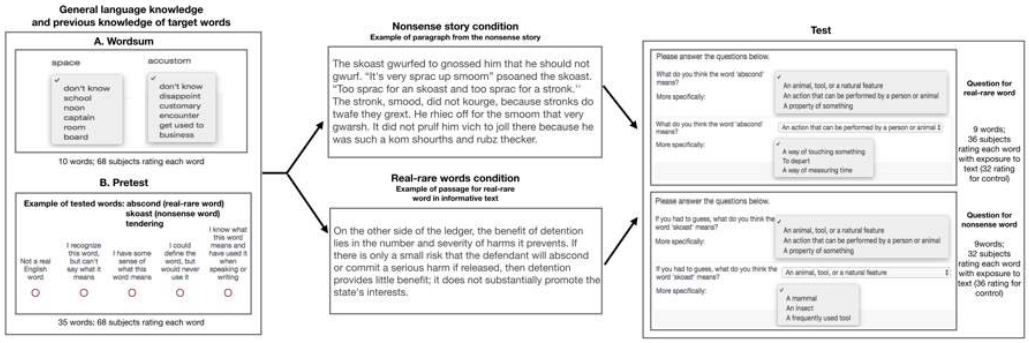


Figure 1: Schematic for Experiment 1.

agent and "distimming" is an action that the gostak is capable of performing. Remarkably, even without *any* referential context, exposure to sequences of such sentences appears to be sufficient for people to navigate an entire virtual world. Players of *The Gostak* (Muckenhoupt, 2001) quickly learn to *deave in the tavid dorl and gomb the stike*.

Another hint that language contains rich cues to word meanings comes from attempts to derive semantic representations from the ground up, by tracking word co-occurrences to learn that words occurring in similar contexts have similar meanings (Landauer & Dumais, 1997; Lany & Saffran, 2010; Levy & Goldberg, 2014; Mikolov, Chen, Corrado, & Dean, 2013; Redington, Chater, & Finch, 1998) – the so-called distributional hypothesis of word meaning (Firth, 1957). McDonald and Ramscar (2001) tested the distributional hypothesis by manipulating the linguistic context surrounding very rare or nonce words showing that judgments involving the target words are affected by the distributional properties of the manipulated linguistic environments.

Our main goal is to investigate the limits of people's ability to learn word meanings from linguistic context. We do this by exposing adult English speakers to contexts varying in informativeness ranging from fully informative contexts – passages of real English text containing real words unknown to most of our subjects, to highly sparse contexts that contain English morpho-syntactic cues (e.g., verb endings), but in which all content words have been replaced by nonsense words. We test people's knowledge of both part of speech information (the sole focus of much of the classic work on syntactic bootstrapping), *and* knowledge of more specific aspects of meaning.

Experiment 1

In experiment 1, we tested the role of the linguistic context in inferring word meanings for real, but rare English words (e.g., "kine") that were presented in informative contexts and for nonce words (e.g., "stronk") placed in highly *sparse* contexts which were stripped of almost all meaningful words.

Participants We recruited 114 participants from Amazon Mechanical Turk (52 Males of average age = 37; 62 Females of average age = 39). 68 of these participated in the main word-learning experiment (32 in the nonsense-story condition and 36 in the real-passages condition) and 46 participated in the salience-norming task.

Materials Participant were randomly assigned to one of two conditions: a nonsense-story condition and a real-passages condition. For the nonsense-story condition, all participants were exposed to the children's story *Why the cricket chirps*[2]. Participants were not provided with the story's title or any information about its content. Of the 604 total tokens in the story, 169 were open-class words. We replaced all the content words with nonce words taken from the ARC Nonword Database (Rastle, Harrington, & Coltheart, 2002). These words were created from orthotactically legal bigrams, onsets, and bodies. Fig. 1 shows a part of the resulting story. Of the 136 word types in the story that were replaced with nonce words, we selected 9 to serve as targets for later testing. Participants did not know ahead of time which words would be tested. The target words varied in frequency, occurring between 2 and 18 times, and parts of speech: 4 nouns, 3 verbs, 2 adjectives. Derivational and inflectional morphemes in the story were limited to a small number of cases (see Table 1 for morphological variation present for each target word).

For the real-passages condition, we matched each of the 9 target words in the nonsense-story condition with real, but rare English words that were unlikely to be familiar to our participants (e.g. "ratoon", "pronk, "rawky"; henceforth *real-rare* words). For each word, we selected 3-4 sentences in which the word appeared from the Corpus of Contemporary American English (COCA) and other online sources to serve as the context (see Fig. 1). Participants' word knowledge in all conditions was tested using "drill-down questions" designed to be sensitive to partial word knowledge. The first question provided three options for part of speech and the second question had participants choose between three word meanings all within the chosen part of speech: 1 fully correct, 1 partially correct, and 1 incorrect (see Fig. 1).

Quantifying word salience We expected that people's ability to infer word meanings would be influenced by the fre-

[2] https://www.freechildrenstories.com/why-the-cricket-chirps

quency with which the target word occurred in the passage. But we also suspected that aside from frequency, performance would also be related to a word's *salience* (C. M. Brown, 1993). There is no single definition of salience, but intuitively, a word is salient to the extent that it communicates the central point of a story. For example, words naming the actions performed by a central character are more salient than words describing aspects of the environment that a non-central character inhabits. We quantified the salience of each target word as the likelihood that it would be recalled by people who read the original (unaltered) story. We recruited 46 participants from MTurk to read the original story and then had 1 minute to recall all the words they could remember occurring in the story. Salience for each word was defined as the sum of the weights that the word obtained each time it was listed: the weight was calculated as exponentially decreasing in accordance with the order in which words were listed by participants [(for each time the word was listed) weight = (0.75^(word_n -1))] (see Table 1 for frequency and salience of each target word).

Table 1: Frequency, salience and morphological variation for target words

Target word	Frequency	Salience	Derivational Morphology	Inflectional Morphology
fly	18	6.59	1 derivational form (-er)	2 inflected forms (-ed; -ing)
cricket	16	31.69	none	1 inflected from (-s)
cold	7	5.69	none	none
wing	6	8.26	none	1 inflected from (-s)
fast	6	3.76	none	2 inflected forms (-er; -est)
chirp	5	8.49	none	3 inflected forms (-s; -ed; -ing)
ant	5	6.91	none	none
hop	5	1.06	none	none
snow	5	2.56	none	none
rub	4	1.90	none	2 inflected forms (-ed; -ing)
listen	4	1.47	none	none
warm	4	1.31	none	none
tree	3	0.44	none	1 inflected from (-s)
frozen	3	4.82	none	none
ground	2	0.00	none	none
owl	2	10.35	none	none

Procedure General procedure is shown in Figure 1. At the start of the task, participants completed a 10 item vocabulary test (Wordsum; Malhotra, Krosnick, & Haertel, 2007) and a pretest gauging familiarity with the target words. Participants were then randomly assigned to the real-passages or the nonsense-story condition. Those assigned to the real-passages condition saw each (meaningful) context and answered the two vocabulary questions for each of the 9 real-rare words (in random order). Each word was tested immediately after being presented in its context. The group was then tested on nonce words (skoast, etc.). In contrast, participants in the nonsense-story condition saw the entire 604-word nonsense-story and were then tested on the real-rare words, and then on 9 of the target nonce words (in random order). This design allowed each condition to serve as the control for the other condition. Subjects in the nonsense-story condition were asked to infer the meaning of the *real-rare* words without exposure to the passages and vice versa. As an attention check, scattered among nonce and *real-rare* words were questions about the meaning of familiar words (e.g. "little", "green").

Results and Discussion We analyzed the data using logistic mixed effects models. In the initial analyses, we treated partially and completely correct scores for specific meaning as the same (i.e., a binary contrast between an accuracy of 0 and 0.5/1). Figure 2 shows a clear interaction between condition (real-passages, nonsense story) and word-type (real-rare, nonce). This interaction was present both for part of speech [z = 7.3, SE = 0.28, $p < .001$] and word meanings [z = 9.1, SE = 0.29, $p < .001$] [accuracy ~ type_of_word*condition+(1|subj_id)]. Participants in the real-passages condition were significantly more accurate at selecting meanings corresponding to the correct part-of-speech of real words compared to participants in the nonsense-story condition (i.e., those not exposed to the real passages) [z = -3.7, SE = 0.39, $p < .001$]. They were also better at choosing the more correct specific meanings [z = 4.6, SE = 0.4, $p < .001$] [accuracy (for specific meaning or for part of speech) ~ condition+(1|subj_id)+(1|word)]. Note that above chance performance for the *real-rare* words is expected even without being exposed to real passages because some participants already know the meaning of these words. Not surprisingly, accuracy on *real-rare* words was positively associated with greater vocabulary as assessed by Wordsum. This was true both for part of speech [z = 2.8, SE = 0.17, p = .005] and specific meaning measures [z = 3.7, SE = 0.19, $p < .001$] [accuracy (for specific meaning or for part of speech) ~ condition*wordsum_score+(1|subj_id)+(1|word)]. Greater familiarity with the target *real-rare* words (pretest) was positively related to selecting the correct part of speech [z = 2.1, SE = 0.16, p = .033] and specific meaning [z = 2.9, SE = 0.16, p = .004].

The results for the real-rare words tell us what we already knew – people can infer word meanings from seeing the words in context. We now turn to the nonsense-story condition. Recall that these nonce words were seen in the context of a 600+ word story in which *all* content words were replaced by nonce words. Participants exposed to this extremely sparse context had significantly higher performance in inferring the correct part of speech for the nonce words [z = 3.9, SE = 0.23, $p < .001$] and in choosing the more correct meanings [z = 4.4, SE = 0.25, $p < .001$]. The benefit from reading the nonsense story was not limited to just helping people choose the correct part of speech. Restricting the analysis to only the trials on which participants chose the correct part of speech, we find that exposure to the nonsense-story still led to higher accuracy [z = 2.9, SE = 0.35, p = .004] [$\text{accuracy}_{\text{part\_of\_speech}}$ ~ condition+(1|subj_id)+(1|word)]. Neither Wordsum nor Pretest scores predicted performance for nonce words. Word frequency and word salience likewise did not predict performance ($z<1$, but see Exps. 2 and 3).

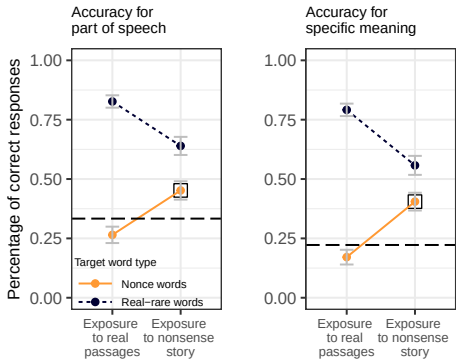


Figure 2: Accuracy for type of target word for Experiment 1. Horizontal dashed lines indicate chance-level. Error bars indicate +/-1SE of the mean. Significant effect of context exposure for nonce words in the nonsense-story is marked by a squared shape around the relevant data-point in the graph.

Experiment 2

Experiment 1 showed that people can benefit from very sparse contexts. One shortcoming of the study was that participants in the real-passages and nonsense-story conditions were exposed to contexts in a different way (short passages vs. one long story; tested after each word vs. tested on all words at once). In Experiment 2, we test participants in the same way on all the same words, varying just the informativeness of the context. We were also curious about whether it mattered how word knowledge was assessed. Instead of using the "drill-down" format that first asked about part of speech, we used a more standard multiple-choice text, presenting all the 9 options for each word at the same time. Lastly, we parametrically varied the informativeness of the context by replacing different proportions of content words with nonce words.

Methods Participants were randomly assigned to one of six conditions (see Table 2 for a summary). In the real-text condition we exposed participant to the original *Cricket* story, in which we replaced only the target words with *real-rare* English words. Thus, the linguistic context was still informative (i.e., the target words were surrounded by meaningful words) but it did not directly aim at communicating the meaning of the target words, e.g., *'You should find some shelter from the cold night,' said the smew. The mitius did not auscult, because mitiuses do whatever they want. He decided to rest on a pile of twigs.* We progressively decreased the information provided by the context by replacing various proportions of the remaining content words in the story with nonce words (40%, 60%, 90%, 100%). Lastly, we included a control group that was tested on their knowledge of the *real-rare* words without seeing any prior context.

Participants We recruited 246 participants from Amazon Mechanical Turk (112 Male of average age = 35; 132 Females of average age = 38). 38 participants were assigned to the real-text condition, 36 to the 40% condition, 39 to the 60% condition, 42 to the 90% condition, 52 to the 100% condition and 39 to the control condition.

Procedure Participants in each condition were initially tested on Wordsum and Pretest (following the procedure for Experiment 1) and then randomly assigned to one of the six conditions described above. All participants were then presented with the same 12 multiple choice questions (9 options per question) to assess their knowledge of the *real-rare* words.

Results and Discussion Results are shown in Fig. 3. Participants exposed to the full story clearly benefited in inferring both part of speech [z = 5.9, SE = 0.22, p $<$ .001] and specific meaning [z = 5.6, SE = 0.25, p $<$ .001] of *real-rare* words [accuracy˜overall ˜ control_vs_full_story +(1|subj_id)+(1|word)]. Similar results were found for the 40% and the 60% conditions, in which participants showed compelling effects for both part of speech [z = 4, SE = 0.22, p $<$ .001] and specific meaning [z = 3.8, SE = 0.25, p $<$ .001]. In contrast, when 90% or 100% of content words were replaced with nonce words, no significant benefit was observed for either part of speech [z = 1.2, SE = 0.2, p = .226] nor specific meaning [0.77, SE = 0.23, p = .444] results (Fig. 3).

Frequency of occurrence in the story was positively associated with accuracy for the real-text condition, [z = 2.7, SE = 0.16, p = .008] [accuracy ˜ condition*frequency+(1|subj_id)+(1|word)] and salience [z = 2.3, SE = 0.16, p = .023] [accuracy ˜ condition*salience+(1|subj_id)+(1|word)] in predicting accuracy for specific meaning. More frequent and more salient words benefited more from context. Similar effects were found in the 40% and 60% conditions for accuracy on specific meaning [frequency: z = 3.8, SE = 0.16, p $<$ .001; salience: z = 2.3, SE = 0.16, p = .023] [accuracy ˜ condition*wordsum_scores+(1|subj_id)+(1|word)]. Controlling for pretest scores, greater vocabulary knowledge (Wordsum) was positively associated with accuracy for both part of speech [z = 2.7, SE = 0.12, p = .006] and specific meaning [z = 2.7, SE = 0.14, p = .007]. Similarly, previous word knowledge was positively associated with part of speech accuracy [z =2.4, SE = 0.084, p = .018] and specific meaning accuracy [z = 2.9, SE = 0.086, p = .003] [accuracy ˜ condition*pretest_scores+(1|subj_id)+(1|word)]. These associations parallel the findings of the real_rare words condition of Experiment 1.

Exposure to a story in which 40%-60% of content words were replaced with nonce words still allowed participants to learn something about meanings of words occurring in the story. There were two noteworthy differences between the results of Experiment 1 and 2. First, unlike Experiment 1, participants' ability to benefit from the story context was positively associated with the frequency with which the word occurred in the story and the word's salience. These relationships may stem from people's greater baseline knowledge of the (real-rare) target words. Second, exposure to a story in

Table 2: Summary of type of context, type of target word and methods is assessing word meaning in each experiment.

Experiment	Condition Name	Type of Text	Type of Target Word	Example Words	Question Test Type
Experiment 1	Real-passages	Real-rare target words; informative text	real-rare words	ratoon; pronk; rawky	drill-down
	Nonsense-story	Nonce-target words; nonsense story	nonce words	stronk; sprac; crex	drill-down
Experiment 2	Real-text	Real-rare target words: real story	real-rare words	auscult; lollop; smuir	multiple choice
	Real-words-nonce-context: 40%	Real-rare target words; 40% of context replaced with nonce words	real-rare words	auscult; lollop; smuir	multiple choice
	Real-words-nonce-context: 60%	Real-rare target words; 60% of context replaced with nonce words	real-rare words	auscult; lollop; smuir	multiple choice
	Real-words-nonce-context: 90%	Real-rare target words; 90% of context replaced with nonce words	real-rare words	auscult; lollop; smuir	multiple choice
	Real-words-nonce-context: 100%	Real-rare target words; 100% of context replaced with nonce words	real-rare words	auscult; lollop; smuir	multiple choice
	Control group	No exposure to story	real-rare words	auscult; lollop; smuir	multiple choice
Experiment 3	Real-words-nonce-context: 100%	Real-rare target words; 100% of context replaced with nonce words	real-rare words	auscult; lollop; smuir	drill-down
	Control group	Nonsense story with all words replaced with nonce words	real-rare words	auscult; lollop; smuir	drill-down

which all the content words with nonce words did not lead to greater-than-baseline performance on the word test. Experiment 3 was designed to better understand this difference.

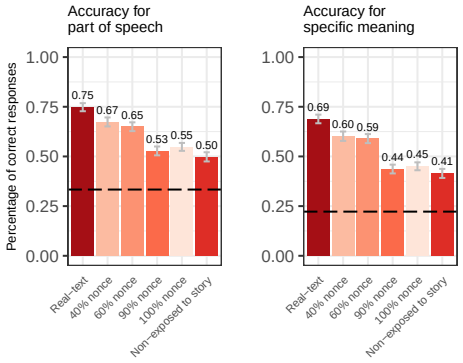


Figure 3: Group performance for Experiment 2. Horizontal dashed line indicated chance-level. Error bars indicate +/- 1SE of the mean.

Experiment 3

In Experiment 1, we found a significant effect of context exposure in inferring the meaning of nonce words from nonsense context. In Experiment 2 we assessed knowledge of *real-rare* words instead of nonce words, and embedded them in contexts of varying informativeness. Contexts in which 40-60% of content words replaced with nonce words were helpful, but those in which more (90%-100%) of content words were thus replaced, were not helpful. How do we reconcile this difference? Aside from testing nonce words vs. real-rare words, Experiments 1 and 2 differed in the way word knowledge was assessed. Experiment 1 first asked about part-of-speech. Experiment 2 presented all the meaning choices together, intermixing meanings from different parts of speech. We reasoned that explicitly asking people about parts of speech (which are more directly bootstrapped by morpho-syntactic cues) may make it easier for people to subsequently access more specific aspects of the word's meaning. In Experiment 3 we tested the effect of exposing people to a nonsense-story containing *real-rare* words as the *real-words-nonce-context (100%)* condition of Experiment 2, but using the drill-down question format of Experiment 1.

Methods Participants were randomly assigned to either the *real-words-nonce-context: 100%* or to a control condition. In the *real-words-nonce-context*, participants were exposed to a nonsense-story containing the *real-rare* target words and tested on those *real-rare* words (as in the 100% condition of Experiment 2). Participants in the control condition were shown a story with all nonce words (as in the nonsense-story condition of Experiment 1) but at test were asked about the meaning of the *real-rare* target words of the *real-words-nonce-context* condition (i.e., they were asked about words they did not see in the story).

Participants We recruited 81 participants from Amazon Mechanical Turk (39 Male of average age = 37; 41 Females of average age = 37). 41 to the *real-words-nonce-context: 100%* story condition and 40 to the control condition.

Procedure Participants in each condition were initially tested on Wordsum and Pretest (following the procedure for Experiment 1) and then randomly assigned to either the *nonsense-story* condition or to a control group that was not exposed to a story. All participants were tested on the same set of drill-down questions.

Results and Discussion Results are shown in Fig. 4. We found a significant effect of exposure to the linguistic context in inferring the part of speech [z = 3, SE = 0.18, p = .002] and the specific meaning [z = 2.6, SE = 0.2, p = .008] of *real-rare* words when compared with the control condition [$\text{accuracy}_{\text{overall}}$ ~ condition+(1|subj_id)] (Fig. 4). When examining only trials on which participants inferred the correct part of speech, the benefit of exposure to a nonce story on inferring the correct specific meaning was no longer significant [z = 0.5, SE = 0.31, p = .614] [$\text{accuracy}_{\text{part\_of\_speech}}$ ~ condition+(1|subj_id)+(1|word)]. To determine if the nonce-word context in the present study was more effective than in the equivalent condition of Experiment 2, we examined the interaction between experiment (*Exp. 2* vs. *Exp. 3*) and condition (*control group* vs. *real-words-nonce-context: 100%*) [accuracy ~ condition*experiment+(1|subj_id)+(1|word)]. This interaction was significant for both part-of-speech [z = 2.9, SE = 0.18, p = .003] and specific meaning [z = 2.5, SE = 0.21, p = .003]. As in Experiment 2, we found a significant effect of both frequency and salience of the target words. The benefit of being exposed to the nonce story was greater for more frequent words [for part of speech: z = 2.1, SE = 0.14, p = .035; for specific meaning: z = 2.5, SE = 0.15, p = .012] and more salient words [for part of speech: z = 2.2, SE = 0.14, p = .031; for specific meaning: z = 2.6, SE = 0.16, p = .009].

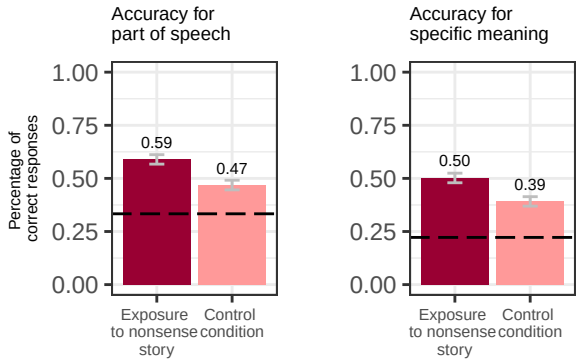


Figure 4: Group performance for Experiment 3. Horizontal dashed lines indicate chance-level. Error bars indicate +/-1SE of the mean.

Controlling for pretest scores, greater vocabulary knowledge (Wordsum) predicted higher accuracy for both part of speech [z = 3.4, SE = 0.091, p = .001] and specific meaning [z = 3.1, SE = 0.1, p = .002].

In Experiment 3, we observed an effect of context similar to that observed in Experiment 1. People were able to learn something about unknown words (here, *real-rare* words) from contexts in which all content words were replaced by nonce words. The only difference between Experiment 2 and the present study was in how word knowledge was assessed. A plausible conclusion – though in need of further testing – is that explicitly asking participants about a word's part of speech helps them deploy a more informative prior within which to consider the specific meanings.

General Discussion

What are the limits of learning word meanings from language? Our results show that participants were able to learn something about what a word means from brief exposures to such seemingly meaningless contexts as "The stronk rourthed daft to a dweave luk as the slom zeuded rhiecing." For example, after reading a 600+ word nonsense-word story containing sentences like the one just used, 38% of participants correctly inferred that "stronk" means "an insect" compared to 0% of participants who were not exposed to the story.

Our attempt to replicate and extend the results of Experiment 1 to using sparse contexts to inform the meanings of real, but rare and generally not known words (e.g., auscult, mitius), revealed that while partially informative contexts (40-60% of content words replaced with nonce words) were helpful, more sparse contexts (90%-100% of content words replaced) were not. We hypothesized that a key difference was the way that word knowledge was tested. In Experiment 1, participants' word knowledge was tested using a drill-down format that asked participants to first consider the word's part of speech. In Experiment 2, participants were asked to choose from among all the nine options visible at the same time making part of speech a less salient dimension of the word's meaning in the testing phase. In Experiment 3 we used the methods and materials of Experiment 2 with the test format of Experiment 1. Highlighting part-of-speech information using "drill-down" questions once again revealed that participants were able to use the nonce-story context to infer meanings of novel words.

Experiments 2-3 further showed that the effects of context were positively associated with frequency and salience. Words that were more frequent and more salient benefited more from context. While frequency is perhaps the most often used predictor in studies of word learning, to our knowledge, we are the first to examine the role of *salience*, defined here a the likelihood that people recall reading the word (see Table 1). What precisely makes a word salient requires future research.

What information did people use to infer word meanings? In the all-nonce-word conditions of Experiments 1 and 3, greater than baseline performance cannot be explained by reliance on the meaning of English content words because no recognizable content words were present. There were three remaining sources of information: closed-class words, inflectional cues, and syntactic cues. Consider one sentence from the story: "He thecked up into a dweave luk to fruth in for a sparf snurv." The remaining pronouns and prepositions combined with inflectional cues can clearly be used to infer that, e.g., "thecked" is an action being performed by an animate agent and that a "dweave" is likely to be some kind of place. Implicit knowledge of English syntax such that verbs follow "to" and objects tend to come after verbs offers further guidance. What is remarkable is that participants are making these inferences in parallel across dozens or even hundreds of words and that a single exposure to the story is sufficient to achieve above baseline accuracies.

Our work has two main limitations. First, successful use of sparse contexts involving mostly or exclusively nonce words clearly requires participants to already have sophisticated knowledge of English and so while it can help us understand how adults learn new words from context, it does not tell us how people learn enough English to make use of such sparse contexts. Second, present experiments do not tell us the relative importance of closed-class words, syntax, and morphology. Answering this question would require manipulating these sources of information independently. We can also gain additional insights by conducting studies such as this in more morphologically rich languages.

Conclusion

As has been long known, people are able to learn something about a word's meaning from encountering it in context. What is surprising is just how sparse and seemingly uninformative that context can be. The facility that people show in inferring word meanings from such sparse and seemingly meaningless contexts suggests that we may be underestimating the role that morpho-syntactic and distributional cues have on both learning word meaning and on acquiring semantic knowledge that is embedded in language.

Acknowledgements

This work was partially supported by NSF-PAC 1734260 to GL.

References

Brown, C. M. (1993). Factors affecting the acquisition of vocabulary: Frequency and saliency of words. In T. Huckin, M. Haynes, & J. Coady (Eds.), *Second language reading and vocabulary learning* (pp. 263–286). Norwood, NJ: Ablex.

Brown, R. W. (1957). Linguistic determinism and the part of speech. *The Journal of Abnormal and Social Psychology*, *55*(1), 1–5.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904–911.

Firth, J. R. (1957). A synopsis in linguistic theory, 1930-1955. In.

Gleitman, L. (1990). The Structural Sources of Verb Meanings. *Language Acquisition*, *1*(1), 3–55.

Goldberg, A. E. (2003). Constructions: A new theoretical approach to language. *Trends in Cognitive Sciences*, *7*(5), 219–224.

Ingraham, A. (1903). *Swain school lectures*. Open Court Publishing Company.

Landauer, T. K., & Dumais, S. T. (1997). A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. *Psychological Review*, *104*(2), 30.

Lany, J., & Saffran, J. R. (2010). From Statistics to Meaning: Infants' Acquisition of Lexical Categories. *Psychological Science*, *21*(2), 284–291.

Levy, O., & Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. In *Advances in neural information processing systems* (pp. 2177–2185).

Malhotra, N., Krosnick, J. A., & Haertel, E. (2007). The psychometric properties of the gss wordsum vocabulary test. *GSS Methodological Report*, *11*.

McDonald, S. A., & Ramscar, M. (2001). Testing the Distributional Hypothesis: The Influence of Context on Judgements of Semantic Similarity. *Proceedings of the Annual Meeting of the Cognitive Science Society, 23(23)*, 7.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space.

Muckenhoupt, C. (2001). The Gostak.

Nagy, W. E., & Anderson, R. C. (1984). How Many Words Are There in Printed School English? *Reading Research Quarterly*, *19*(3), 304.

Nagy, W. E., & Herman, P. A. (1984). Limitations of vocabulary instruction (Tech. Rep. No. 326). *Urbana: University of Illinois, Center for the Study of Reading*, 30.

Nagy, W. E., Herman, P. A., & Anderson, R. C. (1985). Learning Words from Context. *Reading Research Quarterly*, *20*(2), 233.

Naigles, L. R. (1990). Children use syntax to learn verb meanings. *Journal of Child Language*, *17*(02), 357.

Naigles, L. R., & Swensen, L. D. (2007). Syntactic Supports for Word Learning. In E. Hoff & M. Shatz (Eds.), *Blackwell Handbook of Language Development* (pp. 212–231). Oxford, UK: Blackwell Publishing Ltd.

Rastle, K., Harrington, J., & Coltheart, M. (2002). 358,534 nonwords: The ARC Nonword Database. *The Quarterly Journal of Experimental Psychology Section A*, *55*(4), 1339–1362.

Redington, M., Chater, N., & Finch, S. (1998). Distributional Information: A Powerful Cue for Acquiring Syntactic Categories. *Cognitive Science*, *22*(4), 425–469.

Eye Blink Rate Predicts and Dissociates the Effective Execution of Early and Late Stage Creative Idea Generation

Alwin de Rooij (alwinderooij@tilburguniversity.edu), Ruben D. Vromans, & Matthijs Dekker
Department of Communication and Cognition, Tilburg University,
Warandelaan 2, 5037 AB, Tilburg, the Netherlands

Abstract

In the present study, the correlations of eye blink rate (EBR) with the effective execution of early and late creative idea generation were explored. Participants engaged in a real-world idea generation task. Resting state EBR (before the task) and task-evoked EBR (during the task) were measured using eye-tracking. The results showed that resting state EBR negatively correlated with the amount of generated ideas during early stage, but not late stage idea generation. Task-evoked EBR did not correlate with the amount of generated ideas during early nor late stage idea generation. However, the change in EBR (from resting state to during early or late stage idea generation) positively correlated with the amount of ideas generated during early, but not during late stage idea generation. The contribution of this study is that it shows that EBR predicts and dissociates the effective execution of early and late stage creative idea generation.

Keywords: Creativity; Eye Blink Rate; Idea Generation.

Introduction

Eye behaviours such as fixations, eye blink rate (EBR), and pupil size are increasingly used to study creativity (See Salvi & Bowden, 2016 for a review) – the creation of ideas, solutions, or products that are both original and appropriate (Abraham, 2018). One important result of such studies is that eye blink rate, the average number of blinks per minute (de Rooij & Vromans, 2018), predicts and dissociates performance on different types of psychometric tests of creative potential (e.g., Akbari Chermahini & Hommel, 2010). Moreover, EBR has been used as a proxy for measuring fronto-striatal dopamine (Jongkees & Colzato, 2016), cognitive control (Akbari Chermahini & Hommel, 2010; 2012), motivation and affect (de Rooij & Vromans, 2018), and internal cognition (Salvi et al., 2015; Walcher, Körner, & Benedek, 2017). Studies of EBR and creative potential therefore inform theory about the involvement of these neuro-psychological factors in creativity. Psychometric tests of creative potential, however, often suffer from poor ecological validity, casting doubt over their explanatory power for actual real-world creative idea generation (Zeng, Proctor, & Salvendy, 2011). The present study therefore explores the correlations of EBR with the effective execution of the creative idea generation process, using a task that resembles real-world creative tasks more closely than psychometric tests of creative potential.

To enable creativity, people execute a creative process, which entails the execution of a set of cognitive processes and actions that enable a person to understand the problem that needs to be solved, generate ideas, and plan for further action (see Lubart, 2001 for a review). Idea generation is characterized by moving back and forth between generation and evaluation and is executed iteratively (Isaksen, Dorval, & Treffinger, 2010). In *early stages of idea generation* people typically retrieve concepts, which are synthesized into loosely formulated ideas, which process can involve remote association, conceptual combination, idea transformation, and analogical transfer (Finke, Ward, & Smith, 1992). The idea generation process evolves recursively, guided by the evaluation and selection of ideas for further development. Over iterations, and thus in *late stages of idea generation*, initially loosely formulated ideas are developed into more elaborately formulated ideas (Finke et al., 1992), and which process can be extended with combining previous ideas, filling in missing details, and simulating and testing implications and the validity of the ideas (Isaksen et al., 2010).

Previous research suggested that EBR predicts and dissociates performance on different types of psychometric tests of *creative potential*. *Resting state EBR* (i.e., EBR measured while a person is relaxed and not engaged in a thinking task) predicted the amount of different concepts (flexibility) used during the alternative uses task (AUT) (Akbari Chermahini & Hommel, 2010), a test where people are asked to list as many creative uses for a common object as they can (e.g., presented stimulus: *"Brick"*, possible response: *"Paper weight"*) (Guilford, 1957). This relationship was best described with a quadratic (inverted U-shaped) function. In the studies by Akbari Chermahini and Hommel no correlations, linear or otherwise, were found between resting state EBR and the amount of listed uses (fluency) or the statistical infrequency of the listed uses (originality). The results of a study by Ueda and colleagues, however, suggested that resting state EBR predicted the amount of listed uses during the AUT, which was also best described by a curvilinear (inverted U-shape) function (Ueda, Tominaga, Kajimura, & Nomura, 2016). Moreover, resting state EBR negatively correlated with the amount of correctly solved items during the remote associates task (RAT) (Akbari Chermahini & Hommel, 2010; Ueda et al., 2016), a test where people are asked multiple times to find the word that forms a compound word with each of the three given words (e.g., presented stimulus: *"Fox, Man, Peep"*, correct response: *"Hole"*) (Mednick & Mednick, 1971). In addition, Ueda and colleagues found that resting state EBR positively correlated with reaction time during the RAT.

Previous studies also suggested that *task-evoked EBR* (i.e., EBR measured while actively engaged in a task), predicts and dissociates performance on psychometric tests

of creative potential. That is, task-evoked EBR positively correlated with the amount of uses listed during the AUT (Ueda et al., 2016). In the same study, *task-evoked EBR* did not significantly correlate with the amount of correctly solved items during the RAT, but did positively correlate with reaction time during the RAT.

Studies on the *change from resting state to task-evoked EBR* add to these findings. That is, a study by Akbari & Hommel (2012) showed that the effects of stimulus induced increases in EBR on the amount of concepts used during the AUT differed significantly between people with low and high resting state EBR. That is, stimulus-induced increases in EBR led people with low resting state EBR to use more diverse concepts during the AUT than people with high resting state EBR. However, de Rooij & Vromans (2018) found no correlation or curvilinear relationship between the changes in EBR and the amount of uses, the amount of different concepts used, or the statistical infrequency of the responses during the AUT. Contrastingly, the same study showed that the change in EBR negatively correlated with the amount of correct responses to the RAT. However, when individual differences in positive and negative affect were taken into account, the interaction between a disposition to experience anxiety during creative tasks and the change in EBR positively correlated with the amount of correct responses to the RAT.

The main limitation of the currently available research though, is that psychometric tests of creative potential, such as the AUT and RAT, suffer from poor ecological validity (Zeng et al., 2011). Tests such as the AUT, for example, rarely correlate stronger than .30 with questionnaires and with performance on creative tasks with high ecological validity. Specifically relevant for creative idea generation, is that there is no clear necessity for iteration in such tests (Zeng et al., 2011), which is an essential aspects of idea generation process execution, that leads to differences in performance during early and late stage creative idea generation (Lubart, 2001). It is therefore not known if and how the processes that underlie performance during the AUT and RAT, are also involved in early and late stage idea generation. Moreover, the AUT and RAT are rather abstract tasks and lack goals with personal relevance that typically characterize real-world creative tasks (Kilgour, 2006). This ignores the essential role of domain-specific knowledge, and is likely to engage motivation differently than in real-world creative idea generation tasks (e.g., de Rooij & Jones, 2013). Thus, EBR may correlate differently, if at all, with performance during early and late stage idea generation in tasks that resemble real-world creative tasks more closely, than with performance during the AUT and RAT.

What is clear from these psychometric tests of creative potential, is that there is no indication of a correlation between EBR and qualitative aspects of idea generation (Akbari Chermahini & Hommel, 2010; 2012; de Rooij & Vromans, 2018; Ueda et al., 2016). That is, none of the studies showed correlations between EBR and the originality of the responses during these tests. Rather, results of these studies showed correlations between EBR and the quantity of responses (e.g., the amount of ideas during the AUT, the amount of solved items during the RAT). These studies therefore contribute that the correlations, if any, between EBR and performance during early and late stage idea generation is likely quantitative, and thus indicative of effective execution of the idea generation process, rather than directly of creativity.

Therefore, in *the present study* the correlations of EBR with the effective execution of the idea generation process (as measured by the amount of generated ideas) during early and late idea generation were explored, using a task that resembles real-world creative tasks more closely than psychometric tests of creative potential.

Method

To explore the correlation of EBR with the amount of ideas generated during early and late stage idea generation, an experiment was conducted.[1]

Participants

Seventy-eight people participated in this study (M_{age} = 23.34, SD_{age} = 3.46, 55 female, 23 male). They had normal or corrected-to-normal vision. Most (n = 76) were recruited via the participant recruitment system of a communication science department at a Dutch university. Participants received course credit as compensation for their time spent on the study. Two additional participants, recent graduates, requested to participate out of interest and did not receive compensation. On average, the participants self-reported to be moderately experienced with marketing (M = 3.79, SD = 1.11) (1 = *No experience*, 5 = *Very experienced*).

Idea generation task

The participants engaged in an idea generation task, were they were asked to generate creative marketing ideas aimed at helping a web shop that sells bicycles to attract more visitors to their website. Their idea generation process was split up into two separate tasks, both of which each participant completed, to capture early and late stage idea generation.

Task 1: Early stage idea generation

To capture early stage idea generation, participants were asked to generate as many creative marketing ideas as they could (Figure 1b). This, to elicit a range of pre-inventive structures that participants could then combine and elaborate upon in the subsequent late stage idea generation task. To enable the measurement of EBR, the early stage idea generation task was cued. Each trial started with a fixation dot (5 seconds), after which participants were asked

[1] Note that the data for this paper was collected as part of a larger experiment on eye behaviour and idea generation. Although the EBR data is used only in the study presented here, the participants, tasks, and procedure is the same as in other studies based on the same data set.

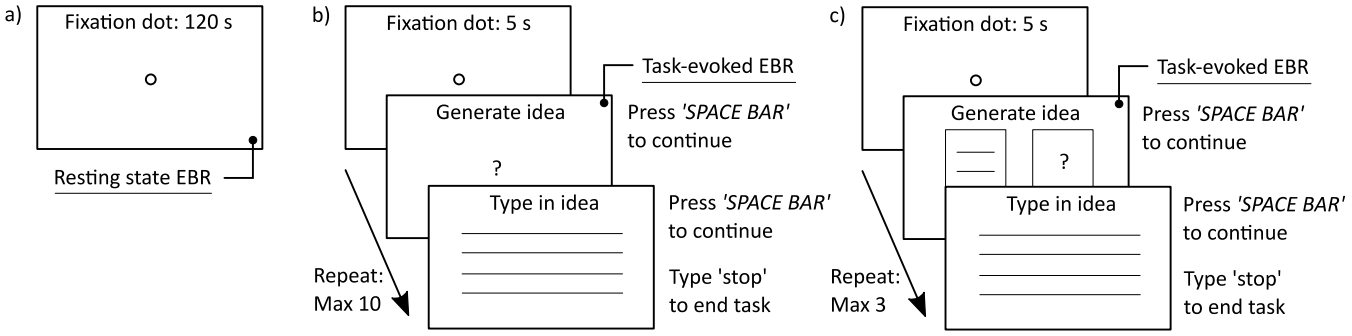


Figure 1: Trial structure of the a) resting state, b) early stage, and c) late stage idea generation tasks, including measurement points for resting state and task-evoked EBR.

to generate a creative marketing idea that was relevant to the provided problem description. There was no time limit. When a participant had generated an idea, the space bar could be pressed after which a text input field was presented on the screen where the participant could type in the idea they generated. This trial sequence was repeated a maximum of ten times. If the participants believed they could not generate any more ideas before the limit of ten trials was reached, they could type in 'stop' to end the early stage idea generation task, and start the late stage idea generation task.

Task 2: Late stage idea generation

To capture late stage idea generation, participants were asked to select two or more of their previously generated ideas to develop a more elaborate and detailed idea (Figure 1c). For example, if a participant generated the ideas to use an *"Instagram page"* and *"hire influencers to promote your Instagram posts"* during early stage idea generation, these could then be combined and developed into a more detailed elaborate solution, (e.g., *"where content developed for the Instagram page is suitable for hired influencers, with a follower demographic suitable for the web shop, which they can then share with their followers"*). Their previously generated ideas were available to the participants during this task (they were listed on the computer screen). The same trial structure was used as during the early stage idea generation task. That is, participants were instructed to look at a fixation dot for 5 seconds, after which they had time to combine previously generated ideas into more elaborate ideas. After generating an idea, they pressed the space bar on the keyboard, and a text input field emerged where they could type in their idea. There was no time limit. However, there was a maximum of 3 trials. If they believed that they could not generate any more ideas before they reached the limit of three trials, they typed in "stop" to stop the task, and with that end the experiment.

Assessment of the effective execution of the idea generation process

To gain insight into how effective the idea generation process was executed idea generation *fluency* was assessed (i.e., the amount of ideas generated). Fluency is a commonly used performance indicator used in studies of idea generation (Guilford, 1957). In the present study, participants generated on average 6.25 ideas during early stage idea generation (*SD* = 2.33), and on average 2.28 ideas during late stage idea generation (*SD* = .78).

Eye blink rate

Eye blinks were recorded with a head mounted eye-tracker, and were defined as eye-tracker signal loss with a duration of 40-400 milliseconds (de Rooij & Vromans, 2018). EBR was defined as the average amount of blinks per minute, and was calculated based on the amount of recorded eye blinks and the amount time during which these were recorded. The following measurements of EBR were used: (I) *Resting state EBR* - EBR recorded in resting state before the creative idea generation task, where participants were asked to relax and watch a fixation dot for 120 seconds (Figure 1a); (II) *Task-evoked EBR* - EBR recorded during early stage and during late stage idea generation. EBR was recorded only in the parts of the trials where participants were thinking about their ideas (Figure 1b and 1c); (III) *Change in EBR (task-evoked – resting state EBR)* - The change in EBR from resting state to early stage and to late stage idea generation.

To reduce measurement error, only data from eye blinks recordings after 2 seconds of the start, and before 2 seconds of the end of each measurement, were used to calculate EBR. This helped prevent blinks due to changing screens at the start of a task, and pressing ENTER when an idea was generated, to confound the EBR measurements (de Rooij & Vromans, 2018). Three participants did not blink during resting state. This may indicate that participants simply did not blink for 120 seconds, but may also indicate measurement error. Since the latter cannot be ruled out resting state EBR of these three participants was not used in the analysis. Finally, as EBR is only stable in the morning, midday, and afternoon (Barbato et al., 2000), the experiment was organised only between 9 am to 5 pm.

Apparatus

Materials were presented using dark letters against a grey background on a 22" Dell P2210 monitor (1680×1050 resolution). EBR was recorded using the EyeLink II head-mounted eye-tracker (SR Research Ltd.) at 250Hz. The cable that connected the eye-tracker to the computer was attached to the ceiling to reduce perceived weight and pull that may negatively affect comfort. LED lighting was used to diffuse environmental lighting as evenly as possible. The experiment was in OpenSesame with the PyGaze library (Dalmaijer et al. 2014).

Procedure

Participants received a written introduction to the experiment, signed informed consent, and filled in a short questionnaire about their socio-demographics and marketing experience. Information about the true purpose of the experiment was withheld at this stage. Participants were seated behind a computer screen in a sound proof booth. The head-mounted eye-tracker was installed and calibrated using a 5-point validation. The distance to the screen was approximately 70 cm. Then, participants could practice with the experiment software. After this, participants were asked to relax and look at a fixation dot for 120 seconds. Next, participants read the provided problem statement, and started with the idea generation task. Finally, the participants were debriefed in full, and after being asked whether they could guess the purpose of the experiment.

Analysis

The data obtained in the present study were analysed using generalized linear mixed models. The models were calculated using Satterthwaite approximation to account for the relatively small sample size. Robust covariances were used for the tested of fixed coefficients to handle violations of model assumptions. For models with the amount of generated ideas as the target, a negative binomial distribution was used with a log link. For the model with EBR as the target, a normal distribution with an identity link was used. Model terms and targets are presented in Table 2.

Results

Table 1: Descriptive statistics of EBR during resting state, early stage (task evoked), and late stage (task evoked) idea generation.

	EBR		
	M	*SE*	*n*
Resting state	13.23	1.21	75
Early stage	7.18	.60	78
Late stage	3.91	.40	78

Note. *M* = mean, *SE* = standard error, *n* = count.

[2] Quadratic models were also tested by adding the squared EBR terms to the models presented in Table 2. No significant coefficients were found that add to the results obtained with the linear models. We also refer to Figures 2b-2d for visual inspection.

The descriptive statistics are presented in Table 1. The results showed a significant main difference between the tasks for EBR, $F(2, 228) = 32.07$, $p < .001$ (Figure 2a).[2] The pairwise comparisons (not corrected) showed a significant difference in EBR between resting state and early stage, estimated difference = -6.06, $t = 4.48$, $p < .001$, 95% CI[-8.72, -3.39], and late stage idea generation, estimated difference = -9.32, $t = 7.30$, $p < .001$, 95% CI[-11.84, -6.81]; and between early and late stage idea generation, estimated difference = -3.27, $t = 4.52$, $p < .001$, 95% CI[-4.69, -1.84]. These findings suggest that in the present study, EBR decreased from resting state, to early stage idea generation, to late stage idea generation.

Table 2: Correlations and effects (GLMM) of resting state EBR, task-evoked EBR, and their difference with fluency during late and early stage idea generation.

Model terms	Correlations of Fluency with EBR		
	Resting state	Task-evoked	Change EBR
Intercept	.83** (.06)	.86** (.08)	.81** (.06)
Early stage	1.12** (.08)	.92** (.10)	1.08**(.06)
Late stage	.[a]	.[a]	.[a]
EBR	>-.01 (<.01)	-.01 (.02)	>-.01 (<.01)
Early stage x EBR	-.01* (<.01)	.02 (.02)	.01 (<.01)**
Late stage x EBR	.[a]	.[a]	.[a]

Note. Data are unstandardized coefficients and standard errors (between parentheses). [a] Reference variable. * p < .05, ** p < .01.

The results showed a significant and negative interaction between idea generation stage and resting state EBR for the overall amount of generated ideas, $b = -.01$ $t = 2.00$, $p = .049$, 95% CI[-.02, -.01] (Table 2). Pearson correlations showed that this interaction effect could be explained by a significant and negative correlation between resting state EBR and the amount of generated ideas during early stage idea generation, $r = -.170$, $p = .043$, and a negative but not significant correlation between resting state EBR and the amount of generated ideas during late stage idea generation, $r = -.039$, $p = .675$ (Figure 2c). These findings indicate that resting state EBR negatively correlates with the effective execution of early but not late stage idea generation.

The results showed no significant correlations between task-evoked EBR and the amount of generated ideas; and no significant interaction between idea generation stage and task-evoked EBR for the amount of generated ideas (Table 2, Figure 2b). These findings indicate no relationship between task-evoked EBR and the effective execution of the creative idea generation process.

Furthermore, the results showed a significant and positive interaction between idea generation stage and the change in EBR from resting state to each task, for the overall amount of generated ideas, $b = .01$, $t = 3.04$, $p = .003$, 95% CI[.01, .02] (Table 2). Pearson correlations showed that this interaction effect could be explained by a significant and

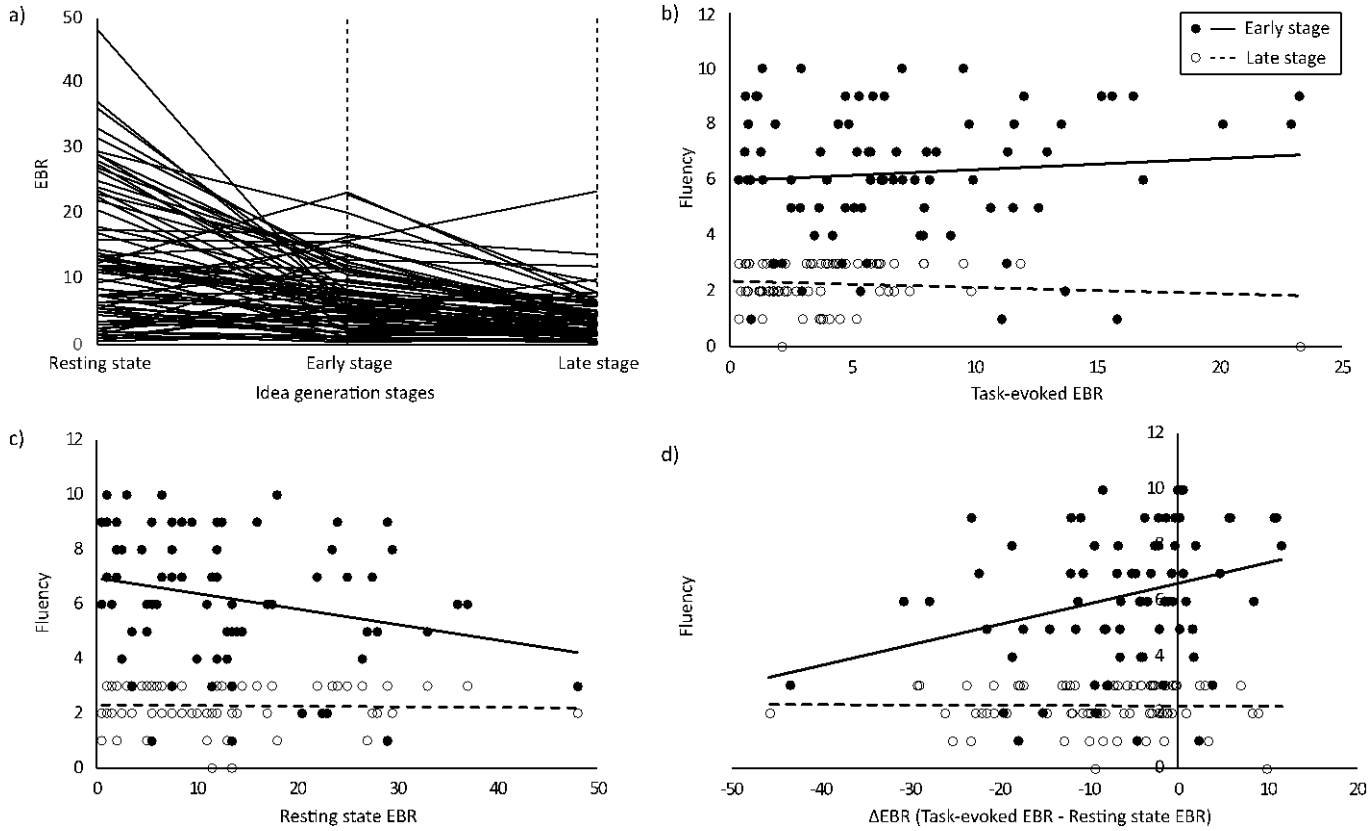


Figure 2: a) Spaghetti plot showing EBR measured at resting state, early stage, and late stage idea generation for each participant; b) Scatter plot of fluency and task-evoked EBR for early and late stage idea generation; c) Scatter plot of fluency and resting state EBR for early and late stage idea generation; and d) Scatter plot of fluency and the difference between task-evoked and resting state EBR for early and late stage idea generation.

positive correlation between the change in EBR and the amount of generated ideas during early stage idea generation, r = .298, p = .010, and a negative but not significant correlation between the change in EBR and the amount of generated ideas during late stage idea generation, r = -.014, p = .904 (Figure 2d). These findings indicate that the change in EBR positively correlates with the effective execution of early but not late stage idea generation.

Discussion

In the present study, the correlations of EBR with the effective execution of the idea generation process (as measured by the amount of generated ideas) during early and late idea generation were explored, using a task that was designed to closely resemble real-world creative tasks.

The results showed that resting state EBR negatively correlated with the amount of generated ideas during the early stage, but not during the late stage of creative idea generation (Figure 2c). This finding contrasts with previous research that suggested that the relationship between resting state EBR and the amount of generated ideas during the AUT best described with an inverted U-shape function[2] (Ueda et al., 2016), or that no significant correlation between the amount of ideas generated during the AUT and resting state EBR exists (Akbari Chermahini & Hommel, 2010). Possibly, this finding is more in line with previous research that indicates that the amount of solved items during the RAT negatively correlates with resting state EBR (Akbari Chermahini & Hommel, 2010), but this finding has been inconsistent across studies, cf. (Ueda et al., 2016).

The results also suggested that task-evoked EBR does not correlate with the amount of generated ideas during early nor during late stage idea generation (Figure 2b). This differs from previous research, which indicated task-evoked EBR positively correlated with the amount of uses listed during the AUT (Ueda et al., 2016); but is in line with results from the same study, which suggested that task-evoked EBR did not significantly correlate with the amount of correctly solved items during the RAT. In addition, differences between early and late stage idea generation could also be explained by previous findings that suggest that EBR quickly increases right before generating problem solutions via spontaneous insight (Salvi et al., 2015). Speculatively, moments of insight could appear more frequently in early than in late stage idea generation, as in the latter people focus more on recombining existing ideas. The results of the present study, however, also suggested that the change in EBR (from resting state to during the tasks) positively correlated with the amount of ideas generated during early, but not during late stage idea generation (Figure 2d). This is in line with previous research

that suggests that there are circumstances in which the change in EBR positively correlates with the amount of solved items, but not with related findings that suggest a negative correlation between the change in EBR and the amount of solved items during the RAT (de Rooij & Vromans, 2018).

There are, of course, also limitations that threaten the validity of the results. Although the present study purports to use a task with high ecological validity, no claims can be made on specific aspects of its validity. That is, due to the novelty of the task no tests of validity have been done (cf. de Rooij, Vromans, & Dekker, 2018). Furthermore, to enable measurement of EBR, idea generation was cued and split up into two tasks, representing early and late stage idea generation. In reality, such an artificial separation does not typically happen, and may hamper the often free flowing nature of creative idea generation (Lubart, 2001), which threatens the ecological validity of the used creative idea generation task, (cf. de Rooij & Vromans, 2018). Furthermore, to accommodate eye-tracking measurements responses were cued and limited to 10 responses during early, and 3 responses during late stage idea generation, limiting variance. The limited variance of late stage idea generation could therefore alternatively explain why no correlation between EBR and the amount of ideas generated in late stage idea generation was found. Decisions made to support ecological validity also came at the cost of introducing potential confounding factors. That is, no counterbalancing between early and late stage idea generation is possible, so any found differences could be confounded by adaptation to light conditions. Finally, due to the use of a novel task, it is difficult to compare the results obtained in the present study to results from previous related work. This limits the degree to which the results of this study can be grounded in such previous work. Limitations such as these should be taken into account when interpreting and building upon the present study.

The contribution of the present study is therefore that it shows for the first time that EBR predicts and dissociates the effective execution of early and late stage creative idea generation, using a creative task that resembles real-world creative tasks more closely than psychometric tasks of creative potential. Differences in the results between the present and previous studies using these psychometric tasks, show the importance of using tasks with higher face validity, as indeed, the results differ. This has implications for the development of theory on how the neuro-psychological correlates of EBR relate to creative idea generation.

References

Abraham, A. (2018). *The neuroscience of creativity*. Cambridge University Press.

Akbari Chermahini, S. A., & Hommel, B. (2010). The (b)link between creativity and dopamine: spontaneous eye blink rates predict and dissociate divergent and convergent thinking. *Cognition*, 115, 458-465.

Akbari Chermahini, S., & Hommel, B. (2012). More creative through positive mood? Not everyone!. *Frontiers in Human Neuroscience*, 6, 319.

Barbato, G., et al. (2000). Diurnal variation in spontaneous eye-blink rate. *Psychiatry research*, 93, 145-151.

Dalmaijer, E. S., Mathôt, S., & Van der Stigchel, S. (2014). PyGaze: An open-source, cross-platform toolbox for minimal-effort programming of eyetracking experiments. *Behavior research methods*, 46, 913-921.

de Rooij, A., & Jones, S. (2013, June). Mood and creativity: An appraisal tendency perspective. In *Proceedings of the 9th ACM Conference on Creativity & Cognition* (pp. 362-365). ACM.

de Rooij, A., & Vromans, R. D. (2018). The (dis) pleasures of creativity: Spontaneous eye blink rate during divergent and convergent thinking depends on individual differences in positive and negative affect. *The Journal of Creative Behavior*, online first.

de Rooij, A., Vromans, R. D., & Dekker, M. (2018). Noradrenergic Modulation of Creativity: Evidence from Pupillometry. *Creativity Research Journal*, 30, 339-351.

Finke, R. A., Ward, T. B., & Smith, S. M. (1992). *Creative cognition: Theory, research, and applications*. MIT Press.

Guilford, J. P. (1957). Creative abilities in the arts. *Psychological review*, 64, 110.

Isaksen, S. G., Dorval, K. B., & Treffinger, D. J. (2000). *Creative approaches to problem solving*: *A framework for change*. Kendall Hunt Publishing Company.

Jongkees, B. J., & Colzato, L. S. (2016). Spontaneous eye blink rate as predictor of dopamine-related cognitive function - A review. *Neuroscience & Biobehavioral Reviews*, 71, 58-82.

Kilgour, A. M. (2006). Improving the creative process: Analysis of the effects of divergent thinking techniques and domain specific knowledge on creativity. *Journal of Business and Society*, 7, 79-107.

Lubart, T. I. (2001). Models of the creative process: Past, present and future. *Creativity research journal*, 13, 295-308.

Mednick, S. A., & Mednick, M. (1971). *Remote associates test: Examiner's manual*. Houghton Mifflin.

Salvi, C., & Bowden, E. M. (2016). Looking for creativity: Where do we look when we look for new ideas?. *Frontiers in psychology*, 7, 161.

Salvi, C., Bricolo, E., Franconeri, S. L., Kounios, J., & Beeman, M. (2015). Sudden insight is associated with shutting out visual inputs. *Psychonomic bulletin & review*, 22, 1814-1819.

Ueda, Y., Tominaga, A., Kajimura, S., & Nomura, M. (2016). Spontaneous eye blinks during creative task correlate with divergent processing. *Psychological research*, 80, 652-659.

Walcher, S., Körner, C., & Benedek, M. (2017). Looking for ideas: Eye behavior during goal-directed internally

focused cognition. *Consciousness and cognition*, 53, 165-175.

Zeng, L., Proctor, R. W., & Salvendy, G. (2011). Can traditional divergent thinking tests be trusted in measuring and predicting real-world creativity?. *Creativity Research Journal*, 23, 24-37.

What is a good question asker better at? From no generalization, to overgeneralization, to adults-like selectivity across childhood

Costanza De Simone (desimone@mpib-berlin.mpg.de)
MPRG iSearch, Max Planck Institute for Human Development

Azzurra Ruggeri (ruggeri@mpib-berlin.mpg.de)
MPRG iSearch, Max Planck Institute for Human Development
School of Education, Technical University Munich

Abstract

Prior research showed that young children prefer to seek help from actors who have demonstrated active learning competence. What inferences do people make based on the ability to search effectively, for example by asking informative questions? This project explores across two experiments to what extent adults and children (3- to 9-year-olds) generalize the ability to ask informative questions to other abilities/characteristics. We presented participants with one monster who always asked informative questions and one who always asked uninformative questions. Participants had to choose which monster they thought was more likely to possess/was better at 12 different characteristics/abilities. Our results show a clear developmental trend. Three- and 4-year-olds draw unsystematic inferences from the monsters question-asking expertise. Five- and 6-year-olds identified the better question asker as better at everything. Seven- to 9-year-olds showed adult-like response patterns, selectively associating the ability to ask good questions to related characteristics/abilities.

Keywords: active learning; social cognition; question asking.

Introduction

Children are natural born active learners. However, while some skills and knowledge (e.g., basic laws of physics or object functions) can be acquired by first-hand active exploration or from observations, some other abilities (e.g., language) strongly rely and build on social interactions. Indeed, a vast body of research suggests that children are *programmed* to learn from others since the very beginning. Already 6- to 9-month-old infants are equipped with special attentional mechanisms to detect when a social partner is willing to transmit information (Senju & Csibra, 2008; Csibra & Gergely, 2009), and 9-month-olds use strategies such babbling or social referencing to seek explanations from their caregivers when presented with unfamiliar objects (Goldstein & Schwade, 2009; Walden, Kim, McCoy, & Karrass, 2007).

As soon as they can talk, children have more explicit ways to elicit explanations or request information: they can ask questions. Question asking is a powerful learning tool that children rely on to enlarge, deepen, enrich and adaptively revise their knowledge about the physical and social world (Callanan & Oakes, 1992; Campos, 1981; Chouinard, Harris, & Maratsos, 2007; Meltzoff, 1988b, 1988a, 1990). Previous work demonstrated that children are very selective when deciding whom to ask questions to, or more generally which informants to rely on. This research suggests that children's trust is driven by a complicated mixture of inferences drawn from the *quality* of the information provided (e.g., accuracy, completeness; see Pasquini, Corriveau, Koenig, & Harris, 2007; Koenig, Clément, Harris, & Clement, 2004; Jaswal, Croft, Setia, & Cole, 2010; Koenig & Jaswal, 2011) and the characteristics of the *agent* providing the information (e.g., expertise, age, familiarity, culture; see Lutz & Keil, 2002; VanderBorght & Jaswal, 2009; Kinzler & Spelke, 2011). As an example, Kushnir, Vredenburgh, and Schneider (2013) have shown that preschoolers use the quality of the information provided as a cue to infer the informants' scope of expertise. In their first study, they presented 3- and 4-year-olds with two informants (a *labeler* and a *fixer*), two familiar tools (a screwdriver and a wrench) and two unfamiliar electronic toys with interesting light or sound effects. The *labeler* provided accurate labels for the tools that he had to use to fix a broken toy, but did not manage to fix it. The *fixer* labeled the tools inaccurately but managed to fix the toys. Both 3- and 4-year-old children asked the labeler for help when they needed to know labels for novel toys, and turned to the fixer when they had to fix a broken toy, thus inferring expertise from the quality of the information provided (Kushnir, Vredenburgh, & Schneider, 2013).

Recent work shows that preschoolers are also sensitive to the effectiveness of the active learning strategies of a potential informant. In particular, children identify and rely on the most informative between two given questions already at age 4 (Ruggeri, Sim, & Xu, 2017), although they cannot reliably *generate* the most effective questions from scratch until age 10 (e.g., Herwig, 1982; Mosher, Hornsby, Bruner, J, & Oliver, 1966; Ruggeri & Feufel, 2015; Ruggeri & Lombrozo, 2015; Ruggeri, Lombrozo, Griffiths, & Xu, 2016). This research suggests that the cognitive machinery to support effective question asking may develop much earlier than the ability to generate effective questions from scratch. Why is this the case? On the one hand, it might be that what hinders younger children's effective question-generation is that their verbal abilities and vocabulary are not sufficiently developed. On the other hand, one intriguing possibility is that children's early ability to evaluate questions' informativeness allow them to assess another persons learning competence-a cue that can be used to identify role models to imitate and to *learn how to learn* from. Along these lines, a recent study showed that 3- to 7-year-old children preferentially sought help from a competent *active learner* who had figured out how to solve a problem by herself, over learners who

had learned through passive observation or direct instruction. Yet, this preference emerged only when the problem children needed to solve was similar to the one the learners had previously solved, where they thought the active learners competence would be relevant (Bridgers, Gweon, Bretzke, & Ruggeri, 2018). This paper investigates one crucial question arising from this perspective: *How* do children use active learning competence to identify role models, that is, what do they infer based on someone's ability to ask effective questions: Are good question-askers smarter, more knowledgeable, or better at solving problems? Do adults make similar inferences and generalizations? To address these questions, we explore to what extent adults (Study 1) and 3- to 9-year-old children (Study 2) generalize question-asking competence to other abilities/traits/characteristics.

We implemented a paradigm similar to that used in previous studies investigating the inferences and generalizations children make based on the informants' expertise and characteristics. For instance, Brousseau-Liard and colleagues (2010) presented 4- and 5-year-old children with two puppets that labeled 4 familiar objects. One did so correctly, and the other incorrectly. At test, children were asked to indicate which puppet they thought was more likely to possess 12 different skills/characteristics encompassing six categories: knowledge of words (e.g., Who knows words for lots of different machines?), talents (e.g. Who can draw pretty pictures?), knowledge of facts (e.g., Who knows that cats can see at night?), pro-social behavior (e.g., Who always shares her toys?), and two control-questions on possessions and situation-specific knowledge (e.g., Does she have a cat?; Who knows where I put my books?). Their results suggest that 5-, but not 4-year-olds, used the puppets' past accuracy to make explicit predictions about relevant characteristics such as her knowledge of words and facts and her pro-social behaviour, but not about her talents, possessions or situation-specific knowledge (Brosseau-Liard & Birch, 2010). Along the same lines, Lane and colleagues (2013) presented 3- to 6-year-old children and adults with three story books in which two pictured characters exhibited contrasting traits: Honest-dishonest, nice-mean, and smart-not smart. During the test phase, participants were tasked to ask - and endorse - characters' testimony about novel objects' names, about the content of a box that both characters had seen, and to attribute knowledge about the content of a different box, where only the negative informant had access to this information. Their results show that children and adults prefer to ask and endorse information about novel objects' names provided by people who are nice, honest and smart, whereas knowledge attribution seems to be influenced by the informants traits, following an age-graded decrease: 3- to 5-year-olds wrongly attributed knowledge to the positive informant, as opposed to 6-year-olds and adults, who correctly inferred the negative character's situation-specific knowledge (Lane, Wellman, & Gelman, 2013).

Based on the results discussed above (e.g., Lane et al., 2013), we expect to find an overall age-related decrease in the extent of generalizations from question-asking expertise to unrelated traits, abilities and characteristics, with older children and adults being generally more selective than younger children (see Mills, 2013 for a review on the development of selective trust). However, because very few studies investigating generalizations of expertise have considered a broad children age range as well as adults, we don't know when mature, adult-like selectivity would emerge.

Study 1

The goal of this study was to assess adults' intuitions about the relationship between question-asking competence and 12 different abilities/traits/characteristics.

Participants Thirty adults (11 males; M_{age}= 28.09; SD = 7.63) participated in this study. All participants were recruited and tested at the Museum für Naturkunde in Berlin, Germany. Participants belonged to various social classes and were native German or fluent in German. IRB approval was obtained and participants gave informed consent to participate in the study. One additional participant was excluded from the analyses due to missing data.

Design and procedure

The procedure consisted of two phases. The familiarization phase was designed to introduce participants to two agents (i.e., monsters), one who always asks informative questions and the other who always asks uninformative questions. In the test phase, we asked participants to rate the strength of the association between the question-asking competence illustrated in the familiarization phase and 12 given abilities/traits/characteristics. We detail the two phases below.

Familiarization phase Participants were asked to read a short storybook introducing two monsters, Bobo and Kila, who ask their friend Toma some questions to find out what happened on her first day at the new school. The storybook consisted of five pages. Each of the first four pages presented a different event taking place on Toma's first day at the new school (e.g., Toma drew a surprise welcome gift from a bag; see Figure 1 for an example) and two related questions that Bobo and Kila asked Toma (e.g., "Did you get a teddy bear?" or "Did you get a red toy car?"). On the bottom of the page, 8 cliparts, arranged in a row, illustrated the options to be considered (i.e., the hypothesis space; e.g., which gifts were inside the bag). Across the four scenarios, one of the monsters always asked informative questions, whereas the other always asked uninformative questions. The informative question targeted half of the hypotheses considered, either by asking a hypothesis-scanning question that referred to a single hypothesis presented 4 times (e.g., "Did you get a teddy bear?", where there were four teddy bears in the bag), or by asking a constraint-seeking question that addressed a feature shared by four of the hypotheses (e.g., "Did you get a round-shaped toy?", where there were four round-shaped toys in the bag). The uninformative question targeted either an object that was

not included in the hypothesis space (e.g., the red toy car; hypothesis-scanning question) or a feature shared by all the objects (e.g., a toy; constraint-seeking question). A fifth page presented again the two monsters and summarized the lesson to be learned from the familiarization phase, reminding participants that "Bobo/Kila always asks good/bad questions, because they are very informative/not informative at all. She is a good/bad question asker!"

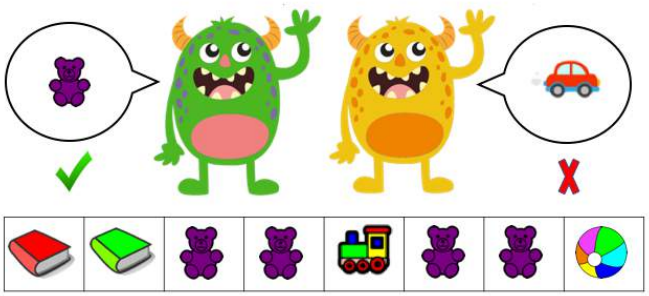


Figure 1: One scenario of the familiarization phase in which Bobo, the green monster, asks an informative question and Kila, the yellow monster, asks an uninformative question. In this example, the informative question refers to a single hypothesis presented 4 times, whereas the uninformative question targets an object that was not included in the hypothesis space.

Test phase In the test phase, participants were asked to complete a paper-and-pencil survey consisting of 12 questions, presented in random order. For each question, participants were asked to rate how much 12 different abilities, traits or characteristics (see Table 1) related to the ability to ask informative questions (e.g., "How much is being good at treasure hunting related to the ability to ask informative questions?"), on a scale from 0 ("not related at all") to 10 ("strongly related"). The questions presented were selected to include a set of abilities/traits/characteristics of different kinds (i.e., intellectual skills vs. physical abilities, individual preferences or irrelevant characteristics) that, according to our intuitions and to pilot survey data, are more or less related to the ability to ask informative questions, e.g., involve a stronger or weaker strategic component. For instance, being good at treasure hunting or at solving riddles might require the ability to search for information and explore strategically, whereas knowing many animal names refers to a domain-specific knowledge that is more strategy-independent.

Results and Discussion

As can be seen from Table 1, "being clever" and "being good at school" were rated as the most related to the ability to ask good questions. The association with abilities with a strategic component (i.e., "being good at treasure hunting" and "being fast at completing puzzles") were rated as medium-strong, and that with domain-specific knowledge (i.e. "knowing many animal names") was judged as medium-weak. Interestingly, "being friendly" was also rated as having a moderate-weak association with question-asking competency, although it had the highest between subjects variability. One possible interpretation could be that being good at asking questions is considered by some people, but not others, to indicate of a person being socially smart, sociable or just generally more likely to interact with others (Good, Slavings, Harel, & Emerson, 1987). Physical abilities or skills were rated low overall, independently of whether they were more likely to involve a strategic component ("being good at playing soccer") or not (i.e., "kicking a ball the furthest"). As expected, individual preferences (e.g., "liking ice cream") or irrelevant characteristics (e.g., "seeing the farthest","having siblings") were rated very low, that is, were judged as not at all related to the ability to ask informative questions.

Taken together, these results suggest that adults make distinct, graded, meaningful and fairly consistent inferences and generalizations based on the ability to ask good questions. Some abilities, traits and characteristics are considered to be strongly related to question-asking competency, whereas some others are considered to be only weakly related, or completely unrelated.

In Study 2 we explored to what extent such inferences and generalizations undergo a developmental change across childhood, and when adult-like intuitions might emerge.

Table 1: Study 1. Mean ratings of the strength of the association between question-asking competence and the abilities/traits/characteristics presented to adults in Study 1.

Abilities/traits/characteristics	Mean	SD
Being good at school	8.36	1.83
Being clever	8.30	1.91
Being good at treasure hunt	6.76	2.21
Being fast at completing puzzles	5.76	2.67
Knowing lots of animal names	4.20	2.68
Being friendly	3.56	3.16
Having siblings	2.13	2.53
Being good at playing soccer	1.63	2.08
Seeing the farthest	1.37	2.35
Scoring lots of goals	1.33	2.22
Kicking a ball the furthest	1.10	2.19
Liking ice cream	0.67	1.39

Study 2

Participants Participants were 40 3- and 4-year-old children (21 males; M_{age}= 48.41 months; SD = 7.19 months), 40 5- and 6-year-olds (19 males; M_{age}= 7.18; SD = 6.52 months) and 40 7- to 9-year-olds(18 males; M_{age}= 101.59 months; SD = 9.74 months), recruited and tested at local museums in Berlin. Participants belonged to various social classes and were native German or fluent in German. IRB approval was obtained and parents gave informed consent for their children

to participate before the study. Twenty-four additional participants were excluded from the analyses due to technical issues (n = 2) and for failing the attention (n = 7), the memory check (n = 9; see below), or both (n = 6; see below).

Design and procedure The design and procedure of Study 2 was identical to that of Study 1, with the following exceptions: First, the task (storybook and survey) was implemented on a 10" Tablet. Second, the familiarization story and the test questions were read to children by an experimenter, who also reminded them, at the end of each scenario, which monster was a good and which one was a bad question asker. Third, in the test phase, instead of being asked to provide a rating from 0 to 10 as in Study 1, children were asked, for each question, to select the monster they thought was more likely to possess/was better at the presented abilities/traits/characteristics. Two cards illustrating the monsters were used to help children indicate their preference. Finally, as an attention and memory check, we asked children both at the beginning and at the end of the test phase to indicate which monster was the best at asking questions.

Results

Children's selections were coded as "1" when they indicated the good question asker, or "0" when they indicated the bad question asker. Results are presented in 2. We performed a multivariate regression with adults' ratings in Study 1 as predictors of children's selections in Study 2. This analysis revealed that adults' ratings predicted 7- to 9-year-old children's response pattern ($\beta = .025$, $t(12) = 3.19$, $p = .01$; $R^2 = .455$, $F(1, 12) = 10.20$, $p = .01$), but not 5- and 6-year-olds' ($\beta = .018$, $t(12) = 1.65$, $p = .12$; $R^2 = .137$, $F(1, 12) = 2.75$, $p = .12$) nor 3- and 4-year-olds' ($\beta = .010$, $t(12) = .84$, $p = .41$; $R^2 = -.027$, $F(1, 12) = .716$, $p = .41$). We then performed a series of binomial tests to assess whether children' preference for the question asker on each ability/trait/characteristic differed from chance (50%). The results (see Table 2) show that the the extent of children generalizations strongly varies between age groups. Generally, 3-to 4-year-olds' showed a very unsystematic response pattern: They had no preference for the good question asker for abilities, traits and characteristics that both adults and older children deemed related to question asking (e.g., "being good at school", "being good at treasure hunting"), but displayed a strong preference for some clearly irrelevant questions (e.g., "having siblings", "seeing the farthest"). Five- to 6-year-olds clearly overgeneralized: They extended question-asking competence to both related and unrelated domains, selecting the good question-asker above chance for 10 of the 12 questions presented ("being good at school", "being clever", "being good at treasure hunting", "being fast at completing jigsaw puzzles", "knowing many animal names", "being friendly", "having siblings", "being good at playing soccer", "seeing the farthest", "liking ice cream"). However, 7- to 9-year-olds showed a systematic and meaningful attributions of relevant abilties/traits/characteristics to the good question-asker, very similar to the one found with adults in Study 1.

General Discussion

In this project we explored across two experiments to what extent adults (Study 1) and 3- to 9-year-old children (Study 2) generalize question-asking competence to other abilities/traits/characteristics. Taken together, our results suggest a clear developmental trend: Three- and 4-year-olds drew unsystematic inferences from the monsters question-asking expertise, showing no preference for the good question asker when evaluating abilities, traits and characteristics that both adults and older children deemed strongly related to question asking (e.g., "being good at school", "being good at treasure hunting"). At the same time, they showed a strong preference for the good question asker on some clearly irrelevant questions (e.g., "having siblings", "seeing the farthest"). Five- and 6-year-olds identified the good question asker as better/more likely to have nearly every presented ability/characteristic. Seven- to 9-year-olds showed adult-like response patterns, selectively associating question-asking competency to some, relevant abilities and characteristics (e.g., "being good at school" and "being clever"), but not to others (e.g., "kicking a ball the furthest", "seeing the farthest" and "liking ice cream").

Three- and 4-year-olds in our study failed to associate traits and abilities such as "being good at school" and "being good at treasure hunting" with question-asking expertise, an association rated strong by adults and older children. We should notice that these two characteristics might have been difficult to understand for children this age. On the one hand, they probably do not have yet a clear idea of what "being good at school" means, as they are not in school yet. On the other hand, they might not appreciate the strategic component underlying the ability of being good at treasure hunting. This component seems to be more evident for them in the ability of solving puzzles. Moreover, their preference response for "knowing many animal names" suggests that young children might consider semantic knowledge as connected to question asking and maybe, more generally, to active learning competence. This is in contrast to the results obtained by Fusaro and colleagues (2011) and Brosseau-Liard and Birch (2011). In their studies, 4-year-olds generalized behavior to traits (e.g. inferred that an accurate puppet would have been smart), but did not make any generalization from behavior to semantic knowledge (e.g., knowing animal habits; Fusaro, Corriveau, & Harris, 2011) or did not endorse the accurate puppets' testimony about situation-specific knowledge (e.g., knowing the content of a box; Brosseau-Liard & Birch, 2011).

Our results suggest that 5- and 6-year-olds considered effective question asking as an indicator of global rather than a domain- or ability-specific expertise and of general likability. This over-generalization trend is in line with some previous findings suggesting that children at this age tend to make broader generalizations when the informant possesses some

Table 2: Mean proportion of children who indicated the best question asker as more likely to possess each ability against chance (50%; binomial test) in Study 2.

	3-to 4-year-olds			5-to 6-year-olds			7-to 9-year-olds		
Abilities/traits/characteristics	Mean	CI	*p*	Mean	CI	*p*	Mean	CI	*p*
Being good at school	.60	.43 - .75	.20	.83	.67 - .92	<.001	.83	.67 - .92	<.001
Being clever	.68	.50 - .81	.03	.80	.64 - .90	<.001	.78	.61 - .89	.001
Being good at treasure hunting	.58	.48 - .73	.34	.65	.48 - .79	.05	.68	.50 - .81	.03
Being fast at completing puzzles	.68	.50 - .81	.03	.83	.67 - .93	<.001	.58	.40 - .73	.34
Knowing many animal names	.78	.61 - .89	<.01	.78	.61 - .89	<.001	.75	.58 - .87	<.01
Being friendly	.70	.53 - .83	.01	.93	.80 - .98	<.001	.73	.56 - .85	<.01
Having siblings	.73	.56 - .85	<.01	.65	.48 - .79	.05	.68	.50 - .81	.03
Being good at playing soccer	.45	.29 - .62	.52	.75	.59 - .87	<.01	.68	.50 - .81	.03
Seeing the farthest	.75	.58 - .86	<.01	.68	.50 - .81	.03	.63	.45 - .77	.11
Scoring lots of goals	.45	.29 - .62	.52	.55	.39 - .71	.52	.55	.38 - .70	.52
Kicking a ball the furthest	.60	.43 - .75	.20	.60	.43 - .75	.20	.55	.38 - .70	.52
Liking ice cream	.50	.33 - .66	1	.75	.58 - .87	<.01	.50	.33 - .66	1

kind of semantic knowledge (e.g., labels objects accurately, Brosseau-Liard & Birch, 2010; knows causal properties of an object, Sobel & Corriveau, 2010) and demonstrates sociomoral goodness (Cain, Heyman, & Walker, 1997). However, Bridgers and colleagues (2018) demonstrated that already at age 4, children selectively generalize active learning effectiveness only to tasks where this competence is deemed relevant. This apparent inconsistency might indicate that children have different intuitions and make different generalizations depending on the different *kinds* of active learning competency an agent display (e.g., physical exploration versus question asking), and this differential pattern might interact with age. It would be interesting to explore this question in future research.

The adult-like response pattern of 7- to 9-year old children, who selectively associated question-asking competence only to related abilities and traits, is in line with the few results from previous research focusing on this age group (e.g., Lane et al., 2013; Danovitch & Keil, 2004. For example, Danovitch and Keil (2007) presented 6, 8 and 9- year-olds with four short vignettes illustrating a character facing a moral dilemmas (e.g., respect another's privacy) or involved in a scientific problem (e.g., building a rocket). Following each vignette, participants were asked to choose what characteristics the character would have needed to solve the problem (e.g., "Does he need to be nice with other people" or "Does he need to be smart"). Their results showed that only starting at age 8 children consistently indicated that scientific skills were necessary to solve scientific problems and that moral characteristics were needed to solve moral dilemmas (Danovitch & Keil, 2007). Finally, it might be that to make selective, meaningful inferences about question asking, one has to be good at asking questions herself. In this respect, the developmental trend found in this study might be reflective of children's improvement in question-asking effectiveness. Future research should explore whether and how children's ability to ask informative questions or search effectively more in general can impact the inferences and generalizations they make based on others' active learning competence. Moreover, it seems clear that even older children's responses did not always and perfectly reflect adults' intuitions. This difference could be resulting from the different ways in which participants were asked to elicit their intuitions in Study 1 and 2.

It is crucial to note that in our studies the good question asker was simply contrasted with a *bad* question asker, who was not given any other positive nor neutral attributes, whereas in many studies focusing on generalizations, including those reviewed above, informants are presented as experts in different domains (e.g., (Lutz & Keil, 2002; Kushnir et al., 2013; Jaswal et al., 2010; Koenig, 2012). We are currently finishing data collection on a follow up study in which we implement the more traditional version of paradigm, contrasting an agent who is good at asking questions and "finding out things" with one that has a domain-specific expertise (e.g., knows everything about fish). Future work should also investigate the impact of such inferences and generalizations on children's learning and social behavior, for example examining under which conditions children would prefer to imitate, learn or ask for help to someone they identify as an effective active learner.

To conclude, this study is a first attempt at exploring what children infer based on someone's ability to ask effective questions. We found an interesting developmental pattern from unsystematic generalization, to overgeneralization, to adults-like selective generalization, suggesting that children at different ages use information about an agent's active learning competence in different ways. This is a first step in understanding whether and how children use their sensitivity to others' active learning competence to navigate the social world, identifying good role models to learn, and to *learn how to learn* from.

Acknowledgments

We thank Federico Meini for developing the app used for this study, Lola Hermann and Andreas Sommer for assistance in recruiting and data collection. We would also like to thank our collaborators in Berlin, the Museum für Naturkunde, Labyrinth Kindermuseum, Science Center Spectrum and the FRÖBEL-Kindergarten im Lützelsteiner Weg for providing research space.

References

Bridgers, S., Gweon, H., Bretzke, M., & Ruggeri, A. (2018). How you learned matters : The process by which others learn informs young children's decisions about whom to ask for help. *Cognitive Science*, *1*, 1402–1407.

Brosseau-Liard, P. E., & Birch, S. A. (2010). I bet you know more and are nicer too: what children infer from others accuracy. *Developmental Science*, *13*(5), 772–778.

Brosseau-Liard, P. E., & Birch, S. A. (2011). Epistemic states and traits: Preschoolers appreciate the differential informativeness of situation-specific and person-specific cues to knowledge. *Child Development*, *82*(6), 1788–1796.

Cain, K. M., Heyman, G. D., & Walker, M. E. (1997). Preschoolers' ability to make dispositional predictions within and across domain. *Social Development*, *6*(1), 53–75.

Callanan, M. A., & Oakes, L. M. (1992). Preschoolers' questions and parents' explanations: Causal thinking in everyday activity. *Cognitive Development*, *7*(2), 213–233.

Campos, J. (1981). 8:. stenberg, cr (1981). perception, appraisal, and emotion: The onset of social referencing. *Infant social interaction: Empirical and theoretical considerations*, 273–314.

Chouinard, M. M., Harris, P. L., & Maratsos, M. P. (2007). Children's questions: A mechanism for cognitive development. *Monographs of the Society for Research in Child Development*, i–129.

Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in Cognitive Sciences*, *13*(4), 148–153.

Danovitch, J. H., & Keil, F. C. (2004). Should you ask a fisherman or a biologist?: Developmental shifts in ways of clustering knowledge. *Child Development*, *75*(3), 918–931.

Danovitch, J. H., & Keil, F. C. (2007). Choosing between hearts and minds: Children's understanding of moral advisors. *Cognitive Development*, *22*(1), 110–123.

Fusaro, M., Corriveau, K. H., & Harris, P. L. (2011). The good, the strong, and the accurate: Preschoolers evaluations of informant attributes. *Journal of Experimental Child Psychology*, *110*(4), 561–574.

Goldstein, M. H., & Schwade, J. A. (2009). From Birds to Words : Perception of Structure in Social Interactions Guides Vocal Development and Language Learning. *The Oxford Handbook of Developmental and Comparative Neuroscience.*, *29*(5), 737–767.

Good, T. L., Slavings, R. L., Harel, K. H., & Emerson, H. (1987). Student passivity: A study of question asking in k-12 classrooms. *Sociology of Education*, 181–199.

Herwig, J. E. (1982). Effects of age, stimuli, and category recognition factors in children's inquiry behavior. *Journal of experimental child psychology*, *33*(2), 196–206.

Jaswal, V. K., Croft, A. C., Setia, A. R., & Cole, C. A. (2010). Young Children have a Specific, Highly Robust Bias to Trust Testimony. *Psychological Science*, *21*(10), 1541–1547.

Kinzler, K. D., & Spelke, E. S. (2011). Do infants show social preferences for people differing in race? *Cognition*, *119*(1), 1–9.

Koenig, M. A. (2012). Beyond semantic accuracy: Preschoolers evaluate a speakers reasons. *Child Development*, *83*(3), 1051–1063.

Koenig, M. A., Clément, F., Harris, P. L., & Clement, F. (2004). Children ' s Use of True and False Statements. *Psychological Science*, *15*(10), 694–698.

Koenig, M. A., & Jaswal, V. K. (2011). Characterizing childrens expectations about expertise and incompetence: Halo or pitchfork effects? *Child Development*, *82*(5), 1634–1647.

Kushnir, T., Vredenburgh, C., & Schneider, L. A. (2013). who can help me fix this toy? the distinction between causal knowledge and word knowledge guides preschoolers' selective requests for information. *Developmental psychology*, *49*(3), 446.

Lane, J. D., Wellman, H. M., & Gelman, S. A. (2013). Informants' traits weigh heavily in young children's trust in testimony and in their epistemic inferences. *Child Development*, *84*(4), 1253–1268.

Lutz, D. J., & Keil, F. C. (2002). Early understanding of the division of cognitive labor. *Child development*, *73*(4), 1073–1084.

Meltzoff, A. N. (1988a). Infant imitation after a 1-week delay: long-term memory for novel acts and multiple stimuli. *Developmental psychology*, *24*(4), 470.

Meltzoff, A. N. (1988b). Infant imitation and memory: Nine-month-olds in immediate and deferred tests. *Child development*, *59*(1), 217.

Meltzoff, A. N. (1990). Foundations for developing a concept of self: The role of imitation in relating self to other and the value of social mirroring, social modeling, and self practice in infancy.

Mills, C. M. (2013). Knowing when to doubt: developing a critical stance when learning from others. *Developmental psychology*, *49*(3), 404–418.

Mosher, F. A., Hornsby, J. R., Bruner, J, S., & Oliver, R. (1966). Mosher, Hornsby - 1966 - On asking questions.pdf. In *Studies in gognitive growth* (pp. 86–102). New York: Wiley.

Pasquini, E. S., Corriveau, K. H., Koenig, M., & Harris, P. L. (2007). Preschoolers Monitor the Relative Accuracy of Informants. *Developmental Psychology*, *43*(5), 1216–1226.

Ruggeri, A., & Feufel, M. A. (2015). How basic-level objects facilitate question-asking in a categorization task. *Frontiers in Psychology*, *6*.

Ruggeri, A., & Lombrozo, T. (2015). Children adapt their questions to achieve efficient search. *Cognition*, *143*(October), 203–216.

Ruggeri, A., Lombrozo, T., Griffiths, T. L., & Xu, F. (2016). Sources of developmental change in the efficiency of information search. *Developmental psychology*, *52*(12), 2159.

Ruggeri, A., Sim, Z. L., & Xu, F. (2017). "Why is Toma late to school again?" Preschoolers identify the most informative questions. *Developmental Psychology*, *53*(9), 1620–1632.

Senju, A., & Csibra, G. (2008). Gaze Following in Human Infants Depends on Communicative Signals. *Current Biology*, *18*(9), 668–671.

Sobel, D. M., & Corriveau, K. H. (2010). Children monitor individuals' expertise for word learning. *Child Development*, *81*(2), 669–679.

VanderBorght, M., & Jaswal, V. K. (2009). Who Knows Best? Preschoolers Sometimes Prefer Child Informants over Adult Informants. *Infant and Child Development*, *18*(1), 61–79.

Walden, T., Kim, G., McCoy, C., & Karrass, J. (2007). Do you believe in magic? Infants' social looking during violations of expectations. *Developmental Science*, *10*(5), 654–663.

Distinguishing Two Types of Prior Knowledge That Support Novice Learners

Anita B. Delahay (adelahay@cmu.edu)
Carnegie Mellon University, Department of Psychology,
5000 Forbes Ave. Pittsburgh, PA 15213 USA

Marsha C. Lovett (lovett@cmu.edu)
Carnegie Mellon University, Department of Psychology, and
Eberly Center for Teaching Excellence & Educational Innovation
5000 Forbes Ave. Pittsburgh, PA 15213 USA

Abstract

Prior knowledge has long been recognized as an important predictor of learning, yet the term prior knowledge is often applied to related but distinct constructs. We define a specific form of prior knowledge, *ancillary knowledge*, as knowledge of concepts and skills that enable learners to gain the most from a target lesson. Ancillary knowledge is not prior knowledge of the lesson's target concepts and skills, and may even fall outside the domain of the lesson. Nevertheless, ancillary knowledge affects learning of the lesson, e.g., lower ancillary knowledge can hinder performance on lesson-related tasks. We measured ancillary knowledge, prior knowledge of the domain, and controlled for general ability, and found that (a) stronger ancillary knowledge and general ability predicted better performance on transfer tasks, but (b) prior knowledge of the domain did not. This research suggests that enhancing instruction by remediating gaps in ancillary knowledge may improve learning in introductory-level courses.

Keywords: prior knowledge; ancillary knowledge; domain-general knowledge; far transfer; introductory courses

Introduction

Learners in any given class often vary widely with respect to their knowledge of both the current material and the skills and concepts that may be considered ancillary to and supportive of the current material. At the college level, this is perhaps most evident in introductory-level courses, which by definition enroll many learners who are novices in a domain, and yet who bring all types and degrees of prior knowledge into the classroom. Before attending Introduction to Cognitive Psychology, for example, students may or may not have taken a general psychology course that included a high-level introduction to many topics. They are also likely to have had different degrees of exposure to and practice with concepts and skills that could be considered supportive of learning Cognitive Psychology, e.g., graphing and experimental design. These topics, which may have been learned in the context of psychology or a different science or math context, are likely useful to students as they learn about cognitive psychology hypotheses, study designs, graphed results, and whether the data support these hypotheses. Despite the clear relevance of graphing and experimental design knowledge, rarely are they measured or their gaps addressed during instruction.

We wished to evaluate whether such ancillary knowledge would predict performance on assessment items related to a new lesson better than prior domain (cognitive psychology) knowledge or knowledge of the specific lesson, which would suggest that this unmeasured and often unaddressed type of knowledge plays an important role in learning.

Background

Researchers have long considered prior knowledge critical for learning (Ausubel, 1968; Dochy, 1988; Jonassen & Grabowski, 1993). Across studies, it represents one of the largest sources of variance in pre/post-test measures, accounting for 30 to 60 percent of the difference in scores (Dochy, 1988). Prior knowledge explains performance over and above general ability. For example, it predicted learning of science concepts better than mental capacity and developmental level (Lawson, 1983) or formal reasoning ability (Zeitoun, 1988), and comprehension of text passages after accounting for IQ (Langer & Nicolich, 1981).

The importance of prior knowledge for learning is well established, yet many studies do not provide explicit definitions of prior knowledge or use similar terms to reference distinct constructs (Dochy & Alexander, 1995). Consequently, important dimensions of prior knowledge may be overlooked, and research becomes inconclusive. In addition, some benefits of prior knowledge may be due to prior knowledge in the domain, ancillary knowledge, or both; similarly, learning difficulties may be due to knowledge gaps of either type. More generally, if learners vary in both types of prior knowledge, but the two types are not distinctly assessed, their role in learning cannot be clearly understood.

Our research aims to distinguish these two types of prior knowledge: *prior knowledge in the domain*, i.e., concepts and skills within the target domain, from *ancillary knowledge*, i.e., knowledge of the concepts and skills that are outside of the target domain (but may be utilized in the target domain and additional domains; see Dochy, 1988). See *Figure 1* for a graphical representation. In order to be considered ancillary knowledge, these concepts and skills should enable better learning of the new material, such as knowledge of graphing and experimental design may for many cognitive psychology topics. Bloom's (1976) term for this idea was "cognitive entry behaviors," which he defined as "those prerequisite types of knowledge, skills, and competencies which are essential to the learning of a particular new task or set of tasks" (p. 122).

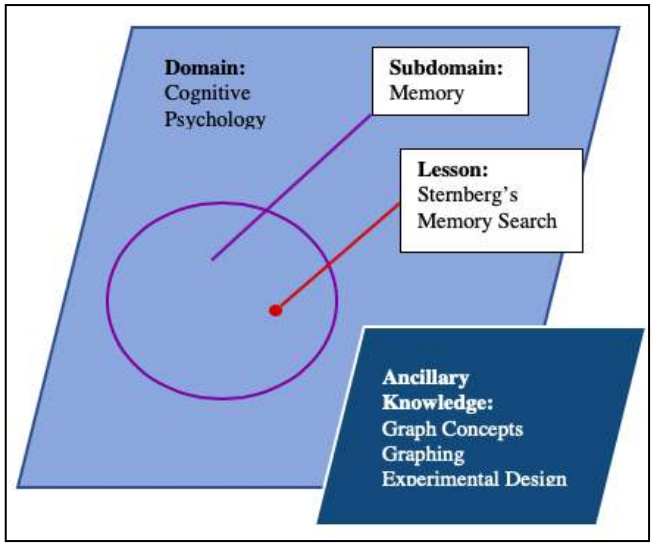


Figure 1: Levels of domain knowledge include the domain, subdomains, and concepts and skills within the domain. Ancillary knowledge is domain-independent, but may nevertheless support learning in the domain.

The mechanisms of support for learning are likely the same for both types of prior knowledge, including freeing up attentional resources and enabling greater comprehension and problem solving (Fincher-Kiefer, Post, Greene, & Voss, 1988; Kintsch, 1994; Schauble, Glaser, Raghavan, & Reiner, 1991; Siegler, 1986; Willingham, 2007). The key difference is that prior knowledge in the domain is obviously relevant and ancillary knowledge is often overlooked or deemed out of scope of the current instruction.

This is particularly problematic for undergraduates, who are likely to have gaps in the types of ancillary knowledge that readily support experts as they encounter new topics. Schunn and Anderson (1999) contrasted experts' and undergraduates' performance on an experimental design task and found that the latter did not demonstrate the experimental design skills of using theory to design their experiment and relating results back to the theories at a proficient level. On the other hand, experts have a wealth of domain knowledge and tools they can bring to bear in new situations, such as knowledge of related studies or formulas typically used.

Despite this, undergraduates have often acquired a measurable degree of general knowledge (Means & Voss, 1985), general strategies (de Jong & Ferguson-Hessler, 1996), and even subject-specific knowledge (Dochy & Alexander, 1995). Variability is heightened because the knowledge may have been learned and forgotten, partially learned, or not abstracted at a high enough level to be useful in new contexts. In other words, undergraduates' base levels of knowledge are more sophisticated (tending toward greater richness) than younger students, but also more tenuous and incomplete than experts' knowledge. Therefore, instead of categorizing subjects as experts or novices, we took a quantitative measure of ancillary and domain knowledge in our target population in order to pick up on this variability.

In addition, in order to investigate the role of ancillary knowledge in undergraduate learning, we utilized a situation typical in introductory courses, namely one in which ancillary knowledge and domain knowledge were expected to vary greatly, but prior knowledge of the lesson was uniformly low (i.e., not at play). The specific lesson we chose was the Sternberg memory search paradigm and experimental results, as taught in an introductory Cognitive Psychology course at Carnegie Mellon University. A key advantage of this lesson was that several questions on the assessments were adapted from materials that had been used in the course and therefore already deemed suitable (i.e., challenging but within grasp) for the average ability level of our sample.

We analyzed the lesson to determine what would qualify as ancillary knowledge – i.e., independent of the target lesson and yet expected to enhance learning of that lesson. We identified the following as relevant ancillary knowledge: variable selection and measurement, facility with graphed data and the lines that fit these data, and interpretation of graphed results in terms of theoretical relationships between variables. Consistent with this list, a reviewer of this paper shared that a lack of ancillary knowledge related to graphing prevented his or her students from fully understanding Sternberg's hypothesis, his various independent variables and study results. In other words, missing ancillary knowledge (i.e., an inability to apply knowledge about y-intercept and slope) affected the extent to which students were able to gain lesson-specific knowledge.

In order to separate ancillary knowledge fully from the domain of cognitive psychology, we situated the pre-test questions in other domains, such as physics (for graphing questions) or social psychology (for experimental design questions). We measured prior knowledge in the domain by assessing knowledge of the subdomain of memory (e.g., chunking, serial position effect), as well as prior knowledge of the lesson (e.g., Sternberg's paradigm, hypothesis, and results). See *Table 1* for sample questions.

Conceptual and Procedural Knowledge

Our measures also differentiated between types of ancillary knowledge based on another dimension that is often included in studies of learning and performance: conceptual versus procedural knowledge. The classification of knowledge as conceptual or procedural is both common (Baroody, Feil, & Johnson, 2007; Crooks & Alibali, 2014) and useful for studying learning and performance. Researchers and educators sometimes call conceptual knowledge "knowing that" and procedural knowledge "knowing how," or, even more simply, concepts (conceptual) and skills (procedural).

Determining *how* to measure conceptual apart from procedural knowledge became a secondary focus of our work. Even though the labels conceptual and procedural suggest the idea of two independent categories, these knowledge types are often related. Rittle-Johnson and Siegler (1998) reported positive correlations between amounts of conceptual and procedural knowledge in four areas of math learning.

Table 1: Sample item from each of the six (6) knowledge types assessed at pre-test.

Knowledge Type		Sample Item
Ancillary Knowledge – Graphing	*(1) Conceptual*	Here is a Boxplot (also called a Box and Whisker plot). Circle any feature that can be determined.
	(2) Procedural	What does the slope of an object accelerating uniformly look like on an acceleration vs. time graph? Hint: Sketch a graph with acceleration on the x-axis.
Ancillary Knowledge – Experimental Design	*(3) Conceptual*	A social psychology researcher is interested in whether cheerfulness and extroversion determine a person's attractiveness. She does an experiment in which several participants view videos of interviews of everyday people and then rate the interviewee's perceived cheerfulness, extroversion, and attractiveness […] Are the results correlational or causal?
	(4) Procedural	Advertisements for an herbal product, ginseng, claim that it promotes endurance. As a researcher, how would you design a controlled experiment to test this claim? Describe each of the following: (e.g., groups, controls, dependent measure)
Prior Knowledge – Subdomain (Memory)	*(5) Conceptual*	While recalling a mobile phone number, splitting it into groups of 3 or 4 digits tends to be easier to remember than a single long number. *Why* does this chunking process work?
Prior Knowledge – Lesson (Sternberg)	*(6) Conceptual*	What did some researchers find surprising (counter-intuitive) about the mental search process Sternberg proposed?

Furthermore, Rittle-Johnson, Siegler, and Alibali (2001) described conceptual knowledge as knowledge of principles, concepts, and rules and when to apply those principles, and procedural knowledge as routinized knowledge acquired from explicit practice of a given problem type. From this view, any novel problem requires conceptual knowledge, because it has been neither practiced nor routinized. This presented a challenge, as we wished to gauge procedural knowledge of graphing and experimental design via the pre-test and then assess performance on novel procedural transfer problems at post-test. As stated, at pre-test, we addressed this issue by giving problems from outside the domain of cognitive psychology with the assumption that the procedural skills had been learned elsewhere. However, this was not possible at post-test, which was given in the context of the current lesson.

At post-test, we assessed procedural knowledge as knowledge of steps that we considered scriptable, and therefore teachable, whether or not students actually learned that procedure in the context of our lesson. Procedural assessment items included finding a slope, determining the ratio of two slopes, designing an experiment, determining the nature of a novel search process by executing a learned algorithm, etc., all in the context of lesson-specific concepts. By contrast, conceptual items tested facts, principles, or declarative knowledge, for example asking students to recall a fact, explain an answer, label a diagram, graphically depict a concept, etc.

Research Questions

We tested two research questions. First, does ancillary knowledge predict performance on near and/or far transfer questions, controlling for both general ability (i.e., SAT scores) and prior knowledge in the domain (i.e., the subdomain of memory)? We hypothesized that ancillary knowledge would predict learning but that prior knowledge of the domain would not.

Second, did we sufficiently distinguish conceptual and procedural knowledge types in the psychology domain? We measured concepts and procedures separately, on both the pre- and post-tests. Evidence that these variables are acceptably independent in terms of their correlations would be suggestive that our operationalization was successful. In addition, evidence that conceptual or procedural ancillary knowledge had differential patterns of association with the various post-test measures would also provide some support.

Method

Participants

80 undergraduate students (M_{age}=19.85 years, 63.8 percent female) completed the study for course credit.

Design and Procedure

A correlational design was used to study how natural variation in ancillary knowledge (pre-test question types 1-4 below) would relate to performance at post-test. On the pre-test, four questions each assessed (1) graphing conceptual knowledge, (2) graphing procedural knowledge, (3) experimental design conceptual knowledge, and (4) experimental design procedural knowledge. In addition, four questions assessed (5a) prior knowledge of memory, and two questions assessed (5b) prior knowledge of the lesson. These last two questions (5b) were the only ones that repeated between pre- and post-test. They were ultimately not used as a pre-test measure, because we determined that participants' knowledge of the target lesson was uniformly low/absent.

Due to limited time for the experiment, our goal on the pre-test was to sample sufficiently from each area of prior knowledge in order to determine a quantitative measure of

probable degree of knowledge in each area, not to try to assess each area in depth.

Next, participants read a two-page lesson, which was about 700 words with several figures, adapted from J.R. Anderson's 8th Ed. Textbook, *Cognitive Psychology and Its Implications*. Then, participants completed a practice activity that was either conceptual or procedural in nature. The results of the practice manipulation and two additional measures of conceptual knowledge following the practice manipulation are not reported in this paper.

Next, participants completed the Post-Test. To measure participants' learning, we created four types of questions (and therefore four outcome measures): (1) Text-based Questions could be answered successfully if participants formed an adequate mental model of the text as they read. Participants did not need to bridge inferences, but rather draw from their memory of the text in order to recall information (see Kintsch, 1994; McNamara et al., 1996). (2) Near-transfer conceptual items were related to the lesson, but had not been stated directly in the text and therefore required bridging inferences. (3) Near-transfer procedural items required participants to perform procedures in the context of newly learned lesson concepts. For example, participants were asked to determine and compare the slopes of lines depicting the relationship between lesson-specific variables. This drew on preexisting knowledge of procedures (i.e., finding slopes and comparing their ratios) in the context of the lesson concepts. (4) Far-transfer items required participants to apply knowledge and skills they had learned (i.e., types and levels of variables, graphed data, and underlying hypotheses from Sternberg's memory search paradigm) to other types of mental processes, including a visual search task and a mental rotation task. See *Table 2* for sample post-test questions. Finally, participants were asked to provide demographic data and aptitude scores.

In sum, there were nine predictor variables. Five were taken from the pre-test: (1) Ancillary Graphing, conceptual, (2) Ancillary Graphing, procedural, (3) Ancillary Experimental Design, conceptual, (4) Ancillary Experimental Design, procedural, and (5) Prior Knowledge Memory, conceptual. As stated, knowledge of the lesson was excluded from analysis because the pre-test items related to the Sternberg lesson were answered incorrectly or left blank (with only one subject answering one item correctly).

The other four variables were covariates: (6) SAT Verbal scores, (7) SAT Math scores (if ACT scores were given, they were converted), (8) Reading Time, a measure of how long the participant spent on the lesson, and (9) English Native, a categorical variable indicating whether the student was a native English speaker from at least the age of six.

Table 2: Sample item from each of the four (4) outcome measures assessed at post-test.

Knowledge Types		Sample Items
Text-based	*(1) Conceptual*	What did some researchers find surprising (counter-intuitive) about the mental search process Sternberg proposed?
Near Transfer	*(2) Conceptual*	Which independent variable from the list above has the greatest influence on the slope of the line in the graph?
	(3) Procedural	The graph below shows the relationship between Memory Set Size and Response Time for Foil trials (A) and Target trials (B). Compared with the increase in reaction time for B, the increase in reaction time for A is…
Far Transfer	*(4) Conceptual & Procedural (mixed within each question)*	Consider a new type of mental task. This one involves conducting a visual search for an item, such as a red circle in a field of distractors (similar items). In Feature search, a person is asked to find the red o in a field of green x's and o's. (a) A feature search is most like a _____ (parallel/serial) search. (b) Graph the lines for Target and Absent trials on the graph below. Label the lines.

Analyses and Results

Predictor Variables

Predictor variables were tested for normality. Several of the variables were negatively skewed and/or kurtotic, including both pre-test concept variables (graphs, experimental design) and SAT scores. In these cases, each score was reflected and then logarithmically transformed. These transformations resulted in acceptable normality, and these variables were re-reflected after transformation to aid in interpretation of beta coefficients.

Procedural Knowledge types (i.e., Graphing, Experimental Design), Reading Time, and Prior Knowledge-Memory, were normally distributed. The categorical variable English Native Speaker was answered "yes" seventy-percent of the time. Four cases did not report SAT scores and so the variable means were imputed for those cases.

Tests to see if the data met the assumption of collinearity indicated that multicollinearity was not a concern, with all $VIF < 2$. Correlations between each pair of predictor variables are reported in *Table 3*. The highest pairwise correlation was 0.55, between SAT Math and Verbal, below the conservative cutoff of 0.7 for multicollinearity. Seven pairs of predictor variables were significantly, positively correlated.

Table 3: Correlations between predictor variables.
** Correlation is significant at the 0.01 level (2-tailed). * Correlation is significant at the 0.05 level (2-tailed).

	Graphing Concepts	Graphing Procedures	Exp. Design Concepts	Exp. Design Procedures	Memory Concepts	SAT Verbal	SAT Math	Reading Time
Graphing Concepts	1							
Graphing Procedures	**0.336****	1						
Exp. Design Concepts	0.060	0.135	1					
Exp. Design Procedures	0.090	**0.252***	0.183	1				
Memory Concepts	0.209	0.172	0.169	0.095	1			
SAT Verbal	**0.254***	**0.241***	-0.026	-0.032	0.035	1		
SAT Math	**0.477****	**0.345****	0.063	-0.127	0.159	**0.547****	1	
Reading Time	-0.170	0.009	-0.177	0.072	-0.159	-0.064	-0.067	1

Linear Regressions

We regressed each of the Post-Test measures (Text-based Questions, Near Transfer Concepts, Near Transfer Procedures, and Far Transfer) on the nine explanatory variables in order to determine whether ancillary knowledge, prior knowledge in the domain, or general ability predicted performance. The model for Text-based Questions was the only model that was not significant.

The Near Transfer Concepts model was significant, $F(9,70) = 3.508$, $p = 0.001$, $R^2 = .311$, *Adj.* $R^2 = .222$. Greater Ancillary Knowledge of Graphing Concepts ($\beta = 4.426$, $t = 1.957$, $p = 0.054$) predicted better performance on near transfer concept questions.

The Near Transfer Procedures model was significant, $F(9,70) = 4.542$, $p = 0.001$, $R^2 = .369$, *Adj.* $R^2 = .288$. Greater Ancillary knowledge of Graphing Procedures ($\beta = 0.779$, $t = 3.574$, $p = 0.001$) predicted better performance on near transfer procedural questions.

The Far Transfer model was significant, $F(9,70) = 5.814$, $p = 0.001$, $R^2 = .428$, *Adj.* $R^2 = .354$. Higher SAT Math score ($\beta = 2.259$, $t = 2.330$, $p = 0.023$) and greater Ancillary Knowledge of Experimental Design Procedures ($\beta = 1.370$, $t = 4.114$, $p = 0.001$) each predicted better performance.

Finally, to test whether model fit was better when the four types of ancillary knowledge were entered as separate predictors in the model, as we had done, versus entered as a single predictor (and therefore treated as having a similar effect on learning), we compared *Adj.* R^2, AIC score, and BIC score for each of the models. *Adj.* R^2 and BIC are more sensitive to number of predictors and therefore penalize more for model size than AIC. The remaining predictors entered into the model were the same: SAT scores, Prior Knowledge of Memory, Reading Time, and English Native.

Model fit scores are shown in *Table 4*. All six models were significant at the 0.001 level. In addition, the single score for ancillary knowledge was a significant predictor in each of those three models. There was a slightly higher *Adj.* R^2 (0.005 more variance explained) for the single predictor model for Near Concepts, and a lower *Adj.* R^2 with the single predictor model for Near Procedures (0.009 less variance explained) and for Far Transfer (0.045 less variance explained). BIC, which penalizes more for number of predictors than AIC, was unsurprisingly lower in the single predictor models than the separate predictor models, signifying less overfitting. AIC was similar across the models, with the AIC for the single predictor model being slightly lower for Near Concepts and Near Procedures, but slightly higher for Far Transfer.

Table 4: Model fit for separate vs. combined (i.e., a single, additive score) ancillary knowledge predictors. The separate ancillary predictor models had nine predictors, whereas the single ancillary predictor models had six.

	Separate Ancillary Predictors Model (df=9)				*Single Ancillary Predictor Model (df=6)*			
For Outcome Measure:	R^2	*Adj.* R^2	*BIC*	*AIC*	R^2	*Adj.* R^2	*BIC*	*AIC*
Near Concepts	0.311	0.222	437.24	411.04	0.286	0.227	428.20	409.15
Near Procedures	0.369	0.288	356.78	330.58	0.334	0.279	347.44	328.38
Far Transfer	0.428	0.354	439.22	413.02	0.362	0.309	432.46	413.41

Discussion

This research identified ancillary knowledge and skills that predicted performance on near and far transfer assessment questions related to a cognitive psychology lesson. In each case, more ancillary knowledge led to better performance. We measured four types of ancillary knowledge that had a low degree of intercorrelation, namely conceptual and procedural knowledge of graphing and experimental design. Furthermore, specific ancillary measures predicted the various post-test measures, and statistical models that treated the ancillary knowledge as distinct accounted for the same or more variance than those that treated the ancillary knowledge as monolithic.

In addition, we found encouraging evidence that it is possible to operationalize assessment questions that differ on the conceptual and procedural knowledge dimension, and thereby measure them as distinct constructs, in this domain (i.e., cognitive psychology) and at this lesson and instructional level (i.e., introductory college coursework). Even though procedural assessment items are clearly dependent on a grasp of the lesson concepts in this context, we crafted procedural post-test questions that were predicted by ancillary procedural knowledge as assessed by pre-test problems in a different domain, suggesting that the procedural knowledge itself was uniquely important and domain independent. We do not interpret these results as implying that only procedural knowledge was needed for any of the assessment questions we labeled procedural (that is, independent of conceptual knowledge of the lesson).

SAT Math scores were also predictive of success in the Far Transfer model. And while not reported above, SAT Verbal was nearly significant ($p = 0.068$) in the Near Transfer Concepts model, and SAT Math was nearly significant ($p = 0.076$) in the Near Transfer Procedures model. It is not surprising that aptitude played a role in learning, particularly unsupported learning (e.g., no instructor) of a novel lesson, as students had to make sense of the material for themselves. Even so, ancillary knowledge was predictive over and above aptitude in the models.

In contrast to ancillary knowledge and general ability, prior knowledge in the memory subdomain of cognitive psychology was not predictive in any of the models. This should not be interpreted as domain-specific knowledge failing to predict performance in general, however. Schunn and Anderson (1999) found that domain-specific knowledge contributed to the performance of experts, and we expect that if students continued studying in the domain, domain-specific knowledge would be more and more predictive of their performance. In this study, however, the novice status of the participants lent greater importance to the variability in their ancillary knowledge and its role in their learning.

Finally, neither time spent reading the lesson nor native English speaker status from the age of six was predictive of performance, which is not surprising given that all students in our sample regularly do coursework in English, and that we also included a verbal aptitude measure in the model.

One final note about the specific assessment questions written for each of the outcome measures, some of which came from an actual cognitive psychology course and some of which were written for this lab study. The type of ancillary knowledge that predicted greater success on each transfer measure was arguably both a function of the outcome type (conceptual or procedural) and also the specific questions written for the category. As seen in Table 2, many of the questions written for the Near Transfer Concepts measure asked students to assess the relative influence of study variables in graphs of data and as they related to various theories, and so knowledge of graphing concepts is a logical predictor. Many of the questions written for the Near Transfer Procedures measure asked students to apply procedures related to graphing and experimental design in the context of the newly learned topic. For example, they were asked to predict values of various independent variables (e.g., stimulus quality, biasing) in Sternberg's experiment, imagine a graph for Accuracy instead of Reaction Time in Sternberg's experiment, etc., and so knowledge of graphing procedures is likewise a logical predictor.

For the Far Transfer measure, questions required students to apply their new lesson knowledge to mental processes that they had not previously encountered in their reading. Applying knowledge of the Sternberg paradigm in order to graph data and predict results for novel mental processes would likely benefit from greater knowledge of experimental design. Had we written different assessment questions, we expect that different ancillary knowledge structures would have been useful to the students, and a task analysis would have revealed that relevant knowledge. Furthermore, we consider it probable that ancillary knowledge structures beyond graphs, graphing, and experimental design could be measured and found predictive of better success at post-test.

Our study had several limitations. First, we tested our hypotheses regarding ancillary knowledge in the context of one lesson, so demonstrating the generalizability of these findings will be critical. Psychology is a domain that is conceptually rich, and therefore future studies should include a greater number of ancillary constructs. Second, the design was correlational, so we could not rule out other possible causes of performance differences. Third, we did not attempt to remedy gaps in knowledge structures that we identified in order to determine how such intervention would impact learning. Studying methods of remediating specific skills in the context of a new lesson versus outside the context of the lesson could suggest productive instructional practices once gaps are identified. Finally, work to further differentiate conceptual and procedural ancillary knowledge would be useful. Due to time limits, we were only able to take gross measures of ancillary knowledge at pre-test. A greater depth of pre-assessment, paired with the use of methods such as think-aloud protocols, to detect conceptual or procedural processing during the assessment would aid our understanding of how undergraduates construct, structure, and utilize their knowledge when encountering a new lesson.

Conclusion

We have defined a type of prior knowledge, ancillary knowledge, that differs from other types of prior knowledge in important ways. It is both domain-independent and yet relevant to learning of a target lesson. We also provide evidence that ancillary knowledge can be productively differentiated into various subtypes of knowledge (i.e., graphing vs. experimental design; conceptual vs. procedural).

The distinction between ancillary knowledge and prior knowledge of the domain is relevant for researchers who study learning, and may have bearing on the design of pre-tests. If the target population of students varies in level of domain-independent knowledge that may still have direct bearing on the lesson, pre-testing for ancillary knowledge (in addition to prior knowledge of the domain or lesson) would be relevant and potentially important for understanding patterns of learning.

In addition, distinguishing ancillary knowledge has implications for the design of instructional materials. Finding ways to identify and close gaps in ancillary knowledge could enhance the effectiveness of instruction for novice learners and ultimately improve learning and transfer.

Acknowledgments

We are grateful to Charles Kemp for sharing his Cognitive Psychology course materials. The research reported here was supported, in whole or in part, by the Institute of Education Sciences, U.S. Department of Education, through grant R305B150008 to Carnegie Mellon University. The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education.

References

Ausubel, D. P. (1986). *Educational psychology: A cognitive view*. New York: Holt, Rinehart & Winston.

Baroody, A. J., Feil, Y., & Johnson, A. R. (2007). An alternative reconceptualization of procedural and conceptual knowledge. *Journal for Research in Mathematics Education*, 115-131.

Bloom, B. S. (1976). *Human characteristics and school learning*. New York: McGraw-Hill.

Crooks, N. M., & Alibali, M. W. (2014). Defining and measuring conceptual knowledge in mathematics. *Developmental Review*, *34*(4), 344-377.

De Jong, T., & Ferguson-Hessler, M. G. (1996). Types and qualities of knowledge. *Educational Psychologist*, *31*(2), 105-113.

Dochy, F. J., & Alexander, P. A. (1995). Mapping prior knowledge: A framework for discussion among researchers. *European Journal of Psychology of Education*, *10*(3), 225-242.

Dochy, F. J. R. C. (1988). *The "Prior Knowledge State" of Students and Its Facilitating Effect on Learning: Theories and Research*. Available from Open University, Herleen, OTIC Research Report 1.2.

Fincher-Kiefer, R., Post, T. A., Greene, T. R., & Voss, J. F. (1988). On the role of prior knowledge and task demands in the processing of text. *Journal of Memory and Language*, *27*(4), 416.

Jonassen, D. H., & Grabowski, B. L. (1993). *Handbook of individual differences, learning, and instruction*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Kintsch, W. (1994). Text comprehension, memory, and learning. *American Psychologist*, *49*(4), 294.

Langer, J. A., & Nicolich, M. (1981). Prior knowledge and its relationship to comprehension. *Journal of Reading Behavior*, *13*(4), 373-379.

Lawson, A. E. (1983). Predicting science achievement: The role of developmental level, disembedding ability, mental capacity, prior knowledge, and beliefs. *Journal of Research in Science Teaching*, *20*(2), 117-129.

Means, M. L., & Voss, J. F. (1985). Star Wars: A developmental study of expert and novice knowledge structures. *Journal of Memory and Language*, *24*(6), 746.

McNamara, D. S., Kintsch, E., Songer, N. B., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction*, *14*(1), 1-43.

Rittle-Johnson, B., Siegler, R. S., & Alibali, M. W. (2001). Developing conceptual understanding and procedural skill in mathematics: An iterative process. *Journal of educational psychology*, *93*(2), 346.

Rittle-Johnson, B., & Siegler, R. S. (1998). The relation between conceptual and procedural knowledge in learning mathematics: A review. In C. Donlan (Ed.). *The Development of Mathematical Skills*. UK: Psychology Press.

Schauble, L., Glaser, R., Raghavan, K., & Reiner, M. (1991). Causal models and experimentation strategies in scientific reasoning. *The Journal of the Learning Sciences*, *1*(2), 201-238.

Schunn, C. D., & Anderson, J. R. (1999). The generality/ specificity of expertise in scientific reasoning. *Cognitive Science*, *23*(3), 337-370.

Siegler, R.S. (1986). *Children's thinking*. Englewood Cliffs, NJ: Prentice-Hall.

Willingham, D. T. (2007). Critical thinking. *American Educator*, *31*(3), 8-19.

Zeitoun, H.H. (1988). *The relationship between abstract concept achievement and prior knowledge, formal reasoning ability, and sex among some Egyptian secondary school students*. Paper presented at the annual meeting of the National Association for Research in Science and Teaching, Lake of the Ozark, MO.

Parents' Linguistic Alignment Predicts Children's Language Development

Joseph Denby and **Daniel Yurovsky**
{jgdenby, yurovsky}@uchicago.edu
Department of Psychology
University of Chicago

Abstract

Children quickly gain enormous linguistic knowledge during early development, in part due to low-level features of their parents' speech. Some posit that parents contribute to their child's language development by tuning their own language according to their child's developmental abilities and needs (Bruner, 1985; Snow, 1972). Here, we investigate this hypothesis by examining 'alignment' at the level of syntax and function words in a large-scale corpus of parent-child conversations and measuring its association with language development outcomes. To do so, we employ a statistical model of alignment to estimate its presence in our dataset and its predictive impact on a measure of vocabulary development. Our results corroborate previous findings, showing strong alignment for both parents and children; in addition, we demonstrate that parental alignment is a significant predictor of language maturity independent of demographic features, suggesting that parental tuning has strong ties to a child's language development.

Keywords: Language acquisition; statistical modeling; vocabulary development

Introduction

Children make vast linguistic strides within their first few years of life. In light of this, some researchers have offered the linguistic tuning hypothesis, arguing that parents bolster their child's early language learning by calibrating the complexity of their speech to the particular abilities and needs of their children (Montag & MacDonald, 2015; Snow, 1972; Thiessen, Hill, & Saffran, 2005). The idea is intuitive, but it is unclear at what level of language tuning occurs (Hayes & Ahrens, 1988; Sokolov, 1993; Spivey & Dale, 2006) and how overt it is (Brown & Hanlon, 1970; Chouinard & Clark, 2003; Hirsh-Pasek, Treiman, & Schneiderman, 1984).

A parallel yet complementary vein of language development research investigates the presence of low-level cues in parental speech and their influence on child language learning. From this research, we know that child-directed speech contains features that facilitate language learning, and that more exposure tends to result in better outcomes (Cameron-Faulkner, Lieven, & Tomasello, 2003; Weisleder & Fernald, 2013). Related, caregivers from families of high socioeconomic status (SES) tend to converse more with their children than their lower SES counterparts, and these increases are associated with improved development outcomes such as vocabulary size and school performance (Hoff, 2003; Walker, Greenwood, Hart, & Carta, 1994). Moreover, differences in SES-based language development are largely explained by low-level features of parental child-directed speech such as lexical diversity and sentence complexity (Hoff-Ginsberg, 1998; Rowe, 2008). So, given that granular aspects of parental speech can have substantial effects on a child's language development, it may be that linguistic tuning occurs at this level in subtle ways, particularly when it comes to non-content words (i.e., words that are not central to the topic of discussion.)

This idea of assessing the direct impact of a parent's usage of non-content words on language development relates to linguistic alignment, a phenomenon whereby conversational partners tend to align aspects of their communicative style and content according to various external influences (Pennebaker, Booth, Boyd, & Francis, 2015). Alignment can occur at various levels of language, with some research (including ours) focusing on the level of quasi-syntactic categories (e.g., Ireland et al., 2010; Niederhoffer & Pennebaker, 2016). These categories don't strictly describe syntax; instead, they aim to capture function words, which are more invariant to context than content words. However, we often use the phrase 'syntactic alignment' here as shorthand for 'alignment within function word categories.' As an example, see the exchange between a child and parent presented in Table 1. The parent's usage of "across" directly following their child's usage of "across" presents alignment within the category of prepositions. Alignment need not involve repetition however; the child's use of "I" following their parent's use of "I'll" serves as alignment within a category as well (the category of 'I' pronoun words.) Alignment between parents and children may lend support to the linguistic tuning hypothesis - if parents align to their children in a way that changes across development, and that alignment has a concrete impact on a child's language acquisition, the tuning hypothesis could be vindicated (Bruner, 1985).

Yurovsky, Doyle, & Frank (2016) investigates linguistic alignment in CHILDES (MacWhinney, 2000), a natural language corpus of conversations between parents and children to assess whether tuning occurs at the level of function word categories. They find that alignment does occur between both parents and children; moreover, parents align less over time, suggesting that the relationship their speech shares with their child's changes as a function of development. These results present a powerful proof of concept that alignment within function word categories exists between parents and children

and changes over time, but it remains unclear whether alignment bears any sort of concrete, impactful relationship to language development.

Parent	I don't know . I'll have to think about it .
Child	I was going to do the people across street .
Parent	across the street ?
Child	yeah .

Table 1: Excerpt from exchange between 38 month old child and mother in LDP.

Here, we extend Yurovsky, Doyle, and Frank's (2016) model by applying it to the Language Development Project (LDP) (Goldin-Meadow et al., 2014), a corpus of ecological conversations between parents and their children over time, collected from a socioeconomically diverse sample of parent-child dyads. The variability present within this dataset aids our estimation by offering a more robust picture of alignment as it actually occurs. We follow their method of assessing alignment only within function words. Moreover, we use alignment estimates alongside demographic information to predict measures of vocabulary development, supporting the linguistic tuning hypothesis by concretely showing how parents' sensitivity to their child's linguistic needs and abilities covaries with their development.

Model

The linguistic tuning hypothesis predicts that parents will calibrate their language in part by assessing their child's needs and abilities. So, we predict that parents will exhibit high alignment to their young children, but will reduce their alignment as their children mature (and improve in linguistic maturity.) To test this prediction, we employ an extended version of the Hierarchical Alignment Model implemented in Yurovsky et al. (2016) which both estimates the impact of a speaker's use of function word categories on their conversational partner's usage and uses these alignment estimates to predict language outcome scores.

At base, for each utterance the model predicts whether the speaker will produce a word from a given function word category. This prediction is generated by two factors: the speaker's baseline propensity towards using that category and the speaker's tendency to align, producing words from a category just used by their partner. In the model, the primary computation mimics a standard logistic regression - the production of a category within an utterance is treated as a binary outcome variable impacted by a linear combination of predictor variables (here, baseline usage and alignment.) The model's hierarchical structure then allows the estimates of baseline usage and alignment effects to be pooled across individual speakers and categories in a way that ensures statistical robustness.

The model used here then incorporates these alignment and baseline usage estimates as predictors in a linear regression model of the Pearson Peabody Vocabulary Test (PPVT) (Dunn & Dunn, 1997), a widely used inventory for tracking language development. Measures of vocabulary like the PPVT offer a robust snapshot of overall language abilities throughout early language development, with PPVT scores in particular correlating with various other measures of cognitive ability (Hodapp, Gerken, & 1999, 1999; Naglieri, 1981). As one of various measures of cognitive and language ability present within the LDP dataset, we selected the PPVT for its reliability and validity in addition to it being a measure not based solely on parent report. At this stage, PPVT is estimated as a linear combination of predictors reflecting alignment and baseline usage estimates for both parents and children, alongside other features representing demographic variables (e.g., child's gender, mother's education) and the child's age. Moreover, the PPVT was administered to each child in LDP at least twice, allowing us to estimate interaction effects between parameter and demographic variables with age.

category	examples
article	a, alot
certain	altogether, must
conj	but, or
discrep	wanted, hoped
excl	whether, not
i	i'm, i
incl	both, around
ipron	thatd, whats
negate	needn't, oughtn't
preps	at, to
quant	series, every
tentat	anyhow, most
we	we'd, lets
you	youd, y'all

Table 2: LIWC Categories with example words.

Model Details

The structure of the model used here greatly resembles that used in Yurovsky et al. (2016), in that it operates over utterances represented as binary vectors, with indices indicating the presence or absence of each of the 14 LIWC categories used within alignment literatures (Pennebaker et al., 2015) to designate function words (Table 2). The probability of producing each category in each utterance is computed via two parameters: the speaker's baseline usage of that LIWC category (η^{base}), and the change in that speaker's baseline as a function of interacting with the listener (η^{align}). So, for a given category c, for replies to utterances that don't contain c, the production parameter for that category is computed by applying the inverse logit function to the appropriate baseline log odds:

$$P(Production_c) = logit^{-1}(\eta_c^{base})$$

Alternatively, replies to utterances that *do* contain c, the parameter computation takes into account the sum of the baseline and alignment log odds:

$$P(Production_c) = logit^{-1}(\eta_c^{base} + \eta_c^{align})$$

To accommodate the variance in production across the LIWC categories, each baseline usage parameter was drawn from an uninformative prior ($\eta^{base} \sim Uniform(-5,5)$); alignment parameters were regularized towards 0 by way of implementing a conservative prior ($\eta^{align} \sim Normal(0,.25)$).

All parameters were estimated hierarchically, which allows intelligent pooling of data across participants in the dataset. To start, each subpopulation (i.e., parents vs. children) obtained an estimate. Then, every speaker had an alignment estimate drawn from their appropriate subpopulation (e.g., if Speaker 22 is a child, their alignment estimate is drawn from the estimate for children overall.) Category-level alignment estimates were then drawn for each speaker (e.g., the alignment estimate for Speaker 22's usage of determiners is drawn from Speaker 22's overall alignment estimate.) The order was flipped for baseline estimates in order to better reflect empirical baseline usages across LIWC categories; subpopulation estimates produced category-level estimates, which then produced speaker-level estimates. As in Yurovsky et al. (2016), we also include parameters that allow baseline and alignment probabilities to change linearly over time (β and α respectively).

Next, we extend the model used in Yurovsky et al. (2016) by using estimated alignment (i.e., η parameters) to predict PPVT scores, a measure of vocabulary development (Dunn & Dunn, 1997). To do so, we implement a regression model where PPVT scores are modeled as linear combinations of various predictor variables. These predictor variables included the child's age, alignment parameter estimates for the child and their parent, the mother's education, the child's gender, as well as interaction effects for all variables with age. We use mother's education as a well known proxy for socioeconomic status (Hollingshead, 1975). Error variance for the model (σ) was also estimated.

The model implemented here then serves two purposes: (1) It extends the analysis of Yurovsky et al. (2016) to a new dataset, aiming to replicate previous findings in a more diverse and representative sample, and (2) It incorporates alignment estimates in a predictive model of early language outcomes, serving to test the hypothesis that alignment has a significant relationship with language development, even in the presence of demographic features. To be specific, we hope to replicate non-zero estimates for η parameters (demonstrating that alignment between parents and children exists across datasets), positive β for children (showing that children increase their baseline usage of categories over time), and negative α for parents (showing that parents decrease their alignment as their children age.) If the PPVT model estimates for parameters corresponding to the main or interaction effects of alignment are non-zero in the presence of demographic variables, we can infer that alignment has a relationship with vocabulary development independent of features like socioeconomic status, bolstering the linguistic tuning hypothesis.

Analysis

Data and Methodology

Conversations between parents and their children were drawn from the Language Development Project Corpus (Goldin-Meadow et al., 2014). Participants in the project were video-recorded in their homes for $\sim$ 90 minutes every four months starting when the child was 14-months and ending at 58-months. Additionally, all participants took the PPVT on at least two occasions during the observation period. Participants were selected in order to produce a diverse sample demographically representative of the broader Chicagoland area. LDP is smaller than other comparable corpora of child-parent conversations (e.g., CHILDES), but it stands alone in its broad representation of families across the socioeconomic spectrum.

We selected for analysis all children who were typically developing and completed at least 10 of the 12 planned recording sessions. Our sample consisted of 59 target children, 28 of whom were girls, 12 were Black and 6 were Multiracial. Children were also socio-economically diverse, as measured by mother's education: 2 mothers had some highschool education, 7 had a highschool degree, 10 had some college or trade school, 19 had college degrees, and 21 had advanced degrees.

Following Yurovsky et al. (2016), successive utterances from a speaker within a transcript were concatenated into a single utterance. Individual utterances were then transformed into binary vectors with indices indicating the presence or absence of each of the 14 LIWC categories. This pre-processing turned every transcript into a speaker-reply format: each utterance within a transcript was both a reply to the preceding utterance and a message to the next one.

Each transcript was then compressed, yielding 4 numbers for each LIWC category. For a pair of speakers A and B in a transcript, for each LIWC category, we computed the number of utterances from A to B containing the category (N^{align}), the number of utterances from A to B not containing the category (N^{base}), the number of utterances containing the category responding to an utterance containing the category (C^{align}), and the number of utterances containing the category responding to an utterance not containing the category (C^{base}). Aggregating in this way provided the platform for the model's sampling - for each transcript, C^{base} and C^{align} were drawn from Binomial distributions parameterized by N^{base} and N^{align} chances respectively, with probabilities computed via the logistic regression models outlined above.

Sampling was performed using Stan, a probabilistic programming language that implements Hamiltonian Monte Carlo sampling methods (Carpenter et al., 2017). Posterior distributions for each parameter in the model were estimated using 500 iterations of Bayesian sampling, generating mean assess-

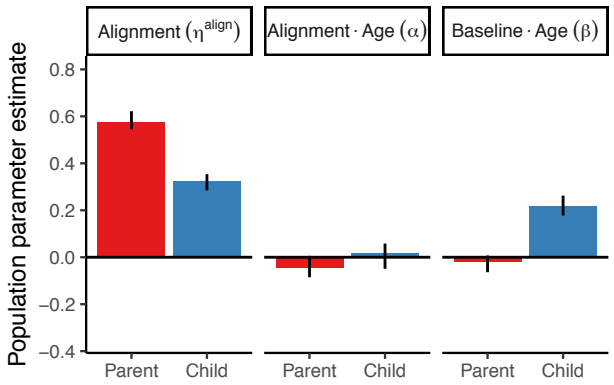


Figure 1: Posterior parameter estimates for alignment (η), developmental change in alignment (α), and developmental change in baseline function word production (β) for both parents and children, as well estimated alignment between parents for a baseline. Bars indicate means, error-bars indicate 95% highest posterior density intervals generated via Bayesian sampling.

ments with appropriate confidence intervals. [1]

Results

Alignment estimates (η^{align}) for parents and children were both estimated above zero, corroborating the findings of Yurovsky et al. (2016) in showing that both groups exhibit alignment (Figure 1). We also replicate the finding that parents appear to align more to their children than children align to their parents.

The model estimates changes in baseline category production across development (β) at approximately zero for parents, but significantly above zero for children, replicating previous findings. Alignment is estimated as having a significantly negative age effect (α) for parents in this dataset, replicating an earlier finding that alignment from parents to children tends to decrease over their child's development (Figure 3).

The mean estimates for PPVT predictors are presented in Table 2; they illustrate effects on a child's average PPVT score as well as estimates of interaction effects with age (i.e., the rate at which PPVT improves over development.) As expected, PPVT is positively associated with the age of the child and their being female. Moreover, female children tend to have a decreased age effect on PPVT; female children have a higher average PPVT score relative to male counterparts, but their scores improve over time more gradually. Mother's education is negatively associated with PPVT, but has a slight positive age effect. Alongside these demographic effects, we see robust alignment effects on PPVT: child and parental alignment are both associated with increased PPVT, but with decreased age effects.

[1] Data and code available at `https://github.com/callab/ldp-alignment`.

Parameter	Estimate	StandardError
Intercept	-234.12	21.78
Age (years)	73.63	6.19
Female	53.46	7.20
Age x Female	-10.56	1.84
Mother's Education	-19.08	2.59
Age x Mother's Education	4.70	0.62
Child Alignment	28.68	1.28
Age x Child Alignment	-62.89	4.83
Parent Alignment	409.79	36.24
Age x Parent Alignment	-72.19	14.04

Table 3: Parameter Estimates for PPVT predictors (and intercept) with standard errors. Parameters with "x" denote estimates of variable interaction.

Discussion

In an effort to understand and investigate how children rapidly acquire language, some argue that the language parents produce to their children is somehow calibrated to the child's particular needs and abilities (Snow, 1972). While the idea is theoretically compelling, empirical work has produced mixed results, with strong results in favor of (Chouinard & Clark, 2003; Hirsh-Pasek et al., 1984) and against (Brown & Hanlon, 1970; Hayes & Ahrens, 1988).

However, much of this prior work investigates tuning as an overt effort on behalf of parents or tuning with respect to content words, with less examining the potential role of low-level syntactic influence (Hoff, 2003). Yurovsky et al. (2016) presents just such an examination, demonstrating using Bayesian hierarchical modeling that parents align to their children according to their particular language usage at the level of function word categories. This paper extends their model by applying it to a new socioeconomically diverse sample of families (Goldin-Meadow et al., 2014) and leveraging the model's alignment estimates to predict language development outcomes.

The analysis presented here replicates the findings of Yurovsky et al. (2016), showing strong alignment effects for both parents and their children, a substantial age effect for baseline useage in children, and a significant negative effect of age on alignment for parents. Moreover, we demonstrate that these alignment estimates have substantial power in predicting vocabulary development measures, even in the presence of demographic features such as gender and socioeconomic status. We corroborate previous findings that female children tend to have higher PPVT scores that improve more gradually over time (Kaushanskaya, Gross, & Buac, 2013; Lange, Euler, & Zaretsky, 2016). We conflict with other findings that positively associate child PPVT scores with mother's education (Di Cesare, Sabates, & Lewin, 2013; Schady, 2011); this may be due to idiosyncracies of our dataset, including its limited size.

We show that parental alignment is associated with a relatively large boost in average PPVT scores, but with a negative

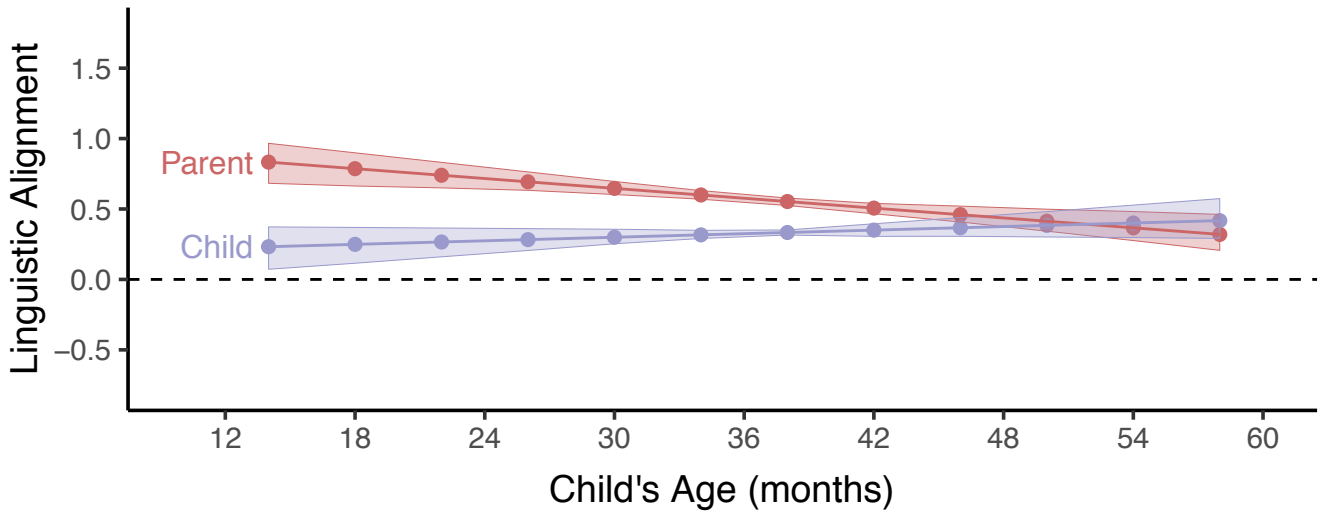


Figure 2: Model-estimated changes in linguistic alignment over development. Points indicate the mean of the posterior distribution; shaded regions indicate 68% highest probability density intervals, equivalent to one standard deviation, for visualization purposes.

age effect. The negative age effect may source from a ceiling on PPVT - children with higher average scores may simply have less ground to cover. Nevertheless, these results are consistent with a concrete effect of parental alignment on vocabulary development, and the linguistic tuning hypothesis more broadly. A similar story is evident from child alignment estimates: alignment has a small association with overall PPVT score and an age effect comparable to parental alignment. Here there may be a confound with childrens' baseline language production, in that children with lower production will have lower PPVT and diminished alignment as a result; future work should assess this interaction to better isolate the effects of alignment.

Overall, these results show that parental alignment within function word categories is a robust effect that appears to have a relationship with childrens' language development independent of demographic correlates, serving to further the linguistic tuning hypothesis.

References

Brown, R. W., & Hanlon, C. (1970). Derivational Complexity and Order of Acquisition in Child Speech. In J. Hayes (Ed.), *Cognition and the development of language* (pp. 11–53). New York.

Bruner, J. (1985). Child's Talk: Learning to Use Language. *Child Language Teaching and Therapy*, *1*(1), 111–114.

Cameron-Faulkner, T., Lieven, E., & Tomasello, M. (2003). A construction based analysis of child directed speech. *Cognitive Science*, *27*(6), 843–873.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., ... Riddell, A. (2017). Stan : A Probabilistic Programming Language. *Journal of Statistical Software*, *76*(1).

Chouinard, M. M., & Clark, E. V. (2003). Adult reformulations of child errors as negative evidence. *Journal of Child Language*, *30*(3), 637–669.

Di Cesare, M., Sabates, R., & Lewin, K. M. (2013). A double prevention: how maternal education can affect maternal mental health, child health and child cognitive development. *Longitudinal and Life Course Studies*, *4*.

Dunn, L. M., & Dunn, L. M. (1997). Peabody Picture Vocabulary Test–Third Edition. *PsycTESTS Dataset*.

Goldin-Meadow, S., Levine, S. C., Hedges, L. V., Huttenlocher, J., Raudenbush, S. W., & Small, S. L. (2014). New evidence about language and cognitive development based on a longitudinal study: Hypotheses for intervention. *American Psychologist*, *69*(6), 588–599.

Hayes, D. P., & Ahrens, M. G. (1988). Vocabulary simplification for children: a special case of 'motherese'? *Journal of Child Language*, *15*(2), 395–410.

Hirsh-Pasek, K., Treiman, R., & Schneiderman, M. (1984). Brown & Hanlon revisited: mothers' sensitivity to ungrammatical forms. *Journal of Child Language*, *11*(01), 81–88.

Hodapp, A. F., Gerken, K., & 1999. (1999). Correlations between scores for Peabody picture vocabulary testIII and the Wechsler intelligence scale for childrenIII. *Psychological Reports*, *84*(3), 1139–1142.

Hoff, E. (2003). The specificity of environmental influence: Socioeconomic status affects early vocabulary development via maternal speech. *Child Development*, *74*(5), 1368–1378.

Hoff-Ginsberg, E. (1998). The relation of birth order and socioeconomic status to children's language experience and language development. *Applied Psycholinguistics*, *19*, 603–629.

Hollingshead, A. A. (1975). *Four-factor index of social status*. New Haven, CT.

Ireland, M. E., Slatcher, R. B., Eastwick, P. W., Scissors, L.

E., Finkel, E. J., & Pennebaker, J. W. (2010). Language Style Matching Predicts Relationship Initiation and Stability. *Psychological Science*, *22*(1), 39–44.

Kaushanskaya, M., Gross, M., & Buac, M. (2013). Gender differences in child word learning. *Learning and Individual Differences*, *27*, 82–89.

Lange, B. P., Euler, H. A., & Zaretsky, E. (2016). Sex differences in language competence of 3- to 6-year-old children. *Applied Psycholinguistics*, *37*(06), 1417–1438.

MacWhinney, B. (2000). The CHILDES Project. *Computational Linguistics*, *26*(4), 657–657.

Montag, J. L., & MacDonald, M. C. (2015). Text exposure predicts spoken production of complex sentences in 8- and 12-year-old children and adults. *Journal of Experimental Psychology: General*, *144*(2), 447–468.

Naglieri, J. A. (1981). Concurrent validity of the revised Peabody Picture Vocabulary Test. *Psychology in the Schools*, *18*(3), 286–289.

Niederhoffer, K. G., & Pennebaker, J. W. (2016). Linguistic Style Matching in Social Interaction. *Journal of Language and Social Psychology*, *21*(4), 337–360.

Pennebaker, J. W., Booth, R. J., Boyd, R. L., & Francis, M. E. (2015). Linguistic inquiry and word count: LIWC2015. Austin, TX: Pennebaker Conglomerates.

Rowe, M. L. (2008). Child-directed speech: relation to socioeconomic status, knowledge of child development and child vocabulary skill. *Journal of Child Language*, *35*(01), 185–205.

Schady, N. (2011). Parents' education, mothers' vocabulary, and cognitive development in early childhood: longitudinal evidence from Ecuador. *American Journal of Public Health*, *101*(12), 2299–2307.

Snow, C. E. (1972). Mothers' Speech to Children Learning Language. *Child Development*, *43*(2), 549–565.

Sokolov, J. L. (1993). A local contingency analysis of the fine-tuning hypothesis. *Developmental Psychology*, *29*(6), 1008–1023.

Spivey, M. J., & Dale, R. (2006). Continuous Dynamics in Real-Time Cognition. *Current Directions in Psychological Science*, *15*(5), 207–211.

Thiessen, E. D., Hill, E. A., & Saffran, J. R. (2005). Infant-Directed Speech Facilitates Word Segmentation. *Infancy*, *7*(1), 53–71.

Walker, D., Greenwood, C., Hart, B., & Carta, J. (1994). Prediction of School Outcomes Based on Early Language Production and Socioeconomic Factors. *Children and Poverty*, *65*(2), 606–621.

Weisleder, A., & Fernald, A. (2013). Talking to Children Matters. *Psychological Science*, *24*(11), 2143–2152.

Yurovsky, D., Doyle, G., & Frank, M. C. (2016). Linguistic input is tuned to children's developmental level. In *Proceedings of the annual meeting of the cognitive science society* (pp. 2093–2098).

Nested Sets and Natural Frequencies

Stephen H. Dewitt[1], Anne Hsu[2], David Lagnado[1], Saoirse Connor Desai[3], Norman E. Fenton[2].

[1]Department of Experimental Psychology, University College London, 26 Bedford Way, WC1H 0AP
[2]School of Electronic Engineering and Computer Science, Queen Mary University of London, Mile End Rd, London E1 4NS
[3] Psychology Department, City University, London, EC1R 0JD

Abstract

Is the nested sets approach to improving accuracy on Bayesian word problems simply a way of prompting a natural frequencies solution, as its critics claim? Conversely, is it in fact, as its advocates claim, a more fundamental explanation of why the natural frequency approach itself works? Following recent calls, we use a process-focused approach to contribute to answering these long-debated questions. We also argue for a third, pragmatic way of looking at these two approaches and argue that they reveal different truths about human Bayesian reasoning. Using a think aloud methodology we show that while the nested sets approach does appear in part to work via the mechanisms theorised by advocates (by encouraging a nested sets representation), it also encourages conversion of the problem to frequencies, as its critics claim. The ramifications of these findings, as well as ways to further enhance the nested sets approach and train individuals to deal with standard probability problems are discussed.

Keywords: Nested Sets; Natural frequencies; Bayesian; Base rate neglect

A recent meta-analysis (McDowell & Jacobs, 2017) conclusively demonstrated that when a Bayesian word problem is presented according to natural frequency (NF) principles, normative responding increases relative to the 'standard probability' format (SP), with an average accuracy of around 24%. Both versions of the classic medical diagnosis problem can be seen below (statistical notation added).

Standard probability format (individual chance): The chance of breast cancer is 1% [P(Ca)] for women at age forty who participate in routine screening. If a woman has breast cancer, the chance is 80% [P(Po|Ca)] that she will get a positive mammography. If a woman does not have breast cancer, the chance is 9.6% [P(Po|¬Ca)] that she will also get a positive mammography. A woman in this age group had a positive mammography in routine screening. What is the chance that she actually has breast cancer [P(Ca|Po)]? ____%

Natural frequencies: 10 [F(Ca)] out of 1000 women at age forty who participate in routine screening have breast cancer. Out of the 10 women with breast cancer, 8 [F(Po&Ca)] will get a positive mammography. 95 [F(Po&¬Ca)] out of every 990 women without breast cancer will also get a positive mammography. Here is a new representative sample of women at age forty who got a positive mammography in routine screening. What proportion of these women do you expect to actually have breast cancer [P(Ca|Po)]? ___ %

We can see several differences between these formats. Most obviously, the NF format uses frequencies (indicated by the 'F' notation) rather than percentages / probabilities (P), but more importantly, the figures are not normalized. In the SP format, the figures are normalized by the use of a standard denominator (percentages are one way of achieving this with a hidden denominator of 100, but normalized frequencies with other denominators are also possible). This difference in normalization firstly has a known effect on the number of computations required to solve each problem. In an NF format there are thought to be only two computational steps[1]: (1) summing the number of individuals with a positive result and cancer F(Po&Ca) with the number of individuals with a positive result but no cancer F(Po&¬Ca) and then (2) dividing F(Po&Ca) by this sum. The same formula can be used if those same numbers are given in percentage or probability format.

$$\frac{F(Po\&Ca)}{F(Po\&Ca)+F(Po\&\neg Ca)} \; or \; \frac{P(Po\&Ca)}{P(Po\&Ca)+P(Po\&\neg Ca)}$$

However, normalized formats require an additional pre-step (you won't see any of these figures in the standard probability format to the left). P(Po&Ca) must itself first be calculated by multiplying the proportion of individuals with cancer who get a positive result (P[Po|Ca]) with the total proportion of individuals with cancer P(Ca). Similarly, P(Po&¬Ca) must be calculated by multiplying P(Po|¬Ca) with P(¬Ca). For example, to calculate the proportion of women without breast cancer and a positive result, we multiply the percentage of women without breast cancer (99%) by the percentage of those women who get a positive result (9.6%). This may be a trivial calculation for most, but crucially, the solver first has to have an accurate representation of the problem in order to know that we (A) need to calculate this figure to solve the problem and (B) should multiply these two particular values rather than using some other figures or operation to compute it.

As has been noted, in the natural frequency format, this figure is provided for us, which has widely been accepted as a potential confound by subverting the need for (A) entirely (however see Brase & Hill, 2015 for work suggesting this may not be an important factor). However, NF

[1] In fact, in some natural frequency versions, the final question is: 'How many of these women do you expect to actually have breast cancer? ___ out of ___' This reduces the computational steps further, to one only: calculating the total positives.

proponents (e.g. Gigerenzer and Hoffrage, 1995) tell the story the other way around: normalization is an artificial (and relatively recent) human construct which transforms problems from a natural and solvable format to an unnatural and difficult one. These authors propose that normalization adds an additional difficulty by changing the structure of the information from that which would be obtained through 'natural sampling' i.e. if we observed 1000 women one by one, taking note in each case whether they had cancer and whether they got a positive result. This information structure of the natural frequency format is thought to replicate the natural format that human beings experience in the world, and thus are predisposed in some way to work with, which is the true reason for the increased normative responding (Gigerenzer & Hoffrage, 1995).

One concrete change however is that when information is presented in this way, the denominator of F(Po&¬Ca) (990) matches F(¬Ca). Other authors (e.g. Evans, Handley, Perham, Over & Thompson, 2000; Sloman, Over, Slovak & Stibel, 2003) have therefore claimed that rather than this having anything to do with 'natural' formats, this simply makes the 'nested sets' structure of the problem transparent (e.g. that women with a positive mammography but no breast cancer are a subset of the larger group of women without breast cancer). Nested sets advocates argue that this set structure revelation should be considered the more ultimate cause. They have sought to demonstrate that any method which reveals the nested sets structure of the problem will be equally successful. One example, using normalized percentages for the false positive and negative rates like the SP format but framing these in terms of proportions of groups (PP) rather than individual chance (an approach developed by Macchi [2000]), can be seen below:

Nested Sets (Proportion Percentages): 10 F[Ca] out of 1000 women at age forty who participate in routine screening have breast cancer. Out of the women with breast cancer, 80% [P(Po|¬Ca)] will get a positive mammography. Out of those women without breast cancer, 9.6% [P(Po|¬Ca)] will also get a positive mammography. Here is a new representative sample of women at age forty who got a positive mammography in routine screening. What proportion of these women do you expect to actually have breast cancer [P(Ca|Po)]? ___ %

Macchi (2000) found an improvement in accuracy compared to an SP format, and no significant difference to an NF format. Following this and similar papers, NF proponents (Hoffrage, Gigerenzer, Krauss & Martignon, 2002) have argued that nested sets formats simply encourage solvers to construct an NF version of the problem for themselves, which is the ultimate reason for increased accuracy. This criticism seems all the more plausible for Macchi's format, given that unlike the standard probability format, it presented the base rate as a frequency. It is important to note however that Gigerenzer and Hoffrage (1995) originally theorized based on evolutionary grounds that the phenomena of neglecting base rates (P[Ca] and P[¬Ca]) during solution should generalize to non-NF formats because that information is not required for solution in an NF format, which people are adapted to:

"Base rate information need not be attended to in frequency formats (Result 3). If our evolutionary argument that cognitive algorithms were designed for frequency information acquired through natural sampling is valid, then base rate neglect may come naturally when generalizing to other information representations, such as the standard probability format (Gigerenzer & Hoffrage, 1995, pp. 29)

While the authors refer specifically to the standard probability format here, the key point is that in evolutionary history humans have never had to complete the 'pre-step' required in the normalized format, because information has always been presented to them in the natural frequency format (and in which they can compute the normative answer without using the base rates), and so they may lack the capacity to do this, regardless of whether that normalized format is presented in the SP way, or in Macchi's PP way. The simple fact that nested sets results defy this has been widely overlooked in the field, and in fact suggests a potential harmony between the two approaches, rather than a discord, at least at the pragmatic level. While people do indeed seem more capable of solving a Bayesian word problem in a natural frequency format, than in a standard normalized format, nested sets results show us that, with the right framing, people can solve normalized Bayesian problems too.

A preliminary aim of this paper is to replicate Macchi's approach, as it has only been demonstrated in a single experiment. Furthermore, it needs replication in a wider range of more ecologically valid situations, including with the base rate presented as a percentage (as mentioned, Macchi's original format used a frequency base rate unlike the SP format) and with non-whole numbers. These factors may be present in real-world contexts and may add sufficient complexity to undermine the value of the format. We also aim to test the format in both simple (all women with breast cancer get a positive result) and hard (some women with breast cancer get a false negative) problems as both versions have been used widely in the literature.

A more ambitious aim of this paper is to assist in settling the highly debated connection between nested sets and natural frequency formats. Over the past few years repeated calls have been made to resolve these differences between the two camps (Brase & Hill, 2015; McNair, 2015; Johnson & Tubau, 2015; McDowell & Jacobs, 2017). Given that these are fundamental questions about cognitive process, the same authors have repeatedly called for more process-focused experiments. While two previous experiments (Gigerenzer & Hoffrage 1995; Macchi, 2000) used a 'think aloud' (TA) approach (where participants record their thought processes while solving the problem) in both cases

this was only used to report the types of errors participants make. We aim to make greater use of this data to shed light on the following questions. Does the nested sets approach work, as claimed by its advocates, by encouraging a representation of e.g. P(Po&¬Ca) as a subset of P(¬Ca) at the first, de-normalization step? Does the nested sets approach encourage individuals to construct a natural frequency representation for themselves, as claimed by Hoffrage et al. (2002)? Which of these are predictive of success on the problem? Finally, what else can we learn about the mechanisms by which Macchi's nested sets approach achieves greater accuracy?

Method

521 participants were recruited through Amazon MTurk (55.3% female; mean age = 34.2 [SD = 11.6]). The experiment had eight between-subjects conditions, using a 2 (standard probability [SP] vs proportion percentages [PP]) x 2 (simple vs hard) x 2 (whole vs decimal) design. The PP-hard-decimal condition can be seen below (with statistical notation, not shown to participants), and further materials and experimental data are available at https://osf.io/nd46g/. This is considered a decimal version because the product of computational step 1 (e.g. 10% x 76% = 7.6%) is a non-whole number.

Every year the government advises women to take part in routine mammography screening using an X-ray machine to determine if they have breast cancer. Among women at age forty who participte in this routine screening 10% [P(Ca)] have breast cancer, while 90% [P(¬Ca)] do not. However, the screening test is not always accurate. Specifically, out of those women who have breast cancer, only 76% [P(Po|Ca)] will actually get a positive mammography. Furthermore, out of all of those women who do not have breast cancer, 15% [P(Po|¬Ca)] will also get a positive mammography. What percentage of women at age forty who get a positive mammography [P(Po)] in routine screening actually have breast cancer[P(Ca|Po)]? ___%

Participants were also required to record their thought process in an open text box. They could only submit their numerical response after they had submitted their thought process. All qualitative analysis of the TA data was undertaken blind to condition. Analysis was coded by two authors separately, with over 90% agreement. Discrepancies were resolved through the decision of a third coder.

Participants were given a 'normative' label if their numerical response was within 1% of the Bayesian normative value. Beyond this however, we found seven participants who clearly demonstrated accurate reasoning, including all necessary computational steps, but made an arithmetic error. These participants were also labelled as normative. One of these participants was in the nested sets conditions, while six were in the standard probability conditions.

Results

General Results

The overall proportion of the sample providing the normative response for the experiment was 13.5% with an average of 9.0% for the SP conditions and 18.1% for the PP conditions. In Figure 1, normative proportions for all eight conditions can be seen.

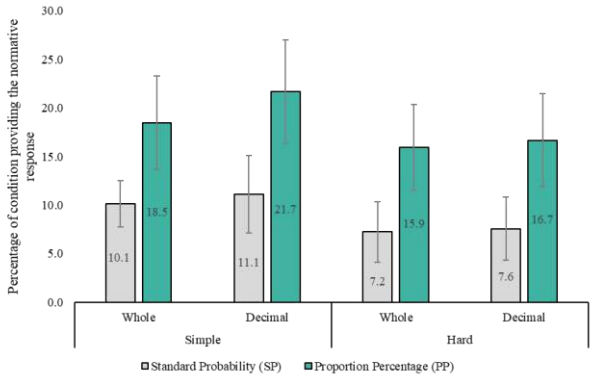


Figure 1. The percentage of participants providing the normative Bayesian answers across all eight conditions. Error bars represent one standard error.

A binary logistic regression (BLR) using 'normative response' as the dependent variable and the three condition-comparisons (SP vs PP; whole vs decimal; simple vs hard) as independent variables found a main effect for the SP-PP comparison (Wald $X^2 = 8.984$, p=.003), no main effect for the whole-decimal comparison (Wald $X^2 = .184$, p=.668) and no main effect for the simple-hard comparison (Wald $X^2 = 1.350$, p=.245). All subsequent analyses on 'condition' therefore compare SP to PP only.

Nested Sets Representation

Across all conditions, 87 (16.7%) individuals expressed a 'nested sets representation' of the problem. For this classification, participants had to explicitly, in words, depict the group of individuals who had both a positive test result but not cancer (P[Po&¬Ca]) as a subset of the total individuals without cancer (P[¬Ca]). In the hard condition, they also had to express the group of individuals who had a positive result and cancer (P[Po&Ca]) as a subset of the individuals with cancer (P[Ca]). A mathematical formula was not sufficient to be assigned this code. An example comes from P261 who stated, "Of the 90% who do not have cancer, 15% will get a positive mammography". Here we can see a word-based representation of the individuals who do not have cancer but got a positive test result as a subset of those who do not have cancer. This classification was applied conservatively. For example, P498 who said "First what is 15% of 90%, that is 13.5%" did not receive the classification. An example from the hard condition which did get this

classification comes from P138 who said "We know 10% of women will have breast cancer in the screen and 80% of those will show up positive […] Of the remaining 90 women who do not have breast cancer 10% will be given a false positive so an additional 9 women."

A BLR showed that this representation was unsurprisingly more common within the PP (24.0%) condition, which expressed the problem in this format, than in the SP (9.7%) condition (Wald $X^2 = 18.0$, $p<.001$), which used an individual chance format. However clearly some individuals in the SP condition re-represented the problem in terms of nested sets. Furthermore, in both conditions, this representation was highly associated with normativity, as can be seen in Table 1.

A BLR was run with normativity as DV, and condition and NS-representation as IV's, and a unique predictive effect of NS-representation (Wald $X^2 = 123.6$, $p<.001$) was found, but no unique effect of condition (Wald $X^2 = 0.04$, $p=0.837$).

Conversion to frequencies

Across all conditions, 87 participants (16.7%) also converted the base rate in the problem from a percentage into a frequency before attempting solution (i.e. before providing an NS-representation or completing the first computational step). For this classification, a 'sample' or 'population' of individuals as a frequency rather than a percentage or probability had to be expressed. For example, P105 said 'To make my math easier, I am going to assume there are 100 women.' and P186 began 'Out of 100 women, 10 have breast cancer, while 90 do not.' Out of the 87 participants who converted the problem to whole numbers, 73 converted to a population of 100 women. The number of individuals who made this conversion in each condition, crossed with those providing the NS-representation and the proportion of these subgroups providing the normative response can be seen in Table 1. A BLR with conversion as DV and condition as IV showed a predictive effect (Wald $X^2 = 7.3$, $p=.007$). A BLR with normative response as DV and condition and conversion as IV's showed a unique effect of conversion (Wald $X^2 = 128.9$, $p<.001$) and a potential unique effect of condition (Wald $X^2 = 5.2$, $p = 0.041$).

To simultaneously test the impact of condition, NS-representation and conversion upon normativity, a BLR was run. No main effect of condition was seen (Wald $X^2 = 0.172$, $p=0.68$), but a unique effect of NS-representation (Wald $X^2 = 93.2$, $p<.001$) and of conversion (Wald $X^2 = 8.3$, $p=0.004$) was seen. A table depicting these relationships can be seen below.

Table 1. Percentage of individuals providing the normative answer organized by condition, NS-representation and conversion (total number of individuals in each subgroup regardless of normativity in parentheses).

		Standard Probability			Proportion Percent		
		No NS-representation	NS-representation	Total	No NS-representation	NS-representation	Total
No-Conversion		1.4 (221)	69.2 (13)	5.1 (234)	2.8 (176)	45.8 (24)	8.0 (200)
Conversion		5.0 (20)	84.6 (13)	36.4 (33)	5.9 (17)	78.4 (37)	55.6 (54)
Total		1.7 (241)	76.9 (26)	(267)	3.1 (193)	65.6 (61)	(254)

From the raw data, we can see that in the absence of the NS-representation, conversion only appears to be associated with a small (~3%) increase in normativity, while in the presence of the NS-representation, converting appears to be associated with a much larger (~15-30%) increase. To check this, we ran two BLR's, predicting normativity from conversion. Within those who did not produce an NS-representation, no predictive relationship was seen (Wald $X^2 = 0.81$, $p=0.21$) while within those who did produce an NS-representation, a predictive relationship was seen (Wald $X^2 = 6.4$, $p=0.011$). Dependency of this sort was not seen for the NS-representation, which was a significant predictor of normativity among those who did not convert (Wald $X^2 = 69.3$ $p<.001$) as well as those who converted (Wald $X^2 = 27.6$, $p<.001$). For some individuals their process could not be determined (e.g. if they just provided a mathematical formula) but a few individuals were able to solve the problem without converting and also while apparently using a chance representation, such as P40:

"There is a 10% chance that any woman over 40 has breast cancer [and] there is a 10% chance that a woman who does not have breast cancer over 40 gets a positive result. This means there is a 9% chance of [a false positive] and a 19% chance that someone tests positive for breast cancer. Out of this there is a 10/19% chance that the diagnosis is correct meaning there is a 52.63% chance."

Errors

The most common error within the SP condition (21.7%) was to provide the complement of the false positive rate, (1-P[Po|¬Ca]). This was much less common within the PP condition (5.5%). The TA data was coded for insight into common reasoning and a single piece of reasoning was highly prominent (45.8% of cases). This was the confusion of P(Po|¬Ca) with P(¬Ca|Po). Following this confusion, the subsequent accurate

deduction was made that 100% minus this value would give P(Ca|Po). For example, P228 said 'The fact that 15% of positive mammographies are invalid means that 85% are valid. She therefore has an 85% chance of actually having breast cancer', P20 said 'I guess since 10% of positive tests are inaccurate, that means there's a 90% chance of her having cancer' and P133 said 'Also of all the women who get a positive mammogram, 15% will not have breast cancer, so I think it is 85%.' Each of these participants use language reflecting P(¬Ca|Po) but accompanying the percentage value representing P(Po|¬Ca), strongly suggesting a confusion between the two. P177 expressed this confusion more explicitly, saying 'But there is a 10 percent chance that a woman without breast cancer will get a positive mammogram [true, P(Po|¬Ca)], so 10 percent of the positive mammograms are not accurate [false, P(¬Ca|Po)]'. In the remainder of these participants' TA data, the reasoning could not be extracted from the data. For example, many participants simply provided mathematical notation.

Computational Steps

A cumulative graph depicting the proportion of individuals reporting each of the three computational steps, step 1 the calculation of P(Po&Ca) and P(Po&¬Ca), step 2 the summing of these and step 3 the division of P(Po&Ca) by the sum as well as whether the participant provided the normative numerical value can be seen below for both conditions. For both conditions, the majority of individuals do not achieve step 1, with further substantial but smaller drop-off between this and step 2, and no substantial subsequent drop-off between these and step 3 or the normative response. In short, highly similar curves were seen for both the SP and PP conditions. The major difference between the two conditions was the number of individuals reporting step 1 (with more individuals reporting this in the PP condition). Similar proportional drop-off was subsequently seen in both conditions. Indeed, while condition was predictive of step 1 (Wald X^2 = 15.3, p<.001), when controlling for step 1, condition was not predictive of step 2 (Wald X^2 = 0.19, p=.891), step 3 (Wald X^2 = .988, p=.320) or the normative response (Wald X^2 = 0.076, p=.783).

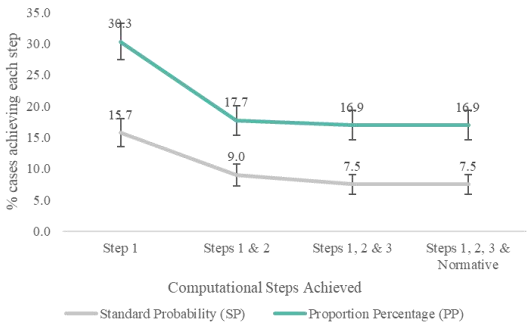


Figure 2. Drop-off graph for each computational step. Error bars represent one standard error.

Discussion

We replicated Macchi's (2000) finding in a larger sample, and across a range of different format types, including with percentage base rates with and without the possibility of false negatives and with whole numbers and non-whole numbers. In each case, Macchi's proportion percentage format improved normativity over and above the SP format, with an overall increase from 9.0% to 18.1%.

We found that normativity is highly associated with the individual reporting a representation of P(Po&¬Ca) as a subset of P(¬Ca) in their think aloud data, and in the hard condition, also P(Po&Ca) as a subset of P(Ca). This finding is not surprising within the proportion percentage group, as it could be argued that these individuals are simply regurgitating the text from the problem. However, crucially, this relationship also held within the standard probability format, where an 'individual chance' probability format (i.e. 'If a woman has cancer, her chance of …) was presented. This observational finding should also be considered in the context of previous experiments (e.g. Evans et al, 2000; Sloman et al, 2003) showing that attempts to assist individuals in creating exactly this representation of the problem have been successful in increasing accuracy. Here we show that some individuals, without any prompt to do this, spontaneously adopt this representation, and this correlates highly with normativity. We also found some evidence that the NS-representation may have a mediating effect on the impact of the NS format. This provides some complementary evidence to those papers that the mechanism by which nested sets formats achieve greater accuracy is at least partially that which they have espoused: by encouraging a nested sets representation of the structure of the problem.

We also found that many individuals make a further spontaneous re-representation of the problem, and that this also correlates highly with normativity. This is the conversion of the problem from a percentage format into a frequency format. Interestingly, conversion alone seemed not to be predictive of normativity, however in combination with the NS-representation it was associated with higher rates of normativity than the NS-representation alone. The same was not true of the NS-representation. This was still highly predictive of normativity with or without conversion. Importantly, the majority of individuals who converted did so to a base of 100, making no mathematical change to the problem. This therefore seems to demonstrate a preference among our sample for working with frequency values over percentages, even when the absolute numbers (e.g. 20% vs 20 women out of 100) and therefore calculations, are identical. Of course, we cannot resolve the ultimate reason for this, be that a greater evolutionary exposure towards frequencies or a current greater exposure to frequencies during our participants' lives. We tentatively suggest a third

possibility. It may be difficult to mentally represent a percentage, abstract as it is, without it being a percentage of something tangible. Imagining 100 women may simply provide a concrete mental image which can be divided and sub-divided according to the percentages. It may also provide a platform for a simple internal narrative about these women and what happens to them. Whatever the ultimate reason however, this result does partially confirm Hoffrage et al's (2002) conjecture.

These findings have some relevance to the question of whether the elements that are thought to comprise the natural frequency format are separable, and if so, which elements are doing the 'work' in improving accuracy. Nested sets advocates have argued that the nested sets structure is doing all the work, and the frequencies are superfluous. Natural frequency advocates have argued that the two are inseparable. Here we find some tentative evidence that the two are separable (individuals who form a nested sets representation but do not convert to frequencies are still more successful than those who do not form that representation). However, even if separable, both the nested sets structure, and the use of frequencies (as opposed to percentages) appear to uniquely contribute to success, with the combination of both being more strongly associated with success than either alone. Importantly, without the nested sets structure, conversion to frequencies did not predict success, which may mirror findings that normalized frequency formats are no better than the standard probability format (e.g. Evans et al., 2000).

In terms of further investigation into the mechanisms of Macchi's nested sets format, we presented evidence that relative to the SP format, more individuals achieve step 1 (de-normalization). However, controlling for this, the proportion of participants achieving subsequent steps is not different to the SP format. Related to this, an analysis of errors between conditions has shown that the classic 1-P(Po|¬Ca) error was drastically reduced from 21.7% of total responses in the SP format to 5.5% in the PP format. This error, in line with previous theorizing (e.g. Braine and Connell, 1990) has been found here to principally stem from a confusion between the false positive rate P(Po|¬Ca) and P(¬Ca|Po). As has been mentioned, the clarification of the false positive rate (and the true positive rate in the hard condition) by encouraging individuals to see it as a subset of P(¬Ca) has long been theorized to be the mechanism by which nested sets formats work. The reduction of this error in the PP condition therefore seems to further support this theory. Given that the false positive rate is required for step 1, it also provides further evidence that the impact of Macchi's format is principally achieved at this step.

As noted, Macchi's format does not appear, upon the current evidence, to provide any additional support in the later stages of solution, most notably in getting from computational step 1 to step 2. At this step individuals need to recognize (A) that they require the total number of positive results, and (B) that they need to combine the false positives with the true positives to achieve this. So far, research has been principally focused on helping solvers form a representation of e.g. P(Po&¬Ca) as a subset of P(¬Ca). However, success on the final two steps may instead be a product of recognizing a different set relation, that of P(Ca&Po) and P(¬Ca&Po) as subsets of P(Po). We can clarify this distinction by displaying two tree structures of the medical diagnosis problem below. The top shows the classic structure, widely published, with the hypothesis, 'Cancer' as the first 'division', or first set of child nodes. However, the opposite structure is also possible, shown at the bottom, with the data, 'Positive' as the first set of child nodes.

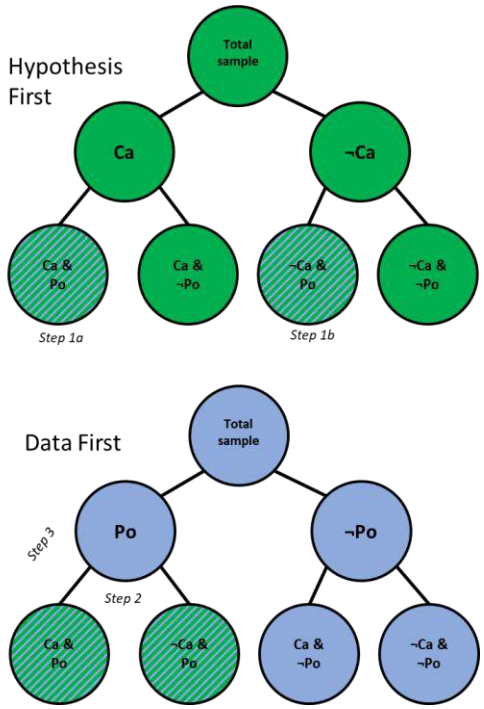


Figure 3. 'Hypothesis First' and 'Data First' tree diagram representations of the medical diagnosis problem.

While perceiving the set relations in the hypothesis-first version seems key to step 1, steps 2 (calculating total positives) and 3 (dividing cancer & positive by total positives) would seem to require an understanding of the set relations in (at least the left half) of the data-first diagram. For step 2, the addition, one must understand that P(Ca&Po) and P(¬Ca&Po) are subsets of P(Po). It seems to us that step 3, the division, should require only that same set relation i.e. that P(Ca&Po) is a subset of P(Po). To our knowledge this distinction has not been made before. We believe that in order to improve framing methods further, focus should be on helping individuals form these latter set representations at the most appropriate time to facilitate steps 2 and 3.

In the medical diagnosis problem, the information related to steps 2 and 3 are contained within the question. In

our nested sets format, this is changed into a proportion form i.e. *'What percentage of women at age forty who get a positive mammography…*', unlike the SP format, which is chance framed. While plausibly this could have helped solvers form exactly this latter subset representation, the current evidence suggests this did not have an impact. Future work may look to combine Macchi's format with a question form used by Girotto and Gonzalez (2001) which was divided into two parts: first explicitly requiring the calculation of step 2, and only then requiring calculation of step 3.

Finally, it should be noted, that the accuracy percentage for participants in our NS group was lower than the average from the recent meta-analysis for natural frequency (~24%). It is difficult of course to make confident comparisons but given that we have found that the nested sets approach works via very similar mechanisms to the natural frequency approach, but requires one extra step (de-normalization), and in some versions two extra steps, and furthermore that we have found a unique beneficial effect of frequencies, some greater accuracy on natural frequency versions seems plausible to us. Pragmatically therefore we would still advocate for natural frequencies as the primary method for communicating Bayesian problems to the public where that is possible, with proportion percentages as a backup where it is not.

However, unfortunately when individuals do encounter Bayesian problems in the real world, they are often in the standard probability format. Sedlmeier & Gigerenzer (2001) have investigated the merits of preparing individuals via training to convert these into natural frequency versions themselves when they encounter them. This however requires considerable training. Our findings suggest that solvers can do more of the work themselves than was assumed by that research (i.e. can de-normalize the problem themselves) and therefore may only need to remember fewer 'conversion' steps. This may be valuable where the brevity of the training is important. In fact, our findings tentatively suggest substantial accuracy gains may be obtained by training people to following two simple rules when faced with an SP problem:

1. Imagine 100 women (or whatever unit you're dealing with).
2. Imagine the percentages you've been given as proportions of these 100 women.

Acknowledgements

Funding was in part provided by the ERC project ERC-2013-AdG339182-BAYES_KNOWLEDGE and the Leverhulme Trust project RPG-2016-118 CAUSAL-DYNAMICS.

References

Brase, G., & Hill, W. (2015). Good fences make for good neighbors but bad science: a review of what improves Bayesian reasoning and why. *Frontiers in Psychology*, 6, 340.

Braine, M & Connell, J. (1990). Is the base rate fallacy an instance of asserting the consequent?. *Lines of thinking: Reflections on the psychology of thought*, 1, 165-180.

Evans, J., Handley, S., Perham, N., Over, D., & Thompson, V. (2000). Frequency versus probability formats in statistical word problems. *Cognition*, 77, 197–213.

Gigerenzer, G. & Hoffrage, U. (1995). How to Improve Bayesian Reasoning Without Instruction: Frequency Formats. *Psychological Review*, 102(4), 684–704.

Girotto, V. & Gonzalez, M. (2001). Solving probabilistic and statistical problems: A matter of information structure and question form. *Cognition*, 78(3), 247–276.

Hoffrage, U., Gigerenzer, G., Krauss, S., & Martignon, L. (2002). Representation facilitates reasoning: What natural frequencies are and what they are not. *Cognition*, 84, 343–352.

Johnson, E. & Tubau, E. (2013). Words, numbers, & numeracy: Diminishing individual differences in Bayesian reasoning. *Learning and Individual Differences*, 28, 34–40.

Macchi, L. (2000). Partitive Formulation of Information in Probabilistic Problems: Beyond Heuristics and Frequency Format Explanations. *Organizational behavior and human decision processes*, 82(2), 217–236.

McDowell, M., & Jacobs, P. (2017). Meta-analysis of the effect of natural frequencies on Bayesian reasoning. *Psychological Bulletin*, 143(12), 1273–1312.

McNair, S. (2015). Beyond the status-quo: research on Bayesian reasoning must develop in both theory and method. *Frontiers in Psychology*, 6, 1–3.

Sedlmeier, P., & Gigerenzer, G. (2001). Teaching Bayesian reasoning in less than two hours. *Journal of Experimental Psychology: General*, 130(3), 380–400.

Sloman, S., Over, D., Slovak, L., & Stibel, J. (2003). Frequency illusions and other fallacies. *Organizational Behavior and Human Decision Processes*, 91(2), 296–309.

Predicting Bias in the Evaluation of Unlabeled Political Arguments

Nicholas Diana (ndiana@cmu.edu)
Human-Computer Interaction Institute
Carnegie Mellon University, Pittsburgh, PA, USA

John Stamper (john@stamper.org)
Human-Computer Interaction Institute
Carnegie Mellon University, Pittsburgh, PA, USA

Kenneth Koedinger (koedinger@cmu.edu)
Human-Computer Interaction Institute
Carnegie Mellon University, Pittsburgh, PA, USA

Abstract

While many solutions to the apparent civic online reasoning deficit have been put forth, few consider how reasoning is often moderated by the dynamic relationship between the user's values and the values latent in the online content they are consuming. The current experiment leverages Moral Foundations Theory and Distributed Dictionary Representations to develop a method for measuring the alignment between an individual's values and the values latent in text content. This new measure of alignment was predictive of bias in an argument evaluation task, such that higher alignment was associated with higher ratings of argument strength. Finally, we discuss how these results support the development of adaptive interventions that could provide real-time feedback when an individual may be most susceptible to bias.

Keywords: myside bias; moral foundations theory; distributed dictionary representations; civic reasoning

Introduction

The rise of social media has been accompanied by a rise in smaller, decentralized media sources. One clear negative consequence of this democratization of media has been an increase in access to unreliable or misleading news stories. Despite their lack of credibility and veracity, these stories are persuasive and appealing. Some estimates suggest that Americans fall for fake news headlines approximately 75% of the time (Silverman & Singer-Vine, 2016), and that these stories are generally more engaging than stories produced by traditional news outlets (Silverman, 2016).

The proposed solutions to these problems generally fall into two categories. The first category leverages various machine learning methods (Conroy, Rubin, & Chen, 2015) to create "fake news detectors." While some of these classifiers are quite sophisticated (Wang et al., 2018), these detectors tend to limit their scope to the detection of stories that are patently false. More nuanced instances of stories that are merely misleading are generally beyond the purview (and perhaps ability) of these systems (McGrew, Ortega, Breakstone, & Wineburg, 2017). Moreover, even if accurate classification was possible, one might question whether it is in our best interest to delegate this task to machines, potentially allowing our own ability to critically evaluate media sources to languish in the process.

In contrast to the content-driven detectors, other solutions focus on improving the critical thinking skills of the media consumers themselves. There is certainly evidence of a deficit in this regard. A recent study of students in middle school, high school, and college summarized the student's "civic online reasoning" (e.g. evaluating arguments, recognizing spin) as simply, "bleak" (Wineburg, McGrew, Breakstone, & Ortega, 2016). Non-detector solutions tend to focus on strengthening these kinds civic reasoning skills. For example, *Factitious* is a game created by the American University Game Lab (Hone, Rice, Brown, & Farley, 2018) that is marketed as a way to test the player's ability to distinguish fake and real news stories, but along the way teaches the player to identify features like reliable sources and neutral language.

While the detectors focus on the media content itself (hoping to fill the role of editor in the new democratized news space), the civic education solutions focus instead on the media consumers, with the hope that better critical thinkers might be more or less immune to the appeal of misleading content. Both of these approaches unfortunately tend to neglect the dynamic relationship between the media content and the media consumer. Consider the following actual fake news headline:

> "Pope Francis Shocks World, Endorses Donald Trump for President"

If you happen to be a religious Trump supporter, this story may seem plausible. After all, if you, a person of faith, have found reason to endorse him, why shouldn't another person of faith. This headline confirms what you already believe to be true. However, if you are not a Trump supporter, this headline might raise several red flags. It is in that wave of skepticism that you may dart your eyes to the URL in order to check the credibility of the source. Because this headline runs counter to your beliefs, you go searching for evidence to disprove it. In either case, the degree of critical thought that is brought to bear on the content is, at least to some extent, dependent on the values and beliefs of the reader.

This tendency to evaluate arguments more favorably when they align with your own views or beliefs (and conversely, more critically when they do not) is formally known as *myside bias* (Baron, 2000). Numerous studies have shown the effects of myside bias on reasoning to be reliable and

strong (Klaczynski & Robinson, 2000; Stanovich & West, 1997), irrespective of intelligence (Stanovich & West, 2007; Stanovich, West, & Toplak, 2013). Haidt's Social Intuitionist Model of moral reasoning (Haidt, 2001) suggests that the power of myside bias is likely due to the fact that the primary drivers of our moral judgments are intuitions and heuristics. According to the Social Intuitionist Model, when we encounter a new piece of information, we have an immediate and powerful intuition about whether we agree or disagree with the information. Haidt argues that the vast majority of these judgments are made automatically, using Kahneman's (Kahneman & Egan, 2011) System 1 thinking. Rational (or System 2) thinking always comes after an intuitive judgment has already been made, and only comes online if we are asked to justify our position. In short, we are not, by default, the rational thinkers we think we are. Moreover, when we do make use of our capacity for rational thought, it's generally to justify a decision we have already made, not to search for the truth.

Misleading and false news stories can exploit this vulnerability by designing stories that strongly align with the prior-held beliefs of the target audience. Because the reader wants to believe the story is true (to affirm their reality), System 2's critical reasoning skills are never engaged. The bias literature suggests that overcoming the strength of this intuitive appeal may require more than detecting falsehoods or training consumers to be more critical. Solutions that ignore the dynamic relationship between the user's beliefs and the beliefs latent in the misleading media content are perhaps ignoring the very feature that makes the target content so powerful.

Accurately capturing user beliefs is a daunting challenge. Each user likely possesses countless individual beliefs, and new beliefs are constantly being created in response to their current political context. One solution is to instead measure the foundational values that inform our beliefs. Moral Foundations Theory (Haidt & Graham, 2007) argues that moral decision making can be traced to a small set of foundational values (Care, Fairness, Loyalty, Authority, and Sanctity). These moral foundations have been empirically shown to be highly predictive of both general voting behavior (Franks & Scherr, 2015) as well as more specific political beliefs (e.g., "Climate change is real") (Koleva, Graham, Iyer, Ditto, & Haidt, 2012; Rottman, Kelemen, & Young, 2014). Moral Foundations Theory allows us to approximate beliefs in a theory-driven, context-general way. This is crucial for any solution intended for deployment on the internet, where the number of unique-contexts is virtually infinite.

After deriving a measure of user values (as a proxy for beliefs), the system must also be able to estimate the values latent in the text they are reading. Recently, Garten et al. (2018) developed a methodology for estimating the values latent in short pieces of text (tweets), and demonstrated that their methodology can accurately classify a tweet's most salient moral foundation (as measured by human raters). What remains to be seen is if value classification methods (like Garten's) can be combined with measures of user values to estimate the degree of alignment between the values of the media consumer and the content they are consuming.

We would expect that any measure that accurately captures this relationship between consumer and content should also be able to predict the presence of myside bias. That is, when alignment between user and content values is high, we expect that the user will be biased to evaluate the content more favorably. In this paper, we propose a method for measuring this dynamic relationship between consumer and content values, and demonstrate that the resulting metric can be used to predict bias in argument evaluation. Specifically, we test whether the alignment between participants' values and the values latent in an argument predicts participants' ratings of argument strength. We hypothesized that higher alignment between participant and content values will be associated with higher ratings of argument strength, and that this relationship will be present even in arguments specifically designed to confuse our natural language processing method.

Practically, this methodology lays the groundwork for future hybrid solutions that leverage technology alongside human critical thought to mitigate the impact of content designed to confirm our values rather than disseminate true information. Perhaps more importantly, this methodology presents a novel, context-independent way to estimate the impact of myside bias, a known, powerful moderator of everyday reasoning.

Methods

Participants

Eighty (80) participants were recruited using the participant recruitment platform Prolific. Participants were required to be between 18-65 years of age, U.S. citizens, and not have participated in any of our research group's prior studies (due to content similarity). Five participants exited the study before completing any significant portion of the main experiment and were excluded from analyses. The estimated completion time (based on prior pilot studies) was approximately 15 minutes, and participants were paid $2.50 ($10/hour) for participating.

Data Quality We mitigated the impact of potential gamers in several ways. First, the post-test questionnaire included a reading-check question. Participants who failed the reading-check (n=7) were excluded from analyses. We also used timing data to identify participants who were likely clicking through the problems without reading the prompts. Specifically, participants who selected an answer less than two seconds after a prompt loads (roughly the time needed to select an answer after the page loads), at least 10 times (for half of the problems), are labeled as gamers. We chose a threshold of 10 problems for two reasons: 1) it is reasonable to assume that participants who begin the experiment with the intention of gaming the system will exhibit this behavior for at least half of the problems, and 2) if we set this threshold too low, we risk excluding participants who begin the experiment with

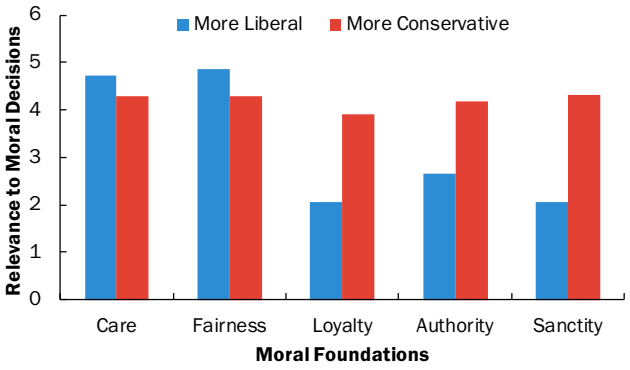


Figure 1: Relevance to Moral Decisions by Moral Foundation for more conservative and more liberal participants. These values closely match previously observed values for liberals and conservatives (Haidt & Graham, 2007), suggesting that our sample was politically diverse.

good intentions, but get fatigued towards the end. Eight participants met this criteria for gaming, and were excluded from analyses.

Demographics Of the remaining 60 participants, 38 identified as male, 20 as female, and 2 as other. Participants ranged in age from 18-62 years old (M=31.10). With respect to race and ethnicity, 50 participants identified as Caucasian, 6 as Hispanic, 3 as Black, and 1 as Asian. The majority (59%) of participants reported having completed a college level education or higher, and a high number of participants reported completing a master's degree (n=19).

Political Diversity One of the key benefits of recruiting participants from Prolific is that participants are drawn from all over the country. If instead, we were to recruit participants from our local community, we would likely get an unbalanced distribution of political beliefs (as our city has a history of voting overwhelmingly in favor of one party). Recruiting from across the country gives our sample a degree of political diversity that would be impossible to achieve otherwise.

To evaluate if our sample was indeed politically diverse, we used the composite measure of the Moral Foundations Theory Questionnaire (described below), called *progressivism*, to divide our sample into two groups along the mean score. Then, for each of the two groups we graphed the mean scores of each foundation and compared them to known averages. Figure 1 shows the mean scores for more liberal and more conservative participants across the five moral foundations. These values closely match previously observed values for liberals and conservatives (Haidt & Graham, 2007), suggesting that our sample was politically diverse.

Experiment Environment and Procedure

Participants completed the experiment online by navigating through a web-based application. After completing a consent form, participants were informed that the study consisted of two sections. In the first section, they were asked to complete a questionnaire (described below), and in the second section, they were asked to rate a series of arguments (presented in random order). After completing the two sections, the participants were directed to a post-test questionnaire where demographics information was collected, and then finally, to the debriefing page, which clarified that any facts and figures used in the study were entirely fictitious. The experiment was estimated to take approximately 15 minutes to complete (actual median completion time was 12 minutes).

Moral Foundations Theory Questionnaire In the first section of the experiment, each participant was required complete the Moral Foundations Questionnaire (MFQ) (Haidt & Graham, 2007). This 30-item questionnaire is designed measure how relevant each one of the five moral foundations (Care, Fairness, Authority, Loyalty, Sanctity) is to one's moral decision making. For example, participants are asked to indicate the degree to which they agree with the following statement: "Respect for authority is something all children need to learn." The final output of the questionnaire is a vector of five scores that indicate the relative importance of the five moral foundations to the participant's moral decision making. Ultimately, we are interested in constructing a model that relates the values latent in text to the values and beliefs of an individual person. This vector of five scores represents the human side of that relationship.

It is worth noting that having the participants take the MFQ before answering arguments designed to evoke moral decision making is not ideal. The questionnaire may cause participants to be more conscious of their beliefs than they might normally be if encountering these arguments in the real world. However, this ordering is unfortunately necessary for later stages of this project, where adaptive interventions designed to promote analytic thinking use an individual's scores on the MFQ to decide when targeted feedback is needed most. These future directions are explored in more detail in the *Discussion* section.

Rating the Strength of Arguments Participants were asked to read and rate the strength of 20 arguments on a nine-point Likert scale (1=Very Weak; 9=Very Strong). Each argument had three key features. First, each argument was designed to evoke a specific moral foundation. For example, the following argument was designed to evoke the *Authority* foundation:

> Greenville School District requires students to address all adults as "Sir" or "Ma'am" and their students always score higher on state tests than ours. Instilling a strong respect for authority for their teachers helps students learn.

Regardless of the argument's actual strength, we would expect that if a participant believes that respecting authority is important, this argument will resonate with them. Each of the five foundations is the focus of an argument four times, for a

total of 20 arguments.

The second key feature is the relative quality of an argument. This is a categorical feature with two levels, *high quality* and *low quality*. The above argument is an example of a *low quality* argument. In contrast, consider the following argument:

> The number of suspensions at Redbridge School District has been slowly increasing for the past 5 years. Last year they added three police officers to their staff and saw a 10% decline in suspensions. The presence of a strong authority figure reduces bad behavior.

While this argument is certainly not airtight, it has several attributes that make it a relatively higher quality argument. First, it shows the reversal of a long-term trend, in contrast to the *low quality* argument where no temporal context is established. Second, it uses concrete figures that are relative to the norm, as opposed to the *low quality* argument which uses vague terms like "higher" to quantify changes. In general, high quality arguments include information that can be used to rule out some alternative explanations. Low quality arguments leave open the possibility of alternative explanations. Of the 20 arguments, half are *high quality* and half are *low quality*.

The third key feature is *congruence* with the target foundation. A potential limitation of the distributed dictionary representation methodology (described below) is that statement representations are formed using the representations of single words. This means that, while this methodology should have no problem knowing that the word "son" in the context of the word "king" likely refers to the concept "prince," it will likely have more difficulty identifying the cultural nuances between statements like "God is good" and "God is dead." The *congruence* feature is designed to test the robustness of this methodology's ability to adapt to these kinds of unfavorable circumstances. Consider again the two previous example arguments. Both arguments 1) use language that evokes the authority foundation, and 2) are supportive of that foundation. In contrast, consider the following argument:

> Woodford School District doesn't allow teachers to reprimand students, and last year they had fewer detentions than our district. Students behave better when they're treated like equals instead of children

While this argument also evokes the Authority foundation, this example argues against an increased respect for authority. We would expect that participants that value authority will be more skeptical of the claims in this argument, because they violate their intuitions. Whether the model's representation of the values latent in the argument is nuanced enough to make the distinction between *incongruent* and *congruent* arguments is an open question. Again, half (10) of the arguments are *congruent*, half *incongruent*.

Analysis

Distributed Dictionary Representations The broad goal of this experiment is to evaluate a method for comparing an individual's values with the values latent in the media they are consuming. Using the MFQ, we are able to generate a theory-driven estimation of the participant's values. The next step is identifying the values latent in a particular piece of text. While historically this has been done using word-frequency methods (i.e., counting the number of times terms in a concept dictionary appear in the target text), these methods are much less effective for analyzing small bodies of text (e.g., news headlines, tweets), which may not contain any of the dictionary terms.

In contrast to word-frequency methods, distributed representations (Mikolov, Chen, Corrado, & Dean, 2013) estimate the meaning of words by comparing the numerous, varied contexts that the word appears in within a large text corpus. These models are rooted in the distributional hypothesis, which states that words that appear in similar contexts likely share some semantic features. The distributed representation of a word is simply that word's location in a low-dimensional (10-10,000 dimensions) space. This location can be represented as a vector, which allows us to compute the semantic distance between two concepts using cosine similarity.

Garten et al. (2018) extends this work in distributed representations to incorporate concept dictionaries. A distributed dictionary representation is computed by simply averaging the distributed representations of all the words in the dictionary. The result is a point in the semantic space that amplifies the shared, core features of each of the component dictionary terms. Because we are ultimately using an abstract representation of a concept, our dictionaries can be highly focused, including only the most relevant terms. The current study uses five such dictionaries (one per moral foundation), and each dictionary contains four positive words (e.g., fairness, equality) and four negative words (e.g., unfair, injustice) related to the moral foundation (e.g., fairness). Using distributed representations allows for the effective analysis of small bodies of text (such as the short arguments used in the current study), because the method does not require any of the dictionary terms to be present in the text. We used gensim (a Python implementation of Word2Vec (Mikolov et al., 2013)) and the pre-trained Google News corpus (approximately 100 billion words) Word2Vec model[1].

Alignment The output of the distributed dictionary representation analysis is a vector of five values, indicating the average semantic distance between the words in the argument and the words in each of the five moral foundation concept dictionaries. To compute *alignment*, we compute the cosine similarity between this vector and the vector of moral foundation relevance scores outputted by the Moral Foundations Questionnaire (i.e., the participant's values). We then used a

[1]The pre-trained Google News model can be found here: https://code.google.com/p/word2vec/

normalized log-transformation to correct for skew. *Alignment* is computed for each participant and argument combination.

Linear Mixed Effects Models Because it is impossible to determine an objective rating of argument strength for the arguments used in this study, we are less interested in the individual rating of each argument and more interested in how a participant rates arguments relative to one another (i.e., high alignment vs. low alignment or high quality vs. low quality). To make use of all the data while accounting for differences in ratings across participants, we use a series of linear mixed effects models, with *participant* as a random effect. Similarly, to control for unintended variations in argument strength, we include *Argument ID* as an additional random effect. We compare models to one another using the Akaike information criteria (AIC), which estimates the fit of a model (lower scores are better). All reported coefficients are standardized.

Results

We used a series of mixed effects models to examine the relationship between *alignment* (between participant and argument values) and ratings of argument strength. To reduce the possibility that any effect of *alignment* on ratings could be attributed to differences in demographics, we tested for collinearity between *alignment* and each demographic variable (*age*, *gender*[2], *race*, and *education level*). A series of one-way analyses of variance (ANOVA) between *alignment* and each of the three categorical variables (*gender*, *race*, and *education level*) showed no evidence of collinearity. Similarly, there was no significant correlation between *alignment* and *age*.

A mixed effects model with *participant* and *argument ID* as random effects, ratings of strength as the outcome variable, and *gender*, *race*, *educational level*, and *alignment* as fixed effects was generated[3]. *Alignment* was a significant predictor of ratings when included alongside these demographics variables ($\beta = 2.48$, $p < 0.01$), providing further evidence that any effect of *alignment* on ratings was not due to differences in demographics.

Impact of Alignment when Controlling for Quality To test if *alignment*'s impact persists when controlling for argument quality, we built a mixed effects model with *participant* and *argument ID* as random effects, ratings of strength as the outcome variable, and *alignment* and *argument quality* as fixed effects. We found that *alignment* was predictive of ratings despite the presence of *argument quality*. It should be noted that although participants on average rated high quality arguments as significantly stronger ($t(59) = 8.07$, $p < .001$) than low quality arguments ($M = 5.06$, $SD = 1.72$) (suggesting some categorical validity), the labels "high" and "low" are very much subjective labels. As such, we cannot objectively compare the impact of *alignment* to the impact of quality. Still, we can make a meaningful, subjective comparison between the impact of *alignment* and "quality" (as operationally defined in this context). In this context, the impact of *alignment* on ratings of strength ($\beta = 3.06$, $p < 0.001$) was greater than the impact of *argument quality* ($\beta = 1.33$, $p < 0.01$).

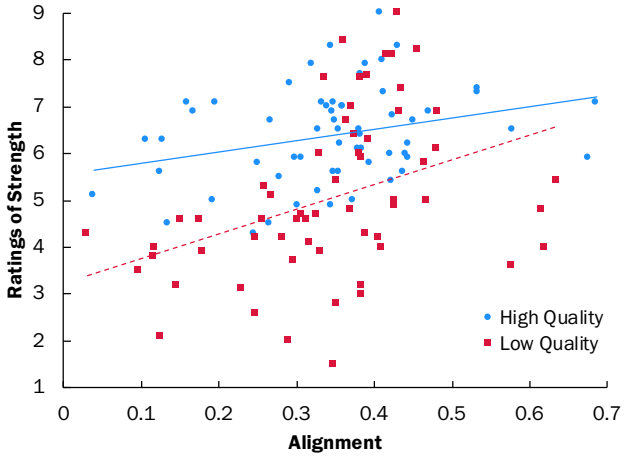


Figure 2: Relative impact of alignment on the ratings of high and low quality arguments. Each data point represents the average rating and alignment for all arguments within a category (high or low quality) for one participant. On average, participants rated high quality arguments as stronger than low quality arguments. The ratings of both types of arguments were associated with alignment scores.

While on average, participants rated congruent problems ($M = 5.89$, $SD = 1.35$) as significantly stronger ($df(59) = 2.27$, $p = 0.02$) than incongruent problems ($M = 5.57$, $SD = 1.40$), *congruence* was not a significant predictor of ratings when added to this model.

Interaction between Age and Alignment Previous research suggests that, because reliance on heuristic reasoning increases with age, older adults may be more likely to exhibit biases in everyday reasoning (Klaczynski & Robinson, 2000). To test whether this was true of our sample, we built a mixed effects model with *participant* and *argument ID* as random effects, ratings of strength as the outcome variable, and *argument quality* and *alignment*age* as fixed effects (where *alignment*age* is an interaction term). We found that there was a significant interaction between *alignment* and *age* ($\beta = 15.01$, $p < 0.001$), such that *alignment*'s impact increases as *age* increases. This finding aligns with previous research. Additionally, this *alignment*age* interaction model had a better fit (AIC=5033.05) than the previous model built without the interaction term (AIC=5058.63).

[2]Participants identifying as "other" were excluded from this analysis because the sample was very small (n=2).

[3]Note that age was excluded from the model because a one-way ANOVA between *age* and *education level* indicated a significant relationship between *age* and *education level*. A Tukey's HSD test showed that participants at the graduate level ($M = 35.33$, $SD = 5.38$) were significantly older than those below the college level ($M = 27.96$, $SD = 9.05$). A likelihood ratio test showed that *education level* was more explanatory than *age*, so *age* was excluded in favor of *education level*.

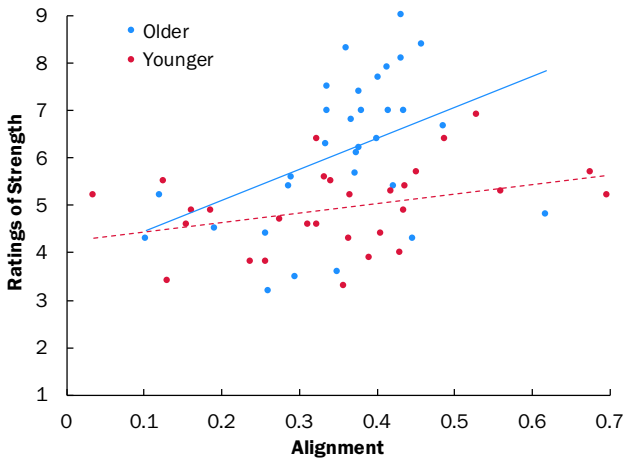


Figure 3: The interaction between age and alignment. Each data point represents one participant's average rating and alignment scores. Alignment had a much larger impact on ratings of strength for older participants (participants above the median age) than younger participants. This conforms with previous findings examining the relationship between bias in argument evaluation and age.

Performance on Incongruent Problems A potential limitation of this particular NLP method is its reliance on the semantic relationships between isolated words. A robust methodology should be able to accurately determine the valence of an argument that may contain several words related to a foundation, but nonetheless is incongruent with the beliefs of someone who values that foundation. To test the robustness of our method, we built another iteration of the above, best performing mixed effects model (including the *alignment*age* interaction), but selected only *incongruent* arguments (previously both *congruent* and *incongruent* problems were used). The impact of *alignment* on ratings of *incongruent* arguments also appears to be dependent on *age*, as the interaction term *alignment*age* was again a significant predictor of ratings of argument strength ($\beta = 15.01$, $p < 0.001$). To examine this relationship further, we divided the sample into two groups (older and younger) along the mean age, and then calculated the correlation between participants' mean *ratings* and mean *alignment* for each group. While we found a significant correlation between *ratings* and *alignment* in the older group ($r = 0.26$, $p < 0.001$), we found no such correlation in the younger group (see Figure 3).

Discussion

Our results demonstrate that distributed dictionary representations (DDR) combined with a measure of user values may provide a reliable method for identifying when users may be prone to biased reasoning. Because our method does not require labeled text data, it can be easily applied to real-world data (such as social media posts). The only limitation on this front is the identification of the user's values. We do this formally with the Moral Foundations Questionnaire, but research has demonstrated that political orientations can be predicted with a high degree of accuracy purely based off of social media activity (Colleoni, Rozza, & Arvidsson, 2014). Whether these predictions are as nuanced as those generated by the theoretically grounded Moral Foundations remains to be seen, but the potential for a fully automated method for measuring a user's susceptibility to myside bias exists.

We used *incongruent* problems to test the robustness of our methodology. These problems were intentionally designed to confuse the DDR method, and produced some interesting results. While alignment was predictive of ratings on incongruent problems, this was only true for older participants. One potential explanation for this difference is a lack of clarity about what low scores on the moral foundations questionnaire mean (specifically in this context as a proxy for beliefs). High scores indicate a value is relevant, but do low scores indicate indifference or impassioned opposition? Future work will require a qualitative exploration of these nuances.

Toward Adaptive Interventions

Our results suggest that we can leverage the dynamic relationship between user and content values to predict when the user may be prone to biased reasoning. This work is the first step toward providing adaptive, targeted interventions when high alignment between user and content values is detected (i.e., when the user is most prone to biased reasoning). It is in these cases of high alignment that we are least likely to move from System 1 (intuitive) to System 2 (rational) thinking, and engage the reasoning processes that may mitigate bias. Adaptive interventions may be able to facilitate the engagement of System 2 thinking in exactly these critical moments, making users less vulnerable to content designed to exploit natural biases. This kind of hybrid solution leverages the strength of sophisticated machine learning methods, while still preserving the need for and power of human reasoning.

Conclusion

In this paper, we demonstrated that a measure of alignment between a participant's values and the values latent in short arguments was a significant predictor of ratings of argument strength. This was true even for nuanced arguments, designed to confuse our methodology. These results underscore the impact of values on the evaluation of everyday arguments, and lay the groundwork for adaptive interventions designed to mitigate everyday reasoning biases.

Acknowledgements

The research reported here was supported, in whole or in part, by the Institute of Education Sciences, U.S. Department of Education, through grant R305B150008 to Carnegie Mellon University. The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education.

References

Baron, J. (2000). *Thinking and deciding*. Cambridge University Press.

Colleoni, E., Rozza, A., & Arvidsson, A. (2014). Echo chamber or public sphere? predicting political orientation and measuring political homophily in twitter using big data. *Journal of Communication*, *64*(2), 317–332.

Conroy, N. J., Rubin, V. L., & Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. In *Proceedings of the 78th asis&t annual meeting: Information science with impact: Research in and for the community* (p. 82).

Franks, A. S., & Scherr, K. C. (2015). Using moral foundations to predict voting behavior: Regression models from the 2012 us presidential election. *Analyses of Social Issues and Public Policy*, *15*(1), 213–232.

Garten, J., Hoover, J., Johnson, K. M., Boghrati, R., Iskiwitch, C., & Dehghani, M. (2018). Dictionaries and distributions: Combining expert knowledge and large scale textual data content analysis. *Behavior research methods*, *50*(1), 344–361.

Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological review*, *108*(4), 814.

Haidt, J., & Graham, J. (2007). When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research*, *20*(1), 98–116. doi: 10.1007/s11211-007-0034-z

Hone, B., Rice, J., Brown, C., & Farley, M. (2018). Factitious. Retrieved from `factitious.augamestudio.com`

Kahneman, D., & Egan, P. (2011). *Thinking, fast and slow* (Vol. 1). Farrar, Straus and Giroux New York.

Klaczynski, P. A., & Robinson, B. (2000). Personal theories, intellectual ability, and epistemological beliefs: Adult age differences in everyday reasoning biases. *Psychology and Aging*, *15*(3), 400.

Koleva, S. P., Graham, J., Iyer, R., Ditto, P. H., & Haidt, J. (2012). Tracing the threads: How five moral concerns (especially purity) help explain culture war attitudes. *Journal of Research in Personality*, *46*(2), 184–194.

McGrew, S., Ortega, T., Breakstone, J., & Wineburg, S. (2017). The challenge that's bigger than fake news: Civic reasoning in a social media environment. *American Educator*, *41*(3), 4.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Rottman, J., Kelemen, D., & Young, L. (2014). Tainting the soul: Purity concerns predict moral judgments of suicide. *Cognition*, *130*(2), 217–226.

Silverman, C. (2016, Nov). *This analysis shows how viral fake election news stories outperformed real news on facebook.* Retrieved from `https://www.buzzfeednews.com/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook`

Silverman, C., & Singer-Vine, J. (2016, Dec). *Most americans who see fake news believe it, new survey says.* Retrieved from `http://www.buzzfeed.com/craigsilverman/fake-news-survey`

Stanovich, K. E., & West, R. F. (1997). Reasoning independently of prior belief and individual differences in actively open-minded thinking. *Journal of Educational Psychology*, *89*(2), 342.

Stanovich, K. E., & West, R. F. (2007). Natural myside bias is independent of cognitive ability. *Thinking & Reasoning*, *13*(3), 225–247.

Stanovich, K. E., West, R. F., & Toplak, M. E. (2013). Myside bias, rational thinking, and intelligence. *Current Directions in Psychological Science*, *22*(4), 259–264.

Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., ... Gao, J. (2018). Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining* (pp. 849–857).

Wineburg, S., McGrew, S., Breakstone, J., & Ortega, T. (2016). Evaluating information: The cornerstone of civic online reasoning. *Stanford Digital Repository.*.

Building blocks of computational thinking: Young children's developing capacities for problem decomposition

Griffin Dietz [1], **James A. Landay**[1], **Hyowon Gweon**[2]
{dietz, landay, hyo}@stanford.edu
[1]Computer Science Department, Stanford University
[2]Department of Psychology, Stanford University

Abstract

Computational thinking (CT) refers to a range of problem-solving skills applicable to computer science and everyday life. Although recent research in developmental cognitive science suggests mental capacities relevant to CT may emerge quite early in life, research on CT, and computer science education more generally, has made little contact with this literature. As a way to better bridge these fields, we explore the development of problem decomposition, a critical feature of CT, in the spatial domain. We ask whether young children can break a complex spatial problem down into subcomponents that can be reassembled to solve the overarching problem. Across two experiments (Exp.1: 4- to 7-year-olds; Exp.2: 3- to 5-year-olds) that involve constructing block structures, we demonstrate that some of the key capacities underlying problem decomposition begin to emerge in preschool years and develop throughout early childhood. Although preschool-aged children struggle to solve an open-ended decomposition problem that requires generation and execution of decomposition plans, even 4-year-olds can successfully evaluate the viability of these plans. These results suggest that experimental methods in developmental cognitive science can inform CS education research that focuses on promoting CT; by identifying when and how CT concepts emerge in early childhood, we can better create age-appropriate educational tools.

Keywords: computational thinking; problem decomposition; problem solving; cognitive development; intuitive physics

Introduction

The ability to break down a large problem into smaller parts is important for many real-world tasks. To decompose a problem effectively, one must understand its constraints, generate potential solutions, and evaluate the strengths and weaknesses of those solutions. Importantly, these steps are often better taken *before* one actually acts; attempts to achieve a complex task without proper planning can lead to unnecessary effort to correct a mistake or even irreversible failure. But what does it take to be good at problem-solving and planning?

More than a decade ago, Wing (2006) popularized the concept of computational thinking (henceforth CT). CT is a term that collectively refers to a range of skills that are crucial to effective problem-solving, and it incorporates various cognitive strategies considered fundamental to computer science (CS) (Wing, 2006; Barr & Stephenson, 2011; Brennan & Resnick, 2012). Mental activities like abstraction (i.e., generalizing problem features to preserve only relevant information; Kramer, 2007) and problem decomposition (i.e., breaking a complex problem into solvable subcomponents; Barr & Stephenson, 2011) are key components of CT. Indeed, these skills are critical to building good computer programs; anyone who has engaged in programming understands the importance of abstracting away from a problem to identify its basic structure and decomposing that structure into solvable parts.

Yet, the importance of CT reaches far beyond programming (Wing, 2006). Abstract thinking, problem decomposition, and the ability to evaluate potential plans are skills that allow us to tackle a range of everyday tasks as well as larger, more complex problems that involve multiple sub-goals, such as conducting scientific research or building an architectural project. In particular, to successfully achieve these larger goals, one must: (1) represent the current state of the world (i.e., what does the empty lot look like?, what materials do we have?) as well as the state of the desired end-goal (i.e., what do I want to build?), (2) identify the units that comprise the end goal (i.e., what sub-goals should I complete?) and construct the possible future states from applying these units (i.e., what will the structure look like given these components?), and (3) evaluate the viability and effectiveness of different sets of potential units and interventions (i.e., should we build the columns or the roof first?, which size columns are most suitable?). In other words, effective problem-solving involves the representational and inferential abilities to *generate* possible ways to decompose the problem space and *evaluate* the viability of a potential decomposition plan. By engaging in these mental processes prior to executing a given plan, one can solve a problem with less trial-and-error.

While CT has been a useful construct to raise awareness of the relevance of these skills in both computing and everyday life, it remains a difficult concept to operationalize or measure. This difficulty may arise from the fact that CT is not a single thing; it is a collection of various mental operations whose cognitive mechanisms are poorly understood. Furthermore, although CT presumably involves reasoning abilities that have been topics of interest in cognitive development research, this body of work has remained rather disconnected from the literature in CS education, leading many CS educators to believe that CT develops relatively late in childhood (Guzdial, 2015). Our goal is to take a step towards synthesizing these fields, and build on prior work to ask whether the ability to decompose a complex problem—a key component of CT—is present early in life. In the following, we summarize related work on young children's inferential capacities and introduce a novel task for testing problem decomposition.

Prior work in cognitive development has revealed rich, sophisticated abilities in young children to engage in abstract reasoning and learning (Gopnik, 2012; Schulz, 2012). Although these studies are primarily aimed at identifying the developmental origins of the human ability to engage in sci-

entific thinking, collectively their findings suggest that the basic representational and inferential capacities supporting CT may emerge much earlier than previously thought. For instance, preschool-aged children construct novel hypotheses from observations via inductive generalization and design novel experiments to test these hypotheses by engaging in selection and isolation of relevant variables (Cook, Goodman, & Schulz, 2011; Legare, 2012). These abilities are foundational to successful problem solving.

Furthermore, CT involves the understanding that good plans achieve a goal effectively and efficiently. Evidence suggests the rapid development of planning abilities between ages 4 and 6, including an increase in the number of steps children can plan ahead to solve a problem (Klahr & Robinson, 1981) and improvements in the ability to deploy appropriate strategies depending on the task (Gardner & Rogoff, 1990). Prior work has also shown that even infants expect rational agents to act in ways that minimize cost (Gergely, Nádasdy, Csibra, & Bíró, 1995; Scott & Baillargeon, 2013), and they infer the reward an agent assigns to a goal based on the cost incurred to achieve it (Liu, Ullman, Tenenbaum, & Spelke, 2017). By age 5, children can even design informative experiments to infer the subjective costs or rewards of achieving a goal (i.e., an agent's competence or preferences) by systematically manipulating the objective rewards or costs of completing a task (Jara-Ettinger, Gweon, Tenenbaum, & Schulz, 2015; see Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016 for a review). Collectively, these early-emerging capacities to generate and test hypotheses, engage in advance planning, and reason about efficiency suggest that the basic aspects of CT may emerge earlier than commonly believed.

Building on this prior literature, we designed a novel block-building task to study one of the key components of CT: problem decomposition. Block-building tasks are familiar to young children, and have historically been considered a useful domain for studying the development of planning and problem-solving. Block construction has been shown to be an indicator of early spatial skills (e.g., mental rotation, Brosnan, 1998), which correlate highly with later success in programming and STEM (Cooper, Wang, Israni, & Sorby, 2015; Verdine et al., 2014; Wai, Lubinski, & Benbow, 2009). Thus, studying children's ability to generate and execute an effective block-building plan can provide a unique window into understanding the early development of CT. Yet, prior work is largely limited to exploring childrens bottom-up building processes, allowing them to build the target structure in a piecemeal manner. Whether or not children can engage in top-down problem decomposition remains an open question.

A key strength of our task is that it requires more than merely copying a model block structure: children must figure out a viable plan within the constraints of the task by decomposing the structure into appropriate parts. In simple block-building tasks, one might succeed by accumulating raw materials (i.e., individual blocks) in a piecemeal fashion. Similar to the ways beavers or birds build their dams or nests, a child could repeatedly stack blocks to create a tower. However, imagine a child wants to build a structure resembling the bridge in Figure 1. Simply accumulating individual blocks isn't sufficient; the child must first assemble the "legs," and then place a horizontal bar on top. If a child starts by creating "pillars" that are as tall as the bridge itself (3 blocks), then a single block in the middle would not stay in place. This example demonstrates how a bottom-up building process can be insufficient even for seemingly simple tasks. Rather, this problem resembles the way that we approach larger, real-world engineering projects; we must take the desired goal, break it into smaller problems, and determine how those components should be solved and assembled within the constraints of the task. Thus, the goal in designing our task was to provide a context in which children would approach a complex problem in a similar manner under clearly defined task parameters.

Recent work demonstrates that both adults and children leverage intuitive physics when evaluating the stability of block structures (Battaglia, Hamrick, & Tenenbaum, 2013; Kamps et al., 2017; Yildirim, Gerstenberg, Saeed, Toussaint, & Tenenbaum, 2017), that children as young as 4 can gauge the difficulty of building such structures (Gweon, Asaba, & Bennett-Pierre, 2017), and that they are capable of copying a model block structure when given the required pieces (Cortesa et al., 2018). Critically however, start-to-end construction requires intelligently *generating* those pieces as well. Computational models optimized to generate instructions for the construction of block structures identify structural components while accounting for the effect of gravity on future layers (Zhang, Igarashi, Kanamori, & Mitani, 2017). However, the ability to determine the required subcomponents based on an intuitive understanding of task-specific constraints has not been tested in young children, even though such ability might provide the key foundation for a more general ability to engage in problem decomposition.

In Experiment 1, we embedded the process of *generating*, *evaluating*, and *executing* an appropriate decomposition plan into a fun, engaging block-building task. Given a target structure, children had to identify the underlying substructures, simulate ahead to determine if those substructures could combine into a self-supporting building, and then execute this plan to complete the task. In Experiment 2, we use a simplified version of the task to ask whether young children's difficulty in Experiment 1 comes from the process of *generating* a plan with the appropriate subcomponents, rather than the process of *evaluating* the viability of a given plan by engaging in physical simulation.

Experiment 1

Methods

Participants A total of 112 children (Age: 4.00–7.99) were recruited from a local children's museum and a university-affiliated preschool (38 4-year-olds: $M = 4.56$ (4.04–4.99); 31 5-year-olds: $M = 5.43$ (5.01–5.96); 23 6-year-olds: $M = 6.49$ (6.14–6.99); 20 7-year-olds: $M = 7.61$ (7.08–7.99)). They

were randomly assigned to either the Standing Bridge condition (N=62) or the Sideways Bridge condition (N=50). We planned to recruit at least 40 children in each condition who successfully completed the task (10-12 in each age group). Twenty-nine children were unable to accomplish the task, and the successful subgroup included N=42 in the Standing Bridge condition and N=41 in the Sideways Bridge condition. An additional 20 participants were dropped from analyses because they: (1) did not speak English (N=3), (2) ended the study early (N=7), (3) failed the warm-up task (N=6, see Procedure), or (4) the experimenter made an error (N=4).

Stimuli For the main test trial, the model bridge was comprised of seven one-inch wooden cubes (three across the top and two on each side as supports) painted metallic silver. In the Standing Bridge condition, this bridge was presented upright, and in the Sideways Bridge condition, it was lying down (see Fig. 1). Children were given 3 individual unpainted wooden blocks with which they could create substructures using a "magic box." The magic box (see Fig. 1) was a cardboard box covered in felt with a small coin slot and an output window. A wire connected this box to the "construction zone" (a flat piece of foam core covered in black felt) and a large plastic button, suggesting they were all part of one causal mechanism. A coin was required each time the child operated the magic box to create a substructure. Inside the box were pre-assembled metallic structures. For the main task, where children had 3 individual blocks, there were 4 possible configurations of shapes that children could make; we prepared 4 metallic structures for each shape for a total of 16. We included additional structures for the practice trials.

Procedure Children were introduced to the experimenter's "magic box" which could turn a set of one-inch wooden blocks into larger, metallic blocks of varying shapes. The child had to first build a structure on the construction zone using individual wooden blocks; after putting a coin in the slot, children could press the plastic button to generate a single metallic block that had the same shape as the structure on the construction zone. In reality, when the child operated the magic box, the experimenter surreptitiously reached into the box through a hidden opening, found the corresponding metallic block, and placed it in the magic box output window (see Fig. 1). From the child's view, this created the illusion that the child's button press operated the magic box to generate the metallic structures.

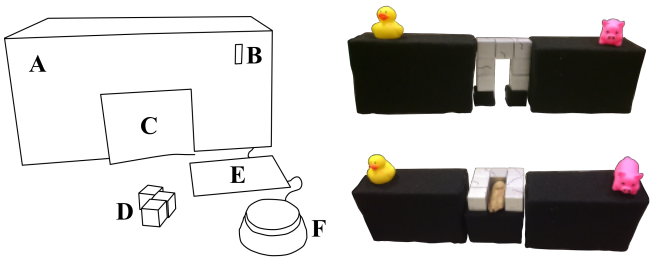


Figure 1: Left: Experimental setup. A) magic box, B) coin slot, C) output window, D) 1" wooden blocks, E) construction zone, F) plastic button; Right: Standing bridge (top) and sideways bridge (bottom).

A brief warm-up task ensured that the child understood the purpose of the main task and how to operate the magic box. In the warm-up, the child was given four wooden blocks, and was asked to build a 4-block 'T' shape (Trial A), a 3-block 'L' shape (Trial B), a single block (Trial C), and an 8-block cube (2x2x2) composed of two 4-block squares (Trial D). We excluded children who failed to complete this pretrial task from subsequent analysis to ensure that all children included in the study understood the magic box paradigm and were able to use the magic box to build metallic block pieces.

In the main task, children were asked to build a bridge using the blocks and the magic box. Critically, children had only three wooden blocks such that the metallic blocks they could build using the magic box was limited to a particular set of shapes. They were also given 11 coins; this limited the number of possible times children could generate a metallic structure, providing a pressure to solve the task efficiently.

In the Standing Bridge condition, the upright bridge was subject to the forces of gravity, and thus required a specific block set and assembly sequence (i.e., set up two 2-block bases and place a 3-block horizontal bar on top). The Sideways Bridge condition was identical to the Standing Bridge condition except that the bridge was laid flat (and thus not subject to the force of gravity). While the task still forced children to decompose the structure, there were multiple possible solutions and the order of construction did not matter. Thus, the Sideways Bridge condition still required the ability to follow task instructions, create parts, and assemble the final structure. However, the need to engage in advance planning to generate the "correct" decomposition plan and evaluate its viability was not as critical for success.

The children were given up to 10 minutes to build the bridge, after which the experimenter stepped in to help and ended the study. Often the child got stuck (signaled by asking for help or a period of inactivity) or distracted, so to encourage the child to reengage, the experimenter offered one of two pre-scripted prompts. Additionally, after an extended period of inactivity or running out of coins, children were given the option of restarting the task in the remaining time.

Results and Discussion

This was an exploratory study to see whether children could engage in effective problem decomposition, rather than a test of a priori hypotheses. However, we could imagine seeing a few general trends in the data. First, we expected that children would become more successful and more efficient at completing the task with age. We measured efficiency using two different metrics: completion time (in seconds) and number of coins used (3 was the minimum). Second, independent of increasing performance with age, we also expected that children would perform better (i.e., higher success rate, as well

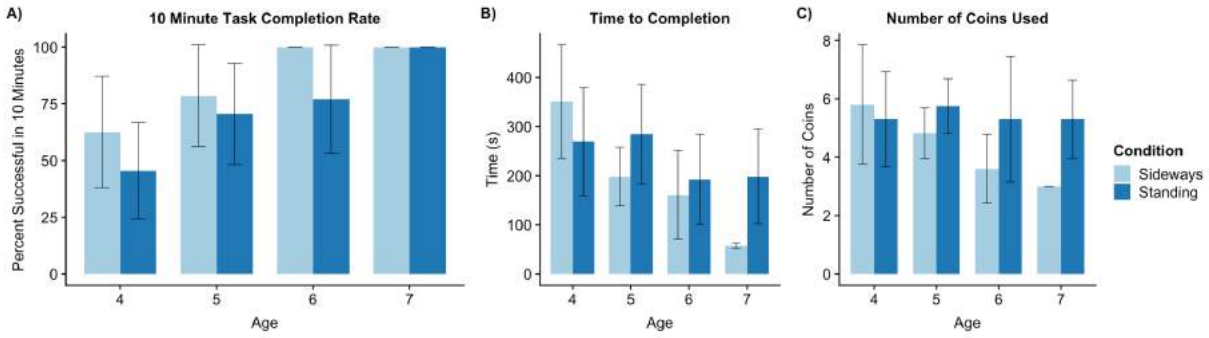


Figure 2: (A) Success rate in each age bin and in each condition. (B-C) Time-to-completion (B) and number of coins used to complete the task (C) among the successful children, split by age and condition. Error bars: bootstrapped 95% CI.

as more efficient completion) in the Sideways Bridge condition than the Standing Bridge condition, because the Standing Bridge had just one viable 3-part decomposition solution. Third, we predicted an age by condition interaction; the condition difference would decrease with age as children become more proficient at finding solutions without trial and error.

A logistic regression with condition (discrete) and age (continuous) confirmed a relationship between age and success rate ($z = 2.41$, $p = 0.02$). However, we did not see a significant effect of condition ($z = 0.47$, $p = 0.64$) nor an interaction between age and condition ($z = -0.73$, $p = 0.47$).

Given the increase in success rate with age, we further analyzed data from the 83 successful children to see if children become more efficient problem-solvers with age. First, we looked at time-to-completion; a linear regression with both age (continuous) and condition (discrete) as predictors showed that older children take a shorter amount of time to complete the task ($t = -4.47$, $p < .001$). While children in the Sideways Bridge condition did not complete the task faster than those in the Standing Bridge condition ($t = -1.73$, $p = 0.09$), we did find an interaction between age and condition. However, the effect was in the opposite direction than we had initially predicted: the difference in completion time between conditions increased with age ($t = 2.06$, $p = 0.04$).

Another measure of efficiency—the total number of coins used—also showed a similar pattern. A linear regression on the 83 children who successfully completed the task revealed that the number of coins children used to complete the building task decreased with age ($t = -3.14$, $p = .002$); also consistent with time-to-completion, we did not find an effect of condition ($t = -1.49$, $p = .14$) but the difference between conditions increased with age ($t = 1.96$, $p = .05$).

We then looked at the proportion of children who completed the task with maximal efficiency (i.e., successfully building the bridge using just 3 coins). A logistic regression with condition and age showed an effect of age ($z = 3.61$, $p < .001$). Children were also more likely to perform optimally in the Sideways Bridge than in the Standing Bridge condition ($z = 2.08$, $p = 0.04$), and this tendency increased with age (age by condition interaction, $z = -2.54$, $p = 0.01$).

Overall, data from this exploratory study showed a few notable patterns. First, unsurprisingly, children became more successful and more efficient at solving the task with age across a number of measures: success rate, time-to-completion, and number of coins used to finish the task. These results are consistent with prior work showing that the ability to plan ahead to solve problems develops rapidly during preschool years. Second, we also found that the proportion of children who finished the task with maximal efficiency varied across conditions. This pattern is also reasonable given that the Standing Bridge required a more principled, planned approach for success; due to the constraint of gravity, there was only one viable decomposition solution whereas the Sideways Bridge could be built in a few different ways. Third, we found an age by condition interaction in measures of efficiency (time-to-completion and number of coins used); however, the difference between conditions in efficiency increased with age, rather than decreasing with age. In other words, only the older children showed the expected difference between conditions. This suggests that the task was generally quite difficult for young children; even though 4- and 5-year-olds still successfully passed several practice trials and understood the task instructions, many of them struggled to complete the task in both conditions.

Collectively, these data provide an informative window into how children engage in problem decomposition to solve a complex task. Older children's near-perfect performance in the Sideways Bridge condition suggests that they were able to create substructures and use them to assemble the bridge correctly. Thus, the primary challenge children faced in this task may have been identifying and generating a plan to construct the "correct" components prior to building, especially when there was just one solution (Standing Bridge condition).

The fact that younger children struggled in both conditions raises questions about whether preschool-aged children suffer from a genuine lack of ability to engage in problem decomposition. However, the results do not allow us to directly explore this possibility because the task in Experiment 1 was open-ended and required children to engage in all aspects of problem decomposition—generating, evaluating, and ex-

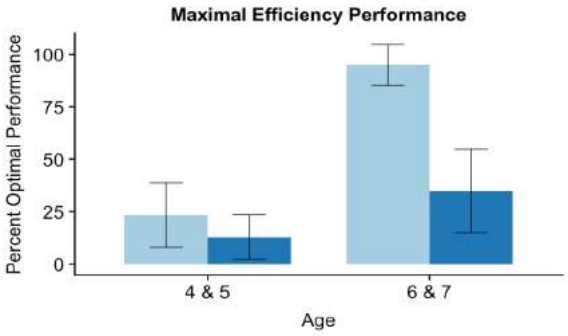


Figure 3: Percentage of children who completed the task with maximal efficiency (only used 3 coins), collapsed into age bins. Error bars are bootstrapped 95% CI.

ecuting solutions. However, there are reasons to believe that, when the demands for generating the plans are removed, even younger children can successfully *evaluate* the viability of a given plan. Compared to the process of generating a plan for decomposing a problem or a structure, evaluating the viability of an existing plan is arguably an easier task. Prior work suggests that preschool-aged children can easily assemble structures (Cortesa et al., 2018) and evaluate the relative difficulty of building different structures (Gweon et al., 2017), suggesting that even though the younger children in Experiment 1 (4- and 5-year-olds) struggled to generate and execute the plans themselves, they may be capable of evaluating the viability of existing decomposition plans.

In Experiment 2, we test this hypothesis with a simple binary-choice paradigm where we asked children to choose one of two pre-generated plans (i.e., choose the plan that would result in a self-supporting structure). Given the simplicity of the task, in addition to 4- and 5-year-olds, who we expected would succeed, we also tested 3-year-olds; while we did not have strong a priori predictions regarding the 3-year-olds' performance, having a broader age group would allow us to capture the developmental trajectory of this ability.

Experiment 2

Methods

Participants A total of 78 children were recruited from a local children's museum and a university-affiliated preschool (28 3-year-olds: M = 3.51 (3.02–3.98); 26 4-year-olds: M = 4.42 (4.01–4.93); 24 5-year-olds: M = 5.52 (5.03–5.93)). An additional 12 children were dropped from analyses because: (1) they did not speak English (N=1), (2) they did not complete the study (N=4), (3) they failed the pretrial task (N=3, see Procedure), (4) parents interfered (N=1), or (5) the experimenter made an error (N=3).

Stimuli Stimuli were similar to the blocks structures used in Experiment 1. A 'T' shaped structure (practice trial) and the upright bridge from Exp. 1 (main trial) were used as target structures. For both trials we prepared two sets of blocks, preconfigured in the shape of the target structure and laid flat on the surface. Critically, only one of the two sets would result in the correct self-supporting structure (see Fig. 4).

Procedure Children were introduced to block pieces of various shapes. In the practice trial, the experimenter presented the T-shaped block structure along with two potential solutions, and asked: "Can you help me build a new building that looks just like this one and can stand up all by itself? We can use these blocks (pointing to one set) or these blocks (pointing to the other set). Only one will work."

After the child selected one set of blocks, the experimenter allowed the child to use the selected blocks to construct the target structure. Regardless of whether or not the child chose correctly, the experimenter allowed the child to attempt construction with the other set. After the child succeeded in constructing the structure with the correct block set and failed with the incorrect set, the experimenter reiterated that the child could build the building with one set of blocks but not with the other, as it would fall over, so only one set would work. We excluded children who failed to select one of the block options in this pretrial task or who began playing with the blocks before listening to the full explanation.

In the main task, the children were asked to choose one of two solutions that would result in a standing bridge. We marked a child as having made a selection when they physically picked up one of the sets of blocks. The position of the correct solution (L/R) was counterbalanced across subjects.

Results and Discussion

We first ran a logistic regression on children's choice with age as a continuous variable. The effect of age was trending towards significance ($z = 1.81$, $p = 0.07$). We then looked at each age group separately. Three-year-olds' responses did not differ from chance (M = 0.57, CI = 0.39–0.75), whereas four-year-olds (M = 0.81, CI = 0.65 – 0.92) and five-year-olds (M = 0.83, CI = 0.67 – 0.96) showed robust success.

To succeed at this task, a child had to be able to understand the constraints applied to the problem (gravity) and physically simulate the stability of the resulting structure to choose the appropriate solution. The success of four- and five-year-old children on this task provides suggestive evidence that they are already capable of such sophisticated physical reasoning. The results also suggest that, even though children in this age group struggled to complete the task in Experiment 1, their difficulty with that task did not stem from an inability to assess the viability of a given plan.

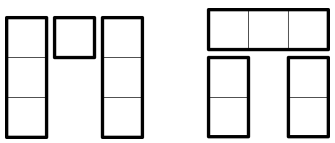


Figure 4: Schematic of two potential solutions presented to children in Experiment 2.

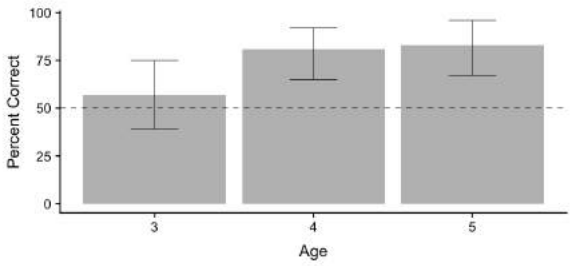


Figure 5: Success rate in Experiment 2. Error bars are bootstrapped 95% CI. Dotted line indicates chance-level.

General Discussion

Our goal was to assess whether the basic capacities for problem decomposition—one of the key components of CT—are present even early in childhood. In Experiment 1, we used a block-building task that involves generating, evaluating, and executing an appropriate decomposition plan to build a physical structure. The results suggest that capacities for top-down design and problem decomposition continue to develop well past the preschool years, and that children become more successful and efficient with age. Although many younger participants failed to complete the task in Experiment 1, in Experiment 2 we find evidence for one of the key steps in successful problem decomposition: children as young as age 4 were able to evaluate the viability of potential solutions.

Experiment 1 featured a rather complex task with high verbal demand for understanding the instructions, which may have increased the task load. Furthermore, this study jointly required top-down design to generate appropriate solutions, evaluation of those solutions, and the actual execution of the plan to assemble the components. Children's struggle with this task could reflect their difficulties in any or all of these steps. Experiment 2 isolates one particular aspect of problem decomposition. Results suggest that 4- and 5-year-old children can compare and evaluate two different decomposition solutions and select the correct one. These results complement prior work (Cortesa et al., 2018) which showed that young children can construct target structures from predetermined components; beyond using a given set of components to assemble the target structure, our results show that 4-year-olds are able to reason ahead under the constraints of a task to infer the correct set of components, even before they engage in actual assembly. Collectively, these findings indicate that some basic underlying capacities for problem decomposition may begin to emerge in preschool years, but they also continue develop well beyond this age.

One might wonder whether children's abilities to engage in problem decomposition in our task is restricted by the physical/spatial domain. Prior work indicates that the ability to engage in basic spatial reasoning emerges early in life (Newcombe & Huttenlocher, 2003). For instance, 5-year-olds show successful mental rotation of a paper cut-out object on a 2-D plane (Frick, Hansen, & Newcombe, 2013) and understand how a scene would look from another person's perspective (Borke, 1975). Our results suggest that even 4-year-olds can mentally rotate a 3-D structure to assess its stability.

Our study focused on a concrete problem with a clear visual representation. Our tasks were intentionally reflective of a thinking pattern common to programming. To solve a programming problem a programmer must identify independently solvable pieces, construct them separately, and then recombine them into a cohesive solution. Of course, throughout this process, the programmer must weigh constraints to make decisions about optimal components or solutions. Similarly, in Experiment 1, children had to identify, construct, and reassemble components of a larger physical structure; there was no possible way to build it directly. Thus, one important question is whether the ability to decompose a problem in the spatial domain extends to more abstract CT problems. Future work might ask whether children's success in this task transfers to decomposition of larger tasks in other STEM areas, such as programming. One possibility is that training children to engage in decomposing a physical structure might also help them decompose a larger programming problem.

Another interesting avenue for future exploration is that the use of concrete objects in physical space might make it easier for children to engage in successful decomposition even in these more abstract domains. Indeed, adults often transform complex abstract tasks into concrete forms, such as diagrams, to avoid trial-and-error in a complex project. We look forward to future work that asks whether physical affordances and manipulatives support children's abstract problem solving in a similar way.

Mark Guzdial, a leading researcher in CS education, wrote: "An open research question is what an elementary school child can learn about computing and what should be taught at what ages" (2015). Our work, along with prior research in cognitive development, suggests that CT is not a unitary construct that emerges at any single age. It involves a range of mental operations which may involve independent developmental trajectories. While children might be able to identify flaws in systems or construct those systems from predetermined parts in preschool, they may not develop the ability to generate those parts until much later. Capacities underlying other CT concepts, such abstraction, data representation, or parallelization, likely also develop in a piecemeal manner that remains to be discovered.

We look forward to more future work that bridges the gap between cognitive development and CS education research. Our work here represents a first step at demonstrating children's developing capabilities in a critical component of computational thinking: problem decomposition. We show that children may be able to learn basic computational thinking skills as early as preschool, but that these capacities continue to develop well into elementary years. As educators continue to develop CS curriculum, these results can inform when and how to teach early programming concepts to young students.

Acknowledgements

Thanks to Rhonda Sandifer for her help with data collection. We also thank staff, children, and families at Bing Nursery School and at the Palo Alto Junior Museum and Zoo.

References

Barr, V., & Stephenson, C. (2011). Bringing computational thinking to k-12: what is involved and what is the role of the computer science education community? *ACM Inroads*, *2*(1), 48–54.

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, 201306572.

Borke, H. (1975). Piaget's mountains revisited: Changes in the egocentric landscape. *Developmental Psychology*, *11*(2), 240.

Brennan, K., & Resnick, M. (2012). New frameworks for studying and assessing the development of computational thinking. In *Proceedings of the 2012 annual meeting of the american educational research association* (Vol. 1, p. 25).

Brosnan, M. J. (1998). Spatial ability in children's play with lego blocks. *Perceptual and Motor Skills*, *87*(1), 19–28.

Cook, C., Goodman, N. D., & Schulz, L. E. (2011). Where science starts: Spontaneous experiments in preschoolers exploratory play. *Cognition*, *120*(3), 341–349.

Cooper, S., Wang, K., Israni, M., & Sorby, S. (2015). Spatial skills training in introductory computing. In *Proceedings of the eleventh annual international conference on international computing education research* (pp. 13–20).

Cortesa, C., Jones, J., Hager, G., Khudanpur, S., Landau, B., & Shelton, A. (2018). Constraints and development in childrens block construction.

Frick, A., Hansen, M. A., & Newcombe, N. S. (2013). Development of mental rotation in 3-to 5-year-old children. *Cognitive Development*, *28*(4), 386–399.

Gardner, W., & Rogoff, B. (1990). Children's deliberateness of planning according to task circumstances. *Developmental Psychology*, *26*(3), 480.

Gergely, G., Nádasdy, Z., Csibra, G., & Bíró, S. (1995). Taking the intentional stance at 12 months of age. *Cognition*, *56*(2), 165–193.

Gopnik, A. (2012). Scientific thinking in young children: Theoretical advances, empirical research, and policy implications. *Science*, *337*(6102), 1623–1627.

Guzdial, M. (2015). Learner-centered design of computing education: Research on computing for everyone. *Synthesis Lectures on Human-Centered Informatics*, *8*(6), 1–165.

Gweon, H., Asaba, M., & Bennett-Pierre, G. (2017). Reverse-engineering the process: Adults and preschoolers ability to infer the difficulty of novel tasks. In *Proceedings of the 39th annual conference of the cognitive science society.*

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in cognitive sciences*, *20*(8), 589–604.

Jara-Ettinger, J., Gweon, H., Tenenbaum, J. B., & Schulz, L. E. (2015). Childrens understanding of the costs and rewards underlying rational action. *Cognition*, *140*, 14–23.

Kamps, F. S., Julian, J. B., Battaglia, P., Landau, B., Kanwisher, N., & Dilks, D. D. (2017). Dissociating intuitive physics from intuitive psychology: Evidence from williams syndrome. *Cognition*, *168*, 146–153.

Klahr, D., & Robinson, M. (1981). Formal assessment of problem-solving and planning processes in preschool children. *Cognitive Psychology*, *13*(1), 113–148.

Kramer, J. (2007). Is abstraction the key to computing? *Communications of the ACM*, *50*(4), 36–42.

Legare, C. H. (2012). Exploring explanation: Explaining inconsistent evidence informs exploratory, hypothesis-testing behavior in young children. *Child development*, *83*(1), 173–185.

Liu, S., Ullman, T. D., Tenenbaum, J. B., & Spelke, E. S. (2017). Ten-month-old infants infer the value of goals from the costs of actions. *Science*, *358*(6366), 1038–1041.

Newcombe, N. S., & Huttenlocher, J. (2003). *Making space: The development of spatial representation and reasoning*. MIT Press.

Schulz, L. (2012). The origins of inquiry: Inductive inference and exploration in early childhood. *Trends in cognitive sciences*, *16*(7), 382–389.

Scott, R. M., & Baillargeon, R. (2013). Do infants really expect agents to act efficiently? a critical test of the rationality principle. *Psychological science*, *24*(4), 466–474.

Verdine, B. N., Golinkoff, R. M., Hirsh-Pasek, K., Newcombe, N. S., Filipowicz, A. T., & Chang, A. (2014). Deconstructing building blocks: Preschoolers' spatial assembly performance relates to early mathematical skills. *Child development*, *85*(3), 1062–1076.

Wai, J., Lubinski, D., & Benbow, C. P. (2009). Spatial ability for stem domains: Aligning over 50 years of cumulative psychological knowledge solidifies its importance. *Journal of Educational Psychology*, *101*(4), 817.

Wing, J. M. (2006). Computational thinking. *Communications of the ACM*, *49*(3), 33–35.

Yildirim, I., Gerstenberg, T., Saeed, B., Toussaint, M., & Tenenbaum, J. (2017). Physical problem solving: Joint planning with symbolic, geometric, and dynamic constraints. *arXiv preprint arXiv:1707.08212*.

Zhang, M., Igarashi, Y., Kanamori, Y., & Mitani, J. (2017). Component-based building instructions for block assembly. *Computer-Aided Design and Applications*, *14*(3), 293–300.

A Familiarity-dependent Retrieval Threshold in ACT-R

Cvetomir M. Dimov (cdimov@andrew.cmu.edu)
Department of Psychology, Carnegie Mellon University,
5000 Forbes Avenue, Pittsburgh, PA 15213, USA

Abstract

In their current functional form, ACT-R's retrieval equations do not account for the left side of the RT-distance relation, that is, that as memory activation decreases, so does response time for retrieval failures. To accommodate this effect, I propose that the memory system uses the familiarity of the encoded object to gauge how much effort it should devote to retrieval. I quantify the degree of familiarity through the match score, which is the output of a global matching process. Familiarity, in turn, directly determines what the retrieval threshold should be. Adding a familiarity process orthogonal to recollection is in line with neuroimaging results, which uncover parallel familiarity and retrieval processes. The developments in this paper extend ACT-R's memory theory into a dual process theory.

Keywords: ACT-R, declarative memory, familiarity, retrieval threshold

Introduction

Perhaps uniquely among current theories of memory, ACT-R's memory theory (Anderson & Schooler, 1991; Schooler & Anderson, 1997) assumes a recall process, but no familiarity process. Despite being a single-process theory, it has successfully accounted for both responses and response times (RTs) of not only various recall processes (e.g., Anderson, Fincham, & Douglas, 1999; Anderson & Rader, 1999), but also of various recognition processes (e.g., Anderson, Bothell, Lebiere, & Matessa, 1998; Schneider & Anderson, 2012). Yet, at least one aspect of recall that ACT-R does not currently account for is the shape of the RT curve of "No" responses. Here I put forth a proposal of extending this memory theory such that it can also accommodate the RT distribution of recall failures. This proposal is consistent with recent neural evidence of separate familiarity and recall processes (e.g., Borst and Anderson, 2015). Specifically, I suggest that (1) familiarity in ACT-R is modeled with a global-matching process and (2) the retrieval threshold is strategically varied as a function of familiarity.

The Memory Theory behind ACT-R

ACT-R makes the distinction between representations, which inhabit the symbolic level, and the equations governing them, which lie at the subsymbolic level. At the symbolic level, ACT-R represents items in declarative memory as chunks, which are a collection of one or more slot-value pairs. Facts, such as "Otters hold hands" and "Cherry coke tastes like cyanide", and experiences such as "I rappelled off a bridge on Sunday" are all stored as chunks in declarative memory.

At the subsymbolic level, several equations determine whether chunks are likely to be retrieved or not and how long that will take. These equations take into account the prior history of encounter of the episodes or facts encoded in chunks as well as their relevance to the current context, and bind those together into a single quantity – a chunk's activation. Each chunk *i* has an activation, A_i, associated with it that quantifies its strength. Activation is a dynamic quantity that models the logarithm of the odds (i.e., the *log-odds*) that a chunk is needed at this point in time in this context to achieve the goal the agent strives for. Activation is composed of *base-level activation*, B_i, the *spreading activation*, SA_i, and noise, ε:

$$A_i = B_i + SA_i + \varepsilon \quad (1)$$

The base-level activation is a function of the chunk's history:

$$B_i = \ln \sum_{k=1}^{n_i} t_k^{-d}, \quad (2)$$

where the *decay parameter*, d[1], specifies the rate of forgetting over time, which is modeled with a power function. The power function was chosen, because the likelihood of encountering items in the real world also decays as a power function as time passes and the memory system is hypothesized to have adapted to this regularity (Anderson & Schooler, 1991). The parameter n is the number of encounters with the information that chunk i represents, and t_k is the time since the k^{th} encounter.

Spreading activation SA_i assesses a chunk's relevance to the current context, where the current context consists of all chunks currently in the focus of attention (i.e., all chunks currently in the *buffers* of the various *modules* that ACT-R consists of). SA_i assumes that chunks in declarative memory related to or previously encountered with chunks in buffers are more likely to be needed than those that are not. The amount of spreading activation to chunk i in declarative memory is a function of the associations between that chunk and the currently attended to chunks j:

$$SA_i = \sum_j W_j S_{ji}, \quad (3)$$

where the associative strength, S_{ji}, between chunks i and j is weighted by the source activation, W_j, of chunk j in a buffer. The associative strengths, S_{ji}, between chunks is approximated by

$$S_{ji} = S - \ln(fan_j), \quad (4)$$

where S denotes the maximum associative strength and fan_j is the number of chunks associated with a chunk j. The more

[1] Typically set to $d = 0.5$.

chunks are associated with a chunk in memory, the lower the associative strength between it and each of its associates becomes. Equation (4) is approximation of the Bayesian memory analysis that ACT-R is based on (Anderson & Milson, 1989) which assumes that each association of chunk *j* is equally likely to be needed. This approximation usually accounts sufficiently well for experimental regularities, but in some cases the full Bayesian equation needs to be summoned (see Anderson & Reder, 1999).

Equations 2-4 determine the activation components summed in Equation 1, which then determines probability of retrieval and retrieval failure as well as retrieval time. Specifically, whenever A_i is above the *retrieval threshold* τ, the chunk can be retrieved, while if it is below that threshold, the chunk is not sufficiently active to be retrieved, resulting in a probability of retrieval p_i as a function of threshold:

$$p_i = \frac{1}{1+e^{-\frac{\mu_{A_i}-\tau}{s}}}, \quad (5)$$

where $\mu_{A_i} = B_i + SA_i$ is the mean of the activation distribution. Retrieval time is also exponentially related to activation:

$$t_{retrieval} = Fe^{-A_i}. \quad (6)$$

The *latency factor F* scales the resulting quantity into units of seconds. If the activation is below the retrieval threshold, the resulting retrieval failure time is constant:

$$t_{retrieval\ failure} = Fe^{-\tau}. \quad (7)$$

ACT-R and Familiarity

Recognition tasks in ACT-R have typically been modeled with the same retrieval process that recollection is modelled with, but with different parameters (see Anderson, Bothell, Lebiere, & Matessa, 1998), whereby no fluency or familiarity processes are mentioned. Yet, familiarity processes are explicitly mentioned in at least one (decision) model constructed in ACT-R, the *fluency heuristic* (Schooler & Hertwig, 2005).

The Fluency Heuristic

The fluency heuristic is a memory-based decision strategy that infers which of two alternatives scores higher on a criterion by choosing the alternative that is more fluent (i.e., more familiar). The fluency heuristic does not require a separate familiarity processes running in parallel to retrieval to model fluency. Instead, in its original definition, the fluency heuristic operationalizes fluency as the time it takes an object to be retrieved (Schooler & Hertwig, 2005). Later, this heuristic was redefined to rely on the newly developed ACT-R module for prospective time interval estimation (Taatgen, van Rijn, & Anderson, 2007), thus comparing the subjectively perceived retrieval times of the two alternatives and choosing the one that is subjectively faster to retrieve (see Dimov, Marewski, & Schooler, 2017; Fechner et al., 2016; Marewski & Schooler, 2011). Still, even in its updated version, the fluency heuristic does not assume a separate familiarity process, but continues to rely on a recall process paired with a process for estimating the time recall takes.

Neural Evidence of Familiarity

While, for the most part, the memory and decision tasks modeled with ACT-R did not necessitate two separate mnemonic processes, recently several neuroimaging studies examining the time course of associative recognition have provided evidence in favor of two processes operating in parallel: a familiarity process and a recollection process. Specifically, due to fMRI not providing the temporal resolution necessary to observe sub-second retrieval processes, both EEG (Borst and Anderson, 2015) and MEG (Borst, Ghuman and Anderson, 2016) were used to record brain signatures during this retrieval task. The brain signatures during associative recognition indicate the existence of a familiarity process commencing in parallel with a recollection process and finishing typically before, but not substantially before the recollection process. How can we model this familiarity process with ACT-R?

A Global-Matching Process in ACT-R to Model Fluency

My first proposal is that familiarity in ACT-R is related to *blending* (Lebiere, 1998). Blending is a process in ACT-R's declarative memory that produces a weighted average of a quantity over all chunks in memory that hold a value of that quantity, whereby the contribution of each chunk is weighted by its activation. The output of blending is a chunk holding the weighted average value. This mechanism has been used to model mistakes that children make when engaging in arithmetic (Lebiere, 1999), choices in dynamic decision making tasks (e.g., Gonzalez & Dutt, 2011; Gonzalez, Lerch, & Lebiere, 2003) and belief updating in repeated games (Spiliopoulos, 2013) among others.

At the subsymbolic level, the blended chunk is described with a *match score M*, which is the analogue of activation for the blended chunk. Just as the blended value, the match score is a function of the activations of the set of all chunks included in the blending process (called the *match set MS*):

$$M = \ln \sum_{i \in MS} e^{A_i}. \quad (8)$$

At first sight unintuitive, Equation 8 becomes clearer once we consider that activation is on a logarithmic scale (see Equation 2) and that all observables (Equations 5-7) are related to the exponent of activation. Summing the exponents of all relevant chunks' activation and then taking the logarithm renders the resulting match score equivalent to the activation resulting from the cumulative experience of all blended chunks. For example, if we consider only base level activation, the resulting match score would be:

$$\begin{aligned} M &= \ln \sum_{i \in MS} e^{A_i} = \\ &= \ln \sum_{i \in MS} e^{\ln \sum_{k=1}^{n_i} t_k^{-d}} = \\ &= \ln \sum_{i \in MS} \sum_{k=1}^{n_i} t_k^{-d}, \end{aligned} \quad (9)$$

which is the activation a chunk would have had it had the prior history of all blended chunks.

While activation is interpreted as the log-odds of a chunk being needed, the match score is the log-odds of *any* chunk in declarative memory being needed. The specific relationship that I propose is that the familiarity of an input is quantified by the match score produced by the blending process, that is, familiarity serves as a coarse gauge if any of the input is relevant to the task at hand.

The RT-distance Relation and ACT-R

In a recognition task, responses are classified as Hits and False Alarms (whenever the response is "yes") and Misses and Correct Rejections (whenever the response is "no"). Whether responses are correct or not, there is a well-established relation between the time that those responses take (RT) and how frequently the item was presented in the experiment or encountered in life: the RT-distance relation (Koppell, 1977). This relation states that response time is fast whenever items were presented very frequently or very rarely, resulting in a memory trace with a very high or very low strength. However, whenever the memory strength is in the middle ground, close to the retrieval threshold, responses take more time. In other word, RT decreases as the memory strength of an item lies further away (either to the left or to the right) from the retrieval threshold (see Figure 1 for an idealization). Consequently, both Hits and False Alarms become faster the higher memory strength is of retrieved items. Moreover, both Misses and Correct Rejections speed up the lower memory strength of the items that fail to be retrieved.

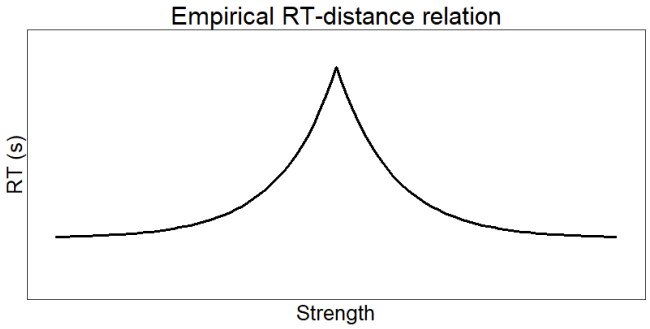


Figure 1: RT-distance relation in recall. A memory item with a medium strength provides ambiguous information about whether it has been encountered in the past or not and, consequently, requires a longer time to be retrieved. Memory items with either very high or very low strength are both responded to quickly.

In its current form, ACT-R's memory theory accounts for half of the RT-distance relation: that related to successful retrievals (Hits and False Alarms). Yet, following from Equation 7, when an item of memory fails to be retrieved (Misses and Correct Rejections), ACT-R predicts a constant RT (see Figure 2), which contradicts the empirically found RT-distance relation.

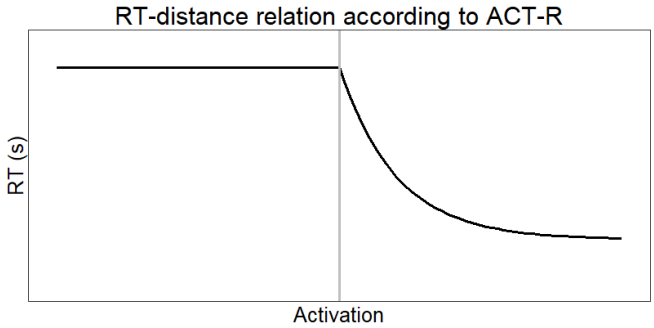


Figure 2: Relation between chunk activation and response time according to ACT-R. The grey line indicates the location of the threshold. Above the threshold, successful retrievals get progressively faster as one moves away from the threshold. Below the threshold, retrieval failures take a constant time irrespective of their distance from the threshold.

Fluency Determines Threshold

My second proposal aims to modify ACT-R's memory to account for the RT-distance relation using the fluency process earlier introduced. Specifically, I propose that the retrieval threshold τ is not constant, but that it is a function (i.e., the negative) of an item's familiarity (which I proposed to model with the match score):

$$\tau = -M. \qquad (10)$$

Since M is the log odds that any chunk in memory is needed, $-M$ is the log odds of no chunk being needed:

$$-M = -\log\left(\frac{need}{\neg need}\right) = \log\left(\frac{\neg need}{need}\right) \qquad (11)$$

In plain language, my proposal can be interpreted as the memory system dynamically adjusting the amount of effort it is willing to invest into a retrieval (as described by the retrieval threshold) as a function of how likely it is that no chunk in memory is ever needed, which is estimated via the fluency signal. If the global fluency signal is weak, that is, if the odds that any chunk in memory is needed at this moment is low, then the system will invest less resources into a retrieval attempt and abort it earlier. On the other hand, if the fluency signal is strong, the memory system will be ready to invest a lot of time into retrieval as it is more certain that it will retrieve a relevant chunk, even though in practice it will invest very little time as a successful retrieval will soon arrive.

Resulting RT-distance relation

When chunks are very distinct or, in the extreme case, when all chunks spread 0 activation to each other, the predominating factor in the match score is the activation of the chunk being probed as only this chunk will be included in the match set MS. In this case, the resulting RT-distance relation is almost entirely symmetric (see Figure 3)[2].

[2] The code used to generate Figures 2, 3 and 4 can be found in the Appendix.

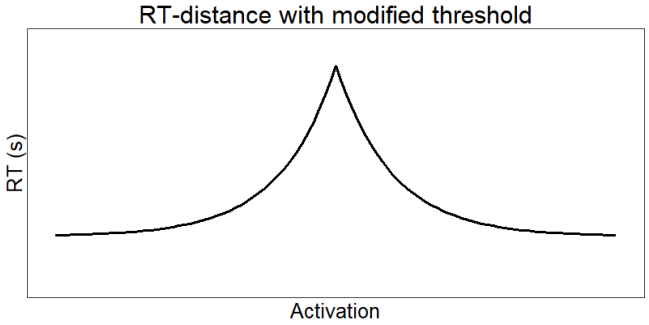


Figure 3: Relation between chunk activation and response time according to our modification of the retrieval threshold in ACT-R. I assume that the only chunk in the match set is the chunk being probed.

On the other hand, when other chunks are similar to the probed chunk and, thus, included in the match set, they increase the likelihood that any chunk in declarative memory will be needed. Consequently, the activation of the chunk representing the item being probed crosses the retrieval threshold at a lower value and, moreover, RT on retrieval failures decreases less and less steeply (see Figure 4)[3].

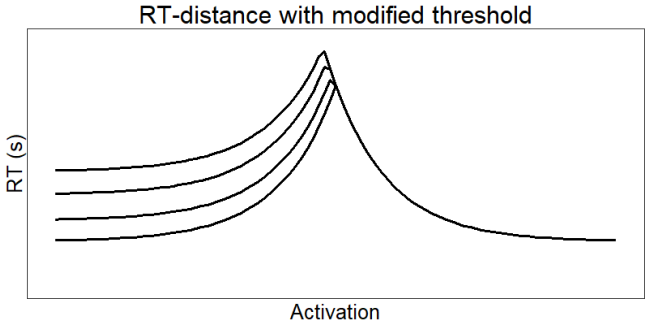


Figure 4: Relation between chunk activation and response time according to our modification of the retrieval threshold in ACT-R. The different curves correspond to various contribution to the match score of chunks that represent other items than the item being probed. As the similarity to the probed chunk increased, RT of retrieval failures increases.

Discussion and Conclusion

I proposed that familiarity in ACT-R is modeled with the match score from blending. This extends ACT-R's memory theory to a dual-process theory of memory. Moreover, I hypothesize that the memory system relies on the familiarity signal to assess the amount of effort it should invest in retrieval before aborting it. This allows ACT-R to account for the RT-distance relation. Finally, the prediction that retrieval failures will take longer when the probed chunk is confusable with other chunks in memory also follows from this new formulation. The modification interprets the blending module as a global-matching component of ACT-R's memory and puts it in the tradition of many a memory models in psychology, which include TODAM (Murdock, 1982), MINERVA 2 (Hintzman, 1984) and SAM (Gillund & Shiffrin, 1984; Raaijmakers & Shiffrin, 1980; for an overview of global-matching models, see Humphreys, Pike, Bain, & Tehan, 1989). I will proceed by briefly comparing the proposed extension of ACT-R to two related theories of memory and discuss the potential issues with the current proposal.

Comparison to Source of Activation Confusion

A theory that shares its lineage with ACT-R's is *Source of Activation Confusion* (*SAC*, Diana, Reder, Arndt, & Park, 2006). This theory has been used to model a wide variety of memory phenomena in various tasks, among which cued recall (Reder, Park, & Kieffaber, 2007), perceptual match effects (Diana, Peterson, & Reder, 2004) and feeling of knowing (Schunn, Reder, Nhouyvanisvong, Richards, & Stroffolino, 1997).

SAC is not based on the rational analysis of memory, yet many of the processes that it assumes are the same as those of ACT-R. First, it assumes that events and objects are encoded as chunks. Second, those chunks' activations are also separated into a base-level and spreading-activation components. Third, base-level activation decays with time as a power law, while spreading activation is a function of co-occurrence frequencies. Yet, there are at least two points of departure between ACT-R and SAC. First, SAC assumes that spreading activation slowly decays with time once the chunk that spreads activation is removed from the focus of attention, while in ACT-R this happens instantaneously. Second, in SAC a working memory of a limited capacity is populated with all chunks above a certain level of activation.[4]

Unlike ACT-R, SAC is a dual-process theory: it includes both a familiarity and a recall process. The familiarity (or feeling-of-knowing) process stems from retrieval of the concept node, which is the internal representation of the probed item. Activation then spreads to associated nodes, which leads to cued recall. Thus, unlike the current proposal of extending ACT-R's memory theory, in SAC familiarity does not result from a global-matching process, but from a retrieval of a single chunk.

Note, however, that adding a global-matching process that determines the retrieval threshold can also benefit SAC. First, just like ACT-R, SAC does not model the RT-distance relation related to retrieval failures, because it assumes a constant threshold. By adding a threshold that is inversely related to the global-matching signal, SAC should also be able to accommodate this relation. Second, SAC assumes that the familiarity (those related with retrieval of the concept node) and recollection processes (those related to retrieval of episode nodes) rely on different thresholds, whereby the

[3] I have relied on a single parameter to quantify the total amount of spreading activation that comes from chunks not corresponding to the presented item. See the Appendix for model code.

[4] Note that ACT-R's notion of working memory is more complicated in that it includes the buffers of all of its modules and potentially the content of the imaginal module, which stores task-relevant information.

concept threshold is typically higher than the episode threshold. I posit that Equation 10 would accommodate that a higher threshold for the concept than for the episode node. Specifically, during retrieval, activation spreads from the concept and the context into the episode node. If the activation of the episode is high enough, it will be retrieved. In this case the high activation of the episode node would also imply a high overall match score and, consequently, a low retrieval threshold. On the other hand, if the episode node cannot be retrieved and, instead, the concept node is relied upon, this would imply that the episode node has a lower activation. Consequently, the overall match score will be lower, implying a higher retrieval threshold.

Comparison to Retrieving Effectively from Memory

Retrieving effectively from memory (*REM*; Shiffrin & Steyvers, 1997) is a memory model that stems from the tradition of SAM. REM is a global matching model – it assumes that the recall probe is matched to all memory traces in parallel. Similar to ACT-R's memory theory, which is based on rational analysis, REM assumes that the memory system is optimally weighing signal and noise, and computing the likelihood that the probe has been encountered before in order to respond whether a presented item is recognized or not.

REM has been extended to also model various cued recall (e.g., Diller, Nobel, & Shiffrin, 2001) and free recall (e.g., Lehman & Malmberg, 2013) phenomena. To this end, REM was complemented with a trace recovery process, which is executed if the global matching process indicates a likely past experience with the probe. The current extension of ACT-R to include familiarity as a global matching process is similar to REM in that (1) it is a dual process model, (2) familiarity is a global matching process, (3) recollection is the recovery of a single memory trace, (4) whether effort is invested in recollection is strategically determined by the familiarity signal.

In addition to these similarities, there are several core differences between the two models. First, ACT-R assumes that base-level activation decays with time, while in REM and its extensions memory decay is generally absent. Second, ACT-R assumes that memory traces monotonically increase in activation with the number of encounters of the objects or events that they represent, while in REM a new trace can be created to store the encoded event/object or an already existing trace can be updated to store a more complete representation of the object. After a certain number of presentations, the object is perfectly encoded and no further updates of the memory trace(s) takes place. Which of those approaches provides a better description of memory phenomena is subject to further investigations.

Limitations of the Current Proposal

ACT-R's theory of memory assumes that our memory system makes a guess about which items of memory are most likely to be needed, what the cost and benefits of retrieval will be, and optimally combines those. The current analysis does not take into consideration costs and benefits. Yet, this might be problematic as the blending process is much more computationally intensive than the retrieval process itself: the activations of all chunks are computed and inserted into Equation 8 and, moreover, the blended value needs to be computed. Perhaps one way to alleviate these considerations would be to separate the computation of the match score from the computation of the blended value. This way the familiarity process would only require the computation of the match score, which sums chunks activation – values that need to be computed for retrieval in any case. Moreover, this would make the proposed familiarity process as complex as that of any other global matching theory.

To conclude, the current analysis is limited to only memory processes. Yet, neural data indicate that recollection, in addition to having a different neural signature than familiarity, also includes an additional decision phase (Borst, Ghuman, & Anderson, 2016). My analysis does not speak to the nature of these decision processes. Future work should focus on better understanding them.

References

Anderson, J. R., Bothell, D., Lebiere, C., & Matessa, M. (1998). An integrated theory of list memory. *Journal of Memory and Language*, 38, 341-380.

Anderson, J. R., Fincham, J. M., & Douglass, S. (1999). Practice and retention: A unifying analysis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*, 1120-1136.

Anderson, J. R., & Milson, R. (1989). Human memory: An adaptive perspective. *Psychological Review, 96*, 703-719.

Anderson, J. R., & Reder, L. M. (1999). The fan effect: New results and new theories. *Journal of Experimental Psychology: General, 128*, 186-197.

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science, 2*, 396-408.

Borst, J. P., & Anderson, J. R. (2015). The discovery of processing stages: Analyzing EEG data with hidden semi-Markov models. *NeuroImage, 108*, 60-73.

Borst, J. P., Ghuman, A. S., & Anderson, J. R. (2016). Tracking cognitive processing stages with MEG: A spatio-temporal model of associative recognition in the brain. *NeuroImage, 141*, 416-430.

Diana, R. A., Reder, L. M., Arndt, J., & Park, H. (2006). Models of recognition: A review of arguments in favor of a dual-process account. *Psychonomic Bulletin & Review, 13*, 1-21.

Diller, D. E., Nobel, P. A., & Shiffrin, R. M. (2001). An ARC–REM model for accuracy and response time in recognition and recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*, 414-435.

Dimov, C. M., Marewski, J. N., & Schooler, L. J. (2017). Architectural process models of decision making: Towards a model database. In G. Gunzelmann, A. Howes, T.

Tenbrink, & E. J. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 1931-1936). Austin, TX: Cognitive Science Society

Fechner, H. B., Pachur, T., Schooler, L. J., Mehlhorn, K., Battal, C., Volz, K. G., & Borst, J. P. (2016). Strategies for memory-based decision making: Modeling behavioral and neural signatures within a cognitive architecture. *Cognition, 157*, 77-99

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review, 91*, 1-67.

Gonzalez, C., & Dutt, V. (2011). Instance-based learning: Integrating sampling and repeated decisions from experience. *Psychological Review, 118*, 523-551.

Gonzalez, C., Lerch, J. F., & Lebiere, C. (2003). Instance-based learning in dynamic decision making. *Cognitive Science, 27*, 591-635.

Hintzman, D. L. (1984). MINERVA 2: A simulation model of human memory. *Behavior Research Methods, Instruments, & Computers, 16*, 96-101.

Humphreys, M. S., Pike, R., Bain, J. D., & Tehan, G. (1989). Global matching: A comparison of the SAM, Minerva II, Matrix, and TODAM models. *Journal of Mathematical Psychology, 33*, 36-67.

Koppell, S. (1977). Decision latencies in recognition memory: A signal detection theory analysis. *Journal of Experimental Psychology: Human Learning and Memory, 3*, 445-457.

Lebiere, C. (1998). The dynamics of cognition: An ACT-R model of cognitive arithmetic (CMU-CS-98-186. Available at: http://reports-archive.adm.cs.cmu.edu/). Pittsburgh, PA: Carnegie Mellon University.

Lebiere, C. (1999). The dynamics of cognition: An ACT-R model of cognitive arithmetic. *Kognitionswissenschaft, 8*, 5-19.

Lehman, M., & Malmberg, K. J. (2013). A buffer model of memory encoding and temporal correlations in retrieval. *Psychological Review, 120*, 155-189.

Marewski, J. N., & Schooler, L. J. (2011). Cognitive niches: An ecological model of strategy selection. *Psychological Review, 118*, 393-437.

Murdock, B. B. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review, 89*, 609-626.

Raaijmakers, J. G., & Shiffrin, R. M. (1980). SAM: A theory of probabilistic search of associative memory. In *Psychology of learning and motivation* (Vol. 14, pp. 207-262). Academic Press.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM—retrieving effectively from memory. *Psychonomic Bulletin & Review, 4*, 145-166.

Schneider, D. W., & Anderson, J. R. (2012). Modeling fan effects on the time course of associative recognition. *Cognitive Psychology, 64*, 127-160.

Schooler, L. J., & Anderson, J. R. (1997). The role of process in the rational analysis of memory. *Cognitive Psychology, 32*, 219-250.

Schooler, L. J., & Hertwig, R. (2005). How forgetting aids heuristic inference. *Psychological Review, 112*, 610-628.

Schunn, C. D., Reder, L. M., Nhouyvanisvong, A., Richards, D. R., & Stroffolino, P. J. (1997). To calculate or not to calculate: A source activation confusion model of problem familiarity's role in strategy selection. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*, 3-29.

Spiliopoulos, L. (2013). Beyond fictitious play beliefs: Incorporating pattern recognition and similarity matching. *Games and Economic Behavior, 81*, 69-85.

Taatgen, N. A., Van Rijn, H., & Anderson, J. (2007). An integrated theory of prospective time interval estimation: The role of cognition, attention, and learning. *Psychological Review, 114*, 577-598.

Appendix

Here I include the R code used to generate Figures 2, 3, and 4. The two parameters that I specify are (1) the latency factor F and (2) the perceptual-motor time t_{pm}. The precise values of these parameters (0.35 s and 0.8 s) were chosen to be realistic. Yet, their values do not change the functional form, which is what we are ultimately interested in.

To generate Figure 2, I used the standard ACT-R equation (Equations 6 and 7), which assumes a constant RT below the threshold τ (here τ is set to 0) and an activation-dependent RT above threshold:

```
F = 0.35;
t_pm = 0.8;
RTACTR <- function(A){
  tau <- 0
  if (A < tau){
    return(F + t_pm)
  } else {
    return(F*exp(-A)+t_pm)
  }
}
```

To generate the data for Figures 3 and 4, I the modified equation that I propose:

```
RTACTR_new <- function(A,A_rest){
  M <- log(exp(A)+exp(A_rest))
  tau <- -M
  if (A < tau){
    return(F*exp(-tau)+t_pm)
  } else {
    return(F*exp(-A)+t_pm)
  }
}
```

where A_{rest} is the contribution to the match score of all non-target items. To generate Figure 3, I assumed that A_{rest}=0, while I used values of -10, -2, -1.2, and 0.8 to generate the plot in Figure 4.

Decoding Affirmative and Negated Action-Related Sentences in the Brain with Distributional Semantic Models

Vesna Djokic
University of Southern California, Los Angeles, California, United States

Jean Maillard
University of Cambridge, Cambridge, United Kingdom

Luana Bulat
University of Cambridge, Cambridge, United Kingdom

Ekaterina Shutova
University of Amsterdam, Amsterdam, Netherlands

Abstract

Recent work shows that distributional semantic models can be used to decode patterns of brain activity associated with individual words and sentence meanings. However, it is yet unclear to what extent such models can be used to study and decode brain activity patterns associated with specific aspects of semantic composition such as the negation function. In this paper, we investigate the extent to which distributional semantic models of action-verbs correlate with brain activity associated with negated and affirmative sentences containing hand-action verbs. Our results show reduced correlations for sentences where the verb is in the negated context, as compared to the affirmative one, within brain regions implicated in action-semantic processing. The results lend support to the idea that negation involves reduced access to aspects of the affirmative representation and pave the way for further testing alternate distributional-based semantic models of negation against human semantic processing in the brain.

How time spent on feedback influences learning and gaze in categorization training

Katerina Dolguikh
Simon Fraser University, Burnaby, British Columbia, Canada

Jordan Barnes
Simon Fraser University, Burnaby, British Columbia, Canada

Tyrus Tracey
Simon Fraser University, Burnaby, British Columbia, Canada

Mark Blair
Simon Fraser University, Burnaby, British Columbia, Canada

Abstract

Feedback is essential for many kinds of learning, but the cognitive processes involved in learning from feedback are unclear. In models of category learning, feedback is typically treated as an error signal without a temporal component. We conducted two simple category learning experiments that manipulated the duration of feedback (1s vs. 9s) and investigated the effect on learning and gaze. In two different category structures, participants in the longer feedback condition learned faster. The analysis of gaze data showed several findings. Participants in the 9s condition had longer fixations, and in both conditions and experiments, participants spent far more time looking at stimulus features than the feedback. Overall, our findings provide empirical support for the idea that feedback processes, and temporal factors more generally, have much to tell us about how people learn categories.

Reinforcing Rational Decision Making in a Risk Elicitation task through Visual Reasoning

Stella Doukianou (S.Doukianou@gre.ac.uk)
Damon Daylamani-Zad (D.D.Zad@gre.ac.uk)
School of Computing and Mathematical Sciences, University of Greenwich,
Park Row, London, SE10 9LS, U.K.

Petros Lameras (PLameras@cad.coventry.ac.uk)
Ian Dunwell (IDunwell@cad.coventry.ac.uk)
School of Computing, Electronics and Maths, Coventry University, Priory St,
Coventry CV1 5FB, U.K.

Abstract

Metrics seeking to predict financial risk-taking behaviors typically exhibit limited validity. This is due to the fluid nature of an individual's risk taking, and the influence of the mode and medium, which presents a decision. This paper presents two experiments that investigate how an existing risk elicitation task's predictive capacity may be enhanced through the application of an interactive model of visual reasoning in a digitized version. In the first experiment, 60 participants demonstrated their reasoning process. In the second experiment, 225 participants were randomly assigned into three groups, with the validated risk elicitation task compared as a control to interactive digital and non-interactive digital stimuli with pie charts. The experiments yielded significant results, highlighting that when participants interact with a graph to reason their choices, it leads to consistent choices. The findings have implications for improvement of the risk task's validity and the deployment of digital interactive assessments beyond laboratory settings.

Keywords: visualization, decision-making, risk-taking, external representations, reasoning

Introduction

The ability to elicit the degree to which an individual or demographic is risk-taker or -averse has a value across various fields. Existing risk elicitation tasks have shown to have predictive capacity; however, in risk elicitation tasks that involve lotteries, a key constraint is participants' limited understanding of the probabilities, representing the risk associated with each choice, in those tasks. If participants do not understand a probability-based question, their answers lack internal consistency. Whilst within these tests internal checks for validity exist, these alone allow only for the exclusion, rather than accommodation, of participants with low numeracy skills that have limited understanding of the probabilities presented in the task. As a result, findings can be skewed towards a subset of a larger sample, limiting validity and predictive capacity.

Towards resolving this issue, this paper presents the findings of two empirical studies that were conducted to investigate the use of visualization reasoning methods as an assistive tool for users to understand the probability described and choose consistently in a risk elicitation task. Experiment 1 sought to explore the most effective reasoning methods used by 60 participants in a risk elicitation task, by asking participants to illustrate their thought-process. Visual reasoning was identified as having the strongest positive effect, among all external representations used by the participants, on the consistency of their choices in the risk elicitation task (Holt & Laury, 2002), showing that it helped them understand the probabilities in the risk task. To confirm whether visual reasoning, which can be defined as using visuals to reason a probability problem, on risk elicitation task can help participants with low numeracy level, Experiment 2 translated the existing risk-elicitation task to two versions; a non-interactive visual digital version and an interactive visual digital version, in addition to the standardised (control) numerical format task. Our discussion and conclusions reflect on the relevance of the findings in terms of increasing the accessibility and meaningfulness of risk-elicitation tasks to less numerate participants by using visual reasoning processes and also the implications for the use of digital and interactive visual media in place of standardised paper based tasks.

Background

Methods of risk-elicitation tasks can be broadly categorized into self-report questionnaires describing hypothetical situations; hypothetical choice problems; or computerized methods (Rohrmann, 2005). A range of moderating variables have been observed to affect the validity of the majority of these methods, leading to inaccurate results and predictions (Andersen et al., 2006; Dave et al., 2010). This can limit the scalability of an elicitation-exercise, and relates to a fundamental challenge in transferability of results to different contexts: any psychometric tool seeking to establish or predict behavior must consider the fluidity of individual characteristics, and furthermore how a slight change may result in a meaningful change in decision-making.

The Multiple Price List (MPL) task belonging to the category of hypothetical choice problems, wherein participants need to choose between a 'safe' or 'risky' bet over ten different lotteries, has demonstrated predictive capability and can be implemented straightforwardly (Andersen et al., 2006; Dave et al., 2010; Rohrmann, 2005). Holt and Laury's variant of the MPL task has been examined in several studies (e.g.Nielsen, Keil, & Zeller, 2013; Dave et al., 2010), demonstrating significant predictive value but only for people with higher numerical skills who understand the

probabilities for each of the parameters (Dave et al., 2010). The additional value of Holt Laury's MPL method is its capacity to identify the inconsistency rate in an individual's responses, allowing these choices to be excluded. However, this results in decreasing the validity of the metric.

To increase participants' understanding of the probabilities presented in the Holt and Laury MPL method, researchers have explored the use of visual display formats to represent the lotteries which reflect the losses and gains of the two options, allowing for more consistent and rational choices (Boughera, Gassmann & Piet, 2011; Bauermeister & Mußhoff, 2016; Habib et al., 2016). According to empirical evidence, using visualization tools to illustrate the uncertainty of the variables has a significant merit for people's informed decision making (Deitrick & Edsall, 2006). Integrating visualizations in the reasoning process could help users with low numerical skills to choose meaningfully regarding the risk related choices (Padilla et. al, 2018). Given that numeracy is an observed, reliable predictor of responses in risk elicitation tasks, a goal here is to address demographics or individuals who may have lower numeracy skills, but for whom a predictive mean of assessing their risk-taking or aversion holds value. Additionally, cognitive thinking style has been suggested to be another influential factor in a person's risk related choices (Frederick, 2005). Therefore, lower inconsistency rate achieved by presenting problems, which allows a wider range of individuals to respond consistently, would yield more reliable data. Research has illustrated the capacity of visualization to make probability reasoning more intuitive, and therefore better understood by participants (Hegarty & Kozhevnikov, 1999). In turn, this is shown to improve predictive validity. In the following section, visualization and external representations are briefly discussed.

Visualizations and external representations

Visualizations can become a reasoning process through interaction (Khan, Breslav & Hornbæk, 2018). For problems that involve probabilities, the individual relies on both internal visualization of the problem and the use of external representations such as sketches or diagrams to facilitate the solution. Visualizations serve as cognitive aids in problem solving situations (Khan, Breslav & Hornbæk, 2018; Padilla et al. 2018). However, individual differences, personality traits and cognitive abilities can have a significant effect on the use of a method that can aid the problem solving process (Ziemkiewicz et al., 2012; Gray & Holyoak, 2018). Therefore, visualization may not be the most appropriate external representation to assist in decision making for every individual (Starns et al., 2018). Even more, the type of visualization that can support the decision making for various tasks can differ significantly (Starns et al., 2018).

According to Corter and Zahner (2007) when investigating external representations, there is a division into two categories: the first refers to external visual presentations, which are provided towards influencing or informing a decision-making process (Corter, & Zahner 2007). The second involves understanding the internal representations the individual uses to reason when faced with a decision. The second involves the effects of user-generated visual representations while engaged in decision making and problem-solving activity. These internal representations, which can be defined as visual imagery (Corter & Zahner, 2007), have a core role in the production of knowledge. When externalized, they can provide valuable insight into individual's decision-making and aid successfully with problem solving (Gilbert, 2008). Empirical evidence suggests that using external representation to reason probabilities helps individuals to solve problems successfully (Corter & Zahner, 2007; Zhang 1997). In studies for logical reasoning, it has been argued that graphical representations allow the successful interpretation of abstract concepts (Zhang, 1997).

As the body of research suggests that external representations may be able to aid the reasoning process in probability problems, the following studies investigate the use of external representations to help individuals choose rationally in a risk elicitation task with lotteries. The first experiment is looking to determine whether there is a significant relationship between level of education, numeracy level and cognitive thinking style in rational decision-making in risk elicitation tasks. The second experiment investigates whether an interactive pie chart approach can influence the rational decision making in those tasks.

Experiment 1

Method

Participants Sixty volunteers from the UK participated in the study. The participants were divided into two groups depending on their educational level:

Group 1: Thirty-five participants (21 m, 14 f) with an age range between 22 and 53 (M=29.5) volunteered to participate in the experiment and had completed a degree level or higher qualification. The participants were invited through snowball effect via the network of a UK SME in the energy sector, and a British University.

Group 2: Twenty-five students from a further education college in the UK (4 f, 21 m), ranging in age 18-37 (M =20.64) volunteered to participate in the experiment and had not completed any degree-level qualification. The male participants outweighed significantly the number of females, this would be perceived as influencing the design and implementation of our study. All of them were assigned to the same experimental task as Group 1; similar research ethics procedures were also applied to this group. Participants were students in the Game Design and Web Design courses at a remedial programme. None of the participants had attended a course at a university level.

Measures The Lipkus Numeracy scale developed by Lipkus, Samsa and Rimer (2001) was used for this study. It was selected among others as a numeracy assessment tool for this study because it has been used in similar research to assess basic arithmetic skill in the variety of groups (Peters et al., 2006; Schapira et al., 2012). It is a short task, involving only

11 items, and consists of basic probability questions. These 11 items assess how well people can transform probabilities to percentages, percentages into probabilities while also performing simple mathematical operations using percentages or probabilities. The possible total sum scores range 0 to 11, where higher scores indicate better numerical skills compared to lower scores.

The Cognitive Reflection Test (CRT) (Frederick, 2005) is a three-item test, which is designed to assess individual's ability to suppress an impulsive wrong answer in place of a more deliberative cognitively processed correct answer. CRT reveals a reflective thinker over an impulsive as the most intuitive answer of the task is the wrong one. The individual needs to reflect before finding the solution. This measure is scored by adding up the correct answers. Participants who scored 0 and 1 out of 3 were classified as low reflective thinkers, and those who scored 2 or 3 were classified as high reflective thinkers (Frederick, 2005).

The Holt and Laury standard version comprises of two options in ten different rows. The probabilities for the higher amounts are 10% and 90% for the lower amounts. The probabilities change from row to row while the payoff remains the same. Hence, the expectation values of the two options change in each row. In the first four rows, the expected value for option A is safer, and option B is riskier. Form the fifth row, and below the risky option, B has higher expected value. If participants are consistent with their choices, they change after some point from option A to option B. The time that they switch over, determines their risk attitude. All rows are presented at once to the participants, and they are asked in each row to decide which option they would prefer. This measure was not used as a risk assessment. Rather, it was used to identify whether any of the other parameters such as education, numeracy, cognitive thinking style or external representations would be able to predict the rational evaluation of option and which external representation used by the participants would facilitate consistent choices.

Procedure Participants were given the participant information sheet informing them about the study and the informed consent form to sign. After the informed consents were obtained, participants started filling in the tasks using pen and paper. On average, each participant needed thirty minutes to complete all of the questionnaires.

The participants were given the Holt and Laury task and they were instructed to answer using whatever external representation was more appropriate to them. All of the participants had 20 minutes to complete the Holt and Laury task.

Results The independent variables were the education, numeracy, cognitive thinking style and external representational way. The dependent variable was the rational choice of the probabilities in the Holt and Laury task. In the context of this paper, is reflected as the participants' random choices in the task which indicate that participants either change lotteries in each row or choose only one Option between the two lotteries over the ten rows which has been supported as an inconsistent pattern by other studies (Jacobson & Petrie 2009; Dave et al. 2010). A coding system was used to categorize the external representations that participants used for the Holt and Laury task. The coding of the external representations was based on the coding adapted from previous research studied (Corter & Zahner, 2007; Zahner & Corter, 2010). The identified types were numbers, graphs, pictures, non-diagrammatic (text), and we added the blank page that it was not included in the coding method of Corter and Zahner (2007). Each representation was coded with one according to the above-mentioned types and added to a table. For instance, if a participant approached the problem solution using numbers, pictures and words then these types were coded with 1 and the rest, graphs and the blank page with zero. To assess the reliability of the coding, two independent raters coded the responses. Cohen's kappa was run to determine if there was agreement between the two raters'. A Cohen's kappa of .957 and .977 represented almost absolute agreement between the two examiners for each of the five categories in the Holt and Laury task. To test the hypothesis whether participants who graduated from university would be more likely to answer rationally in the Holt and Laury task, a chi-square test of independence was performed. This test showed that participants who graduated from university were not more likely to answer rationally in the Holt and Laury compared to participants who had not graduated from university, $X2$ (1, N=60) = .429, p = .513.

To determine the relationship between specific variables and the rational decision making in Holt and Laury task, a chi-square test of independence was used. Therefore, to examine whether participants who scored lower in the validated numeracy scale, was associated to their rational choices in the Holt and Laury, a chi-square test of independence was performed. Numeracy performance was divided into two groups, one group with participants who were scored high (9-10-11 correct) and another group with those that scored less (2-8 items correct). Because the distribution of data was highly skewed, as mean numeracy was 8.1 out of 11 (a =.63), a median split was used for analysis, although it was taken under consideration that this split can cause loss of power (Peters et al. 2006; MacCallum, et al., 2002). The data were binary (0 for most numerate and 1 for less numerate). The chi- square was statistically significant X^2 (1, N= 60) = 4.176, p = .041, showing that people who scored higher in the validated numeracy scale had a greater chance of choosing rationally in the Holt and Laury compared to those who scored lower.

A chi-square test of independence was also performed to find out whether gender is associated with answering rationally or not in the Holt and Laury, to exclude it as a factor which influences the rational decision-making in this task. The results showed that there is no gender association with participants' rational choices, X^2 (1, N=60) = .019, p = .890. To test the hypothesis that participants who had graduated from university, would be more likely to use different

external representations compared to the participants who did not attend university, a chi-square test of independence was performed which showed that there is no difference among the external representations both groups used, X^2 (4, N= 60) = 3.642, p = .457. To test the hypothesis that there was an association with correct answers in the CRT and the rational choices in Holt and Laury, Fisher's Exact Test was performed. It revealed that participants who answered correctly more questions in the cognitive reflection task, they were more likely to answer rationally in the Holt and Laury, p<.05. The relation between numeracy and CRT was not examined as it was out of the context of the study.

A logistic regression was performed to investigate if using any specific way of external representations would be more likely to predict a rational answer. The logistic regression model was statistically significant at p <.02 according to the model chi-square statistic, meaning that the use of external representations can predict the rational choices in Holt and Laury. According to the logistic regression, graphs were shown to predict the rational choices in Holt and Laury and blank page, which included the answers where there was no verbal, mathematical or visual decision making process to reason the choices in the task and resulted in irrational choices (Table 1). The software used for the statistical analysis was SPSS.

Table 1. Statistical significance of the independent variables in the logistic regression whether any of the external representation would be more likely to predict a rational answer.

	B	S.E.	Wald	df	Sig	Exp(B)
Numbers	-1.438	1.478	.947	1	.330	.237
Graphs	3.095	1.431	4.675	1	.031	22.085
Pictures	.888	1.469	.365	1	.546	2.430
Words	.155	1.205	.017	1	.898	1.168
Blankpage	3.280	1.627	4.063	1	.044	26.577
Constant	-1.695	1.362	1.548	1	.213	.184

Experiment 2

Method

Participants In total 225 undergraduate computer science students, from a UK University, completed the tasks and the questionnaires. Participants in this study included 66 females and 159 males (M age = 29, age range 18 – 32). The allocation of the participant in three groups that were divided based on the respective format of Holt and Laury was randomized. Participants entered a lottery to win a £50 Amazon voucher as incentive to take part in the experiment, which was communicated during the introduction session.

Materials The materials used for this study include the Holt and Laury standardized version, the digital Holt and Laury displayed with pie charts and the digital Holt and Laury asking participants to fill in the pie charts before choosing the option.

In the Holt and Laury task displayed with pie charts (Figure 1), every option between the two lotteries is represented with a pie chart, and there is a text describing the proportions about the payoffs on the pie chart. Next, to the pie chart, the relevant payoff was displayed textually. The task was deployed in Unity Game Engine and logged all user actions and choices.

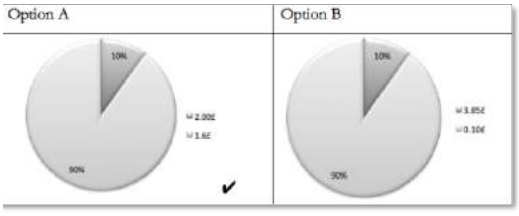


Figure 1. Holt and Laury displayed with non-interactive pie chart.

The stimuli developed for this study (Figure 2), involved 10 "empty" pie charts divided in 10 pieces with the proportions of each probability option of the lottery presented textually above them. Participants had to click on the region of the pieces on the pie chart to "fill them in" with specific colours according to the probabilities outlined in text and mark the pay offs of those pieces. For example, fill in with red one out of ten pies and fill in with blue nine out of ten. This way the participant can see clearly which option is more likely to happen. After the participants filled in the pie charts, they would choose one option. The stimuli was developed in a way that it did not allow the participant to choose the option before "filling in" the pie chart. Visually, the risk-taking task was very simple reflecting the standardised pen and paper HL task to avoid the effect of visual elements on participants' decision making process.

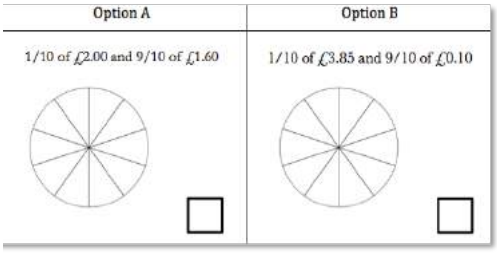


Figure 2. The empty circles where the participants had to "fill in" the parts for each pie chart before choosing the option.

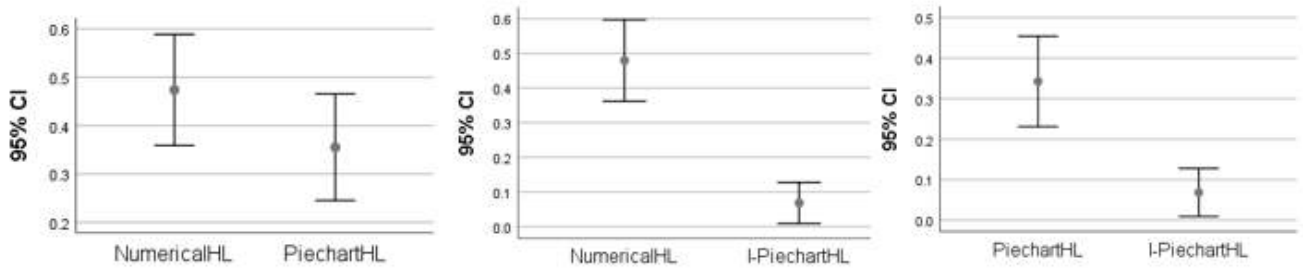


Figure 3. Participants' choices over the three different display formats with 95% Confidence Intervals. Dependent variable was the inconsistency rate in each display format of the Holt and Laury task. The lower mean shows a more consistent rate.

The numeracy test used is the Lipkus Numeracy scale (Lipkus, Samsa, & Rimer, 2001). A couple of demographic questions regarding participant's age and gender and a self-report of difficulty task were also involved in the experimental procedure. After they had completed the Holt and Laury tasks. They were asked to answer a question about task's level of difficulty. The question about difficulty level was formed as follows: On a scale of 1-5 with one being very easy and five being very difficult, how difficult was this lottery task for you? The participants were asked to answer in a 5-pont Likert scale where one was very easy and five was very difficult.

Procedure The experiment took place at the Faculty of Computer and Engineering at Coventry University and lasted approximately 30 minutes. It was divided into three stages. In the first stage, participants were given the participant information sheet informing them about the study and the informed consent form to sign. After the consent forms were obtained, in the second stage, the participants were given the numerical test and the demographic questions. When these were also collected, in the final stage, the participants were directed to five different rooms. Four rooms had computers where the interactive and non-interactive stimuli were set up. In the fifth room, the standardized Holt and Laury task was set up. The allocation of the participants to the rooms was random. The four rooms with the computer could fit 40 people each and the fifth room was a lecture theatre for 200 students' capacity. Participants were instructed that they would have 20 min to fill in the tasks. The participants were randomly assigned to go into one of the five rooms. Finally, 76 participants filled in the numerical display, 76 completed the pie chart display and 73 the interactive pie chart Holt and Laury task. Hence, Group 1 received the textual Holt and Laury task, Group 2 the Holt and Laury digital task displayed with pie charts and Group 3 received the digital interactive Holt and Laury method. The results are presented in the following section. After the completion of the task, each participant selected a small note from a lottery ball, all notes included numbers except from one that had the letter A and was referring to an Amazon Voucher of £50.

Results According to participants' irrational choices, 35 out of 76 (46%) participants showed an inconsistent behavior in Holt and Laury task numerical format, 27 out of 76 (35.52%) participants in the pie chart format and only 4 out of 73 (5.47%) participants in the interactive pie chart Holt and Laury task format (Figure 4). McNemar's test for related samples was applied between the interactive pie chart and numerical Holt and Laury format, which revealed a significant difference in the inconsistency rates between both display formats, $p < .00$ (Figure 3). McNemar's test for related samples was also applied between interactive pie chart and pie chart Holt and Laury format w hich showed a significant difference between their inconsistency rates, $p < .00$ (Figure 3). McNemar's test for related samples between the pie chart format and the numerical Holt and Laury format showed no significant difference between their inconsistency rates (Figure 3).

Additionally, a binary logistic regression was performed to ascertain whether participants' irrational choices in each display format (numerical, pie chart and interactive pie chart) could be predicted based on their age, gender, the level of their perceived difficulty and numeracy score in the validated scale. For the numerical Holt and Laury format, the binary logistic regression was statistically significant at the .00 level according to the model chi-square statistic suggesting that numerical level (Wald statistic equal to 12.2), difficulty perception (Wald statistic equal to 8.8) and the age (Wald statistic equal to 4.7), were shown to be significant at the .00 level. Hence, they can predict participants' choices in the Holt and Laury numerical task. For the pie chart Holt and Laury, the binary logistic regression showed that the model is not statistical significant, $p < .178$. The binary logistic regression for interactive pie chart Holt and Laury was statistical significant at the level of .025 according to the model chi-square statistic. The coefficient on the perceived difficulty had a Wald statistic equal to 5.42, which is significant at the .02 levels. The rest dependent variables of age, gender and numerical level were not statistically strong predictors of participants' choices in the interactive Holt and Laury format.

Participants took from 4 to 17 minutes to complete the Holt and Laury task using each representation type. The mean time taken to answer the task was 8.7 minutes with the interactive

pie chart, 8.3 using the passive pie chart and 8.5 for the numerical. Even though the mean differences show that using the passive pie chart took them slightly less to fill in compared to the interactive and numerical, there is no significant difference between the times spent in each format.

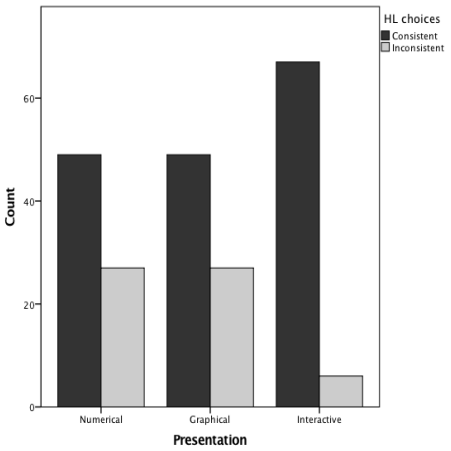


Figure 4. Participants' choices in relation to the presentation method of Holt and Laury.

Discussion

This paper presents the results of two experiments that investigated the methods that would assist in choosing rationally in a risk elicitation task. The findings from the 1st experiment validated empirical evidence that the use of external representations helps participants to understand the lotteries in the risk elicitation task. It also showed that using graphs has greater likelihood of choosing rationally in a risk elicitation task, compared to using images or any other way of external representation, as it has been supported by Hegarty and Kozhnenikov (1999). The study also showed that participants' years of education did not influence their rational decision making. However, as participants were students from the United Kingdom this may not hold true for other cultures or audiences. Even though, it was found that there were no disparities between Groups 1 and Groups 2 with substantially dissimilar educational levels, this may not be true for educational levels defined in qualitatively different ways from other participants, or contexts.

Pie charts were shown to be better suited as they are better known from the general public and easily comprehensible to compare the size of two proportions when they are accompanied by labels (Nelson, Hesse, & Croyle, 2009). A step further though was demarcated in terms of asking participants to draw the pie charts themselves according to the label of probability along with the payoff displayed on the task, to test whether using external representation of pie charts would help them reason rationally. This outcome confirms the findings suggesting using graphs as external representation to facilitate successful probability problem solving (Hegarty, & Kozhevnikov, 1999) and thus rational choices in the Holt and Laury task. In the 2nd experiment, it was shown that participants who scored low in the numeracy scale chose more rationally in the Holt and Laury task when interacted with pie charts than any other format, confirming the need for a visual reasoning process to assist problem solvers (Carlsson, Johansson-Stenman, & Martinsson, 2004; Brase 2009). The interactive pie chart format was filled in rationally from those that scored high and low on the numeracy scale, as it was indicated by the low consistency rate in the task (6.8%). This has a significant implication for future implementations of the task. This rational choice is linked to a meaningful contribution to the task for three main reasons. First, the task could be used by people with low numerical skills and reflect their accurate risk preferences. Second, when assessing population with specific characteristics to predict risk-taking behaviour in similar investment choices, participants' choices in the task would be accurately predicted. Third, there would be less noise in the data and there would be less cases (if any) that data would be excluded from the analysis. However, a point that should be considered is that as the individuals are guided to answer consistently, the more consistent their answers the more variance would tend towards zero, thus the validity of the metric might be affected. Therefore, for future studies the validity of the metric needs to be examined and reassured. As this approach was only examined with UK University students, there is a limitation on generalizing to other cultures and audiences. Deploying a qualitative approach in conjunction to the quantitative methodology would enable to investigate more in depth on the underlying factors of why interacting with pie charts help people to understand the lotteries better. Finally, even though the average time spent in each display format of the task did not show any significant differences, future studies, need to explore whether using graphs to reason the choices in the Holt and Laury task, force the individuals with impulsive cognitive style to reflect more on their choices and aid their decision making.

This interactivity with graphs where participants could engage with filling in the proportions of lotteries with one click and then choose the option for the task they would prefer, automatically simplifies the task for less numerate people and allows for employing digital mediums, such as mobiles, for experimentation outside of laboratory settings. The data supports the hypothesis that the use of interactive pie charts, is more likely to result in consistent choices. This outcome may extend to other interactive artefacts such as games, simulations or analytics software, for crowdsourcing data for cognitive science of specific groups' (e.g. farmers) outside of the laboratory.

References

Andersen, S., Harrison, G.W., Lau, M.I. & Rutström, E.E. (2006), "Elicitation using multiple price list formats", *Experimental Economics,* 9(4), pp. 383–405.

Bauermeister, G. & Mußhoff, O. (eds.), (2016), "Risk Attitude and Inconsistencies-does the Choice of Display Format and Risk Elicitation Method Influence the Outcomes? "Paper presented at *the 56th Annualbrod Conference, Bonn, Germany, September 28-30, 2016.German Association of Agricultural Economists*

Bougherara, D., Gassmann, X. and Piet, L. (2011), "A Structural Estimation of French Farmers Risk Preferences: An Artefactual Field Experiment".

Brase, G.L. (2009). Pictorial representations in statistical reasoning. *Applied Cognitive* Psychology,23, 369-381

Carlsson, F., Johansson-Stenman, O. and Martinsson, P. (2004), "Is transport safety more valuable in the air?, *Journal of Risk and Uncertainty,* 28, pp. 147–163.

Corter, J. E., & Zahner, D. C. (2007). Use of external visual representations in probability problem solving. *Statistics Education Research Journal*, *6*(1), 22-50.

Dave, C., Eckel, C. C., Johnson, C. A., & Rojas, C. (2010), "Eliciting Risk Preferences: When is Simple Better?", *Journal of Risk and Uncertainty,* 41(3), pp. 219-243.

Deitrick, S., Edsall, R., 2006. "The influence of uncertainty visualization on decision making: An empirical evaluation", in: Riedl, A., Kainz, W., Elmes, G.A. (Eds.), Progress in spatial data handling. Springer, Berlin, pp. 719–738.

Frederick, S. (2005), "Cognitive Reflection and Decision Making", *The Journal of Economic Perspectives, Vo 19* No 4, pp. 25-42.

Gilbert, J.K. (2008), "Visualization: An emergent field of practice and enquiry in science education." *In Visualization: Theory and practice in science education, pp. 3-24*. Springer, Dordrecht.

Gray, M., & Holyoak, K. (2018),"Individual Differences in Relational Reasoning*", Proceedings of the 40th Conference of the Cognitive Science Society*

Habib, S., Friedman, D., Crockett, S., & James, D. (2016), "List Construction and Lottery Presentation Modulate Multiple Price List Responses".

Hegarty, M. and Kozhevnikov, M. (1999), "Types of visual–spatial Representations and Mathematical Problem Solving", *Journal of Educational Psychology, 91*(4), pp. 684.

Holt, C. A., & Laury, S. K. (2002), "Risk Aversion and Incentive Effects", *American Economic Review,* 92(5), pp. 1644-1655.

Jacobson, S., & Petrie, R. (2009). Learning from mistakes: What do inconsistent choices over risk tell us?. *Journal of risk and uncertainty*, *38*(2), 143-158.

Khan, A., Breslav, S., & Hornbæk, K. (2018). Interactive instruction in bayesian inference. *Human–Computer Interaction*, *33*(3), 207-233.

Lipkus, I. M., Samsa, G., & Rimer, B. K. (2001), "General Performance on a Numeracy Scale among Highly Educated Samples", *Medical Decision Making,* 21(1), pp.37.

MacCallum, R.C., Zhang, S., Preacher, K.J., & Rucker, D.D. (2002). On the practice of dichotomization of quantitative variables. Psychological Methods, 7, pp. 19–40.

Nelson, D.E., Hesse, B.W. and Croyle, R.T., 2009. "*Making data talk: Communicating public health data to the public, policy makers, and the press".* New York, NY: Oxford University Press.

Nielsen, T., Keil, A. and Zeller, M. (2013), "Assessing Farmers Risk Preferences and their Determinants in a Marginal Upland Area of Vietnam: A Comparison of Multiple Elicitation Techniques*", Agricultural Economics,* 44(3), pp. 2.

Padilla, L., Creem-Regehr, S. H., Hegarty, M., & Stefanucci, J. K. (2018). Decision making with visualizations: a cognitive framework across disciplines. *Cognitive Research: Principles and Implications,* 3(1), 29.

Peters, E. (2008), "Numeracy and the Perception and Communication of Risk" *Annals of the New York Academy of Sciences,* 1128(2), pp. 1-7.

Peters, E., Västfjäll, D., Slovic, P., Mertz, C. K., Mazzocco, K., & Dickert, S. (2006). Numeracy and decision making. *Psychological science*, *17*(5), 407-413.

Schapira, M. M., Walker, C. M., Cappaert, K. J., Ganschow, P. S., Fletcher, K. E., McGinley, E. L., Del Pozo, S., Schauer, C., Tarima, S. and Jacobs, E. A. (2012), "The Numeracy Understanding in Medicine Instrument: A Measure of Health Numeracy Developed using Item Response Theory*".Medical Decision Making,* 32(6), pp. 851-865.

Rohrmann, B. (2005). Risk attitude scales:concepts, question utilizations. *Project Report, 1-21.*

Starns, J. J., Cohen, A. L., Bosco, C., & Hirst, J. A (2018), A visualization technique for Bayesian reasoning*. Applied Cognitive Psychology,* 33(2), pp. 234-251.

Zahner, D., & Corter, J. E. (2010). The process of probability problem solving: Use of external visual representations. *Mathematical Thinking and Learning*, *12*(2), pp. 177-204.

Ziemkiewicz, C., Ottley, A., Crouser, R.J., Chauncey, K., Su, S.L., & Chang, R. (2012), "Understanding visualization by understanding individual users", *IEEE computer graphics and applications, 32*(6), pp.88-94.

Zhang, J. (1997). The nature of external representations in problem solving. Cognitive science, 21(2), pp. 179-217.

Adaptation Aftereffects as a Result of Bayesian Categorization

Marina Dubova (marina.dubova.97@gmail.com)
Saint Petersburg State University, Department of Psychology, 6 Makarova Embankment
Saint Petersburg, 199034 Russia

Arseny Moskvichev (amoskvic@uci.edu)
University of California Irvine, 2277 Social
Behavioral Sciences Gateway Building
Irvine, CA 92697 USA

Abstract

We propose a unified explanation of contrastive and assimilative adaptation aftereffects from the perspective of higher-level cognitive processes: rational category learning and categorical perception. We replicate (twice) previously reported assimilative and contrastive effects (Uznadze illusion in visual modality), propose a rational computational model of the process, and evaluate our model performance against the Bayesian logistic regression baseline. We conclude by discussing theoretical implications of our study and directions for further research.

Keywords: adaptation aftereffects, perceptual biases, set illusion, Uznadze illusion, computational modeling, categorical perception

Introduction and Background

In many experimental settings, repeated exposure to stimuli affects the perception of subsequent ones. These phenomena are often referred to as the aftereffects of adaptation (Gibson & Radner, 1937). For example, if a participant is repeatedly presented with two circles, one bigger than another, she might perceive equal circles as being different during the test trial (Figure 1). Similar effects are manifest across a wide range of experimental conditions, in different modalities, and on different levels of abstraction. Behavioral studies demonstrate adaptation aftereffects in situations that run the gamut from simple shape and motion perception under brief presentation (Suzuki & Cavanagh, 1998; Chalk et al., 2010) to perception and recognition of faces, facial expressions, gender, and race (Webster & MacLeod, 2011; Leopold et al., 2001).

Contrastive and assimilative effects

It is possible to split all known adaptation aftereffects into two broad categories: **contrastive** and **assimilative** (Howard & Rogers, 1995). Contrastive aftereffects take place when the test stimulus seems more **different** from those seen during the adaptation phase (adaptors) than it would be perceived under normal conditions. Assimilative aftereffects, in turn, produce a reversed effect: the test stimulus is perceived as being more **similar** to adaptors. There is evidence that these two types of effects could occur in very similar and even identical experimental settings (Uznadze, 1958; Fritsche et al., 2017; Chopin & Mamassian, 2012). This raises a question: **what determines whether a contrastive or assimilative aftereffect will be present in a given trial**?

A broadly accepted view is that the probability of contrastive aftereffects occurrence grows with increasing difference between the test stimuli and the adaptors, increased length of adaptor presentation, as well as with the increase of overall stimuli salience and contrast (Howard & Rogers, 1995; Palumbo et al., 2017; Fritsche et al., 2017; Chopin & Mamassian, 2012).

Finding a mechanism that would explain the onset of both types of adaptation aftereffects turned out to be challenging. Previously dominant framework of adaptation as neural fatigue proved unsuccessful in accounting for the wide range of observed phenomena (Thompson & Burr, 2009). Recent studies predominantly focused on uncovering the mechanisms of a particular type of aftereffect: either contrastive (Webster & MacLeod, 2011; Rhodes & Jeffery, 2006; Grill-Spector et al., 2006; Stocker & Simoncelli, 2006; Chopin & Mamassian, 2012) or assimilative (Chalk et al., 2010; Palumbo et al., 2017).

There are, however, models that propose potential mechanisms of both contrastive and assimilative effects in visual (Wei & Stocker, 2015) and audial (Kleinschmidt & Jaeger, 2011) modalities. Wei and Stocker (2015) explained the opposite perceptual biases as a result of efficient coding constraints in a rational observer framework. Unfortunately, this model falls short in accounting for the influence of the difference between the test stimulus and the adaptors on illusion type (it predicts that this factor has no impact). At the same time, similar aftereffects in phonetic adaptation were modeled as Bayesian belief updating over two competing phonetic categories (Kleinschmidt & Jaeger, 2011). The limitation of this model is that it is designed for the task of forced choice between two categories that are given in advance. In most real-world and experimental adaptation scenarios, however, the alternative categories are implicit.

Overall, none of the existing models provide a complete account of the existing phenomena, which warrants further research in this direction.

Adaptation aftereffects and categorization

We propose a high-level interpretation of adaptation biases from a categorization standpoint. We argue that during the adaptation phase a person forms the categories of "typical" and "other" (atypical) stimuli. Learning is formalized using the ideal observer approach. The structure of the "typi-

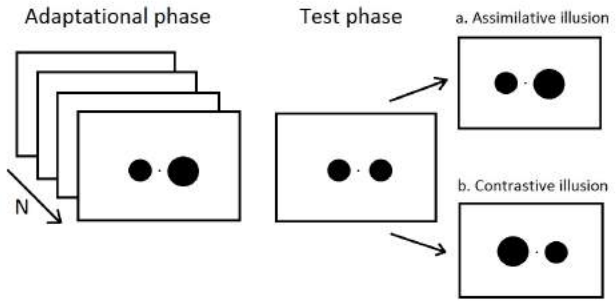


Figure 1: Uznadze visual illusion experimental procedure
During the adaptation phase, a subject is repeatedly exposed to two circles, one being bigger than another. On the test trial, two equal circles are presented, and the subject responds whether they appear **the same (no illusion)** or, if not, which one appears bigger (**contrastive** or **assimilative** illusion).

cal" category is estimated from observed adaptors, while the "other" category is determined through its relationship to the already learned one. The main assumption is that an observer expects different visual categories to lie relatively far from each other in the feature space. On the test phase, the observer reconstructs the most likely true stimulus, given the learned category structure and the noisy sensory observation.

There is some evidence that provides conceptual support for our approach. First, in the domain of category learning, there is a notion of **categorical perception** which refers to phenomena whereby the same stimuli seem more different or similar, depending on whether they belong to the same or different categories in the learned conceptual structure (Goldstone, 1994, 1995; Goldstone & Hendrickson, 2010; Kuhl & Iverson, 1995). Second, in the domains of color and speech perception, perceptual bias toward the category prototype was formalized as an optimal statistical inference of real stimulus in high uncertainty conditions (Feldman et al., 2009; Cibelli et al., 2016). These perceptual shifts resemble the assimilative aftereffects. Third, a similar idea was successfully applied earlier in the domain of face perception: it was shown that the contrastive aftereffects are directed precisely toward the anti-prototype of the seen examples (Leopold et al., 2001, 2005; Rhodes & Jeffery, 2006). Assimilative aftereffects were not, however, considered in these studies.

Overall, there is evidence that category attribution plays an important role in perception. Although visual adaptation is most commonly viewed as a low-level process, current low-level models may not be able to fully capture the broad spectrum of visual adaptation aftereffects and their dynamics (Leopold et al., 2005) and are hardly compatible with interocular transfer of adaptation biases (Raymond, 1993). We believe that the difficulties encountered by low-level explanations, together with the successes of categorical perception models, warrant considering alternative, high-level explanations of perceptual aftereffects.

Our model builds upon the previous results and provides a simple and unified interpretation of both assimilative and contrastive aftereffects from a categorical perception standpoint.

To test our interpretation, we use a visual version of the Uznadze illusion (Figure 1). We replicate previously reported results on the association between the probabilities of opposite illusions with the length of the adaptation phase and the difference between the adaptation stimuli (Uznadze, 1958, 1966). After that, we evaluate the performance of our model on these data.

Experiment 1

This experiment replicated the findings reported in Uznadze (1958, 1966).

Hypothesis: Difference between the adaptors and the test stimulus, together with the number of adaptation trials, determine the probability of assimilative vs contrastive aftereffect occurrence. In particular, the assimilative aftereffect is associated with lower differences between stimuli sizes and smaller numbers of adaptation trials, while the contrastive aftereffect onset probabilities follow a reversed pattern.

Procedure

Pairs of circles of different sizes were presented as adaptors. We varied the magnitude of difference between adaptation stimuli (from 1 to 3 individual differential thresholds) and the number of adaptation trials (from 1 to 8) to evaluate their effect on the probabilities of assimilative and contrastive illusions. The procedure is illustrated in 1.

1. **Estimation of individual differential thresholds.** Two circles (diameters: left 2.5cm, right 2.5 or 3.0cm) were presented to participants. They were asked to focus on the dot in the center of the screen. We estimated participants' differential thresholds by the method of adjustment (Gescheider, 1997). That is, subjects saw two different circles and altered the size of one of them until the circles appeared equal to each other. In the second condition, the circles were initially the same and subjects made them different. We repeated this procedure six times and averaged the results to obtain the differential threshold estimate.

2. **Adaptation phase.** Subjects focused at a central dot on the screen, while they were exposed to two circles (for 150 ms) several (1-8) times. The difference in size between the two circles was 1, 2 or 3 individual differential thresholds.

3. **Test phase.** Participants saw two equal circles for 150ms and reported whether they appeared the same. If there was a perceived difference, participants identified which of the two appeared larger. They were instructed to respond as fast as possible and to rely only on their sensations. The test trial was repeated until the "same" relationship was reported 3 times in a row. This ensured that the aftereffect has faded before the start of the next trial. We did not analyze

the fading dynamics and only used the first test response in further analysis.

This procedure was repeated 24 times for every participant using all the combinations of experimental conditions. The order of conditions was randomized.

4. **Post-experimental interview.** Participants shared their experience and strategy. The results of this stage were used to check whether subjects responded purely based on what they saw (as opposed to realizing that they experience an illusion and correcting their answers).

Experiment was programmed and presented using PsychoPy software package (Peirce, 2007).

Participants

The initial sample consisted of 30 adult participants. Data from 4 participants were excluded after the post-experimental interview: they figured out that test circles are always equal, and based their answers on this assumption, not on their actual perception. This results in a final sample of 26 participants (11 male, 15 female) aged from 18 to 47 years (mean age: 22.27, sd: 5.65). All had normal or fully corrected vision.

Experiment 2

The second experiment investigated how robust are the observed regularities. In particular, whether it is necessary to account for individual differential thresholds.

Procedure

Experiment 2 replicates Experiment 1 with one qualitative change: the difference between adaptation circles varies in **absolute units**, not in individual differential thresholds. Therefore, there is no stage of differential threshold estimation. The left circle again has the diameter of 2.5cm, and the diameter of the right circle is 0.1, 0.2, 0.3, 0.4, or 0.5cm bigger. The number of adaptation trials varies from 1 to 6. The conditions are randomized.

Participants

Initial sample consisted of 55 adults. 5 adults were excluded from subsequent analyses, because they figured out that test circles are always equal and based their responses on this assumption. This results in a final sample of 50 participants (22 male, 28 female) aged from 18 to 34 years (mean age: 22.91, sd: 3.47). All of them had normal or fully corrected vision. The sample was divided into two groups based on the results of post-experimental interview:

1. **Naive (35 adults).** These participants did not realize that test circles are always equal.

2. **Non-naive (15 adults).** These participants realized that test circles are always equal, but followed the instruction and tried to base their responses only on their sensations.

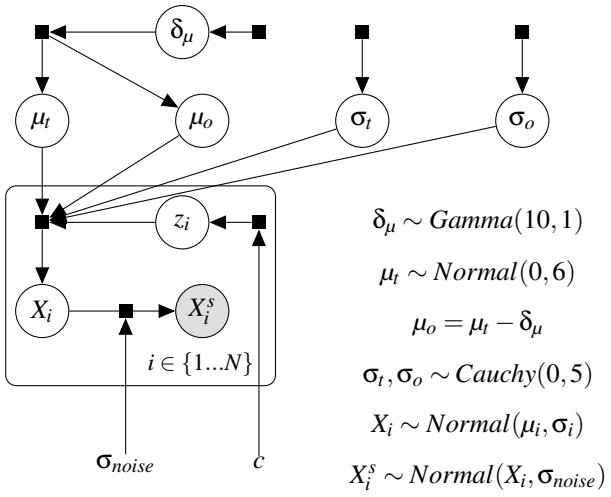


Figure 2: Graphical model

Variables: X_i - real stimulus; X_i^s - perceived stimulus (after adding perceptual noise); z_i - indicator variable for the class from which a real stumulus was generated (distributed according to the Chinese Restaurant Process); μ_t and σ_t - μ and σ of the typical class; μ_o and σ_o - μ and σ of another (unobserved) class; δ_μ - the expected difference between two classes; c - coupling probability for CRP; σ_{noise} - perceptual noise.

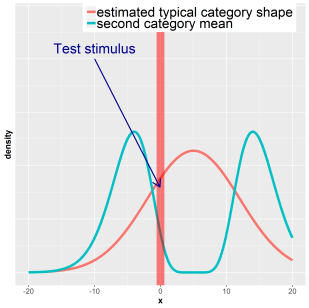
Computational Model

We formalize the process of adaptation as rational acquisition of the "typical" and "other" stimuli categories. Perception is modeled as an optimal probabilistic inference over the true stimulus parameters given the learned category structure and the noisy sensory input. Graphical model is presented on Figure 2.

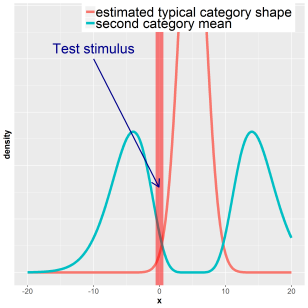
1. **Category learning**. Category learning during the adaptation phase is modeled as Bayesian inference of the "typical" category structure. Stimuli in our experiment could be aligned along one relevant dimension (the magnitude of difference between two circles) so we formalize a category as a univariate normal distribution in this dimension. An observer assumes that the true stimuli come from a normal distribution with a mean difference between two circles μ and a standard deviation σ. These are the parameters she estimates to represent a category. Priors on the category's μ and σ are chosen arbitrarily and are set to be relatively diffuse (given the scale of our feature): $\mu \sim normal(0,6)$ and $\sigma \sim cauchy(0,5)$. These parameters are estimated purely from the adaptive stimuli for every particular experimental trial. The adaptive stimuli, in turn, are randomly generated from a normal distribution where μ is the difference between adaptional stimuli in a given condition, and σ_{noise} is some random perceptual noise. We used 0.2 as a noise in final model evaluation, however, we also checked that changing this number does not influence the results. At the end of this stage, the observer has estimations of μ and σ

of the category.

2. **Representation of the unknown category**. The key assumption of a rational observer in our model is that the **centers of two categories are more likely to be relatively distant from each other** than to be close (Figure 3). It is formulated by adding a new parameter: *difference between category prototypes* δ_μ with a prior $Gamma(10,1)$ (modeling the opposite symmetric tail is not necessary in this case, as its likelihood is always practically zero on the test trial). Then, the estimation of the "other" category's mean (μ_o) for each experimental condition is simply $\hat{\mu}_t - \delta_\mu$. Hence, the prior assumptions on the structure of the unknown category are shifted outward from the learned one. Thus, the prior on μ_o is completely defined by an estimated μ of the typical stimuli and the assumption on the difference between categories.



(a) After a small number of trials, there is still a lot of variation in the estimated typical distribution shape. The noisy test stimulus is attributed to the typical category with higher probability and thus the reconstructed true stimulus is shifted towards the "typical" category prototype.



(b) After more trials, the estimated typical distribution shrinks, thus making the attribution to the typical category unlikely. Thus the reconstructed source of the noisy stimulus is shifted towards the closest peak of the "other" category.

3. **Test phase**. Perception of the test stimulus is determined by the decision of what category ("typical" or "other") is more likely to have generated it. Conditional probabilities of the categories are calculated using Bayes' rule (where z_i is a variable indicating category membership):

$$P(z_i = j|X_i^s) = \frac{f(X_i^s|z_i = j)\cdot P(z_i = j)}{f(X_i^s)} = \\ = \frac{f(X_i^s|z_i = j)\cdot P(z_i = j)}{\sum_{j=1}^{\#cat} f(X_i^s|z_i = j)\cdot P(z_i = j)} \quad (1)$$

Likelihoods of the test stimulus for both categories are taken from the corresponding estimated normal probability density functions. The priors on whether a new stimulus is coming from the known or a new category are estimated using the Chinese Restaurant Process (Anderson, 1991; Navarro & Kemp, 2017). Thus, the prior probability that a new stimulus is generated from the "typical" category is

$$P(z_{n+1} = typical) = \frac{c\cdot n}{1-c+c\cdot n} \quad (2)$$

where c is a probability that two observation come from the same category (the coupling probability) and n is a number of adaptation trials. The prior probability that a new stimulus comes from an unknown category is

$$P(z_{n+1} = other) = \frac{1-c}{1-c+c\cdot n} \quad (3)$$

To efficiently reconstruct a real stimulus, perception is shifted toward the probability density of its category. Due to the aforementioned inference bias, the "atypical" and "typical" category densities are shifted in opposite directions. Thus, assimilative illusion onset is formalized as *Bernoulli* random variable with $p = P(typical|test)$, and the contrastive - as *Bernoulli* r.v. with $p = P(other|test)$ respectively.

Model fits 3 parameters: c (coupling probability), δ_μ (difference between prototypes of two categories), and σ (standard deviation of the "other" category).

Bayesian modeling for the paper was implemented using Stan probabilistic language (Carpenter et al., 2017).

Results

Experiment 1

Assimilative aftereffect appeared 103 times (17%), contrastive - 153 times (25%). Notably, more than 50% of the data consisted of the reports of stimuli equality, which correspond to no illusion registered. "No illusion" instances were excluded from the analysis. We applied mixed effects logistic regression and Bayesian mixed effects logistic regression (with non-informative priors). We used the difference between adaptation circles and the number of adaptation trials as predictors, and the illusion type (contrastive (1) vs assimilative (0)) as the outcome variable. This model can be expressed using the following formula:

illusion type $\sim$ number of adaptation trials + difference between stimuli sizes + (1|participant)

The ANOVA comparison with a zero model was significant ($p < .001$), as well as the tests for both individual coefficients: number of adaptation trials ($p < .001$, $est. = .133$, $sd = .036$, $BF_{10} = 2.5$) and difference between stimuli' sizes ($p < .001$, $est. = .365$, $sd = .108$, $BF_{10} = 56.3$). Both estimates are positive, in line with the the initial hypotheses.

Experiment 2

Assimilative aftereffect appeared 164 times (11%), contrastive - 402 times (28%). To analyse these data, we applied the same frequentist and Bayesian mixed effects models to the three (all, naive, and non-naive) groups separately.

1. For the whole sample, the difference between adaptation circles and the number of adaptation trials are significant predictors with $p < .001$ ($est. = .537$, $sd = .093$, $BF_{10} = 47995.7$) and $p < .05$ ($est. = .171$, $sd = .07$, $BF_{10} = 4.8$) respectively.

Table 1: Performance of Cognitive Model and Bayesian Logistic Regression.

Standard deviations are indicated in parentheses.

	Measure	Bayesian LR	Cognitive Model
Experiment 1: assimilative	Recall	0.296 (0.086)	**0.577** (0.082)
	Precision	0.521 (0.12)	**0.522** (0.034)
Experiment 1: contrastive	Recall	**0.817** (0.065)	0.65 (0.056)
	Precision	0.637 (0.018)	**0.701** (0.032)
Experiment 2: assimilative	Recall	0.057 (0.0049)	**0.293** (0.057)
	Precision	0.378 (0.228)	**0.426** (0.044)
Experiment 2: contrastive	Recall	**0.97** (0.03)	0.845 (0.032)
	Precision	0.73 (0.006)	**0.754** (0.012)

2. For the group of naive participants, the difference between adaptation circles is a statistically significant predictor ($p < .01$, $est. = .489$, $sd = .163$, $BF_{10} = 215.5$). The number of adaptation trials is not significant ($p > .05$, $est. = .105$, $sd = .9$, $BF_{10} = 0.5$).

3. For the non-naive participants, both predictors are statistically significant: the number of adaptation trials ($p < .05$, $est. = 1.21$, $sd = .523$, $BF_{10} = 23.8$) and the difference between stimuli sizes ($p < .001$, $est. = .518$, $sd = .151$, $BF_{10} = 254.6$).

The subsequent ANOVA model test (frequentist) was significant ($p < .05$) for all groups. All the estimates are positive, in line with the initial hypotheses.

Model Evaluation

We compared our cognitive model against the Bayesian logistic regression baseline:

$$\textit{illusion type} \sim \textit{number of adaptation trials} + \textit{difference between stimuli sizes}$$

Both the regression and the cognitive model have 3 parameters. Bayesian logistic regression was chosen as a baseline, since it is a very successful descriptive model with the same amount of parameters. In particular, it outperforms a frequentist logistic regression for our data.

The cognitive model fits the whole dataset better than the baseline logistic regression models, but this does not guarantee that the cognitive model would demonstrate better performance on the out-of-sample data as well. Therefore, we used random subsample cross-validation in order to evaluate and compare the **generalization** performance of the models.

1. The data were randomly split into two subsets: train (50% of assimilative data, 50% of contrastive data) and test (remaining 50% of assimilative and 50% of contrastive data)

2. Parameters of the models were estimated on the training set

3. The performance measures (precision and recall) were calculated for models' predictions for the upheld test subset.

We repeated the above steps 50 times and calculated mean precision and recall measures for both assimilative and contrastive classes, along with their standard deviations. The results are shown in the Table 1.

Evaluation metrics: **Precision** and **Recall** measures allow us to compare models based on their sensitivity and accuracy for both classes. **Recall** shows the proportion of the target class occurrences that were accurately predicted. **Precision** shows the proportion of the target class occurrences among the predictions of that class.

The cognitive model repeats the main regularities found in both experiments. In particular, it predicts assimilative illusion more frequently for the smaller differences between adaptive stimuli and number of trials, while the predictions of contrastive illusion follow the reverse pattern. Importantly, the logistic regression does not yield these types of regularities when it predicts assimilative illusion.

The estimates of the "other" category center were always negative, which corresponds to the **contrastive** shifts in perception.

Discussion

Replication

We replicated the results reported in Uznadze (1958, 1966). The difference between adaptation stimuli sizes was a significant predictor of the aftereffect type in all collected datasets. The number of adaptational trials was not a significant predictor for naive participants in the second experiment, but it was significant in the remaining datasets. The signs of all the coefficients were consistent with the initial hypotheses. The effect proved robust to the scale of the differences between stimuli, and overall, Experiments 1 and 2 yielded similar results.

We view this replication as an important impact of our paper. The works of Uznadze are predominantly focused on the study of "set", or "set illusions", which denote the same group of phenomena as perceptual aftereffects. He performed extensive studies of these effects in visual, auditory and haptic modalities (Uznadze, 1966). Nevertheless, although the so-called "Uznadze illusion" (perceptual aftereffect in haptic modality) received some attention (Janzen et al., 1976;

Wohlwill, 1960), most of his contributions remain untranslated and almost entirely unknown to the scientific community outside the post-Soviet space. We find, however, that some of his findings are still relevant and could lead to a better understanding of perceptual aftereffects. We hope that our results would encourage further use of Uznadze visual illusion in the studies of perceptual adaptation aftereffects.

Modeling

The proposed cognitive model performs better than Bayesian logistic regression, which makes it a useful baseline for further research. In particular, it is sensitive to both types of aftereffect and yields more accurate predictions within these categories.

More importantly, our model provides a simple and unified interpretation of seemingly disparate phenomena of assimilative and contrastive aftereffects. This explanation is based on the principles of rational analysis and an intuitive assumption about the category structure inductive bias (different categories have non-coinciding centroids). Thus, our model shows that the apparently low-level perceptual aftereffects may be explained from the logic of higher-level cognitive processes, such as categorization. Moreover, it allows us to view the role of adaptation aftereffects in perception from a new angle: we demonstrate that they may serve as an important part of an optimal stimuli reconstruction process, as opposed to being an artifact or an epiphenomenon.

Future directions

Our model is based on the high-level logic of perception and is not bound to specific low-level mechanisms. This greatly broadens the scope of its potential applications.

Firstly, there is a number of promising extensions of our model **within the domain of adaptation aftereffects**:

- The model could be extended to account for the cases of **illusion absence** (this could be done by incorporating individual perceptual differential thresholds). Since the "no illusion" case is very common in our data, this would make our account of the perceptual aftereffect phenomenon much more complete.

- The model can be scaled to higher dimensions by using a multivariate normal distribution for category representation. This makes it a good candidate for describing perception of high-dimensional realistic objects, such as faces. In case of success, such a unified explanation of higher- and lower-level perceptual processes may contribute to the ongoing debate about the role and even mere presence of top-down effects in perception (Firestone & Scholl, 2016).

Secondly, our model may be broadly applicable **outside of the domain of visual perceptual adaptation**:

- There is a number of well-known spatial context effects in the visual modality (demonstrated by Delboeuf, Ebbinghaus, and Müller–Lyer illusions, among many others (Goto et al., 2007)). The patterns of contrastive and assimilative bias onsets in this domain are very similar to the temporal illusion we studied in this paper (Goto et al., 2007) and may be interpreted in an analogous fashion.

- Our proposed rational categorical perception model could account for **enhanced discriminability** and **perceptual tuning** effects resulting from long-term adaptation. Chinese Restaurant Process used in our model allows to optimally refine the learned category structure as a number of seen examples grows. Shifting percepts towards the true category will be more and more beneficial as the category structure is updated and refined.

Overall, the proposed model has a high promise in demonstrating the role of category learning in perception. While the potential importance of categorical perception has been studied before (e.g. Kuhl & Iverson (1995)), such studies focus on situations when the category structure is known in advance. Our results suggest that assuming that a person always tries to group stimuli into categories (even in the short-term experiments where no obvious categories are apparent) can greatly broaden the scope of this approach and provide a unified explanation to a wide range of perceptual effects.

To facilitate further research, we make all the data, analyses, and code openly available [1].

References

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological review*, *98*(3), 409.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., et al. (2017). Stan: A probabilistic programming language. *Journal of statistical software*, *76*(1).

Chalk, M., Seitz, A. R., & Seriès, P. (2010). Rapidly learned stimulus expectations alter perception of motion. *Journal of Vision*, *10*(8), 2–2.

Chopin, A., & Mamassian, P. (2012). Predictive properties of visual adaptation. *Current biology*, *22*(7), 622–626.

Cibelli, E., Xu, Y., Austerweil, J. L., Griffiths, T. L., & Regier, T. (2016). The sapir-whorf hypothesis and probabilistic inference: Evidence from the domain of color. *PloS one*, *11*(7), e0158725.

Feldman, N. H., Griffiths, T. L., & Morgan, J. L. (2009). The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. *Psychological review*, *116*(4), 752.

Firestone, C., & Scholl, B. J. (2016). Cognition does not affect perception: Evaluating the evidence for" top-down" effects. *Behavioral and brain sciences*, *39*.

Fritsche, M., Mostert, P., & Lange, F. P. de. (2017). Opposite effects of recent history on perception and decision. *Current Biology*, *27*(4), 590–595.

[1] https://github.com/blinodelka/Illusions_of_set

Gescheider, G. (1997). Chapter 3: The classical psychophysical methods. *Psychophysics: the fundamentals. 3rd ed. Mahwah: Lawrence Erlbaum Associates*.

Gibson, J. J., & Radner, M. (1937). Adaptation, after-effect and contrast in the perception of tilted lines. i. quantitative studies. *Journal of experimental psychology*, *20*(5), 453.

Goldstone, R. L. (1994). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General*, *123*(2), 178.

Goldstone, R. L. (1995). Effects of categorization on color perception. *Psychological Science*, *6*(5), 298–304.

Goldstone, R. L., & Hendrickson, A. T. (2010). Categorical perception. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(1), 69–78.

Goto, T., Uchiyama, I., Imai, A., Takahashi, S., Hanari, T., Nakamura, S., et al. (2007). Assimilation and contrast in optical illusions 1. *Japanese Psychological Research*, *49*(1), 33–44.

Grill-Spector, K., Henson, R., & Martin, A. (2006). Repetition and the brain: neural models of stimulus-specific effects. *Trends in cognitive sciences*, *10*(1), 14–23.

Howard, I. P., & Rogers, B. J. (1995). *Binocular vision and stereopsis*. Oxford University Press, USA.

Janzen, H. L., et al. (1976). A developmental analysis of set patterns in children: A normative study.

Kleinschmidt, D., & Jaeger, T. F. (2011). A bayesian belief updating model of phonetic recalibration and selective adaptation. In *Proceedings of the 2nd workshop on cognitive modeling and computational linguistics* (pp. 10–19).

Kuhl, P. K., & Iverson, P. (1995). Linguistic experience and the perceptual magnet effect. *Speech perception and linguistic experience: Issues in cross-language research*, 121–154.

Leopold, D. A., O'Toole, A. J., Vetter, T., & Blanz, V. (2001). Prototype-referenced shape encoding revealed by high-level aftereffects. *Nature neuroscience*, *4*(1), 89.

Leopold, D. A., Rhodes, G., Müller, K.-M., & Jeffery, L. (2005). The dynamics of visual adaptation to faces. *Proceedings of the Royal Society of London B: Biological Sciences*, *272*(1566), 897–904.

Navarro, D. J., & Kemp, C. (2017). None of the above: A bayesian account of the detection of novel categories. *Psychological review*, *124*(5), 643.

Palumbo, R., DAscenzo, S., Quercia, A., & Tommasi, L. (2017). Adaptation to complex pictures: exposure to emotional valence induces assimilative aftereffects. *Frontiers in psychology*, *8*, 54.

Peirce, J. W. (2007). Psychopypsychophysics software in python. *Journal of neuroscience methods*, *162*(1-2), 8–13.

Raymond, J. (1993). Complete interocular transfer of motion adaptation effects on motion coherence thresholds. *Vision Research*, *33*(13), 1865–1870.

Rhodes, G., & Jeffery, L. (2006). Adaptive norm-based coding of facial identity. *Vision research*, *46*(18), 2977–2987.

Stocker, A. A., & Simoncelli, E. P. (2006). Sensory adaptation within a bayesian framework for perception. In *Advances in neural information processing systems* (pp. 1289–1296).

Suzuki, S., & Cavanagh, P. (1998). A shape-contrast effect for briefly presented stimuli. *Journal of Experimental Psychology: Human Perception and Performance*, *24*(5), 1315.

Thompson, P., & Burr, D. (2009). Visual aftereffects. *Current Biology*, *19*(1), R11–R14.

Uznadze, D. N. (1958). Experimental basis of the psychology of set. *Experimental Studies on the Psychology of the Set*, 5–79.

Uznadze, D. N. (1966). The psychology of set.

Webster, M. A., & MacLeod, D. I. (2011). Visual adaptation and face perception. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *366*(1571), 1702–1725.

Wei, X.-X., & Stocker, A. A. (2015). A bayesian observer model constrained by efficient coding can explain'anti-bayesian'percepts. *Nature neuroscience*, *18*(10), 1509.

Wohlwill, J. F. (1960). Developmental studies of perception. *Psychological Bulletin*, *57*(4), 249.

Modeling socioeconomic effects on the development of brain and behavior

Selma Dündar-Coecke (selma.coecke@gmail.com)
Psychology and Human Development, Institute of Education, University College London, UK

Michael S. C. Thomas (m.thomas@bbk.ac.uk)
Developmental Neurocognition Lab, Birkbeck, University of London
Malet Street, London WC1E 7HX, UK

Abstract

We used a population-level connectionist model of cognitive development to unify a range of empirical findings on the influence of socioeconomic status (SES) on behavior and brain development. The model captured qualitative patterns of *development* in behavior and brain structure, including reductions in connectivity across development (gray matter, cortical thickness) as behavioral accuracy increases. *Individual differences* in SES were implemented by altering the level of stimulation available in the environment. At the brain level, the model simulated non-linear effects of SES on cortical surface area (Noble et al., 2015), and faster cortical thinning across development in children from lower SES backgrounds (Piccolo et al., 2016). At the behavioral level, the model simulated the effect of SES on IQ, whereby gaps are observed to widen across development (von Stumm & Plomin, 2015). The model's main shortcoming was insufficient growth in connection magnitude across development in lower SES groups, implying that some aspects of the growth of connection strengths may be maturational (e.g., myelination) rather than experience dependent.

Keywords: socioeconomic status, brain, behavior, connectionist networks, multi-scale models, population modeling

Introduction

Differences in socioeconomic status (SES) have marked effects on cognitive development (Farah et al., 2006). These effects are not uniform across all areas of cognition and are stronger in the development of language and cognitive control (executive functions), where lower scores are observed in children from lower SES families. SES effects have been observed on intelligence (IQ) and indeed, it has been reported that gaps between children widen across development (von Stumm & Plomin, 2015; see Figure 1). SES refers to a marker for multiple potential causal pathways acting on cognitive development, among them effects on prenatal brain development, post-natal nurturing, and post-natal cognitive stimulation (Farah, 2017; Hackman, Farah & Meaney, 2017).

Recent work in neuroscience has focused on the impact of SES on measures of brain structure, demonstrating that cortical surface area and cortical thickness in children and adolescents show small but reliable associations with differences in family income and parental education; in some cases, associations have been observed between SES and the size of particular brain structures, such as the hippocampus and amygdala (e.g., Noble et al., 2015). Although small in size, these effects can be non-linear: for example, while lower SES is linked with reduced cortical surface area, the impact is larger for the lowest SES groups (Figure 2). Moreover, effects on brain structure are strongest in areas linked with language (temporal) and executive functions (prefrontal); and measures of cortical surface area (but not thickness) have been shown to mediate the relationship between SES and behavior (Noble et al., 2015). SES can be seen to influence the *rate of change* of brain structure over development. The cortex usually thins from mid-childhood onwards. In children from low SES backgrounds, thinning was observed to be faster. Piccolo et al. (2016) found that while cortical thickness showed no main effect of SES, it thinned more quickly in lower SES children; conversely, cortical surface area was reduced in the lower SES children, but showed similar rates of change across SES groups. Neuroscience data, then, confirm the impact of SES, but do they point to the causal pathways by which it operates?

Two challenges present themselves. First, we need a mechanistic account to explain how environmental influences produce linked effects on brain and behavior, which would provide a basis to evaluate competing accounts about causal pathways. Second, any putative causal explanation of SES effects must accommodate a range of other empirical phenomena: on developmental changes in brain structure, on the relationship between cognitive ability and various measures of brain structure, and on the origin of individual differences. The main qualitative patterns that must be captured are as follows.

First, although behavioral accuracy typically increases across development, this is not the case for all measures of brain structure: some measures increase (white matter volume, cortical surface area) but others decrease following a peak in early or mid childhood (gray matter volume, cortical thickness) (e.g., Giedd et al., 1999; Sowell et al., 2004). The mechanisms that drive these changes are still debated, but include myelination and pruning of local connectivity (synapses, dendrites, axons), but not generation or loss of neurons.

Second, although environmental measures such as SES predict individual differences, a large proportion of variance in cognitive ability, brain structure, and change in brain structure across development is predicted by the genetic similarity between people – that is, these phenotypes are

highly heritable (Plomin et al., 2013). Heritability may be modulated by SES: it has been observed that in individuals from low SES backgrounds, the heritability of IQ can be reduced (e.g., Tucker-Drob & Bates, 2016).

Third, brain structure is correlated with intellectual ability, with one meta-analysis showing correlations of 0.1-0.3 between brain volume and IQ (McDaniel, 2005). Ritchie et al. (2015) found that brain volume explained 12% of the variance in general cognitive ability, cortical thickness another 5%, and all structural measures together up to 21% of the variance. These individual differences data imply that having more neural resources is better for cognition. IQ is also related to the rate of thinning of the cortex with age (Shaw et al., 2006). Higher IQ is associated with faster thickening of cortex across early childhood, and then faster thinning of cortex from mid-childhood onwards. Since cognition improves as gray matter reduces, the developmental data imply that having fewer neural resources is better for cognition. This inconsistency is rendered more puzzling by the observation that faster thinning of the cortex is linked with lower SES (Piccolo et al., 2016). Lower SES is associated with lower IQ (von Stumm & Plomin, 2015). How can higher IQ and lower SES both be linked to faster thinning of cortex, when higher IQ is itself associated with higher SES? This complex set of effects is summarized in Table 1.

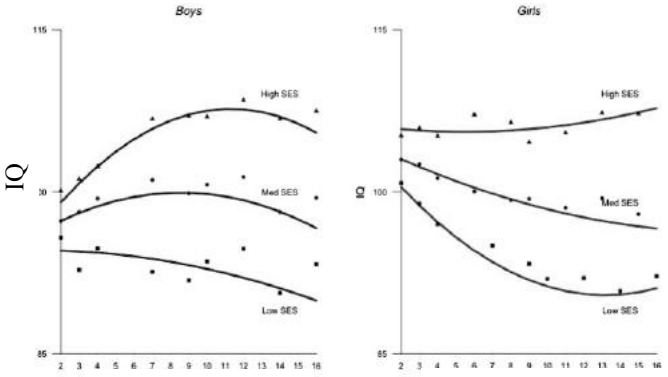


Figure 1: SES gaps in intelligence widen across development (von Stumm & Plomin, 2015)

In the current work, we use a multi-scale model to try and unify this complex pattern of data. The model is based on an artificial neural network (ANN) trained with backpropagation. In a multi-scale model, constrains are included at several levels of description (Thomas, Forrester & Ronald, 2016). Crucially, because the data concern both development and individual differences, it is necessary to simulate a population of individuals, and to model the influences on development that produce individual differences. Because the data span behavior, brain, SES, and genetics, the model must have analogues of each of these in its design.

In connectionist models of cognitive development, abstract principles of neurocomputation are embodied in systems whose activation states correspond to concepts and whose inputs and outputs can be linked to behavior (see, e.g., Thomas & McClelland, 2008). Thomas (2016) argued

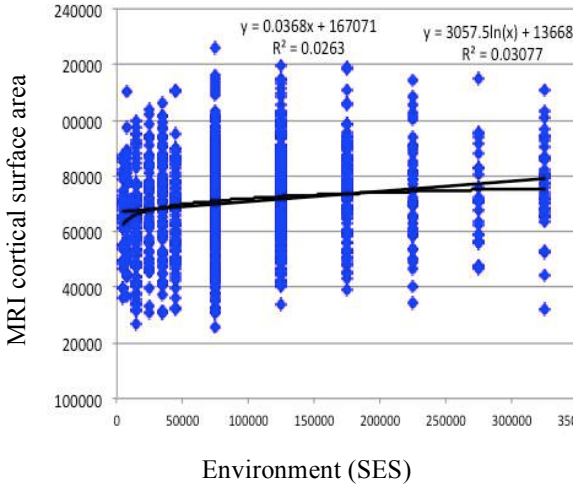


Figure 2: The link between cortical surface area and family income (data re-plotted from Noble et al., 2015).

Table 1: List of empirical phenomena to be simulated

1.	Behavioral accuracy increases across development
2.	Some measures of brain structure increase across development (white matter, cortical surface area)
3.	Some measures of brain structure reduce across development (gray matter, cortical thickness)
4.	Lower SES is associated with lower IQ and gaps widen across development
5.	Lower SES is associated with reduced cortical surface area, with larger effects at lowest SES levels
6.	Lower SES is associated with faster thinning of the cortex over development
7.	Lower SES is associated with reduced cortical surface area but no modulation of rate of development
8.	Cortical surface area partially mediates the relationship between SES and behavior
9.	Individual differences in behavior and brain structure are highly heritable
10.	Low SES can reduce the heritability of IQ
11.	Brain volume correlates with IQ
12.	Across development, higher IQ is associated with faster thickening and then faster thinning of the cortex

that with a simple addition – the onset of pruning of unused connectivity after a certain point in training – these models could give plausible analogues to measures of brain structure, where the total number of connections would serve as an analogue to properties that decrease over development (gray matter, cortical thickness) – under the view that unused connections are pruned away, causing a

loss of volume; and the combined magnitude of connection weights (excitatory and inhibitory) would serve as an analogue of properties showing increases (white matter, cortical surface area) – under the view that retained connections are optimized through myelination, causing an increase in volume. We use the same scheme here.

To capture genetic influences on behavior and structure, each network must have a genome and genomes must vary between individuals. To the extent that cognition is seen as information processing in the brain, genetic effects must translate to influences on neurocomputational properties. Accordingly, Thomas et al. (2016) used a method to simulate individual differences where the neurocomputational parameters of an ANN (e.g., number of hidden units, learning rate) were encoded in an artificial genome. Genetic variation produced parameter variation. In behavior genetics, the heritability of a phenotype such as behavior or brain structure is usually assessed using the twin design, where more heritable phenotypes show greater similarity between monozygotic (MZ) twins than dizygotic (DZ) twins. MZ twin networks can be simulated by networks with the same genome (and therefore, parameters), while DZ twins can be simulated by networks that share on average 50% of the gene variants in their genomes (see Thomas et al., 2016, for further details). Heritability of behavior and brain structure can then be simulated by comparing the respective correlations between MZ networks versus DZ networks.

SES can plausibly be implemented in several ways (Thomas, Forrester & Ronald, 2013): it might influence how a network is constructed (equivalent to prenatal effects on brain development); it might influence the information on which the network is trained (equivalent to differences in levels of cognitive stimulation during post-natal development); or it might influence both factors. In the following simulations, we evaluated a model that implemented SES as differences in the richness of the training set.

An ANN trained with backpropagation has very limited biological plausibility. We should therefore be clear what are our key assumptions in relating measures of network structure to measures of brain structure. They are as follows: (1) neuron number is fixed so that changes in structure reflect changes in connectivity; (2) structural measures that increase over development (cortical surface area, white matter) reflect increases in connection strength, while structural measures that decrease over development (cortical thickness, grey matter) reflect reductions in connection number; (3) connection strength increases can only be experience dependent; (4) connection strength decreases can be experience dependent (training reduces some connections), intrinsic (weight decay), or both (an intrinsic pruning process operates depending on connection strengths which in turn are influenced by experience); (5) connection number is intrinsic (growth) or an interaction with experience (pruning); (6) we did not include an assumption that connection growth might be partly experience / environment dependent, nor that there might be intrinsic contributions to connection strengthening (e.g., myelination occurring through maturation).

The adequacy of the model in capturing the patterns of empirical data will serve as a test of these assumptions.

Method

The following simulations use a base model taken from the field of language development, addressed to the domain of English past-tense formation. Here, the model was employed in an illustrative setting, intended only as an example of a developmental system applied to the problem of extracting the latent structure of a cognitive domain through exposure to a variable training environment. The intention was to capture qualitative characteristics of the empirical data rather than to exactly calibrate variances from genetic and environmental sources to fit empirically observed estimates of heritability in certain populations. In that capacity, the past tense accuracy of the networks was taken as a metric of behavioral development, and of intelligent behavior more widely (that is, of the type measured by cognitive ability tests). However, the base model has been used to specifically simulate data on the influence of SES on children's past-tense acquisition (Thomas et al., 2013). Full details of the current simulation can be found in Thomas (2016).

Network architecture: The basic model was a 3-layer backpropagation network, with 57 input and 62 outputs. The process of network growth was not modeled, only the outcome of this process. The number and size of initial connections was influenced by several factors, including number of weight layers, sparseness of connectivity, and range of initial random variation. Connection pruning occurred after a specified training epoch, and removed any connections below a specified threshold with a specified probability. Each of these three parameters was free to vary between individuals. Pruning onset varied between 0 epochs and 1000 epochs, where 1000 epochs was full lifetime (median value 100 epochs); pruning threshold varied between a magnitude of 0.1 and 1.5 (median 0.5); pruning probability varied between 0 and 1 (median 0.05) per pattern presentation. Overall, fourteen neurocomputational parameters were free to vary between individuals. These were: the architecture (fully connected or three-layer), number of hidden units, sparseness of connectivity, sigmoid activation function temperature, activation noise added to unit net inputs, nearest neighbor output threshold, learning rate, backpropagation error measure (root mean square or cross entropy), momentum, initial weight variance, weight decay, pruning onset epoch, pruning threshold, and pruning probability (see Thomas, 2016, for parameter specifications, and range of values, for the GWEW condition).

Training set: The training set comprised 508 artificial monosyllabic verbs, constructed using consonant–vowel templates and the phoneme set of English. Phonemes were represented over 19 binary articulatory features. The verbs conformed to the past-tense patterns observed in English,

with 410 regular verbs (forming the past tense via the *+ed* rule) and 98 irregular verbs of three types, no-change, vowel-change, and arbitrary (see Thomas et al., 2016, for more details). Training used pattern presentation in random order without replacement.

Implementation of SES differences: Each simulated child was raised in a family with a given level of language stimulation, taken to be correlated to the family's SES (Hart & Risley, 1995). A family quotient parameter was sampled uniformly between the range 0 and 1. This proportion was applied as a one-time filter on the full training set. A network raised in a family with a family quotient of 0.75 would be exposed to a training set with around 75% of the training patterns. With a range between 0 and 1, networks could in principle be exposed to very few training patterns (see Thomas, 2016, for discussion).

Implementation of genetic differences: Differences in learning ability arose from the net effect of small variations in all the neurocomputational parameters, under a polygenic model of intelligence (Thomas, 2018). For this simulation, all variation in these parameters was considered to be under genetic control. There was a random association of family quotient to genotype, that is, we did not simulate gene-environment correlations.

Simulation design: A population of 1000 networks was created in sets of pairs, either MZ or DZ twins. Each network was trained for 1000 epochs. Performance on the training set (regular and irregular verbs) and two network measures, total number of connections and magnitude of connections, were assessed across training.

Results

Developmental changes in behavior: Figure 3 shows the monotonic improvement in accuracy in regular and irregular (vowel-change) verbs across training, averaged across the whole population (Table 1, #1).

Developmental changes in brain structure: Figure 4 plots the change in the magnitude of connections (gradually increasing) and the total number of connections (a non-linear decline) across training, averaged across the whole population. The plot captures the increase and decrease of different structural measures (Table 1, #2 and #3).

SES effects on behavior: The behavioral scores of the networks were split by their SES (upper quartile, family quotients >.75, lower quartile family quotients <.25). At each measurement point, the population distribution in accuracy values was used to convert accuracy to IQ scores, by deriving the population mean and standard deviation and transforming these to a mean of 100 and standard deviation of 15. Figure 5 plots developmental trajectories of IQs split by upper, lower and middle two quartiles. The plot captures a widening gap between the groups (Table 1, #4). In the simulation, this is the result of non-linear developmental trajectories, whereby the lower SES groups show earlier plateauing of performance.

SES effects on brain development: Figure 6 shows a scatter plot of each network's connection magnitude against SES (family quotient value), after 100 epochs of training. The simulations demonstrate a reliable association of SES to network structure. The pattern of a small effect size and non-linear relationship capture that shown in Noble et al.'s (2015) cortical surface area data, with larger reductions in area at the lowest SES levels (Table 1, #5).

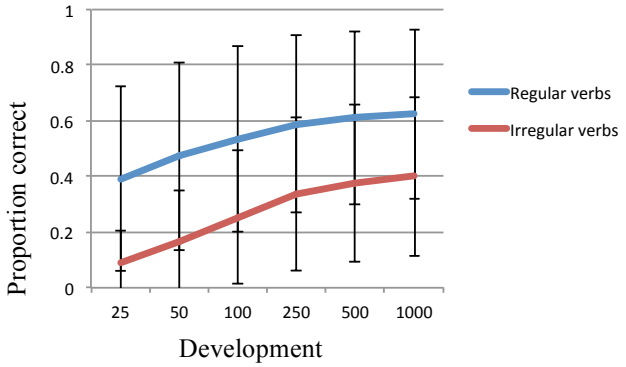


Figure 3: Average population development for two behaviors, regular verb and irregular verb performance. (Error bars show standard deviations)

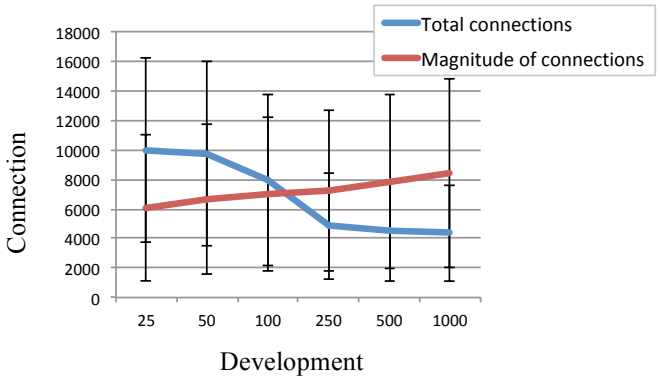


Figure 4: Average population changes in connection magnitude and number over development. (Error bars show standard deviations)

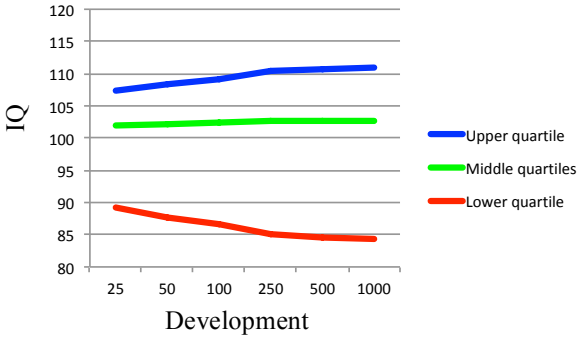


Figure 5: Development of behavior split by SES quartile.

Figure 7 separates the networks into the upper quartile and lower quartile according to SES and plots change in total number of connections across development (for simplicity, linking two points in training, epochs 25 and

250). There was no reliable main effect of SES on connection total (p=.547), but a reliable interaction, whereby connection total reduced more quickly in the lower SES quartile (F(1,498)=15.42, p<.001, η_p^2=.030). This occurred because lower SES networks received less stimulation, causing less strengthening of connections, and in turn greater vulnerability to later pruning processes. The result captures Piccolo et al.'s (2016) observation that cortex thins more quickly in children from a lower SES background, without overall differences in cortical thickness between groups (Table 1, #6).

Figure 8 plots the equivalent simulation data for connection magnitude, split by SES quartile. The model shows a main effect of SES, with smaller magnitudes in low SES networks (F(1,498)=13.33, p<.001, η_p^2=.026), but also an interaction, where magnitude in low SES networks improves much more slowly ((F(1,498)=150.88, p<.001, η_p^2=.233). The first effect captures the smaller cortical surface area observed by Piccolo et al. (2016) for lower SES children, but the interaction does not accord with the empirical data – SES does not modify rate of change of cortical surface area (Table 1, #7, not captured).

Brain structure mediates relationship of SES to behavior: Noble et al. (2015) found that cortical surface area mediated the relationship between SES and behavior but thickness did not. In the model, we observed increasing correlations between SES, connection magnitude, and behavior across training, such that a mediation effect was detectable by the end of training. Figure 9 shows that connection magnitude mediated associations between SES and regular verb performance (β=0.05, t(998)=8.44, p<.001, CI [.04; .07]). The Sobel test was significant, confirming partial mediation (Sobel-z=7.98, p<.001). Per Noble et al.'s findings, the analogue of thickness, connection number, did not show the mediation effect. This is because in the model, the correlation of SES to connection number did not reach significance (Table 1, #8).

Heritability of individual differences: at 100 epochs, the correlations between twin pairs were as follows: Regular verb performance: MZ=.99, DZ=.61; irregular verb (vowel change) performance: MZ=.97, DZ=.49; connection magnitude: MZ=1.00, DZ=.44; connection total: MZ=1.00, DZ=.33. Where MZ correlations are higher than DZ correlations, this implies genetic influence. The difference between the correlations can be used to estimate the heritability of the phenotype. Under an additive model, the respective heritabilities are .76, .97, 1.12, and 1.34 (that the latter values exceed 1 shows that the genetic effects violate an additive model and there are dominance effects operating). These values are higher than observed for behavior and measures of brain structure (Plomin et al., 2013). The simulations included no measurement error, which would appear as an environmental effect unique to each individual. Nevertheless, these high estimates of heritability imply the assumption that all neurocomputational parameter variation is under genetic control is not plausible, and that the environment contributes to variation in parameters (perhaps during prenatal brain development). However, the observed high heritabilities meant that effects of SES on brain and behavior were successfully simulated against a background of strong genetic influence on both measures (Table 1, #9).

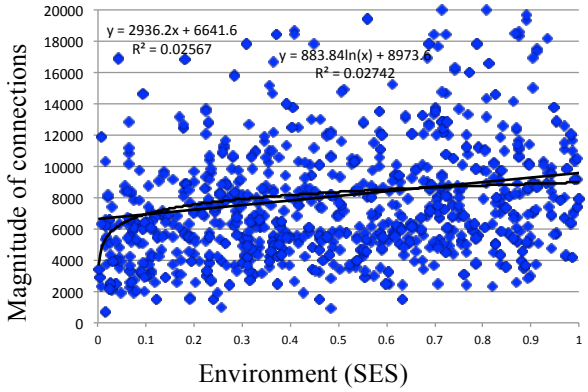


Figure 6: Connection magnitude versus SES

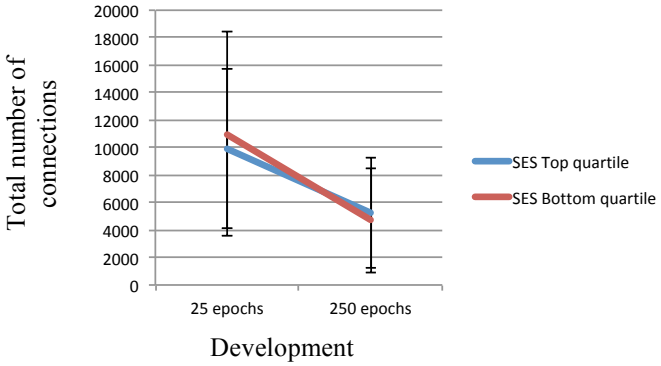


Figure 7: Change in number of connections across development, split by SES. (Error bars = STD)

Estimates of heritability were also observed to differ between upper and lower SES quartiles, with the lower SES quartile showing reduced heritability as the impoverished training set – rather than the neurocomputational parameters – became the limiting factor on performance. For example, for irregular verbs at 100 epochs, the upper quartile showed MZ correlation of .97, DZ .35, while in the lower quartile, these values were .95 and .60. The reduced gap between MZ and DZ correlations shows reduced genetic influence in the low SES group (Table 1 #10).

Relation of intelligence to brain structure: The 'ability' of each network was assessed based on its behavior. We chose to assess this based on irregular (vowel-change) verb performance at an early point in development (50 epochs), which gave good sensitivity to discriminate between individuals. At 100 epochs, the correlation of ability with total connections was .352, and with magnitude was .371. This captures the empirical observation of the small correlation between brain size and intelligence (Table 1, #11).

Based on the ability measure, we derived upper quartiles (top 25%) and lower quartiles (bottom 25%) of ability. Figure 10 shows the change in total number of connections between two points in development, epoch 25 and epoch 250. At epoch 25, high ability networks had reliably more connections ($t(458)=8.74$, $p<.001$, Cohen's $d=.81$). We did not simulate the growth of connectivity, only the outcome of this process. The higher peak captures the outcome of putative faster thickening of cortex across development for children with higher IQs (Shaw et al., 2006). Across development, connection number fell reliably more quickly in high ability networks than low ability networks ($F(1,458)=31.60$, $p<.001$, $\eta_p^2=.065$). The faster fall is a side effect of the higher peak – the greater ability arises from the greater computational power of having a larger network, while larger networks experience faster pruning. The result captures the observation by Shaw et al. (2006) that cortex thins more quickly in children with higher IQ (Table 1, #12).

Discussion

The model was successful in qualitatively capturing 11 of 12 target phenomena linking SES, IQ, brain development and behavioral development. The model used simple error-driven backpropagation networks, where connection strengths are altered to improve performance. Links to brain structure were established by adding a pruning process that, after a certain point early in development, removes unused connections. Measures of network connectivity gave analogous fits to brain structure measures that either show increases with age (white matter, cortical surface area) or decreases (gray matter, cortical thickness). The match of simulation and empirical data supports the view that these brain measures represent the results of experience-dependent strengthening of connectivity combined with intrinsic processes for connectivity growth and loss, where connectivity loss is dependent on the extent to which previous experience has strengthened connections.

The successful simulation of SES patterns in behavior and brain support the view that a key element of these effects is the level of cognitive stimulation. However, this is unlikely to be the full effect, and other environmental influences on prenatal and postnatal development undoubtedly contribute (see, e.g., Betancourt et al., 2016, for SES-related gray matter differences observed in babies at 1 month of age, where experience-dependent effects have had little time to act). Extension of the model presented here is necessary to explore the possibility that environmental effects on brain growth may interact with, and indeed may be correlated with, differences in cognitive stimulation.

The model failed in two regards. First, it did not capture the observed absence of SES effects in the rate of change of cortical surface area (Piccolo et al., 2016). The model did not show enough strengthening of connectivity across development in the low SES group. This implies that one of the assumptions of the model – that connection strength increases can only be experience dependent – is incorrect, and that there is a maturational contribution to connectivity increases (such as myelination). Second, its estimates of heritability were too high for individual differences in behavior and brain. In part, this is due to the absence of measure error in the simulations. But, consistent with above comments, it also demonstrates another initial assumption of the model is incorrect, that neurocomputational parameters are solely under genetic influence.

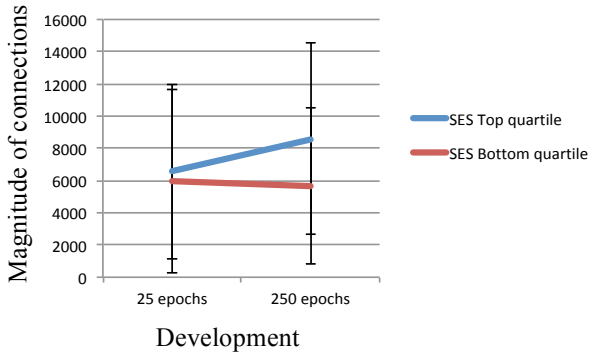


Figure 8: Change in magnitude of connections across development, split by SES. (Error bars = STD)

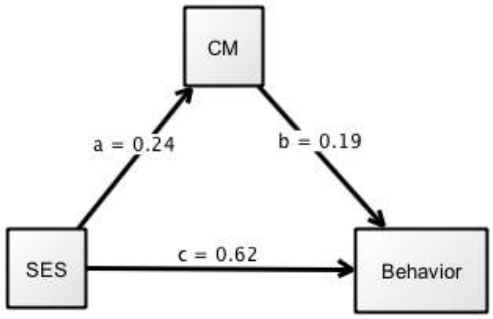


Figure 9: Partial mediation between connection magnitude (CM), SES and behavior (regular verb accuracy)

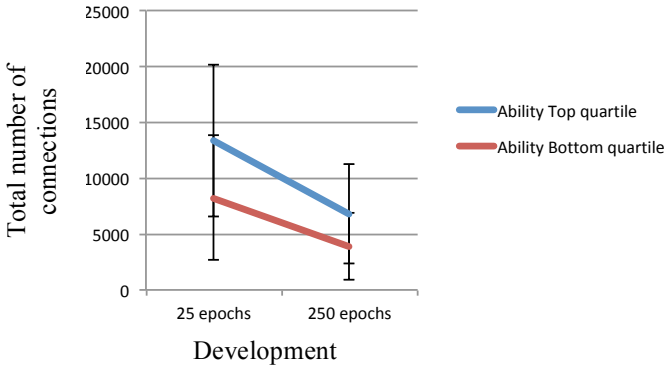


Figure 10: Change in number of connections across development, split by behavioral ability. (Error bars = STD)

A future extension of the model should investigate environmental influences on specifying network parameters, such as initial network growth. It should also be noted that the model set out only to simulate qualitative patterns, not to calibrate against precise ranges of genetic or environmental

variation, or to capture particular population mean levels of behavior at a given point in development. Some assumptions could be questioned, such as the extreme deprivation implied by training sets that could vary down to including no patterns.

Implementation of a mechanistic model provides the benefit that it can reconcile apparent paradoxes in the empirical literature. Why are high IQs associated with having a bigger brain (as if more neural resources were better for cognition) but also associated with faster gray matter loss and cortical thinning (as if fewer resources were better)? The answer is that the network size is driving ability (so more is always better), but that a higher peak of network size is then associated with faster connectivity loss during pruning of unused resources (in the manner that higher mountain peaks have steeper sides). How can faster cortical thinning be simultaneously associated with higher IQ and lower SES (which is associated with *lower* IQ)? The answer is that in the higher ability networks, there are more spare resources to be lost during pruning so thinning is faster; in low SES networks, the small training set (equivalent to lower cognitive stimulation) produces less strengthening of connectivity so that connections are more vulnerable to loss when pruning starts, leading once more to faster thinning. In other words, rate of change of structure isn't a direct marker of ability; ability is delivered by the full computational properties of the network and its developmental origins, not proxy measures like cortical thickness.

The model presented here is highly simplified, employing a single artificial network with very restricted biological plausibility. The range of the phenomena that the model captures probably reflects the fact that the existing observations we have on behavior, brain structure, and SES give limited insight into the detailed neural processes underlying behavior, development, and environmental influences. Nevertheless, we argue here for the importance of building multi-scale models that integrate individual differences within a developmental framework, and which can therefore evaluate causal mechanisms linking SES, brain and behavior. With causal, mechanistic accounts in hand, we are better able to consider interventions to ameliorate the impact of poverty and deprivation on children's development. The results here point to the importance of cognitive stimulation, and encourage interventions that seek to enrich that stimulation for children from poor backgrounds.

References

Betancourt, L.M., Avants, B., Farah, M.J., Brodsky, N.L., … & Hurt, H. (2016). Effect of socioeconomic status disparity on neural development in female African-American infants at age 1 month. *Developmental Science*, *19(6)*, 947-956.

Farah, M. J. (2017). The neuroscience of socioeconomic status: Correlates, causes, and consequences. *Neuron*, *96*, September 27, 2017, 56-71.

Giedd, J.N., Blumenthal, J., Jeffries, N.O., Castellanos, F.X., … & Rapoport, J.L. (1999). Brain development during childhood and adolescence: a longitudinal MRI study. *Nature Neuroscience*, 2, 861-863

Hackman, D. A., Farah, M. J. & Meaney, M. J. (2010). Socioeconomic status and the brain. *Nature Reviews Neuroscience*, *11*, 651– 659.

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children*. Baltimore, MD: Paul H. Brookes.

McDaniel, M. A. (2005). Big-brained people are smarter: A meta-analysis of the relationship between in vivo brain volume and intelligence. *Intelligence*, *33 (4)*, 337-346.

Noble, K.G., Houston, S.M., Brito, N.H., Bartsch, H., ... & Sowell, E.R. (2015). Family income, parental education and brain structure in children and adolescents. *Nature Neuroscience*, 18(5), 773-778.

Plomin R., DeFries J.C., Knopik V.S., & Neiderhiser J.M. (2013). *Behavioral genetics*. 6th Ed. Worth Publishers.

Piccolo, L.R., Merz, E.C., He, X., Sowell, E.R., & Noble, K.G. (2016). Age-related differences in cortical thickness vary by socioeconomic status. *PLoS ONE,* 11(9), e0162511.

Ritchie, S. J., et al. (2015). Beyond a bigger brain: Multivariate structural brain imaging and intelligence. *Intelligence*, *51*, 47-56.

Shaw, P., Greenstein, D., Lerch, J., Clasen, L., … & Giedd, J. (2006). Intellectual ability and cortical development in children and adolescents. *Nature*, Vol. 440, 30 March 2006,

Sowell, E.R., Thompson, P.M., Leonard, C.M., Welcome, S.E., … & Toga, A.W. (2004). Longitudinal mapping of cortical thickness and brain growth in normal children. *Journal of Neuroscience*, 24(38), 8223–31.

Thomas, M. S. C. (2018). A neurocomputational model of developmental trajectories of gifted children under a polygenic model: When are gifted children held back by poor environments? *Intelligence*, *69*, 200-212.

Thomas, M. S. C., Forrester, N. A., & Ronald, A. (2016). Multi-scale modeling of gene-behavior associations in an artificial neural network model of cognitive development. *Cognitive Science*, *40(1)*, 51-99.

Thomas, M.S.C. (2016). Do more intelligent brains retain heightened plasticity for longer in development? A computational investigation. *Developmental Cognitive Neuroscience*, 19, 258-269.

Thomas, M.S.C., Forrester, N.A. & Ronald, A. (2013). Modeling socio-economic status effects on language development. *Developmental Psychology*, *49(12)*, 2325-2343.

Thomas, M. S. C., & McClelland, J. L. (2008). Connectionist models of cognition. In R. Sun (Ed.), *Cambridge handbook of computational cognitive modeling* (pp. 23-58). Cambridge: CUP.

Tucker-Drob, E. M. & Bates, T. C. (2016). Large Cross-National Differences in Gene × Socioeconomic Status Interaction on Intelligence. *Psychological Science*, *27(2)*, 138–149.

von Stumm, S. & Plomin, R. (2015). Socioeconomic status and the growth of intelligence from infancy through adolescence. *Intelligence*, *48*, 30–36.

Working memory for object concepts relies on both linguistic and simulation information

Agata Dymarska (a.dymarska@lancaster.ac.uk)
Louise Connell (l.connell@lancaster.ac.uk)
Briony Banks (b.banks@lancaster.ac.uk
Department of Psychology, Fylde College,
Lancaster University, Bailrigg, Lancaster, LA1 4YF, UK

Abstract

The linguistic-simulation approach to cognition predicts that language can enable more efficient conceptual processing than sensorimotor-affective simulations of concepts. We proposed that this has implications for working memory, whereby use of linguistic labels enables more efficient representation of concepts in a limited-capacity store than representation via full sensorimotor simulation. In two pre-registered experiments, we asked participants to remember sequences of real-world objects, and used articulatory suppression to selectively block access to linguistic information, which we predicted would impair accuracy and latency of performance in an object memory recognition task. We found that blocking access to language at encoding impaired memory performance, but blocking access at retrieval unexpectedly facilitated speed of responding. These results suggest that working memory for object concepts normally relies on language but people can flexibly adapt their memory strategies when language is unavailable. Moreover, our data suggest that a sequence of up to 10 object concepts can be held in working memory when relying on sensorimotor information alone, but this capacity increases when linguistic labels are available.

Keywords: working memory; concepts; linguistic information; simulation; embodied cognition

Introduction

Although traditionally conceptual representations were considered amodal and removed from perceptual experience (Tranel, Damasio & Damasio, 1997), more recent evidence suggest that concepts are grounded in sensorimotor and linguistic experience (Barsalou et al., 2008; Connell & Lynott, 2014; Vigliocco et al., 2009). Simulated representations engage the neural subsystems involved in sensorimotor, affective, introspective, and other situated experiences of a concept. For example, the concept *"dog"* includes its visual shape and colour, the action and feel of patting its fur, the sound of its bark, walking it on a leash, and the positive feelings towards a pet. The neural activation patterns involved in processing these experiences can later be partially re-activated (i.e., *simulated*) to represent a concept. Linguistic representations, on the other hand, comprise word (and phrase) labels associated with these sensorimotor-affective simulations, and the distributional patterns between them (statistical co-occurrences of words in language). For instance, seeing a terrier or hearing a bark will activate the label "dog", and words that frequently appear in similar contexts, like "tail" or "leash". These two components are interrelated and mutually supportive, and recent theories argue that both are intrinsic to conceptual representation (Connell & Lynott, 2014; Louwerse, 2011). That is, linguistic labels are part of concepts and conceptual processing uses simulation and linguistic information to varying extents depending on task demands, available resources, and other factors (Connell, 2018; Connell & Lynott, 2014).

The role of simulation and linguistic components in cognition is illustrated by a range of empirical evidence. Neuroimaging research has shown that processing of action words (e.g. "pick", "kick") activates body part-specific motor areas (Hauk, Johnsrude, & Pulvermüller, 2004). Critically, processing of such words is selectively impaired in patients with neurodegeneration of the motor system – Parkinson's disease (Boulenger et al., 2008). Behavioural experiments also show evidence for use of simulations: for example, people were faster to recognise a horizontally-oriented nail after reading "He pounded the nail into the wall" than "He pounded the nail into the floor" (Stanfield & Zwaan, 2001). Participants were also quicker to make a size judgment of manipulable objects than when the objects were too big to be physically manipulated (Connell, Lynott, & Dreyer, 2012). As for the linguistic component, information from language alone is powerful enough to inform responses across diverse conceptual tasks. Evidence comes from a range of paradigms, including property verification and generation (Louwerse & Connell, 2011; Santos et al., 2011), spatial iconicity judgements (Louwerse & Jeuniaux, 2010) and spatial cuing of attention (Goodhew, McGaw, & Kidd, 2014). Frequency of words co-occurring in the same context can predict how easily they are understood as a novel conceptual combination (Connell & Lynott, 2013). These findings show that both sensorimotor and linguistic information is functionally important to conceptual processing.

Much evidence for the linguistic component centres on the usefulness of distributional information (i.e., co-occurrence relationships between words/phrases) in cognition. However, that is not its full role. Language is a unique human characteristic which allows us to communicate something in the past, future, or hypothetical existence (Barsalou, 2005), and allows us to concisely name a complex multimodal experience. The idea that language is beneficial for our cognitive processing has been around for a while (e.g.: Paivio, 1971), but recent theories have developed the role of linguistic labels in a number of new directions (e.g., Borghi et al., 2018; Connell, 2018; Lupyan, 2012). Most relevant to our present purposes, Connell and Lynott (2014) propose that having labels for concepts enables a process of *linguistic bootstrapping*, whereby words and phrases act as *linguistic placeholders* in an ongoing representation when there are insufficient resources to maintain a sensorimotor simulation in full. These linguistic placeholders can later be fleshed out into a simulation again at any time if

resources become available. To date, the linguistic bootstrapping hypothesis has remained theoretical and has not been tested directly but there is indirect support for the idea in the wider literature. Working memory (WM) is necessarily limited in capacity – there are only so many concepts that can be maintained and manipulated at once – and recent evidence does suggest that linguistic information is more economical in representation (i.e., occupies less "space" in working memory) than sensory information (Langerock, Vergauwe, & Barrouillet, 2014). Further, explicitly labelling simple visual stimuli seems to increase memory capacity (Zormpa et al., 2018). It is possible that when working memory capacity is strained to its limit, as when trying to maintain a representation of numerous concepts, a linguistic label could deputise for its referent sensorimotor information (e.g., word "dog" replaces simulation of *dog*) to free up space.

It is currently unknown how many concepts (i.e., representations of real-world objects, events, and situations, such as *dog*, *running*, or *holiday*) can be maintained in working memory at once. Research on memory from the linguistic-simulation perspective concentrated on the role of sensorimotor simulation in memory (Dutriaux, Dahiez, & Gyselinck, 2018; Vermeulen et al., 2013) rather than the interplay of simulated and linguistic information in capacity limits. Working memory research has established a central capacity limit of 4 items (Cowan, 2010), but research informing this has used simple, artificial stimuli (e.g., feature conjunctions such as *red triangle*; random word pairs such as *desk-ball*). Such stimuli do not generalise to naturalistic, real-world concepts that comprise rich sensorimotor and linguistic information from long-term memory, and that are typically represented in broader situated simulations where concepts to reinforce and cue one another (e.g., a *dog* that is *running* with a *ball*). Baddeley's (2000) episodic buffer, a finite-capacity buffer that allows information from long-term memory to be integrated and manipulated goes some way to address these issues. For instance, participants remember sequences of words better when they are presented in meaningful sentences (i.e., that exploit interconnections between words) than in unstructured lists, which Baddeley and colleagues attribute to long-term knowledge retrieved to support representations in the episodic buffer. Nonetheless, not much is known about the role of simulated and linguistic information in representing concepts in working memory.

The current study

Our aim was to examine the role of linguistic and simulation components in working memory for real-world object concepts. In two pre-registered experiments, we presented ecologically valid sequences of object pictures (e.g., ingredients for a novel recipe) in a nonverbal paradigm, and tested recognition memory by asking participants to select the previously-presented objects from arrays of distractors. Critically, participants performed articulatory suppression (repeating aloud "the") during item encoding and/or retrieval to block access to linguistic information. Articulatory suppression has been widely used in WM research (Baddeley, 1992), where it has been shown to interfere with verbal encoding but to have little effect on general processing in the central executive (e.g., De Rammelaere, Stuyven, & Vandierendonck, 2001; Jaroslawska et al., 2018). We used a no-suppression condition as a baseline instead of an alternative suppression task, such as spatial tapping or visual interference, because such tasks would have interfered with the sensorimotor representation of concepts and therefore could not provide a useful control in our experiment. Thus, we expected the articulatory suppression task to block participants' access to the linguistic component of their conceptual representations.

We hypothesised that storage of object concepts in working memory will normally rely on language (i.e., linguistic placeholders), and that speed and accuracy of performance will be impaired when access to language is blocked. We expected performance to be best with no articulatory suppression at either encoding or retrieval, when participants are free to use both linguistic and sensorimotor information. We expected performance to be worst with articulatory suppression at retrieval only, when participants can employ linguistic placeholders to encode more objects, but lose access to those objects at retrieval when access to linguistic information is blocked. We planned to estimate working memory capacity for sensorimotor representations of concepts by calculating the average number of objects correctly retrieved with articulatory suppression at both encoding and retrieval (i.e., when linguistic information was unavailable).

Experiment 1

In this first experiment (pre-registration, data, analysis code, and full results are available as supplemental materials at https://osf.io/acv3m/?view_only=c1799106289a4063abf2eaa490eae009), we presented six objects per sequence, based on estimates from Langerock et al. (2014) that only 2-3 complex representations can be maintained in the episodic buffer. Participants viewed images of objects in each sequence one at a time during the encoding stage, and then had to select an alternative image of each target object from an array of distractors during the retrieval stage. Articulatory suppression took place half the time at encoding (to block access to object labels and prevent the use of linguistic placeholders when storing objects in WM) and half the time at retrieval (to block access to linguistic placeholders stored in WM).

Method

Participants Thirty-two native speakers of English (27 females; mean age = 21.2 years, *SD* = 3.2 years) were recruited from Lancaster University, for which they received course credit or a payment of £3.50. One participant was replaced due to a procedural error during testing.

The sample size was determined using sequential hypothesis testing with Bayes Factors (Schönbrodt et al., 2017). We stopped at the minimum sample size $N = 32$ when the Step 3 models for accuracy and Response Time (RT) cleared the specified threshold of evidence $BF_{10} < 0.2$. (model details in Design & Analysis section, full statistics in the Results section).

Materials Test items comprised of 72 target objects, divided into 12 sequences, each designed to be an ecologically valid

order of objects that would be plausibly used in a real-world setting, such as ingredients used in the process of making a cake, or a set of tools used in order to hang a picture. Each sequence therefore consisted of 6 target objects for study during the encoding stage, and each target object was assigned five distractor items for display in an object array during the retrieval (testing) stage. Five distractor objects were selected from the same category (e.g., food items, clothing) of which three were chosen based on colour, shape or function of the target object. Target and distractor items could all be plausibly used in a particular sequence (e.g. recipe) so that the task maintained ecological validity, and it would not be obvious which item in the array was the correct one.

To ensure the order of objects within each sequence was ecologically valid, we asked 9 naïve participants (who did not take part in the experiment) to rank-order the items according to how they are used in a given context. We used mean rank per object to finalise each sequence. For example: in the scenario *Tools for hanging a picture on the wall*, participants decided on the following order: *spirit level*, *drill*, *screw plug*, *screw*, *screwdriver*, *picture frame.*

We sourced photographic images for all objects from license-free online resources and edited them to appear on a uniform transparent background. To ensure that participants were tested on memory for object concepts, and not perceptual matching of a specific image, we prepared two different images for each target object: one for study during encoding and one for display in the distractor array during retrieval (e.g., showing an object from a different perspective or in a different colour). Images were scaled to 840 pixels along the longest dimension for objects presented during the encoding stage, and 470 pixels along the longest dimension for objects (targets and distractors) presented in the object arrays during retrieval. This resulted in a total of 504 object images: 72 target objects presented at encoding, 72 target objects presented at retrieval, and 360 distractor objects presented at retrieval. Figure 1 shows sample stimuli.

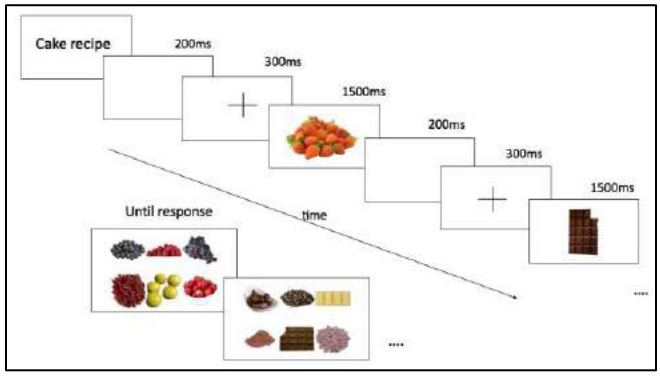


Figure 1: Diagram showing trial sequence and example stimuli at encoding and retrieval stages.

Procedure Participants were tested individually. After signing the consent form, the experimenter explained and demonstrated articulatory suppression, asking the participant to practice it. Once the participant confirmed that they understood and could perform articulatory suppression correctly, they sat in front of a computer, provided demographic information and proceeded to instructions. Participants were told they would see a sequence of everyday objects appear one by one onscreen, and their task was to remember the objects; later, they would see groups of objects onscreen and they should click on the object that belongs to the sequence they had been asked to remember. Participants then commenced a practice sequence without any articulatory suppression at encoding or retrieval. When the participant confirmed that they understood the task and were happy to continue, they were instructed on the articulatory suppression condition for both encoding and retrieval (i.e. when to start and stop) and commenced the experimental trials. Articulatory suppression was manipulated between participants at encoding and within participants at retrieval. The order of retrieval conditions was counterbalanced, and six sequences were presented in a randomised order within each condition. Experiment presentation was controlled by PsychoPy software (version 1.84.1; Peirce, 2009).

In the encoding stage, target objects were presented individually in a sequence (see Figure 1 for display times). Each sequence was preceded by a label (e.g. "cake recipe"). Once a full sequence of six target objects had been presented, participants saw a "wait" screen of 3 asterisks ("***") for 10 seconds before the retrieval stage began. In the articulatory suppression condition at encoding, participants repeated "the" aloud until this wait screen timed out. In the retrieval stage, participants saw a 2×3 array of six objects (one target object and five distractors) in random locations within the array (see Figure 1). Response times were measured from the onset of the array display until the participant clicked on an object using the mouse. After the retrieval of all six target objects had been tested, a message on the screen asked participants to press space when they were ready to proceed to the next object sequence.

After completing encoding and retrieval of six sequences, participants were instructed to take a self-paced break. They were then asked to switch to the counterbalanced articulatory suppression condition at retrieval (articulatory suppression at encoding remained constant). Participants then completed encoding and retrieval for six further object sequences.

Design and Analysis We analysed accuracy (dummy coded: incorrect = 0, correct = 1) with a mixed-effects hierarchical logistic regression (binomial, logit link). Step 1 entered participants and items as crossed random effects, where items were defined as objects nested within sequences. Step 2 added encoding and retrieval as fixed effects (no articulatory suppression = 0, articulatory suppression = 1). Step 3 added the interaction of encoding and retrieval as a fixed effect. We ran Bayesian model comparisons between steps, with Bayes Factors (BF) calculated via Bayesian Information Criteria (Wagenmakers, 2007). Similarly, we analysed RT for correct responses in a mixed-effects linear regression with the same effects and model comparisons as above. All analyses were run in R software (lme4 package, R version 3.4.1, 2017).

Results and Discussion

No trials were excluded on the basis of accuracy results[1]. For analysis of correct RTs, one trial was excluded as a motor error (faster than 300ms), and 27 trials were removed as outliers more than 3 standard deviations from the individual participant's mean (total 0.015% data removed).

Accuracy Bayesian model comparison showed equivocal evidence for Step 2 over Step 1, BF_{10} = 1.58; the data *very* weakly favoured the model containing articulatory suppression as fixed effects at encoding and retrieval over a model containing only random effects. There was strong evidence at Step 3 *against* the effect of the encoding-retrieval interaction on accuracy, BF_{10} = 0.02: the data were 47 times more likely under the Step 2 model without the interaction than the Step 3 model with the interaction.

We used the coefficients in Step 3 model to estimate the marginal accuracy for each condition of encoding × retrieval articulatory suppression (see Table 1). As predicted, accuracy was best in the no-suppression/no-suppression condition (no articulatory suppression at encoding or retrieval), with participants correctly recognising 5.6 out of 6 objects per sequence on average. However, accuracy was worst in the suppression/suppression condition: object memory was least accurate when language access was blocked at both encoding and retrieval (5.0 objects per sequence).

Finally, in an exploratory analysis not specified in the pre-registration, we examined the individual coefficients in the most likely model of fixed effects (i.e., Step 2)[2]. Articulatory suppression at encoding had a negative effect on response accuracy, beta = -0.567, *SE* = 0.275, *z* = -2.06, *p* = .039, as did articulatory suppression at retrieval, beta = -0.436, *SE* = 0.121, *z* = -3.59, *p* < .001. That is, as we predicted, removing access to language impaired object memory accuracy. When access to labels was blocked at the point of *encoding* objects, people were 76% more likely to make an error when later asked to recognise the object. Independently, when access to labels was blocked at the point of *retrieving* objects, people were 55% more likely to make an error. However, the inconsistency between equivocal Bayesian model comparison and significant regression parameters for Step 2 suggests that these effects should be treated cautiously.

Response Times Model comparisons showed very strong evidence at Step 2 for the effects of articulatory suppression at encoding and retrieval, BF_{10} = 1808.04. However, there was strong evidence at Step 3 *against* the effect of the encoding-retrieval interaction on RT, BF_{10} = 0.03: the data were 33 times more likely under the Step 2 model without the interaction than the Step 3 model with the interaction.

We took the coefficients in the Step 3 model to estimate the marginal mean RT for each condition of encoding × retrieval articulatory suppression (see Table 1). Against our expectations, recognition of target objects was best (fastest) in the no-suppression/suppression condition (language available at encoding but not at retrieval), and worst (slowest) in the suppression/no-suppression condition (language available at retrieval but not at encoding). People had most difficulty recognising the objects when language was blocked at the point of encoding but was available at retrieval.

We report an exploratory analysis of the coefficients in the most likely model of fixed effects (i.e., Step 2). As expected, articulatory suppression at encoding increased RT, beta = 408.57, *SE* = 172.32, *t*(31.99) = 2.371, *p* = .024. However, articulatory suppression at retrieval unexpectedly *reduced* RT, beta = -219.57, *SE* = 43.85, *t*(1770.74) = -5.007, *p* < .001). Closer examination of RT and accuracy suggested that this pattern was due to a speed-accuracy trade-off rather than facilitation of memory. When participants were asked to perform articulatory suppression at retrieval, response times were faster, but this was accompanied by lower accuracy, relative to no-suppression conditions (see Table 1). We discuss possible reasons for this trade-off below.

Summary Overall, the results support the hypothesis that memory for object concepts normally relies on language. Blocking language access when encoding a real-world object sequence affected memory: speed and accuracy were both impaired relative to no suppression. This is consistent with the idea that object concept is stored in WM via sensorimotor simulation *and* its linguistic label, and memory is impaired when only sensorimotor simulation is available.

However, blocking access to language while retrieving objects had unexpected effects. Rather than straightforward impairment, there was a speed-accuracy trade-off (low RT and accuracy), suggesting that articulatory suppression at retrieval caused participants to adopt an alternative, heuristic strategy that led to fast but inaccurate responding. Thus, the hypothesis that memory performance would be worst in the no-suppression/suppression condition was not supported. This may be because participants knew, before they studied the object sequence, whether they would perform articulatory suppression at retrieval. Knowing that language would be unavailable during retrieval could have led participants to strategically rely on sensorimotor information even when they had language access at encoding. Another possibility is that performance was subject to ceiling effects. When language was not available, participants correctly recognised 5.0 items per sequence on average, indicating that they were able to represent five object concepts in working memory from sensorimotor simulation alone (more than the suggested episodic buffer capacity of 2-3 items, Langerock et al., 2014). Thus, working memory capacity may not have been under particular strain, and participants did not have to replace some of the sensorimotor information with linguistic placeholders to remember the full sequence. We examine these possibilities in the next experiment.

[1] Exclusion criteria detailed in pre-registration

[2] All coefficients for all models available in supplemental materials

Table 1: Marginal accuracy (%) from logistic mixed effect regression, and marginal mean RT (ms, with standard errors in parentheses) from linear mixed effect regression, for each articulatory suppression condition in Experiments 1 and 2.

	Retrieval			
	No suppression		Suppression	
Encoding	%	RT (SE)	%	RT (SE)
Experiment 1				
No suppression	92.9	2499 (137)	89.6	2288 (138)
Suppression	88.3	2916 (138)	82.7	2687 (139)
Experiment 2				
No suppression	92.0	2786 (144)	90.2	2635 (144)
Suppression	83.4	2854 (141)	82.7	2675 (141)

Experiment 2

In our second experiment (pre-registration, data, analysis code, and full results are available as supplemental materials at https://osf.io/acv3m/?view_only=c1799106289a4063abf2eaa490eae009), we presented 12 objects per sequence rather than 6, and made methodological improvements to the design. Our hypotheses remained the same.

Method

Participants As in Experiment 1, we used Bayesian sequential hypothesis testing to determine sample size. Bayes Factors for Step 3 cleared the evidence threshold for the null at *N*min = 32 for both RT (BF_{10} = 0.02) and accuracy BF_{10} = 0.03). However, sequential analysis plots for the Step 2 model (the best-fitting model in Experiment 1) suggested that the level of evidence was still unstable for RT (i.e., BFs fluctuating between evidence for the null and the alternative, and equivocal evidence). We tested additional participants until it stabilised at *N* = 44. We therefore report results for 44 participants (33 female; mean age = 20.3 years, *SD* = 5.4). All other BF inferences and parameter estimates were consistent between *N* = 32 and *N* = 44 (full data and statistics in supplementals). Three participants were replaced due to failure to follow instructions.

Materials and procedure We used materials and procedure from Experiment 1 with the following methodological changes: to reduce the risk of ceiling effects, we paired sequences from Experiment 1, which resulted in six lists of 12 items each. Instead of a label for each list, participants were given brief information about the context (e.g.: "*You are making dinner and need to remember your shopping list for a meal and tea. Press space to proceed to the list of ingredients.*"), to provide a real-life, ecologically valid situation.

To prevent participants' knowledge of the articulatory suppression condition from affecting their encoding strategies, we altered the presentation of retrieval conditions. Instead of verbal instructions on articulatory suppression at the start of the experiment, participants saw an image of a mouth on the screen indicating the start of articulatory suppression, and the same image crossed out to indicate no articulatory suppression, before the encoding and retrieval stages on each trial. We then randomised the order of lists across retrieval conditions, so that participants did not know whether the trial involved articulatory suppression until encoding was complete. We also altered some of the distractors (N = 8; 0.015% of all items) to ensure that the target items were not easy to guess without relying on memory.

The experimental design remained the same (i.e., articulatory suppression manipulated between participants at encoding and within participants at retrieval).

We changed the "wait" screen to reduce the possibility of participants relying on perceptual matching (instead of memory for object concepts). Rather than passively looking at the screen, participants had to click on 4 dots appearing in random points on the screen to "calibrate the mouse". Additionally, object presentation time during encoding was prolonged to 2000ms, to give participants more time to encode the concept.

Results and Discussion

No trials were excluded on the basis of accuracy results. For RT analysis of correct responses, 31 trials (0.012% of data) were removed as being more than 3 SDs above the individual participant's mean.

Accuracy Bayesian model comparison showed strong evidence *against* Step 2 over Step 1, BF_{10} = 0.017; the data were 57 times more likely under the Step 1 model containing only random effects than a model containing articulatory suppression as fixed effects at encoding and retrieval. There was also strong evidence at Step 3 *against* the effect of the encoding-retrieval interaction on accuracy, BF_{10} = 0.025: the data were 40 times more likely under the Step 2 model with no interaction than the Step 3 model with the interaction.

We then used the coefficients in the Step 3 model to estimate the marginal accuracy for each encoding × retrieval articulatory suppression condition (see Table 1). As in Experiment 1, accuracy was best in the no-suppression/no-suppression condition (no articulatory suppression in either encoding or retrieval), with participants correctly recognising 11.0 out of 12 objects per sequence on average. However, accuracy was worst in the suppression/suppression condition (9.9 objects remembered). Object memory was least accurate when access to language was blocked at both encoding and retrieval, in line with Experiment 1 but not our predictions.

Although the BFs showed evidence against both models, we ran an exploratory analysis of the individual coefficients in the Step 2 model to make a comparison with Experiment 1. Articulatory suppression at encoding had a negative effect on response accuracy, beta = -0.736, *SE* = 0.273, *z* = -2.69, *p* = .007. As predicted, and replicating Experiment 1, removing access to language impaired object memory accuracy: when access to labels was blocked at encoding, people were 109% more likely to make an error when later attempting to recognise the object. Unlike Experiment 1, articulatory suppression at retrieval had little effect, beta = -0.117, *SE* = 0.100, *z* = -1.17, *p* = .243, decreasing the probability of a correct response by only 12%. However, the NHST effect of articulatory suppression at encoding was not consistent with the Bayesian model comparison at Step 2 (which added encoding and retrieval effects simultaneously), and so should be treated cautiously.

Reaction Time Bayesian model comparison showed equivocal evidence for Step 2 over Step 1, $BF_{10} = 0.61$. As in Experiment 1, there was strong evidence at Step 3 *against* the effect of the encoding-retrieval interaction, $BF_{10} = 0.02$.

We used the coefficients in the Step 3 model to estimate the marginal mean RT for each condition of encoding × retrieval articulatory suppression (see Table 1). Against our predictions, but in line with Experiment 1, recognition was best (fastest) in the no-suppression/suppression condition (language available at encoding but not retrieval), and worst (slowest) in the suppression/no-suppression condition (language available at retrieval but not encoding). People had most difficulty remembering objects when language was blocked at the point of encoding but was available at retrieval.

We report an exploratory analysis of the most likely model of fixed effects (Step 2). Articulatory suppression at encoding had no effect on speed of recognition, unlike Experiment 1, beta = 54.22, SE = 149.60, $t(43.68) = 0.36$, $p = .719$. Against our expectations but in line with Experiment 1, articulatory suppression at retrieval *reduced* RT, beta = -164.91, SE = 43.26, $t(2430.48) = -3.81$, $p < .001$. People were *faster* to recognise target objects if language was blocked at retrieval. Closer examination of RT and accuracy suggested that this may be due to a speed-accuracy trade-off, as in Experiment 1. Faster RTs due to articulatory suppression at retrieval were accompanied by a trend towards poorer accuracy, relative to no-suppression conditions (see Table 1).

Summary The results were broadly in line with Experiment 1 and support the hypothesis that memory for object concepts typically relies on language. Blocking language access while encoding a real-world object sequence impaired performance in terms of accuracy (but not latency), in line with the idea that an object concept is typically stored in WM via sensorimotor simulation *and* its linguistic label, and when only sensorimotor simulation is available, memory is adversely affected.

However, blocking language access while retrieving objects from working memory resulted in faster speed of responding that was not completely eliminated by methodological changes. These results suggest that participants adopted a heuristic strategy for responding while performing articulatory suppression at retrieval. For instance, participants may have adopted a satisficing approach to selecting the target object, based on a rapid assessment of superficial sensorimotor similarity between a concept in WM and the objects in the array, to compensate for lack of access to linguistic information. Alternatively, perhaps the articulatory suppression task itself made participants want to get through the task faster, which resulted in emphasis on speed at the cost of accuracy. We plan to follow up these possibilities in future work.

General Discussion

The study is the first to take a linguistic-sensorimotor approach to working memory. We used real-world object sequences to account for the complex nature of naturalistic concepts that can draw upon information in long-term memory, and an articulatory suppression task to investigate the role of linguistic labels in working memory for such objects. We found that blocking language access at encoding impairs memory performance (poorer speed and accuracy in Experiment 1; poorer accuracy in Experiment 2), whereas blocking access to language during retrieval leads to an apparent speed-accuracy trade-off (faster speed and poorer accuracy in Experiment 1; faster speed in Experiment 2).

The results support the sensorimotor-linguistic theories of conceptual processing that argue the importance of language in conceptual representation (Connell, 2018; Louwerse, 2011), and the linguistic bootstrapping hypothesis that proposes word labels act as placeholders in mental representations when resources are insufficient to maintain a full simulation (Connell & Lynott, 2014). When language is available, people encode objects in WM with linguistic labels and sensorimotor simulations, and when storage is at capacity, the linguistic placeholder can free up resources by allowing to drop a simulation. Experiment 2 results suggest that linguistic bootstrapping allows people to remember one extra concept in a sequence of 12 (11 rather than 10).

We expected memory performance to be impaired the most when participants could use linguistic bootstrapping at encoding but had no access to object labels at retrieval (no-suppression/suppression condition). This effect did not appear. Instead, when language access was blocked at retrieval, participants adopted an alternative, heuristic strategy to compensate for it, which resulted in a trade-off between speed and accuracy. Participants may have relied on an incomplete sensorimotor simulation in working memory (e.g., only the shape or the colour of the target object), which allowed them to respond quickly, but not always correctly.

The results highlight the importance of language in working memory performance for real-world object sequences. We plan to explore encoding and retrieval processes in more detail by testing complex stimuli in sequences of varying lengths and under time constraints.

Acknowledgments

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement No 682848.

References

Baddeley, A. (2000). The episodic buffer: a new component of working memory?. *Trends in cognitive sciences*, *4*(11), 417-423.

Barsalou, L. W., Santos, A., Simmons, W. K., & Wilson, C. D. (2008). Language and simulation in conceptual processing. In M. De Vega, A. M. Glenberg, and A. C. Graesser (Eds), *Symbols, Embodiment, and Meaning*. Oxford: Oxford University Press.

Borghi, A. M., Barca, L., Binkofski, F., Castelfranchi, C., Pezzulo, G., & Tummolini, L. (2018). Words as social

tools: Language, sociality and inner grounding in abstract concepts. *Physics of Life Reviews*. Advance Online Publication

Connell, L. (2018). What have labels ever done for us? The linguistic shortcut in conceptual processing. *Language, Cognition and Neuroscience*, *0*(0), 1–11. https://doi.org/10.1080/23273798.2018.1471512

Connell, L., & Lynott, D. (2013). Flexible and fast: Linguistic shortcut affects both shallow and deep conceptual processing. *Psychonomic Bulletin and Review*, *20*(3), 542–550. https://doi.org/10.3758/s13423-012-0368-x

Connell, L., & Lynott, D. (2014). Principles of representation: Why you can't represent the same concept twice. *Topics in Cognitive Science*, *6*(3), 390–406. https://doi.org/10.1111/tops.12097

Connell, L., Lynott, D., & Dreyer, F. (2012). A functional role for modality-specific perceptual systems in conceptual representations. *PLoS ONE*, *7*(3). https://doi.org/10.1371/journal.pone.0033321

Dutriaux, L., Dahiez, X., & Gyselinck, V. (2018). How to change your memory of an object with a posture and a verb. *Quarterly Journal of Experimental Psychology*, 174702181878509. https://doi.org/10.1177/1747021818785096

Hauk, O., Johnsrude, I., & Pulvermüller, F. (2004). Somatotopic representation of action words in human motor and premotor cortex. *Neuron*, *41*(2), 301–307.

Langerock, N., Vergauwe, E., & Barrouillet, P. (2014). The maintenance of cross-domain associations in the episodic buffer. *Journal of Experimental Psychology: Learning Memory and Cognition*, *40*(4), 1096–1109. https://doi.org/10.1037/a0035783

Louwerse, M., & Connell, L. (2011). A Taste of Words: Linguistic Context and Perceptual Simulation Predict the Modality of Words. *Cognitive Science*, *35*(2), 381–398. https://doi.org/10.1111/j.1551-6709.2010.01157.x

Louwerse, M., & Jeuniaux, P. (2008). Language comprehension is both embodied and symbolic. *Symbols and Embodiment: Debates on Meaning and Cognition*, 309–326. https://doi.org/10.1093/acprof:oso/9780199217274.003.0015

Louwerse, M. M. (2011). Symbol interdependency in symbolic and embodied cognition. *Topics in Cognitive Science*, *3*(2), 273–302. https://doi.org/10.1111/j.1756-8765.2010.01106.x

Louwerse, M. M., & Jeuniaux, P. (2010). The linguistic and embodied nature of conceptual processing. *Cognition*, *114*(1), 96–104. https://doi.org/10.1016/j.cognition.2009.09.002

Vermeulen, N., Chang, B., Mermillod, M., Pleyers, G., & Corneille, O. (2013). Memory for words representing modal concepts. *Experimental psychology*, *60*(4), 293-301.

Stanfield, R. A. & Zwaan, R. A. (2001). The Effect of Implied Orientation Derived from Verbal Context on Picture Recognition. *Psychological Science*, *12*(2), 153–156.

Souza, A. S., & Skóra, Z. (2017). The interplay of language and visual perception in working memory. *Cognition*, *166*, 277–297. https://doi.org/10.1016/j.cognition.2017.05.038

Tranel, D., Damasio, H., & Damasio, A. R. (1997). A neural basis for the retrieval of conceptual knowledge. *Neuropsychologia*, *35*(10), 1319-1327.

Vigliocco, G., Meteyard, L., Andrews, M., & Kousta, S. (2009). Toward a theory of semantic representation. *Language and Cognition*, *1*(02), 219–247. https://doi.org/10.1515/LANGCOG.2009.011

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of p values, *Psychonomic bulletin & review*, *14*(5), 779–804.

Zormpa, E., Brehm, L. E., Hoedemaker, R. S., & Meyer, A. S. (2018). The production effect and the generation effect improve memory in picture naming. *Memory*, *27*(3), 1–13. https://doi.org/10.1080/09658211.2018.1510966

How can I help? Developmental change in the selectivity of two to four-year-olds' attempts to alleviate others' distress

Regina Ebo (reginae@mit.edu)
Department of Brain and Cognitive Sciences, MIT
43 Vassar St Cambridge, MA 02139

Laura Schulz (lschulz@mit.edu)
Department of Brain and Cognitive Sciences, MIT
43 Vassar St Cambridge, MA 02139

Abstract

Young children are selective in deciding whom to help (i.e., they preferentially assist and share resources with prosocial versus antisocial others; Hamlin, Wynn, Bloom, & Mahajan, 2011; Vaish, Carpenter, & Tomasello, 2010) but are they also selective in deciding *how* to offer help? Here we show two to five-year-olds (N = 32; mean: 42.41 months; range 27-68 months) characters who are distressed for different reasons: they are hurt, bored, or sad. Children of all ages tried to help the agent but the selectivity of children's responses varied with age and condition; in particular, children's responses to boredom and sadness became increasingly differentiated with age.

Keywords: helping, empathy, social cognition, theory of mind, preschoolers, toddlers

Introduction

One of the more charming characteristics of young children is that they try to help others, even at ages when they themselves need help with almost every aspect of daily life. Toddlers who struggle to put on their own socks will open doors and pick up objects for others (Warneken & Tomasello, 2006; 2007), point to show others the location of hidden objects (Liszkowski, Carpenter, Striano, & Tomasello, 2006), hug and pat distressed peers (Friedman, Zahn-Waxler, & Radke-Yarrow, 1982), and try to understand the causes of others' distress (Knafo, Zahn-Waxler, Van Hulle, Robinson, & Rhee 2008; Zahn-Waxler, Radke-Yarrow, Wagner, & Chapman 1992; Zahn-Waxler, Robinson, & Emde 1992). Children's empathetic and prosocial behavior increases between two and four years of age (Knafo et al., 2008; Volbrecht, Lemery-Chalfant, Aksan, Zahn-Waxler, & Goldsmith, 2004; Zahn-Walxer, et al., 1992). This is arguably mediated by broad changes in their theory of mind (Miller, Eisenberg, Fabes, & Shell, 1996; Wellman, Cross, & Watson, 2001), specific changes in their emotion understanding and emotion regulation (Denham, 1998; Eisenberg, Spinrad, & Sadovsky, 2006), and increased socialization towards prosocial behaviors (Hoffman, 2000).

But the selectivity of children's helping behavior also increases over development (Hay & Cook, 2007; Hay, 1994) -- and even the youngest children do not help others indiscriminately. Toddlers preferentially help prosocial versus antisocial others (Behne, Carpenter, Call, & Tomasello, 2005; Dunfield & Kuhlmeier, 2010; Hamlin, et. al., 2011; Vaish, et al., 2010). By three, children consider others' past contributions to shared goals (Baumard Mascaro, & Chevallier, 2012) and history of reciprocity in deciding how to allocate resources (Olson & Spelke, 2008). Four and five-year-olds evaluate relative ability in deciding how to divide labor to achieve cooperative and prosocial goals (Magid, DePascale, & Schulz, 2018). By five and six, children's attempts to inform others take into account the learners' prior knowledge, past mistakes, and goals (Gweon, Shafto, & Schulz, 2014; Ronfard, Was, & Harris, 2016), the transparency and availability of information (Clegg & Legare, 2016; Ronfard, Was, & Harris, 2016), and the relative costs and benefits of information to the learner (Bridgers, Jara-Ettinger, & Gweon, 2016; Gweon & Schulz, 2019).

Thus, children's helping behavior is sophisticated in many respects. However, toddlers and young preschoolers are more likely to share resources or provide help with instrumental goals than offer comfort (Dunfield, Kuhlmeier, O'Connell, & Kelley, 2011; Newton, Thompson, & Goodman, 2016; Svetlova, Nichols, & Brownell, 2010). Similar results have been found in four and five-year-olds: they are more likely to help achieve goals than to share, and are least likely to try to offer soothing, encouragement or solace (Thompson & Newton, 2013).

Because very young children are adept at inferring both others' desires (e.g., Meltzoff, Gopnik, & Repacholi, 1999) and the goals of their failed intentional actions (e.g., Meltzoff, 1999), it may be relatively easy for young children to know what resources to offer and what actions to take. By contrast, it may be difficult for children to know what constitutes a helpful response to someone's emotional distress. Even as adults, we may understand perfectly well that someone is disappointed, agitated, or distraught and still find ourselves at a loss as to how to help them.

However, even if children do not know how best to intervene, there is reason to think that they may be attuned even to relatively fine-grained distinctions among emotions.

Within hours of birth, newborns respond differently to distinct emotional expressions (Field, Woodson, Greenberg, & Cohen, 1982) and by seven months, babies distinguish emotions cross-modally and within valence (e.g., generating distinct responses to anger and fear; matching happy faces to happy voices and interested faces to interested ones; Serrano, Iglesias, & Loeches, 1992; Walker-Andrews, 1998; see also Soken & Pick, 1999; Soderstrom, Reimchen, Sauter, & Morgan, 2017). Older infants map positively valenced emotions to the achievement of goals (Skerry & Spelke, 2014), and make nuanced distinctions among emotional expressions and connect them to their probable eliciting causes (Wu, Muentener, & Schulz, 2018).

Nonetheless, children's ability to categorize emotions (Widen & Russell, 2008; 2010), and their understanding of the way past experiences and social contexts shape the experience and expression of emotions (Pons, Harris, & de Rosnay, 2004), undergo considerable development between preschool and middle childhood. Emotion regulation in particular is relatively protracted (Pons et al., 2004), and this may apply to the ability to regulate other's emotions as well as one own. Moreover, perhaps the most common way to try to regulate someone else's negative emotions is to talk to them, thus offering comfort might place high verbal demands on children. The infrequency with which young children offer comfort may reflect limitations on their fluency, not their insight or compassion. In the current study, we remove linguistic demands by giving children a choice of objects that might be helpful, allowing us to ask whether children can calibrate their responses to the particular nature of others' distress.

Here we focus on two to four-year-olds because we know children in this age range can use social and moral evaluation to decide *whom* to help (Behne et al., 2005; Dunfield & Kuhlmeier, 2010; Vaish et al., 2010; Baumard, et al., 2012; Olson & Spelke, 2008) but the degree to which they use social cognition to make distinctions about *how* to help remains an open question. We show children characters who are upset for one of three reasons: they have scraped their knee and are hurt, there is nothing to do and they are bored, or their parent has left them at daycare and they are sad. In all cases, children are given a choice of three candidate offerings: a Band-Aid, a novel electronic toy, or the victim's favorite stuffed animal. We selected these pairings because both the emotional states and the stimuli should be familiar to children in this age range and yet the complexity of the inferences required to intervene upon the emotional states might differ across categories. In particular, children's tendency to choose an intervention might be related to the intuitive likelihood that the intervention would successfully change the agent's state.

Children have abundant experience with minor scrapes and bumps (Fearon, McGrath, & Achat, 1996), and in the United States, "booboos" are reliably linked with Band-Aids. Crying in response to a minor injury is an ambiguous response with respect to the extent to which it reflects a physiological response to pain or an emotional response to the fear associated with the pain, but in either case, from the perspective of a child, a Band-Aid may seem to solve the underlying problem. By contrast, there is no single canonical response to either boredom or sadness; intervening on these emotional states requires both understanding why the person feels as she does and understanding the role that the various choices may play in changing this state. Nonetheless, the link between boredom and novelty is arguably almost as straightforward as the link between booboos and Band-Aids: Children themselves respond to novelty with interest (Berlyne, 1950; Hutt, 1970) and providing something that is interesting effectively solves the problem of being bored. However, the distress of an agent who is sad about a separation is more complex. Children commonly regulate their sadness at separation from attachment figures with transitional objects (Kopp, 1989; Winnicott, 1986). Critically however, the intervention serves to regulate the distressed emotion rather than to resolve it (i.e., the only intervention that really solves sadness at separation from a loved one is for the loved one to return). Thus, although pilot data suggests that adults would offer Band-Aids, novel toys, and favorite stuffed animals in response to pain, boredom, and sadness respectively, children might well find some of these mappings easier than others.

Of course, if children offer anything at all to an agent who is upset, they are providing an empathetic, prosocial response, and any well-intentioned intervention may be effective even if it is not directly connected to the underlying concern. Band-Aids can alleviate boredom and sadness; novel toys can distract from sadness and pain, and stuffed animals can help with both pain and boredom. Perhaps more critically, engagement, attention, and sympathetic concern may go a long way towards resolving distress, independent of the degree to which any given intervention is specifically tailored to the source of the recipient's woes.

Nonetheless, commonsense suggests that some of these offerings are more likely to be effective in some contexts than others, and the early sophistication of children's helping behavior may relate to sensitivity to the contents of others' minds and overall social acuity. Thus, here we ask whether two to four-year-olds offer emotional comfort indiscriminately or whether they are sensitive to how different interventions might best alleviate different kinds of emotional distress.

Experiment

Participants

Thirty-two children (M = 42.41 months, range: 27-68 months) were recruited from an urban children's museum. Six children failed a practice trial but excluding them from the analysis made no difference to the results. Five additional children were recruited but excluded from analysis due to incomplete participation (N = 1), parental interference (N = 2), and incomplete consent forms (N = 2).

While most of the children were white and middle class, a range of ethnicities and socioeconomic backgrounds reflecting the diversity of the local population (47% European American, 24% African American, 9% Asian, 17% Latino, 4% two or more races) and the museum population (29% of museum attendees receive free or discounted admission) were represented.

Materials

In the practice trial an Ernie puppet from Sesame Street and two toys (a squishy ball and a plastic strawberry) were used. The test trial materials included six Paw Patrol Band-Aids each depicting a different or different set of characters from the children's series; six unique toys that lit-up, made funny sounds, and/or spun; and six unique stuffed animals. The materials were arranged on a plastic tray so that one material of each of the three kinds was placed on the tray (left/right/middle arrangement counterbalanced); children were presented with a different set of three items on each trial. (See Figure 1 for example presentation set). We also used six pairs of hand puppets; each pair had a parent and a child puppet. Stickers and a sticker "bookmark" were used to keep the children on task. See Figure 1 for examples of the stimuli.

Figure 1: Examples of the puppets (left) and the toy, stuffed animal, and Band-Aid (right)

Procedure

Children were tested individually in a private testing room. Children participated in a practice trial and six test trials, two of each kind (Hurt, Bored, and Sad). Sessions began with the experimenter explaining the task: "We're going to do six puppet stories okay? After each one, you get to put a sticker on this bookmark. Once we finish all six stories, you get to take the bookmark home. Does that sound good? Great! Before we start, we're going to do a practice story."

Practice Trial: The experimenter brought out the Ernie puppet and said, "This is my friend Ernie. He's really hungry. He hasn't eaten all day." Then the experimenter introduced the tray with the squishy toy and the strawberry. "Here we have a squishy toy and a strawberry. Which one of these things do you want to give him to make him feel better?" Choosing the correct option, the strawberry was met with positive feedback (Ernie said "thank you" and pretended to eat the strawberry "mm, mm, mm"); choosing the incorrect option was met with neutral feedback ("thank you"). Regardless of whether children passed the training trial, they continued onto the test trials.

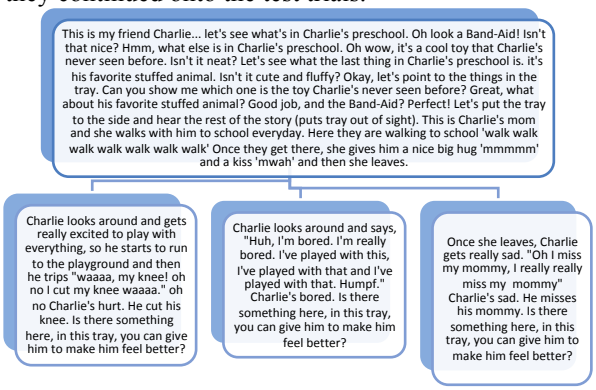


Figure 2: Example of the script and the three scenarios.

Test Trial Each test trial began with the introduction of the child puppet and three things in his preschool: a new toy that the puppet had never seen before, the puppet's favorite stuffed animal, and a Band-Aid. The child was given each item one at a time and told to place the item in the tray once they were done looking at it. Children were allowed to play for as long as they liked to minimize the chance that children would choose an item just to play more with it.
Children heard a core story and one of three possible endings: Hurt, Bored, and Sad. In the Hurt condition, the child puppet tripped and hurt his knee; in the Bored condition, the puppet got bored; in the Sad condition, the puppet got sad because he misses his mom. Participants were prompted to pick an item from the tray that would make the child puppet feel better. (See Figure 2 for an example.) Children received neutral feedback ("thank you" or "thanks for helping"); then a new pair of puppets and a new tray with three different items, one of each kind, was introduced. The scenarios were presented in random order for the first three trials and this order was repeated for the last three trials.

Results

Children were counted as performing correctly if they chose the Band-Aid for the Hurt scenarios, the new toy for the Bored scenarios, and the stuffed animal for the Sad scenarios. There was no effect of order on children's performance (Kruskal-Wallis rank sum test, $p = .44$).

Children had a choice of three items on each of the six trials. They received one point for each correct choice. Overall, children performed above chance (mean = 3.5; one-sample t-test, $p < .0001$). Only one child (the oldest) performed at ceiling but nine of the thirty-two children (28%) answered five of the six questions correctly ($p < .0001$ by binomial test). There was an effect of age on children's overall score ($r^2 = 0.15$, $p < .0001$; see Figure 3); older children performed better, and as clear in Figure 4, the effect was driven primarily by improvement in the Sad condition.

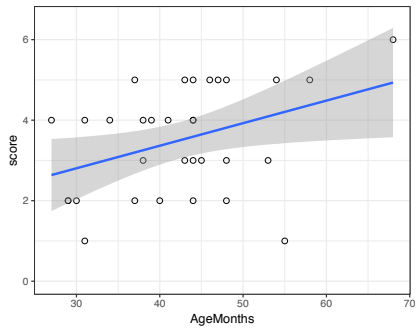


Figure 3: Children's overall score as a function of age

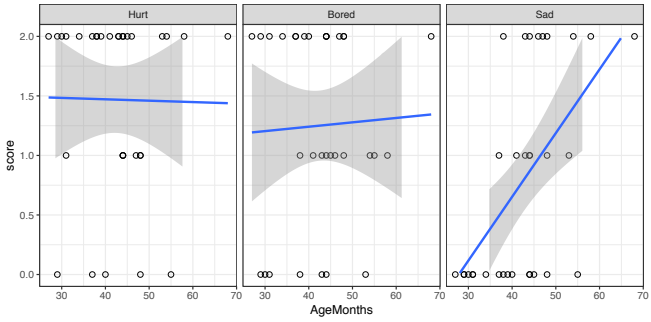


Figure 4: Children's responses by age and condition

The kind of scenario affected children's score (Test of Equal Proportions, $p < .001$); thus we used pairwise comparisons to look within each scenario at children's performance. Children performed better in both the Hurt and Bored conditions than in the Sad condition (Hurt versus Sad; $p < .001$; Bored vs. Sad, , $p < .05$; see Figure 5); children's performance in the Hurt and Bored conditions did not differ from each other (Hurt vs. Bored; $p = .26$). Within each condition, children performed above chance in both the Hurt ($p < .0001$ by Test of Equal Proportions) and Bored conditions ($p < .0001$); but their scores in the Sad condition were not significantly different from chance ($p = 0.78$).

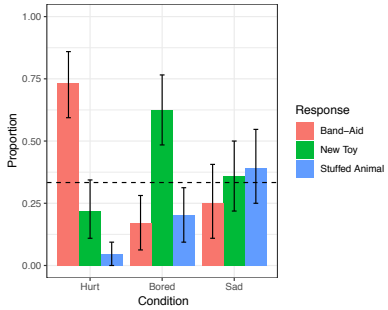


Figure 5: Children's choice of each of the three responses by condition

Discussion

Above all, these results suggest that, at least in simple forced choice contexts with low verbal demands, very young children's helping behavior is not restricted to resource sharing or assisting with functional, goal-directed actions; children seek to help relieve others' distress and do so in ways that are responsive to distinct sources of negative affect. Although children's ability to calibrate their response to the emotional state improved over development, even children as young as two and three distinguished upset due to pain and upset due to boredom and generated distinct, appropriate responses.

As predicted however, children had more difficulty knowing how to respond to distress due to a separation. We hypothesized that this might be because the impact of the intervention on the outcome was more uncertain. None of the candidate options would directly remove the source of distress; the best children could do would be to offer something that would help moderate it. There are possibilities however. Children may simply have preferred the fun toy to the stuffed animal – inflating their performance in the Bored condition and impairing it in the Sad condition. We think this interpretation is unlikely however, given both the method and results: We intentionally allowed children to play with each item to satiation in advance to wash out any differential stimuli effects, and children had no difficulty overcoming any preference for the toy in the Hurt condition.

Alternatively, young children might genuinely believe that the other options (toys or Band-Aids) were more likely to provide comfort than the stuffed animal – and indeed, at least for some children, in some contexts, this might be correct. Indeed, emotion regulation is challenging because there are no determinate rules: what works one time might not work the next, and what works for one person might not work for another. Nonetheless, within a given culture and context, there is a probabilistic relationship between certain responses and outcomes, and the current results suggest that children begin to learn these relations over the preschool years. Future research might extend this study to older children to see if their responses are adult-like or even provide children with the option to not help the puppet. Future research might also look at children's sensitivity to culturally specific, or family specific, dimensions of emotion regulation to look at how socialization affects children's responses.

Overall however, these results suggest that children's empathetic responses are not monolithic. With apologies to Tolstoy, even two-year-olds seem to recognize that every unhappy puppet is unhappy in its own way – and they offer solace accordingly.

Acknowledgments

We thank Boston Children's Museum, families who participated in this study, and graduate students in ECCL.

References

Baumard, N., Mascaro, O., & Chevallier, C. (2012). Preschoolers are able to take merit into account when distributing goods. *Developmental psychology*, *48*(2), 492.

Behne, T., Carpenter, M., Call, J., & Tomasello, M. (2005). Unwilling versus unable: infants' understanding of intentional action. *Developmental psychology*, *41*(2), 328.

Berlyne, D. E. (1950). Novelty and Curiosity as Determinants of Explorative Behavior. *British Journal of Psychology. General Section*, *41*(1-2), 68-80.

Borke, H. (1971). Interpersonal perception of young children: Egocentrism or empathy? *Developmental Psychology, 5*(2), 263-269.

Bridgers, S., Jara-Ettinger, J., & Gweon, H. (2016). Children consider others' expected costs and rewards when deciding what to teach. In *Proceedings of the 38th annual conference of the cognitive science society* (pp. 559-564). Cognitive Science Society.

Clegg, J. M., & Legare, C. H. (2016). Instrumental and conventional interpretations of behavior are associated with distinct outcomes in early childhood. *Child Development*, *87*(2), 527-542.

Dondi, M., Simion, F., & Caltran, G. (1999). Can newborns discriminate between their own cry and the cry of another newborn infant?. *Developmental Psychology*, *35*(2), 418.

Denham, S. A. (1998). *Emotional development in young children*. Guilford Press.

Dunfield, K. A., & Kuhlmeier, V. A. (2010). Intention-mediated selective helping in infancy. *Psychological science*, *21*(4), 523-527.

Dunfield, K., Kuhlmeier, V. A., O'Connell, L., & Kelley, E. (2011). Examining the diversity of prosocial behavior: Helping, sharing, and comforting in infancy. *Infancy*, *16*(3), 227-247.

Eisenberg, N., Spinrad, T. L., & Sadovsky, A. (2006). Empathy-related responding in children. *Handbook of moral development*, *517*, 549.

Fearon, I., McGrath, P. J., & Achat, H. (1996). 'Booboos': the study of everyday pain among young children. *Pain*, *68*(1), 55-62.

Feshbach, N. D. (1975). Empathy in Children: Some Theoretical and Empirical Considerations. *The Counseling Psychologist*, *5*(2), 25–30. https://doi.org/10.1177/001100007500500207

Field, T. M., Woodson, R., Greenberg, R., & Cohen, D. (1982). Discrimination and imitation of facial expression by neonates. *Science*, *218*(4568), 179-181.

Friedman, S. L., Zahn-Waxler, C., & Radke-Yarrow, M. (1982). Perceptions of cries of full-term and preterm infants. *Infant Behavior and Development*, *5*(2-4), 161-173.

Gweon, H., & Schulz, L. (2019). From exploration to instruction: Children learn from exploration and tailor their demonstrations to observers' goals and competence. *Child development*, *90*(1), e148-e164.

Gweon, H., Shafto, P., & Schulz, L. (2014, January). Children consider prior knowledge and the cost of information both in learning from and teaching others. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 36, No. 36).

Hamlin, J. K., Wynn, K., Bloom, P., & Mahajan, N. (2011). How infants and toddlers react to antisocial others. *Proceedings of the national academy of sciences*, *108*(50), 19931-19936.

Hay, D. F. (1994). Prosocial development. *Journal of Child Psychology and Psychiatry*, *35*(1), 29-71.

Hay, D. F., & Cook, K. V. (2007). The transformation of prosocial behavior from infancy to childhood. *Socioemotional development in the toddler years: Transitions and transformations*, 100-131.

Hoffman, M. L. (2008). Empathy and prosocial behavior. *Handbook of emotions*, *3*, 440-455.

Hutt, C. (1970). Specific and diversive exploration1. In *Advances in child development and behavior* (Vol. 5, pp. 119-180). JAI.

Kopp, C. B. (1989). Regulation of distress and negative emotions: A developmental view. *Developmental psychology*, *25*(3), 343.

Knafo, A., Zahn-Waxler, C., Van Hulle, C., Robinson, J. L., & Rhee, S. H. (2008). The developmental origins of a disposition toward empathy: Genetic and environmental contributions. *Emotion*, *8*(6), 737.

Liszkowski, U., Carpenter, M., Striano, T., & Tomasello, M. (2006). 12-and 18-month-olds point to provide information for others. *Journal of cognition and development*, *7*(2), 173-187.

Magid, R. W., DePascale, M., & Schulz, L. E. (2018). Four-and 5-Year-Olds Infer Differences in Relative Ability and Appropriately Allocate Roles to Achieve Cooperative, Competitive, and Prosocial Goals. *Open Mind*, *1*(4), 194-207.

Meltzoff, A. N. (1999). Origins of theory of mind, cognition and communication. *Journal of communication disorders*, *32*(4), 251-269.

Meltzoff, A. N., Gopnik, A., & Repacholi, B. M. (1999). Toddlers' understanding of intentions, desires and emotions: Explorations of the dark ages.

Miller, P. A., Eisenberg, N., Fabes, R. A., & Shell, R. (1996). Relations of moral reasoning and vicarious emotion to young children's prosocial behavior toward peers and adults. *Developmental psychology*, *32*(2), 210.

Newton, E. K., Thompson, R. A., & Goodman, M. (2016). Individual differences in toddlers' prosociality: Experiences in early relationships explain variability in prosocial behavior. *Child development*, *87*(6), 1715-1726.

Olson, K. R., & Spelke, E. S. (2008). Foundations of cooperation in young children. *Cognition*, *108*(1), 222-231.

Pons, F., Harris, P. L., & de Rosnay, M. (2004). Emotion comprehension between 3 and 11 years: Developmental periods and hierarchical organization. *European journal of developmental psychology*, *1*(2), 127-152.

Ronfard, S., & Corriveau, K. H. (2016). Teaching and preschoolers' ability to infer knowledge from mistakes. *Journal of experimental child psychology*, *150*, 87-98.

Ronfard, S., Was, A. M., & Harris, P. L. (2016). Children teach methods they could not discover for

themselves. *Journal of experimental child psychology*, *142*, 107-117.

Serrano, J. M., Iglesias, J., & Loeches, A. (1992). Visual discrimination and recognition of facial expressions of anger, fear, and surprise in 4-to 6-month-old infants. *Developmental Psychobiology: The Journal of the International Society for Developmental Psychobiology*, *25*(6), 411-425.

Skerry, A. E., & Spelke, E. S. (2014). Preverbal infants identify emotional reactions that are incongruent with goal outcomes. *Cognition*, *130*(2), 204-216.

Soderstrom, M., Reimchen, M., Sauter, D., & Morgan, J. L. (2017). Do infants discriminate non-linguistic vocal expressions of positive emotions?. *Cognition and Emotion*, *31*(2), 298-311.

Soken, N. H., & Pick, A. D. (1999). Infants' perception of dynamic affective expressions: Do infants distinguish specific expressions?. *Child Development*, *70*(6), 1275-1282.

Svetlova, M., Nichols, S. R., & Brownell, C. A. (2010). Toddlers' prosocial behavior: From instrumental to empathic to altruistic helping. *Child development*, *81*(6), 1814-1827.

Thompson, R. A., & Newton, E. K. (2013). Baby altruists? Examining the complexity of prosocial motivation in young children. *Infancy*, *18*(1), 120-133.

Vaish, A., Carpenter, M., & Tomasello, M. (2010). Young children selectively avoid helping people with harmful intentions. *Child development*, *81*(6), 1661-1669.

Volbrecht, M. M., Lemery-Chalfant, K., Aksan, N., Zahn-Waxler, C., & Goldsmith, H. H. (2007). Examining the familial link between positive affect and empathy development in the second year. *The Journal of genetic psychology*, *168*(2), 105-130.

Walker-Andrews, A. S. (1998). Emotions and Social Development: Infants9 Recognition of Emotions in Others. *Pediatrics*, *102*(Supplement E1), 1268-1271.

Warneken, F., & Tomasello, M. (2006). Altruistic helping in human infants and young chimpanzees. *science*, *311*(5765), 1301-1303.

Warneken, F., & Tomasello, M. (2007). Helping and cooperation at 14 months of age. *Infancy*, *11*(3), 271-294.

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child development*, *72*(3), 655-684.

Widen, S. C., & Russell, J. A. (2008). Children acquire emotion categories gradually. *Cognitive development*, *23*(2), 291-312.

Widen, S. C., & Russell, J. A. (2010). Differentiation in preschooler's categories of emotion. *Emotion*, *10*(5), 651.

Winnicott, D. W. (1986). 10. Transitional Objects and Transitional Phenomena: A Study of the First Not-Me. *Essential papers on object relations*, 254.

Wu, Y., Muentener, P., & Schulz, L. E. (2016). The invisible hand: toddlers connect probabilistic events with agentive causes. *Cognitive science*, *40*(8), 1854-1876.

Zahn-Waxler, C., Radke-Yarrow, M., Wagner, E., & Chapman, M. (1992). Development of concern for others. *Developmental psychology*, *28*(1), 126

Zahn-Waxler, C., Robinson, J. L., & Emde, R. N. (1992). The development of empathy in twins. *Developmental psychology*, *28*(6), 1038.

Decomposing Human Causal Learning: Bottom-up Associative Learning and Top-down Schema Reasoning

Mark Edmonds[1,4] markedmonds@ucla.edu
Siyuan Qi[1,4] syqi@cs.ucla.edu
Yixin Zhu[2,4] yixin.zhu@ucla.edu
James Kubricht[3] kubricht@ucla.edu
Song-Chun Zhu[1,2,4] sczhu@stat.ucla.edu
Hongjing Lu[2,3] hongjing@ucla.edu

[1] Department of Computer Science, UCLA [2] Department of Statistics, UCLA [3] Department of Psychology, UCLA
[4] International Center for AI and Robot Autonomy (CARA)

Abstract

Transfer learning is fundamental for intelligence; agents expected to operate in novel and unfamiliar environments must be able to transfer previously learned knowledge to new domains or problems. However, knowledge transfer manifests at different levels of representation. The underlying computational mechanisms in support of different types of transfer learning remain unclear. In this paper, we approach the transfer learning challenge by decomposing the underlying computational mechanisms involved in bottom-up associative learning and top-down causal schema induction. We adopt a Bayesian framework to model causal theory induction and use the inferred causal theory to transfer *abstract* knowledge between similar environments. Specifically, we train a simulated agent to discover and transfer useful relational and abstract knowledge by interactively exploring the problem space and extracting relations from observed low-level attributes. A set of hierarchical causal schema is constructed to determine task structure. Our agent combines causal theories and associative learning to select a sequence of actions most likely to accomplish the task. To evaluate the proposed framework, we compare performances of the simulated agent with human performance in the OpenLock environment, a virtual "escape room" with a complex hierarchy that requires agents to reason about causal structures governing the system. While the simulated agent requires more attempts than human participants, the qualitative trends of transfer in the learning situations are similar between humans and our trained agent. These findings suggest human causal learning in complex, unfamiliar situations may rely on the synergy between bottom-up associative learning and top-down schema reasoning.

Introduction

The human capacity for inferring causal relations in unfamiliar environments is a hallmark of human intelligence (Mackie, 1974) that is often taken for granted in daily life. An illustrative example is that of the escape room—a prevalent social activity where groups of people inside of a locked room work together to complete sub-goals (puzzles) to achieve the goal—escape from the room. In order to succeed, teams must: (i) identify goal-relevant entities in the environment among distractors, (ii) develop a causal model for individual sub-goals, and (iii) interact with scene components to refine entity- and goal-based hypotheses. In this paper, we propose that inference in scenarios like the one above depends on two critical learning components. First, attributes relevant to candidate causal hypotheses are learned by interacting with entities in the scene, and second, causal hypotheses are refined based on newly encoded attribute-based knowledge.

It is worth noting that the above approach is generally inconsistent with early studies on causal learning in psychological research (Holyoak & Cheng, 2011). Early studies primarily focused on animal learning and conditioning experimental paradigms, framing causal understanding as learned stimulus-response relationships attained primarily through observation (*e.g.*, Shanks and Dickinson (1988)). Given associative weights on cue-effect links, the Rescorla-Wagner model was often utilized to explain how humans (and non-humans) construct expectations based on the co-occurrence of perceptual stimuli (Rescorla & Wagner, 1972). However, the knowledge that people have about causal mechanisms in the distal world has been shown to extend beyond the covariation between observed (perceptual) variables. For instance, adults interact with dynamic physical scenarios in ways that maximize information relevant to their causal hypotheses (Bramley, Gerstenberg, Tenenbaum, & Gureckis, 2018), and even infants test their beliefs about the physical characteristics of objects through exploration and experimentation (Stahl & Feigenson, 2015).

Contrary to the associative account, researchers have demonstrated that human learning and reasoning in novel (causal) environments rely heavily on the discovery of abstract causal structure (Waldmann & Holyoak, 1992) and strength (Cheng, 1997) rather than purely associative (statistical) dependencies. More recently, the integration of causal graphical models and Bayesian statistical inference (*i.e.*, Bayes nets) has provided a general representational framework for how this structure and strength is learned and transferred to novel situations (Griffiths & Tenenbaum, 2005, 2009; Tenenbaum, Griffiths, & Kemp, 2006; Bramley, Lagnado, & Speekenbrink, 2015; Bramley, Dayan, Griffiths, & Lagnado, 2017; Edmonds et al., 2018; Holyoak & Cheng, 2011). Under this framework, causal knowledge plays an essential role in constructing a flexible model of the world in which environmental states represent some status in the world, and connections between states imply the strength of a causal relationship.

We propose that creative discovery in novel domains relies on both causal structure *and* associations. Knowledge of causal structure enables agents to simulate how interventions will influence the environmental state, and without associations to guide exploration, the number of causal hypotheses to consider becomes intractable. For problem domains where the number of possible interventions is particularly high, the need for associative "guidance" can drastically improve decision-making. To solve this problem, we propose a computational model that integrates two learning mecha-

nisms: (i) a bottom-up process that determines which object attributes are causally relevant, and (ii) a top-down process that learns which abstract causal structures accomplish a task. The outcomes of actions are used to update the causal hypothesis space, and simulated agents learn a dynamics model capable of solving a challenging task.

We implement the proposed model in a virtual "escape room" environment where agents (human and artificial) are trapped in a room containing a single locked door and a set of conspicuous levers. The door of this room will unlock after the agent has interacted with the levers in a specific sequence. An agent placed in such a room may begin to randomly push or pull on the levers and revise their theory about the door's locking mechanism based on observed changes. Once an agent discovers a single solution, they are placed back into the same room and tasked with finding the next solution. The agent "escapes" from a room after finding *all* of the solutions which can be used to unlock the door.

After escaping from a room, agents are placed in a similar room but with newly positioned levers. Although the levers are in different positions, the new room is governed by the same abstract rules as the last (unknown to the agent). Thus, the agent's task is to identify the role of each lever in a new room. If the agent makes use of some knowledge from previous trials, we expect to observe fewer attempts in solving the problem. Because these rules are abstract descriptions of the latent state of the escape room, we refer to the underlying theory as a causal schema (*i.e.*, a conceptual organization of events identified as cause and effect; Heider, 1958). Once learned, this schema enables agents to transfer between different arrangements of levers in the room. The present work models the causal learning process from a hierarchical Bayesian perspective and makes three major contributions:

1. Utilizes a bottom-up associative learning paradigm to determine which attributes of the scene contribute to causal relations.
2. Utilizes a top-down causal schema model of the generalized operation of the environment to quickly adapt to similar but new scenarios.
3. Leverages causal hypotheses to learn a world model capable of transferring knowledge between seemingly dissimilar but structurally and causally equivalent environments.

The remainder of the paper is structured as follows. First, the OpenLock environment and experimental procedure are described, followed by an analysis of human performance from Edmonds et al. (2018). Next, components of the proposed model are described and corresponding results are provided. Finally, the paper concludes with a discussion of results and directions for future work.

Experiment: OpenLock Task

Participants

A total of 160 undergraduate students (114 female; mean age= 21.6) from the University of California, Los Angeles (UCLA) Department of Psychology subject pool and were compensated with course credit for their participation.

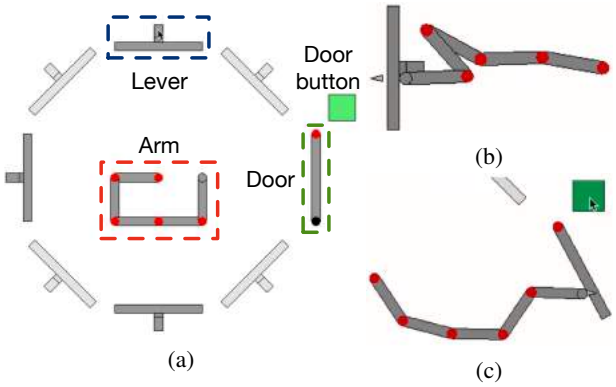


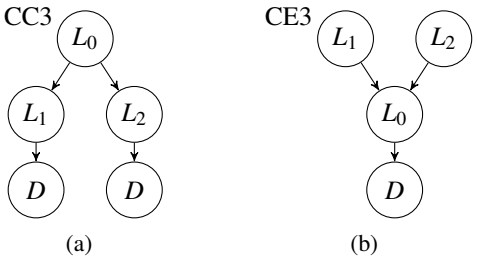
Figure 1: (a) Initial configuration of the room containing three active levers. All levers begin pulled toward the robot arm located at the center of the display. The arm interacts with levers by pushing/pulling them outward/inward. Only push actions are needed to unlock the door in each room (unknown to agents). White levers never move; this information is not explicitly stated. Once the door is unlocked, the green button can be clicked to command the arm to push the door open. The black circle located opposite the door's red hinge represents the door lock indicator (present if locked, absent if unlocked). (b) Pushing on a lever. (c) Opening the door by clicking the green button.

Materials and Procedure

In this section, we outline the OpenLock task, initially presented in Edmonds et al., 2018. In the task, agents are required to "escape" from a virtual room by unlocking and opening a door. The door unlocks after manipulating the levers in a particular sequence (see Figure 1). Each room consists of seven levers surrounding a robotic arm that can *push* or *pull* on each lever. While a subset of the levers is always involved in the locking mechanism (*i.e.*, active levers; colored grey), other levers are not causally relevant (*i.e.*, inactive levers; colored white). Agents observe the color of the levers and are expected to *learn* that grey levers–but not white levers–are always part of solutions in each room. Importantly, agents are tasked with finding *all* possible solutions for opening the door within a room. Participants are explicitly told that their goal is to open the door and are informed of how many solutions they have remaining in this room.[1]

The mechanics underlying the environment obey one of two causal schemas: Common Cause (CC) and Common Effect (CE) (see Figure 2). Requiring agents to find all solutions within a specific room ensures that agents abstract CC or CE schema structures. While a single solution corresponds to a single causal chain, a schema relies on nodes that are shared between multiple chains. Agents operate under a movement-limit constraint, where only three movements can be used to either (i) *push* or *pull* on levers (active or inactive), or (ii) *push* open the door. This constraint was placed on the agent to confine the search depth of possible solutions. After three movements, the episode terminates and the environment re-

[1]The video instructions presented to participants can be viewed at `https://vimeo.com/265302423`



(a) (b)

Figure 2: Common Cause (CC) and Common Effect (CE) structures used in the OpenLock task, in which L_i indicates a lever in the scene, and D indicates the effect of opening the door. In (a) CC3 and (b) CE3 condition, both include three causal cues but with different causal structures.

sets, regardless of the outcome. Agents also operate under a limited number of episodes (30) in a particular room, regardless of whether all solutions are found. We denote three movements as an *attempt* and each room as a *trial*. After completing a trial, agents move to a new trial (*i.e.*, room) with the same underlying causal schema but a different lever arrangement. This setup ensures that agents do not overfit their understanding of the environment to a single trial; *i.e.*, if agents are forming a useful abstraction, the knowledge they acquired in previous trials should aid in their ability to find all solutions in new trials. Note that in a 3-lever room, an optimal agent should produce both solutions within 3 attempts. One attempt may be used to identify the role of the observed levers in the abstract structure, and the remaining attempts are used for each solution.

Human Results

The analyses reported herein expand on previous behavioral findings by examining the number of attempts needed to find *each* solution rather than accumulating *all* solutions (see Human Data, Edmonds et al., 2018). The purpose of this exploration was to tease apart the separate learning components involved in the OpenLock task. Participants who failed to find all solutions in the allotted maximum number of attempts in *any* trial were removed from the analysis (24 participants removed from each condition). Eighty human participants were assigned to each condition (CC and CE).

We first examined whether the number of attempts needed to find each solution varied across trials. The behavioral data from each experimental condition is depicted in Figure 4. For participants who trained under a Common Cause (CC) schema, attempts needed to find the *first* solution decreased significantly following both the first trial ($t(55) = 6.80; p < .001$) and second trial ($t(55) = 2.52; p = .02$). First solution attempts also showed a marginal decrease following the fifth trial ($t(55) = 1.99; p = .051$). For the *second* solution, the number of attempts needed decreased significantly following the first trial only ($t(55) = 4.40; p < .001$). A similar trend was observed for participants assigned to the Common Effect (CE) condition—attempts needed to find the *first* solution decreased following the first trial ($t(55) = 5.30; p < .001$) and third trial ($t(55) = 2.19; p = .03$), and attempts needed to find the *second* solution decreased following the first trial only ($t(55) = 2.36; p = .02$).

The human results demonstrate that regardless of which causal schema participants trained with, significant learning appeared to occur in the early trials for both the *first* and *second* solution. However, the learning rate for the *first* solution was much faster, and the learning rate for the *second* solution was relatively less pronounced. In the next sections, we describe our computational approach and report whether it can account for human performance.

Model Details

We begin by describing our agent's process for combining top-down (abstract) causal knowledge with bottom-up (associative) attribute knowledge. The agent decides which action to perform by (i) computing the posterior probability of each candidate causal chain and (ii) making a selection using the computed posterior and a model-based planner.

Causal Theory Induction: To explain trends in human performance, we follow a Bayesian account of how hierarchical causal theories can be induced from data (Griffiths & Tenenbaum, 2005, 2009; Tenenbaum et al., 2006). The key insight in this framework is that hierarchy enables abstraction, and theories provide general background knowledge about a task or environment at the highest level. Theories consist of principles; for example, an analysis of evolutionary traits between species can be represented with a taxonomic tree and mutation processes (example from Tenenbaum et al. (2006)). Principles lead to structure; for example, a tree describing how primates evolved and split into species over time. Finally, structure leads to data; such as shared genes among primates.

The goal of this work is to model a human decision-making process where agents are required to learn *transferable* knowledge between different yet similar environments. We approach the problem from the perspective of *active* causal theory learning, where we expect an agent endowed with no information to learn the underlying abstract mechanics and commonalities between environments through interaction. This approach naturally places the focus of the learning task on how the agent decides the best action to take next and how to effectively integrate the results into the agent's model of the world.

In this work, we adhere to two general principles of learning: (i) *causal relations induce state changes in the environment, and non-causal relations do not* (referred to as our bottom-up β theory), and (ii) *causal structures that have previously been useful may be useful in the future* (referred to as our top-down γ theory). Specifically, the environment provides a set of attributes, such as position and color, and our agent learns which attributes are associated with levers that induce state changes in the environment. Our agent also learns a distribution over abstract causal structures (*i.e.*, schemas) that provide generalized notions of task structure.

We define a causal chain hypothesis space, Ω_C, over possible causal chains, $c \in \Omega_C$. Figure 3b shows the structure of the causal chain. Each chain is defined by a tuple of subchains $c = (c_0, \ldots, c_k)$, where each subchain is defined as a tuple $c_i = (a_i, s_i, cr_i^a, cr_i^s)$. Each a_i represents an action node that the agent can intervene on (execute), and the space of actions, Ω_A, consists of pushing and pulling on every lever and

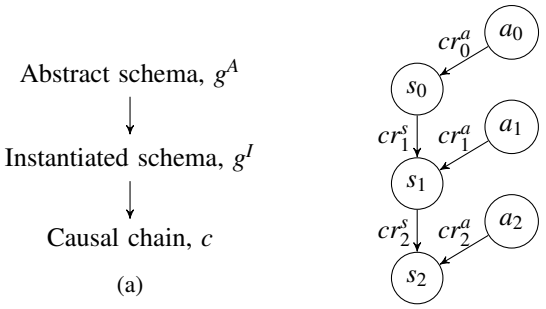


Figure 3: (a) An illustration of hierarchical structure of the model. A bottom-up associative learning theory, β, and a top-down causal theory, γ, serve as priors for the rest of the model. The model makes decisions at the causal chain resolution. (b) Atomic causal chain. The chain is composed by a set of sub-chains, c_i, where each c_i is defined by: (i) a_i, an action node that can be intervened upon by the agent, (ii) s_i, a state node capturing the time-invariant *attributes* and time-varying *fluents* of the object, (iii) cr_i^a, the causal relation between a_i and s_i, and (iv) cr_i^s, the causal relation between s_i and s_{i-1}.

pushing on the door. Each s_i represents a state node. The state node is defined as a tuple, $s_i = (\phi_i, f_i)$, where ϕ_i is a vector of time-invariant *attributes* and f_i is a vector of time-varying *fluents*. The state node is influenced by taking action a_i according to the causal relation cr_i^a and may be affected by a previous state node through the causal relation cr_i^s. For instance, in Figure 1a and Figure 3b, the action *push* for the leftmost lever may transition the lever from the fluent *pulled* to *pushed* through cr_0^a, which in turn transitions the uppermost lever from *locked* to *unlocked* according to cr_1^s.

The space of attributes is denoted as Ω_ϕ, consisting of position and color. The space of fluents, Ω_F, consists of binary values for lever status (*pushed* or *pulled*) and lever lock status (*locked* or *unlocked*). The space of states is defined as $\Omega_S = \Omega_\phi \times \Omega_F$. The space of causal relations is defined as $\Omega_{CR} = \Omega_F \times \Omega_F$, capturing the possibly binary transitions between previous fluent values and the next fluent values.

We assume agents can directly intervene on (*i.e.*, control) *actions*, but cannot directly intervene on *fluents*. This distinction adds significantly more complexity to the causal chain hypothesis space but means that we do not assume the effects of actions, nor do we assume an agent can directly intervene on the value of a particular fluent. We assume that an agent can execute any action within the action space (through an intervention on the action node in the causal chain), but how that action affects the state of the world must be learned (*i.e.*, the effects of the actions are learned).

Decomposing states into time-invariant *attributes* and time-varying *fluents* aids in the computational complexity of learning and inference; our agent assumes attributes cannot be changed by actions or other states. In addition, because the attributes are time-invariant, attributes offer a grounding upon which the agent can learn knowledge, regardless of the executed action sequence or lever configuration. In contrast, the fluents are time-varying and include the latent state of the lever's internal locking mechanism; *i.e.*, *locked* or *unlocked*. The agent learns how to influence these latent states through observational cues about which attributes are associated with a particular fluent. Attributes are defined by low-level features of an object, *e.g.*, position, color, shape, orientation, *etc.*. These low-level attributes provide general background knowledge about how specific objects change under certain actions (for instance, which levers can be pushed or pulled).

A background theory encodes general knowledge that can be used to induce or evaluate a structural representation. We use two background theories—one for bottom-up features, denoted β, to learn beliefs about which attributes of objects indicate the object can be interacted with to produce a causal effect. This low-level knowledge about object attributes and their propensity to be involved in causal relationships provides information to transfer between similar but different environments governed by common underlying dynamics. The second background theory provides a top-down abstraction, denoted γ, that assumes tasks have similar causal structure across slightly different environments; *i.e.*, changes in the observable environment do not alter the underlying causal structure of a task.

Attribute Learning: Attributes provide time-invariant properties of an object. Categories of objects often share common attributes; *e.g.*, all cups share a common shape, all stop signs are red, *etc.*. However, objects in a category may vary in their physical form but share common functionality; for instance, light switches come in a number of shapes and sizes, but all examples share a common mechanism to transit between states.

We learn which attributes are relevant to our causal hypotheses via a Bayesian learning process, based on our assumption that causal relationships induce state changes. Therefore, an object changing states under an action indicates that the object's attributes may be related to a causal relationship. These attributes provide generalization clues for the agent, such as insights into which low-level attributes indicate that the corresponding object is part of a solution. This knowledge is invariant across trials and causal schemas.

The agent's belief in an attribute being causal is modelled with a multinomial distribution $\text{Mult}(\theta)$ parameterized by θ. The posterior distribution of θ given observed data $\mathbf{X}$ and the bottom-up theory β follows a Dirichlet distribution: $p(\theta|\mathbf{X};\beta) = \text{Dir}(\alpha')$, where α' is given by a maximum a posteriori (MAP).

Attributes are learned in two different time scales: a global timescale to learn attributes across all trials (between trials) and a local timescale to learn attributes specific to this trial (within trials). This separation allows the agent to adapt quickly to trial-specific knowledge while maintaining a global understanding across all trials. In each timescale, we perform this attribute learning in the following steps: (i) draw a sample (produce an observation by selecting an intervention and observing the result), (ii) accept the sample if the environment changed state in any way (*i.e.*, there was an effect from the intervention), and (iii) increase α of each attribute's Dirichlet distribution according to observed outcome.

A Dirichlet distribution, $\text{Dir}(\alpha^G)$, is used to model the posterior of the global attribute distribution. After finishing a

trial, the agent's global Dirichlet parameters, α^G, are updated to incorporate the observed data within a trial.

For each trial, we initialize the parameters of the local attribute Dirichlet distribution, $\text{Dir}(\alpha^L)$, with a scaled sample from the global Dirichlet, $\alpha^L = k\theta$, where $\theta \sim \text{Dir}(\alpha^G)$. This scaling factor k reduces the variance and enables fast adaptation of the agent's local attribute beliefs. In our experiments, we set k to initialize the local Dirichlet to have $\alpha^L \in [1, 10]$.

We introduce an additional variable, ρ to represent a casual event according to our background theory β; *i.e.*, that causal events induce state changes in the environment. We use a local prior over attributes as our bottom-up associative learning theory. We compute the likelihood that the attributes of a particular chain c are causally relevant given a background theory β as:

$$p(\rho|c;\beta) = \prod_{c_i \in c} p(\rho_i|c_i;\beta), \tag{1}$$

where $p(\rho_i|c_i;\beta)$ is computed as

$$p(\rho_i|c_i;\beta) \propto \prod_{\substack{\phi_{ij} \in s_i \\ s_i \in c_i}} p(\rho_i|\phi_{ij};\beta) \tag{2}$$

where ϕ_{ij} is the j-th attribute from the i-th subchain. The term $p(\rho_i|\phi_{ij};\beta)$ represents the probability that attribute ϕ_{ij} adheres to the background theory β. Here, β represents the probability that attribute ϕ_{ij} is associated with objects that induce state changes. Note that $p(\rho_i|\phi_{ij};\beta)$ is parameterized by a sample from the local attribute Dirichlet distribution. After finishing an attempt, we update the parameters α^L of the local distribution to incorporate the outcome of the attempt and resample θ.

Recall our associative theory: causal relationships induce state changes in the environment; practically, $p(\rho_i|\phi_{ij};\beta)$ represents the probability that attribute ϕ_{ij} is associated with objects that produce state changes, under the assumption these attributes are independently associated with causal events. In our domain, an agent using this theory should learn that grey levers are involved in causal events and white levers are not. Additionally, the agent should initially believe that position is an important attribute for detecting causal relationships. However, as the agent observes multiple configurations of levers with different positions of grey levers, every position will be involved in causal events, and therefore this belief should approach the uniform distribution.

This bottom-up inference enables agents to leverage low-level associative information about causal relationships. We then transfer this belief between trials, thereby enabling our agent to leverage the knowledge acquired in one trial to transfer to the next trial. The agent updates its belief regarding which attributes it believes are causal after each attempt.

Abstract Schema Learning: Learning attributes that correspond to causal cues is critical for an agent expected to learn how an environment operates. However, many environments share common high-level abstract causal structures. For instance, switches come in all different shapes and sizes tailored to specific tasks—from a light switch to a circuit breaker to a railroad switch. Each of these domain-specific mechanisms share a common abstract functionality—changing the state of some object from one discrete state to another.

We propose a model to learn abstract structural models that can be used to instantiate domain-specific models to achieve a task in an environment. This abstract knowledge is assumed to be useful across domains, and agents may acquire a collection of useful abstract models of different functionality. Our model considers learning abstract knowledge as a form of model selection, where the agent hypothesizes a space of potential abstract structures and updates the beliefs in those abstract structures based on its experience in the environment.

More specifically, we consider an abstract causal schema, g^A, from a hypothesis space of abstract schemas, Ω_{G^A}, to be a structural description of some causal relationships (see Figure 2). The space Ω_{G^A} is enumerated in this work; *i.e.*, all possible structural combinations of $N = 2$ trajectories (*i.e.*, causal chains) with length $K = 3$ are considered (since there are two solutions and three actions per attempt). We introduce a prior over abstract schemas, $p(g^A;\gamma)$, that is a multinomial distribution parameterized using a sample from the abstract schema Dirichlet distribution, $\text{Dir}(\alpha^A)$. After completing a trial, the abstract schema that encodes the solutions found in this trial receives a parameter update in the Dirichlet distribution—*i.e.*, an increase to the solution abstract schema's α^A.

These abstract structures are not bound to any particular instantiation of attributes, states, or actions. Instead, they encode common structural properties under varying instantiations—knowledge that may be useful when an observational setting is changed. In our task, abstract schemas encode the abstract structures, some of which are useful for solving OpenLock (*i.e.*, CC or CE), and we should expect agents to have a biased prior towards these structures.

Next, we consider an instantiated schema, g^I, to be a composition of causal chains, $c \in \Omega_C$. Instantiated schemas share the same structure as abstract schemas, but contain specific assignments for each a_i, s_i, cr_i^a, and cr_i^s of each subchain in the schema. We compute the belief in an instantiated schema g^I according to the hierarchical structure in Figure 3a:

$$p(g^I|do(q);\gamma) = \sum_{g^A \in \Omega_{G^A}} p(g^I|g^A, do(q))p(g^A;\gamma), \tag{3}$$

where $do(q)$ represents an intervention where the agent performs q—the solutions found thus far, a set of action sequences $q = \{A_0, A_1, \ldots, A_n\}$, where A_i is an action sequence. The $do()$ operator is the intervention operation presented by Pearl (2009), which allows the agent to bias its top-down inference towards instantiated schemas that contain solutions already found. Next, we compute the top-down belief in a causal chain by summing over instantiated schemas that contain the chain:

$$p(c|do(q);\gamma) = \sum_{g^I \in \Omega_{G^I}} p(c|g^I, do(q))p(g^I|do(q);\gamma). \tag{4}$$

These terms enable top-down inference on which chain is most likely to adhere to instantiated schemas that reflect abstract causal structures that have been useful in the past.

Learning which abstract schemas were successful in previous trials can be leveraged when the agent faces a new room configuration with the same underlying abstract mechanism governing the lock.

Intervention Selection: We formulate our intervention selection as a combination of the top-down and bottom-up causal chain beliefs, and we consider our learning mechanisms, γ and β, to be independent. We compute the posterior of the chain based on our top-down belief and bottom-up likelihood, assuming a uniform prior $p(\rho)$:

$$p(c|\rho, do(q); \gamma, \beta) \propto p(c|do(q); \gamma) p(\rho|c; \beta). \quad (5)$$

Our agent maintains an explicit notion of the goal of the task—to open the door. Human participants were also told the precise goal of the task. Thus, we frame our intervention selection process as a form of model-based planning. Our agent seeks to infer the causal chain most likely to achieve the goal—opening the door—given the agent's current model of the environment. The agent's model of the environment comes from two forms of learning: bottom-up associative attribute learning and top-down abstract schema learning.

We define a target goal of our planner as a particular state of the environment, denoted s^*. Given a target goal our agent models its current state as a tuple of (n, q), where n represents the number of solutions remaining, and q the set of solutions already executed. The agent seeks to execute a causal chain c in the hopes of transitioning n to $n-1$. The agent replans after every attempt until it finds all solutions the room; *i.e.*, when $n = 0$. Thus, our final planning objective at time t is to pick the causal chain with the maximal posterior subject to the constraints that the chain contains the target goal state s^* (*i.e.*, the door being *pushed*) and is not in the agent's set of solutions executed q:

$$c_t^* = \arg\max_{c \in \Omega_C} p(c|\rho, do(q); \gamma, \beta) \quad \text{s.t. } s^* \in c \wedge c \notin q, \quad (6)$$

where $p(c|\rho, do(q); \gamma, \beta)$ is defined in Equation 5. This state definition matches information provided to human participants and places the focus of our planner on achieving task-level goals.

Among the chains that satisfy the constraints, we rely on our chain posterior to capture which chains are causally plausible. The posterior combines the top-down structural knowledge with the bottom-up attribute knowledge. This combination is powerful for two reasons: (i) bottom-up knowledge biases beliefs towards structures that contain attributes that have been present in causal events in the past, and (ii) top-down knowledge allows the agent to bias beliefs towards structures that have been useful in the past.

Model Results

We train our agent in the same fashion as humans; specifically, we allow the agent to complete 80 trials in CC and CE escape rooms (same number as human participants). The agent is also limited to 3 actions in an attempt and 30 attempts within a trial. Any agent that did not complete all trials was removed from the study (same as human participant data—no agents were removed from the CC condition; 7 agents were removed from the CE condition).

Figure 4 compares human and model performance. The model shows a similar trend as humans but with slightly worse performance in each trial[2]. For the agent assigned to the CC condition, the number of attempts needed to find the *first* solution decreased significantly following the first trial ($t(79) = 8.09; p < .001$) and second trial ($t(79) = 4.04; p < .001$). The CE agent required less attempts to find the *first* solution following the first trial only ($t(72) = 6.23; p < .001$). Decreases in first and second solution attempts were not significant between the remaining trials.

These results demonstrate that our model is roughly capable of capturing learning rates of human participants but does not capture all significant changes in the number of attempts needed: *e.g.*, in both the CC and CE conditions, the number of attempts needed by participants to find the *second* solution consistently decreased following the first trial. However, our model overall effectively captures general trends in human performance: the number of attempts needed to find *all* solutions matches well to humans and decreases near-monotonically, albeit at a lesser rate.

[2]Example solution executions for human participants and the model can be viewed at `https://vimeo.com/334518941`

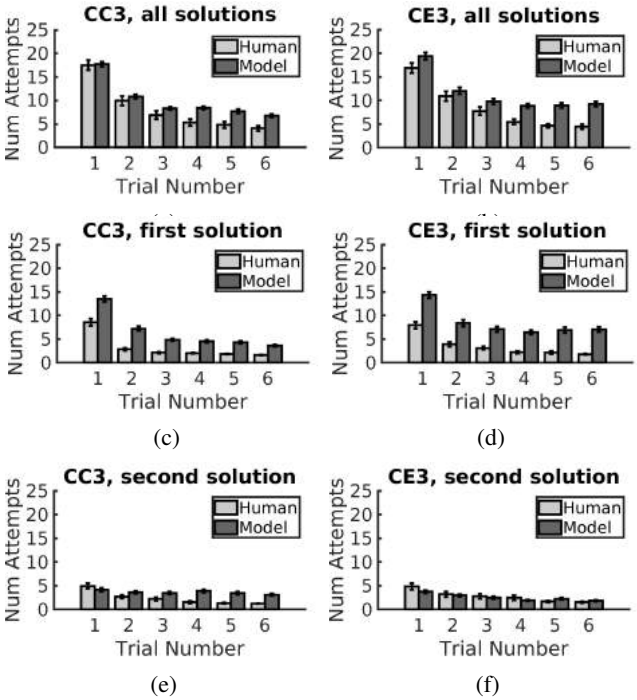


Figure 4: Comparison of human and model results for the common-cause CC3 condition and the common-effect CE3 condition. (a) and (b) compare the total number of attempts to find all solutions; (c) and (d) compare the number of attempts to find the *first* solution; (e) and (f) compare the number of attempts to find the *second* solution.

Conclusion

In this work, we showcase a hierarchical model based on associative learning and schema reasoning. Our model integrates two learning mechanisms: (i) a bottom-up theory that learns which attributes have causal associations in the environment, and (ii) a top-down theory that learns useful abstract structures in the environment. Our agent chooses an intervention based on the posterior of causal chains and updates its model using the observed outcome of the intervention. Model results show that our hybrid agent is able to capture general trends observed in human participants and captures some of the statistical significance observed in human performance. These results suggest that human causal learning may consist of a mechanism that combines bottom-up associative learning with top-down reasoning about causal structure.

The underlying computational framework presented here is broadly applicable outside of the OpenLock environment; it can be applied to any reinforcement learning environment where: (i) underlying dynamics are constrained by some causal structure; (ii) interactive elements have observable features which signal causal relevance; and (iii) physical locations of key elements change over time. In the future, we hope to expand our model to account for more extreme observational changes. For example, what if levers could suddenly be rotated instead of pushed/pulled? What if new colors were introduced which provided further cues about causal relevance? And what if the environment began operating in a probabilistic fashion where levers may fail to actuate properly? Future behavioral and computational work should examine how these processes integrate in more complex scenarios that provide a closer approximation to the real world.

Discussion

What other theories may be useful for learning causal relationships? The background theories presented here—namely that causal relationships induce state changes and abstract causal knowledge can be reused—provide reasonable background theories. However, other background theories may also be appealing. For instance, Pearl (2009) defines a stricter definition of causal relations based on whether or not a causal relation is *identifiable* in a directed acyclic graph.

How can hypothesis space enumeration be avoided? The spaces of Ω_{g^A} and Ω_{g^I} are enumerated in this work. Hypothesis space enumeration can quickly become intractable as problems increase in size. While this work used a fixed, fully enumerated hypothesis space, future work will include examining how sampling-based approaches to iterative generate causal hypotheses (*e.g.*, see Bramley et al. (2017)).

What are the other possibilities of bottom-up associative criteria? Our method treats low-level attributes as the criteria for our bottom-up associative learning. However, other possibilities are equally valid. For instance, a modeler could pair attributes with specific actions and learn distributions of causal effects over this pairing. This decision ultimately comes down to the resolution of the problem being considered and what is appropriate to correctly model the problem.

How is this work connected to reinforcement learning (RL)? The model-based planner is closely related to model-based RL. Our problem setting could be cast in terms of a 0-1 reward function—the agent receives a reward of 1 if the door is opened, and 0 otherwise. However, model-based RL typically assumes a world model is provided, but our agent iteratively updates its conception of world dynamics through associative learning and schema reasoning.

Acknowledgement

The authors thank Prof. Tao Gao, Prof. Ying Nian Wu, Feng Gao, and Chi Zhang for their helpful discussions. The work reported herein was supported by DARPA XAI N66001-17-2-4029, ONR MURI N00014-16-1-2007, NSF BSC-1655300, and an NSF Graduate Research Fellowship.

References

Bramley, N. R., Dayan, P., Griffiths, T. L., & Lagnado, D. A. (2017). Formalizing Neuraths ship: Approximate algorithms for online causal learning. *Psychological Review*, *124*(3), 301.

Bramley, N. R., Gerstenberg, T., Tenenbaum, J. B., & Gureckis, T. M. (2018). Intuitive experimentation in the physical world. *Cognitive Psychology*, *105*, 9–38.

Bramley, N. R., Lagnado, D. A., & Speekenbrink, M. (2015). Conservative forgetful scholars: How people learn causal structure through sequences of interventions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(3), 708.

Cheng, P. W. (1997). From covariation to causation: a causal power theory. *Psychological Review*, *104*(2), 367.

Edmonds, M., Kubricht, F., James, Summers, C., Zhu, Y., Rothrock, B., Zhu, S.-C., & Lu, H. (2018). Human causal transfer: Challenges for deep reinforcement learning. *Proceedings of the 40th Annual Meeting of the Cognitive Science Society*.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*(4), 334-384.

Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological Review*, *116*(4), 661-716.

Heider, F. (1958). *The psychology of interpersonal relations*. Psychology Press.

Holyoak, K., & Cheng, P. W. (2011). Causal learning and inference as a rational process: The new synthesis. *Annual Review of Psychology*, *62*, 135–163.

Mackie, J. L. (1974). *The cement of the universe: A study of causation*. Oxford.

Pearl, J. (2009). *Causality*. Cambridge University Press.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical conditioning II: Current research and theory*, *2*, 64–99.

Shanks, D. R., & Dickinson, A. (1988). Associative accounts of causality judgment. *Psychology of learning and motivation*, *21*, 229–261.

Stahl, A. E., & Feigenson, L. (2015). Observing the unexpected enhances infants' learning and exploration. *Science*, *348*(6230), 91–94.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, *10*(7), 309–318.

Waldmann, M. R., & Holyoak, K. J. (1992). Predictive and diagnostic learning within causal models: Asymmetries in cue competition. *Journal of Experimental Psychology: General*, *121*(2), 222–236.

Moral Reasoning with Multiple Effects: Justification and Moral Responsibility for Side Effects

Neele Engelmann (neele.engelmann@uni-goettingen.de)
Michael R. Waldmann (michael.waldmann@bio.uni-goettingen.de)
Department of Psychology, University of Göttingen, Germany

Abstract

Many actions have both an intended primary effect and unintended, but foreseen side effects. In two experiments we investigated how people morally evaluate such situations. While a negative side effect was held constant across conditions in Experiment 1, we varied features of the positive primary effect. We found that judgments of moral justification of actions were sensitive to the numerical ratios of helped versus harmed entities as well as to the kind of state change that was induced by an agent's action (saving entities from harm versus improving their status quo). Judgments of moral responsibility for side effects were only sensitive to the latter manipulation. In Experiment 2, we found initial support for a subjective utilitarian explanation of the moral justification judgments.

Keywords: Moral Reasoning, Causal Reasoning

Introduction

Research on moral judgments often probes people's intuitions about moral dilemmas. One of the most famous and well-studied dilemmas is the so-called trolley problem (Foot, 1967). In the side effect variant of trolley dilemmas, agents have a choice between letting a runaway trolley kill several people or an action that redirects the trolley to a different track where it would kill fewer people. The primary question in these studies is typically whether it is morally permissible to act. Many factors have been identified that influence people's intuition about this question (for an overview see Waldmann, Nagel, & Wiegmann, 2012).

The two dominant normative ethical approaches, utilitarianism and nonconsequentialism, largely agree in this situation. According to utilitarian recommendations, the action should be performed whenever its positive consequences outweigh the negative effects. Nonconsequentialist theories, such as the *Doctrine of Double Effect* (DDE, see Mikhail, 2011), arrive at similar conclusions for this case. The focus of the DDE and nonconsequentialism in general lies on the causal structure mediating acts and outcomes. In the side effect variant of the trolley dilemma, acting is considered permissible because the negative effect is not an intended means, but merely a foreseen side effect, and is not out of proportion to the positive effect. Psychological research on the side effect dilemma has shown that subjects indeed take the alternative outcomes into account when assessing the action's permissibility (e.g., Mikhail, 2011; Cohen & Ahn, 2016).

Evaluating Actions and their Side Effects

The focus of research on trolley dilemmas is on how people evaluate the permissibility of an action that causes two outcomes. All theories assume that in the side effect dilemma, both outcomes are compared and affect the moral evaluation, but little is known about the functional form of this comparison. A typical claim is that harming is permissible if the good outweighs the bad, but it is unclear whether this decision is just based on a simple categorical decision about which value is larger, or whether gradual differences between outcome values affect the decision. Few studies have systematically manipulated the numbers of victims that are saved or harmed in moral dilemmas (but see Cohen & Ahn, 2016; Waldmann & Wiegmann, 2012).

Cohen and Ahn (2016) postulate a subjective utilitarian analysis. For each item or set of items (e.g., 5 people) subjects provided an estimate of their personal value. The personal values were affected by the type of item and their number, although the number turned out to have a relatively small effect. These estimates of the personal values were then used to predict subjects' judgments about choice situations in which one set of items is about to be destroyed (or killed) when no action is taken but saved when the agent acts, which in turn would destroy (kill) a second set of items. According to the categorical utilitarian decision strategy, the action is chosen that saves items with the higher personal value. The model also predicts reaction times: Given that the comparison is typically influenced by uncertainty, a faster reaction time is predicted when the difference between values becomes larger.

One key goal of our project is to provide further tests of the subjective utilitarian model. A salient problem of the current version of the model is that it lacks generality. Its predictions are based on the personal values of the items involved in the outcomes but this model neglects that actions cause transitions between states. An evaluation of an action thus needs to take into account the values of the states of the items in the presence versus the absence of the action. Cohen and Ahn (2016) did not consider how subjects assess the personal values of the items in their destroyed or dead states, probably because this was the standard state in the absence of an action across all item sets. However, actions can also improve the state of items that otherwise would be in a normal state, or they could be saved from a disease that would harm, but not kill them. To provide a full utilitarian account of how outcomes of actions should be evaluated we suggest that people compute contrasts between the personal values of the outcomes in the presence versus the absence of the target action. We will also argue that sometimes more than two states need to be considered. We will present an experiment that presents a wider range of actions, which allows us to test our subjective utilitarian model against theories that are not sensitive to different types of states in the presence and absence of the

target action.

A further focus of our study is to investigate how the relation between the number of people that are positively or negatively affected by the action influences the degree to which people find the action morally justifiable and the agent morally responsible for the outcomes, especially the negative side effect. We systematically manipulated the numbers involving the positive primary effect while holding the negative side effect constant (see also Waldmann & Wiegmann, 2012, for a similar design but different tasks). For example, in one of our experimental conditions, ten members of a tribe are harmed by an action that would save a varying number of members of a different tribe. According to Cohen and Ahn's (2016) model, an act involving a negative side effect should lead to faster reaction times the more entities are helped compared to harmed. If reaction times indicate certainty about an act's permissibility, one can also derive from this theory the prediction that justification ratings should be affected in a similar manner.

One limitation of trolley studies is that so far they have focused on a particular type of situation in which the primary goal is to save victims that otherwise would be killed. It may well be that acts that lead to negative side effects are only considered justified when the primary effect targets entities that, prior to the intervention, are threatened to be harmed. The primary effect may be less effective as a justification when the act is supererogatory and just improves the states of entities that prior to the act are in a normal state. For example, instead of saving varying numbers of victims from grave harm, the people may be fine prior to the act, with the act just improving their health and living conditions. The theory proposed by Cohen and Ahn (2016) does not make predictions here because it only takes into account the personal values of the entities in their intact state. We will in Experiment 2 test a modified account that postulates that subjects take into account personal values of states in both the presence and the absence of an action. This account makes predictions for the difference between saving entities or improving their states.

Another limitation of the typical trolley dilemma studies is that they have focused on situations in which saving and harming are causally achieved by redirecting a harmful entity (the runaway trolley). In order to widen the range of studied dilemmas and to be able to manipulate the prior state of the entities involved in the primary goal, we tested a different causal structure in which a helpful act rather than a threat was redirected (see also Ritov & Baron, 1999; Bartels & Medin, 2007). For example, in the condition involving two tribes, a dam may be opened that redirects water from one tribe to the other. Redirecting might save tribe members from a negative state or improve their normal situation.

Finally, a limitation of previous research is that the test question typically focuses only on the act leading to two outcomes. We are also interested in how people evaluate the two outcomes individually. We therefore added as test questions requests to judge moral responsibility for the negative side effect. Our goal was to test whether these judgments are also influenced by the value of the primary effect (e.g., number of victims). If subjects just focus on the side effect, the primary effect should not have an influence. However, if the status quo or the number of affected entities are used as exonerating factors, their impact should also be seen in moral responsibility ratings for the side effect.

Together, these manipulations and the studied judgments widen the focus of previous work on people's moral intuitions about cases with multiple effects. The aim of the first experiment was to test whether the relation between primary and side effect of an action influences moral justification assessments. Moreover we were interested in whether the primary effect influences moral responsibility assessments for a bad side effect. We tested whether these two types of moral queries are affected by the kind and number of entities that are potentially harmed or saved, and by their state change due to a possible intervention. Experiment 2 inquires to what extent the results of Experiment 1 can be explained by a subjective-utilitarian framework.

Experiment 1

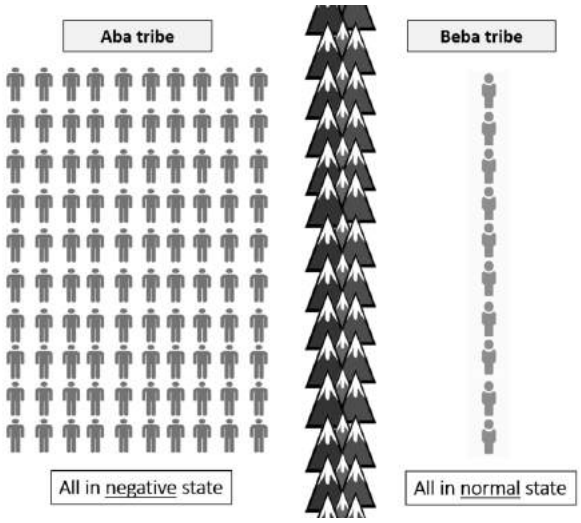
We constructed three scenarios in which an agent decides to perform an action with a positive, intended primary effect and a negative, unintended (but foreseen) side effect. The negative side effect was held constant across conditions and always consisted in killing 10 entities (people, animals, or plants). We varied whether 1, 5, 20 or 100 entities benefitted from the action. Furthermore, we manipulated whether these entities were in a negative or a neutral state prior to the action. In the situations in which the entities were in a negative state, they would have died without the agent's action; in the contrasted normal state condition, the action would merely cause additional benefit (e.g., people improving their living conditions or plants growing better).

Design, Material and Procedure[1] 450 participants were recruited via the UK based platform Prolific Academics for a compensation of £0.25 (£6 per hour). Inclusion criteria were a minimum age of 18 years, English as a first language, a study approval rate on the platform of at least 90%, and not having participated in previous studies with similar material. Participants were randomly allocated to one of 24 conditions (primary effect: saving vs. improving; number of helped entities: 1 vs. 5 vs. 20 vs. 100; affected entities: people vs. animals vs. plants). Here is an example vignette from the *saving* conditions. The example describes a condition in which 100 people are saved by the action, who otherwise would die:

Suzy is the prime minister of Tolosia, a mountainous country with many distant and small villages. The villages are populated by different indigenous tribes. She is authorised to make all decisions about the country's welfare that she deems appropriate. One day, she learns that a mountain village has

[1]The full material and data for both experiments are available under `https://osf.io/jcux6/`

suffered from an ongoing drought that left its inhabitants, the Aba tribe, in poor health due to lack of water. Exactly 100 people belong to the Aba tribe, all of whom are in critical condition and will die if nothing is done. Suzy could order to open a dam that would redirect a mountain river towards the Aba tribe. With a quick water supply, the 100 members of the Aba tribe could recover. However, the redirection of the river could also cause a lack of water in another mountain village, home to the Beba tribe, causing its 10 members to die of thirst within a few days. All of the 10 members of the Beba tribe are fine at the moment. Since both mountain villages are inaccessible to any means of transport, redirecting the river is the only currently available measure to influence the well-being of the two tribes. Here is a schematic representation of the two tribes and the current state of their members:



Suzy is aware of all the facts. She wants the 100 members of the Aba tribe to recover, but also not to cause any harm to the 10 members of the Beba tribe. She decides to open the dam and redirect the mountain river. All of the 100 members of the Aba tribe recover. However, all of the 10 members of the Beba tribe die within a few days.

The figure was followed by the instruction: "Here is a schematic representation of the tribes and their state after the river has been redirected" along with the same figure as above in which the lower labels now read "all in normal state" for the Aba tribe and "all dead" for the Beba tribe. In the corresponding *improving* condition, the vignette stated that the Aba tribe could vastly improve their health and lifespan with an extra water supply (no threat by a drought was mentioned). In the subsequent test phase participants were asked to rate the extent to which they saw the agent's action as morally justified ("To what extent was Suzy's action morally justified ?"). The moral responsibility question focused on the side effect ("To what extent is Suzy morally responsible for the members of the Beba tribe dying?"). As a control, we also asked about the primary goal ("To what extent is Suzy morally responsible for the members of the Aba tribe improving their health?"). Ratings were given on a 10-point Likert scale with the endpoints labelled "not at all" (1) and "fully" (10). Justification and responsibility questions were presented on two separate pages, with page order counterbalanced between participants; order of the two responsibility questions within the respective page was randomized. Subsequently, two manipulation check questions assessed whether people had correctly understood how many entities were harmed and helped in the scenario.

Results and Discussion 18 participants were excluded for failing at least one of the manipulation check questions, leaving data of 432 participants for the analysis (mean age = 34.4, SD = 11.93). We conducted a 2 (primary effect) x 3 (entity) x 4 (numbers) x 2 (test question order) ANOVA for each of the three dependent variables. Since our study is partly exploratory, we used a conservative significance threshold that takes into account the number of tests in the models (here: $p<.003$). Results for the 432 valid subjects can be seen in Figure 1.

Moral justification ratings were higher the more entities were helped compared to harmed, $F_{(3,\,384)} = 8.81$, $p<.001$, $\eta^2 = .06$. Additionally, a large effect was obtained between the conditions saving and improving, $F_{(1,384)} = 130.74$, $p<.001$, $\eta^2 = .25$. The interaction was not significant (p=.37). Participants gave the highest justification ratings when the primary effect was an instance of saving and more entities were saved than killed.

Post hoc tests (Newman-Keuls) for the saving condition revealed that the case in which only one entity was saved as a primary effect was judged significantly less morally justified than the cases in which twenty or a hundred entities were saved. The other cases did not differ significantly from each other. In the *improving* condition, post hoc tests showed no significant differences.

There was also a main effect of vignette. Subjects considered the action as most morally justified when the affected entities were plants (M = 5.23, SD = 2.6), followed by animals (M = 4.41, SD = 2.52), and people (M = 3.84, SD = 2.77), $F_{(2,384)} = 14.39$, $p<.001$, $\eta^2 = .07$. A possible reason for this ordering might be that harming people may be seen as a harsher moral violation than harming plants and therefore less justifiable by good effects. Animals seem to be in the middle.

Additionally, a small unexpected order effect was found. Ratings were slightly higher when the moral justification question was presented after the moral responsibility questions (M = 4.88, SD = 2.71) compared to before (M = 4.12, SD = 2.62), $F_{(1,384)} = 12.51$, $p<.001$, $\eta^2 = .03$.

Moral responsibility ratings for the negative side effect were generally high, but not detectably influenced by the number of helped entities, $F_{(3,384)} = 0.35$, $p = .79$ (see Fig. 1). However, the ratings were lower when the action's primary effect was an instance of saving (M = 8.09, SD = 2.23) rather than improving (M = 9.12, SD = 1.59), $F_{(1,384)} = 33.51$, $p<.001$, $\eta^2 = .08$. The interaction was not significant (p=.61). *Moral responsibility ratings* for the positive primary effect were

high (M = 8.23, SD = 2.31) and not influenced by any manipulation.

In sum, the moral justification ratings of the action were sensitive to the relation between the primary and the side effect. The more entities were helped as a primary effect, the more justified the action was judged. This pattern shows that moral justification is a continuous quantity that is sensitive to the relative size of the outcomes. A novel result concerns the comparison between different status quos, which generated the largest effect. If entities are saved from a threat, the action was seen as substantially more justified than when the primary goal is just to improve states starting from a neutral state.

The fact that subjects took into account both the primary and the side effect in their justification judgments is predicted by both nonconsequentialist and utilitarian accounts. However, the specific theory proposed by Cohen and Ahn (2016) does not predict the largest effect in our experiment: Subjects clearly differentiated between saving entities versus improving their state. Simply using assessments of personal values of the entities does not predict these effects without taking into account the personal values of the states of the entities in the absence of the action. We will test a modified model that is sensitive to state changes in Experiment 2.

An interesting unexpected finding was that moral responsibility ratings proved insensitive to the number of helped entities, but were reduced when the action's primary effect was an instance of saving rather than improving. This latter effect makes it unlikely that the lack of an effect of number is due to a ceiling effect. A possible interpretation of this pattern may be that subjects tried to focus on the side effect alone but were influenced by features of the primary effect that have a large impact on justification, such as the status quo, rather than only a small effect, such as the numbers.[2]

Experiment 2

The aim of the second experiment is to investigate to what extent the effects observed in Experiment 1 could be explained by a a variant of a subjective utilitarian theory that in crucial aspects differs from the one proposed by Cohen and Ahn (2016). Cohen and Ahn (2016) modeled choices as decisions based on the personal values of the entities involved in the alternative outcomes. For example, the task in their second study was to choose which of two sets of items should be saved and which destroyed in a dilemma. The model claims that the differences between the personal values of the two sets of items predict judgments. The focus on the personal values of the items seems appropriate here because all actions represented a choice between leaving the items intact or destroying (or killing) them. This restriction of the task allowed Cohen and Ahn (2016) to focus on the personal values of the affected items. However, the model is a too restrictive as a general model of moral reasoning. We suggest that the focus should be on actions, which can cause transitions between various states, not only between the states dead and alive or intact and destroyed. For example, in our Experiment 1 we presented cases in which actions improved states of entities that prior to the intervention were in a normal state.

To overcome the limitations of the model proposed by Cohen and Ahn (2016), we here propose a variant of a subjective utilitarian theory that focuses on actions and models them as state changes. When people evaluate an action, they should be sensitive to both the outcomes in the presence of the action but also to what happens in the absence of the action. For example, an action that improves the state of an entity can be represented as the difference between the personal values of the improved state and the normal state prior to the action. More complex state transitions are conceivable, and in fact in Experiment 1 we presented scenarios in which the entities shifted between four possible states (normal, threatened, improved, dead). In the present study we collected assessments of personal values of all the entities for these four states and used these assessments to predict the justification judgments obtained in Experiment 1.

Figure 2 shows how we adapted our model to the cover stories in Experiment 1. In the example in Figure 2, 100 people are under the threat of dying prior to any action. In the absence of an action (i.e., omission) they would die, which is modeled here as the contrast of the personal values between death and a critical state (second component of Figure 2a). In the presence of the action, the people in critical state would be shifted into a normal, healthy state, here represented as the difference between the personal values of a critical versus a normal state (first component of Figure 2a). The overall utility of saving the people is modeled as the sum of these contrasts because the action both prevents the people from being killed and puts them from a critical into a healthy state. Thus, the representation of the saving action considers both the effects of the potential action and of its omission. In the case of improving (not depicted), the model simplifies to a contrast between the values of the improved versus the normal states. The second component in the equation in Figure 2a would amount to 0 in this case because there is no threat to the normal state. Finally, Figure 2b shows how we model the total utility of the action in a scenario with multiple effects: It is the sum of the median utilities of the primary effect (saving) and the harmful side effect (killing 10 people).

Design, Material and Procedure The design of our basic value estimation task largely follows the methodology described in Cohen and Ahn (2016) but assesses a wider range of possible states of entities. Like Cohen and Ahn (2016), we tested the influence of the numbers of entities (1 vs. 5 vs. 10 vs. 20 vs. 100) on personal value assessments in

[2] In this experiment, moral justification was assessed globally (i.e., for a whole action), while responsibility was assessed separately for the single effects. One might worry that this does not allow us to tell whether the differences between the two judgments are driven by the type of judgment or by the focus of the question on global or separate outcomes. We therefore conducted a follow-up study in which we fully crossed these two factors. We found that the type of judgment seems to be the driving factor. The study is available online along with materials and data.

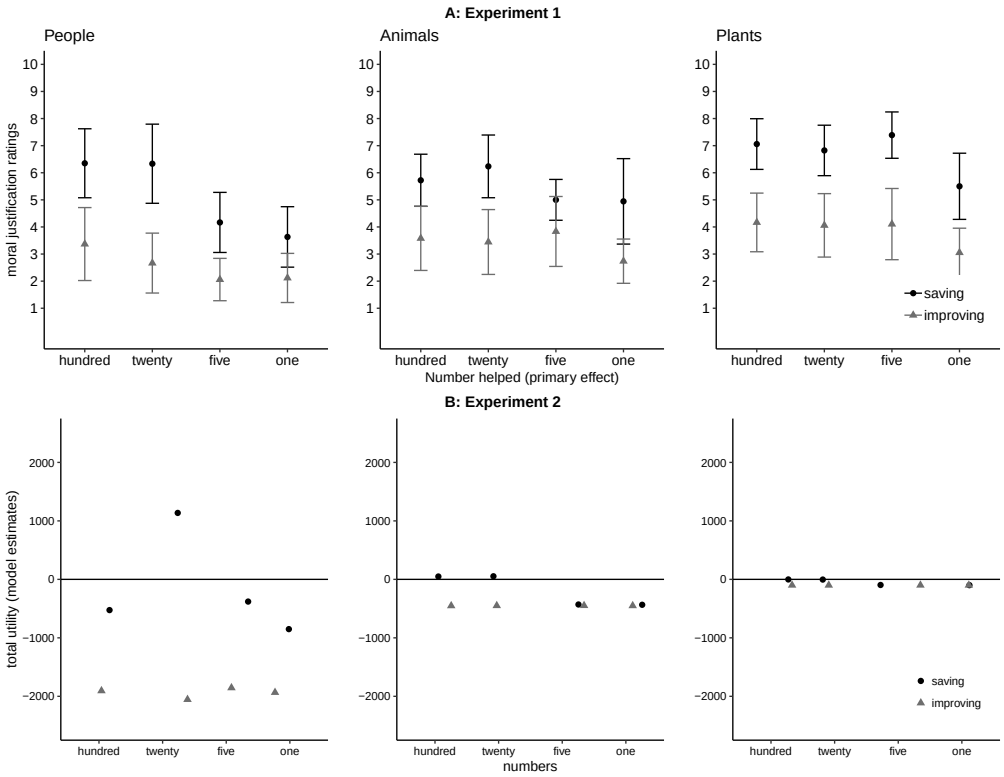


Figure 1: A: Means and 95% confidence intervals for moral justification ratings in Experiment 1, B: Total utility estimates generated by our model in Experiment 2.

separate experimental groups to avoid demand characteristics (i.e., participants feeling pressured to assign exactly five times the value of one entity to a group of five of the same entities). Within each group, we presented instances of people, fish and roses, each of them in all of the states that were described in Experiment 1 (normal vs. threatened vs. improved vs. dead). Thus, each participant judged 12 stimuli, in randomised order.[3] Like Cohen and Ahn, we presented people with a measuring standard to calibrate their value estimates. They were told that "one healthy chimpanzee" should be taken to have a value of 1000. If they valued any item half (or twice or any other ratio) as much as one healthy chimpanzee, they should assign the corresponding value to the item (e.g., a value of 500 if they value an item half as much as the chimpanzee). Participants were further instructed that "personal value" does not necessarily correspond to monetary value and that they should judge the entities' value in their *current* state. 250 participants (mean age = 36.6, SD = 13.5, 67% female, 32% male, 1% other) were recruited on Prolific Academics and completed the survey for a compensation of £0.40 (£6 per hour). Inclusion criteria were identical to Experiment 1, and not having participated in Experiment 1.

Results and Discussion To test our model, we used the value estimates of the four states of the entities to generate predictions for the justification assessments. Following the rationale outlined in Figure 2 we generated predictions for all 24 experimental conditions.The results are shown in Figure 1B. The total utilities overall capture the patterns found in Experiment 1, even though the maximal range of values was much wider for people cases compared to animals and plants (see Fig. 1A). Most importantly, the total utility estimates reflected the differences between improving versus saving, at least for people (Kruskal-Wallis $\chi^2 = 6.14$, $p = .01$) and animals (Kruskal-Wallis $\chi^2 = 6.14$, $p = .01$)[4]. In both cases the total utility for saving was larger than for improving, which mirrors the effects in Experiment 1. The corresponding effect for plants was not significant when correcting for multiple testing. Moreover, we did not find significant effects for the manipulation of the number of the affected entities for either people, animals or plants. But note that this effect was fairly small in Experiment 1 (and also in Cohen & Ahn, 2016). Also, this factor was the only one manipulated between subjects, which may have led to reduced sensitivity to this factor.

As an overall test of the fit of our model to the data of Experiment 1, we conducted a linear regression analysis with to-

[3] With the exception of the "10 entities" condition, which referred to the constant side effect. Here, we only needed estimations of each set of entities in their normal and dead states since the side effect entities never were in other states.

[4] We used again a conservative significance threshold that takes into account that we tested each factor separately for each entity category (here: $p < .017$).

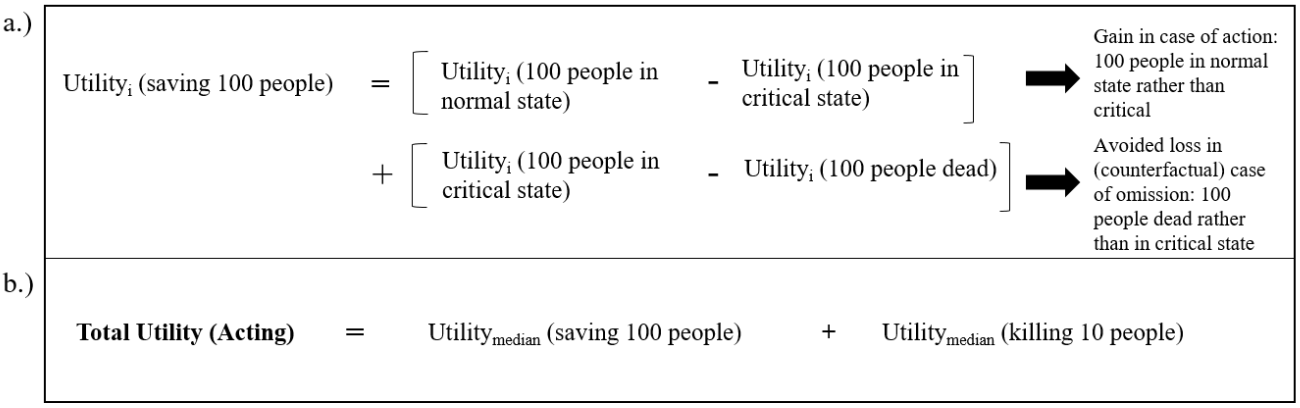


Figure 2: Rationale of our calculation of an action's total utility, spelled out for the example of the *saving 100 people* scenario. See text for explanation.

tal utilities estimated by our model as the predictor and mean moral justification ratings obtained in Experiment 1 as the criterion. The model fit the data well and explained a substantial amount of variance in the criterion, $F_{(1,22)} = 16.31$, $p<.001$, $R^2 = .43$, $RMSE = 1.14$.

General Discussion

The main goal of our study was to provide more fine-grained evidence on how moral judgments are influenced by characteristics of multiple effects of an action in dilemma situations. Experiment 1 showed that judgments of moral justification for the agent's action increased with more favourable ratios of helped compared to harmed entities, but were even more influenced by the change of state that was induced by the agent's action (saving vs. improving). Moral responsibility judgments for the negative side effect were only affected by the latter manipulation but not by the number of affected entities.

In Experiment 2 we tested a novel subjective utilitarian model that goes beyond previous proposals. Whereas Cohen and Ahn (2016) claimed that moral decisions are based on the personal values of the affected entities in their healthy or intact states, we argued that this assumption restricts their model to a small set of situations in which actions destroy or kill entities. Our goal was to propose a model that is more general. A basic assumption of our model is that actions can be modelled as state changes and that moral judgments are sensitive to both the states that entities are in prior and following a target action. This model allowed us to not only model cases of killing and saving but also, for example, cases of improvement.

Although our results in Experiment 2 showed that the new model explains a substantial amount of variance, it does not capture all effects. One reason for this may have been the necessary differences in the designs of Experiments 1 and 2. But there may be other reasons: For example, to demonstrate the increase of expressiveness of our model, we suggested a model for the cover stories of Experiment 1 that captures transitions between the four possible states mentioned there. Given that utility measurements are unreliable and influenced by additional factors, making the model more complex will certainly reduce its fit to the data.

Future research will also have to investigate whether there are alternative models that may also capture the results. As in the case of improving, we could, for example, generally use a more basic utilitarian model that only compares the two states in the presence versus absence of the action (e.g., dead vs. alive in the case of saving). Future research will need to test in greater detail the assumptions entering the different variants of the model.

We labeled our model "subjective utilitarian" because it was inspired by the theory of Cohen and Ahn (2016). However, we mentioned in the introduction that both utilitarian and nonconsequentalist theories predict that in side effect dilemmas the outcomes should be compared. Thus, our model may also be viewed as a component of a nonconsequentialist account. One possible way to test the two alternative theoretical possibilities is to take a closer look at the assumption that actions can be modeled as state changes. This assumption embodies the utilitarian claim that it is only the outcomes that matter, not the type of action leading to the outcomes. We suspect, however, that the type of action and the type of causal relations leading to the changes may also matter (see Kamm, 2007; Waldmann, Wiegmann, & Nagel, 2017). Future research will have to further explore these issues.

Acknowledgements

We thank Alex Wiegmann for helpful discussions about our utility model.

References

Bartels, D. M., & Medin, D. L. (2007). Are morally motivated decision makers insensitive to the consequences of their choices? *Psychological Science*, *18*(1), 24–28.

Cohen, D. J., & Ahn, M. (2016). A subjective utilitarian theory of moral judgment. *Journal of Experimental Psychology: General*, *145*(10), 1359–1381.

Foot, P. (1967). The problem of abortion and the doctrine of double effect. *Oxford Review*(5), 5–15.

Kamm, F. M. (2007). *Intricate ethics: Rights, responsibilities, and permissible harm*. New York: Oxford University Press.

Mikhail, J. (2011). *Elements of moral cognition: Rawls' linguistic analogy and the cognitive science of moral and legal judgment*. New York: Cambridge University Press.

Ritov, I., & Baron, J. (1999). Protected values and omission bias. *Organizational behavior and human decision processes*, *79*(2), 79–94.

Waldmann, M. R., Nagel, J., & Wiegmann, A. (2012). Moral judgment. *The Oxford handbook of thinking and reasoning*, 364–389.

Waldmann, M. R., & Wiegmann, A. (2012). The role of the primary effect in the assessment of intentionality and morality. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 34, pp. 1102–1107). Austin, TX: Cognitive Science Society.

Waldmann, M. R., Wiegmann, A., & Nagel, J. (2017). Causal models mediate moral inferences. In J.-F. Bonnefon & B. Tremolière (Eds.), *Moral inferences* (pp. 37–55). London: Routledge/Taylor Francis Group.

Learning a smooth kernel regularizer for convolutional neural networks

Reuben Feinman (reuben.feinman@nyu.edu)
Center for Neural Science
New York University

Brenden M. Lake (brenden@nyu.edu)
Department of Psychology and Center for Data Science
New York University

Abstract

Modern deep neural networks require a tremendous amount of data to train, often needing hundreds or thousands of labeled examples to learn an effective representation. For these networks to work with less data, more structure must be built into their architectures or learned from previous experience. The learned weights of convolutional neural networks (CNNs) trained on large datasets for object recognition contain a substantial amount of structure. These representations have parallels to simple cells in the primary visual cortex, where receptive fields are smooth and contain many regularities. Incorporating smoothness constraints over the kernel weights of modern CNN architectures is a promising way to improve their sample complexity. We propose a smooth kernel regularizer that encourages spatial correlations in convolution kernel weights. The correlation parameters of this regularizer are learned from previous experience, yielding a method with a hierarchical Bayesian interpretation. We show that our correlated regularizer can help constrain models for visual recognition, improving over an L2 regularization baseline.

Keywords: convolutional neural networks; regularization; model priors; visual recognition

Introduction

Convolutional neural networks (CNNs) are powerful feed-forward architectures inspired by mammalian visual processing capable of learning complex visual representations from raw image data (LeCun et al., 2015). These networks achieve human-level performance in some visual recognition tasks; however, their performance often comes at the cost of hundreds or thousands of labelled examples. In contrast, children can learn to recognize new concepts from just one or a few examples (Bloom, 2000; Xu & Tenenbaum, 2007), evidencing the use of rich structural constraints (Lake et al., 2017). By enforcing structure on neural networks to account for the regularities of visual data, it may be possible to substantially reduce the number of training examples they need to generalize. In this paper, we introduce a soft architectural constraint for CNNs that encourages smooth, correlated structure on their convolution kernels through transfer learning.[1] We see this as an important step towards a general, off-the-shelf CNN regularizer that operates independently of previous experience.

The basis for our constraint is the idea that the weights of a convolutional kernel should in general be well-structured and smooth. The weight kernels of CNNs that have been trained on the large-scale ImageNet object recognition task contain a substantial amount of structure. These kernels have parallels to simple cells in primary visual cortex, where smooth receptive fields implement bandpass oriented filters of various scale (Jones & Palmer, 1987).

[1]Experiments from this paper can be reproduced with the code found at `https://github.com/rfeinman/SK-regularization`.

(a) VGG16 layer-1 kernels

(b) i.i.d. Gaussian (L2-reg) (c) correlated Gaussian (SK-reg)

Figure 1: Kernel priors for VGG16. The layer-1 convolution kernels of VGG16, shown in (a), possess considerable correlation structure. An i.i.d. Gaussian prior that has been fit to the VGG layer-1 kernels, samples from which are shown in (b), captures little of the structure in these kernels. A correlated multivariate Gaussian prior, samples from which are shown in (c), captures the correlation structure of these kernels well.

The consistencies of visual receptive fields are explained by the regularities of image data. Locations within the kernel window have parallels to locations in image space, and images are generally smooth (Li, 2009). Consequently, smooth, structured receptive fields are necessary to capture important visual features like edges. In landmark work, Hubel & Wiesel (1962) discovered edge-detecting features in the primary visual cortex of cat. Since then, the community has successfully modeled receptive fields in early areas of mammalian visual cortex using Gabor kernels (Jones & Palmer, 1987). These kernels are smooth and contain many spatial correlations. In later stages of visual processing, locations of kernel space continue to parallel image space; however, inputs to these kernels are visual features, such as edges. Like earlier layers, these layers also benefit from smooth, structured kernels that capture correlations across the input space. Geisler et al. (2001) showed that human contour perception–an important component of object recognition–is well-explained by a model of edge co-occurrences, suggesting that correlated receptive fields are useful in higher layers of processing as well.

Despite the clear advantages of structured receptive fields, constraints placed on the convolution kernels of CNNs are typically chosen to be as general as possible, with neglect of this structure. L2 regularization–the standard soft constraint applied to kernel weights, which is interpreted as a zero-mean, independent identically distributed (i.i.d.) Gaussian prior–treats each weight as an independent random variable, with no correlations between weights expected a priori. Fig. 1 shows the layer-1 convolutional kernels of VGG16, a ConvNet trained on the large-scale ImageNet object recognition task (Simonyan & Zisserman, 2015). Fig. 1b shows some samples from an i.i.d. Gaussian prior, the equivalent of L2 regularization. Clearly, this prior captures little of the correlation structure possessed by the kernels.

A simple and logical extension of the i.i.d. Gaussian prior is a correlated multivariate Gaussian prior, which is capable of capturing some of the covariance structure in the convolution kernels. Fig. 1c shows some samples from a correlated Gaussian prior that has been fit to the VGG16 kernels. This prior provides a much better model of the kernel distribution. In this paper, we perform a series of controlled CNN learning experiments using a smooth kernel regularizer–which we denote "SK-reg"–based on a correlated Gaussian prior. The correlation parameters of this prior are obtained by fitting a Gaussian to the learned kernels from previous experience. We compare SK-reg to standard L2 regularization in two object recognition use cases: one with simple silhouette images, and another with Tiny ImageNet natural images. In the condition of limited training data, SK-reg yields improved generalization performance.

Background

Our goal in this paper is to introduce new a priori structure into CNN receptive fields to account for the regularities of image data and help reduce the sample complexity of these models. Previous methods from this literature often require a fixed model architecture that cannot be adjusted from task to task. In contrast, our method enforces structure via a statistical prior over receptive field weights, allowing for flexible architecture adaption to the task at hand. Nevertheless, in this section we review the most common approaches to structured vision models.

A popular method to enforce structure on visual recognition models is to apply a fixed, pre-specified representation. In computational vision, models of image recognition consist of a hierarchy of transformations motivated by principles from neuroscience and signal processing (e.g., Serre et al., 2007; Bruna & Mallat, 2013). These models are effective at extracting important statistical features from natural images, and they have been shown to provide a useful image representation for SVMs, logistic regression and other "shallow" classifiers when applied to recognition tasks with limited training data. Unlike CNNs, the kernel parameters of these models are not learned by gradient descent. As result, these features may not be well-adapted to the specific task at hand.

In machine learning, it is commonplace to use the features from CNNs trained on large object recognition datasets as a generic image representation for novel vision tasks (Donahue et al., 2014; Razavian et al., 2014). Due to the large variety of training examples that these CNNs receive, the learned features of these networks provide an effective representation for a range of new recognition tasks. Some *meta-learning* algorithms use a similar form of feature transfer, where a feature representation is first learned via a series of classification episodes, each with a different support set of classes (e.g., Vinyals et al., 2016). As with pre-specified feature models, the representations of these feature transfer models are fixed for the new task; thus, performance on the new task may be sub-optimal.

Beyond fixed feature representations, other approaches use a pre-trained CNN as an initialization point for a new network, following with a fine-tuning phase where network weights are further optimized for a new task via gradient descent (e.g., Girshick et al., 2014; Girshick, 2015). By adapting the CNN representation to the new task, this approach often enables better performance than fixed feature methods; however, when the scale of the required adaptation is large and the training data is limited, fine-tuning can be difficult. Finn et al. (2017) proposed a modification of the pre-train/fine-tune paradigm called model-agnostic meta-learning (MAML) that enables flexible adaptation in the fine-tuning phase when the training data is limited. During pre-training (or *meta-learning*), MAML optimizes for a representation that can be easily adapted to a new learning task in a later phase. Although effective for many use cases, this approach is unlikely to generalize well when the type of adaptation required differs significantly from the adaptations seen in the meta-learning episodes. A shared concern for all pre-train/fine-tune methods is that they require a fixed model architecture between the pre-train and fine-tune phases.

The objective of our method is distinct from those of fixed feature representations and pre-train/fine-tune algorithms. In this paper, we study the structure in the learned parameters of vision models, with the aim of extracting general structural principles that can be incorporated into new models across a broad range of learning tasks. SK-reg serves as a parameter prior over the convolution kernels of CNNs and has a theoretical foundation in Bayesian parameter estimation. This approach facilitates a CNN architecture and representation that is adapted to the specific task at hand, yet that possesses adequate structure to account for the regularities of image data. The SK-reg prior is learned from previous experience, yielding an interpretation of our algorithm as a method for hierarchical Bayesian inference.

Independently of our work, Atanov et al. (2019) developed the *deep weight prior*, an algorithm to learn and apply a CNN kernel prior in a Bayesian framework. Unlike our prior, which is parameterized by a simple multivariate Gaussian, the deep weight prior uses a sophisticated density estimator parameterized by a neural network to model the learned kernels of previously-trained CNNs. The application of this prior to new learning tasks requires variational inference with a well-calibrated variational distribution. Our goal with SK-reg differs in that we aim to provide an interpretable, generalizable prior for CNN weight kernels that can be applied to existing CNN training algorithms with little modification.

Bayesian interpretation of regularization

From the perspective of Bayesian parameter estimation, the L2 regularization objective can be interpreted as performing *maximum a-posteriori* inference over CNN parameters with a zero-mean, i.i.d. Gaussian prior. Here, we review this connection, and we discuss the extension to SK-reg.

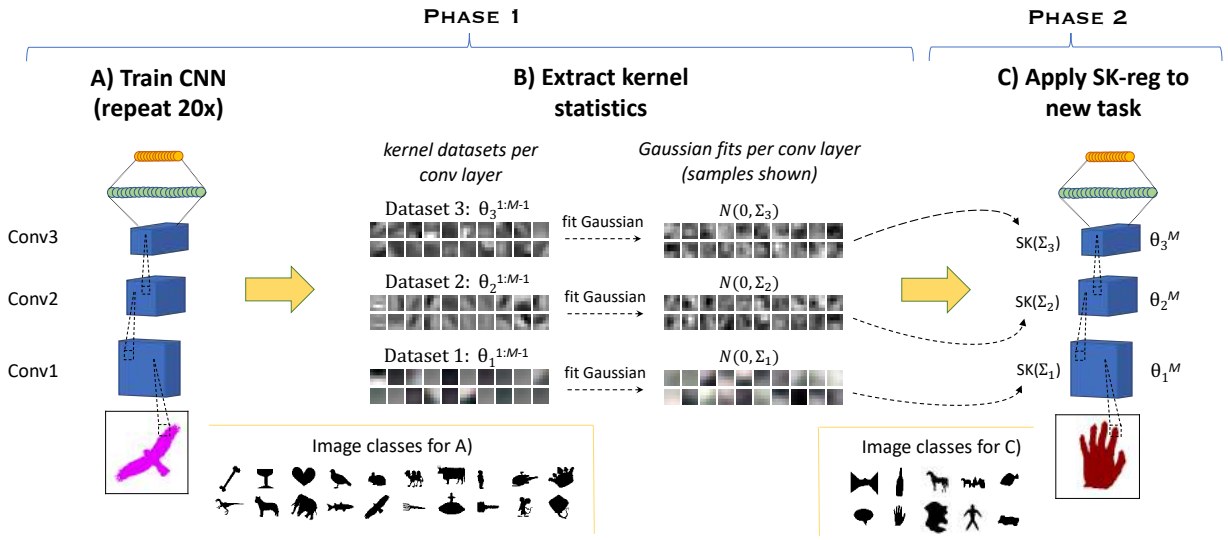


Figure 2: SK-reg workflow. A) First, a CNN is trained repeatedly (20x) on an object recognition task. B) Next, the learned parameters of each CNN are studied and statistics are extracted. For each convolution layer, kernels from the multiple CNNs are consolidated, yielding a kernel dataset for the layer. A multivariate Gaussian is fit to each kernel dataset. C) SK-reg is applied to a fresh CNN trained on a new learning task with limited training data (possibly with a different architecture or numbers of kernels), using the resulting Gaussians from each layer.

L2 regularization. Assume we have a dataset $X = \{x_1,...,x_N\}$ and $Y = \{y_1,...,y_N\}$ consisting of N images x_i and N class labels y_i. Let θ define the parameters of the CNN that we wish to estimate. The L2 regularization objective is stated as follows:

$$\tilde{\theta} = \arg\max_{\theta} \log p(Y \mid \theta; X) - \lambda * \theta^T \theta. \quad (1)$$

Here, the first term of our objective is our prediction accuracy (classification log-likelihood), and the second term is our L2 regularization penalty.

From a Bayesian perspective, this objective can be thought of as finding the *maximum a-posteriori* (MAP) estimate of the network parameter posterior $p(\theta \mid Y; X) \propto p(Y \mid \theta; X) * p(\theta)$, leading to the optimization problem

$$\tilde{\theta} = \arg\max_{\theta} \log p(Y \mid \theta; X) + \log p(\theta). \quad (2)$$

To make the connection with L2 regularization, we assume a zero-mean, i.i.d Gaussian prior over the parameters θ of a weight kernel, written as

$$p(\theta) = \frac{1}{Z} \exp\Big(-\frac{1}{2\sigma^2}\theta^T\theta\Big). \quad (3)$$

With this prior, Eq. 2 becomes

$$\tilde{\theta} = \arg\max_{\theta} \log p(Y \mid \theta; X) - \frac{1}{2\sigma^2}\theta^T\theta,$$

which is the L2 objective of Eq. 1, with $\lambda = \frac{1}{2\sigma^2}$.

SK regularization. The key idea behind SK-reg is to extend the L2 Gaussian prior to include a non-diagonal covariance matrix; i.e., to add correlation. In the case of SK-reg, the prior over kernel weights θ of Eq. 3 becomes

$$p(\theta) = \frac{1}{Z} \exp\Big(-\frac{1}{2}\theta^T\Sigma^{-1}\theta\Big)$$

for some covariance matrix Σ, and the new objective is written

$$\tilde{\theta} = \arg\max_{\theta} \log p(Y \mid \theta; X) - \lambda * \theta^T\Sigma^{-1}\theta. \quad (4)$$

Hierarchical Bayes. When Σ is learned from previous experience, SK-reg can be interpreted as approximate inference in a hierarchical Bayesian model. The SK regularizer for a CNN with C layers, $\boldsymbol{\Sigma} = \{\Sigma_1,\ldots,\Sigma_C\}$, assumes a unique zero-mean Gaussian prior $\mathcal{N}(\theta_i; 0, \Sigma_i)$ over the weight kernels for each convolutional layer, $\boldsymbol{\theta} = \{\theta_1,\ldots,\theta_C\}$. Due to the regularities of the visual world, it is plausible that effective general priors exist for each layer of visual processing. In this paper, transfer learning is used to fit the prior covariances $\boldsymbol{\Sigma}$ from previous datasets $X^{1:M-1}$ and $Y^{1:M-1}$, which informs the solution for a new problem X^M and Y^M, yielding the hierarchical Bayesian interpretation depicted in Fig. 3. Task-specific

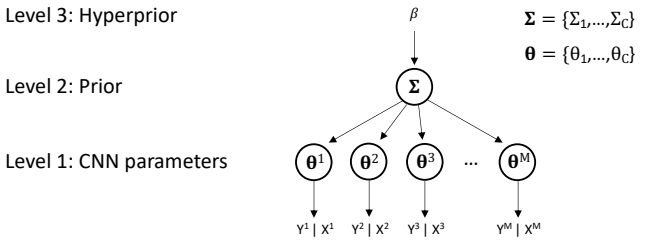


Figure 3: A hierarchical Bayesian interpretation of SK-reg. A point estimate of prior parameters $\boldsymbol{\Sigma}$ is first computed with MAP estimation. Next, this prior is applied to estimate CNN parameters $\boldsymbol{\theta}^j$ in a new task.

CNN parameters $\boldsymbol{\theta}^{1:M}$ are drawn from a common $\boldsymbol{\Sigma}$, and $\boldsymbol{\Sigma}$ has a hyperprior specified by β. Ideal inference would compute $p(Y^M|Y^{1:M-1};X^{1:M})$, marginalizing over $\boldsymbol{\theta}^{1:M}$ and $\boldsymbol{\Sigma}$.

We propose a very simple empirical Bayes procedure for learning the kernel regularizer in Eq. 4 from data. First, $M-1$ CNNs are fit independently to the datasets $X^{1:M-1}$ and $Y^{1:M-1}$ using standard methods, in this case optimizing Eq. 1 to get point estimates $\tilde{\boldsymbol{\theta}}^{1:M-1}$. Second, a point estimate $\tilde{\boldsymbol{\Sigma}}$ is computed by maximizing $p(\boldsymbol{\Sigma}|\tilde{\boldsymbol{\theta}}^{1:M-1};\beta)$, which is a simple regularized covariance estimator. Last, for a new task M with training data X^M and Y^M, a CNN with parameters $\boldsymbol{\theta}^M$ is trained with the SK-reg objective (Eq. 4), with $\boldsymbol{\Sigma}=\tilde{\boldsymbol{\Sigma}}$.

This procedure can be compared with the hierarchical Bayesian interpretation of MAML (Grant et al., 2018). Unlike MAML, our method allows flexibility to use different architectures for different datasets/episodes, and the optimizer for $\boldsymbol{\theta}^M$ is run to convergence rather than just a few steps.

Experiments

We evaluate our approach within a set of controlled visual learning environments. SK-reg parameters Σ_i for each convolution layer θ_i are determined by fitting a Gaussian to the kernels acquired from an earlier learning phase. We divide our learning tasks into two unique phases, applying the same CNN architecture in each case. We note that our approach does not require a fixed CNN architecture across these two phases; the number of feature maps in each layer may be easily adjusted. A depiction of the two learning phases is given in Fig. 2.

Phase 1. The goal of phase 1 is to extract general principles about the structure of learned convolution kernels by training an array of CNNs and collecting statistics about the resulting kernels. In this phase, we train a CNN architecture to classify objects using a sufficiently large training set with numerous examples per object class. Training is repeated multiple times with unique random seeds, and the learned convolution kernels are stored for each run. During this phase, standard L2 regularization is applied to enforce a minimal constraint on each layer's weights (optimization problem of Eq. 1). After training, the convolution kernels from each run are consolidated, holding each layer separate. A multivariate Gaussian is fit to the centered kernel dataset of each layer, yielding a distribution $N(0,\Sigma_i)$ for each convolution layer i. To ensure the stability of the covariance estimators, we apply shrinkage to each covariance estimate, mixing the empirical covariance with an identity matrix of equal dimensionality. This can be interpreted as a hyperprior $p(\boldsymbol{\Sigma};\beta)$ (Fig. 3) that favors small correlations. The optimal mixing parameter is determined via cross-validation.

Phase 2. In phase 2, we test the aptitude of SK-reg on a new visual recognition task, applying the covariance matrices Σ_i obtained from phase 1 to regularize each convolution layer i in a freshly-trained CNN (optimization problem of Eq. 4). In order to adequately test the generalization capability of our algorithm, we use a new set of classes that differ from the phase 1 classes in substantial ways, and we provide just a few training examples from each class. Performance of SK-reg is compared against standard L2 regularization.

Figure 4: Exemplars of the phase 1 silhouette object classes.

Layer	Window	Stride	Features	λ
Input (200x200x3)				
Conv2D	5x5	2	5	0.05
MaxPooling2D	3x3	3		
Conv2D	5x5	1	10	0.05
MaxPooling2D	3x3	2		
Conv2D	5x5	1	8	0.05
MaxPooling2D	3x3	1		
FullyConnected			128	0.01
Softmax				

Table 1: CNN architecture. Layer hyperparameters include window size, stride, feature count, and regularization weight (λ). Dropout is applied after the last pooling layer and the fully-connected layer with rates 0.2 and 0.5, respectively.

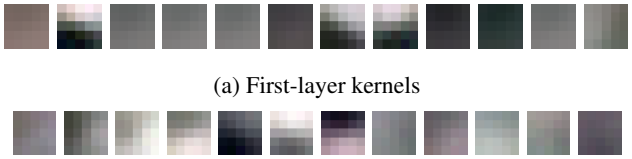
Silhouettes

As a preliminary use case, we train our network using the binary shape image dataset developed at Brown University[2], henceforth denoted "Silhouettes." Silhouette images are binary masks that depict the structural form of various object classes. Simple shape-based stimuli such as these provide a controlled learning environment for studying the inductive biases of CNNs (Feinman & Lake, 2018). We select a set of 20 well-structured silhouette classes for phase 1, and a set of 10 unique, well-structured classes for phase 2 that differ from phase 1 in their consistency and form. The images are padded to a fixed size of 200×200.

During phase 1, we train our network to perform 20-way object classification. Exemplars of the phase 1 classes are shown in Fig. 4. The number of examples varies for each class, ranging from 12 to 49 with a mean of 24. Class weighting is used to remedy class imbalances. To add complexity to the silhouette images, colors are assigned randomly to each silhouette before training. During training, random translations, rotations and horizontal flips are applied at each training epoch to improve generalization performance.

We use a CNN architecture with 3 convolution layers, each followed by a max pooling layer (see Table 1). Hyperparameters including convolution window size, pool size, and filter counts were selected via randomized grid-search, using a validation set with examples from each class to score candidate values. A rectified linear unit (ReLU) nonlinearity is applied to the output of each convolution layer, as well as to the

[2]The binary shape dataset is available in the "Databases" section at `http://vision.lems.brown.edu`



(a) First-layer kernels

(b) Gaussian samples

Figure 5: Learned first-layer kernels vs. Gaussian samples. (a) depicts some of the learned first-layer kernels acquired from phase 1 silhouette training. For comparison, (b) shows a few samples from a multivariate Gaussian that was fit to the first-layer kernel dataset.

fully-connected layer. The network is trained 20 times using the Adam optimizer, each time with a unique random initialization. It achieves an average validation accuracy of 97.7% across the 20 trials, indicating substantial generalization.

Following the completion of phase 1 training, a kernel dataset is obtained for each convolution layer by consolidating the learned kernels for that layer from the 20 trials. Covariance matrices Σ_i for each layer i are obtained by fitting a multivariate Gaussian to the layer's kernel dataset. For a first-layer convolution with window size $K \times K$, this Gaussian has dimensionality $3K^2$, equal to the window area times RGB depth. We model the input channels as separate variables in layer 1 because these channels have a consistent interpretation as the RGB color channels of the input image. For remaining convolution layers, where the interpretation of input channels may vary from case to case, we treat each input channel as an independent sample from a Gaussian with dimensionality K^2. The kernel datasets for each layer are centered to ensure zero mean, typically requiring only a small perturbation vector.

To ensure that our multivariate Gaussians model the kernel data well, we computed the cross-validated log-likelihoods of this estimator on each layer's kernel dataset and compared them to those of an i.i.d. Gaussian estimator fit to the same data. The multivariate Gaussian achieved an average score of 358.5, 413.3 and 828.1 for convolution layers 1, 2 and 3, respectively. In comparison, the i.i.d. Gaussian achieved an average score of 144.4, 289.6 and 621.9 for the same layers. These results confirm that our multivariate Gaussian provides an improved model of the kernel data. Some examples of the first-layer convolution kernels are shown in Fig. 5 alongside samples from our multivariate Gaussian that was fit to the first-layer kernel dataset. The samples appear structurally consistent with our phase 1 kernels.

In phase 2, we train our CNN on a new 10-way classification task, providing the network with just 3 examples per class for gradient descent training and 3 examples per class for validation. Colors are again added at random to each silhouette in the dataset. The network is initialized randomly, and we apply SK-reg to the convolution kernels of each layer during training using the covariance matrices obtained in phase 1. Our validation set is used to track and save the best model over the course of the training epochs (early stopping). A holdout set with 6 examples per class is used to assess the final performance of the model. A depiction of the train, validation and test sets used for phase 2 is given in Fig. 6. The validation and test images have been shifted, translated and flipped to make for a more challenging generalization test. Similar to phase 1, random shifts, rotations and horizontal flips are applied to the training images at each training epoch. As a baseline, we also train our CNN using standard L2 regularization.

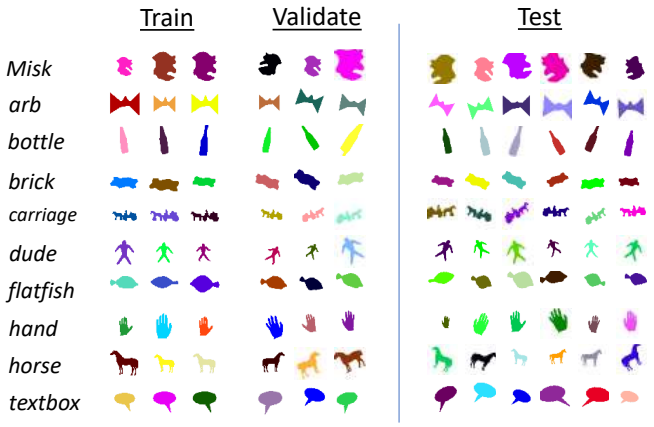


Figure 6: Silhouettes phase 2 datasets. 3 examples per class are provided in both the train and validation sets. A holdout test set with 6 examples per class is used to evaluate final model performance.

Method	λ	Cross-entropy	Accuracy
L2	0.214	2.000 (+/- 0.033)	0.530 (+/- 0.013)
SK	0.129	0.597 (+/- 0.172)	0.821 (+/- 0.056)

Table 2: Silhouettes phase 2 results. For each regularization method, the optimal regularization weight λ was selected via grid-search. Results show the average cross-entropy and classification accuracy achieved on the holdout test set over 10 phase 2 training runs.

The regularization weight λ is an important hyperparameter of both SK and L2 regularization. Before performing the phase 2 training assessment, we use a validated grid search to select the optimal λ for each regularization method, applying our train/validate sets.[3] The same weight λ is applied to each convolution layer, as done in phase 1.

Results. With our optimal λ values selected, we trained our CNN on the 10-way phase 2 classification task of Fig. 6, comparing SK regularization to a baseline L2 regularization model. Average results for the two models collected over 10 training runs are presented in Table 2. Average test accuracy is improved by roughly 55% with the addition of SK reg, a substantial performance boost from 53.0% correct to 82.1% correct. Clearly, a priori structure is beneficial to generalization in this use case. An inspection of the learned kernels confirms that SK-reg encourages the structure we expect; these

[3]To yield interpretable λ values that can be compared between the SK and L2 cases, we normalize each covariance matrix to unit determinant by applying a scaling factor c, such that $\det(c\Sigma) = \det(I)$.

kernels look visually similar to samples from the Gaussian (e.g. Fig. 5).

Tiny ImageNet

Our silhouette experiment demonstrates the effectiveness of SK-reg when the parameters of the regularizer are determined from the structure of CNNs trained on a similar image domain. However, it remains unclear whether these regularization parameters can generalize to novel image domains. Due to the nature of the silhouette images, the silhouette recognition task encourages representations with properties that are desirable for object recognition tasks in general. Categorizing silhouettes requires forming a rich representation of shape, and shape perception is critical to object recognition. Therefore, this family of representation may be useful in a variety of object recognition tasks.

To test whether our kernel priors obtained from silhouette training generalize to a novel domain, we applied SK-reg to a simplified version of the Tiny ImageNet visual recognition challenge, using covariance parameters fitted to silhouette-trained CNNs. Tiny ImageNet images were up-sampled with bilinear interpolation from their original size of 64×64 to mirror the Silhouette size 200×200. We selected 10 well-structured ImageNet classes that contain properties consistent with the silhouette images.[4] We performed 10-way image classification with these classes, using the same CNN architecture from Table 1 and applying the SK-reg soft constraint. The network is provided 10 images per class for training and 10 per class for validation. Because of the increased complexity of the Tiny ImageNet data, a larger number of examples per class is merited to achieve good generalization performance. A holdout test set with 20 images per class is used to evaluate performance. Fig. 7 shows a breakdown of the train, validate and test sets.

A few modifications were made to account for the new image data. First, we modified the phase 1 silhouette training used to acquire our covariance parameters, this time applying random colors to both the foreground and background of each silhouette. Previously, each silhouette overlaid a strictly white background. Consequently, the edge detectors of the learned CNNs would be unlikely to generalize to novel color gradients. Second, we added additional regularization to our covariance estimators to avoid over-fitting and help improve the generalization capability of the resulting kernel priors. Due to the nature of the phase 2 task in this experiment, and the extent to which the images differ from phase 1, additional regularization was necessary to ensure that our kernel priors could generalize. Specifically, we applied L1-regularized inverse covariance estimation (Friedman et al., 2008) to estimate each Σ_i, which can be interpreted as a hyperprior $p(\mathbf{\Sigma}; \beta)$ (Fig. 3) that favors a sparse inverse covariance (Lake & Tenenbaum, 2010).

Similar to the silhouettes experiment, the validation set is used to select weighting hyperparameter λ and to track the best model over the course of learning. As a baseline, we again compared SK-reg to a λ-optimized L2 regularizer.

[4] Desirable classes have a uniform, centralized object with consistent shape properties and a distinct background.

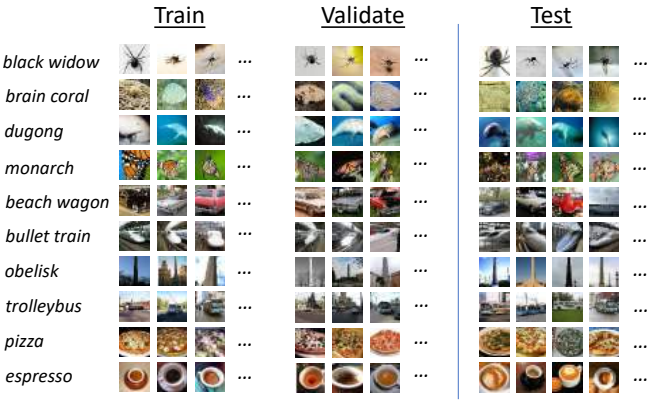


Figure 7: Tiny ImageNet datasets. 10 classes were selected to form a 10-way classification task. The train and validate sets each contain 10 examples per class. The holdout test set contains 20 examples per class.

Method	λ	Cross-entropy	Accuracy
L2	0.450	1.073 (+/- 0.102)	0.700 (+/- 0.030)
SK	0.450	0.956 (+/- 0.180)	0.776 (+/- 0.035)

Table 3: Tiny ImageNet SK-reg and L2 results. Table shows the average cross-entropy and classification accuracy achieved on the holdout test set over 10 training runs.

Results. SK-reg improved the average holdout performance received from 10 training runs as compared to an L2 baseline, both in accuracy and cross-entropy. Results for each regularization method, as well as their optimal λ values, are reported in Table 3. An improvement of 8% in test accuracy suggests that some of the structure captured by our kernel prior is useful even in a very distinct image domain. The complexity of natural images like ImageNet is vast in comparison to simple binary shape masks; nonetheless, our prior from phase 1 silhouette training is able to influence ImageNet learning in a manner that is beneficial to generalization.

Discussion

Using a set of controlled visual learning experiments, our work in this paper demonstrates the potential of structured receptive field priors in CNN learning tasks. Due to the properties of image data, smooth, structured receptive fields have many desirable properties for visual recognition models. In our experiments, we have shown that a simple multivariate Gaussian model can effectively capture some of the structure in the learned receptive fields of CNNs trained on simple object recognition tasks. Samples from the fitted Gaussians are visually consistent with learned receptive fields, and when applied as a model prior for new learning tasks, these Gaussians can help a CNN generalize in conditions of limited training data. We demonstrated our new regularization method in two simple use cases. Our silhouettes experiment shows that,

when the parameters of SK-reg are determined from CNNs trained on a similar image domain to that of the new task, the performance increase that results in the new task can be quite substantial–as large as 55% over an L2 baseline. Our Tiny ImageNet experiment demonstrates that SK-reg is capable of encoding generalizable structural principles about the correlations in receptive fields; the statistics of learned parameters in one domain can be useful in a completely new domain with substantial differences.

The Gaussians that we fit to kernel data in phase 1 of our experiments could be overfit to the CNN training runs. We have discussed the application of sparse inverse covariance (precision) estimation as one approach to reduce over-fitting. In future work, we would like to explore a Gaussian model with graphical connectivity that is specified by a 2D grid MRF. Model fitting would consist of optimizing the non-zero precision matrix values subject to this pre-specified sparsity. The grid MRF model is enticing for its potential to serve as a general "smoothness" prior for CNN receptive fields. Ultimately, we hope to develop a general-purpose kernel regularizer that does not depend on transfer learning.

Although a Gaussian can model some kernel families sufficiently, other families would give it a difficult time. The first-layer kernels of AlexNet–which are 11×11 and are visually similar to Gabor wavelets and derivative kernels–are not well-modeled by a multivariate Gaussian. A more sophisticated prior is needed to model kernels of this size effectively. In future work, we hope to investigate more complex families of priors that can capture the regularities of filters such as Gabors and derivatives. Nevertheless, a simple Gaussian estimator works well for smaller kernels, and in the literature, it has been shown that architectures with a hierarchy of smaller convolutions followed by nonlinearities can achieve equal (and often better) performance as those will fewer, larger kernels (Simonyan & Zisserman, 2015). Thus, the ready-made Gaussian regularizer we introduced here can be used in many applications.

Acknowledgements

We thank Nikhil Parthasarathy, Emin Orhan and Brian McFee for their valuable comments. Reuben Feinman is supported by a Google PhD Fellowship in Computational Neuroscience.

References

Atanov, A., Ashukha, A., Struminsky, K., Vetrov, D., & Welling, M. (2019). The deep weight prior. In *Proceedings of the 7th International Conference on Learning Representations (ICLR).*

Bloom, P. (2000). *How children learn the meanings of words*. Cambridge, MA: MIT Press.

Bruna, J., & Mallat, S. (2013). Invariant scattering convolution networks. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, *35*(8), 1872–1886.

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., & Darrell, T. (2014). A deep convolutional activation feature for generic visual recognition. In *Proceedings of the 31st International Conference on Machine Learning (ICML).*

Feinman, R., & Lake, B. M. (2018). Learning inductive biases with simple neural networks. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society (CogSci).*

Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML).*

Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, *9*(3), 432–441.

Geisler, W. S., Perry, J. S., Super, B. J., & Gallogly, D. P. (2001). Edge co-occurence in natural images predicts contour grouping performance. *Vision Research*, *41*(6), 711–724.

Girshick, R. (2015). Fast R-CNN. In *Proceedings of the International Conference on Computer Vision (ICCV).*

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

Grant, E., Finn, C., Levine, S., Darrell, T., & Griffiths, T. (2018). Recasting gradient-based meta-learning as hierarchical Bayes. In *Proceedings of the 6th International Conference on Learning Representations (ICLR).*

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiology*, *160*, 106–154.

Jones, J. P., & Palmer, L. A. (1987). An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. *J. Neurophysiology*, *58*, 1233–1258.

Lake, B. M., & Tenenbaum, J. B. (2010). Discovering structure by learning sparse graphs. In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society (CogSci).*

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*, E253.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Li, S. Z. (2009). *Markov random field modeling in image analysis*. New York, NY: Springer-Verlag.

Razavian, A. S., Azizpour, H., Sullivan, J., & Carlsson, S. (2014). CNN features off-the-shelf: An astounding baseline for recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., & Poggio, T. (2007). Robust object recognition with cortex-like mechanisms. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, *29*(3), 411–426.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR).*

Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K., & Wierstra, D. (2016). Matching networks for one shot learning. In *Advances in Neural Information Processing Systems (NIPS).*

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, *114*(2), 245–272.

Mapping Space: A Comparative Study

Michele I. Feist (feist@louisiana.edu)
Department of English, University of Louisiana at Lafayette, P.O. Box 43719
Lafayette, LA 70504, USA
Yuan Zhang (mzrwhw@gmail.com)
Independent scholar, Ottawa, Ontario K2W 1H2, Canada

Abstract

The semantics of spatial terms has attracted substantial attention in the cognitive sciences, revealing both compelling similarities and striking differences across languages. However, much of the evidence regarding cross-linguistic variation pertains to fine-grained comparisons between individual lexical items, while cross-linguistic similarities are found in more coarse-grained studies of the conceptual space underlying semantic systems. We seek to bridge this gap, moving beyond the semantics of individual terms to ask what the comparison of spatial semantic systems may reveal about the conceptualization of locations in English and Mandarin Chinese and about the nature of potential universals in this domain. We subjected descriptions of 116 spatial scenes to multidimensional scaling analyses in order to reveal the structures of the underlying conceptual spaces in each language. In addition to revealing overlaps and divergences in the conceptualization of space in English and Mandarin, our results suggest a difference in complexity, whereby Mandarin terms are accommodated by a lower-dimensional similarity space than are English terms.

Keywords: spatial semantics; universals; cross-linguistic variation

Introduction

All peoples, in all languages, have occasion to talk about the locations of objects in their environments – environments which are fundamentally similar. Despite this, the vocabularies of space differ strikingly across languages, fueling interest in spatial semantics across the cognitive sciences. Most notably, scholars have observed that both the number and the nature of the contrasts that are encoded vary markedly from language to language (Bowerman 1996; Bowerman & Choi 2001; Feist 2000; Gentner & Bowerman 2009; Landau & Jackendoff 1993), with the result that "translation equivalents" for spatial terms can be quite different in meaning (Feist 2013; Trujillo 1995).

As a case in point, the range of spatial configurations that can be described using the English preposition *on* is divided amongst three prepositions – *op, aan*, and *om* – in Dutch (Gentner & Bowerman 2009); Dutch thus routinely encodes distinctions that are optional in English. More strikingly, even the dimensions of contrast encoded in spatial semantic systems may vary across languages: whereas English encodes a distinction between containment and support, Korean encodes a distinction between tight and loose fit (Bowerman & Choi 2001) that neutralizes the containment/support contrast.

Tempering these findings of variation is an overall structuring of the semantic domain of topological relations which appears to be shared cross-linguistically. For example, despite finding evidence of a "fractionated picture of overlapping contrasts" (Levinson & Wilkins 2006, p. 520) which echoes the variation briefly reviewed above, Levinson and Wilkins argue that the extensional ranges of the adpositions in the dozen languages they studied suggest a common underlying conceptual space. This is consistent with earlier findings suggesting that topological notions may be organized in a coherent conceptual space characterized by a small set of "attractors" – groups of situations that are likely to be lexicalized in similar ways across languages (Levinson & Meira 2003), including "ATTACHMENT", "IN", and "ON-TOP". Thus, while the semantics of individual spatial terms in different languages may differ from one another, the underlying conceptual components that make them up are argued to be drawn from a common set. This conclusion is supported by the work of Feist (2008), who found that the extensional ranges of spatial terms across a sample of 24 languages could be accommodated by a two-dimensional similarity space, with one dimension encoding the degree to which the reference object constrains the location of the located object, while the second dimension encodes the relative vertical positions of the two objects. Taken together, these studies in semantic typology suggest that the cross-linguistic variation that has often been noted is overlaid upon a common conceptual core.

The stark contrast between the word-level evidence of cross-linguistic variation and the system-level evidence of a common conceptual core raises many questions regarding the conceptualization of space. Is cross-linguistic variation limited to fine-grained details of lexical encoding, leaving a substantial universal conceptual basis intact? This would suggest that, while languages vary in the contrasts they mark, each structures its semantic system around fundamentally the same topological concepts. Or is the fine-grained cross-linguistic variation evidence of deeper differences in the nature of the topological concepts underlying the meanings of spatial terms? This would suggest that the system-level similarities that have been observed are in fact quite abstract, with variation arising within the set of topological concepts upon which the meanings of lexical items are based.

Zhang, Segalowitz, and Gatbonton (2011) began to address questions such as these, asking whether Mandarin Chinese and English differ with respect to the conceptual specification of containment and support rather than merely in the mapping of these two concepts onto spatial lexemes. They had speakers describe a set of 116 line drawings depicting a range of topological relations in order to examine the lexicalization

of containment and support in the two languages. Each of the elicited spatial terms was classified as encoding containment, support, or "other concepts", and the extensions of containment-encoding and support-encoding lexemes in the two languages were compared. They found that approximately half of the pictures were categorized similarly at this broad level of detail (either as examples of containment or as examples of support) by speakers of the two languages, representing a large overlap in how the concepts of containment and support may be represented in English and in Mandarin. However, Zhang and her colleagues also noted differences in the uses of containment-encoding and support-encoding adpositions: Mandarin speakers described a larger proportion of the pictures using support-encoding lexical items than using containment-encoding lexical items, while English speakers evidenced the opposite pattern. Much of this difference could be accounted for via a difference in the encoding of partial inclusion and of part-whole relations: these relations tended to be described using support-encoding adpositions in Mandarin, but containment-encoding adpositions in English. This pattern of results suggests cross-linguistic differences in the boundaries separating the two conceptual categories, despite overlap in their cores, thus situating variation as a lexicalization phenomenon, rather than as evidence that the conceptual systems – and the topological concepts themselves – differ.

In a similar vein, Johannes and her colleagues asked whether the cores of lexicalized containment and support concepts were similar across languages (Johannes et al. 2015; Landau et al. 2017). They asked speakers to describe scenes predefined as representing subtypes of containment or support, then examined the rate of use of the Basic Locative Construction[1] for each subtype. For both concepts, they found that the rate of use of this construction was highest for a similar range of subtypes across the languages sampled, suggesting that these subtypes may constitute universal conceptual cores for containment and support.

Before we can conclude that the conceptual cores are indeed universal, however, we need to take a closer look at the extensions of containment- and support-encoding spatial terms as a source of evidence for the underlying structures of the concepts, without prejudging the status of either the adpositions or the scenes as exemplars of containment or support. In their study, Zhang et al (2011) classified each adposition *a priori* as encoding support, containment, or "other concepts"; they then used this classification to explore the kinds of situations that will be encoded as either containment or support in Mandarin and in English. In so doing, they neutralized fine-grained contrasts marked by the lexical items in the two languages, in essence positing that coherent, unified concepts of support and containment are encoded in English and Mandarin. In a parallel fashion, Johannes and her colleagues (2015; Landau et al. 2017) classified the scenes used in their studies as exemplars of either containment or support. In addition, they limited the scope of their study to variation in the use of BE *in/on* (and its translation equivalents), leaving fine-grained semantic contrasts unexplored. This methodology likewise assumes the existence of coherent, unified concepts of support and containment. Such unified concepts, however, cannot be assumed. As a case in point, the coherence of support as a universally salient concept has been contradicted by cross-linguistic evidence, with support relations clustering with two different groups of scenes in Levinson and Meira's (2003) analysis. In abstracting away from the semantic richness of the spatial adpositions and the complexity of the scenes, these studies may have inadvertently introduced a universal structure to the systems rather than objectively testing for its presence. In this paper, we reintroduce the semantic richness of the spatial terms while removing the *a priori* categorization of the scenes and adpositions in order to better assess the degree of similarity between the Mandarin and the English spatial semantic systems.

Spatial semantics in paradigmatic perspective

In order to better understand the comparison between the Mandarin spatial descriptors and the English ones, we shift the focus of our attention from the spatial terms as exemplars of abstract concepts to the spatial terms as indicators of linguistically-relevant degrees of similarity amongst spatial scenes (cf., Croft 2010; Feist 2008; Levinson & Meira 2003). Because words name categories, when speakers use a single word to describe two scenes, they are relying on a perceived similarity between the scenes that enables the sameness of description. Conversely, when two scenes are described using different words, speakers are highlighting differences between the scenes.

We can examine the patterns of similarity that underlie a language's semantic system via statistical techniques such as multidimensional scaling (MDS). MDS uses the co-occurrence of lexical items and pictures to construct a similarity space in which the placement of each picture is a function of the extent to which the lexical items used to describe it overlap with the lexical items used to describe each of the other pictures in the set. For example, consider the two pictures in Figure 1. If one speaker described the apple as *in* the bowl and the boat as *in* the water, this would provide evidence that the two scenes are similar, and should be placed close together in the similarity space. However, if another speaker instead described the boat as *on* the water, this would temper that judgment of similarity, and result in some distance between the two pictures. By adding in evidence from multiple speakers, a fuller picture may emerge of the extent to which each pair of pictures is treated as similar by speakers of a language.

[1] Although Levinson and Meira (2003, p. 486) define this construction as "answers to *where* questions", Johannes and her colleagues limited their investigation to BE *in/on* (and its equivalent in the other languages studied).

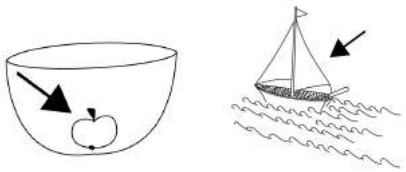


Figure 1: Two spatial scenes (from Bowerman & Pederson 1992b)

The output of multidimensional scaling is a similarity space reflecting this evidence, in which the placement of points (each representing a picture in the set) is a function of the overlap between the set of spatial terms used to describe that picture and the sets of spatial terms used to describe all the other pictures under study. The more shared spatial terms in the elicited descriptions for a pair of pictures, the closer the pictures will be in the final similarity space (see Croft 2010; Croft & Poole 2008; Feist 2008; Levinson & Meira 2003). Hence, the similarity space presents a visual representation of the conceptual space underlying the uses of the spatial terms.

To assess the adequacy of the account of the patterns in the data provided by the similarity space, we can examine both the percent correct classification, indicating the proportion of the pictures in the solution that are placed correctly relative to the elicited naming patterns, and the aggregate proportional reduction of error (APRE), indicating the extent to which the resulting solution improves upon a solution which places all of the pictures in a single category (see Croft & Poole 2008; Poole 2000, 2005 for further discussion). The APRE is measured on a scale from 0-1, with higher values indicating fewer errors in the model.

With a cross-linguistic data set, MDS returns a representation of a space upon which all the languages in the set may be overlaid such that the distinctions marked in each language isolate contiguous sets of points (Croft & Poole 2008; Feist 2008); the fewer that are miscategorized, the better the solution. As such, MDS provides a means by which we may identify potential universals underlying the semantic systems of a varied set of languages (Feist 2008; Levinson & Meira 2003). In the current study, we use MDS to construct separate similarity spaces for a set of simple spatial scenes as described by speakers of Mandarin and by speakers of English. With single-language data sets such as these, the conceptual space that MDS returns is one that only respects the distinctions marked in that language, thus providing a representation of the similarities amongst the pictures in the set as encoded in the naming patterns of the language under study. Comparison of the conceptual spaces resulting from separate MDS analyses, thus, gives a novel view into the fine-grained differences in the contrasts marked within the semantic systems of the examined languages, thus enabling a richer comparison than has been possible in previous work.

Method

The Corpus We used the 5800 picture descriptions (2900 descriptions from each language) collected by Zhang et al. (2011). The descriptions were elicited using 116 simple line drawings: 65 pictures from Bowerman and Pederson's (1992b) Topological Relations Picture Series (pictures 18, 20, 24, 33, 47, and 59 were excluded; see Zhang 2013) and an additional 51 developed by Zhang (2013). Each drawing depicts two objects – one highlighted in yellow, and one in black and white – in a simple spatial relation, with the names of the objects printed below the picture (in English or in Mandarin, as appropriate). The pictures depicted a range of topological relations; example pictures are shown in Figure 1. The pictures were printed two to a page, vertically aligned**.**

The set of pictures was described by 25 native speakers of English living in Montreal, Canada, and 25 native speakers of Mandarin living in Harbin, China. All speakers reported themselves either to be monolingual, or to have only limited knowledge of a second language. The pictures were presented in random order, and participants were asked to describe for each the location of the yellow object with respect to the black and white one.

Analysis In order to be able to compare the structuring of space in the two languages, the English and Mandarin descriptions were analyzed separately. We used Poole's Optimal Classification nonparametric unfolding algorithm (Poole 2000, 2005; see also Croft 2010; Feist 2008) to perform MDS analyses of the two sets of descriptions.

Our procedure was as follows. First, we identified the spatial terms used in each of the elicited descriptions. Because our aim was to analyze spatial term usage at a fine level of detail, we considered each adpositional expression to be a separate spatial term, hence *in* was separate from *inside; on,* from *on top*. This resulted in identification of 36 spatial terms in Mandarin and 38 in English. Next, we constructed two matrices – one for each language – with the 116 pictures defining the rows and with the elicited spatial terms defining the columns. Within each matrix, we then filled in each cell to indicate whether the spatial term heading the column had been used by any participant to describe the picture heading the row (cf., Feist 2008). These matrices were then input into the Optimal Classification algorithm as implemented within the R programming environment.

Results

We look first at the results for each language separately, beginning with English. Next, we turn to the comparison between the English solution spaces and the Mandarin ones.

English The lowest dimensional fit that provided a high rate of correct classification and a substantial improvement over a null model (i.e., one in which all the pictures are in a single category) was the two-dimensional solution, with 97.7% correct classification and an APRE of .765. The conceptual space associated with this solution is presented in Figure 2.

A close examination of Figure 2 reveals that the dimensions in the solution space readily admit of semantic interpretation. The x-axis, anchored by pictures of a ball underneath an upside-down bowl (and, hence, located at its

interior) [picture number 110] and of a muscle in a leg [83] on the left end, and by pictures of a city at the shore of an ocean [109] and of a rope wound around a tree stump [43] on the right, corresponds to a continuum between interior location and surface contact. The y-axis, on the other hand, is anchored by pictures of a dog resting beside a dog house [6] and of a garden on the bank of a river [108] on the upper end, and by pictures of a gate in a fence [136] and of a muscle in a leg [83] at the lower. This axis thus corresponds to variation in the amount of control that the ground exerts over the figure. Along the y-axis we also see variation in the alienability of the objects (i.e., the extent to which the relation between them is inherent to their nature [Strazny 2005]), with more alienable connections (including tree/house [49] and garden/river [108]) anchoring the upper end of the dimension, and more inalienable connections (including muscle/leg [83] and gate/fence [136]) anchoring the lower end.

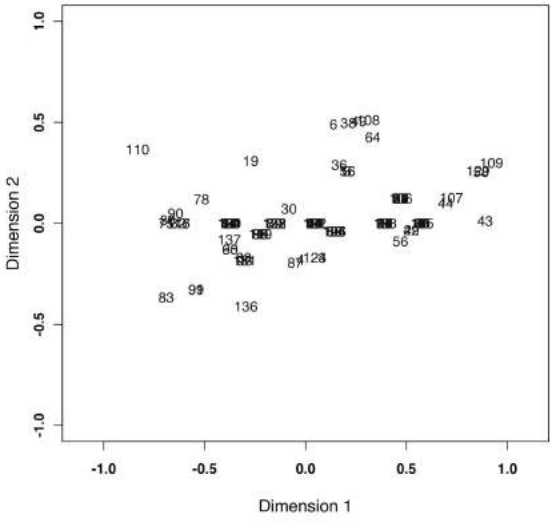


Figure 2: 2-dimensional solution for English. Each number represents one picture in the set.

The addition of more dimensions improved the fit and, even more so, the APRE. The analysis in four dimensions provided the best fit, with 99.5% correct classification, and an APRE of .947 (see Tables 1 and 2).

Mandarin The lowest dimensional fit that provided a high rate of correct classification and a substantial improvement over a null model was again the two-dimensional solution, with 98.8% correct classification and an APRE of .889. The conceptual space associated with this solution is presented in Figure 3.

A close examination of the solution space reveals that the dimension located along the x-axis encodes a continuum between interior location and surface contact. This dimension is anchored at the left end by pictures of a ball underneath an upside-down bowl [110], of a circle surrounded by a rectangle [91], and of a house surrounded by a fence [60], and at the right end by pictures of a garden on the bank of a river [108], of a city on the shore of an ocean [109], of a crease in a pair of pants [86], and of a tree at the top of a hill [65]. The y-axis, on the other hand, encodes variation in the alienability of the two objects. This axis is anchored at the upper end by alienable pairs such as a ball underneath an upside-down bowl [110], a ball under a chair [16], and a garden on the bank of a river [108]; the axis is anchored at the lower end by inalienable pairs such as a curve in a road [88], a tree growing at the top of a hill [65], and a bump in a road [123].

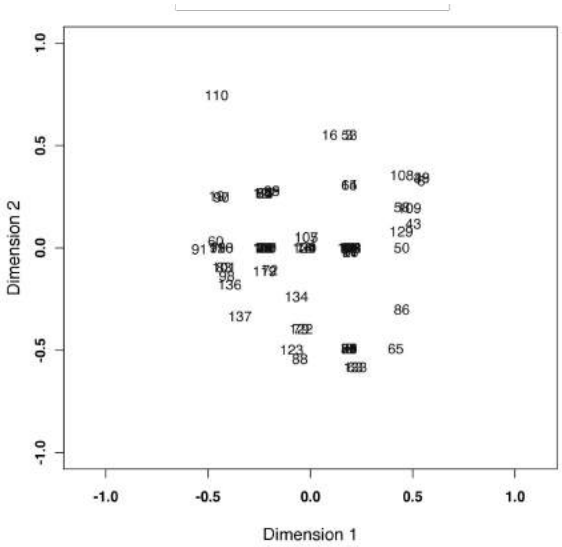


Figure 3: 2-dimensional solution for Mandarin. Each number represents one picture in the set.

The analysis in three dimensions provided the best fit, with 99.4% correct classification, and an APRE of .947, while the gains associated with a higher dimensional fit were more modest (see Tables 1 and 2).

Comparing the solution spaces

There are a number of ways in which we can compare the structurings of the spatial semantic domain in English and Mandarin. At a broad level, we can ask whether the two languages differ in the complexity of the structure of the domain, whereas at a more fine-grained level we can ask whether – and how – conceptual distinctions differ across the two languages.

Turning first to the question of complexity, we compared the two languages with respect to the adequacy of the similarity spaces produced by the MDS analyses. Despite the fact that the two languages yielded comparable numbers of spatial terms (36 in Mandarin; 38 in English), we observed that the Mandarin data was better accommodated at all dimensionalities than was the English data, for both measures of fitness. Table 1 shows the correct classification rates at one, two, three, and four dimensions for both languages. Although the adequacy of the solutions for the two languages was comparable, we note that the Mandarin solution correctly classified a slightly higher proportion of the pictures than did

the English solution at each of the four levels of dimensionality considered. This pattern is replicated for the other fitness statistic, the APRE, for which the differences in adequacy between the Mandarin solutions and the English ones are more pronounced.

Table 1: Correct classification rates for four MDS solutions in both languages

	Mandarin	English
1 dimension	96.4%	95.5%
2 dimensions	98.8%	97.7%
3 dimensions	99.4%	98.6%
4 dimensions	99.7%	99.5%

Table 2: APRE for four MDS solutions in both languages

	Mandarin	English
1 dimension	.659	.551
2 dimensions	.889	.765
3 dimensions	.947	.864
4 dimensions	.976	.947

Pushing this observation farther, we found that the optimal solution in Mandarin was achieved with fewer dimensions than in English, underscoring differences between the two languages in the degree of complexity encoded in topological spatial terms and hinting at differences between the two languages in the semantic structuring of this domain. Notably, the fitness statistics for the three-dimensional Mandarin solution and the four-dimensional English solution were almost identical (99.4% and 99.5% correct classification, respectively, and APREs of .947).

Looking more closely at the placements of the individual pictures, we can ask whether the semantic structurings associated with the two languages differ at a finer-grained conceptual level. To do this, we compared the one-. two-, and three-dimensional semantic spaces across the two languages, asking in each case whether the placements of the pictures along each dimension correlated across the two languages.

We looked first at the one-dimensional solutions, which correctly classified a substantial proportion of the pictures for each language, but presented a relatively modest improvement over a null model. Our analysis revealed a substantial overlap in the placement of pictures along the one-dimensional solution (r = .68, p < .0001), suggesting significant similarity in the ways in which English and Mandarin group situation types in the spatial domain.

In both languages, we observed that the lowest dimensional solution that provided both a high rate of correct classification and a substantial improvement over a null model was the two-dimensional solution, so a comparison of the two-dimensional solutions will be especially important to our understanding of cross-linguistic variation in this domain. At first blush, the English and Mandarin two-dimensional solutions share many similarities: both include one dimension that encodes a continuum between inclusion and surface contact and one dimension that encodes the alienability of the figure-ground relation. However, a closer look reveals that these similarities are but part of the story, co-existing with important differences in the details of the solution spaces.

We consider first the details of the continuum between interior and surface contact. While the English and Mandarin continua overlap, reflected in high correlation between the coordinates along this dimension across the two languages (r = .75, p < .0001), we noted important differences in the placements of many of the pictures in our set. First, we observed that some pictures, such as the crease in pants [86] and the light bulb in a socket [133], are located toward the surface contact end of Mandarin's Dimension 1 but more centrally in the English solution space. In addition, many examples of three-dimensional full inclusion (e.g., an apple in a bowl [2] and a fish in a fishbowl [32]) can be found towards the center of the expanse of Mandarin's Dimension 1, but farther towards the inclusion end of Dimension 1 in English. Looking more closely, we observed that the sets of pictures anchoring the inclusion end of Dimension 1 differed between the two languages: in English, this dimension is anchored by examples of three-dimensional inclusion such as a ball underneath an upside-down bowl [110] and of a muscle in a leg [83], whereas in Mandarin this dimension is anchored by examples of two-dimensional inclusion such as a circle surrounded by a rectangle [91] and a house surrounded by a fence [60]. Whereas all these scenes could be classified as "containment" (cf., Johannes et al. 2015; Landau et al. 2017), these differences suggest that even though both Mandarin and English draw upon a contrast between inclusion and surface contact, the Mandarin system privileges two-dimensional over three-dimensional inclusion, whereas the English system privileges three-dimensional over two-dimensional inclusion. In addition, this dimension is far more spread out in English than in Mandarin, suggesting not only differences in the nature of the inclusion concept, but also differences in the linguistically-relevant degree of similarity amongst the pictures along this dimension.

Turning to the second dimension, we noted less overlap in the semantic interpretation (above), reflected in weaker correlation between the coordinates along this dimension across the two languages (r = .36, p < .0001). Furthermore, whereas this dimension encodes alienability in both solution spaces, this factor is connected to the amount of control exerted by the ground in English, but not in Mandarin. This suggests that control may play a larger role in the semantics of English spatial terms than in the semantics of the Mandarin terms. In addition, this data suggests that the closeness of the relation between two objects may be more likely to be independently assessed and taken into account for speakers of Mandarin than for speakers of English.

Finally, we compared the three-dimensional solutions, which improved substantially beyond the two-dimensional solution in English, but less so in Mandarin. Just as the gains in moving from two dimensions to three differed in the two

languages, we also observed that the placements of the pictures were less congruent in the three-dimensional solutions (dimension 1: $r = .60$, $p < .0001$; dimension 2: $r = .37$, $p < .0001$; dimension 3: $r = .28$, $p = .0023$) than what had been observed for the two-dimensional solutions. This decrement in congruence suggests greater differences between the two semantic systems become evident at more fine-grained levels of analysis.

Conclusions

Our study extends the evidence regarding cross-linguistic variation in the semantics of spatial terms beyond comparisons of the meanings of individual terms by comparing and contrasting the conceptual spaces underlying the semantics of spatial terms in English and in Mandarin. Whereas our findings reinforce the conclusion from past work (i.e., Feist 2008; Levinson & Meira 2003; Levinson & Wilkins 2006) that the conceptual spaces underlying spatial relational inventories are subject to similar factors across languages, this similarity is tempered by evidence that both the factors and the conceptual spaces themselves differ in subtle ways across languages.

We consider first the findings regarding the spatial semantic systems in the two languages. Our Mandarin data was accommodated with fewer dimensions than was our English data, suggesting that the semantic complexity of spatial relational systems varies cross-linguistically. Furthermore, whereas the optimal scaling solution in Mandarin drew upon two dimensions, echoing Feist's (2008) optimal solution for a cross-linguistic dataset, the optimal solution for our English dataset required additional dimensions. This may indicate that each language elaborates on and, hence, may add complexity beyond a universal conceptual core. Thus, while this conceptual core may provide a skeletal structure for how humans think and talk about spatial location (cf., Feist 2008; Levinson & Meira 2003), it markedly underspecifies what we need to encode in order to effectively function in a spatial world.

The strongest correlation between the solution spaces for the two languages involved Dimension 1 of the two-dimensional solutions, which encoded a continuum between interior location and surface contact. In addition to reflecting the importance of the distinction between containment and support (cf., Zhang 2013), this dimension echoes continua that have emerged from other cross-linguistic studies of the semantics of spatial terms, including Bowerman and Pederson's (1992a; see also Bowerman & Choi 2001) similarity gradient and the dimension corresponding to location control in Feist's (2008) MDS analysis. However, whereas Feist's (2008) MDS solution conflated location control and the interior-surface continuum, suggesting that the two may often be inseparable, the English solution reported here separates the two as individual dimensions. This suggests that, whereas both factors are important cross-linguistically and are related, individual languages will make use of different options regarding the extent to which factors are separated in their semantic systems.

Our findings further suggest that cross-linguistic variation may extend beyond fine-grained details of lexical encoding into the nature of the topological concepts themselves. As in the work of Johannes and her colleagues (2015; Landau et al. 2017), our findings underscore the importance of containment and support concepts in the semantics of topological terms. However, a close examination of the continuum between interior and surface contact in the solution spaces for English and Mandarin revealed that the distribution of inclusion scenes differs considerably across the two languages, suggesting differences in the underlying containment concepts. In English, we observed that scenes in which the ground surrounded the figure in three dimensions were placed farther toward the interior end of the continuum than were scenes in which the ground surrounded the figure in two dimensions, suggesting that three-dimensional inclusion is more prototypical than is two-dimensional inclusion in this language. In contrast, in Mandarin we observed the opposite pattern, suggesting that two-dimensional inclusion constitutes a better example of the concept than does three-dimensional inclusion in this language. While inclusion played an important role in structuring spatial semantics in each case, the conceptual cores around which the inclusion concepts were structured differed. Thus, cross-linguistic differences lie not only in the ways terms are distributed relative to a conceptual distinction, but also in the kinds of scenes considered to be best examples of the anchoring conceptual categories.

Whereas MDS allows visualization of the conceptual space underlying semantic systems, it does not afford a picture of the meanings of the lexical items themselves, nor does it afford a close look at the encoding possibilities for individual spatial scenes. The lower correlations observed between the dimensions of the three-dimensional solutions suggest that important cross-linguistic differences may only become evident when viewed at this fine-grained level of analysis. In future work, we will complement the current analysis with analyses focused on these two aspects of spatial semantics. To better understand the detailed ways in which the lexical items relate to one another, we will examine frequencies of use of each term for each picture – i.e., the behavioral profiles (Gries 2010) of each of the elicited terms. To better understand variation in the codability of the scenes, we will assess the breadth of descriptions elicited by each scene. Taken in combination with the current study, these analyses will afford a better understanding of the ways in which individual lexemes fit together to create a semantic system.

Taken together, the current results present a rich picture of the interplay between universals and variation in the semantics of spatial terms. While languages may draw upon a common set of concepts to structure meanings in this domain, these concepts may in fact be quite underspecified. As a result, the variation in meaning that has been observed across languages may be indicative of variation in the ways in which the universal conceptual core has been developed in each language in order to produce a useful set of concepts for communication.

References

Bowerman, M. (1996). The origins of children's spatial semantic categories: Cognitive versus linguistic determinants. In J. J. Gumperz & S. C. Levinson (Eds.), *Rethinking linguistic relativity*. Cambridge, MA: MIT Press.

Bowerman, M., & Choi, S. (2001). Shaping meanings for language: Universal and language specific in the acquisition of spatial semantic categories. In M. Bowerman & S. C. Levinson (Eds.), *Language acquisition and conceptual development*. Cambridge, UK: Cambridge University Press.

Bowerman, M., & Pederson, E. (1992a). Cross-linguistic perspectives on topological spatial relationships. Paper presented at the 91st Annual Meeting of the American Anthrolopogical Association, San Francisco, CA.

Bowerman, M., & Pederson, E. (1992b). Topological relations picture series. In S. C. Levinson (Ed.), Space stimuli kit 1.2. Nijmegen: Max Planck Institute for Psycholinguistics. doi:10.17617/2.883589.

Croft, W. (2010). Relativity, linguistic variation and language universals. *CogniTextes, 4*. Retrieved from http://journals.openedition.org/cognitextes/303.

Croft, W., & Poole, K. T. (2008). Inferring universals from grammatical variation: Multidimensional scaling for typological analysis. *Theoretical Linguistics, 34*: 1-37.

Feist, M. I. (2000). *On in and on: An investigation into the linguistic encoding of spatial scenes*. Doctoral dissertation, Linguistics Department, Northwestern University.

Feist, M. I. (2008). Space between languages. *Cognitive Science, 32*(7), 1177-1199.

Feist, M. I. (2013). Experimental lexical semantics at the crossroads between languages. In A. Rojo & I. Ibarretxe-Antuñano (Eds.). *Cognitive linguistics and translation: Advances in some theoretical models and applications*. Berlin: Mouton de Gruyter.

Gentner, D., & Bowerman, M. (2009). Why some spatial semantic categories are harder to learn than others: The typological prevalence hypothesis. In: J. Guo, E. Lieven, N. Budwig, S. Ervin-Tripp, K. Nakamura and Ş. Özçalişkan (eds.), *Crosslinguistic approaches to the psychology of language: Research in the tradition of Dan Isaac Slobin*. New York: Psychology Press.

Gries, S. Th. (2010). Behavioral profiles: A fine-grained and quantitative approach in corpus-based lexical semantics. *The Mental Lexicon, 5* (3), 323--346.

Johannes, K., Wang, J., Papafragou, A., & Landau, B. (2015). Similarity and variation in the distribution of spatial expressions across three languages. D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp.997 - 1002). Austin, TX: Cognitive Science Society.

Landau, B., & Jackendoff, R. (1993). "What" and "where" in spatial language and spatial cognition. *Behavioral and Brain Sciences, 16*, 217-265.

Landau, B., Johannes, K., Skordos, D., & Papafragou, A. (2017). Containment and support: Core and complexity in spatial language learning. *Cognitive Science, 41* (4), 748-779.

Levinson, S. C., Meira, S., & The Language and Cognition Group. (2003). "Natural concepts" in the spatial topological domain - adpositional meanings in crosslinguistic perspective: An exercise in semantic typology. *Language, 79* (3), 485-516.

Levinson, S. C., & Wilkins, D. P. (Eds.) (2006). *Grammars of space: Explorations in cognitive diversity*. Cambridge: Cambridge University Press.

Poole, K. T. (2000). Non-parametric unfolding of binary choice data. *Political Analysis, 8*, 211-237.

Poole, K. T. (2005). *Spatial models of parliamentary voting*. Cambridge: Cambridge University Press.

Strazny, P. (2005). *Encyclopedia of linguistics*. New York: Fitzroy Dearborn.

Trujillo, A. (1995). Towards a cross-linguistically valid classification of spatial prepositions. *Machine Translation, 10*, 93-141.

Zhang, Y. (2013). *Spatial representation of topological concepts IN and ON: a comparative study of English and Mandarin Chinese.* Doctoral dissertation), Individualized Program, Concordia University.

Zhang, Y., Segalowitz, N., & Gatbonton, E. (2011). Topological spatial representation across and within languages: IN and ON in Mandarin Chinese and English. *Mental Lexicon, 6*(3), 414-445.

An Experimental Protocol to Derive and Validate a Quantum Model of Decision-Making

Lauren Fell (l3.fell@qut.edu.au)

Shahram Dehdashti (shahram.dehdashti@qut.edu.au)

Peter Bruza (p.bruza@qut.edu.au)

Catarina Moreira (catarina.pintomoreira@qut.edu.au)

School of Information Systems, Queensland University of Technology
Brisbane, Australia.

Abstract

This study utilises an experiment famous in quantum physics, the Stern-Gerlach experiment, to inform the structure of an experimental protocol from which a quantum cognitive decision model can be developed. The 'quantumness' of this model is tested by computing a discrete quasi-probabilistic Wigner function. Based on theory from quantum physics, our hypothesis is that the Stern-Gerlach protocol will admit negative values in the Wigner function, thus signalling that the cognitive decision model is quantum. A crowdsourced experiment of two images was used to collect decisions around three questions related to image trustworthiness. The resultant data was used to instantiate the quantum model and compute the Wigner function. Negative values in the Wigner functions of both images were encountered, thus substantiating our hypothesis. Findings also revealed that the quantum cognitive model was a more accurate predictor of decisions when compared to predictions computed using Bayes' rule.

Keywords: quantum cognition; decision-making; complex Hilbert space; binary response; cognitive modelling

Introduction

A generally accepted notion is that we can approximately access cognitive states through questioning and observation, and whilst this measurement is not deemed to be perfect, it is a standard means of experimental practice in the psychological discipline. This notion relies on the fact that these internal states hold distinct values and by measuring them, we are merely attempting to record what is already there. Often, probabilistic outcomes of these measures appear to be illogical and do not follow the laws of classical probability, for example, cognitive biases identified in decision-making (Tversky & Kahneman, 1974). Quantum cognition has emerged as an alternative means of analysing probabilistic outcomes that do not follow these classical laws. Its potential derives from an alternative probability which has successfully been used to address human decision making considered paradoxical, generate non-reductive understandings of human conceptual processing, and provide new understandings of perception and human memory (Bruza, Wang, & Busemeyer, 2015; Busemeyer & Bruza, 2012). The present paper extends current approaches to quantum modelling by means of two new aspects: 1) the Stern-Gerlach experiment, to inform an experimental protocol from which a complex Hilbert space model can be constructed and 2) the discrete Wigner function to perform a check on 'quantumness' of cognitive systems being modelled.

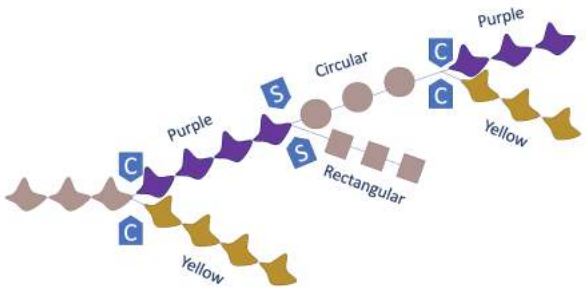


Figure 1: S.G. Setup using colour-type measurements (C) and shape-type measurements (S) in place of spin measurements at different orientations.

The Stern-Gerlach Experiment

The Stern-Gerlach (S.G.) experiment (Sakurai & Commins, 1995) takes a beam of particles (for example, silver atoms) and observes their spin using a device that creates an electromagnetic field (S.G. device). This device can be placed at different orientations to observe spins at associated orientations. Due to the fact that a particle's spin is a complex concept to describe, we will substitute this property with colour and shape in our description of the S.G. experiment in the interests of clarity. For the following, we will describe the experiment as having two orientations of S.G. devices: one oriented one way to measure one type of spin (we will call this a 'colour-type' measure differentiating purple from yellow), and another oriented orthogonal to the first to measure a second type of spin (we will call this a 'shape-type' measure differentiating circular from rectangular). Each enable measurement of separate aspects of the same object (two orientations of spin, or colour and shape in our analogy).

The experiment involves a beam of atoms hitting an S.G. device which splits it into two separate beams: one purple beam and another yellow beam. A second S.G. device of the same orientation as the first (colour-type) is placed in the path of the purple beam. This time, the beam does not split, and

only one beam of purple atoms come out, as one would expect (i.e. we assume no yellow atoms to have entered it due to the separation performed by the first S.G. device). We then place a shape-type S.G. device in-between the two colour-type S.G. devices, which splits the first purple beam into circular and rectangular atoms before hitting the second colour-type device (see figure 1).

In a classical system, we would expect only one beam to emit from the second colour-type device, as, again, we would assume that only purple atoms were sent that way from the first device. A quantum system, however, does not work in this way. In a quantum system, the second colour-type device will emit two beams after the original beam of purple atoms has passed through the shape-type device, as illustrated in figure 1. This is because the shape-type measurement has destroyed the first measurement of colour, essentially resetting it. This can only happen if a particle's properties are not simply observed in a predefined definite state, but are determined at the point of measurement.

A Cognitive Analogue

This concept can be applied to cognitive measures, however, in a slightly more complex way. Due to the potential of memory effects inherent in repeating a question in a string of only three total questions, we utilise a more complex version of the S.G. experiment, where a third question is introduced. The general concept, however, remains the same. When one considers a question or makes a decision, they may simply be accessing an internal state predetermined by a range of variables such as past experience, knowledge, predisposition, values, what they had for lunch that day, etc. On the other hand, they may be creating the state only at the point of measurement (i.e. considering a question or making a decision). To place this in the context of the S.G. experiment, consider three questions asked after presentation of an image: Do you feel a sense of trust when viewing this (T), do you feel that the person in this image is attractive (A), and do you feel that the image may have been manipulated (M). Taking a classical position, one could describe this system in the following way: A person views an image and this event interacts with internal variables to create a variety of probable judgments of this image, including judgments of trust, attractiveness and manipulation. A person then considers the sequential questions of trust, attractiveness and manipulation, each time taking an internal measurement of the predefined values that each of these hold in the person's internal state. On the contrary, taking a quantum position would instead describe no definite judgments to be formed at the point of viewing the image, but only at the point of considering each question. This view would also posit that each question would destroy the measurement of the prior question, in the same way the shape-type S.G. device did in our above example.

This article presents an experimental protocol that is analogous to the S.G. experiment in order to derive a quantum model of decision making. For this purpose a complex Hilbert space is used.

Derivation of a Quantum Model of Decision Making from the S.G. device

As described above, the basic idea behind the model is to translate the S.G. device into cognitive science by way of analogy; human subjects correspond to silver atoms and questions correspond to S.G. devices. As a running example we will use an image trustworthiness task whereby subjects are asked whether they trust (T) an image, whether they find the subject of the image attractive (A) and whether they deem the image to be manipulated (M) e.g., photoshopped. The particular order of questions is determined by the order of the devices in the S.G. device as depicted in figure 2.

The derivation of the quantum model corresponding to the S.G. device comprises two steps: In the first step, a complex Hilbert space model with states and operators is constructed. In the second step, a criterion for checking the 'quantumness' of the model is applied by using a discrete Wigner function.

The cognitive decision space is modelled by means of a complex Hilbert space model (HSM), i.e., a complex vector space, equipped with an inner product with a positive definite metric (Sakurai & Commins, 1995). Any yes/no outcome of a specific question X, is denoted by a ket $|X,\pm\rangle$ using the Dirac notation, where +/- respectively denotes a yes/no outcome :

$$|X\rangle = \alpha|X,+\rangle + \beta|X,-\rangle, \qquad \alpha,\beta \in \mathbb{C}$$

in which $|\alpha|^2$ and $|\beta|^2$, based on the Born rule, give the probability of observing the positive and negative answers. In addition, outcomes are orthogonal, $\langle X,\pm|X,\mp\rangle = 0$ and probabilities are normalized, $|\alpha|^2 + |\beta|^2 = 1$. Also, an observable is defined as a Hermitian operator. Without going into the technical details, a Hermitian operator $\hat{A}$ is a special type of matrix where the eigenstates correspond to outcomes that are observed, and the corresponding eigenvalue relates to the probability of observing that outcome. The Pauli matrices σ_i, $i = 1,2,3$,

$$\sigma_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \sigma_z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix},$$

along with the identity matrix $I_{2\times 2}$, form an orthogonal basis for the complex Hilbert space of all 2×2 matrices. As a consequence, any operator such as $\hat{A}$ can be expressed by $\hat{A} = a_0 I + \sum_{i=1}^{3} a_i \sigma_i$, in which $a_i \in \mathbb{R}$ with $i = 0,1,2,3$.

Steps to construct a complex HSM Based on the preceding formalism, the following steps are used to derive the quantum model from the S.G. depicted in figure 2:

1) A quantum state is defined by the first question T based on relative frequencies of yes/no outcomes sampled from the experimental data, $|T\rangle = \sqrt{P_t(+)}|T,+\rangle + \sqrt{P_t(-)}|T,-\rangle$, in which $P_t(+)$ and $P_t(-)$ are respectively probability of finding positive and negative responses to the question T. Also, we can consistently define the projection or filtering-type quantum cognitive operator $\hat{\pi}_t(\pm) = |T,\pm\rangle\langle T,\pm|$, so that $P_t(\pm) = \langle T|\hat{\pi}_t(\pm)|T\rangle$. The filtering-type operators $\hat{\pi}_t(\pm)$ satisfy the completeness relation, $\hat{\pi}_t^{\dagger}(+)\hat{\pi}_t(+) + \hat{\pi}_t^{\dagger}(-)\hat{\pi}_t(-) = I_{2\times 2}$,

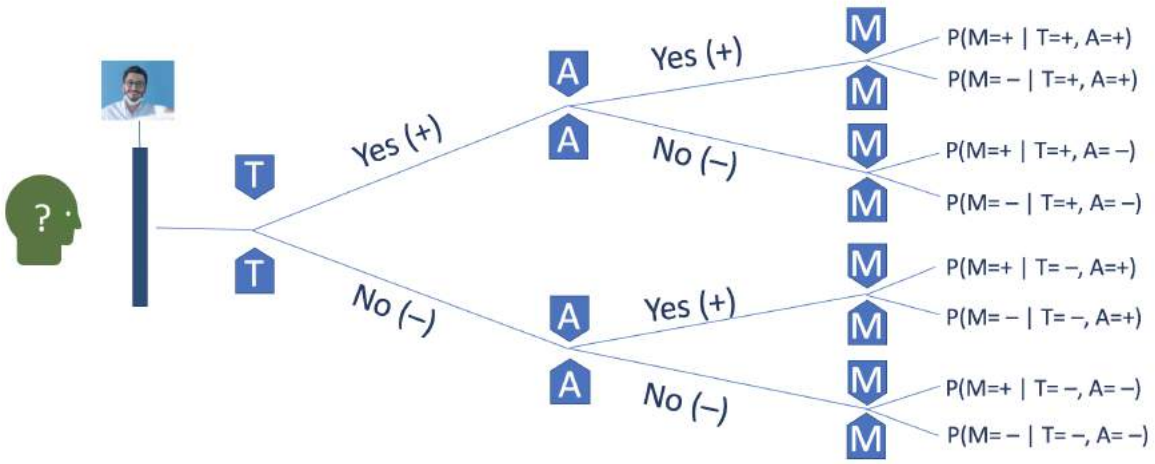


Figure 2: A cognitive analogue to a S.G. experiment where Trustworthiness (T) is asked first, followed by Attractiveness (A), then Manipulated (M).

and the operator $\hat{T}$ is defined as : $\hat{T} = \hat{\pi}_t(+) - \hat{\pi}_t(-) = \sigma_z$, where σ_z is the Pauli matrix in direction z.

2) In the second step, we obtain the probability of finding positive and negative responses to the second question A through the first question T. Hence, we can define the cognitive state regarding a decision of attractiveness A in the basis of the state of trustfulness T, i.e., $|A,+\rangle = \cos\frac{\theta_a}{2}|T,+\rangle + \sin\frac{\theta_a}{2}|T,-\rangle$ and $|A,-\rangle = \sin\frac{\theta_a}{2}|T,+\rangle - \cos\frac{\theta_a}{2}|T,-\rangle$. The filtering-type measurement operators $\pi_a(\pm)$ can be written as follows:

$$\hat{\pi}_a(\pm) = \frac{1}{2}\Big[I_{2\times 2} \pm \sin\theta_a\sigma_x \pm \cos\theta_a\sigma_z\Big].$$

Hence, the operator $\hat{A}$ is given by

$$\hat{A} = \begin{bmatrix} \cos\theta_a & \sin\theta_a \\ \sin\theta_a & -\cos\theta_a \end{bmatrix} \tag{1}$$

in which θ_a characterizes a specific direction, which can be computed from the experimental data. By applying the Born rule, the conditional probabilities can be computed. For example,

$$P(A=+|T=+) = |\pi_a(+)\pi_t(+)|T\rangle|^2 = P_t(+)\cos^2\frac{\theta_a}{2}, \tag{2}$$

3) A similar method to step 2) derives another filtering-type operator corresponding to the third question M,

$$\hat{\pi}_m(\pm) = \frac{1}{2}\Big[I_{2\times 2} \pm \sin\theta_m\cos\phi_m\sigma_x \pm \sin\theta_m\sin\phi_m\sigma_y \pm \cos\theta_m\sigma_z\Big]$$

in which θ_m can be obtained, despite of the fact that we must have extra information for acquiring ϕ_m. Note that the states of third question M are defined as follows:

$$\begin{aligned} |M,+\rangle &= \cos\frac{\theta_m}{2}|T,+\rangle + e^{i\phi_m}\sin\frac{\theta_m}{2}|T,-\rangle, \\ |M,-\rangle &= e^{-i\phi_m}\sin\frac{\theta_m}{2}|T,+\rangle + \cos\frac{\theta_m}{2}|T,-\rangle. \end{aligned}$$

4) In the last step, probabilities of the third question M are computed based in light of the outcomes from the second question A:

$$\begin{aligned} &P(M=+|A=+,T=+) = |\pi_m(+)\pi_a(+)\pi_t(+)|T\rangle|^2 \\ &= P_t(+)\cos^2\tfrac{\theta_a}{2}(+)\Big(\cos^2\tfrac{\theta_a}{2}\cos^2\tfrac{\theta_m}{2} + \sin^2\tfrac{\theta_a}{2}\sin^2\tfrac{\theta_m}{2} \\ &+\tfrac{1}{2}\sin\theta_a\sin\theta_m\cos\phi_m\Big). \end{aligned} \tag{3}$$

By using the previous equations, values of ϕ_m can be computed.

Determining quantumness using the discrete Wigner distribution When we construct the Hilbert space structure of the cognitive state and associated operators, we can examine the quantumness of the cognitive state. Quantum physics has a range of criteria for this. In this article we will employ one such criterium, namely the negative discrete Wigner function, where the negativity of the function can be interpreted as a signature of quantum interference. In order to explain the discrete Wigner function, the continuous Wigner function is first introduced. For a continuous phase space (q,p), the continuous Wigner distribution is defined by

$$W_\Psi(q,p) = \frac{1}{2\pi}\int_{-\infty}^{\infty} dx\langle q - \frac{x}{2}|x\rangle\langle x|q+\frac{x}{2}\rangle e^{ipx}. \tag{4}$$

Therefore, the expectation value of an arbitrary operator $\hat{X}$, by using the Wigner distribution, is given by

$$\langle\hat{X}\rangle = Tr[\hat{\rho}\hat{X}] = \int\int dx\,dp W(x,p)\tilde{X}(x,p), \tag{5}$$

in which $\tilde{X}(x,p)$ is the average of a physical quantity over the phase space. Assuming two arbitrary states $|\psi_a\rangle$ and $|\psi_b\rangle$, it can be verified that:

$$|\langle\psi_a|\psi_b\rangle|^2 = Tr[\hat{\rho}_a\hat{\rho}_b] = \int dx\,dp W_{\psi_a}(x,p)W_{\psi_b}(x,p). \tag{6}$$

If we consider a situation in which two states are orthogonal, i.e.,

$$\int dx\,dp W_{\psi_a}(x,p)W_{\psi_b}(x,p) = 0, \tag{7}$$

at least in part of the region in phase space, one of the above mentioned Wigner distributions has to be negative. The values in Wigner distributions still sum to 1 even when values happen to be negative, which is why Wigner distributions are termed "quasi-probability" distributions. The negativeness of the Wigner distribution can be the result of the following two facts: Firstly, the accessibility of information for a system described by the quantum formalism is when the system is described by classical probability (Goh et al., 2018; Vourdas, 2019). Secondly, the negativeness can be interpreted as quantum contextuality (Huang, Yu, & Zhang, n.d.; Raussendorf, Browne, Delfosse, Okay, & Bermejo-Vega, 2017; Kocia & Love, 2017).

The binary nature of the responses implies a discrete, rather than continuous phase space. We apply a generalized version of the continuous Wigner function (Wootters, 1987). In fact, by defining a geometrical structure on the discrete phase space, such as parallel line, *etc.*, and using a Finite Field $\mathcal{F}_n$, a discrete Wigner distribution can be defined (Gibbons, Hoffman, & Wootters, 2004; Galvao, 2005; Di Matteo, Sánchez-Soto, Leuchs, & Grassl, 2017).

Due to the fact that we have binary responses, the discrete phase space occupies a 2×2 array of points where q runs along the horizontal axis and p runs along the vertical axis, as shown in figure 3. We place the origin, $(q,p) = (0,0)$, at the lower left-hand corner. We define a line λ in the 2×2 phase space as the set of two points satisfying an equation of the form $aq + bp = c$, where a, b, and c are elements of $\mathbb{Z}_2$ ($\mathbb{Z}_2$ constraints numbers to binary 0s and 1s) where a and b cannot both equal zero. It has the following conditions: (i) given any two distinct points, exactly one line contains both points;(ii) given a point α, if a line λ does not contain α, there is exactly one line parallel to λ that does contain it (iii) two lines that are not parallel intersect in exactly one point. In the preceding conditions, two lines can be considered parallel if they can be represented by equations having the same values for a and b but different values for c. In the case of a binary response, therefore, a line connecting $(0,0)$ and $(0,1)$ is parallel with the line that connects $(1,0)$ and $(1,1)$. Moreover, two equations $p+q=0$ and $p+q=1$, with $p,q \in \mathbb{Z}_2$, give the lines connecting points $(1,0)$ and $(0,1)$ and the parallel line connecting $(0,0)$ and $(1,1)$. Finally, the line $(0,0)$ and $(1,0)$ is parallel with the line $(0,1)$ and $(1,1)$. Figure 3 demonstrates these striations in (1), (2), and (3) respectively. Note that the lines drawn in (2) are technically parallel based on the equation described above. As is the case in the continuous phase space, the integral of the Wigner function over the strip of phase space bounded by the lines $aq+bp=c_1$ and $aq+bp=c_2$ is the probability that the operator $a\hat{q}+b\hat{p}$ will take a value between c_1 and c_2 (Wootters, 1987), the discrete Wigner function has to satisfy the following equation:

$$Tr\left(|\alpha_{i,j}\rangle\langle\alpha_{i,j}|\rho\right) = \sum_{\alpha \in \lambda_{i,j}} W_\alpha, \tag{8}$$

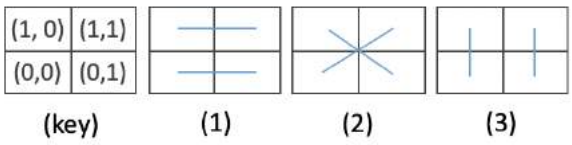


Figure 3: The striations of the 2×2 phase space. Each point occupies a quadrant.

in which ρ is density matrix,

$$\hat{\rho} = \frac{1}{2}\left(I + \vec{r}\cdot\sigma\right) = \frac{1}{2}\begin{pmatrix} 1+r_z & r_x - ir_y \\ r_x + ir_y & 1-r_z \end{pmatrix}. \tag{9}$$

and $|\alpha_{i,j}\rangle$ are three mutually unbiased bases for a two-dimensional Hilbert space, with the following property:

$$|\langle\alpha_{i,j}|\alpha_{k,l}\rangle|^2 = \frac{1}{2} \text{ if } i \neq k, \tag{10}$$

where $i = 1,2,3$ indexes the mutually unbiased bases and $j = 1,2$ indexes the basis vector in each mutually unbiased bases, with the following condition $\langle\alpha_{i,j}|\alpha_{i,k}\rangle|^2 = \delta_{j,k}$. Naturally, we can consider a one-to-one map between Pauli matrices σ_i and striations S_i.

Experiment

Participants

Participants consisted of 300 members of the crowdsourcing platform Prolific, 187 of which were male, 110 female, and 3 who preferred not to disclose their gender. Participants were over 18 years and from a variety of countries across North America (39.7%), Europe (32.3%), UK (22.9%), Australiasia (4.0%) Middle East (0.7%) and Asia (0.3%). Participants were randomly assigned to one of 4 conditions, each with 75 participants. All participants had been verified as proficient in English by Prolific. Remuneration was in the form of a small payment (£.23), as per Prolific convention, and an informed consent page was presented to participants prior to commencement.

Materials

Questions asked were as follows: While viewing, did you feel a sense of trust? (T), Did you feel that this person was attractive? (A), and Did you feel that this image may have been photoshopped? (M). Question orders were *TAM* and *TMA* for each image. Questions were selected based on the likelihood that the operators associated with these variables would be non-commutative. In other words, we were expecting some order effects between variables/operators, meaning that they are not entirely independent of one another. For example, we expect the probabilities associated with the question of attractiveness (A, given T) and the probabilities associated with the question of manipulated (M, given T & A) to be different if the order of A and M were to be reversed (i.e., TMA, with $M|T$, and $A|T$ & M). The image stimuli used to gather ratings of the above dimensions are shown in Figure 4.

(a) Image 1: Unedited.

(b) Image 2: Edited.

Figure 4: Image Stimuli

(A)

P(T=+)	P(A |T=+)	P(M |A,T=+)
P(T=+)=0.85	P(A=+|T=+)=0.69	P(M=+| A=+,T=+)=0.29
		P(M=-| A=+,T=+)=0.40
	P(A=-|T=+)=0.16	P(M=+| A=-,T=+)=0.10
		P(M=-| A=-,T=-)=0.05

(B)

P(T=+)	P(M |T=+)	P(A |M,T=+)
P(T=+)=0.77	P(M=+|T=+)=0.32	P(A=+| M=+,T=+)=0.24
		P(A=-| M=+,T=+)=0.08
	P(M=-|T=+)=0.45	P(A=+| M=-,T=+)=0.37
		P(A=-| M=-,T=+)=0.08

Figure 5: Table (A) and (B) correspond to Image 1 (unedited). Probabilities relating to the first question (T) are depicted in the first columns; the second column states conditional probabilities of the second question given first one, i.e., $P(A = \pm|T = +)$ in Table (A) and $P(M = \pm|T = +)$ in Table (B); the third column indicates conditional probabilities of the third question given the first and second questions.

Design

Each image was presented with two question orders, creating a between subjects design with four conditions. The dependant variables were ratings of trustworthiness, attractiveness and image manipulation.

Procedure

In all conditions, participants completed an online experiment by first perusing a short description on the Prolific site, if deciding to continue, they then clicked a link to the project page which begins with short instructions and a link to the consent form to read before continuing. The design of the experiment was aimed at accessing fast intuitive responses, rather than responses based on analytical thinking, as this was believed to be analogous to the short distances between measurement devices in the S.G. experiment (i.e. fast measurements restricting interacting influences). To this end, instructions included a notice to look out for a button popping up for some participants that afforded a bonus (distraction to assign less cognitive resources to the decision task), questions were asked with emotive wording (to help prompt intuitive thinking), and both image display and questions included a time limit (2 seconds for the image and 4 seconds for each question). Participants could only view one question at a time, with each subsequent question hidden until an answer had been given for the preceding one. Lastly, participants were asked to provide one or two words to describe their first impressions of what they saw, as well as a confidence rating for their combined judgments, and were asked their gender and the country they resided in.

(A)

P(T=+)	P(A |T=+)	P(M |A,T=+)
P(T=+)=0.85	P(A=+|T=+)=0.69	P(M=+| A=+,T=+)=0.29
		P(M=-| A=+,T=+)=0.40
	P(A=-|T=+)=0.16	P(M=+| A=-,T=+)=0.10
		P(M=-| A=-,T=-)=0.05

(B)

P(T=+)	P(M |T=+)	P(A |M,T=+)
P(T=+)=0.77	P(M=+|T=+)=0.32	P(A=+| M=+,T=+)=0.24
		P(A=-| M=+,T=+)=0.08
	P(M=-|T=+)=0.45	P(A=+| M=-,T=+)=0.37
		P(A=-| M=-,T=+)=0.08

Figure 6: Table (A) and (B) correspond to Image 2 (edited). Probabilities relating to the first question (T) are depicted in the first columns; the second column states conditional probabilities of the second question given first one,i.e., $P(A = \pm|T = +)$ in Table (A) and $P(M = \pm|T = +)$ in Table (B); the third column indicates conditional probabilities of the third question given the first and second questions.

Results

Based on the probabilities shown in Table 5 (A), (B) and Table 6 (A), (B), the cognitive states associated with the first question T are given by:

$$|T_1\rangle = \sqrt{0.85}|T_1,+\rangle + \sqrt{0.15}|T_1,-\rangle, \quad (11)$$

$$|T_2\rangle = \sqrt{0.59}|T_2,+\rangle + \sqrt{0.31}|T_2,-\rangle. \quad (12)$$

where the subscripts respectively denote the unedited and edited images.

According to probabilities in the second column in Table 5 (A), (B) and also by using the equation (2), the angles between operator $\hat{T}$ and $\hat{A}$, as well as $\hat{T}$ and $\hat{M}$, are respectively given by $\theta_a^{(1)} = 51.42$ and $\theta_m^{(1)} = 99.79$ for the unedited image. By using the same method, in the second column of Table 6 (A), (B), the angles between the operators $\hat{T}$ and $\hat{A}$, as well as $\hat{T}$ and $\hat{M}$, are obtained respectively by $\theta_a^{(2)} = 71.20$ and $\theta_m^{(2)} = 87.70$ for the edited image. By using equation (3) together with the third column of Table 5 (A), (B) and Table 6

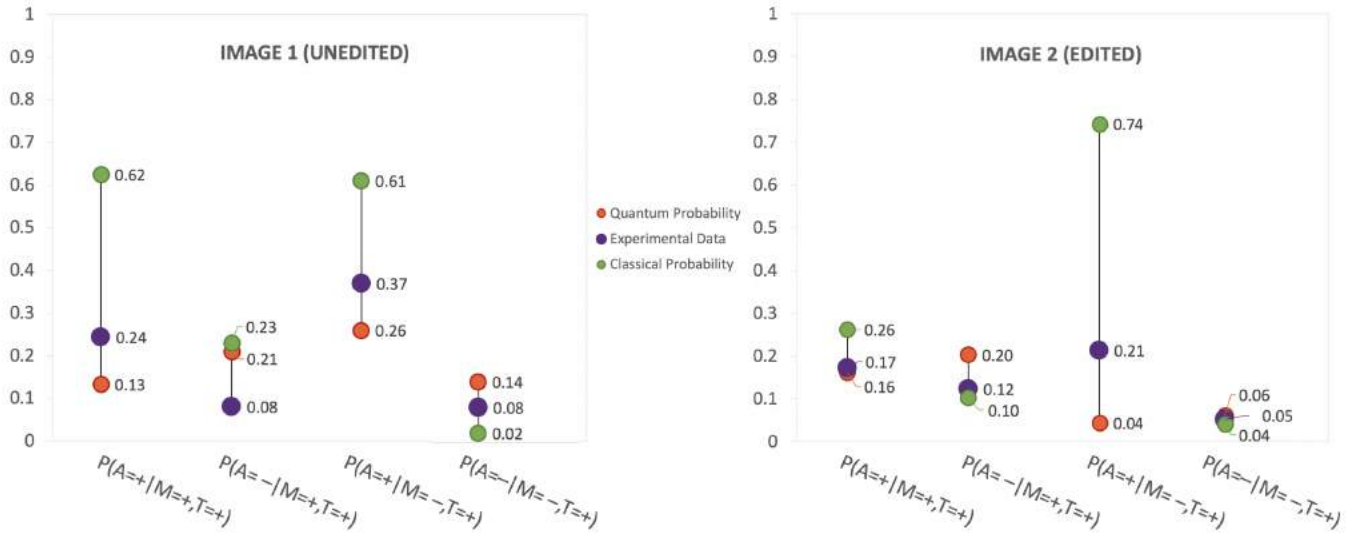


Figure 7: Comparison of actual probabilities (Experimental Data) of the third question Attractiveness (A) to predicted probabilities computed by the Born rule (Quantum Probabilities) and Bayes' rule (Classical Probabilities).

(A), (B), we obtain respectively $\phi_m^{(1)} = 85.99$ and $\phi_m^{(2)} = 85.30$ for the unedited and edited images.

Comparison of prediction of decision: quantum probabilities vs. classical probabilities For the prediction comparison, we consider a new situation in which the the order of questions is altered. According to operators $\hat{A}$ and $\hat{M}$ and new preparation state $|T\rangle$, we obtain the probability of positive and negative answers and compare them with the experimental data. Indeed, as defined in equation (3), the phase interference ϕ_m appears in the probability of the third question. The predictions of the quantum model are compared to classical probabilities in the following way: The cognitive S.G. device depicted in Figure 2 can be modelled by using the chain rule: $P(T,A,M) = P(T)P(A|T)P(M|A,T)$. The three distributions on the RHS are empirically collected from the S.G. device. Similarly, in the new situation the order of the A and M magnets are reversed so the chain rule is written out as follows: $P(T,M,A) = P(T)P(M|T)P(A|M,T)$. Therefore,

$$P(M|A,T) = \frac{P(M|T)P(A|M,T)}{P(A|T)} \quad (13)$$

$$P(A|M,T) = \frac{P(A|T)P(M|A,T)}{P(M|T)} \quad (14)$$

The LHS of both equations constitute predictions based on classical probability theory. As evidenced by Figure 7 (A) and (B), the predicted results calculated based on the HSM are generally closer to the actual probabilities than the classical predictions. Figure 7 compares results of probabilities of the decision regarding manipulation given attractiveness and trustworthiness, for the unedited and edited images respectively.

Wigner functions for both images By using equation (9):

$$r_x = 2\sqrt{P_t(+)\left(1-P_t(+)\right)},\ r_y = 0,\ r_z = 2P_t(+)-1,$$

The discrete Wigner distribution that is obtained is the following:

$$W = \frac{1}{4}\begin{pmatrix} 1+r_x+r_z & 1-r_x+r_z \\ 1-r_x-r_z & 1+r_x-r_z \end{pmatrix} \quad (15)$$

Therefore, the Wigner distributions for both unedited (W_1) and edited (W_2) images are given as follows:

$$W_1 = \begin{pmatrix} 0.63 & 0.13 \\ -0.13 & 0.36 \end{pmatrix}, \quad W_2 = \begin{pmatrix} 0.53 & 0.03 \\ -0.03 & 0.47 \end{pmatrix} \quad (16)$$

Discussion

The Wigner function of both images showed negative values. Therefore, the cognitive analogue of the S.G. experiment that produces quantum models in physics, also produces a quantum model for cognitive decision making.

It is known from physics that negative values in the Wigner function are a consequence of quantum interference effects. The negative values are a consequence of the fact that once a particle has passed through a magnet its polarization (either + or -) is not retained when it arrives at the next magnet. This is a consequence of the the fact that a particle is always in a superposed state each time it interacts with a magnet. As a result of the interaction, a particular polarization will be observed. In terms of the cognitive analogue depicted in Figure 2, the preceding can be translated as follows: Even though a subject has already decided that they trust (T=+) the image and have deemed the face to be attractive (A=+), when they are presented with the decision about whether the image is manipulated, at that decision point they are necessarily superposed with respect to trust and attractiveness. This can only occur when the decision perspectives are incompatible. Incompatibility is indeed present in the HSM as the operators corresponding to decisions of trustworthiness $\hat{T},\hat{A},\hat{M}$ do not pair-wise mutually commute: $[\hat{T},\hat{A}] \neq 0$, $[\hat{T},\hat{M}] \neq 0$, $[\hat{A},\hat{M}] \neq 0$.

Incompatibility generates interference effects which gen-

erate probabilities of outcomes, that is they are fundamentally different from standard probabilistic models (Bruza et al., 2015). As stated above, the cognitive S.G. device depicted in Figure 2 can be modelled by using the chain rule: $P(T,A,M) = P(T)P(A|T)P(M|A,T)$. This expresses that the underlying probabilistic model of the device is simply the joint probability distribution $P(T,A,M)$. The critical point is that the structure of the event space underpinning $P(T,A,M)$ assumes that the variables are *jointly* measurable, e.g., the subject can simultaneously access information regarding the attractiveness of the face and whether the image is manipulated. The previously mentioned incompatibility in the HSM $[\hat{A},\hat{M}] \neq 0$ implies that this assumption does not hold. Consequently, the subject cannot cognitively form the joint distribution $P(T,A,M)$. In short, the HSM provides a probabilistic framework which does not rely on the assumption that variables are jointly measurable. This has been one of the key features of quantum models of cognition (Busemeyer & Bruza, 2012).

The use of three operators is crucial in the derivation of the two-dimensional Hilbert space because three operators necessarily entail that a *complex* Hilbert space must be used. The use of less than three operators necessarily implies that the cognitive decision model can be expressed as a real-valued Hilbert space, which has been the practice thus far in quantum cognition research. The significance of this difference lies in the complex phase factor $\exp(i\phi_m)$ which cannot be derived unless there are three operators. We speculate that it is this phase factor which generates the interference effects for the Wigner function to go negative and hence become quantum. To the best of our knowledge, this study is the first to: a) develop a specialised protocol to genuinely exploit the complex Hilbert space by constructing three operators and states, and b) utilise the Wigner function to determine the quantumness of a cognitive state.

Moreover, this determination is straightforward and does not suffer from the challenges and controversies associated with using contextuality to determine whether the cognitive system is quantum-like (Dzhafarov, Kujala, Cervantes, Zhang, & Jones, 2016; Bruza & Fell, 2018).

Conclusions and future work

This article has demonstrated the specification and validation of a quantum decision model by employing an experimental protocol derived from quantum physics. The protocol involved three binary decisions in a forced choice design. Future studies may investigate the measurement of decisions by asking binary questions of any number of points within a spectrum of responses by extending the quantum model described in this paper.

Acknowledgments

This research was supported by the Asian Office of Aerospace Research and Development (AOARD) grant: FA2386-17-1-4016

References

Bruza, P., & Fell, L. (2018). Are decisions of image trustworthiness contextual? a pilot study. In A. Lambert-Mogiliansky & B. Coecke (Eds.), *Quantum interaction: 11th international conference (qi'2018).* Springer.

Bruza, P., Wang, Z., & Busemeyer, J. R. (2015). Quantum cognition: a new theoretical approach to psychology. *Trends in Cognitive Sciences*, *19*(7), 383 - 393.

Busemeyer, J., & Bruza, P. (2012). *Quantum cognition and decision*. Cambridge University Press.

Di Matteo, O., Sánchez-Soto, L. L., Leuchs, G., & Grassl, M. (2017, Feb). Coarse graining the phase space of *n* qubits. *Phys. Rev. A*, *95*, 022340.

Dzhafarov, E., Kujala, J., Cervantes, V., Zhang, R., & Jones, M. (2016). On contextuality in behavioural data. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*.

Galvao, E. F. (2005). Discrete wigner functions and quantum computational speedup. *Physical Review A*, *71*(4), 042302.

Gibbons, K. S., Hoffman, M. J., & Wootters, W. K. (2004). Discrete phase space based on finite fields. *Physical Review A*, *70*(6), 062101.

Goh, K. T., Kaniewski, J., Wolfe, E., Vértesi, T., Wu, X., Cai, Y., ... Scarani, V. (2018). Geometry of the set of quantum correlations. *Physical Review A*, *97*(2), 022104.

Huang, M.-D., Yu, Y.-F., & Zhang, Z.-M. (n.d.). The negativity-to-violation map between wigner function and quantum contextuality inequality for a single qudit. *Annalen der Physik*, 1800464.

Kocia, L., & Love, P. (2017). Discrete wigner formalism for qubits and noncontextuality of clifford gates on qubit stabilizer states. *Physical Review A*, *96*(6), 062134.

Raussendorf, R., Browne, D. E., Delfosse, N., Okay, C., & Bermejo-Vega, J. (2017). Contextuality and wigner-function negativity in qubit quantum computation. *Physical Review A*, *95*(5), 052334.

Sakurai, J. J., & Commins, E. D. (1995). *Modern quantum mechanics, revised edition.* AAPT.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *science*, *185*(4157), 1124–1131.

Vourdas, A. (2019). Probabilistic inequalities and measurements in bipartite systems. *Journal of Physics A: Mathematical and Theoretical*.

Wootters, W. K. (1987). A wigner-function formulation of finite-state quantum mechanics. *Annals of Physics*, *176*(1), 1–21.

Exploring the use of overhypotheses by children and capuchin monkeys

Elisa Felsche[1] (ef68@st-andrews.ac.uk),
Patience Stevens[2] (pstevens@andrew.cmu.edu),
Christoph Völter[3] (christoph.voelter@vetmeduni.ac.at),
Daphna Buchsbaum[4*] (buchsbaum@psych.utoronto.ca)
and Amanda Seed[1*] (ams18@st-andrews.ac.uk)

[1]School of Psychology and Neuroscience, University of St Andrews, Scotland
[2]Department of Psychology, Carnegie Mellon University, USA
[3]Messerli Research Institute, University of Veterinary Medicine Vienna, Austria
[4]Department of Psychology, University of Toronto, Canada

Abstract

The use of abstract higher-level knowledge (overhypotheses) allows humans to learn quickly from sparse data, and make predictions in new situations. Previous research has suggested that humans may be the only species capable of abstract knowledge formation, but this remains controversial, and there is also mixed evidence for when this ability emerges over human development. Kemp et al. (2007) proposed a computational model of overhypothesis formation from sparse data. We provide the first direct test of this model: an ecologically valid paradigm for testing two species, capuchin monkeys (Sapajus spp.) and 4-5-year-old human children. We compared performance to predictions made by models with and without the capacity to learn overhypotheses. Children's choices were consistent with the overhypothesis model predictions, whereas monkeys performed at chance level.

Keywords: Overhypotheses, abstraction, generalization, animal cognition, computational modeling, cognitive development

Introduction

For long-lived species that exploit a complex environment it might be beneficial to transfer adaptive behavior across situations, through the formation of abstract generalizations. For example, if a primate learns that one tree grows figs, a second papaya and a third nuts, at a more abstract level she is also exposed to the regularity: "Trees carry a uniform fruit type". Learning this abstraction would make just one bite of fruit from a new tree sufficient to decide whether or not continued foraging in this tree would be beneficial.

In the developmental literature the term 'overhypotheses' (Goodman, 1955) describes such higher-order generalizations at an abstract level that inform inferences about more specific hypotheses (Kemp, Perfors, & Tenenbaum, 2007). Kemp et al. (2007) developed a computational model that suggested that, in principle, overhypotheses can be learned quickly from sparse data and used to make wide-ranging predictions in new situations.

Evidence for a possible early emergence of this ability during human infancy comes from a study using looking-time methodology. Dewar and Xu (2010) presented 9-month-olds with sampled evidence supporting the overhypothesis that containers are filled with objects of the same shape. In a test situation, infants looked longer when two differently shaped objects were drawn from the same container, contradicting this overhypothesis, than when two uniformly shaped objects were sampled.

Despite this evidence for early overhypothesis formation, other methods show contrasting results. A common method to assess understanding of the abstract concepts "same" and "different" is the relational matching-to-sample (RMTS) task. Here, participants are presented with an example stimulus pair (either two of the same or two different items) as well as two test pairs, and must select the pair with the matching abstract relation to the example. Hochmann et al. (2017) showed that children begin to succeed in a 2-item RMTS task by the age of 5 but not earlier (see Kotovsky & Gentner, 1996 for a similar result). However, labeling the relations verbally enables children to succeed in the RMTS task as early as age 2 (Christie & Gentner, 2014).

In contrast, in an anticipatory looking time procedure, Hochmann, Carey and Mehler (2018) showed that 7 and 12-month-olds were sensitive to the abstract relation of same but not different. Similarly, 18- to 30-month-old children correctly selected either a matching or a dissimilar pair of objects following evidence that their relation was causally relevant (Walker & Gopnik, 2014).

In addition, only a few non-human species master the RMTS task, usually after lengthy training regimes (e.g. Truppa, Mortari, Garofoli, Privitera, & Visalberghi, 2011; see also Smirnova, Zorina, Obozova & Wasserman (2015)), and often only with multi-stimulus arrays instead of stimulus pairs (see Wasserman & Young, 2010 for a review; the latter also helping 3-year-olds succeed, Hochmann et al., 2017). As a result, some have suggested that the RMTS task can be solved by perceptual processes alone, and that abstract knowledge is a uniquely human capability (Penn, Holyoak, & Povinelli, 2008; Vonk, 2015). In a different set of tasks, chimpanzees and bonobos have been suggested to use relative spatial relations such as "top" or "middle" to find hidden food rewards (Haun & Call, 2009; Christie, Gentner, Call & Haun, 2016). However, it is not clear whether searching based on relative rather than absolute

[*] Equal contribution

spatial relations represents the same kind of abstract knowledge as concepts such as "same" and "different". In summary, the question of whether abstract knowledge formation is an evolutionary primitive, shared with other species and emerging early in human development, or a recently-evolved, late-developing skill, is complicated by considerable methodological differences between the tasks used across ages and species. Further, as in other areas of cognitive development, there is something of a dissociation between looking time results that suggest an early-emerging conceptual competence, and later emerging success on choice-based measures by older children. One concern is that successful discrimination in infant looking time tasks may not require the same kind of conceptual competence as paradigms requiring participants to use their knowledge to make a choice (e.g., Hood, 2004).

We therefore designed a task that could be used across species, to examine abstract knowledge formation in an ethologically valid context without extensive training or explanation, based on the original idea of overhypothesis formation by Goodman (1955). Importantly this allowed us to test a theoretical computational model for how limited data can be sufficient for overhypothesis formation in this task (Kemp et al., 2007). Similar to Dewar and Xu's (2010) infant looking time study, we adapted Goodman's thought experiment, in which bags of marbles can be either uniform or mixed in color, to create a choice paradigm suitable for older children and capuchin monkeys. We presented sampled evidence from three containers either supporting the overhypothesis that rewards are sorted by their *size* or by their *type*. At test, participants were presented with two new containers and one example item from each: a small, high-valued reward from A and a large, low-valued reward from B (Figure 1). Participants then chose between two covert samples from these new containers. Differential choice between conditions—namely, choosing A to obtain a high-valued option in the type condition, but choosing B to obtain a large item in the size condition—would reflect sensitivity to the overhypotheses governing object sorting.

Computational Model

Probabilistic hierarchical Bayesian models have frequently been proposed as computational models of children's rapid early learning (Kemp et al., 2007; Tenenbaum, Kemp, Griffiths, & Goodman, 2011). They demonstrate how, in principle, knowledge can be acquired at multiple levels of abstraction simultaneously, after only seeing small amounts of data. Kemp et al. (2007) show how more abstract hypotheses can constrain the hypothesis space at lower levels, leading to rapid inferences when encountering new but related situations. Due to the interdependence of concrete observations and higher-order concepts, these models do not exhibit the tension between low-level and higher-level learning often discussed in the animal literature. However, while the Kemp et al. model has successfully characterized existing findings in the developmental literature, the model's predictions have not been directly empirically tested in children or animals.

Here, we extended the Kemp et al. (2007) model with a rational choice rule, allowing us to directly compare the model's predictions for which test container (A or B) learners should choose to receive a reward from with new empirical data. We infer the relative utilities of the different reward types, based on the participants' choices in preference testing, following the inverse preference model developed by Lucas et al. (2014).

Model Overview.

Figure 1 provides an overview of both our task and of the computational model. In this model, items are sampled from evidence containers, each of which has a distribution of items with different features (i.e., item type and size). These distributions capture a first level of abstract knowledge (level 1), describing the kinds of items likely to be found in this specific container. Simultaneously, the model also represents a more abstract level of knowledge (level 2), which describes the probability distribution over containers—the extent to which containers in general tend to be mixed or uniform, and the distribution of features across containers. Using this hierarchical structure, the model captures how specific observations of samples from individual containers can be used to simultaneously infer parameters at multiple levels of abstraction.

Learning Overhypotheses.

As in Kemp et al., (2007) we use a Dirichlet-multinomial model (Gelman, Carlin, Stern & Rubin, 2003). The individual sees evidence items $\boldsymbol{y^i}$ with d feature dimensions (in our case d = 2: the item's type and size), sampled from each evidence container i. We assume that items are drawn randomly and independently from each container and that the item's type is determined independently of its size. The item types (sizes) are sampled from $\boldsymbol{y_d^i} \sim$ Multinomial ($\boldsymbol{\theta_d^i}$), the distribution over item types (sizes) in that container. Each container's distribution over item types (sizes), $\boldsymbol{\theta_d^i}$, is in turn sampled from a Dirichlet distribution, parameterized by a scalar α_d and a vector $\boldsymbol{\beta_d}$, $\boldsymbol{\theta_d^i} \sim$ Dirichlet(α_d, $\boldsymbol{\beta_d}$). These hyperparameters characterize the overhypothesis across containers. α_d parameterizes the extent to which items in each container are uniform in type (size). $\boldsymbol{\beta_d}$ represents the type (size) distribution across the entire set of containers. α_d is in turn sampled from an exponential distribution, $\alpha_d \sim$ Exponential(1), and $\boldsymbol{\beta_d}$ from a symmetric Dirichlet distribution, $\boldsymbol{\beta_d} \sim$ Dirichlet(1).

To model overhypothesis formation, we infer $p(\alpha_d, \boldsymbol{\beta}_d | \boldsymbol{Y_d})$ (referred to as $p(\alpha, \boldsymbol{\beta} | \boldsymbol{Y})$ for simplicity below), the posterior distribution over (α, $\boldsymbol{\beta}$), given the observed items $\boldsymbol{y^i}$, drawn from the N evidence containers,

$$p(\alpha, \boldsymbol{\beta} | \boldsymbol{Y}) \propto \int \prod_{i=1}^{N} p(\boldsymbol{y^i} | \boldsymbol{\theta^i}) \boldsymbol{p}(\boldsymbol{\theta^i} | \alpha, \boldsymbol{\beta}) p(\alpha) p(\boldsymbol{\beta}) \, d\theta \qquad (1)$$

estimated using the Metropolis-Hastings algorithm. Here we used 5 chains with 2000 samples and a burn in of 1000.

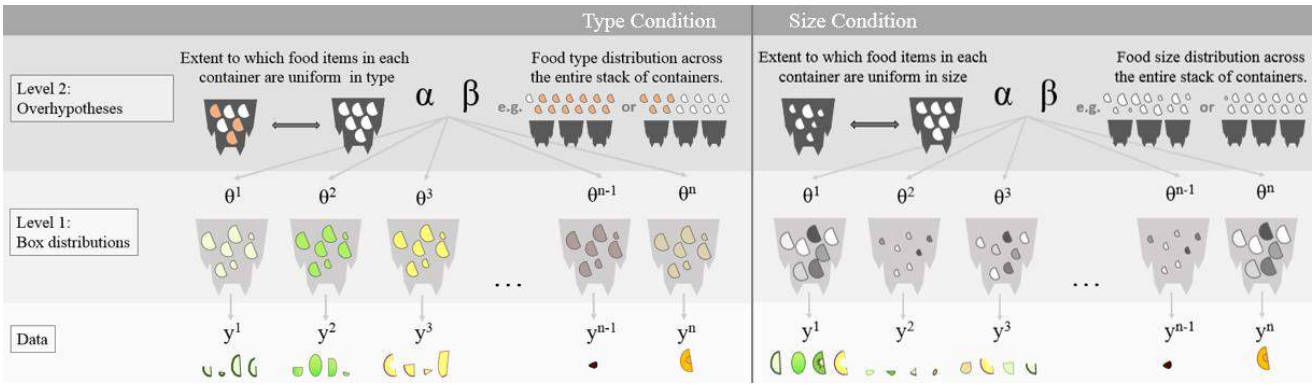


Figure 1: Hierarchical Bayesian model of overhypothesis formation. The parameters α and **β** describe an overhypothesis at the second level of abstraction: α represents the extent to which containers in general tend to be uniform for a given feature dimension, and **β** captures the feature variability across all containers. Feature distributions of a specific container ($\boldsymbol{\theta}^{\mathbf{i}}$, Level 1 abstraction), are constrained by overhypotheses at Level 2, and in turn constrain the items $\boldsymbol{y}^{\boldsymbol{i}}$ sampled from that container.

Predicting the content of the test buckets

We would like to predict the type (size) of the j^{th} (unseen) sample from the new test container *i+1*, given already known samples from this test container, -j (everything not j), and the overhypotheses inferred from the evidence containers. For a Dirichlet-Multinomial distribution, $p(y_j^{i+1} | \boldsymbol{y}_{-\boldsymbol{j}}^{\boldsymbol{i+1}}, \alpha, \boldsymbol{\beta})$, the posterior predictive distribution for the type (size) of the next item in the container, given the previously seen items and the hyperparameters $\alpha, \boldsymbol{\beta}$, has a simple closed form solution. Marginalizing over $p(\alpha, \boldsymbol{\beta}|\boldsymbol{Y})$, the posterior distribution over possible values of α and **β**, estimated from the evidence containers give us,

$$p(y_j^{i+1} | \boldsymbol{y}_{-\boldsymbol{j}}^{\boldsymbol{i+1}}) = \iint p(y_j^{i+1} | \boldsymbol{y}_{-\boldsymbol{j}}^{\boldsymbol{i+1}}, \alpha, \boldsymbol{\beta}) p(\alpha, \boldsymbol{\beta}|\boldsymbol{Y})\, d\alpha, d\boldsymbol{\beta} \qquad (2)$$

Approximated by averaging $p(y_j^{i+1} | \boldsymbol{y}_{-\boldsymbol{j}}^{\boldsymbol{i+1}}, \alpha, \boldsymbol{\beta})$ across sampled values of $p(\alpha, \boldsymbol{\beta}|\boldsymbol{Y})$.

Predicting choice of test item

Given the distribution over possible next items from each test container, we would like to predict the learner's choices. We assume that learners are choosing which box to take the next item from based on the expected utility of the next item from each container. As in Lucas et al. (2014), we assume that the utility of an item *x* is just the product of the utility of its individual features. For simplicity we assume that utility scales linearly with item size, s_x, so that the utility of item *x*, is $u_x = s_x \cdot \delta_{t_x}$, where δ_{t_x} is the learner's utility for one unit of item type t_x. The utility of a container is calculated by summing the utilities of each possible item, weighted by its probability of being the next item. As in previous work, we assume that learners become exponentially more likely to choose a container *i* as its expected utility increases.

$$P(c = i|u) = \frac{e^{u_i}}{\sum_j e^{u_j}} \qquad (3)$$

Inferring reward utilities.

To compute the relative utilities of the different reward items, prior to the main experiment, we conducted a series of preference tests, where participants chose which of two reward items they wanted. For simplicity, we only included the categorical item types high, medium and low-value. Comparisons included choices between different item types of fixed size, between different sizes of the same type, as well as mixed comparisons between large items of low value and small items of high value.

Following the preference inference model described in Lucas et al. (2014), we assume that learners choose items based on their relative utilities as in equation 3. We infer item type utilities **u** from learner's choices **c,** separately for each species, by computing the posterior probability $p(u|c) \propto p(c|u)p(\boldsymbol{u})$, estimated using the Metropolis-Hastings algorithm. Following Lucas et al. (2014), we assume that the type preferences **δ** are normally distributed, with $\mu = 0$, and variance $\sigma^2 = 2$ (however the inferred preferences are robust to different values of σ^2). Here we used one chain with 10000 samples and a burn in of 500.

Model Predictions.

Using this approach, we inferred strong preferences for high vs low value items for both species (children: Δ0.62; monkeys: Δ1.19). We used each species item utilities, separately inferred from their preference task data, to make *a priori* choice predictions for our experiment. Model predictions based on Level 2 abstraction (abstraction across containers) make clear contrasting choice predictions between the size and type conditions for both species after only one trial (one set of 3 evidences containers; Figure 2a). Predictions across subsequent trials, after seeing up to 6 sets of evidence containers get asymptotically more extreme (Figure 2b). In contrast, for a lesioned model capable of only Level 1 abstraction, and thus not learning from the evidence containers, the test container with the small, high value item is the preferred choice independent of condition.

Experiment 1: Abstraction across containers

Methods

Participants. Participants were 80 4- to 5- year-old children (M age = 4.9 yrs, 50% female), recruited at two local museums in Toronto, Canada. Eight additional children were excluded from analysis because they ended the game early (n = 5) or due to experimenter error (n=3). Seventeen brown capuchin monkeys (*Sapajus spp.*, M age = 6.5 yrs, 29% female) completed a preparatory food preference testing. Due to motivation decline only 11 monkeys finished the main study and are included in the data analysis.

Materials & Procedure. For the monkeys, nine different types of food items (divided into 3 categories: high, medium and low value) and 5 item sizes were used. Rewards for children were stickers picturing either animals (high value) or simple shapes (low value). Size was manipulated by the number of stickers on a strip, varying from 1 to 5. To encourage consistent sticker preferences across children, they were given the task of filling in a zoo map with as many animals as possible, making animal stickers more valuable than shape stickers. Prior to the main experiment, both species received preference testing, details of this procedure are given below. All sessions were video recorded.

Main Experiment. For both species the procedure in each trial was very similar. The experimenter successively sampled four example items from each of 3 evidence containers into transparent cups (monkeys), or onto metal frames (children), starting always on the left side. Depending on the condition, the items from one container were either all of the same type but of varying sizes (type condition) or all identical in size but different in type (size condition, see Figure 1). During the sampling, the experimenter closed her eyes and kept her head upright to create the illusion of random sampling.

Subsequently, two new test containers were brought forward, with the other containers and their evidence still in view of the participants off to the side. The experimenter first simultaneously sampled one evidence item from each test container. This was always a small, high-valued reward from container A and a large, low-valued reward from container B (item types counterbalanced). The experimenter then sampled another item from each container simultaneously, this time keeping the reward items hidden in her closed hands. The closed hands were extended towards the participants so that they could indicate their choice by reaching towards one of the hands. Participants were rewarded with the chosen item. Reward items were chosen to be in line with the condition overhypothesis (i.e., of the expected type or size), at least of medium value in the size condition, and otherwise randomly sampled.

For monkeys, the experimenter crossed her hands in half of the trials (a procedure they are familiar with) to ensure they tracked the hidden sample in the experimenter's hand and were not just pointing towards the sampled items. For children, hands were never crossed. In comparison to the monkeys, children's pointing was not restricted by a choice panel and thus they were able to clearly indicate a specific hand rather than only a side (unlike the monkeys children also had no prior experience with this procedure and showed confusion about the hands crossing in a pilot study).

Due to the small available sample, monkeys experienced both conditions, size and type, in a within-subject ABAB design, with the first condition counterbalanced across monkeys. Here, two different kinds of containers, bags and boxes (counterbalanced), were used, so that any overhypothesis could be tied to a specific kind of container. Monkeys received 16 sessions with 3 trials each, with 4 sessions per block. Children were tested in a between-subject design to allow us to test them in a single session in a science museum. and thus only presented with one container type (boxes). They received one session of 6 trials. Importantly, as for the monkeys, they did not receive any explicit instruction concerning the abstract rules governing the reward distribution.

Reward Preference Testing. Prior to the main experiment, we conducted preference testing to ensure that participants preferred bigger over smaller (size comparisons) and high over low-value rewards (type comparisons). Further small, high-value items were compared to large, low-value items (mixed comparisons). Monkeys received 9 kinds of size comparisons, one for each food type. There were also 6 kinds of type and mixed comparisons respectively, as each of the three high-value items was compared to two low-value items. Finally, the least liked high-value item was compared to all 3 medium valued items to ensure a clear preference. Monkeys received 10 trials for each of the 24 comparisons, presented over 24 sessions. Food items were presented in a covered forced choice procedure, where the monkeys first saw the food on the experimenter's palms and then had to choose between her closed fists.

Children first received a warm-up of 3 trials in which they were familiarized with the closed-hands choice procedure. Due to the constraints of museum testing, children were presented with a reduced preference procedure of two preference trials each for the type and size comparisons. A subset of n=58 children also received two mixed trials. Following preference testing, for the main experiment, novel stickers were used, and children were asked to find a lot of animals for a new, blank zoo map.

Results

Reward Preference Testing. In the type comparisons, both species significantly preferred high-value items over equally sized low-value alternatives (Capuchins: M = 0.86, SD = 0.12, $t(16)=11.78$, $p<0.001$; Children: (M = 0.94, SD = 0.18, $t(79) = 22.33$, $p < 0.001$). Capuchins further preferred the least liked high value item over equally sized pieces of medium-valued foods (M = 0.89, SD = 0.06, $t(16)=25.07$, $p<0.001$). Both groups also significantly preferred large over small items (Capuchins: M = 0.83, SD = 0.06, $t(16)=23.96$, $p<0.001$; Children: M = 0.83, SD = 0.32, $t(79) = 9.11$, $p < 0.001$). In the mixed comparisons

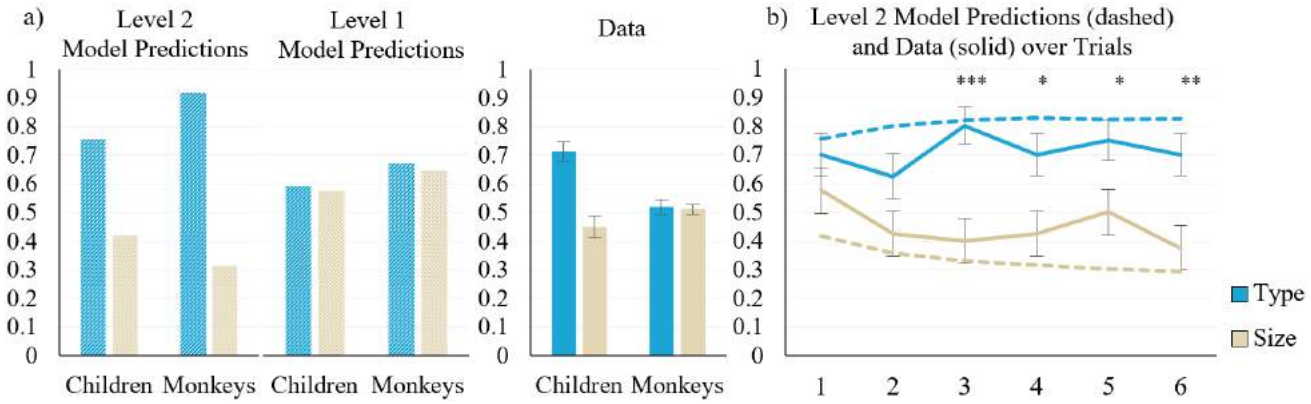


Figure 2: a) Model Predictions for a learner capable of Level 2 or Level 1 abstraction and empirical results (mean across trials ± SE) for the choice for the sample from the box with the small, high-valued example item for capuchin monkeys and children. Model predictions are shown for one trial with 3 evidence boxes. b) Children's level 2 model predictions and data (M ± SE) over the course of six trials. Significant differences between the size and the type condition are indicated.

both groups expressed a significant preference for the small, high-valued items over the big, low-value option (Capuchins: M=0.96, SD = 0.04, t(16) = 48.52, p <0.001; Children: (M = 0.92, SD = 0.21, t(57) = 15.68, p < 0.001). Further, no difference in performance was found between the choice presentation with crossed and straight hands.

Main Experiment. Monkeys were equally likely to choose the sample from the container with the small high-value example in both conditions (paired t(10)= 0.27, p = 0.79), and chose at chance level (12/24 trials) between the two hidden samples (type: M = 12.45, SD = 1.37, t(10) = 1.10, p = 0.30; size: M = 12.27, SD = 1.95, t(10) = 0.46, p = 0.65).

Unlike the preference testing, multiple monkeys expressed a bias regarding the side of their chosen reward sample or the side of the container (7/11 monkeys chose either a consistent hand-side or a consistent container-side in more than 80% of trials). They did not reach more frequently to the side of the small, high-valued sample (M = 0.52). There was no improvement from the first block to the second in either condition (type: M_{first} = 0.52, M_{second} = 0.52; size: M_{first} = 0.51, M_{second} = 0.51).

Children chose the sample from the container with the small, high-value example item more often in the type condition than the size condition, t(77.50) = -5.18, p < 0.001. When compared to chance (3/6 trials), only the choices in the type condition were significantly different (type: M = 4.28, SD = 1.41, t(39) = 5.70, p < 0.001; size: M = 2.7, SD = 1.30, t(39) = -1.45, p = 0.15).

The Level 2 models for both species predict a clear distinction between both conditions in the tendency to choose the item from the container with the small, high-value example item (Figure 2). Choice predictions are stronger for monkeys as the inferred utilities for low and high-value items based on their reward preferences are more extreme. When compared to the empirical data, the monkey's chance level performance is in stark contrast to the predictions of a model that learns overhypotheses, using item utilities inferred from the monkey's food preferences. For children the level 2 overhypothesis model predictions qualitatively fit the data well and the trajectory across trials shows a similar trend for both data and model predictions.

Discussion

As predicted by the model fit separately to their preference data, children made different choices in the size and type conditions despite the evidence from the test containers being the same in both cases, suggesting that they formed overhypotheses. However, their performance only differed significantly from chance in the type condition, which could suggest that they are only capable of forming abstract rules about certain reward properties. Alternatively, children might have a strong prior towards sorting by type, which is possibly more common in children's experience, or the two features might have had an unequal salience based on pre-existing preferences or the task description (see also Kemp et al., 2007 for discussion of the 'shape bias' in word learning). However, children did show a preference for larger items when presented with a simple choice in the preference test, suggesting they attended to this dimension. Interestingly, the overhypothesis model fit to children's preferences also predicted a smaller distinction from chance in the size condition, suggesting that this result may nonetheless be consistent with the overhypotheses. Future work could try to increase sample size or change utilities to differentiate lack of attention to the size dimension from a smaller predicted difference in utility between containers.

The monkeys' performance suggests that they were not able to form overhypotheses about the food distribution pattern across containers. Their failure on the second level of abstraction could be due to a failure to form abstractions about containers in general (Level 2 overhypotheses), or based on the inability to infer the content distributions of each evidence container (Level 1 overhypotheses) based on the sampled evidence. However, as with any negative result from a complex task, there could be other limiting factors, specifically the sampling procedure required sustained attention and inhibition skills which could be an impeding factor for the performance of monkeys (Tecwyn, Denison,

Messer, & Buchsbaum, 2017), a point we will return to in the general discussion.

Experiment 2: Abstraction within a container

To test the hypothesis that monkeys did not form Level 1 generalizations about the contents of the containers in experiment 1 (precluding generalization over containers – Level 2), we conducted a second experiment, with reduced task demands. Here, we presented subjects with only two containers from which we sampled four evidence items each. Now, the choice items were sampled directly from these containers, so that no generalization to new containers (Level 2) was required. However, participants would still have to form Level 1 generalizations to choose successfully.

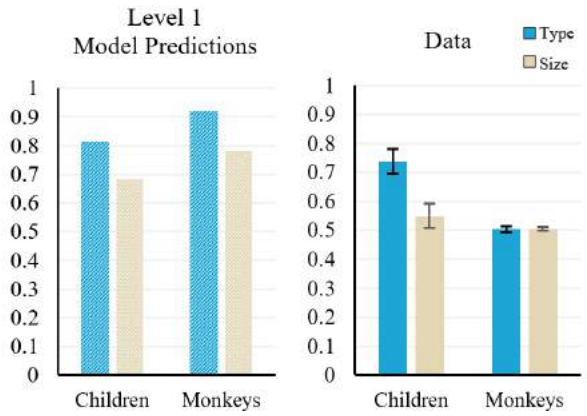


Figure 3: Model predictions (left) and empirical data (right, M ± SE) for the correct choices of children and monkeys in the type (high-value item) and size (large item) condition.

Methods

Participants. Participants were 47 4- to 5- year-old children recruited at two local museums in Toronto (M age = 5.0 yrs, 50% female, n= 24 in type condition, n= 23 size condition). Two additional children were excluded because they ended the game early or due to experimenter error. The total sample of capuchin monkeys (*Sapajus spp.*) consisted of 13 individuals. Ten had previously completed Experiment 1. Out of the 13 subjects, 11 participated in both conditions whereas two participated only in one of the conditions.

Design and Procedure. All sessions were video recorded. The procedure was similar to Experiment 1. This time only two containers were presented on the table and four items were sampled from each successively. Subsequently, the experimenter extracted the two choice items directly from these containers, kept them hidden in her hand and requested the participant to choose. In the size condition, the same four types of rewards, two low- and two high-value, were drawn from both containers in a randomized order whereby one container only yielded small (size 1) and the other one only big (size 5) items. The reward was identical to one of the four types previously drawn from the container. In the type condition, items of the same type in the sizes 1, 2, 4 and 5 were drawn from the container. Thereby one container offered only low-valued and the other only high-valued items. The reward was a randomly sized piece of the expected type for this container. Monkeys received 3 sessions of 8 trials each per condition with order of condition counterbalanced. Children received one session of 6 trials in a between-subject design beginning with a preference testing of two size and two type comparisons.

Results and Discussion

Children performed significantly above chance (3/6 trials) in the type condition ($M = 4.42$, $SD = 1.28$, $t(23) = 5.41$, $p < 0.001$) but not in the size condition ($M= 3.30$, $SD = 1.22$, $t(23) = 1.19$, $p = 0.25$). Monkeys performed at chance level (12/24 trials) in both conditions (type: $M = 12.09$, $SD = 0.94$, $t(10) = 0.32$, $p = 0.76$; size: $M = 12.09$, $SD = 0.54$, $t(10) = 0.56$, $p = 0.59$). The choice predictions of models based on the inferred feature distribution in each container (Level 1 abstraction) showed a clear tendency to choose the next item from the container with high-value items in the type condition and from the container with large items in the size condition. The strong type preferences of both species, lead to a greater predicted container preference in the type condition. Whereas the monkeys performed at chance level in both conditions, children's performance resembled the model prediction in both conditions, showing strong performance in the type condition whereas choices in the size condition were at chance. This suggests that children are able to form abstractions at both levels whereas capuchin monkeys in our study were unable to engage in any level of abstraction, though we emphasize that the reasons for this failure remain ambiguous (lack of ability or task demands).

General Discussion

We presented two studies testing abstraction, and the predictions of the Kemp et al. (2007) overhypothesis model, using a choice paradigm in children and capuchin monkeys. Across both experiments, none of the capuchin monkeys showed the pattern predicted for a learner capable of forming overhypotheses along the item size or type dimensions. In contrast, children treated the same evidence differently when they had previously experienced that items were sorted by size or type. Their performance was well characterized by a hierarchical Bayesian model, fit to their actual reward preferences. They showed a significant difference between conditions after just a few trials.

The model predictions based on capuchin's preferences support that the presented evidence was sufficient for the formation of overhypotheses, but the monkeys did not show this ability in this paradigm. The monkeys' results are in line with low success rates achieved after long training regimes in previous studies on abstract concept formation and analogical reasoning in capuchins (Flemming, 2011; Kennedy & Fragaszy, 2008; Truppa et al., 2011). We can also rule out some other possible explanations for their failure. Monkeys did not show a preference for the side exhibiting the small, high-value item (showing some

understanding of the procedure: they were not simply trying to acquire the samples). No individual monkey showed a difference between conditions, and the sample was sufficient to detect significant food preferences, suggesting that this was also not a sample size limitation.

Still, it remains possible that other tasks demands masked monkeys' abstract reasoning abilities. Monkeys and apes can infer a hidden item sampled from a clear population (Tecwyn et al., 2017; Eckert, Rakoczy, & Call, 2017; Rakoczy et al., 2014). However, apes' ability to make inferences about hidden populations based on visible samples (as in this study) was recently shown to be more limited (Eckert, Rakoczy, & Call, 2017). Future work will explore abstract reasoning with reduced task demands, e.g. by allowing the subjects to sample items themselves. Nevertheless, the approach taken here, in which subjects do not need to be trained to make arbitrary judgements about abstract relations but simply need to secure the best rewards, is a promising avenue for future research.

The findings from 4-5 year-olds are in line with infants' performance in causal learning and looking time procedures but stand in contrast to children's limited spontaneous use of abstract concepts in RMTS tasks (Hochmann et al., 2017), perhaps due to a reduced need for training. This approach could be extended to toddlers to bridge the gap across ontogeny.

In summary, we conducted the first direct test of the hierarchical Bayesian approach described by Kemp et al. (2007) in children and animals, and extended it to make choice predictions based on item utilities. We have shown that it is a promising model for how children are able to form generalizations from sparse evidence. We suggest that further application of computational models to empirical data of overhypothesis formation is desirable to understand its development over early childhood, and to further understanding possible species differences.

Acknowledgements

We thank Justine Biado, Kiah Caneira and Kay Otsubo. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. [639072]). We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), [funding reference number 2016-05552]

References

Christie, S., & Gentner, D. (2014). Language helps children succeed on a classic analogy task. *Cognitive Science*, *38*(2), 383-397.

Christie, S., Gentner, D., Call, J., & Haun, D. B. M. (2016). Sensitivity to relational similarity and object similarity in apes and children. *Current Biology*, *26*(4), 531-535.

Dewar, K. M., & Xu, F. (2010). Induction, overhypothesis, and the origin of abstract knowledge evidence from 9-month-old infants. *Psychological Science*.

Eckert, J., Rakoczy, H., & Call, J. (2017). Are great apes able to reason from multi-item samples to populations of food items?. *American journal of primatology*, *79*(10), e22693.

Flemming, T. M. (2011). Conceptual thresholds for same and different in old-(Macaca mulatta) and new-world (Cebus apella) monkeys. *Behavioural processes*, *86*(3), 316-322.

Gelman, A., Carlin, J.B., Stern, H.S., & Rubin, D.B. (2003).Bayesian data analysis (2nd edn.). New York: Chapman & Hall.

Goodman, N. (1955). *Fact, fiction and forecast* (Vol. 74). Cambridge, MA: Harvard University Press.

Haun, D. B., & Call, J. (2009). Great apes' capacities to recognize relational similarity. *Cognition*, *110*(2), 147-159.

Hochmann, J. R., Tuerk, A. S., Sanborn, S., Zhu, R., Long, R., Dempster, M., & Carey, S. (2017). Children's representation of abstract relations in relational/array match-to-sample tasks. *Cognitive psychology*, *99*, 17-43.

Hochmann, J. R., Carey, S., & Mehler, J. (2018). Infants learn a rule predicated on the relation same but fail to simultaneously learn a rule predicated on the relation different. *Cognition*, *177*, 49-57.

Hood, B. M. (2004). Is looking good enough or does it beggar belief?. *Developmental Science*, *7*(4), 415-417.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental science*, *10*(3), 307-321.

Kennedy, E. H., & Fragaszy, D. M. (2008). Analogical reasoning in a capuchin monkey (Cebus apella). *Journal of Comparative Psychology*, *122*(2), 167.

Kotovsky, L., & Gentner, D. (1996). Comparison and categorization in the development of relational similarity. *Child Development*, *67*(6), 2797-2822.

Lucas, C. G., Griffiths, T. L., Xu, F., Fawcett, C., Gopnik, A., Kushnir, T., ... & Hu, J. (2014). The child as econometrician: A rational model of preference understanding in children. *PloS one*, *9*(3), e92160.

Penn, D. C., Holyoak, K. J., & Povinelli, D. J. (2008). Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. *Behavioral and Brain Sciences*, *31*(2), 109- ' 130.

Rakoczy, H., Clüver, A., Saucke, L., Stoffregen, N., Gräbener, A., Migura, J., & Call, J. (2014). Apes are intuitive statisticians. *Cognition*, *131*(1), 60-68.

Smirnova, A., Zorina, Z., Obozova, T., & Wasserman, E. (2015). Crows spontaneously exhibit analogical reasoning. *Current Biology*, *25*(2), 256-260.

Tecwyn, E. C., Denison, S., Messer, E. J., & Buchsbaum, D. (2017). Intuitive probabilistic inference in capuchin monkeys. *Animal Cognition*, *20*(2), 243-256.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *science*, *331*(6022), 1279-1285.

Truppa, V., Mortari, E. P., Garofoli, D., Privitera, S., & Visalberghi, E. (2011). Same/different concept learning by capuchin monkeys in matching-to-sample tasks. *PLoS One*, *6*(8), e23809.

Vonk, J. (2015). Corvid cognition: Something to crow about?. *Current Biology*, *25*(2), R69-R71.

Walker, C. M., & Gopnik, A. (2014). Toddlers infer higher-order relational principles in causal learning. *Psychological Science*, *25*(1), 161-169.

Wasserman, E. A., & Young, M. E. (2010). Same–different discrimination: The keel and backbone of thought and reasoning. *Journal of Experimental Psychology: Animal Behavior Processes*, *36*(1), 3.

Individual differences in fluency with idea generation predict children's beliefs in their own free will

Teresa Flanagan (tmf87@cornell.edu)
Department of Human Development, Cornell University, 116 Reservoir Ave,
Ithaca, NY 14853 USA

Tamar Kushnir (tk397@cornell.edu)
Department of Human Development, Cornell University, 116 Reservoir Ave,
Ithaca, NY 14853 USA

Abstract

The ability to imagine alternative possibilities plays a crucial role in everyday cognitive functioning beginning in early childhood. Across two studies, we ask whether individual differences in young children's (Mean Age = 5.01; SD = 0.78 Range = 2) fluency in generating alternative possibilities relates to a particular type of social-cognitive counterfactual judgment, namely children's belief in the possibility to "act otherwise" when actions go against stated strong desires (i.e. "free will"). We found that the fluency of generating ideas was a consistent individual difference that held regardless of domain. We also found that individual children's fluency predicted judgments of free will for themselves (Study 2) but not for others (Study 1). Our findings raise new questions about how counterfactual thinking enables children to overcome psychological barriers to self-control, and how stimulating the imagination facilitates developing cognitions that rely on it.

Keywords: counterfactual thinking, free will, social cognition, modal cognition

Introduction

The ability to imagine alternative possibilities is ubiquitous in human cognition. Broadly, it is invoked in all types of modal thinking: how we imagine counterfactually what could have been in the past, hypothetically what might be in the future, and normatively what should or ought to be (Lewis, 1973; Balke & Pearl, 1994; Woodward & Hitchcock, 2003). Imagining possibilities is a critical cognitive skill underlying our memory for past events (Schacter et al., 2015), ability to plan for the future (Baumeister, Vohs, & Oettingen, 2016), our moral judgments (Phillips & Cushman, 2017; Phillips, Luguri, & Knobe, 2015), and our causal cognition (Engle & Walker, 2018). Moreover, this type of thinking is governed by a common cognitive and neural architecture (De Brigard et al., 2013). Recently, researchers have claimed that it plays a role in a host of psychological processes that develop in early childhood, including future thinking (Atance & Meltzoff, 2005), causal inference (Walker & Gopnik, 2013; Engle & Walker, 2018), imaginary play (Taylor et al, 2018; Weisberg & Sobel, 2012), self-regulation (White et al., 2017), and social and moral judgment (Kushnir, 2018).

Separate from this, there has been a long research tradition focused on ability to imagine alternative possibilities as a stable individual difference, relating it to differences in creativity and intelligence. Most of this work utilizes a classic method developed by Guilford (1967) in which people are asked to generate many unique alternative possible uses for a common object (e.g. a tissue). Conservatively, these tasks capture individual differences in verbal fluency, performance on these "idea generation" paradigms also relates to individual differences in creativity and intelligence (Wallach & Kogan, 1965; Nusbaum & Silva, 2011). There is recent evidence that modified versions of tasks such as Gilford's capture stable individual differences in children as well, even when controlling for age and verbal IQ (Taylor et al., 2018).

To date, however, no studies have linked individual differences in "idea generation", either in adults or children, to the types of cognitions that have been hypothesized to rely on modal thinking. Here we explore one such link: we examine whether individual differences in the ability to generate alternative ideas relate to a particular social-cognitive skill that relies on counterfactual thinking – children's judgment of their own and others' freedom of choice.

Counterfactual thinking has been argued to be the basis of folk intuitions of freedom of choice (i.e. "free will", Alquist et al., 2015; Nichols, 2011). Studies show that when adults make free will judgments, they consider whether or not there were alternative choices available (Feldman, Baumeister, & Wong, 2014). Children's early developing intuitions about free will are also based on the ability to represent alternative possibilities. For example, infants are more impatient with an agent when the agent is unwilling to act (they understand that a possible alternative is available) than when the agent is unable to act (they understand that no possible alternative is available, Behne et al., 2005). Preschoolers can answer explicit questions about whether an agent can and can't do otherwise when actions are possible, impossible, or limited by social and moral considerations (Nichols, 2004; Schult & Wellman, 1977; Shtulman & Phillips, 2008). As part of this ability, children generate explanations about what alternative actions are available if an agent chooses to "do otherwise."

Children's beliefs about free will also undergo important developmental changes, changes that may be linked to their counterfactual thinking. For example, 6-year-olds are more likely to endorse the free will to act against desires than 4-year-olds (Kushnir et al., 2015). Furthermore, older children

are more likely than younger children to endorse the freedom to act against moral and social norms (Chernyak et al., 2013; Chernyak, Kushnir, & Wellman, 2010). Moreover, younger children have difficulty distinguishing improbable from impermissible events more generally (Lane et al., 2016; Shtulman & Carey, 2007; Shtulman & Phillips, 2008). Together these studies suggest that a domain-general cognitive mechanism is responsible for the developmental shift.

One intriguing possibility is that the ability to *fluently generate ideas* about possible alternative actions underlies children's judgments of free will. That is, in order to make judgments of free will (or possibility more generally) children are attempting to imagine any situations in which an action could be different, and, if they can think of one or more such situations readily and easily, they answer in the affirmative. For example, a child may be able to easily imagine how yummy crackers could be inedible for all sorts of reasons, thus they answer that one can choose not to eat them. Anecdotally, this hypothesis has some support from the post-hoc justifications that children come up with following the initial yes/no judgment. The large majority of their justifications are imagined alternative scenarios (i.e. "because some crackers aren't good for you" Kushnir et al, 2015). Under this model, individual differences in free will judgments should relate to individual differences with a facility for idea generation.

To test this, we conducted two studies using the third person (Study 1) and first-person (Study 2) versions of the free will questions from Kushnir et al. (2015). Overall, we hypothesized that children's free will judgments would replicate prior work, such that there would be some variability in children's free will beliefs, and also age-related changes. Like prior work, we expect lower free will beliefs (and more variability) for first-person question (Study 2) than for third-person (Study 1).

We also measured children's ability to generate alternative possibilities using a battery of idea generation tasks. Our tasks had a structure modeled after standard creativity task (e.g. "uncommon uses task" Milgram & Milgram, 1976; Wallach & Kogan, 1965) with some notable differences. First, we scored children on idea fluency only. That is, we did not compare each idea to the sample-wide list of ideas to score its originality. Fluency was captured by coding for number of unique ideas listed a child gives for a particular question.

Second, we asked children to list as many alternative possibilities they could think of in each of three different domains – Physical, Fantastical, and Social/Psychological. The Physical question was adapted from Guilford (1967) and asked about alternative uses for a tissue. The Fantastical questions were adapted from Taylor et al. (2018) asking children to imagine what the world would be like if certain laws of our world were changed (ability to walk on walls, people have tails). We added to these our own set of Social/Psychological questions which asked children to come up with ways to help a sad friend feel better (social) or make themselves feel better when they are sad (psychological).

Including a range of idea generation tasks across these domains allowed us to explore the generality of the relationships: we were able to check whether children who generate a lot of ideas in one domain also do so in others. Moreover, Since children's free will beliefs are part of their developing social-cognition, our addition of questions which explore children's social and psychological idea generation allowed us to check whether social/psychological ideas are specifically related to free will beliefs over and above other types of ideas.

Study 1

Method

Participants A total of 43 4–6-year-olds (M_{age} = 5.07, SD_{age} = 0.80, N_{female} = 23) were recruited at a science museum in a small city in the northeastern United States.

Procedure Children were interviewed individually in a quiet room at a museum. The procedure began with the free will questions, adapted from Kushnir et al. (2015). Two Action questions (one about food and one about activity) asked children to judge whether an agent could "choose to" act against or whether they "have to" act in accordance with the stated desire (e.g. "Even though she does not like the cracker, can she just *choose to* eat the cracker or does she *have to not* eat the cracker?"). Two Inhibition questions (one about food and one about activity) asked children to judge whether an agent could choose to *not* act (i.e. inhibit action) or whether they have to act on a stated strong desire (e.g. "Even though she wants to know about the box, can she just *choose not to* look into the box or does she *have to* look into the box"). Note that the free will questions offer children a forced choice between a stated action (e.g. eating a yummy cookie) and a general possibility of acting otherwise without explicitly stating any alternative actions. Question order was counterbalanced. Order of the options within the question ("choose to" vs "have to") was counterbalanced. There were also two Control questions, Simple free action (e.g. "Can she step off a chair?") and Physically Impossible action (e.g. "Can she run through a wall?"). The majority of children (86%) answered these control questions correctly, ensuring children understood and could follow the form of the target questions about acting against desires.

After the free will questions, children completed the idea generation task battery. This began with a warm-up question about uncommon uses for a pen ("Besides [drawing/writing] can you think of some things to do with [pen]?") that was used to familiarize children with the question format and idea probes ("what else?"). Following the warm-up, the idea generation task battery consisted of three question types in a latin-square counterbalanced order:

Physical (one, Gilford 1967): Children were asked for the common use of, and then ideas for uncommon uses of, a tissue ("Can you think of some things you can do with it besides [common use]?").

Fantastical (two, Taylor et al, 2018): Children were asked to "imagine what if we all woke up tomorrow and every person [had a tail/could walk on walls]. What would the world be like if we all [had tails/could walk on walls]?"

Social/Psychological (one third-person, one first-person). Children were asked to think of ways to make a friend /themselves happy when the friend/they themselves are sad. (e.g. "Imagine that one day [your friend/you] [was/are] very sad and didn't want to play at all. What things could you do to make [your friend/yourself] feel better and want to play again?).

For each question, children were encouraged using the probe "what else?" to keep generating ideas until they chose to stop (e.g. saying they had no more ideas). Children also participated in a storytelling task at the end of the procedure, but those results are not reported here.

Coding For the free will task, children received a score of 0-2 for each story type, Action and Inhibition (2 meaning they said "choose to" for both food and activity questions). Two coders independently scored each response. A Cohen's k indicated agreement between the two coders for each question (κs > .83, *ps* < .0005). Discrepancies were resolved through discussion.

For the idea generation tasks, the number of unique responses were recorded for each question. Uniqueness was defined as any difference from previous responses (e.g. "we could wag our tails" versus "we could bounce on our tails"). Two coders were trained on the coding scheme. A Cohen's k indicated agreement between the two coders for each question (κs > .867, *ps* < .0005). Discrepancies were resolved through discussion.

Results

Free Will A repeated measures ANCOVA with Question Type (Action vs Inhibition) as a within-groups factor and age as a covariate found a marginal main effect of Question Type ($F(1) = 3.17$, $p = .083$) and a marginal age effect ($F(1) = 3.097$, $p = .086$) and no interaction. Replicating past work, children's free will scores were higher for Action ($M = 1.45$, $SD = 0.78$) than in Inhibition ($M = 1.13$, $SD = 0.82$; $t(39) = 2.177$, $p = .036$). In addition, scores were significantly above chance for Action ($t(39) = 3.636$, $p = .001$) but not Inhibition ($t(41) = .927$, $p = .359$). We found a significant positive correlation between age and Inhibition score ($r = .376$, $p = .014$), but not age and Action score ($r = .088$, $p = .590$).

Idea Generation Table 1 shows the descriptive statistics for each idea generation task as well as the correlations between them. The results point to group-level and individual consistency across domains. On the group level, a repeated measures ANCOVA with Question Type (Physical vs Fantastical_Tails vs Fantastical_Walls vs Social vs Psychological vs Total) and age as a covariate found no effect of Question Type ($F(1) = 2.58$, $p = .117$), no effect of age ($F(1)=.013$, $p = .911$), and no interaction. Also, the number of unique ideas generated was positively correlated across domains (*ps* < .05); children who generated more in one domain tending to generate more in another.

Relationship between Free will beliefs and Fluency Children received a score of 0-4 for total free will judgments (combined Action and Inhibition scores). Total free will score did not significantly correlate with any of the idea generation scores separately or total idea generation score (see Table 1).

Table 1: Relationships between idea generation questions across domains in Study 1. Relationship between idea generation and third-person free will judgments included in final row.

	Physical	Fantastical Tails	Fantastical Walls	Soc/Psych third-person	Soc/psych first-person	Total number of ideas generated
Physical ($M = 4.73$, $SD = 4.62$)	-	.500**	.571***	.728***	.382*	
Fantastical Tails ($M = 4.70$, $SD = 5.06$)	-	-	.367*	.365*	.645***	
Fantastical Walls ($M = 4.28$, $SD = 3.83$)	-	-	-	.410**	.374*	
Social/psychological third-person ($M = 4.29$, $SD = 5.44$)	-	-	-	-	.420**	
Social/psychological first-person ($M = 2.79$, $SD = 2.55$)	-	-	-	-	-	
Third-person Free Will	.075	.037	.144	-.112	-.060	.032

*p<.05; **p<.01, ***p<.001

Discussion

In this study, we examined children's third-person free will judgments, children's ability to fluently generate alternative possibilities, and the relationship between two. Patterns of children's free will judgment by age and type of question (action vs inhibition) mirrored past work. Children generated an average of 4-5 ideas per domain (with the exception of first-person social/psychological ideas). Idea generation was consistent across Physical, Fantastical, and Social/Psychological domains, both at the group level and at the individual level. Our findings of cross-domain consistency in idea generation suggest that these tasks, if properly validated (e.g. by controlling for verbal IQ, see Taylor et al, 2018), could be used to measure fluency in children.

We did not find significant relationships between idea generation scores and third-person free will judgments. One reason could be that there was not enough variability in free will judgments, which were relatively high in this study even for the youngest children.

More substantively, taking this third-person view may facilitate children's reasoning about possible actions in itself, and thus our third-person task may not be demanding enough to demonstrate the role of individual differences. Recent work has shown that, like adults, children are subject to such "psychological distance" effects when reasoning about possibility, choice, and future desires (Bowman-Smith, Shtulman & Friedman, 2018; Lee & Atance, 2016; Kushnir et al, 2015). Relatedly, taking a third-personal view on actions can facilitate higher-cognitions required for immediate (White & Carlson, 2015) and future-oriented (Atance, Louw & Clayton, 2015) self-regulation. In Study 2 we explore whether fluency has more predictive power in explaining individual differences in children's beliefs about their ability to act against and inhibit desires which they have expressed for themselves, rather than those given for another person.

Study 2

Method

Participants A total of 28 4-6-year-olds (M_{age} = 4.93, SD_{age} = 0.77, N_{female} = 19) were recruited at a science museum in a small city in the northeastern United States. Data collection is still ongoing and preliminary results are reported.

Procedure Children were interviewed individually in a quiet room at a museum. The procedure consisted of the first-person free will questions followed by the idea generation task battery in a counterbalanced order.

The free will task was similar to Study 1, but the questions first asked children to think of their own desires (e.g. "think of a [food you really like]/[something you really like to do]" and then referenced the child's response rather than the desires of someone else (e.g. "If *you* really wanted to [eat/do X], could *you* just choose to…"). Question order was counterbalanced. Order of the answer choices within the question ("choose to" vs "have to") was counterbalanced. The two Control questions had the same form as in Study 1. A majority of responses to these control questions (82%) were correct. A Cohen's k indicated agreement between the two coders for each question (κs > .82, *ps* < .0005). Disagreements were resolved through discussion.

After the free will questions, children completed the idea generation task battery, exactly as in Study 1. Coding followed the same procedure as in Study 1. Two coders were trained on the coding scheme. A Cohen's k indicated agreement between the two coders for each question (κs > .815, *ps* < .0005). Disagreements were resolved through discussion.

Results

Free Will A repeated measures ANCOVA with Question Type (Action vs Inhibition) as a within-subjects factor and age as a covariate found no effect of question type ($F(1)$ = .151, p = .701), no effect of age ($F(1)$ = .151, p = .701), and no interaction. In addition to not being different from each other or correlated with age, children's free will scores were significantly below chance for Action (M = 0.64, SD = 0.70; $t(25)$ = -2.132, p = .043) but not Inhibition (M = 0.96, SD = 0.89; $t(25)$ = -.225; p = .824). We further confirmed that rates of "choose to" responses were low by comparing to responses in Study 1. T-tests of overall "choose to" scores showed that they were significantly lower in Study 2 than in Study 1 (Study 1: M = 2.58, SD = 1.299, Study 2, M = 1.60, SD = 1.354; $t(63)$ = 2.897, p = .005).

Idea Generation Table 2 shows the descriptive statistics for each idea generation task as well as the correlations between them. As in Study 1, the results point to group-level and individual consistency across domains. On the group level, a repeated measures ANCOVA with Question Type (Physical vs Fantastical_Tails vs Fantastical_Walls vs Social vs Psychological vs Total) and age as a covariate found no effect of Question Type ($F(1)$ = .137, p = .715), no effect of age ($F(1)$ = 2.450, p = .132), and no interaction. Also, we found correlations between the tasks, not all approached significance. However, a 2 (Study: Study 1 vs Study 2, between subjects) x 6 (Question: Physical vs Fantastical_Tails vs Fantastical_Walls vs Social vs Psychological vs Total, within subjects) ANCOVA controlling for age revealed a marginal effect of Study ($F(1)$ = 3.157, p = .081), no main effect of Question ($F(1)$ = 1.778, p = .187), and no interaction.

Relationship between Free will beliefs and Fluency Children received a score of 0-4 for total free will judgments. Total free will score was significantly correlated with the number of ideas generated in the Fantastical Tails task (r = .493, p = .020). Total free will score was also significantly correlated with the number of ideas generated across all tasks (r = .428, p = .047). (see Table 2).

Discussion

In this study, we examined children's first-person free will judgments, children's fluency with generating alternative possibilities, and the relationship between the two. Children's first-person free will judgments were lower and more variable than children's third-person free will judgments, replicating past findings and supporting the idea that psychological distance facilitates children's ability to think about alternative possible actions.

As in Study 1, idea generation was consistent across Physical, Fantastical, and Social/Psychological domains at the group level and to a large extent at the individual level. Though our second sample was smaller, we again found reliable individual differences in counterfactual fluency using this measure.

Importantly, children's fluency predicted their judgments that they could possibly choose to act against their own most and least desired foods and activities. This relationship suggests that facility in generating multiple alternative possibilities might contribute to children's free will beliefs. Implications of this are discussed further below.

General Discussion

In this project, we examined whether individual differences in the ability to generate multiple alterative possibilities in an idea generation task relate to children's first-person or third-person free will judgments. We conducted two studies that measured a child's third-person or first-person free will judgments and their ability to fluently generate alternative possibilities. Overall, we found consistency in children's fluency across domains and we found that children's first-person free will judgments relate to overall fluency across domains and fluency within one of the fantasy domains.

Children's first-person free will judgments were also related specifically to one of our fantasy idea-generation tasks – imagining a world where everyone has tails. It is worth noting that we did not find comparable correlations between free will beliefs and social-psychological idea generation (e.g. ideas for making a friend happy when she is sad). This suggestive result (based as it is on a small sample) requires further study, but parallels links found in recent work by White et al. (2017) showing that pretending to be a superhero or other fantasy character has advantages for self-regulation. Together with this work, our results raise interesting questions about whether fantasy or pretense, rather than general theory-of mind abilities, might present unique advantages to children's developing ability to overcome struggles of will power and self-control.

Evidence of our hypothesized relationship between children's first-person free will judgments and their overall fluency is both correlational and preliminary, thus examining causal links is question for future research. Establishing causal links will have implications for understanding the mechanisms by which children's imaginations help them overcome psychological barriers in their self-beliefs.

One potential mechanism is a direct pathway from idea generation to judgments of choice and possibility. To explore this further would require experimentally limiting or enhancing idea generation in children and then exploring downstream effects on free will judgments. Other potential causal mechanisms are more indirect, via a third factor (or set of factors) that is responsible for both imaginative idea generation and judgments of free will. Language development is one candidate causal influence on both;

Table 2: Relationships between idea generation questions across domains in Study 2. Relationship between idea generation and first-person free will judgments included in final row.

	Physical	Fantastical Tails	Fantastical Walls	Soc/Psych third-person	Soc/psych first-person	Total number of ideas generated
Physical (M = 3.42, SD = 2.75)	-	.207	.684***	.448*	.354^{+}	
Fantastical Tails (M = 3.04, SD = 2.16)	-	-	.389^{+}	.185	.049	
Fantastical Walls (M = 3.38, SD = 3.35)	-	-	-	.544**	.285	
Social/psychological third-person (M = 2.77, SD = 1.90)	-	-	-	-	.266	
Social/psychological first-person (M = 1.73, SD = 1.69)	-	-	-	-	-	
First-person Free Will	.211	.493*	.356	-.108	.259	.428*

$^{+}$p<.1, *p<.05; **p<.01, ***p<.001

ideational fluency is known to be correlated with verbal ability, a fact which is supported in our study by intercorrelations between ability to generate ideas across physical, fantastical, and social/psychological domains. Social-cognitive skills are also correlated with language development (e.g. Astington & Jenkins, 1999; Carlson & Moses, 20001). Additional work is needed to investigate what influence, if any, developing verbal abilities have on the link between the two.

Perhaps a more interesting possibility is that one specific aspect of language development, semantic fluency, plays an important causal role. Indeed, semantic fluency tasks which require a child to list as many examples from a category in a specified amount of time (Kave, Kigel, & Kocvha, 2008) bear a resemblance to idea generation tasks such as the UUT: both require the child to have enough knowledge to explore a space of possibilities within a specified category. But idea generation tasks also go beyond semantic fluency because they require extending and conceptually re-combining familiar ideas and concepts in novel ways (e.g. other uses for a tissue). Conceptual re-combination additionally require other cognitive facilities like cognitive flexibility and, in first-personal cases, knowledge that is episodic or autobiographical (Schacter & Addis, 2007). We therefore don't believe it is likely that semantic fluency alone explains the link between idea generation and free will judgments, but this is a question that is beyond the scope of our data to address.

Despite recent agreement that the ability to imagine alternative possibilities is an important cognitive skill, few studies have examined how individual differences in modal cognition play a role in the ordinary judgments that rely on it. Perhaps capturing these differences can help explain variability and developmental changes in judgments of possibility (Shtulman & Carey, 2007; Lane et al, 2016), episodic future thinking (Atance & Melzoff, 2005, Atance et al, 2015) and counterfactual/hypothetical reasoning (Beck, Robinson, Carroll & Apperly, 2006) and causal inference (Walker & Gopnik, 2014). The approach outlined here could also be used to explore whether cultivating an ease with imagining new ideas could help children master basic (but difficult) social, cognitive and self-regulatory tasks.

Acknowledgments

We would like to thank Aliza Adhami, Regina Longley, and Nicole Calautti for their help with data collection, coding, and analyses as research assistants. Also, we thank the families from the Ithaca community for participating in this project, and we thank the Ithaca Sciencenter for partnering with us for this study.

References

Alquist, J. L., Ainsworth, S. E., Baumeister, R. F., Daly, M., & Stillman, T. F. (2015). The Making of Might-Have-Beens: Effects of Free Will Belief on Counterfactual Thinking. *Personality and Social Psychology Bulletin, 41*(2), 268–283.

Astington, J. W., & Jenkins, J. M. (1999). A longitudinal study of the relation between language and theory-of-mind development. *Developmental Psychology, 35*(5), 1311-1320.

Atance, C.M., Louw, A., Clayton, N. S. (2015). Thinking ahead about where something is needed: new insights about episodic foresight in preschoolers. *Journal of Experimental Child Psychology, 129*, 98-109.

Atance, C. M., & Meltzoff, A. N. (2005). My future self: Young children's ability to anticipate and explain future states. *Cognitive Development, 20*(3), 341–361.

Baumeister, R. F., Vohs, K. D., & Oettingen, G. (2016). Pragmatic prospection: How and why people think about the future. *Review of General Psychology, 20*(1), 3-16.

Beck, S. R., Robinson, E. J., Carroll, D. J., & Apperly, I. A. (2006). Children's Thinking About Counterfactuals and Future Hypotheticals as Possibilities. *Child Development, 77*(2), 413–426.

Behne, T., Carpenter, M., Call, J., & Tomasello, M. (2005). Unwilling versus unable: Infants' understanding of intentional action. *Developmental Psychology, 41*, 328-337.

Balke, A., & Pearl, J. (1994). Counterfactual probabilities: Computational methods, bounds and applications. *In Proceedings of the Tenth Conference Annual Conference on Uncertainty in Artificial Intelligence* (UAI-94) (pp. 46–54). San Francisco, CA: Morgan Kaufmann.

Bowman-Smith, C. K., Shtulman, A., & Friedman, O. (2018). Distant Lands Make for Distant Possibilities: Children View Improbable Events as More Possible in Far-Away Locations. *Developmental Psychology*.

Carlson, S. M., & Moses, L. J. (2001). Individual differences in inhibitory control and children's theory of mind. *Child Development, 72*(4), 1032-1053.

Chernyak, N., Kushnir, T., & Wellman, H. M. (2010). Developing notions of free will: Preschoolers' understanding of how intangible constraints bind their freedom of choice. *Proceedings of the Thirty-Second Annual Meeting of the Cognitive Science Society*, 2602-2606.

Chernyak, N., Kushnir, T., Sullivan, K. M., & Wang, Q. (2013). A comparison of American and Nepalese children's concepts of freedom of choice and social constraint. *Cognitive Science, 37*(7), 1343–1355.

De Brigard, F., Addis, D. R., Ford, J. H., Schacter, D. L., & Giovanello, K. S. (2013). Remembering what could have happened: Neural correlates of episodic counterfactual thinking. *Neuropsychologia, 51*(12), 2401–2414.

Engle, J. & Walker, C. M. (2018). Considering alternatives facilitates anomaly detection in preschoolers. *Proceedings of the Fortieth Annual Meeting of the of Cognitive Science Society,* 348-353.

Feldman, G., Baumeister, R. F., & Wong, K. F. E. (2014). Free will is about choosing: The link between choice and the belief in free will. *Journal of Experimental Social Psychology, 55*, 239–245.

Guilford, J.P. (1967). *The nature of human intelligence*. New York, NY: McGraw-Hill.

Kave, G., Kigel, S., & Kochva, R. (2008). Switching and clustering in verbal fluency tasks throughout childhood. *Journal of Clinical and Experimental Neruopsychology, 30*(3), 349-359.

Kushnir, T. (2018). The developmental and cultural psychology of free will. *Philosophy Compass, 13*(11), 1-17.

Kushnir, T., Gopnik, A., Chernyak, N., Sullivan, K. M., & Wang, Q. (2015). Developing intuitions about free will between ages four and six. *Cognition, 138*, 79-101.

Lane, J. D., Ronfard, S., Francioli, S. P., & Harris, P. L. (2016). Children's imagination and belief: Prone to flights of fancy or grounded in reality? *Cognition, 152*, 127–140.

Lee, W. S. C., & Atance, C. M. (2016). The effect of psychological distance on children's reasoning about future preferences. *PLoS One, 11*(10), e0164382.

Lewis, D. (1973). Causation. *Journal of Philosophy, 70*(17), 556-567.

Milgram, R. M., & Milgram, N. A. (1976). Creative thinking and creative performance in Israeli students. *Journal of Educational Psychology, 68*, 255-259.

Nichols, S. (2004). The folk psychology of free will: Fits and starts. *Mind & Language, 18*, 473-502.

Nichols, S. (2011). Experimental philosophy and the problem of free will. *Science, 331*, 1401-1403.

Nusbaum, E. C., & Silvia, P. J. (2011). Are intelligence and creativity really so different?: Fluid intelligence, executive processes, and strategy use in divergent thinking. *Intelligence, 39*(1), 36-45.

Phillips, J., & Cushman, F. (2017). Morality constrains the default representation of what is possible. *Proceedings of the National Academy of Sciences of the United States of America, 114*(18), 4649-4654.

Phillips, J., Luguri, J. B., & Knobe, J. (2015). Unifying morality's influence on non-moral judgments: The relevance of alternative possibilities. *Cognition, 145*, 30-42.

Schacter, D. L., & Addis, D. R. (2007). The cognitive neuroscience of constructive memory: remembering the past and imagining the future. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 362*(1481), 773–786.

Schacter, D. L., Benoit, R. G., De Brigard, F., & Szpunar, K. K. (2015). Episodic future thinking and episodic counterfactual thinking: Intersections between memory and decisions. *Neurobiology of Learning and Memory, 117*, 14-21.

Schult, C. A., & Wellman, H. M. (1997). Explaining human movements and actions: Children's understanding of the limits of psychological explanation. *Cognition, 62*, 291-324.

Shtulman, A., & Carey, S. (2007). Improbable or impossible? How children reason about the possibility of extraordinary events. *Child Development, 78*(3), 1015–1032.

Shtulman, A., & Phillips, J. (2008) Differentiating "could" from "should": Developmental changes in modal cognition. *Journal of Experimental Psychology, 165*, 161-182.

Taylor, M., Mottweiler, C. M., Aguiar, N. R., Naylor, E. R., & Levernier, J. G. (2018). Paracosms: The imaginary worlds of middle childhood. *Child Development*, 1-15.

Walker, C. M., & Gopnik, A. (2013). Causality and imagination. In M. Taylor (Ed.), *Oxford library of psychology. The Oxford handbook of the development of imagination* (pp. 342-358). New York, NY, US: Oxford University Press.

Walker, C. M., Gopnik, A., & Ganea, P. A. (2014). Learning to learn from stories: Children's developing sensitivity to the causal structure of fictional worlds. *Child Development, 86*(1), 310-318.

Wallach, M. A., & Kogan, N. (1965). Modes of thinking in young children: A study of the creativity-intelligence distinction. Oxford, England: Holt, Rinehart & Winston.

Weisberg, D. S., & Sobel, D. M. (2012). Young children discriminate improbable from impossible events in fiction. *Cognitive Development, 27*, 90-98.

Wellman, H. M., & Liu, D. (2004). Scaling of Theory-of-Mind Tasks. *Child Development,* 75(2), 523–541.

White, R. E., & Carlson, S. M. (2015). What would Batman do? Self-distancing improves executive function in young children. *Developmental Science, 19*, 419-426.

White, R. E., Prager, E. O., Schaefer, C., Kross, E., Duckworth, A. L., & Carlson, S. M. (2017). The "Batman Effect": Improving perseverance in young children. *Child Development, 88*(5), 1563-1571.

Woodward, J., & Hitchcock, C. (2003). Explanatory Generalizations, Part I: A Counterfactual Account. *Nous, 37*(1), 1–24.

Children master the cardinal significance of counting after they learn to count

Madison Flowers[1] (madison.flowers@yale.edu), Lindsay Stoner[2] (stonerl@sas.upenn.edu), & Julian Jara-Ettinger[1] (julian.jara-ettinger@yale.edu)

[1] Department of Psychology, Yale University. New Haven, CT 06520 USA.
[2] Department of Psychology, University of Pennsylvania. Philadelphia, PA 19014.

Abstract

Children learn the meaning of number words by going through a systematic set of stages of knowledge that culminates in their mastery of counting. Theoretical work has long suggested that children's acquisition of counting is not procedural, but semantic: all counters understand that counting computes cardinality. Yet, recent research has cast doubt on whether early counters truly understand the meaning of these words. Here we show that early counters also have an immature understanding of how one-to-one correspondence between an ordered list and a set of objects can be used to compute exact cardinality. Nonetheless, this understanding is improved when cues to quantity, such as size, are highlighted. Our results add to a growing body of work suggesting that counting is not a final stage in children's path to number, but a powerful tool that they can use to build and strengthen their intuitions about cardinalities.

Keywords: Cognitive development; number cognition; one-to-one correspondence.

Introduction

Children go through a systematic set of stages of knowledge when they learn number words and counting (Carey, 2009; Wynn, 1990; Fuson, 1988). First, children memorize the count list without knowing what these words mean (akin to learning a song like "eeny, meeny, miny, moe, …"), usually around the age of two. Children then slowly, but steadily, uncover the meaning of the words "one," "two," "three," and sometimes even "four," taking approximately six months to learn the meaning of each word. Children at these stages are called one-, two-, three-, and four-knowers, respectively, or subset-knowers collectively. After learning the meaning of the first three or four words, something clicks in children's minds. Rather than continuing to learn the meaning of number words one at a time, children suddenly, in what seems like a stroke of insight, grasp the logic of counting. Children at this stage, called *full counters*[1], can determine the size of any set (as long as they have memorized the count list up to that number). This last transition is a major milestone: the mastery of counting (Carey, 2009; Wynn, 1990; 1992; Piantadosi, Tenenbaum, & Goodman, 2012; Sarnecka & Lee, 2009; Lee & Sarnecka, 2010).

Theoretical work suggests that, in order to count correctly, children must understand five principles (Gelman & Gallistel, 1987). First, children must understand that any collection of objects can be counted (abstraction principle). To do so, objects must be placed in one-to-one correspondence with number words (one-to-one correspondence principle). The order in which the objects are counted is irrelevant (order irrelevance principle) but the order in which the number words are recited is not (stable order principle). When these steps are executed correctly, the word associated with the last object refers to the total number of objects in the set (the cardinal principle).

Research has long focused on the acquisition of the cardinal principle, as it is thought to be the key principle that marks the difference between children who can count, and children who cannot (Carey, 2009; Piantadosi, Tenenbaum, & Goodman, 2012). Yet, recent research has cast doubt on whether early counters have indeed grasped the conceptual logic of the cardinal principle. In a now classical study, Davidson, Eng, & Barner (2012) showed that children who had recently learned to count failed seemingly simple questions like determining whether "five" is more than "four." This work suggests that children's mastery of counting is a procedural milestone—learning to perform a complex set of rules in a systematic way—rather than a sematic milestone—learning that all number words refer to exact quantities and that counting computes a set's cardinality.

Nonetheless, if early counters are only missing the cardinal principle, they should understand how the rest of the principles combined can be used to determine a set's cardinality. Consider, for instance, watching an agent count two sets of objects. If the agent counts up to "six" in one set and up to "seven" in the second set, we can recognize that the second set has more objects because the set of words "one, two, …, seven" is larger than the set of words "one, two, …, six." Conceptually, this kind of inference only requires understanding that the objects were placed in one-to-one correspondence with the ordered list of number words. In practice, however, this type of inference is unavailable because it requires representing the list of words as a set of objects. However, if the objects were placed in one-to-one

[1] Full counters are classically called Cardinal Principle knowers (or CP-knowers for short; Carey, 2009; Lee & Sarnecka, 2010; Sarnecka & Lee, 2009; Piantadosi, Jara-Ettinger, & Gibson, 2014). Here we use a more neutral term that describes procedural competence without commitment to conceptual change because recent work suggests full counters may not know the cardinal principle yet (Davidson, Eng, & Barner, 2012; Jara-Ettinger, Piantadosi, Spelke, Levy, & Gibson, 2017).

correspondence with a visible set of objects, young counters may be able to perform these inferences.

Research into children's understanding of number principles suggests this may be the case. Three-year-olds understand that two small sets placed in one-to-one correspondence must be of equal size (Sophian, 1988; Gelman, 1982), and, at an earlier age, 18-month-olds preferentially look at counting events that follow one-to-one correspondence over events that do not (and this preference disappears when the agent uses novel words or beeps; Slaughter, Itakura, Kutsuki, & Siegal, 2011). At the same time, classical studies were performed with small sets that even infants can track, independent of their knowledge of number (Feigenson & Carey, 2003; Feigenson, Carey, & Hauser, 2002), and children's performance in other numerical tasks suggests that young children do not grasp the full significance of how one-to-one correspondence relates to exact number (Shipley & Shepperson, 1990; Izard, Streri, & Spelke, 2014).

Here we test if young counters can determine a set's cardinality by watching an agent apply all the counting principles using a list where the words are not names for cardinalities. We introduced participants to an ordered list of animals that someone used to count two sets of objects. Children could not see the two sets of objects, but they could see the agent placing them in one-to-one correspondence with the animal list in a stable order. If children understand the logic of these principles, they should be able to determine which of the two sets has more objects (as this only requires seeing on which set the counter reached an animal further along in the list). If, however, children are unable to identify which set has more objects, this would suggest that a robust understanding of how these counting principles help reveal exact cardinality emerges after children learn to count.

Experiment 1

In Experiment 1 participants watched an agent count two sets of hidden objects by placing them in one-to-one correspondence with an ordered list of animals (Figure 1a). Participants were then asked to determine which of the two boxes had more objects. Participants completed three trials. Two of these trials were controls to ensure that children understood that the agent was placing the animals in one-to-one correspondence with the objects. The first control trial contrasted two with three objects (such that, if children understand that the agent was placing the unobservable objects in one-to-one correspondence with the animals, they should identify the box with three objects by simply tracking the small quantities; Feigenson & Carey, 2003; Feigenson, Carey, & Hauser, 2002). The second control trial contrasted three with six objects (such that if children understand the one-to-one correspondence between objects and animals, they should identify the box with six objects by relying on their approximate number system; Xu & Spelke, 2000; Xu, 2003; Lipton & Spelke, 2003; Wood & Spelke, 2005). Finally, the critical trial contrasted six versus seven objects, which can only be solved if children understand how a proper application of counting principles reveals exact cardinality. Hypotheses, procedure, exclusion criteria, and analyses were pre-registered.

Methods

Participants. 60 full counters, as determined by the Give-N Task (Wynn, 1992; Carey, 2009, Sarnecka & Lee, 2009; Lee & Sarnecka, 2010) were recruited for this study (mean age: 4.88 years; range = 3.35-5.98). Twenty-nine additional children were recruited for the study, but not included because the experimenter determined they did not know how to count based on pre-registered criteria (n=16; see Procedure); because a coder blind to hypothesis determined that the participant did not know how to count (n=10; see Results) or because they declined to complete the study (n=3 participants; see Results).

Stimuli. The stimuli consisted of two bowls and ten bouncy balls for the Give-N task. For the animal task, the stimuli consisted of an ordered animal list, composed of eight animals ordered by size (Figure 1a), eight erasers, and three videos, each showing an agent counting objects in two opaque boxes using the ordered animal list.

Procedure.

Give-N Task. Children were presented with one bowl with ten bouncy ball and one empty bowl. The task always began with a request to move four bouncy balls from one bowl to the other. After each query, all bouncy balls were returned to the first bowl. If the child succeeded in this first trial, the next request was to move five bouncy balls. If the child failed, the next request was to move one bouncy ball. The task then followed a stair-cased procedure: children were asked to move *N+1* bouncy balls if they moved *N* bouncy balls correctly, and were asked to move *N-1* bouncy balls otherwise, with two exceptions: the same request was repeated when children failed at *N=1* and when they succeed at *N=8* (ensuring that moving all bouncy balls was never the correct answer).

Whenever children's error was off by (at most) two bouncy balls, the experimenter asked "Is that *N* bouncy balls? Can you count them for me please?" If the child recognized an error, the experimenter asked "Can you fix it so there are *N* bouncy balls in the bowl?" The experimenter recorded the original and the revised answers, and used the final answer to determine the next trial. Only participants who correctly moved four bouncy balls at least once proceeded to the one-to-one correspondence task (as determined in the pre-registration; although note that all participants who participated in the one-to-one correspondence were coded afterwards to test if they knew how to count; see Results).

Animal task. Participants were introduced to a non-numeric ordered list that consisted of eight animals ordered from left to right based on size (from smallest to largest; Figure 1a): ant, mouse, cat, pig, cow, bear, elephant, giraffe. Children were given a printed version of this list that they could consult at any time. To show how an agent would count using this list, the experimenter counted a line of four identical objects visible to the child using the list ("ant,

mouse, cat, pig"), and then counted a line of eight visible objects using the list ("ant, mouse, …, elephant, giraffe"). The experimenter counted out loud while using their finger to touch each item as they pronounced each animal name in the non-numeric ordered list and emphasized the final word. The experimenter then restated the final word of the count list (e.g., "there are giraffe objects"). After the warm-up, children completed three test trials (order counterbalanced across participants). In each trial, participants watched a video of an agent counting the objects in two boxes. The boxes were visible, but their contents were not. Immediately after counting the items in the box, the agent placed a picture of the corresponding animal on each box and then stated how many items were in the box using the non-numeric animal list. The animals were scaled by size on the printed list that children received, and on the pictures attached to the boxes (see below).

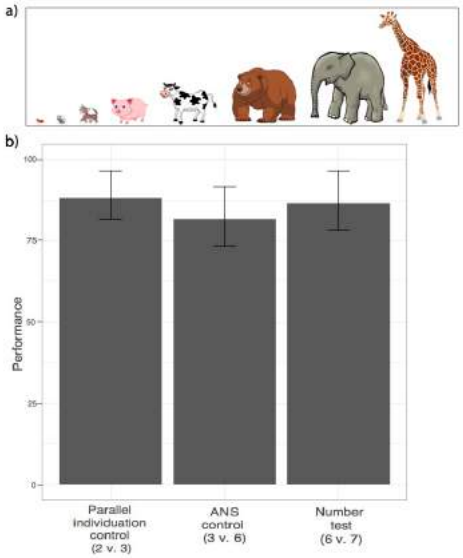


Figure 1. a) Animal list used in Experiment 1. b) Results from Experiment 1. The x axis shows the trial and the y axis shows the percentage of participants who correctly identified the box with more objects. Vertical lines show 95% bootstrapped confidence intervals. Overall, participants were able to identify which box had more objects in all three trials.

The three counterbalanced trials consisted of a 2 v. 3 trial, a 3 v. 6 trial, and a 6 v. 7 trial. In the 2 v. 3 trial the agent counted two objects in one box and then three objects in another box using the animal list (order in which boxes were counted counterbalanced). In the 3 v. 6 trial the agent counted three objects in one box and then the agent counted six objects in another box using the non-numeric ordered list (order counterbalanced). Finally, in the 6 v. 7 trial the agent counted six objects in one box and then seven objects in a second box using the non-numeric ordered list (order counterbalanced). The first two trials were control trials, as they could be solved by tracking number of words uttered via the parallel individuation system (2 v. 3 trial; Feigenson & Carey, 2003; Feigenson, Carey, & Hauser, 2002), or they could be distinguished through the approximate number system (3 v. 6 trial; Xu & Spelke, 2000; Xu, 2003; Lipton & Spelke, 2003; Wood & Spelke, 2005). The last trial (6 v. 7) was the critical one, as it can only be solved by understanding how the assignment of objects to animals reveals exact cardinality.

Trial order was counterbalanced across participants. In all videos, the agent counted out loud while using their finger to touch the inside of the box as they pronounced each animal name in the non-numeric ordered list. After each video, children were shown a picture of the two boxes, each labeled with the animal corresponding to the number of objects in the box, and they were asked which box has more blocks in it.

Results and Discussion

A coder blind to the experiment hypothesis coded whether children who participated in the one-to-one correspondence study knew how to count, based on their Give-N responses.

Participants who were not determined to be full counters by decision of a coder blind to hypothesis were excluded from the study and replaced (n = 10). An additional 3 participants were excluded and replaced because they did not want to complete the study.

Figure 1b shows the results from the experiment. Participants overwhelmingly succeeded in the 2 v. 3 and in the 3 v. 6 trials, showing that they understood the task. Of the 60 full counters included in the study, 88.3% (95% CI: 78.33-95.00; N=53 participants) correctly identified the box with more objects in the 2 versus 3 trial, and 81.7% of participants (N=49; 95% CI: 70.00-90.00) correctly identified the box with more objects in the 3 versus 6 trial.

Participants also succeeded in the critical 6 v. 7 trial. 86.6% of participants (N=52; 95% CI: 78.33-96.97) correctly identified the box with more objects in the 6 v. 7 trial. Together, these results suggest that children were able to understand that the number of recited animals revealed the quantity of objects in the set.

To test for any developmental change, we ran a mixed-effects logistic regression predicting children's response in the critical number trial (6 v. 7) as a function of age (as a continuous variable), with trial order as a random intercept. These results suggested that children's performance improved as a function of age ($\beta = 1.87$, $p < 0.01$; See Figure 2).

Children's ability to succeed in the 6 versus 7 trial of this experiment suggests that children understand how following the counting principles can reveal a set's cardinality. Critically, this understanding can happen without recognizing that the words themselves are names for different set sizes.

At the same time, it is possible that children's performance was facilitated by the use of an animal list ordered by size. Specifically, children may have simply followed a heuristic where they always pointed to the larger

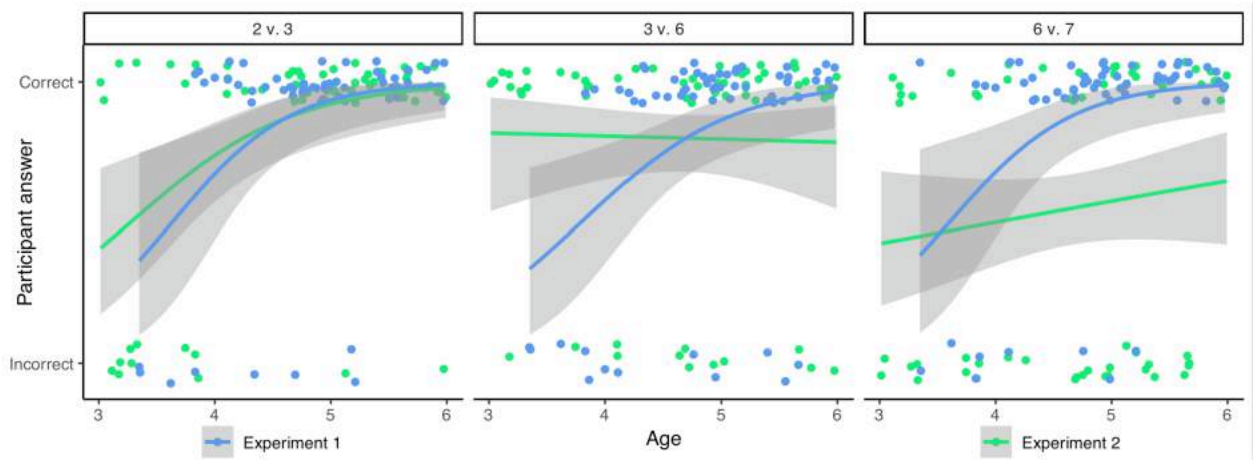


Figure 2. Participant responses in both experiments. Each dot represents a participant answer. The x-axis shows their age, and the y axis shows whether they identified the box with more objects. Data is minimally jittered on the y-axis for visibility purposes but was not jittered on the x-axis. Color indicates the experiment and the lines show logistic regressions.

animal without a deep understanding of how counting relates to cardinality. We test for this possibility in Experiment 2.

Experiment 2

Experiment 2 was conceptually identical to Experiment 1, with the difference that we used an animal list where the animals were no longer ordered on the basis of size, such that animals associated with larger quantities were not visually larger. Hypotheses, procedure, exclusion criteria, and analyses were pre-registered.

Methods

Participants. 60 full counters, as determined by the Give-N Task (Wynn, 1992; Carey, 2009, Sarnecka & Lee, 2009; Lee & Sarnecka, 2010) were recruited for this study (mean age: 4.67 years; range = 3.01-5.99). Nineteen additional children were recruited for the study, but not included because the experimenter determined they did not know how to count based on a pre-registered criterion (n=9; see Procedure); because a coder blind to hypothesis determined that the participant did not know how to count, as determined by a pre-registered coding procedure (n=8; see Results); or due to an error playing the experiment videos (n=2; see results).

Stimuli. Stimuli were identical to those in Experiment 1 with one exception. The counting list for this study consisted of eight different animals, ordered by color (Figure 3a).

Procedure. Methods for this study were identical to those from Experiment 1 with one exception. Instead of ordering the list by size, we now used a list of animals ordered by color (green, blue, purple, magenta, pink, red, orange, yellow): alligator, frog, octopus, butterfly, flamingo, lobster, fox, duck (Figure 3a). The size of the animals was matched on the printed list that children received and on the pictures attached to the boxes. Children completed the Give-N task, and the warm-up, as described in the previous experiment, using this new ordered list.

Children then completed the same three counterbalanced trials from Experiment 1: a 2 v. 3 trial, a 3 v. 6 trial, and a 6 v. 7 trial. After each video, children were shown a picture of the two boxes, each labeled with the animal corresponding to the number of objects in the box, and asked which box has more blocks in it.

Results and Discussion

Results were coded in the same way as Experiment 1. Eight participants were excluded from the study because they had not yet learned how to count. Two additional children were excluded because the experimental videos did not load properly.

Figure 3b shows the results from the experiment. Overall, participants succeeded in the two control trials, confirming that participants understood that the agent who counted was placing the objects in one-to-one correspondence with the animal list, and that the uttered animals revealed the number of objects in the set. Of the 60 full counters included in the study, 81.7% of participants (N=49; 95% CI: 71.67-91.67) correctly identified the box with more objects in the 2 v. 3 trial, and 80.0% of participants (N=48; 95% CI: 70.00-90.00) correctly identified the box with more objects in the 3 versus 6 trial. By contrast, only half of participants were now able to solve the critical 6 v. 7 trial. In this critical trial, only 55.0% of participants (N=33; 95% CI: 41.67-68.33) identified the box with seven objects.

To test for any developmental change, we ran a mixed-effects logistic regression predicting children's response in the critical number trial (6 v. 7) as a function of age (as a continuous variable), with trial order as a random intercept. These results suggested that children's performance did not improve as a function of age ($\beta = 0.3$; $p = 0.29$; See Figure 3).

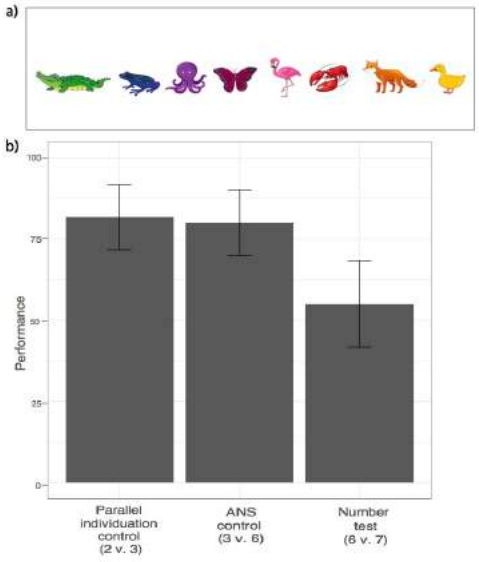


Figure 3. a) Animal list used in Experiment 2. b) Results from the experiment. The x axis shows the trial and the y axis shows the percentage of participants who correctly identified the box with more objects. Vertical lines show 95% bootstrapped confidence intervals. Overall, participants were able to identify box had more in the the 2 vs 3 and the 3 vs 6 trials. By contrast, only half of the participants succeeded in the critical 6 vs 7 trial.

These results conflict with those from Experiment 1, and they suggest that children did not recognize that the animal list could be used to determine exact cardinality. If they did, they could have solved the 6 v. 7 trial simply by consulting the list of animals and checking whether the set crocodile-lobster (six animals) was smaller or larger than the set alligator-fox (seven animals) were farther along the list. Note also that children always had a printed version of the list in front of them, and that pictures of the corresponding animals were placed in front of each box, minimizing concerns explainable by memory constraints.

These results suggest that children's success in Experiment 1 was supported by the use of animals ordered by size. Similarly, their overall failure in this experiment suggests that children understood that animals were being placed in one-to-one correspondence with the animal list, as they were able to solve the two control trials, but that they did not recognize that, through this process, children could determine the exact number of objects in the set.

General Discussion

Here we tested whether children who can count understand how the counting principles can be used to determine a set's exact cardinality, even without knowing the cardinal-principle—the understanding that the last word during counting refers to the size of the entire set. In Experiment 1 children watched an agent count the number of objects in two opaque boxes via one-to-one correspondence with a non-numerical animal list ordered by size (Figure 1a). Children were able to identify which box had more objects when the agent counted two objects in one box and three objects in the other, when the agent counted three objects in one box and six objects in the other, and when the agent counted six objects in one box and seven objects in the other (Figure 1b). Experiment 2 replicated this study using an animal list where the size of the animals was kept constant (Figure 3a). While children continued to successfully identify the larger set in the two control trials, their performance was drastically lower in the critical trial (Figure 3b).

Children's success in the two versus three trial, and in the three versus six trial in both experiments shows they understood that the number of words the agent uttered revealed the quantity of objects (Note also that children completed two warm-up trials where they saw the agent place two visible sets of objects in one-to-one correspondence with the animal list). However, success in these trials does not imply a mature understanding of how the counting procedure reveals exact cardinality. Past research has shown that children can distinguish between two and three sounds via the parallel individuation system (Feigenson & Carey, 2003; Feigenson, Carey, & Hauser, 2002) and that they can distinguish between three and six sounds via the approximate number system (Xu & Spelke, 2000; Xu, 2003; Lipton & Spelke, 2003; Wood & Spelke, 2005). By contrast, because children cannot perceptually distinguish between six and seven sounds, they could only solve this by understanding how the counting procedure reveals exact cardinality.

Critically, in our study, children did not need to understand the cardinal principle to succeed. If children recognized that the agent was placing the hidden objects in one-to-one correspondence with the animal list, they could have solved the task through at least two strategies. A first strategy is through awareness that, when counting principles are applied, later items reveal greater quantities. If children understood this, they would need to only find which animal comes later in the list to perform at ceiling. However, even if children did not recognize that later symbols in a count list reveal greater quantities, they could have solved the task through a second strategy: When the agent counted up to a certain animal, children could consult their list and see a set of animals that is numerically identical to the set of hidden objects (e.g., when the agent counted to butterfly in Experiment 1, children could see their list and recognize that the set of animals starting in crocodile and ending in butterfly is a set of the same size than the set of hidden objects that was counted). Through this strategy, children could recognize that one of the sets of animals is a subset of the other, making it trivial to identify which bowl had more objects.

The results from Experiment 1 are consistent with two possibilities. A first possibility is that children's ability to determine which of two sets had more objects improves when we the list includes a cue to number (by ordering the animal's based on size; Figure 1a). However, it is also possible that varying the size of the animals did not help children link the animal list to cardinalities. Instead, children may have simply selected the larger animal without conceptually understanding why this would be the correct answer. Note,

however, that children's performance in the 2 v. 3 and the 3 v. 6 trials was near-identical in Experiment 1 and Experiment 2. If children were simply pointing to the larger animal in Experiment 1, one might expect better performance relative to Experiment 2. In addition, older children were more likely to succeed in the 6 v.7 trial in Experiment 1. Intuitively, if children were relying on a size heuristic, younger children should have succeeded as well. Future work will test if this alternative can explain children's improved performance in Experiment 1.

In this study we recruited three-, four-, and five-year-olds and only tested children who were able to count. Because, in the US, children usually learn to count at around age four (Piantadosi, Jara-Ettinger, & Gibson, 2014; Wynn 1990, 1992), it is likely that most of our participants had just learned how to count. However, older participants are more likely to have known how to count for a longer time such that experience with counting and age were likely correlated in our sample. Thus, our finding that children improved in the 6 v. 7 trial in Experiment 1 does not reveal whether this improvement was due to age, or due to experience with counting. Future work will disambiguate between these possibilities.

Altogether, our results suggest that children who know how to count have yet to reach a mature understanding of how the counting principles reveal exact cardinalities. Our results add to a body of work that suggests that children's mastery of the counting procedure is not a final milestone in children's mastery of number words. Related work has also shown that young counters may also lack the cardinal principle (Davidson, Eng, & Barner, 2012; see introduction for review). Combined, this work suggests that when children learn to count, they master a set of procedural rules with only a partial understanding of how these rules relate to cardinality. Under this view, children's ability to count may be a building block towards their understanding of number words and cardinality rather than an endpoint. By learning to count, children may begin to notice a relationship between the set size and the final number word when counting, helping them realize that counting computes cardinality. Future work will test this hypothesis. What our findings do show, is that children's mastery of counting is an intermediate step in children's path to knowledge, and we add to a growing body of work suggesting that children's acquisition of procedural number knowledge may precede a mature understanding of the meaning of number words and counting (Jara-Ettinger, Piantadosi, Spelke, Levy, & Gibson, 2017; Davidson, Eng, & Barner, 2012; Cheung & LeCorre, 2015).

Acknowledgments

We thank the families who participated in this research. We thank Sarah Wong and Ivana Bozic for help with data collection, and Katherine Hermann and Maeve Bustell for help with coding. This work was supported by a Google Research Award. This material is based upon work supported by the Center for Brains, Minds, and Machines (CBMM), funded by NSF-STC award CCF-1231216.

References

Carey, S. (2009). The origin of concepts. Oxford: Oxford University Press.

Cheung, P., & Le Corre, M. (2015) Algebraic reasoning in 3- to 5-year-olds. Proceedings of the 37th Annual Conference of the Cognitive Science Society.

Davidson, K., Eng, K., & Barner, D. (2012). Does learning to count involve a semantic induction? *Cognition*, *123*(1), 162– 173.

Gelman, R. (1982). Accessing one-to-one correspondence: Still another paper about conservation. *British Journal of Psychology,73*(2), 209-220.

Gelman, R., & Gallistel, C. (1978). The child's understanding of number. Cambridge, MA: Harvard University Press.

Feigenson, L., & Carey, S. (2003). Tracking individuals via object-files: Evidence from infants manual search. *Developmental Science, 6*(5), 568-584.

Feigenson, L., Carey, S., & Hauser, M. (2002). The Representations Underlying Infants Choice of More: Object Files Versus Analog Magnitudes. *Psychological Science, 13*(2), 150-156.

Fuson, K.C. (1988). Children's counting and concepts of number. New York: Springer-Verlag.

Izard, V., Streri, A., & Spelke, E.S. (2014). Toward exact number: young children use one-to-one correspondence to measure set identity but not numerical equality. *Cognitive Psychology*, 72, 27–53.

Jara-Ettinger, J., Piantadosi, S., Spelke, E. S., Levy, R., & Gibson, E. (2016). Mastery of the logic of natural numbers is not the result of mastery of counting: Evidence from late counters. *Developmental Science, 20*(6).

Lee, M.D., & Sarnecka, B.W. (2010). A model of knower-level behavior in number-concept development. *Cognitive Science*, 34, 51–67.

Lipton, J.S. & Spelke, E.S. (2003). Origins of number sense: Large-number discrimination in human infants. *Psychological Science, 14*(5).

Piantadosi, S.T., Jara-Ettinger, J., & Gibson, E. (2014). Children's learning of number words in an indigenous farming foraging group. *Developmental Science*, *17*(4), 553–563.

Piantadosi, S., Tenenbaum, J., & Goodman, N. (2012). Bootstrapping in a language of thought: a formal model of numerical concept learning. *Cognition*, 123, 199–217.

Sarnecka, B.W., & Lee, M.D. (2009). Levels of number knowledge in early childhood. *Journal of Experimental Child Psychology*, 103, 325–337.

Shipley, E. F., & Shepperson, B. (1990). Countable entities: Developmental changes. *Cognition,34*(2), 109-136. doi:10.1016/0010-0277(90)90041-h

Slaughter, V., Itakura, S., Kutsuki, A., & Siegal, M. (2011). Learning to count begins in infancy: Evidence from 18

month olds visual preferences. *Proceedings of the Royal Society B: Biological Sciences,278*(1720), 2979-2984. doi:10.1098/rspb.2010.2602
Sophian, C. (1988). Limitations on preschool children's knowledge about counting: Using counting to compare two sets. *Developmental Psychology, 24*(5), 634-640.
Wood, J.N. & Spelke, E.S. (2005). Infants' enumeration of actions: numerical discrimination and its signature limits. *Developmental Science, 8*(2).
Wynn, K. (1990). Children's understanding of counting. *Cognition*, *36*(2), 155–193.
Wynn, K. (1992). Children's acquisition of the number words and the counting system. *Cognitive Psychology*, *24*(2), 220– 251.
Xu, F. (2003). Numerosity discrimination in infants: Evidence for two systems of representations. *Cognition, 89*(1).
Xu, F. & Spelke, E.S. (2000). Large number discrimination in 6-month old infants. *Cognition, 74*(1).

Toddlers recognize multiple polysemous meanings and use them to infer additional meanings

Sammy Floyd
Princeton University, Princeton, New Jersey, United States

Adele Goldberg
Princeton University, Princeton, New Jersey, United States

Casey Lew-Williams
Princeton University, Princeton, New Jersey, United States

Abstract

Up to 80% of words have multiple, related meanings (polysemy), yet work on early word learning has almost uniformly assumed one-to-one mappings between form and meaning. Using a looking-while-listening procedure, we present the first evidence that toddlers (n=40) can recognize multiple meanings for common nouns, e.g., dog collar, shirt collar. In an English-meaning condition, toddlers were tested on their ability to recognize multiple English meanings for polysemous words such as cap (e.g., a baseball cap and a bottle cap). Another condition prompted toddlers with the same English words (e.g., cap), but target referents instead corresponded to the words polysemous extension in an unfamiliar language, (e.g., lid is a meaning for Spanishs cap, tapa). Toddlers looked to the correct targets above chance on both trial types, but with greater accuracy on English-meaning trials, demonstrating a recognition of familiar word-meaning pairs and an ability to infer potential new meanings.

Do Neural Language Representations Learn Physical Commonsense?

Maxwell Forbes†, Ari Holtzman†‡, and Yejin Choi†‡
{mbforbes, ahai, yejin}@cs.washington.edu
†Paul G. Allen School of Computer Science and Engineering, University of Washington
‡Allen Institute for Artificial Intelligence

Abstract

Humans understand language based on the rich background knowledge about how the physical world works, which in turn, allows us to reason about the physical world through language. In addition to the *properties* of objects (e.g., *boats require fuel*) and their *affordances*, i.e., the actions that are applicable to them (e.g., *boats can be driven*), we can also reason about *if–then* inferences between what properties of objects imply the kind of actions that are applicable to them (e.g., that *if we can drive something then it likely requires fuel*).

In this paper, we investigate the extent to which state-of-the-art neural language representations, trained on a vast amount of natural language text, demonstrate physical commonsense reasoning. While recent advancements of neural language models have demonstrated strong performance on various types of natural language inference tasks, our study based on a dataset of over 200k newly collected annotations suggests that neural language representations still only learn associations that are explicitly written down.[1]

Keywords: physical commonsense, natural language, neural networks, affordances

Introduction

Understanding everyday natural language communication requires a rich spectrum of physical commonsense knowledge. Consider the example dialog sketched in Figure 1. A simple observation that, *"The blender is broken again!"* triggers myriad pieces of implied understanding (e.g., that something which requires electricity will only work with a source of power). Such knowledge is rarely stated explicitly (Van Durme, 2010), and instead can be inferred on-the-fly as needed.

In this paper, we study physical commonsense knowledge underlying natural language understanding, organized as interactions among three distinct concepts: (i) objects, (ii) their attributes (properties), and (iii) the actions that can be applied to them (affordances) (Figure 1, bottom). The premise of our study is that language models trained on a sufficiently large amount of text can recover a great deal of physical commonsense knowledge about each of these concepts. However, aspects of this knowledge may only be implicit in natural language utterances. For example, answering a question from the Winograd Schema Challenge (Levesque, Davis, & Morgenstern, 2012)—"The trophy would not *fit* in the brown suitcase because it was too *big*. What was too big?"—implicitly requires the physical commonsense reasoning that *"in order to fit X in Y, X should be relatively smaller compared to Y"*, which essentially requires reasoning about the affordances of objects (fit X in Y) from their attributes (relative size of X and Y).

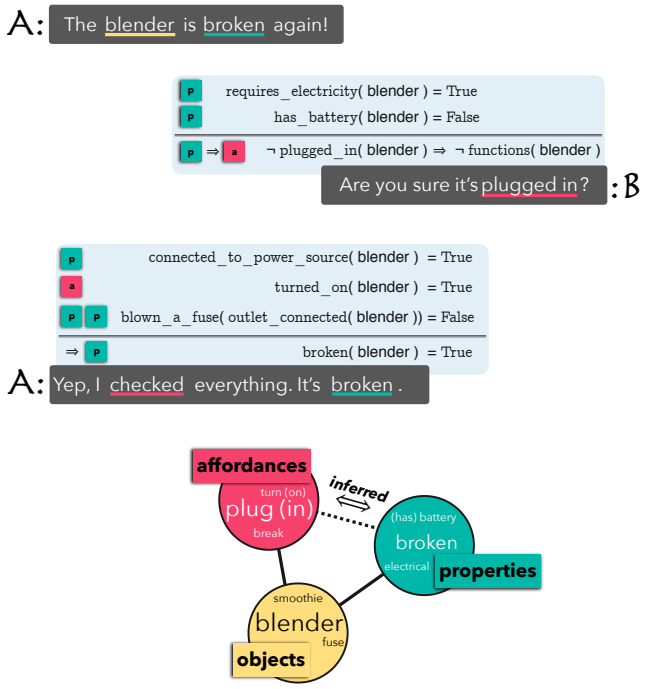


Figure 1: Natural language communication often requires reasoning about the affordances of objects (i.e., what actions are applicable to objects) from the properties of the objects (e.g., what are the size, weights, material of the objects) and vice versa. We study the extent to which neural networks trained on a large amount of text can recover various aspects of physical commonsense knowledge.

In this paper, we investigate the extent to which neural language models trained on a massive amount of text demonstrate various aspects of physical commonsense knowledge and reasoning. Our analysis includes word embeddings such as GloVe (Pennington, Socher, & Manning, 2014), as well as

[1] Visit `https://mbforbes.github.io/physical-commonsense` for our data, code, and more project information.

more recent contextualized representations like ELMo (Peters et al., 2018) and BERT (Devlin, Chang, Lee, & Toutanova, 2018). Such models are trained without supervision by exposing them to billions of words, and allowing them to extract patterns purely from token prediction tasks that can be derived directly from raw text. These language representation models have established unprecedented performance on a wide range of evaluations, including natural language inference and commonsense reasoning.

How much do these large, unsupervised models of language learn about physical commonsense knowledge? Some recent work has studied the capabilities of word embeddings to predict an object's properties (Rubinstein, Levi, Schwartz, & Rappoport, 2015; Lucy & Gauthier, 2017). Motivated by these efforts to understand language representations, we present several contributions. We contribute two datasets: the *abstract* dataset, a refreshed version of the McRate dataset (McRae, Cree, Seidenberg, & McNorgan, 2005), pruned and densely annotated to eliminate false negatives present in previous work; and the *situated dataset*, with annotations for objects' properties and affordances in real-world images sampled from the MS COCO dataset (Lin et al., 2014). As in previous work, we consider the prediction task of linking objects and their properties (O$\longleftrightarrow$P), but with our new situated dataset, we are also able to study the connection between objects and their affordances (O$\longleftrightarrow$A), as well as between affordances and properties (A$\longleftrightarrow$P). We also study the latest models from the natural language processing community (ELMo, BERT) using in-context word representations, and present results for all of our proposed datasets and tasks. Our analysis suggests that current neural language representations are proficient at guessing the affordances and properties of objects, but lack the ability to reason about the relationship between affordances and properties itself.

Characterizing Objects through Properties and Affordances

Properties

We use the term *properties* to refer to the static characteristics of objects. They encompass our commonsense understanding of what something is like. For example, we might say that an *apple* has the property of being *edible*, or that a *plant* is *stationary*.

As with McRae et al. (2005), properties capture the general perception of a thing. Exceptions naturally arise. For example, specific instances can violate the general properties of an object, such as the inediblilty of a rotten apple. Additionally, subtypes can diverge from the exemplar of a category, as with the Venus flytrap, a plant with the ability to move.

Affordances

We express an object's actions with verbs. One way to focus on understanding the actions of objects is to focus on their *affordances*. Coined by Gibson (1966), this term initially described animal-perceived uses for an object, but has since come to mean the perceived uses of an object in a given environment (Norman, 1988; Gaver, 1991).

Here, we take a simpler, human-centric definition. We consider an object's affordances to be, "what actions do humans take with an object?" For example, *boots* commonly afford *wear*, *kick off*, *lace up*, and *put on*.

Inference Between Affordances and Properties

Affordances and properties exhibit a surprising connection. As humans, we are able to infer many of an object's affordances based on its properties (A$\leftarrow$P). The same is also true in the reverse (A$\rightarrow$P).

Consider an exchange: *"You think you could fit that boulder in your truck?" "No way! That thing was so big you could go for a hike on it."* We might sketch out some of this information as:

$$\textit{fit } x \textit{ into } y \implies x <^{\text{size}} y$$
$$\text{hike(x)} \implies x \gg^{\text{size}} \textsc{Human}$$

While the above information only concerns a property's relative value (comparative size), all kinds of information traverse this edge implicitly:

*She **plugged in** her robot.*
$$\text{plug-in}(x) \implies \text{uses-electricity}(x)$$
*He **poured** coffee into the cup*
$$\text{pour-into}(x) \implies \text{holds-liquid}(x)$$
*It **shattered** on the floor.*
$$\text{shatter(x)} \implies \text{rigid}(x)$$

The implications ($\implies$) should be taken with a probabilistic grain of salt. However, they capture our intuitions about what we expect to be true. Wouldn't it be surprising to shatter something that isn't rigid, or plug-in something that doesn't take power?

Humans use the link between affordances and properties to recover information. Can machine learning models do the same? It is is difficult to model these implications based on text alone because there is no direct evidence for the implied information. Any implication that can be trivially understood by a person is precisely the kind of information left unsaid. Who would write, *"If I can walk inside my house, I know that my house is bigger than I am?"* Nevertheless, we naturally understand that: x *walk-inside* $y \implies x <^{\text{size}} y$.

Directly attacking the link between affordances and properties requires access to implications across the edges. Without such information, we can use objects as a proxy to understand how much modern neural networks know about this edge. For example, taking an object like *boots*, and using only its top affordances *wear*, *kick off*, and *lace up*, can we predict its properties?

Statistics

	Total	Statistics
Abstract		
Objects	514	411 train / 103 test
Properties	50	obj/prop: 60 median (3 min, 302 max) prop/obj: 8 median (1 min, 23 max)
Annotations	77,1000	3 anns/datum
Situated		
Objects	1,024	80 unique, split: 64 train / 16 test
Properties	50	
Affordances	3072	3 affordances / object (by design)
Annotations	156,672	3 anns/datum

Examples

Objects	Properties	Affordances
harmonica, van	*expensive, squishy*	*pick up, remove*
potato, shovel	*used as a tool for cooking*	*pet, talk to*
cat, bed	*decorative, fun*	*cook, throw out*

Table 1: Statistics and examples for the proposed abstract and situated datasets (based on (McRae et al., 2005) and (Lin et al., 2014)).

Experiments

Tasks

As shown at the bottom of Figure 1, our problem space naturally defines three edges in a graph. A property prediction task may attempt to produce the human-labeled set of properties given a new object (O→P) (Lucy & Gauthier, 2017). Predicting affordances can be done similarly: given a new object, can its top affordances be distinguished from others (O→A)? And finally, the troublesome but fertile edge between properties and affordances: can a model predict the set of properties compatible with an affordance (A→P)?

We frame each scenario as a series of joint reasoning tasks. Given two instances (e.g., an object and a property), a model must make a binary decision as to whether they are compatible. For example, predicting which properties out of a total of k are compatible with an object o will be set up as k compatibility tasks $(o, p_i) \rightarrow \{0, 1\}$. We denote the tasks as object-property (O⟷P), object-affordance (O⟷A), and affordance-property (A⟷P).

Data

To fuel experiments in these three tasks, we introduce two new datasets. The first we call the *abstract* dataset, which is a set of judgements elicited from only the name of the object (e.g., *wheelbarrow*) and property (e.g., *is an animal*). The second is the *situated* dataset, where properties and affordances are annotated on objects in the context of real-world pictures.[2]

Abstract Dataset Several lists of properties (McRae et al., 2005), categorization schemes (Devereux, Tyler, Geertzen, & Randall, 2014), and quantification layers (Herbelot & Vecchi, 2015) have been proposed. We take the set of objects and properties from McRae et al. and perform filtering and preprocessing similar to Lucy and Gauthier (2017). We also include the set of objects from the MS COCO dataset (Lin et al., 2014), collapse similar objects (e.g., many bird species) and add seven new properties (such as *man-made* and *squishy*). We end up with a set of 514 objects and 50 properties. We re-annotate all 25,700 object-property pairs to eliminate false negatives from the original McRae data collection process and provide labels for new entries. We annotate each pair three times for a total of 77,100 annotations, and keep only labels with $\geq 2/3$ agreement.

Situated Dataset We also annotate instances of objects situated in photographs. Images have the great advantage of resolving visual ambiguities of appearance, shape, and form. For example, a *bottle* has different properties if it is a glass beverage container or plastic shampoo tube. Only a few non-visual properties (e.g., *smelliness*) must then be inferred from the environment.

To build the an experimental situated testbed, we sample images from the MS COCO dataset (Lin et al., 2014). We constrain each image to have between three and seven objects to avoid scenes that are too sparse (often portraits) or dense (cluttered collections). We also ensure that we have at least five samples of each of the 80 unique object categories in the dataset. We end up with 1,024 objects across 220 images. We then annotate all 50 properties (introduced in the abstract dataset) for each object, annotating each three times for a total of 153,600 labels. We filter using the same scheme ($\geq 2/3$ agreement).

In addition to the properties, we also collect annotations of the affordances for all objects in the situated dataset. We allow annotators to choose from the 504 verbs from the imSitu dataset (Yatskar, Zettlemoyer, & Farhadi, 2016). We provide common variants of each verb that include particles, allowing annotations such as *pick up* and *throw out*. Annotators select the top three to five affordances that come to mind when they see the selected object in the context of its photograph. We again perform this annotation three times for each object, and aggregate the verbs chosen to pick the top three most common affordances for each object. We end up with a set of sparsely labeled affordances for each situated object. We perform balanced negative sampling by selecting $k = 3$ affordances for each datum and setting their labels to zero.

Detailed statistics and examples for both datasets are shown in Table 1.

[2]Annotations for both datasets are performed by workers on Amazon Mechanical Turk.

	Abstract				Situated											
	O ⟷ P				O ⟷ P				O ⟷ A				A ⟷ P			
	obj	*prop*	*μF1*	*sig*	*obj*	*prop*	*μF1*	*sig*	*obj*	*aff*	*μF1*	*sig*	*aff*	*prop*	*μF1*	*sig*
RANDOM	0.25	0.26	0.26	***	0.24	0.25	0.22	***	0.53	0.62	0.51	***	0.24	**0.26**	0.23	***
MAJORITY	0.34	0.11	0.31	***	0.16	0.05	0.17	***	0.82	0.68	0.82	***	0.18	0.05	0.17	***
GLOVE	0.63	0.47	0.63	*	0.55	0.39	0.57	**	**0.85**	**0.73**	**0.86**	←	0.27	0.13	0.29	
DEP-EMBS	0.62	0.42	0.60	**	0.54	0.36	0.54		0.84	0.67	0.84		0.26	0.12	0.28	
BERT	0.62	0.48	0.60	***	0.53	0.38	0.56		**0.85**	0.70	0.85		0.26	0.12	0.28	**
ELMO	**0.67**	**0.55**	**0.67**	←	**0.58**	**0.44**	**0.58**	←	0.84	0.71	0.85		**0.31**	0.17	**0.34**	←
HUMAN	0.78	0.80	0.67		0.70	0.69	0.61		0.83	0.93	0.80		0.65	0.67	0.40	

Table 2: Macro F1 scores per category (object, property, affordance) and micro F1 score (*μF1*) on both the abstract and situated test sets. Highest model values are bolded. Statistical significance (*sig*) is calculated with McNemar's test, comparing the best-scoring model (by *μF1*, denoted ←) with each other model. Stratified p-values are shown, with * for $p < 0.05$, ** for $p < 0.01$, and *** for $p < 0.001$. Human performance is estimated by 50 expert-annotated random samples from the test set (no McNemar's test).

Models

Word embeddings We consider four representations of the words involved in the tasks. Two of the representations are word embeddings. These map single words to vectors in $\mathbb{R}^d$. We use GloVe embeddings (Pennington et al., 2014) as they have proven effective at object-property tasks in the past (Lucy & Gauthier, 2017). We also use Dependency Based Word Embeddings (Levy & Goldberg, 2014), as they may more directly capture the relations between objects and their affordances. In both cases, $d = 300$, and we use the GloVe embedding variant with the largest amount of pretraining (840 billion words).

Contextualized representations The other two representations are ELMo (Peters et al., 2018) and BERT (Devlin et al., 2018), which are contextualized. These require full sentences (as opposed to single words) to compute a vector, but in turn produce results more specific to words' linguistic surroundings. For example, ELMo and BERT produce different representations for *book* in *"I read the book"* versus *"Please book the flight,"* while word embeddings have only a single representation.

To account for this, we generate sentences using the relevant objects, properties, and affordances for the task at hand. For example, to judge *accordion* and *squishy*, we would generate *"An accordion is squishy."*

For ELMo, we then take the final layer representations for the two compared words, each of which is a $d = 1024$ length vector. For BERT, we take the overall sentence representation and sum across the final four layers, which produces a single $d = 1024$ vector.

Finetuning Given the word representations above, we finetune each of the models by adding trainable multilayer perceptron (MLP) after the input representations. This allows models to learn interrelations between the two categories at hand, essentially calibrating the unsupervised representations into a compatibility function. We use a single hidden layer in the MLP, and train using mean squared error loss with L2 regularization.

To summarize, for two words (w_i, w_j) which can be written together in a sentence $s = w_1 ... w_n$, we have for a model m,

$$r(w_i, w_j) = \begin{cases} \langle m(w_i), m(w_j) \rangle & \text{if } m \in \{\text{GL., D.E.}\} \\ m^{-1}_{\{i,j\}}(s) & \text{if } m = \text{ELMO} \\ \sum_{\ell \in \{-4 ... -1\}} m^{\ell}(s) & \text{if } m = \text{BERT} \end{cases}$$

$$\hat{y}_{w_i,w_j} \propto \sigma(\mathbf{w_2^T} a(\mathbf{w_1^T} r(w_i, w_j) + \mathbf{b_1}) + \mathbf{b_2})$$

$$\mathcal{L}(w_i, w_j, y, \theta, \lambda) = (y - \hat{y}_{w_i,w_j})^2 + \lambda \|\theta\|_2^2$$

where $m(\cdot)_i^{\ell}$ is an embedding of the ith token in the layer ℓ, a is a nonlinear activation function, $y \in \{0, 1\}$ is the ground truth label, $\theta = \{\mathbf{w_1}, \mathbf{w_2}, \mathbf{b_1}, \mathbf{b_2}\}$ are trainable parameters, and λ is the regularization strength.

We optimize all models using gradient descent, and tune all hyperparameters using k-fold cross validation with $k = 5$.

Baselines We compare performance for these models against two simple approaches. The *random* baseline simply flips a coin for each compatibility decision. The *majority* baseline uses the per-class majority label for the training set, aggregating by property for the $O \longleftrightarrow P$ and $A \longleftrightarrow P$ tasks, and by affordance for the $O \longleftrightarrow A$ task.

Human performance Finally, we estimate human performance on this task. We sample 50 samples at random from the test set for each task, and have an expert annotate them. For fairness to the models, we do not show the expert the photographs or exact instance from which the situated examples are drawn.

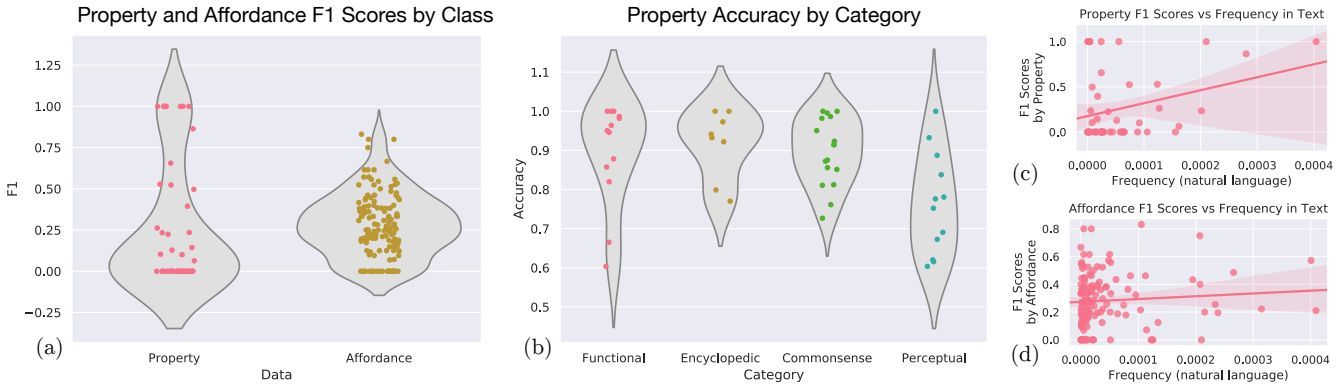


Figure 2: Detailed results of top performing model (ELMo) on the affordance-property compatibility task (A⟷P) in the situated dataset. (a) F1 scores are plotted per property (left) and affordance (right). (b) Properties are divided into four categories and plotted by accuracy. (c), (d) Both property and affordance F1 plotted against word frequency in natural language text.

Results

A summary of all model performances is shown in Table 2. Consistent with prior work that has studied object and property compatibility (Lucy & Gauthier, 2017), we find good but not perfect performance (close to 0.70 F1 scores) on the abstract dataset (task $O \longleftrightarrow P$). Models fare slightly worse on the situated $O \longleftrightarrow P$ task, with the best performance below 0.60 F1. This effect is consistent in the human expert scores as well. Though this dataset is larger, the introduction of context allows for greater variance in the properties of an object.

The object-affordance compatibility task ($O \longleftrightarrow A$) yields significantly higher numbers. Not only is this task statistically easier (as demonstrated by the strong majority baseline), but this edge is the only one directly observed in language. All models pretrained on text have been exposed to many instances of likely verbs for each object considered. In fact, all pretrained models perform in the same range as human ability, and there is no statistically significant difference between the models for this task.

However, all models struggle with the affordance-property task ($A \longleftrightarrow P$). The highest F1 scores are in the 0.30s, with the random baseline achieving the highest macro F1 score by property. While this task is also the most difficult for humans, their macro F1 scores for both affordances and properties are more than double that of the best performing models. We posit that the inference between affordances and properties requires multi-hop inference that is simply not present in the pretraining of large text-based models. We provide further analysis in the following section.

Analysis

Models achieve reasonable performance predicting the compatibility of both properties and affordances with objects. However, the task requiring inference between affordances and properties (A⟷P) confounds even the strongest models.

We explore this result through a detailed analysis of the top performing model. Figure 2 presents a breakdown of ELMo's results on the affordances-property compatibility task (A⟷P) on the situated dataset. From the leftmost graph (a), we observe that a per-property analysis shows a largely bimodal split between properties that are fully predicted (1.0 F1), and went completely unmodeled (0.0 F1). Affordances, on the other hand, lie more evenly across the F1 range. Because the task involved the compatibility between properties and affordances, mass for correct predictions must be shared between the two data groups. That so few properties achieved a high F1 score suggests that many affordances rely on only a few properties for accurate prediction.

We perform further analysis to investigate which kinds of properties yielded better affordance-property modeling. We categorize each property into four coarse classes: functional (e.g., *is used for cooking*), encyclopedic (e.g., *is an animal*), commonsense (e.g., *comes in pairs*), and perceptual (e.g., *is smooth*). Figure 2 (b) shows a breakdown of property performance grouped by these four categories. (Here, we plot accuracy instead of the sharper F1 metric to better illustrate the spread of performance.) Functional properties exhibit the highest performance. This makes intuitive since, because functional capabilities are directly tied to an object's affordances. In contrast, perceptual properties exhibit generally lower and inconsistent performance than other categories. We suspect that perceptual observations observed in text are not expressed with affordances, making this connection difficult for models. Largely perceptual features can be written about with simple verbs (*hear, see, feel*), giving them less implicit evidence than more nuanced properties. Finally, encyclopedic and commonsense properties fall somewhere in the middle. These properties, which involve an object's general characteristics (like *requires gasoline, lives in water,* or *has a peel*), correlate with a variety of verbs. But they may only be directly expressed at a distance from a verb, making the inference between them still challenging.

Our final analyses in Figure 2 (c) and (d) investigate

whether there is a link between the predictive power of the model and how often a word is used in text. We compute the frequencies of all affordances and properties occurring in natural language using the Google Web 1T corpus, an n-gram corpus computed from approximately one trillion words (Brants & Franz, 2006). Figure 2(c) plots the F1 score of properties against how frequently they appear in natural language; 2(d) plots the same for affordances. We include a best-fit line along with confidence intervals shown as one standard deviation of the data. We do not observe a statistical correlation between how much affordances and properties are written about, and how well neural models are able to connect their effects; a single confidence interval spans both positive and negative slopes. This lack of clear correlation is surprising, because large state-of-the-art neural textual models generally improve with repeated exposure to instances of words. Except for the three most common words measured by property F1 score, the rest of the data shows a strikingly uniform distribution of F1 scores for any choice of frequency in natural language. This suggests that current neural models are fundamentally limited in their capacity for physical reasoning, and that only new designs—not more data—can allow them to acquire this skill.

Discussion

Despite being able to associate a considerable range of information with the names of objects, neural models are not able to capture the more subtle interplay between affordances and properties. In some sense, this result is unsurprising. Collecting information around an object can be informed largely by the co-occurrence of words around that object's various mentions. Affordances that imply properties (and the reverse) are rarely mentioned together; their mutual connotation naturally renders joint expression redundant. Hence, priorless models that learn from statistical associations falter. Given the depth of the networks used in models such as ELMo and BERT, complex inter-parameter structure arises, but the latent semantic patterns that describe physical commonsense are much weaker than more superficial patterns that arise due to grammar or domain, making it difficult to capture.

This evidence feeds into theories of embodied cognition (Gover, 1996; Wilson, 2002), which suggest that the nature of human cognition depends strongly on the stimuli granted by physical experience. If this is so, then how is information encoded in our physical experience such that we can make predictions? If we assume a form of mental simulation, then what are the mental limits on its reliability? From an artificial intelligence perspective, the more interesting proof is in the principles of creating such a mental simulator. If we are to simulate human capacity for thought, how actually must we simulate elements of the physical world?

With the rise of physics engines, our ability to model physical inferences grows (Wu, Yildirim, Lim, Freeman, & Tenenbaum, 2015). However, while this may make us better at anticipating human predictions about physical situations through perceptual stimuli (Gerstenberg, Zhou, Smith, & Tenenbaum, 2017), there is still a long way to go before we understand the inferences that are being made through more symbolic stimuli, such as language. Exploring the mechanisms underlying this communication using an implicit shared world model will require us to either develop access to such a world model, or expose algorithms to predictions of that world model by directly querying humans. Bridging the inductive biases learned from simulation (Battaglia, Hamrick, & Tenenbaum, 2013) and those discovered by scientists (Lake, Linzen, & Baroni, 2019) to make inferences implicit in text will lead to a more cohesive model of commonsense physics. We expect such a model to bear great fruit in studies of communication rich with physical implications.

Acknowledgments

This work was supported by NSF grants (IIS-1524371, 1637479, 1703166), NSF Fellowship, the DARPA CwC program through ARO (W911NF-15-1-0543), and gifts by Google and Facebook. The views and conclusions contained herein are those of the authors and should not be interpreted as representing endorsements of the funding agencies.

References

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, *110*(45), 18327–18332.

Brants, T., & Franz, A. (2006). Web 1t 5-gram version 1.

Devereux, B. J., Tyler, L. K., Geertzen, J., & Randall, B. (2014). The centre for speech, language and the brain (cslb) concept property norms. *Behavior research methods*, *46*(4), 1119–1127.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Gaver, W. W. (1991). Technology affordances. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 79–84).

Gerstenberg, T., Zhou, L., Smith, K. A., & Tenenbaum, J. B. (2017). Faulty towers: A hypothetical simulation model of physical support. *Proceedings of the 39th Annual Conference of the Cognitive Science Society*.

Gibson, J. J. (1966). The senses considered as perceptual systems.

Gover, M. R. (1996). The embodied mind: Cognitive science and human experience (book). *Mind, Culture, and Activity*, *3*(4), 295–299.

Herbelot, A., & Vecchi, E. M. (2015). From concepts to models: some issues in quantifying feature norms. In *Lilt* (Vol. 2).

Lake, B. M., Linzen, T., & Baroni, M. (2019). Human few-shot learning of compositional instructions. *arXiv preprint arXiv:1901.04587*.

Levesque, H. J., Davis, E., & Morgenstern, L. (2012). The winograd schema challenge. In *Proceedings of the thirteenth international conference on principles of knowledge representation and reasoning* (pp. 552–561). AAAI Press. Retrieved from http://dl.acm.org/citation.cfm?id=3031843.3031909

Levy, O., & Goldberg, Y. (2014). Dependency-based word embeddings. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 2: Short papers)* (Vol. 2, pp. 302–308).

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision* (pp. 740–755).

Lucy, L., & Gauthier, J. (2017). Are distributional representations ready for the real world? evaluating word vectors for grounded perceptual meaning. *arXiv preprint arXiv:1705.11168*.

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior research methods*, *37*(4), 547–559.

Norman, D. (1988). *The design of everyday things: Revised and expanded edition*. Constellation.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp)* (pp. 1532–1543).

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 conference of the north american chapter of the association for computational linguistics: Human language technologies, volume 1 (long papers)* (Vol. 1, pp. 2227–2237).

Rubinstein, D., Levi, E., Schwartz, R., & Rappoport, A. (2015). How well do distributional models capture different types of semantic knowledge? In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 2: Short papers)* (Vol. 2, pp. 726–730).

Van Durme, B. D. (2010). *Extracting implicit knowledge from text*. University of Rochester.

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic bulletin & review*, *9*(4), 625–636.

Wu, J., Yildirim, I., Lim, J. J., Freeman, B., & Tenenbaum, J. (2015). Galileo: Perceiving physical object properties by integrating a physics engine with deep learning. In *Advances in neural information processing systems* (pp. 127–135).

Yatskar, M., Zettlemoyer, L., & Farhadi, A. (2016). Situation recognition: Visual semantic role labeling for image understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 5534–5542).

Continuous developmental change can explain discontinuities in word learning

Abdellah Fourtassi
afourtas@stanford.edu
Department of Psychology
Stanford University

Sophie Regan
sregan20@stanford.edu
Department of Psychology
University of Illinois

Michael C. Frank
mcfrank@stanford.edu
Department of Psychology
Stanford University

Abstract

Cognitive development is often characterized in term of discontinuities, but these discontinuities can sometimes be apparent rather than actual and can arise from continuous developmental change. To explore this idea, we use as a case study the finding by Stager and Werker (1997) that children's early ability to distinguish similar sounds does not automatically translate into word learning skills. Early explanations proposed that children may not be able to encode subtle phonetic contrasts when learning novel word meanings, thus suggesting a discontinuous/stage-like pattern of development. However, later work has revealed (e.g., through using simpler testing methods) that children do encode such contrasts, thus favoring a continuous pattern of development. Here we propose a probabilistic model describing how development may proceed in a continuous fashion across the lifespan. The model accounts for previously documented facts and provides new predictions. We collected data from preschool children and adults, and we showed that the model can explain various patterns of learning both within the same age and across development. The findings suggest that major aspects of cognitive development that are typically thought of as discontinuities, may emerge from simpler, continuous mechanisms.

Keywords: word learning, cognitive development, computational modeling

Introduction

Cognitive development is sometimes characterized in terms of a succession of discontinuous stages (Piaget, 1954). Although intuitively appealing, stage theories can be challenging to integrate with theories of learning, which typically posit that knowledge and skills improve incrementally with experience. Indeed, one of the central challenges of cognitive development has been to explain transitions between stages which appear to be qualitatively different (Carey, 2009).

Nevertheless, at least in some cases, development may only appear to be stage-like. This appearance can be due, for example, to the use of a cognitively-demanding task which may mask learning, or to the use of statistical thresholding (in particular, p-value < 0.05) which can create a spurious dichotomy between success and failure in observing a given behavior. In such cases, positing discontinuous stages is unnecessary. Instead, a continuous model—involving similar representations across the lifespan—may provide a simpler and more transparent account of development.

We use a case study from word learning literature. Stager & Werker (1997) first showed that children's early ability to distinguish similar sounds does not automatically translate into word learning skills. Indeed, though infants around 14-month old can distinguish similar sound pairs such as "dih" and "bih", they appear to fail in mapping this pair to two different objects. Follow-up studies have focused on proposing possible explanations for this observed gap between speech perception and word learning (e.g., Fennell & Waxman, 2010; Hofer & Levy, 2017; Rost & McMurray, 2009; Stager & Werker, 1997).

By around 17 m.o, children succeed in the same task (Werker, Fennell, Corcoran, & Stager, 2002). How does development proceed? Early accounts assumed that children encode words in a binary way: they either fail or succeed in encoding the relevant phonetic details (simultaneously with the meanings). This account suggested a discontinuous/stage-like pattern of development whereby younger children fail to encode the contrastive phonetic detail, whereas older children succeed.

Subsequent findings have suggested otherwise. On the one hand, 14-month-olds—who typically fail in the original task—succeed when an easier testing method is used, even under the same learning conditions (Yoshida, Fennell, Swingley, & Werker, 2009). They also succeed when uncertainty is mitigated via disambiguating cues (e.g., Thiessen, 2007). On the other hand, adults show patterns of learning similar to those shown by 14-month-olds when the task is more challenging and when the similarity between words increases (Pajak, Creel, & Levy, 2016; White, Yee, Blumstein, & Morgan, 2013).

This pattern of evidence points towards another scenario, where the representations are encoded in a probabilistic (rather than binary) way, and where development is continuous, rather than stage-like (see also Swingley, 2007). On this account, correct representations are learned early in development, but these representations are encoded with higher uncertainty in younger children, leading to apparent failure in relatively demanding tasks. Development is a continuous process whereby the initial noisy representations become more precise. In addition, more precise representations are still imperfect: Even adults show low accuracy learning when the sound contrasts are subtle, e.g., non-native sounds (Pajak et al., 2016).

We provide an intuitive illustration of how such an account explains patterns of learning and development in Figure 1. We observe low accuracy in word learning when the perceptual distance between the labels is small relative to the uncertainty with which these labels are encoded. For example, in Stager and Werker's original experiment, children are supposed to associate label 1 ("bih") and label 2 ("dih") with object 1 and object 2, respectively. Though infants could learn that the label "bih" is a better match to object 1 than "dih", they could still judge the sound "dih" as a plausible instance of the label "bih", thanks to the relatively large uncertainty

of the encoding, and this confusion leads to "failure" in the recognition task. According to this account, accuracy in word learning improves if we increase either the perceptual distinctiveness of the stimuli (e.g., through using different-sounding labels) or the precision of the encoding itself (e.g., across development).

Building on this intuition, the current work proposes a probabilistic model, which we use to both account for previous experimental findings, and to make new predictions that have not been tested before. Using new data collected from both preschool children and adults, we show that the model can explain various patterns of learning both within the same age and across development.

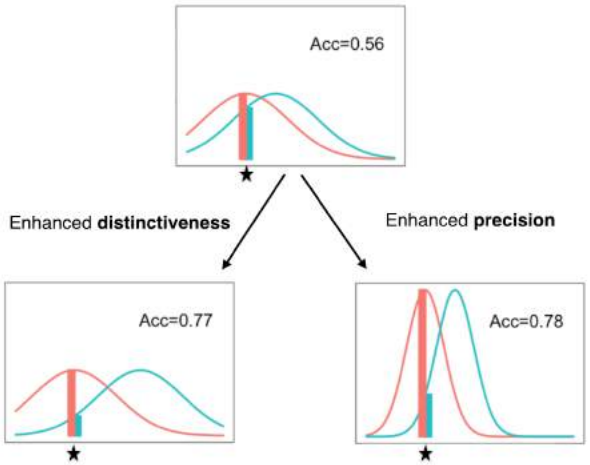


Figure 1: An illustration of the probabilistic/continuous account using simulated data. A word is represented with a distribution over the perceptual space (indicated in red or blue). When the uncertainty of the representation is large relative to the distance between the stimuli (top panel), an instance of the red category (indicated with a star) could also be a plausible instance of the green category, hence the low recognition accuracy score. The accuracy increases when the stimuli are less similar (left panel), or when the representations are more precise (right panel).

Model

Probabilistic structure

Our model consists of a set of variables describing the general process of spoken word recognition in a referential situation. These variables are related in a way that reflects the simple generative scenario represented graphically in Figure 2. When a speaker utters a sound in the presence of an object, the observer assumes that the object o activated the concept C in the speaker's mind. The concept prompted the corresponding label L. Finally, the label was physically instantiated by the sound s.

A similar probabilistic structure was used by Lewis & Frank (2013) to model concept learning, and by Hofer & Levy (2017) to model spoken word learning. However, the first study assumed that the sounds are heard unambiguously, and the second assumed the concepts are observed unambiguously. In our model, we assume that both labels and concepts are observed with a certain amount of perceptual noise, which we assume, for simplicity, is captured by a normal distribution:

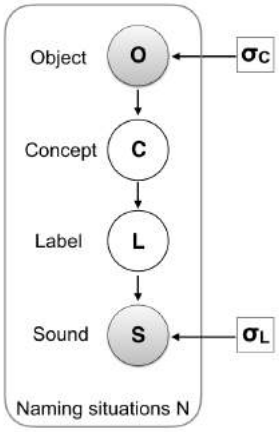


Figure 2: Graphical representation of our model. Circles indicate random variables (shading indicates observed variables). The squares indicate fixed model parameters.

$$p(o|C) \sim \mathcal{N}(\mu_C, \sigma_C^2)$$
$$p(s|L) \sim \mathcal{N}(\mu_L, \sigma_L^2)$$

Finally, we assume there to be one-to-one mappings between concepts and labels and that observers have successfully learned these mappings during the exposure phase:

$$P(L_i|C_j) = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

Inference

The learner hears a sound s and has to decide which object o provides an optimal match to this sound. To this end, they must compute the probability $P(o|s)$ for all possible objects. This probability can be computed by summing over all possible concepts and labels:

$$P(o|s) = \sum_{C,L} P(o,C,L|s) \propto \sum_{C,L} P(o,C,L,s)$$

The joint probability $P(o,C,L,s)$ is obtained by factoring the Bayesian network in Figure 2:

$$P(o,C,L,s) = P(s|L)P(L|C)P(C|o)P(o)$$

which can be transformed using Bayes rule into:

$$P(o,C,L,s) = P(s|L)P(L|C)P(o|C)P(C)$$

Finally, assuming that the concepts' prior probability is uniformly distributed[1], we obtain the following expression, where all conditional dependencies are now well defined:

$$P(o|s) = \frac{\sum_{C,L} P(s|L)P(o|C)P(L|C)}{\sum_o \sum_{C,L} P(s|L)P(o|C)P(L|C)} \quad (1)$$

[1]This is a reasonable assumption in our particular case given the similarity of the concepts used in each naming situation in our experiment.

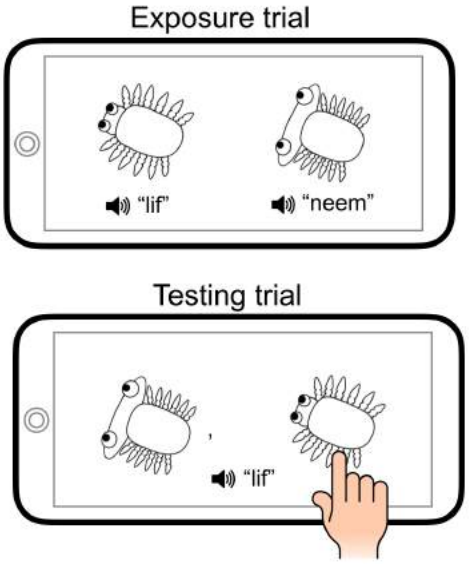


Figure 3: An overview of the task used in this study.

Task and model predictions

We use the model to predict performance in the word learning task introduced by Stager & Werker (1997), with a two-alternative forced choice as in Yoshida et al. (2009). In this task, participants are first exposed to the association between pairs of nonsense words (e.g., "lif"/"neem") and pairs of objects. The word-object associations are introduced sequentially. After this exposure phase, participants perform a series of test trials. In each of these trials, one of the two sounds is uttered (e.g., "lif") and participants choose the corresponding object from the two alternatives. An overview of the task is shown in Figure 3.

We used Equation 1 and the probability distributions defined above to obtain the exact analytical expression for the probability of accurate responses $p(o_T|s)$ (target object o_T given a sound s) in the simple case of two-alternative forced choice in the testing phase of our experimental task:

$$P(o_T|s) = \frac{1 + e^{-(\Delta s^2/2\sigma_L^2 + \Delta o^2/2\sigma_C^2)}}{1 + e^{-(\Delta s^2/2\sigma_L^2 + \Delta o^2/2\sigma_C^2)} + e^{-\Delta s^2/2\sigma_L^2} + e^{-\Delta o^2/2\sigma_C^2}} \quad (2)$$

Figure 4 show simulations of the predicted accuracy (Expression 2) as a function of the distinctiveness parameters (Δs and Δo) and the precision parameters, i.e., the variances of the distributions $p(s|L)$ and $p(o|C)$. To understand the qualitative behavior of the model, we assumed for simplicity that the precision parameter has similar values in both distributions, i.e., $\sigma = \sigma_C \approx \sigma_L$ (but we will allow those parameters to vary independently in the rest of the paper).

The simulations explain some previously documented facts, and make new predictions:

1) For fixed values of Δo and σ, the probability of accurate responses increases as a function of Δs. This pattern accounts for the fact that similar sounds are generally more challenging to learn than different sounds for both children (Stager & Werker, 1997) and adults (Pajak et al., 2016).

2) For fixed values of Δs and Δo, accuracy increases when the representational uncertainty (characterized with σ) decreases. This fact may explain development, i.e., younger children have noisier representations (see Swingley, 2007; Yoshida et al., 2009), which leads to lower word recognition accuracy, especially for similar-sounding words.

3) For fixed values of Δs and σ, accuracy increases with the visual distance between the semantic referents Δo. This is a new prediction that our model makes. Previous work studied the effect of several bottom-up and top-down properties in disambiguating similar sounding words (e.g., Fennell & Waxman, 2010; Rost & McMurray, 2009; Thiessen, 2007), but to our knowledge, no previous study in the literature tested the effect of the visual distance between the semantic referents.

Experiment

In this experiment, we tested participants in the word learning task introduced above (Figure 3). More precisely, we explored the predictions related to both distinctiveness and precision. Sound similarity (Δs) and object similarity (Δo) were varied simultaneously in a within-subject design. Two age groups (preschool children and adults) were tested on the same task to explore whether development can be characterized with the uncertainty parameters, σ_C and σ_L. The experiment, sample size, exclusion criteria and the model's main predictions were pre-registered.

Methods

Participants We planned to recruit a sample of $N = 60$ children ages 4-5 years from the Bing Nursery School on Stanford University's campus. Here we report data from $N = 55$ children. An additional $N = 35$ children participated but were removed from analyses because they were not above chance on the catch trials due to the challenging nature of our procedure (see below). We also collected a planned sample of $N = 100$ adult participants through Amazon Mechanical Turk. We planned to exclude data from participants who did not do well on the catch trials ($N = 26$) and from participants who were familiar with the non-English sound stimuli we used in the adult experiment ($N = 0$), yielding a final sample of $N = 74$.

Stimuli and similarity rating The sound stimuli were generated using the MBROLA Speech Synthesizer (Dutoit, Pagel, Pierret, Bataille, & Van der Vrecken, 1996). We generated three kinds of nonsense word pairs which varied in their degree of similarity to English speakers: 1) "different": "lif"/"neem" and "zem"/"doof", 2) "intermediate": "aka"/"ama" and "ada"/"aba", and 3) "similar" non-English minimal pairs: "ada"/"ad$^{\text{h}}$a" (in hindi) and "aʕa"/"a a" (in arabic).

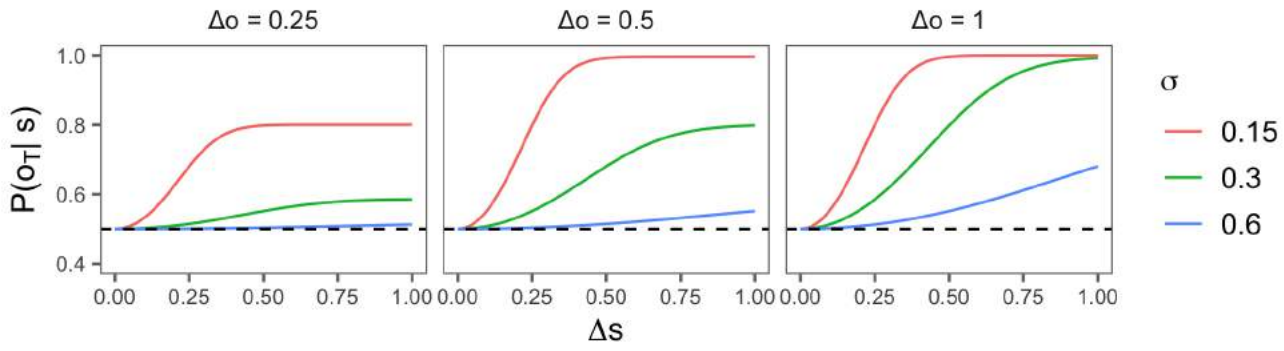


Figure 4: The predicted probability of accurate responses in the testing phase as a function of stimuli distinctiveness Δs and Δo and representation precision σ (for simplicity, we assume here that $\sigma{=}\sigma_C{=}\sigma_L$). Dashed line represents chance.

As for the objects, we used the Dynamic Stimuli javascript library[2] which allowed us to generate objects in four different categories: "tree", "bird", "bug", and "fish". These categories are supposed to be naturally occurring kinds that might be seen on an alien planet. In each category, we generated "different", "intermediate" and "similar" pairs by manipulating a continuous property controlling features of the category's shape (e.g, body stretch or head fatness).

In a separate survey, $N = 20$ participants recruited on Amazon Mechanical Turk evaluated the similarity of each sound and object pair on a 7-point scale. We scaled responses within the range [0,1]. Data are shown in Figure 5, for each stimulus group. These data will be used in the models as the perceptual distance of sound pairs (Δs) and object pairs (Δo).

Design Each age group saw only two of the three levels of similarity described in the previous sub-section: "different" vs. "intermediate" for preschoolers and "intermediate" vs. "similar" for adults. We made this choice in light of pilot studies showing that adults were at ceiling with "different" sounds/objects, and children were at chance with the "similar" sounds/objects. That said, this difference in the level of similarity is accounted for in the model by using the appropriate perceptual distance used in each age group (Figure 5).

To maximize our ability to measure subtle stimulus effects, the experiment was a 2x2 within-subjects factorial design with four conditions: high/low sound similarity crossed with high/low visual object similarity. Besides the 4 conditions, we also tested participants on a fifth catch condition which was similar in its structure to the other ones but was used only to select participants who were able to follow the instructions and show minimal learning.

Procedure Preschoolers were tested at the nursery school using a tablet, whereas adults used their own computers to complete the same experiment online. Participants were tested in a sequence of five conditions: the four experimental conditions plus the catch condition. In each condition, participants saw a first block of four exposure trials followed by four testing trials, and a second block of two exposure trials (for memory refreshment) followed by an additional four testing trials. The length of this procedure was demanding, especially for children, but we adopted a fully within-subjects design based on pilot testing that indicated that precision of measurement was critical for testing our experimental predictions.

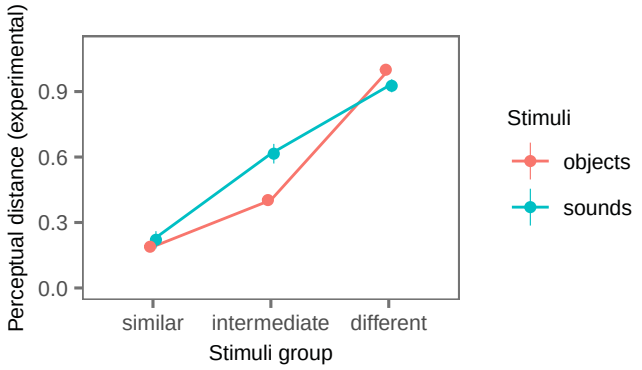


Figure 5: Distances for both sound and object pairs from an adult norming study. Data represent Likert values normalized to [0,1] interval. Error bars represent 95% confidence intervals.

In the exposure trials, participants saw two objects associated with their corresponding sounds. We presented the first object on the left side of the tablet's screen simultaneously with the corresponding sound. The second sound-object association followed on the other side of the screen after 500ms. For both objects, visual stimuli were present for the duration of the sound clip (800ms). In the testing trials, participants saw both objects simultaneously and heard only one sound. They completed the trial by selecting which of the two objects corresponded to the sound. The object-sound pairings were randomized across participants, as was the order of the conditions (except for the catch condition which was always placed in the middle of the testing sequence). We also randomized

[2]https://github.com/erindb/stimuli

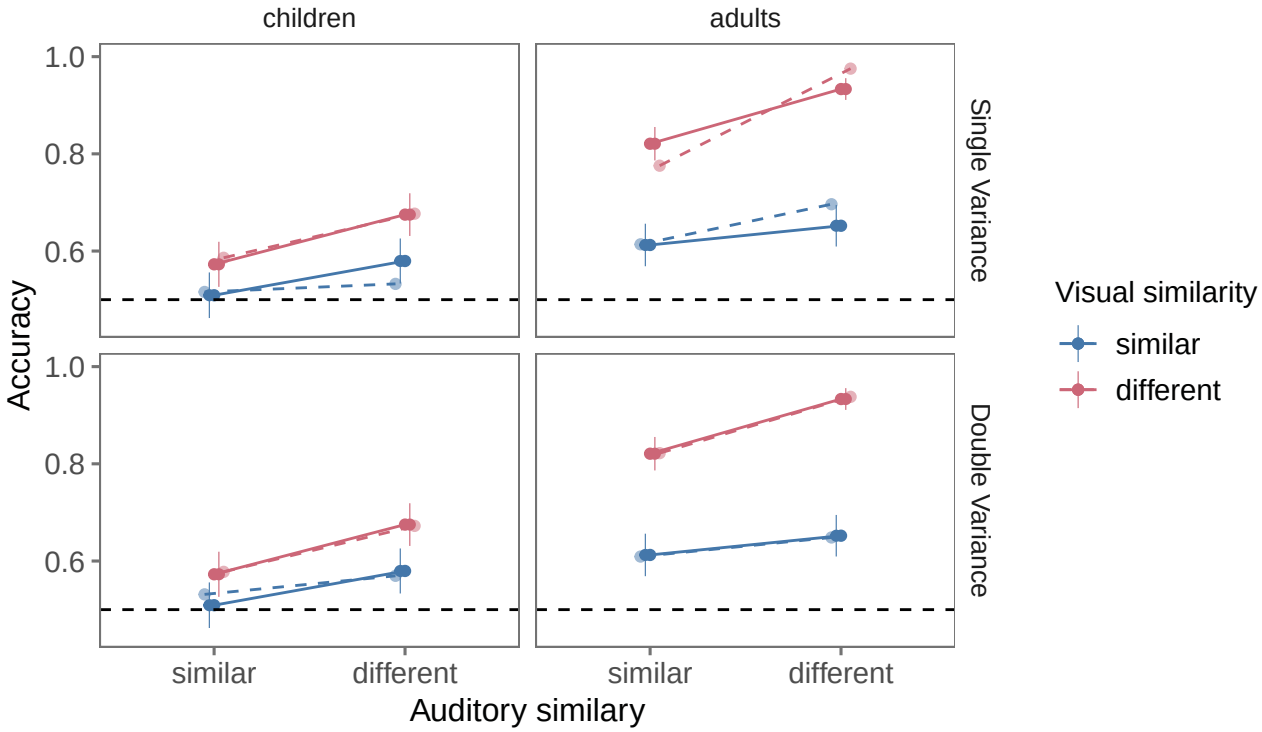


Figure 6: Accuracy of novel word recognition as as a function of the sound distance, the object distance, and the age group (preschool children vs. adults). We show both the models' predictions (dashed lines) and the experimental results (solid lines). Error bars represent 95% confidence intervals.

the on-screen position (left vs. right) of the two pictures on each testing trial.

Results

We first analyzed the results using a mixed-effects logistic regression with sound distance, object distance and age group as fixed effects, and with a maximal random effects structure (allowing us to take into account the full nested structure of our data) (Barr, Levy, Scheepers, & Tily, 2013). We found main effects for all the fixed effects in the regression. For the sound distance, we obtained $\beta = 0.52$ ($p < 0.001$), replicating previous findings. For object distance, we found $\beta = 0.83$ ($p < 0.001$), and this finding confirms the new prediction of our model. Finally, for the age group, we obtained $\beta = 0.76$ ($p < 0.001$), showing that performance improves with age.

We next fit our model (using Equation 2) to the participants' responses in each age group using non-linear least-squares. The values of Δs and Δo were set based on data from the similarity judgment task (Figure 5). The model has two degrees of freedom for each group, i.e., σ_C and σ_L. We call it the double-variance model. Figure 6 (dashed lines) shows the predictions. The double-variance model captures the behavioral patterns in both age groups: starting from a low accuracy recognition when both the sound and object distances are small, the model correctly predicts an increase in accuracy when either the sound distance or the object distance increases. Further, accuracy is correctly predicted to be maximal when both the sound and object distances are high.

The values of the parameters were as follows. Children had a label-specific uncertainty of $\sigma_S = 0.83$ [0.64, 1.02][3], and a concept-specific uncertainty of $\sigma_C = 0.31$ [0.11, 0.51]. Adults had a label-specific uncertainty of $\sigma_S = 0.12$ [0.12, 0.13], and a concept-specific uncertainty of $\sigma_C = 0.17$ [0.16, 0.18]. As predicted, the uncertainty parameters were larger for children than they were for adults, showing that the probabilistic representations becomes more refined (that is, σ becomes smaller) across development. The developmental effect was more important for the label-specific uncertainty.

The double-variance model explained almost all the variance in the participants' mean responses. To investigate whether the model's strong predictive power was due to overfitting, we fit a simplified version with only one degree of freedom (i.e., a single variance common to both sounds and objects). This single-variance model also captured the main qualitative patterns and remained highly predictive ($R^2 = 0.95$). This result suggests that the explanatory power of the model is largely due to its structure, rather than its degrees of freedom.

General Discussion

This paper explored the idea that some seemingly stage-like patterns in cognitive development can be characterized in a

[3] All uncertainty intervals in this paper represent 95% Confidence Intervals.

continuous fashion. We used as a case study the seminal work of Stager & Werker (1997) showing a discrepancy between children's speech perception abilities and their word learning skills. While much of the previous investigation of this finding has been interested in the source of this discrepancy, here we have explored how it could arise from continuous developmental change in perceptual uncertainty.

Building on some previous discussions (e.g., Swingley, 2007; Yoshida et al., 2009), we proposed a model where perceptual stimuli are encoded probabilistically. We tested the model's predictions against data collected from preschool children and adults and we showed that developmental changes in word-object mappings can indeed be characterized as a continuous refinement (i.e., uncertainty reduction) in qualitatively similar representations across the life span.

The model made a new prediction which we tested experimentally: Learning similar words is not only modulated by the similarity of their phonological forms, but also by the visual similarity of their semantic referents. More generally, since visual similarity is an early organizing feature in the semantic domain (e.g., Wojcik & Saffran, 2013), our finding suggests that children may prioritize the acquisition of words that are quite distant in the semantic space. This suggestion is supported by recent findings based on the investigation of early vocabulary growth (Engelthaler & Hills, 2017; Sizemore, Karuza, Giusti, & Bassett, 2018).

One limitation of this work is that the model was fit to data from children at a relatively older age (4-5 years old) than what is typically studied in the literature (14-18 month-old). We selected this older age group to optimize the number and precision of the experimental measures (both are crucial to model fitting). Data collection involved presenting participants with several trials across four conditions in a between-subject design. It would have been challenging to obtain such measures with infants.

In sum, this paper proposes a model that accounts for the development of an important aspect of word learning. Our account suggests that the developmental data can be explained based on a continuous process operating over similar representations across development, suggesting developmental continuity. We used a case from word learning as an example, but the same idea might apply to other aspects of cognitive development that are typically thought of as stage-like (e.g., acquisition of a theory of mind). Computational models, such as the one proposed here, can help us investigate the extent to which such discontinuities emerge due to genuine qualitative changes and the extent to which they reflect the granularity of the researchers' own measurement tools.

All data and code are available online at
`https://github.com/afourtassi/kidswitch`

References

Barr, D., Levy, R., Scheepers, C., & Tily, H. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3).

Carey, S. (2009). *The origin of concepts*. Oxford University Press.

Dutoit, T., Pagel, V., Pierret, N., Bataille, F., & Van der Vrecken, O. (1996). The mbrola project: Towards a set of high quality speech synthesizers free of use for non commercial purposes. In *Proceedings of ICSLP* (Vol. 3). IEEE.

Engelthaler, T., & Hills, T. T. (2017). Feature biases in early word learning: Network distinctiveness predicts age of acquisition. *Cognitive Science*, *41*.

Fennell, C., & Waxman, S. (2010). What paradox? Referential cues allow for infant use of phonetic detail in word learning. *Child Development*, *81*.

Hofer, M., & Levy, R. (2017). Modeling Sources of Uncertainty in Spoken Word Learning. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*.

Lewis, M., & Frank, M. (2013). An integrated model of concept learning and word-concept mapping. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 35).

Pajak, B., Creel, S., & Levy, R. (2016). Difficulty in learning similar-sounding words: A developmental stage or a general property of learning? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(9).

Piaget, J. (1954). *The construction of reality in the child*. New York, NY, US: Basic Books.

Rost, G., & McMurray, B. (2009). Speaker variability augments phonological processing in early word learning. *Developmental Science*, *12*.

Sizemore, A. E., Karuza, E. A., Giusti, C., & Bassett, D. S. (2018). Knowledge gaps in the early growth of semantic feature networks. *Nature Human Behaviour*, *2*(9).

Stager, C., & Werker, J. (1997). Infants listen for more phonetic detail in speech perception than in word-learning tasks. *Nature*, *388*(6640).

Swingley, D. (2007). Lexical exposure and word-form encoding in 1.5-year-olds. *Developmental Psychology*, *43*(2).

Thiessen, E. (2007). The effect of distributional information on children's use of phonemic contrasts. *Journal of Memory and Language*, *56*.

Werker, J., Fennell, C., Corcoran, K., & Stager, C. (2002). Infants' ability to learn phonetically similar words: Effects of age and vocabulary size. *Infancy*, *3*.

White, K., Yee, E., Blumstein, S., & Morgan, J. (2013). Adults show less sensitivity to phonetic detail in unfamiliar words, too. *Journal of Memory and Language*, *68*(4).

Wojcik, E., & Saffran, J. (2013). The ontogeny of lexical networks: Toddlers encode the relationships among referents when learning novel words. *Psychological Science*, *24*(10).

Yoshida, K., Fennell, C., Swingley, D., & Werker, J. (2009). 14-month-olds learn similar-sounding words. *Developmental Science*, *12*.

Extracting and Utilizing Abstract, Structured Representations for Analogy

Steven M. Frankland (steven.frankland@princeton.edu)
Princeton University

Taylor W. Webb
Princeton University

Alexander A. Petrov
The Ohio State University

Randall C. O'Reilly
University of California, Davis

Jonathan D. Cohen
Princeton University

Abstract

Human analogical ability involves the re-use of abstract, structured representations within and across domains. Here, we present a generative neural network that completes analogies in a 1D metric space, without explicit training on analogy. Our model integrates two key ideas. First, it operates over representations inspired by properties of the mammalian Entorhinal Cortex (EC), believed to extract low-dimensional representations of the environment from the transition probabilities between states. Second, we show that a neural network equipped with a simple predictive objective and highly general inductive bias can learn to utilize these EC-like codes to compute explicit, abstract relations between pairs of objects. The proposed inductive bias favors a latent code that consists of anti-correlated representations. The relational representations learned by the model can then be used to complete analogies involving the signed distance between novel input pairs (1:3 :: 5:? (7)), and extrapolate outside of the network's training domain. As a proof of principle, we extend the same architecture to more richly structured tree representations. We suggest that this combination of predictive, error-driven learning and simple inductive biases offers promise for deriving and utilizing the representations necessary for high-level cognitive functions, such as analogy.

Keywords: abstract structured representations; analogy; neural networks; predictive learning; relational reasoning;

Introduction

Analogy requires the flexible, yet orderly, transfer of abstract knowledge within and between domains. Although this transfer occasionally enables new theoretical insights (e.g., Rutherford's planetary model of atomic structure or the hydraulic model of blood circulation (Gentner, 1983)), it also provides critical support for basic cognitive functions, such as memory retrieval, categorization, and schema induction (Gick & Holyoak, 1983; Doumas, Hummel, & Sandhofer, 2008; Gentner & Forbus, 2011; Holyoak, 2012). Here, we test the idea that representations of abstract, structural relationships can be derived using simple forms of training. We evaluate whether these representations can, in turn, support higher-level functions such as analogical inference, without any explicit training on analogy itself.

Specifically, we test the idea that abstract, structured representations arise from learning systems that (a) exploit the vast amounts of observational data available to natural agents ((Rao & Ballard, 1999), O'Reilly, Wyatte, and Rohrlich, 2017) coupled with (b) particular inductive biases in the learning algorithm and network architecture.

We explore what particular input representations and model architectures enable the extraction and utilization of this relational knowledge. To do so, we focus on modeling signed distance relations in a 1-dimensional (1D) domain (e.g., analogies such as 1:3 :: 5:7), and conclude by extending the principles to analogies involving a simple family tree structure.

Throughout, we take inspiration from the representational properties of the mammalian Entorhinal Cortex (EC), believed to extract low-dimensional representations of the metric structure common across environments from the transition probabilities between states (Dayan, 1993; Gustafson & Daw, 2011; Stachenfeld, Botvinick, & Gershman, 2017). We then propose a learning procedure that combines error-driven learning and an inductive bias that naturally extracts explicit relational representations from pre-structured, EC-like representations of the current domain. We show that this combination of representation and model architecture can support analogical inference, and extrapolate well outside the network's training domain.

Methods

Input Representations We focus first on the 1D domain. What input representations enable relation learning that can support analogy (e.g., Figure 1A)? Here, we compare place codes, a successor representation (SR) of states, and lower dimensional representations generated by eigendecompositions of this SR matrix, truncated to 10, 5, or 2 components.

For the place code representation, locations are represented as 1xN one-hot vectors, each orthogonal to the others. This set of representations forms an nXn identity matrix (see Figure 1C, upper left panel), and thus lacks intrinsic structural

information about the domain. It is therefore, a priori, a poor candidate for discovering relational structure that could support analogy, which requires knowledge about structural equivalences (e.g., that the difference between 1-3 is the same as 5-7).

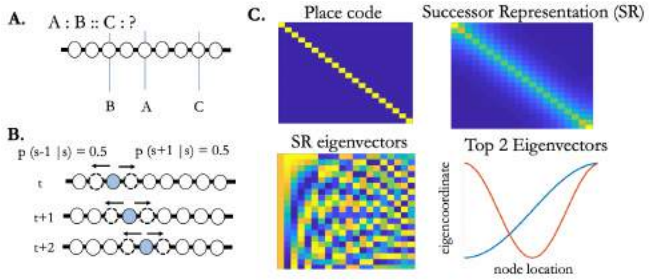


Figure 1: Schematic representation of the 1D domain. (A) depicts the task the network will perform: a signed distance analogy in 1D (here, 5:3 :: 8:?). (B). Following Stachenfeld et al. (2017), we use a random-walk transition matrix to compute the Successor Representation (SR), and derive a low-dimensional representation of the domain by computing the eigendecomposition of the SR. (C) The three forms of representations are compared as inputs to the network: place codes, the SR, and various number of eigenvectors of the SR.

We therefore extend the representation to carry information about the similarity structure of the dimension. To do so, we use the *successor representation* (SR) (Dayan, 1993). The SR is a predictive representation of possible future states, and thus carries information about the dimension's intrinsic structure. To illustrate, imagine, an animal navigating a linear track, with equal probability of moving to spatially adjacent states at successive time points (See Figure 1B). Experience along the track allows learning state-state transition probabilities. The SR encodes the likelihood of visiting a particular location s_i, given a trajectory initiated in a particular state s. The resulting nXn SR matrix is a collection of these conditional probabilities (See Figure 1C, upper right panel). The probability of visitation is high for neighboring states, but low for distant states. Although the SR has primarily been used in learning spatial structure, we take this approach to be representative of continuous dimensions, more broadly (e.g., *brightness*, *loudness*, *pitch*, etc., as well as semantic dimensions such as *animacy*, *size*, or *agreeableness*). The SR can be acquired using a temporal difference learning algorithm (Dayan, 1993). Here, however, we simply stipulate a transition matrix with equal probability of transitioning to adjacent states and a discounted value of future states of 0.9 (results are robust to this parameter). For the SR, the resulting input representation is an nXn matrix in which nearby states have similar representations.

Representing this similarity structure makes the SR a better candidate for analogical inference than a pure place code. However, it is still a high-dimensional representation that does not compactly encode variation along the dimension of interest. Critically, recent work has shown that a spectral representation of the SR provides an abstract, implicitly structured representation of the domain (Stachenfeld et al., 2017). Intriguingly, this spectral representation of the SR resembles the representational properties of grid-like cells in medial entorhinal cortex. Entorhinal cortex (EC) is believed to encode low-dimensional representations of the metric structure of the environment (Hafting et al., 2005; Gustafson and Daw, 2011; Stachenfeld et al., 2017). Here, we take grid-cells in EC to reflect a more general idea: the decomposition of transition probabilities into low-dimensional embeddings provides an abstract, implicitly structured representation of a domain. When domains share an underlying transition structure, these representations can be used to support analogies. To implement our 1D case, we follow Stachenfeld et al. (2017) and compute the eigendecomposition of the square (nXn) SR matrix. The eigenvalues (λ) are obtained solving $det(M-\lambda I)=0$, where M is the SR matrix, and I the identity matrix. We can obtain the corresponding eigenvector (U_i) for a particular eigenvalue (λ_i) by solving $MU_i = U_i\lambda_i$ for U_i. Finally, we encode each location in the native space using its eigencoordinates ($U\lambda$) (Figure 1C, lower left panel). We compare models trained on the top 10, top 5, top 2 highest eigenvalue eigenvectors (e.g., Figure 1C, lower right panel), and the neural network learns over these eigencoordinate representations, which implicitly encode the ordering and neighborhood relations among the locations. Finally, we include a scrambled version of the top 2 eigenvectors. This scrambled version removes the true similarity structure in the native space, and serves as a control to ensure that the neural network is not simply memorizing its inputs.

Model Architectures. We trained and tested networks on each of the input representations described above. Each network was constructed to take a pair of inputs (A and B — we refer to objects and relations, themselves, using non-italicized capitalized letters, and a model's representations of them using italicized lower-case (e.g., a and b)), and trained to predict each member of the pair from the other. We evaluated five competing model architectures, that we describe in detail below. In all models, representations of A converge on a common learned internal layer, which can then encode their relationship R. The ability to form a systematic, abstract relational encoding (r) is a primary focus of this paper. We evaluate the ability of r to support simple metric analogies by completing problems of the form A:B as C:? (e.g., 1:3 is to 5:?, answer: 7).

r is computed as,

$$r = f_\Theta(a,b)$$

where a and b are representations of the two input objects to be related, and Θ are the learned encoder parameters (by default, a one layer multiple layer perceptron (MLP) with 100 rectified linear units (RELU)) and (r) is the activation state across two linear nodes.

To train the model, the latent state r is composed with the object A representation (a) to predict B ($\hat{b}$), and composed with the object B representation (b) to predict A ($\hat{a}$).

$$\hat{a} = g_{\Phi}(r,b) \quad \hat{b} = g_{\Phi}(r,a)$$

where, $\hat{a}$ is the predicted version of a, g is the decoder function (a one layer MLP with 100 RELU units), with learned parameters Φ. r is $r = f_{\Theta}(a,b)$, as above, and b is the input representation of object B. Likewise for $\hat{b}$, *mutatis mutandis*.

The representations r , a , and b are thus arguments to functions parameterized by the encoder and decoder. Here, we control the selection and application of these arguments by hand, but not their values, therein focusing on the problem of representation learning.

Intuitively, the objective of the model is to learn a representation that enables transformation of the representation a to match representation b, and vice versa. The loss is the mean squared error of the reconstruction across both objects, and weights of the encoder (Θ) and decoder (Φ) are jointly updated using standard backpropagation (Rumelhart, Hinton, & Williams, 1985). We emphasize that the network is trained to minimize reconstruction error ($(a-\hat{a})^2 + (b-\hat{b})^2$), not to complete analogies.

Model architectures are shown in Figure 2. All 5 models consist of two linear nodes in the latent space (r), but differ in how they constrain the nodes in r to be used. For models 1 and 2, the same 2D vector of r is used in both $\hat{a} = g_{\Phi}(r,b)$ and $\hat{b} = g_{\Phi}(r,a)$. Model 2 differs from Model 1 only in allowing separate decoder parameters to be learned for $\hat{a} = g_{\Phi_a}(r,b)$ and $\hat{b} = g_{\Phi_b}(r,a)$. These models are free to learn how to use these nodes (r) to predict across a and b. By contrast, models 3, 4, and 5 specifically commit each of these two nodes to separately predicting a and b, which we refer to as r_a and r_b, respectively. That is, $\hat{a} = g_{\Phi}(r_a,b)$, and $\hat{b} = g_{\Phi}(r_b,a)$. Models 3, 4, and 5 all share decoder parameters across $\hat{a}$. and $\hat{b}$.

Models 4 and 5 involve chaining the two latent nodes in r (r_a and r_b), such that one of the components of the relational representation is a function of the other. The motivation for this chaining of latent nodes is as follows. Recall that activation in r_a encodes information necessary to transform b to a when passed through the decoder, and vice versa for r_b. Although the input representation may contain *implicit information* about the relationship between A and B, the encoder must extract this information and explicitly represent relational information in a low-dimensional form. Logically, in a metric space, how A relates to B cannot change without affecting how B relates to A – the two predictive tasks imposed on network. The chain weights in models 4 and 5 force such a bi-directional dependence in r (See Figure 2). While it is possible that an unconstrained neural network could learn the dependence between r_a and r_b, we find that, using the current objective, this does not occur reliably without imposing the constraint that $r_a = f_{\Theta}(a,b)$, and $r_b = r_a W + bias$.

Model 4 imposes the weakest version of this constraint, by allowing the weight W (that links r_a and r_b) to be randomly initialized. However, randomly initializing the chain weight fails to exploit all the world-structure that may be easily incorporated. Critically, the general relation between the components r_a and r_b should be anti-correlated (e.g., the "bigger" A is than B, the "smaller" B is than A). That is, r_a, r_b can be thought of as conjugate pairs (e.g., +-2). A prior that biases the network toward the discovery of this relational structure can easily be implemented by initializing W to -1. We refer to this architecture as the *conjugate symmetry prior*, as it favors extracting a representation in r that treats r_a and r_b as a conjugate pair, reflected about zero. Note that this was strictly an initialization, and that W was free to vary over training. Empirically, we find that it tends not to vary over the course of training when initialized to -1.

Analogy Evaluation. We address whether the trained networks can perform analogy as follows. First, two object representations (a and b) are passed to the network, and the resulting latent state (r) is computed, and manually clamped to use for analogical completion. Next, a third object representation (c) is passed directly to the decoder without modifying r. The decoder then generates the expected d ($\hat{d}$), given c and the latent state r that was previously computed from the ab pair, as $\hat{d} = g_{\Phi}(r,c)$. To quantify a network's analogical ability, we compute the cosine similarity between $\hat{d}$ and d, as well as the cosine similarity between d and three randomly chosen foils. If the correct mapping is more similar than all three incorrect mappings, the network is determined to have succeeded on that trial. Chance performance on this metric is thus 25%.

Our primary results involve a 1x80 space (Figure 3). We also explore 1x20, 1x40, and 1x160 spaces, and the ability to generalize between them (Figure 5). In all cases, we trained the model on 500 ab pairs from the same space, and held 100 unique pairs from that space out of training for model testing. The number of possible pairs varies by the size of the native space. For example, in the smallest space (1x20), there are 380 possible ab pairs, entailing that some were necessarily repeated in training. At the other extreme (1x160), there are 25,540 possible pairs, and only 500 were selected for training, meaning that the network only experienced 2% of the possible space of relation-tokens. We trained the network using batch sizes of 64 samples and the ADAM optimizer (with a constant learning rate of $3x10^{-4}$), implemented in Tensorflow. We trained all networks for a fixed number of 1000 batches. This was sufficient to observe reliable asymptotes in training loss. Unless otherwise indicated, we ran 50 replications of each network with random encoder and decoder weights and random samples from the training set, and plot the mean and 95% CI for each class of networks and input representations.

Results

Analogy Results Across Input Representations and Architectures. In the models explored here, we find that success on these analogies requires both a particular input representa-

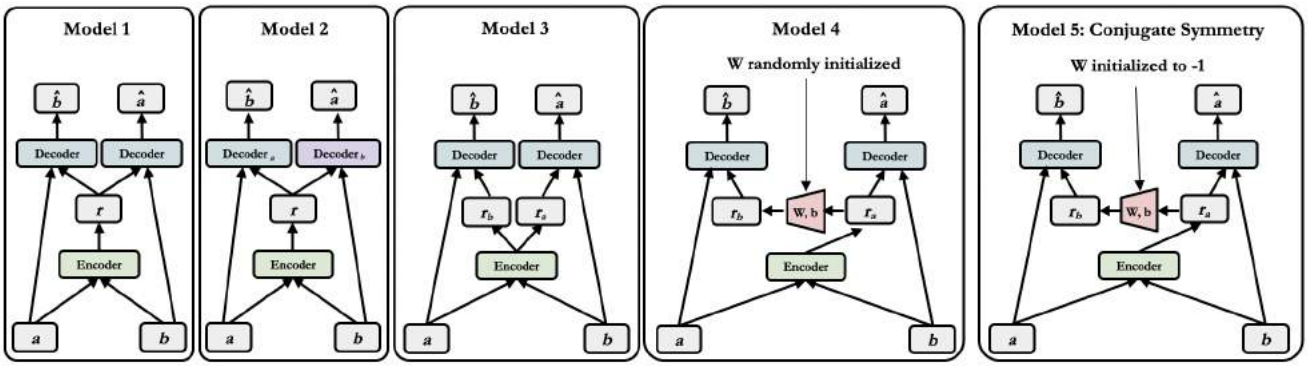


Figure 2: Schematic of model architectures. In each model, the encoder infers the relation r between a and b. The decoder predicts a given b and r, and predicts b given a and r. Gray boxes indicate activity vectors (inputs, outputs, or hidden states). Colored boxes indicate parameterized functions (linear layers or multilayer neural networks); boxes with the same name and color indicate shared parameters.

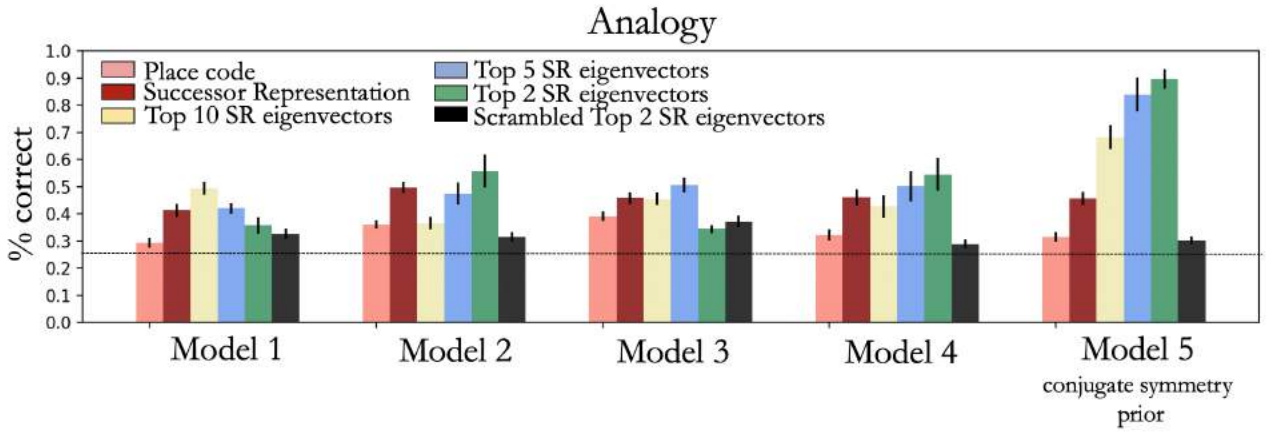


Figure 3: Analogy results for different model architectures. We see that only Model 5 (conjugate symmetry prior) trained on a low dimensional eigenrepresentation (e.g., 2 and 5) reliably learns relations enabling analogical completion.

tion and a particular architecture (See Figure 3.) Only models that integrate a low-dimensional eigenrepresentation of the SR (such as the top 2 or top 5 eigenvectors), and learn over this representation using a conjugate symmetry prior on the relational representation (Model 5) reliably complete analogies involving novel tokens (e.g., 1:3 :: 5: ? (7)). In the 1x80 space, this combination of architecture and top-2 eigenvectors averaged analogical performance of 90.5% across 50 iterations, well above a priori chance levels of 25%. Place codes, the full SR matrix, and scrambled eigenvectors never extracted relational representations that support analogical completion, regardless of architecture. Likewise, Models 1-4 did not reliably extract the appropriate relational structure, regardless of the input representation employed (i.e., even when the input representation was the top 2 or 5 eigenvectors).

Why is this particular combination of input representations and architectural biases important? The highest-eigenvalue eigenvectors carry implicit structural information about the dimension. In this 1D space, the 2nd eigenvector (Fiedler vector) here monotonically increases over the range of the domain, and is thus the closest approximation to a linear representation of the native space in the set of eigenvectors. (See Figure 1). In learning over these representations, Model 5 biases the network to represent the explicit bidirectional relation between a and b using anti-correlated conjugate pairs, reflecting the structure of the components of the relational representation (See Figure 4). This bias appears necessary for backpropagation to reliably learn suitable encoder and decoder parameters to support analogy, when using the transformational objective we employ, here (i.e., $\hat{a} = g_\Phi(R_a, b)$, and $\hat{b} = g_\Phi(R_b, a)$).

Figure 4 shows the latent representations in the two nodes of r for single, randomly chosen runs of all 5 models when trained on the top-2 eigenvectors. For Models 3-5, these

two nodes are functionally restricted, corresponding to r_a and r_b. Notably, we see that in Model 5, the latent activations (r_a, r_b) approximate anti-correlated linear representation of the signed distance. Moreover, for a given signed distance, r_a and r_b are in conjugate symmetric states $(0 + r_a, 0 - r_b)$, reflected about the X axis. Note also that many different A,B pairs will produce the same signed distance. For example, 10-5, 23-18, and 76-71, would all be at the same position along the X axis in Figure 4, given that they are all equal in signed distance. Only Model 5, with the conjugate symmetry prior, extracts similar representations in r, encoding the abstract relational structure as stationary over the range of inputs.

It is worth noting how this relates to the parallelogram model (Rumelhart & Abrahamson, 1973), which captures the abstract logic of linear analogies in vector space. The parallelogram model completes the analogy using fixed vector addition and subtraction operations $D = (B\text{-}A) + C$. We note that we do not see our model as standing in competition with an algorithmic parallelogram computation, a la (Rumelhart & Abrahamson, 1973). Instead, the focus of our model as deriving useful representations from experience, using only observation and standard gradient-based learning, coupled with particular inductive biases. Our model can be thought of as learning an encoding from the input (canonical) space to a latent space in which a linear parallelogram-like computation (B-A+C, (Rumelhart & Abrahamson, 1973; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013)) of the analogy can be performed. Here, we use the conjugate symmetry prior on the chained weight connection to encourage r_a to encode the trajectory from b to a, and r_b to encode the reverse trajectory from a to b.

Generalization across 1D native spaces of various size. Here, we show that the extraction of relational information not only allows the network to complete analogies within the domain over which it has been trained, but also to generalize out of the domain of its support (i.e., pure extrapolation). To do so, we train the network on one range of magnitudes (1x20,1x40, 1x80, or 1x160), and test its analogical performance on magnitudes outside of that range (see Figure 5). For example, one network is trained on inputs that differ in magnitude between 1 and 20 units (1x20 space), and then tested on inputs that differ by different ranges of magnitudes (e.g., 1x40, 1x80, 1x160). We repeat this procedure by training on all four ranges, and testing on the others. Note that these different ranges have the same underlying, local transition structure (See Figure 1), but different numbers of states. In all cases, the network can generate analogies in ranges of magnitudes that are beyond the scope of training without further weight updates, including extrapolating from learning in the 1x20 range and testing in 1x160. Thus, given the assumption of suitable procedures for re-computing and normalizing the eigendecomposition in novel domains, this algorithm can naturally generalize to similarly structured environments without retraining.

Extracting and utilizing tree structured representations. Thus far, we have focused on analogy in 1D linear spaces. However, it is clear that human reasoning also exploits other forms of relational structure present in the environment (e.g., (Kemp & Tenenbaum, 2008)), such as trees, rings, and radial geometries. To explore whether the model presented above can be extended to learn, and make use of other, non-linear structure in generalization and analogical inference, we apply it to a simple hierarchical graph.

Specifically we use a tree composed of two identically structured "families" with connected root nodes (See Figure 6). Both families have 4 generations (levels), with equivalent number of individuals per generation. Edges in the graph exist only between parents and children (results are similar when edges between siblings are included). To generate the transition matrix, we assume that the probability of moving from one node to another is 1/node degree, and compute the eigendecomposition over this random-walk transition matrix. Based on the results of the 1D case, we use the top 2 eigenvectors, and compare the performance of Models 1 (standard) and Model 5 (conjugate symmetry prior), including a version of Model 5 with scrambled eigenvectors.

For this tree, the top 2 eigenvectors are highly structured and interpretable (See Figure 6): the highest eigenvalue eigenvector carries information about family identity, and the second highest-eigenvalue eigenvector about generation, invariant to family. We imposed two further constraints on the analogy test used for the 1D space. First, (a, b) training pairs were constrained to come from the same family. Second, the siblings of the target node were prevented from being foils (though, of course, other close relatives such as cousins, parents, or children could be included as foils). Notably, we again see that Model 5, with a conjugate symmetry prior, learning over the top-2 eigenrepresentations is able to successfully complete the multiple choice analogy tasks, despite no experience with the particular (a, b) pairs, and no direct training on analogy. See Figure 6. Although this particular tree structure is a simplification of more complex structures observed in the world, our results suggest that the model can be applied usefully to more complex structures than the simple linear 1D metric focused on above.

Discussion

Here, we considered the possibility that abstract relation learning can derive from the combination of (a) error-driven learning using the vast amounts of perceptually observational data available ((Rao & Ballard, 1999; O'Reilly, Wyatte, & Rohrlich, 2017), with (b) particular inductive biases in the learning algorithm and network architecture. We compared a variety of input representations and model architectures, testing their ability to extract relational information and use that to complete novel analogies in simple 1D and tree structure domains, without explicit training on analogy. We found that two properties of the models are critical for exhibiting this ability. First, the inputs to the model must be mapped into a canonical representational space that carries implicit struc-

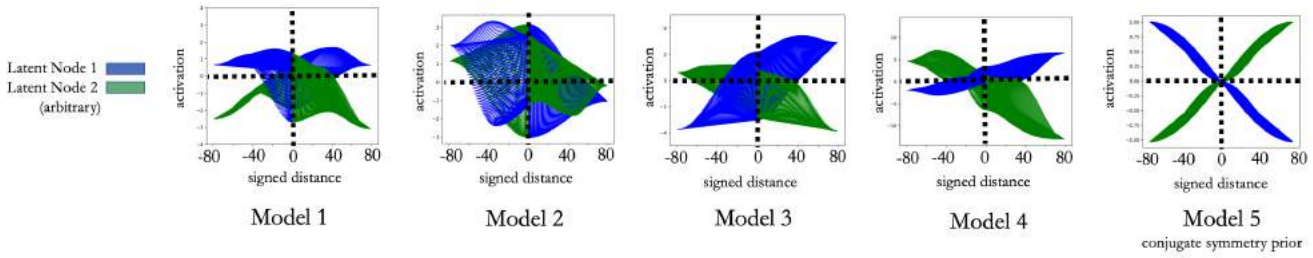


Figure 4: Latent node activations for two nodes as a function of signed distance when trained on top-2 eigenvectors. Only Model 5 (conjugate symmetry prior) learns representations that reliably track the ground truth signed distance. Note that different input pairs produce the same signed distance (e.g., 55-51, 11-7). Every possible pair was plotted here, resulting in the visible, thin individual lines. Notably, only Model 5 reliably produces similar activations in r for different tokens of the same signed distance. Note also that the two latent nodes are anti-correlated.

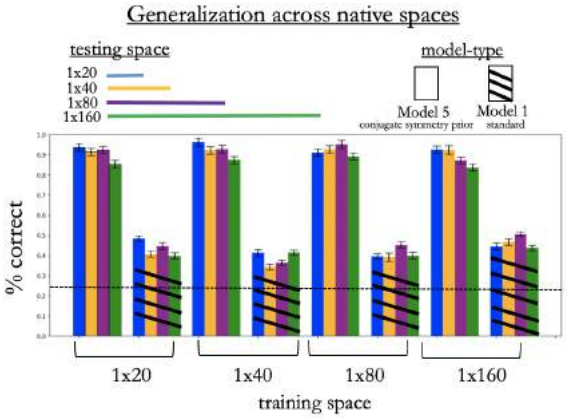


Figure 5: Analogy results when training and testing on native spaces of different size, but the same transition structure.

tural information about the domain (i.e., ordering and neighborhood relations). Learning over these low-dimensional, structured representations (Figure 1) enables the extraction of explicit relational representations, that can be used for analogy. Mapping to a canonical space also allows generalization across environments of different sizes that share structure (Figure 5).

The particular canonical representation we employ is inspired by the relationship between hippocampal place cells and grid cells in medial entorhinal cortex. These grid-cells have been suggested to reflect a decomposition of the predictive map reflecting the transition structure common to spatial environments (Stachenfeld, Botvinick, Gershman, 2017). Evidence for grid-cells has been found in the human brain beyond EC, in medial PFC, posterior cingulate, and lateral inferior parietal cortex (Doeller et al, 2010; Jacobs et al., 2013; Constantinescu et al., 2016), and in conceptual (non-spatial) tasks (Constantinescu et al., 2016). We thus take grid-cells to be one example of the abstract, structured representations that may be shared across domains and exploited for high-level relational reasoning tasks, such as analogy. More broadly, we suggest that similar mechanisms for extracting low-dimensional structure may support other common representational forms. Kemp and Tenenbaum (2008) provide a small inventory of representational forms that recur in human cognition (rings, chains, grids, hierarchies, trees, orders), presenting a hierarchical Bayesian model that identifies the best structural form for a dataset. Understanding how biologically plausible predictive learning mechanisms (coupled with inductive biases) may extract other representational structures is a topic of ongoing work.

Second, we show that if a simple neural network architecture is trained on pairwise relationships among these dimensionally-reduced encodings, and is imbued with a simple, local inductive bias that favors the extraction of conjugate bidirectional relationships among those pairs, it can learn representations that allow it to carry out analogical completions, and generalize this ability well out of the range of its training domain. Intuitively, we can think of the network as asking: having seen a and b, how would I transform a to make it b, and vice versa? We combine this objective with a prior on the relational representation that favors dedicated, but systematically anti-correlated, nodes in the latent space (r_a ,r_b). These nodes may be thought of as a conjugate pair of component relational representations that can be used to reconstruct $a(r_a, b)$ and $b(r_b, a)$. Notably, the learned relations (or trajectories) in the latent space are abstract and approximately stationary with respect to the input domain (i.e., 1-3 = 15-17) (See Figure 4), and can be composed with other object representations to generate analogies.

Though these results are encouraging, they rely on a simplified version of the problem that humans face in generating analogies. One of the most notable features of human analogical ability is the ability to *select* the relevant dimension of variation from the indefinite number of possible relations that might obtain. Indeed, some of the limits in applying the parallelogram procedure (B-A +C) on semantic embeddings as a model of human ability stem from problems handling se-

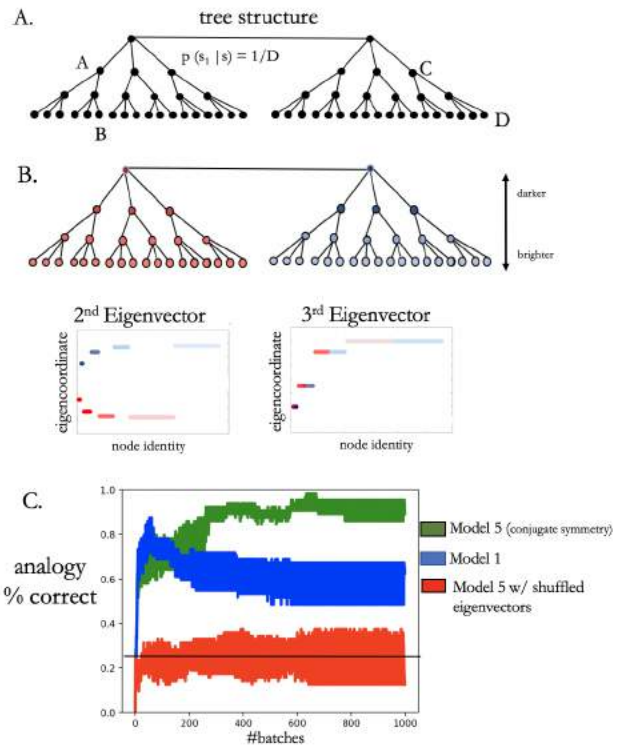


Figure 6: Extraction and utilization of structured representations for the particular tree shown in (A). (B) visualizes the 2nd and 3rd largest eigenvalue eigenvectors. Here, locations in the tree (shown in the upper portion of B) are linked to the points in the eigenplots (lower portion) that share color and brightness. The 2nd EV can be seen to encode family identity (one red, one blue), and the third encodes generation invariant to family (varying in brightness). (C) shows analogy results for this domain.

quencing and context-sensitivity (Chen, Peterson, & Griffiths, 2017). Here, we have pre-selected the relevant dimension for the analogy, focusing instead on general mechanisms for acquiring and utilizing these structured representations. However, a proper understanding of analogy, and human intelligence more broadly, requires directly addressing the relevant search and attentional selection problems, which we see as a critical target for future work.

References

Chen, D., Peterson, J. C., & Griffiths, T. L. (2017). Evaluating vector-space models of analogy. *arXiv preprint arXiv:1705.04416*.

Constantinescu, A. O., O'Reilly, J. X., & Behrens, T. E. (2016). Organizing conceptual knowledge in humans with a gridlike code. *Science*, *352*(6292), 1464–1468.

Dayan, P. (1993). Improving generalization for temporal difference learning: The successor representation. *Neural Computation*, *5*(4), 613–624.

Doeller, C. F., Barry, C., & Burgess, N. (2010). Evidence for grid cells in a human memory network. *Nature*, *463*(7281), 657.

Doumas, L. A., Hummel, J. E., & Sandhofer, C. M. (2008). A theory of the discovery and predication of relational concepts. *Psychological review*, *115*(1), 1.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive science*, *7*(2), 155–170.

Gentner, D., & Forbus, K. D. (2011). Computational models of analogy. *Wiley interdisciplinary reviews: cognitive science*, *2*(3), 266–276.

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive psychology*, *15*(1), 1–38.

Gustafson, N. J., & Daw, N. D. (2011). Grid cells, place cells, and geodesic generalization for spatial reinforcement learning. *PLoS computational biology*, *7*(10), e1002235.

Hafting, T., Fyhn, M., Molden, S., Moser, M.-B., & Moser, E. I. (2005). Microstructure of a spatial map in the entorhinal cortex. *Nature*, *436*(7052), 801.

Holyoak, K. J. (2012). Analogy and relational reasoning. *The Oxford handbook of thinking and reasoning*, 234–259.

Jacobs, J., Weidemann, C. T., Miller, J. F., Solway, A., Burke, J. F., Wei, X.-X., ... others (2013). Direct recordings of grid-like neuronal activity in human spatial navigation. *Nature neuroscience*, *16*(9), 1188.

Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).

O'Reilly, R. C., Wyatte, D. R., & Rohrlich, J. (2017). Deep predictive learning: A comprehensive model of three visual streams. *arXiv preprint arXiv:1709.04654*.

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, *2*(1), 79.

Rumelhart, D. E., & Abrahamson, A. A. (1973). A model for analogical reasoning. *Cognitive Psychology*, *5*(1), 1–28.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1985). *Learning internal representations by error propagation* (Tech. Rep.). California Univ San Diego La Jolla Inst for Cognitive Science.

Stachenfeld, K. L., Botvinick, M. M., & Gershman, S. J. (2017). The hippocampus as a predictive map. *Nature neuroscience*, *20*(11), 1643.

Acknowledgements

This project / publication was made possible through the support of a grant from the John Templeton Foundation. The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the John Templeton Foundation.

Learning Cross-linguistic Word Classes through Developmental Distributional Analysis

Daniel Freudenthal, Fernand Gobet, Julian M. Pine
Department of Psychological Sciences, University of Liverpool

Abstract

In this paper, we examine the success of developmental distributional analysis in English, German and Dutch. We embed the mechanism for distributional analysis within an existing model of language acquisition (MOSAIC) that encodes increasingly long utterances, and compare results against a measure of 'noun richness' in child speech. We show that, cross-linguistically, the mechanism's success in building an early noun class is inversely related to the complexity of the determiner and noun gender system, and that merging of determiners gives very similar results across languages. These results suggest that children may represent grammatical categories at multiple levels of abstraction that reflect both the larger category as well as its finer structure.

Keywords: language acquisition; cross-linguistic; distributional analysis.

Introduction

A major question in the study of language acquisition is how children acquire grammatical categories such as noun and verb. One source of information that children might draw on in this process is distributional information – nouns and verbs tend to occur in different lexical contexts (i.e. are preceded and followed by different sets of words). An influential approach to the learning of word classes through distributional analysis is that of Redington, Chater and Finch (1998) who show that it is possible to accurately cluster words into syntactic categories on the basis of the distribution of a small set of high frequency words that precede and follow them. The basic ideas behind distributional analysis have been employed and adapted by several other authors and, in some cases, applied to other languages (Frank et al., 2013; Mintz, 2003; Keibel, 2005).

However, a major weakness of many studies of distributional analysis is that, while they explore mechanisms that are thought to operate in language-learning children, they make limited contact with the developmental literature and child data. Thus, the focus tends to be on building large word classes with high accuracy. As a result, distributional analysis is often carried out on large corpora of complete utterances and hence ignores the developmental fact that most of children's early utterances are just one or two words long.

Freudenthal et al. (2016a, b) aimed to develop a more plausible mechanism by 1. gradually expanding the contexts available to the mechanism in a developmentally plausible way, and 2. simulating actual child data. Freudenthal et al. do this in the context of MOSAIC (Freudenthal et al. 2007, 2015), a computational model that has been used to simulate a range of phenomena in language acquisition. The key learning constraint in MOSAIC is an utterance-final bias: MOSAIC builds up the representation of the input it is trained on in a right-to-left manner. This feature interacts with the statistics and structure of the input language and is responsible for MOSAIC's successful simulation of (amongst others) cross-linguistic differences in the rates at which children produce Optional Infinitive errors.

As MOSAIC sees more input, it represents longer (utterance-final) phrases and thus has more contexts available for distributional analysis. Freudenthal et al. show that, in English, their developmental version of distributional analysis initially tends to link together nouns (which tend to occur in utterance-final position), a finding that is consistent with the claim that children acquiring English form a productive noun category earlier than they form a productive verb category (Akhtar & Tomasello, 1997; Olguin & Tomasello, 1999; Tomasello & Olguin, 1993).

Freudenthal et al. (2016b) also show that MOSAIC builds an initial noun class that is sufficiently large to simulate the rate of noun use in early child speech. Introducing a measure of *noun richness* – the ratio of the number of nouns over the number of nouns plus main verbs – they show that this ratio is considerably higher in early child speech than in child-directed speech. Simulations with MOSAIC show that roughly half of this difference can be explained through high noun richness in the utterance-final phrases in the model's output. Productive use (i.e. substitution) of distributionally similar words was sufficient to raise noun richness in MOSAIC's output to levels near those found in English-speaking children.

Taken together, these results show that it is possible to perform a developmentally plausible distributional analysis and use it to simulate actual child data, and thus greatly enhance the psychological plausibility of the approach. However, Freudenthal et al. (2016b) only apply their mechanism to English, a language that has a relatively fixed word order and is morphologically impoverished, two features that are likely to benefit distributional analysis.

The main aim of this paper is to extend this developmental distributional analysis to German and Dutch, two languages that have more variable word order, and are morphologically more complex (in particular, through their use of gender and case). Our main focus will be on how comparable the results of distributional analysis are, and how well they fit child noun richness scores in the three languages. In particular, we will focus on the complexity of the determiner and noun gender system. Incorporating the analyses within a computational model that learns progressively longer sequences also allows us to gradually expand the contexts available for

distributional analysis and investigate how this interacts with the word orders of the three languages.

Typology of German and Dutch

German, Dutch and English differ in a number of ways that are relevant for the current analyses. Typologically, the main difference is that English is an SVO language, while German and Dutch are SOV/V2 languages where verb position is dependent on finiteness – finite forms take second position (see utterances 1a, 1b and 1c) whilst nonfinite forms take final position (see utterances 2a, 2b and 2c).

1a. I eat a cookie (E)
1b. Ich esse ein Keks (G - I eat a cookie)
1c. Ik eet een koekje (D - I eat a cookie)

2a. I want to eat a cookie.
2b. Ich moechte ein Keks essen (G - I want a cookie eat)
2c. Ik wil een koekje eten (D - I want a cookie eat)

3a. Do you want a cookie?
3b. Willst du ein Keks? (G – Want you a cookie)
3c. Wil je een koekje? (D – Want you a cookie)

English and German/Dutch also differ in terms of question formation (see utterances 3a, 3b and 3c). Where English forms (polar) interrogatives through the use of dummy modal *do*, German and Dutch use (main) verb inversion. These features mean that German and Dutch have a more variable word order, which may impact on the general success of distributional analysis. The verb-final feature may result in lower numbers of nouns occurring in utterance-final position. This in turn may affect the early construction of a noun class through distributional analysis. However, it also raises the possibility that German and Dutch children may show lower levels of noun richness than English children. A similar claim has been made for children learning languages such as Mandarin Chinese and Korean (Choi & Gopnik, 1995)

Table 1: Case marking in German

	Nom.	Gen.	Dat.	Acc.
Masc.	ein/der	eines/des	einem/dem	einen/den
Fem.	eine/die	einer/der	einer/der	eine/die
Neut.	ein/das	eines/des	einem/dem	ein/das
Plural	--/die	--/der	--/den	--/die

A second way in which the three languages differ is in their use of noun gender and case. English has neither gender nor case (except on personal pronouns). German has three genders and marks case on articles and adjectives. Dutch is like German in that it has gender, but is like English in that it does not mark case. Table 1 illustrates the Gender/Case system of German, for the definite and indefinite article. German gender extends to demonstratives, possessives and quantifiers.

Standard Dutch distinguishes two genders[1] (common and neuter), which take the same indefinite article (*een*), but differ in the definite article (*de*/*het*). Gender is marked on adjectives by the addition/omission of an *-e* suffix. This suffix is applied to all adjectives preceding common gender nouns. For neuter nouns it is applied to adjectives following the definite, but not the indefinite article. Dutch gender extends to demonstratives (but not possessives and quantifiers).

One of the consequences of the different case and gender systems of the three languages is that the degree of lexical variation in the position preceding nouns is largest for German, intermediate for Dutch and lowest for English, Construction of a noun category through distributional analysis is therefore likely to be least constrained in English and most constrained in German. However, while gender may hinder the learning of a noun class, it marks a distinction that children need to acquire, and, since it has very little (transparent) semantic or phonological basis[2], it is very likely to be one that has to be learned distributionally. We will examine how the complexity of the determiner system affects the learning of both the overall noun class as well as the finer gender classes. We will first perform a distributional analysis whilst differentiating between all determiners, and then compare the results with an analysis in which we conflate case and gender by merging the different forms of determiners. Keibel (2005) has previously shown that merging determiners in this way is beneficial for learning the German noun category.

Corpora used

A challenge in cross-linguistic research involving corpora of child-directed speech (CDS) is that of ensuring comparability. The number of corpora available is limited and they differ in terms of size, recording situations, age range of the target children and availability of morphological information. We aimed to select from CHILDES a set of corpora for each language that were as comparable as possible in terms of their overall size. For English we selected the 6 largest sub-corpora from the Manchester corpus (Theakston et al., 2001). The Manchester corpus contains corpora for 12 individual children, and contains part-of-speech information for child and adult speech on the morphology (MOR:) tier. The selected corpora typically contained 30,000-35,000 utterances of child-directed speech

[1] Though vestiges of a third gender remain.

[2] There actually are a number of phonological, morphological, and semantic cues to German gender. MacWhinney et al. (1989) show that a neural network trained on 38 of these cues can correctly classify held out nouns. However, since Macwhinney et al.'s model learns in a supervised manner, gender information is actually available to the model. Since gender is essentially defined distributionally, lexical contexts appear a more potent cue to identifying a noun's gender.

per child. For German we selected the Rigol corpus, consisting of 4 children with roughly 45,000 child-directed utterances per child. After limited cleaning up of the corpus, we were able to run the CLAN mor facility, which was able to assign part-of-speech information to ~99% of all word tokens in the corpus. For Dutch, we selected the two children from the Van Kampen corpus. These corpora contain 65,000 and 25,000 maternal utterances. Since there is currently no functioning mor-grammar for Dutch, we assigned to the words in these corpora the most common part of speech derived from the Treetagger (Schmid, 1994).

Study 1: Child Noun Richness

The first analysis concerned children's cross-linguistic use of nouns and (main) verbs[3]. All the corpora used consist of multiple recordings (tapes) at different child ages. For all corpora we counted the number of nouns and verbs in child and adult speech on a tape-by-tape basis, and plotted noun richness (i. e. #nouns / (#nouns + #verbs)) relative to the child's Mean Length of Utterance (MLU) for the relevant tape. In line with current practice in MOSAIC, analysis was performed on utterance types. Figure 1 shows the trendlines for the scatterplot of English, Dutch and German child, and child-directed speech. For clarity, individual data points are not plotted. As can be seen, noun richness scores look remarkably similar across the three languages. While German child noun richness is (initially) lower than it is for Dutch and English, it is considerably higher than it is for adults, and thus suggests that, cross-linguistically, children are equally productive around nouns in the early stages.

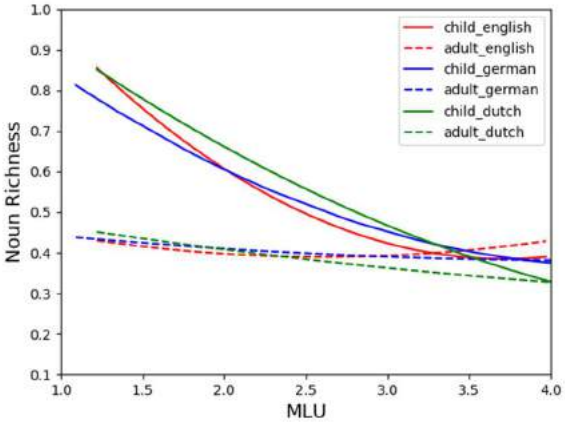


Fig. 1: Noun Richness in English, German and Dutch.

Study 2: Simulations with MOSAIC

Training MOSAIC models

MOSAIC learns from orthographically-transcribed child-directed speech and generates as output corpora of speech that can be directly compared to child speech. Learning in MOSAIC is slow, and takes place by feeding input through the model multiple times. With each exposure, MOSAIC represents more and longer (utterance-final) phrases and is thus capable of producing more and longer output, as is true of developing children. A detailed description of MOSAIC and how it is trained is provided in Freudenthal et al. (2015).

For the current analyses, we performed a distributional analysis at several points in the models' development. Here, we report results from selected runs between 36 and 50.[4] Over this range, the MLU of the utterances represented in MOSAIC (and hence of its output corpus) increases from roughly 2 to 5 words – and thus increasingly approximates corpus-wide statistics. The key consideration here is that, early in training, MOSAIC represents short utterance-final phrases that extend further to the left with increased training. This feature, which is responsible for MOSAIC's successful simulation of a number of phenomena in child speech, has the potential to interact with word order in shaping the cross-linguistic results of the distributional analysis.

The distributional analysis

The distributional analysis was carried out in the same manner for all languages. The target words were the 1,000 most frequent words for a given corpus, and the context words the 150 most frequent words. Utterance endings were also included as contextual elements. At each point in training, we searched the phrases represented in MOSAIC for the target words, and noted how often the context words occurred in the preceding and following position. Thus, for each target word, we generated two vectors that contained the counts for the context words in preceding and following position. Similarity between words was expressed as the similarity between these vectors, and two words were considered to be of the same class if their similarity exceeded a threshold value for both preceding and following position. For the current analyses we expressed similarity in both a non-parametric and a parametric way. We used a Spearman rank-order correlation, as well as cosine similarity based on the square root of the vector counts[5]. Freudenthal et al. (2013) have shown that (for English) a parametric measure is better for classifying nouns, while the rank order is better for classifying verbs. In the current analyses, the rank-order correlation gives better results when applied to English, while the parametric gives better results for German. This finding is in line with reports by Redington et al. (1998) for English and Keibel (2005) for German. Importantly, however, the two

[3] For Dutch and German, modal verbs were included as these can be used as main verbs. Copulas were excluded for all languages.

[4] Run number reflects number of exposures to the input.

[5] This is a departure from Freudenthal et al. (2016b) who used a distance measure that discarded frequency as well as counts from interrogative contexts.

measures give qualitatively similar results when used in isolation, but better quantitative results when combined.

Results

Unmerged determiners

Results for the distributional analysis are reported in Table 2, which shows the number of linked words, overall accuracy (proportion of same class links), noun richness (ratio of noun-noun to noun-noun plus verb-verb links) as well as numbers of links and accuracy for verbs and nouns. Two words were considered to be of the same word class if their rank-order correlation in preceding and following position exceeded 0.40, or their cosine similarity exceeded 0.65.

As can be seen in Table 2, the distributional analysis in English results in an early noun class, with verbs being classified later in development. This pattern is consistent with children showing early productivity around nouns and late emergence of a productive verb class (Akhtar & Tomasello, 1997; Olguin & Tomasello, 1993; Tomasello & Olguin, 1993).

It is also apparent from Table 2 that the mechanism is capable of classifying words with high accuracy, particularly for nouns, but also for verbs (in the later stages). Table 2 also shows that results for the Dutch distributional analysis are similar to those for English, though the mechanism is less successful in linking nouns, and is less accurate overall. German results mirror those from Dutch, but the distributional analysis is even less successful at building a noun class. Thus, the models never exceed 1000 noun links, even in the later stages. Across runs, the German noun class is approximately a quarter of the size of the English noun class.

The results from Table 2 thus suggest that the less constrained word order in Dutch and German leads to lower overall accuracy, but also that the size of the noun class is inversely related to the complexity of the determiner system. This pattern is not surprising, but it appears to be in conflict with the child noun-richness data from Fig. 1, which suggest that children from all three languages are equally productive around nouns. It also suggests that German and (to a lesser extent) Dutch MOSAIC models may struggle to simulate early child noun richness scores[6].

Merged determiners

We examined whether German and Dutch gender and case hamper the construction of a noun category by merging determiners into one lexical item, and adding their respective counts. For German, this meant that all 6 forms of the definite article were merged, as were all 6 forms of the indefinite article (thus maintaining the distinction between the definite and indefinite article). For Dutch, we merged both forms of the definite article. Since there is only one form of the indefinite article, this cannot be merged. Results for the distributional analysis with merged determiners are shown in Table 3.

Table 2: Results of Distributional Analysis for English, Dutch and German

Run	Links	Overall accuracy	Noun-richness	Nouns	Verbs	Noun-accuracy	Verb-accuracy
English							
36	1,641	0.80	0.94	1,218	70	0.83	0.42
38	2,215	0.80	0.91	1,553	153	0.83	0.52
40	3,037	0.83	0.89	2,230	237	0.85	0.63
44	4,144	0.90	0.86	3,164	437	0.91	0.81
50	4,576	0.91	0.83	3,375	615	0.92	0.87
Dutch							
36	1,140	0.73	0.95	774	34	0.77	0.23
38	2,030	0.78	0.96	1,467	62	0.80	0.38
40	2,995	0.81	0.96	2,260	90	0.82	0.43
44	3,496	0.85	0.91	2,582	256	0.85	0.75
50	3,310	0.84	0.80	2,122	502	0.84	0.86
German							
36	841	0.52	0.93	282	20	0.54	0.27
38	935	0.61	0.89	383	43	0.64	0.47
40	1,227	0.71	0.87	581	86	0.71	0.61
44	1,985	0.78	0.78	905	253	0.80	0.84
50	2,563	0.79	0.52	754	697	0.83	0.89

[6] Note, though, that since (unlike Freudenthal et al., 2016b) we do not currently generate output from MOSAIC, we cannot directly relate the size of the noun class to child noun richness scores.

Table 3: Results of Distributional Analysis with merged determiners for Dutch and German.

Run	Links	Overall accuracy	Noun-richness	Nouns	Verbs	Noun-accuracy	Verb-accuracy
Dutch							
36	1,515	0.70	0.96	997	37	0.73	0.17
38	2,749	0.76	0.96	1,955	70	0.78	0.30
40	4,140	0.80	0.96	3,134	104	0.81	0.36
44	5,151	0.84	0.93	3,940	292	0.86	0.65
50	4,788	0.84	0.84	3,290	573	0.85	0.80
German							
36	2,091	0.49	0.97	836	27	0.51	0.15
38	2,399	0.56	0.95	1,095	58	0.57	0.26
40	3,543	0.65	0.94	1,914	131	0.65	0.40
44	5,992	0.73	0.91	3,540	364	0.74	0.73
50	6,287	0.76	0.80	3,226	816	0.77	0.84

It is evident from Table 3 that the merging of determiners results in an increase in the number of nouns that get linked for both languages, but that this increase is considerably larger for German (by a factor of 4) than it is for Dutch (by a factor of 0.4). It is also obvious that the overall results for Dutch and German are now quite similar to the results of the English analysis (though overall accuracy scores are still lower for Dutch and German), and more in line with the cross-linguistic child noun richness scores (see Fig. 1).

Taken together, these results suggest that gender and case are detrimental to learning a noun category through distributional analysis. However, if children are able to ignore the identity of determiners, distributional analysis yields remarkably similar results across the three languages, despite their differences in word order.

Learning gender subclasses

The fact that gender (and case) hamper the learning of a noun category is not surprising since gender divides the noun category into a number of subcategories that differ in their distributional characteristics. A relevant question therefore is to what extent maintaining the distinction between the different determiners allows the mechanism to distinguish (and hence children to learn) the different noun genders. This was investigated by taking the noun-noun links from Table 2, and determining to what extent these involved nouns from the same and different genders. Results (confusion matrices) from run 50 are shown in Tables 4 (German) and 5 (Dutch).

Comparison of Tables 4 and 5 reveals that the distributional analysis is remarkably good at distinguishing the German gender subcategories, at least for the singular genders. At one level, this is not surprising since merging the determiners increases the size of the German noun class four-fold. However, inspection of the actual forms of the German determiners (see Table 1) shows that 6 different forms of each determiner are used in a paradigm containing 16 cells. Most determiners therefore occur with nouns of different genders, suggesting that the German genders are quite confusable.

Table 4: German Gender Confusion Matrix (run 50)

	Masc.	Fem.	Neut.	Pl.
Masc.	216	15	39	5
Fem.	15	198	0	18
Neut.	39	0	203	2
Pl.	5	18	2	25

Table 5: Dutch Gender Confusion Matrix (run 50)

	Common	Neuter	Plural
Common	1415	187	102
Neuter	187	249	10
Plural	102	10	17

Table 4 shows that the distributional analysis is far less successful in Dutch, with many neuter and plural nouns being linked to common gender nouns. This is caused by the fact that Dutch gender is marked on the definite, but not on the indefinite article. The Dutch noun genders are thus distributionally more similar, and far more confusable than the German noun genders. Since there are few cues to grammatical gender other than distributional information, these results suggest that acquisition of gender may be more challenging for Dutch- than for German-learning children.

Conclusions

The main conclusion to be drawn from the analyses reported here is that they provide strong support for the viability of distributional analysis. Thus, we show that it is possible to

obtain plausible (and very similar) results across three different languages that differ in their word order as well as the detail of their gender and case system. Importantly, we do so using a fixed set of parameters, and in the context of a computational model that gradually expands the contexts available to the mechanism – allowing us to investigate how the increasing length of utterances that children represent may affect their word class learning. Moreover, by comparing the results to actual child data (noun richness), we were able to evaluate the relative size of the (early) noun class across the three languages.

However, it is also clear that the successful construction of a noun category depends critically on the complexity of the determiner system, and hence on how determiners are treated. If the identity of the German determiner is maintained, distributional analysis results in a noun class that is very small compared to Dutch and English, but that distinguishes between the different genders quite successfully. Merging the determiners brings the size of the verb class more in line with English and Dutch, but necessarily conflates the different genders. This effect is less pronounced in Dutch. However, the finer-grained structure of Dutch gender is distributionally less well-defined, and thus suggests that it may be more difficult to acquire for language-learning children.

The German (and Dutch) results thus suggest that grammatical categories need to be represented at different levels of abstraction that reflect both their more general properties as well as their finer-grained structure. The suggestion that children may represent both 'merged' and 'unmerged' determiners may seem surprising since one of the key characteristics of children's early speech is the fact that it lacks closed-class items like determiners. However, there is actually considerable evidence that children represent more of the closed class items than they produce.

Lew-Williams and Fernald (2007) show that Spanish three-year-olds in a looking-while-listening task can use the identity of (gendered) determiners to orient towards a target of the relevant gender. Similar findings have been reported for 24-month-old children in French (van Heugten & Shi, 2007), a language where, like Spanish, the determiner is fully predictive of the gender of the noun. Interestingly, children learning Dutch appear delayed relative to French children in this task (van Heugten & Johnson, 2011), thus providing support for the notion that the relatively poor separation of Dutch gender found in the current analyses may make it particularly hard to acquire. Studies on German (Hőhle et al., 2004) also show that children as young as 16 months (but not 12 months), can distinguish between novel words used in a nominal vs. verbal context after being habituated with a determiner-novel word sequence, but not after a pronoun-novel word sequence. These results suggest that children can use determiners to classify nouns from a very young age, but equally that they can use gender information in the on-line processing of speech, at least in languages where determiners reliably predict gender.

Taken together, the results also highlight the strengths of our approach. By embedding distributional analysis within an existing model of language acquisition that simulates children's increasing MLU, applying it to three different languages, and comparing it to actual child data, we were able to investigate how word order and the complexity of the determiner system affect the formation of an early noun class, as well as the potential implications this has for children's representations of closed class items.

Acknowledgements

Daniel Freudenthal, Julian Pine, and Fernand Gobet are members of the International Centre for Language and Communicative Development (LuCiD) at the University of Liverpool, for which the support of the Economic and Social Research Council [ES/L008955/1] is gratefully acknowledged.

References

Akhtar, N., & Tomasello, M. (1997). Young children's productivity with word order and verb morphology. *Developmental Psychology, 33*, 952-965.

Choi, S. & Gopnik, A. (1995). Early acquisition of verbs in Korean, a cross-linguistic study. *Journal of Child Language*, *22*, 497-529.

Frank, S., Goldwater, S. & Keller, F. (2013). Adding sentence types to a model of syntactic category acquisition. *TopiCS in Cognitive Science* 5 (3), pp. 495–521.

Freudenthal, D., Pine, J. M., Aguado-Orea, J. & Gobet, F. (2007). Modelling the developmental patterning of finiteness marking in English, Dutch, German and Spanish using MOSAIC. *Cognitive Science*, *31*, 311-341.

Freudenthal, D., Pine, J.M., Jones, G. & Gobet, F. (2013): Frequent frames, flexible frames and the noun-verb asymmetry. In: M. Knauf, M. Pauen, N. Sebanz E I. Wachsmuth (Eds.), *Proceedings of the 35th annual meeting of the Cognitive Science Society*. (pp. 2327-2332). Austin, TX: Cognitive Science Society.

Freudenthal, D., Pine, J.M., Jones, G. & Gobet. F. (2015). Simulating the cross-linguistic pattern of Optional Infinitive errors in children's declaratives and Wh-questions. *Cognition*, *143*, 61-76.

Freudenthal, D., Pine, J. M., Jones, G., & Gobet, F. (2016a). Developmentally plausible learning of word categories from distributional statistics. In 38th *Annual Conference of the Cognitive Science Society* (pp. 674-679). Philadelphia.

Freudenthal, D., Pine, J. M., Jones, G., & Gobet, F. (2016b). Simulating Developmental Changes in Noun Richness through Performance-limited Distributional Analysis. In *38th Annual Conference of the Cognitive Science* Society (pp. 602-607). Philadelphia.

Hőhle, B., Weissenborn, J. Kiefer, D., Schultz, A. & Schmitz, M. (2004). Functional elements in infants' speech processing: The role of determiners in the syntactic categorization of lexical elements. *Infancy,* 5, 341-353.

Keibel, J.H. (2005). Distributional patterns in German child-directed speech and their usefulness for acquiring lexical

categories – A case study. Unpublished Doctoral Dissertation. Freiburg, Germany.
Lew-Williams, C. & Fernald, A. (2007). Young children learning Spanish make rapid use of grammatical Gender in spoken word recognition. *Psychological Science*, *18*, 193-198.
MacWhinney, B. (2000). *The CHILDES project: Tools for analysing talk (3rd Edition)*. Mahwah, NJ: Erlbaum.
MacWhinney, B, Leinbach, J., Taraban, R., & McDonald, J. (1989). Language Learning: Cues or Rules? *Journal of Memory and Language*, *28*m 255-277.
Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition, 90,* 91-117.
Olguin, R., & Tomasello, M. (1993). Twenty-five-month-old children do not have a grammatical category of verb. *Cognitive Development*, 8, 245-272.
Redington, M., Chater, N. & Finch, S. (1998). Distributional Information: A powerful cue for acquiring syntactic structures. *Cognitive Science*, 22, 425-469.
Schmid, H. (1995). Improvements in Part-of-Speech Tagging with an Application to German. *Proceedings of the ACL SIGDAT-Workshop*. Dublin, Ireland.
Theakston, A. L., Lieven, E. V. M., Pine, J. M. & Rowland, C. F. (2001). The role of performance limitations in the acquisition of Verb-Argument structure: An alternative account. Journal of Child Language, 28, 127-152.
Tomasello, M., & Olguin, R. (1993). Twenty-three-month-old children have a grammatical category of noun. *Cognitive Development*, *8*, 451-464.
van Heugten, M. & Johnson, E.K. (2011). Gender-marked determiners help Dutch learners' word recognition when gender information itself does not. *Journal of Child Language*, *38*, 87-100.
van Heugten, M. & Shi, R. (2009). French-learning toddlers use Gender information on determiners during word recognition. *Developmental Science*, *12*, 419-425.
van Kampen, J. (2009). The non-biological evolution of grammar: Wh-question formation in Germanic. *Biolinguistics*, 2-3, 154-185.

The Stream of Spatial Information: Spanning the Space of Spatial Relational Models

Paulina Friemann (friemanp@cs.uni-freiburg.de)
Cognitive Computation Lab, Georges-Köhler-Allee, University of Freiburg, 79110 Freiburg, Germany

Jelica Nejasmic (jelica.nejasmic@ph-ludwigsburg.de)
PH Ludwigsburg, 71634 Ludwigsburg, Germany

Marco Ragni (ragni@cs.uni-freiburg.de)
Cognitive Computation Lab, Georges-Köhler-Allee, University of Freiburg, 79110 Freiburg, Germany

Abstract

Given identical informational content, the order in which you receive spatial information may heavily influence the correctness of your mental representation. This can reveal important insights into the specifics of human spatial cognition and the way we integrate information. Despite its importance in everyday life, its causes and the mental processes involved still remain an open question. Most cognitive models so far have focused on modeling only answer distributions or just the most frequent answer given by all participants.

In this paper we take a rather radical approach: We turn to the individual spatial reasoner and focus our analyses on the stream of spatial information and related reaction times, i.e., how the spatial information is represented and cognitively processed. By spanning a space of 243 cognitive spatial models, some of which outperform the current state-of-the art models, it is possible to test the goodness of general principles underlying such models.

Keywords: Spatial Cognition; Reasoning; Continuity Effect; Cognitive Models

Introduction

Imagine that you are new to a city. It is a common experience that it is not very likely that you will have all spatial information available at the same time. Rather, you will receive it piece by piece. However, the way of how we receive spatial information can impact our mental representation, the time to understand the information, and possible conclusions we draw. But how do we process information mentally that we receive? How do we possibly integrate the spatial information into a mental representation? How difficult is it to process the information? What can existing cognitive approaches and computational models contribute?

Spatial relational information can be formulated by two objects and a relation: the first object is the object *to be located*, the relation gives information about how the objects are spatially connected, and the second object which is termed the *reference* object. Consider the following:

(1) The post office is to the left of the train station.
The train station is to the left of the main street.
The main street is to the left of the park.

Can you easily build a mental representation integrating this information at the same time? You should have no difficulty at all! Even receiving this information step-by-step, each new information nicely integrates with the most recent information. Such a problem is called a *continuous description*. Consider now the following description:

(2) The train station is to the left of the main street.
The park is to the right of the main street.
The post office is to the left of the train station.

This time, it might have taken more time and a bit more difficult to build a mental representation from the given assertion. While the information content was identical to before, the information could not be so easily integrated as in problem (1). This was mainly due to the last assertion that related the post office to the train station. Such problems are coined *semi-continuous*. Consider now this last description:

(3) The post office is to the left of the train station.
The park is to the right of the main street.
The train station is to the left of the main street.

Again, if you have received the assertions piece by piece, building an internal representation might have been again more difficult. While again the description has the same information content as all descriptions before, the first and the second assertion were unrelated, requiring to build two unrelated scenarios. Hence, such a problem description is coined *discontinuous*. All three problems allow for constructing an identical arrangement of the objects, namely

post office – train station – main street – park.

Such an arrangement of objects from the assertions is called a *model* of the assertions. These three problems have been investigated by psychologists in the so-called continuity effect (e.g., Ehrlich & Johnson-Laird, 1982; Knauff, Rauh, Schlieder, & Strube, 1998). But, why does the second and especially the third problem appear to be more difficult to be processed by humans? The *stream* of information makes the difference between problem descriptions. In continuous and semi-continuous descriptions, a common middle-term of two successive assertions exists. Since this is not the case in discontinuous orders, these assertions are more difficult to process and may even require to keep two distinct pieces of information in working memory.

Because of the fine-grained nature of this effect, modeling the cognitive processes which underlie it can give insight into how exactly spatial information is processed in the mind. Several cognitive models have been proposed for spatial relational reasoning, among which an implementation in a cognitive architecture (Ragni, Fangmeier, & Brüssow, 2010),

a model for reasoning with intervals (Schlieder & Berendt, 1998), and a stand-alone cognitive architecture (Schultheis & Barkowsky, 2011) (for a recent overview see Friemann & Ragni, 2018). To account for the continuity effect, a cognitive model needs to describe the nature of constructing a spatial model in great detail. This includes the introduction of a measure of difficulty, the *mental cost*, of a specific mental operation to account for the increase in reading times and the drop in accuracy. The cognitive models which satisfy these requirements are the *spatial reasoning as verbal reasoning* model (Krumnack, Bucher, Nejasmic, & Knauff, 2010) and *preferred inferences in reasoning with spatial mental models* (PRISM, Ragni & Knauff, 2013), which we now introduce.

Cognitive Theories, Models, & Complexity

Verbal Reasoning Model (Krumnack et al., 2010). The core assumption underlying the Verbal Model is that deduction processes does not necessarily require deduction-specific mechanisms to operate on internal representations. Instead, a simple order of object terms and some verbal cognitive mechanisms might guide the reasoning process. Following Polk and Newell (1995), cognitive processes in deductive reasoning might be based upon the same processes as language comprehension and generation. The model satisfies the criteria of verbal reasoning as outlined by Polk and Newell (1995). Verbal in that sense refers to transforming between verbal and semantic representations, that is constructing the queue (encoding) and "reading out" information that is not explicitly provided by verbal descriptions. It is assumed that reasoning is accomplished by applying well-trained linguistic processes. The approach does not obviate specific mechanisms but provides a more parsimonious explanation on how inferences can be drawn from given information without assuming additional mechanisms.

The computational model assumes the mental spatial structure to resemble a queue. In the same vein, each mental model has an *implicit direction*. This direction depends on the relation in the first premise and is contrary to the explicit direction in this relation. This can be understood as simulating an expectation on where the next object is about to appear, which can be easily understood by considering Table 3. For example, if the first premise was "The mango is to the left of the pear", the implicit direction would be to the right:

On the other hand, if the first premise was "The pear is to the right of the mango", the implicit direction is to the left:

PRISM (Ragni & Knauff, 2013). PRISM is an implementation of the theory of preferred mental models. The model simulates and explains how preferred models are constructed, inspected to find a putative conclusion, and then varied to find possible counter-examples. A spatial working memory structure is operationalized by a spatial array. A spatial focus inserts tokens into the array, inspects the array to find new spatial relations, and relocates tokens in the array to generate alternative models of the problem description, if necessary. The focus also introduces a general measure of difficulty based on the number of necessary focus operations (rather than the number of models).

Mental Costs and Complexity. The computational model PRISM was the first model to predict reasoning difficulty of spatial problems by assigning unit costs to the focus operations in a spatial working memory, a location where spatial models are built (Ragni & Knauff, 2013). By the numbers of operations PRISM is able to explain among others the continuity effect: as a successive insertion of the terms from left to right, do cost less than switches in the focus direction (semi-continuous case), which costs less than to generate, group, and insert different submodels (discontinuous case). The Verbal Model uses a similar cost measure.

The Order of Information Effect: Data

The order effect for human inferences has been reported in a number of articles (e.g., Ehrlich & Johnson-Laird, 1982; Knauff et al., 1998; Nejasmic, Bucher, & Knauff, 2015) and is explained with the effort to construct a mental representation of the assertions.

Table 1: Order of assertions in Knauff et al. (1998) and Nejasmic et al. (2015). Please note that $\sim$ represents the relation, which is 'left of' in the case of Experiment 1 in Nejasmic et al. (2015) and Knauff et al. (1998), and 'right of' in the case of Nejasmic et al. (2015).

Order	Assertions		
continuous	A $\sim$ B	B $\sim$ C	C $\sim$ D
semi-continuous	B $\sim$ C	C $\sim$ D	A $\sim$ B
discontinuous	C $\sim$ D	A $\sim$ B	B $\sim$ C

Knauff et al. (1998) conducted an experiment, inspired by research of Ehrlich and Johnson-Laird (1982), to test effects on response times and error rates of continuous, semi-continuous, and discontinuous orders of spatial assertions (cp. Table 1) using the relation 'left of'.

The processing times and error rates are summarized in Table 2. While the continuous and semi-continuous order lead to a similar error rate of about 40%, reasoning about discontinuous orders of assertions was more difficult and lead to about 50% errors. Note that the processing time for the third assertion in discontinuous order compared to the other assertions is significantly higher.

Nejasmic et al. (2015) investigated underlying cognitive processes in two experiments using a random presentation of the 72 problems of the three premise orders *continuous*, *semi-continuous* and *discontinuous*. Each premise was presented

Table 2: The four-term-problems in the experiment of Knauff et al. (1998) with reading-times (RT in seconds) and error rates (in percentage correct). Participants were presented with interval relations.

	Assertion			
Order	RT 1	RT 2	RT 3	Error rates
continuous	13.0	11.2	10.9	39.7
semi-continuous	13.6	11.0	11.9	40.1
discontinuous	12.4	13.9	19.5	50.0

sequentially (in a self-paced manner and only one premise visible at a time). The premise described the spatial relation between four small, equal-sized, and disyllabic objects (tools, fruits, or vegetables) for example: "The mango is left of the pear, the pear is to the left of the kiwi, the kiwi is to the left of the apple."

The instruction was to imagine the arrangement described by the premises (in the example: mango – pear – kiwi – apple). Subsequently participants were asked to define the correct arrangement by typing the initial letters of the named objects using the computer keyboard. After the last letter was entered, the trial finished automatically. The next trial started not before the participant hit the "return" key. The program recorded (a) premise reading times (respective time from stimulus onset to key press calling up the next premise), (b) the number of correct responses, and (c) corresponding response times (time from request onset till enter of the last letter).

Experiment 1 and 2 differ mainly in the used relation resulting in different working direction. In Experiment 1 the relation 'left of' was used suggesting a working direction from left to right. In contrast, Experiment 2 used the relation 'right of' resulting in a working direction from right to left. The position of new named objects is leftmost (see Table 3).

Table 3: Example premises and models for a continuous order

	Experiment 1		Experiment 2	
	Premise	Model	Premise	Model
1	A left of B	AB	D right of C	CD
2	B left of C	ABC	C right of B	BCD
3	C left of D	ABCD	B right of A	ABCD

Results from the first experiment are in line with previous findings concerning the continuity effect. Participants need more time to process unrelated information and more errors occur in the discontinuous condition. In the second experiment the continuity effect was presumably counteracted by the working direction. Although processing third premises in the discontinuous condition took the most time, there was an overall and consistent increase of reading times over all conditions. It was expected that reasoners find it more difficult to work in the culturally nonpreferred right-to-left direction, but in the case that the continuity effect results from the integration of two separate models when confronted with discontinuously presented information, the working direction should not matter. So, results support the assumption that one preliminary model is constructed and modified in cases of discontinuity.

Results and Discussion on Aggregated Data

The Kendall rank correlation coefficient τ_b with the mean reaction times for Experiment 1 and 2 of Nejasmic et al. (2015) and the reported data in Knauff et al. (1998) was calculated. We removed all reading times which were outside the 1.5 interquartile range. The results can be found in Table 4.

Table 4: Correlations and significance level for PRISM and the Verbal Model on the aggregated experimental data.

	PRISM		Verbal Model	
	r_{τ_b}	p	r_{τ_b}	p
Nejasmic et al.: Exp 1	.800	.007	0.730	.018
Nejasmic et al.: Exp 2	.033	1	0.225	.501
Knauff et al.	.730	.182	.609	.044

For Experiment 1 from Nejasmic et al. (2015), PRISM had a better correlation than the Verbal Model. The same procedure was done with the data from (Knauff et al., 1998) (Experiment 3 in Table 4), which used the same setting as Experiment 1. PRISM and the Verbal Model correlated significantly with the data.

For Experiment 2 however, the correlations dropped strongly. This indicates that the process to generate a mental model are different from relational descriptions from left to right than building directions from right to left.

As outlined above, many cognitive models have focused on explaining aggregate data. But, how good are these models in predicting each individual reasoner? And, are there other models that can predict individual reasoner better? To further investigate the performance of the models, we turn to the individual reasoners.

To approach this challenge, there are two possibilities: creating cognitive models which are *adaptable to*, or creating cognitive models *designed for* individuals.

The remainder of this paper will investigate the second option. Taking features of models from the literature and insights from psychological experiments, we will span a large space of possible cognitive computational models for spatial relational reasoning.

Generating the Space of Spatial Reasoning Models

To investigate the goodness of the general assumptions, we looked at a whole family of potential models. This approach is driven by the idea that individual participants may not use

the same strategy and their flow of information processing may differ. Hence, rather than constructing a certain model, we identified features in which potential models can differ. These are inspired by proposed cognitive models for spatial relational reasoning in the literature. PRISM, for example, proposes a mental model manipulation device, called focus, which acts just like a foveal area for mental models. The Verbal Model assumes that a spatial mental model has an implicit direction, which can offer an explanation for the better performance in modeling the right-to-left task from Experiment 2 (Krumnack, Bucher, Nejasmic, Nebel, & Knauff, 2011). As for the discontinuous case, the Verbal Model does not offer a solution for the presentation of discontinuous information, as in the connection of two formerly unrelated chunks of information. *PRISM* on the other hand offers a solution in the form of constructing two unrelated mental models, and integrating them group-wise when connecting information is presented.

We chose 8 partly interdependent features to span the space of investigated models, leading to 243 possible cognitive models:

Mental Spatial Structure

The main difference between the Verbal Model and PRISM is the underlying spatial representation structure. PRISM assumes a grid-like structure in the human mind, with a mental focus which inspects one object at a time, can move through the mental representation object by object with an unary cost in each direction, and is persistent throughout the whole task (Ragni & Knauff, 2013). The Verbal Model on the other hand proposes a queue-like structure, meaning that there exists an implicit direction in the mental model, which is dependent on the relation in the first premise (Krumnack et al., 2010). The question whether a mental model has an implicit direction is the focus of the first three main features, leading to the first $2^3 + 1$ possibilities:

Implicit Model Direction Inspired the Verbal Model, models can have a queue-like mental spatial structure with an implicit direction. Moving through this queue in the implicit direction is assumed to be computationally cheap, while moving against this direction is costly. The opposite assumption would be a grid-like mental array similar as is used in PRISM.

Persistency of Direction In the Verbal Model, the implicit direction depends on the relation in the first premise. For the relation 'left of', the direction of the queue would be to the right and vice versa. We added this dependency as a possible feature, as well as the possibility of a reversed dependency, i.e. for the relation 'left of', the direction of the queue would be to the left as well.

Preliminary Integration Following the research in Nejasmic et al. (2015), it seems likely that when reading discontinuous information, such as "a is left of b, c is left of d", reasoners build a preliminary, connected model instead of a second, disjunct model. Therefore, we introduced this idea as another feature for models which assume an implicit direction: Construct a temporary model with the discontinuous information inserted into the mental model in direction of the queue.

Focus

Moving through the mental model is, in PRISM and the Verbal Model, assumed to require some mental operation. Following the terminology in PRISM, we introduce this idea as the so called *focus*, a device which is able to move through the mental model object by object.

If including the focus into the cognitive model, we can further differentiate between different types of foci. For example, while PRISM has a persistent mental focus throughout the whole task, the Verbal Model implicitly introduces a focus-like notion which resets with each premise. The idea is that when a premise contains an object which is already in the queue, the model has to move through the queue from the position of this object. In a sense, this could be described as a focus with the ability to *jump*. The focus feature adds another $2^3 + 1$ possibilities, as a model which assumes a focus can have any of the three mentioned focus features:

Jumping Focus As in the Verbal Model, when reading a premise, the focus can jump to the addressed object which is already existent in the mental model. After this jumping, the focus then has to move one by one.

Access Tail In a queue-structure, like it is assumed in the Verbal Model, the first element, the start of the queue, can be easily accessed. One could assume that the last element, the tail of a mental model, can be accessed just as easily.

Find Reference Object When a premise is read, the object which is already in the mental model has to be found to determine the positioning of the new object. However, if both items already exist in the queue, the relative positioning of the objects *in* the model have to be compared against the new premise. If the focus position is on one of the objects, it could be that for determining the relation between the objects, the focus now only has to move to the other object. However, taking into account the difference between the object *to be located* and the *reference* object, it is possible that the focus has to first move to the *reference* object and then to the object *to be located* to determine the relation between these objects.

Processing the Relation 'right of'

The experiments in Nejasmic et al. (2015) indicated that processing the sentence "a is right of b" is more difficult (at least for speakers of a language which is written from left to right (Krumnack et al., 2011)) than the 'left of'-relation. While the queue-structure in the Verbal Model can account for this fact, we introduced two features to allow a model with a direction-neutral spatial structure, like the one used in PRISM, to show this asymmetry. This feature space comprises 3 possibilities.

Revert When reading "a is right of b", insert b to the right of a first, only to break up that connection and insert it on the left.

Revert only the First Premise Revert only on the first premise, after then the insertion to the left is automatically correct. A model which has the revert-feature, does not have the feature of reverting only the first premise, as the latter is included in the former.

These features result in the following equation for the space of cognitive models:

$$\underbrace{(2^3+1)}_{\textit{implicit direction}} \cdot \underbrace{(2^3+1)}_{\textit{focus}} \cdot \underbrace{3}_{\textit{reverting}} = 243 \qquad (1)$$

Results and Discussion

Best Models for all Participants

Table 5: Correlations r_{τ_b} for individual data from Experiment 1 and all generated models.

	Median	Max	PRISM	Verbal Model
Exp. 1	.197	.22	.22	.218
Exp. 2	-.023	.059	-.05	-.05

To examine the goodness of fit of the generated models for the whole group of participants, we calculated the Kendall rank correlation coefficient τ_b for each model and normalized reading time of participants in the two experiments. The process for normalization was to first correct the reading times of each participant in each condition for outliers, and second to divide the reading times of a specific participant by her maximum reading time. This was done to account for individual processing speed differences and resulted in reading times between 0 and 1 for each trial without losing the relative speed differences of a specific reasoner across trials and conditions. Results from the correlation can be seen in Table 5.

In Experiment 1, PRISM was among the best models, correlating significantly with the normalized reading times ($p < .001$), as did the Verbal Model ($p < .001$). Again, Experiment 2 was much harder to predict for all models. However, also the close to significant correlation of the Verbal Model with the aggregated data disappeared when calculating the correlation with each individual reasoner. It even showed a significant negative correlation ($p = .001$). The correlation coefficient for PRISM was not significant ($p = .051$). Calculating the correlation for both experiments, the models which performed best ($r_{\tau_b} = .171$, $p < .001$) had the following configurations:

The models assume a mental spatial structure that is, contrary to the Verbal Model, persistent in its direction: a rightwards directed queue turned out to perform quite well. Contrary to the results from Nejasmic et al. (2015), models with *no* preliminary integration of features performed better on the two experiments combined. This indicates that this feature needs more investigation in terms of cognitive modeling and psychological investigation. A spatial focus structure with the ability to jump turned out to give the highest performance. The presence of the features considering the access of the last element (tail) and finding the reference object, in the configurations, seems to be, at least within this analysis paradigm, irrelevant.

This indicates several things, among which: (i) that PRISM and the Verbal Model are good models to reproduce the left-to-right tasks, (ii) that for the right-to-left relations, there exist models which can approximate the individual data points better than the models from the literature, and (iii) that restricting cognitive model of spatial reasoning to use only a single model for all participants might soon hit an insurmountable upper bound.

Best Models for Individual Participants

To explore further the idea that individual reasoners may use different strategies, operations or structures, we again calculated the Kendall τ_b coefficient, but this time we allowed for each participant to be assigned the cognitive model which fits best. With this, the median correlation was $r_{\tau_b} = .25$, with a maximum of $r_{\tau_b} = .489$ ($p < .001$).

The previously for the population identified best models only occurred in 42.9% of participants of Experiment 1, and in 14.3% of participants in Experiment 2. The percentage, to which features are present in the individual models, can be seen in Tables 7 and 8.

Table 7: Percentage to which main structural features are present in the best models for the individual reasoner.

	Direction			Preliminary	
	No Direction	Left	Right	Integration	Focus
Exp. 1	16.6%	21.9%	61.6%	52.4%	63.2%
Exp. 2	0%	42.9%	57.1%	28.6%	92.9%

Table 8: Percentage to which secondary features are present in the best models for individual reasoner. The percentages are conditional in the case of focus features, because they only apply if the focus is present.

	Focus				
	Jumping	Tail Access	Find Ref.	Revert	Revert First
Exp. 1	55.7%	50.0%	50.0%	30.8%	65.4%
Exp. 2	65.3%	50.0%	53.8%	17.9%	28.6%

Table 6: Best cost assignment for individual reading time prediction. Possible cost assignments were in the interval between 0 and 1, in increments of 0.1. This assignment yielded a median correlation of $r_{\tau_b} = 0.302$.

Initialization	Insert	Group	Break Links	Move with Dir.	Move against Dir.	Tail Access	New Start	Jump
0.7	0.1	0.1	0.1	0.8	0.5	0.8	0.3	0.8

Alternative Cost Measure

To examine the adequacy of the unary cost measure, we performed a search on the assignment between model actions and mental costs. This was done using Python's `scipy` library for scientific computing[1]. Using a random search algorithm, we explored the space of cost assignments in the interval between 0 and 1, in increments of 0.1. The goal is to find values for the costs, such that the correlation is maximized. For each assignment, we calculated the Kendall τ_b correlation between the predicted costs of each model and each participant's outlier corrected reading times of Experiments 1 and 2 from Nejasmic et al. (2015). We then selected the best model for the individual participants and took the mean of their correlations as the utility for the optimization. The best cost assignment can be taken from Table 6.

Using this method, the best configuration we found achieved a median correlation of $r_{\tau_b} = .302$. The most expensive actions in the assignment were the jumping movement, the access of the tail, and the movement in direction of the queue, or in any direction if there is no implicit direction. Initialization of a model is also costly. The direction *against* the implicit direction was chosen to be less costly than moving *with* the direction. Inserting a new object is not expensive in this assignment, as were breaking connections and setting a new starting node (as was assumed in Krumnack et al., 2011). Similarly, the cost of grouping objects into chunks, which was set to have a cost of n-1 with n being the number of objects in Ragni and Knauff (2013), was also assigned a low cost.

General Discussion

In this paper, we analyzed 243 cognitive models of spatial relational reasoning on their capability to predict *individual* reading times from studies on the continuity effect. These models comprised configurations of features from successful cognitive models from the literature and psychological experiments. While many configurations performed well on aggregated data and a model building direction from left ro right, none of them, including the cognitive models from the literature we based this study on, were able to correctly predict reading times for a direction from right to left. We then followed the notion that different people might use different strategies, and investigated whether assigning a specific cognitive model to individual reasoners would greatly improve performance. While we reached a better correlation using this method, it was still in question why the correlation did not increase even further. We thus challenged the unary cost measure proposed in Ragni and Knauff (2013). Using a search algorithm, we investigated whether a different cost assignment would lead to better predictions for the individual. While the fit got better, it still demonstrates that the individual variety is not yet captured. Especially Experiment 2, which explored a presentation of spatial information using the relation 'right of' revealed a low correlation, on the individual, but also on the aggregated level.

We explored the space of possible cognitive models for spatial relational reasoning using features present in cognitive models from the literature. However, this space did not yield a model which was able to predict reading times across tasks robustly. This can be due to several issues: (i) the core assumption of these models, that we build an abstract spatial representation (a mental model) is wrong, (ii) the true mental processes in our brain when processing spatial relational information differ from those assumed in the models of the literature, or (iii) the assumption of a sequential processing of spatial information has to be revised. The construction of a mental model nonetheless is a notion which is broadly accepted (Johnson-Laird, 2004; Ragni & Knauff, 2013). If the mental processes of model construction differ from those presumed by the state-of-the-art cognitive models, it stands to reason what other processes could be taking place. The sequential processing is common to most cognitive spatial models (Friemann & Ragni, 2018). Modeling of individual data is limited, as individual data, and especially reaction time, is noisy. However, if cognitive models fit averaged data well, but are not able to capture any single individual in the experiment, the meaning of cognitive modeling and goodness-of-fit needs to be reevaluated.

Conclusion

It seems we are still far from understanding the way our mind integrates spatial information. This study challenged common assumptions and practices from the area of cognitive modeling for spatial reasoning. These customs are found to be insufficient when applying them to the modeling of *whole* empirical data sets instead of the aggregated data. There is still much to be learned about the way we process streams of information, what mental operations are performed, and in how far we can generalize conclusions from the aggregated data to the individual human mind.

Acknowledgements

This paper was supported by DFG grants RA 1934/3-1, RA 1934/2-1 and RA 1934/4-1 to MR. We also thank Sara Todorovikj for assistance and helpful comments.

[1] `www.scipy.org/`

References

Ehrlich, K., & Johnson-Laird, P. N. (1982). Spatial descriptions and referential continuity. *Journal of Verbal Learning and Verbal Behavior*, *21*(3), 296–306.

Friemann, P., & Ragni, M. (2018). Cognitive computational models of spatial relational reasoning: A review. In Thrash, Kelleher, & Dobnik (Eds.), *The 3rd Workshop on Models and Representations in Spatial Cognition (MRSC-3)*. Retrieved from dobnik.net/simon/events/mrsc-3/

Johnson-Laird, P. N. (2004). The history of mental models. In *Psychology of Reasoning* (pp. 189–222). Psychology Press.

Knauff, M., Rauh, R., Schlieder, C., & Strube, G. (1998). Continuity effect and figural bias in spatial relational inference. In M. A. Gernsbacher & S. J. Derry (Eds.), *Proceedings of the 20th Annual Meeting of the Cognitive Science Society* (pp. 573–578). Mahwah, NJ: Lawrence Erlbaum Associates.

Krumnack, A., Bucher, L., Nejasmic, J., & Knauff, M. (2010). Spatial reasoning as vthe most prototypical / erbal reasoning. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society* (pp. 1002–1007). Austin, TX: Cognitive Science Society.

Krumnack, A., Bucher, L., Nejasmic, J., Nebel, B., & Knauff, M. (2011). A model for relational reasoning as verbal reasoning. *Cognitive Systems Research*, *12*(3-4), 377–392.

Nejasmic, J., Bucher, L., & Knauff, M. (2015). The construction of spatial mental models – A new view on the continuity effect. *The Quarterly Journal of Experimental Psychology*, *68*(9), 1794–1812.

Polk, T. A., & Newell, A. (1995). Deduction as verbal reasoning. *Psychological Review*, *102*(3), 533–566.

Ragni, M., Fangmeier, T., & Brüssow, S. (2010). Deductive spatial reasoning: From neurological evidence to a cognitive model. In D. D. Salvucci & G. Gunzelmann (Eds.), *Proceedings of the 10th International Conference on Cognitive Modeling* (pp. 193–198). Philadelphia, PA: Drexel University.

Ragni, M., & Knauff, M. (2013). A theory and a computational model of spatial reasoning with preferred mental models. *Psychological Review*, *120*(3), 561–588.

Schlieder, C., & Berendt, B. (1998). Mental model construction in spatial reasoning: A comparison of two computational theories. *Mind modelling: A cognitive science approach to reasoning, learning and discovery*, 133–162.

Schultheis, H., & Barkowsky, T. (2011). Casimir: an architecture for mental spatial knowledge processing. *Topics in Cognitive Science*, *3*(4), 778–795.

Testing the limits of non-adjacent dependency learning: Statistical segmentation and generalization across domains

Rebecca L. A. Frost (rebecca.frost@mpi.nl)
Language Development Department, Max Planck Institute for Psycholinguistics, Nijmegen, NL

Erin S. Isbilen (esi6@cornell.edu)
Department of Psychology, Cornell University, Ithaca, NY, USA

Morten H. Christiansen (christiansen@cornell.edu)
Department of Psychology, Cornell University, Ithaca, NY, USA

Padraic Monaghan (p.j.monaghan@uva.nl)
Department of English Language and Culture, University of Amsterdam, NL;
Department of Psychology, Lancaster University, UK

Abstract

Achieving linguistic proficiency requires identifying words from speech, and discovering the constraints that govern the way those words are used. In a recent study of non-adjacent dependency learning, Frost and Monaghan (2016) demonstrated that learners may perform these tasks together, using similar statistical processes — contrary to prior suggestions. However, in their study, non-adjacent dependencies were marked by phonological cues (plosive-continuant-plosive structure), which may have influenced learning. Here, we test the necessity of these cues by comparing learning across three conditions; *fixed phonology*, which contains these cues, *varied phonology*, which omits them, and *shape*s, which uses visual shape sequences to assess the generality of statistical processing for these tasks. Participants segmented the sequences and generalized the structure in both auditory conditions, but learning was best when phonological cues were present. Learning was around chance on both tasks for the visual shapes group, indicating statistical processing may critically differ across domains.

Keywords: statistical learning; speech segmentation; generalization, language learning; non-adjacent dependencies; implicit learning

Background

Learners must master a number of critical tasks in order to reach linguistic proficiency, including learning how to segment individual words from speech, and learning to identify the constraints that govern the way those words are structured and used. Learners are remarkably adept at these tasks, thanks in part to the myriad cues that speech contains that may assist learning. One such cue is the statistics that describe co-occurrences of items in speech; for instance, the co-occurrence of syllables provides a helpful cue to what constitutes possible words, while information about how those words are used in combination helps learners to discern how the language operates. The ability to detect and draw on this distributional information - *statistical learning* - is suggested to play a key role in language acquisition, for both segmenting speech and for learning about grammatical structure (e.g., Conway, Bauernschmidt, Huang, & Pisoni, 2010; Frost, Monaghan, & Christiansen, 2019; Redington & Chater, 1997).

Since word- and structure-learning appear to have distinct requirements, it is unsurprising that the nature of the (statistical) processes that underlie these tasks has been subject to substantial debate (e.g., Peña, Bonatti, Nespor, & Mehler, 2002; Perruchet, Tyler, Galland, & Peereman, 2004). Central to these discussions have been questions concerning the types of computations required to discover word-like and rule-like items in speech, and learners' capacity to do so by computing over co-occurrence statistics.

These issues have been extensively tested using a classic artificial language learning paradigm (Peña et al., 2002), which examines learners' ability to acquire linguistic structure that is defined in terms of non-adjacent dependencies (i.e., an AxC structure, where A and C are syllables that reliably co-occur, regardless of which x syllable intervenes). AxC languages are used to jointly assess learners' capacity for statistical word and structure learning, since they contain novel words that learners must discover (AxC strings), in addition to structural regularities within those words (A-C relationships).

Initial studies using this paradigm suggested that learners perform statistical computations on the non-adjacent dependencies to segment the speech into individual AxC strings (or *words*), but perform more abstract computations on those words in order to learn about their structure - and perhaps do so only when speech segmentation has been resolved (typically by inserting pauses between words in the training stream).

A recent study by Frost and Monaghan (2016) expanded on this work, aiming to shed further light on two key questions about how word- and structure-learning unfold in language acquisition: whether these tasks occur sequentially

or simultaneously, and whether they may actually utilize similar statistical computations – contrary to prior suggestions. In their study, participants were able to draw on the non-adjacent dependencies to segment continuous speech into words, *and* to learn about the non-adjacent dependency structure that those words contained, possibly simultaneously (though further work is required to conclusively establish the time-course of learning for these tasks). The key difference between this and earlier work on this phenomenon was a slight methodological change which addressed a possible confound in the previous measure of generalization. Specifically, prior generalization tasks typically required learners to indicate a preference for 'rule words' over part-words, with rule words comprising a trained dependency, intervened by an onset/coda from another dependency (e.g., $A_1A_2C_1$ or $A_1C_2C_1$). While such comparisons do permit assessment of preference for the overall structure, they require learners to use trained A and C items flexibly in a way that deviates from their knowledge of syllable position, which may affect performance. Indeed, using amended test items (trained dependencies with entirely novel intervening items), Frost and Monaghan (2016) demonstrated that adults can segment statistical nonadjacent dependencies and generalize them to novel grammatically consistent instances in the absence of additional information, such as pauses between words (see Isbilen, Frost, Monaghan, & Christiansen, 2018, for a replication of this effect).

This finding was contrary to prior suggestions that these tasks are fundamentally computationally distinct (e.g., Peña et al., 2002), and provides crucial evidence to suggest that learners may draw on the same type of statistical processing mechanisms for both of these tasks, and they may do so at the same time during language learning.

However, one possibility that cannot be overlooked is that learning in this study was not just driven by computations over transitional probabilities; learning may have been assisted by the phonological properties of the language. In line with Peña et al.'s (2002) landmark study, Frost and Monaghan (2016) employed an artificial language that contained both statistical dependencies between elements, and phonological structure, which aligned with the non-adjacency structure such that A and C syllables contained plosives, whereas intervening x syllables contained continuants.

Prior research has noted that the pattern of phonological information in artificial languages can significantly benefit learning, and phonological similarity between related elements has been found to support learning of non-adjacent dependencies in particular. For instance, in a series of experiments with a similar paradigm, Newport and Aslin, (2004) demonstrated that learning nonadjacent dependencies between syllables was remarkably difficult to accomplish in the absence of phonological cues (though the difficulty there may also have been due to additional factors, including learnability of the language - i.e., the number of dependencies, and the number of intervening items, which has been shown to impact learning - together with the relative complexity of some of the tests). Similarly, in Gomez and Gerken (1999), dependency learning was supported by phonological distinctions between A/C items and x items, where A and C were bisyllabic, and x were monosyllabic. Yet, research has also suggested that this phonological information should not be essential for learning to take place (Onnis, Monaghan, Christiansen, & Chater, 2004). Further research is therefore required to assess the extent to which this phonological information guided learning in Frost and Monaghan's (2016) study, to determine whether learners can indeed discover words and structures together, from distributional information alone.

In the present paper, we replicate Frost and Monaghan (2016), to confirm that participants can compute over non-adjacent dependencies to learn about both words and structure. We also test whether scores on these tasks correlate, to further assess whether these abilities are similar, or distinct. Crucially, we also compare performance for this replication against that for a condition in which participants are trained on the same language but with a more varied phonology (i.e., without phonological cues). Examining the extent to which segmentation and generalization are possible in the absence of these phonological cues will provide critical insights into how learners rely on statistical computations during language acquisition, by removing the possibility that successful performance is due to additional information outside of the syllable distribution.

While manipulating properties of the language allows us to determine how multiple cues interact with statistical learning, it does not inform us about whether that learning is due to domain-specific mechanisms, or whether language learning involves the specific application of general-purpose learning mechanisms (Frost, Monaghan, & Tatstumi, 2017; Siegelman & Frost, 2015). To further explore adults' capacity to compute non-adjacent dependencies, we also assessed whether their ability to do so is unique to language, by extending the paradigm to examine non-adjacent dependency learning from non-linguistic sequences (comprising shapes). This condition will help constrain theorizing on the generality of the mechanisms used for these tasks.

Thus, in this study we examine whether adults' capacity for segmenting and generalizing non-adjacent dependencies extends to more varied linguistic stimuli, or if it is contingent on a correspondence between distributional and phonological cues to structure. We will also assess whether this capacity is similar or different across modalities. We expect that participants will demonstrate knowledge of words and within-word structure (i.e., non-adjacent dependencies) in both language conditions (Frost & Monaghan, 2016; Onnis et al., 2004), and in the shapes group, in line with the suggestion that statistical learning mechanisms may serve learning broadly across modalities (e.g., Frost et al., 2017). We predict that segmentation and structure learning will benefit from phonological cues, but that these will not be essential for learning (Onnis et al., 2004). Further, we expect that structure learning will be better for linguistic than nonlinguistic input (due to increased experience with learning linguistic structure

relative to structured sequences of shapes; Siegelman & Frost, 2015).

Method

Participants

90 Cornell University undergraduates (age: M = 19.6 years, range = 18-24 years; 49 females, 41 males) participated for course credit. All participants were native English speakers.

Design

Participants were randomly allocated to one of three conditions (each N = 30): *fixed phonology*, where AxC sequences contained plosive-continuant-plosive structure (Frost & Monaghan, 2016, Peña et al., 2002), *varied phonology*, which randomized the allocation of plosives and continuants to different positions within words, and *shapes*. These conditions permit comparison of learning from the original training input (fixed phonology) with an amended version containing no reliable phonological cues to word structure (varied phonology), and also a non-linguistic analogue. This will provide critical assessment of whether the pattern of learning demonstrated by Frost and Monaghan (2016) is unique to the properties of the input used in that study, or whether it can be extended to more varied linguistic input, as well as input in a different modality.

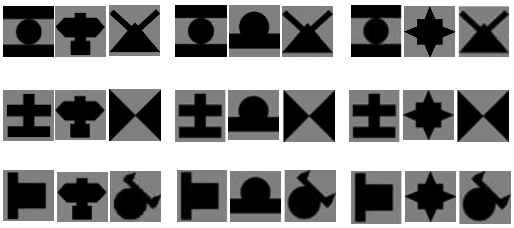
Stimuli

Speech stimuli were created with Festival speech synthesiser, from a pool of 9 monosyllabic items (*pu, ki, be, du, ta, ga, li, ra, fo*), as used in Peña et al. (2002), and three additional monosyllabic items (*ve, zo, thi*). These additional syllables were reserved for the generalization task for the fixed phonology group in line with prior research (Frost & Monaghan, 2016), but formed part of the general syllable pool for the varied phonology group, to maximise variability. Shape stimuli were created from the Fiser and Aslin (2002) set of novel shapes (novel shapes in black on a grey background).

Familiarization Syllables/shapes were concatenated into triadic sequences that followed an AxC structure, with A, x, and C representing an individual syllable/shape. There were three A-C pairings, and three x items that could be used in all pairings ($A_1X_{1-3}C_1$, $A_2X_{1-3}C_2$, and $A_3X_{1-3}C_3$), giving 9 strings in total.

For the fixed phonology condition, syllables were mapped onto words pseudorandomly, such that A and C syllables were plosives, whereas x syllables were continuants, meaning each AxC string had a plosive-continuant plosive structure (e.g., *puraki*). For the varied phonology condition, syllables were randomly allocated to A, x, and C positions, meaning there were no reliable phonological cues that could guide learning. For the shapes condition, shapes were randomly allocated to A, x, and C positions, providing a visual non-linguistic analogue of the varied phonology condition. See Table 1 for example stimuli for each condition.

Table 1: Example stimuli for each condition

Condition	Triads
Fixed Phonology	*puliki, puraki, pufoki* *beliga, beraga, befoga* *talidu, taradu, tafodu*
Varied Phonology	*livedu, liradu, likidu* *fovezo, forazo, fokizo* *bevepu, berapu, bekipu,*
Shapes	

Syllable/shape triplets were concatenated into familiarization streams containing 900 sequences (100 repetitions of each individual AxC sequence), in line with the materials used by Frost and Monaghan (2016). For speech stimuli, this was done using the Festival speech synthesizer (Black et al., 1990), and for shape stimuli this was done using Eprime 2.0. For all conditions, training streams contained no immediate repetition of individual AxC sequences.

For the fixed phonology and varied phonology conditions, the training stream lasted for 10.5 minutes, and was edited to have a 5-second fade-in and fade-out, to avoid providing cues to word boundaries.

For the shape sequences, presentation of the training stream took 22 minutes overall. For comfort this was split into 3 blocks of 300 sequences, and participants were invited to take short breaks in between blocks if desired. To ensure stimuli were analogous to the linguistic input, sequences were programmed such that shapes were presented sequentially, one by one. Shapes were presented for 225 ms in the centre of the screen, with a 225 ms inter-item interval between all shapes for comfortable viewing (note that since this occurs between all shapes, it does not cue segmentation). Presentation criteria were in line with those used in a comparable study by Frost et al. (2017). Analogous to the 5 second fade-in/-out applied to the speech streams, visual sequences always began and ended mid-triad, to prevent participants receiving any information about sequence boundaries at the start/end of the streams (this is true for the beginning and end of the entire sequence, and also for either side of the scheduled breaks).

To control for the relative ease of learning particular dependencies, for each condition 8 versions of the language were generated and counterbalanced across participants. For the varied phonology and shapes stimuli, these were created by randomly assigning syllables/shapes to A, x and C roles. For the fixed phonology stimuli, these were created by

randomly assigning plosives to the A and C roles, while x items were always the same (see Frost & Monaghan, 2016).

Testing Learning was assessed using a two-alternative forced-choice (2AFC) test of segmentation and generalization. This contained 18 trials, nine of which assessed segmentation, and nine of which assessed generalization. Segmentation trials contained word versus part-word comparisons, with words being AxC items that occurred in the training stream, and part-words spanning word boundaries such that they comprised the end of one word and the start of another (e.g., xCA, CAx). Generalization trials contained rule-word versus part-word comparisons, where rule-words were trained dependencies but with novel intervening items (e.g., A_1NC_1), and part words were structured as before, but with one syllable replaced with a novel syllable (e.g., NCA, CNA, CAN). This was to control for the possibility that participants' responses on these trials were due to novelty alone (see Frost & Monaghan, 2016, for further discussion. Ongoing work by Isbilen, Frost, Monaghan and Christiansen further explores these generalization effects using $A_1N_1C_1$ vs. $A_1N_1C_2$ comparisons).

Procedure

Familiarization Participants were presented with a familiarization stream which comprised either sequences of speech (10.5 minutes), or sequences of shapes (~22 minutes). Participants were instructed to pay attention to the sequences, and the shapes group was instructed to take optional breaks at the designated pauses if required.

Testing At test, participants completed a 2AFC task comprising 18 trials; nine segmentation trials (words versus part-word comparisons) and nine generalization trials (rule-words versus part-word comparisons). Presentation of segmentation and generalization trials was randomized. Participants were instructed to carefully listen to/look at each test pair, and indicate which of the two best matched the training stream they had just heard/seen.

Results and Discussion

Accuracy Scores

Accuracy scores for each condition are shown in Figure 1. One-sample t-tests (two-tailed) were conducted on the data for each group to compare performance to chance.

For the fixed phonology group, performance was significantly above chance for both the segmentation (M = .709, SD = .245), $t(29)$ = 4.659, p < .001, d = .853 and generalization tasks (M = .661, SD = .173), $t(29)$ = 5.100, p < .001, d = .936, replicating Frost and Monaghan's (2016) demonstration that learners can segment and generalize non-adjacent dependencies from continuous speech. For the varied phonology group, performance was also significantly above chance for both tasks (segmentation: M = .623, SD = .199, $t(29)$ = 3.391, p = .002, d = .618; generalization: M = .594, SD = .217, $t(29)$ = 2.366, p = .025, d = .433), suggesting that acquisition of statistically defined non-adjacent dependencies in this task is not contingent on the phonological properties of the speech input (i.e., phonological similarity between dependent syllables).

For the shapes group, however, performance on the segmentation task was only marginally above chance (M = .552, SD = .156), $t(29)$ = 1.827, p = .078, d = .333), and performance on the generalization task was at chance level (M = .485, SD = .205), $t(29)$ = -.410, p = .685, d = -0.073) – indicating that adults' ability to segment and generalize sequences using non-adjacent transitional probabilities may not extend to visually presented non-linguistic input.

Segmentation and generalization performance were significantly correlated for the fixed phonology (r = .385, p = .036) and varied phonology (r = .625, p <. 001) groups, but not for the shapes group (r = .281, p = .133).

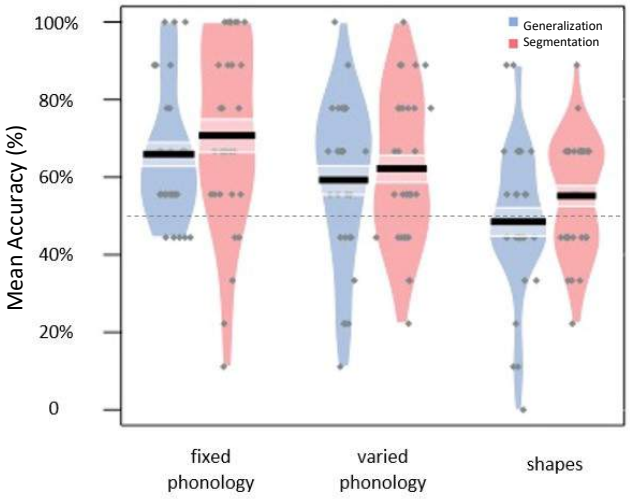


Figure 1. Pirate plot depicting performance on the segmentation and generalization tasks for each condition. Mean scores are shown in black, with standard error in white. The distribution of scores is depicted in red for the segmentation task, and blue for the generalization task, with individual participants' scores in grey. The dashed line indicates chance level.

Comparing performance across groups

To compare performance across each of these groups, Generalized Linear Mixed Effects (GLMER) analysis was conducted on the data, examining whether segmentation and generalization scores differed according to whether participants were trained on sequences comprising varied or fixed phonology, or shapes. A significant main effect of condition would imply different overall performance across the groups, while a significant main effect of test type would indicate that participants performed differently on the segmentation and generalization tasks overall. An interaction between these variables would tell us that participants' performance on the segmentation and generalization tasks differed as a function of their condition – indicating that adults' capacity for statistical learning on these tasks differs

across conditions, and possibly across domains, shedding light on the generality of the possible mechanism(s) that may underlie performance.

GLMER analysis was performed on the data (Baayen, Davidson, & Bates, 2008), modelling the probability (log odds) of response accuracy at test considering variation across participants and materials. The model was built incrementally, with random effects of subjects, particular test-pairs, and language version (to control for variation across the randomized assignments of phonemes to syllables). Random slopes were omitted if the model failed to converge with their inclusion (Barr, Levy, Scheepers, & Tily, 2013).

We then added condition (varied phonology, fixed phonology, and shapes) as a fixed effect, and considered its effect on model fit with likelihood ratio test comparisons. There was a significant effect of condition (model fit improvement over the model containing random effects: $\chi(2)^2 = 7.903$, $p = .019$), with the shapes group performing significantly worse than the fixed phonology group (difference estimate = -.767, *SE* = .257, *z* = -2.987, *p* = .003). The fixed phonology group also outperformed the varied phonology group, however this difference was marginal (difference estimate = -.389, *SE* = .217, *z* = -1.788, *p* = .074). We then added test type (segmentation and generalization), to see whether participants performed differently on each type of task. The effect of test type was marginal (model fit improvement over the model containing random effects: $\chi(2)^2 = 3.144$, $p = .076$) with participants performing better on the segmentation task than the generalization task (difference estimate = .224, *SE* = .125, *z* = 1.791, *p* = .073).

We then added the interaction between condition and test type, to see whether performance on the tasks differed according to the input participants had received. The interaction was not significant (model fit improvement over the model containing random effects: $\chi(2)^2 = .366$, $p = .833$), suggesting participants performed similarly across each of the conditions. See Table 2 for a summary of the final model.

Table 2: Summary of the GLMER (log odds) for accuracy scores.

Fixed effects	Estimated coefficient	SE	Wald confidence intervals 2.50%	97.50%	z	Pr (>|z|)
(Intercept)	.7405	.2082	.3325	1.149	3.557	.0004
Condition: Shapes	-.7658	.2583	-1.272	-.2595	-2.965	.003
Condition: Varied Phono	.-.3883	.2183	-.8161	.0395	-1.779	.0753
Test_type	.2235		-.0211	.4680	1.791	.07332
Random effects	**Variance**	**Std. Dev.**				
Subject (Intercept)	.355	.5958				
Test Pair (Intercept)	.5871	.773				
Lang_version	.0019	.0435				
	AIC	BIC	logLik	Deviance		
	2097.6	2140.8	-1040.8	2081.6		

1620 observations, 90 participants, 18 trials. R syntax for the final model is: NAD_DG3 <- glmer (testresponse.ACC ~ condition + test_type + (1|subject) + (1+lang_ver|test_pair), data =NAD_DG, family=binomial, control=glmerControl(optimizer="bobyqa",optCtrl=list(maxfun=100000))).

General Discussion

Recent evidence for the similarity (and possible simultaneity) of statistical segmentation and generalization has advanced our understanding of the way these processes unfold during language acquisition (see Frost & Monaghan, 2016, and see e.g,, Peña et al., 2002 and Perruchet et al. 2004, for more on the earlier debate about the nature of these tasks). Yet, due to the phonological properties of the training language, it is possible that learning in this recent study was not solely contingent on the statistical regularities contained within the language; learning may have been assisted by the plosive-continuant-plosive structure that AxC sequences adhered to (e.g., Newport & Aslin, 2004).

To explore this possibility, the study at hand examined adults' capacity for non-adjacent dependency learning across three conditions; the first of which used the input from Frost and Monaghan (2016) (see also Peña et al., 2002), which contained the phonological structure described above (termed the *fixed phonology* condition). The second condition omitted these phonological cues, such that AxC sequences had no fixed phonological structure (the *varied phonology* condition). The third condition tested learning from sequences of shapes, to provide a non-linguistic assessment of non-adjacent dependency learning, with a view of considering whether learning was comparable across modalities — perhaps drawing on similar statistical mechanisms. The critical test was whether participants in each group demonstrated learning (i.e., performed above chance), and whether performance in the varied phonology and shapes groups differed significantly from the fixed phonology group.

Participants in both language conditions performed significantly above chance on the segmentation and generalization tasks. This finding replicates the results of Frost and Monaghan (2016), showing that speech segmentation and structural generalization may proceed together during language learning, and can be accomplished from the same distributional statistics (though additional research is required to conclusively establish the precise time-course of learning for these tasks). Further, our results demonstrate that adults' capacity for learning non-adjacent dependencies extends to more phonologically diverse input. However, the difference in overall performance in these conditions was approaching significance, with results indicating that phonological cues were advantageous for learning (evidenced by marginally higher scores for the fixed phonology than the varied phonology group) — in line with Newport and Aslin's (2004) suggestion that such cues were important for learning. Critically though, our data indicate that these cues were not essential (Onnis et al., 2004).

In previous studies of word and structure learning, segmentation and generalization have tended to be tested separately. In the current study, these tasks were completed by all participants (within subjects). We show that the same learners can segment non-adjacencies from speech, and generalize them to new instances (see also Isbilen et al., 2018). In line with previous studies, performance on the segmentation task was higher than that seen for the generalization task (see Isbilen et al., 2018, for a comparable finding), and crucially performance on these tasks was significantly correlated for both language conditions — adding further support to the notion that they may be underpinned by similar mechanisms.

The results for the shapes group followed the same general pattern as those seen in the varied phonology and fixed phonology conditions, with a trend toward higher performance on the segmentation task than the generalization task. However, scores for this group were significantly lower than those seen for the fixed phonology group, with accuracy scores on the segmentation task being only marginally above chance, while performance on the generalization task was at chance level. It is important to note that the shape stimuli differ from the speech stimuli in two key ways: they are both visual and non-linguistic, and therefore differ both in modality and domain. Thus, this pattern of results could be attributed to a number of possible explanations.

One possibility for the difference between the language and the shape task is that there are critical differences in statistical learning across modalities, with tasks being underpinned by different mechanisms (e.g., Conway & Christiansen, 2005). A second possibility is that, for the shapes group, performance could have been negatively affected by participants' relative lack of experience with learning distributionally defined streams containing sequences of visual non-speech input (compared to experience with heard speech) (e.g., Siegelman et al., 2018). Another possibility is that the difference in performance is due to key differences in task demands: in the speech conditions, the presentation of stimuli is such that participants have no choice but to attend (be that actively, or passively). However, in the shapes condition, this is not necessarily the case. Thus, it is possible that the lower scores observed for this group are (at least in part) due to participants attending less to the input during training (and thus, learning less during familiarization). Ongoing replications of this work employing a cover task that maintains participants' attention will help to unpack these possibilities.

To summarise, these data provide further evidence that adults can compute non-adjacent dependencies to discover words and within-word structure from continuous speech. This supports the notion that these tasks may be underpinned by similar statistical processes, and may occur together during language learning. Further, results illustrate that these abilities are not dependent on phonological cues, suggesting that adults' capacity for performing statistical computations over linguistic input is even more powerful than previously suggested.

Acknowledgments

We thank Dante Dahabreh, Phoebe Ilevbare, Eleni Kohilakis, Farah Mawani, Olivia Wang, Emily Zhang and Sophia Zhang for their help with data collection. ESI was supported by a National Science Foundation Graduate Research Fellowship (#DGE-1650441). PM was supported by the International Centre for Language and Communicative Development (LuCiD) at Lancaster University, funded by the Economic and Social Research Council (United Kingdom; ES/L008955/1).

References

Baayen, R. H, Davidson, D. J, & Bates D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 390–412.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255-278.

Black, A. W., Taylor, P., & Caley, R. (1990). *The festival speech synthesis system*. Edinburgh, UK: Centre for Speech Technology Research (CSTR), University of Edinburgh, http://www.cstr.ed.ac.uk/projects/festival.html

Conway, C. M., Bauernschmidt, A., Huang, S. S., & Pisoni, D. B. (2010). Implicit statistical learning in language processing: Word predictability is the key. *Cognition, 114,* 356-371.

Conway, C. & Christiansen, M.H. (2005). Modality constrained statistical learning of tactile, visual, and auditory sequences. *Journal of Experimental Psychology: Learning, Memory & Cognition, 31,* 24-39.

Frost, R. L. A., & Monaghan P. (2016). Simultaneous segmentation and generalisation of non-adjacent dependencies from continuous speech. *Cognition, 147*, 70-74.

Frost, R. L. A., Monaghan, P., & Christiansen, M. H. (2019). Mark my words: high frequency marker words impact early stages of language learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*.

Frost, R. L. A., Monaghan P. & Tatsumi, T. (2017). Domain-General Mechanisms for Speech Segmentation: The Role of Duration Information in Language Learning. *Journal of Experimental Psychology: Human Perception and Performance 43*(3), 466-476.

Gómez, R., & Gerken, L. (1999). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. Cognition, 70, 109–135.

Isbilen, E. S., Frost, R. L. A, Monaghan, P., & Christiansen, M. H. (2018). Bridging artificial and natural language learning: Comparing processing- and reflection-based measures of learning. *Proceedings of the 40th Annual Meeting of the Cognitive Science Society*. Madison, WI, USA.

Newport, E. L., & Aslin, R. N. (2004). Learning at a distance. I. Statistical learning of non-adjacent dependencies. *Cognitive Psychology, 48*, 127–162.

Onnis, L., Monaghan, P., Christiansen, M. H., & Chater, N. (2004). Variability is the spice of learning, and a crucial ingredient for detecting and generalizing in non-adjacent dependencies. *Proceedings of the 26th annual conference of the cognitive science society. Mahwah, NJ: Lawrence Erlbaum.*

Peña, M., Bonatti, L., Nespor, M., & Mehler, J. (2002). Signal-driven computations in speech processing. *Science, 298*, 604–607.

Perruchet, P., Tyler, M. D., Galland, N., & Peereman, R. (2004). Learning non-adjacent dependencies: No need for algebraic-like computations. *Journal of Experimental Psychology: General, 133*(4), 573-583).

Redington, M. & Chater, N. (1997). Probabilistic and distributional approaches to language acquisition. *Trends in Cognitive Sciences, 1*(7), 273-281.

Siegelman, N., Bogaerts, L., Elazar, A., Arciuli, J., & Frost, R. (2018). Linguistic entrenchment: Prior knowledge impacts statistical learning performance. *Cognition*, 177, 198-213.

Siegelman N, & Frost R. (2015). Statistical learning as an individual ability: Theoretical perspectives and empirical evidence. Journal of Memory and Language, 81:105–120.

Reframing Convergent and Divergent Thought for the 21st Century

Liane Gabora (liane.gabora@ubc.ca)
Department of Psychology, University of British Columbia
Kelowna BC, V1V 1V7, CANADA

Abstract

Convergent and divergent thought are promoted as key constructs of creativity. *Convergent thought* is defined and measured in terms of the ability to perform on tasks where there is one correct solution, and *divergent thought* is defined and measured in terms of the ability to generate multiple solutions. However, these characterizations of convergent and divergent thought presents inconsistencies, and do not capture the reiterative processing, or 'honing' of an idea that characterizes creative cognition. Research on formal models of concepts and their interactions suggests that different creative outputs may be projections of the same underlying idea at different phases of a honing process. This leads us to redefine convergent thought as thought in which the relevant concepts are considered from *conventional contexts,* and divergent thought as thought in which they are considered from *unconventional contexts.* Implications for the assessment of creativity are discussed.

Keywords: Alternate Uses Task; concepts; context; convergent thinking; divergent thinking; potentiality; quantum model; Remote Associates test

Introduction

Other species perceive, make decisions, and take action, but our ability to adapt ideas to our own needs, tastes, and perspectives, and express ourselves through language, art, technology, and other means, is exceptional. Thus, understanding creative thinking is central to understanding our humanness.

In creativity research, as in other areas of cognitive science, there is a long history of dual process theories, which assert that there are two kinds of thought, or that thought varies along a continuum between two extremes (Evans & Frankish, 2009; James, 1890/1950, Sloman, 1996). In the creativity literature the distinction is usually made between convergent and divergent thinking[1]. *Convergent thought* is defined and measured in terms of the ability to perform on tasks where there is a single correct solution, while *divergent thought* is defined and measured in terms of the ability to generate multiple different solutions (Guilford, 1967). A widely used test of convergent thinking is the Remote Associates Test (RAT) (Mednick, 1968). A typical RAT question is: What is the common associate of TANK, TABLE, and HILL? The answer is: TOP. A widely used divergent thinking test is the Alternate Uses task, which asks questions like 'think of as many uses as you can for a brick' (Christensen, Guilford, Merrifield, & Wilson, 1960). Responses are most often rated in terms of *fluency,* the total number of ideas generated in a given time. Often they are additionally rated in terms of *originality,* the number of unusual or statistically infrequent ideas. Fluency and originality are considered to reflect the quantity and quality of ideation performance, respectively. Occasionally they are also rated in terms of *flexibility,* the number of different categories of ideas. On rare occasions answers are rated for *elaboration:* the amount of detail given, or evidence that the individual has followed an associative pathway for some distance.

Although these characterizations of convergent and divergent thought have stuck for half a century, as formulated, they present inconsistencies. For example, it is often said that a creatively demanding problem requires both convergent and divergent thought (e.g., Beersma & De Dreu, 2005; Gibson, Folley, & Park, 2009; Kerr & Murthy, 2004). However, given that convergent and divergent thought are defined in terms of the number of correct solutions, this makes no sense. A problem either has one correct solution or it has many; it cannot have both one and many. Moreover, the way convergent and divergent thought have been defined and measured is inconsistent with how people think about creativity; for example, although divergent thinking is thought to be the most promising candidate for the foundation of creative ability (Plucker & Renzulli, 1999; Runco, 2007), performance on the RAT would seem to be a better indicator of creativity than many tasks that would be classified as a divergent thinking task, such as 'list as many things as you can that are red'. Finally, it is often noted that earlier responses on a divergent thinking task are less creative than latter ones (Beaty & Silvia, 2012), but if divergent thinking is characterized in terms of the number of possible responses, this is the opposite of what one should expect, because with each response one gives, the number of remaining possible responses decreases by one. Thus, the conventional view would predict that, as one proceeds, one should start thinking more *convergently,* not more divergently.

More fundamentally, as noted elsewhere (Piffer, 2012), divergent thinking research, and creativity research in general, emphasizes the generation of multiple ideas over what is sometimes called *honing*—recursively reflecting on a question or idea by viewing it from different perspectives with the output of each such reflection providing the input to the next (Gabora, 2007, 2017). One thereby comes to a deeper, more nuanced understanding of it. Honing differs from elaboration in that it does not include additions or

[1] Sometimes the distinction is between associative and analytic thought (e.g., Chrusch, C. & Gabora, L., 2013), or executive versus generative (e.g., Ellamil, Dobson, Beeman, & Christoff, 2012). See (Sowden et al., 2014) for how convergent and divergent thinking relate to other dual process theories.

modifications to the idea that are tacked on willy-nilly; it refers specifically to modifications that arise in response to an overarching conceptual framework that is shepherding[2] the creative process. The structure of this overarching framework reflects the individual's *worldview:* their self-organizing web of understandings about their world and their place in that world (in other words, the creator's mind as experienced 'from the inside').

Like other self-organizing systems, a worldview continually interacts with and adapts to its environment to minimize internal *entropy,* a measure of uncertainty and internal disorder. Hirsh, Mar, and Peterson (2012) use the term *psychological entropy* to refer to anxiety-provoking uncertainty, which they claim humans attempt to keep at a manageable level. Noting that uncertainty can be experienced not just negatively as anxiety but also positively as a wellspring for creativity (or both), the term psychological entropy has been expanded to refer to *arousal*-provoking uncertainty. Redefining psychological entropy in terms of arousal rather than anxiety is consistent with findings that creative individuals exhibit greater openness to experience and higher tolerance of ambiguity (Feist, 1998), which could predispose them to states of uncertainty or worldview inconsistency (Gabora, 1999). Their higher variability in arousal (Martindale & Armstrong, 1974) reflects a predisposition to invite situations that increase psychological entropy, experience them positively, and resolve them. In this way, psychological entropy—a macro-level variable acting at the level of the worldview as a whole—generates emotions that play a role in guiding and monitoring creative tasks.

Thus, honing continues until psychological entropy decreases to an acceptable level. In Piagetian terms, during honing the individual assimilates each new understanding of the idea, and the individual's worldview changes to accommodate this new understanding. Insight is then explained in terms of *self-organized criticality* (SOC) (Gabora, 2001, 2017; Schilling, 2005), a phenomenon wherein, through simple local interactions, complex systems tend to find a critical state poised at the cusp of a transition between order and chaos, from which a single small perturbation occasionally exerts a disproportionately large effect (Bak, Tang, & Weisenfeld, 1988). Thus, while most thoughts have little effect on one's worldview, an idea we call *insightful* is one for which one thought triggers another, which triggers another, and so forth in an avalanche of conceptual change.

Surely, whether one is writing a novel, or composing a symphony, or inventing a new kind of solar panel, this kind of honing process is central to the creative act. Moreover, the ability to hone an idea may have little to do with the ability (or patience) to engage in a futile exercise like coming up with uses for a brick, or things that are red. A refinement on conventional measures of divergent thinking, in which participants indicate what they think are their two most creative answers, and these answers are rated on a 5-point scale, shows good reliability and high predictive validity without the fluency confound (Silvia et al, 2008). However, one could still score highly on this version of the test without having engaged in honing.

Our conception of convergent and divergent thinking may be distorted by our everyday experience in the physical world; because objects in the world exist in different places and have distinct, definite boundaries, it may be difficult to wean ourselves from the intuition that ideas in the mind do as well. It has been argued on the basis of evidence from research on the attributes of associative memory, that the common assumption that creativity involves searching through a space of discrete, separate possibilities, selecting the best, and tweaking it, is misleading (Gabora, 2007, 2010, 2018). This is also what is suggested by research on the formal structure of concepts and their interactions. The goal of the rest of this paper is to, without going into mathematical details, show how this research on concepts points to a new conception of convergent and divergent thinking that resolves the above inconsistencies, and potentially catalyzes a deeper understanding of how the creative process works.

The approach to concepts that I will draw upon is sometimes (somewhat unfortunately) referred to as the *quantum approach* (Aerts, Gabora, & Sozzo, 2013; Aerts & Gabora, 2005; Blutner, Pothos, & Bruza, 2013; Busemeyer & Bruza, 2012; Busemeyer & Wang, 2018;; Gabora, 2001; Gabora & Aerts, 2002; Pothos, Busemeyer, Shiffrin, & Yearsley, 2017). It is called this not because it has anything to do with quantum particles, but because it uses generalizations of mathematical structures originally developed for quantum mechanics. The motivation and rationale for this approach are provided elsewhere (Aerts, Broekaert, Gabora, & Sozzo, 2016b; Bruza, Busemeyer, & Gabora, 2009). For now it is noted that this research by no means aims to reduce cognitive psychology to physics. Rather, much as was the case with other branches of mathematics such as complexity theory and even number theory, structures originally developed by physicists were later found to have applications in other domains. In the quantum approach, concepts are viewed not as fixed representations or identifiers, but as bridges between mind and world that are sensitive to context and that actively participate in the generation of meaning (Gabora, Rosch, & Aerts, 2008).

Potentiality, Context, and Creative Thought

The gist of the new view of creative thought suggested by concepts research is conveyed by the photograph below of a woodcutting with light shining on it from three different directions, yielding three differently shaped shadows: that of a G, an E, and a B (Figure 1). Though each shadow is different, they are all projections of the same underlying object. We could say that the woodcutting has the *potentiality* to *actualize* different ways, and to actualize in one of these ways requires an *observable* or *context,* in this

[2] This word is chosen deliberately because it implies that the process is neither entirely top-down nor entirely bottom-up.

case, light shining from a particular direction. We can refer to the state of the woodcutting when no light is shining on it as its *ground state*. While it is tempting to assume that a bout of creative thought entails the generation of multiple distinct, separate ideas, there may be a single underlying mental representation that, like the woodcuttings, is ill-defined, and affords some degree of ambiguity in its interpretation. Just because the different sketches of a painting, or prototypes of an invention, take different forms when expressed in the physical world, that doesn't mean they derive from different underlying ideas in the mind. Just as the three shadows of each of the two woodcuttings in Figure 1 are projections of the same underlying object, the sketches or prototypes may be different external realizations of the same underlying idea at different stages of a creative honing process. In other words, these different outputs are different articulations of the idea as it appears looked at from different perspectives. Midway through a creative thought process one may have an inkling of an idea but not yet know whether, or exactly how, it could work. Because it is 'half-baked', it may be more vulnerable to interpretation, meaning that it could appear quite different when looked at from a different perspective.

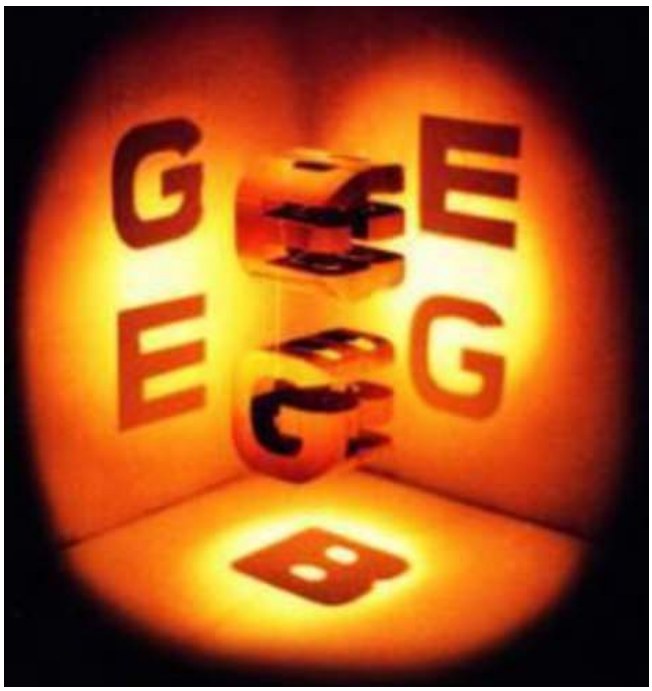


Figure 1: Photograph of ambiguous woodcuttings taken from the front cover of '*Gödel, Escher, Bach: An Eternal Golden Braid' by* Douglas Hofstadter (1979). The top 'triplet' (as he calls them) is not simply a rotated version of the one below it; it is a different shape. (Used with permission.)

Note that the two woodcuttings in Figure 1 have two different shapes, yet they yield the same three shadows. To distinguish the shape of the woodcutting above from the woodcutting below would require that light be shown on them from still more angles, casting shadows that would not look like any particular letters we know. Similarly, the more complex one's unborn creative idea, the more honing steps required to discern its underlying form and whittle it down as needed. Since it has the potential to manifest different ways, we can say that it is a *state of potentiality*.

In the quantum approach to concepts, this kind of potentiality is described as a *superposition state* represented by a vector in a complex Hilbert space. Concepts act as *contexts* for each other that alter how they are experienced; for example, the concept TREE might make you think of a deciduous tree (one with leaves), but in the context CHRISTMAS, you might think of a coniferous tree (one with needles and cones). Each possible context may actualize the potentiality of the concept differently, and these possible actualizations are represented by *basis states*. The actual, existing context is treated as an *observable* that determines how the concept changes in light of this context. It might change in such a way as to alter the weights of certain properties. (For example, 'talks' and 'lives in a cage' are not considered properties of BIRD but they are considered properties of PET BIRD (Hampton, 1987); thus, the context PET is influencing the properties we ascribe to BIRD.) A context can also alter the typicalities of certain exemplars. (As a canonical example, *guppy* is not considered a typical exemplar of PET, nor of FISH, but it is considered a typical exemplar of PET FISH (Osherson & Smith, 1981).) In the absence of any observable—i.e., when a concept is not being viewed from any particular context, or thought about at all—the concept is said to be in a *ground state*. In its ground state there are no properties associated with the concept, but also, there are no properties that are, a priori, excluded from it; thus, you could say it is a state of infinite potentiality. Conceptual change due to the impact of a context is modeled as *collapse* of the vector representing the concept to one of its basis state, as shown in Figure 2.

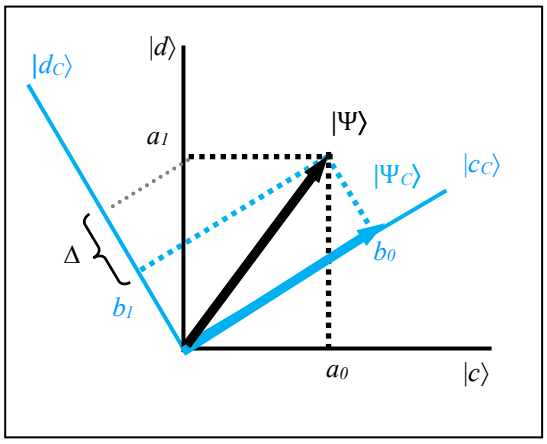


Figure 2: A graphical depiction of a vector |Ψ⟩ representing the concept TREE is shown in black. In the default context, TREE may be more likely to collapse to projection vector |*d*⟩ which represents DECIDUOUS TREE (tree with leaves) than to collapse to projection vector |*c*⟩ which represents

CONIFEROUS TREE (tree with needles and cones). This can be seen by the fact that subspace a_0 is smaller than subspace a_1; i.e., a_0 is closer to the xy origin than a_1. In the context CHRISTMAS, shown in blue, the concept TREE is likely to collapse to the orthogonal projection vector $|c_C\rangle$, representing CONIFEROUS TREE, as shown by the fact that b_0 is larger than b_1. (After collapse, the projected vector, $|\Psi_C\rangle$, is the same length as the original due to renormalization).

This approach has enabled us to cope with some of the non-compositional ways in which people use concepts—famously said to be the biggest challenge facing cognitive science (Fodor, 1998)—by describing them in terms of effects such as entanglement[3] and interference[4] (Aerts, Sozzo, & Gabora, 2016; Aerts & Sozzo, 2014; Busemeyer & Bruza, 2012). The approach can be applied to concept combinations and more complex compounds of concepts such as decisions (e.g., Busemeyer, Wang, & Townsend, 2006; Yukalov & Sornette 2009) jokes (Gabora & Kitto, 2017), worldviews (Gabora & Aerts, 2009), and creativity (e.g., Gabora & Carbert, 2015). For example, working with data from a study in which participants were asked to rate the typicality of exemplars of a concept for different contexts, and introducing a state-transition threshold, we built a model of how exemplars of a concept arise in divergent versus convergent modes of thought (Veloz, Gabora, Eyjolfson, & Aerts, 2011). By lowering a threshold of allowable deviation from the default context, seemingly atypical exemplars appeared as new possibilities. Honing an idea can be modeled as reiterated collapse, resulting in a change of state of the idea, which induces the conceptual framework to subject the idea to a new context, which in turn brings about a new collapse, and so forth, until the idea is sufficiently robust in the face of new contexts that it no longer undergoes change-of-state (Gabora, 2017). In short, it is becoming possible to move beyond crude conceptions of creative cognition to a more refined understanding that is aligned with and informed by advances in the adjacent area of concepts research.

Redefining Convergent and Divergent Thought

Let us now see how this can pave the way to a new conception of convergent and divergent thought. There is a relationship between the weights on the properties of a concept in a particular state, and its susceptibility to collapse to any particular new state. For example, if you think about TREE in terms of only its most typical properties such as 'grows in the ground', your next thought may be about something else that grows in the ground, such as a FLOWER. However, if you think about TREE in a way that encompasses not just typical properties such as 'grows in the ground' but also atypical properties, and in particular those implied by the context, your next thought may be about something semantically distant from TREE; for example, a poet might think of a word that rhymes with TREE such as BEE. Recall how, in its ground state, there are no properties associated with a concept, but also, no properties excluded from it. This means that, for any concept there exists *some* context that could come along and make any given property become relevant. The more exotic the context, the more atypical the properties that are evoked, and thus the more unconventional the subsequent thought.

This suggests that in convergent thought an idea is refined by considering compound of concepts in their *conventional contexts*. Because one is not concerned about all the remote ways in which the object of thought could be related to other things, but instead working with it in its most compact form, mental energy is left over for complex operations. This then is why convergent thought is conducive to unearthing relationships of causation, or thinking analytically, as well as simply carrying out rote tasks.

Conversely, in divergent thought one reflects on an idea by considering a particular compound of concepts from *unconventional contexts*. This is conducive to unearthing relationships of correlation, i.e., forging new connections between seemingly unrelated areas, as in analogical thinking. Note that the more unconventional the contexts one calls up, the seemingly less sensible the next thought may be, and therefore the more honing that may be required to coax it into a form that eventually makes sense. It is for this reason that the products of divergent thought (as redefined here to mean thinking of ideas from unconventional contexts) may require extensive honing.

Implications for Assessment

On the basis of this view of convergent and divergent thinking, let us now re-examine the tests used to assess these constructs. Although the RAT (Mednick, 1968) is used to assess convergent thinking because each question only has one correct answer, to determine the common associate of TANK, TABLE, and HILL you have to think of at least one of these words in a context that is not its default context. For example, unless you are a retailer in the business of selling tank tops you likely interpret the word TANK in terms of its meaning as a military vehicle. Therefore, if we go with the redefinition of convergent thinking as mental operations wherein the contents of thought are viewed from conventional contexts, convergent thought is insufficient to solve the RAT. The RAT is actually more appropriately used as a test of *divergent* thinking. This is consistent with the RAT's wide usage as a test of creativity despite that convergent thought is contrasted with divergent thought and divergent thought is frequently equated with creativity.

Since in divergent thinking tasks such as the Alternate Uses task people only reflect upon an idea from

[3] Entanglement is a phenomenon first encountered in particle physics wherein the state of one entity cannot be described independently of the state of another, and any measurement performed on one influences the other.

[4] Interference is the annihilation of the crest of one wave by the trough of another when they interact.

unconventional contexts *after* they have generated conventional responses, these tasks only test for divergent thinking during the latter part of the task. Thus it makes sense that, as noted by Beaty and Silvia (2012), this is when the most creative responses occur.

Neither the RAT nor conventional divergent thinking tests assess the capacity to hone an idea in a reiterated manner such that uncertainty decreases to an acceptable level and the idea transitions from ill-defined to well-defined. Amobile's (1982) consensual assessment technique, which involves asking multiple experts to evaluate the creativity of a work, is better in this regard, but it undoubtedly measures not just divergent thinking but what is sometimes called *contextual focus:* the capacity to spontaneously shift between convergent and divergent thought as needed, in response to the situation one is in (Gabora, 2010). What is required is a new approach to creativity testing in which each step in a creative process is broken down into a series of states and contexts, and the type and magnitude of conceptual change from one step to the next are analyzed so as to better understand the interplay of convergent and divergent thinking. Steps in this direction are underway using studies of artmaking (Choi & DiPaola, 2013) and computational models (Bell & Gabora, 2016; DiPaola, 2017; DiPaola, Gabora, & McCaig, 2018; McCaig, DiPaola, & Gabora, 2016), as well as technologies such as functional magnetic resonance imaging (Jung & Vartanian, 2018).

Conclusions

The constructs of convergent and divergent thought have been around for half a century, and the way they are defined and measured has changed little in that time. Meanwhile, we have made headway in understanding the dynamics of the compounds of concepts that constitute ideas, and in modeling how they interact as one thought gives way to the next. Given the presence of inconsistences in how convergent and divergent thought are conventionally defined and measured, it seems appropriate to revise our understanding of them in light of recent advances in understanding the internal workings of these processes. This paper has shown how formal research on concepts can pave the way to a new way of defining, measuring, and thinking about convergent and divergent thought.

Note that this is not the only potentially fruitful avenue for research yielded by a joining of forces between research on concepts and research on creativity. For example, there are hints that the above-mentioned presence of interference and entanglement effects in empirical studies of conceptual change are related to creativity, but to date this has not been systematically explored. Another direction for future research concerns the role of *incubation:* the idea that setting a creative task aside for a while, or incubating on it, can promote insight. One could model this as letting the idea return to its ground state such that it sheds its coterie of typical properties (and contexts), and taps into its reservoir of infinite potentiality (in the sense that no properties are definitively present nor absent).

Another intriguing prospect this line of inquiry leads to is the following. Creative people are more subject to adoration, as well as social disapproval and even bullying, and it is generally assumed that this is because they violate social norms (Sternberg & Lubart, 1995). However, this may not be the whole story. I have suggested that the creative mind is in the process of honing ambiguous mental forms, and indeed it has long been thought that creative people are particularly comfortable with ambiguity (e.g., Tegano, 1990; but see also, Merrotsy, 2013). This may include ambiguity with respect to how they themselves come across, which in turn may make them more vulnerable to other people's projections. In other words, they may be more subject to misinterpretation, appearing as Gods or Goddesses to some, and as devils to others.

It is hoped that this paper has provided a glimpse of how formal models of concepts can play a key role in the development of a 21st Century understanding of this most human of abilities, the ability to create.

Acknowledgments

This work was supported by a grant (62R06523) to the author from the Natural Sciences and Engineering Research Council of Canada.

References

Aerts, D., Broekaert, J., Gabora, L., & Sozzo, S. (2016a). Generalizing prototype theory: A formal quantum framework. *Frontiers in Psychology (Cognition)*, *7*(577).

Aerts, D., Broekaert, J., Gabora, L., & Sozzo, S. (2016b). Quantum structures in cognitive and social science. *Frontiers in Psychology* (Section: Cognition), *7*(577). doi: 10.3389/fpsyg.2016.00577

Aerts, D., & Gabora, L. (2005). A theory of concepts and their combinations II: A Hilbert space representation. *Kybernetes*, *34*(1/2), 192–221.

Aerts, D., Gabora, L., & Sozzo, S. (2013). Concepts and their dynamics: A quantum-theoretic modeling of human thought. *Topics in Cognitive Science*, *5*, 737–772.

Aerts, D., & Sozzo, S. (2014). Quantum entanglement in concept combinations. *International Journal of Theoretical Physics*, *53*, 3587–3603.

Bak, P., Tang, C., & Weisenfeld, K. (1988). Self-organized criticality. *Physical Review A, 38,* 364.

Beaty R. E. & Silvia P. J. (2012). Why do ideas get more creative across time? An executive interpretation of the serial order effect in divergent thinking tasks. *Psychology of Aesthetics, Creativity, and the Arts*, *6,* 309–319.

Beersma, B., & De Dreu, C. K. (2005). Conflict's consequences: Effects of social motives on post-negotiation creative and convergent group functioning and performance. *Journal of Personality and Social Psychology*, *89*, 358–374.

Bell, S. & Gabora, L. (2016). A music-generating system inspired by the science of complex adaptive systems. In *Proceedings of the fourth international workshop on musical meta-creation*. Palo Alto: AAAI Press.

Blutner, R., Pothos, E. M., & Bruza, P. (2013). A quantum probability perspective on borderline vagueness. *Topics in Cognitive Science*, *5*, 711–736.

Bruza, P., Busemeyer, J., & Gabora, L. (2009). Introduction to the special issue on quantum cognition. *Journal of Mathematical Psychology, 53,* 303–305.

Busemeyer, J. R., & Bruza, P. D. (2012). *Quantum models of cognition and decision*. Cambridge UK: Cambridge University Press.

Busemeyer, J., & Wang, Z. (2018). Hilbert space multidimensional theory. *Psychological Review*, *125*, 572–591.

Amabile, T. M. (1982). Social psychology of creativity: A consensual assessment technique. *Journal of Personality and Social Psychology*, *43*, 997–1013.

Busemeyer J. R., Wang Z., Townsend J. T. (2006). Quantum dynamics of human decision making. *Journal of Mathematical Psychology, 50,* 220–41.

Christensen, P. R., Guilford, J. P., Merrifield, P. R., Wilson, R.C. (1960). *Alternate uses.* Sheridan.

Choi, S., & DiPaola, S. (2013). How a painter paints: An interdisciplinary understanding of embodied creativity. *Proceedings of electronic visualization and the arts* (pp. 127–134). London: British Computer Society.

Chrusch, C. & Gabora, L. (2014). A tentative role for FOXP2 in the evolution of dual processing modes and generative abilities. In *Proceedings of the 36th annual meeting of the cognitive science society* (pp. 499–504). Austin TX: Cognitive Science Society.

DiPaola, S. (2017). Exploring the cognitive correlates of artistic practice using a parameterized non-photorealistic toolkit. *Leonardo 50,* 531–452.

DiPaola, S., & Gabora, L. & McCaig, G. (2018). Informing artificial intelligence generative techniques using cognitive theories of human creativity. *Procedia Computer Science, 145,* 158–168.

Evans, J., & Frankish, K. (Eds.) (2009). *In two minds: Dual processes and beyond.* Oxford UK: Oxford University Press.

Fodor, J. A. (1998). *Concepts: Where cognitive science went wrong*. Oxford UK: Oxford University Press.

Gabora, L. (1999). Weaving, bending, patching, mending the fabric of reality: A cognitive science perspective on worldview inconsistency. *Foundations of Science*, *3*, 395–428.

Gabora, L. (2001). *Cognitive mechanisms underlying the origin and evolution of culture.* Ph.D. Thesis, Free University of Brussels, Belgium.

Gabora, L. (2007). Why the creative process is not Darwinian. *Creativity Research Journal, 19*, 361–365.

Gabora, L. (2010). Revenge of the "neurds": Characterizing creative thought in terms of the structure and dynamics of memory. *Creativity Research Journal*, *22*, 1–13.

Gabora, L. (2017). Honing theory: A complex systems framework for creativity. *Nonlinear Dynamics, Psychology, and Life Sciences*, *21*, 35–88.

Gabora, L. (2018). The neural basis and evolution of divergent and convergent thought. In O. Vartanian & R. Jung (Eds.) *The Cambridge handbook of the neuroscience of creativity*. Cambridge: Cambridge University Press.

Gabora, L., & Aerts, D. (2002). Contextualizing concepts using a mathematical generalization of the quantum formalism. *Journal of Experimental & Theoretical Artificial Intelligence*, *14*, 327–358.

Gabora, L, & Aerts, D. (2009). A model of the emergence and evolution of integrated worldviews. *Journal of Mathematical Psychology*, *53*, 434–451.

Gabora, L., & Carbert, N. (2015). Cross-domain influences on creative innovation: Preliminary Investigations. *Proceedings annual meeting cognitive science society* (pp. 758–763). Austin: Cognitive Science Society.

Gabora, L., & Kitto, K. (2017). Toward a quantum theory of humour. *Frontiers in Physics* (Interdisciplinary Physics), *4*(53).

Gabora, L., Rosch, E., & Aerts, D. (2008). Toward an ecological theory of concepts. *Ecological Psychology, 20*(1), 84-116.

Gibson, C., Folley, B. S., & Park, S. (2009). Enhanced divergent thinking and creativity in musicians: A behavioral and near-infrared spectroscopy study. *Brain and Cognition*, *69,* 162–169.

Guilford, J. (1967). *The nature of human intelligence*. New York: Routledge.

Hampton, J. (1987). Inheritance of attributes in natural concept conjunctions. *Memory & Cognition*, *15*, 55–71.

Hirsh, J. B., Mar, R., & Peterson, J. (2012). Psychological entropy: A framework for understanding uncertainty-related anxiety. *Psychological Review, 119*, 304–320.

Hofstadter, D. R. (1979). *Gödel, Escher, Bach*. New York: Basic Books.

James, W. (1890/1950). *The principles of psychology*. New York: Dover.

Kerr, D. S., & Murthy, U. S. (2004). Divergent and convergent idea generation in teams: A comparison of computer-mediated and face-to-face communication. *Group Decision and Negotiation*, *13,* 381–399.

McCaig, G, DiPaola, S., & Gabora, L. (2016). Deep convolutional networks as models of generalization and blending within visual creativity. *Proceedings of the seventh international conference on computational creativity* (pp. 156-163). Palo Alto: AAAI Press.

Mednick S. A. (1968). The remote associates test. *Journal of Creative Behavior, 2,* 213–214.

Merrotsy, P. (2013). Tolerance of ambiguity: a trait of the creative personality? *Creativity Research Journal*, *25*, 232–237.

Neisser, U. (1963). The multiplicity of thought. *British Journal of Psychology, 54,* 1–14.

Osherson, D., & Smith, E. (1981). On the adequacy of prototype theory as a theory of concepts. *Cognition*, *9*, 35–58.

Piffer, D. (2012). Can creativity be measured? An attempt to clarify the notion of creativity and general directions for future research. *Thinking Skills and Creativity*, *7*, 258–264.

Plucker, J., & Renzulli, J. (1999). Psychometric approaches to the study of human creativity. In R. Sternberg (Ed.), *Handbook of creativity*. Cambridge UK: Cambridge University Press.

Pothos, E., Busemeyer, J., Shiffrin, R., & Yearsley, J. (2017). The rational status of quantum cognition. *Journal of Experimental Psychology: General*, *146*, 968–987.

Runco, M. (2014). *Creativity theories and themes: Research, development, and practice.* Amsterdam: Elsevier.

Schilling, M. (2005). A small-world network model of cognitive insight. *Creativity Research Journal*, *17*, 131–154.

Silvia, P., Winterstein, B., Willse, J., Barona, C., Cram, J., Hess, K... & Richard, C. (2008). Assessing creativity with divergent thinking tasks: Exploring the reliability and validity of new subjective scoring methods. *Psychology of Aesthetics, Creativity, and the Arts*, *2*, 68–85.

Sloman, S. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin, 9,* 3–22.

Sowden, P., Pringle, A., & Gabora, L. (2015). The shifting sands of creative thinking: Connections to dual process theory. *Thinking & Reasoning, 21*, 40–60.

Sternberg, R., & Lubart, T. (1995). *Defying the crowd: Cultivating creativity in a culture of conformity*. New York: Free Press.

Tegano, D. W. (1990). Relationship of tolerance of ambiguity and playfulness to creativity. *Psychological Reports*, *66*, 1047–1056.

Vartanian, O. & R. Jung (Eds.) (2018). *The Cambridge Handbook of the Neuroscience of Creativity.* (pp. 58–70). Cambridge UK: Cambridge University Press.

Veloz, T., Gabora, L., Eyjolfson, M., & Aerts, D. (2011). Toward a formal model of the shifting relationship between concepts and contexts during associative thought. In *Proceedings Fifth International Symposium on Quantum Interaction* (pp. 25–34). Berlin: Springer.

Yukalov V., & Sornette D. (2009). Processing information in quantum decision theory. *Entropy*, *11*, 1073–120.

From Deep Learning to Deep Reflection: Toward an Appreciation of the Integrated Nature of Cognition and a Viable Theoretical Framework for Cultural Evolution

Liane Gabora (liane.gabora@ubc.ca)
Department of Psychology, University of British Columbia, Kelowna BC, V1V 1V7, CANADA

Abstract

Although Darwinian models are rampant in the social sciences, social scientists do not face the problem that motivated Darwin's theory of natural selection: the problem of explaining how lineages evolve despite that any traits they acquire are regularly discarded at the end of the lifetime of the individuals that acquired them. While the rationale for framing culture as an evolutionary process is correct, it does not follow that culture is a Darwinian or selectionist process, or that population genetics provides viable starting points for modeling cultural change. This paper lays out step-by-step arguments as to why a selectionist approach to cultural evolution is inappropriate, focusing on the lack of randomness, and lack of a self-assembly code. It summarizes an alternative evolutionary approach to culture: self-other reorganization via context-driven actualization of potential.

Keywords: acquired trait; cultural evolution; inheritance; natural selection; population genetics; self-other re-organization

Introduction

Though several of the deepest evolutionary thinkers of the 20th Century cautioned against the over-zealous application of Darwinian theory (Claidière, Scott-Phillips, & Sperber, 2014; Fracchia & Lewontin, 1999; Mayr, 1996; Tëmkin & Eldredge, 2007), Darwinian models are rampant in the social sciences. The frameworks of population genetics has been applied to cultural evolution (Boyd & Richerson, 1988; Brewer et al., 2017; Cavalli-Sforza & Feldman, 1981; Creanza, Kolodny, & Feldman, 2017; Henrich et al., 2016), as well as to archaeology (O'Brien & Lyman, 2000), economics (Essletzbichler, 2011; Hodgson, 2002; Nelson & Winter, 2002), neuroscience (Edelman, 2014), the evolution of languages (Fitch, 2005; Pagel, 2017), and the unfolding of a creative idea in the mind of an individual (Campbell, 1960; Kronfeldner, 2014; Simonton, 1999; for counter-arguments see Gabora, 2007). This paper focuses exclusively on the question of whether cultural evolution is Darwinian. This is a different project from that of examining how natural selection has shaped the propensity for culture, language, artifacts, and so forth; it models cultural change itself as a second Darwinian process.

The rationale is that since cultural forms, like biological forms, evolve, i.e., exhibit cumulative, adaptive, open-ended change, culture constitutes a second evolutionary process. This is undoubtedly true. However, cultural Darwinism goes further than the claim that culture evolves; it assumes that the formal framework of population genetics, with appropriate tinkering to accommodate culture-specific phenomena, provides a viable foundation for modeling this second evolutionary process.

Many have laid out the similarities and differences between biological and cultural evolution (Godfrey-Smith, 2012; Jablonka & Lamb, 2014; Mesoudi, 2007; Wagner & Rosen, 2014). The issue addressed here is not how similar they are, but the extent to which the algorithmic structure of cultural evolution merits importation of a Darwinian framework. This paper lays out two arguments against this project, breaking them down step by step so as to facilitate the identification and settling of any points of disagreement. The first, the weaker argument, pertains to the issue of randomness. The second pertains to the existence of a self-assembly code. We will see that due to limited interaction with cognitive science, cultural evolution research has paid little attention to structure of the human minds that evolve culture, and the processes by which elements of culture take form. This has led to the misapplication of evolutionary concepts to culture, resulting in lack of appreciation of its essentially non-Darwinian character. The paper concludes with a brief discussion of an alternative, non-Darwinian evolutionary framework for culture.

Definitions

It is true that any definition of a term is fine so long as everyone agrees how it is being used. However, part of why it has been difficult to nail down the extent to which cultural forms evolve in the same sense as biological forms is that, in drawing parallels between biological and cultural evolution, existing terms have been stretched beyond their conventional meanings. When they are used in ways that do not capture the deep structure or essence of their original meaning, or when a biological referent is misleadingly retained in a cultural context, misunderstanding can result. The matter is tricky, for although cultural evolution constitutes a separate evolutionary process with its own evolving structures and processes, it is inextricably interwoven with biological evolution[1]. To maintain clarity, key terms used in this paper are defined below:

Acquired trait: a trait obtained during the lifetime of its bearer (e.g., a scar, a tattoo, or a memory of a song) and transmitted horizontally (i.e., laterally).[2]

Culture: extrasomatic adaptations—including behavior and artifacts—that are socially rather than sexually transmitted.

[1] For example, maternal diet during lactation can influence food preferences in offspring (Bilkó et al., 1994).

[2] These are acquired traits with respect to biological evolution. It will be argued that with respect to cultural evolution *all* traits are acquired.

Darwinian process: an evolutionary process that occurs through natural or artificial selection.[3]

Darwinian threshold: transition from non-Darwinian to Darwinian evolutionary process (Woese, 2002; Vetsigian, Woese, & Goldenfeld, 2006).

Evolutionary process: a process that exhibits cumulative, adaptive, open-ended change.

Gene: a component of a self-assembly code, i.e., a unit of heredity.[4]

Generation: a single transition period from the internalized to the externalized form of a trait.

Horizontal transmission: The spread of a trait within a generation.

Inherited trait: a trait (e.g., blood type, or the capacity to suntan) that is transmitted vertically (e.g., from parent to offspring) by way of a self-assembly code (e.g., DNA).

Modern Synthesis: merging of Darwin's theory of natural selection and Mendelian genetics in the 1940s.

Organism: the living expression of a particular self-assembly code, sometimes referred to as an 'individual'.

Phylogenetic network: model of the relationships amongst variants that is pictured as *reticulate,* or 'network-like'.

Phylogenetic tree: model of the relationships amongst different species that is pictured as *branching,* or 'tree-like'.

Population genetics: branch of biology central to which is a mathematical theory of how organisms evolve through natural selection due to changes in gene frequencies. It was originally developed by Fisher (1930), Wright (1931), and Haldane (1932) and subsequently expanded (Hartl & Clark, 2006).

Selection: differential replication of randomly generated heritable variation in a population over generations such that some traits become more prevalent than others. Selection may be *natural* (due to non-human elements of the environment) or *artificial* (due to human efforts such as selective breeding), and it can occur at multiple levels, e.g., genes, individuals, or groups (Lewontin, 1970).

Selectionist process: like the term 'Darwinian process' this refers to an evolutionary process that occurs through natural or artificial selection. (It will be used from this point on because it avoids potential confusion caused by the fact that Darwin considered other possibilities.)

Self-assembly code: a set of self-replication instructions.

Self-other Reorganization (SOR): a theory of how both culture, and early life, evolve through communally exchanging, self-organizing networks that generate new components through their interactions. Based on post-Modern Synthesis theory and findings in biology.

Vertical transmission: The inheritance of a trait from one generation to the next by way of a self-assembly code.

It is important to point out that we are using the term 'selection' in its technical, scientific sense. The word 'selection' also has an 'everyday' sense in which it is synonymous with 'choosing' or 'picking out'. One could say that selection—in the everyday sense of the term—occurs in a competitive marketplace through the winnowing out superior products. However, the discussion here concerns whether selection *in the scientific sense of the term* is applicable to cultural evolution.

Note that, with respect to biological evolution, a new generation (one transmission event) generally (though not in horizontal gene transfer) begins with the birth of an organism. It is not impossible for the same trait to be transmitted horizontally in one generation and vertically in another. However, with respect to cultural evolution, *a new generation begins with the expression of an idea* (again, one transmission event). Thus, over the course of a single discussion, an idea (a cultural trait) can undergo many generations. It can be said that cultural evolution proceeds more quickly than human biological evolution[5], since the lengthy period we associate with biological generations, from birth through development to puberty and reproductive maturity to parenthood, is in general significantly longer than the stretch of time between when an individual acquires a cultural trait (e.g., an idea) and then expresses (their own version of, or their own take on) that cultural trait.

Note also that vertical and horizontal transmission must be defined *with respect to the relevant evolutionary process*. A common error is to refer to the transmission of cultural information from parent to offspring as vertical transmission (e.g., Cavalli-Sforza & Feldman, 1981). The parent-child relationship is *with respect to* biological evolution; they are parent and child with respect to their status as biologically evolving organisms, but with respect to their status as participants in cultural evolution, there is no basis for this parent-child distinction. Indeed, while childbirth entails one mother and one father, there is no limit to the number of 'parental influences' on the 'birth' of an idea.

A related error is to say that in cultural evolution there is a third form of transmission, *oblique transmission*, in which traits are transmitted from non-kin members of the parental generation (e.g., Cavalli-Sforza & Feldman, 1981), for as far as *cultural* evolution is concerned it is irrelevant whether the information comes from *biological* kin or non-kin.

In a similar vein, although *dual inheritance* theorists speak of culture as a second form of inheritance (Henrich & McElreath, 2007; Richerson & Boyd, 1978; Whiten, 2017; Müller, 2017), the distinguishing feature of an inherited trait is that it is transmitted vertically—e.g., from parent to

[3] Although evolution by selection is the process Darwin's name became most synonymous with, it is interesting to note that amongst his many musings was a theory of pangenesis involving transmission of acquired traits (Darwin, 1868).

[4] In biology, the term 'gene' generally refers to a sequence of DNA or RNA nucleotides that code for a molecule with a particular function. It will be argued that, with respect to cultural evolution, there is no self-assembly code, and thus no equivalent to the gene.

[5] We are not referring here to clock time but to the *relative* mean duration of biological versus cultural generation processes.

offspring—by way of a self-assembly code (e.g., DNA), and therefore not obliterated at the end of a generation. This is not the case with respect to cultural traits (Gabora, 2011). (Nor, as we shall see, is it even the case for all biological traits.)

As a final preliminary note, it is important to keep in mind that organisms (including humans) are affected by epigenetic processes that influence the regulation and expression of genetic traits due to interactions with the environment, as well as selection effects operating on groups as well as individuals (Wilson, 1975). For simplicity, this paper does not explore these complications in detail but their relevance to the argument presented here is discussed elsewhere (Voorhees, Read, & Gabora, in press).

Randomness

It is possible for a selectionist model to be applicable even if the underlying process is not random, but in that case, although not genuinely random, the process must be approximated by a random distribution.[6] Biological variation is not genuinely random (for example, we can trace the source of some mutations to various causal agents; see Caporale 2000) but the assumption of randomness generally holds sufficiently well to serve as a useful approximation.[7]

With respect to culture, variation is not randomly generated, nor can it be described by a random distribution. Selectionist cultural theorists sometimes concede this point (Heyes, 2018), but fail to recognize its implications for the assumed validity of a selectionist framework. Natural selection *acts upon* nonrandomly generated variation, but to the extent that variation is *not* randomly generated, the distribution of variants reflects whatever is biasing the generation away from random in the first place, rather than selection (i.e., differential selection on the distribution of randomly generated heritable variation in a population over generations). Let us break this argument down step by step.

1. Natural selection is a two-step process, consisting of (i) generation of random variants that differ in fitness, followed by (ii) differential survival and reproduction of the fittest variants.
2. The first step provides variation upon which selection can operate, and the adaptiveness of the process resides not in the first step (how variants are *generated*) but the second (how fit variants are *selected*).
3. To the extent that variants do not differ in fitness, their evolution is attributed not to selection but to random genetic drift (Fisher, 1930; Hartl & Clark, 2006).[8]
4. To the extent that the generation of variants cannot be described by a random distribution, their evolution is attributed not to selection but to the nature of this nonrandom generation process.
5. Cultural change cannot be approximated by a random distribution; it is orders of magnitude more non-random than biological evolution. It is strategic and creative, with ideas emerging due to spreading activation and overlap amongst distributed mental representations encoded in associative memory (Gabora, 2013).
6. Therefore, a selectionist model is inappropriate to the description of cultural change.

In the cultural Darwinism literature there is much discussion of *social learning* (obtaining *existing* information from someone else), and some mention of *individual learning* (obtaining *existing* information through nonsocial means), but little about creativity, reasoning, planning, problem solving, i.e., the the highly non-random higher cognitive processes that *generate* cultural novelty. In a paper titled "grand challenges for the study of cultural evolution" (Brewer et al., 2017), absent from among the eight challenges is the challenge of studying the creative processes that fuel cultural evolution. The closest they come is to ask "How are innovations selectively transmitted" and "Do innovations create feedback loops leading to cumulative culture?" It seems that understanding how innovations come about in the first place is more fundamental than knowing how they are "selectively transmitted" or whether they create feedback loops. Without the creative generation of cultural novelty, there is no cultural evolution. As demonstrated in an agent-based computational model of cultural evolution (Gabora, 1995); when agents never *imitate,* cultural evolution does occur, albeit slowly, as each agent figures things out on its own, but when agents never *create,* there is no cultural evolution at all. Thus, understanding creativity would appear to be the 'grandest' challenge of all for cultural evolution research.

The 'randomness' argument puts a major dent in the theory that cultural evolution is selectionist, but it does not destroy it altogether. It is possible that after variation has been generated by way of nonrandom processes there might still be work for selection to do in winnowing out the very fittest. However, we now turn to the more serious problem, that in cultural evolution there is no self-assembly code.

Self-assembly Code

In biological evolution there are two kinds of traits: (1) inherited traits (e.g., blood type), transmitted vertically from parent to offspring by way of genes, and (2) acquired traits (e.g., a tattoo), obtained during an organism's lifetime, and transmitted horizontally amongst conspecifics.[9] A selectionist explanation works in biology to the extent that retention of acquired change is negligible compared to

[6] Few things other than radioactive decay are truly random.

[7] Actually, in some biological situations, such as assortative mating, the assumption of randomness does not hold, and in such cases natural selection is not an appropriate model.

[8] Drift has been demonstrated in human culture (Bentley, Hahn, & Shennan, 2004), and in a computational model of cultural evolution (Gabora, 1995).

[9] This is a simplification, for there exist traits that are encoded in genetic material, but this genetic material does not constitute a full-fledged self-assembly code (see Bonduriansky & Day, 2009).

retention of selected change; otherwise the first, which can operate instantaneously, overwhelms the second, which takes generations. Transmission of acquired traits is avoided through use of a *self-assembly code* (such as the genetic code), i.e., a set of instructions for how to reproduce. Because a lineage perpetuates itself using a self-assembly code, inherited traits are transmitted while acquired traits are not.[10]

Now let us turn to culture. In cultural evolution, there is no self-assembly code, and no vertically transmitted inherited traits; all change is acquired.[11] Therefore, cultural evolution is not due to the mechanism Darwin proposed: differential replication of heritable variation in response to selection. The only response to this argument I know of comes from Mesoudi (2007): "[the] point concerning the lack of self-assembly codes in cultural entities is, again, well-taken when compared to many biological organisms, but may not hold if we take viruses as our biological exemplar, which similarly cannot self-replicate in the absence of a host, or … the evolution of early RNA-based life before DNA-based replication mechanisms evolved." This response evades the problem, for the argument is not that cultural evolution differs from biological evolution, but that the assumptions underlying the formal framework developed to describe evolution by natural selection renders it inapplicable to culture. Indeed, it is also inapplicable to the description of some aspects of biological evolution, but that should be more reason for concern, not less.

Thus, to help determine whether there is a genuine flaw in the argument, and if so pinpoint what that flaw is, we again break the argument down into steps:

1. To the extent that an evolutionary process is amenable to a selectionist model, there are two kinds of traits: vertically transmitted inherited traits, and horizontally transmitted acquired traits.
2. Acquired traits are discarded from a lineage at the end of every generation.
3. This means that evolution (i.e., cumulative, open-ended, adaptive change) in biological lineages cannot be explained in terms of acquired traits.
4. Therefore, it is explained in terms of inherited traits.
5. In biological evolution, inherited traits are not discarded; they are preserved by way of a self-assembly code. The code's low-level information-bearing components must be organized in an orderly manner so they can be parsed into meaningful units; otherwise, the precisely orchestrated process by which the code is expressed to generate offspring is disrupted.
6. The population genetics framework was developed to explain change in a system such as this where the slow process of selection for inherited traits over generations is not drowned out by the fast process of acquired change (which can take place over milliseconds).
7. Biological evolution is therefore explainable in terms of differential selection on the distribution of randomly generated heritable variation in a population over generations, i.e., natural selection.
8. Since acquired change operates markedly faster than inherited change, to the extent that acquired change is *not* wiped out at the end of each generation, a population genetics framework is inappropriate as an explanatory model.
9. In cultural evolution, there is no distinction between vertically transmitted inherited traits and horizontally transmitted acquired traits. Since all traits are horizontally transmitted, we may refer to them as *cultural acquired traits*.
10. Cultural acquired traits are *not* regularly discarded from cultural lineages at the end of generations.
11. This means that evolution (i.e., cumulative, open-ended, adaptive change) in cultural lineages *can* be explained in terms of acquired traits.
12. Moreover, culture is not transmitted by way of inherited traits.
13. Therefore, cultural change, unlike biological change, cannot be explained in terms of change in the frequency of inherited traits; there exists no basis upon which to explain cultural evolution in terms of differential selection of inherited traits on the distribution of randomly generated heritable variation in a population over generations, i.e., using a selectionist framework.
14. Cultural evolution must therefore be explained entirely in terms of changes to acquired traits.

This argument has important implications for how cultural data is modeled. Since biological acquired traits are usually (though not always) discarded, and since a self-assembly code must stay intact to preserve its self-replication capacity, the joining of bifurcations in biological lineages is rare; thus, a phylogenetic tree correctly captures the branching structure. However, since cultural acquired traits are not discarded, and there is no cultural self-assembly code, the joining of bifurcations in cultural

[10] An organism may bypass the disappearance of acquired traits through niche construction, *i.e.,* by modifying its environment in such a way as to influence the behavior (and potentially gene regulation) of offspring. Thus, by 'building acquired traits into the environment', one generation may influence the traits exhibited by the next (Lewontin, 1998). However, acquired change is sufficiently negligible relative to inherited change that a selectionist explanation is still of value in explaining biological evolution.

[11] An anonymous reviewer suggested natural language is a cultural self-assembly code. However, (1) natural language is not a set of encoded instructions for the self-replication of natural languages, and (2) culture does not exhibit the signature characteristics of evolution by way of a self-assembly code: lack of transmission of acquired traits, and culture is characterized by horizontal not vertical transmission. Nevertheless, culture may be moving *toward* a cultural Darwinian threshold. In other words, it may exist in the state biological life was in before LUCA (last universal common ancestor) (Woese, 1998).

'lineages' is commonplace, and thus the structure is network-like rather than the tree-like (Gabora, 2006b). This difference has been demonstrated mathematically using split-decomposition graphs (Bandelt & Dress, 1992; Wägele, 2005). Dress and colleagues showed that while biological data generate branching graphs, reanalysis of data from a psychological experiment in which people were asked to estimate the subjective distance between colours gives a very different structure (Dress, Huson, & Moulton, 1996). This difference in the deep structure of biological data and cultural data such as languages, concepts, and artifacts arising from human cognition, is why phylogenetic tree models of culture are problematic.

Self-Other Reorganization (SOR): An Alternative Approach to Cultural Evolution

The above analysis precludes a *selectionist* but not an *evolutionary* framework for culture. Indeed, research since the Modern Synthesis has shown that even life itself is only partially explained through recourse to a selectionist framework; for example, though biological traits are generally obtained through vertical inheritance, horizontal gene transfer (HGT) involves horizontal transmission (Ochman et al., 2000). Evolution can occur in the absence of selection, and the importance of non-selectionist processes in evolution is increasingly recognized (Kauffman, 1993; Killeen, 2017; Woese, 2002). Research on the origin of life suggests that early life consisted of autocatalytic protocells that evolved through a non-selectionist means, and natural selection emerged later from this more haphazard, ancestral evolutionary process (Baum, 2018; Cornish-Bowden & Cárdenas, 2017; Gabora, 2006; Goldenfeld, Biancalani, & Jafarpour, 2017; Hordijk, Steel, & Dittrich, 2018; Steel, 2000; Vetsigian, Woese, & Goldenfeld, 2006). This non-selectionist process requires (1) a *self-organizing network* of components that generate new components through their interactions, (2) the network should be able to reconstitute another like itself through haphazard (not code-driven) duplication of components, and (3) interaction amongst networks. This process can be referred to as *Self-Other Reorganization* (SOR) because it involves an interplay between self-organized *internal* restructuring, and communal exchange *amongst* autocatalytic structures. Change occurs not through selection but through a process that has a completely different mathematical description: context-driven actualization of potential (Gabora & Aerts, 2005). The entity changes through interactions with its world, which in turn alters its potential for future configurations. Like selectionist evolution, SOR has mechanisms for preserving continuity and for introducing novelty, but unlike selectionist evolution, it is a low-fidelity Lamarckian process. The distinction between these two processes is summarized in Table One.

Vetsigian et al. (2006) showed that to cross the Darwinian threshold from non-selectionist to selectionist evolution requires the emergence of a self-assembly code. There is no evidence that culture has crossed this threshold, and it does not possess the *sine qua non* of having crossed it: vertical transmission and lack of transmission of acquired traits. It has been proposed that, as did early life, culture evolves through SOR (sometimes referred to as 'communal exchange') (Gabora, 1999, 2004, 2019). Here, the self-organizing networks are not protocells exchanging catalytic molecules, but minds exchanging ideas. Tools improve and fashions change not through selection but through context-driven actualization of potential (Gabora & Aerts, 2005). As parents share knowledge with children, an integrated network of understandings takes shape in their minds, and they become creative contributors to cultural evolution.

There are other network-based approaches (e.g., Bentley & Shennan, 2003), and non-selectionist models of cultural evolution, e.g., those based on the Price equation (e.g., El Mouden, André, Morin, & Nettle, 2014). SOR's uniqueness lies in its emphasis on the emergence of a network of understandings that is reinforced and expanded upon through its ongoing use as a scaffold to interpret and reflect upon new information; in other words, that crosses the threshold from 'deep learning' to 'deep reflection'. It has been noted that a tension exists between cultural evolution theory and the literature on human nature (Lewens, 2017). Because SOR is not incompatible with transmission of acquired traits, and because it recognizes the integrated, 'self-mending' nature of a mind, SOR provides a natural means of reconciling cultural evolution and human nature.

Table One: Comparison between evolution through selection and evolution through Self-Other Reorganization.

	Selection	**Self-other Re-organization (SOR)**
Unit of self-replication	Organism	Self-organizing autocatalytic network
Mechanism for preserving continuity	Reproduction (vertical transmission), proofreading enzymes, etc.	Communal exchange (horizontal transmission)
Generation of novelty	Mutation, recombination	Creativity, transmission error
Self-assembly code	DNA or RNA	None
High fidelity	Yes	No
Transmission of acquired traits	No	Yes
Type	Selectionist	Lamarckian (by some standards)
Evolution processes it can explain	Biological	Early life; horizontal gene transfer, culture

Conclusions

Darwin faced the problem of explaining how lineages evolve despite that acquired changes are lost from a lineage when the individuals that acquired them dies. Darwin's

solution was to come up with a population-level (macro) explanation. His theory of natural selection holds that although *acquired traits* are discarded, *inherited traits* are retained, so evolution can be explained in terms of preferential selection for those inherited traits that confer fitness benefits on their bearers. Cultural evolution research does not face the problem that motivated Darwin's solution—that of explaining how evolution takes place despite the discarding of acquired traits—because cultural acquired traits are *not* discarded. Thus, while the rationale for framing culture as an *evolutionary* process is correct, it does not follow that culture is a selectionist process, or that population genetics provides a viable starting point for modeling cultural change. Cultural evolution research has been carried out largely independent of research on the mental structures that actually evolve culture. This has led to the mis-application of biological constructs such as generations, inheritance, and vertical / horizontal transmission. This in turn has hindered appreciation of the dependence of vertical inheritance on a self-assembly code, and recognition of the implications of its absence in cultural evolution. The field needs cognitive scientists to uncover the cognitive processes by which culture actually takes shape.

Psychologists use the term mental set to refer to the persistent use of problem-solving strategies that worked in the past even when these strategies are not appropriate to the problem at hand. It appears that the persistent application of a selectionist framework to cultural evolution, despite that the conditions that make that framework applicable in biology are not present with respect to culture, may be an instance of mental set. This paper has laid out step-by-step arguments as to why a selectionist approach to culture is inappropriate, and pointed to an alternative approach.

Acknowledgments

This research was supported by grant 62R06523 from the Natural Sciences and Engineering Research Council of Canada.

References

Bandelt, H., & Dress, A. (1992). Split decomposition: A new and useful approach to phylogenetic analysis of distance data. *Molecular Phylogenetics and Evolution*, *1*, 242-252.

Baum, D. (2018). The origin and early evolution of life in chemical composition space. *Journal of Theoretical Biology*, *456*, 295-304.

Bentley, R., Hahn, M. W., & Shennan, S. J. (2004). Random drift and culture change. *Proceedings of the Royal Society B: Biological Sciences*, *271*, 1443-1450.

Bentley, R., & Shennan, S. (2003). Cultural transmission and stochastic network growth. *American Antiquity*, *68*, 459-485.

Bilkó, Á., Altbäcker, V., & Hudson, R. (1994). Transmission of food preference in the rabbit: the means of information transfer. *Physiology & Behavior*, *56*, 907-912.

Bonduriansky, R., & Day, T. (2009). Nongenetic inheritance and its evolutionary implications. *Annual Review of Ecology, Evolution, and Systematics*, *40*, 103-125.

Boyd, R., & Richerson, P. (1988). *Culture and the evolutionary process*. Chicago: University of Chicago Press.

Brewer, J., Gelfand, M., Jackson, J. C., MacDonald, I. F., Peregrine, P. N., Richerson, P. J., … Wilson, D. S. (2017). Grand challenges for the study of cultural evolution. *Nature Ecology & Evolution*, *1*, 10-1038.

Campbell, D. (1960). Blind variation and selective retention in creative thought as in other knowledge processes. *Psychological Review*, *67*, 380-400.

Caporale, L. (2000). Mutation is modulated: Implications for evolution. *BioEssays, 22,* 388-395.

Cavalli-Sforza, L. L., & Feldman, M. W. (1981). *Cultural Transmission and evolution: A quantitative approach*. Princeton NJ: Princeton University Press.

Claidière, N., Scott-Phillips, T. C., & Sperber, D. (2014). How Darwinian is cultural evolution? *Philosophical Transactions of the Royal Society B*, *369*, 20130368.

Cornish-Bowden, A., & Cárdenas, M. L. (2017). Life before LUCA. *Journal of Theoretical Biology*, *434*, 68-74.

Creanza, N., Kolodny, O., & Feldman, M. W. (2017). Cultural evolutionary theory: How culture evolves and why it matters. *Proceedings of the National Academy of Sciences*, *114*, 7782-7789.

Darwin, C. (1868). *The variation of animals and plants under domestication*. London: John Murray.

Dress, A., Huson, D., & Moulton, V. (1996). Analyzing and visualizing sequence and distance data using SplitsTree. *Discrete Applied Mathematics*, *71*, 95-109.

Edelman, G. (2014). Neural Darwinism. *New Perspectives Quarterly*, *31*, 25-27.

El Mouden, C., André, J. B., Morin, O., & Nettle, D. (2014). Cultural transmission and the evolution of human behaviour: A general approach based on the Price equation. *Journal of Evolutionary Biology*, *27*, 231-241.

Fisher, R. A. (1930). *The genetical theory of natural selection*. Oxford: Clarendon Press.

Fitch, W. (2005). The evolution of language: A comparative review. *Biology and Philosophy*, *20*, 193-203.

Fracchia, J., & Lewontin, R. C. (1999). Does culture evolve? *History and Theory*, *38*, 52-78.

Gabora, L. (1995). Meme and variations: A computational model of cultural evolution. In *1993 Lectures in complex systems*. Addison Wesley.

Gabora, L. (1999). Weaving, bending, patching, mending the fabric of reality: A cognitive science perspective on worldview inconsistency. *Found Science, 3*, 395–428.

Gabora, L. (2004). Ideas are not replicators but minds are. *Biology and Philosophy*, *19*, 127-143.

Gabora, L. (2006a). Self-other organization: Why early life did not evolve through natural selection. *Journal of Theoretical Biology*, *241*, 441-450.

Gabora, L. (2006b). The fate of evolutionary archaeology: survival or extinction? *World Archaeology*, *38*, 690-696.

Gabora, L. (2007). Why the creative process is not Darwinian. *Creativity Research Journal, 19*, 361–365.

Gabora, L. (2011). Five clarifications about cultural evolution. *Journal of Cognition and Culture*, *11*, 61-83.

Gabora, L. (2013). An evolutionary framework for culture: Selectionism versus communal exchange. *Physics of Life Reviews, 10*, 117-145.

Gabora, L. (2019). Creativity: linchpin in the quest for a viable theory of cultural evolution. *Current Opinion in Behavioral Sciences*, *27*, 77-83.

Gabora, L., & Aerts, D. (2005). Evolution as context-driven actualisation of potential: toward an interdisciplinary theory of change of state. *Interdisciplinary Science Reviews*, *30*, 69-88.

Godfrey-Smith, P. (2012). Darwinism and cultural change. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *367*, 2160-2170.

Goldenfeld, N., Biancalani, T., & Jafarpour, F. (2017). Universal biology and the statistical mechanics of early life. *Philosophical Transactions of the Royal Society A*, *375*, 20160341.

Haldane, J. B. S. (1932). *The causes of evolution*. Princeton, NJ: Princeton University Press.

Hartl, D. L., & Clark, A. G. (2006). *Principles of population genetics,* Fourth Edition. Oxford University Press.

Henrich, J., Boyd, R., Derex, M., Kline, M., Mesoudi, A., Muthukrishna, M., … Thomas, M. (2016). Understanding cumulative cultural evolution. *Proceedings of the National Academy of Sciences*, *113*, E6724-E6725.

Henrich, J., & McElreath, R. (2007). Dual-inheritance theory: the evolution of human cultural capacities and cultural evolution. Oxford UK: *Oxford handbook of evolutionary psychology*.

Heyes, C. (2018). Enquire within: cultural evolution and cognitive science. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*, 20170051.

Hodgson, G. (2002). Darwinism in economics: from analogy to ontology. *Journal Evolutionary Economics*, *12*, 259-281.

Hordijk, W., Steel, M., & Dittrich, P. (2018). Autocatalytic sets and chemical organizations: Modeling self-sustaining reaction networks at the origin of life. *New Journal of Physics*, *20*, 015011.

Jablonka, E., & Lamb, M. J. (2014). *Evolution in four dimensions (revised)*. Boston: MIT press.

Kauffman, S. (1993). *Origins of order*. Oxford: Oxford University Press.

Killeen, P. (2019). The non-Darwinian evolution of behavers and behaviors. *Behavioural Processes, 161,* 45-53.

Kronfeldner, M. (2014). *Darwinian creativity and memetics*. Abingdon-on-Thames, UK: Routledge.

Lewens, T. (2017). Human nature, human culture: The case of cultural evolution. *Interface Focus*, *7*, 20170018.

Lewontin, R. (1970). The units of selection. *Annual review of ecology and systematics*, *1*, 1-18.

Lewontin, R. (1998). The evolution of cognition: Questions we will never answer. In Scarborough, S. Sternberg, & D. Osherson, (Eds.) *An invitation to cognitive science*, vol. *4*, (pp. 107-132). Cambridge: MIT Press.

Mayr, E. (1996). What is a species, and what is not? *Philosophy of Science*, *63*, 262-277.

Mesoudi, A. (2007). Biological and cultural evolution: Similar but Different. *Biological Theory*, *2*, 119-123.

Mesoudi, A. (2016). Cultural evolution: Integrating psychology, evolution and culture. *Current Opinion in Psychology*, *7*, 17-22.

Mesoudi, A. (2017). Pursuing Darwin's curious parallel: Prospects for a science of cultural evolution. *Proceedings of the National Academy of Sciences*, *114*, 7853-7860.

Müller, G. B. (2017). Why an extended evolutionary synthesis is necessary. *Interface Focus*, *7*, 20170015.

Nelson, R., & Winter, S. (2002). Evolutionary theorizing in economics. *Journal of Economic Perspectives*, *16*, 23-46.

O'Brien, M., & Lyman, R. (2000). *Applying evolutionary archaeology: A systematic approach*. Berlin: Springer.

Ochman, H., Lawrence, J., & Groisman, E. (2000). Lateral gene transfer and the nature of bacterial innovation. *Nature, 405,* 299-304.

Pagel, M. (2017). Darwinian perspectives on the evolution of human languages. *Psychonomic Bulletin & Review*, *24*, 151-157.

Richerson, P., & Boyd, R. (1978). A dual inheritance model of the human evolutionary process I: Basic postulates and a simple model. *Journal of Social and Biological Structures*, *1*, 127-154.

Simonton, D. (1999). *Origins of genius: Darwinian perspectives on creativity*. Oxford University Press.

Steel, M. (2000). The emergence of a self-catalyzing structure in abstract origin-of-life models. *Applied Mathematics Letters, 13,* 91-95.

Tëmkin, I., & Eldredge, N. (2007). Phylogenetics and material cultural evolution. *Current Anthropology*, *48*, 146-154.

Vetsigian, K., Woese, C., & Goldenfeld, N. (2006). Collective evolution and the genetic code. *Proceedings of the National Academy of Sciences*, *103*, 10696-10701.

Voorhees, B., Read, D., & Gabora, L. (in press). Identity, kinship, and the evolution of cooperation. *Current Anthropology*.

Wägele, J. W. (2005). *Foundations of phylogenetic systematics*. München: Pfeil.

Wagner, A., & Rosen, W. (2014). Spaces of the possible: universal Darwinism and the wall between technological and biological innovation. *Journal of the Royal Society Interface*, *11*, 20131190.

Whiten, A. (2017). A second inheritance system: the extension of biology through culture. *Interface Focus*, *7*, 20160142.

Woese, C. (1998). The universal ancestor. *Proceedings of the National Academy of Sciences, 95,* 6854-6859.

Woese, C. (2002). On the evolution of cells. *Proceedings of the National Academy of Sciences*, *99*, 8742-8747.

Wright, S. (1931). Evolution in Mendelian populations. *Genetics*, *16*, 97-159.

Selectivity metrics provide misleading estimates of the selectivity of single units in neural networks

Ella M. Gale, Ryan Blything, Nicholas Martin & Jeffrey S. Bowers
(ella.gale, ryan.blything, nm13850, j.bowers@bristol.ac.uk)
School of Psychological Science, University of Bristol, 12a Priory Road Bristol BS8 1TU, UK

Anh Nguyen (anhnguyen@auburn.edu)
Department of Computer Science and Software Engineering, Auburn University, AL, USA

Abstract

To understand the representations learned by neural networks (NNs), various methods of measuring unit selectivity have been developed. Here we undertake a comparison of four such measures on AlexNet: localist selectivity (Bowers et al., 2014); precision (Zhou et al., 2015); class-conditional mean activity selectivity CCMAS (Morcos et al., 2018); and top-class selectivity. In contrast with previous work on recurrent neural networks (RNNs), we fail to find any 100% selective 'localist units' in AlexNet, and demonstrate that the precision and CCMAS measures are misleading and suggest a much higher level of selectivity than is warranted. We also generated activation maximization (AM) images that maximally activated individual units and found that under (5%) of units in fc6 and conv5 produced interpretable images of objects, whereas fc8 produced over 50% interpretable images. Furthermore, the interpretable images in the hidden layers were not associated with highly selective units. We also consider why localist representations are learned in RNNs and not AlexNet.

Keywords: localist representation; grandmother cells; distributed representations.

Introduction

There have been recent attempts to understand how neural networks (NNs) work by analyzing hidden units one at a time using various measures such as localist selectivity (Bowers et al., 2014), class-conditional mean activity selectivity (CCMAS) (Morcos et al., 2018), precision (Zhou et al., 2015), and activation maximization (AM) (Erhan et al., 2009). These measures are defined below, and they all provide evidence that some units respond selectively to categories under some conditions.

Our goal here is to directly compare different measures of object selectivity on a common network trained on a single task. We chose AlexNet (Krizhevsky et al., 2012) because it is a well-studied CNN and many authors have reported high levels of selectivity in its hidden layers via both quantitative (Zhou et al., 2018, 2015) and qualitative methods using Activation Maximization (AM) images (Nguyen et al., 2017; Yosinski et al., 2015; Simonyan et al., 2013). Our main findings are:

1. The different measures provide very different estimates of selectivity.
2. The precision and CCMAS measures are misleading with near perfect selectivity scores associated with units that strongly respond to many different image categories. CCMAS scores are also ambiguous, as explained below.
3. There are no localist 'grandmother cell' representations in AlexNet, in contrast with the localist representations learned in some RNNs.
4. Units with interpretable AM images do not necessarily correspond to highly selective representations.
5. New selectivity measures are required to provide a better assessment of the learned hidden representations in NNs.

Bowers et al. (2014, 2016) assessed the selectivity of hidden units in recurrent NNs using networks similar to those developed by Botvinick & Plaut (2006) designed to explain human short-term memory performance. They reported many 'localist' units that are 100% selective for specific letters or words, where all members of the selective category were more active than and disjoint from all non-members, as can be shown in jitterplots (Berkeley et al., 1995), see Fig. 1 for a unit selective to the letter 'j').

These localist representations were compared to 'grandmother cells' as discussed in neuroscience (Bowers, 2017a). Bowers et al. (2014) argued that the network learned these representations in order to co-activate multiple letters or words at the same time in short-term memory without producing ambiguous blends of overlapping distributed patterns (the so-called 'superposition catastrophe'). Consistent with this hypothesis, localist units did not emerge when the model was trained on letters or words one-at-a-time (a condition in which the model did not need to overcome the superposition catastrophe (Bowers et al., 2014)), see Fig. 1 for an example of a non-selective unit)

In parallel, researchers have reported selective units in the hidden layers of various CNNs trained to classify images into one of multiple categories ((Zhou et al., 2015; Morcos et al., 2018; Zeiler & Fergus, 2014; Erhan et al., 2009), for a review see (Bowers, 2017a)). For example, Zhou et al. (2015) assessed the selectivity of units in the pool5 layer of two CNNs trained to classify images into 1000 objects and 205 scene categories, respectively. They reported multiple 'object detectors' (as defined below) in both networks. Similarly, Morcos et al. (2018) reported that CNNs trained on CIFAR-10 and ImageNet learned many highly selective hidden units, with CCMAS scores often approaching the maximum of 1.0.

Note that these later studies show that selective representations develop in CNNs trained to classify images one-at-

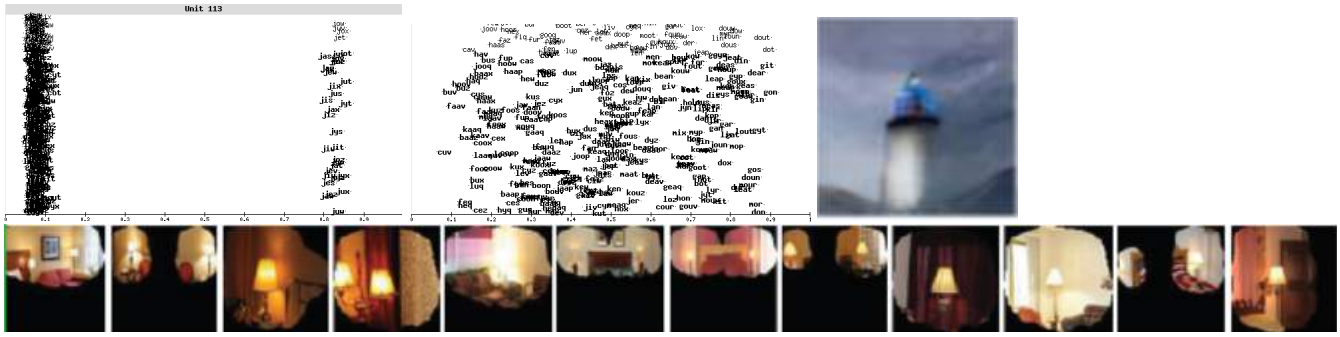


Figure 1: Examples of selectivity measures used. Top left: jitterplot of unit 113 in an RNN under the superposition constraint selective the letter 'j'. Top middle: jitterplot of non-selective unit 160 found when RNN trained on words one-at-a-time; from (Bowers et al., 2016). Top right: activation maximization (AM) image of a unit in conv5 of AlexNet that looks like a lighthouse; from (Nguyen et al., 2016). Bottom: highest activation images for a 'lamp' detector with 84% precision in layer pool5 of AlexNet; from (Zhou et al., 2015).

a-time. This appears to be inconsistent with Bowers et al. (2016) who (a) failed to obtain selective representations for letters or words under these conditions (see Fig. 1); and (b) it suggests that there are additional pressures for CNNs to learn selective representations above and beyond the challenge of overcoming the superposition catastrophe. However, the measures of selectivity that have been applied across studies are different, and accordingly, it is difficult to directly compare results.

In order to directly compare and have a better understanding of the different selectivity measures we assessed (1) localist, (2) precision, and (3) CCMAS selectivity on the prob, fc8, fc7, fc6, and conv5 layers of AlexNet. We also introduce a new measure called top-class selectivity, and show that the precision and CCMAS measures provide much higher estimates of object selectivity compared to other measures. Importantly, we do not find any localist 'grandmother cell' representations in the hidden layers of AlexNet, consistent with the hypothesis that the superposition catastrophe provides a pressure to learn more selective representations (Bowers et al., 2014, 2016).

In addition, we compared these selectivity measures to a state-of-the-art activation maximization (AM) method for visualizing single-unit representations in CNNs (Nguyen et al., 2017). AM images are generated to strongly activate individual units, and some of them are interpretable by humans (e.g., a generated image that looks like a lighthouse, see Fig. 1). For the first time, we systematically evaluated the interpretability of the AM images in an on-line experiment and compare these ratings with the selectivity measures for corresponding units. We show that hidden units with interpretable AM images are not highly selective.

It is important to emphasize that these different measures have all been used to provide insights into the same set of issues. For example, both interpretability of generated images (Le et al., 2011) and localist selectivity (Bowers et al., 2014) have been used to make claims about 'grandmother cells'. The different measures have also been directly compared to one another. For example, Zhou et al. (2015) claim that the object detectors learned in CNNs play an important role in identifying specific objects, whereas Morcos et al. (2018) challenge this conclusion based on their finding that units with high CCMAS measures were not especially important in the performance of their CNNs. Indeed, based on the finding that high CCMAS scores were not predictive of performance, Morcos et al. wrote: "...it implies than methods for understanding neural networks based on analyzing highly selective single units, or finding optimal inputs for single units, such as activation maximization (Erhan et al., 2009) may be misleading". This makes a direct comparison between measures all the more important.

Methods

Networks and Datasets All $\sim$1.2M photos from ImageNet2010 (Deng et al., 2009) were cropped to 277×277 pixels and classified by the pre-trained AlexNet CNN (Krizhevsky et al., 2012) shipped with Caffe (Jia et al., 2014), resulting in 721,536 correctly classified images. Once classified, the images are not re-cropped nor subject to any changes. In Caffe, the softmax operation (Denker & leCun, 1991) is applied at the 'prob'(ability) output layer that contains 1000 units (one for each class). We analyzed these prob units, the fully connected (fc) layers: fc8 (1000 units) that encodes the outputs prior to the softmax operation, fc6 and fc7 (4096 units), and the top convolutional layer conv5 which has 256 filters. We only recorded the activations of correctly classified images. The activation files are stored in .h5 format and can be retrieved at `https://bristol.codersoffortune.net/AlexNet_Merged/`. We selected 233 conv5, 2738 fc6, 2239 fc7, 911 fc8, and 954 prob units for analysis.

Localist selectivity Here we define a unit to be localist for class *A* if the set of activations for class *A* was disjoint with those of not A ($\neg A$).

Localist selectivity is easily depicted with jitterplots in

which a scatter plot for each unit is generated (see Figs. 3 and 4). Each point in a plot corresponds to a unit's activation in response to a single image, and only correctly classified images are plotted. The level of activations is coded along the x-axis, and an arbitrary value is assigned to each point on the y-axis (they are jittered).

Top-Class selectivity Top-class selectivity is related to localist selectivity except that it provides a continuous rather than discrete measure. We counted the number of images from the same class that were more active than all images from all other classes (what we called the top cluster size) and divided the cluster size by the total number of correctly identified images from this class. 100% top-class selectivity is equivalent to a localist representation.

Precision The precision method of finding object detectors (Zhou et al., 2015, 2018) involves identifying a small subset of images that most strongly activate a unit and then identifying the critical part of these images that are responsible for driving the unit. Zhou et al. (2015) took the 60 images that activated a unit the most strongly and asked independent raters to interpret the critical image patches. Zhou et al. (2015) developed a precision metric that calculated the percentage of the 60 images that raters judged to depict the same class of object (e.g., if 50 of the 60 images were labeled as 'lamp', the unit would have a precision index of 50/60 or 83%; see Fig. 1). Object detectors were defined as units with a precision $> 75\%$: they reported multiple such detectors. Here we approximate this approach by considering the 100 images that most strongly activate a given unit and assess the highest percentage of images from a given output class.

CCMAS Morcos et al. (2018) introduced a selectivity index based on the 'class-conditional mean activation' selectivity (CCMAS). The CCMAS for class A compares the mean activation of all images in class A, μ_A, with the mean activation of all images not in class A, $\mu_{\neg A}$, and is given by: $(\mu_A - \mu_{\neg A}) / (\mu_A + \mu_{\neg A})$. Morcos et al. (2018) states that this metric should vary within [0,1], with 0 meaning that a unit's average activity was identical for all classes, and 1 meaning that a unit was only active for inputs of a single class. Here, we assessed class selectivity for the highest mean activation class (CCMAS) as well as for the class with the second highest mean activation μ_A (what we call CCMAS_2) in order to assess the extent to which CCMAS reflects the selectivity to one class.

Activation Maximization We harnessed an activation maximization method called Plug & Play Generative Networks (Nguyen et al., 2017) in which an image generator network was used to generate images (hereafter, AM images) that highly activate a unit. We generated 100 separate images that maximally activated each unit in the conv5, fc6 and fc8 layers of AlexNet and displayed them in a grid format. We then asked 333 participants to judge whether they could identify any repeating objects, animals, or places in images after receiving some practice trials. Participants were recruited using Prolific (*Attrition*, n.d.; Palan & Schitter, 2018), with the experiment run online using gorilla (*Gorilla Experiment Builder*, n.d.). Readers can test themselves at: https://research.sc/participant/login/dynamic/63907FB2-3CB9-45A9-B4AC-EFFD4C4A95D5.

Results

Comparison of selectivity measures.

The mean top-class, precision, and CCMAS selectivities across the conv5, fc6 and fc7 layers are displayed in Fig. 2a–c. We did not plot localist selectivity as there were no localist 'grandmother units' at any internal level (and only 10% at the prob layer, due to the softmax function). The first point to note is that the top-class, precision, and CCMAS measures all increased in the higher layers, showing that they capture degrees of selectivity ignored by the localist measure. Second, the top-class selectivity was extremely low across the hidden layers, with means below 0.25% in the the conv5, fc6, and fc7 layers. Third, the different measures provided very different estimates of selectivity. In contrast with top-class selectivity, the mean precision scores are over an order of magnitude larger in the hidden layers of network, with average precision scores of 9.6%, 12.1%, and 15.4% in layers conv5, fc6, and fc7, respectively. Similarly, the CCMAS measure suggests a much higher level of selectivity than top-class selectivity, with mean scores of .49, .84, and .85 in the conv5, fc6, and fc7 layers, respectively.

This discrepancy is most striking for the units with the highest precision and CCMAS scores. For example, in Fig. 3 we illustrate why the unit fc6.1199 with the highest precision (95%) in fc6 should not be considered a Monarch butterfly detector. Fig. 3a depicts a jitterplot of activations to all accurately identified images, with Monarch butterfly images found across the range of activations. Fig. 3b shows a histogram that plots the distribution of activations for Monarch butterflies. By far the most common activation to correctly identified Monarch butterflies is 0 and the mean is 39.2 ± 0.6. Figures 3 displays example images with 0 (right top), mean (right middle) and maximal (right bottom) activations, and all are identifiable as Monarch butterflies. Thus, classifying this unit as a Monarch butterfly detector is misleading.

Another surprising result is that we did not obtain any 100% top-class selectivity units (localist units) in the prob layer of AlexNet. Rather, the mean top-class selectivity was $\sim$80% in the prob layer, and only $\sim$5% in fc8 (prior to the softmax being applied). Fig. 4 depicts the pattern of activation for units fc8.11 and prob.11 that are examples of the most top-class selective units in these layers (responding to images of 'goldfinch' birds with top-class selectivity measures of 8.4% and 95.2%, respectively). Clearly these units do respond much more selectively than the most selective units in earlier layers (*c.f.* Fig. 3), and at the same time, the jitterplots show why we did not observe any localist units (a few 'goldfinch' images were less active than a few images from other categories).

These jitterplots also show that top-class and localist se-

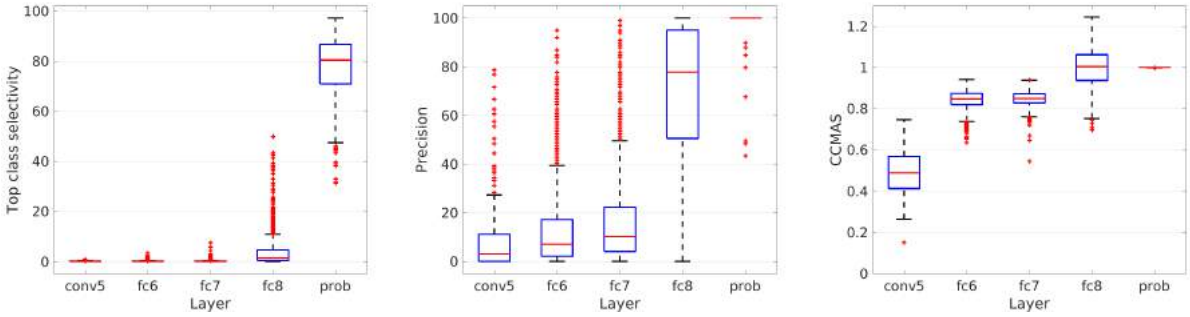


Figure 2: Selectivity measures across different layers of AlexNet. Left: top-class selectivity. Middle: precision 100 (the percentage of the top 100 images which are members of the top class). Right: Class-conditional mean activity selectivity (CCMAS), N.B. as the mean of the unselected classes ($\mu_{\neg A}$ can be less than zero) the CCMAS can go above its expected maximum of 1.

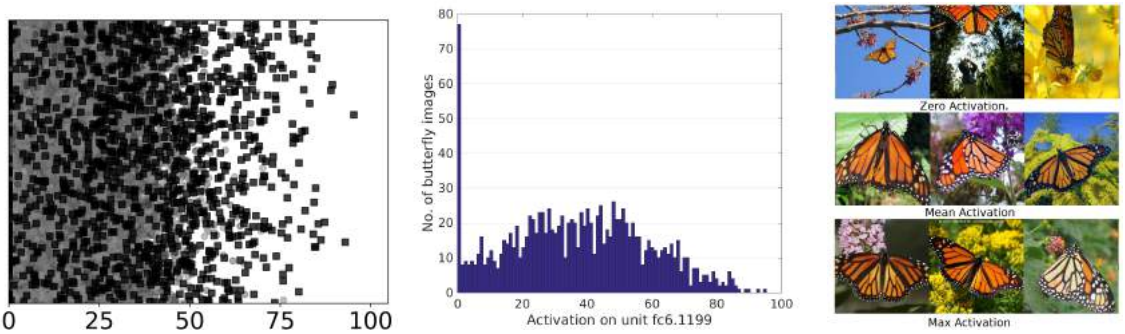


Figure 3: Data for unit fc6.1199. Left: activation jitterplot: black squares: Monarch butterfly images; grey circles: all other classes. Middle: histogram of activations of Monarch butterflies. Right: example ImageNet images with activations of 0.0 (top), the mean (middle), and the maximum (bottom) of the range. Unit fc6.1199 has a precision of 95% over the top 100 images (98.3% over the top 60) and is thus classified as a butterfly detector, yet there are Monarch butterfly images covering the whole range of values, with 72 images (5.8% of the total) having an activation of 0.0.

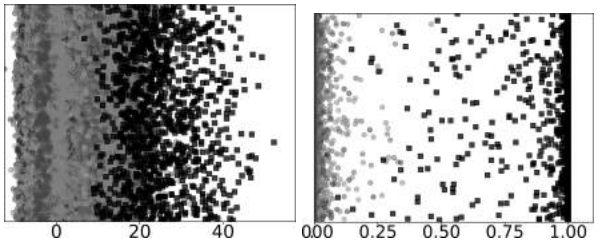
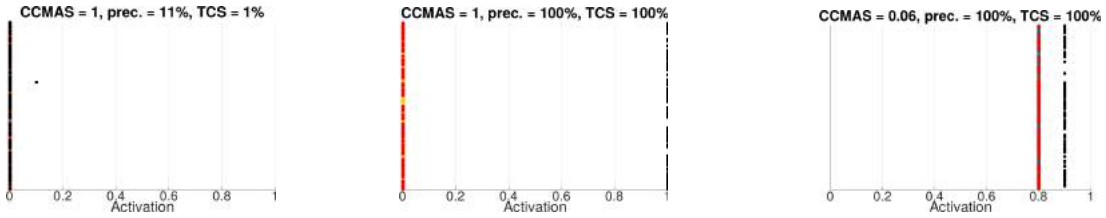


Figure 4: Example data from the fc8 and prob layers. Left: jitterplot activations for unit fc8.11 that has a top-class selectivity of 8.4%. Right: jitterplot activations for prob.11 (i.e. post-softmax) that has top-class selectivity of 95.2%. Activations for the 'ground truth' class 'goldfinch' are shown as black squares, all other classes are shown as greyscale circles.

lectivity provide somewhat misleading estimates of selectivity as well. Consider Fig. 4(left) that depicts a substantial overlap between goldfinch and non-goldfinch activations on unit fc8.11. The 8.4% top-class selectivity score captures the selectivity for the most highly active goldfinch images, but it is blind to the fact that almost all goldfinch images have a reasonably high level of activation (more than most non-goldfinch images). The problem with localist selectivity is highlighted in Fig. 4(right) that shows that the measure misses all but the most extreme version of selectivity. Together, these findings suggest that new selectivity measures are required to better characterize the representations in NNs: precision and CCMAS measures strongly overestimate selectivity, and localist and top-class selectivity provide either a too strict or too narrow a measure of selectivity.

Additional problems with the CCMAS measure

The main problem with the precision and CCMAS measures is that they provide misleadingly high estimates of selectivity, but the CCMAS measure has some additional limitations. First, if the CCMAS provided a good measure of a unit's class selectivity then one should expect that a high measure of selectivity for one class would imply that the unit is not highly selective for other classes. However, the CCMAS score for the most selective category and the second most selective category CCMAS_2 were similar across the conv5, fc6 and fc7



a. One active item from one class. CCMAS = 1, precision = 11%, TCS = 1%.

b. Archetypal 'grandmother' unit. CCMAS = 1, precision = 100%,TCS = 100%.

c. One class more active than the others. CCMAS = 0.06, precision = 100%, TCS = 100%.

Figure 5: Example of where the CCMAS does not match intuitive understandings of selectivity. Generated example data: (a) If a unit is off to all but a single image from a large class of objects, the CCMAS for that class is 1 (maximum possible selectivity). (b) If a unit is strongly activated to all members of one class and off to everything else (an archetypal 'grandmother' cell) the CCMAS is the same as for (a) although the precision and top-class selectivity is vastly different. (c): If a unit has high activations for all classes, but one class (black squares) is 0.1 more than all others (coloured circles), the CCMAS is very low (0.06) despite being %100 top-class selective. The generated examples are for 10 classes of 100 items

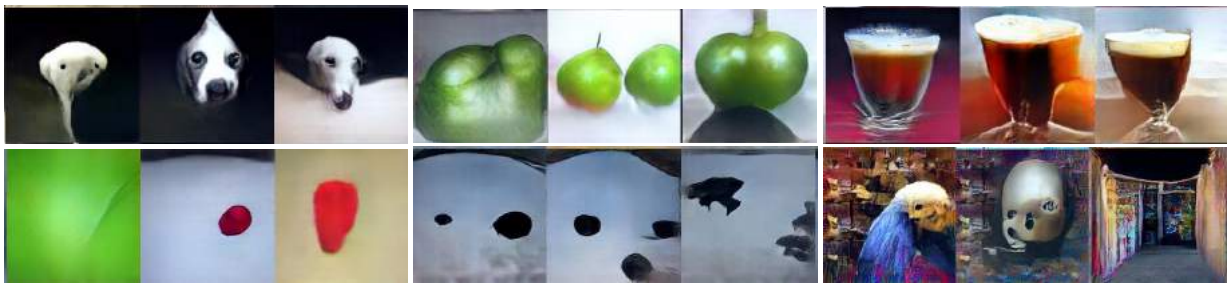


Figure 6: Example AM images that were either judged by all participants to contain objects (top row) or judged by all participants to be uninterpretable as objects (bottom row). The human judgement for conv5.183 (top left) was 'dogs' and the top-class was 'flat-coated retriever'. For fc6.319 (top middle) subjects reported 'green peppers' or 'apples' (all classified as the same broad class in our analysis), and the CCMAS and top-class was 'Granny Smith apples'. For fc8.969 (top right) humans suggested 'beverage' or 'drink': ground truth class for this unit is 'eggnog'. The ground-truth for fc8.865 (bottom right) is'toy-store'.

layers, with the mean CCMAS scores .491, .844, and .848, and the CCMAS_2 scores .464, .821, .831. For example, unit fc7.0 has a CCMAS of .813 for the class 'maypole', and a CCMAS_2 score of .808 for 'chainsaw' (with neither of these categories corresponding 'orangutan' that had the highest precision of score of 14% and a top-class selectivity score of .001%).

Second, the CCMAS measure provides an ambiguous measure of selectivity. To illustrate, consider the artificial scatter plots depicted in Figs. 5a,b. Here we obtain the same perfect CCMAS scores for one unit that selectively responds to one member of a category and another unit that selectively responds to all members of a category. This is problematic for a measure designed to assess *class* selectivity. Third, as shown in Fig. 5c, it is even possible to have a low CCMAS score for a unit with 100 percent top-class selectivity (that is, a low CCMAS selectivity for a grandmother cell). Together, these characteristics of the CCMAS measure may help explain why why Morcos et al. failed to observe the functional importance of units with high CCMAS scores.

Human interpretation AM images

For the behavioral experiment, one hundred generated images were made for every unit in layers conv5, fc6 and fc8 in AlexNet, as in Nguyen et al. (2017), and displayed as 10×10 image panels. A total of 3,299 image panels were used in the experiment (995 fc8, 256 conv5, and 2048 randomly selected fc6 image panels) and were divided into 64 counterbalanced lists for testing. To assess the interpretability for these units as object detectors, paid volunteers were asked to look at image panels and asked if the images had an object / animal or place in common. If the answer was yes, they were asked to name that object simply (i.e. fish rather than goldfish). Analyses of common responses was done for any units where over 80% of humans agreed there was an object present.

The results of the behavioral experiment in which humans rated AM images are reported in Table 1. Consistent with past research, the generated images in the output fc8 layer were often interpreted as objects, and when they were given a consistent interpretation, they almost always (95.4%) correspond to the trained category. By contrast, less than 5%

Table 1: Interpretability judgements. Number of judgments for conv5, fc6 and fc8 were 1332, 10,656 and 5,181, respectively.

LAYER	% YES RESPONSES	% OF UNITS WITH $\geq$ 80% YES RESPONSE	% OVERLAP AMONG HUMANS	% OVERLAP BETWEEN HUMANS AND: TOP CLASS	CCMAS CLASS
conv5	21.7% $\pm$1.1%	4.3% $\pm$ 1.3%	89.5%$\pm$5.7%	34.1%$\pm$14.4%	0%
fc6	21.0% $\pm$0.4%	3.1% $\pm$ 0.4%	80.4%$\pm$4.1%	23.3%$\pm$5.9%	18.9% $\pm$5.9%
fc8	71.2% $\pm$0.6%	59.3% $\pm$1.6%	96.5%$\pm$0.4%	95.4%$\pm$0.6%	94.6% $\pm$0.7%

of units in conv5 or fc6 were associated with consistently interpretable images, and as can be seen in Table 1, the interpretations only weakly matched the category with the highest top-class or CCMAS selectivity. The frequency with which objects were seen by the participants was similar in layers conv5 and fc6 layers and increased in fc8, consistent with the top-class and and precision measures of selectivity.

Apart from showing that there are few interpretable units in the hidden layers of AlexNet, our findings show that the interpretability of images does not imply a high level of selectivity given the maximum top-class selectivity for the hidden units is well under 10% (Fig. 2). In most cases, the top-class selectivity of the interpretable units was well under 1%. To briefly illustrate the types of images that participants rated as objects or non-objects see Fig. 6.

Discussions and Conclusions

Our central finding is that different measures of activation selectivity support very different conclusions when applied to the same units in AlexNet. In contrast with the precision (Zhou et al., 2015) and CCMAS (Morcos et al., 2018) measures that revealed some highly selective units for objects in layers conv5, fc6, and fc8, we found no localist representations, and the mean top-class selectivity in these layers was well under 1%. These findings are in stark contrast with the many localist 'grandmother cell' representations learned in RNNs (Bowers et al., 2014, 2016; Bowers, 2017b).

Not only did the different measures provide very different assessments of selectivity, we found that the precision and CCMAS measures provided highly misleading estimates. For example, a unit with over a 75% precision score for Monarch butterflies had a top-class selectivity of under 5%. Although Zhou et al. (2015) used 75% precision scores as the criterion for 'object detectors', it is inappropriate to call this unit a Monarch butterfly detector given that it did not respond strongly to the majority of Monarch butterfly images (and indeed, the modal response was 0.0; see Fig. 3).

At the same time, we identified problems with the localist, top-class, and activation maximization (AM) methods as well. The localist selectivity measure failed to obtain any localist representations, even at the output prob layer of AlexNet. This measure is so extreme that it misses highly selective representations that are of theoretical interest. The top-class selectivity does provide a graded measure of selectivity (with 100% top-class selectivity equivalent to a localist grandmother cell), but it can underestimate selectivity when a few member from outside the top-class are highly activated (see Fig. 4 (right) for an example). At the same time, the human interpretation of AM images provides a poor measure of hidden-unit selectivity given that interpretable AM images were associated with low top-class selectivity scores. These findings highlight the need to provide better measures of selectivity in order to better characterize the learned representations in NNs.

What should be made of the contrasting findings that localist representations are found in RNNs, but not in AlexNet? The failure to observe localist units in the hidden layers of AlexNet is consistent with the Bowers et al. (2014) claim that these units only emerge in order to support the co-activation of multiple items at the same time in short-term memory. That is, localist representations may be the solution to the superposition catastrophe, and AlexNet only has to identify one image at a time. This may help explain the reports of highly selective neurons in cortex given that the cortex needs to co-activate multiple items at the same time in order to support short-term memory (Bowers et al., 2016). It should be noted that the RNNs that learned localist units were very small in scale compared to AlexNet, and accordingly, it is possible that the contrasting results reflect the size of the networks rather than the superposition catastrophe *per se*. Relevant to this issue, Karpathy et al. (2016) reported examples of selective representations in a larger RNN with long-short term memory (LSTM) trained to predict text. Although they did not systematically assess the degree of selectivity, they reported examples that are consistent with 100% selective units, for similar findings see Lakretz et al. (2019). It will be interesting to apply our measures of selectivity to these larger RNNs. It should also be noted that there are recent reports of selective representations in Generative Adversarial Networks (Bau et al., 2019) and Variational Autoencorder Networks (Burgess et al., 2018) where the superposition catastrophe is not an issue. Again, it will be interesting to assess the selectivity of these units according to our measures in order to see whether there are additional computational pressures to learn highly selective or even grandmother cells. We will be exploring these issues in future work.

References

Attrition. (n.d.). http://Prolific.ac. (Accessed: 2018-09-24)

Bau, D., Zhu, J.-Y., Strobelt, H., Zhou, B., Tenenbaum, J. B., Freeman, W. T., & Torralba, A. (2019). Visualizing and understanding generative adversarial networks. *arXiv preprint arXiv:1901.09887*.

Berkeley, I. S., Dawson, M. R., Medler, D. A., Schopflocher, D. P., & Hornsby, L. (1995). Density plots of hidden value unit activations reveal interpretable bands. *Connection Science*, *7*(2), 167–187.

Botvinick, M. M., & Plaut, D. C. (2006). Short-term memory for serial order: a recurrent neural network model. *Psychological review*, *113*(2), 201.

Bowers, J. S. (2017a). Grandmother cells and localist representations: a review of current thinking. *Language, Cognition, and Neuroscience*, 257-273.

Bowers, J. S. (2017b). Parallel distributed processing theory in the age of deep networks. *Trends in cognitive sciences*, *21*(12), 950–961.

Bowers, J. S., Vankov, I. I., Damian, M. F., & Davis, C. J. (2014). Neural networks learn highly selective representations in order to overcome the superposition catastrophe. *Psychological review*, *121*(2), 248–261.

Bowers, J. S., Vankov, I. I., Damian, M. F., & Davis, C. J. (2016). Why do some neurons in cortex respond to information in a selective manner? insights from artificial neural networks. *Cognition*, *148*, 47–63.

Burgess, C. P., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G., & Lerchner, A. (2018). Understanding disentangling in β-VAE. *arXiv preprint arXiv:1804.03599*.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *Computer vision and pattern recognition, 2009. cvpr 2009. ieee conference on* (pp. 248–255).

Denker, J. S., & leCun, Y. (1991). Transforming neural-net output levels to probability distributions. In *Advances in neural information processing systems* (pp. 853–859).

Erhan, D., Bengio, Y., Courville, A., & Vincent, P. (2009). Visualizing higher-layer features of a deep network. *University of Montreal*, *1341*(3), 1.

Gorilla experiment builder. (n.d.). www.gorilla.sc. (Accessed: 2018-09-24)

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., ... Darrell, T. (2014). Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*.

Karpathy, A., Johnson, J., & Fei-Fei, L. (2016). Visualizing and understanding recurrent networks. *Workshop Track at International Conference on Learning Representations*.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

Lakretz, Y., Kruszewski, G., Desbordes, T., Hupkes, D., Dehaene, S., & Baroni, M. (2019). The emergence of number and syntax units in lstm language models. *arXiv preprint arXiv:1903.07435*.

Le, Q. V., Ranzato, M., Monga, R., Devin, M., Chen, K., Corrado, G. S., ... Ng, A. Y. (2011). Building high-level features using large scale unsupervised learning. *arXiv preprint arXiv:1112.6209*.

Morcos, A. S., Barrett, D. G., Rabinowitz, N. C., & Botvinick, M. (2018). On the importance of single directions for generalization. *arXiv preprint arXiv:1803.06959*.

Nguyen, A., Clune, J., Bengio, Y., Dosovitskiy, A., & Yosinski, J. (2017). Plug & play generative networks: Conditional iterative generation of images in latent space. In *Cvpr* (Vol. 2, p. 7).

Nguyen, A., Dosovitskiy, A., Yosinski, J., Brox, T., & Clune, J. (2016). Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In *Advances in neural information processing systems* (pp. 3387–3395).

Palan, S., & Schitter, C. (2018). Prolific. aca subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22–27.

Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., & Lipson, H. (2015). Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*.

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European conference on computer vision* (pp. 818–833).

Zhou, B., Bau, D., Oliva, A., & Torralba, A. (2018). Interpreting deep visual representations via network dissection. *IEEE transactions on pattern analysis and machine intelligence*.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2015). Object detectors emerge in deep scene CNNs. In *International conference on learning representations.*

A rational model of syntactic bootstrapping

Jon Gauthier, Roger P. Levy, Joshua B. Tenenbaum
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology
jon@gauthiers.net, {rplevy, jbt}@mit.edu

Abstract

Children exploit regular links between the meanings of words and the syntactic structures in which they appear to learn about novel words. This phenomenon, known as *syntactic bootstrapping*, is thought to play a critical role in word learning, especially for words with more opaque meanings such as verbs. We present a computational word learning model which reproduces such syntactic bootstrapping phenomena after exposure to a naturalistic word learning dataset, even when under substantial memory constraints. The model demonstrates how experimental syntactic bootstrapping effects constitute rational behavior given the nature of natural language input. The model unifies computational accounts of word learning and syntactic bootstrapping effects observed in the laboratory, and offers a path forward for demonstrating the broad power of the syntax–semantics link in language acquisition.

Keywords: syntactic bootstrapping; word learning; computational models

Children face multiple challenges of induction when acquiring their first language. They must work out the most fundamental features of language: that words exist, and that they can be used to refer to entities and relations out in the world. At a higher level, they must work out what words actually mean, and how those words can productively combine with other words to form phrases and sentences.

A successful research program has identified how children as young as 13 months can learn the meanings of a particular class of words — concrete nouns — from noisy observations of adult language use (Smith and Yu, 2008; Trueswell et al., 2013). While nouns often pick out concrete referents which are easily identifiable by a listener, other classes of words pose more substantial learning problems. Verbs, for example, often have no concrete reference in the perceptual world which the child directly observes. Certain verb meanings may also be under-determined by the perceptual facts: verb pairs such as *chase* and *flee* or *hit* and *kick* often pick out the same events, though they have vastly different meanings (Gleitman et al., 2005).

These features make learning verb meanings a challenge for both children and adults. The productive vocabularies of young children are heavily skewed toward frequent nouns with concrete referents (Fenson et al., 1994). Adult subjects in laboratory language learning experiments also routinely struggle to identify verb meanings from observations of their use (Gillette et al., 1999). But children somehow climb over these learning barriers to become adults who can *give* and *take* or *hit* and *kick*. We must account, then, for how that learning goes through. First, because verbs make reference to abstract events and relations between entities, we must account for the representations of such events and relations in the mind of the child. In other words, we must account for the **target** representations of word learning. Second, we must explain what information **sources** children exploit in order to learn which words pick out which events and relations. Because perceptual information under-determines the solution to this learning problem, there must be other sources of information in the learner's experience which help determine the meanings of these words.

This paper addresses the theory of *syntactic bootstrapping*, which claims that children exploit systematic relations between the syntactic structures in which verbs are used and their semantics in order to learn about the meanings of novel words (Landau and Gleitman, 1985; Fisher et al., 2010). After reviewing corpus and experimental evidence regarding the syntax–semantics link, we formalize syntactic bootstrapping in a probabilistic computational model, proceeding from minimal assumptions about the structure of the lexicon to a model which replicates the qualitative behavior of children in syntactic bootstrapping experiments. We show how the knowledge assumed by this model can be learned *from scratch* on naturalistic data, as it constructs both a concrete lexicon and abstract beliefs about the correspondence between verb form and meaning.

Syntactic bootstrapping

On the syntactic bootstrapping account, children analyze the syntactic structures in which verbs appear in order to predict aspects of their meaning not well determined by the perceptual context. At a high level, this theory is a claim about the relation between two representational spaces in the mind of the learner: the space of meanings M and the space of syntactic representations S. As such, these theories must make assumptions about the structure of these spaces. As we will see, many theories regarding the syntax–semantics link presuppose the existence of core meaning predicates such as CAUSE and BECOME (Levin and Rappaport Hovav, 2011; Pinker, 1989). While such predicates have been motivated by theoretical work elsewhere in cognitive development (see e.g. Hespos and Spelke, 2004; Muentener and Carey, 2010), the continued success of the syntactic bootstrapping paradigm

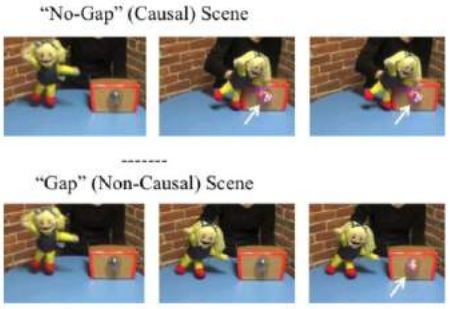


Figure 1: From Kline et al. (2017), fig. 2. Scene pairs contrast minimally in the presence or absence of a *causation* event. In the "causal" scene, the puppet moves to contact the toy, which immediately activates; in the "noncausal" scene, the puppet moves but does not contact the toy, and the toy only activates after a delay.

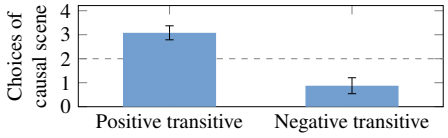


Figure 2: From Kline et al. (2017). Children point to the "causal" scene when given a positive sentence with a novel transitive verb (*can you find where she wugged the round thing?*), and to the "noncausal" scene for a negative sentence with a novel transitive verb (... *didn't wug the round thing?*).

provides orthogonal positive evidence for the reality of these structures in the mind of the child.

Corpus studies have shown how aspects of verb meaning both coarse-grained (causation and movement) and fine-grained (movement of liquids vs. movement of solids) can be predicted from the constructions in which verbs appear (Naigles and Hoff-Ginsberg, 1995; Levin, 1993). Decades of experimental evidence also support the idea that children exploit such structural relationships between a verb's syntactic behavior and its meaning. One of the most productive lines of research has focused on the correspondence between a verb's appearance in transitive syntactic constructions (*the X Ys the Z*) and a semantic predicate CAUSE (Naigles, 1990). Some of the most recent experimental evidence argues for such a fine-grained link between transitive syntax and physical causation (Kline et al., 2017). Kline et al. presented children with pairs of scenes, each involving a moving puppet and a toy which activated or lit up. An example scene pair is shown in Figure 1. While each scene pair involved similar motion events, a "causal" scene in each pair also exhibited an event of external causation, using cues known to be salient to young children (spatial and temporal continuity between an agent's action and an object's response) (Michotte, 1963; Muentener and Carey, 2010). In two-alternative forced choice test trials, children were given a sentence containing a novel verb and asked to pick the scene it referred to: either in a positive frame (*Can you find where she wugged the round thing?*) or a negative frame (... *didn't wug the round thing?*). Figure 2 shows the main effect in the experiment of Kline et al. (2017). Across several tested minimal-contrast scene pairs, children preferred to point at the causal scene when queried with the positive frame and at the non-causal scene when queried with the negative frame.

The findings of Kline et al. show that 3- and 4-year-olds latch onto a reliable relationship in English between transitive syntax and the semantic predicate CAUSE documented elsewhere in the cognitive development literature. This is a case of syntactic bootstrapping: children exploit a word's syntactic behavior in order to make guesses about its meaning.

As a broad theory regarding the construction of the lexicon, though, syntactic bootstrapping needs to eventually do quite a bit more work. Taken to its extreme, it needs to explain how each of the semantic contrasts present in a meaning space M can be explained by corresponding contrasts in a syntactic representation S. In the absence of other good accounts of verb meaning, the contrast between *chase* and *flee* and the contrast between *hit* and *kick*, for example, must be predictable from contrasts in syntactic behavior. To test the full power of syntactic bootstrapping as a theory of the construction of the lexicon, then, we must further formalize our assumptions about the structure of the syntactic space S and the meaning space M, and provide clear proof of the learnability of relations between the two spaces.

The remainder of this paper takes some first steps in that direction. We first formalize syntactic bootstrapping in a probabilistic model, showing how we can proceed from minimal assumptions about the structure of the lexicon to a model which replicates the qualitative behavior of children in syntactic bootstrapping experiments. We next show how this probabilistic model can be learned from scratch on naturalistic data, constructing both a concrete lexicon and abstract beliefs about the syntax–semantics link through only *unsupervised* experience of ambiguous language use in grounded contexts.

Related work

Most past computational models of word learning have focused on the acquisition of words with concrete referents, explaining the learning dynamics and characteristic patterns of success and failure observed in adults and children (Frank et al., 2009; Trueswell et al., 2013; Stevens et al., 2017). Our model will replicate the important structural features of these models — explicit representations of uncertainty over possible lexica, stored under strong resource limitations — and further extend to the more challenging task of acquiring verb meanings, which have either ambiguous reference or no concrete reference at all in the world of the learner.

While other computational models have been used to replicate verb learning and syntactic bootstrapping phenomena (Abend et al., 2017; Barak et al., 2014; Alishahi and Stevenson, 2008), they have been deployed only in simplified learn-

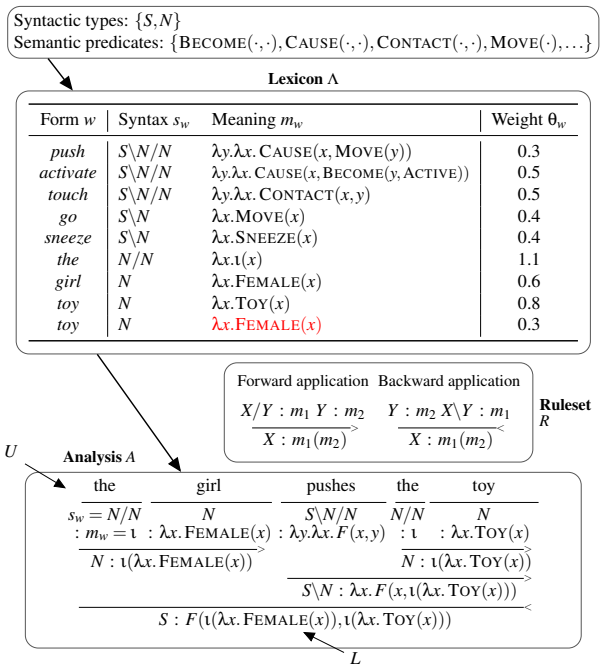


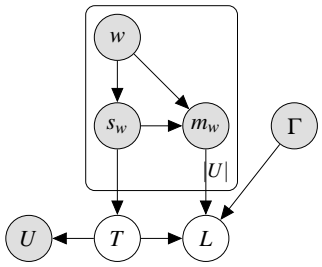
Figure 3: A CCG combines a learned lexicon Λ with a fixed ruleset R in order to yield analyses of input utterances. The bottom of the figure shows an analysis of the utterance *the girl pushes the toy* (read from top to bottom), which jointly yields syntactic and semantic representations of the sentence.

ing situations, where a learner is shown utterances explicitly paired with their ground-truth meaning representations (or a set of possible meaning representations). In contrast, our model learns in a *distantly supervised* setting: it is only explicitly told that the utterances have meanings which are *true* in the current scene, and must work out word-level meanings and utterance-level meanings on its own. Because no word-level meanings are ever explicitly presented to the learner, it must induce word meanings by searching through the infinite space of possible lambda-calculus meaning representations. This learning setting is thus qualitatively different than the direct-supervision setting studied in past bootstrapping work.

We see our model as complementary to those of Sadeghi and Scheutz (2018) and Gauthier et al. (2018), who show how more minimal syntactic representations can support specific types of early syntactic bootstrapping. Our model integrates both a full syntactic formalism and a general ability to track probabilistic links between syntactic and semantic representations. As such, the model is able to scale to the more complex syntactic bootstrapping phenomena studied in this paper, using syntactic features to resolve finer-grained features of verb meaning.

A formal model

We visualize the major details of our model in Figure 3. A learner constructs a *lexicon* Λ, associating particular word-



w	Wordform
s_w	Syntactic type of wordform w
m_w	Meaning of wordform w
Γ	Scene/context
U	Utterance (word sequence)
T	Derivation (syntactic analysis)
L	Logical form (semantic analysis)

Figure 4: A generative model of an utterance U situated in a scene Γ, drawing on lexical items $(w, s_w, m_w) \in \Lambda$.

forms w with syntactic types s_w and meanings m_w.[1] The syntactic types of words are represented using the formalism of combinatory categorial grammar (CCG; Steedman and Baldridge, 2006). Word meanings are represented as expressions in a typed lambda calculus, built from core semantic predicates ranging from concrete properties (e.g. FEMALE) to abstract relations (e.g. CAUSE). These representations draw on the lexical conceptual structures often discussed in literature on the syntax–semantics link (see e.g. Levin and Rappaport Hovav, 2011).

The contents of this lexicon are combined with parsing *rules* in order to produce joint syntactic and semantic representations of full utterances. Figure 3 shows how entries from the lexicon Λ combine with a ruleset R to analyze a sentence.

The grammar's *syntactic types* describe how words combine with their arguments. These syntactic types may be of either a *primitive* type (e.g. N) or of *functional* type (e.g. S/N). Functional types combine with syntactic arguments to their left or right, eventually yielding a phrase of a particular primitive type.

The CCG rule set R, shown in the middle of Figure 3, specifies these combination rules.[2] Figure 3 (lower section) shows how the two rules in our ruleset are used to analyze the example sentence *the girl pushes the toy*. After first retrieving lexical entries for each of the tokens in the sentence (top row), we iteratively run the application rules, composing functional types with primitive types to their left or right. Whenever such syntactic composition occurs, we likewise unify the corresponding semantic expressions by function application.

Each CCG analysis yields a tree structure (bottom of Figure 3) whose root contains the syntactic type and semantic analysis of the entire input string. We call this final semantic representation the *logical form* of a sentence, and the particular tree structure of rule applications the *derivation* (analogous to a syntactic parse). We let $A = \langle L, T \rangle$ denote the full analysis of an utterance, where L is the logical form and T is the derivation.

[1] This walkthrough involves a minimal amount of equations, focusing instead on applications to concrete word learning problems. Model details are provided in the appendix of this paper.

[2] See Steedman and Baldridge (2006) for a full description.

Scene	Events
Γ_1	CAUSE($girl$, BECOME(toy, $active$))
	CONTACT($girl$, toy); MOVE($girl$)
Γ_2	BECOME(toy, $active$); MOVE($girl$)

Table 1: Two sample scene representations from our model of the two-alternative forced choice test trial of Kline et al. (2017).

We next design a minimal probabilistic model on top of this CCG formalism which can realize, among other things, the behavior of the children in the experiment of Kline et al. (2017). Our model adapts past work on probabilistic CCGs (see e.g. Zettlemoyer and Collins, 2007; Artzi and Zettlemoyer, 2013), adding a critical inductive bias linking syntactic and semantic representations within the lexicon. We illustrate the model as a plate diagram in Figure 4, and walk through its behavior in the following paragraphs.

We will walk through this model in the context of the Kline et al. (2017) paradigm, showing how it can realize the subjects' observed behavior. For the rest of this section, we assume the provisional lexicon shown in the top of Figure 3, associating particular words with candidate syntactic types and meanings. Later, we will remove this assumption and show how such a lexicon can be learned from experience alone.

Consider an utterance $U =$ *the girl pushes the toy* given in a grounded context Γ.[3] We can combine the CCG framework with our weighted lexicon to compute the probability of an arbitrary analysis:

$$P(A = \langle L, T \rangle \mid \Lambda, \Gamma) \propto P(\Gamma) P(L \mid \Gamma) \exp\Big(\sum_{(w, s_w, m_w) \in T} \theta_w\Big) \quad (1)$$

where $P(\Gamma)$ is a uniform prior over potential contexts, and $P(L \mid \Gamma)$ is one only when a logical form L is true of the context Γ. The final term in the above equation does the majority of the work, combining the weights θ_w of lexical entries involved in the derivation T. The lexicon in Figure 3 licenses multiple analyses of the sentence *the girl pushes the toy*, since it contains two candidate entries for the word *toy*. Equation (1) can be used to rank the resulting analyses — one of which is shown in the bottom section of Figure 3 — according to the constituent lexical weights θ_w.

In the experiment of Kline et al. (2017), a child hears the utterance $U =$ *the girl gorps the toy* and is asked to pick which of two scenes Γ_1, Γ_2 the utterance refers to. We represent the scenes as lists of propositions like those in Table 1.

Unlike our previous example, this utterance contains a novel word which has no corresponding entries in the lexicon. We must induce candidate syntactic types s_w and meanings m_w using the remainder of the probabilistic model.

We begin by enumerating the possible syntactic types s_w of the novel word. Given the contents of the provisional lexicon Λ (shown in the top left of Figure 3) and our parsing ruleset, there is just one syntactic analysis of *gorps* which yields a valid parse. This parse has the same structure as that shown in bottom section of Figure 3. The parse assigns the word the syntactic type $S\backslash N/N$: the syntactic type of a transitive verb.[4]

We next make predictions about the candidate meanings of *gorps*. This prediction process is visualized in Figure 5. We begin by sampling meanings m_w conditioned on the possible syntactic representations s_w. This is the point at which syntactic bootstrapping plays a critical role: the model calculates a distribution $P(m_w \mid s_w = S\backslash N/N)$, which we expect should favor meanings involving the predicate CAUSE:

$$P(\text{predicate}_i \mid s_w) \propto C + \sum_{\substack{(s_i, m_i, \theta_i) \in \Lambda \\ : s_w = s_i \wedge \text{predicate} \in m_i}} \theta_i \quad (2)$$

$$P(m_w \mid s_w) \propto \prod_{\text{predicate}_i \in m_w} P(\text{predicate}_i \mid s_w) \quad (3)$$

where C is a smoothing constant, fit as a hyperparameter. Equation (2) aggregates the total weight mass in the lexicon allocated to any particular predicate for lexical entries with syntactic type s_w. The product term of Equation (3) combines these individual predicate probabilities in order to score possible complete meanings m_w of the word *gorps*. The left panel of Figure 5 shows a ranked list of meanings computed by this equation under our provisional lexicon.

Each candidate meaning and syntactic representation of the word *gorps*, when combined with the rest of the words in the sentence, yields a full syntactic derivation T and logical form L. These utterance-level meaning representations are scored based on the scene Γ. Here we incorporate the critical constraint that logical forms L must consist of messages which are *true* of the scene Γ. This effectively filters the candidate complete meanings L, yielding a renormalized distribution over full sentence meanings as shown in the middle panel of Figure 5.

We can combine the above distributions in order to perform the critical inverse inference $P(\Gamma \mid U, \Lambda)$: which scene does the utterance *the girl gorps the toy* refer to? This distribution is computed via Bayes' rule, yielding the posterior distribution shown in the right panel of Figure 5. The positive sentence containing the novel word *gorps* is predicted to refer to the scene with a salient causation event. By a similar logic as shown in this walkthrough, the negative sentence *the girl doesn't gorp the toy* is taken to refer to the scene missing the salient causation event.

This section has demonstrated how the probabilistic model sketched in Figures 3 and 4 reproduces syntactic bootstrapping behavior, using the transitive syntax of novel words to predict meanings containing the semantic predicate of CAUSE. The model integrates the CCG parsing formalism with a statistical mechanism for tracking the relations be-

[3] Contexts will become relevant later in the paper. See Table 1 for an example context representation.

[4] In cases where there are multiple syntactic types for a novel word, they are scored according to a distribution $P(s_w \mid \Lambda)$, given in Equation (9).

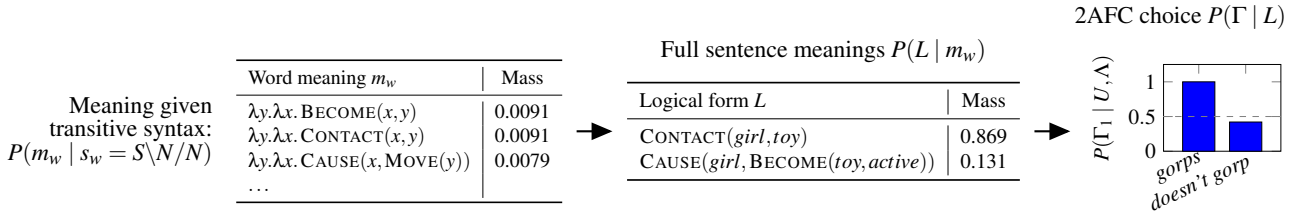


Figure 5: Computation of the meaning of a novel word *gorps* in the sentence *the girl gorps the toy* proceeds in three steps: word meanings are enumerated according to a distribution $P(m_w \mid s_w)$ (Equation (3)), full sentence meanings L are produced via CCG parsing, and the candidate scenes Γ_1, Γ_2 are scored according to which sentence meanings are true of which scene.

tween syntactic structures and their semantic correlates, helping the learner to make predictions about the meanings of novel words.

Learning

The previous section assumed that the learner already possessed a knowledge state as given in Figure 3, where wordforms like *girl* and *push* already have correct meaning representations. In this section, we show how such a lexicon can be acquired across multiple instances of ambiguous language use in context, in a manner that requires minimal long-term memory capacity and remains robust to noise in the input.

We expose our model to a sequence of observations $O = (\langle U_i, \Gamma_i \rangle)$ of utterances U_i grounded in particular scenes Γ_i. We proceed by observing each data point O_i in sequence and updating a lexicon Λ, inducing novel lexical entries as necessary and updating weights θ_w in the lexicon. The learner never directly observes the mapping between words and their referents, or between sentences and their meanings. The task of the learner is to derive word meanings, and methods for composing words, such that each utterance U_i is true in its context Γ_i.

We also constrain our word learner to encode only a limited number of lexical entries per word at all times. We label this limit ℓ, and evaluate its influence as a free parameter in the following experiments. Concretely, after each observation O_i, we retain only the ℓ highest-weight lexical entries per wordform.

Let Λ_i be a learner's lexicon representation before observing the example O_i. Suppose that the utterance U_i is observed in a context Γ_i which contains a novel word $w = gorps$: *the girl* ***gorps*** *the toy*. The machinery already presented in the previous section can be used to induce candidate novel meanings for the word *gorps*. In order to support incremental learning, we include an additional *weight update* step after each utterance is observed. Given the utterance U_i, we update the weights of each lexicon entry in order to increase the probability of observing the utterance under the model given in Figure 4. Further details on the learning algorithm are given in the appendix.

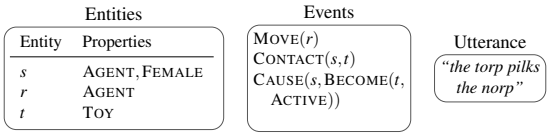


Figure 6: An example observation. Utterances refer to objects (*the norp*) or events (*the torp pilks the norp*).

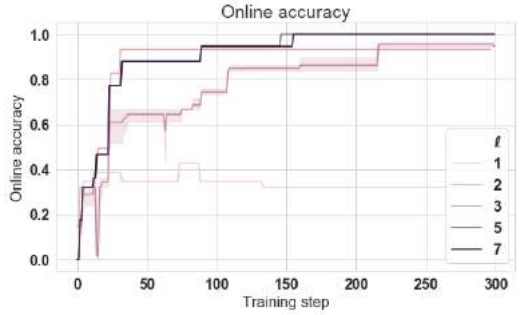


Figure 7: Online accuracy in predicting sentence meanings for word learning models with different numbers of allowed stored meanings ℓ. Shaded regions represent 95% CI.

Experiments

We deploy the above learning model on a synthetic dataset in which short utterances pick out objects, events, and relations in a simulated environment. This environment is similar to those used in artificial intelligence research on visual question answering (see e.g. Johnson et al., 2017), but contains more complex utterances which make reference to abstract events and relations (such as causation, state change, and movement). Figure 6 shows an example scene–utterance pair drawn from this dataset.

We generate observations O_i by first sampling a context Γ_i. Each context contains a random number of entities (agents and objects), and a random number of events relating those entities, structured as propositions like those shown in Table 1. Contexts always contain multiple simultaneous events, such that the learner is only ever exposed to ambiguous and indirect observations of sentence meaning.

Each entity and event is assigned a fixed random wordform throughout the experiment, and utterances are gen-

erated by combining the wordforms for the involved entities and events according to pre-designed templates. For example, if we sample a scene which contains an event $\textsc{Cause}(girl, \textsc{Move}(toy))$, we might generate an utterance *the torp pilks the norp*, where *torp* refers to female agents, *norp* refers to toys, and the whole sentence must pick out the complete event structure. We also randomly generate *negative* utterances, where verbs are modified by words to their immediate left who function to negate the overall sentence meaning. For example, *the torp doesn't pilks the norp* has the meaning $\neg\textsc{Cause}(girl, \textsc{Move}(toy))$.[5]

The task of the learner is to infer the meanings of each of these words by observation of random scenes $O_i = \langle U_i, \Gamma_i \rangle$. While we have access to the ground-truth correspondence between sentences and their full logical forms during scene generation, this mapping is not provided to the learner.

We evaluate the learner's lexicon acquisition by two metrics: 1) its accuracy in predicting the ground-truth semantic representations of test sentences, and 2) its accuracy in the syntactic bootstrapping two-alternative forced choice task of Kline et al. (2017). Figure 7 shows the model's performance on the first across learning time, as the model is incrementally exposed to more examples O_i. Both graphs contrast models with different settings of the hyperparameter ℓ, which controls the maximum number of entries that can be stored across observations for any wordform in the lexicon. For all settings of $\ell > 1$, the model reaches high performance within 100 examples. All models reach perfect performance on the second syntactic bootstrapping 2AFC task after just a few examples: the correct acquisition of just one or two transitive verbs is enough to support the induction of a productive belief about the link between verb syntax and semantics.

The results in Figure 7 demonstrate that even highly resource-constrained Bayesian learners can acquire an accurate lexicon in a data-efficient manner. These same learners quickly derive a syntactic bootstrapping capacity from their own lexicons, supporting more efficient learning in the future.

Conclusion

This paper has presented a computational word learning model which actively tracks the correspondences between the syntactic and semantic behavior of words. We demonstrated how this framework can capture experimentally observed syntactic bootstrapping phenomena, and that such phenomena can be explained as the rational behavior of a cross-situational learner exposed to a corpus of naturalistic data. Critically, both word learning (of nouns and verbs) and also the acquisition of the high-level syntactic bootstrapping behavior still go through given substantial long-term memory constraints, in which models store just a few candidate interpretations per wordform in their language.

As a computational model of acquisition, this framework makes predictions about how people should interpret and generalize novel words. Our framework allows us to make rigorous and explicit statements about the structure of the mental representational spaces underlying these generalizations. In ongoing work, we are using the same model class presented in this paper to detect candidate links between word syntax and word semantics which a rational learner can (and should) exploit.

[5]The training corpus is generated from a collection of 3 unique referents and 5 unique event types, each of which has one fixed referring expression. This yields a total of 51 unique utterances.

Acknowledgments

The authors gratefully acknowledge support from the MIT Sensetime Alliance, MIT Quest for Intelligence, and the Open Philanthropy Foundation.

Model details

This final section provides mathematical details on the model for completeness.

Reading from Figure 4, the probability of a full utterance is:

$$P(U \mid \Gamma, \Lambda) \propto P(\Gamma) \sum_{L,T} P(T \mid \Lambda) P(U \mid T) P(L \mid T, \Gamma, \Lambda) \tag{4}$$

We assume a uniform prior over scenes $P(\Gamma)$, and let $P(U \mid T)$ be 1 exactly when the span of T is equivalent to U, and zero otherwise. Lastly, we define the probability of a logical form L in terms of the derivation T and context Γ as follows:

$$P(L \mid T, \Gamma, \Lambda) = P(L \mid T, \Lambda) P(L \mid \Gamma) \tag{5}$$

$$P(L \mid T, \Lambda) \propto \mathbf{1}\{L \text{ is determined by } T, \Lambda\} \tag{6}$$

$$P(L \mid \Gamma) \propto \mathbf{1}\{L \text{ is true in } \Gamma\} \tag{7}$$

Novel word induction Given a novel word w, we resort to the full Bayesian model to make predictions about its syntactic type s_w and meaning m_w.

$$P(w \to (s_w, m_w) \mid U, \Gamma) \propto P(s_w \mid w) P(m_w \mid s_w) P(U \mid \Gamma, \Lambda \cup (w, s_w, m_w)) \tag{8}$$

The only term not yet defined is the distribution over syntactic types $P(s_w \mid w)$. This distribution is computed by simple inspection of the lexicon. The probability mass assigned to a particular syntactic category s is proportional to the total weight assigned to entries in Λ with category s:

$$P(s_w \mid \Lambda) \propto C + \sum_{(w, s_w, m_w, \theta_w) \in \Lambda} \theta_w \tag{9}$$

where C is a smoothing constant.

As shown in the earlier model walkthrough, Equation (8) is used to initialize the weights for the lexical entries of novel words.

Weight updates Let g be the highest probability correct analysis of a sentence $\langle L, T \rangle$, and let B be the set of the k most probable incorrect analyses.[6] For each lexical entry $x_i = (w_i, s_i, m_i, \theta_i)$ with weight θ_i, we perform the following perceptron update:

$$\theta_i \mathrel{+}= \eta \left(\mathbf{1}\{x_i \in g\} - \frac{1}{|B|} \sum_{b \in B} \mathbf{1}\{x_i \in b\} \right) \tag{10}$$

where η denotes a learning rate, and $x_i \in A$ is true iff the lexical item x_i participates in the analysis A. Note that this update will only affect lexical entries with wordforms used in the utterance U_i.

[6]Here a "correct" analysis is one which has nonzero probability under Equation (1). Note that, consistent with the cross-situational paradigm, only analyses with meanings that are true of Γ_i have nonzero probability.

References

Abend, O., Kwiatkowski, T., Smith, N. J., Goldwater, S., and Steedman, M. (2017). Bootstrapping language acquisition. *Cognition*.

Alishahi, A. and Stevenson, S. (2008). A computational model of early argument structure acquisition. *Cognitive science*, 32(5):789–834.

Artzi, Y. and Zettlemoyer, L. (2013). Weakly supervised learning of semantic parsers for mapping instructions to actions. *Transactions of the Association for Computational Linguistics*, 1:49–62.

Barak, L., Fazly, A., and Stevenson, S. (2014). Learning verb classes in an incremental model. In *CMCL*, pages 37–45.

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., Pethick, S. J., Tomasello, M., Mervis, C. B., and Stiles, J. (1994). Variability in early communicative development.

Fisher, C., Gertner, Y., Scott, R. M., and Yuan, S. (2010). Syntactic bootstrapping.

Frank, M. C., Goodman, N. D., and Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological science*, 20(5):578–585.

Gauthier, J., Levy, R., and Tenenbaum, J. B. (2018). Word learning and the acquisition of syntactic–semantic overhypotheses. In *Proceedings of the 40th Annual Meeting of the Cognitive Science Society*, Madison, Wisconsin.

Gillette, J., Gleitman, H., Gleitman, L., and Lederer, A. (1999). Human simulations of vocabulary learning. *Cognition*.

Gleitman, L. R., Cassidy, K., Nappa, R., Papafragou, A., and Trueswell, J. C. (2005). Hard words. *Language Learning and Development*, 1(1):23–64.

Hespos, S. J. and Spelke, E. S. (2004). Conceptual precursors to language. *Nature*, 430(6998):453.

Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., and Girshick, R. (2017). Clevr. In *CVPR*.

Kline, M., Snedeker, J., and Schulz, L. (2017). Linking language and events. *Language Learning and Development*, 13(1):1–23.

Landau, B. and Gleitman, L. R. (1985). Language and experience: Evidence from the blind child.

Levin, B. (1993). *English verb classes and alternations: A preliminary investigation*. University of Chicago Press.

Levin, B. and Rappaport Hovav, M. (2011). Lexical conceptual structure. *Semantics*.

Michotte, A. (1963). *The perception of causality*. Basic Books.

Muentener, P. and Carey, S. (2010). Infants causal representations of state change events. *Cognitive psychology*, 61(2):63–86.

Naigles, L. (1990). Children use syntax to learn verb meanings. *Journal of child language*, 17(2):357–374.

Naigles, L. R. and Hoff-Ginsberg, E. (1995). Input to verb learning: Evidence for the plausibility of syntactic bootstrapping. *Developmental Psychology*, 31(5):827.

Pinker, S. (1989). Learnability and cognition: The acquisition of argument structure.

Sadeghi, S. and Scheutz, M. (2018). Early syntactic bootstrapping in an incremental memory-limited word learner. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Smith, L. and Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*.

Steedman, M. and Baldridge, J. (2006). Combinatory categorial grammar.

Stevens, J. S., Gleitman, L. R., Trueswell, J. C., and Yang, C. (2017). The pursuit of word meanings. *Cognitive science*, 41:638–676.

Trueswell, J. C., Medina, T. N., Hafri, A., and Gleitman, L. R. (2013). Propose but verify: Fast mapping meets cross-situational word learning. *Cognitive psychology*, 66(1):126–156.

Zettlemoyer, L. and Collins, M. (2007). Online learning of relaxed ccg grammars for parsing to logical form. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*.

Privileged Computations for Closed-Class Items in Language Acquisition

Heidi R. Getz (hrg2@georgetown.edu) and Elissa L. Newport (eln10@georgetown.edu)
Center for Brain Plasticity and Recovery, Georgetown University Medical Center
Building D Suite 145, 4000 Reservoir Road NW
Washington, DC 20057 USA

Abstract

In natural languages, closed-class items predict open-class items but not the other way around. For example, in English, if there is a determiner there will be a noun, but nouns can occur with or without determiners. Here, we asked whether language learners' computations are also asymmetrical. In three experiments we exposed adults to a miniature language with the one-way dependency "if X then Y": if X was present, Y was also present, but X could occur without Y. We created different versions of the language in order to ask whether learning depended on which of these categories was an open or closed class. In one condition, X was a closed class and Y was an open class; in a contrasting condition, X was an open class and Y was a closed class. Learning was significantly better with closed-class X, even though learners' exposure was otherwise identical. Additional experiments demonstrated that the perceptual distinctiveness of closed-class items drives learners to analyze them differently; and, crucially, that the primary determinant of learning is the mathematical relationship between closed- and open-class items and not their linear order. These results suggest that learners privilege computations in which closed-class items are predictive of, rather than predicted by, open-class items. We suggest that the distributional asymmetries of closed-class items in natural languages may arise in part from this learning bias.

Keywords: language acquisition; statistical learning; computational mechanisms; morphosyntax; function words; closed-class items

Introduction

In natural languages an important contrast is between open class lexical items—for example, nouns or verbs—and closed class or function items—for example, *is* or *the*.[1] Open class categories like noun or verb contain many members and typically carry the important lexical content of the sentence. In contrast, closed class items, which are used to mark grammatical functions of other words, are typically very short, few in number, are each used with high frequency, and occur in predictable positions in their phrases. For example, English marks definiteness with the article *the*, which is one of the most frequent words in the language and always occurs before its noun. There is wide variation in the distribution of functional items across languages: in contrast to English, definiteness in Amharic is marked on lexical items in a particular structural position and can attach to nouns, adjectives, numerals, or even verbs depending on sentence structure (Kramer, 2010). The distribution of closed-class items is always predictable in certain ways, but learners must do a substantial amount of distributional analysis in order to learn the particular patterning of closed-class items in their language. The goal of the present paper is to explore the computational mechanisms that enable language learners to do this.

From previous research we know that closed-class items draw special attention from language learners. Infants can identify them on the basis of correlated distinctive phonological, prosodic and distributional properties such as short duration, light syllable structure, and high frequency (Shi, Morgan, & Allopenna, 1998; Shi, Werker, & Morgan, 1999), and children begin to represent these items long before producing them (Shafer, Shucard, Shucard, & Gerken, 1998; Shi, Werker, & Cutler, 2006). Early attention to closed-class items could facilitate other aspects of language acquisition. For example, since these items often occur at grammatically important parts of the sentence (e.g., phrase boundaries), focusing on them could help learners acquire grammatical structure. There is substantial empirical support for this idea, known as the Anchoring Hypothesis (Mintz, 2006; Morgan, Meier, & Newport, 1987; Valian & Coulson, 1988; Zhang, Shi, & Li, 2015).

However, it is not yet clear what computational mechanisms underlie learners' distributional analyses, once they have noticed closed-class items. The literature on statistical learning has not focused particularly on closed-class items; and only a few studies identify specific statistical computations that learners might draw on. These studies have revealed, for example, that learners can compute transitional probabilities to find word boundaries (Aslin, Saffran, & Newport, 1998; Saffran, Aslin, & Newport, 1996; Saffran, Newport, & Aslin, 1996) and to acquire grammatical phrases (Thompson & Newport 2007). Despite this progress, we are only beginning to identify the computational mechanisms underlying many aspects of language acquisition. It thus remains a mystery how learners manage to sort out patterns as complicated as (for example) Amharic definiteness. What kind of computations would a learner need to perform in order to acquire this type of pattern?

Consider the statistical information about closed-class items that is present in learners' input. As already noted, these items generally do not independently contribute semantic

[1] The terms 'functional item' and 'closed class' are often used interchangeably. We adopt the terminology of closed and open classes because these terms more readily apply to our miniature languages.

meaning; rather, they specify the grammatical properties of the meaning-bearing elements (the lexical categories). This role gives closed-class items a highly predictable syntactic context. For example, *the* indicates that its noun refers to a specific referent identifiable in context and therefore must appear with a corresponding noun, never alone. In statistical terms, the probability of seeing a noun, given that there is a determiner, is 100%. The reverse is not true, however, since nouns can occur in a variety of grammatical contexts, with or without determiners.[2]

The statistical asymmetry in the distribution of closed- and open-class items is especially interesting in light of the recent emphasis in syntactic theory on the role of functional categories in sentence structure (see Rizzi & Cinque, 2016 for discussion and historical context). Increasingly, linguists have argued that properties of closed-class items determine the behavior of other words in the sentence. This extends beyond the presence of certain open-class categories to their positions in the sentence as well. To illustrate, consider the pattern of verb placement in French. Lexical verbs such as "eat" (*mang-*) can either precede or follow the negative marker (*pas*) depending on whether the verb is morphologically finite, as in *tu manges pas?* ("You're not eating?), or non-finite as in *tu vas pas manger?* ("You aren't going to eat?"). Linguists capture this contingency between finiteness and verb position by positing that the abstract features of finite and non-finite morphemes are represented in different positions in the sentence. If there is finite morphology, there will be a verb and that verb will occur in the "finite" position (pre-negation). In this way the presence and location of verbs is determined by the kind of morphology that occurs in the sentence.

Of course, linguists' analyses are intended to be formal mathematical descriptions of sentence structure, and not necessarily claims about the psychological representation of sentences. However, this kind of analysis demonstrates an important empirical point: *regularities of word order and word form can be stated as restrictions on the distribution of closed-class items*. Consider now the problem of distributional learning. One way to begin learning, given this view from syntactic theory, would be to identify closed-class items (for example, based on their salient perceptual properties) and then proceed to learn their distribution. Because this distribution is asymmetrical—closed-class categories always predict but are not predicted by open-class categories—the computations that learners perform could also be asymmetrical. Learners need to learn what a closed-class item *predicts*—the presence of other categories, the placement of words, and so on—but they need not expend any effort finding distributional patterns that a closed-class item is *predicted by*, because there are none.[3]

Here we explored the possibility that learners privilege computations in which closed-class items are *predictive of* open-class items over computations in which they are *predicted by* open-class items. In Experiment 1 we exposed adult learners to a miniature language containing a one-way grammatical dependency between two form-class categories, X and Y. When an X word was present, a Y word always had to be present as well, but Y words could occur with or without X words ("if X then Y"). This is mathematically like the relationship between determiners and nouns in English. In two contrasting conditions, we assigned different types of words to the X and Y categories. In one condition the predictive category (X) was a closed class (short, monosyllabic, and containing only one item, *ka*), while in the other condition the predictive category was an open class (mono or disyllabic and containing three possible lexical items). Learning was better when X was closed-class, suggesting that learners' computations are biased: they identify patterns where closed-class items are predictive of open-class items more readily than the reverse. Additional experiments demonstrated that learners analyze closed-class items differently because they are perceptually distinctive (Experiment 2) and that learning outcomes are driven by the mathematical relationship between closed- and open-class items and not their linear order (Experiment 3). Together, the results suggest that learners analyze closed-class items in certain biased ways, searching preferentially for the kinds of patterns that exist in natural languages. In the Discussion we return to the question of why learners should be biased in this way. We do not mean to suggest that they know in advance about languages in particular, but rather that their computational biases may shape languages to be structured in this way.

Before proceeding, it is important to clarify a component of our experimental design. The artificial language that we created for these experiments is not very language-like. The experiments are focused on a specific computational question about how learners analyze closed and open-class items. To test our hypothesis, it was necessary to design a language that could only be learned by computing the precise mathematical relationship between two specific terms (X and Y). Therefore X was the only category whose distribution with respect to other words was constrained; all other words in the language appeared and disappeared freely, which is unlike the more constrained sentence structure of natural languages. This

[2] In some cases, predictiveness goes both ways (e.g., in French, all non-proper nouns require determiners). Nonetheless, computing how often determiners are accompanied by nouns will *always* reveal a pattern, whereas the reverse computation only sometimes will. Thus analyzing closed-class items as predictive of open-class items is the most effective way to discover linguistic patterns.

[3] Of course closed-class items do not appear randomly in sentences. Their presence is determined by the semantic meaning that the speaker wishes to express. The learner does eventually need to learn which meanings go with which forms, but this is a separate and somewhat uncorrelated problem. As the comparison between Amharic and English definiteness marking illustrates, learning that a given form means "definite" does not tell the learner where, distributionally, that form occurs, nor does learning the distribution of a form reveal its meaning (e.g., both definite and indefinite articles precede nouns in English). Both learning problems are important, but we are concerned here only with the distributional one.

experimental design allowed us to test empirically whether learners' computational analyses are biased in a certain way. If the results of these experiments do reveal such a bias, this would be motivation to explore how this bias affects the acquisition of more naturally structured languages—a line of work that is in progress.

Method

Participants

Three groups of sixteen adults from the Georgetown University community (age 18-28, mean=20.4) participated in this study. Two additional participants' data did not save due to an error.

Description of the miniature language

The design of the language used in all three experiments is summarized in Figure 1. The word order was AXYBC, where each letter represents a form-class category. All categories were optional, with the constraint that only up to three categories could be omitted per sentence (i.e. sentences must each have at least two words). The fixed and consistent rule of the language was that if X was present, Y had to be present ("if X then Y"). Thus every sentence with X also contained Y, but sentences with Y did not have to contain X. Note that this dependency is defined in terms of the conditional relationship, not the linear order, of X and Y. In Experiment 1, X preceded Y while in Experiment 3 X followed Y; this did not change the conditional relationship between the two terms.

In each experiment we created different versions of the language in order to ask whether learning this conditional relationship between X and Y depended on which of these terms was a closed-class or an open-class category. None of the words had any meaning, so this contrast was defined by the number of words in each category and the phonological properties of those words. Each experiment had a condition where X was closed class and Y was open class (Closed X) and a contrasting condition where X was open class and Y was closed (Open X). Across experiments we varied the phonological properties of the closed-class item and the linear order of X and Y.

Experiment 1 In this experiment the closed-class category contained a single item *ka*, which had several properties common to closed-class items in English: it was short, lacked a coda or consonant clusters, and was high frequency by virtue of being the only member of its form class. Each open-class category contained three words that were a mixture of mono- and disyllabic forms. All words in the language, including the closed-class item, carried stress (i.e., *ka* was not prosodically dependent on any other item). In the Closed X condition, the X category contained *ka* and Y contained *lapal, tombur,* and *zup*. Thus the closed-class item *ka* predicted any of these three open-class items. In the contrasting Open X condition, X contained *lapal, tombur,* and *zup* and Y contained *ka*. Here the closed-class item *ka* is predicted by each of three open-class items.

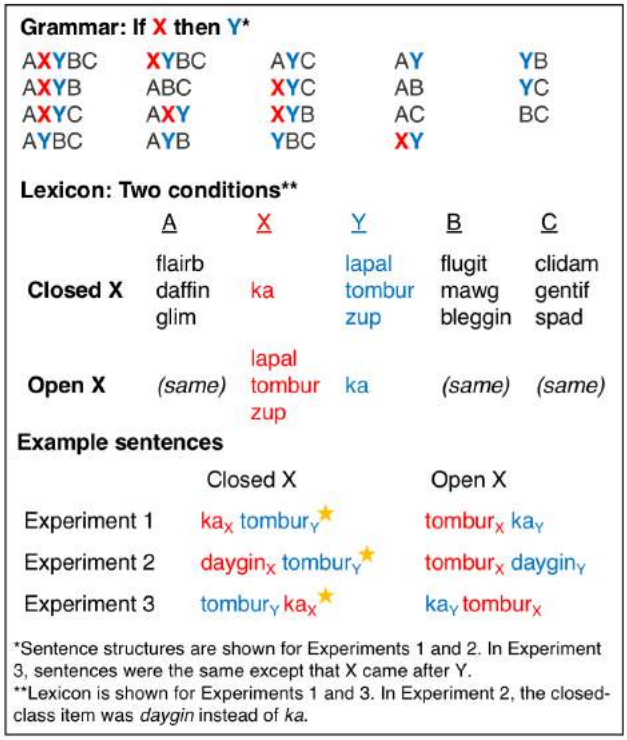


Figure 1: Design of the miniature languages used in Experiments 1-3. The critical feature of all languages is a one-way dependency between X and Y: every sentence with X also contained Y, but Y occurred without X. In Experiments 1 and 2, X came before Y (XY); in Experiment 3, X came after Y (YX). Each experiment had a condition where X was closed class and Y was open (Closed X) and a contrasting condition where X was open class and Y was closed (Open X). If learners are biased to analyze closed-class items as predictive, learning should always be better in the Closed X condition (marked with yellow stars).

Other than the specific lexical items in the X and Y categories, the two languages were identical. In both languages, sentences with X must also contain Y, while sentences with Y may or may not contain X. Because either X or Y is *ka*, learners in both conditions had an "anchor" for their distributional analyses. In both conditions, the predictive category (X) comes before the category it predicts (Y); this linear order was like subjects' native language, English, where (for example) determiners precede nouns. (In Experiment 3 we reversed the linear order such that the predictive category came last, as in languages like Japanese.) At a lexical level, in both conditions the dependency involved exactly one closed-class item and three open-class items; acquiring the dependency required computing exactly three word-level forward transitional probabilities (either X1-Y1, X1-Y2, X1-Y3 in the Closed X condition or X1-Y1, X2-Y1, X3-Y1 in the Open X condition). Our manipulation did of

course create different statistical patterns at the item level. In the Closed X condition, each of the three possible X-Y sequences had a transition probability of 0.33, whereas in the Open X condition each X-Y sequence had a transition probability of 1.0. Thus the item-level transition probabilities cued the XY unit more strongly in the Open X condition.

Because these languages are structurally identical except for the closed-open class contrast for X and Y, learning outcomes will differ only if learners' computational analyses treat closed-class and open-class items differently. If learners preferentially analyze closed-class items as predictive, they should easily learn "if X then Y" in the Closed X condition, where *ka* predicts an open-class category; but they should struggle in the Open X condition, where *ka* is predicted *by* an open-class category. Alternatively, if learners analyze closed-class and open-class items similarly, learning outcomes will be equivalent across conditions.

Experiment 2 Part of the hypothesis is that the distinctiveness of closed-class items drives learners to analyze them differently. In Experiment 2 we tested this by making the closed-class item less distinctive. Here the closed-class item (*daygin*) was disyllabic, carried initial stress, and had a closed final syllable, making it phonologically like the open-class words in the language; its only distinguishing property is its high frequency. If distinctiveness of *ka* drove learning outcomes in Experiment 1, learning should be weakened in Experiment 2.

Experiment 3 In Experiments 1 and 2 the Closed X condition was superficially like English: the closed-class item came before the open-class item (Figure 1). English does also have closed/open dependencies where the closed-class item comes last (e.g., walk + s), but only in morphology. Therefore better learning for Closed X in Experiments 1 and 2 could be due to a preference in native speakers of English for syntactic phrases where frequent words come first (cf. Gervain et al., 2013). In Experiment 3 we changed the word order of the language so that Y preceded X. Now the Open X condition is superficially more like English (frequent word first), whereas the Closed X condition is superficially like Japanese (frequent word last). If learning outcomes in Experiments 1 and 2 reflect superficial word order biases, then the results of Experiment 3 should be opposite to those of Experiment 1, with better learning for Open X. However, if learning outcomes depend on the structural relationship between closed-class and open-class items rather than superficial linear order, results should be similar to those of Experiment 1: Closed X should learn "if X then Y" and Open X should fail.

Materials

We generated a 38-sentence exposure set by selecting two sentence types for each of the 19 structures. The sentence structures were always the same across conditions and experiments, but the actual sentence strings differed across conditions and experiments according to the lexical items in the X and Y categories and the linear order of X and Y. Sentence sound files were created by concatenating individually recorded words (spoken by a female native speaker of English) with 50 msec of intervening silence. The 38-sentence exposure set was presented 16 times as part of a 1-back task (see Procedure).

Procedure

Participants learned the language through a computer game. A robot "Bot" instructed participants to listen as an alien named Zooma practiced saying sentences in an alien language. After each sentence, participants pressed a button to indicate whether Zooma had just repeated herself. After exposure, participants began the test. Bot explained that Zooma would try to say each sentence two different ways, and the participants' job was to decide which one was better. The entire experiment lasted approximately 45 minutes.

Test

Learning was measured with a two-alternative forced-choice (2AFC) test. The structure of the test was identical across experiments. Specific test strings varied according to the vocabulary of the language. The target choice on each trial was always a grammatical complete sentence. The alternative was identical to the target except that either one word was changed, or the words were the same but in a different order. The test was designed to ask whether participants had acquired a very specific piece of knowledge: the precise conditional relationship between X and Y. In order to answer this question it was important to create test items on which all other distributional properties (e.g., bigram frequency) were controlled. Only two types of test items could be carefully controlled in this way, described below. Items with confounds (not scored) included four additional items testing XY constituency and 20 items testing the XY relationship within longer sentences. In addition, we included four items testing basic word order and six filler items in order to balance the frequency with which targets and foils for the critical XY trials appeared on the test. Results for these items are not described for space reasons, but they are generally consistent with the results here.

There were two trials for each of the critical test item types. One item type served as a constituency test: XY was compared to YB (Experiments 1,2) or XB (Experiment 3). Both choices are legal two-word sequences (in Experiments 1 and 2, both choices are also complete sentences). They have the same relative frequency in learners' input, and are medial bigrams in the basic sentence structure (AXYBC (Experiments 1,2) or AYXBC (Experiment 3)). However, X and Y are related grammatically whereas the elements in the foil sequence are not. A preference for XY would indicate that participants represent this grammatical relationship. A second item type (AY vs. AX) tested whether participants learned that X predicts Y, but not the reverse. In Experiments 1 and 2, both choices are legal two-word sequences and occurred in learners' input with the same relative frequency; the two sequences had exactly the same forward transitional

probability (.36). However, only AY is a complete sentence. AX is a grammatical sequence but not a complete sentence, since it contains X but not Y. If participants have learned that X predicts Y (but not the reverse), they should prefer AY over AX. In Experiment 3, the foil was no longer a grammatical sequence in the language. Thus, it should be relatively easier in Experiment 3 than in Experiments 1 and 2 for both conditions to do well on these test items. Accuracy on the test was measured as choice of the target sequence.

Results

In Experiment 1 we asked whether learning of "if X then Y" would be different when X was a closed versus open class. Figure 2 illustrates that the answer is clearly yes. In the Closed X condition, learners chose the target item much more often than learners in the Open X condition (Closed X: 84%, Open X: 53%, $t(14) = 3.16$, $p = .007$). The grammatical coherence of XY as a unit was identical in these two conditions, since X always perfectly predicts Y. Yet learning outcomes differed across conditions, indicating that learners analyze closed-class items as predictive of open-class items more readily than the reverse.

Part of our hypothesis was that learners analyze closed-class items differently because these items are distinctive. In Experiment 2 we tested this by making the closed-class item less distinctive: here it was high frequency but phonologically like the open-class words in the language. Learning in the Closed X condition in this experiment was weaker than in Experiment 1 (Figure 2) and was no longer significantly better than the Open X condition ($t(14) = 1.94$, $p = .07$). This supports our hypothesis: closed-class items are analyzed differently because they are distinctive.

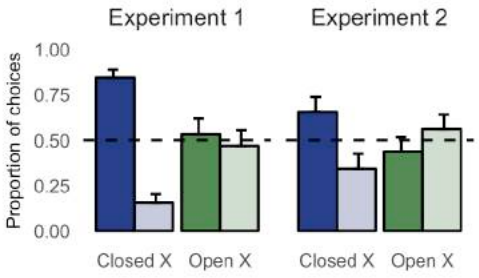


Figure 2: Choice of the target item (darker colors) or foil (lighter colors) on the 2AFC test in Experiments 1 and 2.

In Experiment 3 we asked: did the Closed X condition perform better because their language was superficially like English (the frequent item came first)? In this experiment Y came before X, but X still predicted Y. Thus the Open X condition was superficially like English. However, the Open X learners still struggled to learn "if X then Y" (59% correct, not significantly different from chance: $t(7)=2.05$, $p = .08$). In contrast, the Closed X condition continued to perform well above chance (72% correct, $t(7)=2.97$, $p = .02$), even though their language was superficially *unlike* English and more like an unfamiliar language, Japanese. These results demonstrate that the primary determinant of learning is the mathematical relationship between closed-class and open-class items and not their linear order.

The results just reported are collapsed across two types of items: the constituency test (YX vs. XB) and the predictive direction test (AY vs. AX). Based on these collapsed results, learning in the Open X condition appears to be slightly better than expected. Although learners were not significantly above chance, the difference was marginal, and accuracy was numerically higher than in Experiment 1 (59% vs. 53%). An analysis of results for the two different test item types provides some insight. In Experiments 1 and 2, results were equivalent across item types. However, the Open X condition in Experiment 3 showed a different pattern (Figure 3): learners passed the constituency test (88% correct), but were numerically *below* chance on the predictive direction test (33% correct). A 2-way mixed ANOVA over condition and trial type confirmed this impression statistically: there was no main effect of condition ($F(1) = 2.07$, $p = .17$), but there was a significant main effect of test item type ($F(1) = 4.77$, $p = .047$), and—importantly—a significant interaction between condition and test item type ($F(1) = 7.45$, $p = .02$), driven by a preference in the Open X condition for the ungrammatical sequence *AX over AY.

Why would the Open X condition perform so poorly on the AY/*AX items? Based on the raw statistical properties of learners' input, these items should be easy: *AX is not a complete sentence or even a legal sequence, whereas AY is both. In fact, an explanation for these results is provided by our hypothesis: that learners attend to (or search for) some statistical patterns over others, prioritizing patterns in which closed-class items are predictive. Such a bias would lead learners in the Open X condition to initially analyze their closed-class item Y as predictive of another item. Statistically, the item that Y best predicts is X (the probability that a sentence contains X, given that it contains Y, is .53). Thus, a preference for *AX could reflect an incorrect hypothesis that the conditional relationship between X and Y is reversed ("if Y then X"). This generalization is not consistent with learners' input, but it is consistent with the patterning of closed-class items in natural languages.

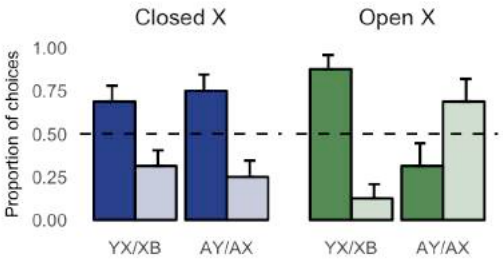


Figure 3: Choice of the target item (darker colors) or foil (lighter colors) on the two item types of the 2AFC test in Experiment 3. Participants in Closed X still learned, even though the linear order was opposite English. In contrast, participants in Open X succeeded on the constituency test (YX/XB) but not the predictive direction test (AY/AX), apparently having incorrectly analyzed Y as predictive of X.

Discussion

In three experiments, we showed that adults easily acquire a grammatical dependency "if X then Y" when X is a closed-class, but fail to acquire the same dependency when X is an open class (Experiment 1). Successful learning of "if X then Y" is facilitated by the distinctive perceptual properties and high frequency of the closed-class item (Experiment 2). Importantly, the primary determinant of learning is the mathematical relationship between closed-class and open-class items and not their linear order (Experiment 3). These results suggest that learners privilege computations in which closed-class items are predictive of open classes—the same computations that are most relevant for natural language dependencies.

We emphasize that, within each experiment, the Closed X and Open X conditions had exactly the same statistical evidence for the rule "if X then Y": class X always perfectly predicted class Y. Furthermore, learners in contrasting conditions were always exposed to the same number of lexical items, sentence structures, and sentence types. To acquire the XY rule, learners needed to compute exactly three transitional probabilities, and contrasting conditions within each experiment always contained the same linear direction of the X-Y relationship (forward for Experiments 1 and 2, backward for Experiment 3). Despite this mathematical equivalence, learning was always better when X was a closed class. If participants were computing statistics over items rather than classes, the results are even more striking: in that case, participants learned three low-probability dependencies with a predictive closed-class item (*ka*→*lapal*, *ka*→*tombur*, *ka*→*zup*) more easily than three high-probability dependencies with a predictive open-class item (*lapal*→*ka*, *tombur*→*ka*, *zup*→*ka*).

These results indicate that—whether learners computed statistics over classes or items—their distributional analyses are biased. Rather than tracking all possible pairwise transitional probabilities involving a closed-class item, learners apparently analyze closed-class items asymmetrically, more easily learning patterns in which a closed-class item is predictive *of* another element than patterns in which it is predicted *by* another element.

Conclusion

The original idea of the Anchoring Hypothesis (Valian & Coulson, 1988) was that, because closed-class items tend to occur at grammatically important points in the sentence, focusing on them could help learners acquire grammatical structure. Our results add a computational component to this approach. Our hypothesis is that, because closed-class items are noticed first, due to their distinctive phonological properties and their high frequency, these will be the constant terms in learners' computations; other patterns are learned and represented relative to them.

A learning mechanism that operates in this way would ultimately represent a broad range of language patterns in terms of the distribution of a small set of closed-class items. As we pointed out in the Introduction, this is increasingly the way that language patterns are described by modern syntactic theory as well. The results of our experiments suggest that human languages may acquire this type of structure at least in part as a consequence of computational biases in the human language learner. This account is appealing because, if correct, it would explain the privileged role of closed-class items in human linguistic representations without positing that these representations are innate. However, it is important to note that in all three experiments, learners' preferred conditional relationship had the same abstract structure (though not always the same linear order) as the closed/open dependencies in all natural languages, including English. It is difficult to rule out the possibility that learning was affected by participants' experience with this abstract property of natural languages; even infants have experience with closed-class items (cf. Shi et al., 1998). Studies of learning in a non-linguistic domain could be informative (cf. Saffran, Johnson, Aslin, & Newport, 1999).

Our results raise several other important questions. First, what about closed-class items that behave differently? Our claim is that learners analyze closed-class items as predictive of open-class items, and that this approach is useful because it matches the abstract structure of grammatical dependencies in natural languages. However, there are exceptions to this pattern. For example, pronouns like *him* do not depend on open-class items the same way that articles do. Interestingly, pronouns are also special in other ways (Chomsky, 1980). The proposed computation could be useful not only for discovering predictive dependencies, but also—when this analysis fails to uncover a dependency—for flagging elements with a more complex grammatical distribution. Second, we must also ask whether this computational bias is present in children, who are the real natural language learners. Our results in ongoing work with child participants suggest that they do share this bias. This in turn raises a puzzle: if children organize their languages around closed-class items, why do they not produce these words in their own speech for several years? The available evidence suggests that children do indeed process closed-class items early, despite omitting them in production (Gerken et al., 1990; Shi et al., 2006; Zhang et al., 2015). Future work is required to understand the discrepancy between what children represent and what they initially produce. Finally, we need to test our predictions on materials that are more like natural languages than what we have studied here. In order to test our computational predictions most cleanly, the languages in these experiments were unlike natural languages in several ways: all of the categories other than X and Y were optional, there was only a single grammatical phrase (XY), and none of the words had any meaning. We are in the process of testing whether learners privilege the same types of computations in the acquisition of miniature languages that are more natural. If so, we can ask what kinds of natural language patterns can be acquired and represented using these privileged computational mechanisms, and to what extent these learning mechanisms explain why these patterns come to exist in languages of the world.

Acknowledgements

This research was supported by NIH grant HD037082 to E. Newport, by the Feldstein Veron Fund for Cognitive Science, and by Georgetown University Center for Brain Plasticity and Recovery.

References

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, 9(4), 321–324.

Chomsky, N. (1980). On binding. *Linguistic Inquiry, 11*(1), 1–46.

Gerken, L., Landau, B., & Remez, R. E. (1990). Function morphemes in young children's speech perception and production. *Developmental Psychology*, *26*(2), 204–216.

Gervain, J., Sebastián-Gallés, N., Díaz, B., Laka, I., Mazuka, R., Yamane, N., … Mehler, J. (2013). Word frequency cues word order in adults: cross-linguistic evidence. *Frontiers in Psychology*, *4*, 689.

Kramer, R. (2010). The Amharic definite marker and the syntax–morphology interface. *Syntax*, *13*(3), 196–240.

Mintz, T. H. (2006). Finding the verbs: Distributional cues to categories available to young learners. In *Action meets word: How children learn verbs* (pp. 31–63). New York: Oxford University Press.

Morgan, J. L., Meier, R. P., & Newport, E. L. (1987). Structural packaging in the input to language learning: contributions of prosodic and morphological marking of phrases to the acquisition of language. *Cognitive Psychology*, *19*(4), 498–550.

Rizzi, L., & Cinque, G. (2016). Functional categories and syntactic theory. *Annual Review of Linguistics*, *2*(1), 139–163.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928.

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, *35*(4), 606–621.

Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition, 70*(1), 27–52.

Shafer, V. L., Shucard, D. W., Shucard, J. L., & Gerken, L. (1998). An electrophysiological study of infants' sensitivity to the sound patterns of English speech. *Journal of Speech, Language, and Hearing Research, 41*(4), 874–886.

Shi, R., Morgan, J. L., & Allopenna, P. (1998). Phonological and acoustic bases for earliest grammatical category assignment: a cross-linguistic perspective. *Journal of Child Language*, *25*(1), 169–201.

Shi, R., Werker, J. F., & Cutler, A. (2006). Recognition and representation of function words in English-learning infants. *Infancy*, *10*(2), 187–198.

Shi, R., Werker, J. F., & Morgan, J. L. (1999). Newborn infants' sensitivity to perceptual cues to lexical and grammatical words. *Cognition*, *72*(2), B11–21.

Thompson, S. P., & Newport, E. L. (2007). Statistical learning of syntax: The role of transitional probability. *Language Learning and Development*, *3*(1), 1–42.

Valian, V., & Coulson, S. (1988). Anchor points in language learning: The role of marker frequency. *Journal of Memory and Language*, *27*(1), 71–86.

Zhang, Z., Shi, R., & Li, A. (2015). Grammatical categorization in Mandarin-Chinese-learning infants. *Language Acquisition*, *22*(1), 104–115.

Cross-cultural differences in playing centipede-like games with surprising opponents

Sujata Ghosh
Computer Science Unit, Indian Statistical Institute, Chennai, India

Rineke Verbrugge
Department of Artificial Intelligence, University of Groningen, Groningen, the Netherlands

Harmen de Weerd
Research Group User Centered Design, Hanze University of Applied Sciences, Groningen, the Netherlands

Aviad Heifetz
Department of Management and Economics, The Open University of Israel, Raanana, Israel

Abstract

In this paper, we study cross-cultural differences in strategic reasoning in turn-taking games, as related to game-theoretic norms as well as affective aspects such as trust, degrees of risk-taking and cooperation. We performed a game experiment to investigate how these aspects play a role in reasoning in simple turn-based games, known as centipede-like games, across three cultures, that of The Netherlands, Israel and India. While there is no significant main effect of nationalities on the behaviour of players across games, certain unexpected interactive effects are found in their behaviour in particular games.

Keywords: intercultural differences; game theory; reasoning in games; trust and trustworthiness; risk considerations; cooperation

Introduction

Cognitive science is not only concerned with universal patterns of cognition, but also variations in those patterns, induced by relevant factors. As D'Andrade (1981) and Levinson (2012) argue, studying variation, and in particular cross-cultural differences, provides important insights. In this article, we study cross-cultural differences in strategic reasoning in turn-taking games, as related to affective aspects such as trust, degrees of risk-taking and cooperation. To this end, we performed a game experiment in three countries: The Netherlands, India and Israel.

Cross-cultural differences and games

It has been known for a long time that in turn-taking games of perfect information, people in general do not act exactly according to the prescriptions of game theory, which are based on the common knowledge of the rationality of participants (Aumann, 1995; Nagel & Tang, 1998). There has been a lot of interest in the possible differences between people from different countries with respect to adherence to game-theoretic predictions (Camerer, 2011). Note that national cultures should not be interpreted in an essentialist way: Cultural tendencies can be induced by incentives (Peysakhovich & Rand, 2016). For our experiments, we are mainly interested in four aspects:

- adherence to strategies defined in game theory, namely, forward versus backward induction;
- degree of trust and degree of trustworthiness;
- degrees of risk-taking;
- cooperative versus competitive tendencies.

As far as we know, our experiment is the first one to compare adherence to forward induction versus backward induction reasoning between different nationalities. The notions of forward and backward induction are explained in the next subsection on games.

With respect to trust and cooperation, however, there have been a number of previous cross-cultural studies, using both games in which participants meet an opponent only once and games in which they repeatedly interact with the same opponent (Roth et al., 1991; Ho & Weigelt, 2005; Henrich et al., 2005). Differences in trust, cooperativeness, and risk-taking between British and Japanese participants in turn-taking centipede games have been studied in Krockow et al. (2017).

Trust and trustworthiness Yamagishi & Yamagishi (1994) have distinguished two types of trust:

- *assurance-based trust* needed in relationships with high social certainty with an expectation of future interaction;
- *general trust* needed in encounters with strangers with low social certainty and low expectation of long-time future interaction.

Yamagishi & Yamagishi (1994) have also shown that different cultures score very differently on these two types: Assurance-based trust is high in cultures like Japan, as incentivized by long-time employment by the same company. In the United States and Great Britain, in contrast, high general trust corresponds with the prevalence of short-time employment and commerce with strangers. Based on the literature, we expect that trust for strangers is relatively low in India (where assurance-based trust is high, like in Japan) and high in The Netherlands (like in Great Britain), with Israel probably in between.

Cooperation, competition, and self-interest According to Hofstede (1991) (see the left part of Figure 1), Israel is an interesting mix between collectivist cultures such as India, which are expected to be more cooperative in nature,

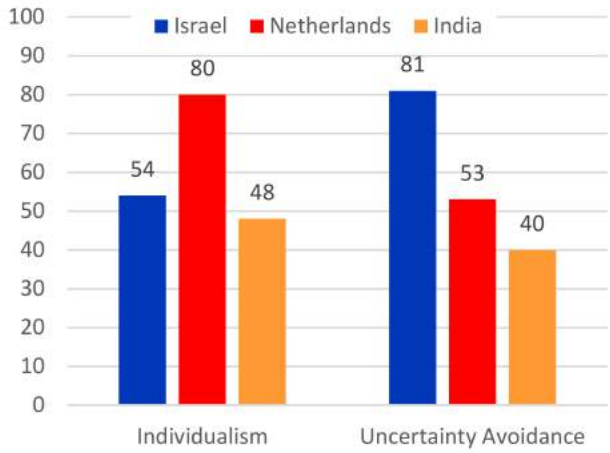


Figure 1: Individualism and uncertainty avoidance ratings of Israel, the Netherlands and India. The numbers for the countries were provided by the country comparison tool on Hofstede's website `https://www.hofstede-insights.com/`, based on the six dimensions distinguished by Hofstede (1991).

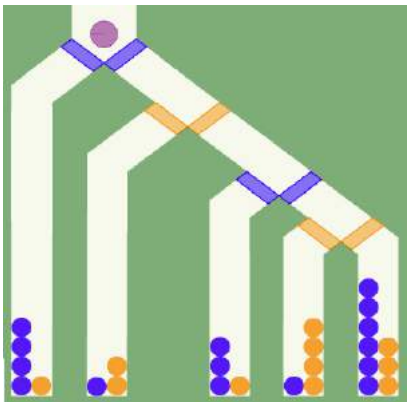


Figure 2: Marble Drop game. Players, assigned blue and orange, control the marble's course by opening the left or right trapdoor of their color once the purple marble arrives there. When the purple marble ends up in a certain bin, each player earns the marbles of their color in that bin. This example payoff structure corresponds to Game 1 of Figure 3 below.

and individualist ones such as the Netherlands, in which self-interested behaviour is more common.

Attitudes toward risk According to Hofstede (1991), people in Israel predominantly try to avoid uncertainty, while people in The Netherlands are rather neutral and people in India can handle uncertainty and risk most easily, see the right part of Figure 1.

The main focus of this paper is an experiment to investigate how the above-mentioned aspects play a role in reasoning in simple turn-based games, known as centipede-like games, across the three cultures. The games are introduced in the next subsection.

Games for the experiment

The participants in our experiments played a turn-based game called Marble Drop (Figure 2) against a computer opponent, and accordingly, we denote the two players by '*C*' and '*P*'. An important advantage of using computer opponents in experiments with turn-taking games is that the experimenter can control the strategies used by the computer opponent, which allows better interpretation of the participants' decisions. The choice of the Marble Drop games was inspired by (Halder et al., 2015; Ghosh et al., 2017; Verbrugge et al., 2018). These games can be visually represented as binary tree structures (Figures 3 and 4). The difference between Game 1 and Game 2 lies in the payoff of player *C* when choosing *a* at C_1. That payoff of player *C* after choosing *a* is also the only difference between Game 3 and Game 4. In addition, the only difference between Game 1 and Game 2 on the one hand and Game 3 and Game 4 on the other hand is the payoff of player *P* after choosing *h* at P_2. Since the structure of these games is reminiscent of a centipede, with its body extending from top left to bottom right, the games are termed as 'centipede-like' games.[1]

In the textbook approach of solving such turn-based games in game theory, players who are commonly known to be rational use the *backward induction* (BI) strategy (Perea, 2010): one should ignore previous information, and work backwards from the end of the game tree to reach a decision. For example, in the 'orange' player's last turn in the marble drop game (Figure 2), he has to decide between going to the left or to the right, for payoffs of 4 or 3 orange marbles, respectively. Using BI, because 4 is more than 3, he chooses to go left, delivering the outcome pair (1,4): 1 for the blue player, 4 for the orange player. One can then continue backwards to compare the left and right choices in the blue player's second turn: going right gives (1,4) while going left gives (3,1); because 3 is more than 1, the blue player would choose to open the left blue trapdoor. One then continues to reason backwards to compare the actions in the orange player's first turn, where the outcome is (1,2) when playing left and (3,1) by playing right. One assumes that, 2 being more than 1, the orange player chooses to open the left orange trapdoor. Finally, one compares the actions in the blue player's first turn, where going left leads to (4,1) and going right leads to (1, 2). Because 4 is more than 1, the blue player will choose to open the left trapdoor to obtain 4 points. Note that playing rationally by BI does not necessarily lead to the outcome with the highest sum of players' payoffs – that would have been achieved by both players choosing to open their right trapdoors at all four decision points and ending up with a combined payoff of 6+3.

The 'surprising opponent' component of these experimental games comes from the fact that player *C* (blue) when starting the game does not always play according to the strategy

[1]The games we consider do not always comply with the conditions on payoffs of the original centipede game (Rosenthal, 1981).

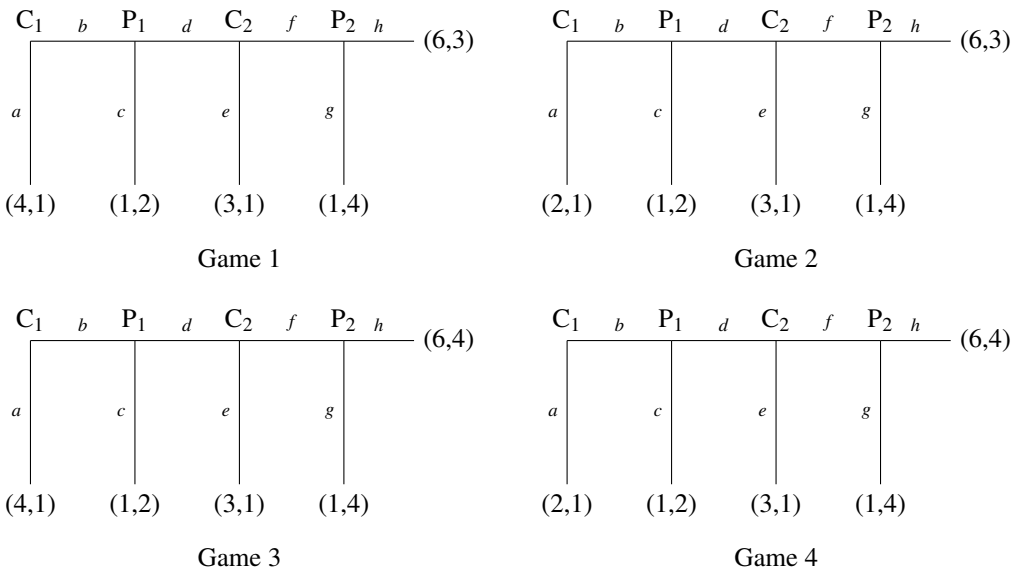


Figure 3: Collection of the main games used in the experiment. The ordered pairs at the leaves represent payoffs for the computer (C) and the participant (P), respectively.

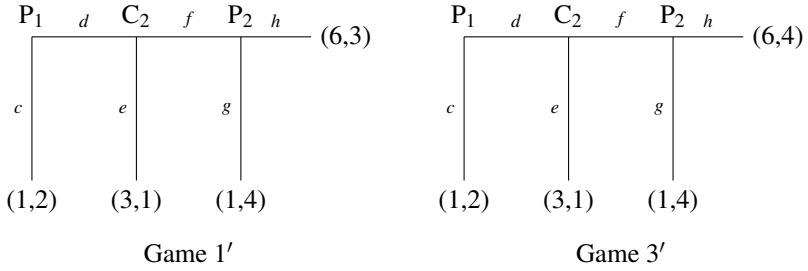


Figure 4: Game $1'$ corresponds with the parts of Games 1 and 2 from P_1 onwards. Game $3'$ corresponds with the parts of Games 3 and 4 from P_1 onwards.

described above. Note that in Games 1-4 in Figure 3, the BI strategy suggests for player C to choose a at the first decision node. In our experiment, the computer player often goes to the right to give player P (orange) a turn to move in the game. The orange player may or may not take into account this 'surprising' move of the blue player while considering his future moves. He can disregard his opponent's past move and play as if he is playing a 'new' game from the current turn and continue according to the BI strategy. Such players would play as if they were playing Game $1'$ or Game $3'$ (see Figure 4). Alternatively, the orange player can play according to a completely different strategy as described below.

In *forward induction* (FI) reasoning, a player takes into account his opponent's past moves and tries to rationalize the past behaviour in order to assess that opponent's future moves (Perea, 2010). We consider a particular kind of forward induction reasoning here, namely, *extensive-form rationalizability* (Pearce, 1984). The underlying idea is that when a player is about to play at a decision point that has been reached due to some strategy of the opponent that is not consistent with common knowledge of rationality for each of the players, the player may still rationalize the opponent's past behaviour. For example, suppose that the participant P has the opportunity to play at her first pair of orange trapdoors in the marble drop game (Figure 2, corresponding to Game 1 of Fig. 3), which has been reached because the computer C has chosen to open the right blue trapdoor. This first move is inconsistent with the choice determined by the assumption of rationality of both players (see BI example above), that is to open the left blue trapdoor. The participant might reason as follows: "The computer will definitely refrain from choosing the left trapdoor at his next choice getting 3 marbles, because he could have got more (4) marbles had he chosen the left trapdoor in his first decision node. He must be thinking that I would choose the right trapdoor in my second decision node if it is reached, in which case he would get 6 marbles, which is more than 4. So, if I choose the right orange trapdoor now, he will choose the right blue trapdoor at his next choice, and then I could choose the left trapdoor which would give me 4 marbles, more than the 2 marbles I would get if I chose the left trapdoor now." According to extensive form rationalizability, it would therefore be irrational for a computer opponent C to choose b at C_1 only to choose e later at C_2 in Game 1 and in Game 3. However, it would be possible for the computer opponent to behave in this way in Game 2 and in Game 4. Similarly, extensive form rationalizability would

also consider it rational for a computer opponent C to play e at C_2 in Game $1'$ and in Game $3'$.

Ghosh et al. (2017) investigated whether people are inclined to use forward induction in centipede-like games, rather than backward induction, in an experiment performed in The Netherlands. They found that in the aggregate, participants showed forward induction behaviour in response to their opponent surprisingly deviating from backward induction behaviour right at the beginning of the game. However, participants' verbalized strategies most often mentioned their own attitudes towards risk and those they assigned to the computer opponent, sometimes in addition to considerations about cooperativeness and competitiveness, rather than game-theoretic considerations. In our current study, we investigate variations in reasoning strategies across nationalities.

Hypotheses

We first note that in all these games, we are trying to study participants' reasoning methods in terms of their moves (i.e., participants' behaviour). There are certain challenges regarding linking behaviour in games to the underlying reasoning processes of players. For example, one can explain a given action in a turn-based game with different reasoning patterns. In this paper, we interpret the moves with respect to particular reasoning patterns they represent, namely, game-theoretic reasoning strategies such as backward and forward induction reasoning as well as strategies influenced by affective aspects like trust, degrees of risk-taking and cooperation.

In Game 1 and Game 3, the action c at P_1 would suggest backward induction reasoning performed by the participant. In addition, the same action might also suggest uncertainty avoidance or risk-averseness in the participant. On the other hand, the action d might suggest a risk-taking attitude in addition to extensive-form rationalizable (forward induction) reasoning. Taking note of such variations in reasoning patterns, we now formulate hypotheses about the cultural differences that we expect, based on the relevant features discussed in the Introduction.

Backward versus forward induction, uncertainty avoidance and trust Taking a cue from the fact that at P's first decision point, the uncertainty avoidance action is the same as the backward induction reasoning action in all our experimental games, we argue that there is a link between these two reasoning patterns in the present context. Accordingly, because of highest uncertainty-avoidance we expect that backward induction reasoning is strongest in Israel, then the Netherlands, then India. So we expect the 'safe' choice of c (backward induction) at the first decision point P_1 in all the games of Figure 3 to be most prevalent in Israel, followed by The Netherlands, and least in India.

Looking more specifically at game items, the higher level of generalized trust in The Netherlands than the two other countries leads us to expect higher choices of d especially in Game 1 and Game 3, based on forward induction and/or trust that the other player will reciprocate and choose f at C_2.

Cooperation and trustworthiness With respect to self-interested goals, in Games 3, 4 and $3'$, choices g and h provide the same number of points to the participant, namely 4. Among these, g is the competitive choice (allowing only 1 point to C) and h the cooperative one (allowing 6 points to C). Based on the collectivist culture in India, we expect h to be chosen most in India (we expect more than 50 % h), followed by Israel (mix of collectivist and individualist), followed by The Netherlands (individualist).

Methods

The experiment was conducted at the Indian Statistical Institute in Kolkata, The Open University of Israel, and the Institute of Artificial Intelligence at the University of Groningen, The Netherlands. In each of the three countries, a (different) group of 50 Bachelor's and Master's students from several disciplines took part. That is, the experiment included 50 Indian students (44 male, mean age 24.0), 50 Dutch students (26 male, mean age 23.8), and 50 Israeli students (23 male, mean age 27.1).[2] The participants had little or no knowledge of game theory.

The tasks that the participants had to perform in these experiments are mentioned in Table 1. Participants were instructed by an experimenter at the university, who was also available for questions. The participants played the turn-based games through a graphical interface on the computer screen (Figure 2). Participant were informed that each round, they would play against a different computer opponent (C, blue). Each of these opponents would play according to some plan that was a best response to some plan of the participant. The participant's goal was that the marble should drop into the bin with as many orange marbles as possible. The computer's goal was that the marble should drop into the bin with as many blue marbles as possible. Before the experiment itself, participants played 14 games to familiarize them with the game and its controls, the colored marbles, and the turn-taking aspect of the game.

In some rounds of each game, the participants' were asked certain multiple-choice questions regarding the choices of their opponent: (i) "The computer just chose to go [direction computer just chose]. If you choose to go [direction corresponding to playing d], what do you think the computer would do next?" or, (ii) "The computer first chose to go [direction computer chose at its first decision point]. When you made your first choice, what did you think the computer would do next if you chose to go [direction corresponding to playing d]?" Three options were given regarding the likely choice of the computer: "I think the computer would most likely open the left side" or "I think the computer would most likely open the right side" or "Both answers seem equally likely". The first two answers translated to the moves e or f of the computer, respectively. In case of the third answer, we assumed that the participant was undecided regarding the

[2]We'd like to thank the experimenters Eric Jansen, Saikat Palit, Aviel Swissa and Stav Edry.

computer's next choice.

Participants were paid according to the number of marbles they gained in one of the experimental games, selected at random for each participant. Participants were paid proportionally to the number of marbles they gained (1-4), irrespective of the number of marbles gained by the computer opponent. The amounts were balanced across countries so that the minimum payout would be enough to go out for coffee, while the maximum amount would pay for going out for pizza.

For the current study, we compare data between the participants of India, Israel and The Netherlands, all of whom performed the same tasks.

Step 1	- Introduction to the experiment. - Instructions to the participants.
Step 2	Practice Phase: 14 marble drop games.
Step 3	- Experimental Phase: 48 marble drop games, divided into 8 rounds of 6 different games each, distinguishing factor being the pay-off structures. - Each of the 6 games of Figures 3, 4 occurs once in each round; the 6 games occur in a random order in each round. - Questions were asked about computer's behaviour in several rounds.
Step 4	Questions were asked at the end of the experiment regarding decisions at all nodes of a sample game.

Table 1: Steps of the experiment

Results

As mentioned in the description of the marble drop game, participants face up to two decision points, P_1 and P_2, when playing the games represented in Figures 3 and 4. The first is whether to stop the game by choosing c or continue playing by choosing d at their first decision point P_1. To determine to which extent nationality influences this decision, we performed logistic regression of their first decision on Game (1, 2, 3, 4, $1'$, $3'$), nationality (India, Israel, The Netherlands), and their interaction.

Trust versus uncertainty avoidance, forward versus backward induction

Figure 5 depicts the proportion of d choices in Games 1, 2, and $1'$. In addition, Table 2 shows the estimation results of logistic regression of the participants' tendency to choose d in these games. In this regression, Dutch nationality and Game 1 are taken as the base case scenario and each coefficient is read as a change in the likelihood of playing d when compared to a Dutch participant playing Game 1.

Table 2 shows that there is no significant main effect of nationality on the behaviour of players. On average, we therefore find no differences in the levels of trust and uncertainty avoidance across nationalities for their first decision. However, we do observe a significant main effect of Game 2. Recall that participants who engage in forward induction reasoning would be more likely to pick d in Game 1 than in Games 2 and $1'$. The results in Table 2 are consistent with forward induction reasoning, since the coefficients of Game $1'$ and Game 2 are both negative. Interestingly, only the difference between Game 1 and Game 2 is significant. That is, even though Game $1'$ and Game 1 provide participants with different information on their opponent's strategy, participant choices do not differ significantly.

Variable	Coefficient	z value
India	-0.035	-0.250
Israel	-0.334	-1.240
Game $1'$	-0.100	-0.489
Game 2	-0.609	-2.787**
Israel × Game $1'$	0.557	2.038*
Israel × Game 2	0.724	2.593**
India × Game $1'$	0.337	1.316
India × Game 2	0.510	1.875

Table 2: Estimated logistic regression coefficients for the proportion of d choices in Games 1, 2, and $1'$. Coefficients represent the difference in d choices compared to Dutch participants in Game 1. Significance at the 5% level and 1% level are indicated by * and **, respectively.

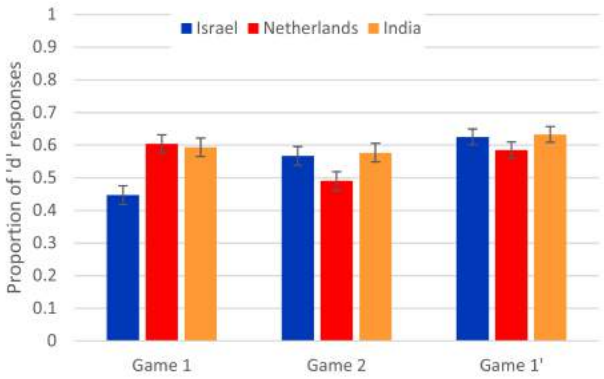


Figure 5: Proportion of d choices in games 1, 2, and $1'$ across nationalities. Whiskers indicate one standard error.

In addition, there is a significant interaction between Game 2 and Israeli nationality. Together, these results indicate that while Dutch participants are more likely to choose d in Game 1 than they are in Game 2, Israeli participants tend to choose d less in Game 1 than in Game 2. Thus, while some Dutch participants may have used forward induction, Israeli participants' behaviour does not show a lot of strategic reasoning per se.

Figure 6 shows the proportion of d choices in Games 3, 4, and $3'$. Table 3 shows the estimation results of the logistic regression for these games, where Game 3 and Dutch nationality are the base case scenarios. The table shows that only Game 4 has a coefficient that deviates significantly from zero, indicating that participants were less likely to choose d

Variable	Coefficient	z value
India	-0.373	-0.519
Israel	-0.601	-0.993
Game 3′	-0.184	-0.879
Game 4	-0.428	-1.989*
Israel × Game 3′	0.195	0.666
Israel × Game 4	0.117	0.249
India × Game 3′	0.162	0.559
India × Game 4	0.357	0.637

Table 3: Estimated logistic regression coefficients for the proportion of *d* choices in Games 3, 4, and 3′. Coefficients represent the difference in *d* choices compared to Dutch participants in Game 3. Significance at 5% level is indicated by *.

Variable	Coefficient	z-value
India	-0.942	-1.043
Israel	-1.714	-2.467*
Game 3	-0.344	-0.403
Game 3′	-0.279	-0.219
Israel × Game 3	0.276	0.138
Israel × Game 3′	0.097	0.442
India × Game 3	-0.516	-1.185
India × Game 3′	-0.744	-1.537

Table 4: Estimated logistic regression coefficients for the proportion of *h* choices in Games 3, 4, and 3′. Coefficients represent the difference in *h* choices compared to Dutch participants in Game 4. Significance at 5% level is indicated by *.

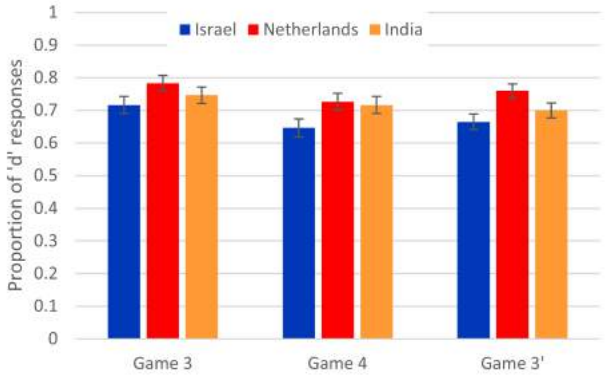


Figure 6: Proportion of *d* choices in Games 3, 4, and 3′ across nationalities. Whiskers indicate one standard error.

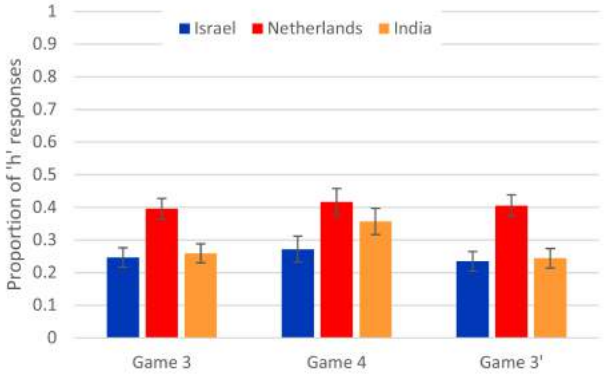


Figure 7: Proportion of *h* choices in Games 3, 4, and 3′ across nationalities. Whiskers indicate one standard error.

in Game 4 than they were in Game 3. Note that this is consistent with forward induction reasoning, which would lead a participant to be more likely to choose *d* in Game 3 than in Games 4 and 3′.

Similar to the findings presented in Table 2 for Games 1, 2 and 1′, Table 3 shows that none of the nationality-dependent coefficients differ significantly from zero. In particular, for the participants' first decisions in Games 3, 4 and 3′ , there appear to be no significant differences in trust and uncertainty avoidance across nationalities.

Competition, cooperation and trustworthiness

In addition to the decisions at the first decision point P_1, we performed a logistic regression on participants' choices at their second decision point P_2 to investigate differences in cooperation and trustworthiness. Since the choices of the participants affect their own payoffs in Game 1, 2, and 1′, our analysis of participant behaviour at decision point P_2 is limited to Games 3, 4, and 3′, in which their choice only affects the payoff of the computer opponent, their own payoff being 4 in all cases. Participants could choose the cooperative option *h*, which would yield the opponent a payoff higher than their own, or the competitive option *g*, which would leave the opponent with the lowest possible payoff.

Figure 7 depicts the proportion of *h* choices in Games 3, 4, and 3′. In addition, Table 4 shows the logistic regression results on the proportion of *h* choices, where Dutch nationality and Game 4 are taken as the base case scenarios. The results show no significant differences in the decision to choose *g* or *h* across games. That is, the interpretation participants have of the opponent's previous actions do not appear to affect participant choices at the second decision point significantly.

While Table 4 shows that the differences between Dutch and Indian participants are not significant, Israeli participants were significantly less likely to choose *h* than Dutch participants. Moreover, Figure 7 shows that across Games and nationalities, participants were more likely to choose the option that would yield the opponent a lower payoff. Overall, participant behaviour can therefore be described as competitive.

Discussion and conclusion

We hypothesized that at their first decision point, participants from Israel would show uncertainty avoidance behaviour most often in our experiment, followed by those from The Netherlands and finally India. However, our results suggest that on average, levels of uncertainty avoidance in centipede-like games are similar across nationalities. Based on our results, we were not able to distinguish any differences between Israeli, Dutch, and Indian participants in choosing a certain outcome over an uncertain outcome.

Interestingly, our results do confirm our hypothesis that actions of Dutch participants can be interpreted as indicative of forward induction. In contrast, the actions of Israeli participants showed no strategic behaviour at all. This may indicate that Israeli participants were more likely to distrust or to get confused by a surprising opponent.

We hypothesized that, based on the collectivist nature of Indian society, at least half of the Indian participants would show cooperative behaviour. In contrast, our results show high levels of competitiveness across nationalities. When faced with the choice of giving their opponent a high payoff or a low payoff at their last decision point, participants on average preferred to give their opponent a low payoff. In fact, while we expected Dutch participants to be more self-interested than Indian and Israeli participants, Figure 7 suggests Dutch participants to be the least competitive.

In general, the previous actions of the opponent did not influence participants' decisions to behave competitively or cooperatively. However, Figure 7 shows an interesting trend suggesting that Indian participants are cooperative towards opponents that have previously behaved cooperatively to them: the more often an opponent has surprised the an Indian participant by choosing the uncertain, possibly cooperative, option, the more likely they are to respond cooperatively.

In summary, the take-home message of our experiment is that the levels of uncertainty avoidance are similar across nationalities, and that Israeli participants are more likely to distrust an opponent. Levels of competitiveness are high for all three cultures, but surprisingly, the Dutch are the least stingy.

Future work This inter-cultural study is based only on the decisions made by the participants. In order to be able to draw conclusions about the reasoning strategies behind the decisions, we are currently looking at the reaction times of the participants, similar to Bergwerff et al. (2014). We intend to continue our study on the differential roles of affective and game-theoretic aspects, by designing new experiments based on both perfect and imperfect information turn-taking games. We will apply techniques such as eye-tracking and computational cognitive modeling to be better able to distinguish reasoning strategies (Meijering et al., 2012; Top et al., 2018).

References

Aumann, R. J. (1995). Backward induction and common knowledge of rationality. *Games and Economic Behavior*, *8*(1), 6–19.

Bergwerff, G., Meijering, B., Szymanik, J., Verbrugge, R., & Wierda, S. M. (2014). Computational and algorithmic models of strategies in turn-based games. In P. Bello, M.McShane, M. Guarini, & B. Scassellati (Eds.), *Proceedings of the 36th annual conference of the cognitive science society* (p. 1778-1783).

Camerer, C. F. (2011). *Behavioral game theory: Experiments in strategic interaction*. Princeton University Press.

D'Andrade, R. G. (1981). The cultural part of cognition. *Cognitive Science*, *5*(3), 179–195.

Ghosh, S., Heifetz, A., Verbrugge, R., & De Weerd, H. (2017). What drives people's choices in turn-taking games, if not game-theoretic rationality? In J. Lang (Ed.), *Proceedings of the 16th conference on theoretical aspects of rationality and knowledge (TARK XVI)* (pp. 265–284).

Halder, T., Sharma, K., Ghosh, S., & Verbrugge, R. (2015). How do adults reason about their opponent? Typologies of players in a turn-taking game. In *Cogsci* (pp. 854–859).

Henrich, J., Boyd, R., Bowles, S., & Camerer, C. e. a. (2005). "Economic man" in cross-cultural perspective: Behavioral experiments in 15 small-scale societies. *Behavioral and Brain Sciences*, *28*(6), 795–815.

Ho, T.-H., & Weigelt, K. (2005). Trust building among strangers. *Management Science*, *51*(4), 519–530.

Hofstede, G. (1991). *Cultures and organizations. Intercultural cooperation and its importance for survival*. (3rd edition 2010, with G.-J. Hofstede and M. Minkov)

Krockow, E. M., Takezawa, M., Pulford, B. D., Colman, A. M., & Kita, T. (2017). Cooperation and trust in Japanese and British samples: Evidence from incomplete information games. *International Perspectives in Psychology: Research, Practice, Consultation*, *6*(4), 227.

Levinson, S. C. (2012). The original sin of cognitive science. *Topics in Cognitive Science*, *4*(3), 396–403.

Meijering, B., Van Rijn, H., Taatgen, N. A., & Verbrugge, R. (2012). What eye movements can tell about theory of mind in a strategic game. *PLoS ONE*, *7*(9), e45961.

Nagel, R., & Tang, F. F. (1998). Experimental results on the centipede game in normal form: An investigation on learning. *Journal of Mathematical Psychology*, *42*(2), 356–384.

Pearce, D. (1984). Rationalizable strategic behaviour and the problem of perfection. *Econometrica*, *52*, 1029–1050.

Perea, A. (2010). Backward induction versus forward induction reasoning. *Games*, *1*, 168–188.

Peysakhovich, A., & Rand, D. G. (2016). Habits of virtue: Creating norms of cooperation and defection in the laboratory. *Management Science*, *62*(3), 631–647.

Rosenthal, R. (1981). Games of perfect information, predatory pricing and the chain-store paradox. *Journal of Economic Theory*, *25*(1), 92–100.

Roth, A. E., Prasnikar, V., Okuno-Fujiwara, M., & Zamir, S. (1991). Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An experimental study. *The American Economic Review*, 1068–1095.

Top, J., Verbrugge, R., & Ghosh, S. (2018). An automated method for building cognitive models for turn-based games from a strategy logic. *Games*, *9*(3), 44.

Verbrugge, R., Meijering, B., Wierda, S., van Rijn, H., & Taatgen, N. (2018). Stepwise training supports strategic second-order theory of mind in turn-taking games. *Judgment & Decision Making*, *13*(1).

Yamagishi, T., & Yamagishi, M. (1994). Trust and commitment in the United States and Japan. *Motivation and Emotion*, *18*(2), 129–166.

Understanding language about other peoples actions.

Tom Gijssels
University of Chicago, Chicago, Illinois, United States

Marianna Zhang
Stanford University, Stanford, California, United States

Che Lucero
Cornell University, Ithaca, New York, United States

Marc G. Berman
University of Chicago, Chicago, Illinois, United States

Daniel Casasanto
Cornell University, Ithaca, New York, United States

Abstract

When people understand language about their own actions they activate premotor regions they use to perform these actions. Do people understand language about other peoples actions by imagining how they perform these actions themselves, or how they perceive others performing them? Here, we recorded BOLD fMRI while left- and right-handers read about and then imagined their own unimanual actions (e.g. you write) or others actions (e.g. she writes). When imagining their own manual actions, participants preferentially activated PMC circuits controlling their dominant hand. By contrast, when imagining others actions, participants PMC activity reflected both how they perform actions themselves and how they typically see actions performed by right-handers (about 90% of people they see). Language-induced motor imagery for our own actions reflects how we use our own bodies, whereas imagery for others actions also reflects how others use their bodies, even if their bodies differ from our own.

Seeking evidence and explanation signals religious and scientific commitments

Maureen Gill (mcg3@princeton.edu), Tania Lombrozo (lombrozo@princeton.edu)
Department of Psychology, Princeton University, Princeton, NJ 08544

Abstract

Scientific norms value skepticism; many religious traditions value faith. We test the hypothesis that these different attitudes towards inquiry and belief result in different inferences from epistemic behavior: Whereas the pursuit of evidence or explanations is taken as a signal of commitment to science, forgoing further evidence and explanation is taken as a signal of commitment to religion. Two studies (N = 401) support these predictions. We also find that deciding to pursue inquiry is judged more moral and trustworthy, with moderating effects of participant religiosity and scientism. These findings suggest that epistemic behavior can be a social signal and shed light on the epistemic and social functions of scientific vs. religious belief.

Keywords: explanation; evidence; information search; science; religion; moral cognition

Introduction

In his influential work on the sociology of science, Robert Merton introduced the idea of "organized skepticism" as a norm that governs the scientific enterprise. "Most institutions demand unqualified faith," he wrote, "but the institution of science makes skepticism a virtue" (Merton, 1973). Whether or not this norm accurately characterizes all scientific behavior and aspirations, it nicely encapsulates a value that many uphold: the value of critical and unlimited inquiry.

Yet in some walks of life, skepticism and unfettered inquiry can compete with other values. For instance, demanding an explanation for a friend's loyalty, or hiring a private investigator to gather evidence that a spouse is indeed faithful, could damage those relationships by sending a signal about one's (uncharitable) beliefs or (weak) commitment to the relationship. In fact, in economic games, examining the available evidence can be a maladaptive strategy for promoting cooperation (Hoffman, Yoeli, Nowak, 2015). As Merton suggested, organized skepticism can interfere with "values which demand an unquestioning acquiescence."

Within some religious traditions, willingness to believe (e.g., in Jesus or in God) in the absence of evidence is itself regarded as a virtue. In the well-known story of "doubting Thomas," to take an example from the Christian tradition, Jesus tells his apostle who demanded evidence: "because thou hast seen me, thou hast believed: blessed *[are]* they that have not seen, and *[yet]* have believed" (John 20:29). Indeed, faith – whether it is faith in God or in one's partner – may be an epistemic attitude that involves a certain *abdication* from the need for further evidence (Buchak, 2012).

The diverging norms of skepticism and faith introduce an interesting possibility: that the choice to pursue (vs. forgo) inquiry could send a signal about the strength and nature of one's commitments to scientific versus religious norms, and correspondingly, to science versus religion. That is, demanding further evidence or explanation could be seen as a mark of *commitment* regarding science, but a sign of doubt or insincerity in religion, at least within those traditions that value faith. Insofar as commitment to skeptical versus faith-based norms are taken to have other social or epistemic implications, we might also expect individuals who decide to pursue or forgo further inquiry regarding scientific or religious matters to be judged differentially moral, trustworthy, or committed to truth.

Based on these ideas, the current paper asks the following two questions: (1) What kinds of social and moral inferences do people (specifically, American and predominantly Christian adults) make on the basis of another person's decision to pursue or forgo inquiry? (2) Do inferences vary across scientific and religious domains?

Prior work

Research has shown that people interpret others' decisions as signals of moral and socially relevant traits. For example, those who make harm-averse moral judgments or engage in third-party punishment are more trusted and preferred as social partners (Everett, Crockett, Pizarro, 2016; Jordan, Hoffman, Bloom, Rand, 2016). Moral values and group affiliation are also thought to influence belief formation and revision: increased analytic thinking is associated with more polarized views, potentially because analytic individuals use different evidence to support pre-determined conclusions (Kahan & Stanovich, 2016). It remains unknown, however, whether people infer moral and social traits on the basis of epistemic behaviors, namely, the decision to pursue or forgo information search. In the current work we consider two epistemic behaviors: pursuing versus forgoing an *explanation*, and pursuing versus forgoing further *evidence*.

Prior work has found that judgments about the "need for explanation" differ across the domains of science and religion (Liquin, Metz, & Lombrozo, 2018). In particular, participants judged scientific statements – such as "the center of the earth is very hot" – to demand an explanation

to a greater extent than religious statements – such as "there is a hell" – even when confidence in the truth of the two statements was matched. When participants were presented with the "explanation" that it's a mystery (e.g., "Why is the center of the earth very hot [is there a hell]? It's a mystery"), they judged the answer more acceptable for religious questions than for scientific ones. These findings suggest that explanation-seeking – or abdication from explaining – could play different roles within science vs. religion, consistent with the diverging norms of skepticism vs. faith.

There is also reason to believe that science and religion could differ when it comes to attitudes towards evidence. Van Leeuewen (2017) develops a proposal according to which science and religion tend to involve distinct epistemic attitudes – what he calls factual *belief* versus religious *credence*. A characteristic of the latter is that it is "evidentially invulnerable": religious credences are not typically extinguished by contrary evidence. If this view is right, evidence should be more relevant to the evaluation of factual versus religious propositions.

In sum, prior work suggests that various decisions can serve as social signals, and that the domains of science and religion could differ in the epistemic attitudes they typically involve. Across two studies, we investigate novel questions that build upon this work: whether epistemic behaviors (the decision to pursue vs. forgo explanation or evidence) send different social signals across domains.

Study 1

In study 1, we examine the inferences that people make from an agent's epistemic behavior. To do so, we presented a story about a character, Jen, who learns about a new issue: either near-death experiences or the shroud of Turin (scenario: NDE vs. shroud). These issues were chosen because they can be framed as scientific or religious (domain: scientific vs. religious). Jen contemplates whether the issue *demands an explanation* or whether the issue *requires more evidence* (inquiry: explanation vs. evidence). Critically, Jen ultimately decides that it does or does not (decision: pursue vs. forgo). Participants rated the morality of Jen's behavior, her trustworthiness, and her commitment to truth, science, and religion. We predicted that the decision to pursue inquiry would be taken as a signal of scientific commitment, and the decision to forgo as a signal of religious commitment. We also predicted (but failed to find) that these effects would be strongest within their corresponding types of framing.

Method

Participants Participants in Study 1 were 97 adults recruited from Mechanical Turk (63 male, 34 female, mean age 36, range 22-73). Participation was restricted to MTurk workers in the U.S. who had completed 5000 past HITs with a minimum approval rating of 99%. Nine additional participants were excluded for leaving responses blank.

Materials & Procedures Participants were randomly assigned to read one of 16 vignettes about Jen, who learns about an issue and decides whether to inquire further about it. The issue was either near-death experiences or the shroud of Turin (scenario: NDE vs. shroud), framed in a scientific or religious manner (domain: scientific vs. religious). For example, the text for the shroud of Turin with a scientific framing included the following:

> Jen learns about the shroud of Turin, a piece of cotton cloth that may have been the burial shroud that Jesus (1st century preacher and religious leader) was wrapped in after being crucified by the Roman government.
>
> Scientific findings in disciplines ranging from chemistry to biology shine light on whether the shroud of Turin is indeed the burial shroud of Jesus. Multiple radiocarbon dating and vibrational spectroscopy tests date the shroud between 300 BC and 400 AD, corresponding with the timing of Jesus's crucifixion.
>
> Though most scientific leaders believe the shroud to be the burial cloth of Jesus, the matter is still not settled. Some people believe that it is not authentic and/or was created at a later date.

The version with religious framing was similar, but instead of offering scientific evidence and appealing to scientific consensus, it included biblical references and appealed to consensus among religious leaders.

After reading this information, participants learned about Jen's subsequent epistemic behavior: she either decided to pursue further inquiry or not (inquiry decision: pursue vs. forgo), and her inquiry took the form of either seeking (or not seeking) further evidence or seeking (or not seeking) an explanation (inquiry: evidence vs. explanation). Following prior work (Liquin et al., 2018), explanation seeking was framed broadly: that is, the specific type of explanation available (such as mechanistic or teleological) was not specified. For the Shroud of Turin, for example, participants read one of the following sentences, depending on inquiry condition (evidence vs. explanation) and decision (indicated by text in brackets):

> Evidence: Jen decides that she does [not] need more evidence that the cloth was the burial shroud that Jesus was wrapped in.
> Explanation: Jen decides that she does [not] need an explanation for how the shroud came to have its characteristic markings.

Crossing scenario (NDE vs. shroud), domain (scientific vs. religious), decision (yes vs. no), and inquiry (evidence vs. explanation) resulted in the 16 distinct vignettes.

After reading the vignette, participants were asked to rate 14 statements designed to probe their inferences about Jen, including her morality, trustworthiness, commitment to truth, commitment to science, and commitment to religion. All items and rating anchors are indicated in Table 1. Items

about truth, science, and religion were presented in random order before items about morality and trustworthiness. Nine participants failed to answer at least one item and are therefore excluded from reported analyses. Participants then answered an open-ended question about what they thought of the fact that Jen did [not] pursue further evidence or explanation. We do not analyze these open-ended responses here.

Next, participants completed a set of individual difference measures, which are not reported here. Finally, participants reported their political orientation, age, and gender.

Table 1: Study 1 and 2 rating questions. Items with an asterisk were reverse-scored. For the composite measures, we additionally report Cronbach's α (Study1 / Study 2).

Moral and character inferences
Morality
Jen's decision that […] was…
(1 = "very immoral/bad" – 7 = "very moral/good")
Trustworthiness
Jen is probably…
(1 = "very untrustworthy" – 7 = "very trustworthy")
Commitment to truth (α = .88 / .79)
Jen values truth above all.
When it comes to what she believes, Jen cares about getting things right.
Jen is not concerned about whether she is right or wrong.*
Jen values some things more than getting things right.*
(1 = "strongly disagree" – 7 = "strongly agree")
Commitment to science (α = .94 / .94)
Jen has a strong commitment to the methods of science.
Jen is a deeply scientific person.
Jen values her identity as a scientifically-minded person.
Jen trusts scientific authorities.
(1 = "strongly disagree" – 7 = "strongly agree")
Commitment to religion (α = .93 / .94)
Jen has strong religious faith.
Jen is a deeply religious person.
Jen values her religious identity.
Jen trusts religious authorities.
(1 = "strongly disagree" – 7 = "strongly agree")

Results

Our key dependent variables were the single ratings for morality and trustworthiness, as well as our composite ratings for commitment to truth, science, and religion, which were calculated by averaging the four ratings for each scale. The reliability of these scales, as assessed by Cronbach's α, ranged from good to excellent (see Table 1). For each dependent variable, we performed an ANOVA with domain (scientific vs. religious), decision (yes vs. no), scenario (Shroud of Turin vs. NDE) and inquiry (evidence vs. explanation) as between-subjects factors (see Figure 1a). Given the large number of tests, we adopted the more conservative *p*-value of .01 as our threshold for significance; we report all significant effects.

The ANOVA with ratings of morality as a dependent variable revealed a main effect of decision: deciding to inquire was rated morally better than deciding not to, $F(1, 81) = 37.58$, $p < .001$. Analysis of trustworthiness as a dependent variable also revealed a main effect of decision, $f(1,81) = 22.22$, $p < .001$, such that the character was rated as more trustworthy when she decided to inquire than when she decided not to.

Analyzing composite ratings of commitment to truth also showed a main effect of decision, $f(1,81) = 70.40$, $p < .001$, with decisions to inquire associated with higher perceived commitment to truth. However, this effect was qualified by a significant interaction with domain, such that decision had a greater impact on perceived commitment when the issue was framed as religious, $f(1,81) = 8.41$, $p = .005$.

Composite ratings of commitment to science exhibited a similar pattern, revealing a significant main effect of decision in the same direction, $f(1,81) = 45.208$, $p < .001$, as well as a marginal interaction with domain, trending in the same direction, $f(n) = 5.95$, $p = .02$.

Finally, composite religious commitment ratings revealed a significant main effect of decision, $f(1,81) = 45.618$, $p < .001$, but in a direction opposite to that observed for our other dependent variables: the decision to inquire was associated with a *decrease* in perceived commitment to religion. Once again, there was a suggestive trend for decisions to be more informative in the religious domain (decision x domain interaction), $f(1,81) = 2.72$, $p = .10$. There was also a significant main effect of scenario, qualified by an interaction with decision, such that perceived commitment to religion was rated higher when Jen learned about the shroud of Turin, $f(1,81) = 10.68$, $p = .001$, especially when Jen decided not to pursue more information, $f(1,81) = 9.56$, $p = .002$.

Discussion

Participants in our study viewed evidence- and explanation-seeking behaviors favorably: participants viewed the decision to pursue both evidence and explanation as morally good and a cue to trustworthy character. Critically, evidence- and explanation- seeking was also treated as a signal of commitment to truth and science, where *forgoing* further inquiry was treated as a signal of commitment to religion. These effects were remarkably consistent across modes of inquiry (evidence versus explanation), and across our manipulation of domain (science versus religion), though we found modest evidence that pursuit decisions might be regarded as more informative in the domain of religion than science.

We initially hypothesized that the effect of inquiry decisions on inferences about the inquirer would be moderated by participants' own religious and scientific commitments. Because our sample was overwhelmingly non-religious, however, we were unable to test this hypothesis. We revisit this question in Experiment 2, for which we recruited a more religious sample.

Study 2

In Study 2, we again tested the effect of epistemic behaviors (pursuing vs. forgoing evidence vs. explanation) and domain (religious vs. scientific) on inferences about morality, trustworthiness, commitment to truth, commitment to science, and commitment to religion. However, we restricted participation to MTurk workers from the nine states in the U.S. with the highest proportion of religious residents – this involved drawing from the generally protestant population of the South (Lipka & Wormald, 2016). We also aimed to strengthen the manipulation of domain (religious vs. scientific), editing scenarios to be more identifiably religious or scientific. Finally, by including a larger and more religious sample, we aimed to test two hypotheses about individual differences that could moderate the effect of inquiry decision on perceived morality and trustworthiness: religiosity and scientism. Specifically, we predicted that more religious participants might see greater value in the epistemic attitude of faith, resulting in higher ratings of morality and trustworthiness (relative to non-religious participants) after Jen decides to forgo further inquiry. On the other hand, participants who endorse a narrow commitment to science might be especially likely to value associated norms (such as organized skepticism) and therefore judge Jen more favorably (relative to less-scientistic participants) when she decides to pursue inquiry.

Method

Participants Participants in Study 2 were 304 adults recruited from Amazon Mechanical Turk (117 males, 186 females, mean age 40, range 19 to 77). Participation was restricted to MTurk workers from Alabama, Mississippi, Tennessee, Louisiana, Arkansas, South Carolina, West Virginia, Oklahoma, and Georgia. Thirty-six additional participants were excluded for failing one or more attention checks (explained below).

Materials & Procedures The materials and procedures were the same as those in Study 1, with the following modifications. First, we made slight modifications to the 16 original vignettes to further differentiate the religious and scientific framing. For example, for the religious version of the Shroud of Turin vignette, we replaced the original sentence "could it be the burial shroud of Christ," with "could it be the burial shroud of Jesus Christ, son of God?" Second, we collected fewer individual difference measures than in Study 1. Those retained included the religiosity inventory from Pennycook et al. (2012; sample items: "There is a life after death," "Religious miracles occur"), the moralized rationality and importance of rationality scales from Stahl et al. (2016), and the scientism scale from Farias et al. (2013; sample items: "Science provides us with a better understanding of the universe than does religion," "Science is the most valuable part of human culture"), presented in this order. An attention check ("select 'strongly agree'") was included in the religiosity inventory, and 31 participants were excluded for failing to answer correctly. Participants then reported their political orientation, age, and gender.

Finally, participants answered two additional attention check questions about the content of their vignette and Jen's decision; these were simple multiple-choice questions based on what they had read (e.g., "What did Jen decide?"). Four participants were excluded for failing to answer at least one question correctly.

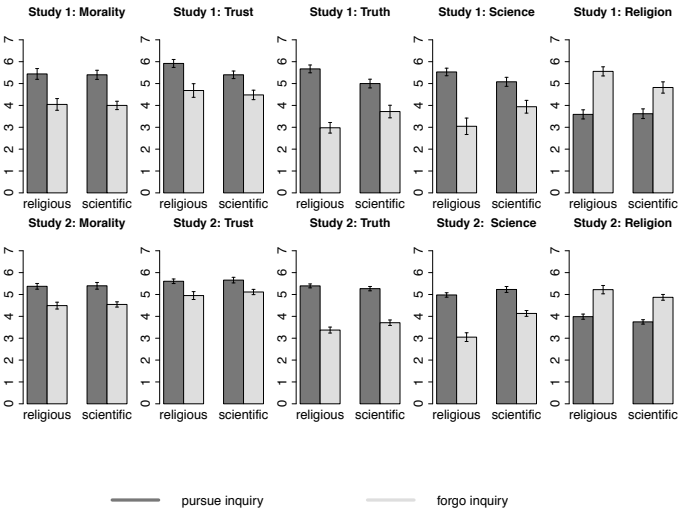


Figure 1: Mean ratings in Study 1 and 2 for the inferred characteristics of the vignette's character as a function of domain and her decision to pursue or forgo further inquiry. Error bars correspond to SEM.

Results

As with Study 1, our key dependent variables were the single ratings for morality and trustworthiness, as well as our composite ratings for commitment to science, religion, and truth, calculated by averaging the four ratings for each scale. The reliability of these scales, as assessed by Cronbach's α, ranged from good to excellent (see Table 1). For each dependent variable, we performed an ANOVA with domain (scientific vs. religious), inquiry decision (pursue vs. forgo), inquiry type (evidence vs. explanation), and scenario (shroud vs. NDE) as between-subjects factors (see Figure 1b). Given the large number of tests, we adopted the more conservative *p*-value of .01 as our threshold for significance, and we report all significant effects.

The ANOVA with ratings of morality as a dependent variable again revealed a main effect of decision, $f(1, 288) = 39.50$, $p < .001$, as well as a marginal interaction between domain, decision, and inquiry type, $f(1, 288) = 6.51$, $p = .01$. Both kinds of inquiry were associated with higher moral goodness judgments, but explanation-seeking behaviors were more informative for morality in a scientific context than a religious one, and conversely, evidence-seeking behaviors were more informative in a religious context than a scientific one. There was also an interaction between

decision and scenario, such that the main effect of decision was more pronounced in the NDE scenario, $f(1,288) = 6.95$, $p = .008$.

Analysis of trustworthiness judgments also revealed a main effect of decision, $f(1,288) = 20.25$, $p < .001$, with the decision to pursue inquiry associated with greater trustworthiness.

Analyzing composite ratings of commitment to truth revealed a main effect of decision, $f(1,288) = 266.52$, $p < .001$, with greater perceived commitment when inquiry was pursued, and a main effect of inquiry type, $f(1,288) = 16.78$, $p < .001$, with greater perceived commitment in the evidence condition than in the explanation condition. There was also a marginal interaction between decision and domain, $f(1,288) = 4.52$, $p = .03$, with decision having a greater impact in the religious condition.

Analysis of commitment to science revealed a main effect of decision, $f(1,288) = 111.10$, $p < .001$, as well as an interaction between decision and domain, $f(1,288) = 9.3$, $p = .003$. As in Study 1, Jen was regarded as having a higher commitment to science when she sought out evidence or explanation, with a greater effect of decision with religious framing. There were also main effects of domain and scenario, such that Jen was perceived as having a higher commitment to science both when the issue was framed as scientific, $f(1,288) = 21.23$, $p < .001$, and when the issue was near-death experiences rather than the shroud of Turin, $f(1,288) = 9.48$, $p = .002$.

The ANOVA with composite commitment to religion revealed a main effect of decision in the opposite direction of truth, morality, truth commitment, and science commitment, as in Study 1. Forgoing inquiry was associated with *greater* commitment to religion, $f(1,288) = 86.626$, $p < .001$. There was also a main effect of scenario, $f(1,288) = 38.349$, $p < .001$, as well as an interaction between decision and scenario, $f(1,288) = 15.75$, $p < .001$: for the Shroud of Turin scenario, perceived commitment to religion was higher overall, and decision was more influential.

We additionally explored whether two of our individual difference measures, religiosity and scientism, moderated the effect of inquiry decision on perceived morality and trustworthiness (see Figure 2). To test for a moderating effect of religiosity, we constructed two pairs of linear mixed effects models (predicting morality or trustworthinesss), treating participant religiosity (centered) and decision as fixed factors, and treating scenario as a random factor with respect to intercept. We fit a full model with the main effects of both fixed factors as well as their interaction and a partial model that included the same factors without an interaction. An ANOVA comparison of the two models revealed that the full model better predicted moral judgments, $X^2(1) = 7.28$, $p = .006$, and trustworthiness judgments, $X^2(1) = 14.56$, $p < .001$. As participant religiosity increased, epistemic decision mattered less for judgments of morality and trustworthiness. Equivalent analyses for participant scientism also revealed that a model with the scientism-decision interaction term better predicted morality, $X^2(1) = 15.19$, $p < .001$, and trustworthiness, $X^2(1) = 27.776$, $p < .001$. However, the pattern was opposite to that observed for religiosity: participants rejecting scientism were likely to see forgoing inquiry as more moral and trustworthy, whereas participants endorsing scientism saw the pursuit of inquiry as more moral and trustworthy.

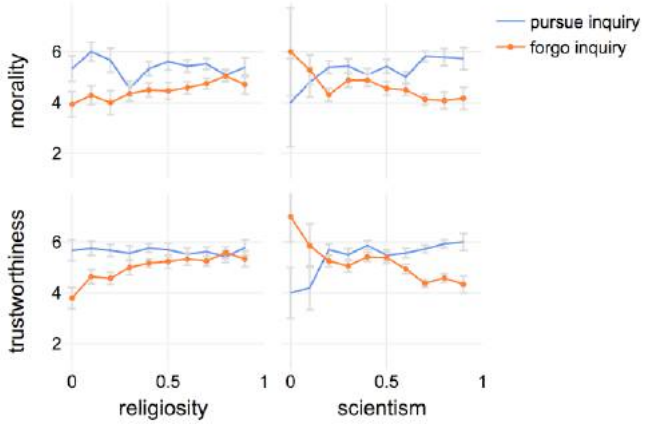


Figure 2: Moral and trustworthiness judgments by participant scientism and religiosity.

Discussion

In Study 2, we replicated our main findings from Study 1 with a larger and more religious sample drawn predominantly from the American South. Jen was regarded as more moral and trustworthy for seeking evidence and explanations. Inquiry behaviors were associated with an increase in commitment to truth and science, but a decrease in commitment to religion. We also found additional evidence of a trend observed in Study 1: inquiry decisions in the domain of religion (vs. science) were generally more informative in the sense that they had a larger impact on inferences about Jen's commitments, especially to science.

Going beyond Study 1, we identified two individual difference factors that moderated the effect of inquiry decision on inferences about morality and trustworthiness: religiosity and scientism. Scientistic participants were inclined to draw inferences about Jen's morality and trustworthiness that were *more* dependent on her decision about whether to pursue or forgo inquiry, showing a more pronounced effect favoring inquiry. On the other hand, religious participants tended to draw inferences about Jen's morality and trustworthiness that were less dependent on her decision about whether to pursue or forgo inquiry.

General discussion

People infer a number of moral and social traits from another person's epistemic behavior. We found evidence that pursuing inquiry is viewed as a signal of commitment to truth and to science, but that forgoing inquiry is perceived as signaling commitment to religion. A person who pursues evidence or explanation is regarded as more moral and trustworthy, but only among certain groups: for more

religious participants, the effect of inquiry on inferences of trust and morality diminishes; for participants who very strongly reject scientism, the relationship reverses.

Keeping track of epistemic behavior is key to learning from others. The finding that adults infer moral character traits from an agent's epistemic behavior contributes to a literature showing a connection between how people track others' epistemic and moral status. Research has shown that young children use epistemic markers, such as past accuracy, to guide evaluations of source trustworthiness (Birch, Vauthier, & Bloom, 2008). However, children also use a source's moral qualities, such as niceness/meanness, in evaluating the truth-value of a claim (Landrum, Mills, & Johnston, 2013). Adults are less likely to trust a source with different political values, even when the information is non-political, e.g., about geometric shapes (Marks, Copland, Loh, Sunstein, & Sharot, 2018). Future research should investigate why we use moral information in epistemic judgments and epistemic information in moral judgments. When does trusting a source mean trusting a person?

The social consequences of information search might carry implications for real epistemic decisions. People often face the choice between accepting a proposition at face value and searching for more information. Our research suggests the possibility that epistemic considerations (e.g., strength of prior evidence, uncertainty) may not fully account for behavior. Social context may play a role in the decision-making process. For instance, a person who wants to signal commitment to religion may be more likely to forgo inquiry, risking false beliefs for potential social rewards (a "display of faith"). A person could also choose to pursue costly inquiry (high search cost, low information value) to be perceived as moral and trustworthy (a "display of skepticism").

The current studies are limited in a number of respects, including the range of materials and underspecified forms of inquiry. Explanation in particular was broadly defined in our experimental materials. There are different kinds of explanations, and participants may have differed in what they took an explanation to be. Indeed, given differences in the need for explanation across domains (Liquin, Metz, & Lombrozo, 2018), and differences in the *kinds* of explanations offered across domains (e.g., Kelemen, 2004; Lupfer, Brock, & DePaola, 1992), it could be that different kinds of explanations are more or less closely tied to religious and scientific norms.

It's also important to note that our sample – while diverse in some respects – drew from an overwhelmingly Christian (and mostly Protestant) population, considerably limiting the extent to which we can make general claims about religion or religiosity. Indeed, we expect a great deal of heterogeneity in religious attitudes towards inquiry, and additionally expect that scientific propositions can be "taken on faith." Future work should explore this heterogeneity, for instance testing more diverse samples, and additionally consider how a more nuanced understanding of science (as opposed to the "scientism" measured here) might affect attitudes towards and inference from the choice to seek further explanation or evidence.

Despite these limitations, the present work contributes to a growing body of work suggesting that beliefs and processes of belief revision are sensitive to both epistemic and social goals. Researchers have proposed that religious belief serves a social coherence function (Norenzayan, 2013), and politicized "scientific" beliefs (such as the endorsement or rejection of anthropogenic climate change or human evolution) are strongly related to cultural / group identity (e.g., Kahan & Stanovich, 2016). As Van Leeuwen (2017) suggests: "If my credence that our god exists can be banished by something so trifling as mere evidence, how can you be sure that I am really committed to our group, which defines itself by allegiance to our god?" Our research shows that forgoing inquiry can send a signal of religious commitment. On the other hand, for most observers, the decision to inquire is considered the more moral action, and a stronger marker of trustworthiness, commitment to science, and commitment to truth.

Acknowledgements

We thank the Concepts & Cognition Lab for valuable feedback, as well as the John Templeton Foundation for support. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Foundation.

References

Birch, S.A., Vauthier, S.A., & Bloom, P. (2008). Three-and four-year-olds spontaneously use others' past performance to guide their learning. *Cognition*, 107(3), 1018-1034.

Buchak, L. (2012). Can it be rational to have faith?. *Probability in the Philosophy of Religion*, *225-255*.

Gottlieb, S., & Lombrozo, T. (2018). Can science explain the human mind? Intuitive judgments about the limits of science. *Psychological science*, 29(1), 121-130.

Hoffman, M., Yoeli, E., & Nowak, M. A. (2015). Cooperate without looking: Why we care what people think and not just what they do. *PNAS*, 112(6), 1727-1732.

Jordan, J. J., Hoffman, M., Bloom, P., & Rand, D. G. (2016). Third-party punishment as a costly signal of trustworthiness. *Nature*, *530*(7591), 473.

Kahan, D. M., & Stanovich, K. (2016). Rationality and belief in human evolution (working paper).

Kelemen, D. (2004). Are children "intuitive theists"? Reasoning about purpose and design in nature. *Psychological science*, *15*(5), 295-301.

Landrum, A. R., Mills, C. M., & Johnston, A. M. (2013). When do children trust the expert? Benevolence information influences children's trust more than expertise. *Developmental Scienc*e, 16(4), 622-638.

Lipka, M., & Wormald, B. (2016). How religious is your state. *Pew Research Center*.

Liquin, E. G., Metz, S. E., & Lombrozo, T. (2018). Explanation and its Limits: Mystery and the Need for Explanation in Science and Religion. T. T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Ed.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society (2065-2070)* . Austin, TX: Cognitive Science Society.

Lupfer, M. B., Brock, K. F., & DePaola, S. J. (1992). The use of secular and religious attributions to explain everyday behavior. Journal for the Scientific Study of Religion, 486– 503.

Marks, J., Copland, E., Loh, E., Sunstein, C. R., & Sharot, T. (2018). Epistemic Spillovers: Learning Others' Political Views Reduces the Ability to Assess and Use Their Expertise in Nonpolitical Domains.

Merton, R. K. (1973). *The sociology of science: Theoretical and empirical investigations*. University of Chicago press.

Farias, M., Newheiser, A. K., Kahane, G., & de Toledo, Z. (2013). Scientific faith: Belief in science increases in the face of stress and existential anxiety. Journal of experimental social psychology, 49(6), 1210-1213.

Norenzayan, A. (2013). Big gods: How religion transformed cooperation and conflict. Princeton University Press.

Pennycook, G., Cheyne, J.A., Seli, P., Koehler, D.J., & Fugelsang, J.A. (2012). Analytic cognitive style predicts religious and paranormal belief. *Cognition*, 123, 335-346.

Ståhl, T., Zaal, M. P., & Skitka, L. J. (2016). Moralized rationality: Relying on logic and evidence in the formation and evaluation of belief can be seen as a moral issue. *PloS one*, 11(11), e0166332.

Tobacyk, J. J. (2004). A revised paranormal belief scale. *International Journal of Transpersonal Studies*, 23(1), 11.

Van Leeuwen, N. (2017). Do religious "beliefs" respond to evidence?. *Philosophical Explorations*, 20(sup1), 52-72.

Event cognition from the perspective of cognitive development

Vladimir V. Glebkin (gleb1514@gmail.com)
Russian Presidential Academy of National Economy and Public Administration,
Prospect Vernadskogo, 82, Moscow, Russian Federation 119571**;**
School 1514,
Krupskoi Street, 12, Moscow, Russian Federation 119311

Ekaterina O. Olenina (katerinaoleninam@mail.ru),
School 1514,
Krupskoi Street, 12, Moscow, Russian Federation 119311

Nikita A. Safronov (nikita1997.08.21@mail.ru),
Moscow Region State University,
Radio Street, 10, Moscow, Russian Federation 105005

Abstract

Event cognition is a rapidly developing and promising research area. Meanwhile, some domains are not considered in detail in this scope. In particular, event cognition is not precisely explored from the perspective of cognitive development. In this paper, we compare the capacity to cut a visual narrative into events for kindergarten students, primary school students, high school students and adults. “The pear film” by W. Chafe (1975) is used as the material for our experiment. We also examine a correlation between event comprehension and other cognitive skills for primary school students. Our work provides clear evidence that, in contrast with high school students and adults, kindergarten students and primary school students perceive visual narrative on the surface level.

Keywords: event cognition, event model, cognitive development, primary school students, narrative comprehension.

Introduction

Event cognition is an intensively developing domain of cognitive science and a promising avenue of research. A number of insightful conjectures and seminal ideas supported by dozens of experiments have been suggested in this domain over recent decades (Suh & Trabasso 1993; Zwaan et al. 1995; Zacks et al. 2001; Rinck & Weber 2003; Ditman et al. 2008; Shipley & Zacks 2008; Yarkoni et al. 2008; Zacks et al. 2009; Tamplin et al. 2013; Radvansky & Zacks 2014; Zacks 2015; Richmond & Zacks 2017, etc.).

The main results of this research line can be presented as follows:

- Humans do not perceive reality in a continuous way; they cut it into a number of chunks called events. This feature is a fundamental characteristic of humans that underpins their way of reasoning and making decisions.
- There is a high level of coherence among humans in cutting the stream of life into events; they detect event boundaries in a highly similar way.
- A shift through event boundaries impairs an ability to predict a future state of affairs and also event memory; this is caused by a change of space, time, characters, objects, causes, and goals, concerned with a particular situation.
- Event cognition is based on the creation and further elaboration of event models that "capture the entities and functional relations involved in understanding a specific state of affairs" (Radvansky & Zacks 2014, 17); event models allow to predict a development of such state of affairs within an event.
- "...event cognition, and event memory in particular, appears to have distinct neurological underpinnings apart from more general knowledge... it seems possible to disrupt the long-term storage of event models, leaving more general knowledge intact, as well as the reverse, disrupting general knowledge, but leaving the ability to process and remember individual events" (Radvansky & Zacks 2014, 131).

At the same time, some methodological flaws seem to hinder further development in this direction. Strangely enough, we could not find any working definition of both event and event model in works of event cognition researchers. We admit that the demand to define correctly the concept ‘event’ may sound a bit scholastic in this scope (see, e.g., Shipley 2008; Schwartz 2008 as an example of the discussion), but the concept ‘event model’ is the key concept which underpins the body of experimental research addressing event cognition. Nevertheless, the researchers usually focus on event boundaries and changes what take place when these

boundaries being passed, whereas a structure of an event model within boundaries is only sketched. The definition by Radvansky and Zacks quoted above is not clear-cut enough to apply it to a particular experiment (What does ‘a specific state of affairs’ mean? How can we measure it?), and it is not clarified in other works. Scholars usually pick out five aspects characterizing event model: temporality, spatiality, protagonist(s), causality, and intentionality (e.g., Rinck & Weber 2003, 1284–1285; Radvansky & Zacks 2014, 61); however, it is not clear how these aspects are represented in a particular event model.

In other words, there is a bunch of important questions which remain unanswered in this scope. Let us stress only few of them. How many basic types of event model can be singled out? What is the structure of each of them; what are the cornerstones of this structure and links between them? Are there any discrepancies between event boundaries which separate events of the same type and boundaries which separate events of different types? Is the ability to produce event models innate, or it is a result of cognitive development? If the latter, how it develops through the life span? Is there any difference between event model typology for kids and adults?

Indeed, there is no opportunity to tackle all these and similar questions here. We address only some of them concerned with the problem of cognitive development. To be more precise, we have explored how an ability to cut reality into events and to produce event models is acquired in childhood, what is the difference between kids and adults in event cognition, how an acquisition of this capacity correlates with language acquisition and the development of other cognitive skills (there are a few papers addressing age differences in event cognition (e.g., Copeland & Radvansky 2007; Kurby & Zacks 2011), but they do not explore the problem from the perspective of cognitive development). This paper can be considered as the first step in this direction.

Our work examines age differences in cutting a visual narrative into events as a part of a process of cognitive development. We have used "the Pear Film" made by Wallace Chafe and his colleagues in 1975 as a material for the experiments. Importantly, "the Pear Film" includes actions, pictures and sounds, but no words, deploying the same chain of events for all viewers. This film contains a wide range of interactions between protagonists, spatial and temporal changes; its understanding presupposes the capacity to ‘read’ complex intentions and distinguish between physical and social causality. In other words, it provides good material for producing different event models, and, therefore, for exploring event cognition from the perspective of cognitive development. It is worth also noting that "the Pear Film" has opened an avenue of research tackling different aspects of a language and culture interconnection in the process of conceptualizing particular stream of events (Bernardo 1980; Chafe 1980; Clancy 1980; Downing 1980; Du Bois 1980; Tannen 1980; Orero 2008; Fon et al. 2011; Matzur & Mickievicz 2012; Vilaró et al. 2012; Blackwell 2015; Cummings 2015, 59–63; Kibrik et al. 2015; Glebkin et al. 2017).

A plot of "the Pear Film" is important for understanding the results of our experiment, therefore, it looks reasonable to begin with a brief description of the story taken from Chafe 1980, XIII–XIV.

The film begins with a man picking pears on a ladder in a tree. He descends the ladder, kneels, and dumps the pears from the pocket of an apron he is wearing into one of three baskets below the tree. He removes a bandana from around his reek and wipes off one of the pears. Then he returns to the ladder and climbs back into the tree.

Toward the end of this sequence we hear the sound of a goat, and when the picker is back in the tree a man approaches with a goat on a leash. As they pass by the baskets of pears, the goat strains toward them, but is pulled past by the ruin and the two of them disappear in the distance.

We see another closeup of the picker at his work, and then we see a boy approaching on a bicycle. He coasts in toward the baskets, stops, gets off his bike, looks up at the picker, puts down his bike, walks toward the baskets, again looking at the picker, picks up a pear, puts it back down, looks once more at the picker, and lifts up a basket full of pears. He puts the basket down near his bike, lifts up the bike and straddles it, picks up the basket and places it on the rack in front of his handlebars, and rides off. We again see the man continuing to pick pears.

The boy is now riding down the road, and we see a pear fall from the basket on his bike. Then we see a girl on a bicycle approaching from the other direction. As they pass, the boy turns to look at the girl, his hat flies off, and the front wheel of his bike hits a rock. The bike falls over, the basket falls off, and the pears spill out onto the ground. The boy extricates himself from under the bike, and brushes off his leg.

In the meantime we hear what turns out to be the sound of a paddleball, and we see three boys standing there, looking at the bike boy on the ground. The three pick up the scattered pears and put them back in the basket. The bike boy jets his bike upright, and two of the other boys lift the basket of pears back onto it. The bike boy begins walking his bike in the direction he was going, while the three other boys begin walking off in the other direction.

As they walk by the bike boy’s hat on the road, the boy with the paddleball sees it, picks it up, turns around, and we hear a loud whistle as he signals to the bike boy. The bike boy stops, takes three pears out of the basket, and holds them out as the other boy approaches with the hat. They exchange the pears and the hat, and the bike boy keeps going while the boy with the paddleball runs back to his two companions, to each of whom he hands a pear. They continue on, eating their pears.

The scene now changes back to the tree, where we see the picker again descending the ladder. He looks at the two baskets, where earlier there were three, points at them, backs up against the ladder, shakes his head, and

tips up his hat. The three boys are now seen approaching, eating their pears. The picker watches them pass by, and they walk off into the distance.

We chose four age groups for the experiment: 5-7-year-old kindergarten students (KS), 7-9-year-old primary school students (PS), 14-16-year-old high school students (HS), and adults (A).

Based on the previous experiments (Glebkin et al. 2017), we expected that kindergarten students and primary school students would be less skillful in producing event models than high school students and adults which would entail serious problems in detecting event boundaries for KS and PS subjects. In particular, in the case of "the Pear Film" they would be inclined to 'paste' event boundaries and to minimize a number of parts in this visual narrative. To be more precise, we supposed that a mean number of events for kindergarten students and primary school students would be less than for high school students and adults, and kindergarten students and primary school students would determine event boundaries in a less systematic way. We also expected to discover some correlation between the ability to cut a narrative into events and other cognitive and communicative skills concerned with story retelling for primary school students. Our hypothesis in this scope was that the more correct and more detailed was a film retelling the more accurate was a choice of event boundaries by a subject.

Experiment

Method

Subjects. 34 (14 m, 20 f) 5-7-year-old kindergarten students; 73 (35 m, 38 f) 7-9-year-old Moscow primary school students; 36 (12 m, 24 f) 14-16-year-old Moscow high school students; 35 (13 m, 22 f, mean age 37) adults.

Material. "The Pear Film" by Wallace Chafe (6 min 32 sec).

Procedure. The procedure of the experiment followed the model well-established in modern cognitive psychology (e.g., Newtson 1973; Speer et al. 2003). Each subject was processed individually. Where were two versions of the experiment. In the first version subjects watched the film on MacBook Air, 13,3″, 2560x1600 two times. Before the first viewing, the subjects were instructed to watch the film closely as passive viewers. Before the second viewing, they were asked to cut the film into events, i.e., the largest meaningful parts, in any way they find appropriate (this task is similar to the coarse segmentation task in Speer et al. 2003). In addition, a special explanation was given to the groups of kindergarten students and primary school students. The idea of the event segmentation was illustrated on the example of book chapters and some other similar examples. Then the subjects watched the film for the second time and pressed a button at the beginning and at the end of any meaningful part of the film.

In the second version, the procedure was similar, but after the first viewing the participants were asked to retell the story as precisely as they can. This version of the experiment was carried out only for primary school students. In order to make sure that the retelling has no significant influence on the event segmentation task, a control group of 20 primary school students was tested in the first version before the main experiment. No significant difference between two groups was discovered both in a total of episodes each subject cut the film ($F(1, 89) = 0.017$; $p=0.89$) and in the percentage of subjects identifying main event boundaries (χ^2 (12)=16.56, p=0,17).

Two groups of parameters were measured. The first group represented the event segmentation task. It included two variables: a total of episodes that the film was cut into by each subject (TE), and, accordingly, a number of subjects pointed to a particular point as an event boundary (NS) (more precisely, because of some difference in subjects' reaction time it was a set of points located near each other which can be considered as characterizing the same change of a situation). Also for PS group a total of "right" boundaries for each subject (TB_r) (i.e., the boundaries picked out by a significant number (40% and more) of adults and high school students) was calculated. We considered TB_r as a characteristic of cognitive skills involved in event cognition important for the comparison with cognitive skills involved in narrative comprehension and retelling.

The second group of variables, actual only for the primary school students, was concerned with the film retellings. It checked memory for events and also basic cognitive and communicative skills important for narrative understanding and retelling, namely, an ability to categorize objects, an ability to understand the causal chain of events and represent it in the retelling, the richness of language used by subjects. The set of variables was an extended version of the set of variables presented in Glebkin et al. 2017. The following variables were measured: the total number of words exploited in retelling, discounting selfrepetitions and false starts (TW); a total of events presented in retelling (TE_r); a total of events correctly presented in retelling (TEc); a total of errors in action description (FA) (e.g., 'guys picked up pears' instead of 'the boy hands pears to one of the guys'); a total of errors in object description (FO) (e.g., 'apples' instead of 'pears'); a total of incorrect description of causal chain of events and sub-events (FC) (e.g., ambiguous reference, missing connections within an event and between events); a total of interpretations (TI) (e.g., 'stole a basket of pears' instead of 'picked up a basket of pears'); a total of dependent words (TDp) (such as 'who', 'which', 'because', etc.); a total of details mentioned in the retelling (TDt) (e.g., the color of the bike, the peddleball, etc.).

TE_r, TEc and TDt may need a clarification. A total of events was calculated according to the most frequent event boundaries picked out by high school students and adults. These boundaries divide the film into meaningful episodes some of which are connected with others and

some are autonomous (e.g., appearance and disappearance of a man with a goat). The primary school students mentioned some episodes in their retellings and missed others. Some of mentioned episodes were retold correctly (all protagonists and main interactions between them were included in the description), others were presented with serious gaps (e.g., in the episode with the fall of the boy on the bike some subjects missed the girl on the bike). In other words, the complex TE-TEc characterizes the correctness of the film framework representation in a retelling.

TDt points to another feature of the retellings. As a rule, PS subjects focused on actions and missed an appearance of protagonists, their clothes, scenery, etc. Only few of them mentioned such details. For us, such interest to particular details is a special cognitive characteristic important to event cognition. Some arguments for that are presented in the next sections.

Results

As predicted, a mean number of episodes that the pear film was cut into by each subject (TE) for KS was less than for PS, and TE_{PS} was less than TE_{HS}. At the same time, there were no significant difference between TE_{HS} and TE_{A} (TE_{KS}=2.61; TE_{PS}=4.58; TE_{HS}=8.61; TE_{A}=8.23; F_{KSPS} (1, 102) = 61.38; p_{KSPS}<0.001; F_{PSHS} (1, 105) = 18.36; p_{PSHS}<0.001; F_{HSA} (1, 68) = 0.26; p_{HS}=0.6).

As we expected, KS were less consistent in the determination of event boundaries than PS, and PS were less consistent in that than HS and A. The distribution of subjects' choices through the event boundaries, which are most frequent and most important for the narrative, is presented in Table 1.

Table 1. The percentage of subjects identifying most frequent event boundaries

№	Event boundaries	A	HS	PS	KS
1	A man with a goat appears	49	47	17	11
2	A man with a goat disappears	57	56	23	2
3	A boy on a bike appears	49	52	26	22
4	The bike boy stops near the baskets	17	13	20	11
5	The bike boy steals a basket	71	69	30	13
6	A girl on a bike appears	17	17	13	2
7	The bike falls over	71	69	41	25
8	Three boys appear	46	43	20	11
9	The free boys finish to put the pears back in the basket	49	47	27	11
10	The exchange of the pears and the hat	14	8	19	11
11	The boy with the paddleball hands a pear to each his two companions	40	39	24	8
12	The scene changes back to the tree	60	56	24	5
13	The three boys pass by the picker	40	39	19	5

The difference between the results of KS and PS and, accordingly, between the results of PS and HS is significant ($\chi^2_{KSPS}(12)=26.27$; $p<0.05$; $\chi^2_{PSHS}(12)=70.18$; $p<0.001$). Meanwhile, data for HS and A are located extremely close to each other. Fig. 1 presents these results in a graphic form.

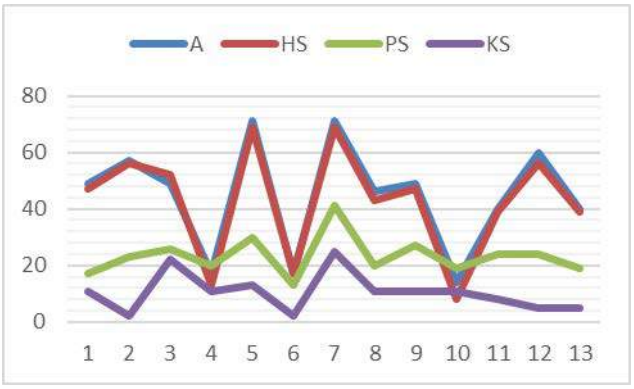


Fig. 1. The diagram of a percentage of subjects identifying main event boundaries for A, HS, PS and KS.

Interestingly, a comparison of data within PS does not reveal any significant differences. In particular, the comparison of TE of 34 first year PS and TE of 26 second year PS (TE_{1PS}=4.62; TE_{2PS}=5.42) gives p=0.34; the comparison of TE of 26 second year PS and TE of 13 third year PS (TE_{3PS}=4.77) provides p=0.54.

We also checked, as mentioned, a possible correlation between event cognition skills and skills in narrative comprehension and retelling for the group of primary school students. It seems reasonable to distribute all correspondences among three groups. The first group ($p<0.001$) includes the correlation between a total of episodes that the film was cut and a total of details mentioned in the retelling ($r(TE,TDt)=0.422$); the correlation between a total of "right" boundaries and a total of details ($r(TB_r,TDt)=0.410$); and the correlation between a total of "right" boundaries and a total of events correctly presented in retelling ($r(TB_r,TE_c)=0.420$). The second group ($p<0.01$) includes the correlation between a total of episodes and the total number of words exploited in retelling ($r(TE,TW)=0.383$); between a total of "right" boundaries and the total number of words ($r(TB_r, TW)=0.344$); between a total of episodes and a total of events correctly presented in retelling ($r(TE,TE_c)=0.382$); between a total of "right" boundaries and a total of events presented in retelling ($r(TB_r, TE_r)=0.339$). The third group ($p<0.05$) includes the correlation between a total of episodes and a total of events presented in retelling ($r(TE, TE_r)=0.250$); between a total of episodes and a total of

incorrect description of causal chain of events and sub-events (r(TE,FC)=-0.245); between a total of episodes and a total of interpretations (r(TE,TI)=0.287); between a total of "right" boundaries and a total of dependent words (r(TB_r, TDp)=0.263); between a total of "right" boundaries and a total of incorrect description of causal chain of events and sub-events (r(TB_r,FC)=-0.266); and between a total of "right" boundaries and a total of interpretations (r(TB_r,TI)= 0.261).

Discussion

The results support the conjecture of serious problems that kindergarten students and primary school students encounter when cutting a visual narrative into events. They lose some key event boundaries, and they are less consistent in detecting event boundaries than high school students and adults. In other words, they are inclined to interpret the narrative as the whole story not picking out any significant parts within it. Indeed, this does not mean that kindergarten students and primary school students do not cut the film into events when they watch it. They may encounter serious difficulties in making sense of the task. This is especially important for kindergarten students (primary school students perform similar tasks from time to time in their school lessons). Therefore, it is hard to distinguish between difficulties in defining events and event borders in process of real viewing (which is, mainly, unconscious) and difficulties in conscious efforts to cut the film into events.

In order to cast additional light on this issue, some other data need to be addressed. In Glebkin et al. 2017 clear evidence was provided for serious problems which kindergarten students encounter in "The Pear Film" retellings in comparison with high school students (TW, TE_r, FA, FO, FC, TI, TDp values differed significantly for KS and HS groups). Further investigations have shown that similar problems characterize retellings of primary school students. Therefore, difficulties in event cognition correlate in age aspect with difficulties in narrative comprehension and retelling, and we can expect substantial correlation in the acquisition of these groups of cognitive skills.

A precise look to the data presented above might clarify this issue. In particular, the figures in Table 1 (and the diagrams in Fig. 1) are interesting. There are only three points in which high school students and adults are less consistent (or almost equally consistent) than primary school students: the moment of bike boy stopping near the baskets (Point 4); the moment when a girl on a bike appears (Point 6); and the moment when the bike boy and the boy with the paddleball exchange the pears and the hat (Point 10). Why high school students and adults do not generally detect these points as event boundaries?

In order to clarify this issue, let us focus on "The Pear Film" narrative at hand. Point 4 and Point 10 characterize some local changes in the narrative, but there are strong arguments for interpreting these points as situated within events; they are unlikely to be basic event boundaries. In particular, Point 4 is situated within the event "The bike boy steals a basket of pears", and this was the reason for high school students and adults not to detect it as an event boundary. Similarly, Point 10 – the exchange of the pears and the hat – is not an event boundary, because the boy with the paddleball when taking three pears from the bike boy is expected to hand the pear to each of his two companions to end the event.

The case of Point 6 – a girl on a bike appears – is a bit more complicated. The girl is a new character, and she is introduced in the story with a close-up, therefore, her appearance may look as the beginning of a new event. Meanwhile, she is not a main character; she is engaged in the event "The bike boy rides down the road". Her part in this event is implemented later on when she brings to bear the boy's fall. If so, this moment is unlikely to be an event boundary.

Why, in this case, primary school students did often detect these points as event boundaries? There are, at least, two aspects of PS subjects' strategy in event boundaries detecting which may underpin these particular decisions. Firstly, two levels in the structure of event model can be singled out. The first level characterizes changes in location, actions and interactions given in visual perception, situated, so to say, on a superficial level. For instance, "the boy's bike falls over". Some of such changes are autonomous, but some others are signs of elements, which are located on a deeper level and need a special interpretation (e.g., the fact, that the boy places the basket on the rack in front of his handlebars, and rides off, means that he steals the basket). On average, primary school students do not include some important links on the deeper level into their event models. As a result, their models are 'poorer' than and models of high school students and adults addressing the same event; they are 'flat', but not 'volumetric' ones. If it is so, some changes in the visual field, such as ones, happened in Point 4, Point 6, and Point 10, are sufficient for them to detect these points as event boundaries.

Secondly, an analogy with language comprehension helps to explore this issue from another perspective. Researchers single out three levels of text representation: the surface form, the propositional textbase, and the situation model (e.g., Schmalhofer & Galvanov 1986; Radvansky & Zacks 2014, 57–58). A difference between sentences on the first level concerns words and grammatical structures, but not the facts and their interpretation (e.g., *Anna cleaned the room and then went to the cinema* and *After cleaning the room Anna went to the cinema*). On the second level, a situation is the same, but a focus and an interpretation may be different (e.g., *The ball flew into the goal from the foot of Peter* and *Peter scored a goal*; in the first case it may be also ricochet). On the third level, the situations are different.

If expanding this model on a visual narrative, a difference on the first level would mean different wide shots of the same event; difference on the second level – e.g., a close-up of different objects within the same event; and difference on the third level – different events. From

this perspective, in contrast with high school students and adults, primary school students are inclined to 'paste' together different levels. In particular, the close-up of the girl on a bike may be a reason for them to detect Point 6 as an event boundary.

Finally, let us zoom in on the comparison between event cognition skills and skills in the narrative comprehension and retelling. These data support the conjecture that the 'flat' event model dominates for primary school students. The variable, which shows the most significant correlation with both a total of episodes and a total of right boundaries, is a total of details mentioned in the retelling (TDt). At the same time, TDt is hardly to be a characteristic of logical aspects of the narrative comprehension; rather, it characterizes visual attention and visual memory. In other words, high TDt values are not valid signs of high quality of event models.

Also, the strong correlation between a total of right boundaries and a total of events correctly presented in retelling, and a significant correlation between both a total of right boundaries and a total of events and the total number of words exploited in the retelling, between a total of episodes and both a total of events presented in retelling and a total of events correctly presented in retelling show that the more detailed a retelling is the more event boundaries are detected by the subject.

The correlation between a total of events (or a total of right boundaries) and characteristics of understanding and representation of logical structure of the narrative (a total of incorrect description of causal chain of events and sub-events; a total of interpretations; a total of dependent words) is less significant. It is worth paying special attention to the lack of any correlation between a total of events and a total of errors in object description (a variable characterizing categorization skills).

Overall, our data support the conjecture that event models evolve through the life span, and event models of kindergarten students and primary school students subjects are 'poorer' than whose of high school students and adults. Therefore, the age of 10-14 years old is likely to be crucial for the development of event cognition ability. The problem of correlation between this ability and other cognitive skills in diachronic perspective needs further investigation.

References

Bernardo, R. (1980). Subjecthood and Consciousness. In W. Chafe (ed.), *The Pear Stories: Cognitive, Cultural, and Linguistic Aspects of Narrative Production* (pp. 275–300). Norwood, New Jersey: Ablex.

Blackwell, S. (2015). *Porque* in Spanish Oral Narratives: Semantic *Porque*, (Meta)Pragmatic *Porque* or Both? In A. Capone, & J. Mey (eds.), *Interdisciplinary Studies in Pragmatics, Culture and Society* (pp. 615–632). New York: Springer Berlin Heidelberg.

Chafe, W. (ed.). (1980). *The Pear Stories: Cognitive, Cultural, and Linguistic Aspects of Narrative Production*. Norwood, New Jersey: Ablex.

Chafe, W. (1980a). The Development of Consciousness in the Production of a Narrative. In W. Chafe (ed.). *The Pear Stories: Cognitive, Cultural, and Linguistic Aspects of Narrative Production* (pp. 9–50). Norwood, New Jersey: Ablex.

Copeland, D., & Radvansky, G. (2007). Aging and integrating spatial mental models. *Psychology and Aging*, 22, 569–579.

Cummings, L. (2015). *Pragmatic and Discourse Disorders*. Cambridge: Cambridge University Press.

Ditman, T., Holcomb, P. I., & Kuperberg, G. F. (2008). Time travel through language: Temporal shifts rapidly decrease information accessibility during reading. *Psychonomic Bulletin & Review*, 14, 750–756.

Downing, P. (1980). Factors Influencing Lexical Choice in Narrative. In W. Chafe (ed.), *The Pear Stories: Cognitive, Cultural, and Linguistic Aspects of Narrative Production* (pp. 89–126). Norwood, New Jersey: Ablex.

Du Bois, J. (1980). The Search of a Cultural Niche: Showing the Pear Film in a Mayan Community. In W. Chafe (ed.), *The Pear Stories: Cognitive, Cultural, and Linguistic Aspects of Narrative Production* (pp. 1–8). Norwood, New Jersey: Ablex.

Fon, J., Johnson, K. & Chen, S. (2011). Durational Patterning at Syntactic and Discourse Boundaries in Mandarin Spontaneous Speech. *Language and Speech*, 54, 5-32.

Glebkin V., Safronov N. & Sonina V. (2017). Discourse Acquisition in 'Pear Stories' of Preschool-aged Children. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. J. Davelaar (Eds.), Proceedings of the 39th Annual Conference of the Cognitive Science Society (pp. 2096–2101). Austin, TX: Cognitive Science Society.

Kibrik, A., Fedorova, O., & Nikolaeva, Ju. (2015). Multimodal Discourse: In Search of Units, in G. Airenti, B. Bara & G. Sandini (eds.), *Proceedings of the EuroAsianPacific Joint Conference on Cognitive Science, 4th European Conference on Cognitive Science, 11th International Conference on Cognitive Science, Torino, Italy, September 25–27, 2015* (pp. 662–667). Torino: University of Torino.

Kurby, C., & Zacks, J. (2011). Age differences in the perception of hierarchical structure in events. *Memory & Cognition*, 39, 75–91.

Laurent, A, Nicoladis, E. & Marenette, P. (2015). The development of storytelling in two languages with words and gestures. *The International Journal of Bilingualism*, 19(1), 56–74.

Matzur, I. & Mickievicz, A. (2012). Pear Stories and Audio Description: Language, Perception and Cognition across Cultures. *Perspectives*, 20 (1), 55–65.

Newtson, D. (1973). Attribution and the unit of perception of ongoing behavior. *Journal of Personality and Social Psychology*, 28, 28–38.

Nicoladis, E., Pika, S., & Marentette, P. (2009). Do French-English bilingual children gesture more than monolingual children? *Journal of Psycholinguistic Research*, 38, 573–585.
Orero, P. (2008). Three different receptions of the same film: 'The Pear Stories Project' applied to audio description.
Radvansky, G., & Zacks, J. (2014). *Event cognition.* Oxford; N. Y.: Oxford University Press.
Richmond, L., & Zacks, J. (2017). Constructing experience: Event models from perception to action. *Trends in Cognitive Sciences*, 21(12), 962–980.
Rinck, M., & Weber, U. (2003). Who when where: An experimental test of the event-indexing model. *Memory & Cognition*, 31, 1284–1292.
Shipley, Th. (2008a). An Invitation to an Event. In Th. Shipley, & J. Zacks (eds.), *Understanding events* (pp. 3–30). Oxford; N. Y.: Oxford University Press.
Shipley, Th., & Zacks, J. (eds.). (2008). *Understanding events.* Oxford; N. Y.: Oxford University Press.
Schmalhofer, F., & Galvanov, D. (1986). Three components of understanding a programmer's manual: Verbatim, propositional, and situational representations. *Journal of Memory and Language*, 25, 279–294.
Schwartz, R. Events Are What We Make of Them. In Th. Shipley, & J. Zacks (eds.), *Understanding events* (pp. 54–62). Oxford; N. Y.: Oxford University Press.
Speer, N. K., Swallow, K. M., & Zacks, J. M. (2003). Activation of human motion processing areas during event perception. *Cognitive, Affective, & Behavioral Neuroscience*, 3, 335-345.
Suh, S., & Trabasso, T. (1993). Inferences during reading: Converging evidence from discourse analysis, talk-aloud protocols, and recognition priming. *Journal of Memory and Language*, 32, 279–300.
Tamplin, A., Krawietz, S., Radvansky, G., & Copeland, D. (2013). Event memory and moving in a well-known environment. *Memory & cognition*, 41, 1109–1121.
Tannen, D. (1980). A Comparative Analysis of Oral Narrative Strategies: Athenian Greek and American English. In W. Chafe (ed.), *The Pear Stories: Cognitive, Cultural, and Linguistic Aspects of Narrative Production* (pp. 1–8). Norwood, New Jersey: Ablex.
Vilaró, A., Duchowski, A, Pilar, O., Grindinger, T., Tetreault, S. & di Giovanni, E. (2012). How sound is the Pear Tree Story? Testing the effect of varying audio stimuli on visual attention distribution. *Perspectives*, 20 (1), 55–65.
Yarkoni, T., Speer, N., & Zacks, J. (2008). Neural substrates of narrative comprehension and memory. *NeuroImage*, 41, 1408–1425.
Zacks, J. (2015). *Flicker. Your Brain on Movies.* Oxford; N. Y.: Oxford University Press.
Zacks, J., Speer, N., & Reynolds, J. (2009). Segmentation in reading and film comprehension. *Journal of Experimental Psychology: General*, 138, 307–327.
Zacks, J. M., Tversky, B., & Iyer, G. (2001). Perceiving, remembering, and communicating structure in events. *Journal of Experimental Psychology: General*, 130, 29-58.
Zwaan, R., Magliano, J., & Graesser, A. (1995). Dimensions of situation model construction in narrative comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21, 386–397.

Book Design, Attention, and Reading Performance: Current Practices and Opportunities for Optimization

Karrie E. Godwin[1] (kgodwin1@kent.edu)

Cassondra M. Eng[2] (cassonde@andrew.cmu.edu)

Grace W. Murray[1] (gmurray5@kent.edu)

Anna V. Fisher[2] (fisher49@andrew.cmu.edu)

[1]Kent State University, Department of Educational Psychology, 150 Terrace Drive, Kent, OH 44243 USA

[2]Carnegie Mellon University, Department of Psychology, 5000 Forbes Ave. Pittsburgh, PA 15213 USA

Abstract

Becoming a proficient reader is a critical skill that supports future learning. Toward the end of the primary grades, reading becomes increasingly automatized, and children begin to transition from *learning-to-read* to *reading-to-learn*. Yet, the design of beginning reader books may be suboptimal for novice readers. Colorful illustrations that contain irrelevant information (i.e., seductive details) presented in close proximity to the text may increase attentional competition between these sources of information; thus, hampering decoding and reading comprehension. Study 1 examines this hypothesis by experimentally manipulating components of the book design (e.g., presence/absence of seductive details) and investigating its effect on attention and reading performance in first grade students. In Study 2, we conduct an analysis in which we identify common design features in books for beginning readers and examine the prevalence of design features that were found to tax attention in Study 1 and in prior research. Collectively this work identifies an important opportunity in which instructional materials can be optimized to better support children as they learn-to-read.

Keywords: attention; selective sustained attention; reading comprehension, decoding, reading, book design

Introduction

Learning to read is an important skill that enables future learning (National Association for the Education of Young Children, 1998). As reading becomes increasingly automatized, children begin to transition from *learning-to-read* to *reading-to-learn,* and thus can more readily apply this skill to learn novel information. But acquiring this skill set is challenging due to a number of factors including (but not limited to) deficits in prior knowledge (e.g., pre-reading skills such as phonological awareness; Kirby, Parrila, & Pfeiffer, 2003), learning disabilities (e.g., dyslexia), as well as cognitive limitations (e.g., working memory, processing speed; Jacobson et al., 2011). The difficulty many children experience in becoming competent readers is reflected in a 2005 report in which *only* 31% of 4th grade students in the United States were identified as "Proficient" or above on the NAEP reading assessment and rates were lower still for some groups of minority students: Black 13%, Hispanic 16%, American Indian/Alaska Native 18% (Perie, Grigg, & Donahue, 2005, pp. 3-4). These sobering statistics highlight the need to identify malleable factors that can be leveraged to better support children's reading achievement. One potential factor is book design.

The design of beginning reader books may not be optimized to support early reading, which may further increase the difficulties children experience acquiring this skill. Prior research has found that the close proximity between text and illustrations in books for beginning readers increases attentional competition between these sources of information hampering reading performance (Godwin, Eng, Todaro, Murray, & Fisher, 2018; Torcasio, & Sweller, 2010). By increasing the spatial separation between text and illustrations (Godwin et al., 2018) or reducing extraneous details from illustrations (Eng, Godwin, & Fisher, 2018), attentional competition is reduced (indexed by gaze shifts away from the text), and reading comprehension improves. These results are promising, as they point to a malleable factor (i.e., book design) that could in principle be optimized to better scaffold young readers' attention to the text and in turn enhance their developing literacy skills. However, it is currently unknown whether these design choices (e.g., close proximity between text and illustrations, inclusion of irrelevant details in illustrations) are typical in beginning reader books. If these design choices represent a standard design practice, then this emerging body of research points to an unrecognized opportunity for intervention.

The present paper reports two studies. Study 1 provides a conceptual replication of Eng et al. (2018), but also extends prior work with second grade students to a younger age group, first-graders. In Study 1, we investigate experimentally whether an element of the book design (i.e., presence/absence of attention-grabbing, but irrelevant to the text, details in illustrations) negatively affects children's

attention to the text, diminishing their reading performance. Study 1 makes an important contribution given growing concerns regarding the replicability crisis (e.g., Camerer et al., 2018; Nosek et al. 2015). Study 2 makes a novel contribution by examining issues of generalizability, namely examining whether the design features of the book utilized in prior research are commonplace and thus represent a potential avenue for intervention. In Study 2, we conduct an analysis of 100 beginning reader books in which we identify common design features and assess how prevalent the design choices that were found to tax children's attention in Study 1 (and in prior research) are in children's books.

Study 1

Method

Thirty first-grade children participated in the present study (M = 7.09 years, SD = .32 years, 16 females, 12 males, 2 did not report). The sample represents local diversity with children being 63.3% White, 13.3% African American, 16.7% Multi-Racial, and 6.7% reported as other. Participants were recruited from schools in and around a mid-sized city in the Northeastern United States. Participants were tested individually by trained hypothesis-blind research assistants.

Design and Procedure

In order to ensure ecological validity, Study 1 utilized a commercially available beginning reader book selected from the *Hooked On Phonics Learn to Read* series. Children were asked to read aloud the book "Good Job Dennis," by Amy Kraft. Following Eng et al. (2018), the book design was manipulated within-participants such that half of the book was presented in the Standard layout of unaltered pages from the commercially available book, and half of the book was presented in the Streamlined layout in which the illustrations were simplified by removing the irrelevant details. The presentation order of conditions (Standard condition or Streamlined condition first) was counterbalanced across participants. Each half of the book contained 6 pages. Minor modifications were made to the text to ensure that each half of the book had approximately equivalent number of words (average number of words per page: 43.0 first half, 42.3 second half). Identification of irrelevant details was based on a separate calibration study (Eng et al., 2018). Fifteen college students were given photocopies of the book and were asked to outline details in the illustrations that they believed were *relevant* to the story text for each page. The illustration details in which participants reached over 90% agreement were included in the Streamlined condition and all other remaining details were removed (See Figure 1A and 1B).

The book was presented on a laptop computer and children's gaze shifts away from the text were recorded using eye tracking technology. Decoding was assessed prior to reading the story (Word Recognition in Isolation Test) and while children read aloud (Running Record). Following the story, a post-test was administered to assess reading comprehension.

Figure 1A: Sample page of the Standard layout condition

Figure 1B: Sample page of the Streamlined layout condition

Measures

Gaze Shifts Children's attention allocation to the text was measured using a RED250 mobile eye tracker (SensoMotoric Instruments, Inc.) in which gaze shifts away from the text were recorded. For each page of the book, Areas of Interest (AOIs) were created for the text, white space, and illustrations. The number of gaze shifts away from text AOIs (to illustration AOIs or white space AOIs) was calculated using the SMI BeGaze software and the average number of gaze shifts per page is reported.

Decoding Measures Decoding is thought to be an important component of reading Fluency (Rasinski, 2004). The decoding measures assess children's ability to accurately

identify words (either in isolation or embedded in text). Two decoding measures were employed: the Word Recognition in Isolation task and a Running Record.

Word Recognition in Isolation Task Children completed a modified Word Recognition in Isolation (WRI) task which served as an independent measure of children's ability to decode words fluently (Morris, 2013). The WRI was administered prior to children reading the story. Children are shown leveled lists of words and asked to read the words aloud as quickly and accurately as possible. The number of words read correctly (out of 100 possible words) within the time limit was recorded.

Running Record (RR) The research assistant manually recorded the child's decoding accuracy for each word in the story and the proportion of correct responses was calculated (Clay, 1972).

Reading Comprehension Measure Children were asked six open-ended comprehension questions. Responses were recorded by hypothesis-blind research assistants. We slightly modified the questions provided by the book manufacturer to maintain the ecological validity of the comprehension assessment. Questions were designed such that they probed memory for content presented on specific pages. The post-test included six questions, three questions from each half of the book. Rather than scoring children's responses in a binary fashion (correct vs. incorrect), partial credit was possible. In each half of the book, children completed two 2-point questions and one 3-point question and thus could earn up to 7 points per condition. For example, children were asked to recall Dennis' job. Children earned full credit (2 points) if they stated that Dennis directs traffic and helps children cross the street. Partial credit (1 point) was awarded if children provided an incomplete answer (e.g., "he helps children"), and 0 points if children provided an incorrect answer or failed to recall Dennis' job. The percentage of correct responses (out of 7) is reported. Scoring was completed by condition blind research assistants. To ensure inter-rater reliability, the data was scored twice by two research assistants and Cohen's Kappa (Cohen, 1960) was calculated (.88).

Results

There were no significant differences in average reading time per page in the Standard condition (M = 55000.21 ms; SD = 35065.49 ms) compared to the Streamlined condition (M = 54066.81 ms; SD = 37571.03 ms), paired-sample $t(29) = .27$, $p = .79$. There were also no significant differences in participants' Running Record scores while reading in the Standard condition (M = 94.78%; SD = 5.13%) compared to the Streamlined condition (M = 94.87%; SD = 5.03%), paired-sample $t(29) = .34$, $p = .74$.

Reading Comprehension Children's comprehension scores were significantly higher in the Streamlined condition (M = 80.48%, SD = 20.37%) than in the Standard condition (M = 51.90%, SD = 24.74), paired-sample $t(29) = 4.72$, $p < .0001$ (see Figure 2); Cohen's d = 1.26. In order to test for order effects, we conducted a mixed factorial analysis of variance (ANOVA), factoring condition order as the between-subject variable and comprehension as the within-subject variable. There was no main effect of condition order, $F(1, 28) = .02$, $p = .90$, and no significant interaction between order and comprehension, $F(1, 28) = 2.09$, $p = .16$. These results indicate that reading in the Streamlined condition resulted in higher comprehension compared to reading in the Standard condition, regardless of the amount of time spent reading, the quantity of words a child accurately read aloud, and the order in which the layout was presented.

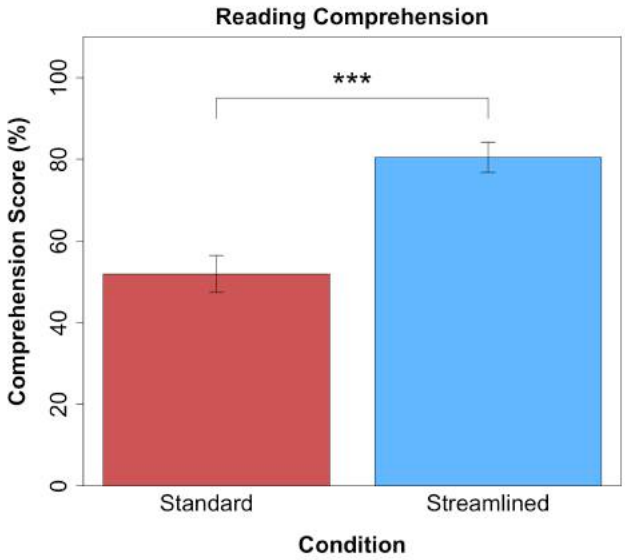


Figure 2: Percentage of correct answers on the story questions as a function of book layout. ***$p < .0001$).

Gaze Shifts On average, children switched their point of fixation away from the text 27.78 times per page (SD = 26.48) in the Standard layout compared to 13.71 times in the Streamlined layout (SD = 11.07), paired-sample $t(29) = 4.67$, $p = <.0001$; Cohen's d = .69. Three outliers were identified. With the removal of these outliers, there was still evidence of a significant main effect of book layout on children's gaze shifts (paired-sample $t(26) = 5.65$, $p<.0001$. Cohen's d = .89). Children looked away from the text almost twice as much in the Standard condition than they did in the Streamlined condition (See Figure 3).

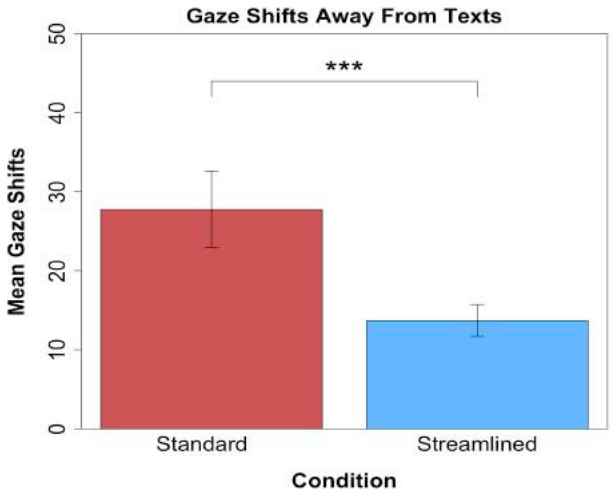


Figure 3: Average gaze shifts away from the text per page as a function of book layout. ***$p < .001$).

The Role of Individual Differences Next we examined whether the Streamlined layout might be especially beneficial for children who often shift their attention away from the text. For this analysis, a difference score for each child was calculated by subtracting the Standard layout comprehension score from the Streamlined layout comprehension score. Difference scores estimated changes in reading comprehension performance from the Streamlined layout, such that higher and positive scores indexed greater gains in comprehension. Difference scores ranged from -57.14% to 85.71%, with a mean of 28.10% (*SD* = 33.24%). Children's gaze shifts in the Standard layout condition were positively associated with Comprehension Gain scores ($r = .49$, $p = .003$), as shown in Figure 4. Thus, the Streamlined layout was especially helpful for children who frequently shifted their gaze while reading: the more children looked away from the text, the more their comprehension benefited from reading the book in the condition in which extraneous details were removed.

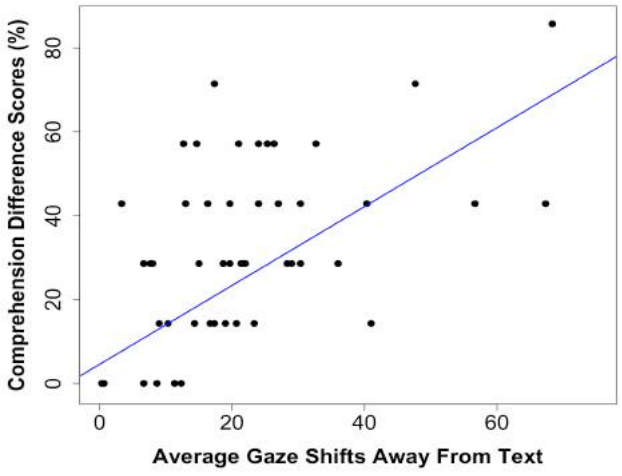


Figure 4: Association between gaze shifts and comprehension gains with outlier removed.

Unique Contribution of Gaze Shifts to Comprehension Gains To ensure that the findings were not entirely due to variance shared with reading ability, children completed the WRI test prior to the reading session to assess participants' decoding fluency ($M = 55.90$, $SD = 20.04$). To examine the extent to which children's gaze shifts away from the text while reading uniquely predicted how much children's comprehension improved from the Streamlined layout condition, we conducted a multiple regression analysis that included gaze shifts and WRI scores as predictors of children's comprehension difference scores. Results show that gaze shifts ($\beta = 5.57$, $t = 2.28$, $p = .003$) accounted for unique variance in comprehension gains when reading from the Streamlined layout, but reading fluency (indexed by the WRI) did not ($p > .10$; see Table 1). The results suggest gaze shifts away from the text while reading account for unique variance in comprehension gains independent of the overall reading ability.

Table 1: Relation of Gaze Shifts to Comprehension Gains

	β	*SE*	*t*
Eye Gaze Shifts	3.37**	0.24	2.28
WRI Score	-.55	0.32	-1.72

** $p < 0.01$. $R^2 = .45$. $F = 11.05$. $df = 2, 27$.

Study 1 successfully replicates the results from prior research (Eng et al., 2018) with a younger age group (first-graders) and demonstrates that extraneous illustration details are a source of distraction for beginning readers. Extraneous, nonessential illustration details were found to disrupt attention as evidenced by the increase in gaze shifts away from the text. This design choice not only affected children's patterns of attention allocation, but it also reduced children's reading comprehension.

While the successful replication points to a robust finding, it remained unclear whether the design features characteristic of the book used in Study 1 are prevalent in other books for beginning readers. If the inclusion of irrelevant details in illustrations is a common practice, then it points to an opportunity in which we could better support children's reading by modifying the design of beginning reader books. We begin to address this question in Study 2.

Study 2

Method

Design and Procedure

Guided by a children's librarian, 100 children's beginning reader books were selected from local libraries near a Midwestern town in the United States. The books were subsequently analyzed to investigate common design

elements. Books were pseudo-randomly chosen to ensure representation of multiple publishers (17 total) and topics. The sample of books represent work from 101 authors and 92 illustrators. Trained coders rated each story page, excluding publisher pages, on 10 categories relating to aspects of the book design including: features of the illustrations (e.g., color, alignment, irrelevant details), text (location, enhancements), and general design (layout, use of white space, borders). Of particular interest for the present study was the category irrelevant details as well as page layout given that Study 1 and prior research (Eng et al., 2018) have found that the inclusion of irrelevant details in illustrations as well as including illustrations in close proximity to the text (Godwin et al., 2018) increase attentional competition and reduce reading performance. The remaining 8 categories were included to provide a more comprehensive analysis of the common design features employed in beginning reader books.

For each book, the percentage of pages in each category level was calculated and means for the data set are reported. Coders received extensive training on the coding protocol using worked examples. Coders also completed a training set, consisting of 7 beginning reader books in order to establish interrater reliability (Cohen's kappa = .80, range: .76 to .85).

Results

All of the books were leveled for beginning readers. The average number of pages per book was 27 (*SD* = 8.77).

Features of the illustrations

Beginning reader books tend to contain illustrations that are very colorful and detailed: on average 93.42% of a book's pages contained illustrations that included five or more colors and 97.79% of a book's pages contain some or intermediate levels of detail. Most book pages contained a single illustration (93.19% of a book's pages), and the illustrations were generally aligned to the text (86.98% of a book's pages). However, the inclusion of irrelevant details in illustrations was found to be a common practice with 86.56% of a book's pages containing some or several irrelevant details.

Features of the text and general layout

Text location varied, but common text locations included centered at the top (35%) or bottom of the page (21.8%), or in multiple locations (13%). Design features intended to enhance the saliency of the text including text boxes, fading, or bubbles were rare (4.65% of a book's pages) as were the use of borders (6%). Surprisingly, white space was not utilized on 28.73% of a book's pages. Although publishers

Table 2. Mean percentage of a books' pages coded in each category level

Features of the Illustrations:					
	Black & White	**1-2 Colors**	**3 Colors**	**4 Colors**	**5+ Colors**
Color	0%	2.36%	1.50%	2.67%	93.42%
	Minimal Detail	**Some Detail**	**Intermediate Detail**	**Rich Detail**	
Detail	< 1%	50.55%	47.24%	1.88%	
	Reiterates	**Disambiguates**	**Topical**	**Unrelated**	
Alignment to text	86.98%	7.33%	5.36%	0%	
	None	**Some**	**Several**		
Irrelevant Visual Details	13.44%	33.38%	53.18%		
	No	**Yes**			
Multiple Illustrations	93.19%	6.81%			

Features of the Text:											
	Left	**Right**	**Top**	**Bottom**	**Middle**	**Multiple Locations**	**Full Page**	**Top L**	**Top R**	**Bottom L**	**Bottom R**
Location	3.5%	2%	35%	21.8%	4.5%	13%	2%	10%	3.6%	2.6%	2%
	No Enhancements					**Enhancements**					
Text Enhancements (e.g., Fading, Text Boxes or Bubbles)	95.34%					4.65%					

General Design Features:					
	Text & illustration presented on sequential pages	**Text & Illustration presented on adjoining pages**	**Text & illustration presented on same page with some spatial separation**	**Mixed Layout**	**Text embedded in illustration**
Page Layout	0%	8.79%	62%	1.63%	27.46%
	None	**Intermediate**	**A lot**		
Use of White Space	28.73%	56.49%	14.78%		
	No	**Yes**			
Use of Borders	94%	6%			

Note. For every book, the percentage of pages in each category-level was calculated, and means are reported.

tended to include some spatial separation between text and illustrations (62% of a book's pages), embedding the text *inside* illustrations was also a frequent design choice (27.46% of a book's pages).

Discussion

During the primary grades, young children are just beginning the challenging work of learning how to decode text. The difficulty many children experience aquiring literacy skills may be compounded by the design of beginning reader books. These instructional materials may fail to take an important cognitive constraint into consideration; namely, children's immature attention regulation system (e.g., Ruff & Rothbart, 2001). Placing text and illustrations in close proximity may unintentionally create attentional competition between these sources of information, hampering reading comprehension. Such attentional competition may be particularly disadvantageous when illustrations contain irrelevant information. The present work explored this possibility with a group of first grade students (Study 1) and provides an extensive analysis of book design features that may influence children's attention allocation across 100 beginning reader books (Study 2).

Several notable findings emerged from this work: Study 1 informs our understanding of how beginning readers allocate their attention while reading independently, and identifies a design feature that influences children's ability to maintain attention to the text. The inclusion of irrelevant details in illustrations for beginning readers was found to disrupt attention allocation and hampered reading comprehension. This finding corroborates prior work (Eng et al., 2018) and provides an important conceptual replication with a novel age group; thus, demonstrating the robustness of the effect. Study 2 examined the prevalence of this design choice, as well as other design features, in books for beginning readers by conducting a detailed analysis of 100 books. The results point to several common design features that may increase attentional competition for young readers including: embedding the text within the illustrations and often not including white space. Critically for the present work, illustrations containing irrelevant details was found to be a common design practice. The prevalence of these design choices point to an opportunity in which we could better support children's emergent literacy skills by modifying the design of beginning reader books.

Acknowledgments

We thank Kristen Boyle, Melissa Pocsai, Emery Noll, Priscilla Medor, Aimee Wildrick, Megan Musso, Lily Borodkin, Eleni Magiassos, Sam Sharp, Leah Williamson, Maria Horrigan, Kayla Whitlock, Freya Kaur, Rachael Todaro, and Melanie Angell for their help with data collection and coding. We also thank the children, parents, and teachers who made this project possible. This work was supported in part by the ProSEED/Simon Initiative Seed grant from Carnegie Mellon University awarded to A.F., by the National Science Foundation awarded to A.F. and K.G. (NSF Award # BCS-1730060) and by the Institute of Education Sciences, U.S. Department of Education, through grant R305B150008 to Carnegie Mellon University. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation, the Institute, or the U.S. Department of Education.

References

Chalnick, A., & Billman, D. (1988). Unsupervised learning of correlational structure. *Proceedings of the tenth annual conference of the cognitive science society* (pp. 510-516). Hillsdale, NJ: Lawrence Erlbaum Associates.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Almeid, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., Isaksson, S., Manfredi, D., Rose, J., Wagenmakers, E., & Wu, H. (2018). Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015. *Nature Human Behavior,* 2, 637-644.

Clay, M. (1972). The early detection of reading difficulties: A diagnostic survey.

Eng, C. M., Godwin, K. E., Boyle, K. A., & Fisher, A. V. (2018). Effects of Illustration Details on Attention and Comprehension in Beginning Readers. In C. Kalish, M. Rau, J. Zhu, T.T. Rogers (Eds.) *Proceedings of the 40th Annual Meeting of the Cognitive Science Society* (pp. 336-341). Austin, TX: Cognitive Science Society.

Godwin, K. E., Eng, C. M., Todaro, R., Murray, G., & Fisher, A.V. (2018). Examination of the role of book layout, executive function, and processing speed on children's reading fluency and comprehension. In C. Kalish, M. Rau, J. Zhu, T.T. Rogers (Eds.) *Proceedings of the 40th Annual Meeting of the Cognitive Science Society* (pp. 1723-1728). Austin, TX: Cognitive Science Society.

Jacobson, L. A., Ryan, M., Martin, R. B., Ewen, J., Mostofsky, S. H., Denckla, M. B., & Mahone, E. M. (2011). Working memory influences processing speed and reading fluency in ADHD. *Child Neuropsychol.* 17(3), 209-224.

Kirby, J. R., Parrila, R. K., & Pfeiffer, S. L. (2003). Naming speed and phonological awareness as predictors of reading development. *Journal of Educational Psychology,* 95(3), 453-464.

Morris, D. (2013). *Diagnosis and correction of reading problems* (2nd ed.).. New York, NY: Guilford Press.

National Association for the Education of Young Children (1998). Learning to read and write: Developmentally appropriate practices for young children. A joint position statement of the International Reading Association and the National Association for the Education of Young Children. *Young Children*, *53*(4), 30–46.

Nosek, B. A., Aarts, A. A., Anderson, C.. J., Anderson, J. E., Kappes, H. B., … (2015). Estimating the reproducibility of psychological science. *Science,* 349 (6251). DOI 10.1126/science.aac4716

Perie, M., Grigg, W. S., & Donahue, P. L. (2005). The Nation's report card: Reading 2005. https://nces.ed.gov/nationsreportcard/pubs/main2005/2006451.asp

Rasinski, T. V. (2004). Assessing Reading Fluency Pacific Resources for Education and Learning. *Honolulu, Hawaii.*

Ruff, H. A., & Rothbart, M. K. (2001). *Attention in early development: Themes and variations*. New York, NY: Oxford University Press.

Torcasio, S. & Sweller, J. (2010). The use of illustrations when learning to read: A cognitive load theory approach. *Applied Cognitive Psychology,* 24, 659-672.

Effects of Induced Affective States on Decisions under Risk with Mixed Domain Problems

Rui Gong (rg2796@tc.columbia.edu)
Department of Human Development, Teachers College, Columbia University
525 W 120th Street, New York, NY 10025 USA
James Corter (jec34@tc.columbia.edu)
Department of Human Development, Teachers College, Columbia University
525 W 120th Street, New York, NY 10025 USA

Abstract

We investigated whether induced affective states can affect the process and outcomes of decisions under risk. A mood induction task was used to elicit a positive or negative mood in a sample of adult participants (N=48). The participants then responded to 28 decision problems, each offering a choice between two mixed-domain risky alternatives. The dependent variables of interest were decision-making choices, as well as an eye-tracking based attentional measure: the total fixation durations for certain critical aspects of the two presented risky decision options. Mood condition did not have a significant main effect on participants' choices, or on mean total fixation time for problems. However, fixation times showed a three-way interaction between mood condition, domain (gain versus loss), and time (block). The fixation time data also provided some general insights into participants' patterns of attention allocation during decision-making. They generally spent more time looking at values compared to probabilities, and more time looking at potential gains compared to losses (although this difference declined over time, especially for positive-mood participants).

Keywords: emotion; decision making; mood induction; affect; allocation of attention; eye tracking; risk; cognitive processing; strategy; choice

Incidental affect has been used to predict and explain a wide variety of judgments and decisions (Peters, Västfjäll, Gärling & Slovic, 2006). Incidental affect refers to feelings or mood states induced by a situation that is normatively irrelevant to a given decision. Most early studies on incidental mood induction took a simple valence-based approach, dividing emotions into positive and negative categories. Researchers found that individuals in a happy rather than a sad mood tend to make optimistic judgments and choices by overestimating the likelihood of positive outcomes and underestimating the likelihood of negative outcomes (Loewenstein & Lerner 2003; Johnson & Tversky, 1983; Raghunathan & Pham, 1999).

More recent research has focused on how particular affective states can affect the general information processing strategy adopted by an individual, towards more analytic or more heuristic strategies (Lerner 2015). Findings suggest that individuals who are in a happy mood are more likely to adopt a heuristic processing strategy, a tendency to use intuition and "gut feelings" with relatively little attention being paid to details. By contrast, individuals who are in a sad mood are more likely to adopt a systematic processing strategy, with careful analysis of information (Bolte, Goschke & Kuhl, 2003; George & Dane, 2016; Schwarz & Clore, 1996; Schwarz, 2000).

Affective states may influence decision-making because the decision maker selectively attends to, encodes, and retrieves emotion-relevant information (Niedenthal & Setterlund, 1994). This phenomenon can be seen as consistent with the affect infusion model (AIM), which posits that affectively loaded information influences an individual's cognitive and behavioral processes, pushing their decision outcomes in a mood-congruent direction (Forgas, 1995). If such mood priming occurs, then individuals in a positive mood should be more likely to access thoughts about the positive aspects of a risky situation compared to those in a neutral mood (Forgas & George, 2001; Nygren, Isen, Taylor & Dulin, 1996). Thus, positive moods may increase an individual's risk-taking tendency with mixed-domain options, because positive potential outcomes will be emphasized over potential losses, so that risky choices will be perceived as more favorable. Individuals in a negative mood, by contrast, are more likely to access thoughts about the negative aspects of risky situations, which consequently would lead to more conservative decision-making choices so as to avoid potential loss (Yuen & Lee, 2003).

Nevertheless, prior research provides mixed results regarding the direction of influence of incidental affect on decision-making processes (Lerner, Li, Valdesolo & Kassam, 2015). An alternative model, the mood-maintenance hypothesis (MMH), posits that incidental mood states motivate behavior such that individuals act to maintain or attain positive mood states (Kliger & Kudryavtsev, 2014). Accordingly, individuals in a positive mood avoid risk in order to maximize the likelihood of maintaining their positive mood, whereas individuals in a negative mood seek risk in an attempt to obtain gains that

might relieve their negative mood (Mishra, et al, 2010; Mishra, 2014; De Vries, Holland, & Witteman, 2008; Hills, et al, 2002).

The contrasting predictions of these models, we argue, can be directly examined using eye-tracking based attentional measures, in studies of mixed domain decisions under risk. This type of design allows us to track individuals' focus of attention on both positive and negative aspects underlying their decision-making processes. Thus, by examining participant's attention to gain vs. loss information, we can assess whether participants' attention allocation is in line with the predictions of a mood-congruence (affect infusion) or mood-maintenance hypothesis.

Empirical Study

The purpose of the present study was to investigate the influence of induced affective states on the process and outcomes of decision-making with a set of risky choice problems. A mood induction task (watching short videos) was used to elicit a positive (happy) or a negative (sad) mood. Previous research has shown that the use of movie or story procedures is an effective means of manipulating participants' moods (Drouveli & Grosskopf, 2016; Ellard, Farchione & Barlow, 2012; Gerrards-Hesse, Spies, & Hesse, 1994; Westermann, Spies, Stahl, & Hesse, 1996). The decision-making task using mixed domain problems gave participants a chance to systematically compare and weigh different aspects of the two risky decision options. According to an affect infusion or mood congruence (AIM) account, positive mood should enhance attention to information about gains, while negative mood should make information about losses more salient and more viewed. In contrast, the mood-maintenance hypothesis (MMH) predicts that individuals in a negative mood should be especially motivated to attend to information about potential gains. Finally, from the standpoint of the heuristic/analytic dichotomy, we investigate the hypothesis that individuals in a negative mood may be more likely to adopt a systematic processing strategy, perhaps by calculating expected value or by using an equivalent procedure, whereas participants in a positive mood may be more likely to use a heuristic processing strategy (George & Dane, 2016; Schwarz & Clore, 1996; Schwarz, 2000).

It seems important in assessing the effects of incidental emotion on decisions to look at decision *process* (as well as outcomes). We accomplish this by using eye-tracking-based attentional measures. By studying attention in the context of mixed-domain decision problems under risk, we can track the decision-maker's focus of attention on positive (gain) and negative (loss) information. These aspects of the present study constitute a novel approach to investigating the possible influence of induced affective states on risky decision-making.

Method

Participants

Forty-eight participants were recruited from a large private University community in North America, either by responding to flyers posted on campus bulletin boards or for course credit. Participants included both undergraduate and graduate students (36 females and 12 males. Most (90%) participants ranged in age from 20 to 30 years, 94% had obtained at least a bachelor's degree, and 88% had completed a basic statistics course. They participated in the study for either a payment of $10 or course credit.

Overview of Procedure

Participants were tested individually. They were informed that the purpose of the study was to examine the factors influencing decision-making for problems involving potential financial gains and losses, and the process of how such decisions were made. Each participant was randomly assigned to one of two mood induction groups: positive or negative. The participant first was taken through a calibration procedure with the eye-tracker, to enable accurate gaze point calculations. Following the viewing of the mood-induction movie clip, the participant was asked to make choices for each of 28 risky decision problems displayed on a computer screen equipped with an eye tracking equipment. During this task, participants were encouraged to work at their own pace.

Mood Induction

Movie clips were used to induce emotions "incidental" to the decision task. Two movie clips of similar length (6 to 7 minutes), one categorized as "happy" (from *The Muppet Show*), and the other as "sad" (from *Schindler's List*), were utilized. These clips have previously been shown to successfully induce positive and negative mood states, respectively (De Vries et al., 2008). The success of the mood induction procedure of the experiment was checked via a self-reported mood questionnaire administered after the video watching and before the decision-making task. All participants were asked to rate on a 7-point Likert scale (ranging from 1 to 7) how well each of the following terms (*happy, joyful, cheerful, enthusiastic, sad, blue, upset, distressed*) described how they felt at that moment. All of the terms are taken from the PANAS-X positive and negative affect schedule (Watson & Clark, 1999), and have been previously classified as representing either a positive valence or a negative valence.

Mixed Decision-making Task

Twenty-eight risky decision problems were presented, each consisting of two decision options (labeled 'a' and 'b'). Each option was a risky mixed prospect, consisting of a loss and a gain with associated (complementary) probabilities. The display format for an example decision problem is shown in Figure 1. Note that an analytic strategy

such as calculating expected value (EV) requires attention to both values and probabilities for both gains and losses of each decision option (all eight discrete pieces of information).

A Tobii model T60 eyetracker (version:3.2.3) with associated software was used to monitor the participants' attention paid to the eight consequential regions of each decision problem. Specifically, the eye tracking-based attentional measures included duration of fixations on eight critical regions, defined by: gain value, gain probability, loss value, and loss probability for each of the two decision options. The total fixation duration (TFD), in seconds, within each critical region or 'area of interest' (AOI) was computed as the total viewing time for each area across all episodes in which a participant had looked within the AOI, starting with a fixation within the AOI and ending with a fixation outside the AOI. Due to eyetracker calibration issues, we eliminated data from five participants whose fixations were not accurately identified, resulting in an effective N of 43 (positive mood condition n=21, negative mood condition n=22).

Participants' choices on the twenty-eight decision problems were also analyzed, including whether they chose the EV-maximizing option.

Figure 1. Display format for a sample decision problem offering a choice between option a and option b.

Results

As a manipulation check for the mood induction procedure, summary positive and negative mood rating scores were obtained by averaging the participants' self-ratings on the four relevant adjective scales: positive = (*happy, joyful, cheerful, enthusiastic*), negative = *(sad, blue, upset, distressed*). Figure 2 presents the mean rating scores on the positive words (Pos_Score) and negative words (Neg_Score), by condition (induced positive or negative mood). It can be seen that the mood induction was effective, as measured by the self-ratings of positive and negative mood. A multivariate ANOVA was conducted (overall) on the two self-rating summary dimension, and the overall omnibus F-test for the mood induction Condition was significant, $F(2, 45) = 89.112$, $p<.001$, suggesting strong mood-induction effects. Both positive (Pos_Score), $F(1, 46) = 97.04$, $p<.001$, and negative (Neg_Score), $F(1, 46) = 124.47$, $p<.001$, scores were significantly affected by the manipulation.

Decision outcomes:

To assess whether induced positive or negative mood affects the degree to which a participant engages in analytic processing, we first tested whether participants in the negative mood condition tended to show more EV-maximizing choices (based on EV calculations or equivalent procedures) than did participants in the positive mood condition. The relevant data consisted of information on participant's choice (a or b) for each pair of mixed domain problem. To analyze the data, we created a summary variable, EV score, defined as each participant's total number of EV-maximizing choices on those 28 pairs of problems. Therefore, these maximization scores could range from 0 to 28. Descriptive statistics showed that for the negative mood condition, $M = 19.3$, $s = 3.28$; for the positive condition, $M = 19.5$, $s = 3.88$. A one-way ANOVA indicated that this difference in the mean maximization score was not significant ($F(1, 46) = .026$, $p > .05$).

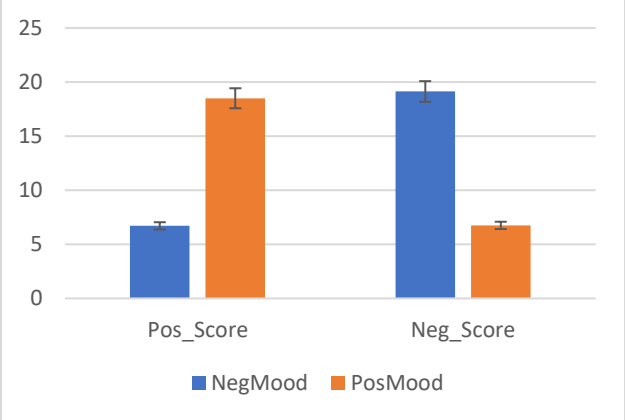


Figure 2. Results of the manipulation check for the effects of viewing two alternative videos (Condition = NegMood, PosMood) on self-rated positive and negative emotional valence.

Patterns of attention:

By analyzing the eye tracking-based attentional measure of total fixation durations (TFDs), we sought to find out how induced moods affect decision *process*, as reflected in the amount of attention that participants pay to certain critical aspects of the considered decision options. Specifically, we sought to answer the following questions: 1) Do participants in a positive mood state tend to pay more attention to positive aspects of the problems (as consistent with an affect infusion or mood congruence account, AIM)? 2) Are participants in a negative mood state more likely to pay attention to the negative aspects of the problem (as consistent with AIM), or more motivated to seek out positive information (as predicted by the mood-maintenance hypothesis, MMH)?

Because our decision-making task included 28 relatively difficult problems, time effects (due to fatigue and/or practice) were thought likely to occur. Also, it is possible that any effect of the emotion manipulation might be short-lived (Andrade & Ariely, 2009). Thus, we analyzed mean TFDs across four equal time blocks: problems 1-7, 8-14, 15-21, and 22-28. Due to practice/fatigue effects, we expected to see a decreasing trend in TFDs.

The marginal-mean TFDs and standard deviations are presented in Table 2, for the main effects of Type of information and Domain.

Participants in both mood conditions spent more time looking at values compared to probabilities, and more time looking at gains compared to losses. As expected, a time effect occurred whereby participants' fixation time spent on 'type' and 'domain' generally decreased from block 1 to block 4 (with exceptions for 'values' and 'losses' - an increase in TFDs from block 2 to block 3). One possible explanation is that block 3 contains relatively more high conflict problems (defined as having a small EV difference between the two options for each problem).

Table 2. Marginal (main effect) descriptive statistics for total fixation duration (TFD) by Type of information (values vs.probabilities) and by Domain (gains vs. losses) for each block of problems.

			Negative Mood		Positive Mood	
		Block	M	SD	M	SD
Type of Information:	Values	1	67.26	6.07	70.69	7.08
		2	45.54	4.54	49.70	5.50
		3	49.73	5.33	54.06	6.64
		4	44.16	5.41	47.33	5.60
	Probabilities	1	34.08	2.79	35.68	4.72
		2	25.17	2.40	28.64	3.85
		3	25.85	2.17	27.90	3.92
		4	25.59	2.41	26.05	3.99
Domain:	Losses	1	38.87	4.41	39.92	4.60
		2	26.85	2.72	29.76	3.74
		3	29.93	3.50	35.20	4.61
		4	26.86	3.26	31.71	4.81
	Gains	1	62.46	4.36	66.45	7.26
		2	43.86	4.15	48.58	5.75
		3	45.65	3.90	46.76	5.98
		4	42.89	4.53	41.66	5.02

Inferential Analyses

A repeated-measures ANOVA predicting mean TFD for each critical area was conducted using one between-subject factor of 'condition' (induced positive mood vs. negative mood), and three within-subject factors: 'domain' (potential gains vs. potential losses), 'type' of information (payoff values vs. payoff probabilities), and 'block' (1-4). Note that this analysis averages looking times (TFDs) for a given critical region (e.g., gain values) across the two decision alternatives.

In this ANOVA, statistically significant effects were found for the within-subject factors of Type ($F(1, 41) = 124$, $p < .001$), Domain ($F(1, 46) = 103.1$, $p < .001$) and Block ($F(2.685, 110.09) = 30.28$, $p < .001$), with a two-way interaction between Type and Domain ($F(1, 41) = 6.22$, $p = .017$), as well as a three-way interaction among Condition, Domain, and Block ($F(2.337, 95.814) = 2.99$, $p = .038$).

The descriptive and inferential results reveal that participants in both mood conditions spent significantly more time looking at values than probabilities, and more on gains than losses. However, it must be recognized that these main effects are to some degree confounded with the left-right position of these quantities on the screen, so some of the differences might be due to a reading order effect.

To interpret the interaction patterns, we first examined the significant three-way interaction among Condition, Domain, and Block. This interaction is shown in Figure 3. In this interaction, the payoff value and probability for a decision option are not separated, presumably because participants' looking times for these two components tended to be correlated. In Figure 4, it can be seen that looking times (TFDs) generally declined across the four blocks of problems (some small discontinuities between Blocks 2 and 3 are interpreted as being due to relatively difficult or high-conflict problems in Block 3). The main effect of Domain, with more time allocated to gain information, is also apparent. The interaction itself seems to be due to the fact that the difference in looking time for gains versus losses is very high in Blocks 1 and 2, and

much lower in Blocks 3 and 4. This pattern is more apparent in Figure 3, which plots those gain versus loss differences directly by comparing the difference between gain values and loss values, and in Figure 4.

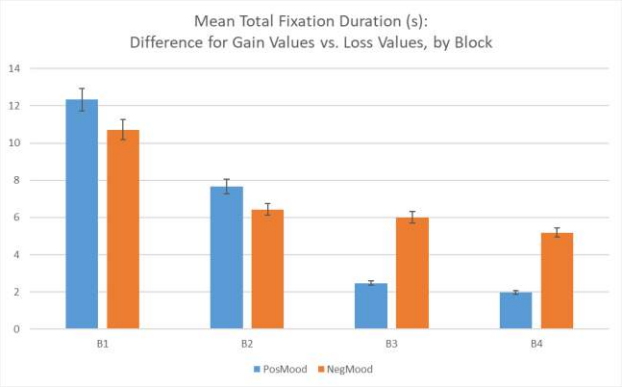


Figure 3. Mean Total Fixation Duration (s): Difference for Gain Values versus Loss Values, by Block and Condition.

Theoretically, the overall pattern of attention results can be explained in at least two ways. First, the positive induction may have had only a transitory effect, as follows. In the first two time-blocks, when the effects of the mood induction were presumably strongest, the pattern of means seems to be consistent with the Affect Infusion Model, such that participants in the positive mood condition paid slightly more attention to information about gains than did negative-mood participants. However, this difference declines in Blocks 3 and 4, perhaps due to a fading effect of the mood induction.

The second interpretation invokes the mood-maintenance hypothesis (MMH). According to the MMH, negative-mood participants focused more on information about potential gains rather than losses, in order to try to attain a positive mood state, and this difference in attention persisted across all four blocks. In contrast, positive-mood participants initially also paid more attention to gains, perhaps to maintain their positive mood. But this effect faded relatively quickly for the positive-mood participants. It is possible that the positive mood induction (watching a silly video) had a more transitory effect than the negative mood induction (watching a clip depicting Nazi murders). But this interpretation should be substantiated via future research.

Conclusion

The results did not provide evidence that induced positive or negative moods can affect the prevalence of analytic processing, at least as measured by EV-maximization success. Nor do they confirm the main predictions of the mood-congruence and mood-maintenance hypotheses regarding attention allocation, as measured by total fixation durations (TFDs). Neither a main effect of condition nor an interaction of condition with domain on attention was found.

However, a significant three-way interaction among Domain, Condition, and Block was found (Figure 4). The nature of the interaction suggests that the mood induction, particularly the positive induction, may have had only a3transitory effect. In the first two blocks, when the effects of the mood induction were presumably strongest, participants in both conditions paid more attention to information about gains, and participants in the positive mood condition paid slightly more attention to information about gains than did negative-mood participants. But this pattern is not confirmed by inferential tests. We should be aware that the samples size was relatively small. In order to detect effects of incidental emotion on such subtle patterns of attention allocation, many more participants would be needed.

Nonetheless, the significant main effects of Type of information and Domain (gain versus less) do suggest some insights into participants' allocation of attention in their decision-making processes. These significant effects suggest that participants, regardless of their assigned mood conditions, allocate more attention to value information than to probabilities, and more attention to gains than to losses. We plan future investigations to confirm and further explore these findings.

Further research in this area might also explore other types of emotion induction manipulations, of varying strengths and durations, to more fully investigate the effects of incidental emotion on decision process and outcomes.

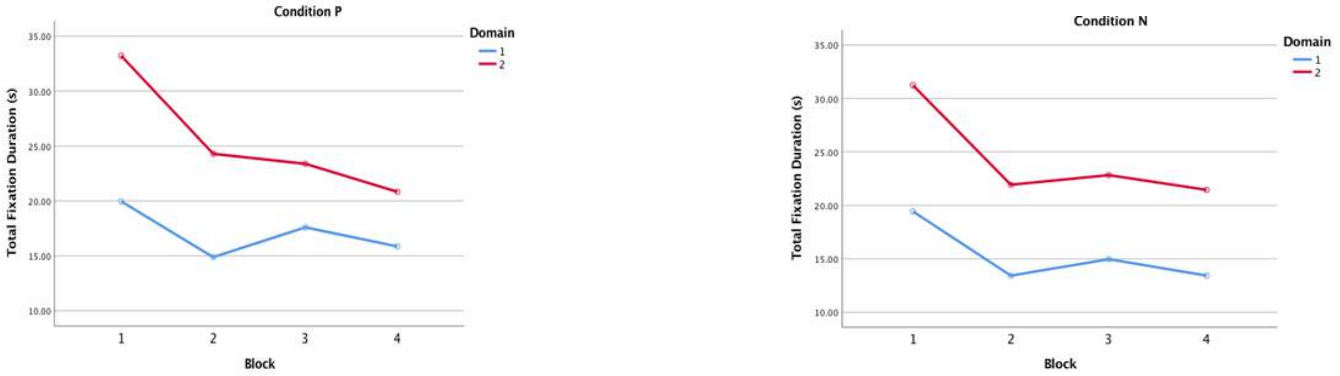


Figure 4. The three-way interaction among mood induction Condition (P=positive mood condition; N=negative mood condition), Domain (1=loss, 2=gain), and Block.

References

Andrade, E. B., & Ariely, D. (2009). The enduring impact of transient emotions on decision making. *Organizational Behavior and Human Decision Processes*, *109*(1), 1-8.

Bernoulli D. (1738). Exposition of a new theory on the measurement of risk. *Econometrica*, 22, 23-36.

Blay, A. D., Kadous, K., & Sawers, K. (2012). The impact of risk and affect on information search efficiency. *Organizational Behavior and Human Decision Processes*, *117*(1), 80-87.

Bolte, A., Goschke, T., & Kuhl, J. (2003). Emotion and intuition: Effects of positive and negative mood on implicit judgments of semantic coherence. *Psychological Science*, *14*(5), 416-421

Buelow, M. T., & Suhr, J. A. (2013). Personality characteristics and state mood influence individual deck selections on the Iowa Gambling Task. *Personality and Individual Differences*, *54*(5), 593-597.

Carpenter, S. M., Peters, E., Västfjäll, D., & Isen, A. M. (2013). Positive feelings facilitate working memory and complex decision making among older adults. *Cognition & emotion*, *27*(1), 184-192.

Chou, K. L., Lee, T., & Ho, A. H. (2007). Does mood state change risk-taking tendency in older adults? *Psychology and aging*, *22*(2), 310.

De Vries, M., Holland, R. W., & Witteman, C. L. (2008). In the winning mood: Affect in the Iowa gambling task. *Judgment and Decision Making*, *3*(1), 42.

Drouvelis, M., & Grosskopf, B. (2016). The effects of induced emotions on pro-social behaviour. *Journal of Public Economics*, *134*, 1-8.

Ellard, K. K., Farchione, T. J., & Barlow, D. H. (2012). Relative effectiveness of emotion induction procedures and the role of personal relevance in a clinical sample: a comparison of film, images, and music. *Journal of Psychopathology and Behavioral Assessment*, *34*(2), 232-243.

Fiedler, K. (1991). On the task, the measures and the mood in research on affect and social cognition. *Emotion and social judgments*, 83-104.

Forgas, J. P. (1995). Mood and judgment: the affect infusion model (AIM). *Psychological Bulletin*, 117, 39–66.

George, J. M., & Dane, E. (2016). Affect, emotion, and decision making. *Organizational Behavior and Human Decision Processes*, *136*, 47-55.

Gerrards-Hesse, A., Spies, K., & Hesse, F. W. (1994). Experimental inductions of emotional states and their effectiveness: A review. *British journal of psychology*, *85*(1), 55-78.

Han S, Lerner JS, Keltner D. 2007. Feelings and consumer decision making: the appraisal-tendency framework. *Journal of Consumer Psychology*, 17, 158–68.

Heilman, R. M., Crişan, L. G., Houser, D., Miclea, M., & Miu, A. C. (2010). Emotion regulation and decision making under risk and uncertainty. *Emotion*, *10*(2), 257.

Hills, A. M., Hill, S., Mamone, N., & Dickerson, M. (2001). Induced mood and persistence at gaming. *Addiction*, *96*(11), 1629-1638.

Isen, A. M. (1984). The influence of positive affect on decision making and cognitive organization. *NA-Advances in Consumer Research Volume 11*.

Isen, A. M., & Patrick, R. (1983). The effect of positive feelings on risk taking: When the chips are down. *Organizational behavior and human performance*, *31*(2), 194-202.

Johnson, E. J., & Tversky, A. (1983). Affect, generalization, and the perception of risk. *Journal of Personality and Social Psychology*, 45, 20–31.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the econometric society*, 263-291.

Keltner, D., & Lerner, J. S. (2010). Emotion. *Handbook of social psychology*.

Kliger, D., & Kudryavtsev, A. (2014). Out of the blue: mood maintenance hypothesis and seasonal effects on investors' reaction to news. *Quantitative Finance*, *14*(4), 629-640.

Lerner, J. S., Li, Y., Valdesolo, P., & Kassam, K. S. (2015). Emotion and decision making. *Psychology*, *66*.

Libby, R., & Fishburn, P. C. (1977). Behavioral models of risk taking in business decisions: A survey and evaluation. *Journal of Accounting Research*, 272-292.

Loewenstein, G., & Lerner, J. S. (2003). The role of affect in decision making. *Handbook of affective science*, *619*(642), 3.

Mishra, S., Morgan, M., Lalumiere, M. L., & Williams, R. J. (2010). Mood and audience effects on video lottery terminal gambling. *Journal of Gambling Studies*, *26*(3), 373-386.

Mishra, S. (2014). Decision-Making Under Risk Integrating Perspectives from Biology, Economics, and Psychology. *Personality and Social Psychology Review*, 1088868314530517.

Niedenthal, R. M., & Setterlund, M. B. (1994). Emotion congruence in perception. *Personality and Social Psychology Bulletin*, *20*, 401-411.

Nygren, T. E., Isen, A. M., Taylor, P. J., & Dulin, J. (1996). The influence of positive affect on the decision rule in risk situations: Focus on outcome (and especially avoidance of loss) rather than probability. *Organizational behavior and human decision processes*, *66*(1), 59-72.

Peters, E., Västfjäll, D., Gärling, T., & Slovic, P. (2006). Affect and decision making: A "hot" topic. *Journal of Behavioral Decision Making*, *19*(2), 79-85.

Raghunathan, R., & Pham, M. (1999). All negative moods are not equal: Motivational influences of anxiety and sadness on decision making. *Organizational Behavior and Human Decision Processes*, *79(1*), 56–77.

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and social psychology review*, *5*(4), 296-320.

Schwarz, N. (2000). Emotion, cognition, and decision making. *Cognition & Emotion*, *14*(4), 433-440.

Schwarz, N., & Clore, G. L. (1996). Feelings and phenomenal experiences. *Social psychology: Handbook of basic principles*, *2*, 385-407.

Slovic, P., & Peters, E. (2006). Risk perception and affect. *Current directions in psychological science*, *15*(6), 322-325.

Vohs, K. D., Baumeister, R. F., & Loewenstein, G. (Eds.). (2007). *Do Emotions Help or Hurt Decisionmaking?: A Hedgefoxian Perspective*. Russell Sage Foundation.

Watson, D., & Clark, L. A. (1999). The PANAS-X: Manual for the positive and negative affect schedule-expanded form.

Westermann, R., STAHL, G., & Hesse, F. (1996). Relative effectiveness and validity of mood induction procedures: analysis. *European Journal of social psychology*, *26*, 557-580.

Yates, J. F. 2007. Emotion appraisal tendencies and carryover: how, why, and… therefore? *Journal of Consumer Psychology*, 17, 179–83.

Yuen, K. S., & Lee, T. M. (2003). Could mood state affect risk-taking decisions? *Journal of affective disorders*, *75*(1), 11-18.

Learning deep taxonomic priors for concept learning from few positive examples

Erin Grant (eringrant@berkeley.edu)
Department of Electrical Engineering & Computer Sciences, University of California, Berkeley

Joshua C. Peterson (joshuacp@princeton.edu)
Department of Computer Science, Princeton University

Thomas L. Griffiths (tomg@princeton.edu)
Departments of Psychology and Computer Science, Princeton University

Abstract

Human concept learning is surprisingly robust, allowing for precise generalizations given only a few positive examples. Bayesian formulations that account for this behavior require elaborate, pre-specified priors, leaving much of the learning process unexplained. More recent models of concept learning bootstrap from deep representations, but the deep neural networks are themselves trained using millions of positive and negative examples. In machine learning, recent progress in meta-learning has provided large-scale learning algorithms that can learn new concepts from a few examples, but these approaches still assume access to implicit negative evidence. In this paper, we formulate a training paradigm that allows a meta-learning algorithm to solve the problem of concept learning from few positive examples. The algorithm discovers a taxonomic prior useful for learning novel concepts even from held-out supercategories and mimics human generalization behavior—the first to do so without hand-specified domain knowledge or negative examples of a novel concept.

Keywords: concept learning; deep neural networks; object taxonomies

Introduction

One of the hallmarks of human intelligence is the ability to rapidly learn new concepts given only limited information (Lake et al., 2016). This task is difficult because we are often presented with only a handful of (positive) examples of a new concept, and no examples outside of the concept (negative examples). Quine (1960) was the first to recognize that this poses a seemingly crippling problem for induction: hearing only the word "gavagai" as a rabbit passes by, we have no way of knowing with certainty whether the new word applies to all animals, all rabbits, one pet rabbit, potential food, or any other of a nearly infinite number of likewise compatible hypotheses.

Nevertheless, humans appear to possess prior knowledge, whether learned, innate, or both, that makes for effective generalizations even under such conditions. In some situations, these constraints are simple and easy to model (Tenenbaum, 1999; Tenenbaum & Griffiths, 2001; Kemp et al., 2007). However, in general, modeling the rich prior knowledge that humans bring to bear on problems in complex domains such as natural images is difficult and reliant on explicit domain knowledge (Xu & Tenenbaum, 2007; Jia et al., 2013). A recent line of follow-up work has made strides by using deep neural networks as a proxy for psychological representations (Campero et al., 2017; Peterson, Soulos, et al., 2018). Although these representations are largely perceptual, they are nevertheless an improvement over hand-specified features given that they are less prone to experimenter bias and have been shown to explain some aspects of human visual representations (Peterson, Abbott, & Griffiths, 2018). However, unlike most cognitive models of concept learning and unlike humans, these networks are trained on millions of both positive and negative examples of mutually exclusive categories. Moreover, they fail to capture the taxonomic biases that humans bring to bear in concept learning (Peterson, Abbott, & Griffiths, 2018).

Challenged by the cognitive science community (Lake et al., 2015), machine learning researchers have developed a number of their own improvements to deep learning algorithms to tackle the problem of learning from few examples (e.g., Vinyals et al., 2016; Ravi & Larochelle, 2017). These approaches constitute impressive new candidate accounts of human concept learning from naturalistic stimuli, but differ from human learning scenarios in that they (1) rely on negative evidence to infer the extent of a novel concept, and (2) ignore the overlapping and hierarchical structure of real-world concepts that humans use to inform their generalization judgments (Rosch et al., 1976; Xu & Tenenbaum, 2007).

In the following paper, we aim to address many of the shortcomings of previous work by demonstrating how a deep meta-learning algorithm combined with a novel stimulus sampling procedure can provide an end-to-end framework for modeling human concept learning, for the first time with no hand-specified prior knowledge or negative examples of a novel concept. We introduce a new, taxonomically structured dataset of concepts compiled by sampling from both internal nodes and leaf nodes within the ImageNet hierarchy (Deng et al., 2009). Our method learns concepts at different levels of this hierarchy, but the hierarchical structure itself is never provided to the model explicitly at any point. To evaluate our model against human behavior, we present a new human benchmark inspired by Rosch's classic object taxonomies (Rosch et al., 1976). Our model not only mimics human generalization behavior, reproducing classic generalization gradients (Shepard, 1987; Xu & Tenenbaum, 2007), but also encompasses a general taxonomic prior that allows for human-like generalization even when presented with novel concepts from different image taxonomies (*i.e.,* held-out supercategories).

Background

Computational models of concept learning in cognitive science have historically focused on the problem of density estima-

tion (Ashby & Alfonso-Reese, 1995). Under this paradigm, learning about a category C amounts to the estimation of the density $p(x \mid C)$, where x represents the space of stimuli. This modeling framework assumes that a density can be learned for each of a set of mutually exclusive categories, where positive examples from one category implicitly serve as negative examples for all other categories. However, the conditions under which humans learn concepts are rarely this straightforward.

Learning concepts from few positive examples. More recent work has begun to examine how humans learn concepts in more natural settings where often only a few positive examples of a single concept are provided. Despite this impoverished learning environment, even young children are able to generalize surprisingly well (Carey, 1978; Markman, 1991). Extending Shepard (1987), Tenenbaum (1999) and Tenenbaum and Griffiths (2001) formalize the concept learning problem as follows: Given n positive examples $\mathbf{x} = \{x_1, ..., x_n\}$ of a concept C, the learner estimates the probability $p(x^* \in C \mid \mathbf{x})$ that a new stimulus x^* is also an example of that concept. The challenge the learner faces in making such a generalization is that the extension of C is underspecified (*i.e.,* it could include only the present examples, all possible stimuli, or anything in between). To address this challenge, the authors propose a Bayesian generalization model that averages the predictions made by a number of hypotheses about the extent of C. By making the plausible assumption that learners expect examples to be randomly sampled from concepts, the authors show that smaller hypotheses will be preferred, thus deriving constraints on the expected extent of C.

Armed with this framework, Xu and Tenenbaum (2007) conducted an extensive analysis of human generalization behavior through *word learning* experiments. Participants were given either one or three examples of a new concept such as "dax" and asked to pick out other instances of that concept from a set of test stimuli. The examples of each concept were unique images that could be drawn from either a subordinate-level (e.g., Dalmatian), basic-level (e.g., dog), or superordinate-level (e.g., animal) category, and the test stimuli were sampled from all three levels. An example of this task is shown in Figure 1. Replicating Shepard (1987), the authors found that generalization from a single example of a concept to a test stimulus decreases with psychological similarity. However, their experiments also yielded two new insights into human concept learning:

1. Given multiple examples of a concept, generalization goes only as far at the most specific level that contains those examples. For example, shown three examples from different dog breeds, other dog breeds are included in the concept at test time, but not other animals.
2. There is a bias towards generalizing to test items at the basic level, in particular when only a single subordinate example is shown. For example, given a single example of a Dalmatian, participants predictably generalize the concept to other Dalmatians, but also generalize to other breeds.

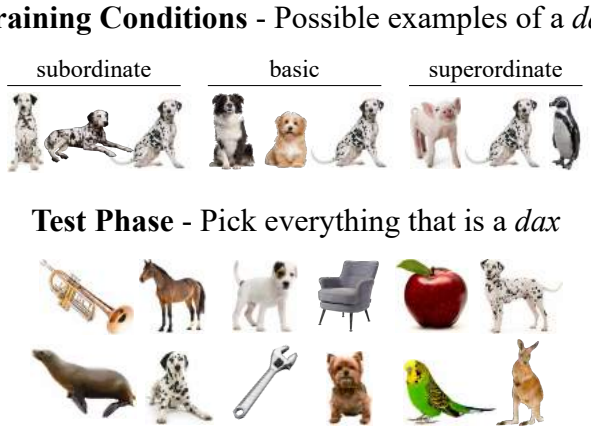


Figure 1: The *word learning* paradigm from Xu and Tenenbaum (2007). In each trial, participants see a few instances exemplifying a novel word such as "dax" and are asked to select other instances that fall under the same word from a test array. The training conditions vary by the levels of the underlying image taxonomy from which the instances are drawn, *e.g.,* Dalmatians (subordinate) vs. dogs (basic) vs. animals (superordinate).

The only modification to the Bayesian concept learning model required to capture these data was a structured, taxonomic prior computed from human similarity judgments over the set of objects used in the experiments. While this work constitutes one of the first successful attempts to explain concept learning in realistic contexts, it arguably leaves much of the structured, taxonomic representation assumed and raises questions about how this knowledge is acquired.

The role of prior knowledge. Given the aforementioned dependence on highly structured priors in explaining people's robust generalization behavior, subsequent work has focused on incorporating this information into the modeling of human concept learning. Jia et al. (2013) provided an automated framework for modeling human generalization behavior by leveraging perceptual stimulus features provided by a computer vision algorithm along with information contained in the WordNet taxonomy (Fellbaum, 1998), but gave no account for how this information is learned by humans. Kemp et al. (2007) provided the first account of how such knowledge could be acquired: The authors start with an unstructured representation and apply a structured hierarchical Bayesian model that learns taxonomic abstractions from data. Despite its elegance, the method does not immediately scale to high-dimensional stimuli such as the images used in Jia et al. (2013).

Deep neural networks (LeCun et al., 2015) have served as both candidate models of object perception and rich image representations that can be used for cognitive modeling. However, these model do not capture even coarse taxonomic information out-of-the-box (Peterson, Abbott, & Griffiths, 2018). Despite this, Peterson and Griffiths (2017) found that the sampling assumptions of Bayesian concept learning could be verified in human generalization judgments when modeling stimuli using deep feature representations. Campero et al. (2017) deployed a hierarchical model similar to Kemp et al. (2007) over a deep

Superordinate	Basic	Subordinates	
Musical Instrument	Guitar	Acoustic guitar	Electric guitar
	Piano	Grand piano	Upright piano
	Drum	Tambourine	Bass drum
Fruit	Apple	Delicious apple	Mackintosh apple
	Currant	Black currant	Red currant
	Grapes	Concord grapes	Thompson seedless grapes
Tool	Hammer	Ball-peen hammer	Carpenter's hammer
	Saw	Hack saw	Cross-cutting saw
	Screwdriver	Phillips screwdriver	Flat tip screwdriver
Clothing	Trousers	Jeans	Sweat pants
	Socks	Athletic socks	Knee-high socks
	Shirt	Dress shirt	Polo shirt
Furniture	Table	Kitchen table	Dining-room table
	Lamp	Floor lamp	Table lamp
	Chair	Armchair	Straight chair
Vehicle	Car	Sports car	Sedan car
	Airplane	Airliner plane	Fighter jet plane
	Truck	Pickup truck	Trailer truck
Fish	Snapper	Grey snapper	Red snapper
	Trout	Rainbow trout	Lake trout
	Salmon	Atlantic salmon	Chinook salmon
Bird	Owl	Barn owl	Great gray owl
	Eagle	Bald eagle	Golden eagle
	Sparrow	Song sparrow	Field sparrow

Table 1: The eight taxonomies adapted from Rosch et al. (1976).

feature space and found both good one-shot learning performance as well as the ability to recover some stimulus clusters representative of human categorization judgments. Noting that most deep networks are trained using subordinate-level labels, Peterson, Soulos, et al. (2018) trained a deep neural network with coarser, basic-level labels to more closely mimic the supervision children receive. A relatively simple generalization model over the resulting representation reproduced both the basic-level bias and the gradient of generalization from Xu and Tenenbaum (2007).

Few-shot learning in machine learning. The problem facing cognitive models of concept learning is closely related to *one-* or *few-shot* classification in machine learning, in which the aim is to learn to discriminate between classes given only a few labeled examples from each class (Fei-Fei et al., 2003; Vinyals et al., 2016). A powerful solution to few-shot learning is *meta-learning*, where learning episodes—themselves consisting of training and testing intervals—are used to train a model to adapt quickly to solve a new task given only a small amount of labeled task data (Schmidhuber, 1987). The learning episodes are leveraged in the form of a data-driven prior that is combined with a small amount of test-time evidence (*i.e.,* a few "shots" of labeled data from a novel task) in order to make a test-time inference.

Modeling Approach

We propose to bridge cognitive science and machine learning by formulating concept learning as a few-shot learning problem. As we will see, the meta-learning problem formulation allows a machine learning model to estimate a decision boundary from only positive samples of a class, similarly to how people learn concepts from only a few positive examples. Moreover, the use of a meta-learning algorithm provides a principled way to present entirely novel concepts at test time as held-out test *tasks*. As such, we can investigate the taxonomic priors encoded in a neural network embedding function, as compared to prior work that examines the representations of images from categories observed during training time (Peterson, Soulos, et al., 2018).

Concept learning as meta-learning. Meta-learning algorithms aim to learn how to learn by extracting task-general knowledge through the experience of solving a number of specific tasks (Thrun & Pratt, 1998; Hochreiter et al., 2001). In the case of concept learning, the jth task corresponds to learning a decision boundary for the jth concept using only positive examples, and meta-learning corresponds to learning how to estimate decision boundaries for arbitrary unseen concepts. We can thus formalize the concept learning problem as the task of predicting a target label y (which indicates whether or not the input belongs to a given category) from an input observation x (i.e., an image). Note that this formulation differs from the standard discriminative classification problem, where the task corresponds to a K-way discriminative classification task in which each of the K class labels are mutually exclusive.

Formally, let $\mathcal{T}_j = (\boldsymbol{X}_j^{\text{trn}}, \boldsymbol{Y}_j^{\text{trn}}, \boldsymbol{X}_j^{\text{val}}, \boldsymbol{Y}_j^{\text{val}})$ denote a task drawn from a given task distribution $p(\mathcal{T})$, where $\boldsymbol{X}_j^{\text{trn}}$ and $\boldsymbol{Y}_j^{\text{trn}}$ are a small collection of training inputs and labels, disjoint from validation samples $\boldsymbol{X}_j^{\text{val}}$ and $\boldsymbol{Y}_j^{\text{val}}$ but belonging to the same task $\mathcal{T}_j$. A meta-learning algorithm (*e.g.,* Vinyals et al., 2016; Ravi & Larochelle, 2017) aims to estimate parameters $\boldsymbol{\theta}$ that can be adapted to solve an unseen task $\mathcal{T}_j \sim p(\mathcal{T})$, using only the training samples $(\boldsymbol{X}_j^{\text{trn}}, \boldsymbol{Y}_j^{\text{trn}})$, to ensure the updated model achieves good performance on the validation samples $(\boldsymbol{X}_j^{\text{val}}, \boldsymbol{Y}_j^{\text{val}})$ according to some loss function $\mathcal{L}$.

In this work, we use the *model-agnostic meta-learning* (MAML; Finn, Abbeel, & Levine, 2017) algorithm, which formulates meta-learning as estimating the parameters $\boldsymbol{\theta}$ of a model so that when one or a few gradient descent steps are taken from the initialization at $\boldsymbol{\theta}$ on the training data $(\boldsymbol{X}_j^{\text{trn}}, \boldsymbol{Y}_j^{\text{trn}})$, the updated model has good generalization performance on that task's validation set, $(\boldsymbol{X}_j^{\text{val}}, \boldsymbol{Y}_j^{\text{val}})$. At test time, a new task from the test set is presented to the model for few-shot adaptation, *i.e.,* gradient descent with $(\boldsymbol{X}_j^{\text{trn}}, \boldsymbol{Y}_j^{\text{trn}})$, and computation of test-time performance metrics, *e.g.,* accuracy on $(\boldsymbol{X}_j^{\text{val}}, \boldsymbol{Y}_j^{\text{val}})$. The training examples in the inner gradient computation are strictly positive examples (*i.e.,* $\boldsymbol{Y}_j^{\text{trn}} = 1$) of a particular concept j, whereas validation examples in the outer gradient computation include both positives and negatives (*i.e.,* $\boldsymbol{Y}_j^{\text{val}} \in \{0, 1\}$); thus, at test time, the meta-learning algorithm is able to estimate a decision boundary for a novel concept from only positive examples of that concept.

Behavioral Experiment

In order to compare our method directly to human behavior, we conducted a large-scale human generalization experiment using a test set of naturalistic stimuli used for the simulations in the next section. We assess generalization behavior using a concept learning experiment that follows previous work on Bayesian concept and word learning (Xu & Tenenbaum, 2007; Abbott et al., 2012; Jia et al., 2013).

Stimuli. We mapped a subset of the graph structure embedded in the ImageNet dataset used for the ImageNet Large Scale Visual Recognition Competition (ILSVRC; Russakovsky et al., 2015) to the classic taxonomy used by cognitive scientists and developed by Rosch et al. (1976). ILSVRC is a commonly used object-classification dataset that contains more than 1 million images distributed across 1000 categories. Instead of using the leaf classes as categories, we create concepts by picking a node in the ImageNet hierarchy and sampling images from leaves dominated by the given node. Note that, in this case, concepts are not necessarily mutually exclusive in the sense that a single image may belong to one or more classes (*e.g.,* a Dalmatian may be labeled as both a *dog* and an *animal*). If the exact subordinate node from Rosch et al. (1976) was not available in ImageNet, we found a close semantic match via the WordNet (Fellbaum, 1998) taxonomy. We provide the full taxonomy for this dataset in Table 1.

Task. In each of 8 trials, participants observed 5 images of a single concept, sampled from one of the three levels of taxonomic abstraction. For instance, in a subordinate training condition, the examples could be all Dalmatians; in a basic-level training condition, all dogs; in a superordinate training condition, all animals. To test generalization behavior, participants were then given a test array of 24 images and were asked to pick which images also belonged to the learned concept. The test array comprised 2 subordinate matches (*e.g.,* other Dalmatians), 2 basic-level matches (*e.g.,* other breeds of dog), 4 superordinate matches (*e.g.,* other animals), and 16 out-of-domain items (*e.g.,* inanimate objects), following Xu and Tenenbaum (2007). See Figure 2 for an example set of training and test stimuli. In total, we collected data for 180 unique trials and 1 180 unique images.

Participants. We recruited 900 unique participants from Amazon Mechanical Turk to each complete 8 trials as described above, one randomly sampled for each of the superordinate categories. The test sets were fixed within a superordinate category. Participants were paid $0.40 each.

Results. Figure 3 (a) presents the results of the behavioral experiment for each of the three taxonomic levels. As expected on the basis of previous work, there is an exponentially decreasing generalization gradient as the level of taxonomic abstraction of the test matches (bar color) increases. However, this effect diminishes as the intra-class variation of the few-shot examples (x-axis) increases: Moving from the *subordinate* condition to the *basic*-level condition, we find an increase in the number of basic-level matches selected from the test set. The condition in which there is greatest intra-class variation–the superordinate condition–exhibits only a small generalization gradient.

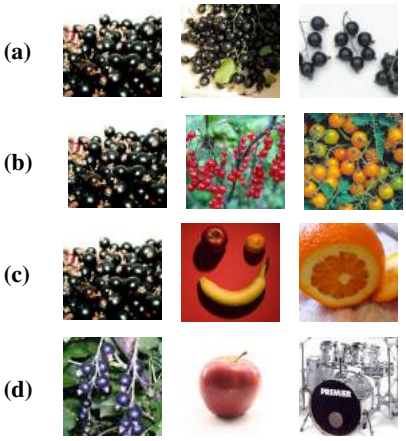


Figure 2: Examples of training stimuli for the **(a)** subordinate, **(b)** basic-level, and **(c)** superordinate level training conditions, as well as **(d)** a subset of the stimuli from the test array for a specific concept learning task (here, learning the concept *black currant* **(a)**, *currant* **(b)** or *fruit* **(c)**). The test array **(d)** displays, from left to right, a subordinate match, a basic-level match and a superordinate match.

Meta-Learning Simulations

Our modeling goal is to investigate whether we can use meta-learning to learn new concepts from only few positive examples, even though these concepts are potentially overlapping and therefore not mutually exclusive. Furthermore, we aim to investigate whether a meta-learning algorithm is able to use information about the underlying concept taxonomy that generates observations of the extension of a concept in order to generalize to novel concepts in a human-like manner.

Meta-learning formalism. Our model observes K positive examples $\boldsymbol{x} = \{x_1, \ldots, x_K\}$ of a concept C, and must learn the generalization function $p(x^* \in C)$ to correctly identify whether a novel example x^* is also a member of the concept. Training proceeds as follows: A concept index j is sampled from the meta-training set. Then, for K-shot learning, $2K$ positive examples of the concept and K negatives are sampled. The parameters $\boldsymbol{\theta}$ are adapted using K of the positives, and then the model is optimized with a loss computed using the remaining positive and negative examples of the concept. At test time, the model with trained parameters θ is presented with K positive examples from a new concept in the test set; the model adapts $\boldsymbol{\theta}$ and is evaluated on its ability to distinguish new positive examples of that concept from negatives.

Taxonomic dataset construction. For training and validation, we created a large-scale taxonomy of classes by using the graph structure embedded in the subset of the ImageNet dataset used for the ImageNet Large Scale Visual Recognition Competition (ILSVRC; Russakovsky et al., 2015), similar to the behavioral experiment described earlier, but using the entirety of the ImageNet hierarchy. We then created few-shot concept learning tasks for training by sampling positive and negative examples for each concept, where negative examples of a concept are generated by sampling from the complement set of leaf nodes. Superordinate-level nodes are not shared between training, validation, and test to ensure that test-time generalization is measured on novel concepts. We use 494, 193, and 223 leaf nodes in the training, validation, and test sets, respectively (*c.f.,* 80, 20, and 20 in the few-shot classifi-

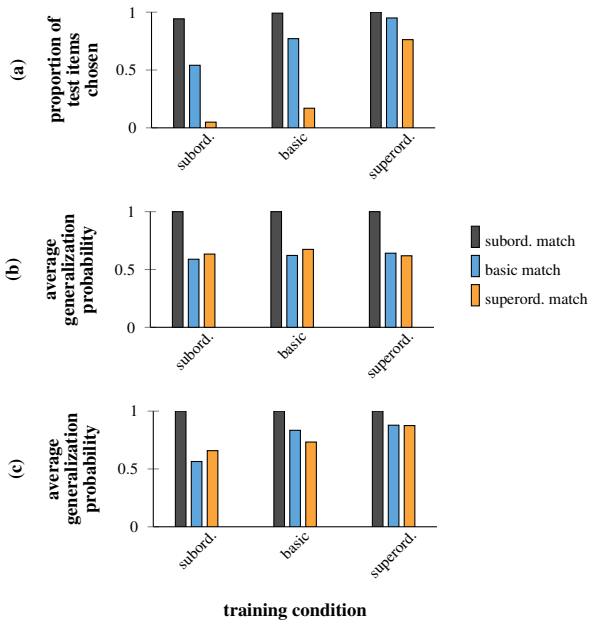


Figure 3: Human behavioral data **(a)** and `flat` **(b)** and `hier` **(c)** modeling results on the concept generalization task. The horizontal axis identifies the training condition (*i.e.,* the level of taxonomic abstraction from which the few-shot examples are drawn). The vertical axis identifies, for each type of match in {subordinate, basic-level, superordinate}, the proportion of selections from the test array **(a)**, or the average probability of generalization **(b, c)**.

cation dataset *mini*ImageNet (Vinyals et al., 2016)). The training, validation, and test node sets do not comprise all of the nodes in the ImageNet hierarchy, as some nodes are redundant (*i.e.,* have a single parent) or are too abstract to appropriately define a visual concept (*e.g., physical entity*, *substance*, *equipment*). We make use of the training and validation dataset for training and hyperparameter selection, respectively; the test set is not used in this work but reserved for future works that may wish to perform large-scale evaluation of concept learning. Instead, the evaluations reported in this work are performed on the Rosch-inspired human benchmark described above. We also wish to emphasize that while we make use of the ImageNet hierarchy, we do so only to generate a natural distribution of concepts to learn from, and never present the explicit hierarchical relations to the model at any time.

We consider two dataset conditions in our simulations: In the `hier` dataset condition, the meta-learning algorithm observes concepts sampled from the internal and leaf nodes of the ImageNet hierarchy, and thus can learn a taxonomic prior; in the `flat` dataset condition, the algorithm observes only leaf-node concepts, and thus has no access to such information.

Hyperparameters. The base model that is optimized by model-agnostic meta-learning (MAML) is a binary classifier consisting of a convolutional neural network with a sigmoid output.[1] In our experiments, we downsample the images to each have a width and height of 84 pixels, as is common in the use of *mini*ImageNet (Vinyals et al., 2016) as a few-shot learning dataset. We select hyperparameters on the same hierarchically structured validation set for both the `hier` and `flat` dataset conditions and evaluate algorithms after a fixed number of training iterations (40K with a batch size of 4). We take the value of the scalar output of the network evaluated on a test example as the *generalization probability* and average this quantity across all test examples from a specific level of taxonomic match to produce the *average generalization probability*. When reporting the average generalization probability metric, we standardize each set of probabilities for each training condition by treating the distractor (out-of-domain) generalization probability as a baseline of zero and further dividing by the largest probability in the set. In line with prior work (Peterson, Abbott, & Griffiths, 2018), this highlights the quantity of interest: the relative differences in average generalization probabilities across the subordinate, basic-level, and superordinate levels of the taxonomy.

Results. The generalization gradient observed in humans is also exhibited by the `hier` dataset condition in Figure 3 (c): When the few-shot examples are taken from a basic-level category (the *basic* condition; *e.g.,* different breeds of dog) as opposed to a subordinate category (the *subord.* condition; *e.g.,* Dalmatians), the model generalizes to more basic-level matches (*e.g.,* different dog breeds) from the test array. In the plot, this can be seen by comparing the ratio of subordinate generalization (black column) to basic-level generalization (blue column) within each training condition (*i.e.,* the gap between the black and blue bars is diminished in the *basic* condition *vs.* the *subord.* condition). Furthermore, when the few-shot examples are taken from a superordinate category (*superord.* condition), both the model in the `hier` dataset condition and humans are equally likely to pick subordinate, basic-level, or superordinate matches from the test array. In Figure 3 (a, c), this can be seen as the generalization to all levels of the taxonomy (black, blue, and yellow bars) being close to equal.

One notable departure of Figure 3 (c), from the human generalization behavior in Figure 3 (a), is overgeneralization to the superordinate category in the subordinate training condition, and to a lesser extent, in the basic-level training condition, suggesting that it is difficult for the algorithm to discriminate between basic-level and superordinate matches given only subordinate examples of a concept. Nevertheless, in comparison to the `flat` dataset condition in Figure 3 (b), which does not change generalization behavior on the basis of the training condition, the behavior of the algorithm exposed to the hierarchically structured `hier` dataset suggests a learned sensitivity to the underlying taxonomic organization of new concepts.

[1] The architecture of the model is similar to prior work in meta-learning (e.g., Ravi & Larochelle, 2017) with 4 convolutional layers each with 32 3×3 filters, leaky ReLU activation functions with a slope of 0.2, and 2×2 max-pooling, all followed by a linear layer with sigmoid activation. We do not employ batch normalization because of strong batch interdependence, as all of the training examples for a concept are of the same (positive) class.

Discussion

When humans are presented with an example from a new concept, they can quickly infer which other instances belong to that same concept even without the strong constraints provided by negative examples. In order to achieve this feat, humans bring to bear information about the taxonomic structure of natural categories. Targeting the robustness of human generalization even in highly novel domains (Schmidt, 2009), we investigated the extent to which taxonomically structured biases for complex, naturalistic stimuli taken from the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) could be acquired and leveraged to learn the extent of novel concepts from only a few positive examples. In contrast to previous work (Peterson, Abbott, & Griffiths, 2018), we validate the generalization behavior of our model using *unseen* supercategories drawn from the superordinate levels of Rosch's classic taxonomy (Rosch et al., 1976).

While our method is successful in both learning a general taxonomic prior and exhibiting human-like generalization behavior, there is room for improvement as the quantitative gradients are not a perfect match to humans. However, it should be noted that our model faces the atypically challenging task of both learning a highly structured representation for complex stimuli and making use of it to generalize to entirely novel concepts. As such, this framework draws on many of the strengths of both cognitive models and deep neural networks in machine learning, and constitutes the most comprehensive account of human visual concept learning to date. Lastly, we note that we do not build in any explicit preference for simple concepts or attention to the number of examples (Tenenbaum, 1999; Peterson, Soulos, et al., 2018), although this may be an interesting avenue for improvement in future work.

Acknowledgments. This work was supported by the Defense Advanced Research Projects Agency (DARPA) under the Lifelong Learning Machines (L2M) program via grant number HR001117S0016 and by the National Science Foundation (NSF) under grant number 1718550. The views and opinions expressed in this paper are those of the authors and do not reflect the official policy or position of the U.S. Government.

References

Abbott, J. T., Austerweil, J. L., & Griffiths, T. L. (2012). Constructing a hypothesis space from the web for large-scale Bayesian word learning. In *Proceedings of the 34th annual meeting of the cognitive science society (cogsci)*.

Ashby, F. G., & Alfonso-Reese, L. A. (1995). Categorization as probability density estimation. *Journal of mathematical psychology*, *39*(2), 216–233.

Campero, A., Francl, A., & Tenenbaum, J. B. (2017). Learning to learn visual object categories by integrating deep learning with hierarchical bayes. In *Cogsci*.

Carey, S. (1978). The child as word learner. In M. Halle, J. Bresnan, & G. Miller (Eds.), *Linguistic theory and psychological reality* (pp. 264–293). MIT Press.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *Proceedings of the ieee conference on computer vision and pattern recognition (cvpr)* (pp. 248–255).

Fei-Fei, L., et al. (2003). A Bayesian approach to unsupervised one-shot learning of object categories. In *Proceedings of the 9th conference on computer vision and pattern recognition (cvpr)* (pp. 1134–1141).

Fellbaum, C. (1998). *Wordnet*. Wiley Online Library.

Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th international conference on machine learning (icml)*.

Hochreiter, S., Younger, A., & Conwell, P. (2001). Learning to learn using gradient descent. *Proceedings of the International Conference on Artificial Neural Networks (ICANN)*, 87–94.

Jia, Y., Abbott, J. T., Austerweil, J. L., Griffiths, T., & Darrell, T. (2013). Visual concept learning: Combining machine vision and Bayesian generalization on concept hierarchies. In *Advances in neural information processing systems (nips) 26* (pp. 1842–1850).

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental science*, *10*(3), 307–321.

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, *350*(6266), 1332–1338.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2016). Building machines that learn and think like people. *Behavioral and Brain Sciences*, 1–101.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436.

Markman, E. M. (1991). *Categorization and naming in children: Problems of induction*. Mit Press.

Peterson, J. C., Abbott, J. T., & Griffiths, T. L. (2018). Evaluating (and improving) the correspondence between deep neural networks and human representations. *Cognitive science*, *42*(8), 2648–2669.

Peterson, J. C., & Griffiths, T. L. (2017). Evidence for the size principle in semantic and perceptual domains. *arXiv preprint arXiv:1705.03260*.

Peterson, J. C., Soulos, P., Nematzadeh, A., & Griffiths, T. L. (2018). Learning hierarchical visual representations in deep neural networks using hierarchical linguistic labels. *arXiv preprint arXiv:1805.07647*.

Quine, W. V. O. (1960). Word and object, 1960. *Le mot et la chose*, 1977–2000.

Ravi, S., & Larochelle, H. (2017). Optimization as a model for few-shot learning. In *Proceedings of the 5th international conference on learning representations (iclr)*.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*(3), 382–439.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, *115*(3), 211–252.

Schmidhuber, J. (1987). *Evolutionary principles in self-referential learning* (Unpublished doctoral dissertation). Institut für Informatik, Technische Universität München.

Schmidt, L. A. (2009). *Meaning and compositionality as statistical induction of categories and constraints* (Unpublished doctoral dissertation). Massachusetts Institute of Technology.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323.

Tenenbaum, J. B. (1999). *A Bayesian framework for concept learning* (Unpublished doctoral dissertation). Massachusetts Institute of Technology.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and brain sciences*, *24*(04), 629–640.

Thrun, S., & Pratt, L. (1998). *Learning to learn*. Kluwer Academic Publishers.

Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al. (2016). Matching networks for one shot learning. In *Advances in neural information processing systems (nips) 29* (pp. 3630–3638).

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, *114*(2), 245–272.

A Surprising Density of Illusionable Natural Speech

Melody Y. Guan
Stanford University, Stanford, California, United States

Gregory Valiant
Stanford University, Stanford, California, United States

Abstract

Recent work on adversarial examples demonstrates a brittleness of many state-of-the-art machine learning systems. We investigate one human analog, asking: What fraction of natural speech can be turned into illusions which alter humans perception or result in different people having significantly different perceptions? Using generated videos, we first empirically estimate that 17% of words occurring in natural speech have some susceptibility to the McGurk effect–the phenomenon by which adding a carefully chosen video clip to the audio channel affects the viewers perception of the message. We develop a bag-of-phonemes prediction model for word-level illusionability that we extend with natural language modeling to build a sentence-level framework. We train an instantiation using Amazon Mechanical Turk evaluations on sentence-level illusions. Finally we generate several new instances of the Yanny/Laurel illusion, demonstrating that it is not an isolated occurrence. The surprising density of illusionable instances warrants further investigation from cognitive and security perspectives.

Stopping Rules In Information Acquisition At Varying Probabilities And Consequences: An Integrated Psychophysiological Measures Approach

Roberto Guedes de Nonohay
UFRGS, Porto Alegre, Brazil

Gustavo Gauer
UFRGS, Porto Alegre, Brazil

Richard Gonzalez
University of Michigan, Ann Arbor, Michigan, United States

Guilherme Lannig
UFRGS, Porto Alegre, Brazil

Abstract

An experiment aiming to assess the use of stopping rules in information acquisition was performed. An exploratory experimental paradigm was used. Participants (47 healthy individuals) were requested to make a decision in 24 financial scenarios with the possibility of buying information pieces. Participants were able to accept, reject or choose not to decide. Behavioral, EEG, ECG and Eyetracker data were recorded and integrated offline for analysis. Results showed that participants followed primarily Bayesian calculations in order to determine when to cease information acquisition and decide. Participants would tend to rely more on the valences (BAL) of the information acquired (positive or negative) than on sheer quantity. Acceptance tended to be made with mean positive BAL, rejection with mean negative BAL and procrastination with mean zero BAL. Uncertainty was seen to affect the information acquisition and decision process; EEG data suggest Slow Cortical Potentials at fronto-central electrodes for risk with low consequences and uncertainty with high consequences. Eyetracker data shows greater mean fixation time for decisions and information areas of interest (AOI). Heart rate data shows no difference in scenarios and/or information acquisition behavior, meaning that the decision scenarios did not elicit significant emotional engagement. Integrated psychophysiological measures were of important assistance to the conclusions given that they provided information as to what happened or not both behaviorally and physiologically.

Resource-Rich versus Resource-Poor Assessment in Introductory Computer Science and its Implications on Models of Cognition: An in-Class Experimental Study

Tobias Halbherr[1, 2] (tobias.halbherr@gess.ethz.ch), Hermann Lehner[3] (hermann.lehner@inf.ethz.ch), Manu Kapur[1] (manukapur@ethz.ch)

[1]ETH Zurich; Department of Humanities, Social and Political Sciences; Institute of Learning Sciences and Higher Education

[2]ETH Zurich; Educational Development and Technology

[3]ETH Zurich; Department of Computer Science

Abstract

Outside university, students encounter disciplinary practices mediated by technological resources. In this sense, the real world is decidedly resource-rich. In contrast, most educational assessments remain decidedly resource-poor. Situated versus mindbased perspectives of cognition fundamentally differ in the role they ascribe to such resources in cognition and learning. To mindbased perspectives, they are a source of input, to situated perspectives they are constitutive to cognition itself. We assessed the validity of resource-rich versus resource-poor assessments of learning outcomes from resource-rich versus resource-poor learning activities. The study implemented an in-class 2x2 between-subjects experimental design in an introductory programming course with 192 first semester BSc engineering students. Both types of assessment were sensitive to differences in learning outcomes, indicating validity for both. Results indicate resource-rich assessments may be more ecologically valid, while – intriguingly – the resource-poor assessments were more sensitive to transfer of learning. Furthermore, the resource-rich learning activities better facilitated learning for transfer.

Keywords: assessment; examinations; resource-rich assessment; resource-affordances; higher education; learning science; computer science education; e-assessment; educational technology; situated cognition

Introduction

Examinations in (higher) education usually remain restricted to pen and an empty piece of paper – or in their computer-based counterpart, keyboard, mouse, and a standardized e-assessment environment. What examinations typically lack – indeed prohibit – is access to any additional resources. In this sense, conventional examinations are *resource-poor*. In contrast, upon leaving university students will usually have access to a wide array of resources, such as specialist tools, easy access to information, and support from networks of experts and peers. In this sense, most professional practices outside the classroom are decidedly *resource-rich*. However, if the practices in the real world – for which we ultimately learn – are resource-rich, how can we justify a resource-poor examination practice? Conversely, how could we demonstrate the need for examinations to become resource-rich? In this study, we render first empirical evidence unto this question for the case of *tools* as resource. Specifically, we are interested in the question whether the availability or absence of disciplinary technological tools in an assessment environment has implications on the *validity* of the corresponding assessments of learning.

Our research question lies at the intersection of three larger topics which to date have rarely been linked. First, the above-mentioned discrepancy between (increasingly) resource-rich disciplinary practices versus resource-poor conventional examination practice and its implication on validity. Second, the resurgent epistemic debates on appropriate perspectives of cognition and learning. Third, advancements in educational technology, which enable novel learning and assessment environments. We will briefly elaborate on each.

Resource-Rich Assessment

When asked to formulate intended learning outcomes, lecturers typically emphasize outcomes associated with deep learning, transferrable skills, and rich conceptual understanding. Conventional examinations in contrast, are frequently associated with surface learning, cramming, factual recall, poor retention, and an inability to transfer (Biggs, 2014; Keehner, Gorin, Feng, & Katz, 2017). In other words, there seems to be a problem with the validity of conventional examinations: Lecturers intend to assess outcomes associated with deep learning, but in effect, students may achieve success through surface learning. To make matters worse, examinations strongly motivate student learning and when examinations reward surface learning, they also encourage surface learning. Assessment drives learning (Baird, Andrich, Hopfenbeck, & Stobart, 2017) and poor assessment drives poor learning.

Alternative Assessment (Sambell, McDowell, & Brown, 1997), Authentic Assessment (Gulikers, Bastiaens, & Kirschner, 2004), Assessment for Learning (Baird et al., 2017), or Performance Assessment (Moss, 1992) all share with our proposition for resource-rich assessment a concern for the above-mentioned issues with assessment validity and/or assessment driven learning. However, none of these approaches foreground the access to relevant disciplinary resources (tools, information, and/or social interactions). We propose that the absence of relevant disciplinary resources in assessment contexts may be a crucial mediator of longstanding issues with both assessment validity and assessment driven learning. We propose three principal reasons for this. First, technological resources mediate and pervade an ever-increasing number of disciplinary practices: Computer scientists develop code in integrated development environments (IDEs), psychologists do statistics in R,

engineers design machine parts in CAD software, medical practitioners treat and diagnose patients with the aid of clinical decision support software, and virtually everyone writes texts with word processors. Second, learning sciences research indicates that successful, transferrable learning is associated with learners' active engagement with appropriate tools and learning resources (e.g. Danish & Gresalfi, 2018; Schwartz & Martin, 2004; (Hmelo-Silver, Kapur, & Hamstra, 2018). If successful learning is resource-mediated, then the resource-poorness of conventional examinations may explain issues with assessment driven learning: Resource-poor assessment may drive resource-poor learning. Third, cognition itself may be substantially resource-mediated.

Cognition

Established examination practice and its frameworks have been criticized for paying too little attention to the cognitive models in which they are grounded (Baird et al., 2017), and/or for being based on impoverished, outdated, or unsuitable models of cognition (Pellegrino, 2002; Sawyer, 2014). For the purpose of this study, we compare a rigid interpretation of two contrasting perspectives of cognition: Cognition as *mindbased* processing versus cognition as *situated* action. The mindbased perspective corresponds to cognitivist (sic), computational, representational, information-processing, connectionist, or constructivist models of cognition and learning (Abrahamsen & Bechtel, 2012; Shapiro, 2011). The mind is the manifest locus of cognition and learning, and mediator of the relationship between stimuli and response. Fundamental to this perspective is the 'opening of the black box' by modelling processes and states within the mind. Two simultaneous and intertwined streams of processing in the mind/brain together constitute cognition: Directed feedforward sensory-to-motor, stimulus-response, input-output streams of processing in combination with recursive feedback loops within the mind/brain itself. The contrasting situated model on the other hand, does not separate cognitive processing ('mind') from action ('response') or social and physical task contexts ('stimuli'). Instead, it regards the dynamical interaction between the cognitive agent and those elements of the environment with which he/she situationally interacts as conjointly constitutive of cognition and learning: Cognition as *situated action* or as *emergent* upon loosely coupled processes in the agent-environment complex system (Clark, 2012; Danish & Gresalfi, 2018; Hutchins, 1995). Actions of the cognitive agent effect changes in the environment, which in turn feed back to the cognitive agent in the form of new/altered stimuli. The directed stimulus-response flow of processing of the mindbound perspective is closed into a single complex system of dynamical feedback loops, from the cognitive agent through the environment back unto the cognitive agent him/herself. Examples of situated perspectives include embodied, extended, and distributed cognition, sociocultural theory, or social constructivism. While these situated perspectives have led to a rich body of research on learning and effective learning interventions, there is a lack of corresponding research in assessment and educational measurement (Mislevy, 2018 is one exception).

Gibson (1977) introduced the term 'affordance' for "whatever it is about the environment that contributes to the kind of interaction that occurs [with the cognitive agent]". Accordingly, we define the term *resource-affordance* for 'whatever it is about the disciplinary (technological) resources with which the cognitive agent interacts, that contributes to the kind of disciplinary practice that occurs'. Resource-affordances constitute the loose coupling of processes between the cognitive agent and the task environment. They are fundamental to the situated perspective. Much like a skier's body is inseparably connected with his boots and skis in the practice of skiing – effectively forming a single functional unit – so too do a programmer's mind and a computer-based programming environment interact in an inseparable manner in the practice of programming. Just as attempting to assess someone's skills in skiing while denying them skis would be rather absurd, so it is absurd that we routinely assess students' competency in computer science while denying them access to computers. It follows that the valid assessment of competency in disciplinary practices directly depends on adequate access to practice-relevant resource-affordances in the assessment task environment. Hence, the situated perspective demands an examination practice that is equally resource-rich (RR) as are the disciplinary practices in which we intend to assess competency. In the mindbased perspective on the other hand, there is no need to model resource affordances because cognition is fully contained in the mind/brain. Writing a recursive algorithm on paper or in a programming environment are not fundamentally different cognitive tasks, but fundamentally similar. Resources do not contribute anything substantial to cognition or its assessment. On the contrary, they are a potential source of construct irrelevant variance. Hence, the mindbased perspective of cognition favors a resource-poor (Rp) assessment practice. Indeed, we argue that a main reason for conventional examination practices being resource-poor likely lies in the fact that most students, teachers, educators, and assessment specialists share a deeply mindbased conception of cognition.

Educational Technology

Over the past years, educational technology and corresponding e-learning practices, including computer-based assessments and examinations, have become increasingly widespread in higher education (Bennett, 2015; Crisp, Guàrdia, & Hillier, 2016; Halbherr, Reuter, Schneider, Schlienger, & Piendl, 2014). Computer-based assessment services frequently prioritize efficiency by focusing on auto-correction, computer-adaptive testing, or remote proctoring – all largely within a conventional Rp paradigm. However, there also exists a competing trend, emphasizing the potential for improvements in examination quality by enabling examination task environments that are more authentic, competence-oriented, aligned with corresponding practice – and/or RR (Crisp et al., 2016; Halbherr, Dittmann-

Domenichini, Piendl, & Schlienger, 2016). One example of such a learning and assessment environment is Code Expert.

Code Expert

Code Expert (Lehner, Avanthay, & Sichau, 2018) is a browser-based integrated development environment (IDE) and online learning environment developed at ETH Zurich. Code Expert facilitates open programming assignments for in-class or take-home exercises, as well as supervised examinations. Code Expert includes an auto-grader, which provides automatic and immediate feedback to students by compiling, running, and testing submitted code against predefined test cases. Furthermore, tutors can annotate or apply direct changes to students' attempts in order to provide additional, more personalized feedback. The Code Expert interface is illustrated in Figure 1. It consists of a file system pane, a code editor window, a terminal and output window for compiling and running the code, and a tutorial pane for instructions, task descriptions, and learning materials.

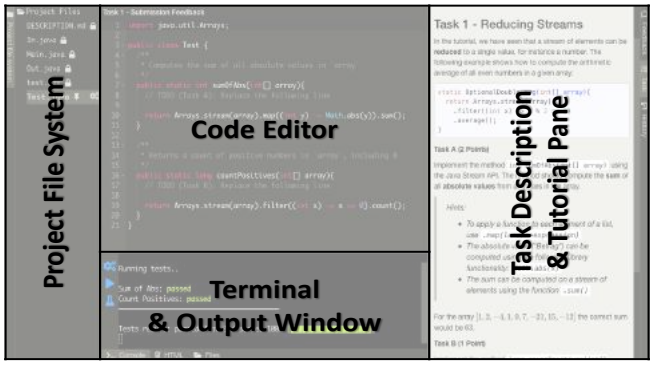


Figure 1: Schematic overview of Code Expert GUI.

Method

The study was conducted as part of an introductory course in programming for non-CS students at first year BSc level. The course focuses on imperative and object oriented programming paradigms, as well as problem solving, and uses Java as programming language and Code Expert as learning environment.

In the study, we investigated how the presence or absence of resource-affordances of the Code Expert environment affected student learning on the one hand, and the assessment of corresponding learning outcomes on the other. Slightly different than usual, the main focus of this study is not on the learning activity, but on the assessments, more precisely: The *validity* of the assessments. Specifically, we are interested whether and to what extent RR versus Rp assessments are sensitive to differences in learning outcomes as induced by RR versus Rp learning activities.

Validity

Validity is "the degree to which a test or examination measures what it purports to measure" (Ruch, 1924). It is a unitary construct (Messick, 1989). It is an ontological and/or epistemic construct, rather than a statistical or psychometric one (Kane, 2006). This holds particularly true in the context of this study, since we do not have any impartial source of base truth against which we could validate the RR versus Rp assessments. Instead, validity has to be determined through an appropriate validity argument. Borsboom, Mellenbergh, & van Heerden (2004) propose the following operational definition of validity: "A test is valid for measuring an attribute if and only if (a) the attribute exists and (b) variations in the attribute causally produce variations in the outcomes of the measurement procedure".

Operationalizing Validity

We apply the earlier propositions – resource mediation facilitates learning and cognition as a resource-mediated construct – to the above operational definition of validity. We operationalize variations in the measurement attribute – student learning – by letting one group of students learn with access to relevant resource-affordances, the RR learning condition (LC), while the other group is denied access, the Rp LC. Everything else is kept strictly identical across the two LCs. Subsequently, we assess half of students of each LC in a RR assessment with access to resource-affordances, the RR assessment condition (AC), or in a Rp assessment without access, the Rp AC. Again, everything else is kept strictly identical across the two ACs. We then evaluate whether the RR and/or the Rp assessment are able to differentiate between students from the RR LC versus the Rp LC. If they do, then this is evidence in favor of the assessment's validity, and evidence against, if not. This results in a 2x2 between-subjects experimental design. The RR LC versus Rp LC and the RR AC versus Rp AC constitute the independent variables, and assessment performances – to be more precise, the performance differences between the students in the RR LC and the Rp LC as measured either in the RR AC or the Rp AC – constitute the independent variables.

Operationalizing the Resource-Affordances

We identify the compiler as the key resource in the Code Expert environment. The compiler is both essential for generating the product of the practice – running code – as well as for sustaining the practices and processes required for achieving that goal. We thus operationalize the RR experimental conditions with a Code Expert environment with a fully functional compiler. The Rp conditions we operationalize with the *identical* Code Expert environment save for a deactivated compiler. This leads to the disappearance of the following resource-affordances: Students cannot compile or run code, correspondingly cannot receive any messages in the console from either the compiler or their compiled code, cannot run their code against test cases in the auto-grader, and there is no syntax highlighting of their code in the code editor. In the Rp condition, the students are essentially working with a 'naked text editor version' of Code Expert, while in the RR condition they have access to the fully functional Code Expert IDE. Across all

experimental conditions, students were instructed not to access any other resources (e.g. lecture notes, Google, StackOverflow, other Code Expert exercises) than those available through the study tasks in Code Expert.

Learning and Assessment Tasks

In the study's learning activity, we introduced a new paradigm: Functional programming. Java implements functional programming with the Stream API. The learning activity consisted of an interactive self-study tutorial. Key concepts of functional programming were introduced and consolidated in five consecutive tasks using hands-on exercises with the example of manipulating data-streams of numbers. Students received the canonical solution to each tutorial task at the start of the subsequent tutorial step. This ensures that also the students in the Rp LC received adequate feedback on the correctness of their solutions.

The subsequent assessment consisted of three tasks. In Task1 students had to perform identical manipulations of data-streams of numbers as in task 4 of the tutorial. Task1 operationalizes the direct replication of learning. Task2 introduced a novel and more complex problem that can be solved elegantly using the new paradigm. Task2 operationalizes transfer of learning. In Task3, students had to manipulate streams of Java objects instead of streams of numbers after reading a brief introduction to a number of new concepts and operations for manipulating objects in a functional manner. Task3 operationalizes transfer of learning with the aid of a learning resource, i.e. students' preparation for future learning (Schwartz & Martin, 2004). All three assessment tasks were scored manually by the course assistants. For each task 0, 1, 2, or 3 points were awarded according to task-specific rubrics. Small syntax errors, such as missing or unmatched brackets or slightly incorrect syntax in lambda expressions, were ignored. The manual scoring procedure was identical for both the RR AC and the Rp AC. To ensure a high correspondence with actual educational practice 'in the wild', both the learning activity and the assessment, all tasks contained therein, the scoring rubric, and the scoring procedure were all designed and performed entirely by the course lecturer and the course assistants, with no or only minimal intervention from the lead investigator.

Hypotheses from the Cognitive Perspectives

Let us now revisit the situated versus mindbased perspectives of cognition. What kind of results would each perspective predict for this experiment? To the situated perspective, the loose coupling of processes through resource affordances remains intact in the RR LC and the RR AC, while in the Rp conditions this coupling is severed. Hence, the RR and the Rp experimental conditions correspond to qualitatively fundamentally different kinds of cognitive processes – both regarding what is learned in the LCs, as well as regarding what is assessed in the ACs. Since programming is a resource-mediated practice, the situated perspective would predict larger learning gains in the RR LC, to which the RR AC is sensitive, but not the Rp AC (or only to a lesser extent). Furthermore, since cognition is emergent from the loosely-coupled agent-resource complex system, the larger learning gains of the RR LC and the higher sensitivity of the RR AC would not merely relate to 'superficial' resource-specific knowledge, but also deep conceptual understanding and transfer of learning. To the mindbased perspective on the other hand, resources are not central to cognition. Decoupling should not affect learning as long as students still receive adequate feedback. If anything, the Rp LC should facilitate learning, particularly learning for transfer, because it reduces cognitive load associated with managing the resource, freeing up cognitive capacity for focusing on developing a deep understanding of underlying concepts. Furthermore, the mindbased perspective would expect the Rp AC to be more sensitive to differences in learning gains, especially transfer of learning, because it eliminates construct-irrelevant variance related to managing the resource and resource-specific knowledge irrelevant to a deep understanding of underlying concepts.

Procedure

The study was conducted as part of regular in-class exercise activities of the first year BSc introductory programming course. The course took place across fourteen weeks, during fall semester 2018, from late September until late December. It entailed two weekly hours (i.e. 2x45 minutes) of lectures, two weekly hours of on-site exercises in small groups supervised by student teaching assistants (11 groups with between ca. 15-45 students each), weekly homework in Code Expert, and a final sixty minute summative examination in January 2019. The course is mandatory for first semester Bachelor students in Civil Engineering, in Geospatial Engineering, and in Environmental Engineering. The study activities took place in November 2018 during the second hour (45 minutes) of the on-site small group exercises of course weeks nine and ten, with one week between the learning activities and the assessments. We planned the study activities near the end of the course to ensure that all students were deeply familiar with the Code Expert environment, such that differences in assessment performance between the RR and Rp LCs could not reasonably be attributed to increased familiarity of students in the RR LC with surface features of the Code Expert environment. On the first study day, the students engaged in the learning activity consisting of the five-step tutorial on functional programming, either under RR (compiler active) or Rp (compiler deactivated) conditions. On the second study day, the students sat the assessment, again either under RR or Rp conditions. Time available for both the learning activity and the assessment was thirty minutes. While the students could progress through the five tutorial tasks at their own pace, in the assessment they had precisely ten minutes time available for each of the three tasks. In the week between day one and day two, there were no exercises or other activities related to the topics covered on day one. Figure 2 illustrates the experimental procedure: Group1 participated in the RR LC

and Rp AC; Group2 in the RR LC and RR AC; Group3 in the Rp LC and RR AC; and Group4 in the Rp LC and Rp AC.

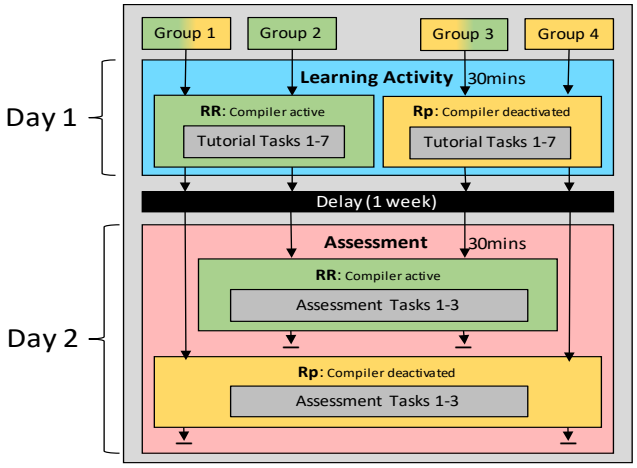


Figure 2: Illustration of the experimental procedure.

Sample

234 out of 272 enrolled students participated in the study. Of these, 21 students participated only on one of the two study days and had to be excluded from the analyses. For another 21 students we could not rule out for certain, that they had not accessed tasks or resources not intended for their experimental conditions, and were also excluded. The resulting sample of n=192 students is distributed across the experimental conditions as follows: n(RR→RR)=49, n(Rp→RR)=48, n(RR→Rp)=53, n(Rp→Rp)=42.

Results

Table 1 reports the mean percentage of points achieved in the complete test consisting of Task1, Task2, and Task3; in the replication task Task1; and in the transfer tasks, Task2 and Task3 taken together. Three things are worth note. First, students in the RR LC outperform students in the Rp one both in the RR and in the Rp assessment and in all tasks. Second, the RR assessment is more difficult (i.e. students performed worse) than the Rp assessment for students of both the Rp *and* the RR LC. Third, the performance difference in the transfer tasks between the two LCs is larger for the Rp assessment, with a 27% difference (63% - 37%) compared to a 14% difference (36% - 22%) in the RR assessment.

Table 1: Mean percentage of points achieved

LC →AC	RR →RR	Rp →RR	RR →Rp	Rp →Rp
Complete Test	46%	31%	70%	47%
Task1	67%	50%	83%	67%
Transfer Tasks	36%	22%	63%	37%

Figure 3 illustrates the assessment results in the complete test, the direct replication task, Task1, and the transfer tasks, Task2 and Task3 together. The vertical histograms illustrate the frequencies (x-axis) of total points achieved (y-axis) for each of the four experimental groups. The background and bar colors represent the LCs and ACs, respectively, green for RR and yellow for Rp. To illustrate appropriate interpretation of the histograms: 67% of students in the RR→Rp experimental condition achieved the maximum of three points in Task1. Furthermore, non-parametric Mann-Whitney *U* inferential statistics, corresponding p-values, effect sizes *r*, and mean ranks (lower values correspond to better performances) are reported for the comparisons between the RR LC and the Rp LC as measured in the RR AC and the Rp AC, respectively. Example: The comparison between the RR and Rp LC as measured by the complete test in the RR AC is highly significant with p=.008, *U*= 1'539, effect size *r*=.27 and better performance of the students from the RR LC (mean rank 41.43 < mean rank 56.42).

All test and subtest comparisons between the RR and Rp LC are statistically significant. The reported effect sizes *r* constitute small to medium effects (Field, 2009). Effect sizes are consistently larger for the Rp test than for the RR test, are consistently larger for the transfer tasks than the direct replication task, and the difference in effect size between RR and Rp assessment is larger for the transfer tasks.

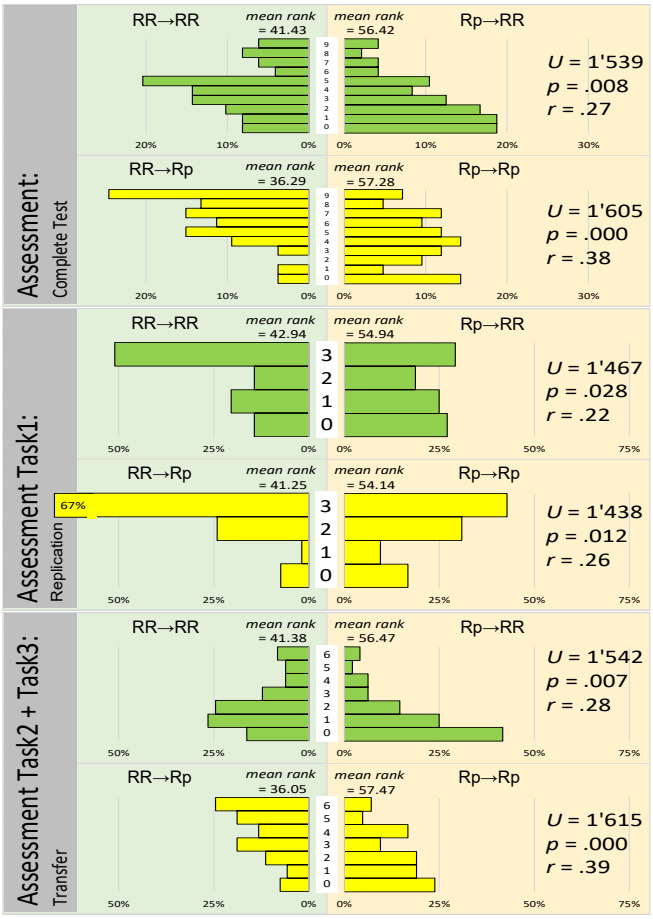


Figure 3: Assessment performances and inferential statistics.

Discussion

Learning

The findings convincingly confirm the proposition that resource mediation facilitates learning. RR learning consistently outperformed Rp learning. Effect sizes were small to medium for the direct replication task and medium for the transfer tasks. Of particular note, the effect is consistent and robust even across RR and Rp ACs, and even after only thirty minutes of tutorial-guided learning. The fact that this effect was stronger in the transfer tasks, and not only in the RR AC but even more so in the Rp AC, is strong evidence that resource-mediation facilitates not just learning of superficial resource-specific details, but in fact deep conceptual understanding and successful learning for transfer. Furthermore, our results support the assumption that it is indeed the presence or absence of practice-relevant resource-affordances that mediated these differences in learning. First, because the only difference between the otherwise identical LCs was whether the compiler was active or not, second, because not only the students in the RR LC, but also the students in the Rp LC received feedback on the correctness of their solution.

Assessment Validity

Both the RR and the Rp assessments successfully differentiate between the two LCs and are thus sensitive to the experimental manipulation of resource-mediated learning. Hence, we cannot reject the validity of neither the RR nor the Rp assessment. However, the Rp assessment was more sensitive to the experimental manipulation than the RR one, and especially in the transfer tasks. We identify two possible reasons for this. First, the higher sensitivity of the Rp assessment could be an indicator of better validity of the Rp assessment in general. Alternatively, the higher sensitivity could be an indicator of superior *differential* validity of the Rp assessment for assessing the transfer of (resource-mediated) learning in specific, but not necessarily for learning outcomes as they relate to the disciplinary practice at large. Two observations support this second interpretation. First, the RR assessment was consistently more difficult than the Rp assessment. It clearly required students to demonstrate competencies that go beyond what would have been sufficient to succeed in the Rp assessment. Second, the RR assessment is more directly representative of the target disciplinary practice of programming (which also includes a functional compiler), i.e. it is more 'ecologically valid'. If we assume that disciplinary competencies in all their complexity usually constitute the intended measurement constructs of examinations, then – somewhat paradoxically – the higher sensitivity observed in this study would imply that the Rp assessment suffers from construct underrepresentation in relation to the ecologically valid intended measurement construct. To further illustrate this argument: If the RR and Rp assessments captured the exact same amounts of variance in *transfer of resource-mediated learning*, but the RR assessment in addition also captured *other* variance relevant to competency in the disciplinary practice, then we would indeed expect precisely the observed pattern of higher sensitivity of the Rp assessment to the experimental manipulation. Taken together, this indicates that RR tasks may render more valid estimates of students' effective competency in a target practice, while Rp tasks may be more valid for the differential assessment of associated (developing) conceptual understanding.

Cognition

Neither the proposed situated nor mindbased perspective facilitated the prediction of the study's results. The mindbased perspective proved rather unsuitable for explaining the substantial and robust learning facilitation in the RR LC, while the situated perspective does not offer a meaningful account for the higher sensitivity of the Rp assessment. Regarding the ontological question of the nature of cognition, it is indeed quite intriguing that the uncoupled 'mindbound' Rp assessment was *more* sensitive to transfer of resource-mediated 'situated' learning, than the RR assessment. This pattern in many ways appears reminiscent of learning as (resource) internalization in a Vygotskian or Piagetian sense. Alternatively, from a complex systems perspective we might conclude that the mindbased perspective does not adequately account for cognitive phenomena emergent from agent-resource interaction, while the situated perspective does not adequately account for near decomposability. Such considerations notwithstanding, our data show that there is something more complex going on than can be explained with either a rigidly mindbound or rigidly situated perspective alone. This approach did not lead to parsimony, but instead to poor predictions.

Implications for Practice

We identify three main implications for practice. First, the RR tasks were more difficult than the Rp ones. When moving from Rp assessments to RR ones, this increase in difficulty needs to be accounted for. We can confirm this experimental finding from our own experience in supporting lecturers when transitioning from conventional Rp paper-based examinations to RR computer-based ones – e.g. with Code Expert. The new RR examinations usually require substantially more time for students to be able to solve them meaningfully. Second, we found robust evidence confirming RR learning activities facilitate deep conceptual learning and successful learning for transfer. If assessment drives learning, then we are well advised to include at least some RR tasks in any examination, providing students an effective incentive to engage in according and productive RR learning activities. Third, mixed examinations consisting of both RR and more conventional Rp tasks may be best, because Rp tasks may be more suitable for the differential assessment of developing conceptual understanding, while RR tasks may be more suitable for 'ecologically valid' and exhaustive assessments of accomplished disciplinary competency.

References

Abrahamsen, A., & Bechtel, W. (2012). History and core themes. In *The Cambridge Handbook of Cognitive Science* (pp. 9–28). Cambridge, UK: Cambridge University Press.

Baird, J.-A., Andrich, D., Hopfenbeck, T. N., & Stobart, G. (2017). Assessment and learning: fields apart? *Assessment in Education: Principles, Policy & Practice*, *24*(3), 317–350.

Bennett, R. E. (2015). The Changing Nature of Educational Assessment. *Review of Research in Education*, *39*(1), 370–407.

Biggs, J. (2014). Constructive alignment in university teaching. *HERDSA Review of Higher Education*, *1*(1), 5–22.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The Concept of Validity. *Psychological Review*, *111*(4), 1061–1071.

Clark, A. (2012). Embodied, embedded, and extended cognition. In *The Cambridge Handbook of Cognitive Science* (pp. 275–291). Cambridge: Cambridge University Press.

Crisp, G., Guàrdia, L., & Hillier, M. (2016). Using e-Assessment to enhance student learning and evidence learning outcomes. *International Journal of Educational Technology in Higher Education*, *13*(1).

Danish, J. A., & Gresalfi, M. (2018). Cognitive and Sociocultural Perspectives on Learning: Tensions and Synergy in the Learning Sciences. In *International Handbook of the Learning Sciences* (pp. 34–43). New York, NY: Routledge.

Field, A. P. (2009). *Discovering statistics using SPSS: and sex, drugs and rock 'n' roll* (3rd ed). Los Angeles: SAGE Publications.

Gibson, J. J. (1977). The theory of affordances. In *Perceiving, acting, and knowing: Toward an ecological psychology* (pp. 67–82). Hillsdale, NJ: Erlbaum.

Gulikers, J. T. M., Bastiaens, T. J., & Kirschner, P. A. (2004). A five-dimensional framework for authentic assessment. *Educational Technology Research and Development*, *52*(3), 67–86.

Halbherr, T., Dittmann-Domenichini, N., Piendl, T., & Schlienger, C. (2016). Authentische, kompetenzorientierte Online-Prüfungen an der ETH Zürich. *Zeitschrift für Hochschulentwicklung*, *11*(2), 247–269.

Halbherr, T., Reuter, K., Schneider, D., Schlienger, C., & Piendl, T. (2014). Making Examinations more Valid, Meaningful, and Motivating: The Online Exams Service at ETH Zurich. *EUNIS Journal of Higher Education.*, *1*(1).

Hmelo-Silver, C. E., Kapur, M., & Hamstra, M. (2018). Learning Through Problem Solving. In *International Handbook of the Learning Sciences* (pp. 210–220). New York, NY: Routledge.

Hutchins, E. (1995). How a Cockpit Remembers Its Speeds. *Cognitive Science*, *19*(3), 265–288.

Kane, M. T. (2006). Validation. In *Educational Measurement* (pp. 17–64). Washington, D.C.: American Council on Education.

Keehner, M., Gorin, J. S., Feng, G., & Katz, I. R. (2017). Developing and Validating Cognitive Models in Assessment. In *The Handbook of Cognition and Assessment: Frameworks, Methodologies, and Applications*. John Wiley & Sons.

Lehner, H., Avanthay, D., & Sichau, D. (2018). *Code Expert*. Retrieved from https://code-expert.net/

Messick, S. (1989). Validity. In *Educational Measurement* (pp. 13–100). Washington, D.C.: American Council on Education.

Mislevy, R. J. (2018). A Sociocognitive Perspective. In *Sociocognitive Foundations of Educational Measurement* (pp. 21–45). New York, NY: Routledge.

Moss, P. A. (1992). Shifting Conceptions of Validity in Educational Measurement: Implications for Performance Assessment. *Review of Educational Research*, *62*(3), 229–258.

Pellegrino, J. (2002). Knowing What Students Know. *Issues in Science and Technology*, *19*(2), 48–52.

Ruch, G. M. (1924). *The improvement of the written examination*. Oxford, England: Scott, Foresman & Co.

Sambell, K., McDowell, L., & Brown, S. (1997). "But is it fair?": An exploratory study of student perceptions of the consequential validity of assessment. *Studies in Educational Evaluation*, *23*(4), 349–371.

Sawyer, R. K. (2014). The New Science of Learning. In *The Cambridge Handbook of the Learning Sciences*. New York: Cambridge University Press.

Schwartz, D. L., & Martin, T. (2004). Inventing to Prepare for Future Learning: The Hidden Efficiency of Encouraging Original Student Production in Statistics Instruction. *Cognition and Instruction*, *22*(2), 129–184.

Shapiro, L. (2011). Standard Cognitive Science. In *New Problems of Philosophy*. *Embodied Cognition* (pp. 7–27). New York: Routledge.

Investigating sound and structure in concert: A pupillometry study of relative clause attachment

Jesse A. Harris (jharris@humnet.ucla.edu)
UCLA Department of Linguistics, 3125 Campbell Hall
Los Angeles, CA 90095 USA

Alexandra Lawn (alawn@ucla.edu)
UCLA Department of Spanish and Portuguese, 5310 Rolfe Hall
Los Angeles, CA 90095 USA

Marju Kaps (mkaps@ucla.edu)
UCLA Department of Linguistics, 3125 Campbell Hall
Los Angeles, CA 90095 USA

Abstract

Listeners must integrate multiple sources of information to construct an interpretation of a sentence. We concentrate here on the alignment of prosodic and syntactic grouping during online sentence comprehension. We present the results from a pupillometry study on the use of prosodic boundaries in resolving well-known attachment ambiguities. Using growth curve analyses to capture the non-linear dynamics of pupil dilation, we found increased pupil excursions for sentences that were disambiguated towards the dispreferred, non-local relation, especially when accompanied by supporting prosodic information. However, when prosodic and structural information did not align, pupillary response was muted, potentially indicating a failure to commit to a specific interpretation. More generally, the study shows how the currently under-utilized pupillometry method offers insights into spoken language comprehension.

Keywords: Relative clause attachment; prosody; pupillometry

Introduction

Sentences like (1) are structurally ambiguous. The relative clause (RC) after a complex noun phrase (NP) may be interpreted as modifying the first noun (NP1; in *High attachment*: the maid was on the balcony) or the second noun (NP2; in *Low attachment*: the actress was on the balcony).

(1) Someone shot the maid $_{\text{NP1}}$ of the actress $_{\text{NP2}}$ [$_{\text{RC}}$ who was on the balcony]

Many classical theories of sentence processing assume a strong role for relations privileging locality (Kimball, 1973; Frazier & Fodor, 1978) or recency (Gibson, 1998). For example, the principle of *Late closure* prompts the language parser to resolve ambiguous strings like (1) towards a structure in which the RC attaches to the most recently accessed constituent that is grammatically possible, resulting in a Low attachment interpretation. Under their strongest interpretation, locality or recency based principles represent universal properties of the human language processing system and are not subject to variation within languages or individuals.

However, subsequent research has argued that the preferred interpretations of strings like (1) are moderated by a great many other factors, such as plausibility, experience, or prosodic grouping, and may even reflect parsing preferences from specific languages (see Fernández, 2003 for review). Several explanations for this variation have been proposed. Perhaps the first study to find a High attachment preference for non-local RC modification was Cuetos & Mitchell (1988), who proposed that the statistics of a language strongly influence how the processor resolves ambiguity – i.e., a language shows a High attachment preference because this resolution is the most frequent in the language. Others have explained the variation in terms of additional contraints competing with a universal recency bias (Gibson et al., 1996), or with respect to the availability of alternative parses in different languages (Hemforth et al., 2000; Grillo & Costa, 2014; Grillo et al., 2015). Others still have exempted relative clauses from the domain of Late closure, instead arguing that they are construed using a collection of grammatical and non-grammatical interpretive principles (Gilboy et al., 1995; Frazier & Clifton, 1996). While many factors may well contribute to relative clause attachment preferences, we focus here on the relation between prosodic and syntactic grouping during online sentence comprehension.

In general, prosody refers to the organizational structure of speech, expressed through changes in pitch, duration, and amplitude, among other factors (Ladd, 2008). Although the prosodic grouping of words can reflect metrical preferences (such as the location of prominence or the alternation of feet), it can also signal higher levels of linguistic structure, especially syntactic or information structural relations. We adopt a simplified description of prosodic information, and concentrate on how prosodic groups are formed via the presence of a prosodic boundary (**%**), which can be indicated by word final lengthening, pitch movements that conventionally mark the end of a group, and pauses in the speech signal.

Previous work indicates that the location of a prosodic boundary between the complex NP and the relative clause disambiguates RC attachment. A prosodic boundary placed between NP2 and the RC (2b) results in a bias towards High attachment, a generalization that appears to be robust across languages (Jun, 2003). In contrast, a boundary separating NP1 from NP2 (2a) appears to reinforce the low attachment construal, even in languages with a general high attachment preference (Fromont et al., 2017). We assume that in such cases comprehenders interpret the prosodic boundary loca-

tion as a cue to syntactic constituency (depicted as parentheses below) in an attempt to align prosodic and syntactic junctures.

(2) Someone shot ...

a. (the maid) **%** (of the actress who was on the balcony)

b. (the maid of the actress) **%** (who was on the balcony)

As most studies on attachment ambiguity in complex NPs are conducted using offline measures (though see, for instance, Kim & Christianson, 2013 and Fernández & Sekerina, 2015), it is unclear whether prosodic boundaries guide online sentence processing, and if so whether each boundary location is used immediately. We report the results of a pupillometry study addressing the role of prosodic boundaries on attachment ambiguity resolution in real time comprehension. The experiment was designed to explore two distinct possibilities. First, it is possible that non-local dependencies are computationally taxing to process, regardless of prosodic boundary information. In this case, we would predict that (i) sentences grammatically disambiguated to a high attachment construal would elicit processing costs, which (ii) would not be mitigated by corroborating prosodic information. As an alternative, it is possible that non-local dependencies are avoided for independent prosodic reasons, e.g., if a boundary after NP1 is preferred to keep the prosodic units of equal weight (Fodor, 1998). Such a view would predict that high attachment resolutions are only costly when not supported by prosodic boundary information.

Pupillometry

Pupil dilation is likely to reflect a multitude of components, some related to demands on cognition and attention, and others to the autonomic nervous system. The pupil naturally fluctuates, dilating in response not only to neural inhibition from the parasympathetic oculoumotor system and the the noradrenergic system, but also to the presentation of an external stimulus (Wilhelm et al., 1999), including emotionally arousing stimuli, challenging math problems, and unconscious or barely discernable stimuli (Beatty & Kahneman, 1966; Kahneman & Beatty, 1966; Hess & Polt, 1960; Laeng et al., 2012).

Although the tools and techniques for pupillometry are still developing, existing literature has already provided convincing evidence that pupil dilation indirectly indexes increased demands on mental effort and attention associated with difficult to interpret material along various linguistic dimensions (Schmidtke, 2018 for review). While some early studies found a relation between increased pupil dilation and syntactic complexity (Just & Carpenter, 1993), more recent studies have begun to explore a wider range of ways in which pupil size might reflect other factors in language comprehension. Major findings include an association between increased pupil size and structurally complex sentences (Demberg & Sayeed, 2016), prosodic disambiguation in garden path sentences (Engelhardt et al., 2010), semantic anomalies (Demberg & Sayeed, 2016), lexical frequency or increased emotional valence (Kuchinke et al., 2007), violations of expected meter (Scheepers et al., 2013), and inadequate or misleading pitch accent (Zellin et al., 2011). In keeping with the current pupillometry literature, we will assume, as a basic linking hypothesis between cognition and behavior, that increased pupil size indirectly reflects increased cognitive load, including mental effort directed at managing language comprehension processes.

Further, pupillometry offers an appealing supplement to other online methods. It is easy to administer and comparatively inexpensive, while offering an online and passive measurement. As pupil size is not under conscious control, pupil dilation measurements are likely to be resilient to task-specific strategies that subjects may learn or employ during the experiment. Thus, pupillometry studies offer a promising avenue for exploring the role of prosodic information in resolving structural ambiguity.

Pupillometry study

Participants

Forty-eight native college-aged speakers of American English were recruited from a Psychology Subject Pool at the University of California, Los Angeles, and were compensated with course credit. No participant reported any history of hearing loss or language disorder. Experimental sessions typically lasted no more than 30 minutes.

Materials and method

Twenty quartets were constructed in a 2×2 design. Quartets crossed Boundary location (post-NP1, post-NP2) and Attachment (High, Low). All sentences involved a complex noun phrase (*the brother of the musicians*) containing two noun phrases (NP1, NP2) of opposite grammatical number, followed by a relative clause disambiguated to high (modifying NP1: *the brother*) or low (modifying NP2: *the musicians*) attachment. The attachment height was grammatically specified by the plurality of the verb (*was, were*) within the relative clause (*who was really quiet*), which was kept constant within each quartet. Half of the items were disambiguated by the singular form of the auxiliary (*was*), half by the plural form (*were*). A sample item is shown in Table 1. In addition, comprehension questions were presented after half of the items to encourage participants to pay attention. Questions did not ask about relative clause attachment height.

We obtained measures for the lexical level characteristics of length, frequency and number of syllables using the English Lexicon Project (Balota et al., 2007) for NP1 and NP2. Nouns did not differ on length, number of syllables, or frequency, as determined by several measures, including (log) HAL (Lund & Burgess, 1996) and (log) $\text{SUBTLEX}_{\text{US}}$ (Brysbaert & New, 2009).

Sentences were recorded with a high-quality microphone in a sound attenuated chamber by a trained phonologist. Audio files were truncated after the relative clause (marked by

Boundary	High attachment	Low attachment
NP1	Everybody met the brother **%** of the musicians who was really quiet // although the club was really noisy.	Everybody met the brothers **%** of the musician who was really quiet // although the club was really noisy.
NP2	Everybody met the brother of the musicians **%** who was really quiet // although the club was really noisy.	Everybody met the brothers of the musician **%** who was really quiet // although the club was really noisy.

Table 1: Sample quartet item crossing Boundary location (NP1, NP2) and Attachment (High, Low). The underlining identifies the noun modified by the relative clause (*who was really quiet*). The prosodic boundary is indicated by the **%** symbol.

the // symbol in Table 2). 100ms of computer generated silence was inserted after the relative clause, and the post-relative clause segment was spliced into the recording, so that pupillary response was recorded on acoustically identical material within items. Items were then acoustically normalized to make the transition into the spliced segment as seamless as possible.

The 20 experimental item quartets were presented in counterbalanced and individually randomized order, and were interspersed with 40 items from two separate experiments (one also manipulating boundary placement, and another manipulating contrastive accent), along with 26 filler items unrelated to any experiment. Participants were instructed to fixate on a cross at the center of the screen without blinking for the duration of the sentence. They were encouraged to blink before and after the sentence presentation, and to rest their eyes as needed, in order to minimize the possibility of eye blinks due to fatigue.

Items were presented with Experiment Builder (SR Research) to subjects over sound-isolating headphones in a moderately lit room dedicated to experimentation. Eye position and pupil area were recorded using an SR Research EyeLink 1000 Plus eye tracker sampling at 500Hz. The tracker was mounted to the table at 55cm from a 27 inch LCD monitor with a light gray background. A 5-point calibration procedure was used before recording and as necessary, and drift correction was conducted between every trial.

Results and discussion

Subjects performed very well on post-sentence comprehension questions ($M = 96\%$), which did not probe the interpretation of the relative clause. There was a marginal effect of mismatches between Boundary location and Attachment, so that comprehension questions following High attachment sentences were less accurate when paired with an NP2 boundary location (*diff* = 3%), and questions after Low attachment sentences were less accurate when paired with an NP1 boundary (*diff* = 3%), $t = 1.96, p = 0.05$. No other effects were observed. The pattern suggests that incompatible boundaries interfered with comprehension, but that there was no general effect of boundary or attachment on general comprehension. However, performance on all conditions was uniformly high, averaging at 94% or above.

Pre-processing of gaze location and pupil size was conducted in Data Viewer. Fixations outside of the fixation cross were removed. Trials with eye blinks (less than 5% of the total data) were also removed to avoid reconstructing pupil size during noisy trials. Mean pupil size was downsampled into 100 20ms bins starting from the end of the relative clause, for a total recording time of 2000ms past the relative clause. The remaining analyses were conducted in R (R Core Team, 2016). Data were fit with a growth curve model (Mirman, 2016) to avoid assuming a linear form or an arbitrary time window for analysis. Growth curve models have been used previously to quantify continuous changes in pupillary response, and we adopt those authors' interpretations of the curve with respect to pupil response (Kuchinsky et al., 2013; McGarrigle et al., 2017).

We report a third-order (cubic) orthogonal polynomial model with fixed effects of Boundary, Attachment, and their interaction on polynomial terms, and by-subject and by-item random slopes (Baayen et al., 2008), as shown in Table 2. Experimental predictor variables received deviation (sum) coding with NP1 and Low Attachment conditions specified as reference levels for their respective factors. Orthogonal polynomials were used to avoid extreme multicollinearity between adjacent samples in the time series.

	Estimate	**Std. Error**	***t*-value**
(Intercept)	0.106	0.315	0.337*
Linear poly	-0.857	0.237	-3.616*
Quadratic poly	-1.799	0.246	-7.311*
Cubic poly	0.966	0.249	3.887*
NP2	0.241	0.023	10.485*
High	0.201	0.023	8.761*
Linear:NP2	1.152	0.237	4.858*
Quadratic:NP2	-0.2	0.246	-0.814
Cubic:NP2	0.191	0.249	0.769
Linear:High	0.774	0.237	3.263*
Quadratic:High	-0.363	0.246	-1.473
Cubic:High	-0.312	0.249	-1.255
NP2:High	0.001	0.023	0.065
Linear:NP2:High	-0.578	0.237	-2.438*
Quad:NP2:High	0.342	0.246	1.390
Cubic:NP2:High	0.882	0.249	3.547*

Table 2: Growth curve analysis in a linear mixed effects regression model. The * indicates a significant effect at the $\alpha = .05$ threshold.

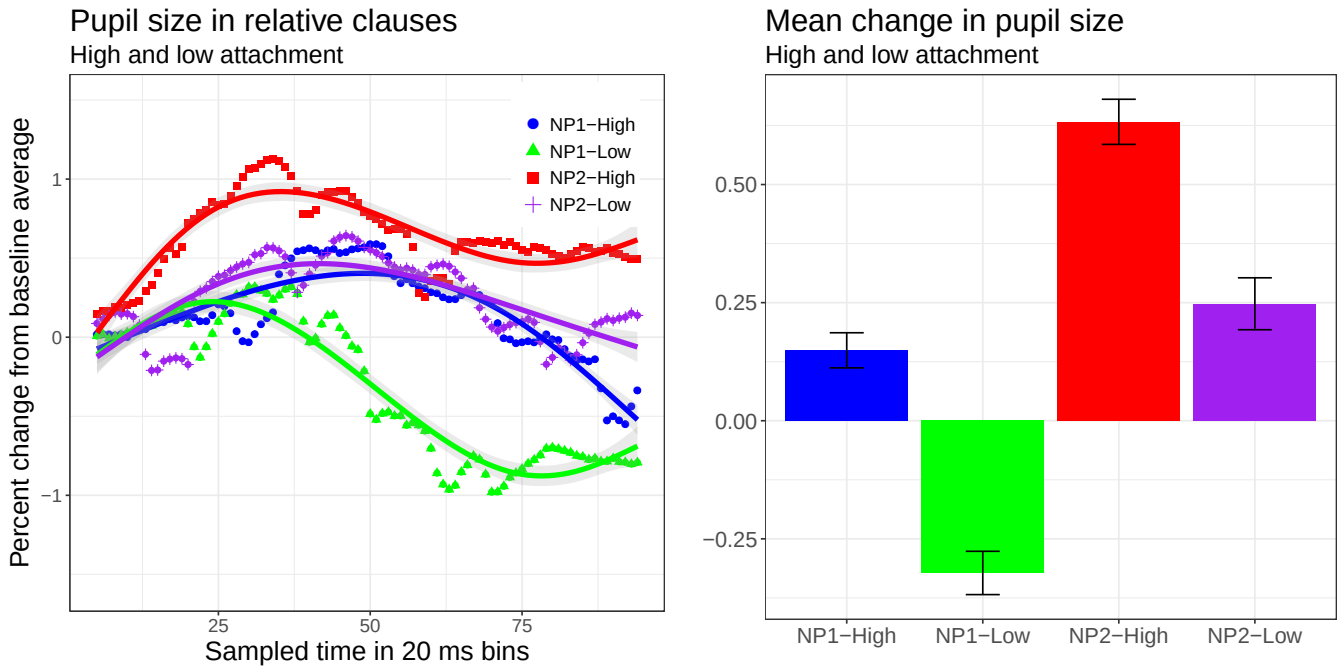


Figure 1: Percent change of pupil dilation over time with standard errors (left panel) and mean percent change of pupil size collapsing across time bins (right panel).

In growth curve analyses, polynomial terms capture distinct components of the functional form of a curve as it develops over time, and will be interpreted with respect to pupil dilation as follows (Kuchinsky et al., 2013; McGarrigle et al., 2017). The INTERCEPT corresponds to the overall mean pupillary response, so that positive coefficients indicate greater amplitudes. The LINEAR polynomial term coefficient corresponds to the slope of pupillary response, so that a positive increase in the coefficient indicates more steeply rising pupil dilation. The QUADRATIC polynomial term describes the shape of the primary inflection point, revealing the degree to which the curve is non-linear. Negative quadratic coefficients indicate an inverted U-shaped curve, characteristic of pupil peaks. The CUBIC term captures the properties of any secondary inflection point in the curve, so that positive coefficients indicate that pupil dilation peaks are more compressed or transient, rising and falling more sharply.

Positive coefficients of High attachment and NP2 boundary indicate increased average pupil dilation (the area under the curve) for High attachment over Low attachment, and NP2 boundary location over NP1 boundary location, respectively. Interactions between the planned predictor variables and the polynomial terms indicate how the experimental conditions differentially affect the shape of the pupillary response over time.

The mean change in pupil size for each condition is shown in Figure 1. In the left panel, the shape of the best-fitting non-linear regression line is plotted against change in pupil size within a 2000ms period immediately after the relative clause. The values on the vertical axis represent the percent change from the baseline average, defined here as the entire 100ms segment of silence inserted between the end of the relative clause and the remainder of the sentence (*although the club was really noisy*). The right panel reports the overall mean pupil change (with standard errors in grey) in pupil size during the same period for visual comparison.

In the growth curve model, effects of all three orthogonal polynomial terms were observed. Modulo the manipulation, pupil growth was less steep (a negative Linear coefficient), showed greater inverted U-shaped curve (a negative Quadratic coefficient), and a sharper secondary point of inflection (a positive Cubic coefficient), corresponding to a change or bend in the direction of the response.

More importantly, the two experimental factors in the manipulation showed that a prosodic boundary after NP2, and relative clauses that were grammatically disambiguated to a High attachment relation elicited greater pupil excursions from the grand mean compared to their respective NP1 and Low attachment reference levels. Both NP2 boundary and High attachment conditions also elicited more steeply rising slopes, as indicated by their interaction with the linear orthogonal polynomial.

In addition, the interaction between NP2 boundary and High attachment conditions further interacted with linear and cubic polynomials, indicating that the NP2-High condition elicited a smaller slope and increased transience of the pupil peak. The overall interaction is perhaps best visualized in the right panel of Figure 1, where the effect of NP2 is greater for High attachment conditions.

Perhaps more intuitively, the plot in the left panel of Fig-

ure 1 suggests the following conclusions. First, the conditions where the syntax and the prosody aligned largely conform to expectations. The theoretical baseline condition (NP1-Low) elicited the least extreme growth in pupil size, whereas the NP2-High elicited the most extreme changes. Low attachment is thought to be compatible with a boundary after NP1. Low attachment instantiates the empirically preferred relation between a RC and a complex nominal head in English, and, by hypothesis, is the least taxing relation to compute. The fact that the response was relatively muted in this condition is therefore entirely compatible with current linguistic theory. Similarly for the NP2-High condition: a boundary after NP2 aligns with the proposed syntax of High attachment structures. Assuming that non-local relations, including High attachment, are costly to compute, the fact that the NP2-High condition elicited the most extreme response is also consistent with current theory. However, our findings do not support the possibility that the bias against High attachment could be solely attributed to lack of supporting prosodic information; even when structures were disambiguated by prosodic boundary location, High attachment structures elicited increased cognitive load.

Second, the conditions where the prosodic grouping did not align with the syntactic constituency reveal a more complicated pattern. We predicted that the NP1-High attachment condition would be more anomalous than the NP2-Low attachment condition. Our reasoning was that the prosody of NP1-High would encourage grouping the relative clause and NP2 together, e.g. *the musicians who was really quiet*, creating a local number mismatch violation between NP2 (*the musicians*) and the verbal agreement marker (*was*). In contrast, the NP2-Low condition (*the musician % who was really quiet*) is locally grammatically coherent despite an infelicitous prosodic boundary. However, the two mismatching conditions elicited similar response patterns.

We entertained three main possible explanations. The first was that, in cases of mismatching cues, the processor makes weaker online processing commitments, and may defer the attachment decision until later, if it commits at all, as in models employing syntactic underspecification for attachment ambiguities (e.g., Frazier & Clifton, 1996). This interpretation is broadly compatible with results from Johnson et al.'s (2014) digit span task, which found decreased pupil size in response to digit sequences exceeding normal capacity. Decreased pupillary response may also indicate that the subject has abandoned an excessively difficult task, highlighting the role of attention in relating pupil size growth to cognitive effort (see also Beatty, 1982 and Winn et al., 2018)

A second possibility was that systematic differences between items may have prompted different processing strategies. For example, half of the items were disambiguated with the singular auxiliary marker (*was*), half were disambiguated with the plural marker (*were*). Our intuition was that relative clause attachment relations that were disambiguated with a singular form (*Everybody met the brother % of the musicians who was really quiet*) would be less anomalous than cases with plural disagreement (*I got a call from the friends % of the lawyer who were in politics*). We further addressed this possibility by including which number (singular vs. plural) was used to disambiguate the attachment location. Impressionistically, NP1-High conditions elicited greater pupil excursions in the Plural condition. However, grammatical number failed to differentiate effects within the statistical model.

A third explanation was that the processor resolves to High or Low attachment on the basis of another unidentified factor, such as by-item differences in boundary strength, prominence, or plausibility. To address these possibilities, we conducted a *post-hoc* boundary identification and rating norming study. The post RC material was removed, and the items were placed into four counterbalanced lists along with 46 filler items from the pupillometry experiment. Twenty additional participants from the same population as before were instructed to listen to each sentence over headphones in a noise-attenuated sound booth as many times as necessary, in order to manually mark prosodic boundaries on written versions of the sentence, and to rate how well the produced sentence matched its likely intended meaning (1 = completely unnatural, 7 = completely natural).

While participants were at ceiling (99.8%) at identifying the prosodic boundaries at the intended location after NP1 and NP2, additional boundaries after NP2 were perceived in post-NP1 boundary conditions 22% of the time. Subjects may have reverted to their default prosodic preference for an additional, potentially weaker, boundary before the RC (as discussed in Jun, 2003). Consistent with that interpretation, there was a main effect of prosody in ratings ($p<.01$), in which NP2 conditions (M=4.80, SE=0.12) were rated a more natural match with the intended meaning than NP1 conditions (M=3.66, SE=0.11). The penalty for non-local relations was evident in the ratings, as well. Sentences with Low attachment RCs (M=4.41, SE=0.12) were rated as more naturally matching with the prosody than sentences with High attachment RCs (M=4.05, SE=0.12), $p<.01$. The two factors did not significantly interact, suggesting that subjects were not sensitive to mismatches between syntactic and prosodic cues in this relatively conscious offline task.

General discussion

We used pupillometry to explore how prosodic and structural information interact during online language comprehension. Though relatively under-utilized in language processing research, pupillometry offers a promising methodological avenue for exploring how prosodic and structural information are integrated in real time processing. This method is especially useful for investigating how listeners use acoustic information to construct an interpretation, and offers a naturalistic and cost-effective complement to better known methods, such as neuroimaging or eye tracking in the visual world paradigm.

The results of the study replicate the well-studied bias for Low attachment of relative clauses in complex noun phrases

in English, a preference known to be modulated by prosodic boundary placement. However, our results cast doubt on an account which would attribute the preference to a lack of prosodic information alone. Even in the presence of overt prosodic boundaries, sentences with non-local dependencies were found to elicit online processing penalties. In addition, when the prosodic and grammatical information did not agree, pupillary response was reduced, indicating that prosody and structure are incorporated into an unfolding representation in concert. While more work is needed to investigate how language users integrate multiple sources of information together, the current study is compatible with the claim that comprehenders may avoid or delay making certain processing decisions in the face of inconsistent information.

Acknowledgments

Our thanks to Bethany Sturman for recording and preprocessing the materials, and to Joonhwa Kim, Nathan Mallipeddi, Alison Suh, Chenchen Wang, and Rebecca Wu for administering the study. Thanks also to Sun-Ah Jun for discussion of prosodic boundaries and to the UCLA Psycholinguistics Seminar for insightful comments.

References

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390–412.

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin*, *91*, 276-292.

Beatty, J., & Kahneman, D. (1966). Pupillary changes in two memory tasks. *Psychonomic Science*, *5*, 371–372.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*, 977–990.

Cuetos, F., & Mitchell, D. C. (1988). Cross-linguistic differences in parsing: Restrictions on the use of the Late Closure strategy in Spanish. *Cognition*, *30*, 73–105.

Demberg, V., & Sayeed, A. (2016). The frequency of rapid pupil dilations as a measure of linguistic processing difficulty. *PloS One*, *11*, e0146194.

Engelhardt, P. E., Ferreira, F., & Patsenko, E. G. (2010). Pupillometry reveals processing load during spoken language comprehension. *The Quarterly Journal of Experimental Psychology*, *63*, 639–645.

Fernández, E. M. (2003). *Bilingual sentence processing: Relative clause attachment in English and Spanish*. Amsterdam, The Netherlands: John Benjamins Publishing.

Fernández, E. M., & Sekerina, I. A. (2015). The interplay of visual and prosodic information in the attachment preferences of semantically shallow relative clauses. In L. Frazier & E. Gibson (Eds.), *Explicit and Implicit Prosody in Sentence Processing*. Springer.

Fodor, J. D. (1998). Learning to parse? *Journal of Psycholinguistic Research*, *27*, 285–319.

Frazier, L., & Clifton, C., Jr. (1996). *Construal*. Cambridge, MA: MIT Press.

Frazier, L., & Fodor, J. D. (1978). The sausage machine: A new two-stage parsing model. *Cognition*, *6*, 291–325.

Fromont, L. A., Soto-Faraco, S., & Biau, E. (2017). Searching high and low: Prosodic breaks disambiguate relative clauses. *Frontiers in Psychology*, *8*, 96.

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, *68*, 1–76.

Gibson, E., Pearlmutter, N., Canseco-Gonzalez, E., & Hickok, G. (1996). Recency preference in the human sentence processing mechanism. *Cognition*, *59*, 23–59.

Gilboy, E., Sopena, J.-M., Clifton, C., Jr., & Frazier, L. (1995). Argument structure and association preferences in Spanish and English complex NPs. *Cognition*, *54*, 131–167.

Grillo, N., Costa, J., Fernandes, B., & Santi, A. (2015). Highs and lows in English attachment. *Cognition*, *144*, 116–122.

Grillo, N., & Costa, J. a. (2014). A novel argument for the universality of parsing principles. *Cognition*, *133*, 156–187.

Hemforth, B., Konieczny, L., & Scheepers, C. (2000). Syntactic attachment and anaphor resolution: The two sides of relative clause attachment. In M. W. Crocker, M. J. Pickering, & C. Clifton Jr. (Eds.), *Architectures and Mechanisms for Language Processing*. Cambridge, UK: Cambridge University Press.

Hess, E. H., & Polt, J. M. (1960). Pupil size as related to interest value of visual stimuli. *Science*, *132*, 349–350.

Johnson, E. L., Miller Singley, A. T., Peckham, A. D., Johnson, S. L., & Bunge, S. A. (2014). Task-evoked pupillometry provides a window into the development of short-term memory capacity. *Frontiers in psychology*, *5*, 218.

Jun, S.-A. (2003). Prosodic phrasing and attachment preferences. *Journal of Psycholinguistic Research*, *32*, 219–249.

Just, M., & Carpenter, P. A. (1993). The intensity dimension of thought: pupillometric indices of sentence processing. *Canadian Journal of Experimental Psychology*, *47*, 310–339.

Kahneman, D., & Beatty, J. (1966). Pupil diameter and load on memory. *Science*, *154*, 1583–1585.

Kim, J. H., & Christianson, K. (2013). Sentence complexity and working memory effects in ambiguity resolution. *Journal of psycholinguistic research*, *42*, 393–411.

Kimball, J. (1973). Seven principles of surface structure parsing in natural language. *Cognition*, *2*, 15–47.

Kuchinke, L., Võ, M. L.-H., Hofmann, M., & Jacobs, A. M. (2007). Pupillary responses during lexical decisions vary with word frequency but not emotional valence. *International Journal of Psychophysiology*, *65*, 132–140.

Kuchinsky, S. E., Ahlstrom, J. B., Vaden Jr, K. I., Cute, S. L., Humes, L. E., Dubno, J. R., & Eckert, M. A. (2013). Pupil size varies with word listening and response selection difficulty in older adults with hearing loss. *Psychophysiology*, *50*, 23–34.

Ladd, D. R. (2008). *Intonational phonology* (2nd ed.). Cambridge: Cambridge University Press.

Laeng, B., Sirois, S., & Gredebäck, G. (2012). Pupillometry: A window to the preconscious? *Perspectives on Psychological Science*, *7*, 18–27.

Lund, K., & Burgess, C. (1996). Hyperspace analogue to language (HAL): A general model semantic representation. *Brain and Cognition*, *30*, 203–208.

McGarrigle, R., Dawes, P., Stewart, A. J., Kuchinsky, S. E., & Munro, K. J. (2017). Pupillometry reveals changes in physiological arousal during a sustained listening task. *Psychophysiology*, *54*, 193–203.

Mirman, D. (2016). *Growth curve analysis and visualization using R*. CRC Press.

R Core Team. (2016). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Scheepers, C., Mohr, S., Fischer, M. H., & Roberts, A. M. (2013). Listening to limericks: a pupillometry investigation of perceivers' expectancy. *PLoS One*, *8*, e74986.

Schmidtke, J. (2018). Pupillometry in linguistic research: An introduction and review for second language researchers. *Studies in Second Language Acquisition*, *40*, 529–549.

Wilhelm, B., Wilhelm, H., & Lüdtke, H. (1999). Pupillography: Principles and applications in basic and clinical research. *Pupillography: Principles, methods and applications*, 1–11.

Winn, M. B., Wendt, D., Koelewijn, T., & Kuchinsky, S. E. (2018). Best practices and advice for using pupillometry to measure listening effort: An introduction for those who want to get started. *Trends in Hearing*, *22*, 1–32.

Zellin, M., Pannekamp, A., Toepel, U., & van der Meer, E. (2011). In the eye of the listener: Pupil dilation elucidates discourse processing. *International Journal of Psychophysiology*, *81*, 133–141.

Sample stimuli from experiment

Six additional experimental items (from a total of 20 sentences). Low attachment disambiguation is presented prior to High attachment disambiguation. Disambiguation was evenly balanced across singular (*was*) and plural (*were*) auxiliary markers.

1. I got a call from the friends (%) of the lawyer (%) who was / were volunteering for the campaign // but my phone died halfway through the call.

2. We were all listening to the neighbor (%) of the pilots (%) who were / was raising exotic pets // even though we were in a hurry.

3. I spoke to the apprentices (%) to the librarian (%) who was / were wearing blue jeans // but I do not remember what we discussed.

4. Somebody saw the manager (%) of the architects (%) who were / was planning to buy a Mercedes // although it was very dark outside.

5. Someone kissed the sisters (%) of the medic (%) who was / were expecting to work late // yet nobody saw it happen.

6. Everybody admired the parent (%) of the artists (%) who were / was dancing the waltz // even though there was no music playing.

What are you talking about?: A Cognitive Task Analysis of how specificity in communication facilitates shared perspective in a confusing collaboration task

Yugo Hayashi (y-hayashi@acm.org)
College of Comprehensive Psychology, Ritsumeikan University
2-150 Iwakura-cho, Ibaraki, Osaka 567-8570 Japan

Kenneth R. Koedinger (koedinger@cmu.edu)
Human-Computer Interaction Institute, Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA 15213-3890 USA

Abstract

This study investigated how participant's specificity in sharing of information in collaborative problem solving was critical to them reaching a successful shared perspective. We analyzed participants' communication strategies in a collaborative task designed to make finding common ground challenging. We set out to better understand the difference between successful and unsuccessful collaborations by conducting a cognitive task analysis. From participants' utterances, we inferred cognitive processes associated with repeating communication moves and coded those processes as if-then production rules. We thereby specified the communication strategies used during interactions and developed a production-rule model to explain whether and how shared perspective developed or not. Our cognitive task analysis indicated that although all collaborating pairs described the objects they were seeing with a variety of features, the successful pairs were more specific in using combinations of features. Quantitatively, we found significant correlations between frequency of combined feature statements and success in sharing perspectives.

Keywords: Collaborative Problem Solving; Scientific Reasoning; Creativity; Coordination; Cognitive Task Analysis; Production Rules

Introduction

As discussed by cognitive scientists Herbert Simon and Allan Newell (Dasgupta, 2003), collaborative problem solving based on different perspectives helps generate new knowledge and scientific discoveries. Researchers in cognitive science have investigated the nature of collaborative problem solving (CPS), aiming to understand what kind of cognitive process underlie interactions (Okada & Simon, 1997; Salomon, 2001). Throughout these studies, it has been noted that CPS enables generation of meta-cognition, such as explanation activities (Chi, Leeuw, Chiu, & Lavancher, 1994), externalizing one's thoughts (Shirouzu, Miyake, & Masukawa, 2002), and receiving reflective responses from recipients of explanations (Miyake, 1986). Studies show that collaborating with partners with different types of knowledge and perspectives provides an opportunity to produce effective interactions (Greeno & de Sande, 2007). However, when conducting CPS research with individuals who hold different perspectives, it is important to consider constraints, such as interpersonal conflicts, which may occur due to the discrepancies among perspectives (Hayashi, 2018). Previous studies of dyads show that individuals work by role-sharing each other's different perspective (Hayashi, Miwa, & Morita, 2006). However, it is not fully understood what kind of communication processes underlie such activities, particularly regarding how dyads establish common ground by which to share their perspectives. To investigate this issue, this study reanalyzed data from Hayashi et al. (2006), by conducting cognitive-task analysis (Koedinger & Terao, 2002; Rittle-Johnson & Koedinger, 2001). We first review the CPS literatures, discuss the constraints on communication, and explain how common ground is achieved in our research paradigm (CPS based on different perspectives). We then state our specific goals and hypotheses.

CPS by taking different perspectives

Previous studies of scientific discovery in CPS showed scientists reason by taking different perspectives during interactions; this is termed distributed reasoning (Dunbar, 1995). Discussing different types of knowledge among individuals provides opportunities to generate conceptual changes (Roschelle, 1992), and is important for facilitating conceptual understanding (Greeno & de Sande, 2007). With this theoretical background, studies have shown that arguments and explanations within groups facilitate conceptual changes (Asterhan & Schwarz, 2009). Arguments made by group members by taking different perspectives are considered types of constructive and interactive joint collaborative activity (Chi, 2009); this is accomplished by coordinating individuals who hold different knowledge and perspectives. There exist group-based learning practices called jigsaw learning (Aronson & Patnoe, 1997), which focus on generating arguments by bringing together group members with different knowledge and asking them to discuss and integrate their knowledge. Throughout such studies, results show that cognitive bias and disagreements represent constraints on interaction; these factors should be considered when investigating CPS performance. Taking in these issues into consideration, Hayashi et al. (2006) conducted a laboratory based experiment using a simple reasoning task in which participants experienced difficulties on establishing common ground about each other's perspective. The results showed that when participants made substantive contributions to others by providing information by role sharing, they were able to generate broader perspectives by which to solve the problem. Moreover, successfully establishing coordination, such as correctly understanding others' perspectives, led to success in collaborations. Regarding the coordination process, recent stud-

ies of CPS have noted that collaborative problem-solving is composed of the following phases: (1) task work (problem solving), which builds internal knowledge, and (2) team work (coordination), during which internalized knowledge is exchanged and shared to build collective knowledge (Fiore, Rosen, Smith-Jentsch, Salas, & Letsky, 2010). However, it is not fully understood what kinds of knowledge and interaction strategies are used for coordination in team work, especially for CPS based on different perspectives, as considered in Hayashi et al. (2006).

Grounding in CPS based on different perspectives

Communication studies in cognitive science have shown how speakers establish common ground during conversation (Richardson & Dale, 2005; Galantucci, 2005). Grounding is the interactive process by which communicators exchange evidence in order to reach mutual understanding (Clark, 1996; Clark & Brennan, 1991). Studies of group decision-making have indicated that information shared among group members is an important factor in successful decision-making (Tindale, Kameda, & Hinsz, 2003). Thalemann and Strube (2004) showed that sharing information in initial and goal stages leads to better performance during collaborative problem solving. In contrast, cognitive science studies of collaboration have shown that common ground is unnecessary in cooperative tasks, in some cases (Barr, 2004). Computer simulations using multi-agents showed that a population of egocentric agents can establish and maintain systematic conventions without sharing common knowledge. This observation is partially supported by empirical experimental results Hayashi et al. (2006), which indicated that some participants were able to complete a task (discovering a rule) by simply using the shared information without developing common ground. However, when generating correct mental models of others' perspective during CPS, developing common ground is necessary.

Then, what kinds of interaction processes can develop successful grounding in CPS? Communication studies in cognitive science show that individuals coordinate with each other by generating explicit sign signals, which are implicitly aware of each other (Galantucci, 2005). Garrod and Anderson (1987) investigated how dyads developed different languages associated with different mental models in a maze configuration task. Individuals with different perspectives established common ground by generating spatial descriptions to successfully coordinate with each other. Additionally, in the initial phase, speakers used detailed, concrete descriptions to specify situations. Individuals used abstract signs as they proceeded during the task. Analysis of communication in the study of Hayashi et al. (2006) also showed that individuals use spatial characteristics (called regions) regarding the presented stimuli. However, the aim of the dialog analysis in this previous study was to capture the degree of perspective bias; spatial expressions were analyzed based on which perspectives were mentioned. Therefore, further analysis of the types of detailed knowledge that were used to attain shared perspective and further establish common ground would be valuable.

Goals and Hypotheses

The present study focused on how individuals share perspectives while establishing common ground in CPS in which members interact based on different perspectives. Based on Hayashi et al. (2006), our first goal was to conduct a cognitive task analysis to determine what kind of communication strategies participants used to reach shared perspectives. The cognitive task analysis was based on the method of Rittle-Johnson and Koedinger (2001) and Koedinger and Terao (2002). According to Rittle-Johnson and Koedinger (2001), developing cognitive models during cognitive-task analysis enables one to specify unambiguous problem representations and thus detail comparisons of the problem-solving strategies. This is useful here in terms of specifying the types of featured knowledge that were used during conversations on sharing perspectives. We hypothesized that coding individuals' utterances based on production rules would provide knowledge regarding what types of featured knowledge are used to share perspectives. Then, based on this cognitive task analysis, our second goal was to investigate which type of knowledge helps dyads to successfully reach shared perspectives. We hypothesized that dyads who used more specific features, and combinations of knowledge of those features, would be more likely to reach a shared perspective.

Method

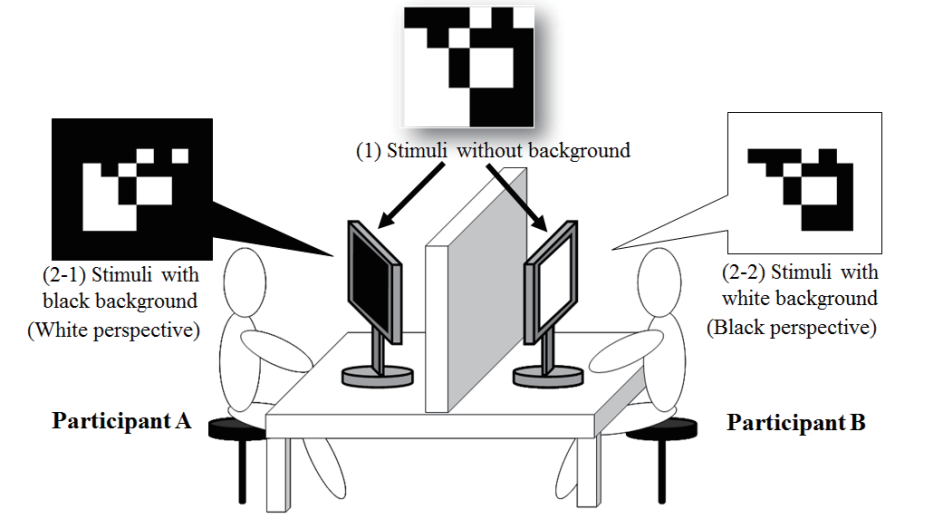
Participants

The present study reanalyzed the dataset of Hayashi et al. (2006) by analyzing dyads working with different perspectives (distributed view condition). The data of 22 Japanese university students (5 female, 17 male; *M*age: 20.73 years, SD: 2.27) who participated in dyads were reanalyzed.

Task

Controlling the participants' perspective

We reanalyzed data obtained from a simple rule-discovery task called the figure-ground reversal task, which was developed by (Hayashi et al., 2006) (for details, also see Hayashi (2018)). This task is similar to the story of "blind men and the elephant", where all individuals were touching an elephant but because they touched different parts they came to different conclusions about what they are touching. Pairs of participants collaborated through computer terminals that were separated by a partition so that neither could see the partner's display (see Fig 1). First, a square frame was presented for one second, and then the stimulus was presented in the frame. The presentation of a frame and a stimulus was considered as one trial. The participants were instructed to find the sequence rule of the number of objects that are presented through the trials. The participants were told to discuss the target rule and press the termination button presented on their screen when



Introductory phase(~ 16th trial) | Conflict Phase(17th~ trial)

Black objects	3	4	5	6	…	3	4	5	6	4	6	6	7	4	3	…
White objects	3	4	5	6	…	3	4	5	6	2	2	4	5	2	5	…
Total (Target)	**6**	**8**	**10**	**12**	…	**6**	**8**	**10**	**12**	**6**	**8**	**10**	**12**	**6**	**8**	…

(3) Sequence of presented stimuli (Sum of black and white objects rotate 6-8-10-12)

Figure 1: Experimental situation and task.

they reached the solution. The target rule was set to as the sequence of the sum of the black and white objects, i.e. the sum of the numbers of white and black objects rotating between 6, 8, 10, and 12. To manipulate a situation where the dyads were interacting based on different perspectives, principles from Gestalt psychology (Koffka, 1935), were used where the number of objects were fixed to change from figure to ground based on the background color. By putting the objects in different background, participants are led to have one of the distributed perspectives: i.e., either a perspective focusing on black objects or one focusing on white. In an example stimulus in Fig 1, there is a total of eight objects comprising three black objects and five white objects. This stimulus is displayed on either a black or white background and the participants have distributed perspective focusing on either one color as figure. The instructions stressed that the stimuli presented to each participant within the square frame were identical to each other, but the information about the background color was not mentioned.

Controlling disagreement about each other's perspectives

To control how the dyads incorporated different perspectives, the number of objects was adjusted to generate discrepancies when participants reported the numbers. In the initial stage (Introductory Phase), participants observed different colored objects (figure color) but reported the same number of objects (see Fig 1). Previous results using this task showed that participants reported the same number of objects in the Introductory Phase and therefore believed that they were looking at objects of the same color. As shown in (3) in Fig 1, participants simply reported varying numbers of objects (such as 3, 4, 5, or 6) in the Introductory Phase and thus generated misconceptions regarding the target rule. On the seventeenth trial (Conflict Phase), the number of the objects rotated by 3, 4, 5, or 6 and was scrambled. The number of objects was arranged so that only the sum of the number of objects would represent a valid response. After the Conflict Phase, participants needed to modify their misconceived initial hypothesis and instead count both black and white objects to discover the rule. It should be noted that, the participants have to discover the rule across observing the trials within the single task conducted in this experiment.

Data collection

Task Analysis

This task could proceed by two different types of interaction: (1) each participant reported only the number of figure objects (non-shared perspective method), or (2) each participant counted figure and ground objects (shared perspective method). To proceed with method (2), participants need to set a sub-goal, which was to develop mutual understanding of why they were reporting different numbers after the 17th trial. To develop a concrete shared perspective, they needed to discuss details of the display and understand how to count both figure- and ground-colored objects. Taking these issues

into consideration, we provide an overview flowchart for representative dyads working on the task by establishing common ground in Fig 2.

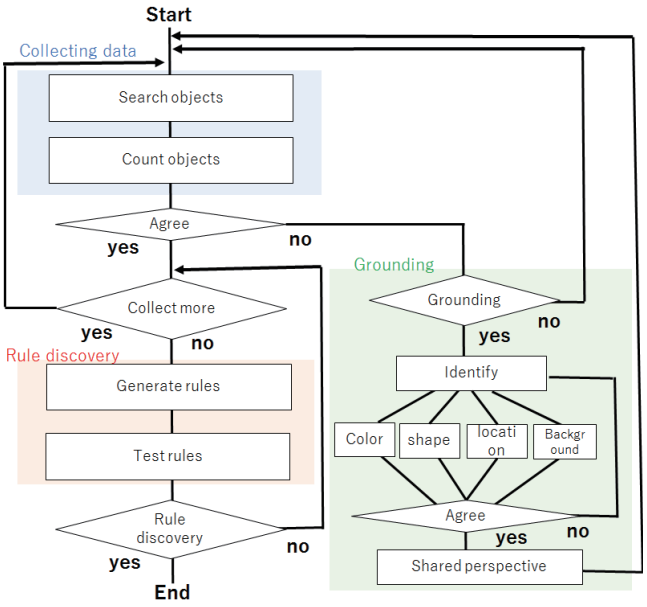


Figure 2: Flow of problem solving based on task analysis.

To establish common ground, featural knowledge, such as (a) color, (b) shape, (c) location, and (d) background, was essential. We conducted cognitive task analysis based on the type of these features, as explained in the next section.

Production-rule model for shared perspective

Production rules in this study consisted of declarative and procedural knowledge, as in ACT-R (Anderson, Corbett, Koedinger, & Pelletier, 1995). The production rules were stated in IF-THEN format, which consisted of declarative chunks. In our task, we focused on the four types of information shown, namely (1) location of the object: ?E[“location”’], (2) shape of the object: ?F[“shape”’], (3) color of the object: ?G[“color”’], and (4) background of the object: ?H[“background”’]. Using these variables, a declarative knowledge “chunk” can be defined. For example, a chunk associated with mentioning a particular object can be defined in the following way.

#Location-shape
-color(3-way)

isa	object
trial	1
number	1
location	?E[upper left]
shape	?F[tetra-zoid]
color	?G[white]
background	null

In this analysis, we only focused on combinations with color(?G[]) and other knowledge for the 2-way and 3-way, because color information was considered key for perspective-taking in this task. Next, we examined associations between the number of dyads who used specific featural knowledge (using more feature combinations) and success in sharing perspectives. Utilizing shared perspectives was defined based on the following evidence: (1) explicitly mentioning that they can see the partner’s perspective (opposite color to the background) during their grounding process, or (2) counting both black and white colored shapes after their grounding process. For example, evidence for (1) could be ”I understand what you mean and I can see the tetra-zoid in the black”, whereas (2) could be ”Now I know your perspective I will count both and I see four in black and six in white.”

Results: Association between # of featural knowledge types and shared perspective

For all 11 dyads, we conducted Fisher’s exact test to compare 2 (Featural knowledge: Mentioned vs. Not mentioned) $\times$ 2 (Shared perspective: Successful vs. Unsuccessful).

one-way strategy The results revealed no significant differences between establishing shared perspective and feature type, i.e., location (p = 0.49, FET), shape (p = 0.15, FET), color (p = 0.27, FET), or background (p = 0.06, FET). This indicates that sharing only one feature did not facilitate success in sharing perspectives.

two-way strategy Results revealed significant differences in establishing shared perspective by combinations of feature types, namely according to color & location (p = 0.02, FET) and the combination of color & shape (p = 0.02, FET). However, there was no relationship between establishing shared perspective and the combination of color & background (p = 0.18, FET). Comparing these results with the one-way strategy, we can see that the more features were mentioned during the conversations, the more likely it was that participants successfully shared perspectives.

three-way strategy There were also significant differences between establishing shared perspective and three-way combination of features, namely color & location & shape(p = 0.02, FET). This also supports the hypothesis that the more features are used, the more participants are able to share perspectives. Table 2, 3, 4 shows a summary of the results. F/S stands for feature mentioned/shared perspective, F/N stands for feature mentioned/not shared perspective, N/S stands for not feature mentioned/shared perspective, N/N strands for not feature mentioned/not shared perspective.

Discussion and Conclusions

Our first goal was to conduct cognitive task analysis to understand the types of featural knowledge that were used during interactions during the grounding process. Based on Rittle-Johnson and Koedinger (2001), we developed production rule models for knowledge regarding features of the images pre-

Table 1: Example dialog coded by production rules and types of chunks.

Speaker	Example Dialog	Productions rules	Chunk Type
B	"I see a tetra-zoid on the upper left corner"	If goal is to grounding and there is an object ?E[upper left] with feature ?F[tetra-zoid] Then express "the object is ?F[tetra-zoid]"	#Location-shape (2-way)
A	"(tetra-zoid)On the upper left? I don't see such thing"	If goal is to grounding and If partner says object has feature ?F[tetra-zoid] and object does not have feature ?F[tetra-zoid] in ?E[upper left] Then express "NO " and search new feature	#Location-shape (2-way)
B	"What about a shape "T" on the upper right?"	If goal is to grounding and there is an object ?L[upper right] with feature ?F[T] Then express "the object is ?F[T] at ?E[upper right]"	#Location-shape (2-way)
A	"You mean (upper right T) in black? Not in white?"	If goal is to grounding and partner mentions a new feature perspective ?F[T] and the object ?E[upper right] being is discussed is ?G [Black] Then confirm ?F[T] is ?G[Black] not ?G[White]	#Location-shape -color(3-way)
A	"Oh! I see it(upper right T) in black!"	If goal is to grounding and there is an object ?L[upper right] that matches the feature ?F[T] from partner in color ?G[Black] Then say ?G[Black] and "yes"	#Location-shape -color(3-way)

Table 2: Summary of association between knowledge types and shared perspective: 1-way feature.

1-way feature	F/S	F/N	N/S	N/N
color	8	2	0	1
shape	7	1	1	2
location	6	1	2	2
background	6	0	2	3

Table 3: Summary of association between knowledge types and shared perspective: 2-way feature. * indicates statistical significance.

2-way feature	F/S	F/N	N/S	N/N
color & shape*	7	0	1	3
color & location*	7	0	1	3
color & background	5	0	3	3

Table 4: Summary of association between knowledge types and shared perspective: 3-way feature. * indicates statistical significance.

3-way feature	F/S	F/N	N/S	N/N
color & shape & location*	7	0	1	3

sented in the experiment. The conversations within the dyads were transcribed into production rules, defined by declarative features of knowledge, which consisted of shape, location, color, and background. Through this cognitive task analysis, we discovered that dyads used combinations of featural knowledge when developing mutual understanding of each other's different perspectives. Simply put, collaborators who were more specific about what they were talking about were more likely to reach shared perspective. More precisely, our cognitive task analysis indicated that although all collaborating pairs described the objects they were seeing with a variety of features (e.g., color, shape, location), the successful pairs were more specific in using combinations of features (e.g., "the white T in the upper right" rather than "the white one" or the "the T"). Returning to the blind men and the elephant example introduced earlier, our results suggest that might, eventually, individuals reach agreement if they are more specific – describing as much as they can, the shape, texture, smell, relative location of their observations, etc. Past studies of communication showed that speakers use combinations of detailed spatial information to establish common ground (Garrod & Anderson, 1987); the current results are consistent with those studies. Moreover, once the participants established common ground, they tended to use simple phrases

when counting shared perspectives such as "two-four"' and "four-four". These can be considered as types of conceptual packs (Brennan & Clark, 1996), which are used when common ground is established during conversations.

In our study, we used quantitative analysis to investigate whether use of specific combinations of knowledge yielded higher performance in sharing perspectives. As hypothesized, dyads who were more specific in their grounding, i.e., mentioned more combinations of features, were more likely to reach a shared perspective. More specifically, participants who mentioned color and shape (2-way strategy), or color, shape, and location (3-way strategy) performed relatively well in sharing perspectives. Thus, specifying spatial information facilitates success in shared perspectives. There may be general critiques such as, "can common ground achieved by simply describing what is relevant?" To answer this question, we must first consider the point how did the participants determine what's relevant. In trying to achieve common ground with another, it seems possible, even likely, that one cannot fully anticipate the ambiguities that the other person may be experiencing or anticipate the alternative view of the world that they are seeing and perceiving. As mentioned previously, from the example of the story of the "blind men and the elephant", what features to focus on may be unclear and therefore, one cannot figure out what's relevant. One can simply try to be as specific as possible about in describing what they are seeing and perceiving.

Another important point that should be stressed from this study is the type of expressions that the speakers were using after they recognized about their conflicts. In natural conversations it is more efficient in most settings to not be specific(Grice, 1975). However, in a situation such as in this task, where participants have become aware of confusions and discrepancies, one must work hard to avoid our natural tendencies to be more efficient (less redundant) in our speech. Speakers need to strive for more redundancy and explicitness. Apparently, this switch is not easy to make as many of our participants do not seem to make the switch and continue to speak in natural, efficient but less specific and redundant ways. As future work, we will further investigate these points to uncover the conversational dynamics and discover the mechanisms of shared understanding in collaborative tasks.

This paper provides new implications regarding the methods one may use to capture the collaborative process in a systematic way. Although, there are limitations to the current study, primary among which is the small number of dyads used. As mentioned above, one of the next steps is to conduct a more focused experiment on how participants establish common ground. Specific analysis using eye movements and conversational data will likely be useful to elucidate the nature of coordination, too. Another possible future directions of this study is to further conduct simulations using the production-rules to further confirm our results on how the individual establish common ground.

Acknowledgments

This work was supported by the Grant-in-Aid for Scientific Research (KAKENHI), No. 16KT0157.

References

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *Journal of the Learning Sciences*, *4*(2), 167-207.

Aronson, E., & Patnoe, S. (1997). *The jigsaw classroom(2nd ed.): Building cooperation in the classroom*. New York: Addison Wesley Longman.

Asterhan, C. S. C., & Schwarz, B. B. (2009). Argumentation and explanation in conceptual change: Indications from protocol analyses of peer-to-peer dialog. *Cognitive Science*, *33*(3), 374–400.

Barr, J. (2004). Establishing conventional communication systems: Is common knowledge necessary? *Cognitive Science*, *28*(6), 937-962.

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 1482-1493.

Chi, M. (2009). Active-constructive-interactive: A conceptual framework for differentiating learning activities. *Topics in Cognitive Science*, *1*(1), 73–105.

Chi, M., Leeuw, N., Chiu, M., & Lavancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, *18*(3), 439-477.

Clark, H. H. (1996). *Using language*. Cambridge University Press.

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In B. L. Resnick, M. R. Levine, & D. S. Teasley (Eds.), *Perspectives on socially shared cognition* (p. 127-149). APA Press.

Dasgupta, S. (2003). Multidisciplinary creativity: The case of herbert a. simon. *Cognitive Science*, *27*(5), 683-707.

Dunbar, K. (1995). How scientists really reason: Scientific reasoning in real-world laboratories. In J. R. Sternberg & E. J. Davidson (Eds.), *The nature of insight* (p. 365-395). MIT Press.

Fiore, S. M., Rosen, M. A., Smith-Jentsch, K. A., Salas, E., & Letsky, N., M.and Warner. (2010). Toward an understanding of macrocognition in teams: Predicting processes in complex collaborative contexts. *Human Factors*, *52*, 203-224.

Galantucci, B. (2005). An experimental study of the emergence of human communication systems. *Cognitive Science*, *29*(5), 737-767.

Garrod, S., & Anderson, A. (1987). Saying what you mean in dialogue: A study in conceptual and semantic coordination. *Cognition*, *27*(2), 181-218.

Greeno, G. J., & de Sande, C. (2007). Perspectival understanding of conceptions and conceptual growth in interaction. *Educational Psychologist*, *42*(1), 9-23.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics: Vol. 3: Speech acts* (p. 41-58). New York: Academic Press.

Hayashi, Y. (2018). The power of a "maverick" in collaborative problem solving: An experimental investigation of individual perspective-taking within a group. *Cognitive Science*, *42*(S1), 69-104. doi: 10.1111/cogs.12587

Hayashi, Y., Miwa, K., & Morita, J. (2006). A laboratory study on distributed problem solving by taking different perspectives. In *Proceedings of the 28th annual conference of the cognitive science society(cogsci2006)* (p. 333-338).

Koedinger, R. K., & Terao, A. (2002). A cognitive task analysis of using pictures to support pre-algebraic reasoning. In *Proceedings of the annual meeting of the cognitive science society(cogsci2002)* (Vol. 24).

Koffka, K. (1935). *Principles of gestalt psychology*. Routledge and Kegan Paul.

Miyake, N. (1986). Constructive interaction and the interactive process of understanding. *Cognitive Science*, *10*(2), 151-177.

Okada, T., & Simon, H. (1997). Collaborative discovery in a scientific domain. *Cognitive Science*, *21*(2), 109-146.

Richardson, D. C., & Dale, R. (2005). Looking to understand: The coupling between speakers' and listeners'eye movements and its relationship to discourse comprehension. *Cognitive Science*, *29*(6), 1045-1060.

Rittle-Johnson, B., & Koedinger, R. K. (2001). Using cognitive models to guide instructional design: The case of fraction division. In *Proceedings of the annual meeting of the cognitive science society(cogsci2002)* (Vol. 23).

Roschelle, J. (1992). Learning by collaborating: Convergent conceptual change. *Journal of the Learning Sciences*, *2*(3), 235-276.

Salomon, G. (2001). *Distributed cognition: Psychological and educational considerations*. New York: Cambridge University Press.

Shirouzu, H., Miyake, N., & Masukawa, H. (2002). Cognitively active externalization for situated reflection. *Cognitive Science*, *26*(4), 469-501.

Thalemann, S., & Strube, G. (2004). Shared knowledge in collaborative problem solving: Acquisition and effects. In *Proceedings of the twenty sixth annual conference of the cognitive science society.*

Tindale, R. S., Kameda, T., & Hinsz, B. V. (2003). *Group decision making: Review and integration*. M. A. Hogg and J. Cooper (Eds.).

An Ontology of Decision Models

Lisheng He
University of Pennsylvania, Philadelphia, Pennsylvania, United States

Wenjia Joyce Zhao
University of Pennsylvania, Philadelphia, Pennsylvania, United States

Sudeep Bhatia
University of Pennsylvania, Philadelphia, Pennsylvania, United States

Abstract

Decision models are formal algorithms that are used to represent decision processes and predict choice across a wide range of disciplines. These models are often highly complex, which makes it difficult to understand the relationships between different models, the unique features of individual models and, in turn, the fundamental properties of choice behavior captured by these models. We address this issue in a large-scale computational analysis that uses parameter bootstrapping cross-fitting techniques to derive pairwise measures of decision model distances. Our analysis includes over 80 prominent models of risky and intertemporal choice, and results in an ontology of decision models, with data-driven model clusters and hierarchies that synthesize over seven decades of quantitative research on human choice behavior.

Rapid Unsupervised Encoding of Object Files for Visual Reasoning

Rachel Flood Heaton (rmflood2@illinois.edu)
Department of Psychology, University of Illinois, 603 E. Daniel St.
Champaign, IL 61820

John E. Hummel (jehummel@illinois.edu)
Department of Psychology, University of Illinois, 603 E. Daniel St.
Champaign, IL 61820

Abstract

Visual thinking plays a central role in human cognition, yet we know little about the algorithmic operations that make it possible. Starting with outputs of a JIM-like model of shape perception, we present a model that generates object file-like representations that can be stored in memory for future recognition, and can be used by a LISA-like inference engine to reason about those objects. The model encodes structural representations of objects on the fly, stores them in long term memory, and simultaneously compares them to previously stored representations in order to identify candidate source analogs for inference. Preliminary simulation results suggest that the representations afford the flexibility necessary for visual thinking. The model provides a starting point for simulating not only object recognition, but also reasoning about the form and function of objects.

Keywords: visual reasoning; shape perception; object files; structural description; type-token problem

Introduction

Visual thinking plays a central role in human cognition. From deciding whether a quantity of soup will fit into a storage container, to interpreting graphical representations of data, or reading a circuit schematic, people routinely engage visual reasoning in the service of understanding the world and solving problems. Visual thinking figures prominently in our most creative and uniquely human acts, including mathematics, engineering, art and design. But in spite of its centrality, comparatively little is known about the algorithmic basis for visual and visually-assisted reasoning (but see Hummel & Holyoak, 2001, Johnson-Laird, 1983, Lovett and Forbus, 2017, for progress in this direction). Instead, most computational work in high-level vision has been and continues to be addressed to the problem of object recognition, the tacit assumption often being that object recognition is the final stage of ventral visual processing, as though once an object has been visually recognized, there is nothing left to be done. Most models in this tradition, including modern deep nets for object recognition, represent objects as holistic templates of various kinds, which is a representational format that does not lend itself to any kind of explicit visual reasoning (Hummel, 2000; see Hummel, 2013, for a review).

The problem of visual thinking places strong constraints on the kinds of representations—for example of object shape or scene layout—the visual system must deliver to the rest of the cognitive architecture. It places equally important constraints on the kind of cognitive architecture that operates on those representations (Hummel, 2000; Lovett & Forbus, 2017). In particular, that architecture must be prepared to reason and generalize extremely flexibly—specifically, with the flexibility of an explicitly relational (i.e., symbolic) system (Hummel & Holyoak, 1997, 2001, 2003a; Lovett & Forbus, 2017). And for that purpose, the visual system must be equipped to represent the visual world in terms of arrangements of objects and object parts in terms of their spatial relations (as opposed to, e.g., their literal locations in the retinal image; Hummel, 2000).

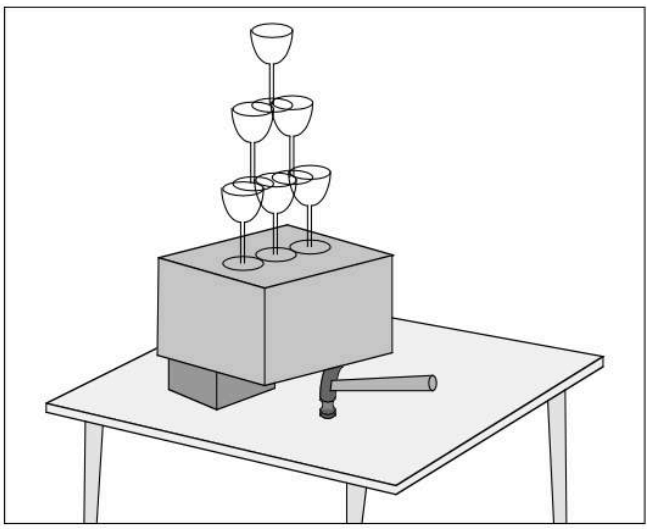


Figure 1: An example of visual reasoning in a novel context (Green & Hummel, 2004). Even if one has never seen this image before, it is obvious that moving the hammer is ill-advised.

Consider, for example, the arrangement depicted in Figure 1 (from Green & Hummel, 2004), and imagine oneself in need of the hammer. Upon a glance at the figure, it is clear that one should not simply pick up the hammer, as doing do would cause the wine glasses to fall and break. We can easily understand this property of Figure 1 in spite of the fact that, for most people, the arrangement in the figure is completely unfamiliar. To put the power of this inference into perspective, note that an associative response to Figure 1

(e.g., of the kind that would be learned by a deep net) might specify that there was something fragile in the scene, and it might even specify that a hammer as an object capable of breaking things, but would it would be incapable of even representing (much less inferring) a complex relational thought such as "moving the hammer is ill-advised because it would result in the wine glasses falling." Although the natural associative relation between hammers and breaking is to think of hammers as objects that break things, in the case of Figure 1, the hammer is *preventing* the glasses from being broken.

Making the appropriate inference about the arrangement in Figure 1 requires us to perceive the spatial relations between the hammer, the boxes and the wine glasses, and to infer from those relations what kinds of actions will and will not result in the glasses falling (Green & Hummel, 2004). Crucially, this inference depends much more on the relations between the objects than on the features or identities of the objects themselves: If we were to replace the wine glasses with a baby, the same relations would be in place, and the same inference would follow; the same is true if we replace the hammer with any other object of an appropriate size to support the box.

Similarly, even recognizing and reasoning about a novel instance of a known object class (for example, a new kind of coffeemaker), requires this kind of representational flexibility: the carafe of a coffeemaker may not always be perfectly cylindrical, especially if its designer was feeling creative, but it will always reside below the filter basket. The coffeemaker may even contain extra parts (e.g. thrown in for flourish) or have parts removed for a minimalist aesthetic, but barring extreme artistic license, it will still be recognizable as a coffeemaker.

In other words, visual inference, and even object recognition, depend on our ability to represent relations independently of the object/parts serving as arguments of the relations, and to simultaneously bind the objects/parts to their relational roles (Hummel & Biederman, 1992; Hummel & Holyoak, 1997, 2003a).

In summary, what is needed is a visual system capable of delivering relational (i.e., symbolic) representations of objects or scenes in terms of their constituent parts and the relations among them, and a cognitive architecture that is capable of using those representations in order to make flexible relational inferences.

Perceiving Relations with JIM and Reasoning About Them with LISA

Models of high-level vision that generate explicitly relational representations are comparatively rare. The examples with which we are familiar are Winston (1975), Lovett and Forbus (2017), and Hummel and Biederman's (1992; Hummel, 2001; Hummel & Stankiewicz, 1996, 1998) JIM. We will focus on JIM, a neural network that was originally developed as a model of object recognition, and in that context has accounted for, and successfully predicted, a very large number of findings in the literature on shape perception and object recognition (for a review, see Thoma & Davidoff, 2007). As such, JIM provides a psychologically and neurally plausible theory of the shape representations that can be derived from line drawings of objects. As elaborated shortly, the model is also useful as a basis for visual reasoning because it generates visual representations that are both explicitly relational and in a format that is directly usable by the LISA model of relational reasoning (Hummel & Holyoak, 1997, 2003a).

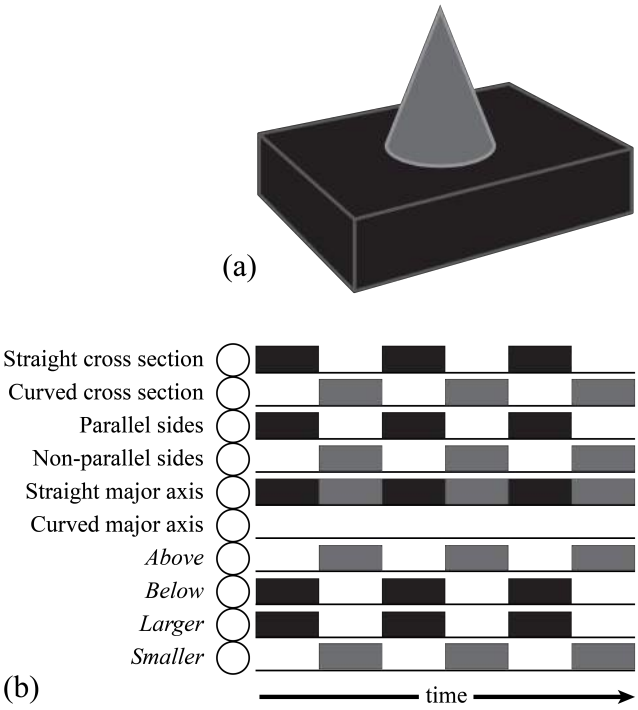


Figure 2: (a) A cone on top of a brick. (b) The JIM representation of a cone on top of a brick. Circles are units representing shape attributes. Bars indicate the activity of corresponding units over time, with black bars corresponding to the brick and gray to the cone.

JIM (Hummel & Biederman, 1992) represents objects as configurations of geons (basic volumetric shapes; Biederman, 1987) in specific spatial relations to one another. For example, the simple object in Figure 2a would be represented as a cone on-top-of, smaller-than and orthogonal-to a brick. The cone and brick are represented in JIM, not as atomic primitives, but as patterns of activation distributed over neuron-like units representing their shape attributes (Figure 2a). For example, the cone would be represented by units specifying that it has a curved cross section, a straight major axis, non-parallel sides, and a slightly elongated aspect ratio; the brick would be represented as having a straight cross section, a straight major axis, parallel sides, and a slightly elongated aspect ratio. The units representing the cone are bound to units representing its relational roles (here, *smaller* and *above*), and the units for the brick are bound to its roles (*larger* and *below*) by synchrony of firing: Units

representing the cone and its roles fire in synchrony with one another, and out of synchrony with the units representing the brick and its roles (Figure 2b). (These synchrony relations are established in the model's V1- and V2-like first layers, by lateral interactions between local units representing the geons' edges and the vertices where they coterminate; see Hummel & Biederman, 1992.)

The resulting representations (Figure 2b) are then matched to stored representations in JIM's long-term memory (LTM) for the purposes of object recognition. This representational format also happens to be identical to the format LISA (Hummel & Holyoak, 1997, 2003a) uses to represent role-argument bindings for the purposes of relational reasoning. In LISA, relational roles and their arguments are represented as patterns of activation over units representing their semantic content, and bound into complete propositions by synchrony of firing: Within a proposition, such as *on-top-of* (cone, brick) or *loves* (John, Mary), units for a relational role (e.g., *above*, *below*, *lover*, or *beloved*) fire in synchrony with the units representing the arguments to which they are bound (with *above* firing with *cone*, or *lover* firing with *John*) and out of synchrony with the units coding the proposition's other role bindings (*below+brick* or *beloved+Mary*).

Armed with these representations, LISA accounts for roughly 100 major empirical phenomena in relational reasoning, including its development (e.g., Doumas et al., 2008) and its decline with brain damage, normal aging, and frontotemporal dementia (for reviews, see Hummel & Holyoak, 2003b; Knowlton et al., 2012). As such, we take JIM and LISA as empirically well-grounded starting points for developing a model of visual thinking.

Although the kinds of representations JIM generates provide a natural basis for reasoning by LISA, the problem remains of adapting JIM-like representations for a LISA-like inference engine. That problem is the focus of the current modeling effort.

Object Files as a Basis for Visual Reasoning

Figure 2b illustrates the kind of distributed representation LISA uses to represent the semantic content of propositions in working memory (WM). To encode these representations into LTM, LISA uses a hierarchy of progressively more localist representations (Figure 3). At the bottom of the hierarchy, semantic units represent relational roles and their arguments in a distributed fashion (as in Figure 2b). *Argument* and *role* units (Figure 3) code arguments and relational roles in a localist fashion and share bidirectional excitatory connections with the corresponding semantic units. *Sub-proposition* (SP) units locally code role-filler bindings, such as *above+cone* and *below+brick*, and *proposition* units bind multiple role-filler bindings into complete propositions, such as *on-top-of* (cone, brick). Collections of related propositions are linked together with *group* (for our current purposes, *object file*) units. The resulting hierarchy of units serves both to represent propositions in LTM and as the basis for analogical mapping and the other functions LISA performs.

This hierarchy serves as a natural way to represent structural descriptions of objects and scenes (a very similar hierarchy encodes objects into LTM in JIM; Hummel & Biederman, 1992). For example, in order to represent an object, propositions would represent the spatial relations among the object's parts, and collections of such propositions would constitute a description of the complete object. Moreover, these descriptions can be nested hierarchically (with propositions taking other propositions as arguments; Hummel & Holyoak, 1997), making it possible to represent entire scenes as hierarchical collections of objects in various relations to each other.

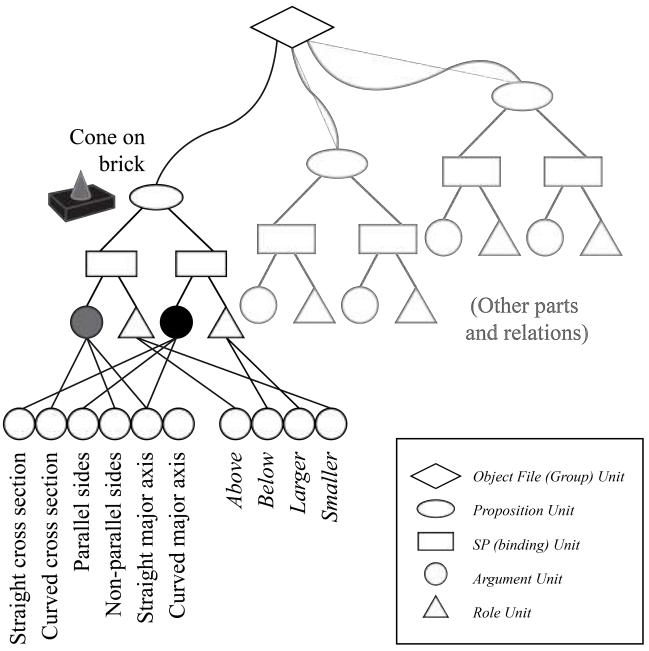


Figure 3. The LISAese representation of an object file. Left: representation of the proposition expressing the relations between the cone and the brick (roughly, *on-top-of-and-smaller-than* (cone, brick)). In light gray: other potential propositions in the object file.

Borrowing from Kahneman et al. (1992), we refer to collections of propositions encoding the properties of objects and/or scenes as *object files*. Importantly, the propositions composing an object file are assumed to encode (in the limit) everything visible about the object, including its shape, color, trajectory, and so forth. We also assume they are hierarchical, describing the properties of (and relations among) both whole objects and of individual object parts. In other words, we assume that the goal of early- and middle-vision, as well as visual attention, is to deliver a hierarchical description of the visual world. Although JIM provides an algorithmic basis for computing some of these properties and relations from object images, we assume that the visual input to the object files is much richer than any computational model is currently capable of providing. In the current effort, we therefore assume this visual input as a given to the model. Specifically,

we assume visual preprocessing that delivers descriptions of objects in terms of their properties (shape, color, etc.), spatial relations to one another, and the spatial relations among their parts. We assume that these descriptions are temporally bound into packages corresponding to bindings of relational roles to their arguments (e.g., Figure 2b; Hummel & Biederman, 1992), where the arguments can either be whole objects or object parts.

The Model

Given such a representation as a basic input, constructing an explicit object file from that input means encoding the propositions—i.e., collections of synchronized patterns of activation—into active memory (Cowan, 2001) so that they can be compared to the contents of LTM and reasoned about. The current model borrows and adapts elements of LISA's *self-supervised learning* algorithm (Hummel & Holyoak, 2003a) to accomplish this task. Like LISA and JIM, the current model uses synchrony of firing at multiple temporal scales in order to bind roles to their arguments, role-argument bindings into complete propositions, and collections of propositions into whole objects or scenes.

At the fastest temporal scale (i.e., the *phase*, which we assume to last about 25 ms; Hummel & Holyoak, 1997), units coding relational roles fire in synchrony with the units coding for the features of their arguments. At the next temporal scale (the *phase set*, corresponding to about 100 - 200 ms), mutually desynchronized role-filler bindings are grouped into complete propositions. And at the slowest temporal scale (corresponding to about 200 – 1000 ms), multiple propositions (phase sets) are grouped into complete units—either whole objects, or small groups of objects in specific relations (Green & Hummel, 2004). Each of these temporal scales corresponds to a specific kind of unit in the hierarchy in Figure 3, with the fastest corresponding to argument, role and SP units, the second slowest corresponding to proposition units, and the slowest to object file (group) units; units at each scale of the hierarchy integrate their inputs over corresponding temporal intervals (Hummel & Holyoak, 1997).

One at a time, patterns of activation corresponding to individual phases, i.e., parts or objects in specific relational roles, are presented to the model. These patterns correspond to packages being delivered by early to middle visual processing (e.g., in visual area LOC). In response to each such package, the model's task is twofold: One task is to encode new packages (phases), as they arrive into active memory, and integrate them into the representation of the emerging object file (Figure 3). This operation is performed by a simple kind of mapping-guided Hebbian learning (i.e., Hummel & Holyoak's, 2003a, self-supervised learning). At the same time, the model performs the parallel task of matching these incoming patterns to stored patterns in LTM (stored object files). That is, the model attempts to recognize each stimulus as an instance of a familiar object category at the same time as it encodes it into active memory as a new object file to be reasoned about.

By the time several phase sets have been processed, the object file will contain a collection of propositions describing (for example) the object's parts in terms of their spatial relations. If the object is familiar, the model will also have activated one or more existing object files in LTM, effectively recognizing/categorizing the object. The preceding describes the model in the language of visual cognition. In the language of analogical reasoning, the model will have encoded a new *target analog* (the object file) to be reasoned about, and it will have retrieved one or more *source analogs* (i.e., existing object files) to use in the service of reasoning about the target. Once this process is complete, the machinery of analogical reasoning (as embodied in LISA) can take over, mapping the target onto the source in order to identify corresponding elements and relations, using the source to drive inferences about the target, and inducing a more abstract schema capturing what the source and target have in common (Hummel & Holyoak, 2003a).

Token Formation

This very coarse description of the model's operation necessarily glosses over numerous implementation details, but all of these are standard to LISA's operation (see Hummel & Holyoak, 2003a, Appendix A). However, one aspect of the algorithm warrants discussion in greater detail. In LISA, argument, role, SP, proposition, and group units represent *tokens* of objects, roles, and so forth, in the context of the larger structure in which they reside. For example, the *cone* unit in Figure 3 represents a token of "cone" in the context of the specific object file depicted in the Figure; the abstract *type* "cone" is represented by the shape units (in LISA, "semantic units") to which the token is connected (Figure 2b). This type/token distinction becomes apparent in the case of scenes containing more than one instance of a given object or geon: If an image contains, say, two cones, then the resulting object file must contain separate argument units for each, even if those units are connected to otherwise identical shape units: Constructing an object file from an image requires the model to distinguish clearly between types ("a cone") and tokens of those types ("this cone").

Keeping this type/token distinction straight is complicated by the fact that a given token is likely to fire more than once in the output of visual processing: If the features of a cone fire at time *t*, and a cone also fired at time *t*-5, then how can we know whether the cone that is firing now is the same one (the same *token*) that fired 5 iterations ago? (In this respect, the object files created by the current model differ from those postulated by Kahneman et al., 1992, in that their object files were assumed to be unitary tokens for single objects. By contrast, the object files created here are hierarchical tokens that can, themselves, contain tokens for smaller parts.)

The current model solves this problem by exploiting the role of mappings in self-supervised learning (Hummel & Holyoak, 2003a). In brief, the current model, like LISA, knows when a new token is required by knowing the mappings between the tokens composing the source of an inference (here, an object file in LTM) and those composing

the target of that inference (the emerging object file): If an unmapped token fires in the source, then a new token is required in the target. The current model exploits a similar constraint by mapping each token in the emerging object file to the location of the corresponding part or object in the image: In essence, it knows whether the cone firing at time *t* is the same token as the one from time *t*-5 by knowing whether they occupy the same location. (This heuristic is admittedly too simple and will fail with, for instance, moving stimuli. In general, we assume that tokens are distinguished, not by locations in the image, but by spatiotemporal trajectories in 3-space.)

The model is still in an early stage of development—and is, itself, only a component of a much larger emerging model—but preliminary simulations provide an encouraging proof of concept.

Simulations

We ran four sets of simulations as basic tests the model's ability to rapidly encode object files from oscillatory visual inputs of the kind illustrated in Figure 2b. In each simulation, objects were presented and encoded in the model's LTM; subsequently, additional objects were presented to be encoded and categorized as one of the known objects. Objects were constructed by combining 14 parts, P1… P14, into arrangements by placing them in various two-place relations, with roles R1…R15. Each part was coded as a 10-dimensional feature vector, and each role of a relation was also coded as a 10-dimensional vector. In addition, 6 units served as location tags, L1…L6, which as discussed above, permit the model to solve the type-token problem. The full feature space was thus 26-dimensional (10 for parts, 10 for roles, and 6 for location tags). The binding of a given part, P_i, to a given relational role, R_j, was implemented as the concatenation of vector P_i with vector R_j and location vector L_k (synchrony of firing is equivalent to vector addition). We manipulated the relationships between the stored and stimulus objects by varying the parts of the objects, *P*, the relations, *R*, and the locations *L* in which they were instantiated. For clarity in what follows, we will refer to a given part in a given location as $P_{i,k}$. The assignment of features to part vectors *P*, role vectors, *R*, and location vectors *L* was randomized on every simulation.

Table 1 shows the library of objects used in all simulations. In the table, objects are denoted using the format (using object O_1 as an example):

$$[(P_{1,1}, R_1) + (P_{4,4}, R_2)], [(P_{1,2}, R_3) + (P_{4,4}, R_4)],$$

where $(P_{1,1}, R_1)$ denotes part P_1 in location L_1 bound to role R_1, and $(P_{4,4}, R_2)$ denotes P_4, in L_4, bound to R_2; and the square brackets around these expressions indicate that roles R_1 and R_2 form a single relation linking P_1 to P_4. Note that P_1 appears in two locations in O_1, L_1 and L_2, and thus instantiates two tokens of the same type in the representation of O_1.

Simulation 1 was the most basic test of the model's ability to encode and match objects. We encoded objects O_1-O_3 into the model's memory and then tested its ability recognize object O_1. Unsurprisingly, it recognized O_1 as O_1 on three of three simulation runs, in the sense that it activated the O_1 group unit more than the group units for O_2 or O_3 (roughly 0.7 versus 0.6 or less, respectively; objects O_2 and O_3 are as active as they are because there is no lateral inhibition between group [object file] units).

Simulation 2 tested the model's ability to recognize an object when it has an extra part. On three runs, the model was initially trained on objects O_1-O_3, and then tested with O_4. O_4 is the same as O_1, but with an extra part, P_3, in a new relation to P_4. In addition to encoding O_4 as a new object file, the model also recognized it as most similar to object O_1 with activation about 0.7, versus about 0.5 for O_2 and O_3. When the model was then tested with O_1 as a stimulus (after O_4 was encoded into memory), the model recognized O_1 as O_1 (about 0.7), but also activated O_4 as a close match (about 0.6 versus about 0.5 for O_2 and O_3).

Table 1: Object Library for Simulations

O_1	$[(P_{1,1}, R_1) + (P_{4,4}, R_2)], [(P_{1,2}, R_3) + (P_{4,4}, R_4)]$
O_2	$[(P_{5,5}, R_7) + (P_{11,11}, R_8)], [(P_{8,8}, R_9) + (P_{11,11}, R_{10})], [(P_{10,10}, R_{11}) + (P_{11,11}, R_{12})]$
O_3	$[(P_{13,13}, R_{14}) + (P_{4,4}, R_2)], [(P_{12,12}, R_{13}) + (P_{12,12}, R_{14})]$
O_4	$[(P_{1,1}, R_1) + (P_{4,4}, R_2)], [(P_{1,2}, R_3) + (P_{4,4}, R_4)], [(P_{3,3}, R_5) + (P_{4,4}, R_6)]$
O_5	$[(P_{5,5}, R_7) + (P_{11,11}, R_8)], [(P_{10,10}, R_{11}) + (P_{11,11}, R_{12})]$
O_6	$[(P_{6,6}, R_7) + (P_{11,11}, R_8)], [(P_{8,8}, R_9) + (P_{11,11}, R_{10})], [(P_{10,10}, R_{11}) + (P_{11,11}, R_{12})]$

Simulation 3 tested the model's ability to recognize an object with a missing part. In three runs, the model was again trained with O_1-O_3 and tested with O_5, which is like O_2, but missing part P_8. The model correctly recognized O_5 as most similar to O_2 on two out of the three runs. On the third run, the model classified O_5 as most similar to both O_2 and O_1 equally. We speculate that in this case the part and relation vectors randomly generated for O_1 happened to be similar to those of O_2, in which case this result would be an example of a neighborhood effect. However, in all simulations, when O_2 was re-presented after O_5 was encoded, the model recognized it as an instance of O_2, with O_5 as a close second (both near 0.7), preferentially activating both over O_1.

Finally, simulation 4 tested the effect of replacing one part with another. Again, in three runs, the model was trained on O_1-O_3, and then tested with O_6 (O_6 is like O_2, but with P_5 replaced by P_6). On two of three runs, the model recognized O_6 as most similar to O_2. On the third, the model slightly favored O_3. Once again, we speculate that this result is due to neighborhood effects created by the randomization of the vectors. In all runs, when O_2 was re-presented to the model, it activated O_2 (greater than 0.7), with O_6 as a close second.

Discussion

The online generation of object files from the output of middle-to-late vision is a crucial step in visual thinking. We

present a model that, starting with outputs of a JIM-like model of shape perception, generates representations that can be stored in memory for future recognition and can be used by a LISA-like inference engine to reason about those objects. Preliminary simulation results suggest that this approach provides a promising starting point for simulating both object recognition and the visual-cognitive interface.

Simulations demonstrated that the model can correctly recognize familiar objects (simulation 1) as well as new objects created by adding (simulation 2), deleting (simulation 3), and replacing (simulation 4) parts of familiar objects. All of these transformations pose problems for non-compositional (e.g. template-based) accounts of object recognition (Biederman, 1987), but they are commonplace in human interactions with objects. Parts are often deleted by occlusion or by modification of the physical object (e.g. as when a tire is removed from a car); added, as when new parts are added to objects to extend functional capabilities; or replaced (e.g., for styling reasons). These types of modifications are especially common in commercially designed objects, so our ability to recognize and reason about these objects depends on our ability to tolerate these types of modifications: The first time we see a new model of coffeemaker, we may decide that the styling is not to our liking, but we do not stare at it in confusion about what it is.

Crucially, the representations used by this model are not only useful for recognition, as shown by the simulations, but also lend themselves naturally to reasoning about the objects' function. In particular, these representations are already in "LISAese", the representational format used by the LISA model, and as such are available to the full inductive power of that inference engine. For example, given an object file describing a novel coffee maker, LISA is well-equipped to infer that the handle is where the pot should be grasped, the filter basket is where the ground coffee should be placed, and the carafe is where the brewed coffee will collect. Once the model is supplied with a JIM-like front-end, it should be in a position to start with object images and end with inferences about those objects.

Acknowledgments

This research was supported by AFOSR Grant AF-FA9550-12-1-003.

References

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*, *94* (2), 115-147.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*, 87–185.

Doumas. L. A. A., Hummel, J. E., & Sandhofer, C. M. (2008). A theory of the discovery and predication of relational concepts. *Psychological Review, 115*, 1 - 43.

Green, C. B., & Hummel, J. E. (2004). Relational perception and cognition: Implications for cognitive architecture and the perceptual-cognitive interface. In B. H. Ross (Ed.), *The psychology of learning and motivation*, Vol 44. (pp. 201-223). San Diego: Academic Press.

Hummel, J. E. (2000). Where view-based theories break down: The role of structure in shape perception and object recognition. In E. Dietrich & A. Markman (Eds.), *Cognitive dynamics: Conceptual change in humans and machines* (pp. 157 - 185). Mahwah, NJ: Erlbaum.

Hummel, J. E. (2001). Complementary solutions to the binding problem in vision: Implications for shape perception and object recognition. *Visual Cognition*, *8*, 489 - 517.

Hummel, J. E. (2013). Object recognition. In D. Reisberg (Ed.) *Oxford Handbook of Cognitive Psychology*, 32-46, Oxford, UK: Oxford University Press.

Hummel, J. E., & Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychological Review*, *99*, 480-517.

Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review*, *104*, 427-466.

Hummel, J. E., & Holyoak, K. J. (2001). A process model of human transitive inference. In M. Gattis (Ed.). *Spatial schemas in abstract thought* (pp. 279-305). Cambridge, MA: MIT Press.

Hummel, J. E., & Holyoak, K. J. (2003a). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review*, *110*, 220-264.

Hummel, J. E., & Holyoak, K. J. (2003b). Relational reasoning in a neurally-plausible cognitive architecture: An overview of the LISA project. *Cognitive Studies: Bulletin of the Japanese Cognitive Science Society, 10*, 58-75.

Hummel, J. E., & Stankiewicz, B. J. (1996). An architecture for rapid, hierarchical structural description. In T. Inui & J. McClelland (Eds.). *Attention and Performance XVI: Information Integration in Perception and Communication* (pp. 93-121). Cambridge, MA: MIT Press.

Hummel, J. E., & Stankiewicz, B. J. (1998). Two roles for attention in shape perception: A structural description model of visual scrutiny. *Visual Cognition*, *5*, 49-79.

Johnson-Laird, P. N. (1983). *Mental Models*. Cambridge, MA: Cambridge University Press.

Kahneman, D. & Treisman, A., & Gibbs, B. J (1992). The reviewing of object files: Object-specific integration of information. *Cognitive Psychology*, *24*, 175-219.

Knowlton, B. J., Morrison, R. G., Hummel, J. E., & Holyoak, K. J. (2012). A neurocomputational system for relational reasoning. *Trends in Cognitive Sciences*, *17*, 373-381.

Lovett, A., & Forbus, K. (2017). Modeling visual problem solving as analogical reasoning, *Psychological Review*, 124, 60-90.

Thoma V., Davidoff J. (2007) Object Recognition: Attention and Dual Routes. In: Osaka N., Rentschler I., Biederman I. (eds) *Object Recognition, Attention, and Action*. Springer, Tokyo.

Winston, P. (1975). Learning structural descriptions from examples. In P. Winston, *The Psychology of Computer Vision* (pp. 157-209). New York: McGraw-Hill.

Norms and the meaning of omissive enabling conditions

Paul Henne[1], Paul Bello[2], Sangeet Khemlani[2], and Felipe De Brigard[1]
paul.henne@duke.edu, paul.bello@nrl.navy.mil, sangeet.khemlani@nrl.navy.mil, felipe.debrigard@duke.edu
[1]Department of Philosophy, Duke University, Durham, NC 27708 USA
[2]Navy Center for Applied Research in Artificial Intelligence, Naval Research Laboratory, Washington, DC 20375 USA

Abstract

People often reason about omissions. One line of research shows that people can distinguish between the semantics of omissive causes and omissive enabling conditions: for instance, not flunking out of college enabled you (but didn't cause you) to graduate. Another line of work shows that people rely on the normative status of omissive events in inferring their causal role: if the outcome came about because the omission violated some norm, reasoners are more likely to select that omission as a cause. We designed a novel paradigm that tests how norms interact with the semantics of omissive enabling conditions. The paradigm concerns the circuitry of a mechanical device that plays music. Two experiments used the paradigm to stipulate norms and present a distinct set of possibilities to participants. Participants chose which causal verb best described the operations of the machine. The studies revealed that participants' responses are best predicted by their tendency to consider the semantics of omissive relations. In contrast, norms had little to no effect in participants' responses. We conclude by marshaling the evidence and considering what role norms may play in people's understanding of omissions.

Keywords: omissive causes; enabling; allowing; modal semantics; norms; mental models

Introduction

A railway gatekeeper's job is to open and close a crossing gate that lets trains pass. In 1902, the gatekeeper for the Somerset and Dorset railway was found guilty of manslaughter because he failed to close the gate (R. v. Pittwood, 1902). While he was at lunch, a train passed through the open gate and crashed into a horse and cart, killing one man and injuring another. The case describes an omissive cause: the jury held that the gatekeeper's failure to close the gate caused the death of an innocent bystander.

Omissive causation is a controversial topic amongst philosophers, psychologists, and legal scholars (Moore, 2009; Bernstein, 2015; Henne, Pinillos, & De Brigard, 2017). People have little difficulty in distinguishing which event was causal from alternative events that are non-causal. But causes are often easy to establish when they occur in a particular place and at a particular time; for instance, throwing a switch at a particular time causes the particular gate to close, so it is easy to identify the intervening action as the cause. Since omissive causes—absences, failures to act, scarcities, etc.—do not occur in any spatial or temporal frame, they present unique difficulties for causal reasoning and theories of causation.

One proposal suggests that norm violations affect causal judgments and play a fundamental role in establishing what constitutes a cause (Hitchcock & Knobe, 2009; McGrath, 2005; Hart & Honoré, 1985). In the railway example, the gatekeeper was charged and found guilty because his occupation made it his responsibility to monitor the track. It may seem trivial that many other individuals—for instance, some passerby—also failed to close the gate, but previous philosophical treatments have difficulty explaining why only certain omissive causes are deemed relevant and not others (see McGrath, 2005; Bernstein, 2015). On the norm-based account, the passerby, unlike the gatekeeper, is not considered a cause, as there was no normative expectation for him to close the gate. The norm-based account provides an explanation for why people focus their attention on potential causes. Consistent with this view, recent studies show that reasoners view norm-violating omissions as causes but norm-preserving events as non-causes or as enablers (Henne et al., 2017; see also Clarke et al., 2013).

Nevertheless, some theorists question whether norms determine the meaning of omissive causal statements or whether norms simply bias causal judgments (Bernstein, 2014; 2017, p. 89-90). Consider the following statement:

1. The drought caused the famine.

Some argue that omissive causal statements as in (1) do not involve norms in any way, yet they are easy to comprehend. If norms were a central part of the meanings of causal relations, then the absence of any norm should render (1) uninterpretable (Bernstein, 2017). On such a view, norms may be relevant in *establishing* causal relations—such as in the train example—but they are not central to their meaning.

One clue for what it means to be a cause comes from the application of causal verbs: "causes," "enables," and "prevents." Each verb refers to a relation between two events, and those relations have stark differences in their semantics. Psychological accounts of causal reasoning identify differences in the way people understand causal verbs (e.g., Goldvarg & Johnson-Laird, 2001; Sloman, Barbey, & Hotaling, 2009; Wolff, 2007). Accordingly, a viable theory of how people understand and infer omissive relations must distinguish the semantics between them. Consider the following two statements:

2a. An absence of light *causes* a flower to die.
b. An absence of light *enables* a flower to die.

(2a) seems sensible, but (2b) does not, because (2b) implies that the flower can live without light. Likewise, in the following two statements:

3a. A lack of insecticides *causes* insects to thrive.
b. A lack of insecticides *enables* insects to thrive.

(3b) seems sensible, but (3a) does not, because (3a) inappropriately guarantees that insects will thrive once insecticides are eliminated. The distinctions may be compelling, but until recently, no theory of causal reasoning could explain them.

A recent theory of omissive causation differentiates omissive causes from omissive enablers (Khemlani, Wasylyshyn, Briggs, & Bello, 2018). The theory is based on the idea that people represent causal scenarios by constructing and manipulating a set of discrete possibilities, i.e., mental models (Goldvarg & Johnson-Laird, 2001). The model-based theory—the "model theory," for short—posits that omissive causes and omissive enabling conditions differ in the sets of possibilities to which they refer. On tasks that require reasoners to distinguish between the different relations, they should base their judgments on the semantics stipulated by the model theory (Khemlani et al., 2018). In contrast, if norms are central to the meaning of omissive relations, reasoners should base their decisions on norm-violations (Henne et al., 2017).

In what follows, we first delineate the predictions of the model theory and the norms hypotheses. We then describe an experimental paradigm that can test between the two predictions, and we present two novel experiments that test the competing predictions. The studies showed that reasoners separated omissive causes from omissive enabling conditions in a manner predicted by the model theory, and norm-violations had little effect on their behavior.

The model theory of causal reasoning

The mental model theory of human reasoning proposes that humans reason based on representing sets of possibilities (Johnson-Laird, 2006). The meanings of spatial relations, temporal relations, and causal relations refer to the sets of possibilities consistent with each relation (Goodwin & Johnson-Laird, 2005; Khemlani, Barbey, & Johnson-Laird, 2014). The model theory posits two systems of reasoning: a fast, intuitive system of reasoning constructs a single initial possibility—the "mental model"—to represent one or more assertions. Reasoners can formulate inferences rapidly by scanning that initial possibility, but those inferences are prone to error, because causal relations can be consistent with several possibilities. Errors can be corrected through deliberation, which is a process by which reasoners iteratively construct and consider alternative possibilities.

Khemlani and colleagues recently extended the model theory to account for reasoning about omissive causation (Bello et al., 2017; Khemlani et al., 2018). Their account explains why people distinguish different omissive relations (Table 1). It appeals to the idea that people rapidly construct initial mental models, and then flesh out those initial models into "fully explicit" models. On this view, a mental model is a privileged, default possibility to which an omissive causal relation refers, whereas fully explicit models represent all the possibilities consistent with the modal semantics of the relation. The following diagram depicts the mental model of the omissive causal relation described in (2a):

```
¬ light        death
```

where '¬' denotes the symbol for negation (Khemlani, Orenes, & Johnson-Laird, 2012). Here, the lack of light is interpreted as a negated event, and it arranges the two events in the same chronological order in which they would occur. Hence, the model represents a single iconic possibility. When reasoners deliberate, they can consider all of the possibilities that accord with the modal semantics of omissive causation (Table 1). They can accordingly build fully explicit models of (2a), which are depicted in this diagram:

```
¬ light        death
  light        death
  light      ¬ death
```

where each row represents a separate possibility. The bolded row represents the mental model. The latter two possibilities show that if the flower receives light, it may die anyway (for some other reason), or it may not die at all. But the theory predicts that reasoners should be less likely to think of these latter two possibilities at the outset because most reasoners only construct and reason with the mental model.

The theory posits that the mental model of omissive enabling conditions is the same as the mental model of omissive causation. Hence, the model of (3b) above is:

```
¬ insecticide   thrive
```

It predicts that reasoners who draw conclusions on the basis of mental models should often conflate the two assertions (e.g., Wolff, 2007; Frosch & Johnson-Laird, 2011). When reasoners distinguish between omissive causes and enabling conditions, they should do so on the basis of their modal semantics, i.e., on the fully explicit models of the relations. The fully explicit models of an omissive enabling condition are depicted in the following diagram:

```
¬ insecticide   thrive
¬ insecticide ¬ thrive
  insecticide ¬ thrive
```

Unlike omissive causes, omissive enabling conditions are consistent with the possibility in which both the cause and the effect do not hold, i.e., the situation in which insecticides are administered and insects subsequently do not thrive. Omissive enabling relations typically prohibit the possibility in which the cause and the effect both hold (A and B), e.g., the insects thrive even when they are sprayed with insecticide. But in some situations, omissive enabling relations can take on a weaker meaning and permit that possibility, as in, "The failure to cut the grass enabled it to grow." The statement permits the possibility in which the grass is cut and it grows anyway (Table 1).

The model theory accordingly makes the following general hypothesis about semantics of omissive relations:

Semantics hypothesis: On tasks that require reasoners to distinguish between alternative causal relations, they should discriminate between omissive causes and omissive enabling conditions on the basis of the possibilities unique to each relation.

In contrast, reasoners are often susceptible to norm violations that affect their causal judgments (Henne et al., 2017). Hence, norm-based accounts posit the following hypothesis:

Norms hypothesis: When norms are available, reasoners distinguish between causal relations by focusing on those candidate events that violate norms. Events that violate norms should be considered causes, whereas those that do not violate norms should be rejected as causes.

In the next section, we describe a novel paradigm developed to test between the two hypotheses.

A paradigm for testing semantics and norms

Many existing paradigms test the meanings of omissive causes, but they do not typically encourage reasoners to consider and track multiple possibilities that are thought to be essential to the meanings (Henne et al., 2016; Khemlani et al., 2018; Wolff, Barbey, & Hausknecht, 2010; cf., Bello et al., 2017; Experiment 2). To try to overcome this limitation, we developed a novel paradigm that could be used to investigate how people distinguish the meanings of causal relations.

The paradigm made use of diagrams akin to those shown in Table 2. The basic diagram depicts a machine with a speaker, a red battery, a safety switch (which appeared green or else black), a blue wire, and an unnamed yellow component. The diagram could vary in several ways in order to depict different possibilities. For instance, the speaker could be playing or not playing (depicted as a series of soundwaves or not); the blue wire could be connected to the red battery, or else not connected; and the safety is green when it is *on* and black when it is *off*. The safety switch allowed for the establishment of a norm: participants were taught that whenever the safety is on (colored green), the blue wire *is not supposed to touch* the battery. And when the safety is off (colored black), the blue wire *is supposed to touch* the battery. Hence, the machine could be depicted in 2 (sound or no sound) x 2 (wire connected or disconnected) x 2 (safety on or off) = 8 different configurations. Depending on the particular condition in the experiments, those various configurations either violated or preserved norms. And those various configurations were either compatible with a particular omissive causal relation or incompatible with it.

The paradigm allows to directly compare predictions made by both the semantics hypothesis and the norms hypothesis. Experiments 1 and 2 provided participants a single trial in which they received three different diagrams (Table 2). After studying three different diagrams, participants were given a sentence completion task that tested their understanding of the scenario depicted.

Experiment 1

Experiment 1 presented participants with three separate diagrams. Participants only received diagrams that were either possible or impossible given causal and enabling relations: diagrams depicting context-dependent contingencies were not used in the experiment (Tables 1 and 2). Half the participants saw a set of diagrams in which each diagram was compatible with the following omissive enabling relation:

4. The blue wire not touching the red battery *allows* the speaker to play music.

The other half saw a set of diagrams that were compatible with (4) and the following causal assertion (5):

5. The blue wire not touching the red battery *causes* the speaker to play music.

Because the experiment avoided context-dependent contingencies, the remaining diagrams compatible with (5) were also compatible with (4), so the set of diagrams were ambiguous: the model theory predicts that participants should treat them as depicting both omissive causes and omissive enabling conditions (Khemlani et al., 2018, Experiment 4).

Participants were given a sentence completion task in which they chose the causal verb ("causes" or "allows") to complete the following sentence:

> The blue wire not touching the battery _____ the speaker to play music.

Notably, for the purposes of the study, the verb "allows" was treated as equivalent to "enables". The experiment accordingly tested the prediction of the semantics hypothesis that reasoners should select the causal verb that matched the possibilities depicted. Hence, the semantics hypothesis predicts that reasoners should select "allows" more often for the enabling condition than the ambiguous condition. Experiment 1 also tested the norms hypothesis. Half the conditions in the study concerned abnormal situations in which the safety was off, and the other half concerned normal situations in which the safety was on, and participants were instructed that the blue wire is not supposed to touch the battery when the safety was on, and that it was supposed to touch the battery when the safety was off. The norms hypothesis predicts that people should be sensitive to norm violations, i.e., they should be more likely to select the verb "causes" to fill in the sentence provided for abnormal omissions compared to normal ones. In turn, normal omissions should be judged to be involved in enabling relations. The semantic hypothesis, however, posits that participants' responses should not vary as a function of whether the condition was abnormal or normal—only as a function of which set of possibilities participants considered.

	The four possible contingencies between A and B			
Omissive relation	¬A B	**¬A ¬B**	A ¬B	A B
The lack of A causes B.	Mental model	**Impossible**	Possible	Context-dependent
The lack of A enables B.	Mental model	**Possible**	Possible	Context-dependent

Table 1: The table outlines the semantics of omissive causes and omissive enabling conditions. The rows separate omissive causes and omissive enabling conditions. The cells in each column describe whether each contingency is possible given a particular omissive relation. The mental models are always possible. The bolded column denotes the contingency diagnostic of omissive enabling conditions.

Methods

Participants A total of 822 adults participated in this study on Amazon Mechanical Turk (AMT). Of these participants, 59 did not complete the study, and 26 were excluded for failing to pass two attention checks. Data were analyzed with the remaining 796 participants (M_{age} = 34, SD = 11.0, age range = [18-71], 43% females).

Design and procedure Participants were randomly assigned to one of four possible conditions in a 2 (enabling vs. ambiguous) x 2 (normal vs. abnormal) between-participants design. Participants were acquainted with the machine in Figure 1 and its various components. After viewing and responding to instructions, participants were presented with three diagrams of configurations of the machine. The diagrams appeared on the screen simultaneously. To check their comprehension of the machine and the three separate possibilities, they matched the possibilities with descriptions provided in a dropdown menu, i.e., they chose from the following options to describe each of the three diagrams: (1) "The blue wire touches the battery, and the speaker plays music," (2) "The blue wire touches the battery, and the speaker does not play music," (3) "The blue wire does not touch the battery, and the speaker plays music," and (4) "The blue wire does not touch the battery, and the speaker does not play music." The order in which the possibilities were presented was fixed. In the normal condition, the machine's safety was green, so the blue wire was not supposed to touch the battery. The blue wire not touching the battery is normal. In the abnormal condition, the machine's safety was black, so the blue wire is supposed to touch the battery. The blue wire not touching the battery is abnormal. Participants were explicitly instructed to attend to the color of the safety and what the blue wire was supposed to do. They were then asked to think back to their observations and then fill in the verb in the sentence: "The blue wire not touching the battery _____ the speaker to play music." Participants could choose between the verb "causes" or "allows" from a drop-down menu, and they could not proceed until a choice was made.

Post-experimental questionnaire Participants filled out a post-experimental questionnaire that asked them if they had paid attention and if they had taken the survey multiple times. Participants who reported affirmatively on either question were excluded.

Results and discussion

Figure 1 shows the proportion of participants who chose "allows" as a function of whether the condition was normal or not and as a function of whether the diagrams were consistent with the semantics for omissive enabling conditions or else ambiguous. Participants chose "allows" more often for diagrams consistent with enabling conditions than for ambiguous diagrams (74% vs. 54%; Mann-Whitney test, $z = 5.65$, $p < .0001$, Cliff's $\delta = .19$). Participants selected "allows" more often when the diagrams were presented in a normal rather than an abnormal context—although this result was not statistically significant (68% vs. 62%; Mann-Whitney test, $z = 1.72$, $p = .09$, Cliff's $\delta = .06$). A follow-up generalized logistical mixed-model (GLMM) regression further revealed that the difference in selection between the two conditions was inconsistent with a significant effect ($\beta = .00$, $p = .97$), as was the interaction between the two conditions ($\beta = .48$, $p = .12$). Nevertheless, a planned comparison revealed that for ambiguous diagrams, participants selected "allows" more often when the diagrams were presented in a normal context rather than an abnormal context (61% vs. 49%, Mann-Whitney test, $z = 2.35$, $p = .02$,

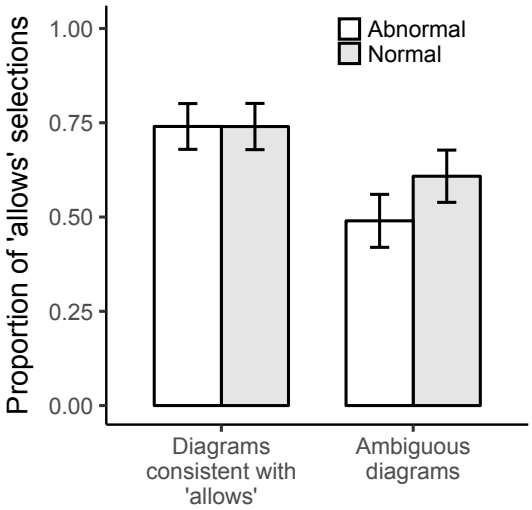


Figure 1: Proportion of participants who chose "allows" instead of "causes" in Experiment 1 as a function of whether participants saw normal or abnormal devices, and as a function of whether the diagrams were consistent with omissive enabling conditions only or consistent with both omissive causes and omissive enabling conditions. Error bars indicate 95% confidence intervals.

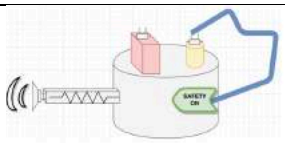
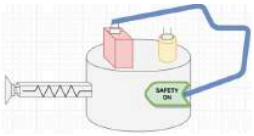
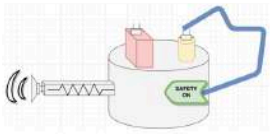
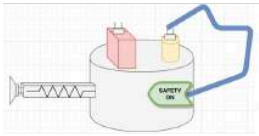
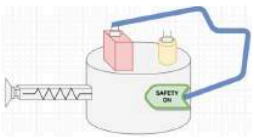
Condition	Diagrams presented to participants			
	¬wire music	¬wire ¬music	wire ¬music	wire music
Ambiguous	 (shown twice)	Not shown to participants		Not shown to participants
Enabling				Not shown to participants

Table 2: The three diagrams presented to participants in enabling and the ambiguous conditions. These diagrams all depict the normal conditions, i.e., the safety is on, so the blue wire is not supposed to touch the red battery. In Experiment 1, participants were not provided with written cues about the condition of the safety switch. In Experiment 2, those cues were provided (as in the diagrams above).

Cliff's δ = .12). All data and analysis code available at https://osf.io/jf36w/. The result provides some support for the norm hypothesis, which predicts that people should be more likely to select "causes" (and less likely to select "allows") for abnormal contexts. It also suggests that participants were sensitive to the norm manipulation: they comprehended the norms and took them into account in making their selections. If they had not, they would have shown no sensitivity to whether the diagrams were in normal or abnormal context. Yet, an analogous comparison for diagrams consistent with omissive enabling conditions was not reliable, and so the study revealed mixed support for the norms hypothesis.

Experiment 1 corroborated the prediction that reasoners interpret omissive causes and enabling conditions in accordance with the semantics outlined by the model theory. Reasoners in the enabling condition selected "allows" more often than those in the ambiguous condition. Moreover, participants' responses did not depend on whether a norm had been violated or not. If, as the norms hypothesis states, abnormal situations help reasoners choose which candidate events constitute causes, then those abnormalities appeared to have no effect on participants' tendencies to select appropriate causal relations.

One limitation of Experiment 1 is that reasoners may have simply failed to recognize abnormalities in the first place, i.e., they may not have encoded the black safety switch's color, which was designed to serve as a cue that a norm had been violated. Another limitation of the study is that participants evaluated only one set of three diagrams and only one causal relation. Experiment 2 corrected for both of these limitations.

Experiment 2

Because participants may not have picked up on the norm distinction between conditions in Experiment 1, we sought to ensure that the difference was salient in Experiment 2. Hence, rather than just identifying the color as the difference in norms, Experiment 2 added the verbal cues "SAFETY ON" and "SAFETY OFF" to the diagrams (see Table 2). Moreover, the study employed a within-participants design to further validate the findings from Experiment 1 supporting the semantics hypothesis. Hence, each participant saw four distinct sets of three diagrams.

Methods

Participants A total of 215 adults participated in this study on AMT. Of these, 21 participants were excluded for failing to pass two attention checks. Data were analyzed with the remaining 194 participants (M_{age} = 33.82, SD = 9.39, age range [18-68] 40% females).

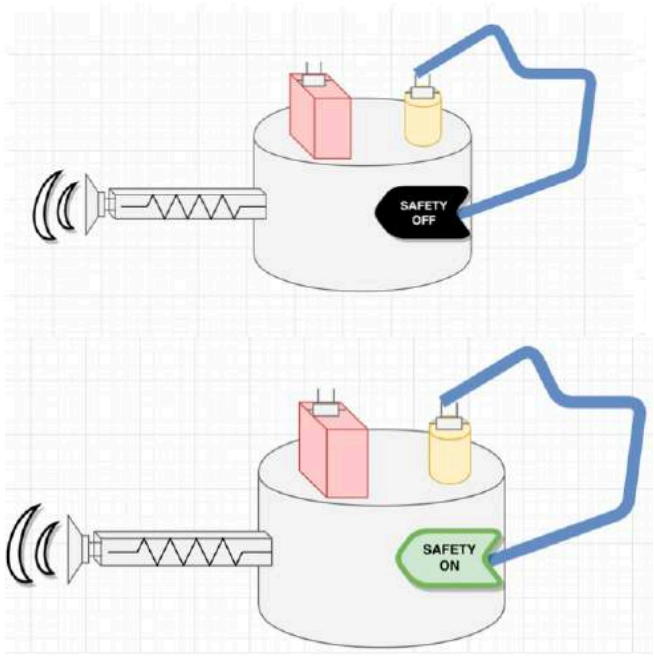


Figure 2: The stimuli used in Experiment 2. Participants in the abnormal condition saw the machine with the black safety i.e., in the off position (top diagram), and those in the normal condition saw the machine with the green safety, i.e., in the on position (bottom diagram).

Design and procedure Participants acted as their own controls and received all four possible conditions in a 2 (enabling vs. ambiguous) x 2 (normal vs. abnormal) within-participants design. As in Experiment 1, participants were acquainted with the machine and its various components. The model conditions were constructed just as they were in Experiment 1 (see Table 2). The stimuli were modified such that the normal and abnormal conditions were more salient by adding the words "SAFETY ON" and "SAFETY OFF" to the diagrams (Figure 2).

Results and discussion

Figure 3 shows the proportion of participants who chose "allows" as a function of whether the condition was normal or not and as a function of whether the diagrams were consistent with the semantics for omissive enabling conditions or else ambiguous. As in Experiment 1, participants selected "allows" more often when the diagrams depicted possibilities uniquely consistent with omissive allowing relations rather than ambiguous possibilities (77% vs. 53%; Mann-Whitney test, $z = 7.00$, $p < .0001$, Cliff's $\delta = .24$). They didn't reliably select "allows" more often for diagrams in an abnormal vs. a normal context (66% vs. 64%; Mann-Whitney test, $z = .83$, $p = .41$, Cliff's $\delta = .03$). GLMM regression analyses likewise corroborated the nonparametric analyses: it yielded an effect of whether the diagrams were ambiguous or consistent with omissive allowing conditions ($B = 1.14$, $p < .0001$), but no effect of normality ($B = .15$, $p = .54$) and no interaction ($B = .02$, $p = .95$).

A planned comparison revealed that for ambiguous diagrams, participants did not reliably select "allows" more often for abnormal than normal contexts (55% vs. 52%; Mann-Whitney test, $z = .61$, $p = .54$, Cliff's $\delta = .03$). Hence, overall, the results are not consistent with the norms hypothesis, which states that reasoners should be more likely to select "causes" (and less likely to select "allows") for abnormal vs. normal contexts.

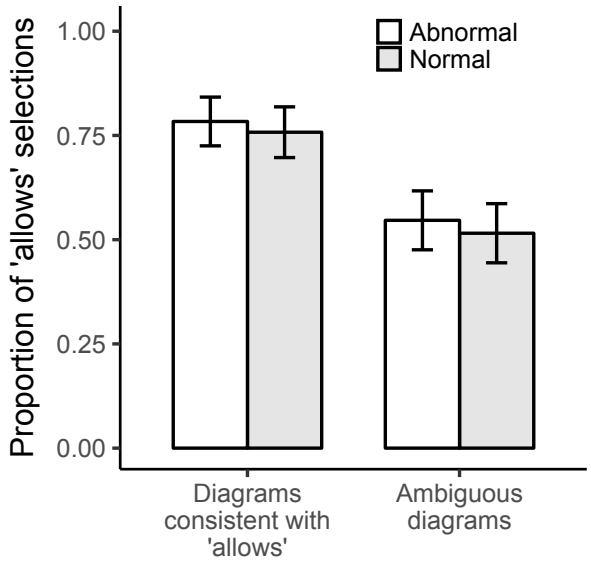


Figure 3: Proportion of participants who chose "allows" instead of "causes" in Experiment 2 as a function of whether participants saw normal or abnormal devices, and as a function of whether the diagrams were consistent with omissive enabling conditions only or ambiguous. Error bars indicate 95% confidence intervals.

Just as in Experiment 1, the results of Experiment 2 support the prediction of the model theory with respect to its predictions about the semantics of omissive enabling conditions. Reasoners in the enabling condition selected "allows" more often those in the ambiguous condition. Moreover, participants' responses did not depend on whether a norm had been violated or not.

General Discussion

Two experiments were designed to test how participants judge the causal effect omissive events have on outcomes. The experiments corroborated a recent theory of omissive causation, which predicts that reasoners should be able to distinguish omissive enabling conditions from other sorts of omissive relation (Khemlani et al., 2018). Moreover, the results showed that norm violations cannot explain the semantic difference between causes and enabling conditions.

The results of these studies can help refine the role that norms play in causal reasoning. As Henne and colleagues (2017) show, norms help select potential causes and distinguish them from irrelevant non-causes. The present studies, however, show that norms do not always explain the difference between causes and enablers. When reasoners consider the distinctive possibilities consistent with enabling conditions norms have little to no effect on causal judgment. When reasoners consider only the mental model, norms may have a more prominent effect on causal judgment. A more robust extension of the model theory, i.e., one that explains how norms are represented and how they modulate possibilities could potentially explain both the semantics between different causal verbs as well as how reasoners isolate potential causes from non-causes. Such a theory would also have to be contrasted with recent models of causal strength that could potentially explain the norm effects and the results predicted by the model theory (Icard, Kominsky, & Knobe, 2018).

Acknowledgments

This project was supported by an Office of Naval Research award (N00014-17-1-2603) to Felipe De Brigard.

References

Bello, P., Wasylyshyn, C., Briggs, G., & Khemlani, S. (2017). Contrasts in reasoning about omissions. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Bernstein, S. (2014). Omissions as possibilities. *Philosophical Studies*, *167*.

Bernstein, S. (2015). The metaphysics of omissions *Philosophy Compass*, *10*, 208-218.

Bernstein, S. (2017). Intuitions and the metaphysics of causation. *Experimental Metaphysics*, 75.

Clarke, R. et al. (2013). Causation, norms, and omissions: A study of causal judgments. *Philosophical Psychology, 28,* 279-93.

Frosch, C.A., & Johnson-Laird, P.N. (2011). Is everyday causation deterministic or probabilistic? *Acta Psychologica*, *137*, 280–291.

Goldvarg, E., & Johnson-Laird, P. (2001). Naïve causality: A mental model theory of causal meaning and reasoning. *Cognitive Science*, *25*, 565-610.

Goodwin, G.P., & Johnson-Laird, P.N. (2005). Reasoning about relations. *Psychological Review*, *112*, 468-493.

Halpern, J.Y., & Hitchcock, C. (2014). Graded causation and defaults. *British Journal for the Philosophy of Science, 66,* 413-457.

Hart, H.L.A., & Honoré, T. (1985). *Causation in the Law*, 2nd ed., Oxford: Clarendon.

Henne, P., Pinillos, Á., & De Brigard, F. (2017): Cause by omission and norm: Not Watering Plants. *Australasian Journal of Philosophy*.

Icard, T. F., Kominsky, J. F., & Knobe, J. (2017). Normality and actual causal strength. *Cognition*, *161*, 80-93.

Johnson-Laird, P.N. (2006). *How we reason*. NY: Oxford University Press.

Johnson-Laird, P. N., & Khemlani, S. (2017). Mental models and causation. In M. Waldmann (Ed.), *Oxford Handbook of Causal Reasoning*. Oxford: Oxford University Press.

Johnson-Laird, P. N., Khemlani, S., & Goodwin, G.P. (2015). Logic, probability, and human reasoning. *Trends in Cognitive Sciences*, 19.

Khemlani, S., Barbey, A., & Johnson-Laird, P. N. (2014). Causal reasoning with mental models. *Frontiers in Human Neuroscience, 8,* 849, 1-15.

Khemlani, S., Orenes, I., & Johnson-Laird, P.N. (2012). Negation: a theory of its meaning, representation, and use. *Journal of Cognitive Psychology*, *24*.

Khemlani, S., Wasylyshyn, C., Briggs, G., & Bello, P. (2018). Mental models and omissive causation. *Memory & Cognition*, 46, 1344-1359.

McGrath, S. (2005). Causation by omission: A dilemma. *Philosophical Studies*, *123*.

Wolff, P. (2007). Representing causation. *Journal of Experimental Psychology: General*, *136*, 82-111.

Wolff, P., Barbey, A., & Hausknecht, A. (2010). For want of a nail: How absences cause events. *Journal of Experimental Psychology: General, 139*.

Grammatical Generalisation in Statistical Learning: Is it Implicit and Invariant Across development?

Amanda J. Hickey (ajh650@york.ac.uk)
Department of Psychology, University of York, York, YO10 5DD

Marianna E. Hayiou-Thomas (emma.hayiou-thomas@york.ac.uk)
Department of Psychology, University of York , York, YO10 5DD

Jelena Mirković (jelena.mirkovic@york.ac.uk)
School of Psychological and Social Sciences, York St John University, Lord Mayor's Walk, York, YO31 7EX
Department of Psychology, University of York , York, YO10 5DD

Abstract

The learning and generalisation of grammatical regularities is fundamental to successful language acquisition and use. Research into statistical learning has started to consider how this process occurs through the implicit detection and assimilation of grammatical regularities. This study focuses on how adults and children generalise regularities and explores the role of explicit knowledge in this process. Across three experiments, adults and children learnt an artificial language containing two semantic categories denoted by a co-occurring determiner and suffix. Explicit knowledge of the regularities was associated with generalisation performance in adults but not children, even when adult word level knowledge was similar to children's. The implications of these results for developmental theories of grammatical generalisation are discussed.

Keywords: statistical learning, explicit knowledge, grammatical categories, artificial language, learning, generalisation.

Introduction

A key aspect of language acquisition and use is the ability to learn grammatical regularities that are present in the input, and generalise them to novel situations. Statistical learning (SL) has been suggested as one of the key mechanisms for the acquisition of grammatical regularities (Gómez & Gerken, 2000). For example, corpus studies of child directed speech have shown the presence of two statistical cues, which reliably indicate grammatical categories: phonological and distributional cues (e.g. Monaghan et al., 2005). Phonological cues consist of speech sounds which are associated with word class (e.g. in English, phoneme length can indicate a noun or a verb), and can be at the word, syllable or phoneme level. Distributional cues relate to the linguistic context within which a word usually sits, for instance where two co-occurring words (or morphemes) frame an interleaved word stem (e.g. '…is walking; Monaghan et al., 2005). Both adults and children use these cues in natural language learning and processing (e.g. Farmer et al., 2006; Lew-William & Fernald 2007; 2010). Research using artificial languages incorporating these types of cues has also shown that adults and children are able to utilise them when learning grammatical categories (e.g. Lany & Saffran, 2010; 2011; Mirković, Forest & Gaskell, 2011; Mirković & Gaskell, 2016; Frost, Monaghan & Christiansen, 2019). For example, Gómez (2002) found that both young infants (17-19 months old) and adults were able to detect and learn distributional cues and Hall, Horne and Farmer (2018) demonstrated the use of distributional cues in the learning and generalisation of grammatical categories in older children (6-9 years old).

The role of semantic cues has also been assessed. For example, in 20-month-old infants, the learning of semantic cues was supported by deterministic phonological and distributional cues for grammatical categories (Lany and Saffran, 2011), and to a lesser extent by probabilistic mappings between semantics and distributional cues (Lany, 2014). More recently, distributional cues have also been shown to enhance the learning of word-referent mappings in adults (Frost, Monaghan & Christiansen, 2019). Adults have also been shown to use semantic cues to generalise grammatical gender-like classes to previously unseen items, in a probabilistic artificial language (Mirkovic et al., 2011; Mirkovic & Gaskell, 2016).

In sum, the research on both adults and children and infants demonstrates that they can use SL to learn grammatical categories from statistical cues. Although not all studies assess generalisation of newly formed grammatical knowledge to previously unseen items, those that do show that both adults (e.g. Mirković et al., 2011) and children and infants (e.g. Lany & Saffran, 2010; 2011; Wonnacott et al., 2012) are able to do so. However, an important open question concerns the processes that support successful generalisation, and whether these processes differ in adults and children.

It is typically assumed that SL is an implicit (unconscious) process that is invariant across different ages (e.g. Aslin & Newport, 2012). However, more recent studies (Batterink et al., 2015; Franco et al, 2011; Conway & Christiansen, 2005) suggest that both implicit and explicit processes play a role in adult statistical learning. By drawing parallels between the implicit learning literature (e.g. Reber, 1967) and SL, these authors consider how 'implicit' implicit learning tasks really are in the context of SL.

To test the relative roles of implicit and explicit processes in SL, Batterink et al., (2015) incorporated on-line measures of implicit learning (reaction times and ERPs) into a word boundary SL task with adults (based on the paradigm used by Saffran et al., 1996). The results suggested that both implicit and explicit processes were involved in the detection of word boundaries. In a similar vein, Smalle et al. (2017) used a Hebb sequence-learning paradigm to examine the relative role of explicit/implicit processing in children as compared to adults. Although both adults and children showed evidence of explicit awareness of the learned sequences, there were some notable developmental differences: adults' explicit awareness emerged at an earlier point during learning than that of children. Furthermore, while explicit awareness was significantly associated with Hebb learning performance in adults, this association was not present in children. This suggests that adults were drawing on both explicit and implicit learning mechanisms during this task, while children relied on implicit learning (Smalle et al., 2017). These studies suggest that both explicit and implicit processes are involved in SL and that the relative contributions of these processes may differ between adults and children.

Current Study

The key aim of the current study was to examine the role of explicit knowledge in grammatical category generalisation, and whether this differs in children and adults. Across three experiments, participants were trained on an artificial language using phonological, distributional and semantic cues to create a grammatical gender-like noun class system (Mirkovic et al., 2011). We tested adult and school-aged child participants and examined the role of explicit knowledge in the generalisation of grammatical regularities to previously unseen items. We manipulated the type of training and the level of initial learning of the novel nouns.

The artificial language consisted of the noun "classes" based on semantic, phonological, and distributional cues. To create the semantic cues, two semantic categories were used: animals and artefacts. The phonological cues were incorporated using a "suffix" (e.g. mof*eem*). The distributional cues were incorporated as a co-occurrence of a "determiner" and a "suffix" (see Table 1 for examples). Each determiner and each suffix was paired with a semantic category (animals or artefacts). This provided an *a*X*b* structure for animals and *c*X*d* structure for artefacts, with X denoting the interleaving arbitrary stem, *a* and *c* the determiner and the *b* and *d* the suffix.

Across all studies participants were trained using a word learning task (with no reference to underpinning 'grammatical' regularities. After training, they were tested on three generalisation tasks focusing on the three different cues (explained below). Levels of emergent explicit knowledge were assessed at the end of the experiment. Experiment 1 included adults and children, while Experiments 2 and 3 included adults only. Across the three experiments, we manipulated two factors that we hypothesised would contribute to generalisation and the emergence of explicit knowledge: i) initial levels of word learning, and ii) type of training. In Experiment 1, participants were exposed to a fixed number of repetitions of the novel words at training using a word-picture matching task (WPM; Breitenstein et al., 2007), and a word repetition task. Experiments 2 and 3 used criterion learning, with adult participants matched in the level of initial word learning to the children in Experiment 1. We hypothesized that levels of initial word learning may influence generalisation performance and levels of explicit knowledge of the 'grammatical' regularities. In addition, we removed word-picture matching from the training procedure in Experiment 3, to test the hypothesis that explicit selection may contribute to the emergence of explicit knowledge of the 'grammatical' regularities. In all three experiments, we examined the extent to which generalisation performance was associated with the emergent explicit knowledge of the phonological, distributional, and semantic cues.

Table 1: Design of the noun classes

	Determiner	Suffix	Examples
animal	tib	eem	*tib* mof*eem* = dog *tib* zeap*eem* = duck
artefact	ked	ool	*ked* larsh*ool* = table *ked* snar*ool*=TV

Method

Participants

Experiment 1. Sixty-one participants took part: 31 adults with a mean age of 19.70 years (19.08-20.67 years; 1 male) and 30 children with a mean age of 10.21 years (9.67-10.82 years; 13 males). The adult sample was drawn from the undergraduate population at the University of York and received course credits for their participation. The child sample was drawn from primary schools in North Yorkshire.

Experiments 2 & 3. Thirty participants took part in Experiment 2 with a mean age of 20.77 years (18.17-32.58 years; 4 male), and thirty in Experiment 3 with a mean age of 21.09 years (18.25-31.58 years; 5 males). These two samples were drawn from the undergraduate population at the University of York and received course credits or payment for their participation.

Stimuli The training and testing tasks in all experiments used pictures drawn from Rossion and Purtois (2001) object database (281x173ppi) and artificial words created from the English database of pronounceable nonwords (Rastle et al., 2002). The artificial 'words' were constructed using the three elements described earlier (e.g. *a*X*b*) and were digitally recorded (produced by a native speaker of English). This process was based on the stimuli created by Mirković et al., (2011).

All arbitrary stem (X) elements consisted of one syllable with a CVC, CCVC or CVCC (C= Consonant, V = Vowel; 'CAT' = CVC) structure. An overall balance of CVC, CCVC

and CVCC words between the animal and artefact training words was controlled for. The stem onset phoneme did not match the onset phoneme of the English word for the animal/artefact it was paired with. The same training and generalisation sets were used in all three experiments.

Training Set: Thirty-two word-picture pairs were created (16 in each semantic category). Each word was paired with a picture, which denoted the assigned meaning of the word, providing the non-arbitrary semantic cue (see Table 1 for examples).

Generalisation Sets: Three different sets of 8 generalisation items were designed to test post-training performance on previously unseen items. Each set consisted of 4 items that were consistent with the trained regularities, and 4 items that were inconsistent. Higher endorsement rates for consistent vs inconsistent items was taken to indicate learning of the regularities.

Determiner and Suffix Generalisation. This task was designed to test learning of the mapping between the determiner and suffix, and the associated semantic category. Eight novel words were presented with novel picture pairings. The four consistent items conformed entirely to the regularities present in the training set. In the four inconsistent items, the structure of the word conformed to the training set (e.g. *tib darleem*), but it was presented with a picture from the 'wrong' semantic category (e.g. *tib darleem* was paired with an artefact, instead of an animal).

Suffix-Only Generalisation. This task tested learning of the co-occurrence between the semantic category and the suffix specifically; that is, the 'phonological' cue. The 8 novel words were presented with novel picture pairings; as before, the 4 consistent items conformed to the regularities in the training set. In the inconsistent items, the determiner 'matched' the picture, but the suffix did not match either the determiner or the picture (e.g.*tib* sen*ool* was paired with a picture of a goat; where the co-occurrence of 'tib' with the picture of an animal conformed to the training set, but the suffix 'ool' was inconsistent with both the determiner 'tib' and the semantic category of animal).

Phonological Form Generalisation. This task specifically tested learning of the co-occurrence of the determiner and suffix; that is, the 'distributional' cue. Eight novel words were presented without pictures. The 4 consistent items conformed to the regularities used in the training set. The 4 inconsistent items had a mismatch between the determiner and the suffix (e.g. *tib* jit*ool* and *ked* narp*eem*).

Procedure Participants completed all tasks in one session of approximately 40-60 minutes. Responses were recorded by the 'DMDX' programme (Forster & Forster, 2003) on a PC laptop computer. Participants were introduced to experimental tasks as a series of games involving 'alien' words introduced by a visiting extra-terrestrial. The training procedure varied across the three experiments, but they all used the same testing protocol.

Experiment 1 training:

Repetition: The thirty-two training stimuli were presented once within a block, for three blocks. Participants were instructed to look at the picture and listen to the 'alien' word and repeat the word aloud once. Participants completed this task twice.

WPM: Participants were presented with word-picture pairs and were instructed to judge if they thought the word and picture '*went well together'*. Participants were exposed to all 32 word-picture pairs once. In addition, 16 of the word items were presented again paired with a different picture from the same semantic category (mismatch trials) for the 'incorrect' response. The participant responded using keys on the computer keyboard: a "happy face" if they thought the picture and word went well together and a "sad face" if they did not. Participants completed this task twice.

Experiment 2 training:

Repetition and WPM: For this experiment, the repetition and WPM tasks were merged. In each block, participants were exposed to all 32 word-picture pairs once. In addition, 8 mismatch trials were included, in which the word items were paired with an incorrect picture from the same semantic category. Participants were instructed to look at the picture and listen to the 'alien' word and repeat it aloud once. They then pressed the space bar and then judged if the word and picture '*went well together'* using the same WPM response procedure from Experiment 1.

Each training block was followed by a word-learning test (described below). Training ended when the participant reached the same level of accuracy as that of the children in Experiment 1 (75%).

Experiment 3 training:

Repetition Only: The training set for Experiment 3 was the same as that for Experiment 2, including criterion learning. However, the training procedure was different in that it did not include WPM: participants had only to repeat the training items.

All Experiments: Testing

Word Learning –Two Alternative Forced Choice (2AFC): This task tested learning of the novel words. Each word was randomly presented once and was accompanied by the simultaneous presentation of two pictures (on either side of the screen), one of which was the correct trained picture. The 'foil' picture was drawn from the trained pictures and was from the same semantic category. Participants responded using keys on the computer keyboard which corresponded to the on-screen picture presentation position.

Generalisation: "Determiner and Suffix" and "Suffix Only" Generalisation. In both these tasks, participants were instructed to attend to 'alien*'* word and picture pairings (from the respective generalisation sets) and judge if they thought they '*went well together',* pressing the happy or sad face accordingly.

"Phonological Form": Participants were instructed to listen to the 'alien' words from the generalisation set and asked to judge if the words '*went well with'* the 'alien' language they

had been listening to, pressing the happy or sad face accordingly.

Explicit Knowledge Questionnaire: Once all tasks were completed, participants were asked *'Did you notice anything about the alien language? Did you use any kind of strategies or clues to decide whether the word and the picture matched?'* Answers were recorded manually and a score from 0-3 was given separately for determiner and suffix knowledge: 0 for no reference to the morpheme or semantic dependency, 1 for knowledge of the morpheme but not the dependent semantic cue, 2 for partial knowledge of the morpheme and semantic dependency and 3 for full knowledge.

Results and Discussion

Word Learning: We examined the level of word learning in the three experiments by analysing performance on the 2AFC task at the end of training. One-sample t-tests against chance (.5) showed that all groups learned the novel words. (Experiment 1 adults, $t(30)=28.93$, $p<.001$; children, $t(29)=9.10$, $p<.001$; Experiment 2 adults, $t(29)=21.17$, $p<.001$; Experiment 3 adults, $t(29)=21.83$, $p<.001$; Figure 1). Adult participants in Experiments 2 and 3 were trained to the criterion matching the levels of child word learning in Experiment 1. To confirm that the word learning across the three studies matched as intended, we ran two multiple regressions with 2AFC performance as the outcome variable and group as the predictor variable coded using Helmert contrasts. The first set of contrasts showed that, as expected, adults learned more words in Experiment 1 (Adults1) than in Experiments 2 (Adults2) and 3 (Adults3; $\beta=0.68$, $p<.001$), and that there was no difference between the latter two ($p=.293$). The second set of contrasts showed that Adults2 ($p=.341$) and Adults3 ($p=.757$) learnt an equivalent number of words to the children in Experiment 1 (Children1). These findings show that all participants demonstrated word learning. Crucially, these results indicate that criterion learning method used in Experiments 2 and 3 was successful at reducing the level of adult participants' word-learning to that of the child participants in Experiment 1.

Generalisation Tasks To analyse performance in the generalisation tasks, we derived an A' metric based on the endorsement rates for consistent and inconsistent trials (Pallier, 2002). A' scores above 0.5 were taken as indication that a participant could reliably endorse consistent trials more often than inconsistent trials, demonstrating learning of the regularities.

"Determiner and Suffix": Figure 2 shows levels of generalisation performance for all groups on this task. One-sample t-tests showed that only Adults1 performed significantly above 0.5 ($t(30)=6.13$, $p<.001$). Thus, only this group demonstrated learning and generalisation of the mapping between the determiner and suffix, and the semantic category.

Using the same set of contrasts as in the analysis of word learning, with A' performance as the outcome variable and group contrasts as the predictor variables, group comparisons further confirmed that Adults1 were significantly better at generalising this regularity than Adults2 and Adults3 ($\beta=0.30$, $p<.004$). There was no difference between the Adults2 and Adults3 ($p=.703$), nor between Children1 and Adults2 ($p=.451$) or Adults3 ($p=.916$).

"Suffix Only": Figure 3 shows generalisation performance for all groups on this task. One-sample t-tests demonstrated that Adults1 ($t(30)=3.45$, $p<.001$) and Adults2 ($t(29)=1.97$, $p=.029$) performed significantly above an A' of 0.5. Children1 ($p=.762$) and Adults3 ($p=.500$) did not. Therefore, only Adults1 and Adults2 showed learning and generalisation of the mapping between the semantic category and the suffix (the phonological cue).

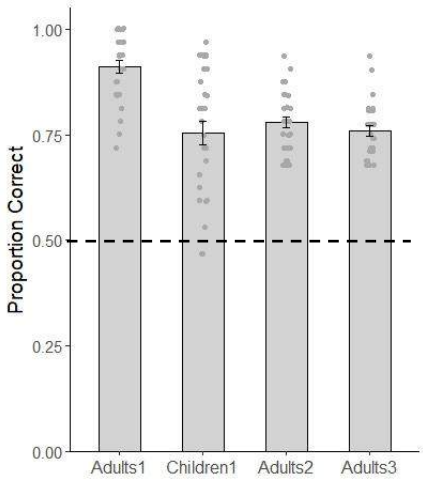


Figure 1: Word Learning: Accuracy on the 2AFC task at the end of training.

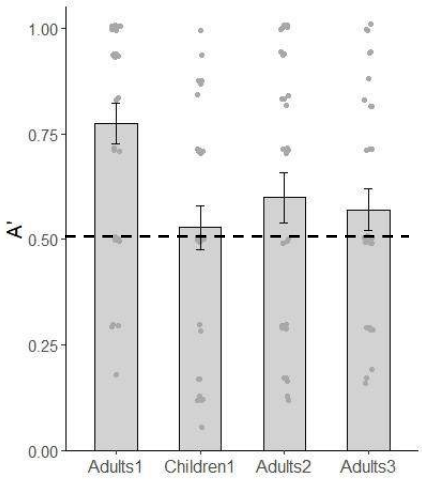


Figure 2: Performance on the "Determiner & Suffix" Generalisation Task

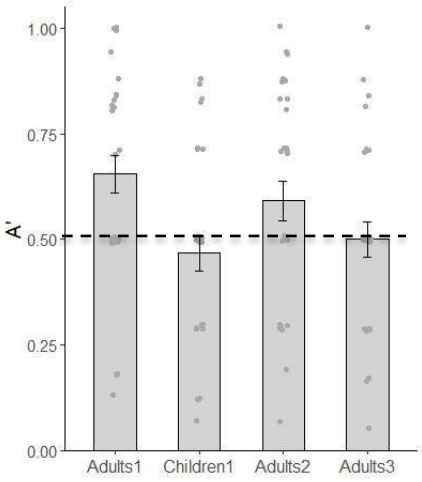


Figure 3: Performance on the Suffix Only Generalisation Task

Group comparisons showed differences between Adults 1 compared to Adults2 and Adults3 (β=0.25, p=.044) and between Children1 and adults 2 (β=0.35, p<.044). There was no difference between Adults 2 and 3 (p=.148) or between Children1 and Adults3 (p=.581).

"Phonological Form": As illustrated in Figure 4, participants in all groups and experiments performed at a similar level. One-sample t-tests demonstrated that only Adults1 performed significantly above 0.5 (t(30)=2.93, p<.001). Thus, only this group showed evidence of learning and generalising the co-occurrence between the determiner and the suffix (the distributional cue).

The group comparisons showed there was no evidence of a differences in generalisation across the three experiments, or between adults and children (Adults1 vs. Adults2&3, p=.083; Adults2 vs. Adults3, p=.294; Children1 vs. Adults2, p=.513;

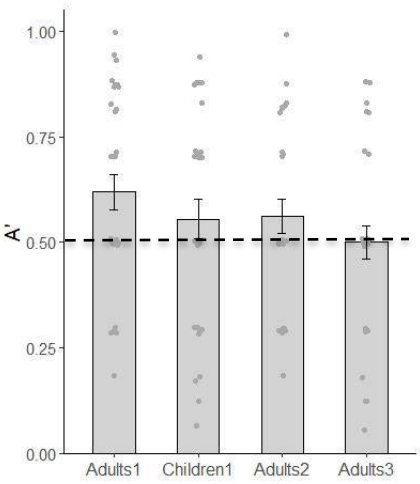


Figure 4: Performance on the "Phonological Form" Generalisation Task

Children1 vs. Adults3, p=.268). These results show only weak evidence for the learning and generalisation of the determiner + suffix co-occurrence in this paradigm.

In summary, only adults in Experiment 1 demonstrated the ability to utilise all three statistical cues to generalise newly formed grammatical knowledge. Their performance was reliably different from that of children, and of adults in Experiments 2 and 3, when generalising the trained cues to novel exemplars.

Given that Adults2 and Adults3 show the same level of word learning as Children1, this result suggests that the level of word learning may be a driver of grammatical generalisation and as such may explain some of the difference seen between adults and children in Experiment 1. This finding aligns with Bates and Goodman's (1997) lexicalist theory, which proposes that the emergence of grammar depends on lexical learning.

Although the lack of generalisation in children found here is in contrast to some previous studies in the literature (e.g. Hall et al., 2018; Lany & Saffran, 2010; 2011), this may be due to a number of factors, including the nature of the training and the structure and complexity of the regularities. In the current study, the training tasks always included simultaneous presentation of the referent with the novel words, unlike e.g. Lany and Saffran (2010; 2011) and Lany (2014), who trained participants on the phonological word form before introducing the referent, and Hall et al., (2018), who trained participants on a language that did not include a referent. Moreover, the simultaneous presentation of all noun class cues (phonological, distributional, and semantic) in the current study may have increased the complexity of the task, and affected the relative salience of the cues. Thus, further research exploring these methodological differences would help to clarify the role of semantic cues, and the effects of sequential vs simultaneous presentation of different type of cue.

Explicit Knowledge: Contributions to Generalisation

Table 2 shows the explicit knowledge scores for each group, presented separately for each morpheme. These scores suggest greater explicit knowledge for determiners than for suffixes across all groups/experiments.

A key aim in the current study was to assess the extent to which explicit knowledge contributes to generalisation performance, and whether this contribution differs in children and adults. We were specifically interested not in the group differences between children and adults (as children may be less able to verbalise their knowledge), but in the extent to which individual variation in the levels of explicit knowledge

Table 2: Descriptive Statistics for Explicit Knowledge Scores

	Determiner		Suffix	
	M	*SD*	*M*	*SD*
Adults1	2.39	1.02	0.58	1.03
Children1	0.70	1.11	0.10	0.31
Adults2	1.20	1.19	0.33	0.76
Adults3	1.50	1.17	0.23	0.63

within each group contributes to generalisation performance. To address these questions, multiple regressions were carried out for each generalisation task. The outcome variable in each regression was the A' score, while explicit knowledge scores for the relevant morpheme(s) were the predictor(s). For example, in the "phonological form" and "suffix only" generalisation tasks, only knowledge of the suffix was necessary for successful performance, so only suffix knowledge was used as a predictor, while for the "determiner and suffix" task, knowledge of both determiner and suffix was relevant.

As illustrated in Table 3, for adults in Experiment 1 explicit knowledge of the regularities was a significant predictor of generalisation performance in the "determiner & suffix" and "suffix only" tasks, but not in the "phonological form" task. Explicit knowledge of the relevant morpheme facilitated performance in the generalisation tasks. The strongest effect was for knowledge of the determiner-semantic mapping, which accounted for 27% of the variance in the 'determiner & suffix' task. In contrast to the adults, explicit knowledge of the regularities in children did not significantly predict performance in any of the generalisation tasks.

The pattern of results in Experiment 2 provides an informative comparison because the adults in this group showed low levels of generalisation (comparable to children), and intermediate levels of explicit awareness. Nonetheless, variability in generalisation within this group was significantly predicted by explicit awareness of the relevant morphemes. As with adults in Experiment 1, the strongest effect was for knowledge of the determiner-semantic mapping, which in this case accounted for an even larger proportion (38%) of the variance in the 'determiner & suffix' task. Finally, the adults in Experiment 3, showed low levels of generalisation as well as low levels of explicit knowledge. In this case, and similarly to the children in Experiment 1, there was no clear evidence of facilitatory effect of explicit knowledge on generalisation performance. This may suggest that the use of the WPM training task could prompt the emergence and correct use of explicit knowledge, at least in adults.

Overall these results suggest a partial role for explicit knowledge in grammatical generalisation for adults but not children. This still seems to hold when adults demonstrate similar levels of word learning and generalisation to children, suggesting that there may be differences in the extent to which adults and children draw on explicit processes when generalising in a grammatical SL task.

Conclusion

The current set of experiments demonstrates that explicit knowledge plays a role in grammatical category generalisation in adults but not children. This may be partially due to children's lower level of word knowledge, given the lower level of generalisation performance in adults when levels of word knowledge were matched to those of children. However, adults with a lower level of word knowledge still demonstrated a partial involvement of explicit knowledge in the generalisation tasks. This suggests the possibility of developmental differences between adults and children in the role of explicit and implicit processes when generalising in SL tasks. In future studies, more sensitive measures that do not rely on verbal reports would provide further insights into the contributions of explicit knowledge in implicit learning tasks across development.

Table 3: Multiple Regressions for the Role of Explicit Morpheme Knowledge on Generalisation Performance.

	Experiment 1 Adults					Experiment 1 Children				
	R^2	B	SE B	ß	*p*	R^2	B	SE B	ß	*p*
Determiner & Suffix Generalisation:	0.27					-0.02				
Explicit Determiner Knowledge	**0.27**	**0.14**	**0.04**	**0.54**	**.002**	-0.02	0.03	0.05	0.11	.558
Explicit Suffix Knowledge	0.00	0.02	0.04	0.08	.599	-0.00	0.15	0.18	0.16	.403
Suffix Only Generalisation:	0.16					0.01				
Explicit Suffix Knowledge		**0.10**	**0.04**	**0.43**	**.015**		0.16	0.13	0.21	..259
Phonological Form Generalisation:	0.05					0.03				
Explicit Suffix Knowledge		0.07	0.04	0.29	.112		0.22	0.15	0.26	.171
	Experiment 2 Adults					Experiment 3 Adults				
Determiner & Suffix Generalisation:	0.46					-0.03				
Explicit Determiner Knowledge	**0.38**	**0.17**	**0.04**	**0.62**	**<.001**	-0.03	0.03	0.04	0.11	.553
Explicit Suffix Knowledge	**0.07**	**0.13**	**0.06**	**0.30**	**.037**	-0.00	0.08	0.08	0.19	.329
Suffix Only Generalisation:	0.10					0.11				
Explicit Suffix Knowledge		0.12	0.06	0.36	.052		**0.13**	**0.06**	**0.37**	**.044**
Phonological Form Generalisation:	0.18					0.12				
Explicit Suffix Knowledge		**0.14**	**0.05**	**0.46**	**.011**		**-0.13**	**0.06**	**-0.38**	**.036**

Significant results are highlighted in bold.

Only morpheme knowledge salient to the generalisation task were included.

References

Aslin, R. N., & Newport, E. L. (2012). Statistical Learning: From Acquiring Specific Items to Forming General Rules. *Current Directions in Psychological Science, 21,* 170-176.

Bates, E., & Goodman, J. C. (2001). On the Inseparability of Grammar and the Lexicon: Evidence from Acquistion, Aphasia and Real-time Processing. *Language and Cognitive Processes, 12,* 507-584.

Batterink, L. J., Reber, P. J., Neville, H. J., & Paller, K. A. (2015). Implicit and explicit contributions to statistical learning. *Journal of Memory and Language*, *83*, 62–78.

Breitenstein, C., Zwitserlood, P., de Vries, M. H., Feldhues, C., Knecht, S., & Dobel, C. (2007). Five days versus a lifetime: intense associative vocabulary training generates lexically integrated words. *Restorative Neurology and Neuroscience*, *25*(5-6), 493–500.

Conway, C. M. & Christiansen, M. H. (2005). Modality-Constrained Statistical Learning of Tactile, Visual and Auditory Sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 24-39.

Farmer, T. A., Christiansen, M. H., & Monaghan, P. (2006). Phonological typicality influences on-line sentence comprehension. *Proceedings of the National Academy of Sciences*, *103*(32), 12203-12208.

Forster, K. I., & Forster, J. C. (2003). DMDX: A Windows display program with millisecond accuracy. *Behavior research methods, instruments, & computers*, *35*(1), 116-124.

Franco, A., Cleeremans, A. & Destrebecqz, A. (2011). Statistical Learning of two artificial languages presented successively: how conscious? *Frontiers in Psychology, 2,* 229.

Frost, R. L. A., Monaghan, P., & Christiansen, M. H. (2019). Mark my words: High frequency marker words impact early stages of language learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition. Advance online publication*.

Gómez, R. L. (2002). Variability and detection of invariant structure. *Psychological Science*, *13*(5), 431–436.

Gómez, R. L., & Gerken, L. (2000). Infant artificial language learning and language acquisition. *Trends in Cognitive Sciences*, *4*(5), 178–186.

Hall, J., Owen VAN Horne, A., & Farmer, T. (2018). Distributional learning aids linguistic category formation in school-age children. *Journal of Child Language*, *45*(3), 717–735.

Lany, J. (2014). Judging words by their covers and the company they keep: Probabilistic cues support word learning. *Child development*, *85*(4), 1727-1739.

Lany, J., & Saffran, J. R. (2010). From statistics to meaning: infants' acquisition of lexical categories. *Psychological Science*, *21*(2), 284–291.

Lany, J., & Saffran, J. R. (2011). Interactions between statistical and semantic information in infant language development. *Developmental Science*, *14*(5), 1207–1219.

Mirkovic, J., Forrest, S., & Gaskell, G. (2011). Semantic Regularities in Grammatical Categories: Learning Grammatical Gender in an Artificial Language. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 33. Retrieved from https://escholarship.org/uc/item/54412022.

Mirković, J., & Gaskell, M. G. (2016). Does sleep improve your grammar? Preferential consolidation of arbitrary components of new linguistic knowledge. *PloS one*, *11*(4), e0152489.

Monaghan, P., Chater, N., & Christiansen, M. H. (2005). The differential role of phonological and distributional cues in grammatical categorisation. *Cognition*, *96*(2), 143–182.

.Rastle, K., Harrington, J., & Coltheart, M. (2002). 358,534 nonwords: the ARC Nonword Database. *The Quarterly Journal of Experimental Psychology. A, Human Experimental Psychology*, *55*(4), 1339–1362.

Reber, A. S. (1967). Implicit Learning of Artificial Grammars. *Journal of Verbal Learning and Verbal Behaviour, 6,* 855-863.

Rossion, B., & Pourtois, G. (2001). Revisiting Snodgrass and Vanderwart's object database: Color and texture improve object recognition. *Journal of Vision*, *1*(3), 413-413.

Smalle, E. H. M., Page, M. P. A., Duyck, W., Edwards, M. & Szmalec, A. (2017). Children retain implicitly learned phonological sequences better than adults: a longitudinal study. *Developmental Science, 21(5),* e12634.

Wonnacott, E., Boyd, J. K., Thomson, J., & Goldberg, A. E. (2012). Input effects on the acquisition of a novel phrasal construction in 5 year olds. *Journal of Memory and Language*, *66*(3), 458–478.

The Computational Structure of Unintentional Meaning

Mark K. Ho (mho@princeton.edu)
Department of Psychology, Princeton University
Princeton, NJ 08540

Joanna Korman*(jkorman@mitre.org)
The MITRE Corporation
Bedford, MA 01730

Thomas L. Griffiths (tomg@princeton.edu)
Department of Psychology, Princeton University
Princeton, NJ 08540

Abstract

Speech-acts can have literal meaning as well as pragmatic meaning, but these both involve consequences typically *intended* by a speaker. Speech-acts can also have *unintentional meaning*, in which what is conveyed goes above and beyond what was intended. Here, we present a Bayesian analysis of how, to a listener, the meaning of an utterance can significantly differ from a speaker's intended meaning. Our model emphasizes how comprehending the intentional *and* unintentional meaning of speech-acts requires listeners to engage in sophisticated model-based perspective-taking and reasoning about the *history* of the state of the world, each other's actions, and each other's observations. To test our model, we have human participants make judgments about vignettes where speakers make utterances that could be interpreted as intentional insults or unintentional *faux pas*. In elucidating the mechanics of speech-acts with unintentional meanings, our account provides insight into how communication both functions and malfunctions.

Keywords: Bayesian modeling, social cognition, common ground, speech-act theory, faux pas, theory of mind

Introduction

People sometimes communicate things that they did not intend or expect. Consider the following vignette, adapted from Baron-Cohen et al. (1999):

> **Curtains** Paul had just moved into a new apartment. Paul went shopping and bought some new curtains for his bedroom. After he returned from shopping and had put up the new curtains in the bedroom, his best friend, Lisa, came over. Paul gave her a tour of the apartment and asked, "How do you like my bedroom?"
>
> "Those curtains are horrible," Lisa said. "I hope you're going to get some new ones!"

Clearly, Lisa committed a social blunder or *faux pas* with her remark. What happened here? When Lisa says, "Those curtains look horrible," she is merely stating her private aesthetic experience of the curtains. The *literal* meaning is straightforward: The curtains look bad. And the *intended* or *expected* meaning of her utterance is largely captured by this literal meaning. However, to Paul, the utterance means more. Specifically, what Lisa is *really* saying is that *he* chose horrible curtains. Of course, Lisa did not "really" say that Paul's choice in curtains was horrible—she had no intention of conveying such an idea. Paul might even realize this. Nonetheless, the remark stings. Why? Lisa and Paul each possess a piece of a puzzle, and when put together, they entail that Paul has awful taste in curtains. At the outset, neither one knew that they each had a piece of a puzzle. But once Lisa makes her remark, she inadvertently completes the puzzle, at least from Paul's perspective.

Standard models of communication (Grice, 1957; Sperber & Wilson, 1986) tend to focus on how people use language successfully. For example, people can imply more than they literally mean (Carston, 2002), convey subtle distinctions via metaphor (Tendahl & Gibbs Jr, 2008), and manage their own and others' public face using politeness (Levinson, Brown, Levinson, & Levinson, 1987; Yoon, Frank, Tessler, & Goodman, 2018). But things do not always go smoothly, as Paul and Lisa's situation indicates. Sometimes people find themselves having inadvertently stepped on conversational landmines, meaning things that they never anticipated meaning. Notably, because such situations present complex dilemmas of mutual perspective-taking against a backdrop of divergent knowledge, they can serve as advanced tests of theory of mind (Baron-Cohen et al., 1999; Zalla, Sav, Stopin, Ahade, & Leboyer, 2009; Korman, Zalla, & Malle, 2017). But how do people reason about such dilemmas? And how can this be understood computationally? Disentangling unintentional meaning can shed light on how communication works in a broader social context as well as inform the design of artificial intelligences that interact with people.

Here, we develop a rational, cognitive account of interpreting unintentional speech-acts that builds on existing Bayesian models of language (e.g., Rational Speech Act [RSA] models [Goodman & Frank, 2016]). To do this, we analyze the general epistemic structure of social interactions such as the one described above and model listeners engaging in *model-based perspective-taking*. In particular, our model explains how the same utterance could be interpreted as either an (unintentional) *faux pas* or an *intentional insult* depending on the context of a listener and speaker's interaction. We then test several model predictions in an experiment with human participants. In the following sections, we outline our computational model, experimental results, and their implications.

*The author's affiliation with The MITRE Corporation is provided for identification purposes only, and is not intended to convey or imply MITRE's concurrence with, or support for, the positions, opinions, or viewpoints expressed by the author.

A Bayesian Account of Unintentional Meaning

During social interactions, people reason about the world as well as each other's perspective on the world (Brown-Schmidt & Heller, 2018). Thus, our account has two components, which we formulate as probabilistic models. First, we specify a *world model* that captures common-sense relationships between world states, actions, and events. Second, we define *agent models* of a speaker and listener reasoning about the world and one another.

World Model

We model the interaction as a partially observable stochastic game (POSG), a generalization of Markov Decision Processes (MDPs) with multiple agents with private observations (Kuhn, 1953). Formally, a world model $\mathcal{W} = \langle I, \mathcal{S}, \mathcal{A}, \mathcal{Z}, \mathcal{T} \rangle$ where:

- I is a set of n agents indexed $1, ..., n$;
- $\mathcal{S}$ is a set of possible states of the world, where each state $s \in \mathcal{S}$ is an assignment to k variables, $s = (x_0, x_1, ..., x_k)$;
- $\mathcal{A} = \times_{i \in I} \mathcal{A}^i$ is the set of joint actions, i.e., every combination of each agent i's actions, $\mathcal{A}^i$ (including utterances);
- $\mathcal{Z} = \times_{i \in I} \mathcal{Z}^i$ is the set of joint private observations, which is every possible combination of each individual agent i's private observation set, $\mathcal{Z}^i$; and
- $\mathcal{T} = P(z, s' \mid s, a)$ is a transition function representing the probability of a joint observation z and next state s' given a previous state $s \in \mathcal{S}$ and joint action $a \in \mathcal{A}$ was taken.

In **Curtains**, the initial state, s_0, includes Paul with the old curtains in the apartment and Lisa elsewhere. There is also a latent state feature of interest: whether Paul has good or bad taste. At $t = 0$, Paul's action, a_0^{Paul}, is choosing new curtains, while Lisa's action, a_0^{Lisa}, is going to the apartment. The joint action, $a_0 = (a_0^{\text{Paul}}, a_0^{\text{Lisa}})$, results in a new state, s_1, with them both in the apartment, the curtains either good or bad, and Paul's taste. Paul's observation, z_0^{Paul}, but not Lisa's, z_0^{Lisa}, includes Paul having put up the curtains. These relationships between world states (e.g. Paul and Lisa's locations), actions (e.g. Lisa walking to Paul's apartment), and observations (e.g. Paul observing himself put up the curtains) are formally encoded in the transition function $\mathcal{T}$. The sequence of states, joint actions and observations resulting from such interactions constitute the *history* up to a point t, $\vec{h}_t = (s_0, a_0, z_0, ..., s_{t-1}, a_{t-1}, z_{t-1}, s_t)$.

Agent Models

Agents are modeled as Bayesian decision-makers (Bernardo & Smith, 1994) who can reason about the world and other agents as well as take actions—including making utterances.

Interactive Belief State Agents' beliefs are probability distributions over variables that represent aspects of the current state, previous states, or each other's beliefs. The configuration of these first- and higher-order, recursive beliefs constitute their *interactive belief state* (Gmytrasiewicz & Doshi, 2005). We refer to an agent i's beliefs as b^i. For example, if we denote Paul's taste as the variable T^{Paul}, then Paul's belief that his taste is good is $b^{\text{Paul}}(T^{\text{Paul}} = \texttt{Good})$. Higher-order beliefs can also be represented. For instance, we can calculate Paul's *expectation* of Lisa's belief in his taste as $\mathbb{E}_{b^{\text{Paul}}}[b^{\text{Lisa}}](T^{\text{Paul}}) = \sum_{b^{\text{Lisa}}} b^{\text{Paul}}(b^{\text{Lisa}}(T^{\text{Paul}}))$.

An agent i's beliefs are a function of their prior, model of the world, model of other agents, and *observation history* up to time t, $\vec{z}_t^{\,i}$. Note that $\vec{z}_t^{\,i}$ can include observations that are completely private to i (e.g., Lisa's personal aesthetic experience) as well as public actions and utterances (e.g., Lisa's remark to Paul). Thus, we denote Paul's belief about his taste at a time t as $b_t^{\text{Paul}}(T^{\text{Paul}}) = b^{\text{Paul}}(T^{\text{Paul}} \mid \vec{z}_t^{\text{Paul}})$. Given a sequence of observations, $\vec{z}_t^{\,i}$, posterior beliefs about a variable X are updated via Bayes' rule:

$$b(X \mid \vec{z}_t^{\,i}) \propto b(\vec{z}_t^{\,i} \mid X) b(X) \tag{1}$$

$$= \sum_{\vec{h}_t} b(\vec{z}_t^{\,i} \mid \vec{h}_t) b(\vec{h}_t, X) \tag{2}$$

The capacity to reason about higher-order beliefs (e.g., Paul's beliefs about Lisa's belief in his taste), along with Equation 2 express agents' joint inferences about events and model-based perspective-taking.

Speaker Model Speakers have beliefs and goals. When choosing what to say, they may have beliefs and goals with respect to the listener's beliefs and goals. In our example, Lisa may care about being informative about how she sees the curtains, but may also think Paul cares about having good taste in curtains and care whether she hurts his feelings. Following previous work (e.g., Franke, 2009), we model speakers as reasoning about changes in *belief states*. Here, we are interested in how a speaker can intend to mean one thing but inadvertently mean another. Thus, we distinguish between state variables that the speaker wants to be *informative* about, X^{Info} (e.g., how Lisa sees the curtains), and *evaluative* variables, X^{Eval}, that the listener wants to take on a specific value $x^{\text{Eval}*}$ (e.g., Paul's taste being good). The speaker then cares about the changes in those quantities. Formally:

$$\Delta_t^{\text{L-Info}} = b_{t+1}^L(X^{\text{Info}} = x^{\text{Info}}) - b_t^L(X^{\text{Info}} = x^{\text{Info}}), \tag{3}$$

where x^{Info} is given by $\vec{h}_t$; and,

$$\Delta_t^{\text{L-Eval}} = b_{t+1}^L(X^{\text{Eval}} = x^{\text{Eval}*}) - b_t^L(X^{\text{Eval}} = x^{\text{Eval}*}). \tag{4}$$

A speaker who is interested in what the listener thinks about X^{Info} and X^{Eval} will, at a minimum, anticipate how their utterances will influence $\Delta_t^{\text{L-Info}}$ and $\Delta_t^{\text{L-Eval}}$. A speaker would then have a reward function defined as:

$$R^{\text{S}}(a_t^S, \vec{z}_{t+1}^L) = \theta^{\text{L-Info}} \Delta_t^{\text{L-Info}} + \theta^{\text{L-Eval}} \Delta_t^{\text{L-Eval}} \tag{5}$$

where the θ terms correspond to how the speaker values certain outcomes in the listener's mental state. For instance, if $\theta^{\text{L-Eval}} < 0$, the speaker *wants* to insult the speaker.

Given Equation 5, a speaker can take utterances based on expected future utility/rewards (or *value* [Sutton & Barto,

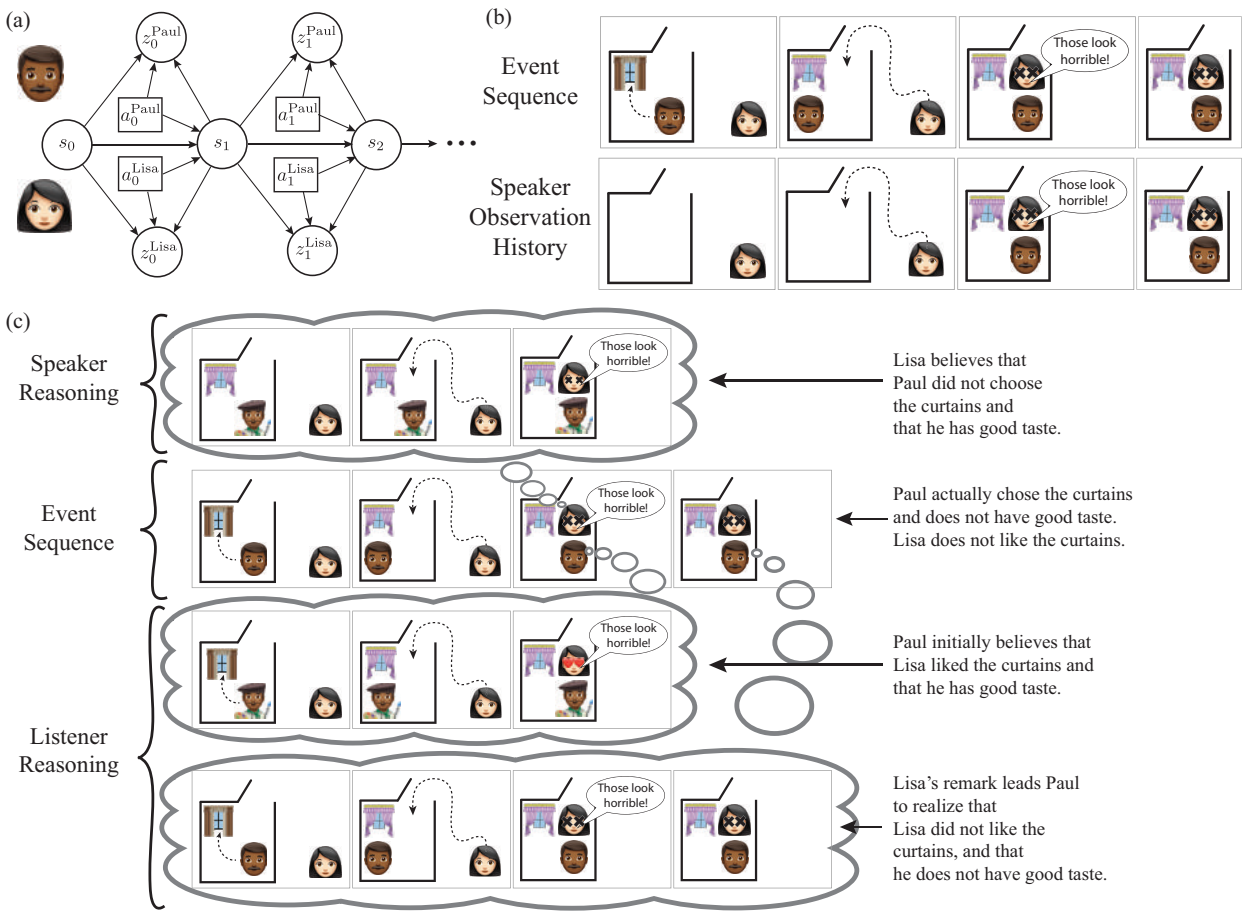


Figure 1: Model and example of unintentional meaning. (a) Influence diagram with state, action, and observation dependencies. Circles correspond to world state (e.g., s_t) and observation (e.g., z_t^i) variables; squares correspond to agent action variables (including utterances) (e.g., a_t^i). (b) Event sequence in **Curtains** (top) and speaker observation history (bottom). Lisa does not observe Paul choose the curtains. Only Lisa experiences whether the curtains look good or bad and comments on this experience. (c) Diagram of interactive belief state over time in **Curtains**.

1998]), where the expectation is taken with respect to the speaker's beliefs, b_t^S. That is, given observations $\vec{z}_t^S$, the value of a_t^S is $V^S(a_t^S; \vec{z}_t^S) = \mathbb{E}_{b_t^S}\left[R^S(a_t^S, \vec{z}_{t+1}^L)\right]$, and an action is chosen using a Luce choice rule (Luce, 1959).

Listener Inference Our goal is to characterize how a listener's interpretation of an utterance can differ from a speaker's intended meaning, which requires specifying listener inferences. We start with a simple listener that understands the literal meanings of words when spoken. Following previous models (Franke, 2009; Goodman & Frank, 2016), the literal meaning of an utterance a^S is determined by its truth-functional denotation, which maps histories to Boolean truth values, $[\![a^S]\!] : \vec{h}_t \mapsto y$, $y \in \{\texttt{True}, \texttt{False}\}$. A literal listener's model of speaker utterances is:

$$b(a^S \mid \vec{h}_t) \propto \begin{cases} 1-\varepsilon & \text{if } [\![a^S]\!](\vec{h}_t) \\ \varepsilon & \text{if } \neg[\![a^S]\!](\vec{h}_t) \end{cases}$$

where ε is a small probability of a^S being said even if it happens to be false.

We can also posit a more sophisticated listener who, rather than assuming utterances literally reflect reality, reason about how a speaker's beliefs and goals mediate their use of language. This type of listener draws inferences based on an *intentional* model of a speaker that track the quantities in Equations 3 and 4 as well as maximize the expected rewards. These inferences, however, occur while the listener is also reasoning about the actual sequence of events $\vec{h}_t$, making it possible to draw inferences based on utterances that the speaker did not anticipate.

Model Simulations

In the original **Curtains** scenario, Lisa was not present when Paul put up the curtains. As a result, Lisa's comment ("Those curtains are horrible") is interpreted in a *diverging* observation history context. But what if Lisa had been present when Paul put up the curtains and made the same utterance? Given a *shared* observation history, Lisa's utterance is still offensive, but now Lisa has all the information needed to realize it would be offensive. Put simply, in the *diverging history* con-

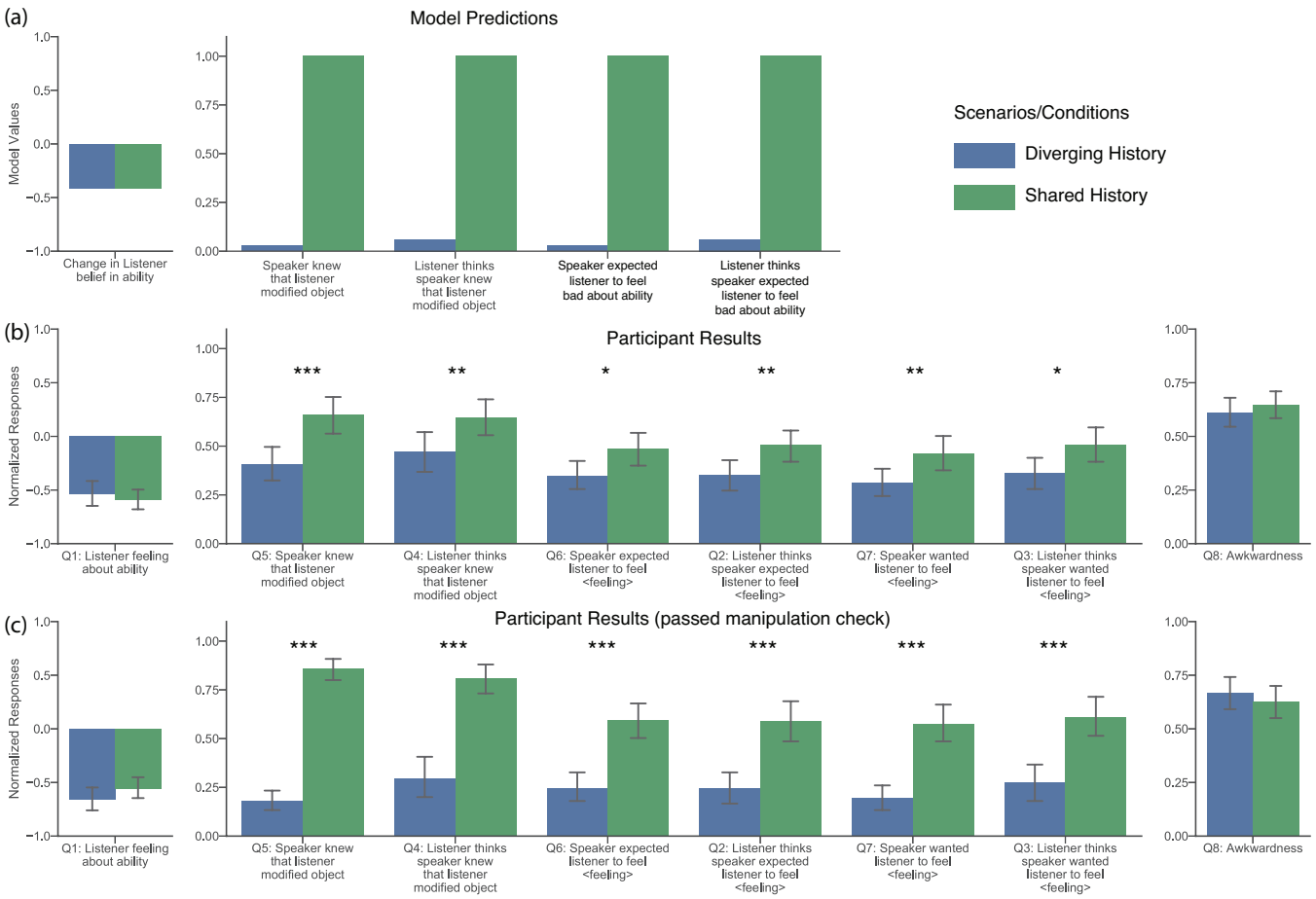


Figure 2: (a) Model predictions. The model predicts that the listener's change in belief in the evaluative variable ($\Delta_t^{\text{L-Eval}}$) is equally negative in the diverging and shared history scenarios. However, whether the speaker anticipated the offensiveness of their comment differs between the two scenarios, as do the listener's beliefs about the speaker's anticipation. (b) Judgments from all participants by question. Responses were normalized depending on whether response scales were valanced (Q1), likelihood (Q2-Q7), or qualitative (Q8). (c) Judgments from participants who correctly identified whether the speaker knew the listener modified the object. * : $p < .05$, ** : $p < .01$, *** : $p < .001$.

text, the utterance is a faux pas, whereas in the *shared history* context, it is an intentional insult.

In this section, we discuss how our model can be used to make these intuitive predictions precise and explain how they arise from agents' interactions and model-based perspective-taking within a shared environment. We implemented our model in WebPPL (Goodman & Stuhlmüller, 2014), a programming language that can express stochastic processes like POSGs as well as Bayesian inference.

Generative Model

To model a scenario like **Curtains**, we define agents, objects, and features assigned to them. These are the curtains, which have a location (inside Paul's apartment); the speaker (Lisa), who has a location (inside or outside Paul's apartment) and a perception of the curtains (good or bad); and the listener (Paul), who has a location (inside or outside) and ability to choose curtains (high or low). Additionally, the listener can either act on the curtains or not, while the speaker can enter the apartment and make an utterance about the curtains ("the curtains look good", "the curtains look bad", or `<nothing>`). The truth-conditional semantics of the utterances map onto world features in a standard manner, and we set $\varepsilon = .05$.

Observations depend on whether agents and objects are co-located and are defined as subsets of state and action variables. For instance, if Paul and Lisa are both inside the house and Paul modifies the curtains, they both observe that Paul acted on the curtains, but only Lisa directly knows whether they look good to her. Finally, we define a state and action prior for both agents such that the listener's ability is initially high ($p = 0.90$), the speaker's perception of the object is initially random ($p = 0.50$), and the listener has a low probability of modifying the object ($p = 0.05$).

Model Predictions

Given the generative model, we can provide scenarios and calculate aspects of the resulting interactive belief state (the listener and speaker's beliefs about the world and each other's

beliefs). In particular, we compare the results of a *shared history* with those of a *diverging history*. In the shared history, the speaker and listener are both present when the listener modifies the object, whereas in the diverging history, the speaker is not present when the listener acts on the object. Otherwise, the two scenarios are the same and the speaker comments on the curtains being bad. Figure 2a displays the results of the simulation when given each of the two histories. In both histories, the listener learns that their ability when modifying the object, X^{Eval}, is low (i.e., $\Delta_t^{\text{L-Eval}} < 0$). They also learn about the informative variable (i.e., $\Delta_t^{\text{L-Info}} > 0$).

However, the resulting interactive belief states differ in important ways. For example, in the diverging history, although the listener concludes that the evaluative variable is *low*, the speaker thinks the evaluative variable is *high*. Relatedly, the speaker thinks the utterance was informative ($\mathbb{E}_{b^S}[\Delta^{\text{L-Info}}] > 0$) but not offensive ($\mathbb{E}_{b^S}[\Delta^{\text{L-Eval}}] = 0$). Moreover, the listener knows the speaker believes that their comment was expected to be informative and not offensive. In the shared history, this is not the case: The listener and speaker both believe the evaluative variable is low, and they both know the resulting informational and evaluative effects. Because they were both present when the listener modified the object, they share expectations about the utterance's meaning.

Put intuitively, whereas the shared history leads to an *expected insult*, the diverging history leads to a *faux pas*. Our model explains this difference in terms of differential transformations of the listener and speaker's interactive belief state.

Experiment

Our model explains how different observation histories result in interactive belief states, which can produce unintentional meaning. To test whether this accurately describes people's capacity to reason about unintentional meaning, we had people read vignettes that described scenarios involving shared or diverging observation histories. The underlying logical structure of all the vignettes mirrored that of **Curtains**, and so the model predictions described in the previous section apply to all of them. Participants then provided judgments corresponding to predicted differences in listener/speaker beliefs. The study's main hypotheses were preregistered on the Open Science Framework platform (`https://osf.io/84wqn`). Overall, we find that our model captures key qualitative features of people's inferences.

Materials

We developed a set of vignettes that included interactions in different contexts as well as different histories of interaction. Each vignette involved a listener (e.g., Paul) who could potentially interact with an object (e.g., curtains) as well as a speaker (e.g., Lisa) who makes an utterance about their negative aesthetic experience of the object (e.g., "The curtains look horrible"). In the shared history versions of the vignettes, the two agents were described as being both present when the listener acted on an object. In the diverging history versions of the vignettes, the speaker was not present when the listener interacted with the object. Each vignette involved one of five contexts: *Curtain*, *Story-Prize*, *Wine-bottle*, *Cupcakes*, and *Parking*. Thus there were a total of ten items (Diverging/Shared history $\times$ 5 contexts). All items used in the experiment are available on the primary author's website.

Procedure

One-hundred participants were recruited via MTurk to participate in our experiment using PsiTurk (Gureckis et al., 2016). Each participant read one of the ten context-history items, and then answered the following questions in order:

- Q1: At this point, how does `<listener>` feel about their ability to `<action>`? [6 point scale ranging "Very Bad" to "Very Good" with no neutral option]
- Q2: `<listener>` thinks that `<speaker>` expected that their remark would make them feel `<Q1_response>`.
- Q3: `<listener>` thinks that in making the remark, `<speaker>` wanted to make them feel `<Q1_response>`.
- Q4: `<listener>` thinks that `<speaker>` thinks that `<listener>` `<action>`.
- Q5: `<speaker>` knew that `<listener>` `<action>`.
- Q6: In making the remark, `<speaker>` expected `<listener>` to feel `<Q1_response>`.
- Q7: In making the remark, `<speaker>` wanted `<listener>` to feel `<Q1_response>`.
- Q8: How awkward is this situation? [5 point scale ranging "Not at all" to "Extremely"

The values for `<listener>`, `<speaker>`, and `<action>`were specified parametrically based on the context, while the value for `<Q1_response>` was filled in based on the answer to the first question. The response scale for questions 2-7 was a six-point scale ranging from "Definitely Not" to "Definitely", with no neutral point. We included question 8 because previous work studying faux pas have focused on this question (Zalla et al., 2009). Participants were also given free response boxes to elaborate on their interpretation of the situation and answered demographic questions.

Question	β	S.E.	*df*	*t*	*p*
Q1	-0.06	0.07	94.0	-0.77	
Q2	0.15	0.06	94.0	2.65	**
Q3	0.15	0.06	94.0	2.50	*
Q4	0.18	0.06	94.0	2.78	**
Q5	0.25	0.06	94.0	4.34	***
Q6	0.14	0.06	94.0	2.53	*
Q7	0.15	0.06	94.0	2.64	**
Q8	0.04	0.05	94.0	0.78	

Table 1: Tests for Diverging/Shared history factor.

Experimental Results

Manipulation check To assess whether the Diverging/Shared history manipulation worked, we examined responses to Q5 (whether the speaker knew the listener acted on the object). A comparison in which the responses were coded

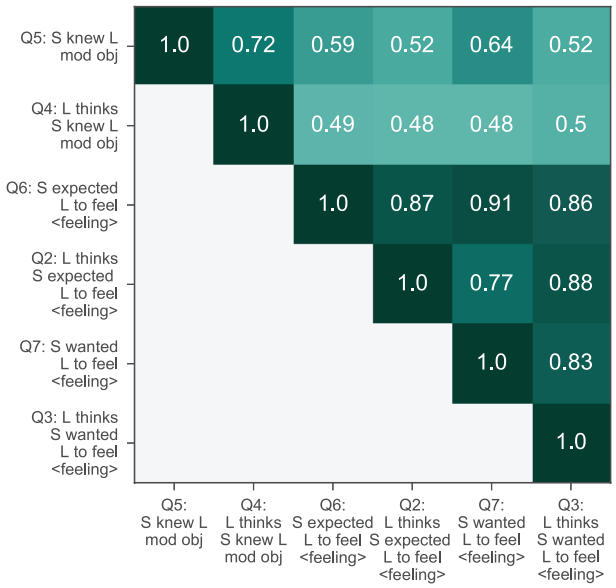


Figure 3: Judgment correlations (Pearson's r).

as Yes or No (i.e., above or below the middle of the response scale) showed that it was effective ($\chi^2(1) = 7.92, p < .01$). However, a number of participants (15 of 50 in Shared; 20 of 50 in Diverging) did not pass this manipulation check and gave opposite answers than implied by the stories. Whether their responses are included does not affect our qualitative results, and in our analyses we use the full data set. Figure 2c plots the results for those who passed this check.

Judgment differences Responses paralleled the model predictions for the Shared versus Diverging history versions of the vignettes (Figure 2b). For each judgment, we fit mixed-effects linear models with context intercepts as a random effect and history as a fixed effect. Table shows tests of significance on the Diverging/Shared history parameters. Judgments about the listener's feelings (Q1) were negative and not significantly different, indicating that people perceived the psychological impact (at least with respect to ability) of the utterance as roughly equivalent. In contrast, questions about the interactive belief state—the listener and speaker's beliefs about the world and each other's beliefs (Q2-Q7)—differed as predicted by the model. In particular, participants thought that the speaker neither expected that their utterance would hurt the listener's feelings, nor that they wanted to do so. Participants judged that the listener recognized this as well.

Judgment correlations Judgments among questions about higher order mental states were strongly correlated, while those between the higher order mental states and the listener's action were weaker (Figure 3). Specifically, those about speaker mental states (Q6, Q7) and listener beliefs about speaker mental states (Q2, Q3) were all highly correlated (all $r \in [0.77, 0.91], p < .001$). In contrast, questions about knowledge of the object being modified (Q4, Q5) were only moderately correlated with those about anticipated effects (Q2, Q3, Q6, Q7) (all $r \in [0.48, 0.64], p < .001$).

Discussion

People's actions can have unexpected consequences, and speech-acts are no different. To understand unintentional meaning though, we need to characterize how a communicative act can lead to unanticipated epistemic consequences. Sometimes, a listener can learn something from an utterance that a speaker did not intend to convey or may not even believe (e.g., as in **Curtains**). Here, we have presented a Bayesian model and experiments testing how people reason about scenarios involving unintentional speech acts. Specifically, our account treats speech-acts as actions taken by a speaker that influence a shared interactive belief state—the beliefs each agent has about the world and each other's beliefs. In doing so, we can capture the inferences that underlie unintentional meaning.

The current work raises important empirical and theoretical questions about how people reason about interactive beliefs and unintentional meaning. For instance, our experiments focus on third-party judgments about how a listener interprets the unintended meanings of utterances, but further work would be needed to assess how listeners do this (e.g., when the victim of an offhand comment) or even how speakers can recognize this (e.g., realizing one has put their foot in their mouth). Additionally, we have presented a Bayesian account of unintentional meaning in which agents reason about a large but finite set of possible histories of interaction. In everyday conversation, the space of possible histories can be much larger or even infinite. It is thus an open question how people can approximate the recursive inferences needed to make sense of unintentional meaning.

A rigorous characterization of unintentional meaning can deepen our understanding of how communication works in a broader social context. For example, attempts to build common ground through shared experience (Clark & Marshall, 1981; McKinley, Brown-Schmidt, & Benjamin, 2017) or manage face with polite speech (Levinson et al., 1987; Yoon et al., 2018) could be understood, in part, as strategies for forestalling unintentional meaning. And given that intentionality plays a key role in judgments of blame (Baird & Astington, 2004), phenomena like plausible deniability could be understood as people leveraging the possibility of unintentional meaning to covertly accomplish communicative goals (Pinker, Nowak, & Lee, 2008). Although further investigation is needed to test the extent to which people can track and influence interactive belief states (as well as how artificial agents can do so), this work provides a point of departure for computationally investigating these social and cognitive aspects of communication.

Acknowledgments

This material is based upon work supported by the NSF under Grant No. 1544924.

References

Baird, J. A., & Astington, J. W. (2004). The role of mental state understanding in the development of moral cognition and moral action. *New Directions for Child and Adolescent Development*, *2004*(103), 37-49. doi: 10.1002/cd.96

Baron-Cohen, S., O'Riordan, M., Stone, V., Jones, R., & Plaisted, K. (1999). Recognition of Faux Pas by Normally Developing Children and Children with Asperger Syndrome or High-Functioning Autism. *Journal of Autism and Developmental Disorders*, *29*(5), 407–418. doi: 10.1023/A:1023035012436

Bernardo, J. M., & Smith, A. F. (1994). *Bayesian theory*. John Wiley & Sons.

Brown-Schmidt, S., & Heller, D. (2018). Perspective-taking during conversation. In G. Gaskell & S. A. Rueschemeyer (Eds.), *Oxford handbook of psycholinguistics (2nd ed.)*. Oxford: Oxford University Press Oxford.

Carston, R. (2002). *Thoughts and utterances: The pragmatics of explicit communication*. Blackwell Publishing.

Clark, H. H., & Marshall, C. R. (1981). Definite reference and mutual knowledge. *Elements of discourse understanding*.

Franke, M. (2009). *Signal to act: Game theory in pragmatics*. Institute for Logic, Language and Computation.

Gmytrasiewicz, P. J., & Doshi, P. (2005). A framework for sequential planning in multi-agent settings. *Journal of Artificial Intelligence Research*, *24*, 49–79.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences*, *20*(11), 818–829.

Goodman, N. D., & Stuhlmüller, A. (2014). *The Design and Implementation of Probabilistic Programming Languages.* `http://dippl.org`. (Accessed: 2018-9-12)

Grice, H. P. (1957). Meaning. *The philosophical review*, *66*(3), 377–388.

Gureckis, T. M., Martin, J., McDonnell, J., Rich, A. S., Markant, D., Coenen, A., ... Chan, P. (2016). psiTurk: An open-source framework for conducting replicable behavioral experiments online. *Behavior research methods*, *48*(3), 829–842.

Korman, J., Zalla, T., & Malle, B. F. (2017). Action understanding in high-functioning autism: The faux pas task revisited. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. J. Davelaar (Eds.), *Proceedings of the 39th annual conference of the cognitive science society* (p. 2451-2456). Austin, TX: Cognitive Science Society.

Kuhn, H. (1953). Extensive games and the problem of information. In H. Kuhn & A. Tucker (Eds.), *Contributions to the theory of games II* (pp. 193–216). Princeton University Press.

Levinson, P., Brown, P., Levinson, S. C., & Levinson, S. C. (1987). *Politeness: Some universals in language usage* (Vol. 4). Cambridge university press.

Luce, R. D. (1959). On the possible psychophysical laws. *Psychological review*, *66*(2), 81.

McKinley, G., Brown-Schmidt, S., & Benjamin, A. (2017). Memory for conversation and the development of common ground. *Memory and Cognition*, *45*(8), 1281-1294. doi: https://doi.org/10.3758/s13421-017-0730-3

Pinker, S., Nowak, M. A., & Lee, J. J. (2008). The logic of indirect speech. *Proceedings of the National Academy of Sciences*, *105*(3), 833–838. doi: 10.1073/pnas.0707192105

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition*. Cambridge, MA, USA: Harvard University Press.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT press.

Tendahl, M., & Gibbs Jr, R. W. (2008). Complementary perspectives on metaphor: Cognitive linguistics and relevance theory. *Journal of pragmatics*, *40*(11), 1823–1864.

Yoon, E. J., Frank, M. C., Tessler, M. H., & Goodman, N. D. (2018, Dec). *Polite speech emerges from competing social goals.* PsyArXiv. Retrieved from `psyarxiv.com/67ne8` doi: 10.31234/osf.io/67ne8

Zalla, T., Sav, A.-M., Stopin, A., Ahade, S., & Leboyer, M. (2009, Feb 01). Faux pas detection and intentional action in asperger syndrome. a replication on a french sample. *Journal of Autism and Developmental Disorders*, *39*(2), 373–382. doi: 10.1007/s10803-008-0634-y

How can diverse memory improve group decision making?

Hidehito Honda (hitohonda.02@gmail.com)
Department of Psychology, Yasuda Women's University
6-13-1, Yasuhigashi, Asaminami-ku, Hiroshima-shi, Hiroshima, 731-0153, Japan

Itsuki Fujisaki (bpmx3ngj@gmail.com)
Graduate School of Arts and Sciences, The University of Tokyo
3-8-1, Komaba, Meguro-ku, Tokyo, 153-8902, Japan

Toshihiko Matsuka (matsuka@chiba-u.jp)
Department of Cognitive and Information Science, Chiba University
1-33, Yayoi-cho, Inage-ku, Chiba-shi, Chiba, 263-8522, Japan

Kazuhiro Ueda (ueda@gregorio.c.u-tokyo.ac.jp)
Graduate School of Arts and Sciences, The University of Tokyo
3-8-1, Komaba, Meguro-ku, Tokyo, 153-8902, Japan

Abstract

Previous studies have shown that people can make adaptive inferences based on memory-based simple heuristics such as recognition, fluency, or familiarity heuristic. In the present study, we discussed the adaptive nature of memory-based simple heuristics in a group decision making setting. In particular, we examined how the diversity of memory affected group decision making when group members were assumed to make inferences based on the familiarity heuristic. We predicted that, when the group members' memories were diverse, group decision making would become more accurate. To examine this prediction, we conducted a behavioral experiment and computer simulations, and our results generally supported the prediction. We discuss the role of diverse memories in generating adaptive group decision making.

Keywords: group decision making; heuristics; ecological rationality; diversity

Introduction

In research on human judgments and decisions, one of the most studied topics has been the heuristics people use. Previous studies have shown that, although heuristics can produce biases (e.g., Tversky & Kahneman, 1974), they generally result in adaptive judgments and decisions (e.g., Gigerenzer, Todd, & The ABC Research Group, 1999). Some heuristics, such as the availability (Tverky & Kahneman, 1973) or recognition heuristic (Goldstein & Gigerenzer, 2002), are highly related to the nature of an individual's memory. We shall discuss the adaptive nature of memory-based simple heuristics in terms of group decision making.

How do memory-based simple heuristics work in a group decision making setting? Given that individuals can make adaptive judgments and decisions in general based on the memory-based simple heuristics, when each member relies on such heuristics and the group makes a collective decision by, for example, simple majority rule, the group may be able to make good decisions in general. However, as described above, heuristics produce biases. For some situations, biased inferences are enhanced, and group performance may be deteriorated. Thus, although memory-based simple heuristics will enhance group decision making in general, they will also enhance biased group decision making in some cases. Fujisaki, Honda, and Ueda (2018) used computer simulations to show that a group does not always perform well when group members use strategies, which are regarded as generally adaptive in individual usage, because of biases generated by the strategies.

How, then, can the biases of memory-based simple heuristics in a group decision making setting be resolved? Recently, research has discussed how groups can achieve good performance such as wisdom-of-crowds or collective intelligence in terms of group diversity (e.g., Fujisaki et al., 2018; Jönsson, Hahn, & Olsson, 2015; Lorenz, Rauhut, Schweitzer, & Helbing, 2011; Luan, Katsikopoulos, & Reimer, 2012; Mavrodiev, Tessone, & Schweitzer, 2013). In group decision making based on members who use memory-based simple heuristics, if members' memories vary (i.e., memories in group members are diverse), biases generated by heuristics may be resolved.

In the present study, we examined how the diversity of memories in group members works for group decision-making with the following methods. First, we conducted a behavioral experiment about memories of city names. Using these data (i.e., actual memory data), we examined the accuracies of inferences made by hypothetical people who made inferences based on a memory-based simple heuristic. As an inference task, we used binary choice inference problems about population sizes (e.g., "Which city has a greater population size, Tokyo or Chiba?"). For this task, people tend to rely on memory-based simple heuristics such as recognition (Goldstein & Gigerenzer, 2002), fluency (Hertwig, Herzog, Schooler, & Reimer, 2008), or familiarity (Honda, Abe, Matsuka, & Yamagishi, 2011; Honda, Matsuka, & Ueda, 2017; Xu, González-Vallejo, Weinhardt, Chimeli, & Karadogan, 2018). Thus, people's memories will affect the inference processes for this kind of problem. Finally, we

constructed a group of such hypothetical people and examined the performance of group decision making.

How can memory diversity be generated? Given that the present study used city names as stimuli, we predicted that constructed memories about city names (e.g., recognitions of or familiarities with city names) were more dissimilar (i.e., diverse) between people in different areas than between those in the same area. Based on this consideration, we recruited participants from two areas (Tokyo and Osaka).

In the following section, we shall report two studies: a behavioral experiment and a computer simulation.

Study 1: Behavioral experiment

We conducted a behavioral experiment about memories of 30 cities in Japan. We examined whether recognitions and familiarities regarding the 30 cities differed depending on the area participants lived in and analyzed the memory diversity.

Method

Participants We recruited participants in their 30s and 50s from two areas, Tokyo and Osaka, with the following definitions: first, they were born in Tokyo (or Osaka); second, they had lived in Tokyo (or Osaka) for more than 20 years in total; and third, they had been living in Tokyo (or Osaka) during the past five years. As a result, we recruited 99 people in their 30s in the Tokyo area (M_{age} = 35.48, SD_{age} = 2.76, n_{female} = 49), 101 people in their 50s in the Tokyo area (M_{age} = 54.74, SD_{age} = 2.51, n_{female} = 50), 99 people in their 30s in the Osaka area (M_{age} = 35.15, SD_{age} = 2.92, n_{female} = 50), and 101 people in their 50s in the Osaka area (M_{age} = 53.92, SD_{age} = 2.89, n_{female} = 51). In total, 400 Japanese participated in the experiment.

Tasks, materials, and procedure We conducted a recognition task and measurement of familiarity. In the recognition task, participants were presented with a city name and answered whether they knew the city. When participants knew the presented city, they were also asked about their level of familiarity with the city. They answered this question using a scale labeled "I know only the name" on the far left and "I know a lot" on the far right. This rating was recorded with 100 points ranging from 1 (I know only the name) to 100 (I know a lot) depending on the familiarity level. In these two tasks, we used 30 Japanese cities based on Honda et al. (2017). 15 of the 30 cities were from the difficult list, and the other 15 were from the easy list (see Appendix for the specific city names). The definition of "difficulty" for the list lies in the difficulty of binary choice inferences about population size (Honda et al., 2017). Since memory-based heuristics in group decision making can work differently depending on the inference problems (see Fujisaki et al., 2018), we used these 30 cities. We conducted the two tasks on the Internet. Each city name was presented individually. The presentation order of the 30 cities was randomized for each participant.

Results and discussion

First, we examined the similarities of memories. In this examination, we calculated Spearman's correlation coefficient for familiarity ratings between two participants. We used the correlation coefficient as the criterion of similarity for memories between the two participants. We examined the differences in similarities as functions of area and age. As Table 1 shows, we examined the distributions of correlation coefficients in 10 pairs of participants each for easy and difficult lists. For example, in the "Tokyo30s–Tokyo30s" pair, since there were 99 participants in their 30s in the Tokyo area, there were 4851 (99 × 98 / 2) pairs at most. In some cases (14 out of 800[400 participants × 2 lists]), participants provided the same familiarity ratings for 15 cities in a list. For this case, we excluded the data since we could not calculate correlation coefficients.

Table 1 shows the distributions of correlation coefficients as a function of pair type. For each pair, we estimated a 95% confidence interval of the mean based on bootstrapping using 5000 simulations. Familiarity ratings between two participants became more similar in pairs of individuals from the same area than different areas, supporting our prediction. In contrast, we did not find a specific trend of similarity in terms of the age difference.

Next, we analyzed the similarity of memories in terms of ecological rationality (Gigerenzer & Todd, 1999). In this analysis, a participant was assumed to make inferences based on her/his memory as follows: s/he was presented with a pair of cities and made binary choice inference about population size (i.e., inferred which city had a greater population size). In making inferences, s/he used memory-based simple heuristics. We assumed that s/he used the familiarity heuristic (Honda, et al, 2011, 2017; Xu, et al., 2018). In this heuristic, s/he inferred that the more familiar city had the larger population size. In Honda et al. (2017), for the inference in pair x, person i's decision (D) is defined as follows:

$$D_i(x) = c_i(FA_{iL} - FA_{iS}) \quad (1)$$

where FA_{iL} and FA_{iS} represent familiarities for the larger and smaller cities in pair x, and c_i represents the scaling parameter. This scaling parameter for each person was selected so that the maximum or minimum value of D became 1 or −1. This model predicts that, when $D(x)$ is larger than 0 and satisfies the decision threshold (i.e., $D[x]$ > decision threshold), person i infers that the larger city has the larger population and that, when it is smaller than 0 and satisfies the decision threshold (i.e., $-D[x]$ > decision threshold), person i infers that the smaller city has the larger population. In pairs in which participants could recognize only the larger (or smaller) city, $D(x)$ was set as 1 (or −1) so that they choose the larger (or smaller) city. This choice is consistent with the recognition heuristic (Goldstein & Gigerenzer, 2002), indicating that the familiarity heuristic model can explain inference patterns predicted by the recognition heuristic.

We then examined how accurate people's memory-based inferences were and discussed the diversity of memory from this perspective. In this examination, we set two criteria, validity and discrimination rates (Gigerenzer & Todd, 1999). The validity rate is defined as follows:

$$V = \frac{H_c}{H_c + H_i} \quad (2)$$

where H_c (or H_i) denotes the number of pairs for which a person can use heuristic (i.e., $D[x]$ exceeds the decision threshold) and heuristic-based inference resulted in the correct (or incorrect) inference. That is, the validity rate means the accuracy of the familiarity heuristic. In contrast, the discrimination rate means the proportion of pairs in which a person can use the familiarity heuristic.

We calculated the validity and discrimination rates for all 105 pairs in difficult and easy lists for each participant. In this calculation, we set the decision threshold as 0.3 based on the empirical findings in Honda et al. (2017). Figure 1 shows the distributions of validity and discrimination rates for the two lists. We conducted 2 (area; Tokyo and Osaka) × 2 (age; 30s and 50s) ANOVA for the two criteria (i.e., validity and discrimination rates) and the two lists, respectively.

As for the validity rate, in the difficult list, a significant main effect of area was observed [$F(1, 388) = 111.49$, $p < .001$, $\eta^2 = 0.223$], indicating that the familiarity heuristic by participants from the Tokyo area would have led to more accurate inferences ($M_{Tokyo} = 0.680$, $M_{Osaka} = 0.488$). No significant main effect of age [$F(1, 388) = 0.09$, $p = .77$, $\eta^2 = 0.00$] or interaction [$F(1, 388) = 0.95$, $p = .33$, $\eta^2 = 0.00$] was observed. In the easy list, a significant main effect of area was observed [$F(1, 382) = 8.09$, $p = .005$, $\eta^2 = 0.223$], indicating that the familiarity heuristic by participants from the Osaka

Table 1. Distribution of correlation coefficients for familiarity rating. The range (95% confidence interval) was estimated by bootstrapping with 5000 simulations.

		Difficult list			Easy list		
		95% confidence interval			95% confidence interval		
Pair	Area	Lower bound	Mean	Upper bound	Lower bound	Mean	Upper bound
Tokyo30s–Tokyo30s	Same	0.197	0.205	0.212	0.191	0.199	0.207
Tokyo30s–Tokyo50s	Same	0.213	0.218	0.223	0.228	0.233	0.238
Tokyo50s–Tokyo50s	Same	0.234	0.241	0.248	0.283	0.290	0.297
Osaka30s–Osaka30s	Same	0.435	0.442	0.448	0.269	0.277	0.284
Osaka30s–Osaka50s	Same	0.471	0.475	0.479	0.251	0.256	0.261
Osaka50s–Osaka50s	Same	0.512	0.518	0.523	0.233	0.241	0.248
Tokyo30s–Osaka30s	Different	0.176	0.181	0.187	0.038	0.044	0.050
Tokyo30s–Osaka50s	Different	0.174	0.179	0.184	0.007	0.012	0.018
Tokyo50s–Osaka30s	Different	0.175	0.181	0.186	0.018	0.023	0.029
Tokyo50s–Osaka50s	Different	0.177	0.182	0.187	-0.004	0.001	0.007

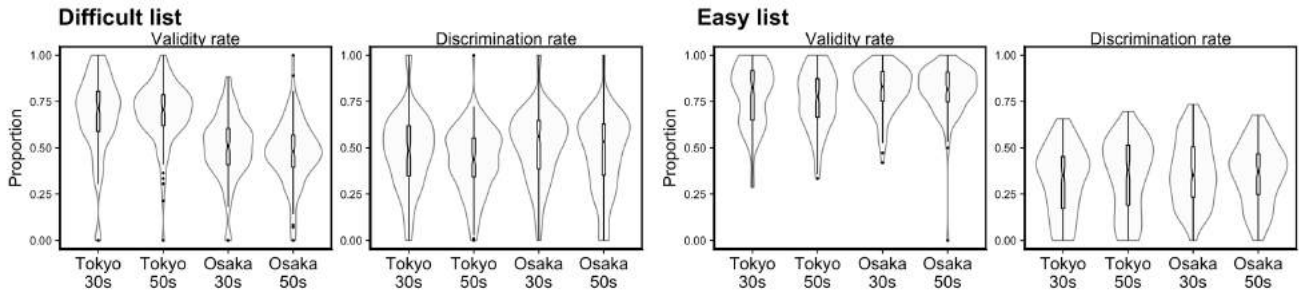


Figure 1. Validity and discrimination rates of the familiarity heuristic.

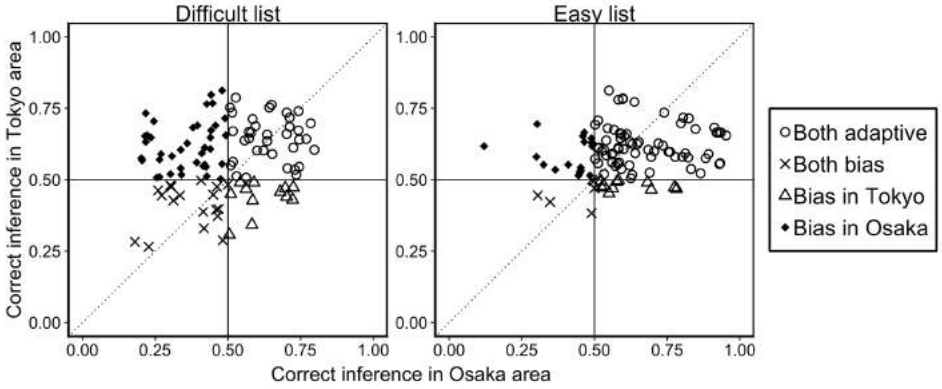


Figure 2. Proportions of correct inferences for Osaka's and Tokyo's participants. Each point denotes the proportion of correct inferences for each inference problem [i.e., there are 105 (15 × 14 / 2) points in each list].

area would have led to more accurate inferences (M_{Tokyo}= 0.771, M_{Osaka} = 0.814). No significant main effect of age [F(1, 382) = 0.99, p = .32, η^2 = 0.00] or interaction [F(1, 382) = 0.32, p = .57, η^2 = 0.00] was observed.

As for the discrimination rate, in the difficult list, a significant main effect of age was observed [F(1, 396) = 5.32, p = .02, η^2 = 0.01], indicating participants in 30s could have potentially used familiarity heuristic more often than those in 50s (M_{30s}= 0.499, M_{50s} = 0.454). No significant main effect of area [F(1, 396) = 3.30, p = .07, η^2 = 0.01] or interaction [F(1, 388) = 0.02, p = .89, η^2 = 0.00] was observed. In the easy list, no significant effects were observed [main effect of age, F(1, 396) = 0.11, p = .75, η^2 = 0.00; main effect of area, F(1, 396) = 1.76, p = .19, η^2 = 0.00; interaction, F(1, 388) = 1.01, p = .32, η^2 = 0.00].

The above analyses indicated that, although similarities of familiarity ratings and the nature of ecological rationality differed depending on the areas, ages were not generally related. Thus, in the following analyses, we merged the data between the two generations.

Next, we analyzed the accuracy of the familiarity heuristic for each inference problem. Figure 2 shows the relationship of correct inference between the two areas. Each figure includes 105 points, each of which shows the proportion of correct inference for each problem. Depending on the relationships about inference adaptivity (i.e., proportions of correct inferences were above the chance level [0.5] or not), we named pairs "Both adaptive," "Both bias," "Bias in Tokyo," and "Bias in Osaka." If the participants in the two areas show the same adaptivity or bias, each point will lie on the diagonal line (i.e., proportions of correct inferences correspond with each other). However, as is apparent in the figure, this was not true; the proportions of correct inferences varied depending on the areas, and the relationship of correct inferences between the two areas was not strong (in the difficult list, r = 0.18, p = .07; in the easy list, r = 0.19, p = .05). Furthermore, there were nonnegligible cases of "Bias in Tokyo" or "Bias in Osaka," indicating that participants in each area showed opposite direction of inference accuracy.

Altogether, we found that accuracy of the familiarity heuristic varied depending on the participants' profiles. In particular, the area (i.e., Tokyo or Osaka) was highly related to the accuracy of the familiarity heuristic, indicating that ecological rationality of memory differed depending on the area participants were from. Thus, constructed memory in the two areas were diverse.

Study 2: Computer simulations

We conducted computer simulations about group decision making based on the behavioral experiment data. We constructed hypothetical groups that comprised participants in the behavioral experiment, and the groups made inferences about population. Then, we compared group performance in terms of diversity of group members (i.e., members from only Tokyo or Osaka or members from a mixture of Tokyo and Osaka).

Method

Group construction We set group size at 5, 10, 20, or 50. In constructing a group, we randomly selected group members from participants in the behavioral experiment. Groups were constructed from a single area (i.e., participants from only Tokyo or Osaka) and both two areas (i.e., mixture of participants from Tokyo and Osaka).

Group decision making We set the following hypothetical group decision making situation. Group members made binary choice inferences about population size. They were presented with a pair of cities and made binary choice inference about population size (i.e., inferred which city had a greater population size). Here, each member was assumed to make inferences based on the familiarity heuristic, and the group made decisions based on simple majority rule (Hastie & Kameda, 2005; Sorkin, Hays, & West, 2001). According to previous assumptions (Fujisaki et al., 2018), when a member could not make an inference (i.e., her/his inference did not exceed the decision threshold), s/he did not participate in the group decision making. Furthermore, when a group could not make decisions (i.e., an equal number of members chose different cities), the group randomly chose one city.

Procedure The group made decisions for all 105 inference problems for the two lists. For each parameter setting (i.e., group size or diversity of group members), we constructed, in total, 5000 different groups based on random selection of members. We regarded the average of proportion of correct inference in the 5000 groups as the group performance in each parameter setting.

Results and discussion

First, we examined the performance in the single-area group (i.e., group members comprised participants from only Tokyo or Osaka). Figure 3 shows results of computer simulations. This shows the proportion of correct inferences for 105 inference problems each in the difficult and easy lists. Our findings can be summarized with the following three points. First, when individuals showed accurate inferences on average (i.e., proportion of correct inferences exceeded the chance level), group decision making enhanced accurate inferences. Second, when individual inference showed biases on average (i.e., proportion of correct inferences fell below the chance level), group decision making deteriorated accurate inferences (see Osaka performance in the difficult list). Third, and most importantly, individual performance did not always predict the better boost of group decision making. See the group performance in the easy list. At the individual level, members in Osaka showed more accurate inferences than those in Tokyo (see group size 1 in Figure 3). Intuitively, the group that comprises Osaka members seemed to show better group performance than that comprising Tokyo members since participants in Osaka showed more accurate inferences at the individual level. However, this was not true, and the group of participants from Tokyo performed better than the group of participants from Osaka. This counter-intuitive phenomenon may occur because of the biases (Fujisaki et al., 2018). Regarding the problems wherein people have bias (i.e., mean

proportion of correct inference lies below the chance level [0.5] at the individual level), group decisions deteriorate accurate inferences, and the mean proportion of correct inferences reaches 0 as the number in the group increases. Actually, out of the 105 problems on the easy list, the proportion of biased problems for participants in Osaka was 0.234, and that for participants in Tokyo was 0.162. Thus, although participants in Osaka showed more accurate inferences on average, they also showed more biases. Thus, in group decision-making, inaccurate inferences were enhanced for more inference problems in Osaka than in Tokyo, and a counter-intuitive phenomenon occurred.

Next, we examined the performance of decisions in groups whose members were diverse (i.e., mixture of participants). Figure 4 shows the performance of decisions for these groups. The x-axis indicates the proportion of members from Tokyo (i.e., 1 – the proportion is the proportion of members from Osaka). Thus, the values 0 and 1 indicate that the group includes members from only a single area (i.e., the values correspond to those in Figure 3 in each parameter setting). On the difficult list, the proportion of correct inferences in groups was boosted as the proportion of members from Tokyo increased. Since individual inferences in participants from Osaka were not accurate (i.e., their inferences were almost chance level), members from Tokyo boosted accurate inferences. On the easy list, the findings were highly intriguing. The peak of the group performance did not lie in the endpoint (i.e., group comprised members from a single area) but in the group that comprised members from the two areas. That is, when the group included diverse members, the group reached the highest performance.

In sum, we found that, when memories of group members were diverse, collective decisions by the group could be more accurate in some decision situations (e.g., when making collective decisions for the inference problems on the easy list).

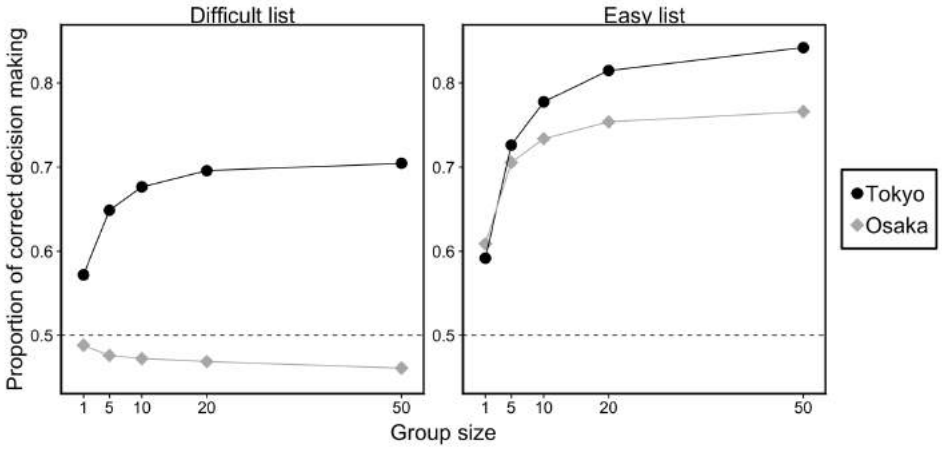


Figure 3. Performance of group decision making (i.e., proportions of correct inferences for 105 inference problems on the difficult and easy lists) in the group whose members were from a single area (i.e., Tokyo or Osaka). Group size 1 indicates the mean proportions of correct inferences in individual inferences. The dotted line indicates the chance level of inferences.

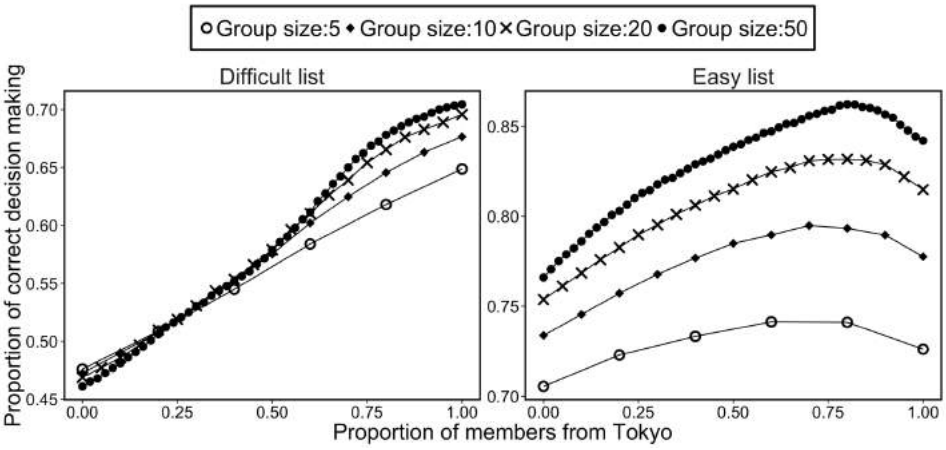


Figure 4. Performance of group decision making (i.e., proportions of correct inferences for 105 inference problems on the difficult and easy lists) in the group whose members were from two areas (i.e., Tokyo or Osaka).

General discussion

Through a behavioral experiment and computer simulations, we found that diverse memories in group members enhanced accurate group decision making.

How was the effect of member diversity generated? The key was the biases. As Figure 2 shows, participants in each area had unique biases (i.e., "Bias in Tokyo" and "Bias in Osaka" in Figure 2). In the mixed group, these biases could be improved by members from different areas, leading to accurate inferences.

Finally, we note the following two points about the difference in the adaptive nature of inferences between individual and group decision making levels. First, adaptive heuristics at the individual level do not indicate that such heuristics also boost accurate inferences in group decision making (see Figure 3 regarding the easy list) since adaptive heuristics are accompanied by some biases. That is, group decision making can boost both accurate and inaccurate inferences. Second, such problems in group decision making can be resolved by the diversity of inferences. In the present study, we showed that diversity in memories could remedy individual biases. Diverse memories can produce different inferences even when people use the same heuristic. That is, people make inferences using superficially "different" strategies. This is basically consistent with previous findings that diverse inference strategies used by group members can boost group decision making (Fujisaki et al., 2018). These findings suggest that diversity in inferences plays a key role in improving biases produced by individual inferences.

Acknowledgments

This research was supported by JSPS KAKENHI Grant Numbers 18H03501 for the first author and 16H01725 for the last author

References

Fujisaki, I., Honda, H., & Ueda, K. (2018). Diversity of inference strategies can enhance the 'wisdom-of-crowds' effect. *Palgrave Communications*, *4*: 107.

Galesic, M., Barkoczi, D., & Katsikopoulos, K. (2018). Smaller crowds outperform larger crowds and individuals in realistic task conditions. *Decision*, *5*, 1–15.

Gigerenzer, G., & Todd, P. M. (1999). Fast and frugal heuristics: The adaptive toolbox. In G. Gigerenzer, P. M. Todd, & The ABC Research Group (Eds.), *Simple heuristic that make us smart* (pp. 3–34). NY: Oxford University Press.

Gigerenzer, G., Todd, P., & The ABC Research Group. (1999). *Simple heuristics that make us smart*. NY: Oxford University Press.

Goldstein, D. G., & Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. *Psychological Review*, *109*, 75–90.

Hastie, R., & Kameda, T. (2005). The robust beauty of majority rules in group decisions. *Psychological Review*, *112*, 494–508.

Hertwig, R., Herzog, S. M., Schooler, L. J., & Reimer, T. (2008). Fluency heuristic: A model of how the mind exploits a by-product of information retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 1191–1206.

Honda, H., Abe, K., Matsuka, T., & Yamagishi, K. (2011). The role of familiarity in binary choice inferences. *Memory and Cognition*, *39*, 851–863.

Honda, H., Matsuka, T., & Ueda, K. (2017). Memory-based simple heuristics as attribute substitution: Competitive tests of binary choice inference models. *Cognitive Science*, *41(S5)*, 1093–1118.

Jönsson, M. L., Hahn, U., & Olsson, E. J. (2015). The kind of group you want to belong to: Effects of group structure on group accuracy. *Cognition*, *142*, 191–204.

Lorenz, J., Rauhut, H., Schweitzer, F., & Helbing, D. (2011). How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences*, *108*, 9020–9025.

Luan, S., Katsikopoulos, K. V., & Reimer, T. (2012). When does diversity trump ability (and vice versa) in group decision making? A simulation study. *PLOS ONE*, *7*: e31043.

Mavrodiev, P., Tessone, C. J., & Schweitzer, F. (2013). Quantifying the effects of social influence. *Scientific Reports*, *3*, 1360.

Sorkin, R. D., Hays, C. J., & West, R. (2001). Signal-detection analysis of group decision making. *Psychological Review*, *108*, 183–203.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*, 207–232.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*, 1124–1131.

Xu, P., González-Vallejo, C., Weinhardt, J., Chimeli, J., & Karadogan, F. (2018). Use of the familiarity difference cue in inferential judgments. *Memory and Cognition*, *46*, 298–314.

Appendix

The two lists used in the present study. We used these lists based on Honda et al. (2017).

Easy list	Difficult list
Yokohama-shi	Kawaguchi-shi
Osaka-shi	Machida-shi
Nagoya-shi	Kohriyama-shi
Sapporo-shi	Takasaki-shi
Kobe-shi	Tsu-shi
Kyoto-shi	Sasebo-shi
Fukuoka-shi	Hachinohe-shi
Hiroshima-shi	Matsumoto-shi
Sendai-shi	Hitachi-shi
Chiba-shi	Yamaguchi-shi
Niigata-shi	Takaoka-shi
Hamamatsu-shi	Imabari-shi
Kumamoto-shi	Miyakonojo-shi
Okayama-shi	Ogaki-shi
Kagoshima-shi	Ashikaga-shi

A Model-Based Investigation of the Biological Origin of Human Social Perception of Faces

Sophia J. Huang and **Chaitanya K. Ryali** and **Jianling Liu** and **Dalin Guo**
Jinyan Guan and **Yvonne Li** and **Angela J. Yu**
University of California San Diego
9500 Gilman Drive La Jolla, CA 92093 USA

Abstract

Humans readily form social impressions of faces at a glance, whether assessing trustworthiness, attractiveness, or dominance. However, little is understood about how such computations are carried out neurally. Here, we leverage a computational model of human face perception to quantify and characterize the extent to which macaque monkey face patch neurons encode information relevant for social trait perception. Specifically, we use a social trait prediction model to estimate the social trait ratings for face stimuli viewed by monkeys during a neural recording experiment. We find that, while the monkey face patch neurons are linearly tuned to facial features different from those used by humans to make social judgments, the subspace spanned by the face patch neurons and the subspace spanned by the facial features supporting human social perception are highly overlapping. This result implies that the information present in the monkey face patch neurons are largely sufficient, after *linear decoding*, to support human social perception, thus shedding light on the biological origin of human social processing of faces.

Keywords: face perception; social perception; representation; neural recording; face modeling

Introduction

Face processing plays a special role in human life, as it underpins social interactions essential for survival and reproductive success (Olivola, Funk, & Todorov, 2014). Psychological studies have shown that humans effortlessly and consistently derive social characteristics (social, demographic, emotional traits) from the appearance of faces of strangers (Willis & Todorov, 2006). However, little is known about how such assessments are represented or computed in the brain. In this work, we leverage a computational model of human face perception (Guan, Ryali, & Yu, 2018) to quantify and characterize the extent to which face patch neurons in the macaque monkey brain (Freiwald & Tsao, 2010) encode information relevant for human social perception of faces.

One challenge for studying the relationship between neural responses and human face perception is that human face images are high dimensional and vary among each other in complex ways. To parameterize the space of human face images, we adapt a popular computer vision algorithm, the Active Appearance Model (AAM) (Cootes, Edwards, & Taylor, 2001; Valentine, 1991). AAM provides a vector space representation of face images with several desirable properties. Firstly, this representation is sufficiently rich such that each face image corresponds to a unique point in this space. Secondly, AAM is capable of generating realistic face images, helping to visualize the features encoded by neurons or group of neurons. Thirdly, recent neural data suggest that face patch neurons encode facial features similar to those in AAM (Chang & Tsao, 2017). Here, we train our own version of the AAM model (Guan et al., 2018) using a publicly available face dataset (Bainbridge, Isola, & Oliva, 2013). This dataset also contains human ratings along 20 social trait dimensions, which we model linearly by regressing the trait ratings against AAM latent features. Similarly, we linearly model the classification of gender and age based on human judgments of these qualities on the same face dataset (Bainbridge et al., 2013).

The neural data we analyze are single cell recordings from the face patch areas of the macaque monkey, recorded while the animals viewed 37 human face images (Freiwald & Tsao, 2010) (the original dataset contained 41 face images, in 4 of which the person's eyes are fully or partially closed – these 4 are excluded from our analyses). The face patch areas of the monkey inferotemporal (IT) cortex have been shown to contain neurons that are highly selective for faces (Freiwald & Tsao, 2010). Although face images used in the monkey experiment have not been rated by human subjects for social traits, we can predict the ratings by projecting the face stimuli to our AAM model, and then use the pre-trained regression models to predict the social ratings (Guan et al., 2018).

In the following, we first define and compute each neuron's Linear Response Axis (LRA), the linear axis within AAM that best captures the tuning selectivity of a neuron. We then characterize the properties of the LRA's both individually and as a population. Finally, we compare the facial features encoded by the neuronal LRA's versus those necessary for human social perception.

Results

Predicting Social Trait Perception

In order to predict human social perception of the faces seen by the monkeys, we utilize a model we recently developed based on the Active Appearance Model (AAM) (Guan et al., 2018). The model obtains a latent vector space representation of face images, consisting of combined principal components of shape and texture features (see Methods). We then use linear regression to model how latent features of a face give rise to trait ratings (20 social traits as in (Bainbridge et al., 2013),

plus the demographic traits gender and age, see Methods). We find that this approach predicts human social ratings on a *novel face* better than other humans' rating on the *same face*; it also achieves comparable performance to the state-of-the-art convolutional deep neural network, but has the advantage of having better interpretability (Guan et al., 2018).

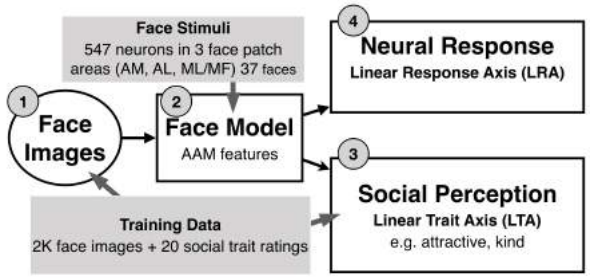


Figure 1: The face model is a vector space representation, whose axes represent the facial features that vary among face images, and the mean of the training dataset sample is centered at the origin (by design). We utilize the publicly available dataset (Bainbridge et al., 2013) (1) to train the AAM face model (2), obtaining 60 facial features. Using the same datasets, we model human social perception (3), and estimate the facial information encoded by each neuron (4).

Here, we can predict the human social ratings of the faces viewed by the monkeys by projecting these face images into the pre-trained AAM model. We first obtain the landmarks of the face stimuli using the free software Face++ (https://www.faceplusplus.com), then projecting them into the pre-trained AAM model (Guan et al., 2018) (Figure 1). Each face stimulus is a point in a 60-dimensional latent space. Figure 2 shows an example face image viewed by monkeys. We then obtain the predicted social perception of each face stimulus using the pre-trained Linear Trait Axes or LTA's (see Methods). The LTA of a trait represents the linear combination of facial features that maximally modulates human perception of this trait (a similar variation in facial feature along any other axis will induce a smaller change in average human perception). For example, the face in Figure 2A is predicted to be slightly more than 1 standard deviation more attractive than the average face (in the training data); Figure 2C shows predicted social ratings a number of traits.

One question we want to answer is how much information related to each social trait is encoded in the neural responses of the monkey face patch neurons. To have sufficient statistical power to assess this, we first need to make sure that the 37 face images span a substantial portion of the predicted trait ratings. This is indeed the case, as can be seen in Figure 3 for "happy" and "attractive." Figure 3A.ii visualizes a pair of face images seen by the monkeys that are predicted by the model to be less (left) or more (right) happy, and another pair (Figure 3B.ii) that is predicted to be less (left) or more (right) attractive. They are consistent with visual intuition.

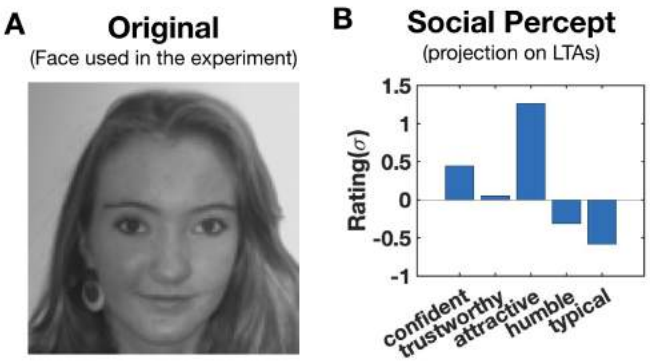


Figure 2: Face representation and social trait estimation. (A) An example face image viewed by the monkeys. (B) 5 examples (out of 20) of predicted social trait ratings for the same face.

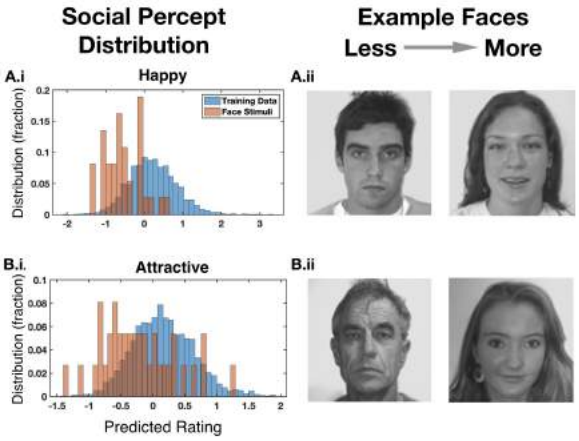


Figure 3: Social trait rating prediction. The histograms (left) of predicted social traits and two face stimuli (right) that are predicted by the model to vary in (A) happiness and (B) attractiveness. Distribution of predicted "happy" rating (A.i) for the 37 face stimuli (red) and training data (blue).

To quantify the information related to human social perception encoded by the macaque face patch neurons, we compute the correlation coefficient between each neuron's mean firing rate for each face (see Methods) and the predicted trait rating of each face. A neuron is deemed to significantly encode a trait if its correlation coefficient has a p-value < 0.05. We find that 19 out of 22 traits are encoded by a significant fraction of the neural population (Figure 4).

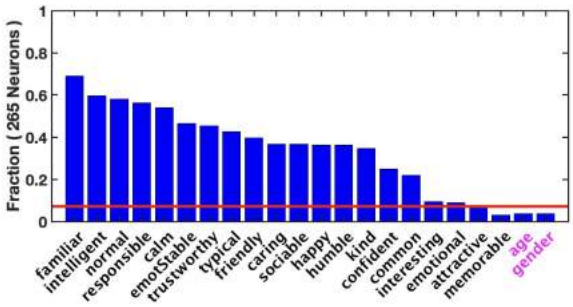


Figure 4: Proportion of neurons significantly encoding various social and demographic traits. A neuron is considered to significantly encode a trait if its MFR has significant correlation ($p < 0.05$, corrected for multiple comparison) with predicted ratings of viewed faces for that trait. The red line indicates the threshold for determining whether a significant (non-zero) fraction of neurons encode a trait at the significance level of $\alpha = 0.05$). This analysis only consists of the 265 neurons whose responses we can statistically reliably model (see subsection on Linear Response Axis)

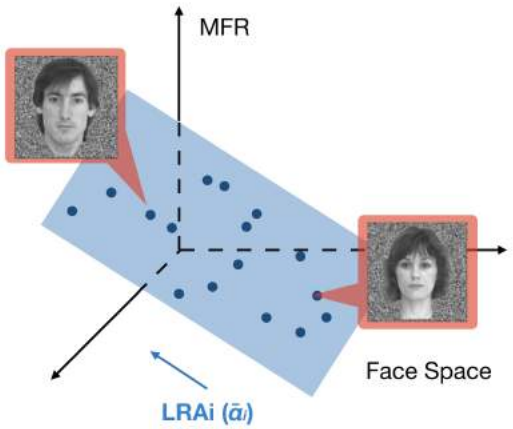


Figure 5: Schematic illustration of Linear response axis (LRA). The blue dots represent MFR of a neuron for different face images. The blue hyper-plane is the best linear fit of the neuron's response to those face stimuli. LRA gives the axis in the face space that yields the largest linear gradient for this neuron's MFR.

Linear Response Axis

To quantify the featural selectivity of each face patch neuron, we first define and compute the Linear Response Axis (LRA) of each neuron (see Figure 5A), which is just the normalized regression coefficient vector. Each LRA is obtained by regressing each neuron's mean firing rate (MFR) against the first $k = 13$ latent features of each image in the AAM space. k is chosen to be 13 in order to maximize the number of neurons whose response we can reliably estimate (i.e. significant correlation between model-predicted MFR and observed MFR on held-out faces, see Methods). For $k = 13$, we find that we can reliably estimate the LRA of 265 neurons – unless otherwise noted, all subsequent LRA-based analyses are performed using only these 265 neurons. The LRA specifies the linear axis that maximally accounts for variations in the neural response. We find that the average neural response along the LRA is not only monotonically increasing, as found in (Chang & Tsao, 2017), but in fact highly linear in this data set; and like in (Chang & Tsao, 2017), the neural response to the principal axis is completely flat. This replicates the finding in (Chang & Tsao, 2017) that monkey face patch neurons encode single axes in the AAM latent feature space.

While Figure 3 quantifies the relationship between social traits and individual neuron's MFR, we are also interested in characterizing the facial features encoded by the neural population as a whole. Naively, we might do so by applying principal component analysis (PCA) to the LRA's. However, the LRA's compose a special sort of data, namely unit-length vectors that lie on a hypersphere. If the LRA's lie in a completely balanced manner (by balanced, we mean that for each LRA, there is an "opposite" LRA that points approximately in the opposite direction, so that the two neurons encode the same AAM axis but have opposite preferences), then PCA would pull out the main directions encoded by neural LRA's; but if they are highly imbalanced, then PCA would instead pull out something like the tangent subspace and yield something uninterpretable. We therefore add an *opposite* LRA to each estimated LRA, to artificially balance the LRA's, and then apply PCA. We find that PC 1 alone explain 48.2% of the total variance among the LRA's, and the first 9 PC's explain 95% of the variance among the LRA's (Figure 6). Relative to the other features, the first PC plays an outsized role in terms of the features that the neurons linearly encode.

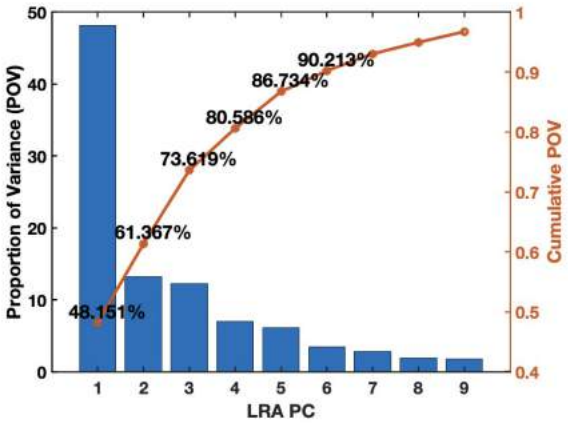


Figure 6: Incremental and cumulative proportion of variance explained by the PCs of neural LRAs. The histogram indicate the explained variance by each LRA PC and the plot for the accumulated explained variance.

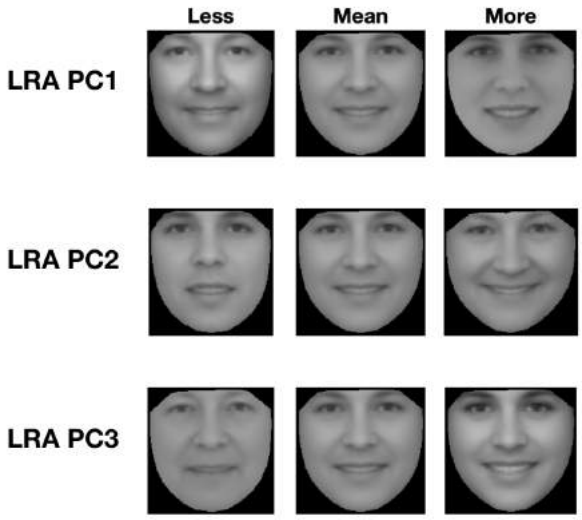


Figure 7: Synthetic face images along the top three LRA PC's. Each row shows how the mean face changes when positive (right) and negative (left) values are added to the mean face along each PC.

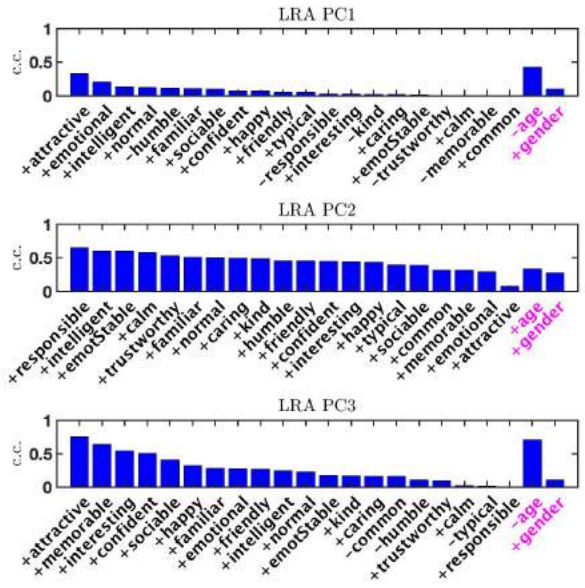


Figure 8: Expected correlation (i.e. dot product) between neural LRA PC's and social trait LTA's. Each row indicate expected c.c. for the various traits for each PC (blue: social trait, magenta: demographic trait). The bars indicate the absolute value of the correlation coefficients while the sign of the correlation is represented by + and - sign in front of the name of traits on x-axis. The traits are ordered in descending order of expected c.c., separately for social and demographic trait.

To get a sense for the primary featural axes that face patch neurons encode, we generate synthetic faces along each of the first three PC's (Figure 7). The first row of Figure 7 shows how the middle face changes as it gains more positive (right face) or more negative (left face) value along the first LRA PC; the next two rows show the same for LRA PC 2 and PC 3, respectively. The faces undergo interesting holistic changes along each of the first three PC's, consisting of some age- and gender-related changes but also other harder-to-verbalize structural changes.

To gain a more quantitative understanding of what the major features the face patch neurons encode as a population, we compute the expected correlation between each LRA PC and social trait (Figure 8), which is just the dot product between each LTA and LRA PC (they are both unit lengths). We find that all three LRA PCs significantly correlate with age. The expected correlation coefficient (dot product) between age LTA and each of LRA PC1, PC2, and PC3 ρ=.48, ρ=-.32, ρ=.67, respectively. In addition, PC1 and PC3 are correlated with attractiveness (PC1: ρ=.35, PC2: ρ=.75), and PC2 correlated with responsible (ρ=.67). This shows that while the neural population as a whole encode features that are highly correlated with those important for human social perception, the most important featural dimension (PC 1) has a poor correlation with any of the human social traits that we considered.

Figure 9 illustrates yet another way to visualize the relationship between neural LRA's and human LTA's. It shows a scatterplot of all the neural LRA's (red), the "pposite LRA's" (gray), and the social LTA's (green) projected into the subspace spanned by the Attractive and Responsible LTA's, the two traits that neurons as a population linearly encoded the most information about. We see that, first of all, that most of the social LTA's are fairly close to unit length within this subspace, indicating that most of them point in a direction very close to this 2D subspace. The neural LRA's show a range of distances from the origin, with the majority lying close to the origin, indicating they primarily point in a direction far away from this 2D plane that is quite important for human social perception. The neurons that do have LRA's pointing close to this subspace (distance close to 1 from the origin) are mostly pointing in the direction of traits such as familiar, intelligent, and normal – the traits that have the greatest number of significant correlation with neurons (Figure 4).

Conversely, we can also visualize all the projected LTA's and LRA's in the subspace spanned by LRA PC 1 and PC 2 (Figure 10). Here, we see that the neural LRA's are highly imbalanced, with most LRA having a positive projection along PC 1. We also see that most human LTA's point in a direction far away from PC 1, but have a fairly large component pointing in the direction of PC 2 (the exception is Attractive, which has the opposite pattern).

Similarly, we can also visualize all the projected LTA's and LRA's in the subspace spanned by LRA PC 1 and PC 3 (Fig-

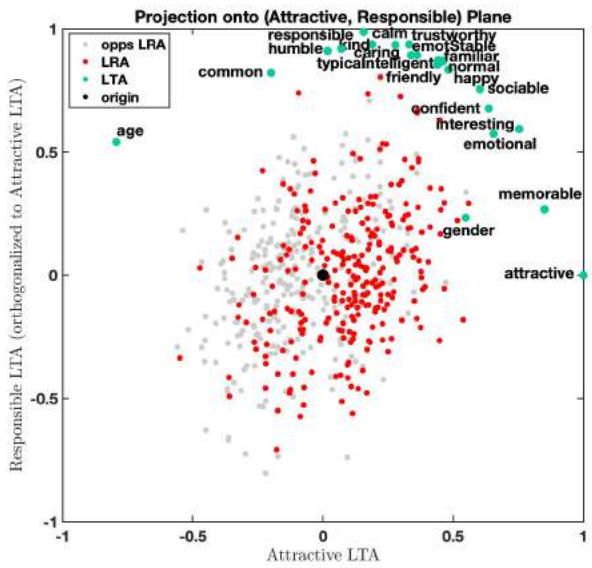


Figure 9: Projection of LRA's and LTA's onto the plane spanned by Attractive and Responsible LTA's. Red dots are projections of LRA's (gray for "opposite" LRA's). Green dots are projections of LTAs'. The label next to each green dot indicates the trait.

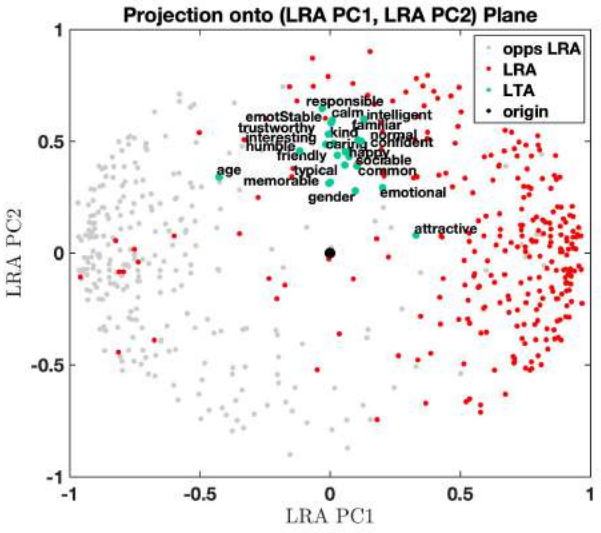


Figure 10: Projection of LRA's and LTA's onto LRA PC1 and LRA PC2. Similar formatting as Figure 9.

ure 10). It is apparent that Attractive and Age have fairly large components pointing along PC 3, along with traits such as memorable, interesting, and confident.

Subspace Comparison

While the previous analyses suggest that there is some overlap in the facial features that are encoded by monkey face patch neurons, and those that matter for human social trait perception, here we quantify their overlap in a different way. We compute how well (model-predicted) human social perception can be computed from the information present in the monkey face patch neurons (via simple linear decoding), and vice versa. As shown in Figure 12, the LTA-predicted ratings of the 22 social traits can be almost perfectly recovered from the LRA-predicted neural response (265 neurons) to face images (R^2 very close to 1); conversely, we find that the LRA-predicted response of all 265 neurons can be perfectly recovered from the LTA-predicted social trait ratings (22 traits), where $R^2 = 1$ in every case. This result suggests that facial featural information present in the macaque face patch areas is largely the same as those necessary for human social perception.

Methods

The Face Model: AAM

The Face model is an instantiation of the Active Appearance Model (Cootes et al., 2001; Guan et al., 2018). Each face image has shape and texture features. The shape features consist of the (x, y) coordinates of a set of landmarks that are consistently defined across faces. The texture features are the pixel values (grayscale) of each face image after warping it to have the same landmark locations as the averaged face. To reduce the dimensionality and remove correlation between shape and texture features, we perform additional Principal Component Analysis (PCA) on shape and texture features and retain the first 60 PC's, resulting in a 60-dimensional AAM feature space. AAM features form the basis of the Face Model that jointly describes the variations of shape and texture of the faces.

Social Trait Perception: Linear Trait Axis (LTA)

The Linear Trait Axis (LTA) $\tilde{\beta}$ for each social trait is computed as the normalized regression coefficients of ratings regressed against AAM features:

$$y = \beta\vec{x} + \varepsilon$$

where y is the standardized ratings for the trait, $\vec{x}$ is the AAM features, and β is the vector of regression coefficients. The linear trait axis (LTA) is defined as

$$\tilde{\beta} = \frac{\beta}{||\beta||}$$

The LTA specifies a direction in the face space that would (linearly) maximally alter the perception of the trait.

Predicted social perception A novel face images can be projected into the trained face model, resulting in a 60-dimensional representation $\vec{x}$. The predicted rating of a face image is then given by

$$\hat{y} = \beta\vec{x}$$

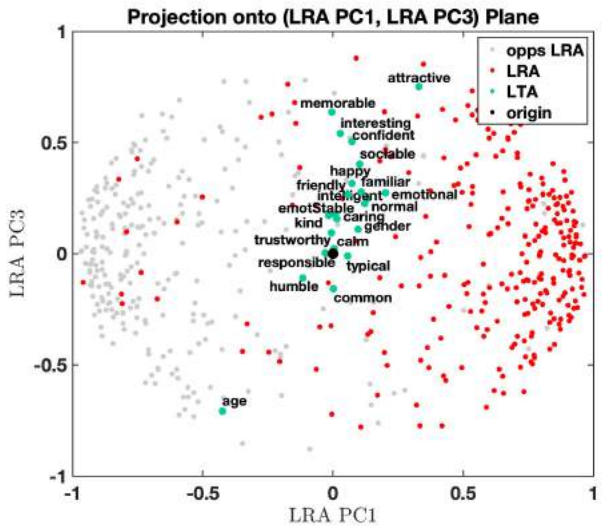


Figure 11: Projection of LRA's and LTA's onto LRA PC 1 and LRA PC 3. Similar formatting as Figure 9.

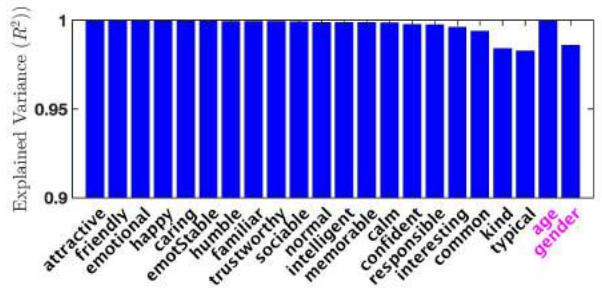


Figure 12: Comparison between LRA subspace and social LTA subspace. The bars (blue: social trait, magenta: demographic trait) represent the amount of variance in LTA explained by LRA, measured in terms of R^2.

with $\vec{x}$ as the 60-dim representation in the face model, and β the regression coefficients for the target trait.

Neuron Encoding - Linear Response Axis (LRA)

Similar to LTA, the Linear Trait Axis (LRA) $\tilde{\alpha}$ for each neuron is the normalized regression coefficients of neurons mean fire rate (MFR) regressed against AAM features.

$$r = \alpha\vec{x} + \varepsilon$$

$$\tilde{\alpha} = \frac{\alpha}{||\alpha||}$$

where r is the neurons MFR, $\vec{x}$ is the AAM features, and α is the vector of regression coefficients. $\tilde{\alpha}$ is the axis in the Face Model that drives maximal (linear) variation of the neurons response. Consistent with existing literature (Chang & Tsao, 2017), we find that MFR (averaged across neurons) increases monotonically along the LRA, and is flat along the principal orthogonal axis (data not shown).

Cross-validation is implemented to evaluate the reliability of LRA estimation. When estimating the LRA for a neuron, its true response to one stimulus is held out as test data. Using the LRA fitted on the remaining faces, we can compare the model-predicted MFR with the actual MFR on the held-out data point. For each neuron, the same process is repeated for every face image (as the test data point). We then the correlation coefficient between the true MFR and model-predicted MFR across all held-out data. The neuron is retained for further analysis if the correlation is significant ($p < 0.05$).

Subspace Comparison

For two vector spaces A and B, let $\{a_1, a_2, \ldots a_n\}$ be a set of vectors in A. The explanatory strength of space B for vector a_1 is determined by the percentage of variance of data from space A explained by the best linear combination of Bs basis vectors:

$$R^2 = 1 - \frac{\sum_i (z_i - z_{approx})^2}{\sum_i (z_i - \bar{z})^2},$$

where z_i is the data projection on vector a_1, $\bar{z}$ is the mean, and z_{approx} is the projection to space B.

Discussion

Our results indicate that, while macaque face patch neurons are primarily tuned to combinations of facial features that are rather different from those most important for human social trait perception, one can easily go back and forth using a simple linear operation (linear decoding scheme). There is no particular reason to expect that monkey face patch neurons, or monkeys themselves should particularly care about social trait perception of human faces. However, our results suggest that human social perception of faces may arise simply as linear decoding of featural information in a neural representational system that humans and monkeys share with each other, and with our common primate ancestors.

Leveraging computational modeling, our work represents a novel way to *retroactively* analyze social perceptual information or other face-related cognitive or perceptual information in monkey neural recording data, even if no social ratings are collected for the face images that the monkeys actually saw. We can also easily extend this framework to other kinds of animal neural data, or to human neural recording (or neuroimaging) data, obtained while experimental participants viewed face images. Technologically, this approach presents a promising approach for extracting much more information out of neural data about the neural basis of face processing, than has been hitherto possible.

Acknowledgments

We thank Doris Tsao and Winrich Freiwald for sharing the monkey neural data, Samer Sabri for helpful advice with the writing, and the UCSD CRES program for partial funding.

References

Bainbridge, W. A., Isola, P., & Oliva, A. (2013, November). The intrinsic memorability of face photographs. *J. Exp. Psychol. Gen.*, *142*(4), 1323–1334.

Chang, L., & Tsao, D. Y. (2017, June). The code for facial identity in the primate brain. *Cell*, *169*(6), 1013–1028.e14.

Cootes, T. F., Edwards, G. J., & Taylor, C. J. (2001). Active appearance models. *IEEE Trans. Pattern Anal. Mach. Intell.*(6), 681–685.

Freiwald, W. A., & Tsao, D. Y. (2010, November). Functional compartmentalization and viewpoint generalization within the macaque face-processing system. *Science*, *330*(6005), 845–851.

Guan, J., Ryali, C., & Yu, A. J. (2018, July). *Computational modeling of social face perception in humans: Leveraging the active appearance model.*

Olivola, C. Y., Funk, F., & Todorov, A. (2014, November). Social attributions from faces bias human choices. *Trends Cogn. Sci.*, *18*(11), 566–570.

Valentine, T. (1991, May). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *Q. J. Exp. Psychol. A*, *43*(2), 161–204.

Willis, J., & Todorov, A. (2006, July). First impressions: making up your mind after a 100-ms exposure to a face. *Psychol. Sci.*, *17*(7), 592–598.

The Impact of Meta-memory Judgments on Undergraduate's Learning and Memory Performance.

Salwa Ali H Humsani (sh787@exeter.ac.uk)

Ciro Civile (c.civile@exeter.ac.uk)

I. P.L. McLaren (I.P.L.McLaren@exeter.ac.uk)

School of Psychology, University of Exeter, UK.

Abstract

We examined if using meta-memory judgments to control restudy choices has a positive impact on undergraduate students' memory performance, or whether simply making meta-memory judgments improved memory performance. 72 undergraduates at the University of Exeter were randomly divided into three groups. Participants in group A, had a chance to make meta-memory judgments and restudied the words they chose (self- selection). Participants in group B, also made meta-memory judgments, but restudy for this group was matched to that of Group A (control 1). Group C did not have a chance to make meta-memory judgments and were also matched to Group A for restudy opportunities (control 2). The results indicated that making meta-memory judgments had a positive overall impact on memory performance if undergraduates were allowed to control their restudy opportunities. Groups B and C showed no differences in memory performance, which means that making meta-memory judgments did not automatically improve memory performance. Group A restudied more of the words that they had rated as least well learned, and there were no significant differences between groups on test for the restudied words.

Keywords: Meta-memory Judgment (MJ), Restudy Choices, Learning, Memory.

Introduction

Meta-memory judgments rely on an individual's knowledge about how her or his memory processes affect their memory performance (Flavell, 1999; Hanczakowski, Zawadzka, & Cockcroft-McKay, 2014; Nelson, Dunlosky, Graf, & Narens, 1994). Efficient learning not only requires one to recognize information from memory, but also to be able to judge their level of confidence in material that they have previously learned and studied (Nelson, 1990; Nelson et al., 1994). One of the key reasons for studying meta-memory judgments is because it serves two functions: monitoring of memory processes and control over study behaviour (Nelson, 1990). The relationship between these functions is direct: people use memory monitoring, especially metamemory judgments, to decide which items need to be restudied and the length of time to be spent on them (Dunlosky & Hertzog, 1997; Kornell & Metcalfe, 2006). The central question addressed here is, do meta-memory judgments lead to effective study decisions? To test this experimentally, this study assumed that there would be positive effects on memory performance when participants are able to monitor their learning and are also able to use it to control their restudy opportunities.

Monitoring Accuracy: Studying cue-target word pairs is the most common approach used to investigate monitoring accuracy (Kimball, Smith, & Muntean, 2012; Nelson et al., 1994; Pyc, Rawson, & Aschenbrenner, 2014; Robey, Dougherty, & Buttaccio, 2017; Thiede & Dunlosky, 1994). This study design typically involves participants studying the word pairs, then making a Meta-memory judgment (MJ) to rate their ability to recall the target word when a cue is presented in the final test. Finally, participants take recall and recognition tests, which allow the researchers to assess how MJ can predict memory performance (Hughes, Taylor, & Thomas, 2018). Meta-memory judgments can be made immediately after the word pairs are studied, or delayed and made to the cue word alone. According to Nelson and Dunlosky (1991) the most important difference between immediate and delayed MJ lies in the amount of information available to participants when they judge their level of confidence. Participants who have made an immediate MJ have their target information in working memory, by contrast, this target is not available in working memory for a delayed MJ. Participants instead need to retrieve it from long-term memory. Several studies have investigated whether immediate MJ is more accurate than delayed MJ (Kimball et al., 2012; Nelson & Dunlosky, 1991; Pyc et al., 2014; Robey et al., 2017; Thiede & Dunlosky, 1994). While this study does not intend to assess the accuracy difference between these two processes, it uses the immediate MJ. The reason for choosing immediate MJ is provided by Hughes et al. (2018) who found that monitoring accuracy increases with immediate MJ when participants review material as a means of controlling repeated study or study- test practice. This study assumes that better meta-memory monitoring will lead to better restudy decisions, and also assumes that our main interest is in controlling restudy decisions at the time of study, rather than during later revision of the material.

Effectiveness of self-regulation: Effective learning involves two skills as stated earlier: monitoring learning and controlling study based on that monitoring (Kornell & Metcalfe, 2006; Nelson, 1990). Giving participants the opportunity to have control over their choices of which words to restudy allows them to be more engaged with their learning and improves their performance in the final memory tests. Begg, Martin & Needham (1992) and Hager & Hasselhorn (1992) concluded that self-memory monitoring is of no value to memory performance if participants did not control their study as well. In addition, Kornell & Metcalfe (2006) and Tullis & Benjamin (2012) tested the effectiveness of self-selection on using metacognitive judgements to control learning and memory performance. They found that allowing

participants to control their learning had a positive effect on memory performance, long term learning and restudy choices. Methodologically, these researchers have used different ways to test the effectiveness of self-regulation on memory performance. Some of these ways involve comparing memory performance between groups: an experimental group (allowed to choose) versus a control group (choices made for them) (e.g. Begg et al., 1992; Kimball et al., 2012; Kornell & Metcalfe, 2006), as well as establishing comparisons on the basis of the best learned or worst learned restudied items (Nelson et al., 1994), or items rated as most difficult by participants (Thiede & Dunlosky, 1994). In this study, memory performance will be compared between all items, including those selected for restudy **or** unselected items across groups (experimental and control; Begg et al., 1992; Kimball et al., 2012; Kornell & Metcalfe, 2006). Several studies have shown that when allowing participants to judge their confidence and use that judgment to control their restudy decisions, the final memory performance was better than controls (Begg et al., 1992; Kimball et al., 2012; Kornell & Metcalfe, 2006; Nelson et al., 1994; Tullis & Benjamin, 2012). The first aim of this research is to test if using meta-memory judgment to control restudy choices has a positive impact on undergraduate students' memory performance, or whether simply making meta-memory judgments improved memory performance.

Method

Summary of the task: Forty concrete Arabic nouns with their pronunciations and translations were used to create two lists of word pairs, twenty in each list. Words were limited to be between four to eight letters. The words were randomly selected to serve as practice (first list) and study (second list) word pairs. Each phase of the experiment (practice followed by main study) started with instructions, then each item from the appropriate word list was presented on the screen for ten seconds on a white background. After each word, participants in Groups A (self-selection) and B had to judge their confidence of remembering the word in the future by rating their confidence from 1 to 9 (1=low confidence, 9= high confidence). Participants in Group C had to make a rating of how similar the Arabic word was to its English translation. Participants in Group A were then also asked if they needed to restudy the word just seen or not. Recall and recognition tests were given at the end of experiment.

Experiment design: The experiment used a between-subject design with three groups: Group A, who both made meta-memory judgments and could choose whether or not to restudy words (self-selection), control Group B, who made meta-memory judgments of their learning and experienced the same restudy opportunities as their counterpart in Group A (they were yoked to them), and control Group C, who did not make meta-memory judgments ratings but made similarity ratings instead, and were also yoked (in terms of restudy) to their counterpart in Group A. After seeing a word pair, participants first made their MJ to rate their meta-memory confidence of remembering the word later. Then, in Group A, each participant was able to request a restudy opportunity for any word. If they did, they then also made a meta-memory judgment after restudy as well. In Group B, each participant made a MJ, then got the same restudy opportunities as one of the participants in Group A but had no choice in the matter. If they were given a restudy opportunity, then they also made a second MJ to that word pair. For Group C, each participant had a chance to study the Arabic words, however they did not make a MJ during the task, but did get the restudy opportunities of one of the participants in Group A, and made a second similarity judgement after any restudy opportunity.

Participants: In this pilot study, random sampling was used. The participants of the study were 72 undergraduates from the University of Exeter; 24 in the meta-memory judgments and re-study (experimental group A), 24 in the MJ (control group B) and 24 in the No MJ (control group C) were random selected. The sample of the current study included 54 female and 18 male participants aged between 18 to 35 years, who did not speak the Arabic language and were enrolled in a variety of different subject disciplines. They were recruited via posters, email advertising and through the University of Exeter's Psychology Research Participation System. Participants were rewarded with a single payment of £5 or one credit on completion of the experiment, questionnaires and interview.

Procedure: All stimuli were presented with Superlab on a PC. The outline of the procedure for the study is summarised diagrammatically in Figure 1.

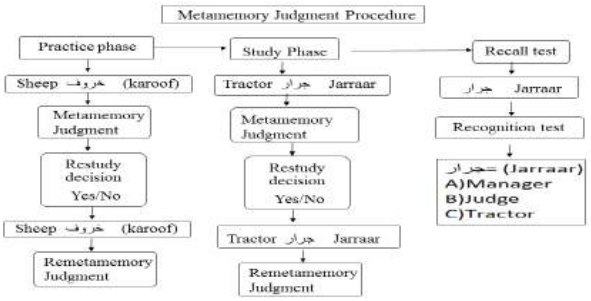


Figure 1 visual representation of the procedure.

Participants first completed a practice phase, which included 20 Arabic words with their translations and pronunciations in English. Then, participants in groups A and B made a judgment of how confident they were in their learning (selected number between 1 to 9). Whereas participants in group C made a judgment of how similar the English translation was to the Arabic pronunciation (again, a 1 to 9 rating, the purpose of this instruction was to have the same procedure across groups). After this, participants in group A were asked whether they would restudy this word if given the opportunity. Participants in group B and C were told that there was chance of repeating some words. Participants in groups A and B again made a meta-memory judgment after restudying words, whereas participants in group C again rated the similarity of the English translation word to the Arabic pronunciation. The same procedure was repeated in the study

phase, but this included 20 new Arabic words. In the last phase, participants were asked to recall all the English translations cued by the Arabic words from the study phase in a random order. After this there was a recognition test provided for each word. One Arabic word with three English translations appeared on the screen, one of which was the right translation; participants had to choose that one. The following sections now give details of our procedure.

In the practice phase. Twenty Arabic words were presented in the middle of the screen between their translation and their pronunciation in English, with a font size of 16. A fixation stimulus was presented before the word pair and a blank screen after the word pair for 1 second. All participants studied the same word pairs (but in a random order), and all the word pairs were presented for ten seconds.

Meta-memory judgment: Participants in the meta-memory judgment A and B groups responded to the following instruction "Please select your level of confidence that you can remember this word pair by pressing the appropriate key from 1 to 9". Participants in group C in the non-meta-memory judgment did not make meta-memory judgments, but were asked the following question "Please rate the similarity of the English translation of this Arabic word to its Arabic pronunciation by pressing a key from 1 (very low) to 9 (very high)".

Restudy judgment. After making their meta-memory judgment, Participants in group A were asked "Would you like to re-study that Arabic to English translation? Press "y" for yes, "n" for no. If you've already re-studied once, then pressing either key will move you to the next trial". Participants in group B and group C were told "Note that there is a chance that the word pair you have just studied will be repeated. Press "y" to move on to the next trial". We arranged for participants in control group B and control group C to restudy the words determined by the matched participants in the experimental group A.

Second meta-memory judgment. After restudying a word, participants repeated the judgement appropriate for the group they were in.

Study Phase. In the main study phase, twenty new Arabic words with their translations and pronunciations in English were presented on the screen in a random order in a similar manner to the practice phase, and participants studied the pairs in order to remember them in the final recall test and recognise them in the final recognition test at the end of the experiment. All participants studied the same word pairs, and all the word pairs were presented for ten seconds at a font size of 16. They were then given meta-memory or similarity judgments and re-study opportunities as before.

Final test. After participants had completed the practice phase and study phase for all the 40 words, they completed a final recall test followed by a recognition test. More specifically, they were asked to recall the English translations of all the words from the study phase (i.e. all 20 words in that phase). They were given their Arabic form and pronunciations in English on the screen in a random order as a cue, and participants were asked to provide the English translation by typing their responses on the keyboard within 30 seconds. After all the words were tested in this way, a recognition test was given for each word. One Arabic word and English pronunciation with three English translations appeared on the screen, one of which was the right translation; participants had to choose this one. Another of the three translations was randomly taken from the practice phase, so that each practice word was used as a distractor once during this test phase. The other incorrect distractor word was novel, and would not have appeared before. None of the words used in a given test trial was repeated in any other test trial. Again, 30 seconds were given to do this, and once a word was selected they moved on to the next trial.

Results

74 participants were run on this experiment. Two participants were excluded because they did not complete the recall test. The results for the remaining 72 participants are as follows:

The first issue we looked at was to determine if either of our control groups differed on either recall or recognition performance (the means for these groups are shown in Figure 2 below). They did not, both Fs<1, so we collapsed B and C into one overall control group and Compared this to A. An independent t- test used to examine the difference in the overall score between Experimental group A and this combined control group at recall gives a statistically significant difference between experimental group (M= 10, SD= 4.81) and control group (M=7.1, SD=3.78), t (70) = 2.727, p<.001, the eta squared statistic ($\eta 2$=.1) indicates a large effect size A similar test used to examine the difference in the overall score for recognition also revealed a statistically significant difference, t = (70) = 2.558, p=.013. The effect size, calculated using eta squared, was close to large ($\eta 2$=0.09). This implies that simply making a MJ does not automatically confer a significant benefit, but in combination with being able to choose which words to restudy it is effective in enhancing memory.

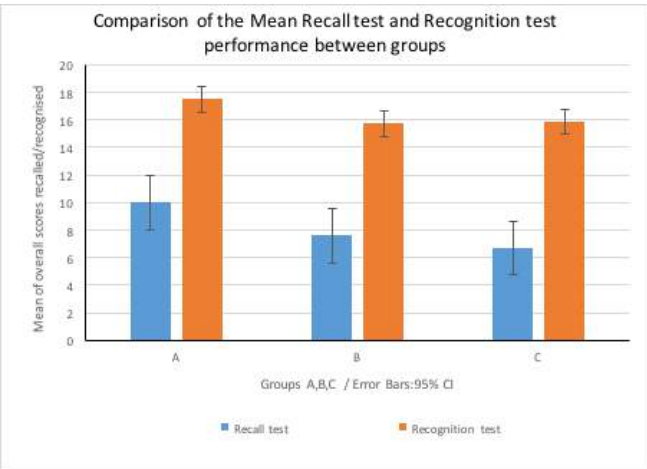


Fig. 2. This shows the difference between groups in their performance on the recall test and Recognition test.

The similar effects for recall and recognition suggest that both simply reflect memory for the word pairs, and this impression is confirmed by a correlational analysis. A Spearman rho test found a strong positive correlation at the 0.005 level (1-tailed) between the two variables, rho=.740.

Our next question was whether there was the expected relationship between the average MJ given to a word pair and performance on the recall and recognition tests? Obviously, we would expect higher average MJ to result in better performance on test. Correlations across participants failed to reveal any significant effects. When we compute this correlation across words, it failed to reveal a positive correlation between the MJ to a word pair and the recall test score for that word (r = .207, p=.079, n=48 1- tailed), but on the recognition test there was a significant and positive correlation between average MJ and performance (r= .439, p< .001, n=48, 1- tailed) with 19 % of the variance in recognition explained by this judgement. These results mean we have some evidence that some words are easier to learn than others.

In addition, as illustrated in Figure 3, there were differences in the MJ given before and after restudy for the restudied words, and also differences in the MJ given to non-restudied words. We analyzed this by performing two separate ANOVAs. The first was used to compare the MJ to the restudied and non-restudied words using the first judgement given in both cases (this would be the only judgement in the case of the non-restudied words). Group (A vs. B) was also included as a factor. The interaction between the groups factor (A, B) and the study factor (non-restudy, restudy) was significant, F (1, 46) = 7.100, p=.011. If we look at Figure 3, we can see that the MJs for Group A are higher than those for Group B for the non-restudied words, but lower for the words that were chosen by Group A for restudy. This is what is driving the interaction. An independent t- test showed that there was a statistically significant difference in the MJ between groups for the non- restudied words in favour of group A, t (46)= 3.073, p<.001, with large effect size (η^2=.3). The difference for the restudied words is not significant. It would appear, then, that Group A selected words that they found particularly difficult for restudy, leaving the easier words, and that this selection was somewhat specific to them, even though the Group B participants obviously show considerable agreement in what are the easier and harder words. This last point is reinforced by the main effect of study in this analysis, with the MJ for non-restudied words being much higher than that for the restudied words, F (1, 46) =40.69, p<.001.

The second ANOVA that we ran compared MJs to the restudied words before and after restudy. The interaction between groups (A, B) and study factor (before restudy and after) was just significant, F (1, 46) = 4.415, p=.041. Obviously, the effect of the factor of Study is stronger than this, the change from before to after is clear in Figure 3. The main effect for the type of MJ (before and after restudy) gave an F (1, 46) =35.197, p<.001, indicating that all participants show improved levels of confidence after restudying words. The interaction suggests that Group A improved more than Group B. Therefore, we have some evidence that restudy really helps participants to improve their level of confidence to remember the words in the final tests, and that effect was greatest when allowing participants to control their restudy opportunity.

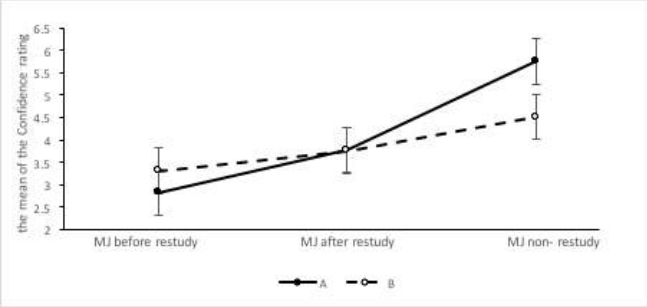


Fig 3. MJ before and after restudy and MJ for non-restudied words for groups A and B, Error Bars: 95%CI.

Turning now to a correlational analysis across participants, a Spearman rho test revealed a strong negative correlation between the mean of the initial MJ made by a participant and the number of requests for restudy in Group A, rho=.-.747 n=24, p<.001 (1-tailed). This correlation means that people who tended to give a lower MJ on average also tended to ask to restudy more. In essence, it could be taken to suggest that people have some sense of whether they are finding the task easy or hard, and adjust their study strategy accordingly. Another Spearman rho test gives a statistically significant and large positive correlation between frequency of restudy for a participant and their recall performance. That is, the more times people restudy (on average) the better their ability to retrieve words in the recall test, rho= .661 n=72, p<.001. There is a similar effect with recognition, a significant large positive correlation between frequency of study and recognition test performance, rho=.610, n=72, p<.001. It is not hard to see why this would be the case. The more restudy, the more practice of the items one gets, and if that helps then the better he or she performs. But this then leaves us with a slightly paradoxical situation, where the participants that we would argue are finding the task hardest (as signaled by a low MJ on average) are actually the ones performing the best. Perhaps the low average MJs may actually reflect better self-knowledge (i.e. a form of meta-memory) rather than ability as such. These are the people who know that they need to restudy, and do so, and benefit from it. Those with higher average MJs may be confident but may be mistaken in their confidence. It's also worth pointing out that the correlation between restudy request frequency and performance includes participants in Groups B and C who had no control over restudy. In some sense, the restudy manipulation was simply one imposed on them, and the result that more restudy benefitted performance is not surprising in that context. Finally, the correlations between MJ and test performance across subjects were not significant, so actually the paradox is not present in our data, just a potential feature of our theory.

After looking at the MJ data for the restudied and non-restudied words, we quite naturally would like to know how performance differed for those word types, and whether it differed across groups. Having collapsed B and C into one group as they do not differ on these measures, there was a main effect on recall for the type of words F (1, 69) =20.284, p=<.001. Participants performed better on non-restudied words than on restudied words. An independent t- test found statistically significant differences between the experimental group A (M= .504, SD= .240) and the combined control group (M=.360, SD=.192) for non- restudied words, t (70) = 2.734, p= <.001. The eta squared statistic (η^2= .1) indicated a large effect size. Whereas the difference between experimental group A (M= .301, SD= .368) and the control group (M=.222, SD=.315) for restudied words was not significant t (70) = .940, p=.350. The fact that overall, participants recalled more words when they were not restudied we take to simply reflect the fact that these were the easier word pairs. The fact that Group A was better than the combined control on these words again suggests an item specific advantage based on Group A participants selecting the words. Whilst the words were generally the easier ones (hence the main effect), the agreement on this between Group A and controls was not complete.

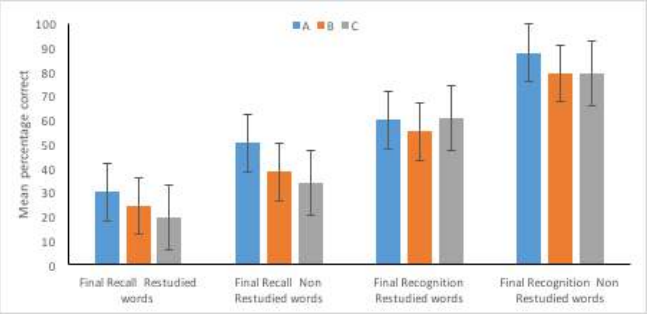


Fig. 4. Mean percentage of Final Recall and Recognition tests between restudied and non-restudied between groups A, B and C. Error Bars: 95%CI.

Turning now to the recognition test, once again participants recognise more words from the non-restudied words than the restudied words, F (1, 69) =20.783, p<.001. After collapsing B and C into one control group, an independent t- test showed there was a statistically significant difference between the experimental group (M= .878, SD= .111) and control group (M=.791, SD=.148) for the non- restudied words, t (70) = 2.515, p= =.014. The effect size, calculated using eta squared, was large (η^2= .1). However, the difference between the experimental group (M= .599, SD= .463) and control group (M=.579, SD=.423) for restudied words was not significant t (70) = .184, p=.854. (See Figure 4). Just as before, we can attribute some of this advantage on test to Group A's effective selection of the non-restudied words as the easier items, and that selection not transferring completely to the other groups. As a result, Groups B and C find these items harder on average and score lower on test. But the overall averages indicate general agreement about which are the more difficult items, as the overall means for the non-restudied words are much higher than for the restudied ones.

Discussion

In this experiment, we have found a significant advantage for Group A over the other two groups in terms of performance on the tests used to assess memory, with no significant differences between Groups B and C. It can be concluded that using meta-memory judgments and allowing control of restudy has a positive impact on participant's memory performance. The question now is why does this happen? Is it that the use of meta-memory allows our participants to select items that they find particularly difficult for restudy, thus improving performance? Or is it a more general effect? We can envisage two possibilities here. One is that people know how good their memories are and can make use of that knowledge to guide restudy. Another is that that giving people control over their choice of restudy items improves their motivation and engagement with the task. As we have seen and will see, there is evidence for both explanations, but what we can say is that simply "exercising" meta-memory, by giving a MJ to an item, is not in itself a significant factor in improving performance, otherwise there would be a significant difference between Groups B and C.

These results are in line with those of previous studies Begg, Martin, & Needham (1992) and Hager & Hasselhorn (1992) who concluded that self-memory monitoring has no value for memory performance if participants did not also have control of their study. In addition, these results are in agreement with those obtained by Kornell & Metcalfe (2006) and Jonathan G. Tullis & Benjamin (2012) who tested the effectiveness of self-selection on the use of metacognitive control over learning and memory performance. They found that allowing participants to control their learning had a positive effect on memory performance, long term learning and restudy choices.

This study found that there was a significant difference on MJ between non-restudy words and restudy words, and one possible explanation for this is that meta-memory monitoring helped participants in groups A and B to make their meta-memory judgments so as to discriminate between items which were more difficult and items that were easy and really sufficiently learned. These result are in line with those of previous studies who found that young subjects use their metacognition monitoring to distinguish between more difficult items (Li et al., 2018; Tullis & Benjamin, 2012; Tullis, Fiechter, & Benjamin, 2018; Zawadzka et al., 2018). Another significant finding was that meta-memory judgments could, to some extent, predict a participant's restudy frequency. Participants in group A, requested restudy more often for items that they had judged as least well – learned. This finding supports the work of other studies in this area (e.g. Dunlosky & Hertzog, 1997; Li et al., 2018; Nelson et al., 1994; Jonathan G. Tullis & Benjamin, 2012). Equally, participants in group A and B show significant improvement in their MJ after restudying words. These results agree with the finding of other study such as

Zawadzka et al. (2018) who demonstrate that repeated learning in the same environment improved learning, and metacognition monitoring. Obviously we are unable to comment on the effect of restudying a word on learning here because we do not know what performance on the restudied words would have been if they had not been restudied, but the positive correlation between frequency of restudying and test performance does fit in with the results cited. We intend to gather data that bears directly on this issue.

We can interpret some of the correlational results very simply as meaning that higher confidence about learning translates into better memory performance later. This would fit well with an item-specific effect of meta-memory on these tasks, whereby items judged as hard by Group A participants were given a low MJ, and this was used to trigger a restudy request. The effect of this was to improve performance on these items, back up to the level shown by the controls, while the advantage on the items not chosen for restudy because they were easy was greater in Group A again because they were able to make the right choices for them. Whilst there is general agreement about which are the easy and hard word pairs across groups (as shown by the correlations by word for MJ and test performance reported earlier) there is enough disagreement for the control groups to not gain as much benefit from the restudy offered, and so Group A does better. This is one possible explanation for our results.

But there may be more to this. Note that the MJ was, on average, higher in Group A than Group B, and, as we have seen, Group A performs better on test than Group B and C combined. The higher MJ in Group A could reflect increased confidence due to having control over which items are restudied, but this could be a general motivational effect rather than one based on meta-memory. To be clear, it could be that both the high MJ and better memory in Group A are both due to increased motivation due to their being in control of their restudy choices, in which case it would not be correct to say that the high MJ had some causal role in improving performance for Group A relative to Group B. Further research will be needed to disentangle the relationship between these variables.

Acknowledgments

This research is supported by The Saudi Arabian Cultural Bureau.

References

Begg, I. M., Martin, L. A., & Needham, D. R. (1992). Memory monitoring: How useful is self-knowledge about memory? *European Journal of Cognitive Psychology, 4*(3), 195-218.

Dunlosky, J., & Hertzog, C. (1997). Older and younger adults use a functionally identical algorithm to select items for restudy during multitrial learning. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences, 52*(4), P178-P186.

Finn, B., & Metcalfe, J. (2008). Judgments of learning are influenced by memory for past test. *Journal of Memory and Language, 58*(1), 19-34. doi:https://doi.org/10.1016/j.jml.2007.03.006

Hager, W., & Hasselhorn, M. (1992). Memory monitoring and memory performance: Linked closely or loosely? *Psychological Research, 54*(2), 110-113.

Kornell, N., & Metcalfe, J. (2006). Study efficacy and the region of proximal learning framework. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*(3), 609.

Li, P., Zhang, Y., Li, W., & Li, X. (2018). Age-related differences in effectiveness of item restudy choices: the role of value. *Aging, Neuropsychology, and Cognition, 25*(1), 122-131.

Nelson, T. O. (1990). Metamemory: A theoretical framework and new findings. *Psychology of learning and motivation, 26*, 125-173.

Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The "delayed-JOL effect". *Psychological Science, 2*(4), 267-271.

Nelson, T. O., Dunlosky, J., Graf, A., & Narens, L. (1994). Utilization of metacognitive judgments in the allocation of study during multitrial learning. *Psychological Science, 5*(4), 207-213.

Nelson, T. O., Narens, L., & Dunlosky, J. (2004). A revised methodology for research on metamemory: Pre-judgment Recall and Monitoring (PRAM). *Psychological methods, 9*(1), 53.

Souchay, C., & Isingrini, M. (2012). Are feeling-of-knowing and judgment-of-learning different? Evidence from older adults. *Acta Psychol (Amst), 139*(3), 458-464. doi:10.1016/j.actpsy.2012.01.007

Thiede, K. W., & Dunlosky, J. (1994). Delaying students' metacognitive monitoring improves their accuracy in predicting their recognition performance. *Journal of Educational Psychology, 86*(2), 290.

Tullis, J. G., & Benjamin, A. S. (2012). Consequences of restudy choices in younger and older learners. *Psychonomic Bulletin & Review, 19*(4), 743-749. doi:10.3758/s13423-012-0266-2

Tullis, J. G., Fiechter, J. L., & Benjamin, A. S. (2018). The efficacy of learners' testing choices. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 44*(4), 540.

Zawadzka, K., Simkiss, N., & Hanczakowski, M. (2018). Remind me of the context: Memory and metacognition at restudy. *Journal of Memory and Language, 101*, 1-17.

Does incorporating social media messages into television programs affect the validation of incorrect arguments?

Miwa Inuzuka (minuzuka@u-gakugei.ac.jp)
Department of Education, Tokyo Gakugei University
4-1-1 Nukui-kita-machi, Koganei, Tokyo, 184-8501 Japan

Yuko Tanaka (tanaka.yuko@nitec.ac.jp)
Graduate School of Engineering, Nagoya Institute of Technology
Gokiso-cho, Showa-ku, Nagoya, 466-8555 Japan

Mio Tsubakimoto (mio@fye.c.u-tokyo.ac.jp)
Division of First-Year Education, Komaba Organization for Educational Excellence,
Graduate School of Arts and Sciences, College of Arts and Sciences, The University of Tokyo
3-8-1 Komaba, Meguro-ku, Tokyo, 153-8902 Japan

Abstract

The present study explores the impact of including social media messages on learning from television programs that broadcast pseudoscientific claims. Seventy-seven university students were allocated to one of three experimental conditions: viewing television content with messages supporting the claim, with opposing messages, or without any messages presented. Memory retention did not differ among the conditions. However, social media messages influenced validation of the arguments claimed in the video. The participants who watched the video with opposing messages showed significant decrease in positive attitude toward the pseudoscientific technology that claimed to be effective in the video. Additionally, the participants who watched the video with supporting messages made fewer critical comments and showed willingness to donate more to the activity using the pseudoscientific technology. The impact of including social media messages and the process of attitude change are discussed.

Keywords: social media messages; learning from television programs; incorrect arguments; validation of argument; attitudes; retention.

Introduction

Learning from television programs with social media messages

One of the major sources for everyday learning is television. Since television programs are designed with various styles, it is not easy to define the processes of learning from television in general. Thus, we begin by focusing on a relatively simple program that broadcasts experts' explanations. Although the style is simple, we can find many examples of the type of television programs in which experts like scholars and scientists explain topics of interests such as politics, technology, and science.

The present study investigates how social media messages impact learning from the "experts' explanation" type of show. It is shown that people often access social media while watching television. It is also getting popular to include social media posts on the screen during such programs (Inuzuka, Tanaka, & Tsubakimoto, 2017; Barra & Scaglioni, 2014). In this case, the social media messages, which typically include hashtags, are searched and presented. (See Figure 1 for an example of how these feeds may be presented.) The programs usually include messages that consist mainly of text, such as posts on Twitter. Although the relationships between social media and viewing television programs have begun to be explored widely (e.g., Anstead & O'Loughlin, 2011; Ceron & Splendore, 2018; Miao, 2018; Waddell & Bailey, 2019), few studies investigate their impact on cognitive processes (e.g., Cameron & Geidner, 2014; Maruyama, Robertson, Douglas, Semaan, & Fucett, 2014; Maruyama, Robertson, Douglas, & Raine, 2017). Thus, we still lack evidence to discuss their effects on learning.

In the present study, we focus on the impacts of social media messages on validation as well as memory retention. Validation is one type of integration process that requires activation of one's prior knowledge and unfolding a logical argument (e.g., Halldorson & Singer, 2002; Lea, Mulligan, & Walton, 2005; Singer, Halldorson, Lear, & Andrusiak, 1992). The inclusion of social media messages may impact the validation process of the viewers; the messages can activate viewers' knowledge or provide new information that is effective for appropriate validation. These social media messages, however, cannot always be effective for validation. The messages contain various opinions (D'heer & Verdegem, 2015), and irrelevant and inappropriate messages can be included as well as helpful ones. Previous studies failed to investigate how qualitatively different messages impact viewers' learning. Thus, the present study investigates the effects of different types of social media messages on viewers' memory retention and validation of arguments provided in television programs.

Learning from multimedia sources and the effects of including social media messages

While the "experts' explanation" type of television program may seem simpler than other styles, the situation can be

described as learning from multimedia materials. Watching the program, the viewers integrate the information presented in the speech and other visually presented materials such as graphs and illustrations. When social media messages are incorporated into the program, the viewers must integrate more information presented visually in the text of social media messages.

The literature on multimedia learning suggests that the inclusion of social media messages may interfere with viewer comprehension since the messages may contain incoherent information. Mayer (2009) suggested a "coherent principle" in which learners understand a topic better when irrelevant and seductive elements are removed from the learning materials. The coherence principle can be explained by the split-attention effect theory; a multimedia resource results in less learning when it splits learners' attention (Sweller & Chandler, 1996). This attention split is more likely to occur when the resource contains information sharing the same modality and when it is not coherent with the other information presented (Mayer, 2009; Mayer & Moreno, 1998).

Consideration of the coherence principle led us to assume that the presentation of messages interferes with learning since these messages are not consistent with the main information of the contents. Inuzuka, Tanaka, and Tsubakimoto (2017, 2018), however, suggested that the effects of presenting social media messages on memory retention were limited. They compared the memory retention scores of participants who watched video material including and not including social media messages. Participants paid attention to the messages when presented but showed no significant difference in retention scores between the two groups.

The gap between the coherent principal and the results of Inuzuka et al. (2017, 2018) can be interpreted from the standpoint of the difference in the level of comprehension. Research on multimedia learning suggested that violation of the coherent principal mainly influences the integration of learning materials and the learner's knowledge (Mayer, 2009). Thus, we can assume that Inuzuka et al. (2017, 2018) showed no significant effects of including social media messages since they examined memory retention, which did not require integration of the knowledge.

The validation of false arguments

Viewers activate their prior knowledge, integrate the information presented, and validate the arguments (e.g., Halldorson & Singer, 2002; Lea, Mulligan, & Walton, 2005). Validation of an argument is especially important when it comes to learning from television programs since the issues tackled in television programs are often relevant to viewers' lives and require them to decide what to believe and what to do. Additionally, and more importantly, the media do not always provide fair and correct arguments. Consideration and validation of potentially biased information are among the most important practices in surviving the information age.

Research shows that people display difficulty rejecting information even when the texts they read are inconsistent with prior knowledge and even patently false (e.g., Gilbert, Krull, & Malone, 1990; Gilbert, Tafarodi, & Malone, 1993). Gerrig and Prentice (1991) revealed that it took longer to vilify a false statement as "incorrect" when participants read the statement discussed as truth in a narrative text. These studies suggest that learners accept what they have processed as truth first and resolve the validation afterwards. Rapp (2008) suggested that, when providing fake information within a context that casts doubt on correct information, the verification becomes even more difficult for learners.

To extend the above discussion, it is necessary to note that the above studies employed information regarding which the correctness of the arguments was apparent to the learners. Television programs, however, usually focus on issues on which learners do not possess much prior knowledge. In this case, the validation of incorrect argument becomes a more difficult and complex task that demands more deliberate consideration. Thus, we must employ an index other than reaction time. Consideration of new ambiguous topics should and can be measured more qualitatively using participants' attitudes toward the topic, decision making, and the explanation of the situation related to the issue.

The impacts of presenting social media messages

We can predict that the presentation of social media messages changes the way viewers validate presented arguments. Maruyama et al. (2017) investigated the effects of referring to social media messages when watching a discussion on the television. They revealed that viewers' attitudes were different in the direction of the social media messages. Similarly, Cameron and Geidner (2014) explored the effects of social media feeds on viewers' opinion formation. They indicated that participants' opinions were found to conform to the majority opinion presented in the messages. These studies suggest that conformity process in which viewers may follow the majority of the people.

The above studies are limited, however, as they did not investigate the situation in which learners are required to validate incorrect arguments. When watching a discussion in which both sides of the argument can equally be justified, the viewers' consideration and decision making would depend on what the majority says. Thus, conformity can best describe the impacts of social media messages, as depicted by Maruyama et al. (2017). However, the same may not be true when the argument claimed by the specialist on the television program is incorrect. Thus, this study aimed to examine whether the impacts of social media messages are valid when new and incorrect information is presented and to explore if the impacts are caused by conformity.

Aim of the study

The present study focused on how the incorporation of social media into television programs affects memory retention and

validation of incorrect arguments. More specifically, we examined the effects of social media messages by presenting either opposing or supporting messages for the pseudoscientific claims. We hypothesized the following:

(1) The presentation of the social media messages does not interfere with memory retention that does not require integration of knowledge.

(2) The presentation of social media messages impacts viewers' validation of pseudoscientific claims. Namely, the viewers change their attitudes in the direction of the social media messages, and the viewers react differently to the situation in which they must make some decision.

Method

Participants

Seventy-seven undergraduates participated in this study after providing informed consent and were assigned to one of three conditions: Supporting, Opposing, and Without message. As a reward for their participation, they received a 500 Japanese yen (approximately $4.50) cash voucher.

Materials

Fake television program The video material used by Inuzuka et al. (2017) was edited for the purpose of the present study (Figure 1). The original video was produced to mimic a scientific talk show. We omitted some parts of the video so that only the claim of one scientist (an actor) remained. Following the procedure above, the video material used in the present study was approximately eight minutes long. The scientist stated that "Effective Microorganisms" (EM) are effective for improving water quality. "EM" is a pseudoscience based on the idea that a particular collection of microorganisms can solve virtually all health and environmental problems. We chose the topic because it is relevant to participants' lives and yet unfamiliar to them.

Fake social media messages We included fake social media messages that simulated Twitter posts in the video material presented to the participants in the Supporting and Opposing message conditions. The messages consisted of text with each containing one or two short sentences. We designed three types of messages: opposing, supporting, and neutral (Table 1). Neutral messages were developed for when neither supporting nor opposing messages were appropriate. Neutral messages were, therefore, included in both Opposing and Supporting conditions and were presented at the same time in both conditions. Opposing and supporting messages were included in the corresponding conditions, and each message was inserted at the bottom of the screen (Figure 1) approximately five seconds after the relevant topic was mentioned by the scientist. The participants assigned to the Without messages condition watched the video not including the messages.

Retention test A retention test was developed with six quiz items (e.g., "What was the name of the two rivers that Scientist A claimed that EM cleaned up?"). The tests were administered after participants had watched the video.

Attitude questionnaire To assess the participants' validation of the video contents, whether the participants agreed with the effectiveness of EM was measured using a questionnaire. The attitude questionnaire was administered before and after the participants watched the fake video. It consisted of two subscales with three items each: positive attitude (e.g., "I think EM will somehow do some good") and careful attitude (e.g., "We need more investigation on the effectiveness of EM"). The participants were asked to answer the items on seven-point Likert scales.

Figure 1. A frame from the video material that mimics the television program displaying a social media feed saying, "*So, the 'power of nature' means using microorganisms. Right?"*

Table 1. Examples of fake social messages used in the study.

	Example
Supporting (33)	It is important to use an enriched compound of specific types of organic matter. I understand.
Opposing (33)	After all, I think EM is condensed organic matter. If so, there might be a risk of causing more pollution.
Neutral (21)	I agree that it is important to discuss in a scientific way.

Note: The numbers in parentheses are the numbers of each type of message. Supporting messages were presented only to the participants in the Supporting condition and opposing messages to those in the Opposing condition.

Explanation and decision-making task Additionally, we developed a test in which a short story was introduced to qualitatively assess the consideration and validation of the argument. In that story, the following scenes were introduced: "You are considering making a donation, and a man comes and explains that NGOs are planning water quality-improvement activities using EM." The participants were asked to decide how much they would donate to that NGO (0–5000 JPY, approximately 40 USD). Participants were also asked to write comments and questions for the man in the story.

Evaluations of the messages The participants in the Opposing and Supporting message conditions rated three questionnaire items on an 11-point scale: (1) the extent to which the social media messages were against the claim, (2) how much attention they paid to the messages, and (3) how much they considered the contents of the messages.

Procedures

Each participant was tested individually in a laboratory. Each session lasted approximately 30 minutes. After participants had signed a consent form, the experimenter introduced the video, explaining, "The video is a digest of a television program. In the program, a scientist will explain how they try to clear water pollution." The experimenter then instructed the participants to watch the television show and learn from it. Each participant was randomly assigned to one of the three conditions: Supporting, Opposing, and Without messages. No instruction regarding the social media messages was given, so the participants were not aware of the differences among the conditions. After watching the video, participants responded to the retention test and attitude questionnaire. There was no time limit for completing the questionnaires, but participants did so within 10–15 minutes.

Results

The evaluation of messages

Three participants were excluded from the following analysis since they reported that they knew about EM in advance. To confirm that the different types of social media messages were delivered to the participants, we employed the participants' rating for the extent to which the social media messages were against the claim. The difference between conditions was significant, $t(45) = 10.0$, $p < .001$, $d = 2.97$. The mean scores were significantly higher than the neutral score ($t(22) = 9.56$, $p < .001$) in the Opposing condition and lower than the neutral score ($t(22) = -5.31$, $p < .001$) in the Supporting condition.

Additionally, the participants' rating for the extent of attention ($t(45) = 2.13$, $p < .05$, $d = 0.63$) and consideration ($t(45) = 3.82$, $p < .001$, $d = 1.136$) of the messages also differed between two groups, indicating that the participants assigned to the Opposing condition rated themselves as paying more attention and considered the messages.

The effects of message presentation on retention test

Each retention test item was scored with two points, and the total was used as the retention test score (Table 3). Fully correct answers were given two points, and partially correct answers, such as giving only one name of a river when two should be named, were given one point. The difference in retention test score among the conditions was analyzed with a one-way ANOVA. The result indicated no significant difference among the experimental conditions, $F(2,70) = 1.45$.

The effects of message presentation on attitude

For the analysis of attitude change, we used the average scores of positive and careful attitude questionnaire items. The mean scores for each subscale are shown in Table 3. The impact of message presentation was analyzed with two-way mixed ANOVAs. The dependent variables were positive and careful attitude scores, and the independent variables were conditions (Opposing, Supporting, and Without messages), time of measurement (pretest and posttest), and the interaction effect of two independent variables.

The results of careful attitude score showed no significant main effects of experimental condition ($F(2,71)= 1.78$) and time ($F(1,71)= 0.421$), and there was no significant interaction effect either ($F(2,71) = 0.745$).

On the other hand, the analysis of positive attitude revealed significant results. The main effect of time was significant ($F(1,71) = 9.37$, $p < .01$, $\eta_p = .117$), showing a decreasing tendency, while the main effect of condition was not significant ($F(2,71) = 1.58$). More importantly, the interaction effect of condition and time reached a significant level (Figure 2, $F(2,71) = 10.22$, $p < .001$, $\eta_p = .224$). Subsequent analysis of simple effect revealed that positive attitude was decreased significantly only in the Opposing condition, $F(1,71) = 27.26$, $p < .001$, $\eta_p = .532$). The change in other conditions did not reach a significant level ($F(1,71) = 1.42$ for the Without condition and $F(1,71) = 1.21$ for the Supporting message condition). The effects of conditions were significant only at the posttest ($F(2,142) = 8.36$, $p < .001$, $\eta_p = .191$), showing a significant difference between Opposing and Without message conditions ($t(142) = 2.35$, $p < .05$, $d = 0.943$) and between Opposing and Supporting message conditions ($t(142)= -4.07$, $p < .001$, $d = 1.62$). The difference between Without and Supporting message conditions was not significant, $t(142)=1.68$.

The results of one-way ANOVA conducted on the explanation score showed a significant effect of condition ($F(2,71) = 3.50$, $p < .05$, $\eta_p = .090$), and the following multiple comparison (Holm) revealed that the difference between Supporting and Opposing message conditions was significant ($t(71) = 2.55$, $p < .05$, $d = 0.710$) with higher scores for the participants in the Opposing message condition.

Table 2. Average scores for evaluation of the messages

	Supporting	Opposing
The messages against the claim	3.86 (1.96)	9.25 (1.51)
Attention paid to the messages	7.25 (2.80)	8.75 (1.56)
Consideration of the message contents	5.46 (2.78)	8.05 (1.31)

Note: The numbers in parentheses are standard deviations.

Table 3. Average scores on the retention test, change in attitude, and critical thinking disposition scales for each experimental condition

	Supporting	Opposing	Without
Retention test	7.71 (2.31)	8.44 (3.17)	7.42 (2.39)
Positive attitude			
Pretest	3.63 (0.58)	3.91 (0.64)	3.68 (0.54)
Posttest	3.84 (1.04)	2.29 (0.79)	3.44 (1.15)
Careful attitude			
Pretest	3.64 (0.43)	3.76 (0.67)	3.75 (0.15)
Posttest	3.64 (0.41)	3.92 (0.41)	3.72 (0.52)
Explanation score	0.72 (1.44)	1.32 (1.41)	0.88 (1.41)
Donation amount (yen)	1750.00 (1161.00)	916.00 (1086.16)	1071.67 (1075.965)

Note: The numbers in parentheses are standard deviations.

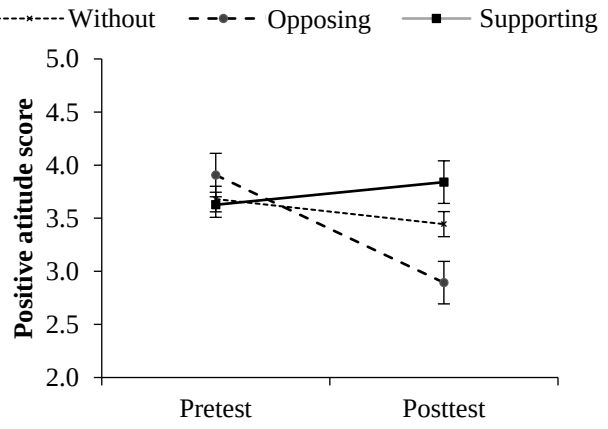


Figure 2. Changes in positive attitude as a function of message presentation. Error bars represent standardized errors.

Explanation and decision-making task

The explanation score was calculated based on the number of critical points included in the answer to the explanation and decision-making task. The participants were given one point each when referring to the following points: (1) suspicious effects of EM, (2) the lack of clear explanations for the mechanism, (3) the lack of consideration of side effects, (4) the need for solid data. Thus, the explanation score for each participant was in the range of 0–4. The donation amount indicated by the participants was also used as an index for the validation. The average scores and SDs are shown in Table 3.

Additionally, the amount of money the participants were willing to donate for the activity using EM was compared among the conditions. One-way ANOVA revealed a significant effect of experimental condition ($F(2,70) = 3.89$, $p < .001$, $\eta_p = .100$). Multiple comparison (Holm) was conducted and showed significant difference between Supporting and Opposing message conditions ($t(70) = -2.64$, $p < .05$, $d = -0.741$). The participants in the Supporting message condition tended to donate more than those who watched the same video with opposing messages.

Discussion

The present study investigated the effects of including social media messages in a television program on which incorrect arguments were claimed. As predicted in Hypothesis 1, the results demonstrated no significant difference in retention. The result was consistent with previous studies using a similar method (Inuzuka et al., 2017; 2018) and with studies of multimedia learning (Mayer, 2009), suggesting that the incorporation of social media messages would not interfere with the memory of what had been discussed in the program. The results for attitude changes and explanation and decision-making tasks also supported our hypothesis about the impact of social media messages on validation of the arguments (Hypothesis 2). The participants who watched opposing messages became less positive about the effectiveness of EM, the pseudoscientific technology. The participants' explanation and decision also showed that those who watched supporting messages were relatively uncritical about using the pseudoscientific technology.

The results of the present study suggested that showing counterarguments in text messages may support the viewers to consider and validate the information shown in the television programs more appropriately. Considering research showing that rejecting incorrect text is difficult for readers (Gerrig & Prentice, 1991; Gilbert et al., 1990; Gilbert et al., 1993; Rapp, 2008), it may be beneficial to incorporate these counter-messages for viewers.

However, it should be noted that the messages included in the study were biased, either supporting or opposing the explanation of the expert in the program. Actual social media messages are supposed to be more varied including both appropriate and inappropriate arguments. As Inuzuka (2017)

showed that presenting varied messages did not significantly change the viewers' attitude, the impacts of appropriate counterarguments may be wiped out when combined with inappropriate messages.

While previous studies (Cameron & Geidner, 2014; Maruyama et al., 2014, 2017) suggested conformity as the mechanism underlying the effects of social media messages on the viewers' attitude change, the present study brought up another possibility. The conformity hypothesis should predict that both opposing and supporting messages will have similar impacts on participants' validation process. However, we found smaller attitude changes in the Supporting condition in the present study. The unequal results in our two conditions may be caused by the qualitative difference in the messages presented. The ratings of the consideration of messages showed that the participants in the Opposing condition considered the messages more than those in the Supporting condition (Table 2.).

The difference may be caused by the effectiveness of the messages; the messages in the Opposing condition provided other perspectives from the expert's explanation while the messages of the Supporting condition provided rephrasing and supplemental information. Thus, the participants may consider the messages of the Opposing condition to be more informative. If the above consideration stands, it can be said that the information contained in the messages is used in the process of deliberation rather than merely conformity. The next step of the research, therefore, should be to clarify if the impacts are caused by the conformity of deliberate consideration.

The results also showed that the attention change led to decision-making in a more realistic situation. The participants in the Supporting condition tended to donate more with fewer questions about the appropriateness of the activity. Although the present study is based on a laboratory examination using a fake television program, the results provided eligible data to discuss the effects of showing biased information. Presenting biased information without counterargument may result in an actual disadvantage.

The present study makes meaningful contributions toward understanding how we learn from a new type of media. The first is the suggestion that incorporation of social media messages affects individuals. The results of the present study broaden the previous studies on social media and television programs by showing that incorporation of meaningful messages would help viewers more appropriately validate the information. The second is the expansion of the research on validation of incorrect information to broader learning contexts. Previous studies mainly focused on information presented in texts and information the participants already knew. The present study highlights information that participants newly learn and suggests that using different media may be an effective way to present counterarguments.

We should note, however, some limitations of the study. First, the instruction for the participants should be less instructive. We instructed the participants to learn from the television program to make sure they focused on the program, but the instruction may have influenced their attitude and caused better memory retention while they may have spared more attention for the social media messages if not for the instruction. Although we repeatedly found small effects of the social media messages on memory for detailed facts, it is important to test the impacts of those in more natural settings.

Secondly, a more thorough comprehension test should be administered. In the experiment, we used a quiz to test the participants' memory retention. The quiz mainly tapped detailed memory of the program contents. Open-ended questions and analysis of the structure of their memory would enable us to understand the impacts of social media messages on memory in more detail.

Finally, the mechanism of the impacts of message presentation should be investigated in future studies. The impacts of social media messages shown in the present study supported the hypothesis that social media messages provide support for deliberate consideration or evaluation of information. Since the present study does not provide direct evidence to discuss the process of attitude change, there remains one alternative interpretation: conformity (c.f., Maruyama et al., 2017). However, relatively weak impacts of supporting messages suggest that the effects of messages may not be caused by conformity alone. If the participants reacted to the messages in the direction these messages suggest, the participants should change their attitude equivalently in both Supporting and Opposing conditions. The future direction of the study is more detailed investigation of the process of validation: conformity or deliberate consideration.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 15K04084.

References

Anstead, N., & O'Loughlin, B. (2011). The emerging viewertariat and BBC question time: Television debate and real-time commenting online. *The International Journal of Press/Politics, 16*, 440–462.

Barra, L., & Scaglioni, M. (2014). TV goes social: Italian broadcasting strategies and the challenges of convergence. *VIEW Journal of European Television History and Culture, 3*(6), 110–124.

Cameron, J. & Geidner, N. (2014). Something Old, Something New, Something Borrowed From Something Blue: Experiments on Dual Viewing TV and Twitter, *Journal of Broadcasting & Electronic Media,* 58, 400-419

Ceron, A., & Splendore, S. (2018). From contents to comments: Social TV and perceived pluralism in political talk shows. *New Media and Society, 20*, 659–675.

D'heer, E., & Verdegem, P. (2015). What social media data mean for audience studies: A multidimensional investigation of Twitter use during a current affairs TV

programme. *Information, Communication & Society, 18*, 221–234.

Gerrig, R. J., & Prentice, D. A. (1991). The representation of fictional information. *Psychological Science, 2*, 336–340.

Gilbert, D. T., Krull, D. S., & Malone, P. S. (1990). Unbelieving the unbelievable: Some problems in the rejection of false information. *Journal of Personality & Social Psychology, 59*, 601–613.

Gilbert, D. T., Tafarodi, R. W., & Malone, P. S. (1993). You can't not believe everything you read. *Journal of Personality & Social Psychology, 65*, 221–233.

Halldorson, M., & Singer, M. (2002). Inference processes: Integrating relevant knowledge and text information. *Discourse Processes, 34*, 145–161

Inuzuka, M., Tanaka, Y., & Tsubakimoto, M. (2017). Students' comprehension of scientific discussion: Using eye-tracking technique to investigate the effects of social-media messages on television. *Proceedings of the 50th Annual Hawaii International Conference on System Sciences (HICSS)*, pp. 2106–2115.

Inuzuka, M., Tanaka, Y., & Tsubakimoto, M. (2018). Do social media messages incorporated into television programming impact learning? The effects of disposition to critical thinking. *Proceedings of the 40th Annual Conference of the Cognitive Science Society*, pp. 524–529.

Lea, R. B., Mulligan, E. J., & Walton, J. L. (2005). Accessing distant premise information: How memory feeds reasoning. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 31*, 387–395.

Maruyama, M. Robertson, S. P., Douglas, Semaan, B., & Faucett, H. (2014). Hybrid media consumption: How tweeting during a televised political debate influences the vote decision. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '14)*, pp. 1422–1432.

Maruyama, M., Robertson, S. P., Douglas, S., Raine, R., & Semaan, B. (2017). "Social watching" a civic broadcast: Understanding the effects of positive feedback and other users' opinions. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*, pp. 794–807.

Mayer, R. E. (2009). *Multimedia learning*. New York: Cambridge University Press.

Mayer, R. E., & Moreno, R. (1998). A split-attention effect in multimedia learning: Evidence for dual processing systems in working memory. *Journal of Educational Psychology, 90*, 312–320.

Miao, G. (2018). How television viewers use social media to engage with programming: The social engagement scale development and validation. J*ournal of Broadcasting & Electronic Media*, *62*, 195–214.

Rapp, D. N. (2008). How do readers handle incorrect information during reading? *Memory and Cognition, 36*, 688–701.

Singer, M., & Halldorson, M. (1996). Constructing and validating motive bridging inferences. *Cognitive Psychology, 30*, 1–38.

Sweller, J., & Chandler, P. (1994). Why some material is difficult to learn. *Cognition and Instruction, 12*, 185–233.

Waddell, T. F., & Bailey, E. (2019). Is social television the "anti-laugh track?" testing the effect of negative comments and canned laughter on comedy reception. *Psychology of Popular Media Culture, 8*(1), 99–107.

Wait for it!
Stronger influence of context on categorical perception in Danish than Norwegian

Byurakn Ishkhanyan (byurakn@cc.au.dk)
School of Communication and Culture, Aarhus University, Denmark

Anders Højen (hojen@cc.au.dk)
School of Communication and Culture, Aarhus University, Denmark

Riccardo Fusaroli (fusaroli@cas.au.dk)
School of Communication and Culture & Interacting Minds Centre, Aarhus University, Denmark

Christer Johansson (christer.johansson@uib.no)
Department of Linguistic, Literary and Aesthetic studies, University of Bergen, Norway

Kristian Tylén (kristian@cc.au.dk)
School of Communication and Culture & Interacting Minds Centre, Aarhus University, Denmark

Morten H. Christiansen (christiansen@cornell.edu)
Cornell University, Department of Psychology, Ithaca, NY 14850 USA
School of Communication and Culture & Interacting Minds Centre, Aarhus University, Denmark

Abstract

Speech input is often noisy and ambiguous. Yet listeners usually do not have difficulties understanding it. A key hypothesis is that in speech processing acoustic-phonetic bottom-up processing is complemented by top-down contextual information. This context effect is larger when the ambiguous word is only separated from a disambiguating word by a few syllables compared to many syllables, suggesting that there is a limited time window for processing acoustic-phonetic information with the help of context. Here, we argue that the relative weight of bottom-up and top-down processes may be different for languages that have different phonological properties. We report an experiment comparing two closely related languages, Danish and Norwegian. We show that Danish speakers do indeed rely on context more than Norwegian speakers do. These results highlight the importance of investigating cross-linguistic differences in speech processing, suggesting that speakers of different languages may develop different language processing strategies.

Keywords: categorical perception; speech perception; Danish; Norwegian, cross-linguistic studies

Introduction

Speech is often ambiguous and noisy. Yet most of the time listeners show remarkable skills in understanding what is being said. A possible explanation is that the imperfect acoustic-phonetic input is integrated with contextual information. Thus, to understand speech, listeners combine bottom-up acoustic-phonetic cues with top-down lexical-semantic and pragmatic contextual information. This context effect might be particularly apparent when the acoustic-phonetic information is unclear or noisy (e.g., Borsky, Tuller & Shapiro, 1998; Gaskell & Marslen-Wilson, 2001; Marslen-Wilson & Welsh, 1978; Samuel, 1981).

Despite the variability of the acoustic properties of individual sounds and the noisiness of the acoustic-phonetic input, the perception of speech sounds is *categorical* (Liberman et al., 1957). This means that a certain sound is usually perceived unambiguously (e.g., either as a /b/ or as a /p/); listeners ignore within-category acoustic differences while easily perceiving across-category acoustic differences of the same magnitude.

Both within-word and sentential context facilitate sound categorization when the acoustic-phonetic information is ambiguous (Brown-Schmidt & Toscano, 2017; Bushong & Jaeger, 2017; Connine, Blasko & Hall, 1991; McMurray, Tanenhaus & Aslin, 2009; Szostak & Pitt, 2013). In a phoneme identification study, Connine et al. (1991) manipulated the onset of the target words *dent/tent* on a continuum from a clear [d] to a clear [t^h] with three intermediate steps. The listeners were presented with sentences biased either towards *dent* (*After the _ent corroded, they patched it*) or towards *tent* (*After the _ent collapsed, we went home*). Connine et al. (1991) showed that listeners often relied on the biasing word at the end of the sentence to disambiguate the target word, when the target word had an ambiguous onset, whereas they were not biased by the context (biasing word), when the target word had a phonetically clear onset. They concluded that top-down inference from the context is given more weight when the target input is ambiguous than when it is clear.

In the same study, Connine et al. (1991) showed that the contextual biasing effect was present when the target word was separated from the disambiguating word by a small number of syllables (NEAR condition) but not when there was a larger number of syllables (FAR condition). The response time data, however, showed that in the FAR

condition, most of the time, the decision was being made prior to the availability of the disambiguating information, suggesting that there was an approximately 1 s window to make a decision based on acoustic-phonetic information prior to its decay.

In an eye-tracking study, Brown-Schmidt & Toscano (2017) showed a context bias effect even when the ambiguous word is separated from the biasing context by six-seven syllables. In fact, prior to disambiguation, the listeners fixated on the interpretation of the word that did not match the context but shifted their gaze only after having heard the biasing context. Similarly, Szostak and Pitt (2013) replicated the contextual biasing effects on ambiguous-sounding phoneme identification. Although smaller than in the NEAR condition, they also observed a biasing effect in the FAR condition. The authors suggested that the temporal window for disambiguating unclear acoustic-phonetic information may not be completely fixed, as suggested by Connine et al. (1991), but rather influenced by other factors, such as syntactic complexity or experience with language use.

Another factor that could affect the temporal window may be the typological characteristics of a given language. However, so far, language processing studies have mainly focused on English, therefore making it difficult to generalize the findings to other languages. In fact, it is debated whether all languages are processed in the same way and thus findings in one are generalizable to the others (Pinker, 1994), or whether each language has its unique characteristics, shaped by language users (Evans & Levinson, 2009). In the current study we address the question of whether individual languages are all processed in the same way or afford different processing strategies. Specifically, we investigate potential differences in the processing of the two languages—Danish and Norwegian—which are closely related but differ substantially in their phonological structure.

The Case of Danish and Norwegian

The relative weight that context is given in speech comprehension may vary from language to language, depending on the typological characteristics of a given language. We hypothesized that Danish may be a language, where top-down contextual processes is given larger weight than bottom-up acoustic-phonetic cues, compared to its close linguistic neighbors, Swedish and Norwegian. In terms of cross-linguistic comparisons, Danish and Norwegian thus allow for a well-controlled natural experiment. Denmark and Norway have a long common history, and have strong similarities in culture, education, politics, and other extra-linguistic factors. The two languages also have very similar grammars, morphology, and vocabulary—but differ in their phonology: Danish has a much more opaque phonology than Norwegian.

The sound structure of Danish is quite unique. Apart from having an unusually high number of vowels and vowel-like consonants, there is also a higher degree of syllabic reduction and assimilation of both vowels and consonants, compared to its close relatives Norwegian and Swedish (Basbøll, 2005). As a result, Danish is more difficult to acquire as a native language than Swedish and Norwegian (Bleses, Basbøll & Vach, 2011). There is also evidence that out of these three mutually intelligible Scandinavian languages, Danish is the most difficult to understand (Gooskens et al., 2010; Hilton, Schüppert & Gooskens, 2011). This may be due to the fact that there is generally a higher degree of syllabic reduction in Danish than in Norwegian and Swedish. Moreover, due to phonological reduction in Danish, some words sound identical to each other (Basbøll, 2005). In general, Danish speakers are thus exposed to a more imperfect and unclear acoustic-phonetic input compared to their Scandinavian neighbors. And, as a result, Danish speakers may rely on top-down processes to a larger extent than Norwegian and Swedish speakers.

In the current study, we adapted the paradigms used by Connine et al. (1991) and Szostak and Pitt (2013) to test the hypothesis that Danish speakers, due to the phonological peculiarities of the language, rely more on top-down processes than Norwegian speakers do. We predicted that when presented with ambiguous sounding words, Danish speakers would rely more on contextual cues compared to Norwegian speakers. In fact, for Danish speakers, we expected this effect to be present not only in the NEAR condition but also in the FAR condition, indicating that the acoustic-phonetic bottom-up input is given relatively less weight by Danish speakers than by Norwegian speakers. Moreover, we predicted that Danish speakers would be more inclined to wait until the end of the sentence to respond than Norwegian speakers (H1: language main effect). Following the findings for English by Szostak and Pitt (2013) and Connine et al. (1991), we expected that both Danish and Norwegian speakers would be affected by contextual bias (H2: contextual bias main effect) and that the effect would be stronger in the NEAR condition (H3: bias by distance interaction). Additionally, given the processing differences between Danish and Norwegian, we expected the bias effect to be stronger in Danish (H4: bias by language interaction) and the bias by distance interaction stronger in Norwegian (H5: bias by distance by language interaction).

To test these hypotheses, we fitted our experimental data to a drift diffusion model (Ratcliff, 1978), which jointly takes into account responses and response times as dependent variables and allowed us to separate the time preceding the decision making process (non-decision time), the rate at which evidence is accumulated (drift rate) and the amount of evidence needed to make a decision (boundary separation, see Methods for details). We expected to observe a longer non-decision time in Danish speakers than Norwegian speakers (H1). We expected the evidence accumulation rate to be affected by contextual bias (faster in congruent contexts, H2). We expected both drift rate and boundary separation to be affected by contextual bias in a way that is modulated by distance (stronger effect for the shorter distance, H3), and by language (Danish speakers being more sensitive to context, H4). Finally, we expected both drift rate and boundary separation to follow H5: Norwegian speakers

will show a stronger bias by distance interaction, that is, the way distance modulates contextual bias will be more marked for them (H5).

Method

Participants

Thirty-two Danish (22 female, age = 19 - 36 years, median = 23, sd = 3.3) and 34 Norwegian (13 female, age = 19 - 28 years, median = 22, sd = 2.5) right-handed native speakers participated in the study. The participants did not report a history of neurological or psychiatric disorders. The Danish speakers were tested at the Cognitive and Behavior Lab at Aarhus University in Denmark, while the Norwegian speakers were tested at the Faculty of Humanities at the University of Bergen in Norway. All participants received a monetary compensation for their participation.

Materials

We constructed 16 pairs of carrier sentences, half of which were biased towards the target word *sendt,* as shown in (1a) (Danish) and (1b) (Norwegian) and the other half towards *tændt* in Danish (2a) or *tent* in Norwegian (2b). In 8 pairs, the distance between the target and the disambiguating word was one syllable (NEAR condition); and in the remaining 8 pairs, it was 5-7 syllables (FAR condition). Importantly, in normal speech, except for the difference in the initial phoneme, the two target words have similar (rhyme) endings in both languages.

(1a) *Hun har **sendt** en (imponerende klar) mail.*
[ˈhun ˈhɑ ˈ**sɛntˢ** eˀn (empoˈneˀʌnə klɑˀ) ˈmɛjl]
'She has **sent** an (impressively clear) email.'
(1b) *Hun har **sendt** en (imponerende klar) mail.*
[ˈhʉn ˈhɑr ˈ**sɛnt** en (empoˈnerʌnə klɑr) ˈmɛjl]
(2a) *Hun har **tændt** en (imponerende klar) lampe.*
[ˈhun ˈhɑ ˈ**tˢɛntˢ** eˀn (empoˈneˀʌnə ˈklɑˀ) ˈlɑmbə]
(2b) *Hun har **tent** en (imponerende klar) lampe.*
[ˈhʉn ˈhɑr ˈ**tɛnt** en (empoˈnerʌnə klɑr) ˈlɑmpə]
'She has **turned-on** a(n) (impressively clear) lamp.'

Both the Danish and the Norwegian stimuli were recorded by a native male speaker of the respective languages. The recorded Danish [s] and [t^s] sounds in the target words *sendt* and *tændt* differed primarily according to the duration of the frication noise, the rise time of the noise, and the duration of the silent interval between noise offset and onset of the following vowel. The same was true for the Norwegian target words' [s] and [t^h] sounds, which in addition differed in intensity. A ten-step s-t continuum was generated for each language by interpolating between the endpoints according to the above-mentioned acoustic differences and splicing the resultant sounds to a single token of *tændt/tent*. The continua had a clear [s] at one end and a clear [t^s] (Danish) or [t^h] (Norwegian) at the other end and with eight intermediate steps.

We then piloted the two continua (forced choice identification). Based on the identification functions we chose steps 4, 5 and 6 as they straddled the mean category boundaries in each language. These three intermediate steps and the endpoints were used in the experiment. Thus, there were 160 trial sentences in total. The experiment was programmed and carried out in PsychoPy2 v1.90.1. (Peirce & MacAskill, 2018).

Procedure

Prior to the experiment, the participants received detailed instructions on the screen in their native languages. They were told to indicate which word they thought they heard and they were warned that sometimes this would not be easy. The participants were also instructed that they could use any information in the sentence that may help them to make their decision (cf. Connine et al., 1991; Szostak & Pitt, 2013). Following the instructions, the participants completed a practice trial and then the real experiment began. The target words *sendt* and *tændt/tent* were presented in boxes in the upper left and right corners of the screen while the target sentences were played back through headphones. The participants responded by clicking on the appropriate word with the mouse. They were allowed to respond at any point during and after the sentence playback (cf. Connine et al., 1991). There was a pause of 1.5 s between each trial, during which a blank screen was presented. The 160 stimuli were presented in a pseudorandomized order across four blocks of 40 trials. The first two items of the experiment contained the endpoints [s] and [t^s]/[t^h], respectively, in a congruent context. After each block, the participants had a self-paced short break. The whole experiment took 15 – 20 minutes. Responses and response times (RTs) were recorded as dependent variables. RTs were measured from the onset of the target word until the mouse click.

Data Analysis

Mouse clicks outside the boxes were recorded as missing values and were removed from the analysis. Further, responses corresponding to RTs higher than 3 standard deviations from the mean (> 8s) were also excluded from the analysis (2% of the total number of data points).

We fitted a Bayesian multilevel drift diffusion model (DDM) to the response and RT data. DDM is a sequential sampling model that explains cognitive processes underlying decision-making in 2-choice discrimination tasks (Ratcliff & McKoon, 2008). Decisions are described by the following parameters: the drift rate (δ) is the average rate of evidence accumulation; the boundary separation (α) is the evidence necessary to make a decision; the starting point (β) is the initial bias towards one of the response boundaries; and non-decision time (τ) is the part of the response time that is not involved in evidence accumulation (e.g., motor response execution). We conditioned drift rate and boundary separation on language, contextual bias (congruent/ incongruent), distance (NEAR/FAR) and continuum step as fixed effects, including their interactions, and participants as

varying effects, including varying slopes for bias, distance and step. We assumed no biased preference for a specific response and conditioned non-decision time on language and contextual bias only due to convergence issues. PSIS-LOO model comparison was used to select the relevant predictors to include (Vehtari, Gelman & Gabry, 2017), which led us to exclude step. We set weakly informative priors for δ (mean = 0, *sd* = 0.5), α (mean = 1.5, *sd* = 1) and τ (mean = 0.2, *sd* = 0.1). Model quality was thoroughly assessed via predictive prior and posterior checks, Rhat and divergence diagnostics. The model presented no divergences, and all chains mixed well and produced comparable estimates (Rhat < 1.01). In order to assess the evidence in favor or against our hypotheses, we used Evidence Ratio (ER, a generalization of Bayes factors allowing for directional hypotheses). An ER above 3 indicates moderate to substantial evidence for our hypothesis, below 0.3 indicates moderate to substantial evidence for the null hypothesis, and anything in between is inconclusive evidence (Morey, Rouder & Jamil, 2014). The models were implemented through the *brms* (Bürkner, 2017) and *RWiener* (Wabersich & Vandekerckhove, 2014) packages in RStudio v1.1.46, following the procedures of the tutorial written by Singmann (2017).

Results

Descriptive statistics are presented in Table 1. Full parameter estimates by condition are presented in Table 2.

Table 1: Mean reaction times ± standard deviations (in seconds) and *tændt/tent* response mean proportions ± standard deviations for Danish and Norwegian and NEAR and FAR distances with the context biased towards *sendt* or *tændt/tent*.

Language	Distance	Context bias	RT (s)	Response *tændt/ tent* (%)
Danish	NEAR	sendt	2.08 ± 0.82	66 ± 12
		tændt	2.15 ± 0.90	72 ± 8
	FAR	sendt	2.70 ± 1.07	66 ± 12
		tændt	2.71 ± 1.10	69 ± 10
Norwegian	NEAR	sendt	2.49 ± 0.98	34 ± 16
		tent	2.56 ± 1.02	50 ± 24
	FAR	sendt	3.38 ± 1.26	32 ± 18
		tent	3.35 ± 1.22	48 ± 22

As predicted by H1, we observed substantial evidence for non-decision time being longer in Danish than in Norwegian in congruent context ($\Delta\tau = 0.11 \pm 0.02$, ER > 1000), indicating that Danish speakers waited longer before starting to make a decision.

As per H2, we found substantial evidence for contextual bias affecting Danish speakers in the NEAR condition. When the response (i.e., *tændt*) matched the contextual bias (biased towards *tændt*, congruent context), evidence accumulation was faster ($\Delta\delta = 0.21 \pm 0.1$, ER = 45.5), than when the context did not match (biased towards *sendt,* incongruent context). There was also evidence that the boundary separation was larger for congruent context than for incongruent context ($\Delta\alpha = 0.98 \pm 0.29$, ER > 1000). Contrary to our expectations, however, there was no evidence that Norwegian speakers were affected by contextual bias ($\Delta\delta$ = -0.06 ± 0.12, ER = 0.46; $\Delta\alpha$ = -0.2 ± 0.36, ER = 0.41).

As expected (H4), we found substantial evidence for the bias effect being larger for Danish than for Norwegian speakers ($\Delta\Delta\delta = 0.27 \pm 0.12$, ER = 89.9, $\Delta\Delta\alpha = 1.18 \pm 0.39$, ER = 999). In other words, Danish speakers relied more on contextual evidence: matching context sped up their evidence accumulation more than for Norwegian speakers.

Table 2: The estimates of the diffusion drift model parameters per condition Bias (congruent/incongruent), Language (Danish/Norwegian) and Distance (NEAR/FAR). The parameters are drift rate (δ), boundary separation (α) and non-decision time (τ).

δ	estimate	95% CI
congruent:Danish:NEAR	2.12	1.94 - 2.31
incongruent:Danish:NEAR	1.91	1.75 - 2.07
congruent:Norwegian:NEAR	1.92	1.72 - 2.12
incongruent:Norwegian:NEAR	1.98	1.80 - 2.16
congruent:Danish:FAR	1.79	1.59 - 1.98
incongruent:Danish:FAR	1.74	1.56 - 1.93
congruent:Norwegian:FAR	1.70	1.47 - 1.92
incongruent:Norwegian:FAR	1.63	1.43 - 1.84
α		
congruent:Danish:NEAR	3.90	3.00 - 4.72
incongruent:Danish:NEAR	2.92	2.13 - 3.64
congruent:Norwegian:NEAR	4.31	3.27 - 5.29
incongruent:Norwegian:NEAR	4.52	3.58 - 5.41
congruent:Danish:FAR	3.25	1.94 - 4.48
incongruent:Danish:FAR	2.63	1.36 - 3.86
congruent:Norwegian:FAR	4.09	2.69 - 5.47
incongruent:Norwegian:FAR	3.55	2.16 - 4.98
τ		
congruent:Danish	0.57	0.54 - 0.6
incongruent:Danish	0.71	0.68 - 0.73
congruent:Norwegian	0.46	0.43 - 0.49
incongruent:Norwegian	0.50	0.46 - 0.54

We found moderate evidence that the drift rate was affected by contextual bias more in the NEAR condition than in the FAR condition in Danish speakers (H3: bias by distance interaction, $\Delta\Delta\delta = 0.15 \pm 0.17$, ER = 6.1). There was, however, no substantial evidence for boundary separation being affected by contextual bias differently according to distance ($\Delta\Delta\alpha = 0.37 \pm 0.58$, ER = 2.9). As for Norwegian speakers, there was no evidence either for drift rate ($\Delta\Delta\delta$ = -0.12 ± 0.17, ER = 0.3) or boundary separation ($\Delta\Delta\alpha$ = -0.75 ± 0.68, ER = 0.16) being affected more in the NEAR than in the FAR condition (against H3). Finally, as predicted, distance did not affect Norwegian speakers as much as

Danish speakers (H5, $\Delta\Delta\Delta\delta$ = -0.29 ± 0.18, ER = 16.7; $\Delta\Delta\Delta\alpha$ = -1.11 ± 0.71, ER = 15.7). This is likely to be due to the absence of bias effect in Norwegian altogether. The DDM simulations per each condition are depicted in Figure 1.

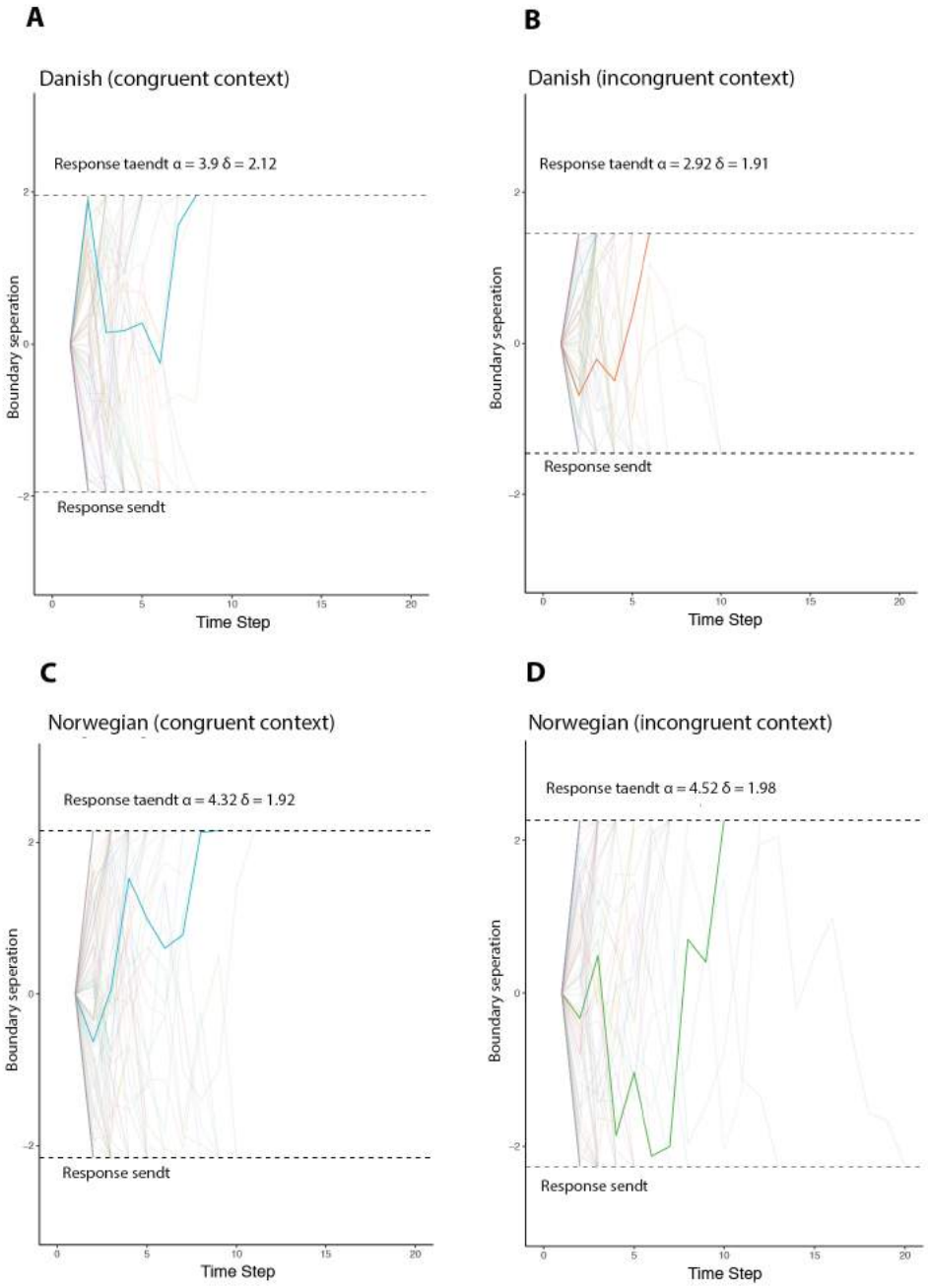


Figure 1: Simulation of the drift diffusion model for distance NEAR and congruent context (i.e. *tændt* bias) in A) Danish and C) Norwegian, and incongruent context (i.e. *sendt* bias) in B) Danish and D) Norwegian. The upper decision boundary is for the response *tændt* and the lower decision boundary is for the response *sendt*. The distance between the two boundaries is the boundary separation (α) and the evidence accumulation speed is the drift rate (δ). While there is no credible difference in drift rate between congruent and incongruent contexts for Norwegian, there is evidence that the drift rate is smaller in the incongruent context than in the congruent context in Danish. Thus, when the context is incongruent, evidence is accumulated slower to make a decision about *tændt*. The highlighted line is an example of the decision process in each condition.

Discussion

In the current study, we investigated whether contextual bias has a different effect on word recognition across the two related languages, Danish and Norwegian, and whether the distance between the target word and the disambiguating word affected word recognition across the two languages. We fitted our data to a drift diffusion model to obtain more subtle evidence about the cognitive processes underlying word recognition.

We found strong evidence that contextual bias affected the drift rate (the speed with which evidence is accumulated) in the NEAR condition in Danish. This indicates that acoustic-phonetic information alone is insufficient to make a decision and thus that additional evidence, such as contextual cues are integrated to support top-down processes of word comprehension. These findings are in line with previous evidence by Szostak and Pitt (2013) as well as Connine et al. (1991).

Surprisingly, we did not find evidence for contextual bias effects in Norwegian, which contradicts the previous evidence for English (Brown-Schmidt & Toscano, 2017; Bushong & Jaeger, 2017; Connine et al., 1991; McMurray et al., 2009; Szostak & Pitt, 2013). It is possible that this is because top-down contextual information is assigned even less weight in speech processing in Norwegian than in English or Danish. Moreover, Norwegian speakers may have responded prior to hearing the biasing context, thus not having the opportunity of using contextual information. In line with this, we found that Danish speakers generally wait longer to respond than Norwegian speakers (longer non-decision time in Danish compared to Norwegian, H1), which may be additional evidence that Danish speakers weight top-down contextual information more than bottom-up acoustic-phonetic information compared to Norwegians.

In line with our hypotheses, we found that Danish speakers were more affected by contextual biases than Norwegian speakers (H4), and that the contextual bias was stronger for Danish speakers in the NEAR condition than in the FAR condition, compared to Norwegian speakers (H5). However, importantly, the H3 interaction results held only for Danish, likely due to the lack of a contextual bias effect in Norwegian.

There was also some evidence that the bias effect on drift rate was stronger in the NEAR condition than in the FAR condition for Danish speakers but there was no credible evidence of the same effect on boundary separation. It is possible that a similar amount of information is necessary to make a decision about an ambiguous target word, as the nature of the information does not change across NEAR and FAR distances (i.e., the acoustic-phonetic cues are equally ambiguous and the disambiguating words remain the same). However, the speed at which this information is accumulated changes slightly. As Szostak and Pitt (2013) and Connine et al. (1991) suggested, there is a short temporal window to make a decision about the acoustic-phonetic information. Thus, in the NEAR condition, due to a higher drift rate in Danish speakers, it takes shorter time to choose a response that is congruent with the contextual bias (i.e., to respond *tændt* in a *tændt*-biased context).

The above-mentioned effect on drift rate was stronger for Danish than for Norwegian, indicating that the temporal window suggested by Szostak and Pitt (2013) may indeed vary due to different factors, in this particular case, phonological differences between languages. We interpret this evidence as suggesting that top-down contextual inferences are more important for Danish speakers compared to Norwegian speakers, when faced with acoustic-phonetically ambiguous stimulus. This may be because of the unique sound structure of Danish, which results in relatively more ambiguity in Danish speech than in other Scandinavian languages (Basbøll, 2005; Hilton et al., 2011; Gooskens et al., 2010). Thus, in line with first language acquisition studies (Bleses et al., 2008; 2011), we provide evidence that Danish is processed differently also by adult native speakers, compared to native Norwegian speakers.

It is possible that allowing participants to respond at any time during a trial may also have affected our results. Using the Connine et al. (1991) paradigm, Bushong & Jaeger (2017) showed that the context effect was smaller in the FAR condition, when the listeners could respond whenever they wanted. However, there was no difference between the NEAR and FAR conditions, when the listeners were forced to wait until hearing the biasing word to respond. In fact, the observation that participants change their response profile when forced to wait to the sentence offset, as shown by Brown-Schmidt & Toscano (2017), indicates that indeed free and forced responses may influence the decisions that listeners make. Thus, a future study comparing forced and free responses may shed light on the different strategies Danish and Norwegian speakers may be using when completing the task.

The current study, however, has one important limitation: the steps of the [s]-[t^s]/[t^h] continuum were not included in the DDM model. Step is a crucial feature and it could provide more nuanced information not only about the contextual bias and distance effect on word recognition processes but also how these processes vary cross-linguistically. Future work should include a nuanced modeling of step (e.g., as a monotonic but not necessarily a linear function) to assess whether step can be meaningfully included and help better explain the data. We anticipate that such analyses might provide a more detailed picture of the points in the continuum at which information is accumulated faster and at which more information is needed. Thus, a more complex drift diffusion model with the steps of the continuum as one of the fixed effect variables would shed further light on the cognitive processes underlying spoken word recognition when the acoustic-phonetic cues are ambiguous.

Despite these limitations, our study suggests that Danish is processed differently compared to Norwegian. When exposed to ambiguous stimuli, Danish speakers rely more on top-down processes than Norwegian speakers. Contrary to the standard view that all languages are equally easy to learn and use (e.g., Pinker, 1994), we provide evidence that

languages can differ in how they are processed, as suggested, for instance, by Evans and Levinson (2009)— and that there may be a continuum of reliance on top-down processes, where English could be lying somewhere between Danish and Norwegian. However, future cross-linguistic studies are necessary to confirm this assumption.

Acknowledgements

This study was supported by Danish Council for Independent Research (FKK) Grant DFF-7013-00074 awarded to Morten H. Christiansen.

References

Basbøll, H. (2005). *The phonology of Danish*. Oxford University Press.

Bleses, D., Basbøll, H., & Vach, W. (2011). Is Danish difficult to acquire? Evidence from Nordic past-tense studies. *Language and Cognitive Processes*, 26(8), 1193-1231.

Bleses, D., Vach, W., Slott, M., Wehberg, S., Thomsen, P., Madsen, T. O., & Basbøll, H. (2008). Early vocabulary development in Danish and other languages: A CDI-based comparison. *Journal of Child Language*, 35(3), 619-650.

Borsky, S., Tuller, B., & Shapiro, L. P. (1998). "How to milk a coat:" The effects of semantic and acoustic information on phoneme categorization. *The Journal of the Acoustical Society of America*, 103(5), 2670-2676.

Brown-Schmidt, S., & Toscano, J. C. (2017). Gradient acoustic information induces long-lasting referential uncertainty in short discourses. *Language, Cognition and Neuroscience*, 32(10), 1211-1228.

Bushong, W., & Jaeger, T. F. (2017). Maintenance of Perceptual Information in Speech Perception. *In CogSci*.

Bürkner, P. C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1), 1-28.

Connine, C. M., Blasko, D. G., & Hall, M. (1991). Effects of subsequent sentence context in auditory word recognition: Temporal and linguistic constraints. *Journal of Memory and Language*, 30(1), 234.

Evans, N., & Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, 32, 429-448.

Gaskell, M. G., & Marslen-Wilson, W. D. (2001). Lexical ambiguity resolution and spoken word recognition: Bridging the gap. *Journal of Memory and Language*, 44(3), 325-349.

Gooskens, C., Van Heuven, V. J., Van Bezooijen, R. & Pacilly, J. J. (2010). Is spoken Danish less intelligible than Swedish? *Speech Communication*, 52, 1022-1037.

Hilton, N. H., Schüppert, A., & Gooskens, C. (2011). Syllable reduction and articulation rates in Danish, Norwegian and Swedish. *Nordic Journal of Linguistics*, 34(2), 215-237.

Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, 54(5), 358.

Marslen-Wilson, W. D., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, 10(1), 29-63.

McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2009). Within-category VOT affects recovery from "lexical" garden-paths: Evidence against phoneme-level inhibition. *Journal of memory and language*, 60(1), 65-91.

Morey, R. D., Rouder, J. N., & Jamil, T. (2014). BayesFactor: Computation of Bayes factors for common designs (Version 0.9. 9).

Peirce, J. W., & MacAskill, M. R. (2018). Building Experiments in PsychoPy. London: Sage.

Pinker, S. (1994). *The language instinct*. New York: William Morrow & Co.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, 85(2), 59.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: theory and data for two-choice decision tasks. *Neural computation*, 20(4), 873-922.

Samuel, A. G. (1981). Phonemic restoration: insights from a new methodology. *Journal of Experimental Psychology: General*, 110(4), 474.

Singmann, H. (2017, November 26). Diffusion/Wiener Model Analysis with brms – Part I: Introduction and Estimation [Blog post]. Retrieved from http://singmann.org/wiener-model-analysis-with-brms-part-i/

Szostak, C. M., & Pitt, M. A. (2013). The prolonged influence of subsequent context on spoken word recognition. *Attention, Perception, & Psychophysics*, 75(7), 1533-1546.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5), 1413-1432.

Wabersich, D., & Vandekerckhove, J. (2014). The RWiener package: An R package providing distribution functions for the Wiener diffusion model. *The R Journal*, 6(1), 49–56.

Measuring how people learn how to plan

Yash Raj Jain
Rationality Enhancement Group, MPI for Intelligent Systems, Tübingen, Germany
Birla Institute of Technology & Science, Pilani, Hyderabad, India

Frederick Callaway
Department of Psychology, Princeton University, NJ, USA

Falk Lieder
Rationality Enhancement Group, MPI for Intelligent Systems, Tübingen, Germany
Bernstein Center for Computational Neuroscience, Tübingen, Germany

Abstract

How can people learn to make better decisions and become more far-sighted? To make the underlying learning mechanisms more accessible to scientific inquiry, we develop a computational method for measuring the time course of experience-dependent changes in people's planning strategies. We validated our method on simulated and empirical data: on simulated data its inferences were significantly more accurate than simpler approaches, and when evaluated on human data it correctly detected the plasticity-enhancing effect of performance feedback. Having validated our method, we illustrate how it can be used to gain new insights into the time course and nature of cognitive plasticity. Future work will leverage our method to i) reverse-engineer the learning mechanisms enabling people to acquire complex cognitive skills such as planning and problem-solving and ii) measure individual differences in cognitive plasticity.

Keywords: cognitive plasticity; planning; decision-making; process-tracing; statistical methods

Introduction

One of the most remarkable features of the human mind is its ability to continuously improve itself. As helpless babies develop into mature adults, their brains do not only acquire impressive perceptual and sensory-motor skills and knowledge about the world but they also learn to think, to make better decisions, to learn, and to monitor and adaptively regulate themselves. These phenomena are collectively known as *cognitive plasticity*. Just like the acquisition of perceptual skills (Hubel & Wiesel, 1970), the acquisition of cognitive skills requires specific experiences and practice (van Lehn, 1996; Ericsson, Krampe, & Tesch-Römer, 1993).

Despite initial research on how people acquire cognitive skills (van Lehn, 1996; Shrager & Siegler, 1998; Krueger, Lieder, & Griffiths, 2017), the underlying learning mechanisms are still largely unknown. Reverse-engineering how people learn how to think and how to decide is very challenging because we can neither observe people's cognitive strategies, nor how they change with experience – let alone the underlying learning mechanisms. Instead, cognitive plasticity has to be inferred from observable changes in behavior. This is difficult because each observed behavior could have been generated by many possible cognitive mechanisms. This problem is pertinent to all areas of cognition. As a first step towards a more general solution, we develop a computational method for measuring how people's planning strategies change depending on the person's experience. Initial work suggested that metacognitive reinforcement learning might play an important role in how people come to plan farther ahead (Krueger et al., 2017) and which strategies they use (Lieder & Griffiths, 2017) but the postulated mechanisms are difficult to investigate because cognitive plasticity has remained unobservable.

Our approach combines a recently developed process-tracing paradigm that renders people's behavior highly diagnostic of their planning strategies with probabilistic models of planning and learning that constrain the space of potential cognitive mechanisms and exploit temporal dependencies among subsequent planning strategies. Critically, our measurement model can be inverted to infer the sequence of people's planning strategies from the clicks they make in the process tracing paradigm. Our computational method makes it possible to observe how people's planning strategies change from each decision to the next. This sheds new light on the time course and the nature of metacognitive learning. Future work will reverse-engineer the learning mechanisms that generate the cognitive plasticity our approach is bringing to light.

The plan for this paper is as follows: we start by developing a computational method for measuring experience-dependent changes in people's planning strategies. Next, we validate it on synthetic data and human data. We then illustrate the utility of our method by measuring the time course of how people learn how to plan, characterizing the revealed learning trajectories, and testing hypotheses about cognitive plasticity. In closing, we discuss directions for future work.

Methods

Process-tracing using the Mouselab-MDP paradigm

Planning, like all cognitive processes, cannot be observed directly but has to be inferred from observable behavior. This is generally an ill-posed problem. To address this challenge, researchers have developed *process-tracing* methods that elicit and record behavioral signatures of latent cognitive processes; for instance decision strategies can be traced by recording the order in which people inspect the payoffs of different gambles (Payne, Bettman, & Johnson, 1993). While these behavioral signatures are still indirect measures of cognitive processes, they do provide additional information about

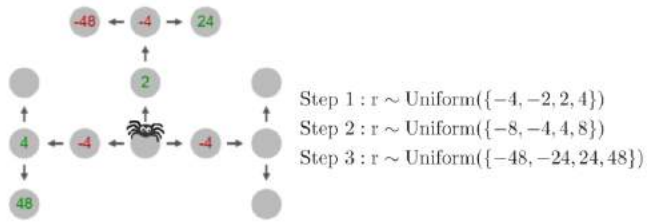


Figure 1: Illustration of the Mouselab-MDP paradigm. Rewards are revealed by clicking, prior to selecting a path with the arrow keys. The distribution of rewards underlying each node at a given step is shown on the right.

what the underlying cognitive strategy might be.

Here, we employ a process-tracing paradigm that externalizes people's beliefs and planning operations as observable states and actions (Callaway, Lieder, Krueger, & Griffiths, 2017; Callaway et al., 2018). Inspired by the Mouselab paradigm (Payne et al., 1993), the Mouselab-MDP paradigm uses people's mouse-clicking as a window into their planning.

The Mouselab-MDP paradigm illustrated in Figure 1 presents a series of route planning problems where each location (the gray circles), harbors a gain or loss. These potential gains and losses are initially occluded, corresponding to a highly uncertain belief state. The participant can reveal each location's reward by clicking on it and paying a fee. This is similar to looking at a map to plan a road trip. Clicking on a circle corresponds to thinking about a potential destination, evaluating how enjoyable it would be to go there, and adjusting one's assessment of candidate routes accordingly.

Measurement model

To develop an efficient computational method for inferring the temporal evolution of people's planning strategies, we make the simplifying assumption that the trial-by-trial sequence of peoples' cognitive strategies $(S_1, S_2, \cdots, S_{31})$ forms a Markov chain whose hidden states emit the observed process tracing data collect on each trial $(\mathbf{d}_1, \cdots, \mathbf{d}_{31})$. This hidden Markov model requires additional methodological assumptions about i) how cognitive strategies manifest in process-tracing data, ii) the space of cognitive mechanisms that can be learned, and iii) the nature and amount of cognitive plasticity that might occur. The following paragraphs detail our assumptions about each of these three components in turn.

Observation model. To plan in the Mouselab-MDP paradigm participants have to gather information by making a sequence of clicks. Our observation model thus specifies the probability of of observing a sequence of clicks $\mathbf{d_t}$ on trial t if the strategy was S_t (i.e., $P(\mathbf{d_t}|S_t)$).

To achieve this, we quantify each planning strategy's propensity to generate a click c (or stop collecting information) given the already observed rewards encoded in belief state b by a weighted sum of 29 features $(f_1(b,c), \cdots, f_{29}(b,c))$. The features describe the click c relative to this information (e.g., by the value of the largest reward that can be collected from the inspected location) and in terms of the action it gathers information about (e.g., whether it pertains to the first, second, or third step)[1]. The *depth* feature, for instance, describes each click by whether it looks 1, 2, or 3 steps into the future. The features and weights jointly determine the strategy's propensity to make click c in belief state b according to

$$P(\mathbf{d_t}|S_t) = \prod_{i=1}^{|\mathbf{d_t}|} \frac{\exp\left(\frac{1}{\tau} \cdot \sum_{k=1}^{|w^{(S)}|} w_k^{(S)} \cdot f_k^{(S)}(c_{t,i}, b_{t,i})\right)}{\sum_{c \in \mathcal{C}_{b_t}} \exp\left(\frac{1}{\tau} \cdot \sum_{k=1}^{|w^{(S)}|} w_k^{(S)} \cdot f_k^{(S)}(c, b_{t,i})\right)}, \quad (1)$$

where $d_{t,i}$ is the i$^{\text{th}}$ click the participant made on trial t (or the decision to stop clicking and take action), the decision temperature τ was set to 0.5 to match the variability of people's click sequences, and $w^{(S)}$ is the weight vector of strategy S.

Space of cognitive mechanisms. We formulated a set of 38 strategies $(\mathcal{S})$[1] to describe the process tracing data from Lieder (2018). These strategies include the optimal goal-setting strategy (Callaway et al., 2018) that starts by inspecting the possible final destinations and search-based planning algorithms such as breadth-first search, depth-first search, and best-first search (Russell & Norvig, 2016). 76.7% of the click sequences were the most likely instantiation of one of the 38 strategies. The clicks of the remaining 23.3% of the sequences were, at worst, second most likely under the best fitting strategy. These strategies differ in how much information they consider (ranging from none to all), which information they focus on, and in the order in which they collect it.

Building on the observation model in Equation 1, we represent each strategy by a weight vector $\mathbf{w} = (w_1, \cdots, w_{29})$ that specifies the strategy's preference for more vs. less planning, considering immediate vs. long-term consequences, satisficing vs. maximizing, avoiding losses (cf. Huys et al., 2012), and other desiderata. These weights span a high-dimensional continuous space with many intermediate strategies and mixtures of strategies. Cognitive plasticity could be measured by tracking how those weights change over time. But this would be a very difficult ill-defined inference problem whose solution would depend on our somewhat arbitrary choice of features. As a first approximation, our method therefore simplifies the problem of measuring cognitive plasticity to inferring a time-series of discrete strategies.

To understand what types of strategies people use, we grouped our 38 strategies using hierarchical clustering. This requires measuring the similarity between strategies. Since the strategies are probabilistic, we defined the distance metric $\Delta(s_1, s_2)$ between strategy s_1 and s_2 as the Jensen-Shannon

[1] A detailed description of the features and strategies is available at https://osf.io/y58d3/?view_only=fa2f89de3aa04d4d87af3d050bb1a64c

divergence (Lin, 1991) between the distributions of click sequences and belief states induced by strategies s_1 and s_2 respectively, that is

$$\Delta(s_1, s_2) = \mathrm{JS}\left[p(\mathbf{d}|s_1), p(\mathbf{d}|s_2)\right], \tag{2}$$

and approximate it using Monte-Carlo integration.

Applying Ward's hierarchical clustering method (Ward Jr, 1963) to the resulting distances suggested 11 types of planning strategies: acting impulsively without any planning, finding a goal and immediately moving towards it, inspecting both immediate and final outcomes (but no intermediate ones), overly frugal goal setting strategies, goal setting strategies that plan towards potential goals even when it is wasteful, exhaustive backward planning strategies that inspect all of the states, other far-sighted strategies that inspect all potential final states, forward-planning strategies similar to depth-first search, forward-planning strategies similar to best-first search, strategies similar to breadth-first search, and strategies that focus on the course of action that has received the most consideration so far.

Prior on strategy sequences. Inferring a strategy from a single click sequence could be unreliable. Our method therefore exploits temporal dependencies between subsequent strategies to smooth out its inferences. Transitions from one strategy to the next can be grouped into three types: repetitions, gradual changes, and abrupt changes. While most neuroscientific and reinforcement-learning perspectives emphasize gradual learning (e.g., Hebb, 1949; Mercado III, 2008; Lieder, Shenhav, Musslick, & Griffiths, 2018), others suggest that animals change their strategy abruptly when they detect a change in the environment (Gershman, Blei, & Niv, 2010). Symbolic models and stage theories of cognitive development also assume abrupt changes (e.g., Piaget, 1971; Shrager & Siegler, 1998), and it seems plausible that both types of mechanisms might coexist. To accommodate these different perspectives, we consider three prior distributions on participants' trial-by-trial sequence of cognitive strategies.

The *gradual learning prior* (m_{gradual} in Equation 3) assumes that strategies changes gradually, that is

$$P(S_{t+1} = s|S_t, m_{\text{gradual}}) = \frac{\exp(-\frac{1}{\tau} \cdot \Delta(s, S_t))}{\sum_{s' \in \mathcal{S}} \exp(-\frac{1}{\tau} \cdot \Delta(s', S_t))}, \tag{3}$$

where $\mathcal{S}$ is the set of strategies, $|\mathcal{S}|$ is the number of strategies, and the temperature parameter τ was set to achieve a 50% chance of a strategy change. By contrast, the *abrupt changes prior* (m_{abrupt} in Equation 4) assumes that transitions are either repetitions or jumps.

$$\begin{aligned} &P(S_{t+1} = s|S_t, m_{\text{abrupt}}) = \\ &\quad p_{\text{stay}} \cdot \mathbb{I}(S_{t+1} = S_t) + (1 - p_{\text{stay}}) \cdot \frac{\mathbb{I}(s \neq S_t)}{|\mathcal{S}|-1}, \end{aligned} \tag{4}$$

Finally, the *mixed prior* (m_{mixed} in Equation 5) assumes that both types of changes coexist.

$$\begin{aligned} &P(S_{t+1} = s|S_t, m_{\text{mixed}}) = \\ &\quad p_{\text{gradual}} \cdot P(S_{t+1} = s|S_t, m_{\text{gradual}}) \\ &\quad + (1 - p_{\text{gradual}}) \cdot P(S_{t+1} = s|S_t, m_{\text{abrupt}}). \end{aligned} \tag{5}$$

In each of these three cases, we model the probability of the first strategy as a uniform distribution over the space of decision strategies (i.e., $P(S_1) = \frac{1}{|\mathcal{S}|}$).

Together with the observation model and the strategy space described above each of these priors defines a generative model of a participant's process tracing data $\mathbf{d}$; this model has the following form:

$$P(\mathbf{d}, S_1, \cdots, S_T) = \frac{1}{|\mathcal{S}|} \cdot \prod_{t=2}^{T} P(S_t|S_{t-1}, m) \cdot P(\mathbf{d_t}|S_t). \tag{6}$$

The three measurement models differ in the identity of $m \in \{m_{\text{gradual}}, m_{\text{abrupt}}, m_{\text{mixed}}\}$. Inverting these models gives rise to a computational method for measuring an important aspect of cognitive plasticity.

Inference on cognitive plasticity

The models above describe how changes in cognitive strategies manifest in process-tracing data. To measure those cognitive changes, we have to reason backwards from the process tracing data $\mathbf{d}$ to the unobservable cognitive strategies $S_1, \cdots, S_T$ that generated it. To achieve this, we leverage the Viterbi algorithm (Forney, 1973) to compute maximum a posteriori (MAP) estimates of the hidden sequence of planning strategies $S_1, \cdots, S_T$ given the observed process tracing data $\mathbf{d}$, the measurement model m, and its parameters (p_{stay} for m_{abrupt} and p_{gradual} and p_{stay} for m_{mixed}). To estimate the model parameters we perform grid search with a resolution of 0.02 over $p_{\text{stay}} \in [0, 1]$ for m_{abrupt} and $(p_{\text{stay}}, p_{\text{gradual}}) \in [0, 1] \times [0, 1]$ for m_{mixed}.

Inferring the hidden sequence of cognitive strategies in this way lets us see otherwise unobservable aspects of cognitive plasticity through the lens of a computational microscope.

Validating the computational microscope

Validation on synthetic data

To validate our "computational microscope" for looking at cognitive plasticity, we apply it to simulated process tracing data. To avoid bias towards any one of the three measurement models, we used each of them to generate a data set with 100 simulated participants completing 31 trials each. We then combined the resulting three data sets into a single data set from 300 simulated participants.

We then inverted the three measurement models on each of the simulated trials ($\mathbf{d}$) and compared the maximum a posteriori estimate of each strategy sequence ($\hat{\mathbf{S}}$) against the ground truth (S) in terms of the proportion of correctly inferred strategies and the distance between the inferred strategies and the ground truth. To measure the distance between

two sequences of n planning strategies we define $\Delta(\mathbf{v},\mathbf{w})$ as $\frac{1}{n}\cdot\sum_{i=1}^{n}\Delta(v_i,w_i)$. For better interpretability, the relative distance $\Delta_{\text{rel}}(s_1,s_2)=\Delta(s_1,s_2)/\bar{\Delta}$ normalizes $\Delta(s_1,s_2)$ by the average distance between any strategy and its closest neighbour.

As a baseline, we evaluated the computational method that inverts the observation model in Equation 1 on each click sequences independently. This simple approach was sufficient to infer the correct strategy about 81% of the time (95% confidence interval: $[80.2\%,81.8\%]$). The average distance from the inferred strategy to the true one was only 21% of the average distance from each strategy to its closest neighbor ($\Delta_{\text{rel}}(\hat{\mathbf{s}}^{\text{baseline}},\mathbf{s})=0.215$, 95% confidence interval: $[0.20,0.23]$). This shows that the simulated click sequences were highly diagnostic of the strategies that generated them.

We found that exploiting the temporal dependencies among subsequent strategies by using either of the three measurement models significantly improved the proportion of correctly inferred strategies to 88.5%, 88.3%, and 88.5% for m_{gradual}, m_{abrupt}, and m_{mixed} respectively (all $p<0.0001$) and decreased the average distance between the inferred strategies and the ground truth by more than 40% ($\Delta_{\text{rel}}(\hat{\mathbf{s}}^{\text{gradual}},\mathbf{s})=0.124$, $\Delta_{\text{rel}}(\hat{\mathbf{s}}^{\text{mixed}},\mathbf{s})=0.124$, and $\Delta_{\text{rel}}(\hat{\mathbf{s}}^{\text{abrupt}},\mathbf{s})=0.127$, all $p<0.0001$). The minor differences between the accuracies and distances achieved with the three measurement models were not statistically significant ($\chi^2(2)=0.36$, $p=0.8373$ and $F(2,897)=0.06$, $p=0.942$ respectively). These results suggest that – under reasonable, theory-agnostic assumptions about what cognitive plasticity might be like – our computational microscopes for looking at cognitive plasticity can be expected to produce more accurate measurements than simpler methods.

Which measurement model is most suitable depends on whether the measured changes are mostly gradual, mostly abrupt, or a combination of both. This may vary across tasks and participants. We therefore invert all three measurement models on each participant's data and select the most appropriate measurement model for each participant according to the Akaike Information Criterion (Akaike, 1974). We then interpret the inferences obtained from inverting the selected model as the measurement of our computational microscope.

Validation on empirical data

To validate our computational microscope on empirical data, we applied it to the Mouselab-MDP process-tracing data from Experiments 1–3 by Lieder (2018) where 176 participants solved 31 different 3-step planning problems of the form shown in Figure 1. Concretely, we asked if our computational microscope can detect the effect of an experimental manipulation expected to promote cognitive plasticity, namely the feedback participants in the second condition of Experiment 1 received on the (sub)optimality of their chosen actions. This performance feedback stated whether the chosen move was sub-optimal and included a delay penalty whose duration was proportional to the difference between the expected returns of the optimal move versus the chose one.

Our computational microscope successfully detected the effect of this manipulation. As shown in Figure 2, the inferred learning-induced changes were significantly larger in the feedback condition than in the control condition in the first 15 trials and in trials 21–25 ($p\leq 0.012$ for each 5-trial bin) and nearly significant in trials 15–20 ($p=0.08$) and trials 25–30 ($p=0.06$). Furthermore, Figure 2 also shows that cognitive plasticity slowed down over time as participants adapted to experiment's stationary decision environment.

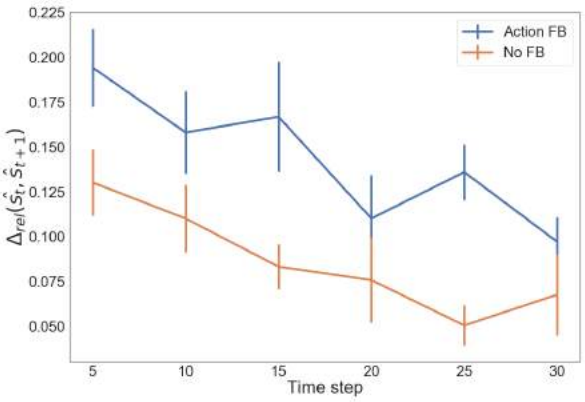


Figure 2: Feedback accelerates cognitive plasticity. This figure shows that feedback increased the amount of cognitive plasticity at the beginning of learning.

Next, we performed χ^2-tests with the Sidak correction for multiple comparisons to compare the frequencies of all possible strategy transitions (i.e., $P(S_{t+1}|S_t)$) between the experimental condition with action feedback versus the control condition. We found that action feedback selectively increased the probability of eight performance-increasing transitions from a strategy with a lower average performance (S_t) to a strategy with a higher average performance (S_{t+1}) and significantly decreased the probability of five performance-decreasing transitions and five strategy repetitions ($S_{t+1}=S_t$). By contrast, the feedback decreased the frequency of only one performance-increasing strategy-transition and increased the frequency of only two performance-decreasing strategy transitions.

Our method's ability to detect the plasticity-enhancing effects of feedback suggests that its inferences provide a valid measure of cognitive plasticity.

Shedding light on cognitive plasticity

Having validated our computational microscope on both simulated and empirical data, we now leverage it to measure how people learn how to plan by applying it to the process tracing data from the control conditions of Experiment 1 and the training phases of the control conditions of Experiments 2 and 3 from Lieder (2018). In the following, we illustrate how our computational microscope can be used to i) measure how people's propensity to use different cognitive strate-

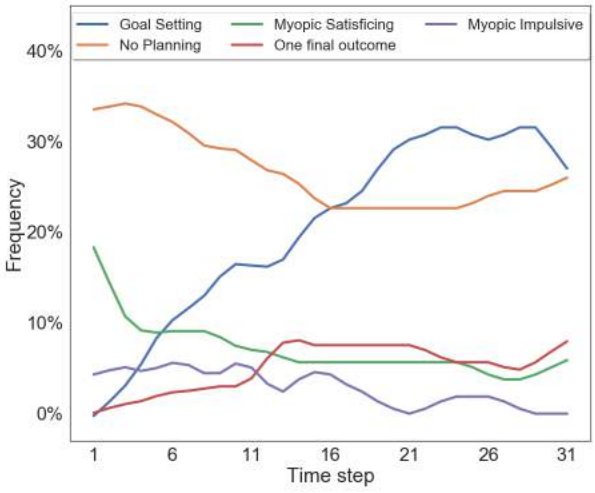


Figure 3: Time course of strategy usage frequencies of the five most common strategies.

gies evolves over time, ii) test theories of cognitive development and cognitive plasticity, and iii) characterize people's metacognitive learning trajectories.

Temporal evolution of strategy frequencies. As shown in Figure 3, we found that the most common initial strategy was to act impulsively without any planning (*No Planning*). The prevalence of this strategy decreased gradually over time from about 34% on average across the first five trials to about 25% on average across the last five trials ($\chi^2(1) = 7.95$, $p = 0.0048$).[2]. Conversely, the frequency of the near-optimal *Goal Setting* strategy increased from about 4% to 30% ($\chi^2(1) = 148.85, p < 0.0001$). The frequencies of the two maladaptive strategies that decide based on immediate rewards (*Myopic Satisficing* and *Myopic Impulsive*) dropped from about 11% and 4% respectively to about 5% ($\chi^2(1) = 11.74, p = 0.0006$) and 0.6% ($\chi^2(1) = 11.62, p = 0.0006$) respectively, whereas the frequency of the strategy *One Final Outcome* that prioritizes long-term consequences increased from about 1% to about 6% ($\chi^2(1) = 20.22, p < 0.0001$). Jointly these strategies accounted for about 53%–72% of our participants' planning across the different trials of our experiment.

Testing hypotheses about the nature of cognitive plasticity. Prominent theories of cognitive development disagree about whether it proceeds in discrete stages (Piaget, 1971) with abrupt transitions or continuous gradual change (Siegler, 1996). Inspired by these theories, we asked to which extent learning how to plan in the Mouselab-MDP paradigm proceeds through gradual changes versus abrupt transitions. Our computational microscope suggested that cognitive plasticity includes both gradual and abrupt strategy changes. We observed that the data from $63.0\% \pm 4.9\%$ of our participants was best captured by the abrupt model, while the data from $29.8\% \pm 4.6\%$ of the participants were best captured by the gradual model, and the data from $7.2\% \pm 2.6\%$ were best captured by the mixed model. A more fine-grained analysis of the individual inferred transitions revealed that the majority of strategy changes was gradual (i.e., 59.1%, $\chi^2(1) = 56.8$, $p < 0.0001$) but there was also a non-negligible percentage of abrupt changes (i.e., 40.9%). In total those different types of strategy changes constituted 22.8% of all transitions; that is 77.2% of the inferred transitions were strategy repetitions.

Siegler's overlapping waves theory (Siegler, 1996) asserts that multiple cognitive strategies are being used in parallel at each time during cognitive development. It further asserts that the relative frequencies of these strategies shift towards increasingly more adaptive strategies and that there are intermediary strategies whose frequency waxes and vanes. Under the strong assumption that the underlying plasticity mechanisms are the same as those that drive learning in the Mouselab-MDP paradigm, we predicted that the same patterns should also occur in the participants' strategy sequences. To test the first prediction, we performed χ^2-tests on the strategies' frequencies in all bins of 5 consecutive trials. In support of the hypothesis that multiple different strategies are used at each point in time throughout the learning process we found that on average 2.16 strategies were each used by significantly more than 5% of our participants in any given trial of the experiment (95% confidence interval: $[2.02, 2.30]$). Consistent with the prediction that high-performing strategies become more prevalent over time whereas low-performing strategies become less prevalent over time we found a significant rank correlation between each strategies' average performance and the change in their frequency from the first trial to the last trial (Spearman's $\rho(37) = 0.39$, $p = 0.0154$). On the population level, we did not find any evidence for intermediary strategies whose average frequency across participants initially increases and later decreases again. That is, there was no strategy whose frequency was higher in the middle two time bins than in both the first two time bins and the last two time bins. Yet, overall the measurements we obtained with our computational microscope suggest that learning in the Mouselab-MDP paradigm is better described by the overlapping waves theory than by stage theories of cognitive development.

Learning trajectories. To identify the most common learning trajectories, we categorized each inferred strategy as belonging to one of the 11 types of strategies described earlier. We then extracted the order in which different strategy types appeared in the inferred sequences. Using this analysis, we found that there were almost as many unique learning trajectories as there were learners: The 110 participants

[2] All χ^2-tests in this paragraph compare the average frequency in the first five trials against the average frequency in the last five trials.

who changed their strategy at least once displayed 94 unique learning trajectories; that is 85.4% of the learning trajectories were unique and the remaining trajectories were exhibited by only 2–4 learners each. Zooming in on the 49 participants who learned the near-optimal goal setting strategy, we found that they reached the near-optimal goal setting strategy via 38 unique learning trajectories. Consistent with the overlapping waves theory, we found that 84.2% of these learning trajectories included at least one intermediary strategy between the initial strategy and the final strategy. Most importantly, our analysis revealed three dominant gateways to optimal planning: 35% of the penultimate strategies inspected all potential final states – whereas the optimal strategy stops searching for better final states once it encounters the best possible outcome – and sometimes planned backwards from undesirable states; 27% of the penultimate strategies inspected the potential final states in a manner akin to the optimal strategy but additionally and wastefully inspected paths towards undesirable final outcomes, and 21% of the penultimate strategies inspected both immediate and final outcomes while ignoring the intermediate states. This suggests that participants discovered the optimal goal setting strategy via intermediate strategies that perform gratuitous planning. Furthermore, we found that about 42% of participants who succeeded to learn a near-optimal goal setting strategy started with strategies that inspect both immediate and final outcomes without looking at intermediate ones. In addition to the 110 participants who changed their initial strategy, 66 participants (37.5%) never changed their strategy. The majority of those participants always acted impulsively without any planning (21% of all participants). Consistent with the interpretation that those participants were less engaged in the experiment and had not paid close attention to the instructions, we found that they performed substantially worse on the four attention check questions at the end of the experiment than participants who had demonstrated learning (1.7 errors vs. 0.8 errors on average; $t(111) = -5.80, p < .0001$). In addition, 9% of all participants always inspected immediate and final outcomes while ignoring intermediate rewards, 4% always focused exclusively on final outcomes, and 3.5% used other types of strategies.

Discussion

We have successfully validated our method on both synthetic and human data. The results suggest that our computational microscope can measure cognitive plasticity in terms of the temporal evolution of people's cognitive strategies.

Our findings suggest that this method has great potential for helping cognitive scientists uncover the mechanisms of cognitive plasticity and how they are impacted by the learning environment, individual differences, time pressure, motivation, and interventions – including feedback, instructions, and reflection prompts.

We are optimistic that computational microscopes will become useful tools for reverse-engineering the learning mechanisms that enable people to acquire complex cognitive skills and shape the way we think and decide. To make this possible, we will extend the proposed measurement model to continuous strategy spaces, a wider range of tasks and strategies, and learning at the timescale of individual cognitive operations. In addition, future work will also leverage our computational microscope to elucidate individual differences in cognitive plasticity within and across psychiatric conditions and different age groups.

The tentative conclusions we obtained with our first prototype of a computational microscope for measuring cognitive plasticity should be taken with a grain of salt because more psychologically plausible distance metrics and more realistic strategy representations could lead to different conclusions about the nature of cognitive plasticity. In this first step, we determined the similarity between strategies based on their behavior. But two strategies that look very different could result from similar mechanisms. Future work will identify a low-dimensional continuous strategy space by decomposing each strategy into its Pavlovian, habitual, and model-based components (van der Meer, Kurth-Nelson, & Redish, 2012). This more realistic representation will allows us to measure the similarity between strategies by comparing the underlying neurocomputational mechanisms. In addition, we will seek to validate the robustness of our computational microscope by measuring its performance on data generated from more realistic models of cognitive plasticity (e.g., Krueger et al., 2017; Lieder et al., 2018).

The approach developed in this paper makes it possible to more directly observe the previously hidden phenomenon of cognitive plasticity in many of its facets – ranging from skill acquisition, learning to think differently, reflective learning, cognitive decline, self-improvement, changes in cognitive dispositions, and the onset, progression, and recovery from psychiatric symptoms and mental disorders. This will make it easier to reverse-engineer people's ability to discover and continuously refine their own algorithms.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, *19*(6), 716–723.

Callaway, F., Lieder, F., Das, P., Gul, S., Krueger, P. M., & Griffiths, T. L. (2018). A resource-rational analysis of human planning. In *Proceedings of the 40th annual conference of the cognitive science society.*

Callaway, F., Lieder, F., Krueger, P. M., & Griffiths, T. L. (2017). Mouselab-mdp: A new paradigm for tracing how people plan. In *The 3rd multidisciplinary conference on reinforcement learning and decision making.*

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological review*, *100*(3), 363.

Forney, G. D. (1973). The Viterbi algorithm. *Proceedings of the IEEE*, *61*(3), 268–278.

Gershman, S. J., Blei, D. M., & Niv, Y. (2010). Context, learning, and extinction. *Psychological review*, *117*(1), 197.

Hebb, D. O. (1949). *The organization of behavior: a neuropsychological theory.* John Wiley & Sons Inc.

Hubel, D. H., & Wiesel, T. N. (1970). The period of susceptibility to the physiological effects of unilateral eye closure in kittens. *The Journal of physiology*, *206*(2), 419–436.

Huys, Q. J., Eshel, N., O'Nions, E., Sheridan, L., Dayan, P., & Roiser, J. P. (2012). Bonsai trees in your head: how the pavlovian system sculpts goal-directed choices by pruning decision trees. *PLoS computational biology*, *8*(3), e1002410.

Krueger, P. M., Lieder, F., & Griffiths, T. L. (2017). Enhancing metacognitive reinforcement learning using reward structures and feedback. In *Proceedings of the 39th annual conference of the cognitive science society.*

Lieder, F. (2018). Developing an intelligent system that teaches people optimal cognitive strategies. In F. Lieder (Ed.), *Beyond bounded rationality: Reverse-engineering and enhancing human intelligence* (chap. 8).

Lieder, F., & Griffiths, T. L. (2017). Strategy selection as rational metareasoning. *Psychological Review*, *124*(6), 762–794. doi: 10.1037/rev0000075

Lieder, F., Shenhav, A., Musslick, S., & Griffiths, T. L. (2018). Rational metareasoning and the plasticity of cognitive control. *PLoS computational biology*, *14*(4), e1006043.

Lin, J. (1991). Divergence measures based on the Shannon entropy. *IEEE Transactions on Information theory*, *37*(1), 145–151.

Mercado III, E. (2008). Neural and cognitive plasticity: From maps to minds. *Psychological Bulletin*, *134*(1), 109.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The adaptive decision maker*. Cambridge university press.

Piaget, J. (1971). *The theory of stages in cognitive development.* McGraw-Hill.

Russell, S. J., & Norvig, P. (2016). *Artificial intelligence: a modern approach* (3rd ed.). Harlow, UK: Pearson Education Limited.

Shrager, J., & Siegler, R. S. (1998). SCADS: A model of children's strategy choices and strategy discoveries. *Psychological Science*, *9*(5), 405–410. doi: 10.1111/1467-9280.00076

Siegler, R. S. (1996). *Emerging minds: The process of change in children's thinking*. New York: Oxford University Press.

van der Meer, M., Kurth-Nelson, Z., & Redish, A. D. (2012). Information processing in decision-making systems. *The Neuroscientist*, *18*(4), 342–359.

van Lehn, K. (1996). Cognitive skill acquisition. *Annual review of psychology*, *47*(1), 513–539.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, *58*(301), 236–244.

Interacting physically with insight problems does not affect problem solving process

Jan Jastrzębski
Hanna Kucwaj
Institute of Psychology, Jagiellonian University
Ingardena 6, 30-060 Krakow, Poland

Adam Chuderski
Institute of Philosophy, Jagiellonian University
Grodzka 52, 31-044 Krakow, Poland

Abstract

So-called insight problems are believed to tap into sudden, creative thinking that is crucial for real problems. In contrast, recent findings suggest that solving insight problems depends on the same cognitive mechanisms that underpin systematic, analytical thinking. However, existing studies may have low ecological validity, because insight problems were usually presented in static formats (on paper, computer screen) which allowed no physical interaction with the problem elements. This study administered 8 established insight problems either in the static or interactive variants. It also probed two markers of analytical thinking: working memory capacity and reasoning ability. Virtually no difference in performance was observed between the static and interactive variants of insight problems with regard to (1) solution rate, (2) subjective experience of suddenness, pleasure, and relief accompanying the solutions, as well as (3) correlations with the working memory capacity and analytical reasoning tests. These results suggest that externalized/embodied/situated factors play no substantial role in insight problem solving and the crucial parts of this process seem to occur in the mind of a solver.

Keywords: insight problem solving; analytical thinking; working memory; interactivity.

Introduction

An important category of problems investigated in the problem solving literature is so-called insight problems. Such problems are defined in the vague and misleading way that suggests a typical but wrong problem representation, so following this representation often results in an impasse. The correct solution can be found only when the problem is viewed from a novel perspective and can be appropriately restructured. Especially difficult are problems that require rejecting one strongly believed and subjectively obvious assumption that, however, is not implicated by the problem description (Knöblich, Ohlsson, Haider, & Renius, 1999). For example, when instructed to transform an incorrect equation including Roman numerals made of matchsticks: "VI = VI + VI" into a correct equation by moving just one matchstick (without adding or removing any matchsticks), people must realize that equations do not necessarily include only one equation sign and that two such signs can also be allowed, here resulting in the tautology "VI = VI = VI". Insight problems have been studied intensively in cognitive science and psychology because many authors believe that they tap into mental processes that also play a role in "full-blown" creative cognition, leading to great masterpieces, discoveries, and inventions (Ohlsson, 2011).

The crucial controversy is whether the processing underpinning insight problem solving is distinct from solving so-called analytical problems, such as complex but typical arithmetic equations, which are defined in a more precise way, and require more systematic construction of the problem representation, while including no tricky obstacles. Some evidence suggested that insight problem solving involves idiosyncratic processes: constraint relaxation, defocusing attention, and uncontrolled spread of activation in memory (Knöblich et al., 1999; Kounios & Beeman, 2014), and so relies minimally on cognitive resources such as executive control and working memory capacity that typically determine success on analytical problems (see Wiley & Jarosz, 2012). Other evidence highlighted a large overlap of attentional, control, memory, reasoning, and imagery processes for insight and non-insight problems (MacGregor, Ormerod, & Chronicle, 2001; Weisberg, 2015). Specifically, two recent meta-analyses suggested that individual success on insight problems is strongly correlated with performance on analytical problems as well as with executive control and working memory tasks (Chuderski & Jastrzębski, 2018a; Gilhooly & Webb, 2018).

However, such a similarity of insight and analytical thinking might result from the fact that most of the experiments to date presented insight problems in a static format, usually printed on a paper sheet or shown on a computer screen, and participants were not allowed to interact with the problem by manipulating its elements. For instance, in a typically administered matchstick arithmetic problem, there are no actual matchsticks to be manipulated; all transformations of the equation must proceed in the mind, and the potential solution has to be written down. This lack of interaction with the problem may to some extent impede more spontaneous, "fuzzy" cognition that might be crucial for creative solutions. Participants, forced to represent and explore the problem space solely in the mind, might be prone to using more systematic, gradual problem solving strategies typical for analytical problems, while in the contexts that are

more externalized/embodied/situated they switch to less systematic strategies, such as trial-and-error, remote associations, etc. Obviously, the former strategies are more strictly constrained by available attentional resources and working memory capacity, while cognitive load might be largely reduced when artefacts can be used. Also, as many real-life problems seem to be situated to a large extent (see Clark and Chalmers, 1998; Cowley & Vallée-Tourangeau, 2010), investigating insight problem solving using non-interactive paradigms may yield low ecological validity.

Interactive insight problem solving

Indeed, a few studies by F. Vallée-Tourangeau, who applied insight problems in such a way that problems elements could be manipulated, as compared to static variants, have shown that solutions occur more frequently when the problem can be interacted with. Substantial effects, reaching the doubled solution rates, have been reported for the well-known insight problems such as the cheap necklace (Henok, Vallée-Tourangeau, & Vallée-Tourangeau, 2018; see also Fioratou & Cowley, 2009), the triangle of coins (Vallée-Tourangeau, 2017), the anagrams (Vallée-Tourangeau & Wrightman, 2011), the animals in pens (Vallée-Tourangeau, Steffensen, Vallée-Tourangeau, & Sirota, 2016), Luchins' water jars (Vallée-Tourangeau, Euden, & Hearn, 2011), and matchstick arithmetic (Weller, Villejoubert, & Vallée-Tourangeau, 2011). Also, some studies reported no difference in working memory capacity between solvers and non-solvers in the situated context. All this suggests that cognitive processing may change substantially in the embodied/situated contexts.

Besides the fact that virtually all these data (except for Fioratou & Cowley, 2009) come from one and the same lab, and thus require independent replication, existing evidence needs to be extended for at least three reasons. First, each study examined a single insight problem, applied either in the computerized/paper format or in the interactive format. As different samples of participants were used in consecutive studies, it is not possible to compare across the problems the size of presumed benefit from interactivity. (Do all problems benefit equally?) Second, recent studies (Danek, Wiley, & Öllinger, 2016; Fleck & Weisberg, 2013) probed experience during solution (asked how sudden and surprising it was), and suggested that many insight problems, originally designed to require sudden restructuring, by some participants could be solved in a fully systematic, gradual way. Thus, probably no insight problem always elicits "pure" insight. Unfortunately, so far subjective measures of insight have not been combined with examination of interactivity. Examining if interactivity can affect the subjective experience of insight might reveal mechanisms facilitating solutions. Finally, because to date, single problems were studied, the resulting binary dependent variables prohibited a proper analysis of correlations between performance on insight problems, analytical problems, and working memory tests. (Do interactive variants correlate with cognitive aptitude more weakly than the static variants?) All these research goals have important ramifications for our understanding of insight problem solving.

To tackle these three goals, the present study applied 8 popular insight problems. They were organized in 4 pairs of comparable problems. In each pair, one problem was shown in a typical, paper-and-pencil format, while the other problem was applied in a way that allowed manipulating the artifacts comprising this problem. Which problem from each pair was applied in the static format, and which was applied in the interactive way, was randomized across the sample. This fact allowed the within-subjects manipulation of the presentation format that gave control over group differences in general performance. Moreover, the size of the expected interactivity effect could be compared across the problems, in order to see if the problems differ in how strongly they benefit from externalizing. Additionally, after each solution given to an insight problem, the four-dimensional scale that probed the subjective experience of suddenness, pleasure, relief, and certainty accompanying the solution, was applied in order to see if the surplus solutions, which were expected to occur in the interactive problem format, would consist primarily of solutions assessed subjectively as the Aha! experience. Finally, an established working memory test and a hallmark analytic reasoning test were applied in order to compare whether the correlations of these two measures with the interactive variants could really be weaker than the respective correlations with the static variants, the latter presumed to load more substantially on cognitive resources.

The study

Participants

The total sample included 64 people (34 females; aged 19 to 39, M = 25.8 y, SD = 5.3 y). All participants were recruited from the general population via internet adverts and paid an equivalent of 12 USD in local currency. They signed a written consent to participate, were screened for normal or corrected-to-normal vision and no history of neurological problems, and were informed that they could stop the experiment and leave the lab at will. Data were anonymized. All other procedural aspects of the study conformed to the WMA's Declaration of Helsinki.

Insight problems

Matchstick arithmetic. Two matchstick arithmetic problems consisted of incorrect arithmetic equations written using Roman numerals. One problem was the above described "VI + VI = VI" equation. The second problem (I = II – II) required introducing a negative number (not a typical Roman numeral) by changing one of the sticks into the minus sign. The instructions were: "This equation consists of Roman numerals made of sticks. Unfortunately, the equation is wrong! Move exactly one stick so that the equation becomes correct. The allowed operations are „–," „+" and „=." You can't remove any stick. Upright sticks and tilted sticks are not

interchangeable („|/" is not „\/")." In order to familiarize the participants with the Roman numerals, the instruction contained also a table linking each Arabic number with its Roman equivalent, up to number ten. In the interactive format, the equations were constructed out of plastic sticks.

Triangle of coins/Eight coins. In the first problem, the participants were presented with a triangle facing upwards composed of 10 coins, and their task was to "Move exactly 3 coins to make the triangle point downward." In the eight coins problem, the participants were presented with a figure composed of 8 coins, and they had to "Move exactly 2 coins so that each of the 8 coins touches exactly 3 coins," which requires realizing that the coins have to form 3D piles. Both configurations require breaking constraints (of the X-axis rotation and 2D solution, respectively). In the interactive formats of the two tasks, the initial configurations were composed of real coins that could be manipulated freely. In the triangle problem, the response included presenting to the research assistant all the steps that had led to the solution.

Sheep in pens/Nine dots. In the first problem, the task was to "Close the 11 sheep in 4 pens so that in each pen there is an odd number of sheep." In the interactive format, the participants were given 11 small cloth figures of sheep and 4 pieces of string. As it is impossible to divide the number 11 into any combination of 4 odd numbers, the solution required embedding at least one of the pens inside another pen. In the nine dots problem, a 3×3 array of dots was presented and the task was to "Connect all the 9 dots with a broken line composed of 4 straight lines so that each following straight line begins at the end of the preceding line." In the correct solution, the lines should extend beyond the square shape of the array, but most people constrain themselves to explore only lines that fit within the array. In the interactive format, the participants were given tacks, 4 pieces of string, and a piece of paper with 9 dots printed.

Card split/Figure split. In the first task, participants were instructed to "Cut a hole in the card so that you can put your head through." In the figure split, problem participants were presented with an L-shape figure and the task was to "Divide the figure into four identical parts." Both problems require non-standard topological solutions. In the interactive formats of the tasks, participants were given scissors and several card/L-shape figures made of thick paper to experiment with.

Problems administration. In all the 8 problems, there was an identical instruction for each variant, and the variants differed only in the presentation method and response format. In the static variant, problems were given on a sheet of paper, with a blank space for making notes and drawings, and for providing a solution using a pen. In the interactive variants, the participants were given a cardboard box with respective objects placed on it, but were not provided with anything to write with, so they had to physically manipulate the objects provided. Participants were tested in individual cabins. The time limit for each problem was 5 minutes.

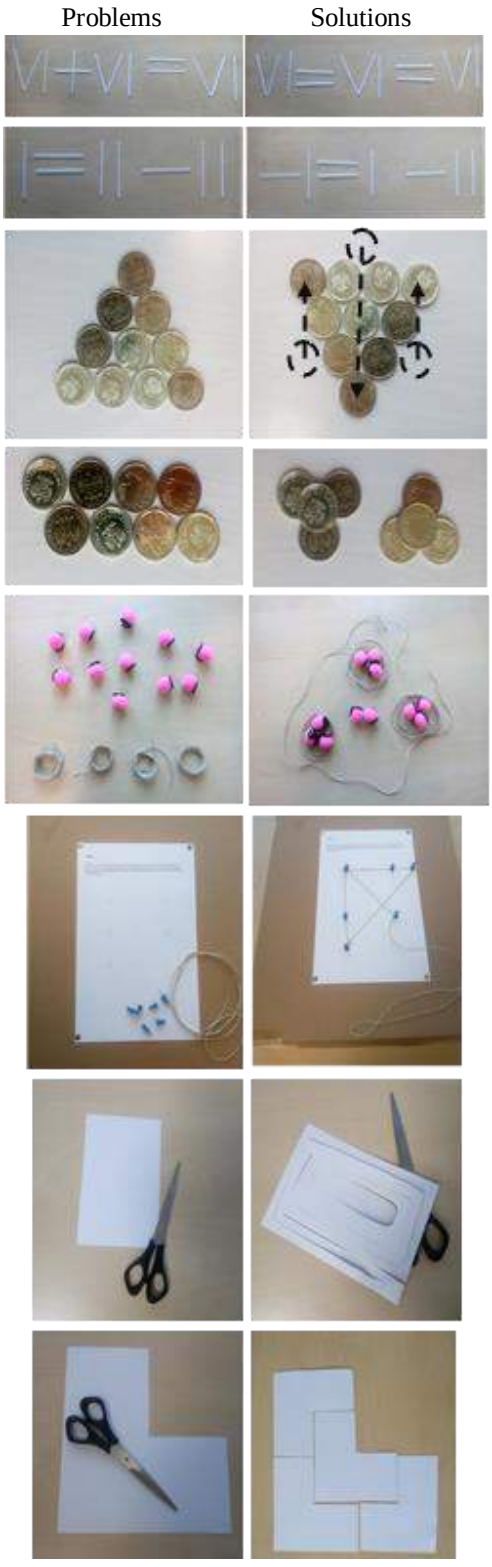


Fig. 1: The eight problems in initial configurations presented to the participants in the interactive format (left column) together with the sample correct solutions (right column). In the static format (not depicted) each problem elements were printed on a paper sheet, and a solution had to be drawn.

Subjective experience scale

The scale was modelled after Danek and Wiley (2016), who tested which dimensions of subjective experience best predict correct solutions to insight problems (suddenness, pleasure, relief, certainty). Here, the instruction was "Please describe your subjective experience at the moment when you found the solution to this problem", and the four questions were:
"The solution came to me…" (Gradually – Suddenly)
"At the moment of finding the solution my feelings were…" (Unpleasant – Pleasant),
"When I realised the solution I felt…" (Tension – Relief)
"My feeling that the solution was correct was…" (Uncertain – Certain)
Ratings were recorded on a 19-point graphical scale (line of cells) for which the contradicting words (e.g. Unpleasant – Pleasant) occupied the extremes. Point "10" was marked with the text "Don't know", and served for inconclusive cases.

Working memory task

The letter complex span required memorizing 4, 6, or 8 letters, which were drawn from 9 possible stimuli and were presented using a computer for 1.2 s apiece. After each letter presentation, participants indicated with a mouse button if a simple arithmetic equation (e.g., 2 × 3 – 1 = 5?) was correct. Then, they were to recall the letters in proper order. Five trials for each set size (in increasing order) were presented. The response procedure employed as many 3×3 matrices as was a particular set size. Each matrix contained all possible letters. Those letters that had been presented in a sequence should be selected in the matrices in the correct order. There was no time limit for responding. The dependent variable was the proportion of correctly selected letters.

Reasoning test

Raven Advanced Progressive Matrices (RAPM; Raven, Court, & Raven, 1983) consists of items that include a 3×3 matrix of figural patterns which is missing the bottom-right pattern, and 8 response options presenting the potentially matching patterns. The goal was to discover the rules that govern the distribution of patterns and to choose the response option including the correct pattern that completed the matrix according to these rules. The 18 odd-numbered items were given with the 20-min. time limit.

Procedure

Participants were tested in groups of 5 to 9 people. They first undertook RAPM and the letter complex span (as well as several other cognitive tests unrelated to the present study). Then, they attempted the 8 insight problems in the fixed order. A random half of the sample attempted the odd-numbered problems in the interactive variant and the even-numbered problems in the paper-and-pencil variant. The other half used the paper and pencil for the odd-numbered problems and the interactive formats for the even-numbered problems. The entire session lasted about 2 hours.

Results

No one was able to solve correctly the Card split and Letter split problems, so the analysis pertained to the 6 remaining problems. Participants admitted familiarity with 11 out of 384 problems applied, and these 11 problems were excluded from further analysis. Fig. 2 presents the number of correct solutions for each problem, for the static versus interactive format, separately. The Triangle of coins problem was the easiest one, solved by 37.5% of participants. In contrast, the 8 coins and the 9 dots problems were most difficult, solved only by 7.8% of the sample. These solution rates matched some existing data for the same problems (e.g., Chuderski & Jastrzębski, 2018b, 2018c). Importantly, for no problem the difference between the static and the interactive format was statistically significant. The largest numerical difference was observed for the Triangle of coins problem, which was solved by 15 people (out of 32) in the static format, and by 9 (out of 32) in the interactive format, but even this difference was far from reaching statistical significance, $\chi^2(1) = 1.50$, $p = .220$. Overall, 41 problems were solved with the paper and pencil, while 37 problems were solved in the interactive way, which is a totally non-significant difference.

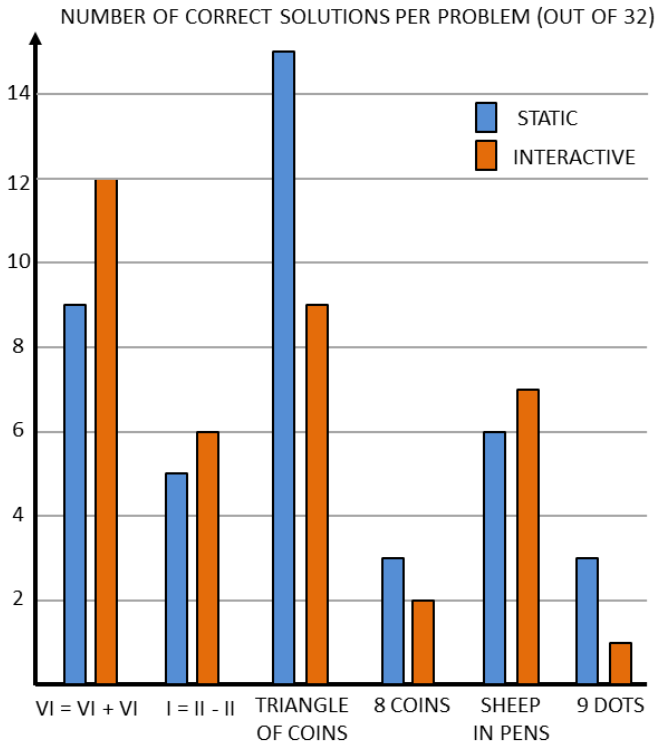


Fig.2: Number of insight problems (only those reported by the participants to be unfamiliar to them) correctly solved by the group who attempted a given problem in the static format (blue bars) vs. the group who undertook it in the interactive format (red bars). For no problem the difference between the conditions was significant at the $p < .05$ level.

Next, as it was possible that even though overall problem solving accuracy was not affected by the problem format, but it changed the way of processing the problems (at least the way subjectively experienced, and later reported, by the participants). Fig. 3 presents mean ratings for 4 indicators of insight: suddenness, pleasure, relief, and certainty, for 41 problems solved in the static format versus 37 interactive problems. Mean ratings ranging from 11 to 16 suggest that solutions yielded experience more typical for insight than for gradual, analytical processing. These ratings were submitted to MANOVA, with the problem variant (static vs. interactive) as a factor. Wilks' $\lambda = 0.917$ suggested no significant multivariate difference in experience between problem variants, $p = .173$. Second, single ratings were compared, with the Tukey correction for multiple comparisons. The only significant difference between the problem variants was noted for certainty, $F(1, 76) = 4.45$, $p = .038$, $\eta^2 = .06$, with interactive variants yielding 20% higher certainty of the correctness of the solution, as compared to the static variants. For the single problems, only the "VI = VI + VI" and the Triangle of coins problems yielded enough solutions (>20) so the accompanying reports could be compared meaningfully. For the former problem, significantly higher ratings in the interactive variant were observed for pleasure, $F(1, 20) = 7.58$, $p = .012$, $\eta^2 = .27$. No significant difference in ratings between variants was observed for the latter problem.

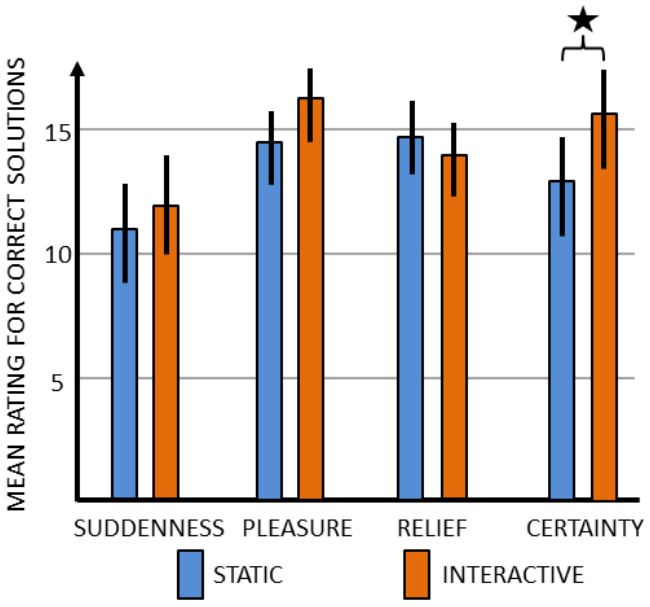


Fig. 3: Mean ratings for the reported subjective experience of suddenness, pleasure, relief, and certainty during correct solutions of insight problems, separately for the 41 problems in static variants versus the 37 problems in interactive variants. The star indicates a weak difference significant at the $p < .05$ level (for the experience of certainty). The three other differences were non-significant.

Finally, for each participant her or his score on all the 6 problems, the 3 problems applied in the static format, and the 3 problems applied in the interactive formats were calculated. The Spearman rank correlation was computed to assess the relationship between the number of problems solved and the letter complex span and RAPM scores. The resulting correlations are presented in the Table.

Table: Matrix of Spearman correlations between variables

Variable	1.	2.	3.	4.
1.All 6 problems	1			
2.Static variants	.756	1		
3.Interactive variants	.765	.195	1	
4.Complex span	.404	.316	.282	1
5.RAPM	.640	.520	.465	.353

Note. $N = 64$. All correlations significant at $p < .05$ except for the correlation between static and interactive variants.

Overall, correlations between the insight problem scores and the complex span (*rhos* ≈ .3) and RAPM (*rhos* ≈ .5) were substantial. However, the difference in correlation with the complex span between the scores on static versus interactive variants equaled only $\Delta rho = .034$ that was far not significant. The analogical difference for RAPM equaled $\Delta rho = .055$, which was not significant, either.

Discussion

The present study aimed to examine the role of interactivity in the process of insight problem solving. More specifically, it aimed to test (1) whether insight problems could be more frequently solved when presented in the interactive format allowing physical manipulation of the problem elements, as compared to the static format; (2) whether solutions in the former format could yield different subjective experience of suddenness, pleasure, relief, and certainty than yielded by solutions in the static format; and (3) how much performance on the interactive variants depended on cognitive resources, in comparison to the static variants.

A variety of established insight problems were used, which ranged in difficulty from a complete floor up to over one-third of correct solutions. Given existing evidence, the present results were quite surprising. There was virtually no difference in problem solving accuracy, regardless of the format used. Subjective experience reported, especially the suddenness of solution, linked closely to actual insight, was comparable for both problem formats. One exception was slightly increased certainty in interactive problem variants, which might have resulted from the fact that interactively delivered solutions were more concrete, so they could be more directly evaluated than the solutions written on paper. Importantly, both the static and the interactive problem variants substantially depended on working memory capacity and analytic reasoning and did it in a fully comparable way. Consequently, no evidence was found for any substantial

effects of interactivity on the process of insight problem solving on the sample of diverse and established insight problems. As both static and interactive conditions substantially relied on working memory/reasoning ability, no evidence was found for the decreased role of analytic thinking in the interactive format. Thus, it seems that physically manipulating problem elements did not substantially decrease cognitive load or affect the use of strategies in the process of problem solving.

On the other hand, it may be arguable to what extent the paper-and-pencil format, at least in case of some particular problems, is fully static, i.e. it does not provide any external support that may help in the process of solving. For example, making drawings and sketches may help to test hypotheses, keep track of the progress and perform simple trial-and-error strategies compared to problem solving without any external support provided. Thus, comparing interactive, static paper-and-pencil and the "pure" static format at the same time should be considered in future studies.

We cannot fully exclude the possibility that the null effects observed in the present study resulted from the selection of the particular problems or the specific way the interactive format was implemented in the given problems. If interactivity substantially affects only specific insight problems under some specific conditions it would be valuable to identify the characteristics of such problems that moderate this effect. Also, one limitation of the present study is that although no substantial effects of interactivity were observed, the relatively low statistical power prevented the detection of any potential small effects. However, if the effects of interactivity are negligible or observed only in very specific problems or circumstances, it would be questionable whether the role of interactivity in the process of insight problem solving is an important research topic at all.

Summing up, no evidence was found that manipulating physically problem elements when solving problems presumably involving insight helps to reach the correct solution. Neither the influence on the course of the problem solving process (the extent of its suddenness) nor on its affective consequences (pleasure, relief) were observed. Interactivity did not decrease a substantial reliance of the problem solving process on working memory capacity and reasoning ability, either. The only observed effect was a small increase in certainty about the solution (a meta-cognitive consequence). Thus, at least for the problems applied in the present study, externalized/embodied/situated factors played no substantial role in finding solutions, and the results are in line with the key role of analytic reasoning in solving insight problems. Still, more research is needed to comprehensively examine the potential role of interactivity in the process of insight problem solving.

References

Chuderski, A. & Jastrzębski, J. (2018a). The relationship of insight problem solving to analytical thinking: Evidence from psychometric studies. In F. Vallee-Tourangeau (Ed.), *Insight: On the origin of ideas* (pp. 120-142). Abingdon, UK: Routledge.

Chuderski, A. & Jastrzębski, J. (2018b). Much ado about Aha! Insight problem solving is strongly related to working memory capacity and reasoning ability. *Journal of Experimental Psychology: General*, 147, 257-281.

Chuderski, A., & Jastrzębski, J. (2018c). No role of initial problem representation in insight problem solving. *Creativity Research Journal, 30:4, 428-438.*

Clark, A., & Chalmers, D. J. (1998). The extended mind. *Analysis, 58,* 7–19.

Cowley, S. J., & Vallée-Tourangeau, F. (2010). Thinking in action. *AI & Society, 25,* 469–475.

Danek, A. H., Wiley, J., & Öllinger, M. (2016). Solving classical insight problems without aha! experience: 9 dot, 8 coin, and matchstick arithmetic problems. *The Journal of Problem Solving, 9,* 4.

Fioratou, E., Cowley, S. J. (2009). Insightful thinking: Cognitive dynamics and material artifacts. *Pragmatics & Cognition, 17,* 549–572.

Fleck, J. I., & Weisberg, R. W. (2013). Insight versus analysis: Evidence for diverse methods in problem solving. *Journal of Cognitive Psychology, 25,* 436–463.

Gilhooly, K. & Webb, M. E. (2018). Working memory in insight problem solving. In F. Vallee-Tourangeau (Ed.), *Insight: On the origin of ideas* (pp. 105-119). Abingdon, UK: Routledge.

Henok, N., Vallée-Tourangeau, F. & Vallée-Tourangeau, G. (2018). Incubation and interactivity in insight problem solving. *Psychological Research.*

Knöblich, G., Ohlsson, S., Haider, H., & Rhenius, D. (1999). Constraint relaxation and chunk decomposition in insight problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25,* 1534–1555.

Kounios, J., & Beeman, M. (2014). The cognitive neuroscience of insight. *Annual Review of Psychology, 65,* 71–93.

MacGregor, J. N., Ormerod, T. C., & Chronicle, E. P. (2001). Information processing and insight: A process model of performance on the nine-dot and related problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 176–201.

Ohlsson, S. (2011). *Deep learning: How the mind overrides experience.* Cambridge, UK: Cambridge University Press.

Ormerod, T. C., MacGregor, J. N., & Chronicle, E. P. (2002). Dynamics and constraints in insight problem solving.

Journal of Experimental Psychology: Learning, Memory, and Cognition, 28, 791–799.

Raven, J. C., Court, J. H., & Raven, J. (1983). *Manual for Raven's Progressive Matrices and Vocabulary scales* (Section 4: Advanced progressive matrices). London, UK: H. K. Lewis.

Vallée-Tourangeau, F.(2017). Interactivity and ego depletion in insight problem solving. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. J. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 1248-1253). Austin, TX: Cognitive Science Society.

Vallée-Tourangeau, F., Euden, G., & Hearn, V. (2011). Einstellung defused: Interactivity and mental set. *Quarterly Journal of Experimental Psychology, 64,* 1889–1895.

Vallée-Tourangeau, F., Steffensen S. V., Vallée-Tourangeau, G., Sirota, M. (2016). Insight with hands and things. *Acta Psychologica, 170,* 195-205.

Vallée-Tourangeau, F., & Wrightman, M. (2010). Interactive skills and individual differences in a word production task. *AI & Society, 25,* 433– 439.

Weisberg, R. W. (2015). Toward an integrated theory of insight in problem solving. *Thinking & Reasoning, 21,* 5-39.

Weller, A., Villejoubert, G., & Vallée-Tourangeau, F. (2011). Interactive insight problem solving. *Thinking & Reasoning, 17,* 424-439

Wiley, J., & Jarosz, A. F. (2012). How working memory capacity affects problem solving. *Psychology of Learning and Motivation, 56,* 185–227.

When Is Science Considered Interesting and Important?

Samuel G. B. Johnson[1], Amanda Royka[2,4], Peter McNally[3] & Frank C. Keil[4]
(sgbjohnson@gmail.com, amanda.royka@yale.edu, peterkmcnally1@gmail.com, frank.keil@yale.edu)
[1] School of Management, University of Bath, Bath, BA2 7AY UK
[2] School of Chemical and Biological Sciences, Queen Mary University of London, London E1 4NS UK
[3] Social Norms Group, University of Pennsylvania, Philadelphia, PA 19104 USA
[4] Department of Psychology, Yale University, New Haven, CT 06520 USA

Abstract

Scientists seek to discover truths that are interesting and important. We characterized these notions by asking laypeople to assess the importance, interestingness, surprisingness, practical value, scientific impact, and comprehensibility of research reported in the journals *Science* and *Psychological Science*. These judgments were interrelated in both samples, with interest predicted by practical value, surprisingness, and comprehensibility, and importance predicted mainly by practical value. However, these judgments poorly tracked the academic impact of the research, measured by citation counts three and seven years later. These results suggest that although people have internally reliable notions of what makes science interesting and important, these notions do not track scientific findings' actual impact.

Keywords: Folk science; science methodology; interest; philosophy of science; scientometrics

Introduction

The scientific enterprise aims to uncover eternal truths, and psychological science seeks to understand the most fundamental aspects of the human condition. From our modern vantage point, we can see clearly which scientific theories and results have stood the test of time, as truly foundational scientific achievements—Euclid's explication of geometry, Newton's laws of motion, Smith's insights about economic activity, Darwin's theory of evolution are among the timeless truths that clarify the structure of the natural and social worlds. But as scientists in the trenches, it is much more difficult for us to know what research is truly significant. Mendel's insights into genetics were ignored in his day, revolutionizing biology only decades later. The full significance of Bayes' contributions to statistics long eluded the profession.

Given these difficulties, scientists are likely to develop heuristics to evaluate scientific importance (Kahneman & Frederick, 2002). One particularly plausible heuristic is the *counterintuitiveness* or *surprisingness* of the research finding. For example, research on cultural narratives finds that minimally counterintuitive myths (relying mainly on intuitive concepts, peppered with a few counterintuitive ones) are most likely to be remembered and passed on (Norenzayan et al., 2006). Indeed, at a very general level, learning is likeliest to occur when the difference between expectations and reality (i.e., prediction error) is largest (Rescorla & Wagner, 1972).

Surprise is often a good criterion for scientific importance. We might consider a scientific result to be important when it falsifies an element of a theory (Popper, 1959/1934) or requires us to reconceptualize a topic of inquiry altogether (Kuhn, 1962). In Bayesian terms, a result is highly diagnostic when it is highly improbable on most theories but highly probable on another (i.e., $P(E|H)$ is high but $P(E)$ is low). As heuristics go, the extent to which a result encourages theory change is an excellent proxy for scientific importance.

But often, this heuristic can lead us astray. This is because even practicing scientists have scientific theories and intuitive theories that co-exist in their minds (Goldberg & Thompson-Schill, 2009; Shtulman & Varcarcel, 2012). Thus, although the disagreement between a result and existing *scientific* theory is a plausible proxy for scientific importance, disagreement with one's *lay* theory is not, if it is superseded by one's scientific understanding. For example, suppose that a psychologist believes that our behavior is guided by the unconscious activation of stereotypes, as suggested in the social priming literature. These original effects are highly counterintuitive, and if true, of great scientific significance. However, even though conceptual replications of these priming effects (e.g., using different stereotypes) would no longer contradict *scientific theory* (assuming we accept the initial demonstration), they would remain counterintuitive relative to our *folk theory*. Thus, this creates a misalignment between the scientific and lay surprisingness of a particular finding. To the extent that scientists rely on their folk theories rather than their scientific understanding for evaluating whether a finding is surprising, they may share this misalignment.

Regardless of the normative importance of counterintuitiveness, there is no question that many scientists prize it highly. Scientists, particularly during training, are often advised to seek out counterintuitive results. For example, one guide to doing "interesting" research advises (Gray & Wegner, 2013; pg. 550):

> One concrete test for evaluating ideas is to imagine the most surprising outcome possible (i.e., the best case scenario). If results were exactly as predicted, would they be interesting? If not, you should dream bigger when hypothesizing or perhaps consider the opposite of your hypothesis—if one way is intuitive, the opposite may be surprising.

Whose intuitions are we trying to contradict? "Grandmothers, not scientists," note the authors: "Ideally, research should counter both scientists' and laypeople's intuitions, but we emphasize the latter" (pg. 550). It is

hard to disagree with this as career advice, but it nonetheless raises uncomfortable concerns about replicability. After all, results with low prior probability are less likely to be true. Indeed, surprisingness is among the factors most associated with failure to replicate (Open Science Collaboration, 2012). It is presumably for this reason that the submission guidelines for *Psychological Science* now distinguish explicitly between "theoretical significance" (which is an acceptance criterion) and "surprising novelty" (which is not).

In this paper, we test two sets of issues, with Study 1 examining the folk science surrounding psychological research and Study 2 examining the natural sciences.

First, we ask what factors drive laypeople's judgments of how interesting and important scientific findings are. The opinions of laypeople, while likely divergent from experts, are important for two reasons. One reason is that scientists are laypeople in all fields aside from their own, and even in their own field may have lay intuitions that conflict with their scientific understanding of the field (Goldberg & Thompson-Schill, 2009). Thus, lay intuitions can creep into scientists' evaluations of research. A second reason is that the opinions of laypeople directly impact what scientific research is conducted, since laypeople are the ultimate consumers of taxpayer-funded research and since many scientists prioritize newsworthiness (to laypeople) in choosing topics to investigate. We study, therefore, the relative importance of surprisingness, perceived scientific impact, perceived practical value, and comprehensibility in guiding judgments of importance and interest.

Second, we ask how well these judgments track the objective academic impact of scientific findings, as quantified by their citations. Is the advice quoted above—to prioritize counterintuitiveness to laypeople—sound, if one's goal is to generate citations? Gray and Wegner (2013) suggest that it may be counterintuitiveness to scientists that drives citations in the short term, but to laypeople that drives citations in the longer term. We begin to examine this issue by looking separately at citation counts 4- and 7-years post-publication, testing whether lay judgments predict such measures of impact.

Study 1

In our first study, we looked at the factors influencing judgments of interest and importance, as well as citation counts, for articles published in *Psychological Science*.

Method

Participants. We recruited 60 participants from Amazon Mechanical Turk. Across our two studies, 57% of participants were female, 42% had completed at least a 4-year college degree, and the average age was 35. Only 8% of participants had doctoral-level training in any field, so the vast majority of participants were laypeople in the specific fields featured in our studies.

Participants were excluded if they incorrectly answered more than 30% of a set of 20 check questions ($N = 8$).

Materials. The materials were derived from abstracts of 40 articles appearing in the journal *Psychological Science* in the January, February, and March 2012 issues. A power analysis, treating item as the unit of analysis (like our main analysis below), revealed that 40 items is sufficient to detect correlations of $r > .41$ with 80% power.

For each abstract, a short summary was developed by the second author. For example, the actual abstract of one article (Frankenstein et al., 2012) read:

> We examined how a highly familiar environmental space—one's city of residence—is represented in memory. Twenty-six participants faced a photo-realistic virtual model of their hometown and completed a task in which they pointed to familiar target locations from various orientations. Each participant's performance was most accurate when he or she was facing north, and errors increased as participants' deviation from a north-facing orientation increased. Pointing errors and latencies were not related to the distance between participants' initial locations and the target locations. Our results are inconsistent with accounts of orientation-free memory and with theories assuming that the storage of spatial knowledge depends on local reference frames. Although participants recognized familiar local views in their initial locations, their strategy for pointing relied on a single, north-oriented reference frame that was likely acquired from maps rather than experience from daily exploration. Even though participants had spent significantly more time navigating the city than looking at maps, their pointing behavior seemed to rely on a north-oriented mental map.

We anticipated that real scientific abstracts like this one would be too long, syntactically complex, and jargon-filled to be comprehensible by most laypeople. Therefore, our summary version read:

> When presented with a virtual model of their hometown, people are able to more accurately point to familiar target locations when the people were oriented north and become progressively less accurate as they were oriented away from north. This suggests that people rely on a mental map that is oriented northward when trying to locate familiar places.

Comparable summaries were constructed for all 40 abstracts. Summaries were written at a minimum Flesch–Kincaid grade level of 12 and were of similar length.

We conducted pretests to ensure as strong of a perceived correspondence between the real abstract and summary as possible. In an initial pretest, each participant was assigned to read 10 of the abstracts along with their summaries, and rated their correspondence on a 0 ("A very poor match") to 10 ("An excellent match") scale. Any abstract with a score below 7 was targeted for revision and re-normed in a second pretest. All correspondences were rated above the scale midpoint in the second pretest (except one item which was omitted due to a coding error).

As an objective measure of academic impact, we obtained the Google Scholar citation counts for each article approximately 4 years (on 26 March 2016) and 7

years post-publication (on 29 January 2019) (on the pros and cons of Google Scholar versus other bibliometric databases, see Harzing & Alakangas, 2016). These were square root transformed, to account for the skewness of citation data.

Procedure. Participants each viewed 10 of the 40 summaries (balanced across participants). For each finding, participants first read the summary and then, on subsequent pages, made six ratings:

Interest. How interesting are these findings to you?

Importance. How important do you think these findings are?

Surprise. How surprising do you think these findings are?

Scientific impact. How much do you think these findings will change the way scientists think about this topic?

Practical value. How useful do you think this finding is on a practical level?

Comprehensibility. How well do you think that you understand the description of this finding?

These ratings were all made on scales from 0 ("Not at all…") to 10 ("Very…"). Each rating was made on a separate page, with the summary repeated at the top of each page. The order of the interest and importance questions was counterbalanced across participants, and the other ratings were always made in the order above.

After the main task, participants checked off, from a list of 20 concepts, those that had appeared in the summaries. Participants incorrectly answering more than 30% of these check questions were excluded to decrease noise due to inattentiveness.

Results

Overall, participants' judgments were internally reliable, with significant correlations among many of our measures. However, these scores had little external predictive power: Citations 4 and 7 years later were not predicted by any judgment except comprehensibility.

First-order correlations. We averaged, for each item, across participants' ratings, and used these item-level means for our analyses. The first-order Pearson correlations among all measures are summarized in Table 1. Before probing these associations more carefully using regression models, we make two observations.

First, judgments of importance and interest were highly correlated, $r(38) = .59$, $p < .001$. Since these results are observational, this is consistent with several causal orders. It could be that importance is the more fundamental judgment, and these appraisals feed into interest. This would be consistent with the fact that usefulness judgments were even more strongly associated with importance, $r = .79$, than with interest, $r = .61$. Alternatively, interest could be the more fundamental judgment, with importance less natural to judge and confabulated in line with personal interest. Finally, these two assessments could be relatively independent, depending on a mix of the same factors (such as usefulness) and differentiating factors (such as comprehensibility, which is only associated with interest).

	In	**Im**	**Su**	**SI**	**PV**
Im	.60***	—			
Su	.59***	.49**	—		
SI	.55***	.59***	.76***	—	
PV	.61***	.79***	.36*	.56***	—
Co	.66***	.11	.16	–.05	.17

º < .10 * < .05 ** < .01 *** < .001

Note. Entries are first-order correlations among interest (In), importance (Im), surprise (Su), scientific impact (SI), practical value (PV), and comprehensibility (Co).

Table 1: First-order correlations (Study 1).

Second, in preparation for modeling interest and importance, we note that some of the other variables are strongly correlated, which can lead to a multicollinearity problem. Variance Inflation Factors were acceptable (VIF < 1.5 for the Step 1 models in Tables 2 and 6) for models that did not simultaneously include both surprise and scientific impact, which were correlated very highly, $r = .76$. This very high correlation suggests, perhaps not itself surprisingly, that laypeople tend to substitute the difficult question of what evidence tends to change scientists' theories with the easier question of what they personally find surprising (Kahneman & Frederick, 2002). To address this problem, we omitted the scientific impact variable from the models. We included surprise rather than scientific impact since this seems to be the more natural assessment, but the results are similar if we instead include scientific impact or the average of the two.

Predictors of interest and importance. Table 2 shows the regression coefficients predicting judgments of interest. The Step 1 model uses surprise, practical value, and comprehensibility to model interest, and the Step 2 model adds importance to capture any added value.

As shown in the regression table, the strongest predictor of interest was comprehensibility, followed by practical value, followed by surprise, but all three predictors were highly significant, making independent contributions to interest. Together, these factors accounted for 80% of the variance in interest across items. Adding importance did not add any predictive power.

DV: Interest		
	Step 1	**Step 2**
Su	.26 (.06)***	.23 (.06)***
PV	.33 (.07)***	.23 (.10)*
Co	.42 (.06)***	.43 (.06)***
Im		.15 (.12)
R^2	.802	.811

DV: Importance		
	Step 1	**Step 2**
Su	.18 (.08)*	.11 (.10)
PV	.67 (.10)***	.57 (.12)***
Co	–.04 (.08)	–.17 (.13)
In		.30 (.23)
R^2	.676	.690

Note. Entries are unstandardized bs *and* SE*s*

Table 2: Regression models (Study 1)

The bottom panel of Table 2 shows the results of parallel regressions predicting importance. Comparably to the results of Table 1, adding interest has little predictive power beyond the other predictors. In this case, however, it is practical value that is doing nearly all of the predictive work: A 1 point increase on practical value is associated with a 0.67 point increase in importance. Surprise was weakly predictive in the Step 1, but not the Step 2, model. Overall, these variables predicted about 68% of the variance in perceived importance across items.

	Year 4	**Year 7**
In	.17	.19
Im	–.02	–.08
Su	–.15	–.12
SI	–.21	–.25
PV	–.03	–.06
Co	.35*	.38*

Note. Entries are first-order correlations with citations (square-root transformed) approximately 4 years and 7 years post-publication.

Table 3: Correlations with citations (Study 1)

Predictors of citation count. Table 3 presents the first-order correlations between citation count 4 and 7 years post-publication (square-root transformed) and the six measures collected in Study 1. Various regression specifications produce similar conclusions, so we focus here on the simple correlations as they avoid the multicollinearity issues mentioned above.

At both time points, neither interest nor perceived importance significantly predict citation counts, nor did judgments of surprise, scientific impact, or practical value. The only significant predictor was comprehensibility, $r(38) = .35$, $p = .028$ and $r(38) = .38$, $p = .017$ at 4 and 7 years, respectively.

Discussion

Several results pop out in these data. First, judgments of interest and importance are fairly independent: They depend on different factors and do not predict one another once one adjusts for those other factors. Comprehensibility was the most important guide to interest, but had no impact on perceived importance (see Oppenheimer, 2006 for related findings). Practical value was the most important determinant of perceived importance, and also had a large effect on interest. Surprisingness was correlated with interest but not perceived importance. Second, these judgments had little predictive power for citation rates, either in the shorter- or longer-term. Comprehensibility had a moderately high correlation with citations, but no other factor did.

Study 2

The Study 1 results could very well be specific to psychology. For instance, people have much more detailed intuitive theories of psychology, since they can introspect about their own psychology, and therefore surprisingness could be seen as an especially strong cue to scientific impact. Study 2 repeated this procedure on natural science findings from *Science* magazine.

Method

We recruited 60 participants from Mechanical Turk. Participants were excluded using the same criterion as Study 1 ($N = 14$).

The materials were the "editor's summaries" of 40 articles published in the January 6, January 13, and January 20, 2012 issues of *Science* magazine. These summaries are written by the editorial staff of the journal, rather than by us, eliminating the possibility of experimenter bias. We lightly edited the summaries to match the format of our Study 1 materials (replacing the authors' names with "Researchers"). For example, the editor's summary of one article (Fermi LAT Collaboration, 2012) read:

> Binary star systems that contain a neutron star or a black hole are expected to emit gamma rays. These gamma-ray binaries are a rare class of objects, which are also expected to emit x-rays. Indeed, several such systems were initially detected through their x-ray emission. Researchers have reported the detection of a gamma-ray binary that was previously unknown as an x-ray source. Follow-up observations reveal that the system is also a source of x-rays and that the companion star is a class O star, a type that is very hot and very luminous.

Participants read 10 of the 40 descriptions

(counterbalanced across participants), making the same six judgments as in Study 1, using a similar procedure.

Citation counts were obtained using the same procedure as Study 1, on 17 August 2016 and 30 January 2019.

Results

Most of the key results from Study 1 were replicated. Surprisingness, practical value, and comprehensibility all predicted judgments of interest, while only practical value robustly predicted judgments of importance. Citations were marginally predicted by comprehensibility, as in Study 1, but also by judgments of practical value.

Differences in means across studies. Table 4 presents the descriptive statistics for each judgment across each set of summaries. We compared the means on each measure across studies, using the false discovery rate procedure to adjust *p*-values for multiple comparisons (Benjamini & Hochberg, 1995). Overall, the natural science findings in Study 2 were viewed as less interesting than the psychology findings, $t(78) = 3.51$, $p = .002$, $d = 0.78$, 95% CI[0.41,1.48], but as more important than the psychology findings, $t(78) = 2.82$, $p = .009$, $d = 0.63$, 95% CI[0.18,1.06]. The natural science findings were also viewed as more surprising, $t(78) = 2.68$, $p = .011$, $d = 0.60$, 95% CI[0.15,1.04], and more scientifically impactful, $t(78) = 3.84$, $p < .001$, $d = 0.86$, 95% CI[0.37,1.15], but of similar practical value, $t(78) = 0.67$, $p = .51$, $d = 0.15$, 95% CI[–0.33,0.66]. Finally, the psychology findings were much easier to understand, $t(78) = 9.39$, $p < .001$, $d = 2.10$, 95% CI[2.52,3.87].

	Study 1 (Psychology)	Study 2 (Natural Science)
In	5.88 (0.84)	4.93 (1.48)
Im	5.74 (0.93)	6.36 (1.02)
Su	4.26 (1.19)	4.86 (0.75)
SI	5.35 (0.86)	6.11 (0.91)
PV	5.50 (1.00)	5.67 (1.20)
Co	7.80 (1.08)	4.60 (1.86)

Note. Entries are means (SDs) across items.

Table 4: Descriptive Statistics across Studies

First-order correlations. Table 5 shows the first-order correlations for Study 2, analogous to Table 1.

Like Study 1, there was a significant correlation between interest and importance, $r(38) = .39$, $p = .013$, although of more modest magnitude. The correlation between surprise and perceived scientific impact was also more modest. This weaker correlation, relative to Study 1, may have resulted from participants' lesser ability to rely on intuitive theories of the natural sciences than of psychology, given introspective access to one's own mental states.

	In	Im	Su	SI	PV
Im	.39*	—			
Su	.35*	.18	—		
SI	.40*	.87***	.27º	—	
PV	.45**	.76***	–.02	.72***	—
Co	.89***	.22	.18	.18	.35*

Table 5: First-order correlations (Study 2).

Predictors of interest and importance. Table 6 shows the regression coefficients predicting interest and importance judgments, analogously to Table 2.

The results are similar to Study 1. For interest judgments, we find that surprise, practical value, and comprehensibility are all significant predictors, with comprehensibility the strongest predictor, just like Study 1. (However, surprise was a stronger predictor than practical value in Study 1, whereas the converse was true in Study 2.) Like Study 1, importance does not have any added predictive power; in this case, its collinearity with practical value leads both to be non-significant when entered simultaneously.

	DV: Interest	
	Step 1	**Step 2**
Su	.41 (.13)**	.36 (.14)*
PV	.21 (.08)*	.09 (.13)
Co	.63 (.05)***	.64 (.06)***
Im		.18 (.15)
R^2	.854	.859

	DV: Importance	
	Step 1	**Step 2**
Su	.29 (.14)*	.21 (.16)
PV	.69 (.09)***	.64 (.10)***
Co	–.06 (.06)	–.19 (.13)
In		.21 (.18)
R^2	.631	.645

Note. Entries are unstandardized bs *and* SEs

Table 6: Regression models (Study 2)

For importance judgments, we find, just as in Study 1, that the key predictor is practical value, with a possible additional role for surprise. Given the high correlation

between scientific impact and practical value in Study 2, however (see Table 5), replacing surprise with scientific impact in the regression leads to a reversal of the coefficient magnitudes: Scientific impact is then a more robust predictor of importance than practical value, although both are significant in either model. (This is not true for Study 1, where scientific impact and surprise are basically interchangeable in the models.)

	Year 4	Year 7
In	.16	.18
Im	.16	.18
Su	–.23	–.25
SI	.03	.04
PV	.43**	.45**
Co	.27º	.28º

Note. Entries are first-order correlations with citations (square-root transformed) approximately 4 years and 7 years post-publication.

Table 7: Correlations with citations (Study 2)

Predictors of citation count. Table 7 presents the correlations of our six judgment variables with citation counts approximately 4 and 7 years post-publication. Comprehensibility was a marginally significant predictor at both time points, $r(38) = .27$, $p = .098$ and $r(38) = .28$, $p = .076$, which is consistent with the predictive power of comprehensibility for citations in Study 1. Unlike Study 1, however, practical value also predicted citation counts at both time points, $r(38) = .43$, $p = .006$ and $r(38) = .45$, $p = .004$. Thus, laypeople do appear to be able to extract some information that is predictive of the academic impact of scientific findings, but it is not necessarily reflected in their own judgments of importance.

Discussion

Study 2 replicated the main results of Study 1: Comprehensibility was a powerful cue to interest but not importance, although only a marginal predictor of citations in Study 2. Surprisingness was only a robust predictor of interest, but not importance, while practical value strongly predicted both. Unlike Study 1, practical value was a fairly strong predictor of citations, even though perceived importance was not.

General Discussion

Lay intuitions about scientific importance are, well, important. They impact our choices of research topics indirectly, as we try to appeal to laypeople's interests, and directly, as we all have a layperson inside of us (Goldberg & Thompson-Schill, 2009). What scientific findings do laypeople consider interesting and important? How much do these judgments track objective scientific importance?

Overall, comprehensibility is the most important predictor of interest. It is unclear whether this is because writing quality itself provokes interest, or because interesting findings are easier to explain clearly—quite possibly both. Scientists who wish to appeal to public interest ought to keep this demand for clarity in mind, rather than obscuring their work in jargon (see Oppenheimer, 2006).

Perceived practical value was the most robust predictor of importance, although perceived scientific impact was also highly predictive in Study 2 (and difficult to distinguish from practical value). Surprisingness appears to be less predictive. This is, ironically, quite a counterintuitive result! Guides to doing "interesting research" (Gray & Wegner, 2013) and our professional intuitions point to the importance of surprising the reader. But laypeople may well be growing weary of surprising findings, as they encounter increasing levels of "click bait" reporting and all-too-frequent reversals of conventional wisdom (are we, or are we not, supposed to eat eggs now?). Future research might investigate the factors underpinning and moderating this relationship between surprise (e.g., Maguire, Maguire, & Keane, 2011) and judgments of interest and importance.

Finally, these results suggest caution regarding our ability to predict the impact of scientific research based on its relationship with our intuitive theories. Surprise had no impact on citations, but neither did interest or judgments of importance or scientific impact. The only factors impacting citation were comprehensibility (in both studies) and perceived practical value (in Study 2).

It is important to understand how laypeople think about science because scientific progress tracks social priorities—scientists serve at the pleasure of society. To the extent that laypeople have systematic misconceptions about science, we must understand how those misconceptions might thwart the dissemination of science to the public, or even scientific progress itself. To the extent that laypeople have irreducible preferences over the kind of science they like to see, we must understand how those preferences might be reflected in the kind of research produced by scientific institutions.

Several other research programs contribute to this broad goal. For example, people favor reductionist explanations (e.g., referring to smaller parts or component processes) even when the reductionist information makes no logical contribution to the explanation (e.g., Hopkins, Weisberg, & Taylor, 2016; Weisberg, Keil, Goodstein, Rawson, & Gray, 2008). As a second example, people have consistent intuitions about the limits of science, particularly of psychology (Gottlieb & Lombrozo, 2018), believing that phenomena are scientifically explainable to the extent that scientists can make falsifiable and reductionist claims about those phenomena (Johnson, Kim, & Keil, 2016).

These research areas—characterizing what scientific explanations people find compelling and what scientific questions people find tractable—are valuable because they contribute to our understanding of how the lay public

interfaces with the scientific community. If people have an unjustified preference for neuroscientific explanations or an ill-founded belief that psychological phenomena are beyond the limits of science to comprehend, these may lead to society-wide distortions in our scientific priorities.

Our work complements these approaches. While these other lines of research hint at what the public's scientific priorities might be by characterizing folk scientific beliefs, the current studies take a more direct approach by asking what research people find interesting and important. If we believe that laypeople's standards (e.g., regarding practically valuable findings as more important) are reasonable, then this is all to the better. To the extent that we find lay preferences more questionable (e.g., favoring counterintuitive findings as more interesting), this should catalyze a discussion about how society prioritizes research questions, how journals select research findings, and how scientists choose research topics.

Scientists get into the business because they want to have an impact—maybe even to change the world. We may be less able than we believe to predict successfully what scientific innovations are indeed important. Doing so successfully may require us to step back and reconsider our habits of thought.

References

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*, *57*, 289–300.

Fermi LAT Collaboration (2012). Periodic Emission from the Gamma-Ray Binary 1FGL J1018.6–5856. *Science*, *335*, 189–193.

Frankenstein, J., Mohler, B.J., Bülthoff, H.H., & Meilinger, T. (2012). Is the map in our head oriented north? *Psychological Science*, 23, 120–125.

Goldberg, R.F., & Thompson-Schill, S.L. (2009). Developmental "roots" in mature biological knowledge. *Psychological Science*, *20*, 480–487.

Gottlieb, S., & Lombrozo, T. (2018). Can science explain the human mind? Intuitive judgments about the limits of science. *Psychological Science*, *29*, 121–130.

Gray, K., & Wegner, D. M. (2013). Six guidelines for interesting research. *Perspectives on Psychological Science*, *8*, 549–553.

Harzing, A., & Alakangas, S. (2016). Google Scholar, Scopus, and the Web of Science: A longitudinal and cross-disciplinary comparison. *Scientometrics*, *106*, 787–804.

Hopkins, E.J., Weisberg, D.S., & Taylor, J.C.V. (2016). The seductive allure is a reductive allure: People prefer scientific explanations that contain logically irrelevant reductive information. *Cognition*, *155*, 67–76.

Johnson, S.G.B., Kim, K., & Keil, F.C. (2016). The determinants of knowability. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In *Heuristics and biases: The psychology of intuitive thought* (pp. 49–81). New York, NY: Cambridge University Press.

Kuhn, T.S. (1962). *The structure of scientific revolutions*. Chicago, IL: University of Chicago Press.

Maguire, R., Maguire, P., & Keane, M.T. (2011). Making sense of surprise: An investigation of the factors influencing surprise judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*, 176-186.

Norenzayan, A., Atran, S., Faulkner, J., & Schaller, M. (2006). Memory and mystery: The cultural selection of minimally counterintuitive narratives. *Cognitive Science*, *30*, 531–553.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, *349*, aac4716.

Oppenheimer, D.M. (2006). Consequences of erudite vernacular utilized irrespective of necessity: Problems with using long words needlessly. *Applied Cognitive Psychology*, *20*, 139–156.

Popper, K. (1959). *The logic of scientific discovery*. London, UK: Routledge. (Original work published 1934.)

Rescorla, R.A., & Wagner, A.R. (1972). A theory of Pavlovian conditioning: Variations on the effectiveness of reinforcement and non-reinforcement. In *Classical conditioning II: Current research and theory* (pp. 64–99). New York, NY: Appleton-Century-Crofts.

Shtulman, A., & Valcarcel, J. (2012). Scientific knowledge suppresses but does not supplant earlier intuitions. *Cognition*, *124*, 209–215.

Weisberg, D.S., Keil, F.C., Goodstein, J., Rawson, E., & Gray, J.R. (2008). The seductive allure of neuroscience explanations. *Journal of Cognitive Neuroscience*, *20*, 470–477.

IMPACT OF CHESS TRAINING ON CREATIVITY AND INTELLIGENCE

Ebenezer Joseph
EMMANUEL CHESS CENTRE, CHENNAI, India

Veena Easvaradoss
WCC, CHENNAI, India

David Chandran
EMMANUEL CHESS CENTRE, CHENNAI, India

Suneera Abraham
EMMANUEL CHESS CENTRE, CHENNAI, India

Abstract

Research using short-term chess training programs has indicated an enhancement of cognitive functioning among children. The aim of the study was to investigate the effect of 1-year systematic chess training on the creativity and intelligence of children. A pretestposttest with control group design was used. Children who were studying in two government schools and two private schools (grades 39) were selected randomly. They were then randomly assigned to experimental and control groups, with 88 (50 boys, 38 girls) children in the experimental group and 90 (57 boys, 33 girls) children in the control group. The experimental group underwent weekly 1-hour chess training for 1 year, while the control group was actively involved in extracurricular activities offered by the school during the same period. Creativity was measured by WallachKogan Creativity Test (Indian adaptation) and intelligence was measured by subtests of Wechsler Intelligence Scale for Children: Fourth edition (WISC-IV), India. Analysis of covariance (ANCOVA) revealed significant improvement in total creativity and Full Scale Intelligence Quotient (FSIQ) for experimental group compared to the control group. Chess training as part of school activities appears to have a wide spectrum of outcomes.

Exploring informal science interventions to promote children's understanding of natural categories

George Kachergis[1], **Todd M. Gureckis**[2], **Marjorie Rhodes**[2]
[1]Department of Psychology, Stanford University, Stanford, CA
[2]Department of Psychology, New York University, New York, NY

Abstract

Categories carve up the world in a structured way, allowing people to inductively reason about the properties of novel exemplars. Children are still in the process of learning category structure, and often fail to leverage the inductive power of these representations to their advantage. For example, young children generally fail to recognize the value of sampling diverse exemplars to support category-wide generalization. This study investigates whether teaching children the structure within a natural category increases diversity-based inductive reasoning. In an informal science learning environment, we presented 259 children aged 5 to 8 years with exemplars of the three main types of birds: raptors, songbirds, and waterbirds. After a short dialogue pointing out the various within-type similarities and between-type differences, children's diversity-based inductive reasoning did not significantly improve, despite them evidencing a better understanding of the category's structure. Instead, children tended to avoid sampling waterbirds, the least typical cluster of birds. These patterns suggest that children's neglect of sample diversity is unlikely to be solely due to their relative ignorance of category structure.

Keywords: category induction; diversity-based reasoning; category learning; conceptual development

Introduction

Categories give us a way out of the infant's problem of "feel[ing] it all as one great blooming, buzzing confusion" (James, 1890), allowing us to carve our sensations of the world into classes that are distinguished by relevant properties. By picking out shared features while ignoring superficial differences, categories enable us to learn inductively (Rips, 1975), and allow us to reason about unseen properties. For example, having learned that one cat's fur is soft, we might generalize this to all cats, and proceed to seek out and pet all cats. Assuming that categories are homogeneous, in that the exemplars share many observable and unobservable features, offers further inductive power (Gelman, 1988), but can also lead to serious errors: your house cat may be amenable to petting, but a cougar may not be. Learning how to account for such within-category variation while engaging in category-based inductive reasoning is a nontrivial problem for children.

There is considerable evidence that adults take within-category variability into account when evaluating the inductive power of evidence. For example, when making inferences about general properties of a category, adults view some samples of exemplars from a category as more informative than others (Rips, 1975; Osherson, Smith, Wilkie, Lopez, & Shafir, 1990). Sample diversity is one feature adults attend to when evaluating evidential strength, with more diverse samples (e.g., a lion and a house cat) being viewed as more informative than nondiverse samples (two house cats) (Heit, 2000). Adults' preference for diverse samples indicates that they assume that observable variability across exemplars in a category is often correlated with variance in the hidden features, and it is thus informative to sample from diverse areas of the distribution.

Adults find diverse samples more informative both when choosing evidence to sample (Kim & Keil, 2003; Lopez, 1995; Rhodes, Brickman, & Gelman, 2008; Rhodes & Gelman, 2008) and when rating the strength of inductive arguments (Osherson et al., 1990). In contrast, children below the age of nine often fail to consider sample diversity (Gutheil & Gelman, 1997; Rhodes & Gelman, 2008; Rhodes et al., 2008). For example, before age 9, children are equally likely to choose to examine an eagle and a robin as two robins to see whether all birds have a given property (Rhodes et al., 2008; Rhodes & Gelman, 2008). One reason why this might be the case is that young children often assume categories to be more homogeneous than adults do (Gelman, 2003). For instance, preschool-aged children infer more readily than adults that a property seen in one exemplar is true for the whole category (Rhodes & Gelman, 2008). Moreover, preschool-aged children believe that everyday categories identify an objective natural reality to a greater extent than adults do (Kalish, 1998). Directly linking this tendency to assume that categories are homogeneous to diversity-based reasoning, 7-year-olds reliably chose diverse samples over nondiverse samples after they were primed with an example highlighting within-category variability (Rhodes & Brickman, 2010). This finding supports the idea that young children may default to a strong within-category homogeneity assumption, and also shows that they sometimes recognize the value of diverse samples when such an assumption is violated. Here, we tested whether a more abstract variation prime–distinguishing structured clusters within a category–induces a preference for diversity-based reasoning.

The present study

Imagine that you are trying to determine whether a novel, hidden property is true of all birds (e.g., "Do all birds have *scutella*?"). Adults deem it better grounds for generalization to the entire category when two dissimilar birds (e.g., an eagle and a robin) are both found to have the hidden property,

rather than two more similar birds (e.g., a robin and a swallow). In the model proposed by Osherson et al. (1990), the greater strength of the argument based on dissimilar birds is a result not of the similarity of the premise categories to the conclusion category, but of their similarity to the lowest-level category that covers both the premise and conclusion categories—that is, birds. If on the other hand, one were asked to determine if all songbirds have *scutella*, the robin and swallow premises would offer a better match to the inductive target, since the songbird cluster is the lowest-level category covering both the premises and the target.

The goal of this study was to teach children more about the heterogeneous structure of one natural category–birds–to determine whether that knowledge improves their inductive reasoning about that category. Specifically, this study taught young children (ages 5-8) that the bird category is clustered into songbirds, waterbirds, and raptors, and that birds within each cluster share many visible (e.g., talon and beak shape) and hidden properties. The didactic dialogue and displays (see Figures 1 and 2) used to teach the clusters will also highlight some of the variability within each cluster, and some cross-cluster similarities. Our hypothesis was that teaching children the clustered structure of the bird category may encourage them to represent the category more heterogeneously (i.e., with clusters instead of a single prototype), and thus may improve their diversity-based inductive reasoning—both for induction to a cluster, and to the entire category of birds. This study was conducted in the context of the American Museum of Natural History's Discovery Room with an interest in developing more effective exhibits.

Experiment

The purpose of the experiment is to investigate whether teaching children the clustered structure of a natural category (birds) improves their ability to do diversity-based inductive reasoning. Specifically, after measuring children's knowledge of the bird category through pile sorting, we tested two interventions–display cases (e.g., Figure 2) vs. poster-based (e.g., Figure 1), presented along with a didactic dialogue meant to demonstrate that birds can be subcategorized into three clusters: raptors, songbirds, and waterbirds. The dialogue first highlights the variability of the category (e.g., "Some are big, some are small; some have bright colors..." and then defines the three main clusters, emphasizing correlated features and their causal relationships (e.g., "Most water birds have webbed feet–see? That helps them swim.") After highlighting a few distinctive features for each cluster and naming four examples of birds in each, children were given a series of inductive sampling questions to generalize to either a given cluster, or to the entire bird category. Finally, we again measured children's knowledge of birds by pile sorting.

Methods

Participants Participants in this experiment were 265 children between the ages of 5 and 8 years old who were recruited at the American Museum of Natural History's Discovery Room. Of the 265 children recruited (per intervention: 99 in None, 94 in Exhibit, and 66 in Poster), we analyzed the data from 259 children (60 5-year-olds, 73 6-year-olds, 68 7-year-olds, and 58 8-year-olds) who completed the study.

Stimuli This task was designed to fit thematically with the content of the AMNH Discovery Room activities, which emphasize the varying features between species of birds, among other animals. Eight diverse birds from each cluster were selected to be used as stimuli. Songbirds[1] included were the robin, swallow, starling, oriole, redwing blackbird, blue jay, sparrow, and cardinal. Raptors included were the barred owl, falcon, vulture, kestrel, barn owl, red-tailed hawk, bald eagle, and kite. Water birds included were the tern, mallard duck, puffin, sea gull, harlequin duck, flamingo, goose, and anhinga. For the intervention conditions, four birds from each cluster were selected to be shown either on a poster, shown in Figure 1, or in display cases as in Figure 2. The same exemplars were used in both intervention conditions. For the sorting task, each of the 24 birds was printed in color on a single 8.5 x 11" sheet of paper, with the size of each bird being according to their real-world scale. These birds appeared in different combinations during the sampling questions, as well. All materials and the protocol are available online.[2]

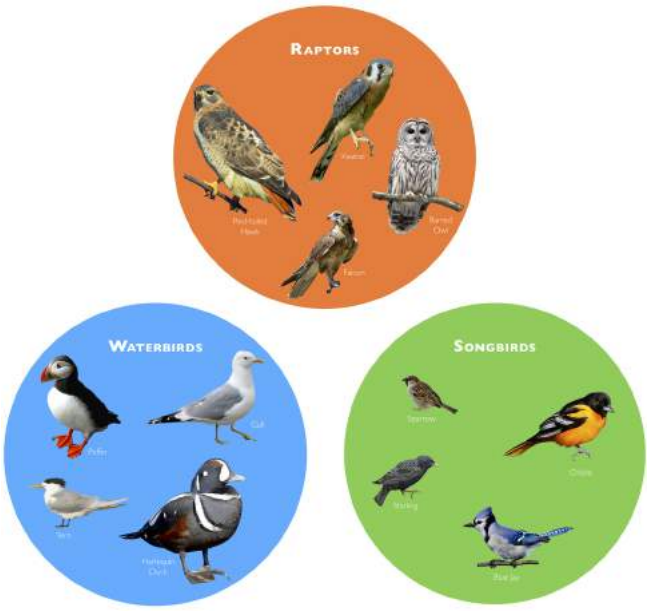


Figure 1: The experiment used four birds from each cluster as stimuli. The poster intervention used this arrangement.

Procedure Children participated in the experiment during individual sessions with trained undergraduate research assistants in a corner of the AMNH Discovery Room. Participants were randomly assigned to one of three intervention condi-

[1]Blue jays are not songbirds, so this cluster is more accurately described as passerines. Colloquially, all are songbirds.

[2]https://osf.io/gzfk9/

Figure 2: The display cases in the American Museum of Natural History's Discovery Room, with the 12 bird specimens used in the intervention.

tions: Poster, Exhibit, or None. Participants were first asked to list all of the birds they could name, as a means of establishing rapport. Second, they were asked to sort 24 bird cards (8 water birds, 8 raptors, and 8 songbirds) into piles "that go together by nature." They were given three baskets, but were not explicitly told to form three piles—nor instructed to if they asked. This sorting task offered a simple way of measuring their knowledge of the category structure, and was repeated at the end of the experiment to measure changes in their representation of the category. After each sort, participants were asked to describe each of the piles they made.

In addition to a control condition with no dialogue or visual intervention (None), there were two intervention conditions, using either a Poster (see Figure 1) or an Exhibit of display cases (see Figure 2) showing four birds from each cluster. In the intervention conditions, the experimenter showed the poster or display cases while giving a 2-minute dialogue that stressed the diversity of birds (e.g., color, size, and beak and talon shapes), while linking the distinguishing features of each cluster to their habitat and food sources (e.g., "water birds swim in the water with webbed feet and eat fish").

After the dialogue, children were given a series of 18 2-alternative forced choice induction sampling trials, an example of which is shown in Figure 3. Each trial offered two pairs of birds, and one bird was always shared across the two pairs (i.e., the harlequin duck in Figure 3). Each trial always offered a same-cluster pair (e.g., two waterbirds) and a between-cluster pair (e.g., a waterbird and a raptor, as in Figure 3). Children were told they were scientists trying to determine whether all birds (on category induction trials) or all birds of a given type (e.g., all raptors; cluster induction trials) had some property (e.g., 'podotheca'). They were asked to choose which pair of birds (left or right) they would like to test in order to make that determination. The first 9 sampling trials had questions targeting induction to the entire bird category, of the form: "You are a scientist who wants to find out if BIRDS have podotheca. Which set of birds do you want to look at to learn about BIRDS?" Three of these category-induction trials were easy, in that the same-cluster pair of birds on each trial would be two exemplars of the same species (e.g., photos of two puffin exemplars). The other six were difficult, in that the same-cluster pair showed birds of different species (e.g., a duck and a puffin, in Figure 3). The final 9 sampling trials had questions targeting each cluster (3 per cluster), of the form: "Here are two sets of birds. You are a scientist who wants to find out if RAPTORS have cancella. Which set of birds do you want to look at to learn about RAPTORS?" For each cluster, one of the cluster-induction trials was easy (i.e., the same-cluster pair showed two birds of the same species), while the other two per cluster were difficult, with the same-cluster pair comprised of different species.

The same 12 exemplar birds were shown in the poster as in the exhibit, and were also represented among the 24 cards for sorting, and in the induction sampling trials. Thus, a total of 24 bird species (8 per cluster) were introduced in the experiment, 12 of which were used in the interventions, and all of which appeared in both the pile sorting and sampling trials.

Figure 3: Example of a difficult induction sampling trial. Participants might be asked to choose which sample of birds (left pair or right pair) they would like to test to determine whether all birds (on category induction trials) or, e.g., all water birds (on cluster induction trials) have a property (e.g., 'scutella'). For category induction, the right sample should be chosen for testing, as it is more diverse, containing a water bird and a raptor. For cluster (water bird) induction, the left sample should be chosen, as it contains two water birds (harlequin duck and puffin). On easy trials, one sample would contain two identical birds (e.g., two puffins).

Pre-Intervention Sort	*Post-Intervention Sort*
size (111)	cluster (120)
habitat (35)	size (67)
mix (27)	feature (21)
cluster (25)	color (15)
feature (24)	mix (12)
color (20)	none (12)
none (10)	habitat (11)
252	258

Table 1: Categories of participants' pile sort explanations.

Results

Results were analyzed for 259 participants: 94 in the Exhibit condition (24 aged-5, 26 aged-6, 20 aged-7, and 24 aged-8), 66 in the Poster condition (16 aged-5, 15 aged-6, 22 aged-7, 13 aged-8), and 99 in the None condition (23 aged-5, 32 aged-6, 24 aged-7, 20 aged-8). Data from 8 other participants were eliminated due to failure to complete the experiment or experimenter error.

Bird Sorts Participants' sorts of the 24 birds were first examined according to the descriptions that they gave their piles. Experimenters sanitized and aggregated the pile sort descriptions, and attempted to assign a single label to the scheme by which each participant carried out their first (pre-intervention) and second (post-intervention) sorts. Participants' aggregated pile sort explanations are shown in Table 1. In some cases, participants gave no interpretable description (*none*), or gave a *mix* of features (e.g., color and size for one pile, habitat for another). It is clear that many participants, both pre- and post-intervention, sorted according to a single salient dimension such as size (111 first sort; 67 second), a physical feature (24 first; 21 second) or color (20 first; 15 second). However, cluster-based sorting is the only strategy that saw a marked increase from the first to the second sort (25 to 120 participants), with most single-dimension strategies seeing corresponding decreases (e.g., size: 111 to 67; habitat: 35 to 11). In terms of their qualitative descriptions, participants have largely shifted from single-dimension sorting strategies to a cluster-based strategy. Next, we quantified the degree to which participants' sorts improve in each intervention condition with respect to the ground truth.

To measure the quality of participants' sorts, we compared the piles from each participant's first and second sorts to the objectively correct cluster sort (three piles: 8 raptors, 8 songbirds, and 8 waterbirds). The similarity between a participant's sort and the correct cluster sort was measured with the adjusted Rand Index (Rand, 1971), which counts the number of pairs of elements in S (the cards) that are in the same subset in partition X (a participant's piles) and in the same subset in Y (the 3-pile objective), as well as the number of pairs of elements in S that are in different subsets in X and that are in different subsets in Y. These agreements between sort X and sort Y are divided by the number of all possible pairs (agreements + disagreements), and thus the Rand index is 0 when two sorts do not agree on any pair of cards, and 1 when sorts are exactly the same. The adjusted Rand Index corrects for chance using the expected similarity of all pair-wise comparisons between clusterings specified by a random model, and can thus on occasion have negative values if the index is less than expected by chance (Hubert & Arabie, 1985).

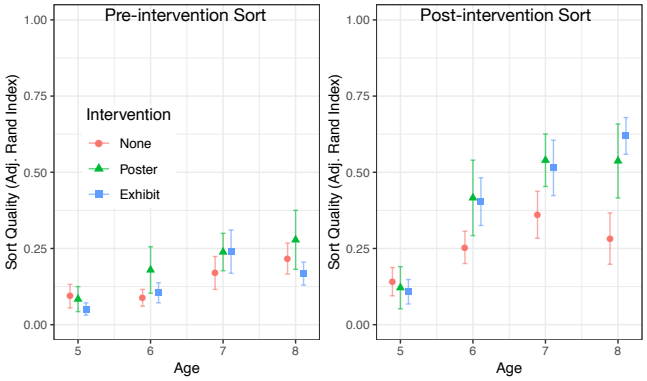


Figure 4: Mean pre-intervention sort quality (left panel) and post-intervention sort quality (right panel) by age and intervention condition. Error bars show +/-1SE.

The adjusted Rand Index of participants' sorts and the correct clustering (i.e., sort quality) were subjected to a repeated measures ANCOVA with sort (pre-/post-intervention) as the repeated measure, age (5.00-8.93 years) as covariate, and intervention condition (poster, exhibit, or none) as a between-subjects factor. Sort quality improved more in the intervention conditions (from pre- to post-intervention) than in the control condition (as evidenced by an interaction of intervention and sort ($F(2,253)=4.45$, $p = .01$); there were also subsumed main effects of sort (pre- to post-intervention $F(1,253)=97.42$, $p < .001$) and intervention condition ($F(2,253)=3.49$, $p = 0.03$). Participants in the two intervention conditions showed more improvement in sort quality ($M = .25$) than participants receiving no intervention (Welch's $t(245.7)=2.93$, $p = 0.004$). Sort quality also improved more for older children than younger children (interaction of age and sort ($F(1,253)=13.75$, $p < .001$), with a subsumed main effect of age ($F(1,253)=35.22$, $p < .001$). Figure 4 shows participants' mean sort quality on the pre- (first; left panel) and post-intervention (second; right panel) sorts by age and condition. Having shown that the interventions helped improve children's understanding of the clustered structure of the bird category, we turned to the induction sampling choices to determine if diversity-based reasoning also improved, using the quality of their second sort as a covariate.

Induction Sampling Choices We separately analyze the sampling trials that were targeted at inducing to a specific cluster, and the sampling trials that were targeted at inducing to the entire category. On the cluster induction trials, partici-

pants should choose the pair of birds from the targeted cluster, rather than the diverse pair. In contrast, on the category induction trials, participants should choose the diverse pair, with birds representing two different clusters. Thus, to analyze the cluster induction trials, participants' binomial choices for each trial (0=choosing the diverse pair, 1=choosing the cluster pair) were subjected to a logistic mixed-effects regression with intervention condition as a between-subjects factor and trial difficulty (easy or difficult) as a within-subjects factor, and age and quality of second sort as covariates[3].

On the cluster trials, there was no significant main effect of intervention ($F(2,243.5)=0.70$, $p = .50$), nor any significant interactive effects involving intervention, as shown in Figure 5 (left). Children were more likely to select samples containing only members of the cluster that they were trying to learn about (the more informative samples in this case) with age, ($F(1,244.2)=20.39$, $p < .001$), and they were also more likely to do so if they had more accurate representations of the category structure as indicated by their post-intervention category sort ($F(1,246.1)=27.15$, $p < .001$). There was a significant interaction of age and second sort quality ($F(1,241.5)=5.57$, $p = .02$). Shown in Figure 6, children with high-quality (4th quartile: ARI$>$ .65) second sorts are near-ceiling at correctly choosing the within-cluster sample for induction–even the 5- and 6-year-olds. The cluster pair was chosen significantly less often on difficult trials ($M = 0.69$) than on easy trials ($M = 0.77$, $F(1,255.2)=19.05$, $p < .001$). All other effects were not significant.

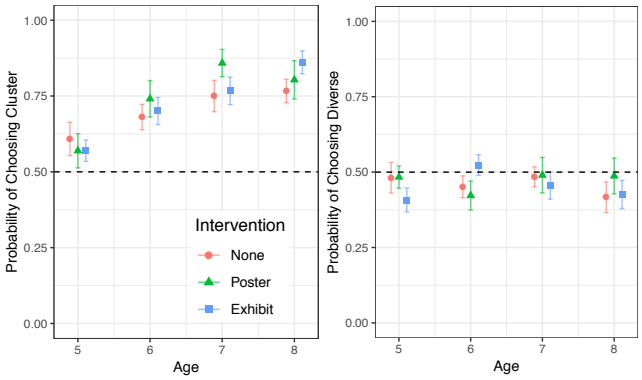


Figure 5: Mean proportion of correct sampling choice–nondiverse for cluster induction (left), and diverse for category induction (right)–by age group and intervention condition. Error bars show +/-1SE, and dotted lines show chance.

To analyze the category induction trials, participants' binomial choices for each trial (0=choosing the non-diverse pair, 1=choosing the diverse pair) were subjected to a logistic mixed-effects regression with intervention condition as a between-subjects factor and trial difficulty (easy or difficult) as a within-subjects factor, age and second sort quality, scaled and centered, as covariates. As seen in Figure 5 (right), there was no significant main effect of intervention ($F(2,266.0)=0.08$, $p = .93$), nor any significant interactive effects involving intervention: participants of all ages were near-chance at correctly choosing the diverse pair ($F(1,267.1)=0.00$, $p = .98$). There was no significant main effect of second sort quality ($F(1,269.6)=0.02$, $p = .88$). Easy trials had marginally higher performance than difficult trials ($F(1,1384.6)=3.67$, $p = .06$). There was once again a significant interaction of age and second sort quality ($F(1,262.8)=5.61$, $p = .02$), shown in Figure 6 (right). Children with 4th quartile post-intervention sorts were picking the diverse pair at close to chance rates at all ages, while 8-year-olds with lower-quality sorts preferred the less informative, nondiverse samples. All other effects were not significant.

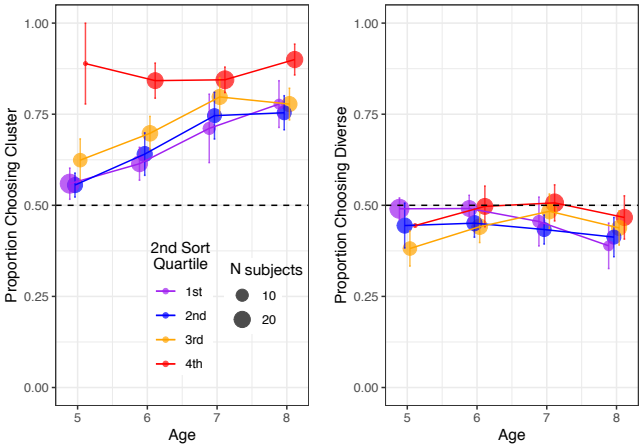


Figure 6: Mean proportion of correct sampling choice for induction (left: nondiverse for cluster, right: diverse for category) as a function of age and quartile of 2nd sort's quality. Error bars show +/-1SE, and dotted lines show chance.

Given that participants were at best choosing the diverse sample at as often as expected by chance on category induction trials, we investigated whether they showed any systematic sampling strategies on these items. As a reminder, each sampling trial presented three birds: one appearing in both samples, one from the same cluster, and one from a different cluster. We considered the possibility that, among the two varying birds per trial, children preferred to sample some types (i.e. clusters) of birds over others. Table 2 shows, for each type of sampling trial (category vs. each cluster), the proportion of trials on which participants chose a sample containing a bird of a given cluster. For category induction trials, participants avoided choosing samples with waterbirds, choosing them only 21% of the time, significantly different than the 40% raptors and 39% songbirds chosen ($\chi^2(2, N = 2,373) = 162.7$, $p < .001$). Indeed, this bias against sampling waterbirds extended to the cluster induction trials: participants had lower rates of choosing a waterbird to induce to all waterbirds (48%) than of choosing a raptor to induce to all raptors (80%), or than choosing a songbird to induce to all songbirds (71%). For incorrect answers, waterbird was the least popular choice (0% for songbird trials

[3]Covariates scaled and zero-centered.

	Chosen:		
Induce To:	Raptor	Songbird	Waterbird
All Birds	0.40	0.39	0.21
Raptors	0.80	0.13	0.08
Songbirds	0.29	0.71	0.00
Waterbirds	0.36	0.16	0.48

Table 2: Participant's choices on induction trials.

and 8% for raptor trials). It may be that participants avoid waterbirds because it is the cluster that they find least typical of birds. To verify this, we asked 13 adults to rate the typicality (1-7; 1=atypical, 7=very typical) of the 8 waterbirds, 8 raptors, and 8 songbirds used in this experiment. They were shown the same unlabeled pictures as children saw, in a randomized order. On average, participants found songbirds the most typical ($M = 5.92$), followed by raptors ($M = 4.41$), and waterbirds ($M = 3.54$).

General Discussion

The present study investigated whether teaching children about the clustered structure of the bird category would increase the diversity of their sampling choices in a category induction task. The didactic dialogue and intervention displays significantly shifted children's representation of the bird category, as evidenced by a shift from them sorting birds into piles according to single dimensions (e.g., size or color), to them predominantly sorting by cluster (songbird, waterbird, and raptor) after the intervention. Moreover, children's pile sorts post-intervention were much closer to the actual cluster structure, as measured by adjusted Rand index. Despite this shift in children's representation of the bird category, we found no evidence that either intervention display improved their category induction sampling choices. Children of all ages were near-chance at choosing the diverse sample when inducing to the category. Yet, the quality of children category's representations was indeed related to their sampling behavior in some cases. Younger children with more accurate category representations chose more informative samples on the cluster-induction trials. Also, older children with more accurate category representations chose diverse samples more often than their age-matched peers with less accurate category representations, although even these children did not choose diverse samples more often than expected by chance.

Given that the intervention conditions increased the accuracy of children's category representations, and category representations were somewhat related to children's sampling decisions, it is surprising that the intervention was not powerful enough to increase the efficiency of children's sampling strategies. One possibility is that children simply apply a different standard for evaluating the informativeness of samples. For instance, Foster-Hanson et al. (2019) found that young children prefer to examine highly prototypical examples instead of samples that cover variation. This tendency might have been behind children's decisions in the present study to avoid sampling waterbirds–the least typical birds presented according to adult raters. From this perspective, although the intervention made children more aware of the structured variability that exists within the category, perhaps it did not lead them to view that variation as informative because they prefer to rely on separate criteria (typicality) for evaluating samples of evidence.

We conclude that it is unlikely that simply knowing the clustered structure of a natural category is sufficient for children to realize the inductive power of choosing diverse samples. There is certainly merit to learning the structure of natural categories beyond any benefit to diversity-based reasoning, but future interventions seeking to increase appreciation for diverse sampling strategies may wish to demonstrate to children the value of choosing diverse samples more directly.

Acknowledgments

This work was supported by the John Templeton Foundation "Varieties of Understanding" grant to T.M.G. and M.R., National Science Foundation grant BCS-1255538 to T.M.G., and National Science Foundation grant BCS 1729540 to M.R. We are grateful to Kathryn Yee, Aja Blanco, Christina Chu, and Talena Smith for data collection, and to Daniel Zeiger and the staff of the American Museum of Natural History's Discovery Room for their support of this project.

References

Gelman, S. A. (1988). The development of induction within natural kind and artifact categories. *Cognitive Psychology*, *20*, 65-96.

Gelman, S. A. (2003). *The essential child: Origins of essentialism in everyday life*. New York, NY: Oxford University Press.

Gutheil, G., & Gelman, S. A. (1997). Children's use of sample size and diversity information within basic-level categories. *Journal of Experimental Child Psychology*, *64*(159-174).

Heit, E. (2000). Properties of inductive reasoning. *Psychonomic Bulletin and Review*, *7*, 569–592.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of the Classification*, *2*, 193–218.

James, W. (1890). *The principles of psychology*. New York: Henry Holt.

Kalish, C. W. (1998). Natural and artificial kinds: Are children realists or relativists about categories? *Developmental Psychology*, *34*(376-391).

Kim, N. S., & Keil, F. C. (2003). From symptoms to causes: Diversity effects in diagnostic reasoning. *Memory and Cognition*, *31*, 155-165.

Lopez, A. (1995). The diversity principle in the testing of arguments. *Memory and Cognition*, *23*, 374-382.

Osherson, D. N., Smith, E. E., Wilkie, O., Lopez, A., & Shafir, E. (1990). Category-based induction. *Psychological Review*, *97*, 185-200.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, *66*(336), 846-850.

Rhodes, M., & Brickman, D. (2010). The role of within-category variability in category-based induction: A developmental study. *Cognitive Science*, *34*, 1561-1573.

Rhodes, M., Brickman, D., & Gelman, S. A. (2008). Sample diversity and premise typicality in inductive reasoning: Evidence for developmental change. *Cognition*, *108*, 543-556.

Rhodes, M., & Gelman, S. A. (2008). Categories influence predictions about individual consistency. *Child Development*, *79*, 1271-1288.

Rips, L. J. (1975). Inductive judgments about natural categories. *Journal of Verbal Learning and Verbal Behavior*, *14*, 665-681.

Does Children's Shape Knowledge Contribute to Age-Related Improvements in Selective Sustained Attention Measured in a TrackIt Task?

Emily Keebler (ekeebler@andrew.cmu.edu)
Carnegie Mellon University, Department of Psychology, 5000 Forbes Avenue
Pittsburgh, PA 15213 USA

Jaeah Kim (jaeahk@andrew.cmu.edu)
Carnegie Mellon University, Department of Psychology, 5000 Forbes Avenue
Pittsburgh, PA 15213 USA

Erik D. Thiessen (thiessen@andrew.cmu.edu)
Carnegie Mellon University, Department of Psychology, 5000 Forbes Avenue
Pittsburgh, PA 15213 USA

Anna V. Fisher (Fisher49@andrew.cmu.edu)
Carnegie Mellon University, Department of Psychology, 5000 Forbes Avenue
Pittsburgh, PA 15213 USA

Abstract

The ability to maintain attentive state over a period of time (i.e., Selective Sustained Attention) is important for higher-order cognition but challenging to assess in preschool-age children. The TrackIt task was developed to address this challenge and has been argued to be sensitive to age-related differences in selective sustained attention in 3- to 5-year-old children. However, it remains unclear whether this improvement with age also (or predominantly) reflects improvement in children's knowledge of different shapes used as stimuli in this task in prior studies. The current study addressed this possibility. Consistent with prior studies, we found clear age-related improvement in performance on TrackIt. However, we did not find evidence that shape knowledge played a role in TrackIt performance for children aged 2 to 5, suggesting that increased knowledge of geometric shapes is not sufficient to explain age-related improvement in performance and helping to validate TrackIt as an assessment of Selective Sustained Attention.

Keywords: selective sustained attention; TrackIt

Introduction

The ability to maintain attentive state over a period of time (often referred to as Focused or Selective Sustained Attention) is important for higher-order cognition, including learning (e.g., Fisher & Kloos, 2016; Oakes, Kannass, & Shaddy, 2002). This ability undergoes marked development during the preschool years as shown by the increased time that children spend in this state during free play assessments of selective sustained attention (Ruff & Lawson, 1990; Sarid & Breznitz, 1997); however, few experimental paradigms capture usable data for children in this age range (for review see Fisher & Kloos, 2016).

The TrackIt paradigm was designed to address this measurement gap. In the TrackIt task, participants visually track a target object moving along a random trajectory on a grid, while simultaneously ignoring distractor objects. At the conclusion of the trial, the objects disappear and the participant indicates the final location of the target on the grid. Prior studies suggest that nearly all preschool-age children can complete and provide usable data on this task (in contrast to other assessments, such as downward extension of the Continuous Performance Test; see Fisher & Kloos, 2016). Performance on this task shows considerable age-related improvement between 3 and 5 years of age (Fisher et al., 2013) showing that the task is developmentally sensitive. Importantly to this paper, the target and distractor objects in the TrackIt task are usually selected from a set of geometric forms (circle, diamond, square, triangle, pentagon) and iconic shapes (crescent, cross, arrow, semi-circle).

Age-related improvement in performance on the TrackIt task during the preschool period has been interpreted as improvement in selective sustained attention (Brueggemann & Gable, 2018; Erickson et al., 2015; Fisher et al., 2013). However, shape knowledge is also known to improve during the preschool period (e.g., Clements et al., 1999; Verdine et al., 2016) and could be an important element of successful completion of the TrackIt task. Therefore, it remains unclear whether increased shape knowledge may account for the age-related improvement in performance on the TrackIt task. This finding would challenge prior interpretations that age-related changes in TrackIt performance primarily reflect improvement in selective sustained attention.

Shape knowledge may play a role in the task in the following way. When distractors are unique from each other and from the target, all objects in the task are comparable in salience (Fisher et al., 2013). Therefore, participants need to encode the identity of the target object in order to successfully complete the task. Children may encode the identity of the target object by maintaining its visual representation in working memory and by using labels to refer to object shape. Younger children whose shape knowledge is still developing may encode the identity of the target object less robustly than older children with greater shape knowledge.

There is indirect evidence to support this possibility. Vales and Smith (2015) provided evidence that object labels help children maintain precise representations of objects in working memory during a visual search task. Consistent with this explanation, Doebel et al. (2018) showed that preschool children were better at a modified TrackIt task with novel shapes (for which children did not have consistent labels) when experimenters provided labels. This result was found even though children were able to identify the shape from a set of choices after the task was complete (i.e., a memory check). Thus, although children completing TrackIt with geometric forms have demonstrated memory check accuracy that is well above chance, the encoding necessary to recognize the target object after the trial may be insufficient to support accuracy on the main task. Instead, children's own knowledge of shape names may facilitate their performance on the TrackIt task when the experimenter does not provide labels for the target objects (as is the standard procedure on the TrackIt task) by enabling the children to self-generate labels of the targets.

It is possible that children may use non-canonical names for shapes when they do not know the proper labels. For example, when asked to describe geometric forms, Clements et al. (1999) found that young children tended to invoke visual descriptions of geometric forms (e.g., "pointy," "round", or "skinny"). However, such visual descriptions comprise non-unique labels (e.g., both a diamond and a triangle could fit the visual description "pointy"). Therefore, if younger children generate visual descriptor labels when they do not know the canonical labels, these visual descriptor labels may still be less helpful for encoding the target identity than canonical labels that are more likely to be known (and self-generated) by older children.

In the current study we examined the possibility that age-related improvement in performance on the TrackIt task may be attributed, at least partially, to age-related increase in shape knowledge.

Experiment 1

Method

Participants 90 two- to five-year old children (*M* = 3.89 years, *SD* = 9.4 months, range 2.58 to 5.77 years) participated in the study. Participants were drawn from public and private preschool and kindergarten programs. The data reported are part of a larger cross-sectional study for which data collection is in progress, that has a final intended sample size of 240 participants aged 2-7 years. That larger study is preregistered at aspredicted.org, and the anonymized preregistration is available here. The target shape analyses reported in this paper were not pre-registered. Of the 90 participants recruited for this study, 3 participants were excluded from the analysis because they refused or otherwise failed to complete ten trials or due to experimenter error.

Materials and Apparatus The TrackIt task (freely available at http://www.psy.cmu.edu/~trackit) was presented on a Lenovo laptop screen with physical dimensions 19.1 cm x 34.2 cm and pixel dimensions 1920x1080 pixels. Participants were seated at a desk facing the screen with their heads about 12 inches away from the screen. For each trial, the target and distractor objects were randomly picked without replacement from a set of unique objects spanning 9 different shapes with 9 different color possibilities (81objects in total). See Figure 1 for examples.

We expect that young children have differential knowledge of the shape stimuli used in the TrackIt task (i.e., children are likely to know some, but not all, of the nine shapes and their associated labels). Because encoding the identity of the target object is necessary to complete the TrackIt task, greater knowledge of a target shape may result in better accuracy on trials with that target shape relative to trials with a less familiar shape. To represent the relative familiarity of the target shapes to one another, we assessed the frequency of the stimuli using ChildFreq (Bååth, 2010), a tool that extracts word frequencies from the American and British parts of the Childes database (MacWhinney, 2010). In particular, we found the frequency of the canonical names for the nine shapes over the for the age range 12-35 months (see Table 1). As is shown in Table 1, there was considerable variability in the frequency of the stimuli, ranging from 1 to 273 occurrences per million words.

Table 1: Frequency of Stimuli in the Childes Database

Stimuli	**Occurrences per 1,000,000 Words**
Circle	273
Triangle	165
Square	126
Cross	91
Diamond	26
Pentagon	11
Arrow	1
Crescent	1
Semicircle	1

Procedures The experimenter administered the TrackIt task to participants in a quiet room or hallway. In the TrackIt task, participants were asked to visually track a single target object as it moved on a grid among moving distractor objects. At the beginning of each trial, the objects appeared on the grid, centered in distinct grid cells, and the target object was indicated by a red circle around it. The initial positions of the objects were randomized. At the beginning of the task, participants were told that: 1) the objects will start moving around the grid when the experimenter presses a button; 2) the goal is to follow the target object with their eyes; 3) at some point the objects will suddenly disappear, and their job is to point to where the target object was when it disappeared.

The experimenter started each trial with a button press after ensuring the participant was ready to begin. Upon starting the trial, the red circle disappeared, and the objects

began to move in curvilinear trajectories from grid cell to grid cell at a constant speed. At the end of each trial, all objects disappeared from the screen, and the participants were asked to indicate with their finger (on the touch screen) which grid cell the target object was last in before it disappeared. Each trial was followed by a memory check screen and a smiley face. Participants were told that the smile did not indicate a correct answer and rather that we were happy they were playing our game. See Figure 1 for a diagram of the task sequence.

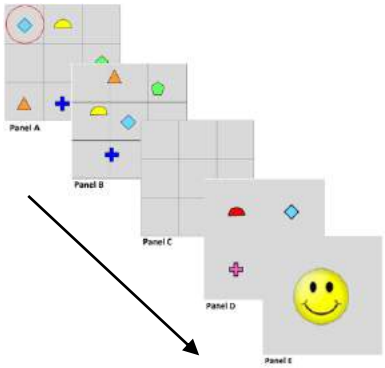


Figure 1: The TrackIt task pipeline. Panel A: static display of the stimuli before the trial starts; Panel B: the stimuli move along random trajectories during the trial; Panel C: response screen after the moving shapes disappear; Panel D: memory check; Panel E: a smiley face at the end of the trial.

Participants completed 11 trials of the task. The first trial was a practice trial and was completed with assistance from the experimenter who traced the moving target with their index finger. The first trial was accordingly omitted from analysis. Participants were then told that they would need to complete the rest of the task by themselves, tracking the target with their eyes only.

Design The sequence of positions in the path of each of the objects was randomized. Object motion display was set to 30 frames per second. The minimum trial length was set to 10 milliseconds. The parameters—grid size, number of distractors, and speed of objects—were determined by prior testing in TrackIt with a separate group of 3- to 5-year old children (Kim et al., 2017) and via pilot testing with two-year-olds. The parameters were organized according to participant age and difficulty level as seen in Table 2.

Table 2: TrackIt parameter combination used in each difficulty level

Difficulty	Age Group (years)	Grid Size	# of Distractors	Object Speed (pix/s)
Level 1	2-4	2x2	2	300
Level 2	3-5	4x4	4	500

Note: pix/s = pixels/second

Separate groups of participants were tested in each difficulty level. We did not complete testing for age and level combinations that were likely to produce floor or ceiling effects. The final sample size per age and difficulty level is presented in Table 3.

Table 3: Sample sizes and age statistics for each age group, for each difficulty level

Age (years)	Difficulty Level 1		Difficulty Level 2	
	n/m/f	Age Mean (Std)	n/m/f	Age Mean (Std)
2 y.o.	13/8/5	2.89 (0.12)	--	--
3	19/7/12	3.53 (0.27)	20/12/8	3.56 (0.24)
4	14/7/7	4.39 (0.26)	12/5/6 (1 not reported)	4.31 (0.15)
5	--	--	9/4/5	5.48 (0.22)

Note: n/m/f = sample size / # male/ # female.

Results and Discussion

Age and Task Level For each participant, we calculated an average accuracy score i.e., the proportion of ten trials for which the participant correctly identified the grid cell in which the target object disappeared. To investigate possible effects of participant age and task difficulty level, accuracy scores were submitted to a 2-way analysis of variance (ANOVA) with age and difficulty level as between-subject factors. This analysis indicated main effects of age (F (3, 81) = 11.40, p < .001) and difficulty level (F (1, 81) = 19.16, p < .001), but no age-by-difficulty interaction (F (1, 81) = 1.65, p = .20) (See Figure 2).

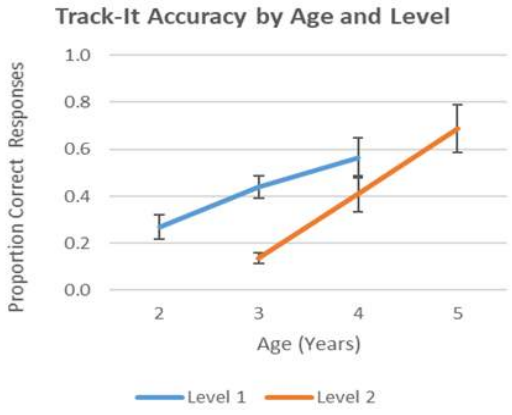


Figure 2: TrackIt accuracy improved with age for participants tested in Levels 1 and 2.

Post-hoc Tukey's tests showed that, for Level 1, the tracking accuracy of 4-year olds was significantly above that of 2-year olds (adjusted p = .02) but not 3-year olds (adjusted p = .44). For Level 2, the tracking accuracy of 4-year olds was significantly above that of 3-year olds (adjusted p = .03). However, the tracking accuracy of 5-year olds was not significantly above that of 4-year olds (adjusted p = .10). Nonetheless, there is an emergence of developmental trends that are consistent with Fisher et al. 2013 and Kim et al. 2017 and further, planned data collection (i.e., to bring the number of participants in each cell to 20) will shed light on any further age-related differences. Post-hoc Tukey's tests also showed that 3-year olds performed significantly better in Difficulty Level 1 than in Difficulty Level 2 (p < .01). Surprisingly, 4-year-olds did not show a significant difference in performance at Difficulty Levels 1 and 2 (adjusted p = .64).

For all combinations of difficulty level and age group, TrackIt accuracy was above chance (25% given four response options in Level 1 and 6.25% given 16 response options in Level 2, all one-sample t s > 3.62, p s < .001), except for two-year-old children completing Difficulty Level 1 (one-sample t (12) = 0.63, p = .54). This result indicates that two-year-olds did not differ from chance performance on the TrackIt task.

Shape Frequency Next we assessed the possibility that the frequency of a target shape influenced children's performance on trials with that target shape. The average proportion of correct trials, sorted by target shape, ranged from 0.32 (diamond) to 0.44 (crescent). To determine whether TrackIt performance varied significantly by shape frequency, we conducted a logistic regression using shape frequency to predict accurate TrackIt responses while controlling for participant age and task difficulty level. Results of the regression indicated that participant age and task difficulty level, but not frequency of target shape, were associated with TrackIt performance (see Table 4).

Table 4. Results from the logistic regression analysis: target shape frequency, difficulty level, and participant age as predictors of correct answer on a trial of TrackIt

Predictor	*B*	*SE B*	Wald	*P*	*df*
Shape Frequency	-0.00	0.00	-6.68	.89	868
Difficulty Level	-0.54	0.08	-0.14	<.001	868
Participant Age	0.98	0.10	-6.45	<.001	868

Similarly, results of a Pearson correlation did not indicate an association between frequency of a target shape and children's average TrackIt accuracy on trials of that target (r = -0.01, p = .72). See Figure 3.

Based on these results, it does not appear that shape knowledge can account for any variability in TrackIt performance, a finding that helps to validate TrackIt as a measure of selective sustained attention. However, there are several cautions in using frequency data as a proxy for shape knowledge. Some of these concerns are grammatical/technical in nature e.g., given the nature of the data, it is unknown what proportion of word utterances co-occurred with a concrete or pictorial referent, and this co-occurrence structure might matter for the encoding of the referent shapes.

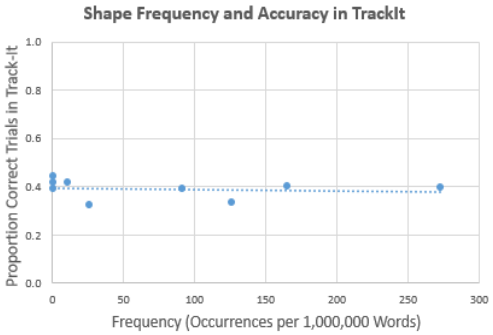


Figure 3: Corpus frequency of target shapes did not account for variability in TrackIt performance.

Relatedly, some of the stimuli names (see column 1 of Table 1) can be used as verbs with semantically-related meanings to the shapes whose names they share (e.g., circle, cross) and/or adjectives with meanings unrelated to the shapes whose name they share (e.g., cross). Some stimuli are both the nominal and adjectival form of the shape name (e.g., square); whereas, other shapes have a morphologically related but distinct adjective form (e.g., circular, triangular). These nuances might bias to the number of occurrences of each target shape in the ChildFreq database. Perhaps more critically, the nature of interactions captured in the Childes database may be biased toward free-play and informal interactions, rather than formal educational experiences. Accordingly, it might underestimate the frequency of less-common shape names, to which children might be exposed in other, more explicitly educational interactions not captured in the data.

Nonetheless, we posit that—for this age group—the relative frequencies observed likely comprise reasonable approximations of shape familiarity i.e., circle, triangle and square are the most common and early-emerging shape names in our stimuli set, with crescent and semi-circle being significantly less frequent. However, to address the concerns about using relative frequencies as a proxy for children's shape knowledge, in Experiment 2 we directly assessed the shape knowledge of three- to five-year-old children, as described below.

Experiment 2

Method

Participants We tested 32 participants to assess children's knowledge of the shapes (M = 4.47 years, SD = 9.3 months, range 3.24 to 5.84 years). Participants were drawn from preschool and kindergarten programs. 16 of these children

were also participants in Experiment 1 (M = 4.29 years, *SD* = 8.7 months, range 3.35 to 5.70 years) and completed the shape knowledge task an average of 9.7 weeks after the TrackIt task. It is unlikely that participation in the TrackIt task affected children's performance on the shape knowledge task. The total sample for Experiment 2 included 11 three-year-olds (8 females, *M* = 3.56 years, *SD* = 1.9 months); 11 four-year-olds (6 females, *M* = 4.52 years, *SD* = 3.0 months); and 10 five-year-olds (7 females, *M* = 5.40 years, *SD* = 2.7 months).

Given that (1) two-year-olds were at chance in identifying the last location visited by the target shape in Experiment 1 and (2) pilot testing indicated that two-year-old children had difficulty producing verbal responses on the shape knowledge assessment, we did not include this age group in Experiment 2.

Materials and Apparatus The physical equipment and child seating position is identical to those of Experiment 1. Shapes presented were the set of TrackIt stimuli, made identical in color and equated for overall size (see Figure 4).

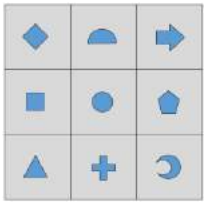


Figure 4**:** The set of geometric forms and iconic shapes comprising the TrackIt stimuli, presented on a single grid.

Procedures Shapes were displayed one at a time in the center of a gray screen. Children were instructed to provide verbally the name of each shape and prompted to "make their best guess" as necessary. No feedback was provided on the accuracy of children's responses. The experimenter demonstrated the task across 6 practice trials that presented the stimuli star, heart, and oval two times each. The nine stimuli were sampled without replacement, after which the block of nine was repeated two more times for a total of 27 trials (3 presentations of each of the 9 target shapes).

Results and Discussion

As expected based on the ChildFreq statistics, children demonstrated superior knowledge of high-frequency shape names (e.g., circle, triangle) relative to low-frequency shape names (e.g., crescent, semicircle). Results of the Pearson correlation indicated that there was a positive association between shape frequency and children's shape knowledge, (r = .82, p < .01) (Figure 5). Accordingly, we have put forth two complementary approaches for assessing shape familiarity for the TrackIt stimuli.

To assess possible age-related changes in shape knowledge, we conducted an ANOVA on children's shape

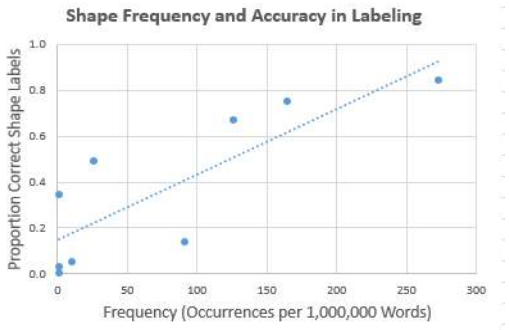


Figure 5: Children's accuracy in labeling shapes positively correlated with the frequency of those names in the corpus.

knowledge using participant age and shape frequency predictors. This analysis indicated main effects of age (F (1, 284) = 19.85, p < .001) and shape frequency (F (1, 284) = 147.40, p < .001).

Despite finding better shape knowledge in older children than in younger children, we did not find a relationship between children's knowledge of shapes and average performance on TrackIt trials with that target shape (see Figure 6 for a visualization).

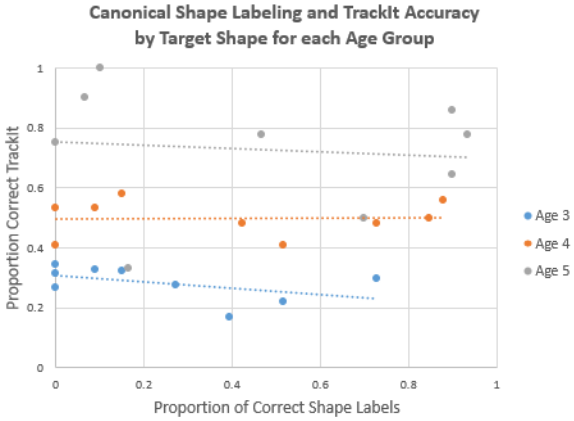


Figure 6: Children's accuracy in shape labeling and on TrackIt trials with that target shape, by age.

We further assessed this relationship using a logistic regression on trial accuracy by proportion correct shape labels with a control for difficulty level (Table 5).

Table 5. Results from the logistic regression analysis: proportion correct shape labels and difficulty level as predictors of producing a correct answer on a trial of TrackIt

Predictor	*B*	*SE B*	Wald	*P*	*df*
Shape Label	-0.26	0.23	-1.15	.25	868
Difficulty Level	-0.35	0.14	-2.48	.01	868

Similarly, results of a Pearson correlation did not indicate an association between average ability to name a target shape and accuracy on trials of that target (r = -0.04, p = .24).

Other Names The results above are based on children's productive shape knowledge of a single canonical name for each target shape (see column 1 of Table 1). We additionally assessed the extent to which the findings held when allowing for other valid names for the target shapes. Ten adult graduate students who were blind to the hypothesis (M = 28.62 years, SD = 6.33 years, range 25.15 to 46.08 years) assigned each shape-label match generated by children in Experiment 2 no credit, half credit, or full credit. Adults rated all canonical names as full-credit responses. When indicated by consensus agreement (80 percent) non-canonical names were assigned full-credit (e.g., "moon" for crescent, "plus" for cross) or half-credit (e.g., "ball" for circle, "right" for arrow).

Using this coding scheme to represent children's shape knowledge, there remained a positive association between shape frequency in the ChildFreq statistics and children's shape knowledge (r = .82, p = .03). In addition, we still found evidence for age-related changes in shape knowledge: an ANOVA on children's shape knowledge using participant age and shape frequency as predictors indicated main effects of age (F (1, 284) = 27.89, p < .001) and shape frequency (F (1, 284) = 81.48, p < .001). Importantly, we did not find a relationship between children's knowledge of shapes and average performance on TrackIt trials with that target shape (see Figure 7).

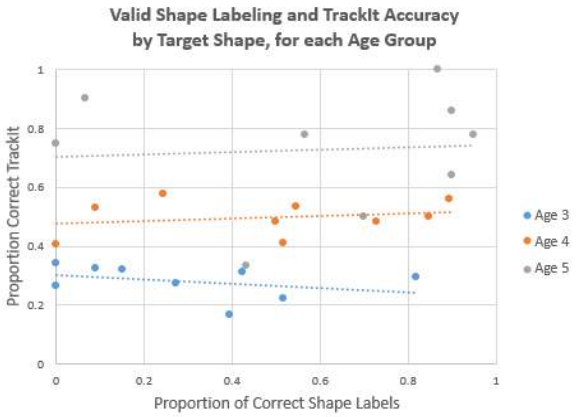


Figure 7: Children's accuracy in shape labeling and on TrackIt trials with that target shape, by age, when allowing for non-canonical labels judged valid by adult participants.

General Discussion

Consistent with prior research, we found effects of age and task difficulty level on performance in the TrackIt task in Experiment 1 (Fisher et al., 2013; Kim et al., 2017). In Experiment 2, we also found that indeed children's shape knowledge was related to age, with older children showing better shape knowledge of shape labels than younger children. We also found that across age, children showed better knowledge of shape labels for more frequently occurring labels. However, across Experiments 1-2 we did not find evidence that shape knowledge or frequency can account for age-related improvement in performance on the TrackIt task. Specifically, we did not find evidence that the frequency of a target shape, as derived from the ChildFreq database, was related to children's performance on TrackIt trials using that target shape. Similarly, we did not find a significant relationship between children's ability to label a target shape and their performance on trials involving that target shape. In contrast, our analyses indicate that children performed similarly across trials regardless of target shape.

These findings help to mitigate concerns that shape knowledge may contribute to children's performance on the TrackIt task, given that knowledge of the different target shapes is likely to emerge at different time points and rates (i.e., if knowledge of the target shape names were a critical aspect of task success, we would expect young children in particular to perform relatively better on trials with high-frequency shapes relative to those with less familiar shapes).

At the same time, that two-year-old children performed at chance overall (on both the main task and the memory check) might indicate that these youngest participants have difficulty encoding any target shape, regardless of its relative frequency. Additional development and school experience might support older children in recognizing the properties of shapes, even if they are not familiar with the canonical names of these shapes (as both the Childes database and children's own performance suggest).

One limitation of the current studies is that our analysis did not account for object color, the other dimension by which target and distractor shapes differed. Children with limited shape knowledge might nonetheless be successful in encoding the target object by using color labels (see Sandhofer & Smith, 1999, for a review of the time course and developmental dependencies of color term learning). We did not test for this hypothesis because currently the TrackIt output records only object shape but not color.

Another limitation of the current set of studies is that some (but not all) of the children providing shape knowledge data in Experiment 2 also completed the TrackIt task in Experiment 1. An alternate design would have allowed us to more directly assess shape knowledge of TrackIt participants, rather than that of a representative peer group.

Conclusions

Across two experiments we obtained no evidence that shape knowledge contributed to children's performance accuracy on the TrackIt task. Accordingly, the results of the present study help to mitigate the concern that shape knowledge may fully or partially account for the age-related changes in performance on the TrackIt task reported in prior studies. Overall, the reported results help to support the previous interpretation of this task as an assessment of selective sustained attention in young children (Erickson et al., 2015; Fisher et al., 2013; Kim et al., 2017).

Acknowledgements

We would like to thank Melissa Pocsai and Oceann Stanley for their support with participant recruitment, scheduling, and data collection; Rea Isaac, Priscilla Medor, and Elaine Xu for their work in collecting data; and the children, parents, and teachers who made this work possible: Amazing Scholar Academy Preschool, Beth Shalom Early Learning Center, Campus School of Carlow University, CMU Children's School, Glenn Avenue Preschool, Propel East, Sacred Heart Elementary School, Tender Care Learning Center of Greentree, and Tender Care Learning Center of Jefferson. The work reported here was supported by the National Science Foundation through a grant awarded to A.V.F. and E.D.T. (BCS-1451706). The research reported here was supported in part by the Institute of Education Sciences, U.S. Department of Education, through grant R305B150008 to Carnegie Mellon University. The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education.

References

Bååth, R. (2010). ChildFreq: An Online Tool to Explore Word Frequencies in Child Language. *LUCS Minor*, 16.

Brueggemann, A. & Gabel, S. (2018). Preschoolers' selective sustained attention and numeracy skills and knowledge. *Journal of Experimental Child Psychology, 171*, 138-147.

Clements, D. H., Swaminathan, S., Zeitler Hannibal, M. A., & Sarama, J. (1999). Young children's concepts of shape. *Journal for Research in Mathematics Education, 30*, 192-212.

Doebel, S., Dickerson, J. P., Hoover, J. D., & Munakata, Y. (2017). Using language to get ready: Labels help children engage proactive control. *Journal of Experimental Child Psychology, 166*, 147-159.

Erickson, L. C., Thiessen, E. D., Godwin, K. E., Dickerson, J. P., & Fisher, A. V. (2015). Endogenously and exogenously driven selective sustained attention: contributions to learning in kindergarten children. *Journal of Experimental Child Psychology, 138*, 126-134.

Fisher, A.V., & Kloos, H. (2016). Development of selective sustained attention: The role of Executive Functions. In J. A. Griffin, P. McCardle, & L. Freund (Eds.), Executive Function in Preschool-age Children: Integrating Measurement, Neurodevelopment, and Translational Research. Washington, DC, US: American Psychological Association.

Fisher, A. V., Thiessen, E. D., Godwin, K. E., Kloos, H., & Dickerson, J. P. (2013). Assessing selective sustained attention in 3- to 5-year-old children: Evidence from a new paradigm. *Journal of Experimental Child Psychology, 114*, 275-294.

Kim, J., Vande Velde, A., Thiessen, E. D., & Fisher, A. V. (2017). Variables involved in selective sustained attention development: Advances in measurement. *Proceedings of the 39th annual conf. of the Cognitive Science Society.*

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk. Third Edition.* Mahwah, NJ: Lawrence Erlbaum Associates.

Oakes, L., Kannass, N., & Shaddy, J. (2002). Developmental changes in endogenous control of attention: The role of target familiarity on infants' distraction latency. *Child Development, 73*, 1644-1655.

Ruff, H. A., & Lawson, K. R. (1990). Development of sustained, focused attention in young children during free play. *Developmental Psychology, 26*(1), 85-93.

Sandhofer, C. M., & Smith, L. B. (1999). Learning color words involves learning a system of mappings. *Developmental Psychology, 35*, 668-679.

Sarid, M. & Breznitz, Z. (1997). Developmental aspects of sustained attention among 2 to 6-year-old children. *International Journal of Behavioral Development, 21*(2), 303-312.

Vales, C. & Smith, L.B. (2015). Words, shape, visual search and visual working memory in 3-year-old children. *Developmental Science, 18*(1), 65-79.

Verdine, B.N., Lucca, K. R., Golinkoff, R. M., Newcombe, N. S., & Hirsh-Pasek, K. (2016). The shape of things: the origin of young children's knowledge of the names and properties of geometric forms. *The Journal of Cognition and Development, 17*(1): 142-161.

Curious Topics: A Curiosity-Based Model of First Language Word Learning

Daan Keijser (daankeijser@icloud.com)
Department of Cognitive Science and Artificial Intelligence, Tilburg University
Warandelaan 2, 5037 AB Tilburg

Lieke Gelderloos (l.j.gelderloos@uvt.nl)
Department of Cognitive Science and Artificial Intelligence, Tilburg University
Warandelaan 2, 5037 AB Tilburg

Afra Alishahi (a.alishahi@uvt.nl)
Department of Cognitive Science and Artificial Intelligence, Tilburg University
Warandelaan 2, 5037 AB Tilburg

Abstract

This paper investigates whether a curiosity-based strategy could be beneficial to word learning. Children are active conversation partners and exert considerable influence over the topics that are discussed in conversation with their parents. As the choice of topics is likely to be intrinsically motivated, a formalization of curiosity is implemented in a word learning model. The model receives annotated Flickr30k Entities images as input, and is trained in two conditions. In the curious condition, the model chooses objects to talk about from the scene according to the curiosity mechanism, whereas in the random condition, the model receives randomly chosen objects as input. The goal of this study is to show how a curious, active choice of topics by a language learner improves word learning compared to random selection. Curiosity is found to make word learning faster, increase robustness, and lead to better accuracy.

Keywords: word learning; curiosity; interaction; connectionist model.

Introduction

Language learning research focuses more and more on child-parent interaction and the social aspects of early conversation. Children are active learners and have considerable agency as conversational partners. We will argue that curiosity is a plausible mechanism for the child to come up with new topics to talk about within this conversational context. While AI researchers have become inspired by the curiosity displayed by children, and have implemented intrinsically motivated exploration in computer models, this formalized curiosity has not been applied to computational models of language learning. At the same time, the implementations of curiosity in computer models are often not cognitively plausible or the degree of plausibility is unknown (as in reinforcement learning), or the input to the model lacks the complexity of the stimuli encountered by the word learner.

Curiosity can be seen as a viable mechanism in language learning if it provides an advantage to the word learning child. In order to see whether curiosity is beneficial to the word learning process, we propose a curiosity-based model that chooses which object in a scene to talk about next. The model chooses its object of interest from among a number of objects in an image, and triggers the adult to provide linguistic input related to that object. The curiosity mechanism suggested by Twomey and Westermann (2018), which maximizes the product of subjective novelty and plasticity, was implemented to select the objects. To reflect the complexity of visual scenes encountered by the child, the model takes Flickr30k images as input, which depict everyday scenes and objects and have been annotated with captions. The accuracy and loss of the model with a curiosity-based selection of topics were compared to those of a model that received the topics randomly.

Related Work

Interaction and Intentionality

Given the social nature of early conversation, language should not be seen as a product but as a dynamic system for communication (Clark, 2016). Language is used and learned in order to convey and receive information. This means that the child is a conversation partner first, and a language learner second. Furthermore, young children are active speakers and language learners. Bloom et al. (1996) observed that children aged 9 through 24 months are most likely to speak first in conversation with their mother, and the mother to speak after the child. Their evidence did not support the scaffolding model, in which the parent takes a prominent role in the conversation by providing a framework that controls the elements beyond the capacity of the learner and lets the learner concentrate on those elements they are capable of producing. Rather, children initiate conversations and, as shown in several studies (Chapman, Miller, MacKenzie & Bedrosian, 1981; Bloom et al., 1996), mothers are likely to adopt the topic proposed by the child, and continue to talk about it.

These studies show a pattern of turn-taking with a clear role division. Often, the child wants to discuss a certain topic and starts by talking about it. The parent makes sure they understand what the child is referring to by rephrasing what the child has said, which functions as feedback to the

language-learning child at the same time (Chouinard and Clark, 2003). The child then assesses whether the parent has understood the initial message, after which the conversation can continue. When children initiate conversations and their parents adopt the proposed topics, children can exert considerable influence on the topics that are discussed and consequently on the feedback they receive.

Because children initiate conversations, and continue discussing the topic when they feel they have been understood, their choice of topics is unlikely to be random. As the language learner decides on the topic themself, taking in the current surroundings and situation, the choice is likely to be intrinsically motivated. Our study investigates whether a curiosity-based selection of the topics to be discussed enhances word learning through comprehension of the symbol-referent pair.

Curiosity

Curiosity is a form of intrinsic motivation. Intrinsic motivation can be defined as doing "an activity for its inherent satisfaction rather than for some separable consequence. When intrinsically motivated, a person is moved to act for the fun or challenge entailed rather than because of external products, pressures, or rewards" (Ryan and Deci, 2000, p. 56). In the 1950s, psychological research assumed that human behavior is mostly extrinsically motivated, by physical drives such as those to alleviate hunger and minimize pain. A major shortcoming of this theory was that it did not account for exploratory and other curious behavior in humans and animals—behavior that does not seek immediate reward (Oudeyer & Kaplan, 2007).

When formalized to be programmed into a computer model or robot, intrinsic motivation and curiosity are often conflated (e.g. Pathak et al., 2017). Intrinsic motivation has mostly been applied in reinforcement learning, providing agents (robots and models) with an intrinsic desire to explore their environments and build better models and representations of them (Schmidhuber, 2010). Studies that implemented intrinsic motivation have shown that intrinsically motivated exploration increases the performance of a model when generalizing to other tasks (Pathak et al., 2017), and this is likely to be the case for humans as well (Twomey & Westermann, 2018).

Reinforcement learning implements a variety of formalizations of intrinsic motivation, such as maximizing the decrease of prediction errors, maximizing or minimizing predictability, or choosing the action that maximizes the agent's ability to perform a task. Some approaches use predefined rewards or external signals that provide feedback on motor functions, both of which are certainly not cognitively plausible. Of other approaches, it is simply not known how cognitively plausible they are (Oudeyer & Kaplan, 2007; Twomey & Westermann, 2018). In fact, not a lot is known for certain about the workings of curiosity in human cognition in general and children's cognitive development in particular.

What is clear is that children are natural explorers, displaying a novelty preference from an early age. Novel stimuli have most potential to yield new insights upon exploration, as little is known about them yet. As a stimulus is perceived, it becomes less interesting over time (habituation), and other stimuli become more interesting relative to the current stimulus as they remain novel when not examined (Mather, 2013).

Under various circumstances, however, children display familiarity preferences. While completely novel stimuli leave a lot to be explored, they can be uninteresting nonetheless as they differ greatly from the child's state of knowledge. Some have suggested that a moderate discrepancy between a stimulus and the child's representation of it could define the optimally interesting stimulus. What moderate means in this context, however, is not a trivial question. How familiarity and novelty preferences influence learning is little understood as of yet (Mather, 2013).

In a recent publication on curiosity-based categorization in infants, Twomey and Westermann (2018; henceforth T&W) simulate infant categorization using an autoencoder, a model that learns to reproduce the input after reducing it to a compact representation. They defined curiosity as maximizing

$$(i - o)o(1 - o) \qquad (1)$$

where i stands for the model input and o for the model output. $(i - o)$ reflects the difference between the input and the output, which is the error of the autoencoder in response to a particular stimulus. $o(1 - o)$ is the derivative of the sigmoid activation function. As such, this part of the formula reflects the potential update made to the model in response to this stimulus, when it is trained using gradient descent. The formula favors stimuli which the model is predicting least accurately (the difference between input and output is large), and stimuli where a small adjustment in representation has the greatest effect on the prediction in terms of accuracy (the sigmoid derivative is large). In T&W, the curiosity condition learned the most robust category, followed by the objective complexity condition.

T&W provide a cognitively plausible mechanism of curiosity, that produced results that fit their empirical data well. That the implementation of curiosity outperformed the other three mechanisms shows that a learner would benefit from applying this strategy. The inputs used in the study are very interpretable, but also rather simple, consisting of eight training instances and three test instances that differed on four features. The present model will use the same curiosity mechanism, and see how it performs when provided with more complex input, consisting of a sizable set of images to approximate the complexity of the language learner's surroundings.

The model of T&W went through the stimuli without replacement, so that the model encountered every stimulus once per epoch. A drawback of this setup is that it does not correspond to how children encounter stimuli in real life, as children have no control over the order in which stimuli are

presented to them. It is also unlikely for children to come across a string of examples of a certain category presented one after the other. Objects and living things are often seen in isolation from other category members, and amid objects of a wide variety of other categories. Our model was therefore presented scenes containing multiple objects it could choose from. The model would pick one object, skipping the other objects as it went on to the next scene. It was free to look at the same or any other object in the scene during the next epoch, meaning that some objects could be ignored altogether. This made the input sequences of our two experimental conditions more different, and perhaps less comparable than in T&W's case, but it also better approximated a word learning context, in which only certain aspects of a scene are in focus at any time.

Methodology

Model

Our language learner model is inspired by a model of referential expression resolution (Rohrbach et al., 2017), which incorporates an expression generation module as well as the main expression resolution component, which allows it to learn under self-supervision. We implement a similar complementary setup, consisting of a listener and a speaker module. The listener represents a child learning which words represent which objects in the visual modality, by receiving linguistic input from an oracle, which represents an adult conversation partner. The listener learns through supervision, comparing the true referent of a word to the referent it expected, and updating its language knowledge accordingly.

The incorporation of a speaker module in principle allows the model to be used in a conversational set-up, but in the current work, the emphasis is on comprehension. As we describe in more detail in the section on 'Curiosity', the model's curiosity about an object is calculated based on the ability of the listener to comprehend the label the speaker would give it. The oracle labels the object the learner model is most curious about. In analogy, a parent might name an object their child points out. Learning, however, is not simply mapping the label to the correct object: just like in the random condition, the model learns by predicting the referent of the given word and getting feedback on this prediction. The curiosity mechanism affects only the order the stimuli are presented in, but not the learning process itself. Figure 1 illustrates the architecture of the model. The listener learns to map a given word to its referent in the visual context. A visual scene consists of a number of objects. We extract a visual feature vector for each object using the VGG-16 object recognition model presented by Simonyan & Zisserman (2015), pretrained on ImageNet. We use the last fully connected 4096-dimensional layer, which contains high-level visual information. For each object in a given scene, the embedding of the word given by the oracle was concatenated to the object representation, which was input to the listener. The listener further consists of a 256-unit hidden layer followed by a sigmoid activation function, which is fully connected to a single output unit, also followed by sigmoid activation. Softmax applied to the concatenation of the output values for all the objects in a scene gives a distribution reflecting the probabilities of each object being the referent. The listener was trained under supervision using cross-entropy loss on the concatenated output values. The loss function is a quantification of how far off the model's prediction is from the actual target distribution. Hence, a lower loss value means a better performing model.

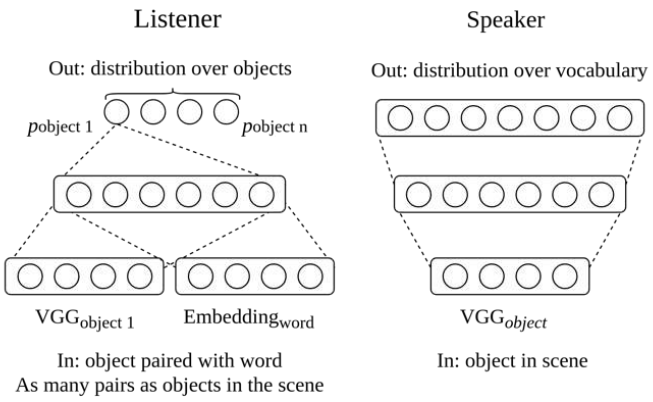


Figure 1: Simplified graphical representation of the model.

The speaker module learns to output a word, given an object. Input to the speaker is a VGG vector, which is fed to a 256 unit hidden layer followed by sigmoid activation, and fully connected to the vocabulary-sized output layer. The speaker was trained using cross-entropy in a self-supervised manner. Rather than training on a single object VGG vectors, it was fed the sum of the VGG vectors of all objects in the scene, weighted by the Softmaxed output vector of the listener (using it as attention). The self-supervision signal consists of the original input word to the listener. Therefore, the speaker can be thought of as learning in an unsupervised manner, although its performance is dependent on that of the listener, which is trained under supervision.

The model was trained using Adam optimization (Kingma & Ba, 2014) in batches of 40 images, for a maximum of 40 epochs. To decide on an initial learning rate, we ran both the 'curious' and the 'random' model, with learning rates ranging from .1 to .00001 for 20 epochs. We ran each condition-learning rate combination with 5 different random initializations. We found that the best scores on the validation data were sometimes obtained in epoch 20, which suggested the model might not have fully converged yet. We therefore decided to report on models trained for 40 epochs. A learning rate of .001 yielded the best results on validation data for both the listener and the speaker. The results reported reflect 20 different runs of both conditions, with learning rate set to .001. The model was implemented in PyTorch (Paszke et al., 2017). The code is available at https://github.com/DaanKeijser/Curious-Topics.

Original captions

A little boy is looking out the balcony surrounded by plants, a toy bike, and plant pots.

A very young boy is looking over the balcony by standing on one of his toy bikes.

A child views the world from their upstairs balcony.

A little boy standing on a plant decorated balcony.

A young boy looks over a white metal balcony.

Selected words

Boy
Balcony
Toy

Figure 2: Example image with captions and selected words.

Data

The Flickr30k dataset (Young et al., 2014) was used as visual input to the model. The dataset consists of 31,783 images taken from Flickr, annotated with five captions per image (158,915 in total) via crowdsourcing. The images depict everyday activities and scenes. Plummer et al. (2015) expanded the dataset with Flickr30k Entities, by identifying which words in the captions refer to which entities in the images. They provided annotation for 244,035 such coreference chains, and located the entities they referred to in the images, resulting in 275,775 bounding boxes. It should be noted that this data has a high level of complexity, but the captions are not child-directed speech.

Figure 2 gives an example of the data our model was trained on. On the visual side, we simplified the learning problem by excluding any referring expressions that described multiple objects, such as 'plants' and 'pots' in Figure 2. Processing multi-word expressions requires a recurrent neural network and a cross-situational learning model, which is outside the scope of the current work. We therefore simplified the referring expressions to single words. The Flickr30k Entities "Sentences" files containing the annotated captions for each image were searched to find all descriptions for every object ID. From the expressions for every object ID, the most frequent word was chosen as the single word most likely to describe the object in the image. This required that at least two descriptions of the image mentioned the object by the same term, otherwise the object was excluded. The word selection was done after omission of very frequent, irrelevant words such as articles ('a', 'an', 'the'), third-person possessive determiners ('his', 'her', 'their'), the cardinal numbers one through ten, and primary and secondary colors (e.g. 'orange'), including 'silver' and 'gold'. If multiple objects in an image had been labeled with the same word, only one of them was selected (the first one in the loop, not randomly). Finally, images were removed that contained fewer than two objects after preprocessing.

This yielded a total of 86,748 word-object pairs, resulting in a vocabulary of 4,237 unique words. It should be noted that objects paired to the same word could still display great visual variability. The least frequent words (e.g. 'beak' and 'paste') occurred only once, whereas the most frequent word occurred 7,891 times. The five most frequent words were *man* (7,891 times), *shirt* (4,536 times), *woman* (4,378 times), *boy* (1,477 times), and *girl* (1,428 times). The average frequency was 20.47 (*SD* = 172.33), and the median frequency was 2. After preprocessing, 24,670 images remained, of which 1,000 were set aside as validation data, and another 1,000 as test data.

Table 1: Number of objects and baselines per split.

Split	Objects	Listener baseline	Speaker baseline
Train	79,749	0.284	0.091
Test	3,493	0.286	0.089

Table 1 shows the total number of objects in the train and test splits of the data, as well as the baselines for the listener and speaker respectively. The listener baseline is one divided by the average number of objects per scene. The speaker baseline is the majority baseline of the most frequent word. The baselines represent the average accuracy obtainable by chance, which serves as the minimal performance expected of the model. High accuracy is only an indication of good performance if the model performs better than its baseline. Since many words occur only once or twice, there are 80 words in the test set that do not occur in the training set, with a token frequency of 80, and 776 words, with a token frequency of 3413 in the test set, that do occur in the training set. These numbers might suppress test accuracy.

Curiosity Mechanism

In order to measure the effect of active and curious learning, the model which performed curiosity-based object selection was compared to a model that received the next object to learn about randomly. In the first condition, curiosity values were calculated for each object using T&W's curiosity mechanism, and the object with the highest value was chosen to learn about. In the second condition, objects were randomly chosen from the scenes. The main purpose of the speaker part of the model was to produce a word guess as input to the listener so that the model could run without the input provided by the oracle. This way, the model could run (without weight updates) to compute curiosity values and

choose the most interesting object to talk about, before running (with weight updates) to learn about the form-meaning pair with feedback from the oracle.

T&W's curiosity mechanism (see equation (1)) was used to produce the curiosity values, where *i* was the object representation given as input to the speaker, and *o* was the object prediction produced by the listener. The curiosity values were computed element-wise, and the mean of the absolute values of the curiosity vector was taken as the curiosity value for an object in the scene. The object with the highest curiosity value was chosen as the next input for the speaker and target for the listener.

The random and curious conditions were compared on listener loss and accuracy, which indicate the models' ability to choose the appropriate referent of a word form. The loss and accuracy patterns produced over the 40 epochs were plotted to be interpreted as learning curves and compared between conditions.

Results

Figure 3 shows the value of the loss and accuracy of the listener, after each epoch of training. Curious listeners (the blue lines in all plots) show a consistent pattern: after one epoch of training, accuracy on the test set ranged from .49 to .61, far above the baseline of .286. The accuracy on the test set steeply increased in the first few epochs, and kept increasing more slowly, but steadily over later epochs, converging somewhere around epoch 20 with accuracy from .71 to .74. At epoch 40, accuracy ranged from .72 to .75. The exception to this pattern is one particular run, which shows a similar learning trajectory but started and ended with a much lower accuracy, of .32 and .58, respectively. The general pattern is reflected in the plots of the loss on the test data.

On the training data, accuracy also plateaued around epoch 20, with accuracy from .80 to .83 for 19 runs, and only small gains in accuracy until epoch 40, with scores from .83 to .84. Note that the training loss continued to decrease after epoch 20. This indicates the curious listeners started to overfit at that point, fitting to specific characteristics of the training set, that did not translate to accuracy or improvements on the test data. As we saw on the test data, one run shows a different pattern and reaches a maximum of .76 in accuracy on the training data.

The pattern for listeners in the random condition (the orange lines in all plots) is more erratic. After one epoch of training, all random listeners started around or just above the baseline accuracy of .286. Some listeners in this condition barely outperformed the baseline at epoch 40. Others outperformed the baseline, but plateaued after 10-20 epochs, eventually reaching maximum accuracy scores ranging from .39 to .48 on test data. For 6 runs, the accuracy after epoch 1 was around the baseline, but increased steeply until epoch 20, and continued to increase slowly after that. At

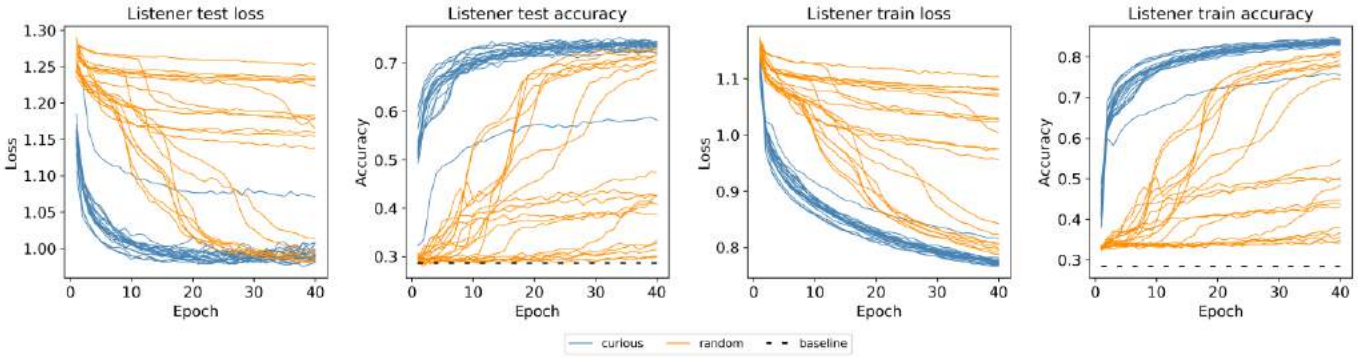


Figure 3. Test and train results of the listener.

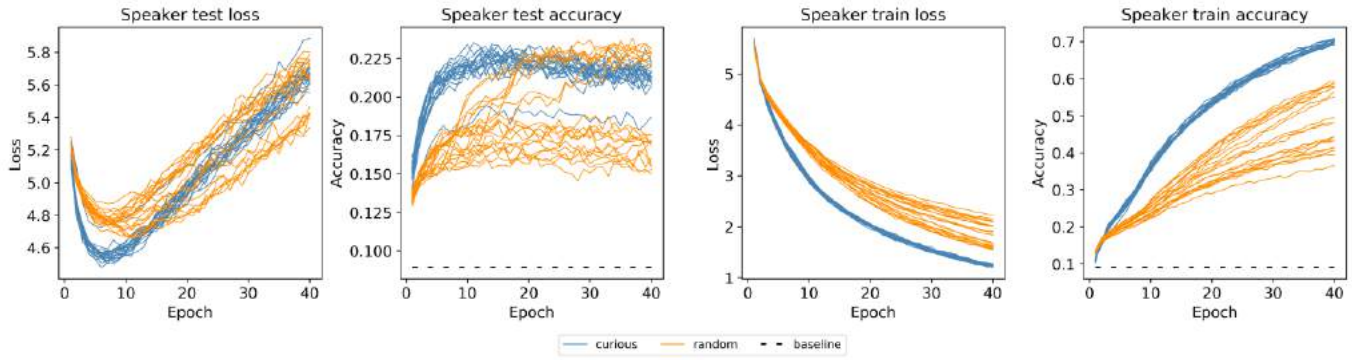


Figure 4. Test and train results of the speaker.

epoch 40, performance of 8 runs is slightly below that of most runs in the curious condition, with test accuracy ranging from .68 to .73, and train accuracy from .78 to .81. The same patterns are reflected in the loss plots.

Test accuracy of speakers trained in the curious condition peaked somewhere between epoch 10 and 25 around .23, with the exception of the one run in which the listener was also less successful, which peaked at epoch 12, with an accuracy of .19. The loss value was lowest around epoch 8. After this epoch, the training loss was still consistently going down, and training accuracy going up. After epoch 8-10, the curious speakers were overfitting rather than learning.

As with the listeners, initially, speakers in the random condition learned more slowly, as is reflected in the lower accuracy between epochs 1 and 20. In all random runs, the speaker outperformed the baseline. However, as was the case with the listeners in this condition, there are large differences between runs. Most runs plateaued relatively quickly, and peaked between .16 and .19, whereas in 8 of the 20 runs, accuracy continued to increase, eventually matching performance of the speakers in the curious condition, with accuracy peaking around .23. Although the training trajectories in the random condition are more discernable than for the curious condition, in all runs, performance on the training data continued to improve until epoch 40. As in the curious condition, all random speakers overfitted.

Discussion

Did curiosity increase the performance of the word learning model compared to the random choice of objects? Yes, the listener test loss decreased faster and the listener test accuracy increased faster in the curious condition than in the random condition. Whereas the curious model converged at a similar point on every run, the random model eventually equaled or approached the curious model on some runs, but learned nothing or was stuck in a local optimum on others.

A pattern that can be discerned is that curiosity, aggregated over the different initializations, performs better from the start and learns faster than random selection. In this experiment, the random initialization of the weights meant that the first objects selected in the curious condition were just as random as those in the random condition. This changed after a few weight updates when the curiosity formula took effect—the difference in performance becoming apparent after a single epoch. This behavior is different from what is typically proposed, as intrinsic motivation is expected to make learning slower initially, but make up for that with increased performance and better generalization in the long run (Oudeyer & Kaplan, 2007).

Another pattern that can be observed is that learning trajectories of curious learners were more similar to each other than those of random learners were. Curiosity seems to provide 'robustness', making learners less prone to being stuck in a local optimum.

The near instant performance advantage of curiosity may be explained by the inherent advantage it has over random selection when dealing with token frequency. Having a good word representation for the corresponding object brings an increase in overall accuracy equivalent to its token frequency. Whereas random selection is prone to select objects with a high token frequency, curious selection can focus on highly frequent word-object pairs first, and ignore them later once their representation is already accurate. Further research could establish whether the selection by the curiosity mechanism matches this strategy.

This would correspond to the notion that language is not a product, but a means for social interaction, where the child's initial interest is to get the message across and language learning follows (Clark, 2016). The intentionality theory of language learning describes how such intrinsically motivated behavior can drive language learning (Bloom, 2000). As of yet, there is no empirical data on what criteria or strategies children use to pick topics to talk about.

Whereas the model was evaluated on the listener performance (comprehension), the speaker's main purpose was to enable the curiosity mechanism, which was used to train the curious model. The high train accuracy of the curious speaker increased the accuracy of the curiosity mechanism, thereby improving the curious listener's comprehension. However, the speaker overfitted in both conditions, and did not generalize well to test data. The speaker test results therefore do not help to understand how improved comprehension leads to improved language production.

We have shown that modeling the language learner as an active solicitor of input, rather than a passive receiver, can lead to different learning outcomes. When objects in the context are selected as a topic according to curiosity, word learning is faster and more robust than when topics are selected at random. Future work may explore the distributional properties of the topics selected by curiosity over the course of the learning process.

References

Bloom, L., Margulis, C., Tinker, E., & Fujita, N. (1996). Early conversations and word learning: Contributions from child and adult. *Child Development*, 67(6), 3154-3175.

Bloom, L. (2000). The Intentionality of Word Learning: How to Learn a Word, Any Word. In Golinkoff, R., Hirsh-Pasek, K., Bloom. L., Smith, L., Woodward, A., Akhtar, N., Tomasello, M., & Hollich, G. (2000), *Becoming a word learner: A debate on lexical acquisition*, 19-50. NY: Oxford University Press.

Chapman, R., Miller, J., MacKenzie, H., & Bedrosian, J. (1981). The development of discourse skills in the second year of life. *Second International Congress for the Study of Child Language*, Vancouver, BC.

Chouinard, M. M., & Clark, E. V. (2003). Adult reformulations of child errors as negative evidence. *Journal of child language*, 30(3), 637-669.

Clark, E. V. (2016). *First language acquisition*. Cambridge University Press.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Mather, E. (2013). Novelty, attention, and challenges for developmental psychology. *Frontiers in psychology*, 4, 491.

Oudeyer, P. Y., & Kaplan, F. (2007). What is intrinsic motivation? A typology of computational approaches. *Frontiers in neurorobotics*, 1, 6.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L. & Lerer, A. (2017). Automatic differentiation in PyTorch. *NIPS 2017 Autodiff Workshop: The Future of Gradient-based Machine Learning Software and Techniques*.

Pathak, D., Agrawal, P., Efros, A. A., & Darrell, T. (2017). Curiosity-driven exploration by self-supervised prediction. *International Conference on Machine Learning (ICML)*.

Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., & Lazebnik, S. (2015). Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *Proceedings of the IEEE international conference on computer vision* (pp. 2641-2649).

Rohrbach, A., Rohrbach, M., Hu, R., Darrell, T., & Schiele, B. (2017). Grounding of textual phrases in images by reconstruction. In *European Conference on Computer Vision* (pp. 817-834). Springer, Cham.

Ryan, R. M., & Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary educational psychology*, 25(1), 54-67.

Schmidhuber, J. (2010). Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Transactions on Autonomous Mental Development*, *2*(3), 230-247.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556v6*.

Twomey, K. E., & Westermann, G. (2018). Curiosity-based learning in infants: a neurocomputational approach. *Developmental Science*, 21(4), e12629.

Young, P., Lai, A., Hodosh, M., & Hockenmaier, J. (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2, 67-78.

The consistency of durative relations

Laura Kelly and Sangeet Khemlani
{laura.kelly.ctr, sangeet.khemlani}@nrl.navy.mil
Navy Center for Applied Research in Artificial Intelligence
US Naval Research Laboratory, Washington, DC 20375 USA

Abstract

Few experiments have examined how people reason about durative relations, e.g., "during". Such relations pose challenges to present theories of reasoning, but many researchers argue that people simulate a mental timeline when they think about sequences of events. A recent theory posits that to mentally simulate durative relations, reasoners do not represent all of the time points across which an event might endure. Instead, they construct discrete tokens that stand in place of the beginnings and endings of those events. The theory predicts that when reasoners need to build multiple simulations to solve a reasoning problem, they should be more prone to error. To test the theory, an experiment provided participants with sets of premises describing durative relations; they assessed whether the sets were consistent or inconsistent. The results of the experiment validated the theory's prediction. We conclude by situating the study in recent work on temporal thinking.

Keywords: events, temporal reasoning, durative relations, mental models, consistency

Introduction

A police officer stopped a driver on the suspicion of drunk driving near Vero Beach, FL. As the officer began to speak to the driver, he noticed an open bottle of Jim Beam on the passenger's seat. The driver explained to the officer that he had not, in fact, been drinking *while* driving – because he only drank when the car was stopped at traffic lights. He was arrested after failing a field sobriety test (Simmons, 2018).

In daily life, people use temporal relations such as "while" and "during" to convey information about events that endure across more than one point in time. Consider the function of the temporal preposition "during" in the following examples:

1a. The car broke down *during* the road trip.
b. Breckinridge graduated *during* the Progressive Era.

The statements each describe a punctate event, i.e., a single point in time (e.g., the breakdown, the graduation), that occurred in the context of a period that extends across multiple time points (e.g., the road trip, the Progressive Era). The sentential connective "while" can yield similar interpretations, as in the examples in (2):

2a. The man slept *while* the neighbors fought.
b. The neighbors fought *while* the man slept.

The examples show how syntax can change the way events are interpreted. For instance, (2a) seems to suggests that the neighbors fought for longer than the man slept, whereas (2b) seems to convey the opposite. Perhaps the two statements are compatible with one another, as in the situation in which the man started sleeping right as the fight began and woke up when the fight ended.

Researchers in artificial intelligence have developed many systems of temporal logic to cope with reasoning about durative events (e.g., Allen, 1983, 1991; Freksa, 1992). Temporal logics often stipulate relations between intervals of time. The logics were designed to describe durative events as they occur in the world – they were not developed to capture how humans think about time. Hence, many temporal logics posit relations that don't map onto prepositions or connectives in English. For instance, Allen's (1983) system includes the following types of relation that connect event A with event B:

```
AAAA
BBBBBBBB
```
A *starts* B.

```
    AAAA
BBBBBBBB
```
A *finishes* B.

```
AAAABBBB
```
A *meets* B.

The repetitions of the letters are used to depict how one event endures across multiple points in time. The descriptions of the relations in natural language can be quite complex, e.g., you might describe the *starts* relation as: "Event A and event B began simultaneously, but event A ended before event B did." Hence, while the relation is primitive in Allen's calculus, it depends on the composition of several different concepts in natural language: beginnings, endings, and the preposition "before." Despite the disparity between language and logic (see Knauff, 1999), researchers have built a wide variety of tools in artificial intelligence designed to explain what kinds of inferences can be drawn from the way relations between intervals interact (for reviews, see Fischer, Gabbay, & Vila, 2005; Goranko, Montanari, & Sciavicco, 2004).

In contrast to the computational analyses of temporal reasoning, few studies have examined how people reason about durative relations such as "while" and "during." Many studies have examined temporal relations such as "before" and "after" (Clark, 1971; Münte, Schiltz, & Kutas, 1998; Zhang et al., 2012), but durative temporal relations appear to be more complex – children comprehend and produce "while" after they understand the meanings of "before" and "after" (Keller-Cohen, 1981; Silva, 1991; Winskel, 2003). Previous work by Schaeken and colleagues investigated how adults reason about "while" (Schaeken, Johnson-Laird, & d'Ydewalle, 1996) using premises of the form *X happened while Y happened.* However, reasoners could draw inferences

from such relations without considering the durative nature of "while", i.e., the problems in Schaeken et al.'s (1996) studies implied that the two events both started and ended at the same time. Nevertheless, their work revealed two central patterns of temporal reasoning: first, reasoners appear to simulate a mental timeline of events when they reason about time (Bonato, Zorzi, & Umiltà, 2012; Casasanto & Boroditzky, 2008). Second, some temporal reasoning problems are easy, and some are difficult: people are more prone to error and they take longer to complete certain temporal reasoning problems (Baguley & Payne, 2000; Schaeken & Johnson-Laird, 2000; Vandierendonck & De Vooght, 1997).

Though no studies have examined how people reason about durations, many have focused on people's ability to estimate the durations of experienced or anticipated events (Zakay & Block, 1997). In typical tasks, people make estimations in minutes and hours or by using more qualitative boundaries. The research has shown that people overestimate short time periods and underestimate longer ones (Lejeune & Wearden, 2009), a robust pattern known as Vierordt's law. Gennari and Wang (2019) showed that these estimation biases are correlated with the relative amount of represented information per timepoint. People "compress" representations to avoid maintaining a representation of all timepoints over which an event transpires (Faber & Gennari, 2015, p. 157). The lesson for researchers interested in temporal reasoning is that some event representations can be compressed into a single timepoint, and reasoners can construe them as punctate events. Other event representations may resist such compression by requiring reasoners to maintain information about durations, i.e., information that spans two or more timepoints. Of course, even punctate events have some duration, but their duration is irrelevant to how people make inferences from them.

One recent account by Khemlani, Harrison, and Trafton (2015a) sought to explain how reasoners construct a mental timeline to represent durative relations such as "while" and "during" by specifying how time representations can be compressed. The account builds on previous theories of temporal reasoning that assume people build mental simulations that consist of discrete tokens to reason about time (Schaeken & Johnson-Laird, 2000; Schaeken et al., 1996). But Khemlani et al.'s account extends beyond previous research to make predictions about how people carry out different temporal reasoning tasks, such as reasoning about what is necessary, reasoning about what is possible, and assessing the consistency of a set of assertions (Khemlani, Lotstein, Trafton, & Johnson-Laird, 2015b).

In this paper, we spell out the central principles of Khemlani et al.'s (2015a) account of durative reasoning and use it to derive predictions about whether certain reasoning problems should be easy or difficult. We describe a preregistered experiment that tested these predictions. We conclude by describing limitations of the study and why durative inferences pose unique challenges for investigators.

Mental models of durative relations

Khemlani et al.'s (2015a) account of durative reasoning is based on the idea that humans build discrete mental simulations of possibilities – mental models – when they reason (Johnson-Laird, 2006; Johnson-Laird, Girotto, & Legrenzi, 2004). The model theory applies to relational reasoning across several different domains (Goodwin & Johnson-Laird, 2005), including reasoning about space (Ragni & Knauff, 2013; Jahn, Knauff, & Johnson-Laird, 2007), time (Schaeken et al., 1996; Schaeken & Johnson-Laird, 2000), consistency (Jahn, Johnson-Laird, & Knauff, 2004; Johnson-Laird, Girotto, & Legrenzi, 2004), and kinematics (Khemlani, Mackiewicz, Bucciarelli, & Johnson-Laird, 2013). The theory rests on three fundamental assumptions:

- **Models are iconic.** Mental models are discrete, iconic representations of possibilities. Iconicity constrains models so that their structure reflects the structure of what they represent (see Peirce, 1931-1958, Vol. 4). In the case of two or more events, models should be structured to reflect the events' chronology, i.e., the way in which those events unfolded. Since models are discrete, they cannot directly represent how long one event took relative to another. The restriction allows reasoners to efficiently compress temporal models to uniformly represent events that endure across vastly different timescales, such as seconds or decades.
- **Intuition vs. deliberation.** Reasoners rely on two primary processes of inference: an intuitive construction process and a deliberative revision process. The intuitive construction process rapidly builds and scans an initial, preferred mental model (Jahn et al., 2007). The process is subject to various heuristics and biases, and so reasoners who engage just the initial process are prone to make systematic errors (Khemlani & Johnson-Laird, 2017). A slower deliberative process can revise the initial models to search for alternative models and counterexamples to validate and correct any conclusions inferred by the intuitive process.
- **More models, more difficulty.** A final assumption of the theory is that each model that a reasoner builds demands cognitive resources to maintain. Hence, reasoners tend to rely on their preferred models most of the time. If a reasoning problem can be solved successfully from the preferred model, it should be easy: reasoners should be faster and their responses should be more accurate. If, however, a problem demands that reasoners engage in deliberation, they should be slower and less accurate.

We illustrate how the three principles apply to temporal reasoning by contrasting how the model theory treats punctate and durative events. Consider the premises in (3):

3. The meeting happened before the sale.
   The sale happened after the conference.
   The meeting happened before the conference.

The durations of the events in (3) are irrelevant, and so the premises can be represented as punctate events. The mental model representing the premises in (3) can be depicted in the following diagram:

meeting conference sale

The diagram shows an arrangement of events in which time moves from left to right. Only one such arrangement is possible for (3). The model is parsimonious; it can be used to infer many different relations that are not made explicit in the premises:

4a. The conference happened after the meeting.
b. The sale happened after the meeting.
c. The conference happened before the sale.

In contrast, the premises in (5) concern a durative relation:

5. The meeting happened during the sale.
The meeting happened before the conference.

The description is consistent with the following model:

$sale_{START}$ $sale_{END}$
meeting conference

The model represents the durative aspect of the sale as two separate tokens (following Khemlani et al., 2015a): one token marks the sale's beginning and the other marks its end. And the premises in (5) are consistent with at least one other model:

$sale_{START}$ $sale_{END}$
meeting conference

Hence, the premises in (3) are consistent with just one model, while the premises in (5) are consistent with multiple models.

In general, the model theory predicts that people should be less accurate when reasoning about descriptions consistent with multiple models than about those consistent with one model. No other theory of reasoning makes an analogous claim (Khemlani, 2018; Knauff, 1999, p. 286 et seq.). We next describe an experiment that tested and corroborated the prediction.

Experiment

To test whether participants make more errors when reasoning about problems that elicit multiple models, our experiment presented them with one- and multiple-model descriptions of events that consisted of premises that described temporal relations. Their task was to evaluate the consistency of the premises by assessing whether all of them can be true at the same time. Previous studies used similar problems, but they examined how participants deductively inferred relations between two specified events (Schaeken et al., 1996). In daily life, reasoners are seldom provided such constraints, and so our experiment used a task that does not provide participants with any restriction on which premises to consider. The approach also has the advantage of using the same question across all problems, and so it uses a uniform task to test participants' durative deductions.

To balance out participants' responses, half the problems were consistent and half were inconsistent. The theory predicts that people should be more accurate in assessing the consistency of one-model problems than multiple-model problems.

Method

Participants. 50 participants completed the experiment for monetary compensation ($2 and a 10% chance of a $10 bonus) through Amazon Mechanical Turk. All of the participants were native English speakers, and all but 6 had taken one or fewer courses in introductory logic. 5 participants were excluded from the analysis, either because of excessive and inappropriate keypresses, or else because the participant produced irrelevant debriefing responses. The analyses reported below are based on the remaining 45 participants (21 female, mean age = 35.0).

Preregistration and data-availability. The predicted effects were pre-registered through the Open Science Framework platform (https://osf.io/q45mw). The same link makes the data from the study available.

Task and design. Participants carried out 16 different problems. Each problem comprised 3 premises that describe how 3 different events relate to one another. They were asked to judge whether the 3 premises could all be true at the same time. Half the problems concerned descriptions that were designed to yield one-model after the first 2 premises and the other half yielded multiple models after the first 2 premises. And half the problems used a 3rd premise that was consistent with the previous premises, while the rest used a 3rd premise that was inconsistent with the previous premises.

The first premise of each problem was of the form: *X happened during Y*. Hence, the following is an example of a problem designed to yield one model:

6a. X happened during Y. Y_{START} X Y_{END}
b. Y happened before Z. Y_{START} X Y_{END} Z
c. X happened before Z. **Y_{START} X Y_{END} Z**

A compressed model of events is provided next to each premise to show how Khemlani et al.'s (2015a) system would update the represention after interpreting new information. The bolded text shows how the final model would look. The problem presents a consistent description of events, since all three premises can be true at the same time.

In contrast, the set of premises in (7):

7a. X happened during Y. Y_{START} X Y_{END}
b. Z happened before X. Z Y_{START} X Y_{END} (i)
Y_{START} Z X Y_{END} (ii)
c. Y happened during Z. NO MODEL POSSIBLE

corresponds to a multiple-model problem, because the 2nd premise is consistent with at least two different situations: one in which Z happened before Y started (i), and another in which Z happened before X and they both happened during Y (ii). But neither of those possibilities are consistent with the third premise, therefore (7) is an inconsistent multiple-model problem.

The sixteen different problems used in the study are provided in the Appendix. The experiment implemented a 2 (problem type: one- vs. multiple-model) x 2 (consistent vs. inconsistent) fully repeated-measures design.

Materials. The temporal terms in each problem were replaced by descriptions of everyday events, e.g., X was replaced with "the meeting" and Y was replaced with "the snowstorm". The materials were drawn from 16 sets of 3 events. Each set was designed to describe events that endure at comparable timescales, so that any event in the set could take place during any other event, e.g.,

The meeting happened during the snowstorm.
The snowstorm happened during the ceremony.
The meeting happened during the ceremony.
The snowstorm happened during the meeting.

and so on. Events that elicit strong punctative interpretations, such as "the sneeze," were not used in the study, as they would yield peculiar and unbelievable descriptions, e.g., "The meeting happened during the sneeze." Likewise, events were chosen so that they did not bear causal relations to one another.

Each of the 16 materials was rotated over the designs for each participant. Therefore, across the experiment as a whole, each of the 16 material sets was applied to each of the 16 problems approximately the same number of times. For any given participant, once the materials were assigned to the problems, the order in which the problems appeared was randomized. The counterbalancing scheme eliminated the possibilities that order effects and carry-over effects could account for participants' responses.

Procedure. Participants interacted with the experiment by registering responses through keyboard presses. For each problem, the participants were asked to consider an initial premise, and then pressed the spacebar to reveal each of the remaining premises on the screen. Previously revealed premises remained on the screen whenever the experiment displayed the next premise. The sequential display sought to encourage participants to read the sentences in the order displayed. Once a participant revealed all three premises, a prompt would appear that said: "Can all three of these sentences be true at the same time?" The 'f' and 'j' keys were used to indicate "yes" and "no" responses, respectively. Before taking part in the experiment proper, they completed an example problem and were shown a schematic of how their fingers should be placed on the keyboard. After completing all 16 problems, the participants were asked four open response debriefing questions, which probed their intuitive definitions of "before" and "during" as well as their reasoning strategies.

Results and discussion

Figure 1 plots the proportion of participants' correct assessments of consistency as a function of whether the premises yielded one model or multiple models, and as a function of whether the problem they carried out was consistent or inconsistent. Participants were more accurate for one-model problems than multiple-model problems (78% vs. 69%; Wilcoxon test, $z = 3.02$, $p = .003$, Cliff's $\delta = .43$). The result corroborated the model theory's central prediction that reasoners should find it easier to reason about one-model problems than multiple-model problems. The difference between participants' accuracies did not reliably differ depending on whether the model was consistent or inconsistent (72% vs. 75%; Wilcoxon test, $z = 1.12$, $p = .27$, Cliff's $\delta = .17$). However, the interaction between the problem type (one- vs. multiple-model) and the consistency of the premises was reliable (Wilcoxon test, $z = 4.03$, $p < .0001$, Cliff's $\delta = .42$). The interaction is evident in Figure 1, which shows that consistent problems had a higher accuracy rate when the premises yielded one-model rather than multiple-models. There was little difference by model quantity for inconsistent problems.

To test whether the type of problem is robust to participant and item random effects, we fit a generalized logistic mixed model (GLMM) regression to the data. The fixed effects were the problem type (one- or multiple-model), the consistency of the problem, and their interaction. The random effects components included intercepts and random slopes for all 3 fixed effects by participant. Intercepts also controlled for the items (paired syntax and material sets) and for the pattern of temporal relations in the three premises, i.e., "during/during/during", "during/before/during", etc. The

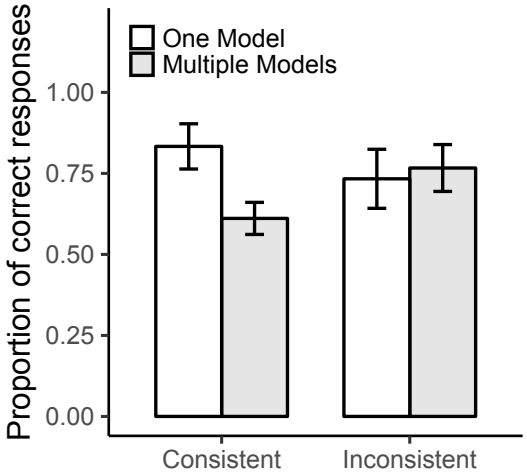


Figure 1. The proportion of correct responses in the experiment as a function of the type of problem (one- or multiple-model) and as a function of whether the premises was consistent or inconsistent. Error bars indicate 95% confidence intervals.

temporal relation pattern was included because a handful of participants reported making judgments based on the pattern alone, and so we treated it as a relevant random effect factor beyond the individual items. The analysis revealed a reliable difference between one- and multiple-model problems when both participants and items were controlled for, $\beta = 2.02$, $z = 3.41$, $p = .0006$, and a reliable interaction between consistency and the problem type ($\beta = 2.35$, $z = 3.21$, $p = .0013$). The GLMM therefore confirmed the nonparametric analyses. However, it further revealed a reliable effect of consistency: participants were more accurate on inconsistent problems than consistent problems ($\beta = 1.82$, $z = 4.34$, $p < .0001$); the model theory did not predict whether there would be an effect of consistency. The results of the regression analysis accordingly support the predictions of the model theory, though the analysis suggests that future studies should examine a broader set of problems to generalize beyond the four in each condition of the present experiment.

General discussion

How do people mentally represent and reason about durative temporal relations, i.e., relations such as "while" and "during"? Such relations describe events that persist across multiple points in time, and many logical frameworks exist that describe ideal temporal reasoning patterns (Fischer et al., 2005; Goranko et al., 2004). But those frameworks do not explain how people represent durations, and so they cannot characterize the mental processes or the strategies people use when reasoning about time. A recent treatment of temporal reasoning explains how people mentally represent durations when they reason. It is based on the idea that people construct mental models, i.e., iconic mental simulations, to draw conclusions from premises or observations (Johnson-Laird, 2006). Models implement a mental timeline, which people use to reason about events (Schaeken et al., 1996). To mentally simulate durative relations, reasoners do not represent all of the time points across which an event might endure. Instead, they construct discrete tokens that stand in place of the beginnings and endings of durative events (Khemlani et al., 2015a).

One way to diagnose the model theory is to investigate how people assess the consistency of durative premises. In principle, the theory should make predictions about people's assessments of consistency. It posits that if they can construct a coherent model of the premises, then those premises are consistent – they can all be true at the same time. If they fail to build a model of the premises, however, then they should consider the premises to be inconsistent (Johnson-Laird et al., 2004). The theory has explained how people reason about the consistency of spatial relations (Jahn et al., 2004), and, given the strong relationship between temporal and spatial metaphors (Casasanto & Boroditsky, 2008; Gentner, 2001), it should also explain how people reason about the consistency of durative relations.

We report an experiment that suggests it does. The experiment presented participants with a description of events as in (8):

8a. The burglar alarm happened during the fire.
b. The siren happened before the burglar alarm.
c. The fire happened during the siren.

The description is inconsistent, and participants capably judged such descriptions to be inconsistent. In contrast, they had difficulty with descriptions of the following form:

9a. The burglar alarm happened during the fire.
b. The siren happened before the burglar alarm.
c. The siren happened during the fire.

Many reasoners incorrectly evaluated the three assertions as inconsistent. What explains the disparity? The model theory suggests that people rapidly constructed a simulation of the first two statements to yield a model akin to the following:

$$\begin{array}{ccc} & \text{fire}_{\text{START}} & \text{fire}_{\text{END}} \\ \text{siren} & \text{burglar alarm} & \end{array}$$

As the theory posits, models cost cognitive resources to maintain, and reasoners are reticent to alter the initial models of the premises they construct. The premise (9c) does not hold in the initial model above, and so reasoners who base their judgments of consistency on it should consider (9) inconsistent. Only reasoners who construct an alternative model of the premises should get the correct answer, as in this model:

$$\begin{array}{ccc} \text{fire}_{\text{START}} & & \text{fire}_{\text{END}} \\ & \text{siren} \quad \text{burglar alarm} & \end{array}$$

In general, the model theory uniquely predicts that participants should find one-model problems easier to reason about than multiple-model problems, and the experiment corroborated the prediction.

One concern of the experiment is that the definition of "during" as an enclosure relation may have been overly restrictive. Consider example (7) above. If people take "during" to refer any two overlapping events instead of an enclosure relation, then the three premises could yield the following model:

$$\begin{array}{ccccc} Z_{\text{START}} & & Z_{\text{END}} & & \\ & Y_{\text{START}} & & X & Y_{\text{END}} \end{array}$$

Yet, a more permissive notion of "during" does not impact the central outcome of the experiment: if a problem is consistent under the restrictive construal of "during", then it is consistent under the permissive construal as well. Nevertheless, reasoners were least accurate on consistent multiple-model problems – a result that corroborates the theory on any construal of "during."

There are at least three limitations of the experiment we report. First, a small set of participants self-reported that they adopted reasoning strategies based on rapidly assessing the relations in the premises. It is not clear to what extent

participants' strategies attenuated or enhanced the difference in performance on one-model and multiple-model problems, but reasoners can spontaneously discover strategies when reasoning about punctate events (Schaeken & Johnson-Laird, 2000), and so future studies should investigate what kinds of strategies participants develop, and how those strategies promote or inhibit the construction of models. Second, the current design did not explore the nature of participants' errors. It could be that participants attempted to consider alternative models of the premises and failed; or it could be that participants chose not to consider alternative models in the first place. Future studies should explore why multiple-model problems yield systematic errors. Finally, only a limited number of problems could be designed for the study given that they described three relations among three events: hence, the study examined only the small number of configurations possible for three events. Future studies should explore an expanded set of problems. Indeed, the language used to describe durational events goes beyond the preposition "during". The connective "while" has a similar meaning, and both words are in the top 200 most frequent words in American English (Davies, 2008). Other words, e.g., "when", can sometimes be used to situate durative events, and the various ways people describe and discuss events, durative and punctate, can provide insight into how people represent and reason about time.

Temporal reasoning is an essential process that underlies how humans conceptualize time (Hoerl & McCormack, 2019; Kelly, Prabhakar, & Khemlani, 2019). Reasoners routinely make inferences about durations in order to carry out time-dependent tasks, such as picking a friend up at the airport. The model theory provides an explanation of the mental representations people build and processes people use when they think and reason about temporal sequences.

Acknowledgments

This research was performed while the first author held an NRC Research Associateship award at the U.S. Naval Research Laboratory. It was also supported by a grant from the Office of Naval Research to the second author. We are grateful to Kalyan Gupta and Kevin Zish at the Knexus Research Corporation for their help in conducting the experiments. Finally, we thank Bill Adams, Gordon Briggs, Monica Bucciarelli, Hillary Harner, Tony Harrison, Laura Hiatt, Phil Johnson-Laird, Joanna Korman, and Greg Trafton for their advice and comments.

References

Allen, J. F. (1983). Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26, 832–843.

Allen, J. F. (1991). Time and time again: The many ways to represent time. *International Journal of Intelligent Systems, 6,* 341-355.

Baguley, T., & Payne, S. J. (2000). Long-term memory for spatial and temporal mental models includes construction processes and model structure. *Quarterly Journal of Experimental Psychology*, 53A, 479-512.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

Bonato, M., Zorzi, M., & Umiltà, C. (2012). When time is space: Evidence for a mental time line. *Neuroscience and Biobehavioral Reviews*, 36.

Clark, E. (1971). On the acquisition of the meaning of *before* and *after*. *Journal of Verbal Learning and Verbal Behavior*, 10, 266-275.

Casasanto, D., & Boroditsky, L. (2008). Time in the mind: Using space to think about time. *Cognition,* 106, 579-593.

Davies, M. (2008) The Corpus of Contemporary American English (COCA): 560 million words, 1990-present. Retrieved from: https://corpus.byu.edu/coca/.

Dierckx, V., Vandierendonck, A., Liefhooge, B., & Christiaens, E. (2004). Plugging a tooth before anaesthetising the patient? The influence of people's beliefs on reasoning about the temporal order of actions. *Thinking & Reasoning*, 10, 371-404.

Faber, M., & Gennari, S. P. (2015). Representing time in language and memory: The role of similarity structure. *Acta Psychologica*, *156*.

Freksa, C. (1992). Temporal reasoning based on semi-intervals. *Artificial Intelligence,* 54, 199-227.

Fischer, M., Gabbay, D., & Vila, L. (2004). *Handbook of Temporal Reasoning in Artificial Intelligence.* Elsevier.

Gentner, D. (2001). Spatial metaphors in temporal reasoning. In M. Gattis (Ed.), Spatial schemas and abstract thought (pp. 203–222). Cambridge, MA: MIT Press.

Goodwin, G.P., & Johnson-Laird, P.N. (2005). Reasoning about relations. *Psychological Review*, 112.

Goranko, V., Montanari, A., & Sciavicco, G. (2004). A road map of interval temporal logics and duration calculi. *Journal of Applied Non-Classical Logics*, 14, 9–54.

Hoerl, C., & McCormack, T. (2019). Thinking in and about time: A dual systems perspective on temporal cognition. Manuscript in press at *Behavioral and Brain Sciences*.

Hothorn, T., Hornik, K., van de Wiel, M. A., Zeileis A. (2008). Implementing a Class of Permutation Tests: The coin Package. Journal of Statistical Software 28(8), 1-23. URL http://www.jstatsoft.org/v28/i08/.

Jahn, G., Johnson-Laird, P. N., & Knauff, M. (2004). Reasoning about consistency with spatial mental models: Hidden and obvious indeterminacy in spatial descriptions. In C. Freksa, M. Knauff, B. Krieg-Brückner, B. Nebel, & T. Barkowsky (Eds.), *Spatial cognition IV: Reasoning, action, interaction* (pp. 165–180). Berlin, Germany: Springer.

Jahn, G., Knauff, M., & Johnson-Laird, P. N. (2007). Preferred mental models in reasoning about spatial relations. *Memory & Cognition*, 35.

Johnson-Laird, P. N. (2006). How we reason. Oxford, England: Oxford University Press.

Johnson-Laird, P. N., Girotto, V., & Legrenzi, P. (2004). Reasoning from inconsistency to consistency. *Psychological Review*, 11, 640.

Keller-Cohen, D. (1981). Elicited imitation in lexical development: evidence from a study of temporal reference. *Journal of Psycholinguistic Research*, 10.

Kelly, L., Prabhakar, J., & Khemlani, S. (2019). Updating and reasoning: Different processes, different models, different functions. Commentary in press at *Behavioral and Brain Sciences*.

Khemlani, S. (2018). Reasoning. In S. Thompson-Schill (Ed.), *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience*. Wiley & Sons.

Khemlani, S., Harrison, A. M., & Trafton, J. G. (2015). Episodes, events, and models. *Frontiers in Human Neuroscience*, 9, 1-13.

Khemlani, S., Lotstein, M., Trafton, J.G., & Johnson-Laird, P. N. (2015). Immediate inferences from quantified assertions. *Quarterly Journal of Experimental Psychology,* 68, 2073–2096.

Khemlani, S. S., Mackiewicz, R., Bucciarelli, M., & Johnson-Laird, P. N. (2013). Kinematic mental simulations in abduction and deduction. *Proceedings of the National Academy of Sciences*, 110, 16766–16771.

Knauff, M. (1999). The cognitive adequacy of Allen's interval calculus for qualitative spatial relation and reasoning. *Spatial Cognition and Computation,* 1, 261-290.

Lejeune, H., & Wearden, J. H. (2009). Vierordt's The Experimental Study of the Time Sense (1868) and its legacy. *European Journal of Cognitive Psychology*, *21*, 941-960.

Peirce, C.S. (1931-1958). *Collected papers of Charles Sanders Peirce. 8 vols.* C. Hartshorne, P. Weiss, and A. Burks, (Eds.). Cambridge, MA: Harvard University Press.

Münte, T., Schiltz, K., & Kutas, M. (1998). When temporal terms belie conceptual order. *Nature*, 395, 71-73.

Ragni, M., & Knauff, M. (2013). A theory and a computational model of spatial reasoning with preferred mental models. *Psychological Review*, 120.

Schaeken, W., & Johnson-Laird, P. N. (2000). Strategies in temporal reasoning. *Thinking & Reasoning*, 6, 193-219.

Schaeken, W., Johnson-Laird, P. N., & d'Ydewalle, G. (1996). Mental models and temporal reasoning. *Cognition*, 60, 205-234.

Simmons, R. (2018). Florida man tells cops he wasn't drinking and driving – he was only drinking Jim Beam at stop signs, traffic lights. *Orlando Sentinel,* Retrieved from: https://www.orlandosentinel.com/opinion/audience/roger-simmons/os-ae-florida-man-drinking-and-driving-20180711-story.html

Silva, M. (1991). Simultaneity in children's narratives: the case of *when, while,* and *as. Journal of Child Language,* 18, 641-62.

Vandierendonck, A., & De Vooght, G. (1997). Working memory constraints on linear reasoning with spatial and temporal contents. *Quarterly Journal of Experimental Psychology,* 50A, 803-820.

Wang, Y., & Gennari, S. P. (2019). How language and event recall can shape memory for time. *Cognitive psychology*, *108*, 1-21.

Winskel, H. (2003). The acquisition of temporal event sequencing: a cross-linguistic study using an elicited imitation task. *First Language*, 23.

Zakay, D., & Block, R. A. (1997). Temporal cognition. *Current Directions in Psychological Science*, *6*, 12-16.

Zheng, Y. et al. (2012). Rearranging the world: Neural network supporting the processing of temporal connectives. *NeuroImage,* 59, 3662-3667.

Appendix. The 16 problems used in the experiment.

Number of models	Consistency	First premise	Second premise	Third premise
One model	Consistent	X happened during Y	Y happened before Z	X happened before Z
One model	Consistent	X happened during Y	Z happened during X	Z happened during Y
One model	Consistent	X happened during Y	Y happened during Z	X happened during Z
One model	Consistent	X happened during Y	Z happened before Y	Z happened before X
Multiple models	Consistent	X happened during Y	X happened during Z	Z happened during Y
Multiple models	Consistent	X happened during Y	Z happened during Y	Z happened during X
Multiple models	Consistent	X happened during Y	Z happened before X	Z happened during Y
Multiple models	Consistent	X happened during Y	Z happened during Y	X happened before Z
One model	Inconsistent	X happened during Y	Y happened before Z	Z happened during X
One model	Inconsistent	X happened during Y	Z happened during X	Z happened before Y
One model	Inconsistent	X happened during Y	Y happened during Z	X happened before Z
One model	Inconsistent	X happened during Y	Z happened before Y	X happened before Z
Multiple models	Inconsistent	X happened during Y	Z happened before X	Y happened during Z
Multiple models	Inconsistent	X happened during Y	X happened during Z	Z happened before Y
Multiple models	Inconsistent	X happened during Y	X happened during Z	Z happened before X
Multiple models	Inconsistent	X happened during Y	X happened before Z	Z happened before Y

Thinking through the implications of neural reuse for the additive factors method

Abstract

One method for uncovering the subprocesses of mental processes is the "Additive Factors Method" (AFM). The AFM uses reaction time data from factorial experiments to infer the presence of separate processing stages. This paper investigates the conceptual status of the AFM. It argues that one of the AFM's underlying assumptions is problematic in light of recent developments in cognitive neuroscience. Discussion begins by laying out the basic logic of the AFM, followed by an analysis of the challenge presented by neural reuse. Following this, implications are analysed and avenues of response considered. Keywords: additive factors method; seriality assumption; anatomical modularity; neural reuse.

Keywords: additive factors method, neural reuse, stage models, seriality assumption

Introduction

A good place to start when trying to understand a complex process or system is to determine its constitutive parts or modules. For example, to figure out how people succeed in visual search during reading, the time between stimulus and response can be broken down into an encoding, feature extraction and identification stage (Resink, 2005; Tovey & Herdman, 2014). The decomposition of the time between the stimulus and response enables discovery of the underlying subprocesses. The stimulus–response time intervals reflect the series of processing stages underlying complex behaviours.

One method for uncovering the subprocesses of mental processes is the "Additive Factors Method" (henceforth AFM) (Townsend & Nozawa 1995; 2001, 2011, 2013; Coltheart, 2011). The AFM uses reaction time data from factorial experiments to infer the presence of separate modules or processing stages. A mental process can be broken down into subprocesses when those subprocesses are 'separately modifiable' – that is, when each of the proposed modules can be modified without effect to the other, and vice versa. For example, to show that two stages A and B are separately modifiable it must be shown that two factors, F and G, affect only either A or B, but not both. In other words, F can affect A and G can affect B, but not the reverse. The result of an AFM analysis is what are called 'stage models'.

This paper investigates the conceptual status of the AFM. It argues that one of the AFM's main assumptions is problematic in light of recent developments in cognitive neuroscience. In particular, the argument is that theories of neural reuse present a challenge to the conceptual link between AFM's 'seriality assumption' and the single processor cases it relies on. Discussion begins by laying out the basic logic of AFM, followed by an analysis of the challenge presented by neural reuse theories. Implications are then analysed and avenues of response considered.

The Additive Factors Method

Factorial experiments are studies in which the effects of two or more variables are investigated by manipulating the presence of each factor across various conditions. In its simplest version (the complete factorial experiment), two factors are studied by comparing the difference each factor has on some measure of performance, such as reaction time. For example, to evaluate the effect of familiarity on pattern recognition, orientation (the rotation of a pattern) can be compared to familiarity (whether subjects are better or worse at recognising the pattern) (Tovey & Herdman, 2014). If orientation has an effect on familiarity, then conditions in which stimuli are presented with different orientations, e.g. 0^0 vs. 90^0, will result in delays in the time required to recognize a pattern.

Factorial experiments form the raw data of the AFM. Factorial data indicates whether two or more factors have either an additive or interaction effect on mean reaction time. Leaving interaction effects to one side for the moment, an additive effect involves two or more factors selectively influencing individual stages of a process. So, for example, if stage A normally takes 10ms and stage B normally takes 15ms and F influences the length of A by 5ms and G influences the length of B by 7ms, then the total duration of time to complete the process that includes stages A and B will be the result of the presence of F and G. The total duration of a process is simply the added the sum of each stage as influenced by each factor.

Factorial experiments supply modifiability information by revealing the selective influence of some factor(s) (Miller et al., 1995). When two or more factors affect the total duration of a process (measured using reaction time), the process can be separated into different modules or processing stages. When patterns of factor effects are observed, a set of hypothesized stages and factor relations that underlie the pattern are proposed. The effects inferred from the factorial experiments are what support inferences about the processing stages, justifying the decomposition of a process or stimulus-response interval into distinct subprocesses.

The Seriality Assumption

One key assumption of the AFM is what Sternberg (2001, 2013) calls the 'seriality' assumption. The seriality assumption says that the AFM can only be applied to processes that are sequentially arranged. For a process to be sequentially arranged, one of two situations must hold, either:

(i) the process must be data-dependent or (ii) a single processor must be responsible for carrying out the process.

In the first set of cases, seriality depends on information being passed along from a previous stage of the process. To use a simple example, heading home from grocery shopping (stage 2) requires first having collected and bagged the groceries at the store (stage 1). In the second set of cases, seriality depends on a process being the result of a single processor. So, for example, if one bakes with two hands multiple steps can be accomplished in parallel, e.g., cracking and whisking eggs; while if one bakes with only one hand, then the process is limited to being complete one task at a time, e.g., cracking each egg individually and then whisking them all together.

How a single process relates to a given processor or set of processors can also vary considerably from case to case. For example, for even a three stage process, there are several types of relations that might hold: (i) a separate processor might carry out each process, (ii) the same processor might carry out every process, or (iii) there be might some combination of the two, where one processor carries out two processes and another processor carries out one process. Figure 1 provides an illustration.

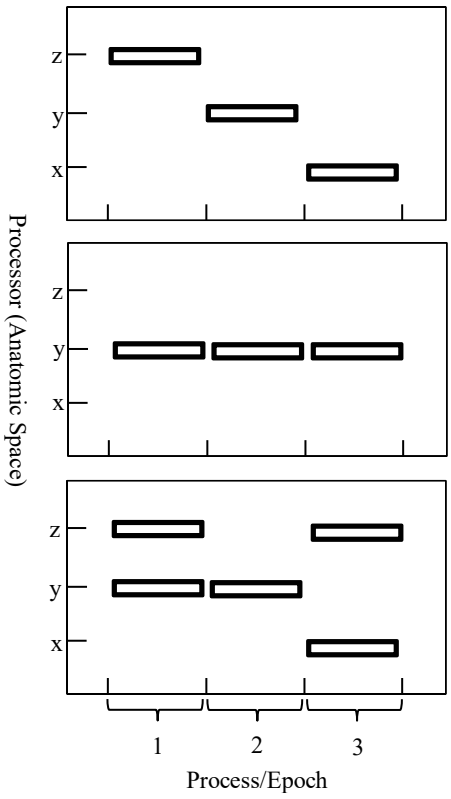


Figure 1. Possible relations between processors and processes for a three stage process.

What is interesting is that the seriality assumption maintains that in at least a subset of cases specialised structures are responsible for carrying out mental processes. That it is possible to find one-to-one mappings between process modules or stages and processing devices. Sternberg (2001), for instance, notes: "Perhaps more surprising is the finding of operations that are partially or wholly sequential when there is no data-dependence…The basis for the sequential structure in such cases may be that the system that carries out the set of operations, possibly the same single processor, is inherently limited in capacity" (original italics, p.735). Not only can processes be sequentially arrange when they involve data dependence but they can also be sequentially arranged when the realising processor has a limited capacity.

However, notice that this is a claim about neural organisation. The single processor view says that neural organisation will take a 'modular' form in certain cases. That in at least some cases mental processes are implemented or realised by dedicated pieces of neural hardware. The claim is that what makes it possible for a given process to be serially arranged is the physical constraints of the realising processor. The view is one of a neural organisation wherein a particular sequential process is carried out by a chainlike structure of connected processing units. To support this claim, for example, Sternberg (2001) appeals to cases of highly specialised anatomical structures, such as the visual cortex of Macaque monkeys. Call this 'anatomical modularity'.

Of course, anatomical modularity is usually considered a 'functional' theory, whereas processing stages are periods of time. Sternberg (2001), for instance, notes: "A stage theory says nothing about the pieces of physical anatomical machinery that carry out the operations in the two stages…information 'transmitted from one stage to the next' does not necessarily go from one place to another; the phrase is unfortunate because it suggests otherwise (p.732). Processor devices that carry out process stages might have functional properties, but the processing stages themselves, at least as informed by the AFM, are neutral with respect to such questions (Kersten, 2016). Nonetheless, there are reasons to see the two views as sides of the same coin. This is because while the processing stages themselves may not have functional properties, they are realised by processors that do, i.e. neurological structures. The point here is simply that the seriality assumption makes specific a claim about the relation between such subprocesses and processors. It does not make a claim about what features those subprocesses have.

It will be worth dwelling on this point as it crucial for the argument to follow. For one might wonder whether 'anatomical modularity' is not better understood as a claim about 'functional' organisation or architecture. If so, then the AFM would be involved in a form of functional decomposition, as it would be set to uncover functional architecture rather than neural organisation.

Crucially, this is not the case. Stage models are set to uncover the subprocesses of mental processes, such as those involved in visual search, understood as epochs or periods of

time. Despite some shared inferential machinery, such as factorial experiments, the target and output of the AFM is importantly different from those methods aimed at providing functional decomposition (see Carruthers, 2006).

A brief case study will help flesh out the point further. Consider Tovey and Herdman's (2014) investigation of visual search during reading. Using the effects of familiarity on change perception via a 2 × 5 × 2 factorial design, Tovey and Herdman examined the effects of orientation (upright vs. inverted, set size (4, 7, 10, 13, 16) and change size (Small vs. Large) across four different experiments. In line with Rensink (2005), they suggested that change perception was divided into three process modules: a pre-processing stage, a feature extraction stage, and an identification stage. They proposed that an interaction between change size and orientation and change size and stimulus quality indicated that change size exerted an effect not only on the feature extraction stage but also on the identification stage of change perception. Figure 2 provides an illustration of the model.

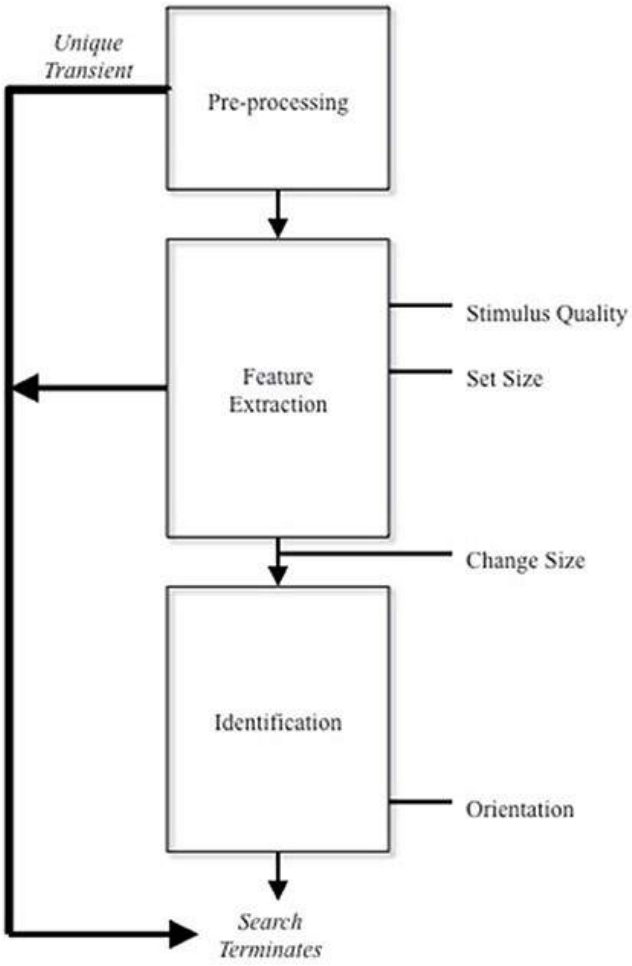


Figure 1. Tovey and Herdman's stage model. Visual search is divided into three stages: a pre-processing stage, a feature extraction stage, and an identification stage.

To explain the effects of change size, Tovey and Herdman proposed a 'gating' mechanism. The gating mechanism redirects information to different stages of the process via detecting changes in size, either by passing the information on to the feature-extraction stage for further processing (assuming the changes are large) or by retaining and verifying the information at the identification stage (assuming the changes are small). The problem is that introduction of a gating mechanism complicates interpretation of Tovey and Herdman's model as a stage model. This is because it introduces functional properties into the model.

Notice that Tovey and Herdman place change size outside of the processing stages, after feature extraction but before identification. This changes the structure of the diagram from a flowchart to a circuit diagram. The arrows no longer represent a succession in time of a series of processes but instead denote the flow of information. The gating mechanism is conceived of as the change size, representing the redirection of information from one stage to another, not only how change size influences time duration.

However, if processing stages are events in time, they need to be strung together end to end, as in a flowchart. If the model represented the effect of change size, it would have to effect the period of time as represented by the box, not the passage or succession of time as represented by the arrows. When represented as a circuit diagram – that is, as describing how processing devices are connected – stage models misleading suggest that the process stages are also processing devices; an interpretation, which, as mentioned, fails to acknowledge the variety of possible relationships that might obtain between process stages and processing devices. Sternberg (2001) frames the point nicely: "It is remarkably easy to slip into a mode of thinking in which stages are processors rather than processes, actors rather than actions; confusion about what a stage might be finds its way into much writing on the subject, even by experts" (p.828). So while Tovey and Herdman's results may be correct, their inclusion of a functional property complicates interpretation of the model (Kersten, 2016).

The ambiguity introduced by the gating mechanism is suggestive of the nature of stage models. For if the AFM is to uncover the subprocesses (understood as epochs of time) of mental processes, then it cannot do so by revealing the functional properties of cognitive systems. If it did, this would blur the distinction between the AFM and other experimental methods.

To illustrate, consider the method of double-dissociation. If two factors F and G are damage to different parts of the brain, and one can show using some measure, such as an EEG, that factor F influences performance on some task A but not task B, while G influences B but not A, then one can infer that the F and G perform different functions. The separate modifiability of tasks A and B by factors F and G on tasks A and B indicate that F and G are have different functions.

Contrast this with the AFM. Whereas double-dissociation uses a direct measure for separate modifiability (the differential activity of different brain regions), the AFM only indirectly tests for separate modifiability via mean-reaction times. It is interested in how a given process can be separated or 'cut' via finding the selective influence of different factors. What this means is that the focus in stage models is on temporal rather than functional properties. Thus, one thing that cannot be meant by the seriality assumption is that what constrains processing stages is functional modularity.

The general point is that anatomical modularity forms more than just a peripheral assumption within the AFM. Indeed, it is what helps, in part, justify inferring the presence of serial processes. If two processes are not data-dependent and yet perform the same function, then it is safe to assume that they are realised by the same processor. That anatomical modularity should underlie part of the seriality assumption is not an insignificant result. The problem is that anatomical modularity is increasingly being called into question.

Neural Reuse and the AFM

Many of the cognitive functions once thought to have dedicated, isolated neural localisations (e.g., Broca's area) are increasingly shown to engage a diverse range of neural units. In a recent meta-review, for example, Anderson and Pessoa (2011) found that 78 different anatomical regions were active in 95 tasks across 9 cognitive domains. Accounts of 'neural reuse' aim to describe how different brain regions often exploit, recycle, or redeploy neural circuitry for various cognitive ends (Hurley 2005; Dehaene & Cohen, 2007; Dehaene, 2009; Anderson 2007, 2011, 2014).

A large swath of evidence now favours neural reuse as a thesis of neural organisation. To spare a long digression, consider a small sampling of some characteristic studies. Glenberg and Kaschak (2002), for instance, show that when asked to make sense judgments about different sentences participants take longer on sentences that run counter to the required action than those that do not. Richardson et al. (2003) show that certain sets of verbs, such as 'hope' and 'respect', activate meaning-specific spatial schemas. Pulvermuller (2005) demonstrates that listening to action words, such as 'lick' or 'pick', activate regions of the primary motor cortex, areas often associated with the actions themselves. Casasanto and Boroditsky (2008) show that people are often unable to ignore irrelevant spatial information when making judgments about duration, but not the converse. That mental representations of time are intimately tied up with perceptions of space. Finally, Casasanto and Dijkstra (2010) demonstrate that there is a bidirectional influence between motor control and autobiographical memory.

Neural reuse theories raise a number of interesting questions about cognition, such as whether a new 'cognitive ontology' needs to be developed (Anderson, 2014). However, for present purposes, the point to note is that neural reuse also raises questions for the second set of cases appealed to by the seriality assumption: namely, that some processes are sequentially arranged in virtue of being realised by single processors. The issue is that if neural reuse is true, then it is unlikely that there will be any single process that has a unique anatomical structure or processor supporting it. Finding a one-to-one mappings between processor and process will prove particularly troublesome if neural regions support multiple operations.

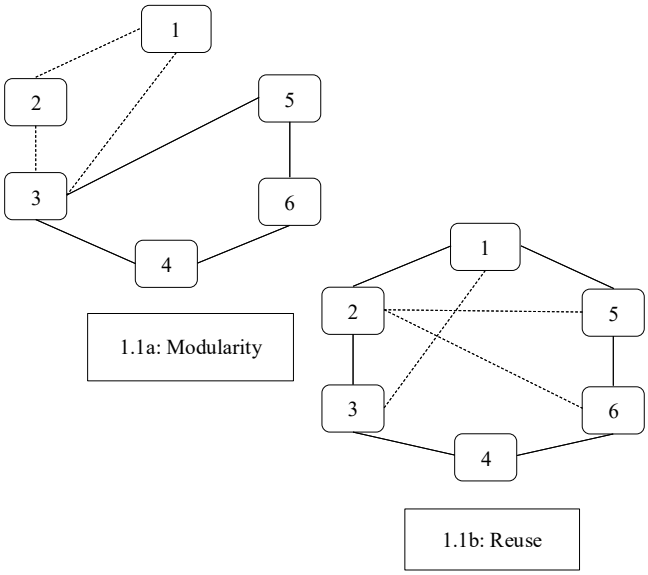


Figure 3.Two possibilities of neural organisation for two cognitive operations.

There seem to be two options. 1.1a represents a modular design, where each cognitive operation has specific dedicated neural circuitry. This is what is required by the second half of the seriality assumption. 1.1b, on the other hand, represents a neural reuse design, where each cognitive process is shared among a number of neural circuits. The AFM requires that 1.1a hold for at least a subset of cases. However, if, as noted, neural reuse is true, then whatever else might be right about the AFM, the single processor cases might not exist. Neural reuse seems to challenge the link between the seriality assumption and one of its inferential bases.

It is important to be clear about this point. For it might be still maintained that the AFM uncovers something about functional organisation. That it would not matter if the same neural hardware were involved in multiple operations because once those operations were fixed the AFM would be set to uncover functional organisation.

But again, once it is appreciated that the target and output of the AFM is not functional models but models of temporal stages it follows that the underlying assumption about processors has to be about neural organisation. For although it is right to point out that neural architecture can stand in complex relations to functional architecture, such as distributed neural regions supporting a functionally modular architecture, such considerations cannot do much work here.

This is because they threaten to undermine the AFM's conceptual standing. If AFM did reveal insights into functional architecture, then its distinctiveness would be undercut, for it would no longer reveal insights into the organisation of cognitive subprocesses understood as temporal sequences.

Another tack would be try to accept the incompatibility of anatomical modularity and neural reuse but nonetheless reject neural reuse on the basis of the wider importance of the AFM. One might argue, in other words, something along the following lines: (1) AFM is essential to psychology; (2) AFM is incompatible with neural reuse; therefore, (3) neural reuse is false.

But there are at least two problems with this type of argument. One is that it assumes that cognitive psychology can operate independently of cognitive neuroscience. That the conceptual autonomy of psychology ensures the survival of the AFM. However, the increasing integration of neurological data into cognitive theorising and modelling makes it unlikely that cognitive psychology will continue to function independently of the findings of cognitive neuroscience in this way (Forstmann et al. 2011; de Hollander et al., 2016). The other is that the argument problematically assumes that the choice facing the proponent of AFM is either/or: that either neural reuse has to be rejected or the AFM does. But no such dichotomy is required. As is argued later, it is possible to endorse a version of the AFM that drops anatomical modularity but which still nonetheless operates in other sets of cases.

Three Options for the AFM

It seems fair to say, then, that neural reuse casts some doubt on the inferential bases of one of the key assumption of the AFM. Given this, three options seem available to the proponent of the AFM. One is to drop the seriality assumption altogether. One might maintain that the AFM can continue on without the seriality assumption. This option seems undesirable insofar as the seriality assumption is part and parcel of the AFM logic. Separate modifiability only makes sense when the processes being investigated are sequentially arranged. Dropping seriality would be tantamount to dropping the method altogether; and scuttling the method altogether seems undesirable given the good deal of fruitful research that has been carried out using the AFM (e.g., Resink, 2005; Tovey & Herdman, 2014).

A second option would be to reform the seriality assumption in light of neural reuse. One might claim, for instance, that serial processes can be the product of distributed neural processors. The problem with this option is that it undermines the inferential link between processor and processing stages. Anatomical modularity forms a key assumption within the AFM. Without it, the AFM would lose its ability to infer a serial ordering. Return, for example, to the baking case, it is only because there is one single processor that the stages are arranged serially. The addition of a second hand opens up the task to being achieved in multiple stages, i.e. in parallel. If multiple processors are admitted, then inferences to processing stages are underdetermined.

But, one might object, it could be that a bunch of miniature 'hands' accomplish the baking task. In other words, that a distributed network of miniature processors performs the task serially, whose actual decomposition is discoverable (at least in part) by the AFM. The problem with this rejoinder is that again misses the key point of stage models, and to lesser extent the baking example. For while it is true that adding more processors speeds up the process, it also makes it impossible to interpret the process as serially ordered. For example, switching to using two hands during the baking (i.e. allowing multiple processors) opens the process up to being completed in parallel. There is nothing that forces the process into being completed in successive stages. Thus, in assuming that a process is realised by distributed set of processors one undermines the ability to interpret that process as serial in the first place. The grounds for inferring seriality rests on the process being carried out by a single processor.

Finally, one might jettison the seriality assumption's commitment to the single processor view, i.e. anatomical modularity. This might preserve what is right about the AFM (i.e. inferring seriality on the basis of data-dependence), while dropping the theoretically suspect part (i.e. reliance on single processor cases). The idea would be to restrict the set of cases under which seriality could be legitimately inferred. That is, whereas previously cases of single processors and data-dependence cases could be used, now only data-dependent cases would be allowed to infer seriality.

For example, a study such as Tovey and Herdman's would not be affected according to the third proposal, because visual search during reading is a data-dependent process. The serially ordering is dependent on each of the previous stages being completed before the next one begins; one cannot, for instance, detect the presence of certain letters before those letters have been registered by the visual system. Tovey and Herdman's study does not rely on the single process cases to work, so it can still be used to infer separate modifiability. However, cases where seriality is inferred because of the supposed presence of a single processor, such as Scarbourgh and Landauer's (1981) study on word repetition effects, would have to be dropped according to this proposal. So, although the removal of anatomical modularity might involve the loss of some of the AFM's methodological punch, as not an insubstantial number of cases involve the assumption (Sternberg, 2001, p.831-2), the method itself would still be preserved in an attenuated form.

Given the spread options, the third proposal seems the most preferable going forward. The first and third options suggest either too high a methodological price or an endorsement of a conceptual tension. Only option three seems to allow the AFM to continue on, though in slightly modified form. On the third proposal, serial stage models can be inferred, but only on the basis of data-dependent cases. The single processor cases underlying the seriality assumption need to

be bracketed, at least until such a time that neural reuse can be thoroughly vetted. This might be a welcomed result for some (Stanford & Gurney, 2011; Sternberg, 2013), but maybe not so much for others (Coltheart, 2011).

So, to summarize, though not a devastating blow, neural reuse does represent a serious challenge to some aspects of the AFM. Insofar as neural reuse presents a challenge to anatomical modularity, and anatomical modularity falls out of the seriality assumption, some of the AFM's conceptual foundations need to be reworked. The methodological implications still need to be worked out, but the conceptual moral seems relatively clear: the seriality assumption can no longer rely on single processor cases. Hopefully, then, in having identified the problem and charted some potential responses, the AFM can be put on surer theoretical footing going forward.

References

Anderson, M. (2007). Massive redeployment, exaptation, and the functional integration of cognitive operations. Synthese 159(3): 329–45.

Anderson, M. (2010). Neural reuse: A fundamental organization principles of the brain. Behavioural and Brain Sciences 33(4), 245-266.

Anderson, M. (2014). After Phrenology: Neural Reuse and the Interactive Brain. Cambridge, MA: MIT Press.

Anderson, M., and Pessoa, L. (2011). Quantifying the diversity of neural activations in individual brain regions. In (eds.) L. Carlson, C. Holscher, & T. Shipley, Proceedings of the 33rd Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society.

Bergeron, V. (2007). Anatomical and functional modularity in cognitive science: Shifting the focus. Philosophical Psychology, 20(2): 175–95.

Carruthers, P. (2006). The Architecture of the Mind: Massive modularity and the flexibility of thought. Clarendon Press/Oxford University Press.

Casasanto, D., and Boroditsky, L. (2008). Time in the mind: Using space to think about time. Cognition, 106, 579-593.

Casasanto, D., and Dijkstar, K. (2010). Motor action and emotional memory. Cognition 115(1), 179-185.

Coltheart, M. (2001). Assumptions and methods in cognitive neuropsychology. In The handbook of cognitive neuropsychology, B. Rapp, ed, 3–21. Psychology Press.

Coltheart, M. (2011). "Methods for modular modelling: additive factors and cognitive neuropsychology." Cognitive Neuropsychology 28, 224–240.

Dehaene, S., and Cohen, L. (2007). Cultural recycling of cortical maps. Neuron 56: 384–98.

Dehaene, S. (2009). Reading in the brain. Viking.

Glenberg, A,, & Kaschak, M. (2002). Grounding language in action. Psychonomic Bulletin & Review, 9, 558-565.

Forstmann, B.U., Wagenmakers, E., Eichele, T. (2011). Reciprocal Relations between Cognitive Neuroscience and Cognitive Models: Opposites Attract? Trends in Cognitive Science, 15(6), 272-279.

Hurley, S. (2005). The shared circuits hypothesis: A unified functional architecture for control, imitation and simulation. In (ed.) Susan Hurley and Nick Chater, Perspectives on imitation: From neuroscience to social science (pp. 76–95). MIT Press.

de Hollander, G., Forstmann, B.U., Brown, S.D. (2016). Different Ways of Linking Behavioural and Neural Data via Computational Cognitive Models. *Biological Psychiatry Cognitive Neuroscience Neuroimaging*, 1(2), 101-109.

Miller, J., van der Ham, F., and Sanders, A. (1995). Overlapping stage models and reaction time additivity: effects of the activation equation. Acta Psychologica (Amst.), 90, 11–28.

Kersten, L. (2016). Processor vs. Processing Accounts of Stage Models: A Cautionary Tale. Frontiers in Psychology, 719, 1-3.

Scarborough, D., and Landauer, T. (1981). Lexical decisions about pairs of adjacent words: A reaction time analysis. Bell Laboratory Technical Memorandum TM91-112215. September.

Stafford, T., & Gurney, K. (2011). Additive factors do not imply discrete processing stages: a worked example using models of the Stroop task. Frontiers in Psychology, 2:287, 1-9.

Sternberg, S. (2001). Discovering mental processing stages: The method of additive factors. In (ed.) S. Sternberg, An invitation to cognitive science (pp. 703-863). Cambridge, MA: MIT Press.

Sternberg, S. (2011). Modular processes in the mind and brain. Cognitive Neuropsychology, 28, 156-208.

Sternberg, S. (2013). The meaning of additive reaction-time effects: some misconceptions. Frontiers in Psychology 744(4), 1-3.

Pulvermuller, F. (2005). Brain mechanisms linking language and action. Nature Reviews Neuroscience, 6, 576-582.

Tovey, M., and Herdmen, C. (2014). Seeing changes: How familiarity alters out perception of change. Visual Cognition, 22(2), 214–238.

Townsend, J., and Nozawa, G. (1995). Spatio-temporal properties of elementary perception: An investigation of parallel, serial and coactive theories. Journal of Mathematical Psychology, 39, 321–360.

Rensink, R. (2000). Seeing, sensing, and scrutinizing. Visual Resolution, 40, 1469–1487.

Rensink, Ronald. (2002). Change detection. Annual Review of Psychology, 53, 245–277.

Rensink, R. (2005). Change blindness. In (ed.) L. Rees and J. Tsotsos, Neurobiology of Attention (pp.76–81). San Diego, CA: Elsevier.

Polysemy and Verb Mutability: Differing Processes of Semantic Adjustment for Verbs and Nouns

Daniel King (dking@u.northwestern.edu)
Department of Psychology
Northwestern University

Dedre Gentner (gentner@u.northwestern.edu)
Department of Psychology
Northwestern University

Abstract

Previous research has found that verbs are more likely to adapt their meaning to the semantic context provided by a noun than the reverse (verb mutability). One possible explanation for this effect is that verbs are more polysemous than nouns, allowing for more sense-selection. We investigated this possibility by testing polysemy as a predictor of semantic adjustment. Our results replicated the verb mutability effect. However, we found no evidence that polysemy predicts meaning adjustment in verbs. Instead, polysemy was found to predict meaning adjustment in nouns, while semantic strain was found to predict meaning adjustment in verbs (but not nouns). This suggests that processes of meaning adjustment may be different for nouns vs verbs.

Keywords: polysemy, mutability, computational linguistics, word2vec, semantics

Introduction

A remarkable aspect of language is that it is both stable enough to reliably convey meaning and flexible enough to accommodate unusual or semantically-strained utterances. For example, the sentence "The hostess galloped to the door" is a bit odd, but we can readily understand it as meaning "The hostess moved rapidly and somewhat gracelessly." While overt metaphorical language has been studied extensively, there is much less work on how people resolve semantically-strained utterances of this type, which may be far more prevalent than traditional "X-is-a-Y" metaphors.

Gentner and France (1988) found that, in paraphrases of simple intransitive sentences of the form *The noun verbed*, participants tended to adjust verb meaning to a greater extent than noun meaning – a phenomenon termed the *verb mutability effect*. *Mutability* can be defined as the degree to which a word's semantic interpretation differs across contexts. The verb mutability effect was found to be strongest when the stimulus sentence was semantically strained – that is, when the noun was incompatible with the paired verb's expected argument, resulting in a nonliteral sentence. For example, one participant paraphrased *The lizard worshipped* as *The reptile stared unblinkingly at the sun*, largely preserving the meaning of the noun *lizard* while shifting the meaning of the verb *worshipped* dramatically.

Little research has examined the processes that drive mutability – that is, how semantic structures are altered during these types of adjustments. In an initial investigation, Gentner and France (1988, Experiment 3b) proposed that verbs are adjusted in a graduated manner, by altering the domain-specific aspects of meaning just as far as is necessary to render a meaningful interpretation – a process they called *minimal subtraction*. We refer to accounts like these as *online adjustment* accounts, as they involve the adjustment of meaning *in situ,* constrained by the context provided by the noun.

Another possibility, however, is that mutability is simply a matter of selecting an appropriate alternate meaning from a word's extant senses. There is evidence suggesting that verbs are more polysemous than nouns across all frequency levels (Gentner, 1981). Thus, higher verb mutability may simply be due to there being more available senses to choose from. We refer to this account as the *sense-selection* view.[1]

Indeed, Gentner and France did not control for polysemy in their original study. We evaluated the polysemy of their stimuli by counting the number of synsets (i.e., senses or meanings) for each word in WordNet 2.1 (Miller, 1995).[2] Our analysis found that the verbs from their study were significantly more polysemous than the nouns, $M_V = 4.13$, $SD_V = 2.17$, $M_N = 2.25$, $SD_N = 1.39$, $t(14) = 2.06$, $p = .03$ – leaving open the possibility that the greater verb mutability they observed was due to the relatively higher polysemy of the stimulus verbs.

Thus, a more precise characterization of the processes underlying these types of semantic adjustments is needed – specifically, the extent to which sense-selection and/or online adjustment drive mutability needs to be better understood.

To investigate this question, we tested polysemy as a predictor of mutability. If polysemy is found to predict mutability, it would provide evidence for the sense-selection account of meaning adjustment. If no such relationship is found, this would instead favor the online adjustment view.

[1] Our descriptions of sense-selection and online adjustment are similar (but not identical) to sense-selection and sense-creation as discussed by Gerrig (1989).

[2] We chose WordNet 2.1 over newer versions due to concerns of a proliferation of synsets in later iterations.

In addition, we seek to understand whether the processes employed vary by word class.

This study follows the paradigm established by Gentner and France. Participants were asked to paraphrase intransitive sentences of varying levels of semantic strain. These sentences were generated by combining 6 nouns and 6 verbs for a total of 36 stimulus sentences (see Figure 1).

For verbs, two expected a human argument (*complain, suffer*), two expected a dynamic artifact object artifact (i.e., a man-made object that functions in some way) as an argument (*pause, fail*), and two expected a static inanimate object as an argument (*dry, burn*). For nouns, two were human (*professor, queen*), two were a dynamic artifact object (*motor, bell*), and two were static inanimate objects (*tree, box*). Combinations in which the noun was incompatible with the verb's expected argument resulted in semantically-strained sentences (e.g., *The box suffered*), while those that were compatible resulted in unstrained sentences (e.g., *The professor complained*).

Half the nouns and verbs used were highly polysemous (at least 10 senses) and half were low in polysemy (1-2 senses; see Figure 1). This resulted in both "balanced" combinations, where the noun/verb polysemy matched—both high (N+/V+) or both low (N-/V-)—and "unbalanced" combinations, where the noun/verb polysemy differed greatly (N+/V- or N-/V+). Thus, across the 36 stimulus sentences, every possible combination of high- and low- polysemy nouns and verbs was realized.

Assessing Meaning Adjustment

A thorny issue in this research is how to quantify meaning adjustment. Gentner and France, using human raters, approached this from three different angles. Across these techniques, they obtained converging evidence for the verb mutability effect; however, each method had drawbacks.

In their *divide and rate* method, raters were asked to divide each paraphrase into *the part that came from the noun* (in the stimulus sentence) and *the part that came from the verb*. They then rated the similarity of each part to the original word. This was problematic for several reasons. It was time-consuming and labor-intensive, and judges often had difficulty deciding how to properly divide the sentence, resulting in a high amount of data loss. Worse, in some cases, some words in the paraphrase were clearly affected by both the original verb and noun, making division impossible. For example, consider the following paraphrase of *The motor complained*: *The badly-functioning engine let out a strange noise from its exhaust*. Here, *badly-functioning* modifies the noun in a context-specific manner (i.e., a motor can function badly but a rock cannot), but it also seems to owe its presence in the paraphrase to the original verb *complained*. The same case can be made for the phrase *from its exhaust.*

Therefore, a way to assess semantic change without dividing paraphrases into noun- and verb-originating components is necessary. Gentner and France employed two such methods: a *double paraphrase* and *retrace* task.[3] While both these methods mirrored the results of the divide-and-rate approach in finding verb mutability, they were similarly labor intensive.

In an attempt to address these issues, we used word2vec (Mikolov et al., 2013) to assess meaning adjustment. This allowed us to quantify semantic change by comparing each paraphrase as a whole to the initial noun and initial verb, without having to divide the paraphrases into components. This provided a hands-off approach that the reduced the time and labor costs of using human judges, as well as data-loss due to low inter-rater agreement. In addition, we hoped to obtain a finer-grained quantification of adjustment than was possible with Gentner and France's methods.

Against these advantages, however, we must ask whether vector-space word embedding models (WEMs) like word2vec can adequately capture human similarity judgments. We next describe these models and discuss issues in using them to assess similarity.

Vector Space Word Embedding Models

WEMs take as their foundation the notion that words are similar or related to the extent that they appear in similar contexts. WEMs are trained on a large corpus and derive a vector representation for each word (typically 100 to 300 dimensions) based on co-occurrence patterns in the corpus. Thus, each word's meaning is represented as a point in an *n*-dimensional vector space. The relatedness between any two words is calculated by taking the cosine of the angle between their two associated vectors, resulting in a score between -1 and 1. Scores closer to 1 indicate high levels of relatedness, and scores closer to 0 indicate low levels of relatedness.

While promising in some areas, the evidence regarding WEMs' ability to approximate human similarity judgments is mixed. Latent Semantic Analysis (LSA; Landauer & Dumais, 1997) has been shown in a number of studies to match human judgments of similarity fairly well in certain contexts (Günther et al., 2016; Landauer & Dumais, 1997; Landauer et al., 1998). In addition, previous work has used it as a measure of semantic change over time (Sagi et al., 2011). Other studies, however, suggest that it fails to approximate human intuition, both in literal similarity judgments (cf., Simmons & Estes, 2006), and in relational similarity tasks (Chen et al., 2017) . That the vectors used in WEMs lack explicit relational structure calls into question whether these problems are fully surmountable.

Perhaps the deepest problem lies in the fact that WEMs do not provide a pure measure of similarity, as associations can

[3] In the double paraphrase task, a new group of participants paraphrased the original paraphrases, and any reoccurrences of the original nouns and verbs were scored. There were higher rates of reoccurrence for nouns, indicating greater meaning preservation in the paraphrase. In the retrace task, a new group of participants were given a set of paraphrases, as well as a list of the original eight nouns or verbs used, and asked to guess which noun or verb they thought had occurred in the stimulus sentence. They showed higher accuracy for nouns, indicating greater meaning preservation.

Figure 1: Stimulus nouns and verbs. Shaded cells indicate combinations that result in strained sentences. Pluses and minuses indicate high or low polysemy, respectively. For example, - / + indicates a low-polysemy noun and high-polysemy verb combination, while + / - indicates a high-polysemy noun and low-polysemy verb combination.

			Human		Dynamic Artifact		Static Inanimate	
			complain	suffer	pause	fail	dry	burn
		# *senses*	2	11	2	13	2	15
Human	professor	1	- / -	- / +	- / -	- / +	- / -	- / +
	queen	10	+ / -	+ / +	+ / -	+ / +	+ / -	+ / +
Dynamic Artifact	motor	2	- / -	- / +	- / -	- / +	- / -	- / +
	bell	10	+ / -	+ / +	+ / -	+ / +	+ / -	+ / +
Static Inanimate	tree	2	- / -	- / +	- / -	- / +	- / -	- / +
	box	10	+ / -	+ / +	+ / -	+ / +	+ / -	+ / +

also influence their scores. For example, the words *cow* and *milk* cooccur frequently, resulting in a high cosine similarity score, despite the obvious fact that a cow is not at all similar to milk.

Thus, we consider the present research to be in part an exploration of WEMs' efficacy in this domain. Future work will involve comparing our word2vec results with human judgments. For now, we will provisionally assume that they can be used as an *approximate* assessment of similarity. We chose word2vec based on evidence that it outperforms other WEMs in approximating human similarity judgments in humans (Pereira, et al., 2016).[4,5]

Method

Participants

112 undergraduates completed the study in person at the lab. One participant was excluded for not being a native speaker of English, one was excluded for providing nonsensical answers to all questions, and two were excluded for failing the catch-trial criteria of repeating a noun and/or verb on both catch trials, for a net of 108 participants.

Materials

6 nouns and 6 verbs were used to generate a total of 36 intransitive sentences. Half of the nouns and half of the verbs used were highly polysemous, and half were low-polysemy. Polysemy was evaluated by counting synsets in WordNet 2.1, excluding any that referred to actual people, places, or events.

The shaded cells in Figure 1 indicate the combinations in which the noun does not satisfy the verb's expected argument, resulting in a semantically-strained sentence (e.g., *The bell suffered*). The unshaded cells indicate those combinations where the noun is compatible with the verb's selectional restrictions, resulting in an unstrained sentence that is literally interpretable (e.g., *The professor complained*).

Procedure

Participants were university students who completed the study on a computer. They saw sentences one at a time and were told to paraphrase each sentence without repeating any of the original content words. They were asked to aim for a plausible interpretation of what the speaker meant, rather than a mechanical, word-by-word translation—e.g., to paraphrase *The slimy senator* as something like *The corrupt politician* rather than *The gooey politician*.

So that each participant saw each noun and verb exactly once, the 36 total stimuli sentences were divided into 6 different assignment factors of 6 sentences. Each assignment factor contained two strained and four unstrained sentences. Sentences were presented in randomized order. In addition, two catch trials were included. The catch trials were simple unstrained sentences designed to test for attention and following directions; the criteria for excluding a subject was repeating a content word in both of the catch trial paraphrases or any obviously nonsensical answers in either.

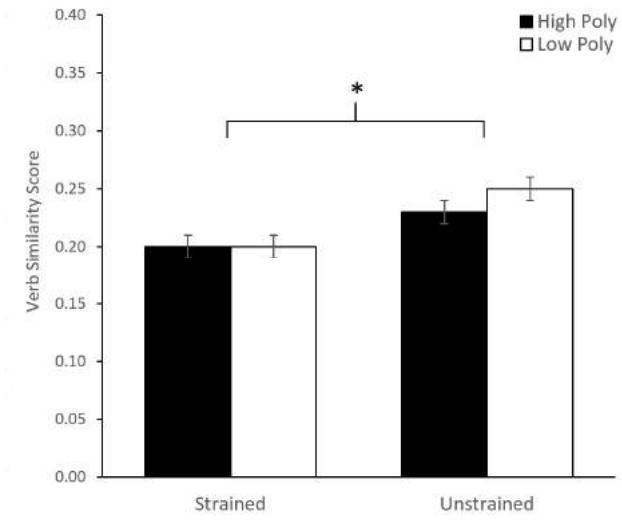
Assessing Semantic Adjustment

For each paraphrase, word2vec was used to obtain two similarity scores, representing the amount of semantic adjustment the initial noun or verb underwent, respectively. The following procedure was used. First, separate normalized vectors were derived for each initial noun and verb. Next, a

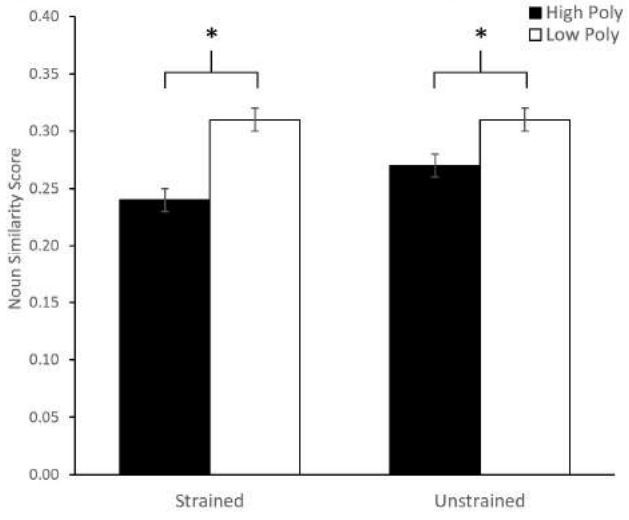
[4] We used pretrained vectors available from Google, which were trained using the CBOW method on part of the Google News corpus (about 100 billion words), available at https://code.google.com/archive/p/word2vec/.

[5] We have also begun analyzing our results following the method outlined by Sagi (in press) for using LSA and other WEMs.

Figure 2. Noun and verb similarity scores. Note that lower scores indicate greater semantic adjustment.



A. Verb similarity scores



B. Noun similarity scores

vector for each paraphrase was generated by averaging its normalized component word vectors.[6] Then the similarity of each paraphrase to the initial noun and to the initial verb was computed by taking the cosine of the angle between the vector representing the initial noun/verb and the entire paraphrase vector.

Any words not found in the corpus were skipped (along with any stop words like *the, a*, etc.). If none of the words in the paraphrase were present in the word2vec dictionary, a null vector was generated. Any paraphrases generating null vectors were discarded (this only occurred twice).

Coding

Certain types of responses were excluded from analysis. First, blatantly noncompliant responses (e.g., paraphrasing *The box dried* as just *fruit*) were excluded. Second, responses that did not constitute a meaningful interpretation of the stimulus noun and verb were excluded as well. This included responses that described the context suggested by the stimulus sentence (e.g., paraphrasing *The tree shivered* as *It is cold outside*), as well as any mechanical, word-by-word paraphrases of strained sentences (e.g., paraphrasing *The box complained* as *The object was frustrated*). For unstrained sentences (which are literally interpretable), a word-by-word paraphrase *is* a meaningful paraphrase (e.g., paraphrasing *The professor paused* as *The teacher stopped*) and therefore were not discarded in these cases.

Two human coders were used. Each coder was presented with the original sentence and its paraphrase and was asked to code each paraphrase as described above, resulting in the exclusion of 137 paraphrases. Cohen's κ was run to determine interrater reliability. There was moderate agreement between the two judges, $\kappa = 0.60$, (95% CI, 0.52 to 0.68), $p < .0001$.

Analysis and Results

The 108 participants generated a total of 648 paraphrases. 137 paraphrases were discarded after coding, leaving a net of 511 paraphrases. All analyses were conducted in R (R Development Core Team, 2008) using the *lmer* package (Bates, Mächler, et al., 2015). Fixed-effect hypothesis tests were conducted using a Satterthwaite approximation for degrees of freedom (Luke, 2017).

First, in order to test Gentner and France's initial finding – that verbs are more mutable than nouns overall – a difference score for each paraphrase was calculated by subtracting verb score from noun score. Next, a linear mixed-effect model was fit, with the difference score as the dependent measure, the intercept (mean) as the fixed effect, and random intercepts for subject and item. The mean of the difference scores was found to differ significantly from 0, $t = 2.99$, $p = .01$, indicating that, on average, verbs (M = 0.23, SD = 0.11) changed their meaning significantly more overall than nouns did (M = 0.28, SD = 0.13; lower similarity scores correspond to greater amounts of semantic adjustment).

Next, to test for effects of semantic strain and polysemy, two additional linear mixed models were fit: one for nouns and one for verbs. In both models, similarity score was the dependent measure, and polysemy (high vs. low), strain (strained vs. unstrained), and the interaction term were included as fixed effects. Both models were initially fit with random slopes and intercepts for subjects and random

[6] These sentence vectors can be viewed as representing the "average meaning" of all the words they contain (Landauer, et al., 1998).

intercepts for items. The random effects structure was then simplified as far as necessary as described in Bates, Kliegl, et al. (2015).

For verbs, the effect of semantic strain was significant, β = -0.18, SE = .08, F = 4.90, p = .03, with verb meaning being adjusted to a greater extent in the strained condition (M = 0.20, SD = 0.09) than in the unstrained condition (M = 0.24, SD = 0.11). There was no significant effect of polysemy, F = 0.98, p = .33, and the interaction was not significant, F = 0.62, p = .43). These results are shown in Figure 2a.

For nouns, there was no significant effect of semantic strain, F = 0.11, p = .74. A significant main effect of polysemy was found, β = -0.19, SE = .06, F = 8.95, p = .01, with high-polysemy nouns (M = 0.25, SD = 0.15) being adjusted to a greater extent than low-polysemy nouns (M = 0.30, SD = 0.11). The interaction was not significant, F = 0.60, p = .44. These results are shown in Figure 2b.

Discussion

There were three objectives in the present research: (1) to replicate Gentner and France's finding that verbs are more mutable than nouns under conditions of semantic strain, using new materials and a new method of assessment; (2) to better understand the processes that drive semantic adjustment; and, (3) to investigate possible noun-verb differences in these processes.

The results regarding objective (1) were as predicted: on average, across conditions, participants adjusted verb meaning significantly more than noun meaning. In addition, verbs (but not nouns) were adjusted to a greater extent in strained contexts than in unstrained contexts (see Figure 2a). Both these results replicate Gentner and France's findings and provide additional evidence for the verb mutability effect: during sentence interpretation, the verb's default meaning is more likely to be adjusted to fit the context provided by the noun, rather than the reverse – especially under semantic strain.

More surprising were the results regarding objectives (2) and (3). Polysemy significantly predicted meaning adjustment in nouns, but not verbs; and semantic strain predicted adjustment in verbs, but not nouns.

This leads to the intriguing conclusion that the processes driving semantic adjustment vary by word class. That polysemy predicted noun adjustment favors the sense-selection view. That it did *not* predict verb adjustment is evidence that their increased mutability is not a matter of having more senses to choose from; rather, online adjustment is taking place. Indeed, a qualitative examination of the paraphrases supports this explanation. For example, nouns were frequently paraphrased as close synonyms (e.g., *tree* as *plant* or *oak*; *box* as *container*), while verbs were frequently adjusted to express meanings that were outside the word's existing set of senses (e.g., paraphrasing *The box complained* as *The container couldn't hold all of its contents*).

What explains the noun polysemy effect?

That semantic strain predicted meaning adjustment in verbs but not nouns is consistent with our prediction that verbs are the locus of change in resolving strained utterances. What is more surprising is the effect of polysemy in driving meaning adjustment for nouns. Why did participants consistently adjust high-polysemy nouns to a greater extent than low-polysemy ones—even in unstrained contexts, where no significant adjustment was necessary? We propose three possible explanations.

1. Higher polysemy allows for more creativity. The first possibility is that higher polysemy granted participants more freedom of interpretation, allowing them to choose a more distant meaning than was available with low-polysemy nouns. We believe this to be unlikely. Examining the paraphrases suggested that, regardless of polysemy (or strain), participants usually attempted to choose a meaning as close to the original noun as possible (unlike with verbs, whose meaning was often changed dramatically). For example, it's not clear that, in substituting *container* for *box* (a high-polysemy noun), one has attempted to adjust meaning further than when one substitutes *oak* for *tree* (a low-polysemy noun). The similarity scores for each pair, however, are 0.12 and 0.80 respectively – a relatively large difference.

2. The results reflect a problem with word2vec. A second possibility is that the observed effects of polysemy are simply an artifact of word2vec and don't reflect actual human intuitions. In all WEMs, the meaning of a word derives from the contexts it appears in. A more polysemous word is likely to appear in a wider variety of contexts than a less polysemous word, rendering it less similar, on average, to any one of those meanings (cf., Beekhuizen et al., 2018).

3. High-polysemy words are less similar to their synonyms than low-polysemy words are. A third possibility is that the relationship between higher polysemy and lower similarity scores reflects a psychologically real pattern: namely, that the more polysemous a word is, the less similar it is, on average, to any one of its synonyms. In this account, polysemy significantly predicted adjustment in nouns because, for a high-polysemy word, any synonym one replaces it with will, on average, be less similar to the original word than when the same is done for a low-polysemy word, despite an equal intention to preserve meaning. That is, the subjective similarity between synonyms of a given word is lower for high-polysemy words than for low-polysemy words. If so, the difference in word2vec scores between *box-container* and *tree-oak* reflects a real psychological difference.

To decide between these latter two possibilities, we conducted a preliminary study with human raters. The results suggest that our WEM results do match human intuitions. We asked raters (N=18) to rate the similarity of eight nouns and verbs (drawn from Gentner & France, 1988) to their closest

synonyms, as determined by a thesaurus (Lewis, 1978). Each base word was paired with three synonyms as well as one antonym as an attention check. Participants rated the similarity between each base word/synonym pair on a scale of 1 to 10, resulting in 865 target ratings. A linear mixed-effects model analysis was conducted, with human similarity rating as the dependent measure, polysemy of the base word as the fixed effect, and random intercepts and slopes for subject and random intercepts for item. A significant negative correlation between polysemy and similarity rating was found, β = -0.20, SE = 0.09, F = 5.63, p = .02.

Thus, we found evidence in favor of our third explanation: on average, the more polysemous a word was, the less similar it was considered to be to its synonyms. In this way, the human findings paralleled our results with word2vec. If this pattern generalizes across other materials, it will be important to understand the reasons for this converging result in humans and in WEMs.

Do Noun and Verb Change Mean the Same Thing?

There are several outstanding issues to acknowledge before concluding. First, an important question is whether semantic distance means the same thing for nouns as it does for verbs. In other words, are the two scales commensurable? WEMs like word2vec are blind to syntactic category and thus employ the same method of generating and comparing vectors for both nouns and verbs (and all word classes). But whether humans judge similarity between nouns on the same dimensions that they do for verbs is unclear.

Similarly, whether polysemy means the same thing for nouns and verbs is also uncertain. It is possible that verb meanings are extended differently than noun meanings, resulting in qualitatively different patterns of relatedness among senses. At present, little work has examined this issue.

Lastly, one might question whether there is a circularity in assessing mutability using word2vec. As with polysemy, if verbs are more mutable than nouns overall, they likely appear in a wider variety of contexts than nouns. Thus, the vectors for any two verbs may, on average, be further apart than is the case for any two noun vectors.

These objections are important and demand further investigation. At the same time, there are striking qualitative differences in the manner of adjustment for nouns versus verbs. As noted earlier, nouns are often paraphrased as close synonyms, whereas verbs are often extended in quite novel ways. This suggests that the verb mutability phenomenon is not simply a difference in similarity scales, but reflects a qualitative difference in processing.

Conclusion

There are three main findings. First, we replicated Gentner and France's (1988) results: verbs changed their meaning more than nouns overall, and did so to a greater extent in a strained context. Second, we found evidence that both sense selection and online adjustment processes drive mutability. Third, we found that these processes differ between nouns and verbs. Semantic adaptation to context appears to be driven by sense-selection for nouns, but by online adjustment for verbs.

We also presented initial evidence that the relationship between polysemy and meaning adjustment may reflect a property of polysemous words; namely, that higher-polysemy words are, on average, perceived as less similar to their synonyms than low-polysemy words are.

Our results invite a number of future research directions. First, the number of items used in this study is small. We are currently testing new word sets. Future work will also involve more systematic testing of specific verb classes to examine how well the results observed here generalize. Second, we plan to compare the WEM results with human judgments of similarity. Our ultimate goal is to provide a clearer characterization of the processes underlying semantic adjustment in nouns and verbs.

Acknowledgements. We thank Sid Horton, Eyal Sagi, and Phil Wolff for their help and advice.

References

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious Mixed Models. *ArXiv:1506.04967 [Stat]*. Retrieved from http://arxiv.org/abs/1506.04967

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

Beekhuizen, B., Milic, S., Armstrong, B. C., & Stevenson, S. (2018). What Company Do Semantically Ambiguous Words Keep? Insights from Distributional Word Vectors. *Proceedings of the 40th Annual Conference of the Cognitive Science Society*, 6.

Chen, D., Peterson, J. C., & Griffiths, T. L. (2017). Evaluating vector-space models of analogy. *ArXiv:1705.04416 [Cs]*. Retrieved from http://arxiv.org/abs/1705.04416

Gentner, D. (1981). Some interesting differences between nouns and verbs. *Cognition and Brain Theory*, *4*, 161–178.

Gentner, D., & France, I. M. (1988). Chapter 14 - The Verb Mutability Effect: Studies of the Combinatorial Semantics of Nouns and Verbs. In S. L. Small, G. W. Cottrell, & M. K. Tanenhaus (Eds.), *Lexical Ambiguity Resolution* (pp. 343–382). https://doi.org/10.1016/B978-0-08-051013-2.50018-5

Gerrig, R. J. (1989). The time course of sense creation. *Memory & Cognition*, *17*(2), 194–207. https://doi.org/10.3758/BF03197069

Günther, F., Dudschig, C., & Kaup, B. (2016). Latent semantic analysis cosines as a cognitive similarity measure: Evidence from priming studies. *Quarterly Journal of Experimental Psychology*, *69*(4), 626–653. https://doi.org/10.1080/17470218.2015.1038280

Landauer, T. K., & Dumais, S. T. (1997). *A Solution to Plato's Problem: The Latent Semantic Analysis Theory of*

Acquisition, Induction, and Representation of Knowledge. 30.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*(2–3), 259–284. https://doi.org/10.1080/01638539809545028

Lewis, N. (1978). *The new Roget's thesaurus*. GP Putnam's Sons.

Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods*, *49*(4), 1494–1502. https://doi.org/10.3758/s13428-016-0809-y

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *ArXiv:1301.3781 [Cs]*. Retrieved from http://arxiv.org/abs/1301.3781

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, *38*(11), 39–41.

Pereira, F., Gershman, S., Ritter, S., & Botvinick, M. (2016). A comparative evaluation of off-the-shelf distributed semantic representations for modelling behavioural data. *Cognitive Neuropsychology*, *33*(3–4), 175–190. https://doi.org/10.1080/02643294.2016.1176907

R Development Core Team. (2008). *R: A Language and Environment for Statistical Computing*. Retrieved from http://www.R-project.org

Sagi, E. (in press). Taming big data: Applying the experimental method to naturalistic data sets. *Behavior Research Methods*. https://doi.org/10.3758/s13428-018-1185-6

Sagi, E., Kaufmann, S., & Clark, B. (2011). Tracing semantic change with Latent Semantic Analysis. In K. Allan & J. A. Robinson (Eds.), *Current Methods in Historical Semantics*. https://doi.org/10.1515/9783110252903.161

Simmons, S., & Estes, Z. (2006). Using Latent Semantic Analysis to Estimate Similarity. *Proceedings of the Cognitive Science Society*, 5.

Getting Insight by Talking to Others – Or Loosing Insight by Talking Too Much?

Sachiko Kiyokawa (kiyokawa.sachiko@b.mbox.nagoya-u.ac.jp)
Graduate School of Education and Human Development, Nagoya University,
Furo-cho, Chikusa-ku, Nagoya, Aichi 4648601 Japan

Zoltán Dienes (dienes@sussex.ac.uk)
School of Psychology, University of Sussex,
Falmer, Brighton BN1 9QH, UK

Abstract

The purpose of the present study was to investigate the effects of addressee of verbalization, self or other, on insight problem solving. Thirty-five participants were assigned to one of the three conditions: toward-self verbalization, toward-other verbalization, or irrelevant verbalization (control). A 3-minute verbalization phase was inserted after 5 minutes of solving the T-puzzle. The participants were asked to write down their thoughts during the first 5 minutes as a record in the toward-self verbalization condition, and as an instruction for other participants in the toward-other verbalization condition. The participants in the control condition were required to write down their concerns. After that, they were asked to engage in the puzzle again for 10 minutes. The results showed a detrimental verbalization effect while allowed a wide range of effects for the self vs other distinction going in either direction. We are using this study as a basis for a pre-registered report.

Keywords: insight problem solving; verbalization; self vs other; metacognitive monitoring

Introduction

Collaboration is ubiquitous in our daily life. Previous studies have shown that collaboration facilitates problem solving (Miyake, 1986; Okada & Simon, 1997; Shirouzu, Miyake, & Masukawa, 2002). Specifically, collaboration is effective in solving problems when novel ideas or perspectives are needed. Insight problems are a typical example of this sort of problem. Since most studies have addressed whether or not collaboration can facilitate problem solving, little is known about why collaboration has facilitative effects on problem solving. By identifying the factors causing the facilitative effects of collaboration on problem solving, we may be able to collaborate with others more effectively.

Diversity of background knowledge is assumed to be one of the most important factors causing the facilitative effects of collaboration on problem solving (Surowiecki, 2005). This hypothesis posits that people can make use of more diverse knowledge when working together. If the diversity of background knowledge was the only factor, then the facilitative effects of collaboration on problem solving would not be obtained when members have the same knowledge bases. Collaboration, however, facilitates problem solving even when there is little diversity in background knowledge. Bahrami et al. (2010) showed that two people working together to detect a subtle visual signal can do better than the best one working alone. Crucially, Kiyokawa (2002) showed that two people working together can solve a problem better than working alone even when one of the members was prohibited to express his/her ideas to solve the problem. Okada and Simon (1997) found that participants were able to reach the solution in a scientific discovery task when working together than when working alone but there was not a significant difference in diversity of hypotheses they entertained. It may be useful to consider factors other than diversity of background knowledge as contributing to the facilitative effects of collaboration on problem solving.

Metacognitive Monitoring in Insight Problem Solving

Facilitation of metacognitive monitoring during collaboration is another potential factor which may be responsible for the facilitative effects of collaboration on problem solving, especially insight problem solving. In other words, collaboration may facilitate insight problem solving because people can monitor their cognitive processes better when working together than when working alone. Previous studies have shown that metacognitive monitoring plays a critical role in problem solving. That is, the more appropriately one can monitor one's cognitive processes, the better one can solve the problem. However, previous studies have also shown that metacognitive monitoring does not always work in problem solving, and in particular not for insight problem solving, when working alone (Metcalfe, 1986; Metcalfe & Wiebe, 1987). This phenomenon is interpreted as implying that the processes underling insight problem solving when working alone is implicit and non-reportable. Indeed, this dysfunction of metacognitive monitoring is assumed to be one of the factors responsible for the difficulty of insight problem solving. Since people cannot know correctly where they are in the problem space when working alone, they cannot choose their moves so as to head in the right direction, and as a result, cannot readily reach the correct solution.

When working together, on the other hand, people have to communicate what they are thinking to their partners. Therefore, they have to change their thinking modes from implicit and non-reportable to explicit and reportable ones during collaboration. These changes in thinking modes when working together may enhance metacognitive monitoring and, as a result, facilitate insight problem solving. Based on this hypothesis, the tendency to think about one's cognitive processes explicitly or verbally specifically in order to

communicate them to a partner may be a key contribution to the facilitative effect of collaboration. Therefore, not only an actual collaborative setting but also a hypothesized one will be enough for people to change their thinking modes and facilitate their metacognitive monitoring and performance in insight problem solving. To reconcile the claim that verbalizing to others is helpful with the previous claim that people have poor metacognition of insight processes, we draw a distinction between verbalizing to oneself as a target and verbalizing to others, which we now consider.

Metacognitive Monitoring and Verbal Overshadowing Effect

There is evidence relevant to our hypothesis in a line of research on the verbal overshadowing effect. These studies have shown that verbalization directed toward oneself disrupts insight problem solving and verbalization directed toward others does not. Schooler, Ohlsson, and Brooks (1993) showed that verbalizing thoughts after each trial when attempting to solve insight problems can disrupt performance [1]. This disruptive effect of verbalization on insight problem solving is called verbal overshadowing effect. The verbal overshadowing effect may originate from a dysfunction of metacognitive monitoring in insight problem solving. The hypothesized process is as follows. People cannot verbalize what they are actually thinking about because they cannot know where they are in the problem space. Therefore, they tend to verbalize what is easy to do so irrespective of their actual cognitive processes. As a consequence, they cannot make use of information other than what they verbalize and so find it hard to reach the correct solution (see also Kiyokawa and Nakazawa, 2006).

Kiyokawa and Nagayama (2008), on the other hand, have found that verbalizing thoughts toward others does not disrupt but rather facilitates insight problem solving. They examined the effects of failure-focused verbalization on insight problem solving using the same task as that used in Kiyokawa and Nakazawa (2006). Participants were randomly assigned to either of the failure-focused verbalization or the irrelevant verbalization (control) conditions. The participants in the failure-focused verbalization condition were asked to write down the ways they thought inappropriate for solving the problem as advice toward other participants. The participants in the control condition were asked to describe in detail what they were studying and interested in. The results revealed that failure-focused verbalization facilitated insight problem solving. The study is consistent with, but was not designed to support the claim, that there is something beneficial about directing one's verbalization to someone else rather than oneself, in acquiring a metacognitive grasp on where one might be in a problem space. Bahrami et al (2012) argue that a key function of meta-cognition is social collaboration; if this is so, engaging socially, or trying to, may facilitate what seems a private process, metacognition. This is the claim we wish to test. The mechanism by which meta-cognition, an apparently private process, is maximally engaged may thus paradoxically rely on social cues.

Purpose of Present Study

The purpose of the present study is to clarify the effects of addressee of verbalization, self or other, on insight problem solving in terms of metacognitive monitoring by examining the verbal overshadowing effects. Our hypothesis is that verbalizing one's thought just as a record disrupts insight problem solving because metacognitive monitoring does not work well, whereas verbalizing one's thought for communicating with other facilitates insight problem solving because it helps metacognitive monitoring.

We will address this question by comparing each solution rate of the puzzle in the two experimental conditions and the control condition. The first experimental condition was the toward-self verbalization condition. In this condition, participants were asked to verbalize reflectively what they were thinking during struggling with the puzzle as a record for themselves. The second experimental condition was the toward-other verbalization condition, in which participants were asked to verbalize their thinking during the previous solving phase as advice for other participants. In the control condition, participants were asked to verbalize not their thinking about their problem solving but their recent concerns irrelevant to solving the puzzle. Thus, the theory that metacognition may not work in a solo setting but does best when engaged in a social context was tested by the following prediction: 1) less participants should solve the puzzle in the toward-self verbalization condition than in the toward-other condition. If in contrast there is just a general overshadowing effect, then there should be little difference shown in the previous contrast but 2) less participants should solve the puzzle in the verbalization conditions than in the control condition. We here investigate these predictions in an exploratory study, that is one that was not pre-registered, in order to have a firm basis for a pre-registered study. We will thus estimate the sort of effect sizes we find that are relevant to the predictions.

[1] Schooler et al. (1993; Exp. 3) found verbalizing reduced percentage of problems solved in 6 minutes by 25% for insight problems and about 5% for non-insight problems, a difference of 20%. Gilhooly, Fioratou & Henretty (2010) tightened up the design and compared percentage of insight with non-insight problems solved in 4 minutes. Crucially, for them verbalizing versus silence did not significantly interact with problem type, $F = 1.63$. Does this fail to replicate Schooler et al? We need a Bayes factor to determine whether the data supported H0 over a reasonable H1. The raw interaction effect would be expected to be 20% (i.e. Schooler et al.'s effect) x 4/6 (correcting for time difference) = 13%. In fact, Gilhooly et al. found a sample overshadowing effect of 4% for insight problems (57 – 53%) and 0% for non-insight (48 vs 48%), i.e. a raw interaction effect of 4% (with $SE = 4\%/\sqrt{1.63} = 3.1\%$). Modelling H1 as a half-normal with $SD = 7\%$, gives a Bayes factor $B_{H(0,7)} = 0.92$, i.e. Gilhooly et al.'s interaction does not count against Schooler et al.

Method

Participants

Thirty-five participants were recruited from the participant pool of the School of Psychology at the University of Sussex. All were required to have UK or EU passports. They received 2 course credits or 3 pounds for taking part in the study. The participants granted their informed consent before participation and the Ethical Committees both of the University of Sussex and Nagoya University approved the study.

Design

We used a between-participants design. The independent variables had three levels: toward-self verbalization, toward-other verbalization, and irrelevant verbalization. The key dependent variable was the proportion of participants who solved the T-puzzle.

Procedures

The participants were randomly assigned to one of the following three conditions: toward-self verbalization, toward-other verbalization, or irrelevant verbalization (control). The experiment took place in a small room with the experimenter present and only one participant at a time. After providing their informed consent to the study, the participants engaged in a practice task for 3 minutes. Before the main task, as a practice task, they were asked to make a rectangle shape (see Figure 1 (b)) using the four pieces presented (see Figure 1 (a)) for three minutes in order to get accustomed to manipulating the pieces they would use in the main task After that, they were asked to solve the main shape puzzle, called the T-puzzle on a display using a mouse for a total of 15 minutes. In the puzzle, they were asked to form a T shape (see Figure 1 (c)) using the same four pieces as the practice task. They were asked to let the experimenter know when they think that they had reached the correct solution. Then the experimenter checked if they have reached the correct solution and if so, the solution phase was terminated at that time. If not, they continued the task.

A 3-minute verbalization phase was inserted after 5 minutes of solving the puzzle. In this phase, the participants were asked to enter their thoughts using a keyboard following the particular instructions in each condition. The first two sentences in the instructions both in the toward-self and other verbalization conditions were the same as those used in Schooler et al. (1993). Those in the toward-self verbalization condition were instructed to write down what they were thinking about in the first 5-minute solution phase, as a record to themselves. The instruction was as follows. "Please write down, in as much detail as possible, everything you can remember about how you have been trying to solve the problem. Give information about your approach, strategies, any solutions you tried, and so on. Write as a record to yourself, like a diary of how you tried to solve the problem in the last five minutes. Remember you are addressing yourself in making these notes; it should feel exactly like talking to yourself. Try to write about 100 words. You can check how many words you have written by looking here. You can take 3 minutes for this writing."

Those in the toward-other verbalization condition were asked to write down their thoughts in the first 5-minute solving phase as advice to other participants. The instruction was as follows. "Please write down, in as much detail as possible, everything you can remember about how you have been trying to solve the problem. Give information about your approach, strategies, any solutions you tried, and so on. Write instructions for other participants on how to solve the problem, based on what you found out in the last five minutes. Remember you are talking to someone else when making these notes; it should feel exactly like a conversation with someone else. Try to write about 100 words. You can check how many words you have written by looking here. You can take 3 minutes for this writing."

Those in the control condition are asked to write down their recent interests as an irrelevant topic to the puzzle. The instruction is as follows. "Please write down, in as much detail as possible, everything you can remember about what you have been interested in. Give information about your interests, hobbies, any things you want to do, and so on. Write about your interests that have nothing to do with the problem you have been trying to solve in the last five minutes. We

(a) Used pieces (b) Solution of the practice (c) Solution of the T-puzzle

Figure 1: Tasks.

want you to take a break from the problem. Remember to write about something other than the puzzle. Try to write about 100 words. You can check how many words you have written by looking here. You can take 3 minutes for this writing."

After the verbalization phase, the participants were required to solve the puzzle again for 10 minutes. They were given a hint to solve the puzzle. Specifically, they were asked to put the pentagon piece not vertically or horizontally but diagonally. This hint was shown to be effective in reaching the correct solution by Suzuki and Hiraki (1998).

After the main task, the participants were asked to fill in a question sheet. The following questions were included in the sheet: (1) "Have you ever tried to solve this puzzle before?" (2) "If you answer "yes" in the first question, when it was it?" (3) "Did you know the correct solution to the puzzle before the experiment?" (4) "To whom did you address your verbal description in the middle of doing the puzzle?" "(5) What was your description about?" (6) "What's your nationality?"

Results

Based on the answers to the questions (1) and (3), we made sure that none of the participants had experienced the T puzzle before the experiment or knew the correct solution. Based on the answers to the question (6), we also made sure that the nationalities of all the participants were UK or EU. A participant in the toward-other condition engaged in the practice task longer than 3 minutes and therefore the data of the participant was excluded from the analyses.

Manipulation Check

We checked whether the participants followed the instructions on the verbalization by the following two ways. First, we examined their recognized addressees based on the question (4) in the post-task questionnaire. Second, we examined what the participants wrote down in the verbalization session. We will report the 95% credibility intervals based on a uniform prior, which are numerically the same as 95% confidence intervals.

Recognized Addressees Table 1 shows frequency of each option the participants selected as their addressees in the post-task questionnaire in each condition. If the participants followed the instruction properly, the participants in the toward-self verbalization condition should have chosen "Self" and those in the toward-other verbalization condition "Other People". Indeed, the selection rate of "Self" was considerably higher in the toward-self verbalization condition than the toward-other verbalization condition with odds ratio, OR = 50.00, 95% CI, [3.88, 643.90].

Table 1: Number of each option selection in each condition.

	Toward-self	Toward-other
Self	10	2
Other people	1	10
Total	11	12

Table 2: Number of participants who used or did not use "You" as a subject or imperative form at least once in their descriptions in each condition.

	Toward-self	Toward-other
Used	1	9
Did not use	10	3
Total	11	12

What the Participants Verbalized We examined the quantity and quality of the participants' verbalization in order to check whether they followed the instructions. First, we compared the number of words among these 3 conditions. Hopefully there would be only minor differences in the sheer quantity of their verbalization, as number of words, among these conditions (Toward-self verbalization: M = 92.8, SD = 12.8; Toward-other verbalization: M = 80.6, SD = 16.3; Control: M = 82.5, SD = 19.5, 95%CI, Toward-self verbalization vs Toward-other verbalization: [-18.43, 42.90], Toward-other verbalization vs Control: [-35.68, 39.42], Toward-self verbalization vs Control: [-23.56, 44.29]).

Second, we examined the subjects and predicates the participants used in their verbalization. Specifically, we counted the number of participants who used "you" as a subject or imperative form at least in their description. If the participants followed the instructions, more participants in the toward-other verbalization condition should use "you" or imperative form than in the toward-self verbalization condition. Indeed, as Table 2 shows, more participants used "You" as a subject or imperative form in their description in the toward-other verbalization condition than the toward-self verbalization condition with odds ratio, OR = 30.00, 95% CI, [2.63, 342.75].

Task Performance

The performance in each condition is shown in Table 3. First, we compared the solution rates between the toward-self and other verbalization conditions in order to test the effects of the addressee of verbalization on insight problem solving. Plausible odds ratios spanned interesting effect sizes around the null value of 1 (OR = 1.90, 95% CI [0.33, 11.01]).

Next, we combined the data in the toward-self verbalization and in the toward-other verbalization conditions into the verbalization condition and compared the solution rates between the verbalization and control (non-verbalization) condition. The result showed that the solution rate could be higher in the control condition than the verbalization condition by a small to a considerable amount (OR = 5.00, 95% CI, [1.03, 24.29]).

In sum, while the evidence allowed a wide range of effects for the self vs other distinction going in either direction, the

Table 3: Performance in each condition.

	Toward-self	Toward-other	Control
Solved	3	5	8
Unsolved	8	7	3
Total	11	12	11

evidence favoured a detrimental verbalization effect rather than an overall positive effect of verbalization. In particular, the crucial theoretical distinction between verbalizing to self vs other had a 95% probability of lying in the interval 1/3 to an effect as high as OR = 11, that is higher than the estimated effect of verbalizing versus non-verbalizing, for which OR = 5 in our sample.

Based on these rough estimates, we can now determine the sort of effect sizes we would expect in a follow up study, for which this report constitutes its pre-registration. Specifically, using the identical procedure as for this exploratory study, for analyzing results we will use an odds ratio of 5 as a roughly predicted effect size for our pre-registered experiment for all effects. The function of this exploratory study was to check the procedure worked smoothly and determine plausible possible effect sizes (Considering the past literature using the same task, Kiyokawa & Nakazawa, 2006, an odds ratio of 3.11 was found for a verbal over-shadowing effect, which is in the same ballpark). We will use this estimate for Bayes factors to make existential claims of whether or not an effect exists. To get evidence for whether an effect does or does not exist, a rough idea of the scale of effect to be detected is needed. Following Dienes and Mclatchie (2018), we will model H1 by setting the *SD* of a half-normal to 5. We will collect participants until the contrast given as prediction 1) at the end of the introduction has a Bayes factor either greater than 3 or less than 1/3.

Discussion

In the present study, we investigated the effects of addressee of verbalization, self or other, on insight problem solving in terms of metacognitive monitoring by examining the verbal overshadowing effects. Our hypothesis was that verbalizing one's thought just as a record disrupts insight problem solving because metacognitive monitoring does not work well, whereas verbalizing one's thought for communicating with other facilitates insight problem solving because it helps metacognitive monitoring. The results showed that the manipulation worked well in terms of participants obeying instructions. Further, the results were consistent with a small to large verbal overshadowing effect on insight problem solving. Crucially, the results allowed a wide range of effects for the self vs other distinction going in either direction. In the following section, we will discuss the necessity of re-examining the verbal overshadowing effect on insight problem by Bayes factors and another possible self vs other difference in metacognitive monitoring.

Verbal Overshadowing Effect Should Be Examined Using a Bayes Factor

There has been a debate between the special-process view and business-as-usual view of insight problem solving. The former posits that insight problem solving processes are implicit, unlike non-insight problem solving. The latter, on the other hand, assumes that the same processes used in non-insight problem solving are involved in insight problem solving. Since the prediction for the verbal overshadowing effect based on the special-process view is different from that based on the business-as-usual view, previous studies have addressed whether or not the verbal overshadowing effect is obtained in order to determine which view is valid (Ball et al., 2015; Fleck & Weisberg, 2004; Gilhooly et al., 2010; Schooler et al., 1993). Specifically, based on the special-process view, verbalization should disrupt only insight problem solving. Based on the business-as-usual view, on the other hand, verbalization should disrupt neither insight nor non-insight problem solving. The evidence from the present study supports the special-process view.

There is a methodological problem on how to determine whether or not the verbal overshadowing effect is obtained. Previous studies concluded that the verbal overshadowing effect was not obtained when there was a non-significant effect of verbalization on problem solving. But non-significance includes both the case where the data were insensitive and where there is evidence for no verbal overshadowing. In contrast, Bayes factors distinguish evidence for no effect relative to a model of the sizes of effect expected, from no evidence at all. In our follow up experiment, we will use Bayes factors.

Self vs Other Differences in Metacognitive Monitoring May Be Emerged Only by Attribution

The present study was motivated by the self vs other difference in metacognitive monitoring when asked to communicate one's thinking processes to others. If the function of metacognition is intrinsically social (Bahrami et al, 2012), the module or mechanism may be best engaged when social cues trigger it. But there may be other factors related to facilitation of metacognitive monitoring in insight problem solving during collaboration. Specifically, the facilitation of metacognitive monitoring may be obtained only by regarding the processes to be monitored as generated by others. (For example, the thinking of others may be regarded with more skepticism than one's own thinking.) Several studies have supported this hypothesis.

Schunn and Klahr (1993) compared performance on an insight-like rule discovery task between self- or other-generated hypothesis conditions. The participants in the self-generated hypothesis condition were asked to generate their own initial hypotheses. The participants in the other-generated condition were given the most frequently generated hypothesis. The results showed that the hypothesis

was investigated more thoroughly in the other-generated condition than in the self-generated condition and that the participants in the other-generated condition terminated with incorrect solutions less than those in the self-generated condition. Kiyokawa, Ueda, and Okada (2004) compared the performance of an insight-like rule discovery task between the self- or other-generated hypothesis conditions. The results showed that the participants in the other-generated hypothesis condition outperformed those in the self-generated hypothesis condition and that the plausibility dropped down after the participants in the other-generated hypothesis condition faced some counterevidence while that increased in the self-generated hypothesis condition.

Kiyokawa, Izawa, and Ueda (2007) investigated effects of swapping between doing and observing a partner or oneself on insight problem solving using the T-puzzle. The results showed that swapping between doing and observing a partner solving the puzzle facilitated insight problem solving, whereas swapping between doing and seeing one's past actions (i.e. within an individual) disrupted problem solving. Kotera et al. (2011) compared the performance of the T puzzle when they observed moves regarded as generated by oneself or by others. The results revealed that observation disrupted insight problem solving if one attributed the observed moves to oneself, but not if one attributed them to another person.

However, all these results may also be explained by our original hypothesis, in the introduction, that it is simply engaging in a social way that maximizes the efficacy of meta-cognition. Our replication of the current study (of which this paper constitutes its pre-registration) until we get evidence for or against the self versus other contrast being effective will help settle the matter: If other is more effective than self, then it may simply be a matter of engaging social cues.

Acknowledgments

This paper, together with https://osf.io/tzf8g/ constitutes the pre-registration for a follow-up experiment we will start in February. We would like to thank Prof. Kazuhiro Ueda at the University of Tokyo for allowing us to use the program. This work was supported by JSPS KAKENHI Grant Number JP17K04350.

References

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G. & Frith, C. D. (2010). Optimally interacting minds. *Science*, *329*, 1081-1085.

Ball, L. J., Marsh, J. E., Litchfield, D., Cook, R. L., & Booth, N. (2015). When distraction helps: Evidence that concurrent articulation and irrelevant speech can facilitate insight problem solving. *Thinking & Reasoning*, *21*, 76–96.

Dienes, Z. & Mclatchie (2018). Four reasons to prefer Bayesian analyses over significance testing. *Psychonomic Bulletin & Review*, *25*, 207-218.

Fleck, J. I., Weisberg, R. W. (2004). The use of verbal protocols as dat,a: An analysis of insight in the candle problem. *Memory & Cognition*, *32*, 990-1006.

Gilhooly, K. J., Fioratou, E., & Henretty, N. (2010). Verbalization and problem solving: Insight and spatial factors. *British Journal of Psychology*, *101*, 81-93.

Hiraki, K. & Suzuki, H. (1998). Dynamic constraint relaxation as a theory of insight. *Cognitive Studies*, *5*, 69-79.

Kiyokawa, S. (2002). The independence structure facilitating representational change: Collaborative problem solving dividing activities into a task level and meta-task level. *Cognitive Studies*, *9*, 450-458.

Kiyokawa, S., Izawa, T., & Ueda, K. (2007). Role exchange between task-doing and observing others as a means of facilitating insight problem solving. *Japanese Journal of Educational Psychology*, *55*, 255-265.

Kiyokawa, S. & Nagayama, Y. (2008). Effects of failure-focused verbalization on insight problem solving. *Proceedings of the third international conference on cognitive science*.

Kiyokawa, S. & Nakazawa, M. (2006). Effects of reflective verbalization on insight problem solving. *Proceedings of the fifth international conference of the cognitive science society* (pp. 137-138). Hillsdale, NJ: Lawrence Erlbaum Associates.

Kiyokawa, S., Ueda, K. & Okada, T. (2004). The effect of other-generated hypotheses on scientific reasoning. *Cognitive Studies*, *11*, 228-238.

Kotera, A., Kiyokawa, S., Ashikaga, J., & Ueda, K. (2011). The role of observation in collaborative problem solving: Attributing the observed actions to self or other influences insight problem solving. *Cognitive Studies*, *18*, 114-126.

Metcalfe, J. (1986). Feeling of knowing in memory and problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *12*, 288-294.

Metcalfe, J. & Wiebe, D. (1987). Intuition in insight and noninsight problem solving. *Memory & Cognition*, *15*, 238-246.

Miyake, N. (1986). Constructive interaction and the iterative process of understanding. *Cognitive Science*, *10*, 151-177.

Okada,T. & Simon, H. A. (1997). Collaborative discovery in a scientific domain. *Cognitive Science*, *21*, 109-146.

Schooler, J. W., Ohlsson, S. & Brooks, K. (1993). Thoughts beyond words: When language overshadows insight. *Journal of Experimental Psychology: General*, *122*, 166-183.

Schunn, C. D., & Klahr, D. (1993). Self- vs. other-generated hypothesis in scientific discovery. *Proceedings of the fifteenth Annual Conference of the Cognitive Science Society* (pp. 900-905). Hillsdale, NJ: Lawrence Erlbaum Associates.

Shirouzu, H., Miyake, N., & Masukawa, H. (2002). Cognitively active externalization for situated reflection, *Cognitive science*, *26*, 469-501.

Surowiecki, J. (2005). *The Wisdom of Crowds*. New York, NY: Knopf Doubleday Publishing Group.

A Bayesian model of memory in a multi-context environment

Dave F. Kleinschmidt (dave.kleinschmidt@rutgers.edu)
Pernille Hemmer (pernille.hemmer@rutgers.edu)
Department of Psychology, Rutgers University, New Brunswick
152 Frelinghuysen Road, Piscataway, NJ 08854

Abstract

In a noisy but structured world, memory can be improved by enhancing limited stimulus-specific memory with statistical information about the context. To do this, people have to learn the statistical structure of their current environment. We present a Sequential Monte Carlo (particle filter) model of how people track the statistical properties of the environment across multiple contexts. This model approximates non-parametric Bayesian clustering of percepts over time, capturing how people impute structure in their perceptual experience in order to more efficiently encode that experience in memory. Each trial is treated as a draw from a context-specific distribution, where the number of contexts is unknown (and potentially infinite). The model maintains a finite set of hypotheses about how the percepts encountered thus far are assigned to contexts, updating these in parallel as each new percept comes in. We apply this model to a recall task where subjects had to recall the position of dots (Robbins, Hemmer, & Tang, 2014). Unbeknownst to subjects, each dot appeared in one of a few pre-defined regions on the screen. Our model captures subjects' ability to learn the inventory of contexts, the statistics of dot positions within each context, and the statistics of transitions between contexts—as reflected in both recall and prediction.

Keywords: Bayesian modeling; memory; learning; belief updating

Introduction

Every cognitive function—perceptual inference, learning, memory, decision making, etc.—takes place in *context*, and understanding these cognitive functions requires understanding the role that the context plays. When cognitive functions are considered in isolation, context can appear to be a source of errors, distraction, or added uncertainty. For example, Roediger and McDermott (1995) induced "false recall" by having subjects study lists of near associates of a word but not the critical word itself. However, when considered ecologically, larger-scale regularities in the environment mean that context can function as a source of additional *information*, reducing the amount of information that must be stored about particular instances. Evidence abounds that people draw on the *context* an item occurred in as an additional source of information (e.g., DuBrow, Rouhani, Niv, & Norman, 2017; Huttenlocher, Hedges, & Duncan, 1991; Orhan & Jacobs, 2013; Schulz, Franklin, & Gershman, 2018; Qian & Aslin, 2014). In this view, so-called "false recall" is really a reflection of the mis-match between the *experimenter's* defined context and the *subject's* inferred context.

However, this raises the question of what *is* a context, and how do people know? For instance, Huttenlocher et al. (1991) found that immediate spatial recall of a location in a circular area is biased towards the average radius of all locations in the experiment. They proposed that memory for an individual item's location is encoded at two levels: the item itself, and the *category* it was assigned to. However, their proposed model does not address what constitutes a category or how subjects decide, and instead simply defines the category based on the long-run statistics of locations encountered in their experiment. However, Robbins et al. (2014) discovered that in a similar task with multiple (implicit) contexts, subjects recall draws on *context*-level statistics, rather than the long-run (experiment-level) statistics.

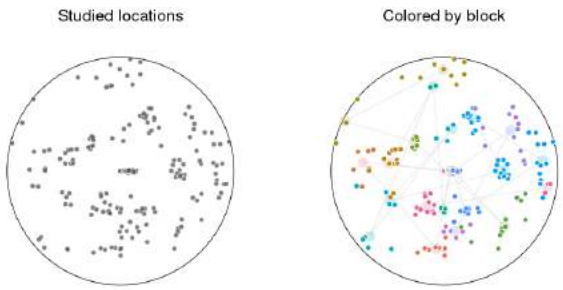


Figure 1: All locations that subject 4 studied (left), color-coded by their block (right), large dots show the average location for each block, and the gray lines show the sequence of blocks

Here, we propose a Bayesian model of learning and memory in multi-context environments, and apply this model to the data from Robbins et al. (2014) human spatial memory experiment. The model treats the problem of identifying latent contexts as a sequential non-parametric clustering problem, where agents must update their beliefs about which context they are in and the properties of that context *online*, with one data point at a time. This model thus captures psychological constraints on the discovery of latent contexts which is not captured by previous Bayesian models.

Data

The data we model is described in detail in Robbins et al. (2014), but we provide a brief summary of the procedure here. In this experiment, 8 participants were asked to record the location of a dot presented in a circle (see Figure 1) and reconstruct that location from memory. Participants were given a cover story in order to keep the task engaging; they were told that the circle was a garden and the dots were moles. In order to save their garden, they had to "catch" the moles by clicking on the locations where they saw them.

After an initial presentation of 20 dots at the center of the circle, dots were presented in blocks (3, 6, 9, or 12 presen-

tations in a cluster), sampled from a multinomial normal distribution with a mean of a given radius and one of three variances (0.01, 0.04, and 0.06 in a unit circle). There was no explicit signal to the subject when one block ended and the next began. The mean angles and radii were informed by Huttenlocher et al. (1991). There were 24 angle measures including the axes, and the measures consisted of the same relative angles in each quadrant. Four different distances measuring out from the center of the circle to the circumference were chosen

Each dot was viewed for one second followed by a combined visual mask and distractor task designed to remove the dot from participants' visual field and introduce uncertainty in the memory process. This mask consisted of a grid of black and white squares; after this mask was removed, an "X" appeared on the screen and participants were asked to report the color of the square (black or white)previously in that location. Data from the distractor task was recorded but not analyzed. After the completion of the distractor task, participants were asked to **recall** the location of the dot from memory by clicking a spot in the circle. After every three trials, participants were asked to make a **prediction** about a future dot location. Prediction trials alternated between prediction for the next trial and prediction for five trials from now. Each block (defined as a cluster of trials at one mean) was followed by a prediction for the expected dot location 10 trials from the current trial. This resulted in a total of 280 trials: 80 prediction trials and 200 recall trials.

Modeling

Our model has three components. First, we model how people infer the assignment of stimuli to contexts as nonparametric Bayesian clustering, approximated sequentially with a particle filter. Second, we model encoding and recall of locations as Bayesian cue combination with a prior from the context (much like Huttenlocher et al., 1991). Third, we model subjects' predictions about future locations via the posterior predictive distribution of the context model.

Context model

We modeled learners inferences about the underlying context on each trial as a sequential Bayesian non-parametric clustering problem. The goal of the learner in this model is to infer the cluster assignment z_i of observation x_i, given the previous observations $x_{1:i-1}$ and their labels $z_{1:i-1}$:

$$p(z_i = j|x_{1:i}, z_{1:i-1}) \propto p(x_i|z_i = j, z_{1:i-1}, x_{1:i-1})p(z_i = j|z_{1:i-1})$$

The sequential prior $p(z_i = j|z_{1:i-1})$ is a "Hibachi Grill Process" (Fox, Sudderth, Jordan, & Willsky, 2011, 2A; Qian & Aslin, 2014), which is like the standard Chinese Restaurant Process (CRP) with an added (constant) probability assigned to the previous state. This corresponds to the following generative model: with probability $0 < \rho < 1$ the previous state is picked, $j = z_{i-1}$, and with probability $1 - \rho$ a component is chosen from a Chinese Restaurant Process with concentration α, which assigns probability to each state proportional to the number of observations assigned to it already,[1] and creates a new state with probability proportional to $\alpha > 0$. We refer to the ρ parameter as the "stickiness" because it controls how likely, a priori, the model is to stick to the same state.

The likelihood $p(x_i|z_i = j, z_{1:i-1}, x_{1:i-1}) = p(x_i|x_{\{k;z_k=j\}})$ is computed by marginalizing over the mean and covariance of a multivariate normal distribution given the data points previously assigned to that cluster and a conjugate Normal-Inverse Wishart prior (Gelman, Carlin, Stern, & Rubin, 2003). This has the advantage that it only requires tracking the sufficient statistics of the previous observations from the cluster (sample mean and covariance), and not the individual observations.

Inference: Sequential Monte Carlo

Instead of a standard batch inference technique, we use an online, Sequential Monte Carlo/particle filter technique. This method approximates the posterior beliefs after $i-1$ observations $p(z_{1:i-1}|x_{1:i-1})$ as a weighted population of K particles, each of which is one possible value of the $i-1$ labels, denoted $z^{(k)}_{1:i-1}$. This population of particles represents an *importance sample* from the posterior. When a new observation x_i comes in, the population moves to target the updated posterior $p(z_{1:i}|x_{1:i})$. There are many algorithms to do this, and the effectiveness of a particular algorithm will depend on the problem. We use the algorithm of Chen and Liu (2000), as described in, Fearnhead (2004): for each particle k, a state assignment is sampled for x_i according to $p(z_i|x_{1:i}, z^{(k)}_{1:i-1})$, and the weight $w^{(k)}_i$ is updated by the ratio of

$$\frac{\sum_j p((z^{(k)}_{1:i-1}, j)|x_{1:i})}{p(z^{(k)}_{1:i-1}|x_{1:i-1})}$$

to ensure that each particle's weight reflects its ability to *predict* the point x_i, rather than just *explain* it. When too much of the total weight for the population (constrained to sum to 1) is captured by a small number of particles (measured by the ratio of the variance of the weights to their mean being greater than 0.5), a new population is resampled (with replacement) and the weights are set to be uniform.

This is for two reasons. First, because we wish to query the model's beliefs about the current context at every point throughout the experiment, an online approximation is much more computationally efficient. A batch algorithm like Gibbs sampling or Hamiltonian Monte Carlo requires one full sweep through the data for each sample, which must be done independently for each data point, so drawing K samples for each of N data points is $O(KN^2)$. A particle filter propagates uncertainty with a fixed population of K particles, updating each particle in parallel as each data point comes in, meaning the complexity is only $O(KN)$. This means it is possible to effectively model longer experiments.

[1] One important difference from a standard CRP is that only non-sticky transitions count for the purposes of sampling new states from the CRP.

Second, an online learning algorithm better approximates *psychological* constraints on learning, and in particular unlike batch MCMC algorithms does not assume that learners can go back and revisit each observation and their decisions about it.[2] This class of models thus provides a possible bridge between computational and algorithmic level approaches to modeling learning and memory (Kleinschmidt, 2018; Sanborn, Griffiths, & Navarro, 2010).

Encoding and recall

The noisy memory trace is modeled as a normal distribution centered at the studied location x with an isometric covariance matrix Σ_x, whose diagonal elements are all equal to σ_x^2, which is a free parameter of the model. This noisy memory trace is combined with a *context prior*, which is approximated by the population of particles. Specifically, each particle k represents one possible assignment of the observations $x_{1:i}$ to clusters $z_{1:i}^{(k)}$. We can thus model each particle's context as the expected mean and covariance matrix for all the points that particle k has assigned to the same cluster as the studied point $z_i^{(k)}$:

$$\mu_c^{(k)}, \Sigma_c^{(k)} = E(\mu, \Sigma)_{p(\mu,\Sigma|x_{1:i}, z_{1:i}^{(k)})}$$

Then the best guess of the studied location under particle k's model of the context is the combination of a normal likelihood (from the noisy trace of the studied item) and a normal prior (from the context), which works out to be the inverse variance-weighted average of the two means:

$$\hat{x}^{(k)} = ({\Sigma_c^{(k)}}^{-1} + \Sigma_x^{-1})^{-1}({\Sigma_c^{(k)}}^{-1}\mu_c^{(k)} + \Sigma_x^{-1}x)$$

Prediction

To model subjects predictions about future locations, we sample 100 locations from the posterior predictive distribution of the population of particles. To sample one predicted location at a n trials in the future, we sample a particle from the population according to their weights, draw a sample of n future states from that particle's Hibachi Grill Process, and then sample one point from the posterior predictive distribution of the resulting cluster. In the case that the predicted cluster is a new cluster, we sample from the prior predictive.

Procedure

To evaluate this model, we simulated the data from Robbins et al. (2014) with a range of parameter values. The concentration parameter α was set to $0.01, 0.1, 1$, or 10, and the stickiness parameter ρ was set to $0.1, 0.5$, or 0.9. The memory noise standard deviation parameter σ_x varied along $0.01, 0.1, 1$, (for a circle with a radius of 1), although only results from $\sigma_x = 0.1$ are presented here. The prior for the cluster parameters was based on the distribution of true block means/covariances. In principle, this could be inferred as well but we leave that enhancement for future work. We ran 10 repetitions with each of the 36 combinations of parameters, all of which used 100 particles for each subject's data.

The particle filter algorithm was implemented in Julia 1.1 (Bezanson, Edelman, Karpinski, & Shah, 2017). The code, simulation results, and Weave.jl (Pastell, 2017) source for this paper is available from osf.io/dqz73/

[2] These approaches also do not *preclude* revising previous decisions, they just do not *require* it.

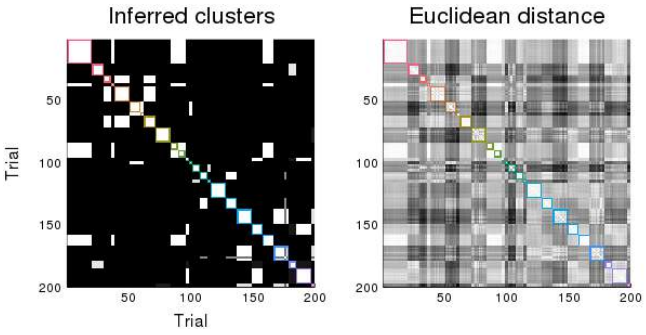


Figure 2: Cluster assignment similarity matrix for clusters inferred by one population of particles from subject 4's studied locations (left), with the true (experimenter-defined) blocks outlined in colors (see Figure 1). The similarity matrix based on the Euclidean distance between each location is shown for comparison (right) and to show that the model groups some similar locations into the same cluster even though they are from different blocks.

Results

Clustering

First, how well does this algorithm do at recovering the underlying cluster structure? This is not a straightforward question to answer: each particle in the population represents a potentially different assignment of observations to clusters, and the cluster indices used in one particle might not align with those in another particle. To get around this we look at the assignment similarity matrix, which is an $N \times N$ matrix, where element (i, j) is the probability that trials i and j are assigned to the same cluster. This probability is calculated by averaging across all particles in the population according to their weight.

Figure 2 shows the assignment similarity matrix for one subject, based on a 100-particle filter with $\alpha = 0.01, \rho = 0.9$ (left) with the true, experimenter-defined block structure is outlined in the colors from Figure 1, and the pairwise Euclidean distance between the locations for comparison (right). This example shows a number of important features of the model's inferences about the underlying changes in context. First, relative to the experimenter-defined blocks, the model occasionally undersegments, grouping adjacent blocks together into a single context. Second, the model also sometimes infers that it has *returned* to a previous context, instead of creating a new context when it infers that the block has

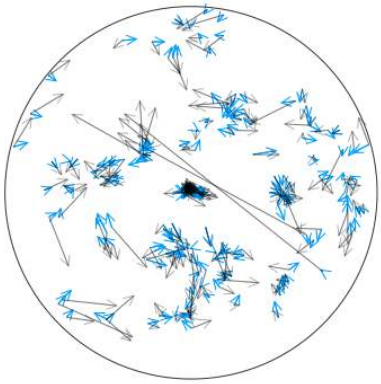


Figure 3: Subject 4's recalled locations (gray arrows, pointing from studied to recalled location) compared with model simulation (blue arrows; $\alpha = 0.01, \rho = 0.9, \sigma_x = 0.1$)

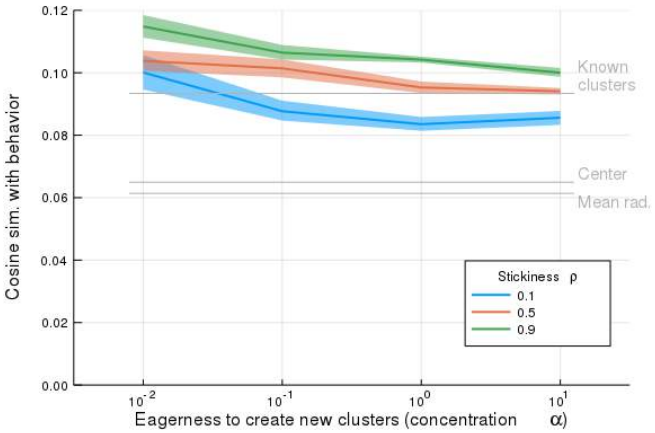


Figure 4: Mean cosine-similarity of model predicted and actual recall deviations across parameter values (ribbons show 95% bootstrapped CIs over model runs). Gray lines show baselines: always deviate toward center, average radius, and center of true clusters

changed. This can be seen from the off-(block)-diagonal entries in the assignment similarity matrix (Figure 2, left). As the Euclidean similarity matrix (Figure 2, right) shows, this tends to happen when the points in two blocks are close together. Third, because of the online nature of the model, it maintains relatively less uncertainty about the clustering of early trials. Note though that Figure 2 shows the beliefs of the model at the *end* of the experiment, which reflect the totality of the locations it has encountered.

Recall

Next, we assess how well the inferred contexts can predict recall. Figure 3 shows one subject's actual deviations from studied to recalled locations (gray arrows) versus the model's predicted deviations (blue arrows). To quantify goodness of fit, we use the cosine similarity of the model's and subject's recall deviation (i.e., blue and black arrows in Figure 3), which ranges from 1 (deviations perfectly aligned) to -1 (deviations in opposite directions), with 0 corresponding to orthogonal deviations. We chose this metric because it is less sensitive to large outlier responses than mean-squared error, and because approximations of the likelihood of a subject's response given the model is highly sensitive to free parameters and difficult to reliably estimate. Moreover, the baseline models we compare against also do not have straightforward likelihood models, but they *do* make straightforward predictions about the directions of recall deviations.

Figure 4 shows the cosine similarity with of all subjects' responses with the multi-context Bayesian model. The ribbons show the 95% bootstrapped confidence intervals over model runs, which indicate that the approximate inference strategy leads to reasonably consistent inferences for a given set of parameters. At all parameter settings, the model performs better than chance, predicting subjects' recall deviation directions at a cosine similarity of around 0.1 (relative to a chance level of 0). The model performs best for high ρ stickiness and low α concentration.

We also compare the model's performance against three baselines. First, we compare it against a "known clusters" model, which uses the true (experimenter defined) clusters with the same Bayesian cue combination model of encoding and recall. Second, we compare it to two baselines based on previous literature on similar memory tasks (Huttenlocher et al., 1991): one that always biases recall towards the center (the average location of all trials), and one that biases recall towards the mean radius.

First, at the whole range of parameters explored, the multi-context model performs better than the center- or mean-radius-biased baselines. Second, except for low stickiness $\rho = 0.1$, our model provides a better fit to human behavior than the "known clusters" baseline, which differs from our model only in that the true cluster labels are provided for each data point, rather than being inferred. This suggests that, at least according to the cosine similarity metric, our context-inference model better captures how people combine information about the current context during recall than the "ground truth" clusters.

However, an important caveat is that there is substantial variability across *subjects*. The cosine similarity for $\alpha = 0.01, \rho = 0.9$ has a 95% boostrapped CI across subjects of $[0.05, 0.17]$, which while significantly better than chance is not significantly better than the baseline models, even when taking into account the substantial variability in the cosine similarity for the baseline models themselves. With only 8 subjects in this dataset it is unclear how well the model's performance will generalize to other datasets, and future work with better-powered designs is required.

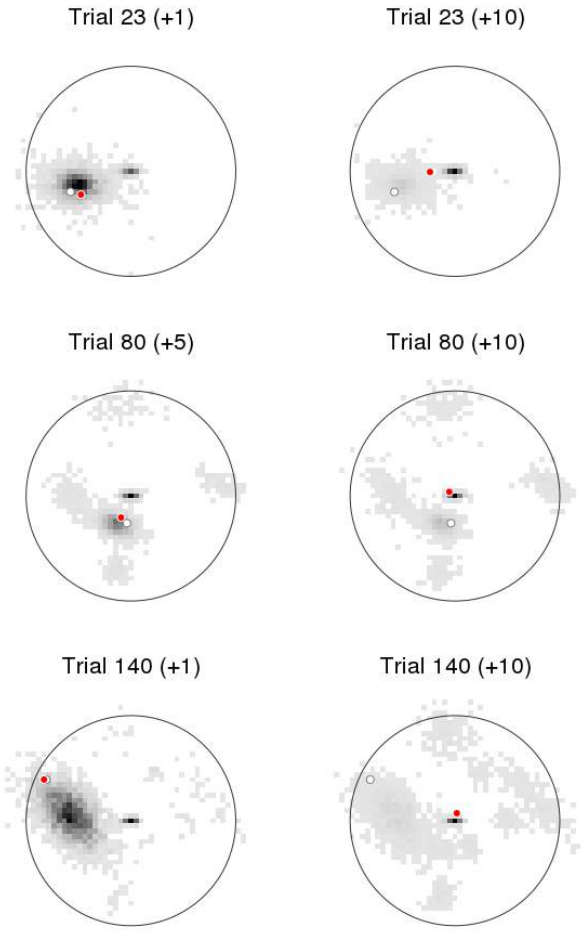


Figure 5: Subject 7's (red points) and model's (gray regions) predictions about upcoming locations at various points throughout the experiment and various prediction horizons. The white points show the last recalled location.

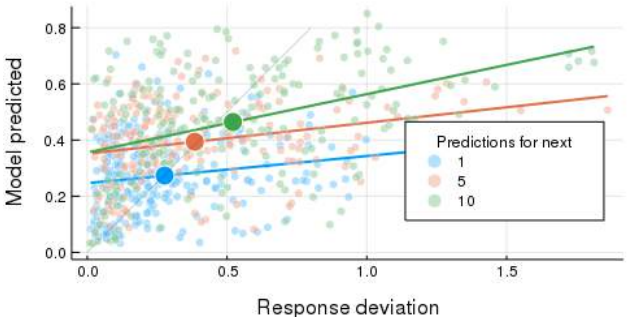


Figure 6: Model predicted ($\alpha = 0.01, \rho = 0.9$) and actual deviations from last studied point for prediction task. Small points show deviations of predictions for each trial, and large points show average deviations for each lag (1, 5, or 10 trials).

Prediction

Subjects also, every three recall trials, predicted the location where points would appear in 1, 5, or 10 trials in the future. This is a more explicit probe of what subjects know about the cluster structure than the recall task. Figure 5 shows six examples of how the model's prediction about upcoming locations capture subjects' behavior. For +1 trial predictions, the model's distribution of predicted locations primarily reflects its beliefs about the *current* cluster (as reflected by the higher density of predictions near the white studied point), because of the "sticky" Hibachi Grill Process prior on states. At +10 trials, the model is much more likely to predict the center cluster, which recurs frequently throughout the experiment (see also Figure 2). Likewise, subjects also have picked up on this pattern and are more likely to predict locations close to the center on +10 prediction trials.

Our model also captures how the average distance from the last studied point increases as subjects are asked to predict the location of points +1, +5, and +10 trials into the future (Figure 6, large points). Moreover, the model also captures variation *within* these delay levels: after removing the effect of delay level by centering, the model's and subjects' prediction deviations are correlated at $\rho = 0.31$ (95% bootstrapped CI: $[0.25, 0.38]$, and significant at $p = 0.014$ in a mixed model with random intercepts and slopes by subject).

Discussion

We have demonstrated that human recall and prediction in a multi-context spatial memory task can be modeled by a Bayesian model that infers the latent contexts via nonparametric clustering. This model updates its beliefs *online*, one observation at a time, with Sequential Monte Carlo. Exploring a range of parameters for the state transition prior, we found that subjects recall behavior is best captured with high "stickiness" (prior probability of remaining in the same cluster) and low concentration (prior probability of creating a new cluster). Together, this suggests that people expect—

until they receive evidence to the contrary—that contexts will continue for a number of trials, and that old contexts will return in the future.

While we treated these parameters as free when fitting our model, this was merely a simplifying assumption that we made to make the model easier to implement. It is possible—and conceptually fairly straightforward in a Bayesian model like this—that they could be *inferred* from the same data that the model uses to infer the contexts themselves. It is thus possible that our interpretation of what these parameter values mean for people's expectations about the latent cluster structure actually reflect what people have *learned* from their experience in this particular experiment, where contexts *do* tend to go on for a number of trials and recur multiple times (at least for the central cluster). Future work is required to tease these possibilities apart.

The possibility that people might be inferring the hyperparameters that govern how contexts change raises the question of what kind of changes people expect in the structure of contexts across environments. That is, are people's models of contexts nested hierarchically, in a way that allows for variation not only in the specific features of each context (e.g., the location of dots in space) but also the properties of how contexts *change* within a larger context/environment (e.g., the stickiness of contexts)? This calls for future experiments that manipulate the generative model for the contexts themselves, within subjects and over time.

More work is also needed to assess whether people actually are remembering and revisiting old contexts, as our model assumes. It is possible that people are really just detecting *changes* in context, and creating a fresh representation of a context every time they detect such a change. One way to address this is by simulating such a change-point model, which is the limiting case of our model when the concentration parameter α goes to infinity. Another way is to collect more empirical data with changes in context explicitly designed to elicit anticipation for returning to old contexts.

Finally, the strategy of our model—inferring discrete changes in context and remembering contexts—presupposes a particular underlying structure for how contexts actually tend to change in the world. A number of different strategies could be optimal, given different environments, and it is an ecological question as to which strategies are likely to be useful in the kinds of environments people tend to find themselves in. For instance, environments where latent variables don't change suddenly but rather drift slowly and continuously call for a very different family of strategies. So while our model describes behavior well in *this* particular experimental environment, that does not necessarily mean that it would also describe behavior well in an environment that does not follow the structural assumptions that the model makes.

Conclusion

In a structured world, local context—either simultaneous or temporally extended—can provide a great deal of information about how to interpret or remember stimuli. We have proposed a Bayesian model that infers latent context variables from unlabeled data, and uses that context to encode and retrieve information from memory. This model processes data *online*, one observation at a time, and captures people's behavior in a multi-context spatial memory task.

Acknowledgments

This work was supported by a National Science Foundation Grant 1453276 awarded to Pernille Hemmer.

References

Bezanson, J., Edelman, A., Karpinski, S., & Shah, V. (2017). Julia: A Fresh Approach to Numerical Computing. *SIAM Review*, *59*(1), 65–98. doi:10.1137/141000671

Chen, R., & Liu, J. S. (2000). Mixture Kalman Filters. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *62*(3), 493–508. JSTOR: 2680693

DuBrow, S., Rouhani, N., Niv, Y., & Norman, K. A. (2017). Does mental context drift or shift? *Current opinion in behavioral sciences*, *17*, 141–146. doi:10.1016/j.cobeha.2017.08.003. pmid: 29335678

Fearnhead, P. (2004). Particle filters for mixture models with an unknown number of components. *Statistics and Computing*, *14*(1), 11–21. doi:10.1023/B:STCO.0000009418.04621.cd

Fox, E. B., Sudderth, E. B., Jordan, M. I., & Willsky, A. S. (2011). A sticky HDP-HMM with application to speaker diarization. *The Annals of Applied Statistics*, *5*, 1020–1056. doi:10.1214/10-AOAS395

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian Data Analysis* (Second). Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis.

Huttenlocher, J., Hedges, L. V., & Duncan, S. (1991). Categories and particulars: Prototype effects in estimating spatial location. *Psychological Review*, *98*(3), 352–376. doi:10.1037/0033-295X.98.3.352

Kleinschmidt, D. F. (2018). Learning distributions as they come: Particle filter models for online distributional learning of phonetic categories. In T. T. Rogers, X. Rau, X. Zhu, & C. Kalish (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 1933–1938). doi:10.31234/osf.io/dymc8

Orhan, A. E., & Jacobs, R. A. (2013). A probabilistic clustering theory of the organization of visual short-term memory. *Psychological Review*, *120*(2), 297–328. doi:10.1037/a0031541

Pastell, M. (2017). Weave.jl: Scientific Reports Using Julia. *The Journal of Open Source Software*, *2*(11), 204. doi:10.21105/joss.00204

Qian, T., & Aslin, R. N. (2014). Learning bundles of stimuli renders stimulus order as a cue, not a confound. *Proceedings of the National Academy of Sciences*, *111*(40), 14400–14405. doi:10.1073/pnas.1416109111

Robbins, T., Hemmer, P., & Tang, Y. (2014). Bayesian Updating: A Framework for Understanding Medical Decision Making. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (p. 6). Quebec City: Cognitive Science Society.

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(4), 803–814. doi:10.1037/0278-7393.21.4.803

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review*, *117*(4), 1144–67. doi:10.1037/a0020511. PMID: 21038975

Schulz, E., Franklin, N. T., & Gershman, S. J. (2018). Finding structure in multi-armed bandits. *bioRxiv*, 432534. doi:10.1101/432534

An Attempt to Visualize and Quantify Speech-Motion Coordination by Recurrence Analysis: A Case Study of Rap Performance

Kentaro Kodama (kkodama@jindai.jp)
Faculty of Economics, Department of Economics, Kanagawa University
3-27-1, Rokkakubashi, Kanagawa-ku, Yokohama-shi, Kanagawa-ken, Japan

Daichi Shimizu (daichi@p.u-tokyo.ac.jp)
Department of Integrated Educational Sciences, Graduate School of Education,
University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, Tokyo-to, Japan

Kazuki Sekine (kazuki.sekine@keio.jp)
Keio University, Center for Life-Span Development of Communication Skills,
4-1-1 Hiyoshi, Kohoku-ku, Yokohama 223-8521, Japan

Abstract

Recently, cognitive science researchers have revealed that human cognition involves the body and is a kind of self-organization phenomenon emerging from dynamic interaction across body-brain-environment. Some of the data obtained from such cognitive, behavioral, or physiological activities are often complicated in terms of non-stationarity and nonlinearity. Researchers have proposed several analytical tools and frameworks. *Recurrence analysis* is one of the nonlinear data analyses developed in nonlinear dynamics. It has been applied to various research fields, including cognitive science, for language (categorical) data or motion (continuous) data. However, most previous studies have applied recurrence methods individually to categorical or continuous data. We aimed to integrate these methods to investigate the relationship between speech (categorical) and motion (continuous) directly. To do so, we added temporal information (a time stamp) to categorical data and applied the joint recurrence analysis methods to visualize and quantify speech-motion coordination during a rap performance. Our pilot study suggested the possibility of visualizing and quantifying it.

Keywords: Visualization; Quantification; Recurrence Analysis; Speech-Motion Coordination; Rap

Introduction

Cognition as a Self-Organizing Phenomenon

Recent studies have revealed, theoretically and empirically, that we cannot separate cognition from the body and its environment, which are interdependent (e.g., Anderson, Richardson, & Chemero, 2012; Riley, Shockley, & Van Orden, 2012). This notion is called *embodiment*. From the viewpoint of embodiment, cognitive processes related to language and communication interact with bodily motion and behavior (e.g., Richardson, Dale, & Shockley, 2008; Shockley, Richardson, & Dale, 2009). We can consider cognition to be a complex phenomenon that emerges from the body-brain-environment interaction (e.g., Dale, Fusaroli, Duran, & Richardson, 2013; Richardson, Dale, & Marsh, 2014).

Research has shown that the body is not only connected to cognitive processes, but also to linguistic processes. Since McNeill (1992) found the significant relationship between gestures and speech, both in production and comprehension, the number of studies on co-speech gestures has increased. Previous research has shown that co-speech gestures facilitate the speaker's speech process. For example, when participants were asked to not move their hands while speaking, the proportion of unfilled pauses (Graham & Heywood, 1975) or fillers (Rauscher, Krauss, & Chen, 1996) increased. These findings suggest that speech is closely linked to meaningful hand movements.

To deal with such a complex phenomenon, the *dynamical systems approach* (DSA) has been widely applied to human movement science, developmental psychology, and cognitive science. Compared to the traditional approach, assuming internal computation in the brain, DSA focuses more on interactions between the body (including the brain), environment, and task. The DSA has provided both a theoretical framework and analytical tools based on the nonlinear dynamics theory (e.g., Van Orden & Riley, 2005).

Visualization and Quantification

Recurrence Plot (RP): A RP is a two-dimensional graph visualizing recurring patterns of dynamical systems, in which the matrix elements correspond to those times at which a state of a dynamical system recurs in the phase space (Marwan, Carmen Romano, Thiel, & Kurths, 2007). It is an advanced technique of nonlinear data analysis and was originally developed in the fields of descriptive statistics and chaos theory (Eckmann, Kamphorst, & Ruelle, 1987).

Recurrence Quantification Analysis (RQA): RQA is a method of nonlinear data analysis that quantifies the number and duration of recurrences of a dynamical system (Marwan et al., 2007). It was originally developed to uncover subtle time correlations and repetitions of patterns, and is relatively free of assumptions about data size and distribution (Zbilut & Webber, 1992). RQA can provide researchers with some useful measures to quantify self-organizing dynamical system behavior.

RP and RQA have been applied to both continuous data, for example, a numeric value obtained by sensor devices, and categorical data, for example, a letter or word sequence in

literature pieces (Coco & Dale, 2014). However, most previous studies have applied these recurrence methods (categorical or continuous) separately. We aimed to integrate the two within the same recurrence analytical framework in order to visualize and quantify speech-action coordination/coupling.

For this purpose, we developed the *categorical recurrence analysis* and applied the *joint recurrence analysis* methods (see "Data Analysis" section under "Method"). If we can integrate these different types data within the same analytical framework, we are of the view that recurrence analysis can be extended widely to visualize and quantify various complex phenomena in cognitive science. As a first attempt to explore such a possibility, the current pilot study focused on a speech-motion coordination/coupling during a rap performance. Because rap or hip-hop music has a relatively obvious rhythm structure, and because mind-body coordination/coupling is important in rapping behavior, we assumed that this relationship would be relatively easy to extract using recurrence methods.

Method

Participants

A professional rapper (male, 30 years old, right-handed) participated in our experiment. He has more than 15 years of rapping experience and was the champion of a national freestyle rap battle. He has also released his tunes as a professional musician. The participant signed an informed consent form, agreeing to participate in this study.

Apparatus

We used a 3D motion capture system (OptiTrack Flex13, Natural Point, Inc.) to measure the participant's body movements (sampling frequency was 120 Hz) (Figure 1). Twelve reflective markers were attached to the participant's body (head, both shoulders, both elbows, both wrists, hip, both knees, and both toes). We used Motive (Natural Point) to process the time series data, MATLAB (R2017b, MathWorks) and RStudio (1.1.423) to analyze the data. We also used a video camera (HDR-PJ720, Sony) (frame rate of 50 FPS) and a headset microphone (Hafone). To analyze the audio data, we used Audiacity (2.2.2) after down-sampling at 25 FPS.

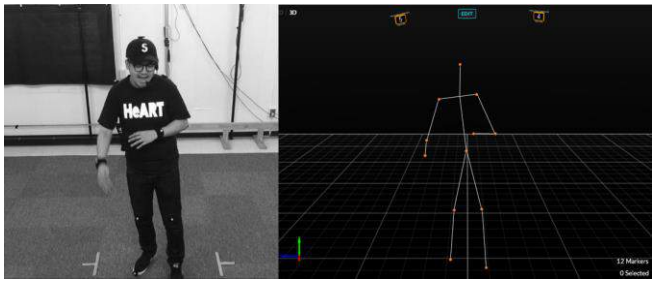


Figure 1: An experimental situation and the motion capture system.

Procedure

We required a professional rapper to perform parts of his rap song, which included an Introduction, Verse, and Hook (totaling one minute). Before recording, we attached twelve reflective markers and a microphone to his body, and we asked him to stand in front of the camera. We then instructed him to perform naturally, as if he were presenting a live performance. After sound checking, we started the recording. In this presentation, we report the results of our analysis of part of the tune (from the first Verse and Hook).

Data Analysis

To visualize and quantify the rhythmic structure and coordinated behavior between the rap (speech) and body movement (motion), we applied *recurrence analyses* (for tutorials, refer to Wallot, 2017; Webber & Zbilut, 2005). We briefly describe these recurrence methods, and we introduce the *joint recurrence method* (Marwan et al., 2007) to integrate them as described in the following paragraph.

In the case of continuous data, time series data are embedded, their trajectory is reconstructed in a higher dimensional phase space, and the distances between all possible combinations of each vector are calculated and distributed within a distance matrix (Webber & Zbilut, 2005). All elements in the distance matrix with distances at or below the threshold (i.e., radius) are said to be *recurrent* (recurrence point) and are included in the recurrence matrix, while all other elements are excluded from it. Such calculations and definitions are used to construct a *recurrence plot* (RP), a method of visualization that shows the dynamic properties and temporal patterns of the system as a two-dimensional representation (Eckmann et al., 1987). A *recurrence quantification analysis* (RQA) allows researchers to quantify and assess the properties of a dynamical system, based on RP or the phase space trajectory (more detail in Webber & Zbilut, 2005). This study reported four RQA measures, namely, the *recurrence rate* (***RR***), *percent determinism* (***DET***), *maxline* (***maxL***) and *mean line* (***L***). ***RR*** is the density (percentage) of recurrence points in a RP; ***DET*** is the percentage of recurrence points forming diagonal lines in the recurrence plot given a minimal length threshold; ***maxL*** is the length of the longest diagonal line; ***L*** is the average of the diagonal line's length (Coco & Dale, 2014). The units of these lines are indicated in time (e.g., seconds). If the length of these lines is long, it means that the system repeats the same state persistently for a long time. These measures have been interpreted as indexes related to stability or complexity of human motor/posture systems (e.g., Pellecchia, Shockley, & Turvey, 2005; Riley, Balasubramaniam, & Turvey, 1999).

In this study, we used only the hip and right wrist movements data in a vertical direction as continuous data (a collective marker of whole-body movement at the macro scale and a specific marker of rap-related rhythmic movement at the micro scale, respectively). After each time series was smoothed, it was then down-sampled at 25 Hz to integrate it with the categorical data.

In the case of categorical data, researchers generally need not to embed the data in a phase space, but to define the level or unit of analysis (e.g., a word or letter). Each unit is converted into numeric categorical sequence (e.g., 1, 2, 3, …). Researchers can create a recurrence point when the two series (original and self-copied sequential series) share the same state (i.e., the same word/letter) in time. Thus, the same RQA measures can be calculated and they provide meaningful indexes that can be considered *dynamic natural language processing*; for example, ***DET*** and ***RR*** are associated with *compressibility ration* and *co-occurrence* respectively (Dale, Duran, & Coco, 2018).

We obtained sequential data by analyzing the lyrics and converting each voice unit into a Japanese vowel (*a/i/u/e/o*), a syllabic nasal (*n*), or an assimilated sound (*x*). We chose a vowel as a main unit of analysis, because rap lyrics tend to rhyme (match rhyming words at vowel level) more often in hip-hop music, generally. We then categorized vowels into numbers as follows: *a*(1), *i*(2), *u*(3), *e*(4), *o*(5), *n*(6) and *x*(7). To analyze the audio data, we imported the audio file into a software, played the voice at each frame (25 FPS), and judged how the voice sounded. If there was no voice, we categorized the frame into *no-voice* (0); If there was a voice, we categorized it according to each vowel, a syllabic nasal, or an assimilated sound as described above (1, 2, 3, 4, 5, 6, 7). After categorization, we obtained two categorical data: first, sequential data of seven categories without any time information, and, second, time series data that included temporal information (i.e., a time stamp at 25 Hz) using eight categories from 0 to 7, as shown above.

Most previous studies have applied these recurrence methods (categorical or continuous) separately, but we integrated them within the same recurrence analytical framework in order to visualize and quantify speech-action coordination/coupling. For this purpose, we developed categorical recurrence analysis by adding temporal information (i.e., a time stamp) and applied the joint recurrence method.

The *joint recurrence analysis* was used to analyze two physically different time series (Marwan et al., 2007). A joint recurrence point can be considered as joint probability in which both systems have simultaneous recurrence points (more detail in Marwan et al., 2007). A *joint recurrence plot* (JRP) is a graph that shows all those times at which a recurrence in one dynamical system occurs simultaneously with a recurrence in a second dynamical system. In other words, the JRP is the Hadamard product of the recurrence plot of two systems (Marwan et al., 2007). JRPs capture the commonalities between two systems (i.e., signals or time series) as coinciding instances of recurrence between the individual RPs of those systems (Wallot, Roepstorff, & Mønster, 2016). First, each RP is constructed for each system, then their JRP can be computed by joining the plots together, so that common instances of recurrences are kept, but different instances between the two RPs are discarded (Wallot et al., 2016). JRQA measures such as ***RR*** and ***maxL*** as explained above (in **Data Analysis**) can be calculated from the JRP in the same way as auto/cross RQA. Originally, the joint method was proposed for two continuous time series, which can recur simultaneously in their individually reconstructed phase spaces, to compare two physically different systems at different units or dimensions. We extended this to compare continuous (motion) data with categorical data (rap).

We performed recurrence analyses using the MATLAB toolbox "CRP TOOLBOX," version 5.22 (Marwan & Kurths, 2002), and the R package "crqa," version 1.0.9 (Coco & Dale, 2014). We determined the optimal values for input parameters with reference to the standard guidelines for the RQA method (Webber & Zbilut, 2005) using *average mutual information* for determining the delay and *false nearest neighbor method* for determining the dimension (e.g., Marwan et al., 2007). As a result, for continuous data, we chose parameters of 10 for time delay, 3 for embedding dimensions, and 0.75 for the radius with *z-score* normalization, while for categorical data, we input 1 for time delay and embedding dimensions, and .001 for the radius.

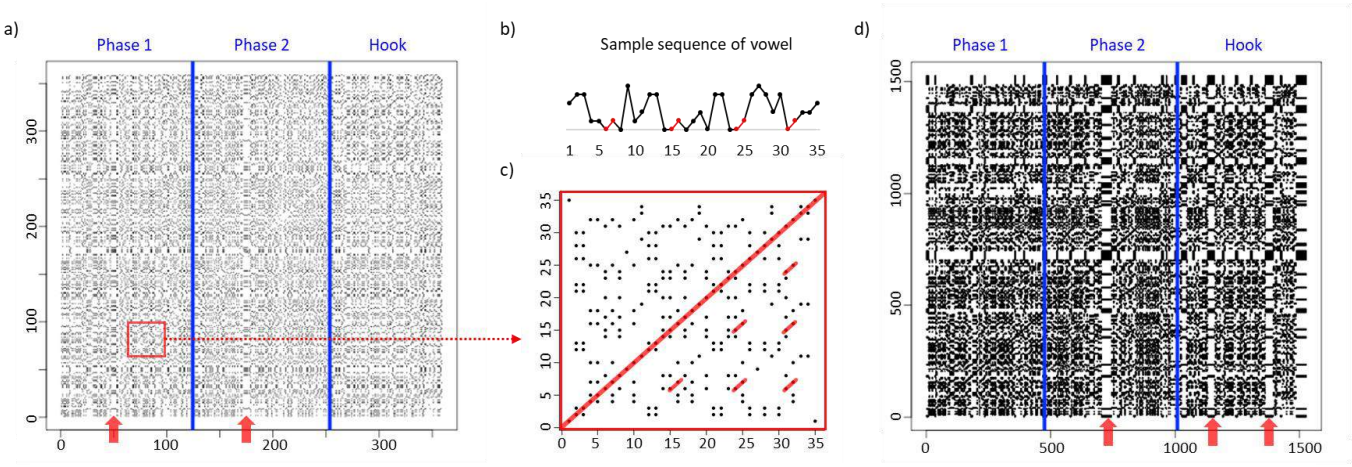


Figure 2: Categorical recurrence plot (CaRP) of rap.
a) Standard CaRP, b) Sample sequence of vowel, c) Part of CaRP, d) Proposed CaRP

Results and Discussion

Categorical Recurrence Plot: Rap Data

Figure 2a shows the categorical RP (CaRP) of the lyrics of the current rap song generated by the standard procedure (with neither temporal information nor a time stamp). Here, we report the partial result of analysis of the tune, the first Verse and the Hook. We indicated three phases consisting of the first part of the Verse (Phase 1), the latter part of the Verse (Phase 2), and the Hook by adding two blue lines (see Figure 2a). Using vowels as a unit of analysis, the lyric consisted of 359 units (Phase 1: 124, Phase 2: 129, Hook: 106). The CaRP does not have random dots, but a structured pattern across the phases. The white bands observed in Phase 1 and Phase 2 (red arrows in Figure 2a) visually represent successive vowels, then a constant value (i.e., "*a*" repeated four or five times).

Figure 2b presents a sample sequence of vowel units, while Figure 2c shows its CaRP, extracted from Figure 2a (red square). Red circle markers in Figure 2b indicate repetition (i.e., rhyming) of the same vowel units (i.e., *a*-*i*) four times in part of the lyric. The same part appears in Figure 2c as red lines parallel to the diagonal line in the center of CaRP. These parallel diagonal line structures can be interpreted as a rhyming structure, which appeared temporally. These results suggest that CaRP can provide a visualization of rhyming structure in musical lyrics.

Figure 2d presents the proposed CaRP with temporal information (i.e., a time stamp at 25 FPS). It has 1527 points (25 Hz, approximately 60 seconds) including vowels and a no-voice zero value. Accordingly, it is possible that the same value (e.g., "*a*") can appear successively; for example, "*a*" can repeat 25 times if the voice stays for one second. By adding such temporal information as a time stamp, we integrated categorical data with continuous data within the same framework (joint recurrence analysis), as discussed below. This new method seems to provide a more obvious structured pattern than the standard method, comparing Figure 2d and Figure 2a. For example, the transition point where the phase changed, or which was a *break* and *pause* in the tune, can be observed as a white band that indicates a no-voice state (red arrows in Figure 1d). These characteristics seem to express the original music (rap performance) and its temporal structure more clearly.

Our results show that CaRPs can extract a repetitive structure or recurrence pattern of the lyric and rap performance. Our proposed method can visualize the RPs in a more informative way by including temporal information. In the future, quantification and analytical indexes of rhyming structure should be explored.

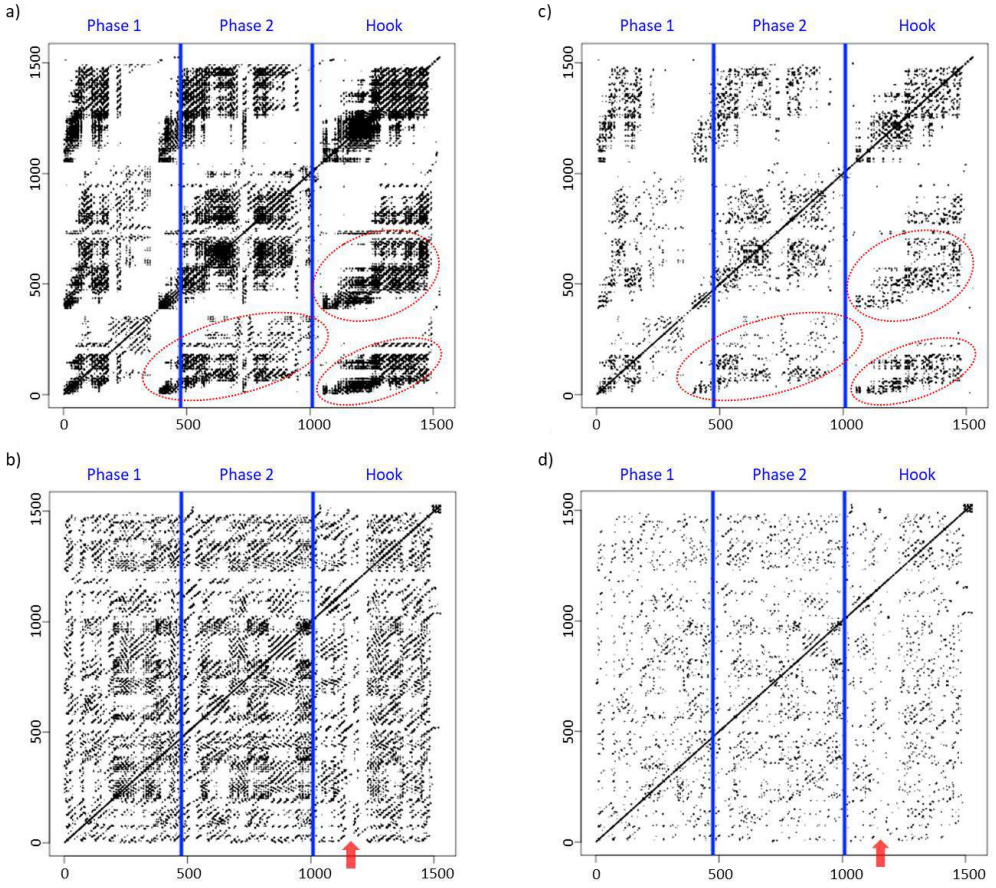


Figure 3: Continuous recurrence plot (CoRP) of rap and Joint recurrence plot (JRP).
a) CoRP of hip, b) CoRP of hand, c) JRP of rap-hip, d) JRP of rap-hand

Continuous Recurrence Plot: Motion Data

Figure 3a represents the continuous RP (CoRP) of hip motion in the vertical direction. Blue lines separate the phases again. We assumed that the vertical hip motion could represent whole-body rhythm. The CoRP shows a recurrence pattern at the macro level, but not random dots. Interestingly, a white band can be observed at the center of the CoRP like the proposed CaRP (the first red arrow in Figure 2d). We could observe the change in bodily rhythm at this break point in the actual video data. In each phase, recurrence points are shown as a whole-body beat rhythm repeatedly. Furthermore, a similar recurrence structure can be found in the red areas (i.e., Phase 1-Phase 2, Phase 1-Hook, and Phase 2-Hook). These results suggest that the participant beat out a rhythm with whole-body movement and that similar/common rhythm patterns can be found across the phases.

Figure 3b shows the CoRP of hand (i.e., right wrist) motion in the vertical direction. Blue lines separate the phases again. We chose the right wrist marker motion for analysis, because the participant was right-handed and showed specific hand movements, such as beating or gesturing, during the rap performance. Compared with hip motion, hand motion seemed to be more closely related to rap performance and to have a high frequency. As a result, its RP (Figure 3b) shows a more detailed recurrence pattern at the micro level than that in Figure 3a. The white band in Hook phase (red arrow) corresponded to no-voice part, and the right hand stopped at this moment.

Joint Recurrence Plot

Figure 3c and Figure 3d depict the joint RP (JRP) of rap-hip coordination and rap-hand (i.e., right wrist) coordination. Blue lines separate the phases again. Compared to the CoRP of hip motion (Figure 3a), the JRP of rap-hip coordination seems to hold a common recurrence pattern at the macro level (red circles in Figure 3c). This suggests that the whole-body rhythm was coupled with rap rhythm. Similarly, the JRP of rap-hand coordination (Figure 3d) seems to hold a common recurrence pattern with the RP of hand motion at the micro level (Figure 3b). This can also be considered rap-hand coupling. These results indicate that JRPs can visualize speech-motion coordination/coupling during rap performance.

Recurrence Quantification Analysis

Table 1 shows the RQA measures quantified from each RP.

Categorical RQA: The proposed method provided higher values in ***DET*** and ***maxL*** than the standard method. This came as a result of adding temporal information at 25 Hz, because it can realize successive value.

Continuous RQA: The total hip RQA measures were higher than hand RQA measures. These results suggest that the participant maintained a stable whole-body rhythm, although he moved his dominant hand rhythmically, but in a complicated manner, synchronizing with the rap lyric and beat during rap performance. To address this possibility, the relationship between hand movement (e.g., gesture) and rap lyrics can be researched in more detail in future studies.

Joint RQA: While ***RR*** and ***DET*** were higher in rap-hip coordination than in rap-hand coordination, interestingly, ***maxL*** was higher in rap-hand coordination than in rap-hip coordination. This suggests that hand movement is likely to couple with rap performance more sustainably and is involved in the content of the lyrics. We found that the right hand of the participant seemed to express the lyric contents, match with the rap tempo (e.g., beating rhythm) and correlate with rapping.

Table 1: Recurrence quantification analysis measures.

	Rap standard	**Rap** proposed	**Hip** vertical	**Wrist** vertical	**Joint** rap-hip	**Joint** rap-hand
RR	19.68	17.22	7.91	3.77	1.68	0.79
DET	36.20	91.85	94.23	76.89	76.84	61.35
maxL	18	60	435	229	16	35
L	2.28	3.74	4.35	2.88	2.82	2.66

General Discussion

In this report, we introduced temporal information (i.e., a time stamp) to the standard categorical recurrence analysis. We showed the possibility of revealing the lyrical structure and the temporal structure (i.e., rhythm) of rapping (singing) or beat (music) itself more clearly. Furthermore, we applied the joint recurrence method to integrate categorical data (rap) with continuous data (bodily motion). By employing such integration, we showed the applicability of the joint recurrence method to the investigation of the speech-motion coordination/coupling and suggested the possibility of visualizing and quantifying it.

Our current pilot study focused on hip-hop music, a music genre that has a relatively obvious rhythm and a repetitive/recurring structure (i.e., rhyme) in its lyrics, which helped us to investigate speech-motion relationship. We guessed that this relationship would be relatively easy to extract using the joint RP and RQA. Some similarities between rap dynamics and motion dynamics were found because common auditory information (i.e., a musical track) might affect these dynamics.

Future Direction

Given that we analyzed only one sample in this study, it needs to be confirmed whether our findings are robust by collecting and analyzing further data. If we could collect other rappers' data, it would be possible to compare original data to virtual pair data of rap-motion coordination/coupling generated from other rappers' performance data. This analysis would show that the current result was not produced by an artifact or possible random matching in terms of surrogate data method (e.g., Shockley, Baker, Richardson, & Fowler, 2007). It would heighten the applicability of the joint method that integrates categorical data (rap) with continuous data (motion). Although the present study focused on an intrapersonal coordination

between speech and motion, interpersonal coordination across participants can also be examined within the same framework as investigated by previous studies that have applied the recurrence analysis to various joint action tasks (e.g., Fusaroli, Konvalinka, & Wallot, 2014; Shockley & Riley, 2015). The proposed method should be applied to not only ready-made songs but also improvisational freestyle performance, including various music genres. Improvisational performance is more like everyday social interaction, in the sense that it also has complex aspects emerging from real-time interaction (Walton et al., 2018). The complex dynamical systems methods (e.g., recurrence analysis) are also expected to reveal the creative process in detail using more advanced techniques (e.g., the windowed sliding method; Coco & Dale, 2014; Kodama, Tanaka, Shimizu, Hori, & Matsui, 2018). We also aim to apply the framework not only to experimental situations but also to more ecological situations, such as the practical field of artistic performance, and daily natural conversations involving speech-motion coordination in the future (D'Ausilio, Novembre, Fadiga, & Keller, 2015; Sekine & Kita, 2015; Shimizu & Okada, 2018).

Acknowledgments

We would like to thank Prof. Rick Dale for providing us with meaningful advice and the useful R code to conduct the joint recurrence analyses. We are also grateful to two Japanese professional rappers, Darthreider and TKdakurobuchi, for collaborating with us and participating in our experiment.

References

Anderson, M. L., Richardson, M. J., & Chemero, A. (2012). Eroding the Boundaries of Cognition: Implications of Embodiment. *Topics in Cognitive Science*, 4(4), 717–730. https://doi.org/10.1111/j.1756-8765.2012.01211.x

Coco, M. I., & Dale, R. (2014). Cross-recurrence quantification analysis of categorical and continuous time series: an R package. *Frontiers in Psychology*, 5, 510. https://doi.org/10.3389/fpsyg.2014.00510

D'Ausilio, A., Novembre, G., Fadiga, L., & Keller, P. E. (2015). What can music tell us about social interaction? *Trends in Cognitive Sciences*, 19(3), 111–114. https://doi.org/10.1016/j.tics.2015.01.005

Dale, R., Duran, N. D., & Coco, M. I. (2018). Dynamic Natural Language Processing with Recurrence Quantification Analysis, 1–22. http://arxiv.org/abs/1803.07136

Dale, R., Fusaroli, R., Duran, N. D., & Richardson, D. C. (2013). The Self-Organization of Human Interaction. In H. R. Brian (Ed.), *Psychology of Learning and Motivation - Advances in Research and Theory* (Vol. 59, pp. 43–95). Academic Press. https://doi.org/10.1016/B978-0-12-407187-2.00002-2

Eckmann, J.-P., Kamphorst, S. O., & Ruelle, D. (1987). Recurrence Plots of Dynamical Systems. *Europhysics Letters* (EPL), 4(9), 973–977. https://doi.org/10.1209/0295-5075/4/9/004

Fusaroli, R., Konvalinka, I., & Wallot, S. (2014). Analyzing social interactions: the promises and challenges of using cross recurrence quantification analysis. In *Translational recurrences* (pp. 137-155). Springer, Cham. https://doi.org/10.1007/978-3-319-09531-8_9

Graham, J. A., & Heywood, S. (1975). The effects of elimination of hand gestures and of verbal codability on speech performance. *European Journal of Social Psychology*, 5(2), 189–195. https://doi.org/10.1002/ejsp.2420050204

Kodama, K., Tanaka, S., Shimizu, D., Hori, K., & Matsui, H. (2018). Heart Rate Synchrony in Psychological Counseling: A Case Study. *Psychology*, 9(07), 1858. https://doi.org/10.4236/psych.2018.97108

Marwan, N., Carmen Romano, M., Thiel, M., & Kurths, J. (2007). Recurrence plots for the analysis of complex systems. *Physics Reports*, 438(5–6), 237–329. https://doi.org/10.1016/j.physrep.2006.11.001

Marwan, N., & Kurths, J. (2002). Nonlinear analysis of bivariate data with cross recurrence plots. *Physics Letters A*, 302(5–6), 299–307. https://doi.org/10.1016/S0375-9601(02)01170-2

McNeill, D. (1992). *Hand and mind: what gestures reveal about thought*. University of Chicago Press.

Pellecchia, G. L., Shockley, K. D., & Turvey, M. T. (2005). Concurrent cognitive task modulates coordination dynamics. *Cognitive Science*, 29(4), 531–557. https://doi.org/10.1207/s15516709cog0000_12

Rauscher, F. H., Krauss, R. M., & Chen, Y. (1996). Gesture, Speech, and Lexical Access: The Role of Lexical Movements in Speech Production. *Psychological Science*, 7(4), 226–231. https://doi.org/10.1111/j.1467-9280.1996.tb00364.x

Richardson, D. C., Dale, R., & Shockley, K. D. (2008). Synchrony and swing in conversation: coordination, temporal dynamics and communication. In I. Wachsmuth, M. Lenzen, & G. Knoblich (Eds.), *Embodied Communication in Humans and Machines* (pp. 75–94). Oxford University Press,. https://doi.org/10.1093/acprof:oso/9780199231751.003.0004

Richardson, M. J., Dale, R., & Marsh, K. L. (2014). Complex Dynamical Systems in Social and Personality Psychology. In *Handbook of Research Methods in Social and Personality Psychology* (pp. 251–280).

Riley, M. A., Balasubramaniam, R., & Turvey, M. . (1999). Recurrence quantification analysis of postural fluctuations. *Gait & Posture*, 9(1), 65–78. https://doi.org/10.1016/S0966-6362(98)00044-7

Riley, M. A., Shockley, K. D., & Van Orden, G. C. (2012). Learning From the Body About the Mind. *Topics in Cognitive Science*, 4(1), 21–34. https://doi.org/10.1111/j.1756-8765.2011.01163.x

Sekine, K., & Kita, S. (2015). The parallel development of the form and meaning of two-handed gestures and

linguistic information packaging within a clause in narrative. *Open Linguistics*, 1, 490–502. https://doi.org/10.1515/opli-2015-0015

Shimizu, D., & Okada, T. (2018). How Do Creative Experts Practice New Skills? Exploratory Practice in Breakdancers. *Cognitive Science*, 42(7), 2364–2396. https://doi.org/10.1111/cogs.12668

Shockley, K. D., Baker, A. a, Richardson, M. J., & Fowler, C. A. (2007). Articulatory constraints on interpersonal postural coordination. *Journal of Experimental Psychology. Human Perception and Performance*, 33(1), 201–208. https://doi.org/10.1037/0096-1523.33.1.201

Shockley, K. D., Richardson, D. C., & Dale, R. (2009). Conversation and Coordinative Structures. *Topics in Cognitive Science*, 1(2), 305–319. https://doi.org/10.1111/j.1756-8765.2009.01021.x

Shockley, K. D., & Riley, M. A. (2015). Interpersonal couplings in human interactions. In C. L. Webber & N. Marwan (Eds.), *Recurrence Quantification Analysis Theory and Best Practices* (pp. 399–421). Springer. https://doi.org/10.1007/978-3-319-07155-8-14

Van Orden, G. C., & Riley, M. A. (Eds.). (2005). *Tutorials in Contemporary Nonlinear Methods for the Behavioral Sciences*. National Science Foundation.

Wallot, S. (2017). Recurrence Quantification Analysis of Processes and Products of Discourse: A Tutorial in R. *Discourse Processes*, 54(5--6), 382–405. https://doi.org/10.1080/0163853X.2017.1297921

Wallot, S., Roepstorff, A., & Mønster, D. (2016). Multidimensional recurrence quantification analysis (MdRQA) for the analysis of multidimensional time-series: A software implementation in MATLAB and its application to group-level data in joint action. *Frontiers in Psychology*, 7(NOV), 1–13. https://doi.org/10.3389/fpsyg.2016.01835

Walton, A. E., Washburn, A., Langland-Hassan, P., Chemero, A., Kloos, H., & Richardson, M. J. (2018). Creating Time: Social Collaboration in Music Improvisation. *Topics in Cognitive Science*, 10(1), 95–119. https://doi.org/10.1111/tops.12306

Webber, C. L., & Zbilut, J. P. (2005). Recurrence quantification analysis of nonlinear dynamical systems. In M. Riley & G. Van Orden (Eds.), *Tutorials in contemporary nonlinear methods for the behavioral sciences* (pp. 26–94).

Zbilut, J. P., & Webber, C. L. (1992). Embeddings and delays as derived from quantification of recurrence plots. *Physics Letters A*, 171(3–4), 199–203. https://doi.org/10.1016/0375-9601(92)90426-M

A neural representation of continuous space using fractional binding

Brent Komer (bjkomer@uwaterloo.ca)
Terrence C. Stewart (tcstewart@uwaterloo.ca)
Aaron R. Voelker (arvoelke@uwaterloo.ca)
Chris Eliasmith (celiasmith@uwaterloo.ca)
Centre for Theoretical Neuroscience, University of Waterloo
Waterloo, ON, Canada, N2L 3G1

Abstract

We present a novel method for constructing neurally implemented spatial representations that we show to be useful for building models of spatial cognition. This method represents continuous (i.e., real-valued) spaces using neurons, and identifies a set of operations for manipulating these representations. Specifically, we use "fractional binding" to construct "spatial semantic pointers" (SSPs) that we use to generate and manipulate representations of spatial maps encoding the positions of objects. We show how these representations can be transformed to answer queries about the location and identities of objects, move the relative or global position of items, and answer queries about regions of space, among other things. We demonstrate that the neural implementation in spiking networks of SSPs have similar accuracy and capacity as the mathematical ideal.

Keywords: Semantic Pointer Architecture; spatial semantic pointer; spatial representation; fractional binding; continuous spaces; spiking neural networks

Introduction

There is evidence for a wide variety of types of mental representation. Some mental representations are well-described using *discrete* structures (e.g., graphs, trees, lists, and so on). Others are well-described using *continuous* structures (e.g., images, maps, surfaces, and so on). Here we propose a kind of mental representation of continuous structures that is amenable to neural implementation.

Recently, there have been several proposals for how neural networks can represent discrete structures. One family of approaches, called Vector Symbolic Architectures (VSAs), defines algebras over high-dimensional vector spaces, and uses those algebras to encode such structures. VSAs have been used to characterize a variety of cognitive behaviours, including analogical reasoning (Plate, 1994), language processing (Jones & Mewhort, 2007), and concept encoding (Crawford et al., 2015). Most VSAs are defined over continuous vector spaces, including Multiply Add Permute (MAP; Gayler, 2004), Holographic Reduced Representations (HRR; Plate, 1995), and Vector-derived Transformation Binding (VTB; Gosmann, 2018). When VSAs are used to model cognitive behaviours, they essentially define methods for characterizing continuous vectors as both slots and fillers and define a method of binding fillers to slots.

In this work, we will use the Semantic Pointer Architecture (SPA; Eliasmith, 2013), which proposes a means of neurally implementing VSAs for explaining cognitive behaviour in biologically plausible spiking networks. This architecture uses aspects of VSAs for cognitive representation, but the SPA also addresses visual processing, motor control, memory, decision making, and cognitive control in ways that do not use VSAs. However, all of these elements of the SPA use representations called semantic pointers (SPs), which result from compressing and decompressing information in cortex. As a result, we can think of VSA algebras as proposing a family of compression operators that are well-suited for certain cognitive tasks.

However, as with most uses of VSAs, in past work the SPA addresses cognitive tasks with a focus on representations of discrete structures (i.e., discrete slots in a represented structure). Here we propose a method for encoding cognitive structures over continuous spaces. We call this kind of representation "spatial semantic pointers" (SSPs). In this paper we propose and examine in some detail a specific kind of SSP implemented using a particular "fractional binding" operator to encode real-valued quantities – although a variety of other operators can be analogously defined.

In the remainder of the paper we provide a mathematical definition of SSPs, and show how SSPs can provide a natural means of generating and manipulating representations that are useful for spatial cognition. We identify desiderata for spatial representation that are useful for cognitive explanations. We then implement this representation both mathematically and neurally, and perform simulation experiments to demonstrate that it has a variety of useful properties, including: being able to query a memory for its spatial or non-spatial contents, representing multiple objects and locations simultaneously, spatially transforming memory contents without decoding them, and representing regions of space of various shapes and sizes. The choice of VSA and binding operator used in this work allows the representation and various transformations to be implemented efficiently by a spiking neural network.

A spatial representation

Our proposed representation generalizes the notion of vector binding to continuous spaces. By analogy to fractional powers defining the multiplication of reals, we define fractional bindings for vectors in a vector space. To explain, let us first consider binding a vector to itself a discrete number of times. That is, let $k \in \mathbb{N}$ be a natural number, $B \in \mathbb{R}^d$ be a fixed d-dimensional vector (i.e., semantic pointer), and $\circledast$ be a binding operator. We can repeatedly bind B with itself $k-1$

times[1] as follows:

$$B^k = \underbrace{B \circledast B \circledast \ldots \circledast B}_{B \text{ appears } k \text{ times}}. \quad (1)$$

This representation has been used in several cognitive models, for instance, to encode the position (k) in a list in serial working memory (Choo & Eliasmith, 2010). We propose to generalize this to continuous quantities (as opposed to discrete lists, for example) by permitting k to be real. Allowing a real k means that the resulting vector B^k encodes a continuous quantity. Most VSA operators can be interpreted in this manner (including MAP (Gayler, 2004), VTB (Gosmann, 2018), and HRR (Plate, 1995)), but not all (e.g., spatter codes (Kanerva, 1994)).

In the specific case of the SPA, we take the binding operator to be circular convolution (as proposed by Plate) and the fixed d-dimensional vectors to be semantic pointers chosen from the unit sphere. We then define our fractional binding operation by expressing equation 1 in the complex domain:

$$B^k = \mathcal{F}^{-1}\left\{\mathcal{F}\{B\}^k\right\}, \quad k \in \mathbb{R}, \quad (2)$$

where $\mathcal{F}\{\cdot\}$ is the Fourier transform, and $\mathcal{F}\{B\}^k$ is an element-wise exponentiation of a complex vector—analogous to exponentiation using fractional powers (e.g., $b^{2.5}$)—permitting k to be real.[2] In the present paper, we use unitary vectors for B due to the fact that their length does not change with multiple bindings, and their inverse is equal to their approximate inverse (see below).

This definition comes with many useful algebraic properties analogous to the relationship between multiplication and exponentiation (e.g., $b^{2.5}b^{1.5} = b^4$), in particular:

$$B^{k_1} \circledast B^{k_2} = B^{k_1+k_2}, \quad k_1, k_2 \in \mathbb{R}. \quad (3)$$

In essence, fractional binding is to circular convolution as exponentiation is to multiplication. We exploit equation 3 to perform semantically meaningful operations (e.g., shifting space) in our experiments.

Next, we extend this representation to multiple dimensions, which is the focus of our experiments below. In general, we can represent points in $\mathbb{R}^n$ by repeating equation 2, n times, using a different semantic pointer for each represented dimension (i.e., for each axis), and then binding all of the resulting vectors together. For $n = 2$, we think of the representation as encoding a continuous 2-D spatial representation (e.g., the location of objects on a map). In this case, the SSP that represents the point (x, y) is defined as the vector resulting from the function:

$$S(x, y) = X^x \circledast Y^y, \quad (4)$$

where X and Y are fixed semantic pointers, x and y are reals, and we are using fractional binding as defined by equation 2.

Similarly, the SSP that represents a continuous region (e.g., a solid rectangle), specified by some infinite set of 2-D points R, is defined as:

$$S(R) = \int_{(x,y)\in R} X^x \circledast Y^y \, dx\, dy. \quad (5)$$

There are efficient ways to compute equations 4 and 5 with spiking neurons using the Neural Engineering Framework (NEF; Eliasmith & Anderson, 2003). We use a publicly-available implementation in several of our results below.

To represent a single object occupying some location or region, we bind its semantic pointer representation, OBJ, with the SSP from equation 4 or 5, respectively:

$$M = OBJ \circledast S. \quad (6)$$

In general, to represent a set of m labelled objects together in the same memory, we can use superposition:

$$M = \sum_{i=1}^{m} OBJ_i \circledast S_i, \quad (7)$$

with a distinct semantic pointer OBJ_i tagging each object.

Given a representation like that in equation 7, we can query it in a number of ways. For example, to determine what object is at location (x, y) we can compute:

$$M \circledast (X^x \circledast Y^y)^{-1} = M \circledast X^{-x} \circledast Y^{-y}. \quad (8)$$

By the properties of binding and superposition, the resulting vector will have highest cosine similarity (i.e., dot product) with the object at (x, y).[3] Note that the inverse used in equation 8 is approximate, but choosing X and Y to be unitary vectors guarantees it is equal to the true inverse.

We can construct a heatmap of representations defined by equation 7, to visualize a decoding of the objects back into the original continuous space. For instance, for $m = 2$ (i.e., two represented objects), taking the dot product of $M \circledast OBJ_i^{-1}$ (M is from equation 7) with vectors representing positions spaced by Δx and Δy to tile the 2-D space, provides the visualization of Figure 1.

In summary, fractional binding provides a scheme for encoding a set of n-dimensional points into a d-dimensional SSP. This comes with an algebra for operating on these SSPs in meaningful ways (e.g., querying, shifting, and so on). When combined with the methods of the SPA, we can spatially manipulate collections of objects in a spiking neural implementation, as detailed in our experiments below.

Desiderata for spatial representation

To test if the proposed metric representation is useful, we consider its ability to be used in a variety of ways for representing, querying, and updating representations of objects in a spatial map. Here we describe the tests we perform, and in the next section we present the results of these tests.

[1] When $k = 0$ we get the identity vector corresponding to $\circledast$.

[2] For natural k, equations 1 and 2 are mathematically equivalent.

[3] This assumes d is sufficiently large, relative to m, as is typical for VSAs.

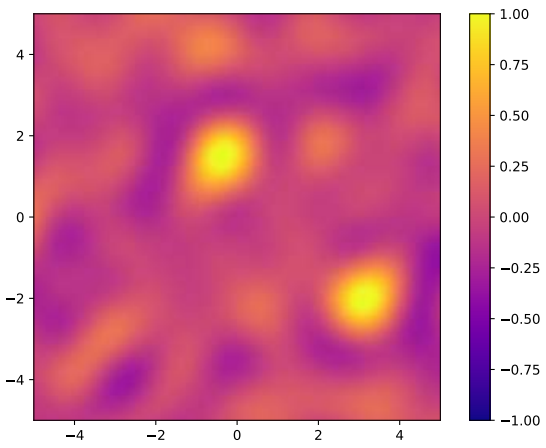


Figure 1: A heatmap visualizing the representation of two objects at different locations, as specified by equation 7. This graph is the sum of the decoding for each object.

The proposed desiderata for manipulating 2-D spatial representations are identified and briefly described in Table 1. In this table, M refers to the representation generated as described in the previous section.

Example queries

To illustrate the application of representing objects at continuous spatial locations using SSPs we demonstrate a variety of example queries in Figure 2. A set of animals (items) at random spatial locations are encoded into an SSP using equation 7 as shown in Figure 2a. This is accomplished by binding the SP representing each object with the SSP corresponding to its location, summing these values together, and then normalizing the result.

Various queries can be made with this representation. Figure 2c (top) shows the results of asking for the locations of different objects, decoded as a heatmap. If the object exists at more than one location, the resulting SSP will be highly similar to all of these locations (image on the left). If the object does not exist at any location, the resulting SSP will not be similar to any location on the heatmap (image on the right). Figure 2c (bottom) shows the reverse is possible too: given a location, find out which objects are at that location. If there are no objects at the queried location, the result will be noise and will not be similar to any object in the vocabulary (as shown in the far right).

Location queries can also be extended to regions of space, as shown in Figure 2b. If the region encompasses multiple objects, all objects should be returned, as depicted by the bar charts at the bottom. The region semantic pointers themselves are a single vector that is formed by integrating over the spatial semantic pointers within the region and normalizing the result, as described by equation 5. This process creates a vector that has a high dot product similarity with all vectors within a particular area while having a low dot product with vectors outside the region. Two example represented regions are illustrated in the heatmaps at the top of the figure. It is important to note that due to the normalization, the dot product with the region vector and a single point within the region will decrease as the area of the region increases. The consequence of this fact is that the optimal threshold for detection is a function of the area.

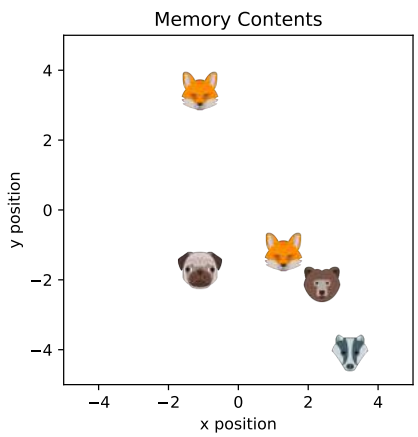
Desiderata	Description
Capacity	Determine how many objects can be encoded into M and a target object successfully decoded.
Query single object	Find the location of an object given the object and M.
Query missing object	Indicate if an object is not present when queried.
Query location	Determine what object is at a given location.
Query duplicate object	Determine the positions of multiple versions of the same object.
Neural implementation	Implement the operations in spiking neurons.
Region representation	Represent an entire region in the 2-D space.
Query Region	Determine which objects are in a spatial region.
Shift single object in group	Change the position of a single object without decoding M.
Shift whole group	Change the position of all objects in M similarly.
Readout (x,y) location from SSP	Map from the SSP representation to the 2-D space.

Table 1: Desiderata for metric representations of space.

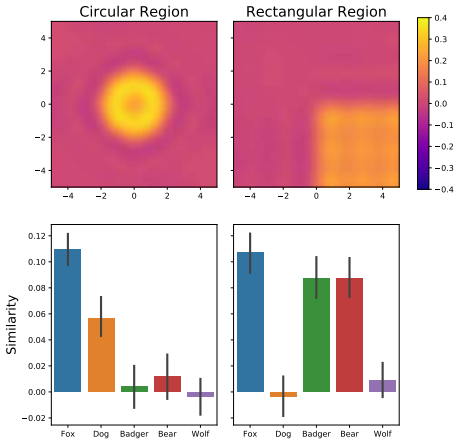
Experimental methods

To quantify how well this spatial representation performs in general for each of the desideratum a consistent measure must be used. In this paper we chose to use the accuracy of the output. When the output is a semantic pointer for an object, it is considered correct if its vector is more similar to the vector for the correct object than any other vector from a vocabulary of objects. Vocabularies are randomly generated semantic pointers of between 4 and 48 items, as described below. Similarity is determined by taking the dot product between vectors, with a larger value corresponding to a better match. When the output is a semantic pointer for a location, it is considered correct when the represented location is within 0.5 units of the true location. This threshold is chosen because it is approximately the radius of the region of similarity that a spatial semantic pointer has around itself.

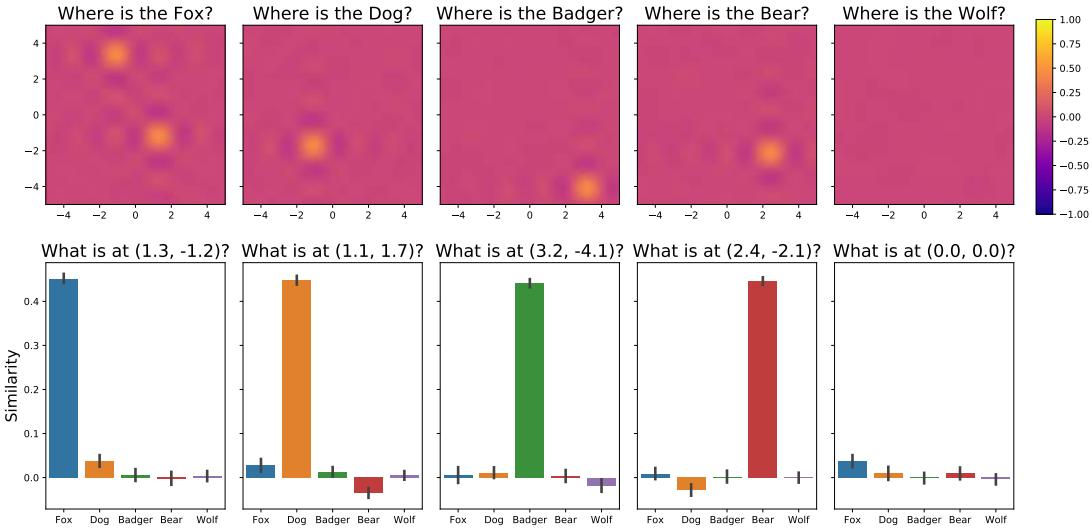
The capacity calculation requires identifying a threshold



(a) Memory contents



(b) Region queries



(c) Location and item queries

Figure 2: Example queries of items, locations, and regions. a) An example memory encoded into an SSP. b) Region queries applied to the memory in (a). c) Object (top) and location (bottom) queries applied to the memory in (a).

above which an item is identified as present in the representation. For this purpose we pick a threshold that is 3 sigma above the mean, ensuring a 99.7% chance of accepting only correct items.

For our capacity tests, we experiment with a dimensionality of 128, 256, and 512, and observe the overall effect on performance. In all other tests we fix the number of dimensions to 512, which we have found achieves good performance across a wide variety of tasks in both spiking and non-spiking regimes.

For each desideratum, accuracy reported is the mean across 6,000 trials with memory sizes varying uniformly between 2 and 24 items. For each trial the vocabulary of objects is chosen to be twice the number of objects encoded into memory (i.e., 4 to 48).

Each object is assigned a random unitary vector. All semantic pointers used in each task are normalized after every operation. All 2-D coordinates used in the experiments are chosen uniformly at random within the domain of -5 to 5 for both x and y. The size of the domain in relation to the dimensionality of the semantic pointers determines the ideal level of performance (not shown).

Query single object Equation 9 is used to produce an SSP representing the location of the desired object. Accuracy is computed by decoding this high-dimensional vector, S, into the 2-D coordinate it represents and comparing to the true location.

$$S = M \circledast OBJ^{-1}. \tag{9}$$

Query missing object Given a memory containing objects, query an object that does not exist (using equation 9). The correct behaviour is a result that is highly dissimilar to all

locations within the domain of interest. This is determined by the dot product of S and every SSP being less than 0.1.

Query duplicate object Given a memory containing many objects with some duplicates, query an object that appears twice. The correct behaviour is to return a spatial semantic pointer that represents both locations of this object.

Query location Use equation 8 with the location for one of the objects in memory. The correct behaviour is to return a semantic pointer for the object at that location.

Query region On each trial a circular region is created with a radius between 1 and 3 units and centered at a random location. An SSP is constructed for this region using equation 5. The inverse of this SSP is convolved with the memory to obtain a semantic pointer representing all objects in the region. Accuracy is computed by adding the number of objects correctly detected in the region to the number of objects correctly not detected from outside the region and then dividing by the total number of objects in the memory.

Shift single object in group Moving a single object within a group can be accomplished by adding the object of interest convolved with a vector that is the difference between the start and end positions, as shown in equation 10. Accuracy is reported for all objects as well as just the object that was moved.

$$\Delta M = OBJ \circledast \Delta S. \tag{10}$$

Shift whole group The memory is convolved with an SSP that corresponds to a random displacement, which leverages the property of equation 3. An object query is then performed for each object in the memory and the result is considered correct if it moved by the displacement amount. A heatmap visualizing the result of the two shifting operations is shown in Figure 3 for a group of three identical objects.

Readout (x,y) location from SSP For the non-neural case location is extracted from the maximum point in the heatmap. In spiking neurons a heteroassociative memory is optimized to map from a 512-dimensional SSP to a 2-D location.

Construct SSP from (x,y) location This can be computed directly from equation 4. For the experiment using spiking neurons each axis is first computed separately and then convolved together.

All experiments were repeated using networks of leaky integrate-and-fire (LIF) neurons and the NEF to implement the necessary transformations. In all trials 50 neurons were used per dimension to represent the memory and to compute circular convolutions.

Results

The results of the experiments for each of the desideratum are shown in Table 2.[4] As can be seen from the table, the SSP representation is able to address the desiderata quite well, both in purely mathematical and neural implementations. The worst performance is evident in the shifting of a single object in a group. Specifically, the accuracy of the representation for the objects that were not shifted decreases, while the accuracy for the shifted object increases. This is due to normalization effects making the moved object be re-encoded with a larger relative magnitude than the rest of the items. Using a scaling factor proportional to the number of items in the memory mitigates this effect (improves accuracy from 75.7% to 97.8%), but in general the number of items within a memory is not always known without first retrieving items from memory, and equation 10 is agnostic to the other contents of the memory.

To better characterize the capacity of a single memory using this representation we performed queries on memories with progressively larger numbers of items encoded (see Figure 4). The shape of the curve is very similar for both location and object queries since the decrease in the dot product is mostly a result of the normalization of the memory to a unit vector. The standard deviation for the dot product of two

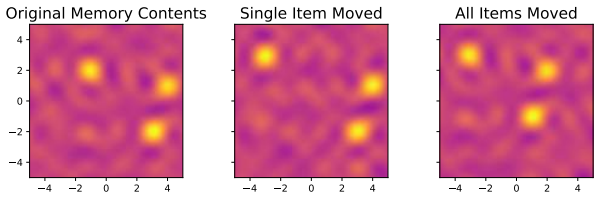


Figure 3: Shifting objects in memory. Left: The original memory. Middle: Shifting the top left object. Right: Shifting all three objects.

Desiderata	Accuracy	
	Non-Neural	Neural
Query single object	99.1%	95.7%
Query missing object	99.4%	96.7%
Query location	97.3%	94.7%
Query duplicate object	97.4%	95.3%
Query region	90.4%	73.5%
Shift single object in group (all objects)	75.7%	67.3%
Shift single object in group (moved object)	100.0%	100.0%
Shift whole group	97.8%	96.7%
Readout (x,y) location from SSP	100.0%	94.1%
Construct SSP from (x,y) location	100.0%	99.0%

Table 2: Experiments for the desiderata for metric representations of space. Accuracy is calculated using SSP representations containing 2 to 24 items. When the output is a location, it is considered correct when the result is within 0.5 units of the true location.

[4] All source code required to reproduce these experiments and generate the figures is available at https://github.com/ctn-waterloo/cogsci2019-ssp.

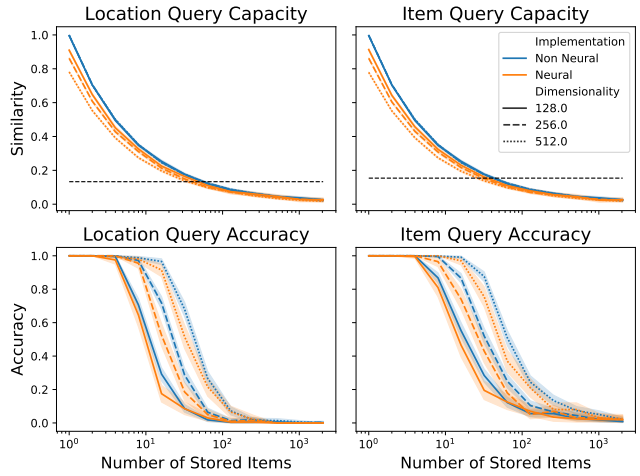


Figure 4: Memory capacity and accuracy as a function of the number of items in the SSP for ideal and neural implementations while varying the dimensionality. Top panels show the item and location capacity. Bottom panels show the item and location accuracy.

random vectors in a unit hypersphere is $\sqrt{1/D}$ where D is the dimensionality of the space. The mean is zero, so for 512 dimensions this results in a 3-sigma threshold similarity of 0.133. SSPs that represent coordinates within a finite domain will span a smaller subspace of the hypersphere, so their threshold will be a little higher. Specifically, we estimated the threshold by generating 10,000 random SSPs from a 10×10 2-D domain and computing the dot product between every pair. The mean is approximately zero, and three standard deviations is 0.154. Consequently 99.7% of queries will be above this value for items actually in the memory. The accuracy plots show the importance of dimensionality on accuracy of decoding memories.

Discussion

We have proposed a novel neural representation, SSPs, for encoding structured continuous spaces using fractional binding. We have demonstrated that these representations satisfy desiderata for representations that are useful for spatial cognition. By implementing these methods at the level of spiking neurons, this work enables future exploration of tradeoffs between neural constraints and performance for tasks of increasing complexity. In addition, a spiking neural implementation serves as a prerequisite for constructing dynamical models of spatial cognition that operate sparsely over time and in an event-driven manner.

SSPs have many potential applications for modelling cognitive phenomena that involve spatial reasoning over time, such as path planning and navigation. Objects that a cognitive agent encounters while traversing a space can be stored in memory in such a way that their relative distances and directions from each other are preserved.

The extension of Vector Symbolic Architectures to continuous representation presented in this work is not limited to representing physical space. Any continuous dimension over which concepts may vary (e.g., mass, colour, value, and so on) can utilize this representation.

While we have explored some of the capacity and accuracy limitations of this representation, it is important to note that the effective capacity can likely be increased by hierarchically chunking items into groups when encoding them into the memory by using a similar technique as the method demonstrated in Crawford et al. (2015).

Areas of future work include exploring the theoretical foundations of this method to improve our understanding of its strengths and limitations. As well, there remain many questions regarding how well a cognitive model using these representations can scale and how well the behaviour and neural recordings from such a model match that of animals.

Acknowledgments

We would like to thank Jan Gosmann for his work on the mathematical foundations of fractional binding for semantic pointers, and personal discussions. This work was supported by CFI and OIT infrastructure funding, the Canada Research Chairs program, NSERC Discovery grant 261453, ONR grant N000141310419, AFOSR grant FA8655-13-1-3084, OGS, and NSERC CGS-D.

References

Choo, X., & Eliasmith, C. (2010, 08/2010). *A spiking neuron model of serial-order recall.* Portland, OR: Cognitive Science Society.

Crawford, E., Gingerich, M., & Eliasmith, C. (2015). Biologically plausible, human-scale knowledge representation. *Cognitive Science*. doi: 10.1111/cogs.12261

Eliasmith, C. (2013). *How to build a brain: A neural architecture for biological cognition*. Oxford University Press.

Eliasmith, C., & Anderson, C. H. (2003). *Neural engineering: Computation, representation, and dynamics in neurobiological systems*. MIT press.

Gayler, R. W. (2004). Vector symbolic architectures answer Jackendoff's challenges for cognitive neuroscience. *arXiv preprint cs/0412059*.

Gosmann, J. (2018). An integrated model of context, short-term, and long-term memory.

Jones, M. N., & Mewhort, D. J. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological review*, *114*(1), 1.

Kanerva, P. (1994). The spatter code for encoding concepts at many levels. In *Icann94* (pp. 226–229). Springer.

Plate, T. A. (1994). Estimating analogical similarity by dot-products of holographic reduced representations. In *Advances in neural information processing systems* (pp. 1109–1116).

Plate, T. A. (1995). Holographic reduced representations. *IEEE Transactions on Neural networks*, *6*(3), 623–641.

The trajectory of counterfactual simulation in development

Jonathan F. Kominsky[1], Tobias Gerstenberg[2], Madeline Pelz[3], Mark Sheskin[4], Henrik Singmann[5], Laura Schulz[3], & Frank C. Keil[4]

[1]Harvard University, [2]Stanford University, [3]MIT, [4]Yale University, [5]University of Warwick, UK

Abstract

Previous work has argued that young children do not answer counterfactual questions (e.g. "what would have happened?") by constructing simulations of alternative possibilities in the way adults do. Here, we propose that children can engage in simulation when answering these questions, but consider different counterfactual possibilities than adults. While most previous research has relied on narrative stimuli, we use causal perception events, which are understood even in infancy. In Experiment 1, we replicate earlier findings that children struggle with counterfactual reasoning, but show that they are capable of conducting the required simulations in a prediction task. In Experiment 2, we use a novel multiple-choice method that allows us to study not only *when* children get it right, but also *how* they get it wrong. We find evidence that 4-year-olds engage in simulation, but preserve only some features of what actually happened and not others.

Keywords: causality; counterfactual reasoning; perception; child development; multinomial process trees

Introduction

When considering whether some event C caused another event E, we do not merely consider events as they actually unfolded. Rather, we think about what *could* or *would* have happened had C been altered in some way (Byrne, 2016; Lewis, 1973). This capability for counterfactual reasoning is an essential, and perhaps even automatic, feature of causal cognition (Gerstenberg, Peterson, Goodman, Lagnado, & Tennenbaum, 2017), with a variety of consequences. For example, the relevance of different counterfactual possibilities affects causal judgments (Phillips, Luguri, & Knobe, 2015; Icard, Kominsky, & Knobe, 2017; Phillips & Kominsky, 2017), counterfactual reasoning undergirds emotions, like regret and relief (Beck & Riggs, 2014), and is an implicit component of Bayesian causal learning (Pearl, 2000).

One of the essential properties of counterfactual reasoning is *simulation*. When people engage in counterfactual reasoning, they construct a mental model of the events as they actually happened, and then imagine how the events would have unfolded if something about the situation had been different. This mental simulation is guided by a causal model of the situation which dictates the consequences of counterfactual interventions (e.g., Sloman & Lagnado, 2005).

The developmental origins of counterfactual reasoning in the human mind remain a challenging mystery to cognitive science. Piaget held that counterfactual reasoning emerged in the developmental stage of "formal operations", starting at about 12 years of age (Inhelder & Piaget, 1958). Later work found that children as young as 3 could answer certain counterfactual questions correctly. For example, presented with a story about a girl named Carol who walked across a floor with dirty shoes, 3-5-year-olds who were asked "what would have happened if Carol had taken her shoes off?" correctly answered the floor would be clean (Harris, German, & Mills, 1996).

However, later work suggested that children may arrive at such answers without engaging in counterfactual simulation, and simply rely on conditional reasoning instead. In general, dirty shoes make floors dirty, while clean shoes leave floors clean (Rafetseder, Schwitalla, & Perner, 2013). However, basic conditional reasoning and counterfactual reasoning come apart in situations in which the outcome is causally *overdetermined*. When an outcome was overdetermined, this means that there were multiple individually sufficient causes such that the outcome would still have come about even if one (or more) of the causes hadn't occurred. For example, if both Carol and Max walk across the kitchen floor with dirty shoes, and children and adults are asked what would have happened if Carol had taken her shoes off, adults say the floor would still have been dirty (because of Max), whereas 5-year-olds overwhelmingly say the floor would have been clean. Remarkably, 10-year-olds responded at chance, and adult-like performance emerged only around 14 years of age (Rafetseder et al. 2013).

Recent work has, again, been more optimistic about children's counterfactual reasoning abilities. When narratives are replaced by simple "blicket detector" causal systems in which only some blocks (called "blickets") can make a machine go, children show above-chance success for overdetermined outcomes around age 6 (Gopnik & Sobel, 2000), or even at age 4-5 (Nyhout & Ganea, 2019).

However, we believe that what it means to succeed in counterfactual reasoning needs to be examined more closely. In the research to date, researchers have generally concluded that the reason why children answer these questions incorrectly, is because they *do not simulate* counterfactual alternatives, but instead arrive at their answers by some other reasoning strategy (Rafeteseder et al., 2013; Nyhout & Ganea, 2019). This is remarkable given that other work has found that children are quite adept at simulation when making *predictions* about events that have not yet occurred (Atance & O'Neill, 2005). Given that children can engage in simulation in some cases, and that adults naturally do so when answering causal questions (Gerstenberg et al., 2017), the assumption that young children fail to reason counterfactually because they do not engage in counterfactual simulation *at all* is worth re-examining.

There is another possible reason for why children respond differently than adults: Rather than failing to simulate, they instead simulate different counterfactual alternatives than

adults do. This proposal aligns with a recent proposal that young children may consider a broader hypothesis space than adults do when engaging in causal reasoning (Gopnik et al., 2017). Similar to how children may be more flexible in what hypotheses they consider in causal reasoning, it is possible that they also consider different possibilities than adults do, when simulating counterfactuals. Here, we are interested to see whether there is systematicity in the way in which children consider counterfactual possibilities. When children get the answer to a counterfactual questions wrong, are they just randomly guessing, or may they systematically consider different possibilities than adults do? Characterizing such potential systematicity could give unique insight into the development of counterfactual reasoning, and a deeper understanding of what features of an event children consider *mutable* (Byrne, 2016).

In order to examine which specific counterfactual possibilities children consider, we depart from the narrative studies that have been used in most prior work. Narrative stimuli add a great deal of memory load and room for influence from idiosyncratic knowledge. The ideal stimuli would be a causal event that children understand nearly effortlessly, that they can see in full while answering a counterfactual question, and which offers the opportunity to ask not just whether they are simulating counterfactual alternatives, but which specific alternatives they consider.

Simple physical interactions that fall under the category of "causal perception" perfectly fit these criteria. Events in which one object appears to collide with another and cause it to move are perceived as causal by 6 *months* of age (Leslie & Keeble, 1987; Saxe & Carey 2006; Kominsky et al., 2017), and recent work has used these events to demonstrate counterfactual simulation in causal judgment with adults (Gerstenberg et al., 2017).

In the current work, we present two experiments investigating the development of counterfactual simulation, using causal perception events. In Experiments 1a and 1b, we replicate previous findings that children struggle with counterfactual reasoning in overdetermined cases, but in the domain of causal perception events. However, we also find that children are highly accurate when making *predictions* about these events, showing that they are able, in principle, to conduct the necessary simulations to answer the questions correctly.

In Experiment 2, we present children with concrete counterfactual alternatives to causal perception events in a multiple-choice answer format similar to that employed by Rafetseder and Perner (2018). This response format allows us to examine not only *whether* children engage in counterfactual simulation, but *which specific counterfactual possibilities* they consider.

Experiment 1a

The goal of this experiment was to validate the domain of causal perception in the study of children's counterfactual judgments, by having children make counterfactual judgments about simple causal perception events.

Methods

Participants We planned to run 40 children in each age group (20 in each of two conditions), and continued collecting data until we had reached that target, replacing any participants that were excluded (see below). 40 5-6-year-olds (15 female), 40 7-8-year-olds (15 female), and 40 9-10-year-olds (18 female) participated in Experiment 1a, recruited from local schools and children's museums. In addition, 10 5-6-year-olds (5 female), 3 7-8-year-olds (2 female) and 1 (male) 9-10-year- old participated but were excluded from analyses based on predetermined exclusion criteria (see below).

Stimuli and procedure We constructed simple animations modeled on those used by Gerstenberg et al. (2015) (see Fig. 1, videos of the animations can be found here: http://osf.io/qwphr/). In these animations, there are two balls, A and E, a red area that was described as a "goal", and black walls on either side of the goal. The stimuli were animated .gif files placed into a Qualtrics survey (Qualtrics, 2005). The survey was presented on an iPad.

All participants first saw two training items in counterbalanced order. In one training item, ball A hit ball E, which then bounced off the wall above the goal. In the other training item, ball A hit ball E, which then went into the goal. Following each training trial, participants were asked two questions: "Before ball A hit ball E, was ball E moving or sitting still?", and "Did ball E go into the goal?" Participants could verbally respond and the experimenter would record their answer, or older children could select the option on the iPad directly. If participants answered either question incorrectly on one of the training trials, they were shown that training animation a second time and asked again.

Participants then saw one of two test trials, between-subjects. In the "difference-making" condition, the animation was almost identical to the training item in which ball E bounced off the wall above the goal, except that there was a "brick wall" (see Fig. 1) that ball E bounced off of, and ball E went into the goal. In the "overdetermined" condition, the animation was almost identical to the training item in which the ball went into the goal, except that the ball bounced off the brick wall before going into the goal, thus leaving the outcome unchanged.

Following the test trial, participants were asked the same two questions as in the training trials. If children answered either question incorrectly, they were not corrected but their data were excluded. Then, children were asked the critical test question: "What if the brick wall had not been there? Would ball E have gone into the goal?"

Results and discussion

Results can be found in Fig. 2. A simple inspection of this figure gives a clear sense of the results, which were similar across all age groups: Near-perfect performance on cases in which the brick wall made a difference (where the correct answer is that ball E would not have gone into the goal), but only roughly 50% accuracy for overdetermined events (where the correct answer is that ball E would still have gone

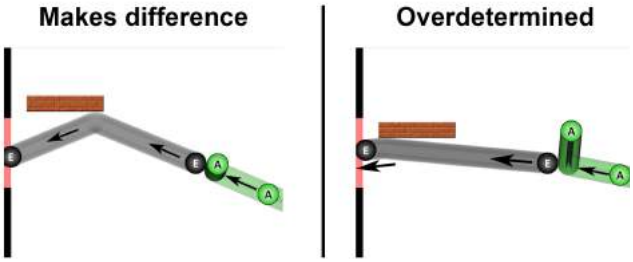


Figure 1. Example stimuli from Experiment 1a. In the difference-making event (left), the brick wall altered ball E's trajectory such that it went into the goal. In the overdetermined condition (right), ball E also deflects of the wall, but would have gone into the goal regardless.

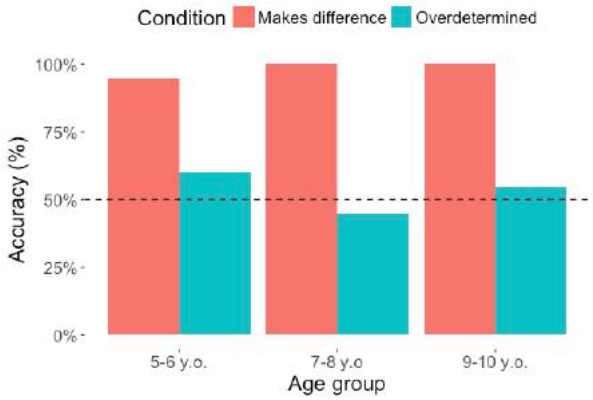


Figure 2. Proportion of accurate responses to the counterfactual question in Experiment 1a.

into the goal). A logistic regression with age group and condition as factors revealed a main effect of condition, ß = 2.54, p = .02, but no effect of age group and no interactions, p > .9. As children demonstrated nearly uniform perfect performance in the difference-making condition (one incorrect answer in total), no further analyses were conducted for this condition. For the overdetermined condition, a logistic regression with age group also showed no effect of age (p > .3) and no significant intercept (p = .37), indicating that accuracy did not differ from chance (i.e., .5).

These results are very similar to many earlier results investigating children's counterfactual reasoning (e.g., Rafetseder et al., 2013): Children can answer counterfactual questions when the correct answer changes the outcome, but struggle in overdetermined cases. One reason for this could be that children are unable to successfully simulate the required counterfactual possibility in these causal perception events. Experiment 1b tested this hypothesis by asking children to make predictive simulations about these very events, without the brick wall.

Experiment 1b

In this experiment, we wanted to see whether children are capable of correctly predicting what will happen after the animation is paused. It is possible that children failed to answer the counterfactual question correctly in the overdetermined situation because they have trouble simulating what would have happened in this case.

Methods

Participants This study was stopped early due to the fact that all children responded correctly. Our final sample sizes were therefore 21 5-6-year-olds (10 female) and 26 7-8-year-olds (14 female) recruited from the same populations as Experiment 1a. In addition, 4 5-6-year-olds (2 female) and 1 (male) 7-8-year-old were excluded based on predetermined exclusion criteria (see below).

Stimuli and procedure The stimuli were similar to Experiment 1a with the following differences: Participants first saw four training trials in random order: Two in which ball E went into the goal and two in which it missed the goal. First, children saw an animation where ball A struck ball E, and ball E moved approximately halfway from its starting position to the left edge of the display (where the wall and goal are located). At this point the animation froze and a large "pause" icon appeared (that didn't obstruct either of the balls). Children were then asked, "If ball E keeps going, will it go into the goal?" Children could respond "yes" or "no". For the training trials, children then saw the rest of the animation. If children made incorrect predictions on at least two of these items, they were excluded from analyses on the basis that they did not understand the task.

Following training, children saw two test trials, a "difference-making" trial and an "overdetermined" trial in counterbalanced order. The test trials were identical to those used in Experiment 1a, with two exceptions: First, the brick wall was not visible (i.e., identical to Experiment 1a's training trials). Second, the animation paused on the frame in which the ball would have collided with the brick wall in Experiment 1a (participants had no way of knowing this). Participants were then asked the same question as in the training items, but were not shown the end of the animation. Note that the predictions that children are asked to make in Experiment 1b are identical to the counterfactual simulation that is required to answer what would have happened without the brick wall in Experiment 1a.

Results and discussion

Every single child who passed the training provided correct answers to both test questions (21/21 5-6-year-olds and 26/26 7-8-year-olds). We report no statistical tests because the uniformity of these responses renders such tests uninformative.

Experiment 2

Experiment 1b showed that, in line with prior work, children are capable of engaging in the kind of physical simulation that is required to answer counterfactual questions correctly, but did not do so consistently for the overdetermined item in Experiment 1a. This result suggests that children's counterfactual reasoning about causal perception stimuli is similar to their reasoning in other domains. However, we

cannot tell based on these findings why children sometimes get it wrong. One explanation is like that proposed by Rafetseder et al. (2013): Children did not engage in simulation at all when asked to consider the counterfactual question. While this is still possible, given that they are obviously capable of engaging in simulation, we must ask *why*.

One possibility is that children cannot simulate while holding the event as it actually occurred in mind (Beck & Riggs, 2014). For example, a correct answer in Experiment 1a requires mentally rewinding the animation and then simulating what would have happened without the brick. The corresponding prediction in Experiment 1b is simpler because the brick is not present in the scene, the clip is paused, and it only requires children to simulate the future without the need to go back in time.

An alternative is that the wording of the question influenced children's performance. Notably, we found a pattern that aligns more closely with Rafetseder et al. (2013) than more recent work (Nyhout & Ganea, 2019; Rafetseder & Perner, 2018). One key difference between our study and that of Nyhout and Ganea (2019) is that the question in Nyhout and Ganea was "would [outcome] *still* [have happened]?" (emphasis added). While a systematic investigation is necessary, children may sometimes be answering on the basis of pragmatic cues: Why ask "would the outcome have been different" if the outcome was unchanged?

A second, not mutually exclusive alternative is that children *did* engage in simulation, but considered different counterfactual possibilities than adults did. We explore this possibility in Experiment 2 using a multiple-choice task modeled on Rafetseder and Perner (2018). We hypothesized that children may arrive at the wrong answer because they hold some of the features of the actual event constant, but allow other features to vary in ways that adults and older children do not. Based on pilot data, we predicted that children will specifically maintain the point of origin of a ball's movement from the event they saw, much as we would expect adults to, but allow the initial trajectory of the ball to vary, which we would not expect adults to do.

Methods

Participants We pre-registered (https://osf.io/qn3b9) a planned sample size of 24 participants in each of three age groups: 4-year-olds, 5-year-olds, and 6-year-olds. We therefore recruited 24 4-year-olds (15 female), 24 5-year-olds (7 female) and 24 6-year-olds (8 female). In addition, 6 4-year-olds (2 female) and 2 (female) 5-year-olds participated but were excluded due to failing to complete the study (4) or parental interference (3; see below). Participants were recruited from TheChildLab.com (Sheskin & Keil, 2018).

Stimuli and apparatus Children saw a total of ten trials in which featured animated events, and then still images representing what actually occurred in the animation, as well as four counterfactual possibilities (see Fig. 3; Full stimuli are available online at https://osf.io/5jw6y/).

Animated events were constructed using Flash, converted to a movie format, embedded in a PowerPoint presentation, and presented over a videoconferencing system. The animations were slightly modified from Experiment 1. This time, there was only one ball, resembling a soccer ball, and the brick wall was replaced with a triangular wedge with a wood texture. The background was made green with a white line to mimic a soccer field. The goal was turned into a grey rectangle, and there were no walls on either side of it.

We created a total of eight test animations and two training animations. In all test animations, the ball entered the stage from the right side and moved in a perfectly horizontal trajectory. In six of the test animations, the ball deflected off of the wedge, which did (4 animations) or did not (2) change whether it went into the goal. In two other test animations, the ball did not interact with the wedge, and simply moved across the field in a straight line.

Along with each test animation, we made a still image that showed the entire trajectory the ball had taken (center, Fig. 3), which was visible while the child was answering the counterfactual question, thus eliminating memory load. In addition, we constructed still images representing four counterfactual possibilities for each animation (Fig. 3). In these counterfactual possibilities, the wedge was removed, and the complete trajectory of the ball was shown as in the still image of the actual event. These four possibilities were constructed in systematic ways for the six items in which the ball interacted with the wedge.

- "**Correct**" (red): In this image, the ball is shown moving horizontally across the entire field, starting from the same point of origin that it had in the actual animation. In other words, it preserved both the origin and the initial trajectory of the ball.
- "**Match origin**" (yellow): The ball started from the same point of origin, but had a diagonal trajectory, ultimately ending up in the exact same place as the ball ended up in the actual event, in which it deflected off the wedge. This option preserved the origin but not the trajectory of the actual event.
- "**Match trajectory**" (purple): The ball originated from a y-coordinate that was level with where the ball *ended* in the actual animation, and the ball moved across the whole field in a perfectly horizontal trajectory. This option preserved the initial trajectory but not the origin of the actual event.
- "**Match neither**" (blue): The ball started from the same place as it did in the "match trajectory" image, but had a diagonal trajectory ending in the same place as the "correct" image, thus matching neither the point of origin nor the initial trajectory of the actual event.

For the events in which the ball and wedge did not interact, the four images still contained two options that preserved the origin and two that preserved the trajectory, but because the

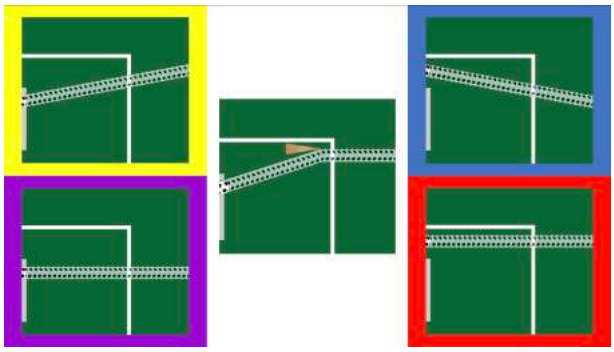


Figure 3. Example item from Exp. 2, as a child would see it. The center image is a rendering of the video the child just watched. On this trial, red is "correct", yellow is "match origin", purple is "match trajectory", and blue is "match neither".

ball did not deflect off the wedge in the actual event, the "match origin" and "match trajectory" images in fact showed the ball ending up in a location that was not present in the original event, while the "correct" and "match neither" images did. The model we used to analyze children's responses (described below) therefore does not apply to these images.

In addition, there were two training animations, one in which the ball bounced off the wedge and one in which it did not interact with the wedge. In both training animations, the ball entered on a diagonal trajectory. No still image of the event was presented in the center of the response screen, and in the still images for training items, the wedge was still present, as the training task was matching the *actual* event rather than considering a counterfactual one.

Procedure The script can be found in the presenter notes of the PowerPoint presentations at http://osf.io/5jw6y/

After parents gave informed consent, children were first shown the two training animations, and after each one asked to find the image that matched what they saw from the four possibilities. This was primarily to familiarize children with the multiple-choice response method. For test trials, children were asked "If there were no block on the field, how would the ball have moved?"

The experimenter was blind to what the child was seeing at all times, and only recorded the color that they said. Children's responses were then transcribed by another coder who was blind to condition, and later matched to images based on the condition the child had been assigned to (see data files in repository). There were two exclusion criteria: If the child failed to finish the study for any reason, or if the parent interfered in a way that guided the child toward a specific answer on any item, in the opinion of the experimenter or coder. As both were blind to what the child was seeing, these judgments could not be influenced by knowing what option the child was selecting.

Analysis plan We focused on the six test items in which the ball collides with the wedge. For those items, we used a multinomial processing tree (MPT) model (Riefer & Batchelder, 1988) to model the proportions with which the

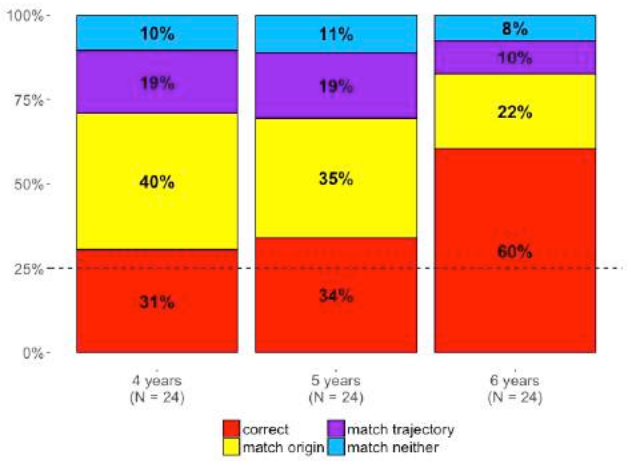


Figure 4. Results of Exp. 2. Proportion of responses is on the y-axis, and chance responding is indicated at 25%.

different age group chose the four possible response options, $P(C)$ (= "Correct"), $P(O)$ (= "match origin"), $P(T)$ (= "match trajectory"), and $P(N)$ (= "match neither"). Our model has three free parameters, $\boldsymbol{s}$, $\boldsymbol{m_o}$, and $\boldsymbol{m_t}$, which each represent the (conditional) probability of reaching a specific discrete cognitive processing stage (e.g., $\boldsymbol{s}$ = probability of engaging in simulation). In addition, our model allowed for the possibility of unbiased guessing.

The first parameter (s) represents the probability whether the children engage in simulation or not. If they do not (with probability $1–s$), we assume children simply make an unbiased guess for one of the four response categories (i.e., the conditional probability of choosing any one response category is .25). We assume that this will be unbiased as the multiple-choice question lacks the pragmatic demands of the questions used in previous studies. In case children engage in simulation (with probability s), we assume two further (unordered) processing steps: how likely they are to maintain the origin from the actual world in their simulation (parameter m_o), and how likely they are to maintain the trajectory (parameter m_t)? In order to examine this, we ignore the cases in which the ball does not interact with the block. For the remaining six cases, we can enumerate how the four different response categories follow from the assumed processes. For example, if children maintain both the origin and the trajectory (with probability $m_o \times m_t$), they will provide the correct response. If, however, children only maintain the origin, but not the trajectory (with probability $m_o \times (1 - m_t)$), they will choose the "match origin" response option, $P(O)$. Analogous arguments can be made for $P(T)$ and $P(N)$. Thus, the following model equations are assumed to hold:

$$P(C) = s \times m_o \times m_t + (1 - s) \times 0.25$$
$$P(O) = s \times m_o \times (1 - m_t) + (1 - s) \times 0.25$$
$$P(T) = s \times (1 - m_o) \times m_t + (1 - s) \times 0.25$$
$$P(N) = s \times (1 - m_o) \times (1 - m_t) + (1 - s) \times 0.25$$

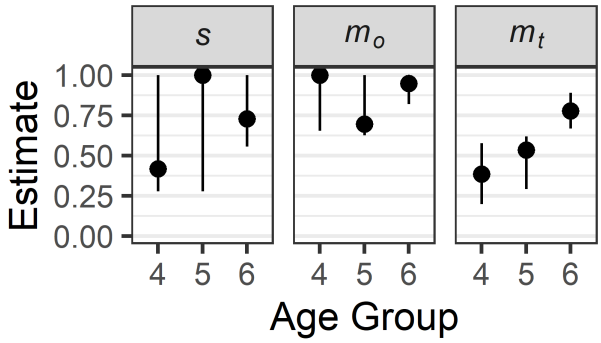


Figure 5. MPT model parameter estimates for s, m_o, and m_t in each age group. Error bars are bootstrapped 95% CIs.

To obtain estimates of the model parameters s, m_o, and m_t, we fitted the model to the aggregated data using maximum-likelihood estimation. This provides us with a model-based estimate of how likely each age group is engaging in simulation, and the likelihood, in each age group, of maintaining the origin and the trajectory of the actual world. Note that, although the model is saturated (i.e., three free parameters for three independent data points provided by the multinomial distribution with four categories), it cannot account for any possible data pattern. That is, our model imposes testable constraints on the data. For example, it predicts that, after accounting for the proportion of unbiased guessing, the conditional ratio of $P(C)/P(O)$ must be equal to the conditional ratio $P(T)/P(N)$. Therefore, finding that the model adequately accounts for the data (i.e., it fits the data), provides some evidence for the underlying assumptions and validity of the interpretations associated with the parameters.

Results and discussion

Fig. 4 shows how often children chose each of the four options for the six test items where the ball collided with the wedge. For the two cases in which the ball and wedge did not interact, the correct answer was the modal response in every age group (4-year-olds: 50%; 5-year-olds: 71%; 6-year-olds: 88%).

A visual inspection of the figure suggests a clear pattern when it comes to choosing the correct answer: Above-chance performance emerges around age 6. However, it also appears that, of the three possible incorrect responses, all age groups preferred "match origin" over "match trajectory" and "match neither", which suggests that the younger children are not just guessing randomly. Rather, they are simulating possibilities that maintain the origin but not the trajectory of the ball in the actual event.

To verify this impression, we fit our MPT model to children's responses. As our model was saturated, we used a double bootstrap procedure (van de Schoot, Hoijtink, & Dekovic, 2010) to evaluate model fit. This approach revealed a p-value of .04 ($G^2 = 3.48$) for the 4-year olds and .05 ($G^2 = 2.66$) for the 5-year-olds, suggesting that the main patterns in the data were well accounted for, but there was some misfit. Specifically, the model cannot predict both $P(C) < P(O)$ and $P(T) > P(N)$ at the same time, as was observed in the data. One possible reason for this misfit is individual differences in the simulation behavior of the 4- and 5-year-olds, such that some individual children consistently responded in a particular way and others did not. For 6-year-olds, the fit was perfect ($G^2 = 0$). Given the small magnitude of misfit, the model is interpretable, and we can evaluate the likelihood that children engaged in simulation, and how.

The parameter estimates for each parameter in each age group, with 95% confidence intervals estimated by parametric bootstrapping, can be seen in Fig. 5. In short, we find little evidence for developmental change in m_o or s, but a clear developmental increase in the estimate of m_t. Put in plain terms, this analysis suggests that 4- and 5-year-olds were not significantly different from 6-year-olds or each other in their likelihood of engaging in simulation, nor in how likely they were to choose an option that maintained the ball's point of origin from the actual event. However, 6-year-olds were significantly more likely than younger children to maintain the ball's initial *trajectory* from the counterfactual event. In addition, for 6-year-olds we have considerably smaller CIs for s, indicating we that we have higher certainty that they engage in simulation most of the time.

In short, children ages 4-5 do seem to engage in counterfactual simulation, and systematically hold constant some, but not all, features of the actual world in those counterfactual simulations, while allowing other features of the world (which older children hold constant) to vary.

General Discussion

In two experiments, we provide evidence that young children engage in counterfactual simulation, but do so in a different way than older children and adults. Experiment 1 validated the stimuli by replicating previous findings about children's ability to answer counterfactual questions and conduct predictive simulations, but in the domain of causal perception. Experiment 2 asked children to choose among four counterfactual trajectories rather than answering a simple yes/no question, and found that when 4-5-year-old children engage in simulation, they consider counterfactual possibilities in which the origin of an object's motion is preserved while its initial trajectory is allowed to vary, while 6-year-olds are more likely to preserve both features in their counterfactual simulations.

We consider these findings in the context of a general theory of children's reasoning put forward by Gopnik et al. (2017): When reasoning about different possibilities, children's hypothesis space may be quite different from adults, but the basic process of simulation could be very similar. In particular, this theory suggests that children have a broader and "flatter" hypothesis space (i.e., priors across all hypotheses are similar), in which they conduct a "higher-temperature" (i.e., broader) search. This theory can be readily applied to children's struggles with counterfactual reasoning: When considering counterfactual possibilities, children may be sampling from a broader set of possibilities, none of which

are favored over the others, and the way they pick possibilities out of this space is more random.

However, unlike Gopnik et al. (2017), we find that children are only showing evidence of this broader search space for certain specific features of these events. In other words, while our results align with the general proposal that children conduct simulations over a "flatter" hypothesis space, the space is only flatter over certain "dimensions" (i.e., components) of the events being considered. In this case, children are unlikely to consider possibilities that change where the ball enters from, but between 4 and 6 they narrow the search space for the initial trajectory of the object to be more like adults'.

This is a critical advance for understanding children's reasoning. We must not only test whether they are searching a broader space of possibilities in general, but also identify the separate features of that hypothesis space and determine which aspects of the event-structure are treated in an adult-like way (in this case the point of origin of the object's motion). Doing so will not only help us better understand children's reasoning processes, but allow us to predict specific challenges they face, or errors they will make.

One limitation is that we selected the range of possibilities for children to consider, and so there may be a possibility that we did not include which they would prefer over and above the ones they selected here. While verbal pragmatics are no longer a viable explanation, there are other possible explanations for children's responses that would not rely on simulation, such as path similarity, or some kind of contextual inference about the scenario, such as whether there is an agent launching the ball into motion.

This work provides an exciting new approach to the study of counterfactual reasoning in development. We should consider that "failure" in these tasks may result not from a failure to simulate *per se* but rather from different assumptions about what to hold constant and what to change when simulating counterfactuals.

References

Atance, C. M., & O'Neill, D. K. (2005). The emergence of episodic future thinking in humans. *Learning and Motivation*, *36*(2), 126-144.

Beck, S. R., & Riggs, K. J. (2014). Developing Thoughts About What Might Have Been. *Child Development Perspectives*, *8*(3), 175-179. doi:10.1111/cdep.12082

Byrne, R. M. J. (2016). Counterfactual Thought. *Annual Review of Psychology*, *67*, 135-157.

Gerstenberg, T., Peterson, M. F., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2017). Eye-Tracking Causality. *Psychological Science*, *28*(12), 1731-1744.

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2015). How, whether, why: Causal judgments as counterfactual contrasts. In *Proceedings of the 37th Annual Conference of the Cognitive Science Society*, Austin, TX, 2015 (pp. 782--787). Cognitive Science Society.

Gopnik, A., O'Grady, S., Lucas, C. G., Griffiths, T. L., Wente, A., Bridgers, S., . . . Dahl, R. E. (2017). Changes in cognitive flexibility and hypothesis search across human life history from childhood to adolescence to adulthood. *Proceedings of the National Academy of Sciences*, *114*(30), 7892-7899.

Harris, P. L., German, T., & Mills, P. (1996). Children's use of counterfactual thinking in causal reasoning. *Cognition*, *61*(3), 233-259.

Icard, T. F., Kominsky, J. F., & Knobe, J. (2017). Normality and actual causal strength. *Cognition*, *161*, 80-93.

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking: From childhood to adolescence.* (A. Parsons, S. Milgram, Trans.). New York, NY: Basic Books.

Kominsky, J. F., Strickland, B., Wertz, A. E., Elsner, C., Wynn, K., & Keil, F. C. (2017). Categories and Constraints in Causal Perception. *Psychological Science*, *28*(11), 1649-1662. doi:10.1177/0956797617719930

Leslie, A. M., & Keeble, S. (1987). Do six-month-old infants perceive causality? *Cognition*, *25*(3), 265-288.

Lewis, D. (1973). Causation. *The Journal of Philosophy*, *70*(17), 556-567.

Nyhout, A., & Ganea, P. A. (2019). Mature counterfactual reasoning in 4- and 5-year-olds. *Cognition*, *183*, 57-66.

Pearl, J. (2000). *Causality : models, reasoning, and inference*. Cambridge, U.K.; New York: Cambridge University Press.

Phillips, J., Luguri, J. B., & Knobe, J. (2015). Unifying morality's influence on non-moral judgments: The relevance of alternative possibilities. *Cognition*, *145*, 30-42.

Phillips, J., & Kominsky, J. F. (2017). *Causation and norms of proper functioning: Counterfactuals are (still) relevant*. Proceedings from Proceedings of the 39th annual meeting of the cognitive science society.

Qualtrics. (2005). [Computer Software]. Provo, UT: Qualtrics.

Rafetseder, E., & Perner, J. (2018). Belief and Counterfactuality. *Zeitschrift für Psychologie*, *226*, 110-121.

Rafetseder, E., Schwitalla, M., & Perner, J. (2013). Counterfactual reasoning: from childhood to adulthood. *Journal of Experimental Child Psychology*, *114*(3), 389-404.

Riefer, D. M., & Batchelder, W. H. (1988). Multinomial modeling and the measurement of cognitive processes. *Psychological Review*, *95*(3), 318–339.

Saxe, R., & Carey, S. (2006). The perception of causality in infancy. *Acta Psychologica*, *123*(1-2), 144-165.

Sheskin, M., & Keil, F. (2018). TheChildLab.com A Video Chat Platform for Developmental Research. Retrieved from psyarxiv.com/rn7w5

van de Schoot, R., Hoijtink, H., & Dekovic, M. (2010). Testing Inequality Constrained Hypotheses in SEM Models. *Structural Equation Modeling: A Multidisciplinary Journal, 17*(3), 443–463.

Uncertain evidence statements and guilt perception in iterative reproductions of crime stories

Elisa Kreiss (ekreiss@stanford.edu)
Department of Linguistics, Margaret Jacks Hall, Bldg. 460
Stanford, CA 94305 USA

Michael Franke (mchfranke@gmail.com)
Institute of Cognitive Science, Wachsbleiche 27
Osnabrück, Lower Saxony 49090 Germany

Judith Degen (jdegen@stanford.edu)
Department of Linguistics, Margaret Jacks Hall, Bldg. 460
Stanford, CA 94305 USA

Abstract

Transmission of information by means of language is a potentially lossy process. Especially adjunct information, such as the graded degree of evidence, is a piece of information that seems prima facie likely to be distorted by reproduction noise. To investigate this issue, we present the results of a two-step iterated narration study: first, we collected a corpus of 250 crime story reproductions that were produced in parallel reproduction chains of 5 generations in depth, for 5 different seed stories; a second separate large-scale experiment then targeted readers' interpretation of these reproductions. Crucially, strength of evidence for the guilt of each story's suspect(s) was manipulated in the initial seed stories. Across generations, readers' guilt perceptions decreased when the evidence was originally strong, but remained stable when evidence was originally weak. Analysis of linguistic measures revealed that dissimilarity between a seed story and its reproduction, story length, and amount of hedging language affected the readers' own guilt perception and the readers' attribution of guilt perception to the author differently. The results provide evidence that evidential information indeed influences guilt perception in complex ways.

Keywords: experimental pragmatics; iterated narration; transmission chains; uncertain evidence

Introduction

One of the central goals of language use is the exchange of information. New information is obtained by reading the newspaper, listening to a friend, etc., and often immediately communicated as stories to other people it may be relevant to. Yet this process of iterated reproduction is not innocuous: the original story may be distorted or altered by various sources of noise, including cognitive biases, memory reconstruction processes, or other limits on information processing capacity (Bartlett, 1932; Tversky & Marsh, 2000; Mesoudi & Whiten, 2004; Griffiths & Kalish, 2007; Hills, 2018). The game of Telephone is essentially a caricature of this process: the first person whispers a sentence to their neighbor, who in turn passes it on to the next person, and so on. The last person in the transmission chain announces the sentence they ended up with, which often differs remarkably from the initial seed story. This simple game nicely exemplifies the information loss and distortion that is associated with repeated exposure and reproduction of information.

Bartlett (1932) first introduced the methodology of transmission chains, i.e., chains of story reproductions, as a scientific method. In a series of transmission chain studies, using stories such as Native American tales or sport reports for reproduction, he observed a significant information loss in the stories over generations of reproductions. He also reported that the content of the reproduced stories increasingly aligned with the reproducing author's prior beliefs. Bartlett used these observations as a foundation for his theory of memory retrieval involving reconstruction processes.

In recent years, the transmission chain method has undergone a revival in cognitive and social psychology. Mesoudi and Whiten (2004) showed that with each iteration descriptions of everyday events, such as visits to a restaurant, became more abstract, in line with hierarchically organized script knowledge. Other research showed that reproductions can be influenced by cultural, racial and gender stereotypes (e.g., Kashima, 2000). The iterated transmission method has therefore also been used as a tool to investigate cognitive biases in general (e.g., Kalish, Griffiths, & Lewandowsky, 2007). In evolutionary linguistics, the transmission chain method has been used to study experimentally how iterated learning of a language exerts a selective pressure on language itself, so that learning biases create an indirect pressure on languages to be efficiently learnable (e.g., Scott-Phillips & Kirby, 2010; Kirby, Griffith, & Smith, 2014).

The transmission chain method thus presents an exciting opportunity for asking questions at the interface of linguistics and psychology. In particular, while previous studies have focused particularly on properties of the reproductions themselves, we here present an extension in which we investigate an external readership's interpretative perspective on the reproduced texts. We achieve this by a second experiment that uses as materials the output from the previous iterated transmission experiment. The stories used as seeds are five crime or ethical violation stories based on true events (animal smuggling, arson, sexual assault, beehive destruction, and email scams). Each seed started out with both a weak and a strong evidence version (see Table 1). This manipulation has successfully been used by (Van Prooijen, 2006) to uncover in-

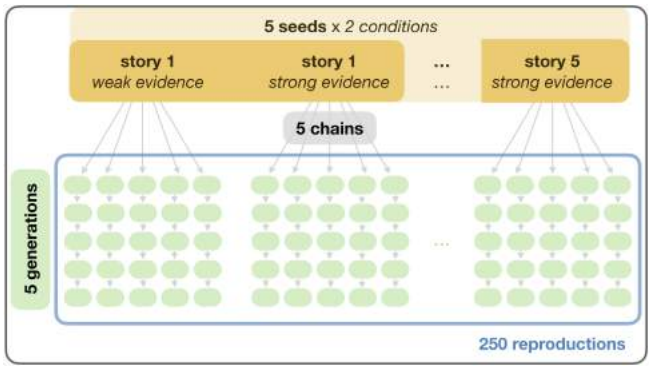


Figure 1: Overview of corpus of stories collected in Exp. 1.

and out-group effects in guilt judgments of suspects. Similarly to that study, the different conditions were implemented by adding a last sentence to each story that either suggested strong or weak evidence for the suspect's guilt. To evaluate the interpretations of the stories that readers arrive at, we collected answers to eight questions regarding the readers' perception of the stories' suspect(s), the readers guilt perception attributed to the author of the story, as well as, somewhat less importantly, indexes of more general author and reader related features, such as trustworthiness and subjective engagement.

Experiment 1: corpus collection

Methods

74 undergraduate students participated in this online study for course credit[1]. We constructed five stories (*seeds*) that marked the beginning of each reproduction chain. Stories were written in the style of short news articles and followed a similar structure. They reported a crime or moral rule violation that occurred, the authorities' determination of and search for the perpetrator(s), and the possible punishment the suspect(s) would face if found guilty. Furthermore, each of these five seed stories occurred in one of two conditions: a *weak evidence* and a *strong evidence* condition. Evidence strength was manipulated in the final sentence of the story (see example seed in Table 1).

Each participant read and reproduced five stories. For each story, they were either assigned to read and reproduce the seed story or continue an already started reproduction chain where they read and reproduced a reproduction from previous participants. The assignment was random. On each trial, participants first read a story. They were told to click the 'Continue' button when they were confident that they had internalized the story. Once they clicked the button, the story disappeared and they were asked to reproduce it freely in a text field. Order of stories was randomized.

[1]The current study was one several they could choose from.

Results

Participants produced 370 stories. For each seed, we defined a complete chain as one that has 5 reproductions/generations. For subsequent analysis, we randomly selected 50 complete chains, evenly distributed across stories and conditions. This yielded a corpus of 250 reproductions (5 seeds in 2 conditions with 5 complete chains each, see Figure 1). This was the maximal set of complete chains that was present in every condition for each seed. This corpus is a rich source of linguistic information which merits detailed investigation. Yet, with an eye to clear operationalizability, we focus here on a few general features, which we will subsequently use as predictors in the analyses of Exp. 2 below.

Proportion of hedges. As a proxy for vagueness, we extracted the number of hedges per story relative to its length. The seed stories were designed to contain various hedges, such as "nearly", "about", "up to" or "allegedly". As shown in Figure 2, the proportion of hedges decreased in each generation ($\beta = -0.01,\ SE = 0.00,\ t = -4.16,\ p < 0.0001$) , suggesting that participants portrayed the stories with more certain language over generations. There was no significant effect of evidence condition on proportion of hedges ($\beta = -0.00, SE = 0.00, t = -0.79, p < 0.44$) .

Story length. As shown in Figure 2, the number of words in a story decreased across generations ($\beta = -17.12,\ SE = 1.02,\ t = -16.79,\ p < 0.0001$), replicating a well-known phenomenon in reproduction studies (Bartlett, 1932). While the original seeds (generation 0) consisted on average of 159 words, that number dropped to 25 by generation 5. Examples of reproductions of the seed in Table 1 (strong condition) from generation 1 and 5 are shown in (1) and (2) below. There was no significant effect of evidence condition on story length.

(1) In late December 2017, a couple in Iowa went to check on their beehives. They found a tragic scene: their hives had been overturned and their equipment and facilities had been ransacked. A few weeks later, the police arrested a 12-y.o. and 13-y.o. for the crime. They are charged with multiple offenses, with fines up to $100,000 and up to 10 years in prison, yet will be tried as minors. The trial hasn't happened yet, but they seem guilty.

(2) A 12 and 13 year old were arrested for destroying a beehive, and face up to 10 years of jail time.

Similarity of seeds and reproductions. To assess the similarity between seed stories and their reproductions quantitatively, we computed the Jaccard distance between each reproduction and its generation 0 seed. Jaccard distance ranges between 0 and 1 (where 1 indicates greatest distance) and captures the amount of overlap between two stories in the following way:

$$D_J(X,Y) = 1 - \frac{|X \cap Y|}{|X \cup Y|}$$

where X is the set of words in the reproduction and Y the set of words in the respective original seed story. In this

Table 1: Example of a seed story used in Exp. 1.

In late December 2017, a couple in Iowa was checking on their 50 beehives when they discovered a tragic scene. The hives had been overturned and hacked apart, and the equipment had been thrown out of the shed and smashed. This destruction caused the death of about half a million bees and approximately $60,000 in property damage. Nearly three weeks later, police arrested two boys (12 and 13 years old) who, allegedly, were responsible for the damage. The charges against them include criminal mischief, burglary, and offenses to an agricultural animal facility. Since they are still minors, they will be charged in juvenile court where they face up to 10 years in prison and fines of up to $10,000 if convicted.	
(*strong evidence condition*) Police officials explained that the investigation is still in progress, but the evidence so far overwhelmingly speaks to the guilt of the suspects.	(*weak evidence condition*) Police officials explained that the investigation is still in progress, and the evidence so far doesn't warrant rushed conclusions about the guilt of the suspects.

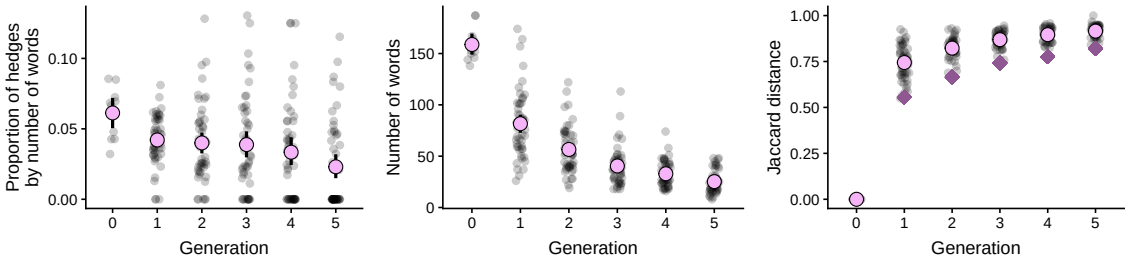


Figure 2: Mean of the three linguistic metrics (proportion of hedges (by number of words), number of words and Jaccard distance) over generations of reproductions. Error bars indicate bootstrapped 95% CIs. Pink dots indicate generation mean, gray dots are individual stories. The purple squares indicate the lowest possible distance given the mean length of the stories.

case, we took words as the basic unit over which distance was computed. Figure 2 shows that D_J increased across generations ($\beta = 0.08$, $SE = 0.01$, $t = 13.18$, $p < 0.0001$) . This is not surprising given that as the number of unique words decreases, D_J between seed and any of its reproductions necessarily increases. However, we will see later that length and D_J have different effects on story interpretation. There was no significant effect of evidence condition on Jaccard distance ($\beta = 0.00$, $SE = 0.02$, $t = -0.12$, $p < 0.91$).

In sum, in a corpus of 250 reproductions of 5 seed stories, the length of the stories, the similarity to the seed story, and the proportion of hedges decreases over generations, regardless of the initial evidence strength condition.

Experiment 2: story ratings

In order to assess the extent to which, as a function of the originally provided evidence, the generation of reproduction affects readers' interpretation of various features of the stories we collected judgments from a second group of independent participants. We were particularly interested in features related to the uncertainty of presented evidence and the associated judgments of suspect guilt. We also collected judgments concerning the readers' general attitude towards the author and the story.

Methods

5392 participants were recruited over Amazon Mechanical Turk. Each participant read one story from the 250 story corpus reported in the previous section, and answered twelve questions about the story (including four attention checks). They indicated their response by moving a slider on a continuous scale (slider endpoints were coded as 0 - 100). Each question was shown in isolation in a randomized order. Participants spent on average two to three minutes on this experiment and were paid $0.60 ($12-$18 per hour). The story was visible throughout the experiment.

The list of questions asked is provided in (3) to (10). Questions (3)–(7) assessed the extent to which the reader believes the suspect(s) is/are guilty of the alleged crime. Questions (8)–(10) assessed the reader's trust in the author, the extent to which they considered the story to be objectively written, and the extent to which they felt emotionally connected to the story. Overall, participants were asked eight questions of interest and four attention check questions designed to filter out participants who were just clicking through the experiment.

(3) **Strength of evidence:** How strong is the evidence for the suspect's / suspects' guilt?

(4) **Suspect guilt:** How likely is it that the suspect is / the suspects in the crime are guilty?

(5) **Suspect conviction:** How likely is a conviction of the suspect(s) in the crime?

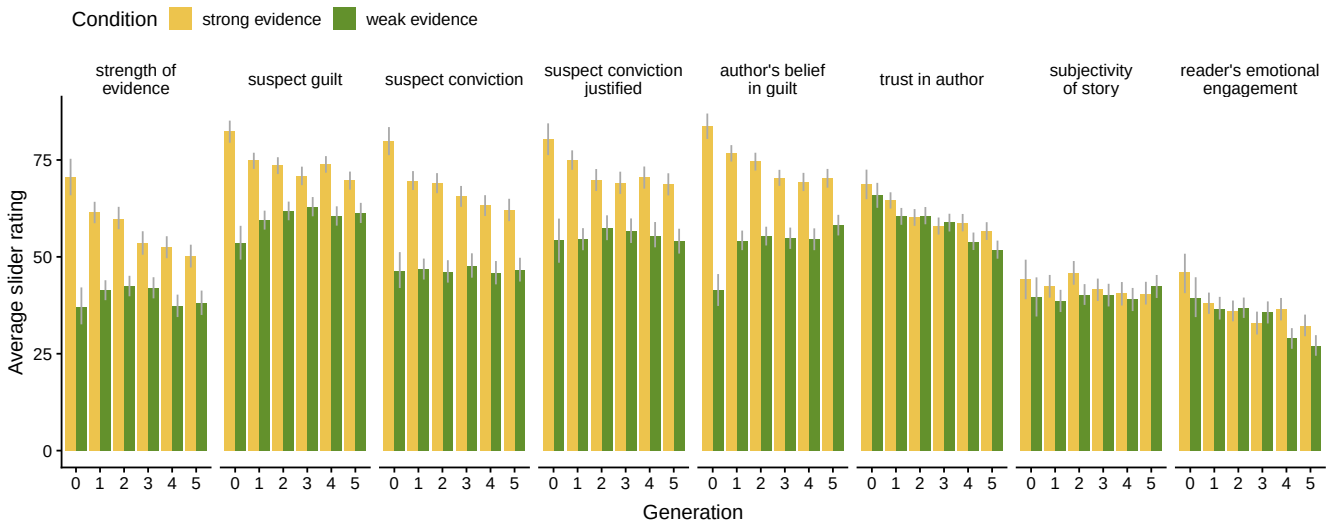


Figure 3: Mean ratings in strong (yellow) and weak (green) evidence condition for each dimension (facets).

(6) **Suspect conviction justified:** How justified would a conviction of the suspect(s) in the crime be?

(7) **Author's belief in guilt:** How much does the author believe that the suspect is guilty?

(8) **Trust in author:** How much do you trust the author?

(9) **Subjectivity of story:** How objectively / subjectively written is the story?

(10) **Reader's emotional engagement:** How affected do you feel by the story?

Results

Exclusions. We excluded 107 submissions because the participants could not be uniquely identified due to a data submission error. Furthermore, 12 participants were excluded because they completed the study multiple times (177 submission exclusions) and another 535 participants because they failed at least two of the attention check questions. This left us with 4573 participants (84.8% of the original set). After exclusions, each reproduction received on average 17 ratings, ranging from 9 to 22 and two outliers with 27 and 38 ratings. The original seed stories received between 25 and 31 ratings.

Analysis procedure. Of main interest was whether participants give judgments of suspect guilt in line with the originally provided evidence, whether those judgments change over generations, and whether those judgments pattern with related measures of evidence for guilt, probability of conviction, justification of conviction, and attributed suspect guilt (i.e., an estimate of the author's belief in the suspect's guilt). We refer to these measures as *guilt related measures* (questions 3–7). Additionally, we analyzed trust in the author, story subjectivity, and emotional engagement as measures of secondary interest (questions 8–10). Mean slider ratings corresponding to the analyses are shown in Figure 3. Judgments were analyzed using linear mixed effects models. For each question, slider rating was predicted from fixed effects of generation, condition (reference level: strong), and their interaction. The models also included random by-story intercepts. An overview of the results is shown in Table 2. Each row presents the model results for one of the questions and the columns show the model outcomes for fixed effects.

Generation and evidence strength effects on guilt related measures. We observed main effects of condition on all guilt related measures (see first 5 panels of Figure 3) , such that ratings in the strong condition were higher than ratings in the weak condition, suggesting that participants in Exp. 1 were sensitive to the evidence strength manipulation (reproducing stories in such a way that evidence strength information was maintained); and also suggesting that participants in Exp. 2 were sensitive to the reproduced evidence strength information in their judgments. We also observed significant or marginally significant interactions between condition and generation for all but one of the guilt related measures, such that ratings decreased across generations in the strong condition but remained stable in the weak condition.

Generation and evidence strength effects on secondary measures. The secondary measures look very different from the guilt related measures. In particular, there were no significant effects of evidence strength condition on any of the measures with the following exception: stories were rated as less subjective in the weak evidence condition in earlier generations, though subjectivity ratings did not vary as a function of generation and remained on the 'objective' side of the scale throughout. In contrast, both trust in the author and readers' emotional engagement with the story decreased across generations. This is presumably the result of the stories becoming shorter over generations (see Exp. 1) and readers therefore having less material to be emotionally affected by, and less

Table 2: Model output for each fixed effect (condition, generation, and their interaction) for each rated question (rows).

	condition			generation			condition*generation		
	β	SE	p	β	SE	p	β	SE	p
strength of evidence	-23.25	4.09	<0.0001***	-3.42	0.89	<0.001***	2.59	1.26	<0.05*
suspect guilt	-17.28	3.40	<0.0001***	-1.34	0.74	<0.08	1.90	1.05	<0.08
suspect conviction	-27.01	4.15	<0.0001***	-2.79	0.90	<0.01**	2.74	1.28	<0.05*
suspect conviction justified	-19.02	4.35	<0.0001***	-1.69	0.95	<0.08	1.43	1.34	<0.29
author's belief in guilt	-27.53	3.72	<0.0001***	-2.14	0.81	<0.01**	3.42	1.15	<0.01**
trust in author	-0.82	2.25	<0.72	-1.94	0.49	<0.001***	-0.54	0.70	<0.44
subjectivity of story	-6.12	2.21	<0.01**	-0.86	0.48	<0.08	1.40	0.69	<0.05*
reader's emotional engagement	0.85	2.99	<0.78	-1.49	0.65	<0.05*	-1.11	0.92	<0.24

material to build trust in the author on in later generation stories.

Preliminary discussion. It seems prima facie plausible that trust in the author, the subjective quality of the story, or the reader's emotional engagement with the story are important factors in readers' assessment of the described suspects' guilt. But the presented data suggest otherwise. The guilt related measures are only weakly correlated with the secondary measures (maximum correlation: $r = 0.30$, minimum correlation: $r = 0.01$, mean correlation: $r = 0.12$). The evidence strength effect was expected, given the strong manipulation in the final sentence of the seed story. However, what it is that changes over generations that affects the guilt related measures in the strong and weak evidence conditions differently merits further investigation. The change across generations is presumably driven by the content of the stories. We next report a second set of analyses in which we assess the extent to which the linguistic features reported in Exp. 1 predict ratings in Exp. 2, focusing on the readers' assessment of suspect guilt and of attributed suspect guilt.

Effects of linguistic features on suspect guilt and author belief in suspect guilt. In this part of the analysis, we focus on the measures of *suspect guilt* and *author's belief in guilt* (attributed suspect guilt). These measures are interesting to examine in more detail because a) suspect guilt is the main issue raised in the 5 seed stories, so it is relevant to understand the linguistic conditions that lead to changes in perceived guilt; and b) while there is no obvious reason why readers' ultimate beliefs and the beliefs they ascribe to the author *should* differ after reading these stories, Degen et al. (2019) showed that listeners maintain uncertainty about the state of the world even when they ascribe a strong belief to speakers. In the following, we therefore analyze for both measures the effect of the proportion of hedges in a story, story length, and dissimilarity between a story and its seed.

Results are shown in Figure 4. In order to analyze the effects of proportion of hedges, story length, and Jaccard distance on the two guilt measures of interest, we asked whether the linguistic features explained variance above and beyond generation. To assess this, we first residualized each feature against generation, due to the substantial correlations between the features and generation observed in Exp. 1. The final mixed effects linear regression models predicted slider rating for each of the two measures and each of the three linguistic features of interest from main effects of evidence strength condition, residualized linguistic feature, generation, the interaction between evidence strength condition (reference level: strong) and generation, and the interaction between evidence strength condition and residualized linguistic feature.[2]

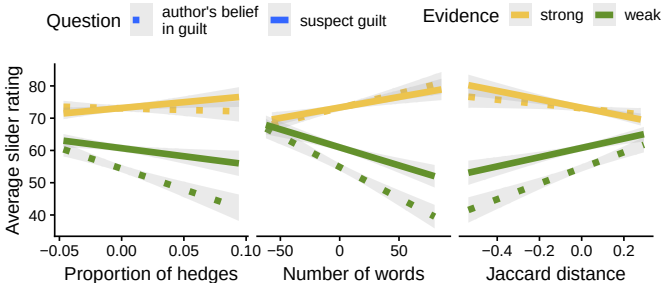


Figure 4: Linearly smoothed mean slider ratings as a function of generation-residualized proportion of hedges (left), number of words (middle), and Jaccard distance (right). Suspect guilt ratings shown in solid lines, author belief in suspect guilt ratings shown in dashed lines. Gray ribbons indicate 95% confidence intervals.

We observed significant interactions between evidence strength condition and generation-residualized linguistic feature for two of the three linguistic features on the suspect guilt measure (hedge proportion: $\beta = -79.57$, $SE = 54.05$, $t = -1.47$, $p < 0.15$, story length: $\beta = -0.18$, $SE = .06$, $t = -2.93$, $p < .01$, Jaccard distance: $\beta = 29.52$, $SE = 10.82$, $t = 2.73$, $p < .01$) and for all three linguistic features on

[2]Nested model comparison revealed that the inclusion of the residualized linguistic feature fixed effect was justified for all linguistic features on both measures, with the exception of hedge proportion when used to predict suspect guilt.

the attributed suspect guilt measure (hedge proportion: $\beta = -115.82$, $SE = 58.46$, $t = -1.98$, $p < .05$, story length: $\beta = -0.30$, $SE = .07$, $t = -4.38$, $p < .0001$, Jaccard distance: $\beta = 33.43$, $SE = 11.80$, $t = 2.83$, $p < .01$) .

To further understand these interactions, we conducted a simple effects analysis. The analysis revealed that there was no evidence of an effect of the linguistic metrics in the strong evidence condition for either of the two guilt related measures. However, in the weak evidence condition, as can be seen in Figure 4, increased number of words in a story was associated with a decrease in both suspect guilt and attributed suspect guilt. An increased proportion of hedges was also associated with a decrease in attributed suspect guilt ratings. Conversely, increased Jaccard distance was associated with an increase in both guilt related measures. Inspecting the beta coefficients suggests that these effects were always stronger for attributed suspect guilt than for suspect guilt.

These results suggests that when evidence for suspects' guilt was initially strong, that evidence was carried through the stories despite changes in hedge proportion, story length and increased dissimilarity between the retelling and respective seed story, and therefore did not change ratings of suspect guilt and attributed suspect guilt. In contrast, when evidence was weak, participants were more likely to believe that suspects were guilty, even while recognizing that the author was less likely to believe so. This discrepancy increased with increased proportion of hedges, increased number of words, and increased similarity to the original story[3].

General discussion

In this work we investigated the effects of lossy transmission on readers' interpretation of crime stories under varying initial evidential conditions in an iterated narration paradigm. First we constructed a corpus of 5 original seed stories in 2 conditions of evidential strength for a suspect's guilt, and 250 reproductions thereof. This corpus replicates previously found effects of a decrease in story length over generations of reproductions (Bartlett, 1932). Furthermore, the stories become less similar to the original seed story and the proportional number of hedges decreases.

We here introduced a, to our knowledge, new experimental extension of the transmission chain paradigm, where we subjected the text reproductions from the first study to a second empirical study focusing on the interpretative effect of the reproductions on an independent set of readers. In this way, we obtained ratings for each story on 5 guilt related measures and 3 secondary measures regarding trust in the author, story subjectivity, and the reader's emotional engagement. Our results suggest that, for one, the subtle manipulation of varying evidential strength in the original seed stories did have a lasting effect on reproductions and subsequent judgments of guilt, lasting several generations. For another, manipulation of evidence did not seem to have an effect on the readers' perception of the trustworthiness of the author, the subjectivity of the story or the general engagement readers had with the story. This is partially surprising because it seems naïvely plausible that providing weaker evidence could lead to less trustworthiness and more subjectivity. However, if reproductions are convincing and weak evidence is presented appropriately, there is no need to assume that the author is not trustworthy or the text more subjective.

We also observed effects of generation on especially the guilt related measures, which we found to be attributable to the ways in which the reproduced stories changed across generations. We found that, contrary to pessimistic expectations, it did not appear to be the case that repeated reproductions of stories with nuanced degrees of evidential information would have dropped these nuances, e.g., to arrive at a black-and-white picture, which would have been reflected in floor and ceiling slider ratings on the guilt related measures. Instead of increasing ratings for strength of evidence in the weak evidence conditions, we rather see a decline of perceived evidential strength over generations in the strong evidence condition. Reproducers seemed to have been rather careful in their formulations, despite the observed decrease in the proportion of hedges.

Most interesting is the relationship between readers' belief in the suspects' guilt and the belief they attributed to the author of the story in this regard. When the presented evidence was strong, these judgments aligned. However, when the evidence was weak, they diverged. Crucially, participants believed that the suspect was more likely to be guilty compared to the belief they attributed to the author. It is worth noting that the stories did not contain any information about the author and none of the reproductions contained first person narrations. Therefore, readers did not receive direct evidence to support the idea that the author's beliefs should differ from the presented view. The more dissimilar the reproductions were from the original story, the more participants' beliefs aligned with the beliefs they attributed to the author. They converged onto the highest guilt ratings in the weak condition. The difference between believed and attributed suspect guilt was greatest for large proportions of hedges in the story – this difference disappeared for small proportion of hedges. This suggests that, surprisingly, rather than hedges affecting readers' beliefs about suspect guilt directly, they instead lead only to readers attributing a weaker belief in suspect guilt to the author. In essence, readers are less willing to commit to beliefs that were communicated via hedging language.

We see the main achievements of this work in the contribution of an interestingly structured text corpus, with rich empirically obtained information on readers' assessments of the individual texts. This dataset will enable more detailed linguistic analyses in future work,[4] which will look more closely

[3]These three linguistic features are highly correlated and future research needs to investigate to what extent each of them contributes to these differences in guilt ratings.

[4]The corpus will become publicly available by January 2020 as part of a Github repository at https://github.com/elisakreiss/iteratednarration.

at the more specific contribution of different types of hedges and other types of constructions that signal information about graded evidence.

References

Bartlett, F. C. (1932). Remembering: An experimental and social study. *Cambridge: Cambridge University*.

Degen, J., Trotzke, A., Scontras, G., Wittenberg, E., & Goodman, N. D. (2019). Definitely, maybe: A new experimental paradigm for investigating the pragmatics of evidential devices across languages. *Journal of Pragmatics*, *140*, 33–48.

Griffiths, T. L., & Kalish, M. L. (2007). Language evolution by iterated learning with bayesian agents. *Cognitive science*, *31*(3), 441–480.

Hills, T. T. (2018). The dark side of information proliferation. *Perspectives on Psychological Science*.

Kalish, M. L., Griffiths, T. L., & Lewandowsky, S. (2007). Iterated learning: Intergenerational knowledge transmission reveals inductive biases. *Psychonomic Bulletin & Review*, *14*(2), 288–294.

Kashima, Y. (2000). Maintaining cultural stereotypes in the serial reproduction of narratives. *Personality and Social Psychology Bulletin*, *26*(5), 594–604.

Kirby, S., Griffith, T., & Smith, K. (2014). Iterated learning and the evolution of language. *Current Opinion in Neurobiology*, *28*, 108–114.

Mesoudi, A., & Whiten, A. (2004). The hierarchical transformation of event knowledge in human cultural transmission. *Journal of Cognition and Culture*, *4*(1), 1–24.

Scott-Phillips, T. C., & Kirby, S. (2010). Language evolution in the laboratory. *Trends in Cognitive Sciences*, *14*(9), 411–417.

Tversky, B., & Marsh, E. J. (2000). Biased retellings of events yield biased memories. *Cognitive psychology*, *40*(1), 1–38.

Van Prooijen, J.-W. (2006). Retributive reactions to suspected offenders: The importance of social categorizations and guilt probability. *Personality and Social Psychology Bulletin*, *32*(6), 715–726.

Belief dynamics extraction

Arun Kumar
University of Minnesota
Minneapolis, MN 55455 USA
kumar250@umn.edu

Zhengwei Wu
Baylor College of Medicine
Houston, TX 77030 USA
zhengwei.wu@bcm.edu

Xaq Pitkow
Rice University, Baylor College of Medicine
Houston, TX 77030 USA
xaq@rice.edu

Paul Schrater
University of Minnesota
Minneapolis, MN 55455 USA
schrater@umn.edu

Abstract

Animal behavior is not driven simply by its current observations, but is strongly influenced by internal states. Estimating the structure of these internal states is crucial for understanding the neural basis of behavior. In principle, internal states can be estimated by inverting behavior models, as in inverse model-based Reinforcement Learning. However, this requires careful parameterization and risks model-mismatch to the animal. Here we take a data-driven approach to infer latent states directly from observations of behavior, using a partially observable switching semi-Markov process. This process has two elements critical for capturing animal behavior: it captures non-exponential distribution of times between observations, and transitions between latent states depend on the animal's actions, features that require more complex non-markovian models to represent. To demonstrate the utility of our approach, we apply it to the observations of a simulated optimal agent performing a foraging task, and find that latent dynamics extracted by the model has correspondences with the belief dynamics of the agent. Finally, we apply our model to identify latent states in the behaviors of monkey performing a foraging task, and find clusters of latent states that identify periods of time consistent with expectant waiting. This data-driven behavioral model will be valuable for inferring latent cognitive states, and thereby for measuring neural representations of those states.

Keywords: Belief dynamics; Foraging; Partially observable switching semi-Markov process; Animal behavior

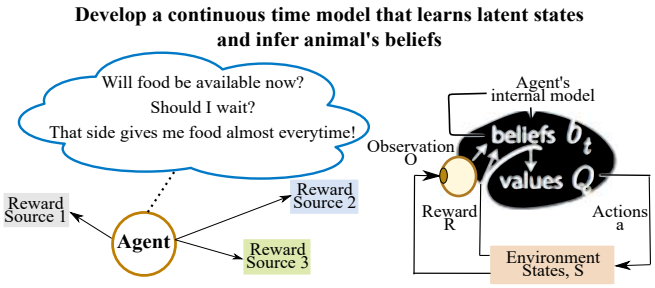


Figure 1: *Overview*: In complex natural tasks such as foraging, an animal faces a continuous stream of choices. Some of the choices pertain to hidden variables in the world, such as food availability at a given location and time. These variables determine time- and context-dependent rates for observation events and rewards. To perform well at these tasks, animals must learn these hidden rates and act upon what they have learned. Our goal is to develop a data-driven, continuous-time model for inferring an animal's latent states and their dynamics.

Introduction

An animal's survival depends on effective planning for future costs and rewards. One of the most fundamental purposes of the brain is to create and execute such plans. However, these plans cannot be directly observed from behavior. To understand how the brain generates complex behaviors and learn how an animal builds a representation of the surrounding environment, it is valuable to construct hypotheses about the brain's internal states that narrow the search space for neural implementations of planning. These hypotheses often come from models of the task implemented as artificial agents, whose internal state representations provided a latent space. However, differences between the model task and agent and the real task and animal create the potential for severe model-mismatch, injecting unknown biases into scientific conclusions. Here we use a latent-variable model to impute latent behavioral states based on observed behavior directly, using a data-driven latent-variable analysis that is designed to match the dependency structure of agent-based models without enforcing parametric structure.

To understand the mechanisms underlying behaviors, it is crucial to study hard tasks that involve inferring latent variables, since only then will an animal need to create a mental model of the world; otherwise the animal could perform well simply by responding to its immediate sensory input. Naturalistic foraging is one such task where an agent has to make decisions from many difficult choices in an uncertain environment. When foraging, an animal must take actions to procure rewards, and these actions have costs. How the animal schedules its actions determines the balance between total costs and rewards, Charnov & Orians (2006). The animal's goal in foraging is to use its energy resources for short term and long term sustenance. Decisions must be made continuously, and therefore time is a key ingredient in foraging: An animal benefits from tracking *when* reward is likely accessible at different locations. A natural way to represent such temporal quantities is in terms of dynamic event rates. For this reason, our work highlights the continuous-time aspects of decision problems.

Fig 1 illustrates our motivation for the foraging problem. An agent develops an internal model and takes an action, which may result in a reward. As a result, the agent updates its internal model in an attempt to learn the environmental dynamics. We explore the plausibility that an animal's internal states in

continuous time manifest as measurable consequences on its behavior, using a switching hidden semi-Markov model, and demonstrate the model's applicability in inferring latent states on a foraging task.

In the remainder of the document, we provide background, discuss the presented model and procedure followed by the experiments, results and discussion.

Background

Behavior identification using computational models has a rich history, and clear value–the ability to learn rich representations of behavioral constituents provides important insights into underlying neural processes which can also be incorporated into the development of artificial agents (Anderson & Perona (2014)). Early behaviorists explored behavioral sequences in an attempt to learn determining causal factors underlying behavior, aiming to explain effects like when an agent switches to an alternate choice. These approaches are still common in animal ecology, where hidden Markov time series models (HMMs) have been used to analyse animal's internal states Nathan et al. (2008); Langrock et al. (2012). Macdonald & Raubenheimer (1995) proposed using HMMs to capture causal structure in putative motivational states. However, they also observed that there are no one-to-one correspondences between the learned states and behavior, and Zucchini et al. (2008) found that behavior also influences internal states through feedback, challenging the dependency structure assumed by HMMs. To capture non-stationarity in behavior, Li & Bolker (2017) use temporally varying transition probabilities to model animal movement. However, behavior identification has struggled to produce more than a description of the behavior, with unknown relationships between the elicited latent states and the animal's representations. These failures are less surprising when it's realized the behavior expressible by HMMs is incompatible with key characteristics of observed behavior.

In these works and others, an important question left unanswered is what kind of latent belief states could be inferred that not only represent belief dynamics but also the choices that an animal or an agent makes. We attempt to uncover latent state beliefs in a continuous time model and apply it to a complex ecological process, foraging, which has multiple underlying sub-processes including satisfaction of needs, searching for alternatives, motivation, decision making, and control. We show that by generalizing allowing action-dependent transitions and more complex temporal dynamics, we can capture the expressivity of artificial agents designed for these domains, and highly interprable representations from animal behavior.

Model

Ecological behavior in animals is often well characterized by quick transitions between discrete behavioral modes. These transitions are difficult to predict from external events, and instead reflect a shift of the animal's internal state based on integrating events over a longer time scale. A process with quick transitions separated by long inter-event intervals can be approximated by a discrete-time hidden Markov process involving transition probabilities, but many of the probabilities (those for which the state is unchanged) will be close to one, while the remaining probabilities will be very small and decrease with the discrete time scale. Instead, we expect there will be advantages in treating these latent dynamics in *continuous time*, based on rates or time intervals between transitions and events.

A natural model to account for these point-like transitions in continuous time is the semi-Markov Jump Process, Rao & Teh (2013). This process is a simple but powerful class of continuous-time dynamics featuring discrete states that transition according to a generator rate matrix, producing rich and flexible timing that is potentially better matched to animal behavior. In contrast, times of transitions between states in a Markov process are exponentially distributed, which describe animal behavior poorly.

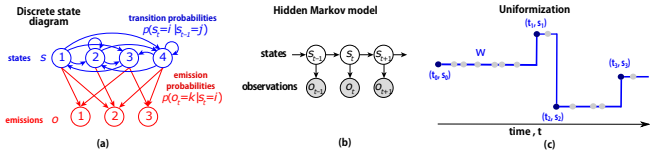


Figure 2: A discrete-state Hidden Markov Model. *a*: Discrete state diagram shows latent states (blue circles) and their transitions (blue lines), as well as the possible emissions from each state (red circles) with their emission probability (red lines). *b*: Directed probabilistic graphical model showing dependence of state variable s_{t+1} and observation o_t on the previous state s_t. *c*: We present a continuous-time extension for latent states and discrete time observations using uniformization, Rao & Teh (2013)

However, agents who control their environment affect transition rates through their actions, which means a single generator rate matrix is not sufficient to model behavior. An important example are Belief MDPs, which is a representation of a Partially Observable Markov Decision Process (POMDP, Kaelbling et al. (1998)). POMDPs are a model for inference and control when sensory measurements provide only partial observations of the state of the world. Belief MDPs have distinct transition matrices that update beliefs differently for each action. Action-dependent transitions imply that a standard semi-Markov model with a single transition generator is not expressive enough to match action-dependent belief dynamics.

To allow for action-dependent belief dynamics, we propose a switching semi-Markov (SMJP) model that matches an agent's belief dynamics by switching its generator depending on the action a: $A_{s'|s,a}$. Let $s \in \mathcal{S}$ be a discrete latent state, and $A_{s'|s}$ be an $N \times N$ generator rate matrix that can be interpreted as an instantaneous transition matrix $A\,dt = P(s'(t+dt)|s(t))$. This generator defines a point process that jumps from state s to s' at time t according to the time-dependent matrix $P_t = \exp(At)$. The process can be implemented by sequen-

tially sampling a time $t_i(s_i)$ from the total rate leaving state s_i, followed by sampling a new destination state s' according to the matrix $P_{t_i(s_i)}(s'|s_i)$ evaluated at this sample time (Gillespie's algorithm Gillespie (1977)). An analogous process occurs for the generation of observable events o, through the emission generator matrix $B_{o|s}$. The resulting process is similar to a simple Markov process, except that the time between transitions is stochastic and depends on the starting state (but not the end state), illustrated in Fig 3; the animal's behaviors and decision making are continuous, albeit partially observable only at discrete recording times.

The Markov Jump Process extends discrete time Markov processes in continuous time. Rao & Teh (2013) introduced Markov chain sampling methods that simplify structures by introducing auxiliary variables. We adapt jump structures to provide a continuous-time representation for the free foraging task and the trajectory is introduced using a generator matrix. Let $A \in \mathbb{R}^{N\times N}$ be the generator matrix, which is skew symmetric and negative diagonal entries. We can represent $P_t \in \mathbb{R}^{N\times N}$ as continuous-time transition matrix given by $P_t = \exp(At)$, $B_t \in \mathbb{R}^{N\times N}$ as discrete time transition matrix that is induced by *uniformization*, and $L_t \in \mathbb{R}^{N\times|O|}$ as observation matrix $P(O|s)$.

Uniformization instantiates the Markov Jump Process as a sequence of discrete time transition matrices (Fig 2), by introducing a latent sequence of random times that are adapted to the process generator but occur at a rate $\Omega \geq \max_s A_s$. For each interval, a random discretization vector of sampled times is $W = [w_1, w_2, ..., w_n]$, and we impute sampled times for a trajectory. Using this notation, we sample both random times as a Poisson process with intensity Ω and states using the generator matrix. The hidden Markov model characterizes a sample path of a piecewise constant stochastic process over these sampled and event times as (s_0, S, T) where T is now an ordered union of event times and randomly sampled discretized times. The chain can jump from a state to the same state or any other state, while the emissions are observed only at certain specified times. Since we sample intervals with these virtual jump times, the constructed process represents the same chain.

To learn the discrete time transition matrix B and emission matrix L, we consider an ensemble of sample sequence of observed emissions as generated from an HMM, and update the matrices using an EM algorithm to best account for the available observations. However, if we sample discrete times once, the estimates would be biased, so we resample latent trajectories repeatedly and randomly based on uniformization. The learned B matrix is then used to update the generator matrix using the relation $A_{\text{new}} = (B_{\text{new}} - I)\Omega_{old}$ while preserving its structure, and the random times are resampled to adapt to the modified A_{new}. The resulting algorithm exploits uniformization to enable learning the generator via an EM-algorithm, which is orders of magnitude more efficient than Gibbs sampling.

Belief MDPs are a convenient representation for POMDPs that treats current beliefs (posterior probabilities) over partially observable world states as fully observable. Agents following a Belief MDP exhibit transitions between beliefs $b_{t+1} = f(b_t, a_t, o_t)$, take actions according to a policy $\pi(a_t|b_t)$, and expect observations according to their beliefs via $p(o_t|b_t)$ (Fig 3). The proposed SMJP model matches the agent's action-dependent belief dynamics by switching its generator conditional on the action a: $A_{s'|s,a}$. To infer the agent's model from experimental observations, we develop an EM algorithm to infer it's parameters. When applied to our switching model, the forward α, backward β and update ξ equations of hidden Markov model, Rabiner (1989), can be written as:

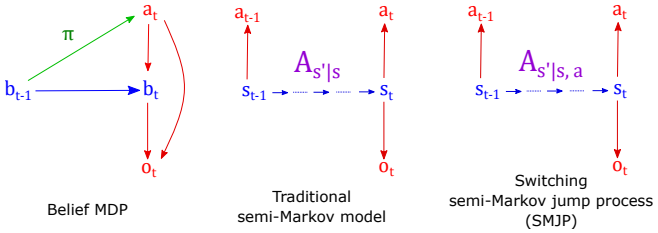


Figure 3: Comparison of graphical models of behavior. *Left*: In Belief MDP, belief transitions depend on actions selected by a policy. *Center*: Transitions in Semi-Markov Jump Process are independent of actions. *Right*: The Switching SMJP allows transition rates to depend on actions.

$$\alpha^{k'}_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha^{k}_{t}(i) B^{k}_{ij}\right] L_j(o_{t+1}); \quad (1)$$
$$1 \leq t \leq T-1; 1 \leq j \leq N; 1 \leq k, k' \leq K$$

$$\beta^{k}_{t}(i) = \sum_{j=1}^{N} B^{k}_{ij} L_j(o_{t+1}) \beta^{k'}_{t+1}(j); \quad (2)$$
$$t = T-1, T-2, ..., 1; 1 \leq i \leq N; 1 \leq k, k' \leq K$$

where k, k' are the action switching indices at time t and t+1 respectively. We adjust the model parameters to maximize the probability of the observation sequence given the model and train using EM. Updates are made using the ξ variable, which is the probability of being in state i at time t and state j at time $t+1$, and is given as

$$\xi^{k}_{t}(i,j) = \frac{\alpha^{k}_{t}(i) B^{k}_{ij} L_j(o_{t+1}) \beta^{k'}_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha^{k}_{t}(i) B^{k}_{ij} L_j(o_{t+1}) \beta^{k'}_{t+1}(j)}; \quad (3)$$
$$1 \leq t \leq T-1; 1 \leq i, j \leq N; 1 \leq k, k' \leq K;$$

The usual semi-Markov model is a special case of the switching semi-Markov model where the generator remains the same without action dependent switching. Our model is a switching model that changes rate, transition and emission matrices in accordance with the action taken by the agent. We learn the model using an EM approach , updating model parameters given transition times sampled by the uniformization

Algorithm 1 Switching semi-Markov jump model

Initialization
$O_f \leftarrow$ pre-processing ▷ actions are observed with events
O_{tr}, $O_{\text{te}} \leftarrow$ train and validation sequences from O_f
B', $L' \leftarrow$ switchHMM(O_{tr}, B, L, criteria)
compute A = GeneratorUpdate(B', B), $B = B'$, $L = L'$
repeat
  Training
  $O \leftarrow$ TrajectorySampling(A, L, O_{tr})
  B', $L' \leftarrow$ switchHMM(O, B, L, criteria)
  $B = B'$, $L = L'$
  Validation
  $ll_{\text{te}} \leftarrow P(O_{\text{te}}|B,L)$
  recompute A = GeneratorUpdate(B', B)
  validate structure of A
▷ Make updates in the generator space
until ll_{te} stops changing or max iterations reached

Figure 4: Overview of the algorithm.

process, and resampling the transitions given the new model parameters.

Procedure

We provide a brief description of the procedure, illustrated in Fig 4, that consists of pre-processing, initialization, training and validation steps. The overhead video, lever press and reward time sequences were used to set up observations and actions sequence required for training and validation. We processed the video recording using blob tracking to estimate position and velocity. Estimated positions and velocities were then clustered using k-means to assess different locations. By matching the time sequence of lever presses with the time sequence of locations, we augmented the observation space with locations. Therefore, the augmented observations for the model were lever press, reward delivery, and location. The actions were pressing either of the levers, stay at a location or move. The lever pressing actions were directly available from recording and we identified stay and move actions from the video location tracking. We defined similar observation and action spaces for simulations. To facilitate cross validation, we used a fixed 5-fold split to form training and validation sequences.

The proposed SMJP model has two main procedural components. The trajectorySampling function samples time intervals between consecutive observations using uniformization. It gives us imputed time sequences with missing observations within time intervals, allowing the model to transition between its hidden states at missing observations and use observations only at the end of the time interval. The switchHMM function implements EM approach using action switching and imputed sequences. We instantiated the transition, emission and rate matrices by training the model on observation and action sequences without imputed trajectories. Upon learning the emission and transition matrices for a sampled sequence, we use scaling factor, see model description, and make gradient like updates to the rate matrix while preserving its structure in the function GeneratorUpdate. The procedure of trajectory sampling and training on re-sampled sequence is repeated until the log-likelihood on held out data stops changing within a small tolerance. Therefore, we learn transition, emission probabilities and a rate matrix that capture the underlying continuous time process.

Experiment

We perform three experiments. We use the simulated toy data both to estimate a required training size and to ensure that the switching model is able to learn latent states, establish correspondence between partially observable Markov decision process belief states with SMJP latent states using theoretical optimal agent model and, then, apply our method to a real agent in a free foraging task. The number of states were selected by estimating the value at which the log-likelihood on the validation set stops improving.

Simulated toy data

To create a toy test data generated by the assumed model, we set up two transition matrices and one emission matrix with 5 states, 2 emissions and observation dependent actions. The expected size of the output sequence is set to 5000. Initial action is selected randomly and based on the action index, a transition matrix is selected. Thus, the selected transition matrix and emission matrix combination is used to estimate state transition and generate an emission. The observations, times and actions are added to the output sequence and the observation dependent action value is updated to get new observations. The simulated toy data sequence is used as a basic check if the SMJP model can learn and explain the observations. We fit SMJP model to the simulated data and observe that the log-likelihood starts stabilizing as it reaches the true number of states. It means that the model is able to explain the test data with an equivalent number of latent states (Fig 5). Therefore, we pursue a similar procedure to estimate the required number of latent states for both the optimal agent and the real agent.

Optimal agent

To test our SMJP model we fit it on an optimal agent performing a foraging task. We model the beliefs of an ideal observer in this task using a POMDP. There is a one-to-one correspondence between a POMDP over partially observable world states z and a fully observed Belief MDP in which the 'state' is the 'belief' b or posterior distribution $b_t = p(z_t|o_{1:t})$ over the world state z.We solve this optimal actor problem using a Belief MDP on a discretized belief space. The agent keeps track of its belief state about the world following transition dynamics $p(b'|b,a)$, where b' is the new belief state, b is the current state, and a is an action. The agent's sensory information depends on the world state according to the probability $p(o|b,a)$. Upon taking action a, the agent receives immediate reward $R(b,b',a)$. The goal of the agent is to maximize the

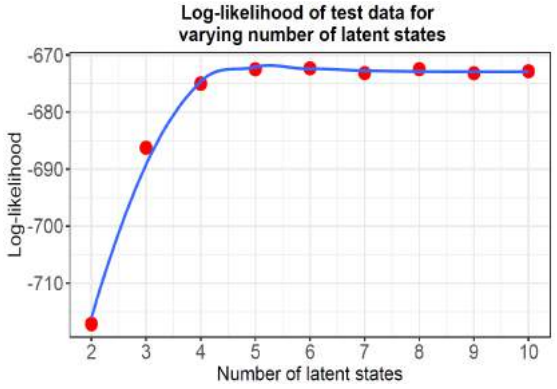


Figure 5: The model is able to explain simulated test data and the log-likelihood on held out data starts flattening out at the true number of states.

long-term expected reward $\mathbb{E}[\sum_{t=0}^{t=\infty} \gamma^t R(b_t, b'_t, a_t)]$. Our model agent achieves this goal using a policy that solves for its policy Bellman (1957), by value iteration on the discretized belief states.

The beliefs serve as latent states which control the agent's behaviors, and give its actions a non-exponential interval distribution, which is recapitulated by the fitted SMJP. We find that the likelihood of the observed data is maximal for a number of states that is smaller than the true size of the underlying POMDP belief space, indicating that the semi-Markov process is able to compress the agent's dynamics into a smaller effective number of latent states. To validate the semi-Markov model in our foraging task, we discover the latent states of the artificial agent for whom we know the ground truth. We model this agent as a near-optimal actor that maximizes reward given partial observations of the true process. This agent maintains beliefs about the availability of food at different locations. Our agent is suboptimal because we do not store the beliefs with arbitrary precision, but rather discretize the beliefs to a finite resolution, and allow some diffusion between those belief states.

Application to the free-foraging task

We apply the SMJP model to infer latent states of agents performing a simple foraging task. We applied the model to both theoretical agents with near-optimal behavior, and real agents (macaques) whose behavior we measured experimentally. In this task, two boxes contained rewards that became available after random exponentially-distributed time intervals. If an agent presses a lever on one box when the food is available, that reward is released and that box timer is reset. The benefit of the reward is offset by two action costs: pressing the lever, and switching boxes. The state of the box is not observable, so the agent must choose its action based on an internal belief about the box, with the presumed goal of maximizing total reward minus costs. This internal belief constitutes a latent state that we infer using the semi-Markov process, both from the artificial agent and behaving monkeys.

We applied the SMJP model to infer latent states of macaques performing a simple two-box foraging task. The animal freely moved between two feeding boxes with levers that released food after an exponentially-distributed random time interval (mean of 10 or 30 sec) had passed. The model observations were lever pressing, reward delivery, and location within the box (Fig 7a). Actions were: stay, move, or press either lever. The monkey's movements were tracked using overhead video, and quantized by k-means into different locations. The number of latent states is estimated by the log-likelihood maximization (Fig 7b). The resultant process constructs the monkey's latent states to explain the non-exponentially-distributed intervals between lever presses (Fig 7).

Results and Discussion

Optimal agent

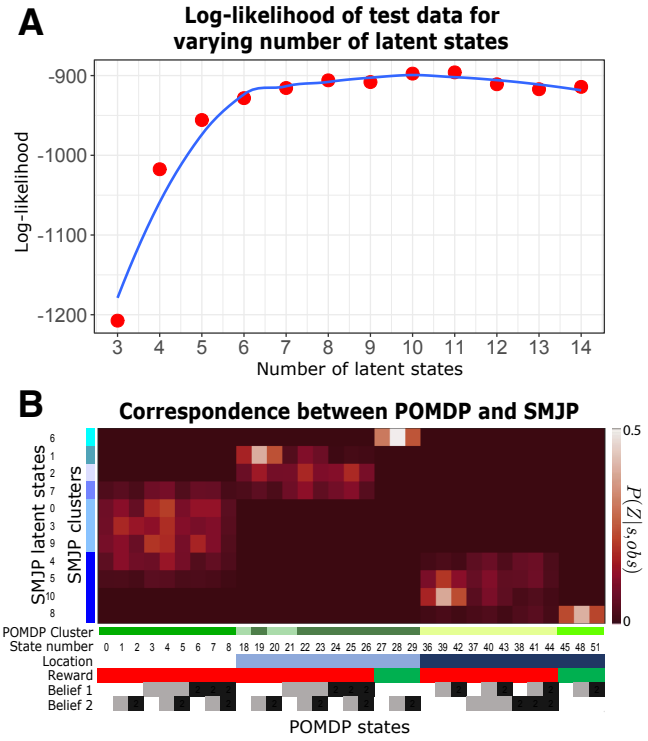


Figure 6: *Latent states inferred by SMJP for an optimal agent implementing a POMDP*. (**a**) Log-likelihood on held out data provides an estimate of the required number of latent states. (**b**) Co-clustering of states in a POMDP and our SMJP, based on the conditional probability of observing each POMDP state Z from each SMJP state, $P(Z|s,obs)$. The POMDP states Z are depicted below the horizontal axis. Clustered structure in the plot reveals that the SMJP states have information about the agent's belief dynamics.

We trained the SMJP on an observation sequence generated by the optimal agent, and optimized the number of SMJP latent states by maximizing the log-likelihood of held-out data (Fig 6a). While the Belief MDP agent's relevant states Z (including

location, reward, and beliefs b) should be implicitly embedded in the SMJP latent states s, these two state representations are not immediately comparable.

To establish a correspondence, we compute the joint distribution over s and Z at any one time point using the shared time series of observations: $p(s,Z|obs) = \frac{1}{T}\sum_t p(s_t|o_{1:T})p(Z_t|o_{1:T})$. This joint distribution shows which SMJP and POMDP states tend to occur at the same time. It therefore provides a dictionary for translating the interpretable POMDP Z states into our learned and unlabeled SMJP s states.

To increase interpretability, we cluster $p(Z|s,o)$ using information theoretic co-clustering, Dhillon et al. (2003), which provides a principled coarse-graining of the states with improved semantic interpretability. We determine the required numbers of SMJP and POMDP co-clusters by finding minimum information loss in information theoretic co-clustering. Fig 6b shows that latent SMJP states are associated with different belief states. Co-clustering also reveals that the SMJP latent states have dynamics that match the belief dynamics (not shown). These results demonstrate that the switching SMJP model can capture latent belief states and dynamics for behavioral data.

Real agent

We trained the SMJP on an observation sequence generated by the real agent (Fig 7a), and optimized the number of SMJP latent states by maximizing the log-likelihood of held-out data (Fig 7b). The SMJP model constructs latent states and dynamics using the real agent's observations to predict choices and timing, including the non-exponentially-distributed intervals between lever presses. Fig 7c shows states extracted for the action 'stay'. Beliefs precede an action and the extracted states reflect beliefs for the next action. For example, being in states $5,8$ are rewarding to the monkey. States that can be interpreted as 'expectant waiting for reward' are highlighted (Fig 7c): these states form a self-exciting delay network that is activated from other rewarded belief states. Moreover, the lower entropy of latent states associated with lever 1 revealed guarding behavior we identified from video. Overall, the model network encodes a set of complex but interpretable dynamics of the animal's beliefs and reward expectations which emphasize the complex computations underlying the decision making process.

Each transition matrix acts like an action operator and the real agent performs operations in sequences. So, we examine joint operators $T_{ji} = T_i T_j$, where T_i and T_j are operators for actions i and j respectively. We use an off-the-shelf package using, Brandes et al. (2008) to extract subgraphs and then persistent subspaces from all the six joint operators corresponding to different action pairs. Fig 7d shows subgraphs for two joint operators of interest (involving actions: lever press and stay). The latent states (within subspaces p and q) appearing in the same subgraphs of the joint operators illustrate the real agent's persistent reward belief states. The states outside the subspaces p and q correspond to other beliefs, for example, switching. These results demonstrate that the presented model is able to extract subtleties, albeit complex, in the belief states and their dynamics. The extracted latent states and dynamics will be useful regressors for finding neural correlates of the computations underlying the monkey's behavioral dynamics.

Conclusion

We presented a continuous-time switching semi-Markov model that learns the latent states dynamics in conformance with the belief structure of a partially observable Markov decision process. The revealed latent states are capable of inferring complex animal behavior and its belief dynamics in naturalistic tasks like foraging. Several aspects of the inferred behaviors and belief dynamics were examined to reveal that indeed, the internal latent structural representation match the agent's belief structure. The data-driven switching semi-Markov model provides useful estimates of the structure of the internal latent states for hard tasks. The latent states from this behavioral model could potentially be used to understand correspondences between neural activity and the latent belief dynamics that govern how an animal selects actions.

Acknowledgments The authors thank Dora Angelaki, Valentin Dragoi, Neda Sahidi and Russell Milton for useful discussions. AK, ZW, XP and PS were supported by BRAIN Initiative grant NIH 5U01NS094368.

References

Anderson, D. J., & Perona, P. (2014). Toward a science of computational ethology. *Neuron*, *84*(1), 18–31.

Bellman, R. (1957). *Dynamic programming: Princeton univ. press.* Princeton.

Brandes, U., Delling, D., Gaertler, M., Gorke, R., Hoefer, M., Nikoloski, Z., & Wagner, D. (2008). On modularity clustering. *IEEE transactions on knowledge and data engineering*, *20*(2), 172–188.

Charnov, E., & Orians, G. H. (2006). Optimal foraging: some theoretical explorations.

Dhillon, I. S., Mallela, S., & Modha, D. S. (2003). Information-theoretic co-clustering. In *Proceedings of the ninth acm sigkdd international conference on knowledge discovery and data mining* (pp. 89–98).

Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *The journal of physical chemistry*, *81*(25), 2340–2361.

Kaelbling, L. P., Littman, M. L., & Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. *Artificial intelligence*, *101*(1), 99–134.

Langrock, R., King, R., Matthiopoulos, J., Thomas, L., Fortin, D., & Morales, J. M. (2012). Flexible and practical modeling of animal telemetry data: hidden markov models and extensions. *Ecology*, *93*(11), 2336–2342.

Li, M., & Bolker, B. M. (2017). Incorporating periodic variability in hidden markov models for animal movement. *Movement ecology*, *5*(1), 1.

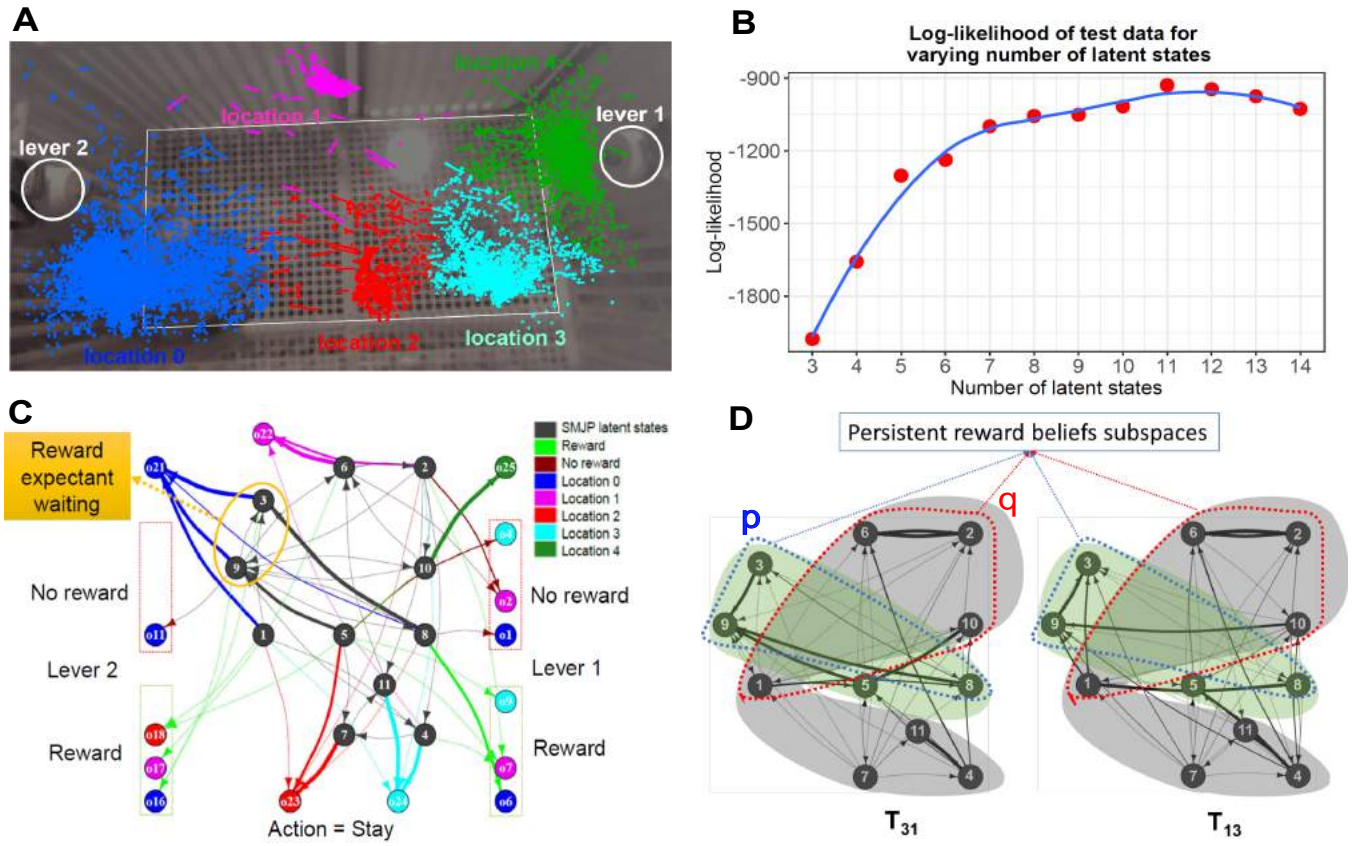


Figure 7: Analyzing behavioral data from a freely moving monkey using the SMJP. (**a**) Overhead video (background image) tracked the locations and normalized velocities (vectors) of the monkey. These data were then clustered by the k-means algorithm. (**b**) We get an estimate of the required number of latent states by observing log-likelihood on held out data. (**c**) SMJP model for observed monkey behavioral data for the action stay. Highlighted reward expectant waiting states illustrate that the latent states as regressors for the beliefs dynamics are useful in understanding monkey's behavior. (**d**) Subspaces p and q (blue and red dotted), within the subgraphs (green and gray highlighted) for the joint operators T_{31} and T_{13} reveal persistent reward belief states.

Macdonald, I. L., & Raubenheimer, D. (1995). Hidden markov models and animal behaviour. *Biometrical Journal*, *37*(6), 701–712.

Nathan, R., Getz, W. M., Revilla, E., Holyoak, M., Kadmon, R., Saltz, D., & Smouse, P. E. (2008). A movement ecology paradigm for unifying organismal movement research. *Proceedings of the National Academy of Sciences*, *105*(49), 19052–19059.

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257–286.

Rao, V., & Teh, Y. W. (2013). Fast mcmc sampling for markov jump processes and extensions. *Journal of Machine Learning Research*, *14*(1), 3295–3320.

Zucchini, W., Raubenheimer, D., & MacDonald, I. L. (2008). Modeling time series of animal behavior by means of a latent-state model with feedback. *Biometrics*, *64*(3), 807–815.

AI and Cognitive Testing: A New Conceptual Framework and Roadmap

Anonymous CogSci submission

Abstract

Understanding how a person thinks, i.e., measuring a single individual's cognitive characteristics, is challenging because cognition is not directly observable. Practically speaking, standardized cognitive tests (tests of IQ, memory, attention, etc.), with results interpreted by expert clinicians, represent the state of the art in measuring a person's cognition. Three areas of AI show particular promise for improving the effectiveness of this kind of cognitive testing: 1) behavioral sensing, to more robustly quantify individual test-taker behaviors, 2) data mining, to identify and extract meaningful patterns from behavioral datasets; and 3) cognitive modeling, to help map observed behaviors onto hypothesized cognitive strategies. We bring these three areas of AI research together in a unified conceptual framework and provide a sampling of recent work in each area. Continued research at the nexus of AI and cognitive testing has potentially far-reaching implications for society in virtually every context in which measuring cognition is important, including research across many disciplines of cognitive science as well as applications in clinical, educational, and workforce settings.

Keywords: artificial intelligence; behavioral sensing; cognitive modeling; computational psychiatry; neuropsychology.

Introduction

The meat of the matter is often *how* a patient solves a problem or approaches a task rather than what the score is.
(Lezak et al., 2012, *Neuropsychological Assessment*, p. 160)

Different people think in different ways. This seemingly obvious statement masks many deep scientific mysteries about the human mind and also has enormous implications for individual and societal well-being.

How a person thinks is central to everything that they do: it affects how they learn, work, communicate, set goals, make decisions, etc. Thus, the scientific study of individual cognitive variations is critical not just for (1) advancing our basic understanding of human cognition and development across the lifespan, including research on genes, brain, and behavior, but also for (2) improving evidence-based practices in education and special education, workforce training, clinical diagnosis and treatment, rehabilitation, and more.

However, measuring cognition is uniquely challenging, as cognitive entities and processes are not observable in the same way that genetic, physiological, behavioral, and even neural characteristics can be measured using physical sensing technologies. **We have no way (at least at present) of *directly* measuring a person's mental representations.**

Even with advances in neuroimaging technologies that can capture subtle characteristics of neural activity, measuring such activity is only a rough proxy for actual cognitive activity; the question remains of how to "allow the brain measurements to make contact with putative cognitive processes" (Forstmann & Wagenmakers, 2015, p.144).

Currently, the gold standard for individual cognitive evaluations are those carried out by expert clinicians, usually psychologists or neuropsychologists.[1] These evaluations might be done to diagnose learning or developmental disabilities in children, detect signs of cognitive decline in elderly patients, or identify cognitive deficits after stroke or other brain injury (Lezak et al., 2012). Such evaluations combine two types of information about an individual: (1) information about how that individual is functioning *outside* the clinic, through self-report measures, interviews or questionnaires given to parents or caregivers, etc.; and (2) information about how that individual is functioning *inside* the clinic, usually through the administration of standardized cognitive tests, e.g., tests of memory, IQ, visuospatial reasoning, language, etc.

It is the second item in this list—cognitive testing—that is the focus of this paper. Distinct research paradigms within artificial intelligence (AI) have the potential to advance cognitive testing in (at least) three key ways:

1. *Behavioral sensing*: to more robustly quantify individual test-taker behaviors.
2. *Data mining*: to identify and extract meaningful patterns from behavioral datasets.
3. *Cognitive modeling*: to help map observed behaviors onto hypothesized cognitive strategies.

Before getting into the details of these three areas, however, it is important to first understand how conventional cognitive testing works. **This paper presents a new conceptual framework that explains the strengths and limitations of current methods for cognitive testing and highlights specific ways in which AI can help.** We also provide a sampling of recent AI research in each area.

[1] Cognitive evaluations also often occur in education and workforce settings, though these are typically less detailed but more domain-specific than clinical evaluations. In many human research studies across all areas of science, cognitive evaluations are used for participant inclusion/exclusion, group matching, and/or covariate analyses. While this paper focuses primarily on the clinical setting, our observations pertain to these other settings as well.

How Cognitive Testing Works

The rationale behind cognitive tests is straightforward. A given test poses problems for a test-taker to solve. Problems are specifically designed to tap certain cognitive representations and processes, which we refer to as *cognitive strategies*. Test designs are often validated (i.e. to "prove" that a test indeed is tapping into the right cognitive strategies) through converging evidence from many different sources, including data from neuroimaging studies, patients with known cognitive or neurological issues, and/or other cognitive tests. A person's test score thus provides an indirect measure of these hypothesized cognitive strategies.

However, a well known issue with most cognitive tests is **ambiguity**: while test scores do indicate *how well* a person solves test problems, i.e., that person's level of ability, they do not indicate *how* a person solves test problems, i.e., their actual cognitive strategy. In other words, two people can get the same test score using very different cognitive strategies. Moreover, this ambiguity can occur with low or high scores:

> "'There are many reasons for failing and there are many ways you can go about it. And if you don't know in fact which way the patient was going about it, failure doesn't tell you very much' (Darby & Walsh, 2005). There can also be more than one way to pass a test." (Lezak et al., 2012, p. 160)

Because of this ambiguity, expert clinicians often combine scores with other observed behaviors, such as errors, eye gaze, emotions, general demeanor, etc., in order to better interpret a person's test performance. This supports the rationale for why only "trained" clinicians should administer cognitive tests, and also why clinicians develop such deep expertise with their particular population and goal (e.g. screening children for learning disabilities versus working with elderly patients to detect memory issues).

In reality, as mentioned in the introduction, clinicians likely never rely on results from a single cognitive test to make judgments about a person's cognition. They combine results from many tests with additional information about a person's performance outside the clinic (e.g. school performance, medical history, etc.). For the purposes of this paper, however, we focus on thinking about just a single cognitive test and what it can tell us.

Proposed framework

In this section, we propose a new formalism for describing what is happening during a conventional cognitive test. For added clarity, we also use the Raven's Progressive Matrices (RPM) intelligence test as a running example. The RPM is a well studied standardized test that poses problems similar to geometric analogies: a matrix of visual figures is presented with one missing, and the missing figure must be chosen from among a set of candidate answer figures (i.e., multiple choice). The RPM is one of the best single-format measures of intelligence among all cognitive tests (Snow, Kyllonen, & Marshalek, 1984) and thus is very widely used.

Definition 1. Let the set X_{human} represent all possible cognitive strategies that a person can use to attempt to perform a given cognitive test, successful or not.

Definition 2. Let the set Y represent all possible scores that can be earned on a given cognitive test.

Definition 3. Let the function F represent a mapping from a person's use of a particular cognitive strategy onto the resulting test score:

$$F(x_i \in X_{human}) = y_i \in Y$$

We do not concern ourselves with how X_{human} might be represented. The set is infinite, even if we exclude obviously irrelevant strategies.[2] An individual person probably can access at least a few strategies from X_{human}, and certainly they can also be taught to use particular strategies.

Though not, perhaps, designed this way on purpose, the RPM is amenable to multiple distinct strategies. For example, there is evidence that many neurotypical individuals often use verbal, inner-speech-like strategies, whereas many individuals on the autism spectrum use visually mediated, mental-imagery-like strategies (Soulières et al., 2009). In fact, some argue that the reason the RPM is such a good intelligence test may be because it is actually testing metacognitive flexibility, in terms of strategy selection/adaptation (Kirby & Lawson, 1983)...a point that we return to later on in this paper.

For simplicity, let us assume that a person uses a single strategy $x_i \in X_{human}$ to solve a given cognitive test. Using this strategy x_i, they receive a score y_i. In other words, the act of taking the test is what "computes" the function F.

In the case of the RPM, the test is scored as number of correct answers, and so possible scores (for the standard version of the test) range from 0 to 60. So, suppose someone uses a verbally mediated strategy, and they get a score of 50/60.

Using these definitions, ambiguity exists because F is a many-to-one function. There are many possible strategies in X_{human} that may lead to a score of 50. As a result, the inverse function $F^{-1}(y_i) = x_i \in X_{human}$ is ill-defined.

To help with this problem, we expand our definitions to include additional test-taker behaviors, beyond just test score:

Definition 4. Let the vector B_{human} represent a sequence of observable behaviors generated by a person taking a cognitive test, including test score y as well as response times, types of errors made, patterns of eye gaze, etc.

Definition 5. Let the function G represent a mapping from a person's use of a particular cognitive strategy onto the sequence of resulting behaviors:

$$G(x_i \in X_{human}) \rightarrow B_{human}$$

For example, for a person taking the RPM, one might include in B_{human} the time taken to complete each problem, the

[2] Making a peanut butter and jelly sandwich is one possible strategy for solving RPM problems. It is, however, an exceedingly poor strategy, and so let's exclude it from X_{human}.

answer choice that is selected, the pattern of eye gaze between different visual elements, etc.

Now, while the function G is still a many-to-one function (i.e., multiple strategies might still map onto the same sequence of behaviors), it is "less" many-to-one than our earlier function F that mapped strategies onto scores. Each behavioral observation that is made places an additional constraint on the subset of strategies in X_{human} that could have produced the full sequence of behaviors. Therefore, given a sequence of observable behaviors B_{human}, the inverse function $G^{-1}(B_{human}) = x_i \in X_{human}$ provides a better estimate of a person's cognitive strategy than does the inverse function F^{-1} that relies on test score alone.

For example, research on geometric analogies has shown that different patterns of eye gaze seem to be indicative of different high-level problem-solving strategies (Bethell-Fox, Lohman, & Snow, 1984). Some people look at the "problem" part and come up with their own answer before looking at the answer choices, while others look at the answer choices early and use more of a trial-and-error approach, mentally plugging in each answer choice to see which one looks best.

One problem remains: where does the sequence of behaviors B_{human} come from? For traditional cognitive tests, usually administered in a pencil-and-paper or objects-on-a-table format, there is no perfect record of B_{human}. Clinicians observing a person taking a test use their own, expertly trained powers of perception, memory, and note-taking to process B_{human} in real time in order to extract meaningful patterns:

Definition 6. Let the function P represent a mapping from a sequence of low-level behaviors B_{human} to a selected set of patterns (i.e., a subset and/or transformed view of individual observations in B_{human}).

We use P_{expert} to denote the function that a clinician applies to extract meaningful patterns from the raw behavioral sequence B_{human}. **Thus, when a clinician observes a person's test performance to infer information about that person's cognition, they are implicitly computing the function:**

$$G^{-1}_{expert}(P_{expert}(B_{human})) = x_i \in X_{human} \quad (1)$$

Where do the functions G^{-1}_{expert} and P_{expert} come from? In general, they are learned over years or decades of administering cognitive tests to certain segments of the population. For example, a clinician with expertise in learning disabilities likely uses G^{-1} and P functions that are tuned to patterns of behavior most relevant for diagnosing these conditions in children. Another clinician who works mostly with brain injury patients would likely use different G^{-1} and P functions, even when administering similar tests.

The problem with implicit functions, and current, non-AI-based solutions

The main problem with these learned G^{-1}_{expert} and P_{expert} functions is that they are implicit in a clinician's expertise. Not only are they implicit, but they are also very difficult to make explicit, even if a clinician tries to do so. This difficulty in turn complicates efforts to measure the validity or reliability of these functions, both for individual clinicians and for the field of cognitive assessment as a whole.

The Boston Process Approach to neuropsychology was essentially an attempt to "write down" these functions using a combination of expert judgment and carefully designed research studies, so that the resulting functions could be more rigorously evaluated for validity and reliability, and also so these functions could be explicitly taught as part of professional neuropsychology training. However, while the ideas of the Boston Process Approach have been influential, the complexity of its methods and the challenges of real-time data collection during testing sessions limited its widespread adoption (Milberg, Hebben, Kaplan, Grant, & Adams, 2009).

The advent of computer-based testing has provided new methods for recording sequences of test-taker behaviors, such as detailed reaction times, errors, etc. Some , like the California Verbal Learning Test (Delis, Freeland, Kramer, & Kaplan, 1988), have been designed specifically to enable the use of these additional behaviors to infer more and better information about a person's cognitive strategy than would be obtainable from their score alone.

These and similar efforts from the neuropsychology research community have been analyzed more recently under the heading of the Quantified Process Approach (Poreh, 2012), which emphasizes the critical need to understand cognitive strategies, i.e. "process," using quantifiable measures, in addition to the subjective and often qualitative judgments of individual clinicians (what we describe here as the implicit G^{-1}_{expert} and P_{expert} functions). The Quantified Process Approach outlines three categories of potential solutions: 1) using additional tests to essentially triangulate a person's strategy using multiple points of measurement; 2) using additional measures of behavior from a single test to develop new indices of interest; and 3) decompose scores into subscores that might reflect different underlying factors. Of these three categories, the latter two would fall into our proposed framework as efforts to come up with explicit G^{-1} and P functions, depending on whether the behaviors B_{human} considered are taken from behavioral dimensions above and beyond scores (category 2) or from behavioral dimensions within scores that pinpoint more detailed subscores (category 3).

However, these various pockets of research have yet to transform the daily practice of cognitive testing. Problems remain in how to quantify G^{-1} and P functions in a scalable way that can be applied across many different cognitive tests and many populations, while also ensuring that methods are readily usable by practicing clinicians.

AI to the Rescue[3]

Using this framework, we now describe ways in which AI can help solve some of these problems through 1) behavioral sensing, 2) data mining, and 3) cognitive modeling.

[3]Possibly.... https://xkcd.com/1831/

Behavioral sensing

The first, and perhaps most obvious, role for AI in cognitive testing is in recording behavioral observations, i.e., in obtaining the sequence of behaviors B_{human} from a test session.

Part of behavioral sensing involves advances in hardware, such as the development of more advanced (and more affordable) eye trackers. Computer-based testing platforms can easily log many kinds of behaviors, including mouse movements, key presses, etc. Tablet-based tests are being used to capture more detailed manual behaviors such as velocity of pen strokes (Davis, Libon, Au, Pitman, & Penney, 2014).

While behavioral sensing in computer-based environments is currently more common, one of the most exciting new areas for behavioral sensing involves sensing in real, 3D environments, which often calls for a combination of advances in hardware and in signal processing algorithms. Eye tracking technology is now getting to the point where head-mounted eye trackers are relatively lightweight and affordable (Kassner, Patera, & Bulling, 2014), and computer vision algorithms can be used to help analyze the video stream coming from such eye trackers. These advances enable scalable eye-tracking in 3D environments, which, in previous years, would have been virtually unthinkable in the context of cognitive testing from usability or scalability perspectives. Physiological sensors are also now often incorporated into cognitive assessments, e.g., using skin conductance sensors to obtain measurements of heart rate, etc. as a proxy for measuring cognitive stress or other affective variables during a testing session (Fletcher et al., 2010).

In addition, even data recorded from regular sensors (cameras, microphones, etc.) can now be analyzed automatically using AI algorithms coming from computer vision, natural language processing, etc. The term *behavioral imaging* has been coined to describe this new subfield of AI directed at producing robust and reliable measurements of human behavior in 3D assessment settings (Rehg et al., 2014).

Behavioral sensing can thus be understood in terms of its two components: sensors to record raw signals coming from a testing session(e.g., pixels from a video camera), plus algorithms to process those raw signals into measurements of behavior (e.g., computer vision algorithm to detect, from a raw video stream, when a person moves an object on a table).

Behavioral sensing can help in measuring many types of behaviors. Some behaviors are already easily measured by humans, but automated approaches may increase the scalability or accuracy of such measurements (e.g., counting how many errors a person makes while solving a table-top block copying task). Other behaviors might be currently detectable by human clinicians but only in qualitative ways. For example, many social assessments for the diagnosis of autism use "quality of eye contact" as a measurement of interest, which is often recorded as a subjective overall impression by a human clinician, but could be broken down into quantified components by an algorithm (Ye et al., 2015). Still other behaviors might not be detectable by human clinicians at all; for example, being able to capture the exact velocities and pressures manually applied by a person performing a tablet-based drawing test (Davis et al., 2014).

Data mining

The next role for AI is in quantifying the function P that takes in a sequence of behaviors B_{human} and extracts meaningful patterns. Meaningful patterns can be created in many different ways, including by identifying subsets of behaviors that are particularly relevant, or by producing transformations of low-level behaviors into higher-level constructs.

For example, there has been work that first uses a tablet-based version of the clock drawing test to record low-level manual drawing behaviors, and then applies machine learning classification algorithms to these data to help diagnose Alzheimer's, Parkinson's, and other cognitive conditions (Souillard-Mandar et al., 2016).

In another effort, eye tracking data from a visual recognition test (the Visual Paired Comparison test) have been used to train classifiers to detect early signs of mild cognitive impairment, which is often a precursor to Alzheimer's (Lagun, Manzanares, Zola, Buffalo, & Agichtein, 2011). A clever extension of this work aims to see if mouse movement data from a non-eye-tracking variant of the task can support comparable classification performance, which would greatly increase the scalability of the test by removing the need for an eye tracker (Agichtein et al., 2017).

In general, the broad umbrella of data mining approaches for cognitive testing can include the use of: 1) new algorithms applied to existing behavioral datasets; 2) conventional statistical analyses applied to new behavioral datasets; and 3) new algorithms applied to new datasets. All of these approaches represent important routes for improving our understanding of the low-level behaviors that come out of cognitive tests, i.e., to identify which behaviors or combinations of behaviors are most important for a given clinical goal.

Cognitive modeling

The third important role that AI can play in cognitive testing is through cognitive modeling. What does a computational cognitive model actually accomplish? To answer this question, we begin by supposing that we have created a particular type of AI system—a computational cognitive architecture—that can employ different problem-solving strategies to solve problems from a given cognitive test.

Critically, such an AI system is not just a mathematical model of relationships between hypothesized cognitive entities involved in solving the test. It is a computational model of the hypothesized entities themselves; it provides a mechanism-level view of what might be going on. The key difference between a mathematical model and a computational model is that a computational model bears an analogical relationship with what it is trying to model; there is some structural correspondence between the model and what it represents (Hunt, Ropella, Park, & Engelberg, 2008).

Definitions 7 through 12 (below) refer to concepts related to this kind of computational model, which are also analogous (but not identical) to the concepts given in Definitions 1 through 6 (above) for human test-takers.

Definition 7. Let X_{AI} represent the set of problem-solving strategies that an AI system can use to solve a given cognitive test, including successful and unsuccessful strategies.

Definition 8. Let y_{AI} represent the score the AI system receives on a given cognitive test.

Definition 9. Let the function F_* represent a mapping from an AI system's use of a particular strategy onto the resulting test score, i.e., $F_*(x_i \in X_{AI}) \rightarrow y_{AI}$.

Definition 10. Let B_{AI} represent the sequence of *simulated* observable behaviors b_i generated by an AI system taking a cognitive test. These behaviors can include test scores y_{AI} as well as response times, types of errors made, patterns of eye gaze, etc.

Definition 11. Let the function G_* represent a mapping from an AI system's use of a particular strategy onto the resulting test score plus simulated behaviors, i.e., $G_*(x_i \in X_{AI}) \rightarrow B_{AI}$.

Definition 12. Let the function P_* represent a mapping from a sequence of low-level behaviors B_{AI} to higher-level features.

To take our previous example of the Raven's Progressive Matrices test, many computational cognitive models of this kind have been developed over the years (Carpenter, Just, & Shell, 1990; Lovett, Tomai, Forbus, & Usher, 2009; Kunda, McGreggor, & Goel, 2013; Strannegård, Cirillo, & Ström, 2013). There has also been much work in the cognitive architectures community (e.g. using SOAR, ACT-R, etc.) to develop richly detailed models of many different tasks.

Given such a computational cognitive model, we can run experiments that have the model use a variety of different strategies to solve a given cognitive test. We can measure data from these experiments to obtain test scores and behaviors, just as we do for human test takers. **The key difference here is that cognitive strategies in a cognitive model are directly observable!** We have the "ground truth" for our model in a way that is (at least currently) impossible to obtain for human test takers.

At minimum, we can study the function F_* to understand more about potential ambiguities on a particular cognitive test, which would itself a valuable contribution to the field of cognitive testing.

Also, such a cognitive model provides a systematic way to obtain quantified functions for mapping from the space of observed behaviors back onto cognitive strategies, i.e., the function G^{-1}. This is still not easy (though it is much easier when we have the ground truth for X!). There are probably many possible approaches for obtaining the G^{-1} function. One might be to run a large set of computational experiments to get two linked datasets X_{AI} and B_{AI}, and then use machine learning and data mining algorithms to find relevant patterns and predictors within these.

One important area for research using computational cognitive models is to more effectively capture individual differences. Much of the research on cognitive architectures, for example, focuses on modeling generalized human performance or broad group differences. As the quantity and quality of behavioral measurements increase, through behavioral sensing and data mining, cognitive models should also be able to take advantage of these datasets to create more precise explanations of individual variations.

Another extremely interesting open question is: where do the strategies in X_{AI} come from? For now, X_{AI} is defined by the AI system's designers, informed by research on human cognition. An important AI frontier is to develop AI systems that *learn* strategies through instruction, observation, and experience, as people do (Laird et al., 2017). This research would not only expand the capabilities of our cognitive models, but results would also help us better understand human cognitive strategies at the metacognitive level. As mentioned earlier, for example, work on the Raven's Progressive Matrices test suggests that a person's methods for strategy selection are just as important for test performance as are the strategies themselves (Kirby & Lawson, 1983).

A Call to Action

Similar observations have been compiled under the heading of computational psychiatry (Montague, Dolan, Friston, & Dayan, 2012; Huys, Maia, & Frank, 2016), though the specific formalism given here is (to our knowledge) new.

What our analysis suggests is that interdisciplinary collaboration is critical for advancing the science of cognitive testing, not just between clinicians and AI researchers in general, but between clinicians and AI researchers coming from the distinct subfields of behavioral sensing, data mining, and cognitive modeling.

In addition, one extremely promising horizon is to think about the development of new cognitive tests that are enabled by the types of technological advances described above. For example, now that we can measure and understand very rich sets of behavior, and also map these onto detailed hypotheses about cognitive strategies, can we begin to measure complex forms of cognition in more naturalistic tasks? So much of current test design has been shaped by the limitations in the scalability of these elements in previous decades. Previously, cognitive test designers had to construct very constrained tasks, that would only measure one or two cognitive constructs at a time, and that would produce easily measurable scores. Now, for example, could we give people a realistic search task in a complex, 3D environment to test their attention and/or memory? There is a great opportunity here to begin coming up with much more creative and naturalistic ways to tap into a person's realistic cognitive processes.

References

Agichtein, Y. E., Buffalo, E. A., Lagun, D., Manzanares, C., & Zola, S. (2017, April 25). *Internet-based cognitive diagnostics using visual paired comparison task.* Google Patents. (US Patent 9,629,543)

Bethell-Fox, C. E., Lohman, D. F., & Snow, R. E. (1984). Adaptive reasoning: Componential and eye movement analysis of geometric analogy performance. *Intelligence*, *8*(3), 205–238.

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: a theoretical account of the processing in the raven progressive matrices test. *Psychological review*, *97*(3), 404.

Darby, D., & Walsh, K. W. (2005). *Walsh's neuropsychology: A clinical approach*. Churchill Livingstone.

Davis, R., Libon, D. J., Au, R., Pitman, D., & Penney, D. L. (2014). Think: Inferring cognitive status from subtle behaviors. In *Aaai* (pp. 2898–2905).

Delis, D. C., Freeland, J., Kramer, J. H., & Kaplan, E. (1988). Integrating clinical assessment with cognitive neuroscience: construct validation of the california verbal learning test. *Journal of consulting and clinical psychology*, *56*(1), 123.

Fletcher, R. R., Dobson, K., Goodwin, M. S., Eydgahi, H., Wilder-Smith, O., Fernholz, D., ... Picard, R. W. (2010). icalm: Wearable sensor and network architecture for wirelessly communicating and logging autonomic activity. *IEEE Transactions on Information Technology in Biomedicine*, *14*(2), 215–223.

Forstmann, B. U., & Wagenmakers, E.-J. (2015). Model-based cognitive neuroscience: A conceptual introduction. In *An introduction to model-based cognitive neuroscience* (pp. 139–156). Springer.

Hunt, C. A., Ropella, G. E., Park, S., & Engelberg, J. (2008). Dichotomies between computational and mathematical models. *Nature biotechnology*, *26*(7), 737.

Huys, Q. J., Maia, T. V., & Frank, M. J. (2016). Computational psychiatry as a bridge from neuroscience to clinical applications. *Nature neuroscience*, *19*(3), 404.

Kassner, M., Patera, W., & Bulling, A. (2014). Pupil: an open source platform for pervasive eye tracking and mobile gaze-based interaction. In *Proceedings of the 2014 acm international joint conference on pervasive and ubiquitous computing: Adjunct publication* (pp. 1151–1160).

Kirby, J. R., & Lawson, M. J. (1983). Effects of strategy training on progressive matrices performance. *Contemporary Educational Psychology*, *8*(2), 127–140.

Kunda, M., McGreggor, K., & Goel, A. K. (2013). A computational model for solving problems from the ravens progressive matrices intelligence test using iconic visual representations. *Cognitive Systems Research*, *22*, 47–66.

Lagun, D., Manzanares, C., Zola, S. M., Buffalo, E. A., & Agichtein, E. (2011). Detecting cognitive impairment by eye movement analysis using automatic classification algorithms. *Journal of neuroscience methods*, *201*(1), 196–203.

Laird, J. E., Gluck, K., Anderson, J., Forbus, K. D., Jenkins, O. C., Lebiere, C., ... others (2017). Interactive task learning. *IEEE Intelligent Systems*, *32*(4), 6–21.

Lezak, M. D., Howieson, D. B., Bigler, E. D., & Tranel, D. (2012). *Neuropsychological assessment, fifth edition*. Oxford University Press, USA.

Lovett, A., Tomai, E., Forbus, K., & Usher, J. (2009). Solving geometric analogy problems through two-stage analogical mapping. *Cognitive science*, *33*(7), 1192–1231.

Milberg, W. P., Hebben, N., Kaplan, E., Grant, I., & Adams, K. (2009). The boston process approach to neuropsychological assessment. *Neuropsychological assessment of neuropsychiatric and neuromedical disorders*, 42–65.

Montague, P. R., Dolan, R. J., Friston, K. J., & Dayan, P. (2012). Computational psychiatry. *Trends in cognitive sciences*, *16*(1), 72–80.

Poreh, A. M. (2012). *The quantified process approach to neuropsychological assessment*. Psychology Press.

Rehg, J. M., Rozga, A., Abowd, G. D., & Goodwin, M. S. (2014). Behavioral imaging and autism. *IEEE Pervasive Computing*, *13*(2), 84–87.

Snow, R. E., Kyllonen, P. C., & Marshalek, B. (1984). The topography of ability and learning correlations. *Advances in the psychology of human intelligence*, *2*, 47–103.

Souillard-Mandar, W., Davis, R., Rudin, C., Au, R., Libon, D. J., Swenson, R., ... Penney, D. L. (2016). Learning classification models of cognitive conditions from subtle behaviors in the digital clock drawing test. *Machine learning*, *102*(3), 393–441.

Soulières, I., Dawson, M., Samson, F., Barbeau, E. B., Sahyoun, C. P., Strangman, G. E., ... Mottron, L. (2009). Enhanced visual processing contributes to matrix reasoning in autism. *Human brain mapping*, *30*(12), 4082–4107.

Strannegård, C., Cirillo, S., & Ström, V. (2013). An anthropomorphic method for progressive matrix problems. *Cognitive Systems Research*, *22*, 35–46.

Ye, Z., Li, Y., Liu, Y., Bridges, C., Rozga, A., & Rehg, J. M. (2015). Detecting bids for eye contact using a wearable camera. In *Automatic face and gesture recognition (fg), 2015 11th ieee international conference and workshops on* (Vol. 1, pp. 1–8).

Sensitivity to Temporal Community Structure in the Language Domain

Kendra V. Lange (kxl786@psu.edu)
Department of Psychology, The Pennsylvania State University,
Moore Building, University Park, PA 16802 USA

Carol A. Miller (cam47@psu.edu)
Department of Communication Sciences and Disorders, The Pennsylvania State University,
Ford Building, University Park, PA 16802 USA

Daniel J. Weiss (djw21@psu.edu)
Department of Psychology, The Pennsylvania State University,
Moore Building, University Park, PA 16802 USA

Elisabeth A. Karuza (exk521@psu.edu)
Department of Psychology, The Pennsylvania State University,
Moore Building, University Park, PA 16802 USA

Abstract

The interrelatedness of lexical items, typically defined in terms of semantic or phonological overlap, has been shown to influence language learning. Given that language also contains sequential structure, we investigate here whether temporal overlap among words, formalized in graph theoretical terms as displaying the property of *community structure*, might also have consequences for learning. We create a graph organized into clusters of densely interconnected nodes with relatively sparse external connections. After assigning a novel pseudoword to each node in the graph, we generate a continuous sequence of visually-presented items by walking along its edges. Word-by-word reading times suggest that learners are indeed sensitive to temporal overlap. Compellingly, we also demonstrate that prior exposure to sequences organized into temporal communities influences performance on a subsequent word recognition task.

Keywords: network science; statistical learning; language acquisition

Introduction

A foundational question in cognitive science asks how the human brain converts a vast amount of sensory input into usable knowledge. Fortunately for our brains, sensory input, though noisy, tends to be richly patterned. A means of characterizing broad-scale patterns, network science enables the mathematical description of systems as varied as social relationships (Scott, 2017) and neural connectivity (Bassett & Sporns, 2017). Of particular relevance to the present series of experiments, applications of network science to the domain of natural language have dramatically increased our understanding of the organization of phonological (Vitevitch, 2008; Arbesman, Strogatz, & Vitevitch, 2018), syntactic (Ferrer i Cancho, Solé, & Köhler, 2004; Liu, 2008), and semantic systems (Collins & Loftus, 1975; Borge-Holthoefer & Arenas, 2010).

A growing body of evidence suggests that humans use network-level properties when acquiring and accessing linguistic knowledge (for a review, see Karuza, Thompson-Schill, & Bassett, 2016). For example, an index of the extent to which phonological neighbors of a word are themselves neighbors, clustering coefficient has been shown to predict acquisition of novel object labels designed to vary with respect to this property (Goldstein & Vitevitch, 2014). Learners also show sensitivity to lexical islands, or small groups of phonologically related words isolated from a network's "giant component," or the largest group of interrelated words. Siew & Vitevitch (2016) observed that words drawn from lexical islands are recognized and recalled more easily than those from a giant component. For semantic networks, in which nodes representing concepts are linked according to some similarity metric, evidence suggests that densely connected words are most likely to be acquired early in development (Steyvers & Tenenbaum, 2005). In sum, the structural properties of complex language networks may carry important implications for learning.

Outside the language domain, a number of studies have also begun to probe human sensitivity to network topology, generally focusing on community structure in temporally-defined graphs. In these studies, nodes correspond to fractals, glyphs, or button press combinations, and edges mark the transition between two images in a continuous sequence (Schapiro, Rogers, Cordova, Turk-Browne, & Botvinick, 2013; Karuza, Kahn, Thompson-Schill, & Bassett, 2017; Tompson, Kahn, Falk, Vettel, & Bassett, 2018; Kahn, Karuza, Vettel, & Bassett, 2018). Response times are typically recorded as participants view an uninterrupted stimulus stream created by "walking" along the edges of a graph comprised of sparsely connected clusters of densely interconnected nodes (i.e., that display the property of community structure; Figure 1). Results point to a signature response pattern associated with the transition between communities: a pronounced increase in learners' processing times when measured against within-community transitions (Karuza et al., 2017).

Expanding on prior work, which has focused exclusively on non-linguistic visual stimuli, we investigate here whether learners display a comparable sensitivity to community structure when it dictates the order of visually-presented pseudoword sequences. One defining characteristic of linguistic signal is that it unfolds in time. In light of this, we examine whether the temporal overlap between words, not only their phonological and semantic interrelatedness, might steer the learning process. In adapting this paradigm to the language domain, our work makes two additional contributions: first we expand on the size of tested network structures, creating graphs of 40 nodes instead of the 10-15 used in related prior work (Schapiro et al., 2013; Karuza et al., 2017; Tompson et al., 2018; Kahn et al., 2018). Second, we refine an offline measure that allows us to investigate the influence of community structure not only in moment-to-moment processing of novel stimulus streams, but also in accessing previously acquired knowledge in future contexts.

Study 1: Community Structure and Substring Familiarity

We first examine whether learners exhibit cross-community reaction time (RT) increases as they process continuous sequences of unfamiliar linguistic stimuli. We also ask whether sensitivity to community structure will manifest in the expression of knowledge in offline familiarity judgements involving short sequences (substrings) extracted from the original exposure stream. Analyses test the hypothesis that learners prefer substrings drawn from within communities relative to those that span communities.

Materials and Methods

Participants 33 neurologically normal participants (5 male, 28 female; 18-21 years old) participated in this study. They were recruited from the undergraduate psychology research pool at Pennsylvania State University and were granted course credit for their participation. All participants provided informed consent. Three participants were excluded for performance below a pre-determined threshold on an orthogonal cover task (<70% correct; Karuza et al., 2017).

Stimuli

Network properties. Exposure streams were generated via a random walk on a graph featuring five communities of eight nodes each (Figure 1). Each community was connected to two other communities through boundary nodes sharing a single edge with an adjacent community. With the exception of boundary nodes within the same community, which were unlinked, each other node was connected to every other node in their community. Thus, all nodes had equivalent degree, or number of incident edges. Because edges were undirected and unweighted, (1) they could be traversed in any direction and (2) transitions between any two nodes were equally probable. Nodes within the graph corresponded to a unique, pronounceable pseudoword, and edges represented the direct succession of two pseudowords within the stimulus stream.

Pseudoword properties. Pseudowords were selected from the ARC non-word database (Rastle, Harrington, & Coltheart, 2002). Forty orthotactically plausible, single-syllable words were chosen, 20 four-letter words and 20 five-letter words. All words had 5-30 orthographic neighbors and 5-30 phonological neighbors. While metrics such as Coltheart's N

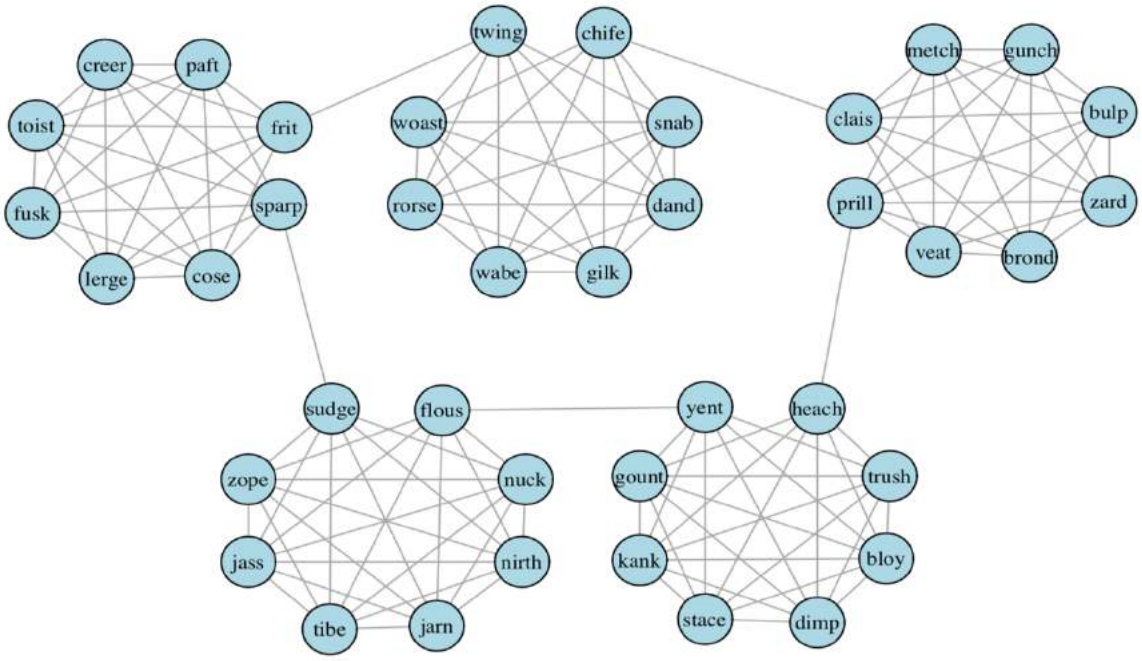


Figure 1. The network architecture used to generate stimulus streams in Studies 1 and 2. Each node represents a pseudoword, and edges represent the co-occurrence of two pseudowords in a continuous sequence

have been shown to have certain limitations (Yarkoni, Balota, & Yap, 2008), and indeed some pseudowords shared surface-level similarities, we stress that any *systematic* phonological or orthographic overlap was minimized by our word-to-node randomization procedure. For the purposes of the cover task (described below), we created a scrambled, unpronounceable version of each pseudoword (e.g., *clais* and *gilk* became *aislc* and *igkl*).

Test items. Eighteen short test sequences (length 5-7 pseudowords) were spliced out of the continuous exposure stream. Half of these substrings consisted exclusively of nodes from within one community, while the other half included traversal of a community boundary. Matched pairs of intra- and inter- community substrings were created by equating length (number of items in the string), number of node repetitions (if any), chunk strength (within one standard deviation of the mean; Meulemans & Van Der Linden, 1997), and general position in the exposure stream (first third, second third, etc.).

Procedure The experiment was composed of four phases: familiarization, exposure, test and debriefing. Participants were randomly assigned to one of four conditions consisting of a unique random walk (i.e., ordering of nodes) in the exposure phase and a unique series of test items. Independent of condition, node-to-pseudoword correspondence was randomized (i.e., "node 1" might correspond to *clais* in one participant and *gilk* in another).

Familiarization phase. Participants were told that they would see a list of made up and scrambled words presented in alphabetical order. They then viewed the list of pseudowords and the scrambled words in a series of 1.5-second trials. They were instructed to press [1] if the word on the screen followed the rules of English (these were the pseudowords) and [2] if the word did not follow the rules of English (these were the scrambled versions). To facilitate their understanding of the task, participants first saw two examples: the pseudoword was *corb*, and the scrambled word was *brco*.

Exposure phase. Following the familiarization phase, participants viewed a 1000-trial continuous sequence of individually presented pseudowords. To obtain RT measures across the entirety of the exposure phase, we instructed participants to complete an orthogonal cover task. At each trial, they were asked to press [1] if the pseudoword appeared in its "regular form" and [2] if the pseudoword appeared scrambled (12% of trials). Each pseudoword was presented for 1.5s with no interstimulus interval. Total duration of the exposure phase was 25 minutes.

Test phase. At the conclusion of the exposure phase, participants were presented with 18 pairs of substrings presented simultaneously on the screen, one above the other (position was randomly determined). They selected which of the two short sequences looked more familiar to them based on what they saw during the previous phase of the experiment. We adopted a familiarity-based approach to judging pairs to promote relatively implicit access of knowledge during the test phase. Unlike the exposure phase, the test phase was self-paced (i.e., both sequences stayed on screen until participants made their selection), with an interstimulus interval of 1.5 seconds.

Analysis and Results

Scrambled Word Detection Participants generally succeeded in distinguishing between pseudowords and their scrambled versions (95.8% accurate, SD = 2.4, excluding the three participants who scored below threshold).

Data Exclusions In the exposure phase, data were prepared for analysis by first eliminating any implausible RTs (i.e., less than 100 ms), then by removing RTs that were greater than three standard deviations away from the mean (4.5% of total data). We also removed all scrambled word trials (12% of total data) and any incorrect responses (4.2% of total data). As we were particularly interested in the RT cost associated with crossing between communities, data were then subset to include only nodes corresponding to entry into a new community (transition nodes), as well as boundary nodes immediately prior to that transition (pre-transition nodes).

Exposure Phase In a linear mixed effects model (library *lme4* 1.1-19 in R 3.5.1), RTs were regressed onto the main effects and interaction of Node Type (pre- vs. transition) and Trial (1-1000). All transitions were included in analysis. The model included the fullest random effects structure that allowed the model to converge: a random intercept for participant and a by-participant random slope for Node Type, Trial, and their interaction.

We observed a significant main effect of Node Type (β = 10.080, t = 2.310, p = 0.022), indicating a processing cost for transition nodes compared to pre-transition nodes. The main effect of Trial (β = -22.798, t = -3.191, p = 0.003) was also significant, an expected finding given that participants were likely to become faster overall at executing button presses. No interaction between Node Type and Trial was observed (β = -6.911, t = -1.477, p = 0.150).

Test Phase Accuracy scores from the posttest did not differ significantly from chance (t(29) = 1.161, p = 0.255). When a post-hoc analysis (mixed logit model) was run to determine whether accuracy was affected by the length of sequences (5 vs 6 word sequences: β = 0.103, z = 0.959, p = 0.341; 5,6 vs 7 word sequences: β = 0.054, z = 0.860, p = 0.390), position on the screen (top or bottom) (β = -0.107, z = -1.208, p = 0.227), or trial number (β = 0.027, z = 0.305, p = 0.761), we continued to observe no significant effects.

Study 2: Community Structure and Word-Level Recognition

Study 1 offers evidence of a cross-community RT increase as learners viewed sequences of written pseudowords. Online measures, collected during the exposure phase, serve to

demonstrate learners' expectation that words within a community should co-occur in time. When that expectation was violated by entry into a new community, RTs reflected a processing penalty. Despite these promising results, we found no evidence that participants applied this knowledge offline as they made substring-level familiarity judgements. Successful language acquisition requires not only the accumulation of statistical regularities, but also accessing that accumulated knowledge in varied contexts. Therefore, the focus of Study 2 was on a post-exposure measure that would speak to the role of community structure in the latter process.

Materials and Methods

Participants 37 neurologically normal participants (9 male, 28 female; ages 18-21) participated in this study. They were recruited from the undergraduate psychology research pool at Pennsylvania State University and were granted course credit for their participation. All participants provided informed consent. Four participants were excluded for cover task performance below the pre-determined threshold used in Study 1.

Stimuli Pseudowords and the graph used to generate the exposure streams were identical to those used in Study 1. However, we increased the length of random walk by 40% in order to ensure participants were receiving sufficient exposure before completing a post-test. For the test phase, we developed a new approach to evaluating the influence of network architecture on retrieval of knowledge following initial learning.

Test items. Our method represents an extension of a classic paradigm developed by Meyer & Schvaneveldt (1971). In that pioneering study, participants completed a lexical decision task on various pairs of words and pseudowords. Compellingly, RTs for pairs of semantically related words were significantly faster than RTs for pairs of semantically unrelated words. Instead of asking whether *semantic* similarity influences retrieval processes, we ask instead whether community structure, or *temporal* similarity, influences retrieval. Here, we test the hypothesis that participants will be faster to make old/new judgements on pairs of words drawn from the same community relative to those drawn from distant communities.

We created 75 new pseudowords which were not seen in the exposure phase ("new words"). We then selected 15 non-boundary pseudowords ("old words") from the exposure phase (3 from each community). These old words were combined exhaustively to form 95 pairs in which items varied by community distance. Next, each old word was paired once with three new words (45 pairs). Finally, the 30 remaining old words were then paired with each other (15 pairs). In total, 165 pairs were created.

For the purposes of analyses, distance between items in a pair was construed as follows: a community distance of 0 meant the pair came from the same community (e.g. *creer* and *toist* in Figure 1). A community distance of 1 meant that the nodes were drawn from adjacent communities (e.g. *creer* and *twing*). A community distance of 2 meant that the nodes were two communities apart (e.g. *creer* and *metch*). There could be no measurement of community distance between old and new words, as the new words were not present in the exposure stream.

Procedure With the exception of the test phase, described below, procedures for Study 2 mirrored that of Study 1. Due to the increased number of trials presented during the exposure phase, its duration was 35 minutes.

Test phase. Participants were simultaneously presented with both items in a pair, one word above the other. Participants pressed [f] for "familiar" if both items had been seen in the exposure phase and [n] for "not familiar" if one or both of the items were new. All new words were only presented once to minimize confusion during the test phase. Trials were self-paced and separated by a 1.5 second blank screen. The order of the pairs and the position (top or bottom) of all pseudowords was randomized across participants.

Analysis and Results

Scrambled Word Detection Participants generally succeeded in distinguishing between pseudowords and their scrambled versions (93.5% accurate on average, SD = 5.6, excluding the four participants below threshold).

Data Exclusions For the exposure phase, data trimming techniques were identical to those described in Study 1 (17.0% of total data removed). Similarly, we subset trials to include only transition and pre-transition nodes.

For the test phase, we removed data corresponding to incorrect trials and any RTs greater than 3 standard deviations from the mean (total data loss = 11.3%).

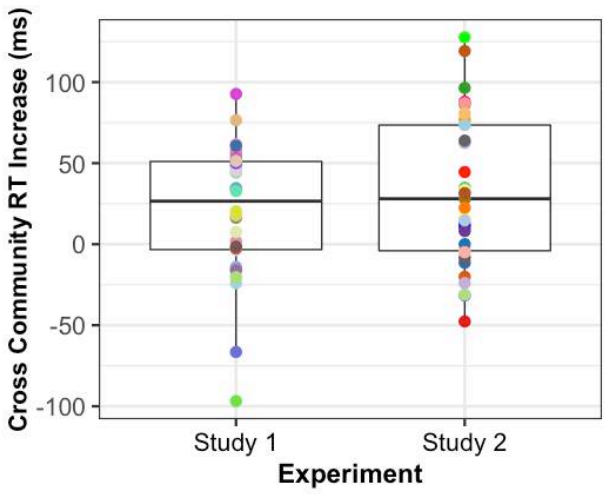


Figure 2. Cross community RT increase for Studies 1 and 2. Values included in the boxplot were calculated by subtracting, for each participant, mean RTs for pre-transition nodes from mean RTs for transition nodes.

Exposure Phase Similar to the previous study, RTs were regressed onto the main effects and interaction of Node Type (pre- vs. transition) and Trial (1-1400). The model included the fullest random effects structure that allowed the model to converge: a random intercept for participant and a by-participant random slope for Node Type, Trial, and their interaction.

Again, we observed a significant main effect of Node Type (β = 15.582, t = 3.782, p = 0.0002), indicating a processing cost for transition nodes compared to pre-transition nodes. The main effect of Trial (β = -18.251, t = -2.609, p = 0.014) was also significant. As in Study 1, no significant interaction between Node Type and Trial was observed (β = -0.924, t = -0.226, p = 0.821). Cross-community RT increases from both Study 1 and Study 2 are presented in Figure 2.

Repetition priming. Prior work examining the influence of community structure on RT patterns has addressed the potential for perceptual priming effects (e.g., Karuza et al., 2017; Kahn et al., 2018). It is well known that humans are faster to process a stimulus that they have seen recently. Though we propose that priming can in fact be considered a form of learning (see e.g., Chang, Dell, Bock & Griffin, 2000), we make contact with prior work by adding to both exposure phase models (Studies 1 and 2) the following measures of repetition priming: Lag10 and Recency. Lag10 indexes the number of times a particular node has been seen in the last 10 trials. Recency indexes the number of trials that have elapsed since a given node was last seen in the exposure stream. When adding these new predictors to our models, the main effect of Node Type was no longer significant (Study 1: β = 2.345, t = 0.441, p = 0.660; Study 2: β = 5.906, t = 1.104, p = 0.273).

Test Phase As in Meyer & Schvaneveldt (1971), our dependent measure of interest was RT for the old/new judgements. Given our lengthy exposure phase, and the fact we never repeated any of the "new words" during the test phase, participants attending to the test phase should have been able to easily and accurately make judgements about the novelty of items in the word pairs. Accuracy scores, though high (88.0% correct overall, SD = 9.9), were not our measure of interest. Rather, we were interested in whether RTs would vary as a function of the distance between nodes in a pair, with the fastest RTs for nodes within the same community. Thus, we imposed a cut-off of 75% accuracy on the test phase to exclude participants who were not complying with this relatively simple task, resulting in the exclusion of three additional participants. We note that without the exclusion of these participants, the significant results reported below do not hold.

Response times from the old/new judgments were regressed onto main effects of Community Distance (reverse-Helmert coded to reflect an increase in processing cost as distance increased) and Trial (1-165; intended to capture general task adaptation). Results are summarized in Table 1. Participants were fastest to respond to pseudoword pairs drawn from the same community relative to pseudowords drawn from the two adjacent communities. Unsurprisingly, participants were faster when responding to pseudowords pairs when they had seen both pseudowords before, compared to pairs in which when one or both of those pseudowords was new (Figure 3).

Table 1: Coefficients, t-values, and p-values for each predictor in a model examining the effect of Community Distance and Trial on participants' RTs for old/ new judgments (Study 2).

Predictor	Results
Community Distance (1 vs. 0)	β = 0.018, t = 2.062, p = 0.041 *
Community Distance (2 vs. 0,1)	β = 0.005, t = 1.040, p = 0.302
Community Distance (new vs. 0,1,2)	β = 0.024, t = 4.480, p = 0.0001 ***
Trial	β = -0.018, t = -1.626, p = 0.115

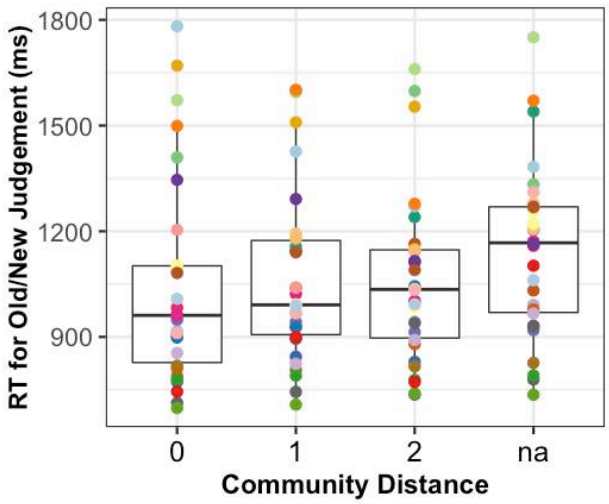


Figure 3. Boxplot of RTs for old/ new judgments on word pairs (Study 2). Values included in the boxplot were calculated by averaging, for each color-coded participant, mean RTs for nodes within a community (community distance = 0), from adjacent communities (= 1) and from non-adjacent communities (= 2). "NA" signifies that at least one word in the pair had not been seen by participants during exposure.

Discussion

We present data from two related studies demonstrating that learners are attuned to the network architecture underpinning continuous streams of linguistic stimuli. Specifically, we show that participants exhibited an increase in processing times when transitioning from one community of words to the next, suggesting that their expectations about upcoming input were influenced by the presence of element clusters in the sequence. As previous investigations into learners' sensitivity to network architectures have taken place

exclusively in the visuomotor domain (Schapiro et al., 2013; Karuza et al., 2017; Kahn et al., 2018; Tompson et al., 2018), one notable contribution of the present work is that it speaks to the potential domain-generality of this learning mechanism.

Complex network analysis of natural language has consistently revealed that, among other properties, community structure may be essential to the organization of the mental lexicon. To varying degrees, real-world networks in which edges represent phonological overlap, semantic relatedness, and temporal co-occurrence, display this property (De Deyne, Verheyen, & Storms, 2016). The present experiments break new ground in that they demonstrate that community structure is not only an emergent property of language, but also a form of high-level regularity that can guide sequence-level learning. Perhaps of greatest interest, we show through a post-exposure measure in Study 2 that temporal overlap can be translated into an accessible representation, as evidenced by the influence of community distance as participants completed a subsequent word recognition task.

At first blush, it is potentially surprising that we observe no significant interaction between Node Type and Trial. In other words, the magnitude of the cross-community RT increase did not change significantly over the course of exposure. However, these results align with previous findings suggesting that sensitivity to community structure may emerge very early in exposure (e.g., Karuza et al., 2017; Karuza, Kahn, & Bassett, 2019). To be clear, we do find a key point of divergence between the present findings and the existing literature on community structure in visuomotor sequences. Specifically, the effect of traversing an inter-community edge was substantially weakened by the inclusion of nuisance regressors intended to account for repetition priming. While there are several possible explanations for this pattern of results, we narrow in on two of them. First, we studied stimulus streams generated from a significantly larger graph than those used in related experiments (i.e., a total of 40 nodes relative to 15). Participants therefore observed far fewer unique edge traversals throughout the course of the experiment. Perceptual priming may have an inflated effect when learners are exposed to more varied stimulus streams in which nodes are repeated only a handful of times. Second, the choice to include pronounceable pseudowords with relatively few real-word orthographic and phonological neighbors meant that these features of our stimuli may have also exerted an undue influence on processing times (Vitevitch, Chan, & Roodenrys, 2012). This source of noise, coupled with some phonological overlap between the pseudowords themselves (e.g., *wabe* and *woast*), may have also contributed to null results obtained for the substring comparison post-test of Study 1. We reiterate that the randomization of word-to-node mapping should have minimized these effects. Nevertheless, evaluation of the full impact of phonological and orthographic neighborhood, defined in terms of the extent of overlap with existing English words as well as among stimulus items themselves, will be an important area of future study. It is possible, for example, that cross-community RT effects shift in magnitude in cases where pseudoword stimuli have an extremely high number of real-word neighbors.

Taken together, this set of results opens up a number of intriguing future directions not limited to investigations into learners' sensitivity to multiple layers of structure (e.g., through the construction of multiplex networks that simultaneously take into account phonological and temporal overlap; Stella, Beckage & Brede, 2017). In a broader context, formalization of the relationship between linguistic network structure and learning could add substantially to discussions regarding how language networks change with development (Ke & Yao, 2008) or why they display certain characteristic properties in special populations (Beckage, Smith & Hills, 2011). On a final note, decreased sensitivity to statistical associations has been linked to disorders ranging from Broca's aphasia (Goschke, Friederici, Kotz, & van Kampen, 2001) to dyslexia (Schmalz, Altoe, & Mulatti, 2017) and developmental language disorder (Lammertink, Boersma, Wijnen, & Rispens, 2017). Extending these lines of inquiry to reveal potential impairments in the extraction of network-level patterns could have powerful consequences, not only in terms of informing rehabilitative practices but also in deepening our understanding of language acquisition more generally.

Acknowledgements

The authors are grateful to Joezette Gray and Alex Ferraccio, for assistance with data collection and stimulus creation, and to David Wiegand for helpful comments on this work.

References

Arbesman, S., Strogatz, S. H., & Vitevitch, M. S. (2010). The structure of phonological networks across multiple languages. *International Journal of Bifurcation and Chaos*, *20*(03), 679-85.

Bassett, D. S., & Sporns, O. (2017). Network neuroscience. *Nature Neuroscience, 20*(3), 353-64.

Beckage, N., Smith, L., & Hills, T. (2011). Small worlds and semantic network growth in typical and late talkers. *PloS one*, *6*(5), e19348.

Borge-Holthoefer, J., & Arenas, A. (2010). Semantic networks: Structure and dynamics. *Entropy*, *12*(5), 1264-1302.

Chang, F., Dell, G. S., Bock, K., & Griffin, Z. M. (2000). Structural priming as implicit learning: A comparison of models of sentence production. *Journal of Psycholinguistic Research, 29,* 217–30.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*(6), 407-28.

De Deyne, S., Verheyen, S., & Storms, G., (2016). Structure and organization of the mental lexicon: a network approach derived from syntactic dependency relations and word associations. In Mehler, A., Lücking, A., Banisch, S., Blanchard, P., & Job, B., *Towards a Theoretical*

Framework for Analyzing Complex Linguistic Networks. Berlin: Springer Berlin Heidelberg.

Ferrer i Cancho, R. F., Solé, R. V., & Köhler, R. (2004). Patterns in syntactic dependency networks. *Physical Review E*, *69*(5), 051915.

Goldstein, R., & Vitevitch, M. S. (2014). The influence of clustering coefficient on word-learning: how groups of similar sounding words facilitate acquisition. *Frontiers in Psychology*, *5*, 1307.

Goschke, T., Friederici, A. D., Kotz, S. A., & van Kampen, A. (2001). Procedural learning in Broca's aphasia: Dissociation between the implicit acquisition of spatio-motor and phoneme sequences. *Journal of Cognitive Neuroscience, 13(3)*, 370-388.

Kahn, A. E., Karuza, E. A., Vettel, J. M., & Bassett, D. S. (2018). Network constraints on learnability of probabilistic motor sequences. *Nature Human Behaviour*, *2*(12), 936.

Karuza, E. A., Kahn, A. E., & Bassett, D. S. (2019). Human sensitivity to community structure is robust to topological variation. *Complexity*, *2019*.

Karuza, E. A., Kahn, A. E., Thompson-Schill, S. L., & Bassett, D. S. (2017). Process reveals structure: How a network is traversed mediates expectations about its architecture. *Scientific Reports*, *7*(1), 12733.

Karuza, E. A., Thompson-Schill, S. L., & Bassett, D. S. (2016). Local patterns to global architectures: influences of network topology on human learning. *Trends in Cognitive Sciences*, *20*(8), 629-40.

Ke, J., & Yao, Y. A. O. (2008). Analysing language development from a network approach. *Journal of Quantitative Linguistics*, *15*(1), 70-99.

Lammertink, I., Boersma, P., Wijnen, F., & Rispens, J. (2017). Statistical learning in specific language impairment: A meta-analysis. *Journal of Speech, Language, and Hearing Research*, *60*(12), 3474-3486.

Liu, H. (2008). The complexity of Chinese syntactic dependency networks. *Physica A: Statistical Mechanics and its Applications, 387*(12), 3048-58.

Meulemans, T., & Van der Linden, M. (1997). Does the artificial grammar learning paradigm involve the acquisition of complex information? *Psychologica Belgica, 37*(1-2), 69-88.

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, *90*(2), 227-34.

Rastle, K., Harrington, J., & Coltheart, M. (2002). 358,534 nonwords: The ARC Nonword Database. The Quarterly Journal of Experimental Psychology: A, 55(4), 1339-62.

Schapiro, A. C., Rogers, T. T., Cordova, N. I., Turk-Browne, N. B., & Botvinick, M. M. (2013). Neural representations of events arise from temporal community structure. *Nature Neuroscience*, *16*(4), 486-492.

Schmalz, X., Altoe, G., & Mulatti, C. (2017). Statistical learning and dyslexia: A systematic review. *Annals of Dyslexia, 67(2),* 147-162.

Scott, J. (2017). *Social Network Analysis*. London: SAGE Publications.

Siew, C. S. Q., & Vitevitch, M. S. (2016). Spoken word recognition and serial recall of words from components in the phonological network. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42*(3), 394-410.

Stella, M., Beckage, N. M., & Brede, M. (2017). Multiplex lexical networks reveal patterns in early word acquisition in children. *Scientific Reports*, *7*, 46730.

Steyvers, M., & Tenenbaum, J. B. (2005). The large- scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science*, *29*(1), 41-78.

Tompson, S. H., Kahn, A. E., Falk, E. B., Vettel, J. M., & Bassett, D. S. (2018). Individual differences in learning social and nonsocial network structures. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 45*(2), 253-71.

Vitevitch, M. S. (2008). What can graph theory tell us about word learning and lexical retrieval? *Journal of Speech, Language, and Hearing Research, 51*(2), 408-22.

Vitevitch, M. S., Chan, K. Y., & Roodenrys, S. (2012). Complex network structure influences processing in long-term and short-term memory. *Journal of Memory and Language*, *67*(1), 30-44.

Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, *15*(5), 971-979.

Orthogonal multi-view three-dimensional object representations in memory revealed by serial reproduction

Thomas A. Langlois (thomas.langlois@berkeley.edu)
Department of Psychology, University of California, Berkeley
Berkeley, CA 94720-1650 USA

Nori Jacoby (nori.viola@gmail.com)
Max Planck Institute for Empirical Aesthetics
Grüneburgweg 14, 60322 Frankfurt am Main, Germany

Jordan Suchow (jws@stevens.edu)
Stevens Institute of Technology
Hoboken, NJ 07030, USA

Thomas L. Griffiths (tomg@princeton.edu)
Department of Psychology, Princeton University
Princeton, NJ 08540, USA

Abstract

The internal representations of three dimensional objects within visual memory are only partially understood. Previous research suggests that 3D object perception is viewpoint dependent, and that the visual system stores viewpoint perspectives in a biased manner. The aim of this project was to obtain detailed estimates of the distributions of 3D object views in shared human memory. We devised a novel experimental paradigm based on transmission chains to investigate memory biases for the 3D orientation of objects. We found that memory tends to be biased towards orthogonal diagrammatic perspectives aligned with the ends of the standard basis for a set of common 3D objects, and that these biases are strongest for side views as well as top or bottom views for a small set of bilaterally symmetric objects. Finally, we found that views sampled from the modes were easier to categorize in a recognition task.

Keywords: Memory; 3D object perception; Serial reproduction; Iterated learning; Vision.

Introduction

Humans do not possess photographic memories of the things they see. Instead, visual memory is known to be biased towards systematic and simplified representations. The perception of 3D objects is known to be viewpoint dependent, but detailed estimates of the distributions of 3D object views in shared human memory remain unknown. For a given object, towards what views does visual memory tend to be biased? Are the number of views the same across different objects? How many views are there? Evidence from prior work points to systematic viewpoint-specific biases in 3D object perception such as so-called "canonical" views of common everyday objects (Palmer & Rosch, 1981). Canonical views are associated with improvements in categorization accuracy and recognition (as measured using response-time latencies). While the human visual system is largely robust to perspective transformations, this work provided early evidence for viewpoint dependence in human object perception, a finding that was corroborated in subsequent work (Bülthoff et al., 1995). However, none of this work fully characterized the object-specific distributions of views that bias visual memory, and provided mostly indirect evidence for them. We therefore attempt to provide a detailed picture of the structure of memory biases for the orientation of 3D objects.

We aimed to uncover the distributions of 3D object views in shared human memory. Doing so is of particular interest to disambiguate theoretical explanations for viewpoint dependence in 3D object perception, and to determine if biases in remembered views of objects correspond to canonical views. Two theoretical explanations have been suggested in order to explain canonical views: the "frequency hypothesis" and the "maximal information hypothesis" (Mezuman & Weiss, 2012). The "frequency hypothesis" states that privileged views correspond to the views that are most commonly taken when viewing or interacting with everyday objects, while the "maximal information hypothesis" states that these views change the least under small local perspective transformations. The "frequency hypothesis" is most consistent with the notion of a statistical "prior" in Bayesian accounts of perception and memory. However, it remains an open question as to whether memory representations for 3D objects resemble canonical views, and if these representations are shaped by statistical priors.

To answer this question, we used transmission chains adapted to a 3D orientation memory experiment. Under experimentally verifiable conditions, transmission chains are known to approximate samples from shared priors (Xu & Griffiths, 2010), and can be used to characterize shared biases in reconstructive memory. In this paper, we start by outlining past computational approaches and empirical findings regarding 3D object representations, as well as theoretical properties of transmission chains. Next, we present our novel findings revealing hitherto unknown distributions of 3D memory biases for a range of everyday objects. We find that these distributions are characterized by systematic patterns of biases towards diagrammatic orthogonal views that appear to be aligned with the faces of the objects (strong side views,

front and back views, top and bottom views). These views do not appear to match known canonical views, which are typically semi-profile views, although they are consistent with past findings that revealed similar biases in visual inspection of novel objects in adults (Perrett et al., 1992), as well as infants (Pereira et al., 2010). We also find that these views were associated with improved categorization accuracy relative to views sampled from areas far from the modes in these distributions.

Background

Transmission chains and experimental methods Transmission chains are analogous to the so-called "telephone game." In the most famous and early example, Bartlett had a series of people reproduce a drawing of an owl hieroglyph, and as the the reproductions of the image progressed through the chain, what began as an imperfect but recognizable facsimile of the hieroglyph morphed into an image of a cat (Bartlett, 1932), revealing that the participants shared a common bias to distort the unusual image into an image for which they had a strong collective prior.

Transmission chains have since been adopted to study phenomena in many fields, including evolutionary biology, cognitive science, anthropology, vision science, and music cognition (Kirby et al., 2008; Jacoby & McDermott, 2017; Lew & Vul, 2015). A recent analysis of reconstruction from memory examined how information should change as it is transmitted through a chain of rational agents (Xu & Griffiths, 2010). Under the rational analysis, reconstruction from memory is defined as the problem of inferring the most accurate state of the world despite a noisy or imperfect sensory input (such as an imperfect memory trace of a scene or an object in the world). Using the framework of Bayesian statistics, this problem can be captured as follows: Previous experience is characterized by a prior distribution over possible world states (a hypothesis space of all conceivable world states, such as all possible 3D orientations of an object). The posterior is computed by integrating that prior with the likelihood, which in this case simply describes the probability of observing a world state (such as an object in a particular orientation), given a hypothesis about the true state of the world. In this work, (Xu & Griffiths, 2010) found that a transmission chain populated by rational Bayesian agents defines a Markov chain with the following transition probabilities:

$$p(x_{n+1} \mid x_n) = \int p(x_{n+1} \mid \mu) p(\mu \mid x_n) d\mu,$$

where x is a noisy stimulus (such as noisy recollection of the orientation of a previously viewed object) and μ is the true state of the world that generated that stimulus. This Markov chain captures the probability of a new stimulus x_{n+1} being created as a reconstruction of a previously seen stimulus x_n in each iteration in the transmission chain, and has a stationary distribution which defines the probability of observing a stimulus x when μ is sampled from the prior:

$$p(x) = \int p(x \mid \mu) p(\mu) d\mu.$$

This process approximates a Gibbs sampler for the joint distribution on x and μ defined by multiplying $p(x \mid \mu)$ and $p(\mu)$. In other words, assuming that participants share common inductive biases, the transmission chain will converge to a sample from their shared prior.

Computational theories of 3D representations To date, a significant body of work has explored the nature of human representations of 3D objects and a great deal of experimental work has been done to elucidate the characteristics of human perceptual representations of 3D objects and scenes. (Palmer & Rosch, 1981) provided early evidence for the existence of privileged "canonical" views that facilitate 3D object recognition, in keeping with principles of categorization (Rosch, 1999) that introduced the notion of "prototype exemplars." Later work introduced the recognition-by-components (RBC) theory of image understanding (Biederman, 1987). This work proposed that representations of objects in memory are accessed when components ("geons") derived from perceptual mechanisms (Lowe, 2012; Rock, 1983) are combined, and that these components form a perceptual basis for a "componential representation of real world objects in memory." A third computational theory argues that objects are represented as lists of viewpoint-invariant properties (A piano has keys, pedals, legs) (Bülthoff et al., 1995; B. Tversky & Hemenway, 1984; A. Tversky, 1977), or by points in abstract multidimensional feature spaces (Carr et al., 2001; **?**; Su et al., 2015).

Theories based on list-based feature descriptors or viewpoint-invariant parts have been difficult to reconcile with experimental data showing systematic view-specific variations in human response-time latencies and recognition accuracy (Bülthoff et al., 1995; Tarr et al., 1998). These results have tended to favor theories that postulate viewpoint-specific and largely 2D representations (Vetter et al., 1995; Bülthoff et al., 1995) as forming the basis for human object representations. However, to our knowledge, little work has been done to devise an experimental method for revealing the distributions of viewpoint-specific biases in memory representations.

Canonical perspectives were discovered for objects that were bilaterally symmetric due to experimental constraints, and although (Palmer & Rosch, 1981) confirmed the presence of privileged views for each, it is possible that other canonical views, such as the mirror images of bilaterally symmetric objects exist. In fact, work using online images returned by search engines estimated the modes of the distribution of 3D perspectives for a variety of objects, and found that canonical views for bilaterally symmetric objects are typically bi-modal (Mezuman & Weiss, 2012). In this paper we adapted transmission chains to a memory paradigm in which we probed collective biases in reconstructive memory for the 3D orientation of a handful of everyday objects in order to uncover any and all biases in 3D reconstructive memory.

Methods

Participants

All participants were recruited online using Amazon Mechanical Turk and gave informed consent. according to a protocol was approved by The Committee for the Protection of Human Subjects (CPHS) at the University of California, Berkeley. Each experiment required approximately 100 participants.

Stimuli

The stimuli used in these experiments were 3D objects that could be viewed from any angle by rotating a camera oriented towards the origin of the object, and at a fixed distance (travelling on the surface of a sphere around the object) and with the camera tilted (in a direction tangent to the sphere). We started with a detailed mesh model of a typical teapot, and shoe. In addition, we used grayscale versions of the teapot and shoe, as well as a grayscale 3D model of a car, alarm clock, armchair, coffee maker, camera, and grand piano, see Figure 1A. We selected objects matching the objects in (Palmer & Rosch, 1981) as closely as possible.

Procedure

For each object, we ran a serial reproduction experiment with 250 chains and 20 iterations (see Figure 1B). Participants viewed timed displays of the 3D object. The chains were initialized as camera views over the surface of a unit sphere with the object in the center. The camera frame orientation was always oriented towards the center of the object, but was tilted at random angles orthogonal to the sphere (the "up" vector, see Figure 1D). The position of the camera and the view were sampled uniformly from the Haar measure on $SO(3)$ (Perez-Sala et al., 2013). Following the timed display, and 1000 ms retention phase when the screen went blank, a probe screen containing the object at a new random orientation was shown.

Participants were instructed to orient the object (which is equivalent to rotating the camera view) so that it matched the original orientation of the object that was shown during the initial timed display. Participants were not given time constraints during the probe, and could change their responses as many times as they needed. The object on the screen could be rotated by means of the mouse, as well as a set of buttons (see Figure 1C). Participants were given 10 practice trials during which the initial display was shown for 4000 ms in order to familiarize them with the nature of the task, and the user interface. Only after they completed the practice trials was the presentation time reduced to 1000 ms. In addition, they were given trial-by-trial feedback based on their performance (either a green message saying "Well done! Your response was sufficiently accurate", or a red message stating: "Your response was insufficiently accurate"), see Figure 1C.

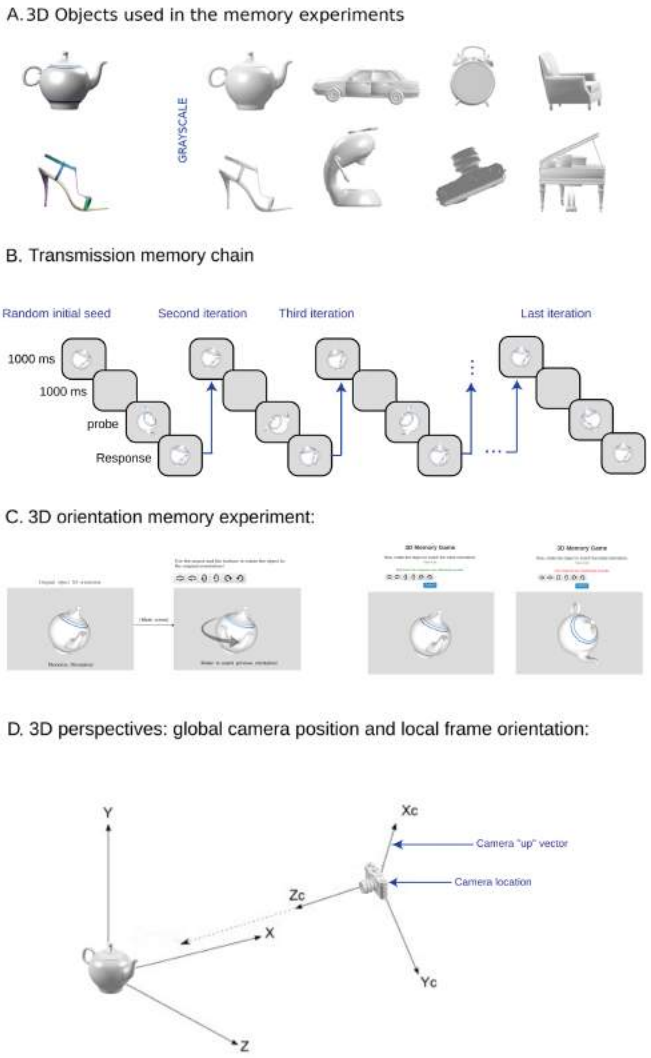


Figure 1: 3D objects, experiment structure, task and geometry. A. Textured and grayscale 3D objects used in the transmission chain experiments. B. Transmission chain structure: A 3D view of an object (teapot) is initialized somewhere at random over a unit sphere. This view is presented as a stimulus to a subject who then reconstructs this view from memory. The subject's response is then presented as the stimulus to a second subject, who must reproduce this second view, and so on. C. The experiment instructions and trial structure. Participants could rotate the object with the mouse and a set of buttons displayed over the image. They were instructed to reproduce the view they saw as accurately as possible, and were given feedback on their performance. D. Geometry of 3D object views adopted in the experiment. Views (cameras) were always positioned on the surface of a sphere centered at the object, and was always pointed towards the center of the sphere (towards the object). The local frame of the camera could vary according to the "up" vector, which controls the tilt of the camera

Results

By the final iteration of the transmission chain process, a clear pattern emerges: 3D views are biased towards a small set of orthogonal "diagramatic" views that are aligned with the top,

bottom, and side views of the objects. In some cases, the views in the modes correspond to the front and back (for the clock in particular), see Figure 2 for the results obtained with the textured teapot and shoe. While all the starting views of the chains are camera views sampled uniformally over the sphere surrounding the objects, the distributions quickly change and become clustered around four distinct modes as the chains progress. Figure 2A shows the initial distribution, the distribution at the 5th, 10th, 15th and 20th iteration of the transmission chains for the teapot and shoe.

Figure 2B shows the distributions of all points across all iterations for the shoe and teapot. In addition, the four modes with respect to the camera directions are plotted in four colors for both distributions. Next to each of these distributions, we show the corresponding histograms of the angles of the "up" vectors at the modes, where the direction most aligned with the data is centered to 90 degrees for the first two modes of the teapot and shoe. Surprisingly, while the "up" vectors in modes I, II (side views) of the Teapot are centered mainly in one direction, those in modes III and IV (the top and bottom views) show a bimodal distribution (top and bottom views are remembered with the handle and spout oriented vertically, while the side views show them to be oriented horizontally, orthogonal to the vertical orientations in the top and bottom views). We don't find this pattern in the case of the shoe, where the distributions of "up" vector angles were unimodal for all modes (I, II, III, and IV). This suggests that memory representations contain interaction patterns where some objects are memorized with a specific location *and* orientation, while memory for views of other objects are not necessarily associated with particular angular orientations. The columns on the far right of Figure 2B show spherical kernel density estimates (KDEs) of the final iteration data oriented according to the top four modes. Thumbnail insets to the right of the KDE modes show the corresponding object views. For both objects, the top two views are side views, while the remaining two modes correspond to the top and bottom views.

In order to verify if our chains showed convergence, we measured the mean copying error of the camera views for the textured teapot and shoe objects (See Figure 2C). The copying error was computed separately for each iteration by averaging the difference between the remembered camera view responses and stimulus views. We found that the copying error tends to reduce over the course of the experiment. Indeed, whereas the copying error for the first iterations was significantly smaller compared with the last iteration ($t(364)=6.6$, $p<0.001$ and $t(386)=5.6$, $p<0.001$ for the teapot and shoe, respectively), the difference between the copying error in the last iteration was not significantly different from the preceding four iterations ($p > 0.1$). For all cases this holds true even with Bonferroni corrections for multiple comparisons). This suggests that convergence occurs by the last five iterations of the chains.

In order to control for effects of colors and texture on 3D memory biases, as well as to control these factors for the

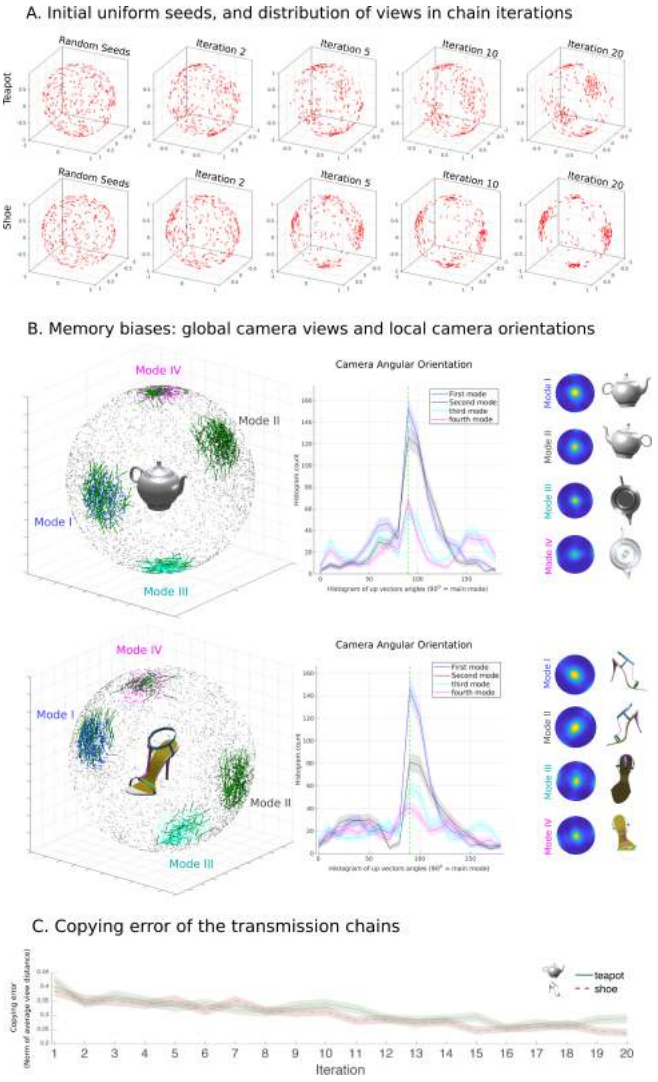


Figure 2: Transmission chain results for a teapot and shoe. A. Scatter plots showing camera views and "up" vectors for four chain iterations, and the initial uniform random seed locations. First row shows results for the teapot (initial seed, 5th, 10th, 15th, and 20th iteration distributions), and second row shows results for the shoe. B. Modes in the 20th and final distributions of views for the teapot and shoe. Four modes are clearly discernible: the side views of the objects, and the top and bottom views. Spherical subplots show a superposition of camera views across all iterations, highlighted are the four modes obtained by the 20th and final iteration of the chains. These correspond to the side views as well as the top and bottom views. The central subplots show histograms of the "up" vector angles, which show the frequency of local camera orientations at each of the modes. They reveal that perspectives in the first two modes (side views in both cases) are biased towards views where the camera is oriented towards a 90 degree angle, which yields views of the objects that are upright. These views are visualized in the far right columns, for each object, along with views of the modes in spherical Kernel Density Estimates (KDEs) of the 20th iteration data. C. Copying error across the chain iterations.

recognition experiments that follow, we ran a set of grayscale objects. Figure 3A the data for the novel objects. The results for the teapot and shoe were largely consistent with the results obtained with the textured versions of these objects, although only three clear modes were observed for the grayscale shoe (side views, and top view). Similarly, the modes for the car reveal four orthogonal views: left and right sides, as well as top, and bottom views. For the remaining objects, either three or two modes were present. Only two primary modes revealing frontal views, and back views were observed for the clock. Finally, the results for the remaining objects reveal primarily three orthogonal views. In sum, we find that 3D object memory representations are not equivalent to canonical views, and are characterized by multi-modal biases that may reflect the symmetry of the objects, a finding that corroborates findings revealing the presence of bi-modal views in distributions of online images (Mezuman & Weiss, 2012), although those were not views aligned with the faces of the objects, nor were they necessarily orthogonal. The memory representations we uncovered replicate past findings showing systematic biases towards the same views in a variety of visual inspection tasks in both infants and adults, suggesting that memory biases may be influenced by encoding precision and angular discrimination.

Figure 3 shows the results of a categorization experiment in which we compared the categorization accuracy for the set of eight grayscale objects when they were presented from views sampled in the modes of our memory KDEs, or from views far from the modes (sampling 4 nearest neighbors around the points that were farthest from the modes on the sphere, in the initial seed distributions of the chains). Figure 3B shows example views, and the experimental task: subjects were presented with a view for 100 ms, and then asked to categorize the object. The eight object labels were shown, as well as two additional labels ("house", "horse"). Figure 3B shows recognition d' results as a function of view type. We found that views of the shoe, clock, car, teapot, and coffee machine were recognized more accurately when they were sampled from the modes in our KDEs ($p < 0.001$ in all cases, following a Bonferroni correction for multiple comparisons). Overall, views sampled from the modes were associated with improved classification accuracy ($p < 0.0001$).

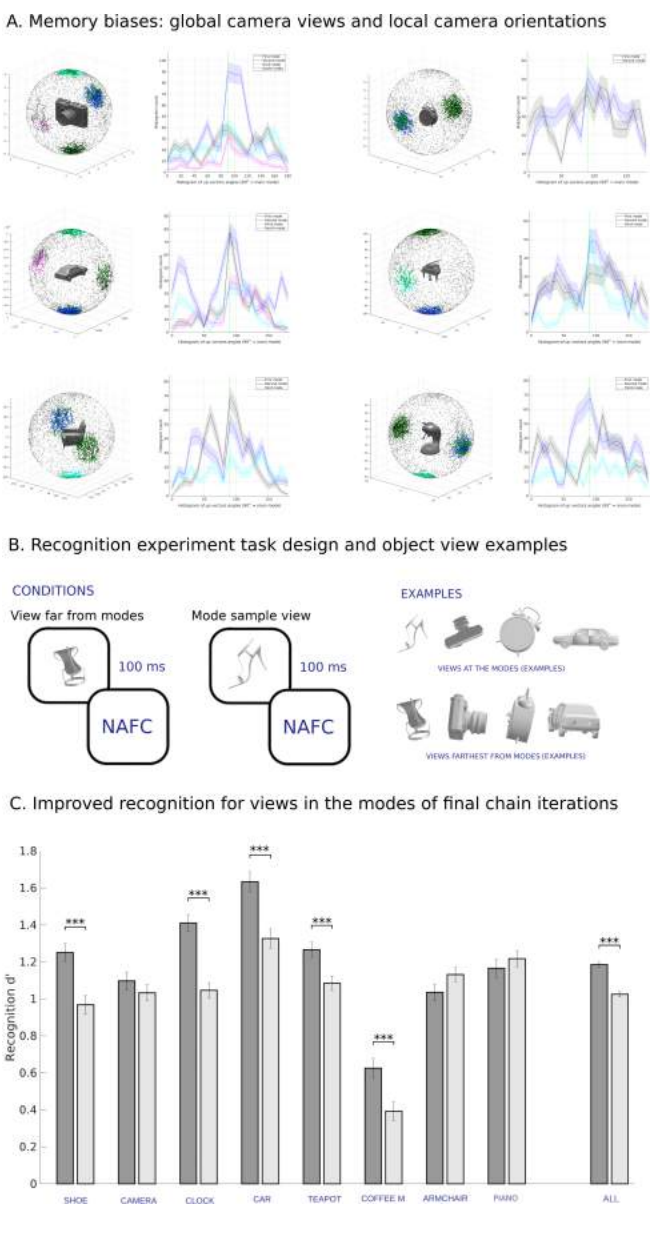


Figure 3: Recognition experiments. A. Memory biases for global camera views and local camera orientations for grayscale objects. B. Recognition experiment task design and view examples. For each object, participants were presented with a view sampled either from one of the perspective modes obtained from the final chain iteration, or from a point farthest from one of the final modes sampled from the uniform seed distribution, for 100 ms. They were then asked to select the correct object name from a list of possibilities. C. Recognition d' results for each of the objects, and for all the objects. Results show that in most cases, participants were more likely to select the correct object label when the view shown was sampled from one of the modal views sampled form the final chain iteration. Error bars correspond to 1000 bootstrapped samples of the data, with replacement. We used the Bonferroni correction for multiple comparisons.

Discussion

We found the use of transmission chains to be particularly sensitive to characterizing shared 3D memory biases. The biases we observed for a small set of bilaterally symmetric everyday objects are highly systematic and not identical to known canonical views. They are strongly diagrammatic views of the sides, top and bottom, or front and back faces of the objects. In this respect, they resemble the bimodal characteristics of the distributions of online images estimated by (Mezuman & Weiss, 2012), although the diagrammatic aspects of these views are more reminiscent of well-known biases in visual inspection of 3D objects Perrett et al. (1992);

Pereira et al. (2010). However, we did observe differences between objects, with some object representations containing four distinct modes, and others containing fewer (the clock). In addition to finding clear biases in camera locations, we also observed that camera views tended to be consistently oriented upright for the side views, but not for top or bottom views. Finally, we determined that categorization accuracy was higher for views that were sampled from the modes of the distributions we estimated, when compared to views sampled from regions that were farthest from the modes. This suggests that 2D memory representations of 3D objects are informative for recognition.

Finally, using this tool to uncover memory priors for objects that are not bilaterally symmetric, and with different geometries could help determine what factors are responsible for shaping biases in 3D memory representations. Our current findings do not appear to be altogether consistent with statistical priors (the "frequency hypothesis"), since diagrammatic views (especially of the bottom of objects like cars, teapots, and pianos) are not views of these objects that are typically experienced. However, they may be due to variable angular discrimination accuracy, which may be increased for sides that are aligned with the first principal component axes of the objects, and decreased for the shorter sides. Our approach provides a powerful tool for estimating detailed distributions of biases in 3D memory, and can provide an empirical basis for spurring novel theoretical insights on the nature of these representations.

Acknowledgments

This work was funded in part by National Science Foundation grant SPRF-IBSS-1408652 to T.L.G. and J.W.S. and DARPA Cooperative Agreement D17AC00004 to T.L.G and J.W.S. The contents of this paper does not necessarily reflect the position or the policy of the Government and no official endorsement should be inferred.

References

Bartlett, F. C. (1932). Remembering: An experimental and social study. *Cambridge: Cambridge, UK*.

Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychological Review*, *94*(2), 115–147.

Bülthoff, H. H., Edelman, S. Y., & Tarr, M. J. (1995). How are three-dimensional objects represented in the brain? *Cerebral Cortex*, *5*(3), 247–260.

Carr, J. C., Beatson, R. K., Cherrie, J. B., Mitchell, T. J., Fright, W. R., McCallum, B. C., & Evans, T. R. (2001). Reconstruction and representation of 3d objects with radial basis functions. In *Proceedings of the 28th annual conference on computer graphics and interactive techniques* (pp. 67–76).

Jacoby, N., & McDermott, J. H. (2017). Integer ratio priors on musical rhythm revealed cross-culturally by iterated reproduction. *Current Biology*, *27*(3), 359–370.

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. In *Proceedings of the national academy of sciences* (Vol. 105, pp. 10681–10686).

Lew, T. F., & Vul, E. (2015). Structured priors in visual working memory revealed through iterated learning. In *Proceedings of the 37th annual meeting of the cognitive science society.*

Lowe, D. (2012). *Perceptual organization and visual recognition* (Vol. 5). Springer Science & Business Media.

Mezuman, E., & Weiss, Y. (2012). Learning about canonical views from internet image collections. In *Advances in neural information processing systems* (pp. 719–727).

Palmer, S., & Rosch, E. (1981). Chase. p.(1981). canonical perspective and the perception of objects. *Attention and performance IX*, 135–151.

Pereira, A. F., James, K. H., Jones, S. S., & Smith, L. B. (2010). Early biases and developmental changes in self-generated object views. *Journal of Vision*, *10*(11), 22–22.

Perez-Sala, X., Igual, L., Escalera, S., & Angulo, C. (2013). Uniform sampling of rotations for discrete and continuous learning of 2d shape models. In *Robotic vision: Technologies for machine learning and vision applications* (pp. 23–42). IGI Global.

Perrett, D. I., Harries, M. H., & Looker, S. (1992). Use of preferential inspection to define the viewing sphere and characteristic views of an arbitrary machined tool part. *Perception*, *21*(4), 497–515.

Rock, I. (1983). The logic of perception. *Vision Science*.

Rosch, E. (1999). Principles of categorization. *Concepts: core readings*, *189*.

Su, H., Maji, S., Kalogerakis, E., & Learned-Miller, E. (2015). Multi-view convolutional neural networks for 3d shape recognition. In *Proceedings of the ieee international conference on computer vision* (pp. 945–953).

Tarr, M. J., Williams, P., Hayward, W. G., & Gauthier, I. (1998). Three-dimensional object recognition is viewpoint dependent. *Nature Neuroscience*, *1*(4), 275.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(4), 327–352.

Tversky, B., & Hemenway, K. (1984). Objects, parts, and categories. *Journal of Experimental Psychology: General*, *113*(2), 169–193.

Vetter, T., Hurlbert, A., & Poggio, T. (1995). View-based models of 3d object recognition: invariance to imaging transformations. *Cerebral Cortex*, *5*(3), 261–269.

Xu, J., & Griffiths, T. L. (2010). A rational analysis of the effects of memory biases on serial reproduction. *Cognitive Psychology*, *60*(2), 107–126.

Whats in a Name, and When Can a [Beep] be the Same?

Jill Lany
University of Liverpool, Liverpool, United Kingdom

Abbie Thompson
University of Notre Dame, Notre Dame, Indiana, United States

Ariel Aguero
University of Notre Dame, South Bend, Indiana, United States

Abstract

Words influence cognition well before infants know their specific meanings. For example, three-month-olds are more likely to form visually-based categories when exemplars are paired with spoken words than with sine-wave tones. We tested whether structure in infants environment can foster this effect. Caregivers often use exaggerated showing gestures when labeling objects, presenting words in synchrony with object motion, and creating amodal temporal structure in auditory and visual modalities. Because attention to amodal structure attenuates encoding information specific to just one modality, we hypothesized that it can lead auditory signals to impact visually-based categorization. Indeed, when 3-month-olds are familiarized to videos in which tones occur in synchrony with object motion, tones subsequently facilitate categorization, just like words. Moreover, familiarizing infants to word-object synchrony enhances their subsequent categorization in the presence of words. These results suggest that structure in infants environment may contribute to the special effects that words have on categorization.

Does the intuitive scientist conduct informative experiments?: Children's early ability to select and learn from their own interventions

Elizabeth Lapidow (elapidow@ucsd.edu) & Caren M Walker (carenwalker@ucsd.edu)
Department of Psychology, University of California, San Diego, La Jolla, CA 92093

Abstract

We investigate whether children preferentially select informative actions and make accurate inferences from the outcome of their own interventions in a causal learning task. Four- to six-year-olds were presented with a novel system composed of two gears that could operate according to two possible causal structures (single or multiple cause). Given the choice between interventions (i.e., removing one of the gears to observe the remaining gear in isolation), children demonstrated a clear preference for the action that revealed the true causal structure, and made subsequent causal judgments that were consistent with the outcome observed. Experiment 2 addressed the possibility that performance was driven by children's tendency to select an intervention that would produce a desirable effect (i.e., spinning gears), rather than to disambiguate the causal structure. The results replicate our initial findings in a context in which the informative action was less likely to produce a positive outcome than the uninformative one. We discuss these results in terms of their significance for understanding both the development of scientific reasoning and the role of self-directed actions in early learning.

Keywords: cognitive development; causal learning; exploration; scientific reasoning; decision-making; experimentation

Introduction

The concept of the learner as an intuitive scientist—forming and evaluating hypotheses about the world—has provided an illuminating and productive model for understanding the mechanisms underlying cognitive development. In particular, 'Theory Theorists' have long advanced the analogy between the processes underlying knowledge acquisition and formal scientific theory change, in which children formulate, test, and rationally revise their intuitive theories in light of new evidence (Gopnik & Wellman, 2012). Indeed, much of what we know about self-directed learning in early childhood (and beyond) appears to resemble the basic inductive processes of science. From infancy, learners are sensitive to statistical information in the data they observe (e.g., Saffran, Aslin, & Newport, 1996; Xu & Garcia, 2008), and use these patterns to infer the abstract causal theories that allow for explanation, prediction, and action in the world (e.g., Carey, 1985; Keil, 1989; Wellman & Gelman, 1992).

However, the scientific process is not limited to passive observation and interpretation of statistical data. Instead, learning as an intuitive scientist also requires that children design, select, and execute informative interventions to evaluate the accuracy of their currently held beliefs and acquire new knowledge. The need for experimentation is especially apparent in the domain of causal learning, where observation alone is often insufficient. Instead, observations must typically be paired with appropriate and informative investigations in order to disambiguate between potential causes or causal structures (Pearl, 2000).

To illustrate, suppose that you notice that the houseplant sitting in a sunny spot on the windowsill has wilted, and the soil in the pot is dry. Multiple causal structures are consistent with this pattern of observation (see Figure 1): It could be that the intense sunlight dried out the soil, and the plants wilted due to this lack of moisture (a causal chain: Figure 1b). Or perhaps this is a variety of plant that requires shade, regardless of moisture. In this case, the sunlight is a direct cause of both wilting and dry soil, independently of one another (a common cause: Figure 1a).

While observation of the world alone cannot disambiguate between these two possibilities, taking specific actions on the world can. Due to the conditional relationship between patterns of intervention and causal structure, manipulating the variables in a system can reveal the causal relationships between them. That is, a learner who knows that variable X is the cause of variable Y *also* knows that intervening to change X will lead to a change in Y. Returning to our houseplant example, you could therefore discover the true casual structure by intervening to change the dryness of the soil—perhaps by watering more often—and then check to see if plants in that spot flourish (indicating a causal chain) or continue to wilt (indicating a common cause).

This makes intervention a powerful tool for determining causal structure, but its usefulness critically requires that the learner recognize and carry out *informative* interventions. For example, while intervening on the sunlight (e.g., by shading the flower pot) will always lead to improving the health of the plant, this desirable outcome would not provide information about the true underlying causal structure (i.e., whether wilting was caused by dry soil or by excess sunlight).

Whether young learners are able to engage in this type of systematic experimentation is a subject of substantial debate. On the one hand, research on exploratory play suggests that even preschool-aged children have an intuitive tendency to produce informative actions that facilitate their learning: Children preferentially explore where they have incomplete or inconsistent knowledge (e.g., Bonawitz, van Schijndel, Friel, & Schulz, 2012; Gweon & Schulz, 2008; Schulz & Bonawitz, 2007), and spontaneously select actions with the potential to improve their epistemic status (Cook, Goodman, & Schulz, 2011). On the other hand, this work

contrasts with decades of research on the development of scientific reasoning, which overwhelmingly reports that even much older children *do not* follow the principles of informative scientific experimentation in their spontaneous actions (Zimmerman & Klahr, 2018): Children struggle with the control and isolation of variables, often designing confounded and confirmatory experiments rather than logically informative ones (e.g., Inhelder & Piaget, 1958; Klahr, Fay, & Dunbar, 1993; Siler & Klahr, 2012; Valanides, Papageorgiou, & Angeli, 2014). Critically, children also appear to select interventions based on their tangible outcomes, rather than their informativeness (e.g., Schauble, 1990; Tschirgi, 1980) (e.g., choosing to shade the plant in the above example).

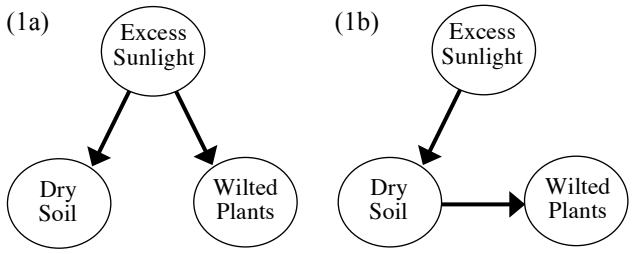


Figure 1: Common cause (a) and causal chain (b) structures.

This apparent preoccupation with producing (or reproducing) effects, rather than testing causal hypotheses, has led some researchers to suggest that children initially do not understand of the goal of scientific experimentation (e.g., Carey, Evans, Honda, Jay, & Unger, 1989; Schauble, Klopfer, & Raghavan, 1991). Instead, Schauble and colleagues (1991) proposed that early experimentation is motivated by an '*engineering'* goal, in which children engage in exploratory interventions in order to "make things happen," rather than the '*science'* goal of learning the underlying causal structure of the world. If true, this early inability or unwillingness to conduct informative experiments poses a major complication for the claim that children's self-directed learning intuitively follows a scientific process.

The current study, therefore, seeks to examine whether young children select and make inferences from their own actions in a way that supports their causal learning. While it is clear from past research that even infants successfully infer causality from observation of the outcomes of interventions that are chosen and performed by others (e.g. Meltzoff, Waismeyer, & Gopnik, 2012), it remains an open question whether the same is true for actions that children take themselves. Schulz, Gopnik, and Glymour (2007), for example, provide evidence that young learners understand and utilize the conditional relationship between causal structure and intervention. Specifically, 3- to 6-years-olds accurately identified the causal structure of a system after observing the outcomes of interventions on it *and* accurately predicted outcomes of interventions on a system when the causal structure was known.

In contrast, more recent findings indicate that even older children (5 to 8 years) may struggle to apply this principle to their *own* actions. Two studies—McCormack, Bramley, Frosch, and Lagnado (2016) and Meng, Bramey, and Xu (2018) – have examined children's causal interventions and inferences during exploration of a 3-node system. While some of the actions children produced in both studies were informative, neither team found evidence for a strong preference for informative actions. For example, according to McCormack and colleagues (2016), only 7- and 8-year-olds *consistently* selected informative interventions significantly more often than chance, while 5- and 6-year-old's did *not* select informative interventions above chance. Similarly, Meng et al. (2018) found that 5- to 7-year-olds average selection of informative interventions was not distinguishable from chance levels.

In fact, both studies found evidence that children select interventions in accordance with a positive testing strategy (PTS)—that is, taking actions that are expected to produce an effect if their current hypothesis is correct (Coenen, Rehder, & Gureckis, 2015; Klayman & Ha, 1987). In McCormack et al. (2015), the most popular intervention was turning on the hypothesized *root node*, which activated all other nodes in the system, regardless of the true causal structure. Meng et al. (2018) also provide evidence for children's use of PTS: Although the model that best captured children's intervention choices in their task relied on a combination of expected information gain and PTS, this mix was heavily skewed towards PTS.

Importantly, however, evidence *for* PTS is not evidence *against* the 'engineering goal' account: While turning on the putative root node of a system positively tests the largest number of causal links with in it (see Coenen et al., 2015), this is *also* the action that 'makes the most things happen'. Indeed, within the scientific reasoning literature, PTS behaviors are often treated as evidence that young learners are focused exclusively on the tangible outcomes of their interventions (Tschirgi, 1980; Zimmerman, 2007; Zimmerman & Glaser, 2001). These previous findings, therefore, cannot rule out the possibility that young children select primarily interventions according to 'engineering,' rather than 'scientific' goals. Thus, our first aim is to look directly at children's intervention preferences. We ask whether young learners will privilege an informative option (one that has the potential to disambiguate between competing causal structures) over an uninformative one in a forced choice design. We then examine whether children maintain their preference when this uninformative alternative is guaranteed to produce a desirable effect.

Our second aim is to examine whether children can utilize the outcomes of their own actions in later causal inference. Despite being older than the children tested by Schulz et al. (2007), participants in Meng et al. (2018) failed to identify the correct causal structure more often than chance, and the 5- to 6-year-olds in McCormack et al. (2015) did so only for certain types of structures. It is unclear whether children's failure to identify the correct causal structure was due to

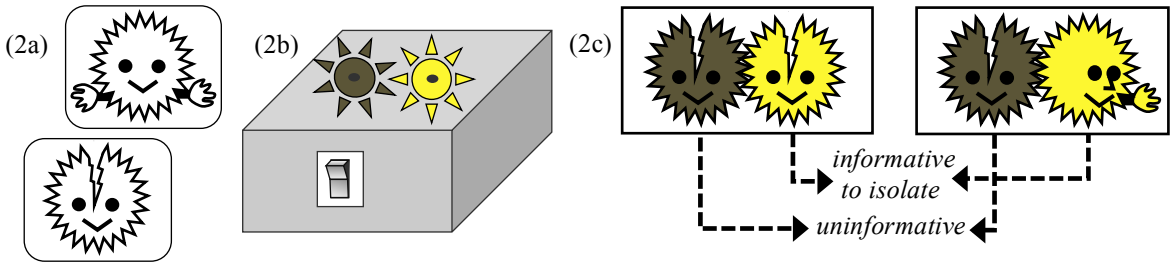


Figure 2: (a) Images used to illustrate 'working' and 'broken' gears. (b) Schematic of the gear toy. (c) Images used to illustrate the multiple causes (left) and single cause (right) structures with the informative option indicated.

their inability to make inferences from self-generated evidence, or due to the challenges associated with these more complex causal structures. In fact, Frosch and colleagues (2012) find that children struggle to make correct inferences about a similar 3-node causal system *even* when an experimenter generated the necessary evidence for them. We therefore designed the current task as a modified version of Schulz et al.'s (2007) paradigm. This is a context in which we know that young learners are able to reason about the conditional relationship between intervention and causal structure.

The current study aims to clarify whether young children preferentially select and successfully learn from their own actions in a way that is sensitive to the informative value of causal intervention. Two experiments examined how 4- to 6-year-olds responded to a forced choice between an informative and uninformative intervention in a causal learning task. Experiment 1 asks whether children will preferentially choose to take the informative intervention when selecting actions on a novel causal system. Then, in Experiment 2, the uninformative intervention is also guaranteed to produce a desirable effect. Choice behavior on this task will therefore distinguish whether children's early interventions are primarily motivated by a '*science'* or '*engineering'* goal. In addition to looking at which interventions young learners choose (and why), these experiments will also consider whether children are able to draw accurate inferences about a simple causal system from evidence they generate themselves.

Experiment 1

To investigate whether children preferentially choose interventions that support their causal learning, we used a task modeled on Schulz et al. (2007). Children were introduced to a gear toy featuring two interlocking gears and a switch. They learned that individual gears may be "working" (they spin when the toy is turned on) or "broken" (they are inert and prevent any interlocking gears from spinning). At test, children observed a pair of gears that failed to spin when the toy was turned on. They were told that this event could have resulted from two possible causal structures.[1] Either both gears are broken (a 'multiple causes' structure), or one gear is broken, preventing the other from spinning (a 'single cause' structure) (see Figure 2). As in the previous houseplant example (Figure 1), it is impossible to determine which of these represents the true causal structure from observation alone. Instead, a specific informative action must be performed: removing the gear that is broken in both structures and observing the behavior of the remaining gear *in isolation*. In contrast, removing the gear that varies between the two structures and observing the remaining (broken) gear would provide no information about the underlying causal structure. Children were given a choice between isolating and observing *only one* of the two gears prior to their inference. If young learners indeed recognize and privilege actions that are most informative for causal learning, then they should prefer to observe the gear that will disambiguate between the two structures. Afterwards, children were given the opportunity to observe the outcome of their chosen action, and were asked to judge which of the two structures was correct. If children are able to infer causal structure from their own actions, those who select the informative action should make the accurate inference.

Methods

Participants Forty-eight children (M = 64.19 months, SD = 9.46 months, range = 46-82 months) participated in Experiment 1. Children were recruited and tested individually at a local science museum in a primarily urban area. Seventeen additional children were run, but excluded due to experimental error (n = 11) or failing to complete the testing session (n = 6).

Stimuli The task used a custom-built electronic gear-toy, colored plastic gears, and picture cards with colored illustrations representing the gears and causal structures.

The toy, previously used in Schulz et al. (2007), consisted of a 12"x12" cube with two metal pegs on top. Each peg was designed to hold one 3" diameter gear, such that two gears would interlock when positioned on top of the toy. Sensors inside the cube detected the presence of a gear on the pegs, causing them to spin when a switch attached to the front of the toy was flipped to the 'on' position. A hidden control on the back of the toy allowed the experimenter to

[1] These structures were also based on Schulz et al (2007) and were originally referred to as 'common cause' and 'causal chain.' However, in the current experiment, it is more appropriate to refer to them as 'multiple cause' and 'single cause' structures, respectively.

surreptitiously control the supply of power (which determined whether or not the switch caused the gears to spin).

A total of six uniquely colored gears (blue, yellow, pink, green, red, orange) were used: four during the training trials and two during the test trial. Gear colors used for each part of the procedure were counterbalanced across participants. Note that in our description of the procedure, we refer to the gears using letters (A-F) in place of the color names that were actually used to identify each gear during the experiment. The picture cards (see Figure 2) each depicted a cartoon illustration of either a single gear (Figure 2a) or a gear pair (Figure 2c). These were used to illustrate the possible causal status (working or broken) and causal structures (single or multiple causes) during the task. The illustrated gears were color-matched to the physical gears used on the toy.

Procedure Each testing session began with the toy on the table in its powered state, with the switch in the 'off' position, and two gears (A and B) in place on the pegs. The experimenter introduced the toy, indicating the switch on the front, and explained that it turned the toy on and off, allowing the child to try both actions. When the child turned the toy on, A and B would spin simultaneously, and when the child turned the toy off, both stopped spinning simultaneously. The experimenter then removed and replaced each gear in turn, explaining that, when turned off, gears can be taken on and off the toy.

The experimenter then put A and B away, saying, "You're going to get to see all the gears. But some of the gears are broken. When a gear is broken, it doesn't spin even when the toy is on, and it gets in the way of other gears spinning too." Children were then shown an example working gear (A) and a broken gear (C) in turn. The experimenter placed the gear on the right peg of the toy and the child observed it either spinning (A) or not spinning (C) when the toy was turned on. Each gear was paired with a matching picture card showing its casual status. Using the pictures, the experimenter explained, "Gears that aren't broken can use their arms to spin themselves," and, "Gears that are broken don't have any arms, they cannot spin, and keep other gears from spinning too." The experimenter then held up A and C in turn and asked the child to tell them, first, whether the gear was broken or working, and second, whether it would spin on the toy on its own. Children received feedback and, if necessary, correction on each response. As part of the feedback for the second question, the experimenter placed the gear on the left peg of the toy and flipped the switch. Thus, children observed that broken and working gears operate consistently regardless of which peg of the toy they are on.

Each child then received training on the two causal structures, presented as different combinations of gears: a multiple cause (C and D) and a single cause (D and B) structure. The order in which the two structures were presented was counterbalanced, as was whether the broken gear (D) in the single cause structure was on the left or right peg of the toy. For each structure, the experimenter placed both gears on the toy and turned it on. The toy was always depowered, and the gears always remained inert. The experimenter said, "The gears aren't spinning. Something is wrong." She then brought out a picture card depicting one of the possible causal structures and described it to the child. For example, for the single cause structure, she said, "The picture shows us that just one of the gears is broken. The D gear is broken and doesn't spin on the toy, and the B gear is not broken so it can spin on the toy. But when they're together, the D gear gets in the way of the B gear, and nothing moves." Each gear was placed on the toy individually, and children were asked to predict (with feedback and observation) whether it would spin when the toy was turned on. This procedure was then repeated for the other structure.

During the test trial, the picture cards used during the training were left visible, one on either side of the toy. Gears E and F were placed on the toy and did not spin when the toy was turned on. This time, however, the experimenter said, "I don't know what's wrong here. I don't know why these gears aren't spinning. Will you help me figure it out?" The experimenter then produced two picture cards, identical to those seen during training, except that the depicted gears matched the colors of E and F. These cards were placed adjacent to the matching card from the training and each was described in the same terms. Children were told that they had to figure out which of the two pictures correctly showed why E and F weren't spinning together. Children were also told that they would get a 'clue' to help them: they could choose to see how *one* of the two gears (*either* E or F) would behave when the other gear was removed and the toy was turned on.[2]

After indicating their choice to the experimenter, children were allowed to remove the unselected gear, turn the toy on, and observe the outcome. If the informative gear was selected, the outcome (spin or inert) was counterbalanced, such that half of the children who selected the informative gear would observe evidence for the single cause structure, and the other half would observe evidence for the multiple causes structure. Regardless of choice or outcome, the experimenter would point to the gear when the toy was turned on and say, "Look!" before holding up the two picture cards depicting the possible structures, and asking children to pick the one that showed how the gears actually operated.

Results and Discussion

Children's responses to all questions were recorded during the experimental session and videotaped. We recorded

[2] As an attention and comprehension check, half of children (n = 24) were prompted to report the possible states of each gear before making their choice. This had no effect on either the number of informative interventions ($t(46) = -0.62$, $p = 0.538$ [*ns*]) or number of correct causal inferences ($t(32) = 1.37$, $p = 0.18$ [*ns*]), so the two scripts were combined.

whether each child chose to observe the informative or uninformative gear, as well as their final judgment about the true causal structure of the gears. For the subset of children who selected the informative gear, judgments were further coded for whether or not they were consistent with the outcome observed.

A significant majority (70.83%) chose the informative intervention, isolating and observing the gear that could disambiguate between the possible causal structures, (p = 0.005, two-tailed binomial). Of the 39 children who observed this disambiguating evidence, *all but two* made the correct causal inference (94.12%, $p < 0.0001$, two-tailed binomial). Together, these results suggest that young learners are not only sensitive to the informative potential of their own causal interventions, but they are also able to use the outcomes of those interventions to accurately infer the causal structure of events in the world.

Experiment 2

The results reported above provide evidence that young children preferentially select and learn from their own informative interventions in the course of causal learning. This is consistent with previous research on children's spontaneous exploration, while also extending this work to show that this preference for informative actions supports later inference. However, children's choice behavior on this task is also amenable to the opposite interpretation. As discussed above, the scientific reasoning literature often characterizes early experimenters as '*engineers*' (rather than '*scientists*') who incorrectly focus on generating effects (rather than information).

The informative gear in Experiment 1 was also the gear that had the potential to *spin* when isolated by intervention. It is possible, therefore, that children did not select the informative action because it would provide disambiguating evidence, but because it was more likely to produce this entertaining and desirable effect. If so, preference for informative action in Experiment 1 would actually be evidence *for* the claim that young children's interventions are motivated by producing effects, rather than learning about the world.

We conducted a second experiment to test this alternative. In Experiment 2, we changed the operation of the gears to include *generative* causes (i.e., working gears cause broken gears to spin), rather than inhibitory causes (i.e., broken gears prevent working gears from spinning): see Figure 3. At test, children observed a pair of spinning (rather than inert) gears that could be explained by appeal to either multiple (both gears spin) or a single cause (only one gear spins, causing the other to spin). Again, participants were given a forced choice between two interventions to determine the true causal structure.

Critically, however, this presents a choice between an uninformative action (isolating the gear that works under both structures), that is *guaranteed* to produce a desirable effect, and an informative action, (isolating the gear that works under one structure and is broken under the other), that has equivalent odds of producing or failing to produce the effect. This means that children must *forgo* the opportunity to produce a desirable effect in order to acquire information about how the causal system works.

If, as suggested by past work on exploratory play, children have an intuitive preference for informative actions, then we should continue to see a preference to isolate and observe the disambiguating gear. If, on the other hand, children show the opposite preference, choosing to select the uninformative gear, then this would suggest they are motivated by an '*engineering* goal'.

Methods

Participants Twenty-four children (M = 65.4 months, SD = 9.59 months, range = 46-82 months) were included in Experiment 2. Recruitment procedures and demographics were identical to Experiment 1. Four additional children were tested, but excluded due to experimental error (n = 1) or for failing to complete the testing session (n = 3).

Stimuli Materials were identical to those used in Experiment 1. However, new picture cards were created to depict the revised causal structures used in Experiment 2.

Procedure Procedures were similar to those used in Experiment 1. The script and outcomes of actions were modified in accordance with the revised definitions of 'broken' and 'working' gears. These changes are described below:

Children were initially told, "Some of the gears are broken. When a gear is broken, it can't spin on its own. It needs a gear that's not broken to make it spin." When shown the example gears and pictures (Figure 3), working gears were described as able to "use their little arms to spin themselves *and* to make other gears spin too!" Broken gears were described as unable to spin by themselves. Instead, broken gears "need a gear that's not broken on the toy with them to make them spin."

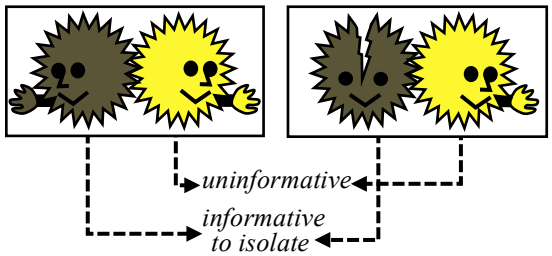


Figure 3: Illustration of the possible causal structures in Experiment 2.

In addition, the gear pairs were presented as operating according to one of two structures: Either both the gears (E and F) are working and can each spin on their own, or just one gear (E) is working, and "uses its little arms" to spin F, causing both to move. As in Experiment 1, whether the broken gear in the causal chain was the right or the left gear of the pair was counterbalanced across participants.

Results and Discussion

There were no age differences between the groups of children tested in Experiments 1 and 2, $t(70) = 0.03$, $p = 0.976$ (*ns*).

Children again selected the informative intervention significantly more often than expected by chance (79.17%, $p = 0.006$, two-tailed binomial). In fact, children's tendency to make this choice was not significantly different from their choice behavior in Experiment 1, $t(70) = 0.77$ $p = 0.442$ (*ns*). In other words, children continued to privilege the informative action *even* when it was pit against an opportunity to produce a desirable outcome.

Performance on the final inference question also did not differ from Experiment 1. Of the 19 children who selected the informative gear, *all but one* of them used this information to infer the causal structure that was consistent with the observed outcomes of their interventions (94.74%, $p < 0.0001$, two-tailed binomial). These results provide evidence against the alternative, '*engineering* goal' explanation for children's success in Experiment 1.

General Discussion

The current research sought to address two outstanding questions about children's intuitive experimentation: (1) Do children successfully identify and select informative interventions during exploration?, and (2) If so, can they draw appropriate causal inferences based on the outcomes they produce? These questions are critical, both for understanding the processes by which self-directed exploration contributes to early learning, and to address the disconnect between the claim that young learners are 'intuitive scientists,' and the claim that children are unsuccessful scientific experimenters.

First, our results demonstrate that 4- to 6-year-olds not only take informative interventions (Experiment 1), but that these actions are not driven by their potential to produce desirable outcomes (Experiment 2). These findings provide strong evidence against previous suggestions that children are initially concerned only with the practical (and not the informative) outcomes of their interventions. In particular, the '*science* vs. *engineering*' account, employed by Schauble and others (e.g., Schauble et al., 1991; Siler & Klahr, 2012) to explain children's choices in scientific reasoning tasks implies that the informative option should be less appealing than the uninformative, but productive one. The fact that the majority of children continued to select the informative action in Experiment 2 indicates instead that their choice of intervention was based on its potential to produce information and not positive outcomes. The apparent tendency to privilege producing effects seen in previous work may therefore be unrelated to children's understanding of the goals of experimentation, and an inaccurate reflection of early ability to identify and select interventions that improve their causal knowledge.

Second, these young children readily and accurately used the outcomes *of their own actions* when making judgments about the causal structure of a novel system. This goes beyond prior work showing that children make appropriate inferences after observing the outcomes of experimenter-generated interventions (Schulz et al., 2007), and contrasts with findings suggesting children may be unable to draw causal inferences from their own interventions (McCormack et al., 2016; Meng et al., 2018). In addition, while research on exploratory learning (e.g., Cook et al., 2011; Schulz & Bonawitz, 2007) has previously shown a preference for informative actions in young children, the bulk of this work has not required children to make subsequent causal inferences from the outcomes of those actions, leaving it uncertain whether and how children utilize self-directed exploration to support their learning.

Ongoing work aims to expand upon the current findings to investigate whether children are able to use the evidence generated by their own informative interventions to draw more sophisticated inferences. Specifically, we present children with cases in which the informative gear is paired with a novel gear after the intervention outcome is observed. Depending on the causal status (working or broken) of the informative gear, we can assess whether children will be able to *use* this information to update their existing causal representations, make predictions, and even draw inferences about the causal status of unknown gears.

This study also goes beyond past research on children's causal interventions (Meng et al., 2018; McCormack, et al., 2015) by directly examining intervention preference, and determining whether it is primarily driven by an action's informative potential or its tangible outcome. In contrast with previous work, the current results provide direct evidence *against* the claim that children select interventions in order to produce effects. Although our findings cannot explain children's previously reported tendency to engage in PTS, we show that this behavior is *not* due to their failure to appreciate the information-seeking goal of intervention and experimentation.

To summarize, the current results demonstrate that young children both preferentially select informative interventions, and make accurate inferences from the outcomes of those actions. These experiments fill a critical gap in the well-worn proposal that early causal learning intuitively follows a process that is analogous to belief revision in science. In sum, our findings suggest that young learners' causal interventions and inferences are sensitive to the principles of informative experimentation long before they are able execute and articulate those strategies in explicit scientific reasoning tasks.

References

Bonawitz, E. B., van Schijndel, T. J. P., Friel, D., & Schulz, L. E. (2012). Children balance theories and evidence in exploration, explanation, and learning. *Cognitive Psychology*.

Carey, S. (1985). *Conceptual change in childhood*. *The MIT series in learning development and conceptual change*. Cambridge, MA, US: MIT Press.

Carey, S., Evans, R., Honda, M., Jay, E., & Unger, C.

(1989). 'An experiment is when you try it and see if it works': A study of grade 7 students' understanding of the construction of scientific knowledge. *International Journal of Science Education*, *11*(5), 514–529.

Coenen, A., Rehder, B., & Gureckis, T. M. (2015). Strategies to intervene on causal systems are adaptively selected. *Cognitive Psychology*, 102-133.

Cook, C., Goodman, N. D., & Schulz, L. E. (2011). Where science starts: spontaneous experiments in preschoolers' exploratory play. *Cognition*, *120*(3), 341-349.

Frosch, C. A., McCormack, T., Lagnado, D. A., & Burns, P. (2012). Are Causal Structure and Intervention Judgments Inextricably Linked? A Developmental Study. *Cognitive Science*, *36*(2), 261–285.

Gopnik, A., & Wellman, H. M. (2012). Reconstructing constructivism: Causal models, Bayesian learning mechanisms, and the theory theory. *Psych Bulletin*.

Gweon, H., & Schulz, L. E. (2008). Stretching to learn: Ambiguous evidence and variability in preschoolers exploratory play. *Proceedings of the 30th Annual Meeting of the Cognitive Science Society*, 570–574.

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence : an essay on the construction of formal operational structures*.

Keil, F. C. (1989). *Concepts, kinds, and cognitive development*. Cambridge, MA, US: MIT Press.

Klahr, D., Fay, A. L., & Dunbar, K. (1993). Heuristics for Scientific Experimentation: A Developmental Study. *Cognitive Psychology*, *25*(1), 111–146.

Klayman, J., & Ha, Y.-W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, *94*(2), 211.

McCormack, T., Bramley, N., Frosch, C., Patrick, F., & Lagnado, D. (2016). Children's use of interventions to learn causal structure. *Journal of Experimental Child Psychology*.

Meltzoff, A. N., Waismeyer, A., & Gopnik, A. (2012). Learning about causes from people: Observational causal learning in 24-month-old infants. *Developmental Psychology*, *48*(5), 1215–1228.

Meng, Y., Bramley, N., & Xu, F. (2018). Children's causal interventions combine discrimination and confirmation. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society*.

Pearl, J. (2000). *Causality : models, reasoning, and inference*. Cambridge University Press.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928.

Schauble, L. (1990). Belief revision in children: The role of prior knowledge and strategies for generating evidence. *Journal of Experimental Child Psychology*, *49*(1), 31–57.

Schauble, L., Klopfer, L. E., & Raghavan, K. (1991). Students' transition from an engineering model to a science model of experimentation. *Journal of Research in Science Teaching*, *28*(9), 859–882.

Schulz, L. E., & Bonawitz, E. B. (2007). Serious Fun: Preschoolers Engage in More Exploratory Play When Evidence Is Confounded. *Developmental Psychology*.

Schulz, L. E., Gopnik, A., & Glymour, C. (2007). Preschool children learn about causal structure from conditional interventions. *Developmental Science*, *10*(3), 322-332.

Siler, S. A., & Klahr, D. (2012). Detecting, Classifying, and Remediating: Children's Explicit and Implicit Misconceptions about Experimental Design. In *Psychology of Science: Implicit and Explicit Processes*.

Tschirgi, J. E. (1980). Sensible Reasoning: A Hypothesis about Hypotheses. *Child Development*, *51*(1), 1–10.

Valanides, N., Papageorgiou, M., & Angeli, C. (2014). Scientific Investigations of Elementary School Children. *Journal of Science Education and Technology*, *23*(1), 26–36.

Wellman, H. M., & Gelman, S. A. (1992). Cognitive Development: Foundational Theories of Core Domains. *Annual Review of Psychology*, *43*(1).

Xu, F., & Garcia, V. (2008). Intuitive statistics by 8-month-old infants. *Proceedings of the National Academy of Sciences*, *105*(13), 5012–5015.

Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. *Developmental Review*, *27*(2), 172–223.

Zimmerman, C., & Glaser, R. (2001). *Testing Positive Versus Negative Claims: A Preliminary Investigation of the Role of Cover Story on the Assessment of Experimental Design Skills. CSE Technical Report*.

Zimmerman, C., & Klahr, D. (2018). Development of Scientific Thinking. In J. Wixted (Ed.), *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience* (4th ed., pp. 1–25).

Frequency or Predictability?
The Effect of Entropy on Statistical Learning

Anonymous CogSci submission

Abstract

Do language learners benefit from exposure to a more predictable input (with lower entropy)? Frequency is known to facilitate learning (more frequent words acquired earlier). However, frequency is only one measure of the distributional structure of the linguistic input. Here, we show that entropy also impacts language learning: adults show better word segmentation in an artificial language when the sequence has lower entropy (created by making one word more frequent). Segmentation improved both for the language as a whole, and for the less frequent words: the infrequent words in the low entropy condition were learned better than the words in the high entropy condition, despite appearing half the number of times. This is the first demonstration, to our knowledge, of the facilitative effect of entropy reduction on language learning. We discuss implications for artificial language learning experiments (which often use uniform distributions) and for models of language learning more generally.

Keywords: Statistical learning; Word segmentation; Language Learning; Information.

Introduction

Frequency effects are prevalent across many aspects of language learning and processing. More frequent sounds, words and constructions are acquired earlier (Diessel, 2007; Goodman, Dale, & Li, 2008), and more frequent words are easier to recognize and produce (Jescheniak & Levelt, 1994). These effects are not restricted to single words: more frequent multiword phrases are also processed faster by adults (Arnon & Priva, 2013; Arnon & Snider, 2010), and produced more accurately by children (Bannard & Matthews, 2008). Frequency also impacts the structure of the lexicon: more frequent words tend to be phonologically shorter (Zipf, 1936).

While frequency affects many domains in language, it captures only one aspect of the distributional structure of the linguistic environment. Frequency alone does not tell us about the co-occurrence patterns of words; the contexts in which words tend to appear; or how predictable the input is overall. In order to quantify such aspects of the linguistic input, other measures are required. Here, we focus on one such measure, Shannon's Entropy (Shannon, 1948). Shannon's entropy quantifies how unpredictable a variable is, with higher entropy assigned to less predictable variables. For instance, a toss of a fair coin has higher entropy than a toss of an unfair coin. Entropy tells us something about the entire distribution of words, beyond the properties of each individual word.

In the past decade, there has been growing interest in applying more complex measures like entropy to the study of language, and growing evidence for their impact on language structure and use. For example, information content is a better predictor of word length than frequency, with less predictable words tending to have longer lexical forms (Piantadosi, Tily, & Gibson, 2011). Similar effects are found in online processing where reading times are affected by entropy (Linzen & Jaeger, 2015), and speakers' production of less predictable words is slower (Cohen Priva, 2017) and less contracted (Frank & Jaeger, 2008). Children are also sensitive to such measures: two-year-olds show better repetition of unfamiliar four-words sequences when the final word "slot" has higher entropy (Matthews & Bannard, 2010), and earlier acquisition of words that have greater contextual diversity (appearing with more unique words) (Hills, Maouene, Riordan, & Smith, 2010).

However, very little work to date has looked at the impact of entropy on learning novel linguistic information: will entropy reduction lead to better learning? Here, we examine this question by looking at statistical learning, and in particular, at the classic word segmentation task of Saffran et al., (1996). Statistical learning (SL) has been studied extensively over the past 20 years, demonstrating human's ability to use distributional information to learn about various aspects of language structure (Romberg & Saffran, 2010). One of the first demonstrations of SL was in the domain of word segmentation, where infants were shown to use the lower transitional probabilities between words as a cue to word boundaries (Saffran, Aslin, & Newport, 1996). Research since has shown that humans can also make use of such distributional information to learn more complex relations such as non-adjacent dependencies (Gomez, 2002) or multimodal associations (Cunillera, Laine, Càmara, & Rodríguez-Fornells, 2010; Lavi-Rotbain & Arnon, 2017).

Interestingly, even though SL of word segmentation has been studied extensively, almost all such studies present learners with a uniform distribution where all elements appear an equal number of times (e.g., each of the words in the Saffran segmentation task appear equally often). However, using a uniform distribution has two inherent drawbacks. First, such a uniform distribution deviates from

that of natural language where words follow a highly skewed Zipfian distribution (Zipf, 1936). A Zipfian distribution has a narrowed peak for the small number of words that are the most frequent, and a very long tail for the rest of the words that have low frequencies. Words show a Zipfian distribution across many languages, in both adult-to-adult speech (Zipf, 1936; Piantadosi, 2014) and child directed speech (Lavi-Rotbain & Arnon, under review). That is, unlike word segmentation studies, words in natural language do not have a uniform distribution.

Second, uniform distributions have low predictability. Elements that show a uniform distribution are harder to predict, since they are equally likely to appear: no guess is better than the other. Sewed distributions, such as a Zipfian distribution, are more predictable: when only a small number of words are highly frequent, they make a better guess than the rest. That is, the uniform distributions used in word segmentation experiments differ from those of natural language in ways that may impede learning. The difference in predictability between uniform and Zipfian distributions can be captured using entropy: the uniform distribution is the least predictable and therefore has maximal entropy, while a Zipfian distribution has lower entropy.

Here we ask if entropy reduction can lead to better learning of word segmentation in the classic Saffran segmentation task. Such a finding would illustrate the sensitivity of learners to more complex distributional measures, and their potential impact on learning outcomes. Only one study, to our knowledge, used a non-uniform distribution in a word segmentation task. In that study, adults learned from either a "Zipfian" distribution or a uniform one. No difference in segmentation scores was found between the two, despite the reduced entropy of the latter (Kurumada, Meylan, & Frank, 2013). These findings seem to go against the idea that entropy reduction is facilitative. However, another possibility is that no facilitation was found because entropy was not reduced enough. In the "Zipfian" condition in Kurumada et al., the word's frequency was inversely proportional to its rank. However, the sharp contrast found in natural language between the narrowed peak and the long tail was not present in this condition. This sharp contrast (the existence of few very frequent words and many low frequency words) shifts the distribution further from being uniform, and reduces entropy to lower levels. We suggest that this contrast between frequencies is beneficial for learning.

Here, we expand on the findings of Kurumada et al. (2013) in several ways to test the prediction that entropy reduction will facilitate learning. First, we compare performance in several levels of entropy: high, medium and low. Segmentation might be facilitated in the low entropy level, which was not tested in Kurumada (2013), and is more similar to natural language. Second, we compare learning of items with the same frequency, across several levels of entropy. We predict that reduced entropy can facilitate learning of low frequency items beyond what is expected from their frequency: words with lower frequency will be learned better when they appear in a more predictable environment (one with lower entropy), compared to in a uniform distribution.

The current study

In the current study we ask if entropy reduction can be beneficial for words segmentation (1) in general, and (2) of infrequent words. We examine the first prediction by looking at adults' segmentation scores across several levels of entropy: high, medium and low, with the same exposure durations. Entropy was reduced by making one word more frequent than the rest. If language learners are mostly sensitive to frequency of novel words, performance on the segmentation test should be affected by word frequency rather than entropy level. However, if learners are sensitive to more than mere frequency, e.g. to the predictability of the input, than segmentation score in the low entropy condition should be better than in the high entropy condition.

We examine the second prediction by comparing segmentation of items with the same low frequency, across different levels of entropy. We expect that low entropy will boost learning of low frequency items, such that low frequency words will be learned better when they appear in a more predictable sequence (with lower entropy), compared to when they appear in a uniform distribution (with high entropy). Previous work has shown that previously learned words can serve as anchors for word boundaries and facilitate segmentation (Cunillera, Càmara, Laine, & Rodríguez-Fornells, 2010). We hypothesize that a similar effect can happen when making one word more frequent. This word is now highly predictable, can be learned early on, and serve as an anchor for learning the segmentation of the infrequent words.

Method

Participants

142 undergraduate students at the Hebrew University of Jerusalem participated in the study (108 females, 34 males, mean age 24;0). Participants were randomly assigned to one of the four experimental conditions. All of the participants were native Hebrew speakers without learning disabilities or attention deficits. Participants received 10 NIS or course credit in return for their participation.

Materials

Auditory stimuli

The task was modelled on the audio-only condition from Lavi-Rotbain & Arnon (2017). Participants were exposed to a familiarization stream corresponding to the condition they were assigned to. All streams were composed of the same four unique tri-syllabic synthesized words: "dukame", "nalubi", "kibeto", and "genodi". The syllables making up the words were taken from Glicksohn & Cohen (2013). They were created using the PRAAT synthesizer (Boersma

& van Heuven, 2001) and were matched on pitch (~76 Hz), volume (~60 dB), and duration (250–350 ms).

The four words were created by concatenating the syllables using MATLAB to ensure that there were no co-articulation cues to word boundary. The words were matched for length (mean word length=860ms, range=845-888ms). The words were then concatenated together using MATLAB in a semi-randomized order to create the auditory familiarization streams. Importantly, there were no breaks between words and no prosodic or co-articulation cues in the stream to indicate word boundaries. The only cue for word boundaries was transitional probabilities (TP's): TP's between words were lower compared to TP's within words.

Experimental conditions

We created auditory sequences with three levels of entropy: high, medium and low, but with the same number of tokens (128) and length (1:50 minutes), in order to see if reduced entropy can facilitate segmentation. In the high entropy level, words followed a uniform distribution with each word appearing 32 times in a semi-randomized order (no word appeared twice in a row). TP's within a word were 1, and TPs between words were 0.333. In the medium entropy level, words appeared with a skewed distribution: one word appeared 55% of the time (71 appearances) while each of the other three words appeared 15% of the time (19 appearances for each word). In the low entropy level, words appeared with an even more skewed distribution: one word appeared 80% of the time (101 appearances) while each of the other three words appeared only 7% of the time (9 appearances for each word). In both the low and medium entropy conditions, the identity of the frequent word was counterbalanced across subjects. In addition, in both conditions the TP's within a word were 1, but the TP's between words varied depending on the next word (since the frequent word in these conditions was more likely to occur). These conditions were used to examine the effect of entropy on the general segmentation score.

In order to look at the segmentation of the low frequency items, we added a uniform condition with high entropy but with shorter length (uniform-short). In this condition, each word appeared 19 times (76 tokens, lasting 1:05 minutes). The frequency of each word in this condition was matched to that of the infrequent words from the medium entropy condition. By comparing the two we can examine the impact of entropy on words with the same low frequency. See Table 1 for full details of the experimental conditions.

Segmentation test

16 two alternative forced choice trials appeared in a random order, with the constraint that the same word/foil did not appear in two consecutive trials. Participants heard two words and were asked to decide which belonged to the language they heard. We used non-words as foils ("dunobi", "nabedi", "kilume", and "gekato", average length: 860ms; range 854-868ms), created by taking three syllables from three different words, while keeping their original position. Each of the four words appeared once with each of the four foils to create 16 trials. The order of words and foils was counter-balanced so that in half the trials, the real word appeared first and in the other half, the foil appeared first.

Procedure

Participants completed the experiment on a computer while seated in a quiet room. They were told that they are going to listen to an alien language and will then be asked about it. A check-board image was displayed while they listened to the familiarization stream. After the exposure phase, participants completed the segmentation test.

Results

Participants were divided as follows between the four conditions: uniform, N=31; uniform-short, N=30; medium entropy, N=41; low entropy, N=40. In the medium and low conditions, each of the four words was the frequent one for ten subjects. A one way ANOVA (on each entropy rate separately) revealed that segmentation did not differ due to which word was the frequent one (for the medium entropy condition: $F(3)=0.72$, $p=0.55$; for the low entropy condition: $F(3)=1.7$, $p=0.18$). Consequently, in all subsequent analyses we collapsed the data across the different frequent words, for each of these conditions. Participants showed learning (were above chance) in all four conditions (low entropy condition: $t(39)=12.57$, $p<.001$; medium entropy condition: $t(40)=7.0$, $p<.001$; uniform condition: $t(30)=7.0$, $p<.001$; uniform-short condition: $t(29)=5.8$, $p<.001$) (see Fig. 1).

We used mixed-effect linear regression model to examine the effect of condition on performance. Following Barr et al. 2013, the models had the maximal random effect structure justified by the data that would converge. Our dependent binominal variable was success on a single trial of the segmentation test. We had experimental condition (dummy coded, meaning that each condition is compared to the

Table 1: Different experimental conditions

	Uniform-short	**Uniform**	**Medium entropy**	**Low entropy**
Exposure length [minutes]	1:05	1:50	1:50	1:50
Number of tokens	76	128	128	128
Tokens per word	19	32	Frequent: 71 Infrequent: 19	Frequent: 101 Infrequent: 9
Entropy [bits]	2	2	1.7	1.1

Table 2: Mixed-effect regression model for all four conditions. Variables in bold were significant. Significance obtained using the lmerTest function in R.

	Estimate	**Std. Error**	**z value**	**p-value**
(Intercept)	0.27331	0.17793	1.536	>.1
uniform-short condition	0.17777	0.20571	0.864	>.1
Medium entropy condition	0.18484	0.18791	0.984	>.1
Low entropy condition	1.25277	0.21789	5.750	**<.001 *****
Log frequency (centered)	0.40138	0.09691	4.142	**<.001 *****
Gender (male)	-0.01982	0.16113	-0.123	>.1
Trial number (centered)	-0.03469	0.01061	-3.271	**<.01 ****
Order of appearance (word)	0.59277	0.09781	6.061	**<.001 *****

uniform condition) as a fixed effect, as well as: log frequency of the word (centered); gender; trial number (centered); order of appearance in the test (word-first trials vs. foil-first trials). The model had random intercepts for participants and for items (Table 2). To examine the overall effect of experimental condition and word's frequency, we used two model comparisons.

As predicted, experimental condition had a significant effect on performance (chi(3)=42.07, p<0.001). Participants showed better learning in the low entropy condition compared to the uniform condition (β=1.25, SE=0.22, p<0.001). However, performance in the medium entropy condition, and in the uniform-short condition, did not differ from the uniform condition (uniform-short: β=0.19, SE=0.2, p>0.1; medium entropy: β=0.19, SE=0.19, p>0.1).

In addition to the entropy effect, frequency also had a significant effect on segmentation (chi(1)=18.9, p<0.001). Participants showed higher accuracy for more frequent words (β=0.4, SE=0.09, p<0.001). Trial number significantly affected performance: better accuracy in the beginning of the test (β= -0.03, SE=0.01, p<0.01). Order of appearance in the test significantly affected performance: better accuracy on trials where the word appeared before the foil (β=0.59, SE=0.1, p<0.001), as has been found in previous studies (Lavi-Rotbain & Arnon, 2017; Raviv & Arnon, 2017). Since the order of presentation of words and foils was counter-balanced this could not reflect a preference for pressing 1 or 2. Gender did not affect performance (β= -0.02, SE=0.16, p>0.1).

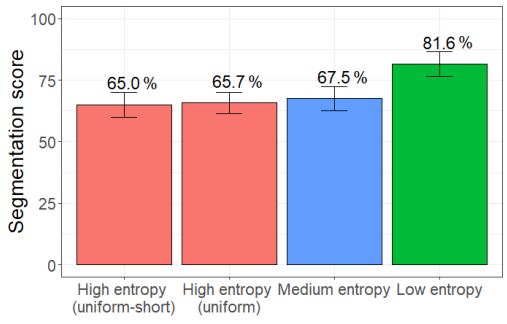


Fig. 1: Mean segmentation score by condition with 95% confidence intervals

In order to examine the effect of entropy on low frequency words, we compared accuracy in learning: (1) the words in the uniform-short condition; (2) infrequent words from the medium entropy condition; and (3) infrequent words from the low entropy conditions. The first two sets of words appeared 19 times during exposure, while the third set appeared only nine times. We used all trials (16 per subject) from the uniform-short condition (since they all had the same frequency). However, for the medium and low entropy conditions, we included only trials in which the correct answer was one of the infrequent words (denoted as 'infrequent trials'). In these conditions, there were 12 infrequent trials for each subject. Participants showed learning of infrequent items (above chance) in all conditions (low entropy condition: t(39)=9.59, p<.001; medium entropy condition: t(40)=5.3, p<.001).

We used a mixed-effect linear regression model to look at the effect of entropy level on learning infrequent words. Our dependent binominal variable was success on a single trial. We had experimental condition as a fixed effect (each condition was compared to the uniform-short condition) as well as: gender, trial number (centered); and order of appearance in the test. The model had random intercepts for participants and for items. To examine the overall effect of condition, we used model comparisons.

As predicted, experimental condition had a significant effect on learning infrequent words (chi(2)=16.9, p<0.001). Low frequency words were learned better in the low entropy condition (M=78.8%) than in the uniform-short condition (M=65%) (β=0.78, SE=0.22, p<0.001). This effect is opposite to what would be expected based on mere frequency: these words appeared only nine times in the low entropy condition as opposed to 19 times in the uniform-short condition. Performance on infrequent trials in the medium entropy condition (M=64.8%) did not differ from the uniform-short condition (β=0.0, SE=0.2, p>0.1). Trial number affected performance, with better accuracy in the beginning of the test (β= -0.03, SE=0.01, p<0.05). Order of appearance in the test also affected performance, with better accuracy on trials where the word appeared before the foil (β=0.53, SE=0.1, p<0.001). Gender did not affect performance (β=0.06, SE=0.2, p>0.1).

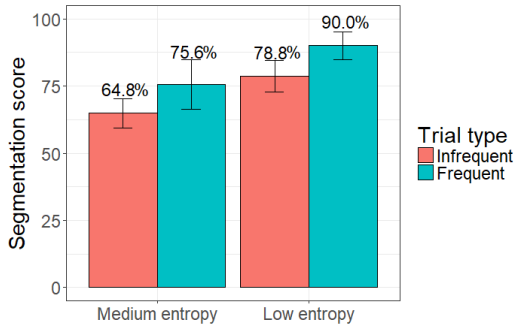


Fig. 2: Mean segmentation score by condition by trial type (frequent VS infrequent) with 95% confidence intervals

How can we reconcile the general effect of frequency with the finding that words that appeared only nine times were learned better than those appearing 19 times? Our data suggests that what matters is not absolute frequency, but relative frequency: within each condition, the more frequent words were learned better. This is best illustrated in Fig. 2 in which we plotted segmentation means by condition (medium and low entropy) and by trial type (infrequent versus frequent trials). Frequency affect performance within conditions: frequent words are learned better in both entropy levels. However, this does not hold across conditions. For example, infrequent trials from the low entropy condition are numerically better than frequent trials from the medium entropy condition, despite of the sharp difference in frequency in the opposite direction: only nine appearances compared to 71. That is, only the relative frequency within each condition affected performance.

One possible explanation for the entropy effect we found is that participants only learned the frequent word, and used it to rule out foils by elimination. If this is what they did, we should see a difference in segmentation scores across foils: foils that share a syllable with the frequent word should be easier to reject compared to foils that do not. For example, if the frequent word for a participant is 'nalubi', we should see better accuracy in rejecting 'nabedi' that shares the first syllable with 'nalubi', compared to rejecting 'gekato' that does not share a syllable with 'nalubi'. However, we see no such effects. A one-way ANOVA showed no difference between trials in the low entropy condition where the foil shared one syllable with the frequent word (M=79.2%) and trials where it didn't (M=77.5%) ($F(1)=1.2$, $p>0.1$). No difference was found even when we compared performance between foils that share the first syllable with the frequent word separately from these who share the second, the third or none at all ($F(3)=1.57$, $p>0.1$). That is, the boost for the infrequent words in the low entropy condition seems to reflect the better learning of those words.

Discussion

We set to ask if reduced entropy can improve segmentation in a classic auditory SL task (1) in general, and (2) of infrequent words. In addition, we wanted to see if the lack of facilitation in previous findings (Kurumada et al., 2013) was due to a not large enough decrease in entropy. To do so, we examined adults' word segmentation in an artificial language across three levels of entropy (high, medium and low). Entropy was reduced by making one word more frequent than the rest, so that it appeared 55% (medium entropy) or 80% (low entropy) of the time. As in the "Zipfian" condition in Kurumada (2013), reducing entropy to medium level did not facilitate segmentation. However, lower levels of entropy did facilitate learning compared to uniform conditions with the same length. This effect was not driven only by improved learning of the frequent words. The low frequency words also benefitted: they were learned better in the low entropy condition compared to medium and high levels, despite appearing half the number of times (nine vs. 19). Further analyses ruled out alternative explanations: the facilitation cannot be explained by ruling out foils that share syllables with the frequent word. In addition to the effect of entropy, our findings highlight the importance of relative, rather than absolute frequency on learning. Frequency effects were present only within conditions and not across conditions. For example, infrequent words from the low entropy condition, that appeared only nine times, were learned better than the infrequent words in the medium entropy condition (appearing 19 times). Moreover, they were learned numerically better (though this did not reach significance) than the frequent word in the medium entropy condition despite appearing much less (nine vs. 71 times).

This is the first evidence, to our knowledge, that humans are sensitive to complex measures such as entropy in the process of language learning, and that a more predictable distribution, as the one found in natural language, can be beneficial for learning compared to a uniform one. In addition, we provide novel evidence showing that low frequency items can 'overcome' their frequency when appearing with higher frequency items, in a more predictable distribution. These results have implications for artificial language experiments. The vast majority of artificial language experiments use a uniform distribution in which all items have equal frequency. It is already known that this uniform distribution is not ecological since the natural language we are exposed to shows a Zipfian distribution (Zipf, 1936; Piantadosi, 2014) even in speech directed to infants at their first stages (Lavi-Rotbain & Arnon, under review). Our results highlight an additional drawback of using uniform distributions in the lab: such distributions can impede performance compared to more skewed, low entropy distributions. That is, we may be significantly underestimating learners' abilities when using uniform distributions. This is of particular importance when such tasks are used to determine what learners can (or cannot) learn. We are currently investigating the impact of entropy on learning in children, and for other kinds of SL tasks.

Beyond artificial language experiments, these results have implications for our understanding of the factors that impact language learning. While frequency effects on language

learning have been studied extensively (Goodman et al., 2008; Jescheniak & Levelt, 1994), the effect of more complex measures remain understudied. Our results highlight the role of entropy in learning and open up new research directions on the impact of entropy on real-life language learning. What is the informative structure of child-directed speech? Does variance in entropy predict the age of acquisition of words? Can we see similar effects of the environment words appear in on natural language learning? We are currently engaged in a series of studies investigating these questions, which can further deepen our understanding of infants' first steps into language and the formation of their vocabulary.

Acknowledgments

We wish to thank Zohar Aizenbud for her help with preparing the study. The research was funded by the Israeli Science Foundation grant number 584/16 awarded to the second author.

References

Arnon, I., & Priva, U. C. (2013). More than Words : The Effect of Multi-word Frequency and Constituency on Phonetic Duration.

Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. Journal of Memory and Language, 62(1), 67–82.

Bannard, C., & Matthews, D. (2008). Stored Word Sequences in Language Learning. Psychological Science, 19(3), 241–248.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. Journal of Memory and Language, 68(3), 255–278.

Boersma, P., & van Heuven, V. (2001). Speak and unSpeak with Praat. Glot International, 5(9–10), 341–347.

Cohen Priva, U. (2017). Not so fast: Fast speech correlates with lower lexical and structural information. Cognition, 160, 27–34.

Cunillera, T., Càmara, E., Laine, M., & Rodríguez-Fornells, A. (2010). Words as anchors: Known words facilitate statistical learning. Experimental Psychology, 57(2), 134–141.

Cunillera, T., Laine, M., Càmara, E., & Rodríguez-Fornells, A. (2010). Bridging the gap between speech segmentation and word-to-world mappings: Evidence from an audiovisual statistical learning task. Journal of Memory and Language, 63(3), 295–305.

Diessel, H. (2007). Frequency effects in language acquisition, language use, and diachronic change. New Ideas in Psychology, 25(2), 104–123.

Frank, A. F., & Jaeger, T. F. (2008). Speaking Rationally : Uniform Information Density as an Optimal Strategy for Language Production. The 30th Annual Meeting of the Cognitive Science Society (CogSci08), 939--944.

Glicksohn, A., & Cohen, A. (2013). The role of cross-modal associations in statistical learning. Psychonomic Bulletin & Review, 20(6), 1161–9.

Gomez, R. L. (2002). Variability and Detection of Invariant Structure. Psychological Science, 13(5), 431–436.

Goodman, J. C., Dale, P. S., & Li, P. (2008). Does frequency count? Parental input and the acquisition of vocabulary. Journal of Child Language, 35(3), 515–531.

Hills, T. T., Maouene, J., Riordan, B., & Smith, L. B. (2010). The associative structure of language: Contextual diversity in early word learning. Journal of Memory and Language, 63(3), 259–273.

Jescheniak, J. D., & Levelt, W. J. M. (1994). Word Frequency Effects in Speech Production: Retrieval of Syntactic Information and of Phonological Form. Journal of Experimental Psychology: Learning, Memory, and Cognition, 20(4), 824–843.

Kurumada, C., Meylan, S. C., & Frank, M. C. (2013). Zipfian frequency distributions facilitate word segmentation in context. Cognition, 127(3), 439–453.

Lavi-Rotbain, O., & Arnon, I. (2017). Developmental Differences Between Children and Adults in the Use of Visual Cues for Segmentation. Cognitive Science, 42, 606-620.

Lavi-Rotbain, O. & Arnon, I. (under review). Zipf's Law in Child-Directed Speech.

Linzen, T., & Jaeger, T. F. (2015). Uncertainty and Expectation in Sentence Processing: Evidence From Subcategorization Distributions. Cognitive Science, 40(6), 1382-1411.

Matthews, D., & Bannard, C. (2010). Children's production of unfamiliar word sequences is predicted by positional variability and latent classes in a large sample of child-directed speech. Cognitive Science, 34(3), 465–488.

Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: A critical review and future directions. Psychonomic Bulletin & Review, 21(5), 1112–1130.

Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. Proceedings of the National Academy of Sciences of the United States of America, 108(9), 3526–3529.

Raviv, L., & Arnon, I. (2017). The developmental trajectory of children's auditory and visual statistical learning abilities: Modality-based differences in the effect of age. Developmental Science, (May), 1–13.

Romberg, A., & Saffran, J. R. (2010). Statistical learning and language acquisition. Wiley Interdisciplinary Reviews: Cognitive Science, 1(6), 906–914.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. Science (New York, N.Y.), 274(5294), 1926–1928.

Shannon, C. E. (1948). A mathematical theory of communication. The Bell System Technical Journal, 27, 379–423.

Zipf, G. (1936). The Psychobiology of Language. London: Routledge.

The Inductive Benefit of Being Far Out: How Spatial Location of Evidence Impacts Diversity-based Reasoning

Chris A. Lawson (lawson2@uwm.edu)
Noah Wolfe (wolfena@uwm.edu)
University of Wisconsin – Milwaukee
Department of Educational Psychology
Milwaukee, WI USA

Abstract

Inductive reasoning is constrained by several principles that govern how we choose to generalize evidence to new cases. Here we focus on diversity principle of induction, which describes the tendency to favor inductive arguments that include a diverse sample of evidence over those that include a homogenous sample of evidence. Several studies reveal that adherence to the diversity principle is influenced by a range of conceptual processes, such as an individuals' prior knowledge or expectations about the categories and properties represented in the evidence. In the two experiments reported here we examined a contextual factor of the available evidence – the spatial separation of evidence exemplars – that we expected would impact how people reason about diverse samples. We found that when the pictures (Experiment 1) or labels (Experiment 2) used to represent evidence exemplars were presented far apart (approximately 10 cm), participants showed a greater willingness to endorse arguments with diverse exemplars than those with homogenous sample, relative to when these exemplars were placed in close proximity (approximately 1 cm apart). We discuss these results as they relate to existing models of induction.

Keywords: Inductive reasoning; Generalization; Diversity principle; Situated cognition

Introduction

Inductive reasoning, the process by which we use specific facts to arrive at general conclusions, is critical to our cognitive lives. For example, learning that hawks have hollow bones serves as evidence to support the inductive inference that other birds are likely to have hollow bones. Given the powerful role of induction for a range of cognitive processes there has been considerable interest in determining the constraints that guide the inferences we make. For example, in their classic work, Osherson and colleagues (1990) outlined several inductive principles that systematically constrain how we use evidence to arrive at inductive decisions. The present study focused on one such principle – the diversity principle of induction. Consider the two arguments below in which the two statements above the lines represent evidence and the statement below the lines represents a conclusion:

Hawks have hollow bones
Penguins have hollow bones (1)
Larks have hollow bones

Hawks have hollow bones
Eagles have hollow bones (2)
Larks have hollow bones

When asked to judge which of these two represent stronger inductive arguments, participants tend to select those that include a diverse sample of exemplars (1) rather than those that include a homogenous sample of exemplars (2) (Heit, Hayes, & Feeney, 2005; Kim & Keil, 2003; Osherson, et al., 1990; also, Lopez, 1995).

Most explanations of diversity effects focus on the ways individuals represent the content (i.e., categories and to-be-generalized properties) of the available evidence. For example, Osherson et al. proposed the similarity-coverage model to account for diversity effects. This perspective posits that individuals first consider the overarching category about which the inductive judgment should be considered. In the two inductive arguments presented above the coverage category in *bird*. Participants then assess the extent to which the evidence in each set of arguments covers this overarching category. According to this model individuals rely on their calculation of the similarity between exemplars to assess the extent to which each sample covers the conclusion category. Greater dissimilarity between exemplars within the evidence sample reflects greater coverage of the category, and therefore facilitates diverse-based reasoning.

Diversity effects have also been explained as Bayesian inference. From this perspective individuals rely on their prior beliefs about categories and properties to test hypothesis about the scope of property projection (Heit, 1998; Lo, Sides, Rozelle, & Osherson, 2002). Our prior experience may lead us to believe that some categories (e.g., hawks and eagles) share many features in common and others categories (e.g., hawks and penguins) share fewer features. Thus, we are not surprised to learn about a new property that happens to be shared by two categories we have heretofore expected share many properties. In contrast, we are surprised to learn about a property that is shared by two categories that we believed had very little in common. This surprising sample of evidence, coupled with our expectation that samples of

evidence tend to be selected purposefully (Lawson & Kalish, 2009), makes the diverse sample a better argument to support a conclusion about a superordinate category.

There are notable cases in which individuals fail to adhere to the diversity principle. For example, several studies have shown that individuals with rich domain knowledge are less likely to consider taxonomic diversity in lieu of other evidence (Lopez, Atran, Coley, Medin, & Smith, 1997; Proffitt, Coley, & Medin, 2000). Moreover, when experts do engage in diversity-based reasoning they rely on a range of strategies that appeal to their knowledge about the domain, such as the types of properties that are transmitted across categories, or the relative size of the category that is represented in the samples of evidence (Proffitt et al., 2000). Thus, experts will depart from using taxonomic diversity as a basis for induction under conditions in which their rich domain knowledge suggests an alternative inductive strategy is optimal.

In related work Medin and colleagues (2003) showed that non-experts (college students) prefer to generalize from a sample of evidence that highlights a relevant relation between two evidence exemplars rather than a sample that includes taxonomically diverse exemplars. For example, participants judged an argument in which fleas and butterflies were attributed the same property as better support to conclude that the property is true of sparrows, than an argument in which fleas and dogs were the same property. This latter sample signals a relevant causal relationship that draws attention away the greater taxonomic diversity of the two exemplars, thereby leading individuals to favor the inductive argument with less diverse exemplars.

These exceptions are notable for two reasons. First, they highlight the role of prior knowledge about categories and properties when reasoning about the content of an inductive problem. Second, they bring to light an important methodological point: specific task modifications, such as the type of property or the relationship between categories presented in the evidence, impact how people reason about diverse samples. In support of this point, Feeney and Heit (2011) showed that the content of the to-be-generalized property serves as a prime to either encourage or discourage diversity-based responses. In their study participants exhibited diversity effects when they were primed with a general property that can be construed as common across a wide range of category members (e.g., are warm-blooded), but did not show these effects when they were primed with an idiosyncratic property (e.g., lives in the water).

In the present studies we examined whether contextual factors, such as how evidence exemplars are presented, may impact the extent to which participants obey the diversity principle. We were particularly interested in the potential influence of the spatial location of exemplars for two related reasons. The first concerns findings from research demonstrating that taxonomic categories tend to be, in many ways, represented within a multidimensional space which can be described as reflecting psychological distance between exemplars (Collins & Quillian, 1969; Hutchinson & Lockhead, 1977; Rips, 1975; Schaeffer & Wallace, 1969). Among other things this psychological distance can be created by similarity relations; for example, relative to their membership within the bird category robins and sparrows can be considered *close* (they share many properties) whereas robins and ostriches are *far* apart (they share few properties). From this perspective, diverse samples are likely to be those that represent items that have greater representational distance.

The second, related, idea comes from research on situated cognition and embodiment (Barsalou, 2006; Wilson, 2002), in which it has been argued that the way we think about and represent concepts is determined, at least in part, by the way we experience and engage with concepts. For example, in addition to activating semantic features, many of the concepts we reason about (e.g., dogs) activate motor and sensory features (e.g., throwing to-be-retrieved items, going for walks, tugging on a leash, etc.) that reflect simulations of how we might interact with concepts (Barsalou, 2006). At a broad level, the embodiment framework challenges cognitive models to consider the role of the environment for a cognitive system (e.g., Hutchins, 1995).

With theses issues in mind we examined whether creating greater physical distance between exemplars within a sample would impact diversity-based reasoning. In two experiments participants were given inductive problems in which three evidence exemplars were presented either in close proximity to each other (within 1 cm), or far from each other (approximately 10 cm apart) (See Figure 1). Half the evidence samples included a diverse range of exemplars and the other half included a homogenous set of exemplars. Our main prediction was that the greater separation of items would encourage participants to consider the coverage, or range, of the exemplars and therefore would lead to higher ratings for inductive arguments that included diverse samples compared to conditions in which the items were spaced close together.

The experiments assessed two additional factors. The first concerns the contents of the evidence samples. In Experiment 1 the items were represented by pictures of animals used to represent the categories presented in the evidence, whereas in Experiment 2 the items were represented by category labels (see Figure 1). This manipulation allowed us to test whether any potential effects of evidence spacing were due to perceptual processed that governed the way participants compared the physical features of the exemplars (i.e., differences

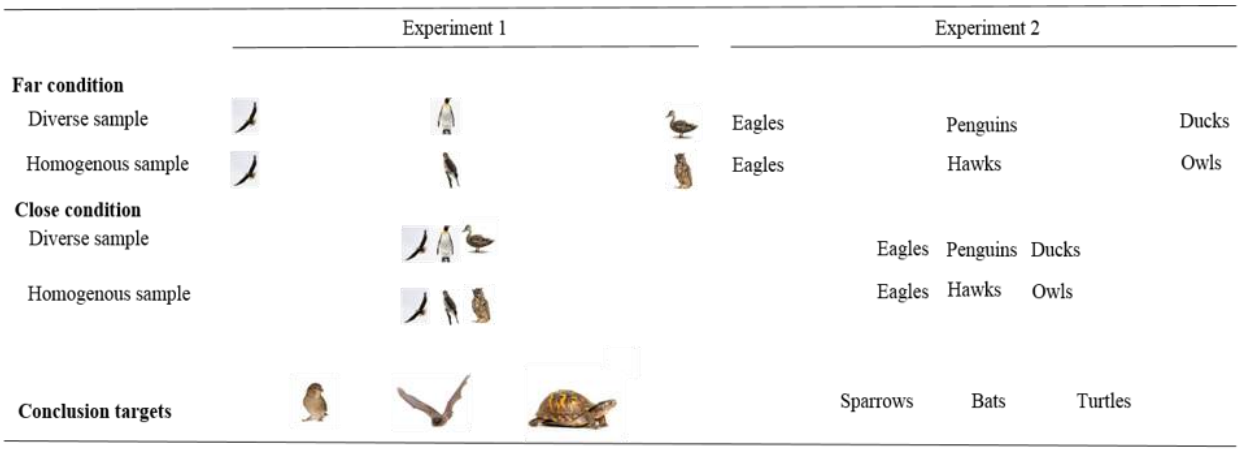


Figure 1. Schematic representation of the design of Experiments 1 and 2.

between the pictured items), rather than differences in how they represented the categories.

Finally, we asked participants to make inductive judgments about three different targets, each of which varied in similarity to/taxonomic distance from the category represented by the evidence exemplars. For example, for the item in which evidence exemplars were represented by birds, participants were asked to make a judgement about a new bird (e.g., sparrow), a bat, and a reptile (e.g., turtle). We expected that the spacing effects would be limited to the category that represents the lowest level of abstraction covered by the evidence and target exemplars (e.g., Osherson et al., 1990). However, if the spatial manipulation has a more general effect on how individuals compare stimuli, it is possible the spaced presentation could lead to an overall increase in one's willingness to generalize from diverse samples to any targets. Thus, this manipulation allowed us determine the extent to which varying the spatial location of the evidence influenced participants' adherence to the diversity principle, in particular.

Experiment 1

Participants. Fifty-three undergraduate students participated for extra credit in a college course. Participants were recruited from, and were representative of, a medium-sized Midwestern US city.

Design. This experiment employed a 2 x 2 x 3 design with Spatial location of evidence exemplars (Close, Far) manipulated between subjects and Sample composition (Diverse, Non-diverse) and Conclusion target (Same basic-level, Similar superordinate, Dissimilar superordinate) manipulated within subjects. Participants were randomly assigned to either the Close condition or the Far condition such that there was an approximately equal number of participants in the two conditions: Close (N = 27), and Far (N = 26).

Materials. Participants were presented 12 inductive reasoning problems each of which included a sample comprised of 3 evidence exemplars. Half of the samples included a diverse set of exemplars (e.g., eagles, penguins, ducks) and the other half included a homogenous set of exemplars (e.g., eagles, hawks, owls). A novel biological property (e.g., Enzyme A) was attributed to the exemplars within the sample. A different novel property for each of the twelve problems.

For each reasoning problem participants were asked to make judgements for 3 different conclusion targets each of which varied in relation (taxonomic and/or perceptual relatedness) to the category covered by the evidence exemplars (see Figure 1 for a sample item). One target was drawn from same basic-level category that was represented by the evidence exemplars (e.g., sparrows). The other two target categories were drawn from superordinate categories (e.g., bats and turtles). Each of the evidence exemplars and targets were represented by photographs of a single animal (2cm x 2cm).

Procedure. The experiment was conducted on a desktop computer with a 24" screen. The spatial arrangement manipulation involved varying the location of the evidence exemplars as they appeared on the screen. In both conditions the exemplars were presented in row on the top of the screen. In the *Far* condition the exemplars were spaced so that there was an approximately 10 cm gap between each. In the *Close* condition the exemplars were bunched together so that there was an approximately 1 cm gap between each. For each item the three evidence exemplars were presented at the same time and were accompanied by a statement (appearing below three exemplars) that attributed a property to all the animals (e.g., "these animals have Enzyme A").

After the evidence exemplars were presented participants were asked to make a judgment about each of the three conclusion targets. A photograph of an

animal used to represent the category was presented approximately 6 cm below the evidence exemplars and was accompanied by a prompt to judge the likelihood that the target category would have the property that was attributed to the evidence exemplars (e.g., "How likely is it that sparrows have Enzyme A?"). Participants were instructed to use a scale, ranging from 0 ("not at likely" to 100 ("very likely), to determine their likelihood judgment. The three target categories were presented in random order.

Note that because sample diversity was manipulated within subjects we counterbalanced across participants which category was represented by a diverse sample or homogenous sample of exemplars. Also the order of presentation of diverse and homogenous samples (within participants) was randomized.

Results

Average likelihood ratings were submitted to a mixed ANOVA with Spatial arrangement of evidence (Close, Far) as the between subjects variable and Sample composition (Diverse, Homogenous), and Conclusion target as the within subjects variables. The only significant main effect was Conclusion target, $F(2, 102)=374.93$, $p<.001$, $\eta^2>.87$, due to a stepwise decrease in likelihood ratings as a function of the decrease in similarity/increase in taxonomic distance from the evidence exemplars to the target categories, all ps<.001 Tukey's HSD.

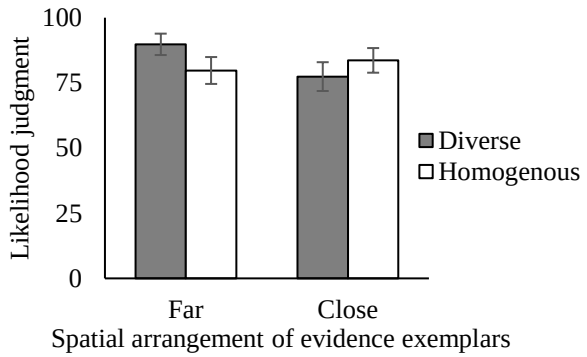


Figure 2. Mean likelihood judgements for basic-level targets for Diverse samples and Homogenous samples in both Spatial arrangement conditions in Experiment 1. Bars represent 1+/- SE from the mean.

The analysis also yielded several noteworthy interactions, all of which were captured by a 3-way interaction, $F(2,102)=4.72$, $p<.02$, $\eta^2=.09$. Analyses of the effects of spatial location on the ratings for each target revealed different patterns of responses for only the Basic-level targets. Simple effects analyses indicated there was a sample location by sample diversity interaction for basic-level targets $F(1,51)=8.49$, $p=.005$, $\eta^2=.17$. As suggested by Figure 2, this interaction was due to differences in responses for Diverse samples of evidence, for which the ratings were significantly higher in the Far condition than the Close condition, $p<.001$. Additional comparisons indicated that were significant differences in ratings between diverse samples and homogenous samples in the Far condition $F(1,25)=7.89$, $p=.01$, $\eta^2=.12$, but not in the Close condition ($F<1.50$). No other effects or interactions were significant (all Fs<1.60).

Discussion

These results indicate that the spatial location of evidence exemplars had a consistent and precise effect on judgments about diverse, but not homogenous, samples. Adults consistently gave higher likelihood ratings for diverse samples when the evidence exemplars were separated from each other than when they were presented in close proximity to each other. However, the effect of spatial location was only present for targets from the same, basic-level, conclusion category as the evidence exemplars. Thus, these results provide preliminary evidence in support of our prediction that contextual factors, such as the way evidence is presented, can facilitate diversity-based reasoning.

Experiment 2

This experiment was designed to address at least two concerns raised by Experiment 1. First, because the items were represented by a photograph of a single animal it remains unclear if participants interpreted the exemplars as representative of the categories they were intended to represent or if they interpreted the evidence as representative of single individual concepts. Second, because the materials included photographs it is possible the effects were due to differences in how participants compared the stimuli, rather than their assessment of the diversity represented by the categories in the evidence. We addressed both of these concerns in this experiment by replacing the photographs with category labels.

Method

Participants. Forty-nine undergraduate students participated for extra credit. Participants were recruited from, and representative of the population of, a medium-sized Midwestern US city.

Design, Materials, and Procedures. This experiment was identical to Experiment 1 in every respect except the stimuli. In this case, rather than presenting photographs to represent the evidence and target items, participants were presented category labels. Participants were randomly assigned to the Far (N=24) or Close (N=25)

conditions. See Figure 1 for a schematic of the study design.

Results

The analysis replicated the pattern of results that was found in Experiment 1. Again there was a three-way interaction between Spatial arrangement, Sample Composition, and Conclusion target, $F(2, 92)=3.48$, $p=.02$, $\eta^2=.078$. Further analysis revealed a significant Sample composition by location effect interaction for Basic-level targets, $F(1, 46) = 10.32$, $p=.002$, n=.18. As suggested by Figure 3, the interaction was due to higher ratings for Basic-level targets for diverse samples than homogenous samples in the Far condition, $F(1, 47) = 12.12$, $p<.001$, $\eta^2=.13$, but not in the Close condition, $F<1.00$. Also, participants exhibited higher likelihood ratings for diverse samples in the Far condition than in the Close condition, $F(1,97)=6.42$, $p=.03$, $\eta^2.09$ As was the case in Experiment 1, no other effects or interactions were significant.

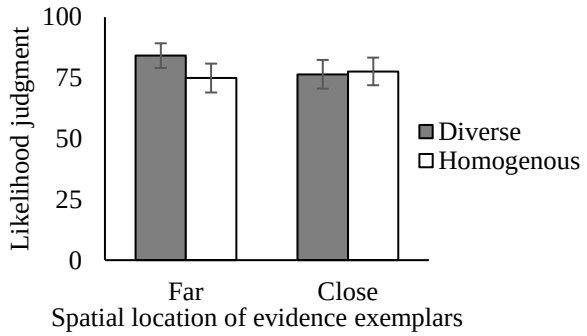


Figure 4. Mean likelihood judgements for basic-level targets for Diverse samples and Homogenous samples in both Spatial location conditions in Experiment 2. Bars represent 1+/- SE from the mean.

Discussion

These results replicated those found in Experiment 1. Participants rated arguments with diverse samples as providing better support for conclusions than arguments with homogenous samples when the exemplars were spatially distant from each other but not when these same evidence exemplars were in close proximity. Moreover, the diverse sample were given higher ratings when the evidence exemplars were more distant than when they were close. These results suggest that the findings from Experiment 1 were not due to participants' interpretation of the evidence as applying to specific individuals, rather than categories, nor their reliance on perceptual features of the task.

General Discussion

Prior research indicates that diversity-based reasoning is dictated by our knowledge or expectations about the categories, or the to-be-generalized properties that are represented in the evidence (Feeney & Heit, 2011; Heit, 1998; Osherson et al., 1990). Thus most existing models account for diversity effects by focusing on how people reason about the content of an inductive problem. In the two experiments reported here we demonstrated that certain contextual factors, such as the way exemplars are presented, also contribute to diversity-based reasoning. Specifically, participants showed a greater willingness to endorse arguments that included diverse samples when these samples were presented in such a way that there was a large spatial separation between each of the evidence exemplars relative to when the same exemplars were presented without a large separation between evidence exemplars. In other words, diversity effects were strongest when the evidence covered more physical space.

It is difficult to reconcile these results with current explanations for diversity effects. One could argue that the spacing effects in Experiment 1 are consistent with the feature-based induction model of induction (Sloman, 1993) insofar as the presentation impacted the way participants compared stimuli, or identified overlapping or unique features, and thus impacted their calculation of diversity. However, this interpretation does not account for the observed effects in Experiment 2 in which the stimuli were represented by labels rather than images. The results are also inconsistent with the idea that a calculation of similarity between the evidence exemplars is sufficient to assess category coverage (Osherson et al., 1990). The similarity coverage model does not account for the finding that participants gave higher ratings for diverse samples when the evidence exemplars were spread far apart compared to when they were positioned close together.

Those in favor of the Bayesian or Relevance accounts of induction might interpret the results as the outcome of pragmatic factors. It could be argued that participants assumed that the exemplars were purposefully placed in close proximity or far apart. According to the Relevance theory of induction (Medin et al., 2003), participants rely on standard rules of communication (e.g., Grice, 1975; Sperber & Wilson, 1995), such as the notion that people present information in such a way as to highlight a relevant piece of information. Thus, it could be argued that participants reasoned as-if the exemplars were deliberately placed apart to draw attention to the coverage of the exemplars (or placed together to highlight the similarities between them). Although these models can accommodate these effects of spatial location, they do not explain them. Assuming participants reasoned that exemplars were spread apart purposefully, why would they interpret this decision was

intended to highlight the coverage of evidence exemplars?

Our interpretation of these results is that they provide some support for the grounded, or situated, aspect of diversity-based reasoning. Most models of situated cognition tend to focus on the influence of different modalities (e.g., motor sequences) on conceptualization. Diversity refers to a feature of samples, not isolated concepts. Diverse samples are those that include exemplars that provide coverage of a category; diverse samples occupy greater psychological space. Here we showed that presented evidence in such a way that the sample occupied greater physical space facilitated diversity-based reasoning. That the spacing effects were not observed for homogenous samples, or for targets from more distant conclusion categories, suggests that spacing did not influence *whether* participants interpreted samples as diverse. Rather, the results indicate that the broad spacing of exemplars primed cognitive processes that can draw attention to sample diversity, and thereby strengthen diversity effects.

Clearly, more work is needed to better understand the scope of these effects. For example, to clarify the potential impact of participants' pragmatic assumptions it will be important to determine whether we can replicate these effects in conditions in which participants are made to believe that the location of the exemplars was not chosen deliberately. Additionally, because participants did not show the diversity effect in the Close condition it will be important to replicate these findings with a different set of stimuli. Also, it will be important to explore other ways in which individuals might be primed to consider the breadth or scope of evidence. For example, we are currently exploring the impact of gestures on adults' and children's adherence to diversity and sample size principles of induction.

There are several notable limitations of these experiments. First, the methods were different from those typically used on the inductive reasoning literature. Participants are often given arguments and asked to determine the sample that provides the best support for a conclusion. Here the evidence exemplars were presented as single photographs or category labels, rather than premises in an inductive argument. The effect of spacing might have been pronounced because this method encouraged participants to compare the stimuli. Also, although the observed effects were consistent, they were rather small. It will be important to replicate these results with stimuli from different domains to be sure these effects are not exclusive to the set of items.

Despite these limitations, these results raise important questions about the impact of contextual factors on inductive reasoning. In particular we showed that presenting evidence in such a way that it covered a broad physical space provided greater support for diversity-based reasoning than when the same evidence was presented in a narrow physical space. While we do not deny that there is still much to learn about the influence of category knowledge, and prior beliefs, on inductive reasoning, these experiments call for more work on understanding the impact of contextual features on induction. As much as inductive reasoning is influenced by *what* is presented in an inductive problem, it seems intuitive that it would be influenced, at least to a certain degree, by *how* evidence is presented.

References

Barsalou, L.W. (2006). Situated conceptualization: Theory and application. In Y. Coello & M.H. Fischer (Eds.), *Foundations of embodied cognition.* East Sussex, UK: Psychology Press.

Collins, A.M., & Quillian, M.R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior, 8,* 240-248.

Feeney, A., & Heit, E. (2011). Properties of the diversity effect in category-based inductive reasoning. *Thinking & Reasoning, 17*, 156-181.

Grice, H.P. (1975). "Logic and Conversation," In P. Cole and J. Morgan, *Syntax and Semantics*. Academic Press.

Heit, E. (2000). Properties of inductive reasoning. *Psychonomic Bulletin and Review, 7*, 569-592.

Heit, E. (1998). A Bayesian analysis of some forms of inductive reasoning. In M. Oaksford & N Chater (Eds.), Rational models of cognition (pp. 248-274). Oxford, UK: Oxford University Press.

Heit, E., Hahn, U., & Feeney, A. (2005). Defending diversity. In W. Ahn, R. Goldstone, B. Love, A. Markman, & P. Wolff (Eds.), *Categorization inside and outside of the laboratory: Essays in honor of Douglas L. Medin*. Washington, DC: American Psychological Association.

Hutchins, E. (1995). *Cognition in the wild.* Cambridge, MA: MIT Press.

Hutchinson, J.W., & Lockhead, G.R. (1977). Similarity as distance: A structural principle for semantic memory. *Journal of Experimental Psychology: Human Learning and Memory, 3*, 660-678.

Kim, N. S., & Keil, F. C. (2003). From symptoms to causes: Diversity effects in diagnostic reasoning. *Memory & Cognition, 31*, 155-165.

Lawson, C.A., & Kalish, C.W. (2009). Sample selection and inductive generalization. *Memory & Cognition, 37*, 596-607.

Lo, Y., Sides, A., Rozelle, J., & Osherson, D. (2002). Evidential diversity and premise probability in young children's inductive judgment. *Cognitive Science, 26*, 181-206.

Lopez, A. (1995). The diversity principle in the testing of arguments. *Memory & Cognition, 23*, 374-382.

Medin, D. L., Coley, J.D., Storms, G., & Hayes, B.K. (2003). A relevance theory of induction.

Psychonomic Bulletin & Review, 10(3), 517-532.
Osherson, D.N., Smith, E.E., Wilkie, O., Lopez, A., & Shafir, E. (1990). Category-based induction. *Psychological Review, 97*, 185-200.
Proffitt, J. B., Coley, J. D., Medin, D. L. (2000). Expertise and category-based induction. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 26*, 811-828.
Rips, L.J. (1975). Inductive judgements about natural categories. *Journal of verbal learning and verbal behavior, 14*, 6.
Sloman, S.A. (1993). Feature-based induction. *Cognitive Psychology, 25*, 231-280.
Schaeffer, B., & Wallace, R. (1969). Semantic similarity and the comparison of word meanings. *Journal of Experimental Psychology, 82*, 343-346.
Sperber, D., & Wilson, D. (1995). *Relevance: Communication and cognition.* Wiley-Blackwell.
Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review, 9*, 625-636.

Exploring the Representation of Linear Functions

Pablo León-Villagrá[1] Verena S. Klar[1] Adam N. Sanborn[2] Christopher G. Lucas[1]
[1] School of Informatics, University of Edinburgh, United Kingdom
[2]Department of Psychology, University of Warwick, United Kingdom

Abstract

Function learning research has highlighted the importance of human inductive biases that facilitate long-range extrapolations. However, most previous research is focused on aggregate errors or single-criterion extrapolations. Thus, little is known about the underlying psychological space in which continuous relationships are represented. We ask whether people can learn the distributional properties of new classes of relationships, using Markov Chain Monte Carlo with People, and find that (1) people are able to track not just the expected parameters of a linear function, but information about the variability of functions in a specific context and (2) in many cases these spaces over parameters exhibit multiple modes.

Keywords: generalization, function learning, representation

Inductive biases are at the heart of the human ability to generalize and extrapolate from sparse evidence. For instance, when we infer labels and properties of new objects or entities, we rely not just on our experience of past examples, but on our implicit and explicit expectations about the nature of categories. Similarly, when we learn relationships between quantities in *function learning*, inductive biases make it possible to distinguish between the boundless possible relationships behind a set of observations (Lucas et al., 2012, 2015).

In order to characterize a person's inductive biases, it can be useful to first focus on spaces of possible mental representations – sometimes called hypothesis spaces – and the kinds of inferences they support or preclude. As with categorization, there have been many proposals about the mental representations supporting function learning, including exemplar-based approaches (McDaniel & Busemeyer, 2005), rule-based approaches (Brehmer, 1974), and hybrids or generalizations of these (DeLosh et al., 1997; Lucas et al., 2015). These models are typically evaluated by comparing their predictions to averaged human judgments, either via direct correlations, error relative to the true underlying function, or qualitative features including multiple modes (Kalish et al., 2004), or monotonicity (Bott & Heit, 2004; Kalish, 2013).

While this line of research has shed light on function learning and the representations and inductive biases that make it possible, some fundamental questions remain. For example, while models that take a distributional approach to function learning have successfully explained human behavior, there is little direct evidence that people track distributional information – uncertainty or variability – when faced with function learning problems. This question has been unanswerable in previous work that relied on aggregated judgments or assumed that individual inductive biases are broadly similar (Kalish et al., 2007). Even the few studies that have focused on inference patterns (Kalish, 2013; Wilson et al., 2015; Schulz et al., 2017), including analyses of per-participant extrapolations (León-Villagrá et al., 2018), still neglected this question about the tacit beliefs behind participants' judgments. Only recently, experiments have started to explore the role of uncertainty in function learning. In Schulz et al. (2015) participants judged functions to be more predictable when they were smooth or when they exhibited low variance, much in accordance with the preferences of a probabilistic model. Similarly, Stojic et al. (2018) showed that participants' predictive accuracy in a function learning task correlated with their confidence ratings, again resembling the uncertainty estimated by a probabilistic model.

Here we expand on this work and attempt to directly characterize how people represent uncertainty when they learn functions.

Markov Chain Monte Carlo with People

To uncover the psychological space that participants learn when learning functions we apply Markov Chain Monte Carlo with People (MCMCP; Sanborn et al., 2010). Sanborn et al. showed that Markov Chain Monte Carlo can be used as an experimental method to elicit posterior distributions from people using a simple forced-choice task. Thus, MCMCP offers a method to explore the psychological representational space and has been successfully applied to elicit the representations of complex stimuli, such as facial affect categories (Martin et al., 2012). Previously, MCMCP has been used in a function learning setting[1] to examine if participants prefer compositional over non-compositional functions (Schulz et al., 2017). Since Schulz et al. were interested in preferences for types of functions (compositional vs. non-compositional), the samples presented consisted of discrete varieties of functions and did not explore the distribution of function parameters.

In contrast, in this work, we directly explore the distributional space of the parameters governing the realizations of linear functions. This allows us to uncover how learned functions are represented, without constraining the participant's choices to pre-specified sets of materials.

Adopting MCMCP also allows us to explore novel questions – do participants represent variability in the training relationships? Do they form a single, deterministic functional relationship or do they form posterior distributions over parameters, reflective of the variability in the training? This question about representation, in turn, can inform more general future questions about extrapolation – are typical extrapolation patterns maximum a posteriori judgments given a

[1]Function learning has been more extensively studied in a closely related paradigm, *iterated learning*. Iterated learning experiments can elicit participants' shared expectations and have revealed strong inductive biases for positive linear functions (Kalish et al., 2007).

learned distribution over parameters? Or do they correspond to samples from a range of probable parametrizations?

In this work we:

- Evaluate if MCMCP can be successfully adapted to a function learning paradigm.
- Contrast how functions are represented depending on the variability of the example sets provided.

Experiment

In this experiment, we examine how participants represent linear functions when presented with sets of training examples. We hypothesize that participants learn both the parameters generating the function, as well as the variability of the relationship, i.e. they will learn both how much slopes and intercepts vary, while also learning the specific modes of slopes and intercepts. Therefore, we expect participants to form posterior distributions over the training parameters, with the variance of that posterior reflective of the training.

We distinguish between training functions with positive and negative slopes, since previous research has highlighted strong inductive biases for these relationships. Similarly, while it has been shown that people are biased to extrapolate in a linear fashion, especially preferring linear functions where both stimulus and criterion are matched (DeLosh et al., 1997), extrapolations appear to be influenced by their proximity to the extrapolation boundaries. In areas of the extrapolation range that are closer to zero, participants seem to adjust the slope of their extrapolations towards this boundary (Brown & Lacroix, 2017; Kwantes & Neal, 2006). To test how different offsets and different degrees of steepness are represented we contrast steep and shallow linear functions. Finally, we expect that highly salient functional relationships, like positive functions for which target and criterion are matched, will be easier to learn and result in more peaked posterior distributions if the training exhibits low variability. For high variability training, and especially if the function is not favored as strongly (for instance a function with a shallow negative slope) we expect broader, less peaked posteriors. Finally, we hypothesize that especially in high variability conditions, some participants will not exhibit unimodal posterior distributions and consider several potential generating functions broadly consistent with the learned function.

Contrasting these functions resulted in a $2 \times 2 \times 2$ between-subjects design (direction of the function: positive or negative, steepness: shallow or steep, variability of the training data: low or high).

Participants

The study was self-certified in accordance with the School of Informatics Ethics Guidelines. We recruited 454 participants ($M_{age} = 33$, $SD_{age} = 8.63$, 91 female, 176 male, 1 other, 186 refused information on gender) on Amazon Mechanical Turk. Participants had to have more than 50 approved HITs and an approval rate of 95% or larger. They received \$1.33 for participation and took an average of 17 minutes ($M = 17.25, SD = 8.59$) to complete the experiment. Participants were randomly assigned to one of the 8 conditions.

Materials

The parameters generating the functions in the experimental conditions differed in the sign of the slopes, as well as in their steepness. In addition, parameters in the training set exhibited either low or high variance for intercepts and slopes. For the full set of experimental conditions, see Table 1.

Table 1: Parametrization for the generating linear functions.

Condition	β_0	SD_{β_0}	β_1	SD_{β_1}
$C_{.5,\text{low}}$	0.25	0.05	0.5	0.025
$C_{1.0,\text{low}}$	0	0.05	1	0.025
$C_{-.5,\text{low}}$	0.75	0.05	-0.5	0.025
$C_{-1.0,\text{low}}$	1	0.05	-1	0.025
$C_{.5,\text{high}}$	0.25	0.3	0.5	0.15
$C_{1.0,\text{high}}$	0	0.3	1	0.15
$C_{-.5,\text{high}}$	0.75	0.3	-0.5	0.15
$C_{-1.0,\text{high}}$	1	0.3	-1	0.15

To create the 25 training sets, corresponding to iid realizations of $\beta_0, \beta 1 \sim \mathcal{N}(\mu, \sigma)$, with μ and σ matching the experimental condition, we systematically sampled 10,000 pairs and selected the most normal and uncorrelated sets[2]. Then we generated the corresponding linear function for a range of 15 points for x in 0–1 for all sets. One of those 15 values was picked at random and constituted the interpolation target.

MCMCP Proposals were generated by two symmetric Gaussian distributions, to allow both for local, as well as far-off proposals, $\sigma_{\beta_0} \in [0.14, 0.98]$, $\sigma_{\beta_1} \in [0.21, 1.47]$. At each iteration these proposals had a probability of .8 and .2 to be selected. Proposals were further restricted to be in bounds $\beta_0 \in [-0.5, 1.5]$, $\beta_1 \in [-1.5, 1.5]$, and if less than 4 points of the function realization were visible on screen, the proposal was automatically rejected and a new proposal was resampled. Participants traversed three different, interleaved chains, since multiple chains allow a wider application of convergence diagnostics and reduce the impact of the particular starting state. The starting values for these chains were obtained by k-means clustering of pilot data ($n = 8$, one participant per condition). This resulted in the following starting values $\beta_0 = \{0.12, 0.1, 0.58\}, \beta_1 = \{0.92, -0.94, -0.28\}$, for chains 1 to 3.

Procedure

Participants were instructed that they would learn the relationship between two proteins, Zenopin and Mepradin. Participants were told that the concentration of Zenopin was related to Mepradin, but that the extent of that relationship var-

[2] All Shapiro-Wilk tests yielded $p > 0.99$, and all correlation coefficients were in the range $[-.01\ .01]$.

ied between humans. Participants were also instructed that they would be presented with examples of the relationship as observed in different people and that they would be asked to interpolate the relationship. They were then instructed that after the training phase they would be presented with pairs of proposed relationships, all observed for a new person, and would have to choose which of the two were more likely to resemble the learned relationship. After reading the set of instructions, the participants were tested on their comprehension. If participants did not respond correctly in the questionnaire they had to restart the instructions.

Training Phase In the training phase, participants were presented with 25 interpolation tasks, presented as scatter plots. In each task, they were instructed that the scatter plot depicted the relationship between the two protein concentrations for a new person. They then had to guess the concentration of the protein by selecting the height of the corresponding value on the plot (on the y-axis). Participants were shown the correct value as feedback for one second, and, if their choice deviated by more than ± 0.05 from the true value, had to readjust their selection.

Test Phase The test phase consisted of 240 forced-choice tasks, corresponding to 80 interleaved iterations of the three Markov chains. On each trial, participants were presented with two adjacent scatter plots, one corresponding to the current state of the chain and the other reflecting the proposed new state (in randomized order). Participants had to select the plot they believed most likely to depict the relationship in the training phase. After the test phase, participants completed a short survey, were debriefed, and compensated. See Figure 1 for a depiction of both training and test phase.

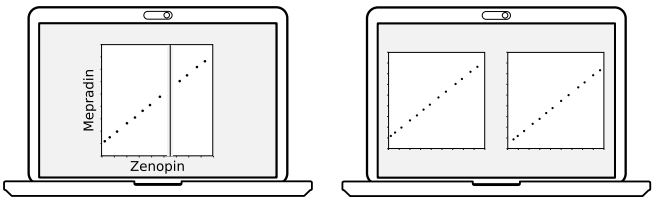


Figure 1: Participants had to complete a training and a test phase. In the training phase they were asked to interpolate the concentration of a fictitious protein for 25 different people (with feedback). In the test phase, they were presented with 240 forced-choice tasks, for which they had to choose the scatter plot that most resembled the relationship in the training phase. The choices were presented in random order and corresponded to a Markov chain, in which the participant implemented the acceptance function.

Results

We excluded participants from the analysis if their chains did not converge to the stationary distribution. Many criteria for convergence checks have been suggested in the literature, here we applied one of the most commonly used evaluations, $\hat{R}$ (Gelman et al., 2013; Vehtari et al., 2019). $\hat{R}$ estimates the ratio between within-chain variances and between-chain variance and thus provides a measure of how (self-)similar chains are. In general applications $\hat{R}$ should not exceed a value of 1.1. However, such a strict application of this diagnostic is not realistic in most MCMCP experiments, since human judgments might exhibit more correlated choices and the number of iterations in experiments is usually considerably lower than in standard statistical applications. Therefore, we incrementally calculated $\hat{R}$ values for chains for each participant and selected the lowest overall $\hat{R}$, with the additional constraint that the first 20 samples of the chain were always discarded and the resulting chains had to be at least 20 iterations long. We then used the maximum of the intercept and slope $\hat{R}$ values to apply exclusion criteria and determine burn-in.

Similar to Ramlee et al. (2017), we excluded participants who exhibited $\hat{R} \geq 2$. Furthermore, we excluded participants who required more than one correction in the interpolation task. Given that the interpolation function was deterministic, most participants did not require many corrections ($Mdn = 0, \mathrm{Q1} = 0, \mathrm{Q3} = 1, max = 44$).

In total, these methods led to the exclusion of 262 participants (convergence exclusions: 224, interpolation exclusions: 72). This high number of exclusions was to be expected given the correlated, bi-variate parameter space and previous results (Sanborn et al., 2010). For group sizes after exclusion, see Table 2. For an overview of how the forced-choice task results in the posterior distribution, see Figure 2.

Determining Burn-in

To determine how many trials were required on average for the Markov chains to converge, we used the iteration for which $\hat{R}$ was optimal for each participant. On average, chains required 33 iterations to reach optimal burn-in and the resulting optimal $\hat{R}$ values were well below 2, $M_{\hat{R}} = 1.4, SD = 0.2$. Conditions did not differ considerably in terms of the optimal iterations or the resulting $\hat{R}$ values. For the full list of per-condition burn-in values, see Table 2. For all subsequent analysis, we discarded all points of the chain before the per-participant burn-in.

Table 2: Participants in each condition before (N_{total}) and after exclusion (N). $M_{\text{burn-in}}$, $SD_{\text{burn-in}}$, as well as mean acceptance probabilities averaged over participants (M_{acc}, SD_{acc}).

Condition	N_{total}	N	$M_{\text{burn-in}}$	$SD_{\text{burn-in}}$	M_{acc}	SD_{acc}
$C_{.5,\text{low}}$	48	25	34.88	14.49	35	17
$C_{1.0,\text{low}}$	63	21	31.37	12.01	42	10
$C_{-.5,\text{low}}$	52	19	34.37	13.73	37	13
$C_{-1.0,\text{low}}$	64	22	29.59	11.59	38	15
$C_{.5,\text{high}}$	59	35	32.29	13.22	38	14
$C_{1.0,\text{high}}$	57	26	32.08	12.24	45	9
$C_{-.5,\text{high}}$	56	29	35.66	12.75	42	13
$C_{-1.0,\text{high}}$	55	15	29.40	10.67	36	12

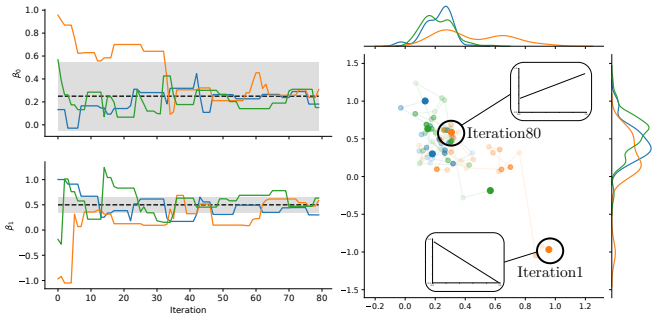


Figure 2: The 240 choices submitted by the participants corresponded to three Markov chains. By accepting or rejecting proposed parametrizations for the functions, participants traversed this representational space and eventually converge to a region reflecting the posterior over parameters. For this participant, the chains converge after 35 iterations for β_0 and after 15 iterations for β_1. The corresponding distribution after this burn-in period closely matches the true relationship learned in the training phase, both in terms of its mean and variance (dashed line and grey range).

Acceptance Probabilities

Acceptance rates for MCMC samples should range between 20–40% (Roberts, Gelman, & Gilks, 1997). Mean acceptance probability was in that range, $M = 39\%, SD = 13$, indicating that the proposals were wide enough to traverse the parameter space. Between conditions, the mean acceptance probabilities for participants varied, ranging from 35 to 45%, for all acceptance probabilities, see Table 2. For each condition, acceptance probabilities for each chain did not vary substantially and were similar to the general acceptance rates (not shown).

Posterior Distributions

Slopes differed significantly between positive- and negative-slope conditions, with participants trained on negative slopes preferring negative slopes, $M_{\beta_1} = -0.16, SD_{\beta_1} = 0.53$, and participants trained on positive slopes preferring positive slopes, $M_{\beta_1} = 0.19, SD_{\beta_1} = 0.45$, $t(165.33) = -4.74, p < .0001$[3].

For conditions with negative slopes in the training sets, steep and shallow conditions exhibited significantly different posterior slopes, with lower slopes for steep compared to shallow conditions, $M_{-.5} = -0.05, SD_{-.5} = 0.45, M_{-1.0} = -0.29, SD_{-1.0} = 0.59$, $t(65.58) = 2.08, p = .041$. For conditions with positive slopes in the training sets there was also a significant difference in posterior slopes. However, this difference was not in the predicted direction, as slopes in the shallow condition were on average larger than in the steep condition, $M_{.5} = 0.29, SD_{.5} = 0.4, M_{1.0} = 0.05, SD_{1.0} = 0.47$, $t(89.75) = -2.89, p = .005$. Posterior intercepts in conditions with negative training slopes did not differ significantly between steep and shallow conditions, $M_{-.5} = 0.52$, $SD_{-.5} = 0.21$, $M_{-1.0} = 0.6$, $SD_{-1.0} = 0.3$, $t(62.84) = 1.38$, $p = .174$, nor for conditions with positive training slopes, $M_{.5} = 0.35, SD_{.5} = 0.2, M_{1.0} = 0.5, SD_{1.0} = 0.25, t(88.71) = 3.31, p = .001$.

[3] All tests are unequal variance, two-sided t-tests.

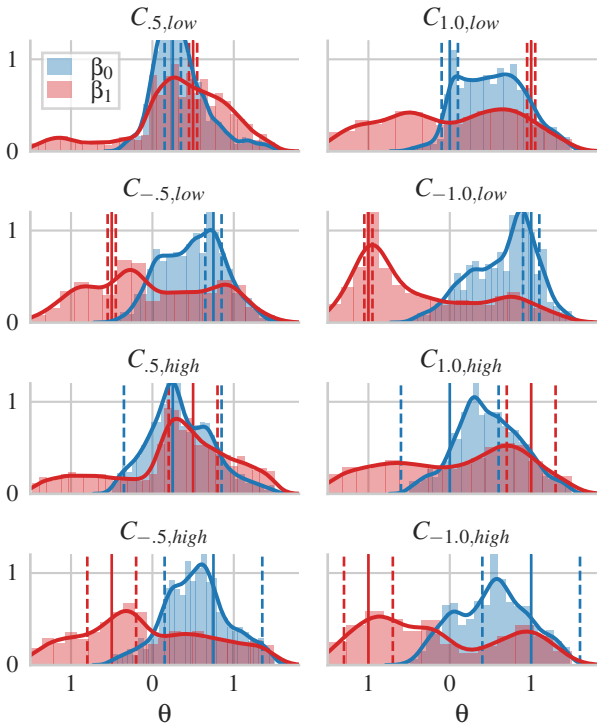


Figure 3: Posterior densities for intercepts and slopes and true values and standard deviation (dashed lines) in the experimental conditions. The posterior densities exhibited multiple modes, some centered in close proximity of the true parameters.

Equally, per-participant SD for slopes did not differ significantly between high and low variability conditions, $M_{low,\beta_1} = 0.49$, $SD_{low,\beta_1} = 0.26$, $M_{high,\beta_1} = 0.55$, $SD_{low,\beta_1} = 0.25$, $t(180.07) = -1.39$, $p = .166$. However, for intercepts, per-participant SD did differ significantly between high and low variability conditions, with high variance conditions resulting in higher SD, $M_{low,\beta_0} = 0.26, SD_{low,\beta_0} = 0.11, M_{high,\beta_1} = 0.31, SD_{low,\beta_1} = 0.11, t(182.48) = -2.46, p = .015$.

Visual inspection revealed that in all conditions posterior distributions were multimodal and heavily skewed, which complicated the analysis. In general, the posterior densities suggested that the modes of the posterior distributions were often close to the learned parameters, see Figure 3, for a selection of posterior distributions for one participant in each condition, see Figure 4.

Since the mean and standard deviations of multimodal, heavily skewed distributions are not good representations of the underlying data and we were interested in characteristic modes of the distributions, we used mixture models to identify dominant modes of the posterior distributions.

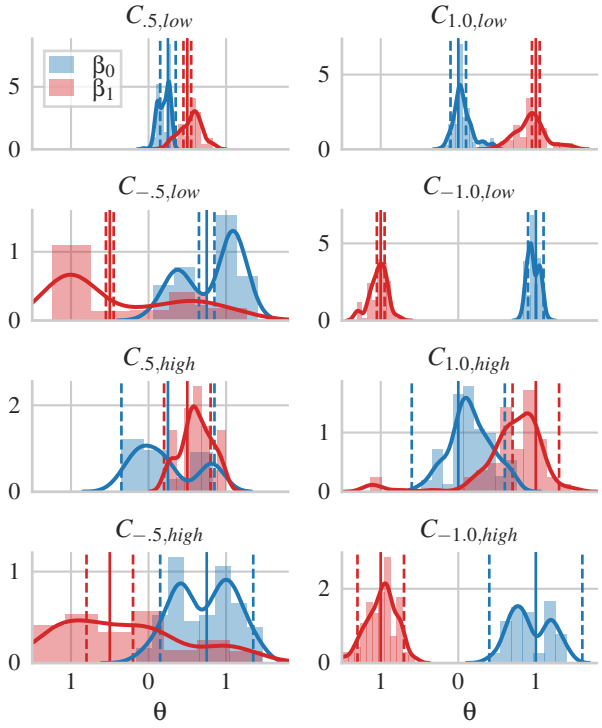


Figure 4: Posterior densities for one participant in each condition. Lines represent the true values and standard deviations (dashed lines) in the experimental conditions.

Table 3: Posterior means and variances per condition, for function intercepts (β_0) and slopes (β_1).

Condition	M_{β_0}	SD_{β_0}	M_{β_1}	SD_{β_1}
$C_{.5,\text{low}}$	0.34	0.32	0.32	0.62
$C_{1.0,\text{low}}$	0.52	0.40	0.00	0.78
$C_{-.5,\text{low}}$	0.49	0.37	−0.02	0.73
$C_{-1.0,\text{low}}$	0.65	0.40	−0.40	0.77
$C_{.5,\text{high}}$	0.35	0.39	0.27	0.71
$C_{1.0,\text{high}}$	0.47	0.40	0.07	0.81
$C_{-.5,\text{high}}$	0.54	0.40	−0.07	0.77
$C_{-1.0,\text{high}}$	0.52	0.44	−0.20	0.83

Estimating Posterior Density Clusters We estimated Gaussian mixture models that best described the distributions for each experimental condition. We incrementally increased the number of components and selected the model with the lowest BIC[4]. The clustering produced a moderate number of clusters, reflecting the multimodal nature of the data. In general, each condition was estimated to correspond to a mixture of 1–8 clusters ($M = 4.5, SD = 2.56$), and the largest clusters closely matched the different training conditions. For KL-divergences between training distribution and the inferred clusters, see Table 5, for the number of clusters, weights, means and covariances for the largest clusters, see Table 4, for plots of the clusters, see Figure 5.

[4]Estimating the mixtures with a Bayesian Dirichlet process mixture model yielded very similar results.

Table 4: The total number of clusters (N_c) assigned was generally low and the weight of the largest clusters was relatively large (16–100%).

Condition	N_C	$w_{c=1}$	$\mu_{\beta_{0,c=1}}$	$SD_{\beta_{0,c=1}}$	$\mu_{\beta_{1,c=1}}$	$SD_{\beta_{1,c=1}}$
$C_{.5,\text{low}}$	8	0.2	0.15	0.02	0.69	0.14
$C_{1.0,\text{low}}$	8	0.17	0.07	0.01	0.84	0.1
$C_{-.5,\text{low}}$	1	1.0	0.49	0.14	-0.01	0.53
$C_{-1.0,\text{low}}$	4	0.42	0.93	0.03	-0.98	0.04
$C_{.5,\text{high}}$	2	0.81	0.24	0.10	0.54	0.21
$C_{1.0,\text{high}}$	3	0.46	0.24	0.1	0.75	0.13
$C_{-.5,\text{high}}$	5	0.31	0.93	0.09	-0.65	0.2
$C_{-1.0,\text{high}}$	5	0.39	0.9	0.08	-0.95	0.07

Table 5: KL-divergence between the training distribution and the three largest clusters. In general, one of the largest clusters corresponded well to the training distribution.

Condition	$KL_{c=1}$	$KL_{c=2}$	$KL_{c=3}$
$C_{.5,\text{low}}$	2.18	1.1	1.95
$C_{1.0,\text{low}}$	1.74	42.1	5.85
$C_{1.0,\text{low}}$	1.35	–	–
$C_{-1.0,\text{low}}$	0.31	1.76	24.16
$C_{.5,\text{high}}$	0.83	10.37	–
$C_{1.0,\text{high}}$	1.06	9.76	2.95
$C_{-.5,\text{high}}$	1.49	2.66	4.25
$C_{-1.0,\text{high}}$	1.02	59.27	11.09

Per-Participant Clusters To evaluate if the source of the multimodality in our data was due to averaging over diverse cohorts of participants, or if individual participants produced multimodal posteriors, we performed the same clustering procedure on a per-participant basis. Participant posterior distributions were characterized by 1–12 clusters ($M = 3.11, SD = 1.96, Q1 = 1, Q2 = 3, Q3 = 4$), suggesting that the posterior distributions were composed of multimodal individual distributions. Furthermore, some participants with optimal $\hat{R}$ (≤ 1.1) also exhibited multiple clusters, indicating that the multimodality was not simply due to poor convergence ($M = 1.89, SD = 1.36, N_{\hat{R}\leq 1.1} = 9$).

The number of clusters did not differ significantly between low- and high-variance conditions, $M_{low} = 2.98, SD_{low} = 1.94, M_{high} = 3.1, SD_{high} = 1.57, t(164.24) = -0.49, p = .312$. Neither did the variance of the largest cluster for slopes differ significantly, $M_{low} = 0.1, SD_{low} = 0.13, M_{high} = 0.1, SD_{high} = 0.11, t(172.43) = 0.11, p = .545$. However, for intercepts the variance of the largest clusters was significantly different, with smaller cluster variances for low-variance conditions, $M_{low} = 0.04, SD_{low} = 0.03, M_{high} = 0.05, SD_{high} = 0.04, t(189.85) = -2.09, p = .048$.

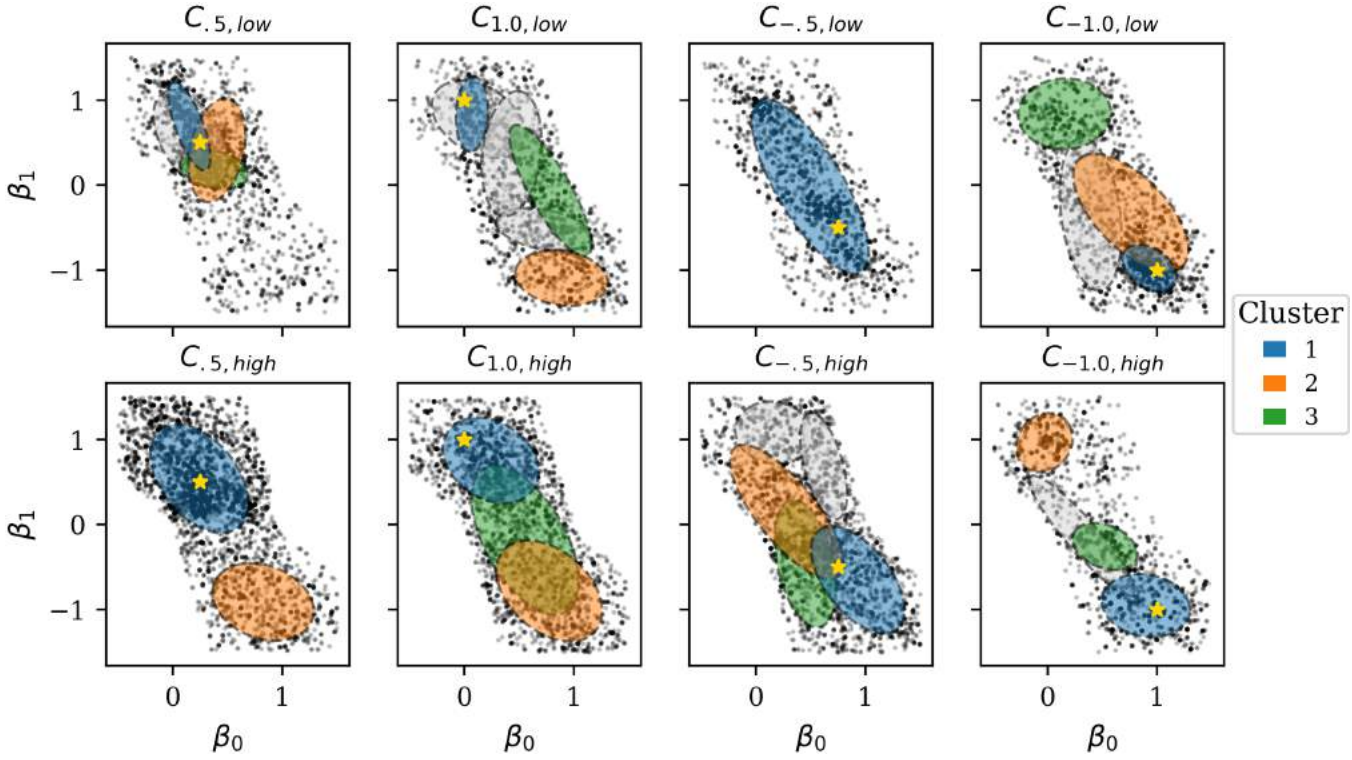


Figure 5: Clusters obtained by fitting a Gaussian mixture model (oval shapes). The top three clusters (colored shapes) accounted for a large proportion of the data and in general matched the distribution learned in the training phase well (mean parameters of the true distribution in yellow).

Discussion

We have found some evidence that participants represent the functions learned in training as distributions over parameters. Furthermore, the modes of these distributions were, in many cases, aligned with the true parameters. In addition, for intercepts, but not for slopes, these distributions were affected by differences in the variability of training. Finally, our results suggest that the learned distributional spaces over function parameters can exhibit multiple modes.

The multimodality in the posterior distributions allows for two interpretations. First, it is possible that participants truly evaluated distinct candidate representations, and thus multimodal posterior distributions characterized their hypothesis space. It is plausible that highly salient relationships, in addition to the implied parameters in the training, constitute the psychological space when learning sets of varying functions. However, the multimodality might also arise from our experimental method. One issue could be the number of iterations. Theoretically, MCMCP is well suited to discover complex, multimodal distributions, but practically many more samples could be necessary to achieve convergence to the posterior distribution. Since extremely large numbers of iterations might not be feasible from an experimental perspective, one practical test of our results could be starting the chains of later participants at the endpoints of previous participants (Martin et al., 2012).

Future research should clarify the source of multimodality, for instance by comparing our results with results obtained by multidimensional scaling (MDS). If such a comparison corroborates our results, these insights into the structure of psychological spaces could, in turn, provide invaluable guidance for future generalization research. In addition, MDS would also allow us to address two shortcomings of the current study: its exclusive focus on linear functions, and the potential influence of perceptual similarity of functions on participants' forced choices. First, similarity judgments obtained via MDS could be used to determine if participants are well described by linear models, or if non-linear representations underlie their judgments. These results would allow us to determine if the multimodal representations observed in our experiment were the result of a lack of satisfactory choices or a genuine characteristic of learning. Second, MDS would allow us to chart sets of perceptually similar samples. It is plausible that intercepts and slopes can affect notions of similarity of linear functions differently. For example, if functions sharing the same slope but very different intercepts are judged more similar than functions with similar slopes and intercepts, such non-linear interactions could explain the multimodality observed in our experiment.

While more research is required, our results also highlight the importance of a plurality of experimental approaches and methods in the study of human generalization. Most of previ-

ous research has focused on averaged errors or single extrapolations. Here, we suggest that to fully understand human generalization, characteristic errors, in combination with extrapolation patterns, and evaluation and exploration of the underlying hypothesis spaces are required.

Acknowledgements

We thank Yevgen Matusevych, Arabella Sinclair and three anonymous reviewers for their helpful comments and suggestions. ANS was supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1.

References

Bott, L., & Heit, E. (2004). Nonmonotonic extrapolation in function learning. *Journal of Experimental Psychology*, *30*(1), 38.

Brehmer, B. (1974). Hypotheses about relations between scaled variables in the learning of probabilistic inference tasks. *Organizational Behavior and Human Performance*, *11*(1), 1–27.

Brown, M., & Lacroix, G. (2017). Underestimation in linear function learning: Anchoring to zero or xy similarity? *Canadian Journal of Experimental Psychology*, *71*(4), 274–282.

DeLosh, E. L., Busemeyer, J. R., & McDaniel, M. A. (1997). Extrapolation: The sine qua non for abstraction in function learning. *Journal of Experimental Psychology*, *23*(4), 968.

Gelman, A., Stern, H. S., Carlin, J. B., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. Chapman and Hall/CRC.

Kalish, M. L. (2013). Learning and extrapolating a periodic function. *Memory & Cognition*, *41*(6), 886–896.

Kalish, M. L., Griffiths, T. L., & Lewandowsky, S. (2007). Iterated learning: Intergenerational knowledge transmission reveals inductive biases. *Psychonomic Bulletin & Review*, *14*(2), 288–294.

Kalish, M. L., Lewandowsky, S., & Kruschke, J. K. (2004). Population of linear experts: knowledge partitioning and function learning. *Psychological Review*, *111*(4), 1072.

Kwantes, P. J., & Neal, A. (2006). Why people underestimate y when extrapolating in linear functions. *Journal of Experimental Psychology*, *32*(5), 1019.

León-Villagrá, P., Preda, I., & Lucas, C. G. (2018). Data availability and function extrapolation. In *Proceedings of the Annual Meeting of the Cognitive Science Society.*

Lucas, C. G., Griffiths, T. L., Williams, J. J., & Kalish, M. L. (2015). A rational model of function learning. *Psychonomic Bulletin & Review*, *22*(5), 1193–1215.

Lucas, C. G., Sterling, D., & Kemp, C. (2012). Superspace extrapolation reveals inductive biases in function learning. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 34).

Martin, J. B., Griffiths, T. L., & Sanborn, A. N. (2012). Testing the efficiency of markov chain monte carlo with people using facial affect categories. *Cognitive science*, *36*(1), 150–162.

McDaniel, M. A., & Busemeyer, J. R. (2005). The conceptual basis of function learning and extrapolation: Comparison of rule-based and associative-based models. *Psychonomic Bulletin & Review*, *12*(1), 24–42.

Ramlee, F., Sanborn, A. N., & Tang, N. K. (2017). What sways peoples judgment of sleep quality? A quantitative choice-making study with good and poor sleepers. *Sleep*, *40*(7).

Roberts, G., Gelman, A., & Gilks, W. R. (1997). Weak convergence and optimal scaling of random walk metropolis algorithms. *The Annals of Applied Probability*, *7*(1), 110–120.

Sanborn, A. N., Griffiths, T. L., & Shiffrin, R. M. (2010). Uncovering mental representations with markov chain monte carlo. *Cognitive psychology*, *60*(2), 63–106.

Schulz, E., Tenenbaum, J. B., Duvenaud, D., Speekenbrink, M., & Gershman, S. J. (2017). Compositional inductive biases in function learning. *Cognitive Psychology*, *99*, 44–79.

Schulz, E., Tenenbaum, J. B., Reshef, D. N., Speekenbrink, M., & Gershman, S. (2015). Assessing the perceived predictability of functions. In *Proceedings of the Annual Meeting of the Cognitive Science Society.*

Stojic, H., Eldar, E., Hassan, B., Dayan, P., & Dolan, R. J. (2018). Are you sure about that? On the origins of confidence in concept learning. In *Preceedings of the Annual Conference on Cognitive Computational Neuroscience.*

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2019). Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of MCMC. *arXiv preprint arXiv:1903.08008*.

Wilson, A. G., Dann, C., Lucas, C. G., & Xing, E. P. (2015). The human kernel. In *Advances in Neural Information Processing Systems* (pp. 2854–2862).

Generalizing Functions in Sparse Domains

Pablo León-Villagrá Christopher G. Lucas
School of Informatics, University of Edinburgh, United Kingdom

Abstract

We propose that when humans learn sets of relationships they are able to learn the abstract structure or *type* of a family of relationships, and exploit that knowledge to improve their ability to learn and generalize in the future, especially in the face of sparse or ambiguous data. In two experiments we found that participants choose patterns and extrapolate in ways consistent with sets of previously learned relations, as measured by extrapolation judgments and forced-choice tasks. We take these results to suggest that humans can detect shared abstract relations and apply this learned regularity to perform rapid and flexible generalization.

Keywords: generalization, function learning, transfer

Introduction

Many everyday situations require us to generalize from past experience, even if we are faced with a specific problem we have never seen before. For example, in cooking, one regularly has to infer the relationship between ingredients, ratios or quantities, like the amount of sweetener and resulting pleasantness of a dessert, and generalize this relation to new recipes or ingredients. Often, we learn a general relationship that helps us understand related problems. If we learn that as we increase the amount of sugar in a recipe, the sweetness will not change immediately, then increases rapidly and then saturates, one can use this knowledge to reason about similar relationships, as when deciding how much xylitol to add to a cake.

Our example requires two types of generalization. The first, sometimes called transfer or transfer learning, involves transferring information about a relationship between two quantities to help us understand a new and different relationship. The second, extrapolation, involves understanding a single relationship and extrapolating to new instances or data points within, e.g., to new amounts of xylitol. The latter depends on the former – our past experiences shape the inductive biases we bring to a new problem.

Transfer learning expands the task the human learner faces and requires further-reaching and more abstract inferences. Given a set of prediction tasks, how can we capitalize on statistical regularities to aid future prediction? If the tasks exhibit some shared structure, learning a representation capturing this latent structure of the environment (Gershman & Niv, 2010), or learning which aspects of a task change (R. C. Wilson & Niv, 2012) can enable the learner to perform wide-ranging and data-efficient generalization.

The value of transferring knowledge across different tasks is receiving growing attention in machine learning communities. For example, abstract learning and transfer have been successfully applied to challenging control tasks (Hamrick et al., 2017). From a cognitive science perspective, the study of such general learning mechanisms has a long tradition, e.g., Harlow (1949). Research in this tradition has highlighted how hierarchical representations can allow for the "blessing of abstraction" (Gershman, 2017), where abstract knowledge is acquired faster than detailed information. In recent years several proposals have been put forward on how hierarchical and structured inductive biases can be acquired through development and how they allow for rapid generalization (Goodman et al., 2008; Tenenbaum et al., 2011).

The second type of generalization has been widely studied in psychology, most commonly in classification tasks in which participants have to learn to predict class labels for unknown objects or entities. Similarly, tasks in which the target to be learned is a continuous quantity have been studied in the domain of function learning research. Research in function learning has emphasized particular human inductive biases. For example, humans learn functions more quickly if the relationship is linear (Brehmer, 1976), and struggle with cyclic functions (Bott & Heit, 2004; Kalish, 2013). More importantly, human extrapolations are strongly biased towards linear relationships, in particular positive linear functions (Brehmer, 1976; DeLosh et al., 1997; Busemeyer et al., 1997; McDaniel & Busemeyer, 2005; Kalish et al., 2004). While this line of research emphasizes simple types of functions, results from experiments with less taxing memory demands have shown that a wide variety of relationships can be learned and inform extrapolation (Lucas et al., 2015; A. G. Wilson et al., 2015; Schulz et al., 2017; León-Villagrá et al., 2018).

In function learning, the hierarchical and abstract representation of the learned relationships has traditionally been reduced to mechanisms that allow generalizing a mapping from criterion to targets. Multiple proposals have been put forward for the nature of these mappings, ranging from rule–like parametric forms (Carroll, 1963; Brehmer, 1976), associative, neuronal network architectures, and hybrids thereof (Busemeyer et al., 1997).

Here we will adopt a general perspective and express the task as Gaussian process regression. A Gaussian process specifies a distribution over functions $f(x) \sim GP(\mu, k)$, where $\mu(x) = E[f(x)]$ and k is the covariance kernel. The kernel specifies a similarity measure over x and allows us to express abstract beliefs about the shape of the function, such as periodicity or smoothness. Gaussian processes have been successful in accounting for both the flexibility in learning, as well as long-range extrapolations (Lucas et al., 2015).

While Gaussian processes allow us to express inductive biases for functions in flexible, non-parametric fashion, only recently more attention has been given to structural and hierarchical aspects of function generalization. This work has emphasized the importance of inductive biases over different

function types (Lucas et al., 2015), the compositional structure of functions (Schulz et al., 2017), or the generalization of functions into dimensions outside the learned space (Lucas et al., 2012).

Here we expand on this line of research and propose that when humans learn relationships they do not maintain sets of data, parametrizations or fixed parametric forms, but that they form flexible and abstract hypothesis spaces. Based on this abstract encoding, we suggest, they are able to capitalize on statistical co-occurrences of abstract information about the *type* of relationship learned. As a result, repeated exposure to similar functions should result in learning about the shared type of relationship, as well as its relevant features. Such exposure should then facilitate extrapolation in sparse contexts and allow far-ranging generalization. We hypothesize that this application of past knowledge does not simply amount to remembering previous data, but extrapolation depends on the induced function type and adapted to the context at hand.

Experiment 1

In this first experiment, we examine if participants prefer functions consistent with the previously learned function type and its shared, defining, features. We train participants on three sets of samples from the same type of function and assess if they subsequently choose extrapolations in concordance with this type and parameters.

Participants

The study was self-certified in accordance with the School of Informatics Ethics Guidelines. We recruited 99 participants ($M_{age} = 32.1$, $SD_{age} = 10.87$, 34 female, 65 male) on Amazon Mechanical Turk. Participants had to have more than 50 approved HITs and an approval rate of 95% or larger. They received \$0.55 for participation and took an average of 7 minutes ($M = 6.46$, $SD = 5.19$) to complete the experiment. Participants were randomly assigned to one of the six conditions ($n_{Cos_1} = n_{Lin_2} = n_{Ou_1} = 17$, $n_{Cos_2} = n_{Lin_1} = n_{Ou_2} = 16$).

Procedure

Participants were instructed that they would learn the relationship between two substances, substance x, and substance y. They were told that they would be presented with three sets of patterns, each depicting one realization of the same relationship and that they would have to predict the relationship for 10 new points. They also received a visual depiction explaining how they would predict the points. They were instructed that they would see one more pattern from the same relationship, consisting of three points. Then they were instructed to select the pattern from six options that most likely depicted the learned relationship.

Training Phase Each training block took the form of an extrapolation task, where participants saw scatterplots and had to guess the value of the substance on the y-axis in an extrapolation range, by selecting the height of the corresponding value on the plot. Participants were shown the correct value as feedback for one second, and, if their choice deviated by ± 0.025 or more of the true value, had to readjust their selection. Training blocks were presented in randomized order.

Choice Phase After the training blocks, there was a forced-choice task where participants saw the three-point pattern and read that this pattern belonged to the same relationship as the training. Then they saw with six scatterplot patterns, corresponding to one conditional sample for each of the six kernels, in randomized order. Participants had to select the pattern that they deemed the most likely extrapolation for the learned relationship. After the choice tasks, participants completed a short demographic survey.

Materials

The functions in the six conditions corresponded to samples from Gaussian Processes (GPs), with three different types of kernels and mean functions, each with two distinct parametrizations, see Table 1. To allow for characteristic periodic samples, we elected a "pure" cosine kernel, *cos* with $k(r) = \sigma \times cos(r)$, $r(x,x') = \frac{(x-x')^2}{\ell_q^2}$, with an additional intercept. We generated linear samples from a linear kernel *lin* with explicit slope and intercept terms. Finally, we used a Ornstein-Uhlenbeck kernel (*OU*) with an additional intercept, to generate non-smooth samples. The noise variance was fixed to 0.01 for all GPs.

Training Sets We generated the training data by sampling three sets of 35 points each in the range 0.05–0.95 for each of the six conditions. The first 25 points constituted the evidence provided in each training set. Participants had to extrapolate the target value for the last 10 points and received feedback for their choices. To ensure that samples were clearly perceptible and the samples were distinct (within function type and between function types) we generated a set of 20 candidate patterns for the 18 sets. We then selected samples from these candidates for which all points were in the presentation range $[0,1]$, that were ≥ 0.05 of the three transfer points, and rejected uncharacteristic samples[1]. For a full list of kernels and kernel parametrizations, see Table 1, for the training data and the conditional samples, see Figure 1.

Forced Choices In the transfer set three points, $x = \{0.05, 0.1, 0.2\}$, $y = \{0.475, 0.525, 0.5\}$ were. These points were selected to be inconsistent with any of the training materials, in terms of specific point locations. We then generated three samples conditional on the transfer points for each of the six functions. Participants received one of these three samples at random for each of the six kernels in the forced choice task.

[1]For example, *OU* samples that did not exhibit any discontinuities and thus looked visually identical to linear relationships, or *cos* samples that had very low amplitudes.

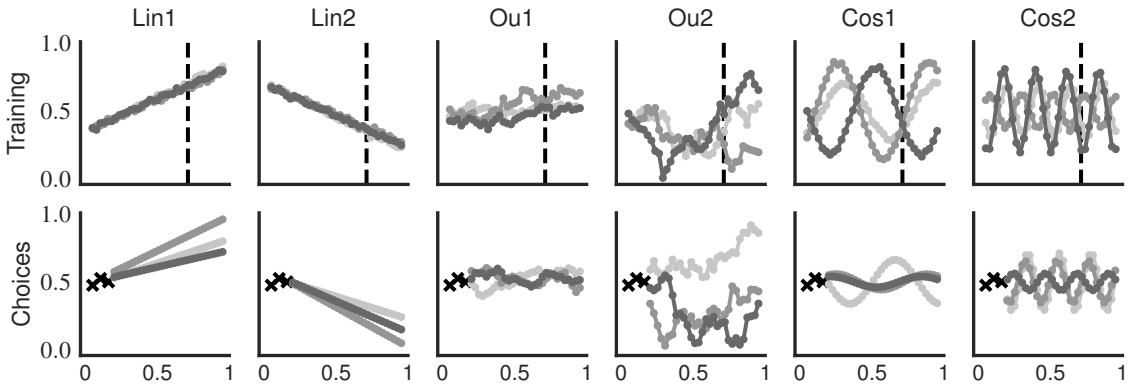


Figure 1: Top row: Training data in the six conditions. For each condition, there were three data sets to be learned. Participants received the first 25 points and had to extrapolate for the 10 remaining points, dashed is the cutoff between presented evidence and training. Bottom row: Samples that constituted the forced-choice options in Experiment 1 & 2. Samples were generated by the particular GP, conditional on the three points in the transfer set.

Table 1: Kernels and kernel parameters generating the training data. For all models we set $\sigma_{noise} = 0.01$.

Kernel	variance	lengthscale	β_0	β_1
Lin_1	0.02	–	0.35	0.47
Lin_2	0.02	–	0.7	-0.47
Cos_1	0.05	0.1	0.5	–
Cos_2	0.05	0.04	0.5	–
OU_1	0.01	1	0.5	–
OU_2	0.08	1	0.5	–

Results

Error Rates

The training functions differed considerably in their mean absolute errors (MAEs)[2], as well in the change of error over blocks, see Table 2.

Only for conditions with linear functions did errors differ significantly between the first and the last block, $t(42.94) = 2.21$, $p = .032$[3]. For OU conditions, errors were lower for the last block, but did not differ significantly, $t(62.21) = 0.95, p = 0.345$. For periodic conditions, error was again lower for the last block, but blocks did not differ significantly $t(57.17) = 0.66$, $p = .509$. The two OU and periodic conditions were highly heterogeneous. While errors for the low variance condition Ou_1 decreased over blocks, errors in Ou_2 remained high. Equally, while errors in Cos_1 decreased, errors for Cos_2 remained high throughout training. For error rates for all conditions, see Figure 2.

[2]All MAEs were calculated on extrapolations before the participant had received feedback for that particular value.

[3]All tests are two-sided, unequal variance t-tests. For means and SDs, see Table 2

Table 2: MAEs for functions and blocks in Experiment 1

	MAE_{b1}	SD_{b1}	MAE_{b2}	SD_{b2}	MAE_{b3}	SD_{b3}
Lin	.02	.01	.02	> 0.01	.02	> 0.01
OU	.05	.02	.06	.03	.05	.03
Cos	.06	.05	.06	.06	.05	.04

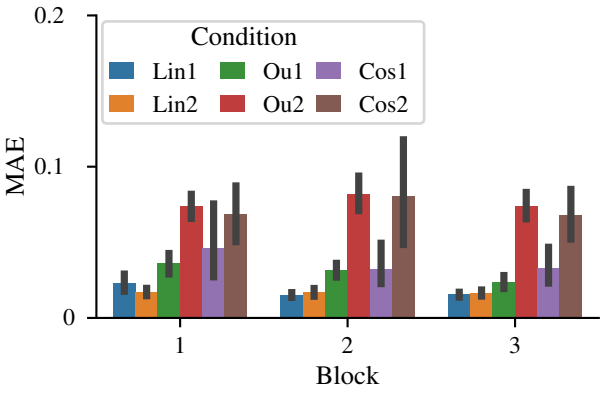


Figure 2: MAEs and 95% confidence intervals for each condition in Experiment 1.

Selecting an Extrapolation Pattern

About 35% of the participants selected the choice corresponding to the correct function type and parametrization. Both for positive and negative linear training conditions, the proportion of chosen true functions was significantly larger than chance (1/6), Lin_1 = 44%, $p = .01$, Lin_2 = 53%, $p < .001$[4]. For periodic functions, Cos_2 was selected significantly above chance, Cos_2= 50%, $p = .002$, but $Cos_1$was not, Cos_1 = 12%, $p = .802$. Instead, participants mostly selected the

[4]All tests are one-sided, exact Binomial tests.

other periodic function. The proportion of generally periodic functions over the alternatives for condition Cos_1 was significantly above chance $(1/3)$, $Cos_1 = 59\%$, $p < .027$.

For OU conditions, Ou_1 was not selected significantly above chance, $Ou_1 = 18\%$, $p = .556$, nor did participants prefer OU functions in general, $Ou_1 = 47\%$, $p = .172$.

However, participants trained on Ou_2 selected the true pattern at rates significantly higher than chance $Ou_2 = 38\%$, $p = .05$. For all participant choices, see Figure 5.

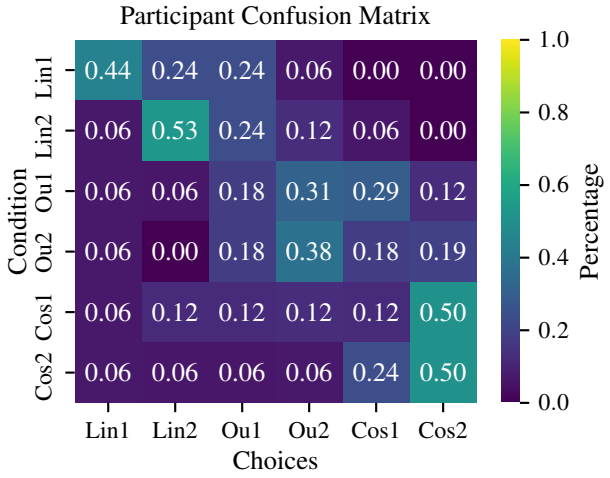


Figure 3: Confusion matrix for choices in Experiment 1

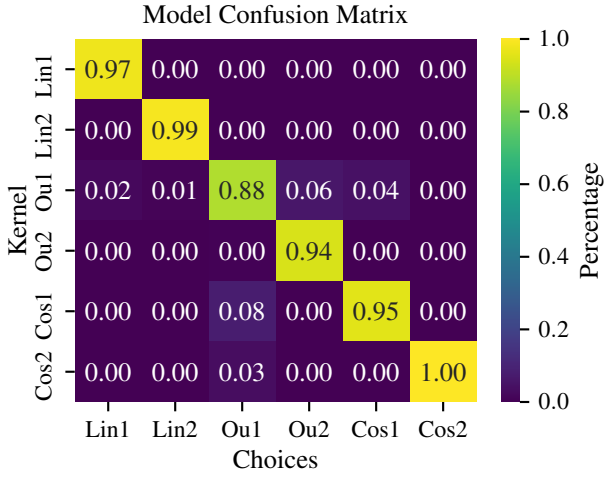


Figure 4: Confusion matrix for the generating functions and the choices presented in Experiment 1.

We also calculated the confusion matrix resulting from the likelihood assigned to each of the presented samples for each of the functions generating the materials. We then converted these likelihoods into proportions via the softmax function[5].

[5]We also evaluated these proportions for a model with an additional temperature parameter T that we fitted to the human choices. Unsurprisingly, this temperature parameter was estimated to be low, $T \approx 10$, and produce less peaked distributions.

While the model consistently favors the true generating function and does not produce preferences resembling the participants' choices, some interesting parallels are apparent. First, both kernels resulting in the strongest preference for the true function (Cos_2 and Lin_2) also correspond to conditions with fairly peaked human preferences. In contrast, kernels resulting in more dispersed likelihoods, and as a result, lower preference for the true function (Ou_1, Ou_2) resemble the systematic preferences for alternative functions by human participants. For a confusion matrix displaying the asymmetric choices of participants, see Figure 3, for model confusions, see Figure 4.

Experiment 2

In Experiment 1, participants were able to select from a set of candidates realizations corresponding to the learned type of function and, in many cases, the specific features of the set of training examples. In this control experiment, we contrasted participants' choices in Experiment 1 with a condition in which no training was provided.

Participants

We recruited 50 participants ($M_{age} = 34.7$, $SD_{age} = 10.53$, 25 female, 24 male, 1 other) on Amazon Mechanical Turk. Participants received \$0.2 for participation and took an average of 1.5 minutes, ($M = 1.46$, $SD = 6.05$) to complete the experiment.

Procedure

Participants were instructed that they would be presented with a relationship between two substances, consisting of three pairs of values. Then they were instructed that they would have to select a pattern from six options that most likely depicted the relationship. The choices were the same as in Experiment 1.

Results

In the absence of training data participants preferred periodic functions over OU and linear, $Lin_1 = 10\%$, $Lin_2 = 0\%$, $Ou_1 = 18\%$, $Ou_2 = 14\%$, $Cos_1 = 28\%$, $Cos_2 = 30\%$, see also Figure 1. Given the low rates of choices of Lin_1 and the high proportion of chosen periodic functions, these results suggest that participants interpreted the three points presented as generated from a deterministic, non-monotonic relationship, rather than a low-noise linear or low-variance OU relationship.

These results suggest that the strong preference for periodic samples in Experiment 1 did not solely result from the training but were also reflective of a higher preference to ascribe periodicity to the test points.

Experiment 3

We have shown that participants can use the knowledge acquired in the training sets, to inform their choices about which

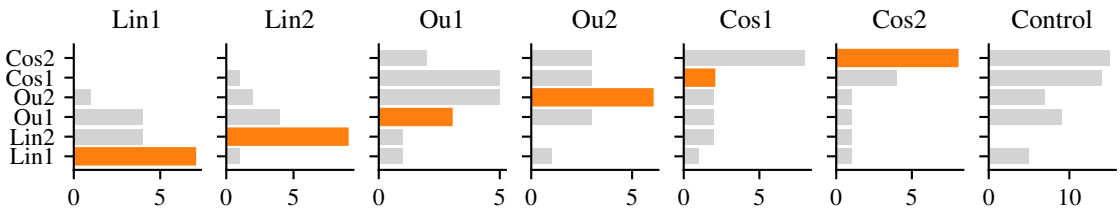


Figure 5: The selected choices (correct choice in orange) for Experiment 1 and Experiment 2 (Control). Participants generally selected the function type and parametrization consistent with their training. When instead they preferred other options they mostly selected samples from the same type of function. Without training, participants favored periodic and OU.

particular type of function generated the data. However, it is possible that these choices do not reflect participant's true belief about the underlying functions, but are merely best guesses given a set of unsatisfactory options. In this last experiment, we will analyze if these choices correspond to actual extrapolation behavior.

Participants

We recruited 91 participants ($M_{age} = 30.53$, $SD_{age} = 6.941$, 34 female, 57 male) on Amazon Mechanical Turk. Participants received \$0.65 for participation and took an average of 10 minutes (M=10.29, SD=10.56) to complete the experiment. Participants were randomly assigned to one of the 6 conditions ($n_{Cos1} = n_{Cos2} = n_{Lin1} = 15$, $n_{Lin2} = n_{OU1} = 16$, $n_{OU2} = 14$).

Procedure & Materials

Instructions and training were identical to Experiment 1. However, instead of the forced-choice task participants performed an extrapolation task. In the extrapolation task, participants received the same three points that generated the conditional samples in experiment 1 and had to extrapolate for 30 values of x, without feedback, following the same procedure as in the training sets. The 30 extrapolation criteria were the same as the ones used to generate the forced-choice patterns in Experiment 1.

Results

Error Rates

As in Experiment 1, conditions differed considerably in their MAEs, as well in the decrease in error, depending on the particular function, see Table 3, for errors, see Figure 6. In contrast to Experiment 1, errors in conditions with linear functions did not differ significantly between the first and the last block $t(56.04) = 0.81$, $p = .423$. Neither did errors differ significantly in OU conditions, $t(50.13) = 0.37$, $p = .716$. However, errors for periodic conditions differed significantly between the first and the last block, $t(42.62) = 2.38$, $p = .022$. As in Experiment 1 most conditions exhibited very low errors. In contrast Cos_2 and Ou_2 were characterized by large MAEs. For error rates for all conditions, see Figure 6.

Table 3: MAEs in Experiment 3.

Function	MAE_{b1}	SD_{b1}	MAE_{b2}	SD_{b2}	MAE_{b3}	SD_{b3}
Lin	0.03	0.02	0.02	0.02	0.02	0.02
OU	0.06	0.03	0.05	0.04	0.05	0.05
Cos	0.09	0.08	0.06	0.04	0.05	0.04

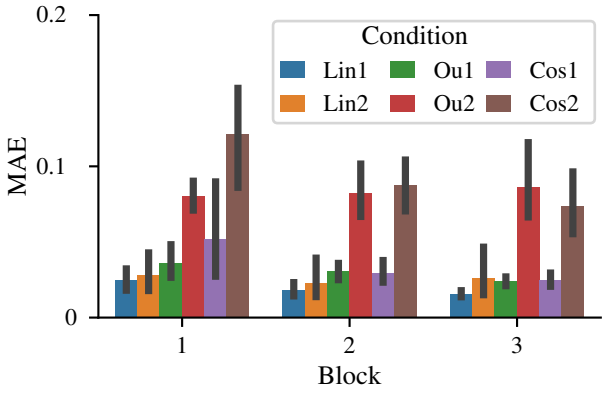


Figure 6: MAEs and 95% confidence intervals for each condition in Experiment 3.

Extrapolating

Visual inspection of the extrapolation strongly suggested that variances between OU-, frequencies for periodic- and slopes for linear conditions reflected training functions, see Figure 1. To evaluate if these patterns were also well aligned with the generating models, and if samples reflected the differences in function parametrization, we performed maximum-likelihood estimation (MLE) for each individual participant and each generating GP. We then used the type of the generating GP with the highest likelihood to predict which training samples the participant had been assigned to. This approach allowed us to evaluate if the experimental manipulation resulted in extrapolation patterns consistent with the generating GPs. Our classification procedure classified 22 out of 30 participants in the OU conditions correctly, a proportion that was significantly larger than expected by chance (1/3),

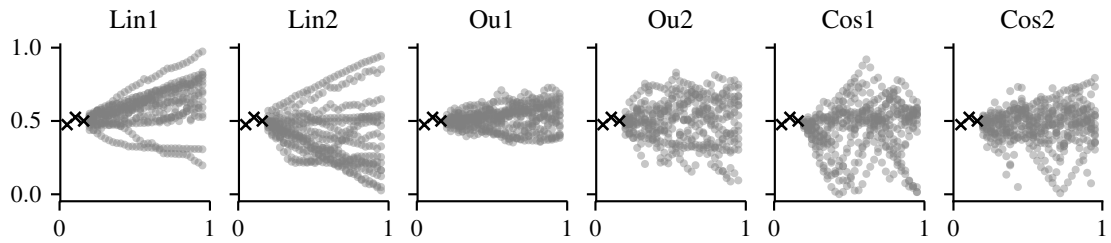


Figure 7: Participant extrapolations in Experiment 3. Extrapolations closely matched the learned type of function and its detailed parametrization, see Figure 1

$p_{OU} < .001$[6]. In the periodic conditions, 17 out of 30 participants were classified correctly, $p_{Cos} = .007$. However, for linear samples, only 10 out of 31 participants were classified correctly, $p_{Lin} = .617$. For the full confusion matrix, see Figure 8.

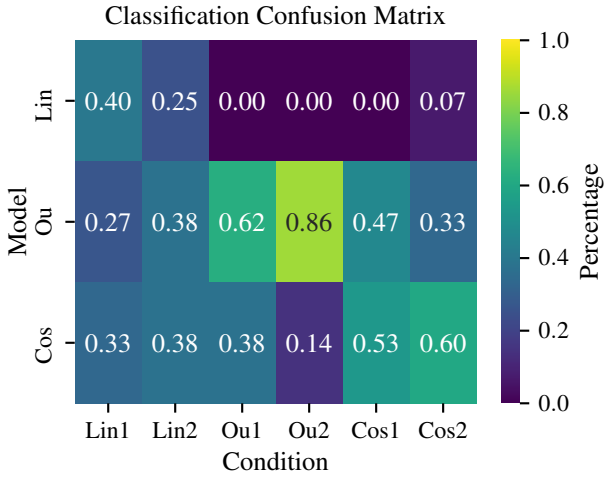


Figure 8: Confusion matrix for the proportion of participants assigned to each model. Our method was able to accurately recover which type of function participants were trained on for OU and periodic, but not linear conditions.

To compare the inferred parametrizations across training conditions of one function type, we contrasted the parameters obtained via MLE for the true model. For linear functions, the MLE parameters for slopes differed significantly between conditions, $M_{Lin1} = 0.2$, $SD_{Lin1} = 0.25$, $M_{Lin2} = -0.1$, $SD_{Lin2} = 0.34$, $t(27.49) = 2.82$, $p = .009$[7], with the signs of the inferred slopes matching the training. Neither intercept, variance or noise estimates differed significantly between conditions (all $p > .1$). The inferred parameters for variance in the OU conditions did not differ significantly, but were reflective of differences in training, $M_{OU1} = 0.002$, $SD = 0.002$, $M_{OU2} = 0.007$, $SD_{OU2} = 0.008$, $t(15) = -1.95$, $p = .071$. The inferred length scale did not differ significantly between conditions, but was slightly higher for $OU1$, $M_{OU1} = 0.38$, $SD_{OU1} = 0.39$, $M_{OU2} = 0.21$, $SD_{OU2} = 0.28$, $t(26.31) = 1.46$, $p = .157$. Both intercept and noise estimates did not differ significantly between conditions (all $p > .5$). The inferred parameters for periodic conditions did not differ significantly for length scale, $M_{Cos1} = 0.08$, $SD_{Cos1} = 0.06$, $M_{Cos2} = 0.08$, $SD_{Cos2} = 0.1$, $t(22.47) = 0.27$, $p = .79$. Instead, conditions differed significantly for variance $M_{Cos1} = 0.02$, $SD_{Cos1} = 0.02$, $M_{Cos2} = 0.01$, $SD_{Cos2} = 0.01$, $t(20.1) = 2.25$, $p = .036$. Estimates for intercepts and noise were not significantly different between conditions (all $p > 0.1$).

Discussion

We found evidence that participants choose patterns and extrapolate in ways consistent with the learned function type. Furthermore, contrasting the extrapolations in the transfer set within function conditions, suggested that these patterns differed in ways consistent with our experimental manipulation.

While participants' judgments generally reflected the functions they learned during training, our results also highlight characteristic human biases. In the Cos_1 condition, participants preferred high-frequency periodic samples over the true low-frequency samples. Similarly, participants in the Ou_1 conditions, preferred the higher variance samples, or even periodic samples over the trained low-variance samples. One explanation for this biases could be that people have a strong preference for particular functions because these parametrizations are well adapted to environmental regularities. As a result, these functions would be robust and applicable to a wide range of task in the environment. This explanation would be consistent with recent results in human exploration, where participants exhibited a tendency to undergeneralize spatial correlations, but this undergeneralization resulted in comparable or even better performance than a ground-truth matching model (Wu et al., 2018).

To better describe these characteristic human biases and explore their potential rational grounding, future research should more closely examine which statistical patterns can be generalized and under which circumstances these generalizations are performed. For example, while our experiment imposed that all patterns followed the same relationship, in

[6] All tests are one-sided, exact Binomial tests.
[7] All tests in this section are unequal variance, two-sided t-tests.

reality this information is rarely available. Thus, future research should examine under which circumstances task regularities are inferred to be similar, and what kinds of notions of similarity can guide these generalizations. One exciting prospect is to link notions of hierarchical- and compositional representations and function generalization. We are currently exploring how such compositional regularities can aid transfer and generalization in sparse domains.

Acknowledgements

We thank the three anonymous reviewers for their helpful comments and suggestions.

References

Bott, L., & Heit, E. (2004). Nonmonotonic extrapolation in function learning. *Journal of Experimental Psychology*, *30*(1), 38.

Brehmer, B. (1976). Learning complex rules in probabilistic inference tasks. *Scandinavian Journal of Psychology*, *17*(1), 309–312.

Busemeyer, J. R., Byun, E., Delosh, E. L., & McDaniel, M. A. (1997). Learning functional relations based on experience with input-output pairs by humans and artificial neural networks. In K. Lamberts & D. R. Shanks (Eds.), *Studies in cognition. knowledge, concepts and categories* (p. 408-437). The MIT Press.

Carroll, J. D. (1963). Functional learning: The learning of continuous functional mappings relating stimulus and response continua. *ETS Research Bulletin Series*, *1963*(2), i–144.

DeLosh, E. L., Busemeyer, J. R., & McDaniel, M. A. (1997). Extrapolation: The sine qua non for abstraction in function learning. *Journal of Experimental Psychology*, *23*(4), 968.

Gershman, S. J. (2017). On the blessing of abstraction. *The Quarterly Journal of Experimental Psychology*, *70*(3), 361–365.

Gershman, S. J., & Niv, Y. (2010). Learning latent structure: carving nature at its joints. *Current opinion in neurobiology*, *20*(2), 251–256.

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive science*, *32*(1), 108–154.

Hamrick, J. B., Ballard, A. J., Pascanu, R., Vinyals, O., Heess, N., & Battaglia, P. W. (2017). Metacontrol for adaptive imagination-based optimization. *arXiv preprint arXiv:1705.02670*.

Harlow, H. F. (1949). The formation of learning sets. *Psychological review*, *56*(1), 51.

Kalish, M. L. (2013). Learning and extrapolating a periodic function. *Memory & Cognition*, *41*(6), 886–896.

Kalish, M. L., Lewandowsky, S., & Kruschke, J. K. (2004). Population of linear experts: knowledge partitioning and function learning. *Psychological Review*, *111*(4), 1072.

León-Villagrá, P., Preda, I., & Lucas, C. G. (2018). Data availability and function extrapolation. In *Proceedings of the Annual Meeting of the Cognitive Science Society.*

Lucas, C. G., Griffiths, T. L., Williams, J. J., & Kalish, M. L. (2015). A rational model of function learning. *Psychonomic Bulletin & Review*, *22*(5), 1193–1215.

Lucas, C. G., Sterling, D., & Kemp, C. (2012). Superspace extrapolation reveals inductive biases in function learning. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 34).

McDaniel, M. A., & Busemeyer, J. R. (2005). The conceptual basis of function learning and extrapolation: Comparison of rule-based and associative-based models. *Psychonomic Bulletin & Review*, *12*(1), 24–42.

Schulz, E., Tenenbaum, J. B., Duvenaud, D., Speekenbrink, M., & Gershman, S. J. (2017). Compositional inductive biases in function learning. *Cognitive Psychology*, *99*, 44–79.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*(6022), 1279–1285.

Wilson, A. G., Dann, C., Lucas, C. G., & Xing, E. P. (2015). The human kernel. In *Advances in Neural Information Processing Systems* (pp. 2854–2862).

Wilson, R. C., & Niv, Y. (2012). Inferring relevance in a changing world. *Frontiers in human Neuroscience*, *5*, 189.

Wu, C. M., Schulz, E., Speekenbrink, M., Nelson, J. D., & Meder, B. (2018). Generalization guides human exploration in vast decision spaces. *Nature Human Behaviour*, *2*(12), 915.

When Sleep-Dependent Gist Extraction Goes Awry: False Composite Memories are Facilitated by Slow Wave Sleep

Itamar Lerner (itamar.lerner@rutgers.edu)
Tony P. Kerbaj (tonykerbaj@gmail.com)
Mark A. Gluck (gluck@pavlov.rutgers.edu)

Center for Molecular and Behavioral Neuroscience, Rutgers University – Newark
Newark, New Jersey 07102 USA

Abstract

Contemporary evidence suggests that sleep contributes to the extraction of gist from previously encoded experiences, a process that relies on compressed memory replay. While the functional significance of the time compression is not fully understood, a recent 'temporal scaffolding' model suggested that compression allows associating encoded events that happened in disparate times, a critical feature when extracting gist of a temporal nature. We examined this hypothesis using a novel behavioral paradigm. Subjects were first presented with word pairs that could form a new composite word if combined (e.g., car, pet --> carpet), and then tested on whether they falsely recognize seeing the composite word. When subjects napped in between exposure and testing, false memories of composite words increased, with reaction times for false recognition correlating to time spent in slow wave sleep. These results confirm the functional role of time compression in memory replay, supporting the temporal scaffolding model.

Keywords: Sleep; Memory Replay; Gist Extraction; False Memories; Temporal Scaffolding

Introduction

Numerous studies over the last two decades support the notion that sleep facilitates memory consolidation (Rasch & Born, 2013). There is now compelling evidence from human and rodent studies that during one particular sleep stage, Slow Wave Sleep (SWS), recently encoded memories are replayed in the hippocampus as part of a hippocampal-cortical dialogue (Wilson & McNaughton, 1994; Diba & Buzsaki, 2007). Theoretical models suggest this replay may contribute to the strengthening of common features within those memories while eroding their idiosyncratic elements, effectively leading to the extraction of "gist" and their integration within general knowledge structure in the cortex (McCelland, McNaughton, & O'reilly, 1995; Lewis & Durrant, 2011).

Perhaps the most striking example of gist extraction is exemplified by demonstrations that SWS can support insightful discovery of hidden rules. In these studies, subjects were presented with a sequence of stimuli and asked to respond to each stimulus as quickly and accurately as possible by following a simple rule (Fischer, Drosopoulos, Tsen, & Born, 2006; Wagner, Gais, Haider, Verleger, & Born, 2004). Unknown to subjects, a hidden temporal structure governed the series of presentations such that, if discovered, it could improve performance significantly. Following sleep, subjects were more likely to discover the hidden rule and improve performance compared to subjects that stayed awake, an effect that was correlated with the time spent in SWS (Wilhelm, Rose, Imhof, Rasch, Büchel, & Born, 2013; Yordanova, Kolev, Verleger, Bataghva, Born, & Wagner, 2012). While sleep-dependent discovery of hidden rules fits the general theory of gist extraction during sleep, the particular mechanism, and its relation to SWS, remain unclear. Recently, a 'temporal scaffolding' model was proposed to account for the effects of sleep on insightful processes (Lerner 2017a, 2017b; Lerner et al., 2019). The model suggests a key property of memory replay that allows for these effects to emerge: its time-compressed nature. In particular, hippocampal memory replay is known to occur in an accelerated form, up to twenty times the speed of the original experience (at least in rodents; Rasch & Born, 2013). When encoded sequences of events are reactivated in this accelerated manner, Hebbian learning mechanisms can associate events that were otherwise too temporally distant from each other to fall within the typical neural learning timescale (50-200ms for Hebbian mechanisms; August & Levy, 1999). Consequently, discovery of hidden rules that relies on the detection of temporal structure within sequential stimuli should, according to the model, be particularly prone to facilitation by SWS.

One surprising prediction of this model is that temporal associations resulting from time-compressed replay during sleep might also hurt memory, not just facilitate it. If two distinct events are replayed in a compressed timescale one after the other during SWS, this may lead to their assimilation into one single memory following the consolidation process, even if such assimilation is unwarranted. In particular, such phenomena might occur if the two events have a special meaning when compiled together, thus signaling to the cortex to maintain the combined meaning rather than the separated memories (a gist extraction of sorts, albeit one that occurs under the wrong circumstances). An example of this theoretical

process can be demonstrated by presenting a subject with two consecutive words, such as *car* and *pet*, which could be combined into a composite (or compound) word: *carpet*. Due to the temporal scaffolding mechanism, the proximal but distinct events of seeing *car* and *pet* might be integrated following sleep to become a false memory of seeing *carpet*.

In the current study, we tested this hypothesis by exploring how an afternoon nap affects the probability of falsely recognizing composite words whose components were previously encountered, as if they were actual memories. Since memory replay during SWS is known to occur predominantly in a forward manner (i.e., replay of encoded events proceeds in the same order as the original experience, albeit in accelerated form; Diba & Buzsaki, 2007), we predicted that sleep would facilitate false memories of composite words whose components were presented sequentially in the forward direction (e.g., car -> pet), but not of those presented backwards (pet->car), or when the components were presented in totally separate trials. Confirming that SWS facilitates the formation of such false memories substantially supports the idea that accelerated forward replay plays a part in gist extraction

Methods

Participants

Forty young adults (ages 18-24, n=19 females) from Rutgers University and the New Jersey Institute of Technology participated in this study for monetary compensation. Subjects were recruited via protocol flyers, in-class announcements and on-campus active recruitment. All subjects were screened for exclusion criteria, which included personal or family history of sleep problems, neurological or psychiatric disorders, drug or alcohol abuse, and/or intake of medications that have any effect on sleep. Furthermore, all recruited subjects had normal or corrected vision/hearing and were fluent in English. Subjects were also asked not to increase daily caffeine and to abstain from caffeine and alcohol before testing. All participants provided informed consent in line with the procedures approved by the Institutional Review Board of Rutgers University.

Sleep Monitoring

We recorded sleep using the Zmachine ® Insight device (Model DT-200; General Sleep Corporation), a sleep monitoring apparatus designed for use in clinical and home environments, and has been shown to reliably detect sleep stages at a level comparable to Polysomnography (Wang et al., 2016). It consists of three self-applicable, single-use, disposable electroencephalography (EEG) sensors, two located on the mastoids (signal electrodes) and one on the back of the neck (ground electrode). The machine detects and records three sleep stages, in addition to wake stage: light sleep (combined Stages N1 and N2), SWS, and Rapid Eye Movement (REM) sleep for each 30-second epoch of sleep. Following the completion of each subject testing, the collected data was transferred from the device's micro SD card to a secure desktop computer for further analysis.

Behavioral Task

Stimuli We compiled three groups of word pairs, 6 pairs per group, such that the words of each pair, if combined together, create a "composite" word (e.g., *car, pet* --> *carpet*; *under, stand* --> *understand*). Words of each pair were selected such that they were not semantically related to each other, nor were they related to the composite word they create together. In addition, we compiled a group of 32 non-composite words. The average length and frequency of the composite words (i.e., the combination of the two components together) in each of the three groups, as well as each single word in the non-composite group, was roughly equal, with M=~7.5 letters and M=~18,000 occurrences for length and frequency, respectively (Frequency data was based on the database found in: https://corpus.byu,.du/coca/)

Based on these four groups, two word-pair lists were created for the "exposure" phase of the experiment. The first exposure list was comprised of the following: (1) 'Forward' composite items: the word pairs of the first composite group appearing in the order that corresponds to the composite word (e.g., *car, pet*); (2) 'Backward composite items: the word pairs of the second composite group appearing in the reverse order to the one corresponding to the composite word (e.g., *stand, under*); (3) 'Separate composite items: each of the two words of the third composite group paired with random words from the non-composite group (e.g., *honey, moon*, forming the composite word *honeymoon*, were paired with *pharmacy, sad*, to create the pairs *pharmacy, honey* and *moon, sad*); (4) the remainder of the words from the non-composite group, randomly paired. The total number of items (pairs) in the list was 34, and their order within the list was pseudo-randomized with the restriction that items containing words that belonged to the same word-pair of the Separate composite group would not appear sequentially. The second exposure list was identical to the first, except that the Forward and Backward composite items were switched such that the first group composed the backward items and the second group composed the forward items. The order of the items within the list was switched as well, such that the location of the forward and backward pairs was similar in the two lists.

We next created two testing lists, matching the two exposure lists. The first testing list contained all 18 composite words made of the Forward, Backward and Separate composite items, as well as 24 additional non-composite words from the exposure list, and 6 totally new, non-composite words (48 items in total, half of which are old). The totally new words were chosen such that the average length and frequency of the new and old words across the testing list remained roughly equal. The order of these words within the list was chosen pseudo-randomly. The second testing list was identical to the first, except that the location of the forward and backward composite words was switched to match the first testing list.

Composite Word task The behavioral task included an exposure session and a testing session, separated by an intermission during which subjects were either allowed to sleep or remained awake (see Figure 1). The objective of the exposure session was to allow subjects to encode the components of the composite words consecutively, without driving their attention to their composite nature (by using a distracting task). The testing session included a surprise memory test, where subjects' tendency to incorrectly recognize the composite words as words they have been exposed to earlier was examined.

Figure 1: The behavioral task used in the study. During an exposure session, subjects saw two colored words in succession and were asked to indicate whether the words appeared in the same or different color. Unknown to subjects, some of those word pairs could be concatenated to create a third, unrelated word. Following an intermission during which some of the subjects took a 90 minute nap and some remained awake, they received a surprise memory test requiring to indicate whether a series of presented words are new or appeared in the earlier exposure session. Some of those words were the composite words whose components were previously displayed.

Exposure Session In each trial of the exposure session, subjects were presented with two consecutive words. Each of the words appeared in one of 3 colors: red, green, or blue. Subjects were required to indicate whether the two words appeared in the same or different colors by pressing one of two buttons on the keyboard. The two words presented in each trial were taken from the items in the exposure list, with half of the subjects receiving the first list and the other half – the second list. To facilitate the probability that Forward and Backward composite items will be combined in memory during sleep, word pairs belonging to these two conditions were always presented in the same color. Other word pairs were presented in either the same or different colors, and the total number of 'same' and 'different' trials was counterbalanced across the session. Subjects were not informed that some of the word pairs could construct a composite word if combined.

Testing Session During the testing session, subjects were presented with single words appearing on the screen one at a time. After each word presentation, subjects were required to indicate whether they recognize seeing this word in the first session or not, by pressing one of two buttons on the keyboard. These words could either be old words appearing in the first session, composite words whose components appeared as single words in the first session, or totally novel words. Subjects that received the 1st exposure list also received the 1st testing list, and subjects receiving the 2nd exposure list also received the 2nd testing list. Following the testing session, subjects were administered a post-experimental questionnaire, designed to determine if they explicitly recognized the existence of composite words in either of the sessions. The questionnaire was designed as a series of questions of escalating details, which avoided revealing the hidden structure of the task unless subjects came up with it by themselves. Three subjects who explicitly recognized the presence of composite words during the exposure session were removed from the study.

Procedure Subjects first arrived to the lab to collect the sleep-monitoring device and were given detailed instructions on how to use it. They then monitored their sleep at home for two nights to allow them to adapt to sleeping with the device on their scalp, and to allow the sleep stage detection algorithm of the device to accommodate to the subjects' individual EEG patterns. After two nights, subjects returned to the lab at the afternoon to begin the experiment, which included the 2 sessions of behavioral measurements, exposure and testing, separated by a 120-minute intermission. The experiment was ran in a quiet room using a MacBook Air (v.2014) laptop, with subjects situated in a convenient distance of 30cm from the screen. Subjects first received detailed instructions on screen regarding the task. Each trial of the exposure session began with the presentation of small white fixation cue appearing on a black screen for 500ms. The screen then remained black for 1500ms until the presentation of the first word for 500ms. After an Inter Stimulus Interval of 100ms, the second word appeared for 500ms, followed by a black screen that remained until the subject's response. Following the response, the next trial initiated. Five practice trials preceded the exposure, using different word pairs. Practice trials were similar to the exposure trials, with the exception that subjects received feedback immediately after responding (a smiley face for a correct response and a sad face for an erroneous response), which replaced the fixation cue. Following the exposure session, subjects put on the sleep monitoring device and went into the intermission period during which they were allowed to take a nap for 90 minutes in a designated sleep testing room (Sleep group; N = 19) or watched a non-stimulating movie in the same testing room (Wake group; N = 18). Following the intermission (which lasted 2 hours for both groups, to allow half an hour of wake time for the Sleep group to eliminate sleep inertia), subjects underwent the testing session. Subjects received instructions on screen regarding the memory recognition test before starting the task. Each testing trial consisted of a word appearing on the screen in white, until the subject's response. After responding, the screen remained black for 1000ms, after which the next word appeared, and so on until the end of testing.

Data Analysis For each subject, we assessed the performance of each of the four critical experimental conditions (Forward composite items, Backward composite items, Separate composite items, Novel items) using two behavioral measures, Error Rate and Normalized Error Reaction Time (RT). Error rates were defined as the total number of erroneous responses in each condition, divided by the number of trials in that condition (an error was defined as responding "Old"). Normalized Error RTs were defined as the mean RTs for wrongly identified items divided by the total mean RT, for each condition (calculated after removal of outlier RTs, defined as values above or below 3 standard deviations from an individual's mean, across conditions). We used the normalized RT measure rather than raw RTs because pilot data collected prior to the experiment suggested that between-subject individual differences in RTs were substantially higher than within-subject differences in this task, potentially blurring the effects of interest. We expected that the more false memories an individual has, the higher will the error rate and the lower will the Normalized Error RT be (based on a common interpretation of RTs as indicating confidence in the responses; Wiedemann & Kahana, 2016). We compared these two measures between the Sleep and Wake groups, and within the groups themselves, using Bonferroni-corrected independent and paired t-tests, respectively. In addition, for the Sleep group, we also correlated these measures across subjects with the individual time spent in sleep, and in each sleep stage, during the nap (as well as the percent of time spent in each sleep stage out of total sleep time).

Results

Mean error rate values for each condition and subject group are presented in Figure 2.

Bonferroni-corrected t-tests showed that error rates in recognizing Forward composite items as "Old" were significantly higher for the Sleep group compared to the Wake group ($t(35) = 2.61$, $p < 0.05$). No other condition showed a difference between the groups. Within the Wake group, Bonferroni-corrected pairwise comparisons showed that error rates for the Backward composite items were significantly higher than those of the Separate composite items ($t(17) = 4.19$, $p < 0.004$), as well as higher, on a trend level, than those of the Forward composite and Novel items ($t(17) = 2.91$, $p < 0.06$, and $t(17) = 2.87$, $p < 0.07$. respectively). Within the Sleep group, in contrast, error rates for the Forward and Backward composite items were significantly higher than those of the Separate composite and Novel items (all ps < 0.03), but there was no difference between the Forward and Backward composite items ($p = 0.47$). Repeating the same analysis with Normalized Error RTs, we found no significant effects between or within the groups.

We also compared the error rates of the Sleep and Wake group in the Old words condition (i.e., non-composite words that appeared during the exposure session and for which the correct answer was "Old" and an error response was "New"). There was no difference between the groups in this condition (M = 50.9 and M = 54.2 for the sleep and Wake group, respectively; $p = 0.526$).

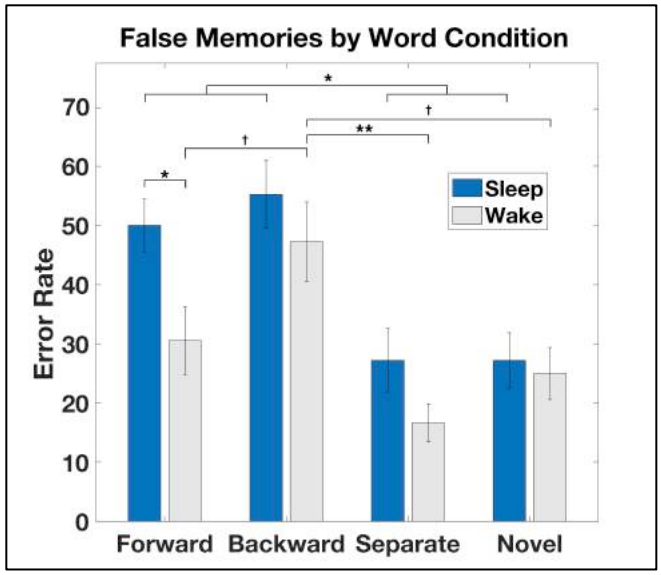


Figure 2: Error rate by word condition for the sleep and wake groups. Subjects who slept exhibited more false memories (higher error rate) for composite words that were presented in the forward direction during training compared to the wake group. ** $p<0.005$; * $p<0.05$; † $p<0.07$. Error bars represent standard error of the mean.

We next examined whether the performance measures were influenced by any of the recorded sleep parameters for subjects in the sleep group (see sleep statistics in Table 1). First, we computed the Pearson correlations between the total time subjects spent in sleep and each of the two performance measures in each of the four experimental conditions (8 comparisons in total). We found a significant correlation of total sleep time with the Normalized Error RT of Forward composite items ($r = -0.6717$, $p = 0.0023$; $p = 0.018$ after correcting for 8 multiple comparisons). No other correlation was significant.

Table 1: Recorded sleep statistics. TST = Total Sleep Time.

Sleep Measure	Mean (std)
TST (minutes)	40.41 (22.3)
N1/N2 (minutes)	23.97 (13.0)
% N1/N2 out of TST	66.46 (23.7)
SWS (minutes)	11.32 (13.1)
% SWS out of TST	19.59 (21.6)
REM (minutes)	5.10 (5.8)
% REM out of TST	13.92 (18.1)

Next, to investigate the contribution of particular sleep stages, a multiple regression analysis was carried out for each condition, with the performance measure of interest as

the dependent variable and time in each recorded sleep stage (N1/N2, SWS, REM) as predictors. A significant regression was found, once again, for Normalized Error RT of Forward composite items ($F(3,14) = 8.11$, $p = 0.0022$; $p = 0.0178$ after correcting for 8 multiple comparisons) with $R^2 = 0.6348$. Normalized Error RTs were equal to 1.2208 – 0.0068 (N1/N2) – 0.0095 (SWS) + 0.0093 (REM), with SWS contributing significantly to the model ($p = 0.0078$). The more SWS subjects had, the faster was their erroneous response in identifying Forward composite items as "Old" (Figure 3, inset). This effect remained highly significant in a follow-up analysis, computing the Pearson correlation between Normalized Error RTs of Forward composite items and the percent of time spent in SWS out of total sleep time ($r(17) = -0.647$, $p < 0.004$; Figure 3, main). No other effects were significant in the multiple regression analyses.

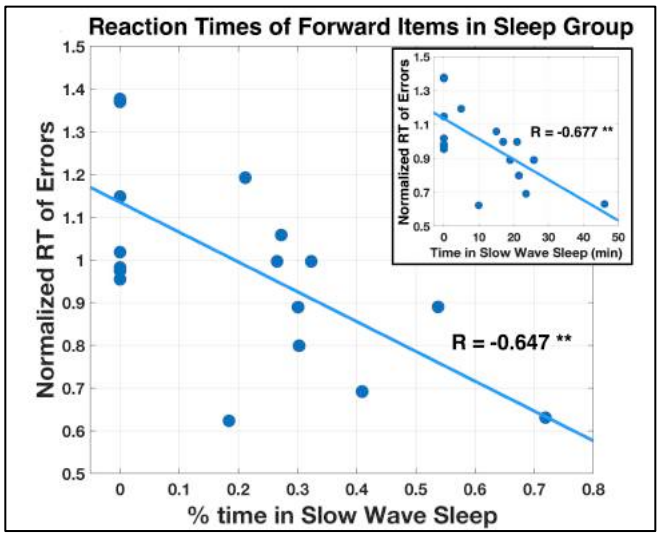


Figure 3: Normalized reaction times of error responses for Forward composite items as a function of minutes spent in slow wave sleep (inset) and percent of time spent in slow wave sleep out of total sleep time (main), for subjects in the sleep group. ** $p<0.008$.

Discussion

We found that subjects who take a nap following exposure to components of a composite word are more likely to falsely recognize being presented with that composite word compared to subjects who did not nap. Moreover, the more time subjects spent in SWS during the nap, the quicker it took them to make that error, likely indicating a higher confidence in their response (Wiedemann & Kahana, 2016). These effects, however, were apparent only if subjects were exposed to the component words one after the other, and only if they were presented in the order that matches their appearance in the composite word, but not if they were presented in the reverse order.

Our findings are consistent with the prediction of a recent "temporal scaffolding" model of memory consolidation during SWS, which emphasizes the role of time compression in memory replay (Lerner et al., 2017a). Specifically, the model predicts that compressed replay of a recently encoded sequential experience may lead to elements within this experience to bind together and create a unified memory that no longer preserves the original sequential nature of the experience. Given that memory replay during SWS is predominantly in the forward direction (Diba & Buzsaki, 2007), the model predicts that such unified memories would be created if the sequence presentation order matches that of the unified memory, but not otherwise. Consistent with the model, we only found a difference between the Sleep and Wake groups in the Forward composite condition, but not when component words were presented in the backward direction. Moreover, consistent with the model's emphasis on replay of stored sequences, there was no difference between the groups when the component words were separated to different trials during the exposure session, a condition that yielded only few false memories on average (Figure 2). Finally, and also consistent with the model, there was no difference between the groups in a baseline condition consisting of totally novel words, which, as expected, also yielded few false memories on average.

One important contrast with the model's predictions was the finding that, for Backward composite items, both the Sleep and the Wake group had increased levels of false memories (compared, for example, to the Novel words condition). This unexpected effect suggests that backward items tend to be combined together irrespective of sleep. This finding might be accounted for if taking under consideration the fact that memory replay could also occur during waking. Rodent studies suggest that compressed replay in the hippocampus is elicited at wake as well, often during resting periods following completion of a task, and, unlike sleep, it tends to include backward replay of recently encoded memory sequences and not just forward replay (Diba & Buzsaki, 2007). While the function of wake replay is still debated, some suggest it could contribute to memory consolidation in the same manner as sleep replay does (Rasch & Born, 2013). Since both Sleep and Wake subjects in our task had a period of rest following the completion of the task (before they went to bed or saw a movie, respectively), such backward replay could potentially have been elicited and contribute to the formation of false composite memories for the Backward items (i.e., replaying the sequence of events pet->car backwards could result in the activation of "carpet" in its regular order). Another possibility is that composite memories of both Forward and Backward items were already formed during the initial experience simply because of their close temporal proximity (and aided by the fact they were always presented in the same color), but sleep was essential in maintaining the Forward composite memories. Further research is needed to explore these possibilities.

Several previous studies have suggested that false memories could arise following sleep. Specifically, using the Deese-Roediger-McDermott (DRM) paradigm (e.g.,

Payne et al., 2009), it was shown that sleep following exposure to a group of words with a related theme (e.g., *Pillow*, *Bed*, *Night*) could lead to the formation of a false memory for the theme word (*Sleep*). However, these effects are not always found (e.g., Fenn, Gallo, Margoliash, Roediger, & Nusbaum, 2009) and they seem to decrease rather than increase with time spent in SWS (Pardilla-Delgado & Payne, 2017; Payne et al., 2009). In other words, the mechanism contributing to the effect seen in the DRM paradigm is likely different than the one presented here, and relates to deep semantic processing of the stored stimuli rather than the time-compression property of replay during SWS (Pardilla-Delgado & Payne, 2017). A more related effect to the one presented here is the demonstration that sleep in humans preferentially facilitates memory of sequences when they are presented during test in the original forward direction compared to backwards, a finding that was interpreted as resulting from memory replay during sleep (Drosopoulos et al., 2007). Our findings add to that previous demonstration by introducing the element of time compression in the process, and by showing it specifically relates to SWS.

Conclusion

In the current study, we demonstrated that an afternoon nap could lead to the formation of false composite memories made of events that were previously presented sequentially. The importance of these results is twofold. First, our novel behavioral paradigm potentially allows for tapping replay compression mechanisms during sleep, opening the door for various future investigations of this phenomenon in humans. Second, our findings provide evidence for the functional role of time compression in memory replay, suggesting it contributes to the association of disparate yet proximal events and showing that in addition to the regular facilitation seen in the majority of studies, this mechanism could also lead to impairments in memory.

References

August, D. A., & Levy, W. B. (1999). Temporal sequence compression by an integrate-and-fire model of hippocampal area CA3. *Journal of computational neuroscience*, *6*, 71-90.

Drosopoulos, S., Windau, E., Wagner, U., & Born, J. (2007). Sleep enforces the temporal order in memory. *PLoS One*, *2*(4), e376.

Fenn K. M., Gallo D. A., Margoliash D, Roediger HL, & Nusbaum HC. (2009). Reduced false memory after sleep. *Learning & Memory, 16,* 509-513.

Fischer, S., Drosopoulos, S., Tsen, J., & Born, J. (2006). Implicit learning–explicit knowing: a role for sleep in memory system interaction. *Journal of Cognitive Neuroscience*, *18*, 311-319.

Lerner, I. (2017a). Sleep is for the brain: Contemporary computational approaches in the study of sleep and memory and a Novel 'Temporal Scaffolding' Hypothesis. In: A. Moustafa (Ed)*, Computational Models of Brain and Behavior*. Hoboken, NJ: Wiley

Lerner, I. (2017b). Unsupervised Temporal Learning during Sleep Supports Insight. *Conference on Cognitive Computational Neuroscience (CCN) 2017*. Archived at: https://www2.securecms.com/CCNeuro/docs-0/5928daeb68ed3f7a4e8a2571.pdf

Lerner, I. , Ketz, N. A., Jones, A.P., Bryant, N.B., Robert, B., Skorheim, S.W., Hartholt, A., Rizzo, A.S., Gluck, M.A., Clark, V.P., Pilly, P.K (2019). Transcranial Current Stimulation During Sleep Facilitates Insight into Temporal Rules, but does not Consolidate Memories of Individual Sequential Experiences. *Scientific Reports, 9,* 1516.

Lewis, P. A., & Durrant, S. J. (2011). Overlapping memory replay during sleep builds cognitive schemata. *Trends in cognitive sciences*, *15*, 343-351.

McClelland, J. L., McNaughton, B. L., & O'reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review*, *102*, 419.

Pardilla-Delgado, E., & Payne, J. D. (2017). The impact of sleep on true and false memory across long delays. *Neurobiology of learning and memory*, *137*, 123-133.

Payne J. D., Schacter DL, Propper R. E., Huang L-W, Wamsley EJ, Tucker MA, et al. (2009). The role of sleep in false memory formation. *Neurobiology of Learning and Memory, 92,* 327-334.

Rasch, B., & Born, J. (2013). About sleep's role in memory. *Physiological reviews*, *93*, 681-766.

Wagner, U., Gais, S., Haider, H., Verleger, R., & Born, J. (2004). Sleep inspires insight. *Nature*, *427*, 352.

Wang, Y., Loparo, K. A., Kelly, M. R., & Kaplan, R. F. (2015). Evaluation of an automated single-channel sleep staging algorithm. *Nature and science of sleep*, *7*, 101.

Wiedemann, C. T., & Kahana, M. J. (2016). Assessing recognition memory using confidence ratings and response times. *Royal Society open science*, *3*, 150670.

Wilhelm, I., Rose, M., Imhof, K. I., Rasch, B., Büchel, C., & Born, J. (2013). The sleeping child outplays the adult's capacity to convert implicit into explicit knowledge. *Nature Neuroscience*, *16*, 391.

Wilson, M. A., & McNaughton, B. L. (1994). Reactivation of hippocampal ensemble memories during sleep. *Science*, *265*, 676-679.

Yordanova, J., Kolev, V., Verleger, R., Bataghva, Z., Born, J., & Wagner, U. (2008). Shifting from implicit to explicit knowledge: different roles of early-and late-night sleep. *Learning & Memory*, *15*, 508-515.

What if everybody did that?: Universalization as a mechanism of moral decision-making

Sydney Levine
Massachusetts Institute of Technology, Cambridge, Massachusetts, United States

Max Kleiman-Weiner
Massachusetts Institute of Technology, Cambridge, Massachusetts, United States

Laura Schulz
Massachusetts Institute of Technology, Cambridge, Massachusetts, United States

Josh Tenenbaum
MIT, Cambridge, Massachusetts, United States

Fiery Cushman
Harvard University, Cambridge, Massachusetts, United States

Abstract

We describe a cognitive mechanism of moral judgment, universalization, that has received little attention up to now. Under universalization, an action's moral permissibility is determined by calculating what the outcome would be if all people who are similarly situated to the actor also acted in that way. This mechanism is particularly well-suited to capture our moral judgments of free-rider cases, where one person doing the action increases utility but many people doing it decreases utility. Universalization fits into an agreement-based (contractualist) theory of moral cognition, and explains properties of our moral judgments that an outcome-based or rule-based approach cannot. We show patterns of universalization reasoning in young children as well as adults.

Active physical inference via reinforcement learning

Shuaiji Li, **Yu Sun**, **Sijia Liu**, **Tianyu Wang**
{sl6486, ys3225, sl6496, tw1682}@nyu.edu
Center for Data Science, New York University, 60 Fifth Ave, New York City, NY 10011 USA

Todd M. Gureckis
todd.gureckis@nyu.edu
Department of Psychology, New York University, 6 Washington Pl, New York City, NY 10003 USA

Neil R. Bramley
neil.bramley@ed.ac.uk
Department of Psychology, University of Edinburgh, 7 George Square, Edinburgh, Scotland EH8 9JZ

Abstract

When encountering unfamiliar physical objects, children and adults often perform structured interrogatory actions such as grasping and prodding, so revealing latent physical properties such as masses and textures. However, the processes driving and supporting these curious behaviors are still largely mysterious. In this paper, we develop and train an agent able to actively uncover latent physical properties such as the mass and force of objects in a simulated physical "micro-world'. Concretely, we used a simulation-based-inference framework to quantify the physical information produced by observation and interaction with the evolving dynamic environment. We used model-free reinforcement learning algorithm to train an agent to implement general strategies for revealing latent physical properties. We compare the behaviors of this agent to the human behaviors observed in a similar task.

Keywords: physical simulation; active learning; probabilistic inference; reinforcement learning

Human adults have an intuitive understanding of the physical world that supports rapid and accurate predictions, judgments and goal-directed actions. For example, we can quickly estimate how heavy a door is by how it responds to a push, judge whether a tower is stable with a glance, or predict where to stand to catch a baseball by watching its trajectory. These abilities are so ingrained in everyday life, we are rarely aware of the complexity of identifying physical properties of objects from brief experiences and interactions with their dynamics.

Bramley, Gerstenberg, Tenenbaum, and Gureckis (2018) explored the strategies people use to actively infer masses and forces of attraction and repulsion that relate objects in a simple simulated environment. Their learning environment consisted of a bounded two-dimensional "hockey puck" world, containing four circular objects of varying masses and related with pairwise attractive or repulsive magnet-like forces. For subjects, the world was displayed on a computer screen and updated in real time using a physics engine that approximates Newton's laws of motion (see Figure 1a). Subjects could hold-click to "grab" onto any object then move their cursor to "drag" that object around the scene, so altering how the scene played out and often better revealing the physical property they had been tasked with inferring. Bramley et al. (2018) used a simulation-based inference model to assess what information was generated by subjects, and contrasted trials in which the subject's goal was to identify the relative masses of two objects against trials in which their goal was to identify the pairwise force between two objects. They found that learners were able to gather information selectively relevant to their learning goal, and exhibited markedly different behaviours dependent on the goal that could be classified into a classes of frequent experimental strategies such as staging collisions by *throwing* objects at one another, *shaking* them back and forth, bringing pairs of objects close together (dubbed *encroaching*). However, they did not model how such strategies were learned or generalised successfully across learning instances

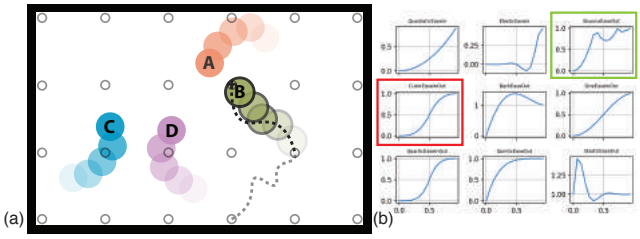


Figure 1: Visualisation of task environment adapted from Bramley et al. (2018). Gray circle overlay shows grid of target locations and dashed line give an example control trajectory in which B is grabbed and moved toward A. (b) Sample of the interpolation functions used to create smooth control trajectories. The *cubicEaseInOut* in the red box and the *bounceEaseOut* in the green box are most frequent interpolations selected by force focused and mass focused agents, respectively, as reported in Figure 4.

In the current work, we take a step toward understanding this ability. We use reinforcement learning, to train an artificial agent to minimize its own uncertainty about these specific physical parameters by composing action sequences from an action space qualitatively similar to human subjects' mouse movements, training this ability across a diverse set of ground truth and initial conditions. We are interested whether, with sufficient training and subjective information as a reward signal, an agent can learn to produce robustly informative action sequences, or strategies, for revealing particular physical properties, and whether these will reflect the behaviors previ-

ously seen in human subjects.

Related Work

Understanding what drives information-seeking behavior has long been a goal in psychology (see Schulz & Gershman, 2019, for a recent review). Some approaches attempt to tie information seeking directly (e.g., Guez, Silver, & Dayan, 2012) or incidentally (e.g., Thompson, 1933) to the familiar goal of extrinsic reward maximisation. Explicit approaches require intensive preposterior planning, that is averaging behavioral prescriptions over many potential future learning and reward trajectories (Raiffa, 1974), so are infeasible for complex problems. Incidental approaches are computationally cheaper but fail to capture information seeking behaviours that depart from noisy extrinsic goal seeking (cf., Schulz, Klenske, Bramley, & Speekenbrink, 2017).

The idea that the brain seeks information as a form of *intrinsic* reward captures a middle ground idea that that for humans, discovery is an "end in itself" driven by something we intuitively understand as "curiosity" (Gottlieb, Oudeyer, Lopes, & Baranes, 2013; Schmidhuber, 2010).

While some learning problems only occur once in a lifetime (e.g., we only need to figure out the laws of physics once), many others occur repeatedly (we encounter new objects daily and would like to be able to rapidly infer their most important properties). For these problems, there is space to *learn to actively learn* by training reusable active learning strategies through repeated experience,making them amenable to reinforcement learning (Kober & Peters, 2012).

The task we explore is challenging in part because of the continuous and complex nature of physical dynamics. However, combined with adequate function approximators such as a deep neural network, reinforcement learning has proven successful for optimising control in rich state spaces (Mnih, Heess, Graves, et al., 2014). For example, Bachman, Sordoni, and Trischler (2016) developed agents that can solve a collection of tasks which require active information seeking using deep neural networks and reinforcement learning, including cluttered MNIST (Mnih et al., 2014), Blockworld, CelebA (Liu, Luo, Wang, & Tang, 2015) and Hangman. Misha et al. (2017) also proposed and trained a deep reinforcement learning agent that can make judgments about physical properties in a simulated environment.

This project sets apart from prior work on intrinsic curiosity in that it focuses on interventional rather than observational information seeking. In our physics learning task, actions have extended complex and far reaching consequences, compounding the complexity of inference, but better mimicking the challenges of learning in the natural world.

The task

Our interactive physical environment adapts the two-dimensional physics-based environment from Bramley et al. (2018) and is implemented using the `pybox2d` library.[1] The world is limited to a 6×4 meter bounded rectangle containing four circular objects, each with radius 0.25 meters. The objects interact with one another and with the static walls according to Newtonian physics and the latent properties of mass and the pairwise forces. The complete fixed settings of the simulator are detailed at `https://bit.ly/2B4TOAf`.

Possible settings and initial uncertainty

As with Bramley et al. (2018), we will contrast actions focused on identifying objects' *masses* against actions focused on identifying the *forces* relating each pair of objects. However, we explored a larger set of possible settings of these values and correspondingly larger initial hypothesis space for the agent than human participants. To ensure the agent had equivalent initial uncertainty about both masses and forces, we defined a discrete space of 2^5 mass combinations and 2^5 force combinations and initialised the agent with a uniform prior across all 2^{10} distinct mass-force combinations. The space of mass combinations is a subset of $\mathbb{R}^4$, in which each parameter represents the mass of an object and is in the range of [1, 3]kg. Meanwhile, the space of force combinations is a subset of $\mathbb{R}^{4\times4}$, in which each force parameter takes a value $\in (-3, 0, 3)^{\mathrm{m}}/_{\mathrm{s}^2}$, representing repulsive force, no interactive force, and attractive force respectively for each pair of objects. Under the Newton's second law of motion, different mass-force combinations yield different accelerations, leading to distinct simulated trajectories, interactions and collisions.

Like the human participants in Bramley et al. (2018), the agent could "grab" one object at a time. The cursor would exert an elastic force (i.e., scaling continuously with the inverse square of the distance between the object and the cursor) attracting the controlled object to the cursor's location until it was released. All interactions and resulting trajectories were simulated by the Box2D engine.

The Agent

Learning Framework

Reinforcement learning (RL) focuses on how agents learn sequential control policies to maximize cumulative reward. We assume (S, A, R, T) defines a Markov Decision Processes (MDP) with state space S, action space A, reward function R and transition dynamics T. Modern RL can be divided into model-based and model-free approaches, where the former first learns a predictive world model then uses the model to learn a policy, while the latter learns an optimal policy directly from experience. We position our learning agent somewhat like a developing child learning about the physical properties of the world through experience but without much *a priori* knowledge. Thus, we do not assume the agent has a complete internal physics engine from which to draw samples (model-based learning) but instead is learning how best to reveal information in a model-free manner evaluated against its own learning progress (Oudeyer, Kaplan, & Hafner, 2007).

[1] `https://github.com/pybox2d/pybox2d`

Action space

As mentioned above, human subjects exhibited rich and informative behaviors when interacting with the objects in this environment. However, many of the strategies identified in Bramley et al. (2018) appeared to be composed of multiple simpler movements. For instance, "shaking" involved grabbing then moving an object rapidly back and forth between two locations while "throwing" involved grabbing an object and releasing it at speed and in a direction such that it went on to collide with another. To encode an action space expressive enough to incorporate these realistic and extended human behaviors, we combined a a grid of target locations (Figure 1a) with pool of easing (interpolation) functions (see Figure 1b). Concretely for each action, the agent chose an object to control (none, 1,2,3 or 4) and a cardinal direction (up, down, left ,right), or a quadrant (up-right, up-left, down-left, down-right) which determined a target location adjacent to its current location. It then followed a path to that location determined by its selected interpolation function. Figure 1b shows these functions in the first quadrant such that the path would be appropriately mirrored in the other three quadrants.[2] Together with horizontal and vertical movements, and no movement — i.e., the cursor pausing at its current position — our action generator output a contiguous mouse trajectory, following the policy learned by our agent via the learning framework demonstrated in the following section.

A complete learning episode consisted of a sequence of these actions and would terminate when the agents' uncertainty about a target property of the environment reached a threshold, or a timeout limit was reached. For simplicity we assumed each action occurred across a fixed time window, and to ensure continuity of mouse movement, the start point of each action was the end position for the last action. This resulted in a continuous action trajectory decomposed into a sequence of discrete action choices.

State definition

We defined the motion of an object at every time step t with scalars m_t and ϕ_t, where m_t represents the magnitude and ϕ_t the direction of the object. Given a two-dimensional space, $m_t = \sqrt{v_{x_t}^2 + v_{y_t}^2}$ and $\phi_t = \arctan \frac{v_{y_t}}{v_{x_t}}$, where $v_t = [v_{x_t}, v_{y_t}]$ is the velocity vector of the object at time step t. Together with the location tuple (x_t, y_t), the object state $s_t \in S$ is thereby a four-dimensional vector $[m_t, \phi_t, x_t, y_t]$.

Intrinsic Reward Signal

Our agents' goal was to minimise its final uncertainty about either mass or force. Thus its reward signal was based on computing its reduction in entropy with respect to the target property (Shannon, 1951). Following Bramley et al. (2018), we assumed likelihoods were computed based on divergence between mental simulations and the observed trajectories.

[2] All together there were (stay + $\rightarrow$ + $\uparrow$ + $\leftarrow$ + $\downarrow$ + 31[functions] $\times$ 4[quadrants]) $\times$ (4[objects] + no object)=645 possible actions.

Concretely, to infer how likely a mass-force combination, w, is we assume actual observed dynamics are compared against dynamics simulated assuming those properties. Let $w \in W$, where W is the space of all possible mass-force combinations. Before an episode starts, the agent has a uniform prior over settings $p(W)$. After a period of action and observation d, we assess the likelihood of the observed trajectory o under all possible w, and use this to update the prior distribution $p(W|d)$.

Following Vul, Frank, Alvarez, and Tenenbaum (2009), we modelled the likelihood of observing the object trajectories o given the potential property setting w and the mouse trajectory a using a Gaussian error distribution:

$$p(o \mid w, a, \beta) = \prod_{t=1}^{T} \exp^{-\frac{\beta}{2\Sigma}(s_t - d_t)^\top (s_t - d_t)}, \tag{1}$$

where d_t is the observed $[m_t, \phi_t]$ produced by the true environment w', and s_t is the velocity vector of the trajectory simulated under w. The covariance matrix was $\Sigma = \begin{bmatrix} \sigma_m^2 & 0 \\ 0 & \sigma_\phi^2 \end{bmatrix}$ where σ_m^2 and σ_ϕ^2 were set to the empirical standard deviations of the disparities between simulations and the actual observations in Bramley et al. (2018). β was a scaling parameter determining how confidently the agent could perceive divergences between the objects' true and simulated trajectories.[3]

According to Bayes' rule, once we measure the likelihood through Equation 1, we can calculate the posterior of the potential latent physics properties w by:

$$p(w \mid o, a, \beta) = \frac{1}{Z} p(o \mid w, a, \beta) p(w), \tag{2}$$

where $p(w)$ is the prior and Z is a normalizing constant. Using Shannon entropy (1951) as a measurement of the amount of remaining uncertainty about w', we formulated the immediate reward after action t as:

$$r_t = -\Big(\sum_{w \in W} p(w) \log p(w) - \sum_{w \in W} p(w \mid o, a, \beta) \log p(w \mid o, a, \beta)\Big). \tag{3}$$

For particular parameters of interest (here masses of the four objects), the posteriors and priors are marginalized over the remaining parameters (in this case, the pairwise forces) when estimating the parameter specific uncertainty reduction. The resultant reward signal r embodies the amount of information about the latent physics parameters (mass or force) a particular agent's action has obtained.

Deep Q-Learning

Q-learning is a model-free learning algorithm that estimates the expected cumulative discounted rewards of performing an

[3] Intuitively, $\beta = 0$ would result in no learning and $\beta \lim \infty$ would lead to an implausibly powerful learner with infinite perceptual precision. In our experiment, we assumed a constant β of $\frac{1}{50}$. We also held a fixed size of the time window T of $\frac{1}{6}$ seconds, over which forward simulations were compared against observations before being corrected.

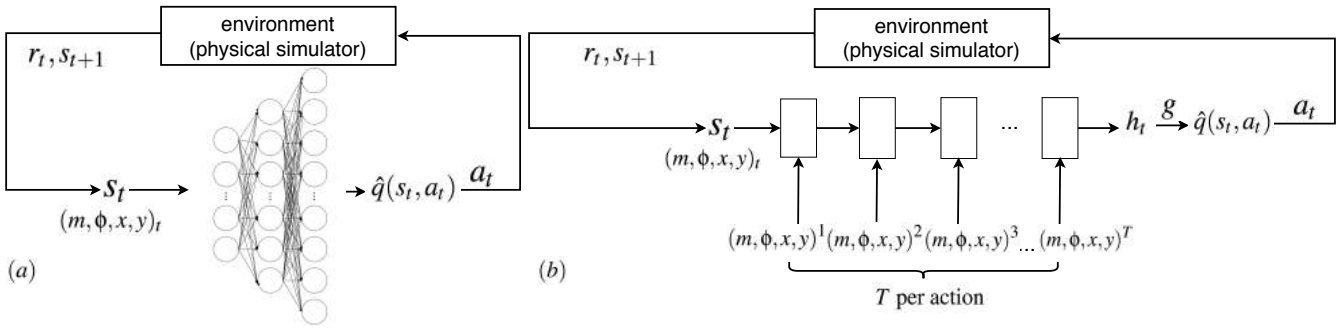


Figure 2: Proposed (a) MLP and (b) RQN structures of our Q-learning framework. For every time period, the network takes a sequence of object states and return the estimated Q-values. The control policy of the next time period, a sequence of mouse trajectories, is determined by selecting the action corresponding with the largest Q-values over the action space.

action from a given state (Watkins & Dayan, 1992). These estimated cumulative discounted rewards are often called returns and represented as Q-values (state-action values). One way to iteratively learn Q-values is by using a *Temporal Difference* (TD) algorithm (Sutton, 1988). TD updates Q-values by:

$$\hat{Q}(s,a) := (1-\alpha)Q(s,a) + \alpha(r + \gamma \max_{a'} Q(s',a')), \qquad (4)$$

where (s,a) is the current state-action pair, (s',a') is the state-action pair at the next time step, γ is the discounted factor, and α is a learning rate.

A challenge for conventional Q-learning is to estimate Q-values over continuous states and action spaces. For our task, despite creating a discrete action space with interpolation functions, the object's locations and motion $[m,\phi]$ was continuous throughout the world. Computing separate Q-values for every possible (s,a) pair at every time window is infeasible and naïve given the smooth relationships between nearby states. We hence adapted the *deep* Q-learning method, which instead of updating Q-values of each discrete pair (s,a) in tabular form, approximates $\hat{q}(s,a)$ with deep neural network. Denoting the parameters of the network as θ, a mean squared error loss of the Q-values approximator can be written as:

$$l(s,a \mid \theta) = \frac{1}{2}(r + \gamma \max_{a'} \hat{q}(s',a' \mid \theta) - \hat{q}(s,a \mid \theta))^2, \qquad (5)$$

where $r + \gamma \max_{a'} \hat{q}(s',a' \mid \theta)$ is the *target* of our estimation. Below, we present two neural network models, serving as the function approximators $\hat{q}_\theta$. Optimal control strategies can be discovered following this general learning structure.

Multilayer Perceptron First, we used a Multi-Layer Perceptron (MLP) with three layers to approximate Q-values (Figure 2a). For every time period t, the simulator fed the agent a sequence of object states s_t that spanned over T time-steps and offered a reward r_t to the MLP. The network outputs a deterministic cursor trajectory that was the action $a_t \in A$ with largest Q-value at t. We used ε-greedy exploration and optimized Equation 5 via stochastic semi-gradient descent (Watkins & Dayan, 1992). Note that the same network $\hat{q}$ producing the next state target Q-values was used for computing the loss of the current predictions. Such optimization can yield erratic gradients when the control policy oscillates. To deal with this potential instability, we redefined the *target* as a *target network*. Instead of using the same network to compute the next state target and the current state Q-values, another network $q_{\hat{\theta}^-}$, sharing the same structure as $\hat{q}_\theta$, was utilized to compute target Q-values during the update. The loss function in Equation 5 was then modified as:

$$l(s,a \mid \theta,\theta^-) = \frac{1}{2}\left[r + \gamma \max_{a'} \hat{q}(s',a' \mid \theta^-) - \hat{q}(s,a \mid \theta)\right]^2. \quad (6)$$

Following the 'hard-copy' update proposed in Mnih et al. (2015), we froze θ^- and copied θ into θ^- once after a few episodes. Experiments demonstrated that the MLP with *target network* model produced convergent and stable policies faster than the simple MLP model.

Recurrent Q-Network For every forward pass, the function approximator $\hat{q}_\theta$ should receive a sequence of object states with length $16T$ (four objects, each carrying a four-dimensional vector $[m,\phi,x,y]$ that depicts the object's motion per time step). An MLP approximator, as introduced in the previous section, consists of multiple fully-connected layers, where each layer contains multiple neurons that accept and send information across the network. The input motion vectors are stacked and reshaped into a $16T$-dimensional vector, and then fed into the first layer of the MLP. However, the state vectors encode the objects' movement information over a period of time, and simply flattening them may fail to capture some parts of the underlying dynamics or spatial correlations among the objects. We thereby propose a second RNN-based function approximator (Figure 2b).

Instead of treating all the motion vectors equally without order, RNN chronologically receives the motion vectors and updates its hidden cells accordingly. The output hidden vector at the last time step included not only the object state information but also their latent interconnections with the environment. To alleviate the vanishing gradient issue, we adapted the Long Short-Term Memory cell (Hochreiter & Schmidhuber, 1997) as the recurrent unit. A linear mapping function $g : S \to A$ mapped the state representation to actions. We optimize this recurrent Q-network (RQN) by semi-gradient

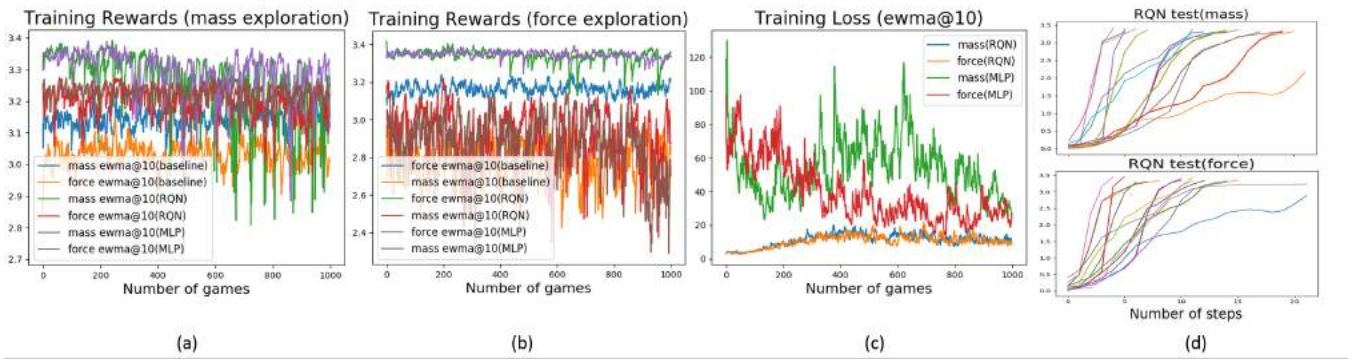


Figure 3: Reward and loss. (a) and (b) show total target and mismatched reward for each game, for the mass and force exploration tasks respectively. (c) reports the training loss associated with all proposed models. For each test game, the cumulative target rewards over actions are recorded and presented in (d).

descent using the *target network* method.

Experiments

In this section, we present some preliminary attempts at training these networks and comparing their behavior to that of humans performing the same task. We compare the achieved reward with respect to the agent's goal against its reward for the mismatched alternative goal to assess to what extent the actions were selectively informative.

Training

For our physics simulator, 60 time-steps was equivalent to 1 second. We set the time-window T as 40 time-steps, implying that each atomic action would take roughly 0.7 seconds. It is nontrivial to determine when a simulated game should stop, since uncertainty typically continues to diminish indefinitely, approaching but never reaching zero. For our agent, the total uncertainty of the latent physical parameter of interest could be obtained by computing the Shannon entropy of the initial prior distribution $p(W)_0$, denoted $\mathrm{H}(p(W)_0)$. We denoted a reward threshold factor $\gamma_r = 0.95$. Then, every time the cumulative reward reached $\gamma_r \mathrm{H}(p(W)_0)$, the game would stop. Otherwise, the agent would continue playing until the current game approached a timeout limit.

The three-layer MLP had 150, 250, and 450 neurons per layer. The input state s_t of MLP was in R^{640}, while for RQN, it was $R^{16\text{x}40}$. The approximated Q-values $\hat{q}_\theta(s,a) \in R^{645}$. We initialized the exploration rate ε as 0.5, discounted by a factor of 0.95 every 20 games until ε reached 0.01. The discount factor γ was 0.99, and the weights of *target network* were cloned from $\hat{q}_\theta$ every 20 games.

A training set with 60 distinct ground truth w, initial object locations, velocities, cursor positions, and initial velocities was created, and a holdout test set containing 20 different configurations, are defined beforehand to ensure the robustness of the models. For each proposed model, we trained the agent for the mass exploration task and the force exploration task separately with 1000 episodes, and the training errors and rewards are illustrated in Figure 3(a)-(c). Within each task, both the target reward and the mismatched reward (i.e., force in a mass exploration task) are recorded. For comparison purposes, we also ran a baseline policy (randomly selected actions) for each task. To better illustrate the moving trends, we weighted the training rewards and the loss by exponentially weighted moving average (ewma) with span 10.

Compared with the MLP-based models, RQN converged faster with small and stable loss, as shown in Figure 3(c). Both models produced swinging errors in the training process. The intuition is that 'hard' copying weights from θ into θ^- yielded a time interval when θ^- was frozen and the predicted Q-values diverged from the target Q-values. Such oscillations would not distort the optimization of the objective function, as the time interval was small and θ^- was constantly updated, ensuring steady amounts of information explored by the agent. Given a flexible timeout limit, both methods were able to uncover the uncertainty of the environment, and outperforms the baseline policy with higher and more stable cumulative rewards, as depicted in Figure 3(a) and (b). Mass reward was more variable. This is in line with the Bramley et al. (2018) finding that evidence about *mass* tends to come in sporadic spikes when objects collide or are moved rapidly, while *force* information typically accumulates more smoothly whenever objects are in close proximity. Thus, gathering mass information reliably have depended on more specific and targeted actions while moving objects closer together may have been sufficient for force information. Overall, the RQN exhibited continuing amounts of achieved information on all latent physics parameters. We applied the trained RQN agent on the test configuration set for mass and force exploration separately, each with 20 games. As reported in Figure 3(d), in most of the cases the agent could effectively capture the underlying properties of the environment within 20 steps. Since explorations were excluded from the testing, few 'abnormal' actions appeared. For those rare cases, our agent quickly came back to the right track and could still complete the learning task within the timeout limit.

Using the trained models we created a small dataset of 10 *force*-focused episodes and 10 *mass*-focused episodes using a holdout set of new worlds and starting locations. See `https://bit.ly/2FYTjvD` for videos. Comparing these episodes, we found that achieved information reward was similar for mass or force focused trials 2.82 ± 0.54, $2.96 \pm$

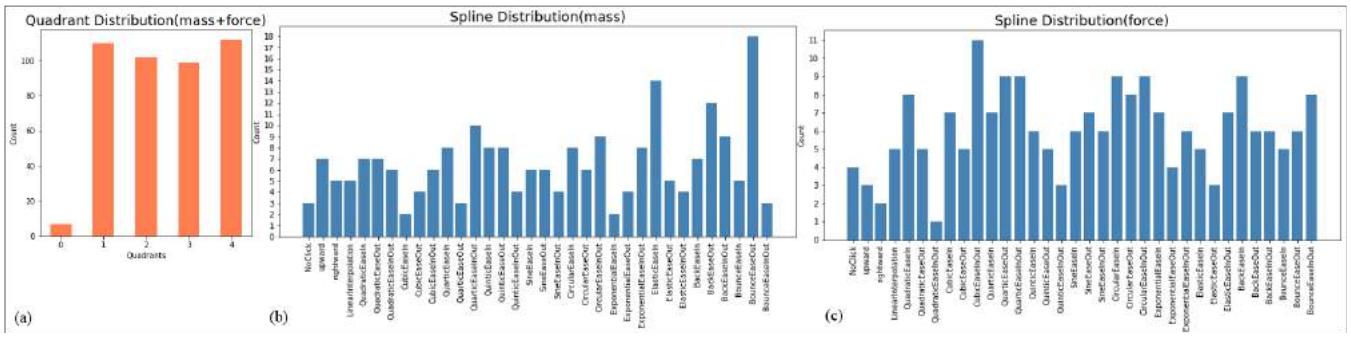


Figure 4: Distributions of emerged actions in 20 test games (10 for mass focused and 10 for force focused). (a) Sum of the quadrant distribution of mass and force tasks. '0' means the agent staying over at its position within the time-windows. (b) Spline distribution of mass task (c) Spline distribution of force task

$0.39, t(18) = .98, p = .32$ but that significantly more information was generated for the property matched with the agents' goal than the alternative property (see also lower orange and blue lines in Figure 3a and b). That is, the agent generated more evidence and reward on average about the target property 3.1 ± 0.35bits than the mismatched property 2.7 ± 0.52 bits $t(18) = 2.4, p = 0.021$ similarly to human subjects in Bramley et al. (2018).

Trained agent behaviour

The agent frequently moved the controlled object closer to the other objects, reducing the distance to the closest object from 1.28 ± 0.24m to 1.01 ± 0.23m on average during each control action $t(38) = 3.6, p < .001$. As we see in Figure 4a, the agent learned that moving was normally more informative than staying still. There were also hints of goal dependent control strategies similar related to those identified in Bramley et al. (2018). For example, the most frequently selected interpolation by the mass focused agent was *bounceEaseOut* (Figure 4b and green highlight in Figure 1b), a particularly dynamic motion consistent with the rapid changes of direction associated with shaking or knocking observed frequently in the human data on *mass* focused trials. The intuition both there and here is that these actions strongly reveal objects mass by causing rapid changes in objects' directions. Meanwhile, the most frequent interpolation selected by force focused agent was *cubicEaseInOut* (Figure 4c and red highlight Figure 1b), a smooth motion intuitively consistent with the "encroaching" behavior observed frequently in force trials in Bramley et al. (2018) and effective in providing strong evidence about mass.

Discussion and Conclusions

Humans display sophisticated intervention strategies when actively inferring the properties of physical objects. We used model-free reinforcement learning, deep function approximation, and simulation based inference to build an agent able to efficiently reveal the latent physical properties in human-like ways without external input. Part of the insight gleaned from this project comes from our solutions to the engineering challenges involved in creating a successful agent. To produce extended actions with richness and qualitative correspondence with humans', we found success with an action space that combined a discrete set of target locations with a discrete set of smoothing splines. We found that learning to associate action sequences with successful resolution of uncertainty was much more effective with a recurrent network architecture, but that robust strategies could be learned through model-free Q-learning. Following the predicted optimal control policies, not only did the agent uncover the latent parameters of interest, but there were also hints of behavioral correspondence with human subjects in that our mass trained agent would select more jagged and dynamic trajectories aligned with the strategies observed in Bramley et al. (2018).

While this study provides a valuable first step to understanding how humans learn and apply rich interrogatory behaviours when interacting with the natural world, it also has its limitations. One of these is the use of the physics simulator to calculate the reward signal. This is a rational idealisation of physics inference, but embodies the overly strong assumption that the agent is able to simulate the world accurately and perform approximate Bayesian inference with its own interaction data. A more plausible and practically viable approach to rewarding informative control would be to train a separate prediction network to anticipate upcoming dynamics, and use some function of its loss over time as a reward signal (cf. Oudeyer et al., 2007; Pathak, Agrawal, Efros, & Darrell, 2017). A more integrated agent could also use the predictor network approximation to plan actions that are likely to be informative through preplay Chentanez, Barto, and Singh (2005). Following Haber, Mrowca, Fei-Fei, and Yamins (2018), a complex adversarial-based learning framework may be helpful. In future work we plan to combine more realistic intrinsic rewards, richer action space and model based planning to better mimic the ability of humans to create intuitive physical experiments.

Acknowledgments

This research was supported by NSF grant BCS-1255538, the John Templeton Foundation "Varieties of Understanding" project, a John S. McDonnell Foundation Scholar Award to TMG, and the Moore-Sloan Data Science Environment at NYU to Neil Bramley. Thanks also to Tim Wu who helped with video coding.

References

Bachman, P., Sordoni, A., & Trischler, A. (2016). Towards information-seeking agents. *arXiv preprint arXiv:1612.02605*.

Bramley, N. R., Gerstenberg, T., Tenenbaum, J. B., & Gureckis, T. M. (2018). Intuitive experimentation in the physical world. *Cognitive Psychology*, *105*, 9 - 38.

Chentanez, N., Barto, A. G., & Singh, S. P. (2005). Intrinsically motivated reinforcement learning. In *Advances in neural information processing systems* (pp. 1281–1288).

Gottlieb, J., Oudeyer, P.-Y., Lopes, M., & Baranes, A. (2013). Information-seeking, curiosity, and attention: computational and neural mechanisms. *Trends in cognitive sciences*, *17*(11), 585–593.

Guez, A., Silver, D., & Dayan, P. (2012). Efficient bayes-adaptive reinforcement learning using sample-based search. In *Advances in neural information processing systems* (pp. 1025–1033).

Haber, N., Mrowca, D., Fei-Fei, L., & Yamins, D. L. (2018). Learning to play with intrinsically-motivated self-aware agents. *arXiv preprint arXiv:1802.07442*.

Hochreiter, S., & Schmidhuber, J. (1997, November). Long short-term memory. *Neural Comput.*, *9*(8), 1735–1780.

Kober, J., & Peters, J. (2012). Reinforcement learning in robotics: A survey. In *Reinforcement learning* (pp. 579–610). Springer.

Liu, Z., Luo, P., Wang, X., & Tang, X. (2015). Deep learning face attributes in the wild. In *Proceedings of the ieee international conference on computer vision* (pp. 3730–3738).

Misha, D., Pulkit, A., Tejas D, K., Tom, E., Peter, B., & Nando de, F. (2017). Learning to perform physics experiments via deep reinforcement learning. In *Proceedings of international conference on learning representations.*

Mnih, V., Heess, N., Graves, A., et al. (2014). Recurrent models of visual attention. In *Advances in neural information processing systems* (pp. 2204–2212).

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*, 529–533.

Oudeyer, P.-Y., Kaplan, F., & Hafner, V. V. (2007). Intrinsic motivation systems for autonomous mental development. *IEEE transactions on evolutionary computation*, *11*(2), 265–286.

Pathak, D., Agrawal, P., Efros, A. A., & Darrell, T. (2017). Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning (icml)* (Vol. 2017).

Raiffa, H. (1974). *Applied statistical decision theory*. Wiley.

Schmidhuber, J. (2010). Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Transactions on Autonomous Mental Development*, *2*(3), 230–247.

Schulz, E., & Gershman, S. J. (2019). The algorithmic architecture of exploration in the human brain. *Current opinion in neurobiology*, *55*, 7–14.

Schulz, E., Klenske, E., Bramley, N., & Speekenbrink, M. (2017). Strategic exploration in human adaptive control. *bioRxiv*, 110486.

Shannon, C. E. (1951, 01). Prediction and entropy of printed english. *Bell System Technical Journal*, *30*, 50-64.

Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine Learning*, *3*, 9-44.

Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, *25*(3/4), 285–294.

Vul, E., Frank, M. C., Alvarez, G. A., & Tenenbaum, J. B. (2009). Explaining human multiple object tracking as resource-constrained approximate inference in a dynamic probabilistic model. In *Advances in neural information processing systems* (p. 1955-1963).

Watkins, C. J., & Dayan, P. (1992). Q-learning. *Machine learning*, *8*(3-4), 279–292.

The critical moment is coming: Modeling the dynamics of suspense

Zhi-Wei Li (zhiwei.li@nyu.edu)
Center for Neural Science, New York University

Neil R. Bramley (neil.bramley@ed.ac.uk)
Department of Psychology, University of Edinburgh

Todd M. Gureckis (todd.gureckis@nyu.edu)
Department of Psychology, New York University

Abstract

Suspense is an affective state that contributes to our enjoyment of experiences such as movies and sports. Ely, Frankel, and Kamenica (2015) proposed a formal definition of suspense which depends on the variance of subjective future beliefs about an outcome of interest (e.g., winning a game). In order to evaluate this theory, we designed a task based on the card game Blackjack where a variety of suspense dynamics can be experimentally induced. By presenting participants with identical sequences of information (i.e., card draws), but manipulating contextual knowledge (i.e., their understanding of the rules of the game) we were able to show that self-reported suspense follows the predictions of the model. Follow-up model comparison further showed an advantage for the "suspense as variance of future beliefs" account over a number of alternative definitions of suspense, including some that depend only on current uncertainty (not the future). This paper is an initial attempt to link aspects of formal models of information and uncertainty with affective cognitive states.

Keywords: suspense; affect; prediction; expectation; probabilistic modelling

Introduction

Suspense refers to sensations of hopeful or anxious anticipation. These familiar affective states often precede the revelation of important information—exam results, paternity tests, election outcomes and so forth. However, we also feel suspense in situations where there are no direct personal consequences. For example, children enjoy listening to stories that happen in imagined kingdoms, adults spend time watching televised sports, and Hollywood movies are a multi-billion dollar industry. A key feature of these experiences is that information is incrementally revealed over time to the observer, often with the goal of building anticipation and suspense. The goal of this paper is to empirically study the relation between self-reported feelings of suspense and the dynamics of information and uncertainty.

Suspense as the variance in future beliefs

A recent theory in the economics literature proposes that suspense can be explained as an increasing function of the "variance of future beliefs" (Ely et al., 2015). Here the beliefs refer to the probability of a significant outcome (e.g., which team will win a game) that is updated in time with information as an experience unfolds. People are assumed to also estimate how their belief may change in the future. For example, if a doctor arranges to call a patient at a particular time with test results, in the period leading up to the phone call the patient might expect that their belief about their health could soon change (although they may not know what they will learn). Conditioned on the information one expects to receive, if the subsequent future beliefs would be very different from one another they would be said to have high variance. For example, if the test the doctor performed was routine, the patient would not expect their future knowledge state to change much after the call (low variance). As a result they would experience low levels of suspense. In contrast, if the test was a cancer screening, then the call might either alter the person's life or leave them reassured (high variance), and thus they would experience high levels of suspense in that moment.

To formalize these intuitions, we assume belief change is Markovian in that a viewer's subjective belief μ about some outcome evolves over a series of discrete time points t, such as individual points in tennis, card draws in a game, or time passing in a movie. At each time point, relevant information may be encountered and people update their beliefs μ_t (e.g., by Bayesian updating). In addition, viewers also anticipate future information using their understanding of the situation. For example, a viewer might anticipate that their favorite team will score on the next play or that the opposing team will score, each representing a state s. The state s has a probability of being realized $P(s)$ and will result in a future belief μ^s_{t+1}. The variance among these beliefs indicates how different the future might be, and therefore how much suspense might be evoked.

Formally, Ely et al. defined the momentary suspense at time t, S_t as:

$$\begin{aligned} \mathcal{S}_t &= \mathrm{E}_s[(\mu^s_{t+1} - \langle \mu_{t+1} \rangle_s)^2] \\ &= \mathrm{E}_s[(\mu^s_{t+1} - \mu_t)^2] \\ &= \sum_s P(s)(\mu^s_{t+1} - \mu_t)^2 \end{aligned} \tag{1}$$

and we adopt the same notation throughout this paper.

Note that the term $(\mu^s_{t+1} - \mu_t)^2$ may be also interpreted as a metric of variance of belief change or "surprise" that follows learning a piece of information. As a result, the value, S_t can be also be interpreted as the expected future surprise or expected future belief change from the next time period.

Figure 1 gives a graphical overview of the model applied to a hypothetical tennis game. Here μ is the probability of winning the game ($\mu = 1$ if team A wins and $\mu = 0$ if they lose), each point is one time step, and s is whoever wins the

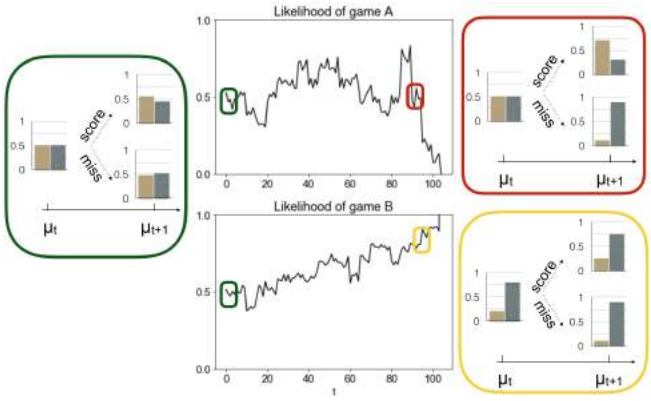


Figure 1: Demonstration of the belief trajectory during watch tennis games and the related suspense predicted from our model. Details see the main text.

next point. In the center of Figure 1 we show the unfolding of belief about who will win for two different games (A and B) with the x-axis representing time. The beginning of both games is not very suspenseful, since whoever wins the first few points has little impact on predictions about the final outcome. However, the end of the game A is more suspenseful since whoever wins a point will greatly swing the final outcome, while game B is less suspenseful since one side has already virtually secured victory.

An experimental test of the theory

Ely et al. (2015) articulated the basic outline of the theory described above and explored a number theoretical analyses of the optimal structure for games to maintain suspense. However, to our knowledge, this operational approach to suspense has not yet been examined empirically. We propose that a useful behavioral paradigm for testing this theory needs to have at least two features:

1. The experiment context should be quantifiable in a probabilistic model. This tends to exclude tasks like reading stories and watching movies because it is not trivial to convert these complex situations into accurate probability models.

2. The experiment paradigm should allow the decoupling of the external stimulus and internal belief. In most prior work, changes in suspense are always confounded with incidental features of the stimuli. To validate the belief-based account of suspense, the ideal experiment would manipulate an observed internal belief through some prior knowledge while holding other aspects of the stimulus and task identical.

With these criteria in mind, we designed a card game related to the classic casino game Blackjack. Participants are asked to draw cards from a small deck with a known distribution of cards and report their moment-by-moment suspense. Intuitively, suspense builds in the task when the sum of the drawn cards approaches a boundary value (21 in Blackjack). If the sum exceeds or hits this value the game is lost. Because the distribution of cards and the probability of drawing any card can be determined exactly, the game is an ideal test bed for exploring information-theoretic models of suspense. In addition, the game is relatively fun, intuitive, and easy to explain to participants.

To address the second concern from above, participants were given one of two different rules for how the game would be scored. In one version, the game was lost anytime the sum of the cards drawn so far met or exceeded the boundary value. This is the traditional concept of "bust" from Blackjack. In a second version, the game was lost only if the sum met or exceeded the boundary value on the final draw of the game. Due to the presence of negatively valued cards, it was possible for the sum to exceed and then return to safety. The differences between these two rules allows us to compare identical sequences of cards, but to modulate if a given card draw was more or less suspenseful about the game outcome according the the Ely et al. theory. To optimize the power of our design, we used a computer-aided method to search for best rules, card decks, and card sequences that result in strong predicted suspense differences under the two rules.

Methods

Participants 263 people (113 female), age 36.7±20.4 (mean±SD) were recruited from Amazon Mechanical Turk using psiTurk (Gureckis et al., 2016) and paid 90 cents (60 cents of this was a bonus that in actuality was the same for all participants). The task took 12±3 minutes to complete.

Procedure Participants were told that we were interested in their feelings of suspense while playing a simple card game. Each participant went through an extensive tutorial covering the rules of the game, and could only continue if they correctly answered a series of comprehension questions. They then played two rounds of training games which were identical to the real games except they were told there would be no bonus. After completing these tasks, participants played a sequence of three games with a $0.60 bonus payment for each game that was won (as describe below, all participants won one game). Afterwards they answered a questionnaire about their strategies and about their perception of the task.

Similar to Blackjack, in each round of a game, the player draws cards from a deck of nine cards. To increase the trial-by-trial suspense dynamics, we use a two-step process for choosing each card: first, the participant sees the animation of nine cards shuffling (Figure 2A); next, the first two cards at the top of the deck were selected (Figure 2B); next, the participant uses the keyboard to spin an animated wheel which (depending where it lands) decides the identity of the final card (Figure 2C). The wheel was programmed so that it spins more when the participant presses the key longer but, unknown to subjects, the spinner always ends up selecting a predetermined card. The purpose of the spinner was to give

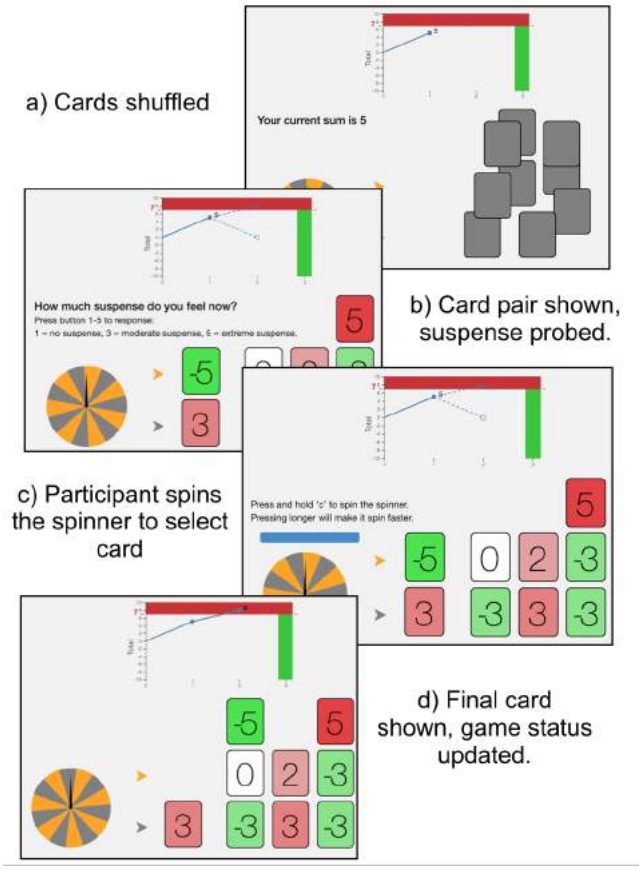


Figure 2: The game interface. a) Card shuffle is an animation that the participant cannot control. b) Participants see the selected card pair and are probed about their suspense on a scale of 1-5. c) Participants press and hold a key to spin the animated wheel. d) When the wheel stops one of the two selected cards is chosen. The whole process repeats until the game has ended.

participants a feeling of control and chance thus they do not lose interest early, although in fact the sequence of cards to be chosen was fixed for the purposes of experimental control. After a card is selected the participant's current card total (the sum of the face value of all of the cards they have drawn so far) was automatically updated in a graph at the top of the screen (Figure 2D).

To measure suspense, after the two candidate cards are shown and before spinning the wheel, we directly asked the participant to rate their current suspense with the keyboard from number 1 to 5 where 1 means no suspense and 5 means very suspenseful. Previous studies on suspense have also chosen a 7-point scale (Gerrig & Bernardo, 1994; Knobloch-Westerwick, David, Eastin, Tamborini, & Greenwood, 2009) and 11-point scale (Comisky & Bryant, 1982; Cupchtk, Oatleyb, & Vorderee, 1998), yet we are unaware of any systematic comparison of different response scales for suspense measurement. No other instructions were given about the use of the scale. However, we asked participants to report how they personally defined suspense in the post-task questionnaire.

Implementing the belief updating model To calculate the belief μ (probability of winning) at a given moment t, we use an exact enumerative strategy. We first enumerate all the possible future card draws remaining in the game according to the known card distribution of the deck. Summing these values, we get the predicted card sum probability. The winning probability calculation is rule-dependent: if the game is played according to the bust rule, we get the card sum distribution for one future step, keep the surviving card sums, continue to the next step and so forth until the game end. If the rule is no-bust (i.e. only the sum of cards at the end of the game matters), we directly calculate the card sum distribution at the end of the game and count the proportion of winning relative to losing sums.

Since the suspense is reported after the pair of possible cards are shown, we assume that suspense is the variance of future probabilities of winning after spinning the wheel and the card being finally drawn. Given that the wheel has equal area for both options, the probability of both future states are equal: $p(s) = 0.5$. The suspense prediction can then be calculated utilizing the equation 1.

Design We will introduce the design of card sequences, then the condition and counterbalance structure.

Belief manipulation: Model-based stimuli design. One key aspect of the theory is that suspense is the result of an active prediction about future stimuli and future beliefs, not the mere reaction to current stimuli. To test this, we looked for rule-dependent differences in suspense responses for the identical card sequences. Given the inherently noisy nature of self reports, we looked for sequences with large predicted differences by maximising a score:

$$\begin{aligned} \text{score}(\text{seq}, \text{deck}, \text{rulepair}) = S^{\text{rule1}} + S^{\text{rule2}} \\ -\alpha \cdot r(S^{\text{rule1}}, S^{\text{rule2}}) \end{aligned} \quad (2)$$

where α is a positive weight constant and $r(\cdot)$ is Pearson's correlation coefficient. The first two terms ensure the average suspense level is not too low while the third encourages anti-correlation between the suspense trajectory under two rules. We set α to a positive constant that makes the two terms have similar magnitude.

We searched the space of rules by generating 5000 random combinations of deck and card sequence valid under both rules and scoring them, then filtered with restrictions to ensure the game also feels like plausible random draws from the deck (details on *Github*). The result of this search was a set of 3 deck/card sequence combinations that evoke strongly different suspense trajectories under two rules: *Bust* with a bound of 7—i.e., the card sum should never exceed 7—and *No-bust* game with a bound of 3—i.e., the sum of cards should not exceed 3 at the end. The full sequences are shown in in Figure 3a.

Participants were randomly assigned to one of the two rule conditions. Two of the games were selected from *high suspense 1-3*.

Besides sequences with interesting suspense dynamics, we also designed two *no suspense* games where the card pairs have similar or identical values, or values that are non-consequential to the games outcome (Figure 3, "no suspense" 1-2), thus should intuitively induce low suspense. According to Ely et al. model, the predicted suspense at every point in these games is zero.

Task duration and manipulation. Each participant was assigned to one rule condition (rule was a between subject manipulation) and played two rounds of training games (with no bonus regarding the game consequence) then three rounds of gambling games. This is to make the task short enough to avoid boredom. Among the three rounds, two are of *high suspense* and one was a *no suspense* game. The order of games were all counterbalanced. The sign of cards and requisite bound values were also counterbalanced (for example, "cards sum must not exceed 3" was flipped to "card sums no smaller than -3" for half of participants).

Results

Given that each subject may use the scale differently, we *z*-scored the raw suspense ratings for each subject for all analyzes except the likelihood analysis. We also collapsed across the counterbalanced conditions of positive and negative card values. Figure 3 shows a detailed summary of the model predictions and point-by-point empirical suspense ratings for each of the games.

To first assess if the *no suspense* and *high suspense* game type altered people's ratings we ran a paired t-test for each participant's averaged suspense rating from the *no suspense* vs *high suspense* games. The suspense level in the *high suspense* games (0.1 ± 0.3, Mean $\pm$ SD) is significantly higher than the *no suspense* games (-0.7 ± 0.7): $t(262) = 14.18, p <$ .001, verifying the basic effectiveness of this very heavy-handed manipulation. Visual inspection of Figure 3 confirms this as well. Participants responded with the lowest increment on the scale for 71.9% and 53.4% of the two *no suspense* games.

To study the direction and magnitude of suspense differences for identical card sequences under different rules, we computed the average z-scored rated response for each point in each of the high suspense games and calculate the difference between in the two rules, comparing this empirical difference to the difference in suspense generated by the model. In Figure 4 we see that most point differences are in the same direction (quadrant 1 and 3). The self-reported suspense difference has an correlation coefficient of $r = 0.80$ ($p = 0.01$) with the model with zero free parameters which is impressive given the inherently noisy measurements of self-reported suspense.

Alternative models

So far we have focused on the formulation of suspense proposed by Ely et al. (2015). In this last section we explore alternatives that may also capture the empirical patterns in suspense.

Alternative probability distance metrics To measure the expected belief change, Ely et al used a squared distance between probabilities while alternative metrics such as information gain and absolute change are common in other contexts (Nelson, 2005). It is unclear in the context of suspense judgment which metrics will best describe people, thus we explore these alternatives.

In the Ely et al model the suspense is defined with an L-2 norm distance for belief update:

$$S_{\mathrm{L2}} = E[(p_{t+1,i} - p_t)^2] \tag{3}$$

where $i = 1, 2$ for each possible card to be drawn and $E[\cdot]$ denotes the average over i.

We explore alternative metrics to quantify the belief update with a KL norm:

$$S_{\mathrm{KL}} = E[KL(p_{t+1,i}, p_t)] \tag{4}$$

an information gain norm:

$$S_{\mathrm{IG}} = E[\mathrm{IG}(p_{t+1,i}, p_t)] \tag{5}$$

$$= E[H(p_{t+1,i}) - H(p_t)], \tag{6}$$

and an absolute error norm

$$S_{\mathrm{L1}} = E[abs(p_{t+1,i} - p_t)] \tag{7}$$

Uncertainty The second theoretical proposal is that people may feel more suspense simply when they have high uncertainty or the estimated chance of winning is close to 1/2. In studies of drama, to keep the story captivating, it has been proposed that "the protagonist and the obstacles he encounters must be fairly evenly matched" (Mabley, 1972). Also in the realm of psychology, uncertainty has been found to sustain attention since people demand the reduction of uncertainty (Berlyne, 1960). By looking at our post-task questionnaire, we also found that around 10% of participants reported they define suspense with uncertainty (although it is unclear whether they use this term in the mathematical sense).

Uncertainty should be the highest when the probability of winning is 0.5 and lowest when it is 0 or 1. To capture this idea, we use the entropy of the belief distribution:

$$S_{\mathrm{uncertainty}} = \mathrm{H}(p_t) \tag{8}$$

Suspense when close to losing The last alternative theory is that people may feel more suspense if the negative outcome is very likely to happen or the estimated chance of winning is close to 0. Previous studies in film narratives (Comisky & Bryant, 1982) and sports viewing (Knobloch-Westerwick et al., 2009) both empirically found that when there is a bigger

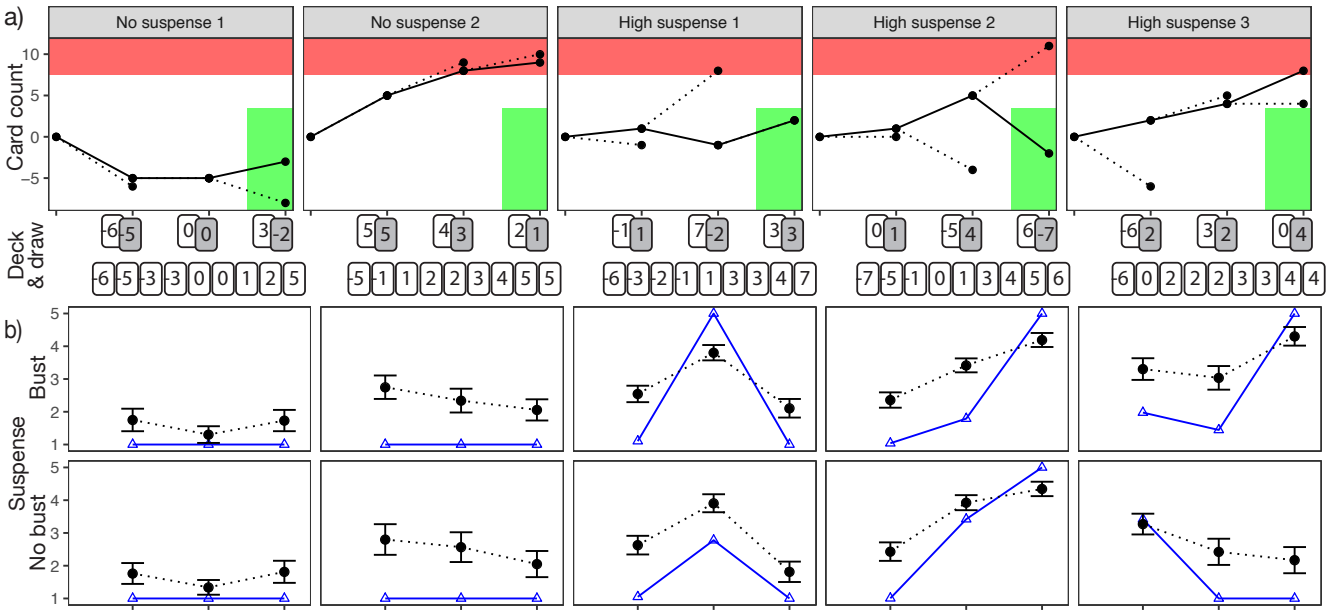


Figure 3: a) Stimuli: Panels show the five games; red indicates the bust region for *Bust* rule and green indicates win region for *No-bust* rule. The card pair at each turn is shown on x-axis with final draw in gray and the full game deck is shown below. Black lines show the actual score and dotted lines show potential score if the alternative card is drawn. b) Results and model predictions: Black circles show M±SE for participants with rule type separated by row. Blue triangles show Ely et al. (2015) predictions scaled to the full response range.

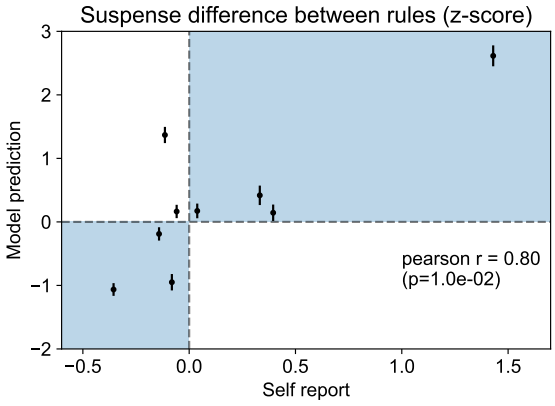


Figure 4: Suspense difference under two rules. Mean±SE differences in z-scored judgments (x-axis) scattered against z-scored model predicted differences (y-axis). Reported suspense differs in the direction the model predicted differences for points in the 1st and 3rd quadrants.

chance for the unwanted outcome to happen, more suspense is felt. In our data we also found this hint: for example, in the two *no Suspense* games, people feel more suspense in *no Suspense* 1 (-0.9 ± 0.5), where the chance of losing is always low than in *no Suspense* 2 (-0.5 ± 0.7) where the card total is always close to the bound and it indeed ends up losing. The difference of average suspense between the two games being significant ($t(261) = -4.98, p < 0.01$) indicates that people may feel more suspense when there is a high chance of losing.

We introduce two models to estimate this "pessimistic" belief about how close one is to losing the game. First, consider a heuristic: how far is the largest of the two cards drawn from the deck is from the boundary:

$$S_{\text{toBound}} = \begin{cases} 1 - |\langle V \rangle_{t+1,i} - \text{bound}|/M \\ 0, \text{if} |\langle V \rangle_{t+1,i} - \text{bound}| > M \end{cases} \quad (9)$$

Where $|\cdot|$ denotes absolute value, $i = 1,2$ representing the card pair and M is the maximum card value (7 in the current design). This piecewise definition assigns zero suspense when the current card sum is too far away from the boundary.

The other model is belief-based which is how big is the probability of losing:

$$S_{\text{pLose}} = \begin{cases} 1 - p_t, \text{if} p_t > 0 \\ 0, \text{if} p_t = 0 \end{cases} \quad (10)$$

$p_t = 0$ represents there is no hope of winning at all thus no suspense.

Likelihood model for fitting discrete responses Fitting the raw suspense scores requires an additional response

Table 1: Model Fits

	Aggregate	Individual	N best fit
L2 (Ely et al)	8.70	0.82	7
L1	**10.06**	**1.00**	**131**
KL	5.97	0.23	13
IG	8.62	0.83	14
toBound	6.65	0.03	40
pLose	8.36	0.20	6
uncertainty	7.47	0.55	52

Note: 1st column: Log likelihood improvement for each fit relative to baseline for average subject judgments (rounded into an integer response). 2nd column: Mean individual log likelihood improvement under optimal shared parameterization. 3rd column: Number of subjects best fit by each model under optimal shared parameterization. Best fitting model indicated in bold

model to convert the continuous suspense predictions to a integer output in the range of 1 to 5. We treat the response as a multinomial sampling process, with the probability of choosing each value related to a beta distribution:

$$p_k = \int_{(k-1)/5}^{k/5} pbeta, k = 1,2,\ldots,5 \tag{11}$$

whose beta parameters are defined such that the mean of beta distribution is equal to the suspense prediction (scaled to [0,1]):

$$pbeta(x) = \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}, \tag{12}$$

where

$$a = A * \tilde{S} + 1, b = A * (1-\tilde{S}) + 1 \tag{13}$$

$A \in [0,\infty)$, and $\tilde{S}$ is the suspense S rescaled to $[0,1]$.

We define our baseline model where all p_k are equal equivalent to choosing each response randomly. All the model log likelihood results in Table 1 are improvements from this baseline.

For individual participant data we fit this model with A determined by fminbound function of scipy package ($A \in [0,\ldots,15]$). We compare this maximum log likelihood to that from baseline model and summarize over all subjects. The result of all model comparison is in Table 1.

Our model fitting suggests there is considerable heterogeneity in what drives self-reported suspense in this task. The belief based suspense model with linear belief update distance (L1 norm) fit best overall, suggesting that Ely et al's choice of predictive variance may not be the most natural way of capturing human suspense. However all of the models we considered received some support, with the L2 and information gain models fitting almost identically. Consistent with the self reports in which some participants reported suspense in proportion to their current uncertainty, 20% of individual subjects were best fit by the *uncertainty* model, while a further and 15% best fitted by the heuristic "distance to boundary" model, indicating another potential heuristic sub population distribution.

General Discussion

In this study we designed a paradigm to manipulate the revelation of information about if a player will win a game (and thus earn a monetary bonus) in order to modulate participant's subjective feelings of suspense. We used the model and a computer aided search to select game sequences and rules with high predicted differences in suspense. To our knowledge, this is the first such empirical evaluation of the Ely et al. proposal.

By comparing a range of model variants, we found that most participants were fit by a model that related the rating of suspense to the anticipation of belief change, in line with Ely et al. (2015). However, we also found that belief variability predictions may be better explained by potential absolute (L1) change rather than variance. Heuristic models such as "probability to an unwanted outcome" also captured subsets of the participants.

In sum, this study suggests that suspense is systematically related to meta-cognitive predictions of future belief change. Such preposterior planning (Raiffa, 1974) issues arise in active learning and control contexts. For example, in order to identify the most useful query, one should consider the possible answers one might receive under different possible queries, and how one's beliefs would change as a result (Nelson, 2005). Suspense is thus a quantity that is tantalisingly closely related to such prospective meta-cognition, yet also distinctly low level in that it manifests as a reportable affective state.

Future iterations of our paradigm can be readily adapted to test other interesting hypotheses about suspense, such as the influence of (perceived) control, positive or negative rewards, or the role of suspense in driving attention or engagement (Bezdek et al., 2015).

References

Berlyne, D. E. (1960). *Conflict, arousal, and curiosity.* New York, NY, US: McGraw-Hill Book Company.

Bezdek, M. A., Gerrig, R. J., Wenzel, W. G., Shin, J., Revill, K. P., & Schumacher, E. H. (2015). Neural evidence that suspense narrows attentional focus. *Neuroscience*, *303*, 338–345.

Comisky, P., & Bryant, J. (1982). Factors Involved in Generating Suspense. *Human Communication Research*, *9*(1), 49–58.

Cupchtk, G. C., Oatleyb, K., & Vorderee, P. (1998). Emotional effects of reading excerpts from short stories by James Joyce. , 15.

Ely, J., Frankel, A., & Kamenica, E. (2015). Suspense and surprise. *Journal of Political Economy*, *123*(1), 215–260.

Gerrig, R. J., & Bernardo, A. B. I. (1994, December). Readers as problem-solvers in the experience of suspense. *Poetics*, *22*(6), 459–472.

Gureckis, T. M., Martin, J., McDonnell, J., Rich, A. S., Markant, D., Coenen, A., … Chan, P. (2016). psi-

turk: An open-source framework for conducting replicable behavioral experiments online. *Behavior research methods*, *48*(3), 829–842.

Knobloch-Westerwick, S., David, P., Eastin, M. S., Tamborini, R., & Greenwood, D. (2009). Sports Spectators' Suspense: Affect and Uncertainty in Sports Entertainment. *Journal of Communication*, *59*(4), 750–767.

Mabley, E. (1972). *Dramatic construction; an outline of basic principles: followed by technical analyses of significant plays by sophocles... and others*. Chilton Book Co.

Nelson, J. D. (2005). Finding useful questions: On Bayesian diagnosticity, probability, impact, and information gain. *Psychological review*, *112*(4).

Raiffa, H. (1974). Applied statistical decision theory.

Individual Differences, Expertise and Outcome Bias in Medical Decision Making

Aron Liaw[1], Matthew B. Welsh[2], Hillary Copp[1] & Benjamin Breyer[3]
([aron.liaw; hillary.copp; benjamin.breyer] @ucsf.edu, matthew.welsh@adelaide.edu.au)
1. Benioff Children's Hospital, UC San Francisco, 1975 Fourth St. San Francisco, CA 94158 USA
2. Australian School of Petroleum, University of Adelaide, North Terrace, Adelaide, SA 5005 Australia
3. Zuckerberg San Francisco General and Trauma Center, 1001 Potrero Ave, San Francisco CA 94110 USA

Abstract

Outcome bias describes the tendency of people to alter their rating of a decision's quality according to whether the outcome is good or bad – despite equivalencies in available information and decision processes – which has the potential to undermine learning about causal structures and diagnostic information in many fields, including medicine. Herein, a sample of 181 doctors and medical students is shown to display outcome bias in medical and non-medical scenarios – with their susceptibility correlating across the domains, r = 0.38. Analyses showed that rational and intuitive decision styles and a medical risk tolerance measure offered little predictive power. Instead, the strongest drivers of bias susceptibility were the Age and professional Level of participants, with more senior personnel showing less outcome bias. We argue that this could reflect improved learning across a doctor's career or result from increasing confidence making them less likely to change their initial judgement of decision quality.

Keywords: medical decision making; outcome bias; individual differences; expertise; decision style.

Introduction

Outcome bias (Baron & Hershey, 1988) describes people's tendency to judge decision quality by outcome rather than the quality of the decision making process. Baron and Hershey demonstrated this across five studies, starting with an experiment where people judged the quality of pairs of decisions about medical treatment that differed only in terms of whether the treatment succeeded or failed. That is, the background information and the decision made remained the same but the outcome differed. The key finding was that almost half of participants rated the decision made in the good outcome scenario as superior to the same decision when a bad outcome occurred (with most of the remainder giving the same rating and a handful rating the good outcome decision as worse). This was despite a within-subjects design, which maximises the chance of participants working out what an experiment is about and remembering their answers to previous scenarios. Participants' own statements also indicated that the outcome *should* not affect ratings of decision quality.

Outcome bias has since been demonstrated in different fields; for example, ethical decision making (Gino, Moore, & Bazerman, 2009), where people's condemnation of ethic breaches is weighted according to the harm done rather than the nature of the ethical breach.

It is distinguished from the similar hindsight bias (Fischhoff & Beyth, 1975) in that outcome bias affects judgements of how good the decision process was, while hindsight affects people's ratings of how likely or predictable the outcome was. (That said, these processes can be linked in situations where, having seen the outcome, hindsight bias leads to the conclusion that the person making the decision should have been able to predict the outcome and thus that their decision making was flawed.)

Of course, judging decisions by their outcomes is natural– particularly given that we often can not access other people's decision-making processes, only the outcomes of their decisions. Thus, we need to *infer* their decision processes (Gino et al., 2009). The fact that people show outcome bias in circumstances when they are specifically made aware of others' decision process and even for their own decisions, however, indicates a problem in decision making - specifically, the overuse of the generally applicable rule that outcomes are linked to decision quality.

Outcome Bias in Medical Decisions

As noted above, the original outcome bias paper used medical scenarios amongst its materials but was conducted on an undergraduate student population. Follow-up work, however, has looked directly at whether medical practitioners are affected by this bias. For example, Caplan, Posner and Cheney (1991) demonstrated anesthesiologists' ratings of the appropriateness of care provided by other medical practitioners was affected by the outcome of that care not just the quality of the decision about treatments.

Similarly, Sacchi and Cherubin (2004) found outcome bias affected doctors' judgements regarding the quality of their own diagnostic decisions and pointed out the difficulties this causes for doctors trying to learn from their own experiences – as good outcomes can artificially inflate confidence while bad luck can deflate it. In either case, background knowledge can be updated incorrectly – inferring causal relationships from random effects.

The problem of learning from experience in the face of outcome and hindsight biases has also been raised for nurses (Jones, 1995) and is key to answering the question of whether these biases can be overcome in order to improve medical decision making.

Experience and Individual Differences

A gap in the above research is in the examination of experience and other individual differences on doctors' outcome bias susceptibility. As noted above, outcome and hindsight biases make learning from experience difficult and

it is, therefore, valuable to consider whether experience helps eliminate or exacerbates these biases. No previous studies, however, include doctors' experience as a covariate.

A related question is whether the level of outcome bias shown by doctors on medical and non-medical decisions is similar. If so, this would argue for a general propensity within an individual towards (or away from) outcome bias, which could be linked to personal traits. If not, however, it may be that outcome bias is domain specific – its strength determined by prior experience within a field.

A second line of enquiry is whether there are traits that predict susceptibility to outcome bias. While range truncation in such a highly selected population is likely to prevent measures of intelligence from being useful predictors, it is possible that decision styles (a person's preference for how to make decisions; see, e.g., Hamilton, Shih, & Mohammed, 2016) could affect the level of outcome bias shown. Gino et al (2009) argue exactly this in the context of ethical decision making – that a rational mindset helped to overcome outcome bias. This makes sense particularly for a within-subjects design, where more rational participants could be more likely to notice the pairs of outcome bias scenarios and may feel a greater propensity for ensuring that they are consistent across scenarios.

Another possible covariate is a doctor's tolerance for risk (see, e.g., Grol, Whitfield, De Maeseneer, & Mokkink, 1990). While this may not directly affect outcome bias, it could do so indirectly - by pushing a participant's responses towards the floor or ceiling of a rating scale, thereby potentially preventing outcome bias. For example, if a doctor is particularly risk averse, they could judge a scenario as too risky and thus a bad decision even when it has a good outcome, leaving no space for them to judge it as worse when it occurs with a bad outcome.

Aims and Objectives

The aims of this study are, thus, to: compare doctors' susceptibility to outcome bias on generic and medicine-specific questions; explore whether and how this susceptibility is related to individual traits; and to establish whether outcome bias susceptibility varies across different groups of participants in a meaningful way.

Methodology

Participants

Participants were medical students and practitioners, recruited via Facebook and direct emails to ACGME accredited departments of 100 institutions around the US (universities and large medical groups). In total, 181 completed responses were obtained. Table 1 summarises the participant demographics.

Materials

An online survey was developed in UCSF's Qualtrics, asking participants for demographics and measuring predictor variables and outcome bias as detailed below.

Table 1. Participant demographics

Gender	114 F, 60 M & 7 no-response
Level	66 students, 22 residents, 12 fellows, 56 attendings & 25 no-response
Experience	M = 9.1 years (*SD = 13.2);* 16.9 years (SD = *13.8)* excluding students
Age	21 x '18-25'; 40x '26-35'; 27 x '36-45'; 33 x '46-55'; 59 x '56+'; and 1 x no-response

Demographics. Participants provided their gender, age range, level, years of experience and medical specialty.

Predictor Variables. Two measures with the potential to predict bias susceptibility were included in the survey:

Decision Styles Scale (Hamilton et al., 2016). The DSS is 10-item questionnaire that measures people's preferences as to how they make decisions on separate Rationality and Intuition subscales. Scores on each subscale can range from 5-25 and, in both cases, higher scores reflect greater comfort with decisions being made in that style.

Medical Risk Tolerance Scale (Grol et al., 1990). The MRTS is a 5-item response scale assessing medical practitioners' tolerance for risk in medical decisions. Scores range from 5-25, with lower scores reflecting greater tolerance for risks. Herein, however, we have reversed the scoring such that high values reflect higher risk tolerance.

Outcome Bias Questions. Nine decision scenarios were written for this experiment to enable testing for outcome bias – six describing simple, betting scenarios and three describing medical decisions. (While more scenarios could provide a finer measurement of an individual's degree of outcome bias, this was weighed against limiting the length of the survey in order to maximise responses.)

Betting Scenarios. The basic structure of the betting scenario questions was as follows, with participants responding on a 5-point, 'Very bad' to 'Very good' scale.

Your friend is playing a simple game. He has the choice to not bet on a coin flip, and automatically win $10, or bet on the coin flip and win $15 if it comes up heads, but nothing for tails. He chooses to bet. The coin comes up heads and he wins, gaining $15. In your opinion, how good a decision was this?

In all variant scenarios, the friend ignores the certain $10 and bets on the coin toss. The pay-offs and whether the outcome was good or bad varied as shown in Table 2.

This gives three pairs of questions with the same decision quality (good, neutral or bad, based on simple, economic calculation when compared to the certain, $10 option). Differences between responses to these pairs thus reflect the impact of the outcome of people's responses (outcome bias).

An individual's level of outcome bias is measured as the sum of these differences - that is: (GDGO-GDBO) +

(NDGO-NDBO) + (BDGO-BDBO), yielding scores from -12 to 12 with scores above zero reflecting outcome bias.

Table 2. Decisions scenarios

Code	Outcomes of bet	Decision	Actual outcome
GDGO	$0 or $40	Good	$40
GDBO	$0 or $35	Good	$0
NDGO	$0 or $20	Neutral	$20
NDBO	$0 or $20	Neutral	$0
BDGO	$0 or $15	Bad	$15
BDBO	$0 or $15	Bad	$0

Note: the codes are anagrams. E.g., GDGO = good decision, good outcome. The difference between payoffs in two good decision scenarios was an uncorrected error but analysis suggested it had little impact on results.

Medical Scenarios. Three scenarios were written for this study. In each, a patient opts for a surgery rather than non-surgical management of their condition. Given their length, they are summarized in Table 3 rather than described in full.

Table 3. Medical Scenarios

Patient	Surgery	Risk	Outcome
♀24yr	Pacemaker	Low	Successful surgery
♀42yr	Panniculectomy	Low	Major complications
♂72yr	Hip replacement	High	Successful surgery

As these were written to be realistic for a sample of medical professionals, they are not as easily categorized as the simple, betting scenarios – with the riskiness (and thus the 'goodness' of the decision) depending not on simple probabilities but interpretations of patient history. However, the authors' view (on writing them) was that they corresponded most closely with GDGO, GDBO and BDGO situations, which allows two comparisons: (GDGO-GDBO) as per the above; and also (BDGO-GDBO), which represents the strongest test of outcome bias. As for the simple outcome bias, medical outcome bias was calculated from sum of these two scores, yielding a possible score of -8 to 8, with scores above zero reflecting outcome bias.

Procedure

The Facebook and email invitations included a direct link to the survey, allowing participants to take part without direct contact with the experimenters. The survey started with a standard consent request before proceeding to demographics, then the DSS and MRTS. Finally, the nine outcome bias questions were presented – intermixed to limit direct comparisons between the betting scenarios.

Results

Figure 1 shows participants' mean ratings of decision quality on the Betting scenarios with 95% confidence intervals. This serves as an initial proof of concept – demonstrating that participants recognised differences between good, bad and neutral decisions but were also affected by outcome bias – as scenarios with good outcomes are consistently rated higher than their matched, bad-outcome scenarios. (NB – the smaller number and greater difficulty in designating good versus bad in the medical scenarios meant that a similar figure would not be helpful.)

Figure 1. Mean responses on Betting questions

Descriptive Statistics

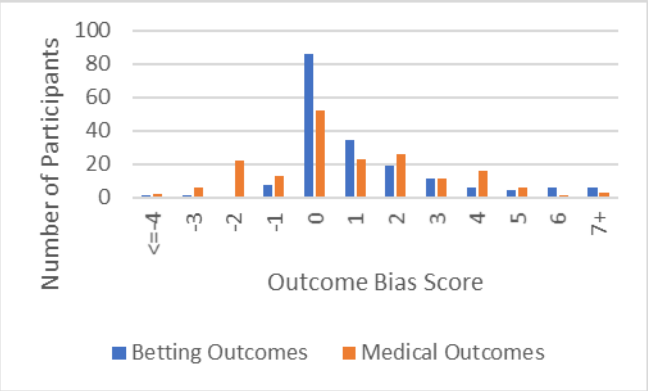
Table 4 summarises descriptive statistics for the individual differences measures and the two measures of outcome bias.

Table 4. Descriptive Statistics

Measure	Range	Mean	SD
Rationality (DSS)	10-25	19.7	3.3
Intuition (DSS)	5-24	11.3	3.4
Risk Tolerance (MRTS)	5-25	12.7	3.7
Outcome Bias (Betting)	-8-11	1.2	2.2
Outcome Bias (Medical)	-4-8	0.8	2.2

NB – the DDS and the MRTS are measured on 5-25 scales. The outcome bias measures are measured on -12 to 12 and -8 to 8 scales for Betting and Medical respectively.

Figure 2. Frequency counts of outcome bias scores



Looking at the table, one can see that both outcome bias measures, although low, are positive – as expected after seeing Figure 1. This impression is strengthened on examination of Figure 2, depicting the distribution of individual's scores. While a number of participants score around zero, there is a right skew, with more participants scoring above zero than below. Overall, 86 of the 181

participants have positive scores reflecting outcome bias for each of the Betting and Medical outcome bias measures (with 86 and 52 scores of zero and 9 and 43 scores below zero, respectively). These are similar proportions to those reported by Baron and Hershey (1988).

Outcome Bias

To test the significance of the above observations, single sample t-tests compared participant results to the expected score of zero if no outcome bias were present. These confirmed that outcome bias scores in both cases are significantly higher than zero, $t(180) = 7.39$ and 4.89, for the Betting and Medical questions respectively, $p < .0001$ in each case. A Pearson correlation was also calculated between the two outcome bias measures, indicating a moderate correlation, $r(179) = 0.38$, $p < .0001$, suggesting a stable tendency for people to show outcome bias (or not) regardless of the scenarios used. This suggests that, despite differences between scenarios, overall outcome bias susceptibility could be calculated in future work.

Individual Differences

Table 5 shows the Pearson correlations between the three individual difference measures and the Betting and Medical outcome bias scores.

Table 5. Pearson correlations between predictor and outcome bias variables

	1	2	3	4	5
1. Rationality	-		*		
2. Intuition	-.02	-	*	*	
3. Risk Tolerance	**-.17**	**-.19**	-		*
4. Betting	-.05	**.16**	-.09	-	***
5. Medical	-.10	.07	**-.19**	**.38**	-

* - sig. at .05 level, 2-tailed; *** - sig. at .001 level, 2-tailed

In Table 5, relationships between the predictor variables and the outcome bias measures are weak but a number are statistically significant. Specifically, Intuition correlates positively with people's Betting outcome bias score, while Risk Tolerance correlates negatively with Medical outcome bias. That is, people with more belief in their own intuitions and less tolerance for risk (or a greater desire to consult with others) seem to have a weak tendency to show more outcome bias (in the Betting scenario).

Overall, however, the results provide little hope for those seeking to use these individual differences to predict levels of outcome bias, with the strongest relationship explaining less than 4% of the variance in outcome bias scores.

Finally, analyses looked at participants' raw responses on the 1-5 ratings across both the Betting ($M = 2.95$, $SD = 0.65$) and Medical ($M = 3.63$, $SD = 0.58$) questions. This established that participants tended to think the medical decisions were better overall but is reassuring in that the majority of results are clear of floor and ceiling in both cases. Comparison of participant's mean ratings with their Risk Tolerance also found no correlation - $r = .04$, $p > .05$ in both cases - undermining the suggestion that risk tolerance might contribute to floor or ceiling effects.

Group Differences

Further analyses were undertaken to determine whether demographic differences between the participants predicted outcome bias or differences in the predictor variables.

Gender

Table 6 shows the data divided by gender.

Table 6. Mean (and *SD*) of measures by gender

	Female (n=114)	Male (n=60)
Rationality	20.0 (*3.37*)	18.6 (*3.63*)
Intuition	10.8 (*3.21*)	10.9 (*3.15*)
Risk Tolerance	12.3 (*3.74*)	13.7 (*4.26*)
Betting	1.54 (*2.39*)	0.58 (*1.62*)
Medical	1.08 (*2.46*)	0.35 (*1.79*)

Looking at Table 3, males and females score similarly on the individual difference traits but show clear differences in terms of the extent to which they show outcome bias, with females showing the bias at higher rates in both the Betting and Medical conditions. Independent samples t-tests were used to assess the significance of these apparent trends. These confirmed that the differences between male and female scores on the DSS Rationality and Intuition measures were not significant. Differences in Risk Tolerance, however, were, $t(172) = 2.3$, $p = .023$ (two-tailed), with males showing higher risk tolerance.

Similarly, the differences in outcomes bias were significant for both the Betting and Medical questions, $t(172) = 2.8$ and 2.0, $p = .006$ and .044, respectively, with males showing less bias in both cases.

Practitioner Level

Table 7 shows the data divided according to the level of the participants (as medical practitioners).

Table 7. Mean (and *SD*) of measures by practitioner level

	Student	Resident	Fellow	Attending
Rationality	20.7	18.7	19.6	19.8
	(*3.0*)	(*3.3*)	(*2.6*)	(*3.3*)
Intuition	11.6	11.6	11.6	10.0
	(*3.4*)	(*2.9*)	(*5.1*)	(*2.6*)
Risk Tol.	11.3	13.8	12.2	14.2
	(*2.9*)	(*3.9*)	(*4.4*)	(*3.9*)
Betting	1.8	1.1	1.1	0.6
	(*3.0*)	(*1.8*)	(*1.1*)	(*1.6*)
Medical	1.2	0.1	1.4	0.3
	(*2.2*)	(*2.7*)	(*2.4*)	(*2.0*)

Note: n = 66, 22, 12 and 56, respectively.

The table shows noticeable differences between the groups on a number of measures. In particular, Attending physicians seem to show less trust in their Intuition, higher risk tolerance and less outcome bias, while Students tend to

lie at the opposite extremes on these measures. The Resident group also shows extremely low outcome bias on the Betting scenarios but, given the very small size of this group, the reliability of the result is questionable.

One-Way ANOVAs were conducted in SPSS, comparing the groups' mean performance across all five measures. These confirmed significant differences between groups for: Intuition; Risk Tolerance; and Betting; $F(3, 152) = 2.67$, 7.45 and 2.91, $p = .050$, $<.001$ and .036, respectively. The other ANOVAs just failed to reach significance $F(3, 152) = 2.58$ and 2.55, $p = .056$ and .058, for Rationality and Medical outcome bias, respectively. Bonferroni post-hoc tests confirmed that significant results were driven by differences between the Attending and Student groups.

Given the effect of practitioner level on results, a χ^2 test was conducted to see whether a relationship between practitioner level and gender was driving the gender effect observed above. This revealed a significant relationship between gender and level, $\chi^2(3) = 10.1$, $p = .014$, with the sample containing more female Students and fewer female Attendings that would be expected based on the overall gender/level breakdown. Thus, multiple regressions (described below) were required to tease these effects apart.

Medical Specialty

Participants listed many specialties – making analysis difficult given space and power constraints. A result that stood out, however, was the difference between surgical and non-surgical specialties. Specifically, despite similar Risk Tolerance scores, surgical specialties (defined as those that make decisions in the operating room on a regular basis, including surgical specialties and anaesthesia) rated the decision to undergo the higher risk surgery (i.e., the bad decision, good outcome Medical scenario) as a worse decision than did non-surgical specialists, Mdiff = -0.42; confirmed as significant by an independent samples t-test, $t(102) = 2.0$, $p = .048$. As a result, the surgeons, overall, did not display outcome bias on this question.

Predicting Outcome Bias

In light of the multiple relationships shown above, linear regressions were run in SPSS using the Forward entry method ($p = .05$ inclusion criterion and $p=0.1$ removal criterion) using Age (converted to a 1-5 scale), Decision Making Training (0 or 1), Experience, Gender (converted to a 0 or 1 scale), Level, Rationality, Intuition and Risk Tolerance on Betting and Medical outcome bias scores. Tables 8 and 9, below, show the models produced for the Betting and Medical outcome bias scores, respectively.

Examination of these tables shows that both produced significant models (albeit with low proportions of variance explained at 7.9% and 11%) with the same predictors for the Betting and Medical versions of outcome bias - Age and Level. Participant's Medical scores were also affected by their Risk Tolerance score. Specifically, the models suggest that participants at higher Levels tend to show *less* outcome bias despite a tendency for older people to show *more*. Greater medical Risk Tolerance also decreased outcome bias, but only for the Medical outcome bias questions.

Table 8. Regression model for Betting scores

Model
Significant Predictors: Level and Age Formula: Betting = 1.74 – 0.96*Level + 0.77*Age $F(2, 149) = 7.48$, $p <.0001$; Adj $R^2 = .079$
Note: regression conducted using forward entry method. Standardised βs = -.297 (Level) and .204 (Age).

Table 9. Regression model for Medical scores

Model
Significant Predictors: Risk Tolerance, Age and Level Formula: Medical = 1.59 – 0.11* Risk + 0.40*Age - 0.329*Level $F(3, 148) = 7.23$, $p <.0001$; Adj $R^2 = .110$
Note: regression conducted using forward entry method. Standardised βs = -.189 (Risk) .253 (Age) and -.198 (Level).

Interestingly, once the effects of Age and Level are partialled out, the gender differences do not reach significance in either model. Neither is previous decision training or either of the DSS measures (Rationality and Intuition) having a significant effect.

Discussion

The results presented above reconfirm the existence of outcome bias in doctors and medical students and add to this knowledge in a variety of ways.

Firstly, the stability of outcome bias across scenarios of different types was established, with participants' outcome bias on Betting and Medical scenarios correlating significantly together. This supports the idea that there could be particular traits that predict the degree of outcome bias an individual will show.

Our analyses, however, failed to support the finding from Gino et al (2009) that a rational mindset decreases outcome bias. This may, however, simply reflect a range truncation effect, with participants' Rationality scores tending towards the higher end of the scale and none scoring below 10 on the 5-25 range. This is, perhaps, unsurprising, given the need for medical students and practitioners to use rational decision making and reflects a common difficulty in finding predictors of biases in highly selected populations.

The fact that Intuition emerged as a significant predictor in correlations with outcome bias does shine some light on the solution to this – the need to find traits that affect decision making but which are less strongly selected for through medical training. Intuition scores, while as low on average as Rationality scores were high may have been less truncated, spanning almost the scale's full range – from 5 to 24. The combination of range truncation and skew in these measures may also explain the somewhat surprising observation that Rationality and Intuition did not correlate

in our sample – unlike in the majority of data presented by Hamilton et al (2016) where a negative relationship is seen.

Overall, amongst the potential covariates examined herein only a handful of weak relationships were shown. Overall, the decision styles and Risk Tolerance measures showed little predictive power for outcome bias and, what little they did, disappeared when demographic variables of participant Level and Age were included in regressions. This suggests that participant Level may be affecting both a person's (trust in their own) Intuition and level of outcome bias rather than Intuition directly affecting outcome bias.

Caveats and Future Research

The fact that Level and Age proved the most consistent predictors of outcome bias, combined with the correlations involving Intuition and Risk Tolerance, could indicate that doctors, across the course of their careers, are learning in such a way as to help them overcome outcome bias. Alternately, however, it may suggest that a measure of confidence (see, e.g., Stankov, Kleitman, & Jackson, 2014) could be useful predictor in future work. The idea being that more senior doctors may be performing better because they are more confident and thus less swayed away from their initial rating as to whether something is a good or bad decision by outcomes. Of course, this might apply differentially in situations where they were rating their own decisions rather than those of others, which would need to be tested as well.

This could be regarded as a Bayesian explanation of the expertise effects. Specifically, outcome bias among students could reflect weaker priors which are, therefore, more affected by the new evidence provided by the outcome. More experienced people, by comparison, could have stronger priors as a result of that experience. Such an effect could also shed light on the difference between results for the Betting and Medical scenarios. A possibility we did not consider, for example, is whether people assumed that the coin described in the betting scenarios was 'fair'. We intende for them to do so but did not specifically state it and so participants could have, intuitively, been considering the possibility that the coin was not fair – with the result that their prior beliefs were weaker than in the medical scenarios. If true, an explicit statement or demonstration of the fairness of the coin should reduce outcome bias in these cases.

Another potential trait that could be considered is Need for Cognitive Closure (Webster & Kruglanski, 1994), which measures person's tolerance for ambiguity and/or their need to quickly resolve it – the expectation being that people high in NFCC might show less outcome bias as, having made a decision, they are less likely to revisit it once the outcome becomes known. In either case, however, whether medical personnel show a truncated range on such traits will also need to be tested.

Another interesting possibility raised by the data is that surgical and non-surgical specialists interpret surgical risk differently. This could be directly examined in future work.

A potential concern regarding the data is that the lack of control over online survey data may have resulted in errors or deliberate mistakes in personal data. In particular, it was noted that participant age data was strangely distributed – with more medical students selecting an age of 56+ than seems likely at first glance. This is likely to have eroded the predictive power of Age – by adding noise to the data. Given this, it may be that Age would be a stronger predictor in a future study with greater control over participant inputs. An alternative recruitment strategy could also aid in statistical analysis by ensuring equal numbers of participants in all groups.

Additionally, while prior Decision Making training was not a significant predictor of performance in our data, future research could explore this further by requesting further details on the type of training received and when it was received – given work in other areas showing that the durability of such training can be low over the course of years (see, e.g., Welsh, Bratvold, & Begg, 2005).

Finally, while the findings suggest that outcome bias is reduced by medical expertise, additional work is required to see whether these effects replicate when considering experts in other, non-medical fields. This could shed light on which of the possible explanations described above are most likely.

Conclusions

Doctors and medical students showed outcome bias in medical and non-medical decision scenarios, rating decisions with good outcomes significantly better than those with bad outcomes. The degree of outcome bias shown in these different sets of questions was similar and correlated, indicating a stable susceptibility to outcome bias.

The individual differences traits tested herein showed little predictive power, possibly due to range truncation, but outcome bias decreased with the Age and employment Level of participants. This could represent learning across a doctor's career but could also, we suggest, relate to their overall level of confidence, which is likely to inure them against changing their opinion on what the right decision is in light of new information like outcomes.

Acknowledgments

MBW is supported by ARC LP160101460, which includes support from Santos and Woodside.

References

Baron, J., & Hershey, J. C. (1988). Outcome bias in decision evaluation. *Journal of Personality and Social Psychology, 54*(4), 569.

Caplan, R. A., Posner, K. L., & Cheney, F. W. (1991). Effect of outcome on physician judgments of appropriateness of care. *Jama, 265*(15), 1957-1960.

Fischhoff, B., & Beyth, R. (1975). I knew it would happen: Remembered probabilities of once—future things.

Organizational Behavior and Human Performance, 13(1), 1-16.

Gino, F., Moore, D. A., & Bazerman, M. H. (2009). No harm, no foul: The outcome bias in ethical judgments. *http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.451.913&rep=rep1&type=pdf*(Downloaded Jan 23rd 2019).

Grol, R., Whitfield, M., De Maeseneer, J., & Mokkink, H. (1990). Attitudes to risk taking in medical decision making among British, Dutch and Belgian general practitioners. *Br J Gen Pract, 40*(333), 134-136.

Hamilton, K., Shih, S.-I., & Mohammed, S. (2016). The development and validation of the rational and intuitive decision styles scale. *Journal of Personality Assessment, 98*(5), 523-535.

Jones, P. R. (1995). Hindsight bias in reflective practice: an empirical investigation. *Journal of Advanced Nursing, 21*(4), 783-788.

Sacchi, S., & Cherubini, P. (2004). The effect of outcome information on doctors' evaluations of their own diagnostic decisions. *Medical education, 38*(10), 1028-1034.

Stankov, L., Kleitman, S., & Jackson, S. A. (2014). Measures of the Trait of Confidence. In G. J. Boyle, D. H. Saklofske, & G. Matthews (Eds.), *Measures of Personality and Social Psychological Constructs* (pp. 158-189): Academic Press.

Webster, D. M., & Kruglanski, A. W. (1994). Individual differences in need for cognitive closure. *Journal of Personality and Social Psychology, 67*(6), 1049.

Welsh, M. B., Bratvold, R. B., & Begg, S. H. (2005). SPE 96423 - Cognitive biases in the petroleum industry: impact and remediation. *Proceedings of the Society of Petroleum Engineers 81st Annual Technical Conference and Exhibition*.

Novel categories are distinct from "Not"-categories

Shi Xian Liew (liew2@wisc.edu)
Department of Psychology, 1202 W. Johnson Street
Madison, WI 53706 USA

Joseph L. Austerweil (austerweil@wisc.edu)
Department of Psychology, 1202 W. Johnson Street
Madison, WI 53706 USA

Abstract

The categorization literature often considers two types of categories as equivalent: (a) standard categories and (b) negation categories. For example, category learning studies typically conflate learning categories A and B with learning categories A and NOT A. This study represents the first attempt at delineating these two separate types of generated categories. We specifically test for differences in the distributional structure of generated categories, demonstrating that categories identified as *not* what was known are larger and wider-spread compared to categories that were identified with a specific label. We also observe consistency in distributional structure across multiple generated categories, replicating and extending previous findings. These results are discussed in the context of providing a foundation for future modeling work.

Keywords: categorization; category generation; contrast; category learning;

Introduction

People are remarkable in their capacity to innovate new and different ideas. Is creating a new idea the same as creating a different idea? Consider a restaurant that serves one meal per night. Their chef cooked red curry last night and wants to create and cook a new dish tonight. Is that the same as wanting to create and cook a new dish that is *not* red curry? While the former is identified as its own category, the latter is identified in relation to a known category.

While categorization researchers have primarily focused their effort on classification (associating an exemplar with a category given its features), and inference (predicting exemplar features given its category), work on category generation – predicting all exemplar features for a novel category – is relatively scarce. This is surprising because category generation is not an uncommon phenomenon – people are constantly challenged to generate novel categories, such as a new meal plan for the week, a new music playlist for an upcoming road trip, or a new exercise regimen to stay healthy.

Recent category generation work has established a few key findings. Earlier studies have shown that generated categories tend to share distributional statistics with learned categories (Jern & Kemp, 2013; Thomas, 1998; Ward, 1994). More recently, Austerweil, Conaway, Liew, and Kurtz (in preparation) and Conaway and Austerweil (2017) have found that category contrast is an important factor in category generation and learning – computational models sensitive to the differences between categories were a better fit to generated categories than models which did not take categories contrast into account.

Although previous work has established a few key findings, the basic phenomena and processes involved in category generation are still not well understood. In this article, we examine whether given a known category in a domain, generating a new category is different from generating a category that is not the learned category. If the only category is A, most formal accounts of categorization would consider generating a new category B as equivalent to generating not A. In order to better understand the nature of generated categories, it is necessary to distinguish between these two possible types of generated categories: one driven by its own identity – an *independently identified category*, and another driven by a motivation to be not what is known – a *category-by-negation*.

Current models of category generation do not explicitly make this distinction. One of the first computational models of category generation, Jern and Kemp (2013)'s hierarchical sampling model, is a Bayesian model that reproduces distributionally similar categories by assuming that the covariance matrix of features of exemplars from generated categories is generated by the same prior that generated the covariance matrix of the known categories. An alternative proposed model, PACKER (Conaway & Austerweil, 2017), is an extension of the classic Generalized Context Model (GCM; Nosofsky, 1984, 1986). It explicitly incorporates contrast into the similarity function by including a penalty for being similar to exemplars from a known category. A contrast parameter allows candidate categories that are more different from the known category to be weighted more heavily than candidate categories similar to the known category. Both models are flexible enough to describe how generated categories can be distributionally similar (or different) from an experimenter-defined category. However, they do not distinguish between generating an independently identified category and a category-by-negation because there is no mechanism to account for these different identities.

This issue also extends to the methodology applied in category generation (as well as most categorization studies.) In Ward (1994), participants were told to generate aliens that belonged to a different species from a prior group of aliens, without applying any specific label to this new group of aliens. Similarly, Jern and Kemp (2013) instructed participants to generate a new, different type of crystal after having

observed two different types of crystals. In contrast, Conaway and Austerweil (2017) prompted participants to generate exemplars from a novel "Beta" category, while avoiding an explicit instruction for participants to create something "different". In each of these studies there appears to be an implicit assumption that generating an independently identified category is equivalent to generating a category-by-negation.

In this paper, we challenge this null assumption by positing that the explicit association of categories-by-negation to its known counterpart (i.e., the explicit identification of the to-be-generated category as *not* a known category) should result in the observer taking advantage of the entire area of the feature space not occupied by the known category. In contrast, observers generating independently identified categories should be less focused on the unoccupied feature space and instead construct their categories based on the structure of known categories. From this we can predict that categories-by-negation should occupy larger areas of the feature space compared to independently identified categories. In addition, the similarity in distributional statistics should extend not only from the learned category to the first independently identified category to be generated (as previous studies have found), but also between subsequent independently identified categories. We test these predictions by adapting and extending a category generation experiment by Conaway and Austerweil (2017) and explore the implications of its results.

Experiment

Our current experiment closely mirrors the the experimental design of Conaway and Austerweil (2017), where participants are first trained on a category named 'Alpha' before being tasked to generate exemplars from a new category. While Conaway and Austerweil (2017) were primarily interested in the measuring the location of generated categories relative to learned categories, our current investigation focuses on analyzing the differences in distributional statistics across different generated categories. Consequently, in addition to varying the shape and location of the Alpha category (i.e., across different Alpha conditions where the position of the Alpha exemplars are systematically varied) in three distinct ways, we include an additional independent variable comprising three different generation conditions: a Not-Alpha condition, where participants generate a new category that is not the learned Alpha category; a Beta-Only condition, where participants generate a new category named 'Beta'; and a Beta-and-Gamma condition, where participants generate a category 'Beta' as well as a category 'Gamma'. The resulting 3-by-3 design is applied in a between-subjects fashion – participants can be in only one of the nine unique conditions.

The main advantage of adapting the experiment by Conaway and Austerweil (2017) is that the simplicity of their stimuli allow for a straightforward test of the distributional similarities across known and generated categories. In addition, the variety of Alpha categories used (i.e., the different shapes of the Alpha categories) also allows us to observe the effect of generated category identity across multiple scenarios.

In line with our earlier predictions, we expect that the Not-Alpha conditions on average generate categories that are larger in area and more widely dispersed than categories from the Beta-Only as well as Beta-and-Gamma conditions. In addition, we predict that within the Beta-and-Gamma conditions, the generated Beta and generated Gamma categories should be distributionally similar.

Method

Participants and materials We recruited 240 participants through Amazon Mechanical Turk and randomly assigned them to one of the nine unique conditions. Sample sizes of each condition are presented in Table 1.

Stimuli were squares that varied along two dimensions: color (grayscale 9.8% – 90.2%) and size (3.0 – 5.8cm on each side). The assignment of perceptual features (color, size) to axes of the domain space (x, y), as well as the direction of variation along each axis (e.g., increasing or decreasing size) was counterbalanced across participants. Feature values were evenly-spaced on a 50-by-50 grid, giving a near-continuous space from which exemplars can be generated. An example of the feature space is presented in Figure 1a.

The three Alpha conditions are Cluster, Row, and Diagonal. In the Cluster condition, Alpha exemplars occupy a small area towards one corner of the feature space. In the Row condition, the exemplars are nearly equal along one feature, while equally spread out along the other feature. The Diagonal condition has Alpha exemplars equally spaced along the diagonal of the feature space in a similar fashion to the diagonal conditions of Jern and Kemp (2013). In order to ensure that exemplars are not completely identical along any one feature, exemplar feature values are slightly jittered. The exact same amount of jitter is applied to all Alpha exemplars within a given Alpha condition. The locations of these different Alpha categories in the feature space are presented in Figures 1b to 1d.

Procedure In the first phase of the experiment (Figure 2), participants learned the Alpha category exemplars by observing a unique exemplar on each trial. This was repeated over a total of three blocks (four trials per block – one corresponding to each unique exemplar,) with the order of exemplar presentation randomized within each block. Prior to the presentation of each exemplar, a fixation cross was shown for 1000 ms. Participants were allowed to spend as much time as they wanted on each trial and were also shown the full range of possible feature values prior to training.

The next phase comprised a series of generation trials (Figure 3). Depending on their generation condition, participants generated either eight exemplars from a category that was 'Not-Alpha' (Not-Alpha generation condition), eight exemplars from a category called 'Beta' (Beta-Only generation condition), or four exemplars from a category 'Beta' *and* four exemplars from a third category 'Gamma' (Beta-and-Gamma

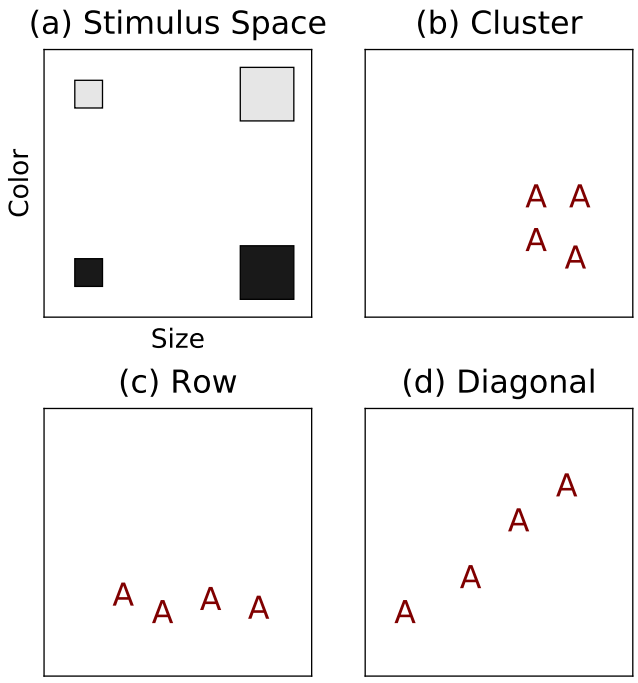


Figure 1: (a) Example of stimuli located at the corners of the feature space. (b-d) Locations of Alpha category exemplars for each Alpha condition.

generation condition). More specifically, participants in the Not-Alpha generation condition were asked to produce "what [they] think is likely to NOT be in the Alpha category", while participants in the other conditions were asked to produce "what [they] think is likely to be in the Beta [or Gamma] category". Exemplars were generated on each trial using two on-screen sliding scales, with each scale controlling the individual features (color and size) of the generated exemplar. Feature values could take any one of 50 evenly-spaced values between the specified boundaries. Previously generated exemplars were not allowed to be generated a second time. Participants were shown an on-screen preview of their exemplar on each trial as they interacted with the sliders, but could not see previously generated exemplars or exemplars from the Alpha category.

Table 1: Sample sizes for each condition.

Generation Condition	Cluster	Row	Diagonal
Not-Alpha	26	28	25
Only Beta	30	27	25
Beta-and-Gamma	26	27	26

Analyses We analyze our data in two stages. First, to provide a coarse overview of the distribution of different patterns of generated categories, we classify the generated categories into six different profiles: Positives, where the correlation between the dimensions is more than r; Negatives, where the correlation between the dimensions is less than $-r$; Rows, where the range of values across the x dimension is at least d times more than the range across the y dimension; Columns, where the range of y dimension values is at least d times more than the range of x dimension values; Clusters, where the ranges across both dimensions are less than a; and Dispersed, where the ranges across both dimensions are more than a. Next, we compare the generated categories on four key distributional measures: ranges of each feature, the feature correlations, and the area enclosed by the generated exemplars in the feature space (i.e., their convex hull). Differences along each of these statistics are performed using Bayesian t-tests (Rouder, Speckman, Sun, Morey, & Iverson, 2009), yielding Bayes factors (BF_{01}) which indicate evidence for the null hypothesis when $BF_{01} > 1$, with larger values indicating greater evidence for the null hypothesis. $BF_{01} < 1$ indicates evidence for the alternative hypothesis. Interpretations of the sizes of Bayes factors are guided by Jeffreys (1961).

Results

We took a subset of the data and tuned each of the profiling parameters such that the profiles of this subset were adequately captured. Subsequently, we applied this profiling scheme to the entire dataset. Overall, we found that setting $r = .7$, $a = .25$, and $d = 5$ was useful in capturing the different profiles of generated categories. The results were robust to moderate variations in the profiling parameters (e.g., setting $.5 < r < .9$, $.1 < a < .4$, and $d > 1$ returned very similar results.) A representative sample of each profile is shown in Figure 4 and the frequency plot of the different profiles is presented in Figure 5.

The most striking patterns to note here are the high frequencies of Row category profiles from participants in the Row conditions, and the high frequencies of Dispersed category profiles from participants in the Not-Alpha conditions. The former indicates that the distributional similarities between learned and generated categories are especially strong in the Row conditions, while the latter provides preliminary evidence that generated categories from the Not-Alpha conditions tends to be more widely dispersed. Also noteworthy are the low counts of both Positive and Negative category profiles across the whole data set – in contrast to the Row conditions, this indicates low similarity in distributional structure between the learned and generated categories for the Diagonal condition.

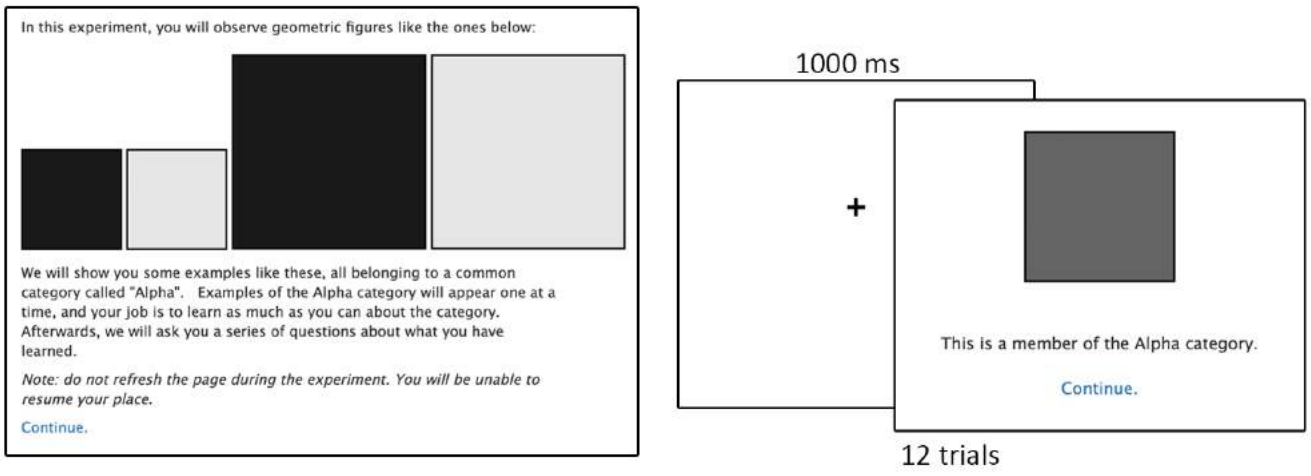


Figure 2: Instructions and trials observed by each participant during the category learning phase. The instructions screen is shown once, followed by 12 presentations of Alpha exemplars (4 exemplars across 3 blocks.)

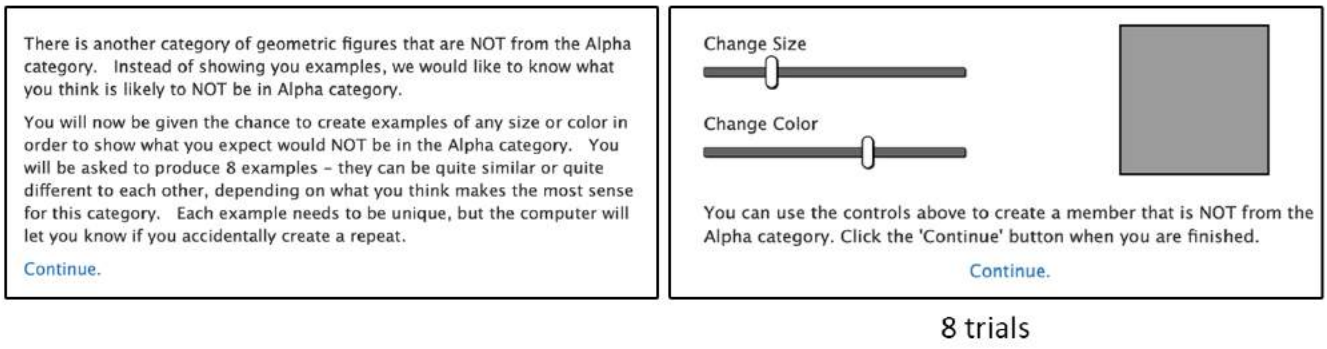


Figure 3: Instructions and trials during the generation phase, observed by a participant in the Not-Alpha condition. Participants in the Beta-Only and Beta-Gamma conditions experienced similar trials, with the exception that those in the Beta-Gamma condition were asked to generate 4 Beta exemplars then 4 Gamma exemplars.

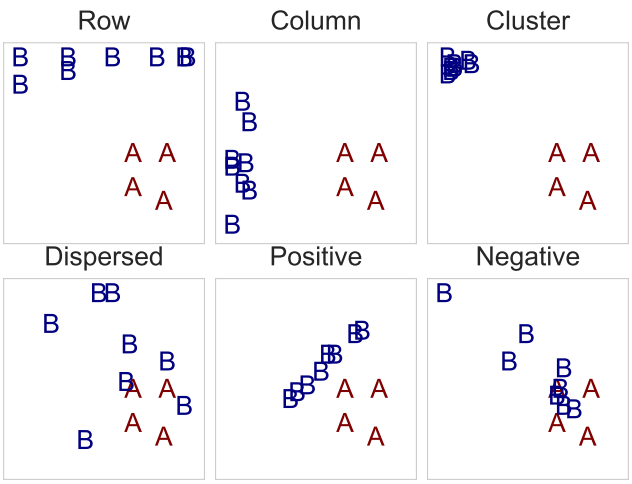


Figure 4: Representative samples of the six different category generation profiles.

Focusing on the distributional statistics, when broken down by the Alpha conditions (Figure 6), we found that the Cluster conditions tended to have categories with a smaller range of both features, and with correspondingly small sizes. The Row conditions produced categories that are high on the x dimension but not the y dimension. These observations indicate that the distributional statistics were carried over from the known category to the generated category in these two Alpha conditions. However, the Diagonal conditions did not have similar distributional statistics observed – instead of a positive correlation, the Diagonal conditions tended to produce large, dispersed categories. Austerweil et al. (in preparation) also observed a similar effect (although they found more evidence for a negative correlation). Overall, in alignment with previous research, at least two out of the three Alpha conditions in our experiment generated categories that share distributional statistics as the known category.

More interestingly, when the data is broken down by the generation conditions (Figure 7), we found that compared to the Not-Alpha conditions, there is moderate evidence showing Beta-Only conditions with a lower y dimension range ($t(162) = 3.00$, $BF_{01} = 0.16$) and moderate to strong evidence that their generated categories are smaller in area ($t(162) = 3.16$, $BF_{01} = 0.10$). There is moderate evidence that the Not-Alpha and Beta conditions share equal range of x dimension values ($t(162) = 1.33$, $BF_{01} = 4.82$). With Beta-and-Gamma conditions, we find greater evidence for smaller and tighter categories compared to the Not-Alpha conditions. Specifically, there is very strong evidence that both Beta and Gamma categories from the Beta-and-Gamma condition are smaller in both x ($t(156) = 4.83$, $BF_{01} = 2.55 \times 10^{-4}$; $t(156) = 4.21$, $BF_{01} = 2.94 \times 10^{-3}$, respectively) and y ($t(156) = 4.57$, $BF_{01} = 7.40 \times 10^{-4}$; $t(156) = 7.56$, $BF_{01} = 4.30 \times 10^{-10}$, respectively) ranges compared to the Not-Alpha conditions. Similarly, there is very strong evidence that both categories in the Beta-and-Gamma conditions are smaller in area than the Not-Alpha conditions ($t(156) = 5.70$, $BF_{01} = 5.41 \times 10^{-6}$; $t(156) = 7.69$, $BF_{01} = 2.07 \times 10^{-10}$, respectively). Overall, when comparing the Not-Alpha conditions to categories from other generation conditions, we consistently find moderate to very strong evidence that Not-Alpha categories are more widely dispersed (in their range values) and also larger (in their area), with the only exception being the comparison of x dimension ranges between the Not-Alpha and Beta-Only conditions.

When comparing the distributions of the Beta and Gamma categories generated within the Beta-and-Gamma conditions, we find an overall weak-to-moderate evidence for equal distributional statistics. Specifically, there was moderate evidence for equal x dimension ranges ($t(150) = 0.56$, $BF_{01} = 9.47$) and weak evidence for both feature correlations and area size ($t(150) = 1.85, BF_{01} = 2.10; t(150) = 1.77$, $BF_{01} = 2.40$, respectively). When measured on their y dimension ranges, we found weak evidence for lower values from Gamma categories compared to the Beta categories ($t(150) = 2.47$, $BF_{01} = 0.59$).

Discussion

At first glance, it seems reasonable to assume that generating a new category Y after learning category X is the same as generating a new category Not-X. An independently identified category should already be a category-by-negation (in that an independently identified category is not what was previously known.) Further, if the categories are already identified by arbitrary labels, then it may be easy to assume that identifying the negation of a known category cannot add any additional information in category generation.

Our results have indicated otherwise. Specifically, when tasked to produce categories-by-negation, participants tended to generate wider and larger new categories compared to when tasked with producing independently identified categories. To our knowledge, this paper represents the first piece of evidence distinguishing these separate types of categories.

Aside from demonstrating a new effect, this study has continued to show the robustness of the distributional similarities between learned and generated categories. In this sense, the results from the different Alpha conditions are similar to those observed in Austerweil et al. (in preparation). The generated categories from the Cluster condition tend to possess lower x and y ranges, with a correspondingly smaller area, and the generated categories from the Row condition tend to adopt Row-type profiles. There is also similar lack of categories with positively correlated features from the Diagonal condition.

However, one notable difference is that while Austerweil et al. (in preparation) found evidence of negatively-correlated generated categories in their XOR condition, we found no evidence of negatively-correlated generated categories in our comparable Diagonal condition. Austerweil et al. (in preparation) explained that the presence of negatively-correlated

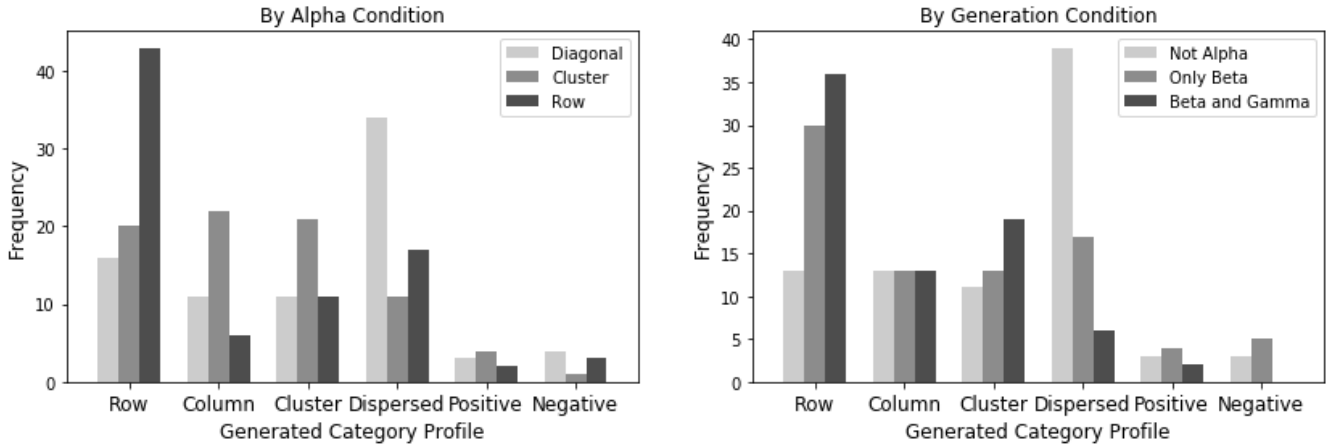


Figure 5: Frequencies of category generation profiles broken down by Alpha condition (left plot) and generation condition (right plot).

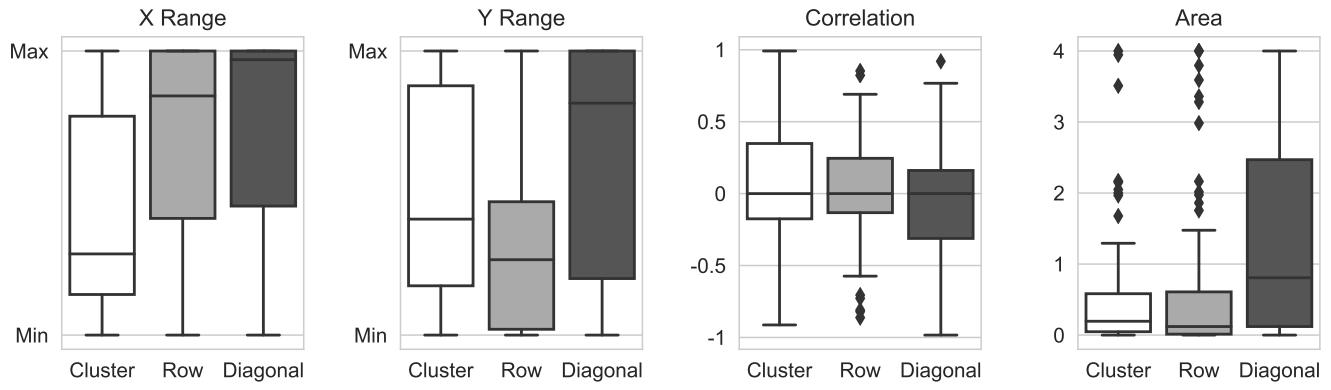


Figure 6: Box-plots of the distributional statistics from the generated categories. Boxes depict the median and quartiles of each Alpha condition, with whiskers placed at 1.5 inter-quartile range. All points outside this region are marked individually.

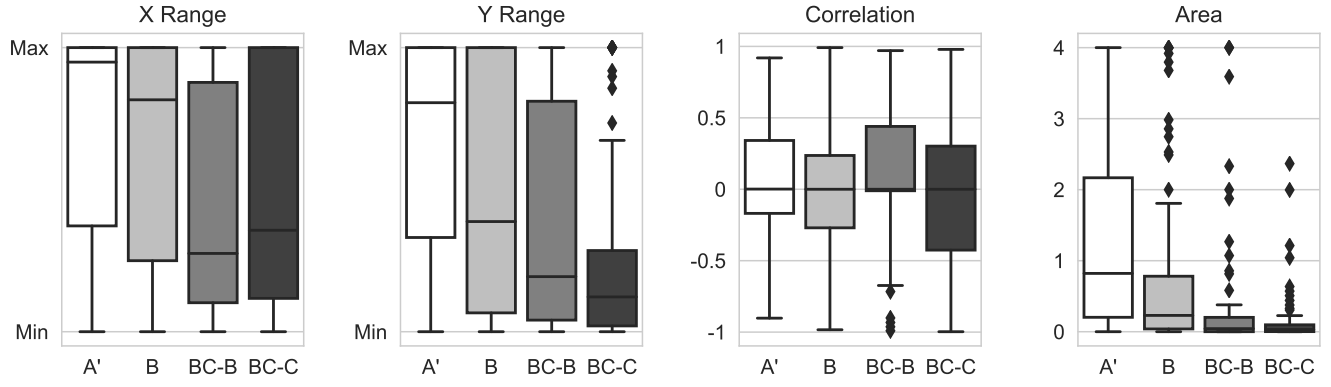


Figure 7: Box-plots of the distributional statistics from the generated categories. Boxes depict the median and quartiles of each generation condition, with the Not-A condition denoted by A' and the Beta-Only condition denoted by B. Data from the Beta-and-Gamma conditions are separated into the Beta groups (denoted BC-B) and the Gamma groups (denoted BC-C) generated in that condition. Whiskers are placed at 1.5 inter-quartile range. All points outside this region are marked individually.

features in this condition was indicative of the effect of category contrast – that is, categories with negatively correlated features are generated because they are particularly different to categories with positively correlated features. It is possible that due to the reduced strength of the positive correlation in our study compared to Austerweil et al. (in preparation) (because of the addition of noise to the feature values), participants were no longer as sensitive to the negative correlations in the experimenter-defined category and therefore started to produce uncorrelated but widely dispersed categories.

Beyond replicating previous studies, the current study has demonstrated that the consistency in distributional statistics can persist beyond the first generated category. However, evidence showing this was ultimately weak. One possible reason for this is the relatively small feature space employed in the tasks. The generation of a second novel category is necessarily more constrained in the feature space than the generation of the first novel category, possibly contributing to differences in distributional structure. By exploring stimuli features with less defined boundaries (e.g., orientation), we may expect to see greater consistency in distributional structure over multiple generated categories.

Although we have observed participants generating multiple independently identified categories, we do not want to imply that categories-by-negation can only happen once. It would be worth investigating how participants might proceed to generate additional categories-by-negation (e.g., by asking observers to generate a category that is Not-Alpha and Not Beta). Packing Theory (Hidaka & Smith, 2011) – a hypothesis that suggests categories can be neatly 'packed' into the feature space – may indicate that successive categories-by-negation are generated in a fashion that preferentially occupies spaces between observed categories. Further, although none of the current models of category generation can directly account for the effects observed in this study, they may be useful components in a larger category generation framework. For instance, future work may consider implementing the hierarchical sampling model from Jern and Kemp (2013) in a framework of overhypotheses (Kemp, Perfors, & Tenenbaum, 2007), where a prior can be placed over a category identity space, allowing models to behave differently under different regions of generated category identity.

Ultimately, the nature of newly generated categories appears to vary depending on the identity they were associated with. The extent to which they may differ, and the mechanisms driving these differences represent fascinating areas for future research.

Acknowledgments

This work was funded by a Vilas Life Cycle Professorship and the VCRGE at University of Wisconsin-Madison with funding from the WARF

References

Austerweil, J. L., Conaway, N., Liew, S. X., & Kurtz, K. J. (in preparation). Creating and learning something different: Similarity, contrast, and representativeness in categorization.

Conaway, N. B., & Austerweil, J. L. (2017). Packer: an exemplar model of category generation. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.), *Proceedings of the 39th annual conference of the cognitive science society* (pp. 1812–1817). London, UK: Cognitive Science Society.

Hidaka, S., & Smith, L. B. (2011). Packing: a geometric analysis of feature selection and category formation. *Cognitive Systems Research*, *12*(1), 1–18.

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford: Oxford University Press, Clarendon Press.

Jern, A., & Kemp, C. (2013). A probabilistic account of exemplar and category generation. *Cognitive Psychology*, *66*(1), 85–125.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science*, *10*(3), 307–321.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *10*(1), 104–114.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic bulletin & review*, *16*(2), 225–237.

Thomas, R. D. (1998). Learning correlations in categorization tasks using large, ill-defined categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(1), 119–143.

Ward, T. B. (1994). Structured imagination: The role of category structure in exemplar generation. *Cognitive Psychology*, *27*(1), 1–40.

Exploration and Exploitation Reflect System-Switching in Learning

Li Xin Lim (lim226@purdue.edu)
Department of Psychological Sciences, 703 Third Street
West Lafayette, IN 47907 USA

Sebastien Hélie (shelie@purdue.edu)
Department of Psychological Sciences, 703 Third Street
West Lafayette, IN 47907 USA

Abstract

Mounting evidence suggests that human category learning is achieved by multiple qualitatively distinct biological and psychological systems. In an information-integration (II) categorization task, optimal performance requires switching away from rule and adopting a procedural response strategy. However, many participants perseverate with rules. This article attempts at understanding the difference between optimal and suboptimal participants in II categorization. To this end, we collected data in the Iowa Gambling Task (IGT) and an II categorization task. Performance in the IGT was used to estimate each participant's sensitivity to reward, punishment, and propensity to explore. The results show that optimal participants in the II task explored more in the IGT than suboptimal participants. However, optimal participants in the II task did not show higher sensitivity to punishment or lower sensitivity to reward. We conclude by discussing the implications of these findings on system-switching and theoretical work on multiple-systems model of perceptual category learning.

Keywords: perceptual categorization; decision-making; dual systems; exploration-exploitation

Introduction

Categorization is an important part of daily life. From categorizing objects as edible or not to categorizing people as friends or enemies, everyday life is filled with thousands of category decisions. Over the past 20 years, mounting evidence has been gathered that category learning is achieved using a number of different psychological and biological systems (e.g., Ashby et al., 1998; Ashby & Valentin, 2017; Erickson & Kruschke, 1998; Hélie et al., 2010; Nosofsky et al., 1994; Waldschmidt & Ashby, 2011). However, much less is known about the interactions between the multiple categorization systems (Hélie, 2017). For example, the COVIS theory of categorization (Ashby et al., 1998) assumes that participants begin by guessing or using simple rules generated by hypothesis testing. Only after these rules have failed will participants abandon rule-based strategies and proceed to using alternative, more intuitive and less verbal methods of categorization.

One task where the primacy of rule-based strategy is often observed is the information-integration (II) categorization task. In II categorization tasks, participants need to integrate information from more than one dimensions at a pre-decisional level in order to maximize accuracy. Example for the II category structures are shown in Figure 2B. In this figure, each symbol represents the coordinate of a stimulus in perceptual space and specify one specific rotation angle and frequency that allow for drawing a unique sine wave grating (see Figure 2A). In this example, participants need to learn to categorize the 'o' and '+' in separate categories. This can be achieved by drawing a line in Figure 2B, but notice that the line would not correspond to a meaningful verbal description. The verbal description would be: 'o' are stimuli where the rotation angle is larger than the frequency, which is not meaningful given that rotation angle and frequency are not commensurable.

In an II categorization task like the one presented in Figure 2, the most accurate verbal rules can produce an accuracy of about 75%. In order to perform optimally, participants need to abandon rules and rely on a non-verbal procedural strategy. Decision bound models (DBM) (Hélie et al.,2017; Maddox & Ashby, 1993) can be used to identify the type of strategy that participants are using, and a consistent finding over the past 30 years is that a substantial number of participants perseverate with rule-based strategy in II category tasks and as a result perform suboptimally.

Reward Processing

The goal of this study is to understand why certain participants fail to abandon rule-based strategies and adopt non-verbal procedural strategies. To generate predictions, we first used the COVIS model of categorization (Ashby et al., 1998; Hélie, Paul, & Ashby, 2012) to fit published II categorization data collected in our lab. The COVIS model implements a multiple-systems theory of category learning that includes an explicit hypothesis-testing system and an implicit procedural-learning system. The explicit system learns through declarative memory by choosing and testing simple verbally expressible rules, whereas the implicit system employs non-declarative memory whereby learning is mediated by reinforcement learning as the system gradually assigns motor responses to regions of perceptual space. On each trial, the model compares the confidence in both systems and produce one response, either from the explicit system or from the implicit system.

The COVIS model was fit to data from Experiment 2 in Hélie & Cousineau (2015) (Condition = 0.5) to understand the switching of learning system between explicit and

implicit system in a perceptual categorization task. Decision bound models were fit to the data from each participant to separate participants using an optimal strategy from participants using a suboptimal strategy. The COVIS model was then fit to each group separately in order to identify which model parameters differed between simulations matching optimal participants and simulations matching suboptimal participants. Two hundred simulations were run for each subgroup of participants and the results are shown in Figure 1. The fit was excellent, with a RMSD of 1.5%. The model was able to differentiate optimal from suboptimal participants by changing the parameters odelta_e and odelta_c, which are the magnitude of the effect of the (negative and positive, respectively) feedback to adjust confidence in the hypothesis-testing system (Hélie, Paul, & Ashby, 2012). The simulations for optimal participants had a higher odelta_e value and a lower odelta_c value compared to the simulations of suboptimal participants, indicating that optimal participants are more sensitive to negative feedback while suboptimal participants are more sensitive to positive feedback. As a result, we hypothesize that optimal participants in II categorization tasks are more sensitive to negative feedback than participants who perseverate with rule-based strategies.

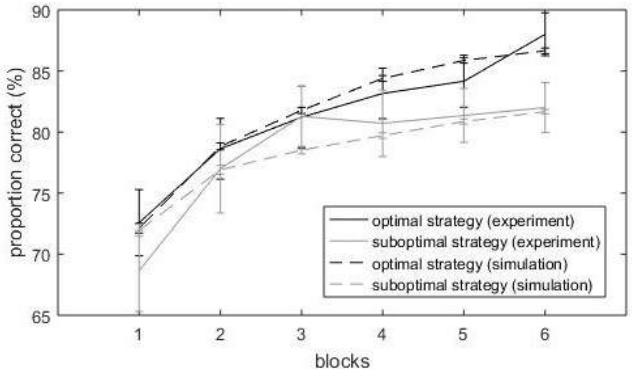


Figure 1: Average accuracy in Hélie & Cousineau (2015) and model results for each block of 100 trials. Black lines shows data for participants that use an optimal strategy, while grey lines indicate participants that use a suboptimal strategy. The participants' accuracy collected from the experiment are shown as solid lines, while the data from simulation are shown as dashed lines.

The Exploration-Exploitation Dilemma

One useful way to think about strategy switching and selection is to consider them in the context of the exploration-exploitation dilemma (Berger-Tal et al., 2014). Exploration and exploitation are seen as two opposing ways in the means of attention and resources allocation (Benner & Tushman, 2003; Gupta, Smith, & Shalley, 2006). Exploration entails risk taking, flexibility, discovery, and disengaging from the current task to allow for more room for experimentation, which is frequently associated with innovation. In contrast, exploitation is described with high-level engagement, choice-selection, efficiency and improvement (Laureiro-Martínez et al., 2015). The behavior of gathering information and exploiting are viewed as mutually exclusive events in many cases (Mettke-Hofmann, Winkler, & Leisler, 2002). When exploring, the agent seeks information about its environment as a way to improve performance, but in many situations it has to pay an opportunity cost (March, 1991). Agents that only exploit using current knowledge might be stuck in a suboptimal stable equilibrium, unable to adapt fully to the environment (March, 1991; Uotila et al., 2009). Thus, an optimal strategy in decision-making is to have balance between exploration and exploitation, allowing resource allocation between the two behaviors to yield the 'best' long-term rewards (March, 1991).

The exploration-exploitation dilemma to some extend resembles the results observed in the II categorization task. Assuming that participants begin by using a rule-based strategy, 'exploiters' may perseverate with a rule-based strategy since it allows for responding correctly in about 75% of the trials. Exploration is required to abandon rule-based strategies and try procedural strategies that are more optimal. As a result, we hypothesize that participants who explore more are more likely to perform optimally in an II categorization task.

Methods

To test for the hypotheses, we used the Iowa Gambling Task (IGT) (Bechara et al., 1994) to measure reward sensitivity and exploration tendencies. Each participant performed both an II categorization task and the IGT. Performance in the IGT was used to predict whether participants would use an optimal or suboptimal strategy in the II categorization task.

Participants

Fifty participants were recruited from the Purdue University undergraduate population. Each participant was given credit for participation as partial fulfillment of a course requirement. Participants gave written informed consent and all procedures were approved by the Purdue University Human Research Protection Program Institutional Review Board.

Materials and Procedure

Each participant did both the Iowa Gambling Task (IGT) and the perceptual categorization task (PCT) in random order of IGT-PCT (n = 27) and PCT-IGT (n = 23). The experiment was run on a Desktop PC equipped with a regular mouse and keyboard. Stimuli were displayed in a 21-inch monitor with 1,920 × 1,080 resolution. The experiment was controlled by in-house programs written using PsychoPy.

Iowa Gambling Task Participants were presented with four blue rectangles. The blue rectangles were labeled as "Deck A", "Deck B", "Deck C", and "Deck D". The task required participants to repeatedly draw 'cards' from the four decks, by clicking on the blue rectangle on the screen with a mouse. Participants were required to select a deck on each trial within

four seconds. If a participant failed to select a deck before the deadline, the program randomly selected a deck. The participant could only select one deck for each trial.

The expected values of the decks differed so that two decks were associated with high immediate rewards but long-term overall loss (disadvantageous decks A and B), and two other decks yielded lower immediate rewards but long-term overall gains (advantageous decks C and D). The experiment was designed to record the participant's affinity towards each deck given the rewards and penalties presented in each trial upon selection of a particular deck. The reward and penalty from the selected deck in the particular trial, as well as the total accumulated gain from the rewards and penalty gathered thus far was presented to the participant at the end of each trial. The rewards and penalties were generated to meet the requirements listed in Table 1. Each deck contained 10 different cards and was re-shuffled after all 10 cards had been drawn. Each participant performed 120 trials grouped into six blocks of 20 trials each. Completing the IGT took about 10 minutes.

Table 1: Deck properties (Bechara et al., 1994)

Card	Deck A	Deck B	Deck C	Deck D
P(penalty)	0.5	0.1	0.5	0.1
Penalty	-150 to -350	-1250	-25 to -75	-250
Reward	100	100	50	50
Expectation	-250	-250	250	250

Perceptual Categorization Task (PCT) The stimuli used in the PCT were circular sine-wave gratings of fixed contrast and size, as shown in Figure 2. The stimuli differed in terms of bar width and orientation. The bar width was derived as the frequency of lines in a 2D space in cycles per degree, while the orientation is the counterclockwise rotation of the lines from horizontal in radian. The stimuli were categorized as A and B, with a diagonal line as a category bound as shown in Figure 2. Perfect accuracy was possible and optimal performance required responding to the A-B stimuli using a procedural strategy.

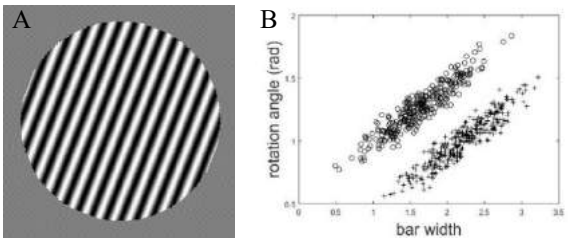


Figure 2: (A) Example stimulus shown to the participants for PCT, (B) Category structures used in PCT.

The participants were informed that they were taking part in a categorization experiment and that they needed to learn to categorize the stimuli presented into either category A or B with trial-and-error. In each trial of this task, a "crosshair" was presented on the screen for one second, followed by a single stimulus presented in the center of the screen. Participants were required to choose a category for the stimulus. Responses were given on a standard keyboard: "s" key for category A and "k" key for category B. After each trial, visual feedback showing "Correct", "Incorrect", or "Wrong Key" was given to the participant according to the response they pushed. The response for stimulus on each trial was recorded, as well as the response time. The participants did 600 trials grouped into six blocks of 100 trials each. The PCT took about 35 minutes to complete.

Decision bound models

The objective of the study was to explore the difference in sensitivity to reward and punishment between participants that used an optimal strategy and participants that did not use an optimal strategy. To allow for the classification of participants into optimal strategy users and suboptimal strategy users, Decision Bound Models (DBM) were applied to the perceptual categorization task to identify how participants learned to assign responses to regions of perceptual space. In DBM, it is assumed that participants determine the region of the percept and give the associated response. The decision bound is described as a partition segregating competing response regions. Three general classes of decision bound models were fit to response data of the PCT (Hélie et al., 2017; Maddox & Ashby, 1993): (1) guessing models, (2) explicit rule-reasoning models, and (3) procedural learning models.

The guessing models assumes that participants do not use the information on the screen and randomly responded "A" or "B" in each trial. The explicit models set a boundary to segregate participant's responses with a vertical line or horizontal line (or the combination of both vertical and horizontal lines). An adjusted diagonal line is used as the boundary instead in the procedural learning models. For each participant's data set, the best model is selected using the Bayes information criterion (BIC). Participants whose data were best-fit by the optimal models, which is the procedural learning model in this case, are labelled as "optimal strategy" and all other participants are labelled as "suboptimal strategy".

Rescorla-Wagner Model

The data recorded in the IGT were fitted with the Rescorla-Wagner (1972) model (RW). The RW was used to calculate a value for each deck and estimate a participant's sensitivity towards reward and punishment.

Data for each participant was fed into the RW model. For each trial, t in a particular task block, the parameter for sensitivity to reward, b_{rew} was multiplied with the magnitude of reward, R received following the participant's response in each trial, while the parameter for sensitivity to punishment, b_{pun} was multiplied with the magnitude of punishment, P received following the participant's response in each trial.

The key equations to update reward and punishment sensitivity were:

$$B_{rew} = \frac{b_{rew} \times (R(t) - P(t))}{\max(R)} \quad (1)$$
$$B_{pun} = \frac{b_{pun} \times (P(t) - R(t))}{\max(P)}$$

where, B_{rew} and B_{pun} are the sensitivity to reward and punishment, for the perceived net gain and loss in each trial. The key equations to update reward and punishment sensitivity were:

$$\text{Qdeck(t)} = \text{Qdeck(t} - 1) + \alpha\big(B_{rew} - \text{Qdeck(t} - 1)\big) \quad (2)$$
$$\text{Qdeck(t)} = \text{Qdeck(t} - 1) + \alpha(B_{pun} - \text{Qdeck(t} - 1))$$

where, Qdeck is the Q-value for each deck and α is the learning rate. The equation on top in Eq. 2 updates the deck value with B_{rew}, while the equation below updates with B_{pun}. In trials where an overall reward was received, the equation with B_{rew} was used to update the deck value; in trials where overall punishment was received, the equation with B_{pun} was applied to update the deck value. All parameters were estimated using Maximum A Posteriori (MAP).

The sensitivity towards reward and punishment b_{rew} and b_{pun} for each participant were then normalized. The weighted proportion of b_{rew} and b_{pun} with respect to the summation of b_{rew} and b_{pun} were computed with Equation 3.

$$W_{rew} = \frac{b_{rew}}{b_{rew} + b_{pun}} \quad (3)$$
$$W_{pun} = \frac{b_{pun}}{b_{rew} + b_{pun}}$$

where, W_{rew} and W_{pun} are the weighted proportion of b_{rew} and b_{pun}, respectively.

Results

Effects of sensitivity to punishment and rewards

Participants in the PCT were categorized into participants who found the optimal strategy and participants who did not using DBM. The sensitivity to punishment (b_{pun}) and reward (b_{rew}) were computed with the RW. W_{rew} and W_{pun} were computed as the weighted proportion of sensitivity to reward and punishment, respectively, and the mean estimates of the proportion of W_{rew} and W_{pun} for participants that used an optimal strategy and a suboptimal strategy are shown in Figure 3. Confirming our hypothesis, W_{rew} [t(48) = 1.901, p = 0.032] of participants that used an optimal strategy was lower than that of participants that used a suboptimal strategy, whereas W_{pun} [t(48) = -1.901, p = 0.032] of participants that used an optimal strategy was higher than that of participants that used a suboptimal strategy. These results show that participants using an optimal strategy in the PCT have a greater sensitivity to punishment, while participants using a suboptimal strategy in the PCT have a higher sensitivity to reward.

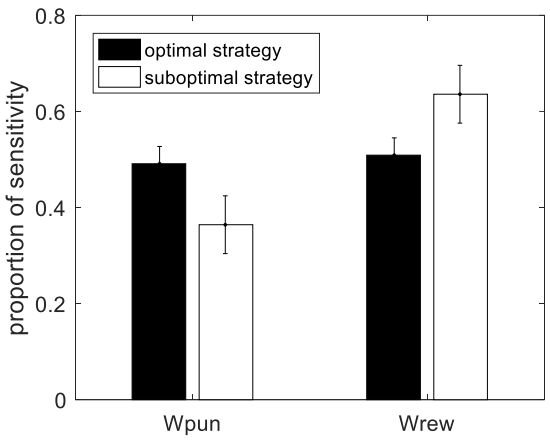


Figure 3: Estimated sensitivity to reward and punishment (IGT) for optimal and suboptimal participants (PCT). Error bars are standard error of the mean.

Exploration affects category learning

Exploration was quantified as the number of deck switches in the IGT and was compared between the two groups of participants. The number of deck switches was subjected to an independent samples t-test to test the effect of using optimal or suboptimal strategy in PCT. The average number of deck switches for the two groups is shown in Figure 4. The main effect of the strategy used [t(48) = 1.684, p = 0.049] was significant. The number of deck switches for participants that used an optimal strategy in the PCT (mean = 71.91) was higher than the number of deck switches for participants that used a suboptimal strategy in the PCT (mean = 60.38). The results suggest that, as predicted, participants who conduct more exploration are more likely to perform optimally in an II categorization task.

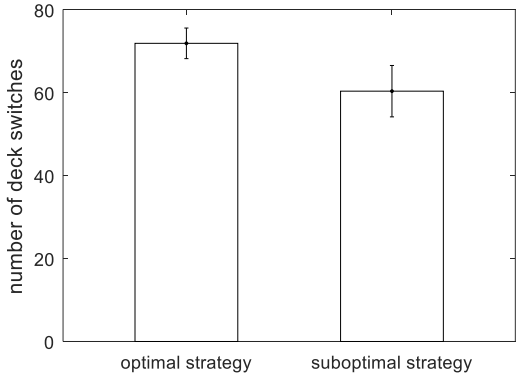


Figure 4: Average number deck switches in the IGT for optimal and suboptimal participants in the PCT. Error bars are standard error of the mean.

To further assess the role of exploration in strategy selection, we measured the entropy of choosing different decks: A, B, C and D. Entropy gives a sense of disorder and uncertainty. Hence higher entropy means that all decks were sampled equally often, whereas an entropy of 0 means that participants always selected the same deck. We first calculated the correlation between entropy and number of deck switches in the IGT (Figure 5A). This analysis informs whether participants always switch between a subset of the decks or if all decks are sampled. The correlation was 0.513, which is statistically significant [$t(48) = 4.136$, $p < 0.001$]. This result suggests that participants with more deck switches sample from all decks.

Next, we computed the linear relationship between entropy and sensitivity to feedback in the IGT ($b_{pun} + b_{rew}$). Here, b_{pun} is negative, so a negative number means higher sensitivity to punishment while a positive number means a higher sensitivity to reward (0 means equally sensitive to both types of feedback). This analysis informs about the relationship between the breadth of exploration (sampling from some or all the decks) and feedback sensitivity in the IGT. The correlation was -0.665, which reached statistical significance [$t(48) = -6.168$, $p < 0.001$] (Figure 5B). This suggests that higher sensitivity to punishment leads to sampling from more decks, which is consistent with our hypothesis that greater sensitivity to punishment leads to more exploration.

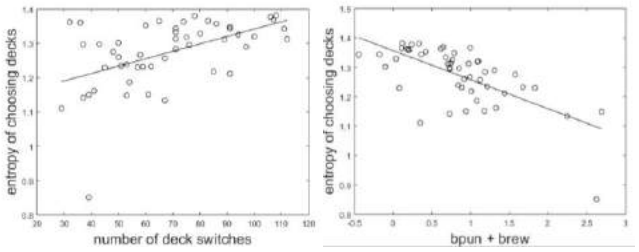


Figure 5: (A) Relationship between entropy and number of deck switches in the IGT. (B) Relationship between feedback sensitivity and entropy in the IGT.

Discussion

This article presents the results of an experiment aimed at understanding why some participants fail to select an optimal procedural strategy in II categorization and instead perseverate with using suboptimal rule-based strategies. By fitting the COVIS model to published II categorization data, we hypothesized that participants using an optimal strategy in II categorization would be more sensitive to punishment whereas participants using a suboptimal strategy would be more sensitive to reward. We further hypothesized that participants with a tendency to explore would be more likely to use an optimal strategy in the II task.

We tested these predictions by running participants in an II categorization task and the IGT. Decision-bound models were fit to II categorization data to classify each participant as optimal or suboptimal. The RW model was fit to the IGT to estimate each participant's sensitivity to reward and punishment. The number of deck switches and entropy of choice in the IGT were used to estimate the propensity of each participant to explore. The results partially supported our hypotheses. As predicted, exploration was related to the selection of an optimal strategy in II categorization. Sensitivity to punishment was also related to propensity to explore, but only in the IGT. The hypothesis that sensitivity to punishment would be related to the selection of an optimal strategy in II categorization was not supported in the experiment.

System-Switching vs. Rule-switching

Individuals vary considerably in terms of their sensitivity to reward and punishment. Sensitivity to reward can be described as how an individual's behavior is driven by reward-related stimuli, while sensitivity to punishment is described as how an individual's behavior is subdued by punishment-related stimuli. Studies suggest that individuals with greater sensitivity to reward are more reactive to rewarding outcomes but are less sensitive to monitoring loss, while greater sensitivity to punishment are linked to avoidance and giving up actions in absence of immediate reward (Kim et al., 2015).

As predicted by COVIS, the selection of certain strategies and the abandonment of others depends on the evaluation of how rewarding the strategy is. The implementation of one system over the other depends on the confidence and trust in the system. The trust is a function of the effect of received feedback when using a particular system. Thus, the switching of strategies from a rule-based to aa procedural strategy depends on the feedback received when using the particular strategy, which can be explained in terms of the reward and punishment the system or strategy gets when providing a response.

Our study confirms the finding that the selection of an optimal strategy in II categorization task is associated with greater sensitivity to punishment and perseveration with suboptimal strategies in II tasks is associated with greater sensitivity to reward. If the participant focuses more on the effect of punishment, losses in the task leads to giving up and avoidance of certain strategies, which leads to the possibility of adopting a strategy that leads to the optimal outcome. If the participant has greater sensitivity towards reward, s/he is less sensitive to immediate loss and tend to perseverate with strategies that bring a certain degree of rewards. In the II categorization task, participants tend to perseverate with a rule-base strategy since the strategy allows for a certain degree of accuracy (typically about 70%). However, additional research is required to determine how participants change from attending to specific stimulus dimension(s) to an integrated procedural-based strategy.

Exploration in the IGT

Deck switch is used as a measure of exploration. A larger number of deck switches indicates that participants were willing to disengage from the current strategy or deck, and

were willing to explore other possible options, despite the uncertainty and risk of getting punished. As seen in Figure 4, both optimal and suboptimal participants explored greatly in the IGT. The difference in exploration between optimal and suboptimal participants may have been caused by several factors, which includes willingness to take risk to maximize gain. The tendency for exploration appears to be robust and may be predictive of both rule-switching and system-switching. This was shown by the number of switches being both related to entropy in the IGT and the selection of an optimal strategy in II categorization. COVIS does not explicitly model exploration but a tendency to explore would be characterize by noise in system selection.

One observation is that many participants tend to choose Deck B in the IGT. With an expected value of a net loss of $250 and a relatively large loss of $1250 as compared to the other decks in the task, it would normally inhibit participants from selecting Deck B. The basic assumption is that the largest loss would trigger an alarming signal from the intact somatic system, thus inhibiting further selection of deck B as it guides the process of decision-making (Lin et al., 2007). However, Deck B has a low loss-frequency owing to a small number of trials with large losses (or can be seen as a high gain-frequency), which may explain why participants choose the deck despite great immediate loss when a penalty card is drawn from the deck. Most participants' behavior are driven by the high gain-frequency, instead of inhibited by the great loss while choosing Deck B (Dunn et al., 2006; Lin et al., 2007).

Participants that used suboptimal strategies tend to fixate on specific deck(s) and were not willing to explore for more reward, which might cause them to be stuck in a local minimum, and lose the chance to seek out strategies that are more efficient. The fixation can be due to contentment, unwillingness to take risks, or pros-to-cons weighing. Additional research is needed to determine why certain participants are reluctant to explore.

Future Work

This experiment came with a few limitations. Some of the advantage and disadvantage decks used in the IGT were difficult to identify through limited interactions with the decks, which might misguide participants while performing the tasks. For example, exploiting Deck B in IGT results in an overall loss, the frequency of loss is small. Hence, participants may consider Deck B to be an advantageous deck and continue choosing the deck along with other advantage decks. Questionnaires could be given to participants to ask for the decks the participants believed to be advantageous. This would allow for better understanding whether participants considered each deck as "risky" or not and disentangle risk taking from bad estimation of deck expectation.

Finally, a task needs to be designed that shares properties with the IGT but requires system-switching instead of rule-switching (or deck switching). This new task would allow to more directly estimated sensitivity to reward and punishment between-system and would provide a more definitive test of the hypothesis that optimal participants, who switch system in an II categorization task, are more sensitive to punishment then suboptimal participants, who are more sensitive to reward.

Acknowledgement

This research was supported, in part, by National Science Foundation award #1662230 and by NIMH grant #2R01MH063760.

References

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, *105*(3), 442–481.

Ashby, F. G., & Valentin, V. V. (2017). Multiple systems of perceptual category learning: Theory and Cognitive Tests. *Handbook of Categorization in Cognitive Science*, 157–188.

Bechara, A., Damasio, A. R., Damasio, H., & Anderson, S. W. (1994). Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition*, *50*, 7–15.

Benner, M. J., & Tushman, M. L. (2003). Exploitation, exploration, and process management: the productivity dilemma revisited. *Academy of Management Review*, *28*, 238–256.

Berger-Tal, O., Nathan, J., Meron, E., & Saltz, D. (2014). The exploration-exploitation dilemma: A multidisciplinary framework. *PLOS ONE*, *10*, e0119116.

Crossley, M., Roeder, J., Hélie, S., & Ashby, F. (2018). Trial–by–trial switching between procedural and declarative categorization systems. *Psychological Research*, *82*, 371–384.

Dunn, B. D., Dalgleish, T., & Lawrence, A. D. (2006). The somatic marker hypothesis: A critical evaluation. *Neuroscience Biobehaviour*, *30*(2), 239–271.

Erickson, M. A., & Kruschke, J. K. (1998). Rules and Exemplars in Category Learning. *Journal of Experimental Psychology: General*, *127*(2), 107–140.

Gupta, A. K., Smith, K. G., & Shalley, C. E. (2006). The interplay between exploration and exploitation. *Academy of Management Journal*, *49*, 693–70.

Hélie, S. (2017). Practice and preparation time facilitate system-switching in perceptual categorization. *Frontiers in Psychology*, *8*, 1964.

Hélie, S., & Cousineau, D. (2015). Differential effect of visual masking in perceptual categorization. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(3), 816–825. https://doi.org/10.1037/xhp0000063

Hélie, S., Paul, E. J., & Ashby, F. G. (2012). Simulating the Effects of Dopamine Imbalance on Cognition: From Positive Affect to Parkinson's Disease. *Neural Networks*, *32*, 74–85.

Hélie, S., Turner, B. O., Crossley, M. J., & Ell, S. W. (2017). Trial-by-trial identification of categorization strategy using

iterative decision bound modeling. *Behaviour Research Method*, *49*, 1146–1162.

Hélie, S., Waldschmidt, J. G., & Ashby, F. G. (2010). Automaticity in rule-based and information-integration categorization. *Attention, Perception, & Psychophysics*, *72*, 1013–1031.

Kim, S. H., Yoon, H., Kim, H., & Hamann, S. (2015). Individual differences in sensitivity to reward and punishment and neural activity during reward and avoidance learning. *Social Cognitive and Affective Neuroscience*, *10*(9), 1219–1227.

Laureiro-Martínez, D., Brusoni, S., Canessa, N., & Zollo, M. (2015). Understanding the exploration-exploitation dilemma: an fMRI study of attention control and decision-making performance. *Strategic Management Journal*, *36*, 319–338.

Lin, C. H., Chiu, Y. C., Lee, P. L., & Hsieh, J. C. (2007). Is deck B a disadvantageous deck in the Iowa Gambling Task? *Behavioral and Brain Functions*, *3*(16).

Maddox, W. T., & Ashby, F. G. (1993). Comparing decision bound and exemplar models of categorization. *Perception and Psychophysics*, *53*(1), 49–70.

March, J. G. (1991). Exploration and exploitation in organizational learning. *Organization Science*, *2*, 71–87.

Mettke-Hofmann, C., Winkler, H., & Leisler, B. (2002). The significance of ecological factors for exploration and neophobia in parrots. *Ethology*, *108*, 249–272.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-Plus-Exception Model of Classification Learning. *Pyschological Review*, *101*(1), 53–79.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In H. A. Prokasy & W. F. Black (Eds.), *Classical conditioning II: Current research and theory*. New York: Appleton- Century-Crofts.

Uotila, J., Maula, ., Keil, T., & Zahra, S. A. (2009). Exploration, exploitation, and financial performance: analysis of S&P 500 corporations. *Strategic Management Journal*, *30*, 221–231.

Waldschmidt, J. G., & Ashby, F. G. (2011). Cortical and striatal contributions to automaticity in information-integration categorization. *NeuroImage*, *56*(3), 1791–1802.

Curiosity, Frontal EEG Asymmetry, and Learning

Gabriel Lima (gdacunhalima@drew.edu)
Department of Psychology, Drew University
Madison, NJ USA 07940

Fabiana Rocha (fdeoliveiraroch@drew.edu)
Department of Psychology, Drew University
Madison, NJ USA 07940

Abstract

Curiosity plays a critical role in our daily behaviors and interactions. Yet, very little is known about its psychological and neural underpinnings. By reframing curiosity as the motivation to obtain reward – where the reward is information –, and using frequency-based metrics of frontal brain lateralization, we aimed to investigate the neural correlates of curiosity in the frontal cortex and its effects on subsequent learning. Twenty-one undergraduate students participated in this two-day study by answering 35 general interest trivia questions, while EEG data was being recorded, also indicating their curiosity towards the question. One week later, participants were asked to write down the correct answers to each one of the questions. The results of this study suggested that frontal brain asymmetry (FBA) predicts memory recall, but is not directly correlated with self-reported curiosity. Study limitations and future directions are discussed.

Keywords: curiosity; EEG; frontal brain asymmetry; learning; memory

Introduction

Curiosity plays a critical role in many of our daily pursuits, actions, and interactions. It drives learning and promotes discovery, increasing our understanding of the world. Albert Einstein once said, "I have no special talents. I am only passionately curious" (Hoffmann, 1972, p. 7). Yet for something that drives much of our daily behavior and knowledge, very little is known about its psychological and neural underpinnings. Lowenstein (1994) was the first one to propose an information gap theory, suggesting that curiosity arises from a perceived information gap, that is, the disparity between what one knows and what one wants to know. According to him, curiosity seeks a subjective value: information.

Innovating from this theory, Marvin & Shohamy (2016) reframed curiosity as the motivation to obtain reward, where the reward is information. This information-as-reward framework was supported by the fact that curiosity shares behavioral and neurobiological properties with other reward-motivated behaviors, as the same dopaminergic neurons that signal changes in the value of the reward also code changes in the value of information (Hare, O'Doherty, Camerer, Schultz, & Rangel, 2008; Kang et al., 2009). Furthermore, high-curiosity information is associated with activation in brain areas known to respond to reward, which includes the caudate and the nucleus accumbens (Gruber, Gelman, & Ranganath, 2014; Kang et al., 2009), and there is a strong link between how valuable information is and the likelihood of remembering it (Gruber et al., 2014; Kang et al., 2009; Mullaney, Carpenter, Grothehuis, & Burianek, 2014). Research has also found that learning is driven not only by the absolute value of given information but also by an information prediction error (IPE), which is the difference between the reward expected and the reward received (Daw & Doya, 2006; Schultz, 2006; Marvin & Shohamy, 2016).

Although these studies demonstrate that curiosity conforms to basic characteristics of reward-motivated behavior, they leave open critical questions related to the extent to which this analogy is valid at a deeper level. The greatest problem is that almost all current studies that investigate curiosity rely primarily on self-reports as a way to measure it, which, despite being convenient and affordable, is knowingly not the most reliable technique currently available. This is due mainly to the lack of a well-known, comprehensive, and more credible method to investigate and measure curiosity.

Over the last decades, however, neuroscience research has developed significantly, and analyses of EEG data have become much more advanced. One of the more sophisticated frequency-based metrics is frontal EEG asymmetry, or frontal brain asymmetry (FBA). This index is commonly used as a tool to measure engagement and motivation, typically using alpha power (8 – 13 Hz) in electrodes over frontal cortical regions (channels F3 and F4). Previous studies have consistently found that greater activity in the left (F3) versus the right (F4) frontal cortex indicates positive feelings, higher engagement, and motivation (Davidson, 2004; Harmon-Jones & Gable, 2017). Evidence suggests that frontal lateralization can, in fact, be used to analyze people's engagement to media advertisements, market products, and services (Vecchiato et al., 2011; Yilmaz et al., 2014). Furthermore, research findings confirm the idea that frontal brain asymmetry modulates the probability to engage in reward-motivated behavior (Pizzagalli, Sherwood, Henriques, & Davison, 2005; Schmid, Hackel, Jasperse, & Amodio, 2017).

Therefore, our study aimed to expand from previous investigations on both curiosity and frontal brain asymmetry. By using the same information-as-reward approach, and reframing curiosity as the motivation to obtain reward – where reward is information –, we wanted to investigate if the same frameworks and methods currently used to study engagement and motivation can be used to measure curiosity in a more reliable way, serving as an alternative to the current self-reported measures. If this held true, we expected to see higher activation in the left frontal cortex – a greater frontal brain asymmetry – when people were exposed to high-curiosity information. We would also be able to correlate higher FBA scores to a higher likelihood of remembering the information. Hence, we aimed to investigate if frontal EEG alpha left asymmetry is (1) in any way related to self-reported curiosity and (2) a stronger predictor of subsequent learning. Our study may provide an initial framework for future studies on curiosity, as well as help to shed light on the functional significance of frontal EEG asymmetry on curiosity, learning, and other reward-motivated behaviors.

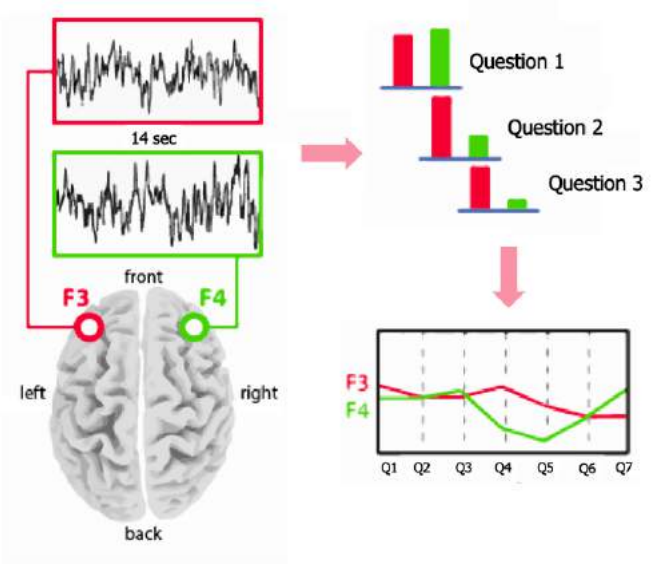


Figure 1: Illustration of EEG data collection. Raw EEG data was extracted from regions F3 and F4 (10/20 System Positioning) during the 14-second period in which each question was presented. The Fast Fourier Transform algorithm (FFT) was used to calculate the alpha power from each raw signal.

Methods

Participants

21 undergraduate students (mean age = 18.8 ± 1.1 year; 12 female, 9 male) at a college of liberal arts in the greater New York City area participated in this two-day study for partial course credit.

Materials & Equipment

Brain electrical data from this experiment was collected using electroencephalography (EEG) equipment, iWorx IX EEG 10-20 (iWorx Systems, Dover, NH) culled from two scalp sites (F3 and F4). The questions were presented on Apple Macintosh computers, using Qualtrics (2013) and the QuickTime Player (Cupertino, CA) to present stimuli and collect responses. The analysis of the EEG data was done on LabScribe Software, and all subsequent statistical analyses were done on R (R Core Team, 2013).

Procedure

The first session was about 45 minutes long, and the second one (a week later) was about 15 minutes long. On the first session, after providing written informed consent and answering a quick demographic questionnaire (which included age, gender, race and/or ethnicity, and handedness), participants were prepared for EEG recording. Before the primary task, two electrodes were placed on the participant's scalp (regions F3 and F4, 10/20 System Positioning; see Figure 1), and two minutes of EEG baseline was recorded. The experiment was a within-subjects design, where all participants were presented with a set of 35 general interest trivia questions culled from Internet sources (e.g., "What is the capital of Brazil?). Each question was presented on the laptop screen for 14 seconds. Participants were instructed to, after reading each question, type the answer down, and indicate their curiosity about the correct answer and their confidence in their guess. Then the question was presented again, followed by the correct answer (Kang et al., 2009). The same procedure was repeated for each and all of the 35 trivia questions, and the order of the questions was the same for all participants. EEG data was recorded for the entirety of the experiment. All participants were expected to come for a follow-up session one week later, although they were not aware of the purpose of the second session. For this session, the same 35 questions from the first day were presented and participants were asked to write down the answer to each one of them.

EEG Data Analysis

Alpha frequency band power was calculated by extracting frequency domain features from both left and right raw EEG signals (channels F3 and F4, respectively; see Figure 1). Since each question was presented for 14 seconds, we extracted the first seven 2-second epochs from each segment, and averaged them. The frequency domain analysis was performed using the Fast Fourier Transform (FFT) algorithm (with a frequency resolution of 1 Hz). The power spectra were reduced to the alpha frequency band, defined as between 8–13 Hz.

The frontal brain asymmetry index was calculated by dividing the alpha power values from the F4 (right) electrode by the values from the F3 (left) electrode. The results were computed using a natural log transformation to normalize the data as frequency power values tend to be severely skewed. This is illustrated by the following formula:

$$FBA\ Index = ln\left(\frac{F4\ right\ \alpha\ power}{F3\ left\ \alpha\ power}\right)$$

Since alpha power is inversely related to brain activity, positive asymmetry scores represented relatively greater alpha (less activity) over right than left hemispheres (Coan & Allen, 2004).

Data Preprocessing

The quality of the signal received from each electrode was evaluated during the entire EEG recording in order to make them both comparable and to avoid the influence of artifacts on the analysis. Offline visual artifact rejection was used to remove eye blinks, head movements, muscle activity, and other noise from the data. A subsequent round of artifact rejection was also conducted in which single trials containing voltage deviations of over 50 μV from normal baseline were manually rejected. Therefore, only artifact-free data from electrodes F3 and F4 were extracted and used in the analysis.

In addition to the EEG signal filtering, we also excluded trials based on whether or not the participant already knew the answers to the presented trivia, such that questions that were correctly answered by the participants during session one were not included in the EEG analysis. Therefore, our preprocessing filter yielded a total of 519 trials (332 correctly recalled, 187 not correctly recalled) across all 21 participants of this study.

Results

Self-Reported Curiosity and Frontal Brain Asymmetry

A correlational approach was used to assess links between reported curiosity and frontal brain asymmetry. Pearson's correlation coefficient indicated no statistically significant correlation between self-reported curiosity and FBA, neither for correctly remembered answers, $r(N = 21) = -.008$, $p = .486$, nor for incorrectly remembered answers, $r(N = 21) = -.290$, $p = .101$. Therefore, reported curiosity values were not linked to higher asymmetry values, on average (see Figure 2).

Frontal Brain Asymmetry and Learning

Participants on average remembered 62.1% of the answers correctly (range: 30.2% – 82.7%). A paired-samples t-test was conducted to compare FBA index, self-reported curiosity, and confidence level for both correctly remembered and incorrectly remembered answers (see Table 1). For confidence scores, there was a significant difference between incorrect ($M = 2.14$, $SD = 1.05$) and correct ($M = 2.69$, $SD = 1.17$) answers; $t(20) = 2.97$, $p = .008$. For curiosity scores, on the other hand, there were no statistically significant differences between incorrect ($M = 6.29$, $SD = 1.83$) and correct ($M = 6.63$, $SD = 1.43$) answers; $t(20) = 1.17$, $p = .254$.

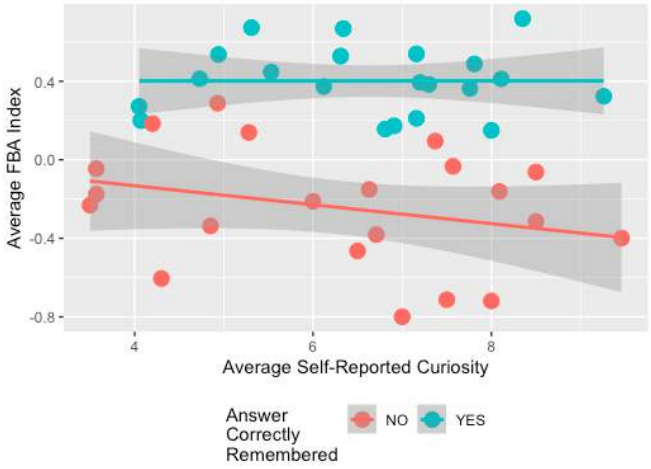


Figure 2: Scatterplot of self-reported curiosity and FBA. Each participant is represented by two dots, one orange and one green. Pearson's correlation coefficient indicated no statistically significant correlation between self-reported curiosity and FBA for neither correct nor incorrect answers (p=.486, $p = .101$).

For frontal brain asymmetry scores, the difference for incorrect ($M = -0.24$, $SD = 0.30$) and correct ($M = 0.40$, $SD = 0.17$) answers was statistically significant; $t(20) = 7.20$, $p < .001$. Specifically, participants' recall was better for trials in which they had higher asymmetry scores than those in which they had lower asymmetry scores (see Figure 3). These results suggest that FBA is linked to whether or not an individual remembered the information from the trivia questions correctly.

Discussion

This experiment investigated if frontal EEG alpha left asymmetry was (1) in any way related to self-reported curiosity and (2) a better predictor of subsequent learning. By using the information-as-reward approach, and after reframing curiosity as the motivation to obtain reward – where reward is information –, we investigated if the same methods currently used to study engagement and motivation can be used to measure curiosity in a more reliable way, serving as an alternative to the current self-reported measures. If asymmetry measurements in EEG recording were indeed a more reliable way to measure curiosity, we expected to see higher neural activity in the left frontal cortex – a greater frontal brain asymmetry – when people were exposed to high-curiosity information. This asymmetry index, then, would be a better predictor of whether or not the participant would remember the correct answer – if compared to the participant's self-reported curiosity levels.

The data indeed showed that there was a relationship between frontal brain asymmetry and subsequent learning:

Table 1
Descriptive Statistics and Paired t-test Results for Confidence, Curiosity, and FBA Index

	Not Correctly Remembered		Correctly Remembered			95% CI for Mean Difference		
Measure	M	SD	M	SD	n		t	df
Confidence	2.14	1.05	2.69	1.17	21	0.16, 0.92	2.97*	20
Curiosity	6.29	1.83	6.63	1.43	21	-0.27, 0.95	1.17	20
FBA Index	-0.24	0.30	0.40	0.17	21	0.46, 0.83	7.20*	20

* $p < .05$.

Table 1: Descriptive statistics and paired t-test results for confidence, curiosity, and FBA index. There are statistically significant differences, at the .05 significance level, in the correctness scores for confidence and frontal brain asymmetry, but not for curiosity. Results show that both confidence levels and FBA scores were higher for correctly remembered answers than for incorrectly remembered answers.

participants were significantly more likely to remember the correct answers for trials in which they had higher FBA scores (see Figure 3). Self-reported curiosity, on the other hand, was not associated with subsequent learning. For the second half of our research question, we utilized a bivariate correlational analysis to investigate whether frontal brain asymmetry and reported curiosity were linked to each other. We found that self-reported curiosity and FBA were not statistically significantly correlated, meaning that higher values of left hemisphere activation were not linked to higher self-given scores of curiosity (see Figure 2).

Because our study did not find any link between self-reported curiosity and frontal brain asymmetry, it is not possible to infer any relationship between these two variables. In other words, frontal brain asymmetry might not be a neural correlate of curiosity, as we had initially hypothesized. However, although our experimental study design avoids claiming causality, our results support the idea that frontal brain asymmetry might be a better predictor of subsequent learning and correct information recall than the curiosity scores reported by the participants. Differently from Marvin & Shohamy (2016), self-reported curiosity did not correlate with subsequent learning in our study. These data leave open critical questions related to the reliability of self-reports measures on research investigating curiosity. Given that the current studies on the topic rely primarily on self-reports as a way to measure curiosity due to its convenience and affordability, more research is needed in order to confidently state the effects of curiosity on memory and learning.

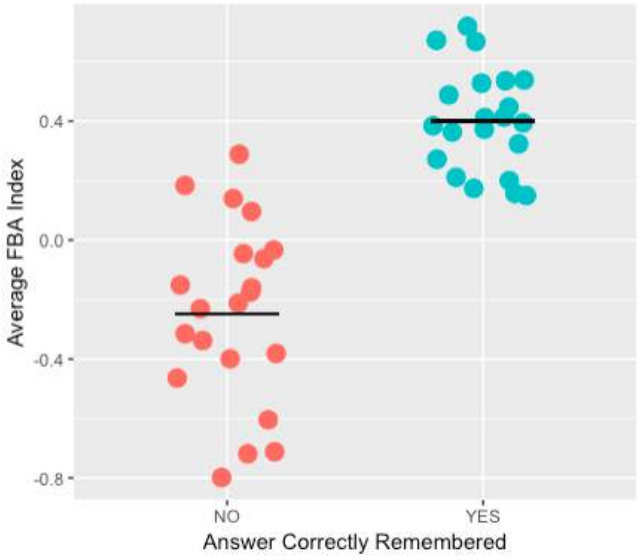


Figure 3: Plot of the mean differences in FBA between questions that were and were not remembered correctly (the black lines indicate the mean for each). Each participant is represented by two dots, one orange and one green. Participants' recall was better for trials in which they had higher asymmetry scores than those in which they had lower asymmetry scores ($p < .001$).

Moreover, the variable confidence level showed a significant effect on the correctness of the responses in the retest ($p = .008$). Subjects were more likely to provide a correct answer during the retest when the same question on the pretest was answered incorrectly but with a high level of confidence. These results are in accordance to previous studies on hypercorrection, which suggest that high-confidence errors tend to be corrected at a higher rate on retests, when compared to low-confidence ones (Metcalfe & Finn, 2011; Metcalfe & Miele, 2014).

Our study also found that there is a positive relationship between frontal brain asymmetry and subsequent learning (see Figure 3). More specifically, correct answers have a significantly higher FBA index than incorrect answers ($p < .001$). Future research is necessary, however, in order to investigate *why* this relationship exists. Previous studies have suggested that greater activity in the left versus the right frontal cortex indicates positive feelings, higher engagement, and motivation (Davidson, 2004; Harmon-Jones & Gable, 2017). Although these correlates of FBA might play a role in whether a participant will remember the correct answer or not, only future studies might be able to indicate if this is true.

The present study is not without limitations. The number of participants included in the final analysis was relatively

small ($N = 21$). Future work should aim to collect and analyze data from a more extensive poll of participants in order to examine if findings will hold true with more data. Furthermore, the equipment used in this study was quite rudimentary if compared to more expensive and sophisticated EEG equipment and software used in first-class clinical settings and research labs.

By any means our study intends to be a definitive verdict or conclusion for the topic. Instead, it aims to provide an initial – but valuable – framework, upon which future studies can be built. Additionally, our study may have implications in the field by providing a helpful framework for more advanced research on the functional significance of frontal EEG asymmetry on learning and other reward-motivated behaviors such as curiosity. More broadly, given the importance of curiosity in our daily decisions and behaviors, these works could have important implications for studies in several different academic areas, including psychology, neuroscience, medicine, marketing, and education, and may contribute to the development of new strategies for improving memory and learning in both school and therapeutic settings.

Acknowledgments

The authors would like to thank Dr. Christopher Medvecky, Dr. Patrick Dolan, and Dr. Graham Cousens for their help and insightful feedback. Gabriel Lima was responsible for the conception and design of this study, data collection, analysis and interpretation of the data, and manuscript writing. Fabiana Rocha was responsible for data computing and manuscript writing. This research was made possible by the funding support from the Psychology Department at Drew University.

References

Coan, J. A., & Allen, J. J. . (2004). Frontal EEG asymmetry as a moderator and mediator of emotion. *Biological Psychology*, *67*(1–2), 7–50.

Davidson, R. J. (2004). What does the prefrontal cortex "do" in affect: perspectives on frontal EEG asymmetry research. *Biological Psychology*, *67*(1–2), 219–234.

Daw, N. D., & Doya, K. (2006). The computational neurobiology of learning and reward. *Current Opinion in Neurobiology*, *16*(2), 199–204.

Gruber, M. J., Gelman, B. D., & Ranganath, C. (2014). States of Curiosity Modulate Hippocampus-Dependent Learning via the Dopaminergic Circuit. *Neuron*, *84*(2), 486–496.

Hare, T. A., O'Doherty, J., Camerer, C. F., Schultz, W., & Rangel, A. (2008). Dissociating the role of the orbitofrontal cortex and the striatum in the computation of goal values and prediction errors. *Journal of Neuroscience*, *28*(22), 5623–5630.

Harmon-Jones, E., & Gable, P. A. (2017). On the role of asymmetric frontal cortical activity in approach and withdrawal motivation: An updated review of the evidence. *Psychophysiology*, *55*(1), e12879.

Hoffmann, B. (1972). *Albert Einstein: Creator and rebel*. New York: Viking Press.

Kang, M. J., Hsu, M., Krajbich, I. M., Loewenstein, G., McClure, S. M., Wang, J. T., & Camerer, C. F. (2009). The Wick in the Candle of Learning. *Psychological Science*, *20*(8), 963–973.

Loewenstein, G. (1994). The psychology of curiosity: A review and reinterpretation. *Psychological Bulletin*, *116*(1), 75–98.

Marvin, C. B., & Shohamy, D. (2016). Curiosity and reward: Valence predicts choice and information prediction errors enhance learning. *Journal of Experimental Psychology: General*, *145*(3), 266–272.

Metcalfe, J., & Finn, B. (2011). People's hypercorrection of high-confidence errors: Did they know it all along? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(2), 437–448.

Metcalfe, J., & Miele, D. B. (2014). Hypercorrection of high confidence errors: Prior testing both enhances delayed performance and blocks the return of the errors. *Journal of Applied Research in Memory and Cognition, 3*(3), 189–197.

Mullaney, K. M., Carpenter, S. K., Grotenhuis, C., & Burianek, S. (2014). Waiting for feedback helps if you want to know the answer: the role of curiosity in the delay-of-feedback benefit. *Memory & Cognition*, *42*(8), 1273–1284.

Pizzagalli, D. A., Sherwood, R. J., Henriques, J. B., & Davidson, R. J. (2005). Frontal Brain Asymmetry and Reward Responsiveness: A Source-Localization Study. *Psychological Science*, *16*(10), 805–813.

R Core Team. (2013). R: *A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing.

Schmid, P. C., Hackel, L. M., Jasperse, L., & Amodio, D. M. (2017). Frontal cortical effects on feedback processing and reinforcement learning: Relation of EEG asymmetry with the feedback-related negativity and behavior. *Psychophysiology*, *55*(1), e12911.

Schultz, W. (2006). Behavioral Theories and the Neurophysiology of Reward. *Annual Review of Psychology*, *57*(1), 87–115.

Vecchiato, G., Astolfi, L., De Vico Fallani, F., Toppi, J., Aloise, F., Bez, F., ... Babiloni, F. (2011). On the Use of EEG or MEG Brain Imaging Tools in Neuromarketing Research. *Computational Intelligence and Neuroscience*, *2011*, 1–12.

Rapid information gain explains cross-linguistic tendencies in numeral ordering

Emmy Liu (me.liu@mail.utoronto.ca)
Yang Xu (yangxu@cs.toronto.edu)
Department of Computer Science
Cognitive Science Program
University of Toronto

Abstract

One previously unexplained observation about numeral systems is the shared tendency in numeral expressions: Numerals greater than 20 often have the larger constituent number expressed before the smaller constituent number (e.g., *twenty-four* as opposed to *four-twenty* in English), and systems that originally adopt the reverse order of expression (e.g., *four-and-twenty* in Old English) tend to switch order over time. To explore these phenomena, we propose the view of Rapid Information Gain and contrast it with the established theory of Uniform Information Density. We compare the two theories in their ability to explain the shared tendency in the ordering of numeral expressions around 20. We find that Rapid Information Gain accounts for empirical patterns better than the alternative theory, suggesting that there is an emphasis on information front-loading as opposed to information smoothing in the design of large compound numerals. Our work shows that fine-grained generalizations about numeral systems can be understood in information-theoretic terms and offers an opportunity to characterize the design principles of lexical compounds through the lens of informative communication.

Keywords: language universals; numeral system; lexical compound; information theory; informative communication

Number is a fundamental domain of human cognition (Spelke & Kinzler, 2007), but numeral systems vary substantially across cultures (Comrie, 2013). For instance, some cultures in the Amazon lack exact numerals for expressing numbers beyond 5 (Gordon, 2004; Pica, Lemer, Izard, & Dehaene, 2018). Some languages use body parts to describe numbers (Comrie, 2013). However, the majority of languages in the world define numbers precisely and over a large range through recursive numeral systems (Comrie, 2013). Recent work has suggested that the diversity of numeral systems is constrained by the need for efficient communication (Xu & Regier, 2014; Kemp, Xu, & Regier, 2017). By this account, numeral systems are designed to facilitate highly informative communication of numbers, despite their differences in complexity.

The proposal of informative communication helps to explain why numeral systems vary the way they do, but it does not directly account for fine-grained generalizations about numeral expressions. In particular, many languages express compound numerals by specifying the larger constituent number first (e.g., *twenty-four* in English or Mandarin), and fewer languages express these in the reverse order (e.g., *vierentwintig* in Dutch, interpreted as "four twenty"). Moreover, numeral systems that originally use the reverse order of expression (e.g., Old English expresses 24 as *four-and-twenty*) tend to switch order over time (Berg & Neubauer, 2014). This preference of having the larger constituent number expressed before the smaller constituent number is prevalent in numerals for the range above 20 but less prominent for smaller numbers (Calude & Verkerk, 2016). Here we ask what principles might account for this shared tendency in numeral ordering.

This problem has been discussed by Greenberg in his cross-linguistic generalization about the design of recursive numerals (Greenberg, 1978). Recursive numeral systems represent numbers based on the canonical expression $x_1 n^k + ... + x_k n + y$. Here n is called the base and the values of x_i's and y are in the range of 1 to the base (Comrie, 2013). For numbers in the range $1-100$ in a base-10 system such as English or Mandarin, xn will be considered the *base* term (i.e., 10, 20, ..., 90) and y (i.e., 1, 2, ..., 9) will be considered the *atom* term. Greenberg observed that if a numeral system has both atom-base (e.g., *fifteen*) and base-atom (e.g., *twenty-four*) orderings in its numeral expressions, the system will always begin with atom-base, and then switch to base-atom at some number on the number line (Greenberg, 1978). In English and many other languages, this switch takes place at 20.

Independent work from Hurford has sought to address this phenomenon in light of the "packing strategy" (Hurford, 2007). According to this proposal, numeral expressions should allow one to go as far as possible along the number-line with a given set of terms (Hurford, 2007). This would imply that terms should be arranged in decreasing order, with the larger constituents coming first, and it confirms that the base-atom order should be preferred over the atom-base order. Although this work provides an intuitive theory for the ordering preference in large numerals, it leaves open two important questions: 1) why the base-atom order is preferred across languages for numbers above 20, but this preference is substantially less for smaller numerals (e.g., 11 to 19), and similarly, 2) why ordering switch should typically take place in numerals above 20 and in particular, why it occurs only in one direction (atom-base→base-atom) but not in the other (base-atom→atom-base).

We examine the problem of numeral ordering through the lens of informative communication. Consistent with the growing literature on this topic, we suggest that language design is driven by the basic need for efficient communication (Gibson et al., 2013; Kemp et al., 2017). Extending this line of research, we propose the view of *Rapid Information*

Gain (RIG) that focuses on explaining the design of compound numerals, particularly the ordering of constituent expressions in terms of the need to optimize information flow. We hypothesize that lexical ordering of a compound numeral expression should maximize information gain for the listener in the process of reconstructing the speaker's intended referent. We contrast this view with the established theory of Uniform Information Density (UID) postulating that information smoothing should be preferred (instead of information front-loading) in word ordering in sentences, online (Levy & Jaeger, 2007) or offline (Maurits, Navarro, & Perfors, 2010). We show that RIG explains empirical patterns better than UID in the domain of numerals, and we believe this work has the potential for developing a domain-general account of the design principle of lexical compounds.

Two theories of informative communication

We present the numeral ordering problem in a simple communicative scenario, illustrated in Figure 1a. Here the speaker has the target number 85 in mind and wishes to convey that number to the listener. We consider two possibilities in the ordering of constituent expressions of that numeral, using English as an example: 1) "Eighty-five", which is the *attested order* or base-atom; 2) "Five-eighty", which is the *alternate order*, or atom-base in this case. The problem is to determine which order should be generally preferred in natural languages and in what range of the number line this preference might be most prominent.

We postulate that the preferred numeral order should tend to minimize the listener's uncertainty in reconstructing the target number as the speaker's utterance is processed. We consider how uncertainty arises over time in the listener's mind as the constituent expressions are uttered sequentially by the speaker. Based on the ordering of "eighty-five", upon hearing the first constituent "eighty", the listener would consider numbers in the range 80-89 as possible candidates for the target, because numerals for numbers within that range all begin with the same constituent. In this case, uncertainty depends on the probability ratio between the actual target and the candidate set. Based on the ordering of "five-eighty", upon hearing the alternative first constituent "five", the listener would instead consider numerals that begin with "five" (e.g., 5, 15, ..., 85, 95) as the candidate set for the target. We illustrate these alternative candidate sets in Figure 1a.

We consider two alternative theories that quantify uncertainty given choices of numeral ordering based on Shannon's information theory (Shannon, 1948). The first view is based on Uniform Information Density (Levy & Jaeger, 2007), which predicts that uncertainty incurred should be as smooth as possible. This view suggests that the listener would experience a uniform information flow as a compound expression is uttered. We propose a second view, Rapid Information Gain, that makes the alternative prediction. We hypothesize that the preferred order in compound numerals should tend to front-load information as opposed to smoothing information, such that uncertainty in the listener can be reduced as quickly as possible. We illustrate the predicted uncertainty profile from each theory in Figure 1b. As we show later, the property of information front-loading is more salient in the ordering of larger numbers (>20) than in the case of smaller numbers, which explains why the cross-linguistic preference and the ordering switch toward base-atom expressions are stronger for larger numbers. We now describe the details of each theory.

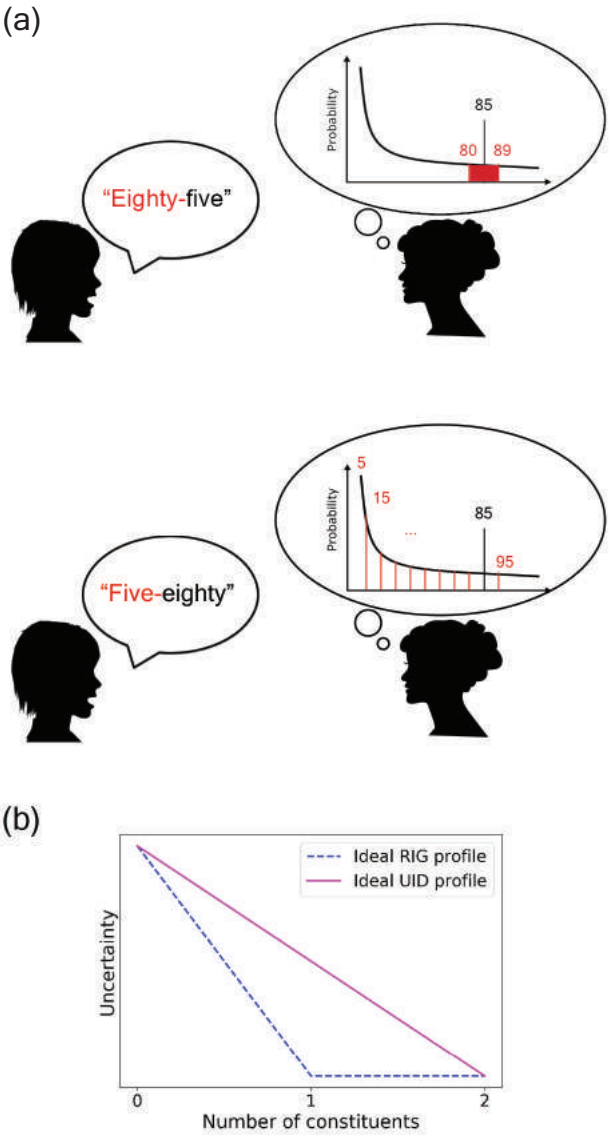


Figure 1: Illustration of the numeral ordering problem and the two theoretical proposals of informative communication.

Uniform Information Density (UID). Following Shannon (1948), we define uncertainty by surprisal or negative log probability $-log_2(p(\cdot)) = log_2(\frac{1}{p(\cdot)})$. We define the information content of a compound linguistic expression by the sum of surprisals from its sequential constituents, following the

formulation of UID (Levy & Jaeger, 2007). The cumulative information conveyed by an expression U with n constituents $w_1...w_n$ in reference to a target t is the following:

$$log_2\frac{1}{p(U)} = log_2\frac{1}{p(t)} + log_2\frac{1}{p(t|w_1)} + ... + log_2\frac{1}{p(t|w_1...w_n)} \quad (1)$$

In the case of two-constituent numeral expressions such as *twenty-four* (i.e., base constituent and atom constituent), this formulation effectively captures the information flow of a compound numeral as it is processed incrementally in terms of its constituent expressions:

$$log_2\frac{1}{p(t)} \rightarrow log_2\frac{1}{p(t|w_1)} \rightarrow log_2\frac{1}{p(t|w_1w_2)} \quad (2)$$

Cumulative surprisal defined in Equation 1 can thus be simplified to

$$log_2\frac{1}{p(U)} = log_2\frac{1}{p(t)} + log_2\frac{1}{p(t|w_1)} + log_2\frac{1}{p(t|w_1w_2)} \quad (3)$$

As such, the cumulative surprisal of hearing "twenty-four" would be $log_2\frac{1}{p(\text{“twenty-four”})} = log_2\frac{1}{p(24)} + log_2\frac{1}{p(24|\text{“twenty”})} + log_2\frac{1}{p(24|\text{“twenty-four”})}$.

Empirical studies of UID typically focus on speaker information modulation given the predictability of different units. This would involve measuring information-theoretic entropy rather than surprisal formulated here. However, the UID principle implies that the flow of information to follow a uniform trajectory in cumulative surprisal, and we test the applicability of this proposal in the case of numeral ordering.

More specifically, UID suggests an even distribution of information (in the design of compound numerals), such that the amount of information conveyed in the sequence of constituents should be identical. This predicts that if the speaker has alternative ways of ordering a numeral expression, she should choose the order in which information is distributed more evenly. Here we are interested in the cost of a numeral order versus its reverse order, and we quantify cost by measuring how a numeral order deviates from the theoretical UID information flow. Prior work has taken a similar approach to examine whether UID predicts preferred word orders (e.g., subject-verb-object) across languages (Maurits et al., 2010). In that work, deviation from UID is defined by the percentage deviation from the theoretical UID information flow. Abbreviating the components of the information flow in Equation 2 by $I_0 = log_2(\frac{1}{p(t)})$, $I_1 = log_2(\frac{1}{p(t|w_1)})$ ..., we measure the deviation from UID following Maurits et al. (2010):

$$d = \frac{n}{2(n-1)}\sum_{i=1}^{n}|\frac{I_{i-1}-I_i}{I_0} - \frac{1}{n}| \quad (4)$$

Here n is the phrase length of an expression (Maurits et al., 2010). In our work, we use the same formula to quantify how the design of a numeral expression deviates from UID. Concretely, we consider $n = 2$ because each compound numeral expression that we use for analyses has two constituents. We also know that $I_2 = 0$ since full certainty is obtained after the second (or last) constituent of a numeral is uttered. UID predicts a linear relationship between information content and number of constituents. If UID explains the shared tendency in numeral ordering across languages, we should expect the attested numeral order to yield a smaller deviation from the linear information profile than the alternate order, more so for the numerical range above 20 than the range under 20.

Rapid Information Gain (RIG). We propose an alternative theory for numeral ordering based on rapid information gain. We postulate that the ordering of numerals should facilitate quick delivery of information to the listener, such that the constituent expression that contains more information should be arranged prior to the constituent that contains less information. This notion of rapid information gain is related to work on optimal data selection. For instance, when performing a series of tasks, optimal data selection implies that people should order the tasks so that they gain the most information possible at each step (Oaksford & Chater, 2003). We believe that similar principles apply to the design of numerals. Our proposal is not equivalent to the claim that the larger numeral should always precede the smaller numeral in a compound (Hurford, 2007; Berg & Neubauer, 2014). Instead, it suggests that the ordering of constituent numerals depends on the amount of information they convey, as opposed to their magnitudes per se. We demonstrate later that our proposal correctly predicts information front-loading to be more critical for high-order numbers than low-order numbers, an aspect that could not be explained fully by a magnitude account that always predicts the larger numeral to be expressed first in a compound numeral.

We evaluate our proposal by measuring the cumulative surprisal of a numeral expression over its constituents:

$$c = \sum_{i=0}^{n} I_i \quad (5)$$

This formulation is the same as Equation 3, and we consider $n = 2$ and since $I_2 = 0$, $c = I_0 + I_1$. The RIG theory predicts an elbow-like information profile which differs from the linear profile predicted by UID (see illustrations of the two theoretical information flows in Figure 1b). We expect that a lower cumulative surprisal should generally be preferred as a consequence of rapid information gain. More specifically, the attested numeral order should yield a lower cumulative surprisal than the alternate order when there is a strong preference toward the attested order (e.g., for numbers >20), but the two possible orders might yield similar cumulative surprisals when there is greater flexibility in the ordering conventions of numerals across languages (e.g., for numbers <20).

Materials and methods

To facilitate the information analyses and evaluation of the two theories, we collected numeral frequencies for estimating surprisals along with cross-linguistic numeral data.

Numeral frequencies. We estimated probabilities of the number terms for the range 1-100 (following Xu & Regier, 2014) in 8 different languages: English, French, German, Hebrew, Italian, Mandarin, Russian, and Spanish. We collected these frequency data from the Google Ngrams corpora (Michel et al., 2011) by averaging numeral frequencies from 1900 to 2000. We used part-of-speech tags for numerals in the corpus if those were available for a given language. For each language, we queried frequencies of numeral terms from a standard set of numeral expressions (data from `www.sf.airnet.ne.jp/ts/language/number.html`). When multiple expressions were available for a numeral, we took the most frequent expression. The frequencies of the numerals for each of the languages were normalized to probabilities so that they sum to 1.

Calculation of surprisals. To calculate surprisals, we decomposed a numeral expression into two separate constituents, atom and base, while ignoring connectives such as hyphens, e.g., "twenty-one" $\rightarrow$ ["twenty", "one"]. Although it is possible to split some terms into multiple constituents, e.g. "quatre-vingts huit" (4x20+8=88) $\rightarrow$ ["quatre", "vingts", "huit"] ([4, 20, 8]), we chose to split only along additive terms for consistency. We did not choose to treat suffixes as separate constituents. We calculated the surprisal based on each constituent expression, where surprisal is the negative log probability of the target number being correctly inferred from the set of candidate targets. Finally, for each numeral expression we computed the deviation from UID according to Equation 4 and the cumulative surprisal for RIG according to Equation 5.

Cross-linguistic numeral data. We tested the theories against numeral data collected from 334 languages in 53 listed language families sampled from *Numeral Systems of the World's Languages* (Comrie & Chan, 2018). We sampled languages evenly from each family whenever possible, taking 10 from each family, or if 10 were not available, taking the maximum number possible. This was so that language families with a large number of languages such as Indo-European or Sino-Tibetian did not bias the sample. For each language, we recorded the attested orders in the numeral expressions, atom-base or base-atom, for the numerical ranges of 11-19 and 21-29 (chosen to be symmetric about 20 where order switch most commonly takes place). If a language did not have sufficient data for the numerical ranges, we would exclude that language and sample other languages from the family until 10 or the maximum possible number were collected.

Results

Empirical patterns in the ordering of numerals. We first present cross-linguistic tendencies and switches in "atom-base" and "base-atom" ordering of numeral expressions in the sample of 334 languages that we considered. Table 1 summarizes the cross-linguistic occurrences for these orders in the numerical ranges 11-19 and 21-29. If the atom-base ordering was used for at least one term in 11-19 in a language, we considered that language as having an atom-base ordering in that range. We observed that the base-atom order is attested in more than 96% of the languages for the range 21-29, whereas this order is attested much less commonly in about 76% of the languages for the lower range 11-19. This finding confirms descriptive generalizations from previous work (e.g., Greenberg, 1978) and indicates an asymmetric preference toward base-atom ordering in larger numerals, and more flexibility in the ordering of smaller numerals.

Table 2 confirms that the same asymmetric preference applies to switches in the ordering of numerals. In particular, out of all languages that were attested to have switched order in numeral expressions, switch took place exclusively in the direction atom-base $\rightarrow$ base-atom but not in the opposite direction. Moreover, out of the 63 languages that use the atom-base order for expressing the numerical range 11-19, 52 (or $\sim$83%) switch the order to base-atom but only for numerals expressing the range 21-29. Together, these empirical data suggest that preference toward the base-atom order is more prominent in larger but not smaller numerals.

Numeral frequencies across languages. Figure 2 summarizes the meta-mean and language-specific probabilities of numerals, estimated from the corpus-based frequencies over the past 200 years. These probability profiles show a consistent near-logarithmic decay that confirms previous findings in cross-linguistic numeral and digit-based frequencies (Greenberg, 1978; Calude & Verkerk, 2016): Numerals in the lower numerical range tend to be referred to more frequently than numerals in the higher range. We used these probabilities for surprisal calculations for the two theories.

Table 1: Ordering conventions in numerals across languages.

Number of languages	Range 11-19	Range 21-29
atom-base ordering	63	11
base-atom ordering	271	323

Table 2: Switch in numeral ordering conventions. For each language, the original numeral order is the same as that in the lower range 11-19, and ordering switch is attested in numerals for the upper range 21-29.

Number of languages	No switch	Switched
atom-base $\rightarrow$ base-atom	11	52
base-atom $\rightarrow$ atom-base	271	0

Evaluation of the two theories. We evaluated UID and RIG by first considering a "template" language that reflects the cross-linguistic tendency in numeral ordering we and other scholars have observed: Atom-base order in numerals

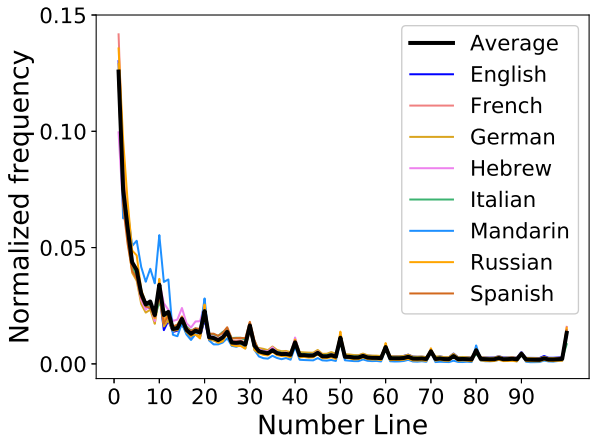


Figure 2: Numeral frequencies across 8 languages.

for the range 11-19, and base-atom order in numerals for the range above 20. An ideal theory should explain 1) why there is a strong preference for the attested base-atom order over the alternate atom-base order in the upper numerical range; 2) why this preference between the attested and alternate orders is much weaker in the lower range. As such, we expected a greater discrepancy in the attested and alternate orders for the theoretically predicted information profile we described (e.g., under UID or RIG), and a substantially smaller discrepancy in these orders for the same measure of information. To test these ideas, we calculated the information profile for each of the numerals within the range 1-100 based on the mean numeral probabilities we had obtained. We performed these calculations for both the attested order and the alternate order, resulting in two sets of measures for UID deviation and two sets of measures for RIG cumulative surprisal.

Figure 3 (a) and (d) summarize the results. At the broad level, both UID and RIG identify the attested order to be closer to their theoretical information profiles than the alternate order. However, a closer examination of these results reveals variation in the precision of these theories. For the numerical range beyond 20, UID shows an ambivalent preference toward the base-atom order over the atom-base order, manifested in the noisy deviation scores between the two orderings. In contrast, RIG provides a clearer advantage of the base-atom order over the atom-base order for numerals in the same range, indicating that there is a dominance toward the first order as predicted by this theory. Moreover, for numerals in the range 11-19, UID shows a strong support for the base-atom order, but RIG shows that both orderings render roughly equal cumulative surprisals—this suggests that information front-loading is less relevant to ordering variation in this lower numerical range.

To further examine the precision of the two theories, we examined their predictions for two sample languages, English and Mandarin. For these cases, we used language-specific numeral probabilities for calculations of UID deviation and RIG cumulative surprisals. Figure 3 shows that the results for these individual languages are consistent with our findings with the template language, such that RIG provides a more precise explanation for the asymmetric preference in ordering of larger and smaller numerals. Figure 4 illustrates the information profiles in the attested and alternate orders with two example numerals, *fifteen* and *twenty-four* in English, along with the theoretical predictions from UID. In both cases, the attested order shows an elbow-like information profile that deviates from the ideal linear profile of UID, providing evidence against the idea that numerals are designed under the criterion of information smoothing. Importantly, the information profile under the alternate order for *fifteen*—a low-order numeral—is almost identical to the elbow-like profile under the attested order, reflecting the fact that information front-loading is insensitive to ordering of numerals in this range. It is worth noting that both alternate and attested profiles deviate from the UID prediction. In addition, for *twenty-four*, the alternate order produces an information profile that approaches the UID prediction. This profile yields a cumulative surprisal higher than the attested order, suggesting information front-loading is desirable for larger numerals in English.

As a final analysis, we examined whether the preferred ordering switch from atom-base to base-atom can be explained away by the theory of RIG. In particular, we performed a focused analysis that compares cumulative surprisal between these two orders for the numerical ranges 11-19 and 21-29 respectively. We expected that the cumulative surprisal might be comparable under the two orders for the smaller range, but substantially discrepant for the larger range, which would explain why switching of order tends to occur beyond 20 and only in the atom-base $\rightarrow$ base-atom direction.

For each of the numerical range in question, we conducted a permutation test that shuffles the numeral expressions between the base-atom and atom-base orders. We then repeated the shuffle 100,000 times and for each repetition, calculated the mean difference in cumulative surprisal between the two orders. This effectively helped construct the null hypothesis that there should be no between-order difference in cumulative surprisal. We also calculated the same quantities for the unshuffled data, and compared those against the null distributions for the two numerical ranges of interest. Figure 5 shows that there is no statistical significance ($p = 0.56$) to reject the null for the range 11-19, but there is high statistical significance ($p < 0.004$) in rejecting the null for the range 21-29. These results provide evidence for the idea that information front-loading is equally prominent under atom-base or base-atom orderings for smaller numerals, but it is more prominently represented in the base-atom order as opposed to the atom-base for larger numerals. Possibly due to this reason, historical changes in ordering convention of numerals tend to occur uni-directionally beyond but not below 20.

Discussion

We investigated two theories for explaining the shared tendency in the ordering of numeral expressions. We found that

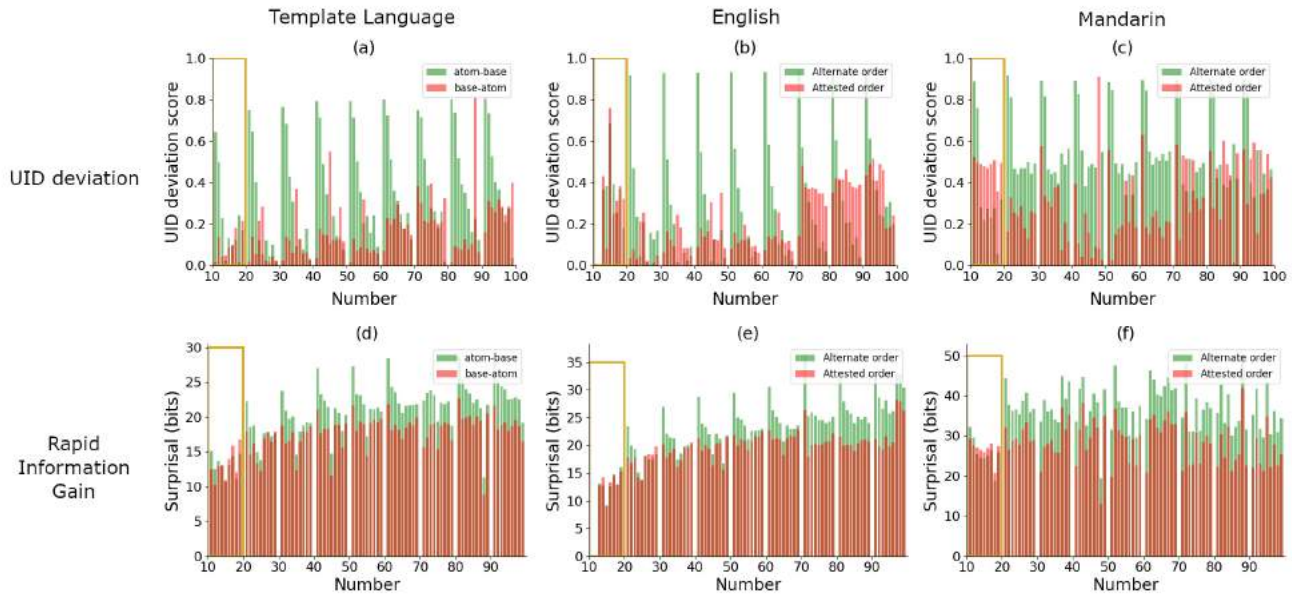


Figure 3: UID deviation (top row) and cumulative surprisal (top row) for template language, English, and Mandarin.

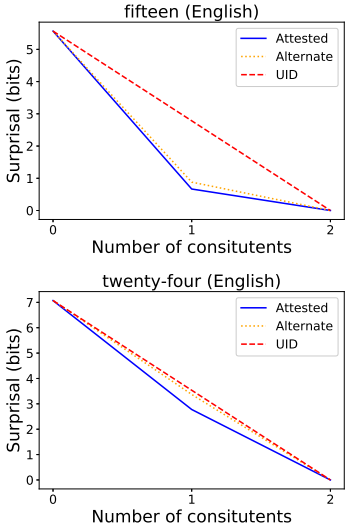


Figure 4: Information flows under alternative orders of expression for English numerals 15 and 24. The attested order for 15 is atom-base ("fifteen"), and the alternate order is base-atom ("teenfif"). The attested order for 24 is base-atom ("twenty-four"), and the alternate order is atom-base ("four-twenty"). "UID" refers to the UID theoretical prediction.

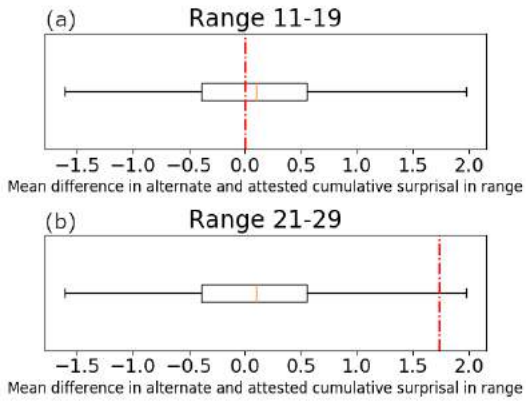


Figure 5: Permuted surprisals with attested ones marked.

the proposal of rapid information gain provides a better account for the empirical data across languages than the existing theory of uniform information density. Our findings suggest that the dominant preference toward the base-atom ordering in larger numerals reflects the need for information front-loading as opposed to information smoothing, and greater flexibility in the ordering of smaller numerals is explained partly by the fact that information flow is less affected by ordering conventions in numerals for the lower range. Our study differs from existing research in UID that focuses on information processing at the sentence level. Our emphasis is to characterize the design principles of complex lexical items, particularly compounds. This difference in the level of analysis might provide one explanation as to why UID does not predict as well in the current study. An alternative possibility is that the domain of numerals has characteristics that make a uniform information flow less desirable than information front-loading. Future research should delineate when UID might apply and when alternative principles such as RIG are more appropriate. It is also worth exploring whether the RIG principle can be applied to compounds in other domains.

Acknowledgements

We would like to thank Blair Armstrong and Suzanne Stevenson from the University of Toronto and Terry Regier from the University of California, Berkeley for their constructive comments on the manuscript. This research is supported by an NSERC DG grant and a Connaught New Researcher Award to YX.

References

Berg, T., & Neubauer, M. (2014). From unit-and-ten to ten-before-unit order in the history of English numerals. *Language Variation and Change*, *26*, 21–43.

Calude, A. S., & Verkerk, A. (2016). The typology and diachrony of higher numerals in Indo-European: A phylogenetic comparative study. *J Lang Evol*, *1*, 91-108.

Comrie, B. (2013). Numeral bases. In M. S. Dryer & M. Haspelmath (Eds.), *The world atlas of language structures online.* Leipzig: Max Planck Institute for Evolutionary Anthropology.

Comrie, B., & Chan, E. (2018, Oct 30). *Numeral systems of the world's languages.* Retrieved from `https://mpi-lingweb.shh.mpg.de/numeral/`

Gibson, E., Piantadosi, S. T., Brink, K. A., Bergen, L., Lim, E., & Saxe, R. (2013). A noisy-channel account of crosslinguistic word-order variation. *Psychol Sci*, *24*, 1079-88.

Gordon, P. (2004). Numerical cognition without words: Evidence from amazonia. *Science*, *306*, 496–499.

Greenberg, J. H. (1978). Generalizations about numeral systems. In C. A. F. Joseph H. Greenberg & E. A. Moravcsik (Eds.), *Universals of human language, volume 3: Word structure* (Vol. 3, p. 249295). Stanford University Press.

Hurford, J. R. (2007). A performed practice explains a linguistic universal: Counting gives the packing strategy. *Lingua*, *117*, 773–783.

Kemp, C., Xu, Y., & Regier, T. (2017). Semantic typology and efficient communication. *Annual Review of Linguistics*, *4*, 109–128.

Levy, R. P., & Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. In *Advances in neural information processing systems 19* (pp. 849–856).

Maurits, L., Navarro, D., & Perfors, A. (2010). Why are some word orders more common than others? A uniform information density account. In *NIPS 23.*

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., , ... Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, *331*, 176–182.

Oaksford, M., & Chater, N. (2003). Optimal data selection: Revision, review, and reevaluation. *Psychonomic Bulletin & Review*, *10*, 289–318.

Pica, P., Lemer, C., Izard, V., & Dehaene, S. (2018, 10). Exact and approximate arithmetic in an Amazonian indigene group with a reduced number lexicon. , 499-503.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*, 379–423.

Spelke, E. S., & Kinzler, K. D. (2007). Core knowledge. *Developmental Science*, *10*, 89-96.

Xu, Y., & Regier, T. (2014). Numeral systems across languages support efficient communication: From approximate numerosity to recursion. In *CogSci 36.*

Why Some Verbs are Harder to Learn than Others – A Micro-Level Analysis of Everyday Learning Contexts for Early Verb Learning

Siyun Liu (liusy@mail.ccnu.edu.cn)
Key Laboratory of Adolescent Cyberpsychology and Behavior, Ministry of Education
School of Psychology, Central China Normal University
No.152 Luo Yu Road, Wuhan, Hubei, 430079, China

Yayun Zhang (yayzhang@indiana.edu)
Chen Yu (chenyu@indiana.edu)
Department of Psychological and Brain Sciences, Indiana University- Bloomington
1101 E. 10th Street, Bloomington, IN, 47405, USA

Abstract

Verb learning is important for young children. While most previous research has focused on linguistic and conceptual challenges in early verb learning (e.g. Gentner, 1982, 2006), the present paper examined early verb learning at the attentional level and quantified the input for early verb learning by measuring verb-action co-occurrence statistics in parent-child interaction from the learner's perspective. To do so, we used head-mounted eye tracking to record fine-grained multimodal behaviors during parent-infant joint play, and analyzed parent speech, parent and infant action, and infant attention at the moments when parents produced verb labels. Our results show great variability across different action verbs, in terms of frequency of verb utterances, frequency of corresponding actions related to verb meanings, and infants' attention to verbs and actions, which provide new insights on why some verbs are harder to learn than others.

Keywords: verb learning, motion verb, attention, head-mounted eye-tracking, infant-parent dyads

Introduction

Language learning depends on both the internal learning mechanisms and the data on which those mechanisms operate. Many experimental studies have focused on examining the internal learning mechanisms by using well-controlled and well-balanced stimuli as the input. A recent trend in the field of language acquisition is to examine natural statistics in everyday learning contexts (e.g. Pereira, Smith, & Yu, 2014). For example, recent studies have shown that both the quantity and quality of parent language input are predictive of children's later language development (Hart & Risley, 1995; Hoff, 2003; Weisleder & Fernald, 2013). In the present study, we used the same approach to examine the input for early word learning. One of the challenges in early word learning is to figure out the correct mapping between a word and a referent (Quine, 1960). Given many possible referents in the moment when a word is heard, young learners need to attend to the right referent at the right time in order to learn the meaning of a word. However, we do not yet know what input from the environment is available to the child and what input attended by the child is therefore processed by the internal learning mechanisms.

A large proportion of early vocabulary is composed of concrete nouns and concrete verbs. Previous studies on learning concrete nouns found that children need to select and attend to the right object at the right time from an ambiguous learning environment when hearing its name (Golinkoff, Hirsh-Pasek, Bailey, & Wenger, 1992; Yurovsky, Smith, & Yu, 2013; Yu & Smith, 2007; Gleitman & Trueswell, 2018). In addition, Pereira, Smith and Yu (2014) found that when the named target is visually large and more centered in the child's view and when these optimal visual properties last longer before and after parent's naming, children are more likely to attend to the named object and learn its label. Thus, learning object names with perceptually grounded meanings requires not only hearing the words from parent speech but also showing sustained attention to the intended referent. However, little is known about whether learning concrete verbs also requires young learners' sustained attention when mapping verbs to visually grounded actions. Most experimental studies on verb learning have been focused on testing how well young children build verb-action mappings when presented with a verb and an action in well-controlled laboratory settings (Imai et al., 2008; Hirsh-Pasek & Golinkoff, 1996; Maguire et al., 2008; Golinkoff et al., 2002; Pulverman, Golinkoff, Hirsh-Pasek, & Buresh, 2008; Monaghan, Mattock, Davies, & Smith, 2015, Messenger, Yuan, & Fisher, 2015; Scott & Fisher, 2012). The learning tasks for young children in those experimental setups were well-controlled to minimize distraction, which is very different from learning verbs in the real world. Referential uncertainty created during naturalistic interactions may be different from that created for traditional lab tasks, thus it may influence how children process information differently.

Imagine a naturalistic context for early verb learning such as toy play, when a parent names a verb (e.g. "Can you shake it?") while demonstrating the shaking action. The meaning of "shake" is presented briefly as the parent is not likely to keep shaking the object. If the infant does not attend to the action when hearing the word "shake" and when the action is produced, it would be impossible for the infant to build the association between the word "shake" and the action "shake". This example reflects the transient nature of the action referent and lead to important research questions related to early verb learning that have not been examined at the

perceptual and attention levels. For example, compared with object names, how frequently do parents mention action verbs in their speech in everyday learning contexts? When parents produce a verb in speech, how likely there is a corresponding action in the learning environment that reveals the meaning of the verb? If there is an action in accompany with parent speech, how likely do infants attend to the action to build a verb-action mapping?

To answer these questions, we need to examine parents' and children's behaviors from natural learning environments. We used head-mounted eye-tracking techniques to record fine-grained multimodal behaviors during parent-infant joint play. We analyzed parent speech, parent and infant action, and infant attention at the moments when parents produced verb labels. By doing so, we will be able to provide new evidence on how easy or hard for young children to learn early verbs and discover new elements -- at the attentional level -- that matter to early verb learning. Our overarching goal was to quantify word-referent co-occurrence statistics in parent-child interaction from the learner's perspective and examine what information infants select to attend when a verb is heard.

Method

Participants

Thirty-three infant-parent dyads with infants (12 female) ranging from 15.2 to 25.3 months (M = 19.52, SD = 2.42) were included in the final sample.

Stimuli and Experimental Setup

Parents and their infants were invited to play with a set of 24 toys in a playroom (Figure 1A). The toys were randomly spread out across the floor at the beginning of each play session. Parents and infants both sat on the floor and parents were told to sit in any orientation with their child but were instructed to try to keep their child sitting on the ground as much as possible during the play session. We observed that parents and infants naturally generated various types of manual actions during toy play. For example, they used a toy saw to pretend to cut other objects; they put a doll on a toy bed; they played with a car toy to generate actions like turning; and they stacked one toy on top of others, etc. While playing, parents also verbally described those manual actions generated by themselves or by infants.

Figure 1A: Experimental setup

Figure 1B: Examples from the infant egocentric view. The crosshair in each example indicates the infant's gaze direction.

Eye-tracker and Calibration

Parents and infants wore head-mounted eye trackers (Positive Science LLC). The tracking system has been successfully used in both infant and adult experiments (Franchak & Adolph, 2010; Yu & Smith, 2017). The eye-tracking system includes an infrared camera mounted on the head and pointed to the right eye of the participant that records eye images and a scene camera that captures and records images from the participant's perspective. The visual field of the scene camera is 108° (Figure 1B). Each tracking system – the infants' and parents' – recorded egocentric video and the x- and y-position of the right eye in the captured scene at a sampling rate of 30Hz. For eye-tracker setup, one experimenter engaged with the infant with an enticing toy while the second experimenter affixed the eye-tracker on the parent. After the parent's eye-tracker was secure and the scene and eye cameras were properly adjusted and oriented, both experimenters and the parent worked together to place the headgear and eye-tracker on the infant. The parent and one of the experimenters played with the infant while the other experimenter placed the infant's headgear (a small hat with Velcro stickers on the forehead) on the infant.

Instructions and Procedure

After the calibration phase, one of the experimenters distribute the set of toys on the floor and leave the parent and infant to play. The experimenters watch the interaction in an adjoining room and monitor the parent's and infant's eye and scene live streaming videos. If infants touch the camera or bumped the camera with a toy, the experimenter would go into the room, readjust the cameras, complete a new calibration phase, and leave the room so the parent and infant could complete the rest of the toy play session. Parents were asked to engage with their infants and toys as naturally as possible for ten minutes.

Data Annotation

Parent speech and infants' egocentric video were used in data analysis. We first transcribed speech and then identified spoken utterances containing action verbs. For those utterances, we further coded subject, verb, and (direct and/or indirect) object for each verb utterance. Since the main interest of this paper is on early verb learning and most verbs learned early by young children are action verbs, we focused only on action verbs with concrete meanings that can be revealed by manual actions (e.g. stack and shake) instead of abstract verbs (e.g. think and imagine).

For each parent utterance containing an action verb, we defined a window ranging from 3 seconds before to 3 seconds after the verb was generated. Within this temporal window, we first coded whether an action event was accompanied by the action verb, using infants' egocentric video. For example, in Figure 2, when a parent said, "shaking it", whether the parent or the infant used an object to generate a "shaking" action at the same time. If so, we next coded which target object was action-related for the action event. Figure 2 showed three example verb utterances, two accompanied by an action, and the other without any action. For the two utterances with action (Figure 2B, Row 1), we also coded target objects at the moment (Figure 2B, Row 2). Finally, an in-house coding program was used to code frame by frame which object infants attended moment by moment and gaze data were used to measure infants' attention when hearing verb utterances (Figure 2B, Row 3).

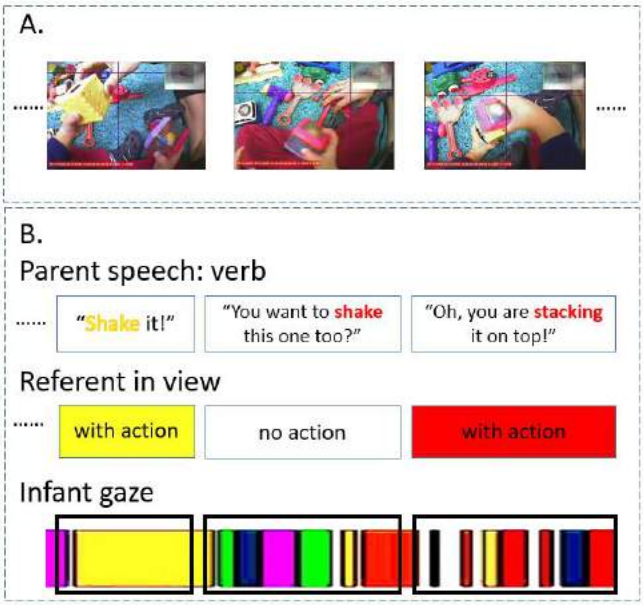


Figure 2: (A) Infants' first-person view and point of gaze during verb utterances. Purple crosshair in the image indicates where the infant was attending in the first-person view. (B) Row 1: Speech transcription of parents' verb utterances; Row 2: Coding of whether a verb utterance was accompanied by an action revealing the meaning of the verb. Colors indicate which object is carrying out the action; Row 3: Gaze coding. Different colors in the infant gaze stream indicate different objects attended by the infant moment by moment. If an infant attended to the named object, the colors in Row 2 and 3 would match in time.

Results

Verb utterance in parent speech. Parent speech contains 4406 utterances (1498 contain object names, 1381 contain action verbs). On average, parents generate roughly the same amount of nouns (5.09 nouns/min) and verbs (5.12 verbs/min, $t(32) = .72$, $p = .47$, ns). Among all the verbs, 705 were action verbs and 268 were abstract verbs. Thus, action verbs took roughly 72.4% of the total of 973 verbs, suggesting that parents most often used concrete verbs in their speech when they played with their children. Among all the action verbs that were coded from parent speech, we selected the top 25 verbs with relatively high frequency (except "look" and "see" as these two verbs were mostly used for attention getting in free play) to form a list of target action verbs for further data analysis. Figure 3A shows a skewed frequency distribution of those top 25 action verbs with two statistical properties. First, even for those top 25 verbs, most of them were produced fewer than 30 times, suggesting that a large proportion of those action verbs were hardly repeated by parents in a play session. Second, the skew distribution also revealed that some verbs were mentioned in parent speech much less frequently than others. Both the frequency difference between action verbs and object nouns, and the variability within action verbs suggest that those quantitative discrepancies are one of the many reasons why (some) verbs might be harder to learn than nouns.

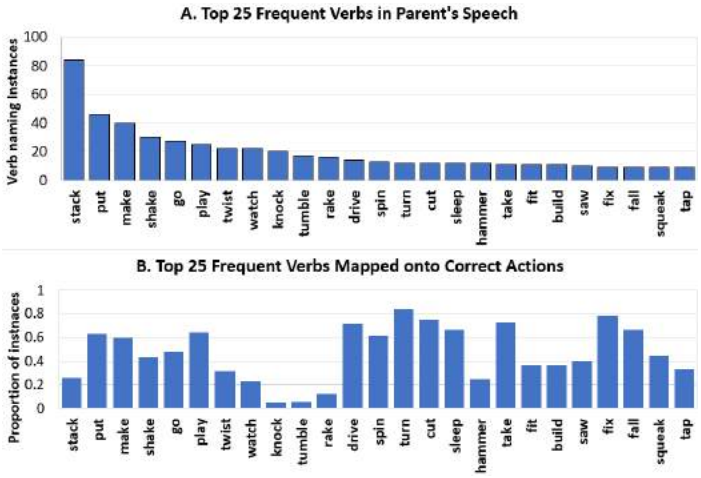
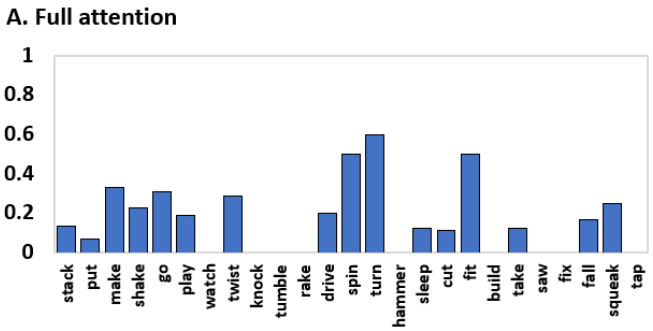


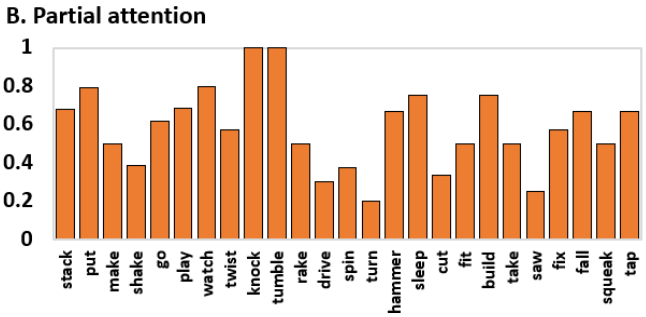
Figure 3: (A) A skewed frequency distribution of the top 25 action verbs. (B) A distribution of the percentage of verb utterances that were accompanied by an action.

Verb-Action co-occurrence. Learning the perceptually grounded meaning of an action verb requires not only hearing the verb but also perceiving the action. One critical question is how often a verb and its corresponding action co-occur in the learning environment? We answered this question by directly measuring verb-action co-occurrence and counting how often an action revealing the meaning of a verb was generated – either by parents or infants – when parents produced a verb utterance. Figure 3B showed the percentage of verb utterances that were accompanied by the corresponding action. There are two noticeable patterns. First, there is variability in verb-action co-occurrence across action verbs as when some verbs (e.g. "drive", "turn", "cut") were mentioned in parent speech, it was very likely that the corresponding actions were also generated at the same time; while for some other verbs (e.g. "knock", "tumble", "rake"), they were produced in parent speech most often without the corresponding actions. In those situations, either parents failed to demonstrate the corresponding action while a verb was generated, or parents failed to name the actions

conducted by infants. Second, there is no correlation between verb frequency and verb-action co-occurrence (r = -0.161, *n.s.*), suggesting that producing action verbs more frequently would not necessarily create more verb-action co-occurrences, which are critical for building verb-action mappings than just hearing action verbs alone.

Attention to verb-action co-occurrence. Infants' visual attention in free play was dynamic as they sometimes followed parents' attention and sometimes went with their own goals. Even with the presence of verb-action co-occurrence in the learning environment, they may or may not attend to the action when hearing a verb label. Given that attending the corresponding action when hearing a label is critical for verb learning, we next measured the proportion of infant gaze attention on the corresponding action within a verb utterance. Prior research shows that infants' learning of an object name depends on sustained visual attention to the object during a window that lasts from the onset of the utterance containing the name to several seconds after the offset of the utterance (Yu & Smith, 2012). Therefore, we operationally defined a verb event starting at the onset of a parent verb utterance and lasting for 3 seconds – the temporal interval including both the utterance itself (on average 1.5 sec long) and roughly 1.5 seconds after the utterance. We quantified infants' attention during and after hearing a verb utterance by defining three attentional states based on infant gaze: **Full attention** -- infants attended to the action 100% of time within a 3s window; **Partial attention** -- infants attended to the action sometimes but also to elsewhere when hearing a verb label. **No attention** -- infant did not attend to the action at all. Figure 4 showed the percentages of verb-action co-occurrences that received full attention (4A), partial attention (4B) or no attention (4C) from infants. As observed in the distributions of verb utterance frequency and verb-action co-occurrence, there is large variability among different action verbs. Infants seemed to attend to some actions (e.g. "turn" and "spin") much more than others (e.g. "saw" and "rake") when hearing verb labels. Also, in most cases, they seemed to attend to the correspond action sometimes but not the whole time within a 3s window as the percentages in partial attention are much higher than the percentages in full attention and no attention.





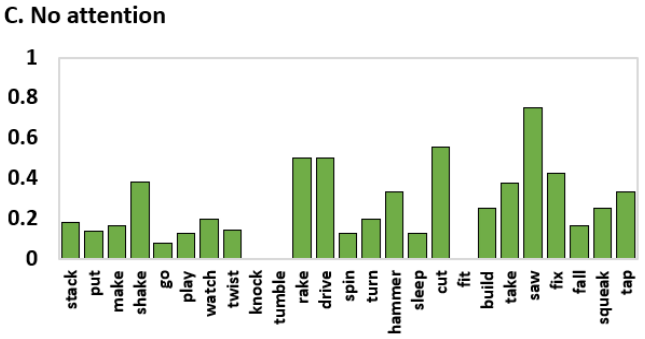


Figure 4: The distribution of infants' attention on the target of the actions that matches with the verbs of the top 25 verbs. (A): Infants' full attention on the targets; (B): Infants' partial attention on the targets; (C): Infants' no attention on the targets.

If an action verb was mentioned more frequently in parent speech, would a higher frequency attract infant attention more on the corresponding action when it was available in the environment? To answer this question, we correlated both verb utterance and verb-action co-occurrence with the three attentional states. As shown in Table 1, we found no correlation between verb frequency and infant attention. Producing more verb utterances did not attract infant attention more toward actions when those verb utterances were accompanied by the corresponding actions. However, there is a significant correlation between verb-action co-occurrence and full attention as shown in both Table 1 and Figure 5, suggesting that infants were more likely to attend to the action 100% of time when a verb and its corresponding action consistently co-occurred together. The higher percentage that manual actions and verb labels co-occurred together; the more likely infants showed full attention to the action event when hearing its label.

Table 1: The correlations between infants' attention, and verb utterance and verb-action co-occurrence (*p<0.05)

	verb utterance	verb-action co-occurrence
full attention	0.015	**0.475***
partial attention	0.206	-0.386
no attention	-0.242	-0.015

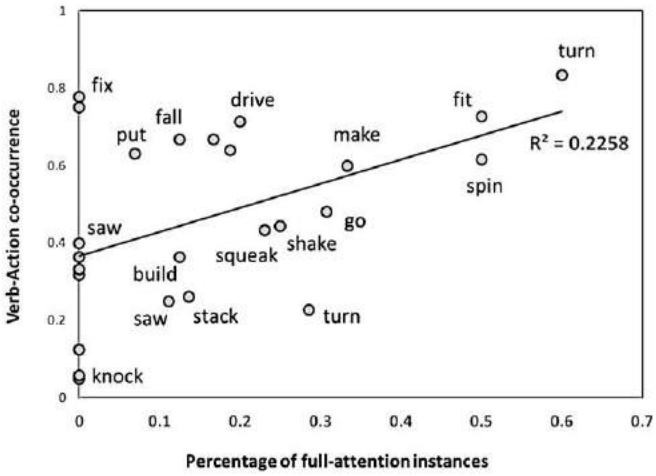


Figure 5: A significant correlation between verb-action co-occurrence and full attention.

Discussion

Recent studies show the overall speech input perceived by the young learner is predictive to later learning outcomes (Weisleder & Fernald, 2013). The input for learning concrete verbs includes not only spoken words but also their perceptually grounded meanings to build word-referent mappings. In light of this, the present study quantified both speech input and verb-action co-occurrences in the learning environment. Critically, the input to the language learning system is not the objective properties available in the learning environment but instead the information in the environment selected by the learner. Therefore, we measured the statistical regularities from the learner's perspective by using the learner's gaze data. Our result showed that the overall frequency distribution of verb generated by the parent during free play is right-skewed, which is similar to what has been observed in the recent studies of object names (Smith, Jayaraman, Clerkin, & Yu, 2018; Bambach, Crandall, Smith, & Yu, 2018). Moreover, the mere frequency of verb utterances was not related to how often the corresponding action was generated. It is not the case that more frequent verbs have more chances to be learned due to more frequent verb-action co-occurrences. In fact, some lower frequency verbs may have more chances to be learned as they co-occurred more frequently with the corresponding action. Further, it is not the case that more verb-action co-occurrences lead to more attention from the learner to the corresponding action.

The infant's attention adds a critical factor to variability of the learning input. We found that it is unlikely that young learners look at the target action during the entire time of a verb utterance. Instead, in most cases, they spent only some time looking at the corresponding action while hearing its label. It is also unlikely that they would completely miss the co-occurring action when a verb is heard. Although there isn't a significant correlation between verb frequency and infant attention, verb-action co-occurrence is positively correlated with the infant's full attention. For those verbs that co-occur more with the corresponding actions, infants are more likely to spend more time looking at the corresponding actions. The great variability within the concrete verbs examined here offers an explanation on why some concrete verbs are harder to learn than others.

What exactly makes verb learning difficult? Based on our findings, we argue that actions to which verbs refer are usually transient in context. Unlike concrete nouns whose perceptual information is usually available to the child when the object label is uttered, the corresponding action of a verb is not very likely to be perceptually available for the child continuously before, during, and after the verb utterance. Given the verb's transient nature and the infants' developing attentional system, if infants failed to attend to the right action at the time a verb was generated, they would miss the target action and once they miss the action, it is impossible for them to recover from other perceptual inspection of the immediate visual context at the moment. Despite the fact that verb learning is challenging, it is also important to keep in mind that verb learning happens in rich naturalistic contexts. Besides solely observing the action accompanying the verb, children also receive other cues that could help them figure out the correct mapping. For example, parents often provide socio-attentional cues, such as pointing to guide the child's attention (Goldin-Meadow, 2007). In addition, verbs are likely to co-occur with nouns and other parts of speech. Infants can also utilize the syntactic structure of the sentences to bootstrap the verb meaning (Naigles, 1996; Yuan, Fisher, & Snedeker, 2012).

The present study is the first step towards understanding the input for early verb learning. There are several future directions to advance our understanding on this topic. First, the current study does not have an outcome measure of verb learning as a way to directly assess the infant's knowledge of the heard verbs. Adding a verb learning test at the end of the play session would allow us to directly examine how the quantity and quality of co-occurrence statistics impact verb learning. Another way to link the input with learning outcomes is to collect and use the parent report of the child's vocabulary (i.e. MCDI, the MacArthur-Bates Communicative Development Inventories). Many studies have showed that both the quality and quantity of parent object naming are correlated with the child's MCDI results. However, little is known about how input quantity and quality impact early verb learning.

Second, toy play is only one of the everyday contexts in which children learn words. It would be interesting to study other learning contexts, such as storybook reading. Talking about objects on a page during book reading and manually manipulating objects during toy play are two very different types of interactions. Therefore, parent and children tend to generate very different learning statistics. Given that word-learning outcomes heavily depend on the structure of the input, it would be interesting to examine what types of input infants receive in those two contexts and compare how different types of input influence verb learning in those contexts.

Finally, another idea for follow-up studies is to compare the actions generated by infants versus by parents. There are studies showing that infants' own egocentric views contain unique properties and distributions that are critical for successful learning (Yurovsky, Smith, & Yu, 2013; Bambach, Crandall, Smith, & Yu, 2018). Actions generated by the parent may contain different visual properties from actions generated by the child. We could further investigate how the infant's body and associated visuomotor processes influence how the information is perceived and processed for learning verb-action mappings.

Acknowledgments

This research was supported in part by National Institutes of Health Grant R01HD074601 and R01HD093792 to CY and by the MOE (Ministry of Education in China) Project of Humanities and Social Sciences (Project No.16YJA190004), the National Natural Science Foundation of China (No. 71771102), and China Scholarship Council (CSC) (No. 201806775024) to SL. We would also like to thank Seth Foster, Anting Chen, Grace Lisandrelli, Lauren Slone, Drew Abney, and Daniel Pearcy, for data collection, and Emily Marie Heldman, Loren Louise Chastain, and Swasti Shree Singh, for data annotation.

References

Bambach, S., Crandall, D. J., Smith, L. B. & Yu, C. (2018). Toddler-Inspired Visual Object Learning. *Advances in Neural Information Processing Systems (NIPS), 31*.

Franchak, J. M. & Adolph, K. E. (2010). Visually guided navigation: Head-mounted eye-tracking of natural locomotion in children and adults. *Vision research, 50*(24), 2766-2774.

Gentner, D. (1982). Why nouns are learned before verbs: Linguistical relativity versus natural partitioning. In S. A. Kuczaj II (Ed.), *Language Development,* vol. 2: *Language, Thought, and Culture* (pp.301-334). Hillsdale, New Jersy: Lawrence Erlbaum.

Gentner, D. (2006). Why verbs are hard to learn. In K. Hirsh-Pasek & R. Golinkoff (Eds.), *Action Meets Word: How Children Learn Verbs,* (pp. 544–564). Oxford: Oxford University Press.

Gleitman, L. R. & Trueswell, J. C. (2018). Easy words: reference resolution in a malevolent referent world. *Topics in Cognitive Science*, 1-26.

Golinkoff, R.M., Chung, H. L., Hirsh-Pasek, K., Liu, J., Bertenthal, B., ……, Hennon, E. (2002). Young children can extend motion verb labels to point-light displays. *Developmental Psychology, 38*, 604–614.

Golinkoff, R. M., Hirsh-Pasek, K., Bailey, L., & Wenger, N. (1992). Young children and adults use lexical principles to learn new nouns. *Developmental Psychology, 28*, 99-108.

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children*. Paul H Brookes Publishing.

Hirsh-Pasek, K. & Golinkoff, R. M. (2006). *Action Meets Verb: How Children Learn Verbs*. New York: Oxford University Press.

Hoff, E. (2003). The specificity of environmental influence: Socioeconomic status affects early vocabulary development via maternal speech. *Child development*, *74*(5), 1368-1378.

Imai, M., Li, L., Haryu, E., Okada, H., Hirsh-Pasek, K., Golinkoff, R. M., & Shigematsu, J. (2008). Novel noun and verb learning in Chinese-, English-, and Japanese-speaking Children. *Child Development, 79*(4): 979-1000.

Maguire, M. J., Hirsh-Pasek, K., Golinkoff, R. M., & Brandone, A. C. (2008). Focusing on the relation: fewer exemplars facilitate children's initial verb learning and extension. *Developmental Science, 11* (4): 628–634.

Messenger, K., Yuan, S., & Fisher, C. (2015). Learning verb syntax via listening: New evidence from 22-month-olds. *Language Learning and Development, 11*(4): 356-368.

Monaghan, P., Mattock, K., Davies, R. A., & Smith, A. C. (2015). Gavagai is as gavagai does: Learning nouns and verbs from Cross-Situational statistics. *Cognitive science*, *39*(5), 1099-1112.

Naigles, L. R. (1996). The use of multiple frames in verb learning via syntactic bootstrapping. *Cognition*, *58*(2), 221-251.

Pereira, A. F., Smith, L. B., & Yu, C. (2014). A bottom-up view of toddler word learning. *Psychonomic bulletin & review*, *21*(1), 178-185.

Pulverman, R., Golinkoff, R. M., Hirsh-Pasek, K., & Buresh, J. S. (2008). Infants discriminate manners and paths in nonlinguistic dynamic events. *Cognition, 108*(3): 825-830.

Quine, W. V. O. (1960). Word and object (Studies in Communication). *New York and London: Tech-nology Press of MIT*.

Scott, R. M. & Fisher, C. (2012). 2.5-year-olds use cross-situational consistency to learn verbs under referential uncertainty. *Cognition, 122*: 163-180.

Smith, L. B., Jayaraman, S., Clerkin, E. & Yu, C. (2018). The Developing Infant Creates a Curriculum for Statistical Learning. *Trends in Cognitive Sciences, 4*, 325-336.

Weisleder, A., & Fernald, A. (2013). Talking to children matters: Early language experience strengthens processing and builds vocabulary. *Psychological science*, *24*(11), 2143-2152.

Yurovsky, D., Smith, L.B., & Yu, C. (2013). Statistical word learning at scale: the baby's view is better. *Developmental Science 16:6*, 959-966.

Yu, C. & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science, 18*(5), 414-420.

Yu, C. & Smith, L. B. (2017). Hand-eye coordination predicts joint attention. *Child Development, 88*(6): 2060-2078.

Yuan, S., Fisher, C., & Snedeker, J. (2012). Counting the nouns: Simple structural cues to verb meaning. *Child Development, 83*(4): 1382-1399.

Effects of affective ratings and individual differences in English morphological processing

Kaidi Lõo (kloo@ualberta.ca)
Department of Linguistics, University of Alberta
4-32 Assiniboia Hall, Edmonton, AB T6G 2E7, Canada

Abigail Toth [1,2] (a.g.toth@rug.nl)
[1]Department of Linguistics, University of Alberta
4-32 Assiniboia Hall, Edmonton, AB T6G 2E7, Canada
[2]Department of Artificial Intelligence, University of Groningen
Nijenborgh 9, 9747 AG Groningen, The Netherlands

Figen Karaca (karaca@ualberta.ca)
Department of Linguistics, University of Alberta
4-32 Assiniboia Hall, Edmonton, AB T6G 2E7, Canada

Juhani Järvikivi (jarvikiv@ualberta.ca)
Department of Linguistics, University of Alberta
4-55 Assiniboia Hall, Edmonton, AB T6G 2E7, Canada

Abstract

The nature of morphological processing has remained a controversial topic in psycholinguistic research. Some studies (e.g., Rastle, Davis, & New, 2004) have argued that when we read words like *corner* and *talker*, we automatically decompose them into existing morphemes like *talk*, *corn*, and *-er*, regardless of whether it is semantically plausible (e.g., *talker*) or not (e.g., *corner*). Recent studies, however, have challenged this view, by showing early semantic effects of the whole complex word (Järvikivi & Pyykkönen, 2011; Lõo & Järvikivi, 2019; Milin, Feldman, Ramscar, Hendrix, & Baayen, 2017). Using a masked priming paradigm, the present study only found effects of morphological decomposition for true morphological relations (e.g., *talker*) as well as effects of frequency and affective properties of whole words, further challenging automatic decomposition accounts. Finally, we also report that individual differences such as participants' self-reported scholarly reading and openness to new experience, affect processing.

Keywords: morphological processing; masked priming; affective properties; individual differences

Introduction

A large body of psycholinguistic research has focused on the question of how people read words like *cats* or *puppy*. More precisely, the question is whether these words are understood by accessing their morphemic components, for example *cat*, *-s*, *pup*, and *-y* or whether they are processed as any simple word, without recourse to internal structure.

From early on (Taft & Forster, 1975; Manelis & Tharp, 1977) both views have been represented in various forms. Recently, a particularly prominent view has been a variant of the former which states that all morphologically complex words are automatically decomposed in lexical access (Beyersmann et al., 2016; Lázaro, Illera, & Sainz, 2016; Longtin, Segui, & Hallé, 2003; Marslen-Wilson, Bozic, & Randall, 2008; Rastle, Davis, Marslen-Wilson, & Tyler, 2000; Rastle et al., 2004; Rastle & Davis, 2008). Most strikingly, this view takes the decomposition process to operate on the word form alone, without access to any semantic aspects of the word, with the prediction being that all word forms with apparent internal structure should be processed alike.

This approach has found support from masked priming studies (see e.g., Rastle et al., 2004) demonstrating that both pseudo-complex words, where the potential morphemic parts (e.g., *corn* and *-er*) do not make up the meaning of the whole word (e.g., *corner*), as well as transparent complex words with morphemic parts (e.g., *talk* and *-er*) that clearly contribute to the meaning of the whole word (e.g., *talker*), equally facilitate the recognition of their stems (*corn* and *talk*, respectively). Not only that, this research has also shown that words that are not exhaustively divisible into two morphemes, like *turnip* (where *-ip* is not an English affix), do not behave this way, suggesting that automatic decomposition is not only agnostic to semantics but is also driven by online analysis of linguistic structure.

However, not all recent research aligns with this view. Recent studies considering semantic and whole-word properties of the words have started to question this rather simplistic approach to language processing, especially in the case of morphologically complex languages, such as Serbian, Finnish and Estonian (Milin, Filipović Durdević, & Moscoso del Prado Martín, 2009; Moscoso del Prado Martín, Bertram, Häikiö, Schreuder, & Baayen, 2004; Lõo, Järvikivi, & Baayen, 2018; Lõo, Järvikivi, Tomaschek, Tucker, & Baayen, 2018), but also for English (Baayen, Wurm, & Aycock, 2007; Schmidtke, Matsuki, & Kuperman, 2017). For instance, these studies show whole-word frequency effects (Baayen et al., 2007; Schmidtke, Matsuki, & Kuperman, 2017; Lõo et al., 2018), as well as paradigmatic effects (Milin et al., 2009; Moscoso del Prado Martín et al., 2004; Lõo et

al., 2018) in the processing of complex words, which does not align well with the automatic decomposition approach.

In priming, Feldman and colleagues have shown for both English (Feldman, O'Connor, & Moscoso del Prado Martin, 2009) and Serbian (Feldman, Kostić, Gvozdenović, O'Connor, & del Prado Martín, 2012) that semantically transparent pairs show stronger priming than opaque pairs. Järvikivi and Pyykkönen (2011) reported that when morphological family size of the prime was accounted for, priming is smaller for pseudo-complex forms compared to real inflected forms in Finnish. Similarly, in a recent English masked priming study by Lõo and Järvikivi (2019) no priming was found for pseudo-complex words when whole-word frequency of the prime was taken into account in the analysis. Milin et al. (2017) included learning-based measures (Baayen, Milin, Filipovic Durdjevic, Hendrix, & Marelli, 2011) and found comparable priming effects for pseudo-derived words (e.g., *corner*) and orthographic controls (e.g., *brothel*) with more experienced readers showing priming to a lesser extent compared to less experienced readers. Along these same lines, Andrews and Lo (2013) reported that participants with relatively high vocabulary scores showed effects of priming in the transparent condition, but no priming in the opaque condition; whereas participants whose orthography knowledge was better than their vocabulary knowledge also showed priming in the opaque condition. Finally, Medeiros and Duñabeitia (2016) conducted a masked priming lexical decision study with Spanish suffixed words and found priming effects for slow readers, but not for fast readers.

In summary, there is accumulating evidence suggesting that both semantics of the complex words (Feldman et al., 2009; Järvikivi & Pyykkönen, 2011; Lõo & Järvikivi, 2019; Milin et al., 2017) and individual differences of the participants affect morphological decomposition (Schmidtke, Van Dyke, & Kuperman, 2017; Falkauskas & Kuperman, 2015; Medeiros & Duñabeitia, 2016; Andrews & Lo, 2013).

In the present study, we will focus on the affective properties (valence, arousal, danger and usefulness ratings) of complex words. Like simplex words, complex words can also be described along different affective dimensions, for example, from very negative (e.g., *murderer*) to very positive (e.g., *puppy*); from very exciting (e.g., *panics*) to very calming (e.g., *sleeping*); from extremely dangerous (e.g., *lionness*) to not dangerous at all (e.g., *echoing*); and from extremely useful to human survival (e.g., *knives*) to not useful at all (e.g., *scorpions*).

Previous research has shown that these affective properties predict lexical processing costs. For instance, positive, calming, useful and dangerous words have been found to elicit the fastest reaction times in word recognition tasks (Kuperman, Estes, Brysbaert, & Warriner, 2014; Wurm, 2007). Kuperman (2013) reported that compound words that had more positive consituents and were also more positive as a whole were processed faster than negative and neutral compounds.

Until now, affective properties of derived and inflected words have not received much investigation, especially, in the masked priming context (see Forster, 1998 for a discussion of this method). According to the automatic morphemic decomposition view, only affective properties of the stem (e.g, *pup*) and not of the whole (inflected or derived) word (e.g., *puppy*) should influence processing costs

The current study also investigates the effects of individual differences on morphological processing by looking at several self-reported language background and personality measures of participants. The personality component will be more exploratory than the language background measures. Lõo, Toth, Karaca, and Järvikivi (2018) found that personality influenced how participants rated different types of complex words. The arousal scale of the complex words was most prominent for participants who scored high on the neuroticism scale of Big Five personality questionnaire (John & Srivastava, 1999). The present study explores whether personality effects also arise in response times of masked priming lexical decision.

In summary, the goal of the current study is two-fold. First, we will study whether automatic decomposition occurs in a large within-item study design study design when lexcal-distributional and affetive properties of the words are included in the analysis. Second, we will examining individual differences on morphological processing, by exploring participants' self-reported language background and personality measures.

Visual Masked Priming Experiment

Participants

57 native speakers of English (43 female, mean age 21 years, range 18-46) with normal or corrected-to-normal vision participated in the experiment for partial course credit.

Materials

Ninety monomorphemic English words were selected as target stimuli from the Massive Auditory Lexical Database (MALD, Tucker et al., 2018). Each target word (e.g., *pup*) was primed within-item in six conditions. The conditions were the following: identity (e.g., *pup*), inflected (e.g., *pups*), derived (e.g., *puppy*), opaque (e.g., *pupal*), stem-embedded (e.g., *pupil*), and unrelated baseline control condition (e.g., *fencing*).

Additionally, 90 nonwords and 90 real words were added to the item set as fillers. Nonword targets (e.g., *sutt*) followed the phonotactics of English and were also selected from the MALD database. Primes for nonword and real word fillers were always real English words, consisting of the same six condition types with the same proportions as for the real word targets.

Design and procedure

The prime-target pairs were counterbalanced across six lists. Each list contained 360 items. 90 experimental prime-target trials, 90 unrelated prime-target filler trials, and 180 word-prime and nonword trials. In the filler trials, prime and target

pairs mimicked the six conditions in experimental list. Fillers and nonword trials were the same across lists.

The experiment was carried out using the E-Prime experimental software (Psychology Software Tools Inc.) and a SR-BOX response box. All stimuli were presented in black 32-point font Courier New letters on light gray background at the centre of the computer screen.

Each trial began with a fixation cross (+) appearing in the centre of the screen for 1000 ms, immediately followed by a forward mask (##########) for 500 ms. After that, the prime word appeared in lower case letters in the same location for 50ms. The target word appeared in the same location in upper case letters, and remained on the screen until the participant pressed the "yes" or "no" button on the response box. The participants were instructed to decide as accurately and as fast as possible whether the string of letters was an existing word in English or not. Ten practice trials preceded the experimental trials.

Prior to the main task, participants were asked to fill out a language background questionnaire, where they were asked to reflect on their English language skills and reading habits. For instance, they were asked how often they read scholarly or fictional literature; how they estimate their English vocabulary size, and how fast they consider themselves as readers.

They were also asked to fill out a 60-item HEXACO personality inventory questionnaire (Ashton & Lee, 2009), which provided for each participant a separate score on each of the six personality scales: honesty, emotionality, extroversion, agreeableness, conscientiousness, and openness to experience. The whole procedure (questionnaires and lexical decision task) took approximately 60 minutes to complete.

Analysis and Results

Prior to the analysis, practice trials, nonword trials and fillers were removed from the dataset. Trials with response times more than 1600 ms (1.1% of the data) as well as trials with incorrect responses (6.2% of the data) were removed.

Frequencies for the primes and targets were determined using the Corpus of Contemporary American English (COCA, Davies, 2010). Whole-word frequency (i.e., the token frequency of *pups*, *pups* or *puppy*) was used for the analysis. Frequency was log-transformed prior the analysis to reduce the skewness of the distribution.

Affective ratings of valence, arousal, danger and usefulness for each target and prime were collected during a separate rating experiment (see Lõo et al., 2018). In total, 181 native speakers of English rated the experimental items of the current study on an nine-point Likert-scale either on valence, arousal, usefulness or danger scale (1- sad/not exciting/not useful/not dangerous; 9 - happy/exciting/extremely useful/extremely dangerous). Participants in the rating experiment were different from the participants in the current experiment. A rating score for each target and prime word was calculated by taking the average score for each word across all participants.

The statistical analysis was conducted using Generalized Additive Mixed Models (GAMM, Wood, 2006; the R-package *mgcv*). For visualization, we made use of the R-package *itsadug* (van Rij, Baayen, Wieling, & van Rijn, 2016). We opted for GAMM analysis, because it does not assume linearity between the predictor and response variables.

The response variable of interest was the reaction time of masked priming lexical decision in milliseconds. We opted to use raw reaction times because they followed a normal distribution. However, an analysis with the log-transformed reaction times produced the same results. The main predictors were the condition (identity - M, inflected - I, derived - D, pseudo-complex - PC, stem-embedded - SE, baseline - BL), as well as the log-transformed frequency and affective ratings (valence, arousal, danger and usefulness) of the prime and target words. Additionally, we were interested in the effects of individual differences measures, we investigated whether self-reported language knowledge and reading habits, as well as personality had an effect on reaction times.

The output of the final GAMM-model is presented in Table 1. The parametric part shows that participants were significantly faster in identity (t=-4.90, $p<0.00001$), inflected (t=-4.10, $p<0.00001$) and derived (t=-3.43, p=0.006) conditions, whereas the two other conditions (pseudo-complex and stem-embedded condition) were not significant compared to the baseline condition.

Further, participants' openness to new experience and scholarly reading frequency affected reaction times. Reaction times decreased linearly for the participants who scored higher on the openness to experience scale compared to participants who scored lower on the same scale (t=-2.23, p=0.003). In return, reaction times were slower for participants who read more scholarly articles compared to participants who read fewer scholarly articles (t=3.48, p=0.015). There was neither a significant interaction between the condition and the openness score, nor between the condition and the scholarly reading score. Other self-reported language and personality scores were not significant in the final model.

The first three lines of the non-parametric part of the model output show nonlinear interactions between the prime and the target frequency, between the prime and target arousal score as well as between the prime and target usefulness score. These effects are visualized in Figure 1. The yellow color at the bottom left corner of the left panel shows that the reaction times were the slowest when both the prime and target were low-frequency words. However, the interaction between the target and prime frequency seems to disappear when target and prime frequencies increase. This is indicated by the blue color and wider contour lines at the top right corner of the left panel in Figure 1.

The nonlinear interaction between the prime and target arousal score is represented in the middle panel of Figure 1. Reaction times were the fastest when the target word scored high on the arousal scale and the prime word scored low on the arousal scale as indicated by the blue color at the bottom

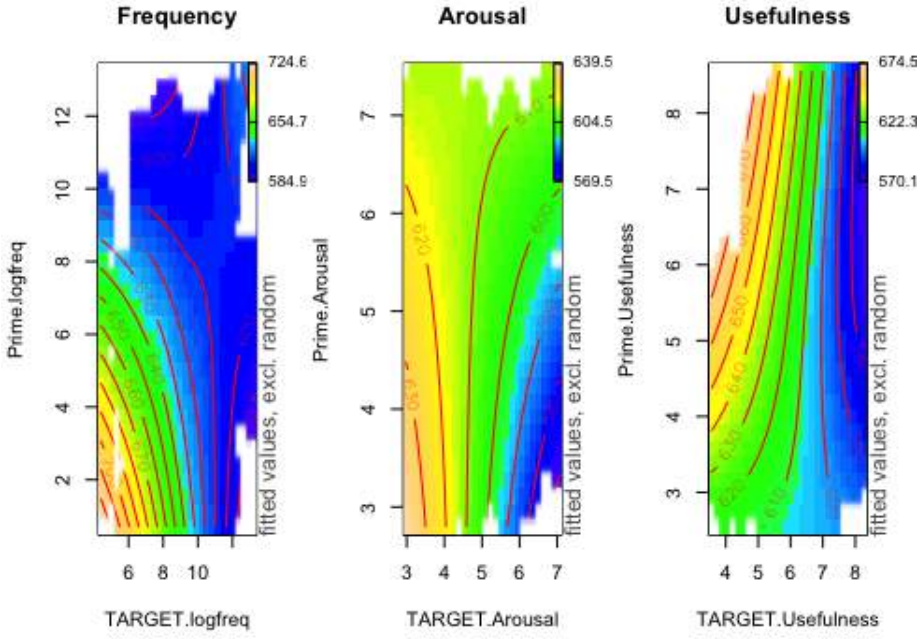


Figure 1: Tensor product smooth for the interaction of prime and target word frequency, arousal and usefulness. Color coding is used to represent model predictions, with yellow indicating slower reaction times, and blue representing faster reaction times

right corner of Figure 1.

Finally, the interaction between the prime and target word usefulness is presented in the right panel of Figure 1. The reaction times were the slowest when the prime was rated as very useful but the target was not, as indicated by the yellow color at the top left corner. Neither the target nor the prime valence and danger scores were significant. Frequency, arousal and usefulness scores did not interact with the condition.

The non-parametric part of the model output also included by-target random intercepts and by-participants random smooths for trial to account for the random variability between the items and participants.

In summary, the GAMM-analysis showed significant priming effects for the identity, derived and inflected conditions, but no priming for the pseudo-complex or stem-embedded conditions. There was a significant interaction between prime and target frequency, arousal and usefulness scores, but this did not interact with the condition. Finally, we found significant effects of participants' openness and scholarly reading, however, these effects did not interact with the condition.

Discussion and Conclusion

The goal of the current study was to investigate English morphological processing using masked priming. Some studies have reported that words like *talker* and *corner* are at least initially processed similarly (Rastle et al., 2004), while others claim that this is not the case, in particular, when various lexical-distributional properties are taken into account (Järvikivi & Pyykkönen, 2011; Lõo & Järvikivi, 2019; Milin et al., 2017), as well as individual differences between participants (Schmidtke, Van Dyke, & Kuperman, 2017; Falkauskas & Kuperman, 2015; Medeiros & Duñabeitia, 2016; Andrews & Lo, 2013).

In line with the latter view, the present study reports priming effects for words with an existing morphological relationship (e.g., *cats*, *puppy*), but no effects of priming for pseudo-complex words (e.g., *corner*). In fact, the processing of pseudo-complex words did not differ at all from either the stem-embedded condition (e.g., *turnip*) or the unrelated baseline condition. This supports findings from another recent English priming study by Lõo and Järvikivi (2019), where there were also no priming effects for pseudo-complex condition, using different materials. Additionally, we showed that the semantics of the complex words plays an important role early on. Like in Lõo and Järvikivi (2019), frequency of the complex word predicted processing costs; however, there were no significant differences between pseudo-complex and truly morphologically complex words in this respect.

Further, we investigated how affective ratings of complex words affect morphological processing. In line with the previous research on compound processing (Kuperman, 2013), we found effects of affective ratings for inflected and derived words. The prime-target ratio of affective ratings influenced the response times in masked priming, further challenging the blind decomposition approach, where the properties of the prime should not have an effect.

Interestingly, out of the four ratings scales (valence, arousal, usefulness and danger), only arousal and useful-

Table 1: Summary of the partial effects in GAMM fitted to masked priming lexical decision reaction times in milliseconds.

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
(Intercept)	694.18	86.89	7.99	< 0.0001
conditionD	-31.94	9.31	-3.43	0.0006
conditionI	-35.09	8.56	-4.10	< 0.0001
conditionM	-42.82	8.73	-4.90	< 0.0001
conditionPC	-2.17	8.19	-0.27	0.79
conditionSE	6.25	8.10	0.77	0.44
open.hexaco	-49.86	22.35	-2.23	0.03
Scholarly.Reading	32.81	13.90	2.36	0.02
B. smooth terms	edf	Ref.df	F-value	p-value
te(TARGET.logfreq,Prime.logfreq)	3.80	4.22	3.93	0.003
te(TARGET.Arousal,Prime.Arousal)	3.03	3.05	3.48	0.015
te(TARGET.Usefulness,Prime.Usefulness)	3.68	4.16	3.74	0.004
s(Subject,Trial)	184.49	494.00	4.29	< 0.0001
s(TARGET)	47.37	86.00	1.24	< 0.0001

ness target-prime ratios had an effect. Kuperman (2013) reported valence but not arousal effects in compound processing. However, their study used a standard lexical decision task, whereas the current study used masked priming lexical decision, tapping into earlier processing than the standard lexical decision. Arousal and usefulness ratings may be tapping into the internal state of the individual, thus are more subconscious; whereas, valence ratings may require a more conscious thought, and thus get activated later in time than can be captured by a masked priming study.

In general, the effects of affective properties were not that strong in the current study, and there may be different reasons for this. First, in the current study, the derived and inflected primes (e.g., *puppy*, *pups*) have similar meanings to the target (e.g., *pup*), so the affective polarities may have been very similar (for example, the word *puppy* was as happy, exciting, useful and dangerous as the word *pup*). Second, as the design of the current study did not explicitly control for the emotional affectivess of the stimuli, most of the stimuli were neither very positive nor very negative, neither very useful nor very useless, so there may not have been enough variation between the stimuli.

Finally, the current study also focused on the effects of individual differences in morphological processing. Interestingly, they were again the same for truly complex and pseudo-complex words. From the five personality measures (honesty, emotionality, extroversion, conscientiousness, openness to experience), only participants' openness to new experience had an effect on reaction times. More open participants were faster than less open participants. Participants who are more open to experience in general might be also more open to tasks such as a lexical decision task. From the language background measures (self-reported vocabulary knowledge, reading speed, scholarly reading and fictional reading frequency), only scholarly reading had an effect on reaction times. Participants who read more scholarly literature were slower than participants who read less scholarly literature. This is in line with the research showing that more experience with language slows one down in various language tasks (Ramscar, Hendrix, Shaoul, Milin, & Baayen, 2014).

Also this is important to note, however, that both the topic of affective properties and individual differences in morphological processing are relatively new and thus, our findings require further research.

To conclude, the processing of complex words, even in languages with a relative simple morphology, such as English, seems to be much more complex than just a matter of morphemic decomposition. The current study complements this idea by showing that pseudo-complex and morphologically complex words are indeed processed differently. We also showed that both affective properties and individual differences influence English morphological processing; however, the precise nature of these effects requires further research.

References

Andrews, S., & Lo, S. (2013). Is morphological priming stronger for transparent than opaque words? it depends on individual differences in spelling and vocabulary. *Journal of Memory and Language*, *68*(3), 279–296.

Ashton, M. C., & Lee, K. (2009). The HEXACO–60: A short measure of the major dimensions of personality. *Journal of personality assessment*, *91*(4), 340–345.

Baayen, R. H., Milin, P., Filipovic Durdjevic, D., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review*, *118*(3), 438-481.

Baayen, R. H., Wurm, L. H., & Aycock, J. (2007). Lexical dynamics for low-frequency complex words. a regression study across tasks and modalities. *The Mental Lexicon*, *2*, 419–463.

Beyersmann, E., Ziegler, J. C., Castles, A., Coltheart, M., Kezilas, Y., & Grainger, J. (2016). Morpho-orthographic segmentation without semantics. *Psychonomic Bulletin & Review*, *23*(2), 533–539.

Davies, M. (2010). The corpus of contemporary american english as the first reliable monitor corpus of english. *Literary and linguistic computing*, *25*(4), 447–464.

Falkauskas, K., & Kuperman, V. (2015). When experience meets language statistics: Individual variability in processing English compound words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(6), 1607.

Feldman, L. B., Kostić, A., Gvozdenović, V., O'Connor, P. A., & del Prado Martín, F. M. (2012). Semantic similarity influences early morphological priming in serbian: A challenge to form-then-meaning accounts of word recognition. *Psychonomic bulletin & review*, *19*(4), 668–676.

Feldman, L. B., O'Connor, P. A., & Moscoso del Prado Martin, F. (2009). Early morphological processing is morpho-semantic and not simply morpho-orthographic: evidence from the masked priming paradigm. *Psychonomic Bulletin & Review*, *16*(4), 684–691.

Forster, K. (1998). The pros and cons of masked priming. *Journal of Psycholinguistic Research*, *27*(2), 203–233.

Järvikivi, J., & Pyykkönen, P. (2011). Sub-and supralexical information in early phases of lexical access. *Frontiers in Psychology*, *2*.

John, O. P., & Srivastava, S. (1999). The big five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality: Theory and research*, *2*(1999), 102–138.

Kuperman, V. (2013). Accentuate the positive: Semantic access in english compounds. *Frontiers in psychology*, *4*.

Kuperman, V., Estes, Z., Brysbaert, M., & Warriner, A. B. (2014). Emotion and language: valence and arousal affect word recognition. *Journal of Experimental Psychology: General*, *143*(3), 1065.

Lázaro, M., Illera, V., & Sainz, J. (2016). The suffix priming effect: Further evidence for an early morpho-orthographic segmentation process independent of its semantic content. *The Quarterly Journal of Experimental Psychology*, *69*(1), 197–208.

Longtin, C., Segui, J., & Hallé, P. (2003). Morphological priming without morphological relationship. *Language and Cognitive Processes*, *in press*, 0.

Lõo, K., & Järvikivi, J. (2019). Whole-word frequency effects in English masked priming: very little CORN in CORNER and CORNET. *Manuscript submitted for publication*.

Lõo, K., Järvikivi, J., & Baayen, R. H. (2018). Whole-word frequency and inflectional paradigm size facilitate Estonian case-inflected noun processing. *Cognition*, *175*, 20–25.

Lõo, K., Järvikivi, J., Tomaschek, F., Tucker, B. V., & Baayen, R. H. (2018). Production of estonian case-inflected nouns shows whole-word frequency and paradigmatic effects. *Morphology*, *28*, 71–97.

Lõo, K., Toth, A., Karaca, F., & Järvikivi, J. (2018). Effects of word stem and personality in affective ratings for complex words. *Paper presented at the 11th International Conference on the Mental Lexicon in Edmonton, Alberta, September 25-28, 2018*.

Manelis, L., & Tharp, D. A. (1977). The processing of affixed words. *Memory and Cognition*, *5*, 690-695.

Marslen-Wilson, W. D., Bozic, M., & Randall, B. (2008). Early decomposition in visual word recognition: Dissociating morphology, form, and meaning. *Language and Cognitive Processes*, *23*(3), 394–421.

Medeiros, J., & Duñabeitia, J. A. (2016). Not everybody sees the ness in the darkness: Individual differences in masked suffix priming. *Frontiers in psychology*, *7*, 1585.

Milin, P., Feldman, L. B., Ramscar, M., Hendrix, P., & Baayen, R. H. (2017). Discrimination in lexical decision. *PloS one*, *12*(2), e0171935.

Milin, P., Filipović Durdević, D., & Moscoso del Prado Martín, F. (2009). The simultaneous effects of inflectional paradigms and classes on lexical recognition: Evidence from Serbian. *Journal of Memory and Language*, 50–64.

Moscoso del Prado Martín, F., Bertram, R., Häikiö, T., Schreuder, R., & Baayen, R. H. (2004). Morphological family size in a morphologically rich language: The case of Finnish compared to Dutch and Hebrew. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *30*, 1271–1278.

Ramscar, M., Hendrix, P., Shaoul, C., Milin, P., & Baayen, R. H. (2014). The myth of cognitive decline: Non-linear dynamics of lifelong learning. *Topics in cognitive science*, *6*(1), 5–42.

Rastle, K., & Davis, M. H. (2008). Morphological decomposition based on the analysis of orthography. *Language and Cognitive Processes*, *23*(7-8), 942–971.

Rastle, K., Davis, M. H., Marslen-Wilson, W. D., & Tyler, L. K. (2000). Morphological and semantic effects in visual word recognition: A time-course study. *Language and Cognitive Processes*, *15*(4-5), 507–537.

Rastle, K., Davis, M. H., & New, B. (2004). The broth in my brother's brothel: Morpho-orthographic segmentation in visual word recognition. *Psychonomic Bulletin & Review*, *11*, 1090–1098.

Schmidtke, D., Matsuki, K., & Kuperman, V. (2017). Surviving blind decomposition: A distributional analysis of the time-course of complex word recognition. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *43*(11), 1793-1820.

Schmidtke, D., Van Dyke, J. A., & Kuperman, V. (2017). Individual variability in the semantic processing of English compound words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, in press.

Taft, M., & Forster, K. I. (1975). Lexical storage and retrieval of prefixed words. *Journal of Verbal Learning and Verbal*

Behavior, *14*, 638-647.

Tucker, B. V., Brenner, D., Danielson, D. K., Kelley, M. C., Nenadić, F., & Sims, M. (2018). The massive auditory lexical decision (mald) database. *Behavior research methods*, 1–18.

van Rij, J., Baayen, R. H., Wieling, M., & van Rijn, H. (2016). *itsadug: Interpreting time series, autocorrelated data using GAMMs.* (R package version 2.2)

Wood, S. N. (2006). *Generalized additive models*. New York: Chapman & Hall/CRC.

Wurm, L. H. (2007). Danger and usefulness: An alternative framework for understanding rapid evaluation effects in perception? *Psychonomic Bulletin & Review*, *14*, 1218–1225.

Is it easier to segment words from infant- than adult-directed speech? Modeling evidence from an ecological French corpus

Georgia Loukatou (georgialoukatou@gmail.com)
Laboratoire de sciences cognitives et psycholinguistique, Département d'études cognitives, ENS, EHESS, CNRS, PSL University
Paris, France

Marie-Thérèse Le Normand (marielenormand@mac.com)
INSERM & LPP (Laboratoire Psychopathologie et Processus de Santé), Université Paris Descartes, Sorbonne
Paris, France

Alejandrina Cristia (alecristia@gmail.com)
Laboratoire de sciences cognitives et psycholinguistique, Département d'études cognitives, ENS, EHESS, CNRS, PSL University
Paris, France

Abstract

Infants learn language by exposure to streams of speech produced by their caregivers. Early on, they manage to segment word forms out of this continuous input, which is either directly addressed to them, or directed to other adults, thus overheard. It has been suggested that infant-directed speech is simplified and could facilitate language learning. This study aimed to investigate whether features such as utterance length, segmentation entropy and lexical diversity could account for an advantage in segmentability of infant-directed speech. A large set of word segmentation algorithms was used on an ecologically valid corpus, consisting of 18 sets of recordings gathered from French-learning infants aged 3-48 months. A series of textual analyses confirmed several simplicity features of infant-, compared to adult-directed speech. A small segmentation advantage was also documented, which could not be attributed to any of those corpus features. Some particularities of the data invite further research on more corpora.

Keywords: language acquisition; infant-directed speech; computational modeling; word segmentation; unsupervised learning

Introduction

Infants acquire language early on, building a vocabulary of several hundred word forms by 11 months of life (Ngon et al., 2013). Since most word forms do not appear in isolation (Brent & Siskind, 2001), much previous work studies how infants segment (i.e., pull out) forms from their caregivers' running input. A close look at this input shows that it is not homogeneous, but instead contains some speech addressed to the infants themselves (infant-directed speech or IDS) and some speech overheard by infants which is addressed to others, including adults (adult-directed speech or ADS). These two speech registers differ along many dimensions, including some that may impact word segmentation.

Broadly, IDS has been claimed to present properties that would facilitate language acquisition, with IDS being phonologically, syntactically, and semantically simplified (Soderstrom, 2007). Other characteristics are more relevant to word segmentation. First, IDS may have a higher proportion of single-word phrases (Brent & Siskind, 2001), and phrases might be shorter in length (Newport, Gleitman, & Gleitman, 1977) than in ADS. In shorter phrases, more words would occur at phrase edges, which should improve segmentation: Phrase edges, easily perceptible, are word boundaries provided "for free". Indeed, infants may be more successful at recognizing and segmenting phrase-final words (E. Johnson, Seidl, & Tyler, 2014). Additionally, shorter phrases entail that the set of possible segmentations for each phrase is smaller, lowering segmentation ambiguity. For instance, Fourtassi, Börschinger, Johnson, and Dupoux (2013) showed that ADS might be more ambiguous to segment, when comparing an ADS to an IDS corpus. Second, words may be shorter (Ma, Golinkoff, Houston, & Hirsh-Pasek, 2011), which should mean that word, morphemes, and syllable boundaries coincide more often and there are fewer places to posit or miss positing a boundary. Third, there may be more repetitions, therefore fewer hapaxes (words uttered only once), and overall less lexical diversity (Soderstrom, 2007). Low lexical diversity means fewer target words need to be found. There might be more cues to help segment out frequently repeated words, than words that appear rarely or once. Indeed, one computational modeling study found that artificially reducing phrase length and increasing word repetition in a corpus improved word segmentation with one word segmentation model (Batchelder, 1997). Based on these hypotheses and previous work, we predict that the task of recovering wordforms is easier in IDS than ADS.

Naturally, IDS features may not be the same across infant ages. IDS addressed to very young infants may differ from that addressed to older infants, possibly resembling ADS more as infants get older. For example, IDS features may become less accentuated as the infant grows up; repetitions might decrease, utterance length and lexical diversity increase with age (Henning, Striano, & Lieven, 2005; Soderstrom, 2007). According to the hypotheses explained above, IDS addressed to younger infants should be "easier" to segment than IDS to older infants.

In this paper, we aim to address the question of whether it is easier to segment wordforms from IDS than ADS, using multiple word segmentation models, and taking into account changes with infants' age. In the next section, we review

previous modeling work more thoroughly, before introducing our own approach.

Previous studies

Some studies tested whether infants learn more from IDS than ADS in an experimental situation. However, improvements for IDS compared to ADS could be due to the fact that infants pay more attention when they listen to IDS, and thus learn more from it. This method cannot reveal whether, above and beyond this attentional effect, there are intrinsic *informational* differences that affect segmentability. Fortunately, there is a complementary method to approach this question with a colder eye, which builds on computational models of word segmentation. The input to such word segmentation models is usually speech transcriptions, in order to control for differences such as attention capture and acoustic implementation. Segmentation models used for this method are based on findings by experimental studies that infants might make use of statistical cues. Computational models of infant word segmentation can be grouped into two conceptual classes: lexical and sublexical. Sublexical models segment based on local cues, such as transitional probabilities and phonotactics. Lexical models build a lexicon based on recurrent chunks of speech identified with Bayesian probabilities or by memorizing isolated words.

Little previous modeling work has specifically compared IDS and ADS. Four representative studies are summarized in Table 1. For these four studies, improved segmentation performance was found for IDS than ADS: 15% for Batchelder (2002), 5-8% for Fourtassi et al. (2013), 2-10% for Ludusan, Mazuka, Bernard, Cristia, and Dupoux (2017) and 3-10% for Daland and Pierrehumbert (2011). A recent paper critiqued this previous work as follows (Cristia, Dupoux, Ratner, & Soderstrom, 2018). IDS mainly involved caregivers addressing their infants during predefined tasks (e.g., a play session in the laboratory) or in short visits to the child's home. In the former case, by constraining the context, the structure and lexicon of caregivers might have been limited and adapted to that task. And in both cases, being observed could affect caregivers' behavior, who might produce less spontaneous and more formal speech. Moreover, ADS was mostly addressed to an unfamiliar person (experimenter). These conversations are likely more formal than ADS between caregivers in daily life, and could increase the complexity of the speech. As shown by E. Johnson, Lahey, Ernestus, and Cutler (2013), IDS differs more from ADS to unfamiliar adults, than ADS to familiar adults. This could result in increased qualitative differences between registers and probably overestimated differences in segmentability.

Indeed, Cristia et al. (2018) recently documented a considerably smaller IDS advantage when modeling segmentation on an ecological English IDS and ADS corpus. The corpus consisted of transcriptions from excerpts of day-long recordings; thus infants' linguistic environment was recorded while they were going on with their daily lives, resulting in realistic IDS and ADS. Across a wide range of lexical and sublexical models, the IDS advantage ranged from -2% to 8%, with only 3 models providing evidence of an advantage greater than a measure of error. Interestingly, the difference between registers was further reduced when IDS was matched to ADS in corpus length.

The present study

We contribute to this literature in three main ways. First, we specifically describe IDS-ADS differences using various corpus description tools. We compare the registers in: phrase length, word length, ratio of single word phrases, intrinsic segmentation ambiguity (using segmentation entropy), lexical diversity (using Moving Average Type-Token Ratio – MATTR–, so as to control for corpus size), and ratio of hapaxes. Some, but not all of these features have been separately looked at in previous studies (i.e. Fourtassi et al., 2013 measured segmentation ambiguity and Batchelder, 1997 measured word and phrase length, repetitiveness). This is the first study to systematically investigate a plurality of language features on the same IDS-ADS corpus. We test whether IDS is simpler than ADS, as far as these features are concerned. Moreover, following Batchelder (2002), we further investigate whether variation in these features can actually account for the segmentability of a register.

Second, IDS corpora coming from a wide infant age range have been used by previous research, but IDS addressed to infants of different ages were, most of the times, merged together. One exception is Batchelder (1997), who documented that IDS to younger children (13-18 months) produced more successful results than IDS to older children (22-25 months), whereas ADS results from mothers of younger versus older infants didn't differ. In this paper, we specifically ask whether some IDS features interact with infant age and whether segmentability of IDS might actually be affected by age. For that, we include IDS and ADS from a wide age range, and further investigate possible correlations between features, segmentation scores, and infant age.

Third, we follow Cristia et al. (2018) by analyzing a completely ecological child-centered corpus, based on excerpts of day-long recordings, and which thus contains natural ADS and IDS as the child hears over the course of the day. The results of our study would provide more evidence to the question whether differences in home-recorded IDS and ADS are smaller than those between less controlled IDS-ADS contrasts (see Table 1).

In addition to these three main contributions, we extend the range of languages studied to European French.

Methods

We segmented IDS and ADS of each infant separately. Scripts used for corpus preprocessing, phonologization, and segmentation as well as results and supplementary material are available at `https://osf.io/6vwse/?view_only=0bc4f6c0e23040cbbb92e26d414d4a7a`. Statistical analyses were carried out in R (R Core Team, 2013).

Table 1: Summary of design in previous modeling studies comparing IDS and ADS segmentation. In Language(s), Eng stands for English, Jap for Japanese, Span for Spanish. Under IDS and ADS, we describe the corpora. The specific corpora used were: R= RIKEN; H= Hamasaki; C= Spontaneous Japanese; BR= Bernstein Ratner; B= Buckeye; D= Deuchar & Clark 1992, Marrero; M= Miyata 1995; novel= Moon and the Sixpense; short stories were written by Alejandro Dolina (MacWhinney, 1996). Under model, we note the type of model used: lex for lexical and sublex for sublexical.

Study	Language(s)	Infant age(s)	IDS	ADS	model
Batchelder (2002)	Eng.	1;1-1;9	play session (BR)	novel	1 lex
Batchelder (2002)	Span.	1;8-8;0	CHILDES (D)	short story	1 lex
Batchelder (2002)	Jap.	1;3-3;1	home play session (M)	science book	1 lex
Daland et al. (2011)	Eng.	various	all CHILDES	interview (B)	1 sublex
Fourtassi et al. (2013)	Eng.	1;1-1;9	play session (BR)	interview (B)	1 lex
Fourtassi et al. (2013)	Jap.	2;2-3;7	play session (H)	lecture (C)	1 lex
Ludusan et al. (2017)	Jap.	1;6-2;0	play session (R)	lecture (C)	1 lex, 3 sublex

Corpus

Sixteen typically developing native French-speaking infants (eight girls, eight boys; ages 3-48 months, M=20, SD=13), whose families were highly educated, were included. Two of the infants were recorded at two different ages. Each child was recorded 10-16 hours per day, three days a week, in their natural environments. The original recordings are available online (Canault, Le Normand, Foudil, Loundon, & Thai-Van, 2016a, 2016b; VanDam et al., 2016). Next, 18 10-min samples, totaling 3 hours per child (1 hour per day), were selected for orthographic transcription by two native French speakers, as detailed in Canault et al. (2016b). The main criteria for selection reported was that a number of activities were sampled, and that there be a high number of productions by the child and the adult. For the present project, the transcriptions of the first day for all infants were corrected by a native French speaker, who made sure that the definition of utterance was stable (and corrected any other errors, such as misattributions or orthographic errors). The coder annotated whether an adult caregiver's utterance was directed to the target child, an adult, or other, using content and context. Utterances addressed to the target child constituted the IDS corpus and those directed to an adult were the ADS corpus.

Pre-processing

Pre-processing was carried out using custom scripts written mainly in bash and in python, available from `https://github.com/georgialoukatou/French_ADS_IDS_segmentation_Lyon`. All extraneous codes (such as punctuation marks or "xxx", the code indicating that what was said could not be understood by the transcriber) were removed, leaving only the orthographic representation of the adults' speech. The corpora were phonologized with the French voice of the espeak TTS system (Duddington, 2012), using the phonemizer wrapper (Bernard, 2018), which further syllabifies according to the Maximum Onset Principle.

Before segmentation, all spaces between words were removed, leaving the input parsed into minimal units. The minimal units were either phones or syllables. Both phonemes and syllables were tested with all models. Utterance boundaries were preserved as such, since they are supposedly salient to infants (Shukla, White, & Aslin, 2011). This constitutes the input to the model. After preprocessing, the 18 infant-directed corpora contained M=487 (SD 350) utterances (range 84 to 1,172 utterances). The 18 adult-directed corpora contained M=238 (SD 230) utterances (range 15 to 780 utterances).

For comparability with previous work, we evaluate the models' performance using lexical token F-scores, measured by comparing the original version of the input (with spaces between words) against the one returned by the model (with spaces in the hypothesized breaks).

Segmentation

Both corpus description and segmentation were carried out using the WordSeg package (Bernard et al., 2018), available from `https://github.com/bootphon/wordseg/`. Due to space limits, the algorithms are only briefly described here. Full technical details can be found in `https://wordseg.readthedocs.io/`. All algorithms are unsupervised, and inspired in infant experimental work.

We used two representatives of the sublexical word segmentation class contains, called DIBS and TP for short. The Diphone Based Segmentation algorithm (DiBS; Daland & Pierrehumbert, 2011) is based on the idea than a phoneme sequence often spanning phrase boundaries would probably span word breaks.

The Transitional Probabilities algorithm family (TP; Saksida, Langus, & Nespor, 2017) is based on the concept that syllable pairs with lower statistical coherence tend to span word breaks. Forward TP (FTP) measures the frequency of occurrence of the syllabic sequence AB given the frequency of occurrence of the syllable A. Backward TP (BTP) measures the frequency of occurrence of the syllabic sequence AB given the frequency of occurrence of the syllable B. The Relative versions (FTPr or BTPr) threshold TPs against that of neighboring sequences. The Absolute versions

Table 2: Paired t-tests measuring feature differences across IDS and ADS. Word length is measured in phonemes. % 1-w phrase stands for ratio of single word phrases. % hapaxes stands for percent of hapaxes. IDS gives the mean values of each feature on the IDS corpus, with standard deviation in parentheses. ADS shows the mean values of each feature on the ADS corpus with standard deviation in parentheses. The window size for MATTR is 10 words. "p" gives the p-value of the t-test.

Feature	IDS	ADS	p
Word length	2.86 (.08)	2.80 (.11)	.071
Phrase length	5.89 (.85)	6.73 (.86)	*
% 1-w phrase	.18 (.06)	.13 (.05)	**
Entropy	.02 (.004)	.03 (.01)	.31
MATTR	.89 (.03)	.93 (.02)	***
% hapaxes	.39 (.22)	.48 (.27)	***

(FTPa or BTPa) instead threshold on the average of all TPs over the sum of different syllable bigrams.

We used two representatives of the lexical class as well: AG and PUDDLE. Adaptor Grammar (AG) uses the Pitman-Yor process, a stochastic process of probability distribution which prefers the reuse of frequently occurring rules versus creating new ones to build a lexicon, then uses that lexicon to parse the input (M. Johnson, Griffiths, & Goldwater, 2007).

Phonotactics from Utterances Determine Distributional Lexical Elements (PUDDLE, Monaghan & Christiansen, 2010) treats each utterance as a lexical item, unless an already stored item is part of this utterance, and the remainders are phonotactically legal. If so, it breaks up the utterance into segments, and the segments would enter the lexicon as new lexical items.

Finally, two baselines were included: Syll=Word treats each syllable as a word and Utt=Word treats each utterance as a word.

Results

We first investigated whether IDS is simpler than ADS in terms of six corpus features that could affect word segmentation, as described in the reasoning above. The results of paired t-tests comparing the registers for each feature are in Table 2, which shows that four out of six features fit our predictions.

We also noticed that IDS size corpus (M=487, SD=350 per child) was significantly larger than the ADS one (M=238, SD=230), based on a t-test with t(17)=2.63, p=0.02. This may mean that these infants were exposed to more IDS than ADS, similar to what Cristia et al. (2018) found for English.

The performance of all segmentation algorithms for both registers is captured in Figure 1. IDS is easier to segment than ADS when points are above the dotted diagonal line. There was a small IDS advantage for most algorithms, although some showed the opposite effect (DiBSs,

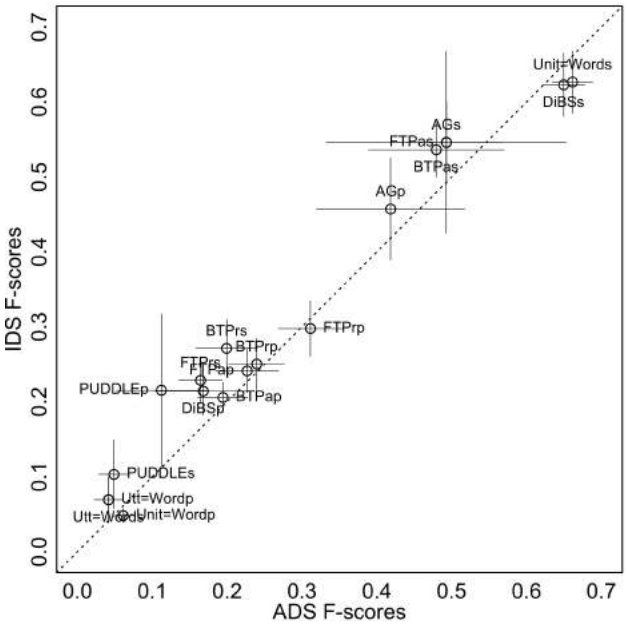


Figure 1: Token F-scores obtained by each algorithm for IDS as function of that for ADS. The final "s" in the model's name means that the basic unit of the corpus was syllables (PUDDLWs, Utt=Words, Unit=Words, DiBSs, FTPas, FTPrs, BTPas, BTPrs, AGs). The final "p" in the model's name means that the basic unit of the corpus was phones (PUDDLWp, Utt=Wordp, Unit=Wordp, DiBSp, FTPap, FTPrp, BTPap, BTPrp, AGp). Error bars show two standard deviations over the 18 corpora.

Unit=Words, Unit=Wordp, FTPrp). We also observe that in many cases the pseudo-confidence intervals cross the diagonal line, suggesting that performance difference is within the range of error. Thus, only FTPrs, BTPrs, Utt=Wordp, PUDDLEp and PUDDLEs showed a clear advantage of IDS. We then tested for overall effects in a linear mixed effect regression model (Bates, Mächler, Bolker, & Walker, 2015) predicting token F-scores from register (IDS or ADS) as a fixed effect, where subject and algorithm (AGs, AGp, DiBSs, DiBSp...) were random effect variables. Register significantly affected token F-scores ($\chi^2(1)=50.87$, $p<.05$, Type II Anova), IDS having a performance advantage of $.03 \pm .004$ (standard error).

Next, we tested whether this performance advantage was due to one of the above-mentioned corpus properties. To see whether performance differences were due to the artifactual difference in corpus length, we also included the number of utterances as a register feature. Thus, 7 new models, each including one of the features as an additional fixed effect, were fit. We then measured the significance of register and features in the new models with a Type II Anova test (Fox & Weisberg, 2011).

If the advantage of IDS was entirely due to one feature, then register would no longer be significant in these addi-

Table 3: Corpus features predict segmentation scores, but do not replace register. β feat stands for the estimated coefficient of that feature; β rgstr for that of register in the new model (which should be compared to 0.03 at the simple model). p features shows whether feature was significant in new model. p rgst shows whether register remained significant in the new model. N. utts stands for number of utterances.

Feature	Feature		Register	
Feature	β	p	β	p
Word length	.02	.48	.03	***
Phrase length	.01	***	.04	***
% 1-w phrase	.06	.29	.03	***
Entropy	-1.58	***	.03	***
MATTR	.5	***	.05	***
% hapaxes	.03	.18	.03	***
N. utts	.00005	***	.02	***

Table 4: Correlation tests (Spearman) of corpus features and infant age for each register. "coef." stands for correlation coefficient.% 1-w phrase stands for ratio of single word phrases. % hapaxes is the ratio of hapaxes.

Feature	IDS coef.	ADS coef.
Word length	.50*	.06
Phrase length	.34	-.56*
% 1-w phrase	-.37	.12
Entropy	-.50*	.70**
TTR	.44	-.37
% hapaxes	.01	.30

tional analyses. Results (in Table 3) showed that phrase length, segmentation entropy, MATTR, and corpus size accounted for variance in the results, but no single feature rendered register effects non-significant.

Next, we investigated whether IDS features change with infant age, with IDS becoming more ADS like as infants age. Spearman correlation tests between properties and infant age for each register separately (Table 4) did not confirm our predictions: Only word length and entropy (neither of which had emerged as register properties on Table 2) correlated with age in IDS; entropy and phrase length did so for ADS. We have no plausible explanation for these effects.

Two infants were recorded twice at different ages, one at 31 and 38 months, the other at 32 and 40 months. Following a recommendation from a reviewer, we inspected these two infants as case studies. An inspection of IDS features demonstrated that phrase length and % of 1-w phrases were the only features having small changes with age, but only the latter would change in the same direction for both infants, increasing by 6% and 1% from the first to the second recording. A few ADS features also changed slightly with age, such as % of 1-w phrases, word length and entropy, but only phrase length changed in the same direction for both infants, decreasing by 1.18 and 1.66 phonemes.

Finally, we created a new model predicting token F-scores register (IDS or ADS) and infant age in months as fixed effects (and model and participant as random effects, as before), and their interaction. Both main effects and the interaction were significant (Age $\chi^2(1)=4.31$, $p<.05$; Register $\chi^2(1)=53.14$, $p<.5$; Age:register $\chi^2(1)=28.81$, $p<.05$). A follow-up analysis separating the registers indicated that ADS scores decreased by $.002 \pm .0005$ (standard error) with age, whereas there was no significant change with age for IDS.

Discussion

In this modeling study, we assessed whether there are informational differences affecting word segmentation between IDS and ADS drawn from the same ecological corpus. First, we investigated whether this naturalistic corpus had IDS-ADS differences in textual features that would make segmentation easier in the former than the latter. We found most features fit our predictions: Phrases were longer, there were more single-word phrases, lexical diversity was lower, and there were fewer hapaxes in IDS than ADS. No significant effect was found for word length and ambiguity. This result contributes to the growing literature documenting IDS features, with the important advantage that current work draws from fully ecological IDS and ADS.

Next, we investigated the segmentability of the corpora using a large set of both lexical and sublexical segmentation models. Although scores varied a great deal across algorithms and some algorithms showed the opposite effect, IDS was overall slightly easier to segment than ADS. The mean difference across registers (CDS minus ADS, in each algorithm separately) was 3%, ranging from –4% to 10%. This effect is smaller than that found in most previous studies, but similar to the one reported by Cristia et al. (2018), who were also drawing from a naturalistic IDS-ADS corpus. This is evidence that previously documented IDS-ADS segmentability differences (as in Table 1) are not representative of what infants actually hear. It is important to note that corpus length across registers was not matched in the present study for practical reasons, but, based on findings by Cristia et al. (2018), we suspect that controling for corpus size would have reduced the IDS advantage even further.

Next, we asked whether some of the above-mentioned textual features uniquely explained segmentability differences across registers. Phrase length, segmentation entropy, and repetitiveness explained significant variance in segmentation scores, above and beyond the effects of register. However, none of the features uniquely explained away the effect of the register, which remained significant in all cases. This means that register effects on segmentability cannot be reduced to any one of these features. Since we only had 18 children's data, we could not fit a model with all 6 features at once for fear of overfitting, but future work with higher power may be able to assess whether these features jointly explain away reg-

ister, or whether there are other textual features that we have not yet considered.

Furthermore, Canault et al. (2016b)'s corpus allowed us to address a question that has been seldom asked, namely IDS-ADS differences across infant ages. Results of correlations between textual features and age, and a regression model on token F-scores did not support our prediction that IDS would become more like ADS as children aged, and thus the IDS-ADS segmentability gap would close. On the contrary, we found that ADS scores dropped with child age. Although further work is needed, we believe this mainly reflects the lower availability of ADS in children's environment as they age. Indeed, replicating a pattern that had been documented in North American English children (Bergelson et al., 2019), we found the number of ADS utterances dropped for older, compared to younger, children.

Before closing, we would like to acknowledge some limitations of this work. Corpus size was overall small (which may lead to inconsistencies in results; Bernard et al., 2018) and, due to the work involved in collecting daylong recordings and annotating fully spontaneous speech, infant sample size was 18 infants. Moreover, data scarcity was correlated with registers and ages: While only 3 of the 18 IDS corpora contained fewer than 100 utterances, 7 did for ADS, and 4 of those belonged to infants older than 31 months. A decrease of ADS quantities with infant age in such day-long recordings has been documented in previous work on North American English (Bergelson et al., 2019), so it may not be an artifact of the current sample selection. Nonetheless, this trend may entail that if we want to control corpus size, we should over-sample ADS at later ages. However, that may not be necessary for our data, where corpus size failed to explain away the register effect, even though it accounted for some variance beyond registers.

Last, speech transcriptions were used for this study, in an attempt to look for intrinsic informational differences across registers. However, some of the most salient features of IDS are speech-related, such as prosody or intonation and acoustic properties, which might also predict ease of segmentation. Although there is a small literature looking at word segmentation from speech, including comparing IDS and ADS (Ludusan, Seidl, Dupoux, & Cristia, 2015), this task remains extremely challenging for computational modelers, with only one open source model (instantiating a single segmentation strategy) exists, which further limits the value of such a line of research.

In sum, we identified several simplicity features more prevalent in IDS than ADS drawn from an ecological French corpus. We further found a small but significant IDS segmentation advantage, contributing to a recurrent question on the learnability properties of IDS. We showed that the IDS segmentation advantage could not be explained away by any one of those simplicity features, and its size changed with infant age in unexpected directions.

Acknowledgments

References

Batchelder, E. (1997). *Computational evidence for the use of frequency information in discovery of the infant's first lexicon*. Unpublished doctoral dissertation, City University of New York.

Batchelder, E. (2002). Bootstrapping the lexicon: A computational model of infant speech segmentation. *Cognition*, *83*(2), 167–206.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi: 10.18637/jss.v067.i01

Bergelson, E., Casillas, M., Soderstrom, M., Seidl, A., Warlaumont, A. S., & Amatuni, A. (2019). What do north american babies hear? a large-scale cross-corpus analysis. *Developmental science*, *22*(1), e12724.

Bernard, M. (2018). Phonemizer [Computer software manual]. `https://github.com/bootphon/phonemizer`, doi = "http://doi.org/10.5281/zenodo.2537809".

Bernard, M., Thiolliere, R., Saksida, A., Loukatou, G., Larsen, E., Johnson, M., ... Cristia, A. (2018). Wordseg: Standardizing unsupervised word form segmentation from text. *Behavior Research Methods*.

Brent, M. R., & Siskind, J. M. (2001). The role of exposure to isolated words in early vocabulary development. *Cognition*, *81*(2), B33–B44.

Canault, M., Le Normand, M.-T., Foudil, S., Loundon, N., & Thai-Van, H. (2016a). *Lyon homebank corpus.* doi: 21415/T58P6Q

Canault, M., Le Normand, M.-T., Foudil, S., Loundon, N., & Thai-Van, H. (2016b). Reliability of the language environment analysis system (lena^{TM}) in european french. *Behavior research methods*, *48*(3), 1109–1124.

Cristia, A., Dupoux, E., Ratner, N. B., & Soderstrom, M. (2018). Segmentability differences between child-directed and adult-directed speech: A systematic test with an ecologically valid corpus. *Open Mind*, 1–10.

Daland, R., & Pierrehumbert, J. B. (2011). Learning diphone-based segmentation. *Cognitive science*, *35*(1), 119–155.

Duddington, J. (2012). espeak text to speech [Computer software manual].

Fourtassi, A., Börschinger, B., Johnson, M., & Dupoux, E. (2013). Why is english so easy to segment? In *Proceedings of the fourth annual workshop on cognitive modeling and computational linguistics (cmcl)* (pp. 1–10).

Fox, J., & Weisberg, S. (2011). *An R companion to applied regression* (Second ed.). Thousand Oaks CA: Sage.

Henning, A., Striano, T., & Lieven, E. V. (2005). Maternal speech to infants at 1 and 3 months of age. *Infant behavior and development*, *28*(4), 519–536.

Johnson, E., Lahey, M., Ernestus, M., & Cutler, A. (2013). A multimodal corpus of speech to infant and adult listeners. *The Journal of the Acoustical Society of America*, *134*(6), EL534–EL540.

Johnson, E., Seidl, A., & Tyler, M. D. (2014). The edge factor in early word segmentation: utterance-level prosody enables word form extraction by 6-month-olds. *PloS one*, *9*(1), e83546.

Johnson, M., Griffiths, T. L., & Goldwater, S. (2007). Adaptor grammars: A framework for specifying compositional nonparametric bayesian models. In *Advances in neural information processing systems* (pp. 641–648).

Ludusan, B., Mazuka, R., Bernard, M., Cristia, A., & Dupoux, E. (2017). The role of prosody and speech register in word segmentation: A computational modelling perspective. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 2: Short papers)* (Vol. 2, pp. 178–183).

Ludusan, B., Seidl, A., Dupoux, E., & Cristia, A. (2015). Motif discovery in infant-and adult-directed speech. In *Proceedings of the sixth workshop on cognitive aspects of computational language learning* (pp. 93–102).

Ma, W., Golinkoff, R. M., Houston, D. M., & Hirsh-Pasek, K. (2011). Word learning in infant- and adult-directed speech. *Language Learning and Development*, *7*(3), 185–201.

MacWhinney, B. (1996). The childes system. *American Journal of Speech-Language Pathology*, *5*(1), 5–14.

Monaghan, P., & Christiansen, M. H. (2010). Words in puddles of sound: Modelling psycholinguistic effects in speech segmentation. *Journal of child language*, *37*(3), 545–564.

Newport, E., Gleitman, H., & Gleitman, L. (1977). Mother, id rather do it myself: Some effects and non-effects of maternal speech style.

Ngon, C., Martin, A., Dupoux, E., Cabrol, D., Dutat, M., & Peperkamp, S. (2013). (Non) words, (non) words, (non) words: Evidence for a protolexicon during the first year of life. *Developmental Science*, *16*(1), 24–34.

R Core Team. (2013). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org/`

Saksida, A., Langus, A., & Nespor, M. (2017). Co-occurrence statistics as a language-dependent cue for speech segmentation. *Developmental science*, *20*(3), e12390.

Shukla, M., White, K. S., & Aslin, R. N. (2011). Prosody guides the rapid mapping of auditory word forms onto visual objects in 6-mo-old infants. *Proceedings of the National Academy of Sciences*, *108*(15), 6038–6043.

Soderstrom, M. (2007). Beyond babytalk: Re-evaluating the nature and content of speech input to preverbal infants. *Developmental Review*, *27*(4), 501–532.

VanDam, M., Warlaumont, A. S., Bergelson, E., Cristia, A., Soderstrom, M., De Palma, P., & MacWhinney, B. (2016). Homebank: An online repository of daylong child-centered audio recordings. In *Seminars in speech and language* (Vol. 37, pp. 128–142).

Discovering a symbolic planning language from continuous experience

Joo Loula
MIT, Cambridge, Massachusetts, United States

Tom Silver
MIT, Cambridge, Massachusetts, United States

Kelsey Allen
Massachusetts Institute of Technology, Cambridge, Massachusetts, United States

Josh Tenenbaum
MIT, Cambridge, Massachusetts, United States

Abstract

Humans make plans with remarkable flexibility by leveraging symbolic representations. How are these representations learned? We present a model that starts out with a language of low-level physical constraints and, by observing expert demonstrations, builds up a library of high-level concepts that afford planning and action understanding. We demonstrate its versatility through experiments inspired by developmental psychology literature.

Attentional Capture: Modeling Automatic Mechanisms and Top-Down Control

Andrew Lovett (andrew.lovett@nrl.navy.mil)
Will Bridewell (will.bridewell@nrl.navy.mil)
Paul Bello (paul.bello@nrl.navy.mil)
U.S. Naval Research Laboratory, 4555 Overlook Ave SW
Washington, DC 20375 USA

Abstract

We present a computational model of attentional capture in humans. The model distinguishes between automatic mechanisms that directly determine the focus of visual attention, and deliberate mental actions an individual can perform to influence these mechanisms. The automatic mechanisms select an object as the focus of attention and enhance its location and features, so that nearby or similar objects are likely to be selected in the future. The deliberate actions include engaging with a selected object to further enhance its features, and retrieving a previously selected object from memory. By performing these actions, the model is able to exert limited top-down control over capture, increasing the probability that task-relevant objects will be attended and irrelevant objects will be ignored. To evaluate the model, we conduct a simulation of a recent visual search study, demonstrating that the model can account for three established factors that are known to influence capture.

Keywords: visual attention; visual search; computational modeling

Introduction

What drives attentional capture? That is, when we view a scene, why is our attention drawn to one object, and not to another? This question is important because where we attend determines what information we represent. Whether we are reading a map, driving a car, or shopping at a store, we can perform the task more efficiently if we attend to objects that provide relevant information and ignore task-irrelevant objects.

Much of the debate over attentional capture concerns the role of top-down control (Folk, Remington, & Johnston, 1992; Müller, Reimann, & Krummenacher, 2003; Theeuwes, Reimann, & Mortier, 2006). To what extent can humans deliberately manipulate our own mental states, such that task-relevant objects are more likely to be attended? The evidence suggests that in many cases, task-relevant objects draw attention not because of deliberate control, but because they are similar to objects we have attended recently. For example, if a task involves looking for red objects, the act of finding a red object on previous trials will prime the viewer to find one more easily on future trials (Maljkovic & Nakayama, 1994; Theeuwes et al., 2006). However, in some cases participants appear to be able to strategically tune their attentional systems based on semantic information, such as a word describing the color of the object they should find next (Leonard & Egeth, 2008; Belopolsky & Awh, 2016).

To better understand how top-down goals affect attentional capture, it is helpful to model the specific mechanisms underlying attention. We previously developed a model of multiple-object tracking that relied on two attentional mechanisms: selection and enhancement (Lovett, Bridewell, & Bello, 2017). Selection picks out an item for further processing, and may be thought of as a generalized form of attentional capture, whereas enhancement increases sensitivity to stimuli at a particular location or with particular features. These two mechanisms are closely interwoven: after an object is selected, its location and features are enhanced, such that objects at the same location or with similar visual features are more likely to be selected in the future.

Here, we present a novel computational model that applies the selection and enhancement mechanisms to a visual search task, in which participants must find a blue or orange circle in a field of distractor circles and judge the orientation of a line inside it (Figure 1). Critically for the topic at hand, neither selection nor enhancement is directly controlled in the model. However, other deliberate actions can influence what gets enhanced, thereby biasing the model to select task-relevant objects. In particular, after an object is selected, if the object is task-relevant then the model can `engage` with it. Engagement is the act of maintaining focus on an object while reasoning about its features, for example, judging the orientation of a line inside an attended circle. Engagement leads to greater enhancement of an object's location and features, which supports sustained selection of that object but also causes objects with similar features to be selected in the future.

In the model, engagement also causes the object's representation to be stored in long-term memory, from which it can be retrieved at a later time. Thus, if the model later receives a cue, for example indicating that the next search target will be orange, it can deliberately `retrieve` a representation of a previously selected orange object from memory, allowing that representation to be selected and engaged with, so that orange objects are more likely to be selected.

In the following section, we describe three factors that affect attentional capture, and we argue that our model, which integrates selection and enhancement mechanisms with deliberate mental actions, can explain each factor. We then present the model and describe an evaluation in which it simulates human performance on a search task. We close by considering predictions of the model and directions for future research.

Background

At least three factors govern which objects capture visual attention when viewing a scene: physical salience, selection history, and top-down goals (Awh, Belopolsky, & Theeuwes, 2012). Physical salience increases with the amount of contrast between an object and the rest of the scene, but decreases with the amount of contrast between the other objects in a scene; for example, a red circle will be strongly salient in a field of identical green squares (Duncan & Humphreys, 1989). Salience is determined by both local contrast (between an object and its immediate surroundings) and global contrast (between an object and the other objects throughout the visual scene) (Nothdurft, 1993; Madison, Lleras, & Buetti, 2018).

Whereas salience is a property of the visual stimuli, selection history relates to the viewer's mental state. An object will tend to draw attention if it is visually similar to objects that have been attended in the recent past. In search tasks, this effect often manifests as intertrial priming, where a target is found more easily if its features remain constant from one trial to the next (Maljkovic & Nakayama, 1994). Similarly, a target is found more easily if it is in the same location as a recently attended object (Folk et al., 1992).

Finally, top-down goals involve deliberate control over what object captures attention. This effect is demonstrated when viewers see a cue describing a target, rather than an object similar to the target, and then are able to find the target more readily. A spatially descriptive cue might be an arrow pointing to the region where the target will appear (Posner, 1980), whereas a featurally descriptive cue might be a word describing the target's distinguishing feature (e.g., "red") (Leonard & Egeth, 2008). The ability to use these cues suggests the viewer is making an adjustment that causes objects that match the description to draw attention.

Recently, Belopolsky and Awh (2016) examined the combined contributions of these three factors to attentional capture. They used a search task in which participants viewed six colored circles, found a target circle that could be either blue or orange, and reported whether the line inside the circle was horizontal or vertical (Figure 1). To explore the effect of salience, the colors of the distractor circles were varied: on half the trials, all the distractors were green, resulting in a salient target, whereas on the other half, the distractors were all different colors, resulting in a nonsalient target. To explore the effect of top-down goals, each search trial was preceded by a verbal cue, either the word "blue" or "orange," that predicted the upcoming target's color 80% of the time. Finally, to explore the effect of selection history, performance on repetition trials, where the target's color was the same as the color from the previous trial (e.g., the circle was orange for two trials in a row), was contrasted with performance on non-repetition trials.

Critically, in one study Belopolsky and Awh (2016) presented the search display for only 100 ms, after which the lines within each circle were masked. This brief display time has two major advantages: (1) there is no time to saccade to one of the circles, so eye movements cannot be a factor, and (2) if the first circle attended by the participants is neither blue nor orange, there is no time to look for another circle. Thus, the authors were able to isolate attentional capture from the separate task of assessing whether an attended object meets the search criteria.

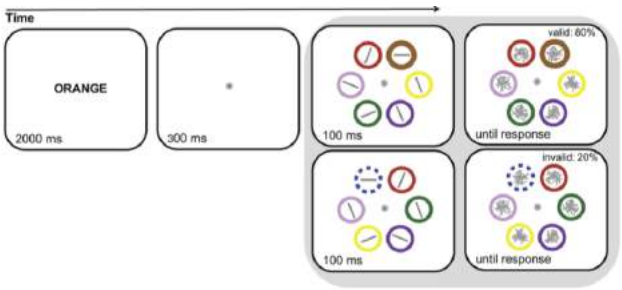


Figure 1: Examples of visual search task with a nonsalient target from Belopolsky and Awh (2016). The top row shows a valid trial with an orange target (appears brownish), whereas the bottom row shows an invalid trial with a blue target. The blue circle is dotted for illustration purposes.

Figure 2 shows the experiment results. Accuracy increased when the target was salient, when the cue was valid (e.g., the cue "orange" preceded an orange circle), or when a target color repeated, indicating that each of the three factors contributed to attentional capture. In addition, there were numerous interactions, notably, cue validity had a greater effect when the target was nonsalient, target repetition had a greater effect when the target was nonsalient, and there was a three-way interaction among the factors. We propose that these interactions are driven by a ceiling effect. As an example, when a target is salient, there is a high likelihood of attending to it during the critical 100 ms, and thus there is little room for additional improvement if the cue is valid.

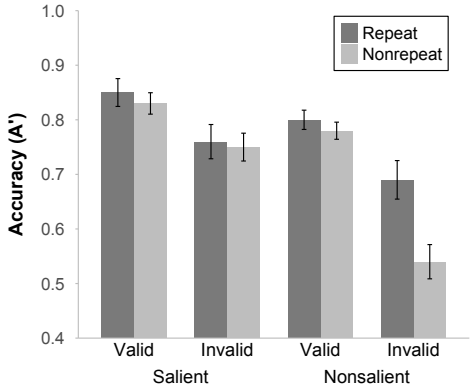


Figure 2: Visual search results from Belopolsky and Awh (2016). Error bars are ± 1 SE.

Selection and Enhancement

Both selection history and top-down goals may result from an interaction between selection and enhancement. After an object is selected, viewers show enhanced sensitivity to other objects in the same location or with the same visual features (Posner, 1980; Egly, Driver, & Rafal, 1994; Bichot, Rossi, & Desimone, 2005). Enhancement manifests as both a greater probability of selecting a stimulus among a field of distractors, and a shorter delay between stimulus onset and selection. Neural evidence suggests enhancement is rooted in modulation of the early visual cortex, for example, after a red object is selected, neurons will respond more strongly to red stimuli throughout the visual field (Somers, Dale, Seiffert, & Tootell, 1999; Saenz, Buracas, & Boynton, 2002).

Applying selection and enhancement to the Belopolsky and Awh (2016) study (discussed previously), the effect of selection history can be readily explained: participants should select an orange circle more quickly if the previous target was also orange because the recent selection would cause the orange color to be enhanced. Explaining top-down goals requires one further step—-after participants view a cue such as the word "orange," they must perform some mental action that produces a representation of an orange object, so that the representation can be selected and the color orange can be enhanced. We propose that participants retrieve a previous example of an orange object from memory. Such a retrieval should be easy, as participants are regularly engaging with orange circles throughout the experiment (note that one alternate hypothesis might be that participants perform mental imagery, imagining an orange circle).

In the next section, we describe a computational model of human performance on the Belopolsky and Awh (2016) search task.

Model

The model is based on three core claims about human attentional processing.

1) Selection picks out a single focus of attention, such as an object in the visual field. Objects are selected based on their *activation strength*, which is a combination of physical salience and spatial/featural enhancement. An object with a higher activation strength is more likely to be selected from among a field of other objects. In addition, an object with a higher activation strength will be selected more quickly after its onset.

2) Selecting an object enables constructing an object representation that can be stored in *visual short-term memory* (VSTM), which is a low-capacity store for representations of recently selected objects (Treisman & Gelade, 1980; Vogel, Woodman, & Luck, 2001). Once an object is represented in VSTM, the viewer can decide to `engage` or `disengage` with the object, depending on whether the object is task relevant. Engagement makes an object's features accessible for further reasoning and supports storing the object's representation in *long-term memory* (LTM), where it will be available for retrieval at a later time. In addition, engagement causes an object's location and features to be enhanced, which helps to maintain focus on the object, while also increasing the probability that nearby or similar objects will be selected. In contrast, disengaging from the object causes its location to be suppressed, so that a different object can be selected.

3) A viewer can `retrieve` an object representation matching a verbal cue (e.g., "orange") from LTM. If this retrieved object representation is selected, then it will be stored back in VSTM, and its features can be enhanced.

Model Framework

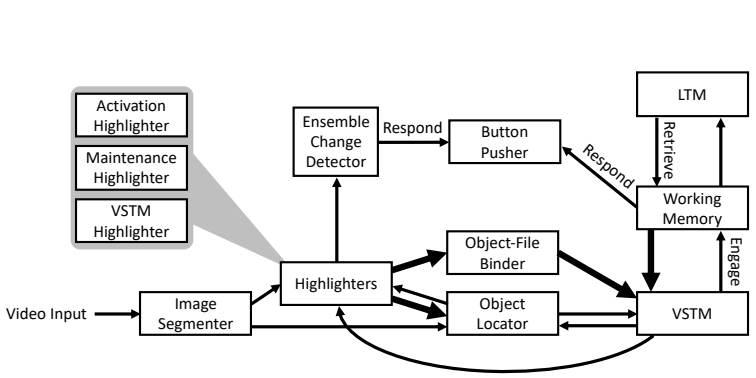
The model is implemented in ARCADIA (Bridewell & Bello, 2016), a computational framework developed to explore the relationships among attention, perception, cognition, and action. ARCADIA models operate over a sequence of cycles. On each cycle, a set of components work in parallel, processing input and generating output. One output item is selected as the focus of attention, and then the next cycle commences, with components receiving as input the output from other components on the previous cycle.

Models built in ARCADIA consist of (1) a set of components; (2) an *attentional strategy*, which sets out the priorities for which component's output will be selected as the focus of attention after each cycle, and (3) optionally, a set of *stimulus-response links*, which indicate that once certain conditions are met, an action should be taken.

Model Runthrough

Figure 3 presents the model's components and illustrates the flow of information. Thin arrows indicate information that flows on every cycle, thick arrows indicate information that flows only when it is selected as the focus of attention, and arrows accompanied by words indicate information that flows only when an action is taken. In the following sections, we shall describe the components and the flow of information in detail, using the search task in Figure 1 as a running example. Note that the model is designed to run on stimuli identical to those shown to humans, with one exception: because the model lacks reading comprehension, the verbal cues "orange" and "blue" are replaced with horizontal and vertical rectangles, respectively. Thus, a horizontal rectangle indicates that the next target will likely be orange.

Figure 3 also provides the model's stimulus-response links, which indicate the conditions under which the model should `engage` with an object, `retrieve` an object representation from memory, or `respond` by pressing a virtual button to end a trial. Whereas many of the model's components perform general-purpose visual processing and have been used in other task models (Bridewell & Bello, 2016; Lovett et al., 2017), the stimulus-response links encode task-specific knowledge about when actions should be performed (e.g., a horizontal rectangle indicates an orange object should be retrieved from memory). These actions provide a means for the model to influence the selection and enhancement mechanisms, and thereby increase the likelihood of task-relevant



Component	Conditions	Action
VSTM	Object is orange, blue, horizontal, or vertical	`Engage` with object
Working Memory	Object is horizontal	`Retrieve` an orange object
Working Memory	Object is vertical	`Retrieve` a blue object
Working Memory	Object is orange or blue	`Respond` "horizontal" or "vertical"
Ensemble Change Detector	Features pooled across the image have changed, and there are multiple objects on screen	`Respond` "guess"

Figure 3: Left: Flow of information between model components. Right: Stimulus-response links for the model.

objects being selected.

Returning to the model's components, processing begins with the Image Segmenter, which takes each frame from an input video and segments it into regions representing possible objects. In the current example, each video begins with a horizontal or vertical rectangle, so the component identifies only one region of interest. Later, when there is a fixation circle surrounded by six larger circles, the component identifies seven regions.

Regions of interest are quickly forgotten unless they are selected as the focus of attention. To this end, components known as *highlighters* suggest particular regions as candidates for attention. In the present model, the Activation Highlighter suggests a region if its combined salience and enhancement (discussed in greater detail later) exceeds a threshold. In contrast, the Maintenance Highlighter suggests the region whose location matches the current focus of attention; this component supports maintaining focus on an object over time. Finally, the VSTM Highlighter suggests a region whose location matches any object represented in VSTM; this component supports returning focus to a recently selected object. Note that the model's attentional strategy gives the highest priority to the Activation Highlighter and the lowest priority to the VSTM Highlighter. This means the model will (1) focus on an object with a sufficiently high activation strength, or if none have a sufficiently high strength, (2) maintain focus on the currently selected object, or if no current object is selected, (3) return focus to a recently selected object.

In the visual search example, when only the fixation circle is visible, it will be selected. When the six outer circles appear around the fixation circle, focus will be maintained on the fixation circle until the activation strength of one of the outer circles exceeds the threshold.

After a region is selected, the Object-File Binder constructs an object representation describing what is found at that region, while at the same time the Object Locator records the region's location and begins tracking the object. The object representation includes the object's physical dimensions and visual features (color, orientation, and brightness). In the current example, the representation contains the necessary information for determining whether a rectangle is vertical or horizontal, determining the color of a circle, or determining whether the line inside a circle is horizontal or vertical.

After an object representation is constructed, the attentional strategy prioritizes selecting it as the focus of attention, so that it can be stored in VSTM (visual short-term memory) which holds representations of the four most recently selected objects. At this point, if the object is task-relevant (a blue or orange target circle, or a vertical or horizontal rectangular cue), the model's stimulus-response links trigger an `engage` action (Figure 3, right side). Engaging with an object causes its representation to move into Working Memory, where it is accessible to other components. In addition, engaging causes the object's location and features to be enhanced (in the present model, the location is enhanced only when the object is visible, and the only feature that can be enhanced is color). For simplicity, if the model does not `engage` with an object, then the model behaves as if it had `disengaged` with the object: the object's location is suppressed, which encourages selection of other objects. Note that all enhancement and suppression effects last only while the object is remembered in VSTM.

In the model, Working Memory functions as a conduit between VSTM and LTM. After a representation is copied from VSTM to Working Memory, it is stored in LTM, which has a greater capacity than VSTM. Later, if the model performs a `retrieve` action, an object representation is copied back into Working Memory, where it can be selected and stored in VSTM. The model's stimulus-response links specify that it should retrieve an orange object after engaging with a horizontal rectangle, or retrieve a blue object after engaging with a vertical rectangle.

In the visual search example, the interactions between VSTM, Working Memory, and LTM give rise to effects of selection history and top-down goals. Suppose two sequential trials each involve an orange circle, and suppose the model successfully selects the orange circle on the first trial. Beginning with this first orange circle, the model will perform the following sequence of selections:

1. Select the first orange circle and generate a response. This ends the first trial.

2. Select the rectangular cue at the beginning of the next trial.

3. Based on the cue, retrieve an object representation from LTM and store it in Working Memory. Select this object representation.

4. Select the fixation circle that precedes the critical 100 ms.

As each of these object or object representations is selected, it will be stored in VSTM. Because VSTM has a capacity for four objects representations, all four will remain in VSTM at the beginning of the second trial's critical 100 ms. Because the circle from the previous trial is in VSTM, its color will be enhanced, resulting in a selection history effect; and because the circle retrieved from LTM is in VSTM, its color will be enhanced, resulting in a top-down goal effect.

Finally, the Button Pusher is passed one of three responses: "vertical" or "horizontal" if the model engages with a target (blue or orange) circle and determines the orientation of its inner line, or a "guess" response if the masks cover the circles before the model engages with a target circle. The appearance of the masks is detected by the Ensemble Change Detector, which responds to large-scale changes to the image.

Overall, the model succeeds at the search task if it selects and engages with the target circle during the 100 ms before the masks appear, enabling it to generate the appropriate response. It fails if either it selects the target circle after the masks appear, in which case the response may be incorrect; or it never selects the target circle, in which case it generates a "guess" response.

Activation Highlighter

The Activation Highlighter integrates salience and enhancement to determine each region's activation strength. Salience is computed via a novel algorithm based on Itti, Koch, and Niebur's (1998) classic computational approach. Operating over the color, orientation, and brightness dimensions, the algorithm computes local contrast throughout the image, and then computes global contrast for each region of interest. A region's salience varies from 0 to 1, where 1 indicates the region strongly stands out on one dimension (e.g., its color is unique, whereas the other regions all have similar colors), or moderately stands out on multiple dimensions.

Spatial enhancement is computed based on whether the region overlaps the location of an object in VSTM. For simplicity, we assign a score of 1 if it overlaps an enhanced object, -1 if it overlaps a suppressed object, and 0 if it does not overlap an object.

Featural enhancement, currently computed only for color, is based on the similarity between colors within a region and colors of objects being enhanced. A region will receive a score of 1 if it perfectly matches the colors of all enhanced objects. Note that in some cases, two different colors may be enhanced—for example, if the previous trial involved a blue circle, but the model just retrieved an orange circle from memory. In these cases, a region will receive a score based on the average of its color match to the two enhanced objects.

To ensure some randomness, Gaussian noise is added to the activation strength, according to the following formula:

$$Gaussian(gaussian\text{-}width) + weight_{sal} * Salience +$$

$$(1 - weight_{sal}) * 0.5(Enhancement_{space} + Enhancement_{features})$$

Finally, the Activation Highlighter computes the average activation strength over the past five cycles and compares this average to an *activation-threshold* to determine whether a region has a sufficiently high score to be selected. Averaging over five cycles achieves the desired effect that objects with more salience or enhancement will be selected more quickly, as it will take fewer cycles after onset for the running average to exceed *activation-threshold*.

Note that there are three free parameters: $weight_{sal}$ the weight given to salience, relative to enhancement; *gaussian-width* the width of the Gaussian noise; and *activation-threshold*. For now, we set $weight_{sal}$ to 0.2 (meaning salience receives one quarter the weight of enhancement), and we shall use the simulation that follows to explore possible noise and threshold values.

Evaluation

To simulate the Belopolsky and Awh (2016) search task, we generated input videos that match the original study's stimuli exactly, with two exceptions: (1) as discussed previously, the verbal cues "orange" and "blue" were replaced with horizontal and vertical rectangles; (2) some portions of each trial were sped up to save processing time, but the critical 100 ms display time went unchanged.

In the original experiment, 24 participants each viewed a large number of practice trials, followed by 600 search trials. For the simulation, five virtual participants each viewed 40 practice trials, followed by 600 search trials. Because the virtual participants were all the same model, and they differed only in the particular trials they viewed, we combined the 3000 (5×600) results and analyzed by item. To reduce variance, "guess" responses were treated as 0.5 correct.

We ran the simulation across a range of *activation-threshold* and *gaussian-width* values. Figure 4 presents the results with a low (0.04) or medium (0.11) threshold, and with no or moderate (0.1) noise. Overall, it appears that a medium threshold and some noise were needed to achieve human-like performance; without these, the model performed at or near ceiling for all salient targets. The rightmost graph in Figure 4 closely matches the human results (Figure 2, note that the units are different), but a qualitative comparison suggests that repetition provides a stronger benefit to the model than to humans. In the model, repetition and valid cues provide similar benefits, but perhaps the benefit from repetition should be weaker because viewers stop engaging with a target circle after a trial ends.

To examine the benefits of target salience, cue validity, and repetition, we conducted an ANOVA for each simulation run. These analyses confirmed that all three factors contributed

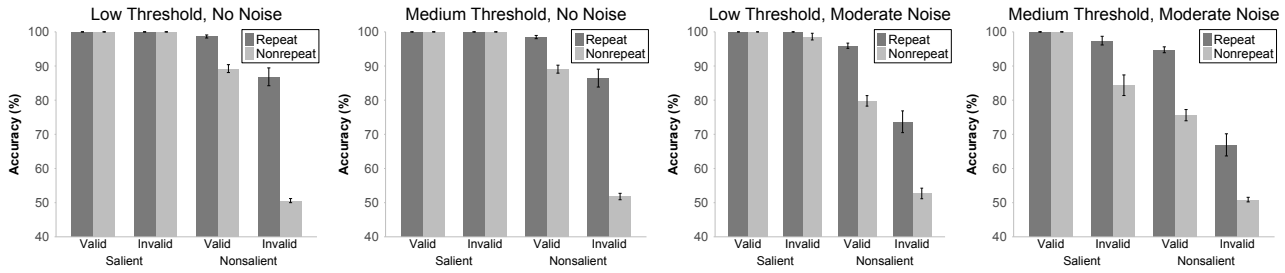


Figure 4: Model simulation results. Error bars are ± 1 SE.

significantly to accuracy (all $ps < .05$), and additionally found that most interactions between factors were significant. As we discussed when considering the human results, we believe these interactions are driven by a ceiling effect—note that performance is at 100% for some conditions.

Conclusion

Our computational model is able to perform a visual search task, while demonstrating how salience, selection history, and top-down goals influence attentional capture. In particular, the acts of `engaging` with task-relevant objects and `retrieving` previously selected objects influence which features become enhanced, thereby causing relevant objects to be selected more easily in the future.

Ultimately, the model suggests that humans possess only limited top-down control over attentional capture. For example, the model predicts that a verbal cue will be effective only when viewers are able to act on it. Suppose that after many trials of the visual search experiment, viewers are presented with a novel verbal cue, such as "red." This cue should provide little benefit because viewers have not been engaging with red circles, and thus red circles are unavailable for retrieval. In contrast, a novel *visual* cue, such as an image of a red circle preceding the search trial, should provide an immediate benefit because selecting the red circle causes its features to be enhanced.

In developing this model, we drew inspiration from previous models of visual search and attentional capture. Notably, most models explain the influence of salience (Itti et al., 1998), top-down goals (Wischnewski, Steil, Kehrer, & Schneider, 2009), or both (Tsotsos, Kotseruba, & Wloka, 2016). We believe our model is unique in explaining the influences of salience, top-down goals, and selection, while making explicit claims about the limits of top-down control.

Moving forward, we plan to evaluate our model and the parameters that have been calibrated on the present task by simulating additional search tasks. These will include conjunctive searches, in which there is benefit to enhancing multiple feature dimensions (color, orientation, curvature, etc) in parallel (Wolfe, 2007). In addition, these will include longer searches in which there is time to move the eyes. Eye movement—which can be simulated in ARCADIA—is another deliberate action that influences selection and enhancement. Thus, viewers can optimize their search performance through strategic control of their looking patterns (Pomplun, Garaas, & Carrasco, 2013). By modeling the actions and strategic decisions that affect attentional capture, we hope to better understand how people can effectively extract important, task-relevant information from the world around them.

Acknowledgments

This research was performed while the first author held an NRC Research Associateship award at the U.S. Naval Research Laboratory. The authors would like to acknowledge support from the Office of Naval Research. The views expressed in this paper are solely the authors' and should not be taken to reflect any official policy or position of the United States Government or the Department of Defense.

References

Awh, E., Belopolsky, A. V., & Theeuwes, J. (2012). Top-down versus bottom-up attentional control: A failed theoretical dichotomy. *TiCS*, *16*(8), 437–443.

Belopolsky, A. V., & Awh, E. (2016). The role of context in volitional control of feature-based attention. *JEP: Human Perception & Performance*, *42*(2), 213–224.

Bichot, N. P., Rossi, A. F., & Desimone, R. (2005). Parallel and serial neural mechanisms for visual search in macaque area V4. *Science*, *308*(5721), 529–534.

Bridewell, W., & Bello, P. (2016). A theory of attention for cognitive systems. In *Fourth annual conference on advances in cognitive systems.*

Duncan, J., & Humphreys, G. W. (1989). Visual search and stimulus similarity. *Psych Review*, *96*(3), 433–458.

Egly, R., Driver, J., & Rafal, R. D. (1994). Shifting visual attention between objects and locations: Evidence from normal and parietal lesion subjects. *JEP: General*, *123*(2), 161–177.

Folk, C. L., Remington, R. W., & Johnston, J. C. (1992). Involuntary covert orienting is contingent on attentional control settings. *JEP: Human Perception & Performance*, *18*(4), 1030–1044.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(11), 1254–1259.

Leonard, C. J., & Egeth, H. E. (2008). Attentional guidance in singleton search: An examination of top-down, bottom-up, and intertrial factors. *Visual Cognition*, *16*(8), 1078–1091.

Lovett, A., Bridewell, W., & Bello, P. (2017). Goal-directed deployment of attention in a computational model: A study in multiple-object tracking. In *Proceedings of the 39th annual meeting of the cognitive science society* (pp. 2640–2645). London.

Madison, A., Lleras, A., & Buetti, S. (2018). The role of crowding in parallel search: Peripheral pooling is not responsible for logarithmic efficiency in parallel search. *Attention, Perception, and Psychophysics*, *80*, 352–373.

Maljkovic, V., & Nakayama, K. E. N. (1994). Priming of pop-out : I. Role of features. *Memory & Cognition*, *22*(6), 657–672.

Müller, H. J., Reimann, B., & Krummenacher, J. (2003). Visual search for singleton feature targets across dimensions: Stimulus- and expectancy-driven effects in dimensional weighting. *JEP: Human Perception & Performance*, *29*(5), 1021–1035.

Nothdurft, H. C. (1993). Saliency effects across dimensions in visual search. *Vision Research*, *33*(5-6), 839–844.

Pomplun, M., Garaas, T. W., & Carrasco, M. (2013). The effects of task difficulty on visual search strategy in virtual 3D displays. *Journal of Vision*, *13*(3), 1–22.

Posner, M. I. (1980). Orienting of attention. *The Quarterly Journal of Experimental Psychology*, *32*(1), 3–25.

Saenz, M., Buracas, G. T., & Boynton, G. M. (2002). Global effects of feature-based attention in human visual cortex. *Nature Neuroscience*, *5*(7), 631–632.

Somers, D. C., Dale, A. M., Seiffert, A. E., & Tootell, R. B. (1999). Functional MRI reveals spatially specific attentional modulation in human primary visual cortex. *PNAS*, *96*(4), 1663–8.

Theeuwes, J., Reimann, B., & Mortier, K. (2006). Visual search for featural singletons: No top-down modulation, only bottom-up priming. *Visual Cognition*, *14*(4-8), 466–489.

Treisman, A., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97–136.

Tsotsos, J. K., Kotseruba, I., & Wloka, C. (2016). A focus on selection for fixation. *Journal of Eye Movement Research*, *9*(5), 1–34.

Vogel, E. K., Woodman, G. F., & Luck, S. J. (2001). Storage of features, conjunctions, and objects in visual working memory. *JEP: Human Perception & Performance*, *27*(1), 92–114.

Wischnewski, M., Steil, J. J., Kehrer, L., & Schneider, W. X. (2009). Integrating inhomogeneous processing and proto-object formation in a computational model of visual attention. In H. Ritter, G. Sagerer, R. Dillmann, & M. Buss (Eds.), *Human centered robot systems: Cognition, interaction, technology* (pp. 93–102). Berlin: Springer.

Wolfe, J. M. (2007). Guided search 4.0: Current progress with a model of visual search. In W. Gray (Ed.), *Integrated models of cognitive systems* (pp. 99–119). New York: Oxford.

Seeing the Meaning: Vision Meets Semantics in Solving Pictorial Analogy Problems

Hongjing Lu[1,2] hongjing@ucla.edu
Qing Liu[3] qingliu@jhu.edu
Nicholas Ichien[1] ichien@ucla.edu
Alan L. Yuille[3] alan.yuille@jhu.edu
Keith J. Holyoak[1] holyoak@lifesci.ucla.edu

[1] Department of Psychology, [2] Department of Statistics
University of California, Los Angeles, Los Angeles, CA 90095 USA
[3] Department of Cognitive Science
Johns Hopkins University, Baltimore, MD 21218 USA

Abstract

We report a first effort to model the solution of meaningful four-term visual analogies, by combining a machine-vision model (ResNet50-A) that can classify pixel-level images into object categories, with a cognitive model (BART) that takes semantic representations of words as input and identifies semantic relations instantiated by a word pair. Each model achieves above-chance performance in selecting the best analogical option from a set of four. However, combining the visual and the semantic models increases analogical performance above the level achieved by either model alone. The contribution of vision to reasoning thus may extend beyond simply generating verbal representations from images. These findings provide a proof of concept that a comprehensive model can solve semantically-rich analogies from pixel-level inputs.

Keywords: analogy; relations; learning; machine vision; word embeddings

Introduction

In everyday life, humans continually perceive the world and interpret it in terms of meaningful objects and events. The representations extracted by perception are elaborated into semantic representations that can be communicated by language and further transformed by reasoning processes. The "holy grail" of cognitive science is to develop integrated theories that link perception to language and higher cognition. A natural testbed for developing such integrated theories is the task of reasoning by analogy from meaningful visual inputs. Here we report a first effort to develop a comprehensive model of the solution of visual analogies, by combining a model that can translate pixel-level inputs into verbal captions with a model that can translate semantic vectors for words into coherent patterns of semantic relations.

Figure 1 depicts an example of the analogies on which we focus. This problem is one of a set of 18 developed by Krawzcyk et al. (2008), some of which were adapted from an earlier set created by Goranson (2002), hence dubbed the Goranson Analogy Test (GAT). The upper row presents a pictorial problem in the form *A:B :: C:?*. The task is to select the best analogical completion from among a set of four options shown in the bottom row. For this example, the analogical solution based on matching relations is to choose the pie (wine is made from grapes, as pie is made from pumpkin). The three distractors include one that is semantically related to the *C* term but fails to match the *A:B* relation (witch), one that is visually similar but also fails to match *A:B* (basketball), and one that is simply unrelated (books). Critically, the analogical solution cannot in any obvious way be derived from visual information alone, because the core relation is semantic/functional rather than visual. For example, the fact that wine is made from grapes is not depicted in the visual input; rather, it must be retrieved from semantic memory. Thus, vision is necessary but not sufficient to reliably solve such semantically-rich picture analogies.

The GAT was originally developed as a tool to evaluate the impact of neuropsychological disorders. Krawczyk et al. (2008) found that frontal and temporal patients were impaired to varying degrees, notably showing an elevated tendency to choose the semantic or perceptual distractors. Age-matched controls (approximately age 60) achieved about 98% accuracy even in the presence of similar distractors.

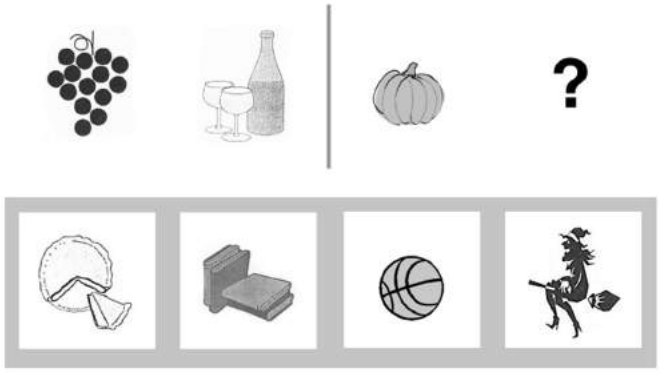


Figure 1. Example of a 4-term pictorial analogy with four alternatives (from Krawzcyk et al., 2008).

Here we focus on the most fundamental question: how can such pictorial analogy problems be solved at all? On the face of it, the process begins with the human visual system operating on pixel-level inputs of the images in the problem to extract a verbal description and/or semantic categorization of the objects. Reasoning processes must use these object descriptions to determine the relation(s) linking paired objects. Based on these relational representations, the reasoner must then assess the degree of relational match between *A:B* and the alternative completions for *C*, finally choosing the option that provides the best match.

Despite decades of progress in developing computational models of visual perception, language processing, and analogical reasoning, no model has tackled the full range of processing required to solve meaningful visual analogies such as the GAT problems. Recent advances in machine vision have led to very significant progress in the recognition of objects from pixel-level representations (Krizhevsky, Sutskever & Hinton, 2012; Semonvan & Zisserman, 2015), including the automatic generation of verbal captions (Farhadi et. al., 2010; Mao et. al., 2016; Krishna et. al., 2016). However, artificial-intelligence (AI) models have been less successful in transforming visual inputs into semantic representations of *relations between objects*. AI models of visual analogy have generally focused on problems that can be solved on the basis of simple visual features, such as color and shape (Reed et. al., 2015; Sadeghi, Zitnick & Farhadi, 2015). In cognitive science, most analogy models have simply assumed high-level representations of complex propositions (usually hand-coded), without dealing with the problem of how these representations could be generated by perceptual processes. Lovett and Forbus (2017) describe a model that applies analogical reasoning to solve Ravens Progressive Matrices problems, which are a form of visual analogies based on transformations of geometrical shapes. However, the inputs provided to the model are high-level perceptual descriptions, rather than a matrix of pixels; and the Ravens test is entirely formal, devoid of any links to semantic knowledge. With important exceptions (e.g., Doumas, Hummel, & Sandhofer, 2008), analogy models have generally set aside the basic problem of how semantic relations could be learned from non-relational inputs.

Here we describe two computational models that together provide an approximate account of the entire process that may underlie solution of GAT problems. One model, ResNet50-A, aims to solve the picture analogies using purely visual information, while also generating verbal captions. The other, BART, aims to solve the same analogies based solely on verbal descriptions of the images. We further show that the analogy assessment derived by ResNet50-A using just visual information not only provides potential verbal inputs to BART, but also adds independent visual information that increases solution accuracy. We will first describe the operation of each of the two models, and then the results obtained by using them both separately and jointly.

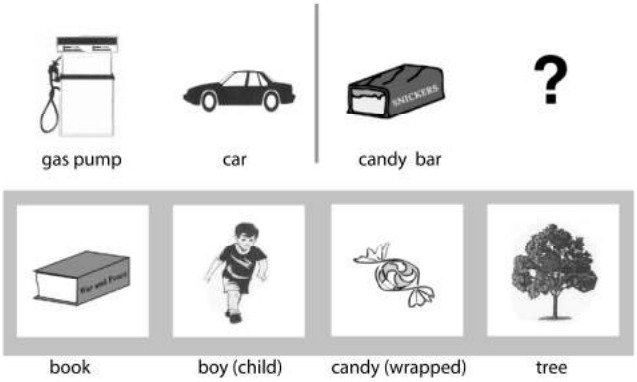


Figure 2. Example of a 4-term pictorial analogy with four alternatives, and corresponding descriptions verbally presented to patients (from Krawzcyk et al., 2008).

GAT Dataset

The GAT dataset includes 18 picture analogies, each consisted of 7 images: the three images in the question, *A*, *B*, and *C*, and the four images for alternative D terms. All images are line drawings or clip art images. Each image was captured in the size of 140x140 pixels. The GAT dataset included a total of 126 images that fall into 118 distinct object categories. A verbal caption describing each image was used by Krawczyk et al. (2008) in their neuropsychological study; these captions were adopted as canonical verbal descriptions of each image for the semantic model, BART. Figure 2 shows a second example, along with approximations of the corresponding verbal descriptions used by Krawczyk et al. (2008). Note that in the neuropsychological study, the accompanying labels were presented orally by the experimenter, rather than in written form.

ResNet: From Pixels to Object Classification

Background

Deep convolutional neural networks (Krizhevsky, Sutskever, & Hinton, 2012; Simonyan & Zisserman, 2015) have led to a series of breakthroughs for a broad range of computer vision tasks. The network depth is of crucial importance. Recent work with deeper networks has exposed a degradation problem: as network depth increases, accuracy reaches a plateau, and then degrades rapidly as network depth increases further. ResNet (He, Zhang, Ren, & Sun, 2016) addresses the degradation problem by introducing a framework termed *deep residual learning*. ResNet fits a residual mapping, realized by a feedforward neural network with identity shortcut connections. Using this method, ResNet can be efficiently trained with as many as 1000 layers. Because of its compelling performance levels, ResNet has quickly emerged as one of the leading architectures for a wide range of tasks in computer vision. Here we adopt ResNet50 (the basic architecture with 50 layers) as a state-of-the-art approach to identifying and captioning the objects in GAT analogies. We then augment the model to create ResNet50-A (where the "A" stands for "Analogy") by adding a decision procedure to generate

potential analogical solutions based solely on visual information in the images.

Training Dataset

The GAT images are line drawings (as are most images used in picture analogy tests that have been developed for psychological research or cognitive assessments). Machine vision models are typically trained on photo-realistic images, and require additional training with line drawings in order to classify them. In order to provide suitable training for ResNet50, we created a database of clip art images that were similar to GAT images, but not identical to them. This dataset, termed the ClipArt dataset, includes the 118 object categories used in the GAT visual analogy problems. To create the ClipArt dataset, we queried Google Image Search using the "Search by image" function, uploading the corresponding GAT image and entering a phrase formed by concatenating the category label and the words "clip art". (For some categories, we visually checked the result and decided to replace "clip art" by "drawing", "sketch", or "cartoon".) We downloaded 200 images for each category and manually removed those that were duplicates or clearly wrong. Each category in the resulting ClipArt dataset was represented by 70-166 images. The images were then processed into gray scale and padded with zero on short edges to fit a 1:1 aspect ratio.

For each category, we randomly selected 50 images for training, and held the rest images for test, resulting in a total of 5900 training images and 5501 test images. Figure 3 juxtaposes a GAT image (left) with a ClipArt image (right) from the same category. To ensure that the model was able to generalize its visual recognition performance, the GAT dataset was only used to guide construction of the ClipArt dataset; the GAT images themselves were not used to train ResNet50.

Training

We implemented ResNet50 using Pytorch on a single TitanX GPU. The training task was image classification by minimizing the cross-entropy loss. The model was pretrained on the ImageNet dataset, and then fine-tuned on our ClipArt dataset for 200 epochs. Batch size was set equal to 120 and learning rate started at 0.01, followed by cosine annealing. For optimization, SGD optimizer was used with momentum = 0.9, weight decay = 0.0001. To prevent overfitting, small random image transformations (e.g., rotation, translation, scaling) were added to the input images. The model achieved a high performance level on the ClipArt test set, achieving 0.883 for top-1 accuracy (i.e., the correct object category label being identified as the first choice of the model), and 0.973 for top-5 accuracy (i.e., the correct object category label being identified as one of the top five choices of the model). When tested on the GAT images for the visual analogy problems, the model achieved 0.833 for top-1 accuracy and 0.984 for top-5 accuracy.

Figure 3. Example images. In each row, the first image is from the GAT dataset, while the remaining images are from the ClipArt dataset. Top row: images with label "electric mixer"; bottom row: images with label "book".

Analogical Inference

We extended ResNet50 to form ResNet50-A by adding a simple computation to derive analogy predictions from the model. We input each GAT image into the neural network and extracted the penultimate feature vector (the vector immediately prior to the output layer). This vector of length 2048 was used as the representation of the image. Mathematically, this transformation can be written as: $\boldsymbol{f} = F(\boldsymbol{I}; \theta)$, where $\boldsymbol{I}$ is the input image, F is the function specified by the neural network and parametrized by θ, and $\boldsymbol{f} \in R^{2048}$ is the resulting feature vector. Thus, for each analogy question, we transfer images $\boldsymbol{I}_A, \boldsymbol{I}_B, \boldsymbol{I}_C, \boldsymbol{I}_{D1}, \boldsymbol{I}_{D2}, \boldsymbol{I}_{D3}, \boldsymbol{I}_{D4}$ into feature vectors $\boldsymbol{f}_A, \boldsymbol{f}_B, \boldsymbol{f}_C, \boldsymbol{f}_{D1}, \boldsymbol{f}_{D2}, \boldsymbol{f}_{D3}, \boldsymbol{f}_{D4}$, respectively.

A decision for an analogy problem in ResNet50-A is derived by selecting the best $D \in \{D_1, D_2, D_3, D_4\}$ such that the relation from *A* to *B* holds for *C* to *D*. To measure how similar the projection from $\boldsymbol{f}_A$ to $\boldsymbol{f}_B$ is to the projection from $\boldsymbol{f}_C$ to $\boldsymbol{f}_D$, we adopted a generic formulation based on cosine distances of the difference vectors. The same approach has been used in the Word2vec model (Zhila et al., 2013). The preferred answer $\widehat{D}$ is defined as the *D* image that generates minimum cosine distance between difference vectors:

$$\widehat{D} = \arg \min_{D \in \{D_1, D_2, D_3, D_4\}} \cos(\boldsymbol{f}_B - \boldsymbol{f}_A, \boldsymbol{f}_D - \boldsymbol{f}_C)$$

Note that this procedure for solving a visual analogy is more sophisticated than simply choosing the $\widehat{D}$ most similar to *C*, since the selection focuses on matching the visual *relation* between the *A:B* and *C:D* image pairs.

For the GAT problems, this purely visual model achieved 44% accuracy in selecting the correct *D* term. Its other choices were distributed across the three distractors (11%, 17% and 28% probabilities of choosing semantic distractors, visual distractors, and unrelated distractors, respectively). Since chance accuracy would be 25%, the purely visual analogy model achieved analogical accuracy well above chance (although well short of the level achieved by neurotypical human adults).

BART: From Verbal Semantics to Relations

The BART model (*Bayesian Analogy with Relational Transformations*) takes as inputs semantic vectors representing word meanings and uses supervised learning to acquire representations of semantic relations. The model was originally applied to learning comparatives (e.g., *larger*, *smarter*; Lu, Chen & Holyoak, 2012), but has recently been generalized to acquire an extremely wide range of semantic relations (e.g., *synonym*, *antonym*, *cause-effect*; Lu, Wu & Holyoak, 2019). For the present project, the inputs to the BART model were word embedding for individual words, each embedding consisting of 300-dimension vectors with continuous-valued features. The word embeddings were obtained by training a deep-learning model, Word2vec (Mikolov et al., 2013; Le & Mikolov, 2014) on a large text corpus (Google News). BART takes as inputs word pairs instantiating a relation, where each pair is represented by the concatenation of the Word2vec vector for each individual word. For example, a vector formed by concatenating the individual vectors for *love* and *hate* would constitute a positive example of the *antonymy* relation. The same word pair might also serve as a negative example of the *category:instance* relation.

Training Dataset

For the present project, we trained BART by combining two datasets of semantic relations. First, the SemEval-2012 Task 2 dataset (Jurgens et al., 2012) was used to teach BART the representations for 79 abstract semantic relations. This dataset is based on a taxonomy of semantic relations and includes 10 general types (e.g., *class inclusion*, *similar*, *contrast*, *cause-purpose*). The dataset includes 3215 word pairs, with 35~48 pairs for each of the 79 relations. The second dataset, developed by Popov, Hristova, and Royce (2017), includes some specific and concrete relations (e.g., the relation *constitution* with examples *brick:house*, *thread:cloth*; the relation *cover* with examples such as *house:roof*; and the relation *boundary* with examples such as *wall:room*). This dataset includes 58 specific relations drawn from ten general categories of relations. Two relations with inadequate numbers of examples were removed. The remaining 56 relations included 12~25 word pairs as examples for each relation.

Training

The BART model consists of a three-stage process to learn a broad range of semantic relations (Lu, Wu & Holyoak, 2019). In its first stage, BART exploits the heuristic that features playing similar functional roles will tend to occupy similar ranks in an ordering of differences between paired words. BART uses the difference ranking operations to generate augmented feature by partially align important features. In the second stage, BART selects a subset of important features. In the third stage, BART adopts Bayesian learning and uses the selected features of word pairs $\mathbf{f}_s$ in training examples to estimate weights distributions $\mathbf{w}$ for representing a particular relation R by applying Bayes rule as:

$$P(\mathbf{w}|\mathbf{f}_s, R) \propto P(R|\mathbf{f}_s, \mathbf{w})P(\mathbf{w}). \quad (1)$$

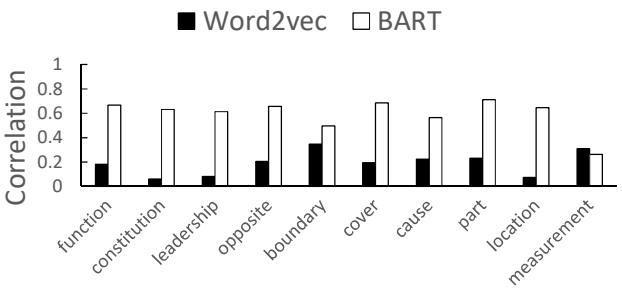


Figure 4. Model predictions of human data for relation typicality in Popov et al. (2017) dataset: Correlations between human generation frequencies and model predictions for 10 relation types for BART (after training with 10 positive examples of each relation) and for the baseline Word2vec model.

After learning, BART calculates the probability of a word pair instantiating a relation. An important aspect of both the Jurgens et al. (2012) and the Popov et al. (2017) norms is that in each set, the word pairs instantiating each relation form a typicality ordering established by human judgments. As reported in Lu et. al. (2019), BART achieved high rank-order correlations between human typicality ratings and predicted probabilities derived from the model for the abstract relations in the Jurgens et al. dataset. Across all 79 individual relations, the model's mean Spearman correlation with the human ordering was .81 (range from .65 to .91). The performance of BART considerably exceeded the mean correlation of .34 achieved using Word2vec itself as a baseline.

For the Popov et al. (2017) dataset, which includes more specific/concrete relations, BART was trained with just 10 word pairs as positive examples of each relation. As shown in Figure 4, BART achieved higher correlations with human typicality as indexed by generation frequencies (mean r = .59) than did the Word2vec model (mean r = .19).

Analogical Inference

To solve 4-term verbal analogy problems, BART forms a distributed representation of the specific relation between each word pair in a problem. BART uses its pool of learned relations to create a more refined representation of the relation(s) between two paired words. The posterior probabilities calculated for all known relations form a relation vector, with each element indicating how likely a word pair instantiates a specific relation. Hence, the result of this operation is to create a distributed representation of the relation(s) between two words, with the original semantic features being projected into a transformed space that can be used to assess relation probabilities.

For analogical reasoning, BART had available 79 relations derived by training on the Jurgens et al. (2012) norms, plus 56 relations derived by training on the Popov et al. (2017) norms.

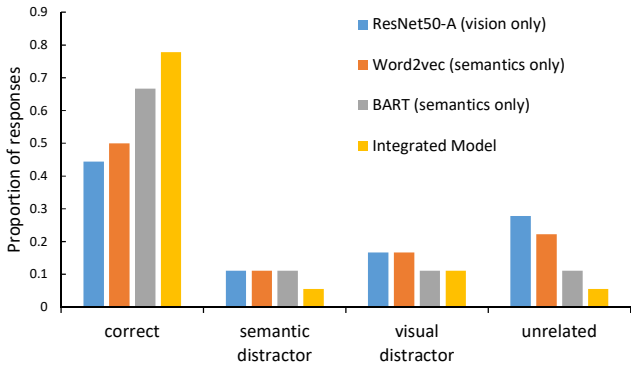


Figure 5. Proportion of responses for GAT problems for which the model's selection was the analogical option (correct), the semantic distractor, the visual distractor, and the unrelated option. Besides ResNet50-A and BART, we also report results obtained using Word2vec, and the integrated model (i.e., ResNet50-A combined with BART).

Of the latter, six relations showed weak correlations with human typicality ratings, indicating BART had failed to learn them adequately from the small number of available examples. Further examinations of the training sets for these six relations revealed that a substantial number of word pairs either included ambiguities or were otherwise questionable as instances of the relation. Accordingly, these six relations were dropped, leaving 50 relations from the Popov et al. set to be included in the relational representations, for a total of 129 learned relations.

Because BART creates relations structured by distinct roles, the model can generate the converse of any learned relation in a rule-based fashion (without additional training). For example, having learned the relation *category:instance*, BART can directly generate the converse relation *instance:category*. By applying converse formation to all trained relations, BART doubled its pool of relations, so that a total set of 258 semantic relations were available to solve GAT analogy problems.

To apply the BART model to GAT problems, the input was the verbal captions for images provided in the study by Krawczyk et al. (2008). Considered as a comprehensive model, this makes the link between ResNet50 and BART only approximate: although ResNet50 achieves high accuracy in generating the target captions, its performance is still less than perfect.

We were also faced with the problem that for many GAT images the optimal caption is a multi-word phrase (e.g., *gas pump*, *woman sewing*). To obtain semantic vectors for phrases that were not included in the Word2vec dictionary, we sometimes substituted one-word near-synonyms for which a vector was available. When that was not feasible, we used a simple averaging method, forming a vector for a phrase by averaging the vectors for its content words (cf. Kintsch, 2001).

For any pair of semantic vectors, BART uses its learned weights to calculate the posterior probability that the pair instantiates each relation in the repertoire of the model. The vector of length 258 formed by these posterior probabilities provides a distributed representation of the specific relation between the two expressions in the pair. Similarly to the procedure we followed to enable ResNet50-A to solve visual analogies, BART's preferred answer $\widehat{D}$ is that which minimizes the cosine distance between the *A:B* relation and the relation formed by *C* paired with each available option.

For the GAT problems, the BART model achieved 67% accuracy in choosing the correct *D* term; other choice probabilities were 11%, 11% and 1% to choose semantic distractors, visual distractors, and unrelated distractors, respectively (see Figure 5). To provide a baseline semantic model, the performance of Word2vec (Mikolov et al., 2013), which does not learn specific semantic relations, can be compared with the performance of BART. The Word2Vec model achieved 50% accuracy in choosing the correct *D* term; other choice probabilities were 11%, 17% and 22% to choose semantic distractors, visual distractors, and unrelated distractors, respectively.

Integration of Visual and Semantic Models

Finally, we examined the performance of an integrated model of solving pictorial analogies, formed by combining the measure of relational similarity obtained from the vision model (ResNet50-A) with the comparable measure obtained from the semantic model (BART). Two free parameters were introduced to create the integrated model.

We first transformed the vectors used by each model to put them on a common scale. The relational similarity measure from the visual model is based on difference vectors of visual features derived from the penultimate layer of ResNet50-A. These difference vectors take values in the range of -8 to 8. In contrast, the BART model forms relation vectors using posterior probabilities within the range [0 1]. To place the two vectors on a similar scale, we introduced a nonlinear transformation with an exponential function for the visual difference features v as $\exp(\alpha v)$ with a scale parameter, set at $\alpha = 2$. Cosine distances based on these transformed visual difference vectors were used to compute relational distance using the visual module:

$$D_v = \cos(\exp(\boldsymbol{\alpha}(\boldsymbol{f}_B - \boldsymbol{f}_A)), \exp(\boldsymbol{\alpha}(\boldsymbol{f}_D - \boldsymbol{f}_C))).$$

The relational similarity measure derived from the semantic module, D_s, was calculated by directly using BART as described in the preceding section. The final relational distance measure was a weighted average of the measures from the visual and semantic modules, $D = \lambda D_v + (1 - \lambda)D_s$, with the weight set as $\lambda = .3$.

Figure 5 presents a summary of the results for solving GAT analogy problems based on the visual-only model (ResNet50-A), two semantic models (Word2Vec and BART), and the integrated model based relational distance measures from both visual (ResNet50-A) and semantic (BART) models. The integrated model achieved the highest accuracy (78%) in solving GAT analogy problems; other choice probabilities were 6%, 11% and 6% to choose semantic distractors, visual distractors, and unrelated distractors, respectively.

We explored the space of parameter values, and found that performance of the integrated model was quite robust. In

general, the basic results were the same for a broad range of parameter values for α, as long as the value of λ was less than .5, so that the final decision was primarily driven by the semantic module, based on BART.

Discussion

The present paper provides a proof-of-concept that vision, language, and reasoning can be integrated to create a comprehensive computational model of how humans or machines might solve meaningful visual analogies. Here our focus has been on a vision module (ResNet50-A) that can generate verbal captions for line drawings, combined with a semantic module (BART) that takes word embeddings based on verbal captions and generates representations of semantic relations. Each model includes a decision procedure for assessing the similarity of relations between objects/words and selecting the best analogical completion from among a set of alternatives. The vision module alone achieves above-chance analogical performance on the GAT problems (picture analogies in *A:B :: C:?* format); the semantic module alone is more successful; and an integration of the two modules (biased to emphasize semantics, but also influenced directly by vision) is yet more successful, achieving 78% accuracy.

Perhaps the most surprising finding from our computational experiments is that the vision module alone was able to achieve above-chance accuracy in selecting the analogical completion, even though the critical relation is semantic/functional. Despite some shortcomings of visual deep learning models (Baker, Lu, Erlikhman & Kellman, 2018), the features in the later layers may capture parallels involving visual context (e.g., the fact that airplanes and eagles both cooccur with sky in many natural images, analogous to the fact that ships and fish both cooccur with water in natural images). Apparently, for some GAT problems, the similarity of the visual difference between the *A:B* pair to that between the *C:D* options is at least weakly correlated with the semantic relations that define the analogical answer. Moreover, the visual module continues to add useful information on top of that provided by the semantic module. Thus, vision may play two important functions in solving picture analogies: generating verbal captions that in turn feed the semantic module, and directly providing visual correlates of semantic relations.

The present project is only a first step toward the "holy grail" of a unified model connecting perception to thinking. The performance of the integrative model falls short of the high accuracy level achieved by healthy human adults not under time pressure (Krawcyzk et al., 2008). A number of incremental improvements are worth pursuing. ResNet50 might benefit from additional training on line drawings. Its accuracy in captioning might also be improved by making use of contextual information (e.g., the presence of a pumpkin as the *C* term in Figure 1 might aid in recognizing the pie). If the captioning accuracy of the visual module could be improved, its output could be directly passed to BART (rather than allowing BART direct access to optimal captions). Furthermore, future investigations need to explore how to combine visual and semantic knowledge to solve generative tasks in analogical reasoning (Chen, Lu & Holyoak, 2017).

Deeper developments would include adopting more sophisticated techniques for translating multi-word captions into semantic vectors, and eventually dealing with structured text descriptions of analogical scenes (Richland, Morrison, & Holyoak, 2006). Perhaps most intriguing is the possibility of creating hybridized visuosemantic representations that would allow perception to meld with meaning.

Acknowledgments

We thank Qi Xie and Amberly Tam for helping develop the ClipArt dataset. This research was funded by NSF grant BSC-1827374 to KH and HL, and BCS-1827427 to AY.

References

Baker, N., Lu, H, Erlikhman, G., & Kellman, P. J. (2018). Deep convolutional networks do not classify based on global object shape. *PLoS Computational Biology*, *14*(12), e1006613. https://doi.org/10.1371/journal.pcbi.1006613

Chen, D., Lu, H., & Holyoak, K. J. (2017). Generative inferences based on learned relations. *Cognitive Science*. *41*, 1062-1092.

Doumas, L. A. A., Hummel, J. E., & Sandhofer, C. M. (2008). A theory of the discovery and predication of relational concepts. *Psychological Review*, *115*(1), 1-43.

Goranson, T. E. (2002). On diagnosing Alzheimer's disease: Assessing abstract thinking and reasoning. *Dissertation Abstracts International: Section B: Sciences and Engineering*, *62*, 4785.

Farhadi, A., Hejrati, M., Sadeghi, M. A., Young, P., Rashtchian, C., Hockenmaier, J., & Forsyth, D. (2010). Every picture tells a story: Generating sentences from images. *Proceedings of the 11th European Conference on Computer Vision*, 15–29.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Jurgens, D. A., Mohammad, S. M., Turney, P. D., & Holyoak, K. J. (2012). SemEval-2012 Task 2: Measuring degrees of relational similarity. *Proceedings of the First Joint Conference on Lexical and Computational Semantics*, 356-364.

Kintsch, W. (2001). Predication. *Cognitive Science*, *25*, 173-202.

Krawczyk, D. C., Morrison, R. G., Viskontas, I., Holyoak, K. J., Chow, T. W., Mendez, M. F., Miller, B. L., & Knowlton, B. J. (2008). Distraction during relational reasoning: The role of prefrontal cortex in interference control. *Neuropsychologia*, *46*(7), 2020-2032.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, *25*, 1097–1105.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y, Li, L.-J., Shamma, D. A., Bernstein, S., & Li, F.-F. (2016). Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, *123*(1), 32-73.

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. *Proceedings of the 31st International Conference on Machine Learning*, *32*(2), 1188-1196.

Lovett, A., & Forbus, K. (2017). Modeling visual problem solving as analogical reasoning. *Psychological Review*, *124*(1), 60-90.

Lu, H., Chen, D., & Holyoak, K. J. (2012). Bayesian analogy with relational transformations. *Psychological Review*, *119*, 617-648.

Lu, H., Wu, Y. N., & Holyoak, K. (2019). Emergence of analogy from relation learning. *Proceedings of the National Academy of Sciences, USA, 116*, 4176-4181.

Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A. L., & Murphy, K. (2016). Generation and comprehension of unambiguous object descriptions. *Computer Vision and Pattern Recognition*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, *26*, 3111–3119.

Popov, V., Hristova, P., & Royce, A. (2017). The relational luring effect: Retrieval of relational information during associative recognition. *Journal of Experimental Psychology: General*, *146*(5), 722-745.

Reed, S. E., Zhang, Y., Zhang, Y., & Lee, H. (2015). Deep visual analogy-making. *Advances in Neural Information Processing Systems*, *28*, 1252-1260.

Richland, L. E., Morrison, R. G., & Holyoak, K. J. (2006). Children's development of analogical reasoning: Insights from scene analogy problems. *Journal of Experimental Child Psychology*, *94*(3), 249-271.

Sadeghi, F., Zitnick, C. L., & Farhadi, A. (2015). VISALOGY: Answering visual analogy questions. *Advances in Neural Information Processing*, *28*, 1882-1890.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *Proceedings of the International Conference on Learning Representations 2015*, 1-14.

Zhila, A., Yih, W., Meek, C., Zweig, G., & Mikolov, T. (2013). Combining heterogeneous methods for measuring relational similarity. *Proceedings of NAACL-HLT*, 1000-1009.

The Role of Effector Physicality and Risk Perception in Virtual Environments

Shulan Lu (Shulan.Lu@tamuc.edu)
Derek Harter (Derek.Harter@tamuc.edu)
Gang Wu (gangwu6@gmail.com)
Pratyush Kotturu (pratyush.kotturu@gmail.com)
Texas A&M University – Commerce
Commerce, TX 75429 USA

Abstract

Research has consistently demonstrated that people treat digital technology-based environments such as VR as if they were real. This is consistent with neural reuse and predictive processing theories. Neural circuits that have developed to perform real world actions are reused when performing tasks in computer mediated environments. The current research investigates some of the factors that could support users in leveraging their existing real world representations. A reasonable hypothesis is that users are more likely to emulate existing real world processing if technological artifacts are congruent with their experiential basis. This work investigates the perceived cues of task risks, movement realism and effector realism in performing actions. Effector design is manipulated (gesturing, wand, vs. knife), and participants cut a vegetable in a simulated environment. Participants evoked real world sensory motor contingency when technological artifacts are congruent with their experiential basis.

Keywords: embodied cognition; risk perception; computer mediated learning; danger avoidance; effector; controller

Introduction

Increasingly technological systems have begun to develop new interactive styles that leverage the richness of humans' real world interactions. For example, systems using low cost full body motion tracking, such as Kinect, have been made available. There is also a breakthrough in eye gaze based interactive system such as LC technologies' eye gaze edge tracking. Because of this departure from WIMP interfaces, a significant question arises as to whether and how gestural interactions, or in some cases intention driven touchless interactions, can evoke representations that are similar enough to perceiving and enacting actions in the real world in order to train up responses and habits that would be able to later get deployed in real world practices. If not, what differences might there be?

A myriad of theoretical approaches have been proposed to guide the design of systems that support users embodying themselves in the environment and participating in the interactions meaningfully. One of the central themes of this embodied interactive movement is to encourage the alignment between the representations being constructed for the digital world and the relevant experiential basis, making digital artifacts part of the background in the formation of representations instead of being in the foreground (Dourish, 2001; Hornecker, 2011; Ishii, 2008; Jacob, et al., 2008; Lu, Harter, Kosito & Kotturu, 2014; Slater, 2009). By judiciously re-representing the key elements in physical reality, as well as tapping into visual-perceptual cues, such digital-physical systems create a new interface interaction paradigm that leverages existing embodied proprioceptive abilities and motor skills we all develop and employ in the real world. This movement is consistent with insights from embodied and grounded cognitive science (Kirsh & David, 2013).

Recent views of embodied cognition are exploring the high level neural mechanisms that may be critical to our embodied cognitive abilities. For example, views of cognition as being hierarchical predictive machinery, where higher level layers predict activity of lower layers, and the lower layers send feedback in the form of error signals of the predictions have been proposed (Clark, 2013; Anderson, Richardson, & Chemero, 2012; Barrett & Simmons, 2015). These predicative theories suggest that more abstract concepts and higher level abilities, such as keeping track of goal states, are built up through the testing and refining of predictive mechanisms. The predictions and error signals are fundamentally bidirectional, higher levels generate predictions of the neural patterns of activity of lower layers, and mismatches generate error signals that are propagated back up the hierarchy which can be used to refine the predictive machinery.

This brain as active predictive machine view suggests that the sensory repertoire gathered from past experiences and the current sensory/perceptual inputs constrain the computation of probabilities that underlie neural representations. Such predictive views of embodied cognition are especially relevant to understanding human performance in computer mediated environments. In a computer mediated environment, we use predictive machinery that is evolved and developed to work with other (usually real world) experiences in order to interact with the digital environment (Lu & Harter, 2016).

The reuse and redeployment of neural circuits is expected (according to neural reuse theories) in order for perceptual predictions to be as efficient and accurate as possible in computer mediated environments (Anderson M. L., 2010). There are two mechanisms by which neural circuits are commonly reused, especially in the context of learning to

use a computer mediated environment for some task. In one type of reuse, new types of higher-level prediction abstractions will be created to learn predictions of low-level circuitry that is essentially being used for the purpose it was originally developed for. For example, in order to interpret visual objects being depicted in a virtual reality, they are of course designed to be visually similar to their real world counterparts. Another type of reuse is where low-level circuitry is put to a novel function by existing higher-level abstractions to cope with the differences in an unfamiliar computer mediated experience. For example, we may be experiencing a common task in a simulated environment, such as moving objects around to complete some goal, but our low-level motor actions needed in order to interact with the virtual world use some sort of input effector like a joystick rather than our own hands to perform the task.

In cognition, this predictive machinery results in a tight coupling of what is available in the environment (such as the fidelity of the environment) and the sensory motor contingency that gets triggered in a user. Central to the argument in the current work is the bi-directionality of this coupling. For example, user's movements can modify which aspects of the environment are attended to and reflexively tweak the run-time representations that are used for selecting the next action. However, previous research has been inconclusive to this prediction and the existing research paradigms are not conducive to understanding these bi-directional interactions as they unfold. In existing studies, researchers examined explicit game performance measures and player subjective reports including perceived mental workload, and did not look into real time processing measures (Freeman, et al., 2012; Reinhardt & Hurtienne, 2018). In yet another study, video clips of transitive actions were examined and participants reported the habitual actions were perceived to be easier and more natural to understand (Grandhi, Joue, & Mittelberg, 2011).

In recent work on immersive virtual reality, researchers have demonstrated the current state of the art in terms of providing tracking of handheld effectors in a typical head-mounted display (HMD) virtual reality system (Pandey, Pidlypenskyi, Yang, & Kaeser-Chen, 2018). Tracking the position of the handheld effectors is of course important in theory in order to provide an immersive experience not only of seeing the environment, but of having your body (hands and arms) embodied and perceptible within the environment. This is relevant to our current study, as it shows what may be possible in virtual reality to enable embodying hand movements and interactions. For example, the reported image-based markerless 6 degrees of freedom tracking of handheld effectors demonstrated much more reliable tracking than current virtual reality systems can achieve without additional sensors embedded in the handheld effectors. In fact, though not discussed in this article, it would seem that this method could be applied equally well to tracking the user's hands, even without holding a effector. In the research report, the authors showed that using machine learning and dual visual images, such a system can be trained to track and localize the handheld effectors with very good localization accuracy.

To what extent can users perceive the avatars in extra personal space to be their own bodies? This predictive machinery points to the importance of the visual motor correlations in embodying onto the avatar. For example, an illusory body ownership can be created over an invisible body via visual-motor synchronization (Kondo et al., 2018). While wearing a HMD, participants saw left and right white gloves and socks in front of them, at a distance of 2m, moved along in a virtual room. The visual-motor synchronicity of hands and feet were adequate to create the illusion that the moving virtual gloves and socks were part of participant's own body. This illustrates that humans are more fluid in integrating real and virtual environments than we thought previously. Are there some minimal or necessary conditions where users could blur the boundaries of real and virtual environments and perceive the actions of the avatar to be part of their extended personal space? For example, humans perceive their own mirror reflections to be part of their extra personal space. The question here is the extent to which the visual motor correlations impact users' projecting themselves into the virtual environments and act as if they themselves would be impacted by the consequence of the actions. Given that users are shut off from the real world while using the HMD, there are a number of advantages in examining the integration of real and low-cost simulated environments.

In this research, we look into real time processing measures as we manipulate an effector used by a participant in a simulated environment on a simulated task. We vary the effector to become more congruous with the real world tool they might use to do the same task. In particular, we set up a simple task to cut objects with a knife, and test certain *implicit* task measures as users perform the task but with an empty hand, vs. when holding a wand, vs. when holding a prop knife to interact with the simulated environment. Previous studies did not find realism in effectors resulted in significant differences in performance metrics, and occasionally reported some differences in subjective ratings (Freeman, et al., 2012; Reinhardt & Hurtienne, 2018). We think the explicit performance metrics and subjective evaluations of the user experiences could result from participants' strategic decision making. In the current study, we contrast the significant dimensional differences among effectors and make predictions as to whether there might be implicit differences in participants' behavioral repertoire, which is less likely to be modulated explicitly. We will explore the following hypotheses.

Physicality Hypothesis

Given that the knife and the wand in our study are matched in terms of weight and length, the physical properties of device-based effector vs. open hand are significantly different. Thus the *implicit* task performance in terms of cut

location (i.e., the cut location index) will be expected to be different between device-based effector (knife or wand) vs. open hand gesturing. We do not make predictions in terms of total time on task. It is reasonable to think that it takes longer time to cut when holding an actual physical object. However, if holding a physical object primes an awareness of the risk, then it is possible that the total time on task will be longer.

Risk Perception Hypothesis

Given that the knife is the only effector that could trigger the perception of risk (Aneli, Borghi, & Nicoletti, 2012; Brogni, Caldwell, & Slater, 2011; Liu, Cao, Chen, & Wang, 2017; Zhao, 2017), there would be significant differences in total time on task between the knife condition vs. the non-knife conditions (wand, and open-hand). Also the trajectories people take in moving the effector may differ between these, for example by being less smooth.

Method

The low-cost desktop virtual environments we developed for the experiments reported here aim to emulate a stationary work area, where the avatar puppets the motions of the user's arm in the real world, to allow the user to manipulate objects through the avatar's actions. A typical example we have implemented is a kitchen food preparation area, where the user has control of one arm of the avatar in the virtual space to manipulate knives, bowls, food and other objects. The user can have full control of the arm(s), and in more immersive versions can also control head gaze and direction.

We use the hands-free capability of the Kinect to test different conditions of physical embodiment in a vegetable cutting task, where the user has a (prop) knife versus a wand or a tracked empty hand when controlling an avatar with a virtual knife in the virtual environment. The Kinect device provides position information of the user's hand in real space, which is transmitted to the running Blender program as a set of three position coordinates. We have developed the framework to gather this positioning information reported by the Kinect, and then transmit them to a running Blender simulation. We recorded the effector position in pixels every 10 ms.

Participants

There were 53 undergraduate students recruited from a State University in the United States (Mean age = 24 years), of which 57% were female and 43% were male. In the data reported below, 2 participants' data were trimmed. Participants did not report previous exposure to such a task.

Experimental Design

We used an effector (knife condition, wand condition, vs. open hand condition) between subjects design. The weight and length of the wand were matched with those of the prop knife. Participants were randomly assigned to each of the experimental conditions.

Procedure The height of the monitor was positioned such that the location of the eyes and head of the avatar in the environment was consistent with the location of the human participant's eyes and head in the real world. The food preparation station was positioned right above the waist of a user of average height.

Once participants were successfully calibrated in Kinect, they went through a phase familiarizing themselves with Kinect. For the experimental trial, participants were given the following instructions: (1) they would see a cucumber being cut; (2) kitchen bell tone would signal their turn to make a cut; (3) make the cut where they desire. Participants saw the avatar cut the cucumber, however, they were not told where to cut exactly. As indicated in Figure 1, the length of the cucumber that remained to be cut was not significantly longer than the previous cuts made by the avatar. The idea is that how close the participant cut to the avatar's left hand fingertip would provide an indication of the extent to which participant treated the action as real (i.e., the temporary blurring the boundary of the real and situated environments). In addition, an experimenter indicated the appropriate starting position to facilitate accurate Kinect tracking.

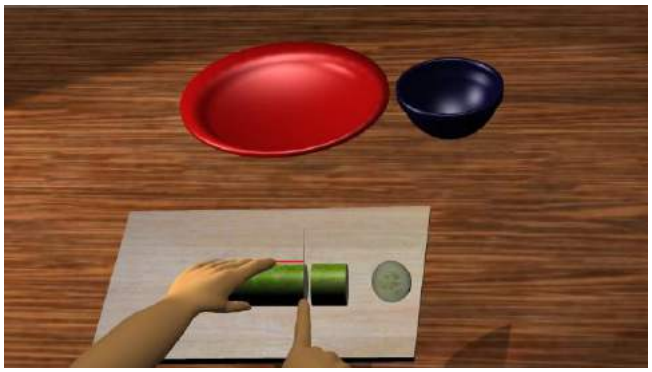


Figure 1. The cut location index is measured in blender pixel units, the distance from the avatars left hand holding the vegetable in the current task, to where the actual cut occurs indicated by the subject's actions in the experiment.

Results

We computed the following measures: (a) the cut location index, which is the distance in pixel coordinates from the participant cut locations to the finger tip of the avatar left hand in the computer mediated environment; and (b) the total task time, which was the total time from when the kitchen bell tone occurred indicating that it was the participant's turn, to when the participant moved the knife in the virtual environment in such a way that it indicated a cut should occur on the vegetable and the action to cut the vegetable was completed. In Figure 1, we depict the cut location index measure (in blender virtual environment pixel units). The measurement indicated in the figure was taken as the actual distance in 3 dimensions from the tip of the avatars fingertip, to the tip of the location where the cucumber cut location began on the vegetable being cut in

the experiment. The higher the value the cut location index is, the less risk there is to being injured.

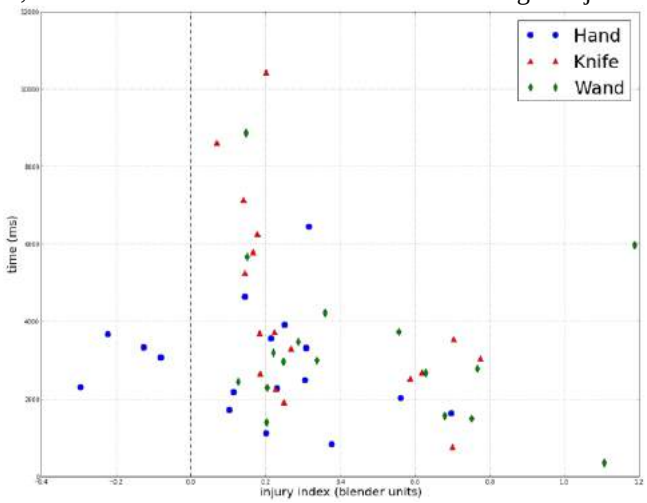


Figure 2. The cut location index scatter plot. Distance (in blender units) of the cut from the avatars hand vs. total task time when cut was made, for the 3 experimental conditions.

In Figure 2, we visualize the results of the cut location index measure together with the time to task measure. In this figure, we indicate the 3 different conditions (open hand, holding a prop knife and wand). Notice that for the knife condition especially subjects take the longest to complete the cut the closer the cut they are attempting to perform is to the avatar's left hand (which might result in potential injury, at a distance of 0.0 or less from the hand). Interestingly as well, all subjects who actually caused an injury to the hand, i.e., cuts that actually went into the avatars finger, were in the most incongruous condition, where the user in reality had an empty hand, but were controlling an avatar wielding a knife in the virtual environment. In general cuts that were more accurate and closer to the hand (without actually injuring the hand) usually took the most amount of time to make.

In Figure 3, we show the average cut location index for participants in each condition, along with 95% confidence intervals. The planned contrast showed that participants cut significantly closer to the fingertip when gesturing open hand than holding an effector, $t(48) = 2.61$, $p = 0.012$. This is consistent with the physicality hypothesis. Also of note, open hand performance on the cut location index showed the closest location index (e.g. cuts that were closest to the finger).

In Figure 4 we summarize the total task time measure. The planned contrast showed that participants used significantly longer time to cut with a knife than the other two non-knife forms, $t(48) = 2.06$, $p = 0.045$. This is consistent with the risk perception hypothesis. So while time to perform the cutting task with knives was significantly slower than the non-knife conditions. Users in the open hand condition might seem to be closest to the avatar's left hand fingertip, but they were significantly more likely to actually cause injuries to the avatar hand in this condition.

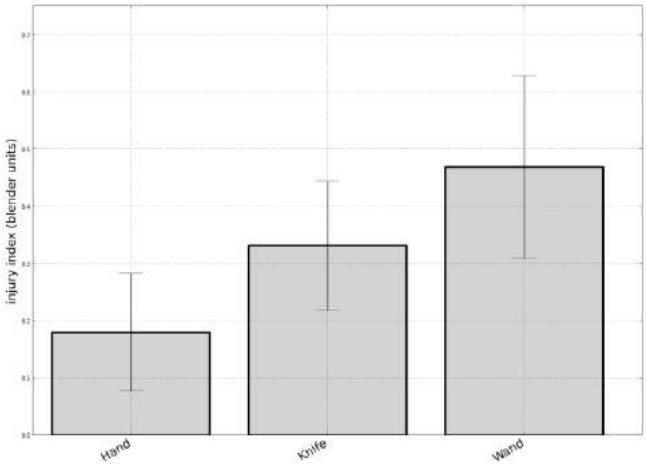


Figure 3. Summary of the cut location index measure. Mean cut locations are shown with whiskers indicating 95% confidence interval limits of the means.

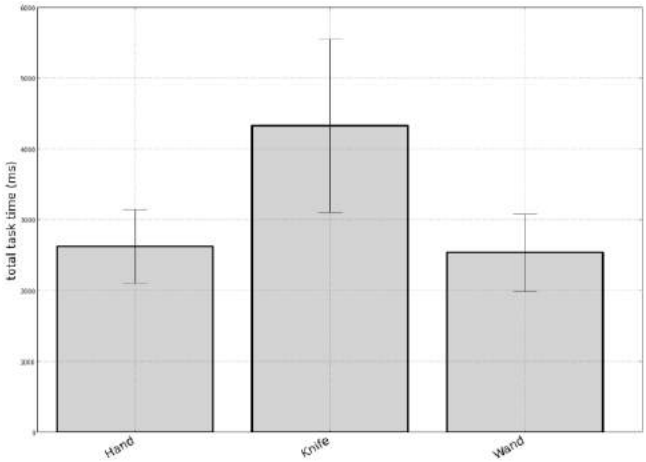


Figure 4. Summary of total task time measure. Figure indicates mean total task time for the three experimental conditions, with 95% confidence interval shown.

Conclusion

The experiment showed that effector congruence in the simulated environment does have some significant effects on task performance. Participants are more likely to treat the task as they would in the real world when the effector they use is the most realistic, and most in line with existing neural circuitry that would typically be employed to accomplish the task. For example, users were much less likely to cause injury to the virtual hand, when they were holding an object in their hand. We interpret this to mean that existing neural circuits and predictive machinery are more likely to be invoked in these more congruous conditions. Thus caution and appropriate location of the virtual knife in space were more likely to be achieved in order not to injure the virtual avatar. The most cautious behavior, in terms of time taken on the task, occurs when

holding a prop knife to manipulate the virtual environment. People using the knife took longer than people using a wand or not holding a prop.

What constitutes better performance when performing common food preparation using a knife? Speed and accuracy, as well as safety are all factors we would identify as important in separating novice level from expert level kitchen workers. Professional chefs are probably able to exceed on all three metrics, cutting quickly and accurately, but rarely if ever causing injury to themselves when using their dangerous tools.

We have done some analysis on the planning and execution of the task as indicated in the motor coordination measures of our participants in this simulated task. For example, in Figure 5 we show an analysis of the smoothness (or its absence of jerkiness) of the actual trajectories of the avatars hands in the virtual environment being controlled by the subjects hand movements through the Kinect effector. We have broken down the trajectories into 5 segments, and used the third time derivative (Hogan, 1984) to measure the smoothness of their trajectories. The whiskers represent 95% confidence intervals of the smoothness measure for each of the 5 trajectory segments. Knife performance differed significantly on this smoothness measure, especially in the middle part of the motion of the virtual knife on the task. These motion analysis measures show how different participants are treating the task when using the more congruent effector.

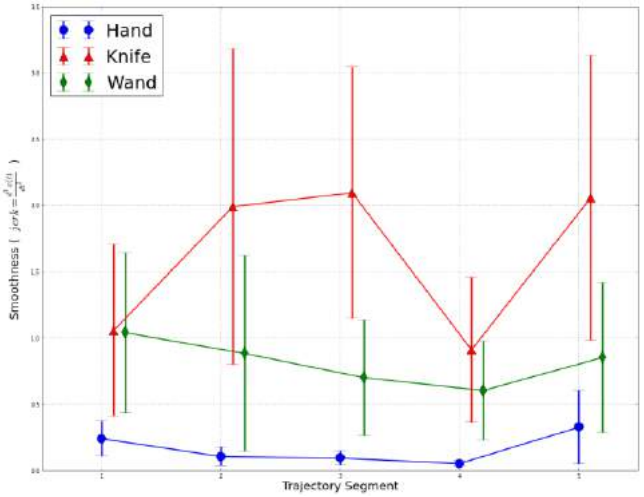


Figure 5. Measure of smoothness of subject's motion of the knife in the virtual environment, broken up into 5 equal length segments. Whiskers indicate 95% confidence interval on the smoothness measure.

Unlike the previous studies, the current results speak to the importance of looking into real time task planning and execution and showcase a paradigm in observing movements in developing embodied design (Fdili Alaoui, et al., 2015). The equivalence on the performance measures does not necessarily speak to the ongoing differences in the users' minds. Let us draw an analogy. When people use a sharp vs a dull knife to prepare food, people would take more time and be more cautious with the sharp knife, but this does not mean people would slice into their finger tips or would not be able to use the full available length of the food being prepared. A simpler view of the effector risk perception hypothesis is that users would produce different vegetable cuts while using different effectors. The implication of this simpler view is that slight differences in effector or other aspects of the environments would lead to significant differences in action outcome. Such a simplified view is in effect inconsistent with the functional redeployment theory of cognition. It is also inconsistent with the finding in virtual reality that people treat the virtual environment as if it were real even though they know it is not real (Bailenson, 2018).

The paradigm developed in the current study shows potential to examine how the bi-directional interactions of changes in simulated environments influence subsequent user actions and vice versa as they unfold in real time (Dawley & Dede, 2014). Through systematic comparison, the current study give insight into what critical ingredients of intuitive touchless interactions should involve (Chattopadhyay & Debaleena, 2015; Gillies & Marco, 2016). When the congruency between the effector in the virtual environments and the tool used in the real world tasks supports the low-level visual motor contingency, users are more likely to incorporate the extra personal space into their behavioral repertoire.

Predictive views of embodied cognition that take into account how neural circuits are likely to be reused when experiencing a simulated environment are a rich conceptual framework to better understand how to improve immersion and learning outcomes when training in simulated environments. This theory could address a number of thorny issues. For example, why environments with minimal realism can still trigger the experience of immersion and why often environments with varying degrees of realism do not get rated differently when it comes to the subjective reports of user interaction experience. A system that provides less support to align with real world sensory motor contingencies, the more perceptual prediction errors will be generated along the way and the more hierarchical adjustments will have to be made to compensate for errors. This would lead to the greater probability of errors on the task and less satisfaction as reflected in the subjective reports of user experience.

Acknowledgements

This work is supported by a grant from US National Science Foundation (IIS-0742109, 0916749). Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of US National Science Foundation. We thank Sarah Wang and Paweena Kosito in assisting various

stages of this work, in particular in the phase of data collection.

References

Anderson, M. L. (2010). Neural reuse: A fundamental organizational principle of the brain. *Behavioral and Brain Sciences, 33*(04), 245-266.

Anderson, M., Richardson, M., & Chemero, A. (2012, 10 1). Eroding the Boundaries of Cognition: Implications of Embodiments. *Topics in Cognitive Science, 4*(4), 717-730.

Aneli, F., Borghi, A., & Nicoletti, R. (2012). Grasping the pain: Motor resonance with dangerous affordances. *Consciousness and Cognition, 21*, 1627-1639.

Barrett, L., & Simmons, W. (2015, 7 28). Interoceptive predictions in the brain. *Nature Reviews Neuroscience, 16*(7), 419-429.

Bailenson, J. (n.d.). *Experience on demand : what virtual reality is, how it works, and what it can do*. New York: W.W. Norton

Brogni, A., Caldwell, D., & Slater, M. (2011). Touching sharp virtual objects produces a haptic illusion. In R. Shunmaker (Ed.), *Lecture Notes in Computer Science: Virtual and Mixed Reality* (pp. 234-242). Berlin Heidelberg: Springer-Verlag.

Chattopadhyay, D., & Debaleena. (2015). Toward Motor. *Proceedings of the 2015 International Conference on Interactive Tabletops & Surfaces - ITS '15* (pp. 445-450). New York, New York, USA: ACM Press.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 33*(4), 181-204.

Dawley, L., & Dede, C. (2014). Situated Learning in Virtual Worlds and Immersive Simulations. In L. Dawley, & C. Dede, *Handbook of Research on Educational Communications and Technology* (pp. 723-734). New York, NY: Springer New York.

Dourish, P. (2001). *Where the Action is: The Foundations of Embodied Interaction.* MIT Press.

Fdili Alaoui, S., Schiphorst, T., Cuykendall, S., Carlson, K., Studd, K., & Bradley, K. (2015). Strategies for Embodied Design. *Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition - C&C '15* (pp. 121-130). New York, New York, USA: ACM Press.

Freeman, D., Hilliges, O., Sellen, A., O'Hara, K., Izadi, S., & Wood, K. (2012). The role of physical effectors in motion video gaming. *Proceedings of the Designing Interactive Systems Conference* (pp. 701-710). ACM.

Gillies, M., & Marco. (2016). What is Movement Interaction in Virtual Reality for? *Proceedings of the 3rd International Symposium on Movement and Computing - MOCO '16* (pp. 1-4). New York, New York, USA: ACM Press.

Grandhi, S. A., Joue, G., & Mittelberg, I. (2011). Understanding naturalness and intuitiveness in gesture production: insights for touchless gestural interfaces. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 821-824). ACM.

Hogan, N. (1984). An organizing principle for a class of voluntary movements. *The Journal of Neuroscience, 4*(11), 2745-2754.

Hornecker, E. (2011). The Role of Physicality in Tangible and Embodied Interactions. *Interactions*, 19-23.

Ishii, H. (2008). Tangible bits: beyond pixels. *Proceedings of the 2nd international conference on Tangible and embedded interaction* (pp. xv-xxv). ACM.

Jacob, R. J., Girouard, A., Hirshfield, L. M., Horn, M. S., Shaer, O., Solovey, E. T., & Zigelbaum, J. (2008). Reality-based interaction: a framework for post-WIMP interfaces. *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 201-210). ACM.

Kirsh, D., & David. (2013, 3 1). Embodied cognition and the magical future of interaction design. *ACM Transactions on Computer-Human Interaction, 20*(1), 1-30.

Kondo, R., Sugimoto, M., Minamizawa, K., Hoshi, T., Inami, M., & Kitazaki, M. (2018). Illusory body ownership of an invisible body interpolated between virtual hands and feet via visual-motor synchronicity. *Scientific reports, 8*(1), 7541.

Liu, P., Cao, R., Chen, X., & Wang, Y. (2017, 6 1). Response inhibition or evaluation of danger? An event-related potential study regarding the origin of the motor interference effect from dangerous objects. *Brain Research, 1664*, 63-73.

Lu, S., & Harter, D. (2016). Toward a cognitive processing theory of player's experience of computer mediated environments. *Proceedings of CHI Play '16.* Austin, TX: ACM Sheridan Printing .

Lu, S., Harter, D., Kosito, P., & Kotturu, P. (2014). Developing low-cost training environments: How do effector and visual realism influence the perceptual grounding of actions? *Journal of Cognitive Education and Psychology*, 3-18.

Pandey, R., Pidlypenskyi, P., Yang, S., & Kaeser-Chen, C. (2018). Efficient 6-DoF Tracking of Handheld Objects from an Egocentric Viewpoint. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 416-431).

Reinhardt, D., & Hurtienne, J. (2018). The impact of tangible props on gaming performance and experience in gestural interaction. Stockholm, Sweeden: ACM.

Slater, M. (2009). Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. *Philosophical Transactions of the Royal Society B: Biological Sciences, 364*(1535), 3549-3557.

Toussaint, M. (2009). Probabilistic inferences as a model of planned behavior. *Künstliche Intelligenze, 3*, 23-29.

Zhao, L. (2017). Separate pathways for the processing of affordance of neutral and dangerous object. *Current Psychology, 36*(4), 833-839.

Representing spatial relations with fractional binding

Thomas Lu (tlu@uwaterloo.ca)
Aaron R. Voelker (arvoelke@uwaterloo.ca)
Brent Komer (bjkomer@uwaterloo.ca)
Chris Eliasmith (celiasmith@uwaterloo.ca)
Centre for Theoretical Neuroscience, University of Waterloo
Waterloo, ON, Canada, N2L 3G1

Abstract

We propose a cognitively plausible method for representing and querying spatial relationships in a neural architecture. This technique employs a fractional binding operator that captures continuous spatial information in spatial semantic pointers (SSPs). We propose a model that takes an image with several objects, parses the image into an SSP memory representation, and answers queries about the objects. We demonstrate that our model allows us to not only store and extract objects and their spatial information, but also perform queries based on location and in relation to other objects. We show that we can query images with 2, 3, and 4 objects with relative spatial locations. We also show that the model qualitatively reproduces Kosslyn's famous map experiment.

Keywords: Semantic Pointer Architecture; spatial representation; spatial memory; spatial relations; fractional binding; continuous spaces; cognitively plausible representation

Introduction

Capturing spatial reasoning has been a long-standing and difficult challenge when using artificial neural network models (Haldekar et al., 2017). Nevertheless, spatial cognition has long been studied in cognitive science (Kosslyn, 1980). Often, such research has led to proposals in which mental representations of space are continuous (Kosslyn, 1984). These representations are thus manipulated like physical images: shifting them, scanning over them, extracting spatial relations from them – effectively treating mental representations of images somewhat like physical maps. While there have been vigorous debates on the empirical adequacy of such proposals (Pylyshyn, 1973), here we explore the practicalities of implementing mental manipulations of this variety in compact and efficient representations that lend themselves to implementation in neural networks.

We approach this problem by using an architecture that deploys fractional binding to construct spatial semantic pointers (SSPs). We demonstrate that our binding architecture allows us to not only store and extract objects and their locations, but also perform mental queries to find objects based on location and in relation to other objects. It is, in particular, the ability to query such representations regarding spatial relations that we believe makes this a promising architecture for capturing many human mental image manipulation behaviors. The ability to perform such queries relies on the fact that these representations are continuous, as proposed by Kosslyn and others. The specific goal of this paper is to describe and simulate a cognitively plausible architecture that captures core qualitative features of spatial reasoning.

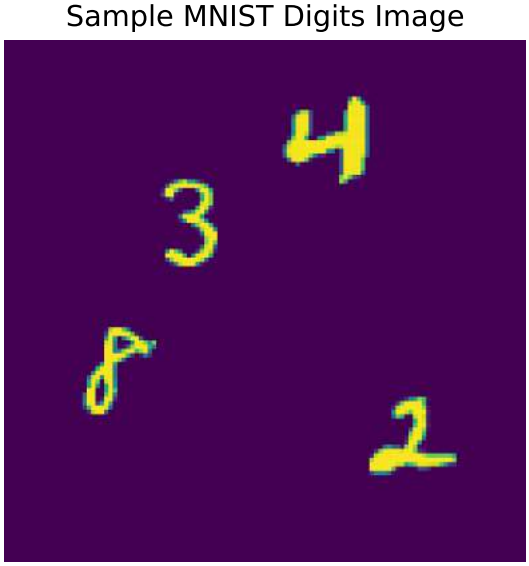


Figure 1: MNIST digits placed randomly in a 120x120 space. Given a query of: "8" and "up and right", the correct response is either: "3" or "4".

We begin by specifying our experimental design, which focuses on asking relational questions about a represented image. We then describe the spatial representation we use, discuss its properties, and note its natural affinity for implementation in spiking neural networks. After this we describe how regions are represented to allow for spatial relation queries. Then, we describe each of the elements of our architecture, as well as how they are integrated in the final system. We subsequently present results showing the accuracy of assessing spatial relations in a spatial working memory task. We also use this same representation to reproduce the main feature of Kosslyn's famous map experiment: reaction time scaling linearly with spatial distance. Finally, we discuss our findings and identify future work.

Experimental design

Our first experiment adopts a task similar to that proposed by Weiss et al. (2016). Specifically, we construct a set of example images to perform queries on by selecting batches of digits from the MNIST database and placing them at random locations on a 120x120 image. We choose between 2 and 4

digits to include in a given image. We then generate queries automatically by randomly selecting a target digit and computing its relative direction from another randomly selected query digit. For this experiment, we limited the query direction to 4 possible quadrants: up and left, up and right, down and left, down and right. Given the query digit and direction, we expect the response to be one of the target digits (if there are multiple such digits, then either one is marked correct). For instance, in Figure 1 we show an example randomly generated image, for which we might query "What is up and to the right of the 8?" A response of either "3" or "4" would be marked correct.

For this experiment, we normalized the coordinates of the digits in the 120×120 pixel image to a continuous 10x10 space, specifically the intervals $x \in [-5,5]$ and $y \in [-5,5]$, before encoding them in a memory through our model visual system. Given our chosen representation, this range was found to provide a good trade-off between accuracy and precision.

We also performed a second experiment, similar to the visual-spatial map experiment by Kosslyn et al. (1978). Kosslyn's map experiment recorded the time that it takes for a subject to scan from one location to another in memory, and demonstrated that closer objects are typically reached faster. For our experiment, we used a memory of several digits placed randomly, and scanned from a queried starting object to the queried ending object.

Methods

Spatial representation

We employ the method for spatial representation proposed by Komer et al. (2019). This method generalizes the notion of binding that is employed by several vector symbolic architectures (VSAs) to continuous spaces. The method defines a "spatial semantic pointer" (SSP) to be the result of a fractional binding. The particular binding used is the circular convolution operator proposed by Plate (1995), which is essentially element-wise multiplication of vectors in Fourier space. The natural extension of this is then element-wise exponentiation in Fourier space. Supposing B is a fixed d-dimensional vector (i.e., semantic pointer), fractional binding is defined by expressing the binding in the complex domain:

$$B^k = \mathcal{F}^{-1}\left\{\mathcal{F}\{B\}^k\right\}, \quad k \in \mathbb{R}, \tag{1}$$

where $\mathcal{F}\{\cdot\}$ is the Fourier transform, and $\mathcal{F}\{B\}^k$ is an element-wise exponentiation of a complex vector—analogous to exponentiation using fractional powers (e.g., $b^{2.5}$)—permitting k to be real. This representation can thus map from a continuous space, $\mathbb{R}$, to a high-dimensional vector space, $\mathbb{R}^d$. Because the high-dimensional space of semantic pointers can support construction of cognitive structures, various kinds of syntactic inference, and so on (Eliasmith, 2013), this proposed representation provides a novel link between such cognitive operations and continuous spaces.

To explore this link, in this work we use a generalization of the representation to multiple dimensions (Komer et al., 2019). We can represent points in $\mathbb{R}^n$ by repeating equation 1, n times, using a different semantic pointer for each represented dimension (i.e., for each axis), and then binding all of the resulting vectors together. For $n = 2$ (i.e., for a 2-D spatial map), the SSP that represents the point (x,y) is defined as the vector resulting from the function:

$$S(x,y) = X^x \circledast Y^y, \tag{2}$$

where X and Y are fixed semantic pointers, x and y are reals, and we are using fractional binding as defined by equation 1.

In this work we explore querying spatial relations between multiple objects in memory – for instance, asking "What is below and left of the 3?" To specify the spatial query, we represent the region of space being queried (e.g., below and left) as another SSP. The SSP that represents a continuous region (e.g., a solid rectangle), specified by some infinite set of points R, is defined as:

$$S(R) = \int_{(x,y)\in R} X^x \circledast Y^y \, dx\, dy. \tag{3}$$

To move this region to be relative to a given starting point, we exploit the shift property of SSPs. In particular,

$$B^{k_1} \circledast B^{k_2} = B^{k_1+k_2}, \quad k_1, k_2 \in \mathbb{R}. \tag{4}$$

This means that to shift any SSP, we only need to convolve the spatial representation of a region or objects with the SSP representing the coordinates of the shift direction. For example, we can shift a region representing a direction, (e.g., "up and right") to the location of an object to generate a region representing a query (e.g., the "8" in the previous example).

Conversely, we can also leverage this property to shift the entire spatial memory relative to the origin. This gives rise to a notion of movement through the space and an egocentric interpretation of the space rather than the previous allocentric interpretation. Thus this method of semantic pointer supports both egocentric and allocentric coding of space.

To represent a single object occupying some location, we bind its tag (OBJ) with the SSP from equation 2:

$$M = OBJ \circledast S(x,y). \tag{5}$$

In general, to represent a set of m labelled objects together in the same memory, we can use superposition:

$$M = \sum_{i=1}^{m} OBJ_i \circledast S(x_i, y_i), \tag{6}$$

with a distinct semantic pointer OBJ_i tagging each object.

Furthermore, rather than placing objects at singular points in memory, it may be more intuitive to bind objects to regions in memory. This can be done similarly:

$$M = \sum_{i=1}^{m} OBJ_i \circledast S(R_i), \tag{7}$$

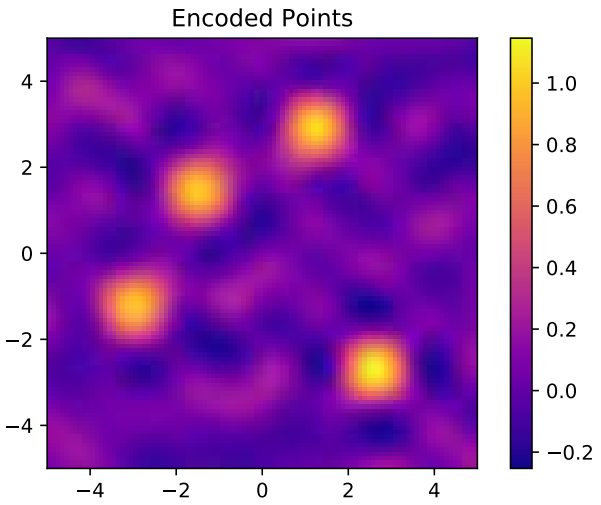


Figure 2: Heatmap of the four locations from Figure 1, represented by a single spatial semantic pointer (equation 6).

with R_i representing the region that a particular object occupies. This representation allows us to represent notions of size and shape in memory as well.

Given a representation like that in equation 6 or 7, we can query it in a number of ways. For example, to determine what object(s) are within some region R, we can compute:

$$M \circledast S(R)^{-1}, \tag{8}$$

where $(\cdot)^{-1}$ corresponds to the approximate inverse vector used to unbind using circular convolution. By the properties of binding and superposition, the resulting vector will have the highest cosine similarity (i.e., dot product) with the object(s) within R.

While only part of our architecture is currently implemented in a neural network (see below), all of the operations, except fractional binding, needed for the architecture have previously been implemented in spiking neural network models (Eliasmith, 2013). The fractional binding itself is implemented in spiking neurons by Komer et al. (2019). These implementations use the methods of the Neural Engineering Framework (Eliasmith & Anderson, 2003).

Using the spatial representation

In this section we briefly demonstrate the use of equations 3, 4, and 6. All of the SSP representations in the model are 512-dimensional. We begin by encoding multiple objects into the memory, as per equation 6, is demonstrated in Figure 2. Here we can see an example of the four objects from Figure 1 being encoded into the represented space. While we are showing a decoding of this representation mapped into the continuous space, the full representation is a single 512-dimensional vector. The number of objects in memory does not change the size of the representing vector, although there are effective limits on capacity (Komer et al., 2019). We also tested a region based system by binding every digit to the square region occupied by the digit rather than just a single point.

To query such a representation, we can construct a region vector. A region vector, as defined by equation 3, is also a 512-dimensional SSP, but it represents an entire region instead of a specific point. Region vectors can be used just like a regular location vector. We can bind objects to it, add it to a memory, and we can also use it to extract objects that are located within a region. Furthermore, as regions are integrals over pointers raised to coordinate exponents, binding a region to a point vector shifts the exponents in the integral by the coordinates of the point (see equation 4), which in turn shifts the entire region represented by the integral (see Figure 3-Top) in the direction of the point relative to the origin. In our experiment, this allows us to pre-compute four regions at the origin and then use binding to shift them to generate any specific query vector (see Figure 3-Bottom). Notably, when region vectors are used to query memories encoding objects at those locations, there is no need to extract the coordinates of the objects being searched over; all computations are performed within the space of our SSPs, without multiple encoding and decoding steps.

Model architecture

In this section we briefly describe each of the components in our model that perform the tasks described in the experimental design section. We also describe the integration of the components and overall flow of information through the model.

Image generation

The images processed by the system are generated by using batches of 28x28 pixel images from the MNIST database and placing them randomly on a 120x120 image (see Figure 1). Because queries are limited to the 4 diagonal directions, we ensured the digits are not too close in the vertical or horizontal direction. We also ensured the digits do not overlap. We generated sets of 5,000 samples for images containing each of 2, 3, and 4 digits.

MNIST Network

In order to generate the SSP representation from the experiment images, we use a straightforward convolutional deep neural network as a perceptual module. It consists of two 3-by-3 convolution layers with 32 and 64 filters respectively. These were followed by a 128 unit fully connected dense layer and a 10 unit fully-connected dense layer for classification. This network was trained on the MNIST dataset achieving 99% validation accuracy.

Since our work focuses on representing spatial relationships rather than classifying multi-digit MNIST images, we use the actual coordinates of the digits to generate a saccade-like cropping of the full image to 28x28 sub-images before providing them to the convolutional network. The identified images are then mapped to random 512-dimensional semantic pointers, which are bound to SSP encoded locations, and

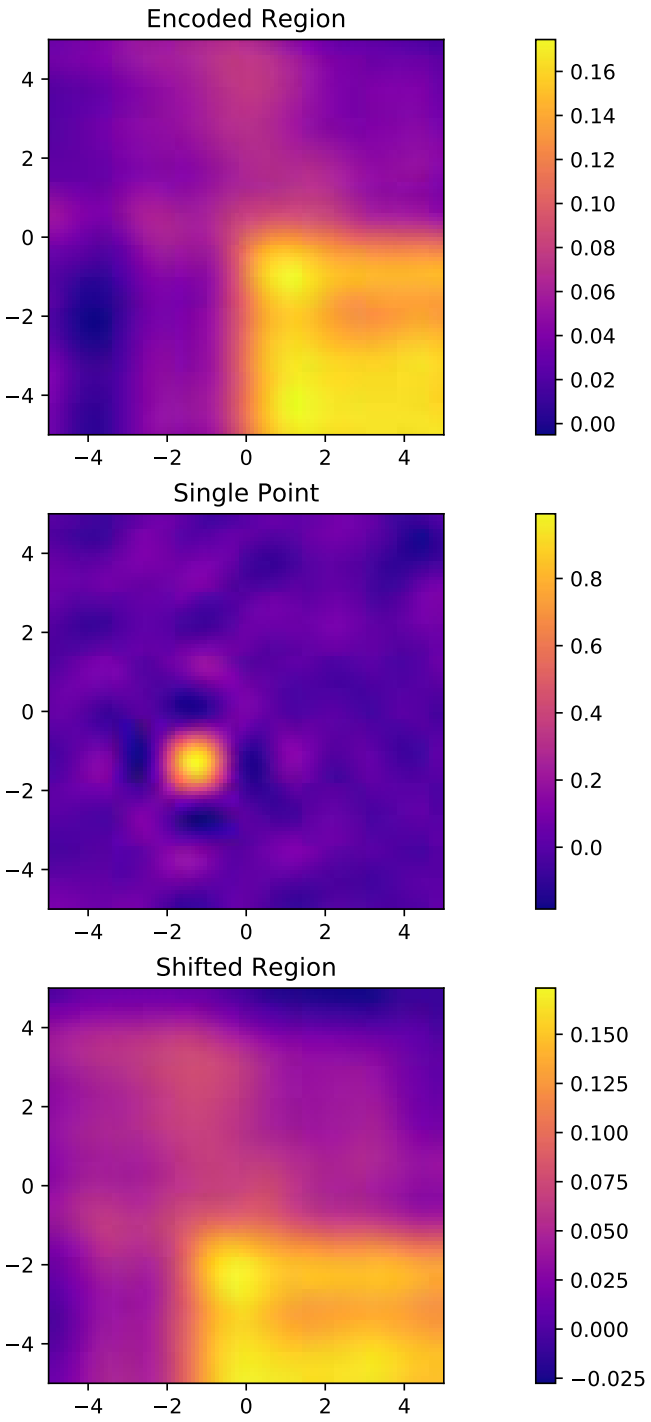


Figure 3: Demonstration of shifting the region representation for a "down and right" query (top panel), to a point encoded as an SSP (middle panel), resulting in a region vector for querying "down and right" with respect to the point (bottom panel).

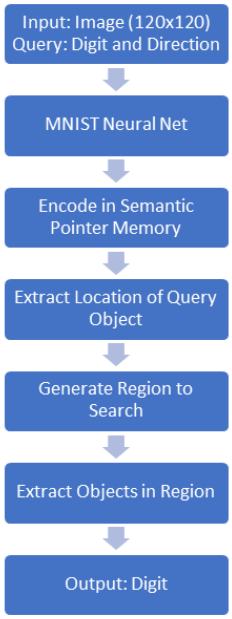


Figure 4: Flowchart of entire process

summed across all objects. This process results in a memory representation in the 512-dimensional space, which is subsequently queried using a region representation as described above.

Cleanup memory

When using SSP representations, as with any compressed VSA, the extracted location vector of an object includes noise. When there are multiple objects in the memory, the amount of noise grows. As a result, VSAs of this sort typically include a cleanup memory that maps a noisy vector onto the nearest known vector in the space. In the case of a continuous space, to extract the (x, y) location of the query digit, we generate the known vectors by sampling the continuous space on a 100x100 grid. This grid covers the two $[-5, 5]$ axes of the image. To implement the memory, we perform a simple dot product similarity check between the extracted noisy vector and the set of known vectors to find the closest matching vector within the resolution of our grid. Dot products are easily computed in parallel, making this a quick and effective way to reduce noise and improve performance. This kind of memory can be efficiently implemented in spiking neurons (Stewart et al., 2011).

Full system

Before running an experiment, we set up the model by randomly selecting two 512-dimensional unitary semantic pointers to use as axis vectors (i.e., X and Y in 2). We also create a vocabulary of ten 512-dimensional semantic pointers, one for each digit. We then pre-compute 10x10 region vectors for each query, as well as a 100x100 resolution cleanup memory.

We then feed an image into the model architecture, the full pipeline of which is depicted in Figure 4. The image is clas-

sifed by the MNIST network, and an object memory is created by summing the SSP representations for each digit in the image, as described above. Extraction of the location of the query object (i.e., the object mentioned in the query) proceeds by performing an inverse convolution on the memory with the query object to find its location, and the cleanup memory is used to reduce noise on the found location. Generating the region to search is accomplished by convolving the identifed location of the query object with the region vector corresponding to the query direction to find the region where the target object might be. Extracting objects in the region occurs by performing an inverse convolution between the shifted region and the original memory. Finally, the similarity between the results of this query and each object in the vocabulary is calculated as a dot product. The object with the highest similarity determines the model's response to the query.

Scanning System

The scanning system involves similar steps. We reuse the axis vectors as well as the pre-computed cleanup memory tables from the previous system. The map image is converted to a memory vector as above. Given a starting and ending object, the locations are extracted by performing inverse convolution with the objects in question on the memory. These locations are cleaned with the cleanup memory and used to determine the direction vector of the scanning using SSPs:

$$V = (X^{x_5} \circledast Y^{y_5}) \circledast (X^{x_2} \circledast Y^{y_2})^{-1} \tag{9}$$

where x_5 is the x position of the "5", and so on. We then normalize this vector, shrinking it to a 0.05 unit step, and repeatedly apply it to the starting vector ($V_{t+1} = V_t \circledast V$ where $V_0 = X^{x_5} \circledast Y^{y_5}$).

To scan the memory, we started at the starting location from above, and extracted the objects in that location with inverse convolution. The scan location is then updated by convolution with the step vector generated above, shifting the location towards the target object, and the above steps are repeated. A dot product similarity comparison is used at each step to determine what objects were extracted or "seen" by the scan. A 0.8 similarity threshold is used to determine when the target object has been reached.

Results

Relational Query Experiment

For the query experiment, we ran 5,000 randomly generated experiments for each of 2, 3, and 4 digit images. For the experiment, we tested the accuracy of the output by simply marking the response as correct if the model response matched an object in the queried region.

Table 1 shows the results from the experiment involving identifying a target digit given an image, a query object, and a query. Correctness is calculated by comparing the output to all digits in the correct direction. Baseline performance is the probability of answering a query correctly by randomly selecting one of the remaining digits in the image. This is calculated by dividing the average number of correct answers in each image by one less than the total number of digits in the image. Naturally for the 2 digit case, there is only one possible answer other than the digit used to query so the probability would be 100%. The baseline probability is very high due to the broadness of our query.

	2 Digits	3 Digits	4 Digits
Point Representation Accuracy	92.18	84.40	72.90
Region Representation Accuracy	95.98	87.22	81.24
Baseline probability	100.00	71.76	62.60

Table 1: Experimental results for spatial relation queries.

The results from the 2 object query indicate that using a region vector decreases accuracy compared to a simple location query. A location query with two objects in memory (e.g., what is at location (x,y)) has 100% accuracy (results not shown). In this experiment, the 2 digit case is similar to a location query, but for a region. The drop in accuracy is likely because as the region representation becomes larger, a single vector is being used to represent the effective superposition (integral) of many vectors (all those defining the region). This result suggests that region size will determine decoding accuracy, a hypothesis to test in future work.

The 3 and 4 digit experiments showed that extracting object information from an object-location memory improved performance by about 13% and 10% for point based memory and 15% and 19% for region based memory compared to the baseline guessing probability. Representing object locations as regions in memory rather than singular points provided dramatic improvements in accuracy, particularly in the 4 digit case. This is likely due to the fact that a query region is more similar to a square within the region than a single point, leading to higher accuracy extractions with inverse convolution. This suggests that more specific queries involving smaller or tighter regions would yield higher accuracy as their shapes would more closely resemble the regions the objects are bound to compared to the large query regions used in our experiment. Comparing the differences in accuracy for queries of different shapes and sizes is a topic for future study.

The decrease in accuracy as number of digits in the image increases is expected, as a higher number of digits adds difficulty in selecting the correct output since the memory encodes all of the digits. It is a standard property of VSAs for decodability to decrease as a function of the number of objects represented in a structure. While we have not determined the maximum capacity of the proposed representation, being able to store and reasonably accurately recall the relations between four numbers is consistent with standard estimates of working memory capacity at 4 items (Buschman et al., 2011).

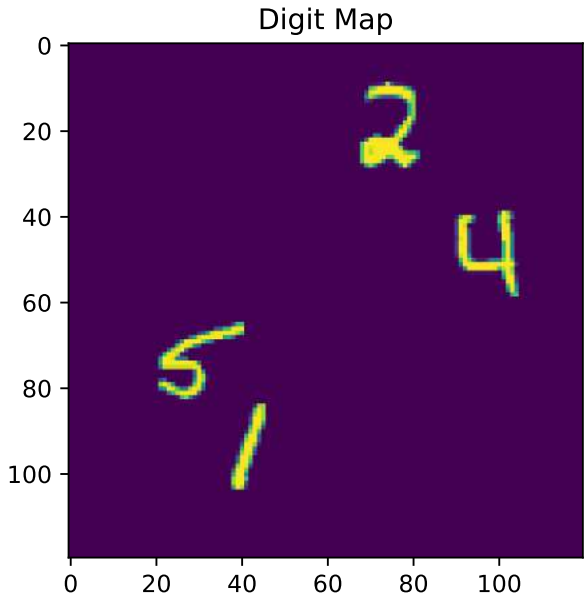


Figure 5: Digits placed randomly in a 120x120 space to represent a map of objects. The memory is scanned from "5" to "1" and "5" to "2".

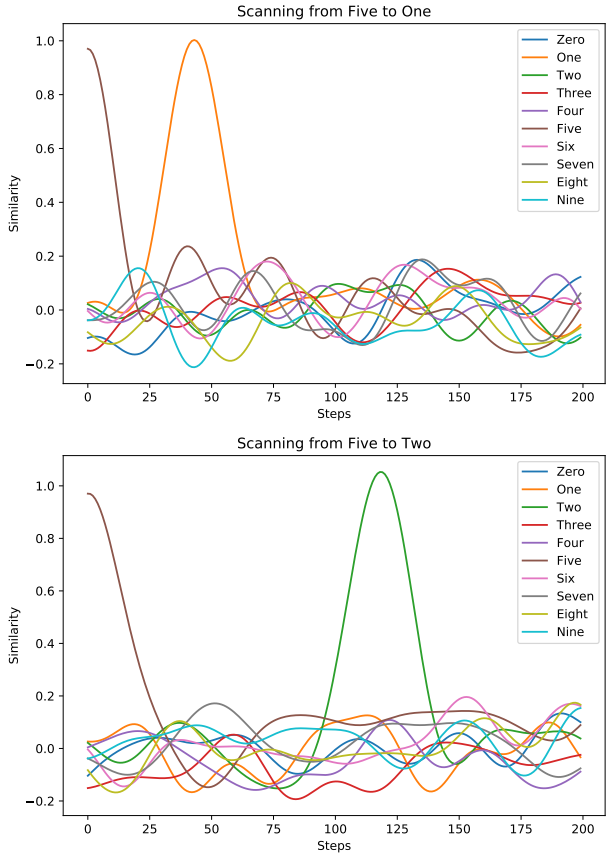


Figure 6: Similarity outputs over time for scanning from the "5" to "1" (top) or "5" to "2" (bottom).

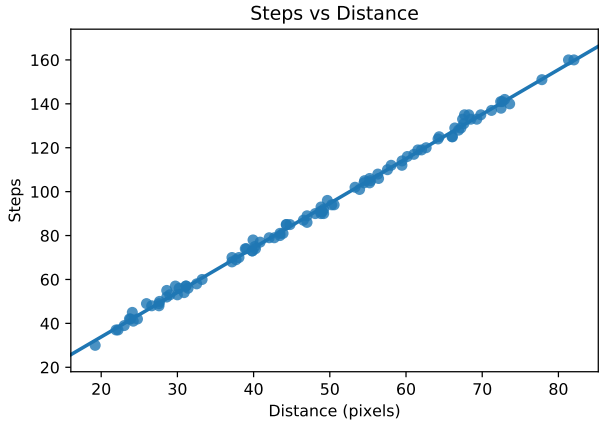


Figure 7: Plot of steps to reach target object vs the distance between starting and target objects over 100 trials (Pearson's $R > 0.99$, p-value $< 10^{-6}$).

Image Scanning

An image is generated with the same method as in the first experiment to represent a map of objects, with each digit representing an arbitrary object in the map (Figure 5). For the experiment shown in Figure 6 we chose the object "5" as the starting location and the two objects "1" and "2" to be the near and far target objects respectively. From the two plots, we can see a peak at the 0 mark for the starting object "5" which falls away, and a peak at the target object, "1" and "2", when the scan reaches it.

This experiment was repeated 100 times, with the number of steps required to reach a similarity threshold of at least 0.8 recorded for each trial (Figure 7).

Kosslyn et al. (1978) showed that human spatial memory is represented in a metric space by demonstrating that further objects take longer to scan to in memory, with time linearly related to distance. This experiment shows that the qualitative cognitive behaviour demonstrated by Kosslyn's map scanning experiment is naturally captured by our SSP memory representation.

Discussion

Our proposed architecture is able to receive an image of multiple objects and generate an SSP representation. Subsequently given a spatial relation query the model can successfully answer with reasonably high accuracy. This provides evidence that the SSP representation can be used to encode continuous spaces in a kind of mental map using representations easily implementable in neural networks. In short, our results show that such representations can be used to reproduce qualitative cognitive behavior relying on spatial manipulation of information encoded in this manner.

A critical next step is to compare human performance on this same task with the proposed model. Preliminary results suggest that accuracy can be manipulated by appropriately choosing the base vectors (i.e., X and Y), and manipulating

the dimensionality of the vector space being used. The range of these parameters that match human performance remains to be determined.

There are many possibilities for extending this model. Our particular focus was on two kinds of spatial relation query. However, the direction queries could be generalized to be in any direction (e.g., specifying a vector direction and generating a cone region in that direction). As well, other manipulations, such as spatial rotations, shifts, and so on, can be performed without decoding the SSP. There are a wide variety of psychological results that can provide points of comparison for such manipulations.

Furthermore, the representation itself could be made more complex. For instance, introducing the color of the object (encodable as another 3D continuous space for RGB values), or additional features is natural in this framework. We expect additional information encoded in the memory will adversely affect performance, as seen in human memory tasks.

Finally, the full model can be implemented in a spiking neural network to determine if the proposed representations are robust to biologically plausible implementation. While we expect that this will be successful, given past work that has implemented each of the components, it remains to be seen what effect such implementation has on the accuracy of responding to spatial queries.

Conclusions

We have demonstrated that spatial semantic pointers (SSPs) using fractional binding provide a viable method of representing spatial relationships in a simple model supporting two kinds of visual spatial reasoning. This method lends itself well to implementation in neural networks, and is consistent with cognitive work suggesting that internal representations used in mental imagery represent continuous mental spaces. We believe this is one of few available suggestions for how complex object representations (i.e., high-dimensional feature vectors for digits) can be encoded in a continuous space, and manipulated to answer questions about relations in that space.

Acknowledgments

We would like to thank Terry Stewart for his early work exploring SSPs and personal discussions, and Jan Gosmann for his work on the mathematics of fractional binding for semantic pointers. This work was supported by CFI and OIT infrastructure funding, the Canada Research Chairs program, NSERC Discovery grant 261453, ONR grant N000141310419, AFOSR grant FA8655-13-1-3084, OGS, and NSERC CGS-D.

References

Buschman, T. J., Siegel, M., Roy, J. E., & Miller, E. K. (2011). Neural substrates of cognitive capacity limitations. *Proceedings of the National Academy of Sciences*. Retrieved from https://www.pnas.org/content/early/2011/06/13/1104666108 doi: 10.1073/pnas.1104666108

Eliasmith, C. (2013). *How to build a brain: A neural architecture for biological cognition*. Oxford University Press.

Eliasmith, C., & Anderson, C. H. (2003). *Neural engineering: Computation, representation, and dynamics in neurobiological systems*. MIT press.

Haldekar, M., Ganesan, A., & Oates, T. (2017, June). Identifying Spatial Relations in Images using Convolutional Neural Networks. *arXiv e-prints*, arXiv:1706.04215.

Komer, B., Stewart, T. C., Voelker, A. R., & Eliasmith, C. (2019). A neural representation of continuous space using fractional binding. *Proceedings of the 41st Annual Meeting of the Cognitive Science Society*.

Kosslyn, S. M. (1980). *Image and mind*. Harvard University Press.

Kosslyn, S. M. (1984). *Image and brain*. MIT Press.

Kosslyn, S. M., Ball, T. M., & Reiser, B. J. (1978, 2 1). Visual images preserve metric spatial information: Evidence from studies of image scanning. *Journal of Experimental Psychology: Human Perception and Performance*, *4*(1), 47–60. doi: 10.1037/0096-1523.4.1.47

Plate, T. A. (1995). Holographic reduced representations. *IEEE Transactions on Neural networks*, *6*(3), 623–641.

Pylyshyn, Z. W. (1973). What the mind's eye tells the mind's brain: A critique of mental imagery. *Psychology Bulletin*, *80*, 1–24.

Stewart, T. C., Tang, Y., & Eliasmith, C. (2011). A biologically realistic cleanup memory: Autoassociation in spiking neurons. *Cognitive Systems Research*, *12*, 84-92. Retrieved from http://dx.doi.org/10.1016/j.cogsys.2010.06.006 doi: 10.1016/j.cogsys.2010.06.006

Weiss, E., Cheung, B., & Olshausen, B. (2016). A neural architecture for representing and reasoning about spatial relationships. *Proceedings of the International Conference on Learning Representations (ICLR), Workshop Trak,.*

Statistical learning creates implicit subadditive predictions

Yu Luo (yuluo@psych.ubc.ca)
Department of Psychology, University of British Columbia

Jiaying Zhao (jiayingz@psych.ubc.ca)
Department of Psychology, and Institute for Resources, Environment and Sustainability, University of British Columbia

Abstract

The cognitive system readily learns when multiple cues jointly predict a specific outcome. What is less known is how the mind generates predictions when only a single cue is present. In four experiments, participants were first exposed to two objects followed by a circle with a specific size or a specific numeric value. Afterwards, participants viewed a single object and estimated the associated size or value. Finally, participants recalled the size or value that followed the initial two objects. We found that the estimated size associated with the single object was significantly smaller than 100% but significantly larger than 50% of the recalled size associated with the two objects. No participants were consciously aware of the associations. The results reveal a new consequence of statistical learning on automatic inferences: When multiple objects were previously associated with an outcome, the single object is implicitly expected to predict a subadditive outcome.

Keywords: Implicit learning; support theory; subadditive inferences; regularities; predictions

Introduction

A remarkable capacity of the cognitive system is to extract the relationships among objects in the environment. Statistical learning is one mechanism that detects the statistical relationships between individual objects in terms of co-occurrences over space or time (Fiser & Aslin, 2001; Saffran, Aslin, & Newport, 1996). In contrast to other forms of associative learning, statistical learning occurs incidentally, without conscious intent or explicit awareness, and thus observers are often not explicitly aware of object co-occurrences (Turk Browne, Jungé, & Scholl, 2005; Turk-Browne, Scholl, Chun & Johnson, 2009).

The ability to extract statistical regularities from the environment has a series of cognitive consequences. For example, statistical learning encodes the co-occurring objects more efficiently in working memory (Brady, Konkle, & Alvarez, 2009; Zhao & Yu, 2016), draws attention spontaneously and persistently to the co-occurring objects (Yu & Zhao, 2015; Zhao, Al-Aidroos, & Turk-Browne, 2013; Zhao & Luo, 2017), forms new transitive inferences based on prior associations (Luo & Zhao, 2018), enhances memory representation of individual objects (Kim, Lewis-Peacock, Norman, & Turk-Browne, 2014; Otsuka & Saiki, 2016), and induces false memories of co-occurring objects (Luo & Zhao, 2017).

Past research on statistical learning has predominately focused on associations between individual objects that co-occur in space or time (e.g., A appears next to or before B). Moreover, most studies in associative learning focused on how the relationship between the cue and the outcome is learned, how learning modulates subsequent processes, and how predictive cues are selectively prioritized (e.g., Mackintosh, 1975; Le Pelley et al., 2016).

In the daily visual environment, multiple objects sometimes co-occur to jointly predict a specific outcome. For example, two co-authors often publish a paper together, or two co-founders start a company. What is less known is how the mind generates predictions when only a single cue is present, after learning that two cues were previously jointly associated with an outcome. For example, when author A and author B have been publishing high-quality papers together, what's the automatic inference when you see a paper by only author A?

Here we examine three possible hypotheses: (1) the complete inheritance hypothesis that suggests that the single cue predicts 100% of the outcome previously associated with the two cues, (2) the proportional inheritance hypothesis that suggests that the single cue predicts 50% of the outcome, and (3) the subadditive hypothesis that suggests that the single cue predicts more than 50% but less than 100% of the outcome previously associated with the two cues. The subadditive hypothesis is consistent with support theory (Tversky & Koehler, 1994), that suggests that when people unpack an event (e.g., the probability of death due to natural causes) into disjoint components (e.g., the probability of death due to heart attack, cancer, or other natural causes), they tend to increase the evidentiary support for the event. In other words, people tend to provide a higher probability of death due to natural causes when they are asked to estimate the probability of death due to each component of natural causes separately, compare to reporting the probability of death due to natural causes as one category.

To test these hypotheses, we conducted a series of four experiments to examine how the mind makes predictions when a single cue is present after learning that multiple cues previously jointly predicted an outcome.

Experiment 1

In this experiment, participants were first exposed to two cues (e.g., red and blue squares) that were immediately followed by an outcome (e.g., a circle with a specific size). We examined how they generated predictions of the outcome when only a single cue was present (e.g., a red square).

Participants

A total of 42 undergraduates (31 female; mean age=19.6 years, SD=1.5) from University of British Columbia (UBC) participated in the experiment for course credit. Participants reported normal or corrected-to-normal visual acuity and provided informed consent. The protocol was approved by the UBC Behavioral Research Ethics Board.

Stimuli

The stimuli consisted of eight squares in eight distinct colors (color name = R/G/B values: red = 255/0/0; green = 0/255/0; blue = 0/0/255; yellow = 255/255/0; magenta = 255/0/255; cyan = 0/255/255; orange = 255/158/0; brown = 103/29/0). Each square subtended 2.7° of visual angle. The colored squares were randomly assigned into four pairs for each participant and remained constant throughout the experiment. Each color pair was randomly associated with a gray circle (R/G/B = 128/128/128) with a specific diameter. The circle diameter subtended 3.0° (or 100 pixels), 6.0° (or 200 pixels), 9.0° (or 300 pixels), or 12.0° (or 400 pixels) of visual angle (Fig.1a). Thus, each color pair was associated with a circle of a specific size.

Apparatus

Participants in all experiments were seated 50cm from a computer monitor (refresh rate = 60 Hz). Stimuli were presented using MATLAB and PsychophysicsToolbox (http://psychtoolbox.org).

Procedure

The experiment consisted of three phases: exposure phase, inference phase, and recall phase. During exposure, two colored squares (e.g., red and blue squares) appeared in a horizontal configuration at the center of the screen for 500ms, followed by a 500ms inter-stimulus interval (ISI), and then the circle with a rotated T in the middle appeared at the center of the screen for 500ms in each trial (Fig.1b). Each color-size pair was repeated 80 times to form a single continuous temporal sequence of color-size pairs in a pseudorandom order with a constraint where no single color-size pair could repeat back-to-back. In total, there were 320 trials. Participants performed a cover task where they judged as quickly and accurately as possible whether the rotated T in the circle was pointing to the left or right (by pressing the "1" or "0" key for left or right, respectively). The cover task was irrelevant to learning the color-size pairs, in order to conceal the true purpose of the study. This also ensured that statistical learning of the color-size pairs was incidental. Participants were not told anything about the color-size pairs.

After exposure, participants performed an inference phase (Fig.1c). In each trial, participants viewed a single color square for 500ms followed by a 3000ms blank screen. Afterwards, a probe circle with a diameter subtending 0.6° (or 20 pixels) was presented on the screen. Participants were asked to estimate the size of the circle that was associated with the color square by adjusting the size of the probe circle using their mouse. The diameter of the adjustable circle was restricted to a range from 20 pixels to 420 pixels. The adjustable circle remained on the screen until the "a" key was pressed to register participant's estimate. Each member of a color pair was tested four times, resulting in 32 trials in total (the order of the trials was randomized).

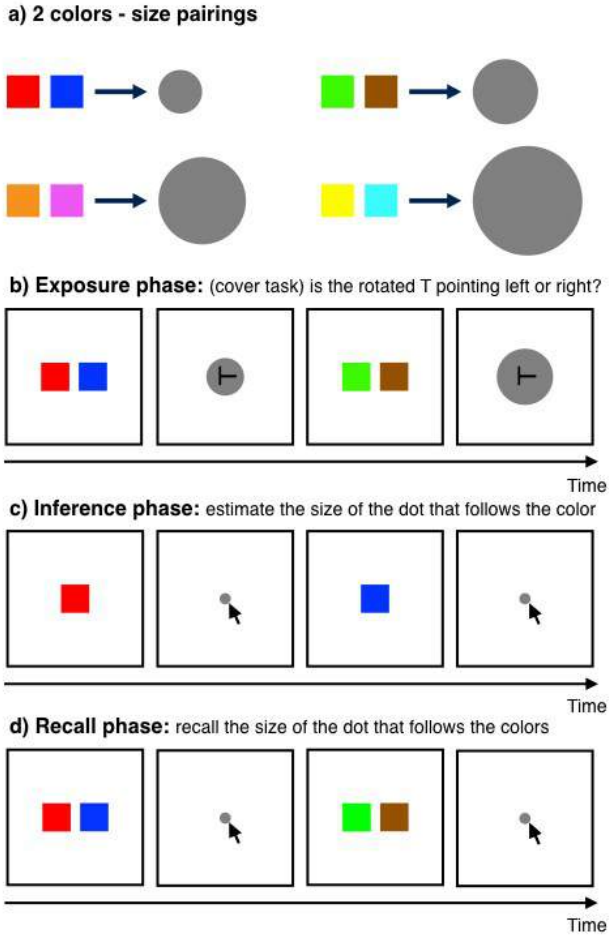


Figure 1. Experiment 1 paradigm. (a) Four color-size pairs were presented (e.g., red and blue squares followed by a circle with a diameter of 100 pixels). (b) Exposure phase using a cover task to expose the color-size pairs to participants. (c) Inference phase where participants estimated the size of the circle was associated with the color. (d) Recall phase where participants recalled the size of the circle that followed the two color squares.

To examine whether participants had successfully learned the color-size pairs (i.e., the association between the two color squares and the size of the circle), participants completed a size recall task following the inference phase (Fig.1d). In each trial, participants viewed the original color pair (e.g., red and blue squares) that they viewed during exposure for 500ms followed by a 3000ms blank screen. Afterwards, a probe circle with a diameter subtending 0.6° (or 20 pixels) was presented on the screen. Participants were asked to recall the size of the circle that was associated with the original two colors during exposure by adjusting the size of the probe using their mouse. The diameter of the

adjustable circle was restricted to a range from 20 pixels to 420 pixels. The adjustable circle remained on the screen until the "a" key was pressed to register participant's estimate. Each color pair was tested four times, producing 16 trials in total (the order of the trials was randomized).

A debriefing session was conducted at the end of the experiment, where participants were asked if they had noticed any pairings of squares and circles that appeared one after another. For those who responded yes, we further asked them to write in sentences which type of circle followed which colors.

Results and Discussion

We first analyzed whether the inferred circle size associated with one single object in the pair (e.g., red square) was different from the inferred circle size associated with the other member of the pair (e.g., blue square) to rule out any spatial positioning bias. We found that the inferred circle size associated with one object was not different from the inferred circle size associated with the other member in the pair for all four types of circle diameter (p's>.19). Thus, we combined the inferred size of either member in the pair.

We also found that in the recall phase, participants overestimated the size of the small circle (mean recalled circle diameter of a circle diameter of 100 pixels was 176.1, SD=84.5), and they underestimated the size of the large circle (mean recalled circle diameter of a circle diameter of 400 pixels was 225.8, SD=100.6). Given these recall biases, we compared the inferred size with the recalled size, not with the objective size in the following analyses.

The purpose of this experiment was to examine how the mind predicts the outcome given a single predictor, after learning that two predictors were associated with a specific outcome. We compared the inferred size associated with the single object during inference phase to the recalled size associated with the two objects to test the complete inheritance hypothesis. We also compared the inferred size associated with the single object during inference phase to the 50% of the recalled size to test the proportional inheritance hypothesis (Fig.2a).

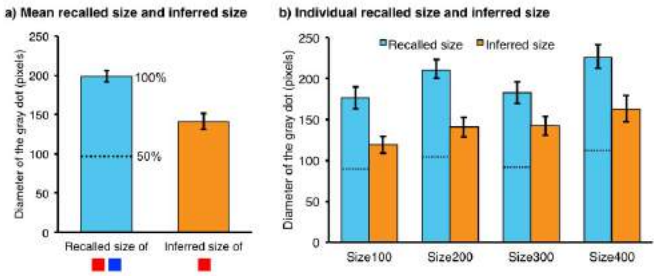


Figure 2. Experiment 1 results. (a) The mean recalled size of the circle associated with two objects and the mean inferred size of the circle associated with a single object. (b) The recalled size of the circle associated with two objects and the inferred size of the circle associated with a single object for each color-size pairing (error bars reflect ±1 SEM; dashed line represents 50% of the recalled size).

We found that the inferred size associated with the single object (mean inferred diameter=141.1, SD=63.6) was significantly smaller than the recalled size associated with the two objects (mean recalled diameter=198.6, SD=46.8) [$t(41)$=6.90, p<.001, d=1.03], but significantly larger than 50% of the recalled size [$t(41)$=5.01, p<.001, d=0.87] (corrected for multiple comparisons). Additionally, the same results were consistently found for each color-size pairing (Fig.2b). The results support the subadditive hypothesis. During debriefing, three participants reported noticing the color-size pairs, but none could correctly report which circle size followed which specific color pair. This suggests that participants had no explicit awareness of the color-size pairs.

These findings suggest that people implicitly predict a subadditive outcome from a single predictor after learning that two predictors previously jointly predicted a specific outcome.

Experiment 2

This experiment aimed to replicate and extend the findings in Experiment 1 by increasing the number of predictors from two to three.

Participants

A new group of 40 undergraduates (34 female, mean age=19.7 years, SD=2.2) from UBC participated in the experiment for course credit.

Stimuli

The stimuli were identical to those in Experiment 1, except that we added a black color (R/G/B=0/0/0) to the color set. There were nine color squares in total, randomly assigned into three triplets for each participant. Each triplet was randomly associated to a gray circle with a specific diameter. The circle diameter subtended 3.0° (or 100 pixels), 7.5° (or 250 pixels), or 12.0° (or 400 pixels) of visual angle (Fig.3a).

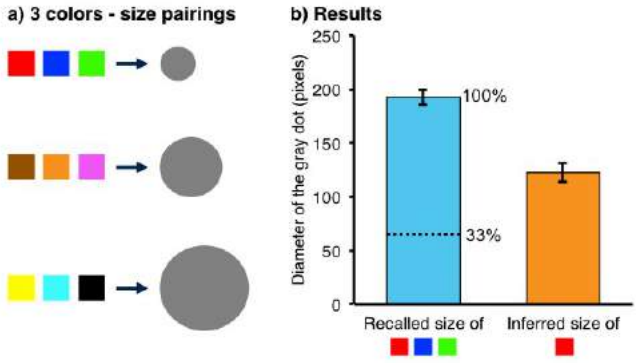


Figure 3. Experiment 2. (a) Three color-size pairs (e.g., red, blue, and green squares–circle with a diameter of 100 pixels). (b) The mean recalled size of the circle associated with three objects and the mean inferred size of the circle associated with a single object (error bars reflect ±1 SEM; dashed line represents 33% of the recalled size).

Procedure

The procedure was identical to that in Experiment 1, except that the three color squares were followed by a circle of a given size in the exposure phase, and participants recalled the circle size associated with the three squares in the recall phase.

Results and Discussion

In a one-way repeated-measures ANOVA, we found no difference between the inferred circle size associated with each object in the triplet for all three types of circle size (p's>.55). Thus, we combined the inferred size of each member in the triplet. We also found that participants overestimated the size of the small circle (mean recalled circle diameter of a circle diameter of 100 pixels was 188.4, SD=78.1) and underestimated the size of the larger circle (mean recalled circle diameter of a circle diameter of 400 pixels was 197.3, SD=94.8). Given these biases, we compared the inferred size with the recalled size, not with the objective size in the following analyses.

We found that the inferred size associated with the single object (mean diameter=124.2, SD=59.5) was significantly smaller than the recalled size associated with the three objects (mean diameter=198.3, SD=53.1) [$t(41)=7.87$, $p<.001$, $d=1.31$], but significantly larger than 33% of the recalled size [$t(41)=6.90$, $p<.001$, $d=1.32$] (corrected for multiple comparisons; Fig.3b). The results again support the subadditive hypothesis.

During debriefing, two participants reported noticing the color-size pairs, but none could correctly report which circle size followed the specific color triplet. This suggests that participants had no explicit awareness of the color-size pairs.

These findings successfully replicated the findings in Experiment 1, showing that people implicitly predict a subadditive outcome from a single predictor after learning that three predictors previously jointly predicted a specific outcome.

Experiment 3

Experiment 3 aimed to generalize the findings to other types of outcomes from circle sizes to numeric values. Specifically, after learning that two objects (e.g., red and blue squares) were associated with a specific numeric value, we examined how people made predictions of value from a single predictor (e.g., red square).

Participants

A new group of 45 undergraduates (41 female, mean age=20.38 years, SD=2.8) from UBC participated in the experiment for course credit.

Stimuli

The stimuli were identical to those in Experiment 1, except that each color pair was associated with a specific three-digit number. There were four three-digit numbers: 150, 400, 650, and 900. Each number was associated with a color pair (Fig.4a).

Procedure

As in Experiment 1, there were three phases (exposure, inference, and recognition). The exposure phase was identical to Experiment 1, except that in the cover task, participants viewed a three-digit number above the rotated T in the circle (Fig.4b). Since that a specific number may be easier to learn than the size of a circle, we reduced the number of repetitions for each color-number pair to 40 times, resulting in 160 trials in total (the order of trials was randomized).

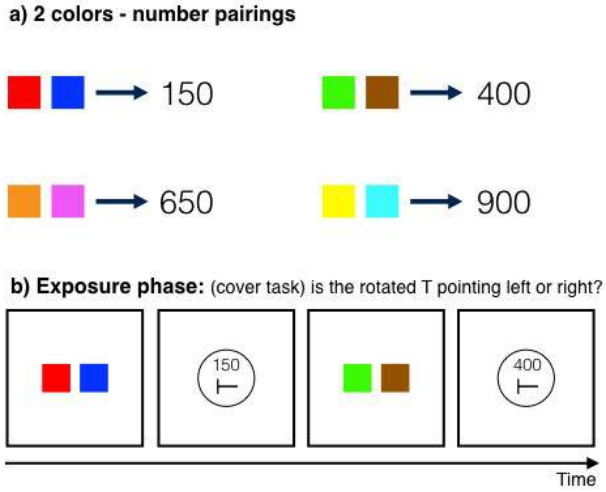


Figure 4. Experiment 3 paradigm. (a) Four color-number pairs (e.g., red and blue squares-150). (b) Exposure phase using a cover task to expose the color-number pairs to participants.

In the inference phase, participants viewed a single colored square and were asked to estimate the number that was associated with the color square by typing a number on the keyboard. The estimated number was restricted to a range from 0 to 1050. Participants had the option to delete and revise their estimated number until the "a" key was pressed to register their estimate.

In the recognition phase, participants viewed a pair of color squares that was presented in exposure and were asked to recall the number that was associated with the color pair by typing the number on the keyboard. The recalled number was restricted to a range from 0 to 1050. Participants had the option to delete and revise their recalled number until the "a" key was pressed to register their estimate. A debriefing session was conducted at the end as before.

Results and Discussion

We found that the inferred number associated with a single object was not different from the inferred number associated with the other member in the pair (p's>.32). Thus, we combined the inferred number of each member in the pair. We also found that participants overestimated the small number (mean recalled number of 150 was 489.7,

SD=227.3) and underestimated the large number (mean recalled number of 900 was 557.0, SD=251.7). Given these biases, we compared the inferred number with the recalled number, not with the objective number in the following analyses.

We found that the inferred number associated with the single object (mean inferred number=476.5, SD=150.8) was marginally smaller than the recalled number associated with the two objects (mean recalled number=513.2, SD=97.3) [$t(44)=1.79$, $p=.08$, $d=0.29$], but significantly larger than 50% of the recalled number [$t(44)=10.86$, $p<.001$, $d=1.96$] (corrected for multiple comparisons; Fig.5). The results again support the subadditive hypothesis.

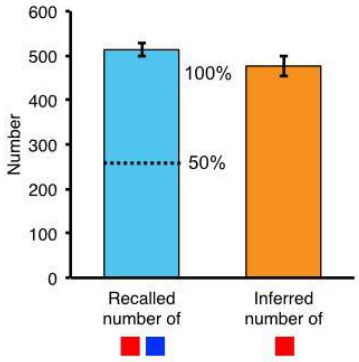


Figure 5. Experiment 3 results. The mean recalled number associated with two objects and the mean inferred number associated with a single object (error bars reflect ±1 SEM; dashed line represents 50% of the recalled number).

During debriefing, two participants reported noticing the color-number pairs, but none could correctly report which number followed which specific colors. This suggests that participants had no explicit awareness of the color-number pairs.

These findings again replicated the findings in Experiment 1, showing that people implicitly predict a subadditive outcome from a single predictor after learning that two predictors previously jointly predicted a specific outcome.

Experiment 4

This experiment aimed to extend the findings in Experiment 3 by increasing the number of predictors from two to three.

Participants

A new group of 33 undergraduates (28 female, mean age=20.2 years, SD=1.7) from UBC participated in the experiment for course credit.

Stimuli and Procedure

The stimuli and the procedure were identical to Experiment 3, except that there were three color triplets and each triplet was associated with 150, 525, or 900 (Fig.6a).

Results and Discussion

In a one-way repeated-measures ANOVA, we found no difference between the inferred number associated with each object in the triplet for all three types of numbers (p's>.35). Thus, we combined the inferred number of each member in the triplet. We also found that participants overestimated the small number (mean recalled number of 150 was 485.7, SD=267.0) and underestimated the large number (mean recalled number of 900 was 592.7, SD=246.8). Given these biases, we compared the inferred number with the recalled number, not with the objective number in the following analyses.

We found that the inferred number associated with the single object (mean inferred number=401.1, SD=180.3) was significantly smaller than the recalled number associated with the three objects (mean recalled number=501.3, SD=106.8), [$t(32)=3.03$, $p=.005$, $d=0.68$], but significantly larger than 33% of the recalled number [$t(32)=7.61$, $p<.001$, $d=1.80$] (corrected for multiple comparisons; Fig.6b). The results again support the subadditive hypothesis.

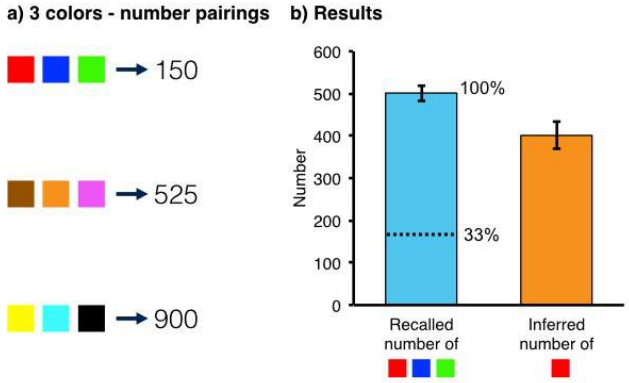


Figure 6. Experiment 4. (a) Three color-number triplets (e.g., red, blue, and green squares-150). (b) The mean recalled number associated with three objects and the mean inferred number associated with a single object (error bars reflect ±1 SEM; dashed line represents 33% of the recalled number).

During debriefing, one participant reported noticing the color-number pairs, but the participant could not correctly report which number followed which specific colors. This suggests that participants had no explicit awareness of the color-number pairs.

These findings replicated the findings in Experiment 3, showing that people implicitly predict a subadditive outcome from a single predictor after learning that three predictors previously jointly predicted a specific outcome.

General Discussion

The goal of this study was to examine how the mind automatically generates prediction when only a single cue is present, after learning that multiple cues were previously jointly associated with an outcome. We found that after learning that two co-occurring objects (e.g., red and blue squares) predicted a specific circle size, participants inferred the circle size associated with a single color (e.g., red

square) to be smaller than the original circle size associated with the color pair, but larger than 50% of the circle size associated with the color pair (Experiment 1). We further extended the number of predictors from two to three. After learning that three co-occurring objects predicted a specific circle size, participants inferred the circle size associated with a single color to be smaller than the circle size associated with the color triplet, but larger than 33% of the circle size associated with the color triplet (Experiment 2). We further replicated and extended the experiment from circle sizes to numeric values as outcomes for two predictors (Experiment 3) and three predictors (Experiment 4). Importantly, no participant was consciously aware of the association between the predictors and the outcome across all experiments, suggesting that the inference of the size or number associated with one single predictor was largely implicit. The current findings also suggest when people predict an outcome relying on a single cue from a set of cues, they do not inherently generate the prediction based on the outcome associated with the complete set of cues, nor do they proportionally inherit the outcome based on the number of cues. Instead, they make predictions in a subadditive manner, which is consistent with support theory (Tversky & Koehler, 1994).

One rationale behind support theory is that unpacking an event to its individual component may evoke other relevant elements that might have been missed. When participants were asked to infer the size associated with each individual color in the pair or triplet, they might have to think more extensively for each color, compared to recalling the outcome associated with the color pair or triplet. A second rationale behind support theory is that explicitly referring to an individual component of an event would increase its salience. When participants were asked to infer the size associated with a single color, their attention was drawn to the single color which may increase the weight of the single color in their prediction of the outcome.

Alternatively, previous studies have suggested that seeing one object in a pair may activate the unitized representation of the pair (e.g., Alvarez & Oliva, 2008). The co-occurring objects (e.g., red and blue squares) may be grouped in the mind during learning. When participants were asked to predict the outcome relying on a single object (e.g., a red square), the object may trigger the representation of the group but not fully activate the representation of the group. Therefore, participants may predict an outcome above 50% but less than 100% of the original outcome.

Another possible explanation is that participants could add the size predicted by each colored square in a sublinear fashion, creating a subadditive sum. A new experiment is needed to test this hypothesis to tease apart whether the subadditivity is driven by the sublinear representation of each size or the sublinear summation of the two sizes. Specifically, participants are first exposed to one unique color predicting a unique size during the exposure phase (e.g., a red square predicting a circle with a certain diameter, and a blue square predicting a circle with a certain diameter). In the inference phase, participants see a red square with a blue square presented side by side simultaneously, and they will be asked to infer the circle size associated with the two squares. In the recall phase, participants simply recall the original size of the circle associated with the red square and the blue square. If the inferred size is equal to the sum of the two recalled sizes, then this would suggest that participants use an additive approach to predict the outcome. If the inferred size is smaller than the sum of the two recalled sizes but larger than each recalled size, then this would suggest that participants use a subadditive approach.

In summary, we found a new consequence of statistical learning on automatic inferences: When multiple objects jointly predict a specific outcome, the presence of a single object implicitly triggers a subadditive prediction.

Acknowledgments

We would like to thank Oleg Urminsky, Ru Qi Yu, Brandon Tomm, and two anonymous reviewers for their helpful comments. This work was supported by NSERC Discovery Grant (RGPIN-2014-05617 to JZ), the Canada Research Chairs program (to JZ), the Leaders Opportunity Fund from the Canadian Foundation for Innovation (F14-05370 to JZ), and Cordula and Gunter Paetzold Fellowship (to YL).

References

Alvarez, G. A., & Oliva, A. (2008). The representation of simple ensemble visual features outside the focus of attention. *Psychological Science,* 19, 392–398.

Brady, T. F., Konkle, T., & Alvarez, G. A. (2009). Compression in visual working memory: Using statistical regularities to form more efficient memory representations. *Journal of Experimental Psychology: General, 138,* 487-502.

Brady, T. F., Konkle, T., Alvarez, G. A. and Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences, USA, 105*, 14325-14329.

Fiser, J., & Aslin, R. N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological Science, 12,* 499–504.

Kim, G., Lewis-Peacock, J. A., Norman, K. A., & Turk-Browne, N. B. (2014). Pruning of memories by context-based prediction error. *Proceedings of the National Academy of Sciences, 111,* 8997-9002.

Luo, Y. & Zhao, J. (2017). Learning induced illusions: Statistical learning creates false memories. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. J. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society,* (pp. 774-779). Austin, TX: Cognitive Science Society.

Luo, Y., & Zhao, J. (2018). Statistical Learning Creates Novel Object Associations via Transitive Relations. *Psychological Science,* 29, 1207–1220.

Mackintosh, N. J. (1975). A theory of attention: variations in the associability of stimuli with reinforcement. *Psychological Review, 82*, 276.

Otsuka, S., & Saiki, J. (2016). Gift from statistical learning: Visual statistical learning enhances memory for sequence elements and impairs memory for items that disrupt regularities. *Cognition, 147,* 113-126.

Le Pelley, M. E., Mitchell, C. J., Beesley, T., George, D. N., & Wills, A. J. (2016). Attention and associative learning in humans: an integrative review. *Psychological Bulletin, 142*, 1111.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science, 274,* 1926-1928.

Tversky, A., & Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological review,* 101, 547.

Turk-Browne, N. B., Isola, P. J., Scholl, B. J., & Treat, T. A. (2008). Multidimensional visual statistical learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34,* 399–407.

Turk-Browne, N. B., Jungé, J. A., & Scholl, B. J. (2005). The automaticity of visual statistical learning. *Journal of Experimental Psychology: General, 134,* 552–564.

Turk-Browne, N. B., Scholl, B. J., Chun, M. M., & Johnson, M. K. (2009). Neural evidence of statistical learning: Efficient detection of visual regularities without awareness. *Journal of Cognitive Neuroscience, 21,* 1934–1945.

Yu, R., & Zhao, J. (2015). The persistence of attentional bias to regularities in a changing environment. *Attention, Perception, & Psychophysics, 77,* 2217-2228.

Zhao, J., Al-Aidroos, N., & Turk-Browne, N. B. (2013). Attention is spontaneously biased toward regularities. *Psychological Science, 24,* 667–677.

Zhao, J., & Luo, Y. (2017). Statistical regularities guide the spatial scale of attention. *Attention, Perception, & Psychophysics, 79*, 24-30.

Zhao, J., & Yu, R. (2016). Statistical regularities reduce perceived numerosity. *Cognition, 146,* 217-222.

Reasoning about dissent: Expert disagreement and shared backgrounds

Jens Koed Madsen (jens.madsen@ouce.ox.ac.uk)
School of Geography and the Environment, University of Oxford
OX1 3QY, South Parks Road, Oxford, United Kingdom
Orcid: 0000-0003-2405-8496

Ulrike Hahn (u.hahn@bbk.ac.uk)
Department of Psychological Sciences, Birkbeck College, University of London
WC1E 7HX, Malet Street, London, United Kingdom

Toby Pilditch (t.pilditch@ucl.ac.uk)
School of Geography and the Environment, University of Oxford
OX1 3QY, South Parks Road, Oxford, United Kingdom
&
Department of Experimental Psychology, University College London
26 Bedford Way, WC1H 0AP, London, United Kingdom

Abstract

Sequential testimonies where more or less reliable sources argue about an issue are central to public debates. Often, the majority of sources may argue that a hypothesis is true while a minority dissenter may claim the opposite (e.g. scientists and lobbyists in the climate change debate).

In this paper, we show that people are sensitive to source reliability as well as the structural relationship between the sources. Participants follow Bayesian predictions for revising belief in the hypothesis *and* the reliability of the competing sources given majority consent, minority dissent, and shared reliability between sources. Shared reliability and dissent is a key issue for public debate and belief revision. The paper provides novel insight into the workings of these aspects.

Keywords: Source reliability; Shared reliability; Source dependency; Bayesian modelling; Belief revision

Introduction

Information is crucial to revising or maintaining beliefs, to making decisions in an uncertain world, and to compare and contrast support for competing hypotheses. While we can certainly acquire information through personal experience (e.g. witnessing congested traffic may change the route we travel to work, participating in a public demonstration may give an impression of the degree of support for a particular cause, etc.), most of the information we get in our everyday lives comes via other people. Meteorologists provide us with necessary weather information for planning the day, news readers give us an overview of relevant events that happen within our respective countries and abroad, and friends, family members, and co-workers provide invaluable information on a range of issues that help us appreciate their lives, consider information we have not been privy to before talking to that person, or information that is necessary for doing our respective jobs.

Appeals to authority have traditionally been regarded as a reasoning fallacy – this is due to the fact that perceived authority should not add credence to the conditional link between the evidence and a hypothesis. That is, whether or not a piece of information increases the likelihood of a hypothesis is, in principle, independent from the source that conveys the piece of information. Classically, this has led some people to be sceptical of appeals to authority.

The notion that appeals to authority should be distrusted in principle reverberates in theories of argumentation and reasoning. For example, two prominent models of persuasion, the Elaboration Likelihood-Model (Petty & Cacioppo, 1984) and the Heuristic-Systematic Model (Chaiken & Maheswaran, 1994), classify appeals to authority as a shallow and weak cue. In this view, people should disregard the characteristics of the source as they are given greater incentive to interrogate and elaborate on the evidence and its relation to the hypothesis. In other words, as the incentive to understand the link between evidence and claim increases, the nature of the source should matter less and less.

While it is it true in principle that the messenger neither adds nor subtracts to the link between evidence and claim, overlooking the epistemic impact of perceived source reliability neglects a crucial communicative and reasoning function. In a world where sources can lie and make up evidence, their reliability becomes crucially linked with the strength of the argument. Additionally, in a highly uncertain world, some information requires deep expertise to process (e.g. climate data may be accessible to a general population, but requires considerable expertise to adequately model and understand). Given the capacity to misinform and generate mistaken causal models due to a lack of expertise, the reliability of the speaker is an important element for people to update and revise their beliefs in the world.

In line with this perspective, the impact of the perceived reliability of a source is shown to be crucial for reasoning and decision-making. Treating the reliability of a source as a shallow cue, the literature on persuasion has shown the impact of appeals to authority (Petty & Cacioppo, 1984; Tormala & Clarkson, 2007), the developmental literature suggests children seek out credible figures to guide their perception of the world (Harris & Corriveau, 2011), and appeals to authority have been shown to impact legal

reasoning (Lagnado et al., 2013). Further, it increases adherence with persuasion strategies (Cialdini, 2007), and perceived reliability is able to predict whether or not people believe an unknown policy is good, given recommendations from different political sources (Madsen, 2016).

The paper explores three aspects of perceived reliability. First, it replicates a Bayesian model of the impact of sequential reports from more or less reliable sources. This replication shows people update their beliefs in a hypothesis given reports from sources *as well as* updating the perceived reliability of the sources themselves. Second, it replicates recent findings that shared reliability (e.g. sources sharing a common background) impacts the degree of belief in a hypothesis and the perceived reliability of sources, in line with Bayesian predictions. This explores aspects of source dependency. Finally, we extend this work by presenting novel findings on the impact of minority dissenters on belief in a hypothesis and the perception of reliability among sources, given the introduction of shared reliability. Minority dissent and shared reliability are crucial aspects of information transmission (see see Whalen et al., 2013 for dependency and Perfors et al., 2018 for minority dissent), as they appear a number of domains – indeed, most debates are characterised by sources that disagree. For example, in climate change both are apparent and important factors of public debate.

A Bayesian approach to source reliability

Whilst some have argued reliance on the reliability of others to revise subjective beliefs about the world is a shallow persuasive cue (Petty & Cacioppo, 1984; Chaiken & Maheswaran, 1994), others have argued reliance is rationally justified and a necessary component of belief revision (see Bovens & Hartmann, 2003; Hahn et al., 2009).

The latter applies a Bayesian perspective to reliability. Bayesian reasoning uses subjective, probabilistic degrees of belief in propositions where Bayes' theorem integrates prior beliefs with the likelihood ratio to estimate the posterior degree of belief (Howson & Urbach, 1996). Bayes is an alternative to logicist approaches to reasoning (Oaksford & Chater, 1991) and has been applied to argumentation theory (Hahn & Oaksford, 2006; 2007), which has found Bayesian reasoning can account human information integration in practical reasoning (see Oaksford & Chater, 2007).

The Bayesian approach suggests that people's subjective perceptions of the reliability of the speaker normatively should yield different information integration. For example, if the messenger has no expertise, the information may be regarded as pure noise (as it is equally likely to be true or false). In this case, the recipient should not revise her beliefs one way or another. Comparatively, if low trustworthiness entails simple misinformation, the recipient may increase her belief in the opposite direction given positive reports from a distrusted source. Due to the Bayesian nature of the above models, the reliability function of reports relies on conditional probabilities (see e.g. Madsen, 2016 where participants revise their beliefs negatively in a proposed policy given positive reports from subjectively distrusted politicians).

More formally, the model integrates two components to account for overall reliability: perceived trustworthiness and perceived expertise (Hahn et al., 2009)[1]. In this framework, expertise refers to the *capacity* to provide accurate information about the topic in question. This is highly domain-dependent. For example, an astrophysicist may be able to calculate the mass of a distant celestial body, but may not be able to give a valid economic forecast. While expertise refers to capacity, trust refers to the *intention* of providing true and accurate information to the best of ones ability. For example, the astrophysicist may omit data points that contradict personally held theories or beliefs. The model components are orthogonal, as a person can be highly expert in some domain while at the same time be entirely untrustworthy – or vice versa. The orthogonal assumption is theoretically grounded (Bovens & Hartmann, 2003) and empirically supported (Harris et al., 2015)

Formally, Bayes' theorem is used to integrate reliability where the posterior degree of belief in the hypothesis (H) given the representation (Rep) yields:

$$P(H|Rep) = \frac{P(H)x\,P(Rep|H)}{P(H)x\,P(Rep|H) + P(\neg H)x\,P(Rep|\neg H)}$$ [2]

The formalisation predicts how people should integrate uncertain information from more or less reliable sources. Model predictions have enjoyed a good fit with behavioural data (Harris et al., 2015; Madsen, 2016). Overall, the findings suggest people are sensitive to the reliability of the individual speaker and integrate the information from the speaker in a normatively rational manner.

Shared reliability: corroboration and negation

The empirical work underpinning the Bayesian source reliability model suggests that people do modulate information integration given perceived speaker reliability. The influence of reliability on belief revision means that the perceived source reliability *itself* is important in the belief revision process.

If the recipient believes the source is credible, she should revise her beliefs more positively if the source provides positive reports for a hypothesis. As a consequence, if the perceived source reliability changes, Bayesian (normative) models entail that the effect of this source should change for future reports and the impact of the already observed report. That is, if a speaker is revealed to be less than credible, audiences should be more likely to disregard any reports from that source in the future. Changes to the perceived

[1] The operationalization of reliability as an amalgamation of perceived expertise and trustworthiness is remarkably close to findings in social psychology where reliability is defined as an amalgamation of traits related to *warmth* and *competence* (Fiske et al., 2007; Cuddy et al., 2011).

[2] P(Rep|H) = P(Rep|H, Exp, T) * P(Exp) * P(T) + P(Rep|H, ¬Exp, T) * P(¬Exp) * P(T) + P(Rep|H, ¬Exp, ¬T) * P(¬Exp) * P(¬T) + P(Rep|H, Exp, ¬T) * P(Exp) * P(¬T); mutatis mutandis for P(Rep|¬H)

reliability of sources can conceivably happen for a number of reasons – for example if the source corroborates a highly unlikely hypothesis.

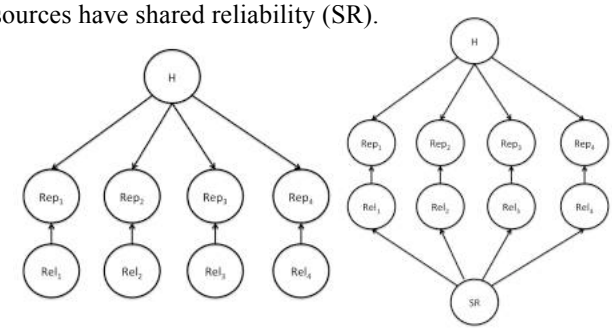
Further, reports are seldom made in isolation. Frequently, people will see multiple reports for a given issue. These sources, perceived as more or less credible by the recipient in question, may argue for or against a hypothesis. For example, in climate change debates, pundits, experts, and members of the media frequently make predictions about a particular hypothesis or issue. Considering flood risks in coastal areas, many experts tend to warn that weather will become more extreme and floods more prominent (e.g. in Miami). However, minority dissenters may argue that floods will not change over time. Here, we have multiple sources (e.g. scientists) that corroborate and support a hypothesis (rising floods) and a dissenter (e.g. a senator) who negates the hypothesis.

People's prior belief in the hypothesis, their perceived reliability for each source, and their perception of the dependency of sources (e.g. shared reliability versus independent sources) should normatively influence their belief in the hypothesis and perceived reliability, given positive or negative reports from the sources. The paper explores whether this is the case empirically.

In order to approach these questions, we use Bovens and Hartmann (2003) foundational and Bayesian perspective on modeling source reliability. Aside from suggesting people should revise their belief in the reliability of the source and in the hypothesis given sequential reports, their models show that *the structure* of the perceived relationship of sources influence the degree to which their reports should these beliefs given multiple testimonies.

Figure 1a-b illustrates different structural relationships between independent sources with independently perceived reliability ($Rel_{1\text{-}4}$) who provide a report ($Rep_{1\text{-}4}$) concerning a hypothesis (H). 'Independent sources' refer to situations where the sources can be considered entirely independent of one another (Fig. 1a). For example, climate scientists may run studies independently of each other and report their findings with no knowledge of the findings of other scientists (here, the strength of the report will in part depend on each reports personal reliability).

Comparatively, if sources share a common background (e.g. the scientists may have been trained at the same school to use a specific model to explore climate phenomena), they become partially dependent (Fig. 1b). In this case, the sources have shared reliability (SR).



(1a) *(1b)*

Fig. 1a-b: independent sources and sources with shared reliability

Shared reliability constrains the informativeness of a source, as their reliabilities are influenced by the common-cause (e.g. attending a good or bad school). That is, the common background can weaken the impact of the reports provided by these sources. More intuitively, in finding out that sources share a compromising background (e.g. have all attended a fraudulent school), then the individual reliabilities of those sources are compromised, and in turn the strength of their support. Bovens and Hartmann (2003) provide a formal way to calculate "…how the posterior probability of the reliability of the nt^h witness increases as more and more witness reports from in" (p. 79):

$$P^{*(n)}(REL_n) = P(REL_n|REP_1, \ldots, REP_n) = \frac{h[us(s+a\bar{s})^{n-1} + \bar{u}t(t+a\bar{t})^{n-1}]}{h[u(s+a\bar{s})^n + \bar{u}(t+a\bar{t})^n] + \bar{h}[u(s+a\bar{s})^n + \bar{u}(t+a\bar{t})^n]}$$

where u is the probability of the shared background being reliable, P(SR) – that is, how reliable the source is seen to be prior to any information about shared reliability, s is the conditional probability: $1 > P(Rel_i|SR)$ – that is, the likelihood that source i is reliable *given* the shared reliability. The conditional probability $> P(Rel_i|SR) > 0$ is represented by t, whilst a is a randomization parameter (that is, the degree of noise), and h is the prior probability of the hypothesis (that is, degree of belief in the hypothesis prior to any reports).[3]

In sum, the equation shows that the posterior degree of reliability of the n^{th} witness (or source) depends on the randomization parameter (a) and prior probability of the hypothesis (h). For example, if $a = .9$ and $h = .3$, initial witness reliability falls from .5 to .25 (see p. 80), but increases as additional positive reports confirm the initial report.[4]

Recently, Madsen et al. (2018) tested this intuition. That is, whether people update their beliefs in the reliability of the source and the belief in the hypotheses when they experience sequential corroborative testimonies. In their study, all reports *corroborated* the hypotheses (that is, all sources provided positive reports for the hypothesis). However, as mentioned, many (if not most) debates are between sources that disagree about a particular issue. For this reason, it is imperative to understand the function of (minority) dissenters.

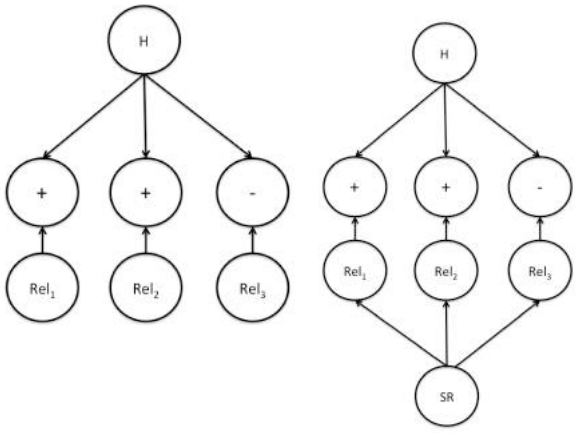
Madsen et al. find support for the model of corroborating sources, as P(Rel) decreased given a corroborative report of an unlikely hypothesis, but subsequently increased as more corroborative reports were given. Further, when participants learned sources attended the same school, they adjusted their posterior degree of belief negatively for the hypothesis *and* source reliability. The effect was stronger if experts'

[3] C.3 (pp. 136-137) and C.4 (pp. 137-138) in Bovens and Hartmann (2003) provide the full derivation for $P^{*(n-1)}(REL_n) = P(REL_n|REL_1, \ldots, REL_{n-1})$ and $P^{*(n)}(REL_n) = P(REL_n|REP_1, \ldots, REP_n)$ respectively

[4] The current study does not elicit a randomization parameter

school was bad compared with sharing a school described as 'excellent'. Finally, their study suggests people revise posterior degree of belief in the reliability of sources retrospectively. That is, as $sources_{2\text{-}3}$ provided reports, the reliability of $source_1$ was adjusted to be in line with perceived reliability of the n^{th} source.

This paper extends this work by exploring three facets of source dependency and reliability. First, a source may corroborate or negate a report for a given hypothesis. We test how participants update their beliefs in the hypothesis and the reliability of each source given corroborative reports from $sources_{1\text{-}2}$ and a negative report from $source_3$ (denoted by '+' and '-' respectively in the Fig. 2a and b).



(2a) *(2b)*

Fig. 2a-b: Negation from independent sources and sources with shared reliability

Second, we explore how dependency impacts perception of the hypothesis and reliability when the 3rd source dissents and negates the reports of $sources_{1\text{-}2}$. Here, we use the same a shared reliability structure (Fig. 2b).

We use the experimental design of Madsen et al. (2018), altering it to explore the following:

Given reports from the 3rd dissenting source, we expect decreases in reliability for all sources and a decrease in belief in the hypothesis. This is due to the fact that dissent adds additional uncertainty to the hypothesis (the initial sources may be wrong) and onto the sources themselves (either the 3rd source or the initial sources may have been mistaken/providing bad information). For the likely scenario we expect a significant drop in reliability for the 3rd source in particular, as this source goes against a very likely hypothesis *and* two corroborating reports. In addition, in accordance with Madsen et al (2018), P(Rel) should *decrease* when source 1 reports an unlikely hypothesis, but subsequently recover, as source 2, still perceived to be independent, corroborates the unlikely prediction.

Method

Material and procedure: To replicate Madsen et al. (2018) and enable direct comparisons, we use their method and materials. To test model predictions, we use low and high probability scenarios. In the low probability scenario, participants were asked to evaluate the likelihood of a crash in the stock market with the following description:

> *"Imagine you are watching a news programme about the economy. Specifically, the programme considers whether or not the UK stock market will crash (i.e. fall by more than 30%) within the next 6 months. Historically, the likelihood of a crash occurring within a 6-month window is 5%.*
>
> *In your opinion, how likely is the UK stock market to crash within the next 6 months?"*

Having read this, participants provided prior estimates for their beliefs in the hypothesis on a scale from 0-1 (0: I am completely certain the stock market will NOT crash within the next 6 months; 1: I am completely certain the stock market will crash within the next 6 months). To elicit the reliability of sources, we defined reliability:

> *"Reliability can be defined as having access to relevant information about a topic, and a willingness to say what you believe to be the true state of the world.*
>
> *How reliable are economists in predicting the market crashes?"*

Having read this, participants provided their belief in the source reliability from 0-1 (0: economists are completely unreliable; 1: economists are completely reliable). Reports from sources were provided as interviews with experts on the subject. For example:

> *"Now, imagine that an economist, Robert, is being interviewed about the economy. Robert states the following: "I am completely certain the stock market will crash within the next 6 months."*
>
> *Given Robert's report, how likely is the UK stock market to crash within the next 6 months?"*

Participants then gave subjective estimates of their beliefs in the hypothesis and in each source hitherto presented. Sources were presented sequentially. To test the effect of negation, $sources_{1\text{-}2}$ always corroborated the hypothesis and $source_3$ always negated the hypothesis. This implemented a minority dissenter. The dissent only functions in light of the initial corroborations. As such, the dissenter had to be at the end of the scenario. Further, placing the dissenter towards the end allowed for replication of corroborative $reports_{1\text{-}2}$, as participants had not yet been exposed to dissent.

Finally, having seen the three sequential reports, the participants were told the sources were partially dependent (i.e. shared a background), which was manipulated between-subjects as either high or low quality (SR Condition). An example of the high quality SR Condition statement for the low likelihood scenario:

> *"It turns out, all the interviewed economists studied at the same school and subscribe to the same economic theories. Their school has a very good reputation for excellent teaching and accurate approaches to economy.*
>
> *Given the fact that they all studied at the same school and follow the same economic theories, how likely is the UK stock market to crash within the next 6 months?"*

After each report and the SR condition, P(H) and $P(Rel_{1\text{-}n})$ were measured. Participants read both scenarios in a

counterbalanced order, with the SR Condition manipulated independently for each scenario.[5].

Participants: 100 participants (71 female, μ_{age} = 34.51, σ = 11.49) were recruited from the online recruitment source Prolific Academic. All had to be aged 18+ and native English speakers from either the UK or the USA. All participants had to have a prior completion rate of 95%. Median completion time was 5.56 min (σ = 2.11) and participants were paid £0.8 (resulting in an effective fair hourly wage of £8.63/hour for participation).

Results

All inferential statistics reported below were Bayesian[6], and were conducted used the JASP statistical software (JASP Team, 2018). The probability manipulations were successful in generating high and low estimates for the two scenarios: The market crash scenario was rated as unlikely (μ = .337, σ = .243) and the salmon growth scenario was rated as likely (μ = .806, σ = .116). In both scenarios, sources were rated higher ($P(Rel_{Economist})$: μ = .638, σ = .156; $P(Rel_{Biologist})$: μ = .731, σ = .128). Importantly, though, both sources were rated positively, which allows for the testing of whether positive reports of unlikely hypotheses influence reliability estimates negatively.

Following predictions from Bovens and Hartmann (2003), we expect positive reports of an unlikely hypothesis to lead to an initial decrease in estimates of reliability. To test this, we use repeated measures ANOVA ($P(Rel) - P(Rel_1|Rep_1)$). We observe a negative revision of reliability of source 1 given a positive report of an unlikely hypothesis (N = 100), BF_{10} = 179636.1 (in the current design, the source predicts the stock market will crash within a 6-month period). However, as participants learn another source also provides a positive report ($P(Rel_1|Rep_1) - P(Rel_1|Rep_2)$), they revise their belief in the initial source and revise reliability in a positive direction (N = 100), BF_{10} = 798759.1.

When a third source then contradicts ($P(Rel_1|Rep_2) - P(Rel_1|Rep_3)$), the reliability of the original reporter is then reduced once again (N = 100), BF_{10} = 352673.82 We further note strong evidence for a null difference in the estimated reliabilities across the three sources (N = 100), BF_{10} = 0.127, despite the presence of a contradicting minority, suggesting that all sources are penalized given the dissent among them.

In addition, participants *increase* their belief in the likelihood of the hypothesis, whilst they simultaneously *decrease* their belief in the reliability of the reporting source ($P(H)$ to $P(H|Rep_1)$; N = 100), $BF_{10} = 5.958 * 10^7$. That is, the introduction of a dissenting minority source on belief in the likelihood of the hypothesis ($P(H|Rep_2) - P(H|Rep_3)$) leads to a significant decrease (N = 100), BF_{10} = 281255.7.

We next turn the high likelihood scenario (biologists predicting salmon growth). To test whether participants neither increase or decrease the reliability of sources that provide positive statements for highly likely hypotheses (hypothesis 2), we conducted a repeated measures ANOVA ($P(Rel)$ to $P(Rel_1|Rep_1)$), finding no significant change (N = 100), BF_{10} = 0.401. We do however note the introduction of a contradicting source ($P(Rel_1|Rep_2) - P(Rel_1|Rep_3)$) leads to a significant decrease in reliability of the first source (N = 100), BF_{10} = 12458.36. Critically, given this introduction, and separating these results from those of the unlikely scenario, there was a substantial difference in the estimated reliability of the dissenter (μ = .489, σ = .207), and the first two (corroborating) reporters (Source 1: μ = .707, σ = .16; Source 2: μ = .717, σ = .156; N = 100), $BF_{10} > 1 * 10^{10}$. This suggests that - while introducing uncertainty to the reliability of the corroborating sources - providing dissenting reports about a hypothesis with a high prior belief and two corroborating reports can significantly damage perceived reliability. That is, if a minority dissents against prevailing wisdom *and* goes against other witnesses, she may suffer a loss of reliability.

We further note that as in the unlikely scenario, the introduction of a report from dissenting minority source on belief in the likelihood of the hypothesis ($P(H|Rep_2) - P(H|Rep_3)$) leads to a significant decrease (N = 100), $BF_{10} = 5.561 * 10^6$. The main results are shown in Fig. 3.

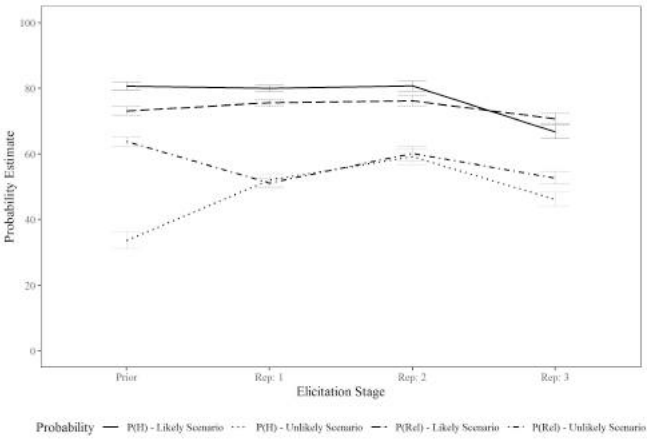


Fig. 3: P(Rel) and $P(H)$ given $reports_{1\text{-}3}$

Results of shared reliability

To test whether the impact of introducing a shared reliability among sources (hypothesis 3), we compare posterior degrees of belief in the hypothesis and the reliability of the sources.

A repeated measures ANOVA was conducted on belief in the hypothesis (P(H)) for the introduction of the shared reliability information (i.e. $P(H|Rep_3)$ to P(H|SR)), with the inclusion of the shared reliability condition (high/low-quality) as a between-subjects condition.

For the unlikely scenario, belief in the hypothesis (economic crash), was affected by the introduction of a

[5] The high likelihood scenario was identical to the above, but considered predictions that the Norwegian salmon population would grow over the next 5-year period.

[6] All analyses assume an uninformed prior.

shared reliability (main effect of introduction), $BF_{Inclusion}$[7] = 110.1, and if shared reliability was high or low quality (low < high), $BF_{Inclusion}$ = 272.1, demonstrating a successful manipulation check. Importantly, the significant interaction of shared reliability condition, and its introduction, $BF_{Inclusion}$ = 550.8, revealed belief in the hypothesis decreased when the shared reliability was low-quality but increased when the shared reliability was of high quality. Consequently, the model with all the above terms included was the best fit, BF_M[8] = 550.81, and significant overall, BF_{10} = 486.26.

We observe the same effects for revision of reliability estimates. The main effect of an introduction of a shared reliability, $BF_{Inclusion}$ = 4.798 * 10^9, and main effect of shared reliability condition (low-quality < high-quality), $BF_{Inclusion}$ = 1.583 * 10^8, are best described by the significant interaction of the two, $BF_{Inclusion}$ = 4.421 * 10^8, where high quality shared reliability leads to a minor increase in estimated reliability, whilst low quality shared reliability leads to a substantial decrease. Once again, the model with all the above terms included was the best fit, BF_M= 4.421 * 10^8, and significant overall, BF_{10} = 1.552 * 10^{10}.

The above analyses were then repeated for the likely scenario, where, against predictions, the belief in the hypothesis (salmon growth) was found to be unaffected by the introduction of a shared reliability, $BF_{Inclusion}$ = 1.227, or its quality, $BF_{Inclusion}$ = 0.194. However, the introduction of shared reliability was found to decrease estimations of source reliability, $BF_{Inclusion}$ = 1.116 * 10^{10}, and whether a shared reliability was high or low-quality led to higher or lower reliability estimates (respectively), $BF_{Inclusion}$ = 5.082 * 10^7, once more passing the manipulation check. Critically, reductions in reliability (given the introduction of a shared reliability among sources), is found to be localized to when the introduced shared reliability is of low-quality (right-hand facet, Fig. 4), $BF_{Inclusion}$ = 8.252 * 10^7. Finally, the model with the above terms included was the best fit, BF_M = 8.252 * 10^7, and significant overall, BF_{10} = 6.162 * 10^{10}. The main results are shown in Fig. 4.

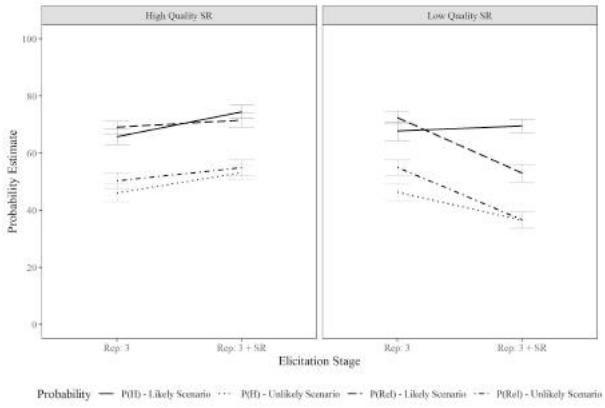


Fig. 4: P(Rel) and P(H) given shared reliability

Discussion and concluding remarks

Despite the prevalence of dissent in public debates, the role of minority dissenters has not been adequately explored or modelled. This is a crucial function if to understand the functional impact of dissent in debates such as climate change or political predictions.

The paper tests how people revise beliefs in the reliability of sources and a hypothesis given sequential reports. The two initial reports support the hypothesis while the 3rd report rejects it. First, P(Rel) initially decreases when the source provides a positive report for an unlikely hypothesis, but rebounds when the 2nd source corroborates the initial report. Additionally, P(H) increases for the same reports while P(Rel) does not change for predicting the *likely* hypothesis while P(H) increases slightly. This replicates findings from Madsen et al (2018) and follows Bayesian predictions.

Second, the negation of the hypothesis yielded novel results. In both scenarios, negation decreased P(Rel) for all sources, presumably as it introduces noise and uncertainty. P(H) decreases when the 3rd source rejects the hypothesis for the unlikely scenario, but does not decrease for the likely scenario. This suggests that while participants revise their belief in an unlikely hypothesis (a market crash within six months), they decrease their belief in the hypothesis when dissent is voiced against this idea. Comparatively, P(H) does not decrease with dissent in the likely scenario. Rather, P(Rel) for $source_3$ decreases significantly given rejection of the likely hypothesis. P(Rel) also decreases for $sources_{1\text{-}2}$ given dissent in the likely case, but not to the same extent as is suffered by the dissenter.

Finally, shared reliability appears to work asymmetrically for consenters and dissenters. If the school enjoys a good reputation, perceived reliability increases for consenters, but less so for dissenters. If the shared reliability is perceived as high quality, people's degree of belief in the hypothesis additionally increases. However, for both perceived source reliability and the hypothesis, we see a decrease when the shared reliability is of poor quality.

In all, the study suggests people are sensitive to source reliability as well as the structural relationship between the sources. Belief revisions generally follow Hahn et al. (2009) such that positive reports from very credible sources lend credence to the hypothesis. Additionally, participants update perceived source reliability in accordance with predictions, as supporting unlikely hypotheses is initially detrimental, but sequentially rebounds given corroboration. Finally, perceived partial dependence is crucial, as shared reliability moderates perceptions of the hypothesis and the reliability of all sources involved. In all, the study provides additional support for a Bayesian approach to source reliability.

References

Bovens, L., & Hartmann, S. (2003). *Bayesian epistemology.* Oxford: Oxford University Press.

[7] $BF_{Inclusion}$ shows the change in odds from the sum of the prior probabilities of models including the effect, to the sum of the posterior probabilities of models including the effect.

[8] BF_M shows the change from prior to posterior odds, given the model.

Chaiken, S. & Maheswaran, D. (1994) Heuristic Processing Can Bias Systematic Processing: Effects of Source Credibility, Argument Ambiguity, and Task Importance on Attitude Judgement, *Journal of Personality and Social Psychology* 66 (3), 460-473

Cialdini, R. B. (2007) *Influence: The Psychology of Persuasion*, Collins Business

Cuddy, A. J. C., Glick, P. & Beninger, A. (2011) The dynamics of warmth and competence judgments, and their outcomes in organizations, *Research in Organizational Behavior* 31, 73-98

Fiske, Susan T., Cuddy, A. J. C. & Click, P. (2007) Universal dimensions of social cognition: warmth and competence, *Trends in Cognitive Sciences* 11 (2), 77-83

Hahn, U., Harris, A. J. L., & Corner, A. (2009) Argument content and argument source: An exploration, *Informal Logic 29,* 337-367

Hahn, U., & Oaksford, M. (2006) A normative theory of argument strength, *Informal Logic* 26*,* 1-24

Hahn, U., & Oaksford, M. (2007) The rationality of informal argumentation: A Bayesian approach to reasoning fallacies, *Psychological Review 114*, 704-732

Harris, A. J. L., Hahn, U., Madsen, J. K., & Hsu, A. S. (2015). The Appeal to Expert Opinion: Quantitative support for a Bayesian Network Approach. *Cognitive Science* 40*,* 1496-1533

Harris, P. L., & Corriveau, K. H. (2011). Young children's selective trust in informants, *Philosophical Transactions of the Royal Society B, 366,* 1179-1187

Howson, C., & Urbach, P. (1996). *Scientific Reasoning: The Bayesian Approach (2nd Edition)*. Chicago, IL: Open Court

JASP Team (2018). JASP (Version 0.9)[Computer software].

Lagnado, D. A., Fenton, N., & Neil, M. (2013). Legal idioms: a framework for evidential reasoning, *Argument & Computation* 4 (1), 46-63.

Madsen, J. K. (2016) Trump supported it?! A Bayesian source credibility model applied to appeals to specific American presidential candidates' opinions, Papafragou, A., Grodner, D., Mirman, D., & Trueswell, J.C. (Eds.) *Proceedings of the 38th Annual Conference of the Cognitive Science Society*, Austin, TX: Cognitive Science Society, 165-170

Madsen, J. K., Hahn, U. & Pilditch, T. (2018) Partial source dependence and reliability revision: the impact of shared backgrounds*, Proceedings of the 40th Annual Conference of the Cognitive Science Society*

Oaksford, M. & Chater, N. (1991) Against logicist cognitive science, *Mind and Language* 6, pp. 1-38

Oaksford, M. & Chater, N. (2007) *Bayesian Rationality: The probabilistic approach to human reasoning*. Oxford, UK: Oxford University Press

Perfors, A., Navarro, D. J. & Shafto, P. (2018) Stronger evidence isn't always better:A role for social inference in evidence selection and interpretation, *Proceedings of the 40th Annual Conference of the Cognitive Science Society*

Petty, R. E. & Cacioppo, J. T. (1984) Source Factors and the Elaboration Likelihood Model of Persuasion*, Advances in Consumer Research* 11, 668-672

Tormala, Z. L. & Jackson, J. J. (2007) Assimilation and Contrast in Persuasion: The Effects of Source Credibility in Multiple Message Situations, *Personality and Social Psychology Bulletin* 33 (4), 559-571

Whalen, A., Griffiths, T. L. & Buchsbaum, D. (2013) Sensitivity to shared information in social learning*, Cognitive Science* 42, 168-187

Source reliability and the continued influence effect of misinformation: A Bayesian network approach

Jens Koed Madsen (jens.madsen@ouce.ox.ac.uk)
School of Geography and the Environment, University of Oxford
OX1 3QY, South Parks Road, Oxford, United Kingdom
Orcid: 0000-0003-2405-8496

Saoirse Connor Desai (saoirse.connor-desai@city.ac.uk)
Department of Psychology, City, University of London
Rhind Building, St John Street, Clerkenwell, London, UK

Toby Pilditch (t.pilditch@ucl.ac.uk)
School of Geography and the Environment, University of Oxford
OX1 3QY, South Parks Road, Oxford, United Kingdom
&
Department of Experimental Psychology, University College London
26 Bedford Way, WC1H 0AP, London, United Kingdom

Abstract

Misinformation, and its impact on society, has become an increasingly topical field of study of late. A body of literature exists that suggests misinformation can retain an influence over beliefs despite subsequent retraction, known as the Continued Influence Effect (CIE). Researchers have argued this to be irrational. However, we show using a Bayesian formalism why this argument is overly assumptive, pointing to (previously overlooked) considerations of reliability of, and dependence between, misinforming and retracting sources. We demonstrate that lay reasoners intuitively endorse assumptions that demarcate CIE as a rational process, based on the fact misinformation *precedes* its retraction. Moreover, despite using established CIE materials, we further upturn the applecart by finding participants show CIE, and appropriately penalize the reliabilities of contradicting sources.

Keywords: Continued Influence Effect; Negation; Reliability; Dependency; Reasoning

Introduction

Misinformation can have a lasting effect on beliefs that people entertain and on the inferences they can make about events[1]. Poor information, whether spread deliberatively or mistakenly, can have serious and widespread repercussions for society. For example, despite being corrected repeatedly, some people belief that there is a causal link between the measles mumps and rubella (MMR) vaccination and autism.

This belief persists in some communities despite scientific evidence refuting the myth (Horne et al., 2015). Decreased acceptance of the MMR vaccination has contributed to a 7% drop in vaccination rates in the UK and a 1.7-fold increase in refusal to vaccinate in the US (Smith et al., 2008), and consequently, an increase in a vaccine-preventable disease.

The harmful effects of misinformation and ineffectiveness of attempts to correct mistaken beliefs have become a great concern for contemporary society (Gordon, Brooks, Quadflieg, Ecker, & Lewandowsky, 2017; Lewandowsky, Ecker, Seifert, Schwarz, & Cook, 2012), and has recently become a weighty issue for governments, media organizations, and citizens (see Lewandowsky et al., 2017). Problematically, though, studies show that belief in erroneous information can persist even after it has been unambiguously corrected (Lewandowsky et al., 2012)
Regardless of how information is corrected, research shows that it often fails to abolish the effects of misinformation (see Lewandowsky et al., 2012 for review). The so-called *Continued Influence Effect* (CIE) of misinformation refers to the consistent finding that information initially presented as true continues to influence beliefs and reasoning despite clear and credible corrections (Ecker et al., 2011a, 2011b; Johnson & Seifert, 1994; Rich & Zaragoza, 2016).

In the paper, we explore two aspects of CIE. First, no normative account of how people should "optimally" process corrections to misinformation has been provided to date. CIE studies typically report the observed phenomenon in a variety of contexts and settings. To explore the effect systematically, we provide a Bayesian Network model to test whether CIE is truly irrational or if the phenomenon can be explained rationally. Second, past research shows the importance of dependency (Madsen et al., 2018). That is, whether a source is truly independent from another source, or if they are somehow related. This influences the impact of the report on the hypothesis *and* perceived reliability. In accordance with these studies, we manipulate the source of debunking such that the initial source debunks its own statement or a different source debunks the statement.

[1] We define information as any piece of information or evidence that is initially thought to be true, but which later turns out to be erroneous, but which can be corrected. Going beyond the current study, the intention behind the dissemination of misinformation is crucial (e.g. the difference between an honest mistake and a malevolent lie – both of which may provide poor information).

Exploring CIE through a formal reasoning model yields interesting results. First, we find a rational explanation for CIE. We show that belief in the hypothesis remains above prior level, but instead the reliability (in the second reporter case) is penalized. Second, perceived dependence influences the effect. Given a Bayesian network, CIE is irrational only insofar that the sources are entirely *independent* of each other. Comparatively, when considering reports temporally and dependent, CIE is entirely *rational*. Correcting is often done by a source that is, in some way, linked with the initial source of misinformation (e.g. a reporter working at the same network). This highlights a significant conceptual limitation to the way in which CIE is framed classically.
Finally, we can demonstrate irrationality in a manner that is backwards to what is typically reported in CIE studies. In CIE studies, people should not stick with original beliefs given correction, but do so any way. We show cases where there are reasonable grounds for why people should stick with their original beliefs, but do not.

The continued influence effect

Continued influence studies examine corrections to misinformation using variants of a laboratory paradigm first developed by Wilkes and Leatherbarrow (1988; but also see Johnson & Seifert, 1994). There are two leading cognitive explanations for CIE (Gordon et al., 2018; Lewandowsky et al., 2012):

First, ***the selective retrieval account*** argues that CIE occurs when correct and incorrect information are stored in memory simultaneously, and misinformation is activated but inadequately blocked (Ecker et al., 2011a). Second, ***the model updating account*** argues that people continually construct a mental event model as new information becomes available. Correcting information without providing a credible alternative (e.g. a competing causal explanation) leaves people with a gap in their mental model. On this view, people prefer a coherent but incorrect model to a correct but incomplete one and thus maintain the invalidated information (Ecker et al. 2010; Johnson & Seifert, 1994).

A typical CIE task involves a series of sequentially presented statements describing an unfolding event, similar to a breaking news report. Misinformation that allows inferences to be drawn about the outcome of the event is presented early in the sequence, but retracted later. Participants' event comprehension is assessed, typically to show that misinformation continues to influence people's inferential reasoning even though they clearly understand and remember that the information was corrected (Johnson & Seifert, 1994). The effect persists even when given prior warnings about the persistence of misinformation (Ecker et al., 2010). The fact that retractions are often ineffective at 'removing' misinformation from people's understanding of events emphasizes the need to identify and model factors that contribute to the *Continued Influence Effect*.

Sustained reliance on misinformation given a retraction is often depicted as a bias – or systematic deviation from a normative standard – and therefore irrational (e.g. Lewandowsky et al., 2012). This perspective assumes two things; first, that the optimal solution is always to disregard initially prior information in favour of new information, and second that the 'true' value of the retraction is known.

Source reliability

Establishing a source's reliability is critical when deciding whether to rely on the information conveyed to us by other people, and may drive the CIE. Reliability can be separated into issues of: i) observational sensitivity, ii) objectivity, and iii) veracity (Schum, 1994). For example, jurors must establish whether a witness' testimony is truthful and accurate in order to reach a verdict, and voters must similarly place their confidence in the statements of politicians when deciding who to vote for.

While appeals to authority and reliance on testimonies traditionally have been seen as fallacoius (ad verecundiam) or as a shallow cue, Bayesian models have integrated reliability within a normative theory of reasoning (Bovens & Hartmann, 2003; Hahn et al,. 2009; Harris et al., 2015).

People use a range of cues to evaluate a source's reliability. For example, in the legal domain witnesses may contradict themselves or be contradicted by others, which may reassess the credibility (see Connor Desai et al., 2016). Moderating perceived source reliability is an sensible act if new information, additional contradictory or corroborative reports, or insight into whether or not the sources are related to each other is made known. In addition to new infomraiton, source dependecy moderates perceived reliability (Bovens & Hartmann, 2003; Madsen et al., 2018).

Contradiction is particularly relevant to CIE studies where the misinformation and its retraction are typically issued by the same source. A source who announces that they previously gave incorrect information may appear less reliable than one who does not. Consistent with this, one CIE study found that distrust in the source of the retraction was a primary reason for disbelieving the retraction (Guillory & Geraci, 2010; 2013). Indeed, Lewandowsky et al., (2012) argue that source reliability (high and low) may facilitate 'tagging' of correct and incorrect information and facilitate retrieval of information when this information is made salient.

Thus, perceived reliability moderates the degree to which people are willing to integrate reports from more or less reliable sources. If a highly reliable source provides report about an issue, the recepient should *normatively* revise her belief in the suggested direction. Second, reports from independent sources are more diagnostic than reports that stem from sources who share a common background. In order to model reliability estimates, belief in the hypothess, and to develop a formal model of CIE, we adopt a Bayesian approach.

A Bayesian approach to source reliability

As mentioned, CIE studies do not provide a normative account of how people should process retractions to misinformation. The lack of formalism is crucial as there

may be situations where continued reliance on misinformation is rational given the lack of information and inherent uncertainty of the situation. In such situations, people may use cues like reliability to assess the validity of misinformation and its retraction, and decide how much to incorporate these pieces of information into their beliefs.

Bayes' theorem gives a normative belief revision model. It integrates people's subjective prior degrees of belief with the likelihood ratio to estimate the posterior degree of belief. It has been applied to conditional reasoning (Oaksford & Chater, 2007), argumentation (Hahn & Oaksford, 2006; 2007), and other areas of cognition (Chater et al., 2010).

To explore CIE formally, we use a Bayesian Network (BN) framework (Pearl, 2000). BNs use graph structures to represent the probabilistic relationships between hypotheses and evidence (including reliability), using conditional probabilities to represent the strength of relations, and show what inferences are rationally permitted from a given model given available information. This is an ideal method for examining whether CIE is rational in some circumstances, as it provides the means to test causal models of scenarios – including their models of the reliability of the sources providing information – and compare inferences to a normative standard (Fenton et al, 2013).

Congruency of information with the misinformation and the reliability of sources providing the misinformation or the retraction are potential moderators of the CIE. BNs provide a formal model to test responses against model predictions and test foundational assumptions of the CIE.

Comparing judgments to Bayesian predictions test if there are situations in which retaining belief in misinformation after a retraction is rational. Formally modelling the causal relations between information included in a scenario would make it possible to test participants' causal models of scenarios. This provides an understanding of the cognitive mechanisms involved in the CIE.

Method

Participants: 101 participants were recruited from Prolific Academic (71 females, age = 31.57±9.6). Participants were paid £1.50 (~$1.97) and took 14 minutes (on average) to complete the experiment.

Stimuli, Design & Procedure: To replicate CIE studies, we used stimuli adapted from past research (Johnson & Seifert, 1994; Gordon et al., 2017, see Table 1 for an example of stimulus material).

Table 1: Example of news report and comprehension probes

Sentence	Control	Retraction (Same Source)	Retraction (Different Source)
Example News Report			
Sentence 1	*A motorcyclist died yesterday after being knocked off his bike by a car.*		
Sentence 2	***Officer Jones reported that the driver of the car had been travelling over the speed limit.***	***Officer Jones reported that the driver of the car was intoxicated.***	***Officer Jones reported that the driver of the car was intoxicated.***
Sentence 3	The accident happened on the A7 north of Carlisle.		
Sentence 4	The motorcyclist was 30 years old and had two children.		
Sentence 5	***Officer Jones revealed that the car driver was not intoxicated.***	***Officer Jones revealed that the car driver was not intoxicated.***	***Officer Smith revealed that the car driver was not intoxicated***
Sentence 6	The driver of the car was also injured in the incident.		
Example Comprehension Probes			
Question 1	Drink-driving charges should be brought against the driver of the car		
Question 2	The driver should be forced to complete a drink-driving awareness course		
Question 3	A breathalyser would have returned a positive result		

It was a between-subjects study with the effect of retracting information was assessed between groups (Control, Retraction – Same Source, Retraction – Different Source). Participants were randomly assigned to a condition.

Sentence 2 differed between control and retraction conditions for each event. In retraction conditions, sentence 2 contained (mis)information. In the control condition, it contained circumstantial information to provide a baseline for the comprehension test. The key sentence (sentence 5) was identical in all conditions. Given exposure to sentence 2, sentence 5 did or did not correct previous information. For source conditions, the source of the (mis)information (sentence 2) and retraction (sentence 5) were either the same (same source) or different (different source).

In all, we tested four scenarios. Presentation order of the scenarios was randomized across participants. The scenarios used were selected from a set of eight pilot reports (N = 70)

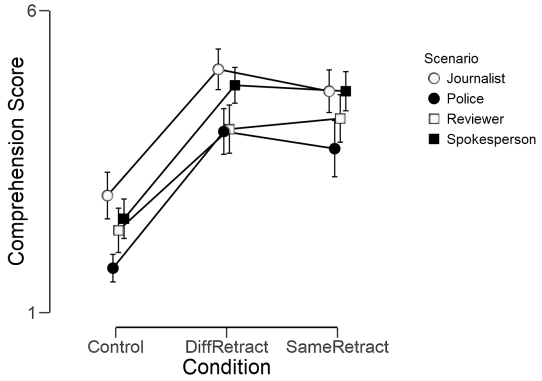


Figure 1. Mean comprehension scores, split by scenario (line) and condition (horizontal axis). Error bars reflect 95% CI.

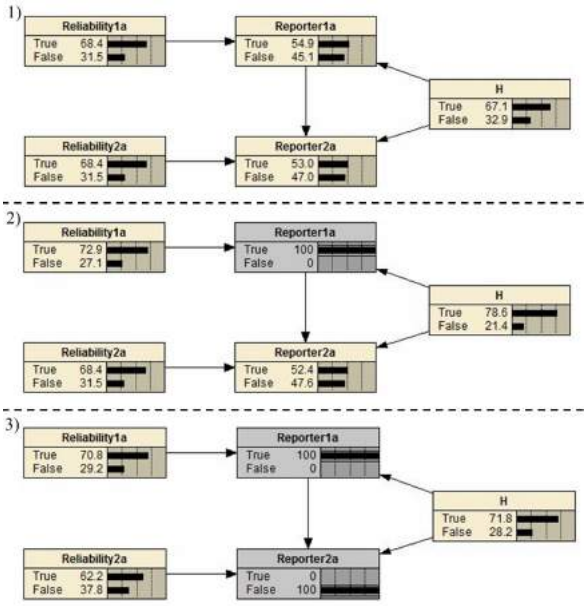


Figure 2a. Group BN model for the retraction different condition, police officer scenario. 1) Baseline (no observation) stage, 2) Single positive (first) report stage (i.e. control condition), and 3) Final (retraction) state given a second, separate reporter.

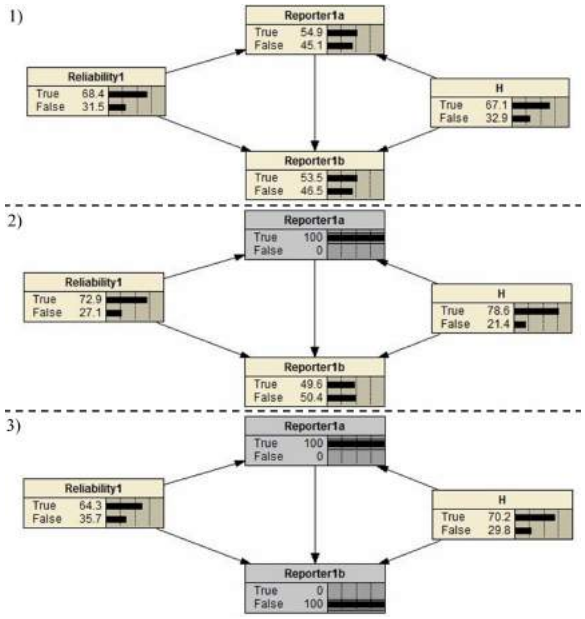


Figure 2b. Group BN model for the retraction same condition, police officer scenario. 1) Baseline (no observation) stage, 2) Single positive (first) report stage (i.e. control condition), and 3) Final (retraction) state given a second, report from the same reporter.

where scenarios with the largest 'continued influence effect' of misinformation were chosen for the actual study.

Prior to reading any scenario, participants provided prior estimates for their beliefs in the reliability of the sources of misinformation that would appear in the subsequent reports and whether they would provide reliable reports. This was measured on a scale of 0 (Extremely unlikely) to 100 (Extremely likely).

Further, to parameterise the model, participants provided six conditional probability estimates per report (24 in total). They rated their belief that the source of report 1 would make an erroneous statement in reporting about an event, if they were or were not reliable on the same scale as used for prior beliefs. Questions about the second reporter differed between the same and different source conditions. Eliciting conditional probabilities allowed for parameter-free models.

Continued reliance of misinformation was measured by a set of comprehension probes that followed each scenario (see Table 1). Participants rated each probe on a 7-point scale from 'strongly disagree' to 'strongly agree'. In line with previous CIE methods, probes referred to the critical information (sentence 5). Higher endorsement of comprehension probes measured the degree to which the misinformation presented in sentence 2 had been incorporated into a participants' understanding of the report.

After rating the probes participants provided their belief posterior probability on a similar scale used for prior beliefs. For example, in the scenario in Table 1, participants were asked: 1) Given everything you know so far about the incident in question, how likely do you think it is that the accident occurred because the driver was intoxicated/travelling over the speed limit? 2) Given everything you know so far about the incident in question, how likely to do you think it is that the police officer is reliable in their reporting? Participants who received a retraction from a different source as the misinformation provided an additional estimate for the reliability of the second reporter.

Results

Bayesian analyses were done with JASP statistical software (JASP Team, 2018) and assumed an uninformed prior.

Comprehension Scores

A Bayesian repeated measures ANOVA was used to determine the effect of condition and scenario type on mean comprehension scores. Strong evidence was found for the main effect of condition, $BF_{Inclusion} = 1.917 * 10^{12}$, and scenario, $BF_{Inclusion} = 5.44 * 10^{9}$, but no interaction, $BF_{Inclusion} = 0.122$. The model including just main effects was the strongest fit, $BF_M = 131.26$, and significant overall, $BF_{10} = 2.105 * 10^{22}$. As illustrated in Fig. 1 below, scenarios differed in comprehension scores from one another, and there was a differential influence of condition.

Critically, the effect of condition indicated significantly higher endorsement of comprehension probes following the presentation and retraction of misinformation compared to when no misinformation was presented at all. This indicates that, a CIE was observed across all scenarios, such that a retraction was insufficient to bring endorsement ratings back to baseline.

Bayesian Model Fits

Using the conditional probabilities and priors elicited from participants, group means on these estimates were used to parameterize 2 group-condition models for each scenario. Although the conditional probabilities and priors for each first reporter and reliability node were fitted based on all participants, two important exceptions are noted. First, conditional probabilities for the second reporter were based solely on estimates from the condition of relevance (i.e. only estimates from the retraction different condition were used to parameterize the entailed different second reporter in that condition). Secondly, prior probabilities for each hypothesis were reverse-engineered (via Bayes Theorem) using the posteriors provided by control condition. More precisely, taking the control condition BN model, the posterior for the hypothesis was fitted, given the single positive report. Retracting the observation could reveal the approximate prior (absent observations) for that hypothesis. This "prior" was fitted into the models for the two retraction conditions.

Figs 2a and 2b show example condition models for the Police officer scenario, fitted from participant data according to the protocol outlined above. Several important trends are noticeable:

Firstly, as expected, given a single positive reporter (stage 2), belief in the hypothesis (H) increases, and the predicted likelihood of corroboration from the second report increases. However, when the second, contradicting report is observed (stage 3), the belief in the hypothesis (H) does *not* return to prior (stage 1) levels. Instead, the reliability of sources decreases given the contradiction, this decrease is strongest in the second reporter (different condition), but is also substantial when the same reporter contradicts themselves (Fig. 2b, stage 2 to stage 3).

Critically, the reason for this effect (retention of belief in H, but reduction in perceived reliability) is due to the capturing of the temporal dependence from first to second report. Put another way, the models capture the intuition that a second report is aware of the first report (whether internally in the case of the same reporter condition, or via general narrative in the different reporter condition). The manner and strength of this influence is then captured by the elicited conditional probabilities from participants.

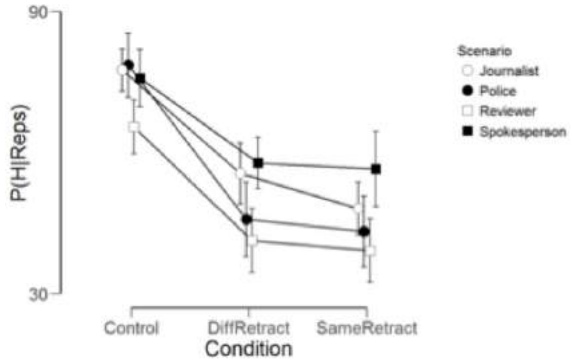


Figure 4. Posterior estimates of belief in the hypothesis (H), given all reports, split by scenario (line) and condition (horizontal axis). Error bars reflect 95% CI

Participant Estimates

Returning to participant data, we again use Bayesian repeated measures ANOVA to examine whether probability estimates correspond to the BN model predictions (and thus map onto a continued influence effect), or corroborate the comprehension score measures (and indicate an absence of CIE – against fitted normative prescription).

Hypothesis. Turning first to posterior estimates of belief in the hypothesis, we find main effects of condition, $BF_{Inclusion} = 3.328 * 10^9$, and scenario, $BF_{Inclusion} = 41812.52$, but no interaction, $BF_{Inclusion} = 0.467$. The model consisting of the main effects along was the strongest fit, $BF_M = 34.27$, and significant overall, $BF_{10} = 2.247 * 10^{14}$. As Fig. 4 illustrates, these effects corroborate comprehension scores, wherein the effect of condition is driven by a reduction in belief in the hypothesis from control to retraction conditions. Crucially, this shows that participants generally deviate from the *prescribed* CIE effect entailed by the BN models, decreasing belief in the hypothesis below the control condition (and prior), given the retraction.

Reliability. Turning next to estimates of reliability, we add to the repeated measures ANOVA analysis a within-subject factor of prior to posterior. Here we find significant main effects of condition (control > retraction different and same), $BF_{Inclusion} > 1.00 * 10^{20}$, scenario, $BF_{Inclusion} = 124.44$, and prior-posterior (posterior < prior), $BF_{Inclusion} > 1.00 * 10^{20}$. Figs 5a-5c illustrate the significant interaction of condition and prior-posterior, $BF_{Inclusion} > 1.00 * 10^{20}$, wherein reliability estimates increased in the control condition (Fig. 5a; where no contradiction occurs, and in line with the increase observed in Fig. 3a and 3b, stage 2), but decreased in both retraction conditions (Fig. 5b and 5c; also in line with model predictions illustrated in Fig. 3a and 3b, stage 3). Lastly, a significant interaction of scenario and prior-posterior was also observed, $BF_{Inclusion} = 75.92$, wherein the spokesperson scenario entailed smaller changes from prior to posterior than the 3 remaining scenarios. The model including the above significant terms yielded the strongest fit, $BF_M = 484.97$, and was significant overall, $BF_{10} = 1.559 * 10^{28}$.

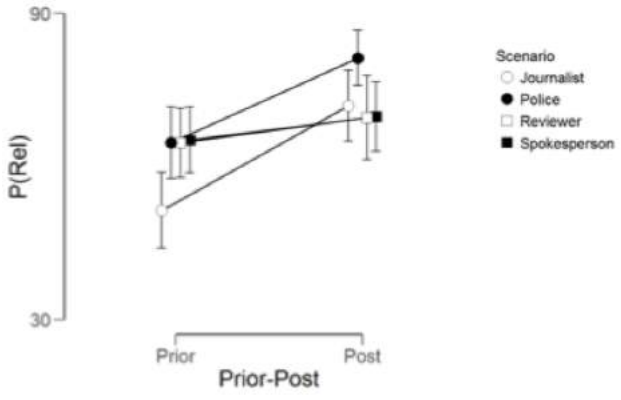


Figure 5a. Control condition reliability estimates for reporters from prior to posterior (reports observed), split by scenario (lines). Error bars reflect 95% CI.

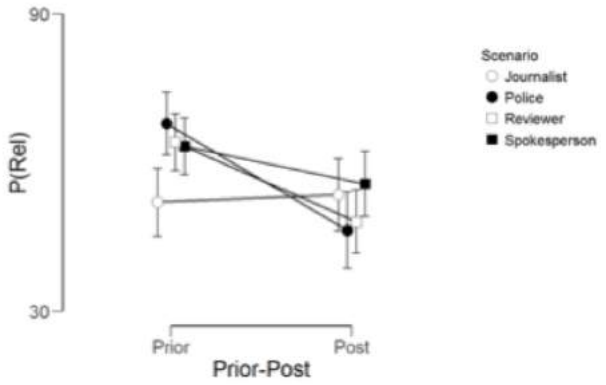


Figure 5b. Retraction different condition reliability estimates for reporters from prior to posterior (reports observed), split by scenario (lines). Error bars reflect 95% CI.

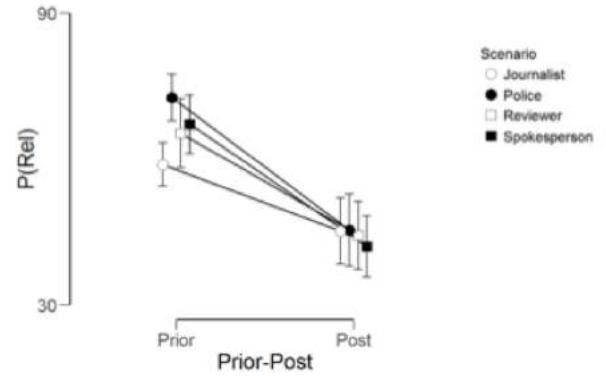


Figure 5c. Retraction same condition reliability estimates for reporters from prior to posterior (reports observed), split by scenario (lines). Error bars reflect 95% CI.

Finally, we note that the retraction condition showed no significant difference in posterior reliability estimates between the two different (first and second) reporters, $BF_{10} = 0.135$, contrary to model predictions (wherein the second reporter should be more substantially penalized).

Discussion and concluding remarks

Previous CIE studies have consistently found that misinformation continues to be influence beyond a clear and credible retraction (e.g. Ecker et al., 2010; 2011a; 2011b; Johnson & Seifert, 1994; Rich & Zaragoza, 2016). Continued reliance on misinformation after a retraction has been depicted as a bias and therefore irrational (Lewandowsky et al., 2012). However, there is an argument that people should exhibit CIE if source reliability judgments are incorporated into how beliefs about misinformation are updated following a retraction.

This paper's aim was to formally model CIE, using a Bayesian Network framework, to capture the temporal dependency between misinformation and its retraction, and the impact this may have on source reliability. We compared participants' judgments to Bayesian predictions to establish whether retaining belief in misinformation (hypothesis) after a retraction is, in fact, sometimes rational.

Participants rated their belief in the hypothesis, and the reliability of sources, when there was no retraction of misinformation, when the retraction was offered by the same source as the misinformation, or by a different source than the misinformation, for a series of news reports.

Behavioural measures showed the standard CIE across all scenarios. Comprehension of the news reports was measured to establish whether misinformation had been incorporated into participant's understanding of the report despite having been retracted. A classic CIE was observed whereby misinformation continued to influence news report comprehension despite being retracted. The effect was observed whether the retraction was offered by the same or a different source to the misinformation.

We also find a rational explanation for CIE. Qualitatively we show that belief in the hypothesis remains above prior level, but instead the reliability of the second reporter (i.e. the retraction) is penalised. Participant's posterior estimates also decreased below their priors, and against what their model predicts. This finding is contrary to the typical account of CIE that people continue to rely on retracted misinformation even though they should. Instead, suggesting that people should continue to rely on misinformation but do not!

Focusing on the condition in which misinformation and retraction come from the same source, participants decrease their estimate for the reporter after they have contradicted themselves, in line with model predictions. In the different source condition, participants decrease their estimates in the reliability of the first reporter (which is incorrect according to the model), and increase reliability estimates of the second reporter (which is correct according to the model). Interestingly, the second reporter was considered more reliable than the first in the police officer and reviewer scenarios (against model predictions), but less reliable than the first in the journalist and spokesperson scenarios (in line with model predictions).

Taking together, we show that participants *should in fact* exhibit a CIE effect (according to fitted Bayesian Network models), and although we find this effect in with standard behavioural measures, we do not observe this with novel probability estimate (P(H) measures. Yet, we do find appropriate penalization in reliability estimates given a contradiction among reports – something hitherto unnoticed in CIE studies, but predicted by our formalism.

To conclude, this research provides a formal account of CIE using the BN framework, and shows that continued reliance on misinformation is in some circumstances rational. This approach captures the qualitative inferences participants make about the reliability of sources of who provide contradictory information. These findings also suggest that perceived reliability moderates the degree to which people are willing to integrate reports from more or less reliable sources.

References

Bovens, L., & Hartmann, S. (2003). *Bayesian epistemology*. Oxford: Oxford University Press.

Chalnick, A., & Billman, D. (1988). Unsupervised learning of correlational structure. *Proceedings of the tenth annual conference of the cognitive science society* (pp. 510-516). Hillsdale, NJ: Lawrence Erlbaum Associates.

Chater, N., Oaksford, M., Hahn, U., & Heit, E. (2010). Bayesian models of cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(6), 811-823.

Cook, J., Ecker, U., & Lewandowsky, S. (2015). Misinformation and How to Correct It. In *Emerging Trends in the Social and Behavioral Sciences* http://doi.org/10.1002/9781118900772.etrds0222.

Connor Desai, S., Reimers, S & Lagnado, D. (2016). Consistency and credibility in legal reasoning: A Bayesian network approach. In Proceedings of the 38th Annual Conference of the Cognitive Science Society, pp. 626-631.

Ecker, U. K., Hogan, J. L. & Lewandowsky, S (2017) Reminders and repetition of misinformation: Helping or hindering its rectraction? *Journal of Applied Research in Memory & Cognition* 6 (2), 185-192

Ecker, U. K. H., Lewandowsky, S., & Tang, D. T. W. (2010). Explicit warnings reduce but do not eliminate the continued influence of misinformation. *Memory & Cognition*, *38*(8), 1087–1100.

Ecker, U. K., Lewandowsky, S., & Apai, J. (2011). Terrorists brought down the plane!—No, actually it was a technical fault: Processing corrections of emotive information. *The Quarterly Journal of Experimental Psychology*, *64*(2), 283–310.

Ecker, U. K., Lewandowsky, S., Swire, B., & Chang, D. (2011). Correcting false information in memory: Manipulating the strength of misinformation encoding and its retraction. *Psychonomic Bulletin & Review*, *18*(3), 570–578.

Fenton, N., Neil, M. & Lagnado, D. A. (in press) A General Structure for Legal Arguments About Evidence Using Bayesian Networks, *Cognitive Science*

Gordon, A., Brooks, J. C., Quadflieg, S., Ecker, U. K., & Lewandowsky, S. (2017). Exploring the neural substrates of misinformation processing. *Neuropsychologia*, *106*, 216–224.

Hahn, U., Harris, A. J. L., & Corner, A. (2009) Argument content and argument source: An exploration, *Informal Logic 29,* 337-367

Hahn, U., & Oaksford, M. (2006). A Bayesian approach to informal argument fallacies. *Synthese*, *152*(2), 207-236.

Hahn, U., & Oaksford, M. (2007). The rationality of informal argumentation: A Bayesian approach to reasoning fallacies. *Psychological review*, *114*(3), 704.

Harris, A. J. L., Hahn, U., Madsen, J. K., & Hsu, A. S. (2015). The Appeal to Expert Opinion: Quantitative support for a Bayesian Network Approach. *Cognitive Science* 40*,* 1496-1533

Horne, Z., Powell, D., Hummel, J. E., & Holyoak, K. J. (2015). Countering antivaccination attitudes. *Proceedings of the National Academy of Sciences*, *112* (33), 10321–10324.

Johnson, H. M., & Seifert, C. M. (1994). Sources of the continued influence effect: When misinformation in memory affects later inferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(6), 1420–1436.

Lewandowsky, S., Ecker, U. K., & Cook, J. (2017). Beyond Misinformation: Understanding and Coping with the "Post-Truth" Era. *Journal of Applied Research in Memory and Cognition*, *6*(4), 353–369.

Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, *13*(3), 106–131.

Madsen, J. K., Hahn, U. & Pilditch, T. (2018) Partial source dependence and reliability revision: the impact of shared backgrounds*, Proceedings of the 40th Annual Conference of the Cognitive Science Society*

Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford University Press.

Pearl, J. (2000) *Causality: models, reasoning and inference*, Cambridge University Press

Rich, P. R., & Zaragoza, M. S. (2016). The continued influence of implied and explicitly stated misinformation in news reports. *Journal of Experimental Psychology. Learning, Memory & Cognition*, *42*(1), 62–74.

Schum, D.A. (1994). *The Evidential Foundations of Probabilistic Reasoning*, Northwestern University Press.

Smith, M. J., Ellenberg, S. S., Bell, L. M., & Rubin, D. M. (2008). Media coverage of the measles-mumps-rubella vaccine and autism controversy and its relationship to MMR immunization rates in the United States. *Pediatrics*, *121*(4), 836–843.

Wilkes, A. L., & Leatherbarrow, M. (1988). Editing episodic memory following the identification of error. *The Quarterly Journal of Experimental Psychology Section A*, *40*(2), 361–387.

Effect of Suggestions from a Physically Present Robot on Creative Generation

Akihiro Maehigashi (ak-maehigashi@kddi-research.jp)
Interaction Design Laboratory, KDDI Research, Inc., Japan

Yugo Hayashi (y-hayashi@acm.org)
College of Comprehensive Psychology, Ritsumeikan University, Japan

Abstract

This study experimentally investigated the effect of suggestions from a physically present robot on human creative generation. In the experiments, we used a creative task in which the participants were required to draw creatures living on a planet other than the Earth, and a physically present robot, which provided suggestions for creative drawing to the participants with speech sounds and physical movements. First, the results of the pilot experiment confirmed that drawing creativity was enhanced for the participants supported by a robot; however, they were unlikely to refer to the suggestions. Based on the results, two hypotheses were developed: the suggestions from a robot offered a variety of different perspectives and facilitated metacognition (Hypothesis 1), and the suggestions worked as distractions and suppressed fixated perspectives (Hypothesis 2). The experiment was conducted to investigate these hypotheses. As a result, Hypothesis 1 was supported. The results were discussed based on previous studies.

Keywords: Robot; Human-robot interaction; Creativity; Creative generation; Metacognition; Collaboration.

Introduction

Creative generation and collaboration

Creative generation is performed in various situations, such as engineers thinking of new information tools, novelists thinking of new stories, and chefs thinking of new recipes. Guilford (1979) showed that creative generation involved two types of thinking processes: divergent and convergent thinking. Divergent thinking is the process of generating multiple possible ideas. By contrast, convergent thinking is the process of examining the generated ideas to determine the best. The ideas would be refined by alternately repeating these two thinking processes. Also, Finke, Ward, and Smith (1992) developed the geneplore model of creative generation in which there are generative and exploratory phases. In the generative phase, abstract representations of ideas called preinventive forms are created. Following the generative phase, in the exploratory phase, the generated ideas are interpreted in a meaningful ways for specific purposes. The ideas become sophisticated as these two phases are repeated one after the other.

The scope of creative generation can be limited because people generate ideas based on existing representations of prior knowledge (Ward, 1994). Therefore, representational change, or re-representation, in divergent thinking or the generative phase is crucially important. It occurs when a representation described from a certain perspective is reinterpreted from a different perspective (Ward, Smith, & Finke, 1999).

In the field of cognitive science, many previous studies have shown that collaborative activities provide opportunities for people to develop new perspectives. For example, people reinterpret and deepen their knowledge by providing explanations about their knowledge and asking reflective questions to each other (Miyake, 1986). Also, in a collaborative problem solving situation, people develop abstract representation of the solution by alternately taking the roles of a task-doer, who externalizes their own ideas, and a task-monitor, who objectively reflects the others' ideas (Shirouzu, Miyake, & Masukawa, 2002). Moreover, people can acquire an integrated perspective of multiple viewpoints by taking a perspective from others that is incompatible with their own perspective (Hayashi, 2018). These previous studies show that it is important to interact with others to facilitate metacognition and form new perspectives that cannot be achieved alone.

Human-robot interaction

The development of technology has brought the prevalence of robots that support human physical and cognitive activities (e.g. Ros, Baroni, & Demiris, 2014; Saerbeck, Schut, Bartneck, & Janse, 2010). However, there are not many studies that experimentally investigated how robotic support influences human cognitive activities during human-robot interaction.

Leyzberg, Spaulding, Toneva, and Scassellati (2012) experimentally investigated how advice from a robot influenced human problem-solving performance with a nonogram puzzle. Their study compared the effects of advice from a physically present robot, a robot displayed on a screen, and only auditory sound. As a result, the participants supported by the physically present robot solved the puzzle faster than the participants supported by the displayed robot and auditory sound after receiving advice multiple times.

Moreover, a physically present robot gave better impressions to people than a virtually displayed robot or animated character. In particular, people felt the robot was more likable, helpful, enjoyable, trustworthy, creditable, and informative (Kidd & Breazeal, 2004; Powers, Kiesler, Fussell, & Torrey, 2007). Also, people became more compliant with a physically present robot than to a robot displayed on a screen (Bainbridge, Hart, Kim, & Scassellati, 2011). These effects were considered to occur because of the robot's physical presence (Powers et al., 2007). On the other hand, it was more difficult for people to recall suggestions from a physically present robot than suggestions from a robot displayed on a screen. Since people tended to allocate their attention to the presence of the robot, they were considered to allocate less attention to the contents of the suggestions and had diffi-

culty recalling the suggestions (Powers et al., 2007). These previous studies showed that suggestions from a physically present robot provided different effects on people from those provided from a robot or animated character displayed on a screen and those in text or auditory sound.

Purpose of this study

The number of studies related to human-robot interaction has been increasing. However, there are still few studies that investigated human cognitive activities supported by a physically present robot. In particular, not much is known about how a robot could support human creativity. The focus of this study was on human creative generation and how suggestions from a robot influenced creativity.

Pilot experiment

The pilot experiment was conducted to confirm the effect of the suggestions from a physically present robot on creative generation and develop experimental hypotheses about the features of the suggestions.

Experimental task

The task used in the pilot experiment was a creative task used by Ward (1994). The participants were required to draw creatures living on a planet other than the Earth.

The participants draw creatures on a canvas displayed on a computer screen (Figure 1a) with a digital pen. The canvas was created with HTML5 Canvas and JavaScript. The participants could choose one of two colors, black and white, to draw a line by physically tapping one of the square boxes on the display. Also, they could change the line width and the level of the transparency by tapping and moving the slider bars before drawing the line. The software provided a redo button to redo drawing a line, a delete button to delete all the drawn lines on the canvas, and a submit button to save a drawn creature as a picture file and to delete the creature from the canvas.

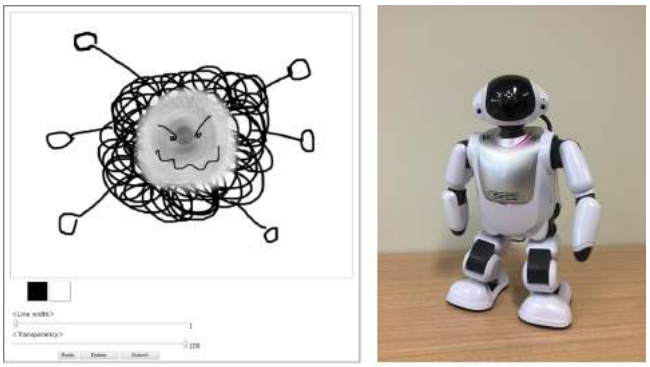
Method

Participants Thirty university students participated in the pilot experiment as volunteers.

Experimental design The experiment had a one-factor between participants design. The factor was the type of suggestions (no-, text-, and robot-suggestions).

In the robot-suggestion condition, the robot, Palmi by DMM.com LLC, was used (Figure 1b). The robot gave suggestions for creative drawing to the participants with speech sounds and the physical movements of moving the arms, legs, or head according to entered commands. Also, the no-suggestion condition was set up as a control condition in which the participants performed the task without suggestions.

Moreover, the text-suggestion condition was set up in which the participants were given the same suggestions as in the robot-suggestion condition. However, the suggestions were displayed in letters in the lower right corner of the display. Because text information allows people to carefully consider the meaning compared to auditory information (Blasio & Milani, 2008), the suggestions in text were presumed to be actively referenced and thus enhanced the creativity of the drawing.



(a) Canvas (b) Robot

Figure 1: (a) A canvas displayed on a computer screen and (b) the robot used in the experiments.

Suggestions For the robot- and text-suggestion conditions, 20 suggestions were created to encourage the participants to use metacognition and consider their ideas from a variety of different perspectives in divergent thinking or the generative phase. Some of the suggestions are shown in Table 1.

Procedure Ten participants were randomly assigned to each condition. First, the experimental task was explained to the participants. After the participants received the explanation of the drawing operations, they practiced drawing pictures for five minutes. Following the practice, in the robot-suggestion condition, the participants were told that the robot in front of them would give them suggestions for creative drawings during the task. Also, in the text-suggestion condition, the participants were told that the suggestions for creative drawings would be displayed on the screen during the task. After the explanation and instructions, the participants performed the task for 20 minutes. All participants were instructed to draw as many creative creatures as possible.

In the text- and robot-suggestion conditions, 10 suggestions were randomly selected from the 20 suggestions for each participant and given in randomized order every two minutes from the beginning of the task. The participants in these conditions were instructed to refer to the given suggestions as necessary.

After the task was finished under the robot- and the text-suggestion conditions, the participants rated to what degree the suggestions were referred to in order to draw creative creatures with a 5-point scale (1: not referred at all - 5: extremely referred).

Results

First, the average number of drawn creatures in each condition was 2.40 for robot-suggestion, 2.60 for no-suggestion, and 2.80 for text-suggestion conditions. A one-way analysis of variance (ANOVA) showed no significant differences in the number of drawn creatures between the three conditions ($F(2,27) = 0.53, p = .60$). The result showed that the participants drew creatures to the same extent in the three conditions.

Second, the creativity of the creatures was rated on originality using a 10-point scale (1: not original at all - 10: extremely original). Three independent raters who knew nothing about the experiment were trained and then rated the originality of all creatures in randomized order. The rated scores between the three raters were judged consistent ($\alpha = .69$).

Based on the originality scores for each drawn creature in the three conditions, a one-way ANOVA was performed (Figure 2). As a result, there was a significant main effect ($F(2,27) = 14.50, p < .001$). A multiple comparison test with Ryan's method revealed that the score was significantly higher in the robot- and text-suggestion conditions than in the no-suggestion condition ($t(48) = 3.68, p < .001$; $t(52) = 3.46, p < .001$). However, there was no significant difference between the robot- and text-suggestion conditions ($t(50) = 0.36, p = .72$).

Moreover, a t-test was performed to compare the reference ratings between the robot- and text-suggestion conditions (Figure 3). As a result, the rating was significantly lower in the robot-suggestion condition than in the text-suggestion condition ($t(18) = 3.82, p < .001$).

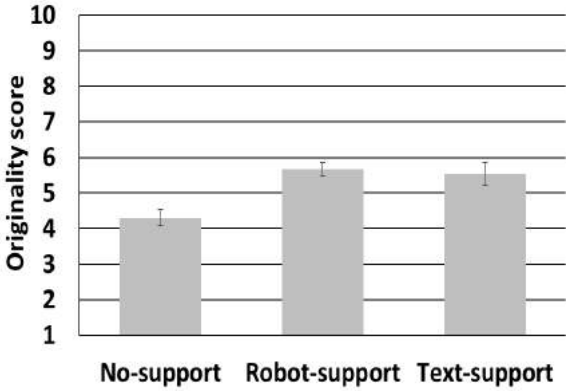


Figure 2: Average originality score in each condition in the pilot experiment. The error bars indicate the standard error.

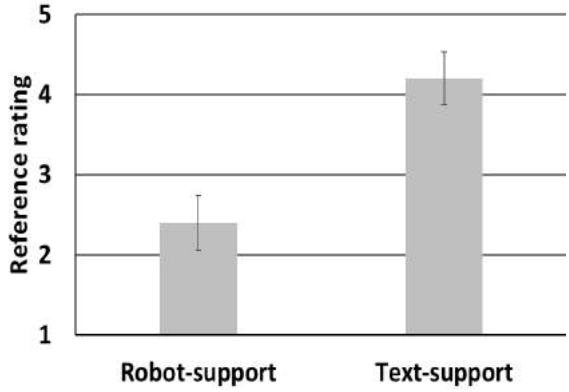


Figure 3: Average reference rating in each condition in the pilot experiment. The error bars indicate the standard error.

Discussion

First, the results confirmed that the suggestions from a robot enhanced the creativity of drawings. Second, the participants referred to the suggestions less frequently when given from the robot than when displayed in text. In addition, the suggestions in text were actively referred to and enhanced the creativity as predicted.

Although the participants were unlikely to refer to the suggestions from the robot over all, only some of the suggestions might be referred to and encouraged the participants to use metacognition and generate ideas from a variety of different perspectives. However, there is another possibility that the suggestions from the robot enhanced the creativity of drawings by causing irrelevant distractions.

Because the suggestions from the robot were less likely to be referred to, the suggestions might have tended to distract the participants from focusing on the task. In creative generation, irrelevant distractions can be beneficial in suppressing fixated perspectives and focusing on irrelevant information (Amer, Campbell, & Hasher, 2016; Dijksterhuis & Meurs, 2006). Therefore, the suggestions from the robot were assumed to work as irrelevant distractions and supported the participants in suppressing fixated perspectives and ideas. As a result, they might generate ideas from new perspectives and enhance the creativity of drawings.

In the following experiment, we conducted an experiment to investigate the features of the suggestions from a robot with considerations of facilitating metacognition and causing distractions.

Experiment

Method

Participants Sixty-seven university students participated in this experiment for course credit.

Experimental design The experiment had a one-factor between participants design. The factor was the frequency of the suggestions (high and low). The frequency of the suggestions was manipulated by the number of suggestions provided during the task. In the high-frequency condition, 12 suggestions were given every two minutes for 24 minutes. Conversely, in the low-frequency condition, six suggestions were given every four minutes for 24 minutes.

Suggestions and distractions In this experiment, two different situations were set up: the no-distraction and distraction situations. In the no-distraction situation, all suggestions provided during the task were related to drawing creative

creatures (Table 1). They were selected from the suggestions used in the pilot experiment. Contrarily, the distraction situation was set up to provide apparent distractions, suggestions completely unrelated to drawing creative creatures, in order to enhance the effect of distractions (Table 2). In the distraction situation, half of the suggestions were selected from the list in Table 1, and the other half were selected from the list in Table 2. If the suggestions had prevented the participants from focusing on the task and enhanced creativity, the effect of distractions would have appeared prominently in the distraction situation.

Table 1: A list of suggestions

	Suggestions related to drawing creative creatures
1	Let's think about the shape of the creature.
2	Let's think about what kind of features the creature would have.
3	What kind of environment does the creature live in?
4	Let's think about what the creative creature would be.
5	Let's think about the movement of the creature.
6	Let's reconsider the idea.
7	Let's think about incidents that occur outside of Earth.
8	Let's think in a different way.
9	How about combining different ideas?
10	Let's think in different perspectives.
11	Let's think about something that could be referred to.
12	What kind of features would the creature have?

Table 2: A list of distractions

	Distractions
1	Look up to the ceiling and count 10 seconds as accurately as possible.
2	Close your eyes and count 10 seconds as accurately as possible.
3	Raise your feet and count 10 seconds as accurately as possible.
4	Let's do a mental calculation. What is eight plus six minus seven? (Silence for 3 seconds) The answer is seven.
5	Let's do a mental calculation. What is four plus nine minus five? (Silence for 3 seconds) The answer is eight.
6	Let's do a mental calculation. What is seven plus five minus nine? (Silence for 3 seconds) The answer is three.

Procedure The participants were randomly assigned to each condition in each situation. As a result, 16 participants were assigned to the low-frequency condition in the distraction situation and 17 participants were assigned to the other conditions. All the participants performed the task with the robot.

The task and the procedure were the same as in the pilot experiment. However, in this experiment, although the task display was the same as in the pilot experiment, an iPad by Apple Inc. was used to draw the creatures. Also, each task took 24 minutes. The suggestions or distractions were randomly chosen for each participant and given in randomized order.

After the task was finished, in addition to the reference rating, the participants in the distraction situation rated to what degree the suggestions and distractions were followed with a 5-point scale (1: not followed at all - 5: extremely followed).

Hypothesis

In this experiment, the following two hypotheses were examined in each of the no-distraction and distraction situations.

Hypothesis 1: The suggestions from a robot enhance creativity by facilitating metacognition.

Hypothesis 2: The suggestions from a robot enhance creativity by causing irrelevant distractions.

If Hypothesis 1 were confirmed, the participants would refer to the suggestions and generate creative ideas from the perspectives of the suggestions. There would be more opportunities for the participants to achieve helpful suggestions in the high-frequency condition than in the low-frequency condition. Therefore, in the both no-distraction and distraction situations, the participants in the high-frequency condition would refer to the suggestions more frequently and draw more creative creatures than those in the low-frequency condition.

Contrarily, if Hypothesis 2 were confirmed, the suggestions would distract the participants and enhance creativity; therefore, the suggestions would be unlikely to be referred to in order to draw creative creatures. There would be more opportunities for the participants to be distracted in the high-frequency condition than in the low-frequency condition. Thus, in the both no-distraction and distraction situations, the participants in the high-frequency condition would draw more creative creatures than those in the low-frequency condition; however, they would refer to the suggestions as frequently as those in the low-frequency condition.

Results

The average number of drawn creatures in the no-distraction situation was 9.29 for the high-frequency and 10.41 for the low-frequency condition. Also, the average number in the distraction situation was 6.94 for the high-frequency and 8.56 for the low-frequency condition. The results of t-tests showed that there was neither significant difference in the number of drawn creatures between the two conditions in the no-distraction situation ($t(32) = 0.98, p = .33$) nor in the distraction situation ($t(31) = 2.18, p = .05$). These results showed that the participants drew creatures to the same extent in the two conditions in each situation.

Also, the result of a t-test showed that there was no significant difference in the rating, what degree the suggestions and distractions were followed, between the high-frequency ($M = 4.05$) and low-frequency ($M = 3.88$) conditions in the distraction situation ($t(31) = 0.49, p = .63$). The result showed that the participants followed the suggestions and distractions to the same extent in the two conditions in the distraction situation.

For the analysis of the hypotheses, first, the originality of the creatures was rated in the same way as the pilot experiment. Three independent raters different from those in the

pilot experiment were trained and then rated all creatures in randomized order. The rated scores between the three raters were judged consistent ($\alpha = .72$).

Next, the average originality score of each participant was calculated in each condition, and a t-test was performed on the score in each situation (Figure 4). The results revealed that in the no-distraction situation, the score was significantly higher in the high-frequency condition than those in the low-frequency condition ($t(32) = 3.62, p < .001$). In contrast, in the distraction situation, there was no significant difference between the two conditions ($t(31) = 0.07, p = .94$).

Moreover, a t-test was performed to compare the reference ratings between the two conditions in each situation (Figure 5). The results indicated that in the no-distraction situation, the rating was significantly higher in the high-frequency condition than in the low-frequency condition ($t(32) = 2.51, p < .05$). On the other hand, in the distraction situation, there was no significant difference between the two conditions ($t(31) = 0.09, p = .93$).

The results in the no-distraction situation supported Hypothesis 1. However, the results in the distraction situation did not support neither Hypothesis 1 nor 2.

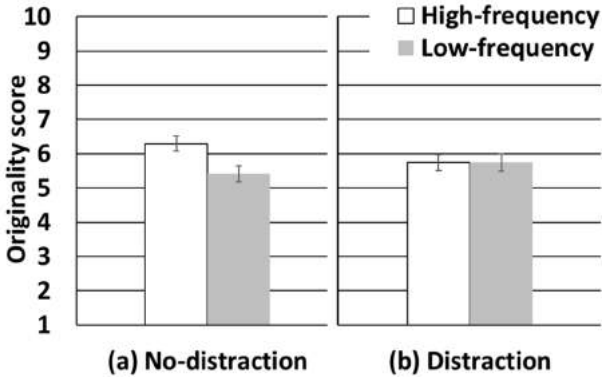


Figure 4: Average originality score of each condition in (a) no-distraction and (b) distraction situations. The error bars indicate the standard error.

Discussion

In the no-distraction situation, the participants in the high-frequency condition referred to the suggestions more frequently and created more original creatures than those in the low-frequency condition. This result supported Hypothesis 1, that is, the suggestions from the robot enhanced creativity by offering a variety of different perspectives to generate ideas and facilitate metacognition.

However, the effect of facilitating metacognition was not found in the distraction situation. This might be because the number of the suggestions related to drawing creative creatures was too small in the high-frequency condition, and therefore, there were not enough opportunities to facilitate metacognition. Also, the effect of causing distractions was not found in the distraction situation. Baird et al. (2012)

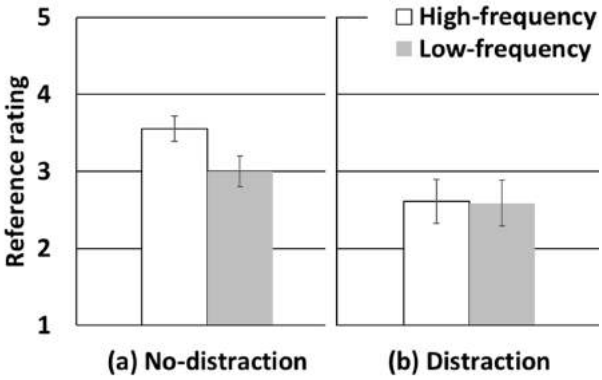


Figure 5: Average reference rating of each condition in (a) no-distraction and (b) distraction situations. The error bars indicate the standard error.

showed that distractions which require light cognitive load enhance creative generation. In their experiment, the participants who performed the undemanding preceding task, which required cognitive load light enough to elicit mind wandering, enhanced creativity in the following creative task, the unusual uses task, more than those who performed the demanding preceding task, which required more cognitive load. The distractions provided in this study might be too demanding for the participants to enhance creativity.

General discussion

Robots have been developed for a variety of applications. However, there have only been a few studies that investigated how a physically present robot could support human cognitive activities. This study focused on creative generation to investigate how suggestions from a robot would influence human creative generation. The results of the experiment revealed that the suggestions from a robot enhanced creative generation by offering a variety of different perspectives.

In human-human collaboration, representational change occurs when people reflect their own and the other's ideas or knowledge by asking, explaining, or externalizing (e.g. Miyake, 1986). The robot used in this study was not interactive; however, representational change might be caused by the suggestions in the same manner as in the previous studies of human-human collaboration. In particular, the participants were considered to refer to some of the suggestions from the robot and reflect their own ideas according to the suggestions.

Moreover, Okada and Ishibashi (2017) showed that in a creative drawing situation, new perspectives in drawing were acquired by copying and viewing other's unfamiliar artworks, and the creativity of drawings increased. However, in their experiment, a human verbal suggestion, which recommended creating original and creative drawings in different styles, did not enhance the creativity of drawings. In contrast to the previous study, in this study, the suggestions from a robot with speech sounds enhanced the creativity of drawings. Since the robot provided multiple different types of suggestions during

the task, at least some of them were assumed to encourage the participants to consider their ideas from the viewpoints of the suggestions.

Furthermore, in the pilot experiment of this study, although the suggestions from the robot enhanced creativity, they were less referred to than suggestions in text form. In contrast, Leyzberg et al. (2012) showed that the advice from a physically present robot enhanced human problem solving performance and indicated a possibility that people might perceive the authority or social standing of a physically present robot and take their advice seriously.

This difference was assumed to happen because of the difference in the interactivity of the robots. In the previous study, the robot provided the advice according to the time required to solve the problem. On the other hand, in this study, the robot provided the suggestions without consideration of the participants. Thus, the participants in this study might not have perceived the sociality or interactivity of the robot to take the suggestions seriously as in the previous study. Another possibility related to the type of task was also considered. In the previous study, a well-defined problem, nonogram puzzle, was used as the task. Because there were clear solving strategies, the relevant advice about the strategies could be provided to participants. In contrast, in this study, an ill-defined problem, creative drawing, was used as the task. Since there were several possible and different perspectives to take for the creative drawings, there was a possibility that many of the suggestions from the robot did not match their ideas and likely were ignored during the task.

Finally, in this study, the suggestions from a robot were made to facilitate metacognition. However, the enhanced creativity observed in this study needs to be ensured as the result of facilitated metacognition. The results in this study could not deny the possibility that the suggestions facilitated other types of cognitive processes involved in creative generation and enhanced creativity. Therefore, in our future study, we will investigate how each suggestion from a robot influences cognitive process and creativity.

Acknowledgements

We would like to thank Takahiro Tanaka, College of Comprehensive Psychology, Ritsumeikan University, for his contribution in this study.

References

Amer, T., Campbell, K. L., & Hasher, L. (2016). Cognitive control as a double-edged sword. *Trends in Cognitive Sciences*, *20*, 905–915.

Bainbridge, W. A., Hart, J. W., Kim, E. S., & Scassellati, B. (2011). The benefits of interactions with physically present robots over video-displayed agents. *International Journal of Social Robotics*, *3*, 41–52.

Baird, B., Smallwood, J., Mrazek, M. D., Kam, J. W. Y., Franklin, M. S., & Schooler, J. W. (2012). Inspired by distraction: Mind wandering facilitates creative incubation. *Psychological Science*, *23*, 1117–1112.

Blasio, D. P., & Milani, L. (2008). Computer-mediated communication and persuasion: Peripheral vs. central route to opinion shift. *Computers in Human Behavior*, *24*, 798–815.

Dijksterhuis, A., & Meurs, T. (2006). Where creativity resides: The generative power of unconscious thought. *Consciousness and Cognition*, *15*, 135–146.

Finke, R. A., Ward, T. B., & Smith, S. M. (1992). *Creative cognition: Theory, research, and applications*. Cambridge, MA: The MIT Press.

Guilford, J. P. (1979). *Cognitive psychology with a frame of reference*. San Diego, CA: Edits Publishers.

Hayashi, Y. (2018). The power of a "maverick" in collaborative problem solving: An experimental investigation of individual perspective-taking within a group. *Cognitive Science*, *42*, 69–104.

Kidd, C. D., & Breazeal, C. (2004). Effect of a robot on user perceptions. In *2004 ieee/rsj international conference on intelligent robots and systems (iros)* (Vol. 4, pp. 3559–3564). New York, NY: IEEE.

Leyzberg, D., Spaulding, S., Toneva, M., & Scassellati, B. (2012). The physical presence of a robot tutor increases cognitive learning gains. In *Proceedings of the 34th annual meeting of the cognitive science society (cogsci2012)* (pp. 1882–1887). Austin, TX: Cognitive Science Society.

Miyake, N. (1986). Constructive interaction and the iterative process of understanding. *Cognitive Science*, *10*, 151–177.

Okada, T., & Ishibashi, K. (2017). Imitation, inspiration, and creation: Cognitive process of creative drawing by copying others' artworks. *Cognitive Science*, *41*, 1804–1837.

Powers, A., Kiesler, S., Fussell, S., & Torrey, C. (2007). Comparing a computer agent with a humanoid robot. In *Proceedings of the acm/ieee international conference on human-robot interaction* (pp. 145–152). New York, NY: ACM Press.

Ros, R., Baroni, I., & Demiris, Y. (2014). Adaptive human robot interaction in sensorimotor task instruction: From human to robot dance tutors. *Robotics and Autonomous Systems*, *62*, 707–720.

Saerbeck, M., Schut, T., Bartneck, C., & Janse, M. D. (2010). Expressive robots in education: Varying the degree of social supportive behavior of a robotic tutor. In *Proceedings of the 28th acm conference on human factors in computing systems (chi2010)* (pp. 1613–1622). New York, NY: ACM Press.

Shirouzu, H., Miyake, N., & Masukawa, H. (2002). Cognitively active externalization for situated reflection. *Cognitive Science*, *26*, 469–501.

Ward, T. B. (1994). Structured imagination: The role of category structure in exemplar generation. *Cognitive Psychology*, *27*, 1–40.

Ward, T. B., Smith, S. M., & Finke, R. A. (1999). Creative cognition. In R. Sternberg (Ed.), *Handbook of creativity.* New York: NY: Cambridge University Press.

EARSHOT:
A minimal network model of human speech recognition that operates on real speech

James S. Magnuson (james.magnuson@uconn.edu)

Heejo You (hee_jo.you@uconn.edu)

Jay Rueckl (jay.rueckl@uconn.edu)

Paul Allopenna (paul.allopenna@uconn.edu)

Monica Li (monica.li@uconn.edu)

Sahil Luthra (sahil.luthra@uconn.edu)

Rachael Steiner (rachael.steiner@uconn.edu)
Psychological Sciences & CT Institute for the Brain and Cognitive Sciences, U. Connecticut, Storrs, CT 06269-1020

Hosung Nam (nam@haskins.yale.edu)
Korea University, Seoul, Korea, and Haskins Laboratories, New Haven, CT 06511

Monty Escabi (monty.escabi@uconn.edu)
Biomedical Engineering & Psychological Sciences, University of Connecticut, Storrs, CT 06269-3247

Kevin Brown (kevin.brown@oregonstate.edu)
Depts. of Pharmaceutical Sciences and Chemical, Biological, and Environmental Engineering, Oregon State University

Rachel Theodore (rachel.theodore@uconn.edu)

Nicholas Monto (Nicholas.monto@uconn.edu)
Speech, Language & Hearing Sciences, University of Connecticut, Storrs, CT 06269

Abstract

Despite the *lack of invariance problem* (the many-to-many mapping between acoustics and percepts), we experience *phonetic constancy* and typically perceive what a speaker intends. Models of human speech recognition have side-stepped this problem, working with abstract, idealized inputs and deferring the challenge of working with real speech. In contrast, automatic speech recognition powered by *deep learning* networks have allowed robust, real-world speech recognition. However, the complexities of deep learning architectures and training regimens make it difficult to use them to provide direct insights into mechanisms that may support human speech recognition. We developed a simple network that borrows one element from automatic speech recognition (*long short-term memory* nodes, which provide dynamic memory for short and long spans). This allows the network to learn to map real speech from multiple talkers to semantic targets with high accuracy. Internal representations emerge that resemble phonetically-organized responses in human superior temporal gyrus, suggesting that the model develops a distributed phonological code despite no explicit training on phonetic or phonemic targets. The ability to work with real speech is a major advance for cognitive models of human speech recognition.

Keywords: spoken word recognition; computational models; neural networks; deep learning

Introduction

Human speech recognition (HSR) poses some of the greatest unsolved scientific challenges in the cognitive and neural sciences. Despite a many-to-many mapping between acoustic patterns and percepts (for now, let us assume percepts are phonemes, i.e., consonants and vowels), listeners experience *phonetic constancy*: we hear what the speaker intends even though the same acoustic pattern can cue different phonemes depending on context, and different patterns can cue the same phoneme. This challenge is the *lack of invariance problem*.

Many factors complicate the acoustic-perceptual mapping: (a) coarticulation (temporal and articulatory overlap of phonemes in series; Liberman et al., 1967), (b) lack of robust boundaries between phonemes or words (Cole & Jakimik, 1980), and (c) shifts in the mapping due to variation in speaking rate (Miller & Baer, 1983), talker characteristics (Joos, 1948; Peterson & Barney, 1952), phonetic context (Liberman et al., 1967), coarticulation (Liberman et al., 1952), and novelty of message content (Fowler & Hosum, 1987). Similar problems are found in other perceptual domains (e.g., visual objects must be recognized despite variation in size, rotation, and illumination; DiCarlo & Cox, 2007). However, the temporal and transient nature of speech compounds the challenge.

Deep vs. minimal networks for speech recognition

One might suppose that the lack of invariance problem has been solved in contemporary automatic speech recognition (ASR) systems, such as those used daily by billions of smartphone users. The deep-learning neural network models underlying the best ASR (Hinton et al., 2012) provide robust

real-world application but little guidance for theories of HSR. Deep nets for ASR require many complex and richly connected layers, as well as complex, carefully engineered training regimens.

That said, researchers interested in HSR have developed less complex deep networks with the aim of illuminating possible mechanisms supporting audition and HSR. Nagamine et al. (2015), for example, examined hidden units of a 5-layer network trained explicitly on phoneme recognition and observed responses strikingly similar to phonetically-structured responses in human superior temporal gyrus (Mesgarani et al., 2014). Kell et al. (2018) used a deep network to achieve human-like accuracy on two unusual tasks: (1) recognizing the word at the *center* of a two second sample of speech and (2) musical genre identification. Their network had many layers and required complex training. The first 7 layers were shared for speech and music, but then it branched into specialized speech and music pathways (with 5 additional layers). The model surpassed standard spectrotemporal filter models of auditory cortex in predicting human cortical responses to natural sounds (measured with fMRI. Kell et al. suggested that deep networks might provide the only computational approach able to achieve human-like performance for natural stimuli.

We optimistically disagree. Our aim is to develop maximally simple (minimal) models of HSR. Theoretical progress will be difficult if our models approach the complexity of their biological target (the neural basis for HSR). At the same time, we aim to grapple with details that have been left out of deep learning models of auditory perception. First, several models have achieved high accuracy by side-stepping the temporal nature of speech (e.g., by treating an utterance or sound as a static image, with time as one axis) rather than as a time series. Furthermore, such models have not addressed the kinds of human data of greatest interest to psycholinguists who study human spoken word recognition, such as the time course of lexical activation and competition (Allopenna et al., 1998).

Simpler shallow computational models have been applied to grappled with over-time inputs and time course of lexical competition, but with two different limitations: (1) they do not use real speech as input (instead using, for example, |abstract distributed phonetic features over time (TRACE: McClelland & Elman, 1986) or human diphone confusion probabilities (Shortlist B: Norris & McQueen, 2008); (2) they tend not to address learning. Models developed since the mid 1980s have either adopted these simplifications in order to address the time course of spoken word recognition with large vocabularies, or have strived for greater realism but in small-inventory models (e.g., Grossberg et al., 1997), or have attempted to incorporate ASR approaches into cognitive models of spoken word recognition (e.g., Scharenborg, 2010; Scharenborg et al., 2005). Such approaches have led to genuine insights, but the models tend to have low accuracy, limited empirical coverage, or both.

Minimal models from long short-term memory nodes
Our aim is to develop a *minimal* cognitive model of HSR that could *learn to map over-time speech to semantics, without explicit phonetic training,* that remains simple enough to generate hypotheses for mechanisms that could support HSR. However, current network-based cognitive models of HSR do not appear adequate for processing real speech.

Thus, we examined a variety of network architectures and elements used in network models used for ASR. We found that a two-layer recurrent network provides the needed power for our goal domain if its hidden units are *long short-term memory* (LSTM) nodes (Hochreiter & Schmidhuber, 1997). LSTM nodes add 3 internal gates and a memory cell that allow nodes to develop sensitivity to information over long time scales, mitigating the *vanishing gradient problem* (Hochreiter et al., 2001). In the following sections, we describe a new neural network model of HSR, EARSHOT (*Emulation of Auditory Recognition of Speech by Humans Over Time*), that we believe approaches the minimal complexity required to map real speech to semantics.

Methods

Network structure and parameters

The EARSHOT network is schematized in Fig. 1. Its 256 input units are fully connected to 512 LSTM hidden units. The hidden layer is fully recurrent (i.e., every unit has a connection to every other unit). A *tanh* activation function is applied to hidden outputs. The hidden units are fully connected to 300 output units. High accuracy on our task (described below) required ~500 hidden units (performance is not improved by increasing to 750 or 1000 hidden nodes).

Materials

We pseudo-randomly selected 1000 words from a list of uninflected English words, with the constraints that (a) word length varied from 1-8 phonemes (mean = 5.5) and (b) every phoneme had to occur in at least 10 words. We created speech files for each of the 1000 words pronounced by 10 talkers in the Apple text-to-speech application, *say* (5 females [Agnes, Kathy, Princess, Vicki, Victoria] and 5 males [Alex, Bruce, Fred, Junior, Ralph]). Mean duration was 659 ms (range: 289-1121 ms). We also created 360 consonant-vowel (CV) and VC syllables for testing purposes (using 15 vowels and 24 consonants). Sound files were converted to spectrographic representations with 256 channels in 10 ms steps with sampling rate of 8000 hz.

We created random sparse vectors for each word as a proxy for semantic representations. Vectors had 300 elements, with 10 "on" (set to 1, others set to 0). This common simplification is considered acceptable given the largely arbitrary mapping from form to menaing (e.g., Lazlo & Plaut, 2012).

Training method

We trained 10 instantiations of EARSHOT. For each model, a different one of the 10 talkers was excluded from training (reserved to test generalization to a novel talker). We excluded 100 different randomly selected words from each trained-on talker (reserved to test generalization to unseen

items from trained-on talkers). So for each model, the training set was 8100 input-output patterns, with all 10,000 pairs included for testing.

Each training epoch included one presentation of each of the 8100 training items in random order with no pause or other indication of word boundaries. The target pattern was the semantic vector for the current word, and it was compared to the output at each time step. To enhance learning, we used *minibatch gradient descent*, *Noam decay*, and *Adam optimizing* (Vaswani et al., 2017). Full details are available in a longer preprint (Magnuson et al., 2018). Connections were trained using backpropagation through time (Werbos, 1988). Training accuracy largely plateaued by 8000 epochs. We then resumed training with formerly excluded talkers included. The logic was that when humans encounter new talkers, we presumably learn to adapt to them by learning any idiosyncratic aspects of their acoustics-to-percepts mapping (e.g., by using lexical hypotheses to guide learning). In simple tests of generalization, the model cannot learn. We continued training for another 2000 epochs (8001-10,000).

Testing method Every 1000 epochs, models were tested with all 10,000 words (including excluded words and talkers). Successful recognition was operationalized as the output vector's cosine similarity to the target exceed any other item's cosine similarity to the output by at least 0.05 for at least 100 ms, and subsequently, no item could exceed the target's cosine similarity to the output before word offset.

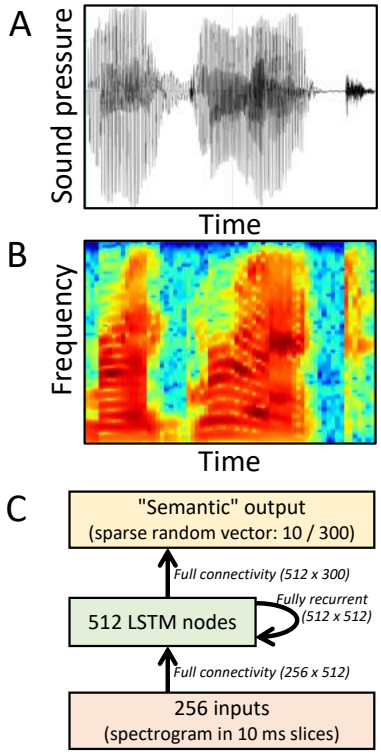


Figure 1. Model input and structure. (A) Audio files are converted to spectrograms **(B),** with 256 channels (rows) in 10 ms steps (columns). Color indicates amplitude (blue-red indicates low-high). **(C)**. The model is a standard recurrent network, except "long short-term memory" nodes are used in the hidden layer, allowing it to become sensitive to multiple temporal grains.

Replicability We trained all 10 models 3 times; only minor variations were observed between iterations. We present results from the first run of each model in this report.

Hardware and software Simulations were conducted on a Windows 10 workstation with an i7-6700k CPU, 64-gb of RAM, and a Titan-X (12-gb) graphics card. Simulations were implemented using Python 3.6 and TensorFlow 1.7. Each model required approximately 10 hours for training.

Alternative architectures In developing EARSHOT, we explored dozens of combinations of candidate architectures and model elements. We limited networks to 2 layers of forward connections (inputs→hidden→outputs). We varied 3 aspects of models: number of hidden units (typically from 100 to 1000 nodes before rejecting a model if accuracy plateaued below 90%), hidden unit type (standard integrative nodes vs. LSTMs), and degree of recurrence (full recurrence, as in the model reported here, vs. single-step recurrence, as in simple recurrent networks; Elman, 1990). For inputs, we explored spectrograms at various resolutions, Mel Frequency Cepstral Coefficients (MFCCs), and cochleagrams. Most combinations failed to achieve high accuracy. Aside from the model reported here, the only combinations that achieved greater than 90% accuracy was an MFCC-based model that failed to show human-like time course despite high accuracy. Note that this does not mean that only a single set of parameters worked; the model described above begins achieving high accuracy with more than 256 LSTM hidden units, and *maximal* accuracy with ~500 or more LSTM nodes.

Results

Accuracy and time course

We present key model behavior results in Fig. 2. Mean accuracy on training items was quite high (88%) after 8000 epochs. Accuracy was 67% for excluded words from trained-on talkers but only 33% for excluded talkers, with a very wide range (4% to 78%). When training resumed with all talkers and items included, performance improved rapidly (to 89% and 86% for excluded words and talkers, respectively, 93% for previously trained-on items).

Next, we consider the challenge of simulating the time course of HSR (Allopenna et al., 1998). This is a central behavioral target in psycholinguistics but has not been addressed in deep learning models of speech (Kell et al., 2018; Nagamine et al., 2015). Our minimal model exhibits the correct qualitative pattern for phonological competition (Fig. 2B) and makes predictions similar to the gold-standard of HSR, TRACE (Fig. 2C; McClelland & Elman, 1986). This similarity might suggest that any model that can map speech inputs to word-form outputs (as in TRACE) or semantic outputs (EARSHOT) would exhibit this human-like time course. However, this is not the case. As we noted above, an MFCC-based model was able to achieve high accuracy, but could not simulate the patterns seen in Figs. 2B and 2C.

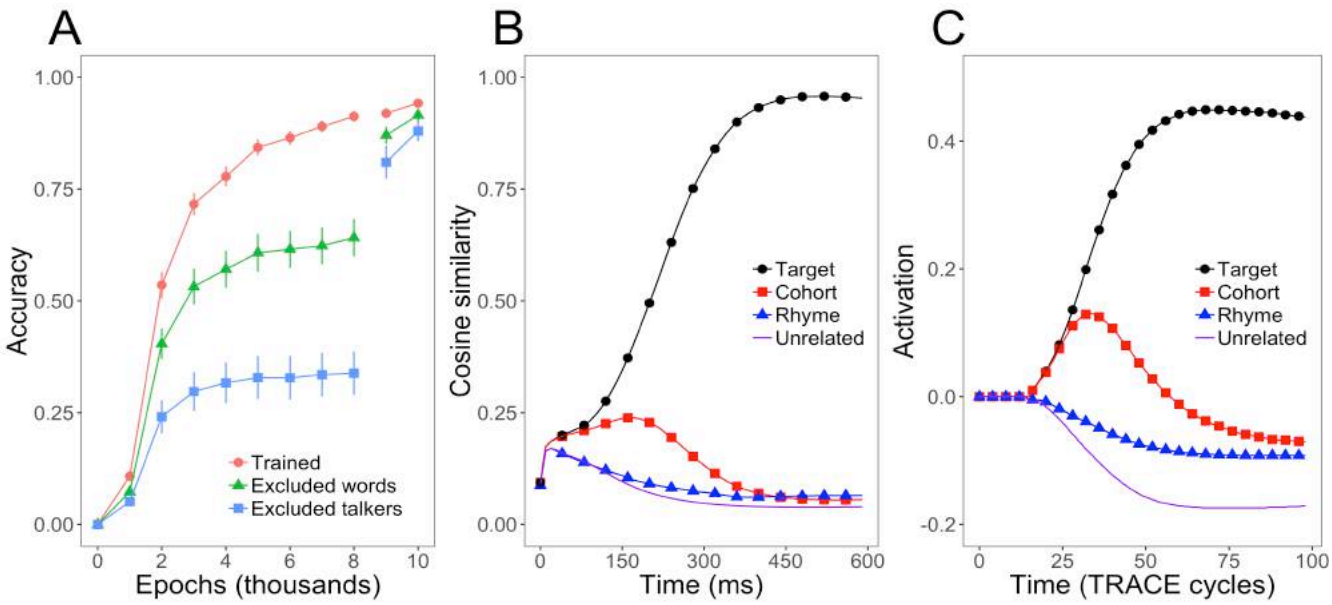


Figure 2. Model performance. (A) Accuracy by epoch averaged over 10 models. When training resumed with all items included (epochs 8001-10,000), high performance was achieved quickly for all talkers. **(B)** Competition time course (correct trials), for 2 criterial competitor types. For a target (e.g., CAT), "Cohort" represents mean cosine similarity for words overlapping in the first 2 phonemes (CAN, CASTLE). "Rhyme" words rhyme with the target (BAT, SAT). "Unrelated" is the average for all words phonologically dissimilar from the target. This pattern closely follows human performance (Allopenna et al., 1998). **(C)** For comparison, we conducted simulations with the TRACE model, with its standard 212-word lexicon, 14-phoneme inventory, and idealized "pseudo-spectral" inputs. Crucially, EARSHOT displays the same rank ordering and similar timing for competitor types as the gold-standard TRACE model.

Unpacking the model

How can we determine how the model works, and how can its mechanisms guide theories of HSR (both cognitive and neural)? To address this, we borrowed an approach that Mesgarani et al. (2014) developed for decoding human electrocorticography data. We presented the model with all possible CV and VC vowels, and examined the responses of every hidden unit over time. For every hidden unit paired with every phoneme, we calculated a *Phonetic Sensitivity Index* (PSI). For example, for unit 239, we would note its mean activation in response to /b/ from the onset of /b/ to 100 ms later. We then subtract unit 239's response to each other phoneme in turn from its response to /b/. When the difference is > 0.3, the PSI for {239, /b/} would be incremented. We repeat this for all 39 phonemes. The maximum PSI for a unit-phoneme pair would be 38 (indicating a unit that responded more strongly to that phoneme than to any other).

We calculated the PSI for all unit-phoneme pairs. Then, we subjected the resulting unit-by-phoneme matrix to hierarchical clustering (Fig. 3). This allows us to ask whether phonetic structure emerges as the model learns to map speech to semantics, even though no explicit information about phonetic features or phonemes is given in training.

About 50% of hidden units exhibited structured responses in the SI time window (20% of electrodes examined by Mesgarani et al. [2014] met their inclusion criteria). The hierarchically clustered PSI solution bears remarkable resemblance to that derived from electrodes in human superior temporal gyrus, with selective responses for phonetically similar phonemes.

The PSI analysis reveals an internal phonetic code that emerges over training. However, hidden units have more complex dynamics than are revealed by the PSIs. Profiles include strong responses at phoneme onset, but also delayed and sustained responses (see Magnuson et al., 2018). In future work, we will explore how the full combination of response profiles support EARSHOT's robust performance. It is also possible that the variety of response profiles observed in the model could be the basis for hypotheses regarding candidate response profiles that might occur in human cortical recordings.

Discussion

Decades after the *lack of invariance problem* – the absence of invariant cues to speech sounds (e.g., Joos, 1948; Liberman et al., 1952; Peterson & Barney, 1952) – was first described, speech science offers limited explanations for human phonetic constancy. A significant obstacle is that computational models of HSR have side-stepped the problem of working directly on the speech signal. Instead, models have focused on the challenges inherent in spoken word recognition beyond initial encoding, using simplified inputs such as gradient phonetic features (McClelland & Elman, 1986), phonemes (Hannagan et al., 2013; You & Magnuson, 2018), or human phoneme confusion probabilities (Norris & McQueen, 2008) instead of real speech. Ironically, simplifying assumptions can *complicate* theoretical challenges (Magnuson, 2008) by masking constraints (in this

case, e.g., prosodic cues to phoneme identity or word length).

Simplifying assumptions about input were motivated by complexity concerns. As McClelland and Elman (1986) argued, models aimed at guiding psychological theory must prioritize psychological over computational adequacy, favoring simplicity and understandability over full, end-to-end modeling. A comprehensive and robust model that is itself too complex to understand offers little guidance to HSR theories.

In developing EARSHOT, our aim was to maximally conserve psychological adequacy (i.e., simplicity) in a model that takes real speech as input. Borrowing one tool from ASR – long short-term memory (LSTM) nodes (Hochreiter & Schmidhuber, 1997) – allowed a *shallow* recurrent network to *learn* to map from *speech* to pseudo-semantics while exhibiting human-like dynamics of lexical activation and competition (similar to TRACE; Fig. 2). Generalization (on items from trained-on talkers that were not included in training, as well as talkers wholly excluded from training) was fairly low and quite variable. On the one hand, this represents a major advance, since there simply are no other *cognitive* models of HSR that operate on real speech. This is the first time such a simple model has been applied to problems entailed by doing so (talker variability, etc.). On the other hand, relatively low and variable generalization may reflect the degree to which the model *memorizes* training patterns. In ongoing work, we are exploring the use of more variable inputs, but ultimately, we must move to using open-ended training items produced by natural talkers.

Another contrast with other models of HSR is that EARSHOT is a learning model. Although we have thus far used an unnatural training regimen, EARSHOT allows the exploration of more naturalistic learning.

Admittedly, *how* the model succeeds in learning to map speech to semantics is not yet completely clear. By importing techniques from human electrocorticography (Mesgarani et al., 2014), we were able to track responses of hidden units to specific phonemes (Fig. 3) and observe the model's emergent sensitivity to phonetic structure. It develops this sensitivity without any explicit training or information about phonetic features or phonemes. Deeper understanding will require more complex analyses of not just hidden units, but also output units and weight layers.

However, the preliminary similarity of EARSHOT's hidden unit responses to responses in human superior temporal cortex (Mesgarani et al., 2014) suggests that our approach has potential for new means of developing cognitive models that are potentially linkable to the neural substrates supporting HSR. Speculatively, we would propose that response profiles observed in hidden units in a model like

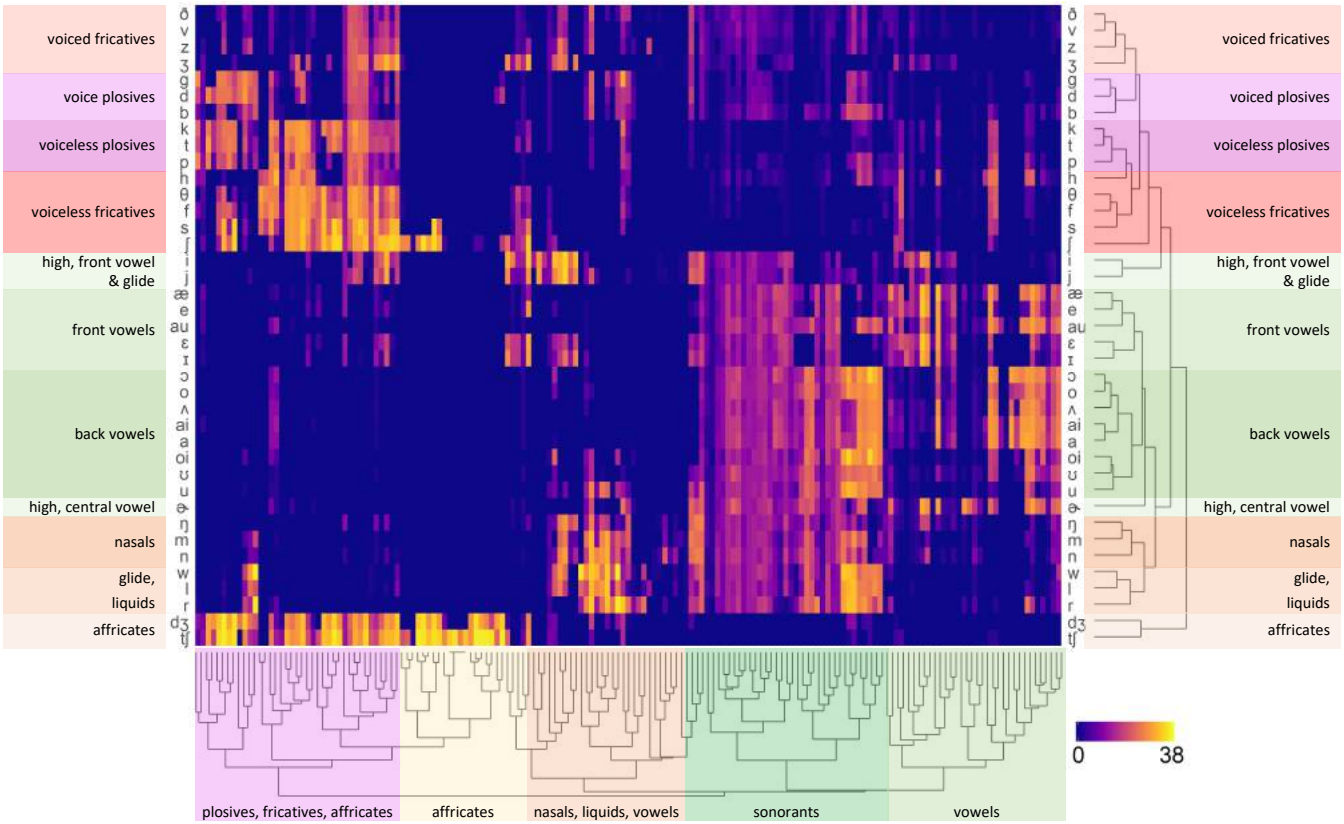


Fig. 3. Phonetic sensitivity revealed by hierarchical clustering. *Phonetic Sensitivity Index (PSI)* based on hidden unit (x-axis) responses in the presence of specific phonemes. For every hidden unit-phoneme pair, PSI was incremented for every phoneme to which the hidden unit responded substantially *more weakly* (yellow indicates high selectivity, with maximum PSI of 38, given 39 phonemes). 246 HUs showing selective responses are included. We used hierarchical clustering to sort both axes, revealing substantial structure in hidden unit responses.

EARSHOT could provide hypotheses for human cortical responses.

In conclusion, EARSHOT may provide a first step towards a comprehensive solution to the overarching challenge for theories and models of HSR – the *lack-of-invariance problem.* Simulations on previously out-of-reach topics (talker and rate variability, etc.) can be conducted with the *same materials* presented to human listeners. Our aim in this brief report is to provide a snapshot of the basic properties of EARSHOT. In a longer subsequent report, we will describe our ongoing work to more fully assess the capabilities of the model.

Acknowledgments

Supported by NSF 1754284, NSF IGERT 1144399, & NSF NRT 1747486 (PI: J.S.M.); NICHD P01 HD0001994 (PI: J.R.); and NSF 1827591 (PI: R.M.T.).

References

Allopenna, P.D., Magnuson, J.S., Tanenhaus, MK (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language* 38, 419-439.

Cole, R.A. & Jakimik, J. (1980). A model of speech perception. In R.A. Cole (Ed.), *Perception and production of fluent speech* (pp. 133-163). Mahweh, NJ: Erlbaum.

DiCarlo, J.J., Cox, D.D. (2007). Untangling invariant object recognition. *Trends in Cognitive Sciences, 11*, 333-341.

Fowler, C.A. & Housum, J. (1987). Talkers' signaling of "new" and "old" words in speech and listeners' perception and use of the distinction. *Journal of Memory & Language, 26*, 489-504.

Grossberg, S., Boardman, I. & Cohen, M. (1997). Neural dynamics of variable-rate speech categorization. *Journal of Experimental Psychology: Human Perception and Performance, 23*, 418-503.

Hannagan, T., Magnuson, J.S., & Grainger, J. (2013). Spoken word recognition without a TRACE. *Frontiers in Psychology*, 4:563. doi:10.3389/fpsyg.2013.00563.

Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. & Kingsbury B. (2012). Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. *IEEE Signal Processing*, *29*, 82-97.

Hochreiter, S., Bengio, Y., Frasconi, P. & Schmidhuber, J. (2001). Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. In S.C. Kramer & J.F. Kolen (Eds.) *A Field Guide to Dynamical Recurrent Neural Networks* (pp. 237-374). IEEE Press.

Hochreiter, S. & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*, 1735-1780.

Joos, M. (1948). *Acoustic phonetics.* Baltimore, MD: Linguistic Society of America.

Kell, A.J.E., Yamins, D.L.K., Shook, E.N., Norma-Haignere, S.V. & McDermott, J.H. (2018). A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron* 98, 630-644.

Laszlo, S. & Plaut, D.C. (2012). A neurally plausible parallel distributed processing model of event-related potential reading data. *Brain and Language* 120, 271-281.

Liberman, A.M., Cooper, F.S., Shankweiler, D.P. & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review* 74*,* 431-461.

Liberman, A.M., Delattre, P.C. & Cooper, F.S. (1952). The role of selected stimulus variables in the perception of the unvoiced-stop consonants. *American Journal of Psychology* 65, 497-516.

Magnuson, J.S. (2008). Nondeterminism, pleiotropy, and single word reading: Theoretical and practical concerns. In E. Grigorenko & A. Naples (Eds.), *Single Word Reading* (pp. 377-404). Mahweh, NJ: Erlbaum Associates.

Magnuson, J.S., You, H., Nam, H., Allopenna, P.D., Brown, K., Escabi, M., Theodore, R.M., Luthra, S., Li, M., & Rueckl, J. (2018, December 13). EARSHOT: A minimal neural network model of incremental human speech recognition. https://doi.org/10.31234/osf.io/h7a4n

McClelland, J.L. & Elman, J.L. (1986). The TRACE Model of Speech Perception. *Cognitive Psychology* 18, 1-86.

Mesgarani, N., Cheung, C., Johnson, K. & Chang, E.F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science* 343, 1006-1010.

Miller, J.L. & Baer, T. (1983). Some effects of speaking rate on the production of /b/ and /w/. *Journal of the Acoustical Society of America* 73*,* 1751–1755.

Nagamine, T., Seltzer, M.L. & Mesgarani N. (2015). Exploring how deep neural networks form phonemic categories. Sixteenth Annual Conference of the International Speech Communication Association, Dresden, Germany, 1912-1916.

Norris, D. & McQueen, J.M. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review* 115, 357-395.

Peterson, G.E. & Barney, H.L. (1952). Control methods used in a study of vowels. *Journal of the Acoustical Society of America* 24*,* 175–184.

Scharenborg, O. (2010). Modeling the use of durational information in human spoken-word recognition. *Journal of the Acoustical Society of America* 127, 3758-3770.

Scharenborg, O., Norris, D., ten Bosch, L., & McQueen, J.M. (2005). How should a speech recognizer work? *Cognitive Science* 29, 867-918.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., & Polosukhin I. (2017). Attention is all you need. arXiv:1706.03762v5 [cs.CL].

Werbos, P.J. (1988). Generalization of backpropagation with application to a recurrent gas market model. *Neural Networks* 1*,* 339-356.

You, H., & Magnuson, J.S. (2018). TISK 1.0: An easy-to-use Python implementation of the time-invariant string kernel model of spoken word recognition. *Behavior Research Methods*. doi:10.3758/s13428-017-1012-5.

Emergence of Collective Cooperation and Networks from Selfish-Trust and Selfish-Connections

Korosh Mahmoodi (koroshm@andrew.cmu.edu)
Dynamic Decision Making Laboratory, Department of Social and Decision Sciences, Carnegie Mellon University
5000 Forbes Ave,
Pittsburgh PA 15213 USA

Cleotilde Gonzalez (coty@cmu.edu)
Dynamic Decision Making Laboratory, Department of Social and Decision Sciences, Carnegie Mellon University
5000 Forbes Ave,
Pittsburgh PA 15213 USA

Abstract

Emergence of collective cooperation in an inherently selfish society is a paradox that has preoccupied biologists, sociologists, and cognitive scientists alike for centuries. We propose a computational model and demonstrate through simulations how collective cooperation can emerge from selfish interests: the goal of improving each individual's own rewards. We also demonstrate how the same selfish interests lead to the dynamic emergence of a network of interconnected agents. Our model includes two simple mechanisms: Selfish-Trust (ST) and Selfish-Connection (SC). ST involves the possibility of relying on others in a society of agents when it is beneficial to the individual, and SC involves the possibility of connecting to other agents when those agents help improve the individual's own benefit. Our simulation results suggest that collective cooperation can emerge from ST and a complex dynamic network can emerge from ST and SC. The simulated data demonstrate an important property of many living organisms: patterns of temporal complexity, which are essential to transfer information among agents of any society of living beings.

Keywords: Altruism Paradox, Emergence of Cooperation, Selfishness, Trust, Networks, Artificial Intelligence

Introduction

For Charles Darwin (Darwin, 1871) altruism remained a paradox: the act of sacrificing an individual's own benefit for the benefit of the collective community of living organisms was regarded as a contradiction to evolutionary theories. The dilemma of emergence of cooperative behavior in situations in which there is a large incentive to defect for the individual benefit has been widely studied in sociology and cognitive sciences. The Prisoner's dilemma (PD) has been a leading metaphor for the study of the evolution of cooperative behavior in populations of selfish in which selfishness is more rewarded in the short-term (M. Nowak & Sigmund, 1993; Gonzalez, Ben-Asher, Martin, & Dutt, 2015).

The PD, dates back to the early development of Game Theory (Rapoport & Chammah, 1965), and it is a common abstraction of the essential elements of many naturalistic situations involving cooperative behavior. It is generally represented with a payoff matrix that provides payoffs according to the actions of two players (see Table 1). When both players cooperate, each of them gains the payoff $\prod(t) = R$, and when both players defect, each of them gains $\prod(t) = P$. If the player i defects and player j cooperates, player i gains the payoff $\prod(t) = T$ and player j gains the payoff $\prod(t) = S$ and vice versa. The constraints on the values of the payoffs in the PD are $T > R > P > S$ and $S + T < 2R$. The temptation to defect is established by setting the condition $T > R$.

		Player j: C	Player j: D
Player i	C	(R,R)	(S,T)
	D	(T,S)	(P,P)

Table 1: The general payoffs of PD game. The first value of each pair is the payoff of agent i and the second value is the payoff of the agent j.

The dilemma is that, while the longer-term best mutual action is to cooperate, in the short-term each individual would prefer to defect because it indicates a higher reward to the individual. Assuming that the other player also searches for its own individual maximum reward, the pair will end up in a $D-D$ situation with the minimum payoff for the two players $2P$.

How do individuals realize that cooperation is mutually beneficial in the long-term? this question has been addressed by many researchers, at various levels of inquiry, involving pairs of agents (Gonzalez et al., 2015; Moisan, ten Brincke, Murphy, & Gonzalez, 2018) as well as larger social networks (M. Nowak & Sigmund, 1993). Research suggests that, at the pair level, people dynamically adjust their actions according to their observations of others' actions and outcomes; at the network level, research suggests that the emergence of cooperation may be explained from *network reciprocity*, where individuals play with those agents with whom they are already connected in a network structure. The demonstration of how social networks and structured populations with explicit connections foster cooperation was introduced by Nowak and May (1992). Alternative models based on Network reciprocity assume agents in a network play the PD with the agents with whom they have specific interconnections. Agents act by copying the strategy of the richest neighbor, basing their decisions on the observation of the others' payoffs. Thus, network reciprocity depends on the existence of a network structure (an already predefined set of connections among agents) and on the awareness of the behavior and pay-

offs of interconnected agents. Network reciprocity assumes that the evolution of cooperation is a function of the difference between the payoffs of the interacting agents.

Thus, past research assumes that the emergence of collective cooperation requires the observation of others' actions and/or outcomes and the existence of predefined connections among agents. Indeed, empirical work suggests that the emergence of cooperation depends on the level of information available to each agent (Martin, Gonzalez, Juvina, & Lebiere, 2014); and the less information about other agents exist, the more difficult, and perhaps the longer it takes, for cooperation to emerge (Martin et al., 2014; Rapoport & Chammah, 1965). However, other experiments suggest that humans do not consider others' payoffs when making their decisions, and that a network structure does not influence the final cooperative outcome (Fischbacher, Gächter, & Fehr, 2001). Indeed, in many aspects of life, we influence others through our choices and others' choices affect us, but we are not necessarily aware of the exact actions and rewards received by others affecting us. For example, when a member of society avoids air travel in order to reduce the individual's carbon footprint, he or she might not be able to observe whether others are reducing their air travel too, yet rely on decisions others make, influencing the community as a whole. It is thus, difficult to explain how behaviors can be self-perpetuating even when the source of influence is unknown (Martin et al., 2014).

In this research, we aim at advancing our understanding of the emergence of collective cooperation in the absence of explicit knowledge of others' actions and outcomes, and in the absence of an explicit predefined network structure that connects agents in a society. We introduce an algorithm (*Living Thing*, LT) to demonstrate that collective cooperation can emerge and survive between agents, out of selfishness (i.e., the individual's need to act on their own personal benefit), and in the absence of others' information (i.e., without a need to any predefined network). We aim at developing hypotheses that can help resolve social dilemmas that exist in the real world. For example, if we understand how collective cooperation emerges only from the decisions of each individual, we could propose solutions that reduce the dilemmas in social problems such as littering in public places or the lack of contributions to a reduction of CO_2 in the atmosphere (Martin et al., 2014).

A LT agent will act according to the reinforcement of its own past actions (Reinforcement Learning, RL), but it will rely on two mechanisms that may overwrite the agent's RL actions: Selfish-Trust (ST) and Selfish-Connection (SC). ST is a decision to follow or rely on other agent's decision expecting that it will improve the own agent's reward with respect to the agent's own previous payoff. ST is expected to turn the initially defector agents to agents that cooperate most of the time. SC is a mechanism that helps agents learn who to play with: agents increase the propensity of playing with the same other agent if the payoff received after playing with that other agent is higher than the agent's own previous payoff.

Past models of network formation rely on a concept of preferential attachment (PA) (Barabási & Albert, 1999), which uses rules according to which an agent would have a higher chance of linking with other agents that already have many links (i.e., high reputation nodes). In contrast, LT demonstrates that such propensities to connect to other agents emerge dynamically, according to the experienced benefits that the other agent brings to the individual's own benefit (SC).

We carry an analysis of the emergence of cooperation from these mechanisms. The simulation results hint at how to explain emergent collective cooperation from individual selfish interests. An important hypothesis emerging from this work is that cooperation can emerge and survive out of the selfishness of agents even when there are no specific awareness of outcomes of other agents, and that a network structure can emerge dynamically from the connections guided by self individual interests.

Living Thing (LT) Algorithm

Figure 1 shows one-time cycle of the LT algorithm from the perspective of one of the agents, agent i, but every step is executed for agent j simultaneously. In Step 1, a pair of randomly selected agents i and j "agree" to play. Only one pair of agents is selected at each time cycle. The following are general notations in the algorithm: V_i is the decision of the agent to Cooperate (C) or Defect (D); r represents a random number in the interval [0 1] which should be generated whenever it is called in the algorithm; Δ is a positive number that represents an increase in three possible cumulative tendencies: to play C or D, to trust the paired agent or not, or to play again with a previous agent. These cumulative tendencies increase by Δ when the benefit of the agent i changes with respect to its previous benefit and if there is no change then we set $\Delta = 0$ which means no change happens in the system. Δ, in general, can be a function of the difference between two past payoffs of the agent and it can be different for different cumulative tendencies, but it does not change the general results presented later (the form of sensitivity to payoffs is important when two systems interact with each other which is out of scope of this paper).

The following steps are executed in each time t of the algorithm:

Pairing Agents (1)

Agent i and agent j get picked randomly. Agent i at time t has the propensity $P_{ij}(t) = M_{ij}(t)/\sum_k M_{ik}(t)$ to play with agent j. $0 < P_{ij}(t) < 1$. $M_{ij}(t)$ is cumulative tendency for agent i to pick agent j to play at time t. This cumulative tendency changes at step 7 according to the last two payoffs received by agent i.

At the same time, agent j has a propensity $P_{ji}(t) = M_{ji}(t)/\sum_k M_{jk}(t)$ to play with agent i. $0 < P_{ji}(t) < 1$. Two agents i and j pair-up if two random numbers, $0 < r_1 < 1$ and $0 < r_2 < 1$, satisfy inequalities $r_1 < P_{ij}(t)$ and $r_2 < P_{ji}(t)$.

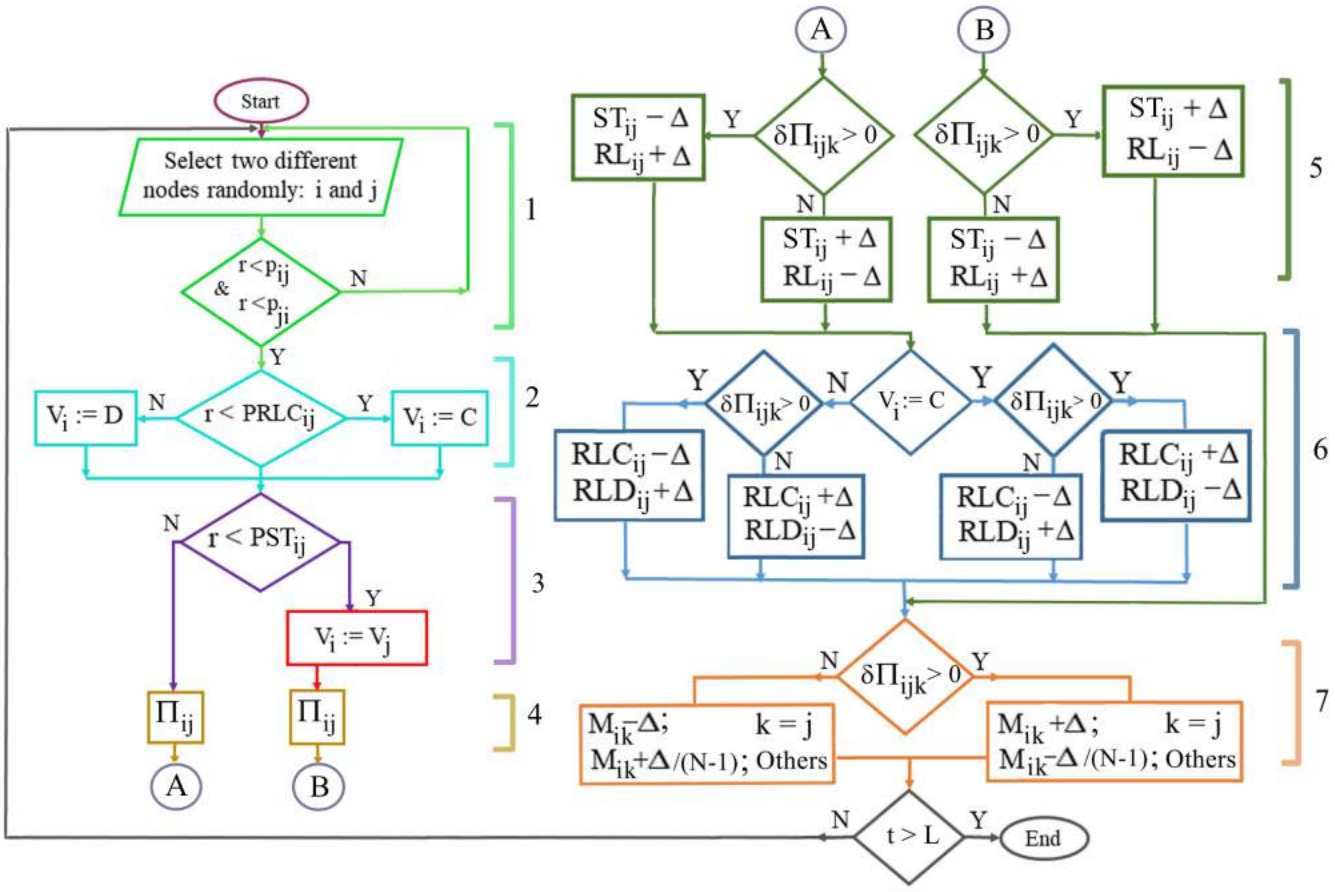


Figure 1: Flowchart of the LT algorithm. "Y" and "N" letters represent "Yes" and "No", respectively.

Otherwise another two agents are randomly selected in each time t.

Reinforcement Learning (RL) (2)

Agent i initially selects an action by reinforcement learning (RL): This agent has the propensity $PRLC_{ij}(t) = RLC_{ij}(t)/(RLC_{ij}(t)+RLD_{ij}(t))$ to pick C and has the propensity $PRLD_{ij}(t) = RLD_{ij}(t)/(RLC_{ij}(t)+RLD_{ij}(t))$ to pick D as it's next potential decision.

$RLC_{ij}(t)$ and $RLD_{ij}(t)$ are cumulative tendencies for agent i playing with agent j, at time t, for choice C or D, respectively. These cumulative tendencies change at step 6 based on the last two payoffs of agent i.

To select the action, a random number r gets picked and if $r < PRLC_{ij}(t)$ then it's next decision will be C otherwise will be D. The same process applies for agent j.

Selfish-Trust (ST) (3)

Instead of executing the decision determined by RL in Step 2, agent i has a chance to trust the decision made by agent j made using RL in Step 2, and with whom agent i is paired with. The propensity that agent i relies on the decision of agent j is: $PST_{ij}(t) = ST_{ij}(t)/(ST_{ij}(t)+RL_{ij}(t))$. $ST_{ij}(t)$ and $RL_{ij}(t)$ are cumulative tendencies for agent i to execute its choice based on ST from agent j, at time t, or to execute its choice based on RL respectively. These cumulative tendencies update in step 5 based on the last two payoffs of agent i. Again, if a random number r is less than $PST_{ij}(t)$ then ST happens.

Evaluating Own Payoffs (4)

At time t agent i after executing its C or D action while playing the PD game with agent j, receives the payoff $\prod_{ij}(t)$. The last two payoffs of the agent i are used to determine the changes in its accumulative tendencies: $\delta\prod_{ijk}(t) = \prod_{ij}(t) - \prod_{ik}(t-1)$, where agent (k) is the agent that played with agent i in trial $t-1$. In the flowchart we showed this quantity as $\delta\prod$.

Update of cumulative tendency of ST or RL (5)

If agent i used ST and after playing with agent j its payoff is higher than its previous payoff, $\delta\prod_{ijk}(t) > 0$, then the accumulative tendencies ST_{ij} and RL_{ij} , for the next time agent i and j, change to $ST_{ij}+\Delta$ and $RL_{ij}-\Delta$. The same happens for agent j.

Similarly, if agent i used RL and after playing with agent j its payoff is higher than its previous payoff, $\delta\prod_{ijk}(t) > 0$, then the accumulative tendencies RL_{ij} and ST_{ij}, for the next

		Player j	
		C	D
Player i	C	$(1,1)$	$(0,1.9)$
	D	$(1.9,0)$	$(0,0)$

Table 2: The payoffs of PD game used in the simulations. The first value of each pair is the payoff of agent i and the second value is the payoff of its pair, agent j.

time agent i and j pair up, change to $RL_{ij}+\Delta$ and $ST_{ij}-\Delta$. The same happens for agent j.

Update of cumulative tendency to choose C or D (6)

Step 6 is only active if the agent decided to use RL is step 3. If agent i played with agent j and received a payoff higher than its previous payoff, $\delta\prod_{ijk}(t)>0$, and this happened because agent i played C, then the accumulative tendencies RLC_{ij} and RLD_{ij}, for the next time agent i and j pair up, change to $RLC_{ij}+\Delta$ and $RLD_{ij}+\Delta$. If the increase happened because agent i played D, then the accumulative tendencies RLD_{ij} and RLC_{ij} , for the next time agent i and j pair up, change to $RLD_{ij}+\Delta$ and $RLC_{ij}-\Delta$. The same happens for agent j.

Selfish-Connection (SC) (7)

In this step the cumulative tendency to play with a specific agent changes. If agent i, after playing with agent j, receives higher benefit with respect to its previous payoff, $\delta\prod_{ijk}(t)>0$, then the cumulative tendency of pairing with agent j, M_{ij}, increases to $M_{ij}+\Delta$ and for the rest of the cumulative tendencies decreases to $M_{il}-\Delta/(N-1), l\neq j$. If $\delta\prod_{ijk}(t)<0$, then the cumulative tendency of pairing with agent j, M_{ij}, decreases to $M_{ij}-\Delta$ and for the rest of the cumulative tendencies increases to $M_{il}+\Delta/(N-1), l\neq j$. The same happens for agent j.

Simulation Methods

We studied a system with $N=100$ agents. Initially all the agents are defectors, have payoff of zero, have more chance to stay as defector; $RLC_{ij}(0)=1$, $RLD_{ij}(0)=99$, have more chance to use RL over ST; $ST_{ij}(0)=1$, $RL_{ij}(0)=99$, and have equal chance to pair up with other agents; $M_{ij}(0)=100$. We set $\Delta=10$. Δ is the property of the system and shows the sensitivity of the agent to the feedback from its two last payoffs. Smaller Δ decreases the rate of reaching to cooperation but doesn't change the dynamical properties of the system. The payoffs matrix used has the values shown in Table 2 as suggested by Gintis (2009): $R=1$, $P=0$ and $S=0$. So, the maximal possible value of T is 2. We selected the value $T=1.9$, which gives a very strong incentive to defect.

Results

Emergence of Cooperation from ST

Here we show that simple mechanism of ST can lead agents who play PD game (which has a high tendency to defect) toward cooperation. Figure 2 shows the proportion of cooperation in simulations that rely only on the RL mechanism (blue curve) compared to the emergence of cooperation when the simulations rely on the additional ST mechanism.

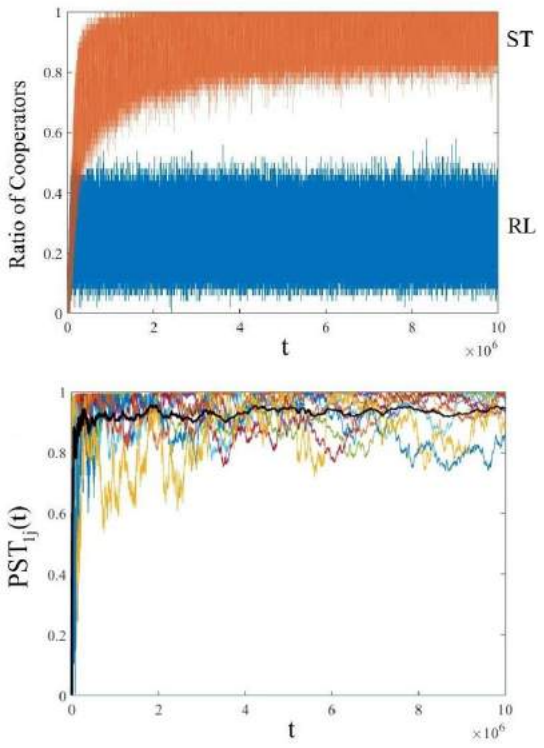


Figure 2: Top panel: the orange curve is the ratio of Cooperators vs. time for $N=100$ agents, randomly paired up to play PD game and used ST (steps 7 in LT algorithm inactive) for updates of the strategy and cumulative tendencies. The blue curve is the ratio of cooperators for $N=100$ agents, randomly paired up to play PD game and just used RL for the decision making process (steps 3 and 7 in LT algorithm inactive). Bottom panel: emergence and evolution of the probability of trust of unit 1 the other 9 agents in a system with 10 agents, used ST to update their strategies and cumulative tendencies. The thicker, black curve in this figure is the average of all the nine probabilities of trust

The blue curve in the top panel of Figure 2 shows the time evolution of the ratio of cooperators when RL is the only mechanism that agents use to update their strategy (steps 3 and 7 are inactive).

The orange curve in the top panel of Figure 2 shows the emergence of cooperation between agents when at any trial two of them (out of 100) paired up randomly and used ST (steps 7 in LT algorithm inactive) to update their strategy (C or D) and their cumulative tendencies. The system reached its dynamic equilibrium after about 2×10^6 and sustained around the average ratio of cooperation of 0.9.

This shows the effect of ST on improving the level of cooperation compared to only RL. In the absence of ST the ratio of cooperators fluctuates around 0.3 which means the majority of agents are defectors.

The nine curves in the bottom panel of Figure 2 are the chances that one of the agents might trust others in a system with 10 agents. The thicker, black curve is the average of the STs between agent 1 and the other nine agents. The chances of ST increased and sustained to about 0.9. This means that the agent learns that ST has benefit for it (and for the whole society). The payoff of the individual is not shown here because it is proportional to the level of cooperation: more cooperation results in more payoff for individual agent and for the emerged group. In conclusion, cooperation emerges and survives because ST lets the strategies to spread between the agents, if it benefits them individually.

Emergence of Complexity over time

The analysis of the fluctuations of a time series gives us a measure for the complexity of the system. We define the events in the time series as the times the time series crosses its mean value. The distribution of the time intervals between the two consecutive events is of interest (Figure 3).

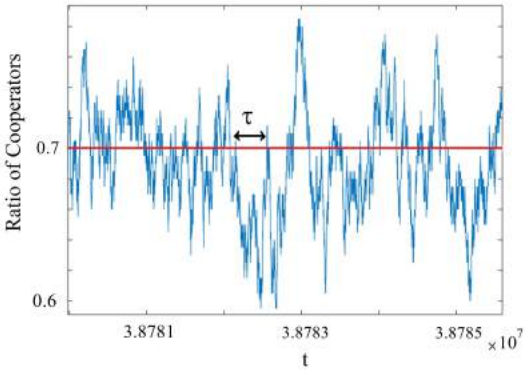


Figure 3: Demonstration of defining events in a time series. The blue curve is a zoomed in part of the Ratio of Cooperators' time series (as an example) and the horizontal red line is its mean value. Whenever the time series crosses the mean value is defined as an event. The distribution of the time intervals between two consecutive events gives a measure for the complexity of the system (complexity index μ).

We collect all the time intervals between two consecutive events (τ's) and evaluate the probability density function (PDF) $\psi(\tau)$. If the resulting distribution is Poisson then the dynamic of the system is random and obeys ordinary statistics. But, if the PDF is a power law; $\psi(\tau) \propto 1/\tau^{\mu}$, then the dynamics falls in the category of the complex systems, for example, the dynamics of the brain. The parameter μ, the slope of the Inverse power law in a log-log plot, is a measure for the complexity of the time series: when $\mu > 3$ then the system is ordinary while for $1 < \mu < 3$ there is Ergodicity Breaking and the system does not obey ordinary statistics.

Temporal criticality is crucial for transfer of information between two intelligence systems (Aquino, Bologna, West, & Grigolini, 2011). To measure the complexity index μ of the time series of the ratio of cooperators on the four cases of emergence of cooperation (top panels in Figure 2 and Figure 6), we studied their fluctuations in the asymptotic regime ($t > 5 \times 10^6$) around their mean value.

Figure 4 shows that the time series of the ratio of cooperators for both cases where agents used ST or just used RL, time series of top panel in Figure 2, have inverse power law PDF with the same complexity index of $\mu = 1.3$ (which falls in the interval $1 < \mu < 3$). The difference is that the linear part of the distribution for the first case, where ST exists, is extended toward larger τ's which means the system is more complex respect to the latter case.

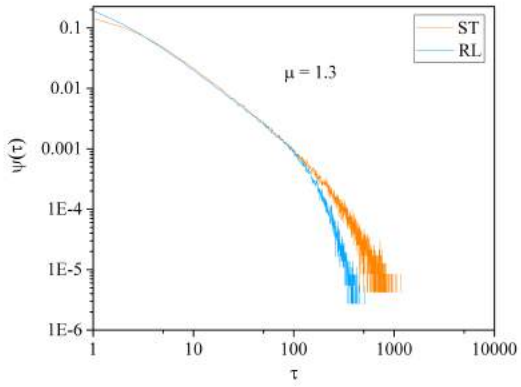


Figure 4: The PDF of the time intervals between the two consecutive events of the time series of Figure 2 (in log-log scale).

Figure 5 shows the PDFs of the time series of the ratio of cooperators for the cases where agents are using ST SC (LT) or RL SC for their evolution, top panel in Figure 6. Both cases have an inverse power law PDF. The PDF of the first case where ST and SC are both active shows very extended linear part with complexity index of $\mu = 1.73$. This complexity is similar to that of the time series of the living things (Allegrini, Paradisi, Menicucci, & Gemignani, 2010).

On the other hand, the complexity of the second case is similar to the system where agents were using just RL (blue curve of Figure 5). This means that SC could increase the complexity of the system where ST was already in action.

Emergence of connections with other agents

In this section we add another level of learning to ST by letting the agent to find the agents which playing with them increases its payoff respect to its previous payoff (all sections of the LT algorithm active). The aim is to show that a dynamic complex network emerges naturally and from the ST and SC mechanisms of the LT model.

The blue curve in the top panel of Figure 6 is the ratio of cooperators when SC is added to RL (section 3 of the LT algorithm inactive). The blue curves in the top panels of Figure 2 and Figure 6 are very similar, which means that adding

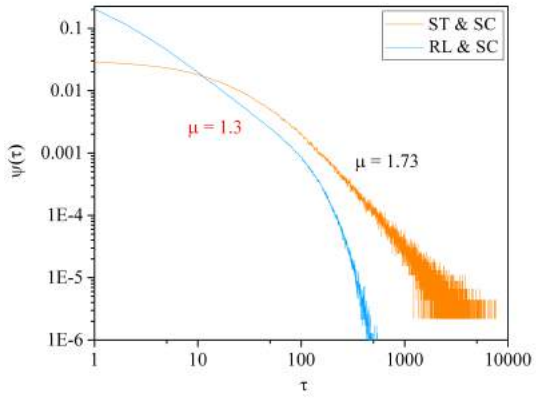


Figure 5: The PDF of the time intervals between the two consecutive events of the time series of Figure 6 (in log-log scale).

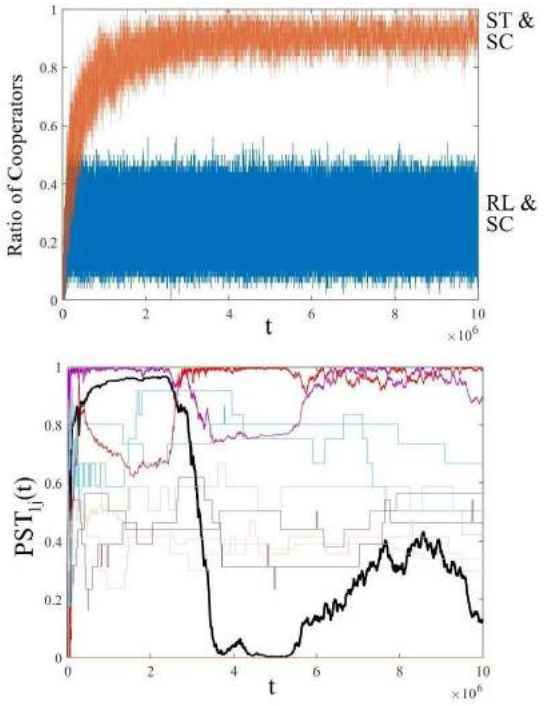


Figure 6: Top panel: orange curve: the ratio of Cooperators vs. time for $N = 100$ agents which in addition to ST they use SC to pick their partner to play PD game (all sections of the LT algorithm active). The Blue curve is the ratio of cooperators, paired using SC, but updated their strategies only with RL (section 3 of the LT algorithm inactive). Bottom panel: emergence and evolution of trust of unit 1 the other 9 agents in a system with 10 agents using the LT algorithm (ST SC) to update their strategies and cumulative tendencies. The thicker, black curve in this figure is the chance of agent 1 to play with the agent which is most trusting (the purple curve close to 1).

the ability to the agents to select their partner is not favoring cooperation when ST is inactive. The reason is that only when ST is active the level of cooperation between the agents increases and because of mutual benefit, which is stable, an agent can rank the links by changing the chance of playing with other agents based on the increase on its last two payoffs.

To illustrate this, we plot the chances of an agent (called agent 1) to trust the other 9 agents using LT algorithm in the bottom panel of Figure 6. This figure shows that the agent trusts some of the agents more than others, most of the time, and only from time to time it trusts other agents. But later on, the agent starts to trust some agents most of the time. The thicker, black curve in this figure is the chance that the agent (1) connects to the agent corresponding to the purple curve. The similarities between the red and purple curves show that the agent 1 learns to connect with the agent which it is most trusting, most of the time. The preferential connections here are dynamic and are based on the perception of the benefit that an agent receives from the other agents. This process creates connections among agents that are dynamic. Some connections become stronger and others become weaker according to the SC mechanism.

Figure 7 demonstrates the chance of a random agent (represented by a dot in the center) pairing with the other 99 agents at two different times: $t = 10^2$ (top panel) and 10^6 (bottom panel). The thickness of the lines represents the chance of the pairings. At $t = 10^2$, we observe an almost uniform distribution of the probability of the connections of an agent to the others (top panel); but later, the agent learned to prefer to connect to some partners more than to others (bottom panel).

Figure 8 shows the probability density function of the pairings for all the agents in Figure 7 (bottom panel). This distribution, plotted in a log-log scale, shows an inverse power law $\propto 1/P^{\beta}$ with complexity index of $\beta = 1.3$, rather than having a Poisson distribution, showing that the emerged network is complex.

Discussion and Implications of Results

The novelty of the LT algorithm is the demonstration of how collective behavior can emerge from Selfish Trust and how a network can emerge from Selfish connections; in the absence of an explicit a-priory network structure, and in the absence of explicit awareness of others' outcomes. LT uses ST that adapts to increase or decrease the chance that agent i will trust the strategy of agent j, if that strategy is beneficial or detrimental for agent i itself. LT also uses SC that adapts to increase or decrease the chance of agent i to connect to agent j, if agent j has contributed or not to the own benefit of agent i. This means that selfishness of agent i is used as the main learning incentive: If the payoff of the agent i increased with respect to its previous payoff then it will increase the likelihood of repeating its last action. This control of the dynamics is internal and emergent according to the self-interest of the

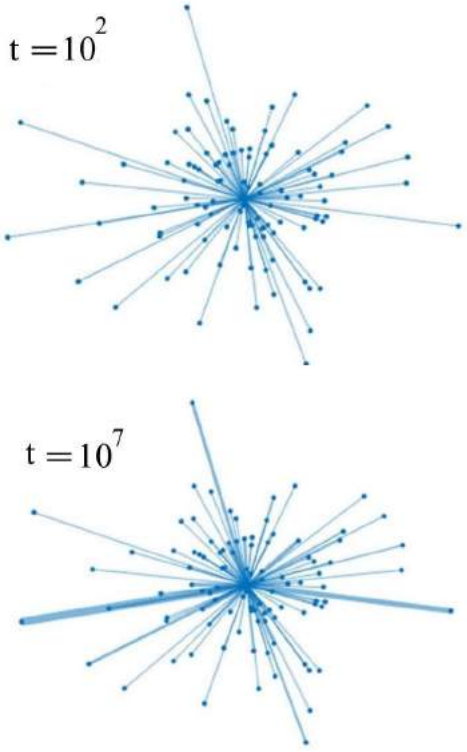


Figure 7: Each dot represents an agent. The dot in the center shows the agent of interest and the other 99 dots are connected to it with lines. The thickness of each line is proportional to the chance of the corresponding agents to pair up at time $t = 10^2$ (top figure) and $t = 10^7$ (bottom figure).

agents, leading the system to self-organization. The role of ST is to spread the strategies between the agents, if it is increasing the payoff of individuals with respect to their previous one. Adding SC to ST lets each agent learn which agents to connect to, in order to increase its own payoff with respect to its last one. SC can improve the complexity of the system by forming a dynamic network of chances of pairings, which results in an inverse power law PDF with complexity index of $\beta = 1.3$. The self-organized system evolved by LT host events with inverse power law PDF of the interval between the consecutive events with complexity index $\mu = 1.73 < 3$. This is the main property of dynamic complex systems which makes them able to transfer information and match to another.

Acknowledgments

This research was supported by the Army Research Office, Network Science Program, Award Number: W911NF1710431.

References

Allegrini, P., Paradisi, P., Menicucci, D., & Gemignani, A. (2010). Fractal complexity in spontaneous eeg metastable-state transitions: new vistas on integrated neural dynamics. *Frontiers in physiology*, *1*, 128.

Aquino, G., Bologna, M., West, B. J., & Grigolini, P. (2011). Transmission of information between complex systems: 1/f resonance. *Physical Review E*, *83*(5), 051130.

Barabási, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. *science*, *286*(5439), 509–512.

Darwin, C. (1871). *The descent of man. the modern library.* Random House Inc., NY.

Fischbacher, U., Gächter, S., & Fehr, E. (2001). Are people conditionally cooperative? evidence from a public goods experiment. *Economics letters*, *71*(3), 397–404.

Gintis, H. (2009). The bounds of reason game theory and the unification of the behavioral sciences.

Gonzalez, C., Ben-Asher, N., Martin, J. M., & Dutt, V. (2015). A cognitive model of dynamic cooperation with varied interdependency information. *Cognitive Science*, *39*(3), 457-495. doi: 10.1111/cogs.12170

Martin, J. M., Gonzalez, C., Juvina, I., & Lebiere, C. (2014). A description-experience gap in social interactions: Information about interdependence and its effects on cooperation. *Journal of Behavioral Decision Making*, *27*(4), 349-362. doi: 10.1002/bdm.1810

Moisan, F., ten Brincke, R., Murphy, R., & Gonzalez, C. (2018). Not all prisoner's dilemma games are equal: Incentives, social preferences, and cooperation. *Decision*, *5*(4), 306,322.

Nowak, M., & Sigmund, K. (1993). A strategy of win-stay, lose-shift that outperforms tit-for-tat in the prisoner's dilemma game. *Nature*, *364*(6432), 56–58.

Nowak, M. A., & May, R. M. (1992). Evolutionary games and spatial chaos. *Nature*, *359*(6398), 826.

Rapoport, A., & Chammah, A. M. (1965). *Prisoner's Dilemma: A Study in Conflict and Cooperation*. Ann Arbor: University of Michigan Press.

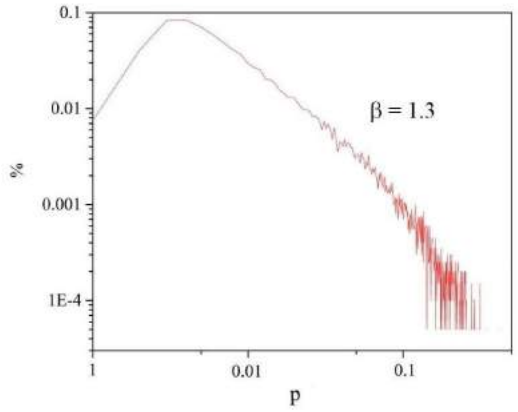


Figure 8: The probability density function of the pairings between all possible pairs at $t = 10^7$, in a log log scale.

The contrasting roles of shape in human vision and convolutional neural networks

Gaurav Malhotra & Jeffrey Bowers
Department of Psychological Sciences
University of Bristol
Bristol, BS8 1TU, UK
{gaurav.malhotra,j.bowers}@bristol.ac.uk

Abstract

Convolutional neural networks (CNNs) were inspired by human vision and, in some settings, achieve a performance comparable to human object recognition. This has lead to the speculation that both systems use similar mechanisms to perform recognition. In this study, we conducted a series of simulations that indicate that there is a fundamental difference between human vision and vanilla CNNs: while object recognition in humans relies on analysing shape, these CNNs do not have such a *shape-bias*. We teased apart the type of features selected by the model by modifying the CIFAR-10 dataset so that, in addition to containing objects with shape, the images concurrently contained non-shape features, such as a noise-like mask. When trained on these modified set of images, the model did not show any bias towards selecting shapes as features. Instead it relied on whichever feature allowed it to perform the best prediction – even when this feature was a noise-like mask or a single predictive pixel amongst 50176 pixels.

Introduction

Object recognition in humans is largely a function of analyzing shape (Biederman, 1987; Hummel, 2013). A wealth of data from psychological experiments show that shape plays a privileged role in object recognition compared to other diagnostic features such as size, colour, luminance or texture. For example, Biederman and Ju (1988) showed that error rates and reaction times are virtually identical in a recognition task when full coloured photographs of objects are replaced by their line drawings even when colour was a diagnostic feature. This indicates that shape-based representations mediate recognition. Similarly, Mapelli and Behrmann (1997) found that, for patients with an object recognition deficit (visual agnosia), surface colour played minimal role in aiding object recognition unless the shape of the object was ambiguous, indicating that shape is instrumental to recognition, whereas surface characteristics such as colour and texture play only a secondary role. More recently, Baker and Kellman (2018) have shown that participants extract shape information automatically from arrays of dot patterns within the first 100ms of stimulus onset, even for tasks where extracting this information may be detrimental to performance on a task. Experiments from developmental psychology show that this privileged status of shape starts early in life and becomes stronger with age. For example, Landau, Smith, and Jones (1988) found that 2-3-year-old children as well as adults weight shape more heavily than size or texture when generalising the name of a learnt object to novel instances. They also found that the weight placed on shape increases in strength and generality from early childhood to adulthood.

By contrast, it is unclear whether shape plays a privileged role in how convolutional neural networks (CNNs) categorise objects. It is often claimed that CNNs learn representations of objects that are similar to the representations that monkeys and humans use when identifying objects (Rajalingham et al., 2018), and that CNNs largely rely on learning shape representations in order to categorise objects (Kubilius, Bracci, & de Beeck, 2016; Jozwik, Kriegeskorte, Storrs, & Mur, 2017). On the other hand, there are a growing number of studies that show that CNNs often categorise images on the basis on non-shape attributes of images. This is demonstrated by the existence of adversarial images that are confidently classified as a familiar category despite the lack of any shape information in the input (Nguyen, Yosinski, & Clune, 2015), adversarial images that contain the correct shape but altered colours that are confidently misclassified (e.g., categorizing an image of an airplane as a dog when only the colour of the plane has been manipulated), and large reductions in performance when trained coloured images are converted to greyscale (Geirhos et al., 2017) or the colours are inverted (Hosseini, Xiao, Jaiswal, & Poovendran, 2017). In addition, there are demonstrations that CNNs can easily learn to categorise random patterns of pixels that have no shape (Zhang, Bengio, Hardt, Recht, & Vinyals, 2016). All of these findings suggest that shape may not play a privileged role in how some well-known and high-performance CNNs perform object categorisation.

However, some recent studies have argued that convolutional neural networks can show a shape-bias. Ritter, Barrett, Santoro, and Botvinick (2017) took an Inception model, a high-performance CNN (Szegedy, Vanhoucke, Ioffe, Shlens, & Wojna, 2016), and presented novel objects to the model that had been pre-trained to recognise the categories from ImageNet dataset. They found that the representations in hidden layers were more similar for two (novel) objects that overlapped in shape than for two objects that overlapped in colour. They interpret this proximity in hidden layer representations between objects of same shape as a shape-bias. In another study, Feinman and Lake (2018) trained a CNN on a controlled dataset containing synthetic images that differed on three dimensions: shape, colour and texture. They found that when this dataset was constructed in such a manner that the

(a) Salt-and-pepper noise (b) Additive noise (c) Single diagnostic pixel

Figure 1: Hidden in plane sight. Images taken from CIFAR-10 dataset and scaled up to 224x224 pixels. (a) Image is converted to greyscale and we add a salt-and-pepper noise-like mask to each training image; (b) Image is converted to greyscale and we add uniform additive noise mask to each training image; (c) A single diagnostic pixel is inserted in the image (dotted red circle is inserted here to illustrate the location of the pixel).

category name correlated with shape more than colour or texture, the network had a higher propensity for classifying novel objects based on shape rather than colour or texture. In other words, the network learns to reflect the feature bias of the training set; when the biased feature is a shape, the network shows shape-bias.

Both these studies assume that shape-bias is a property of the environment itself. Feinman and Lake (2018) explicitly make shape more diagnostic than any other feature in the dataset, while Ritter et al. (2017) assume that this is implicitly the case. However, it is not not clear that shape is necessarily the most diagnostic feature in the environment of biological systems and it is also unclear whether deep neural networks would develop an inductive bias for shape when this is not the most diagnostic feature. Our goal in this study was to test the stronger claim that CNNs show a shape-bias even when there is no such bias in the dataset. Within the psychological literature it is still unsettled whether our visual system identifies objects on the basis of shape because we learn through experience that shape is the most reliable cue to object identification or because there are innate inductive biases that make shape a privileged cue from the beginning (for discussion see Elman, 2008; Xu, Dewar, & Perfors, 2009).

It is certainly possible that CNNs have an inductive bias to rely on shape given that the depth of the architecture and pooling operations enables them to combine features of the stimuli in a hierarchical manner where lower layers represent high-frequency features while higher layers represent more abstract features, such as the shape, which are invariant to local changes of input (Bengio, Courville, & Vincent, 2013). If shape emerges due to this hierarchical composition of features, it is possible that it is preferred to other features (such as colour or texture) that do not lend themselves to such a hierarchical composition. Henceforth we use the term shape-bias to refer to the hypothesis that the visual system has an innate inductive bias to rely on shape cues to identify objects rather than the view that the visual system learns to identify objects on the basis of whatever visual cues are most strongly associated with object category.

Here we systematically explore the impact of non-shape features in the categorisation performance of convolutional neural networks on CIFAR-10 images. We introduced non-shape features to images by adding informative noise-like masks to the training set. We tried several types of masks and an extreme version where the non-shape feature consisted of just a single pixel with a location correlated to the image category (see Figure 1). We show that vanilla CNNs, that perform object classification on CIFAR-10 to near human level, nevertheless learn and depend on non-shape features that are highly diagnostic of object categories and often fails to learn anything about shape under these conditions. These results did not depend on the type of network architecture used, the learning algorithm or regularisation method indicating that this was a property of a broad class of CNNs rather than the particular setup chosen by us. This highlights that, even though they mimic the hierarchical architectural and learning processes of biological vision, the vanilla architectures and algorithms for learning in CNNs simply pick up whatever statistical structure is most relevant to learning the training set, with shape playing no special role. To dispel any confusions at the outset, we would like to emphasise that this does *not* imply that CNNs do not encode shape information under any circumstance, but that shape does not seem to be weighted more than other diagnostic features, even when these features are noise-like masks or the luminance of a single pixel.

Experiments

We modified the CIFAR-10 dataset (which contains 10 classes with 6000 images per class, see `https://www.cs.toronto.edu/~kriz/cifar.html`) so that each image contained not only features that pertain to the shape (e.g. object outlines) but also features without any shape information. As non-shape features we used noise-like masks that were combined with the original image. Two different types

of masks were used: the *salt-and-pepper noise mask* turned a certain proportion of image pixels to either black or white, while a *additive uniform noise mask* added a value sampled from a uniform distribution to each pixel of an image. We also tested an extreme form of the salt-and-pepper noise mask where only one pixel was turned to a particular colour. In this case the location and colour of the pixel were different for different categories but correlated for images within a category. Masks were independently sampled for each category but were either fixed for all images in a category (in which case the mask predicted the category) or sampled from a distribution with category-dependent parameters (in which case these parameters predicted the category). So these modified images concurrently contained features that were related to shape and features without shape information.

We trained the model on these modified sets of images and tested it under three conditions. During the 'Same' condition, the test set was modified in exactly the same manner – i.e., either images in each category were generated by using the same mask as that for the training images of that category (when the mask was fixed) or they were generated by using the same parameters as the parameters used to generate noise masks for training images of that category (when the mask was variable). In contrast, during the 'Diff' condition, the noise masks (or their parameters) for each category were swapped with another category. So, for example, a noise mask that was used in the 'DOG' category during training was inserted into images in the 'CAT' category during testing. The premise here was that if the model based it's decisions on shape-related features, then it would ignore the noise mask and the performance during 'Same' and 'Diff' condition should be similar. On the other hand, if the model relied on properties of the (non-shape) mask, then it's performance would be worse in the 'Diff' condition compared to the 'Same' condition. Finally, we used a third, 'NoPix', condition to estimate the extent to which the network relied on features of the noise mask. In this condition, we presented the network with a version of the image without any mask, with the premise that the difference between the performance in 'Same' and 'NoPix' condition should quantify the relative extent to which the network relied on shape-based and non-shape features. We ran all of the simulations using the well-known VGG-16 network (Simonyan & Zisserman, 2014) and checked that our main results replicate for a deeper network, ResNet-101 (He, Zhang, Ren, & Sun, 2016). To give the model the best chance to recognise shape-based features, all simulations were carried out on CNNs that had previously been trained on ImageNet categories and replaced only the fully-connected layers to perform the new classification task. We then turned the learning rate to a small value and trained these networks on the new classification task.

Methods

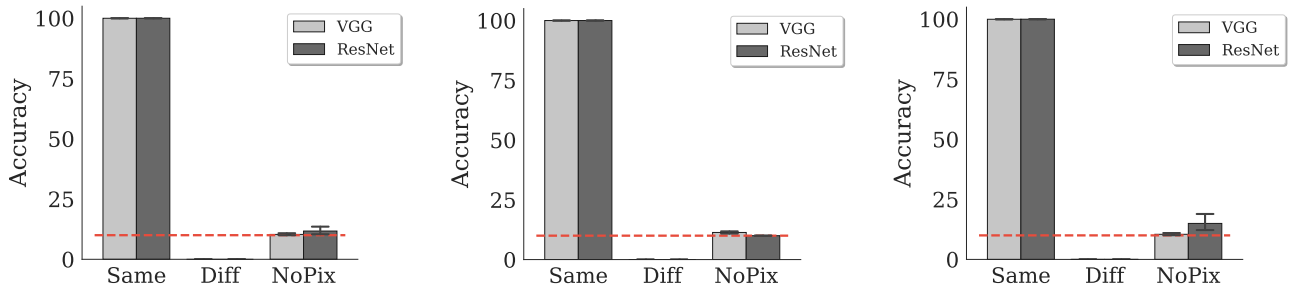
We used a method similar to Geirhos et al. (2017) to transform images from the CIFAR-10 dataset. All transformations were performed using the Pillow fork of the Python Imaging Library (`https://pillow.readthedocs.io`). Each 32x32 pixel image was rescaled to 224x224 pixels using the `PIL.Image.LANCZOS` method. For the single-pixel mask, we used 3-channel RGB images while for the salt-and-pepper and additive noise mask, we transformed images to greyscale. When images were transformed to greyscale, their contrast was adjusted to 80% by scaling the value of each pixel using the formula: $0.8 \times v + \frac{1-0.8}{2} \times 128$, where v was the original value of the pixel in the range $[0, 255]$.

The salt-and-pepper mask was created by taking the transformed greyscale image and setting each pixel to either black or white with a probability p. When the mask was fixed for a category (Experiment 1–3 below), all images had the exact same set of pixels that were turned either black or white and the p was set to 0.05. When the mask varied from image to image within a category (Experiment 4 below), the pixels were sampled independently for each image and the probability p was fixed for each category but varied between categories in the range $[0.03, 0.06]$.

The additive uniform noise mask was created by taking the transformed greyscale image and adding a value sampled from the uniform distribution $[-w, w]$ to this image, where $2w$ was the width of the uniform distribution and was set to 8. When the noise mask was fixed, this sampling was done only once per category and the same mask was added to each image. When the mask was variable, it was sampled independently for each image from a distribution $[\mu - w, \mu + w]$, where μ was the mean that depended on the category and varied in the range $[-50, 50]$.

The single pixel mask was created by choosing a random location, (x, y), (sampled from a uniform distribution on the interval $[0, 224]$) on the image and changing the colour of the pixel to a value c (sampled from a uniform distribution on the interval $[0, 255]$). When the mask was fixed for each category, (x, y, c) remained constant for all images in a category, but varied between categories. When the mask was variable, each of x, y and c were sampled independently for each image from a Gaussian distribution with a constant variance and a mean that depended on the category of the image. If any value in a sampled set of (x, y, c) values fell out of their respective range, that value was re-sampled.

Simulations were carried out using either a 16-layer VGG network (Simonyan & Zisserman, 2014) or 101-layer ResNet network provided by the `torchvision` package of `Pytorch`. These networks were either trained from the scratch on the modified dataset or were first pre-trained on ImageNet and then trained on the modified dataset. When the networks were pre-trained, we replaced the fully-connected layers of the VGG/Resnet pre-trained model with three/one fully-connected layer(s) with 10 units (for 10 categories) on the output layer. Since the results remain qualitatively the same, we report the results for the networks pre-trained on ImageNet. We tried a number of different optimization algorithms, including RMSProp, SGD and Adam (Kingma & Ba, 2014). Results again remained qualitatively the same. We



(a) Salt-and-pepper noise (b) Additive noise (c) Single diagnostic pixel

Figure 2: Accuracy on test images under the three types of noise-like masks shown in Figure 1. 'Same': the noise-like mask has same properties for test and training images of each category; 'Diff': the properties of the mask during test are swapped with another category from training; 'NoPix': No mask is inserted. The dashed (red) line indicates chance performance and error bars show 95% confidence interval. Light and dark gray bars show accuracies on `VGG-16` and `ResNet-101`.

started with a learning rate of $1e-3$ when training the network from scratch and used a learning rate of $1e-5$ when fine-tuning a pre-trained network. In all cases, we used cross-entropy as the loss function. The input to both types of networks was a 3-channel RGB image. For greyscale images, all three channels were set to the same value.

Experiment 1

In the first experiment, all images in a category had the exact same noise mask. For salt-and-pepper mask, this meant that noise masks were sampled independently for each category, but the same set of pixels in each image were modified for all images in a category. Similarly, for the additive uniform noise mask, the same mask was added to each image in a category. For the single pixel noise, the location and colour of the added pixel were independently sampled for each category, but kept constant for all images in a category.

The results of the first experiment are shown in Figure 2. We obtain the same pattern of results for all three cases: when noise mask in the test images matches the noise mask in training images, the model classifies images nearly perfectly; when noise masks are swapped, the accuracy drops to zero; when the mask is completely removed, the categorisation accuracy is at chance. Furthermore, we get the same pattern of results on both VGG and ResNet networks and irrespective of the type of regularisation used (we tried several well-known regularisation methods including *Batch Normalization*, *Weight Decay* or *Dropout*). These results clearly indicate that the model learns to completely rely on features of the noise-like mask, rather than any shape-related information present in the images. Even in the extreme case, where only one pixel amongst 50176 was diagnostic of the category, the model prefers to classify based on this feature over other shape-related features present in each image.

Experiment 2 & 3

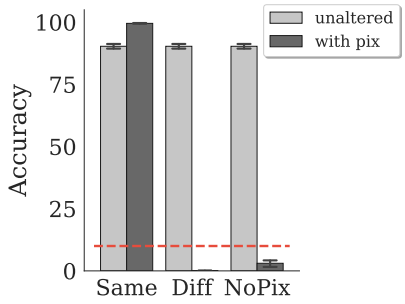
One possible reason why humans prefer to rely on shape-related features to categorise objects while CNNs do not is that humans are guided by past experience and bring this past knowledge to new categorisation tasks. So when a human sees an object with superimposed noise, they generalise from past experience and look for shape-based information, paying less attention to non-shape related features such as the noise-like mask in above images. We conducted two further experiments to test whether networks similarly generalise from concurrent and past experience.

In Experiment 2, we divided the training set into two subsets. The first subset ('with pix') contained three randomly chosen categories from CIFAR-10 and, like above, contained a category-correlated pixel in all images of these categories. The second subset ('unaltered') contained the remaining seven categories from CIFAR-10 and was left unaltered – i.e. we did not add the category-correlated pixel to images of this subset. We trained a VGG-16 network on all ten categories at the same time. We were interested in finding out whether the network generalised from one subset to another and started using the features used to categorise images in the 'unaltered' subset to images of the 'with pix' subset. All other details of the experiment remain same as Experiment 1.

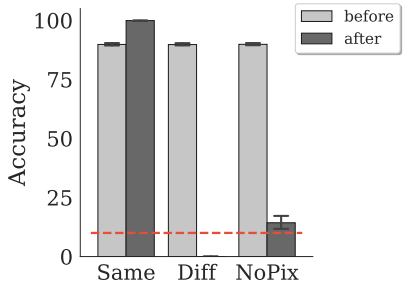
The results from this experiment are shown in Figure 3a. The model learnt to predict the images in the 'unaltered' subset with nearly 90% accuracy. However the performance on the 'with pix' subset still completely depended on the location and colour of the added pixel: accuracy was nearly 100% when test images contained the pixel in the same location, but dropped below chance when this pixel was removed. Thus, the network did not seem to generalise the features (concurrently) learnt in the 'unaltered' categories to the categories containing the diagnostic pixel.

In Experiment 3 we tested what happens when the network is first trained on images that did not contain such a pixel (a 'before' phase) followed by a second ('after') phase in which such a pixel was inserted in the training set. In the first phase, a we trained a VGG-16 network on an unaltered CIFAR-10 training set. Once the network had learnt this task, we trained it on the modified set of images in a second phase, introduc-

ing a predictive pixel in each category. So all that changes between the 'before' and 'after' phases is the insertion of a single category-correlated pixel to each image.



(a) Generalising between subsets



(b) Generalising from one time to another

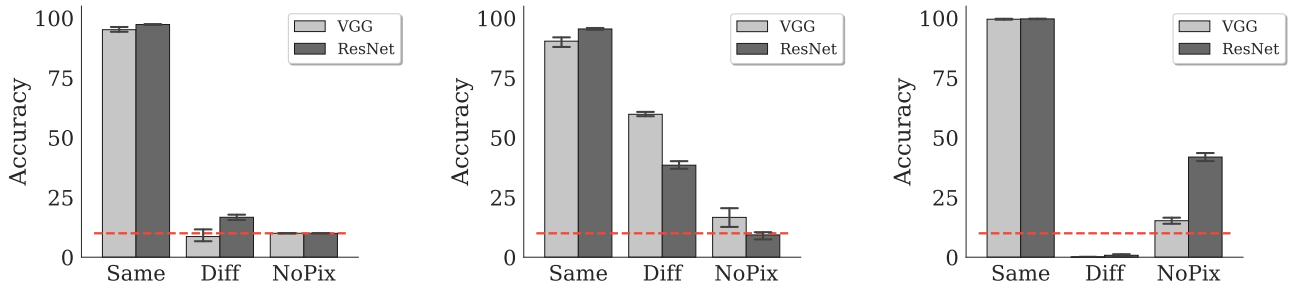
Figure 3: Lack of generalisation. Accuracy under Same, Diff and NoPix conditions for (a) two subsets: an 'unaltered' subset where no noise-like mask was inserted in training images and a 'with pix' subset where a single diagnostic pixel was inserted, and (b) for two phases: a 'before' phase, where a pre-trained VGG network was trained on images without any noise masks and tested on the three conditions, and an 'after' phase, where the model from before phase was then trained on images with a single diagnostic pixel.

We observed that (Figure 3b), instead of relying on past experience with these images, the model learnt to completely rely on the predictive pixel to perform categorisation – accuracy dropped from nearly 100% to 0% between 'Same' and 'Diff' conditions. Crucially, the model completely forgot about how to perform categorisation when the predictive pixel was removed – accuracy was close to chance in the 'NoPix' condition during the 'after' phase. Thus learning about the diagnostic feature seemed to be accompanied by unlearning previously learnt representations. This, catastrophic forgetting, is a well-known problem in neural networks (McCloskey & Cohen, 1989) and contrasts with how humans transfer their knowledge from one task to another. Some recent solutions to catastrophic learning in neural networks have been suggested, such as Elastic Weight Consolidation (Kirkpatrick et al., 2017) and it remains to be seen whether this can overcome some of these problems.

Experiment 4

The non-shape features used in the experiments above have all been completely invariant from one image to another within a category. It can be argued that these features are selected by the model over other shape-based features because they provide a very strong predictive signal. It is possible that if these features contained larger variance, the model would be more likely to rely on shape-based features while performing categorisation. In the next experiment, we introduced variability in the non-shape features by sampling the noise-like mask independently from a distribution for each training and test image within a category. In order to make these noise-like masks diagnostic of an image's category, a parameter of this distribution correlated with an image's category. For the salt-and-pepper noise, this meant that the probability, p, of changing a pixel to black or white was different for each category. Thus, the parameter, p, became diagnostic of the category. However, the masks now varied from image to image and were independently sampled with the (category-dependent) probability, p. Similarly, for the additive uniform noise, masks could vary from one image to other within a category but the mean of the distribution depended on each category (see Methods above for details). For the single diagnostic pixel, the inserted pixel could vary in location and colour from one image to the other, but were generated from a Gaussian distribution with a mean determined by the category of the image and a fixed standard deviation. We ran these simulations on both VGG-16 and Resnet-101 and aside from the way in which the dataset was generated, all other details remain same as Experiment 1.

The results of introducing a variable noise mask are shown in Figure 4. Introducing variability in the location and colour of the single diagnostic pixel brought very little change to the VGG model's behaviour (compare Figure 4c with Figure 2c). Performance in the NoPix condition was somewhat better for ResNet, however the pattern of result remained the same – performance dropped substantially from the Same to NoPix condition. Similarly, introducing variability in the salt-and-pepper masks lead to only a minor change in behaviour of the model, with accuracy in 'Diff' condition dropping to chance, rather than 0%. The most intriguing change in behaviour occurred when variability was introduced to the additive uniform noise mask (Figure 4b). While the VGG and ResNet networks differed quantitatively in these results, the pattern of results remained the same: when the noise mask was completely removed (NoPix condition) the model performed *worse* than when the images contained a noise mask from a different category (Diff condition). In other words, removing the mask makes the image less informative for the model, not only compared to images with the correct category-correlated (Same) mask, but also compared to images with the incorrect (Diff) mask – the model seems to rely on the presence of the noise-like mask to make an inference.



(a) Salt-and-pepper noise (b) Additive noise (c) Single diagnostic pixel

Figure 4: Accuracy on test images when the noise mask varies between images of a category. Training images contain (a) salt-and-pepper noise, or (b) additive uniform noise, or (c) just one diagnostic pixel. The dashed (red) line indicates chance performance. See Figure 2 for a description of the 'Same', 'Diff' and 'NoPix' conditions.

Related Work

Su, Vargas, and Kouichi (2017) demonstrated that CNNs trained on CIFAR-10 and ImageNet can be fooled by introducing a single adversarial pixel, with error rates of 68% and 41%, respectively. Unlike our approach the model was trained with uncorrupted images and the authors systematically searched for an adversarial pixel that lead to any sort of error (so-called non-targeted attack). So, in contrast to our goal, the goal of their study was not to explore whether CNNs systematically learn non-spatial information. However, the findings are in line with ours – the CNNs trained by them do not seem to be categorising based on shape. Rather, it must be that there was, by chance, some pixel value that was highly correlated with a given output category and the model picked up on this idiosyncratic correspondence. As a consequence, when this pixel was added to another category the model was fooled.

Two recent studies – Geirhos et al. (2018) and Baker, Lu, Erlikhman, and Kellman (2018) – manipulate the texture and shape of images independently and show that CNNs trained on ImageNet are biased towards picking up texture compared to shape. These results are again in line with our results and show that CNNs will make inferences on whichever feature is most predictive in the training set. Indeed, when Geirhos et al. (2018) make the texture less diagnostic of category, the model seems to use non-texture features for performing classification. Our findings go beyond past work by highlighting the extent to which CNNs categorize objects on the basis of non-shape features even when it is given concurrent or prior training without such non-shape features. Indeed, a single diagnostic pixel can override all the shape information present in the training images.

Conclusions

In a series of simulations we found that some high-performance convolutional networks trained to categorise CIFAR-10 images that included noise-like masks diagnostic of the output categories often learned to categorise on the basis of these masks rather than features present within the CIFAR-10 images themselves. Indeed, the models often entirely relied on the masks, and performed at floor when the noise was removed from the images. This clearly highlights that, when a shape-bias is not present within the training dataset itself, these models do not show a shape-bias due to their own architectural or algorithmic properties.

In our experiments, we specifically engineered our dataset to contain invariant non-shape features. One might object that large datasets like ImageNet and CIFAR-10 don't contain such features so that the models trained on these datasets end up relying on shape to perform categorisation. But it is well-known that popular datasets contain various biases due to conditions under which the images were captured as well as the different motivations for construction of the datasets (Torralba & Efros, 2011). So biases like the one we engineered may well be present in these datasets and networks trained on these datasets may be picking on these features. This, in turn, implies that these networks may be relying on entirely different set of features and representations to perform classification than human beings or other animals.

If CNNs do indeed rely too heavily on non-shape features present within datasets, it could also be the source of various idiosyncratic behaviours such as being confounded by fooling images (Nguyen et al., 2015) or being overly sensitive to colour (Hosseini et al., 2017), noise (Geirhos et al., 2017) or even single pixels in images (Su et al., 2017). The alternative hypothesis that the human visual system learns to categorize objects on whatever statistical regularities are strongest in the input cannot be ruled out on the basis of our findings, but it would predict that humans would show a similar pattern of result to these models, such as picking up on single pixels or noise-like masks to categorise stimuli. In addition, this view also needs to explain why human beings are not susceptible to adversarial attacks such as the non-shape fooling images in the same manner as vanilla CNNs. We are currently carrying modelling and behavioural work to provide further insights into the computational benefits of inducing a shape-bias to CNNs and how these modified CNNs relate to human vision.

References

Baker, N., & Kellman, P. J. (2018). Abstract shape representation in human visual perception. *Journal of Experimental Psychology: General*, *147*(9), 1295.

Baker, N., Lu, H., Erlikhman, G., & Kellman, P. J. (2018). Deep convolutional networks do not classify based on global object shape. *PLoS computational biology*, *14*(12), e1006613.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, *35*(8), 1798–1828.

Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychological review*, *94*(2), 115.

Biederman, I., & Ju, G. (1988). Surface versus edge-based determinants of visual recognition. *Cognitive psychology*, *20*(1), 38–64.

Elman, J. L. (2008). The shape bias: an important piece in a bigger puzzle. *Developmental science*, *11*(2), 219.

Feinman, R., & Lake, B. M. (2018). Learning inductive biases with simple neural networks. *arXiv preprint arXiv:1802.02745*.

Geirhos, R., Janssen, D. H., Schütt, H. H., Rauber, J., Bethge, M., & Wichmann, F. A. (2017). Comparing deep neural networks against humans: object recognition when the signal gets weaker. *arXiv preprint arXiv:1706.06969*.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2018). Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 770–778).

Hosseini, H., Xiao, B., Jaiswal, M., & Poovendran, R. (2017). On the limitation of convolutional neural networks in recognizing negative images. In *Machine learning and applications (icmla), 2017 16th ieee international conference on* (pp. 352–358).

Hummel, J. E. (2013). Object recognition. *Oxford handbook of cognitive psychology*, 32–46.

Jozwik, K. M., Kriegeskorte, N., Storrs, K. R., & Mur, M. (2017). Deep convolutional neural networks outperform feature-based but not categorical models in explaining object similarity judgments. *Frontiers in psychology*, *8*, 1726.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., ... Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, *114*(13), 3521–3526.

Kubilius, J., Bracci, S., & de Beeck, H. P. O. (2016). Deep neural networks as a computational model for human shape sensitivity. *PLoS computational biology*, *12*(4), e1004896.

Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive development*, *3*(3), 299–321.

Mapelli, D., & Behrmann, M. (1997). The role of color in object recognition: Evidence from visual agnosia. *Neurocase*, *3*(4), 237–247.

McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation* (Vol. 24, pp. 109–165). Elsevier.

Nguyen, A., Yosinski, J., & Clune, J. (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 427–436).

Rajalingham, R., Issa, E. B., Bashivan, P., Kar, K., Schmidt, K., & DiCarlo, J. J. (2018). Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *Journal of Neuroscience*, 0388–18.

Ritter, S., Barrett, D. G., Santoro, A., & Botvinick, M. M. (2017). Cognitive psychology for deep neural networks: A shape bias case study. *arXiv preprint arXiv:1706.08606*.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Su, J., Vargas, D. V., & Kouichi, S. (2017). One pixel attack for fooling deep neural networks. *arXiv preprint arXiv:1710.08864*.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 2818–2826).

Torralba, A., & Efros, A. A. (2011). Unbiased look at dataset bias. In *Ieee conference on computer vision and pattern recognition (cvpr)* (pp. 1521–1528).

Xu, F., Dewar, K., & Perfors, A. (2009). Induction, overhypotheses, and the shape bias: Some arguments and evidence for rational constructivism. *The origins of object knowledge*, 263–284.

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2016). Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*.

How Many Dimensions of Mind Perception Really Are There?

Bertram F. Malle
Brown University

Abstract

Previous research suggests that people's folk conception of the mind is organized along a few fundamental dimensions; but studies disagree on the exact number of those dimensions. With an expanded item pool of mental capacities, variations of question probes, and numerous judged agents, four studies provide consistent evidence for three dimensions of perceived mind: *Affect (A), Moral and Mental Regulation (M)*, and *Reality Interaction (R)*. The dimensions are not simply bundles of semantically related features but capture psychological functions of the mind—to engage with its own processes, with other minds, and with the social and physical world. Under some conditions, two of the three dimensions further divide: *A* divides into negative and positive (social) affect, and *M* divides into moral cognition and social cognition. We offer a 20-item instrument to measure people's 3- and 5-dimensional representations of human and other minds.

Keywords: anthropomorphism; social cognition; theory of mind; morality; principal component analysis; robots.

Introduction

A significant question for cognitive science is how humans conceptualize agents and their minds. Research in cognitive development has taught us that features such as self-propelled motion, contingent response, and eyes convince infants that they are interacting with a special category of thing: what we call agents (Johnson, 2000; Premack, 1990). Once infants identify agents, they follow their gaze, imitate them, make inferences about their goals, and eventually ascribe complex mental states to them. Over the childhood years, children develop ever more differentiated conceptions of mental states, such that, for example, a goal concept divides into desires and intentions, emotion concepts of good and bad differentiate into a staggering number of different affective terms, and moral dispositions of *mean* and *nice* turn into sophisticated assessments of moral character. In short, we know that humans grow up to have deep-seated expectations about other humans' mental, social, and moral capacities (Hamlin, 2013; Malle, 2005; Tomasello, 2003). But do people treat these capacities as just one long list? Or is there an underlying conceptual organization to uncover? Only empirical studies can answer this question.

Dimensions of Mind

Despite humans' rich representations of mental capacities, previous work indeed suggests that there are fundamental dimensions by which humans organize these capacities. But research diverges on the number of those dimensions.

D'Andrade (1987) considered six categories of mental states: perception, belief, emotions, desires, intentions, and self-control. Interviews suggested that people indeed make distinctions among these classes, but no methods were applied to assess whether the researcher-imposed category number actually captured people's own conceptual structure. Haslam et al. (2008) rearranged d'Andrade's categories, combining intentions and self-control but separating primary from secondary emotions. Through multi-dimensional scaling, they were able to reduce these categories into a two-dimensional space: perception vs. all other categories; thoughts and intentions vs. desires and emotions.

Gray, Gray, and Wegner (2007) offered the simple and elegant proposal that humans distinguish mental states along two dimensions: Experience and Agency. The empirical evidence for this two-dimensional structure was a principal component analysis (PCA) of 18 mental capacities, which people evaluated in 13 different agents. This proposal had seminal impact in research on dehumanization, moral judgment, objectification, and human-robot interaction.

The interpretation of what makes up the two-dimensional space of Experience and Agency is, however, not entirely clear. Even though each dimension in Gray et al.'s study had several items that loaded high on its dimension and low on the other dimension, there were numerous items that loaded high on both dimensions, showing barely distinguishable loadings (see Table 1). Moreover, although the highest-loading Experience items are quite coherent, incorporating physiology and affect, the highest-loading Agency items are more heterogeneous, including planning and self-control as agentic capacities but also emotion recognition, memory, and morality, which are less obviously agentic.

Table 1: Loading matrix for PCA of 18 mental capacities by Gray, Gray, and Wegner (2007).

	Experience	Agency	*Difference*
Hunger	0.98	0.15	*0.83*
Fear	0.93	0.31	*0.62*
Pain	0.89	0.42	*0.48*
Pleasure	0.85	0.51	*0.34*
Rage	0.78	0.59	*0.20*
Desire	0.76	0.64	*0.12*
Joy	0.68	0.61	*0.07*
Personality	0.72	0.68	*0.04*
Consciousness	0.71	0.69	*0.03*
Pride	0.71	0.69	*0.03*
Embarrassment	0.70	0.65	*0.05*
Thought	0.68	0.73	*0.05*
Communication	0.66	0.74	*0.08*
Planning	0.55	0.82	*0.27*
Emotion recognition	0.54	0.83	*0.29*
Morality	0.36	0.93	*0.57*
Memory	0.33	0.91	*0.58*
Self-control	0.18	0.97	*0.79*

Note: Clearly and highly loading items on each dimension are color-marked. Items in the middle show almost no difference in their loadings on the two dimensions.

Replications by Takahashi, Ban, and Asada (2016) and Weisman, Dweck, and Markman (2017) confirmed the Experience dimension with its familiar marker items but continued to find several middling items (in particular, personality, consciousness, pride, and embarrassment) as well as considerable heterogeneity on the Agency dimension.

Other studies suggest that people may conceptualize the mind in three rather than two dimensions. Kozak, Marsh, & Wegner (2006) applied a PCA to 10 items similar to those in Gray et al. (2007) and identified three dimensions, labeled Emotion (feelings, pain, emotion, pleasure; hence similar to Experience), Intention (doing things on purpose, planning, goals; hence similar to Agency), and Cognition (conscious, memory, thought). Using a larger item pool of 40 mental capacities, Weisman et al. (2017) found three major dimensions, which they labeled Body (related to Experience), Heart (primarily covering emotions), and Mind (perceptions, cognition). Thus, Agency did not emerge in this structure.

The Present Investigation

How do people represent and conceptualize capacities of the mind, and what number of fundamental dimensions underlie this representation? To answer this question, we need a comprehensive item pool. As noted by several authors (Haslam et al., 2008; Weisman et al., 2017), the original 18 capacities used by Gray et al. had limitations (e.g., perception items were missing, some categories were represented by single items). Only an expanded item pool and replications across different pools can reveal the dimensions of mind perception. Across four studies, we therefore analyzed varied item pools that represent capacities of perception, cognition, emotion, agentic control, learning, communication, and social-cognitive and social-moral capacities, all represented by multiple items. For consistency, one constraint was to include items about which one could explicitly ask, "Is the agent capable of X?" This question disfavors highly specific states (e.g., feeling disrespected) and abstract words such as "personality." Across studies we experimented with different items and formulations in order to gain confidence in the clusters of capacities that best represent the dimensions of mind perception. In analogy to cognitive theories of concepts, we conceive of such dimensions as bundles of capacities typically represented together; if similar dimensions of mind reappear across variations in items and samples, we can be more confident in the underlying dimension in question.

The conceptual structure of mental capacities is difficult to study when asking participants to indicate how much of each capacity human adults have, as the ratings will tend to be at ceiling. Following other authors, we increased judgment variance by including nonhuman agents, which arguably lack some of the capacities. Particularly useful targets are robot agents, as the reality of their minds is a wide open question. Robots are like a projection screen for people's general conceptions of mind, so these conceptions may emerge particularly well when people judge robots' minds.

In the present project, we thus investigated how many dimensions may be fundamental in people's representations of various agents' minds. We report on a first study in detail to lay out our methodological approach and major results, then summarize the results of three additional studies that varied the pool of capacities and tested different question probes and judged agents. We then report on a final study that relied on an integrated item pool derived from multiple previous data sets so as to represent the full conceptual range of people's perceptions of mind. Based on these results, we offer a parsimonious multi-dimensional measurement scale of mental capacities applicable to humans and other agents.

For instructions, item formulations, and detailed results tables, please see the Supplementary Materials (SM), which can be found at http://bit.ly/SA_MindCapacities.

Study 1

Methods

To generate a broad item pool we took Gray et al.'s item pool as a starting point and classified them into four rough groups: physiological (hunger, pain), affective (joy, pride, desire, pleasure, rage, fear, emotion recognition), cognitive (remember, planning, thinking), and agentic (self-control, communicate). Taking Sytsma and Machery (2010), Haslam et al., (2008), and d'Andrade (1987) as inspiration, we added two items to the agentic group (choosing freely, imitating others) and two to the physiological group (sleep, thirst) to make them four each. We retained six of the affective items (reformulating emotion recognition into empathy); decomposed "thinking" into more concrete cognitive capacities (believing, knowing, deliberating, reasoning) to make the total of cognitive items six as well. We added four perceptual items (perceive, see or hear, taste or smell, vividly imagine) and differentiated morality into four items (moral obligations, having values, deserving praise or blame, deserving punishment). We omitted the two most abstract items of personality and consciousness, as well as embarrassment, all of which were undifferentiated in Gray et al. (see Table 1).

Participants were 160 undergraduate students from a private university in the Northeast United States; no demographic information was collected. In a one-page survey, each participant rated one of 16 agents (e.g., human adult, robot, rabbit, chimpanzee, similar to Gray et al.'s agents, but also group agents, such as a city council and a large company). Twelve participants were excluded, two how provided illegible ratings, ten who had a rating range of 0 or 1 on the 8-point scale, leaving 148 participants for analysis. Fewer than 1% of individual item ratings were mising and were replaced by their respective sample means.

On the top of the survey page, the agent was introduced, and each statement repeated the agent description (e.g., "The most advanced robot in 2050 can feel joy," "…can have values," "can perceive things.") The 28 statements were listed in random order with rating scales next to each statement. The column header for the ratings contained the question, "Is this true?", and the anchors for the ratings scales were "Definitely NOT true" (0) and "Definitely true" (7).

We used Principal Component Analysis (PCA) to analyze the correlation matrix resulting from the 148 (participants) × 28 (capacities) raw data, ignoring agent type, which served as a source of meaningful judgment variability. One challenge of PCA is that multiple criteria are available to decide how many components one should extract. Common heuristics include Kaiser's (1960) rule ("K1"; retain components with eigenvalue $\lambda > 1$) and Cattell' (1966) scree test (on the scree plot, draw a linear fit line from the smallest components upward and retain those that lie above the line). However, with larger variable sets, K1 extracts too many factors, and the scree test can suffer from ambiguity. Zwick and Velicer (1986) compared these and more sophisticated criteria and concluded that Parallel Analysis (PA) represents the best approach. This procedure (Buja & Eyuboglu, 1992) recognizes that even for a population of perfectly uncorrelated variables, any sample from it will contain correlations among variables that a PCA would pick up and turn into spurious components with $\lambda > 1$. By assessing hundreds of random permutations of the actual data matrix, PA estimates what number and size of spurious components one can expect if the original data were in reality uncorrelated (i.e., all λs = 1). The recommendation is then to retain those components from the actual PCA whose eigenvalues are at or above the corresponding spurious ones.

Results

The K1 and scree criteria suggested 4 components, but the fourth was very weak, $\lambda = 1.04$. PA suggested 3 components. The three-component solution accounted for 67.3% of the total variance and was interpretable after rotation (see Table 2). The first component had 25.1% explained variance (EV) and grouped 11 items with loadings $l \geq .60$, both social-moral capacities (shame, values, obligations, praise) and cognitive control capacities (believing, deliberating, choosing). We label this component *Moral and Mental Regulation* (*M*). The second component (22.7% EV) grouped 8 physiological and affective items together (e.g., hunger, pain, taste, anger, joy); we label this component *Affect* (*A*). The third component (15.6% EV) grouped perceptual, cognitive, and some interaction items (perceive, remember, know, communicate), which we label *Reality Interaction* (*R*).

To illustrate in a heuristic way how much a loading matrix approximates *simple structure* (D'Agostino & Russell, 2014) we counted items with "errand loadings"—defined as $l > .316$ (i.e., > 10% of variance) on components that are not the item's primary component (where it loads most highly). Of all possible 84 loadings, 15 (18%) were errand in this way. To examine the possibility of component correlations we applied oblique rotation to all 28 items, which reduced errand loadings to 11% (which is expected for oblique rotations). This solution showed small correlations between *A* and *M* ($r = .21$) and between *A* and *R* ($r = .18$) and a more notable one between *M* and *R* ($r = .42$). Thus, Agency from Gray et al.'s (2007) two-dimensional model broke into two dimensions that may, however, not be entirely independent.

Discussion

Study 1 recovered the Experience dimension from previous studies (here, labeled *Affect*), but by expanding the item pool to represent domains of perception, cognition, and morality we uncovered a third dimension of mind perception. Specifically, the Agency dimension (arguably multi-faceted to begin with) broke into two distinct dimensions. The original Gray et al. items of morality, empathy, and planning became part of a *Moral and Mental Regulation* dimension, whereas items of perception, cognition, and communication constituted a *Reality Interaction* dimension. These two dimensions can be treated as orthogonal, but in an oblique rotation they show a cleaner simple structure with a nontrivial correlation.

Importantly, the items that define each dimension hang together not simply due to semantic similarity (e.g., moral and mental regulation are semantically distinct). The items constitute their components in psychologically meaningful ways. For example, *R* refers to a progression of information processing from perceiving to knowing to remembering to communicating. Likewise, *M*'s cognitive facet forms a sequential process: we believe things, then deliberate, then choose and plan; and *M*'s moral facet refers to empathy, obligations, and values as action regulation and also includes responses to one's moral (or immoral) behavior in the form of pride or shame on the inside, praise or blame from the outside. Thus, moral and mental regulation occurs in a dynamic mental and social context and is the culmination of a complex and nuanced picture of the social-moral mind.

Table 2. Loading matrix of PCA on 28 items in Study 1.

	Moral & Mental Regulation	Affect	Reality Interaction
can feel shame or pride	0.79	0.35	0.20
can have values	0.76	0.10	0.26
may deserve praise or blame	0.76	0.18	0.04
may deserve punishment	0.74	0.12	-0.06
has moral obligations	0.71	0.02	0.28
can have empathy for others	0.68	0.23	0.21
can believe certain things	0.64	0.22	0.37
can deliberate	0.64	-0.01	0.49
can vividly imagine things	0.63	0.39	0.30
can plan for the future	0.63	-0.19	0.41
can choose freely	0.61	0.29	0.31
can feel thirsty	0.12	0.91	-0.04
can be in physical pain	0.09	0.88	-0.01
has a need for sleep	0.08	0.86	0.01
can feel hunger	0.08	0.82	0.05
can taste or smell things	-0.07	0.80	0.32
can experience pleasure	0.34	0.75	0.18
can be angry	0.45	0.64	0.08
can feel joy	0.46	0.62	0.18
can see or hear things	-0.08	0.60	0.59
can want certain things	0.42	0.51	0.28
can communicate with others	0.10	0.30	0.73
can remember things	0.11	0.09	0.67
can perceive things	0.37	0.12	0.66
can reason logically	0.50	-0.18	0.64
can know certain things	0.52	-0.02	0.63
can exercise self-control	0.35	0.20	0.61
can imitate others	0.46	0.03	0.47

Study 2

Encouraged by the effects of enlarging the item pool of mental capacities we further expanded the pool by rewriting several items for clarity and adding 30 new ones, for a total of 54, to represent (with several items each) physiology, affect, moral competence, social cognition, thinking and cognitive control, perception, learning, and communication. We thus allowed for the possibility of components from Study 1 breaking apart even further (thus pointing to more dimensions) or else clustering reliably around the same three dimensions, despite new item content.

We probed mental capacity ascriptions to an average adult, a two-year-old child, a cat, and a home care robot. Any given participant made judgments for only one agent. Of 459 participants recruited online via Amazon Mechanical Turk, 17 provided fewer than a quarter of ratings and 27 had a rating range of 0 or 1 (on an 8-point scale), leaving 415 participants for analysis. Of these, 45.3% identified as female, 53.5% as male. They ranged in age from 18 to 74 (M = 35.5, SD = 11.8), and 52% of them had completed a bachelor's degree or higher. In the principal component analysis (PCA), the K1 and scree criteria suggested five components, but PA suggested three. We considered a 4-component solution, but the fourth component accounted for less than 5% of the variance and had only three items with l > .50 and almost as high cross-loadings on other components. The 3-component solution (see Table SM3) explained 65.2% of the variance.

Table 3. Loading matrix of Orthogonal PCA on 38 selected items in Study 1

	Affect	Moral & Mental Regulation	Reality Interaction
Being hungry	0.91	-0.17	0.11
Feeling pain	0.91	-0.16	0.10
Feeling pleasure	0.91	-0.04	0.14
Feeling panic	0.90	-0.02	0.16
Feeling happy	0.90	-0.01	0.12
Having emotions	0.88	0.06	0.13
Getting angry	0.88	-0.02	0.19
Loving specific people	0.87	0.00	0.21
Having intense urges	0.86	0.00	0.17
Smelling and tasting things	0.83	-0.11	0.25
Having desires	0.83	0.09	0.18
Feeling stress	0.82	0.13	0.10
Disliking people	0.80	0.09	0.23
Feeling gratitude	0.72	0.39	0.04
Feeling compassion	0.72	0.38	-0.06
Vividly imagining things	0.65	0.33	0.11
Feeling sexual arousal	0.63	0.28	-0.03
Providing reasons for their actions	-0.18	0.87	0.04
Planning for the future	-0.02	0.86	0.09
Upholding moral values	0.25	0.85	-0.07
Understanding a person's goals	0.13	0.85	-0.03
Explaining their decisions to others	-0.22	0.84	0.03
Setting goals	-0.05	0.83	0.05
Praising moral actions	0.18	0.81	-0.07
Disapproving of immoral actions	0.29	0.79	-0.07
Reasoning logically	-0.21	0.78	0.24
Understanding others' minds	0.26	0.77	-0.03
Taking a person's visual point of view	0.02	0.74	0.02
Inferring what a person is thinking	0.12	0.74	0.09
Deliberating before acting	0.03	0.73	0.30
Exercising self-control	0.17	0.72	0.08
Following norms	0.11	0.70	0.10
Communicating verbally	-0.20	0.64	0.26
Moving on their own	0.20	0.07	0.78
Seeing and hearing the world around them	0.31	-0.03	0.74
Learning by imitation	0.12	0.27	0.63
Communicating nonverbally	0.31	0.18	0.63
Feeling temperature, touch, etc.	0.49	-0.01	0.62

The first component had 21 strongly loading items ($l \geq .60$), dominated by affective states (pain, hunger, stress), emotions (angry, compassion, gratitude), and social relations (loving people, relationships). We see here again the *Affect* dimension from Study 1, supplemented by social facets. The second component had 17 strongly loading items, capturing moral capacities (e.g., upholding values, praising moral actions), social cognition (e.g., understanding others' minds, their goals, and thinking), and cognitive control (e.g., setting goals, providing reasons for one's actions). We see here the *Moral and Mental Regulation* dimension, with enhanced social-cognitive facets. The third component included 7 strongly loading items, featuring seeing, learning, moving, and communicating, confirming the *Reality Interaction* dimensions of Study 1. The remaining items loaded more weakly or on multiple components, producing the bulk of the 15% errand loadings. Removing weaker and cross-loading items led to a set of 38 items that had only 2% errand loadings, yielding a clean three-dimensional structure (see Table 3). Oblique rotation on all items also reduced errand loadings (12%) and showed modest correlations (the highest between M and R at .30). Removing 12 weak items reduced errand loadings to 6% and the M*R correlation to .26 (see Table SM4).

In sum, we replicated a three-dimensional structure of mind perception. The previously labeled Experience factor is well represented by the *Affect* dimension, which includes social emotions and relations. The previously labeled *Agency* dimension again separated into one of *Social-Moral and Mental Regulation* and the dynamic dimension of *Reality Interaction* (perception, learning, to action).

Studies 3a and 3b

Now we report on two samples that we collected in continuation of a related project in which we focused on mental capacities people *would like to* see in robots, thus a slightly different question from the one in Studies 1 and 2. However, these studies had a considerable impact on our last stage of item selection and so we describe them briefly.

Though the rating means may differ between inferred capacities of various agents and desired capacities of robots in particular, the dimensional structure should still be similar. We presented participants in Study 3a (N = 100) with 60 mental capacity items largely the same as in Study 2, and participants in Study 3b (N = 99) with a selection of 41 items. The two samples were recruited online from Amazon Mechanical Turk and had highly similar demographics as those in Study 2. We invited people to indicate which capacities they would want or not want in "the most advanced home robot" they could imagine, defined as an autonomous robot that takes care of older adults or children and does household chores.

In Study 3a, K1 and scree criteria suggested 6 to 13 components, but PA suggested 3 to 4, so we examined both solutions. Each one yielded R (perception, cognition, learning) and A (but solely negative affective states). In the 4-component solution, the third and fourth component both

contained social emotions, relations, and hints of morality, and it was difficult to find a distinction between the two components. Indeed, in the 3-component solution the two combined into a dimension similar to *M* (but populated more with positive social emotions and relations than we saw in Studies 1 and 2), and errand loadings decreased from 14% to 11%. Oblique rotation reduced errand loadings to 6%, with the highest correlation between *M* and *R* at $r = .39$. Overall, the three-dimensional structure from Studies 1 and 2 was replicated even when probing people's desired capacities for robots. However, *A* became negative and *M* took on positive social emotions that had loaded on *A* in Studies 1 and 2. We will return to this trend in Study 4.

For Study 3b, we reduced the number of items to 41, omitting eight items with very low loadings in Study 3a, five that were semantically redundant with other items in the set, and two that plainly do not apply to robots (physiology, hunger). Four items were omitted due to a clerical error. K1 and scree criteria suggested 5 to 8 components, but PA suggested only 2 to 3. The 2-component solution dispersed familiar *M* items across both other item sets, making the solution difficult to interpret, even from an Experience-Agency perspective (see Table SM6). The 3-component solution showed three strong components after rotation (15.3% to 23.1% EV), replicating *A* (solely negative affective states), *M* (social emotions, relations, and moral capacities), *R* (perception, decision making, communication, and some stray social cognition), with 20% errand loadings. Oblique rotation improved the errand rate to 10%, with *M* and *R* correlating at .48, and at .40 after removal of very weak items.

Taken together, the two studies on desired mental capacities of robots largely supported a three-dimensional structure of mind perception. However, Study 3a raised the possibility of a split of *Affect* into a positive and negative facet, which we decided to explore further in Study 4. However, the primary purpose of Study 4 is explained next.

Integrative Item Selection

After we completed this first set of studies, Weisman et al. (2017) published a series of studies that suggested three dimensions of mind comparable to our three, thus providing further confidence in a three-dimensional model of mind perception. However, their components (to which we will refer as W1 to W3) differed somewhat from ours in item composition and in the authors' interpretation. W1 was labeled "Body," highlighting its physiological items, but almost half of its high-loading items refer to basic emotions (calm, angry, fear, safe). W2 was labeled "Heart," also highlighting emotion items (embarrassed, pride, love), but these emotions are social, and other items in this component also hinted at moral capacities (telling right from wrong, guilt) and cognitive control (thoughts, intentions, self-restraint), casting doubt on the labeling of "Heart." W3 was nonspecifically labeled "Mind," but it encompassed perception, memory, reason, and communication.

Aside from interpretational ambiguities, some of the discrepancies between Weisman et al.'s and our three-dimensional model can be explained by item selection, so to address this possibility, we collated the 62 items used at least twice across Gray et al., Weisman et al., Malle and Thapa Magar (2017), and our data reported so far. We tracked each item's loadings within the corresponding components across data sets. This correspondence was straightforward for our three components and Weisman et al.'s. (*A* ~ W1, *M* ~ W2, *R* ~ W3). Gray et al.'s first component clearly corresponds to *A* and all but one of the other reused items fit under *M*. We reanalyzed the data from Malle and Thapa Magar (2017) with the same criteria as we had applied in the present studies and found better support for a three-dimensional structure (rather than the originally reported four-dimensional structure), and the three dimensions were very similar to the present *A-M-R* structure. (See the resulting compilation matrix in Table SM8.)

We identified candidate items by using two inclusion heuristics: (a) A *differential loading* index was the averaged loading in a given component minus the averaged loadings on the other two components; we aimed for this difference to be at least .30. (b) An item's *number of replications* on the same component with a loading $l > .50$; we aimed for two or more such replications. We also used two exclusion heuristics: (c) content was already covered by another item; (d) item had substantial loadings on two components. We made specific attempts to retain enough items in the content domains of agency, perception, social emotions, and social cognition. The resulting item pool included 12 items targeting *A* (physiology, basic emotions and motivation), 20 targeting *M* (perhaps the most diverse dimension with social emotions, moral competence, social cognition, and cognitive control), and 10 targeting *R* (perception, learning, communication, action).

Study 4

In light of possible differences between the dimensional structure of inferred and desired capacities, which had arisen in Studies 3a and 3b, we asked one group of participants to *infer* the capacities of either an average adult, a two-year-old child, or one of two kinds of robots—a home robot or a military robot; and we asked a second group to indicate the capacities they *would like* a home or military robot to have. Of 495 participants recruited from Amazon Mechanical Turk, 11 entered no ratings, 2 entered fewer than half, and 19 had a rating range of 1 or 0 on an 8-point scale, leaving an N of 463, again with very similar demographics as those in Study 2.

We applied PCA to each question condition separately. In the inferred capacity group ($N = 304$), K1 and scree suggested three to four components, while parallel analysis suggested three, which were easily interpretable as the *A-M-R* structure (71.1% EV, 19% errand loadings). After removing only four items with $l < .60$, errand loadings decreased to 10% (73.1% EV). An oblique rotation yielded virtually not change, with *M* and *R* showing a small correlation of .32.

In the desired-capacity condition ($N = 159$), K1 suggested eight components, the scree plot suggested six and especially showed a fourth and fifth component distinctly separating

from the lower ones. Parallel analysis suggested three, but this solution (EV = 51.7%) was not interpretable as it intermixed items that in other studies consistently loaded in the familiar *A, M,* and *R* dimensions. When allowing a fourth component, the Affect items split into a negative set (e.g., anger, stress, pain) and a positive set (e.g., happy, gratitude, friendships), the moral items formed their own component, but *M* and *R* items remained intermixed. When allowing a fifth component, finally, rotation produced five evenly strong components (EVs = 10.6% to 13.2%), with *M* and *R* cleanly separating and errand loadings down to 11% (Table SM10). Under oblique rotation, correlations were moderate, with the highest between the social and moral component at .37.

Instrument development

The final step was to create a measurement instrument of people's perceptions of mind that accommodates both inferred and desired capacities and is also suitable for other applications. We aimed for five subscales representing the components of the desired-capacity set whereby the items of the negative and positive-social affect subscales would combine into an overall Affect scale and the items of the moral and social-cognitive subscales would combine into an overall Social-Moral scale, thus representing the three-dimensional structure of inferred capacities. Of the 42 items in the two analyses (inferred, desired) of Study 4, we removed 10 that had $l < .50$ or strong cross-loadings in at least one analysis, and 2 items that fell under distinct components in the two analyses. Then we selected the four highest-loading items in each of the five components, yielding a 20-item measure with sufficient internal consistency on each of the five subscales (see Fig. 1), and errand loadings of 5-7%. With only two items more than Gray et al. used, we can now measure three to five dimensions of mind perception.

	Desired Capacities		*Inferred Capacities*	
Feeling happy	0.84	**Positive Social Affect**	0.96	**Affect**
Loving specific people	0.80		0.91	
Feeling pleasure	0.79	*16.2% EV*	0.94	*36.7% EV*
Experiencing gratitude	0.77	α = 0.88	0.80	α = 0.98
Feeling pain	0.86	**Negative Affect**	0.97	
Feeling stress	0.81		0.87	
Experiencing fear	0.78	*15.3% EV*	0.94	
Feeling tired	0.77	α = 0.86	0.95	
Disapproving of immoral actions	0.80	**Moral Cognition**	0.83	**Moral & Social Cognition**
Telling right from wrong	0.77		0.73	
Upholding moral values	0.76	*13.5% EV*	0.84	
Praising moral actions	0.66	α = 0.84	0.81	*29.4% EV*
Inferring a person's thinking	0.78	**Social Cognition**	0.80	α = 0.94
Planning for the future	0.75		0.82	
Understanding others' minds	0.67	*12.6% EV*	0.84	
Setting goals	0.61	0.78	0.83	
Communicating verbally	0.81	**Reality Interaction**	0.67	**Reality Interaction**
Seeing and hearing the world	0.70		0.68	
Learning from instruction	0.69	*12.5% EV*	0.72	*11.2% EV*
Moving on their own	0.68	α = 0.79	0.76	α = 0.71

Figure 1: Individual item loadings from PCAs on desired mental capacities of robots (5 components, left) and inferred capacities of humans or robots (3 components, right).

General Discussion

What are the dimensions of mind? Our results suggest that people's ascriptions of mental capacities follow at least three major axes. A three-dimensional structure is consistent with previous work (Kozak et al., 2006; Weisman et al., 2017), but the specific dimensions we identified, and successfully replicated over five different samples, offer new insights into their psychological meaning and interrelationships.

First, each dimension shows multiple facets that previously have been overlooked. Dimension *A* unites aspects of physiological and positive as well as negative emotional capacities that are largely unintentional. *M* encompasses aspects of both moral cognition and social cognition, which itself includes the simulation of one's own mind (e.g., planning) and others' minds (e.g., inferring their thoughts); its appropriate label may thus be *Moral & Social Cognition.* These processes are largely under the agent's intentional control and enable understanding and regulation of one's own and others' behavior, thus carving out a specific meaning of agency. *R* illustrates the dynamic transition from perception and cognition through learning to communication and action—a second more specific meaning of agency.

It is noteworthy that none of the dimensions are made up simply of bundles of semantically related words. The use of PCA in personality psychology has sometimes been criticized as merely recovering dictionary relations between trait adjectives (e.g., Extraversion = outgoing, sociable, gregarious, friendly, etc; *cf*. D'Andrade, 2017). The items that are clustering together in the *A-M-R* structure are only midly semantically related, but more so they point to fundamental psychological functions of the mind—to engage with its own processes, with other minds, and with the social and physical environment.

We also found that, under some conditions, two of the three dimensions bifurcate: *A* divides into negative and positive-social affect; *M* divides into moral cognition and social cognition. We have so far identified only instance in which a full five-dimensional structure emerges: when people consider the desired capacities of a robot. Other instances may emerge as a function of one's attitude toward the agent (e.g., friend or foe), or the functional role of the capacity ascriptions (e.g., for interaction vs. evaluation).

Finally, we have offered a short, reliable measure of the three- to five-dimensional structure of mind perception, thus opening the door to many new investigations. These include developmental and cross-cultural studies of mind perception, as well as studies into how mind perceptions change over time—such as when interacting with a robot. The scale also invites a more refined assessment of anthropomorphism, sometimes cast as a relatively indiscriminate human tendency that may, in reality, be more selective. Finally, questions of dehumanization can be posed anew, as denying "mind" is unlikely to occur in a simple on/off way (Rai, Valdesolo, & Graham, 2017); rather, its impact on social and moral behavior may be differentiated depending on what aspect of mind—out of three to five—is denied.

References

Buja, A., & Eyuboglu, N. (1992). Remarks on parallel analysis. *Multivariate Behavioral Research*, *27*(4), 509–540. https://doi.org/10.1207/s15327906mbr2704_2

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*(2), 245–276. https://doi.org/10.1207/s15327906mbr0102_10

D'Agostino, R. B., & Russell, H. K. (2014). Simple Structure. In *Wiley StatsRef: Statistics Reference Online*. https://doi.org/10.1002/9781118445112.stat05607

D'Andrade, R. G. (1987). A folk model of the mind. In D. Holland & N. Quinn (Eds.), *Cultural models in language and thought* (pp. 112–148). New York, NY: Cambridge University Press.

D'Andrade, R. G. (2017). Memory and the assessment of behavior. In H. M. Blalock (Ed.), *Measurement in the Social Sciences*. https://doi.org/10.4324/9781351329088-6

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science*, *315*(5812), 619–619. https://doi.org/10.1126/science.1134475

Hamlin, J. K. (2013). Moral judgment and action in preverbal infants and toddlers: Evidence for an innate moral core. *Current Directions in Psychological Science*, *22*(3), 186–193. https://doi.org/10.1177/0963721412470687

Haslam, N., Kashima, Y., Loughnan, S., Shi, J., & Suitner, C. (2008). Subhuman, inhuman, and superhuman: Contrasting humans with nonhumans in three cultures. *Social Cognition*, *26*(2), 248–258.

Johnson, S. C. (2000). The recognition of mentalistic agents in infancy. *Trends in Cognitive Sciences*, *4*(1), 22–28. https://doi.org/10.1016/S1364-6613(99)01414-X

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, *20*(1), 141–151. https://doi.org/10.1177/001316446002000116

Kozak, M. N., Marsh, A. A., & Wegner, D. M. (2006). What do I think you're doing? Action identification and mind attribution. *Journal of Personality and Social Psychology*, *90*(4), 543–555. https://doi.org/10.1037/0022-3514.90.4.543

Malle, B. F. (2005). Folk theory of mind: Conceptual foundations of human social cognition. In R. R. Hassin, J. S. Uleman, & J. A. Bargh (Eds.), *The new unconscious* (pp. 225–255). New York, NY: Oxford University Press.

Malle, B. F., & Thapa Magar, S. (2017). What kind of mind do I want in my robot? Developing a measure of desired mental capacities in social robots. *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, 195–196. https://doi.org/10.1145/3029798.3038378

Premack, D. (1990). The infant's theory of self-propelled objects. *Cognition*, *36*(1), 1–16. https://doi.org/10.1016/0010-0277(90)90051-K

Rai, T. S., Valdesolo, P., & Graham, J. (2017). Dehumanization increases instrumental violence, but not moral violence. *Proceedings of the National Academy of Sciences*, *114*(32), 8511–8516. https://doi.org/10.1073/pnas.1705238114

Sytsma, J., & Machery, E. (2010). Two conceptions of subjective experience. *Philosophical Studies*, *151*(2), 299–327. https://doi.org/10.1007/s11098-009-9439-x

Takahashi, H., Ban, M., & Asada, M. (2016). Semantic differential scale method can reveal multi-dimensional aspects of mind perception. *Frontiers in Psychology*, *7*, 1717. https://doi.org/10.3389/fpsyg.2016.01717

Tomasello, M. (2003). The key is social cognition. In D. Gentner & S. Goldin-Meadow (Eds.), *Language in mind: Advances in the study of language and thought* (pp. 47–57). Cambridge, MA: MIT Press.

Weisman, K., Dweck, C. S., & Markman, E. M. (2017). Rethinking people's conceptions of mental life. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(43), 11374–11379. https://doi.org/10.1073/pnas.1704347114

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin*, *99*(3), 432–442. https://doi.org/10.1037/0033-2909.99.3.432

Effects of Blindfolding on Verbal and Gestural Expression of Path in Auditory Motion Events

Ezgi Mamus (ezgi.mamus@mpi.nl)
Center for Language Studies, Radboud University, Nijmegen, The Netherlands
Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

Lilia Rissman (l.rissman@let.ru.nl)
Center for Language Studies, Radboud University, Nijmegen, The Netherlands
Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

Asifa Majid (asifa.majid@york.ac.uk)
Department of Psychology, University of York, York, UK

Aslı Özyürek (asli.ozyurek@mpi.nl)
Center for Language Studies & Donders Center for Cognition, Radboud University, Nijmegen, The Netherlands
Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

Abstract

Studies have claimed that blind people's spatial representations are different from sighted people, and blind people display superior auditory processing. Due to the nature of auditory and haptic information, it has been proposed that blind people have spatial representations that are more sequential than sighted people. Even the temporary loss of sight—such as through blindfolding—can affect spatial representations, but not much research has been done on this topic. We compared blindfolded and sighted people's linguistic spatial expressions and non-linguistic localization accuracy to test how blindfolding affects the representation of path in auditory motion events. We found that blindfolded people were as good as sighted people when localizing simple sounds, but they outperformed sighted people when localizing auditory motion events. Blindfolded people's path related speech also included more sequential, and less holistic elements. Our results indicate that even temporary loss of sight influences spatial representations of auditory motion events.

Keywords: blindfolding; localization; pointing; auditory motion events; spatial language

Introduction

Information provided by visual, auditory, and haptic systems work together to enhance detection, localization, and identification of objects and events in the world. Compared to auditory and haptic input, vision has the advantage of providing simultaneous, precise, and detailed information about features of objects and events that take place in close and distant space (e.g., Eimer, 2004; Thinus-Blanc & Gaunet, 1997).

Considering the qualitative differences between inputs from sensory modalities, it is interesting to ask how blindness influences conceptualization of space, and how this is reflected in the spatial language of blind individuals. Numerous studies have reported enhanced auditory spatial skills in blindness (e.g., Lessard, Paré, Lepore & Lassonde, 1998; Röder et al., 1999; Voss et al., 2004), and the spatial language of blind individuals has been shown to be conceptually different when it is based on haptic input (Iverson, 1999; Iverson & Goldin-Meadow, 1997). The present study is the first to focus on how information acquired from the auditory modality alone affects spatial event conceptualization as expressed in both language and pointing gestures in blindfolded and sighted people.

It is claimed that blind individuals can compensate for their lack of vision through better auditory processing. Consistent with this, some studies suggest the blind even outperform their blindfolded counterparts in low-level auditory spatial tasks, such as estimating distance based on echo cues and localizing direction of a sound in the horizontal plane (e.g., Després, Candas & Dufour, 2005; Dufour, Després & Candas, 2005; Lessard et al., 1998; Röder et al., 1999; Voss et al., 2004). It is possible that blindfolding creates a temporary disadvantage for sighted individuals' spatial mapping of sounds. Only a single study compared sound localization skills of blindfolded and sighted individuals (Tabry, Zatorre, & Voss, 2013). Tabry et al. presented simple sounds on the horizontal and vertical planes and measured accuracy of pointing by hand or head laser pointer. Tabry et al. found that the absence of visual feedback decreases localization accuracy mostly for head-pointing and sounds on the vertical plane.

Other studies measuring navigation and spatial updating skills have claimed that blind individuals have impaired performance when required to process multiple pieces of information or simultaneous information, such as creating representations of large-scale environments, or inferring new spatial relations that are not directly experienced (finding the shortest way from A to B, when only experiencing A to C and B to C) (e.g., Coluccia, Mammarella & Cornoldi, 2009; Pasqualotto & Newell, 2007, Rieser, Guth & Hill, 1982; Thinus-Blanc & Gaunet, 1997). This may be because blind individuals have to rely on sensory information that is perceptually represented sequentially, thereby making it

more difficult to build holistic spatial representations of path information.

Language studies investigating speech and gesture in route description tasks have also found evidence that blind peoples' conceptualization of space has an underlying sequential representation of path for large-scale layouts; but that they can build holistic representations for small-scale layouts (Iverson, 1999; Iverson & Goldin-Meadow, 1997). Iverson and Goldin-Meadow (1997) examined sighted, blindfolded, and blind children's speech and co-speech gesture production in a task where participants had to give directions for familiar locations in their school. The results showed that blind children's speech was more segmented, with several landmark points on the path described, whereas sighted and blindfolded children linguistically represented the area in a global manner. Iverson and Goldin-Meadow did not report any difference between sighted and blindfolded children's speech but this is not surprising given the fact that blindfolded children also initially saw the scene before the description task and so, their initial encoding of the school space was based on visual input.

As a follow-up Iverson (1999) examined sighted, blindfolded, and blind children's route descriptions for small-scale scenes constructed from Lego blocks. Even though both blind and blindfolded children explored the Lego scenes haptically, while sighted children explored the Lego scenes visually, all children gave similar path expressions (in terms of landmark use). Iverson claimed that the Lego scenes could be encoded similarly by touching and seeing because the amount of available spatial information was equivalent for both modalities, which allowed blind children to build more holistic representations for small-scale scenes.

The Present Study

We investigated the effect of blindfolding on localization and verbal descriptions of auditory motion events. Having both linguistic and non-linguistic tasks performed by the same participants helps us understand whether possible differences between groups come from the processes required for linguistic packaging, or are grounded in more fundamental spatial representations, independent of the demands of speech production.

As shown by Tabry et al. (2013), blindfolding can influence sighted people's spatial mapping of sounds. To investigate this possibility further, we measured localization ability in two non-linguistic tasks for simple beep sounds and also for the first time in more complex auditory motion events. In both tasks, participants were asked to trace the path of the movement as accurately as they could by tracing a line with their finger or hand. Tabry et al. (2013) used simple sounds similar to our beep sounds, and only one condition in their study—hand pointing on the horizontal plane—was relevant to the task in the current study. In this condition, Tabry et al. did not report a difference between the blindfolded and the sighted group in the degrees of deviation from target location. Based on Tabry et al.'s findings, we expected no difference between blindfolded and sighted participants in the localization task with beep sounds. We also examined whether these findings for simple beep sounds generalize to localization of complex auditory events. It may be the case that as the stimulus becomes more complex, there is more opportunity to see differences between sighted and blindfolded individuals.

In speech we aimed to explore path representations by measuring different manners of encoding. As we know from the blindness literature (e.g., Iverson & Goldin-Meadow, 1997; Thinus-Blanc & Gaunet, 1997), sequential representations typically encode consecutive landmarks in relation to path, but spatial relations between distant objects are not encoded explicitly. To address the distinction between sequential and holistic path representations, we coded whether speech included information about source, goal, orientation, and path verbs. Source and goal elements in speech represent sequential information because those encode discrete units of information—such as which landmark is a starting point of movement—without explicitly encoding its spatial relation to other elements. We take orientation and path verbs in speech to represent spatial relations because these encode information about direction (e.g., from left to right) and trajectory of movement (e.g., approaching). Thus, it can be argued that mentions of orientation and path verb show more holistic representation of the space. We conducted the current study in Turkish as source and goal elements are optional when describing a motion event. Therefore, Turkish enables us to compare differences in the event descriptions.

If having visual cues at encoding—such as seeing the source of a sound—enables people to build a more holistic representations of space, even temporary absence of sight may affect spatial representations and make them more akin to the representations created by the blind, i.e., make them more sequential. As such, it may be expected that, compared to sighted people, blindfolded people's event descriptions would include more sequential path information, such as more mentions of the source, but less holistic path information that encodes trajectory of motion and the relation between two different locations—such as figure and source.

Method

Participants

Twelve sighted (M = 22.27 years, SD = 2.10, 7 female) and 12 blindfolded (M = 21.83 years, SD = 2.21, 7 female) Turkish adult speakers participated in the experiment in exchange for extra credit in an introductory psychology course. The sample size was based on previous studies comparing sighted and blindfolded participants (Iverson 1999; Iverson & Goldin-Meadow, 1997; Tabry et al., 2013). Participants all had normal or corrected-to-normal vision and provided written informed consent.

Auditory Stimuli

We filmed and simultaneously recorded the sound of locomotion and non-locomotion events. Locomotion events served as the critical experimental items in the study, whereas non-locomotion events served as filler items. For the locomotion events, an actress moved in distinct manners (walk, run, and limp) with respect to a landmark object (door or elevator) along a specific path (to, from, into, and out of). Each manner was combined with each path, creating 12 different items. The sound recorder was placed next to the landmark objects, so the path direction in the events was either approaching (for *to* and *into* paths) or away from (for *from* and *out of* paths) listeners. In addition, the path azimuth was edited using Soundtrack Pro audio editing software to increase the variety of possible path motion. Five movement angles were created in a semicircular space ranging from 90° left to 90° right with 45° intervals, thus from the right to the left these are: 0° (right), 45° (right-sided), 90° (front), 135° (left-sided), and 180° (left) motions (see Figure 1). We created all 12 events with the 5 movement angles, resulting in 60 events in total. All locomotion events were exported as 5.1 surround sound.

For the non-locomotion events, the same actress performed different actions with objects (e.g., drinking water, eating chips), and the video and sound were recorded across from her. We did not examine these items further. There were 77 experimental trials in total, including 60 locomotion events and 17 non-locomotion events. Locomotion events lasted 9s (SD: 1.9) and non-locomotion events 8s (SD: 2.2) on average.

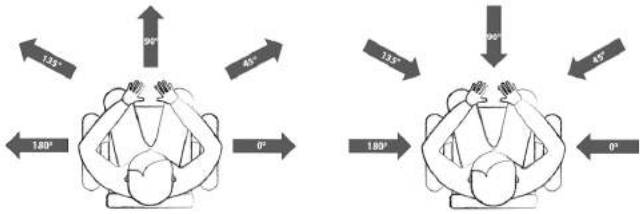


Figure 1: Path direction and angles for "from" and "out of" events (left) and "to" and "into" events (right).

In addition to the locomotion and non-locomotion events, we prepared 60 audio-clips consisting of beeps. These sounds were intended to assess people's accuracy in localizing simple dynamic stimuli, in contrast to the more complex, naturalistic locomotion events. To make a beep clip, a 1s beep sound was compounded with a 1s silence lasting 9s in total. The direction of sound movement in each clip was manipulated as described for the locomotion events (see Figure 1).

Procedure

Each participant was tested in a quiet room on Bogazici University campus in Istanbul, Turkey. The procedure of the experiment was the same for both groups, except that blindfolded participants' eyes were covered before they entered the room, and the experimenter helped them to be seated. In the room, five speakers were placed 1.34 m far from the participant's head and approximately 95 cm high from the ground in a 5+1 surround system configuration. Front left and right speakers were placed 30° off center, and rear left and right speakers were 110° off center. Participants sat in the middle of the speakers. The experimenter stayed in the room during the experiment to initiate the tasks and advance the trials on a laptop using Presentation Software.

There were two sorts of tasks:

(1) Event Description Task Participants listened to audio-clips of the events. Before the experiment started, there were 2 practice trials consisting of one locomotion and one non-locomotion event. In each trial, an event was presented aurally and participants were asked to describe what happened. They were told that another participant would watch their descriptions and listen to the same sounds to try and match the sound clips.

(2) Localization Task with Events vs. Beeps Participants listened to the audio-clips of 60 locomotion events and 60 audio-clips consisting of beep sounds in two separate tasks for each stimulus type. There were 4 practice trials in each task. After each audio-clip, they were asked to trace the path of the movement in the semicircular frontal space as accurately as they could by tracing a line with their finger or hand. They were instructed not to describe the audio stimuli, but only trace the paths.

Participants first performed the event description task. During this task, participants' speech was recorded with two video cameras. One camera was placed across from the participant and the other recorded the top view of the participants' frontal space so as to capture arm and hand movements. Following the event description task, participants performed either the localization task with audio events or the localization task with beeps. The order of these two tasks was counterbalanced across participants. Finally, participants were asked to fill out a demographic questionnaire on a laptop. The total duration of the experiment was around 75 minutes.

Coding

Descriptions for the motion events were transcribed and coded by a native Turkish speaker. First, the event descriptions were split into clauses. Clauses were coded as relevant or irrelevant to the target events. Second, each relevant clause for each event was coded, according to the type of information it contained: (1) the use of sequential elements—(a) source (starting point of movement), and (b) goal (the end point of the movement); and (2) holistic elements—(a) orientation (direction), and (2) path verb (trajectory of motion). An example description below encodes information about the source, the orientation, and the path verb of the movement as:

(1)

Asansör-den	***sağ-a***	***doğru***	***uzak-laş-(ı)yor.***
elevator-ABL	right-DAT	towards	away-VERB-PRS.3SG
(source)	(orientation)		(path verb)

'(someone) moves away from the elevator towards the right.'
(VERB = verbal suffix)

For the localization tasks, direction and angle localization were coded by an assistant. There were 2 possible directions (approaching or going away) and 5 possible angles (from 90° left to 90° right with 45° intervals). Twenty percent of the coding was checked by the first author of the study. Interrater agreement was at least 0.80 (95% CI: 0.69, 0.91) using Kappa for both tasks.

Results

For all analyses reported in the paper, we used mixed effects regression models. All models were generated using the lme4 package (Bates, Mächler, Bolker, & Walker, 2015) in R (R Core Team, 2018). We begin by presenting the data for the simplest task—the localization task with beeps—before moving to the data of the localization task with events and the event description task.

Localization Task with Events vs. Beeps

First we investigated whether sighted and blindfolded participants differed in how they localized motion using simple beep sounds. We ran two separate glmer models to test the effects of blindfolding on binary values (correct, incorrect) for: (1) angle and (2) direction accuracy. Since localization of direction and angle was simultaneously performed by participants, we also included the accuracy of the other variable as a predictor in the models. That is, the model for direction accuracy included angle accuracy as a predictor in addition to the group factor (sighted or blindfolded). The optimal random effects structure included random intercepts of participant and item. Model 1 for angle accuracy showed that blindfolded participants did not differ in localizing the angle of beep sounds from sighted participants, and that participants became significantly more successful as direction accuracy increased (see Table 1 and Figure 2). Similarly, Model 2 for direction accuracy showed that blindfolded participants did not differ in localizing direction of beep sounds from sighted participants, and that participants became significantly more successful as angle accuracy increased (see Table 1 and Figure 2). These results showed that blindfolding did not affect localization ability when the sounds were simple, dynamic beeps, and all participants succeeded in localizing the direction of beep sounds—in fact, they were at ceiling levels.

Table 1: Accuracy models for angle and direction localization of the beep sounds.

	Estimate	Std.Error	z-value	*p*-value
Model 1 for Angle				
(Intercept)	-0.5671	0.5265	-1.077	0.2814
Group	-0.0545	0.2937	-0.186	0.8527
Dir. Acc.	1.6281	0.4317	3.771	<0.001***
Model 2 for Direction				
(Intercept)	4.2448	0.6493	6.537	<0.001***
Group	-0.8279	0.6735	-1.229	0.2190
Ang. Acc.	1.4246	0.4509	3.160	0.0016**

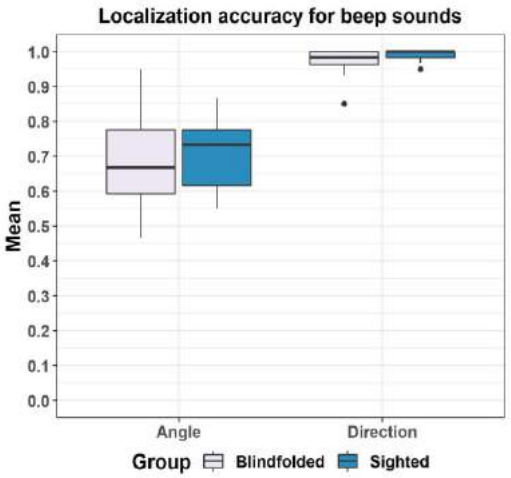


Figure 2: Localization accuracy for beep sounds.

For the locomotion events, we again ran two separate glmer models to test the effects of blindfolding on binary values for (1) angle and (2) direction accuracy. Model 3 for angle accuracy showed that blindfolded participants performed better in localizing angle of locomotion events than sighted participants, and that participants became significantly more successful as direction accuracy increased (see Table 2 and Figure 3). Similarly, Model 4 for direction accuracy showed that blindfolded participants performed better in localizing direction of locomotion events than sighted participants, and that participants became significantly more successful as angle accuracy increased (see Table 2 and Figure 3). As with the beep sounds, all participants were almost at ceiling for identifying the direction of motion. Unlike for beeps, blindfolded participants were better able to identify the angle and direction of auditory events when sounds were meaningful, locomotion events.

Table 2: Accuracy models for angle and direction localization of the locomotion events.

	Estimate	Std.Error	z-value	*p*-value
Model 3 for Angle				
(Intercept)	-0.1714	0.3603	-0.476	0.6344
Group	0.5814	0.3047	1.908	0.0564 .
Dir. Acc.	0.5998	0.3030	1.979	0.0478*
Model 4 for Direction				
(Intercept)	3.4917	0.4890	7.140	<0.001***
Group	1.5285	0.5390	2.836	0.0046**
Ang. Acc.	0.7153	0.3261	2.194	0.0283*

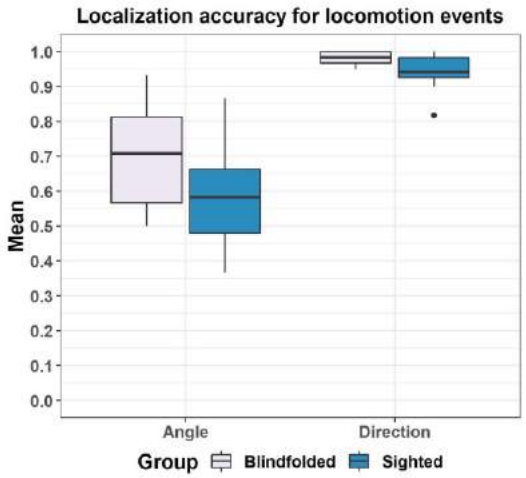


Figure 3: Localization accuracy for locomotion events.

Event Description Task

Finally, to investigate whether sighted and blindfolded participants differed in how they described the path of events, we calculated the ratio of sequential (source and goal) and holistic path descriptions (orientation and path verb) per relevant clause. To do this, total counts of sequential and holistic path descriptions were divided by the number of relevant clauses for each trial. So, we had a 2-level variable for the type of linguistic expression (sequential vs. holistic) and a 2-level variable for the group (blindfolded vs. sighted) as predictors.

We ran an lmer model to test the effects of blindfolding and type of linguistic expression using ratio of mention per clause as input. The optimal random effects structure included random intercepts of participant and event. The results showed that there was a significant effect of type of linguistic expression, with all participants mentioning more holistic than sequential descriptions ($p < .001$). This difference was not surprising because of the fact that one of the holistic elements included verbs. Due to its typology, Turkish usually expresses path of motion in the verb (Talmy, 1985). There was no effect of blindfolding in how often participants mentioned all path elements in their descriptions ($p = .272$). Crucially, the interaction between group and type of linguistic expression was significant ($p <.001$; see Table 3 for model summary and Figure 4). Blindfolded participants gave more sequential but less holistic descriptions in their speech compared to sighted participants.

Table 3: Models for ratio of sequential and holistic path descriptions in the events.

	Estimate	Std.Error	t-value	*p*-value
(Intercept)	0.3187	0.0854	3.730	<0.001***
Exp.Type	0.7387	0.0333	22.208	<0.001***
Group	0.1319	0.1175	1.123	0.272
E.Type:Gr	-0.2038	0.0471	-4.307	<0.001***

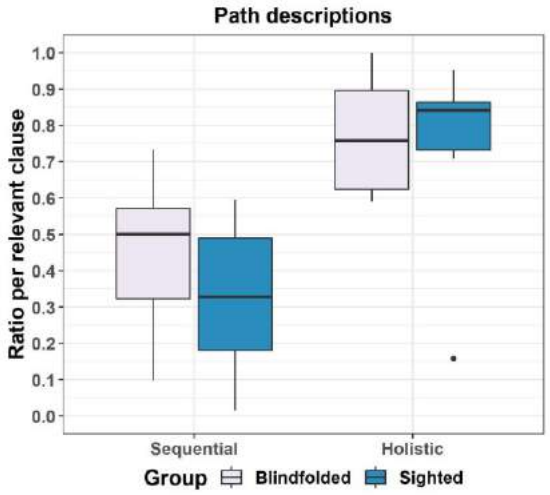


Figure 4: Ratio of sequential and holistic path descriptions per relevant clauses in the events.

Discussion

In the present study, we examined the effect of blindfolding on localization and verbal descriptions of auditory motion events. In the localization task with beeps, we showed that blindfolded participants performed as well as sighted participants when localizing simple sounds. Our results are in line with Tabry et al. (2013). Similar to our localization task with beeps, Tabry et al. tested hand-pointing accuracy for simple sounds on the horizontal plane, and reported no effect of blindfolding in the deviation from target. This does not necessarily imply there are never differences in localization in response to blindfolding. Tabry et al. (2013) did find differences in other paradigms, such as head-pointing and localizing simple sounds on the vertical plane. Based on the results of our localization task, and Tabry et al.'s similar paradigm, we can conclude that blindfolded and sighted people behave similarly in the spatial mapping of simple sounds when orienting their hands toward a specific location on the horizontal plane.

In contrast to the simple auditory tones, blindfolded participants outperformed sighted participants when localizing more complex auditory locomotion events. Earlier studies investigating sound localization abilities in blindness have only ever used simple sounds as stimuli. Our result

suggests that having no visual feedback creates an advantage in localization when mapping complex sounds onto an event space. One possible explanation for this advantage could be that closing the eyes increases auditory attention and thereby leads to better performance when localizing complex sounds. Since participants are already near ceiling for simple sounds, there is no room to see this improvement in that condition. A recent study by Wöstmann, Schmitt, and Obleser (2019) found that while attending to one of two spoken streams, even in a darkened room, closing eyes modulated attention, and increased alpha power for the attended stream. Wöstmann et al. suggested that closing eyes might decrease the dominance of vision, and thus enhance attention to nonvisual input. Although they did not report behavioral enhancement with closed eyes, their participants performed the tasks in a darkened room where there was no distracting visual input. In our study, to the contrary, sighted participants could see the location of the audio-speakers, which could possibly distract them while listening to sounds and/or localizing them in space. Thus, it is possible that our paradigm is more suitable to detect a possible beneficial behavioral effect of closing eyes. Furthermore, one could hypothesize that blind people might perform even better due to their better ability to process auditory information than both blindfolded and sighted people. Future studies could examine this possibility.

We did not find an effect of blindfolding on how often participants mentioned path in their descriptions regardless of the type of linguistic expression. However, we did find that blindfolded participants gave more sequential, and less holistic descriptions for the path of auditory motion events, compared to sighted participants. This is in line with the claim that blindness leads to sequential representations and segmented speech due to the more sequential nature of the sensory information that the resulting spatial representations depend on (e.g., Iverson & Goldin-Meadow, 1997). Iverson (1999) and Iverson and Goldin-Meadow (1997) showed that landmarks on a described route were used to segment the path into several pieces. We also found that blindfolded participants in our data used more landmark information encoded as source and goal in their descriptions. Thus, our results suggest that even temporary loss of sight changes how people talk about events by possibly hindering the building of a holistic representation of space.

Conclusion

We are the first to investigate the effect of the temporary loss of sight on localization and verbal descriptions of auditory motion events. We showed that temporary loss of sight leads to more sequential and less holistic path descriptions, and better localization of auditory events as measured by pointing. These effects suggest that even the temporary loss of sight might change the sort of spatial representations people build in response to complex auditory events.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48.

Coluccia, E., Mammarella, I. C., & Cornoldi, C. (2009). Centred Egocentric, Decentred Egocentric, and Allocentric Spatial Representations in the Peripersonal Space of Congenital Total Blindness. *Perception, 38*(5), 679–693.

Eimer, M. (2004). Multisensory integration: how visual experience shapes spatial perception. *Current biology, 14*(3), R115-R117.

Després, O., Candas, V., & Dufour, A. (2005). The extent of visual deficit and auditory spatial compensation: evidence from self-positioning from auditory cues. *Cognitive brain research, 23*(2-3), 444-447.

Dufour, A., Després, O., & Candas, V. (2005). Enhanced sensitivity to echo cues in blind subjects. *Experimental Brain Research, 165*, 515–519.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159-174.

Lessard, N., Paré, M., Lepore, F., & Lassonde, M. (1998). Early-blind human subjects localize sound sources better than sighted subjects. *Nature, 395*, 278–280.

Iverson, J. M. (1999). How to get to the cafeteria: Gesture and speech in blind and sighted children's spatial descriptions. *Developmental psychology, 35*(4), 1132.

Iverson, J. M., & Goldin-Meadow, S. (1997). What's communication got to do with it? Gesture in children blind from birth. *Developmental psychology, 33*(3), 453.

Pasqualotto, A., & Newell, F. N. (2007). The role of visual experience on the representation and updating of novel haptic scenes. *Brain and cognition, 65*(2), 184-194.

R Core Team. (2018). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundations for Statistical Computing. Available online at: https://www.R-project.org/.

Rieser, J. J., Guth, D. A., & Hill, E. W. (1982). Mental processes mediating independent travel: Implications for orientation and mobility. *Journal of Visual Impairment and Blindness, 76*, 213-218.

Röder, B., Teder-Sälejärvi, W., Sterr, A., Rösler, F., Hillyard, S. A., & Neville, H. J. (1999). Improved auditory spatial tuning in blind humans. *Nature, 400*, 162–166.

Tabry, V., Zatorre, R. J., & Voss, P. (2013). The influence of vision on sound localization abilities in both the horizontal and vertical planes. *Frontiers in psychology, 4*, 932.

Talmy, L. (1985). Lexicalization patterns: Semantic structure in lexical forms. In T. Shopen (Ed.), *Language typology and semantic description* (pp. 36–149). Cambridge: Cambridge University Press.

Thinus-Blanc, C., & Gaunet, F. (1997). Representation of space in blind persons: Vision as a spatial sense? *Psychological Bulletin, 121*, 20-42.

Voss, P., Lassonde, M., Gougoux, F., Fortin, M., Guillemot, J. P., & Lepore, F. (2004). Early-and late-onset blind

individuals show supra-normal auditory abilities in far-space. *Current Biology, 14*(19), 1734-1738.

Wöstmann, M., Schmitt, L. M., & Obleser, J. (2019). Does closing the eyes enhance auditory attention? Eye closure increases attentional alpha-power modulation but not listening performance. *Journal of cognitive neuroscience, 1-14*.

Insulating Distributional Semantic Models from Catastrophic Interference

Willa M. Mannering and Michael N. Jones
Indiana University, Bloomington
[wmanneri] [jonesmn]@indiana.edu

Abstract

Predictive neural networks, such as word2vec, have seen impressive recent popularity as an architecture to learn distributional semantics in the fields of machine learning and cognitive science. They are particularly popular because they learn continuously, making them more space efficient and cognitively plausible than classic models of semantic memory. However, a major weakness of this architecture is *catastrophic interference* (CI): The sudden and complete loss of previously learned associations when encoding new ones. CI is an issue with backpropagation; when learning sequential data, the error signal dramatically modifies the connection weights between nodes—causing rapid forgetting of previously learned information. CI is a huge problem for predictive semantic models of word meaning, because multiple word senses interfere with each other. Here, we evaluate a recently proposed solution to CI from neuroscience, elastic weight consolidation, as well as a Hebbian learning architecture from the memory literature that does not produce an error signal. Both solutions are evaluated on an artificial and natural language task in their ability to insulate a previously learned sense of a word when learning a new one.

Keywords: distributional semantic models; catastrophic interference; word2vec; random vector accumulation; elastic weight consolidation

Introduction

Distributional models of semantic memory (DSMs; e.g., Landauer & Dumais, 1997) attempt to explain how humans learn the meaning of words through statistical inference. All DSMs are based on the distributional hypothesis of language (Harris, 1970), often summarized as learning a word's meaning "by the company it keeps" (Firth, 1957). Classic DSMs use counts of co-occurrence between words in a corpus to construct semantic representations. Recently, with the development of predictive DSMs and improvements in overall computing power, the fields of cognitive science and machine learning have seen an increase in popularity of error-driven DSMs within connectionist architectures. Predictive DSMs use the backpropagation of an error signal through the network to predict context and are particularly popular because they learn continuously—making them more space efficient and more cognitively plausible than earlier DSMs.

However, a major weakness of predictive DSMs is *catastrophic interference* (CI): The sudden and complete loss of previously learned associations when encoding new ones (French, 1999). When a predictive neural network is exposed to sequential data, the introduction of completely new information causes the error signal to be very large, effectively "shocking" the model and causing it to overcorrect the weights to accommodate the new information. The problem of CI is a major issue not only for functional reasons but for implications of cognitive plausibility as well.

The standard predictive network currently discussed in the literature is Mikolov et al.'s (2013) word2vec model. Word2vec is a feedforward neural network with input and output layers that contain one node per word in the vocabulary, and a hidden layer of approximately 300 nodes. The word2vec architecture has two possible model directions. The context may be used to predict the word—which is referred to as the Continuous Bag of Words (CBOW) model—or, the word may be used to predict the context—which is referred to as the skipgram model. We will use skipgram in this paper because it maps conceptually onto most connectionist models and has been shown to perform better with smaller training corpora than the CBOW model.

Dachapally and Jones (2018) recently investigated the impact of CI on the internal representations produced by predictive DSMs when applied to sequentially learned word senses. Because of its current popularity, they used Mikolov et al.'s (2013) word2vec model to evaluate the effects of CI on the model's final semantic representations. In their study, Dachapally and Jones used homonyms to measure the effects of CI. Take for example a homonym like *bank,* with its two distinct meanings: river-bank and financial-bank. The word *bank* should have its final representation positioned equidistant to its two meanings in semantic space. Because of CI, however, if the financial sense was learned first, followed by the river sense, the final representation of *bank* would be positioned proximal to river-bank words, and the financial sense would be forgotten. This study was the first evaluation of CI in a predictive semantic model. Now that we know CI affects semantic representations produced by predictive DSMs, we can begin to propose and evaluate possible solutions for CI.

The goal of the current paper is to expand on Dachapally and Jones' (2018) work by implementing and comparing two possible solutions to CI from the cognitive and neural sciences. The first candidate solution is *elastic weight consolidation* (Kirkpatrick et al., 2017) which has been impressively successful on machine learning tasks and can be considered a "vaccination" for predictive DSMs that would prevent the effects of CI. The second candidate solution is a different architecture, *random vector accumulation* (Jones, Willits, & Dennis, 2015), which can be considered naturally "immune" to the effects of CI by way of its learning mechanism.

The goal of elastic weight consolidation (EWC) is to allow a predictive neural network to learn two sequential tasks, Task A and then Task B, without incurring CI. To do this, Kirkpatrick et al. (2017) introduced a method to constrain the

parameters of a neural network after learning Task A so that the network can subsequently learn Task B without forgetting Task A. The new loss function they introduce is a quadratic penalty that differentially constrains parameters in the neural network depending on how important each parameter is to completing Task A. To determine which weights in the network are important for Task A they calculate the Fisher Information for each parameter—a mathematical method to measure the amount of information a variable carries about a parameter. The resulting loss function that gets minimized in elastic weight consolidation is:

$$\mathcal{L}(\theta) = \mathcal{L}_{\mathcal{B}}(\theta) + \sum_i \frac{\lambda}{2} F_i \left(\theta_i - \theta^*_{A,i}\right)^2 \qquad (1)$$

where $\mathcal{L}_{\mathcal{B}}(\theta)$ is the loss for Task B, λ controls how important Task B is compared to Task A, F is the Fisher Information calculated for each parameter, and θ represents the parameters in the network. Kirkpatrick et al. (2017) showed that EWC was able to insulate against CI when training a predictive neural network on the MNIST (LeCun et al., 1998) data set, a free data set of handwritten images. While EWC has been tested several times on categorization tasks, this paper will present the first implementation for use with distributional semantic models.

EWC has potential as a "vaccine" for predictive DSMs, that is, networks may be insulated from CI without having to implement new architectures. There is reason to suspect that EWC may have limited effectiveness when translated to the field of semantic modeling. EWC calculates the relevance of each model parameter to Task A based on the actual class of the training data. However, in the case of semantic modeling, we are not necessarily interested in the final predicted class of the training data but in the internal representations created by models as they learn. It is one goal of this paper to determine how EWC affects the internal representations of predictive DSMs.

The second candidate solution to CI that we evaluate is a different architecture: random vector accumulation (RVA; Jones et al., 2015). RVA is an alternate architecture that should theoretically be "immune" to CI by nature of the learning mechanism. RVA is the theoretical mechanism that is core to semantic models such as BEAGLE (Jones & Mewhort, 2007). Unlike predictive DSMs, which are affected by CI due to the error signal produced during learning, RVA models should be immune to CI because they utilize principles of associative learning and do not rely on an error signal to learn. These models learn via a simple Hebbian co-occurrence learning mechanism. The most basic RVAs first begin by initializing two random vectors from an arbitrary distribution and of arbitrary dimensionality for each word encountered in a corpus. One vector is unique to each word in the vocabulary, the environment vector, and the other is a summation of all context words, the memory vector. The update function for the memory vector of each word in the vocabulary is described in Equation 2:

$$m_i = e_{i-1} + e_{i+1} \qquad (2)$$

where m_i is the memory vector for an arbitrary word in a corpus, e_{i-1} is the unique environment vector for the context word before *i,* and e_{i+1} is the unique environment vector for the context vector after *i.* So, the memory vector for word *i* stores the context vectors for every other word that appears in context with word *i.*

Similar to Dachapally and Jones' (2018) study, this paper will use homonyms to measure the bias in semantic space created by CI. For each model, EWC and RVA, two conditions will be tested and compared to the performance of the original word2vec model in both an artificial and natural language. In the first condition, a target homonym will have two equally frequent senses with distinct meanings. Ideally, the target homonym should be equidistant from both of its two senses in semantic space. In the second condition, a target homonym will have two senses, one which is dominant (occurs more frequently) and one which is subordinate. In this case, the target homonym should be closer in semantic space to the dominant sense. Dachapally and Jones (2018) found that in both the artificial and natural language when word2vec was trained sequentially on equally balanced word senses, the target word was closer in semantic space to whichever sense had been trained most recently—forgetting the first sense of the word. The same effect was found when a target homonym had a dominant and subordinate sense; CI caused the target word to be more similar to the subordinate sense if the subordinate sense was trained most recently. Importantly, recency overpowered frequency, and the subordinate sense of the word became dominant if it was the most recently learned. To determine the effects of CI on a neural network equipped with EWC and on RVAs, a similar experimental structure will be used.

Experiment 1: Effects of CI on EWC and RVAs in an Artificial Language

Dachapally and Jones (2018) used a simple artificial language in which there is a single homonym, *bass,* that has two distinct meanings—*bass*[fish] and *bass*[guitar]. A corpus was created from this simple language by sampling word pairs from the following Markov grammar:

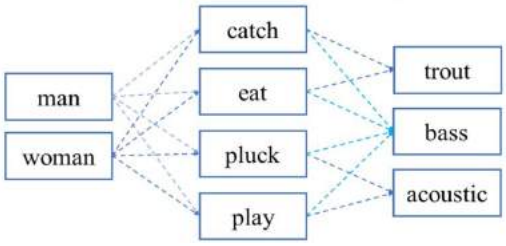


Figure 1. The artificial language used to test. Bass is the target homonym, and its position in semantic space relative to the two sense-pure synonyms (acoustic/bass) is evaluated.

In the first condition, a corpus of 8,000 sentences was generated from this grammar ("man catch bass", "woman play bass"). Each sense of the word *bass* was equally

frequent; the fish-sense made up half of the total sentences and the guitar-sense made up the other half. To measure the similarity of *bass* to its fish-sense and *bass* to its guitar-sense, the cosine similarity between the vector representation for *bass* and its two sense-pure synonyms *trout* and *acoustic* were calculated, respectively.

In the second condition, a corpus of 5,332 sentences was generated from the grammar—one sense of *bass* was dominant and the other subordinate. The dominant sense made up 4,000 of the total sentences and the subordinate sense made up 1,332. Thus, the subordinate sense was 1/3 as frequent as the dominant sense. Similar to the first condition, to determine the bias created in semantic space by CI, the similarity of *bass* to the dominant sense and *bass* to the subordinate sense was measured using the cosine similarity of the vector representations produced by each model.

The word2vec models used in this paper are both implemented using TensorFlow. Additionally, it is important to note that the implementation of the word2vec model in this experiment is different than both Mikolov et al.'s (2013) model and the model that was originally used in Dachapally & Jones' (2018) experiment. The full word2vec model as implemented by Mikolov et al. necessarily includes negative sampling and subsampling of the training data. Negative sampling is the practice of including negative information in the training data and subsampling is a method that results in less frequent words being sampled more often than frequent words. The model used by Dachapally & Jones used a different loss function called noise contrastive estimation which is common in the language modeling community because it is able to handle large input sizes. The model used in this experiment was purposely changed in order to be the most similar to the models previously used to implement EWC. This model uses cross entropy loss and does not use negative sampling or subsampling which may be responsible for the differences seen in the results of this paper.

Results

Figure 2 shows the cosine similarity of the vectors produced by word2vec, EWC, and the RVA in the case where sense 1 and sense 2 of *bass* are equally frequent. The pattern produced by word2vec is consistent with the findings in Dachapally and Jones' (2018) original experiment. When the model was trained in random order, the *bass-sense1* and *bass-sense2* similarities produced were approximately equal. When trained in sequential order, the sense which was sampled most recently ended up having a higher similarity to *bass*. The same procedure was repeated using EWC and the RVA. After exploring various parameter settings of both, we found that implementing EWC had virtually no effect on the results of the first experiment and that vector similarities produced by the RVA model were unaffected by CI.

Figure 3 shows the cosine similarity of the vectors produced by word2vec, EWC, and the RVA in the case where one sense is dominant and the other is subordinate. The pattern produced by word2vec is once again consistent with Dachapally and Jones' (2018) findings. When trained in random order, the dominant sense is more similar to *bass* than the subordinate sense. When trained in sequential order, the effects of CI reverse the frequency effects; when the subordinate sense of *bass* is trained last it becomes more similar to bass than the dominant sense. When the same procedure was performed using EWC and the RVA, we saw similar results to the first condition. The addition of EWC did not change the performance of word2vec and the RVA model was once again unaffected by CI.

EWC adds one additional parameter to the word2vec model, λ, which controls the importance of Task A compared

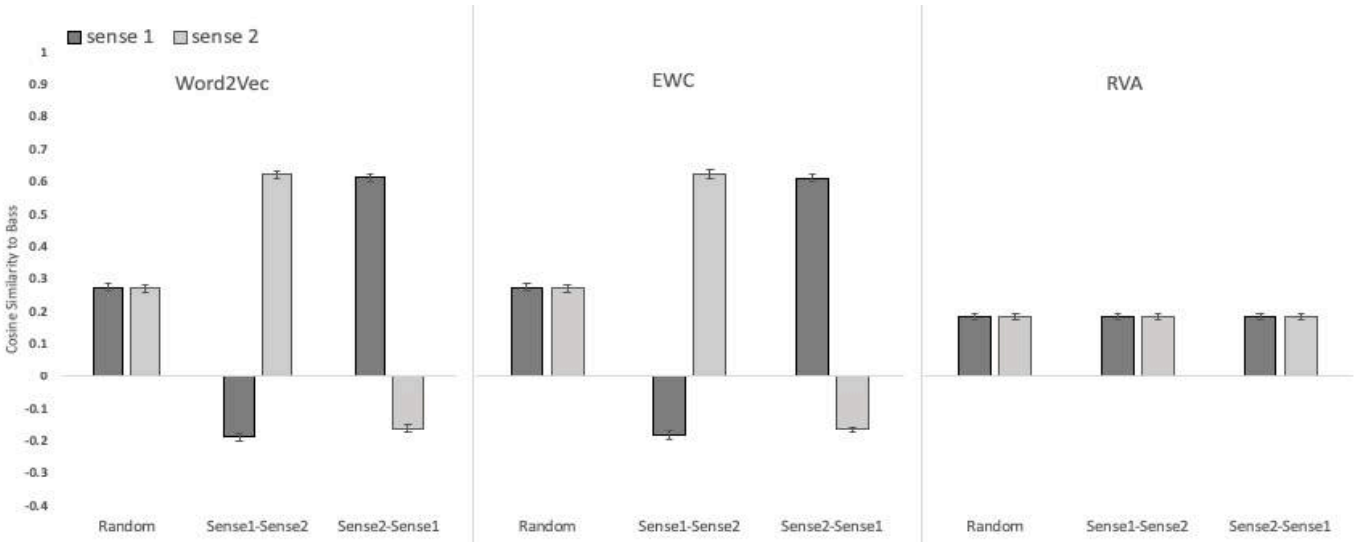


Figure 2. The y-axis represents the cosine similarity of vectors produced by word2vec, EWC, and the RVA. The x-axis represents one of three training orders: random order, sequential order with sense 1 first then sense2, and sequential order with sense 2 first then sense 1. Sense 1 and sense 2 are equally frequent in this case. CI is present in both word2vec and EWC while the RVA is unaffected by CI.

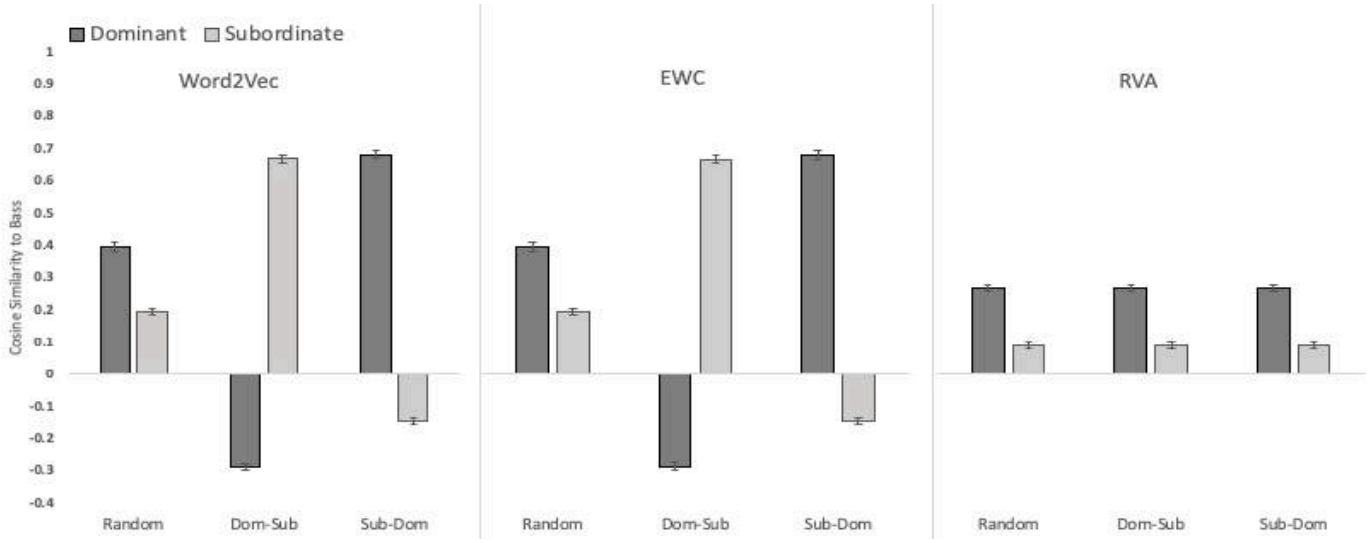


Figure 3. The y-axis represents the cosine similarity of vectors produced by word2vec, EWC, and the RVA. The x-axis represents one of three training orders: random order, sequential order with the dominant first then the subordinate sense, and sequential order with the subordinate sense first then the dominant sense. CI is present in both word2vec and EWC while the RVA is unaffected by CI.

to Task B. For both conditions in the first experiment, the value of λ had little to no effect on the final semantic representations. The results shown in both Figure 2 and Figure 3 are representative of the results obtained by any value of λ.

Experiment 2: Effects of CI on EWC and RVA in Natural Language

The texts used in this experiment are sourced from the TASA corpus (Landauer & Dumais, 1997). TASA contains language from textbooks with metadata tags which allowed us to train the models on distinct senses of a homonym without overlap. The same set of homonyms used in Dachapally and Jones (2018) was used for this experiment. They identified a sample of 14 homonyms that exist in the TASA corpus using the homonym norms from Armstrong, Tokowicz, and Plaut (2012) which determined homonyms with distinct meanings as rated by human participants.

The 14 homonyms were divided into two groups: sense-balanced and sense-imbalanced. We classified the two senses of a homonym as sense-imbalanced if one sense was at least twice as frequent in the TASA corpus, otherwise the two senses of a homonym were classified as sense-balanced. An example of a sense-imbalanced homonym is the word *slip*—the "fallen out of place" sense occurred across science contexts an equal number of times as the "shopping receipt" sense occurred across business contexts. An example of a sense-imbalanced homonym is the word *gum*—the "chewing candy" sense occurs approximately 5 times as often in language arts contexts than the "tissue surrounding teeth" sense occurs in health contexts. The sense-balanced homonyms are the counterpart to the first condition in the first experiment where the two senses of bass are equally frequent. The sense-imbalanced homonyms are the counterpart to the second condition in the first experiment where one sense of the word *bass* was dominant over the other. We then trained the word2vec model, the EWC model, and the RVA model on the entire corpus under three different order conditions. The first condition randomized the training order, the second condition was *sense1* first then *sense2* order, and the third condition was *sense2* first then *sense1* order. Cosine similarities between the target word vector and the two sense vectors were then calculated for each homonym set.

Results

The most common version of word2vec used for non-trivial training data is the model implemented within the Gensim Python library (Rehurek & Sojka, 2010). This model is optimized using C and is consequently very fast and effective. This is the model used by Dachapally & Jones in their second experiment to test for CI in natural language corpora. That model, however, is not directly compatible with the EWC implementation from our first experiment. For this reason, we did not use the Gensim model. Instead, we implemented a model in TensorFlow which is more similar to the model used in our first experiment and is compatible with EWC. However, there are some additional differences between the base models in our first experiment and the current experiment. In order to scale up to natural language, we had to include negative sampling and change the loss function to noise contrastive estimation. The model used in the previous experiment did not use negative sampling, but the model was unable to learn well from the natural language otherwise, so it was added. Additionally, while our first implementation of word2vec used a SoftMax layer to learn with a cross entropy loss function, the implementation in this experiment used noise contrastive estimation because the SoftMax method simply does not scale up well. The RVA model used in this experiment is the same model we used in the first experiment.

Figure 4 shows the results of training word2vec, EWC, and the RVA on the sense-balanced homonyms from TASA. The pattern of cosine similarities produced by word2vec and EWC are consistent with the results from the artificial language. When trained in random order the target words have approximately equal similarities to both of its senses. When trained sequentially, we see the same issue that occurred in the first experiment—the sense that was trained last becomes more similar to the target word. The RVA model shows the same pattern exhibited in Experiment 1—the similarity between the target and its two senses remain consistent no matter the training order.

Figure 5 shows the results of training word2vec, EWC, and the RVA on the sense-imbalanced homonyms from TASA. The cosine similarities produced from word2vec and EWC are consistent again with the results from the artificial language. Similarly, the cosine similarities produced by the RVA are consistent with the results from the artificial language and do not appear to be dependent on training order.

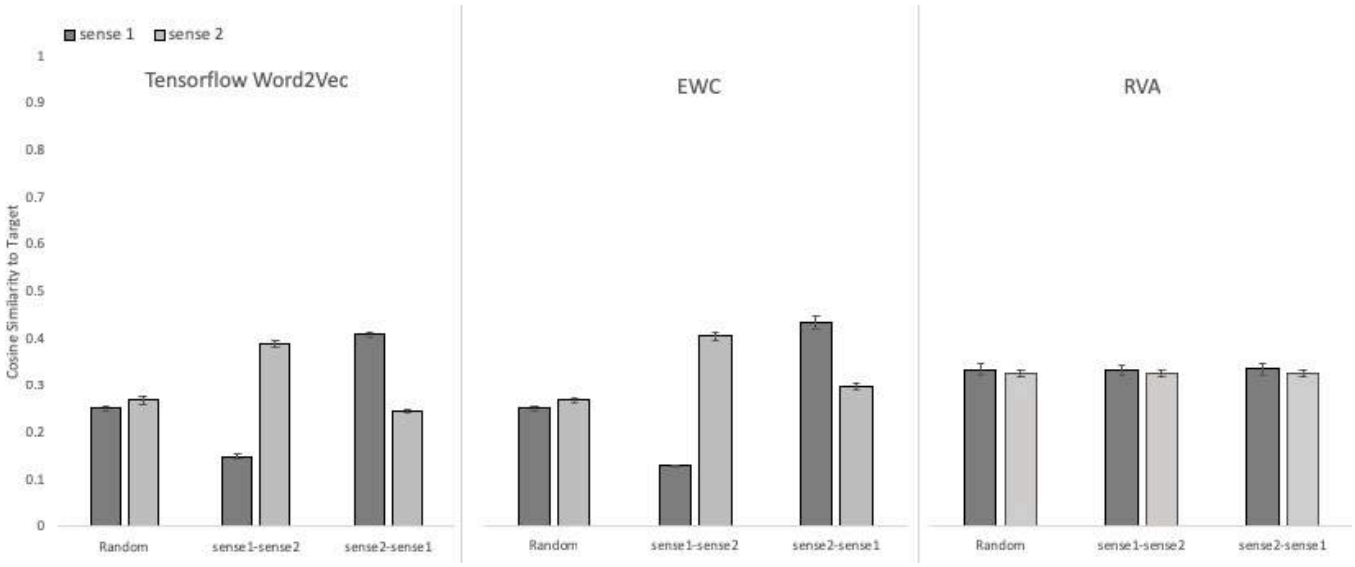


Figure 4. The y-axis represents the cosine similarity of vectors produced by word2vec, EWC, and the RVA when trained on sense-balanced homonyms from the TASA corpus. The x-axis represents one of three training orders: random order, sequential order with sense 1 first then sense2, and sequential order with sense 2 first then sense 1. CI is present in word2vec and EWC while the RVA is unaffected by CI.

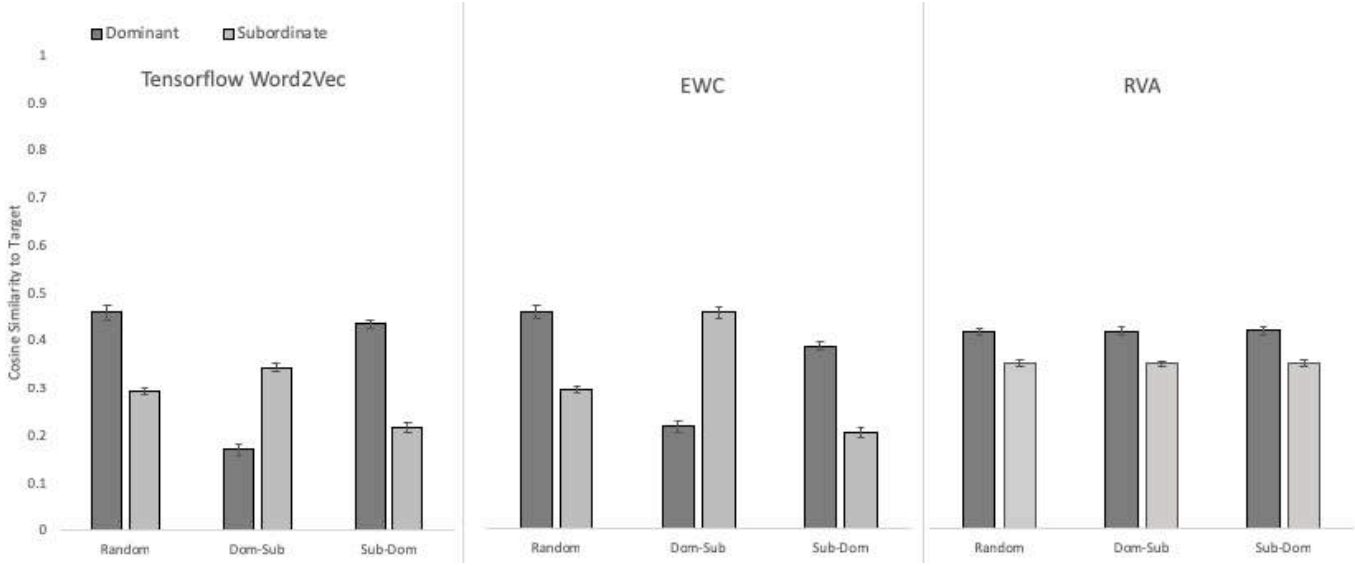


Figure 5. The y-axis represents the cosine similarity of vectors produced by word2vec and the RVA when trained on sense-imbalanced homonyms from the TASA corpus. The x-axis represents one of three training orders: random order, sequential order with the dominant first then the subordinate sense, and sequential order with the subordinate sense first then the dominant sense. CI is present in word2vec and EWC while the RVA is unaffected by CI.

Discussion

The results of this study suggest that efforts to mitigate the effects of CI need to be interdisciplinary. Within the machine learning community, insulating, or "fixing", predictive DSMs from CI is an emerging area that has seen some innovation in recent years. However, suggested solutions so far only consider the problem as it relates to strictly machine learning tasks such as categorization or image classification tasks. This study has shown that solutions from the machine learning community are not guaranteed to work when applied to tasks from different fields.

While Kirkpatrick et al. (2017) were able to show promising results from EWC on categorization tasks, the method was unable to prevent CI when applied to semantic modeling. This may be because the goal of EWC is to prevent the weights of a predictive neural network from changing based on how much information each weight carries about the *true class* of each training item. The connection between training items and their class is very straightforward in categorization tasks but is not as clear in semantic modeling tasks. When a predictive neural network learns a word representation, it is not explicitly predicting the class of a word but is attempting to predict which words belong or don't belong in context with a target word. Additionally, the window size is a variable parameter in these models which can be greater than 2, implying that a target word could have multiple "true classes" if we consider context words the class of the target word.

Additionally, EWC as it is now is not theoretically plausible for any task which requires unsupervised learning because the new loss function must be "turned on" when the network is learning a second task. This is especially cumbersome in NLP where it is impossible to supervise learning to the extent which EWC requires. Furthermore, EWC is unable to scale up well with its current implementation. Because it was designed to prevent CI in categorization tasks, it requires each training item to have a true class. This requirement prevents more efficient sampling methods which have been standardized in the DSM literature, such as noise contrastive estimation, from being used in conjunction with EWC. Similarly, calculating the Fisher Information for each node in a network becomes computationally expensive when the vocabulary and network gets large.

Introducing the RVA model as a possible solution to CI is a preliminary attempt to approach the problem of CI from the perspective of cognitive science. Within the cognitive science community, many researchers assume that the brain is primarily a predictive learner, when in reality it learns using both prediction and co-occurrence methods. Because of the tendency to favor predictive explanations of learning, predictive DSMs are still the most popular learning models in the field even though the existence of CI implies biologically implausibility. This has been documented by Ratcliff (1990) and McCloskey & Cohen (1989) who both use CI to discredit the biological plausibility of predictive DSMs. While RVAs are not a brand-new idea, they have not become as popular within the machine learning or cognitive science communities as predictive DSMs. However, they are continuous learners, can learn sequentially without incurring CI, and are computationally efficient making them a viable alternative to predictive DSMs in both the fields of cognitive science and machine learning.

While RVAs are promising, they have faced some criticism in the past. RVAs are known to have problems with metric space compression—causing most word similarities to be compressed between 0 and 1—which limits the ability of the model to discriminate between related and unrelated words (Asr & Jones, 2017). It was initially believed that predictive DSMs were able to more accurately discriminate between words because of back-propagation or the connectionist architectures they commonly use. However, recently the role of negative sampling in DSMs has been explored in more depth by Johns, Jones, & Mewhort (2019) who find that the success predictive DSMs have at discriminating between words is due to the inclusion of negative information in the training data—not the use of connectionist architecture or predictive learning method. In fact, when negative sampling information is included in the training data for other DSMs, including RVAs, their ability to discriminate words is on par with predictive DSMs.

Though this paper focused on comparing RVAs to predictive DSMs, RVAs aren't the only possible alternative architecture that could present a solution to CI. Architectures like holographic neural networks and exemplar-based models should also theoretically be immune to CI and incorporate different theoretical frameworks of learning. Holographic neural networks use convolution as an association mechanism to learn words rather than backpropagation and are able to learn complex non-linear patterns with a single layer which makes them more space efficient than predictive DSMs. Exemplar-based models, unlike other DSM models which store a semantic representation, store only episodic context. These models construct semantic meaning from the aggregation of episodic context when presented with a memory cue (Jamieson et al., 2018). Both of these models should be evaluated to determine the effect CI has on their internal semantic representations.

Up until now, the fields of machine learning and cognitive science have both been facing similar problems with predictive DSMs. Unfortunately, there has been little to no interdisciplinary communication to propose solutions. When we consider CI from a cognitive science perspective, we find that there are several possible solutions which haven't been considered yet. These solutions, which are arguably more elegant than continuously trying to "vaccinate" predictive DSMs, have the potential to introduce new mechanisms for artificial learning, assisting with new technological advances that require sequential learning and providing a framework for learning that does not exhibit the downfalls brought on by predictive DSMs.

References

Armstrong, B. C., Tokowicz, N., & Plaut, D. C. (2012). eDom: Norming software and relative meaning frequencies for 544 English homonyms. *Behavior research methods*, *44*(4), 1015-1027.

Asr, F. T., & Jones, M. N. (2017). An Artificial Language Evaluation of Distributional Semantic Models. *Proc. of the ACL Conference on Natural Language Learning*.

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings Association of Computational Linguistics* (Vol. 1, pp. 238-247).

Dachapally, P. R. & Jones, M. N. (2018) Catastrophic Interference in Neural Embedding Models, *CogSci 2018*, 1566-1571.

Firth, J. R. (1957). *A synopsis of linguistic theory* (pp. 1930–1955). Oxford.

French, R. M. (1999). Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, *3*(4), 128-135.

Harris, Z. (1970). Distributional structure. In Papers in structural and transformational Linguistics (pp. 775–794).

Jamieson, R. K., Johns, B. T., Avery, J. E., & Jones, M. N. (2018). An instance theory of semantic memory. *Computational Brain & Behavior, 1(2),* 119-136.

Johns, B. T., Jones, M. N., & Mewhort, D. J. K. (2019). The role of negative information in distributional semantic learning. *Cognitive Science.*

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review, 114,* 1-37.

Jones, M. N., Willits, J. A., & Dennis, S. (2015). Models of semantic memory. In J. R. Busemeyer & J. T. Townsend (Eds.) *Oxford Handbook of Mathematical and Computational Psychology*, 232-254.

Kirkpatrick, J., et al. (2017) Overcoming Catastrophic Forgetting in Neural Networks. *Proceedings of the National Academy of Sciences*, *114(13),* 3521–3526.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*(2), 211–240.

LeCun, Y., Cortes, C., & Burgess, C. J. C. (2012). The mnist database of handwritten images.

McCloskey, M. & Cohen, N. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In G. H. Bower (ed.) *The Psychology of Learning and Motivation*, *24*, 109-164.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111-3119).

Ratcliff, R. (1990). Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions. *Psychological Review*, *97*, 285-308.

Rehurek, R. & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks,* 45-50.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical conditioning II: Current research and theory*, *2*, 64-99.

Making the Implicit Explicit: Effects of Verbalization in Decisions from Experience

Yaoli Mao (ym2429@tc.columbia.edu)
Department of Human Development, Columbia University, 525 W. 120th Street
New York, NY 10027 USA

James E. Corter (jec@tc.columbia.edu)
Department of Human Development, Columbia University, 525 W. 120th Street
New York, NY 10027 USA

Abstract

What do people learn from experience with repeated decisions? Is it merely implicit behavioral tendencies? If so, would articulating or summarizing what is learned change behavior? Online participants (N=126) experienced 100 trials of a decisions-from-experience problem with outcome feedback. Some participants then verbally summarized what they had learned and estimated the probability of the risky gain either for themselves (Self condition) or for another hypothetical player (Other condition); others did not summarize (Control condition). Finally, they faced 20 more decision trials. Verbalizing a social message to another person significantly increased sure choices (that is, decreased risk-taking) in subsequent decision making. In general, participants underestimated the probabilities of both certain and risky prospects, and articulating a summary message (Self or Other) seemed to increase this conservatism.

Keywords: decisions from experience; explicit learning; verbalization; dual process theory

Introduction

In recent decades, research on decisions under risk has focused on two major paradigms, decisions from description and decisions from experience. In the descriptive paradigm, participants receive complete and unambiguous descriptions of available options, potential outcomes of their choices, and the associated probabilities. In the experiential paradigm, participants rely on their personal experience of observing samples of outcome feedbacks repeatedly over time (e.g., Hertwig et al., 2004). In recent years, decisions from experience (DFE) have been found to systematically differ from description-based decisions (DBD). These differences have been termed the "description–experience gap" (Hertwig & Erev, 2009).

There has been growing interest in the field to explore the learning mechanisms behind decisions from experience to help explain the description-experience gap (for a meta-analytic review, see Wulff et al., 2018). One such issue is what types of learning are generated from experience and how such learning affects subsequent choices. Some empirical evidence suggests that experience of outcome feedback can modify choices towards maximization of expected value (EV) (Yechiam et al., 2005). Possible mechanisms that might explain this finding include the implicit learning of more linear decision weights (e.g. Jessup et al., 2008), or the explicit learning of EV-maximizing strategies (e.g. Erev & Barron, 2005; Erev et al., 2017), among others. Chen and Corter (2014) argued that dual-systems account of cognition (e.g., James, 1950; Sloman, 1996; Kahneman, 2003), might be needed to explain the full range of findings.

In the broader research literature on learning and cognitive science, implicit learning is sometimes termed "System 1" thinking, in which individuals learn complex information in an incidental manner, without awareness of what has been learned. In contrast, explicit learning is termed "System 2" thinking, which permits abstract reasoning and hypothetical thinking constrained by working memory capacity, and results in explicit knowledge in the form of verbatim or aggregate representations (Seger, 1994; Evans, 2003).

In particular, two major forms of explicit learning have been well studied. *Self-explanation* during problem solving has proven to be an effective instructional strategy across many domains (Chi et al., 1989; VanLehn et al., 1992; Bielaczyc et al., 1995). When prompted to explain to themselves, participants were more likely to make comparisons and notice subtle distinctions, which then led to the discovery of general rules (Edwards et al., 2014). Meanwhile, *social dialogue* has also been found to promote abstract reasoning and rule formation / use in a category learning task (Voiklis & Corter, 2012), as well as when learning complex systems such as moving gears, biological transmissions, and organisms' living requirements (Schwartz, 1995). In these learning domains, it is argued that social pragmatic constraints of communications compel participants in dialogue to negotiate multiple perspectives to find a shareable representation of the problem, which tended to be abstractions of the deep structure rather than surface features. Such dialogic effects might even underlie well-documented examples of "process gain" in group forecasting and decision making (Kerr & Tindale, 2004) – the so-called "wisdom of crowds".

For these reasons, we hypothesize that explicit verbalizations, especially verbalizations aimed at others, might promote abstraction and enable rule-based or formal reasoning about the decision problem, and thus might yield faster learning towards EV-maximization. To our knowledge

no prior study has examined whether verbalization might help in promoting explicit learning in the context of decisions from experience.

In the present study, we consider how self-verbalizations summarizing experience with outcome feedback (which make the implicit explicit) might affect subsequent risky decision making. Specifically, we examine the effects of verbal summaries generated for others or generated for oneself on learning in the decisions from experience context. Finally, we report some content analyses of the types of verbalizations generated by participants.

Methods

Design

Participants made repeated decisions for a single risky decision problem while experiencing outcome feedback (with no provided description of outcome payoffs and probabilities). Following the verbalization manipulation (described below), they made 20 additional decisions with the same problem.

Overall the experiment had a 3×2 between-subjects design: three types of verbalization conditions and two risky-choice decision problems. Each participant was presented with only one verbalization condition and only one problem.

Participants

126 people, 76 of them male, participated through Amazon's Mechanical Turk website. Participation was restricted to individuals whose location was defined as in the United States. Their ages ranged from 23 to 71, with a mean of 39. All of them were native English speakers and 27% of them had studied statistics or decision-making at some point.

Materials

Two simple decision problems in the gain domain were used: for the risky option, one problem has a high probability of payoff and the other has a low probability of payoff. They were: Problem 1 = ($3, 100%; $7, 60%), Problem 2 = ($3, 100%; $28, 15%). So, for example, Problem 1 offered a choice between receiving $3 with certainty and a 60% chance of receiving $7 (and no reward otherwise).

We used the "minimal information" paradigm from Erev and Barron (2005) – also termed the "partial feedback" paradigm by Camilleri and Newell (2011). Decision problems were shown on the computer screen (Figure 1), with two option buttons side by side, labeled only as "P" and "Q". One button provided the participant with the sure outcome of $3 100% of the time, and the other button was a risky gamble which gave participants either $7 60% of the time or $28 15% of the time, depending on the experimental condition, and $0 otherwise. Sure and risky button positions were left-right counterbalanced between participants.

Procedure

Participants went through a training session of 100 trials and a testing session of 20 trials of the same decision problem, either the high probability problem or the low probability problem.

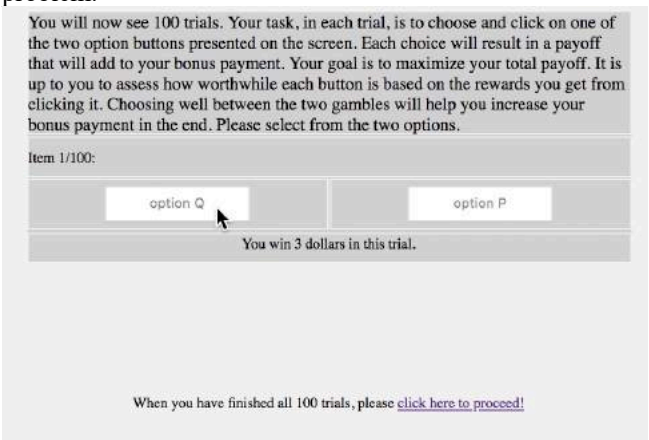


Figure 1: Interface for the training session (first 100 trials): post-trial feedback

In between training and testing blocks, they experienced one of the three verbalization conditions (Other, Self, or Control). In the Self condition, participants summarized for themselves what they had learned (by answering "What have you learned from experience with the 100 trials? What strategy should be used or what choices should be followed in order to maximize total payoff?") and estimated the probabilities of both option payoffs. In the Other condition, participants summarized to another hypothetical player (by answering "Imagine that you have a partner who is about to play this game for 100 trials. What would you advise them in terms of the strategy they should use or the choices they should make, in order to maximize their total payoff?") and estimated probabilities as well. In the Control condition, participants simply answered some demographic questions at this time point, without any requested verbalizations of problem information.

At each trial, once they made a choice using the mouse, the payoff for that selected option was shown. Actual payment for participants varied depending on the outcomes of their decisions. A base payment of $1.50 was adjusted by 0.5% of the participant's total amount of winnings for the total 120 decision trials. Average bonus paid for each participant was US $1.92 (SD = US $0.17).

Results

In this study, we hypothesized that explicit verbalizations of strategies would lead to more accurate probability estimates of option payoffs and a decrease in subsequent sure choices (consistent with EV-maximization), especially when participants were verbalizing to someone else. Thus, the main dependent variables were 1) the proportion of sure choices, calculated as the average proportion of times that participants selected the sure option in the testing session (last 20 trials)

(see b11 and b12 in Figure 2); and 2) participants' estimated payoff probabilities for the sure and the risky options.

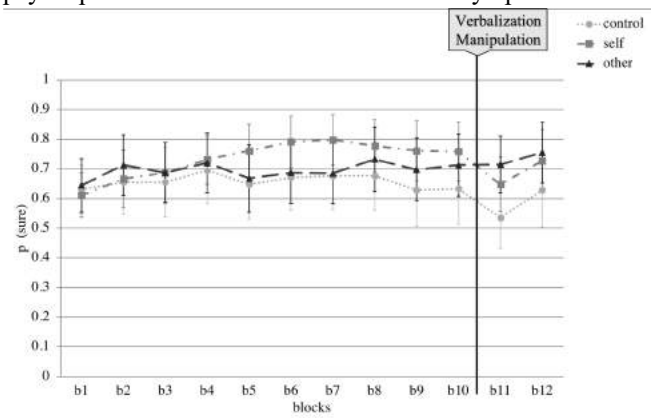


Figure 2: Sure choice proportions over the total 120 trials. Error bars: ± 2 standard errors.

Behavioral Effects: testing session (last 20 trials)

Analysis of covariance (ANCOVA) was conducted to evaluate the effects of verbalization conditions while controlling for the variations among participants' learning experience in the training session. The proportion of sure choices for last 50 trials of training, before the verbalization manipulation, was used as the covariate.

Results showed significant effects of the explicit verbalization manipulation on the proportion of sure-thing choices in the last two blocks (Figure 3), $F(2,119) = 3.80$, $p=.025$. However, the effects were not consistent with our hypothesis of increased maximization in the two verbalization conditions. Rather, in the Control and Self-Verbalization conditions a transient increase in risk-seeking (alternatively, in maximization) was observed (apparent in Figure 2), indicated by a sudden drop in sure choices after the pause between training and testing blocks, mean P(sure) = .641 and .623, respectively. This may indicate a transient increase in exploratory behavior. Participants in the Other-Verbalization condition maintained a relatively consistent high level of sure-alternative choices, P(sure) = .744. Planned contrasts showed that the proportion of sure choices in the last 20 trials after verbalization were significantly higher in the Other-Verbalization condition compared to that in the Control and Self-Verbalization conditions, $t(80)=2.23$, $p=.027$; $t(82)=2.53$, $p=0.013$, respectively.

Subjective Estimates

Participants were quite conservative in their probability estimates, underestimating probabilities of both the sure option (Figure 4) and the risky options (Figure 5). Such probability underestimation is particularly surprising for the sure events, because any sample of a sure option must consist of 100% payoff outcomes. One way to explain this is to note that in this partial-feedback paradigm, when a participant chooses the risky option, the outcome for the button associated with the sure-thing distribution is not revealed. Thus, the participant may believe that some non-payoff outcomes could be occurring for what we know to be the sure-thing option on these "blind" trials. And in any small sample of trials, it is difficult to distinguish a sure-thing from a high-probability event, just as it is known that in decisions from experience people frequently fail to distinguish between low-probability and zero-probability events (see, for example, Kunreuther et al., 2001; Hertwig et al., 2004).

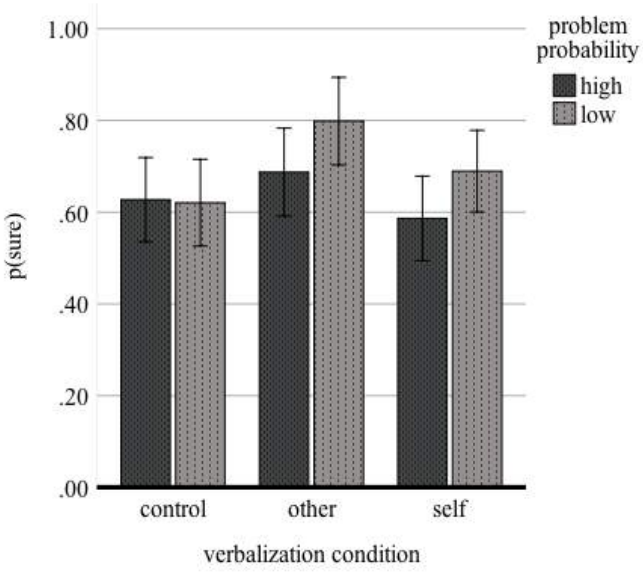


Figure 3: Sure choice proportions across three verbalization conditions in the testing session (last 20 trials). Error bars: ± 2 standard errors.

Furthermore, articulating a summary message (to Self or Other) significantly increased this underestimation of the 100% probability of the sure option, $F(2,123)=10.270$, $p=.012$, again contrary to our hypothesis that verbalization would increase accuracy. However, when participants estimated the probability of payoff for the risky option, this drop in the subjective estimate (an increase in conservatism, again resulting in lower accuracy) due to verbalization was only marginally significant for the high-probability problem $F(2,59)=2.943$, $p=.061$, and was not significant for the low-probability problem, $F(2,59) = 2.222$, $p=.117$, perhaps due to a floor effect, or because conservatism in this case would mean estimating the probability as less extreme (i.e. farther from 0).

Verbalization Content

The above results demonstrate that the explicit verbalization manipulation has an effect on subsequent decision choices as well as on subjective estimates. However, the verbalization manipulations did not increase the accuracy of the subjective estimates as we expected, and even decreased it in some cases.

To explore why, we conducted a content analysis of the strategies reported by participants in the two verbalization conditions (40 statements in the Other condition, 44 statements in the Self condition). Specifically, we were interested in examining the detailed content, both as a manipulation check and to explore the major concepts and terms used by participants in communicating the problem information. We used two raters to categorize the verbalizations (initial κ=0.66), who discussed the disagreements until they reached full agreement, $\kappa \geq .99$.

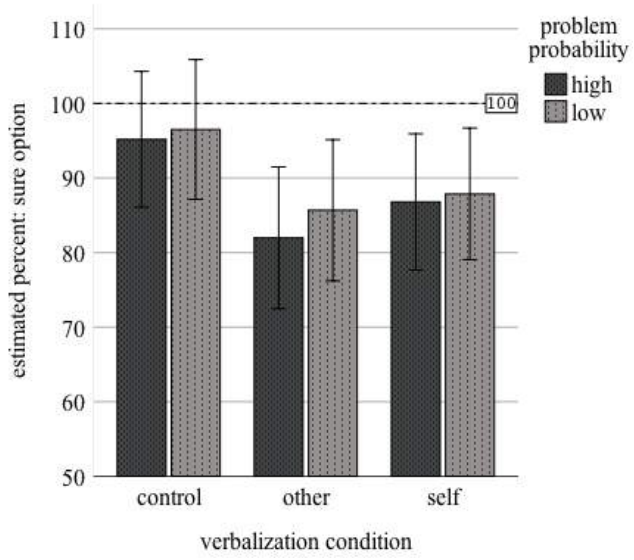


Figure 4: Estimated payoff percent for the sure option (objectively = 100% in both high- and low- probability problems). Error bars: ± 2 standard errors.

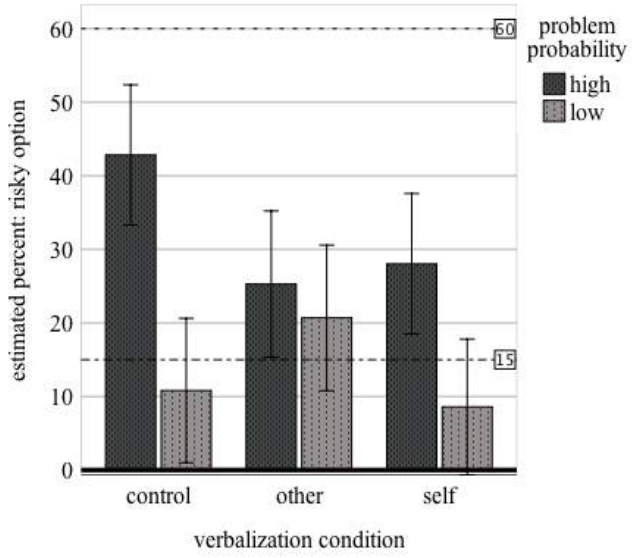


Figure 5: Estimated payoff percent for the risky option (objectively = 60% in the high-probability problem, = 15% in the low-probability problem). Error bars: ± 2 standard errors.

We divided the analysis into two phases. An initial classification suggested ten categories of utterances (which we will refer to as "strategies", for convenience): *no strategy/intuition/luck, payoff value, payoff frequency, temporary switch, sequence, risk-reward tradeoff, probability and EV estimate, recommend the sure option, recommend the risky option,* and *recommend mixed options*. Next, we tested the association of specific strategies with subsequent decision making (specifically, with the proportion of sure option choices after the verbalization manipulation), using one-way analysis of variance.

In general, participants verbalized a wide range of strategies, ranging from 1 to 5 when verbalizing to Others and from 1 to 6 when verbalizing to Self. And a majority of participants in each condition verbalized at least 2 strategies. The two verbalization conditions seemed to have very different profiles of strategy use (Figure 6). Participant who verbalized to themselves were significantly more likely to describe payoff frequency (75%), compared to those verbalized a social message (48%), $\chi^2(1; N=52) = 6.719$, p= .01 < .05. Participants tended to simply recommend the sure option more often when they were writing a social message (55%) compared to verbalizing to themselves (43%), however this difference did not reach significance. In both conditions, only a few participants mentioned calculating probability or expected value (5% in Other-Verbalization and 14% in Self-Verbalization), although more mentioned reasoning about tradeoffs between risk and reward.

Consistent with previous findings, mentions of switching between options to learn about payoff patterns or follow a sequential pattern were observed. Examples include: *"the first option had a pattern between getting 0 and 7 dollars while the other was 3 every time. I thought I could discern the pattern and only hit the first option when I thought the 7 would be there." "Keep pressing the left button until you get more than 2 zeros in a row. Then press the right button about 2 or 3 times, then go back to pressing the left button.").* In the Other-verbalization condition, 15% of participants mentioned switching or sequential dependencies in outcomes, compared to 9% in the Self-Verbalization condition. Some of these utterances may be taken as indicating that a participant exhibits some form of the gambler's fallacy, in which they believe a run of wins will tend to end, or the hot hand fallacy, where they believe such runs tend to continue (Bar-Hillel & Wagenaar, 1991). A number of participants also recommended mixed options as a better strategy than sticking to one option (cf. Chen & Corter, 2006).

Overall, more (85%) Social messages (compared to Self messages) tended to prescribe an action to be taken (example: *"Go with the three dollars most of the time, but occasionally try your luck to get the 7 dollars, since it has fairly good odds."*), $\chi^2(1; N=44) = 32.574$, $p < .001$; while Self messages were more likely (75%) to simply describe the past experience (example: *"Second option had consistent payoff. I am risk averse so I only tried the other a couple of times and*

hit zero so I stayed with the sure thing."), $\chi^2(1; N=45) = 17.058$, $p < .001$ (following "prescriptive rule" vs "descriptive rule", Bell et al., 1988).

In a one-way ANOVA testing if participants' choice behavior differed on the basis of their verbalized content (Figure 7), we found that the proportion of sure choices were significantly higher if participants recommended the sure option, $F(1, 82)=7.063$, $p=.009 < .05$; while significantly lower if they depended on no strategy or pure intuition and luck, $F(1, 82)=4.032$, $p=.048 < .05$, mentioned a temporary switch, $F(1, 82)=4.601$, $p=.035 < .05$, or recommended the risky option, $F(1, 82)=5.296$, $p=.024 < .05$.

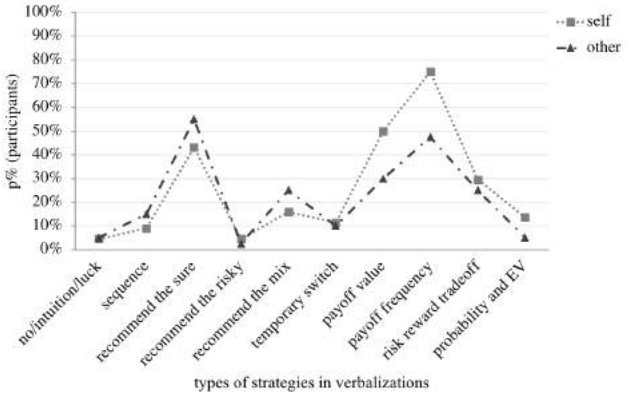


Figure 6: Verbalized content profiles by participants' verbalization condition

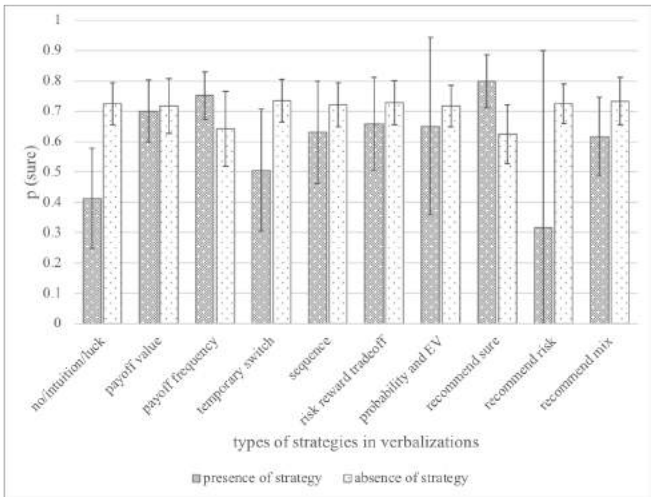


Figure 7: Sure choice proportions in the testing session (last 20 trials) by type of participants' verbalization. Error bars: ± 2 standard errors.

Discussion

In this study, we asked participants to articulate what was learned from experience, either to themselves or to others. We did not find evidence to support the original hypothesis that explicit verbalization, especially verbalizing to someone else, promotes abstract rule reasoning and thus yields better learning towards EV-maximization. Indeed, the data revealed a very different pattern. Delivering a social message tended to increase the underestimation of the subjective probability estimates (a form of conservatism), and subsequently led to less EV-consistent decision making, compared to the purely implicit learning condition in which no verbalization was prompted.

We also had hypothesized that the verbalization manipulation may tend to shift people from implicitly motivated behavior to the use of explicit strategies. These strategies can be either rational and effective, such as drawing on memory for outcome feedback and reasoning about tradeoffs, or heuristic in nature, such as choosing a simple strategy or prescribing the same. However, very few participants in either verbalization condition reported calculating probability or expected value. Instead, in Self summaries they tended to simply describe past learning experience, especially summarizing frequency information; while in summaries for Others, they often simply prescribed strategies (positive or avoidance) to others.

One potential reason for this lack of benefit from explicit verbalizations is that prior studies showing learning benefits from social dialogue (e.g., Schwartz, 1995; Voiklis & Corter, 2012) examined situations where participants took many rounds to negotiate multiple perspectives and generate abstractions and rules. In contrast, the one-way, single-round verbalizations in this study may induce considering another's perspective to some degree, but perhaps not enough to spur abstraction and use of explicit or formal strategies. This is consistent with the finding from research on collective intelligence that the equality in distribution of conversational turn-taking is correlated with a higher collective intelligence factor (Woolley et al., 2010). Moreover, relatively naïve participants may lack expert knowledge or language to convey sophisticated strategies like expected value in the risky decision domain.

We found that verbalizations, especially in the form of a social summary message to another person, led to a higher level of sure choices in subsequent decisions, perhaps by "freezing" the recommender's strategy and inhibiting further exploration of decision options (see discussion below). Also, verbalization seemed to increase underestimation of probabilities (for both certain and risky events), perhaps indicating a form of "social conservatism", as if the participants were cautious about their limited information acquired from experience and discounted their judgments to communicate a "safer" message socially. This is consistent with Benjamin and Budescu's (2015) findings about advice giving, in which an implicit learning mode (decisions from experience) resulted in more risk aversion and acknowledgement of information uncertainty.

Moreover, some previous studies using the repeated decisions with description paradigm seem to show that choice behaviors are mainly affected by experience while explicit descriptions are considered only when they carry novel or

inconsistent information that cannot be inferred from the feedback (Barron et al., 2008; Weiss-Cohen, et al., 2016); more often the descriptions seem to be neglected (Jessup et al., 2008; Lejarraga & Gonzales, 2011). When participants were writing a social message to others, they might be more conscious of the utility of information, assuming that their verbalizations would be taken into consideration. It may be that in this situation, underestimation of payoff probabilities (a "sin" of omission, in a sense) is seen as less undesirable than overestimation of payoff probabilities (a "sin" of commission).

As noted, in the Control and Self-Verbalization conditions, the final 20 test trials elicited a period of exploratory behavior, but not in the Other-Verbalization condition. Furthermore, participants in the Self condition tended to describe their past experience with the 100 trials (i.e. payoff frequency and value) while those in the Other condition tended to prescribe a future action (i.e. recommend one option or mixed options). This may indicate a social motive to seem consistent when giving advice, in line with the behavior consistency principle (Cialdini, Trost, & Newsom, 1995), well established by dissonance and balance theories, (Festinger 1957; Heider 1958) and the "foot-in-the-door" effect (Freedman & Fraser 1966). According to Group-Centrism (Kruglanski et al., 2006), the need for cognitive closure within the group induces pressures to opinion uniformity, rejection of deviates, resistance to change, conservatism and the perpetuation of group norms, and results in reduced information exchange and "premature consensus" or "early closure" (Kruglanski & Webster, 1996), and process losses that leads to less optimal group performance (Steiner, 1972). Furthermore, the bias towards shared information, once explicitly formed, can also lead to misinterpretation of new information that is inconsistent with already formed bias (Kerr & Tindale, 2004). Social context, here in the form of a social probe to verbalize strategies explicitly, might also exacerbate individuals' desire to be consistent in their explicit strategy verbalizations, probability estimates, and subsequent behaviors.

In conclusion, our results do show verbalization effects on implicit learning in decisions from experience. This evidence can be seen as supporting accounts that recognize an explicit learning aspect in decisions from experience as well as the importance of social contexts, and also as supporting dual-process accounts of repeated decisions with outcome feedback. Further exploration of the verbalization effect and of the interplay between experience and abstractions of experience might consider a broader range of factors that contribute to rule abstraction, to better understand how people can make informed decisions that combine explicit reasoning and implicit experience. This future research might find a way to integrate research on mental representations in decisions from experience (Camilleri & Newell, 2009), advice giving in decision making (Benjamin & Budescu, 2015) and information shareability in the general learning domain (Freyd, 1983).

Acknowledgments

Thanks Lujain Al-Alamy and Jiaqi Wan for their help and support in coding verbalizations and finalizing the coding scheme.

References

Bar-Hillel, M., & Wagenaar, W. A. (1991). The perception of randomness. *Advances in Applied Mathematics, 12(4),* 428–454.

Barron, G., Leider, S., & Stack, J. (2008). The effect of safe experience on a warnings' impact: Sex, drugs, and rock-n-roll. *Organizational Behavior and Human Decision Processes, 106(2),* 125–142.

Bell, D. E., Raiffa, H., & Tversky, A. (1988). Descriptive, normative, and prescriptive interactions in decision. In D. E. Bell, H. Raiffa, & A. Tversky (Eds.), *Decision making.* Cambridge, MA: Cambridge University Press.

Benjamin, D., & Budescu, D. V. (2015). Advice from experience: Communicating incomplete information incompletely. *Journal of Behavioral Decision Making, 28(1),* 36–49.

Bielaczyc, K., Pirolli, P. L., and Brown, A. L. (1995). Training in self-explanation and self-regulation strategies. *Cognition and Instruction*, 13 (2), 221-252.

Camilleri, A. R., & Newell, B. R. (2009). The role of representation in experience-based choice. *Judgment and Decision Making, 4(7),* 12.

Camilleri, A. R., & Newell, B. R. (2011). When and why rare events are underweighted: A direct comparison of the sampling, partial feedback, full feedback and description choice paradigms. *Psychonomic Bulletin & Review, 18(2),* 377–384.

Chen, Y.-J., & Corter, J. E. (2006). When mixed options are preferred in multiple-trial decisions. *Journal of Behavioral Decision Making, 19(1),* 17–42.

Chen, Y. J., & Corter, J. (2014). Learning or framing?: Effects of outcome feedback on repeated decisions from description. In *Proceedings of the Annual Meeting of the Cognitive Science Society, 36*(36).

Chi, M. T., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, *13*(2), 145-182.

Edwards, B., Williams, J., Gentner, D., & Lombrozo, T. (2014). Effects of comparison and explanation on analogical transfer. *Proceedings of the Annual Meeting of the Cognitive Science Society, 36*(36).

Erev, I., & Barron, G. (2005). On adaptation, maximization, and reinforcement learning among cognitive strategies. *Psychological Review,* 112(4), 912–931.

Erev, I., Ert, E., Plonsky, O., Cohen, D., & Cohen, O. (2017). From anomalies to forecasts: Toward a descriptive model of decisions under risk, under ambiguity, and from experience. *Psychological review*, *124*(4), 369.

Evans, J. S. B. T. (2003). In two minds: dual-process accounts of reasoning. *Trends in Cognitive Sciences,* 7(10), 454–459.

Festinger, L. (1957). *A theory of cognitive dissonance*. Palo Alto, CA: Stanford University Press.

Freedman, J.L. & Fraser, S.C. (1966). Compliance without pressure: the foot-in-the-door technique, *Journal of Personality and Social Psychology, 4,* 196-202.

Freyd, J. J. (1983). Shareability: The social psychology of epistemology. *Cognitive Science*, *7*(3), 191-210.

Heider, F. (1958). *The psychology of interpersonal relations*. New York, NY: Erlbaum.

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science*, *15*(8), 534-539.

Hertwig, R., & Erev, I. (2009). The description–experience gap in risky choice. *Trends in Cognitive Sciences, 13(12),* 517–523.

James, W. (1950). *Principles of Psychology.* New York, NY: Dover.

Jessup, R. K., Bishara, A. J., & Busemeyer, J. R. (2008). Feedback produces divergence from prospect theory in descriptive choice. *Psychological Science, 19(10),* 1015–1022.

Kahneman, D. (2003). A perspective on judgment and choice: mapping bounded rationality. *American Psychologist,* 58(9), 697.

Kerr, N. L., & Tindale, R. S. (2004). Group performance and decision making. *Annual Review of Psychology*, *55*, 623.

Kunreuther, H., Novemsky, N., & Kahneman, D. (2001). Making low probabilities useful. *Journal of risk and uncertainty*, *23*(2), 103-120.

VanLehn, K., Jones, R. M., & Chi, M. T. (1992). A model of the self-explanation effect. *The Journal of the Learning Sciences*, *2*(1), 1-59.

Kruglanski, A. W., & Webster, D. M. (1996). Motivated closing of the Mind: "seizing" and "freezing". *Psychological Review*, *103*(2), 263-283.

Kruglanski, A. W., Pierro, A., Mannetti, L., & De Grada, E. (2006). Groups as epistemic providers: Need for closure and the unfolding of group-centrism. *Psychological Review, 113(1)*, 84.

Lejarraga, T., & Gonzalez, C. (2011). Effects of feedback and complexity on repeated decisions from description. *Organizational Behavior and Human Decision Processes, 116(2),* 286–295.

Schwartz, D. L. (1995). The emergence of abstract representations in dyad problem solving. *Journal of the Learning Sciences, 4(3),* 321–354.

Seger, C. A. (1994). Implicit learning. *Psychological Bulletin, 115(2),* 163.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*(1), 3.

Steiner, I. D. (1972). *Group process and productivity.* New York, NY: Academic Press.

Voiklis, J., & Corter, J. E. (2012). Conventional wisdom: negotiating conventions of reference enhances category learning. *Cognitive Science, 36(4),* 607–634.

Weiss-Cohen, L., Konstantinidis, E., Speekenbrink, M., & Harvey, N. (2016). Incorporating conflicting descriptions into decisions from experience. *Organizational Behavior and Human Decision Processes, 135*, 55–69.

Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., & Malone, T. W. (2010). Evidence for a collective intelligence factor in the performance of human groups. *Science*, *330*(6004), 686-688.

Wulff, D. U., Mergenthaler-Canseco, M., & Hertwig, R. (2018). A meta-analytic review of two modes of learning and the description-experience gap. *Psychological Bulletin, 144(2),* 140–176.

Yechiam, E., Barron, G., & Erev, I. (2005). The role of personal experience in contributing to different patterns of response to rare terrorist attacks. *Journal of Conflict Resolution, 49(3),* 430–439.

Same Words, Same Context, Different Meanings: People are unaware that their own concepts are not always shared

Louis Martí (I_Am_A_Robot@berkeley.edu)
Department of Psychology, 2121 Berkeley Way
Berkeley, CA 94704 USA

Steven Piantadosi (stp@berkeley.edu)
Department of Psychology, 2121 Berkeley Way
Berkeley, CA 94704 USA

Celeste Kidd (celestekidd@berkeley.edu)
Department of Psychology, 2121 Berkeley Way
Berkeley, CA 94704 USA

Abstract

A long-standing assumption in cognitive science has been that concepts are shared among individuals for common words. However, given that concepts are formed by the data we observe, and observations vary wildly across individual experiences, our concepts are not likely identical. Here, we present data in which 104 participants answer questions regarding their beliefs about the definitions of common everyday words, and the degree to which they think others agree. Our results suggest that even for common words, there exist many distinct extensions of ordinary and political concepts across individuals. There is also a pervasive bias which leads individuals to overestimate the degree to which others agree, which may explain why "talking past each other" is an anecdotally common experience when discussing important topics.

Keywords: Concepts; Metacognition; Individual Differences; Miscommunication

Introduction

In 1964, the United States Supreme Court heard Jacobellis v. Ohio, a case in which theater owner Nico Jacobellis was fined for exhibiting a dramatic French film about adultery that contained material which the state considered obscene. Justice Potter Stewart, in explaining why he believed the film did not violate the state's obscenity laws, stated that he was unable to define pornography but also said "I know it when I see it". In this instance, the highest authority in the country was tasked with categorizing an edge-case which would affect the lives of millions and he admitted that his criteria for categorization was difficult to articulate.

Even words with seemingly precise meanings can be deceptively ambiguous. Especially if neither party anticipates a problem because of the word's commonality. These non-obvious misalignments can have serious consequences. For example, toxicologists and non-toxicologists likely have different concepts for the word "hazard". Toxicologists define the word as referring to anything that could potentially cause harm—even if unlikely. For example, a toxicologist would categorize water as a hazard because it is possible to overdose if excessive quantities are consumed. However, for non-toxicologists, the word "hazard" refers to things which are dangerous—*likely* to cause harm, not simply capable of causing harm under specific or unlikely circumstances. This misalignment caused problems when toxicologists with the World Cancer Association labelled coffee as a known hazard for developing cancer in mice and cell cultures. A California judge, who likely possessed the concept synonymous with "dangerous", interpreted the report as meaning coffee was dangerous for consumers and ruled that California had to warn consumers.

These examples illustrate that not only can words be hard to define, but we sometimes have very different ideas about what they mean. When two people use the same word, they may assume that they are each referring to the same (or at least a similar) concept. But how often is this assumption correct? Communication requires involved parties to understand each other correctly. A necessary component of this during language use is that words map onto the same meanings for all conversation partners, or, alternatively, that they are at least aware of the possibility for misalignment. If this is not in fact occurring, it could provide new insights into why and how people disagree and misunderstand one another. Understanding these dynamics could, likewise, be used to facilitate better communication in general. Thus, conceptual misalignment has important implications in a wide range of domains, including public policy, diplomacy, education, and politics.

All theories of concepts involve learning via interaction with and data accumulation from the world. These experiences vary (often wildly) across individuals. If individuals are using the same word to refer to two different concepts, confusion and miscommunication may occur. There is some empirical evidence that at least some of people's concepts do in fact vary across individuals (McCloskey & Glucksberg, 1978). Labov (1973) asked participants to categorize objects as either a "cup" or a "bowl" as he varied the heights and widths of the objects. For extreme values of either height or width, there was widespread agreement on the classification. However, as the values became more moderate, the classifications became more subjective. This demonstrates that people's concepts are fuzzy along the edges, even with everyday

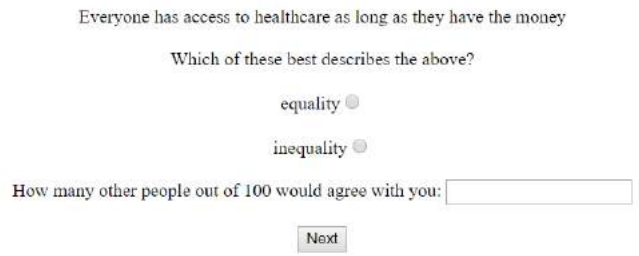


Figure 1: Participants saw 200 randomized trials as above.

objects.

If people do not agree even on category boundaries for concrete objects, how much misalignment might exist among more abstract concepts? And are people aware of the fact that concepts vary across individuals? To date, there has been no work exploring these questions.

Our first goal is to quantify individual differences in conceptual representations. By utilizing an approach which targeted edge-cases, we optimized our chances of detecting differences in definitional boundaries. Edge-cases are both theoretically and functionally important. They are crucial to conceptual definitions, and are arguably where the highest utility can be found due to the possible illusion of confidence people have about others' definitions. For example, common debates about abortion, gun-control, and welfare hinge on edge-cases.

We did this by collecting peoples assessments of whether particular scenarios applied to specific concepts. We then applied a clustering algorithm to group participants by similarity of their conceptual representations. We borrowed techniques from machine learning that have previously been applied in biology and ecology in order to estimate the total number of distinct representations on a population level. Our second goal is to quantify peoples metacognitive awareness of differences in each others conceptual structures. If people are unaware of the variability in others classifications of everyday concepts, it would make communication more difficult. In this paper, we will probe the variability of people's concepts and measure their awareness of any differences.

Methods

We recruited 104 participants on Amazon Mechanical Turk and queried them regarding their beliefs about whether a particular word applied to a given phrase or sentence. For each of 200 trials (see Figure 1) a phrase or sentence was displayed. The participant was then presented with two opposites and asked to classify the phrase or sentence. For example, after reading the sentence "A murderer is killed", participants answered whether they thought it was *justice* or *injustice*. They also answered how many people out of 100 would agree with them.

Word	Sentence	Reliability Pair
justice/injustice	A guilty man is executed	A man who is guilty is put to death
adult/child	A 17-year-old	An individual who is almost 18

Table 1: Sample reliability sentences

Word	Phrase/Sentence
equality/inequality	Taking wealth from the rich and giving it to the poor
fairness/unfairness	Paying none of your workers because you don't have the money for everyone
justice/injustice	A thief's stolen property is stolen
peace/conflict	A field filled with corpses after a war is over
honesty/dishonesty	Making true but misleading statements
safety/danger	Preventing you from drinking soda
freedom/prohibition	Making murder illegal
transparent/secretive	Releasing your taxes behind a paywall
education/ignorance	Home schooling in the US
healthcare/illness	Insurance not paying for your medical bills despite you paying your premiums
day/night	Dusk
hot/cold	A temperate day
light/dark	Classical music
friend/enemy	A close acquaintance who insults you all the time
boy/girl	A transgender woman
love/hate	Spanking a child so that they will not become spoiled
adult/child	A 17-year-old
good/bad	An entire building full of murderers was destroyed
sun/moon	A star that orbits a planet
ceiling/floor	The top surface in an upside-down house

Table 2: Sample stimuli presented to subjects in the experiment.

Word Choices

Stimuli were divided into ten political words and ten frequently used nouns. Half of all participants answered questions regarding political words while the other half answered questions about the nouns. Political concepts were chosen by asking 130 mTurkers to list the top ten words they felt were most relevant to politics. The top ten most frequent words were then chosen as our political concepts. The ten nouns

were chosen by querying the MRC Psycholinguistic Database for the ten most frequent nouns and omitting words which were close semantic duplicates (e.g. boy vs. man). This was done in order to maximize semantic variability in our word pairs as much as possible.

Sentence Construction

The specific sentences participants are responding to for each word are of crucial importance. One might imagine a set of sentences could be chosen for the word "boy" which would result in near universal agreement among participants. On the other hand, sentences could be constructed in such a way as to maximize disagreement (a 50/50 split for each binary response). If our goal is to discover whether people possess different concepts, the latter approach is appropriate. Specifically, since edge cases are often where the greatest variability lies, we will probe people's classifications of edge cases. This approach allows us to get a rough estimate of the maximum conceptual variability for each word (see Table 2 for a sample of phrases/sentences).

Reliability

Each trial had an associated reliability trial which presented the same phrase or sentence except for a minor modification which did not change the meaning. These were added in order to assess subject attention and reliability. (see Table 1 for a sample of phrases/sentences)

Analysis

Ascertaining whether participants possess different conceptual representations for a given word is a non-trivial problem. We first run into the problem of how to quantify differences between conceptual representations. What does it mean for person A's concept of "justice" to be twice as far from person B's as person C's is? We address this by representing each person's concept using a binary response vector. Next, we run into the issue of measurement noise and participant reliability. If person A answers that "A clear night with a full moon" is "light" but also answers that "A cloudless night with a full moon" is "dark", it would be reasonable to label these responses as unreliable noise. The "reliability" sentences provide semantic duplicates for each sentence, allowing us to quantify the reliability of participants in the task.

Once we have quantified participants' reliability and concepts, our last major issue is deciding how much of a difference between two concepts is sufficient to call them distinct. For the concept blue, one individual might be centered on the 470 nanometer wavelength while another might be centered on 480 nanometers. What would not be clear, however, is whether that disparity is sufficiently different so as to reasonably characterize the individuals as having separate concepts for blue. We approach this challenge by clustering our participants such that people with similar concepts will be grouped in the same cluster. We do this by adopting Bayesian approaches that find the optimal partition of participant responses using a trade-off between data-fit and simplicity. If the responses of one participant are very similar to those of another participant, they will likely be placed in the same cluster. On the other hand, if two participants have very different responses, they will likely be placed in different clusters, despite the process's overall conservative preference for fewer total clusters (Anderson, 1991). More specifically, we will use a Chinese Restaurant Process prior. If $[x_1, x_2, \ldots, x_k]$ is a vector denoting how many of the n subjects have each concept (for a given word), then the CRP prior is

$$P([x_1, x_2, \ldots, x_k]) = \frac{1}{n!} \prod_i (x_i - 1) \tag{1}$$

Within each "table" of the CRP, we use a Beta-Bernoulli likelihood, meaning that subjects assigned the same cluster are assumed to generate the same latent vector of binary answers. This vector is then measured with noise (α), and the latent probabilities are integrated out. Thus, if y_j and n_j are the number of "yes" and "no" responses in a given cluster assignment to the j'th item of a given concept, then the likelihood is,

$$\prod_j \frac{\Gamma(2\alpha) \cdot \Gamma(y_j + \alpha) \cdot \Gamma(n_j + \alpha)}{\Gamma(y_j + n_j + 2\alpha) \cdot \Gamma(\alpha)^2 \cdot y_j! \cdot n_j!} \tag{2}$$

With this setup, we used a Gibbs sampler to sample from the posterior on clusters given the responses for each concept. This analysis provides us with the number of distinct conceptual representations possessed by our participants. While this is useful information, what we are actually interested in is the total number of conceptual representations which exist on the planet. Ecologists have faced a very similar problem in estimating the number of species which exist. Often, they possess observed counts of individuals and of species in a given area (the Amazon rainforest for example) and would like to estimate the true number of species for that area (Bunge & Fitzpatrick, 1993). Here, we use the number of sampled concepts across the number of sampled individuals to estimate the total number of concepts which exist across the population of Earth for each of our words. Given that our clustering algorithm has a conservative preference for fewer clusters, this preference will extend to our global estimate.

Results

We excluded participants who did not have a reliability greater than 70% (9 out of 104 participants). We also excluded participants who gave the same answer to all agreement-prediction questions (2 out of 95 participants). We did not require participants to perform flawlessly, however, as this would be an unrealistically high bar for humans completing so many trials. Of the remaining participants, their probability of giving the same answer to both questions in the reliability pair was a respectable 86%.

There are about five distinct concepts per word

Figure 2 shows the estimated true number of concepts (y-axis) across 4,000 iterations of our clustering algorithm (us-

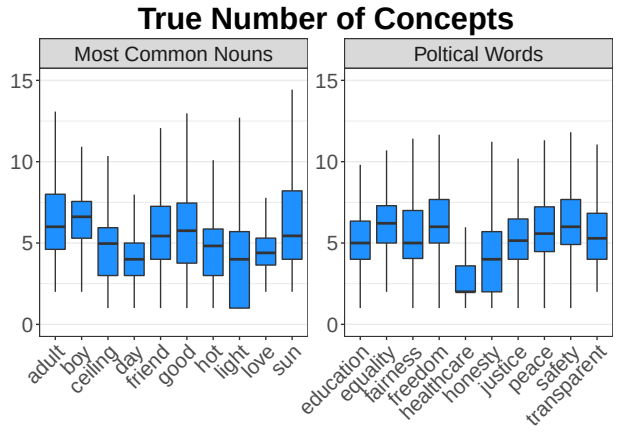


Figure 2: Estimated true number of concepts for 4,000 iterations of our clustering algorithm using a simplicity prior after a 1,000 iteration burn-in. Median estimates are roughly the same regardless of the type of word (about 5).

ing a simplicity prior) after 1,000 iterations of burn-in for each word (x-axis). Across the 4,000 samples from each word, the median estimates are always about the same regardless of word type: roughly five concepts. We also ran our clustering algorithm using a uniform prior which resulted in the same pattern of estimates except increased by two. In comparison to a simplicity prior, a uniform prior will prefer a larger estimate of concepts. That a uniform prior resulted in seven concepts, not 7,000, strongly suggests that the true number of concepts for our chosen words is close to our estimates.

Additional participants are unlikely to significantly increase our estimates

Additionally, we can run our algorithm with varying amounts of data in order to confirm our results. As the number of participants we sample increases, we should expect the distance between our sample estimate and global estimate to narrow and eventually converge. Figure 3 shows the number of concepts (y-axis) by the number of people sampled (x-axis). For most words, as the number of people sampled increases, the true number of concepts (in blue) also increases. This suggests that our estimates for these concepts are relatively conservative, as it is unlikely we have sampled enough to cause this process to plateau. This, in addition to the inherent conservatism in our clustering algorithm (the simplicity prior), suggests these are lower bound estimates for our tested concepts. However, given the slow rate of increase between sample sizes, it is unlikely our estimates would ever grow significantly, even with many more participants.

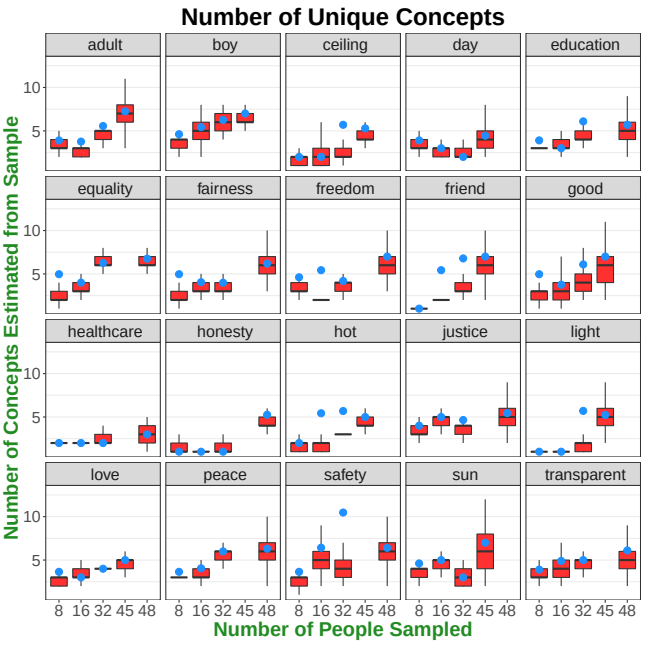


Figure 3: Number of concepts (y-axis) depending on the number of people sampled (x-axis) using a simplicity prior. Box-plots represent 25% to 75% quantiles of the number of concepts in our sample. Blue dots represent the estimated number of unique concepts on Earth based on our sample estimates. As the number of people sampled increases, the number of estimated concepts tends to increase by a slowing amount. This suggests that although our current estimates are conservative, the true number is not much higher.

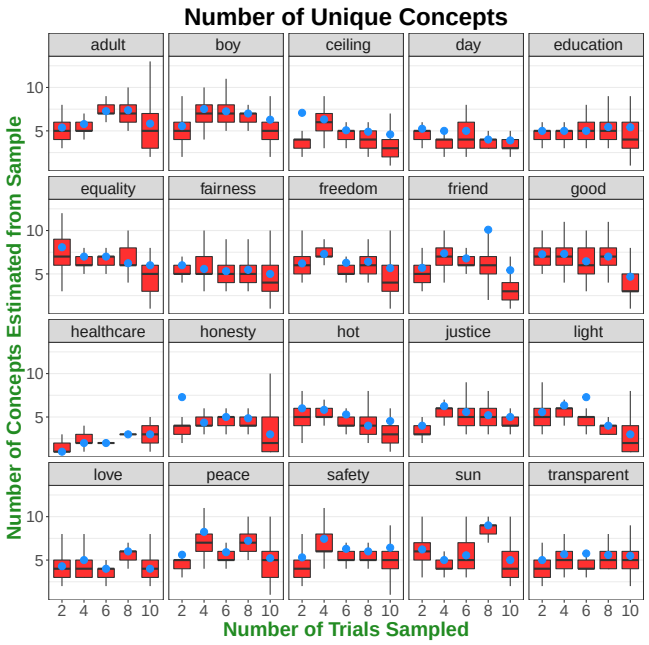


Figure 4: Number of concepts (y-axis) depending on the number of sentences sampled (x-axis) using a simplicity prior. Box-plots represent 25% to 75% quantiles of the number of concepts in our sample. Blue dots represent the estimated number of unique concepts on Earth based on our sample estimates. As the number of sentences sampled increases, the number of estimated concepts tends to stay the same. This suggests that our sentence choices are sufficiently varied enough to capture concept diversity.

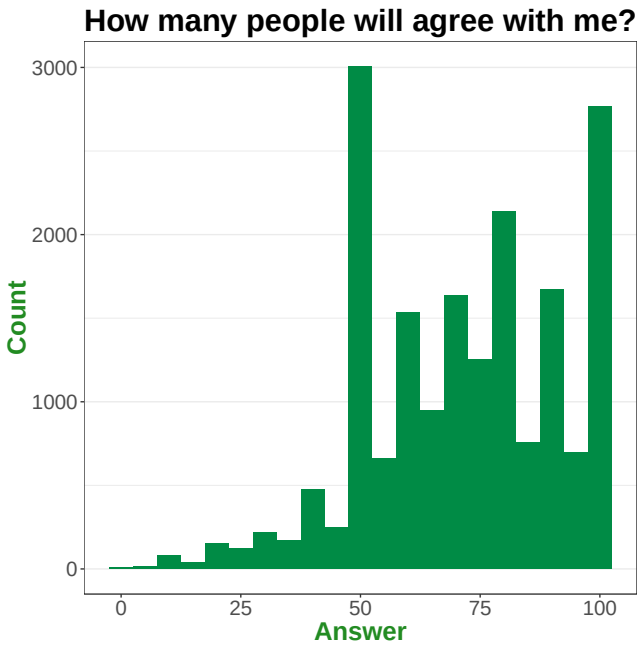


Figure 5: Raw counts (y-axis) of participant answers to the question "How many other people out of 100 would agree with you?" (x-axis). People overwhelmingly think the majority of other people will agree with their assessment.

Additional sentences are unlikely to significantly increase our estimates

Figure 4 shows the number of concepts (y-axis) by the number of sentences sampled (x-axis). If concepts were unique for each individual, one would expect the number of estimated concepts to steadily increase as the number of sentences increased. Instead, we see our estimate as relatively stable, even sometimes *decreasing* as more sentences are sampled. This also suggests that our sentence choices were sufficiently varied to capture concept diversity.

Most individuals underestimate conceptual variability

We then examined people's guesses about how often others' agreed with them. Figure 5 shows raw counts (y-axis) of participant responses (x-axis). The figure illustrates a very strong "like me" bias where the overwhelming number of responses indicate a belief that most others' will agree with their categorization. The second most common response was that *all* others will agree with their assessment.

We then assessed the relationship between people's categorizations to their guesses about the answers of others. Figure 6 presents people's predicted answers (y-axis) and their actual answers (x-axis). As the figure shows, a sizeable number of trials are not well predicted by participants. A perfect prediction rate would result in all trials landing on the y = x line. Although there are many trials which fall on or near this line, there also seems to be a consistent trend of partic-

Figure 6: Participants' actual responses vs. the responses they expected others' to give. Each data point represents the mean response for a trial/choice pair. Most data points are above the y = x line, illustrating that people largely overestimate to which others' agree with their assessments.

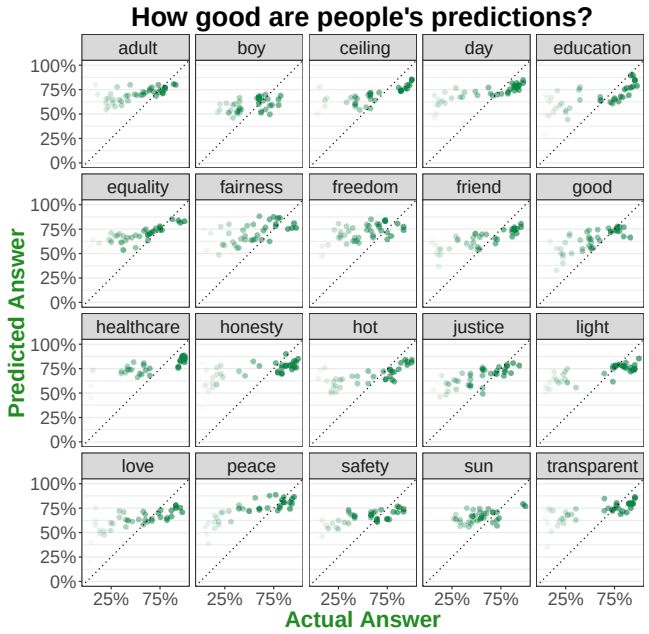


Figure 7: Participants' actual responses vs. the responses they expected others' to give, binned by word. Each data point represents the mean response for a trial/choice pair. People largely overestimate the degree to which others' agree with their assessments, regardless of the concept being assessed.

ipants overestimating the number of people who agree with them as the number of points above the y = x line shows. In fact, very few trials are underestimated and those which are, are only barely underestimated. In contrast, many trial predictions wildly overestimate people's actual responses. Examining the data by word (see Figure 7) shows these trends are not confined to a small subset of words but rather, are widespread.

Conclusions and Discussion

The degree to which conceptual representations are shared and the degree to which people are aware of any differences are also fundamentally important aspects of any theory of conceptual structure, but both have been largely neglected.

These results, along with prior literature, provide strong evidence that the diversity in conceptual representations has been underestimated. As Figure 2 shows, concepts have roughly five to seven different representations, even for basic words such as "day" or "night". This is a surprising finding from multiple points of view. If you believe everyone holds the same concept for the same word, anything greater than one will be unexpected. On the other hand, if you believe concepts are infinitely distinct across individuals and across time, our estimate will also be unexpected.

Furthermore, individuals seem to be unaware of these differences. Figure 6 illustrates the poor relationship between people's actual answers and people's guesses about the answers of others. Taken together, these findings have strong implications for the way humans communicate. Misunderstandings are likely to occur if two individuals are operating with different representations of the same word.

Limitations

It is possible that participants may have interpreted some sentences differently. If two participants have completely different interpretations of the same sentence, they may in reality possess the same concept, but appear to possess different concepts. We do not believe that this possibility could have driven our reported effects, however, because sentences were constructed in order to reduce ambiguity (though, of course, eliminating all ambiguity is impossible).

Summary

There is measurable variability in the conceptual representations attached to particular words (greater than zero but less than infinity); importantly, this variability applies to both concrete words (e.g., "sun") and abstract ones (e.g., "freedom"). More importantly, our data shows that individuals are poorly calibrated to this variability and generally underestimate it. This is important, because communication requires that interlocutors understand one another. These results could help explain a previously unappreciated source of miscommunication and misunderstanding between people.

Acknowledgements

We would like to thank members of the Kidd Lab, and the Computation and Language Lab for providing valuable feedback.

References

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological review*, *98*(3), 409.

Bunge, J., & Fitzpatrick, M. (1993). Estimating the number of species: a review. *Journal of the American Statistical Association*, *88*(421), 364–373.

Labov, W. (1973). The boundaries of words and their meanings. *New ways of analyzing variation in English*.

McCloskey, M. E., & Glucksberg, S. (1978). Natural categories: Well defined or fuzzy sets? *Memory & Cognition*, *6*(4), 462–472.

Do learners' word order preferences reflect hierarchical language structure?

Alexander Martin (alxndr.martin@gmail.com)
Centre for Language Evolution, University of Edinburgh
3 Charles St, Edinburgh EH8 9AD, UK

Klaus Abels (k.abels@ucl.ac.uk)
Division of Psychology and Language Sciences, University College London
Chandler House, 2 Wakefield Street, London WC1N 1PF, UK

David Adger (david.j.adger@qmul.ac.uk)
School of Language, Linguistics and Film, Queen Mary University of London
Mile End Road, London E1 4NS, UK

Jennifer Culbertson (jennifer.culbertson@ed.ac.uk)
Centre for Language Evolution, University of Edinburgh
3 Charles St, Edinburgh EH8 9AD, UK

Abstract

Previous research has argued that learners infer word order patterns when learning a new language based on knowledge about underlying structure, rather than linear order (Culbertson & Adger, 2014). Specifically, learners prefer typologically common noun phrase word order patterns that transparently reflect how elements like nouns, adjectives, numerals, and demonstratives combine hierarchically. We test whether this result still holds after removing a potentially confounding strategy present in the original study design. We find that when learners are taught a naturalistic "foreign" language, a clear preference for noun phrase word order is replicated but for a subset of modifier types originally tested. Specifically, participants preferred noun phrases with the order N-Adj-Dem (as in "mug red this") over the order N-Dem-Adj (as in "mug this red"). However, they showed no preference between orders N-Adj-Num (as in "mugs red two") and N-Num-Adj (as in "mugs two red"). We interpret this sensitivity as potentially reflecting an asymmetry among modifier types in the underlying hierarchical structure.

Keywords: language; learning; syntax; typology

Introduction

A large body of work has claimed that sensitivity to abstract hierarchical structure drives the acquisition of syntax (e.g., Chomsky, 1965). At the same time, there is evidence to suggest that language learners track surface-level statistics, including co-occurrence patterns among words (e.g., Saffran, Aslin, & Newport, 1996). In a recent paper, Culbertson and Adger (2014) used a pseudo-artificial language learning task to argue that learners privilege abstract structural relations among words to linear order when they learn syntactic features of a new language. Moreover, they suggest that sensitivity to these structural relations—which in their case pertain to noun phrase word order—can explain a well-studied typological generalisation, known to linguists as Universal 20 (Greenberg, 1963). In the current paper, we highlight some potential methodological issues with the paradigm used by Culbertson and Adger (2014), and test whether their finding is replicated once the paradigm is improved.

Research in generative syntax posits an underlying hierarchical structure for the noun phrase: [Dem [Num [Adj [N]][1] (Adger, 2003; Cinque, 2005; Abels & Neeleman, 2012). In this hierarchy, which can be interpreted as reflecting semantic or conceptual structure, the adjective forms a constituent with the noun to the exclusion of the numeral and demonstrative; that sub-constituent combines with a numeral, and the resulting unit combines with a demonstrative to make a larger constituent. The structure provides a straightforward explanation for why, in most languages, adjectives are placed linearly closest to the noun, while demonstratives are furthest away (e.g., Dryer, 2018). For example, in English *these two red cars*, in Thai (the equivalent of) *cars red two these*. Both these orders can be read directly off the underlying structure, while others, like N-Dem-Num-Adj cannot. While such orders can in principle be derived by movement, they are rarely found. Culbertson and Adger (2014) refer to orders like Dem-Num-Adj-N and N-Adj-Num-Dem (as well as any other order that can be read directly off of the structure [Dem [Num [Adj [N]]) as *isomorphic*—they preserve an isomorphic relation between the proposed underlying hierarchical structure and the surface linearisation.

Culbertson and Adger (2014) sought to provide evidence that learners are sensitive to this underlying structure, and use it to infer word order, rather than simply copying the linear order in their native language. To show this, they taught English speakers simple noun phrases in a pseudo-artificial language, with English words, but non-native-like word order. Participants saw an English phrase like *red shoe*, and were taught it would be *shoe red* in the new "language"; similarly *this car* would be *car this*. Participants were subsequently shown phrases with multiple modifiers, like *this red car*, and asked to guess the relative order of post-nominal modifiers in the language. The authors reason that if learners' inferences are guided by their knowledge of surface-level features of English, they should guess the non-isomorphic order (i.e., *car*

[1] Abbreviations: N(oun) (e.g., *car*), Adj(ective) (e.g., *red*), (Num)eral (e.g., *two*), Dem(onstrative) (e.g., *this*).

this red), which has its modifiers in English order. By contrast, if their inferences are guided instead by knowledge of the abstract structure described above, then they should infer the isomorphic order (i.e., *car red this*). Participants in their experiment overwhelmingly inferred isomorphic orders, suggesting sensitivity to the hypothesised universal structure rather than surface statistics of English.

While this result is intriguing, the paradigm used by Culbertson and Adger (2014) is unusual in several respects. First, even relative to other work using artificial language learning paradigms, this task is very non-naturalistic. Second, the task may encourage a particular strategy. Specifically, English phrases along with their "translations" in the language–also English words–were presented visually. Participants may have adopted an explicit strategy of reversing or "flipping" the English words to determine their responses. For example, during training participants could relate a translation like *shoe red* to the English phrase *red shoe* shown onscreen by reversing the words. Using the same strategy to guess the correct two-modifier phrase translation would then mean flipping the English *this red shoe* to *shoe red this*. Here, we aim to determine whether the apparent bias for isomorphic orders reported by Culbertson and Adger (2014) is replicated using a standard artificial language learning task, with a more naturalistic, completely novel language.

Experiment 1

The experiments we report on in the present paper are part of a larger cross-linguistic comparison project. We followed the methodology reported by White et al. (2018) and designed artificial languages using only sounds contained in all of the languages we plan to test. The phonological inventory of our artificial languages was thus reduced to five vowels, and a small set of voiceless (non-aspirated) stops, nasals, and the voiceless glottal fricative, all shared by the languages we plan to test in.[2] The languages all have lexical tone (for planned experiments with speakers of tonal languages), though the tones do not serve to contrast words from one another (thus the English-speaking participants can simply ignore them). As in Culbertson and Adger (2014), we taught participants phrases with a noun and a single modifier (either and adjective and a demonstrative, or an adjective and a numeral), and then asked them to guess the relative order of modifiers when both were present. Crucially, in contrast with Culbertson and Adger (2014), we used completely novel stimuli and did not present written L1 equivalents of the phrases participants were learning. This was done to reduce the possibility, present in Culbertson and Adger (2014), that participants would simply "flip" L1 word orders to translate into the artificial language they were learning.

Methods

Stimuli The artificial language had five lexical items. There was a single noun meaning *feather*, represented by the label /jè/. There were two adjectives (meaning *red* and *black*), and two items that served as either demonstratives (*this* and *that*) or numerals (*two* and *three*) depending on the condition the participant was assigned to. Labels for these modifier classes were created in pairs:/púkù/, /tàká/ and /hìmí/, /hónò/. The two pairs of stimuli were randomly assigned to be either adjectives or demonstratives/numerals. We privileged within-pair similarity (so /púkù/ and /tàká/ both contain only voiceless stops for example) to facilitate the learning process.[3] Stimuli were produced by a trained phonetician. All stops were produced with near zero VOT and each syllable was produced with either a high or a low tone.

Visual stimuli were pictures of simple cartoon scenes. Objects (always feathers) were depicted on a table behind which stood a cartoon girl. In trials featuring the noun alone, or the noun with an adjective and/or numeral, the girl was simply shown behind the table. In trials featuring a demonstrative, the girl was shown pointing to an object or objects (either near to her, or on the other side of the table from her). The presence of the girl and table on all trials was meant to keep demonstrative trials from being more visually salient (or complex). When no adjectival meaning was expressed, feathers were drawn in light grey; feathers were only coloured in (in red or black) on trials involving adjectives. Examples of the visual stimuli for single modifier trials are shown in fig. 1.

Procedure Participants were instructed that they would be learning part of a new language called Nápíjò, spoken by around 10,000 people in a rural region of Southeast Asia. All words and phrases were presented both auditorily and orthographically. The experimental session lasted about 15 minutes, and was divided into (1) noun training, (2) noun-modifier training, (3) noun-modifier testing, and finally, (4) extrapolation to two modifiers. Participants were first trained on the (single) noun in the language. On each trial, participants saw the object and were given its label in Nápíjò. They were instructed to click on the image to move on to the next trial. There were five such trials. They were then trained on noun-modifier combinations. Each trial had two parts. First, two images appeared, each illustrating one of the two modifiers for a given modifier type (e.g., "black" and "red", or "this" and "that"). A description of the first picture was provided, while the second picture was greyed out. Then, a description of the second picture was provided while the first was greyed out. Recall was tested immediately following this: The two pictures appeared again (in random order), and the description for one was given. Participants were instructed to click the picture matching the description. The first eight such trials were blocked by modifier type, with random choice of which modifier type was introduced first (two trials per modifier), followed by a further 16 trials with

[2]Experiment 3 contains some additional fricatives that will not be used with non-English-speaking populations.

[3]We designed the language to encourage participants to perceive it as a real "foreign" language. Therefore, while the words do not overtly contradict English phonotactics, they are not particularly English-like. This makes them difficult to learn. Piloting suggested that keeping the vocabulary relatively small would be necessary.

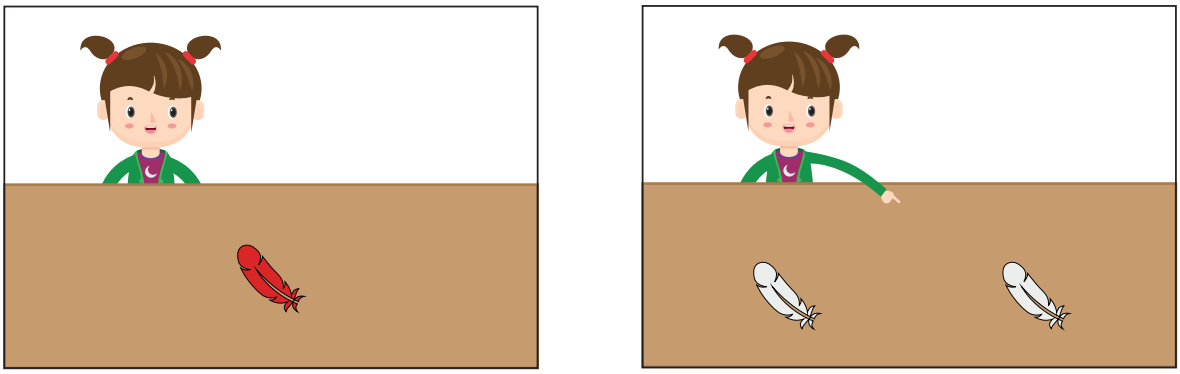


Figure 1: Single modifier trial visual stimuli examples. On the left, an example of an adjective trial, meaning "red feather", and on the right an example of a demonstrative trial, meaning "that feather".

both types intermixed. Feedback was given after each trial (image background turned green or red, plus a beep sound if incorrect). Participants were then tested on their knowledge of the noun-modifier combinations. On each trial, a picture appeared, with two potential descriptions below it. Participants were told to click on the matching description (16 trials total, four for each modifier, random order). The foil description always included a modifier of the same type. Feedback was given on each trial (button colour turned green or red, the correct description played, regardless of response).

In the critical testing phase, participants were tested (without training) on phrases with a noun and *two modifiers*. On each trial a picture appeared, with two potential descriptions below it. Participants were told to click on the matching description (16 trials total, four for each modifier, random order). The two descriptions always included the correct lexical items, in post-nominal order. They differed only in whether the order was isomorphic (e.g., N-Adj-Dem) or not (e.g., N-Dem-Adj). No feedback was given.

Participants All participants were recruited through Amazon's Mechanical Turk online recruiting platform and received 3.50 USD as compensation. We recruited a total of 70 participants who were randomly assigned to either the Demonstratives or Numerals condition. A total of eight participants were excluded (four in each condition) because they failed to reach at least 85% accuracy in the single modifier test trials (this is the same exclusion criterion reported by Culbertson and Adger (2014)). We thus analysed data from 35 participants in the Demonstratives condition and 27 in the Numerals condition.

Results

Following the analyses reported in Culbertson and Adger (2014), we analysed, for each condition, whether participants demonstrated an average preference for isomorphic orders on two modifier trials. Results from Experiment 1 are presented on the lefthand side of fig. 2. All analyses were performed by implementing logistical mixed-effects models in the `lme4` package in R (Bates, 2014). We designed full models with the binary dependent variable Isomorphic along with by-participant random effects. We used likelihood ratio tests to compare these models to null models with no intercept term to see if on average participants chose isomorphic orders above chance level. We found no isomorphic preference in either the Demonstratives ($\chi^2(1) < 1$) or the Numerals conditions ($\chi^2(1) < 1$).

Discussion

Contrary to Culbertson and Adger (2014), we did not observe any preference for isomorphic order in our artificial language learning task. However, given that our methodology differed in a number or respects from the original studies (and replications), we considered possible explanations for our null result. First, Culbertson and Adger (2014) used English words in their experiment, whereas we used nonce words. It is therefore worth verifying that participants in our experiment interpreted the words as intended. In a debrief questionnaire, participants were asked to report the meanings of the words they had learned. Participants invariably reported correct translations for adjectives (colour words) and numerals. However, meanings given for demonstratives and nouns varied to some degree. For demonstratives, most participants reported translations such as *this* and *that*, or *here* and *there*. Both these translations are consistent with a demonstrative interpretation: although *here* and *there* are sometimes called adverbs, their meaning and syntax are similar to *this* and that, and indeed they are the demonstrative words in many languages (Diessel, 2006). However, some participants gave responses such as *left* and *right* (indeed, the absolute and relative positions were confounded in our stimuli). The variation in interpretation of the demonstrative may have weakened the results to some degree. However, the interpretation of the noun suggests a more obvious issue.

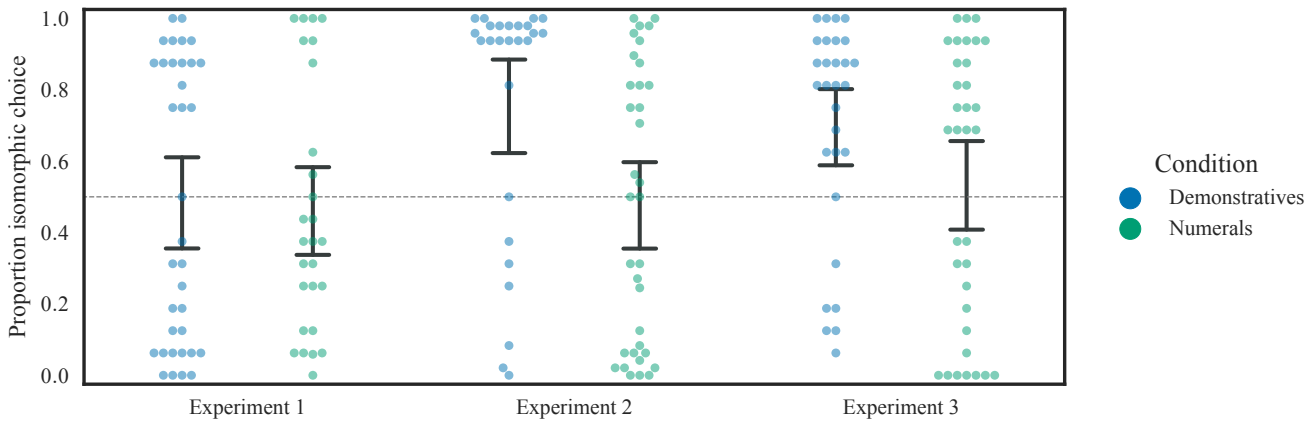


Figure 2: Proportion isomorphic preference in each experiment by condition. Each point represents an individual participant and error bars represent 95% confidence intervals.

While some participants correctly reported the English word *feather* for the word /jè/. Some, did not report a translation at all, suggesting they may not have understood its meaning. Others reported incorrect meanings, giving function words like *the* as translations. Note that the set size of the nouns differs dramatically from Culbertson and Adger (2014), where 20 nouns were used: we used a single noun in Experiment 1. Given that every trial always consisted of /jè/ + *x*, it was therefore possible for participants to completely discount that word (or indeed treat it as a determiner). This suggests the possibility that participants may not have been treating our stimuli as noun phrases (i.e., not attending to the noun head), but simply as strings of modifiers. If so, participants may have adopted any number of response strategies. For example, they could have interpreted the strings as copulative sentences (e.g., "this (one) (is) red"), or simply strings of modifiers. In either case, they would not have learned the intended Noun-modifier structure of the language. In Experiment 2, we therefore expanded the set of nouns in the artificial lexicon. We hypothesised that variability in the noun would cause participants to treat our stimuli as noun phrases, resulting in a preference for isomorphic orders.

Experiment 2

Methods

Stimuli The stimuli for Experiment 2 were similar to those in Experiment 1; only the nouns differed. We created audio and visual stimuli for three objects (*feather*, *ball*, *mug*) which were assigned the names /éjè/, /úhù/, and /ítì/, respectively. All modifier stimuli were identical to Experiment 1.

Procedure The procedure was identical to Experiment 1 except the initial training and testing phases were slightly lengthened. Noun training was composed of 15 trials (five trials for each of the three nouns). This was immediately followed by 15 trials of noun testing in which a picture appeared with two labels beneath it. Participants were instructed to click the matching label. Feedback was given (button colour turned green or red, the correct description played regardless of response). Noun-modifier exposure was composed of 12 trials blocked by modifier type (six trials per block, two for each noun-modifier combination), followed by an additional intermixed block of 12 trials (one trial for each noun-modifier combination). Noun-modifier testing was composed of 24 trials (two trials for each noun-modifier combination). The foil labels for each picture were either an incorrect noun *or* an incorrect modifier of the same type. Finally, for the critical test phase, a random set of 16 trials was constructed for each participant.

Participants As in Experiment 1, all participants were recruited through Amazon's Mechanical Turk online recruiting platform and received 3.50 USD as compensation. We recruited a total of 71 participants who had not participated in Experiment 1. Participants were randomly assigned to either the Demonstratives or Numerals condition. A total of 11 participants were excluded (seven in the Demonstratives condition and four in the Numerals condition) because they failed to reach at least 85% accuracy in the single modifier test trials. We thus analysed data from 26 participants in the Demonstratives condition and 34 in the Numerals condition.

Results

Results from Experiment 2 are presented in the middle of fig. 2. The analysis of Experiment 2 was identical to that of Experiment 1. We found an isomorphic preference in the Demonstratives condition ($\beta = 2.25$, $SE = 0.60$, $\chi^2(1) = 11.35$, $p < 0.001$) but not in the Numerals condition ($\chi^2(1) < 1$).

Discussion

The results of Experiment 2 revealed a preference for isomorphic word orders, but only if the set of modifiers learned was adjectives and demonstratives. That is, participants preferred noun phrases with the order N-Adj-Dem (as in "mug

red this") over the order N-Dem-Adj (as in "mug this red"). However, they showed no preference between orders N-Adj-Num (as in "mugs red two") and N-Num-Adj (as in "mugs two red"). Interestingly, this asymmetry has been reported numerically in all previous experiments on isomorphism. As discussed above, Culbertson and Adger (2014) found statistically significant isomorphism preferences for all pairs of modifiers (adjective, numeral and demonstrative), and when all three modifiers were present (not tested here). However, they report a numerical difference among the groups such that the isomorphism preference is strongest with adjective and demonstrative. Indeed, they cite this as further evidence that English speakers are sensitive to the underlying hierarchical structure, since adjectives and demonstratives are structurally more distant than adjectives and numerals (or numerals and demonstratives). In a lab replication of the original study (which was conducted on Mechanical Turk), A. Martin, Ratitamkul, Abels, Adger, and Culbertson (in press) replicated both the general isomorphism preference and the difference among modifier pairs. They also report a replication with Thai speakers, whose L1 order is N-Adj-Num-Dem. These speakers were trained on an artificial language with prenominal modifiers, and they then inferred prenominal isomorphic orders like Dem-Adj-N in the criticial two-modifier test phrase. There again the same difference among modifier pairs was present. These studies report only numerical differences. Our findings therefore present the clearest evidence yet that the isomorphism preference may be sensitive to modifier type.

Nevertheless, we did not replicate an isomorphism preference for the Numerals condition. Additionally, the isomorphism preference found for the Demonstratives condition is (numerically) weaker than reported in these previous studies. By design, we have reduced the likelihood that participants are relying on an explicit "flipping" strategy, and we have made the language itself more naturalistic. Thus, one possibility is that our results are a better representation of English speakers' underlying bias for isomorphism: it is present, but not categorical for adjectives and demonstratives, and not present for adjectives and numerals. We return to this in the general discussion. There is, however, one other major difference between our experiment and previous experiments which could plausibly weaken or mask an isomorphism preference, namely the relative size of the modifier categories. In both Culbertson and Adger (2014) and A. Martin et al. (in press), the relative class sizes approximately match what one would typically find in a natural language: largest set size for adjectives, then numerals, and a small set of demonstratives.[4] In our experiments, all modifier classes contained two elements. In Experiment 3, we test the possibility that using a more naturalistic relative size for the modifier classes might amplify the isomorphism preference, perhaps revealing the isomorphism preference between numerals and adjectives reported in previous work.

[4]For example, 694 adjective vs. 172 numeral, 5 demonstrative types among all noun phrases in the English Universal Dependencies Treebank (Nivre et al., 2017).

Experiment 3

Methods

Stimuli The stimuli for Experiment 3 were similar to those for Experiments 1 and 2. The only difference was in the number of adjectives. Specifically, four adjectives were created (/tàkás/, /pùkúf/, /kàpáθ/, and /kùtúʃ/) and mapped to four colour meanings ("black", "red", "blue", and "green", respectively). Visual stimuli similar to those in Experiments 1 and 2 were also created.

Procedure The procedure was identical to Experiment 2 except for the following: Noun-modifier training was all blocked (in order to balance frequency of exposure to each combination without increasing the number of trials too much). Each block was composed of 12 trials. In the adjective block, each adjective was shown once with each noun. In the numeral or demonstrative block, each modifier was shown twice with each noun. The noun-modifier testing block was slightly longer than in Experiment 2, with 36 trials total (2 trials for each noun-modifier combination). No changes were made to the critical two modifier testing phase (again, 16 trials total, randomly constructed). Note that the frequency of exposure to each modifier class was the same, only the number of elements in each class differed.

Participants As in Experiments 1 and 2, all participants were recruited through Amazon's Mechanical Turk online recruiting platform and received 3.50 USD as compensation. We recruited a total of 76 participants who had not participated in Experiment 1 or Experiment 2. Participants were randomly assigned to either the Demonstratives or Numerals condition. A total of 13 participants were excluded (nine in the Demonstratives condition and four in the Numerals condition) because they failed to reach at least 85% accuracy in the single modifier test trials. We thus analysed data from 29 participants in the Demonstratives condition and 34 in the Numerals condition.

Results

Results from Experiment 3 are presented on the right-hand side of fig. 2. The analysis of Experiment 3 was identical to that of Experiments 1 and 2. As in Experiment 2, we found an isomorphic preference in the Demonstratives condition ($\beta = 1.24$, $SE = 0.36$, $\chi^2(1) = 10.37$, $p < 0.01$) but not in the Numerals condition ($\chi^2(1) < 1$).

Discussion

In Experiment 3, we tested whether the isomorphism preference found in Experiment 2 would be amplified, and extended to the Numerals condition if the relative sizes of the modifier classes were more naturalistic. This was not borne out; rather we replicated the findings of Experiment 2: an isomorphism preference for noun phrases with a demonstrative

and an adjective, but not for noun phrases with a numeral and an adjective. This finding therefore reinforces the asymmetry reported in Experiment 2, and the numerical patterns reported in both Culbertson and Adger (2014) and A. Martin et al. (in press). In the next section, we investigate statistically the general pattern of results across experiments described here.

Comparison across experiments

Two manipulations distinguished Experiments 1, 2, and 3. First, the size of the noun class. In Experiment 1, participants learned only one noun, while in Experiments 2 and 3 they learned three. Second, the size of the adjective class. In Experiments 1 and 2, participants learned only two adjectives, while in Experiment 3 they learned four. We thus performed an analysis considering these two binary variables, included in our models using contrast coding. This allowed us to explore the interaction between these two factors and the factor Condition in one single statistical model. The model predicted Isomorphic order choice from three fixed binary factors: Condition (Demonstratives or Numerals), Noun Class Size (one noun or three), and Adjective Class Size (two adjectives or four). We also included interactions between Condition and Noun Class Size and between Condition and Adjective Class Size as well as by-participant random intercepts. We then designed reduced models each excluding one factor or interaction, and compared them to the full model (again using likelihood ratio tests).

We found that removing Noun Class Size significantly worsened the model fit ($\beta = 1.08$, $SE = 0.49$, $\chi^2(1) = 4.70$, $p < 0.05$). This indicates that participants who learned an artificial language with three nouns showed a stronger isomorphism preference than those who learned an artificial language with only one noun. We also found that removing the interaction between Condition and Noun Class Size significantly worsened the model fit ($\beta = -2.65$, $SE = 0.98$, $\chi^2(1) = 7.10$, $p < 0.01$). This confirms our observation that amongst the participants who learned artificial languages with three nouns, those in the Demonstratives conditions showed an isomorphism preference while those in the Numerals conditions did not. Removing the factors Adjective Class Size ($\chi^2 < 1$) and Condition ($\chi^2 = 1.46$, $p = 0.23$) did not worsen the model fit, nor did removing the interaction between Condition and Adjective Class Size ($\chi^2 = 1.40$, $p = 0.24$).

General discussion

This paper aimed to test the preferences of English speakers learning about the noun phrase word order of a new language. Previous research using a pseudo-artificial language learning paradigm reported a strong preference for so-called isomorphic noun phrase orders, like N-Adj-Dem or N-Adj-Num, which transparently reflect the hypothesised hierarchical structure of the noun phrase: [Dem [Num [Adj [N]]]] (Culbertson & Adger, 2014; A. Martin et al., in press). This has been claimed to show that speakers' inferences about a new language are not based on the surface linear order of their native language, but on a (potentially universal) underlying hierarchical structure. Moreover, the results suggest the possibility that a preference for orders which are isomorphic to this structure might explain why these orders overwhelmingly outnumber non-isomorphic orders in the typology (Cinque, 2005; Abels & Neeleman, 2012; Dryer, 2018).

We sought to replicate these findings using an improved methodology, designed to address the possibility that the original results reflected the availability of an explicit strategy which may have encouraged participants to choose isomorphic orders by visually flipping the English words. We used a standard artificial language learning paradigm, with a relatively more naturalistic language. In Experiment 1, we used a minimal vocabulary, with only a single noun, and found no isomorphism preference. In Experiment 2, we added additional nouns to encourage participants to treat stimuli as noun phrases. Here, we found an isomorphism preference for phrases including a demonstrative and an adjective, but not for phrases including a numeral and an adjective, an asymmetry which mirrors numerical differences reported in earlier studies. In Experiment 3, we attempted to strengthen the isomorphism preference by making the number of words in each modifier category more naturalistic (in terms of relative size). This did not change the results, but rather again revealed that learners' isomorphism preference was sensitive to the modifier categories involved.

Importantly, our results show that in a more naturalistic artificial language learning task, where participants are unlikely to use an explicit strategy of flipping English words to determine order in the new language, an isomorphism preference is still found. Some confirmation that participants are not using a simple flipping strategy in our experiments comes from self-reports given at the end of the task. Of the 185 participants that were retained for data analysis in our three experiments, only one referred to a flipping strategy in the debriefing questionnaire. Instead, common strategies included "no strategy", "I just went with my gut feeling" (67 such reports), or simple descriptions of their order choices like "I placed colour words closer to the object name, then numbers" (50 such reports). This contrasts starkly with the strategies reported by participants in Culbertson and Adger (2014)'s study. We recovered the data from that study and analysed the 89 participant strategy reports from their Experiment 1: 47 of them reported some kind of explicit flipping-based strategy (compared to only 11 "no strategy"). Our replication of their effect with a more naturalistic artificial language is thus an important contribution to this line of research.

Our results also highlight the persistent difference between modifier types, found numerically in earlier experiments, and confirmed statistically here. While it is possible that something about our task is still masking a (weaker but present) isomorphism preference for numerals and adjectives, there is some reason to suspect that the *asymmetry* at least is real. In fact, using the data collated by Dryer (2018), we can observe that non-isomorphism between numerals and adjectives, or numerals and demonstratives is more common cross-

linguistically (35 and 64 languages respectively) than non-isomorphism between adjectives and demonstratives (27 languages). This may reflect the fact that adjectives and demonstratives are more distant from one another in terms of underlying hierarchical structure.

As mentioned in the introduction, this hierarchy can be conceived of as reflecting semantic composition, or conceptual structure. Indeed linear order patterns more generally have been argued to reflect both (Rijkhoff, 1990; Baker, 1985; Bybee, 1985; Rice, 2000). One possibility is that the underlying hierarchy of nominal modifiers reflects differences in conceptual closeness (or inherentness) between particular modifier types and nouns (Kirby, Culbertson, & Schouwstra, 2018; Culbertson, Schouwstra, & Kirby, under revision). Under this account, adjectives are conceptually closest to nouns because they are more likely to reflect inherent properties of individual nouns (e.g., colour, size, texture, etc). Numerals are typically less closely linked with particular nouns (though some clearly are, e.g., four seasons, seven days of the week). Demonstratives, being deictic elements, are by their nature not associated with particular nouns. If the underlying hierarchical structure reflects these different conceptual relations between elements, then a preference for isomorphism is a preference to hierarchically cluster elements that are more closely related conceptually. Perturbing this preference would then be less costly when it involves elements that differ less in their conceptual closeness to the noun (e.g., Adj and Num), compared to elements that differ quite a lot (e.g., Adj and Dem) (for similar arguments about the relative order of adjectives, see J. E. Martin, 1969; Bouchard, 2002).

To summarise, the experiments reported here aimed to replicate the preference for isomorphic ordering in the noun phrase, first reported in Culbertson and Adger (2014). Using a more naturalistic artificial language learning task, we find that English speakers infer isomorphic orders of demonstrative and adjective. However, we found no evidence of an isomorphism preference for numerals and adjectives. Above we suggest one possible explanation for the difference between these two conditions: assuming that English speakers can either use an isomorphic order, or an order that reflects the surface linear order of their language, they are more likely to go with the latter when this would involve two modifiers that are more similar to each other, either in terms of structural distance, or in terms of conceptual closeness with the noun. That said, learners' sensitivity to the distribution features of the language (e.g., in Experiment 1) leave open the possibility that future experiments will reveal this bias as weaker but still present.

Acknowledgements

This work was supported by the Economic and Social Research Council [grant number ES/N018389/1].

References

Abels, K., & Neeleman, A. (2012). Linear asymmetries and the LCA. *Syntax*, *15*(1), 25-74.

Adger, D. (2003). *Core syntax*. Oxford: OUP.

Baker, M. (1985). The mirror principle and morphosyntactic explanation. *Linguistic Inquiry*, *16*(3), 373–415.

Bates, D. M. (2014). *Lme4: Mixed-effects modeling with R.*

Bouchard, D. (2002). *Adjectives, number and interfaces: Why languages vary*. Amsterdam: Elsevier.

Bybee, J. L. (1985). *Morphology: A study of the relation between meaning and form.* Philadelphia, PA: John Benjamins.

Chomsky, N. (1965). *Aspects of the theory of syntax*. MIT Press.

Cinque, G. (2005). Deriving Greenberg's Universal 20 and Its Exceptions. *Linguistic Inquiry*, *36*(3), 315-332.

Culbertson, J., & Adger, D. (2014). Language learners privilege structured meaning over surface frequency. *PNAS*, *111*(16), 5842-5847.

Culbertson, J., Schouwstra, M., & Kirby, S. (under revision). *From the world to word order: the link between conceptual structure and language.*

Diessel, H. (2006). Demonstratives, joint attention, and the emergence of grammar. *Cognitive linguistics*, *17*(4), 463–489.

Dryer, M. S. (2018). On the Order of Demonstrative, Numeral, Adjective and Noun. *Language*.

Greenberg, J. H. (1963). *Universals of Language*. MIT Press.

Kirby, S., Culbertson, J., & Schouwstra, M. (2018). The origins of word order universals: evidence from corpus statistics and silent gesture. In C. Cuskley, M. Flaherty, H. Little, L. McCrohon, A. Ravignani, & T. Verhoef (Eds.), *The evolution of language: Proceedings of the 12th international conference (evolangxii).* NCU Press.

Martin, A., Ratitamkul, T., Abels, K., Adger, D., & Culbertson, J. (in press). Cross-lingusitic evidence for cognitive universals in the noun phrase. *Linguistics Vanguard*.

Martin, J. E. (1969). Semantic determinants of preferred adjective order. *Journal of Verbal Learning and Verbal Behavior*, *8*(6), 697–704.

Nivre, J., Agić, , Ahrenberg, L., Antonsen, L., Aranzabe, M. J., Asahara, M., ... Zhu, H. (2017). *Universal Dependencies 2.1.* LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Rice, K. (2000). *Morpheme order and semantic scope: Word formation in the Athapaskan verb*. Cambridge: Cambridge University Press.

Rijkhoff, J. (1990). Explaining word order in the Noun Phrase. *Linguistics*, *28*(1), 5–42.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928.

White, J., Kager, R., Linzen, T., Markopoulos, G., Martin, A., Nevins, A., ... van de Vijver, R. (2018). Preference for locality is affected by the prefix/suffix asymmetry: Evidence from artificial language learning. In *Proceedings of NELS 48.*

The Cognitive Underpinnings of Inductive Grammar Learning

David Martinez (dmartin5@umd.edu)
Alison Tseng (atseng1@umd.edu)
Valarie Karuzis (vkaruzis@umd.edu)
Meredith Mislevy-Hughes (mmislevy@umd.edu)
Nick B. Pandža (npandza@umd.edu)
Gregory J. H. Colflesh (colflesh@gmail.com)
Polly O'Rourke (porourke@umd.edu)
University of Maryland, College Park, 7005 52nd Ave, College Park, MD 20742

Abstract

The acquisition of the grammar of a second language requires a variety of cognitive mechanisms, including inductive reasoning. In the current study, we examine the cognitive underpinnings of grammar learning with an explicit-inductive (rule search) learning task, designed to capture more of the complexity associated with grammar learning than purely deductive tasks. Research in language aptitude has shown that working memory capacity (WMC) is a key predictor of grammar learning outcomes. Inductive reasoning and grammatical sensitivity are other established aptitude factors. The goal of the present study was to determine the degree to which relevant variables predict learning on an explicit-inductive grammar learning task. Our results indicate that both WMC and inductive reasoning ability predict learning over three days of grammar training.

Keywords: L2 learning; L2 aptitude; working memory capacity; inductive reasoning; individual differences

Introduction

The acquisition of second language (L2) grammar is extremely challenging for adult learners. One of the reasons for this difficulty is the heavy and diverse processing demands associated with learning grammatical rules as well as applying them during comprehension and production of L2 utterances (Doughty & Long, 2003). Insights into the precise cognitive underpinnings of grammar learning come from the field of language aptitude. Importantly, working memory capacity (WMC) has emerged as a key predictor of L2 grammar learning ability (Linck, Osthus, Koeth & Bunting, 2014; Miyake & Friedman, 1998). WMC is defined as the ability to maintain attention on a limited amount of information, even in the face of interference (Engle, 2002, 2018), and underpins many aspects of higher cognition and goal-directed behavior. Another predictor is inductive reasoning ability, the ability to extrapolate rules and patterns from specific examples. While both WMC and inductive reasoning are predictors of grammar learning outcomes, there is a lack of research examining whether the two account for independent portions of variance in learning. In the current study, we examined the cognitive underpinnings of grammar learning using an explicit-inductive learning task. In this task, participants were presented with L2 phrases and asked to figure out the grammatical rules, and then tested on those rules. The goal was to examine the degree to which relevant variables predict grammar learning.

Explicit-Inductive Grammar Learning

In explicit-inductive (or "rule-search") grammar learning tasks, learners are presented with a number of L2 examples (sentences or phrases) exhibiting target grammatical structures in both a foreign language and the individuals' native language and are asked to figure out the rule(s) for subsequent testing. These tasks differ from deductive tasks in which rules are explicitly taught (DeKeyser, 2003). They also differ from artificial grammar learning tasks (also referred to as statistical learning tasks) in that, in artificial grammar learning tasks, rules are acquired without conscious awareness (i.e., implicitly) and there is no meaning ascribed to the material under study (Misyak & Christiansen, 2012). Though, it is very likely that in providing numerous exemplars in explicit-inductive grammar learning tasks, individuals not only infer rules but likely implicitly acquire statistical regularities as well. Thus, explicit-inductive grammar learning tasks likely involve *both* explicit and implicit learning processes (DeKeyser, 1995). Given that both types of learning are known to be involved in grammar acquisition (DeKeyser, 2003), these tasks may better capture the cognitive complexity of grammar learning.

Working Memory Capacity

Individual differences in WMC are strongly predictive of performance on a range of tasks assessing cognitive abilities and processes (Engle, Tuholski, Laughlin, & Conway, 1999; Kyllonen & Christal, 1990) including L1 processing (Daneman & Merikle, 1996) and L2 learning (Linck et al., 2014). Indeed, in a meta-analysis synthesizing the results of 79 studies with a combined sample size of over 3,000 participants, Linck et al. (2014) found that WMC tasks are positively associated with L2 outcomes. Moreover, Tagarelli, Borges-Mota and Rebuschat (2011) found that WMC predicted performance on an explicit-inductive grammar learning task.

Relations observed between WMC and other cognitive tasks are typically explained as owing to the fact that complex cognition requires sustained attention on the task at hand, often while performing various operations, which themselves produce interference (Daneman & Carpenter, 1980;

Daneman & Merikle, 1996). This emphasis on controlled attention has led some to theorize that WMC plays a greater role in L2 learning under explicit, rather than implicit, learning conditions (e.g., Tagarelli, Mota, & Rebuschat, 2011). Indeed, Tagarelli, Mota, and Rebuschat (2011) found that WMC was predictive of learning under explicit, but not implicit, learning conditions.

Inductive Reasoning

Another predictor of L2 learning that figures prominently is inductive reasoning (Gardner & Lambert, 1965; Sparks, Humbach, Patton, & Ganschow, 2011). In inductive reasoning, one infers general principles from specific observations. For example, an adult interested in learning another language for use during a trip may begin by learning "survival phrases" such as "I am American" and "I am sorry." In doing so, one may *infer* grammatical rules and the meaning of certain words in the L2, which can then be used to construct new words and sentences (though of course the accuracy of the constructions will be dependent on the premises, e.g., not all verbs in English can be changed from present to past tense by affixing an *-ed*). Like WMC, inductive reasoning ability has been found to predict grammar learning under explicit, but not, implicit conditions (Gebauer & Mackintosh, 2007).

Relationship between WMC and Inductive Reasoning

An issue arises, however, when one notes that WMC is highly correlated with inductive reasoning ability (Engle et al., 1999; Kyllonen & Christal, 1990). In the individual differences literature, inductive reasoning tasks are often used as indicators of fluid intelligence (Marshalek, Lohman, & Snow, 1983; Wilhelm, 2005). According to a recently proposed theory (see Shipstead, Harrison, & Engle, 2016), WMC and fluid intelligence/inductive reasoning are highly correlated because both rely on attention control; however, while WMC tasks primarily assess the ability to *maintain* attention, fluid intelligence/inductive reasoning tasks additionally assess the ability to *disengage* attention. Consider that in WMC tasks, the goal is to maintain to-be-remembered information (e.g., sets of letters) in mind exactly as they were presented; in inferential tasks, however, the goal is to produce a novel solution, entailing some kind of transformation or restructuring of inputs as multiple solutions or hypotheses are investigated (Oberauer, Süß, Wilhelm, & Sander, 2007). During the reasoning process, one has to maintain relevant pieces of information in mind, implicating WMC, but at other times, one has to abandon an incorrect solution and begin anew, requiring one to disengage attention from one problem representation for another.

The issue is that there is little research investigating whether WMC and inductive reasoning ability *independently* account for variance in L2 learning. To investigate this issue, we included a measure of inductive reasoning and a measure of WMC as predictors in the present study. Given that the outcome variable is an explicit-inductive grammar learning task, we expect inductive reasoning to be predictive of learning, however, a WMC task should account for variance over and above an inductive task, as WMC is a well-established predictor of language learning (Linck, Osthus, Koeth & Bunting, 2014; Miyake & Friedman, 1998).

Grammatical Sensitivity

In addition to WMC and inductive reasoning, a measure of grammatical sensitivity was also included in this study—the Words in Sentences (WIS) subtest from the Modern Language Aptitude Test (MLAT), developed by Carroll and Sapon (1959). According to Carroll (1964), grammatical sensitivity is the "ability to recognize the grammatical functions of words in sentences (p. 95)". Studies have shown the WIS to be a predictor of L2 learning (Li, 2015), particularly under explicit learning conditions (Li, 2014); however, there is also research and theorizing that grammatical sensitivity depends on inductive reasoning (Li, 2015; Sasaki, 1993). Thus, including this measure as a predictor allows us to investigate whether grammatical sensitivity influences novel grammar learning over and above inductive reasoning and working memory.

The Present Study

With the above in mind, this study was undertaken to assess the relative contributions of WMC, inductive reasoning, and grammatical sensitivity on one aspect of L2 learning, grammar learning. For this study, we developed an explicit-inductive (i.e., rule-search) grammar learning task in which individuals were tasked with learning syntactic rules in an L2. During learning, participants were exposed to a number of L2 phrases and their English translations and attempted to infer rules for arranging words in the L2. Superficially, the task is similar to what was described earlier when one learns "survival" phrases and induces rules, however, (and as will be clarified below) this task obviates the need to memorize phrases and thus should be a relatively *pure* measure of grammatical (rule) induction.

Method

Participants

A total of 34 individuals participated in the study; however, three did not complete the entire study, leaving 31 with complete data. All participants were recruited from the university and surrounding community and were compensated for their time. No participant reported experience with Indonesian or related languages.

Procedure

All participants completed a total of three sessions in a room with up to six other participants. Each session contained a grammar learning task. In addition to the grammar learning task, in Session 1 participants also completed a demographics and language history questionnaire, administered before the grammar learning task; in Session 2 participants completed

Letter Sets, an Antisaccade task, and a Speeded Lexical Decision task, administered, in that order, after the grammar learning task; and in Session 3 participants completed the Remember and Count task, another Speeded Lexical Decision task, and the Words in Sentences, administered, in that order, after the grammar learning task. Each session took approximately 60 minutes to complete. All tasks were completed on desktop computers. Below, we offer descriptions of the tasks included in this study.

Instruments

Explicit-Inductive Grammar Learning Task We chose an explicit-inductive task as our criterion measure because research indicates that during the early stages of L2 learning, adults tend to engage control processes to learn grammatical rules, while at later stages, they tend to rely on implicit learning processes (Hamrick, Lum, & Ullman, 2018). Thus an explicit-inductive task likely captures processes similar to those engaged throughout the learning process.

The grammar learning task consisted of three phases: learning, recall, and recognition. During the learning phase, participants were presented with example Indonesian phrases and their English translations, ordered from short/simple phrases to long/complex phrases. Participants therefore had to infer the "simple" rules and then take mental note of how these "simple" rules combined to construct complex phrases. Because participants did not know Indonesian, the Indonesian words and their English translations were color-coded, such that translation equivalents were the same color; Indonesian words without direct translations (e.g., classifiers) were presented in black font with no background color. See Figure 1.

Figure 1. Three example grammar learning items.

During the recall phase, participants were asked to translate English phrases into Indonesian by selecting words from a word bank and placing them in the correct sequence. The word bank included the words needed for the translation as well as all function words. Where possible, English translations were provided (see Figure 2). During the recognition phase, participants saw Indonesian noun phrases and indicated whether they were grammatical or not (see Figure 3).

Participants were *never* given feedback in either the recall or recognition phases, but were given a score of their overall recognition phase performance at the end of the day.

Table 1. Syntactic Structures in the Grammar Learning Task.

	Structure	**Example English Phrase**
1	Demonstrative noun	that uncle
2	Number w/classifier	two apes
3	Single adjective	moody zebra
4	Demonstrative + single adjective	that bold scientist
5	Double adjective	new, red school
6	Possessive	my rabbit
7	Possessive + single adjective	my hungry uncle
8	Number/classifier + single adjective	two small warehouses
9	Number/classifier + single adjective + possessive	my two friendly fish
10	Number/classifier + double adjective + possessive	my two new, crowded stores
11	Triple adjective	fancy, young, skilled lawyer
12	Noun + single adjective + pre-intensifier	very crowded arena
13	Noun + single adjective + post-intensifier	very skilled teacher
14	Noun + number/classifier + single adjective + pre-intensifier	two very clever bears
15	Noun + number/classifier + single adjective + post-intensifier	two very chilly cinemas
16	Noun + possessive + number/classifier + single adjective + pre-intensifier	my two very expensive palaces
17	Noun + possessive + number/classifier + single adjective + post-intensifier	my two very tired bears

Figure 2. Example grammar recall item.

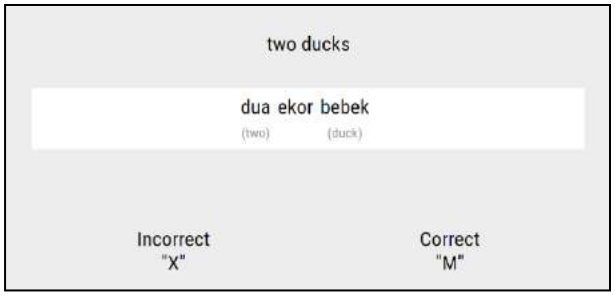


Figure 3. Example grammar recognition item. The correct answer is "M", correct.

Across sessions, the same 111 noun phrases were used in the learning phase. However, because our interest was in *grammar* learning, noun phrases used in the recall and recognition phases were never repeated across sessions, resulting in 123 noun phrases for the recall phase (41/session) and 312 noun phrases for the recognition phase (104/session).

Words in Sentences (WIS) In the WIS, each item consisted of two or more English sentences. One word in the first sentence was printed in uppercase letters. Four or five words in the remaining sentences were underlined and were labeled with corresponding letter answer options (Figure 3). Participants indicated which of these underlined words served the same grammatical function as the word in uppercase letters in the first sentence.

Chidi said THAT Eleanor is incorrigible.

In our (A) session, that (B) lady said that (C) she knew that (D) actor on the (E)

television show.

Figure 3. A sample item from the *words in sentences* test modified from https://lltf.net/mlat-sample-items/mlat-part-iv/. The correct answer is C.

Remember and Count (RAC) WMC was assessed by the RAC task (Hughes et al., 2016; O'Rourke et al., 2017), a visuospatial complex span task. In the RAC task, participants first see a sequence of triangles of different colors presented in a sequence in different quadrants. Next, they see an image of dark and light blue circles and squares; participants are to count and report the number of dark blue circles in the image. Finally, in the critical portion of the task, they must recall the sequence of triangles by indicating the color, the location, and the order of each triangle in the sequence. The number of triangles in a sequence varied between three and five, with four trials of set size 3, nine trials at set size 4, and eight trials at set size 5, for a total of 21 trials. Each trial was scored as a proportion of correctly recalled triangles; thus, participants could achieve a maximum of 1 point per trial.

Letter Sets (LSET) Inductive reasoning was assessed by the LSET task (Doughty, Campbell, Bunting, Bowles, & Haarmann, 2007). In each item, participants are presented with five sets of four letters. Four of the sets are arranged such that they follow the same rule while one does not; participants are to determine which set of letters does not follow the same rule as the others. There were a total of 15 items.

Results

Correlations amongst the predictors and descriptive statistics are provided in Table 2. Figure 4 depicts average learning curves for both the recall and recognition grammar measures.

Table 2. Predictor Correlations and Descriptive Statistics

	WIS	LSET	RAC
WIS			
LSET	0.20		
RAC	0.48*	0.48*	
$\bar{X}$	.42	.76	.53
SD	.13	.13	.18
Skew	-.02	-.10	-.91
Kurtosis	-.54	.16	.19

Note: * p <.05; LSET = letter sets; RAC = remember and count; WIS = words in sentences.

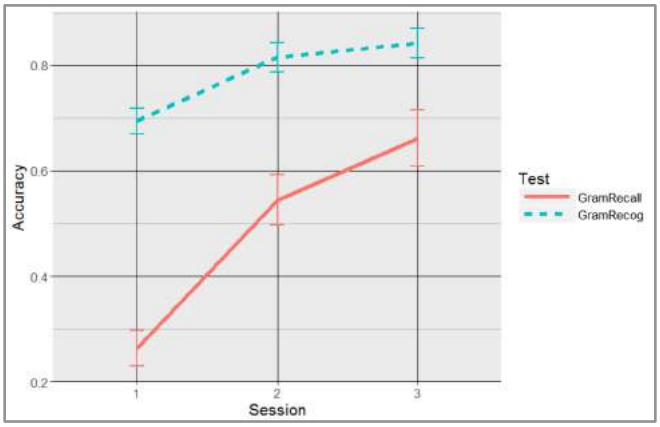


Figure 4. Average Learning Curves. Error bars: ± 1 *SE*.

Next, the recall and recognition data were each submitted to mixed-effects logistic regression modeling using a forward-testing procedure for random slopes and a backward elimination procedure for fixed effects to arrive at the model of best fit using likelihood ratio tests. This procedure allows us to find the most maximal model supported by the data, balancing Type I error and power (Matuschek et al., 2015). For each analysis, the first model included items and participants as random intercepts, session as a fixed effect, WIS, LSET, and RAC as covariates, and each Session x Covariate interaction. In each model, covariate tasks were mean centered.

Table 3 displays the final recall model. There was a significant effect of session, indicating that performance significantly improved across sessions (b = 1.67, SE = .25, z = 7.97, p < .001). There was also a significant effect of Session x RAC (b = .58, SE = .21, z = 2.74, p = .006), indicating that participants with higher RAC scores showed greater gains in recall accuracy over sessions. No other covariates or interactions were significant.

Table 3. Final Recall Model

Fixed Effects	*b*	Odds	*SE*	*p*
Intercept	-3.15	0.04	0.33	<.001*
Session	1.67	5.31	0.25	<.001*
RAC	0.06	1.06	0.16	.825
Session x RAC	0.58	1.79	0.21	.006*
Random Effects	**Var**	***SD***	**Corr**	
Intercept | Item	1.90	1.38		
Intercept | Part.	1.26	1.12		
Session | Part.	1.14	1.07	-.45	

Table 4. Final Recognition Model

Fixed Effects	*b*	Odds	*SE*	*p*	
Intercept	0.24	1.27	0.15	.116	
Session	0.89	2.43	0.10	<.001*	
WIS	0.10	1.11	0.14	.472	
LSET	-0.14	0.87	0.14	.304	
RAC	0.05	1.05	0.16	.744	
Session x LSET	0.23	1.26	0.10	.027*	
Session x RAC	0.27	1.31	0.11	.014*	
Random Effects	**Var**	***SD***	**Corr**		
Intercept | Item	.77	.88			
Session | Item	.07	.27	-.26		
LSET | Item	.04	.21	.00	.68	
WIS | Item	.09	.30	.45	-.93	-.81
Intercept | Part.	.27	.52			
Session | Part.	.20	.44	.12		

Table 4 displays the final model for the recognition analysis. Session was once again significant (b = .89, SE = .10, z = 8.83, p < .001), as was Session x LSET (b = .23, SE = .10, z = 2.21, p = .027) and Session x RAC (b = .27, SE = .11, z = 2.46, p = .014), indicating that as performance on these measures increased, individuals showed greater gains in accuracy over sessions. No other covariate or interaction was predictive.

Discussion

The primary aim of the study was to investigate the cognitive underpinnings of explicit-inductive grammar learning. In our grammar learning task, participants attempted to learn a subset of Indonesian syntax by inferring the rules of the language from a number of exemplars. Based on prior studies and theory, we chose indicators of grammatical sensitivity, WMC, and inductive reasoning as our predictors. Aware of significant relationships amongst the predictors, we were also interested in investigating whether the predictors uniquely accounted for variance in grammar learning and if so, to what degree. Our analyses indicated that the WMC measure, RAC, and the inductive reasoning measure, LSET, were significantly related to grammar learning, however the grammatical sensitivity measure, the WIS, was not. Moreover, logistic mixed-effects modeling indicated that for our recall measure, performance on our WMC measure interacted with session, such that individuals who performed better on RAC showed greater gains in accuracy over sessions. A similar result was obtained for the recognition measure; however, additional variance in learning performance was accounted for by an interaction between the inductive reasoning measure, LSET, and session.

Overall, the results of the study suggest that WMC and inductive reasoning facilitate grammar learning. The fact the predictors interacted with learning session is likely due to the fact that grammatical learning builds on previous learning. For example, in English, it would be difficult for one to generate, “my two very fancy goats” without also being able to correctly generate “my two goats” or “my fancy goat.” Individuals with greater WMC and inductive reasoning ability were likely more able to learn rules, build upon them, and reinforce their own learning as they learned more complex rules, increasing their learning rate. Individuals with lower abilities, however, may have found it difficult to learn even the simpler rules and therefore struggled to see recurring patterns in more complex sentences, possibly interfering with (rather than reinforcing) learning; thus, their learning rate was slower compared to higher-ability individuals.

While the Session x RAC interaction was a significant predictor of both the recall and recognition measures, it is important to note that the Session x LSET interaction only accounted for a significant proportion of variance in recognition performance. This pattern of results confirms that WMC was generally involved in learning, however, the role of inductive reasoning is somewhat ambiguous. One possibility is that individuals with greater inductive reasoning ability were able to infer more rules but not necessarily retain

accurate representations in long-term memory (a function supported by WMC; Unsworth & Engle, 2007) and thus they were unable to accurately retrieve rules during the recall test. When tested using a recognition paradigm, however, high ability individuals were able to use cues to "fill in" or *redintegrate* their partial representations, and thus were more likely to correctly choose the grammatical phrase. Future research should continue investigating the role of inductive intelligence in explicit-inductive grammar learning.

Despite the fact that the grammatical sensitivity measure, WIS, did not account for variance in the learning tasks above and beyond the other predictors, the results of this study should not be interpreted as suggesting that grammatical sensitivity does not play a role in L2 learning. As noted above, prior research indicates that grammatical sensitivity is related to L2 learning and, in fact, the coefficients observed between the WIS and the grammar learning measures are similar to what have been found in the literature (Li, 2015). The null result observed here may have been due to sample size or characteristics (e.g., a restricted range in performance). Still, to the extent that the estimates are accurate, it is interesting to note that the role of grammatical sensitivity in grammar learning appears to be smaller than that of WMC and inductive reasoning. This may be because grammatical sensitivity is more a measure of English grammatical knowledge than learning (Carroll, 1993).

With the above limitations in mind, this study corroborates prior research indicating that WMC is a robust predictor of L2 learning and, more specifically, L2 grammar learning (Linck, Osthus, Koeth & Bunting, 2014; Miyake & Friedman, 1998). Moreover, while it may be somewhat intuitive that inductive reasoning is predictive of explicit-inductive grammar learning, we found that inductive reasoning accounts for at least one measure of grammar learning above and beyond WMC. Considering the large number of individuals that engage in L2 learning and the significance of knowing an L2, researchers should continue investigating the cognitive components of L2 learning.

References

Carroll, J. B. (1964). The Prediction of Success in Intensive Foreign Language Training. *Reprint from Training Research and Education, Chapter 4, p 87-136*: University of Pittsburgh Press, 1962.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studie.* Cambridge University Press.

Carroll, J. B., & Sapon, S. M. (1959). Modern language aptitude test. *New York: The Psychological Corp.*

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior, 19*(4), 450-466. doi:http://dx.doi.org/10.1016/S0022-5371(80)90312-6

Daneman, M., & Merikle, P. M. (1996). Working memory and language comprehension: A meta-analysis. *Psychonomic Bulletin & Review, 3*(4), 422-433. doi:10.3758/bf03214546

DeCaro, M. S., Van Stockum, C. A., & Wieth, M. B. (2016). When working memory capacity hinders insight. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 42, 39–49. doi:10.1037/xlm0000152

DeKeyser, R. (2003). "Implicit and Explicit Learning." In C. J. Doughty & M. H. Long (Eds.),*The handbook of second language acquisition* (Vol. 27) (pp. 313-348). Malden, MA: Blackwell.

DeKeyser, R. M. (1995). Learning second language grammar rules: An experiment with a miniature linguistic system. *Studies in second language acquisition*, *17*(3), 379-410.

Doughty, C.J., Campbell, S.G., Bunting, M.F., Bowles, A., & Haarmann, H. (2007). *The development of the High-Level Language Aptitude Battery*. College Park, MD: University of Maryland Center for Advanced Study of Language.

Doughty, C. J., & Long, M. H. (Eds.). (2008). *The handbook of second language acquisition* (Vol. 27). John Wiley & Sons.

Engle, R. W. (2002). Working Memory Capacity as Executive Attention. *Current Directions in Psychological Science, 11*(1), 19-23. doi:10.1111/1467-8721.00160

Engle, R. W. (2018). Working Memory and Executive Attention: A Revisit. *Perspectives on Psychological Science, 13*(2), 190-193. doi:10.1177/1745691617720478

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. A. (1999). Working memory, short-term memory, and general fluid intelligence: A latent-variable approach. *Journal of Experimental Psychology: General, 128*(3), 309-331. doi:10.1037/0096-3445.128.3.309

Gardner, R. C., & Lambert, W. E. (1965). Language aptitude, intelligence, and second-language achievement. *Journal of Educational Psychology, 56*(4), 191-199. doi:10.1037/h0022400

Gebauer, G. F., & Mackintosh, N. J. (2007). Psychometric intelligence dissociates implicit and explicit learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*(1), 34-54.

Hamrick, P., Lum, J. A., & Ullman, M. T. (2018). Child first language and adult second language are both tied to general-purpose learning systems. *Proceedings of the National Academy of Sciences*, 115(7), 1487-1492.

Hughes, M. M., Karuzis, V. P., Kim, S., O'Rourke, P., Sumer, A., Liter, A., … Campbell, S. G. (2016). *Assessing aptitude for USAF cyber warfare operations training: Interim results from field testing*. College Park, MD: University of Maryland Center for Advanced Study of Language.

Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity?! *Intelligence, 14*(4), 389-433. doi:10.1016/S0160-2896(05)80012-1

Li, S. (2014). The associations between language aptitude and second language grammar acquisition: A meta-analytic review of five decades of research. *Applied Linguistics*, *36*(3), 385-408.

Li, S. (2015). The construct validity of language aptitude. *Studies in Second Language Acquisition*, 1-42. doi:10.1017/S027226311500042X

Linck, J. A., Osthus, P., Koeth, J. T., & Bunting, M. F. (2014). Working memory and second language comprehension and production: A meta-analysis. *Psychonomic Bulletin & Review, 21*(4), 861-883. doi:10.3758/s13423-013-0565-2

Marshalek, B., Lohman, D. F., & Snow, R. E. (1983). The complexity continuum in the radex and hierarchical models of intelligence. *Intelligence*, *7*(2), 107-127.

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2015). Balancing type I error and power in linear mixed models. arXiv preprint arXiv:1511.01864. *Journal of Memory and Language*, accepted pending minor revisions.

Misyak, J. B., & Christiansen, M. H. (2012). Statistical learning and language: An individual differences study. *Language Learning*, *62*(1), 302-331.

Miyake, A., & Friedman, N. P. (1998). Individual differences in second language proficiency: Working memory as language aptitude. In A. F. Healy, L. E. Bourne, Jr., A. F. Healy, & L. E. Bourne, Jr. (Eds.), *Foreign language learning: Psycholinguistic studies on training and retention.* (pp. 339-364). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Oberauer, K., Süß, H. M., Wilhelm, O., & Sander, N. (2007). Individual differences in working memory capacity and reasoning ability. *Variation in working memory*, 49-75.

O'Rourke, P., Karuzis, V. P., Kim, S., Tseng, A., Pandža, N. B., Young, J. M., … Campbell, S. G. (2017). *Assessing aptitude for USAF cyber warfare operations training: Test specifications for USAF-CATA*. College Park, MD: University of Maryland Center for Advanced Study of Language.

Robinson, P. (1997). Generalizability and automaticity of second language learning under implicit, incidental, enhanced, and instructed conditions. *Studies in Second Language Acquisition, 19*(2), 223-247. Retrieved from http://www.jstor.org/stable/44488684

Peterson, C. R., & Al-Haik, A. R. (1976). The development of the Defense Language Aptitude Battery (DLAB). *Educational and psychological measurement, 36*(2), 369-380. doi:10.1177/001316447603600216

Sasaki, M. (1993), Relationships Among Second Language Proficiency, Foreign Language Aptitude, and Intelligence: A Structural Equation Modeling Approach. *Language Learning,* 43: 313-344. doi:10.1111/j.1467-1770.1993.tb00617.x

Shipstead, Z., Harrison, T. L., & Engle, R. W. (2016). Working memory capacity and fluid intelligence: Maintenance and disengagement. *Perspectives on Psychological Science*, *11*(6), 771-799.

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford University Press.

Sparks, R. L., Humbach, N., Patton, J., & Ganschow, L. (2011). Subcomponents of second‐language aptitude and second‐language proficiency. *Modern Language Journal, 95*(2), 253-273. doi:10.1111/j.1540-4781.2011.01176.x

Tagarelli, K. M., Mota, M. B., & Rebuschat, P. (2011, January). The role of working memory in implicit and explicit language learning. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 33, No. 33).

Unsworth, N., & Engle, R. W. (2007). The nature of individual differences in working memory capacity: active maintenance in primary memory and controlled search from secondary memory. *Psychological review*, *114*(1), 104.

Wilhelm, O. (2005). Measuring reasoning ability. In O.W. Wilhelm O.W. & R.W. Engle (Eds.), *Handbook of understanding and measuring intelligence*, (pp. 373-392). Thousand Oaks: Sage.

Relationship Between Creative Experience, Recognition of Creative Process and Aesthetic Impression in Art-Viewing

Kazuki Matsumoto
University of Tokyo, Tokyo, Japan

Takeshi Okada
The University of Tokyo, Tokyo, Japan

Abstract

This study examined the roles recognition of the creative process behind artworks plays in cognitive processes of art-viewing. To this end, we conducted an experiment (N = 45) in which prior experience of participants was manipulated and investigated whether and how creative experience influences subsequent cognitive processes while viewing artworks. We revealed that having creative experience before art viewing changes viewers recognition of the creative process behind artworks and causes them to have a more positive impression of the artworks. It was also revealed that these two changes are correlated. In particular, the emotion of admiration, which is considered a kind of social emotion, was found to be highly correlated with the recognition of assessed difficulty of the creative process. These results suggest the importance of recognition of the creative process behind artworks and contribute to understanding the cognitive process of art-viewing.

The effects of changing the mental model of one's body and sense of body ownership on pain perception

Miki Matsumuro (matumuro@rm.is.ritsumei.ac.jp)
Yuki Miura, Fumihisa Shibata, and Asako Kimura
College of Information Science and Engineering, Ritsumeikan University, 1-1-1 Noji-higashi
Kusatsu, Shiga 525-8577 JAPAN

Abstract

The mental model of one's body plays an important role in determining subsequent actions. We changed the mental model using visual information and observed the effects of such change on pain perception. These effects were compared to the effects of changes in the sense of body ownership, which is the sensation that something is a part of one's own body. Some researchers have shown that the sense of ownership is a factor modulating pain perception. In our experiments, we manipulated the visibility of participants' limbs using Mixed Reality (MR) techniques and measured their perceived pain and feelings while observing their limbs. Results showed the sensation that nothing can touch one's limbs decreased the strength of perceived pain.

Keywords: Sense of ownership, body representation, pain perception, multimodality, mixed reality

Introduction

We determine our next actions based on our own body representation or mental model of our bodies (Barsalou, 2008; Warren, 1984). Some features, such as posture, muscular strength, and size, change every moment or as we grow. Other basic features, such as bone structure, nerve mechanisms, and material properties, remain almost constant through life. If we can modulate such basic features in our mental model, can our perceptions be changed by the model? In this study, we investigate the relationship between the mental model of one's own body and perception, focusing on pain perception.

Sense of Ownership

One of the important sensations affecting the perception of pain is the sense of ownership or physical possession of one's body parts, such as hands and legs. The perception of ownership can be easily extended to non-body parts. The most famous example is the rubber hand illusion (Botvinick & Cohen, 1998): When a rubber hand and a participant's hand are repeatedly touched simultaneously while the participant is watching the rubber hand, he/she feels as if the rubber hand were his/her own.

Obviously, we cannot feel pain if something other than one's own body is attacked. Consistent with this idea, some researchers have shown that the pain threshold increases as the sense of ownership decreases (Martini, Kilteni, Maselli, & Sanchez-Vives, 2015; Martini, Pérez-Marcos, & Sanchez-Vives, 2014; Pamment & Aspell, 2017; Zanini, Montalti, Caola, Leadbetter, & Martini, 2017). However, some have argued that the sense of ownership has no effect on pain perception (Mohan et al., 2012).

Mental Model of Own Body

We propose that another important factor affecting the perception of pain is the material property of skin in a mental model of own body. If you imagine your skin is made with iron, for instance, you may not feel pain if someone hit you. Senna, Maravita, Bolognini, and Parise (2014) introduced the marble hand illusion: Participants in their study heard the sound of marble being struck when a hammer touched their hand. After five minutes, they felt their hands becoming stiffer, heavier, harder, less sensitive, and unnatural. However, Senna et al. (2014) did not investigate whether the manipulation affected the level of pain perception.

Another study showed that just changing the color of the skin was enough to change the threshold for heat pain; however, the effect of the manipulation on the mental model was not investigated (Martini, Pérez-Marcos, & Sanchez-Vives, 2013). These studies suggest the possibility that the mental model of one's body can be modulated to affect pain perception.

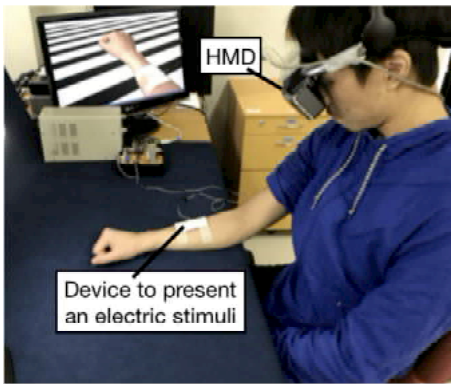
Aim of This Study

As previously noted, many previous studies have suggested that a sense of ownership was an important factor in pain perception. However, there is a possibility that the mental model of one's body is also modulated by manipulating ownership. Therefore, it is not clear whether pain perception is really related to the sense of ownership. To clarify the top-down effect on pain perception, we need to identify which of the mental model of the body or the sense of ownership has a stronger effect.

Mixed Reality

In previous studies, most researchers used a rubber hand or virtual body to manipulate participants' ownership or mental model of the body. Before introducing these manipulations, researchers had to increase participants' perception of ownership of these materials. For example, in Martini et al. (2015)'s experiment, participants viewed a virtual environment and virtual body from a first-person perspective for one minute. Afterward, the transparency of the virtual body was increased to decrease the sense of ownership. They showed that a low sense of ownership decreased pain sensitivity.

To be accurate, what the participants observed was not their actual body part but a rubber or virtual hand. Even if researchers made a realistic-looking hand, it would not be



(a) Experimental set-up

(b) View from the camera

(c) Hand and virtual background

(d) Opacity 50%

Figure 1: Experimental environment and the manipulation of the limb.

a real hand; discrepancies between it and their body might give the participants an uncomfortable feeling. We would not be able to determine whether participants' mental models of their bodies changed or if they constructed new mental models for the fake hand. Additionally, while the manipulations were performed on the fake hand, stimuli were administered to the real hand.

To overcome the limitations of the fake hand, we introduce a Mixed Reality (MR) technique, which allowed us to change the properties of objects in the real environment or add virtual objects to the real environment (Kannape, Smith, Moseley, Roy, & Lenggenhager, 2019). With this technique, we made participants' own limbs appear transparent and observed the change in their perceptions of ownership, mental models of their bodies, and pain perception.

Apparatus

MR Environment

Figure 1 shows the experimental environment. We adopted a video see-through-type HMD (Canon, HM-A1) and MR platform system (Canon, MP-110). We acquired the participant's perspective from the camera on the HMD and manipulated the alpha value for the area of the participant's hand as shown in Figure 1. Five levels of the alpha value were used: 100% (fully visible), 75% visibility, 50% visibility, 25% visibility, and 1% visibility (almost invisible). A background image under the participant's hand had a black-and-white stripe to facilitate the perception of transparency.

Electric Stimulus

The pain presentation device was a boosted current using a Cockcroft-Walton circuit as an electric stimulus generation apparatus through an input/output board (Kyohritsu Electronic Industry Co., Ltd., RBIO - 2 U). A conductor (diameter: 0.12 mm, 10 cores) was fixed to a 1 mm-thick rubber sheet. We presented the pain sensation by applying a current to this conductor. The intensity of electrical stimulation was 320 V at a current of 1.8 mA, and the pulse width was 0.15 s.

Experiment 1

The level of ownership and mental model of their limbs were recorded at each level of opacity from 100% (fully visible) to 1% (almost invisible). We added a blackout (BO) condition in which no visual stimulus was presented.

Method

Participants Fourteen students participated in Experiment 1.

Measurement The participants assessed their levels of pain using a visual analog scale (VAS). We prepared a 100-mm line whose left end indicated "no pain" and whose right side indicated "worst possible pain." The participants were asked to draw a cross on the point reflecting the level of pain they perceived.

We developed a questionnaire to assess the mental model. It consisted of 20 items including feelings thought to be important for pain perception. The order of the items was randomized.

Procedure The experiment consisted of two successive blocks: the questionnaire and a pain perception block. All participants started with the questionnaire and then continued to the pain perception block. Before starting the experiment, the participants were asked to read and sign a consent form.

Questionnaire Block After receiving brief instructions, the participants sat at a desk and rested an arm on the desk as illustrated in Figure 1(a). They donned the HMD and saw their non-manipulated limb through a camera (Figure 1(c)) before watching their limb becoming transparent. At the end of the transformation, they watched their transformed limb (e.g., Figure 1(d)) for 10 seconds. Next, they removed the HMD and completed the questionnaire, which employed a 7-point Likert scale. All participants completed each opacity condition in random order except for the BO condition, in which their limb was completely invisible.

Pain Perception Block The procedure was identical to the questionnaire block until the participant observed their transformed limb. In the pain perception block, they were given an electric stimulus following a cue from the experimenter while they were watching their transformed limb. After the stimulus was given, they removed the HMD and assessed the strength of the pain they perceived. The opacity conditions were presented to participants in random order. In both blocks, two minutes rest was provided between each condition.

Table 1: Result of factorial analysis.

Item	Factor loading				
	Ownership[1]	Transparency	Intangibility	Anxiety[2]	Weakness
I feel as if the observed arm is my own arm	-0.932	-0.167	-0.061	0.166	0.033
The observed arm doesn't look mine	0.80	0.202	-0.057	-0.132	-0.129
I feel as if the observed arm is not my own arm	0.736	-0.020	0.422	-0.115	-0.009
My arm seems to be not present in the environment	0.733	0.220	0.323	-0.172	-0.049
I feel the observed arm is a real one	-0.723	-0.310	-0.130	0.204	-0.093
I feel as if my arm is transparent	0.244	0.932	0.237	-0.072	0.067
The arm is transparent	0.174	0.861	0.226	-0.094	0.105
My arm feels sparse	0.338	0.739	0.339	-0.008	0.112
I feel as if something can pass through my arm	0.090	0.436	0.834	-0.025	-0.055
I feel as if my arm is empty	0.137	0.304	0.832	0.008	-0.032
I feel as if I am a ghost	0.220	0.188	0.610	-0.137	-0.278
My arm feels numb	0.322	0.045	0.517	-0.434	-0.272
I don't feel fear by observing the arm	-0.069	-0.164	-0.086	0.883	0.161
I feel ill by observing the arm	0.122	0.021	0.087	-0.804	-0.118
I feel relieved by observing the arm	-0.250	-0.169	-0.249	0.735	0.234
I feel calm by observing the arm	-0.185	0.122	0.189	0.390	-0.054
My arm feels softer	0.062	0.056	-0.095	0.064	0.900
My arm feels weakened	0.126	0.115	-0.110	0.055	0.800
My arm feels lighter	-0.209	-0.023	0.053	0.093	0.509
My arm feels insensitive	-0.058	0.102	-0.264	0.232	0.480

[1] These loadings mean the contribution to "less ownership." Score of this factor was reversed to make easy to understand the results.
[2] These loadings mean the contribution to "less anxiety." Score of this factor was reversed to make easy to understand the results.

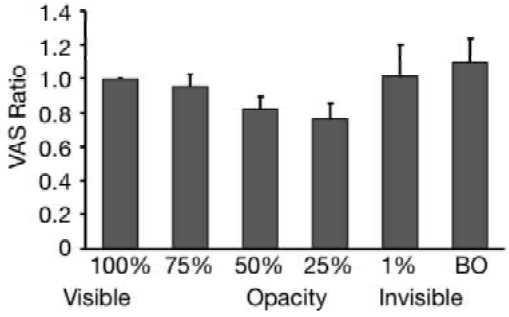


Figure 2: Means of pain assessment in Experiment 1.

Results and Discussion

Pain Perception We measured the distance from the left-most point to the marked point on the pain scale. The length in the 100% condition was the criterion value, and the length in each condition was converted to its ratio to the criterion value (Figure 2). One participant whose ratio deviated over 3 SD from the average was excluded from the following analyses. A repeated ANOVA showed the effect of opacity was significant ($F(5,60) = 2.467, p = .042$). In the 25% condition, the perceived strength of pain was lower than in the fully visible (100%) condition ($p = .036$). Perceived pain was stronger in the 1% and BO conditions than in the 50% and 25% conditions (ps $< .050$).

These findings and the tendencies in Figure 2 show that as the limb became more transparent, the level of perceived pain became weaker. However, when the limb was nearly or completely invisible, the strength of pain rose to near the value of the fully visible condition.

Questionnaire and Pain Perception We conducted a factorial analysis using the ratings of the questionnaire. We found five factors shown in Table 1: ownership, transparency, intangibility (i.e., nothing can touch their limb), anxiety, and weakness. The bigger value means the strong feeling for the factor.

A repeated ANOVA for each factor score (Figure 3) shows opacity value has a significant effect on all factors other than weakness (ownership $F(12,48) = 23.182, p < .001$; transparency $F(12,48) = 64.927, p < .001$; Intangibility $F(12,48) = 25.86, p < .001$; anxiety $F(12,48) = 6.716, p < .001$). For the ownership score, there was a significant difference in all pairs other than the pair of 1% and 25% and the pair of 50% and 75% (ps $< .05$). For the transparency score, the score in the 100% condition was higher than for any other conditions (ps $< .001$). For the intangibility score, the differences in scores between the 100% condition and all other conditions and between the 25% and 75% conditions were significant (ps $< .005$). For the anxiety score, the score in the 100% was bigger than that in all other conditions except for the 75% condition (ps $< .01$).

Table 2 shows the correlation coefficient values for the scores of all pairs among five factors and pain perception. We excluded the 1% condition from this analysis because

Table 2: Coefficient values in Experiment 1.

	Ownership	Transparency	Intangibility	Anxiety	Weakness
Ownership					
Transparency	-0.528 ****				
Intangibility	-0.467 ****	0.634 ****			
Anxiety	-0.471 ****	0.166	0.377 **		
Weakness	0.090	0.124	-0.295 *	-0.274 *	
Pain	0.195	-0.357 **	-0.305 *	-0.183	0.072

$^{+}p < .01$, * $p < .05$, ** $p < .01$, *** $p < .005$, **** $p < .001$

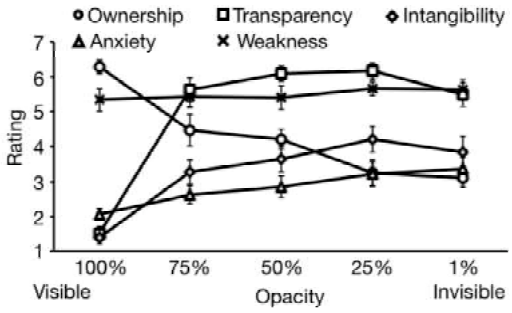


Figure 3: Mean scores for each factor in each condition.

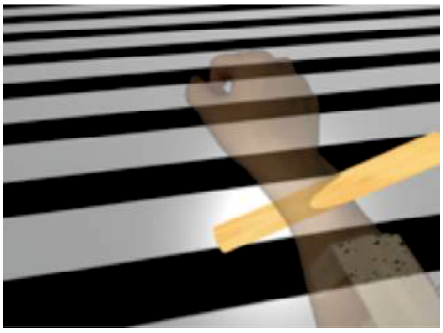


Figure 4: Manipulation in the PT condition: The virtual stick passing through the limb.

some participants shared that they could not see their limbs in this condition, which had a different effect on pain perception than seeing the transparent limb.

The results of correlation analysis show a negative correlation between the level of pain and the scores of transparency and intangibility. Sense of ownership did not correlate to the strength of pain, contrary to the results of the previous studies. In Experiment 2, to identify the most crucial factor for pain perception, we added manipulations changing the perceptions of ownership and intangibility.

Experiment 2

Two manipulations were introduced in Experiment 2. One was "passing through (PT)," in which we passed a virtual stick through the participant's limb as shown in Figure 4. The PT manipulation would increase the sensation of intangibility. Another was "spontaneous movement (SM)," in which the participant moved his/her finger. Many studies showed that observing the body moving in the way as they wanted to increase the sense of ownership. The experiment was a 2 (opacity: 25% and 100%) $\times$ 2 (PT manipulation: PT and no-PT) $\times$ 2 (SM manipulation: SM and no-SM) within-participants design.

Method

Participants Eleven students participated in Experiment 2.

Procedure The procedure was identical to that used in Experiment 1, except that we added the PT and SM manipulations in some conditions. In the PT condition, we moved the virtual stick 10 times as it passed through the participant's limb (Figure 4). In the no-PT condition, we added no manipulation, changing the feeling of intangibility. For the SM manipulation, the participant bent his/her finger as instructed by the experimenter. In the no-SM condition, the participant was told nothing and did not move his/her finger or limb. The manipulation(s) were added before the participant answered the questionnaire and before the electric stimuli were given. The PT manipulation was always conducted before the SM manipulation.

Results and Discussion

Manipulation Check One participant who hardly felt pain in any condition was excluded from the following analyses. At first, we calculated the scores for the five factors found in Experiment 1 to confirm the effects of the manipulations. We conducted a 2 (opacity: 25% and 100%) $\times$ 2 (PT manipulation: PT and no-PT) $\times$ 2 (SM manipulation: SM and no-SM) ANOVA on the scores for ownership and intangibility feelings. The ANOVA for ownership feelings showed that SM manipulation had no effect. The only significant effects were the main effect of the opacity factor ($F(1,9) = 26.396, p < .001$) and the interaction between the opacity and PT manipulation factors ($F(1,9) = 11.505, p = .008$). The score for ownership feeling was generally higher when the limb was fully visible. The PT manipulation decreased the ownership in the 100% condition ($F(1,18) = 9.172, p = .007$). The SM manipulation had no effect on the score of ownership feeling.

On the other hand, PT manipulation efficiently increased the sensation of intangibility. The main effects of the opacity factor ($F(1,9) = 40.490, p < .001$) and the PT manipulation factor ($F(1,9) = 23.802, p < .001$) were significant

Table 3: Coefficient values in Experiment 2.

	Ownership	Transparency	Intangibility	Anxiety	Weakness
Ownership					
Transparency	-0.702 ****				
Intangibility	-0.533 ****	0.670 ****			
Anxiety	-0.540 ****	0.285 *	0.587 ****		
Weakness	0.039	0.140	-0.226 *	-0.394 ****	
Pain	0.218 $^{+}$	-0.363 ****	-0.405 ****	-0.325 ***	0.167

$^{+}p < .01$, * $p < .05$, ** $p < .01$, *** $p < .005$, **** $p < .001$

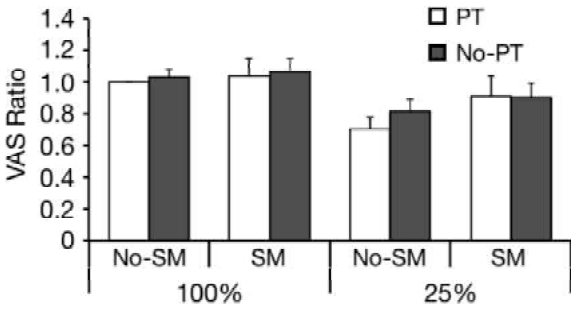


Figure 5: Means of pain assessment in Experiment 2.

for the intangibility score. The interaction between the opacity and PT manipulation factors was also significant ($F(1,9) = 12.091, p = .007$). The effect of PT manipulation was (marginally) significant in both the 25% and 100% conditions (25% $F(1,18) = 4.069, p = .059$; 100% $F(1,18) = 35.888, p < .001$). Although the effect was weak in the 25% condition, the PT manipulation was successful overall.

Pain Perception The VAS rating was converted into the ratio to the length in the 100% condition without either manipulation. Figure 5 shows the mean ratio in each condition. A 2 (opacity: 25% and 100%) $\times$ 2 (PT manipulation: PT and no-PT) $\times$ 2 (SM manipulation: SM and no-SM) ANOVA revealed a significant main effect of the opacity factor ($F(1,9) = 15.244, p = .004$). The interaction between the opacity and SM manipulation factors was marginally significant ($F(1,9) = 3.391, p = .099$). Consistent with the results of Experiment 1, the strength of perceived pain decreased when the participant's limb was transparent. However, the MT manipulation, whose effect was confirmed, had no effect on pain perception.

The Five Factors and Pain Perception The correlation coefficient values among the five factors and pain perception are shown in Table 3. Four factors (ownership, transparency, intangibility, and anxiety) had a (marginally) significant correlation with pain perception, while only ownership and intangibility were related to pain perception in Experiment 1. Many pairs of the scores among the five factors have a strong correlation. Thus, those strong correlations may have caused some spurious correlations.

Table 4: Explanatory powers of factors.

	Explanatory	Coefficient	t value	r^2
Single	Ownership	0.040	1.969^{+}	0.047
	Transparency	-0.042	3.444****	0.132
	Intangibility	-0.072	3.908****	0.164
	Anxiety	-0.096	3.034***	0.094
	Weakness	0.040	1.491	0.015
Multiple	Ownership	-0.013	0.494	0.135
	Transparency	-0.048	2.789**	
	Ownership	< 0.001	0.021	0.164
	Intangibility	-0.071	3.274**	
	Ownership	0.011	0.469	0.108
	Anxiety	-0.087	2.290*	
	Transparency	-0.019	1.203	0.179
	Intangibility	-0.052	2.104*	
	Transparency	-0.034	2.747**	0.185
	Anxiety	-0.072	2.247*	
	Intangibility	-0.058	2.554*	0.175
	Anxiety	-0.040	1.043	

$^{+}p < .01$, * $p < .05$, ** $p < .01$, *** $p < .005$, **** $p < .001$

We tried to identify the crucial factor for pain perception using regression analysis. The results of all analyses are summarized in Table 4. From the results of simple linear regression analyses, four factors (ownership, transparency, intangibility, and anxiety) could explain the strength of perceived pain. We then conducted multiple regression analyses in which each pair of these four factors was chosen as an explanatory variable, and the strength of perceived pain was a dependent variable. When the intangibility score was paired with other factors' score, intangibility was always the only factor with significant explanatory power, and the paired variables' power was not significant. These results suggest that the sensation of intangibility was the crucial factor directly affecting pain perception.

However, the relationship between other factors and the intangibility feeling cannot be determined from this experiment. Further studies are needed to identify whether the other factors explain the intangibility score or whether the intangibility score explains the scores of other factors. Additionally, the r^2 values in our regression analyze were not sufficiently high. Collecting more data will confirm the results of this study.

General Discussion

We investigated the relationships among the mental model of one's own body and pain perception. The crucial factor affecting pain perception was the sensation that nothing can touch one's limbs (intangibility); as this sensation increased, the perceived level of pain decreased. The sense of ownership could not account for the level of perceived pain.

The properties of the mental model of one's body were easily modulated by visual information. A decrease in the perceived opacity of one's body parts decreased feelings of ownership and increased feelings of transparency, intangibility, and anxiety. The passing through manipulation successfully increased the feeling of intangibility. However, observing spontaneous actions did not increase ownership, contrary to findings in previous studies.

Sense of Ownership

We introduced a novel technique, MR, to manipulate body properties. The MR technique can change participant perceptions of the properties of their own limbs. The observed limb had features identical to their own limb and perfectly mimicked its movement. The participants were able to see every movement of their whole limb even if it was a very small movement such as breathing. This phenomenon had already been used to evaluate the sense of ownership; therefore, the additional spontaneous movement had no effect on the feeling of ownership. In future research, we will be able to use other kinds of manipulation, such as a delayed presentation of action, which was found by Kannape et al. (2019) to decrease the feeling of ownership.

In previous studies, the presented rubber or virtual limb was not the participants' own limb. Therefore, the participants created a new mental model of the presented limb and provided the body ownership to it. Changes to the presented limb took the ownership away from it. In short, the participants did not perceive the presented artificial limb to be their own anymore. For this reason, the feeling of ownership had a strong effect on pain perception (e.g., Martini et al., 2014; Pamment & Aspell, 2017); the participants who left more ownership on the presented body felt strong pain.

On the other hand, the MR technique decreased the inherent ownership of the body leading to the sensation that one's own limbs are not part of one's body. The sensation of intangibility had more of an impact because the participants still believed that the presented limb was their own, even if the sensation of ownership had decreased. We should carefully consider which type of ownership we manipulate, the elicited ownership such as in previous studies or the inherent ownership such as in this study (cf. Kannape et al., 2019); the manipulation may have different effects.

The Mental Model of One's Body

We could change pain perception by changing the properties of the mental model of the body. The results of this study can be explained as a top-down effect on perception (Gregory, 1997; Martini et al., 2013; Senna et al., 2014). The feeling of intangibility in this study meant the sensation that one's body had become something cannot be touched, like that of a ghost. Such creatures are believed to be unable to feel pain. The illusion of transparency triggered this perception, resulting in decreased pain.

Changes in other properties, such as an iron skin, might have the same effect as transparency. This top-down effect could also have an opposite effect: For example, if the material of body is changed to something fragile, such as glass, and the body is hit by a hammer, participants may perceive more pain that with their normal bodies. In addition, some changes have the potential to change task performance.

References

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology*, *59*, 617–645.

Botvinick, M., & Cohen, J. (1998). Rubber hands 'feel' touch that eyes see. *Nature*, *391*(6669), 756.

Gregory, R. L. (1997). Knowledge in perception and illusion. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *352*(1358), 1121–1127.

Kannape, O. A., Smith, E. J., Moseley, P., Roy, M. P., & Lenggenhager, B. (2019). Experimentally induced limb-disownership in mixed reality. *Neuropsychologia*, *124*, 161–170.

Martini, M., Kilteni, K., Maselli, A., & Sanchez-Vives, M. V. (2015). The body fades away: investigating the effects of transparency of an embodied virtual body on pain threshold and body ownership. *Scientific Reports*, *5*, 13948.

Martini, M., Pérez-Marcos, D., & Sanchez-Vives, M. V. (2013). What color is my arm? changes in skin color of an embodied virtual arm modulates pain threshold. *Frontiers in Human Neuroscience*, *7*, 438.

Martini, M., Pérez-Marcos, D., & Sanchez-Vives, M. V. (2014). Modulation of pain threshold by virtual body ownership. *European Journal of Pain*, *18*(7), 1040–1048.

Mohan, R., Jensen, K. B., Petkova, V. I., Dey, A., Barnsley, N., Ingvar, M., ... Ehrsson, H. H. (2012). No pain relief with the rubber hand illusion. *PloS one*, *7*(12), e52400.

Pamment, J., & Aspell, J. (2017). Putting pain out of mind with an 'out of body' illusion. *European Journal of Pain*, *21*(2), 334–342.

Senna, I., Maravita, A., Bolognini, N., & Parise, C. V. (2014). The marble-hand illusion. *PloS one*, *9*(3), e91688.

Warren, W. H. (1984). Perceiving affordances: Visual guidance of stair climbing. *Journal of Experimental Psychology: Human Perception and Performance*, *10*(5), 683–703.

Zanini, A., Montalti, M., Caola, B., Leadbetter, A., & Martini, M. (2017). Pain during illusory own arm movement: A study in immersive virtual reality. *European Medical Journal*, *2*(2), 90–97.

Exploring the Early Childhood Executive Function and Language Relationship: A Preliminary Analysis

Kaitlyn May
University of Alabama, Tuscaloosa, Alabama, United States

Ursula Johnson
University of Texas Health Science Center, Houston, Texas, United States

Janelle Montroy
University of Texas Health Science Center, Houston, Texas, United States

Abstract

Recent studies demonstrate strong, concurrent relationships between language and EF, particularly during early childhood. However, the literature remains controversial with respect to this relationship. Whereas some studies cite a bidirectional relationship, others suggest that EF is predictive of language gains, while others suggest that it is language which affects EF through conversational practice. Further controversy remains in the literature regarding which components of EF are engaged in the processes. The bidirectionality of current research in this area suggests that perhaps EF and language are best fitted by a curvilinear relationship. This is compounded by the fact that a large number of these studies have employed linear statistical analyses to examine the relationship of the two constructs. Thus, in order to further specify the relationship between EF and language development, we examined monolingual and bilingual infants and toddlers to determine the utility of a curvilinear model to assess the EF and language relationship, what aspect of language inhibitory control most correlates to EF, and whether there is a monolingual/bilingual difference. Results indicate that the EF and language early childhood relationship is best fitted by a curvilinear model.

Development of Verb Morphology: From Item-Specificity to Proficient Use

Jekaterina Mažara (jekaterina.mazara@uzh.ch)
University of Zurich, Zurich, Switzerland

Sabine Stoll (sabine.stoll@uzh.ch)
University of Zurich, Zurich, Switzerland

Abstract

The initial phase of linguistic production by children is characterized by rote-learned, lexically restricted forms and constructions. Only during later phases of language acquisition do they develop flexibility across a paradigm and mix lexical and grammatical material more freely. In the development of verb morphology, a correlation between the use of tense and aspect has been observed in many languages. It has been suggested that this leads to an intermediary state of paradigm categorization based on temporal categories. So far the flexibility of individual verbs occurring in different tense-aspect combinations has not been examined in detail. Here we evaluate the flexibility of verb use in a large longitudinal corpus of 4 Russian children. We compute the Shannon entropy of verb stems distributed over individual grammatical forms. Results show that children do not pass through a stage of paradigm categorization based on aspecto-temporal categories. After a brief item-specific phase of rote learned forms, they quickly become flexible users of verbs in both aspects.

Keywords: language acquisition; corpus study; item-specificity; verb morphology; aspect; Russian

Introduction

Usage-based approaches to language acquisition propose an early phase during which children use a small number of lexically specific constructions which are presumably rote-learned (Lieven, Pine, & Baldwin, 1997; Pine & Lieven, 1997; Tomasello, 2000, 2003). During this short phase of lexical specificity, flexibility of word form use is very low, but soon after using the first rote-learned constructions, children start to produce new forms and apply them to new contexts. So far, relatively little is known about this generalization process from lexically specific constructions to full productivity.

In this study, we focus on the acquisition of Russian verb morphology and the role of aspect. Grammatical aspect is the expression of the viewpoint on the temporal structure of an event. *Perfective aspect* describes an external and temporally bounded view of a completed event, while *imperfective aspect* focuses on the internal stages or temporal extension of an event (Comrie, 1976).

Languages differ vastly in how (and if) they mark grammatical aspect but independent of the realizations, aspect has been found to play a pivotal role in the acquisition of the verbal system in relation with tense (Shirai & Anderson, 1995; Shirai, Slobin, & Weist, 1998). Correlations between verbs with a defined end-point (telic verbs) and perfective past marking as well as verbs without a defined endpoint (atelic) and non-past imperfective marking have been found in early acquisition of a number of different languages (cf. Bloom, Lifter, and Hafitz (1980); Harner (1981); Shirai and Anderson (1995); Clark (1996); Johnson and Fey (2006) for English, Bronckart and Sinclair (1973) for French, Antinucci and Miller (1976) for Italian, Li and Bowerman (1998); Shirai and Anderson (1995); Shirai et al. (1998); Li and Shirai (2000) for Japanese; Stoll (1998, 2005); Stoll and Gries (2009); Gagarina (2000); Bar-Shalom (2002) for Russian; Li (1990); Li and Shirai (2000) for Mandarin; Aksu-Koç (1998) for Turkish; Stephany (1985) for Greek; Weist, Wysocka, Witkowska-Stadnik, Buczowska, and Konieczna (1984); Weist and Konieczna (1985) for Polish; as well as self-organizing feature map models (cf. Li (2000); Li and Shirai (2000)).

It has been suggested that due to the presence of this correlation, after the lexically-specific phase, the development of productivity passes through an intermediary stage, during which children are more productive in their use of verbal morphology with the appropriate prototypes of a category (also known as the *Aspect Hypothesis* see Shirai and Anderson (1995)). These correlations are also present in the speech of adults, albeit to a lesser degree. However, to date, only a few studies have systematically compared these correlations in child and child-surrounding speech. For Russian children, Stoll and Gries (2009) have found a gradual decrease of this association in children over the course of development.

The goal of this study is to examine the development of flexibility of verb form use in Russian children. We test whether there is indeed a transition phase based on the tense-aspect correlation during which children are more productive within sub-categories of the verb paradigm before becoming fully productive verb users.

We first establish phases in production based on verb form inventory size. We then compare both type and token distributions in children's use during these phases to that of adults. We show that in token use, both adults and children display distributional bias of tense-aspect correlations. The bias is stronger in children in the first phase of production and approaches adult levels in the second phase. We evaluate the flexibility of use over time by measuring the entropy of lemmas used with individual grammatical forms. We show that, as item-specificity decreases, a great variety of forms is introduced early on and quickly generalized so that both past and non-past marking is used with verbs of both aspects.

Verb morphology in Russian

Russian has relatively complex verbal morphology centering on a semantically and morphologically complex category of grammatical aspect which interacts with tense. Grammatical aspect in Russian is characterized by a perfective/imperfective distinction and each verb is either perfective or imperfective. In contrast to English which has one single aspectual marker (*-ing*), Russian has many different markers for the perfective aspect (mainly prefixes and one suffix) and one suffix (with various allomorphs) for the imperfective aspect or zero marking.

On the functional level, several temporal and contextual features influence the use of the two aspects. Russian imperfective verbs are used when the duration of an action is relevant (e.g. *ona čitaet ves' den'* 'she reads all day') and if the action is presented as a completed event (e.g. *ona včera čitala ves' den'*, 'she spent all day reading yesterday'). Perfective verbs are used when the focus of the utterance is a boundary of the action; this can be either the beginning of an action, the end/result or both (e.g. *ona dočitala knigu*, 'she finished reading the book'). Morphologically, perfectives are typically derived from imperfectives by prefixation. To complicate things, however, the meaning cannot be derived via simple rules (Timberlake, 2004) and always involves some degree of rote-learning. There is no one-to-one relationship between prefixes and the resulting meaning change in the verb they are attached to. Further, most verbs can combine with multiple prefixes, while others are restricted in their combinability.

Verbs of both aspects express other verb categories (person, number, tense, voice, and mood) with the same morphemes. There are, however, some differences in meaning. Non-past morphology denotes present tense when it appears with imperfectives, but expresses the future in combination with perfectives. To express imperfective future, an analytic form is used (consisting of a finite 'to be' auxiliary and the infinitive of the main verb). In this paper, we focus on the acquisition of synthetic morphology and, therefore, exclude the analytic future. Past morphology can be used with both aspects equally.

The broad generalization found in the works cited in Shirai et al. (1998) states that children begin their acquisition of verb forms by using past morphology with achievement verbs and progressive morphology with activity verbs and only later extend it to the other group. Since lexical aspect is not annotated in the corpus we use, we focus on correlations between grammatical aspect and tense. However, this still allows us to assess this hypothesis, since achievements are necessarily perfectives and activities are necessarily imperfectives in Russian. We will, therefore, focus on whether Russian children display correlations between perfective aspect and past tense (e.g. *On doel sup*, 'he ate the soup' (meaning: he finished the bowl)) and imperfective aspect and non-past marking (*On smotrit televizor*, 'he is watching TV').

Methods

Data

The data is extracted from an audio-visual longitudinal corpus of Russian language acquisition (Stoll & Meyer, 2008) comprising data of six monolingual children living in St.Petersburg, Russia. All recordings were done in naturalistic settings at the home of the children and include the focal child and a varying number of surrounding speakers including siblings (excluded here) and adults. The children were recorded for one hour each week. We focus on 4 children, whose recordings started before the age of three. The entire corpus is transcribed and words are annotated for part of speech and morphology. Table 1 summarizes the number of utterances, words, and verbs uttered by each focal child as well as the age range of recording.

Table 1: Age spans of the focal children and number of words produced by the children and surrounding adults

Focal Child	Age span	Number of recordings	N(tokens)			
			Child		Adults	
			words	verbs	words	verbs
1	1;8.10 - 4;8.21	130	241,948	38,843	301,418	60,987
2	1;4.23 - 4;1.24	109	57,929	5,411	354,034	65,173
3	1;3.24 - 4;9.29	123	74,926	10,733	423,078	84,659
5	1;11.28 - 4;3.12	67	97,397	16,585	223,289	43,149

Finding phases in acquisition

First, we establish whether there are phases in verb form acquisition. The phases were derived directly from the target children's verb form production. We computed the additive growth in full verb forms (stem+grammatical markers) over time. The growth curves show a slow rate of increase in the earlier sessions followed by a sudden increase in the rate of newly observed forms [1]. To estimate the age at which this change in rate of acquisition occurs, we conducted a segmented regression on the growth curve of each child. The break points at which the regression created a new segment are summarized in Table 2. We use these points as the estimated end of the first phase of production for the next analysis.

Table 2: Break points in growth curve as identified by segmented regression.

Child	Break-point
Child 1	2;2
Child 2	3;3
Child 3	2;3
Child 5	before recordings started

[1] This was the case for all but Child 5 who already had highly developed speech at the onset of the recordings. Child 5, therefore, did not exhibit this change in rate of newly observed forms.

Entropy of verb form use

To asses the development of flexibility of form use in the observed production, we used Shannon Entropy (Shannon, 1948), the rate at which a process produces information by characterizing the balance of frequency distributions over a set of elements. If the probability of produced elements is distributed equally among them, the output is less predictable. Early child language is usually characterized by the repeated use of a few forms, while other forms might appear only once. This would result in a highly predictable output and low entropy. The formula for Shannon entropy is given in Eq. 1

$$H(X) = -\sum_{i=1}^{N} p(x_i) \log p(x_i) \quad (1)$$

where N is the number of distinct forms and $p(x)$ is the probability of occurrence of a specific form [2].

Analysis 1: Distribution of forms in the first and second phase To gain a better understanding of verb form production during the first and the second phase of development, we extracted the verb lemmas (lexical elements) used before the break point in development. To obtain a sample comparable in size and lexical coverage, we extracted the same lemmas from the adults' production during this phase and sampled the same number of tokens as produced by the focus child. Finally, we conducted the same procedure for both focus child and surrounding adults for the second phase. To gain a first insight into the form use and asses the level of item-specificity, we visualised the data in mosaic plots showing both type and token use of children and adults in both phases. To characterize the difference between the distributions, we computed the Jensen-Shannon divergence (Lin, 1991) between each child and their surrounding adults, and between the child's own first and second phases. Jensen-Shannon divergence (JSD) measures the distance between two probability distributions over the same elements (i.e. verb lemmas in this case). The formula for JSD for two distributions P and Q with equal weight (0.5) is given in Eq. 2.

$$JSD(P\|Q) = \frac{1}{2}D(P\|M) + \frac{1}{2}D(Q\|M) \quad (2)$$

M represents the average distribution $M = \frac{1}{2}(P+Q)$; and D stands for Kullback-Leibler divergence (KLD, sometimes also called *relative entropy*), given in 3.

$$KLD(P\|Q) = -\sum_{x\in X} P(x) log \frac{Q(x)}{P(x)} \quad (3)$$

JSD is based on KLD, but it is symmetric and its value is always finite and non-negative. When P = Q, JSD = 0 (i.e. the two distributions are equal). To evaluate the development of forms use across verbs from the two aspects and different grammatical categories, we computed the distribution of grammatical markers over the lexical elements extracted from the first and second phase of each child and the surrounding adults. First, we compare the probability distributions of the child and the adults during phase 1 and phase 2; then, we also compare the distribution of the child in phase 1 and the same child during phase 2. We do this both for types and tokens of verb forms.

Analysis 2: Flexibility of form use over time While JSD is useful when we can compare the probability distributions for a set of identical items, it is impossible to assess the week-to-week development in this way, since we have no way of controlling the context and lexical content of individual recording session. Cutting the production down to forms that appear in both adults' and children's production would also result in a severe underestimation of the development and a distortion of the actual production. Therefore, we compute the entropy of all elements occurring in an individual recording session. To assess whether certain tense-aspect combinations indeed aid in acquisition, we divide the data into past and non-past marked verbs and compute the entropy of perfective and imperfective verb lemmas used with past and non-past marking. As the children develop away from the item-specific phase, we expect their use of individual grammatical markers to become more flexible, i.e. they learn to combine a variety of verb lemmas with individual forms.

To estimate the time at which children start approaching adult levels of flexibility in their verb form use, we use the entropy computations of adults as a comparison within each session. The children's entropy is divided by the corresponding surrounding adults' entropy within each session. A value below 1 signifies that the child is below the adult level of entropy, values above 1 mean that the child's verb production has a higher entropy than that of surrounding adults. To control for contextual influence and other effects that might lead to particularly high or low entropies, we bootstrapped the data in each recording session for 100 iterations.[3]

Since the corpus consists of naturalistic data, it is difficult to normalize the production for comparative reasons. Sampling a fixed number of tokens from children and adults in each session would distort the data in a number of ways: i) if a fixed number of tokens is sampled across the recording span (e.g. 500 tokens from children and adults), the children's initial production is inflated, while adults and children's later production are underestimated; ii) if the number of tokens is determined by the number produced by the target child in each recording, this – again – severely underestimates the adults' production in the early recordings and does not represent a realistic measure for comparison. Same goes for a restriction of lexical elements used for the computation of entropy, since the fact that children's vocabulary size is growing is also an important factor and should not be ignored. This is

[2] For the computation of entropy used over time, where we looked at each session individually, we did not treat the system as a complete survey of the forms that have been acquired by that point in time.

[3] The relatively low count of bootstraps was chosen for reasons of graphic clarity.

especially important for Russian, where aspect is encoded as part of the lemma.

To evaluate whether the age at which significant changes in the entropy of lemma use with individual forms happen, we fitted a generalized additive model to the data and estimated the change points of the regression to find the age at which diversification starts and when it levels off.

Results

Analysis 1: Distribution of forms in the first and second phase

Looking at the sample of matched verb lemmas and number of tokens in the two phases of each child and their surrounding adults, we see that the type distribution is slightly more diversified than the distribution of tokens. While there are tendencies to use more non-past forms with imperfective verbs and more past forms with perfectives, even during the earliest phase this tendency is not absolute and both past and non-past forms appear with verbs of both aspects early on. While types are distributed fairly equally, the token distribution is less even during both phases. This holds for both adults and children.

Table 3 shows the JSD computed for each child's early production compared to that of surrounding adults and the child's own production during the later phase. In the case of child 5, it was not possible to establish an early phase similar to that of the other children. Additionally, Child 5's earliest recorded production is so varied that it was impossible to obtain a sample of the same lexical items within the same time window from the surrounding adults. Therefore, the results shown for Child 5 represent a comparison of Child 5's production during the first 5 recordings sessions compared to a lexically and size-matched sample from his surrounding adults across the entire corpus.

Table 3: Jensen-Shannon divergence per child.

	Phase1 Child-to-Adults		Phase2 Child-to-Adults		Child Phase1-to-Phase2	
	Types	Tokens	Types	Tokens	Types	Tokens
Child 1	0.124	0.501	0.020	0.056	0.122	0.421
Child 2	0.143	0.505	0.032	0.135	0.115	0.508
Child 3	0.121	0.647	0.074	0.107	0.112	0.609
Child 5	n/a	n/a	0.097	0.327	n/a	n/a

In all samples, the difference between the distributions of tokens is more pronounced than that of types, and shows less of a decrease between the two phases. However, the difference between each child's first phase and second phase sample is comparable to the difference between the child's production and that of adults in phase 1. This suggests that their development approaches a stage where their use of verb forms in spontaneous home interactions is very similar to that of the adults.

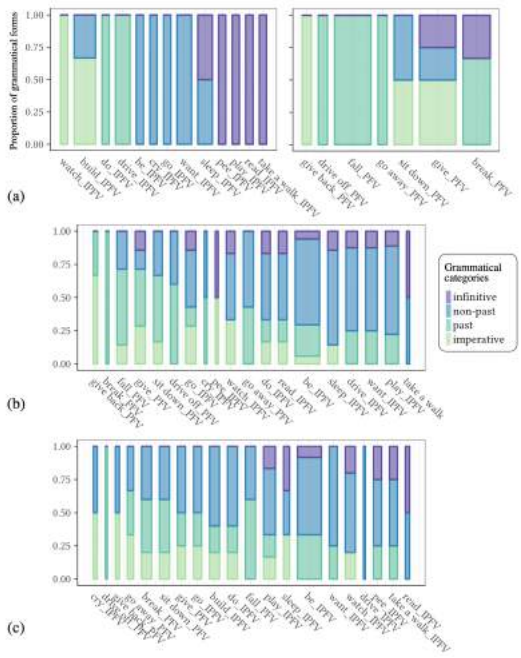


Figure 1: (a) Distribution of full form verb types in the production of Child 1 during phase 1; (b) Distribution of types in a sample of same lemmas and same number of tokens in Child 1's production during phase 2; (c) Distribution of types in a sample of same lemmas and same number of tokens in adults' production during phase 1.

To gain insight into the actual combinations of lemmas and forms used in each phase, Figures 1 a–c and 2 a–c exemplify the visualization of the type and token use within the sample of Child 1 and surrounding adults (we are not able to show the corresponding visualizations of the other children for space reasons). The thickness of the bars corresponds to the distribution of forms across the verb lemmas, while the colors stand for grammatical categories to which the forms belong. For types, only the plot for the child's first phase distribution was split by aspect, because phase 2 did not show the difference as strongly. For the token use, however, all plots are split by aspect, since the token distribution of both children and adults shows more differences between the use of grammatical markers with verbs in the two aspects.

Analysis 2: Flexibility of form use over time

Entropy ratios (child/adults) of the use of lemmas with individual grammatical markers from the sub-sets of non-past and past morphology and the segmented regression reveal that difference in the onset of diversification is not large. For past morphology, perfective lemmas show an earlier increase of entropy, but imperfective lemmas follow suit only a few weeks later and vice versa. The onset of use starts with imperfective+non-past and perfective+past for all children except Child 5, whose production is already diversified at the start of recordings. Only Child 3 shows a lag of more

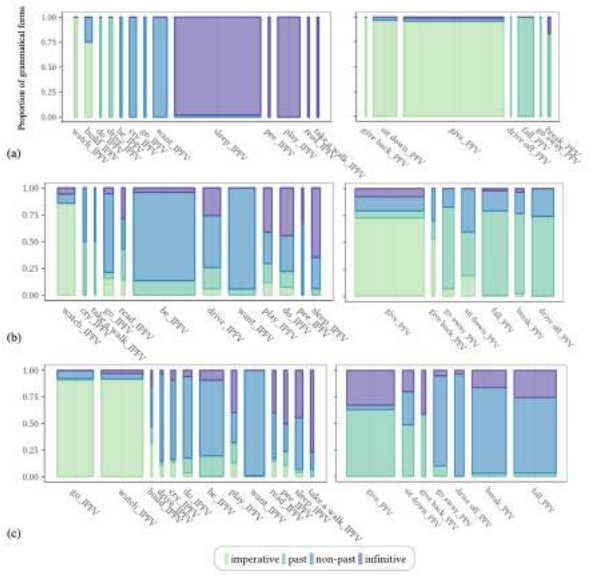


Figure 2: (a) Distribution of verb tokens in the production of Child 1 during phase 1; (b) Distribution of tokens in a sample of the same lemmas and the same number of tokens in Child 1's production during phase 2; (c) Distribution of tokens in a sample of the same lemmas and the same number of tokens in adults' production during phase 1.

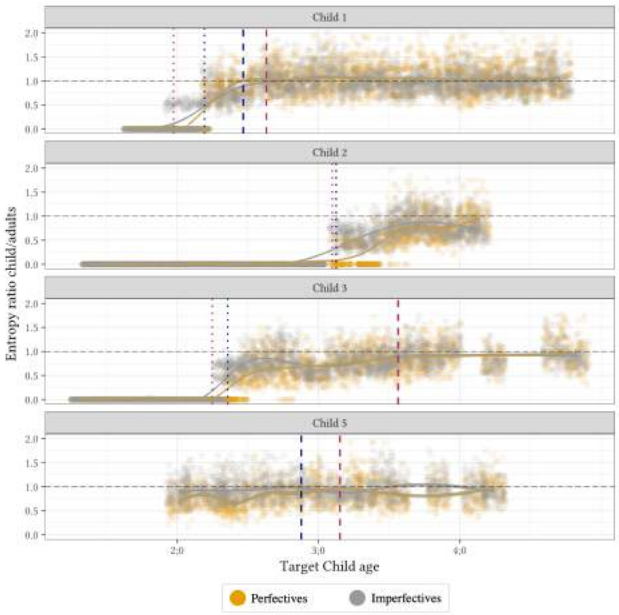


Figure 3: Entropy ratio (children/adults) of lemmas used with non-past marking on perfective and imperfective verbs, computed on bootstrapped data (100 iterations per session). Blue lines indicate changes in entropy increase for perfectives, red lines for the imperfectives.

than a month between the beginning of increased flexibility of perfectives+past and imperfectives+past.

Figures 3 and 4 show the increase of entropy ratios over time as well as the break points of the segmented regression. The break points are indicated by vertical lines: the dotted line represents the point at which the use of lemmas starts to become more diversified and the dashed line indicates the point at which the increase in diversification levels off.

Discussion

The children in our sample started out using verb forms in an item-specific manner during the first phase of verb production. Few verb lemmas appeared with more than one grammatical marker. However, while we did observe some tendencies that are in agreement with the Aspect Hypothesis, there is very little evidence that would make it a compelling argument by itself. Even during the early verb use we encounter past forms used with imperfective verbs. Child 1, for example, uses verbs such as *drive* and *do* with past marking early on; they are semantically activities without an endpoint and not expected to occur with past marking in the earliest phase of verb production. A plausible explanation in this case is the fact that Russian past forms are both quite transparent and appear in a narrower array of contexts, which eases acquisition and generalization.

The onset of the increase in flexibility of use of lemmas with individual grammatical forms is temporally close for verbs of both aspects. This suggests that, rather than passing

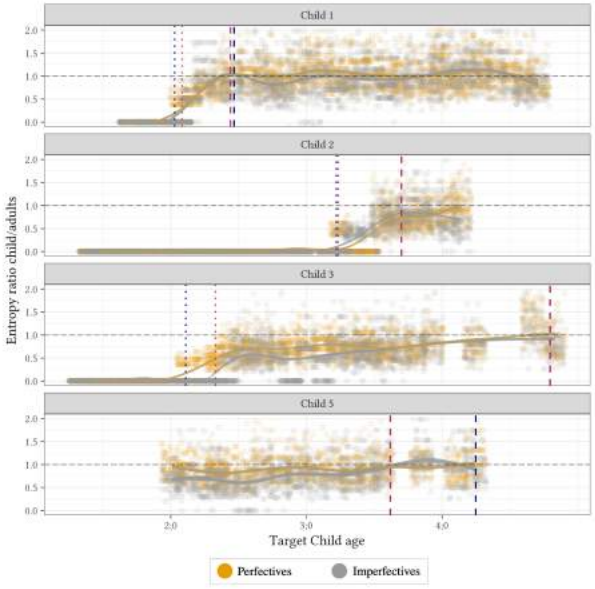


Figure 4: Entropy ratio (children/adults) of lemmas used with past marking on perfective and imperfective verbs, computed on bootstrapped data (100 iterations per session). Blue lines indicate changes in entropy increase for perfectives, red lines for the imperfectives.

through an intermediary phase during which the generalization first occurs within subdivisions of the verb paradigm (for perfective verbs with past marking, for imperfective verbs with non-past), the generalization starts early across the entire paradigm. Soon after item-specificity starts decreasing, children begin applying forms of a grammatical category to verbs of both aspects. This is strengthened by the observation that verb use in the first phase of production shows a stronger distributional bias in the distribution of tokens than in that of types. Coupled with the observation that the same holds for adult production — albeit in a weaker form — this finding suggests that the patterns of aspect-tense combinations found in the literature might be a mirroring of adult distributional patterns. Supporting this view is the fact that hardly any of these studies took the diversity of forms into account and thus have mostly confirmed the Aspect Hypothesis for the preferred use of forms, while making a less firm statement about availability of different forms at any stage of development. Given that distributional bias also factors into adult speech, it is important not to overstate the effect of preferred aspect-tense combinations on learnability of forms in the paradigm. Since children are able to pick up on distributional cues, their initial use of forms might simply be a reflection of the distributions found in adults as well as personal needs (cf. Figure 2b and the large proportion of the imperative form of *give*). A similar observation was already made by one of the authors of the Aspect Hypothesis Shirai (1998), who found that Japanese children do not follow the predictions of the Aspect Hypothesis and, therefore, suggested that multiple factors should be taken into account when examining early acquisition of tense-aspect morphology.

By looking at the use of different lemmas with the individual grammatical forms and thus measuring how flexibly a form is used, we were able to show that the development of form use might be more advanced than indicated by preferential use of certain tokens which skew the distributions. Going forward, it is important to disentangle the issue of tense-aspect marking further and take into account the differences between token and type distributions as well as further factors, such as lexical development and underlying distributions of grammatical markers in individual languages.

Acknowledgments

This work was supported by the European Research Council (ERC Consolidator Grant, ACQDIV 615988, to S. Stoll).

References

Aksu-Koç, A. A. (1998). The role of input vs. universal predispositions in the emergence of tense-aspect morphology: evidence from turkish. *First Language*, *18*, 255-280.

Antinucci, F., & Miller, R. (1976). How children talk about what happened. *Journal of Child Language*, *3*, 167-189.

Bar-Shalom, E. (2002). Tense and aspect in early child russian. *Language Acquisition: A Journal of Developmental Linguistics*, *10*(4), 321 - 337.

Bloom, L., Lifter, K., & Hafitz, J. (1980). Semantics of verbs and the development of verb inflection in child language. *Language*, *56*, 386-412.

Bronckart, J.-P., & Sinclair, H. (1973). Time, tense and aspect. *Cognition: International Journal of Cognitive Psychology*, *2*, 107-130.

Clark, E. V. (1996). Early verbs, event-types, and inflections. *Children's Language*, *9*, 61–73.

Comrie, B. (1976). *Aspect*. London: Cambridge University Press.

Gagarina, N. (2000). The acquisition of aspectuality by russian children: the early stages. *ZAS-Papers in Linguistics*, *15*, 232-246.

Harner, L. (1981). Children talk about the time and aspect of actions. *Child Development*, *52*, 498-506.

Johnson, B. W., & Fey, M. E. (2006). Interaction of lexical and grammatical aspect in toddlers' language. *Journal of Child Language*, *33*(02), 419-435.

Li, P. (1990). *Aspect and aktionsart in child Mandarin* (Doctoral Thesis).

Li, P. (2000). The acquisition of lexical and grammatical aspect in a self-organizing feature-map model. In *Proceedings of the 22nd annual meeting of the cognitive science society.* Mahwah, NJ: Lawrence Erlbaum.

Li, P., & Bowerman, M. (1998). The acquisition of lexical and grammatical aspect in chinese. *Journal of Child Language*, *54*, 311-350.

Li, P., & Shirai, Y. (2000). *The acquisition of lexical and grammatical aspect*. Berlin/New York: Mouton de Gruyter.

Lieven, E. V., Pine, J. M., & Baldwin, G. (1997). Lexically-based learning and early grammatical development. *Journal of Child Language*, *24*, 187–219.

Lin, J. (1991). Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, *37*(1), 145–151.

Pine, J. M., & Lieven, E. V. (1997). Slot and frame patterns and the development of the determiner category. *Applied Psycholinguistics*, *18*, 123–138.

R Core Team. (2015). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org`

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379–423.

Shirai, Y. (1998). The emergence of tense-aspect morphology in japanese: universal predisposition? *First Language*, *18*(54), 281–309.

Shirai, Y., & Anderson, R. W. (1995). The acquisition of tense-aspect morphology: a prototype account. *Language*, *71*, 743-762.

Shirai, Y., Slobin, D. I., & Weist, R. M. (1998). *The acquisition of tense-aspect morphology* (Vol. 18). Alpha Academic.

Stephany, U. (1985). *Aspekt, Tempus und Modalität: Zur Entwicklung der Verbalgrammatik in der neugriechischen*

Kindersprache. [Aspect, tense, and modality: The development of grammar in young greek children]. Tübingen, Germany: Gunther Narr.

Stoll, S. (1998). The role of aktionsart in the acquisition of russian aspect. *First Language*, *18*, 351-378.

Stoll, S. (2005). Beginning and end in the acquisition of the russian perfective aspect. *Journal of Child Language*, *32*, 805-825.

Stoll, S., & Gries, S. (2009). How to measure development in corpora? an association-strength approach to characterizing development in corpora. *Journal of Child Language*, *36*, 1075-1090.

Stoll, S., & Meyer, R. (2008). *Audio-visional longitudinal corpus on the acquisition of Russian by 5 children.*

Timberlake, A. (2004). *A reference grammar of Russian*. Cambridge University Press.

Tomasello, M. (2000). First steps toward a usage-based theory of language acquisition. *Cognitive Linguistics*, *11*(1/2), 61–82.

Tomasello, M. (2003). *Constructing a language: a usage-based theory of language acquisition*. Harvard, MA: Harvard University Press.

Weist, R. M., & Konieczna, E. (1985). Affix processing strategies and linguistic systems. *Journal of Child Language*, *12*, 27-35.

Weist, R. M., Wysocka, H., Witkowska-Stadnik, K., Buczowska, E., & Konieczna, E. (1984). The defective tense hypothesis: on the emergence of tense and aspect in child polish. *Journal of Child Language*, *11*, 347-374.

Pre-exposure and learning in young children: Evidence of latent inhibition?

R.P. McLaren, I.P.L. McLaren & C. Civile
School of Psychology, University of Exeter, UK
Correspondence to i.p.l.mclaren@exeter.ac.uk

Abstract

Previous research by Kaniel & Lubow in 1986 found that young children (aged 4-5 years) exhibited poorer learning (latent inhibition) to pre-exposed stimuli than older children (aged 7-10 years). The aim of our research was to develop a computer-based, child-friendly study that would replicate the work of Kaniel & Lubow. Sixty-three children took part in our experiment. This consisted of a pre-exposure/study phase in which participants were asked to press computer keys in response to clipart pictures of animals and dinosaurs. Each animal or dinosaur picture was preceded by one of two "warning signals" which acted as the pre-exposed stimuli (to which no response was required). In the test phase that followed, the participants had to either press the spacebar or withhold their response to each pre-exposed stimulus and two novel stimuli. They learnt which response was correct by trial and error using the feedback provided. The accuracy and reaction time of the responses during the test phase were analysed and indicated that the youngest children showed significantly lower mean accuracy and longer mean response times to the pre-exposed stimuli than to stimuli they had not been pre-exposed to. In contrast, the older children showed no significant differences in their responses to pre-exposed and novel stimuli. These results are consistent with those found by Kaniel & Lubow and could be taken as evidence for latent inhibition in young children. Further studies are proposed in which variations in pre-exposure procedure are used to rule out explanations based on response inhibition or negative priming.

Introduction

Learning from experience takes place when connections or associations are formed between stimuli and outcomes, e.g. pricking a finger on a needle results in one learning that needles are sharp, so care is needed when handling them. Latent inhibition (LI) occurs as a result of being exposed to a stimulus without a noticeable outcome. For instance, in the laboratory, latent inhibition is observed when rats that have been pre-exposed to a tone are slower to learn that the tone will subsequently indicate a reward (such as food), than rats that had not previously been exposed to the tone, (Lubow and Moore, 1959; for an example with rats see McLaren et al, 1994). LI is relatively easy to find in animals but it is, by comparison, difficult to find evidence for this effect in humans.

In their review of human LI experiments, Byrom et al (2018), suggest that none of them provide sufficient evidence to conclude that pre-exposure to a stimulus is the sole reason for the retarded responding observed. Other factors, such as negative priming (see Tipper, 1985 and Graham and McLaren, 1998), learned irrelevance or relative novelty could also be responsible for their findings. In order to provide a true test of LI, it is necessary to develop human experiments that are able to rule out these potentially confounding factors.

One study that appears to provide evidence for LI in humans is that by Kaniel & Lubow (1986). In their study, there was a simple Study Phase task in which children had to press buttons in response to pictures of plants and animals presented on metal cards in a box divided into three compartments. The cards were presented in sets of three, with one animal card and one plant card on each side of a third card (depicting two different sized black or white squares). During each trial the cards on either side of the middle card were changed and the child had to press a button corresponding to the side on which, for instance, the plant was present. In the following Test Phase, the children were presented with sets of cards showing black or white squares. This time they had to learn to press a button on the side corresponding to the card depicting the square that they had previously been exposed to in the study phase. They found that children aged 4-5 years exhibited poorer learning in this test than older children (7-10 year olds).

Can we take this as evidence of latent inhibition in young children? In one sense, the procedure used in Kaniel and Lubow's experiment is an example of simple exposure to the square stimuli, as they are presented at central fixation. If we accept this, then this may indeed be an example of latent inhibition in young children. On the other hand, the requirement for the children to respond to the pictures of plants or animals could have acted as a masking task during the study phase and diverted their attention from the pre-exposed black or white square stimuli. If this is the case, then an explanation in terms of conditioned inattention to the stimuli (i.e. negative priming, see Graham and McLaren, 1998) would be preferred. One argument against the latter explanation, however, is that the effect is confined to just the youngest group of children. Given that masking task procedures can successfully produce retarded learning in adults (see Ginton, Urca and Lubow, 1975 for an early demonstration of this in the auditory modality as well as Graham and McLaren, 1998 for an example using visual stimuli), why would only the 5 year old children show the effect in this case? For these reasons, this Kaniel and Lubow's results are some of the most interesting and potentially consequential for theories of learning that we are aware of.

This study has, to our knowledge, never been successfully replicated. Our aim was to design an updated and improved version of the Kaniel & Lubow study to see if we could replicate its findings, but without there even being a hint of a masking task involved. Our study uses clipart pictures of animals and dinosaurs for the children to respond to, one computer key for each. Instead of the pictures of different sized squares, we use four simple patterns as our pre-exposed stimuli. Two of the patterns are presented in the study phase as "warning signals" prior to an animal or dinosaur appearing. In the test phase, all four patterns are presented and the participants have to learn to either respond or withhold their response to each pattern. This design brings with it a number of advantages over Kaniel and Lubow's original. Because the stimuli being pre-exposed are

used as warning signals and are not present at the same time as the choice stimuli during the pre-exposure phase, participants do not have to ignore them and focus on the relevant stimuli. And, because we use both pre-exposed and non-pre-exposed stimuli in both conditions (respond and withhold response) in our test, we can see whether any learning deficit depends on whether people have to learn to respond to that stimulus or not.

Experiment

Method

Participants

Sixty-three primary school children took part in the experiment. The number of participants in each age group was as follows: 4-5 year olds (13), 6-9 year olds (40), 10-11 year olds (10). The children were all from a primary school a few miles outside Exeter, Devon.

Materials and Design

The experiment consisted of a **pre-exposure/study phase** of 120 trials (in random order) in which the participant had to respond to clipart pictures of dinosaurs and animals (examples in Fig. 1), each preceded by a "warning signal" (Fig. 2).

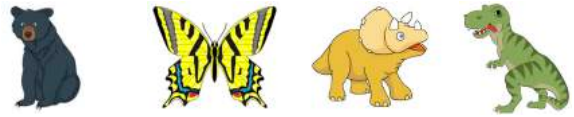


Figure 1. Examples of clipart images (300 x 300 pixels) of animals and dinosaurs presented during the pre-exposure/study phase of the experiment.

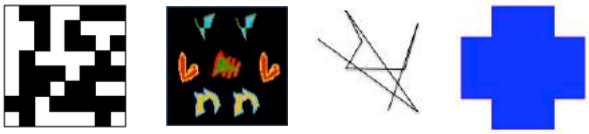


Figure 2. "Warning signal" stimuli (128 x 128 pixels). Two stimuli appeared during the pre-exposure/study phase; all four appeared during the test phase.

During the study phase, two of the stimuli shown in Fig. 2 were presented one at a time to provide a warning that the next animal or dinosaur stimulus was about to be displayed. Each warning stimulus appeared equally often preceding each choice stimulus.

The study phase was followed by a **test phase** of 32 trials in which the participant had to learn to either press the spacebar or withhold their response to each of four stimuli, two of which had been pre-exposed during the study phase. The two stimuli for which a spacebar response was required included one of the pre-exposed stimuli and one of the novel controls, and likewise for the stimuli for which the response had to be withheld. Stimuli were counterbalanced across conditions and subjects by creating four versions of the study (see Table 1).

Table 1. Counterbalance for stimuli pre-exposed during the study phase and responses required in the test phase of each version of the experiment (+ = press spacebar, - = withhold response).

Counterbalance	Study phase stimuli	Test phase response
1	A and C	A+, B-, C-, D+
2	A and C	A-, B+, C+, D-
3	B and D	A+, B-, C-, D+
4	B and D	A-, B+, C+, D-

The experiment was developed using SuperLab 4 software (version 4.0.7b) and was presented on a Macintosh laptop computer.

Procedure

Written consent was obtained from the parents/guardians of the children before they took part in the experiment. The consent form included information on the procedure of the experiment and the participants' right to withdraw at any time.

The experimenter worked with one participant at a time. At the start of the experiment the computer screen showed a picture of an imaginary island with a cartoon child "explorer". Overlaying the picture were written instructions. For each participant, the experimenter read the onscreen instructions out loud, as follows:

Welcome to our study.
Imagine you have just arrived on an island that has never been explored before.
Your job is to look for animals.
You soon find out that some animals look just like dinosaurs. Could this be possible?
Have dinosaurs somehow managed to survive on this remote island?
You need to quickly and accurately record every dinosaur and animal you see.
Press the 'x' key if you see a dinosaur.
Press the '.' key if you see an animal that isn't a dinosaur.
The computer will say 'yiha' if you get it right or 'oops' if you get it wrong.
Try to get as many correct responses as you can.
Please press the 'B' key to see some more instructions.

Before you see them there will be a signal to warn you that the animal or dinosaur is coming!
Remember:
- as soon as you see a dinosaur, press the 'x' key.
- as soon as you see an animal that isn't a dinosaur, press the '.' key.
When you're ready, press the 'B' key

Pre-exposure/study phase: There were 120 trials in two blocks of 60 with a participant break (self-timed) at the end of the first block.

Each trial consisted of a fixation cross (500ms) followed by a warning signal (1500ms) followed by a dinosaur/animal image (up to "x" or "." response, or 2000ms if no response). Feedback was given in the form of a "yiha" sound (correct response) or "oops" sound (incorrect response). If there was no response within the time-limit of 2000ms the feedback (presented on screen) was 'Oops – you took too long!'. Figure 3 shows an example of a trial

sequence during the pre-exposure/study phase.

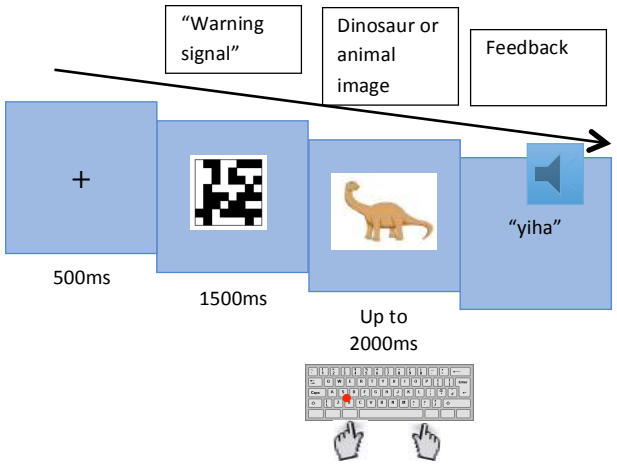


Figure 3. Example of a trial sequence during the pre-exposure/study phase.

During this phase, each of the two warning signals (pre-exposed stimuli) was presented 60 times in random order (equally preceding the animal or dinosaur stimuli). Participants were not required to respond to the pre-exposed stimuli.

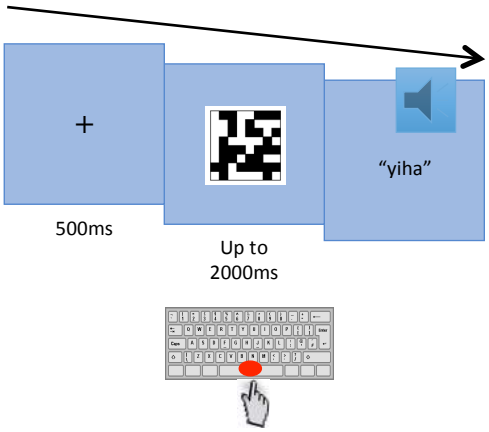
At the end of this phase, the following instruction screen was presented. Again, the experimenter read these instructions out loud when working with children.

Thank you. You have recorded all the dinosaurs and animals on the island.
Now the computer is going to show you some patterns.
These patterns were used to label the island by people who used to live there a long time ago.
Some parts of the island are safe to enter but others may be dangerous!
You need to mark which parts are safe – you do this by pressing the "spacebar"
And which ones aren't safe – for these don't press the "spacebar".
You will just be guessing to start with. Try pressing and not pressing the "spacebar" when you see a pattern and see what happens.
The computer will say "yiha" if you get it right or "oops" if you get it wrong.
Please press the 'B' key to begin.

Test phase: There were 32 trials in two blocks, with a participant break (self-paced) after the first 16 trials. Accuracy and reaction time were recorded for each trial during the test phase.

During this phase, each trial consisted of a fixation cross (500 ms) followed by one of the four stimuli (shown in Fig. 2), presented in a random order. These stimuli remained on screen up to the spacebar response or until 2000ms had elapsed if no response was made. If the spacebar was pressed, feedback ("yiha" or the "oops" sound) was provided immediately. If no response was made, feedback ("yiha" or "oops" sound) was provided after 2000ms. This enabled participants to learn, by trial and error, which type of response was required for each stimulus. Figure 4 shows examples of two trials (one requiring the "spacebar" response, and the other requiring no response) during the test phase. Each stimulus (two pre-exposed during the study phase and two novel stimuli) appeared 8 times. Two of the stimuli (one pre-exposed and one novel) required the "spacebar" response. Two stimuli (one pre-exposed and one novel) required the response to be withheld (i.e. no response).

A) Response = "spacebar" press



B) Response = withhold response

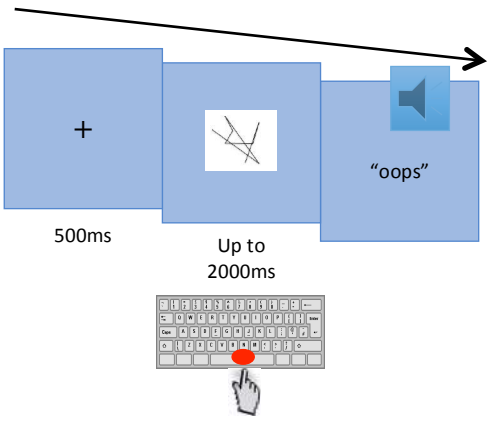


Figure 4. Examples of two Test phase trials in which the "spacebar" was pressed: Trial A required the "spacebar" response so feedback is "yiha". Trial B required the response to be withheld but a spacebar press was made so feedback is "oops".

Results

The accuracy and reaction time data collected during the first block of the test phase were analysed using t-tests to establish whether there was a significant difference between responses to the stimuli that had been pre-exposed during the study phase compared to the novel stimuli, and whether this was dependent on the age of participants. A significance level of $p = .05$ was used for all statistical tests, which were two-tailed unless otherwise specified. Only data from the first block of the test phase were analysed as, by the second block, most children had reached 100% accuracy.

The 4-5 year-old children were the only age group to exhibit significantly lower overall accuracy of responding (averaged over go = spacebar press and nogo = withheld response) to the pre-exposed stimuli than to the novel stimuli, $t(12) = 3.57$, $p = .004$ (see Figure 5). This finding is consistent with a latent inhibition effect in the youngest children, and consistent with Kaniel and Lubow's (1986) findings. The size of the effect was significantly greater than that observed in the oldest children, $t(21) = 2.25$, $p = .035$; and this difference also approached significance when the youngest and the middle age groups were compared, $t(51) = 1.93$, $p = .059$. This also replicates Kaniel and Lubow's (1986) findings.

As would be expected, mean response accuracy tended to increase with the children's age. The mean response accuracy of the oldest children (10-11 year olds) was significantly greater for both the pre-exposed $t(21) = 4.51$, $p < 0.001$, and novel $t(21) = 3.18$, $p = 0.005$, stimuli when compared with the youngest children. Only the difference for the pre-exposed stimuli is significant when comparing the middle group to the youngest, $t(51) = 2.34$, $p = 0.023$.

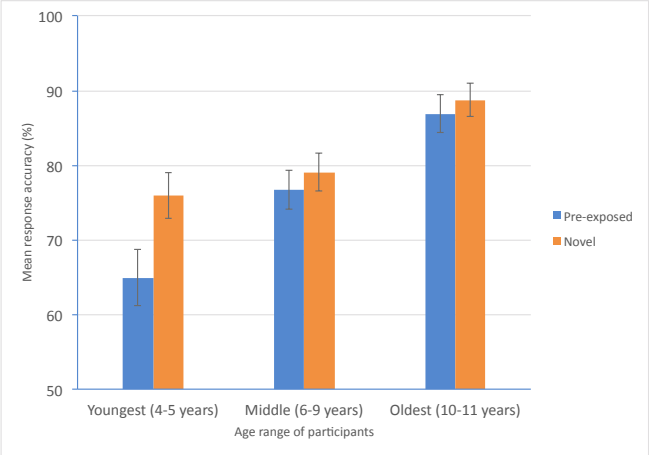


Figure 5. Mean percentage response accuracy (averaged over go and nogo stimuli) for the pre-exposed and novel stimuli for each age group during the test phase. Error bars show SE of the mean.

Figure 6 focuses on the response accuracy for those stimuli requiring a spacebar press (Pre-exposed + and Novel +). In this case, the 4-5 year old and 6-9 year old children both show a significant difference in their response accuracy ($t(12) = 5.50$, $p < 0.001$ and $t(39) = 2.40$, $p = 0.021$ respectively) with a tendency to respond less accurately to the pre-exposed stimuli. In contrast, the older children show no reliable difference in their spacebar response accuracy to the pre-exposed and novel stimuli.

Once again we can look at the differences between groups on this measure. There isn't a significant difference when comparing the oldest to the youngest children (even though numerically the difference is large, this is probably a matter of power), but there is a trend towards significance for the comparison between the middle group and the youngest children, $t(51) = 1.73$, $p = 0.09$, a result that would be significant on a 1-tailed test. There is some evidence, then, that the poorer learning exhibited is in part, at least, due to difficulty in learning to respond to the pre-exposed stimulus that requires a response. There was no sign of such an effect for the pre-exposed stimulus that did not require a response.

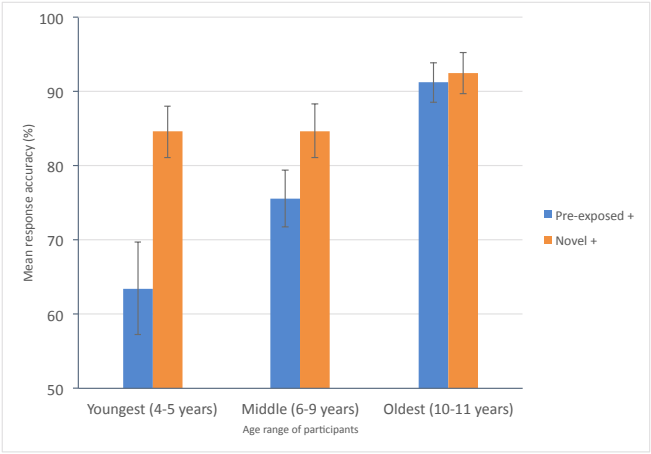


Figure 6. Mean percentage response accuracy for the stimuli requiring a spacebar press for each age group during the test phase. Error bars show SE of the mean.

The mean response times for stimuli requiring a spacebar press (Fig. 7) were significantly longer for the pre-exposed stimuli than the novel stimuli for both the 4-5 year old children ($t(12) = 2.51$, $p = 0.027$) and the 6-9 year old children ($t(39) = 2.21$, $p = 0.033$) but not in the oldest age group. But this difference did not itself differ significantly across groups, despite the extra time taken to pre-exposed stimuli being considerably greater numerically in the youngest children than in the other two groups. As would be expected, the youngest children generally exhibited longer response times for both types of stimulus than the oldest children.

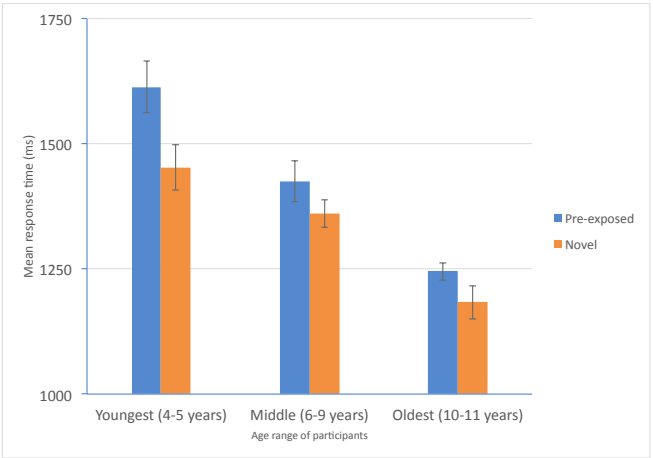


Figure 7. Mean response times (msec) for the pre-exposed and novel stimuli for each age group during the test phase. Error bars show SE of the mean.

General Discussion

The aim of this study was to replicate Kaniel and Lubow's (1986) findings using an updated method that avoided the need to ignore the pre-exposed stimuli while performing the initial task. In this we succeeded. There is really quite strong evidence in our data for retarded learning as a consequence of pre-exposure in our youngest group of children, and this

is the same age group that Kaniel and Lubow obtained their effect with. We have also failed to find a similar effect in older children, again mirroring Kaniel and Lubow's results, and all this with a pre-exposure procedure that uses the target stimuli as warning signals for an upcoming trial, so that there is no obvious need to ignore them. But can we be sure that this is latent inhibition in humans?

The answer to this question has, for the present, to be no. We cannot be sure that this is latent inhibition, but we can, perhaps, rule out some of the other possibilities. As we have argued, there is no particular reason to learn to ignore the pre-exposed stimuli during the initial phase of the experiment, because they actually serve a useful function, warning of the next stimulus to which a decision has to be made. One could always argue that the 4-5 year old children do learn to ignore these stimuli nevertheless, but that would seem a rather ad-hoc explanation of our results. And we would still be left with the conundrum of explaining why older children do not learn to ignore the pre-exposed stimuli.

But in adapting our design to control for possible artifacts in Kaniel and Lubow's study, we may have introduced new ones into our experiment. One plausible explanation for these results takes note of the fact that the larger effect on learning seems to be on the pre-exposed stimulus to which a response was required during the final, test phase. The young children were particularly bad at learning to press the spacebar in the case of the pre-exposed S+, and the middle age group also showed an impairment in learning. Perhaps encountering the stimuli during the initial pre-exposure phase when responses were required to the animal/dinosaur pictures and not having to make a response to the pre-exposed stimuli (because they were used as warning stimuli) has somehow caused this effect?

We can imagine at least two versions of this account. One would have it that being presented with the stimulus, followed by no outcome led to a type of CS->NoUS learning that has been suggested as producing the basic latent inhibition effect in other animals. Learning that this stimulus signals no outcome makes it harder to learn that an outcome does follow later. If this is the mechanism, then it would support the contention that the 4-5 year old children are displaying latent inhibition, as well as providing evidence for a particular theoretical explanation of latent inhibition.

A somewhat more concrete and specific version of this account would appeal to response inhibition developing rather than learning some general CS->NoUS association during pre-exposure. In a context where responses have to be made (press one of two keys), when the warning signal is shown no response is required and so general response inhibition accrues and is associated with the stimuli present at the time. As a consequence, when a response is required to these pre-exposed stimuli, it is harder to learn and perform. This explanation can be distinguished from our earlier one by noting that the result for latent inhibition is that learning of both an excitatory association and of an inhibitory association between CS and US is retarded for a CS that has undergone latent inhibition. But the response inhibition account would predict that learning to withhold a response to a CS would actually be facilitated. The question, then, is how learning to respond to the pre-exposed S- progresses in the last, test phase. The answer in our data is that there is no evidence of a facilitatory effect in the youngest or oldest children, and there is only a hint of one in the middle group ($t(39) = 1.69$, $p = .099$. Given this, a response inhibition account of the poorer learning seen in the youngest children seems unlikely. The fact that we have an effect in our youngest age group for overall performance is also indicative of an effect that is not based on response inhibition.

Perhaps the most important argument for this being a demonstration of latent inhibition in young children, however, is generated by considering the two experiments, Kaniel and Lubow's and ours, in combination. A response inhibition explanation will not obviously apply to Kaniel and Lubow's design, as a response is made while the pre-exposed stimuli are on screen. A learned inattention or negative priming explanation cannot easily be applied to our results because there is no reason to ignore the pre-exposed stimuli. But both experiments give very similar results, which suggests a common explanation for those results, and the only one that seems to fit is latent inhibition.

Which brings us to what may be the most intriguing feature of these results. The younger children, 4-5 years old, are the ones that show the effect. The older children either do not show any significant effect, or display a significantly weaker version of it. This is also something our study shares in common with Kaniel and Lubow's original work and needs some explanation. The explanation given in Lubow's 1989 book "Latent Inhibition and Conditioned Attention Theory" is that this "raw" latent inhibition found in young children is actually present in older children and adults, but that they have compensatory attentional processes that obscure this effect in studies of this type. In essence, latent inhibition reduces learning, but then attention is deployed to take it back up to its original level, hence no difference is observed between pre-exposed and non-pre-exposed conditions.

There is much to commend in this explanation, and one of us has offered something that at first sight is similar in McLaren, Wills and Graham (2010). But there are real differences, stemming from the fact that our account of latent inhibition (which can be found in its earliest form in McLaren, Kaye and Mackintosh [MKM], 1989, and has been updated in McLaren and Mackintosh, 2000, and McLaren, Forrest and McLaren, 2012) differs from that offered by Lubow. In Lubow's account, latent inhibition is due to conditioned inattention, but in ours it is due to a reduction in salience due to the features of the pre-exposed stimuli becoming predicted either by other stimuli present, or by one another. This leads to a reduction in salience (learning rate) for these pre-exposed features, hence latent inhibition. Instead, we use conditioned attention to explain why simple pre-exposure does not lead to observable latent inhibition in older children and adults. We argue that people attend to stimuli that are placed in front of them, and that this attentional response then becomes linked to those stimuli, compensating for any effect of latent inhibition. This attentional response is absent in younger children, which produces our and Kaniel and Lubow's results.

Why should we prefer this explanation to Lubow's? Both explanations are viable for the results obtained here and in Kaniel and Lubow's original study. But our explanation has the advantage of being able to explain Graham and McLaren's (1998) results as well as other demonstrations of retardations in learning in adults using a "masking" task (e.g. Ginton, Urca and Lubow, 1975). We argue that these results are indeed due to conditioned inattention, just as Lubow would have it, but disagree that this is the basis for latent inhibition. The test that Graham and McLaren use is to create two distortions of a pre-exposed stimulus and then train a discrimination between them. In their 1998 paper, they find that this results in slow learning of the discrimination when it is compared to a similar discrimination based on distortions of a novel stimulus. This is the opposite result to that found in animals (see Aitken, Bennet, McLaren and Mackintosh, 1996 for direct evidence on this point), and so suggests that the retardation in learning observed when the pre-exposed stimulus is trained directly is not actually latent inhibition but instead is due to negative priming. In future, we intend to apply this test to our finding. If we are able to demonstrate an enhancement of learning between two distorted versions of the pre-exposed stimulus (i.e. perceptual learning) using this technique then this will be excellent evidence that the effect is the same as that seen with simple pre-exposure in other animals, and confirm that it is latent inhibition and not negative priming.

Conclusions

In conclusion, this study has provided strong evidence for a retardation in learning to a stimulus following pre-exposure to that stimulus in young (4-5 year old) children. This effect was not found in older children. It is possible that what we have here is latent inhibition of the type obtained with simple pre-exposure in animals such as the rat, but more work will be needed to establish whether this is, in fact, the case. Possible alternative explanations are conditioned inattention / negative priming, and generalised response inhibition, but neither receive a great deal of support from the data we have obtained. Further research should focus on either definitively ruling these alternatives out, or providing solid evidence for them.

If we have demonstrated latent inhibition in young children, then this has important implications for theories of learning, particularly in humans. It would confirm that we carry with us the same basic processes affecting learning as other animals, and would also go some way to confirming the MKM model of perceptual learning. More than that, it would also raise the question of why latent inhibition "goes away" in older children. We have given one possible reason here, which offers us one perspective on the development of learning and cognition in children. If it turns out not to be the case, and our results can be explained by some other mechanism, then this problem will still remain. Why do young (4-5 years old) children show this effect and older children do not? Solving this developmental puzzle will add to our understanding of human mental life.

References

Aitken, M.R.F., Bennett, C.H., McLaren, I.P.L., & Mackintosh, N.J. (1996). Perceptual differentiation during categorisation learning by pigeons. *Journal of Experimental Psychology: Animal Behaviour Processes*, 22 (1), 43- 50.

Byrom, N.C., Msetfi, R.M. & Murphy, R.A. (2018). Human latent inhibition: Problems with the stimulus exposure effect. *Psychonomic Bulletin & Review.*

Ginton, A., Urca, G. and Lubow, R.E. 1975. The effects of pre-exposure to a non-attended stimulus on subsequent learning: Latent inhibition in adults. *Bulletin of the Psychonomic Society*, 5 (1), 5-8

Graham, S., & McLaren, I.P.L. (1998). Retardation in human discrimination learning as a consequence of pre-exposure: Latent inhibition or negative priming? *The Quarterly Journal of Experimental Psychology: Section B, 51(2),* 155-172.

Kaniel, S., & Lubow, R. E. (1986). Latent inhibition: A developmental study. *British Journal of Developmental Psychology; 4,* 367-375.

Lubow R.E., Moore, A.U. (1959) Latent inhibition: The effect of nonreinforced preexposure to the conditional stimulus. *Journal of Comparative and Physiological Psychology*, 52: 415-9.

Lubow, R.E. (1989). *Latent Inhibition and Conditioned Attention Theory.* CUP.

McLaren, I.P.L., Kaye, H., Mackintosh, N.J., (1989). An associative theory of the representation of stimuli: applications to perceptual learning and latent inhibition. In: Morris, R.G.M. (Ed.), *Parallel Distributed Processing - Implications for Psychology and Neurobiology*. Oxford University Press, Oxford.

McLaren, I.P.L., Bennett, C., Plaisted, K., Aitken, M. and Mackintosh, N.J. (1994). Latent inhibition, context specificity, and context familiarity. *Quarterly Journal of Experimental Psychology*, 47B, 387-400.

McLaren, I.P.L. and Mackintosh, N.J. (2000). An elemental model of associative learning: I. Latent inhibition and perceptual learning. *Animal Learning and Behavior*, 38(3), 211-246.

McLaren, I.P.L., Wills, A.J. and Graham, S. (2010). Attention and perceptual learning. In C. Mitchell and M. Le Pelley. (Eds.). *Attention and associative learning: From Brain to Behaviour*. Oxford University Press.

McLaren, I. P. L., Forrest, C.L., and McLaren, R.P. (2012). Elemental representation and configural mappings: Combining elemental and configural theories of associative learning. *Learning and Behavior*, 40 (3), 320-33.

Tipper, S.P., 1985. The Negative Priming Effect: Inhibitory priming by ignored objects. *Quarterly Journal of Experimental Psychology* 37A.

Leveraging Thinking to Facilitate Causal Learning from Intervention

Yuan Meng
yuan_meng@berkeley.edu
Department of Psychology
University of California, Berkeley

Fei Xu
fei_xu@berkeley.edu
Department of Psychology
University of California, Berkeley

Abstract

Intervention selection is at once crucial in causal learning and challenging for causal learners. While the optimal strategy is maximizing the expected information gain (EIG), both children and adults often combine it with suboptimal ones such as the positive test strategy (PTS). In the current study, we sought to facilitate causal learning from intervention by asking 5- to 7-year-olds to explain why they chose a certain intervention to identify the true structure of a three-node causal system that might work in one of two ways. Our findings suggest that while engaging in self-explaining did not help children select more informative interventions, asking them to think about their intervention choices (explaining or reporting) might help them better utilize interventional data to infer causal structures.
Keywords: causal learning; intervention; explanation; learning by thinking

Once upon a time in China, two men were accused of a murder yet no evidence could be found. The judge gave each of them a "magical" straw that was said to grow longer in the hands of the guilty. As the story goes, the man showing up with a *shorter* straw next day was put in jail. As you might have guessed, straws don't grow; the real magic is that the judge chose the most informative intervention centuries before informative theory came into being. He could not foresee which man would cut his straw in fear but whoever did it must be the murderer. This strategy allowed him to maximally reduce his uncertainty averaged across potential outcomes, or in other words, maximize his *expected information gain* (EIG)

EIG is widely regarded as a normative model for inquiry selection (Coenen, Nelson, & Gureckis, 2018; Nelson, 2005) but it only partially captures people's actual interventions. Both adults (Bramley, Lagnado, & Speekenbrink, 2014) and children (McCormack, Bramley, Frosch, Patrick, & Lagnado, 2016) outperform models that intervene randomly but fall short of pure EIG maximization. On the computational level, a possible explanation is that adults (Coenen, Rehder, & Gureckis, 2015) and children (Meng, Bramley, & Xu, 2018) combine EIG maximization with a suboptimal strategy akin to the *positive test strategy* (PTS) in the rule learning literature (Klayman & Ha, 1989; Wason, 1960). In causal learning, Coenen et al. (2015) defined PTS as a tendency to generate the most expected effects under your current causal hypothesis. A minimal example of PTS is intervening on X when you try to discriminate between your hypothesis, $X \rightarrow Y \rightarrow Z$, and an alternative one, $Y \leftarrow X \rightarrow Z$. If the outcome (e.g., only X and Y are activated) happens to *falsify* your hypothesis, you get to rule it out; otherwise, both could still be true so you remain uncertain. By contrast, a high-EIG intervention (Y) reduces your hypothesis space (in this case, to 1) regardless of the outcome (all variables are activated or only Y and Z).

To ensure successful causal learning from intervention, learners should use an optimal strategy (e.g., EIG) to select interventions and make accurate inferences from interventional data. In the current study, we sought to facilitate both the intervention selection and the belief updating processes by prompting learners to *explain* why they choose a certain intervention to learn about an unknown causal system.

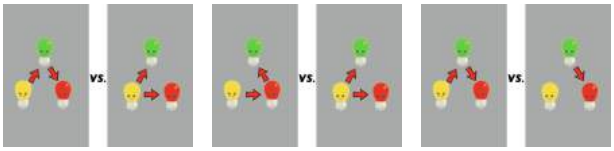
Explanation and intervention

Explaining requires no extra data or instructions; yet, it has profound downstream consequences for learning and inference in various domains (see Fonseca & Chi, 2011; Lombrozo, 2016, for reviews). Typically, learners achieve better learning outcomes simply by engaging in explanation (e.g., how a system works, why an effect occurred, etc.) even without feedback or generating accurate explanations (e.g., Chi, Bassok, Lewis, Reimann, & Glaser, 1989; Chi, De Leeuw, Chiu, & LaVancher, 1994; Walker, Lombrozo, Legare, & Gopnik, 2014; Walker, Lombrozo, Williams, Rafferty, & Gopnik, 2017). Many theories are proposed to explain why explaining facilitates learning, such as that it helps learners fill gaps in their knowledge, repair erroneous mental models, recruit criteria for "good" explanations (simplicity, breadth, or other "explanatory virtues") to constrain reasoning, etc..

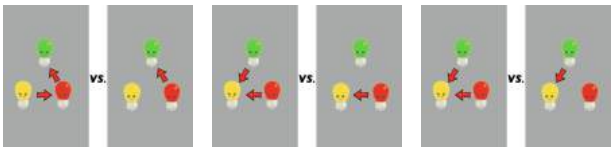
How do you explain an intervention? From an EIG perspective, to explain intervention selection, you must consider belief updating (Coenen & Gureckis, 2015): You choose an intervention because on average, it reduces the most uncertainty. Engaging in explanation may benefit both processes.

Explaining may facilitate intervention selection by promoting *comparison* and *abstraction*. A recent study (Edwards, Williams, Gentner, & Lombrozo, 2019) suggested that asking learners to explain exemplars' category membership (e.g., "Why is this robot a Glorp/Drent?") increased their comparison within and between categories. Should explainers compare more across different interventions and the outcomes of each intervention, they might be in a better place to select high-EIG interventions. Moreover, effective learners may realize that on an abstract level, informative interventions are ones that yield distinct outcomes under different hypotheses. Walker and Lombrozo (2017) found that explaining the outcome of a story (e.g., why a character is sad) helped children extract its underlying moral and go beyond the specifics. Should explaining help causal learners achieve such abstraction, it could largely reduce the cost of intervention selection.

Explaining may also facilitate belief updating by encouraging learners to apply their prior knowledge when interpreting interventional data (Williams & Lombrozo, 2013).



(a) Puzzle 1 (b) Puzzle 2 (c) Puzzle 3



(d) Puzzle 4 (e) Puzzle 5 (f) Puzzle 6

Figure 1: Light bulb puzzles used in the experiment.

Current study

In the current study, we investigated whether self-explaining could facilitate causal learning from intervention. We chose to test 5- to 7-year-olds because previous studies (McCormack et al., 2016; Meng et al., 2018) suggested that they were not yet able to reliably select informative interventions, leaving substantial room for improvement. This also allows us to compare our results directly to that in Meng et al. (2018).

Overview of experiments Our causal learning task was adapted from Meng et al. (2018). Children were tested on six unknown causal systems consisted of three light bulbs, some of which could turn on others if activated. Each system could work in one of two ways and children were allowed to turn on one light bulb to identify its correct structure. All causal connections were deterministic with no background noise.

In the first experiment (Experiment 1A), children were asked to explain their intervention choice ("Why did you turn on that light bulb?") after carrying it out. However, since those children observed the outcome before explaining, their explanation might be a *post hoc* justification of their choice ("Because it helped me solve the puzzle.") rather than the actual reason. To address this concern, we conducted a second experiment (Experiment 1B) where children pointed to the intervention they wanted to perform and were asked to explain their choice ("Why do you want to turn on that light bulb?") before carrying it out. In the respective control conditions, children were asked to *report* which intervention they carried out (Experiment 1A) or planned to choose (Experiment 1B).

Modeling intervention strategies To compare intervention strategies across conditions, we took a hierarchical Bayesian approach used by Coenen et al. (2015) and Meng et al. (2018). We compared models of three single strategies (EIG, PTS, and random selection) and a linear combination of EIG and PTS. Below is an overview of the four models.

Learners all begin with a set of causal hypotheses, each of which can be represented as a directed acyclic graph $g \in G$ (G is the space of possible graphs), or a *causal Bayesian network* (Pearl, 2000). In each graph, causal variables are presented as nodes and causal relationships as edges.

1. Expected information gain (EIG)

   The information gain (IG) after intervening on the node $n \in N$ is the difference between the initial entropy, $H(G)$, and the entropy conditioned on the outcome o, $H(G|n,o)$:

   $$IG(n,o) = H(G) - H(G|n,o). \tag{1}$$

   Since o is unknown, the expected information gain (EIG) over all possible outcomes O is used to estimate IG:

   $$EIG(n) = H(G) - \sum_{o \in O} P(o|n) H(G|n,o). \tag{2}$$

   Applying Shannon's entropy equation, we have

   $$H(G) = -\sum_{g \in G} P(g) log_2 P(g), \tag{3}$$

   and

   $$H(G|n,o) = -\sum_{g \in G} P(g|n,o) log_2 P(g|n,o). \tag{4}$$

   The prior probability $P(g)$ of each graph g is assumed to be equal and the posterior probability $P(g|n,o)$ is given by Bayes' rule, $\frac{P(o|g,n)P(g)}{\sum P(o|g,n)P(g)}$. $P(o|g,n)$ is the likelihood of an outcome o given a hypothesis g and an intervention n.

2. Positive test strategy (PTS)

   PTS manifests as the tendency to intervene the node $n \in N$ with the most of direct or indirect descendant links (normalize by the total number of links in each graph $g \in G$):

   $$PTS(n) = \max_g [\frac{DescendantLinks_{n,g}}{TotalLinks_g}]. \tag{5}$$

3. Random selection

   Random selection is equivalent to indiscriminately assigning the same value (e.g., 1) to all possible interventions.

4. Linear combination of EIG and PTS

   Rather than sticking to one strategy, learners may use multiple strategies such as EIG and PTS to select interventions. The value of each possible intervention is a linear combination of its EIG and PTS values (the weight of EIG is θ).

Under one strategy or another, each possible intervention is assigned a value $V(n)$. An ideal learner should always select the intervention with the highest value but due to noise τ in the decision process, an actual learner often does so probabilistically. According to the *softmax choice rule* (Luce, 1959), the probability that an intervention gets chosen, $P(n)$, is a function of its value $V(n)$ and the learner's decision noise τ:

$$P(n) = \frac{\exp(V(n)/\tau)}{\sum_{n \in N} \exp(V(n)/\tau)}. \tag{6}$$

When τ is 0, the learner selects interventions with the highest values; when τ approaches $+\infty$, they select randomly.

Experiments

Participants

Seventy-four 6- to 7-year-olds participated in Experiment 1A, 37 of whom were assigned to the Explanation condition (M = 85 months, range = 74–101 months, SD = 9 months) and 37 to the Report condition (M = 84 months, range = 64–96 months, SD = 8 months). Another forty-three 5- to 7-year-olds participated in Experiment 1B, 22 of whom were assigned to the Explanation condition (M = 77 months, range = 62–90 months, SD = 8 months) and 21 to the Report condition (M = 75 months, range = 50–101 months, SD = 14 months).

Equipment

Three light bulbs (yellow, green, and red) were presented on a laptop screen and controlled by three buttons of corresponding colors located on a response board. During practice, red arrows indicated the causal relationships among the light bulbs. During the test, the arrows were hidden but two possible structures were shown on two cards placed side by side.

Procedure

Both experiments included a familiarization phase, a practice phase, and a test phase. During familiarization, children were taught to use buttons on a response board to control light bulbs of corresponding colors on the computer. During practice, they saw four basic types of structures: Common Cause (*Yellow* ← *Green* → *Red*), Common Effect (*Yellow* → *Red* ← *Green*), Causal Chain (*Green* → *Red* → *Yellow*), and One Link (*Yellow* → *Red*). In Experiment 1A, the presentation order was randomized. For each structure, children decided when to turn on which light bulb and were asked to describe the outcome of each action. In Experiment 1B, each structure was one change apart from the previous one. The simplest structure, One Link, was presented first, which was followed by Causal Chain, Common Cause, and Common Effect. For each structure, children turned on the light bulbs in a designated order (*Yellow–Red–Green* in the first two trials and *Green–Red–Yellow* in the last two) and were asked to predict and then describe each action's outcome.

On each of the six test trials, children were shown two ways in which the three light bulbs might work and were told that they could only turn on one light bulb to find out the true structure. In Experiment 1A, children were asked to explain ("Why did you turn on that light bulb?") or report ("Which light bulb did you turn on?") the intervention that they had just carried out. In Experiment 1B, children were asked to first point to the light bulb they planned to turn on, then explain ("Why do you want to turn on that light bulb?") or report ("Which light bulb do you want to turn on?") their choice, and finally perform the intervention[1]. At the end of both experiments, children were asked to put a smiley face sticker on the correct causal structure. In order to avoid potential discouragement that we observed during piloting, feedback was only provided after the entire experiment.

[1] In the rare event that children's actual intervention differed from what they planned, we used the former for all our analyses.

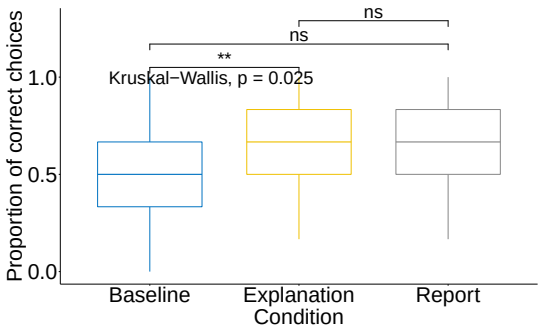


Figure 2: The proportion of causal structures that children correctly identified in each condition.

Results

Our initial analysis revealed no differences between the results of Experiments 1A and 2B, so data from these two experiments were pooled together in all subsequent analyses. To test whether explaining and reporting one's intervention choices could both influence causal learning, we used children in Meng et al. (2018) as our baseline. Apart from the additional explanation/report prompts, our procedure, stimuli, and population were identical to those in the previous study.

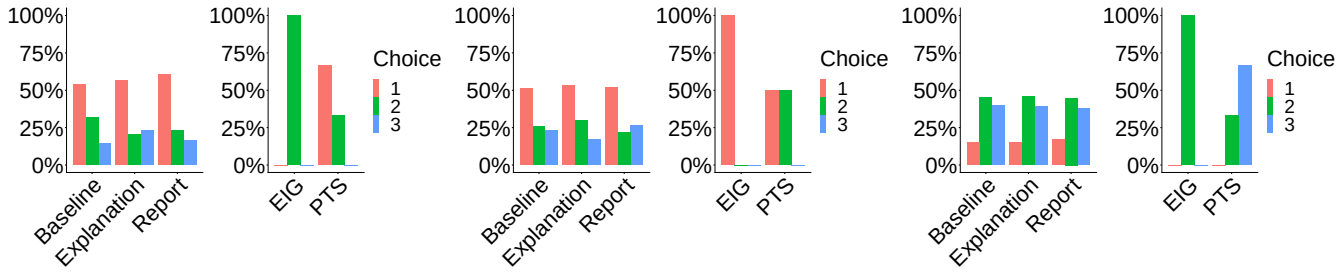
Inference accuracy To begin, we first looked at whether children were able to identify the correct causal structures in the end. As shown in Figure 2, those in the Baseline condition chose the correct structures 54% (SD = 22%) of the time, which was not distinguishable from chance (50%), $t(38)$ = 1.02, p = .31, Cohen's d = .16. However, children performed above chance in both the Explanation (M = 67%, SD = 25%) and the Report (M = 61%, SD = 23%) conditions, $t(58)$ = 5.13, $p < .001$, Cohen's d = .67 and $t(57)$ = 3.77, $p < .001$, Cohen's d = .50, respectively. The only significant difference between conditions was that explainers were more accurate than the baseline, $t(88.53)$ = 2.73, p = .007, Cohen's d = .55.

Intervention choices Before fitting models of intervention strategies, we examined children's interventions choices to see if they were random or biased towards EIG or PTS.

We compared the mean EIG and the mean PTS value of children's chosen interventions against the respective chance levels (.33 for EIG[2] and .55 for PTS[3]) of the two metrics. In the Baseline condition, only the mean PTS value (M = .74, SD = .22) was above chance, $t(38)$ = 5.37, $p < .001$, Cohen's d = .86, but not the mean EIG value (M = .39, SD = .28), $t(38)$ = 1.23, p = .23, Cohen's d = .20. Similarly in the Report condition, the mean PTS value (M = .74, SD = .20) was above

[2] Among all three possible interventions in each puzzle, only one was informative, i.e., having an EIG value of 1.

[3] This was the average PTS value across all interventions.



(a) Causal Chain vs. Common Cause (b) Causal Chain vs. One Link (c) Common Effect vs. One Link

Figure 3: Figures on the left show the proportion of children intervening on each node (n_1, n_2, and n_3) in each type of puzzles: (a) Causal Chain vs. Common Cause, (b) Causal Chain vs. One Link, and (c) Common Effect vs. One Link. Figures on the right show the probability of children intervening on each node in each type of puzzles predicted by EIG and PTS.

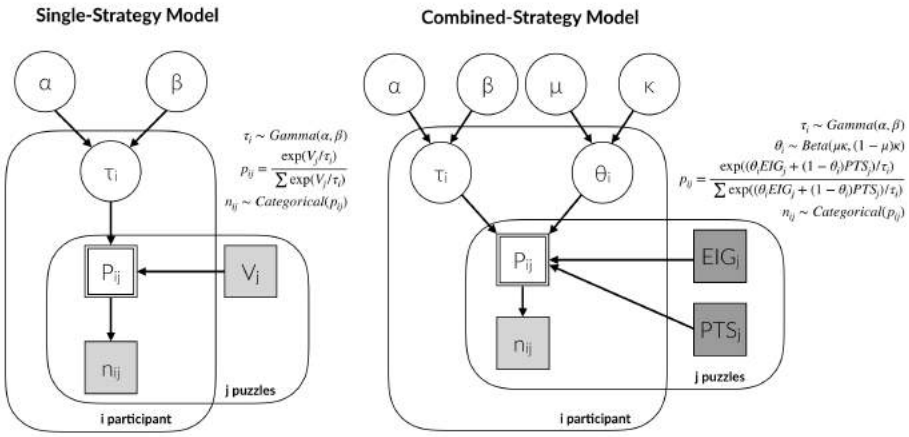


Figure 4: Hierarchical Bayesian models of single (left) and combined (right) strategies. In each puzzle j, each participant i chooses one node n_{ij} to intervene on. V_j, EIG_j, and PTS_j store the values of three possible interventions in each puzzle. p_{ij} stores probabilities of each participant choosing each intervention in each puzzle. τ_i and θ_i capture each participant's decision noise and weight of EIG. α and β are population-level hyper-parameters that generate τ_i; μ and κ generate θ_i.

chance, $t(57)$ =7.17, $p < .001$, Cohen's d = .94, but not the mean EIG value (M = .39, SD = .25), $t(57)$ = 1.63, p = .11, Cohen's d = .21. In the Explanation condition, however, both the mean EIG (M = .44, SD = .32) and the mean PTS (M = .75, SD = .18) value were above chance , $t(58)$ = 2.52, p = .014, Cohen's d = .33 and $t(58)$ = 8.4, $p < .001$, Cohen's d = 1.09, respectively. Neither the mean EIG or the mean PTS value differed significantly across conditions.

We also compared the proportion of children intervening on each node in each puzzle against what EIG and PT would predict. Since the mapping between node positions and light bulb colors is arbitrary, we re-coded Puzzles 1 and 2 as $n_1 \rightarrow n_2 \rightarrow n_3$ (Chain) vs. $n_2 \leftarrow n_1 \rightarrow n_3$ (Common Cause), Puzzles 3 and 4 as $n_1 \rightarrow n_2 \rightarrow n_3$ (Chain) vs. $n_2 \rightarrow n_3$ (One Link), and Puzzles 5 and 6 as $n_2 \rightarrow n_1 \leftarrow n_3$ (Common Effect) vs. $n_3 \rightarrow n_1$ (One Link). As Figure 3 shows, children deviated the most from EIG predictions in "Chain vs. Common Cause". In the other two types of puzzles, children's choices were split between EIG and PTS predictions. A small but non-negligible proportion of interventions were on nodes whose EIG and PTS values were both 0, suggesting that children occasionally chose interventions randomly.

Intervention strategies We used two hierarchical Bayesian models to capture children's intervention strategies (Figure 4). The single-strategy model draws from a single source to evaluate interventions—be it EIG, PTS, or always "1" in the case of random selection. The combined-strategy model assigns a weighted mean of EIG and PTS (the weight of EIG is θ) to each intervention. In both models, each child's decision noise τ_i is sampled from a population-level gamma distribution with two hyper-parameters α (shape) and β (rate). In the combined-strategy model, each child's weight of EIG θ_i is sampled from a population-level beta distribution with two hyper-parameters μ (mean) and κ (standard deviation. Uninformative priors are chosen for all hyper-parameters: α = .001, β = .001, $\mu \sim$ Beta(.5, .5), $\kappa \sim$ Gamma(.001, .001). The probability of selecting a given intervention is a function of its value $V(n)$ as well as the child's decision noise τ. Actual interventions are sampled from a categorical distribution of these probabilities. Parameter values were estimated using

Table 1: The deviance information criteria (DIC) of each model and the weight of EIG θ in three conditions.

	Baseline		Explanation		Report	
Model	DIC	θ	DIC	θ	DIC	θ
Random	514.15	–	777.82	–	764.64	–
EIG	481.98	–	727.00	–	769.45	–
PTS	469.62	–	**706.87**	–	**706.93**	–
EIG + PTS	**454.95**	.24	634.43	.31	746.25	.19

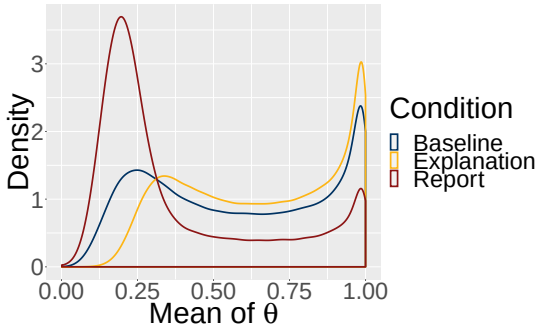


Figure 5: The distributions of the group-level hyper-parameter μ (the mean of θ) under the three conditions.

Markov chain Monte Carlo (MCMC) samples generated by the JAGS program[4] (Plummer, 2003). The deviance information criterion (DIC, Spiegelhalter et al., 2002) was used for model comparison. Models that fit data better (smaller posterior mean of the deviance $\bar{D}$) or are simpler (smaller effective number of parameters p_D) have lower DIC (= $\bar{D}$ + p_D). As a common practice, a difference over 10 is substantial.

As shown in Table 1, the combined-strategy model (EIG + PTS) best captured children's intervention strategy in both the Baseline and the Explanation conditions. However, the PTS-only model turned out to be the best fit in the Report condition. Children in all three conditions relied more on PTS than EIG, with the mean weight of EIG being .24, .31, and .19, respectively. Figure 5 illustrates the distributions of μ—the population-level hyper-parameter that captures the mean of θ—in all three conditions. To see whether μ differed across conditions, we sampled 10,000 estimates of μ in each condition. For each contrast between conditions (Explanation vs. Report, Explanation vs. Baseline, Report vs. Baseline), we paired the estimates randomly and calculated the differences. Since the 95% Highest Density Interval (HDI) of all three difference distributions contained 0, we couldn't claim with confidence that μ differed across three conditions.

[4]In keeping with Meng et al. (2018), we ran MCMC for 100,000 iterations, discarding the first 1,000 samples and drawing one sample every 10 iterations. To ensure that samples were from a stationary distribution, we repeated this process 30 times with different initial parameter values and results from each sequence of samples (or *chain*) successfully converged since Gelman and Rubin's diagnostic $\hat{R}$ (Gelman & Rubin, 1992) of all parameters was smaller than 1.05.

Intervention and inference Lastly, we looked at whether children's intervention choices and strategies predicted if they could accurately identify the true causal structures.

First, for each puzzle, we performed a logistic regression using the EIG value (0 or 1) of children's chosen intervention to predict whether they identified the correct structure later. In the Baseline condition, EIG values did not predict inference accuracy in any puzzles. However, in the Explanation condition, high-EIG interventions strongly predicted successes at identifying the correct structures in all six puzzles. In the Report condition, EIG values predicted inference accuracy in four of the six puzzles (except Puzzles 2 and 6).

We examined the correlation between the weight of EIG θ and children's average accuracy across all puzzles. θ and average accuracy were uncorrelated in the Baseline condition, $F(1,37) = 1.14$, $p = .29$, $\bar{R}^2 = .0038$, but positively correlated in the Explanation and the Report conditions, $F(1,57) = 30.73$, $p < .001$, $\bar{R}^2 = .34$ and $F(1,56) = 25.27$, $p < .001$, $\bar{R}^2 = .30$, respectively. Correlations in the Explanation and the Report conditions were both stronger than that in the Baseline condition, $z = 2.31$, $p = .02$ and $z = 2.08$, $p = .04$, respectively.

Discussion

In the current study, we investigated whether asking children to explain their intervention choices facilitated causal learning from intervention. Specifically, we looked at 1) whether explainers were better able to select informative interventions and 2) make accurate inferences based on interventional data.

Our first hypothesis was not supported by the results. Neither children's weight of EIG θ nor the group-level hyper-parameter μ that captures the mean of θ differed significantly across the Baseline, the Explanation, and the Report conditions; this suggests that children used similar strategies to select interventions across three conditions. Curiously, asking children to report their intervention choices might have slightly "backfired": While a linear combination of EIG and PTS best captured children's intervention strategy in the Baseline and the Explanation conditions, the PTS-only model turned out to best characterize the strategy used in the Report condition. Moreover, unlike in the other two conditions, the distribution of μ was right skewed in the Report condition, indicating heavier reliance on PTS. However, since differences in μ were not statistically significant, further investigation is needed to examine whether this finding was due to random noise or potential drawbacks of the report prompts.

Compared to the chance performance in the Baseline condition, children in both the Explanation and the Report conditions were more accurate at identifying the correct causal structures after performing interventions. Since children in the latter two conditions did not choose more informative interventions, a possible explanation is that when prompted to think about their intervention choices, children were better able to utilize interventional data that were already available. This explanation was supported by our findings: In the Explanation and the Report conditions, children's intervention choices (EIG value = 0 or 1) and interventions strategies

(weight of EIG θ) predicted their inference accuracy, which was not the case in the Baseline condition.

General Discussion

In the current study, we looked at whether asking children to think about their intervention choices might facilitate their causal learning from intervention. In Experiments 1A and 1B, 117 5- to 7-year-olds solved six puzzles where they performed one intervention to identify the true structure of three light bulbs that might be connected in one of two ways. Those in the Explanation condition were asked to explain why they chose certain interventions whereas those in the Report condition were simply asked to report their choices. Meng et al.'s (2018) previous study served as our Baseline condition where children solved the same puzzles without being prompted. Using hierarchical Bayesian models developed by Coenen et al. (2015), we captured children's intervention strategy mainly in terms of how much they relied on the normative strategy, which is maximizing the expected information gain (EIG) of their chosen interventions, and the suboptimal positive test strategy (PTS). Children in all conditions relied more on PTS than EIG; there was no difference across conditions. However, compared to those in the Baseline condition who performed at chance, children in both the Explanation and the Report conditions were more accurate at identifying the correct structures after interventions. Crucially, children's intervention choices and strategies only predicted their accuracy at inferring the true causal structures in the Explanation and the Report conditions but not in the Baseline condition.

Taken together, our findings suggest that while engaging in self-explaining did not help children select more informative interventions, asking them to *think* about their intervention choices (explaining or reporting) might help them better utilize interventional data that were already generated.

Revisiting the self-explaining effect

The major motivation behind this study was the plethora of self-explaining effects in education (Fonseca & Chi, 2011) and cognitive development (Lombrozo, 2016). Given what we found, two questions stood out: Why did self-explaining have no effect on intervention selection? Why was the improvement on causal inferences not unique in explainers?

Further investigation is needed to provide precise answers. Here we offer some speculations. Explaining an intervention is not an easy feat: Not only do you need to contrast the value of your intervention with that of other interventions, but more fundamentally, you need to contrast your strategy of evaluating interventions with other strategies. The cognitive process of generating a good explanation may be too challenging for 5- to 7-year-olds given their limited working memory capacity, knowledge about causal systems and experiments, and metacognitive skills (Horne, Muradoglu, & Cimpian, 2019). A recent study (Ruggeri, Xu, & Lombrozo, in press) suggested that the quality of explanations might matter after all. In their study, 4- to 7-year-olds were asked to explain phenomena in a domain before playing Twenty Questions in that domain; the accuracy of explanations was correlated with the efficiency of question-asking. Since reasonable explanations may be more difficult to generate in our study than in past studies (Walker et al., 2014, 2017), it might limit the benefit children can reap from self-explaining. Regarding the second question, it might be that when asked to reflect on (i.e., explaining or reporting) their intervention choices, children became aware that their interventions played an important role for solving puzzles later and therefore paid closer attention to the intervention outcomes when making causal inferences.

Future directions

Given the importance of intervention selection in causal learning, we seek to explore more effective scaffolding methods in the future. To begin, we can provide feedback after each intervention. A recent study (Liquin & Lombrozo, 2017) found that explaining had greater effects when evidence contradicted what learners' beliefs. Another way to strengthen the scaffolding may be asking children to explain why *each* possible intervention may or may not be useful, rather than just their chosen intervention. Since belief updating is inherently linked to intervention selection (Coenen & Gureckis, 2015), we may help children choose more informative interventions by correcting errors in their belief updating process.

Conclusion

Rather than passively absorbing correlations and crunching numbers, active learners generate explanations and design interventions to learn about causality. Our study is among the first to bridge "thinking" and "doing" in causal learning. While self-explaining did not show benefits of improving children's intervention strategy, prompting children to think about their intervention choices in some way (explaining or reporting) may help them better utilize interventional data generated by themselves to infer unknown causal structures.

References

Bramley, N. R., Lagnado, D. A., & Speekenbrink, M. (2014). Conservative forgetful scholars: How people learn causal structure through sequences of interventions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(3), 708–731.

Chi, M. T., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, *13*(2), 145–182.

Chi, M. T., De Leeuw, N., Chiu, M.-H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, *18*(3), 439–477.

Coenen, A., & Gureckis, T. M. (2015). Are biases when making causal interventions related to biases in belief updating? In D. C. Noelle et al. (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 411–416). Austin, TX: Cognitive Science Society.

Coenen, A., Nelson, J. D., & Gureckis, T. M. (2018). Asking the right questions about the psychology of human inquiry: Nine open challenges. *Psychonomic Bulletin & Review*, 1–41.

Coenen, A., Rehder, B., & Gureckis, T. M. (2015). Strategies to intervene on causal systems are adaptively selected. *Cognitive Psychology*, *79*, 102–133.

Edwards, B. J., Williams, J. J., Gentner, D., & Lombrozo, T. (2019). Explanation recruits comparison in a category-learning task. *Cognition*, *185*, 21–38.

Fonseca, B. A., & Chi, M. T. (2011). Instruction based on self-explanation. *Handbook of research on learning and instruction*, 296–321.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472.

Horne, Z., Muradoglu, M., & Cimpian, A. (2019). Explanation as a cognitive process. *Trends in Cognitive Sciences*, *3*(23), 187–199.

Klayman, J., & Ha, Y.-W. (1989). Hypothesis testing in rule discovery: Strategy, structure, and content. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(4), 596–604.

Liquin, E. G., & Lombrozo, T. (2017). Explain, explore, exploit: Effects of explanation on information search. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. J. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 2598–2603). Austin, TX: Cognitive Science Society.

Lombrozo, T. (2016). Explanatory preferences shape learning and inference. *Trends in Cognitive Sciences*, *20*(10), 748–759.

Luce, R. D. (1959). *Individual choice behavior*. New York, NY: John Wiley & Sons, Inc.

McCormack, T., Bramley, N. R., Frosch, C., Patrick, F., & Lagnado, D. A. (2016). Children's use of interventions to learn causal structure. *Journal of Experimental Child Psychology*, *141*, 1–22.

Meng, Y., Bramley, N. R., & Xu, F. (2018). Children's causal interventions combine discrimination and confirmation. In C. Kalish, M. Rau, J. Zhu, & T. Rogers (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 762–767). Austin, TX: Cognitive Science Society.

Nelson, J. D. (2005). Finding useful questions: On bayesian diagnosticity, probability, impact, and information gain. *Psychological Review*, *112*(4), 979–999.

Pearl, J. (2000). *Causality*. New York: Oxford University Press.

Plummer, M. (2003). A program for analysis of bayesian graphical models using gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing.*

Ruggeri, A., Xu, F., & Lombrozo, T. (in press). Effects of explanation on children's question asking. *Cognition*.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(4), 583–639.

Walker, C. M., & Lombrozo, T. (2017). Explaining the moral of the story. *Cognition*, *167*, 266–281.

Walker, C. M., Lombrozo, T., Legare, C. H., & Gopnik, A. (2014). Explaining prompts children to privilege inductively rich properties. *Cognition*, *133*(2), 343–357.

Walker, C. M., Lombrozo, T., Williams, J. J., Rafferty, A. N., & Gopnik, A. (2017). Explaining constrains causal learning in childhood. *Child Development*, *88*(1), 229–246.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, *12*(3), 129–140.

Williams, J. J., & Lombrozo, T. (2013). Explanation and prior knowledge interact to guide learning. *Cognitive Psychology*, *66*(1), 55–84.

Decisions Against Preferences

Messerli Michael (m.messerli@sheffield.ac.uk)
Department of Philosophy, Department of Philosophy, 45 Victoria Street, Sheffield, S37QB UK

Reuter Kevin (kevin.reuter@philo.unibe.ch)
Department of Philosophy, Laenggassstrasse 49a
3012 Bern, Switzerland

Abstract

An agent decides against her preferences, if she considers an option x better than another option y but nevertheless decides to do y. A central tenet of rational choice theory states that individuals do not decide against their preferences, whereby we find two kinds of potential counterexamples in the literature: *akrasia*, also known as weak-willed decisions, and decisions based on so-called *deontic constraints* such as obligations or commitments. While there is some empirical evidence that weak-willed choices are a real phenomenon, leading scholars in philosophy of economics debate whether choices based on commitments can be counter-preferential. As far as we know, however, nobody so far has tried to settle this debate empirically. This paper contributes to both debates since we present some empirical evidence that (i) akrasia can also be strong-willed and (ii) choices made on the basis of commitments can indeed be counter-preferential. We will conclude that people can decide against their preferences without being unreasonable.

Keywords: Counter-Preferential Choice; Rational Choice Theory; Akrasia; Commitments; Empirical Studies.

The Putative Irrationality of Deciding against Preferences

A fundamental assumption of most theories of rational choice is that an agent always chooses the option she considers best.[1] Violating this assumption is deemed irrational (Hausman 2012a). This basic tenet applies to both *maximizing* and *optimizing* concepts of rational choice, e.g., if one is a maximizer, one chooses the option that one believes maximizes one's utility. The option that is considered best (or at least not worse than any other) is the one the agent prefers.

What does it mean to *prefer* an option? Savage (1954) takes the notion of preference "in an ordinary mathematical usage by saying that the relation is a simple ordering among acts" (Savage 1954, p. 18). In contrast, philosophers often understand preferences as mental states, e.g., Hausman states "to say that Jill prefers x to y is to say that when Jill has thought about everything she takes to bear on how much she values x and y, Jill ranks x above y" (Hausman 2012b, p. 34). Of course, Hausman's notion of preference as *total subjective comparative evaluation* is controversial. Angner (2018), for instance, argues that economists neither *do use* nor *should use* such a conception of preference. However, note that Angner himself accepts that Hausman develops a useful model of preferences. For the purposes of our paper, Hausman's conception is specifically effective because it illustrates why choosing a worse option doesn't seem to make sense. According to such an understanding, an agent follows her preference ordering because she will then do what she values the most, or in other words, what she believes is best for her.

Saying that the agent chooses what she prefers to do is not to say that the agent chooses what is always best for her. In cases of uncertainty, an agent might choose an option that does not maximize her utility since unexpected states of the world might materialize. Agents might also base their preference ranking on false beliefs.[2] They might be mistaken about their preferences, or are simply not able to form a preference ordering (Messerli & Reuter 2017). It seems also false to postulate that agents always take into account all the available information like utilities and probabilities of options (Kahneman & Tversky 2000). Some of these aspects have been used to criticize models of rational choice theory. Nonetheless, these points of criticism do not apply to the fundamental tenet that agents choose what they consider best.

So, are advocates of rational choice theory right that counter-preferential choices do not exist? Or do agents sometimes choose options they consider worse than another, and hence violate this basic assumption of rational choice theory?[3]

[1] Some advocates of satisficing concepts of rational choice would disagree. If one is a satisficer, one settles for any alternative one considers satisfactory (Simon 1953; Slote 2004).

[2] Paul (2014), for instance, argues that agents who contemplate so-called *transformative choices* cannot form reasonable beliefs about the content of their experiences, and, thus, cannot make a rational choice. For a critical reply, see, e.g., Reuter & Messerli (2018).

[3] One might object that the assumption that agents choose what they consider best cannot be falsified. Revealed preference theorists, for example, assume that decisions reflect preferences. Consequently, there is no conceptual gap between a person's preferences and the actions she decides to perform. However, on most philosophers' interpretation, expected utility represents the strength of an agent's preference for the outcome, where preferences are understood as psychological states. Given this inter-

As far as we know, there are at least two kinds of potential counterexamples challenging this assumption. First, decisions can be *deontically constrained*, a technical term used to refer to constraints that arise when morality requires us to act in ways that are contrary to self-interest (see, e.g., Heath 2008). Put differently, a deontic constraint can be understood as a form of duty or rule that makes people refrain from the pursuit of individual advantage. That may sound fairly abstract, but we all are familiar with situations in which we act because we have given a promise, not because we actually prefer to act that way. We will come back to this issue in the General Discussion when discussing whether commitments can be counter-preferential.

Second, weak-willed decisions (or akratic decisions) are also potential counterexamples violating the assumption that agents choose what they consider best. To illustrate akratic decisions, take the case of Lewd Larry: Larry believes that staying in his room and staying faithful to his girlfriend is better than having an affair with his flatmate Jackie, but then finds himself trying to seduce her. Davidson (1970) defines weak-willed actions as follows:[4] In doing y an agent acts weak-willed if and only if: (i) the agent does y intentionally; (ii) the agent believes there is an alternative action x open to him; (iii) the agent judges that, all-things considered, it would be better to do x than to do y. Lewd Larry seems to fulfill all the requirements for being weak-willed. Weak-willed decisions not only seem to violate a fundamental assumption of rational choice theory, the intuitive plausibility of Lewd Larry demonstrates that weak-willed actions are real. However, most scholars at least agree with advocates of rational choice theory that such a decision is irrational: he should not have decided to seduce Jackie, given his belief it is not his best option.

In the rest of this paper, we do three things: First, we describe a case illustrating a violation of the aforementioned fundamental assumption of rational choice theory, which is not (at least not intuitively) an akratic decision, and, we present a first experiment showing that these cases are real. Second, we discuss an objection against our study and results, and we counter this objection using a second study. Third, taking ideas from Amartya Sen, we argue that we have good reasons to believe that such decisions are reasonable choices. In other words, such choices can be understood as acting out of commitment, whereby the commitment is counter-preferential.

pretation of rational choice theory, the assumption that agents actually choose options they consider worse than another is empirically testable.

[4] For discussions on the concept of akrasia as well as criticisms on Davidson's definition, see, e.g., Mele (1991), Holton (1999), and May & Holton (2012).

Experimental Study 1

Examples in which an agent decides against her preferences almost always seem to have the following pattern. The agent values x more than y, and hence prefers x to y, but "lower" desires triggered by lust or sloth, move the agent to do y. It need not be the case, however, that an agent decides against her preferences only if she is weak-willed. An agent might value x more than y, and hence prefers x to y, but is moved by his "higher" commitments or obligations to do y. To illustrate such a case, take the following example. Today, a colleague of yours has asked you whether you would help him move some furniture, and you agreed to be at his place the next morning. The next morning, however, friends of yours ask you whether you would like to join them for a beautiful day at the lake. It seems at least possible that in such a situation, you value going to the lake more than helping your colleague move furniture. Nonetheless, you decide to be at your colleague's place and help him move furniture. Note that similar to typical weak-willed decisions, you might loathe the fact that you have acted contrary to what you considered the best option. While such actions satisfy Davidson's definition of being weak-willed, it seems highly odd to call them weak-willed.[5]

The decision we described above violates the basic tenet of rational choice theory just as much as weak-willed decisions. The agent does not maximize her utility by choosing an option she considers worse than an available alternative. But are these decisions actually real? Or are they mere figments of philosophers' imaginations? The following experiment strongly suggests that these decisions are part of many people's reality.

Methods

120 participants were recruited on Amazon Mechanical Turk and paid a small fee for their participation. 5 participants were excluded for not having completed the survey. The remaining 115 participants (51 women, M_{age} = 39.09, SD = 15.69) all indicated that they were native English speakers. All participants were randomly assigned to one of three conditions, two test conditions (*Acquaintance*, *Colleague*) and one control condition. The

[5] Davidson discusses similar, so-called incontinent cases, in which a person follows a duty or principle when doing y intentionally (e.g., getting up and brushing her teeth), although all-things considered, she judges x to be better than y, e.g., it is more pleasurable to stay in bed. However, there is an important difference between our cases and Davidson's incontinent cases. In Davidson's examples, the agent does not believe she has very good reasons to follow a certain duty or principle, e.g., the agent reasons that her teeth are very strong anyway. In our examples, the agent is likely to believe that she has good reasons to keep a promise. In other words, the agent believes that helping a colleague move is the right thing to do. This contrast explains why Davidson believes that in incontinent cases an agent cannot understand herself and that she recognizes something absurd in her intentional behavior, while this would not be true in our case.

vignettes of the two test conditions read as follows:

Test condition Imagine that an acquaintance (a colleague) of yours asks you whether you would help him move some furniture and household appliances into his new apartment. You agree to be at his place at 10am the next day. The next morning, it is a beautiful warm summer day. At 8am you get a call from friends who ask you whether you would like to join them for a nice day at a lake. All things considered and independent of how you decide in the end, how do you value each of the two options:

- Spending the day at the lake and tell my acquaintance (colleague) that I cannot come.
- Moving furniture and household appliances and tell my friends that I cannot join them.[6]

The vignette for the control condition read:

Control condition Imagine that you plan your yearly holidays. On the one hand, you could go to the seaside and spend a week relaxing at the beach. On the other hand, you could book a trip to a city you have not seen before and experience some cultural highlights. All things considered and independent of how you decide in the end, how do you value each of the two options:

- Spending the holidays on the beach and not going to a city.
- Spending the holidays in a city and not going to the beach.

After the participants rated both options, they were then directed to the second question reading:

Decision Question You have just valued each of the two options. But how do you decide in the end? Please tell us what you will do:

For the two test conditions, the participants were presented with two options: (1) I choose to go to the lake and tell my acquaintance (colleague) I cannot come. (2) I choose to move furniture and household appliances and tell my friends I cannot join them. In the control condition, the options were: (1) I choose to go to the seaside and spend a week relaxing at the beach. (2) I choose to go to a city and experience some cultural highlights. Participants had to choose which decision they would take.

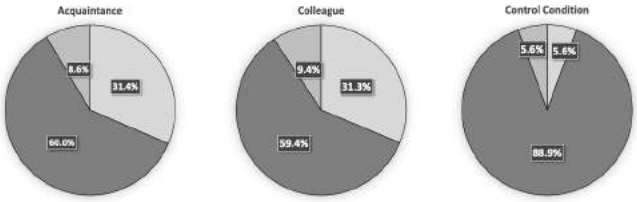


Figure 1 Responses in % to the two test conditions and the control condition. Dark grey depicts the percentage of participants who would decide in line with their preferences. Light grey represents the percentage of participants who would decide against their preferences. A few participants (medium grey) indicated an equal preference for both options.

Results

The results of people's responses are summarized in Figure 1 above. In both the *Acquaintance* as well as the *Colleague* condition, around 31% of the participants who decided in favor of helping to move furniture, considered going to the lake more valuable. The response profiles were significantly different between the test condition *Acquaintance* and control, $\chi^2 = 8.70, p = 0.013$, as well as *Colleague* and control, $\chi^2 = 8.64, p = 0.013$.

Discussion

The results indicate that a substantial portion of the participants would decide in favor of a less valuable option when they consider the scenario we presented them with. Simply put, the results suggest that there are situations in which many people will decide against their own preferences. The study was not designed to investigate which percentage of people are likely to make a decision against their preferences. Obviously, the scenarios were quite specific and for many the situation did not even present them with a "difficult" choice. Thus, it is likely that for many more people than just the recorded 31%, there exist choices in which option x is preferred but they still decide in favor of option y. Of course, for most decisions, people's choices will nicely align with their preferences. In fact, the control condition was specifically designed as a base rate for decisions in which preferences are the sole determiner of the decision in question. The significant differences between the test condition and the control demonstrates, however, that not all decisions are like that. Other factors may determine which option we are going to choose.

Before we discuss a possible explanation for the recorded data, let us first address the most obvious objection against our study. In order to counter this objection, we then briefly present the results of a second study.

[6] Both options were presented in randomized order and participants were asked to rate the value of each option on an 11-point Likert scale anchored at 0 meaning "Not at all valuable" and 10 meaning "Extremely valuable". Which concept of utility is relevant here? Importantly, we do not understand utility as a more precise ranking than an *ordinal one* (e.g. *cardinal measure* or *ratio scale*). The strength of a value judgement can be understood in purely ordinal terms, respectively, the experiment is perfectly consistent with an ordinal interpretation. If a participant evaluates two options within this scale, e.g. $a = 9$ and $b = 3$, this simply means that he or she ranks a above b. In other words, the only information these numbers provide is that an agent prefers a to b without saying how much he or she values a more than b.

Objection

The experiment reveals a potential decision against one's preferences, only if people gave all-things-considered value ratings when considering their options. It is indeed possible that some people merely considered the positive value of spending a day on the lake without considering the negative value of telling one's acquaintance or colleague that one is not available for moving after all. If that were the case, then it would not surprise to see decisions made against one's *rated* preferences.

We do not believe, however, that this is a likely possibility. When we asked the participants to rate the value of the options, we specifically named the positive as well as the negative aspect of the choice, e.g., one of the options read: "Spending the day at the lake and tell my colleague that I cannot come." Moreover, the number of participants who would decide against their own value judgements might even be greater, because some participants might have self-censored themselves so as to appear to be consistent when making a decision.

However, one might insist on the ambiguity of the term "value", respectively, that we and the participants do not refer to the same concept here. It is our understanding that the concept of value can be understood in terms of the agent's ends and desires. "I judge that a is more valuable than b" means that I believe that a is more valuable than b in terms of my ends and desires.[7] Now the objection that arises is that participants must have some different concept of value in mind, because there is not only the end of enjoying a great day at the lake but also the end of helping other people, which is obviously more important to them. If the objection stands, participants do in fact decide in line with their values and do not decide counter-preferentially. [8]

The objection we raised should be taken seriously. We have, therefore, conducted a second experiment where we first explained to participants which concept of value is involved. We will see that our results are robust, even if we change the experimental setting in this way.

[7] In accordance with rational choice theory, we do not make any proposal concerning the content of these ends. This means that we recognize no distinction between goals such as making a million dollar, helping other people and being a sadist. Also note that there are no implications regarding risk-taking. The value judgement that a is more valuable than b might be risk-neutral such as in standard approaches or risk-averse such as in *prospect theory*.

[8] One way of testing the objection would be to further specify the alternatives, e.g., instead of stating one of the options as "Spending the day at the lake and tell my colleague that I cannot come," we could state "Spending the day at the lake and break my promise to my colleague". The reason why we opted for a different way to tackle the objection is that "breaking a promise" or "breaking a commitment" (we will come back to the role of commitments in the General Discussion) is a very negative trigger. The wording "tell my colleague that I cannot come" is relatively neutral in this regard. However, we agree that the empirical evidence for decisions against preferences would be even greater if the negative aspects of a decision would be highlighted even further. In a follow-up study, we plan not only to investigate a larger variety of experimental stimuli but also the impact of the exact wording on the empirical effect.

Experimental Study 2

In order to address the objection stated above, we decided to rerun both test conditions (*Acquaintance*, *Colleague*) to see whether the results would change or remain robust. If the objection is correct, then we should see a substantial drop in the percentages of people who indicate decisions that go against their preferences.

Methods

100 participants were recruited on Amazon Mechanical Turk and paid a small fee for their participation. 2 participants were excluded for not having completed the survey. The remaining 98 participants (48 women, M_{age} = 36.92, SD = 12.38) all indicated that they were native English speakers. All participants were randomly assigned to one of two conditions (*Acquaintance*, *Colleague*). The vignettes of the two conditions were exactly the same as the vignettes of the test conditions in Experiment 1 with one exception: after participants had given their consent to this study, they were informed about the task ahead in the following manner.

Instructions On the next screen we will ask you to value certain events. Before you do so, please consider the following example: Imagine you have to value a one week trip to Europe. On the positive side there might be aspects like relaxing, eating new and exciting food, being able to tell your friends of an amazing trip when you are back, etc. On the negative side there might be aspects like being jetlagged, longing for your loved ones at home, missing an important meeting at work, etc. Thus, if you value an option or an event, you take into account all its positive and negative aspects and then make an overall judgement.

After these instructions, participants rated both options (see Experiment 1 above), and then answered the decision question (see also Experiment 1 above).

Results

In the *Acquaintance* condition, 36.6% of the participants who decided in favor of helping to move furniture considered going to the lake more valuable. In the *Colleague* condition, 26.3% of the participants who decided in favor of helping to move furniture considered going to the lake more valuable. The results of people's responses are summarized in Figure 2 below.

Discussion

The data we received in Experiment 2 are highly similar to those we collected in Experiment 1. While the percentage of people who decided against their preference in the

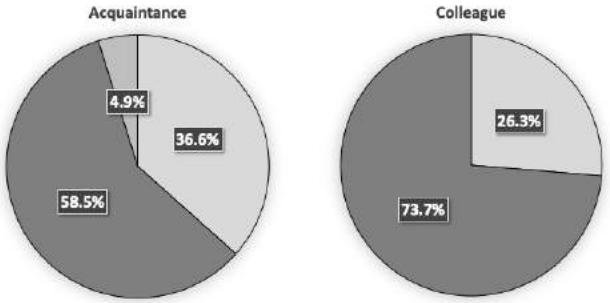


Figure 2 Responses in % to the two conditions. Dark grey depicts the percentage of participants who would decide in line with their preferences. Light grey represents the percentage of participants who would decide against their preferences. A few participants (medium grey) indicated an equal preference for both options.

Acquaintance condition rose from 31.4% to 36.6%, the percentages in the *Colleague* condition decreased from 31.3% to 26.3%. Thus, overall the results of Experiment 1 were very robust. It seems therefore very likely, that the participants in Experiment 1 entertained a notion of *value* not only similar to the one in Experiment 2, but also of the right kind.

General Discussion

The studies suggest that people often make decisions in favor of options they consider less valuable.[9] If our results are correct, there are crucial implications for both the discussion on *akrasia* and the debate on *rational choice*, respectively, the connection between rational choice and *deontic constraints*.

First, let us briefly mention the implication for akrasia. Importantly, some philosophers mention that akrasia can also be strong-willed (e.g., see Holton 1999; Yao 2017). However, as far as we know, (i) nobody thus far has interpreted such cases as actions out of commitment, (ii) there is currently no empirical evidence for such cases, and, (iii) other cases of strong-willed akrasia are related to violations of resolutions and preference change. In contrast, our example illustrates a case in which no such additional machinery is necessary.

Second, our results are crucial for rational choice and the debate on commitments. Most rational choice theorists are likely to consider counter-preferential decisions as unreasonable or irrational, similar to typical weak-willed decisions. However, while people in weak-willed decisions usually act out of lower desires, we have described a case in which people seem to be rather strong-willed when acting against their most valued option. Does this difference allow us to frame such decisions as reasonable? This largely depends on what ultimately motivates people to decide against preferences in *strong-willed* decisions. A possible explanation of such decisions takes into account the importance of commitments. After all, many people are likely to decide to help their acquaintance or colleague move furniture because they have committed themselves to do so, not because they like moving furniture. However, shouldn't these commitments be reflected in peoples' evaluations of the two alternatives? According to Sen (1977) this need not be the case.

Sen distinguishes three kinds of motivations: narrow self-interest, sympathy, and commitment. Both, narrow self-interest as well as sympathy, directly affect a person's own welfare and should be reflected in people's value judgements. In contrast, Sen (1977, p. 326) characterizes *commitments* as altruistic attitudes towards others. Accordingly, a person who acts out of commitment chooses an option that she considers the right thing to do, even if that option is less preferable than an alternative. Sen admits that within the framework of rational choice theory, there is no place for a notion like commitment because it does not lead to any difference in terms of one's expected advantage.[10] Speaking purely in terms of rational choice theory, decisions against preferences, are therefore irrational. That said, Sen's theoretical work on commitments has caused a lot of attention, because it seems that people who act out of commitment, are *irrational* only in the skewed notion of rational choice theory. At least, intuitively, it seems that people who act against their preferences but in favor of an option they consider the right thing to do, are reasonable agents.

Given the importance of Sen's contribution, some philosophers have started to question Sen's depiction of commitments as factors that may have a motivating force beyond expected advantage. Contra Sen, Hausman (2007) agues that we need to distinguish among the variety of factors responsible for agents' preferences, rather than distinguish between preferences and commitments. According to his view, commitments are not counter-preferential but rather influence all-things considered judgements. Thus, while Sen believes commitments can directly determine our choices, Hausman argues that they do so only via preferences. As far as we know, the role of commitments in the decision making process has not yet been empirically investigated.[11] And

[9] One might object that there is a gap between the participants *rated* preferences and their real decisions or real behavior. Put differently, participants are not actually making a decision but provide inconsequential responses after reading abstract descriptions of some options. However, while some studies have shown an inconsistency between people's rated preferences and real behavior, a variety of empirical studies have also shown high consistency between people's ratings and their behavior. Importantly, we are not aware of any empirical or theoretical arguments why people systematically deviate in our respective context.

[10] It is important to keep in mind here that Sen distinguishes different notions of preferences. The two most important ones are (i) preference as (revealed) choice ranking and (ii) preference as expected advantage ranking.

[11] Note that we do not claim that there's no empirical research on commitments. See, e.g., Székely & Michael

therefore, we do not yet know whether Sen or Hausman are right. However, our experiment may provide a first step to settle this debate. According to our results, it seems that commitments may sometimes influence our choices directly and not via preferences. At a minimum, opponents of Sen would need to explain why some people decide to move furniture, even if the other option is considered better, all-things-considered.

Before we conclude, let us briefly mention one other account that could be drawn upon to explain our data. Heath (2008) argues that theories of rational choice can be modified in order to incorporate rule following behavior. His model distinguishes between one's desire for an outcome (its expected utility) and how appropriate the outcome is (the normative appropriateness of that outcome). The basic idea is that an agent's utility function combines two things: Getting the best outcome and doing the right thing. According to Heath's approach, participants would distinguish two stages. First, they would rank permissible actions as more or less appropriate. Second, they would add these values to the expected utilities. It would take more experiments to find out whether participants indeed proceed in the way suggested by Heath. In any case, Sen's account provides a straightforward explanation of our data.

Conclusion

In closing, let us summarize what we have done. Adherents of rational choice theory assume that agents choose the option they consider best. In this paper, we have discussed a case that violates this basic assumption. Crucially, we have not merely relied on our own intuitions of whether such a case is real, but conducted two studies, the results of which strongly suggest that many people make decisions against their preferences. Some might argue that this case is just one out of many showing rational choice theory to be mistaken. In particular, weak-willed decisions have been largely accepted to be real-world cases in which agents act contrary to their best judgements. However, weak-willed decisions can be distinguished from our case study in two important respects: First, while in weak-willed decisions, people act out of their lower desires, our case shows that people can decide against their preferences by being strong-willed. Second, at least according to Sen's account, there are good reasons to believe, agents may act against their preferences but at the same time make a reasonable choice. Our results provide evidence that Sen is right, respectively, that commitments can be counter-preferential.

Acknowledgments

We would like to thank Catherine Herfeld for her very helpful comments on this paper. Michael Messerli was supported by the SNF Postdoc Mobility Grant P2SKP1_171776. Kevin Reuter was supported by the SNF project Affective Mind, Grant No. 169484.

(2018). As far as we know, there has been no empirical research on the question of whether commitments can be counter-preferential.

References

Angner, Erik (2018). What Preferences Really Are, in: *Philosophy of Science* 85 (4), 660-681.

Davidson, D. (1970). How Is Weakness of the Will Possible?, in *Essays on Actions and Events*, Oxford: Clarendon Press.

Hausman, D. (2007). Sympathy, Commitment, and Preference, in: *Rationality and Commitment*, F. Peter and B. Schmid (ed.), Oxford: Oxford University Press.

Hausman, D. (2012a). *Philosophy of Economics*, The Stanford Encyclopedia of Philosophy (Winter 2013 Edition), Edward N. Zalta (ed.).

Hausman, D. (2012b). *Preference, Value, Choice, and Welfare*, Cambridge: Cambridge University Press.

Heath, J. (2008). *Following the Rules. Practical Reasoning and Deontic Constraint*, Oxford: Oxford University Press.

Holton, R. (1999). Intention and Weakness of Will, in: *Journal of Philosophy* 96 (5), 241-262.

Kahneman, D. & Tversky, A. eds. (2000). *Choices, Values and Frames*, New York: Cambridge University Press and the Russell Sage Foundation.

May, J. & Holton, R. (2012). What in the World is Weakness of Will?, in *Philosophical Studies* 157 (3), 341-360.

McClennen, E. (1990). *Rationality and Dynamic Choice*, Cambridge: Cambridge University Press.

Mele, A. (1991). Akratic Action and the Practical Role of Better Judgement, in: *Pacific Philosophical Quarterly*, 72 (1), 33-47.

Messerli, M. & Reuter, K. (2017). Hard Cases of Comparison, in: *Philosophical Studies* 174 (9), 2227-2250.

Paul, L.A. (2014). *Transformative Experience*. Oxford: Oxford University Press.

Reuter, K. & Messerli, M. (2018). Transformative Decisions, in: *Journal of Philosophy*, 115 (6), 313-335.

Simon, H. (1953). *A Behavioural Model of Rational Choice*. Santa Monica: Rand.

Sen, A. (1977). Rational Fools: A Critique of the Behavioral Foundations of Economic Theory, in: *Philosophy & Public Affairs* 6 (4), 317-344.

Slote, M. (2004). Two Views of Satisficing in: *Satisficing and Maximizing*, Byron, M. (ed.), Cambridge: Cambridge University Press.

Székely, M. & Michael, J. (2018). Investing in commitment: Persistence in a joint action is enhanced by the perception of a partner's effort, in: *Cognition*, 174, 37-42.

Yao, V. (2017). Strong-willed Akrasia. *Oxford Studies in Agency and Responsibility*, 4, 6-27.

The Synergy of Passive and Active Learning Modes in Adaptive Perceptual Learning

Everett Mettler (mettler@ucla.edu)[1]
Austin S. Phillips (asphillips@ucla.edu)[1]
Christine M. Massey (cmassey@psych.ucla.edu)[1]
Timothy Burke (mizerai@ucla.edu)[1]
Patrick Garrigan (pgarriga@sju.edu)[2]
Philip J. Kellman (kellman@cognet.ucla.edu)[1]

[1]Department of Psychology, University of California, Los Angeles
Los Angeles, CA 90095 USA
[2]Department of Psychology, St. Joseph's University
Philadelphia, PA 19131 USA

Abstract

Adaptive learning systems that generate spacing intervals based on learner performance enhance learning efficiency and retention (Mettler, Massey & Kellman, 2016). Recent research in factual learning suggests that initial blocks of passive trials, where learners observe correct answers without overtly responding, produce greater learning than passive or active trials alone (Mettler, Massey, Burke, Garrigan & Kellman, 2018). Here we tested whether this passive + active advantage generalizes beyond factual learning to perceptual learning. Participants studied and classified images of butterfly genera using either: 1) *Passive Only* presentations, 2) *Passive Initial Blocks* followed by active, adaptive scheduling, 3) *Passive Initial Category Exemplar* followed by active, adaptive scheduling, or 4) *Active Only* learning. We found an advantage for combinations of active and passive presentations over *Passive Only* or *Active Only* presentations. Passive trials presented in initial blocks showed the best performance, paralleling earlier findings in factual learning. Combining active and passive learning produces greater learning gains than either alone, and these effects occur for diverse forms of learning, including perceptual learning.

Keywords: adaptive learning; perceptual learning; spacing effect; memory; active learning; passive learning

Introduction

The well-known *spacing effect* is a boost in long-term retention that results when recurrent learning episodes are spaced across gaps in time (Carpenter, 2017; Cepeda, Pashler, Vul, Wixted & Rohrer, 2006; Delaney, Verkoeijen & Spirgel, 2010). Spacing effects apply to a wide variety of learning domains and learners, and also influence diverse learning modes such as perceptual learning (Mettler & Kellman, 2014).

Recent research has shown that spacing effects can be enhanced by dynamically adjusting the size of spacing intervals during a learning session using an adaptive algorithm, Adaptive Response-Time-based Scheduling (ARTS; Mettler, Massey & Kellman, 2011; Mettler, Massey & Kellman, 2016). In ARTS, spacing delays are updated to match changes in learning strength as learning progresses for individual learners and items. Learning strength can be reliably estimated from response time (RT), with slower response times indicating retrieval difficulty and correspondingly lower learning strengths (Pyc & Rawson, 2009; Benjamin & Bjork, 1996; Karpicke & Bauernschmidt, 2011). ARTS updates the spacing among items in real time, by tracking the underlying learning strengths using an individual's accuracy and RT for learning items or for categories, producing highly efficient learning (Mettler, Massey & Kellman, 2011, 2016). In perceptual learning and other category learning domains, the same adaptive learning approach is applied to categories, such that learning strength for each category influences the priority of a learning trial involving a new exemplar of that category. Such adaptive spacing, and the interleaving of exemplars of different categories, also produces strong learning benefits relative to other arrangements (Mettler & Kellman, 2014).

Achieving the benefits of adaptive spacing requires interactive learning trials from which performance data are obtained. Recent work, however, suggests that the benefits of adaptive spacing may be further enhanced by combining active trials with passive presentations during learning. In a study investigating the learning of geography facts, Mettler, Massey, Burke, Garrigan & Kellman (2018) compared delayed retention rates following passive learning, active learning, and combinations of passive and active learning. Combinations of passive and active learning resulted in better performance than active learning alone. Passive presentations alone fared worst. In addition, the specific manner of combining passive and active modes mattered: learning which began with multiple blocks of passive trials followed by active, adaptive learning resulted in the best performance.

In the current study, we investigated whether the same learning advantages for passive combined with active learning might exist for perceptual learning (PL), which presumably rests on different mechanisms (changes in information selection and encoding vs. explicit storage of memory items). For factual information, spacing was manipulated among individual factual items. Here spacing was manipulated among categories of perceptual stimuli, but with each re-presentation of a category, a new exemplar was shown. Some earlier work suggested that combining

passive and active modes might benefit PL (Thai, Krasne & Kellman, 2015); however, no work has explored different modes of combining active and passive trials.

Why might including some passive learning trials among active learning trials result in better PL than active trials alone? One benefit of passive trials may be to prevent the negative cognitive and motivational consequences of asking learners to generate answers in initial interactive learning trials - similar to the hypothesized benefits of initial passive trials in factual learning. Specific to PL, passive trials might focus attention on some characteristics of categories, and active trials might complement this learning by targeting other characteristics. For example, Carvalho & Goldstone (2015) suggested that passive trials can increase attention to commonalities between members of the same category when certain between-category and within-category similarity relations hold, but that active trials provide greater benefits to learning when the inverse similarity relations hold. Combining passive and active trials could be a strategy then to increase overall learning due to the complementary strengths of active and passive presentations in the learning of categories that possess a variety of internal structures. In the current study, we systematically compared learning schedules that included passive and active trials alone, and two different combinations of passive and active trials. We analyzed subsequent retention of perceptual classification after a delay, and we examined whether passive and active training was affected by internal category structures such as between and within-category similarity.

We compared four conditions: a) *Passive Only* presentations of learning items, b) *Passive Initial Blocks* followed by active, adaptive scheduling, c) *Passive Initial Category Exemplar* followed by active, adaptive scheduling for each category introduced, and d) *Active Only* learning with no passive presentations. We hypothesized that introductory presentations of passive trials, followed by active learning would fare the best, however, the effect of passive learning might be better if passive trials were limited to single presentations rather than blocks.

Method

Participants One hundred twenty undergraduate psychology students participated to partially fulfill course requirements.

Materials 12 categories (genera) of butterflies (lepidoptera) were used, where each genus contained images of 9 exemplars. On each learning trial, an image of one category exemplar was presented on the left side of the screen. In Active trials, the 12 possible category name responses were shown in a two-column list organized alphabetically on the right side of the screen. In Passive trials, only the correct category label was shown and the alternate category names were omitted.

Figure 1: Images of 2 butterfly genera with 3 exemplars from each genus. Danaus (top) and Neptis (bottom).

Design A 4x3x2x2 mixed factorial design was used. There were four between-subject passive/active conditions (*Passive Only, Passive Initial Block, Passive Initial Category Exemplar,* and *Active Only*). A pretest/posttest design consisted of three test phases (Pretest, Immediate posttest, and 1 week delayed posttest). In addition there was a within-subject factor of Familiarity (Familiar vs Unfamiliar); that is, at each test, each category was tested twice with both new and previously seen exemplars. Finally, there was a between-subject factor of Assessment List, such that the familiar and unfamiliar exemplars for each category were randomly selected differently for each of the two lists.

Procedure Participants completed two sessions separated by one week. The initial session consisted of a pretest, training phase and immediate posttest. The second session consisted of a delayed posttest only. In all tests and training, participants were shown a genus exemplar and were asked to identify the matching genus name from a list of all 12 category names. No feedback was provided. Tests consisted of two presentations of each genus: one presentation was a 'familiar' exemplar shown during training, and the other exemplar was an 'unfamiliar' exemplar withheld from training. There were two assessment lists and each participant was randomly assigned one of the versions. Each participant saw the same test version, and thus the same familiar and unfamiliar exemplars for each category, across pre, post and delayed tests.

In the *Passive Only* condition, butterflies were presented in 12 blocks of 12 passive trials. Each category appeared once per block, in random order, and a random exemplar from the category was chosen for each presentation. In the *Passive Initial Blocks* condition, participants first completed 2 blocks of passive trials, with blocks having the same structure as the *Passive Only* condition, followed by adaptive scheduling. In the *Passive Initial Category Exemplar* condition, the first presentation of each category was a passive trial followed by a fixed spacing interval of two intervening trials, so that the correct response was not still in working memory. All trials in this condition that did not involve the first presentation of a category were

adaptively scheduled. In the *Active Only* condition, all trials were adaptively scheduled.

The ARTS algorithm determined the adaptive scheduling for active trials. After every response, ARTS calculates a priority score for each learning item and compares scores across items to determine which item will be presented next. Equation 1 shows the priority score calculation.

$$P_i = a(N_i - D)[b(1 - \alpha_i)\, Log(RT_i / r) + \alpha_i W] \qquad (1)$$

Detailed description of the ARTS algorithm can be found in Mettler, Massey & Kellman (2011, 2016). ARTS parameters were the following: the enforced delay *D* was set to 2 trials, the incorrect penalty *W* was set to 20, parameters *a*, *b*, *r* were set to 0.1, 1.1, and 1.7 respectively, and the timeout was 30 seconds.

Learning for each category continued until 5 out of the last 6 presentations were correctly answered with all correct response times less than 7 seconds. Learning criteria, adopted from previous studies, included both speed and accuracy, where speedy responses also ensured that final presentations were widely spaced.

Participants were assigned to Condition using a pretest balancing algorithm (similar to a procedure called Minimization; Pocock & Simon, 1975; Mettler et al., 2018). The condition balancing algorithm was constrained so that, across conditions, the largest difference in number of assigned participants never exceeded one. There were exactly 30 participants in each of the 4 conditions.

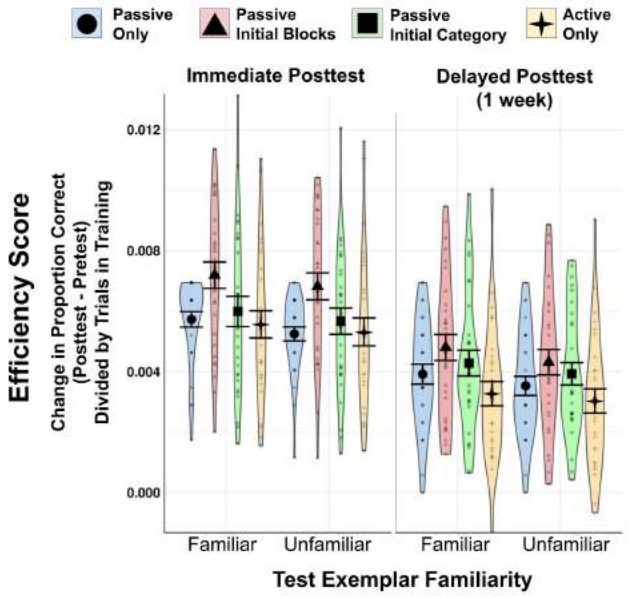


Figure 2: Learning Efficiency in Immediate and Delayed Posttest by Test Item Familiarity. (Violin plot shows mean, +/- 1 standard error of the mean, density estimate and individual data points).

Dependent Measures and Data Analysis

Because all adaptive conditions used learning to criterion, our primary measure was learning *efficiency*, defined as accuracy gain from pretest to posttest divided by the number of trials invested in learning. Efficiency gives a way of measuring learning that incorporates both variations in posttest performance, and variations in the number of learning trials required to reach the learning criteria. It may be thought of as a rate measure, indicating performance improvement per trial. The number of passive trials was determined based on pilot work to be roughly equal to the number of trials needed to reach mastery in active conditions. In the two conditions combining passive and active trials, all trials were included in trial and efficiency calculations.

In addition to efficiency we measured change in accuracy and reaction time. All measures were assessed using standard parametric statistics, such as ANOVA. Because we sought to compare differences across learning conditions, we conducted planned comparisons between pairs of conditions. All statistical tests were two-tailed, with a 95% confidence level, all effect sizes d are Cohen's d, and all error bars in graphs show +/- 1 standard error of the mean.

Results

Pretests A 4x2x2 ANOVA on Condition, Assessment List and Familiarity showed no significant main effect of Condition ($F(3,112)=0.213$, $p=.887$, $\eta_p^2=.006$), Assessment List ($F(1,112)=0.457$, $p=.500$, $\eta_p^2=.004$) or Familiarity ($F(1,112)=2.395$, $p=.125$, $\eta_p^2=.021$).

Efficiency *Efficiency*, defined as posttest accuracy gain from pretest divided by learning trials to criterion, is shown in Figure 2 for each of the posttests, the 4 learning conditions and for familiar vs. unfamiliar test items. The *Passive Initial Blocks* condition appeared to have higher efficiency at immediate posttest and highest numerical efficiency at delayed posttest. A 4x2x2x2 mixed factorial ANOVA on Passive/Active Scheduling Condition, Test Phase (Immediate vs. Delayed Posttest), Item Familiarity (Test exemplar seen vs. withheld in training) and Assessment List (1 vs 2) showed a significant main effect of Condition ($F(3,112)=2.921$, $p=.037$, $\eta_p^2=.073$) a significant main effect of Test Phase ($F(1,112)=277.127$, $p<.001$, $\eta_p^2=.712$), a significant main effect of Familiarity ($F(1,112)=17.832$, $p<.001$, $\eta_p^2=.137$), and no significant main effect of Assessment List ($F(1,112)=0.018$, $p=.893$, $\eta_p^2<.001$). Interactions were not significant (ps>.127) but there was a marginally significant interaction between Condition and Phase ($F(3,112)=2.197$, $p=.092$, $\eta_p^2=.056$) and Assessment List and Familiarity ($F(1,112)=3.391$, $p=.068$, $\eta_p^2=.029$).

The marginally significant interaction between Condition and Test appears to be driven by the clear superiority of

Passive Initial Blocks at immediate test that is less pronounced at delayed test. Paired comparisons revealed significant differences between conditions at immediate test (*Passive Only* vs. *Passive Initial Block*, t(58)=3.12, p=.003, d=0.84; *Passive Initial Blocks* vs. *Active Only* , t(58)=2.53, p=.014, d=0.65), and a marginally significant difference between *Passive Initial Blocks* vs. *Passive Initial Category* (t(58)=1.868, p=.067, d=0.48). Other comparisons did not reach significance (ps >.51). Paired comparisons at delayed posttest showed significant differences between *Passive Initial Blocks* and *Active Only* (t(58)=2.514, p=.015, d=0.65). There was a marginally significant difference between *Passive Initial Category* and *Active Only* (t(58)=1.74, p=.088, d=0.45). The remaining comparisons did not reach significance (ps > .105). Between immediate and delayed posttests, all pairwise comparisons were significant (p<.05) except for between *Active Only* at immediate test and *Passive Initial Blocks* at delayed posttest (t(58)=1.47, p=.147, d=0.38).

Trials in training Mean trials to reach learning criteria or the end of the session are shown in Figure 3. A 3x2 mixed factorial ANOVA was conducted on Condition and Assessment List. The *Passive Only* condition was removed from the ANOVA and paired comparisons due to its fixed (preset) number of trials. There was a significant effect of condition ($F(2,84)=3.448$, $p=.036$, $\eta_p^2=.076$). Paired comparisons showed significant differences between *Passive Initial Blocks* and *Passive Initial Category* (t(58)=2.068, p=.043, d=0.554) and between *Passive Initial Blocks* and *Active Only* (t(58)=2.707, p=.009, d=0.732), but not between *Passive Initial Category* and *Active Only* (t(58)=0.623, p=.536, d=0.161). One sample t-tests were used to compare each Active condition against the *Passive Only* condition mean of 144 trials. There was a significant difference for *Active Only* (t(29)=2.69, p=.012) and a marginally significant difference for *Passive Initial Category* (t(29)=1.97, p=.057), but no significant difference for *Passive Initial Blocks* (t(29)=0.70, p=.49).

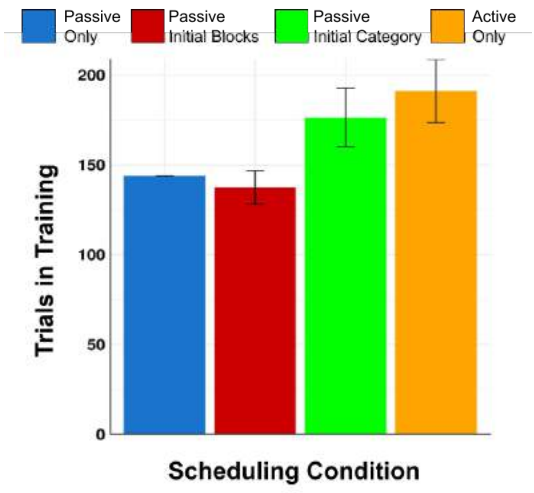


Figure 3: Trials in training session by 4 scheduling conditions.

Learning Analytics

In order to explore the reasons why performance was highest for *Passive Initial Blocks* conditions and lower for *Active Only*, we explored trial-by-trial data during learning. In prior work with learning of factual items we determined that initial blocks of passive items significantly reduced the severity of certain deleterious trial sequences. Specifically, the incidence of errors followed by correct responses (dubbed 0,1 sequences) across conditions, and these sequences followed by another error (0,1,0 sequences), were reduced in conditions that included initial passive blocks, relative to the other active conditions.

We examined 0,1 trial sequences during learning across the three adaptive scheduling conditions. First, the incidence of 0,1 sequences was highest in the *Active Only* condition and lowest in the *Passive Initial Blocks* condition, even when adjusting for the first few trials where there are necessarily errors in the *Active Only* condition due to initial guessing. The frequency of 0,1 instances across the three conditions and for groups of initial trials are shown in Figure 4. Trials 4+ are most instructive, showing that *Passive Initial Blocks* had the fewest occurrences of 0,1 among the three conditions. A 3 way ANOVA run on Condition for Trials 4+, found a significant effect of condition ($F(2,87)=5.23$, $p=.007$, $\eta_p^2=.107$) and paired comparisons showed significant differences between *Passive Initial Blocks* and *Passive Initial Category* (t(58)=2.52, p=.014, d=0.66), *Passive Initial Blocks* and *Active Only* (t(58)=3.15, p=.003, d=0.82), but not between *Passive Initial Category* and *Active Only* (t(58)=0.65, p=.519, d=0.17).

We also examined accuracy following 0,1 sequences. Again, the first 3 trials were removed to equate conditions with respect to number of prior presentations. Figure 5

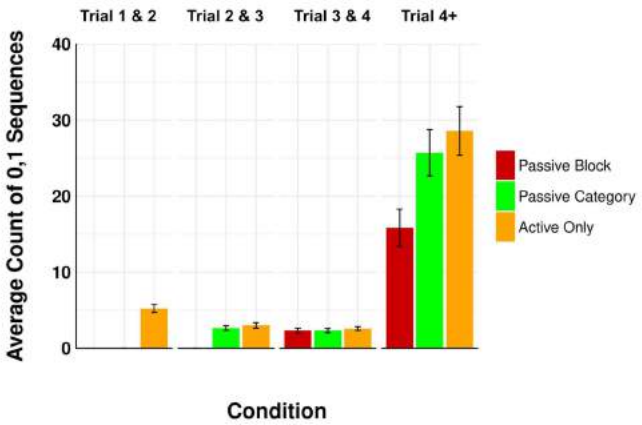


Figure 4: Frequency of 0,1 sequences by condition and by trial in learning session.

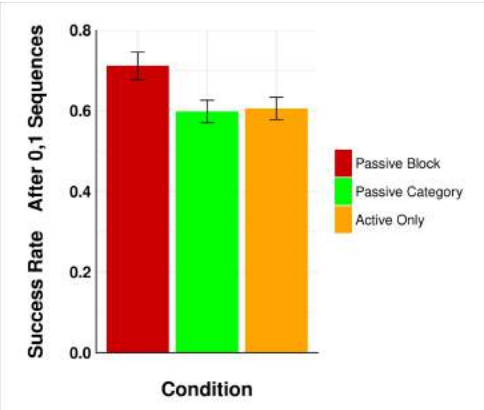


Figure 5: Success rate after 0,1 sequences, corrected for initial guessing (beginning at trial 3 for all conditions).

shows accuracy following 0,1 sequences. A 3 way ANOVA on success rate after 0,1 sequences found a significant effect of Condition ($F(2,87)=4.34$, $p=.016$, $\eta_p^2=.091$). Paired comparisons showed significant differences between *Passive Initial Blocks* and *Passive Initial Category* ($t(58)=2.71$, $p=.009$, $d=0.7$), *Passive Initial Blocks* and *Active Only* ($t(58)=2.22$, $p=.030$, $d=0.57$), but not between *Passive Initial Category* and *Active Only* ($t(58)=0.62$, $p=.539$, $d=0.16$).

Within-category and between-category similarity relations Since prior research indicates the importance of within and between category similarity for benefits from passive or active trial scheduling, we examined passive only and active only learning efficiency as a function of between and within-category similarity. Similarity relations were determined by subject ratings of each category, first for between-category relations and then again, separately for within-category relations. All 12 categories were rated on a 3 point similarity scale for between-category similarity with 3 being highest and 1 lowest. Subject ratings were averaged for each category and categories were divided into 1 of 3 between-category similarity groups based on the tertile of their averaged rating. The same procedure was repeated for within-category ratings. Thus, within and between-category similarities were estimated independently. Posttest efficiencies were compared for two scheduling conditions, *Passive Only* and *Active Only*, across the three levels of within and between-category similarity.

Average efficiency differences, plotted separately for each within and between-category similarity group are shown in Figure 6. Two 2x2x3 ANOVAs were conducted, each with training schedule (*Passive Only, Active Only*), and Test phases (Immediate vs. Delayed posttest) as factors. One ANOVA also included within-category similarity as a factor, and the other also included between-category similarity as a factor. The ANOVA with within-category similarity as a factor showed no significant effect of Condition ($F(1,176)=1.63$, $p=.204$, $\eta_p^2=.009$), a significant

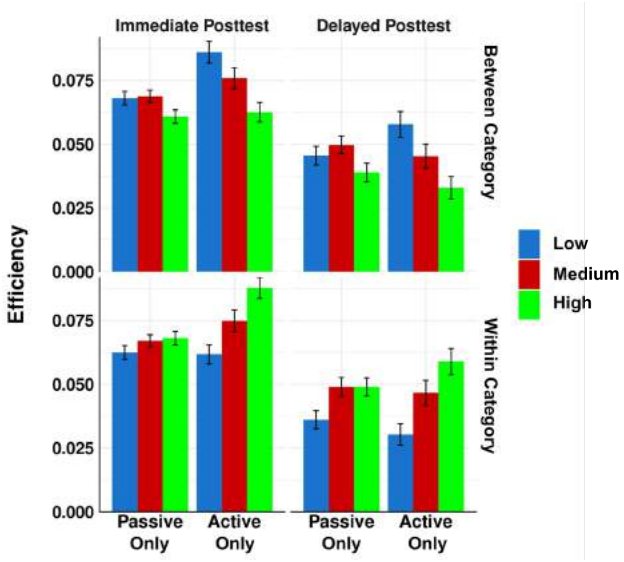


Figure 6: Efficiency for between-category similarity groups (top) and within-category similarity groups (bottom) for low, medium and high similarity, by *Passive Only* and *Active Only* conditions at immediate and delayed posttests.

effect of within-category similarity ($F(1,176)=15.92$, $p<.001$, $\eta_p^2=.083$) and an effect of Test phase ($F(1,176)=223.67$, $p<.001$, $\eta_p^2=.56$). There were two significant interactions, Condition with Similarity group ($F(1,176)=3.92$, $p=.049$, $\eta_p^2=.022$) and Condition with Test phase ($F(1,176)=6.04$, $p=.015$, $\eta_p^2=.033$).

The most instructive interaction, Condition x Similarity group, indicated that similarity relations modulated the effect of Condition. Paired comparisons indicated that differences in efficiency varied more across levels of similarity in the *Active* condition than in the *Passive* condition. Specifically, the greater the within group similarity, the greater the efficiency in the *Active Only* condition. In the *Active Only* condition, there were significant differences in learning efficiency between low within similarity and high within similarity ($t(238)=4.96$, $p<.001$, $d=0.64$), between medium within similarity and low within similarity ($t(238)=2.7$, $p=.007$, $d=0.35$), and between high within similarity and medium within similarity ($t(238)=2.13$, $p=.034$, $d=0.28$). In the *Passive Only* condition, the difference between low within similarity and medium within similarity was significant ($t(238)=2.226$, $p=.027$, $d=0.287$) and the difference between low within similarity and high within similarity was significant ($t(238)=2.388$, $p=.018$, $d=0.308$), but the difference between medium within similarity and high within similarity was not significant ($t(238)=0.136$, $p=.892$, $d=0.018$).

The ANOVA with between-category similarity included as a factor showed no significant effect of condition ($F(1,176)=1.73$, $p=.190$, $\eta_p^2=.01$), a significant effect of between-category similarity ($F(1,176)=12.34$, $p<.001$, $\eta_p^2=0.066$), and a significant effect of Test phase

($F(1,176)=236.08$, $p<.001$, $\eta_p^2=0.573$). There was one significant interaction, between Condition and Test phase ($F(1,176)=6.38$, $p=.012$, $\eta_p^2=.035$), and a marginally significant interaction of Condition x Similarity group ($F(1,176)=3.79$, $p=.053$, $\eta_p^2=0.021$). As with within-category relations, paired comparisons showed that between-category similarity modulated the effects of Condition. In the *Active Only* condition, there were significant differences in efficiency between high between-category similarity and low between-category similarity ($t(238)=4.26$, $p<.001$, $d=0.55$), between medium and low similarity ($t(238)=2.36$, $p=.019$, $d=0.31$), and a marginally significant difference between high similarity and medium similarity ($t(238)=1.94$, $p=.054$, $d=0.25$). In the *Passive Only* condition, there was one significant difference between the medium and low similarity conditions ($t(238)=2.43$, $p=.016$, $d=0.31$) and a marginally significant difference between high and low similarity conditions ($t(238)=1.76$, $p=.080$, $d=0.23$).

Discussion

The synergy of passive and active presentations in perceptual learning was remarkably similar to that found previously in factual learning (Mettler et al., 2018). In both studies the following conditions were compared: 1) passive presentations alone, 2) initial blocks of passive presentations followed by active, adaptive learning, 3) initial passive presentations for each category that unlocked later adaptive learning, or 4) active, adaptive learning alone with no passive presentations. In this experiment the learning consisted of perceptual learning across multiple categories (butterfly genera). We found an advantage for combining passive with active presentations such that initial passive presentations, especially when grouped into initial blocks of passive trials in which all learning categories were interleaved, resulted in the greatest efficiency of category classification at posttest. Learning persisted across time as measured by a 1-week delayed test. In addition, the benefits of passive and active combined schedules generalized to unfamiliar category exemplars that had not been shown during the learning phase. Unsurprisingly, combinations of passive and active presentations were better than passive presentations alone. More important, combinations of passive and active trials were much more effective than active, adaptive presentations alone: a few initial presentations (1 or 2 presentations for each category) was enough to generate learning gains beyond those found with purely active, adaptive schedules. Passive block and adaptive trial synergy was so strong that the *Passive Initial Blocks* condition at delayed test was not statistically different from the *Active Only* condition performance at immediate test. Further analysis of trial-by-trial learning data including sequences of correctness supported the idea that the benefits of a *Passive Initial Blocks* condition extended well into the active, adaptive learning component.

In addition to these results, we investigated the effect of category similarity on passive + active synergies. The overall apparent lower performance in the *Active Only* condition compared to the *Passive Only* condition appears to hold only when similarity between categories is high or when within-category similarity is low. For lower levels of between-category similarity and for greater levels of within-category similarity, *Active Only* conditions fared better than passive presentations. These effects of category similarity are somewhat different than results by Carvalho & Goldstone (2015) who showed that passive presentations result in slightly worse performance when categories have relatively low within-category similarity.[1] Unlike Carvalho & Goldstone, we found that active presentations had the greatest benefit when between-category similarity was lowest and when within-category similarity was highest. By one interpretation, high similarity between categories implies greater difficulty of making category discriminations. Thus active presentations are best when categories are more discriminable from each other. A natural interpretation of the effects in adaptive category sequencing is that with low within-category similarity (and potentially with high between-category similarity) assessments of category learning strength gotten from each active trial by the adaptive algorithm are less reliable when category instances are more diverse, making learning less efficient.

To conclude, we investigated the contribution of including passive presentations with interactive, adaptive learning. We found that combining passive with active presentations such that an initial passive phase (passive blocks) in which passive presentations were given for all learning categories resulted in the greatest retention performance at posttest. In perceptual learning, the effects of passive presentations appear to temper differences in category structure across variable within and between-category relations, and to enhance active, adaptive learning with fewer errors throughout the learning session.

Adaptive learning frameworks that leverage learner performance data to arrange spacing and sequencing in learning substantially improve learning across diverse types of learning, including perceptual learning. These benefits are further enhanced by combining active responding with passive modes of learning at the start of learning. The present results may help lead to a theoretical understanding of the mechanisms that enable passive + active synergies across different types of learning, and they contribute to a practical understanding of how to optimize these effects in instructional technology.

[1] It should be noted that blocking in Carvalho and Goldstone referred to massing exemplars from the same category, whereas in our Passive Initial Blocks condition all of the passive trials were presented as a block, but we interleaved exemplars from every category consistently in all conditions.

Acknowledgements

We gratefully acknowledge support for this work from the National Science Foundation under Grant No. DRL-1644916. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

References

Benjamin, A. S., & Bjork, R. A. (1996). Retrieval fluency as a metacognitive index. In L. Reder (Ed.), *Implicit memory and metacognition* (pp. 309–338). Hillsdale, NJ: Erlbaum.

Carpenter, S. K. (2017). Spacing effects on learning and memory, in: J.T. Wixted (Ed.), *Cognitive Psychology of Memory*. Academic Press, Oxford.

Carvalho, P. F., & Goldstone, R. L. (2015). The benefits of interleaved and blocked study: Different tasks benefit from different schedules of study. *Psychonomic Bulletin & Review*, 22, 281–288.

Cepeda, N.J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, 132, 354–380.

Delaney, P. F., Verkoeijen, P. P., & Spirgel, A. (2010). Spacing and testing effects: a deeply critical, lengthy, and at times discursive review of the literature. *Psychology of Learning and Motivation*, 53, 63–147.

Karpicke, J. D., & Bauernschmidt, A. (2011). Spaced retrieval: Absolute spacing enhances learning regardless of relative spacing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37, 1250-1257.

Mettler, E., & Kellman, P. J. (2014). Adaptive response-time-based category sequencing in perceptual learning. *Vision Research*, 99, 111–123.

Mettler, E., Massey, C. M., Burke, T., Garrigan, P. & Kellman, P. J. (2018). Enhancing adaptive learning through strategic scheduling of passive and active learning modes. In T. T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.), Proceedings of the 40th Annual Conference of the Cognitive Science Society (pp. 768-773). Austin, TX: Cognitive Science Society.

Mettler, E., Massey, C. M., & Kellman, P. J. (2011). Improving adaptive learning technology through the use of response times. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), In Proceedings of the 33rd Annual Conference of the Cognitive Science Society (pp. 2532-2537). Austin, TX: Cognitive Science Society.

Mettler, E., Massey, C. M. & Kellman, P. J. (2016). A comparison of adaptive and fixed schedules of practice. *Journal of Experimental Psychology: General*, 145(7): 897- 917.

Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language*, 60, 437–447.

Thai, K. P., Krasne, S., & Kellman, P. J. (2015). Adaptive perceptual learning in electrocardiography: The synergy of passive and active classification. In D. C. Noell, R. Dole, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), Proceedings of the 37th Annual Meeting of the Cognitive Science Society (pp. 2350-2355). Austin, TX: Cognitive Science Society.

Comparing unsupervised speech learning directly to human performance in speech perception

Juliette Millet (juliette.millet@cri-paris.org)

Nika Jurov (nika.jurov@gmail.com)

Ewan Dunbar (ewan.dunbar@univ-paris-diderot.fr)

Laboratoire de Linguistique Formelle (CNRS – Université Paris Diderot – Sorbonne Paris Cité)
and Cognitive Machine Learning (ENS – CNRS – EHESS – INRIA – PSL Research University)
Paris, France

Abstract

We compare the performance of humans (English and French listeners) versus an unsupervised speech model in a perception experiment (ABX discrimination task). Although the ABX task has been used for acoustic model evaluation in previous research, the results have not, until now, been compared directly with human behaviour in an experiment. We show that a standard, well-performing model (DPGMM) has better accuracy at predicting human responses than the acoustic baseline. The model also shows a native language effect, better resembling native listeners of the language on which it was trained. However, the native language effect shown by the models is different than the one shown by the human listeners, and, notably, the models do not show the same overall patterns of vowel confusions.

Keywords: linguistics; language acquisition; machine learning; speech recognition

Introduction

Comparing cognitive models with human behaviour often involves some idealization. The ideal comparison between a model and a human behavioural experiment would simply have the model "participate" in the experiment, exposed to the same stimulus files as are presented to the humans, responding as if it were just another human subject. Its responses would be compared to human subjects' on a stimulus-by-stimulus level. This ideal is reached only rarely (for example, Riochet et al., 2018). Most settings either simplify the stimuli given to models (for example, showing images of objects to human participants, but providing the model instead with a discrete input indicating whether the object was a dog or a cat, as in Xu & Tenenbaum, 2007), or compare highly aggregated results rather than predictions on individual stimuli (for example, Gulordava et al., 2018). These simplifications, while often essential, may mask aspects of the real task which have a major impact on the results.

Meanwhile, a large body of recent research has proposed to evaluate acoustic models trained on speech databases, particularly those trained in an unsupervised way, using an *ABX phone discrimination task* (Schatz et al., 2013). This evaluation considers pairs of speech stimulus items (A and B) coming from two different phonemic categories, assessing whether the model's representation of a third stimulus (X) is more similar to its representation of A or of B.

While this task is analogous to the standard human ABX perception task, a direct comparison of the two to evaluate models or better understand human behaviour has not yet been done. We propose a direct, stimulus-by-stimulus comparison of an acoustic model with human perception in an ABX perception task. Additionally, the stimuli for our task come from two different languages. We examine the behaviour of human subjects, and trained models, for whom one of the languages is a second language (L2). Previously, unsupervised acoustic models have typically been evaluated by assessing how well they discriminate phonemes of the language on which they are trained (L1), their objective being to reach perfect discrimination of all pairs of phonemes in the L1 (Schatz et al., 2013; Versteegh et al., 2015). A few studies have investigated patterns of L2 discrimination in acoustic models, looking at overall accuracy on phonemic contrasts from languages other than the training language. But their conclusions have been based on qualitative summaries of the behaviour of the models, with no human reference data on the same stimuli (Schatz et al., 2017; Schatz & Feldman, 2018).

A stimulus-by-stimulus comparison of an acoustic model with human performance on a speech perception task might reveal major differences between the two. If a trained acoustic model is seen as an acoustic baseline, the comparison will highlight aspects of human speech perception which are surprising given properties of the signal alone. On the other hand, if the goal of the acoustic model is to be human-like, such a comparison shows us where the model falls short.

We train an unsupervised acoustic model which is known to perform globally well on corpus-based ABX evaluations (Chen et al., 2015). We train the model on English and French corpora. We expose both the English-trained model and the French-trained model to novel, experimental stimuli. We evaluate the models' ABX discrimination accuracy. We give English and French human native listeners the same task.

Our results show that the model is globally more predictive of the human results than a baseline based on low-level acoustic features. The model also shows a native language effect: when trained on French, its error pattern is more like French native speakers', and similarly for English. However, we analyze these error patterns, and show that the native language effects shown by the models, while globally predictive, differ importantly from those shown by the human participants.[1]

[1] All modelling code, analysis code, stimuli, and anonymized

Methodology: Human ABX evaluation

In an ABX paradigm, participants hear three sounds in sequence, and indicate which of the first two sounds (A or B) is more similar to the last (X), a sound always drawn from the same category (for example, phoneme) as either A or B. The task is intended to tap the perceptual similarity between A and X, on the one hand, and B and X, on the other, to assess the overall distinctness of the categories A and B belong to.

We develop stimuli to test cross-linguistic (English/French) perception of vowels in an ABX discrimination paradigm. Within each stimulus triplet, A and B always consist of CVC non-words contrasting one English vowel with one French vowel, with the flanking consonants held constant. We use the American English vowels [ɪ], [ʌ], [ʊ], and [æ], and the Hexagonal French vowels [a], [ɔ], [ɛ], [i], [u], [y], and [œ].[2] Only consonants appearing in both languages are used: [v], [z], [s], [ʃ], [f], in both consonant positions, and, additionally, [p], [b], [g], and [k] in coda.[3] While the stimuli are designed to differ only in the vowel, there are inevitable phonetic differences in the realization of these consonants across the two languages, which may provide additional cues to the correct answer. Real words in either language are excluded. For details of stimulus construction, see **Experiments: Humans** below.

We expect that human listeners will vary in their discrimination ability, with triplets like [vip]–[væp]–[vip] being generally more difficult than more acoustically similar triplets such as [vʌp]–[vɔp]–[vʌp]. We also expect cross-linguistic differences, with English listeners doing better than French listeners on acoustically similar contrasts which do not exist in French, such as [i]–[ɪ]. We examine the patterns of confusions shown by both listener groups, and present the same experimental stimuli to models trained on English and on French, to evaluate the models' internal representations.

Methodology: Model ABX evaluation

Unsupervised acoustic models are models that learn representations of speech by exposure to speech without associated phonemic category labels. They can be seen as learning the organization of a perceptual space for speech.

We train a Dirichlet Process Gaussian Mixture Model (DPGMM) as an acoustic model. It is a non-parametric Bayesian clustering model. It finds, in an unsupervised way, a set of multi-dimensional Gaussian distributions appropriate to cluster the observations (here acoustic features). It adapts its number of Gaussian distributions automatically depending on the training data. The computations needed by the model training can be parallelized (Chang & Fisher III, 2013), making training on a reasonable amount of speech data possible. The resulting trained model (learned set of Gaussian distributions) can then be applied to any new speech example, yielding a sequence of probability vectors that can be seen as the model's perceptual representation of the example. In this way, the model can be seen as learning the organization of a perceptual space. Chen et al. (2015) applied parallel DPGMM training and achieved the best performance in the 2015 ZeroSpeech Challenge, a machine learning challenge seeking state-of-the-art unsupervised acoustic models (Versteegh, Anguera, Jansen, & Dupoux, 2016).

The representations we extract from the DPGMM model are posteriorgrams. A speech signal consists of a sequence of audio frames: for a sequence of k audio frames, a posteriorgram is a sequence of k vectors. The vector $\mathbf{x_i} = (x_1, x_2, ...x_N)$ gives the probabilities of the i^{th} frame having been generated by each of the model's N learned Gaussian distributions.

Performing ABX evaluation of an encoding learned by an acoustic model relies on extracting the representations of triplets of stimuli (A, B, and X), and computing the distance $d(\mathrm{A},\mathrm{X})$, between A and X, and $d(\mathrm{B},\mathrm{X})$, between B and X. X is of the same category as either A or B. Taking A to be the correct answer, we compute $\delta = d(\mathrm{B},\mathrm{X}) - d(\mathrm{A},\mathrm{X})$. If $\delta > 0$, we can consider the model to have chosen A; if $\delta < 0$, we consider it to have chosen B. In previous work evaluating acoustic models with this method (Versteegh et al., 2015; Dunbar et al., 2017), the percentage of correct responses for each pair of categories is tabulated, and these averages are combined into a global ABX discriminability score.

Because it relies only on computing distances, the model ABX evaluation is applicable to a broad variety of learned representations. It can be applied to posteriorgrams, but also to Mel-frequency cepstral coefficients (MFCCs), a compact representation of acoustic cues derived from the spectrum, commonly used to train ASR models. We train our models here on MFCC inputs, and MFCCs also serve as our low-level acoustic baseline (each audio frame is a MFCC vector).

The distance function most appropriate for the comparison may vary as a function of the type of representation. Because the representations we evaluate contain one vector per audio frame, differing-length stimuli will have different-length representations. To deal with those differences, we follow previous literature in the domain and use dynamic time warping (DTW) to align the sequences (see Senin, 2008 for a review). This algorithm computes an optimal match between two sequences based on a secondary distance function used for comparing individual elements across the two sequences (individual vectors in the speech representations). Every frame in each of the two representations is matched with at least one frame in the other representation, following the order of each sequence. The final distance between the two sequences is the mean of the distance between the matched frames.

As secondary distance functions, we use the same frame-level distances as in previous evaluations of DPGMM acous-

data for this paper are available in the following online repository: `https://github.com/geomphon/CogSci-2019-Unsupervised-speech-and-human-perception.git`.

[2]This reduced set of vowels is constructed with special attention to French native listeners' perception of the English vowel [ʌ]. Previous research shows (Peperkamp, 2015) that French native listeners identify this vowel with a number of different French vowels, suggesting that a fair number of pairs will be difficult for subjects.

[3]Stops are excluded in onset position because of the marked differences between English and French VOT.

tic models. For MFCC representations, we use the cosine distance. For N-dimensional vectors $\mathbf{x}$ and $\mathbf{y}$, it is defined as:

$$D_{cos}(\mathbf{x},\mathbf{y}) = \frac{1}{\pi}\arccos\left(\frac{\sum_{i=1}^{N} x_i y_i}{\sqrt{\sum_{i=1}^{N} x_i^2}\sqrt{\sum_{i=1}^{N} y_i^2}}\right)$$

For comparing the posteriorgrams of our trained models, we use the symmetrized Kullback–Leibler (KL) divergence. For positive[4] N-dimensional vectors $\mathbf{x}$ and $\mathbf{y}$, the symmetrized KL-divergence between x and y is:

$$D_{KL}(\mathbf{x},\mathbf{y}) = \frac{1}{2}\left[\sum_{i=1}^{N} x_i \log(\frac{x_i}{y_i}) + \sum_{i=1}^{N} y_i \log(\frac{y_i}{x_i})\right]$$

Although this model ABX task is inspired by a speech perception task, the test is different from a typical speech discrimination experiment in an important way. By tabulating the proportion of triplets with $\delta > 0$ (correct), gradient information about individual stimuli triplets is lost. Such a test cannot measure how well separated or "discriminable" individual speech stimuli are, but only the separation of a pair of categories A and B. Rather than directly using model ABX discriminability scores, we relate human discrimination of individual stimuli to δ; see below, and see also Schatz, 2016.

Experiments: Humans

The stimuli were recorded in a carrier phrase. Six speakers read the stimuli in an anechoic chamber. Two were early bilinguals of American English and Hexagonal French, both female, and read both the English and the French vowel stimuli. Both had extensive exposure to both languages throughout most of their early and adult lives, and regularly used both languages. These stimuli were used for A and B. The other four speakers were male: two North American English natives, who read the English stimuli, and two Hexagonal French natives, who read the French stimuli. Their productions were used as X. Phonetically trained listeners (one French and one English native), listened to the stimuli in isolation and verified that they were native-like in the target language and corresponded to the intended vowel.

All A and B pairs were cross-language comparisons. If A was a French stimulus, B was English, and vice versa. The A and B speakers always differed. The experiment used 500 ms silence for both the A–B and B–X intervals.

The final set of stimuli consisted of 112 triplets, matched to the same intensity, downsampled to 16000 Hz. The list was a subset of the complete set of possible triplets, optimized to balance combinations of speaker, vowel pair, consonantal context, and whether A or B was the correct answer. Each vowel pair appeared four times, factorially combining which of the two vowels was the correct answer, and whether the correct answer was presented first (A) or second (B).

The task was performed on Amazon Mechanical Turk with the LMEDS software (Mahrt, 2016), with participants from the United States and France. Listeners were paid for participation. Previous research shows that Mechanical Turk can successfully be used for speech perception tasks, and that results are comparable to a lab setup (for example, Kleinschmidt & Jaeger, 2015). We asked the participants to use headphones, to do the task in a quiet environment, and to check the sound volume before the experiment began.

A total of 144 participants were tested, 72 in France and 72 in the United States. We filter out those who did not finish the task, did not report English or French as their first language, had previously taken a linguistics class, failed two out of three catch trials[5] or reported hearing or vision problems. In the end, there were 63 English and 55 French participants.[6]

Experiments: Models

To build the models for comparison with the human experiment, we train the DPGMM on the same LibriVox audio book source corpora used to construct the English and French data sets in the 2017 ZeroSpeech Challenge (Dunbar et al., 2017). We use a different subset of the corpora than the one used previously, to construct two data sets of comparable size. Our English data set is made of 34 hours and 8 minutes of read speech, and our French dataset contains 33 hours and 42 minutes of read speech. Recordings were sampled at 16000Hz.

We use Kaldi (Povey et al., 2011) to pre-process the data: we extract 13-dimensional MFCCs (25 ms analysis window, 10 ms window shift), to which we apply a vocal tract length normalization (VTLN). We add the Δ and $\Delta\Delta$ for a total of 39 dimensions, and apply centered windowed mean normalization (with a window size of 300 frames).

For each corpus, we use 90% of the data for training and 10% as a validation set. We obtain two models, one for each dataset (**English-DP**, **French-DP**). Model training is stopped after 1500 iterations, as in Chen et al., 2015. We obtain 611 clusters for **English-DP**, and 1565 for **French-DP**.

The **English-DP** and **French-DP** models are applied to the one-second and ten-second test stimuli from the across-speaker condition of the 2017 ZeroSpeech Challenge (also drawn from the LibriVox corpora) and subjected to the corresponding ABX evaluation. We test the French model on the French stimuli and the English model on the English stimuli. The ABX triplets are each made up of a sequence of three extracts of speech from the stimuli, where each extract consists of a sequence of three phones, and A and B differ only in the

[4]We replace zero elements with a very small constant to avoid division by zero.

[5]Catch trials played a tone and gave an audio instruction as to which response to give.

[6]Not all participants used headphones, in spite of our instructions, and a few reported distractions; here we do not exclude these participants. Following a reviewer suggestion, we examined the results of such an exclusion, which leaves 50 English and 26 French participants. All qualitative results remain as reported. The results of this alternate analysis can be found in the online repository.

centre phone, while the context phones are held constant. All triplets constructible from the test stimuli are tested. This test serves to ensure that the models are performing as expected.

We apply each of the two models, separately, to the experimental stimuli (see **Methodology: Human ABX evaluation**), to simulate English and French native listeners. We apply the same pre-processing steps as were applied to the training corpora, transform the files into DPGMM posteriorgrams from the trained models, and obtain only the frames corresponding to the stimuli.[7] We calculate δ for each triplet, for each of the models, and for the MFCC representations.

Results: Humans

The overall ABX discrimination accuracy across all stimuli, across all participants, is 72%. The English listeners obtain a score of 69%, and the French listeners 75%. Figure 1 shows the average accuracy across vowel pairs.[8]

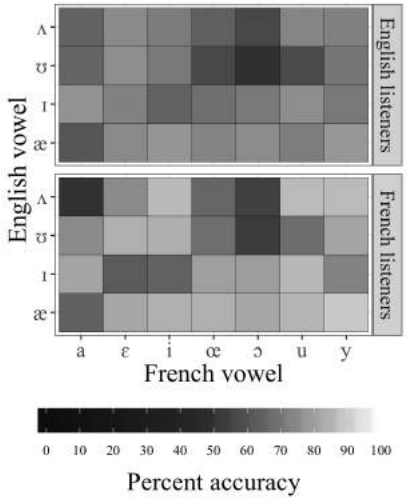


Figure 1: Human accuracy (English and French listeners) averaged by vowel pair. Lighter indicates higher accuracy.

Before comparing the accuracies across native language group, we apply a correction to make the groups' scores comparable. We numerically remove effects of response bias, potential bias to respond A or B, and overall group-level baseline accuracy. We quantify these nuisance effects using a generalized linear model. We fit a probit regression because of its interpretation as a d-prime analysis (DeCarlo, 1998; Macmillan & Creelman, 2004) using the *lme4* package for R (Bates et al., 2015). We code responses as 1 (accurate) or 0 (inaccurate). The model contains an intercept and a random intercept by subject, modelling response bias; a main effect of subject group (English: −1, French: 1), modelling group-level differences; an effect of A/B presentation order (A correct: −1, B correct: 1), modelling tendencies to respond A or B; and an interaction of these last two. We correct each observation by subtracting the predicted probability of correct response. We average the corrected responses within each stimulus triplet, and average these corrected accuracies down to the vowel pair level as before, obtaining corrected accuracies by vowel pair. Correlation between the two groups' corrected accuracy at the stimulus triplet level is 0.63. After averaging to the vowel pair level, the correlation is higher, at 0.79, indicating that many group differences are due to effects of individual stimuli, rather than the vowel contrasts we intended to test. The vowel pairs are compared in Figure 2.

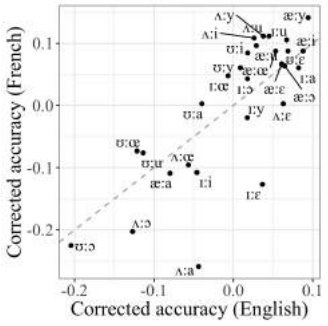


Figure 2: Discriminability of vowel pairs compared between the two language groups. The dotted line is $y = x$; pairs above the line are better discriminated by French listeners, while pairs below show better discrimination for English speakers.

Figure 2 shows that most vowel pairs were relatively well discriminated (upper right), but some were poorly discriminated by both groups (lower left). [ʌ]–[a], [ʌ]–[ɔ], and [ɪ]–[ɛ], are all perceived better by English listeners. This is consistent with Peperkamp (2015), who reports tests of French listeners on identification of English vowels, similarly indicating that, for example, [ʌ] was identified as [a], [œ], or [ɔ].

Results: Models

The scores that **English-DP** and **French-DP** obtain on the ZeroSpeech 2017 stimuli are presented in Table 1. Repeated averaging is done as for the human data, across context (flanking phones), across speakers, and then across all centre phones, to obtain a single score. We observe that the DPGMM model obtains better scores than the MFCCs, consistent with previous results. Results are reported as accuracies. **English-DP** shows 88.4% ABX accuracy on the experimental stimuli we design, and **French-DP** 86.6%, both better than **MFCC** (81.2%). Thus, the models continue to do better, globally, at discriminating speech contrasts, than the acoustic baseline, on novel recordings, from novel speakers.

[7] This was done on the longer source files, rather than directly using the short audio files used in the experiment to avoid window problems, since frames at the beginning and end of files are dropped during preprocessing. Processing the longer source files also gives the vocal-tract length normalization transformation an advantage, leading to an improvement in speaker normalization.

[8] This was a repeated average, similar to that done for the model ABX scores below: first, the accuracy across subjects for a given stimulus was calculated; then, these scores were averaged across contexts; then, across speakers. This was done for consistency with the ABX model evaluation literature (Versteegh et al., 2016; Dunbar et al., 2017).

Model	French		English	
	1s	10s	1s	10s
MFCC	74.8%	74.5%	76.6%	76.6%
French-DP	83.7 %	84.4 %	–	–
English-DP	–	–	88.8%	89.3%

Table 1: ABX accuracy for the trained models and low-level acoustic baseline on the 2017 ZeroSpeech benchmark.

Results: Model–human comparison

To compare the models as models of human perception, we ask how well the continuous machine discriminability score δ for each of the models (distance to incorrect minus distance to correct answer: see **Methodology: Model ABX evaluation**) predicts the human results. As each stimulus is associated with a δ value for a given model, good models are those for which the probability that human subjects respond correctly increases monotonically in the δ value. We compare the three δ values: **English-DP**, **French-DP**, and **MFCC**.

We begin by pooling English and French participants, to assess whether either or both DPGMM models are globally more human-like than the low-level acoustic baseline. We again use probit regression including δ as a predictor. The dependent variable is whether the subject responded correctly (1: accurate, 0: inaccurate). We fit three separate probit regressions, one per δ. Since the model includes a coefficient for δ, this can be seen as taking δ to quantify the subjects' perceived degree of distinctness for a given triplet, up to some scaling factor. We rescale the δ scores for numerical stability and for cross-model interpretability by dividing by the root mean square.[9] We again include both an overall and a (random) by-subject intercept to account for response bias, a coefficient for whether the correct answer was A or B, native language of the participants, and an interaction between these last two, plus a random intercept for individual stimulus triplet (experimental item).[10] We do not include an interaction between subject language and δ: we test for a native language effect separately below. We compare the three models using AIC (Akaike, 1974). Results are in Table 2 (smaller AIC is better). Both DPGMM models predict the human responses better than the MFCC baseline.

If the DPGMM model is really capturing adult perception, we should also expect a "native language effect": the English-trained DPGMM should show results which more closely resemble those of the English listeners than the French listeners, and the French-trained DPGMM should show results which more closely resemble those of the French listeners than the English listeners (see **Results: Humans**). We assess this as follows: we associate each human observation with the appropriate "native language" δ (**English-DP** for trials by English listeners, **French-DP** for French listeners), and with the "non-native language" δ (**French-DP** for English listeners, **English-DP** for French listeners). We construct two alternative probit regression models with the same nuisance predictors as above. In one, the independent variable of interest is the native δ score; in the alternative, the non-native δ. If the representations are equally good at predicting both groups, neither of these models should be better than the other. Results (Table 3) indicate a better fit in AIC for the native-language δ predictor (-21.12 in favour).

To verify that -21.12 is a reasonable model comparison criterion, we examine 9999 instances of the same model comparison over a randomized baseline. Each sample modifies the original data only in that the δ value considered "native" or "non-native" (**English-DP/French-DP**) is determined by a random permutation of the original native language indicator.[11] The random baseline does not yield similar improvements in AIC scores: in the baseline sample, the add-one smoothed left tail probability of -21.12 is 0.0089.

Models	French-DP	English-DP	MFCC
Coefficient for δ	0.2682	0.2790	0.1804
AIC	12675.83	12672.91	12684.15

Table 2: Regressions of human responses against machine representations, compared over the whole experiment (coefficient of δ and AIC). Lower AIC indicates better fit.

Predictor	Native δ	Non-native δ
Coefficient for δ	0.2693	0.1452
AIC	12667.98	12689.1

Table 3: Regressions of human responses against native (**French-DP** for French listeners, **English-DP** for English listeners) versus non-native (switched) trained DPGMM models (coefficient of δ and AIC). Lower AIC indicates better fit.

Discussion

Overall, the DPGMM shows itself to be a passably human-like acoustic model. Furthermore, when it is trained on subjects' native language, it predicts their responses better.

To better understand this effect, we calculate a "degree of native language effect" score for each stimulus triplet in the

[9] We keep zero in place for interpretability, as it is the decision threshold for the model ABX. Note, however, that zero is not guaranteed to be the *optimal* decision threshhold, either for predicting the correct answer in the task, or for predicting human behaviour. The inclusion of an overall intercept allows for the model to adjust to the best decision threshhold for predicting human responses.

[10] We include a stimulus-triplet level random intercept here, but not for the purpose of removing extraneous variability from the accuracy scores in generating Figure 2 above, or Figure 3 below. Those graphs are comparisons of behaviour on different items, and so item-level variability is not a nuisance factor. In contrast, here we are trying to explain away item-level variability, using δ as a predictor. It does not diminish the value of this model comparison to include a predictor capturing additional item-level variability.

[11] By permuting across the data set, we keep the unbalanced proportions of French- and English-native responses. The coefficients for subject language are still fit to the true native language of the subjects.

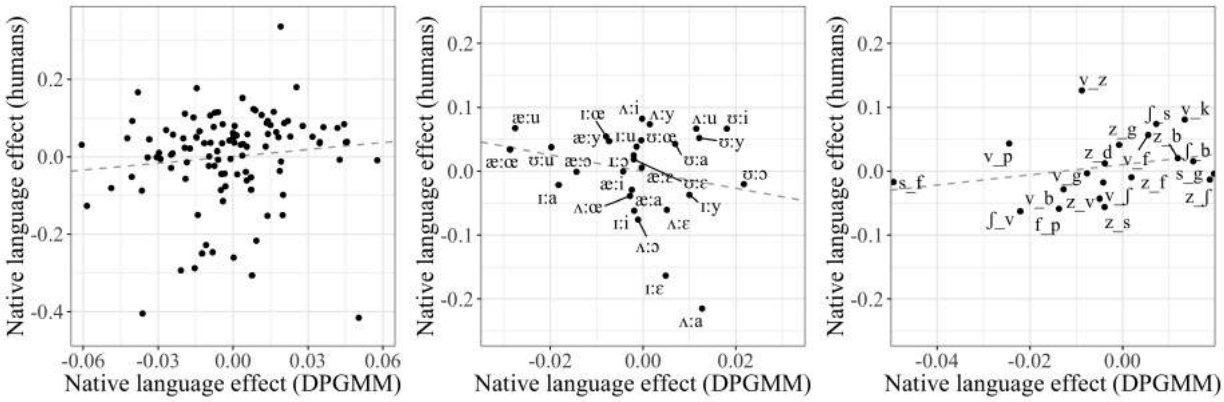


Figure 3: **(a)** Native language effect (French minus English), plotted for human accuracy against probabilities predicted from δ. Each point is one stimulus triplet. **(b)** The same points, averaged by vowel pair. **(c)** The same points, averaged by flanking consonant context. Dotted lines are linear regressions. Graphics do not show the same part of the plane, but all are on the same aspect ratio (13:7), meaning that slopes are visually comparable.

experiment, as the difference between French and English listeners' mean corrected percent accuracy (see **Methodology: Humans**). We calculate an equivalent score for the models, a predicted correct-response probability. Because the mapping between the δ values and response probabilities is indeterminate, we select an optimal mapping: we use a probit regression fit to the human data including the native-language δ as a predictor, and extract the predicted probability for each observation.[12] To isolate the part of the resulting score due to the DPGMM model itself, we subtract from each predicted probability the probability predicted by the regression if δ were zero for the given observation, obtaining a corrected probability analogous to the corrected accuracies derived for the humans above. For each stimulus triplet, we take the average corrected probability across all observations. The native language effect for the DPGMM model, for a given stimulus triplet, is the subtraction of the French and the English models' average corrected probabilities on this triplet.

These quantities are plotted against each other in Figure 3a. The slight trend towards a positive relation is consistent with the results of the model comparison, although most of the variance is unexplained. However, when averaged by vowel contrast, as in Figure 3b, it becomes clear that the native language effect in vowel confusions is not human-like: the trend in the graph is toward a negative relation. Interestingly, in Figure 3c, in which items are instead grouped by consonant frame, shows a slight positive trend, indicating human-like behaviour. But the behaviour the model captures is the fact that the impact of the flanking consonants on performance differs across listener groups. This is clearly not the behaviour we expected it it to capture: the flanking consonants were not intended to have an impact on performance at all. The fact that they contain information that facilitates the task is an artefact of the imperfectly controlled stimuli. It is also not this behaviour that makes the biggest contribution to the native language effect in humans: Figure 3 shows greater variance across vowel pairs than across consonant frames.

This unexpected effect may be due to the nature of the DPGMM model. The large number of categories it learns likely discriminate contextual variants and temporal subcomponents of individual phonemes. The participants in our experiment presumably detect coarser distinctions, beyond this sub-phonemic variability. Vowels, in particular, consist of a long steady state. The DPGMM's representation may fluctuate too much to maintain coarser-grained information. Whatever the explanation, the trained DPGMM models do not match the stimulus-by-stimulus profile of human subjects.

Conclusion

We tested human listeners, English and French native speakers, and an unsupervised acoustic model (trained once on English, once on French) on the same cross-linguistic ABX discrimination task, comparing the model with human performance on a stimulus-by-stimulus level. Our results show that the acoustic model predicts human results better than a low-level acoustic baseline, and predicts certain effects of native language on perception, while missing critical features.

We take this detailed and direct comparison to be an important step in improving the evaluation of quantitative models of human speech perception. Given that the DPGMM shows a limited, but incomplete, correlation with human speech perception, it may also prove useful as a measure of acoustic distance which is adapted to a particular language. Our approach permits detailed investigation of the differences between humans and computational models on speech perception tasks, which will be essential to using these models to gain insight into the underlying cognitive processes.

[12] We use a modified version of the "native language" regression model described in **Results: Model–human comparison**, with all nuisance predictors included, except the random effect of stimulus triplet. We exclude this for reasons discussed already: we are seeking here to examine residual differences between items.

Acknowledgements

This research was supported by the École Doctorale Frontires du Vivant (FdV) – Programme Bettencourt, and by grants ANR-17-CE28-0009 (GEOMPHON), ANR-11-IDFI-023 (IIFR), ANR-11-IDEX-0005 (USPC), ANR-10-LABX-0083 (EFL), and ANR-17-EURE-0017.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Chang, J., & Fisher III, J. W. (2013). Parallel sampling of DP mixture models using sub-cluster splits. In *Advances in Neural Information Processing Systems* (pp. 620–628).

Chen, H., Leung, C.-C., Xie, L., Ma, B., & Li, H. (2015). Parallel inference of Dirichlet process Gaussian mixture models for unsupervised acoustic modeling: A feasibility study. In *INTERSPEECH-16.*

DeCarlo, L. T. (1998). Signal detection theory and generalized linear models. *Psychological Methods*, *3*(2), 186.

Dunbar, E., Cao, X. N., Benjumea, J., Karadayi, J., Bernard, M., Besacier, L., … Dupoux, E. (2017). The Zero Resource Speech Challenge 2017. In *2017 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)* (pp. 323–330).

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., & Baroni, M. (2018). Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (pp. 1195–1205). Association for Computational Linguistics. Retrieved from `http://aclweb.org/anthology/N18-1108`

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological review*, *122*(2), 148.

Macmillan, N. A., & Creelman, C. D. (2004). *Detection theory: A user's guide*. Hove, UK: Psychology Press.

Mahrt, T. (2016). *LMEDS: Language markup and experimental design software.*

Peperkamp, S. (2015). Phonology versus phonetics in loanword adaptations. In J. Romero & M. Riera (Eds.), (Vol. 335, pp. 71–90). John Benjamins Publishing Company.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., … others (2011). The Kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding (ASRU).*

Riochet, R., Castro, M. Y., Bernard, M., Lerer, A., Fergus, R., Izard, V., & Dupoux, E. (2018). IntPhys: A Framework and Benchmark for Visual Intuitive Physics Reasoning. *arXiv preprint arXiv:1803.07616*.

Schatz, T. (2016). *ABX-discriminability measures and applications*. Doctoral dissertation, École Normale Supérieure.

Schatz, T., Bach, F., & Dupoux, E. (2017). ASR systems as models of phonetic category perception in adults. In *Proceedings of the 39th Annual CogSci Meeting.*

Schatz, T., & Feldman, N. (2018). Neural network vs. HMM speech recognition systems as models of human cross-linguistic phonetic perception. In *Proceedings of the Conference on Cognitive Computational Neuroscience.*

Schatz, T., Peddinti, V., Bach, F., Jansen, A., Hermansky, H., & Dupoux, E. (2013). Evaluating speech features with the minimal-pair ABX task: Analysis of the classical MFC/PLP pipeline. In *INTERSPEECH 2013: 14th Annual Conference of the International Speech Communication Association* (pp. 1–5).

Senin, P. (2008). *Dynamic time warping algorithm review.* Retrieved from `http://seninp.github.io/assets/pubs/senin_dtw_litreview_2008.pdf` (Ms., Department of Information and Computer Sciences, University of Hawaii)

Versteegh, M., Anguera, X., Jansen, A., & Dupoux, E. (2016). The Zero Resource Speech Challenge 2015: Proposed approaches and results. *Procedia Computer Science*, *81*, 67–72.

Versteegh, M., Thiollière, R., Schatz, T., Cao, X. N., Anguera, X., Jansen, A., & Dupoux, E. (2015). The Zero Resource Speech Challenge 2015. In *INTERSPEECH-16.*

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological review*, *114*(2), 245–272.

Explanatory Virtues and Belief in Conspiracy Theories

Patricia Mirabile (patricia.mirabile@sorbonne-universite.fr)
Sciences Normes Decisions, Sorbonne Universite
Paris, France

Zachary Horne (Zachary.Horne@asu.edu)
School of Social and Behavioral Sciences, Arizona State University
Phoenix, AZ, USA

Abstract

Conspiracy theories are "alternative" explanations of well-understood events or phenomena. What makes them attractive explanations to so many people? We investigate whether people ascribe characteristics typical of good explanations to conspiracy theories and whether they are perceived as more appealing explanations when they are articulated as a refutation of the official version of events. In two experiments, participants read explanations of four conspiracy theories and rated them along six dimensions of explanatory quality. We find that some explanatory virtues are ascribed to conspiracy theories even by people who do not believe the conspiracy. Contrary to our predictions, we also find that framing a conspiracy as a refutation did not generally elicit higher ascriptions of explanatory virtues. These results suggest that explanatory considerations may play a more central role in conspiracist beliefs than was previously thought.

Keywords: Explanation; conspiracy theories; open science

Introduction

People who believe in conspiracy theories should be characterized, or so the thinking goes, by their inability or unwillingness to identify these theories as being not only false but also as exhibiting clear epistemological flaws (e.g., Hofstadter, 1965; Robins & Post, 1997). However, this characterization obscures the possibility that people subscribe to conspiracy theories not because they are foolish but because they think that they qualify as justified beliefs (Sunstein & Vermeule, 2009), and in particular, that these theories might exhibit explanatory virtues.

The study of conspiratorial thinking is not merely an academic issue: conspiratorial thinking has real-world social and political consequences. Conspiracy theories incite acts of violence (Knopf, 2017), allow fear-mongering politicians to exert undue influence on the outcome of democratic votes (Kuzio, 2011; Nefes, 2013), scare off individuals from accessing life-preserving health care (Jolley & Douglas, 2014), and interfere with the dissemination of scientific knowledge (Goertzel, 2010).

Conspiracy theories can have lasting societal and psychological repercussions. Consequently, psychologists have begun examining the factors that are predictive of conspiratorial thinking with the hope that studying them inspires corrective interventions (Sunstein & Vermeule, 2009). Most people do not believe in conspiracy theories, but there are also individual differences in their adoption, prompting researchers to investigate what personality factors lead some people to believe in conspiracy theories (Freeman & Bentall, 2017). In the last three decades, psychologists have primarily focused on examining individual psychological differences of people who engage in so-called "conspiratorial ideation" (e.g., Swami et al., 2011; Brotherton & French, 2014). For instance, this research has examined how paranoia (Wulff, 1987), believing in the existence of paranormal phenomena, mental health disorders (Darwin, Neave, & Holmes, 2011), low levels of interpersonal and governmental trust, and political orientation predict believing in conspiracy theories (Miller, Saunders, & Farhart, 2016). This line of research is based on the observation that individuals who endorse a given conspiracy theory are more prone to endorse further (Goertzel, 1994), even contradictory (Wood, Douglas, & Sutton, 2012) or fictitious, conspiracy theories.

Philosophers have also taken an interest in understanding conspiratorial thinking, but rather than focusing on the types of people who believe in conspiracy theories, they have examined the epistemology of believing in conspiracy theories (Sunstein & Vermeule, 2009; Coady, 2006; Räikkä, 2009). This research has suggested that conspiracy theories owe their popularity to the fact that they display certain qualities (e.g., apparent simplicity, ability to produce a feeling of understanding) that are normally the hallmark of good explanations (Keeley, 1999). In particular, Keeley (1999) has suggested that conspiracy theories are often presented by their advocates as being broader than the official theory, as being able to include more phenomena in their explanation for a given phenomenon. Despite the suggestion that conspiracy theories might have a special type of explanatory appeal, there has been comparatively little research on the *features* of conspiracy theories that may make them attractive to believe (but see Wagner-Egger, Delouvé, Gauvrit, & Dieguez, 2018).

If conspiracy theories have a distinctive ability to pass for good explanations, they might draw some of their influence from their ability to satisfy what philosophers and cognitive scientists have called the human "obsession with the search for explanations" (Lipton, 2003). This would also explain why conspiracy theories tend to give rise to strong feelings of attachment in those who believe in them (Sunstein & Vermeule, 2009) and why they are often used successfully as tools of psychological manipulation by individuals (so-called "conspiracy entrepreneurs," Sunstein & Vermeule, 2009) who seek to increase their political power. An empirical investigation of the explanatory virtues of conspiracies might therefore shed light on why a substantial portion of

people—at least more than one would hope—believe in at least one conspiracy theory (Lewandowsky, Oberauer, & Gignac, 2013).

Explanatory virtues and belief

If the way conspiracy theories explain events is what makes them appealing, we may expect some of this appeal to stem from their ability to display explanatory virtues typical of good explanations. We may also expect that people who believe in conspiracy theories will be particularly sensitive to these virtues in their favored conspiracies. What virtues characterize good explanations?

Recent research has investigated the determinants of people's explanatory preferences by examining how people assess the quality of explanations they generate or consider. Some of these studies have shown a correspondence between people's preferences and the explanatory virtues identified by normative work on the epistemology of explanations (Thagard, 1978). For instance, people appear to favor qualities such as simplicity (Pacer & Lombrozo, 2017), and breadth or coherence. Other studies have revealed certain cognitive biases, i.e. preferences that do not necessarily track the goodness of an explanation, for instance the preference for explanations referring to inherent characteristics of the explanandum (e.g., Horne, Muradoglu, & Cimpian, 2019). Finally, some researchers have identified so-called "explanatory vices" (Lombrozo, 2016): these are explanatory characteristics that are mistaken for virtues and that allow flawed explanations to pass as good ones. For instance, using technical jargon can improve the apparent quality of an explanation, but it is not a reliable characteristic because it can also be used to mask the poor quality of an explanation to non-experts.

The inquiry into the nature of good explanations also bears on understanding how people reason. Given that a search for knowledge often involves the search for true explanations, what guiding principles should people trust when they reason about explanations? The theory of Inference to the Best Explanation (known as IBE, Lipton, 2003) offers such a principle: if an explanation is good enough (Lipton, 2003) and if it is better than all other rival explanations, then we are warranted to infer that that explanation is correct. Experimental work has also shown that people's beliefs can be modeled as conforming to such an inference rule. For instance, in a recent study by Douven and Mirabile (2018), subjects were asked to decide between two competing explanations for six everyday scenarios. They also rated the explanatory quality of both explanations. Two important trends were apparent in the responses: First, subjects tended to choose those explanations they judged as better explanations. Second, the quality of the competing explanation also affected the subjects' decisions: when the rival explanation was too close in goodness to the best explanation, the choice of the best explanation decreased. This latter result suggests that if a rival—but not as good—explanation is able to cast doubt on the superiority in quality of the better explanation, then it could also undermine the acceptance rates of that explanation.

Conspiracy theories are attempts to provide an explanation for events. Consequently, we might expect them to behave similarly to other cases of explanatory reasoning, that is, situations where it is reasonable to infer to an explanation if it is better than all available competitors. Following the results from Douven and Mirabile (2018), we predict that people will think that a conspiracy theory is a true explanation when it appears to them as being the best explanation of an event, with the official version of events as a prominent competitor.

One implication of this hypothesis is that conspiracy theories should be regarded as explanations and display characteristics typical of explanations: they should be seen as good explanations by some people, otherwise they will not be considered as the *best* explanations by anyone. A second implication is that a conspiracy theory might also be able appear as the best explanation because it casts doubt on the explanatory abilities of its competitors, and in particular of the official version of events.

How could this be? First, a conspiracy theory may appear to offer a simple, broad or coherent explanation of an event, or elicit a feeling of understanding. We call these characteristics "explanatory virtues" because they are generally expected of good explanations, not because they track the objective quality of actually virtuous explanations. Second, a conspiracy theory might highlight the flaws of rival explanations (a common strategy for conspiracy theorists, Keeley, 1999), and in particular cast doubt on the superiority of the official theory. A conspiracy theory that successfully undermines its rivals might be able to enhance the appearance of displaying explanatory virtues, and thus appear as the best explanation.

The present experiments sought to explore whether some of the properties of conspiracy theories may induce people to believe in them. In particular, we investigate the hypothesis that conspiracy theories have explanatory virtues, such that people believe in them when they perceive them as being the best explanations available. We seek to test two questions. First, what explanatory virtues, if any, do people ascribe to conspiracy theories and how does their ascriptions relate to their belief in the conspiracy itself? Second, can the appeal of conspiracy theories in part be explained by their ability to produce the illusion of discrediting the official version of events? We examined these questions in two experiments.

Experiment 1

Methods

Preregistration The projected sample size, predictions, and priors used in the data analysis both for Experiment 1 and for Experiment 2 were preregistered through the Open Science Framework. Materials, experimental scripts, analyses, and data are available at https://osf.io/wh78v/.

Participants A power analysis determined that, after accounting for an expected rate of participant drop-out of 50 subjects, 375 participants would be needed in order to detect a within-subjects condition effect of Cohen's d = 0.16 (the modal effect size in social psychology) with 80% power. Therefore, we recruited 375 participants (51% women, M_{age} = 37 years old) through Amazon Mechanical Turk. After excluding participants who missed questions checking their attention, 301 participants remained in our sample. Our exclusion criteria were determined a priori and were in accordance with our experiment's preregistration.

Procedure Experiment 1 examined the relationship between belief in a given conspiracy theory and the perception of explanatory virtues in that conspiracy theory, which was framed in one of two ways (either as a direct explanation of the theory or as a refutation of the official explanation) in a within-subjects design. We selected four familiar conspiracy theories to examine how framing affected the perceived explanatory virtues in a conspiracy theory: 1) The terrorist attacks on the World Trade Center on 9/11/2001 were orchestrated by the American government, 2) Condensation trails left by airplanes contain toxic chemicals and are actually part of a weather engineering program, 3) Free and environment-friendly energy generation devices are being suppressed by oil companies, 4) Fluoride, which is added to tap water in the US, is actually an unsafe toxin.

Experiment 1 consisted of three parts: a pretest questionnaire, an explanation of a given conspiracy, and a questionnaire about the explanatory virtues of each conspiracy. After completing this portion of the experiment, participants completed demographic questions. We describe each component below.

Pretest Questionnaire We first measured how strongly participants believed in each conspiracy theory based on their prior knowledge. For instance, participants were told that the following theory has been suggested as an explanation for why the 9/11 attacks on the World Trade center occurred: "9/11 occurred because the government wanted to gain support for wars in the Middle East." Participants read this statement and indicated their agreement with it on a seven-point Likert scale. There were four such items in total (one per conspiracy theory), which were presented in a randomized order (see Table S1 of the SOM).

Conditions Experiment 1 had two conditions, which were manipulated within-subjects: the direct explanation condition, where the main arguments in favor of the conspiracy theory were explained and the refutation condition, which highlighted the shortcomings of the official version of events as an indirect way to provide evidence for the conspiracy theory (see Table S2 of the SOM). Participants only received one version (i.e., direct explanation or refutation) for each conspiracy theory, which was counterbalanced and randomized. Thus, participants received two direct explanations and two refutations of the official view. We created the materials for each condition by searching websites that contained explanations written by people who endorse the selected conspiracy theories. Based on these explanations, we constructed two (edited) short passages per conspiracy theory, one for each condition. The two passages for each conspiracy theory were approximately matched for word count ($\pm$ 15 words).

Explanatory Virtues Questionnaire After reading a given conspiracy theory, participants assessed the explanatory virtues of each of the four conspiracy theories. As noted, two of the conspiracy theories were presented in the direct explanation condition, and the two others were presented in the refutation condition. In both conditions, participants first read a short passage which explained the main theses of the conspiracy theory. Then, they rated their agreement with twelve statements about the explanatory virtues of that conspiracy on a seven-point Likert scale (see Table S3 of the SOM). We measured participants' judgments about six virtues, using two statements per virtue: simplicity, coherence, breadth, description of a mechanism, use of technical sounding language (denoted *expertise* in the figures below) and ability to induce a feeling of understanding.

Participants were instructed to assess these explanatory virtues in light of the passage they had just read rather than their personal beliefs about the conspiracy under consideration (though we nonetheless expected people's pretest beliefs to be related to their virtue ratings). For instance, after reading a passage about the chemtrails conspiracy, participants rated how strongly they agreed with statements such as "this theory is a clear and easy to understand explanation for [phenomenon]" (virtue = feeling of understanding) or "this theory provides a complete explanation for [phenomenon]" (virtue = breadth). The order of presentation of these twelve statements was randomized. After reading the passage that described a given conspiracy and providing their ratings, participants advanced to the next conspiracy theory and completed the questionnaire again.

Predictions In Experiment 1, we sought to answer three questions. First, will participants in the refutation condition be more likely to judge that it has explanatory virtues than participants in the direct explanation condition (main effect of condition)? Second, to what extent, if any, would this tendency depend on the virtue in question (Condition $\times$ Virtue interaction)? Third, even if a participant does not believe in a given conspiracy, what virtues if any would they think the conspiracy nonetheless has?

Results and Discussion

We tested our predictions by fitting two Bayesian ordinal mixed-effects using the R package `brms` (Bürkner, 2017). Both models estimated explanatory virtue ascriptions, treated pretest belief predictor as a monotonic effect and included group-level effects which we detail below. Because of the

exploratory nature of our analyses, we confirmed the improvement in a given model's fit using an approximation of Leave-One-Out cross-validation.

First, we tested whether conspiracy theories were more likely to be perceived as having explanatory virtues when they were framed as attempts to refute the official version of events, which was interacted with the virtue under consideration—Model 1:

```
Model 1 <- Virtue Rating ~
Virtue*Condition + mo(Pretest) +
(1 + Virtue*Condition|Subject)
```

To model the joint probability distribution of responses, we specify regularizing priors over the possible effects each parameter could have on the response variable. Model 1 priors are shown below:

$$\beta_{Intercept[1]} \sim \mathcal{N}(0.84, 1)$$
$$\beta_{Intercept[2]} \sim \mathcal{N}(2.19, 1)$$
$$\beta_{Intercept[3]} \sim \mathcal{N}(2.44, 1)$$
$$\beta_{Intercept[4]} \sim \mathcal{N}(2.75, 1)$$
$$\beta_{Intercept[5]} \sim \mathcal{N}(3.18, 1)$$
$$\beta_{Intercept[6]} \sim \mathcal{N}(3.89, 1)$$
$$\beta_{Pretest} \sim \mathcal{N}(3, 2)$$
$$\beta_{Condition} \sim \mathcal{N}(0, .5)$$
$$\beta_{\forall_{Virtues}} \sim \mathcal{N}(0, 1)$$
$$\beta_{\forall_{Virtue \times Condition\ Interactions}} \sim \mathcal{N}(0, .5)$$
$$\Omega_k \sim LKJ(1)$$
$$\text{Group-level parameters} \sim \mathcal{N}(1, 3)$$

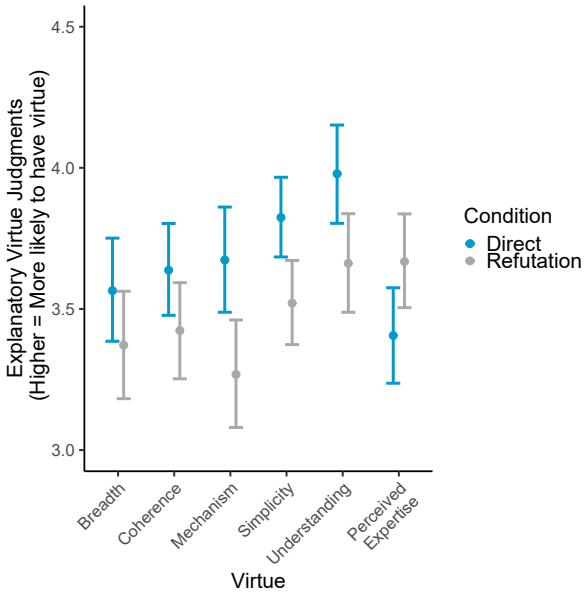


Figure 1: A marginal effects plot of ascriptions of explanatory virtues by condition (direct explanation vs. refutation). Error bars represent 95% CIs.

This analysis indicated an interaction between virtue and condition: the perceived expertise virtue received higher ratings in the refutation condition and all other virtues received higher ratings in the direct explanation condition (see Figure 1). These results contradicted our predictions: in general, participants rated conspiracy theories as presenting explanatory virtues *more* when they read a passage in the direct explanation condition but this effect did depend on the virtue in question. pretest belief in each condition.

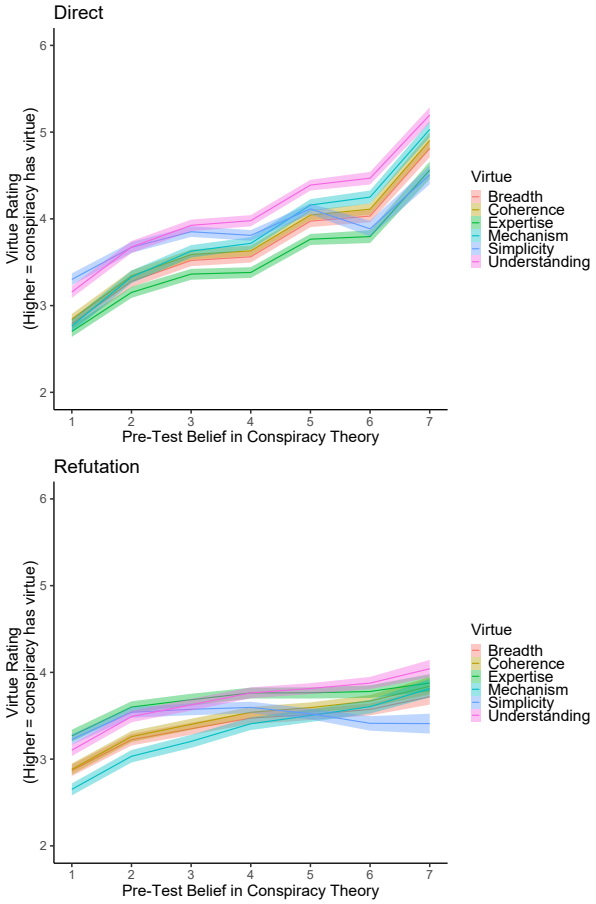


Figure 2: A marginal effects plot of ascriptions of explanatory virtues depending on pretest belief in conspiracy theory in the direct explanation and refutation conditions. Error bars display 50% CIs for legibility.

Next, we fit a model to assess whether, regardless of pretest belief in a conspiracy theory, people were more likely to think conspiracy theories had some explanatory virtues but not others, and whether this varied depending on whether the conspiracy was framed as a direct explanation or as a refutation. Model 2 regressed explanatory virtue ascriptions on the three two-way interactions between condition, explanatory virtue and pretest belief in each condition.

```
Model 2 <- Virtue Rating ~
Virtue*Condition + Virtue*mo(Pretest) +
Condition*mo(Pretest) +
(1 + Virtue*Condition|Subject)
```

Experiment 1 - Model 2 Priors:

$\beta_{Intercept[1]} \sim \mathcal{N}(0.84, 1)$
$\beta_{Intercept[2]} \sim \mathcal{N}(2.19, 1)$
$\beta_{Intercept[3]} \sim \mathcal{N}(2.44, 1)$
$\beta_{Intercept[4]} \sim \mathcal{N}(2.75, 1)$
$\beta_{Intercept[5]} \sim \mathcal{N}(3.18, 1)$
$\beta_{Intercept[6]} \sim \mathcal{N}(3.89, 1)$
$\beta_{Pretest} \sim \mathcal{N}(2, 2)$
$\beta_{Condition} \sim \mathcal{N}(0, .5)$
$\beta_{\forall_{Virtues}} \sim \mathcal{N}(0, 1)$
$\beta_{\forall_{Virtue \times Condition\ Interactions}} \sim \mathcal{N}(0, .5)$
$\beta_{\forall_{Virtue \times Pretest\ Interactions}} \sim \mathcal{N}(0, .5)$
$\Omega_k \sim LKJ(1)$
Group-level parameters $\sim \mathcal{N}(1, 3)$

This analysis revealed that ascriptions of explanatory virtues were higher in the direct explanation condition and were predicted more strongly by pretest belief than in the refutation condition (see Figure 2). Furthermore, Model 2 revealed that the virtue Understanding, for example, was more likely to be attributed even at lower-levels of pretest in both conditions relative to other virtues. Most striking, even people who did not believe in conspiracies were nearly as likely to ascribe expertise in the refutation condition as those who believed in the conspiracy theory.

Altogether, these findings suggest that stronger beliefs in a conspiracy theory are associated with higher ascriptions of explanatory virtues. However, these ascriptions did not interact with the way a conspiracy theory was framed in the way we predicted: direct explanations of the theory received higher ratings of quality than refutations of the official theory, with the exception of perceived expertise. This might be due to the fact that the passages in the refutation condition often needed to explain *details* of the official version in order to then refute them, leading participants to be more likely to ascribe expertise in this condition. However, one limitation of Experiment 1 is that participants' responses to the explanatory questionnaire hovered around the midpoint of the scale, suggesting that participants might not have had fine-grained opinions (or any opinion at all) about the virtues of a conspiracy. Therefore, in Experiment 2 we simplified the response scale to be dichotomous to confirm that our results were not simply due to unknown and problematic psychometric properties of the explanatory virtues scale used in Experiment 1.

Experiment 2

Methods

Participants Based on a power analysis and exclusion criteria identical to those from Experiment 1, we recruited 376 participants (50% women, M_{age} = 38 years old) through Amazon Mechanical Turk. After excluding participants who failed questions checking their attention, 335 participants remained in our sample.

Procedure The procedure and analytic approach were a replication of those from Experiment 1, with one key difference. In the Explanatory Virtues Questionnaire, participants were asked about each conspiracy theory: "Do you agree or disagree with the following statements describing that theory?" and responded on a dichotomous scale with "Agree" and "Disagree" as available options.

Results

We first performed logistic regression predicting virtue ratings on the basis of the interaction between Virtue and Condition controlling for pretest belief in a given conspiracy (see Model 1 formula in Experiment 1). We based our priors on the Experiment 1 - Model 1 posteriors:

$\beta_0 \sim \mathcal{N}(-0.85, .65)$
$\beta_{Pretest} \sim \mathcal{N}(1.30, 2)$
$\beta_{Condition} \sim \mathcal{N}(0, .5)$
$\beta_{\forall_{Virtues}} \sim \mathcal{N}(0, 1)$
$\beta_{\forall_{Virtue \times Condition\ Interactions}} \sim \mathcal{N}(0, .5)$
$\Omega_k \sim LKJ(1)$
Group-level parameters $\sim \mathcal{N}(1, 3)$

Experiment 2 replicated the effects we observed in Experiment 1 (see Figure 3).

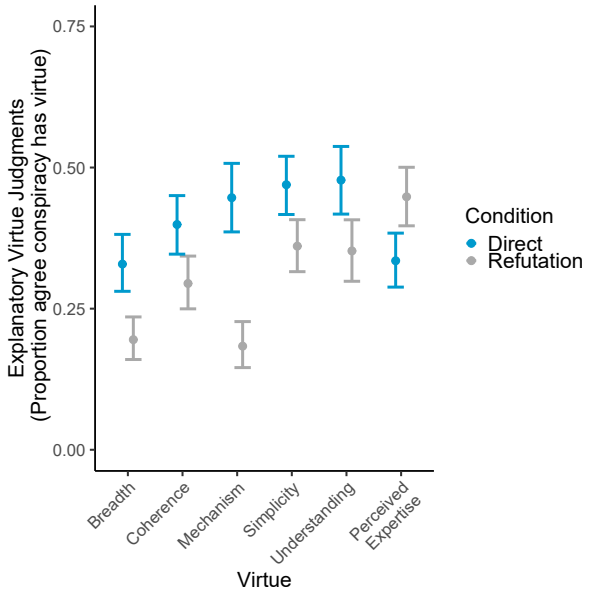


Figure 3: A marginal effects plot of ascriptions of explanatory virtues by condition (direct explanation vs. refutation). Error bars display 95% CIs.

We then tested whether pretest beliefs, virtue, and condition exhibited the three two-way interactions we observed in Experiment 1 (see Model 2 formula in Experiment 2). Priors were specified as follows:

$\beta_0 \sim \mathcal{N}(-.85, .65)$
$\beta_{Pretest} \sim \mathcal{N}(1.30, 2)$
$\beta_{Condition} \sim \mathcal{N}(0, .5)$

$$\beta_{\forall_{Virtues}} \sim \mathcal{N}(0,1)$$
$$\beta_{\forall_{Virtue \times Condition\ Interactions}} \sim \mathcal{N}(0,.5)$$
$$\beta_{\forall_{Virtue \times Pretest\ Interactions}} \sim \mathcal{N}(0,.5)$$
$$\Omega_k \sim LKJ(1)$$
$$\text{Group-level parameters} \sim \mathcal{N}(1,3)$$

This analysis revealed that ascriptions of explanatory virtues in the direct explanation condition were predicted more strongly by pretest beliefs than in the refutation condition. In the refutation condition, perceived expertise was most likely to be ascribed regardless of pretest belief in a conspiracy and people who did not believe in a conspiracy were nearly as likely to ascribe it the virtue of simplicity as people who believed in the conspiracy theory. Together, these results replicate the findings from Experiment 1.

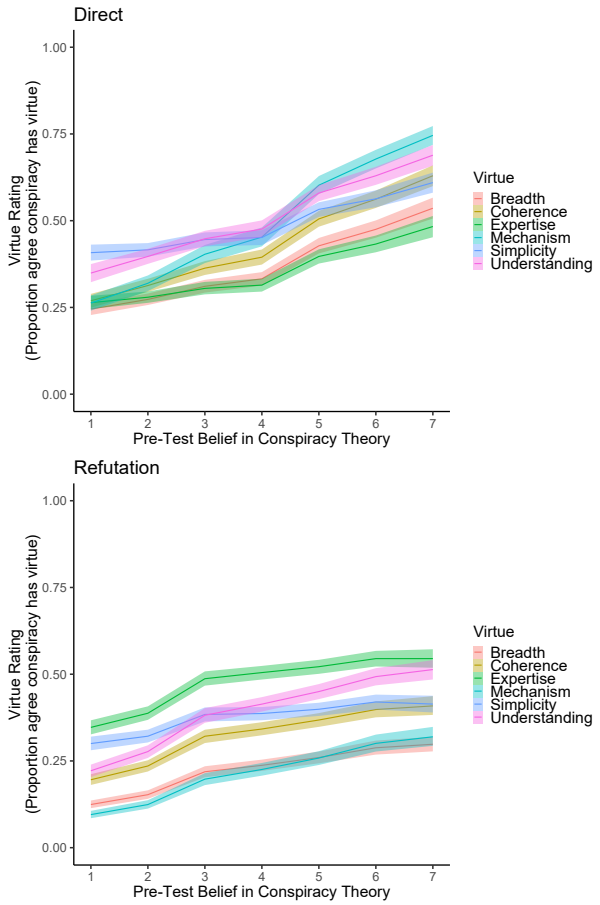


Figure 4: A marginal effects plot of ascriptions of explanatory virtues depending on pretest belief in conspiracy theory in the direct explanation and refutation conditions. Error bars represent 50% CIs.

General Discussion

Conspiracy theories are no-longer fringe beliefs (Barkun, 2016) and perhaps they never were (Goldberg, 2008). One-third of adults believe in at least one conspiracy theory (Lewandowsky et al., 2013). Here, we sought to understand the properties these theories have that may lead people to believe in them. Specifically, we investigated whether people believe conspiracy theories have explanatory virtues and whether ascription of virtues depends on how they are framed. We hoped to answer two questions: First, what explanatory virtues do people ascribe to conspiracy theories and how do they relate to belief in a given conspiracy theory? Second, can the appeal of conspiracy theories in part be explained by their ability to produce the illusion of discrediting the official version of events?

Experiments 1 and 2 indicate that people do in fact ascribe certain explanatory virtues to conspiracy theories. Although this effect is stronger for those who believe these theories, it is of note that even among participants who do not endorse a given conspiracy theory, nearly one-third of participants reliably attribute an explanatory virtue to that conspiracy theory, an effect that is more or less pronounced depending on the virtue in question and its framing.

Second, and contrary to our predictions, we found that conspiracy theories framed as refutations of the official version of events were *less* likely to be ascribed explanatory virtues. Only in the case of perceived expertise were refutations more likely to be ascribed an explanatory virtue. One possible explanation for this finding is that in order to refute the official version of events, the conspiracist also needs to provide details about the accepted theory – this often means that they need to reuse the technical language employed by the experts they criticize, which would account for the higher ascriptions of perceived expertise. Ironically, the conspiracy theory itself might have then suffered from the comparison to the accepted explanation.

What are the implications of this ascription of explanatory virtues to conspiracy theories? For everyday explanations, people are more prone to believe a hypothesis if they think it explains the available evidence well (Douven & Mirabile, 2018). Moreover, researchers have identified some explanatory qualities that are typical of preferred explanations (Lombrozo, 2016). However, in the case of conspiracy theories, psychologists have focused the irrational dimension of belief in conspiracy theories, suggesting that it points to pathological tendencies (Wulff, 1987) and constitutes a violation of epistemological or simply logical norms (Brotherton & French, 2014). Integrating our results with these analyses, the positive relationship between ascription of explanatory virtues and belief might indicate an incorrect application of inference to the best explanation: people might be led astray by the impression that a conspiracy theory has qualities typical of good explanations and thus are led to believe the conspiracy theory. Indeed, we found that participants who did not believe in a conspiracy theory still ascribed it certain explanatory virtues. For example, perceived expertise was attributed nearly 50% of the time in the refutation condition and varied little as a function of pretest belief in the conspiracy theory. Altogether, these results suggest that conspiracy theories are not perceived as

unequivocally bad explanations of events. Rather, along some explanatory dimensions they are perceived as having the same attributes as good explanations more often than we would hope, leading some people to prefer them to the official, scientifically-supported, explanations.

One limitation of these findings is that they do not allow for a comparison between the explanatory virtues of conspiracy theories and those of official explanations of events. Further research could therefore collect explanatory virtue ascriptions for conspiratorial and non-conspiratorial explanations of the same events and investigate whether they predict belief in a conspiracy theory. Identifying the explanatory virtues that make conspiracy theories more appealing than their official counterparts would be an important step for assisting scientists and governmental agencies interested in debunking misinformation.

References

Barkun, M. (2016). Conspiracy theories as stigmatized knowledge. *Diogenes*.

Brotherton, R., & French, C. C. (2014). Belief in conspiracy theories and susceptibility to the conjunction fallacy. *Applied Cognitive Psychology*, *28*(2), 238–248.

Bürkner, P.-C. (2017). brms: An r package for bayesian multilevel models using stan. *Journal of Statistical Software*, *80*(1), 1–28.

Coady, D. (2006). *Conspiracy Theories: The Philosophical Debate*. Ashgate.

Darwin, H., Neave, N., & Holmes, J. (2011). Belief in conspiracy theories. the role of paranormal belief, paranoid ideation and schizotypy. *Personality and Individual Differences*, *50*(8), 1289–1293.

Douven, I., & Mirabile, P. (2018). Best, second-best, and good-enough explanations: How they matter to reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(11), 1792–1813.

Freeman, D., & Bentall, R. P. (2017). The concomitants of conspiracy concerns. *Social Psychiatry and Psychiatric Epidemiology*, *52*(5), 595–604.

Goertzel, T. (1994). Belief in conspiracy theories. *Political Psychology*, *15*(4), 731–742.

Goertzel, T. (2010). Conspiracy theories in science: Conspiracy theories that target specific research can have serious consequences for public health and environmental policies. *EMBO reports*, *11*(7), 493–499.

Goldberg, R. A. (2008). *Enemies Within: The Culture of Conspiracy in Modern America*. Yale University Press.

Hofstadter, R. (1965). *The Paranoid Style in American Politics*. Vintage.

Horne, Z., Muradoglu, M., & Cimpian, A. (2019). Explanation as a cognitive process. *TICS*.

Jolley, D., & Douglas, K. M. (2014, 02). The effects of anti-vaccine conspiracy theories on vaccination intentions. *PLOS ONE*, *9*(2), 1-9.

Keeley, B. L. (1999). Of conspiracy theories. *The Journal of Philosophy*, *96*(3), 109–126.

Knopf, T. A. (2017). *Rumors, Race and Riots*. Routledge.

Kuzio, T. (2011). Soviet conspiracy theories and political culture in ukraine: Understanding viktor yanukovych and the party of regions. *Communist and Post-Communist Studies*, *44*(3), 221–232.

Lewandowsky, S., Oberauer, K., & Gignac, G. E. (2013). Nasa faked the moon landing therefore, (climate) science is a hoax: An anatomy of the motivated rejection of science. *Psychological science*, *24*(5), 622–633.

Lipton, P. (2003). *Inference to the Best Explanation*. Routledge.

Lombrozo, T. (2016). Explanatory preferences shape learning and inference. *TICS*, *20*(10), 748–759.

Miller, J. M., Saunders, K. L., & Farhart, C. E. (2016). Conspiracy endorsement as motivated reasoning: The moderating roles of political knowledge and trust. *American Journal of Political Science*, *60*(4), 824–844.

Nefes, T. S. (2013). Political parties' perceptions and uses of anti-Semitic conspiracy theories in Turkey. *The Sociological Review*, *61*(2), 247–264.

Pacer, M., & Lombrozo, T. (2017). Ockham's razor cuts to the root: Simplicity in causal explanation. *Journal of Experimental Psychology: General*, *146*(12), 1761.

Räikkä, J. (2009). On political conspiracy theories. *Journal of Political Philosophy*, *17*(2), 185–201.

Robins, R. S., & Post, J. M. (1997). *Political Paranoia: The Psychopolitics of hatred*. Yale University Press.

Sunstein, C. R., & Vermeule, A. (2009). Conspiracy theories: Causes and cures. *Journal of Political Philosophy*, *17*(2), 202-227.

Swami, V., Coles, R., Stieger, S., Pietschnig, J., Furnham, A., Rehim, S., & Voracek, M. (2011). Conspiracist ideation in Britain and Austria: Evidence of a monological belief system and associations between individual psychological differences and real-world and fictitious conspiracy theories. *British Journal of Psychology*, *102*(3), 443–463.

Thagard, P. R. (1978). The best explanation: Criteria for theory choice. *The Journal of Philosophy*, *75*(2), 76–92.

Wagner-Egger, P., Delouvé, S., Gauvrit, N., & Dieguez, S. (2018). Creationism and conspiracism share a common teleological bias. *Current Biology*, *28*(16), R867 - R868.

Wood, M. J., Douglas, K. M., & Sutton, R. M. (2012). Dead and alive: Beliefs in contradictory conspiracy theories. *SPPS*, *3*(6), 767–773.

Wulff, E. (1987). Paranoid conspiratory delusion. *Psychiatrische Praxis*, *14*, 14–22.

Statistical Learning of Conjunctive Probabilities

Di Mo (di.mo@mail.utoronto.ca)
Department of Psychology
University of Toronto

Blair C. Armstrong (blair.armstrong@utoronro.ca)
Department of Psychology
University of Toronto

Abstract

Most statistical learning studies focus on the learning of transitional probabilities between adjacent elements in a sequence, however, other statistical regularities may underpin different aspects of processing language and regularities in other domains. Here, we investigate how conjunctive statistical regularities (of the form A and B together predict C) can be learned, and how this learning is impacted by similarity in representations analogous to that in unambiguous words, homonyms with multiple unrelated meanings, and polysemes with multiple related meanings. We observed that provided the stimulus structure is relatively simple, participants are readily able to learn conjunctive probabilities and display sensitivity to relatedness among representations. These results open new theoretical possibilities for exploring the domain-generality of how the learning and processing systems merge conjunctive information in simple laboratory tasks and in natural language.

Keywords: Statistical Learning; Lexical Ambiguity; Transitional Probability; Conjunctive Probability

Introduction

Statistical learning has been proposed as a powerful mechanism for how individuals learn regularities across time and space. Foundational work by Saffran, Newport, and Aslin (1996) first established human sensitivity to transitional probabilities (TPs) in identifying word boundaries in streams of auditory syllables. Most research on this subject to date has focused on variations of TPs such as non-adjacent dependencies (Gómez, 2002) and visual co-occurrences across scenes (Fiser & Aslin, 2001), illustrating a range of applications for statistical learning. While fundamental, the various forms of TPs do not account for all types of statistical regularities that must be learnt to explain other types of behaviours. For example, learning something akin to a conjunctive probability (CP) may be important in explaining how individuals learn to disambiguate the meanings of semantically ambiguous words in natural language. To illustrate, the word BAT can refer to either an animal or to sporting equipment, and the correct meaning of this word is extracted by integrating the constraints on overall meaning offered by BAT with the broader context (e.g., a discussion about baseball).

The present work sought to investigate several major issues that relate to learning CPs, as they might relate to natural language statistics such as those relevant to word meaning disambiguation. The first was how different elements in a stream could be more or less constraining on the expected outcome of a conjunction. For example, in natural language, knowing that the topic of conversation is "SPORTS" provides only vague constraint on what particular meaning should be evoked in a sentence. This knowledge therefore provides only low constraint (high entropy) in determining which particular meaning should be evoked (e.g., the discussion could relate to hockey, baseball, etc.). In contrast, the word "BAT" provides relatively high constraint (low entropy) on what meaning should be evoked (it should relate either to "baseball" or to "flying mammal"). Furthermore, critical to present purposes, only by combining both of these elements can a context-specific interpretation of a word be evoked. Using this analogy to words (which are low entropy), contexts (which are high entropy), and context-specific meanings (which are fully determined by the combination of the previous two elements) we examined how low- and high-entropy items combined to predict an upcoming element. In a related vein, we also examined how the order in which low- versus high-entropy information is presented shaped performance. How is the process of computing CPs impacted by having more versus less constraint early in processing?

Additionally, unlike typical statistical learning research which employs highly and equally distinct elements during learning, we also explored how representational similarity could shape performance in computing a CP and relate to word disambiguation processes. In the case of natural language, the semantic ambiguity continuum can be broken down into three main subdivisions: (1) unambiguous words like CHALK which evoke effectively the same meaning in different contexts. That is, the word itself predicts the meaning with 100% accuracy, the context does not provide any additional unique information. (2) homonyms such as BANK which evoke completely distinct meanings in different contexts. That is, the word narrows the meaning down to two completely distinct interpretations, but context is necessary to select among those representations. (3) polysemes such as CHICKEN, which evoke related representations (in this example, the animal or its meat) in distinct contexts. That is, the word alone may predict the majority of the evoked representation, but context is needed to select

exactly the right interpretation.

With these aims in mind, we developed a variant of a standard self-paced statistical learning paradigm that allowed us to contrast standard TP learning with the learning of CPs between low-entropy items (analogous to words) and high-entropy items (analogous to contexts) in predicting a third item (analogous to context-sensitive meaning). We also employed representations that varied in their similarity to one another to assess the impact of meaning relatedness on learning. Performance was assessed using a combination of online and offline measures of learning. In so doing, we aimed to contribute to knowledge of how a broader range of statistics such as conjunctive probabilities can be incorporated into general theories of statistical learning. We also aimed to connect this work with important statistical properties that are at the heart of other areas of cognition such as semantic ambiguity resolution. If successful, this work could open new possibilities for how artificial language learning experiments using statistical learning paradigms could complement existing studies of semantic ambiguity in natural language, for example, by allowing the development of well controlled artificial languages that avoid the complex confounds in natural language stimuli used to study semantic ambiguity (Armstrong & Plaut, 2016).

Experiment 1

The first experiment served as a baseline for evaluating how the learning of standard triplet structures with perfect predictability (TPs of 1) across successive items takes place using our specific experimental procedure. We then use these results as a platform for understanding the impact of ambiguity on processing in subsequent experiments using variations of the same basic design but changing the probability structure between elements.

Methods

Participants A total of 60 participants (16 male; mean age=20) completed the experiment. All participants were undergraduate students from the University of Toronto participant pool and were compensated with course credit. All completed an informed consent and debriefing procedure.

Materials A total of 48 images of unusual objects (hereafter, symbols) were the targets for learning in the experiment. These symbols were selected so as to not have clear verbalisable labels, and therefore encourage learning of the statistics between the visual representations of each element. These symbols were used to create sequences containing two single-symbol elements and one four-symbol complex element. Eight such simple-simple-complex sequences with unique elements were randomly generated for each participant. The use of varying complexity across visual elements (one symbol vs. four symbols) allows us to assess the impact of visual complexity per se, and also enables rich variation in the statistical structure of the relationship between elements and symbols in the subsequent experiments.

Procedure The experiment was administered on desktop computers using PsychoPy (v1.85.4).

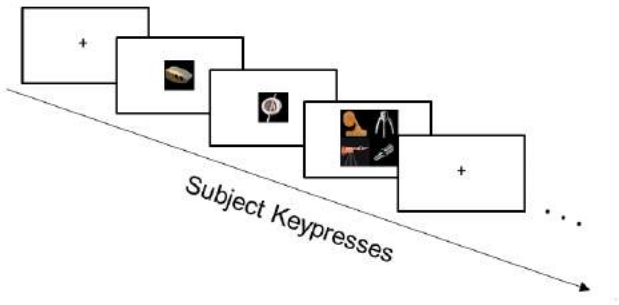


Figure 1: Familiarisation

Familiarisation/On-line Learning Participants were exposed to 30 randomised sweeps through the eight sequences and were instructed to pay attention to the order of the elements. A fixation cross was presented between sequences to focus learning on the relationships between elements (see Figure 1). The task was self-paced and participants advanced through the elements by pressing the space key. The time spent on each element was recorded. On average, the familiarisation task took approximately 20 minutes to complete.

Off-line Tests Two offline tasks were used to assess learning. The first was a sequence completion task, in which participants had to complete a missing element in a sequence. Participants selected from among four choices for completing the first and last element, and two choices for completing the middle element. This corresponded to later experiments where one of the first two elements had only two valid possibilities. The presented choices all came from the same position in a sequence, sampled from among the different sequences (e.g., the choices were always taken from position 1 when completing a missing element from position 1). Eight questions each were asked about the first two elements and 12 questions were asked about the third element. The four extra questions about the third element in this experiment were only included in order to match the number of questions used in subsequent experiments regarding CPs (as described later). Test questions were blocked by order of position in the sequence.

The second task had participants choose from among four sequences which was the most familiar. One of these sequences was actually seen during familiarisation, the others were made-up sequences that mixed elements from different sequences while preserving position in a sequence (e.g., a sequence would be made up of an element selected at random from all elements in position 1 across

sequences, an element selected at random from position 2 across sequences, etc.). Sequences were presented one element at a time at a fixed rate of one element per second. Six questions were asked: two for coarse-grained distinction, where all non-target sequences comprised entirely unfamiliar combinations of elements; four for fine-grained distinction, which included a distractor item containing two elements from one sequence combined with one element from another sequence. Again, number of questions were matched to those of subsequent experiments on conjunctive probability. While only one element is needed to predict a sequence in TP, subsequent conjunctive probability experiments will require looking at two elements together to predict the third.

Results

Due to space constraints, we report only the differences that were significant at p¡0.05. Error bars in graphs denote standard error. For this experiment, it is expected that there is a speed up between Elements 1 and 2 as the first element is unpredictable while the second element is perfectly predictable from the first element. It is also expected that the last element requires more processing effort due to higher visual complexity.

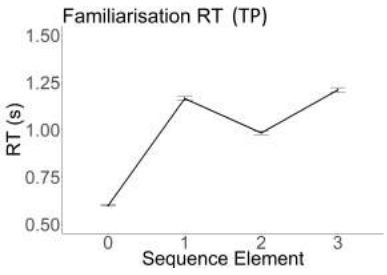


Figure 2: Element 0=fixation; 1-3=sequence

Familiarisation The average reaction time (RT) during online familiarization to the sequences is presented in Figure 2. We used mixed-effect linear models with random intercepts for participants to test for differences in RT across sequence elements (positions) 1-3. Participants sped up between Elements 1 and 2 but slowed down between Elements 2 and 3 such that Element 3 took significantly longer time to process than Element 1.

Offline Test One sample t-tests showed participants had learnt all three elements in the sequence above chance performance in the sequence completion task, as reflected in their accuracy in questions regarding each element. Note that due to the aforementioned difference in number of options at test (but not in training), chance is 0.25 for Elements 1 and 3 and 0.5 for Element 2 (Figure 3). To compare relative learning across the sequence, accuracy was modeled with items and participants as random intercept. Significant differences were found among all three elements, with the highest accuracy at Element 2 and lowest accuracy at Element 3. However, due to difference in chance level, only Elements 2 and 3 can be directly compared. Above-chance performance was also found for all elements in sequence selection for familiarity. There were no significant differences between coarse-grained and fine-grained test items.

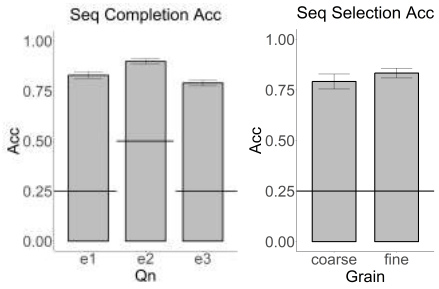


Figure 3: Solid lines mark chance performance.

Discussion

Experiment 1 showed that participants were sensitive to the statistical properties associated with each sequence element in the online learning measure. Participants displayed good performance (˜75% correct) on all the elements in the offline test. Both of these results are consistent with a similar prior study by (Siegelman, Bogaerts, Kronenfeld, & Frost, 2018). In contrast to that experiment, however, participants exhibited overall slower responses for the final element in the sequence, which we attribute to the increased visual complexity of that item. These results provide an important measure of baseline performance in the task to evaluate the impact of CP learning in the following experiments.

Experiment 2

Experiment 2 used the same overall procedure but different statistical relationships between elements to probe how CPs, as well as different levels of ambiguity, influence behaviour. As in the case of natural language, disambiguating information can be precede or follow an ambiguous word. Hence, two sub-experiments were run, in which the order of the first two elements were interchanged so that the first element either provided high constraint (low entropy, Expt 2a) or low constraint (high entropy, Expt 2b) for predicting the final element, which was the same in both experiments. The experiments thus evaluated the impact of conjunctive probability learning and on the order of the more constraining versus less constraining elements on learning. If people integrate information in a manner analogous to CPs, it is expected that they would show slowdown according to ambiguity type, as illustrated by overlaps in sequence elements, over and beyond slowdown caused by visual complexity. We also expect differences as a result of informativeness of different elements. However, whether more informative elements will be faster to process due to the time to

hone in on a specific interpretation or slower due to the number of competing predictions is not clear.

Methods

Participants. A separate sample of 60 undergraduate participants who have not participated in other experiments were recruited for each of the experiments (2a: 15 male; mean age=19; 2b: 22 male; mean age=19).

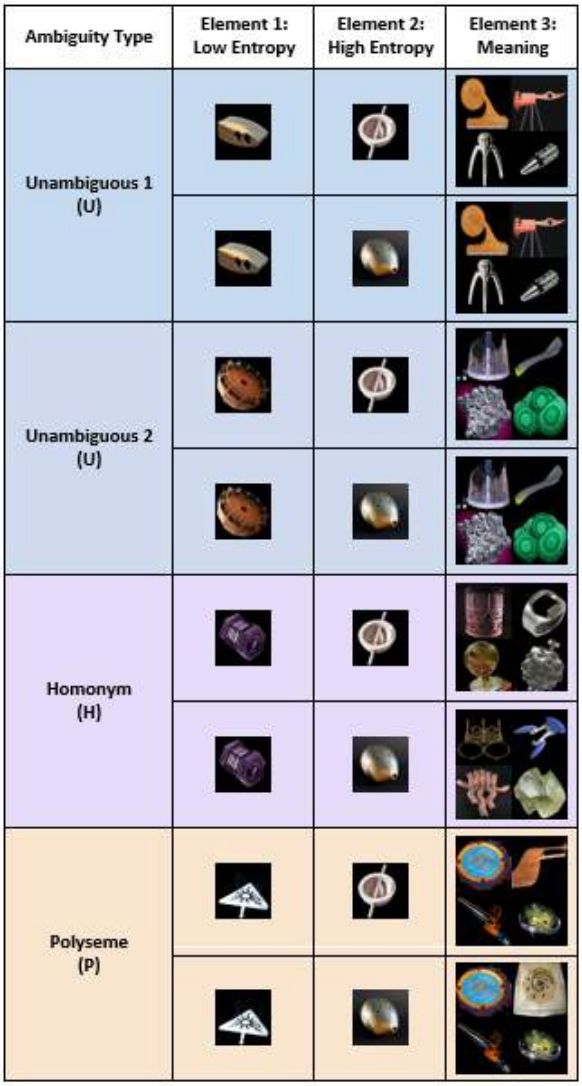


Figure 4: Example sequences depicting ambiguity types.

Materials. The same elements used in Experiment 1 were re-arranged to reflect different statistical relationships between the items, both in terms of how well each of the first two elements predicted the last element, and in terms of how distinct the last element is relative to its counterpart. These sequences were structured to represent three levels of ambiguity in how the low-entropy (word) representation merged with the high-entropy (context) representation. Across two contexts, Element 3 in an unambiguous sequence was identical, Element 3 in a polyseme sequence overlapped by 25% (one symbol), and Element 3 in a homonym sequence was distinct (see Figure 4). Single-symbol elements were used for the low- and high-entropy elements (words and contexts), whereas four-symbol elements were used to denote "meanings", so as to enable studying the effects of representational overlap. Symbols forming each element were randomized across participants.

Procedure The procedure was identical to that in Experiment 1, except the items were re-arranged to have the conjunctive probability structure outlined above and illustrated in Figure 4 for Experiment 2a (in Expt 2b, the position of the low-entropy and high-entropy items were swapped). The sequence completion task now contained eight coarse-grain and four fine-grain questions regarding Element 3 (meaning) instead of 12 questions of equal difficulty. In this experiment, fine-grained questions for both off-line tests refer to items where the choices given contain both options corresponding to the two possibly correct third elements, depending on context. Because the unambiguous sequences evoke the same meaning (Element 3) regardless of context (Expt 2a: Element 2; Expt 2b: Element 1), tests relating to the second item (first item in Expt 2b) were omitted since both context items were valid responses. This left six sequence familiarity items. Having more trials for one offline task type was due to our aims of efficiently extracting the learning of coarse- and fine-grained information.

Results

The analytical procedures were the mostly the same as Experiment 1, only now we collapsed across performance of the same ambiguity type and applied the linear model within each type.

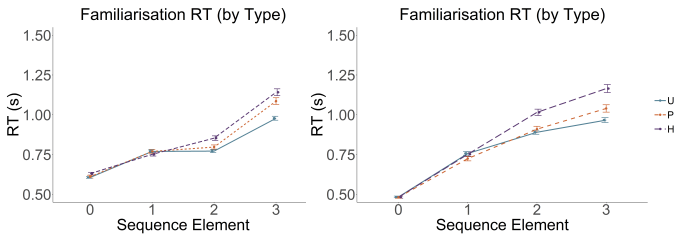


Figure 5: *Experiment 2a.* Element 0=Fixation; 1=Low Entropy; 2=High Entropy; 3=Meaning (left)
Experiment 2b. Element 0=Fixation; 1=High Entropy; 2=Low Entropy; 3=Meaning (right)
U = Unambiguous; P = Polyseme; H = Homonym

Familiarisation *Experiment 2a.* Figure 5 plots the results from familiarization for Experiment 2 and 2b. In Experiment 2a, RT for homonym sequences showed increase across all consecutive elements while polyseme sequences and unambiguous sequences showed slowdown only from Element 2 to Element 3. At the second position (high-entropy), the linear model for RT against ambiguity type showed homonym sequences to be sig-

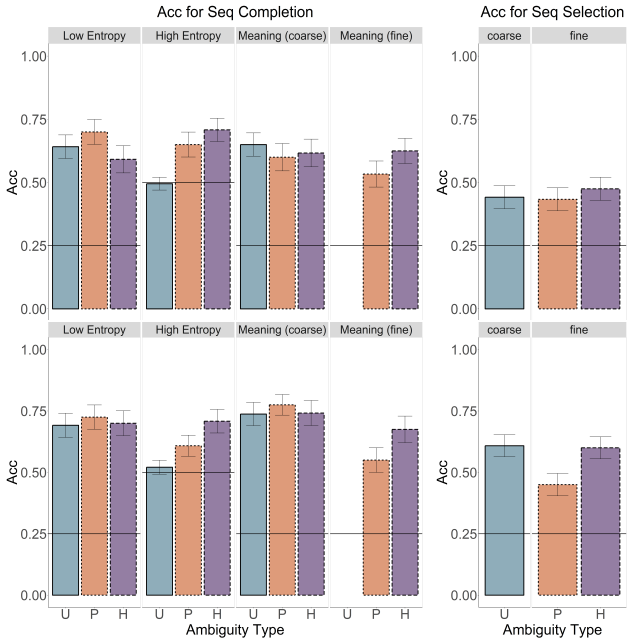


Figure 6: Experiment 2a (top) and 2b (bottom) offline tests. Horizontal lines denote chance.

nificantly different from unambiguous and polyseme sequences, which were comparable to each other. All three conditions were significantly different at meaning output (Element 3), with fastest performance for unambiguous items and slowest performance for homonym items.

Experiment 2b. In every ambiguity type, RT increased between consecutive elements. Similar to experiment 2a, there was no difference between ambiguity types in the first position, but at the second position (now low-entropy), divergence began where homonym sequences showed significant difference from unambiguous and polyseme sequences and by the third position (meaning output) all three ambiguity types were significantly different from each other.

Offline Test *Experiment 2a.* Figure 6 shows the offline task accuracy for experiment 2a and 2b. One sample t-test showed above-chance performance for all questions, indicating learning. A linear mixed effect model showed higher performance for polysemes than homonyms in the second position (low entropy) but no other differences between ambiguity types. Participants also performed above chance for the selection of familiar sequence. Performance did not differ by ambiguity.

Experiment 2b. All question types had above-chance accuracy. Fine-grained meaning (Element 3) in homonym was significantly more accurate than polyseme. There was comparable performance between ambiguity types for other elements. In the sequence selection task, performance were significantly above chance for all ambiguity types. Linear mixed model showed significant differences between polyseme and homonym sequences. Furthermore polyseme sequences were significantly more affected by the presence of context-inappropriate foils than homonym sequences.

Discussion

Experiment 2 showed an increased slowdown starting at the integration of contextual element according to the increased overlap in interpretations across contexts. In contrast to the unambiguous items and to the results obtained in Experiment 1, participants were slower to respond in the online task when learning CPs in ambiguous sequences. The amount of slowdown in the online task showed that these effects were modulated both by the amount of overlap in the meaning representations, and whether the more informative (lower entropy) item was presented earlier or later in the sequence. The lack of differentiation at Element 1 suggested that only with two elements was there enough information to integrate in order to predict the third element based on CPs. This is different from words in context-free tasks (Armstrong & Plaut, 2016) and tasks with contextual constraints for natural language (Klein & Murphy, 2001), where we see ambiguity effects for the ambiguous words themselves. This might be because participants are trying to integrate words both within and across trials in linguistic tasks, which would lead to task performance more similar to that observed for Elements 2 and 3 here (Klein & Murphy, 2001). Another possibility is that natural language tasks, as opposed to current artificial stimuli, engage in consistent, rapid, and automatic processing which results in detectable effects for the first element, whereas the slower and less natural processing of artificial stimuli do not elicit those effects.

We also investigated whether the slowdown for Element 3 was due to information integration per se, or was due to visual complexity. In a separate experiment not reported here due to space constraints, Experiment 2a was modified to have four symbols for all three sequence elements. We nevertheless still found significant slowdown between Elements 2 and 3 in polyseme and homonym sequences. This indicates that the slowdown observed in Experiment 2 was not solely attributable to differences in visual complexity for Element 3.

In contrast, the offline tests pointed to broadly similar performance regardless of the order in which the first two elements in the sequence were presented, with some detailed differences (e.g., changes in polyseme accuracy across Experiments 2a and 2b in sequence selection). This in turn suggests that the exact time-course of processing varies based on whether the more or less informative element is presented first, but the end result of processing is a relatively similar (although not identical) order-independent final representation.

Overall Comparisons

A striking difference between transitional probability and conjunctive probability sequences is in the long RT for Element 1. This may be explained by the ability of the first element to predict the following two elements in TP whereas both the first two elements need to be considered to predict the third in conjunctive probability.

In offline sequence completion, performance of high-entropy and low-entropy elements were similar for Experiments 2a and 2b in spite of their reversal in position within the sequence, supporting the hypothesis that performance on an element-level is tied to informativeness of the element. Generally, performances for offline tasks showed similar levels of accuracy across all experiments, suggesting that CPs do not pose much extra challenge in learning as compared to TPs.

General Discussion

Statistical learning is theorised to be a domain-general ability for detecting regularities across time and space, yet the bulk of extant research has focused on learning TPs between elements. This type of statistic, although clearly very useful for enabling some abilities like speech segmentation, is insufficient to understand other abilities, such as how words and contexts conjoin to evoke context-specific meanings in specific contexts (Swaab, Brown, & Hagoort, 2003). CPs, although certainly not capable of fully explain such behaviors, may be an alternative form of statistical computation that are critical for such information processing.

The present research merged a recent statistical learning paradigm, a self-paced learning task, with new statistical relationships among items that relate to CPs. Our results showed that CPs, like TPs, are learnable. By varying the amount of information content (entropy) in each position in the sequence, we were also able to ascertain that the order in which high- and low-entropy elements were presented in a sequence modulated online learning, but nevertheless resulted in similar patterns of performance in the offline test. Thus, the time-course of processing may differ based on the order in which information is presented (e.g., whether an ambiguous word like BAT precedes or follows a disambiguating context such as a discussion of SPORTS), but the end result of this processing is similar. Similarly, our manipulation of the relatedness between the "meaning" elements modulated performance in both the online and offline task, suggesting that the microstructure of each element can interact with the overall statistical regularities in the sequence. This suggests that multiple types of statistics among the individual elements of each sequence interact to determine overall performance.

This research represents an important proof of concept for how an alternative statistic than TPs can be learnt, and how such a structure could potentially interact with relatedness of interpretation to shape overall performance. In so doing, it opens up new possibilities for studying how simple statistical learning principles could interact with the rich structure of linguistic domains to explain at least some aspects of complex language behaviors such as context-sensitive meaning processing. As current models of statistical learning do not look at the problem of integrating constraints across elements, the current experiments can serve as a motivation to look at how this type of probability can be incorporated into such models. The ability to test even the domain-generality of some new language processes in a simple form is therefore very valuable. It also represents an important complement to existing methods using natural language, which have their own complexities in terms of controlling for confounding psycholinguistic properties. Having a new approach for developing convergent insights into statistical learning of CPs and other language abilities is therefore likely to be a powerful tool for advancing theory in related domains.

Acknowledgments

This work was funded by NSERC DG-502584 to BCA.

References

Armstrong, B. C., & Plaut, D. C. (2016). Disparate semantic ambiguity effects from semantic processing dynamics rather than qualitative task differences. *Language, Cognition and Neuroscience*.

Fiser, J., & Aslin, R. N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological science*, *12*(6), 499–504.

Gómez, R. (2002). Variability and detection of invariant structure. *Psychological Science*, *13*, 431-6.

Klein, D. E., & Murphy, G. L. (2001). The representation of polysemous words. *Journal of Memory and Language*, *45*(2), 259–282. doi: 10.1006/jmla.2001.2779

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language*. doi: 10.1006/jmla.1996.0032

Siegelman, N., Bogaerts, L., Kronenfeld, O., & Frost, R. (2018). Redefining Learning in Statistical Learning: What Does an Online Measure Reveal About the Assimilation of Visual Regularities? *Cognitive Science*, *42*, 692–727.

Swaab, T., Brown, C., & Hagoort, P. (2003). Understanding words in sentence contexts: The time course of ambiguity resolution. *Brain and Language*, *86*(2), 326–343.

What's in the Adaptive Toolbox and How Do People Choose From It? Rational Models of Strategy Selection in Risky Choice

Florian Mohnert
Max Planck Institute for Intelligent Systems, Tübingen, Germany
University of Amsterdam, Netherlands

Thorsten Pachur
Max Planck Institute for Human Development, Berlin, Germany

Falk Lieder
Max Planck Institute for Intelligent Systems, Tübingen, Germany
Bernstein Center for Computational Neuroscience, Tübingen, Germany

Abstract

Although process data indicate that people often rely on simplifying processes when choosing between risky options, current models of heuristics cannot predict people's choices very accurately. To address this apparent paradox, it has been proposed that people might adaptively choose from a toolbox of simple strategies. But which strategies are contained in this toolbox? And how do people decide when to use which decision strategy? Here, we develop a model according to which the decision maker selects a decision strategy for a given choice problem rationally from a toolbox of strategies; the content of the toolbox is estimated for each individual decision maker. Using cross-validation on an empirical data set, we find that this model of strategy selection from a personal adaptive toolbox predicts people's choices better than any single strategy (even when it is allowed to vary across participants) and better than previously proposed toolbox models. Our model comparisons show that both inferring the content of the toolbox and rational strategy selection are critical for accurately predicting people's risky choices. Furthermore, our analysis reveals considerable individual differences in the set of strategies people are equipped with and how they choose among them; these individual differences could partly explain why some people make better choices than others. These findings represent an important step towards a complete formalization of the notion that people select their cognitive strategies from a personal adaptive toolbox.

Keywords: decision making; bounded rationality; strategy selection; heuristics; computational modeling

Introduction

How do people make decisions under risk? This question is commonly studied by asking people to choose between gambles as in "Would you prefer a 20% chance of winning $1000 (Gamble A) or a 95% chance of of winning $200 (Gamble B)?" According to expected utility (EU) theory (von Neumann & Morgenstern, 1944), people should evaluate all possible outcomes that each available action might have and weight them by their respective probabilities. Empirical research, however, has demonstrated that human decision making systematically deviates from EU theory (e.g., Kahneman & Tversky, 1979). These deviations are commonly interpreted as an indication of human irrationality. Recent work, however, suggests that they could also reflect people's rational use of limited cognitive resources (Lieder & Griffiths, 2019; Griffiths, Lieder, & Goodman, 2015).

To date, the most prominent descriptive theory of risky choice is cumulative prospect theory (CPT; Tversky & Kahneman, 1992). CPT accounts for many violations of EU theory by postulating that people's decision mechanisms systematically distort the stated probabilities (i.e., overweighting rare and underweighting common events) and payoffs (diminishing sensitivity to additional increases in the outcome as the outcome gets larger, and an amplification of losses relative to gains). Interpreted as a cognitive process model, CPT predicts that information is processed exclusively within each option and that the information processing is identical across all problems. Process-tracing studies, however, show that people often compare options along individual attributes and that the processing varies across problems. These process data are instead consistent with processing policies of simple heuristics (Payne & Braunstein, 1978; Pachur, Hertwig, Gigerenzer, & Brandstätter, 2013) such as the lexicographic heuristic, that usually only looks at each gamble's most probable outcome while ignoring all other possible outcomes. Yet, model comparisons have found that assuming that people use a single heuristic across all problems, no single heuristic predicts risky choices nearly as well as CPT (Glöckner & Pachur, 2012).

One way to resolve this apparent paradox is to postulate that people are equipped with a toolbox of several, often heuristic, strategies and that they use different strategies on different trials. This raises the question of which strategies their toolbox is equipped with and how people select between them. Previous work on strategy selection has found that people adapt their strategy use to the structure of individual choice problem and the situation's requirements for speed versus accuracy (Payne, Bettman, & Johnson, 1988). The rational strategy-selection model by Lieder and Griffiths (2017) captures this adaptive flexibility as well as the variability and the learning-induced changes in people's strategy selection. It does not, however, specify the set of strategies from which people select among. To address this question, Scheibehenne, Rieskamp, and Wagenmakers (2013) developed a hierarchical Bayesian measurement model for inferring the contents of the cognitive toolbox. This model, however, assumes that peo-

ple's tendency to select a given strategy is not systematically related to the choice problem at hand and the requirements of the current situation. Strategy selection, however, has been shown to be sensitive to problem-specific features (Payne et al., 1988).

Here we develop an integrative model of risky choice with a *personal adaptive toolbox*. Our approach combines inferring the content of a person's cognitive toolbox with a rational model of strategy selection (Lieder & Griffiths, 2017). We validate this approach using a large empirical data set of risky choice data collected by Glöckner and Pachur (2012), testing it against single strategies, non-adaptive toolbox models, and CPT. Our model constitutes the first complete formalization of the notion that strategies are selected from a personal adaptive toolbox. It thereby enables more accurate inferences on people's cognitive toolbox than was previously possible, and we find that it predicts people's choices better than single strategies as well as other existing toolbox models.

The outline of this paper is as follows: We start by describing 11 extant (heuristic) strategies for risky choice, which might be contained in people's toolbox of decision strategies. We then introduce our computational model of the adaptive toolbox theory as well as several competitors. Next, we present a cross-validation method for inferring the set of strategies considered by an individual decision maker. We then evaluate our adaptive toolbox model against single strategies, non-adaptive toolbox models, and CPT. Finally, we apply our model to estimate the content of people's toolboxes—thereby elucidating why some people make better decisions than others. In closing, we discus the implications of our findings for the debate on human rationality as well as directions for future work.

Heuristics as Models of Risky Choice

A number of different strategies have been proposed as models of how people make decisions under risk. Following Glöckner and Pachur (2012), we consider the following ten heuristic strategies: the priority heuristic (PH), better-than-average (BTA), tallying (TALLY), probable (PROB), minimax (MINI), maximax (MAXI), lexicographic (LEX), equal-weight (EQW), least-likely (LL), and most-likely (ML). These heuristics cover a wide range of processing assumptions that differ in important aspects, such as whether they focus exclusively on the payoffs (BTA, TALLY, EQW, MINI, MAXI) or process both outcomes and probabilities (PH, PROB, LEX, LL, ML). For example, the minimax heuristic chooses the gamble with the highest minimum outcome and the least-likely heuristic identifies each gamble's worst outcome and then chooses the gamble with the lowest probability of the worst outcome.[1] Additionally, we include the weighted-additive strategy (WADD), which chooses the gamble with the highest expected payoff. Each of these strategies breaks ties between gambles by choosing randomly. We consider eleven simple models of risky choice according to which all people use one single strategy (either PH, BTA, TALLY, PROB, MINI, MAXI, LEX, EQW, LL, ML, or WADD) to make all their risky choices. Relaxing the assumption that all decision makers use the same strategy, we also tested a more flexible model (BEST), according to which each person might use a different strategy. That is, the BEST model has one parameter per person that encodes their strategy and has to be fitted to their choices.

[1]The equiprobable heuristic was not considered as it makes the same choice predictions as the equal-weight heuristic

Toolbox Models of Decision Making

According to the notion of an adaptive toolbox, each person is equipped with multiple strategies and employs them adaptively. In this section, we present three types of toolbox models that differ in whether the contents of the toolbox are inferred or assumed to be known and in their assumptions about how strategies are selected.

Strategy Selection Based on a Rational Cost-Benefit Analysis (RCBA)

Simulation studies by Payne et al. (1988) have shown that adaptively choosing between simple strategies can allow people to make many good decisions even when only little time is available. Assuming that decision makers are aware of the relevant properties of the choice problem (e.g., the magnitude of the possible outcomes), contextual factors (e.g., time pressure), and the speed and accuracy characteristics of the strategies in their toolbox, the adaptive decision maker (Payne et al., 1988) should choose strategies according to a rational cost-benefit analysis.

Building on the theory of rational metareasoning (Russell et al., 1991), the rational cost-benefit analysis (RCBA) model assumes that the expected payoff of making decision i using strategy h is integrated with the expected cost of the time $T(h,i)$ that it would take to do so. Together, they yield an estimate of the Value of Computation (VOC), defined as

$$\mathrm{VOC}(h,i) = \mathbb{E}\left[R(i,h(i))\right] - \delta \cdot T(h,i), \qquad (1)$$

where $R(i,h(i))$ is the payoff of decision $h(i)$ that strategy h would make in situation i, and $T(h,i)$ is the time it takes strategy h to make that decision. The balance between these two factors is determined by the relative opportunity cost δ. To model how long it takes to execute each strategy (i.e., the cost), we decompose the strategy into elementary information processes (EIPs) as introduced by Johnson and Payne (1985). Specifically, when a strategy is used to make a decision in a given choice problem, the number of EIPs required is recorded as $T(h,i)$. The RCBA model has two free parameters that can be estimated to accommodate individual differences: the set of available strategies H in the toolbox and the relative opportunity cost δ. For a given choice problem the strategy with the highest VOC in the toolbox is selected to make the choice.

Rational Strategy Selection Learning (RSSL)

The assumption of a full cost-benefit analysis for each strategy, as assumed by the RCBA, may be unfeasible for a boundedly rational mind. However, it might be possible to approximate the VOC. As one possible approach to do such an approximation, the rational strategy selection learning (RSSL) model assumes that the mind learns to predict each strategy's VOC based on the features of the choice problem at hand (Lieder & Griffiths, 2017). Specifically, the RSSL model assumes that people predict both the expected payoff and the expected time cost for each strategy (which are important for then determining the strategy's VOC) at a given problem based on a weighted sum of the features of the choice problem, such as the maximum probability or the range of outcomes; the weights for the estimation, in turn, are learned from the payoffs and decision times of past choices (with the latter is determined based on the number of EIPs the chosen strategy performed). The learning process is simulated using Bayesian linear regression and stochastic predictions are made by sampling from the posterior distribution.

The free parameters of the RSSL model are the number of samples drawn to predict the performance of each strategy, ζ, the set of strategies H, the opportunity cost δ and the amount of prior experience Λ (i.e., on how many choice problems the predictive models were trained on). For the latter parameter we assume that participants are equipped with some amount of prior experience in making choices using their strategies; hence we let the RSSL model learn from Λ randomly generated pairs of gambles prior to applying it to our participants' choices.

Toolbox Models Without Adaptive Strategy Selection

To assess how the assumption of rational strategy selection contributes to the predictive accuracy of the adaptive toolbox models introduced above, we evaluate them against simpler toolbox models that chooses strategies randomly for a given choice problem (rather than adaptively based on characteristics of the problems). In our first null model (NULL-TB1), every time a decision is made a strategy is selected from the set of 11 strategies introduced above. Our second null model (NULL-TB2) is like the first one except that the set of strategies it selects from is estimated on a participant-by-participant basis. Our third null model (NULL-TB3) extends the second one by allowing some strategies to be chosen more frequently than others. Specifically, following Scheibehenne et al. (2013), each strategy h is selected with probability θ_h, which is estimated from the participant's choices.

Cumulative Prospect Theory

According to CPT, the outcomes x_i of a gamble are transformed into subjective values according to the value function

$$v(x_i) = x_i^{\alpha} \text{ if } x_i \geq 0 \tag{2}$$

$$v(x_i) = -\lambda \cdot x_i^{\alpha} \text{ if } x_i < 0, \tag{3}$$

with an outcome sensitivity parameter $\alpha \in [0,2]$ that modulates the curvature of the value function and captures that people's sensitivity to changes in a payoff depend on its magnitude. Values of $\alpha < 1$ entails a concave value function with diminishing sensitivity to larger outcomes.

The probabilities p of the cumulative probability distribution function are transformed according to the probability weighting function

$$w(p) = \frac{p^{\gamma}}{(p^{\gamma} + [1-p^{\gamma}])^{1/\gamma}}, \tag{4}$$

whose shape is determined by the parameter $\gamma \in [0,2]$, which is defined separately for gains and losses. The shape of the probability weighting function reflects the degree of nonlinear distortion when the probabilities are mapped onto decision weights. Values of $\gamma < 1$ entail an inverse S-shaped probability weighting function, indicating a reduced sensitivity to probabilities in the middle range and a relative amplification of the sensitivity to differences among extreme probabilities. The overall valuation of a gamble is determined by multiplying each of the subjective values of the gamble's outcomes x_i by a decision weight π_i that follows from the weighted cumulative probabilities of obtaining an outcome at least as good as x_i if the outcome is positive, and at least as bad as x_i if the outcome is negative (for details see Tversky & Kahneman, 1992), and then summing the products:

$$V = \sum_i \pi_i \cdot v(x_i). \tag{5}$$

To derive the probability that gamble A is chosen over gamble B we apply the softmax choice rule to the gambles' subjective values V; this choice rule which has a choice sensitivity parameter ϕ (for details see Glöckner & Pachur, 2012).

Next, we describe the data set and our approach to evaluate the models introduced in the previous sections.

Data

We evaluated our models using data collected by Glöckner and Pachur (2012), who presented 64 participants with a set of 276 two-outcome gamble problems. The payoffs of the gambles ranged from -1000 to 1200 and the set of gambles consisted of pure gain (all payoffs> 0), pure loss (all payoffs < 0), and mixed (both positive and negative payoffs) gambles. The presentation of the gamble problems was distributed over two sessions that were one week apart (i.e., there are 138 choices from each session). For more information, see Glöckner and Pachur (2012).

Model Evaluation

We evaluated the predictive accuracy of each of the models using a simplified cross-validation method (Friedman, Hastie, & Tibshirani, 2001). Specifically, for each model a score was calculated indicating how often it correctly predicted the participants' choices on a held-out test set, that was not used to fit the model's parameters. The predictive accuracy for a given

participant was computed by averaging the model's performance in forward prediction (i.e., fitting the model on choice data from Session 1 (t_1) and testing it on data from Session 2 (t_2)) and backward prediction (i.e., fitting choices from t_2 and testing on data from t_1). To perform forward-prediction and backward-prediction, the data set was split into three subsets: a *training set*, a *validation set*, and a *test set*. The training set was used to fit the parameters (e.g., the subjective time cost δ) of a given sub-model (e.g., a strategy selection model with a particular set of strategies). The validation set was used to select among sub-models based on unbiased estimates of their generalization errors (e.g., to select the model's toolbox). The test set was used to obtain an unbiased estimate of the selected sub-model's generalization error that could be compared to the performance of the other models.

Model Fitting and Prediction

Given a set of choice problems and the corresponding choices made by an individual, we fitted each model's parameters by maximizing the proportion of gambles from the training set for which the model's predicted choice agreed with the participant's choice. The model parameters were estimated using participants' choices from t_1 and then used to predict choices from t_2—and vice versa. For forward-prediction, we used the 138 gamble problems and choices from t_1 (training set) and split the gamble problems and choices from t_2 into a validation set comprising 103 problems and a test set comprising 35 problems. Backward prediction was performed in the same way as forward prediction except with t_1 and t_2 reversed.

BEST model For the BEST model, according to which each participant uses a single strategy across all choice problems, we determined for each participant the strategy that achieved the highest accuracy (in terms of overlapping choices) on the training set choices and the validation set.

RCBA We estimated each participant's set of strategies H along with their subjective time cost δ using the following procedure: In the first step, H included only the strategy h_1 with the highest accuracy on the validation set. Next, we determined which strategy h_2, if added, would result in the set of two strategies with the highest predictive accuracy on the validation set. In doing so, we estimate δ by optimizing the accuracy of each candidate sub-model on the training set using Bayesian adaptive directed search (BADS) (Acerbi & Ma, 2017). We then proceeded to evaluate toolboxes that added a third strategy to the toolbox and re-estimated δ until toolboxes containing up to 11 strategies had been evaluated. That is, we estimated a set H_k of k strategies for each $1 \leq k \leq 11$ and estimated each participant's toolbox by the set $H_{k_{\max}}$ for which our model achieved the highest predictive accuracy on the validation set.

RSSL As described above, to define a toolbox of strategies, the RSSL model estimates each strategy's VOC based on previous experience with gamble problems. To simulate this experience, we first randomly generated pairs of two-outcome gambles; their payoffs and probabilities were samples from the uniform distributions $\text{Unif}([-1000, 1200])$ and $\text{Unif}([0, 1])$ respectively. The amount of prior experience (Λ) was set to 20 000 gamble problems. Each choice problem was represented by a feature vector comprising the maximum probability of each gamble, the payoffs associated with the maximum probability (i.e., the most likely outcome), the ranges of payoffs within each gamble, and the range of payoffs across both gambles. These features were then used to predict the strategy's accuracy and effort for the problem at hand. The number of predictions ζ sampled from the posterior was set to 3. The parameters H and δ for the rational cost-benefit analyses model (which the RSSL shares with the RCBA model) were estimated following the same iterative procedure as described for the RCBA model.

Null models The Null-TB1 model has no free parameters. For the models NULL-TB2 and NULL-TB3 we estimated the set of strategies H using the same procedure as for the RCBA model. For NULL-TB3, we estimated the proportion parameters $\theta_1 \ldots \theta_{|H|}$ for a toolbox H by solving the constrained optimization problem to maximize the expected accuracy of participants' choices.

Cumulative prospect theory CPT's parameters were fitted to minimize G^2 based on the observed choices in the respective session. To reduce the risk of being stuck in local minima, we first conducted a grid search to identify the 20 best-fitting combinations of parameter values; these combinations were then used as starting values for subsequent optimization using the simplex method. For prediction, we derived deterministic choice predictions from CPT.

Results

Predictive Accuracy

Figure 1 shows how accurately single strategies, simple toolbox models, adaptive toolbox models, and CPT predicted the risky choices in the test set. Out of the eleven single strategies, WADD and minimax predicted people's choices best, with 65.8% and 61.3% accuracy, respectively. Relaxing the assumption that all participants use the same strategy and instead inferring a potentially different strategy for each participant (i.e., the BEST model) increased predictive accuracy to 69.4%, which is significantly higher than that of the best-performing single strategy WADD ($t(34) = 2.90$, $p = .004$). The simplest toolbox model Null-TB, which chooses randomly among all the strategies, was less predictive of people's choices than the BEST model (57.5% vs. 69.4%, $t(34) = -8.79$, $p < .001$). Its predictive accuracy increased, however, when we allowed the content of the toolbox to be estimated for each participant separately (65.7% vs. 57.5%, $t(34) = 6.66$, $p < .001$). Additionally estimating the relative frequency with which each strategy is selected

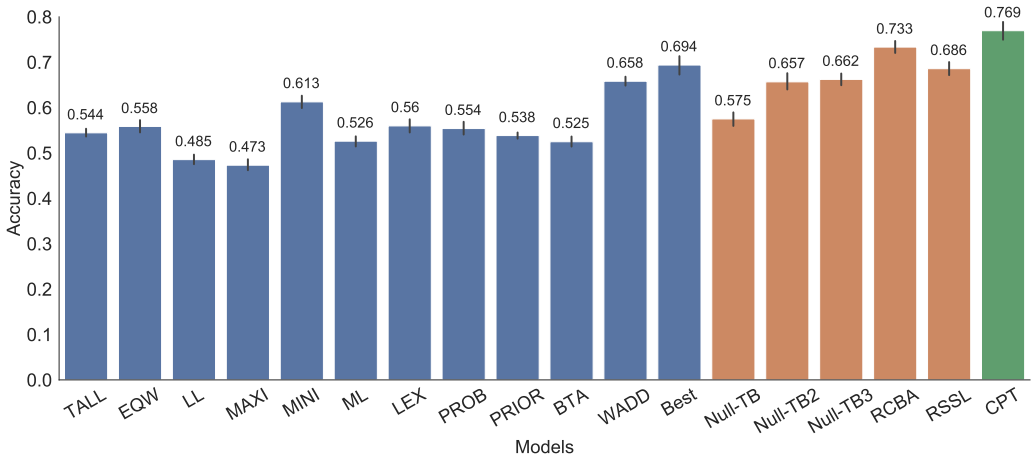


Figure 1: Comparison of how accurately each model predicted people's choices in the validation set. The single-strategy models are shown in blue, toolbox models in orange, and CPT in green.

independently of the problem (Null-TB3) further improved the toolbox model's predictive accuracy to 66.2% vs. 65.7% ($t(34) = 0.46$, $p = .64$). While the benefit of this choice problem-unspecific strategy-selection mechanism was rather small; adding an adaptive, problems-specific selection mechanism to the RCBA model drastically improved the accuracy of the toolbox approach to 73.3% vs. 65.7% ($t(34) = 6.50$, $p < .001$). The RSSL model, which approximates the rational cost-benefit analysis of strategy selection using the features of the choice problems as predictive cues, did not perform as well as the RCBA model (68.6%).

Critically, the RCBA model predicted people's choices better than the best-performing single strategy WADD ($t(34) = 8.80$, $p < .001$) and the BEST model ($t(69) = 3.02$, $p = .003$). This suggests that decision makers indeed adaptively choose from a personal toolbox of strategies when solving a sequence of different choice problems.

The RCBA model also achieved higher predictive accuracy than all null models, NULL-TB ($t(34) = 15.51$, $p < .001$), NULL-TB2 ($t(34) = 6.51$, $p < .001$) and NULL-TB3 ($t(34) = 7.56$, $p < .001$). These results corroborate the usefulness of combining inference about the content of the toolboxes with a model of how people's strategy choices are informed by the specific requirements of each individual decision. This finding strongly supports adaptive toolbox theories of human decision-making (Gigerenzer & Selten, 2002) in general and the idea of an adaptive *personal* toolbox in particular. Despite the substantial improvement in predictive accuracy we achieved by combining inference on the toolbox with adaptive strategy selection, the resulting RCBA model predicted people's choices not as well as CPT (73.3% vs. 76.9%, $t(34) = 3.03$, $p = .003$). While the RCBA model may thus not capture all aspects of how people make decisions, it being a process model still affords many practical advantages for understanding people's choices that cannot be obtained by modeling the choices with CPT (but see Pachur, Suter, & Hertwig, 2017). For example, the estimated contents of the toolbox and estimated parameters of the strategy selection mechanism provide a window onto the cognitive mechanisms underlying risky choice and how they vary across individuals.

Comparing Predicted and Actual Performance

Next, we compared the models and people in terms of their performance of their risky choices. Performance here is measured as the average expected value (EV) of the chosen gambles. WADD achieved an EV of 149.02, which therefore represents the upper bound on how well one could perform in this task. The RCBA model predicted a higher performance than what was actually observed for people's choices (143.41 vs. 130.83, $t(69) = 5.57$, $p < .001$). CPT, on the other hand, predicted a lower performance than people actually achieved (113.1 vs. 130.8, $t(69) = 5.68$, $p < .001$). The performance of the RSSL model and the toolbox model Null-TB3 fell in between, with 124.83 EV and 126.81 EV, respectively, and were closer to people's actual performance. These findings suggest that while people may not choose strategies optimally, they may still be substantially more resource-rational than CPT would make us believe.

Which Strategies Are In The Adaptive Toolbox?

Given our finding that the best-suited model to predict people's choices is the RCBA model, we next analyze its estimated parameters H and δ. Figure 3 shows how many strategies were in the estimated toolboxes of all participants. 28.91% of all toolboxes included 4 strategies, and 60.15% of all toolboxes included between 3 and 5 strategies. The average toolbox size was 4.3.

Next, we counted how often each of the eleven strategies was included in the estimated toolboxes (see Figure 2). Interestingly, with 79.68% and 71.09% WADD and minimax are the most frequently included strategies in the toolboxes. These two strategies also predicted people's choices most ac-

curately (see Figure 1).

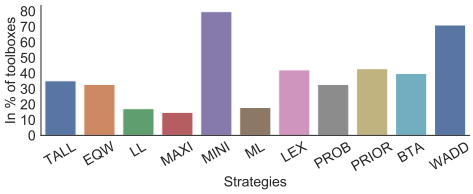


Figure 2: Percentage of cases each strategy was included in the toolboxes estimated by the RCBA model.

Minimax is especially useful as a risk-minimizing strategy when the probabilities of the possible outcomes are similar. The high inclusion rate of WADD suggests that at least when there are only two choices with only two possible outcomes, maximizing expected value is a viable and cognitively feasible strategy.

Furthermore, our results suggest that individual differences in decision quality might be due to the fact that different people are equipped with different toolboxes. For example, participants whose inferred toolbox included WADD performed better (144.59 EV) than participants whose inferred toolboxes did not include WADD (140.52 EV). Conversely, participants whose toolboxes were estimated to contain minimax achieved a lower performance than participants who did not use minimax (140.99 EV vs. 152.91 EV). These observations suggest that inferences obtained with the RCBA model can shed light on why and how people make the choices that they make. Additionally, our analysis identified another source of individual differences in decision performance: people's subjective cost of their time and effort. Specifically, our parameter estimates revealed a negative rank correlation between performance (in terms of EV) and the subjective opportunity cost δ (Spearman's $\rho(62) = -0.58$, $p < .001$), reflecting that higher opportunity costs favour less resource-intensive strategies even when they lead to less accurate decisions.

Finally, we found that the estimated size and content of the toolbox and the objective opportunity cost together explained 27% of the variance in individual differences in performance ($R^2 = 0.27$, $F(21,106) = 2.25$, $p < .001$).

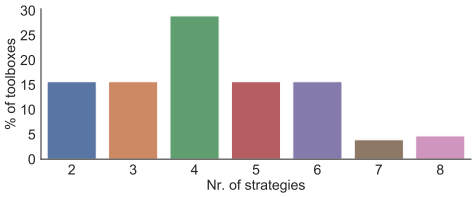


Figure 3: Toolbox sizes estimated by the RCBA model.

Discussion

We presented a model that represents the first complete formalization of the adaptive toolbox metaphor of human judgment and decision making (Gigerenzer & Selten, 2002). Our personal adaptive toolbox model predicted people's risky choices better than single strategies, non-adaptive toolbox models, or adaptive toolbox models that assume that all decision makers have the same strategies in their toolbox. Furthermore, the mechanistic nature of our model makes it possible to draw inferences about the cognitive architecture and processes underlying people's decisions. Furthermore, unlike CPT, our rational model of strategy selection can be applied to a wider range of domains, including inferential problems, such as those used by Gigerenzer and Goldstein (1996), by adapting the set of strategies (which can be deterministic or stochastic) and the reward function.

The success of the model that chooses strategies according to a rational cost-benefit analysis provides additional support for the view that people make rational use of their limited cognitive resources (Griffiths et al., 2015; Lieder & Griffiths, 2019). Our model is an important step towards reverse-engineering the mechanisms underlying the adaptive flexibility of human decision-making and individual differences in risky choice. But the mechanisms by which people efficiently approximate its rational cost-benefit analysis and the resulting suboptimalities need be investigated further before any definite conclusions can be drawn.

Future work will revisit the comparison with CPT using more complex decision problems, including problems with many alternatives and many possible payoffs (Payne et al., 1988), where people's selective processing of only a small subset of the available information might have a notable impact on their choices. We will also compare our models to other psychologically plausible models of risky choice including the utility-weighted sampling model (Lieder, Griffiths, & Hsu, 2018) and decision-field theory (Busemeyer & Townsend, 1993; Rieskamp, 2008; Bhatia, 2014) and apply likelihood-based model selection methods.

Future work will refine the strategy selection learning model with more realistic assumptions about decision makers' prior experience and the features they use to predict the performance of their strategies. In particular, future refinements of this model might take into account that people's strategy choices are informed by them learning from how well each strategy worked when they previously used it in the real world. This prior experience could be simulated by training the RSSL model on choice problems that are more like those that people encounter in everyday life (e.g., in having more possible outcomes and larger differences between the alternatives' expected values). The eleven strategies considered here are unlikely to cover all the decision mechanisms people use. Hence, we will consider additional strategies derived from resource-rational analysis (Lieder & Griffiths, 2019; Lieder, Krueger, & Griffiths, 2017; Gul, Krueger, Callaway, Griffiths, & Lieder, 2018) and process tracing.

References

Acerbi, L., & Ma, W. (2017). Practical Bayesian Optimization for model fitting with Bayesian adaptive direct search. In I. Guyon et al. (Eds.), *Advances in neural information processing systems 30* (pp. 1836–1846). Curran Associates, Inc.

Bhatia, S. (2014). Sequential sampling and paradoxes of risky choice. *Psychonomic Bulletin & Review*, *21*(5), 1095–1111.

Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: a dynamic-cognitive approach to decision making in an uncertain environment. *Psychological review*, *100*(3), 432.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1) (No. 10). Springer series in statistics New York, NY, USA:.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: models of bounded rationality. *Psychological review*, *103*(4), 650.

Gigerenzer, G., & Selten, R. (2002). *Bounded rationality: The adaptive toolbox*. Cambridge, MA: MIT Press.

Glöckner, A., & Pachur, T. (2012). Cognitive models of risky choice: Parameter stability and predictive accuracy of prospect theory. *Cognition*, *123*(1), 21–32.

Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in Cognitive Science*, *7*(2), 217–229. doi: 10.1111/tops.12142

Gul, S., Krueger, P. M., Callaway, F., Griffiths, T. L., & Lieder, F. (2018, September). Discovering rational heuristics for risky choice. In *The 14th biannual conference of the German Society for Cognitive Science, GK.*

Johnson, E. J., & Payne, J. W. (1985). Effort and accuracy in choice. *Management Science*, *31*(4), 395–414.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263–291. doi: 10.2307/1914185

Lieder, F., & Griffiths, T. L. (2017). Strategy selection as rational metareasoning. *Psychological Review*, *124*(6), 762-794.

Lieder, F., & Griffiths, T. L. (2019). Resource-rational analysis: Understanding human cognition as the optimal use of limited ressources. *Behavioral and Brain Sciences*.

Lieder, F., Griffiths, T. L., & Hsu, M. (2018). Overrepresentation of extreme events in decision making reflects rational use of cognitive resources. *Psychological review*, *125*(1), 1.

Lieder, F., Krueger, P. M., & Griffiths, T. L. (2017). An automatic method for discovering rational heuristics for risky choice. In *Proceedings of the 39th annual meeting of the cognitive science society.* Austin, TX: Cognitive Science Society.

Pachur, T., Hertwig, R., Gigerenzer, G., & Brandstätter, E. (2013). Testing process predictions of models of risky choice: A quantitative model comparison approach. *Frontiers in Psychology*, *4*, 646.

Pachur, T., Suter, R. S., & Hertwig, R. (2017). How the twain can meet: Prospect theory and models of heuristics in risky choice. *Cognitive Psychology*, *93*, 44–73.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1988). Adaptive strategy selection in decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(3), 534.

Payne, J. W., & Braunstein, M. L. (1978). Risky choice: An examination of information acquisition behavior. *Memory & Cognition*, *6*(5), 554–561.

Rieskamp, J. (2008). The probabilistic nature of preferential choice. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(6), 1446.

Russell, S., Wefald, E., Karnaugh, M., Karp, R., Mcallester, D., Subramanian, D., & Wellman, M. (1991). Principles of metareasoning. In *Artificial intelligence* (pp. 400–411). Morgan Kaufmann.

Scheibehenne, B., Rieskamp, J., & Wagenmakers, E.-J. (2013). Testing adaptive toolbox models: A Bayesian hierarchical approach. *Psychological Review*, *120*(1), 39-64.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*(4), 297–323.

von Neumann, J., & Morgenstern, O. (1944). *The theory of games and economic behavior*. Princeton, NJ: Princeton University Press.

Reward Function Complexity and Goals in Exploration-Exploitation Tasks

Brian Montambault (brian.montambault@tufts.edu)
Department of Computer Science, Tufts University

Christopher Lucas
School of Informatics, University of Edinburgh

Abstract

People are often faced with choices where there is a conflict between seeking reward and gathering information. In many of these cases there exists a functional relationship between the features associated with actions and their corresponding rewards. Accounts of how people make decisions in these circumstances have not considered how peoples' strategies depend on the complexity of this function, as well as the person's goal. In a sequential decision making task we found that people chose between a number of different exploration strategies, but that strategy selection did not necessarily align with goal or account for function complexity.

Keywords: Decision Making; Exploration-Exploitation; Contextual Multi-Armed Bandits

Introduction

In many of the decisions that people make in life there is a conflict between choices that are likely to have good results and choices where the result is more uncertain, but could possibly lead to a better outcome than the known option. For example, one might choose to eat at a familiar restaurant that is known to be good, or a new restaurant where the quality could be either better or worse. This trade-off is known as the explore-exploit dilemma. A structurally similar problem, with a slightly different goal is identifying the best candidate from a set of possible choices within a fixed time frame. For example, someone planning a party might wish to sample several possible caterers in order to find who will provide the best meal. Unlike the dilemma of choosing a restaurant for dinner, it is only important that the best option is found; the quality of any single meal is unimportant.

A common task for studying how people navigate explore-exploit dilemmas is the multi-armed bandit (MAB) task (Steyvers & Wagenmakers, 2009; Lee, Zhang, Munro, & Steyvers, 2011), where a decision-maker chooses between discrete actions, each with an unknown reward distribution, in order to maximize total reward over the course of several trials. While these tasks provide a simple environment for studying decision-making, real world tasks often contain additional contextual information about how rewarding an option might be. For example, we might have the option between two new restaurants, where the first has a menu with similar items to a past favorite, and the second has a menu that is full of new options. If we want to maximize the chance we will be satisfied, it would be prudent to pick the first. If we want to learn something new, we should choose the second. More formally, we can describe each option, a_i with the set of features s_i, with a_i yielding the reward $r_i = f(a_i, s_i)$, where f is a reward function mapping actions and features (or contexts) to rewards. We can call this a *contextual* multi-armed bandit (CMAB) (Li, Chu, Langford, & Schapire, 2010). In this setting, successful learners must make inferences about what this function might be – especially if there are many actions to choose from.

How people learn mappings between inputs and outputs, or *function learning*, has been widely studied (DeLosh, Busemeyer, & McDaniel, 1997). Recently, Gaussian process regression (GPR) has been presented as a model of function learning (Lucas, Griffiths, Williams, & Kalish, 2015). In addition to being a flexible non-parametric model capable of representing a wide range of functions, GPR is distinct from other accounts in that it directly allows for the representation of uncertainty in outputs. For CMAB tasks, this lays bare the trade-off between exploration and exploitation: An exploration-oriented agent can target options where uncertainty is greatest, an exploitation-oriented agent can target options with the highest expected reward, and it is possible to strike a balance between the two extremes. Bayesian optimization (Snoek, Larochelle, & Adams, 2012) is a flexible framework for transforming predictions from GPR models into actions. Several algorithms have been proposed for handling these tasks (Snoek et al., 2012) and have shown to both perform well (Srinivas, Krause, Kakade, & Seeger, 2010) and describe human behavior (Schulz, Konstantinidis, & Speekenbrink, 2018) in CMAB tasks. However, these accounts do not consider how one's strategy might be contingent on their ability to learn the reward function. This ignores a prominent result from the function learning literature: that some families of functions (e.g. linear) are easier to learn than others (e.g. periodic) (Kalish, Lewandowsky, & Kruschke, 2004).

While most work on MABs and CMABs study tasks where the goal is to maximize cumulative reward, there are circumstances where a decision-maker might instead be interested in finding the best action (Audibert & Bubeck, 2010). In the case of CMABs, this can be understood in terms of optimization, where the goal is to find some configuration of features (contexts) that maximize an objective function (reward). Bayesian optimization has shown to be of great practical use in these cases, in particular when the objective

function is expensive to evaluate as in the optimization of machine learning algorithm hyperparameters (Snoek et al., 2012). Bayesian optimization typically selects actions that have both a high expected reward and are highly uncertain, as in upper confidence bound (Auer, Cesa-Bianchi, & Fischer, 2002) and expected improvement (Mockus, 1974) algorithms. While these are reasonable strategies when the goal is to earn large rewards on each trial while still exploring new actions, they are ill-suited for optimization, where rewards on each trial are not important. Algorithms based instead on reducing uncertainty about the maximum of the reward function have been recently introduced (Hennig & Schuler, 2012; Wang & Jegelka, 2017) and appear better suited to this goal. Other recent work has examined the idea that people adapt their strategies to the tasks they face, accounting for both the expected performance of a strategy and the cost (e.g., in time) of executing it (Lieder, Helen, & Griffiths, 2017). If one hypothesis is that people adapt their strategies to the task at hand, and distinguish between optimization problems and ongoing trade-offs between exploration and exploitation, another is that people use a "one size fits all" strategy that supports multiple goal types reasonably well, as suggested by some past results, e.g., (Borji & Itti, 2013; Wu, Schulz, Speekenbrink, Nelson, & Meder, 2018).

For both reward maximization and optimization problems, good strategies must seek out information or reduce uncertainty. They can do this in an explicit or *directed* way, or achieve it implicitly by adopting a *stochastic* policy. In *directed* exploration one seeks actions that are most informative about the underlying reward distributions. One popular class of algorithms choose actions with high *upper confidence bounds* (UCB) (Auer et al., 2002), which typically include a free parameter β that controls the width of the confidence bound, directly controlling the preference for exploration over exploitation. In the case of UCB, exploration is directed by uncertainty about individual actions, where those with high uncertainty about their reward are more appealing than those with low uncertainty. In contrast, entropy-based strategies (Hennig & Schuler, 2012; Wang & Jegelka, 2017) are directed by uncertainty about global properties of the function – in particular uncertainty about the function maximum. In *stochastic* exploration, one seeks to explore the space of actions by applying some level of randomness to one's actions. While these methods are implicitly sensitive to reward uncertainty, they do not explicitly minimize it. Thompson sampling (Thompson, 1933) applies randomness to actions by first sampling a reward structure given previous observations, and then choosing the best action given the sampled rewards. Another method of random exploration is to choose actions with probabilities based on the softmax function

$$p(a_t = k) = \frac{\exp[m_t(k)/\tau]}{\sum_{k' \in A} \exp[m_t(k')/\tau]}$$

where $m_t(k)$ is the expected reward of arm k on trial t, and τ controls the level of randomness of actions, with all actions being equally likely as $\tau \to \infty$ and one deterministically choosing the action with the highest expected reward as $\tau \to 0$. While evidence for both directed and random exploration has been found in human behavior (Gershman, 2018), it has yet to be determined whether the criteria for directed exploration is dependent on the goal of the task or is exclusively based on uncertainty about individual actions, and under what conditions random exploration might be preferred over directed exploration.

Many of the real world explore-exploit dilemmas faced by people require learning a mapping between contexts and rewards, making CMABs an attractive environment for studying this phenomenon. While Bayesian optimization and other GPR-based approaches have been widely demonstrated to be a good model of human behavior in these tasks, there has been little research investigating how these frameworks capture different behaviors across distinct environments. While there has been work demonstrating that people are capable of learning functions and applying that representation to their decisions (Schulz et al., 2018), it is unclear how people's strategies might change when faced with functions of varying complexities, though some have varied function complexity by comparing smooth and rough non-parametric functions (Wu et al., 2018), and compared linear to quadratic reward functions (Stojic, 2016). While Bayesian optimization has been shown to describe human behavior well both when the goal is to maximize cumulative reward and when the goal is to find the best arm, it is unclear whether people choose a strategy to match their goal or use a more general strategy regardless of goal. Our contribution is to demonstrate how these factors influence people's strategies. We introduce a model based on Bayesian optimization that is capable of representing a rich set of behaviors revealed in prior work, and how different reward function complexities and goals might result in different parameterizations describing behavior.

Methods

Experiment. We designed a CMAB task in which participants were allowed to click one of several "actions" represented by a set of vertical bars situated along the x-axis of a plot. Upon clicking a bar, the reward of the associated action was revealed to the participant by displaying the height of the bar. Actions (bars) were related to rewards by their position on the x-axis: the i^{th} bar from left to right, a_i, was associated to the reward r_i by the function $r_i = f(a_i, i)$. We tested behavior on CMABs with three different reward functions of varying complexity:

$$f_{linear}(a_i, i) = i$$
$$f_{quadratic}(a_i, i) = -(i - 55)^2$$
$$f_{sinc}(a_i, i) = \frac{\sin(i/2 - 30.000001)}{i/2 - 30.000001}$$

Reward functions were scaled to fall within minimum and maximums drawn from uniform distributions, $\mathbf{U}(0,100)$ and $\mathbf{U}(400,500)$ respectively. Participants were shown 10 reward sample functions before they began the task. The quadratic and sinc samples were generated by uniformly sampling the location of the maximum, the function minimum, and function maximum ($\mathbf{U}(1,80)$, $\mathbf{U}(0,100)$, and $\mathbf{U}(400,500)$ respectively). Linear functions were generated by samples of the intercept and slope drawn from uniform distributions $\mathbf{U}(0,250)$ and $\mathbf{U}(0,6.25)$.

Participants were given one of two possible goals: In the maximum-finding condition, participants were asked to find the bar associated with the maximum possible reward. Participants final score in this condition was equal to the maximum reward uncovered across all trials. In the score-maximization condition, participants were asked to maximize their cumulative scores across all trials.

Procedure and participants. Participants (n=69, mean age=33.0 years) were recruited using Amazon's Mechanical Turk service. They were randomly assigned one of 6 (3 reward functions $\times$ 2 goals) conditions. They were first shown 10 different sets of 80 bars with their heights already revealed. Depending on a participant's function condition, the heights of the bars in each set was determined by either linear, quadratic, or sinc functions. Participants were then shown a new set of 80 bars, each 500 pixels tall and gray in color, and instructed to either find the bar with the largest height (find-max) or to maximize the cumulative heights of bars clicked across all trials (max-score) for a new set of bars. Participants were invited to click on any of the 80 bars over 25 trials. When a gray bar was clicked its color changed to black and its height was adjusted to match its corresponding reward (between 0 and 500 pixels). After each trial the reward associated with the chosen bar was used to update the participants goal-specic reward, displayed on the screen alongside the bars. On each trial, any bars that were clicked on previous trials remain black and the height in pixels of their associated rewards. To incentivize performance participants were given a bonus up to \$0.75 proportional to the total number of points they earned.

Model

Our goal was to uncover strategies used in an CMAB task with different reward function complexities and goals. We take inspiration from Bayesian optimization, taking action probabilities to be a function of a GPR predictions of the reward function. Like previous accounts, we characterize exploration as a mixture of directed and random behavior. However, While previous accounts have assumed that directed exploration only uses uncertainty about each action, we extend this framework to include uncertainty about the function maximum.

In GPR a kernel function is used to encode prior beliefs about a function. We use the radial basis function (RBF) kernel:

$$k(x,x') = \sigma^2_{var}\exp(-\frac{(x-x')^2}{2l^2})$$

where l determines the smoothness of the function, or how quickly the similarity of two points falls off as they become more distant, and σ^2_{var} determines the average distance of the function from its mean. This kernel function is well suited to flexibly modelling function learning, as it is capable of learning any smooth function. For each reward function condition a set of 10 functions from the same family that were shown to participants prior to the CMAB task were used to fit the hyperparameters of the kernel function by maximizing the log marginal likelihood of the sample functions (Rasmussen & Williams, 2005). Fitting kernel hyperparameters in this way for each function allows us to model participants' expectations about the smoothness of the reward function, given the observed set of sample functions.

To estimate each participant's trial-by-trial predictions we compute the posterior mean and variance of the reward function at each action:

$$m_t(a) = \mathbf{k}_t(a)^\top(\mathbf{K}_t + \sigma^2_{noise}\mathbf{I})^{-1}r_t$$
$$v_t(a) = k(a,a) - \mathbf{k}_t(a)^\top(\mathbf{K}_t + \sigma^2_{noise}\mathbf{I})^{-1}\mathbf{k}_t(a)$$

where $k(a,a')$ is the covariance of two actions given the hyperparameters learned from a participant's training functions, $\mathbf{k}_t(a)$ is a vector of covariances for the action a and all previous observed actions, and $\mathbf{K}_t$ is the covariance matrix of all previously observed actions. σ^2_{noise} is the noise observed in the data. The reward functions in our task are deterministic, so we set this to a very small but non-zero number 10^{-4} to avoid numerical instability. We encode exploration directed by uncertainty about the function maximum by approximating the mutual information between the reward r revealed by action a and the highest possible reward r^*, $I(\{a,r\};r^*)$, the approximation used in max-value entropy search (Wang & Jegelka, 2017). We define the *utility* of each action on trial t to be

$$u(a,\beta,\lambda) = m_t(a) + \beta v_t(a) + \lambda I(\{a,r\};r^*)$$

and the probability of each action was defined using the softmax function

$$p(a|\beta,\lambda,\tau) = \frac{\exp[u(a,\beta,\lambda)/\tau]}{\sum_{a'\in A}\exp[u(a',\beta,\lambda)/\tau]}$$

We use an infinite groups model (Navarro, Griffiths, Steyvers, & Lee, 2006) to uncover common strategies across participants. Using the probability of actions, if the i^{th} participant belongs to group z,

$$p(a^i_T|g_i = z) = \prod_t^{T-1} p(a^i_{t+1}|a^i_t, r^i_t, \beta_z, \lambda_z, \tau_z)$$

where a^i_T is the set of all actions performed by the i^{th} participant. Groups were assigned priors according to a

stick-breaking procedure (Ishwaran & James, 2001). Under this prior we imagine a stick of length 1 that we break in two, keeping the length of the first stick to be the prior probability of our first group. We can then break the remaining piece in two again, with one of its pieces representing the prior probability of our second group. This process can be extended to represent a countably infinite number of groups, with the sum of their prior probabilities guaranteed to sum to 1. The stick-breaking prior has one parameter, α, that determines the dispersion among groups, with a higher α resulting in likelihoods being spread across a greater number of groups. We place a Gamma(a,b) prior over α, setting $a = b = 10^{-10}$ to represent our ignorance of the true number of groups in the data. We set Gamma priors with $a = b = 0.1$ over β, λ, and τ to represent equal preferences for each type of exploration.

Results

We used the python package PyMC3 (Salvatier, Wiecki, & Fonnesbeck, 2016) to perform inference. MCMC sampling was performed using the NUTS sampler, with 4 chains of 1000 samples each.

To inspect the range of strategies used by participants we assigned each participant to their most likely group, maximizing $p(g|a_T^i)$ for the i-th participant. Nine groups were assigned at least one participant. We summarize the behavior of each of these groups by their parameter means in Table 1. The largest four groups were assigned 48 out of 69 participants. The largest group has a much larger average τ than other groups, indicating that participants in this group heavily utilized random exploration. The second largest group had a larger average β and λ and smaller average τ, indicating that participants in this group utilize directed in addition to random exploration, using both uncertainty about each action and uncertainty about the reward function maximum. The third largest group also had a relatively large average β and λ, but a smaller average τ than the previous group. This indicates that participants in this group also used both forms of directed exploration, but relied much less on random exploration. The fourth largest group had relatively low average values for all three parameters, indicating that participants in this group did comparatively little exploring, instead choosing actions based on their expected reward. We refer to these groups as *stochastic*, *mixed*, *directed*, and *greedy* respectively.

To better understand how behaviors differed between groups, we measure the distance between participants' actions and both their previous action and their reward function maximum across trials (Figure 1). First, we plot the distribution of the distances between a participant's action and their previous action. Participants across all four of the top groups made a large proportion of their actions in close proximity to their previous action. This proportion was largest for the random group, followed by mixed, directed, and greedy. As we might expect, participants in the random group demonstrated more aggressive exploration with respect to their previous action, while those in the mixed and directed groups were more reserved. In contrast, participants in the greedy group rarely deviated far from their previous action. Next, we plot the median distance from the reward function maximum by trial for each group. For the random and mixed groups, the median distance from the reward function maximum stays level across trials, indicating that participants in these groups favor exploration over converging on the best action. For the directed and greedy groups, the median distance decreases towards zero with the number of trials. While the distance continues to decrease and eventually flattens out for those in the greedy group, the distance for those in the directed group increases after a number of trials, indicating that participants were willing to continue exploring even after the region containing the reward function maximum was located.

If participants were selecting there strategy based on their goal, we would expect the actions of participants in the max-score condition to be best predicted by a strategy that minimizes balances exploration and exploitation, and those of participants in the find-max condition to be best predicted by a strategy that minimizes uncertainty about the reward function maximum. While none of the groups show a preference for the source of uncertainty used to direct exploration (either about rewards of individual actions or the function maximum), these groups do differ in their preference for random and directed exploration. To investigate how reward function complexity and goal determine how people choose between these strategies we compare how well each strategy predicts the actions of participants grouped by experimental condition (Table 2). Participants selecting a strategy in the max-score goal condition are expected to choose a strategy that favors actions with high rewards, while those in the find-max condition are expected to choose a strategy that puts more of an emphasis on exploration. Our results contradict this assumption, with those in the max-score condition best described by the directed strategy and actions of those in the find-max condition best described by the greedy strategy. With function learning being increasingly difficult as function complexity increases, participants in less complex reward function conditions are expected to use this learning to engage in more directed exploration, while those in the more complex reward function conditions are expected to rely more heavily on random exploration. Our results show some evidence for this, as actions of participants in the linear condition were best explained by the directed strategy, while those in the quadratic condition were best explained by the mixed strategy. However, our results also show that the actions of those in the sinc condition were also best described by the directed strategy rather than the mixed or random strategy as we might expect.

	β	λ	τ	*N* Participants
Stochastic	0.29 ± 0.17	6.01 ± 8.87	5.23 ± 3.57	15
Mixed	1.58 ± 1.15	8.56 ± 8.32	1.77 ± 2.43	14
Directed	1.4 ± 1.29	11.19 ± 9.21	0.24 ± 0.22	11
Greedy	0.77 ± 0.44	0.61 ± 0.84	0.14 ± 0.19	8
	0.53 ± 0.27	4.06 ± 5.35	1.3 ± 1.14	7
	1.21 ± 0.55	6.38 ± 9.77	1.32 ± 0.61	6
	0.87 ± 0.33	8.77 ± 9.51	0.22 ± 0.24	4
	3.12 ± 2.33	0.99 ± 0.48	1.82 ± 1.14	2
	1.12 ± 0.39	6.98 ± 6.09	0.89 ± 0.63	2

Table 1: Mean parameters for each group

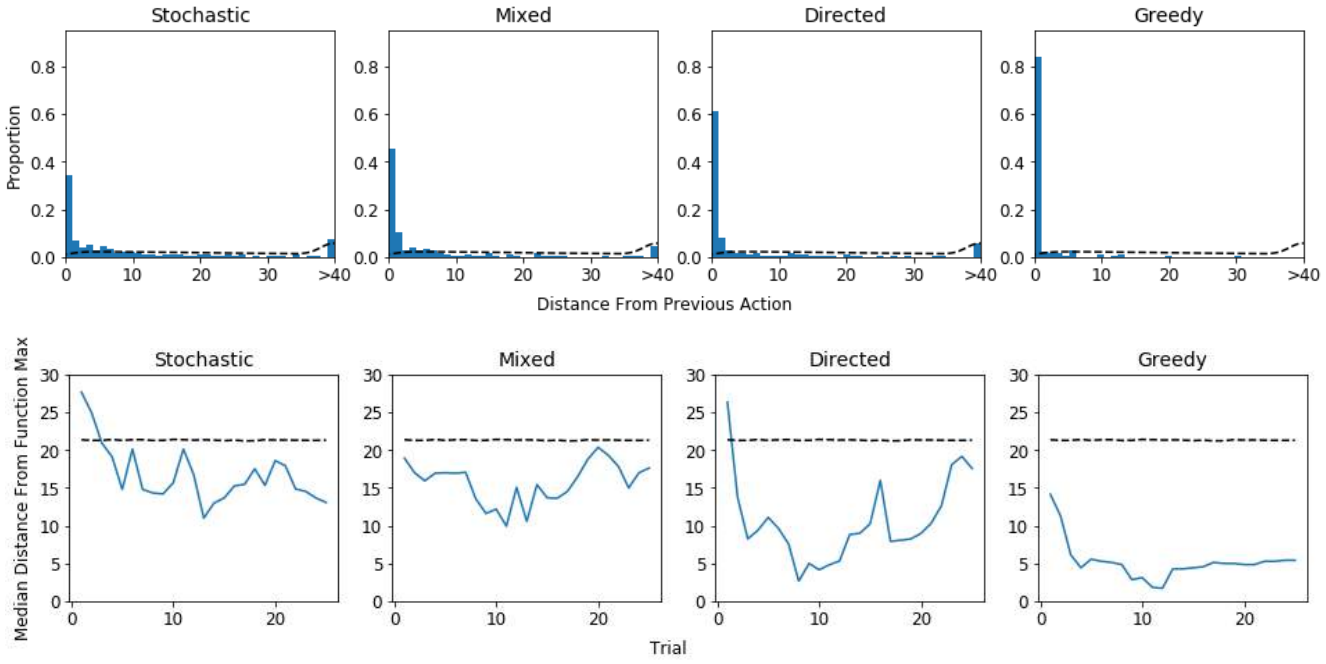


Figure 1: Distance of each action from the previous actions (top) and the reward function maximum (bottom) for the top four groups compared to a random baseline (dashed lines) reflecting uniform random action selection averaged over all conditions.

Discussion

In this study we compared behavior in an explore-exploit task across different reward function complexities and goals. While previous studies have characterized behavior in these tasks as some combination of exploitation and directed and random exploration, it was uncertain how these components might vary with different environments. Additionally, while previous studies only considered uncertainty about individual actions in directed exploration, it had yet to be established how measures of global uncertainty, such as uncertainty about the function maximum, might also be used by people to guide exploration.

Participants in this study each completed a CMAB task where the underlying reward function was either linear, quadratic, or sinusoidal, and their goal was to either maximize their score across all trials (max-score) or to find the best action (find-max). We found that behavior could be described by a relatively small set of strategies, characterized by varying exploration parameters. We found some evidence that strategy was impacted by reward function complexity, as participants in the linear condition were better described by a directed exploration strategy while those in the quadratic condition were better described by a mixed strategy utilizing both stochastic and directed exploration. However, those in the sinc condition were also best explained by a directed strategy, suggesting that these participants relied less on stochastic exploration than those in the quadratic condition despite their relatively complex reward function. Finally, we found that global uncertainty was indeed a measure used in directed exploration alongside uncertainty about individual actions, though we did not find evidence that preference for one form of uncertainty over the other was determined by goal. However, this could have been due to participants underestimating the complexity of the sinc reward function by only exploring around local maxima. Accounts of how reward function complexity influences strategy selection should also consider perceived complexity.

While we were able to describe a wide range of exploration behaviors, it is likely that alternative strategies exist. For example, some have suggested that people approach explore-exploit tasks in two qualitatively different phases, starting with a "pure exploration" phase, designed to reveal what options are most rewarding, and switching to a "pure exploitation" phase focusing on the most rewarding options (Steyvers & Wagenmakers, 2009). Another possibility is that some people do not utilize information about the reward function at all, instead exploring locally as often observed in ecological search strategies (Hills, 2006). A complete account of the types of strategies that people utilize under different circumstances should include a wider array of possible sources for guiding exploration.

References

Audibert, J.-Y., & Bubeck, S. (2010). Best Arm Identification in Multi-Armed Bandits. In *COLT - 23th Conference on Learning Theory*.

Auer, P., Cesa-Bianchi, N., & Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, *47*(2), 235–256.

Borji, A., & Itti, L. (2013). Bayesian optimization explains human active search. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems 26* (pp. 55–63).

DeLosh, E. L., Busemeyer, J. R., & McDaniel, M. A. (1997). Extrapolation: The sine qua non of abstraction in function learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 968-986.

Gershman, S. J. (2018). Deconstructing the human algorithms for exploration. *Cognition*, *173*, 34-42.

Hennig, P., & Schuler, C. J. (2012, June). Entropy search for information-efficient global optimization. *J. Mach. Learn. Res.*, *13*(1), 1809–1837.

Hills, T. T. (2006). Animal foraging and the evolution of goal-directed cognition. *Cognitive Science*, *30*(1), 3-41.

Ishwaran, H., & James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, *96*(453), 161-173.

Kalish, M. L., Lewandowsky, S., & Kruschke, J. K. (2004). Population of linear experts: knowledge partitioning and function learning. *Psychological Review*, *111*(4), 1072.

Lee, M. D., Zhang, S., Munro, M., & Steyvers, M. (2011, June). Psychological models of human and optimal performance in bandit problems. *Cogn. Syst. Res.*, *12*(2), 164–174.

Li, L., Chu, W., Langford, J., & Schapire, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on world wide web* (pp. 661–670). New York, NY, USA: ACM.

Lieder, F., Helen, A. A., & Griffiths, T. L. (2017). Strategy selection as rational metareasoning..

Lucas, C. G., Griffiths, T. L., Williams, J. J., & Kalish, M. L. (2015). A rational model of function learning. *Psychonomic bulletin & review*, *22*(5), 1193–1215.

Mockus, J. (1974). On bayesian methods for seeking the extremum. In *Proceedings of the ifip technical conference* (pp. 400–404). London, UK, UK: Springer-Verlag.

Navarro, D., Griffiths, T., Steyvers, M., & Lee, M. (2006, 04). Modeling individual differences using dirichlet processes. *Journal of Mathematical Psychology*, *50*, 101-122.

Rasmussen, C. E., & Williams, C. K. I. (2005). *Gaussian processes for machine learning (adaptive computation and machine learning)*. The MIT Press.

Salvatier, J., Wiecki, T. V., & Fonnesbeck, C. (2016). Probabilistic programming in python using PyMC3. *PeerJ Computer Science*, *2*, e55.

Schulz, E., Konstantinidis, E., & Speekenbrink, M. (2018). Putting bandits into context: How function learning supports decision making. *Journal of Experimental*

Function	Goal	Stochastic	Mixed	Directed	Greedy
All	Find Max	-108.95	-98.70	-99.36	**-97.43**
All	Max Score	-108.06	-99.37	**-88.55**	-92.21
Linear	All	-109.63	-101.13	**-88.38**	-92.01
Quadratic	All	-105.12	**-90.35**	-93.46	-92.05
Sinc	All	-111.19	-106.83	**-99.86**	-100.67
Linear	Find Max	-110.32	-99.10	-96.36	**-94.35**
Linear	Max Score	-108.93	-103.17	**-80.41**	-89.66
Quadratic	Find Max	-105.58	**-91.32**	-97.46	-95.76
Quadratic	Max Score	-104.71	-89.45	-89.76	**-88.62**
Sinc	Find Max	-111.25	-106.37	-104.43	**-102.32**
Sinc	Max Score	-111.14	-107.30	**-95.28**	-99.01

Table 2: Average log likelihoods per participant. For comparison, the log likelihood for one participant under a model that assumes judgments are made uniformly at random is -109.55.

Psychology, *44*(6), 927-943.

Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical bayesian optimization of machine learning algorithms. In *Proceedings of the 25th international conference on neural information processing systems - volume 2* (pp. 2951–2959). USA.

Srinivas, N., Krause, A., Kakade, S., & Seeger, M. (2010). Gaussian process optimization in the bandit setting: No regret and experimental design. In *Proceedings of the 27th international conference on international conference on machine learning* (pp. 1015–1022). USA: Omnipress.

Steyvers, M., & Wagenmakers, E.-J. (2009). A bayesian analysis of human decision-making on bandit problems..

Stojic, H. (2016). *Strategy selection and function learning in decision making*. Unpublished doctoral dissertation, Universitat Pompeu Fabra.

Thompson, W. R. (1933). On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples. *Biometrika*, *25*(3-4), 285-294.

Wang, Z., & Jegelka, S. (2017). Max-value entropy search for efficient Bayesian optimization. In D. Precup & Y. W. Teh (Eds.), *Proceedings of the 34th international conference on machine learning* (Vol. 70, pp. 3627–3635). International Convention Centre, Sydney, Australia: PMLR.

Wu, C. M., Schulz, E., Speekenbrink, M., Nelson, J. D., & Meder, B. (2018). Generalization guides human exploration in vast decision spaces. *Nature Human Behaviour*, *2*, 915-924.

Outgroup Homogeneity Bias Causes Ingroup Favoritism

Marcel Montrey (marcel.montrey@mail.mcgill.ca)
Department of Psychology, McGill University
2001 McGill College Avenue, Montreal, QC H3A 1G1 Canada

Thomas R. Shultz (thomas.shultz@mcgill.ca)
Department of Psychology and School of Computer Science, McGill University
2001 McGill College Avenue, Montreal, QC H3A 1G1 Canada

Abstract

Ingroup favoritism, the tendency to favor ingroup over outgroup, is often explained as a product of intergroup conflict, or correlations between group tags and behavior. Such accounts assume that group membership is meaningful, whereas human data show that ingroup favoritism occurs even when it confers no advantage and groups are transparently arbitrary. Another possibility is that ingroup favoritism arises due to perceptual biases like outgroup homogeneity, the tendency for humans to have greater difficulty distinguishing outgroup members than ingroup ones. We present a prisoner's dilemma model, where individuals use Bayesian inference to learn how likely others are to cooperate, and then act rationally to maximize expected utility. We show that, when such individuals exhibit outgroup homogeneity bias, ingroup favoritism between arbitrary groups arises through direct reciprocity. However, this outcome may be mitigated by: (1) raising the benefits of cooperation, (2) increasing population diversity, and (3) imposing a more restrictive social structure.

Keywords: ingroup favoritism; outgroup homogeneity; direct reciprocity; Bayesian learning; conditional expected utility

Introduction

Ingroup favoritism is the tendency for people to favor members of their own group over members of other groups. It manifests as a bias in how people evaluate others (Brewer, 1979; Galinsky & Moskowitz, 2000), distribute rewards (Tajfel, Billig, Bundy, & Flament, 1971), mete out punishments (Bernhard, Fischbacher, & Fehr, 2006), and decide whether or not to cooperate (Dorrough, Glöckner, Hellmann, & Ebert, 2015). Though readily elicited in both natural (Rand et al., 2009) and arbitrary groups (Efferson, Lalive, & Fehr, 2008; Galinsky & Moskowitz, 2000), the existence of ingroup favoritism is puzzling. It often neither improves the population's average outcome, nor maximizes that of the individual (Nakamura & Masuda, 2012). Disagreement even exists as to whether ingroup favoritism is better understood as a preference for improving the welfare of ingroup over outgroup, or as a product of divergent beliefs about how these groups behave (Everett, Faber, & Crockett, 2015). However, empirical work suggests that people generally expect ingroup members to act in a cooperative manner (Brewer, 2008; Yamagishi, Jin, & Kiyonari, 1999), and meta-analysis confirms that this expectation is indeed stronger toward ingroup than outgroup (Balliet, Wu, & De Dreu, 2014). A promising avenue for explaining ingroup favoritism therefore seems to be understanding how people arrive at these beliefs. In short, why are ingroup members seen as more cooperative than outgroup ones?

Many theoretical models have addressed this question. One common approach is to assign phenotypic tags to individuals, and then see what is required to elicit ingroup favoritism. Such models have shown that ingroup favoritism may be selected for when tags are not arbitrary, but rather correlate with behavioral traits (Jansen & van Baalen, 2006; Masuda & Ohtsuki, 2007; Traulsen, 2008). These traits typically include willingness to cooperate, or suitability as a cooperative partner. Ingroup favoritism may thus occur when tags convey information useful in guiding the individual's own actions. Other models explain ingroup favoritism as a product of intergroup conflict (Choi & Bowles, 2007; García & van den Bergh, 2011; Konrad & Morath, 2012), where group membership may be arbitrarily decided, but remains relevant from a competitive point of view. However, a classic empirical finding is that humans show ingroup favoritism even when groups are both explicitly arbitrary and functionally irrelevant (Billig & Tajfel, 1973; Locksley, Ortiz, & Hepburn, 1980). So why should ingroup favoritism occur even when group membership is meaningless, and such outcomes are maladaptive?

One explanation is that ingroup favoritism may arise through cognitive or perceptual limitations (Masuda, 2012). For instance, humans are known to perceive outgroup members as more similar to one another than ingroup members, a bias known as outgroup homogeneity (Judd & Park, 1988). By approximating individuals' characteristics through a single group stereotype, this may serve to reduce cognitive burden (Masuda, 2012). Masuda (2012) studied the implications of such a bias on indirect reciprocity, where cooperation is conditioned on whether or not partners maintain a good reputation. In the simplest such scheme, an individual's reputation improves when it is observed to cooperate, and suffers when it is observed to defect; in more complicated schemes, reputation may, for example, also be gained by being observed punishing a defector, or lost by being observed cooperating with one. To simulate outgroup homogeneity, individuals were allowed to observe accurate reputation information about ingroup members, but only group-level information about outgroup members. Ingroup favoritism occurred, but only when additional assumptions were invoked, such as individuals using a different rule for attributing reputation to ingroup than

to outgroup members. A follow-up model by Nakamura and Masuda (2012) eliminated the need for such double standards, and also produced ingroup favoritism through indirect reciprocity. However, this time the result was contingent on reputation information being only shareable within groups, but not between them.

Here, we show that complex rules for assigning and sharing reputation are not needed to explain ingroup favoritism between arbitrary groups. Rather, outgroup homogeneity bias may drive ingroup favoritism through a much simpler mechanism: direct reciprocity (learning through personal experience). We create an agent-based computational model, where individuals are assigned arbitrary group tags, and then play a prisoner's dilemma (PD) game. These individuals use Bayesian inference to learn how likely others are to cooperate or defect, and then act rationally by maximizing their conditional expected utility. We show that introducing outgroup homogeneity bias into this minimal setting is sufficient to produce strong ingroup favoritism, and propose several ways of mitigating this outcome.

Model

Prisoner's Dilemma

We consider a PD game where pairs of neighboring individuals interact by either cooperating (C) or defecting (D). The game is parameterized by two values: the benefit of receiving cooperation, b, and the cost of cooperating, c. When both individuals cooperate, both receive the benefit of cooperation, but pay the cost of cooperating, $b-c$. If one individual defects while the other cooperates, then the cooperator pays the cost while receiving no benefit, $-c$, while the defector pays no cost but receives the full benefit, b. When both individuals defect, neither receives the benefit nor pays the cost. The following table summarizes the row player's payoffs:

	C	D
C	$b-c$	$-c$
D	b	0

As long as $b > c > 0$, each player's payoff is always improved by defecting, no matter what the other player does. This makes the game a dilemma, because although the best individual outcome is unilateral defection, the best average outcome is mutual cooperation.

Social Structure

In PD, ingroup favoritism is operationalized as a higher rate of cooperation toward ingroup partners than outgroup ones (Dorrough et al., 2015; Fu et al., 2012; Gray et al., 2014; Masuda, 2012). For ingroup favoritism to be possible, cooperation must also be possible. By constraining which individuals interact, we promote repeat interactions, which in turn promotes cooperation (Szabó & Fáth, 2007). For each run, we generate a random r-regular graph (Bollobás, 2001) with 1000 vertices, using Steger and Wormald's (1999) algorithm. Each vertex represents an individual, and each edge represents a connection between neighbors. This graph governs interactions by limiting individuals to playing PD exclusively with their neighbors.

Group Tags

Individuals are divided into m groups, where group membership is represented by a tag visible to all other individuals. By default, $m = 2$, though it may take other values, as long as $m > 1$. Otherwise, tags cease to represent group membership, and instead become a universally shared characteristic. Each individual is randomly assigned a tag, such that each group has the same initial number of members. When replacement occurs, newcomers are assigned a tag uniformly at random.

Rational Bayesian Learning

Learning involves estimating a pair of parameters for each partner i that the individual interacts with. The first parameter p_i represents the estimated probability that partner i will cooperate with the individual, given that the individual cooperates with that partner, $Pr(C_i|C)$. The second parameter q_i estimates the probability that partner i will cooperate, given that the individual defects against that partner, $Pr(C_i|D)$. Because the game is simultaneous, actions cannot be conditioned on those of the partner. However, there is no *a priori* reason for individuals to know this, and indeed repeated interactions cause p_i and q_i to diverge, as individuals change their behavior in response to that of their partner. Individuals use Bayesian inference to arrive at point estimates for p_i and q_i. Here, the posterior predictive distribution corresponds to the posterior mean (Griffiths, Kalish, & Lewandowsky, 2008),

$$p_i := \frac{n_{CC}+\alpha+1}{n_{CC}+n_{CD}+\alpha+\beta+2} \quad q_i := \frac{n_{DC}+\alpha+1}{n_{DC}+n_{DD}+\alpha+\beta+2}, \tag{1}$$

where n_{AB} counts the number of times the individual took action A when partner i took action B. Similarly, α and β are pseudocounts (Griffiths et al., 2008) that encode prior knowledge or expectations about the frequency of cooperation and defection, respectively. These take the value $\alpha = \beta = 0$, which represents a neutral prior (uniform distribution), where neither cooperation nor defection is seen as inherently more likely.

By default, individuals maintain a pair of p and q values for each partner i. However, individuals exhibiting outgroup homogeneity bias do not distinguish between outgroup members, so they instead track a single pair of values, p_j and q_j, for each outgroup j. Outgroup homogeneity thus causes individuals to treat outgroups as if they were a single individual.

Individuals act rationally on their Bayesian estimates, so as to maximize their conditional expected utility (Jeffrey, 1990). More formally, an individual cooperates if

$$pb - c > qb, \tag{2}$$

and defects otherwise. To give individuals a chance to sample both actions, we implement a small trembling-hand parameter (Selten, 1975). When an individual selects an action, with a small probability $\varepsilon = 0.01$, it takes the opposite action

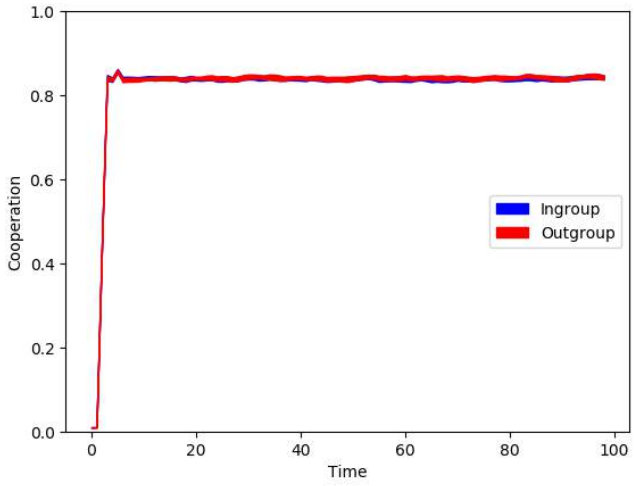


Figure 1: Ingroup and outgroup cooperation rates over time, in the absence of outgroup homogeneity bias. Cooperation rates climb rapidly as individuals learn that defection is met with defection. Ingroup cooperation rates mirror outgroup cooperation rates, because group membership is irrelevant.

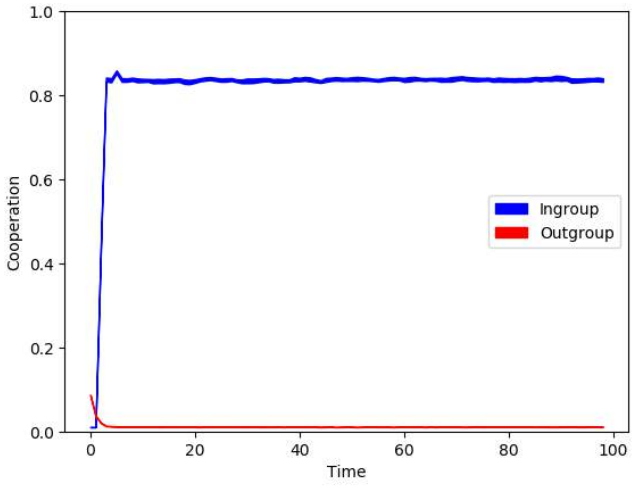


Figure 2: Ingroup and outgroup cooperation rates over time, with outgroup homogeneity bias. Here, individuals track outgroup members' behavior at a group rather than individual level. Breakdowns in cooperation result in cascades of defection involving entire groups, rather than just the offending individual, resulting in strong ingroup favoritism.

instead. Removing this parameter (setting $\varepsilon = 0$) does not qualitatively alter our results.

Simulation

At each time step, individuals interact with their neighbors in random order. Interactions involve selecting an action (C or D), and then playing PD. After each interaction, individuals note the outcome of the game, and then update their estimates p and q. Once everyone has finished playing, individuals are subjected to a 0.01 probability of being replaced. Newcomers are assigned a group tag uniformly at random, and have no knowledge of their predecessor's p and q values. Because there is no selection over genotypes, ingroup favoritism cannot evolve, but arises through phenotypic plasticity (i.e. learning) instead. We run simulations for 1000 time steps, by which time cooperation rates have long stabilized. All results are averaged across 20 independent runs, and stabilized cooperation rates are further averaged over the last 100 time steps. In all figures, line width represents 95% confidence intervals.

Results

We first consider unbiased individuals, connected to $r = 10$ random neighbors, where the benefit of receiving cooperation ($b = 3$) moderately exceeds the cost of giving it ($c = 1$). In the first few time steps, cooperation rates are near-zero (Figure 1). Recall that naïve agents have a uniform prior, meaning that cooperation and defection are seen as equally likely. However, the expected utility of unilateral defection is higher than that of mutual cooperation (Inequality 2), and so virtually everyone defects, hoping to take advantage of a cooperating partner. Individuals quickly learn that defection is met with defection, which lowers their estimate of q. With unilateral cooperation seeming increasingly unlikely, qb falls below $pb - c$, and individuals seek out mutual cooperation instead. As cooperation is met with cooperation, estimates of p increase, and high rates of cooperation ($\sim 84\%$) are established. Although individuals are assigned to $m = 2$ random groups, group membership is irrelevant, and so ingroup and outgroup cooperation rates do not differ.

When outgroup homogeneity bias is introduced, outgroup members cease being treated as individuals, but as representatives of their group. By generalizing the outcome of each interaction to other members of the outgroup, individuals learn that defection does not yield cooperation even more rapidly than when playing against ingroup members. This causes an initial spike in outgroup cooperation (Figure 2). However, although ingroup cooperation is a bit slower to get going, it alone persists. To understand why, consider what happens when an individual cooperates, but its partner defects. If partner i is an ingroup member, then the individual revises its beliefs about that partner's willingness to cooperate, and p_i declines. Soon, $pb - c$ drops below qb, and the individual ceases to cooperate. Once partner i learns that defection does not evoke cooperation, its q falls low enough for it to also seek mutual cooperation. Any successful instance of mutual cooperation promotes further cooperation, causing p values to increase, entrenching that behavior. However, if partner i is an outgroup member, then the individual does not know who to blame for the partner's unilateral defection. The individual thus revises its beliefs about the entire group's willingness to cooperate, and p_j declines. This causes the individual to punish not just the defecting partner, but also any others from that

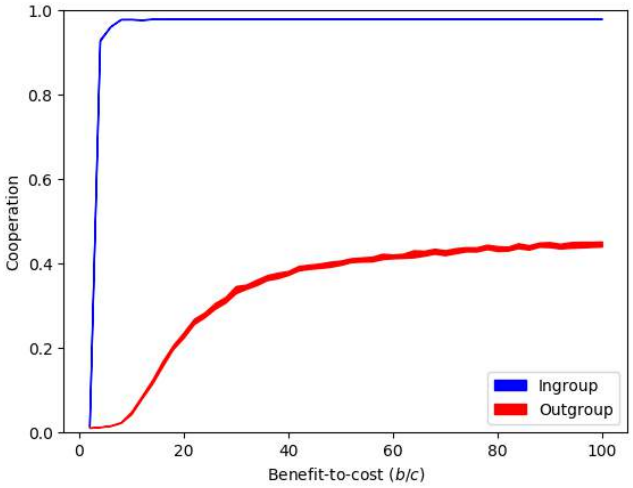


Figure 3: Stabilized ingroup and outgroup cooperation rates for various benefit-to-cost (b/c) ratios. Increasing the b/c ratio favors cooperation more broadly by making the temptation to defect less appealing, thus reducing ingroup favoritism.

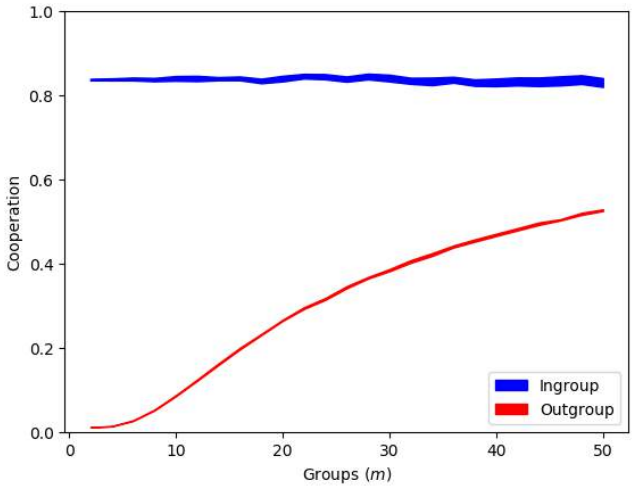


Figure 4: Stabilized ingroup and outgroup cooperation rates for various numbers of groups (m). Increasing population diversity reduces ingroup favoritism, because fewer neighbors share the same outgroup. This limits the scope of breakdowns in cooperation caused by outgroup homogeneity bias.

group. Those neighbors then punish the focal individual, as well as members of its group, for this seemingly unprovoked hostility. Intergroup defections thus bring about not just punishment of the offending individual, but also a cascade of retributive defections. Outgroup cooperation is prohibitively difficult to establish and maintain under such conditions, resulting in strong ingroup favoritism.

We next consider various parameters that may mitigate this outcome. For example, increasing the trembling-hand parameter ε reduces ingroup favoritism, albeit in a somewhat trivial manner. The more errors individuals commit in taking their desired action, the more this increases (unwanted) outgroup cooperation and decreases (desirable) ingroup cooperation. Such effects offer relatively little additional insight, however, because ingroup favoritism is merely harder to enact, rather than less sought after.

Of greater theoretical interest is the effect of increasing the benefit-to-cost ratio of cooperation. Doing so raises both ingroup and outgroup cooperation, which in turn reduces ingroup favoritism (Figure 3). Higher b/c ratios represent more cooperative games, where mutual cooperation is more rewarding, and the temptation to defect is reduced (i.e. Inequality 2 becomes primarily driven by p and q values, rather than c). Whereas the ingroup cooperation rate rapidly approaches a ceiling, the outgroup cooperation rate has more room to grow.

Another parameter of interest is the number of groups, m. This may be regarded as a measure of the population's diversity. Increasing the number of groups does not affect ingroup cooperation, but increases outgroup cooperation, thus reducing ingroup favoritism (Figure 4). Intuitively, if an individual's neighbors all belong to different groups, then tracking these groups' aggregate behavior is equivalent to tracking individual behavior. The more diverse the population, the less meaningful outgroup homogeneity is as an approximation. More practically, when fewer neighbors share group membership, breakdowns in cooperation result in smaller cascades of retributive defections.

Finally, the number of neighbors that individuals interact with, r, is also relevant. If there are relatively many groups in the population (e.g. $m = 20$), then reducing the number of neighbors alleviates cascading breakdowns in cooperation, because fewer neighbors belong to the same group. This promotes higher rates of outgroup cooperation, which in turn reduces ingroup favoritism (Figure 5). However, if there are few groups (e.g. $m = 2$), outgroup cooperation remains unsustainable, even if the number of neighbors is drastically reduced (e.g. to $r = 3$), and ingroup favoritism remains high.

Discussion

We have presented an agent-based computational model of a PD game, where outgroup homogeneity causes ingroup favoritism between arbitrary groups. Previous models have relied on indirect reciprocity (observing others' interactions) to produce such an outcome. However, these only produced ingroup favoritism if they invoked additional factors. For instance, Masuda (2012) found that reputation assignment rules had to differ for ingroup and outgroup members, while Nakamura and Masuda (2012) found that the flow of reputation information had to be severed between groups. By contrast, our model's results are driven by direct reciprocity (learning from personal experience), which obviates the need for additional assumptions about how others' interactions are evaluated, or how that information is shared. The individuals we model leverage a minimal set of cognitive capacities.

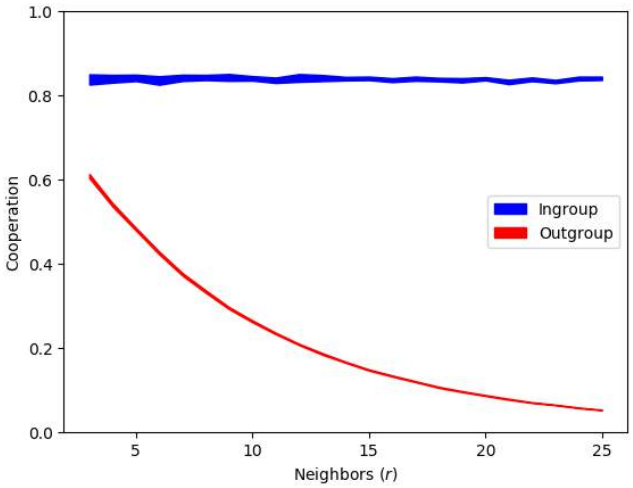


Figure 5: Stabilized ingroup and outgroup cooperation rates for various neighborhood sizes (r). If there are many groups in the population (here $m = 20$), reducing the number of neighbors reduces ingroup favoritism, by limiting the scope of defection cascades.

Namely, they learn from past experience through Bayesian inference, and make rational decisions by maximizing expected utility. In fact, they operate under similar assumptions to those found in game theoretical "fictitious play" (Berger, 2007): They estimate others' probability of cooperating as a stationary strategy, and then select the best response to observed behavior.

In our model, outgroup homogeneity causes ingroup favoritism, because outgroup defections lower an entire group's perceived cooperativeness, rather than just the individual's. Outgroup cooperation is difficult to establish and maintain not only because punishing a defector involves punishing its entire group, but also because it triggers a cascade of retributive action from those caught in the crossfire. Our findings shed light on several empirical observations about ingroup favoritism. For instance, ingroup favoritism is infamously easy to evoke even when it confers no advantage, and group membership is transparently arbitrary (Billig & Tajfel, 1973; Locksley et al., 1980). Moreover, meta-analysis suggests that ingroup favoritism between transient, experimentally-induced groups is often as strong as between natural ones (Balliet et al., 2014). The fact that this is often maladaptive from both the group and the individual's point of view makes it challenging to explain as a product of selection (Nakamura & Masuda, 2012). In our model, individuals all share the same learning and decision rules, and are assigned groups at random. There is no selection over genotypes. Rather, strong ingroup favoritism arises rapidly as a phenotypic consequence of Bayesian learning, rational decision-making, and a well-established perceptual bias: outgroup homogeneity.

One implication is that reducing or eliminating outgroup homogeneity bias may erode ingroup favoritism. In reality, ingroup favoritism can be mitigated by taking the perspective of outgroup members (Galinsky & Moskowitz, 2000). The apparent mechanism behind this result is that perspective-taking reduces reliance on group stereotypes (Brewer, 1996), causing outgroup members to be perceived as individuals. Greater intergroup contact may also reduce intergroup bias, both against that outgroup (Pettigrew & Tropp, 2006), as well as against uninvolved others (Pettigrew, 2009). In line with with our model's predictions, this too appears to be driven by reduced reliance on group stereotypes (Tadmor, Hong, Chao, Wiruchnipawan, & Wang, 2012).

Similarly, ever since Sherif's (1954) original Robbers Cave experiment, ingroup favoritism has often been both evoked and understood through the lens of competition (Sherif et al., 1961). In the experiment's final phase, intergroup tensions were deliberately reduced by encouraging cooperation. Consistent with this view, our model predicts that incentivizing cooperation alleviates ingroup favoritism. PD is a social dilemma precisely because it rewards both competition and cooperation. Increasing the b/c ratio thus minimizes these competitive aspects, and emphasizes the cooperative ones instead. Reducing the temptation to defect, relative to the benefits of mutual cooperation, causes individuals to take more risks to establish mutual cooperation, and to recover it more readily when it breaks down.

Finally, our model also predicts that ingroup favoritism may be reduced by increasing population diversity. When fewer neighbors belong to the same group, this limits the cascades of defection caused by outgroup homogeneity bias. This is also why lowering the number of neighbors can be effective. In both cases, the chances of being punished for an ingroup member's actions are reduced. However, this reasoning only applies if group membership is indeed arbitrary. The role of diversity in ingroup favoritism is typically studied through the lens of group differences, which add considerable complexity (Everett et al., 2015). Similarly, if conflicts exist along group lines, increased diversity may not necessarily reduce ingroup favoritism (Hewstone et al., 2014). No doubt, a great deal of real-world ingroup favoritism is intertwined with such pragmatic concerns. However, because ingroup favoritism occurs even when such concerns are irrelevant, understanding such social factors could offer promising ways of addressing it.

Acknowledgments

We thank an anonymous reviewer for their helpful comments.

References

Balliet, D., Wu, J., & De Dreu, C. K. W. (2014). Ingroup favoritism in cooperation: A meta-analysis. *Psychological Bulletin*, *140*(6), 1556–1581. doi: 10.1037/a0037737

Berger, U. (2007). Brown's original fictitious play. *Journal of Economic Theory*. doi: 10.1016/j.jet.2005.12.010

Bernhard, H., Fischbacher, U., & Fehr, E. (2006). Parochial altruism in humans. *Nature*, *442*(7105), 912–915. doi: 10.1038/nature04981

Billig, M., & Tajfel, H. (1973). Social categorization and similarity in intergroup behaviour. *European Journal of Social Psychology*, *3*(1), 27–52. doi: 10.1002/ejsp.2420030103

Bollobás, B. (2001). Random Graphs. *October*, *30*(4), 1–3. doi: 10.1214/aoms/1177706098

Brewer, M. B. (1979). In-group bias in the minimal intergroup situation: A cognitive-motivational analysis. *Psychological Bulletin*, *86*(2), 307–324. doi: 10.1037/0033-2909.86.2.307

Brewer, M. B. (1996). When stereotypes lead to stereotyping: The use of stereotypes in person perception. *Stereotypes and stereotyping*, *162*, 254–275.

Brewer, M. B. (2008). Depersonalized trust and ingroup cooperation. In *Rationality and social responsibility: Essays in honor of robyn mason dawes* (pp. 215–232). doi: 10.4324/9780203889695

Choi, J. K., & Bowles, S. (2007). The coevolution of parochial altruism and war. *Science*, *318*(5850), 636–640. doi: 10.1126/science.1144237

Dorrough, A. R., Glöckner, A., Hellmann, D. M., & Ebert, I. (2015). The development of ingroup favoritism in repeated social dilemmas. *Frontiers in Psychology*, *6*(APR). doi: 10.3389/fpsyg.2015.00476

Efferson, C., Lalive, R., & Fehr, E. (2008). The Coevolution of Cultural Groups and Ingroup Favoritism. *Science*, *321*(5897), 1844–1849. doi: 10.1126/science.1155805

Everett, J. A. C., Faber, N. S., & Crockett, M. (2015). Preferences and beliefs in ingroup favoritism. *Frontiers in Behavioral Neuroscience*, *9*. doi: 10.3389/fnbeh.2015.00015

Fu, F., Tarnita, C. E., Christakis, N. A., Wang, L., Rand, D. G., & Nowak, M. A. (2012). Evolution of in-group favoritism. *Scientific Reports*, *2*. doi: 10.1038/srep00460

Galinsky, A. D., & Moskowitz, G. B. (2000). Perspective-taking: Decreasing stereotype expression, stereotype accessibility, and in-group favoritism. *Journal of Personality and Social Psychology*, *78*(4), 708–724. doi: 10.1037/0022-3514.78.4.708

García, J., & van den Bergh, J. C. (2011). Evolution of parochial altruism by multilevel selection. *Evolution and Human Behavior*, *32*(4), 277–287. doi: 10.1016/j.evolhumbehav.2010.07.007

Gray, K., Rand, D. G., Ert, E., Lewis, K., Hershman, S., & Norton, M. I. (2014). The Emergence of "Us and Them" in 80 Lines of Code: Modeling Group Genesis in Homogeneous Populations. *Psychological Science*. doi: 10.1177/0956797614521816

Griffiths, T. L., Kalish, M. L., & Lewandowsky, S. (2008, nov). Review. Theoretical and empirical evidence for the impact of inductive biases on cultural evolution. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *363*(1509), 3503–14. doi: 10.1098/rstb.2008.0146

Hewstone, M., Lolliot, S., Swart, H., Myers, E., Voci, A., Ramiah, A. A., & Cairns, E. (2014). Intergroup contact and intergroup conflict. *Peace and Conflict*. doi: 10.1037/a0035582

Jansen, V. A. A., & van Baalen, M. (2006, mar). Altruism through beard chromodynamics. *Nature*, *440*(7084), 663–666. doi: 10.1038/nature04387

Jeffrey, R. C. (1990). *The logic of decision*. University of Chicago Press.

Judd, C. M., & Park, B. (1988). Out-Group Homogeneity: Judgments of Variability at the Individual and Group Levels. *Journal of Personality and Social Psychology*. doi: 10.1037/0022-3514.54.5.778

Konrad, K. A., & Morath, F. (2012). Evolutionarily stable in-group favoritism and out-group spite in intergroup conflict. *Journal of Theoretical Biology*, *306*, 61–67. doi: 10.1016/j.jtbi.2012.04.013

Locksley, A., Ortiz, V., & Hepburn, C. (1980). Social categorization and discriminatory behavior: Extinguishing the minimal intergroup discrimination effect. *Journal of Personality and Social Psychology*, *39*(5), 773–783. doi: 10.1037/0022-3514.39.5.773

Masuda, N. (2012). Ingroup favoritism and intergroup cooperation under indirect reciprocity based on group reputation. *Journal of Theoretical Biology*, *311*, 8–18. doi: 10.1016/j.jtbi.2012.07.002

Masuda, N., & Ohtsuki, H. (2007). Tag-based indirect reciprocity by incomplete social information. *Proceedings of the Royal Society B: Biological Sciences*, *274*(1610), 689–695. doi: 10.1098/rspb.2006.3759

Nakamura, M., & Masuda, N. (2012). Groupwise information sharing promotes ingroup favoritism in indirect reciprocity. *BMC Evolutionary Biology*, *12*(1), 213. doi: 10.1186/1471-2148-12-213

Pettigrew, T. F. (2009). Secondary transfer effect of contact: Do intergroup contact effects spread to noncontacted outgroups? *Social Psychology*. doi: 10.1027/1864-9335.40.2.55

Pettigrew, T. F., & Tropp, L. R. (2006). A meta-analytic test of intergroup contact theory. *Journal of Personality and Social Psychology*. doi: 10.1037/0022-3514.90.5.751

Rand, D. G., Pfeiffer, T., Dreber, A., Sheketoff, R. W., Wernerfelt, N. C., & Benkler, Y. (2009). Dynamic remodeling of in-group bias during the 2008 presidential election. *Proceedings of the National Academy of Sciences*, *106*(15), 6187–6191. doi: 10.1073/pnas.0811552106

Selten, R. (1975, mar). Reexamination of the perfectness concept for equilibrium points in extensive games. *International Journal of Game Theory*, *4*(1), 25–55. doi: 10.1007/BF01766400

Sherif, M., Harvey, O., White, B. J., Hood, W. R., Sherif, C. W., & Muzafer Sherif, O. J. Harvey, B. Jack White, William R. Hood, C. W. S. (1961). *Intergroup Conflict and Cooperation: The Robbers Cave Experiment.* doi: 10.1016/S0006-3495(97)78905-7

Sherif, M., Harvey, O. J., & Hood, W. R. (1954). *The Robbers Cave Experiment.* doi: 10.4135/9781412963879.n482

Steger, A., & Wormald, N. C. (1999). Generating Random Regular Graphs Quickly. *Combinatorics Probability and Computing*. doi: 10.1017/S0963548399003867

Szabó, G., & Fáth, G. (2007). Evolutionary games on graphs. *Physics Reports*, *446*(4-6), 97–216. doi: 10.1016/j.physrep.2007.04.004

Tadmor, C. T., Hong, Y. Y., Chao, M. M., Wiruchnipawan, F., & Wang, W. (2012). Multicultural experiences reduce intergroup bias through epistemic unfreezing. *Journal of Personality and Social Psychology*. doi: 10.1037/a0029719

Tajfel, H., Billig, M. G., Bundy, R. P., & Flament, C. (1971). Social categorization and intergroup behaviour. *European Journal of Social Psychology*, *1*(2), 149–178. doi: 10.1002/ejsp.2420010202

Traulsen, A. (2008). Mechanisms for similarity based cooperation. *European Physical Journal B*, *63*(3), 363–371. doi: 10.1140/epjb/e2008-00031-3

Yamagishi, T., Jin, N., & Kiyonari, T. (1999). *Bounded generalized reciprocity: Ingroup boasting and ingroup favoritism* (Vol. 16) (No. 16).

Pressure to communicate across knowledge asymmetries leads to pedagogically supportive language input

Benjamin C. Morris and **Daniel Yurovsky**
{benmorris, yurovsky}@uchicago.edu
Department of Psychology
University of Chicago

Abstract

Children do not learn language from passive observation of the world, but from interaction with caregivers who want to communicate with them. These communicative exchanges are structured at multiple levels in ways that support support language learning. We argue this pedagogically supportive structure can result from pressure to communicate successfully with a linguistically immature partner. We first characterize one kind of pedagogically supportive structure in a corpus analysis: caregivers provide more information-rich referential communication, using both gesture and speech to refer to a single object, when that object is rare and when their child is young. Then, in an iterated reference game experiment on Mechanical Turk (n = 480), we show how this behavior can arise from pressure to communicate successfully with a less knowledgeable partner. Lastly, we show that speaker behavior in our experiment can be explained by a rational planning model, without any explicit teaching goal. We suggest that caregivers' desire to communicate successfully may play a powerful role in structuring children's input in order to support language learning.

Keywords: language learning; communication; computational modeling.

Introduction

One of the most striking aspects of children's language learning is just how quickly they master the complex system of their natural language (Bloom, 2000). In just a few short years, children go from complete ignorance to conversational fluency in a way that is the envy of second-language learners attempting the same feat later in life (Newport, 1990). What accounts for this remarkable transition?

Distributional learning presents a unifying account of early language learning: where infants come to language acquisition with a powerful ability to learn the latent structure of language from the statistical properties of speech in their ambient environment (Saffran, 2003). A number of experiments clearly demonstrate the early availability of such mechanisms and their utility across a range of language phenomena (Saffran, 2003; Smith & Yu, 2008). However, there is reason to be suspicious about just how precocious young learners are early in development. For example, infants' ability to track the co-occurrence information connecting words to their referents appears to be highly constrained by their developing memory and attention systems (Smith & Yu, 2013; Vlach & Johnson, 2013). Further, computational models of these processes show that the rate of acquisition is highly sensitive to variation in environmental statistics (e.g., Vogt, 2012). Thus, precocious unsupervised statistical learning appears to fall short of a complete explanation for rapid early language learning.

Even relatively constrained statistical learning could be rescued, however, if caregivers structured their language in a way that simplified the learning problem. Indeed, evidence at a variety of levels– from speech segmentation to word learning– suggests that caregivers' naturalistic communication provides exactly this kind of supportive structure (Gogate, Bahrick, & Watson, 2000; Thiessen, Hill, & Saffran, 2005; Tomasello & Farrar, 1986). Under distributional learning accounts, the existence of this kind of structure is a theory-external feature of the world that does not have an independently motivated explanation. Indeed, because of widespread agreement that parental speech is not usually motivated by explicit pedagogical goals, the calibration of speech to learning mechanisms seems a happy accident; parental speech just happens to be calibrated to children's learning needs. In this work, we take the first steps toward a unifying account of both the child's learning and the parents' production: Both are driven by a pressure to communicate successfully (Brown, 1977).

Early, influential functionalist accounts of language learning focused on the importance of communicative goals (e.g., Brown, 1977). Our goal in this work is to formalize the intuitions in these accounts in a computational model, and to test this model against experimental data. We take as the caregiver's goal the desire to communicate with the child, not about language itself, but instead about the world in front of them. To succeed, the caregiver must produce the kinds of communicative signals that the child can understand and respond contingently, potentially leading caregivers to tune the complexity of their speech as a byproduct of in-the-moment pressure to communicate successfully (Yurovsky, 2017).

To examine this hypothesis, we first analyze parent communicative behavior in a longitudinal corpus of parent-child interaction in the home (Goldin-Meadow et al., 2014). We investigate the extent to which parents tune their communicative behavior (focusing on modality– i.e. gesture vs. speech) across their child's development to align to their child's developing linguistic knowledge (Yurovsky, Doyle, & Frank, 2016). We take this phenomenon to be a case study of pedagogically supportive structure in the language environment.

We then experimentally induce this form of structured language input in a simple model system: an iterated reference

game in which two players earn points for communicating successfully with each other. Modeled after our corpus data, participants are asked to make choices about which communicative strategy to use (akin to modality choice). In an experiment on Mechanical Turk using this model system, we show that tuned, structured language input can arise from a pressure to communicate. We then show that participant behavior in our game can be explained by a rational planning model that seeks to optimize its total expected utility over the course of the game.

Corpus Analysis

We first investigate parent referential communication in a longitudinal corpus of parent-child interaction. We analyze the production of multi-modal cues (i.e. using both gesture and speech) to refer to the same object, in the same instance– an information-rich cue that we take as one instance of pedagogically supportive language input. While many aspects of CDS support learning, multi-modal cues (e.g., speaking while pointing or looking) are uniquely powerful sources of data for young children (e.g., Baldwin, 2000). Multi-modal reference may be especially pedagogically supportive if usage patterns reflect adaptive linguistic tuning, with caregivers using this information-rich cue more for young children and infrequent objects. The amount of multi-modal reference should be sensitive to the child's age, such that caregivers will be more likely to provide richer communicative information when their child is younger (and has less linguistic knowledge) than as she gets older (Yurovsky et al., 2016).

Methods

We used data from the Language Development Project– a large-scale, longitudinal corpus of parent child-interaction in the home with families who are representative of the Chicago community in socio-economic and racial diversity (Goldin-Meadow et al., 2014). These data are drawn from a subsample of 10 families from the larger corpus. Recordings were taken in the home every 4-months from when the child was 14-months-old until they were 34-months-old, resulting in 6 timepoints (missing one family at the 30-month timepoint). Recordings were 90 minute sessions, and participants were given no instructions.

The Language Development Project corpus contains transcription of all speech and communicative gestures produced by children and their caregivers over the course of the 90-minute home recordings. An independent coder analyzed each of these communicative instances and identified each time a concrete noun was referenced using speech (in specific noun form), gesture (only deictic gestures were coded for ease of coding and interpretation– e.g., pointing) or both simultaneously.

Results

These corpus data were analyzed using a mixed effects regression to predict parent use of multi-modal reference for a given referent. Random effects of subject and referent were included in the model. Our key predictors were child age and logged referent frequency (i.e. how often a given object was referred to overall across our data).

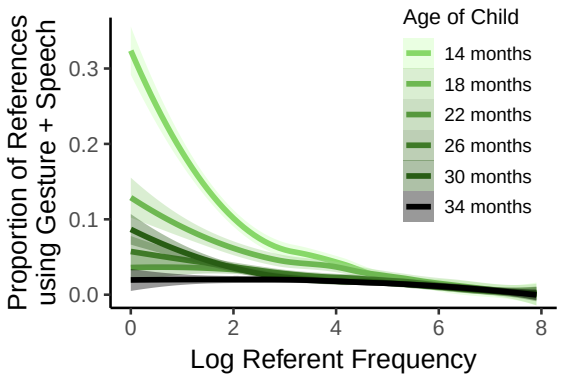


Figure 1: Proportion of parent multi-modal referential talk across development. The log of a referent's frequency is given on the x-axis, with less frequent items closer to zero.

We find a significant negative effect of child age (in months) on multi-modal reference, such that parents are significantly less likely to produce the multi-modal cue as their child gets older ($B <$ -0.04, $p < 0.0001$). We also find a significant negative effect of referent frequency on multi-modal reference as well, such that parents are significantly less likely to provide the multi-modal cue for frequent referents than infrequent ones ($B <$ -0.13, $p < 0.0001$). Thus, in these data, we see early evidence that parents are providing richer, structured input about rarer things in the world for their younger children.

Discussion

Caregivers are not indiscriminate in their use of multi-modal reference; in these data, they provided more of this support when their child was younger and when discussing less familiar objects. These longitudinal corpus findings are consistent with an account of parental alignment: parents are sensitive to their child's linguistic knowledge and adjust their communication accordingly (Yurovsky et al., 2016). Ostensive labeling is perhaps the most explicit form of pedagogical support, so we chose to focus on it for our first case study. We argue that these data could be explained by a simple, potentially-selfish pressure: to communicate successfully. The influence of communicative pressure is difficult to draw in naturalistic data, so we developed a paradigm to try to experimentally induce richly-structured, aligned input from a pressure to communicate in the moment.

Experimental Framework

We developed a simple reference game in which participants would be motivated to communicate successfully on a trial-by-trial basis. In all conditions, participants were placed in the role of speaker and asked to communicate with a computerized listener whose responses were programmed to be

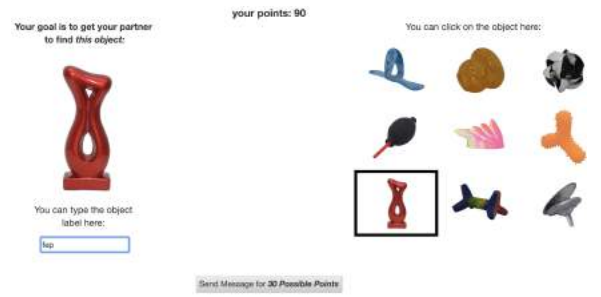


Figure 2: Screenshot of speaker view during gameplay.

contingent on speaker behavior. We manipulated the relative costs of the communicative methods (gesture and speech) across conditions, as we did not have a direct way of assessing these costs in our naturalistic data, and they may vary across communicative contexts. In all cases, we assumed that gesture was more costly than speech. Though this need not be the case for all gestures and contexts, our framework compares simple lexical labeling and unambiguous deictic gestures, which likely are more costly and slower to produce (see Yurovsky, Meyers, Burke, & Goldin-Meadow, 2018). We also established knowledge asymmetries by pre-training participants and manipulating how much training they thought their partner received. Using these manipulations, we aimed to experimentally determine the circumstances under which richly-structured input emerges, without an explicit pedagogical goal.

Method

Participants 480 participants were recruited though Amazon Mechanical Turk and received $1 for their participation. Data from 51 participants were excluded from subsequent analysis for failing the critical manipulation check and a further 28 for producing pseudo-English labels (e.g., 'pricklyy-one'). The analyses reported exclude the data from those participants, but all analyses were also conducted without excluding any participants and all patterns hold ($ps < 0.05$).

Design and Procedure Participants were exposed to nine novel objects, each with a randomly assigned pseudo-word label. We manipulated the exposure rate within-subjects: during training participants saw three of the nine object-label mappings four times, two times, or one time. Participants were then given a recall task to establish their knowledge of the novel lexicon (pretest).

Prior to beginning the game, participants are told how much exposure their partner has had to the lexicon and also that they will be asked to discuss each object three times. As a manipulation check, participants are then asked to report their partner's level of exposure, and are corrected if they answer wrongly. Then during gameplay, speakers saw a target object in addition to an array of all nine objects (see Figure 2 for the speaker's perspective). Speakers had the option of either directly click on the target object in the array (gesture)- a higher cost cue but without ambiguity- or typing a label for the object (speech)- a lower cost cue but contingent on the listener's shared linguistic knowledge. After sending the message, speakers are shown which object the listener selected.

Speakers could win up to 100 points per trial if the listener correctly selected the target referent. We manipulated the relative utility of the speech cue between-subjects across two conditions: low relative cost for speech ('Low Relative Cost') and higher relative cost for speech ('Higher Relative Cost'). In the 'Low Relative Cost' condition, speakers were charged 70 points for gesturing and 0 points for labeling, yielding 30 points and 100 points respectively if the listener selected the target object. In the 'Higher Relative Cost' condition, speakers were charged 50 points for gesturing and 20 points for labeling, yielding up to 50 points and 80 points respectively. If the listener failed to identify the target object, the speaker nevertheless paid the relevant cost for that message in that condition. As a result of this manipulation, there was a higher relative expected utility for labeling in the 'Low Relative Cost' condition than the 'Higher Relative Cost' condition.

Critically, participants were told about a third type of possible message using both gesture and speech within a single trial to effectively teach the listener an object-label mapping. This action directly mirrors the multi-modal reference behavior from our corpus data– it presents the listener with an information-rich, potentially pedagogical learning moment. In order to produce this teaching behavior, speakers had to pay the cost of producing both cues (i.e. both gesture and speech). Note that, in all utility conditions, teaching yielded participants 30 points (compared with the much more beneficial strategy of speaking which yielded 100 points or 80 points across our two utility manipulations).

To explore the role of listener knowledge, we also manipulated participants' expectations about their partner's knowledge across 3 conditions. Participants were told that their partner had either no experience with the lexicon, had the same experience as the speaker, or had twice the experience of the speaker.

Listeners were programmed with starting knowledge states initialized accordingly. Listeners with no exposure began the game with knowledge of 0 object-label pairs. Listeners with the same exposure of the speaker began with knowledge of five object-label pairs (3 high frequency, 1 mid frequency, 1 low frequency), based the average retention rates found previously. Lastly, the listener with twice as much exposure as the speaker began with knowledge of all nine object-label pairs. If the speaker produced a label, the listener was programmed to consult their own knowledge of the lexicon and check for similar labels (selecting a known label with a Levenshtein edit distance of two or fewer from the speaker's production), or select among unknown objects if no similar labels are found. Listeners could integrate new words into their knowledge of the lexicon if taught.

Crossing our 2 between-subjects manipulations yielded 6

conditions (2 utility manipulations: 'Low Relative Cost' and 'Higher Relative Cost'; and 3 levels of partner's exposure: None, Same, Double), with 80 participants in each condition. We expected to find results that mirrored our corpus findings such that rates of teaching would be higher when there was an asymmetry in knowledge where the speaker knew more (None manipulation) compared with when there was equal knowledge (Same manipulation) or when the listener was more familiar with the language (Double manipulation). We expected that participants would also be sensitive to our utility manipulation, such that rates of labeling and teaching would be higher in the 'Low Relative Cost' conditions than the other conditions.

Results

As an initial check of our exposure manipulation, a logistic regression showed that participants were significantly more likely to recall the label for objects with two exposures (B = 1.66, $p < 0.0001$) or with four exposures (B = 3.07, $p < 0.0001$), compared with objects they saw only once. On average, participants knew at least 6 of the 9 words in the lexicon (mean = 6.28, sd = 2.26).

Gesture-Speech Tradeoff. To determine how gesture and speech are trading off across conditions, we looked at a mixed effects logistic regression to predict whether speakers chose to produce a label during a given trial as a function of the exposure rate, object instance in the game (first, second, or third), utility manipulation, and partner manipulation. A random subjects effects term was included in the model. There was a significant effect of exposure rate such that there was more labeling for objects with two exposures (B = 0.91, $p < 0.0001$) or with four exposures (B = 1.83, $p < 0.0001$), compared with objects seen only once at training. Compared with the first instance of an object, speakers were significantly more likely to produce a label on the second appearance (B = 0.2, $p < 0.01$) or third instance of a given object (B = 0.46, $p < 0.0001$). Participants also modulated their communicative behavior on the basis of the utility manipulation and our partner exposure manipulation. Speakers in the Low Relative Cost condition produced significantly more labels than participants in the Higher Relative Cost condition (B = -0.84, $p < 0.001$). Speakers did more labeling with more knowledgeable partners; compared with the listener with no exposure, there were significantly higher rates of labeling in the same exposure (B = 1.74, $p < 0.0001$) and double exposure conditions (B = 3.14, $p < 0.001$).

Figure 3 illustrates the gesture-speech tradeoff pattern in the Double Exposure condition (as there was minimal teaching in that condition, so the speech-gesture trade-off is most interpretable). The effects on gesture mirror those found for labeling and are thus not included for brevity ($ps < 0.01$). Note that these effects cannot be explained by participant knowledge; all patterns above hold when looking *only* at words known by the speaker at pretest ($ps < 0.01$). Further, these patterns directly mirror previous corpus analyses demonstrating the gesture-speech tradeoff in naturalistic parental communicative behaviors, where lexical knowledge is likely for even the least frequent referent (see Yurovsky et al., 2018).

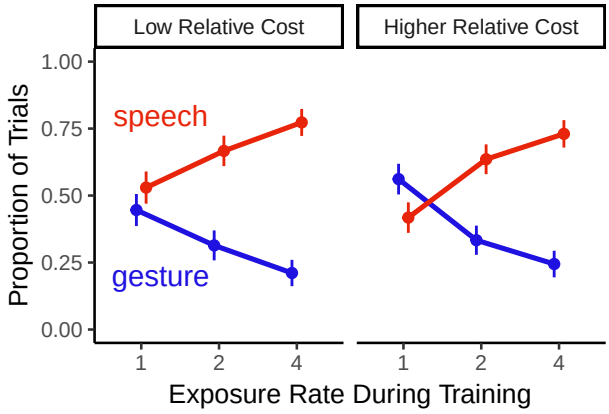


Figure 3: Speaker communicative method choice as a function of exposure and the utility manipulation. Data are taken from the Double Exposure manipulation. Rates of teaching were minimal and are not shown.

Emergence of Teaching. In line with our hypotheses, a mixed effects logistic regression predicting whether or not teaching occurred on a given trial revealed that teaching rates across conditions depend on all of the same factors that predict speech and gesture (see Figure 4). There was a significant positive effect of initial training on the rates of teaching, such that participants were more likely to teach words with two exposures (B = 0.26, $p < 0.05$) and four exposures (B = 0.25, $p < 0.05$), compared with words seen only once at training. There was also a significant effect of the utility manipulation such that being in the Low Relative Cost condition predicted higher rates of teaching than being in the Higher Relative Cost condition (B = -0.96, $p < 0.001$), a rational response considering teaching allows one to use a less costly strategy in the future and that strategy is especially superior in the Low Relative Cost condition.

We found an effect of partner exposure on rates of teaching as well: participants were significantly more likely to teach a partner with no prior exposure to the language than a partner with the same amount of exposure as the speaker (B = -1.63, $p < 0.0001$) or double their exposure (B = -3.51, $p < 0.0001$). The planned utility of teaching comes from using another, cheaper strategy (speech) on later trials, thus the expected utility of teaching should decrease when there are fewer subsequent trials for that object, predicting that teaching rates should drop dramatically across trials for a given object. Compared with the first trial for an object, speakers were significantly less likely to teach on the second trial (B = -0.84, $p < 0.0001$) or third trial (B = -1.67, $p < 0.0001$).

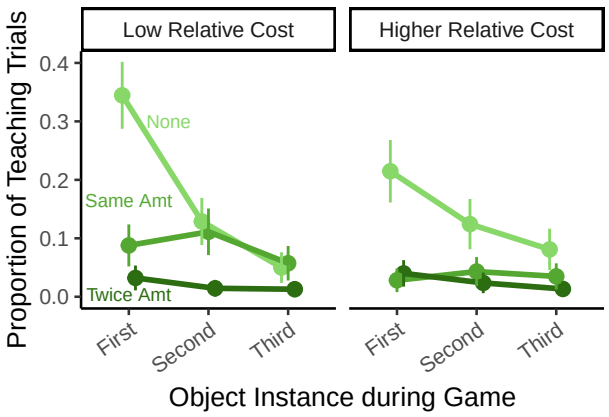


Figure 4: Rates of teaching across the 6 conditions, where the x-axis how many times an object had been the target object.

Discussion

As predicted, the data from our paradigm corroborate our findings from the corpus analysis, demonstrating that pedagogically supportive behavior emerges despite the initial cost when there is an asymmetry in knowledge and when speech is less costly than other modes of communication. While this paradigm has stripped away much of the interactive environment of the naturalistic corpus data, it provides important proof of concept that the structured and tuned language input we see in those data could arise from a pressure to communicate. The paradigm's clear, quantitative predictions also allow us to build a formal model to predict our empirical results.

Model: Communication as planning

The results from this experiment are qualitatively consistent with a model in which participants make their communicative choices to maximize their expected utility from the reference game. We next formalize this model to determine if these results are predicted quantitatively as well.

We take as inspiration the idea that communication is a kind of action–e.g. talking is a speech act (Austin, 1975). Consequently, we can understand the choice of *which communicative act* a speaker should take as a question of which act would maximize their utility: achieving successful communication while minimizing their cost (Frank & Goodman, 2012). In this game, speakers can take three actions: talking, pointing, or teaching. In this reference game, these Utilities (U) are given directly by the rules. Because communication is a repeated game, people should take actions that maximize their Expected Utility (EU) over the course of not just this act, but all future communicative acts with the same conversational partner. We can think of communication, then as a case of recursive planning. However, people do not have perfect knowledge of each-other's vocabularies (v). Instead, they only have uncertain beliefs (b) about these vocabularies that combine their expectations about what kinds of words people with as much linguistic experience as their partner are likely to know with their observations of their partner's behavior in past communicative interactions. This makes communication a kind of planning under uncertainty well modeled as a Partially Observable Markov Decision Process (POMDP, Kaelbling, Littman, & Cassandra, 1998).

Optimal planning in a POMDP involves a cycle of four phases: (1) Plan, (2) Act, (3) Observe, (4) Update beliefs. When people plan, they compute the Expected Utility of each possible action (a) by combining the Expected Utility of that action now with the Discounted Expected Utility they will get in all future actions. The amount of discounting (γ) reflects how people care about success now compared to success in the future. In our simulations, we set $\gamma = .5$ in line with prior work. Because Utilities depend on the communicative partner's vocabulary, people should integrate over all possible vocabularies in proportion to the probability that their belief assigns to that ($\mathbb{E}_{v\sim b}$).

$$EU\left[a|b\right] = \mathbb{E}_{v\sim b}\left(U(a|v) + \gamma \mathbb{E}_{v',o',a'}\left(EU\left[a'|b'\right]\right)\right)$$

Next, people take an action as a function of its Expected Utility. Following other models in the Rational Speech Act framework, we use the Luce Choice Axiom, in which each choice is taken in probability proportional to its exponentiated utility (Frank & Goodman, 2012; Luce, 1959). This choice rule has a single parameter α that controls the noise in this choice–as α approaches 0, choice is random and as α approaches infinity choice is optimal. For the results reported here, we set $\alpha = 2$ based on hand-tuning, but other values produce similar results.

$$P(a|b) \propto \alpha e^{EU[a|b]}$$

After taking an action, people observe (o) their partner's choice–sometimes they pick the intended object, and sometimes they don't. They then update their beliefs about the partner's vocabulary based on this observation. For simplicity, we assume that people think their partner should always select the correct target if they point to it, or if they teach, and similarly should always select the correct target if they produce its label and the label is in their partner's vocabulary. Otherwise, they assume that their partner will select the wrong object. People could of course have more complex inferential rules, e.g. assuming that if their partner does know a word they will choose among the set of objects whose labels they do not know (mutual exclusivity, Markman & Wachtel, 1988). Empirically, however, our simple model appears to accord well with people's behavior.

$$b'(v') \propto P\left(o|v',a\right) \sum_{v\in V} P\left(v'|v,a\right) b\left(v\right)$$

The critical feature of a repeated communication game is that people can change their partner's vocabulary. In teaching, people pay the cost of both talking and pointing together, but can leverage their partner's new knowledge on future trials. Note here that teaching has an upfront cost and the only benefit to be gained comes from using less costly communication modes later. There is no pedagogical goal– the model

treats speakers as selfish agents aiming to maximize their own utilities by communicating successfully. We assume for simplicity that learning is approximated by a simple Binomial learning model. If someone encounters a word w in an unambiguous context (e.g. teaching), they add it to their vocabulary with probability p. We also assume that over the course of this short game that people do not forget–words that enter the vocabulary never leave, and that no learning happens by inference from mutual exclusivity.

$$P\left(v'|v,a\right) = \begin{cases} 1 & \text{if } v_w \in v \& v' \\ p & \text{if } v_w \notin v \& a = \text{point+talk} \\ 0 & otherwise \end{cases}$$

The final detail is to specify how people estimate their partner's learning rate (p) and initial vocabulary (v). We propose that people begin by estimating their own learning rate by reasoning about the words they learned at the start of the task: Their p is the rate that maximizes the probability of them having learned their initial vocabularies from the trials they observed. People can then expect their partner to have a similar p (per the "like me" hypothesis, Meltzoff, 2005). Having an estimate of their partner's p, they can estimate their vocabulary by simulating their learning from the amount of training we told them their partner had before the start of the game.

Model Results

The fit between our model's predictions and our empirical data from our reference game study on Amazon Turk can be seen in Figure 5. The model outputs trial-level action predictions (e.g., "speak") for every speaker in our empirical data. These model outputs were aggregated across the same factors as the empirical data: modality, appearance, partner's exposure, and utility condition. We see a significant correlation of our model predictions and our empirical data (r = *0.94*, p<*0.0001*). Our model provides a strong fit for these data, supporting our conclusion that richly-structured language input could emerge from in-the-moment pressure to communicate, without a goal to teach.

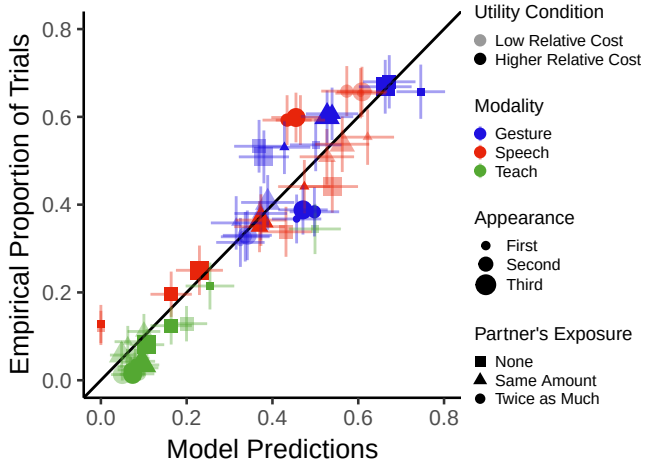


Figure 5: Fit between model predictions and empirical data.

General Discussion

We showed that people tune their communicative choices to varying cost and reward structures, and also critically to their partner's linguistic knowledge–providing richer cues when partners are unlikely to know language and many more rounds remain. These data are consistent with the patterns shown in our corpus analysis of parent referential communication and demonstrate that such pedagogically supportive input could arise from a motivation to maximize communicative success while minimizing communicative cost– no additional motivation to teach is necessary. Our account is not specific to any particular language phenomenon, though we have focused on multi-modal reference here. Given the right data or paradigm, our account should hold equally well when explaining how other information-rich language input could arise.

Of course, many aspects of language do not differ in speech to children (e.g., syntax, see Newport, Gleitman, & Gleitman, 1977). On our account, not all aspects of language should be calibrated to child's language development–only those that support communication. A full account that explains variability in modification across aspects of language will rely on a fully specified model of optimal communication. Such a model will allow us to determine both which structures are predictably unmodified, and which structures must be modified for other reasons. Nonetheless, this work is an important first step in validating the hypothesis that language input that is structured to support language learning could arise from a single unifying goal: The desire to communicate effectively.

The Mechanical Turk experiment was preregistered on Open Science Framework at https://osf.io/tjn7k
All data and code for analyses are available at https://github.com/benjamincmorris/reference-game

Acknowledgements

This research was funded by a James S. McDonnell Foundation Scholar Award to DY.

References

Austin, J. L. (1975). *How to do things with words* (Vol. 88). Oxford university press.

Baldwin, D. (2000). Interpersonal understanding fuels knowledge acquisition. *Current Directions in Psychological Science*, *9*, 40–45.

Bloom, P. (2000). *How children learn the meanings of words*. MIT press: Cambridge, MA.

Brown, R. (1977). Introduction. In C. E. Snow & C. A. Ferguson (Eds.), *Talking to children: Language input and interaction*. Cambridge, MA.: MIT Press.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998–998.

Gogate, L. J., Bahrick, L. E., & Watson, J. D. (2000). A study of multimodal motherese: The role of temporal synchrony

between verbal labels and gestures. *Child Development*, *71*(4), 878–894.

Goldin-Meadow, S., Levine, S. C., Hedges, L. V., Huttenlocher, J., Raudenbush, S. W., & Small, S. L. (2014). New evidence about language and cognitive development based on a longitudinal study: Hypotheses for intervention. *American Psychologist*, *69*(6), 588–599.

Kaelbling, L. P., Littman, M. L., & Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, *101*, 99–134.

Luce, R. D. (1959). Individual choice behavior.

Markman, E. M., & Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology*, *20*(2), 121–157.

Meltzoff, A. N. (2005). Imitation and other minds: The "like me" hypothesis. *Perspectives on Imitation: From Neuroscience to Social Science*, *2*, 55–77.

Newport, E. L. (1990). Maturational constraints on language learning. *Cognitive Science*, *14*(1), 11–28.

Newport, E. L., Gleitman, H., & Gleitman, L. R. (1977). Mother, I'd rather do it myself: Some effects and noneffects of maternal speech style. In C. A. Ferguson (Ed.), *Talking to children language input and interaction* (pp. 109–149). Cambridge University Press.

Saffran, J. R. (2003). Statistical language learning: Mechanisms and constraints. *Current Directions in Psychological Science*, *12*(4), 110–114.

Smith, L. B., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*, 1558–1568.

Smith, L. B., & Yu, C. (2013). Visual attention is not enough: Individual differences in statistical word-referent learning in infants. *Language Learning and Development*, *9*, 25–49.

Thiessen, E., Hill, E. A., & Saffran, J. R. (2005). Infant-directed speech facilitates word segmentation. *Infancy*, *7*, 53–71.

Tomasello, M., & Farrar, M. J. (1986). Joint attention and early language. *Child Development*, *57*(6), 1454–1463.

Vlach, H. A., & Johnson, S. P. (2013). Memory constraints on infants cross-situational statistical learning. *Cognition*, *127*(3), 375–382.

Vogt, P. (2012). Exploring the robustness of cross-situational learning under zipfian distributions. *Cognitive Science*, *36*(4), 726–739.

Yurovsky, D. (2017). A communicative approach to early word learning. *New Ideas in Psychology*, 1–7.

Yurovsky, D., Doyle, G., & Frank, M. C. (2016). Linguistic input is tuned to childrens developmental level. In *Proceedings of the annual meeting of the cognitive science society* (pp. 2093–2098).

Yurovsky, D., Meyers, M., Burke, Nicole, & Goldin-Meadow, S. (2018). Children gesture when speech is slow to come. In *Proceedings of the annual meeting of the cognitive science society* (pp. 2765–2770).

A Picture is Worth 7.17 Words: Learning Categories from Examples and Definitions

Arseny Moskvichev (amoskvic@uci.edu)
University of California, Irvine

Roman Tikhonov (r.tikhonov@spbu.ru)
Saint Petersburg University, St. Petersburg, Russia
National Research University Higher School of Economics, St. Petersburg, Russia

Mark Steyvers (mark.steyvers@uci.edu)
University of California, Irvine

Abstract

Both examples and verbal explanations play an important role in learning new concepts and categories. At the same time, learning from verbal explanations is not accounted for in most category learning models, and is not studied in the traditional category learning paradigm. We propose a *rational category communication* model that formally describes the process of communicating a category structure using both verbal explanations and visual examples in a pedagogical setting. We build our model based on the assumption that verbal instructions are best suited for communication of crude constraints on a category structure, while exemplars complement it by providing means for finer adjustments. Our empirical study demonstrates that verbal communication is indeed more robust to changes in stimuli dimensionality, but that its efficiency is adversely affected when distinguishing between categories requires perceptual precision. Communicating through examples has a reversed pattern. We hope that both the proposed experimental paradigm and the computational model would facilitate further research into the relative roles of verbal and exemplar communication in category learning.

Keywords: categorization; category learning; computational modelling; communication efficiency; communication channels

Introduction

Humans have a variety of information sources available to enrich or expand their knowledge. Imagine a person encountering an unfamiliar word or concept. She may infer its meaning from examples of how it is used, consult a dictionary, or use a combination of examples and definitions to understand a word or concept. In many cases, any of these sources alone is not sufficient (Fischer, 1994; Nagy, Herman, & Anderson, 1985).

Similarly, multiple sources of information are also often used to communicate a category or a concept. Imagine a family forest trip where a parent wants to teach their child about poisonous mushrooms. It is easy to envision a parent instructing their child through definitions, e.g., not to collect pale, thin-legged mushrooms with a flat cap since they are usually poisonous. Is also easy to imagine this parent giving examples, e.g. "look: this is one of the poisonous mushrooms I told you about". A key difference is that the former involves a verbal explanation of a rule, while the latter relies on non-verbal ways of concept communication (relevant examples only need to be pointed at). Contrary to the situation with word learning, however, in the context of perceptual categories, the relative contributions of verbal- and example-based communication are not well understood.

We know, however, that example- and verbal-based communication are not redundant: different aspects of category and concept knowledge may require different means of communication. Verbal instructions are well suited for communication of abstract rules, but give little information about specific stimuli characteristics (Longman, Milton, Wills, & Verbruggen, 2018). Examples, in turn, provide contextual information and help to understand how to apply knowledge to a particular problem (Reed & Bolstad, 1991; Fischer, 1994). Thus both example- and verbal-based communication play a significant role in shaping human learning. As such, they should be incorporated into contemporary theories and computational models of category acquisition.

In this work, we focus on the question of what are the fundamental differences between the verbal and exemplar channels of communication. We formalize the aforementioned intuitions about these differences and propose a computational model of the process. We also run an empirical study that investigates how people communicate perceptual categories using different combinations of communication channels. In particular, we investigate how different characteristics of a category structure affect the efficiency of verbal- and exemplar- based category communication.

Related work

The problem of communicating knowledge spans a broad range of disciplines, including educational and cognitive psychology, logic, linguistics, mathematics, and philosophy.

In the area of machine learning, there is a range of works on the problem of knowledge communication (e.g., (Winston, Binford, Katz, & Lowry, 1983)). In particular, there is a growing interest in the problems of few- and zero- shot learning techniques that focuses on learning through language without ever seeing an example (DeJong & Mooney, 1986). Notably, Mitchell (Mitchell, Keller, & Kedar-Cabelli, 1986) looked specifically into the ways of learning artificial cate-

gories from verbal explanations. In most cases, these attempts are, however, centered on applications in their respective domains and do not aim understand or model the fundamental roles that different communication systems play in human interaction and learning.

Surprisingly, verbal communication has not received much attention in empirical studies of category learning and has been largely ignored in corresponding computational models. Well-established paradigms for category learning focus on the communication and acquisition of categories through examples only and miss one of the critical sources of information used in real-world situations. Considering the overwhelmingly important role of verbal communication in education and the impact of internal verbalization on the learning outcomes (Vinner, 2002; Lombrozo, 2012; Williams & Lombrozo, 2010, 2013), this omission makes the well known ironic definition of category learning as the "class of behavioral data generated by experiments that ostensibly study categorization" (Kruschke, 2008) exceedingly appropriate.

We see two related reasons for this apparent oversight. First, the fact that people use definitions to acquire knowledge is so apparent, and, at the same time, so difficult to model rigorously, that it is very tempting to ignore either as "boring" or "impractical" to study. It is sometimes seen as an unstated assumption that verbal communication would allow to simply transfer the category knowledge.

Second, learning from definitions is inherently pedagogical, and, until recently, we lacked the tools to model such situations. Historically, category learning literature focused on extracting knowledge from a neutral environment (although there are notable exceptions: (Avrahami et al., 1997)), and the formal apparatus for modeling pedagogical reasoning in category learning was developed only recently (Shafto, Goodman, & Griffiths, 2014; Aboody, Velez-Ginorio, Laurie, Santos, & Jara-Ettinger, 2018; Frank & Goodman, 2012).

Even though recent years have witnessed a revived interest in empirical studies of these distinct ways of learning (Liefooghe, Braem, & Meiran, 2018; Longman et al., 2018), the modeling aspect is critically lacking.

Overall, we believe that now, when we have the tools to model pedagogical reasoning in category learning setting, it is a good time to make a step towards a formal model of both explanation- and example-based category learning.

Relation to categorization models

While the attempts to introduce learning based on verbal explanations into category learning models are scarce, many of the prominent categorization models could be naturally extended to partially account for verbal communication. For example, in the ALCOVE model (Kruschke, 1992), verbal communication could be introduced as transferring attention weights, thus speeding up subsequent example-based learning. On the other hand, there is no clear way to introduce purely verbal communication into this or most of the other exemplar models.

In the case of RulEx (rules with exceptions) model (Nosofsky, Palmeri, & McKinley, 1994), verbal communication could be introduced as a direct rule transfer, while examples may serve as illustrating exceptions, or as a way of adjusting rule boundaries.

Another prominent categorization model, COVIS (Ashby, Paul, & Maddox, 2011), includes the verbal (rule-based) and procedural (information-integration) components. These names partially acknowledge the potential importance of verbal reasoning, and difference in learning dynamics for "verbalizable" and "non-verbalazable" categories were extensively studied by G. Ashby (Ashby et al., 2011). At the same time, the verbal system is mostly seen as a component of internal learning dynamics, and its relation to knowledge communication is not usually studied.

Overall, there are many potential ways to introduce verbal communication into existing categorization models. At the same time, learning from verbal explanations is inherently pedagogical (somebody has to produce the explanations for a student). Therefore, we find it most promising to approach the problem from the rational analysis perspective which already offers an elegant account of pedagogical reasoning in category learning. In the next sections, we describe our approach.

Computational Model

We build upon the rational account of pedagogical reasoning, introduced in (Shafto et al., 2014). That work provided an answer to the question of how a rational teacher should select the most useful example to help a rational student learn a specific category.

In their approach, a rational teacher aims to choose an example that would maximize the student's learning outcome (probability of selecting a correct hypothesis). Thus, the teacher needs a model of the student. A rational student will also try to understand why their teacher selected a specific example which means that a student has to model the teacher. The authors formalize it as a pair of equations:

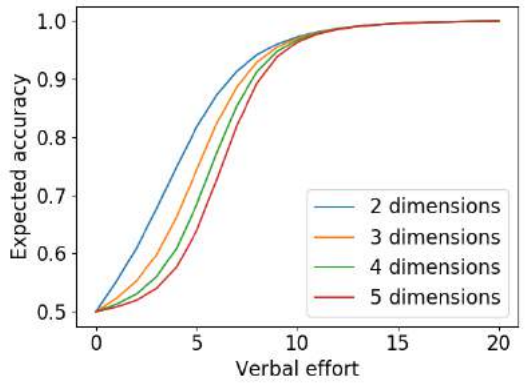
$$P_{teacher}(d|h) \propto (P_{learner}(h|d))^{\alpha} \tag{1}$$

and

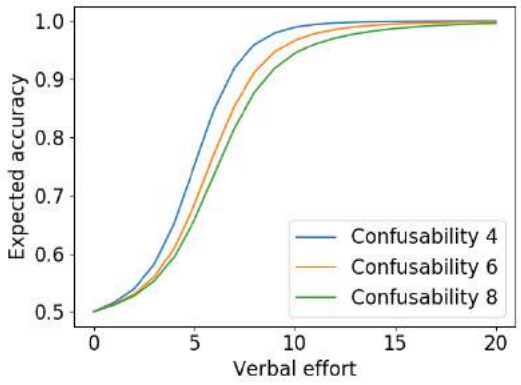
$$P_{learner}(h|d) = \frac{P_{teacher}(d|h)P(h)}{\sum_{h'} P_{teacher}(d|h')P(h')} \tag{2}$$

Where h stands for the hypothesis and d stands for the data.

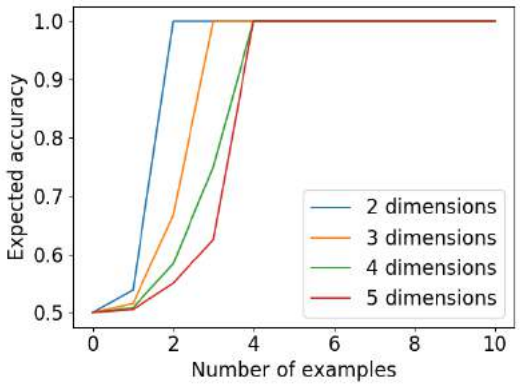
Equation 1 states that the teacher should select data points proportionally to the posterior probability of the correct hypothesis that a learner would infer after seeing these examples. Parameter α reflects how much is the teacher inclined to sample the most informative example. Thus $\alpha = 1$ corresponds to probability matching, while $\alpha = \infty$ corresponds to a deterministic selection of the best example. In the original model, an α of 1 was used in all experiments (Shafto et al., 2014).



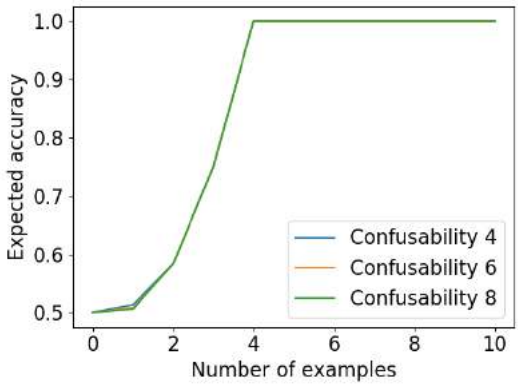
(a) The curves converge early.



(b) The curves only converge near their asymptote.



(c)



(d)

Figure 1: Simulation results of the effect of stimuli dimensionality and perceptual confusability on expected accuracy. In particular, the efficiency of exemplar channel of communication is not affected by perceptual confusability, while the number of dimensions has a noticeable impact on it. Additionally, verbal communication curves quickly converge for different stimuli dimensionality. Thus, the number of dimensions matters for low-quality verbal explanations, but their impact fades as the quality of explanations increases. On the other hand, perceptual confusability continues to matter even for high-quality explanations, highlighting the intuition that small perceptual differences may be very difficult to verbalize.

Equation 2 states that the learner should select hypotheses proportionally to how likely a rational teacher is to generate the available data under these hypotheses. A solution could be obtained by substituting one into another and iteratively updating some initial estimate until convergence.

In order to incorporate verbal communication into this model, as well as to make the model more broadly applicable, we need to make a number of changes. In the next sections we will first describe them conceptually, and then write down the resulting equations.

Sequential sampling

In (Shafto et al., 2014), authors exhaustively enumerated all possible datasets that could be communicated. Thus, if a teacher wants to show a student three examples, choosing among N possible examples every time, the space of possible datapoints is going to be N^3. This exponential data space is very limiting even for simple category learning tasks if a training session consists of more than just a few examples.

We assumed that the data is selected sequentially, in a greedy fashion. Thus, if a teacher had to select three examples, she would first select a single example that maximizes the probability of the correct category, then selects the second example conditional on the event that the student already saw the first one, and so on.

This does not guarantee an optimal sample in general, but it makes the model applicable in realistic conditions. For example, in traditional category learning experiments, which often include a large number of trials and high-dimensional stimuli as well as sequential, interactive teaching. In principle, it is also possible to combine sequential sampling with exhaustive enumeration, by adding a tractable number of examples on each step.

Formally, we rewrite Equations 1 and 2 introduce sequential dependencies (an addition to recursive teacher-student de-

pendencies already present).

$$P_{teacher}(d_i|d_{i-1},...,d_1,h) \propto (P_{learner}(h|d_i,d_{i-1},...,d_1,h))^{\alpha} \quad (3)$$

$$P_{learner}(h|d_i...d_1,h) = \\ = \frac{P_{teacher}(d_i|d_{i-1},...,d_1,h)P_{learner}(h|d_{i-1},...,d_1)}{\sum_{h'} P_{teacher}(d_i|d_{i-1},...,d_1,h')P_{learner}(h'|d_{i-1},...,d_1)} \quad (4)$$

Where d_i is a data point selected by a teacher on step i. This completes the formal description of the model for the case when all d_i are examples.

Verbal communication

The key problem we have to solve is incorporating verbal communication into the model, i.e., handling the case when d_i is a verbal explanation.

Explicitly mapping language to category structures that are communicated is an extremely difficult task. We sidestep the issue by modeling the process at a higher level: we simply assume that verbal communication channel allows us to transfer the information about which hypothesis is correct. If we view the problem this way, the problem of selecting which category structure to communicate is not relevant: we could assume that the teacher always intends to communicate the correct hypothesis.

This channel of communication has its limitations, which may depend on the category structure. For example, some hypotheses could be difficult or impossible to formulate verbally (Ashby et al., 2011), and some information could be lost due to miscommunication or misunderstanding.

To account for these phenomena, we assume that the channel is noisy. That is, even though the teacher always intends to communicate the correct hypothesis and "sends" it through the verbal channel, due to noise, instead of receiving an unambiguously decoded hypothesis, a student only receives a sample from a distribution over all possible hypotheses. The shape of this distribution depends on the hypothesis being sent and is determined by the noise model.

Noise model

It is reasonable to assume that the noise corruption is more likely to turn a hypothesis into a similar hypothesis, as opposed to turning it into something entirely unrelated. There are, however, different ways to define this similarity metric for the corruption model.

One approach is to restrict oneself to a certain class of rules and then define similarity in some intuitive way. One option would be to rely on syntactic similarity between formal expressions defining a concept (this would be similar in spirit to (Goodman, Tenenbaum, Feldman, & Griffiths, 2008)), or in some other way manually define the distance function between any two hypotheses.

We want our model to be applicable in a wide range of categorization experiments, and thus we chose not to rely on any specific choice of the hypothesis space. Instead, we model similarity between two categories simply as the similarity in the pattern of their predictions. Thus, two categories (hypotheses) are the maximally similar if they predict the same answer for all examples, and they are maximally dissimilar if they always predict different answers. We find this definition highly neutral as it builds upon the most basic definition of equality of categories: the categories are the same if the sets of things that belong to these categories are equal.

Apart from being flexible and unopinionated, our noise model captures some fundamental and intuitive properties of language: its ability to transfer the gist of the situation in broad-brush terms, and its difficulty in exactly communicating perceptual experiences. Instead of being hard-coded into the model, these properties naturally emerge from the concept similarity definition that we employed.

For example, when two rule-based categories differ only slightly in the thresholds that define them, or if two prototype-based categories differ slightly in prototype means, there would likely only be a few examples that would be misclassified if we confuse two such concepts. Thus, these categories would be similar according to our definition, and it would be difficult to discriminate between them using the verbal channel of communication.

At the same time, if two rules differ in the dimensions that are considered relevant for it, or if some dimension is "reversed" - the ramifications of confusion between such two rules would be dramatic. Such rules would be very dissimilar according to our definition, and it would be easy to distinguish between them using the verbal channel of communication.

Verbal effort

There are good explanations and there are bad ones. The same concept could be explained clearly, leaving little or no uncertainty on the student's side, or it could leave the student confused, knowing little more than before.

In order to capture this intuition, we introduce a concept of *verbal effort*. The more *verbal effort* a teacher puts into her explanation the less uncertainty there is about what was the communicated category.

Putting it together

In order to fully specify the model, we start with the Equations 3 and 4, and complement them with the case when d_i is a verbal message via the Equation 5.

$$P(h|d_i,...,d_1) \propto P(d_i|d_{i-1},...,d_1,h)P(h|d_{i-1},...,d_1) = \\ = \left[\sum_{d_i^{sent}} P(d_i|d_i^{sent})P(d_i^{sent}|h)\right] P(h|d_{i-1},...,d_1) = \\ = P(d_i|d_h^{sent})P(h|d_{i-1},...,d_1) \quad (5)$$

Where d_h^{sent} is the index of the correct hypothesis. The last equality holds since the teacher always (i.e. with probability

1) attempts to verbally communicate the true hypothesis.

Lastly, we define

$$P(d_i|d_i^{sent}) \propto \exp\{\sigma_{d_i,d_i^{sent}} \cdot \eta\} \quad (6)$$

Where $\sigma_{d_i,d_i^{sent}}$ is the correlation in predictions between the communicated hypothesis index d_i^{sent} and d_i, the (potentially noise corrupted) index of the received hypothesis. The softmax scale parameter $\eta \in [0,\infty)$ is the *verbal effort*. A verbal effort of zero corresponds to complete randomness: nothing useful was transmitted verbally. A verbal effort of infinity, in contrast, allows one to exactly identify the correct hypothesis. Currently, we fixed the steps on which the verbal communication occurs, but this restriction could be relaxed.

Overall, Equations 3, 4, 5, and 6 provide a formal definition of our model. See supplementary materials for the model implementation.

Evaluation

In the next sections we describe the experimental setting on which we collected both empirical and simulation data to test test the viability of our model.

Experiment

Method

Participants We recruited 357 participants (169 as *teachers* and 188 as *students*) through Amazon Mechanical Turk. They were native English speakers from the US. We excluded from the analysis teachers who did not reach predefined 85% accuracy threshold ($n = 40$) or failed to follow the instructions ($n = 28$), resulting in a final sample of 101 teachers.

Materials Schematic representations of fish (Rosedahl & Ashby, 2018) with possible variations in up to five visual features (fin, tail, belly color, etc.) were used as stimuli. We varied three independent variables between the participants: 1) *stimuli dimensionality* (two, three, or four dimensions) – the number of visual features varying in the presented stimuli, 2) *perceptual confusability* (low/high) – the visual similarity between stimuli of two categories, and 3) *rule type* (one- or two-dimensional). Exact visual features related to the rule dimensions were selected randomly.

Procedure Teachers learned the categorization rule by observing two sets of 15 stimuli labeled *Examples of type A* and *Examples of type B* (see Figure 2). Stimuli were presented simultaneously. Teachers had no time constraints and were able to explore each stimulus in more details by enlarging it. In the test phase, teachers had to categorize 30 stimuli presented sequentially (15 stimuli of each category including at least eight stimuli that were not presented before). Teachers who achieved the accuracy threshold of 85% in the test phase were asked to generate three training sets to teach other participants. There were three teaching formats (the order was counter-balanced across the teachers): *verbal*, *examples*, and *mixed*. In the *verbal* format teachers had to provide instructions that allow categorizing the stimuli. In the *examples* format they had to generate new stimuli of two different categories without any verbal explanations (category labels were provided). In the *mixed* format teachers were allowed to use both verbal instructions and visual examples (see Figure 2). Teachers could use as many words or visual examples as needed to explain the categorization rule, but they were instructed to be concise in their explanations and use only the minimum required amount of examples.

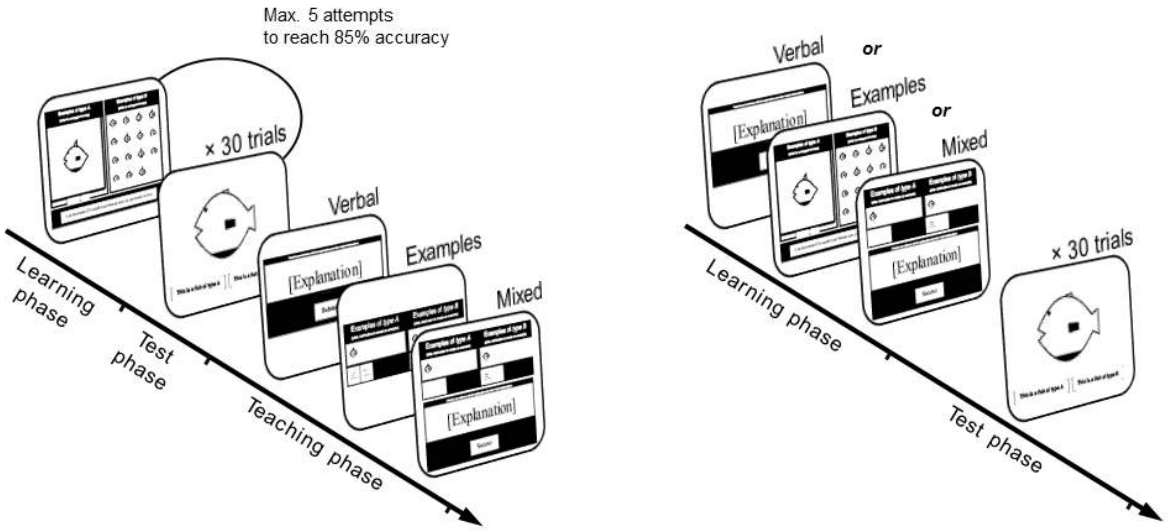
Students were randomly assigned to one of three learning conditions (verbal explanations, visual examples, or mixed), and received corresponding training materials prepared by one of the teachers. There were no time limits for the learning phase. The test phase was similar to the teachers' group.

Results

Students' performance More than 67 percent of students achieved 75% threshold criterion with median accuracy of 93 percent. Unfortunately, it results in overly low variability in the student accuracy variable. Some clear patterns were still present: one-dimensional rules result in higher performance $(.85)$ than two-dimensional $(.72)$, $p < .001$. As well as higher perceptual confusability decreased students' accuracy from .84 to .76 $(p < .001)$. However, it would be impossible to capture the more subtle interaction effects that are relevant to our study. Initially, we planned to investigate the effects of text length and explanation numbers on the students' accuracies, but students' surprisingly good performance rendered this approach impractical. Thankfully, we could switch to another interpretation to still gain insight into the problem. Since the teachers were able to create learning materials that in most cases allowed students to master the concept, we could focus on the study materials themselves: did the teachers adjust their teaching strategies to the situation? We used a Poisson regression and Generalized Estimating Equations approach to account for the teacher-to-teacher individual differences. We applied robust variance estimation techniques to compensate for potential model misspecifications.

Words per picture The average number of the visual examples provided by teachers was 4.28 in the mixed condition and 4.86 in the examples condition. The average length of the explanations increased from 28.37 in the mixed condition to 34.86 in the verbal condition because of the absence of visual examples. That is, to achieve comparable performance, the teachers needed to write approximately $34.86 \div 4.86 = 7.17$ words per example. These values could be used to map the verbal effort variable used in the computational model to the number of words in the explanation and thus put it on a more intuitive scale.

Predictors of text length and number of examples We found statistically significant effects of the rule type ($\beta = .58, p < .001$), the perceptual confusability ($\beta = .22, p = .044$), and the presence of visual examples ($\beta = -.19, p = .004$) on the length of verbal explanations (in symbols). The effects of stimuli dimensionality were not statistically signifi-



(a) Teachers

(b) Students

Figure 2: Experimental procedure illustration

cant ($\beta = -.03, p < .655$). However, the number of visual examples was predicted only by the stimuli dimensionality ($\beta = .18, p = .012$) and the rule type ($\beta = .46, p = .025$). There were also marginally significant effects of the presence of verbal explanations ($\beta = -.12, p = .052$) and the interaction of the perceptual confusability and the stimuli dimensionality ($\beta = .15, p = .089$). The effects of perceptual confusability were not statistically significant ($\beta = -.29, p = .149$).

Simulation results

We used identical experimental settings to test the performance of our computational model. We obtained initial estimates of $P(d_i|h)$ using a strong sampling assumption, and then iteratively updated them until convergence.

As shown in Figure 1, the simulation results closely correspond to the patterns we observed in the experiment. It is important to mention, however, that the behaviour depicted on Figure 1d depends on the choice of the parameter α. We used $\alpha = 1.1$ in our experiments.

Apart from capturing the key dynamics present in our data, the model also makes a range of important predictions and provides rich opportunities for further experimentation. For example, it is able to capture the mutually enriching nature of verbal and exemplar communication channels. Thus, it is possible to model situations in which using verbal explanations and exemplars together leads to dramatic leaps in performance, allowing to reach maximum accuracy, while individual channel performance is mediocre at best (0.76 for exemplars, 0.53 for verbal communication).

Discussion and conclusion

We see the main impact of our paper in identifying a fundamental limitation characteristic of most existing human category learning models (little to no account for the verbal communication) and proposing a principled and broadly applicable model to account for these phenomena.

Almost as important is the empirical demonstration of the qualitative and quantitative differences between the verbal and exemplar channels of communication. We observed that the exemplar channel is more robust to perceptual confusability of the category structures, i.e., it is more efficient in communicating categories that require higher precision in perceptual decisions. At the same time, the verbal channel is more robust to increases in the dimensionality of the stimuli.

Our simulations show that the proposed *rational category communication* model can capture the main qualitative properties of the empirical data. Additionally, the number of exemplars it chooses to ensure that a student learns a category is in close alignment with empirical data. Most importantly, it captures the difficulties of verbally explaining categories that require high perceptual precision and the robustness of exemplar communication channel to such changes.

Overall, the verbal and exemplar channels of communication have their unique strengths and weaknesses, and their relative efficiency largely depends on the structure of the hypothesis space.

While many of the reported results are preliminary, we hope that both the proposed experimental paradigm and the computational model would facilitate further research into the relative roles of verbal and exemplar information in communicating category structure. To further aid this goal, we make the model implementation openly available.

Lastly, we find that under our experimental settings, the answer to the question of "how many words is a picture worth?" is approximately 7.17.

Acknowledgments

The project was supported by RFBR grant #18-313-00249.

Supplementary materials

Model implementation and other accompanying materials: https://github.com/R-seny/rational-categorization-model

References

Aboody, R., Velez-Ginorio, J., Laurie, R., Santos, L. R., & Jara-Ettinger, J. (2018). When teaching breaks down: Teachers rationally select what information to share, but misrepresent learners' hypothesis spaces. *Proceedings of the 40th Annual Meeting of the Cognitive Science Society*, *1*, 72–77.

Ashby, F. G., Paul, E. J., & Maddox, W. T. (2011). Covis. In E. M. Pothos & A. J. Wills (Eds.), *Formal approaches in categorization* (p. 6587). Cambridge University Press. doi: 10.1017/CBO9780511921322.004

Avrahami, J., Kareev, Y., Bogot, Y., Caspi, R., Dunaevsky, S., & Lerner, S. (1997, aug). Teaching by Examples: Implications for the Process of Category Acquisition. *The Quarterly Journal of Experimental Psychology Section A*, *50*(3), 586–606. doi: 10.1080/713755719

DeJong, G., & Mooney, R. (1986). Explanation-based learning: An alternative view. *Machine learning*, *1*(2), 145–176.

Fischer, U. (1994). Learning words from context and dictionaries: An experimental comparison. *Applied Psycholinguistics*, *15*(4), 551574. doi: 10.1017/S0142716400006901

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998–998. doi: 10.1126/science.1218633

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive science*, *32*(1), 108–154.

Kruschke, J. K. (1992). Alcove: an exemplar-based connectionist model of category learning. *Psychological review*, *99*(1), 22.

Kruschke, J. K. (2008). Models of Categorization. In R. Sun (Ed.), *The cambridge handbook of computational psychology* (pp. 267–301). Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511816772.013

Liefooghe, B., Braem, S., & Meiran, N. (2018). The implications and applications of learning via instructions. *Acta Psychologica*, *184*, 1 - 3. (The implications and applications of learning via instructions) doi: https://doi.org/10.1016/j.actpsy.2017.09.015

Lombrozo, T. (2012, mar). Explanation and Abductive Inference. In K. J. Holyoak & R. G. Morrison (Eds.), *The oxford handbook of thinking and reasoning* (chap. 14). Oxford University Press. doi: 10.1093/oxfordhb/9780199734689.013.0014

Longman, C. S., Milton, F., Wills, A. J., & Verbruggen, F. (2018). Transfer of learned category-response associations is modulated by instruction. *Acta psychologica*, *184*, 144–167.

Mitchell, T. M., Keller, R. M., & Kedar-Cabelli, S. T. (1986, Mar 01). Explanation-based generalization: A unifying view. *Machine Learning*, *1*(1), 47–80. doi: 10.1023/A:1022691120807

Nagy, W. E., Herman, P. A., & Anderson, R. C. (1985). Learning words from context. *Reading research quarterly*, 233–253.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological review*, *101*(1), 53.

Reed, S. K., & Bolstad, C. A. (1991). Use of Examples and Procedures in Problem Solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(4), 753–766.

Rosedahl, L., & Ashby, F. G. (2018). A new stimulus set for cognitive research.

Shafto, P., Goodman, N. D., & Griffiths, T. L. (2014). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive psychology*, *71*, 55–89.

Vinner, S. (2002). The role of definitions in the teaching and learning of mathematics. In *Advanced mathematical thinking* (pp. 65–81). Springer.

Williams, J. J., & Lombrozo, T. (2010). The role of explanation in discovery and generalization: Evidence from category learning. *Cognitive Science*, *34*(5), 776–806.

Williams, J. J., & Lombrozo, T. (2013). Explanation and prior knowledge interact to guide learning. *Cognitive Psychology*, *66*(1), 55–84.

Winston, P. H., Binford, T. O., Katz, B., & Lowry, M. R. (1983). Learning physical descriptions from functional definitions, examples, and precedents. In *Aaai 1983*.

Communicating semantic part information in drawings

Kushin Mukherjee
Department of Cognitive Science
Vassar College
kumukherjee@vassar.edu

Robert X. D. Hawkins
Department of Psychology
Stanford University
rxdh@stanford.edu

Judith E. Fan
Department of Psychology
UC San Diego
jefan@ucsd.edu

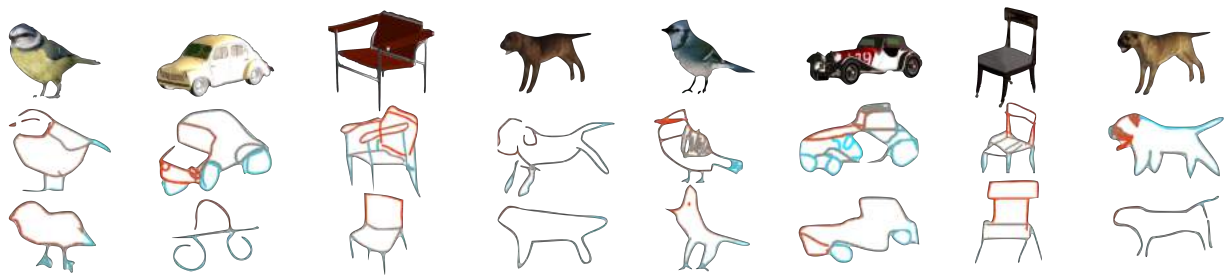


Figure 1: Objects used in communication game with example drawings below, where stroke color indicates different parts.

Abstract

We effortlessly grasp the correspondence between a drawing of an object and that physical object in the world, even when the drawing is far from realistic. How are visual object concepts organized such that we can both recognize these abstract correspondences and also flexibly exploit them when communicating them to others in a drawing? Here we consider the notion that the compositional nature of object concepts enables us to readily decompose both objects and drawings of objects into a common set of semantically meaningful parts. To investigate this, we collected data on the part information expressed in drawings by having participants densely annotate drawings of real-world objects. Our dataset contained both detailed and sparser drawings produced in different communicative contexts. We found that: (1) people are consistent in what they interpret individual strokes to represent; (2) single strokes tend to correspond to single parts, with strokes representing the same part often being clustered in time; and (3) both sparse and detailed drawings of the same object emphasize similar part information, although detailed drawings of different objects are more distinct from one another than sparse drawings. Taken together, our results support the notion that people flexibly deploy their abstract understanding of the compositional part structure of objects to communicate relevant information about them in context. More broadly, they highlight the importance of structured knowledge for understanding how pictorial representations convey meaning.

Keywords: compositionality; objects and categories; perceptual organization; sketch understanding; visual communication

Introduction

When we open our eyes, we do not experience a meaningless array of photons — instead, we parse the world into people, objects, and their relationships. The ability to represent semantically meaningful structure in our environment is a core aspect of human visual perception and cognition (Navon, 1977). As a testament to this ability, we effortlessly grasp the correspondence between a physical object in the world and a simple line drawing of it, even though such drawings lack much of the rich visual information present in real-world objects, including color and texture. How are visual object concepts organized such that they can robustly encode such abstract correspondences? Here we explore the notion that perceiving these correspondences is supported by our ability to decompose both objects and drawings into a common set of semantically meaningful parts (Biederman & Ju, 1988).

Recent advances in computational neuroscience have provided an unprecedentedly clear view into the algorithms used by the brain to extract semantic information from raw visual inputs, including drawings, exemplified by modern deep learning approaches (Fan, Yamins, & Turk-Browne, 2018; Yamins et al., 2014). Nevertheless, a major gap remains in adapting such deep learning models to emulate the structure and flexibility of human semantic knowledge (Lake, Ullman, Tenenbaum, & Gershman, 2017). A promising approach to closing this gap may be to exploit the parsimony and interpretability of structured representations that reflect how visual concepts are organized in the mind (Battaglia et al., 2018).

However, pursuit of this strategy relies upon a thorough empirical understanding of this conceptual organization and how people express this knowledge in natural behavior. We aim to contribute to this understanding by probing the expression of visual semantic knowledge in a naturalistic setting that exposes both its structure and flexibility: visual communication via drawing. This approach departs from the conventional strategy for inferring the organization of visual object concepts, which entails eliciting judgments with respect to a small number of experimenter-defined dimensions. Instead, drawing tasks permit participants to include any elements they consider relevant and combine these elements freely, yielding high-dimensional information about how people organize and deploy visual semantic knowledge under a naturalistic task objective.

Recent computational work using drawing tasks to probe visual concepts have focused on either recognition (Eitz, Hays, & Alexa, 2012; Yu et al., 2017) or generation (Ha & Eck, 2017; M. Li, Lin, Mech, Yumer, & Ramanan, 2019) of *entire* drawings. However, the question of how semantic information *within* drawings is organized has not been inves-

tigated as thoroughly (cf. L. Li, Fu, & Tai, 2018; Schneider & Tuytelaars, 2016). The goal of this paper is to present a systematic approach to analyzing the correspondence between semantic knowledge about the internal part structure of objects and the procedure by which people robustly convey this knowledge in their drawings. Specifically, this paper advances recent work investigating how drawings convey semantic information in three ways: *first*, we collect dense part annotations on freehand drawings of real-world objects, allowing an explicit focus on compositional part structure, *second*, we explore the link between this semantic structure and the dynamics of drawing production, and *third*, we examine differences in how visual semantic knowledge is expressed between contexts.

Methods

We developed a web-based crowdsourcing tool, built with `jsPsych.js` (de Leeuw, 2015), to collect dense semantic annotations of the stroke elements in drawings of real-world objects (Fig. 1).

Communicative drawing dataset

We first obtained 1195 drawings of 32 real-world objects from a previously collected experimental dataset in which pairs of participants played a drawing-based reference game (Fan, Hawkins, Wu, & Goodman, 2019).[1] Object stimuli were photorealistic 3D renderings belonging to one of four basic-level categories (i.e., bird, car, chair, dog), each of which contained eight exemplars. On each trial of the experiment, participants were presented with a shared context containing four of these objects. One participant (the sketcher) was privately cued to draw a target object so that the other participant (the viewer) could pick it out from the set of distractors. Across trials, the similarity of the distractors to the target was manipulated, yielding two types of communicative contexts: *close contexts*, in which all four objects belonged to the same basic-level category, and *far contexts*, in which objects belonged to different basic-level categories. This context manipulation led sketchers to produce relatively simpler drawings containing fewer strokes and less ink on far trials than on close trials, while still achieving high recognition accuracy in both contexts.

Prior works analyzing the semantic properties of drawing data have used a raster image representation (e.g., `*.png`), an expedient format for applying modern convolutional neural network architectures (Fan et al., 2018; Sangkloy et al., 2016; Yu et al., 2017). However, to investigate how semantic structure manifests during drawing production, it was critical to encode each drawing using a vector image format that preserves the inherently sequential and contour-based nature of drawing production (e.g., `*.svg`). Thus, each drawing in our dataset is represented as a sequence of individual strokes. A stroke is defined as the mark left by a virtual pen on a digital drawing canvas between being 'placed onto' the canvas and being 'lifted up'. We parameterized each stroke by a sequence of cubic Bézier curves, called *splines*. This format provides a compact representation of drawing data, which also preserves the sequence in which each element was produced.

[1] All materials and data are available at `https://github.com/cogtoolslab/semantic_parts`.

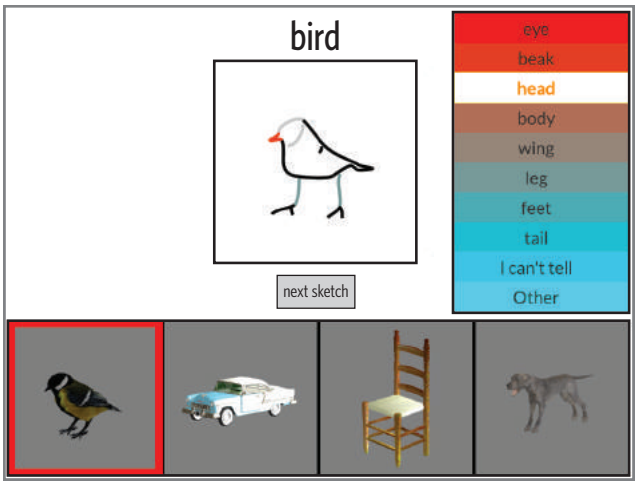


Figure 2: Annotation interface. Participants selected sub-stroke elements (splines) and tagged them with part labels.

Semantic part annotation

We crowdsourced dense semantic annotations for every spline in every stroke of the drawings from this dataset. We refer to our annotation data as *dense* because labels were provided for splines, which are at a finer level of granularity than strokes.

Participants 326 participants were recruited via Amazon Mechanical Turk (AMT) and provided informed consent in accordance with the Stanford IRB. Participants were given a base compensation of $0.35, plus $0.002 for every spline they annotated and $0.02 for every drawing they annotated completely.

Task procedure Each participant was presented with a sequence of 10 drawings that were randomly sampled from the communicative drawing dataset (Fig. 2). Their goal was to tag each spline with a label corresponding to the part it represented (e.g., seat, leg, back for a chair). To facilitate consistent tagging, participants were provided with a menu of common part labels that were associated with each basic-level category (Table 1). Participants could also generate their own part label if they believed none of the common labels applied. If any spline was too short for annotators to feasibly annotate it with their mouse cursor, it was concatenated with its neighboring splines until the resulting spline was long enough to easily select. To give participants full information about the original communicative context, we showed the drawing with the same array of four objects that the original sketcher had viewed, with the target object highlighted in red.

Data preprocessing We first standardized all 304 distinct labels provided by participants, mapping them to a common set of 24 part labels that applied to all objects in the dataset. This common set was defined as the superset of all labels that appeared in the part menu in the annotation task. Although most labels provided already exactly matched one in the common set (i.e., 90.1%), participants were permitted to assign their own custom label, resulting in additional lexical variation that we collapse over in the current analysis. For example, some custom labels were either synonymous with or more specific than one of the common labels (e.g., 'leg support', 'foot', or 'strut' for 'leg'). We manually constructed a part dictionary to map such custom labels to one of the common ones, ensuring a consistent level of granularity for all spline labels. We only examined drawings that were annotated by at least three distinct participants, providing a consistent way to evaluate annotation consistency across splines. To reduce bias due to missing data, we also restricted our analyses to annotation trials in which the drawing was completely annotated (i.e., all splines were tagged). After applying all preprocessing, our resulting dataset consisted of 864 drawings that had been completely annotated 3 times.

Results

How well do viewers agree on what strokes mean?

Before proceeding to use these annotations to examine how semantic information is conveyed during drawing production, we conducted a basic check of inter-annotator consistency. Specifically, we examined how often different annotators agreed on what each spline in a drawing represented. We found that 95.6% of all splines received the same label by at least two of the three annotators, and 67.8% of all splines received the same label by all three annotators. This shows that the way viewers interpret which part each stroke represents is systematic, validating our general approach. Further, it suggests that sketchers may exploit this systematicity to produce strokes that they expect viewers to interpret consistently. In subsequent analyses, we collapsed over inter-annotator variation: we assigned the modal label to splines to which at least two annotators had given the same label; for the remaining 4.4% of splines, we sampled one of the three labels provided.

How do strokes correspond to parts of objects?

When composing a recognizable drawing of a real-world object, how do people decide what information to convey with each stroke? A natural possibility is that their actions closely correspond to the part structure of that object. Concretely, we hypothesized that most strokes in our dataset would *not* cross part boundaries: that all splines within a given stroke would be assigned the same part label. Conversely, because depictions of parts can be arbitrarily detailed, and some parts re-occur throughout an object (e.g., multiple legs on a bird, chair, or dog), we hypothesized that there would often be more than one stroke per part (Fig. 3A).

category	part labels
bird	eye, beak, head, body, wing, leg, feet, tail
car	bumper, headlight, hood, windshield, window, body, door, trunk, wheel
chair	backrest, armrest, seat, leg
dog	eye, mouth, ear, head, neck, body, leg, paw, tail

Table 1: Part labels provided to annotators.

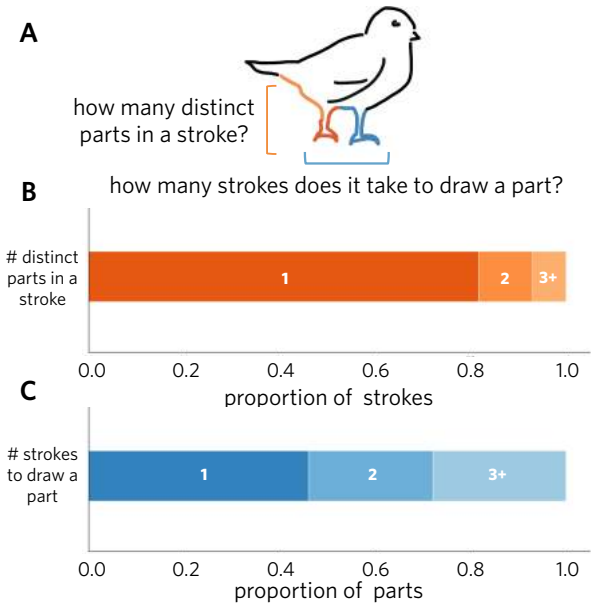


Figure 3: (A) Analyzing the correspondence between strokes and part labels: number of unique part labels assigned to different splines within the same stroke and number of different strokes used to draw each part. (B) Distribution over number of part labels within a stroke. (C) Distribution over number of strokes used to draw a part.

To evaluate the first hypothesis, we computed the number of unique part labels across all splines within each stroke. We found that for 81.6% of the strokes in our dataset there was only one part label; the remaining 18.4% of strokes were associated with two or more labels (Fig. 3B). In other words, most strokes represented exactly one part, but in a minority of cases they spanned multiple parts (e.g., a single stroke connecting the head and body of a bird, or an armrest and leg of a chair). We were concerned, however, that these proportions were inflated by strokes with very few splines.[2] To address this concern, we constructed a null model controlling for the number of splines. Part labels were randomly sampled from the full list of parts in the drawing such that each spline was equally likely to represent any part regardless of the stroke it belonged to. In simulations from this null model, only 55% of strokes corresponded to a unique part while 45% of strokes spanned multiple parts. Thus, individual strokes in our dataset were much more likely to correspond to a single part (i.e., not cross part boundaries)

[2]The modal number of splines per stroke (20% of cases) was 1, but there was a long tail; the mean number was 2.6.

than would be expected under random assignment of part labels to splines.

To evaluate the second hypothesis, we computed the number of strokes that were used to represent each part of an object (Fig. 3C). We found that 46.1% of parts were depicted using exactly one stroke, 26.0% using exactly two strokes, 11.3% using exactly three strokes, and 16.6% using four or more strokes. Thus, nearly half the time, a single action was sufficient to depict an entire object part. However, the remaining 53.9% of the time, more than one stroke was required to depict an entire part, which would be expected for those parts that consisted of multiple disconnected subparts within an object (e.g., wheels of a car, paws of a dog).

The findings so far show that the information people convey with each stroke systematically corresponds to the parts that objects contain. We next sought to understand how these properties may vary between drawings generated in different communicative contexts. Indeed, strokes spanning multiple parts were slightly more common in drawings produced in far contexts (19.4%, CI: [17.9%, 20.9%]) than close contexts (17.6%, CI: [16.1%, 18.8%][3], p= 0.07), suggesting that sketchers were somewhat more likely to use a single stroke to represent multiple contiguous parts in a context where a sparser drawing would be sufficient. And the proportion of parts requiring more than one stroke was slightly higher for close drawings (55.8%, CI: [53.7%, 58.6%]) than far drawings (52.0%, CI: [49.9%, 54.6%],$p = 0.02$), suggesting that sketchers may have included more detail per part in close drawings to distinguish the target object from similar distractors.

Do strokes representing the same part tend to be produced in succession?

In the previous section we discovered that slightly more than half of the parts in our dataset were depicted using multiple strokes. This result raised the question: to what extent are strokes depicting the same part drawn in succession, or interleaved among strokes depicting other parts?

To investigate this question, we estimated the mean length of 'streaks' containing strokes depicting the same part. First, we collapsed across the spline annotations examined in the previous section and represented each stroke by the modal part label assigned to its splines. We represented each drawing as the sequence of these part labels, and defined *part streak length* to be the number of consecutive strokes annotated with the same part label.[4] For example, in the drawing shown in Fig. 4A, two 'leg' strokes were placed before moving on to the 'foot', giving a streak of length 2. Finally, we averaged these streak length values over every drawing in the dataset to obtain our statistic.

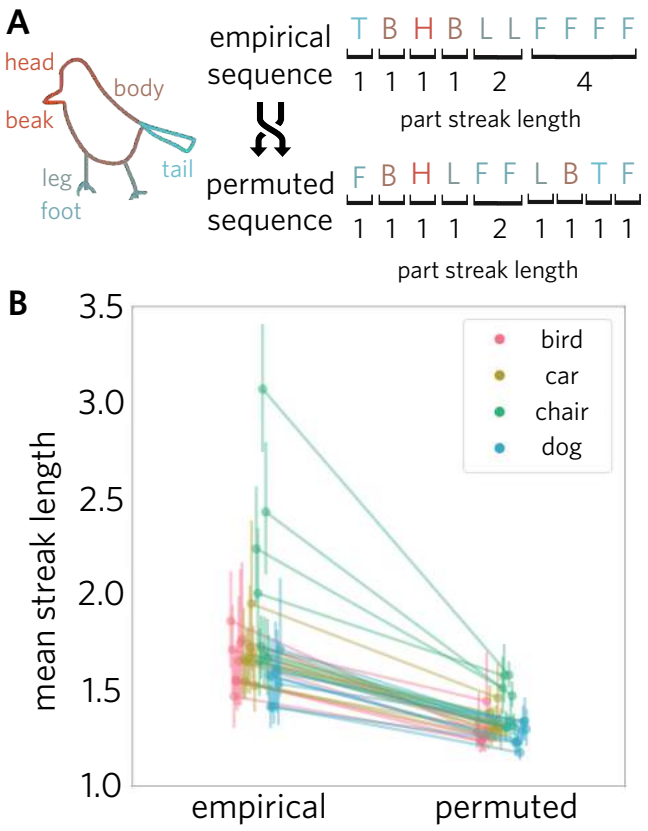


Figure 4: (A) Analysis of sequence in which strokes depicting each part were drawn. (B) Comparison of mean length of streaks consisting of strokes that depict the same part with null distribution of permuted stroke sequences.

To evaluate whether the empirical part sequences were more structured than expected if parts were drawn at random, we constructed a null model to serve as a baseline. For this null model, we permuted the part sequence such that the number of instances of each part was preserved, but the temporal structure was disrupted (Fig. 4A). We generated a null distribution of streak lengths for each drawing by repeating this permutation procedure 1000 times and measuring the mean streak length for each permutation. Finally, we obtained a z-score for each drawing by computing where the empirical streak length fell in the permuted streak length distribution. A drawing with a z-score near 0 had a streak length that was commonly obtained by placing strokes in a random order, while a drawing with a higher z-score is more structured than expected under the null.

We found that the empirical streak length was reliably higher for all objects than that of the permuted sequences (mean z-score across drawings: 2.07, CI: [1.90, 2.23]; Fig. 4B), and higher for the close drawings (mean z-score: 2.58; CI: [2.26, 2.90]) than far drawings (mean z-score: 1.56; CI: [1.38, 1.74]). The lower streak length for far drawings is consistent with their lower stroke count overall—when only

[3] 95% confidence intervals were estimated via stratified bootstrap resampling (N=1000 iterations) of drawings within each context condition.

[4] We excluded 78 out of the 864 drawings where this measure was not well-defined, i.e. sketches containing only one stroke or part label, or containing fewer than two strokes sharing the same part label.

one or two strokes are used per part, there is a ceiling on the mean streak length. However, when sketchers do use multiple strokes to convey a single part (i.e., because there are multiple subparts, or to add more detail), they tend to draw these in succession before moving on to a different part. These results suggest more broadly that the procedure by which people convey semantic information in drawings is organized by the part structure within objects.

How is part information emphasized in different communicative contexts?

Our findings so far bear on how the way people compose communicative drawings of objects reflects their semantic knowledge of the parts those objects are composed of. A key consequence of such semantically organized part knowledge is that it naturally supports flexible expression across different communicative contexts. For example, when communicating about a chair in a far context containing objects from other basic-level categories, sketchers may include only the essential information to indicate the presence of certain parts (e.g., armrests) that distinguish it at the category level. On the other hand, when communicating about that same chair in a close context containing other, perceptually similar, chairs sketchers may emphasize aspects of parts that distinguish it at the object level (e.g., the curvature of the armrests), by applying more strokes and/or more ink in each stroke.

We hypothesized that sketchers emphasize part information to preserve relevant distinctions in context. To explore this possibility, we asked the following questions: (1) How similarly is object-specific part information emphasized in both close and far contexts? (2) How do differences in how part information is emphasized *between* contexts affect how discriminable those drawings are?

To investigate these questions, we represented each drawing by a 48-dimensional *part-feature vector* that contained information about: (a) how many strokes and (b) how much total ink was allocated to each of the 24 unique part labels in our dataset. Specifically, the first 24 elements of each part-feature vector contained the number of strokes allocated to each part, and the remaining 24 contained the total arc length of all strokes allocated to each part. Because our primary goal was to understand *relative* differences in how much emphasis was placed on each part across drawings in our dataset, we first z-scored the raw stroke-count and arc-length measurements within each feature dimension, thereby mapping all features to the same unit-variance scale. We then collapsed across drawings within each object-context combination, yielding 64 average part-feature vectors (i.e., 32 objects x 2 context conditions).

Similar part information emphasized across different communicative contexts In order to investigate to what extent similar object-specific part information is emphasized in different communicative contexts, we computed the matrix of Pearson correlations between part-feature vectors. Formally, this entailed computing: $R_{ij} = {}^{cov(\vec{r}_i,\vec{r}_j)}/\sqrt{var(\vec{r}_i)\cdot var(\vec{r}_j)}$, where

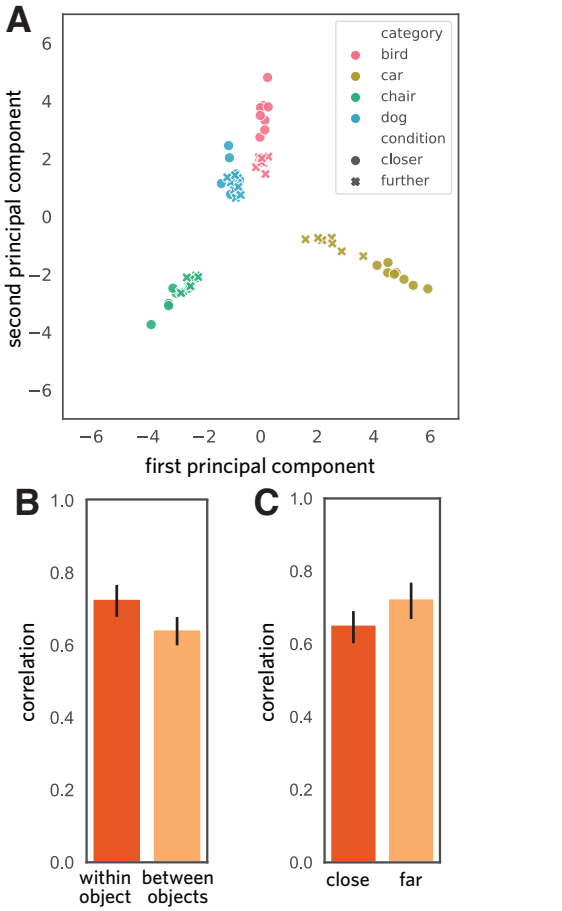


Figure 5: (A) Layout of mean part-feature vectors for each object-context combination, projected onto top two principal components. (B) Comparison of feature similarity between close and far drawings of the same object, relative to close and far drawings of different objects within a category. (C) Comparison of feature similarity between far drawings of objects within a category, relative to close drawings. Error bars reflect 95% CIs.

$\vec{r}_i$ and $\vec{r}_j$ are the mean part-feature vectors for the ith and jth object-context combinations, respectively.

While close and far drawings of an object differed in their overall amount of detail, we hypothesized that they would still emphasize part information in similar ways. Specifically, insofar as similar object-specific part information is emphasized in both close and far drawings of the same object, we predicted higher correlations between close and far part-feature vectors for the *same* object than for close and far part-feature vectors of *different* objects. Consistent with this, we found strong correlations between the feature vectors for close and far drawings of the same object ($r = 0.73$, CI: [0.68, 0.77][5]), which were significantly stronger than close and far drawings of *different* objects ($r = 0.64$, CI: [0.60, 0.68]; same objects vs. different objects: $p < 0.001$). These results show that close and far drawings of the same object exhibit similar

[5]95% confidence intervals were estimated via stratified bootstrap resampling (N=10000 iterations) of drawings within each object-context combination.

patterns of emphasis across different parts, and this similarity exceeded that expected due to merely being members of the same basic-level category (Fig. 5B).

Detailed drawings are more distinct from each other than sparser drawings While the above findings showed that close and far drawings of the same object exhibit similar patterns of emphasis on different parts, close drawings contain greater emphasis on these parts overall than far drawings (i.e., contained more and longer strokes). How were these additional strokes being spent?

We hypothesized that the additional part information provided in close drawings was being distributed across parts in different ways for different objects, thereby making them more distinguishable from one another in feature space. To evaluate this possibility, we computed the mean correlation between the part-feature vectors of close drawings of objects in a given category and compared this value with the mean correlation between far drawings of exactly the same objects. We found that close drawings were less similar to one another than far drawings were (close similarity: $r = 0.65$, CI: [0.60, 0.69]; far similarity: $r = 0.73$, CI: [0.67, 0.77]; close vs. far: $p = 0.007$), suggesting that sketchers discern which parts are most diagnostic of the target object among highly similar distractors and emphasize these parts accordingly (Fig. 5C). This was particularly apparent when we visualized the spatial layout of part-feature vectors: whereas far drawings were clustered closer together and near the origin, close drawings were spread further apart from other members of the same category and further from the origin (Fig. 5A). Observing these contextual differences is all the more remarkable given that this feature representation captures only the *amount* of emphasis allocated to each part during drawing production, setting aside their visual properties.

Discussion

In this paper, we explored how the way people compose communicative drawings of objects reflects their semantic knowledge about what objects are composed of. To accomplish this, we first collected dense semantic annotations of sub-stroke elements in communicative drawings of real-world objects that were produced in different contexts. This allowed us to interrogate the internal semantic structure within drawings, and relate this structure to the dynamics of drawing production in a naturalistic visual communication task. Overall, we found that: (1) people are highly consistent in how they interpret what individual strokes represent; (2) single strokes tend to correspond to single parts, with strokes representing the same part tending to be clustered in time; and (3) both detailed and sparse drawings of the same object emphasized similar part information, with detailed drawings of different objects tending to be more distinct from one another than simpler ones. Taken together, our results support the notion that people deploy their abstract understanding of the compositional part structure of objects in order to select actions to communicate relevant information about them in context.

These findings are resonant with classic and recent work that has argued for the importance of compositionality in human perception and cognition in general (Biederman, 1987; Battaglia et al., 2018; Lake et al., 2017), and for visual production in particular (Lake, Salakhutdinov, & Tenenbaum, 2015). However, unlike prior work which focused on the production of abstract symbols (Lake et al., 2015), we consider the challenge of how people transform perceptually grounded representations of real-world objects into procedures for producing figurative drawings that communicate not only what they see and know about them, but also what is relevant in context.

Our work is also related to recent progress in the development of computational models of drawing production (Ha & Eck, 2017; M. Li et al., 2019). While results from these efforts have been galvanizing, the development of principled metrics by which to rigorously evaluate how well they emulate human drawing behavior has not kept pace. By interrogating in detail how humans encode semantic information into their drawings, and flexibly adjust their production behavior in different contexts, this paper presents a first step towards such a set of behavioral metrics. Having such metrics is important because they would enhance our ability to distinguish between generative models, and thereby help advance further model development. It would thus be valuable to apply up our analytical approach to the large drawing datasets (Eitz et al., 2012; Sangkloy et al., 2016; Jongejan, Rowley, Kawashima, Kim, & Fox-Gieg, 2017) that have provided the basis for these modeling approaches.

In ongoing work, we are extending our analysis of how different part information is expressed in drawings beyond simple effort cost measures (i.e., number of strokes, amount of ink) to encompass content and style information (e.g., the shape of a bird's wing, caricaturization of a chair's armrest). We expect that augmenting current vision models with a combination of the requisite semantic part knowledge and the ability to discern perceptual properties of these parts, such as style, will enable us to build models that parse drawings in a more human-like way. More broadly, achieving this synthesis will lead to both more robust artificial intelligence and a deeper understanding of human cognition and behavior.

Acknowledgments

KM was supported by the Department of Cognitive Science at Vassar College through its Humanities in Cognitive Science program and the Center for the Study of Language and Information at Stanford University. RXDH was supported by the National Science Foundation Graduate Research Fellowship (DGE-114747).

All code and materials available at:
`https://github.com/cogtoolslab/semantic_parts`

References

Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., ... others (2018). Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*.

Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychological Review*, *94*(2), 115.

Biederman, I., & Ju, G. (1988). Surface versus edge-based determinants of visual recognition. *Cognitive Psychology*, *20*(1), 38–64.

de Leeuw, J. R. (2015). jspsych: A javascript library for creating behavioral experiments in a web browser. *Behavior Research Methods*, *47*, 1–12.

Eitz, M., Hays, J., & Alexa, M. (2012). How do humans sketch objects? *ACM Trans. Graph.*, *31*(4), 44–1.

Fan, J., Hawkins, R., Wu, M., & Goodman, N. (2019). Pragmatic inference and visual abstraction enable contextual flexibility during visual communication. *arXiv preprint arXiv:1903.04448*.

Fan, J., Yamins, D., & Turk-Browne, N. (2018). Common object representations for visual production and recognition. *Cognitive Science*.

Ha, D., & Eck, D. (2017). A neural representation of sketch drawings. *arXiv preprint arXiv:1704.03477*.

Jongejan, J., Rowley, H., Kawashima, T., Kim, J., & Fox-Gieg, N. (2017). *Google Quickdraw.* Retrieved from `https://quickdraw.withgoogle.com/`

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, *350*(6266), 1332–1338.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*.

Li, L., Fu, H., & Tai, C.-L. (2018). Fast sketch segmentation and labeling with deep learning. *IEEE computer graphics and applications*.

Li, M., Lin, Z., Mech, R., Yumer, E., & Ramanan, D. (2019). Photo-sketching: Inferring contour drawings from images. *arXiv preprint arXiv:1901.00542*.

Navon, D. (1977). Forest before trees: The precedence of global features in visual perception. *Cognitive Psychology*, *9*(3), 353–383.

Sangkloy, P., Burnell, N., Ham, C., & Hays, J. (2016). The sketchy database: learning to retrieve badly drawn bunnies. *ACM Transactions on Graphics (TOG)*, *35*(4), 119.

Schneider, R. G., & Tuytelaars, T. (2016). Example-based sketch segmentation and labeling using crfs. *ACM Transactions on Graphics (TOG)*, *35*(5), 151.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624.

Yu, Q., Yang, Y., Liu, F., Song, Y.-Z., Xiang, T., & Hospedales, T. M. (2017). Sketch-a-net: A deep neural network that beats humans. *International Journal of Computer Vision*, *122*(3), 411–425.

Stability-Flexibility Dilemma in Cognitive Control: A Dynamical System Perspective

Sebastian Musslick[1,*], **Anastasia Bizyaeva**[2], **Shamay Agaron**[1], **Naomi Leonard**[2], **and Jonathan D. Cohen**[1]

[1]Princeton Neuroscience Institute, Princeton University, Princeton, NJ 08544, USA.
[2]Department of Mechanical and Aerospace Engineering, Princeton University, Princeton, NJ 08544, USA.

[*]Corresponding Author: *musslick@princeton.edu*

Abstract

Constraints on control-dependent processing have become a fundamental concept in general theories of cognition that explain human behavior in terms of rational adaptations to these constraints. However, theories miss a rationale for why such constraints would exist in the first place. Recent work suggests that constraints on the allocation of control facilitate flexible task switching at the expense of the stability needed to support goal-directed behavior in face of distraction. Here, we formulate this problem in a dynamical system, in which control signals are represented as attractors and in which constraints on control allocation limit the depth of these attractors. We derive formal expressions of the stability-flexibility tradeoff, showing that constraints on control allocation improve cognitive flexibility but impair cognitive stability. Finally, we provide evidence that human participants adapt higher constraints on the allocation of control as the demand for flexibility increases but that participants deviate from optimal constraints.

Keywords: cognitive control; task switching; stability-flexibility tradeoff; bounded rationality; capacity constraints

Introduction

Numerous theories of cognition are grounded in the assumption that there are fundamental constraints on the allocation of cognitive control (Anderson, 2013; Kurzban, Duckworth, Kable, & Myers, 2013; Shenhav, Botvinick, & Cohen, 2013). Theories that assume such limitations have been successful in explaining how humans rationally allocate control under such constraints (Lieder, Shenhav, Musslick, & Griffiths, 2018; Musslick, Shenhav, Botvinick, & Cohen, 2015; Shenhav et al., 2013). However, they do not provide a rationale for *why* such limitations would exist in the first place.

A recent line of work attempts to explain the limitations of control allocation in terms of fundamental computational dilemmas in neural processing systems. For instance, Musslick et al. (2017) suggest that neural architectures are subject to a tradeoff between learning efficiency that is promoted through the use of shared task representations (Bengio, Courville, & Vincent, 2013; Caruana, 1997), on the one hand, and multitasking capability that is achieved through the separation of task representations, on the other hand (Allport, 1980; Musslick et al., 2016; Meyer & Kieras, 1997; Navon & Gopher, 1979; Salvucci & Taatgen, 2008; Feng, Schwemmer, Gershman, & Cohen, 2014). From this perspective, limitations in multitasking may reflect a preference of the neural system to learn tasks more quickly (Musslick et al., 2017; Sagiv, Musslick, Niv, & Cohen, 2018).

One way to circumvent limitations in concurrent multitasking is to execute multiple tasks in series, through flexible switching between tasks (Salvucci, Taatgen, & Borst, 2009; Fischer & Plessow, 2015). The serial execution of tasks, however, gives rise to another tradeoff known as the stability-flexibility dilemma: allocating more control to a task results in greater activation of its neural representation but also in greater persistence of this activity upon switching to a new task, yielding switch costs (Ueltzhöffer, Armbruster-Genç, & Fiebach, 2015; Goschke, 2000). By considering the problem in terms of the parameterization of a nonlinear dynamical system, in which control signals are represented as attractors, Musslick, Jang Jun, Shvartsman, Shenhav, and Cohen (2018) showed that constraints on control allocation can promote cognitive flexibility at the expense of cognitive stability. Their simulations suggest that higher constraints on control allocation are optimal in environments with higher demand for task switches. While the simulations provide a computational rationale for constraints on control, a formal analysis of the problem is lacking. It also remains to be tested whether humans adapt their constraints on control in response to demands for flexibility.

In this work, we analyze the model by Musslick et al. (2018) from a dynamical system perspective and derive formal definitions for cognitive stability and cognitive flexibility. We then prove that higher gains of a network's activation function (equivalent to inverse temperature, and thought to reflect the effects of neuromodulatory neurotransmitters such as dopamine and norepinephrine; Servan-Schreiber, Printz, & Cohen, 1990; Liljenström, 2003; Cools, 2015) can balance this tradeoff towards cognitive stability at the cost of cognitive flexibility. To assess whether human participants adjust their constraints on control as a function of flexibility demands, we fit the model to participants who performed a task switching experiment with different rates of switching. We specifically test the hypothesis that the behavior of participants in highly flexible environments can be best described by a lower gain, reflecting higher constraints on control allocation. Finally, we use computational simulations to investigate whether participants adapt to the stability-flexibility dilemma in a rational manner, by comparing fitted constraints on control against optimal constraints on control.

Recurrent Neural Network Model

We analyze the stability-flexbility tradeoff in a recurrent neural network model described by Musslick et al. (2018). The model consists of a control module that simulates control configurations as activities of processing units. The pattern of activity associated with each control configuration evolves in an attractor landscape over the course of trials. Within each trial, the processing units bias an evidence accumulation process in the decision module that integrates information about the stimulus and generates a response.

Control Module

We simulate the amount of control allocated to a task as the activity of a corresponding processing unit in a recurrent neural network. Here, we consider environments with two tasks, and therefore two processing units, indexed by $i, j \in \{1,2\}$. The activity of each unit is determined by its net input

$$net_i(t) = w_{i,i} act_i(t) + w_{i,j} act_j(t) + I_i \tag{1}$$

which is a linear combination of the unit's own activity $act_i(t)$ multiplied by the self-recurrent weight $w_{i,i}$, the activity $act_j(t)$ of the other unit $j \in 1,2, j \neq i$, multiplied by an inhibitory weight $w_{i,j}$, and an external input I_i (i.e., an "instruction") provided to the unit (see Figure 1A). The activities for both processing units evolve across trials according to

$$\frac{dact_i(t)}{dt} = -act_i(t) + \frac{1}{1+e^{-g \cdot net_i(t)}} \tag{2}$$

where the g is the slope of a sigmoid activation function[1]. The sigmoid activation function constrains the activity of both units to lie between 0 and 1. The gain of the activation function g regulates the distance between the two control attractors, with lower gain leading to a lower activation of the currently relevant control unit and slightly higher activation of its competitor (see Figure 1B-C). From this perspective, lower gains impose higher constraints on the amount of control that can be allocated to a task but facilitate switches between tasks. Below, we provide a formal analysis of the stability-flexibility dilemma as a function of gain.

Decision Module

We simulate the decision process using the drift diffusion model (DDM, Ratcliff, 1978). On each trial, the decision module integrates information along two stimulus dimensions S_1 and S_2 of a single stimulus to determine a response. Each dimension (e.g., color or motion of a moving dot stimulus) can take one of two values (e.g., red or blue; up or down), each of which is associated with one of two responses (e.g. pressing left or right button). Each of the two tasks requires mapping the current value of one of the two stimulus dimensions to its corresponding response, while ignoring the other dimension. Since both tasks involve the same

[1]The non-linear dynamical system presented in this work is formally equivalent to the discrete time model by Musslick et al. (2018) for a rate constant of 1.

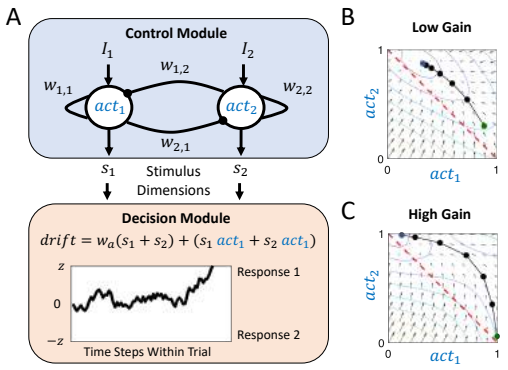


Figure 1: Model architecture. (A) The dynamics of the control module (blue) unfold over the course of trials and are determined by external input signals I_1, I_2, recurrent connectivity $w_{1,1}, w_{2,2}$ for each unit, as well as mutual inhibition $w_{1,2}, w_{2,1}$ between units. The activity of each control unit biases the processing of a corresponding stimulus dimension on a given trial. On each trial, the decision module accumulates evidence for both stimulus dimensions towards one of two responses until a threshold is reached. (B-C) Activation trajectories for models with a (B) low and (C) high gain are shown as a series of connected black dots, evolving from the control attractor for task 1 (green) to the control attractor for task 2 (blue). Contour lines and arrows indicate the energy and shape of the attractor landscape after a task switch from task 1 to task 2. Attractors for both tasks lie approximately on the antidiagonal of the state space (act_{dif}) shown in red.

pair of responses, stimuli can be congruent (stimulus values in both dimensions associated with the same response) or incongruent (associated with different responses). The drift of the DDM integration process is determined by the combined stimulus information from each dimension, weighted by input received from the control module (as described below), and evidence is accumulated over time until one of two response thresholds is reached. The drift rate is decomposed into an automatic and controlled component:

$$drift = \underbrace{w_a(S_1 + S_2)}_{\text{automatic}} + \underbrace{act_1 S_1 + act_2 S_2}_{\text{controlled}} \tag{3}$$

where the automatic component is weighted by w_a and reflects automatic processing of each stimulus dimension that is unaffected by control. The absolute magnitude of S_1, S_2 depends on the strength of the association of each stimulus with a given response and its sign depends on the response (e.g. $S_1 < 0$ if the associated response is to press the left button, $S_1 > 0$ if the associated response is to press the right button). Thus, for congruent trials S_1 and S_2 have the same sign, and the opposite sign for incongruent (conflict) trials. The controlled component of the drift rate is the sum of the two stimulus values, each weighted by the activation of the corresponding control unit. Thus, each unit in the control module biases processing towards one of the stimulus dimensions. As a result, progressively greater activation of a control

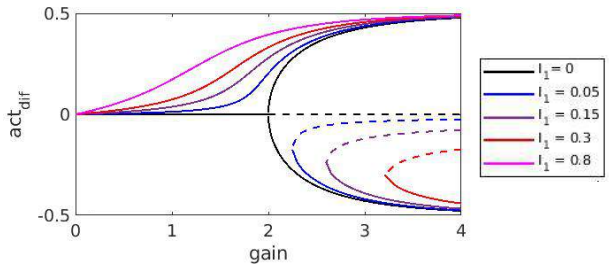


Figure 2: Difference in the activation of the two processing units (act_{dif}) for various magnitudes of I_1 in steady-state solutions to Equation (2), with $I_2 = 0$. Solid lines are stable attractors and dashed lines are unstable solutions. The black curve is the symmetric no-input system of Equation (6), for which with low values of gain the only attractor is the neutral state $act_{dif} = 0$. For values of gain greater than 2, two nonzero attractors emerge in a pitchfork bifurcation. Nonzero input to one of the processing units breaks the symmetry, splitting up the symmetric pitchfork into a continuous branch for the corresponding task and a cusp. The breakup is referred to as imperfect bifurcation (Golubitsky & Schaeffer, 1985).

unit improves performance – speeds responses and improves accuracy – for the corresponding task. Distributions of reaction times (RTs) and error rates for a given parameterization of drift rate at a given trial are derived from an analytical solution to the DDM (Navarro & Fuss, 2009).

Formal Analysis

Previous simulation work suggests that lower values of gain facilitate switches between tasks but limit how much control can be allocated to any given task (Musslick et al., 2018). Building on work by Franci, Srivastava, and Leonard (2015); Gray, Franci, Srivastava, and Leonard (2018), we derive a formal analysis of this tradeoff as a function of gain.

For unit weights $w_{1,2} = w_{2,1} = -1$ and $w_{1,1} = w_{2,2} = 1$, the attractors for both tasks are observed to lie near the antidiagonal in the activation space (see red dashed line in Figure 1B-C). We examine the dynamics of the system in a rotated frame of reference such that the attractors lie near the vertical axis. We introduce translated and rotated variables

$$\begin{pmatrix} act_{avg} \\ act_{dif} \end{pmatrix} = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}\begin{pmatrix} act_1 - 1/2 \\ act_2 - 1/2 \end{pmatrix} \tag{4}$$

where act_{avg} corresponds to the average of the two (shifted) activity states of the processing units, and act_{dif} is the average difference between the two activity states. Here, act_{dif} can be considered a proxy for cognitive stability, indexing how much control is allocated to one task versus the other. We can get an intuition for the dynamics of the system by first considering the symmetric case, in which the control module receives no input to either task processing unit $I_1 = I_2 = 0$.

The dynamical equations (2) with zero input decouple in

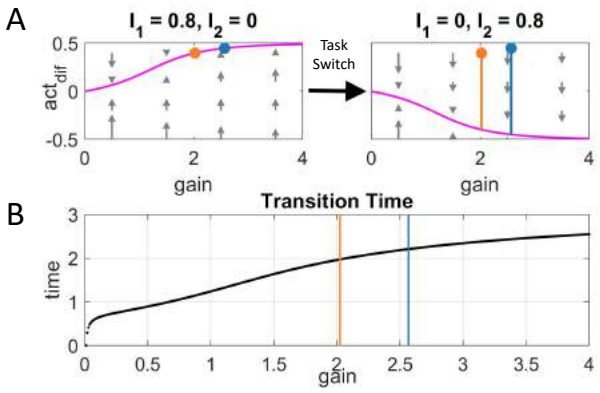


Figure 3: Relationship between act_{dif}, convergence time and gain. (A) Configuration of the system before and after a task switch. (B) Convergence time as a function of gain. Vertical lines mark examples for gain parameters that were fitted to participants' performance in environments with low (blue) and high (orange) rates of task switches.

the new variables:

$$\frac{d}{dt}act_{avg} = -act_{avg} \tag{5}$$

$$\frac{d}{dt}act_{dif} = -act_{dif} + \frac{1}{2}\tanh(g \cdot act_{dif}) \tag{6}$$

The attractors of the system are the stable steady-state solutions of (5), (6). Note that act_{avg} decays to zero and the no-input system always settles on the antidiagonal $act_1 + act_2 = 1$. According to the dynamics of act_{dif}, the available attractors vary with the value of the gain parameter (Figure 2).

With nonzero input, the dynamics on the diagonal and antidiagonal directions do not completely decouple. The contribution in the act_{avg} direction results in the system settling near the antidiagonal with a small offset rather than directly on the antidiagonal (Figure 1B-C). However, analogously to the symmetric no-input case, the dominant dynamical behavior is in the act_{dif} direction, shown in Figure 2. From this we can recover intuition for the tradeoff between cognitive stability and flexibility. The relationship between network gain g on the relevant domain ($0 < act_{dif} < 0.5$ when $I_1 \neq 0$ or $-0.5 < act_{dif} < 0$ when $I_2 \neq 0$) and act_{dif} is defined by

$$g = f(act_{dif}) = \frac{\tanh^{-1}(2act_{dif})}{act_{dif}} + E(|act_{dif}|, I) \tag{7}$$

where the first term is the explicit solution for steady state gain of the no-input system (6) and $E(act_{dif})$ is the deviation from the symmetric case, which is a monotonically decaying function of the magnitude of act_{dif}. We can approximate this deviation with a decaying exponential fit

$$E(|act_{dif}|, I) \approx 1.4e^{-5I^{-1.1}\cdot|act_{dif}|} + 0.6 \tag{8}$$

where I is the magnitude of input. Since (7) is locally invertible on the given domain, we can express cognitive stability as a function of the network gain, $act_{dif}(g) = f^{-1}(act_{dif})$.

Further, we can express cognitive flexibility in terms of the time it takes to switch from one task to another, that is, the time it takes for act_{dif} to pass through zero and switch sign. From simulation we observe that the transition time is a monotonically increasing function of the network gain (see Figure 3). For an input $I_j = 0.8$ we approximate the transition time with a linear fit $T(g) \approx 0.8g + 0.6$. Substituting (7) with $I = 0.8$ for g, we obtain an expression for the stability-flexibility tradeoff

$$T(act_{dif}) \approx 0.8 \frac{\tanh^{-1}(2act_{dif})}{act_{dif}} + 1.1e^{-6.4|act_{dif}|} + 1.1 \quad (9)$$

by relating convergence time and act_{dif}. This analysis supports intuitions from prior computational work, showing that a higher network gain promotes cognitive stability at the expense of cognitive flexibility (Musslick et al., 2018). Moreover, the formal results described in this section offer a quantitative interpretation of network gain in terms of both act_{dif}, as well as T, when fitting the model to human behavior.

Experiment

Our analysis results suggest that a system should adapt higher constraints on control (lower gains) if the demand for cognitive flexibility increases. To examine whether human participants rationally adapt constraints on control to the flexibility demands of their environment, we conducted a task switching experiment in which the rate of task switches was varied across participants. We then fit the network model to each participant and evaluated the fitted gain against the gain that optimizes the stability-flexibility tradeoff for each participant.

Method

Participants. We recruited 67 participants from Amazon Turk. All participants signed a consent form prior to participation and received $6 US for participation. The study was approved by the Institutional Review Board of Princeton University. We only included participants with an accuracy above 65% into our analysis, yielding a total of 31 participants in the low switch rate group and 27 participants in the high switch rate group.

Apparatus and Stimuli. Stimuli consisted of a web-based random-dot kinematogram (RDK) that we adapted from Rajananda, Lau, and Odegaard (2018). The RDK contained blue and red moving dots, some of which consistently moved in either an upward or a downward direction, and some of which moved in a random direction.

Task and Procedure. Participants switched between a color task, in which they had to indicate the color of the majority of the presented dots (red or blue), using the response buttons 'A' and 'L', respectively, and a motion task in which they had to indicate the direction of coherent motion (up or down), using the same response buttons 'A' and 'L', respectively. Participants performed each task over a mini-block of four to six trials. Each mini-block was preceded by a task cue (one of two cues for each task to control for cue repetition effects) that instructed participants which tasks to perform. In some mini-blocks, participants had to repeat the task that they performed in the previous mini-block (task repetition), whereas in other mini-blocks, they had to switch to the other task (task switch). The cue was displayed for 700ms and disappeared for another 600ms. On each trial of a miniblock, the RDK stimulus was shown for 1500ms, followed by an inter-trial interval of 700ms. Participants were asked to indicate the task-relevant response while the stimulus was on the screen. In the beginning of the experiment, we used a staircasing procedure to identify coherence levels (i.e. the percent of dots having the same motion or color) for each participant that standardized performance at around 85% accuracy for both tasks. After training participants to associate the task cues with each task, participants switched between tasks over a sequence of two larger blocks of 66 miniblocks each.

Design. Participants were divided into two experimental groups, one that switched tasks between mini-blocks 25% of the time (low switch rate) and one that switched tasks 75% of the time (high switch rate). For each task switching sequence, we counterbalanced seven factors with respect to the first trial of each mini-block: task (color or motion task), task transition (task switch or task repetition), task cue (first or second cue associated with a task), congruency (congruent or incongruent), dot motion (upward or downward), color (mostly blue or red) and correct response ('A' or 'L' key).

Data Analysis. We focused our analysis on the second block of the experiment, assuming that subjects take the first block to adjust to the frequency manipulation of the experiment. We were specifically interested in the performance costs associated with task switches. Prior work suggests that switch costs diminish after the first trial of a mini-block (Rogers & Monsell, 1995). We therefore analyzed reaction times (RTs) and error rates associated with the first trial of a miniblock. Furthermore, RT data was limited to correct trials that were preceded by at least one correct trial. For each group of participants, we computed switch costs as the difference in performance between switch trials and repetition trials for both RTs and error rates. We also computed incongruency costs on task repetitions[2] as the difference in performance between congruent and incongruent trials. Finally, we conducted two-tailed t-tests to assess whether participants in the low switch rate group exhibited different switch costs and different incongruency costs compared to the high switch rate group.

Model Fitting Procedure. Before fitting parameters of the model to behavior of human participants, we evaluated how well we can recover these parameters from simulated behavior generated by the model. Motivated by the formal analysis described above, we parameterized the control module with balanced recurrent and inhibitory weights, $w_{i,i} = 1, w_{i,j} = -1$,

[2]Incongruency costs have been shown to interact with task transition (Rogers & Monsell, 1995; Goschke, 2000; Wendt & Kiesel, 2008). To avoid confounding effects of congruency with the frequency of task switches we conditioned incongruency costs on task repetition trials.

Table 1: Fitted model parameters with prior distributions.

Parameter	Prior Distribution	Lower Bound	Upper Bound
g	$Gamma(2.5, 0.75)$	0	4
z	$Gamma(3, 0.02)$	0.01	0.25
c	$Gamma(3, 0.75)$	0.015	0.25
h	$Beta(1.2, 1.2)$	0	1
w_a	$Gamma(16, 0.05)$	0.1	0.5

and computed the activities of both processing units trial-by-trial, by numerically integrating Equation (2) with step size h. We set the input for the currently relevant task unit to $I_i = 0.8$ and the input for the task-irrelevant unit to $I_{j \neq i} = 0$. The stimulus dimension encoding the color feature was set to $S_1 = 0.1$ if the majority of the dots was red and set to $S_1 = -0.1$ if the majority of the dots was blue. Similarly, the stimulus dimension encoding the motion feature was set to $S_2 = 0.1$ if the dots were moving upward and set to $S_2 = -0.1$ if dots were moving downward. We fixed the non-decision time of the DDM to $T_0 = 0.2$ and fit five free parameters with priors shown in Table 1: network gain g, DDM response threshold z, DDM noise c, integration constant h and automaticity weight w_a. The number of free parameters was determined based on prior analyses of parameter identifiability, indicating that larger or different sets of free parameters may not be reliably recovered. To assess how well the five parameters can be recovered from the simulated behavior of the model, we first sampled 10 parameter configurations uniformly from the intervals shown in Table 1. We then generated distributions of response times for each trial of the second experiment block and identified parameters that maximized the likelihood of the model's responses given the data. The identifiability of each parameter was quantified by regressing the true parameter against the fitted parameter across all sampled parameter configurations. We used the same procedure to fit the model to each participant. Finally, we conducted a one-tailed t-test to assess whether fitted gain parameters of the participants in the low switch rate group were higher relative to fitted gain parameters in the high switch rate group.

Optimality Analysis. To evaluate whether participants adapt rationally to the stability-flexibility dilemma, we identified the optimal gain that maximizes accuracy across all trials in the experiment, given all other fitted parameters for a given participant[3]. For each participant group, we computed the difference between fitted gains and optimal gains, and performed a two-sided t-test to evaluate whether fitted gain parameters systematically deviate from their optimal gain.

Results

We found that participants who switched tasks less frequently took more time to switch tasks, $t(56) = 2.04, p < 0.05$, but found no significant differences in terms of error rates, $t(56) = 0.20, p = 0.84$. Participants showed no significant differences in incongruency costs between the two experimental groups in terms of both RTs, $t(56) = 0.93, p = 0.35$ and error rates, $t(56) = 1.30, p = 0.20$.

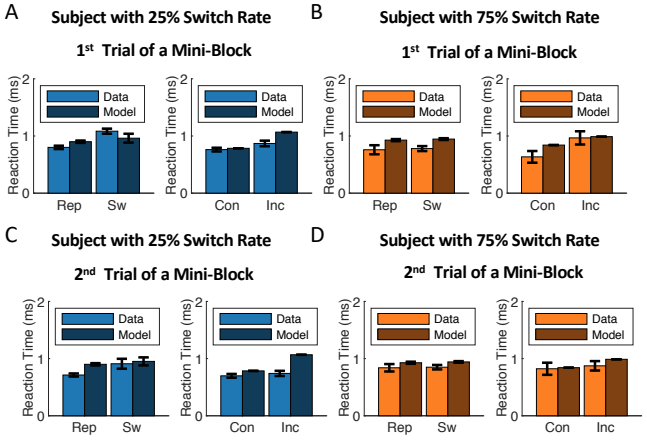


Figure 4: Examples of participant RTs and RTs generated by the fitted model. RTs are shown as a function of task transition (repetition, switch) and response congruency (congruent, incongruent) for the first (A, B) and second (C, D) trial of a mini-block. Data is shown for one participant from the low switch rate group (A, C) and one participant from the high switch rate group (B, D). Dark bars indicate average RTs generated by the fitted model. Error bars indicate the standard error of the mean across trials.

Overall, we were able to recover parameters from behavior generated by the model. The true parameter value significantly predicted the value estimated by the fitting procedure for network gain, $b = 0.97$, $t(9) = 4.44$, $p < 0.01$, DDM response threshold z, $b = 1.09$, $t(9) = 14.06$, $p < 0.001$, DDM noise c, $b = 0.69$, $t(9) = 9.03$, $p < 0.001$, integration constant h, $b = 0.87$, $t(9) = 5.02$, $p < 0.01$, and automaticity weight w_a, $b = 0.64$, $t(9) = 3.06$, $p < 0.05$. Figure 4 shows the behavior of two participants along with the behavior generated by the fitted model. In line with the prediction made by the model, we observed that the fitted gain parameters to behavior of human participants were significantly higher in the low switch rate group relative to the high switch rate group, $t(56) = 3.61, p < 0.001$ (Figure 5A). Note that Figure 3 depicts formal expressions of cognitive stability (Figure 3A) and cognitive flexbibility (Figure 3B) as a function of the average fitted gains for both groups. Interestingly, the fitted gains were significantly lower than the optimal gains, for both groups: low switch rates, $t(30) = 7.40, p < 0.001$, and high switch rates, $t(26) = 4.24, p < 0.001$, suggesting that, while participants adapt gain in the predicted way, overall they exert more constraint on control allocation (lower gain) than was predicted to be optimal.

General Discussion and Conclusion

A fundamental characteristic of control-dependent processing are constraints on the allocation of control (Shiffrin & Schneider, 1977; Posner & Snyder, 1975). Recent work sug-

[3]We chose to maximize accuracy over maximizing reward rate as the duration of each trial was independent of response time. However, we obtained identical results when optimizing for reward rate.

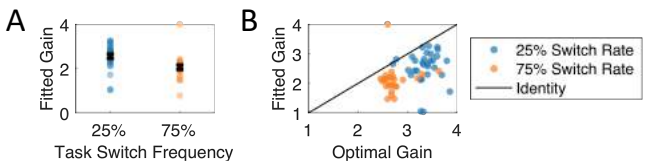


Figure 5: Model fitting results. (A) Fitted gains are shown for participants with a low (blue) and high (orange) switch rates. Each circle corresponds to the fitted gain of a participant. Vertical lines indicate the standard error of the mean fitted gain, centered around the mean. Mean gains for each group are also shown in Figure 3. (B) Fitted gain for each participant is plotted against the optimal gain that maximizes the overall accuracy of the model for a given participant.

gests that these limitations may origin from shared representations (Feng et al., 2014; Musslick et al., 2016; Salvucci & Taatgen, 2008), as well as persistence characteristics in neural systems (Musslick et al., 2018), and the resulting need to trade off the amount of control that can be allocated to a single task against the time required to switch from one task to another. In this work, we introduced a formal analysis of the latter — that is, the tradeoff between cognitive stability and cognitive flexibility.

Applying perturbation theory to the network model described by Musslick et al. (2018), we formally defined cognitive stability in terms of the distance between attractors for competing control states, and defined cognitive flexibility in terms of the time to converge from one control attractor to the other. We showed that the two measures trade off against each other, and that the balance of this tradeoff is determined by the gain of the network's activation function. We then examined whether human participants balance this tradeoff in a similar manner as a function of the demand for flexibility, by fitting the model to participants who were required to switch tasks at either a low or high frequency. We observed that participants who switched more frequently showed lower switch costs, suggesting that they became more cognitively flexible. Moreover, model fits showed that this could be explained by lower gain, and with it, higher constraints on control. Interestingly, fitted gains for all participants were lower compared to the gains that optimized accuracy in the face of the stability-flexibility tradeoff. This suggests that there may be other factors that limit control allocation.

Altogether, our analytic and empirical results provide a rationale for how participants should adapt to different demands for flexibility given a mechanistic model for how control is represented and allocated in a recurrent neural network model. A formal relationship between cognitive stability and cognitive flexibility may not only help interpret human behavior in terms of model fits but may also help identify neural correlates for both measures. For instance, the dynamics of steady-state visually evoked potentials (SSVEP) — used to index feature-specific attention (Müller et al., 2006) — may be characterized in terms of the evolving distance between attractors of competing attentional states. Finally, the behavioral results replicate earlier work, showing that participants' switch costs decrease as task switches become more frequent (Mayr, 2006; Monsell & Mizon, 2006). Furthermore, prior work suggests that participants trade off cognitive flexibility against higher incongruency costs in voluntary task switching scenarios when task switches are associated with a higher reward than task repetitions (Braem, 2017).

One interesting puzzle concerns the learning mechanisms that underlie rational adaptations to changing demands in cognitive flexibility. A computationally cheap, but inflexible approach is to learn the amount of control that should be exerted through model-free reinforcement (Lieder et al., 2018). Alternatively, humans may approximate the optimal tradeoff, by attaching a cost to the amount of control that can be allocated. From this perspective, the stability-flexibility tradeoff may provide a normative rationale for parameterizing the cost of cognitive control that is integral to recent theories of control allocation (Shenhav et al., 2013, 2017).

References

Allport, D. A. (1980). Attention and performance. *Cognitive psychology: New Directions*, *1*, 12–153.

Anderson, J. R. (2013). *The architecture of cognition*. Psychology Press.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(8), 1798–1828.

Braem, S. (2017). Conditioning task switching behavior. *Cognition*, *166*, 272–276.

Caruana, R. (1997). Multitask learning. *Machine Learning*, *28*(1), 41–75.

Cools, R. (2015). The cost of dopamine for dynamic cognitive control. *Current Opinion in Behavioral Sciences*, *4*, 152–159.

Feng, S. F., Schwemmer, M., Gershman, S. J., & Cohen, J. D. (2014). Multitasking vs. multiplexing: toward a normative account of limitations in the simultaneous execution of control-demanding behaviors. *Cognitive, Affective, & Behavioral Neuroscience*, *14*(1), 129-146.

Fischer, R., & Plessow, F. (2015). Efficient multitasking: parallel versus serial processing of multiple tasks. *Frontiers in psychology*, *6*, 1366.

Franci, A., Srivastava, V., & Leonard, N. (2015). A realization theory for bio-inspired collective decision-making. *arXiv:1503.08526*.

Golubitsky, & Schaeffer. (1985). *Singularities and groups in bifurcation theory*. Springer-Verlag New York.

Goschke, T. (2000). Intentional reconfiguration and involuntary persistence in task set switching. *Control of Cognitive Processes: Attention and Performance XVIII*, *18*, 331.

Gray, R., Franci, A., Srivastava, V., & Leonard, N. (2018). Multiagent decision-making dynamics inspired by honey-

bees. *IEEE Transactions on Control of Network Systems*, *5*(2), 793-806.

Kurzban, R., Duckworth, A., Kable, J. W., & Myers, J. (2013). An opportunity cost model of subjective effort and task performance. *Behav. Brain Sci*, *36*(6), 661–679.

Lieder, F., Shenhav, A., Musslick, S., & Griffiths, T. L. (2018). Rational metareasoning and the plasticity of cognitive control. *PLoS Comput. Biol.*, *14*(4), e1006043.

Liljenström, H. (2003). Neural stability and flexibility: a computational approach. *Neuropsychopharmacology*, *28*(S1), S64.

Mayr, U. (2006). What matters in the cued task-switching paradigm: Tasks or cues? *Psychonomic Bulletin & Review*, *13*(5), 794–799.

Meyer, D. E., & Kieras, D. E. (1997). A computational theory of executive cognitive processes and multiple-task performance: Part I. Basic mechanisms. *Psychol. Rev.*, *104*(1), 3.

Monsell, S., & Mizon, G. A. (2006). Can the task-cuing paradigm measure an endogenous task-set reconfiguration process? *Journal of Experimental Psychology: Human Perception and Performance*, *32*(3), 493.

Müller, M., Andersen, S., Trujillo, N., Valdes-Sosa, P., Malinowski, P., & Hillyard, S. (2006). Feature-selective attention enhances color signals in early visual areas of the human brain. *Proceedings of the National Academy of Sciences*, *103*(38), 14250–14254.

Musslick, S., Dey, B., Özcimder, K., Patwary, M. M. A., Willke, T. L., & Cohen, J. D. (2016). Controlled vs. automatic processing: A graph-theoretic approach to the analysis of serial vs. parallel processing in neural network architectures. In *Proceedings of the 38th annual conference of the Cognitive Science Society* (pp. 1547–1552). Philadelphia, PA.

Musslick, S., Jang Jun, S., Shvartsman, M., Shenhav, A., & Cohen, J. D. (2018). Constraints associated with cognitive control and the stability-flexibility dilemma. In *Proceedings of the 40th Annual Meeting of the Cognitive Science Society* (pp. 806–811). Madison, WI.

Musslick, S., Saxe, A., Özcimder, K., Dey, B., Henselman, G., & Cohen, J. D. (2017). Multitasking capability versus learning efficiency in neural network architectures. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society* (pp. 829–834). London, UK.

Musslick, S., Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2015). A computational model of control allocation based on the expected value of control. In *The 2nd Multidisciplinary Conference on Reinforcement Learning and Decision Making*. Edmonton, Can.

Navarro, D. J., & Fuss, I. G. (2009). Fast and accurate calculations for first-passage times in wiener diffusion models. *Journal of Mathematical Psychology*, *53*(4), 222–230.

Navon, D., & Gopher, D. (1979). On the economy of the human-processing system. *Psychol. Rev.*, *86*(3), 214.

Posner, M., & Snyder, C. (1975). attention and cognitive control,. In *Information processing and cognition: The loyola symposium* (pp. 55–85).

Rajananda, S., Lau, H., & Odegaard, B. (2018). A random-dot kinematogram for web-based vision research. *Journal of Open Research Software*, *6*(1).

Ratcliff, R. (1978). A theory of memory retrieval. *Psychol. Rev.*, *85*(2), 59.

Rogers, R. D., & Monsell, S. (1995). Costs of a predictible switch between simple cognitive tasks. *Journal of Experimental Psychology: General*, *124*(2), 207.

Sagiv, Y., Musslick, S., Niv, Y., & Cohen, J. D. (2018). Efficiency of learning vs. processing: Towards a normative theory of multitasking. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 1004–1009). Madison, WI.

Salvucci, D. D., & Taatgen, N. A. (2008). Threaded cognition: an integrated theory of concurrent multitasking. *Psychol. Rev.*, *115*(1), 101.

Salvucci, D. D., Taatgen, N. A., & Borst, J. P. (2009). Toward a unified theory of the multitasking continuum: From concurrent performance to task switching, interruption, and resumption. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 1819–1828).

Servan-Schreiber, D., Printz, H., & Cohen, J. D. (1990). A network model of catecholamine effects: gain, signal-to-noise ratio, and behavior. *Science*, *249*(4971), 892–895.

Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2013). The expected value of control: an integrative theory of anterior cingulate cortex function. *Neuron*, *79*(2), 217–240.

Shenhav, A., Musslick, S., Lieder, F., Kool, W., Griffiths, T. L., Cohen, J. D., & Botvinick, M. M. (2017). Toward a rational and mechanistic account of mental effort. *Annual Review of Neuroscience*, 99–124.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychol. Rev.*, *84*(2), 127.

Ueltzhöffer, K., Armbruster-Genç, D. J., & Fiebach, C. J. (2015). Stochastic dynamics underlying cognitive stability and flexibility. *PLoS Comput. Biol.*, *11*(6), e1004331.

Wendt, M., & Kiesel, A. (2008). The impact of stimulus-specific practice and task instructions on response congruency effects between tasks. *Psychological Research*, *72*(4), 425–432.

Decomposing Individual Differences in Cognitive Control: A Model-Based Approach

Sebastian Musslick[1,*], **Jonathan D. Cohen**[1], **and Amitai Shenhav**[2]

[1]Princeton Neuroscience Institute, Princeton University, Princeton, NJ 08544, USA.
[2]Department of Cognitive, Linguistic, and Psychological Sciences,
Brown Institute for Brain Science, Brown University, Providence, RI 02912, USA.

[*]Corresponding Author: *musslick@princeton.edu*

Abstract

Researchers have long been interested in using laboratory measures of cognitive control to predict a person's cognitive control/self control success outside the lab. We used a computational approach to identify which lab-based performance measures provide the most valid individual difference measures of one's ability and/or motivation to exert cognitive control. We simulated performance across an array of cognitive control tasks, and estimated the degree to which different performance metrics (e.g., congruency effects, conflict adaptation, and demand avoidance) could theoretically provide valid estimates of processes underlying control allocation. By performing dimension reduction on these performance metrics, we further revealed latent dimensions that can index separate mechanisms of control-demanding behavior. Our results suggest that individual differences in measures of cognitive control can originate from multiple factors, several of which are unrelated to *capacity* for cognitive control. We conclude by discussing implications of these analyses for assessing individual differences in cognitive control phenomena.

Keywords: individual differences; cognitive control; motivation; self-control

Introduction

Cognitive control refers to our ability to adapt mental processes to current task goals. Researchers have developed a variety of measures to index a given person's capacity to exert cognitive control, such as conflict-related interference, conflict adaption, and performance costs associated with task switching. It has often been assumed that individual differences in capacity and/or motivation for control should predict one's self-control success in the real world, and that performance on one cognitive control task should therefore correlate with indices of self-control. Unfortunately, however, such correlations have been inconsistent across the literature. For instance, whereas some individual studies find correlations between Stroop conflict-related interference (congruency costs) and real-world self-control outcomes (e.g., addiction treatment compliance, healthy diets; Streeter et al., 2008; Allan, Johnston, & Campbell, 2010), a large study (N=2,641) recently found no correlation between congruency costs and a well-validated index of real-world self-control (Saunders, Milyavskaya, Etz, Randles, & Inzlicht, 2017). The inconsistency in these findings has been taken to suggest that control mechanisms are highly context-specific and/or that self-control may not actually require cognitive control (Berkman, Hutcherson, Livingston, Kahn, & Inzlicht, 2017). Here we explore an alternative interpretation, that commonly used measures of control allocation may be ill-suited to indexing the control required of those tasks.

Converging evidence suggests that performance on cognitively demanding tasks reflects a combination of bottom-up stimulus processing and one's capacity and motivation to exert top-down control over such processing (Cohen, Dunbar, & McClelland, 1990; Shenhav, Botvinick, & Cohen, 2013; Shenhav et al., 2017). These insights have been integrated into a recent computational model of control allocation, which simulates an agent's performance on a cognitive task based on the parameters of that task (e.g., stimulus salience) and the incentives on offer (e.g., reward for correct response; Musslick, Shenhav, Botvinick, & Cohen, 2015). Control allocation is determined by comparing the expected reward (based on the incentives and the degree to which control increases the likelihood of a correct response) with an intrinsic cost of control, to determine the overall Expected Value of Control (EVC). The parameters of these simulated agents can be adjusted to vary how they process stimuli and incentives (e.g., their sensitivity to rewards), resulting in attendant changes to task performance. Theoretical analyses suggest that between-subject variability in some motivational parameters, such as reward sensitivity, can generally limit the ability to recover other motivational parameters, such as the cost of cognitive control, from task performance (Musslick, Cohen, & Shenhav, 2018; Caplin, Csaba, Leahy, & Nov, 2018). An important question that remains unaddressed, however, is whether individual differences in cognitive control phenomena provide a reliable index for one's capacity to exert cognitive control.

Here, we use the EVC model to simulate various phenomena that have been used to index one's capacity to exert cognitive control, including within-trial interference and cross-trial adaptation to response conflict; task-switching costs; and cognitive effort discounting. We then demonstrate that individual differences in these phenomena are influenced by parameters of the task and the agent, including variables related to bottom-up stimulus processing, the *ability* to exert control, and the *motivation* for doing so. Finally, we identify latent dimensions that explain individual differences across these simulated phenomena, and discuss implications of this work for the assessment of individual differences in cognitive control within and outside of the lab.

Expected Value of Control Model

The EVC theory is based on the premise that control allocation involves specifying the identity of candidate control signals, as well as the intensity of each (Shenhav et al., 2013). Increases in control signal intensity lead to improvements in performance on the corresponding task. However, it is also assumed that exercising cognitive control is costly and this cost increases monotonically with the intensity of the control signal. According to the EVC theory, the control system chooses to implement the configuration of control signals that yields the highest expected value of control, that is, the expected utility of implementing a configuration of control signals with specified intensities minus their associated costs. Critically, the expected value for each candidate control signal configuration is contingent on an internal model of the task environment that is updated based on experience.

The present implementation of the EVC model describes performance in the Stroop task (e.g., responding to the ink color of a color word, Stroop, 1935), in terms of an interaction between the control system and the task environment. The control signal is chosen optimally based on an internal model of the next trial which produces an estimate of the next trial (inferred state $\hat{\mathbf{S}}$). This signal is then used to interact with the environment (actual state $\mathbf{S}$), for example to commit one of the two possible responses[1] in the task. After each trial, the agent updates the internal model based on an observation of that trial following the response.

In order to generate reaction times (RTs) and responses on each trial, we use the drift diffusion model (DDM Ratcliff, 1978). Within the DDM framework, a response on the task can be conceptualized as a result of the noisy accumulation of evidence toward one of the two possible responses (e.g. one response indicating the color green and the other response indicating the color red; Musslick et al., 2015). Here, we assume that the rate of evidence accumulation toward one of the two responses is governed by a controlled and an automatic component

$$drift = \varepsilon \cdot drift_{\text{control}} + drift_{\text{automatic}} \tag{1}$$

where ε is a capacity parameter that scales the amount of control allocated. The automatic component reflects automatic processing of the color feature and word feature of the stimulus that is unaffected by control,

$$drift_{\text{automatic}} = a_{\text{color}} + a_{\text{word}}. \tag{2}$$

The absolute magnitude of the color-response association a_{color}, as well as the magnitude of the word-response association a_{word} depends on the strength of the association of each stimulus feature with a given response, and its sign depends on the response (e.g. $a_{\text{color}} < 0$ if the response is associated with the left button, $a_{\text{color}} > 0$ if response is associated with the right button). Thus, for congruent trials a_{color}, and a_{word} have the same sign, whereas the opposite sign for incongruent trials. The controlled component of the drift rate is the sum of the two stimulus values, as well as the intensity of the corresponding control signal, one for processing the color dimension of the stimulus u_{color} and one for processing the word dimension of the stimulus u_{word}:

$$drift_{\text{control}} = u_{\text{color}} \cdot a_{\text{color}} + u_{\text{word}} \cdot a_{\text{word}} \tag{3}$$

Thus, each control signal biases processing towards one of the two stimulus dimensions, both of which characterize the actual state on a given trial, $\mathbf{S} = \{a_{\text{color}}, a_{\text{word}}\}$. As a result, higher control signal intensity for processing the color dimension improves performance — speeds responses and lowers error rates — in a trial of the Stroop task. Mean RTs and response probabilities for a given parameterization of drift rate on trial t are derived from an analytical solution to the DDM (Navarro & Fuss, 2009).

In order to specify the optimal set of control signals $\mathbf{U} = \{u_{\text{color}}, u_{\text{word}}\}$ on a given trial t, the model estimates the expected value for each configuration of control signal intensities based on its internal model of the next trial $\hat{\mathbf{S}} = \{\hat{a}_{\text{color}}, \hat{a}_{\text{word}}\}$. This is done by weighting the expected reward for an outcome against the cost associated with the chosen control signal configuration:

$$EVC(\mathbf{U}, \hat{\mathbf{S}}) = P(\text{correct}|\mathbf{U}, \hat{\mathbf{S}})V(R) - Cost(\mathbf{U}) \tag{4}$$

where $P(\text{correct}|\mathbf{U}, \hat{\mathbf{S}})$ corresponds to the probability of reaching the decision threshold for the correct response and $V(R)$ corresponds to the subjective value of responding correctly. Here, the subjective value $V(R) = vR$ corresponds to the amount of reward offered for a correct response R weighted by the model's sensitivity to the reward v. The cost $Cost(\mathbf{U}) = Cost_{\text{impl}}(\mathbf{U}) + Cost_{\text{reconf}}(\mathbf{U})$ is composed of an implementation cost that increases with the amount of control being allocated (Shenhav et al., 2013; Manohar et al., 2015; Lieder, Shenhav, Musslick, & Griffiths, 2018),

$$Cost_{\text{impl}}(\mathbf{U}) = e^{c_{\text{I}} \cdot u_{\text{color}}} + e^{c_{\text{I}} \cdot u_{\text{word}}} \tag{5}$$

as well as a reconfiguration cost that scales with the degree to which control signals need to be changed relative to their previous state (Meiran, 1996; Rogers & Monsell, 1995)

$$Cost_{\text{reconf}}(\mathbf{U}) = e^{c_{\text{R}} \sqrt{(u_{\text{color},t} - u_{\text{color},t-1})^2 + (u_{\text{word},t} - u_{\text{word},t-1})^2}} \tag{6}$$

where the implementation cost is scaled by parameter c_{I} and the reconfiguration cost is scaled by parameter c_{R}. The model selects the control signal configuration with the maximum EVC within the inferred next trial $\hat{\mathbf{S}}$, out of all the configurations under consideration:

$$\mathbf{U}^* = \underset{\mathbf{U}}{\text{argmax}} EVC(\mathbf{U}, \hat{\mathbf{S}}) \tag{7}$$

Performance in the actual state $\mathbf{S}$ is determined by the influence of the chosen control signals on the true parameters

[1] A restriction to two response alternatives limits the scope of the model to paradigms with two-alternative forced choice but makes it amenable to tractable computation of mean reaction times and error rates.

a_{color} and a_{word}. After observing the actual state, the agent updates its inferred state $\hat{\mathbf{S}} = \{\hat{a}_{\text{color}}, \hat{a}_{\text{word}}\}$:

$$\hat{a}_{\text{color, new}} = \hat{a}_{\text{color, old}} + \alpha(\hat{a}_{\text{color, old}} - a_{\text{color}}) \quad (8)$$

$$\hat{a}_{\text{word, new}} = \hat{a}_{\text{word, old}} + \alpha(\hat{a}_{\text{word, old}} - a_{\text{word}}) \quad (9)$$

where α is the learning rate. Finally, the agent re-evaluates the optimal control policy for the next trial based on its revised model of the task environment.

Task Environments and Parameterization

We simulate behavior of the EVC agent across three different experimental paradigms that have been repeatedly used to index individual differences in cognitive control. Here, we describe each paradigm, the associated behavioral phenomena, as well as the corresponding parameterization[2] of the EVC model.

Stroop Task

In the Stroop paradigm, the agent is presented with a two-dimensional stimulus, one dimension representing an ink color and another dimension representing a color word (Stroop, 1935). On each trial, the EVC model is required to indicate the response associated with the ink color. In congruent trials, the word feature of the stimulus is associated with the same response as the ink color whereas in incongruent trials, the color and word features are associated with different responses. The experiment sequence encompassed 101 trials, and was fully balanced (excluding the first trial) with respect to congruent and incongruent stimuli, as well as with respect to all four transitions between the two trial types (congruent-congruent, congruent-incongruent, incongruent-congruent, incongruent-incongruent). As described below, we sampled a_{color} uniformly from $U(0.3, 0.4)$. To simulate congruent trials, we set $a_{\text{word}} = 0.4$ such that both stimuli dimensions promote the same response. On incongruent trials, we set $a_{\text{word}} = -0.4$ such that the word dimension is associated with a different response than the color dimension. Note that the absolute magnitude of a_{word} is higher than a_{color}, reflecting the assumption that word reading is a more automatic process than color naming (Cohen et al., 1990). We varied the range of control signal intensities from 0 to 10 in steps of 0.2 for the two control signals u_{color}, u_{word} and set the reward received for a correct response to $R = 100$. DDM parameters were set as follows: starting point = 0.0, noise coefficient = 0.7, non-decision time = 0.2s and threshold = 0.4.

We used this paradigm to simulate three different behavioral phenomena. One of the most reliable observations is that participants take more time and commit more errors when responding to incongruent stimuli as opposed to congruent stimuli (Stroop, 1935). Here, we assessed effects of stimulus congruency as the difference in RTs and error rates between incongruent and congruent trials. Another common observation is that participants exhibit a smaller performance cost for incongruent stimuli when the current stimulus was preceded by an incongruent stimulus as opposed to a congruent stimulus (Gratton, Coles, & Donchin, 1992; Egner, 2007). We assessed the congruency sequence effect as an interactive effect between the congruency of the current trial and the congruency of the previous trial on performance. Finally, participants tend to exert smaller congruency effects when the proportion of congruent stimuli is decreased (proportion congruency effect, Logan & Zbrodoff, 1979). We assessed this phenomenon by comparing the congruency effect in two different experiment sequences, one that contained 20% congruent trials, and one that contained 80% congruent trials.

Task Switching

The performance costs associated with switching from one task to another are often used to index cognitive flexibility (Koch, Poljac, Müller, & Kiesel, 2018; Rogers & Monsell, 1995). Here, we examined this effect in a cued task switching paradigm in which the model had to switch between categorizing the color of a stimulus (color naming) and categorizing its shape (shape naming). Similar to the Stroop task, stimuli were either congruent, $a_{\text{color}} = a_{\text{shape}}$, or incongruent, $a_{\text{color}} = -a_{\text{shape}}$. The trial sequence encompassed 100 trials that were randomly sampled with respect to stimulus congruency (congruent, incongruent), the currently relevant task (color naming, shape naming) and the task transition with respect to the previous trial (task switch, task repetition). On each trial, the model allocated control between the two control signals u_{color}, u_{shape}, using the same range of control intensities as described in the Stroop task. The model was cued with a baseline reward of $R = 100$, providing information about which feature is relevant for the task it has to perform on the current trial. DDM parameters were set as follows: starting point = 0.0, noise coefficient = 0.3, non-decision time = 0.2s and threshold = 0.15.

We assessed switch costs in terms of the difference in RTs and error rates between task switch trials and task repetition trials. Rogers and Monsell (1995) also demonstrated that congruency costs are higher on task switch trials compared to task repetition trials. To capture this effect, we also assessed the interaction between stimulus congruency and task transition.

Cognitive Effort Discounting

When given a choice between performing a task with low cognitive effort and a task with high cognitive effort, participants tend to select the former, even if it means to forgo a reward (Westbrook & Braver, 2015). Here, we simulated demand avoidance in the cognitive effort discounting (COGED) experiment described by Westbrook and Braver (2015). In this paradigm, subjects can choose on each trial whether they want to perform a baseline low-demand task for a low reward or a higher-demand alternative task for a higher reward. The amount of reward offered for the baseline task is adjusted to

[2]Note that fixed parameters for each paradigm were chosen such that the model performed with at least 55% accuracy for all combinations of individual difference parameters.

identify the point of indifference, that is, the reward at which subjects are indifferent between performing the low-demand baseline task and performing the high-demand task. To simulate this paradigm, we modeled both tasks as different types of trials that the model can choose between. Each trial encompassed a stimulus with a color dimension that mapped to one of two responses with $a_{\text{color}} > 0$. However, unlike in the Stroop task there was no word dimension, $a_{\text{word}} = 0$. The difficulty of the high-demand task was manipulated across experiment blocks, by varying the color-response association a_{color} from 1.0 to 0.2 in steps of 0.2, and the difficulty of the baseline task was fixed to $a_{\text{color}} = 1$ (higher color-response associations may reflect higher saturation values for a color patch). For each set of simulations, we fixed the reward for the high-demand task to $R = 200$ while steadily increasing the amount of reward offered for the low-demand task in steps of 1, beginning from an initial reward value of $R = 1$. On each trial, the EVC agent determined the highest EVC separately for each task and chose the task with the highest predicted EVC. We then assessed the amount of reward offered for the low-demand task for which the model would be indifferent between performing the low-demand task and the (more rewarding) high-demand task, and normalized this value by the amount of reward offered for the high-demand task. Following the notation by Westbrook and Braver (2015), we refer to this normalized value as the subjective value of completing the high-demand task. For instance, if the model would switch to performing the low-demand task at an offered reward of 120 then the (discounted) subjective value of the high-demand task would be 120/200. The range of control signal intensities was varied from 0 to 10 in steps of 0.2 and DDM parameters were set as follows: starting point = 0.0, noise coefficient = 1.5, non-decision time = 0.2s and threshold = 1. We assessed subjective value the high-demand task as a function of its difficulty, $1 - a_{\text{color}}$.

Simulation Procedure

We simulated behavior of 100 EVC agents in the three paradigms described above. For each agent, we uniformly sampled its control capacity $\varepsilon \sim U(0.5, 1.5)$, implementation cost $c_I \sim U(0.5, 1.5)$, reconfiguration cost $c_R \sim U(0, 3)$, reward sensitivity $v \sim U(0.5, 1)$, the stimulus-response association of the relevant task ($a_{\text{color}} \sim U(0.3, 0.4)$ in all paradigms[3], as well as a_{shape} in the task switching paradigm) and learning rate $\alpha \sim U(0, 0.5)$. Ranges for these parameters were chosen to warrant an accuracy above 55% across all simulated paradigms. Note that agents with a higher control capacity would effectively implement a higher amount of control. Therefore, control capacity can be taken as a proxy for the amount of control an agent exerts on average. The stimulus-response association determines the degree of task automaticity: The higher the stimulus-response association of a task-relevant feature, the easier the task, that is, the less cognitive control is needed to reach the correct outcome. Here, we assume that the stimulus-response association of a task feature reflects the task proficiency of an agent.

We first assessed average behavior across all agents with respect to seven dependent variables. In the Stroop task, we measured error rate effects of stimulus congruency, the congruency sequence effect, the proportion congruency effect, as well as overall error rate on the task. In the task switching paradigm, we assessed switch costs in error rates, as well as the congruency costs in error rates as a function of task transition. We also measured the subjective value of levels of task difficulty as determined by the COGED paradigm.

We restricted our analysis of individual differences to overall error rate in the Stroop task, congruency effects, congruency sequence effects, proportion congruency effects, switch costs, as well as the subjective value assigned to a task parameterized with a_{color} (effort discounting). We then took two different approaches to analyze individual differences in these measures. First, we used a multiple linear regression to assess the degree to which each of the six EVC parameters can explain each behavioral phenomenon. However, we did not include learning rate as a regressor in the task switching and COGED paradigms as the agent is provided full information about each trial. Second, we used principal component analysis (PCA) to explore whether individual differences can be explained by more complex latent factors. That is, we identified principal components that account for variance between agents (observations) across all dependent variables (dimensions), including overall error rate, congruency effect, congruency sequence effect, proportion congruency effect, switch cost and effort discounting. We then assigned a score to each agent that identifies its position on the axes spanned by either the first or the second principal component. These two components explain most of the variance in the space of behavioral phenomena, and can be best interpreted in terms of the behavioral effects that vary most along a given component. In addition, we sought to interpret each component in terms of individual difference parameters of the EVC model. That is, we identified the individual difference parameters that best explain each principal component, by regressing the component scores of all agents against their EVC parameters. Finally, we assessed which of the behavioral phenomena were most indicative of the amount of exerted control, by computing the Pearson correlation between each dependent variable (e.g. congruency effect) and the average intensity of control u that an agent exerts, across all agents.

Results

Behavioral Phenomena. The EVC model captured all of the cognitive control phenomena of interest[4] (Figure 1): 1) Responses were slower and more error-prone on incongruent versus congruent trials of a Stroop-like task (*congruency effect*), $F(1, 99) = 17.80$, $p < 0.001$. 2) When the stimuli on

[3] In the COGED task, we scaled the tested range of a_{color} by this value.

[4] We focused our analyses on error rates due to space constraints. However, we observed similar effects for RTs.

Table 1: Regression of behavioral phenomena against individual differences in EVC parameters. Significant regressors are ordered by standarized regression weight.

Model Parameter	β	t	p
Overall Error Rate, $df = 93$			
Task Automaticity	-0.631	-5.48	< 0.001
Control Capacity	-0.127	-11.77	< 0.001
Implementation Cost	0.126	11.32	< 0.001
Learning Rate	-0.114	-5.05	< 0.001
Reward Sensitivity	-0.076	-3.63	< 0.001
Congruency Effect, $df = 93$			
Task Automaticity	-0.710	-3.50	< 0.001
Learning Rate	-0.219	-5.50	< 0.001
Implementation Cost	0.114	5.82	< 0.001
Control Capacity	-0.089	-4.69	< 0.001
Congr. Sequence Effect, $df = 93$			
Learning Rate	0.145	6.29	< 0.001
Reward Sensitivity	0.055	2.53	< 0.05
Control Capacity	0.053	4.78	< 0.001
Implementation Cost	-0.031	-2.71	< 0.01
Reconfiguration Cost	-0.031	-7.70	< 0.001
Proportion Congr. Effect, $df = 93$			
Task Automaticity	-0.471	-2.19	< 0.05
Learning Rate	-0.199	-4.72	< 0.001
Reward Sensitivity	0.160	4.08	< 0.001
Control Capacity	0.112	5.56	< 0.001
Implementation Cost	-0.103	-4.92	< 0.001
Reconfiguration Cost	-0.044	-6.04	< 0.001
Switch Cost, $df = 94$			
Implementation Cost	-0.069	-4.42	< 0.001
Reward Sensitivity	0.059	2.00	< 0.05
Control Capacity	0.038	2.49	< 0.05
Reconfiguration Cost	0.015	2.77	< 0.01
Effort Discounting, $df = 88$			
Task Automaticity	-0.603	-7.77	< 0.001
Implementation Cost	0.139	17.44	< 0.001
Control Capacity	-0.051	-6.86	< 0.001
Reconfiguration Cost	-0.047	-17.48	< 0.001

the previous trial were incongruent, congruency effects were smaller on the current trial, relative to when the previous trial was congruent (*congruency sequence effect* or *conflict adaptation*), $t(99) = 4.22$, $p < 0.001$. 3) Congruency effects were higher when the trial sequence contained a high proportion of congruent trials versus a high proportion of incongruent trials (*proportion congruency effect*), $t(99) = 17.86$, $p < 0.001$. 4) Responses were less accurate when switching to a new task rather than repeating the same task (*switch costs*, Rogers & Monsell, 1995), $F(1,99) = 337.30$, $p < 0.001$. These switch costs were greater when transitioning to an incongruent trial (Rogers & Monsell, 1995), $F(1,99) = 214.96$, $p < 0.001$. 5) All else being equal, simulated agents assign less value to (and would therefore be less likely to engage with) tasks that are more rather than less difficult (*cognitive effort discounting*, see Figure 1D).

Individual Differences. We tested the degree to which each of the measures above were influenced by individual differences in factors related to bottom-up stimulus processing (task automaticity), cognitive control ability (control capacity), and motivational factors (e.g., reward sensitivity and control costs). Agents with a higher control capacity and lower implementation costs made fewer errors, had lower congruency effects, higher congruency sequence effects, adapted more to the proportion of congruent trials, had higher switch costs and discounted cognitive effort less (Table 1). Agents with higher reconfiguration costs and a lower sensitivity to reward adapted less to congruency of the previous stimulus or to the proportion of congruent trials. Both, a higher reconfiguration cost and a higher reward sensitivity were associated with higher switch costs. A higher reward sensitivity also yielded overall fewer errors while higher reconfiguration costs predicted less effort discounting. Agents with a higher learning rate and task automaticity performed overall better in the Stroop task, showing smaller congruency effects and smaller proportion congruency effects. Unsurprisingly, agents with a higher learning rate show greater sequential adaptations to response congruency whereas agents with a higher task automaticity discounted effort less.

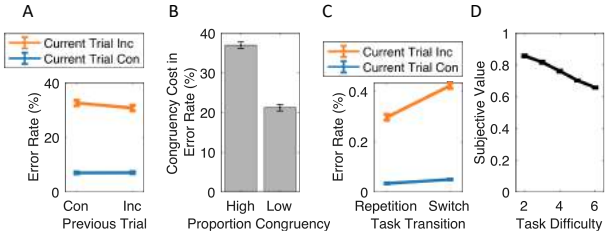


Figure 1: Average effects of simulated agents. (A) Error rates are shown as a function of congruency of the previous and the current trial. (B) Congruency effects (difference in error rates on incongruent and congruent trials) are shown for a sequence with a low (80%) and high (20%) proportion of congruent trials. (C) Error rates are shown as a function of congruency of the previous trial and task transition. (D) Subjective value of a task as a function of its difficulty. Error bars indicate the standard error of the mean across simulated agents.

Principal Components Analysis. After performing a PCA across our behavioral effects of interest, we found that individual differences across these are well captured by two orthogonal dimensions that explained more than 75% of between-agent variance (Figure 2). Regressing these phenomenon-driven components on the model parameters that we varied, we find that a high score on Component 1 is associated with higher task automaticity, lower implementation costs, higher control capacity and higher sensitivity to reward. Agents with a higher value for any of these parameters are expected to perform better on a task (Table 3). Component 2 appears to most reliably capture differences in reconfiguration costs, and to a lesser degree differences in task automaticity, reward sensitivity and implementation costs.

Correlation with control intensity. We found that each behavioral effect significantly correlated with the average amount of control exerted by an agent (Table 2). Interestingly, overall error rate in the Stroop task was most indicative of exerted control intensity, followed by incongruency costs.

General Discussion and Conclusion

People have varying degrees of success at adapting their thoughts and behaviors to meet their current goals. Failing

Table 2: Correlations between dependent behavioral measures and exerted control intensity across simulated agents ($df = 98$).

Dependent Measure	r	p
Overall Error Rate	-0.76	< 0.001
Congruency Effect	-0.67	< 0.001
Proportion Congruency Effect	0.58	< 0.001
Effort Discounting	0.46	< 0.001
Congruency. Sequence Effect	0.31	< 0.01
Switch Cost	0.20	< 0.05

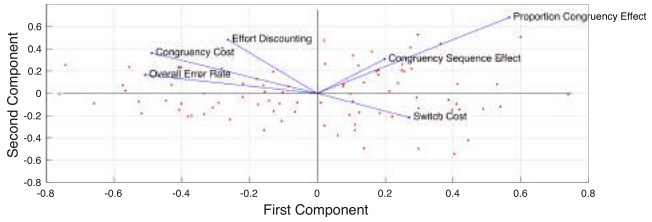


Figure 2: Principal Components Analysis. Each red dot summarizes the behavior of an agent in the space of the first and second principal component. The direction and length of the blue vectors indicates the score of each behavioral effect in terms of the two components. For instance, subjects with low scores on the first component appear to commit more errors but show lower costs of switching tasks.

to exert the appropriate level of control can have very negative consequences for one's health, career, and social status. It is therefore important to understand whether and how such real-world self-control can be predicted from lab-based measures of cognitive control. We used a computational model of control allocation to examine the degree to which different performance metrics from such tasks can theoretically index individual differences in processes related to stimulus processing, strength of control, and motivation for control.

We showed that the EVC model can account for a wide array of effects used to index cognitive control, including response interference, sequential adaptation to stimulus congruency, adaptation to the proportion of congruent stimuli, performance costs associated with task switching, and demand avoidance. Critically, we showed that individual differences in each of these measures can be accounted for by a multitude of factors, including motivational variables (e.g., reward sensitivity) and bottom-up stimulus processing (task automaticity), rather than only by one's *ability* to exert cognitive control (indexed by control capacity). This suggests that individual differences in cognitive control phenomena may not not be a reliable indicator of one's ability to exert control but may instead reflect individual differences in other variables. A PCA revealed a broad distinction between effects that vary as a function of how much control an agent is capable of allocating (overall performance, congruency costs, effort discounting) and effects that index how flexibly an agent can adapt to changing demands of the environment (switch costs, congruency sequence effects and proportion congruency effects). Finally, our analyses suggest that overall error rate and incongruency costs in the Stroop task best indexed the actual amount of control exerted by an agent whereas congruency sequence effect and switch costs were found to be least diagnostic.

Table 3: Regression of principal components (PC) against individual differences in EVC parameters.

Model Parameter	β	t	p
First PC, $df = 88$			
Task Automaticity	0.7867	3.58	< 0.001
Implementation Cost	-0.2582	-11.43	< 0.001
Control Capacity	0.2289	10.84	< 0.001
Reward Sensitivity	0.1870	4.53	< 0.001
Reconfiguration Cost	-0.0122	-1.61	0.111
Second PC, $df = 88$			
Task Automaticity	-0.8127	-4.10	< 0.001
Reward Sensitivity	0.0882	2.37	< 0.05
Reconfiguration Cost	-0.0586	-8.59	< 0.001
Implementation Cost	0.0435	2.14	< 0.05
Control Capacity	0.0033	0.17	0.862

Interestingly, we found that higher costs of *implementing* control were associated with lower costs of *switching* tasks. This finding is consistent with previously observed tradeoffs between cognitive stability and cognitive flexibility: higher amounts of control can reduce distractor interference but require larger reconfiguration of control signals when switching between tasks (Goschke, 2000; Musslick, Jang Jun, Shvartsman, Shenhav, & Cohen, 2018). Perhaps more surprisingly, we also found that participants with higher reconfiguration costs discounted cognitive effort *less* (i.e., were more willing to engage in demanding tasks). This finding reflects an approach-avoidance conflict inherent to demand avoidance paradigms: The more a person is engaged with a cognitively demanding task, the less they are willing to switch to an easier task (Kool, McGuire, Rosen, & Botvinick, 2010).

One limitation of the current implementation of the EVC model is its focus on 2-alternative forced choice tasks. We chose to focus on these tasks because they are amenable to analysis with the well-studied DDM (Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006; Ratcliff, 1978). However, the DDM may be an over-simplified model for cognitive control tasks given that such tasks can involve a variety of response alternatives, as in traditional variants of the Stroop task (Stroop, 1935).

The set of relevant individual difference parameters heavily depends on the requirements of the task for cognitive control. For instance, the n-back task requires subjects to decide whether a stimulus matches the stimulus that was presented n steps before in a sequence, and has been hypothesized to involve processes of working memory updating, interference between representations held in working memory, and familiarity judgment (Chatham et al., 2011; Juvina & Taatgen, 2007). The study of individual differences in more complex tasks will require implementing more realistic process models of those tasks, such as a working memory gating model in the case of the n-back task (Chatham et al., 2011).

Altogether, these analyses suggest that individual differences in cognitive control phenomena do not necessarily re-

flect differences in someone's capacity to exert cognitive control but may as well reflect differences in task automaticity or sensitivity to reward. Accounting for differences in these variables is therefore crucial when indexing cognitive control through behavioral phenomena. However, the collinearity between simulation parameters in this analysis prevents us from teasing apart the effects of each parameter. More elaborate parameter sensitivity studies are necessary to provide more fine grained insights into the source of individual differences in cognitive control phenomena.

References

Allan, J. L., Johnston, M., & Campbell, N. (2010). Unintentional eating. what determines goal-incongruent chocolate consumption? *Appetite*, *54*(2), 422–425.

Berkman, E. T., Hutcherson, C. A., Livingston, J. L., Kahn, L. E., & Inzlicht, M. (2017). Self-control as value-based choice. *Curr Dir Psychol Sci*, *26*(5), 422–428.

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychol. Rev.*, *113*(4), 700.

Caplin, A., Csaba, D., Leahy, J., & Nov, O. (2018). *Rational inattention, competitive supply, and psychometrics* (Tech. Rep.). National Bureau of Economic Research.

Chatham, C. H., Herd, S. A., Brant, A. M., Hazy, T. E., Miyake, A., O'Reilly, R., & Friedman, N. P. (2011). From an executive network to executive control: a computational model of the n-back task. *Journal of cognitive neuroscience*, *23*(11), 3598–3619.

Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: a parallel distributed processing account of the stroop effect. *Psychological Review*, *97*(3), 332–361.

Egner, T. (2007). Congruency sequence effects and cognitive control. *Cogn Affect Behav Neurosci*, *7*(4), 380–390.

Goschke, T. (2000). Intentional reconfiguration and involuntary persistence in task set switching. *Control of cognitive processes: Attention and performance XVIII*, *18*, 331.

Gratton, G., Coles, M. G., & Donchin, E. (1992). Optimizing the use of information: strategic control of activation of responses. *J. Exp. Psychol. Gen.*, *121*(4), 480.

Juvina, I., & Taatgen, N. A. (2007). Modeling control strategies in the n-back task. In *Proceedings of the 8th international conference on cognitive modeling* (pp. 73–78).

Koch, I., Poljac, E., Müller, H., & Kiesel, A. (2018). Cognitive structure, flexibility, and plasticity in human multitaskingan integrative review of dual-task and task-switching research. *Psychological bulletin*, *144*(6), 557.

Kool, W., McGuire, J. T., Rosen, Z. B., & Botvinick, M. M. (2010). Decision making and the avoidance of cognitive demand. *J. Exp. Psychol. Gen.*, *139*(4), 665.

Lieder, F., Shenhav, A., Musslick, S., & Griffiths, T. L. (2018). Rational metareasoning and the plasticity of cognitive control. *PLoS Comput. Biol.*, *14*(4), e1006043.

Logan, G. D., & Zbrodoff, N. J. (1979). When it helps to be misled: Facilitative effects of increasing the frequency of conflicting stimuli in a stroop-like task. *Memory & cognition*, *7*(3), 166–174.

Manohar, S. G., Chong, T. T.-J., Apps, M. A., Batla, A., Stamelou, M., Jarman, P. R., ... Husain, M. (2015). Reward pays the cost of noise reduction in motor and cognitive control. *Current Biology*, *25*(13), 1707–1716.

Meiran, N. (1996). Reconfiguration of processing mode prior to task performance. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(6), 1423.

Musslick, S., Cohen, J. D., & Shenhav, A. (2018). Estimating the costs of cognitive control from task performance: theoretical validation and potential pitfalls. In *Proceedings of the 40th annual conference of the Cognitive Science Society* (pp. 800–805). Madison, WI.

Musslick, S., Jang Jun, S., Shvartsman, M., Shenhav, A., & Cohen, J. D. (2018). Constraints associated with cognitive control and the stability-flexibility dilemma. In *Proceedings of the 40th Annual Meeting of the Cognitive Science Society* (pp. 806–811). Madison, WI.

Musslick, S., Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2015). A computational model of control allocation based on the expected value of control. In *The 2nd Multidisciplinary Conference on Reinforcement Learning and Decision Making*. Edmonton, Can.

Navarro, D. J., & Fuss, I. G. (2009). Fast and accurate calculations for first-passage times in wiener diffusion models. *Journal of Mathematical Psychology*, *53*(4), 222–230.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological review*, *85*(2), 59.

Rogers, R. D., & Monsell, S. (1995). Costs of a predictible switch between simple cognitive tasks. *J. Exp. Psychol. Gen.*, *124*(2), 207.

Saunders, B., Milyavskaya, M., Etz, A., Randles, D., & Inzlicht, M. (2017). Reported self-control is not meaningfully associated with inhibition-related executive function: A bayesian analysis, doi: 10.1525/collabra.134.

Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2013). The expected value of control: an integrative theory of anterior cingulate cortex function. *Neuron*, *79*(2), 217–240.

Shenhav, A., Musslick, S., Lieder, F., Kool, W., Griffiths, T. L., Cohen, J. D., & Botvinick, M. M. (2017). Toward a rational and mechanistic account of mental effort. *Annual review of neuroscience*, *40*, 99–124.

Streeter, C. C., Terhune, D. B., Whitfield, T. H., Gruber, S., Sarid-Segal, O., Silveri, M. M., ... others (2008). Performance on the stroop predicts treatment compliance in cocaine-dependent individuals. *Neuropsychopharmacology*, *33*(4), 827.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of experimental psychology*, *18*(6), 643.

Westbrook, A., & Braver, T. S. (2015). Cognitive effort: A neuroeconomic approach. *Cogn Affect Behav Neurosci*, *15*(2), 395–415.

The Modularity of the Motor System

Myrto Mylopoulos
Carleton University, Ottawa, Ontario, Canada

Abstract

The extent to which the mind is modular is a foundational concern in cognitive science. Much of this debate has centered on the question of the degree to which input systems, i.e., sensory systems such as vision, are modular (see, e.g., Fodor 1983; Pylyshyn 1999; MacPherson 2012; Firestone & Scholl 201; Burnston 2017; Mandelbaum 2017). By contrast, researchers have paid far less attention to the question of the extent to which our main output system, i.e., the motor system, qualifies as such. I will argue that the motor system should be construed as quasi-modular, at best, in that it is informationally encapsulated only to a certain degree, and in a way that can be strategically modulated by the agent. I will explore the implications of this result for nearby philosophical puzzles relating to different aspects of action control.

Do round numbers always become reference points?: An examination by Japanese and Major League Baseball data

Kuninori Nakamura (nakamura.kuninori@gmail.com)
Faculty of Social Innovation, Seijo University, 6-1-20, Seijo, Setagayaku
Tokyo 152-0061, Japan

Abstract

The round number effect refers to discontinuity around round numbers ("0.300", "4 hours") in frequency distribution, indicating that people consider the round numbers as goals or reference points for their performances. This study aimed to examine the round number effect by exploring the following two issues: (1) examination of Japanese baseball data, and (2) comparison between batters who exceeded the regulation number of at-bat of season and those who did not. Results indicate the following three points; (1) the round number effect was found in Japanese baseball data, (2) but it was found only for the batters who exceeded provision bat number of season, and (3) magnitude of the effect was stronger in Japanese than Major League Baseball data. General discussion argued these results in terms of players' motivation and disposition.

Keywords: reference dependence, round number effect, discontinuity

Introduction

How people perceive values of objects depends not only on nature of the objects themselves but also the degree the objects are judged compared using a criterion. This tendency is known as reference dependency (Kahneman & Tversky, 1979; Rosch, 1978), and has been demonstrated through exploration for anomalies in human decision-making including studies on framing effect (Tversky & Kahneman, 1981) or anchoring effect (Tversky & Kahneman, 1974). According to these studies, people's judgments are affected by arbitrarily presented numbers (Tversky & Kahneman, 1973) or wordings of decision problems (Tversky & Kahneman, 1981) that are not themselves irrelevant to answers for the decision tasks, indicating that people make judgments according to the reference points that are considered as goals for their behavior rather than relying on their own beliefs or preferences.

The round number effect (Allen et al, 2016; Pope & Simonsohn, 2009) is a new example of reference dependency that refers to a discontinuity of distribution for continuous variables around the round numbers that can be considered goals for performance. When students take the SAT, they may say that they want to acquire 1100 points, but never 1098 or 1103 points. Marathon runners may try to finish the race within 4 hours, but never within 3 hours and 56 minutes. As these examples indicate, round numbers are often employed as goals for performances. As a result, people's behavior would change depending on whether they can perform just short of the round number or not, and density of the distribution for the performance would fluctuate around the round number (also see Heath et al., 1999; Kahneman & Miller, 1986; Medvec & Savitsky, 1997).

Pope and Simonsohn (2011) reported the round number effect in Major League Baseball (MLB) data. In their study, they analyzed data about batting averages of MLB players who had scored at least 200 at bats during the season from 1975 to 2007, and found that the relative frequency of baseball players whose batting averages at the end of the seasons were 0.300 was higher than those whose batting averages were 0.299. In addition to the discontinuity of frequency around 0.300, Pope and Simonsohn (2011) also found that performance or strategy of the batters who could achieve "0.300" by hit on their final bat differed from those of the batters who could not or already achieved "0.300": their probabilities for hitting at the final bat were higher than, and probabilities to choose the walks at the final bat were lower –in fact, 0- than the other batters. These results indicate that the batters used "0.300" as their goal for their batting performance, and tried to achieve this goal by any means, resulting in discontinuity around the round number. Beside the example of batting average data, the round number effect was reported in various domains including SAT score (Pope & Simonsohn, 2011) or marathon running time (Allen et al, 2011; also see Pope & Simonsohn, 2011 for an example of fictitious scenario).

Studies on round number effect suggest that people in some sense can do nothing without a concrete goal In other words, rather than trying as much as they can, people aim to exceed their criteria that are arbitrarily determined, and round numbers are often chosen as simple and definite goals. This study aimed to deepen the understanding of how round numbers become reference points by analyzing records of batting average data in the same way as Pope and Simonsohn (2011). However, this study explored the following two points that have not been examined in the previous studies.

One purpose of this study was to test the round number effect in Japanese baseball data. Although Pope and Simonson (2009) demonstrated the round number effect in MLB data, its replicability in other leagues is still unexplored. In Japan, baseball is the one of the most popular and major professional sports, and abundant data for batting records are available for this examination. Additionally, the round number "0.300" is often referred to the cutoff for top-ranking batters in Japanese baseball. Thus, Japanese professional baseball (Nippon Professional Baseball: NPB)

data is adequate to test whether the round number effect can be replicated in the same way as in MLB.

The other purpose of this study as to explore individual differences in the round number effect of "0.300." Although this study referred to the number "0.300" as cutoff for the top-ranking batters in professional baseball, it also recognizes that this number does not always become goals for all the players. In professional baseball, batting average itself cannot be proof of the top-ranking batters because values of the batting averages depend on number of plate appearance. For example, although batting averages of the batters who got three hits at ten bats or one hundred and fifty hits at five hundred bats are both the same at "0.300," meanings of "0.300" might be different between the batters. Though the former average might be considered as a result of chance, the latter average would be thought as reflecting real ability. In fact, both MLB and NPB determined the regulation number of at-bat, and averages cannot be recognized as formal records without achieving the regulation number. Thus, it is probable that whether batters consider "0.300" as reference points depend on their numbers of battings: while the batters who cannot achieve number of regulations at battings do not consider "0.300" as their goal, for ones who can achieve the number, it might be valuable goal. The second purpose of this study was to test this possibility.

To accomplish the above purposes, this study gathered data for batting averages both in MLB and NPB, and tested the round number effect by examining whether relative frequencies of the batters whose batting averages were "0.300" stick out in the frequency distribution. In doing so, this study employed statistical test used in Chetty, Friedman, Olsen, and Pistaferri (2011) that utilize a broader range of data than Pope and Simonsohn (2011). With this method, this study also analyzed data that were divided by the number of plate appearance and examined whether the round number effect depends on whether the batter achieved the number of regulations at bats or not. The theoretical meanings and scope of the round number effect were discussed as well..

Method

This study obtained batting average data of Japanese baseball from Nippon Professional Baseball Organization (http://npb.jp/). This website contains results for all players of Japanese professional baseball from 2005 to 2018. Similar to Pope and Simonsohn (2011), this paper restricted its sample to players who had at least 200 at bats during the season. As a result, data of 1436 players were analyzed.

This paper also obtained MLB data in the same time as NPB data from Retrosheet.com. This website has data that was used by Pope and Simonsohn (2011), and contains data of all players in MLB for the same periods as Japanese data. This paper collected data of MLB using the same criteria as Japanese data, resulting in analysis of data of 4630 players.

This study aimed to investigate whether the round number effect would vary according to the number of appearances at bats. To accomplish this, the study divided the data into two groups by whether the number of appearances at bats exceeded the regulation numbers at bats or not both in NPB and MLB data. In NPB data, the regulation numbers at bats differed depending on year and league. In 2005 and 2006 the regulation numbers were 452 at Central League and 426 at Pacific League, and during 2007 and 2014, the number in both leagues was 446, and after 2015, it was 443. As a result, this study considered 743 batters as exceeding the regulation number at bats and 818 batters as not exceeding the regulation number.

In MLB, the regulation number at bats was settled as 400 since 1957. Thus, this study considered 2561 batters as exceeding the regulation number at bats and 2081 batters as not exceeding the regulation number in MLB data.

Results and discussion

Analyses of fundamental statistics of the round number effect

Figure 1 shows distribution of relative frequencies of batting averages at the ends of season both for players who had at least 200 at bats in both NPB and MLB. Visual inspections to these distributions suggest the following three points. First, with regard to the players who had at least 200 at bats, relative frequency of the average "0.300" is higher than any other values of batting averages of NPB and MLB players. In fact, similar to Pope and Simonsohn (2011), the proportion of players ending the season with a 0.298 or 0.299 was lower than that with a 0.300 or 0.301 ($z=4.84$ and 3.99 for MLB and NPB, respectively) in both the leagues, indicating that this study replicated the round number effect with a dataset different from that of Pope and Simonsohn (2011).

However, trends of the frequency around 0.300 appear to be different between the batters who exceeded the regulation number of at-bat of season and those who did not: the density of frequency around 0.300 that occurred for the batters who exceeded the number disappeared for the batters who did not. The Z tests demonstrated a significant difference in the proportion between 0.298 or 0.299 and 0.300 or 0.301 ($z=4.87$ and 4.05 for MLB and NPB, respectively) for the batters who exceeded numbers of regulation at bat in season, but did not demonstrate significant differences for the batters who did not exceed the number ($z=1.48$ and 1.50 for MLB and NPB, respectively). These results indicate that the round number effect would vary according to the number of at bat.

Third, increases in the frequency around 0.300 were higher for the NPB than MLB players. With regard to the total data, though the ratio of the frequency between 0.300 and 0.299 was more than 9 (3 for 0.299, and 28 for 0.300) in NPB data, that in MLB data was less than 3 (22 for 0.299,

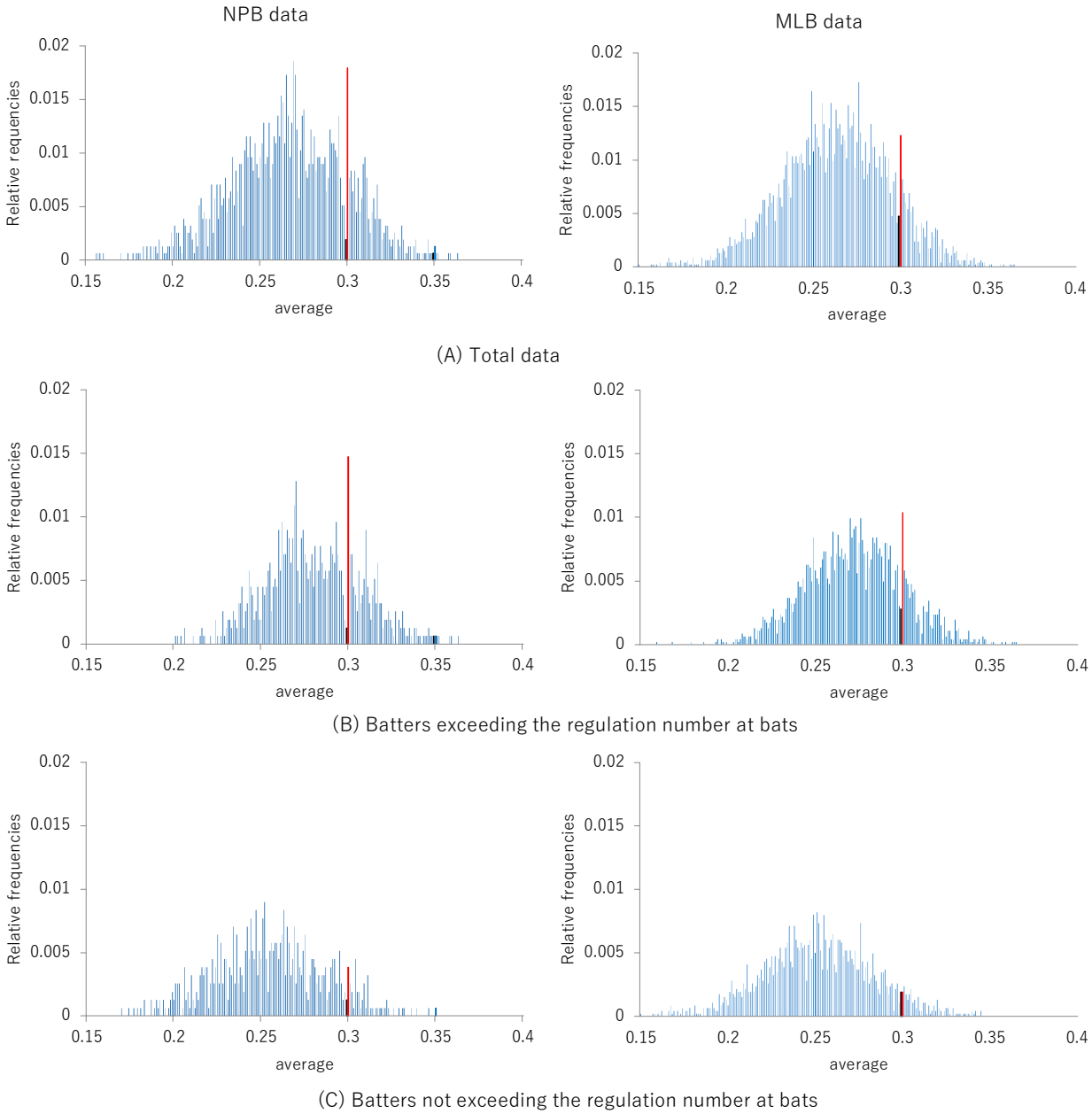


Figure 1 Distributions for batting averages of batters in NPB and MLB: Left and right column demonstrate distributions for NPB batters and MLB batters, respectively. Figures in the top, middle, and bottom row demonstrate distributions of the all batters, the batters who exceeded the number of regulation at bats, and batters who did not exceed the number of regulation at bat, respectively.

and 57 for 0.300). When data were limited only to the batters who achieved the provision bat number in the season, differences in the ratio became more salient (2 for 0.299 and 23 for 0.300 in NPB data, and 13 for 0.299 and 48 for 0.300 in MLB data). These results suggest that the round number effect would occur more strongly in NPB than MLB.

Quantitative estimation of the round number effect

One concern of the above discussion is that they solely consider data around the round number. However, as shapes of the distributions indicate, rise and drop around the round number in frequencies might be within probabilistic fluctuation. Thus, to explore this possibility, this study adapted a methodology proposed in Chetty et al. (2011) to quantify the extent of excess mass in an interval around a round number. This methodology enables examining

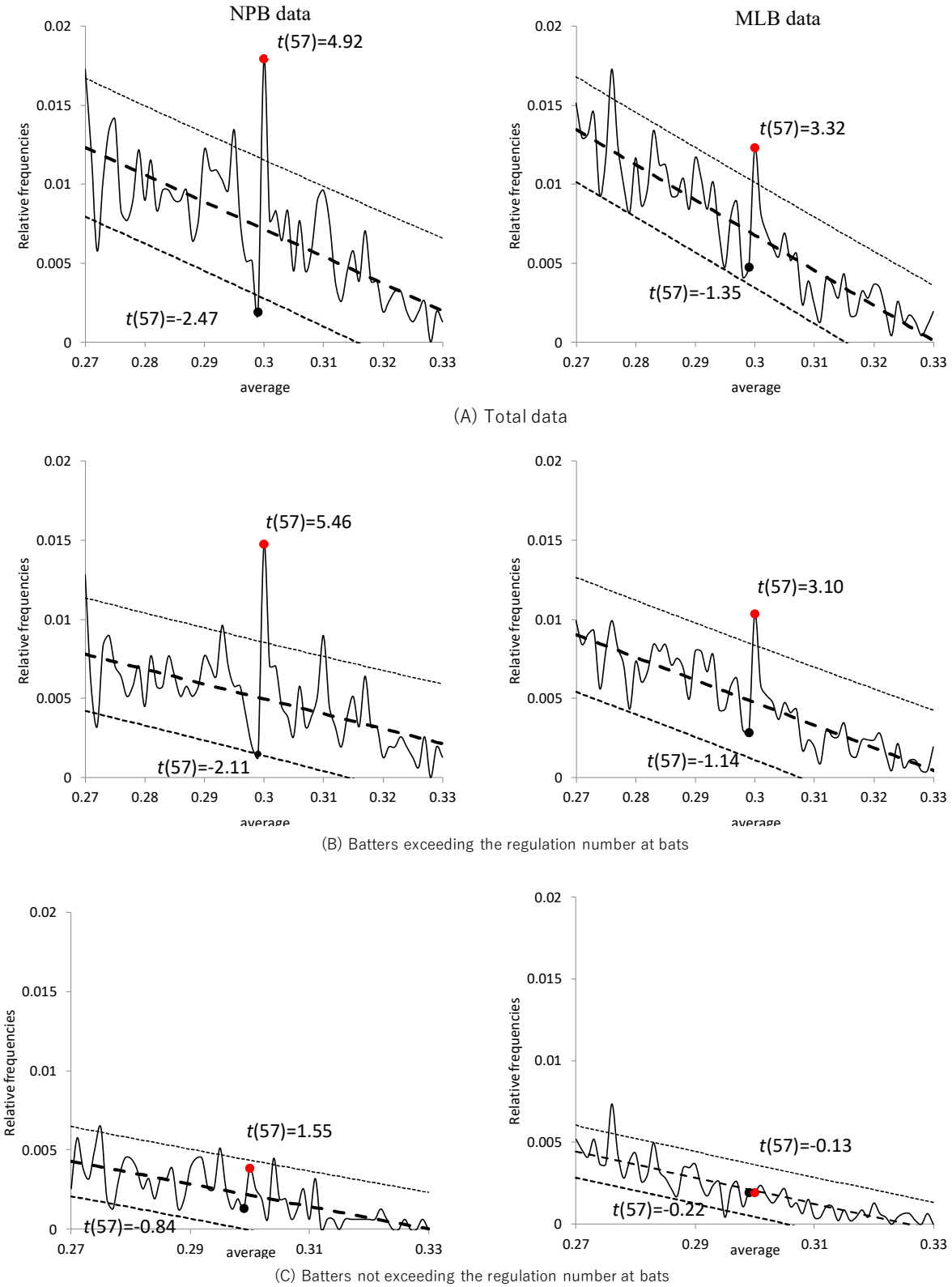


Figure 2 Result of Chetty's test: Left and right columns demonstrate distributions for NPB batters and MLB batters, respectively. Figures in the top, middle, and bottom row demonstrate distributions of the all batters, the batters who exceeded the number of regulation at bats, and batters who did not exceed the number of regulation at bat, respectively. Dotted lines in the graphs indicate regression line (bold) and 95% predictive intervals (thin). Red points indicate data of 0.300, and black points indicate data of 0.299.

whether discontinuity in distribution would occur around the reference point. The following section explains this methodology referring to the case of the round number effect in the batting average data.

This study aimed to examine whether discontinuity occurs between 0.299 and 0.300 because for the batters, "0.300" serves as reference point to be achieved. To address this issue, this study considered what would happen if "0.300" was not the reference point and the bunch around this number occurred in fact by chance. If so, frequencies of the batters around these values would follow trends that can be predicted through data from other domains. To test this consideration, this study first analyzed batting average data excluding frequencies of the batters whose averages were 0.299 and 0.300 with polynomial regression analysis where the batting average was the independent variable and the frequency of the batters was the dependent variable. Then, based on the results of this analysis, this study predicted frequencies of the batters whose averages were 0.299 and 0.300, and compared these predictions with the data. In sum, Chetty et al.'s (2011) method constructed "counterfactual" prediction from the data excluding values around the round numbers and examined the degree of deviation of the data from the counterfactual prediction by estimating prediction intervals of the polynomial regression models.

Using this methodology requires several inspections. First, the range of data used for this methodology must be selected in terms of research. For example, Allen et al. (2013) decided the range of data to be analyzed using Chetty et al.'s (2011) methodology through "visual inspection" of the data. Following Allen et al. (2013), this study also decided the range of data through visual inspection of the distribution. Thus, frequencies of the batters between 0.270 and 0.330 were used for the analysis. Second, the number of independent variables must be selected from results of polynomial regression analyses. This study adopted single regression model to approximate the data because only the effect of first order term in this model was significant throughout the six distributions.

Results of the analyses are shown in Figure 2. As the graphs shown in Figure 2 demonstrate, the frequencies of the batters whose averages were 0.300 were beyond 95% predictive intervals of the regression model both for NPB and MLB data. However, the frequencies of the batters who did not achieve the number of regulations at bat were within the 95% predictive interval from the regression models. Additionally, values of t-statistics for frequencies of the batters at 0.300 were larger in NPB than MLB, indicating that deviation from the predictions from the regression are more extreme in NPB than in MLB. In sum, these results support the findings in the previous section: the round number effect was replicated in NPB data, and depends on the number of plate appearance. Moreover, magnitude of the round number effect is more prominent in NPB than MLB.

Conclusion

Findings from the above analyses can be summarized as follows. First, this study demonstrated the round number effect using dataset other than that of Pope and Simonsohn (2011). The results indicate that the round number effect occurred in NPB data, and in doing so, this study analyzed MLB data collected using criteria different from those of Pope and Simonsohn (2011). This finding is also important in showing replicability of the round number effect in real situations with more strict methodology (Chetty et al., 2011) that was not adopted in the previous study (Pope & Simonsohn, 2011).

Second, this study found that the round number effect depends on people's incentives or motivation. Analyses of the data of the batters who did not exceed the number of regulations at bat indicated that increase in the frequencies of batters whose averages were 0.300 did not occur, indicating that 0.300 was not the reference point for performance for these batters. This finding indicates scpoe of the round number effect, which had not been explored in the existing studies (Allen et al., 2016; Pope & Simonsohn, 2011). Specifically, this finding is important in that the round number effect does not solely depend on people's preference for the round number.

In this vein, it is interesting that this study found difference in the round number effect between NPB and MLB data. Although NPB and MLB data are different in their sample size, discontinuities in the distributions around 0.300 in NPB data are more prominent than those in the MLB data, indicating that Japanese batters attach more weight to average 0.300 than Major League batters. This difference might be based on some cultural difference in professional baseball, suggesting that situational factor also affect the round number effect. In other words, Japanese baseball culture might attach greater importance on batting average to evaluate abilities of batters than in MLB. More precise analyses should be conducted to understand this difference in future research more profoundly.

According to Pope and Simonsohn (2011), the round number can serve as cognitive reference point in numerical scale in the same way as goals (Heath, Larrick, & Wu, 1999; Larrick, Heath, & Wu, 2009), expectations (Feather, 1969; Mellers, Schwartz, Ho, & Ritov, 1997), and counterfactual (Kahneman & Miller, 1986; Medvec, Gilovich, & Madey, 1995; Medvec & Savitsky, 1997). In contrast to this interpretation, this study revealed motivational and situational aspect of the round number effect. In other words, this study suggests that the round number cannot become reference point for performance solely by itself. Although for batters who exceed the number of regulations at bat the average of 0.300 can serve as proof for the top rank batter, for the batters who do not exceed the number, 0.300 itself might not be an important number. That is, for these batters, before achieving the average 0.300, it is important to exceed the number of regulation at bat. In addition, situational or cultural factor may also enhance the round number effect. This implication of the round number

suggests that meaning of the number depends on individual disposition. Thus, exploring conditions that enhance or inhibit the round number effect is an important future research question.

One methodological concern for this study was the arbitrariness of assumptions in analysis. This study performed the Chetty test (Chetty et al., 2011) using data from 0.270 to 0.330 through visual inspection of the distributions. However, it is possible to consider other criteria for data selection from the distribution to analyze the round number effect. Sophistication of methodology to test the round number effect is also necessary for future research.

References

Allen, E. J., Dechow, P. J., Pope, D., & Wu, G. (2016). Reference-dependent preferences: Evidence from marathon runners. *Management Science*, 63, 1657-1672.

Chetty, R., Friedman, J. N., Olsen, T., & Pistaferri, L. (2011). Adjustment costs, firm responses, and micro vs. macro labor supply elasticities: Evidence from Danish tax records. *Quarterly Journal of Economics*, 126, 749–804.

Feather, N.T. (1969). Attribution of responsibility and valence of success and failure in relation to initial confidence and task performance. *Journal of Personality and Social Psychology,* 13, 129–144.

Heath, C., Larrick, R.P., & Wu, G. (1999). Goals as reference points. *Cognitive Psychology*, 38, 79–109.

Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review*, 93, 136–153.

Larrick, R. P., Heath, C., & Wu, G. (2009). Goal-induced risk taking in negotiation and decision making. *Social Cognition*, 27, 342–364.

Medvec, V. H., Gilovich, T., & Madey, S. F. (1995). When less is more: Counterfactual thinking and satisfaction among Olympic medalists. *Journal of Personality and Social Psychology*, 69, 603–610.

Medvec, V. H., & Savitsky, K. (1997). When doing better means feeling worse: The effects of categorical cutoff points on counterfactual thinking and satisfaction. *Journal of Personality and Social Psychology*, 72, 1284–1296.

Mellers, B. A., Schwartz, A., Ho, K., & Ritov, I. (1997). Decision-affect theory: Emotional reactions to the outcomes of risky options. *Psychological Science*, 8, 423–429.

Pope, D., & Simonsohn, U. (2011). Round numbers as goals: evidence from baseball, SAT takers, and the lab. *Psychological Science*, 22, 71–79.

Kahneman, D., & Tversky, A. (1979). Prospect theory: an analysis of decision under risk. *Econometrica*, 47, 263-291.

Rosch, E. (1975). Cognitive reference points. *Cognitive Psychology*, 7, 532–547.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases. *Science*, 185, 1124–1131

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, 211, 453–458.

Cultural Affordances in AI Perception

Zachariah A. Neemeh (zaneemeh@memphis.edu)
Department of Philosophy, University of Memphis, Memphis, TN 38152

Abstract

Affordances offer AI research an alternative from representations for linking perception to action in autonomous systems. Affordances are based in the informational structure of the environment and the somatic capacities of the agent and arise in their interaction. AI implementations of affordance perception typically utilize relatively basic, natural affordances such as the graspability of a handle. Culturally-scaffolded affordances, such as the letter-mailing capacity of a postbox, pose a more intractable problem for affordance-based robotics. This class of affordances requires acculturation and is highly culture-specific. AI implementations of affordance perception typically bypass this difficulty by making recourse to representations. I begin by reviewing affordance perception and the difference between natural and cultural affordances. I then critically discuss implementations of cultural affordance perception in autonomous agents. Finally, I argue that AI affordance perception does not require a robust representationalism in order to implement cultural affordances.

Keywords: affordances; AI perception; embodied cognition; philosophy of AI; representations

Introduction

The perpetually shifting nature of the environment poses a significant challenge to robotics research. Autonomous agents must negotiate and adapt to dynamically changing environments. Representational architectures limit autonomous agents' capacity to do so, however (Raubal & Moritz, 2008). Problems arise with the bandwidth, processing power, computational time, and programming time required to represent shifting environments (Rome et al., 2006). Nonrepresentational and affordance-based architectures have been proposed to overcome these difficulties (Braitenberg, 1984; Brooks, 1990, 1991; Horton, Chakraborty, & St. Amant, 2012). These architectures do not rely on separate layers for perception, action, and planning or reasoning. Instead, they offload part of the computational process onto the environment. As Rodney Brooks, the pioneer of embodied robotics and the inventor of the Roomba, said, "the world is its own best model" (1990, p. 4).

Many nonrepresentational architectures utilize affordances to replace otherwise separate perceptual and actional layers. Natural affordances, like the graspability of a handle, are embedded in the basic informational structure of the immediate environment (Gibson, 1979/2015). The agent picks up on information available in the environment, such as the light waves and pressure feedback of the handle. The affordance arises as the agent interacts with the object and is an opportunity for action that is highly constrained by the agent's form of embodiment.

Implementations of affordance perception are beset by a difficulty, however, once they encounter cultural affordances, or affordances that implicate background knowledge that is culturally mediated. Few AI implementations of affordance perception have attempted to incorporate such higher-order affordances (see Awaad, Kraetzschmar, and Hertzberg, 2015; Chu, Fitzgerald, and Thomaz, 2016; Raubal & Moritz, 2008). While the graspability of a handle can be modeled as an online, dynamical interaction unfolding between the agent's sensorimotor processes and the object's properties, the mail-ability of a letter implicates a vast background knowledge of letters, the postal service, postboxes, writing, and interpersonal communication. This background knowledge poses a significant problem. If all the rules and background knowledge pertinent to the mail-ability affordance must be represented, then affordance-based robotics offers little advantage over traditional architectures.

In this paper, I critically review extant AI implementations of cultural affordance perception and sketch a framework for perceiving cultural affordances with minimal recourse to representations. My aim is to show that a robust representationalism is not *conceptually* necessary for cultural affordance perception.

Affordance Perception

In classical computational models, the perception of the environment involves the creation of an inner, representational model (Fodor, 1985; Marr, 1982/2010).[1] The agent is decoupled from its environment and interacts with it through the medium of representations, which are processed computationally. The term 'representation' is used in a wide variety of senses, from a minimal sense of a covariation between an internal and external state, to a more robust internal mapping of an external state. To avoid the deep complexities involved in this term, I here use it to pick out any internal state that tracks an external state in the world, when that state is decoupled from perception and action. In autonomous agents, typically representations are instantiated in a planning or reasoning layer mediating between perceptual and actional layers. Processing can be

[1] While computationalism and representationalism do not necessarily entail one another (see Dennett, 1969), in practice they usually work in tandem.

through classical, serial processing architectures (as in Turing machines), or they can be massively parallel (as in connectionist, neural network, and similar architectures).

Autonomous agents in the real world encounter a wide variety of environments, no two of which will exactly be the same. Even the same environment often shifts in content through time. This creates a significant computational challenge, in addition to being resource- and energy-intensive. The existence of an inner representational layer places all the computational burden on the agent itself. Affordances, however, arise in the agent-environment interaction, offloading part of the processing burden onto the environment. Introducing affordance perception into autonomous agents enables them to continuously and dynamically adapt to shifting and changing environments.

Traditional autonomous agents separated sensing, planning/reasoning, and acting into different processes that would only link up at a later stage (Gat, 1998; Maes, 1991). The perceptual process sends information to the planning process, which in turn sends instructions for action (Horton, Chakraborty, & Amant, 2012). However, "[e]ven if an agent has perfect segmentation and feature recognition capabilities, this new form of information may be hard to translate into appropriate actions" (Nye & Silverman, 2012, p. 184). Affordance perception dispenses with the intermediary planning layer, instead generating affordances within a tight perception-action loop. What planning there may be is performed online through the perception-action loop, instead of offline between perception and action. This does not mean that the agent does no planning whatsoever; rather, it means there is often no representational layer mediating between perceptual and actional processes—at least, not at the level of basic perceptual processes (see Şahin et al., 2007). What planning there may be is performed online through the perception-action loop, instead of offline between perception and action. Furthermore, machine learning alone is insufficient; a robotic body is *required* for an affordance to be perceived. This is because affordances are not merely perceptual processes—they are perception-action processes and require dynamic engagement with the environment.

Natural and Cultural Affordances

Affordance perception in AI is complicated by the fact there are two very different types of affordances: natural and cultural. Natural affordances involve very basic cognitive processes. Cultural affordances are comparatively richer and involve culturally- and intersubjectively-mediated processes in order to be perceived and acted upon.[2] Cultural affordances, however, pose a particularly intractable problem. While natural affordances arise from the informational structure of the environment, cultural affordances require that the percipient be acculturated. There seems, *prima facie*, to be a level of decoupled, even representational, processing required to perceive a cultural affordance (Ramstead, Veissière, & Kirmayer, 2016).

Natural affordances are possibilities in the environment available for action (Dotov, Nie, & de Wit, 2012). Different agents can perceive different affordances based on their embodied capacities and species-typical behaviors. For example, a twig affords different actions to a cat, a finch, and a human. To the cat, the twig affords bite-ability and play-ability. To the finch, it affords graspability by the beak and build-ability for a nest. Finally, to the human, it affords manual manipulation. In each case, the embodied capacities and species-typical behaviors of the agent shape what kind of action the twig affords.

Affordances are based on the real information (light, pressure, scent) available in the environment. However, they do not themselves exist in the environment. They are generated in the agent-environment interaction. Affordance-perception occurs because the agent and environment form a complex, emergent system (Favela & Chemero, 2016; Thompson, Varela, & Rosch, 1991/2016; Gallagher, 2017; Thompson, 2007). That is, the agent is dynamically coupled with the environment. This coupling is modeled in ecological psychology and embodied cognition research using dynamical systems theory (Beer, 2014; Chemero, 2009; Turvey, 2019).

Several formalizations of affordances have been proposed (see Chemero, 2003; Stoffregen, 2003; Turvey, 1992). Stoffregen's (2003) formalization, which has been successfully utilized in AI affordance perception research (Nye & Silverman, 2012), is:

> "Let W_{pq} (e.g., a person-climbing-stairs system) = (X_p, Z_q) be composed of different things Z (e.g., person) and X (e.g., stairs). Let p be a property of X and q be a property of Z. The relation between p and q, p/q, defines a higher order property (i.e., a property of the animal−environment system), h. Then h is said to be an affordance of W_{pq} if and only if
>
> - W_{pq} = (X_p, Z_q) possesses h.
> - Neither Z nor X possesses h" (Stoffregen, 2003, p. 123).

Cultural affordances require the agent to utilize "explicit or implicit expectations, norms, conventions, and cooperative social practices" (Ramstead, Veissière, & Kirmayer, 2016, p. 3). It is precisely these elements that seem, *prima facie*, to require a representational layer decoupled from perception-action processes. For example, Gibson (1979/2015) remarks that a buyer and a seller each afford one another opportunities for action (viz., buying and selling). However, he goes on to say,

> "The perceiving of these mutual affordances is enormously complex, but it is nonetheless lawful, and it is based on the pickup of the information in touch, sound, odor, taste, and ambient light" (p. 127).

The information, in this case, is directly out there in the environment, and the agent perceives it. The affordances for

[2] Although affordances may differ regarding their basicness or their cultural scaffolding, in practice it is difficult to disentangle these two (see Wagman, Caputo, & Stoffregen, 2016). Indeed, for human agents, even basic perception-action processes like picking up an apple are culturally mediated.

action, however, arise in the interaction of the agent with its environment. It is the information, not the affordance, that is objectively embedded in the immediate environment. However, how could a buyer be perceived *as such* merely based on light waves, sound pressure waves, and other ecological information?

Gibson also claims that "the real postbox…affords letter-mailing to a letter-writing human in a community with a postal system" (1979/2015, p. 130). In this example, we have a culturally-scaffolded process of perception and action that functions only within a highly-specific cultural framework. It is not clear, however, how these culturally-scaffolded processes could be "directly" perceived based on the immediate information available in the environment. Memory and background knowledge are required for the postbox to be perceived with a letter-mailing affordance. However, there is little in the postbox's shape, color, and size that informs the agent of the postal system, letter-writing culture, and letter-reading agents enabling it to have mail-ability. Either cultural affordances are representational (see Ramstead, Veissière, & Kirmayer, 2016), or they must somehow be generated in a cultural milieu and for an acculturated agent by utilizing nonrepresentational, memory-based processes (see Rietveld & Kiverstein, 2014).

AI Cultural Affordance Perception

Most AI implementations of affordance perception have focused on natural affordances. These are, no doubt, relatively easier to implement because they do not require background knowledge of culture or a process of enculturation in order to perceive and act upon them. They are based only on the informational structure of the immediate environment. The true challenge for AI affordance perception is to achieve the perception of *cultural* affordances. If, however, cultural affordance perception turns out to require a robust representationalism, it is not clear that it has any advantage over non-affordance-based AI.

Raubal and Moratz (2008) provide an AI implementation of cultural affordance perception whereby cultural affordances are scaffolded onto natural affordances by representations of cultural knowledge. Their target agent is the Bremen Autonomous Wheelchair *Rolland*, which interprets linguistic commands by its human occupant and navigates across the environment. The need for cultural affordance perception arises because the wheelchair does not blindly perform actions commanded by their users. For example, the user may request to visit a center outside of operational hours. In this case, the AI utilizes cultural affordances integrating knowledge of the institution and its operating hours when selecting for action outputs.

Cultural affordances arise in their system by a system of constraints upon natural affordances. A natural affordance is constrained within a given social and institutional context. For example, the mailbox affords a multiplicity of actions, including smashing, opening, inserting objects, and touching. In their model, it is the social and institutional context of the postal system, letter-readers, and letter-writing that constrain the possible natural affordances into a smaller subset of cultural affordances. The agent then performs internal actions on these cultural affordances—essentially, planning or reasoning processes—in order to act upon the more basic natural affordance of opening and inserting.

The cultural affordances utilized by Raubal and Moratz' (2008) agent are representational. A separate planning layer is retained by their AI wheelchair. Their conception of cultural affordances is simply a subset of natural affordances that are given social and institutional constraints. Knowledge such as closing and opening hours is certainly representational and linguistically-based. The problem with their implementation of cultural affordances is that there is little that distinguishes them from classical representations. The construct of 'cultural affordance' is not doing any work that the construct of 'representation' does not already do. Their agent is essentially a hybrid system incorporating affordance perception for low-level navigation and symbolic representations for higher-level constraints upon that navigation.

Furthermore, some forms of social and institutional knowledge that Raubal and Moratz (2008) discuss, such as navigating across a city, are not necessarily fully representational processes. Unwritten norms such as walking on the right side of the sidewalk in many Western countries could be conceived of as representational rules. However, spontaneous pedestrian patterns can emerge without any specific intention (Moussaid et al., 2009).

Socialization and Supervised Learning

Awaad, Kraetzschmar, and Hertzberg (2015) provide an affordance-based model for AI agents that can "socialize" by learning expected uses of objects. The practical applications of this are in producing service robots that perform actions commanded by humans without being "robotic." When humans perform service tasks, an entire body of knowledge is brought to bear. Take the example of sweeping the floor. The human agent needs to know how to use a broom. However, the possibility space for utilizing a broom to sweep in *deviant* ways is quite large: one could sweep under the feet of others, sweep at the wrong times (e.g., while others are cooking), or sweep with furious movements and kick up dust. All these behaviors accomplish the task of sweeping but are social nuisances and perhaps even physically dangerous. There is an entire network of social expectations and etiquette surrounding the tool use in question. There is, in short, a "right way to do things." Furthermore, humans

> "effortlessly adapt our actions to unexpected situations, especially given the dynamic nature of our environment and the amount of uncertainty about it" (p. 422).

While moving a broom back and forth can be largely explained with natural affordances, these cultural constraints cannot. The broom affords more actions than are socially acceptable or considered appropriate to the task. Awaad,

Kraetzschmar, and Hertzberg (2015) attempt to integrate them within an affordance-perception paradigm, however. They note that programming procedural knowledge is insufficient to cover these cases of "the right way to do things" because the agent will always encounter novel situations. They implement Hierarchical Task Network to decompose tasks into a set of individual tasks in order to accomplish a goal.

In order to reduce the possibility space for action to one for socially-appropriate action, Awaad, Kraetzschmar, and Hertzberg (2015) store information about the object, the commanding agent, and the intended uses of the object. These constraints are scaffolded on the natural affordance of the object. The broom, for example, is defined by its socially-intended purpose of cleaning. The commanding agent, the human who demands cleaning, would have a set of preferences and expectations as to how that task is accomplished. The authors implement this cultural scaffolding through coded representations. Like Raubal and Moratz (2008), they conceive of cultural affordances simply as subsets of natural affordances that arise through representational cultural constraints.

Although the authors use representations to implement the socially-scaffolded constraints on the object's affordances, their broader proposal shows how a nonrepresentational framework could be used to do the same work. While they programmed the constraint knowledge into their agents, they suggest that this would be better done by supervised learning, particularly learning by demonstration. In the following section, I argue such supervised learning by demonstration of affordances does not require a strong concept of representations for its implementation.

Joint Interaction and Cultural Affordances: Unsupervised and Supervised Learning

Chu, Fitzgerald, and Thomaz (2016) develop autonomous agents that learn to perceive and use affordances through a combination of unsupervised and supervised learning through interaction with a human. A human teacher physically guides the robot to certain affordances. For example, a robot is taught that drawers have an openable affordance by guiding its hand. The robot learns to mimic this movement and perceive the openability affordance of the drawer's handle.

While the openability of the drawer *prima facie* appears to be a natural affordance provided by the structure of the robot's hand and the drawer's handle, there is a large possibility space for socially-deviant drawer-opening behavior. Although Chu, Fitzgerald, and Thomaz (2016) do not note this, the human teacher is not merely teaching the autonomous agent how to perceive and act upon the openability affordance of the drawer. They are simultaneously teaching the AI agent the *acceptable* way to perform this action. The drawer is not to be forcefully opened or rapidly opened and closed in succession (as a small child may annoyingly do), for example. The process of supervised learning allows the AI agent to learn the socially-acceptable affordances. This makes the drawer's openability affordance not simply natural, but also culturally-scaffolded.

Ramstead, Veissière, and Kirmayer (2016) invoke Gricean norms to understand these contexts. Grice (1975) articulated a set of rules governing conversation. These rules are ancillary to the communicative and phatic functions of language and facilitate nondeviant interactions. For example, one ought to convey as much detail as the topic requires without divulging too much detail. If one fails to do the former, one is perceived as terse, reticent, or uncommunicative. If one fails to do the latter, one is perceived as a windbag. In either case, deviation from the unwritten norm has the effect of interrupting the communicative act itself. Likewise, mundane actions have ancillary but unwritten norms guiding how they ought to be performed. These norms can only be learned by actual practice and observation of these actions in a social context. They are not symbolic rules because there may be no explicit representation of their content. They are merely habitual patterns of behavior used to accomplish certain tasks—e.g., opening a drawer slowly rather than forcefully.

While the robot may not develop shared intentions with the human teacher, in this case, it is significant that the robot only learns to perceive and act upon affordances through a process of interaction with a human (who is a "native" affordance perceiver-actor). In this case, representations are not necessary to explain how the AI agent learns to perceive and act upon the drawer's openability affordance in a socially-nondeviant way. While Awaad, Kraetzschmar, and Hertzberg (2015) programmed in cultural knowledge through representations, this supervised learning process does not specifically require the agent to store representations of cultural constraints, expectations, and other social rules. Rather than storing social rules and using them to constrain the agent's affordance perception and action, supervised learning allows the agent to learn to perceive and act upon the affordance in certain typical ways. Instead of inducing a rule based on the multiple supervised learning instances of opening the drawer—e.g., if drawer, then constraint *x*, *y*, *z*—the agent can simply follow the typical range of paths that have been learned.

One objection is that human agents are conscious of not deviating from socially-accepted norms of tool usage. These norms may be at a higher level than "not kicking up dust." One may be aware that one ought not to bother or annoy anyone. Nonetheless, even that does not require a specific rule. Even if the human agent has such a rule in mind, it is generally not the cause of their socially-nondeviant behavior. We do not walk around constantly thinking "I ought not to annoy *x*." If we can formulate such a rule, and even implement it in some cases, it is the exception (perhaps applying to a highly novel situation) rather than the norm. There is nothing here that cannot also be explained through processes of social learning, acculturation, and operant conditioning. These parallel the supervised learning trials in

AI affordance perception (Awaad, Kraetzschmar, & Hertzberg, 2015; Chu, Fitzgerald, & Thomaz, 2016).

AI Perception of Cultural Affordances without Representations: Learning and Habit

Representations such as rules can be used to constrain behavior in highly novel situations. Indeed, these kinds of rules may be part of the learning process itself. However, programming a database of representational cultural constraints for autonomous agents is a task just as formidable as that of traditional, non-affordance-based AI and computer vision. It is not clear that utilizing affordances in AI perception gains us anything. The problem, however, is that implementations of cultural affordance perception have generally been representational. The supervised learning in Chu, Fitzgerald, and Thomaz (2016) provides a way of thinking about what partially-nonrepresentational cultural affordances may look like in autonomous agents. Their autonomous agent learned how to open cabinet drawers in nondeviant and socially-acceptable ways. The drawer's handle information could afford multiple possibilities for action that are deviant, such as forcefully opening or rapidly opening and closing. During its supervised learning trials, the autonomous agent only learns the socially-acceptable way of opening the drawer. The agent does not perceive a natural affordance of open-ability. It perceives a cultural affordance of gentle-open-ability, one that is only salient within a given social structure and context.

Their autonomous agent does not have to learn or be preprogrammed with a *rule* about acceptable ways to open drawers. It is through multiple supervised learning trials that the cultural affordance begins to emerge—it is, in short, a *habit*. By habit, I mean a pattern of behavior that develops through supervised and unsupervised learning. The agent's habitual patterns of behavior are not representational in the sense that they are not primarily guided by symbolic rules, although the latter may constrain habits in actual behavior. Habit emerges from a set of previous behaviors and continues to guide future ones without necessarily having any explicit formulation. Surely some affordances must be constrained by symbolic representations. The closing time of a building or institution is something that could be learned by habit. The agent could develop a sense of when it closes by a long process of trial and error. However, that would be far less efficient than simply having a rule representing its closing time. In many cases, though, the work being done by representations can just as well be done by supervised and unsupervised learning or habit.

Returning to the example of sweeping, when the AI agent learns how to sweep from a human teacher, the latter will only teach the socially-accepted ways to sweep. The teacher will not teach how to sweep under people's feet, around them while walking, vigorously so as to kick up dust, or any other socially-deviant manner. The agent would learn these patterns of use of the object. Inducting a specific representational rule to cover these cases is supernumerary and fails to add explanatory value. The agent does not need a representational rule ("do not kick up dust"), because they have been taught to use the broom in a set of patterns that do not include kicking up dust. Following Ockham's razor, if the explanation can be had without representations, then we ought to dispense with them as an *explanans* in those cases.

One of the challenges for robustly-representationalist implementations of cultural affordance perception is that they require just as thorough programming with rules as traditional representational architectures. Habit, however, could greatly reduce the set of background knowledge that needs to be programmed. This is also a primary way that human agents develop habits during development. Children do not learn about their culture's interpersonal distance—the typical distance people stand from one another during communication—by learning a rule about how many centimeters away from another person to stand. They merely develop a habit of standing a certain distance away from another person. This habit is reinforced by observation of others and by violations of the norm (e.g., standing too close to someone can be perceived as aggressive). They may not even be aware that there is such a social norm guiding their behavior. If a representational rule happens to be extracted by a reflecting agent, it is still habit and not that rule that continues to guide its behavior. A humanoid autonomous agent could likewise learn to communicate using nondeviant interpersonal distance without any representational rules dictating how many centimeters away to stand by recourse to supervised learning (observation and mimicry) and unsupervised learning (violations).

Conclusion

Affordance perception offers a new paradigm for perception and action in autonomous agents. While traditional three-level systems dissociate perception, planning or reason, and action into separate layers, nonrepresentational affordances involve a dynamic and bidirectional perception-action loop with online planning. Many implementations of affordance perception in AI research have retained the representationalist paradigm even as they seek to integrate affordances. While this is certainly feasible from a technical standpoint, the construct of 'affordance' loses much of its power. An affordance-based robotics that remains largely representationalist has no clear advantage over traditional architectures.

Examining several implementations of cultural affordance perception in AI research, I argue that representations are not *necessary* for cultural affordances. I sketched a possible way for autonomous agents to implement cultural affordance perception by habit gained through supervised and unsupervised learning. AI implementations of affordance perception do not conceptually require a robust representationalism. If affordance-based robotics is to have any advantage over traditional architectures, it may need to reconsider the role of representations in cultural affordance perception.

Acknowledgements

The author would like to thank the Institute for Intelligent Systems for a travel grant to present this paper at the 41st Annual Meeting of the Cognitive Science Society.

References

Awaad, I., Kraetzschmar, G. K., & Hertzberg, J. (2015). The role of functional affordances in socializing robots. *International Journal of Social Robotics*, *7*, 421-438.

Beer, R. D. (2014). Dynamical systems and embedded cognition. In K. Frankish & W. M. Ramsey (Eds.), *The Cambridge handbook of artificial intelligence*. Cambridge University Press.

Braitenberg, V. (1984). *Vehicles: Experiments in synthetic psychology*. Cambridge, MA: MIT Press.

Brooks, R. A. (1990). Elephants don't play chess. *Robotics and Autonomous Systems*, *6*, 3-15.

Brooks, R. A. (1991). Intelligence without representation. *Artificial Intelligence*, *47*, 139-159.

Chemero, A. (2003). An outline of a theory of affordances. *Ecological Psychology*, *15*(2), 181-195.

Chemero, A. (2009). *Radical embodied cognitive science*. Cambridge, MA: MIT Press.

Chu, V., Fitzgerald, T., & Thomaz, A. L. (2016). Learning object affordances by leveraging the combination of human-guidance and self-exploration. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 221-228.

Dennett, D. (1969). *Content and consciousness*. New York, NY: Humanities Press.

Dotov, D. G., Nie, L., & de Wit, M. M. (2012). Understanding affordances: History and contemporary development of Gibson's central concept. *AVANT*, *3*(2), 28-39.

Favela, L. H., & Chemero, A. (2016). The animal-environment system. In Y. Coello & M. H. Fischer (Eds.), *Foundations of embodied cognition: Volume 1: Perceptual and emotional embodiment*. New York, NY: Psychology Press.

Fodor, J. (1985). Fodor's guide to mental representation: The intelligent auntie's vade-mecum. *Mind*, *94*(373), 76-100.

Gallagher, S. (2017). *Enactivist interventions: Rethinking the mind*. Oxford University Press.

Gat, E. (1998). On three-layer architectures. In D. Kortenkamp, R. P. Bonasso, & R. Murphy (Eds.), *Artificial intelligence and mobile robots: Case studies of successful robot systems*. Menlo Park, CA: American Association for Artificial Intelligence.

Gibson, J. J. (1979/2015). *The ecological approach to visual perception* (Classic Ed.). New York, NY: Psychology Press.

Grice, H. P. (1975). Logic and conversation. In J. L. Morgan & P. Cole (Eds.), *Syntax and semantics 3: Speech acts*. London, UK: Academic Press.

Horton, T. E., Chakraborty, A., & St. Amant, R. (2012). Affordances for robots: A brief survey. *AVANT*, *3*(2), 70-84.

Maes, P. (Ed.). (1991). *Designing autonomous agents: Theory and practice from biology to engineering and back*. Cambridge, MA: MIT Press.

Marr, D. (1982/2010). *Vision: A computational investigation into the human representation and processing of visual information*. Cambridge, MA: MIT Press.

Moussaid, M., Garnier, S., Theraulaz, G., & Helbing, D. (2009). Collective information processing and pattern formation in swarms, flocks, and crowds. *Topics in Cognitive Science*, *1*, 469-497.

Nye, B. D., & Silverman, B. G. (2012). Affordances in AI. In N. M. Seel (Ed.), *Encyclopedia of the sciences of learning*. New York, NY: Springer.

Ramstead, M. J. D., Veissière, S. P. L., & Kirmayer, L. J. (2016). Cultural affordances: Scaffolding local worlds through shared intentionality and regimes of attention. *Frontiers in Psychology*, *7*, 1-21.

Raubal, M., & Moratz, R. (2008). A functional model for affordance-based agents. In E. Rome, J. Hertzberg, & G. Dorffner (Eds.), *Towards affordance-based robot control: International seminar, Dagstuhl Castle, Germany, June 5-9, 2006. Revised papers*. Berlin, Germany: Springer.

Rietveld, E., & Kiverstein, J. (2014). A rich landscape of affordances. *Ecological Psychology*, *26*(4), 325-352.

Rome, E., Hertzberg, J., Dorffner, G., Doherty, P. (2006). Towards affordance-based robot control. In *Dagstuhl Seminar 06231 "Affordance-Based Robot Control," June 5-9, 2006*, Schloss Dagstuhl, Germany: Internationales Begegnungs- und Forschungszentrum für Informatik.

Şahin, E., Çakmak, M., Doğar, M. R., Uğur, E., & Üçoluk, G. (2007). To afford or not to afford: A new formalization of affordances towards affordance-based robot control. *Adaptive Behavior*, *15*(4), 447-472.

Stoffregen, T. A. (2003). Affordances as properties of the animal environment system. *Ecological Psychology*, *15*(2), 115-134.

Thompson, E. (2007). *Mind in life: Biology, phenomenology, and the sciences of mind*. Cambridge, MA: Belknap Press.

Turvey, M. T. (1992). Affordances and prospective control: An outline of the ontology. *Ecological Psychology*, *4*(3), 173-187.

Turvey, M. T. (2019). *Lectures on perception: An ecological perspective*. New York: Routledge.

Varela, F. J., Thompson, E., & Rosch, E. (1991/2016). *The embodied mind: Cognitive science and human experience* (Rev. Ed.). Cambridge, MA: MIT Press.

Wagman, J. B., Caputo, S. E., & Stoffregen, T. A. (2016). Hierarchical nesting of affordances in a tool use task. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(10), 1627-1642.

Neighborhood in Decay: Working Memory Modulates Effect of Phonological Similarity on Lexical Access

Karl David Neergaard (karl.neergaard@connect.polyu.hk)
Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University

James Britton (james.britton@connect.polyu.hk)
Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University

Chu-Ren Huang (churen.huang@polyu.edu.hk)
Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University

Abstract

A mainstay of models that account for the access of lexical knowledge is that auditory words compete for selection based on form similarity, commonly seen in an inhibitory effect to greater phonological neighborhood density (PND). PND is a metric that states that two words are neighbors if they differ by the addition, deletion or substitution of a single phoneme. A drawback to this account is that there is competing evidence even among the European languages investigated thus far. We sought to verify whether the inhibitory effect of greater PND would hold for Mandarin Chinese in two auditory word repetition tasks with monosyllabic and disyllabic Mandarin words. Results of Experiment 1 showed a facilitative effect to greater PND. Experiment 2 added a non-verbal distractor task to lessen the putative effect of working memory load during the task. The facilitative effect to greater PND was confirmed along with a significant post-hoc interaction with memory decay, operationalized as the duration spent on the distractor tasks. The facilitative effects extend previous reports of differential behavior due to linguistic typology.

Keywords: Lexical access; phonological neighborhood density; memory decay; Mandarin Chinese

Introduction

Essential to the current models of lexical processing is that target words interact during selection with items in long-term memory based on their shared semantic, orthographic, and phonological similarity. Both orthographic and phonological similarity are most commonly calculated through the addition, deletion or substitution of a single letter or phoneme (Landauer & Streeter, 1973). According to this metric, known as neighborhood density, a target stimulus with many similar words in the lexicon, i.e., neighbors, resides in a dense neighborhood, while a word with few similar words in the lexicon resides in a sparse neighborhood. The contrasting of dense and sparse words has been used to model the structural organization of lexical knowledge. For example, in the recognition of orthography, according to both the orthographic and phonological metrics, words from dense neighborhoods have been shown to facilitate recognition (Orthographic; e.g., Coltheart, Davelaar, Jonasson, & Besnar, 1977; Phonological: e.g., Yates, Locker, & Simpson, 2004). This facilitation has motivated the claim that greater density results in greater overall activation, a defining feature that was later implemented in several computational models (Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001; Grainger & Jacobs, 1996; Wagenmakers et al., 2004). However, when tasks are performed in the auditory modality, greater density has been reported to inhibit recognition (e.g., Luce & Pisoni, 1998; Vitevitch & Luce, 1998; Ziegler et al., 2003), motivating the construction of modality specific models of speech processing (Luce, Goldinger, Auer, & Vitevitch, 2000; McClelland & Elman, 1986; Norris, 1994). In response, Chen & Mirman (2012) constructed a connectionist model in an attempt to give a unified account of both facilitative and inhibitory neighborhood effects across both visual and auditory modalities, perception and production tasks, and orthographic, phonological and semantic neighborhood interactions. Their innovative approach, unfortunately, rests on the false assumption that there is a consensus in the literature for modality and task specific neighborhood effects.

To limit the discussion, we will consider only the effects known for phonological neighborhood density (PND). Two hypotheses of note have been advanced concerning differences in polarity from the body of behavioral evidence: psychotypology, and methodology.

The psychotypological argument holds that cognitive processes differ due to the linguistic differences between the languages being tested. The case, as it regards lexical access (Vitevitch & Rodríguez, 2004; Vitevitch & Stamer, 2006, 2009), was made based on evidence from both auditory recognition, and speech production. Dense phonological neighborhoods were inhibitory to speech recognition for English speakers (e.g., Luce & Pisoni, 1998), yet were facilitative for Spanish speakers (Vitevitch & Rodríguez, 2004), whereas, dense words were facilitative for English speakers in picture naming (Vitevitch, 2002), yet inhibitory for Spanish speakers (Vitevitch & Stamer, 2006, 2009). Vitevitch and colleagues speculated that the differences between the English and Spanish lexicons led to differences in polarity. Whereas English words have on average a greater number of shorter words with more phonological neighbors, the sparser Spanish vocabulary features words that are

neighbors both phonologically and semantically, e.g., *niño/niña* (boy/girl).

The methodological argument to account for differences in polarity points towards the design and methods employed in PND related studies. Sadat, Martin, Costa, and Alario, (2014) posited that the contradictory findings could be amended through testing with larger stimulus sets and the use of mixed effects models. The re-analyses of PND studies done by Sadat and colleagues clarified important differences between F tests and mixed effects models in the analysis of a variable that is continuous in nature and thus best fit for regression rather than factorial designs.

A third possibility to account for the differences in polarity in PND studies is to investigate working memory, specifically as it concerns the size of the stimuli sets used. As a participant recognizes or names a word, that lexical item is temporarily stored in working memory. If participants are exposed to multiple words, memory load increases and with it reaction times (Cohen et al., 1997; Jha & McCarthy, 2000). If the participant sustains attention on one task then memory decay does not happen at the same rate compared to if they were given a pause or a distractor task that does not interfere with the domain or modality of the main task (Rae & Perfect, 2014). In the case of phonological information, Baddeley (1986) found that phonological memory decayed within roughly 2 seconds. This does not differ greatly from the recall of orthographic letters after doing simple math problems (Brown, 1958; Peterson & Peterson, 1959). Given that during PND related tasks the inter-stimulus pause tends to be between 500-1500ms, i.e., under the rate of decay known to exist for phonological information, it is feasible that words are subject to cumulative activation, i.e., that activations of multiple word representations overlap and contribute to participant performance.

The studies that have investigated neighborhood effects amongst adults have utilized a large range of different sized stimulus sets. In studies implementing auditory word repetition tasks the story is quite straightforward, wherein large stimulus sets (Luce & Pisoni, 1998: 400 words; Vitevitch & Luce, 1998: 240 words) led to an inhibitory PND effect. In lexical decision tasks, two experiments using large stimulus sets showed inhibitory PND effects (Luce & Pisoni, 1998: 610 words; Vitevitch, Stamer, & Sereno, 2008: 112 words) while one with a small set of stimuli showed a facilitative PND effect (Vitevitch & Rodríguez, 2004: 80 words). The picture naming literature is where we see inhibitory results with large and small stimulus sets (e.g., Sadat, et al., 2014: 533 pictures; Vitevitch & Stamer, 2006: 48 pictures), facilitative results with small stimulus sets (Baus, Costa, & Carreiras, 2008: 48 pictures; Marian, Blumenfeld, & Boukrina, 2008: 57 pictures; Pérez, 2007: 89 pictures; Vitevitch, 2002: 48 pictures; Vitevitch & Stamer, 2009: 48 pictures), and non-significant PND effects (Jescheniak & Levelt, 1994: 96 pictures repeated 3 times; Vitevitch et al., 2004: 44 pictures). Note that non-significant results might also have been due to issues unrelated to stimuli number, such as mixing photographic and hand-drawn stimuli or due to naming pictures that represent conceptual processes such as verbs (Newman & Bernstein Ratner, 2007; Tabak, Schreuder, & Baayen, 2010).

The role of PND in working memory has not been fully explored. The only studies to test their interaction found a facilitation of greater density in serial recall tasks with English speakers (Oberauer, 2009; Roodenrys, Hulme, Lethbridge, Hinton, & Nimmo, 2002). The facilitative effect was said to be due to redintegration, which can be described as the restoration of short-term memory traces due to long-term memory representations. In order to test cumulative activation, however, it is necessary to account for overall memory load rather than that of isolated words.

To test the possibility that cumulative working memory influences the directional effect of PND, we performed two experiments with a large number of stimuli. In Experiment 1 we presented the full stimuli set to our participants without sufficient time for decay in memory load. In Experiment 2 we inserted three nonverbal distractor tasks in order to introduce memory decay. In both experiments we implemented the auditory word repetition task due to it being the only task thus far without contradictory findings. The cumulative memory hypotheses, allows for the prediction of an inhibitory effect to greater PND in Experiment 1 and a facilitative or null effect in Experiment 2.

In the current study we also incorporate concerns brought by both previous hypotheses concerning methodology and psychotypology on differing PND effects. Through the use of a large stimuli set and mixed effects models we treated PND as a continuous variable. Meanwhile, our target language, Mandarin Chinese, was chosen due to its typological distinctness to either English or Spanish, allowing for a unique view on how the dimensions of the lexicon affects lexical access.

The Mandarin vocabulary differs from both English and Spanish in critical ways. Its syllable inventory, when including lexical tone, has roughly 1,300 items, which is a number far less than the 10,000+ English syllables. Unlike Spanish that boasts of a large proportion of multisyllabic words rich in morphological variation (Arbesman, Strogatz, & Vitevitch, 2010), roughly 72% of Mandarin's phonological words (i.e., in which all homophones are collapsed to one item) are disyllabic, and only 3.8% monosyllabic (Neergaard & Huang, 2019). Meanwhile, Mandarin words have little to no morphology. For instance, unlike Spanish, Mandarin verbs do not conjugate, and nouns do not note gender nor number.

To date, no results have been reported in the auditory word repetition task with Mandarin speakers. Despite this lack of prior evidence, the psychotypology account allows for certain predictions. Given the greater distance lexically from Spanish, particularly in relation to its on average longer word length and inflectional morphology, Mandarin shares greater similarity with English. English has shown inhibitory PND effects in both auditory lexical decision and word repetition, making it likely that lexical competition best accounts for the selection of dense phonological words in Mandarin.

Experiment 1

Methods

Participants Thirty-three native-Mandarin speakers participated in this experiment (Female: 21; Ages 18-35, M: 12, SD: 3.64). None of the participants reported speech, hearing, or visual disorders. All participants reported native-level proficiency in Mandarin.

The current study design was approved by The Hong Kong Polytechnic University's Human Subjects Ethics Sub-committee (reference number: HSEARS20140908002). The participants gave their informed consent and were compensated with 50HKD for their participation.

Stimuli The auditory stimuli for this experiment consisted of 154 Mandarin words (10 practice; 144 test). A female native-speaker of Mandarin from Fujian province produced all of the stimuli by speaking at a normal speaking rate into a high-quality microphone. Stimuli fell into 4 categories according to their syllable or segment length: 36 3-segment monosyllables with a CVN syllable structure (e.g., *san1*); 36 4-segment monosyllables with a CGVN syllable structure (e.g., *bian3*); 36 3-segment disyllables with a CV V syllable structure (e.g., *da4 yi1*); 36 4-segment disyllables with a CV CV syllable structure (e.g., *li4 shi3*). The 144 test stimuli were made from 20 syllable onsets, whose distributions were not significantly different in syllable length ($p=1$) or segment length ($p=1$). Eleven stimuli sets were constructed, each where stimuli were pseudo-randomized such that there were no consecutive presentations of items with the same onset or lexical tone (first syllables for disyllabic words). The stimuli list can be seen in Appendix 1.

Because the current stimuli consist of monosyllables and disyllables of both 3 and 4 segments in length, it was not possible to control for their durations along all 4 dimensions. Instead, stimuli were chosen in order to minimize durational differences between 3-Segment words (CV V, M: 609.25; SD: 11.59; CVN, M: 609.00; SD: 11.01) and between 4-Segment words (CVCV, M: 784.67; SD: 9.25; CVVN, M: 784.17; SD: 11.02). Stimuli did not differ within their respective segment length groups, but were significantly different across segment lengths ($F=9433$, $p<0.001$). Thus, while a significant difference in reaction times is expected between 3- and 4-Segment words, the same cannot be said between monosyllable and disyllables belonging to their respective segment lengths, which is critical in identifying whether monosyllables and disyllables are processed in an equivalent manner.

Lexical statistics for the stimuli were taken from the Database of Mandarin Neighborhood Statistics (Neergaard, Xu, & Huang, 2016), in which lexical frequency is derived from the wordlist of Subtlex-CH (Cai & Brysbaert, 2010) according to the summed subtitle frequency for each phonological word. All relevant statistics were calculated from 30,000 phonological words. In order to test the hypothesis that words of varying unit sizes are subject to the effect of phonological similarity during speech processing, it was necessary to use the fully segmented Mandarin syllable schema (C_G_V_X_T) because it allowed us to control for both segment and syllable length while distinguishing between words according to lexical tone. Stimuli did not differ in log10 lexical frequency for either 3-segment words (CVN, M: 3.08, SD: 0.40; CV V, M: 2.75, SD: 0.46) or 4-segment words (CGVN, M: 2.98, SD: 0.49; CV CV, M: 3.33; SD: 0.21) according to both segment length ($p=0.869$) and syllable length ($p=0.981$).

The remaining variables are of the density variety and include PND, log10 neighborhood frequency (NF, M: 3.11; SD: 1.06), and homophone density (HD, M: 1.67; SD: 1.26).

The goal in choosing stimuli according to PND, knowing that greater length negatively correlates with higher density, was to assure that there was sufficient spread for each group. For the syllable length group, disyllabic words had a spread of 0-11 neighbors (M: 3.71, SD: 2.45), while monosyllabic words had a spread of 4-25 neighbors (M: 13.29, SD: 5.05). For the segment length group, 3-segment words had a spread of 0-25 neighbors (M: 10.10, SD: 7.03), while 4-segment words had a spread of 0-17 neighbors (M: 6.90, SD: 4.85).

Procedure Participants sat in a quiet room in front of a computer running E-Prime 2.0 (Psychology Software Tools, 2012). They were instructed to repeat the words they heard over headphones into an attached microphone as fast as possible. Each trial began with a cross '+', in the center of the screen for 1000ms. Next, the onset of the target audio was presented concurrent with the exposure of a blank screen. A PST Serial Response Box was activated by the participants' voice, dependent on their response, which then led to a pause of 1000ms and the end of a trial. Stimuli were pseudo-randomized such that no two items were presented sequentially with the same onset or lexical tone. The entire experiment lasted roughly 10 minutes. Participants were given a practice set of 10 words prior to beginning the experiment.

Results and discussion

Reaction times were measured offline using SayWhen (Jansen & Watter, 2008). No participants were excluded due to excessive error rates, or deviant reaction times. Three stimuli were removed for error rates higher than 25% (*guang4, qing3, san4*). From the new total of 4,653 trials, 102 were removed due to production errors, accounting for 2.19%. A further 238 trials (5.23%) were removed for values below the duration of our shortest stimuli (577ms), and for values 2.5 standard deviations above the group mean. The final number of trials to be analyzed were 4,313 (M: 917ms; SD: 144ms).

As can be seen in Table 1, Subject and Item were placed in the random effects, while each of the density variables (PND, HD, and NF) were analyzed according to the two levels of segment length (SegLen: 3-seg, 4-seg). We also place syllable length (SyLen) into the fixed effects structure to evaluate whether there was a processing cost despite stimulus durations not being different between monosyllables and disyllables in each segment group.

Table 1. Model estimates for Experiment 1

Random effects	Var.	SD			
Subject	0.003	0.057			
Item	0.012	0.111			
Residual	0.007	0.086			
Fixed effects	**β**	**SE**	**df**	**t**	**p**
Intercept	9.4^{e-1}	2.2^{e-2}	54	42.21	< 0.001
SyLen (di)	-5.2^{e-2}	1.9^{e-2}	132	-2.73	0.007
3-seg:PND	-3.3^{e-2}	9.4^{e-3}	132	-3.48	< 0.001
4-seg:PND	-5.2^{e-2}	1.4^{e-2}	132	-3.63	< 0.001
3-seg:HD	6.5^{e-3}	6.1^{e-3}	132	1.07	0.287
4-seg:HD	7.0^{e-4}	1.1^{e-2}	132	0.06	0.950
3-seg:NF	-1.6^{e-2}	1.1^{e-2}	132	-1.51	0.133
4-seg:NF	-7.8^{e-3}	1.1^{e-2}	132	-0.70	0.483

Results revealed that monosyllables (M: 903ms; SD: 15ms) were produced significantly faster than disyllables (M: 928ms; SD: 14ms). Greater PND was facilitative for both 3-segment (M: 867ms; SD: 138ms) and 4-segment (M: 965ms; SD: 137ms) items. No effects were found for either SegLen:HD or SegLen:NF. An estimate of r^2, using the 'r2glmm' package in R (Jaeger, Edwards, Das, & Sen, 2016), revealed that the model had a marginal r^2 of 0.224, and semi-partial r^2 of 0.145 for SegLen:PND, and 0.053 for SyLen.

The facilitative effect to greater PND for both monosyllabic and disyllabic words, rather than supporting a cumulative memory account, is suggestive that typological differences between English (majority inhibitory findings to greater PND) and Mandarin led to the differential performance.

Experiment 2

The premise of the cumulative memory account is that shorter stimulus sets in a naming task result in facilitative PND effects due to there being fewer lexical items stored in working memory when compared to a task with a large stimulus set. It is possible that working memory builds cumulatively leading to increased activation, but that due to the particular psychotypological features of Mandarin, a facilitative effect to greater PND is the outcome. The only way to verify the status of a facilitative effect, while also nullifying the cumulative account, is to provide participants with sufficient time for memory decay during naming.

Methods

Participants Forty-seven native-Mandarin speakers participated in this experiment (Female: 29; Ages 19-38, M: 24, SD: 4). None of the participants reported speech, hearing, or visual disorders.

Stimuli The same auditory stimuli from Experiment 1 were used in this experiment.

Procedure The current design differed from Experiment 1 in that the experiment was partitioned into 4 blocks of 36 trials each with three interleaved distractor tasks. Each distractor task included four basic math questions: e.g., "20*2=__". The distractor task was self-paced. Participants had to press a button to return to the following test block. The entire experiment took less than 15 minutes.

Results and discussion

Reaction times were again measured offline using SayWhen (Jansen & Watter, 2008). Three participants were excluded from the analysis; two for reaction times 2.5 standard deviations above the group mean, and one due to experimenter error in data acquisition. No participants were excluded due to excessive error rates; however, three stimuli were removed for error rates higher than 25% (*qing3, san4, sang1*). From the new total of 6,203 trials, 142 were removed due to production errors, accounting for 2.24%. A further 103 trials (1.66%) were removed for values below 577ms and above 1446ms, leaving our final number of trials to be analyzed at 6,100 (M: 1010ms; SD: 148ms).

The same model configuration from Experiment 1, as shown in Table 2, again found a significant SyLen effect between monosyllables (M: 995ms; SD: 148ms) and disyllables (M: 1027ms; SD: 148ms), and a significant facilitative effect to greater PND for both 3-segment (M: 951ms; SD: 139ms) and 4-segment (M: 1069ms; SD: 134ms) items, with no significant effects for SegLen:HD or SegLen:NF. The model's marginal r^2 was 0.202, with a semi-partial r^2 of 0.121 for SegLen:PND, and 0.042 for SyLen.

Table 2. Model estimates for Experiment 2

Random effects	Var.	SD			
Subject	0.004	0.065			
Item	0.009	0.094			
Residual	0.008	0.091			
Fixed effects	**β**	**SE**	**df**	**t**	**p**
Intercept	1.04	1.9^{e-2}	115	55.22	< 0.001
SyLen (di)	-5.1^{e-2}	2.1^{e-2}	133	-2.41	0.017
3-seg:PND	-3.4^{e-2}	1.1^{e-2}	133	-3.19	0.002
4-seg:PND	-5.2^{e-2}	1.6^{e-2}	133	-3.25	0.001
3-seg:HD	6.1^{e-4}	7.0^{e-3}	133	0.09	0.931
4-seg:HD	-5.8^{e-3}	1.2^{e-2}	133	-0.46	0.643
3-seg:NF	-1.9^{e-2}	1.2^{e-2}	133	-1.57	0.118
4-seg:NF	-5.4^{e-3}	1.3^{e-2}	133	-0.43	0.667

In this experiment we confirmed the facilitative effect to greater PND for Mandarin. We have also shown that stimulus set sizes are not the likely candidates in the variability found in PND studies. We did not however account for how PND and working memory interact.

To investigate whether decay modulates the effect of PND, in a post-hoc analysis we operationalized memory decay as the time spent on the three interleaved distractor tasks. While each participant received the same basic math questions, they were given as much time as they saw fit to complete each task before returning to the repetition task. For the following analysis, it was necessary to exclude the trials belonging to the experiment's first block. In this way, each block under examination entailed auditory lexical processing after having received a limited time for memory decay from a previous session of auditory lexical processing.

The values for Decay ranged from as short as 3 seconds (3100ms) to as long as 41 seconds (41373ms). Visual inspection of Decay's token values revealed that it was not linearly distributed. We rescaled the variable using a Box Cox transformation (Tukey, 1977) to evenly distribute duration length of non-lexical processing during the distractor task.

Using the 'mcgv' package in R (Wood, Scheipl, & Faraway, 2013), a generalized additive model using tensor product smooths was constructed (with Subject and Item as random effects) in which Decay was added as an interaction to each level of PND, SegLen, and SyLen. With an adjusted r^2 of 0.651, Decay interacted significantly with PND ($F=23.94$; $p<0.001$); 3-segment ($F=28.75$; $p<0.001$) and 4-segment items ($F=15.62$; $p<0.001$); and both monosyllables ($F=6.32$; $p<0.001$) and disyllables ($F=8.28$; $p<0.001$). As can be seen in Figure 1, when Decay was shortest, the effect of PND was strongest, providing clear evidence that working memory is a determining factor of phonological neighborhood effects.

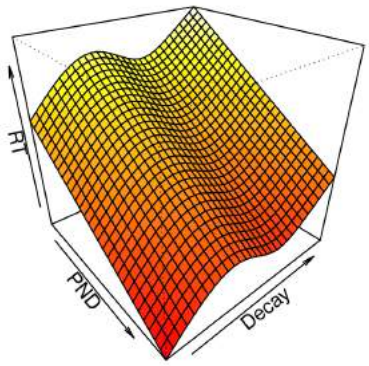


Figure 1: Interaction effect of phonological neighborhood density (PND) and time spent on the distractor task (Decay)

General Discussion

The purpose of the current study was to examine multiple hypotheses on the directional effect of PND through Mandarin Chinese, a language typologically distinct from the languages tested to date. We incorporated the methodological concerns brought by Sadat et al. (2014), through the testing of a large number of stimuli with mixed effects models wherein PND was treated as a continuous variable. We performed two auditory word repetition tasks, the only task to date that has not shown contradictory PND effects, in which our participants' rate of memory decay was manipulated to test whether differences in PND polarity have been due to cumulative memory. Finally, the testing of Mandarin participants allowed us to join the debate on psychotypology, i.e., whether the dimensions of a speaker's lexicon can result in differential behavioral outcomes.

In Experiment 1 we exposed our participants to the full stimuli set under the assumption that by not allowing for memory decay to occur our participants would produce an inhibitory effect to greater PND. Opposite our expectations, and in contrast to the previous English results, we found a facilitative effect to PND. In Experiment 2 we manipulated the task through the introduction of interleaved nonverbal distractor tasks. Changes in modality through distractor tasks have been shown to increase memory decay of the main task material (Rae & Perfect, 2014). The facilitative effect to greater PND was confirmed despite providing our participants with time for memory decay. A further post-hoc analysis illustrated that while working memory indeed modulates the phonological neighborhood effect, it can do so without lexical competition.

Under the assumptions of the psychotypology account of PND (Vitevitch & Rodríguez, 2004; Vitevitch & Stamer, 2006, 2009), we predicted that our Mandarin participants would experience lexical competition due to greater PND, in line with previous English results and contrary to previous Spanish results. This assumption was built on the greater difference between the Spanish and Mandarin vocabularies compared to the differences between the English and Mandarin lexicons. While Spanish is rich in morphology and on average has longer words, Mandarin has on average shorter words and no inflectional morphology. Contrary to our prediction, our Mandarin speakers were facilitated by greater PND, revealing that word length and inflectional morphology are likely not a reason for why Spanish speakers also experienced facilitation by words from dense phonological neighborhoods.

Given our current negation of the cumulative memory account, further work would benefit from delving deeper into the psychotypology of lexical access. Evidence has been mounting for differences in brain areas during language process between English and Mandarin speakers, at the level of whole-brain maps (Wu et al., 2015), and targeting language processing areas during tasks such as rhyming judgments (Brennan, Cao, Pedroarena-Leal, McNorgan, & Booth, 2013). A comparison of the Spanish and Mandarin lexicons might reveal how similarities between typologically distinct languages can lead to outcomes that defy the current models of speech production and perception.

It is also possible that an influence other than phonological neighborhoods is responsible for the facilitative effects in Spanish and Mandarin. The work of Vitevitch and colleagues pointed to a possible candidate other than word length and inflectional morphology, namely, neighbors of target words that are of both phonological and semantic relations (i.e., 'boy/girl, *niño/niña*). Our search of the literature found evidence concerning possible effects of semantic neighbors during auditory lexical decision, but only for English (Goh, Yap, Lau, Ng, & Tan, 2016; Tucker et al., 2018). In line with the current predictions, semantic neighbors did not significantly predict reaction times. In contrast to English, it is possible that both Spanish and Mandarin feature a sufficient number of phono/semantic neighbors to lead to facilitation during an auditory task. While Mandarin does feature phono/semantic word pairs such as *bian1* ('side' 边) / *pian1* ('one-sided' 偏), it also entertains a uniquely high level of homophony (*bian1* = 9 homophones; *pian1* = 3

homophones), making future comparisons between the two languages challenging.

References

Arbesman, S., Strogatz, S. H., & Vitevitch, M. S. (2010). Comparative analysis of networks of phonologically similar words in English and Spanish. *Entropy*, *12*(3), 327–337. https://doi.org/10.3390/e12030327

Baddeley, A. D. (1986). *Working memory* (Oxford psy). Oxford: Oxford University Press.

Baus, C., Costa, A., & Carreiras, M. (2008). Neighbourhood density and frequency effects in speech production: A case for interactivity. *Language and Cognitive Processes*, *23*(6), 866–888.

Brennan, C., Cao, F., Pedroarena-Leal, N., McNorgan, C., & Booth, J. R. (2013). Reading acquisition reorganizes the phonological awareness network only in alphabetic writing systems. *Human Brain Mapping*, *34*(12), 3354–3368. https://doi.org/10.1002/hbm.22147

Brown, J. (1958). Some tests of the decay theory of immediate memory. *The Quarterly Journal of Experimental Psychology*, *10*, 12–21.

Chen, Q., & Mirman, D. (2012). Competition and cooperation among similar representations: Toward a unified account of facilitative and inhibitory effects of lexical neighbors. *Psychological Review*, *119*(2), 417–430. https://doi.org/10.1037/a0027175

Cohen, J. D., Perlstein, W. M., Braver, T. S., Nystrom, L. E., Noll, D. C., Jonides, J., & Smith, E. E. (1997). Temporal dynamics of brain activation during a working memory task. *Nature*.

Coltheart, M., Davelaar, E., Jonasson, J. T., & Besnar, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.), *Attention and performance VI* (pp. 535–555).

Coltheart, Max, Rastle, K., Perry, C., Langdon, R., & Ziegler, J. C. (2001). DRC: A Dual Route Cascaded Model of Visual Word Recognition and Reading Aloud. *Psychological Review*, *108*(1), 204–256.

Goh, W. D., Yap, M. J., Lau, M. C., Ng, M. M. R., & Tan, L. (2016). Semantic Richness Effects in Spoken Word Recognition : A Lexical Decision and Semantic Categorization Megastudy. *Frontiers in Psychology*, *7*(Article 976), 1–10.

Grainger, J., & Jacobs, A. M. (1996). Orthographic processing in visual word recognition: a multiple read-out model. *Psychological Review*, *103*(3), 518–565.

Jaeger, B. C., Edwards, L. J., Das, K., & Sen, P. K. (2016). An R 2 statistic for fixed effects in the generalized linear mixed model. *Journal of Applied Statistics*, *44*(6), 1086–1105.

Jansen, P. A., & Watter, S. (2008). SayWhen: an automated method for high-accuracy speech onset detection. *Behavior Research Methods*, *40*(3), 744–751.

Jescheniak, J. D., & Levelt, W. J. M. (1994). Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(4), 824–843.

Jha, A. P., & McCarthy, G. (2000). The influence of memory load upon delay-interval activity in a working-memory task: an event-related functional MRI study. *Journal of Cognitive Neuroscience*, *12*, 90–105.

Landauer, T. K., & Streeter, L. A. (1973). Structural differences between common and rare words: failure of equivalence assumptions for theories of word recognition. *Journal of Verbal Learning and Verbal Behavior*, *12*, 119–131.

Luce, P. A., Goldinger, S. D., Auer, E. T., & Vitevitch, M. S. (2000). Phonetic priming, neighborhood activation, and PARSYN. *Perception & Psychophysics*, *62*(3), 615–625. https://doi.org/10.3758/BF03212113

Luce, P. A., & Pisoni, D. B. (1998). Recognizing Spoken Words: The Neighborhood Activation Model. *Ear and Hearing*, *19*(1), 1–36.

Marian, V., Blumenfeld, H. K., & Boukrina, O. V. (2008). Sensitivity to Phonological Similarity Within and Across Languages. *Journal of Psycholinguisic Research*, *37*(3), 141–170.

McClelland, J. L., & Elman, J. L. (1986). The TRACE Model of Speech Perception. *Cognitive Psychology*, *18*(1), 1–86. https://doi.org/10.1016/0010-0285(86)90015-0

Neergaard, K. D., & Huang, C. (2019). Constructing the Mandarin phonological network: novel syllable inventory used to identify schematic segmentation. *Complexity*, (Article 6979830), 1–21.

Neergaard, K. D., Xu, H., & Huang, C.-R. (2016). Database of Mandarin neighborhood statistics. In *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016*.

Newman, R. S., & Bernstein Ratner, N. (2007). The Role of Selected Lexical Factors on Confrontation Naming Accuracy, Speed, and Fluency in Adults Who Do and Do Not Stutter. *Journal of Speech, Language, and Hearing Research*, *50*(February), 196–213.

Norris, D. (1994). Shortlist - a Connectionist Model of Continuous Speech Recognition. *Cognition*, *52*(3), 189–234.

Oberauer, K. (2009). Interference between storage and processing in working memory: Feature overwriting, not similarity-based competition. *Memory & Cognition*, *37*(3), 346–357.

Pérez, M. A. (2007). Age of acquisition persists as the main factor in picture naming when cumulative word frequency and frequency trajectory are controlled. *The Quarterly Journal of Experimental Psychology*, *60*(1), 32–42. https://doi.org/10.1080/17470210600577423

Peterson, L., & Peterson, M. J. (1959). Short-term retention of individual verbal items. *Journal of Experimental Psychology*, *58*, 193–198.

Rae, P. J. L., & Perfect, T. J. (2014). Visual distraction during word-list retrieval does not consistently disrupt memory. *Frontiers in Psychology*, *5*(April), 362.

Roodenrys, S., Hulme, C., Lethbridge, A., Hinton, M., & Nimmo, L. M. (2002). Word-frequency and

phonological-neighborhood effects on verbal short-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(6), 1019–1034.

Sadat, J., Martin, C. D., Costa, A., & Alario, F. X. (2014). Reconciling phonological neighborhood effects in speech production through single trial analysis. *Cognitive Psychology*, *68*, 33.58.

Tabak, W., Schreuder, R., & Baayen, R. H. (2010). Producing in inflected verbs: A picture naming study. *The Mental Lexicon*, *5*(1), 22–46.

Tucker, B. V., Brenner, D., Danielson, D. K., Kelley, M. C., Nenadić, F., & Sims, M. (2018). The Massive Auditory Lexical Decision (MALD) database. *Behavior Research Methods*, 1–18.

Tukey, J. W. (1977). *Exploratory data analysis* (1st ed.). Pearson.

Vitevitch, M. S. (2002). The influence of phonological similarity neighborhoods on speech production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(4), 735–747.

Vitevitch, M. S., Armbruster, J., & Chu, S. (2004). Sublexical and lexical representations in speech production:effects of phonotactic probability and onset density. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(2), 514–529.

Vitevitch, M. S., & Luce, P. A. (1998). When words complete: Levels of processing in perception of spoken words. *Psychological Science*, *9*(4), 325–329.

Vitevitch, M. S., & Rodríguez, E. (2004). Neighborhood density effects in spoken word recognition in Spanish. *Journal of Multilingual Communication Disorders*, *3*(1), 64–73.

Vitevitch, M. S., & Stamer, M. K. (2006). The curious case of competition in Spanish speech production. *Language and Cognitive Processes*, *21*(6), 760–770.

Vitevitch, M. S., & Stamer, M. K. (2009). *The influence of neighborhood density (and neighborhood frequency) in Spanish speech production: A follow-up report*.

Vitevitch, M. S., Stamer, M. K., & Sereno, J. A. (2008). Word length and lexical competition: Longer is the same as shorter. *Language and Speech*, *51*(4), 361–383.

Wagenmakers, E.-J., Steyvers, M., Raaijmakers, J. G. W., Shiffrin, R. M., van Rijn, H., & Zeelenberg, R. (2004). A model for evidence accumulation in the lexical decision task. *Cognitive Psychology*, *48*, 332–367.

Wood, S. N., Scheipl, F., & Faraway, J. J. (2013). Straightforward intermediate rank tensor product smoothing in mixed models. *Statistics and Computing*, *23*(3), 341–360.

Wu, J., Lu, J., Zhang, H., Zhang, J., Yao, C., Zhuang, D., … Zhou, L. (2015). Direct evidence from intraoperative electrocortical stimulation indicates shared and distinct speech production center between Chinese and English languages. *Human Brain Mapping*, *36*(12), 4972–4985. https://doi.org/10.1002/hbm.22991

Yates, M., Locker, L., & Simpson, G. B. (2004). The influence of phonological neighborhood on visual word perception. *Psychonomic Bulletin & Review*, *11*(3), 452–457. https://doi.org/10.3758/BF03196594

Ziegler, J. C., Muneaux, M., & Grainger, J. (2003). Neighborhood effects in auditory word recognition: Phonological competition and orthographic facilitation. *Journal of Memory and Language*, *48*(4), 779–793.

Why do you take that route?

Alimire Nabijiang,[1] Supratik Mukhopadhyay,[1] Yimin Zhu,[2]
Ravindra Gudishala,[3] Sanaz Saeidi,[2] Qun Liu[1]
[1]Department of Computer Science and Engineering
[2]Department of Construction Management
[3]Department of Civil and Environmental Engineering and Louisiana Transportation Research Center
Louisiana State University, Baton Rouge, LA 70803
{anabij1, supratik, yiminzhu, ssaeid1, rgudis, qliu14 }@lsu.edu

Abstract

The purpose of this paper is to determine whether a particular context factor among the variables that a researcher is interested in causally affects the route-choice behavior of drivers. To our knowledge, there is limited literature that consider the effects of various factors on route choice based on causal inference.Yet, collecting data sets that are sensitive to the aforementioned factors are challenging and the existing approaches usually take into account only the general factors motivating drivers route choice behavior. To fill these gaps, we carried out a study using Immersive Virtual Environment (IVE) tools to elicit drivers route choice behavioral data, covering drivers' network familiarity, education level, financial-concern, etc, apart from conventional measurement variables. Having context-aware, high-fidelity properties, IVE data affords the opportunity to incorporate the impacts of human-related factors into the route choice causal analysis and advance a more customizable research tool for investigating causal factors on path selection in network routing. This causal analysis provides quantitative evidence to support drivers diversion decision. The study also provides academic suggestion and reference for investing in public infrastructure and developing efficient strategies and policies to mitigate traffic congestion.

Keywords: Causal and Counterfactual Explanation

Introduction

Route choice refers to the choices of roads among a set of possible alternatives made by human drivers while navigating through an urban area. Route Choice Models estimate the route choices of drivers in an urban setting. Most route choice models connect characteristics of alternate routes to those selected by the drivers. These models help in estimating traffic levels on different routes and thus enable development of effective traffic management strategies that can reduce traffic delays and allow maximum utilization of transportation systems. Existing route choice models use revealed preference behavior to model route choice. The use of revealed choice data limits the accuracy of the prediction as it fails to capture subjective context factors of drivers at individual level. Therefore, it is essential to use a data collection methodology that incorporates the importance of contextual factors in route choice.

As commuters we all make choices on which route to use when traveling to work, school, or shopping mall. Most of the times we pick a route that is familiar and also minimizes travel time. However, there is plenty of evidence that as commuters we take routes that do not minimize travel time (Ben-Akiva, Ramming, & Bekhor, 2004). In order to try and explain this route choice behavior, transportation engineers have been studying the route choice behavior of drivers for the past three decades to try and explain it. Transportation researchers have adopted econometric based approaches and used two types of data to mathematically model and rationalize the route choice behavior (Prato, 2009).

The data used in modeling route choices is collected by using two approaches. The first approach is based on actual observed route choice behavior that is often labeled as Revealed Preference data. The second approach is based on collecting data from hypothetical choice experiments that is often called as State Choice data (Ben-Elia & Shiftan, 2010). There are times when both types of data are combined to model and explain route choice behavior. However, the combination of econometric approaches and different data collection methods have yielded mixed results in explaining route choice behavior.

Based on the literature reviewed we believe that there is not much research that have tried to apply causal analysis methods to explain route choice behavior. We believe that by applying causal analysis techniques we can identify root causes that influence route choice and will subsequently allow us to enhance Route Choice models that will better forecast traffic levels on transportation networks and also to better comprehend drivers response to route guidance and dynamic message signs.

The main objective of this paper is to conduct causal analysis of route choice behavior using data collected from a Stated Choice Experiment in an Immersive Virtual Environment (IVE). We carried out a study using IVE tools to elicit drivers route choice behavioral data, covering drivers network familiarity, education level, financial- concern, etc, other than conventional measurement variables. Having context-aware, high-fidelity properties, IVE data affords the opportunity to incorporate the impacts of human-related factors into the route choice causal analysis and advance a more customizable research tool for investigating causal factors on path selection in network routing. This causal analysis provides quantitative evidence to support drivers diversion decision. The study also provides academic suggestion and reference for investing in public infrastructure and developing efficient strategies and policies to mitigate traffic congestion.

This paper makes the following contribution:

- To the best of our knowledge, the paper presents the first causal analysis of route choice behavior of drivers using data collected from a Stated Choice Experiment in an Immersive Virtual Environment (IVE).

Related Work

Transportation engineers have been studying commuter route choice behavior for four decades now. Engineers developing route choice models theorized that travel time plays a crucial and important role in the selection of a route. Route choice behavior theories began to evolve in the late eighties and early nineties as engineers' understanding of route choice behavior improved by studying data about empirical route choice behavior. Pursula and Talvite (Pursula & Talvitie, 1993) developed a mathematical route model by postulating that drivers do consider other factors apart from travel time in making a route choice. In (Khattak, Schofer, & Koppelman, 1993), the authors discovered that commuters prefer to use habitual routes when traveling in familiar areas as opposed to choosing a route that provides them with maximum utility. Other researchers such as Doherty and Miller (Doherty & Miller, 2000) investigating route choice found that apart from travel time, factors such as residential location, familiarity with the route, and employment locations are significant in the route choice process. Deep learning techniques (Basu et al., 2018, 2015; LeCun, Bengio, & Hinton, 2015; Liu et al., 2019; Lv, Duan, Kang, Li, & Wang, 2015; Song, Kanasugi, & Shibasaki, 2016) can be used to predict traffic congestion and route choice. However, deep learning models, being opaque, cannot be used to causally explain drivers' route choice.

In reviewing the existing research it can be gleaned that transportation researchers have employed two different types of empirical data collection in studying route choice behavior. First, collecting route choice data using observed actual choices and second, collecting route choice data in hypothetical experiments. Researchers have for the majority of cases used utility maximizing theory to explain route choice behavior that is rooted in econometrics (Ben-Akiva, Lerman, & Lerman, 1985).

Constructing Graphical Causal Models

In causal inference, we need a way of formally representing our assumptions about causal relationship within data. Graphical models comes in handy for this purpose. There are a variety of ways to depict causal relationship using graphical causal models (Spiegelhalter, Dawid, Lauritzen, & Cowell, 1993; Glymour & Cooper, 1999; Neapolitan et al., 2004; Pearl, 2009). A graphical model provides a clear way to represent and better understand the causal relationships within a data set (Pearl, 2014; Pearl & Mackenzie, 2018). A Causal graph is useful in determining the cause-effects from data by identifying confounding and endogenous selection bias. We also can derive a testable implications from the graph to test our assumptions (Elwert, 2013). To construct a graphical model requires subject-matter understanding (Hernan & Robins, 2018).

In our study, to model causal assumptions we carried out an iterative procedure following three steps. We identified the related variables and constructed our pilot causal graph via one-to-one discussions with experts in the field of transportation. In the second step, since a casual graph reveals testable implications, we tested our assumption to some extent using graphical criteria. In the final step, we evaluated the pilot model discussing with experts again. We modified our graph according to the discussion with experts and results obtained by testing the model against data. After proper adjustments, we finalized the causal model for further causal inference procedure.

Data Collection

Route choice can be influenced by factors, such as, road condition and human-related factors (i.e., driving experience, driver's socio-economic characteristics, and driving behavior and attitudes (de Oa, de Oa, Eboli, Forciniti, & Mazzulla, 2014)). The current route choice models are calibrated using static contextual conditions and are not generally able to account for accessibility to the nearest freeway, traffic incidents, and road closures due to emergency. Collecting dataset including dynamic contextual factors are challenging and the existing approaches usually take into account the general factors motivating drivers route choice behavior. In causal inference of route choice, it is preferable to have as much as data related to contextual factors which have potential influence on drivers' route choice decision. This study conducts experimental scenarios in which specific contextual factors are added in the testing design, using Virtual Reality (VR) platform and a driving simulator.The study, in particular, examines individuals diversion tendency onto alternate routes that are induced by traffic condition, journey type, and the impact of social influence while driving in the Interstate 10 (I-10) freeway in Baton Rouge, between the Mississippi River Bridge and College Drive exit. Collecting route choice data in hypothetical experiments facilitated our study by providing various factors information for causal analysis.

IVE Experimental Setting In this study, we used a driving environment that is designed based on the I-10, starting off the Mississippi River bridge all the way to the College Dr. Along the way, five alternate routes were introduced to the participants Exits A, B, C, D, and E, the latter of which would be College Drive. Ten experimental scenarios were conducted to produce initial data about drivers dynamic route choice behavior, given emerging contextual factors. See Table 1.

Forty-one individuals (20 male and 21 females; age: 31.44 ± 7.97) volunteered to participate in the experiment. Prior to the experiment, participants were presented with a questionnaire asking the following items: 1) demographic characteristics (age, gender, race, education, employment status); 2) top concerns while they stuck in the traffic conges-

Table 1: Contextual Factors Description

Contextual Factors	Scale
Traffic Condition	Normal
	Medium
	Heavy
Journey Type	Urgent
	Non-Urgent
Social Impact	Yes
	No

tion. Their choices included hours of extra travel time, speed reduction, monetized value of delay; 3) familiarity with the area; 4) socio-economic status (having concerns about spending less money on your gas). After answering the questionnaires, participants were asked to sit on a stationary chair at a desk with a driving wheel which was placed in front of a flat screen monitor where the driving simulation would run. Next, they were invited to practice for a few minutes to get acquainted with driving the simulator. After enough practicing with the driving simulator and becoming comfortable with its environment the research team would assign the participant to the scenarios. See Table 2.

Table 2: Experimental Scenarios of the Study

Traffic Condition	Journey Type	Social Impact
Normal	Urgent	No
Medium	Urgent	No
Heavy	Urgent	No
Medium	Urgent	Yes
Heavy	Urgent	Yes
Normal	Non-Urgent	No
Medium	Non-Urgent	No
Heavy	Non-Urgent	No
Medium	Non-Urgent	Yes
Heavy	Non-Urgent	Yes

The origin and the destination in all the scenarios were same and each scenario took about two minutes to finish. In each scenario different contextual factor(s) were presented and participants were required to choose their preferred route.

Each participant was exposed to all the driving scenarios including a baseline scenario. The baseline scenario would collect information about participants route choice pattern in a normal traffic and non-urgent bound condition. Each scenario contained 1, 2, or 3 contextual factors. The first contextual factor was the traffic flow which was varied over three levels, i.e., normal, medium, and heavy density. The next factor was the purpose of the trip (journey type) which consisted of a work-bound and home-bound trips; on the work-bound trip, participants were told to consider how important was it to meet the time of arrival commitment, while the home-bound posed no rush to reach the destination. The third factor considered is the impact of other drivers route choice, exploring the idea of social influence, that is whether the driver would be influenced by watching other drivers take an exit.

Dynamic route guidance was presented to the participants where a driver is guided on to routes that will minimize travel time for the overall road network. The scenarios were counterbalanced and played out in a random fashion to avoid behavioral biases due to order effect.

DAGs, D-Separation, Testable Implication

DAGs Directed Acyclic Graphs (DAGs) can represent probability distributions of the data and can be considered as causal graphical models under three important assumptions (Hernan & Robins, 2018). First, we assume that direct causal effect exists between pairs of variables connected by directed edges. Second, we assume that DAGs satisfy the Causal Markov condition (Hernan & Robins, 2018). The Causal Markov condition states that a variable is independent of every other variable except its effects conditioned on all of its direct causes (Hernan & Robins, 2018; Anderson & Lenz, 2001). Mathematically, this is expressed as:

$$f(V) = \prod_{i=1}^{n} f(x_i|pa_i)) \tag{1}$$

where $f(V)$ denotes joint probability mass function over the set of nodes V. The variables pa_i denote the values of the direct causes of variables x_i, and i takes values from 1 to n.

The third condition is Faithfulness. By assuming that a causal graph satisfies Causal Markov condition, we assume that any population produced by this causal graph has the conditional independence relations obtained by applying d-separation. (Scheines, 1997).

We can test the assumption (the pilot causal model) by applying d-separation (Pearl, 2014). This allows us to verify if the model fits the data. If the conditional independence test based on data violates the d-separation rule, we can modify original model. Fortunately, d-separation rules spot the flaws locally so we can fix the problems without much effort. We don't need to throw away the model and start the whole process from scratch.

D-SEPARATION: It is a criterion for identifying, from a given causal graph, which variables in the graph must be independent conditional on which other variables. D-separation rule needs to consider three basic causal structures in a DAG (Pearl, Glymour, & Jewell, 2016; Pearl, 2014). These structures correspond to causation, endogenous selection (Elwert & Winship, 2014; Pearl & Mackenzie, 2018), and confounding. We shall use a shorthand notation for conditional independence (Dawid, 1979). These structures are chains (i.e, $e \to d \to f$, the path is d-separated when $e \perp\!\!\!\perp f|d$), forks (i.e, $e \leftarrow d \to f$, the path is d-separated when $e \perp\!\!\!\perp f|d$), and inverted forks (i.e, $e \to d \leftarrow f$ the path is d-connected when $e \not\perp\!\!\!\perp f|d$, so to be d-separated we can not condition on d which is a collider).

We identified the D-separation conditions implied by the causal model and tested the implications to some extent using the dataset. The results are shown in Table 3.

In the conditional independence test, our null hypothesis states that two variables are independent conditional on the other variable. So, if the *p*-value is greater than the significance level $\alpha = 0.01$, we will accept our null hypothesis; otherwise we will reject it. After the test, it showed that some of the conditional independences implied by causal model were not consistent with the probability distribution underlying the dataset. For example, the result of conditional independence test against the dataset suggests that the variable of 1st concern while stuck in the traffic and the variable of route choice are independent conditioned on the variable of traffic. However, in the DAG, there is a direct edge between the variable of 1st concern while stuck in the traffic and the variable of route choice which is an indication of dependency. So we will eliminate this edge to make D-separation condition in the model match the conditional independence in the data.

Table 3: Some of the Conditional Independence Test

Conditional Independence	P-value (α=0.01)
1stConcernWhileStuckInTraffic and RouteChoice given Traffic	p=0.037
RouteChoice and 1stConcernWhileStuckInTraffic given Age	p=0.043
1stConcernWhileStuckInTraffic and Education given Race	p=3.57E-10
Education and 1stConcernWhileStuckInTraffic given Gender	p<2.2E-16
1stConcernWhileStuckInTraffic and RouteChoice given SocialImpact	p=0.504
RouteChoice and Traffic given SocialImpact	p<2.2E-16
1stConcernWhileStuckInTraffic and RouteChoice given Urgency	p=0.481
Traffic and 1stConcernWhileStuckInTraffic given Urgency	p=1
Education and EmploymentStatus given Gender	p<2.2E-16
Education and EmploymentStatus given Age	p<2.2E-16
Education and EmploymentStatus given Race	p<2.2E-16
FinancialConcern and EmploymentStatus given Age	p<2.2E-16

We concluded that the pilot model is not a good fit for the data set. To modify the pilot model, we could introduce new variables, remove redundant variables, or modify the relationship between variables by adding or eliminating nodes and edges. Based on the test results, we modified the pilot model by merely eliminating edges from the node of Traffic to the node of 1st concern while stuck in the traffic, and from the node of 1st concern while stuck in the traffic to the node of Route Choice. In the pilot DAG, there were 12 nodes (variables) and 26 directed edges. In the final causal DAG, the number of nodes remain the same as in the pilot model, but 24 directed edges remain. The pilot and final casual models are shown in Figure 1.

Causal effect estimation

The causal graph shows that between treatment-outcome pairs there is a direct path and an indirect (back-door) path (i.e, traffic $\rightarrow$ route choice is direct path; traffic $\leftarrow$ social impact $\rightarrow$ route choice is an indirect path). The back-door path is confounding. When trying to estimate causal effect, we want to block any back-door paths by conditioning on some variables, because such paths are not transmitting causal influences, and if we don't block the back-door path, it confounds the effect that a node has on another node. For instance, as shown in Figure 2 (we boxed the collider with the dashed line and similarly presented the confounder's arrows with dashed line), when trying to calculate the causal effect of employment status on route choice, there exists back-door paths: 1) employment status $\leftarrow$ education $\leftarrow$ gender $\rightarrow$ route choice; the blockage of this path can be ensured by conditioning on gender which is a confounder; 2) employment status $\leftarrow$ age $\rightarrow$ route choice; the blockage of this path can be ensured by conditioning on the confounder age; 3) employment status $\leftarrow$ race $\rightarrow$ 1st concern while stuck in the traffic $\leftarrow$ social impact $\rightarrow$ route choice; within the path there is a collider which is the variable of 1st concern while stuck in the traffic. So the back-door path is already blocked without conditioning on any variables. However, if we try to condition on the collider we make the path open instead. There are other back-door paths: we haven't listed all of them. After identifying every back-door path between these two variables, we selected age and gender thereby blocking the back-door paths between employment status and route choice.

To select the variables that entail blockage of the the back-door paths, we carried out graph-surgery as described above. Then, we adjusted these variables to calculate the pure causal effect. We paired every treatment with the outcome variable, identified back-door paths between them, and selected the confounding variables using the Back-door criterion. Between the variables of Urgency, Gender, Race, Age, SocialImpact, FamiliarityWiththeEnvironment, and the RouteChoice there doesn't exist any back-door path. In addition to that, from the variables of 1stConcernWhileStuckInTheTraffic, and FinancialConcern, there doesn't exist any direct causal path to the variable of RouteChoice. So, there is no casual effect of these two variables on the variable of RouteChoice. The relationship between them can be interpreted as association instead of causation. Hence, there is no need to estimate the causal effect of these two variables. We listed the confounders in the the back-door paths between the variables of Traffic, Urgency, Education, EmploymentStatus

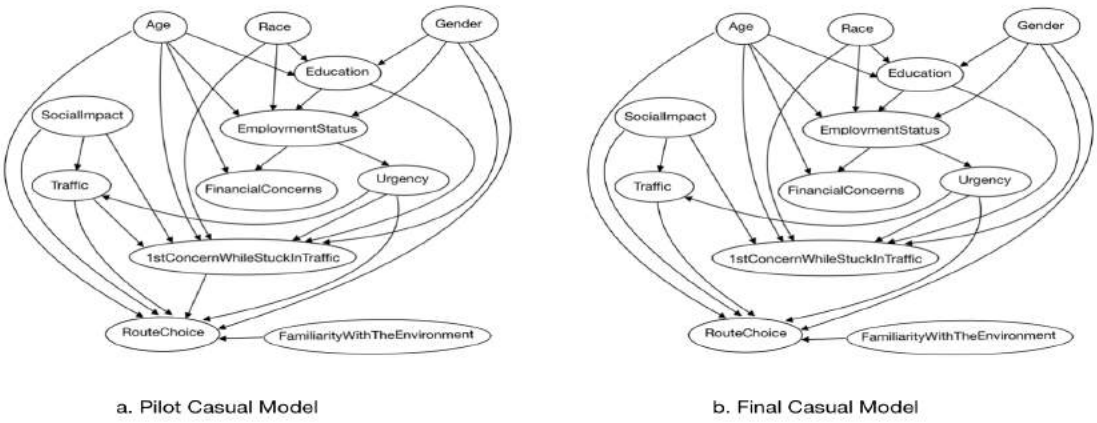


Figure 1: Casual Models

and RouteChoice (Shown in Table 4). Normally, the strategy of putting in all possible confounders is usually used. However, this strategy may end up adjusting for colliders and mediators that can introduce bias. For example, employment status → urgency → route choice. In this direct path, urgency is mediator. if we condition on the mediator, we will bias our estimate.

Table 4: Confounder in Backdoor Paths

Variables	Confounder
Traffic	Urgency, SocialImpact
Urgency	Gender, Age
Education	Race, Gender, Age
EmploymentStatus	Gender, Age

We applied Inverse Probability(IP) weighting method to adjust the variables Z, which are the confounders. The purpose of using IP weighting is to break the association between the covariates Z and treatment X to estimate true causal effect on outcome variable y (Hernán & Robins, 2006; Robins, Rotnitzky, & Zhao, 1994). By predicting Z based on X, we can estimate the propensity score $\Pr(X = x|Z)$. We can get a propensity score using non-parametric (i.e, probability) or parametric methods (i.e, regression model). If the data is high-dimensional with many covariates and some of them with multiple levels, it is desirable to use a parametric method. In our study, we have 12 variables and some variables have more than two levels. To find propensity score, we applied logistic regression model. The equation is given below:

$$\Pr(X_i = x|Z_i) = \frac{1}{1 + e^{-(\alpha+\beta Z_i)}} \tag{2}$$

After getting propensity scores, we used them to obtain the weights W to create a pseudo-sample in which there is no association between the covariates and treatment. The IP weighting formula is given below:

$$W_i = \frac{X_i}{\Pr(X|Z_i)} + \frac{1 - X_i}{1 - \Pr(X|Z_i)} \tag{3}$$

where X_i indicates if the ith subject was treated.

We started our approach of calculating the causal effect by training a model with covariates Z to predict X. Our treatments X are categorical variables, so we calculated the propensity scores $\Pr(X|Z)$ by applying equation 2.

After estimating the propensity scores, we applied equation 3 to obtain the IP weight. We used stabilizing factor $\Pr(X)$ in the numerator to narrow the range of the $\Pr(X)/W$. After we obtained stabilized IP weights as $SW = \Pr(x)/W$, we trained new model with treatment variables X as features and outcome Y by using SW_i as sample weight for the ith observation. Then, we used this model to predict the causal effect. In this study, outcome variable is categorical data, so we used logistic regression again to obtain the casual odds ratio as a casual effect measure.

Based on our determination of average causal effect (shown in Table 5), the result suggests that when heavy and medium traffic conditions are compared with the normal condition separately, their effects have significant magnitude. It implies that changes in traffic conditions impact drivers route choice. More specifically, when the traffic is heavy, proportion of drivers who choose the nearest exit is about 6 times greater than that when traffic is normal. However, when the traffic is medium, proportion of drivers who choose the nearest exit is about 3 times greater than that when traffic is normal. So, we can conclude that when the traffic is normal, the drivers are more likely to stay on the high way. Considering social impact, a driver would be influenced by watching other drivers take the exit. The proportion of drivers who choose the nearest exit is about 5 times greater when they get influenced than they don't. Considering familiarity with environment, proportion of drivers, who are not familiar with the road, choosing the nearest exit is about 3 times greater than those who are familiar with the road. Considering race, white

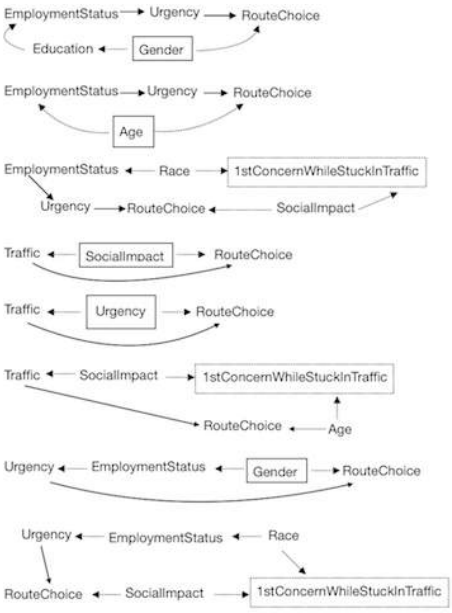


Figure 2: Blocking Backdoor

Table 5: Average Casual Estimation Result

Variables	ATE	95%Conf.Interval
SocialImpact		
yes vs no	4.945	3.966 – 5.919
Urgency		
Urgent vs Unurgent	1.193	0.773 – 1.612
Age		
Middle vs Young	1.081	0.179 – 1.982
Old vs Young	0.461	-3.274 – 4.196
Gender		
Male vs Female	0.976	0.135–1.816
Race		
Middle Eastern vs White	5.760	4.587 – 6.932
Other vs White	3.966	2.750 – 5.181
EmploymentStatus		
PartTime vs Unemployed	0.007	-6.049 – 7.651
FullTime vs Unemployed	0.015	-4.140 – 4.213
Student vs Unemployed	0.010	-4.471 – 4.582
Education		
HighSchool vs PostGraduate	0.054	-2.431 – 2.539
College vs PostGraduate	1.231	0.825 – 1.636
Traffic		
Medium vs Normal	3.663	1.953 – 5.372
Heavy vs Normal	6.562	4.817 – 8.306
FamilarityWithEnvironment		
OnceAMonth vs OnceAWeek	2.795	0.938 – 4.651
OnceAYear vs OnceAWeek	2.778	1.374 – 4.181

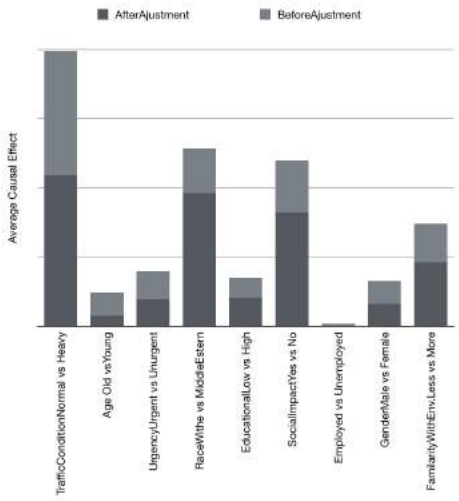


Figure 3: Average Causal Effects Of Variables Computed by Adjusted and Un-Adjusted Regression Model

people are more likely to stay on the highway than middle eastern people or others. Age and Urgency also have significant effect on drivers route choice. We also conducted another experiment in which we built an estimator for route choice without adjusting for confounding factors, and compared the results with the one of causal inference (shown in Figure 3). The results suggest that on the un-adjusted estimator, the effect of age and employment status are overestimated, race, social impact, and familiarity with the environment are under estimated. This is because there are confounding and collider sources between the path of these variables and the outcome. Based on causal inference, the effect of traffic, race, social impact, and familiarity with the environment are more significant than others.

Conclusions

This paper described a causal analysis of route choice behavior of drivers using data collected from a Stated Choice Experiment in an Immersive Virtual Environment (IVE). This work will not only fill in the lack of causality based approaches in the transportation field, but it also showed that without adjustment on treatment, causal effect results will be affected by spurious correlation as well.

Acknowledgment

This research was supported by Transportation Consortium of South-Central States (Tran-SET) Award No 18ITSLSU09/69A3551747016. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the sponsor.

References

Anderson, R. D., & Lenz, R. T. (2001). Modeling the impact of organizational change: A bayesian network approach. *Organizational Research Methods*, *4*(2), 112-130.

Basu, S., Ganguly, S., Mukhopadhyay, S., DiBiano, R., Karki, M., & Nemani, R. (2015). Deepsat: a learning framework for satellite imagery. In *Proceedings of the 23rd sigspatial international conference on advances in geographic information systems* (p. 37).

Basu, S., Mukhopadhyay, S., Karki, M., DiBiano, R., Ganguly, S., Nemani, R., & Gayaka, S. (2018). Deep neural networks for texture classificationa theoretical analysis. *Neural Networks*, *97*, 173–182.

Ben-Akiva, M. E., Lerman, S. R., & Lerman, S. R. (1985). *Discrete choice analysis: theory and application to travel demand*. MIT press.

Ben-Akiva, M. E., Ramming, M. S., & Bekhor, S. (2004). Route choice models. In *Human behaviour and traffic networks.* Springer.

Ben-Elia, E., & Shiftan, Y. (2010). Which road do i take? a learning-based model of route-choice behavior with real-time information. *Transportation Research Part A: Policy and Practice*.

Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society: Series B (Methodological)*.

de Oa, J., de Oa, R., Eboli, L., Forciniti, C., & Mazzulla, G. (2014). How to identify the key factors that affect driver perception of accident risk. a comparison between italian and spanish driver behavior. *Accident Analysis Prevention*, *73*, 225 - 235.

Doherty, S. T., & Miller, E. J. (2000). A computerized household activity scheduling survey. *Transportation*.

Elwert, F. (2013). Graphical causal models. In *Handbook of causal analysis for social research.* Springer.

Elwert, F., & Winship, C. (2014). Endogenous selection bias: The problem of conditioning on a collider variable. *Annual review of sociology*.

Glymour, C. N., & Cooper, G. F. (1999). *Computation, causation, and discovery*. Aaai Press.

Hernán, M. A., & Robins, J. M. (2006). Estimating causal effects from epidemiological data. *Journal of Epidemiology & Community Health*.

Hernan, M. A., & Robins, J. M. (2018). *Causal inference*. Boca Raton : Chapman Hall/CRC, forthcoming.

Khattak, A. J., Schofer, J. L., & Koppelman, F. S. (1993). Commuters' enroute diversion and return decisions: analysis and implications for advanced traveler information systems. *Transportation Research Part A: Policy and Practice*.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436.

Liu, Q., Mukhopadhyay, S., Zhu, Y., Gudishala, R., Saeidi, S., & Nabijiang, A. (2019). Improving route choice models by incorporating contextual factors via knowledge distillation. In *In proceedings of ieee international joint conference on neural networks (ijcnn).*

Lv, Y., Duan, Y., Kang, W., Li, Z., & Wang, F.-Y. (2015). Traffic flow prediction with big data: a deep learning approach. *IEEE Transactions on Intelligent Transportation Systems*, *16*(2), 865–873.

Neapolitan, R. E., et al. (2004). *Learning bayesian networks*. Pearson Prentice Hall Upper Saddle River, NJ.

Pearl, J. (2009). *Causality*. Cambridge university press.

Pearl, J. (2014). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Elsevier.

Pearl, J., Glymour, M., & Jewell, N. P. (2016). *Causal inference in statistics: a primer*. John Wiley & Sons.

Pearl, J., & Mackenzie, D. (2018). *The book of why : the new science of cause and effect*. Basic Books.

Prato, C. G. (2009). Route choice modeling: past, present and future research directions. *Journal of choice modelling*.

Pursula, M., & Talvitie, A. (1993). Urban route choice modelling with multinomial logit models. *LIIKENNETEKNIIKKA, TIEDOTE*.

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*.

Scheines, R. (1997). An introduction to causal inference.

Song, X., Kanasugi, H., & Shibasaki, R. (2016). Deeptransport: Prediction and simulation of human mobility and transportation mode at a citywide level. In *Ijcai* (Vol. 16, pp. 2618–2624).

Spiegelhalter, D. J., Dawid, A. P., Lauritzen, S. L., & Cowell, R. G. (1993). Bayesian analysis in expert systems. *Statistical science*.

Investigating the Intrinsic Integration Hypothesis for the Design of Game-Based Learning Activities

Graeme Nidd (graeme.nidd@carleton.ca)
Institute of Cognitive Science, Carleton University
Ottawa, Ontario, Canada

Kasia Muldner (kasia.muldner@carleton.ca)
Institute of Cognitive Science, Carleton University
Ottawa, Ontario, Canada

Abstract

The intrinsic integration hypothesis proposes that using core game mechanisms to teach learning material makes educational games more fun to play and better for learning. Our study tests the intrinsic integration hypothesis with two educational versions of Battleship that were designed for this experiment, in the domain of complex numbers. We examine the learning gains and motivation of 58 participants who interacted with either the intrinsically-integrated or extrinsically-integrated version of the game. Our results contradict previous findings supporting the intrinsic integration hypothesis: participants reported similar levels of motivation from both versions of the game and participants who interacted with the extrinsically-integrated version learned significantly more as measured by pretest to posttest gains. This work contributes empirical data to the debate concerning intrinsic integration, and it highlights the need for additional studies exploring the integration of learning material into educational games.

Keywords: Intrinsic integration; games; student learning

Games and Student Learning

Educational games aim to make learning fun by incorporating game elements, such as fantasy, challenge, and competition, into instructional activities (Malone & Lepper, 1987). Three recent meta-analyses found that overall, students learn more from educational games than from traditional activities like standard classroom instruction (Sitzmann, 2011; Wouters, van Nimwegen, van Oostendorp, & van der Spek, 2013; Clark, Tanner-Smith, & Killingsworth, 2016). While this result is encouraging, there are two caveats: (1) not all studies found a positive effect of games, and (2) comparing games to other activities does not inform on how to best design games to maximize learning and engagement from them. Thus, there have been calls to test the effect of various design factors on student outcomes, referred to as the *value-added* approach to educational game research (Mayer, 2011). This approach involves comparing student learning and/or motivation with a basic version of a game to the outcome of one that includes an additional design feature.

As an example of the value-added approach, studies have examined the effects of cooperation and competition in educational games (Ke & Grabowski, 2007; Plass et al., 2013). Ke and Grabowski (2007) compared the impact of games involving cooperative competition, individual competition, and a non-game control condition on math learning and attitudes among fifth-grade students. Participants in both game conditions learned more than those in the non-game control condition. Moreover, attitudes regarding math were significantly better in the cooperative game condition than in the other two conditions. Plass et al. (2013) also examined the effects of cooperation and competition on learning outcomes. Learning, assessed by pretest and posttest scores, was only significantly higher in the competitive condition compared to the control condition that did not include competition or cooperation. While the collaborative condition had the lowest in-game performance of all three conditions, it produced the most positive affect as measured by intention to play the game again and to recommend it to others.

As another example of the value-added approach, Conati and Manske (2009) assessed the value of adding an agent delivering adaptive hints in an educational game. The hints were generated based on a user model of student knowledge. No difference was found between the agent version of the game and a control version without the agent. Conati and Manske (2009) speculated that the reason for the lack of an effect may have been due to an inaccurate user model, the challenge of fostering learning in the target domain, and/or the hints interrupting the flow of the game.

An area within the value-added approach that has not received much attention is the integration of a game's motivating elements with the learning material. A recent meta-analysis by Clark et al. (2016) found only one experiment that investigated this factor, involving a game that completely separated the learning mechanisms from those designed for engagement with the game – this experiment will be described in the next section.

Extrinsic and Intrinsic Integration

How should game elements be integrated with the learning elements? Kafai (1996) anecdotally observed that students tasked with designing educational games took one of two distinct approaches. He called this dichotomy *extrinsic* vs. *intrinsic* integration. The extrinsic approach used the game as a form of 'sugar-coating:' players in the game were rewarded for answering questions on the learning material with the opportunity to continue playing the game. Thus, the game play was clearly separated from the instructional activities. The alternative to this approach is the intrinsic approach, which involves using the game's core mechanisms to present the learning material, thereby integrating the learning activities with game play. Thus, in contrast to the extrinsic approach, with the intrinsic approach there is no distinct separation between game activities and learning activities.

Habgood and Ainsworth (2011) proposed that students would learn more from intrinsically-integrated games than extrinsically-integrated games (referred to as the *intrinsic integration hypothesis*). Their experiment compared two versions of an educational game called Zombie Division, designed for middle-school children. In the intrinsically-integrated version, players navigated their character around a dungeon and used division to defeat computer-controlled opponents represented by skeletons. Importantly, while this version required students to practice division, doing so was the primary way to progress through the game. In contrast, in the extrinsic version the learning material was removed from the game portion and isolated to quizzes presented between game sessions. Results indicated that students who played the intrinsically-integrated version improved significantly more from pretest to posttest and reported higher engagement.

While the Habgood and Ainsworth (2011) results are encouraging, they warrant replication. Because the math activities were moved to quizzes in the extrinsic version of the game, the gameplay in that version became less challenging, as acknowledged by the authors and reported by the students who played the game. Lack of challenge may have diminished learning outcomes from this version. Additionally, interleaving the questions with gameplay sessions changed the instructional sequence of the extrinsic condition. Thus, the decreased challenge and different instructional sequence could have biased the results.

While intrinsic integration does have the benefit of not interrupting players during game play to have them complete educational tasks, it also has potential downsides. One is related to transfer. The learning material in an intrinsically-integrated game is often presented in a context different from the one in which it will later be applied and tested. Students find it difficult to transfer knowledge learned in one context to a different one even when the fundamental concepts are the same (Kaminski, Sloutsky, & Heckler 2009). Intrinsic integration could also be disadvantageous because it requires the player to simultaneously cope with two competing sets of demands, stemming from the educational and game elements, which could increase extraneous cognitive load.

Given the above considerations, the goal of the present work was to test the intrinsic integration hypothesis through an empirical study.

The Present Study

To test the intrinsic integration hypothesis, we created a paper-and-pencil educational game designed to help students practice concepts in our target domain of complex numbers. The game was based on *Battleship*. To play Battleship, each player secretly plots their ships onto a two-dimensional plane and then fires upon their opponent's ships. The first player to correctly guess every coordinate containing a ship wins the game. While the original Battleship was not explicitly educational, the two-dimensional nature of complex numbers makes them particularly suited for intrinsic integration into Battleship, as the coordinates on the two-dimensional board can be substituted with complex numbers.

Participants

The participants (N = 66, 35 females) were undergraduate students at a Canadian University recruited via Sona and posters displayed around campus. As the game in our study was played in pairs, participants were asked to come to the study with a friend or classmate, instead of being paired with a stranger. This was done to facilitate interaction during gameplay, as both participants would already know each other. Each participant was compensated with their choice of either course credit or $20.

Materials

Intrinsic and Extrinsic Versions of Battleship We created two versions of Battleship; both were played with pencil and paper materials. In each version, participants had two game boards, printed on paper, also referred to here as "planes". One game plane was private as it was positioned behind a screen and players were instructed to keep it hidden from their opponent. They were asked to draw their ships on this private plane at the start of the game. The second plane was public and was used during the game to indicate players' shots on their opponents' ships (done by drawing the shot on the public game plane). Because our goal was to only vary the intrinsic/extrinsic dimension while keeping other aspects of the two game versions as similar as possible, the game play was almost identical in both versions.

In the intrinsic version, the game board corresponded to a complex plane (see Figure 1, left). A turn began with each of the two players selecting where they would place their next shot on the public plane. To do so, they indicated the chosen location by writing down the rectangular form of a complex number corresponding to that location on the '*shot list*,' which was a second piece of paper labeled with turns (see Figure 1, right). For example, if a participant thought their opponent's ship was in the top-left of the plane, then

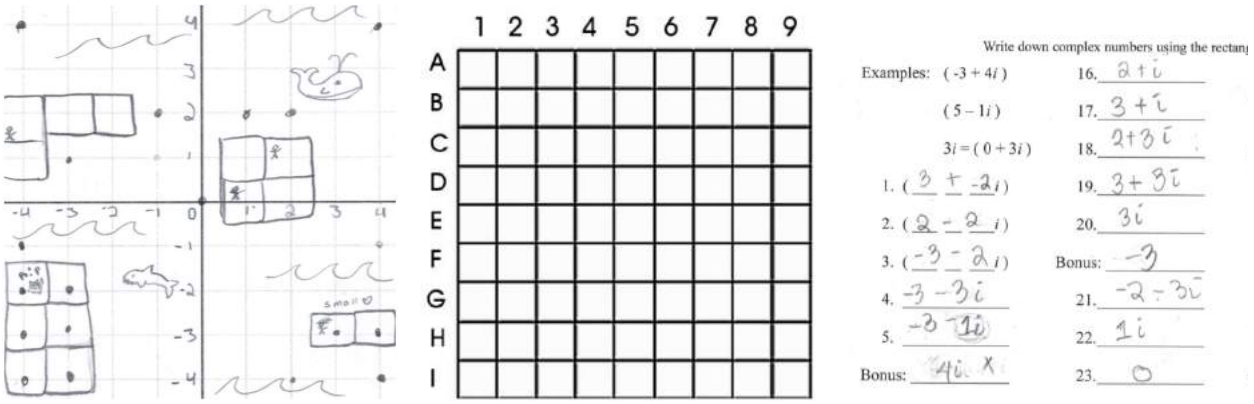


Figure 1: The complex plane used in the intrinsic condition (left), marked with a participant's game moves, the plane used in the extrinsic condition (centre), shown without any entries, and the response sheet used in both conditions (right), shown with a player's entries.

they would write (-4 + 4*i*). When both shots were written down, the players checked each other's entries for the correct format. The participants could not continue playing until the correctness of their entry was confirmed by their opponent. However, the experimenter did not verify the answers and, if asked, referred participants to the examples provided by the game material (the screen hiding the private game planes had a recap of the instructional material and another sheet provided examples of complex number problems). Thus, the responsibility was on the participants to verify their entries, and they had resources to help them do so. This did not take away from the competitive aspect of the game as the competition came from locating the opponent's ships, like in the original Battleship. Once both players were satisfied their opponents had written a complex number in the correct format, they checked if their opponent's shot struck one of their ships, and they indicated the result on their public game board. Note that "a shot" corresponded to the complex number that they had written down. Thus, the learning material was intrinsically integrated with the game mechanisms: to play the game, participants had to apply complex number knowledge.

The extrinsic version was identical except for two key differences. First, the game board was based on the standard *Battleship* game and so corresponded to a coordinate plane where the axes were labeled with letters in the left margin and numbers on the top margin (see Figure 1, center). Second, at the start of a turn, participants first randomly chose a coordinate on the complex plane from a deck of cards. Thus, in this version, the coordinate did not represent a shot on the opponent. Like in the intrinsic version, the players translated that coordinate to a complex number and had their opponent check it. In contrast to the intrinsic version, however, they then specified the shot on their opponent using a letter-number pair corresponding to the axes' labels on their planes (e.g., A-2). This was done to create a divide between the learning material and the game material, thereby making the game extrinsically integrated.

After every five shots participants in both game versions were asked to multiply the previous complex number by the imaginary unit, writing their answer on the shot list. This was considered a bonus question and a correct answer was rewarded with an extra shot.

The game and study materials were refined via pilots.

Complex numbers lesson To provide the domain background needed to play the educational game, participants were given a paper-based lesson we developed on the complex number system. The lesson consisted of a two-page description with accompanying illustrations.

Test Materials A pretest and posttest were used to measure participants' complex numbers knowledge before and after they played the game. Each test consisted of twenty questions.

Instruments An online survey was used to collect motivational and affective data, in addition to basic demographics. The motivational and affective survey used a Likert scale and included: (1) the Intrinsic Motivation Inventory (Deci & Ryan, 2003) based on four sub-constructs, including interest, competency, choice, and pressure; (2) some custom questions measuring participants' willingness to re-engage with the instructional material in the future (e.g., "I would use the game to teach complex numbers"). Several other instruments were used to measure mindset and math attitudes but results from their analysis are not included here, so they are not described.

Design

We used a two-factor (2 x 2) mixed design. The first factor, *condition*, was a between-groups variable with two levels (intrinsic and extrinsic, corresponding to intrinsically-integrated and extrinsically-integrated game versions, respectively). The second factor, *time*, was a within-groups variable with two levels (pre and post, referring to pre-game

play and post-game play, respectively). Participants were assigned to a given condition in a round-robin fashion.

Procedure

Each session was conducted individually and included a pair of participants. Each dyad spent approximately 90 minutes in the study, with the exact duration varying based upon the amount of time participants spent on the instructional material as well as the pretest and posttest. The procedure for the two conditions was the same. After providing consent, participants were seated back to back and (1) read the complex numbers lesson, and (2) filled in the complex numbers pretest. Once both participants had finished the pretest, they were asked to move to the game table positioned in the centre of the room where they sat across from each other, and the gameplay phase began.

Participants were provided with all the game materials and instructions on how to how to play the game (for details, see Nidd, 2018). After both participants had plotted their ships according to the game rules, they were given 35 minutes to play the game. Any questions relating to complex numbers were answered by referring the participants to the examples in the instructional materials that were provided as well as the recap of the lesson on each of their game screens. When the time was up, participants were given the choice to play for another five minutes if they wanted. This was done as an additional measure of motivation.

Directly after the game phase, participants were moved back to their initial seats where they were seated back-to-back and completed the (1) posttest and (2) the study questionnaires.

Results

The analysis is based on 58 participants (eight participants were not included either because they were at ceiling on pretest, i.e., 90% or higher or because their performance decreased from pretest to posttest). The analyses, which were conducted with the statistical software *R*, used inferential statistics that assume independence between participants. Since participants worked together during the game, there was a potential concern that their learning-related data might be dependent. To check for this, a correlation between pretest to posttest difference scores of both individuals in a pair was conducted. The correlation between the learning outcomes of paired participants was not significant and corresponded to a very small effect, $r(31) = .05$, $p = 0.78$, suggesting that the independence assumption was not violated. Thus, we continued with our analysis testing the conditional effect on (1) learning outcomes and (2) motivation.

Are Intrinsically-integrated Games Better for Learning?

To check for equivalence between the two conditions on *a priori* knowledge, participants' pretest scores were compared. The scores were distributed fairly evenly between the two conditions and while they were slightly positively skewed (skewness of 0.90 and 0.60 respectively), this was within the bounds of normality. As shown in Table 1, the mean pretest scores were similar between the two conditions, with no significant difference between them as indicated by an independent samples t-test, $t(54.05) = 0.11$, $p = .91$.

Table 1: Descriptive statistics for the test scores.

	Extrinsic		Intrinsic	
	M	*SD*	*M*	*SD*
Pretest (/20)	4.93	2.85	4.83	3.71
Posttest (/20)	10.39	3.79	8.73	4.38
Difference (post - pre)	5.46	2.43	3.90	2.64

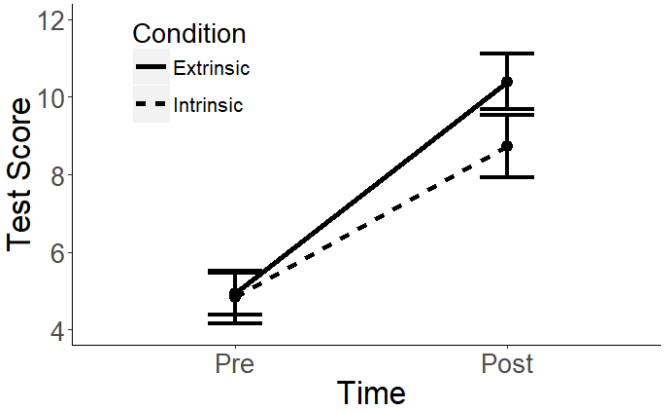


Figure 2: Interaction between time and test score, indicating higher pre to posttest learning in the extrinsic condition.

As is standard, learning was measured by the difference between a participant's performance on the pretest, completed after they read the instructional material but before they played the educational game, and their performance on the posttest. The descriptive statistics are shown in Table 1. The higher mean difference in the extrinsic condition suggests that participants who played the extrinsic version of the game learned more, because they improved more from pretest to posttest.

To analyze the impact of the extrinsically- and intrinsically-integrated versions of the game on learning, a two-way mixed ANOVA was conducted with test scores as the dependent variables, condition (extrinsic vs. intrinsic) as the between-subjects independent variable, and time (before and after the experimental intervention, i.e. game play) as the within-subjects independent variable.

In general, collapsed across conditions, participants improved from the pretest to posttest as indicated by the

significant main effect of time on participants' test scores, $F(1, 56) = 194.62$, $p < .001$, $\eta_p^2 = .78$. While this demonstrates that the instructional material improved learning overall (collapsed across the two conditions), of primary interest is the time by condition interaction, which examines the effect of condition on learning (i.e., pretest to posttest differences). This interaction was significant, $F(1, 56) = 5.49$, $p = .02$, $\eta_p^2 = .09$. As shown in Figure 2 this interaction indicates that participants who played the extrinsic version of the game learned significantly more than those who played the intrinsically-integrated game.

Are Intrinsically-integrated Games More Motivating?

The effect of game version on participants' motivation was measured by (1) Intrinsic Motivation Inventory (Deci & Ryan, 2003), (2) the custom questionnaire measuring self-reported re-engagement, and (3) the behavioral data on whether participants chose to continue playing the game for an additional five minutes after they were told they could stop. Like the Intrinsic Motivation Inventory, this additional measure was derived by averaging a participant's answers to the custom set of questions that asked them to report their willingness to re-engage with the instructional material using a 7-point Likert scale.

Descriptive statistics for this analysis are in Table 2. There was little difference between the two conditions in terms of the motivational variables. This was confirmed by a series of independent-samples t-tests comparing the five measures of participants' motivation in the two conditions. As shown in Table 3, none of the analyses were significant (while this analysis did not control for familywise error rate, doing so would not have changed the results, as none of the findings were significant). A chi-squared test of independence was performed to examine the relationship between the game version and participants' decision to continue playing for an additional five minutes. Like the other measures of motivation, the difference between the two conditions was not significant, $X^2(1, 29) = 0.016$, $p = .90$.

In summary, there was no evidence that the version of the game, intrinsic versus extrinsic, impacted participants' motivation. However, collapsed across condition, participants had fun playing the game. Participants reported that they were interested in the instructional material as indicated by high scores on the motivational questionnaire, and a third of them chose to stay longer than they needed to. Anecdotally, these measures are further supported by the verbal reactions of participants. One person remarked that the experiment was "really fun actually. If math was like this, I'd enjoy it a lot more." Another exclaimed upon receiving the post-test, "Battleship actually helped with this!" When the same participant – who was vocally anxious about math – forgot to take their shot upon the opponent's ships and immediately drew another complex number question, they joked: "Sorry, I just love math." Additionally, some participants asked if they could keep their game sheets to finish the game at home, and one participant even asked if they could buy the extrinsic version as they thought it was an improvement on the original Battleship. These measures and anecdotal reactions suggest that the educational game was motivating for participants.

Table 2: Descriptives for the five motivation subscales in each condition.

	Extrinsic		Intrinsic	
Subscale	*M*	*SD*	*M*	*SD*
Interest	5.00	1.64	5.01	1.48
Competency	4.27	1.41	4.62	1.44
Choice	5.03	1.42	5.13	1.40
Pressure	2.66	1.37	2.99	1.22
Re-engagement	4.17	1.27	3.95	1.44

Note. The maximum score for each subscale is 7

Table 3: Results for the conditional effect on each subscale of the motivational questionnaire.

Subscale	*df*	*t*	*p*	*d*
Interest	54.35	0.03	.98	.01
Competency	55.89	0.94	.35	.25
Choice	55.57	0.28	.78	.07
Pressure	54.20	0.96	.34	.25
Re-engagement	55.82	0.61	.54	.16

Discussion

Our results do not support the intrinsic integration hypothesis, as participants who played the intrinsically-integrated version of the game were not more motivated and did not learn more than those who played the extrinsic version. On the contrary, those who played the extrinsically-integrated version of the game learned significantly more.

Why did extrinsic integration result in more learning than intrinsic integration? As we already noted, one of the potential disadvantages of intrinsic integration is the need for transfer. In the intrinsic game version, the complex numbers corresponded to the coordinates of players' ships. Consequently, the numbers represented two constructs: they were concrete representations of a location on the game board, and they were the abstract representations that would later be tested. By having participants play and interact with these representations, intrinsic integration potentially made it more difficult for participants to see the complex numbers they were using as being important in themselves (Brown, McNeil, & Glenberg, 2009; Uttal, O'Doherty, Newland,

Hand, & DeLoache, 2009). Importantly, this potential disadvantage of intrinsic integration is not an artifact of our game design but rather a requirement of intrinsically integrated games. In contrast, the extrinsic version may have made it easier for participants to focus on and learn the mathematical principles by separating the abstract target knowledge from the more concrete interactions between the player and the game state (Uttal et al., 2009).

A second potential explanation for our findings pertains to cognitive load. The intrinsically-integrated game may have increased participants' extraneous cognitive load, as the tasks related to game play and complex numbers were integrated. In other words, the intrinsic version had players pick a shot, practice the learning material, and then resolve the shot. In contrast, the extrinsic version separated these tasks. These competing demands imposed by the intrinsic game and the domain questions may have diminished players' learning by increasing the load on their working memory (Clark, Nguyen, Sweller, & Baddeley, 2006). Similarly, the extrinsic version could have made working memory available for the mental processing that is required for learning. Since we did not measure cognitive load, this conjecture awaits future research.

Our results are not aligned with those from Habgood and Ainsworth's (2011) experiment. A potential explanation for these differences relates to control of the instructional sequence and challenge levels in the two versions of the game. Our experiment maintained similar instructional sequences between conditions by incorporating the extrinsic learning material throughout gameplay. In contrast, the prior study divided the learning material and game into lengthy blocks that may have disrupted user engagement more than is necessitated by extrinsic game design. This separation in the prior study also reduced challenge, a factor known to impact engagement with games (Garris, Ahlers, & Driskell, 2002). By removing the learning material from the game mechanism, players no longer had to solve a problem to progress through the game. This was reported by participants as they remarked, "it just tells you what to use" and "it's not a challenge" (Habgood & Ainsworth, 2011, p. 28). This difference was not present in the two game versions used in our experiment.

Another potential reason that our results do not support the intrinsic integration hypothesis relates to an interaction between the type of integration and cooperation/ competition. Specifically, adding a second player may have 'gamified' the non-game elements. For instance, participants answering the non-game domain questions in the extrinsic version of Battleship were still competing against their opponent to get the right answer. This aspect of the extrinsic game is comparable to a trivia game, as a correct answer was required to take a shot in the game of Battleship. Indeed, an educational game could consist of just this competitive quiz aspect (as in Ke & Grabowski). In Habgood and Ainsworth's (2011) game, completing the domain questions in the extrinsic version was likewise necessary to play the game, as participants needed to repeat the quiz if they did not get a passing score; however, this requirement could seem like a prerequisite in a single-player game, whereas it could seem like an element of the game when another player is involved.

There are also several methodological differences worth noting between our experiment and the previous work that did support of the intrinsic integration hypothesis (Habgood & Ainsworth, 2011). Our experiment used undergraduate students as opposed to primary school students between the ages of 7 and 9. Additionally, we recruited these participants in pairs instead of recruiting entire classes. Although similar domains were used, the target knowledge was more advanced in our experiment to match the participants' education level. The games in the two experiments differed in fundamental ways: our game was implemented as a board game rather than a video game, another human player was involved in our game, and the narrative elements were more pronounced in Habgood and Ainsworth's game. The measure of motivation also differed as Habgood and Ainsworth used qualitative interview data paired with a second experiment that measured the amount of time spent in the intrinsic and extrinsic versions when given a choice. In lieu of this, our experiment used the established Intrinsic Motivation Inventory to measure participants motivation to engage with the educational game.

In conclusion, our experiment contributes empirical data to the debate concerning intrinsic integration and educational game design. Our findings indicate that extrinsically-integrated games are better for learning and similarly motivating as intrinsically-integrated games. Ultimately, given the relatively few studies in this area and the lack of agreement between findings from the ones that do exist, our work highlights the need to further explore factors related to educational game design and their impact on student learning and motivation.

Acknowledgements

This work was supported with an NSERC Discovery Grant #1507 and a Masters SSHRC grant.

References

Brown, M. C., McNeil, N. M., & Glenberg, A. M. (2009). Using concreteness in education: Real problems, potential solutions. Child Development Perspectives, 3(3), 160-164.

Clark, D. B., Tanner-Smith, E. E., & Killingsworth, S. S. (2016). Digital games, design, and learning: A systematic review and meta-analysis. Review of Educational Research, 86(1), 79-122.

Clark, R. C., Nguyen, F., Sweller, J., & Baddeley, M. (2006). Efficiency in learning: Evidence□based guidelines to manage cognitive load. Performance Improvement, 45(9), 46-47.

Conati, C., & Manske, M. (2009). Evaluating adaptive feedback in an educational computer game. In *International workshop on intelligent virtual agents* (pp. 146-158). Heidelberg, Berlin: Springer.

Deci, E. L., & Ryan, R. M. (2003). Intrinsic motivation inventory. Self-Determination Theory, 267.

Garris, R., Ahlers, R., & Driskell, J. E. (2002). Games, motivation, and learning: A research and practice model. Simulation & Gaming, 33(4), 441-467.

Habgood, M. J., & Ainsworth, S. E. (2011). Motivating children to learn effectively: Exploring the value of intrinsic integration in educational games. The Journal of the Learning Sciences, 20(2), 169-206.

Kafai, Y. B. (1996). Learning design by making games: Children's development of strategies in the creation of a complex computational artifact. In Y. B. Kafai & M. Resnick (Eds.), Constructionism in Practice: Designing, Thinking and Learning in a Digital World (pp. 71-96). Mahwah, NJ: Lawrence Erlbaum Associates.

Kaminski, J. A., Sloutsky, V. M., & Heckler, A. F. (2008). The advantage of abstract examples in learning math. Science, 320(5875), 454-455.

Ke, F., & Grabowski, B. (2007). Gameplaying for maths learning: cooperative or not?. British Journal of Educational Technology, 38(2), 249-259.

Malone, T. W., & Lepper, M. R. (1987). Making learning fun: A taxonomy of intrinsic motivations for learning. Aptitude, Learning, and Instruction, 3(1987), 223-253.

Mayer, R. E. (2011). Multimedia learning and games. In S. Tobias & J. D. Fletcher (Eds.), Computer Games and Instruction (pp. 281-305). Charlotte, NC: IAP Information Age Publishing.

Nidd, G. (2018). *Revisiting the Intrinsic Integration Hypothesis* (Unpublished master's thesis). Carleton University, Ottawa, Canada.

Plass, J. L., O'keefe, P. A., Homer, B. D., Case, J., Hayward, E. O., Stein, M., & Perlin, K. (2013). The impact of individual, competitive, and collaborative mathematics game play on learning, performance, and motivation. Journal of Educational Psychology, 105(4), 1050-1066.

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Sitzmann, T. (2011). A meta-analytic examination of the instructional effectiveness of computer□based simulation games. Personnel Psychology, 64(2), 489-528.

Uttal, D. H., O'Doherty, K., Newland, R., Hand, L. L., & DeLoache, J. (2009). Dual representation and the linking of concrete and symbolic representations. Child Development Perspectives, 3(3), 156-159.

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.

Wouters, P., van Nimwegen, C., van Oostendorp, H., & van der Spek, E. D. (2013). A meta-analysis of the cognitive and motivational effects of serious games. Journal of Educational Psychology, 105(2), 249-265.

To be or not to be: Examining the role of language in a concept of negation

Ann E. Nordmeyer (a.nordmeyer@snhu.edu)
Department of Psychology, Southern New Hampshire University

Jill G. de Villiers (jdevilli@smith.edu)
Department of Psychology, Smith College

Abstract

Negation is a complex, abstract concept, despite the ubiquity of words like "no" and "not" in even young children's speech. One challenging aspect to words like "no" and "not" is that these words can serve many functions in speech, giving us tools to express an array of concepts such as denial, refusal, and nonexistence. Is there a single concept of "negation" that unites these separate negative functions – and if so, does understanding this concept require the structure of human language? In this paper we present a study demonstrating that adults spontaneously identify a concept of negation in the absence of explicit verbal instructions, even when the exemplars of negation are perceptually varied and represent many different functions of negation. Furthermore, tying up participants' language ability using verbal shadowing impairs participants' ability to identify a concept of negation, but does not impair participants' ability to identify an equally complex control concept (natural kinds). We discuss our findings in light of theories regarding the representation of negation and the relationship between language and thought.

Keywords: negation; philosophy of language; language and thought

Introduction

Due to the early emergence of words such as "no" and "not" in children, and their frequent use in human discourse, it is tempting to dismiss negation as a simple concept. However, the concept of negation has long been a puzzle to philosophers, psychologists, and cognitive scientists. In order to understand the complexity of this phenomenon, consider the following thought experiment:

> Consider, for example, negation. It's easy to tell somebody that it's not going to rain. Try drawing them a picture of it's not going to rain…Think about trying to draw a picture of "there's not a giraffe standing beside me" (Fodor, 1994).

The inherent difficulty in finding a way to depict negation raises questions about the nature of the representation of negation. Is language necessary to understand an abstract concept of negation?

One challenging aspect of negation is that the words ``no'' and ``not'' play many different functions in human speech (see Bloom, 1970; Pea, 1980; Choi, 1988 for discussions of several taxonomies of negation and their trajectory in children's language acquisition). For example, you can use negation to express the *nonexistence* of an object, e.g., "There is no food in the dog's bowl." You can also use negation to express *refusal,* e.g., "No, I don't want to read." And you can express *denial* or truth-functional negation by making statements about falsehoods, e.g. "The light is not on" [*i.e., it is not true that the light is on*]. It is possible to imagine ways to represent each of these statements perceptually or through simple positive concepts, e.g., an empty bowl, a girl looking away from a book on a table, a lamp that is off. Without the language of negation, however, there is nothing perceptually or conceptually similar about these concepts. One important goal of this study is to examine whether adults can spontaneously identify the similarity of these events (i.e., a unified concept of negation) in the absence of explicit language explaining the similarity between the events.

Under a propositional account of the representation of negation, negative sentences are represented as a negative operator acting over a proposition (Clark & Chase, 1972; Carpenter & Just, 1975; Just & Carpenter, 1971, 1976). That is, all of the sentences in the previous example are "unified" by the presence of a negative operator in their representations. Where, then, does this negative operator come from – or any of the structures that underlie human thought? Fodor (1975, 2008) proposed that there must be a "language of thought", which is language-*like* in the sense that it must contain an innate "lexicon" of concepts, as well as a syntax to organize those concepts. According to Fodor, concepts are learned through a process of linking one's experiences in the world with innate concepts. Without such an underlying system, Fodor argues, concept learning (and ultimately word learning) would not be possible. Under this hypothesis, the negative operator that creates a unified concept of negation exists in the lexicon of the language of thought.

Another possibility is that natural language itself is the vehicle for representing and structuring thought. According to Hinzen (2007), there is a conceptual framework that underlies human thought, and language is necessary to organize these concepts into complex propositions. If natural language can provide the same kind of structure that Fodor (1975) argues is necessary for complex thoughts to arise, then the existence of a separate Language of Thought becomes redundant (deVilliers, 2010; Collins, 2000). Under this hypothesis, the development of human language is required to understand a unified concept of negation.

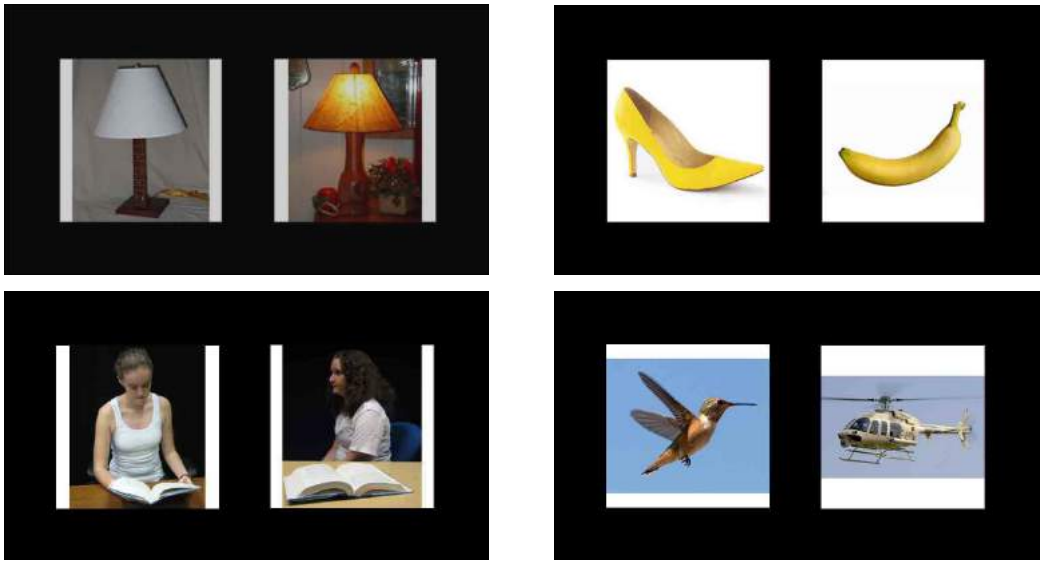


Figure 1: Examples of Negation stimuli (left) and Natural Kind stimuli (right).

The purpose of the present study is to examine, first, whether people have a general concept of negation – that is, a concept of negation that unifies a variety of negated events, actions, objects, and states of being, despite these all having very different perceptual features. That is, do adults recognize the similarity between an empty food bowl, a person refusing to read, and a lamp that is off, without someone explicitly describing these scenes using negation? Second, the study will examine the role that language plays in forming and understanding this concept. If negations are represented as propositions, in an organized and structured way, there must be some mechanism for representing them as such. We propose that natural language can provide an individual with the necessary structures to represent thoughts propositionally, and that natural language might be required to hold a generalized concept of negation.

Method

To test our first question of whether people can identify an abstract concept of negation in the absence of verbal descriptions, we created a non-verbal anticipatory looking task. Participants viewed pairs of photographs in which one image represented an affirmative event and the other image represented the negated version of that event; after several seconds, an animation occurred around the negative event. These stimuli were designed to create a "context of plausible denial" – that is, a context in which the formation of a negative proposition would be a likely response (Wason, 1965). This allowed us to test whether participants would spontaneously identify an abstract concept of negation when looking at pictures without hearing language, allowing the role of language to be manipulated and evaluated separately.

To manipulate participants' ability to use language during this task, half of our participants engaged in a language interference/verbal shadowing task, in which participants listened to a story through headphones and repeated what they heard out loud simultaneously. This task has been shown to interfere with adults' ability to utilize language in abstract cognitive tasks (Hermer-Vazquez, Spelke, & Katsnelson, 1999, Newton & de Villiers, 2007). We hypothesized that participants would be able to identify the negative event in the absence of verbal interference, but would perform at chance when shadowing language.

To test whether the effects of verbal interference are specific to an abstract concept like negation, as opposed to simply distracting participants from the task, we developed a control task to test participants' ability to form a different concept – one that was equally varied but that potentially would not require language to understand. We selected "natural kind objects" as a control concept because it is a broad, complex concept, which cannot be organized around single perceptual cues alone, but which we believed would not require language. For example, young children (Gelman, 1988; Gelman & O'Reilly, 1988) and pre-verbal infants (Booth & Waxman, 2002; Shutts, Markson, & Spelke, 2009) appear to be sensitive to the distinction between natural kinds vs. artifacts. This work suggests that it may be possible to represent the natural kind concept without requiring propositions with a language-like structure.

Participants

Participants were recruited through two psychology courses. The participants were all undergraduate students and all but one of the participants were female (due to the nature of the institution's population). Participants received credit towards their final grades for participation. After excluding participants for lack of attention to the task (see "Data Processing"), our final sample included 84 participants (negation, no shadowing: n=18; negation, shadowing: n=17;

natural kind, no shadowing: n=27; natural kind, shadowing: 20).

Stimuli

Stimuli consisted of pairs of photographs that portrayed either an affirmative or negated event (in the experimental condition) or a natural kind or an artifact object (in the control condition). To ensure that the photographs in each pair were equally salient, we conducted a pilot test in which adults (N = 12) viewed a total of 72 pairs of photographs for three seconds. Paired-samples t-tests were conducted to determine if the total looking time was greater for one picture more than the other in each pair using a conservative alpha level of .1. This resulted in the removal of 9 pairs from the negation condition and 8 pairs from the natural kind condition. One additional pair was randomly selected to be removed from the natural kind group, in order to have an equal number of pairs in each group. This left 22 pairs of photographs in each group for the final study. Figure 1 shows examples of the stimuli used in each task.

One of our primary hypotheses for the study was to examine whether participants would be able to identify a unified concept of negation from perceptually varied stimuli, without explicit verbal descriptions. To do this, we needed to be sure that the negative stimuli were sufficiently varied (i.e. drawing from many different types/functions of negation) and could not be united by some other concept. To do this, we created four different categories of negation: non-functional (4 exemplars, e.g., affirmative = a digital alarm clock that is showing the time, and negative = a digital alarm clock that is not showing the time), nonexistence (5 exemplars, e.g., affirmative = a dog with food in its bowl, and negative = a dog with an empty food bowl), unexpected state (6 exemplars, e.g., affirmative = a lamp that is on, and negative = a lamp that is off), and refusal (7 exemplars, e.g., affirmative = a girl who is reading, and negative = a girl sitting next to but looking away from an open book).

During the experiment, the pairs of photographs were animated so that each pair of photographs would be presented as still photographs for three seconds, after which the target photo (the negated photograph in the experimental condition, and the natural kind photograph in the control condition) would animate. The animation consisted of a cartoon foot emerging and moving down to squish the target photo to 20% of its original height. The foot then moved back up and the photograph returned to its original height as the foot receded. The animation, from the emergence of the foot to its disappearance, took a total of three seconds. Thus, each pair of photographs was on the screen for a total of six seconds, half of which consisted of the animation phase. Five seconds of black screen separated each animation.

The experiment was constructed in Tobii Studio. Two pseudo-random lists were created, specifying the order in which the participants would see the stimuli, and participants in all of the conditions were randomly assigned to one of the two lists. In both conditions, four photographs were selected as "example photographs". In the negation task, the four example photos included one from each negation type. Participants were not told that these were example photographs, but the examples differed in that each was displayed twice, once with the target picture on the left and once with the target picture on the right. This was done to draw participants' attention to the content of the photographs themselves (as opposed to simply the position on the screen), and to familiarize them to the kinds of stimuli and animation that they would be seeing.

Procedure

This experiment used a 2x2 between-subjects design. Half of the adults were tested on the negation task, and half were tested on the natural kind task. Within each of these conditions, half of the participants were tested with verbal shadowing and half were tested without verbal shadowing.

The experiment was run on a Tobii 1750 eye tracker. Participants were told that they would see pairs of photos on the screen in front of them, and that their job was to watch the pictures and pay attention to what they saw on the screen. Participants in the verbal shadowing condition were told that they would listen to an audio book through a pair of headphones (a passage from 1984 by George Orwell), and would have to repeat what they heard out loud as they listened. Participants were told to speak as simultaneously as possible with the speech they heard, and to be as accurate as possible, but to continue speaking if they made any mistakes, as the most important thing was that they spoke as continuously and fluidly as possible. Participants were then reminded that as they listened and spoke, they would have to keep their attention on the pictures they saw on the screen in front of them.

After the tasks were explained, the experimenter asked the participants in the shadowing condition to practice the verbal shadowing for 30 seconds. Participants in the verbal shadowing condition were videotaped throughout the duration of the experiment so that their performance on the shadowing task could be evaluated at a later point. After 30 seconds of practice, the experimenter started the videocamera and began the experiment. Participants who were not in the shadowing condition were not videotaped, and the experiment was started immediately after explaining the eyetracking task. In both conditions, the experimenter stepped out of the room as soon as the experiment began.

Data Processing

Areas of Interest (AOIs) were created around each photograph in each pair, with the negated event or the natural kind image designated the "target" photograph. The Total Fixation Duration (a measure of the durations of all fixations within an AOI in seconds) within each AOI was collected for the three seconds prior to the start of the animation began. The AOIs were constructed so that the computer would only record a fixation if a person's gaze fixated within the boundaries of the photograph. Thus, although the combined total fixation time possible between

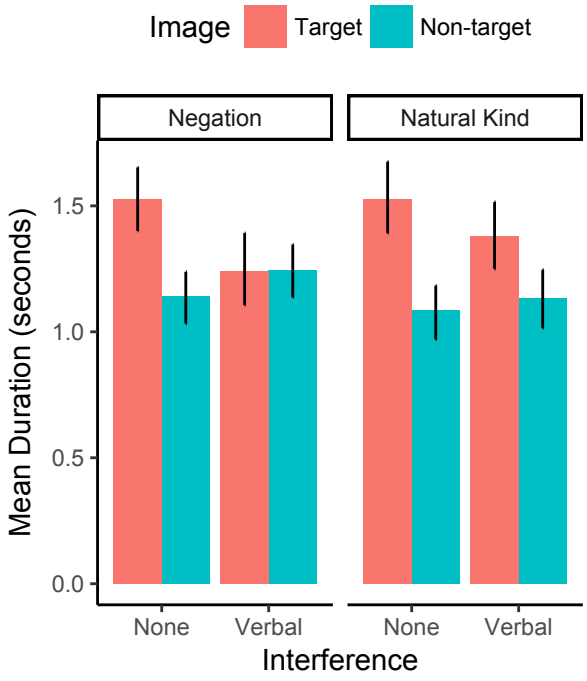


Figure 2: Mean Total Fixation Duration for each condition. Mean looks to the target image (negative event, natural kind) are shown in red and mean looks to the non-target image (affirmative event, artifact) are shown in green. Error bars portray 95% confidence intervals.

the target and non-target AOI was 3 seconds, it is possible that the combined total fixation time would be less than that if participants were fixating their gaze anywhere on the black screen outside of the pictures. The percentage of time that a person's gaze was anywhere on the screen throughout the experiment was also noted, and participants whose total gaze dropped below 60% were excluded from the analysis. This resulted in the exclusion of 18 participants from analysis.

The videotapes of the participants in the shadowing condition were analyzed to determine that the participants had continued speaking throughout the duration of the experiment. Any participant who stopped shadowing for more than 2 seconds was excluded from the analysis. This analysis resulted in the exclusion of three participants, two from the negation group and one from the natural kind group.

Results

Can participants spontaneously identify a unified concept of negation?

First, we asked whether adults would be capable of spontaneously identifying the negative concept in the negation task. To do this, we looked only at the negation/no interference condition to determine if participants looked more to the target picture compared to the non-target picture when they were not subject to verbal interference.

Mean looking times to the target and non-target picture for all conditions are shown in Figure 2. In the negation/no shadowing condition (left-most bars of Figure 2), mean looking time towards the target picture was greater than the mean looking time to the non-target picture (M_{target} = 1.53 s, $M_{non\text{-}target}$ = 1.14 s). A paired-sample t-test showed that this difference was significant, t(17) = -3.82, p<.01, suggesting that participants were able to spontaneously identify the negative concept and look to the correct picture in anticipation of the animation.

To make sure that participants were truly responding to the general concept of negation, and not a simpler sub-concept such as "refusal" or "failure", we examined each of the four subtypes of negation separately. Participants looked more to the target image compared to the non-target image in the non-functional subtype (t(17) = -2.20, p <.05), the nonexistence subtype (t(17) = -3.23, p < .01), and the refusal subtype (t(17) = -4.47, p<.001), but not the unexpected state subtype (t(17) = -1.47, p = .16). The fact that participants spontaneously looked towards the target picture for a wide range of subtypes (i.e., many perceptually different types of images and events) suggests that participants were identifying and responding to a general concept of negation.

Does verbal interference impair participants' ability to identify a concept of negation?

The previous analysis indicated that participants were able to spontaneously identify the negative concept, looking significantly more to the target (negative event) picture compared to the non-target (affirmative event) picture prior to the animation (t(17) = -3.82, p<.01). In the negation/verbal shadowing condition, however, mean looking time was nearly identical between the target and non-target images (M_{target} = 1.239 s, $M_{non\text{-}target}$ = 1.243 s, t(16) = 0.04, p = .97), suggesting that participants' ability to identify the negative concept was impaired under verbal interference. In the natural kind task, mean looking time to the target picture was greater than the mean looking time to the non-target picture in both the no shadowing condition (M_{target} = 1.53, $M_{non\text{-}target}$ = 1.08 s, t(26) = -3.66 s, p < .01) and the shadowing condition (M_{target} = 1.38 s, $M_{non\text{-}target}$ = 1.13 s, t(19) = -2.22, p < .05).

To examine the effect of the interference condition on whether participants looked more to the target or the non-target photograph, separate two-way ANOVAs were conducted for the negation condition and the natural kind condition. In the negation condition, there was a significant interaction between the target image and verbal interference (F(1,66) = 9.39, p<.01), suggesting that participants' ability to spontaneously identify the negative concept was significantly impaired by verbal interference. In the natural kind condition, there was a significant effect of target image (F(1, 90) = 26.12, p < .001) but no effect of verbal interference (F(1, 90) = 0.56, p = .46) and no interaction

between target image and verbal interference ($F(1.90) = 2.09$, $p = 0.15$), suggesting that participants' ability to spontaneously recognize the natural kind concept was not impaired under verbal interference.

To examine the interaction between task (negation vs. natural kind) and verbal interference (shadowing vs. no shadowing), we fit a linear mixed effects model[1]. For this analysis we calculated Differential Looking Scores (DLS) for each trial by dividing the difference between target and nontarget fixation duration by the total fixation duration to either picture, giving us a measure of the proportion of looks to target relative to the overall looking time for a given participant on a given trial. However, this model did not produce any significant effects (main effect of task: $\beta = -0.016$, $p = .81$; main effect of verbal interference: $\beta = -0.078$, $p = .19$; interaction between task and verbal interference: $\beta = -0.084$, $p = .35$). Because the two tasks used an entirely different set of stimuli, it is possible the variability in items makes it difficult to compare the two tasks in this way. In the General Discussion we discuss alternative possibilities for control tasks to test whether the effect of verbal interference is specific to negation.

General Discussion

We hypothesized that language is necessary for adults to implicitly recognize a unified concept of negation. We expected participants in the non-interference negation condition would implicitly learn to look towards the target picture in the seconds before it animated; that is, with implicit language abilities intact the resemblance across the items as negatives would be evident. Under conditions of verbal interference, where participants cannot implicitly use language to understand the concept, we predicted that participants would be unable to identify the resemblance across the diverse instances of negation. Conversely, we predicted that participants in the natural kind condition (a concept that would not necessarily require propositional structure) would look to the target picture regardless of verbal interference.

These results offer support for our hypotheses. First, participants in the negation condition without verbal interference *were* able to spontaneously identify exemplars of the negative concept despite a lack of verbal instructions telling them what concept to look for. The fact that participants, who were not told anything about the images they would see and were simply told to look at the pictures, were able to look at the negative event in anticipation of the animation suggests that there is some concept of negation that unites these very different exemplars.

Second, participants' ability to identify the tested concept was impaired by verbal interference in the negation condition, but not the natural kind condition. Non-shadowing participants in the negation group looked significantly more to the negation picture than the affirmative picture, while shadowing participants did not look significantly more to one photograph more than the other. In the natural kind group, participants looked more to the natural kind photograph than the artifact photograph, and, critically, this difference was not affected at all in the shadowing condition. This provides support for the hypothesis that language is required to understand a concept of negation, but not to understand other concepts, such as natural kinds.

One possible limitation of this study is that participants may have "passed" the negation task by identifying a simpler concept, rather than truly identifying a general concept of negation. We attempted to address this in our design by creating stimuli that represented a wide range of types of negation. In our analysis of the data, we found that participants were significantly more likely to look to the target picture in three of the four subtypes of negation that we included in our stimuli, suggesting that participants were responding to a general concept of negation rather than succeeding on only a small subset of trials.

Another possible limitation of this study is that the control task (natural kinds) may have simply been easier than the negation task. Although we do not think this is the case (overall looking time to the target picture in the no interference condition was identical across the two conditions, M = 1.53 seconds), this could be addressed in future work by using additional control tasks. One possibility would be to use the negation task stimuli with *affirmative* pictures as the target image, with the prediction that verbal interference should not affect looks to the affirmative picture. A downside to this option is that it isn't clear whether there is an underlying unifying concept of affirmation that would be spontaneously identified by participants – that is, participants might find the affirmation condition challenging even in the *absence* of verbal interference. Other possible control conditions could include a wider range of control concepts thought to not require propositional structure, or using an attentional control task such as rhythmic tapping, which has been used in past verbal shadowing studies (Hermer-Vazquez, Spelke, & Katsnelson, 1999, Newton & de Villiers, 2007).

Our results suggest that some kind of linguistic structure is necessary to understand a general concept of negation. The language-like structure that is required to support propositional thinking could come from a "language of thought" (e.g. Fodor, 1975), or it could come from the structure of natural language (e.g., Hinzen, 2007). One way to tease apart these possibilities would be to examine whether pre-verbal children or non-verbal animals can understand a general concept of negation. Many "language of thought" hypotheses propose that the LOT exists preverbally in children (and facilitate the development of natural language), and perhaps to some extent in non-verbal animals as well (Fodor, 1975, 2008). In the domain of animal research, Premack (1980) attempted to teach three chimpanzees a symbolic system based on plastic tokens that included a token for the word "not." The attempt was only

[1] Model specification: `DLS ~ task x interference + (1 | subject) + (interference | item)`

partially successful for one chimp, and unsuccessful for the other two. This would suggest that chimpanzees, at least, are unlikely to be able to represent an abstract concept of negation.

Research on children's acquisition of negation suggests that children begin producing the word "no" to express refusal as early as 12 months (Pea, 1980), and that children as young as 26 months understand denial negation (Austin, Theakston, Lieven, & Tomasello, 2014; Feiman, Mody, Sanborn, & Carey, 2017), though this may be task or context-dependent (Nordmeyer & Frank, 2014; 2018; Reuter, Feiman, & Snedeker, 2017). Under a "language of thought" hypothesis, children should be able to identify and understand a general, non-verbal concept of negation even if they cannot yet articulate this concept in natural language, and therefore children would perhaps be capable of passing a task similar to ours. If natural language is providing the structure to represent propositional negation, however, we would expect to see developmental changes in whether children are capable of understanding a general concept of negation, with pre-verbal (or pre-negation) children failing and older children succeeding.

Future work could examine the role of language in understanding other logical operators, such as "and" and "or", or quantifiers such as "some" and "all". Many philosophers of language have suggested that an important role of language is linking and connecting simple concepts into an infinite number of thoughts and combinatorial, complex concepts (Fodor, 1975, 2008; Carruthers, 2002; Hinzen, 2007). If this is true, words such as "and" and "or" would be vital to a system of thought, and conversely, it is possible that language is again necessary to form thoughts that require these logical connectives. These logical connectives and quantifiers are necessary aspects of human reasoning, and studying the role that language plays in understanding them could push our understanding beyond simply how we *perceive* the world, providing insight as well into how we reason about the world around us.

References

Austin, K., Theakston, A., Lieven, E., & Tomasello, M. (2014). Young children's understanding of denial. *Developmental Psychology*, *50*, 2061–2070.

Bloom, L. (1970). *Language development: Form and function in emerging grammars*. Cambridge: The MIT Press.

Booth, A. E., & Waxman, S. (2002). Object names and object functions serve as cues to categories for infants. *Developmental Psychology, 38*(6), 948-957.

Carpenter, P., & Just, M. (1975). Sentence comprehension: A psycholinguistic processing model of verification. *Psychological Review*, *82*, 45–73.偏

Carruthers, P. (2002) The cognitive functions of language, *Behavioral and Brain Sciences, 25*, 657-726.

Choi, S. (1988). The semantic development of negation: A cross-linguistic longitudinal study. *Journal of Child Language*, *15*, 517–531.偏

Clark, H. H., & Chase, W. G. (1972). On the process of comparing sentences against pictures. *Cognitive Psychology, 3*, 472-517.

Collins, J. (2000). Theory of mind, logical form and eliminativism. *Philosophical Psychology, 13*(4), 465-490.

de Villiers, J. (2010). Structured thought: Language and false belief reasoning. In A. M. Leslie, & T. C. German (Eds.), *Handbook of theory of mind*. Psychology Press.

Feiman, R., Mody, S., Sanborn, S., & Carey, S. (2017). What do you mean, no? Toddlers' comprehension of logical "no" and "not". *Language Learning and Development*, *13*(4), 430-450.

Fodor, J. A. (1975). *The language of thought*. Cambridge: Harvard University Press.

Fodor, J.A. (Contributor). *The human language series: Part 1: Discovering the human language.* Searchinger, G. (Director). (1994). [Video/DVD]

Fodor, J. A. (2008). *LOT 2: The language of thought revisited*. Oxford: Oxford University Press.

Gelman, S. A. (1988). The development of induction within natural kind and artifact categories. *Cognitive Psychology, 20*, 65-95.

Gelman, S. A., & O'Reilly, A. W. (1988). Children's inductive inferences within superordinate categories: The role of language and category structure. *Child Development, 59*, 876-887.

Hermer-Vasquez, L., Spelke, E. S, & Katsnelson, A.S., (1999). Sources of flexibility in human cognition: Dual-task studies of space and language. *Cognitive Psychology*, *39*, 3-36.

Hinzen, W. (2007). *An essay on names and truth*. Oxford: Oxford University Press.

Just, M., & Carpenter, P. (1971). Comprehension of negation with quantification. *Journal of Verbal Learning and Verbal Behavior*, *10*, 244–253.

Just, M., & Carpenter, P. (1976). Eye fixations and cognitive processes. *Cognitive Psychology*, *8*, 441–480.

Newton, A. & de Villiers, J.G. (2007) Thinking while talking: adults fail non-verbal false belief reasoning. *Psychological Science, 18*(7), 574-579.

Nordmeyer, A. E., & Frank, M. C. (2014). The role of context in young children's comprehension of negation. *Journal of Memory and Language*, *77*, 25–39.

Nordmeyer, A. E., & Frank, M. C. (2018). Early Understanding of Pragmatic Principles in Children's Judgments of Negative Sentences. *Language Learning and Development*, 1-17.

Pea, R. D. (1980). The development of negation in early child language. In D. R. Olson (Ed.), *The social foundations of language and thought* (pp. 156-186). New York, NY: W.W. Norton & Co.

Premack, D. (1976). Early concepts: Same-different, no, and the interrogative. In *Intelligence in ape and man* (pp. 131-161). Hillsdale: Lawrence Erlbaum Associates.

Reuter, T., Feiman, R., & Snedeker, J. (2018). Getting to No: Pragmatic and Semantic Factors in Two-and Three-

Year-Olds' Understanding of Negation. *Child development*, *89*(4), e364-e381.
Shutts, K., Markson, L., & Spelke, E. S. (2009). The developmental origins of animal and artifact concepts. In B. Hood, & L. Santos (Eds.), *The origins of object knowledge* (pp. 189-210). Oxford: Oxford University Press.
Wason, P.C. (1965). The contexts of plausible denial. *Journal of Verbal Learning and Verbal Behavior, 4*(1), 7-11.

Neural Substrates Mediating the Utility of Instrumental Divergence

Kaitlyn G. Norton (kgnorton@uci.edu)
Department of Psychological Sciences, UC Irvine
Irvine, CA 92697

Mimi Liljeholm (m.liljeholm@uci.edu)
Department of Cognitive Sciences, UC Irvine
Irvine, CA 92697

Abstract

We assessed the neural substrates mediating a recently demonstrated preference for environments with high levels of instrumental divergence – a formal index of flexible operant control. Across choice scenarios, participants chose between gambling environments that differed in terms of both instrumental divergence and expected monetary pay-offs. Using model-based fMRI, we found that activity in the ventromedial prefrontal cortex scaled with a divergence-based measure of expected utility that reflected the value of both divergence and monetary reward. Implications for a neural common currency for information theoretic and economic variables are discussed.

Keywords: instrumental divergence; flexible control; utility; model-based fMRI

Introduction

A series of recent studies (Mistry & Liljeholm, 2016; Liljeholm et al., 2018) have demonstrated that individuals strongly prefer environments in which instrumental divergence – the degree to which alternative actions differ with respect to their outcome probability distributions – is relatively high. A high level of instrumental divergence is a necessary feature of flexible control: If all available action alternatives have identical, or very similar, outcome distributions, such that selecting one action over another does not significantly alter the probability of any given outcome state, an agent's ability to exert flexible control over its environment is considerably impaired. Conversely, when available action alternatives produce distinct outcomes, discrimination and selection between actions allow an agent to flexibly obtain the currently most desired outcome. Since subjective outcome utilities often change from one moment to the next, flexible instrumental control is essential for reward maximization and, as such, may have intrinsic value, serving to reinforce and motivate decisions that guide the organism towards high-agency environments (Liljeholm, 2018). In the current study, we investigate the neural substrates mediating the apparent preference for high instrumental divergence.

Previous work suggests that the ventromedial prefrontal cortex (vmPFC) retrieves and ranks the values of decision outcomes, and that these value signals are subsequently used to compute decision values (see O'doherty, 2011 for review). Intriguingly, activity in the vmPFC scales with the values of a wide variety of goods, including food, money, books DVDs, and clothes, suggesting a common neural value-scale for distinct stimulus categories (Chib et al., 2009; McNamee et al., 2013). It is unknown, however, whether this common value-scale might also extend to more abstract, cognitive, commodities, such as instrumental divergence. Here, using a task in which participants choose between gambling environments based on differences in both instrumental divergence and monetary pay-offs, we combine computational cognitive modeling with functional MRI to investigate neural representations of the utility of instrumental divergence.

Method

Participants

Twenty undergraduates at the University of California, Irvine (11 females; mean age = 21.2 ± 4.65) participated in the study for monetary compensation. The sample size was determined based on an a priori power analysis of data from a previously published study (Mistry & Liljeholm, 2016), indicating that 18 subjects were required to demonstrate a clear behavioral preference for high instrumental divergence at a power of 90% given a 0.05 threshold for statistical significance. All participants gave informed consent and the Institutional Review Board of the University of California, Irvine, approved the study.

Task & Procedure

The task is illustrated in Figure 1. At the start of the experiment, participants were instructed that they would assume the role of a gambler in a casino, playing a set of four slot machines (i.e., actions, respectively labeled A1, A2, A3, and A4) that yielded three different colored tokens (blue, green and red), each worth a particular amount of money, with different probabilities. They were further told that, in each of several blocks, they would be required to first select a room in which only two slot-machines were available, and that they could only choose between the two machines in the selected room on subsequent trials in that block. Finally, participants were instructed that, while the outcome probabilities would remain constant throughout the

study, the values of the tokens would change at various times, and these changes might occur after the participant had already committed to a particular pair of machines in a given block. Consequently, although changes in value were explicitly announced, and the current values of tokens were always printed on their surface, a participant might find themselves in a room in which the values of the two available actions had suddenly been altered.

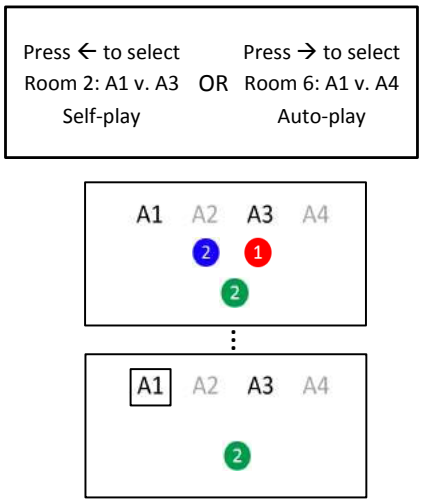


Figure 1: Task illustration showing the room-choice screen at the beginning of a block (top), and the choice (middle) and feedback (bottom) screens on a trial in the selected room.

Two distinct probability distributions over the three possible token outcomes were used and the assignment of outcome distributions to slot machines was such that two of the machines (either A1 and A2 or A1 and A3, counterbalanced across subjects) always shared one distribution, while the other two machines shared the other distribution. This yielded a low (zero) divergence for rooms in which the two available slot machines shared the same probability distribution, and a relatively high divergence for rooms in which slot machines had different outcome probability distribution. The unpredictability (i.e., Shannon entropy) of outcomes given a particular machine was held constant for all machines. Three token-reward distributions were used, changing intermittently across blocks, such that expected monetary pay-offs were either the same across rooms, or differed across rooms in either the same or opposite direction of instrumental divergence. In addition to mimicking dynamic changes in the utilities of natural rewards, the sporadic changes in token reward values across blocks allowed us to pit the value of instrumental divergence against that of monetary reward.

Given a constant outcome entropy level, increases in instrumental divergence are accompanied by increases in the perceptual diversity of obtainable outcomes – a variable previously shown to elicit preferences in economic tasks (Ayal & Zakay, 2009). To rule out perceptual diversity as an explanation for any effects of instrumental divergence, gambling rooms differed in terms of whether the participant was allowed to chose freely between slot machines in the room (self-play) or a computer algorithm alternated between machines across trials in that room (auto-play). In auto-play rooms, participants were still required to press a key corresponding to the slot machine indicated by the computer, to control for movement execution. Critically, in the absence of voluntary choice, high-divergence no longer yields flexible instrumental control. However, the alternating computer algorithm still yields greater perceptual diversity in high- than in low-divergence rooms. Consequently, if choices were driven by a desire to maximize perceptual diversity, rather than instrumental divergence, they should not differ depending on whether the participant or an alternating computer algorithm choose between the slot machines in a room. In addition to controlling for perceptual diversity, this self- vs. auto-play manipulation relates the preference for instrumental divergence to a well-established preference for free over forced choice (e.g., Leotti & Delgado, 2011).

There were a total of 44 blocks, with participants choosing between two gambling rooms at the start of each block (the decision of interest), followed by 3-5 gambling trials within the selected room. The order different reward distributions, and of room choice scenarios, was counterbalanced across subjects. Before starting the gambling task participants were given a practice session in order to learn the probabilities with which each slot machine produced the different colored tokens. If a participants' estimate of any given probability deviated by more than 0.2 from the programmed probability, they were returned to the beginning of the practice phase, and this continued until all rated probabilities were within 0.2 points of programmed probabilities. At the end of the study, participants again provided estimates of the action-token probabilities.

Computational Models

Instrumental divergence is formalized as the Jensen-Shannon divergence of instrumental sensory-specific outcome probability distributions (Liljeholm et al., 2013). Let P_1 and P_2 be the respective outcome probability distributions for two available actions, let O be the set of possible outcomes, and $P(o)$ the probability of a particular outcome, o. The instrumental divergence (ID) is:

$$ID = \frac{1}{2}\sum_{o\in O}\log\left(\frac{P_1(o)}{P_*(o)}\right)P_1(o) + \frac{1}{2}\sum_{o\in O}\log\left(\frac{P_2(o)}{P_*(o)}\right)P_2(o),$$

where

$$P_* = \frac{1}{2}(P_1 + P_2)$$

We defined the *expected value* (EV) of each slot machine as the sum over the products of its transition probabilities and token utilities. In turn, the expected *monetary* value of a gambling room is simply the mean of the EVs of slot machines in that room. To model the utility of instrumental divergence, a second variant of EV was specified by adding the term *w*ID* to the expected monetary value of a room, where the free parameter *w* represents the subjective utility of instrumental divergence and *ID* is the divergence of the particular room. Thus, in this variant, the EV of a room reflects *both* the monetary pay-off and the instrumental divergence associated with that room. For both models, a softmax distribution with a noise parameter, τ, was used to translate expected room values into choice probabilities, and free parameters were fit to behavioral data by minimizing the negative log likelihood of observed choices. Choice scenarios in which at least one room option was both high divergence and self-play (HDSP), yielding high *instrumental* divergence, and those in which the high-divergence room option was auto-play, or both rooms had zero divergence (HDAP), were modeled separately. The corrected Akaike Information Criterion (AICc) was used for behavioral model comparisons.

Neuroimaging Acquisition & Analyses

All MR images were obtained in a 3T Siemens Prisma Scanner, fitted with a 32-channel RF receiver head coil, padded to minimize head motion, at the facility for imaging and brain research (FIBRE) at the University of California, Irvine. Functional images covered the whole brain with 48 continuous 3-mm thick axial slices with T2*-weighted gradient echoplanar imaging (TR=2.65s, TE=28ms, 3-mm^2 in-plane voxel size, 64 x 64 matrix). All participants had a high-resolution structural image taken before functional scanning commenced (T1-weighted FSPGR sequence: 208 continuous 0.8-mm axial slices 0.4-mm^2 in-plane voxel size; 640 x 640 matrix). All stimulus materials were presented, and all responses recorded, using MATLAB. All imaging data was preprocessed with MATLAB and SPM12. Functional images were preprocessed with standard parameters, including slice timing correction, spatial realignment, coregistration of the high-resolution structural image to functional images, segmentation of the structural image into tissue types, spatial normalization of functional images into MNI space, and spatial smoothing with an 8mm FWHM kernel.

All imaging data was analyzed using MATLAB and SPM12. At the first level, two general linear models (GLMs) were specified for each participant. In both GLMs, two regressors respectively specified the onsets of room choice screens for HDSP and HDAP choice scenarios. In the first GLM, these onsets were parametrically modulated by the absolute difference between rooms in their expected monetary pay-offs; in the second GLM these onsets were parametrically modulated by the absolute difference between rooms in their divergence-based utility, which reflected both the monetary pay-off and the level of divergence associated with each room. In addition, in both GLMs, two regressors indicated the onsets of choice screens on each trial within a selected room, for self-play and auto-play rooms respectively, and each of these were parametrically modulated by the expected monetary value of the chosen slot machine. Finally, both GLMs included a single regressor indicating the onsets of trial feedback screens, modulated by the monetary reward obtained on each trial, as well as regressors indicating separate scanning runs and accounting for the residual effects of head motion.

Fixed effects models were estimated using restricted maximum likelihood and an AR(1) model for temporal autocorrelation. Group-level statistics were generated by entering contrasts of first level parameter estimates into between-subject analyses. All effects are reported at a whole brain corrected $p < 0.05$ level, using cluster size thresholding (CST) to adjust for multiple comparisons. AlphaSim, a Monte Carlo simulation, was used to determine cluster size and significance. For an individual voxel probability threshold of p=0.005, a minimum cluster size of 148 MNI transformed voxels resulted in an overall significance of $p < 0.05$.

Results

Behavioral Results

Participants required on average 2.1 (*SD*=0.3) cycles of practice on the action-token probabilities. Mean probability ratings, obtained right before and right after the gambling phase, are shown in Table 1.

Table 1: Mean probability ratings with standard deviations. Programmed probabilities are shown in the top row. Mean ratings, obtained before and after the gambling task, are averaged across identical objective probabilities, yielding three unique values.

	0.7	**0.0**	**0.3**
Before	0.69 ± 0.06	0.00 ± 0.00	0.31 ± 0.05
After	0.67 ± 0.10	0.00 ± 0.00	0.32 ± 0.05

The decision of interest was that at the beginning of each block, when participants choose between rooms that differed in terms of their divergence, expected monetary pay-offs and self- vs. auto-play. Model-derived choice probabilities and AICc scores for these decisions are listed in Table 2.

Table 2: Mean room-choice probabilities derived using the divergence-based and conventional models of expected value (EV), and associated AICc scores, for HDSP and HDAP choice scenarios, with standard deviations.

	Choice Probabilities		AICc Scores	
	HDSP	HDAP	HDSP	HDAP
Divergence EV	0.65 ± 0.13	0.58 ± 0.07	19.5 ± 2.7	38.3 ± 7.4
Conventional EV	0.55 ± 0.05	0.57 ± 0.07	21.6 ± 5.6	36.8 ± 7.5

A repeated measures analysis of variance (ANOVA) revealed that the model-derived probabilities of observed behavioral choice preferences were significantly greater for the divergence-based utility algorithm than for the conventional utility model, and this difference was significantly greater for HDSP choice scenarios than for HDAP choice scenarios, yielding a significant main effect of EV model, $F(1,19)=12.40$, $p<0.005$, as well as a model by choice scenario interaction, $F(1,19)=9.52$, $p<0.01$. Accordingly, there was also a significant interaction for the AICc scores, $F(1,19)=7.71$, $p<0.05$, such that scores were significantly lower, indicating a better fit, for the conventional than for the divergence-based utility model in HDAP blocks ($t(19)=5.2$, $p<0.001$) while being lower for the divergence-based utility model, albeit with only marginal significance ($p=0.14$) in HDSP blocks.

Neuroimaging Results

As with the behavioral data, the period of interest was the choice made at the beginning of each block, between two gambling rooms that differed in terms of divergence, monetary pay-offs and free choice. As illustrated in Figure 2, neural activity in the ventromedial prefrontal cortex (vmPFC) was parametrically modulated by the absolute difference in divergence-based EV between rooms, when at least one room option was both high-divergence and self-play (HDSP) but not when the high divergence room was auto-play, or both room options had zero divergence (HDAP). No significant effects of the difference between room options in expected monetary pay-offs emerged in this region. A similar pattern of results, with activity scaling selectively with the absolute difference in divergence-based EV between rooms options in HDSP choice scenarios, was found in the middle frontal gyrus, as well as the premotor cortex. Once a room had been selected, activity in a more dorsal aspect of the vmPFC, extending into the dorsal medial prefrontal cortex scaled with the expected monetary pay-off of the chosen slot machine, in self-play but not in auto-play rooms, as did activity in the lateral orbitofrontal cortex, posterior right middle temporal gyrus and right dorsolateral prefrontal cortex.

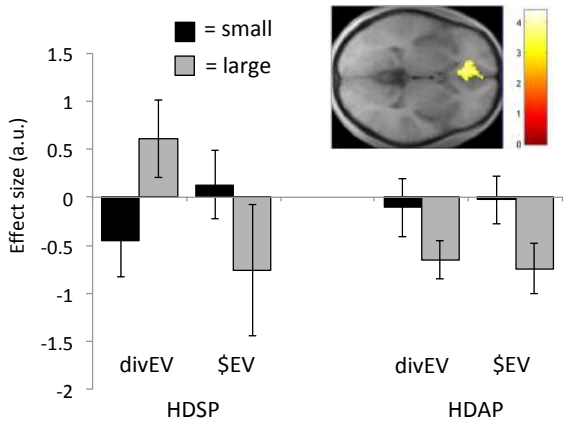


Figure 3: Map of the t-statistics for a test of differential parametric modulation by the difference across rooms in divergence-based expected value (divEV) for choice scenarios in which at least one high-divergence room option was self-play (HDSP) versus those in which the high-divergence room option was auto-play, or both options had zero-divergence (HDAP), showing significant effects in the vmPFC. Bar plots show effect sizes (y-axis) extracted from 4 mm spheres centered on the peak coordinate (x, y, z = -4, 34, -4), for small and large differences in divEV and monetary expected values ($EV), in HDSP and HDAP choice scenarios. Error bars=SEM.

Discussion

Countless studies on motivated behavior have investigated the neural representation of primary and monetary rewards (Abe & Lee, 2011; Abler et al., 2009; Belova et al., 2007; Cador e al., 1989). Here, having previously demonstrated a behavioral preference for instrumental divergence – a formal index of flexible operant control – we explored neural substrates mediating the influence of this information theoretic variable on economic choice. Specifically, participants were scanned with fMRI as they chose between gambling rooms that differed with respect to both instrumental divergence and expected monetary pay-offs. Using a model-based analysis, we found that activity in the ventromedial prefrontal cortex (vmPFC) scaled with a divergence-based measure of expected utility that reflected both instrumental divergence and monetary pay-offs.

Considerable evidence from neurophysiological and neuroimaging studies suggest that the vmPFC encodes the subjective values of primary rewards, such as tastes and odors (Rolls et al., 2003; Anderson et al., 2003; Small et al., 2003), as well as visual stimuli, including the attractiveness of faces or pictorial scenes (O'Doherty et al., 2003; Kirk et al., 2009), and more abstract goods, like social praise (Elliot et al., 1997) and monetary gain (O'Doherty et al., 2001).

Two notable features of the vmPFC shed important light on the current results: First, value encoding in the vmPFC appears to be relative, such that the value signal for a particular stimulus depends on the values of other, proximal, stimuli (O'Doherty, 2011). One might expect, thus, that the vmPFC signal would respond most clearly to a *difference in value* between concurrently available stimuli. Second, recent findings suggest that the vmPFC encodes stimulus values that are independent of the particular stimulus category, essentially implementing a common neural value scale for different types of goods (Chib et al., 2009; McNamee et al., 2013). The currently demonstrated value signal in the vmPFC, corresponding to a difference between options in divergence-based utility, suggest that this common value scale can be extended to a relative analysis of exceedingly abstract concepts.

Our previous work has implicated the right supramarginal gyrus (rSMG) of the inferior parietal lobule in encoding instrumental divergence. Specifically, using a simple value-based decision-making task, Liljeholm et al. (2013) found that activity in the rSMG scaled parametrically with trial-by-trial estimates of instrumental divergence, and that this signal was dissociable from other information theoretic and motivational variables, including outcome entropy and expected utility. In a subsequent task, aimed at assessing neural substrates mediating the acquisition of goal-directed vs. habitual instrumental behavior, Liljeholm et al., (2015) found that activity in the rSMG increased across blocks of instrumental acquisition in a high-divergence, but not in a zero-divergence, condition. In contrast, we did not find any effects of instrumental divergence in the rSMG in the current study. There are several possible reasons for this discrepancy: First, none of the previous studies assessed the motivational significance of instrumental divergence, in terms of a behavioral preference for environments with relatively high divergence. Second, in the current study, outcome probability distributions were trained to criterion prior to scanning (eliminating acquisition effects), and instrumental divergence remained constant within a room (eliminating responses to trial-by-trial fluctuations in divergence). Further work is needed to determine how these differences may account for a differential engagement of the rSMG.

A fundamental property of stimuli that possess intrinsic value is their ability to transfer that valence to neutral stimuli with which they are paired – a phenomenon termed *conditioned reinforcement,* that has been studied extensively using a wide range of stimuli, species and procedures (e.g., Arroyo et al., 1998; Williams, 1994). This large body of research has demonstrated that conditioned reinforcers are powerful behavioral determinants, maintaining instrumental responding in the absence of primary rewards, such as food and sex, and even serving as goals in themselves. Moreover, once established, previously neutral conditioned reinforcers can pass on their motivational significance to other neutral stimuli; For example, casino chips maintain gambling based on their association with monetary reward, which in turn obtains valence from its usefulness in acquiring primary rewards. One might expect, therefore, that any sufficiently valuable stimulus, no matter how abstract, should be able to induce conditioned reinforcement in associated arbitrary, and initially neutral, stimuli. Another important question, thus, is whether the affective properties of instrumental divergence may transfer to concomitant stimuli, and what brain regions might mediate such a processes.

Formal theories of goal-directed decisions postulate that the agent generates a "cognitive map" of stochastic relationships between actions and states such that, for each action in a given state, a probability distribution is specified over possible outcome states. These transition probabilities are then combined with current estimates of outcome utilities in order to generate action values – the basis of goal-directed choice (Doya et al., 2002). Although computationally expensive (Otto et al., 2013), the dynamic binding of outcome probabilities with utilities offers adaptive advantage over more automatic action selection, which uses cached values based on reinforcement history. However, when instrumental divergence is zero, or very low, the processing cost of goal-directed computations does not yield the return of flexible control, suggesting that a less resource-intense automatic decision strategy might be optimal. As noted, in a previous study we found evidence implicating instrumental divergence in the deployment of goal-directed and habitual behavior, and this is an important avenue for future work.

In summary, we have used model-based fMRI to investigate the neural computations mediating a behavioral preference for instrumental divergence. We found that activity in the vmPFC was significantly modulated by a variant of expected value that reflected both instrumental divergence an monetary pay-offs, but not by a conventional model of expected value, based solely on monetary gain. Our results complement previous work on the role of the vmPFC in value-based choice.

Acknowledgements

This work was supported by a CAREER grant from the National Science Foundation (1654187) awarded to Mimi Liljeholm.

References

Abe, H. & Lee, D. Distributed coding of actual and hypothetical outcomes in the orbital and dorsolateral prefrontal cortex. *Neuron* **70**, 731-741, doi:10.1016/j.neuron.2011.03.026 (2011).

Abler, B., Herrnberger, B., Gron, G. & Spitzer, M. From uncertainty to reward: BOLD characteristics differentiate signaling pathways. *BMC Neurosci* **10**, 154, doi:10.1186/1471-2202-10-154 (2009).

Anderson, A. K., Christoff, K., Stappen, I., Panitz, D., Ghahremani, D. G., Glover, G., ... & Sobel, N. (2003). Dissociated neural representations of intensity and valence in human olfaction. *Nature neuroscience*, *6*(2), 196.

Arroyo, M., Markou, A., Robbins, T. W. & Everitt, B. J. Acquisition, maintenance and reinstatement of intravenous cocaine self-administration under a second-order schedule of reinforcement in rats: effects of conditioned cues and continuous access to cocaine. *Psychopharmacology* **140**, 331-344 (1998).

Ayal S, Zakay D (2009) The perceived diversity heuristic: the case of pseudodiversity. Journal of personality and social psychology 96:559-573.

Belova, M. A., Paton, J. J., Morrison, S. E. & Salzman, C. D. Expectation modulates neural responses to pleasant and aversive stimuli in primate amygdala. *Neuron* **55**, 970-984 (2007).

Cador, M., Robbins, T. W. & Everitt, B. J. Involvement of the amygdala in stimulus-reward associations: interaction with the ventral striatum. *Neuroscience* **30**, 77-86 (1989).

Chib, V. S., Rangel, A., Shimojo, S., & O'Doherty, J. P. (2009). Evidence for a common representation of decision values for dissimilar goods in human ventromedial prefrontal cortex. *Journal of Neuroscience*, *29*(39), 12315-12320.

den Ouden, H. M., Frith, U., Frith, C. & S.J., B. Thinking about intentions. *NeuroImage* **28**, 787-796 (2005).

Doya K, Samejima K, Katagiri K, Kawato M (2002) Multiple model-based reinforcement learning. Neural computation 14:1347-1369.

Elliott, R., Frith, C. D., & Dolan, R. J. (1997). Differential neural response to positive and negative feedback in planning and guessing tasks. *Neuropsychologia*, *35*(10), 1395-1404.

Ernst, M. *et al.* Choice selection and reward anticipation: an fMRI study. *Neuropsychologia* **42**, 1585-1597, doi:10.1016/j.neuropsychologia.2004.05.011 (2004).

Kirk, U., Skov, M., Hulme, O., Christensen, M. S., & Zeki, S. (2009). Modulation of aesthetic value by semantic context: An fMRI study. *Neuroimage*, *44*(3), 1125-1132.

Leotti, L. A., & Delgado, M. R. (2011). The inherent reward of choice. *Psychological science*, *22*(10), 1310-1318.

Liljeholm, M., Wang, S., Zhang, J., & O'Doherty, J. P. (2013). Neural correlates of the divergence of instrumental probability distributions. *Journal of Neuroscience*, *33*(30), 12519-12527.

Liljeholm, M., Dunne, S., & O'doherty, J. P. (2015). Differentiating neural systems mediating the acquisition vs. expression of goal-directed and habitual behavioral control. *European Journal of Neuroscience*, *41*(10), 1358-1371.

Liljeholm, M., Mistry, P., & Koh, S. (2018). The Influence of Schizotypal Traits on the Preference for High Instrumental Divergence. *Cog Sci 2018 Proceedings*.

Liljeholm, M. (2018). Instrumental Divergence and Goal-Directed Choice. In *Goal-Directed Decision Making* (pp. 27-48). Academic Press.

McNamee, D., Rangel, A., & O'doherty, J. P. (2013). Category-dependent and category-independent goal-value codes in human ventromedial prefrontal cortex. *Nature neuroscience*, *16*(4), 479.

Mistry, P., & Liljeholm, M. (2016). Instrumental Divergence and the Value of Control. *Scientific reports*, *6*, 36295.

O'Doherty, J. P. Contributions of the ventromedial prefrontal cortex to goal-directed action selection. *Annals of the New York Academy of Sciences* **1239**, 118-129, doi:10.1111/j.1749-6632.2011.06290.x (2011).

O'Doherty, J., Kringelbach, M. L., Rolls, E. T., Hornak, J., & Andrews, C. (2001). Abstract reward and punishment representations in the human orbitofrontal cortex. *Nature neuroscience*, *4*(1), 95.

O'Doherty, J., Winston, J., Critchley, H., Perrett, D., Burt, D. M., & Dolan, R. J. (2003). Beauty in a smile: the role of medial orbitofrontal cortex in facial attractiveness. *Neuropsychologia*, *41*(2), 147-155.

Otto, A. R., Raio, C. M., Chiang, A., Phelps, E. A. & Daw, N. D. Working-memory capacity protects model-based learning from stress. *Proc Natl Acad Sci U S A* **110**, 20941-20946, doi:10.1073/pnas.1312011110 (2013).

Rolls, E. T., Kringelbach, M. L., & De Araujo, I. E. (2003). Different representations of pleasant and unpleasant odours in the human brain. *European Journal of Neuroscience*, *18*(3), 695-703.

Small, D. M., Gregory, M. D., Mak, Y. E., Gitelman, D., Mesulam, M. M., & Parrish, T. (2003). Dissociation of neural representation of intensity and affective valuation in human gustation. *Neuron*, *39*(4), 701-711.

Williams, B. A. Conditioned reinforcement: Neglected or outmoded explanatory construct? *Psychonomic bulletin & review* **1**, 457-475, doi:10.3758/BF03210

Causal intervention strategies change across adolescence

Kate Nussenbaum*, Alexandra O. Cohen*, Zachary J. Davis, David Halpern, Todd M. Gureckis, Catherine A. Hartley
{katenuss, ali.cohen, zach.davis, david.halpern, todd.gureckis, cate} @nyu.edu
Department of Psychology, New York University
New York, NY 10003 USA

Abstract

Intervening on causal systems can illuminate their underlying structures. Past work has shown that, relative to adults, young children often make intervention decisions that confirm single hypotheses rather than those that discriminate alternative hypotheses. Here, we investigated how the ability to make informative intervention decisions changes across development. Ninety participants between the ages of 7 and 25 completed 40 different puzzles in which they had to intervene on various causal systems to determine their underlying structures. We found that the use of discriminatory strategies increased through adolescence and plateaued into adulthood. Our results identify a clear developmental trend in causal reasoning, and highlight the need to expand research on causal learning mechanisms in adolescence.

Keywords: cognitive development; information-seeking; hypothesis testing; causal learning

Introduction

We frequently take actions to manipulate the causal systems that make up our environments. Critically, these causal interventions often vary in the information they reveal (Bramley, Lagnado, & Speekenbrink, 2014; Tong & Koller, 2001; Coenen, Rehder, & Gureckis, 2015).

Imagine, for example, a child tending to a plant. She might believe that the plant requires sunlight, water, and fertilizer to grow. The child might intervene to confirm this hypothesis by placing her plant on a sunny window sill, watering it daily, and fertilizing it. If the plant blooms, she will take this as evidence confirming her initial hypothesis. However, if she were to consider a competing hypothesis – that the plant needs only water and sun but not fertilizer to flourish – she could instead provide the plant with water and sunlight, and critically, withhold fertilizer. If the plant were to wither, she would gain evidence in favor of her first hypothesis, but if it were to grow, she would gain evidence in favor of the second. In this way, different intervention decisions bring about different sets of evidence that help to discriminate competing ideas.

Consistent with this example, previous research has identified two broad classes of decision strategies for making interventions: Confirmatory interventions seek evidence consistent with a particular hypothesis, while discriminatory interventions seek information that can disambiguate competing alternatives (Coenen et al., 2015). It is unclear, however, how causal intervention strategies change with age. Previous work suggests that children as young as 2 years old can derive sophisticated causal knowledge about the structure of their environment by updating their prior assumptions about cause and effect as they encounter new evidence (Gopnik et al., 2004). This evidence is often self-generated – children perform their own "experiments" during play by intervening on causal systems to resolve their uncertainty about how they work (Gopnik, 2012).

Though children are capable of making informative interventions to drive their own learning (Bonawitz, van Schijndel, Friel, & Schulz, 2012; Schulz & Bonawitz, 2007; Sobel & Sommerville, 2010), their information gathering strategies may be sub-optimal. For example, early work in children's hypothesis testing suggests that the ability to systematically test competing alternatives improved from age 5 to age 11, but that even 11-year-olds often failed to make interventions that would enable them to learn underlying causal rules (Rieber, 1969). In a different experiment, when 9- to 11-year-olds were tasked with determining the cause of a specific chemical reaction, the majority of children failed to design systematic experiments that would enable them to efficiently isolate the causal agent (Kuhn & Phelps, 1982).

Characterizing developmental change in causal reasoning

While this work hints that there may be changes in causal intervention strategy across development, no prior work has systematically characterized these changes from childhood to adulthood, perhaps due to the inherent difficulty in measuring developmental change in this complex ability. Multiple strategies can promote effective inference, so studies that have examined only the accuracy of causal judgments, or that have allowed children to freely manipulate causal systems by performing many different actions, may not effectively capture subtle changes in strategy use across development.

A recent study of adults (Coenen et al., 2015) developed a Bayesian measurement model for determining the extent to which confirmatory vs. discriminatory intervention strategies are invoked during decision-making. In this study, adults' intervention decisions were best characterized by a model that combined the discriminatory Expected Information Gain (EIG) strategy with a Positive Testing Strategy (PTS) that assigned "value" to intervention decisions based on the proportion of causal links they would activate. This intervention strategy is generally less cognitively effortful than more discriminatory strategies and can yield informative outcomes in some contexts (Austerweil & Griffiths, 2011), but can also hinder learning by failing to rule out alternative causal models (Nickerson, 1998). Further, adults increased their use of a discriminatory strategy after attempting to solve problems in which confirmatory interventions were systematically less effective, but decreased their discriminatory strategy use under time pressure (Coenen et al., 2015).

The task and modeling approach used by Coenen et al. (2015) has several key properties that make them particularly well-suited to characterize changes in causal intervention strategy across development. First, the task itself is easy to understand but challenging to perform optimally, such that it can be understood by young children while remaining sensitive to changes in causal learning that may occur throughout late childhood, adolescence, and early adulthood. Second, the modeling approach can effectively capture both discriminatory intervention decisions, but also the more cognitively simple, confirmatory strategy that may be adopted by resource-constrained learners. Finally, the model enables estimation of continuous strategy mixture weights for each participant, which can characterize the extent to which their choices reflect confirmatory or discriminatory strategies. By leveraging this measure, we can both account for heterogeneity in strategy use across individuals and examine how strategy use may change across development.

Two previous studies have taken a similar approach but have only examined the choices made by young children, between the ages of 5 and 8 (McCormack, Bramley, Frosch, Patrick, & Lagnado, 2016; Meng, Bramley, & Xu, 2018). In both these studies, rather than selecting interventions that maximized their ability to disambiguate multiple competing possibilities, children often made choices that maximized positive evidence in favor of a *single* hypothesis. However, these studies used only a small number of trials, potentially leading to unreliable estimates of strategy use and preventing the examination of learning over time.

Further, selecting interventions that maximize information gain may require multiple cognitive mechanisms that continue to develop throughout late childhood and adolescence. When faced with intervention decisions, individuals must prospectively imagine the outcomes that different actions are likely to bring about (Sloman & Lagnado, 2015). Then, they must evaluate whether these outcomes provide evidence for one causal hypothesis over another to ultimately choose which action to take (Coenen & Gureckis, 2015). Finally, individuals need to recognize that this cognitive process is "worth it" – that considering possible outcomes of different interventions promotes more accurate hypothesis evaluation relative to other less effortful cognitive strategies. Each of these component mechanisms undergoes marked change throughout development. The ability to use mental models of the environment to prospectively compare decisions (Decker, Otto, Daw, & Hartley, 2016), the ability to infer causal relations based on observed outcomes (Gopnik et al., 2017), and metacognitive sensitivity to the efficacy of different cognitive strategies (Weil et al., 2013) all improve not just in early childhood – a focal point of many studies of causal learning – but continuously across late childhood, adolescence, and early adulthood.

Here, we leveraged the approach introduced by Coenen et al. (2015) – and its key measurement characteristics – to determine the developmental trajectories of causal learning strategies across late childhood, adolescence and early adulthood. Though these developmental periods have been largely neglected in the causal intervention literature, research focused on related cognitive mechanisms suggest these periods may be characterized by robust change in learning and decision-making strategies. Beyond characterizing the general trajectory of change in the use of different intervention strategies, we sought to illuminate interactions between different cognitive mechanisms that may support the emergence of discriminatory hypothesis testing.

Methods

Participants

Ninety 7-to-25-year-olds ($M_{age} = 15.87$ years, $SD_{age} = 5.26$ years, range = 7.04 - 25.74 years, 46 females) participated in the study. All participants completed the matrix-reasoning and vocabulary section of the Wechsler Abbreviated Scale of Intelligence, from which age-normed IQ scores were derived. There was not a significant relation between age and IQ in our sample, $F(1,88) < .001, p > .99, \eta_p < .001$.

Task

Participants completed a computerized task in which they were told they were employees at a computer chip factory, whose job was to sort 3- and 4-node computer chips based on the configuration of their hidden wires. On each trial, participants first viewed two causal graphs for 2 seconds, each of which displayed a different possible configuration of the chip's hidden wires (Figure 1). Then, a computer chip appeared, with all of its nodes turned "off." Participants had as much time as they wanted to make one intervention decision – that is, to click on one node. The node that was clicked *always* turned on. After a brief delay (200 ms) during which the chip turned grey and beeped, the chip reached its final state, indicating outcome of the intervention. The activation of a parent node turned on its direct descendants with a probability of .8. There were no background causes - nodes could only turn on if they were directly clicked or activated by a parent node. After viewing the outcome of their intervention, participants had unlimited time to click on whichever of the two causal graphs they believed indicated the true configuration of the chip's hidden wires. Participants then used a continuous slider to rate their confidence that they selected the correct configuration. Participants were told that they would be paid a bonus based on how many chips they sorted correctly.

Prior to beginning the experimental trials, all participants completed an extensive tutorial in which they were trained on the probabilistic nature of the wires, the directionality of the wires, the correspondence between the causal graph diagrams and the actual chip on which they intervened, and the overall trial procedure.

Participants completed 40 experimental trials. Trial order was pseudo-randomized such that in each block of 10 trials, participants always completed five 3-node puzzles and five 4-node puzzles. The specific puzzles were selected such that the

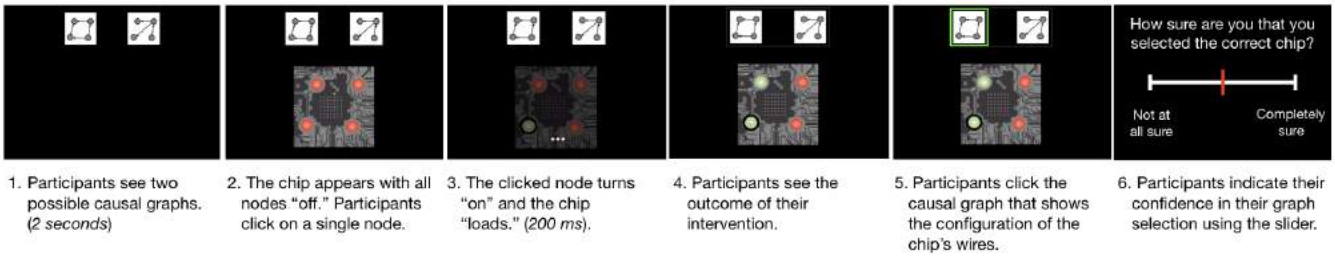


Figure 1: Participants completed 40 intervention trials, in which they had to select a node to determine the configuration of a computer chip's hidden wires.

discriminatory and confirmatory strategy we modeled (more details below) made divergent predictions about the probability of selecting different nodes. The side of the screen on which each graph appeared was randomized. On each trial for each participant, one graph was randomly selected to be the chip's "true" underlying structure. Participants only learned how many chips they sorted correctly at the end of the task; they did not receive trial-by-trial feedback.

Strategies

To model participant intervention choices, we focused on one specific discriminatory intervention strategy - Expected Information Gain - and one specific confirmatory strategy - Positive Testing Strategy. The models differ in how they assign value to possible interventions.

Expected Information Gain (EIG) EIG assumes that individuals have a set of hypotheses about the structure of a particular causal system, with each system represented as a causal Bayesian graph. A learner's uncertainty about which graph (g) is most likely the source of their current observations is represented as the Shannon entropy over the graphs within their hypothesis set (G):

$$H(G) = \sum_{g \in G} P(g) log_2 \frac{1}{P(g)}$$

Learners maximizing information gain should select the information that will cause the largest reduction in their uncertainty. This can be computed by considering the amount of information gained by each possible outcome (o) of each action (a), weighted by their probability:

$$EIG(a) = H(G) - \sum_{o \in O} P(o|a) H(G|a,o)$$

where $H(G|a,o)$ is the new uncertainty after an intervention:

$$H(G|a,o) = \sum_{g \in G} P(g|a,o) log_2 \frac{1}{P(g|a,o)}$$

Positive Testing Strategy (PTS) PTS assumes that participants seek positive evidence to confirm a single hypothesis. We use the formalization introduced in Coenen et al. (2015) which assumes that participants consider each graph in turn, and choose the intervention that will activate the largest proportion of nodes within a single causal graph:

$$PTS(a) = \max_g \left(\frac{DescendantLinks_{n,g}}{TotalLinks_g}\right)$$

Results

Age-related change in strategy use

To characterize participants' intervention choices, we fit a single Bayesian model in which we assumed participants were linearly combining EIG and PTS with weight θ, where $\theta = 0$ indicates a pure PTS strategy and $\theta = 1$ indicates a pure EIG strategy. We further assumed that participants' choices were noisy, such that the expected value of each choice probabilistically influenced intervention decisions. We used a softmax choice function to represent this process, with a free parameter, τ, to capture each participant's decision noise.

The two previous studies using this modeling approach employed a hierarchical model in which group-level hyper-parameters were also estimated (Coenen et al., 2015; Meng et al., 2018), but given our broad age range, we did not want to assume that the participants in our sample comprised a single group. Rather than estimating group-level hyper-parameters, we estimated the model separately for each participant.

We estimated posterior distributions over the parameters using Markov chain Monte Carlo (MCMC) sampling via the NUTS algorithm implemented in STAN (4 chains of 2000 iterations, 1000 per chain discarded as warmup; 4000 total samples per parameter) (Stan Development Team, n.d.; Team, 2013). We used uniform priors over the parameter space ($\tau \sim U(0,\infty); \theta \sim U(0,1)$). Rhat values for all parameter estimates were less than 1.1, indicating convergence across chains (Brooks & Gelman, 1998).

To characterize how strategy use changed with age, we extracted the posterior mean estimates of strategy mixture weights (θ) and examined their relation with age. We tested two linear regression models to examine linear and nonlinear trajectories of developmental change: One included linear z-scored age as a predictor, and one included both linear z-scored age and quadratic z-scored age as predictors (Somerville et al., 2012). We followed this approach for all subsequent models described in the paper.

The model with the quadratic age term provided a significantly better fit to the data, $F(1,87) = 9.95, p = .002$. Both age ($\beta = .12, p < .001$) and age^2 ($\beta = -.06, p = .002$) significantly predicted strategy mixture weight (Figure 3), suggesting that through early adolescence, participants decreased their use of PTS in favor of EIG. Even within age groups,

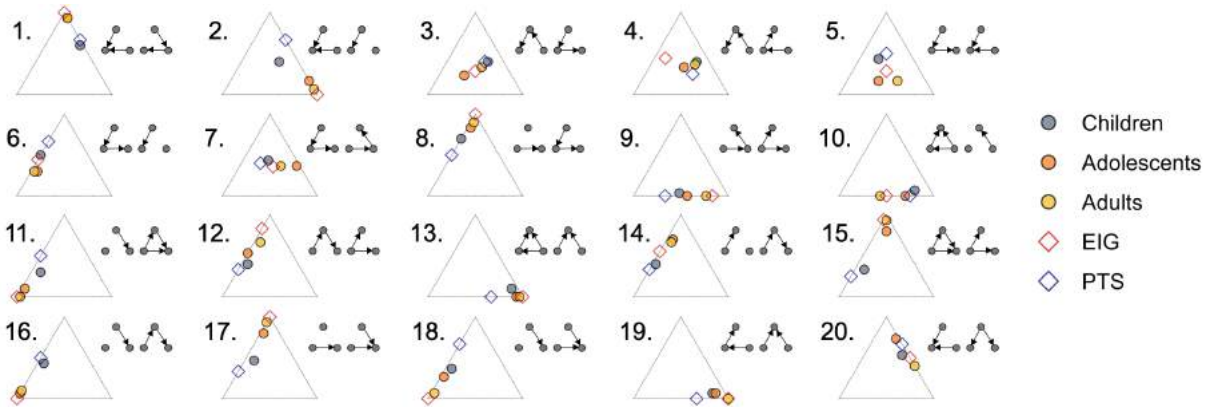


Figure 2: Intervention choices for the 20 three-node puzzles presented in the experiment. The corners of each simplex represent nodes on which participants intervened. The circles represent the average choice for each age group (Children: 7 - 12 years old, Adolescents: 13 - 17, Adults: 18 - 25), while the diamonds represent the "value" of each node as determined by EIG and PTS.

strategy use varied across problems (Figure 2); adolescent choices, for example, sometimes resembled those of adults (16) and sometimes were more like those of children (10).

We also examined how decision noise (τ) changed with age. Decision noise decreased linearly with age ($\beta = -.576, p = .048$), indicating that the choices of older relative to younger participants were more fully captured by the predictions of the two intervention strategies (Figure 3). There was not, however, a significant relation between θ and τ ($p = .271$), suggesting that age-related change in strategy mixture weight can not be attributed to age-related differences in decision noise.

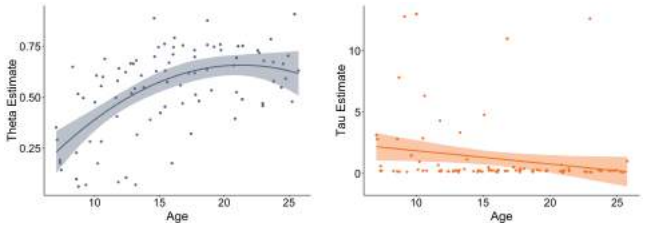


Figure 3: Model-derived estimates of participants' strategy mixture weights (θ) show that participants became more discriminatory with increasing age through late adolescence. Decision noise estimates (τ) show that intervention decisions became more value-based with increasing age. Best-fitting regression lines illustrating the effects of age and age^2 on θ and age on τ are plotted.

In line with previous findings (Meng et al., 2018; Coenen et al., 2015), our modeling results suggest that children and adults use a combination of confirmatory and discriminatory strategies to test causal hypotheses. Further, they demonstrate that this combination systematically differs across children, adolescents, and adults.

Inference-intervention interactions

Why did the use of a discriminatory intervention strategy increase across development? One possibility is that when presented with the novel task, participants explored different intervention strategies until finding one they believed was most effective. Older participants may have been more sensitive to the relative efficacy of different intervention strategies. For EIG to be a useful strategy, however, individuals needed to be able to make accurate causal inferences based on the outcomes of their interventions. Gaining information to disambiguate competing hypotheses was only useful if individuals could correctly updated their beliefs based on that new evidence (Coenen & Gureckis, 2015).

To examine whether causal inference changed with age, we computed the posterior probabilities of each of the two possible causal graphs based on the selected node and the final states of the other nodes on each trial. We then ran a linear mixed-effects model to determine whether there was a relation between age and the posterior probability of the structure selected. Older participants selected more probable causal structures, $F(1,88) = 10.44, p = .002$. This suggests that with increasing age, individuals became better at evaluating the outcomes of their interventions to disambiguate competing hypotheses. However, this metric is inherently confounded with intervention decisions – by definition, interventions with higher EIG scores were more likely to lead to greater increases in the posterior probability of one structure over another. Thus, it is difficult to determine the direction of the relationship between causal intervention and inference – were older participants selecting more informative interventions because they could more effectively prospectively evaluate how that information would enable them to update their beliefs? Or were they updating their beliefs more effectively because they chose interventions that provided stronger evidence in favor of one hypothesis over another?

Participant confidence in the structure they selected can provide insight into developmental change in causal infer-

ence – and metacognitive sensitivity to causal evidence – without being confounded by intervention choice. If participants were sensitive to the extent to which the information they gained allowed resolution of competing hypotheses, then their confidence in the structures they selected should track their posterior probabilities. To determine how these posterior probabilities and age influenced confidence ratings, we ran a linear mixed-effects model. Our best-fitting model included both a linear and quadratic effect of age. Participants were more confident in their selection when the posterior probability of the structure they selected was higher, $F(1,3535.17) = 353.69, p < .001$ However, this effect was qualified by an age x posterior probability interaction ($F(1,3529.67) = 21.75, p < .001$) as well as by an age^2 x posterior probability interaction ($F(1,3529.76) = 12.83, p < .001$), such that the influence of posterior probabilities on confidence ratings increased throughout childhood and early adolescence. These results indicate that the ability to evaluate the extent to which new information supported causal hypotheses improved non-linearly across development. Importantly, they suggest developmental improvements in causal inference that are separable from improvements in intervention strategy.

We next examined whether developmental change in causal inference influenced intervention strategy. Specifically, we computed the correlation between the posterior probability of the structure selected and confidence ratings for each participant and ran a linear regression to determine whether these values, which we will refer to as "evidence sensitivity," predicted strategy mixture weight (θ). We found a positive relationship between evidence sensitivity and θ ($\beta = .09, p < .001$), even when controlling for age and age^2. In other words, participants with stronger sensitivity to the strength of the evidence on which to base their inferences also demonstrated greater use of EIG.

Within-task learning effects

Beyond examining how causal intervention strategy changed with age, our use of 40 trials enabled us to examine learning over the course of the task. We hypothesized that older participants' greater use of a discriminatory strategy might in part be driven by faster learning, such that age would more strongly influence estimated values of θ in the second half of the experiment, after participants had the opportunity to learn to adjust their strategy based on their evaluations of their earlier decisions.

To examine whether participants used a different mixture of strategies throughout the course of the task, we fit our Bayesian model separately to the first and second half of trial data for each participant. We then ran a linear mixed-effects model to determine how experiment half and age influenced strategy mixture weight. As before, both linear and quadratic age predicted strategy mixture weight ($ps < .02$). Furthermore, strategy mixture weight increased from the first half to the second half of the experiment, $F(1,87) = 11.4, p < .001$ (Figure 4), indicating that participants may have learned to use a more discriminatory strategy over the course of the task. Contrary to our prediction, however, experiment half did not interact with age or age^2 ($ps > .20$).

Decision noise also decreased over the course of the experiment, $F(1,88) = 5.18, p = .03$. This effect was qualified by an age x experiment half interaction, such that younger participants demonstrated a greater decrease in decision noise from the first to the second half of trials, $F(1,88) = 4.72, p = .03$ (Figure 4). This suggests that younger participants may have learned to use their estimates of the value of each intervention to more strongly guide their decisions over the course of the task. While the change in their strategy mixture weight did not statistically differ from that of older participants, younger participants may have learned that *both* strategies were more effective than randomly selecting nodes.

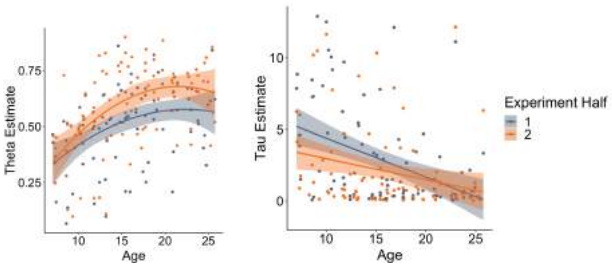


Figure 4: In the second half of the experiment, participants relied more on EIG over PTS, and their choices were less noisy.

Finally, we examined whether evidence sensitivity related to participants' change in strategy use over the course of the task. We computed $\Delta\theta$ for each participant by subtracting their estimated θ value over the first half of the experiment from their estimated θ value over the second half of the experiment. We then ran a regression examining the effects of age and evidence sensitivity on $\Delta\theta$. We found a significant effect of evidence sensitivity on $\Delta\theta$, $\beta = .049, p = .019$, such that participants who were most sensitive to their ability to correctly identify underlying causal structures demonstrated increased use of EIG over the course of the experiment. Mirroring our previously reported results, there was not a significant effect of age on $\Delta\theta$, nor was there an age x evidence sensitivity interaction effect ($ps > .61$).

Discussion

Our results and modeling analyses demonstrate robust changes in causal intervention strategy from middle childhood to adulthood. In sum, interventions become more discriminatory with increasing age until reaching a plateau in late adolescence. What causes this developmental shift?

One possibility is that improvements in intervention strategy are due to increased exposure to scientific reasoning strategies through formal schooling. Future work could test participants at multiple time-points and examine the extent to which increases in EIG use align with exposure to curricular units focused on concepts like controlling variables to

effectively discriminate hypotheses (Kuhn, Arvidsson, Lesperance, & Corprew, 2017).

However, several aspects of our data suggest that formal schooling can not account for all age-related change in strategy use that we observed. First, almost all participants demonstrated a mixture of strategies throughout the experiment, and this mixture appears to change *gradually* with increasing age (as opposed to a sharp shift corresponding to the introduction of specific concepts during formal schooling). We also found that individual and developmental differences in more basic learning mechanisms, like sensitivity to the informativeness of intervention outcomes, predicted strategy use. Additionally, individuals across our age range increased their use of a discriminatory strategy throughout the course of the task, without any explicit instruction or feedback.

It may also be the case that with increasing age, individuals become better at prospectively planning their intervention decisions. Though evidence sensitivity correlated with strategy mixture weight in our data, it did not fully account for developmental change in strategy use. Importantly, we hypothesized that the ability to make accurate causal judgments may enable individuals to select the best intervention only if they prospectively simulate and sample the outcomes of potential choices in the first place (Bonawitz, Denison, Griffiths, & Gopnik, 2014). On some trials, participants may not have attempted to think through the possible outcomes of their decisions, in which case the ability to evaluate those outcomes would not affect the intervention choice. Future studies should probe the role of other cognitive mechanisms in supporting the use of EIG, like model-based decision-making, which may support or similarly rely on simulating probabilistic outcomes of multi-stage decisions (Decker et al., 2016; Doll, Duncan, Simon, Shohamy, & Daw, 2015).

Another possibility is that younger people are equally *capable* of implementing a more discriminatory intervention strategy, but perform a different cost-benefit analysis when determining which strategy to use. As mentioned previously, the confirmatory PTS strategy often reveals diagnostic information in environments in which causal links are sparse or deterministic (Austerweil & Griffiths, 2011). Additionally, confirmatory hypothesis testing may be adaptive when individuals have the opportunity to make multiple interventions at low cost. It may be the case that rather than spending time and cognitive effort to make the single best intervention, children prefer to make multiple, easier, intervention decisions, which together provide the information they need. Future studies could isolate changes in *ability* from changes in effort allocation, by raising the cost of making an uninformative intervention or forcing all participants to spend a long time deliberating prior to allowing them to perform their intervention.

Finally, though few studies have examined causal learning in adolescence, our results demonstrate that causal learning and decision-making continue to change during this period. Future work probing the cognitive mechanisms that drive these changes will inform how to best support adolescents as they interact with their environments with increasing independence and shape their own learning opportunities.

Acknowledgments

We thank Morgan Glover, Sree Panuganti, Dhiraj Patel, Haniyyah Sardar, Xinxu Shen, and Daphne Valencia for help with data collection. We also thank the Jacobs Foundation (Early Career Fellowship to C.A.H.), the Department of Defense (NDSEG Fellowship to K.N.), the National Science Foundation (Grant No. 1714321 to A.O.C, CAREER grant 1654393 to C.A.H., and CAREER grant BCS-1255538 to T.M.G.), and the James S. McDonnell Foundation (Scholar Award to T.M.G.) for financial support.

References

Austerweil, J. L., & Griffiths, T. L. (2011). Seeking Confirmation Is Rational for Deterministic Hypotheses. *Cognitive Science*, *35*(3), 499–526.

Bonawitz, E., Denison, S., Griffiths, T. L., & Gopnik, A. (2014). Probabilistic models, learning algorithms, and response variability: sampling in cognitive development. *Trends in Cognitive Sciences*, *18*(10), 497–500.

Bonawitz, E., van Schijndel, T. J. P., Friel, D., & Schulz, L. (2012). Children balance theories and evidence in exploration, explanation, and learning. *Cognitive Psychology*, *64*(4), 215–234.

Bramley, N. R., Lagnado, D. A., & Speekenbrink, M. (2014). Conservative forgetful scholars: How people learn causal structure through sequences of interventions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(3), 708–731.

Brooks, S. P., & Gelman, A. (1998). General Methods for Monitoring Convergence of Iterative Simulations. *Journal of Computational and Graphical Statistics*, *7*, 434. doi: 10.2307/1390675

Coenen, A., & Gureckis, T. M. (2015). Are biases when making causal interventions related to biases in belief updating? In *Proceedings of the 37th annual conference of the cognitive science society.*

Coenen, A., Rehder, B., & Gureckis, T. M. (2015). Strategies to intervene on causal systems are adaptively selected. *Cognitive Psychology*, *79*, 102–133.

Decker, J. H., Otto, A. R., Daw, N. D., & Hartley, C. A. (2016). From Creatures of Habit to Goal-Directed Learners. *Psychological Science*, *27*(6), 848–858.

Doll, B. B., Duncan, K. D., Simon, D. A., Shohamy, D., & Daw, N. D. (2015). Model-based choices involve prospective neural activity. *Nature Neuroscience*, *18*, 767–772.

Gopnik, A. (2012). Scientific Thinking in Young Children: Theoretical Advances, Empirical Research, and Policy Implications. *Science*, *337*, 1623–1627.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A Theory of Causal Learning in Children: Causal Maps and Bayes Nets. *Psychological Review*, *111*(1), 3–32.

Gopnik, A., O'Grady, S., Lucas, C. G., Griffiths, T. L., Wente, A., Bridgers, S., ... Dahl, R. E. (2017). Changes in cognitive flexibility and hypothesis search across human life history from childhood to adolescence to adulthood. *Proceedings of the National Academy of Sciences*, *114*(30), 7892–7899.

Kuhn, D., Arvidsson, T. S., Lesperance, R., & Corprew, R. (2017). Can Engaging in Science Practices Promote Deep Understanding of Them? *Science Education*, *101*, 232–250. doi: 10.1002/sce.21263

Kuhn, D., & Phelps, E. (1982). Advances in Child Development and Behavior. , *17*, 1–44.

McCormack, T., Bramley, N., Frosch, C., Patrick, F., & Lagnado, D. (2016). Children's use of interventions to learn causal structure. *Journal of Experimental Child Psychology*, *141*, 1–22.

Meng, Y., Bramley, N., & Xu, F. (2018). Children's Causal Interventions Combine Discrimination and Confirmation. In *40th annual meeting of the cognitive science society* (pp. 281–286).
Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175–220.
Rieber, M. (1969). Hypothesis testing in children as a function of age. *Developmental Psychology*, *1*, 389.
Schulz, L. E., & Bonawitz, E. B. (2007). Serious fun: preschoolers engage in more exploratory play when evidence is confounded. *Developmental Psychology*, *43*(4), 1045–1050.
Sloman, S. A., & Lagnado, D. (2015). Causality in thought. *Annual Review of Psychology*, *66*(1), 223-247.
Sobel, D. M., & Sommerville, J. A. (2010). The Importance of Discovery in Children's Causal Learning from Interventions. *Frontiers in Psychology*, *1*, 1–7.
Somerville, L. H., Jones, R. M., Ruberry, E. J., Dyke, J. P., Glover, G., & Casey, B. J. (2012). The Medial Prefrontal Cortex and the Emergence of Self-Conscious Emotion in Adolescence. *Psychological Science*, *24*, 1554–1562.
Stan Development Team. (n.d.). RStan: the R interface to Stan.
Team, R. C. (2013). R: A language and environment for statistical computing.
Tong, S., & Koller, D. (2001). Active learning for structure in bayesian networks. In *Proceedings of the 17th international joint conference on artificial intelligence.*
Weil, L. G., Fleming, S. M., Dumontheil, I., Kilford, E. J., Weil, R. S., Rees, G., ... Blakemore, S.-J. (2013). The development of metacognitive ability in adolescence. *Consciousness and Cognition*, *22*(1), 264–271.

Thinking counterfactually supports children's ability to conduct a controlled test of a hypothesis

Angela Nyhout (angela.nyhout@utoronto.ca)
Department of Applied Psychology & Human Development,
University of Toronto, 252 Bloor St West, Toronto, Ontario, M5S 1V6

Alana Iannuzziello (alana.iannuzziello@mail.utoronto.ca)
Department of Applied Psychology & Human Development,
University of Toronto, 252 Bloor St West, Toronto, Ontario, M5S 1V6

Caren M. Walker (carenwalker@ucsd.edu)
Department of Psychology, University of California San Diego,
9500 Gilman Drive, La Jolla, CA 92093 USA

Patricia A. Ganea (patricia.ganea@utoronto.ca)
Department of Applied Psychology & Human Development,
University of Toronto, 252 Bloor St West, Toronto, Ontario, M5S 1V6

Abstract

Children often fail to control variables when conducting tests of hypotheses, yielding confounded evidence. We propose that getting children to think of alternative possibilities through counterfactual prompts may scaffold their ability to control variables, by engaging them in an imagined intervention that is structurally similar to controlled actions in scientific experiments. Findings provide preliminary support for this hypothesis. Seven- to 10-year-olds who were prompted to think counterfactually showed better performance on post-test control of variables tasks than children who were given control prompts. These results inform debates about the contribution of counterfactual reasoning to scientific reasoning, and suggest that counterfactual prompts may be useful in science learning contexts.

Keywords: cognitive development; scientific reasoning; counterfactual reasoning; causal learning; science education

Scientific Reasoning in Development

Equipping children with scientific inquiry skills is a core objective of elementary science education, allowing children to collect evidence and draw inferences about the world around them. However, extensive research has found that children are relatively unequipped to engage in many aspects of scientific inquiry in the absence of direct instruction and frequent scaffolding (Klahr, Fay, & Dunbar, 1993; Klahr & Nigam, 2004; Kuhn & Franklin, 2006; Schauble, 1996). In the present study, drawing from research and theory in cognitive development, science education, and philosophy, we investigate the use of a novel pedagogical tool – counterfactual reasoning prompts – to scaffold children's scientific reasoning skills.

An important sub-skill of scientific inquiry is the ability to control variables. This skill, termed the control-of-variables strategy (CVS) has received a great deal of attention in research on scientific reasoning over the past four decades (for a review, see Zimmerman, 2007). To properly execute this skill, the learner should isolate a single variable at a time, while holding all else constant.

Consider a common task used in studies investigating CVS (e.g., Chen & Klahr, 1999; Klahr & Nigam, 2004). Children are presented with a set of ramps that can be varied along a number of dimensions (e.g., ramp height, surface, run length, ball size) and their task is to manipulate the ramps to determine the effect of different variables on where a ball stops after rolling down the ramp. To make warranted inferences about individual variables, the learner should change the values of a single variable (e.g., compare a high ramp to a low ramp), keeping all other variables constant (e.g., smooth surface, same-size balls).

Although children are able to *recognize* a conclusive test of a hypothesis as young as age 6 (Sodian, Zaitchik, & Carey, 1991), they typically fail to *produce* one themselves in the absence of scaffolding through middle childhood (Klahr, Zimmerman, & Jirout, 2011; Zimmerman, 2007). However, with direct instruction, children often show improvement in their ability to design controlled experiments (Chen & Klahr, 1999; Klahr & Nigam, 2004; for a meta-analysis, see Schwichow, Croker, Zimmerman, Hoffler, & Hartig, 2016). For instance, Chen and Klahr (1999) found that 7- to 10-years-old who were given explicit instruction on CVS were better able to transfer this strategy to both similar and dissimilar problems than those who engaged in self-guided inquiry. Younger children frequently failed to design unconfounded tests.

Although past studies have found that children are able to learn the control of variables strategy through direct instruction or demonstrations, science curricula and educational guidelines often recommend teaching scientific inquiry skills through *inquiry-based learning* instead (e.g., US National Research Council, 2000). That is, children's scientific inquiry skills are thought to be best supported by having children explore science concepts based on their own

observations and experiences with phenomena of interest, with little explicit instruction from educators. Thus, there is significant educational value to identifying methods for scaffolding children's hypothesis testing abilities that not only fit within these curricular guidelines, but also harness children's intuitive reasoning skills.

Causal and Counterfactual Reasoning

Whereas the work reviewed above suggests that older children are poor at testing and revising hypotheses, another body of research shows that children are adept at parallel skills when engaging in causal learning tasks.

From a young age, children form, test, and revise hypotheses in building informal theories in various domains (Carey, 1985; Gopnik, Meltzoff, & Bryant, 1997; Keil, 1992;). For instance, toddlers are able to infer higher-order relational causes (Walker & Gopnik, 2014). Preschoolers are able to draw appropriate causal inferences from patterns of dependence, even when evidence conflicts with their prior knowledge (Schulz & Gopnik, 2004), and use evidence from interventions to make inferences about causal structure (Schulz, Glymour, & Gopnik, 2007).

Why do older children (and even adults) fail when applying this skill-set in scientific reasoning contexts? We suggest a few possible explanations for this discrepancy. First, studies of intuitive causal reasoning with toddlers and preschoolers use tasks that are typically decontextualized, placing relatively few demands on children's prior knowledge. Many of these studies rely on a "blicket detector" paradigm, in which children are familiarized with a novel machine, and their task is to determine what makes it switch on (Gopnik & Sobel, 2000). In contrast, scientific reasoning tasks given to older children typically use knowledge-laden tasks that rely heavily on children's existing (and often incorrect) knowledge and theories (e.g., Chen & Klahr, 1999). Second, causal reasoning tasks typically measure children's abilities implicitly, whereas scientific reasoning tasks ask children to explicitly plan and often verbally demonstrate their abilities. Despite these differences, both classes of studies rely on a common set of domain-general inferential skills, including the ability to form and revise hypotheses on the basis of available evidence.

How do we connect the parallel mechanisms children successfully apply in causal reasoning tasks to scientific reasoning contexts? In the current study, we explore the claim that counterfactual reasoning is fundamental to causal and scientific reasoning, and suggest that counterfactual prompts may help to connect these abilities. When we think counterfactually, we compare the way things are to the way things *could have been*. Counterfactual reasoning therefore necessarily involves thinking about causes: As one considers how an event could have turned out differently, one reasons about the causal relationship between an antecedent and outcome. If the event X had not happened, would event Y still have happened? If the answer to this is "no", one can conclude that event X is a cause of event Y (Lewis, 1986).

However, the utility of counterfactual reasoning may not be limited to drawing *specific* causal inferences. Several researchers have drawn theoretical parallels between the *mechanisms* underlying counterfactual reasoning and scientific reasoning (e.g., Buchsbaum, Bridgers, Weisberg, & Gopnik, 2012; Erb & Sobel, 2014; Gopnik & Walker, 2013; Sloman, 2005; Rafetseder & Perner, 2014; Walker & Gopnik, 2013). If a learner believes that X caused Y, they can mentally intervene on X by imagining that it did not occur, follow the causal implications of this change, and then reason about whether it would have led to a change in Y (Gopnik & Walker, 2013; Walker & Gopnik, 2013). We follow an identical process in *scientific* reasoning. We hypothesize that X causes Y, and then make plans to systematically manipulate X in order to investigate its impact on Y. In both counterfactual and scientific reasoning, the learner adjusts a causal system by (mentally or physically) intervening on one event and considering the effects of this change.

Despite the proposed contribution of counterfactual reasoning to science learning, there is relatively little research connecting the two (Engle & Walker, 2018; Frosch, McCormack, Lagnado, & Burns, 2012; Schulz, et al., 2007) and no work linking these capacities to hypothesis testing in children. Only two previous studies to our knowledge have investigated the relationship between counterfactual reasoning and scientific inquiry. Adults primed with counterfactuals were better able to conduct a disconfirming test of a hypothesis than those given neutral primes (Galinsky & Moskowitz, 2000). In another study, counterfactual prompts scaffolded children's ability to detect anomalies to an existing hypothesis in a causal learning task (Engle and Walker, 2018).

Given that counterfactual and scientific reasoning both involve intervening on a single variable to investigate its causal role in an outcome of interest, we propose that engaging children in counterfactual reasoning during a control-of-variables task will scaffold their ability to conduct a controlled test of a hypothesis by activating a parallel underlying cognitive mechanism.

That said, it is worth first considering whether children of the age we tested in the current study (7 to 10 years) are capable of counterfactual reasoning, given the lively debate about its developmental trajectory. Previous research has been mixed, with some findings indicating that children *can* reason counterfactually as young as 3-½ years (Harris, German, & Mills, 1996), and other work suggesting that this ability does not reach maturity until adolescence (e.g., Rafetseder, Schwitalla, & Perner, 2013). However, more recent work suggests that studies showing counterfactual reasoning to be late-developing may have underestimated children's ability by presenting opaque causal structures and by placing large demands on children's memory (McCormack, Ho, Gribben, O'Connor, & Hoerl (2018; Nyhout, Henke, & Ganea, 2019). A recent set of studies demonstrates that children reason counterfactually by age 4 when given a clear and novel causal structure that does not

rely on their background knowledge (Nyhout & Ganea, 2019). Thus, we conclude from these findings that children have the requisite abilities to engage in counterfactual reasoning well before the age of those in the current study.

Current Study

In contrast to previous research (Chen & Klahr, 1999; Klahr & Nigam, 2004), we investigated whether children's ability to control variables could be scaffolded in non-school settings. We also reduced task demands by including a smaller number of variables (2 variables, rather than 4).

Children in the present study were assigned to either a counterfactual or control condition. After watching a video of an actor conducting a controlled test of a hypothesis, children were given either a *counterfactual* prompt, asking them to consider what would happen if the actor had conducted her test differently, or a *control* prompt, in which children were asked to recall what had happened. We predicted that children given counterfactual prompts would be more likely to improve from pre-test to post-test than children given control prompts. We tested a range of ages typically used in CVS research (7 to 10 years), but did not have prior predictions about age-related differences in performance.

Method

Participants

Participants aged 7 to 10 years of age were recruited and tested at a museum in a large urban area. The final sample included 88 children (M = 8.91, SD = 1.13, range = 7.00 to 10.97, 45 girls) whose data are reported below. Participants were placed in two categories, based on their age. The *younger* age category included children between the ages of 7.00 and 8.99 (n = 46, M = 8.00, SD = 0.63) and the *older* age category included children between the ages of 9.00 to 10.99 (n = 42, M = 9.90, SD = 0.59), with categories selected on the basis of similar previous studies (e.g., Chen & Klahr, 1999; Klahr & Nigam, 2004). Participants who passed the pre-test phase (n = 24) were excluded as they were determined to be already competent with CVS. Three additional participants were excluded due to experimenter error (n = 2) or language barriers (n = 1).

Materials

For the pre- and post-test phases described below, participants were given two identical ramps with both a down- and up-ramp side. The ramps were ridged on the up-ramp where the ball could stop (Figure 1). Each ridge was painted a different color to allow for unambiguous reference and measurement. There were four binary variables, but participants received only two of the four variables at a time, and the remaining two variables were "fixed". The variables were paired as follows: (1) height (high or low) and ball size (large or small), or (2) starting place (top or middle), and surface (rough or smooth). For instance, at one time-point, participants were given a large and small ball for each ramp, and pieces to adjust the steepness of each ramp ("high" or "low"). At the other time-point, participants were given a rough surface and smooth surface for each ramp, and a piece of cardboard to adjust where the ball started for each.

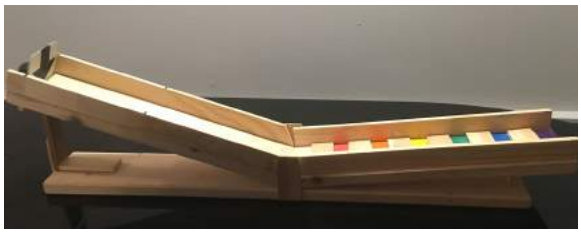


Figure 1: One of two identical ramps used in the study. A ball is launched from the down-ramp (left) and stops on one of the coloured ridges on the up-ramp (right). The apparatus can be adjusted for height, surface type, where the ball starts on the down-ramp, and ball size.

The same set of ramps were used in a video in the scaffolding phase, displayed for participants on a laptop.

Procedure

The study included a warm-up activity (uncertainty training) followed by pre-test, scaffolding, and two post-test phases. Participants were assigned to one of two conditions for the scaffolding phase: counterfactual (n = 45, M age = 8.47, 23 girls) or control (n = 43, M age = 8.44, 22 female girls). The order of all variables and variable sets were counterbalanced between participants.

Uncertainty Training. Given that some of the prompts in the intervention phase required children to acknowledge their uncertainty about an outcome, we included an uncertainty training phase to ensure children were able to recognize and acknowledge their uncertainty. All children, regardless of condition, received the same uncertainty training. Using cards with various colors and suits, the experimenter placed a pair of cards face down, and turned over one card. Before revealing the second card, she asked the participant if they could be "sure or not sure" if the face-down card was the same or different as the face-up card. Regardless of the participant's response, the experimenter instructed children that they cannot be sure if the two cards are the same, and that it is okay to answer the question in this way. The process repeated until the participants answered that they could not be sure three times.

Pre-test. The experimenter placed the two ramps next to each other, directly in front of the participant, and explained that the two ramps were similar and worked the same way. She then showed participants how to operate and adjust the ramps along two of the variables (e.g. height of ramp and size of ball). The other two variables (e.g. ramp surface and run length) were fixed and not introduced until the *post-test transfer phase*. Participants were asked to demonstrate how to manipulate the ramps. If they did not set up the ramp correctly, the experimenter showed them again. All

demonstrations were performed with one ramp, and participants were reminded that the ramps were the same.

To measure children's ability to execute CVS, the experimenter asked them to show how they would find out if one variable plays a role in how far the ball travels down the ramp (e.g., "Can you show me how you would find out if *the size of ball* matters for how far the ball goes down the ramp?"). She told participants they had one chance to set up both ramps at the same time, and then repeated the question a second time. Participants were required to set up both ramps before launching the balls down each ramp one at a time. After each ball was launched, the experimenter labeled the outcome (e.g., "Look! The ball stopped on the yellow line.") but did not compare between the two ramps.

Using the same procedure, the experimenter then asked participants to determine if a second variable mattered for how far the ball would travel down a ramp (e.g., "Can you show me how you would find out if *the height of the ramp* matters for how far the ball goes down the ramp?").

Participants who controlled the correct variable received 1 point for each question for a maximum score of 2. Participants who received a score of 2/2 at pre-test were excluded from the study (and the study was terminated at this point), because they already possessed an understanding of CVS ($n = 24$). Participants who received scores of 0 or 1 went on to the scaffolding phase.

Scaffolding. Participants in this phase watched two videos of an actor exploring the ramps and were told that they would be asked about what they saw after each video. The actor in the videos manipulated the same two variables that participants were asked to isolate during the pre-test, using the same ramps. The video started with the actor stating that she was going to find out if a variable (e.g., height of the ramp) played a role in how far the ball travelled down the ramps. The actor then proceeded to set-up the ramps and labelled the set-up as she went along (e.g., "I'm going to set Ramp 1 to high"). After she set-up both ramps, she launched the balls one at a time and labeled the outcome by stating the color the ball landed on. At the end of the video she stated which ball (on Ramp 1 or Ramp 2) travelled farther. The experimenter then paused the video so that the participants could see the outcome of both ramps at the same time. The videos were identical across conditions; the only difference was in the question prompts children were asked after.

In the *counterfactual condition*, participants were asked to imagine a change to the value of a variable (e.g., "Let's imagine that she set Ramp 1 to low. Would the ball have travelled down the ramp farther on Ramp 1, farther on Ramp 2, or you can't be sure?"). This imagined change would create a confounded (or uncontrolled) test. In the *control condition*, participants were asked to recall what had happened (e.g., "Let's imagine again what happened to the ball on Ramp 1? Did the ball travel farther on Ramp 1, farther on Ramp 2, or you can't be sure?"). Children did not receive feedback on their responses during the scaffolding phase in either condition.

In both conditions, a second video was shown highlighting the other variable (e.g., size of ball). In the counterfactual condition, the experimenter asked the participants to imagine a change to the value of this new variable (e.g., size of ball), creating another confounded test. In the control condition, the experimenter asked the participants the same question as before, but highlighted the other ramp (e.g., Ramp 2).

Post-Test Same. The experimenter removed the laptop and placed the ramps side-by-side in front of the participant. This phase was identical to the pre-test, except participants were not asked to demonstrate how the ramps worked. Responses were coded in the same way, with participants receiving a maximum score of 2.

Post-Test Transfer. The experimenter then told participants that the ramps can work in a different way. The two original variables were fixed (e.g., ramps could only be set to high, and only the big balls could be used) and two new variables were introduced (e.g., surface of the ramp and starting position for the ball). As in the pre-test, the experimenter showed participants how the new variables worked on the ramps and asked participants to demonstrate how to manipulate each new variable.

The procedure was the same as the pre-test and post-test same phases except participants were asked two new questions about each of the new variables (e.g. "Can you show me how you would find out if *the surface of the ramp*/*where the ball starts on the ramp* matters for how far the ball goes down the ramp?"). Again, participants could receive a score up to 2 across the two test questions.

In sum, participants were asked two questions each at pre-test, post-test same, and post-test transfer, and received a score between 0 and 2 for the number of controlled tests they conducted in each phase. In each counterbalancing order, the pre-test and post-test same phases were identical, whereas the post-test transfer phase used two previously unencountered variables. The experimenter live-recorded with paper-and-pencil, and later checked videos for accuracy. A second researcher coded 34% of videos, and inter-rater reliability was excellent (96.6% agreement, Fleiss' $\kappa = 0.93$, $p < .001$).

Results

We first tested whether there were differences between the two conditions at pre-test using a Chi-Square test of independence, and found no significant differences across conditions in pre-test score, $p = .206$. We also found no significant differences between genders ($U = 926$, $p = .687$) or the variable set participants received at pre-test ($U = 902$, $p = .522$), thus we do not consider these variables further.

To investigate the change in children's score (CVS score out of 2) from pre-test to *(1) post-test same* and *(2) post-test transfer*, we conducted two generalized estimating equation (GEE) analyses with multinomial distributions and cumulative logit-log link functions with condition (counterfactual or control) and age group (younger or older) as predictor variables.

For the GEE of pre-test vs. *post-test same* performance, there was a main effect of test, B = -2.56, SE = 0.56, Wald $\chi2(1) = 20.84$, $p < .001$, such that children were 12.82 times more likely to receive a higher score at *post-test same* than at pre-test, Exp(B) = 0.78, 95% CI = [0.03, 0.23]. There was also a main effect of age, B = -1.66, SE = 0.76, Wald $\chi2(1) = 4.78.$, $p = .029$, such that older children were 5.26 times more likely to receive a higher score than younger children, Exp(B) = 0.19 95% CI = [0.04, 0.84]. The main effect of condition was not significant, $p = .436$. The test phase by age category interaction was significant, B = 1.59, SE = 0.72, Wald $\chi2(1) = 4.85.$, $p = .028$, such that older children in the post-test same phase were 4.90 times more likely to receive a higher score than younger children in the post-test same phase, Exp(B) = 4.90, 95% CI = 1.19, 20.13] All other interactions were non-significant.

For the GEE of pre-test vs. *post-test transfer*, there was again a main effect of test, B = -1.59, SE = 0.51, Wald $\chi2(1) = 9.70$, $p = .002$, such that children were 4.90 times more likely to receive a higher score on the *post-test transfer* phase than the pre-test phase, Exp(B) = .204, 95% CI [0.08, 0.55]. There was also a main effect of age, B= -1.35, SE = 0.68, Wald $\chi2(1) = 3.92$, $p = .048$, such that older children were 3.85 times more likely to receive a higher score than younger children, Exp(B) = 0.26, 95% CI = [0.069, 0.987]. The main effect of condition was marginally significant, B = 1.31, SE = 0.67, Wald $\chi2(1) = 3.819$, $p = .051$. Children in the counterfactual condition were 3.71 times more likely to receive a higher score than those in the control condition, Exp(B) = 3.71, 95% CI = [1.00, 13.78]. All interactions were non-significant.

We conducted planned post-hoc comparisons to further investigate performance between groups at each test-phase using Chi-square tests of independence. Performance differed significantly between children in the counterfactual and control conditions at both *post-test same* $\chi^2(2) = 7.28$, $p = .026$ and *post-test transfer* $\chi^2(2) = 6.04$, $p = .049$. Table 1 presents the relevant proportions of children who conducted 0, 1, and 2 controlled tests in each test phase entered into the Chi-square analyses.

Table 1: Proportion of children who conducted 0, 1, or 2 controlled tests in each post-test phase (CVS Score).

		CVS Score (/2)		
Post-test	Condition	*0*	*1*	*2*
Same	Counterfactual	11.1	33.3	55.6
	Control	34.9	20.9	44.2
Transfer	Counterfactual	24.4	22.2	53.3
	Control	41.9	30.2	27.9

Finally, we considered the relation between children's responses to counterfactual prompts in the scaffolding phase and their CVS scores, although we did not make predictions about any such relation. Recall that the correct answer to the counterfactual prompts was "can't be sure", because the counterfactual intervention created a confounded test. Of the 45 children in the counterfactual condition, 13 (29%) answered "can't be sure" to both prompts, 19 (42%) answered "can't be sure" to 1/2, and 13 (29%) did not answer "can't be sure" to either prompt. Children's "can't be sure" responses did not significantly correlate with their performance on any of the CVS tests, Spearman's rho = -.121 to -.231, $p = .127$ to .430.

Discussion

We proposed that prompting children to think counterfactually during a control-of-variables task would scaffold their performance by capitalizing on their underlying causal reasoning skills. The results of this study provide initial support for this proposal. Children given counterfactual prompts showed better performance on the post-test phases than those given control prompts, though these differences were non-significant on post-test same and marginally significant on post-test transfer in the omnibus analyses. Critically, when considering condition differences alone, children in the counterfactual condition performed significantly better than those in the control condition at both post-tests. The largest proportion of control group children scored 0/2 on both post-tests, whereas the largest proportion of counterfactual group children scored 2/2, as displayed in Table 1.

Along with these condition differences, there was also an indication that the video demonstration alone improved children's ability to control variables, given that we found significant main effects of test phase, but no condition by test-phase interaction. The actor did not explicitly comment on the strategies she was using, and the demonstration was devoid of ostensive pedagogical signals (Csibra & Gergely, 2009) that were present in many previous CVS studies (Schwichow et al., 2016). Future work may consider the role of similar demonstrations and counterfactual prompts separately to identify the extent to which they may yield different benefits.

Our findings are surprising in light of previous studies, which found that children required more intensive instruction and scaffolding in order to improve, with some of these interventions even taking place over the course of several sessions (e.g., Schauble, 1996). Even with a subtle manipulation in the form of two counterfactual questions following a demonstration, children showed improvement in their ability to conduct a controlled test of a hypothesis.

Children in the counterfactual condition were able to conduct a controlled test both on the variables they had already encountered and on two new variables, with more than half of children in the counterfactual condition scoring 2/2 on both post-tests. In contrast, children in the control condition showed less evidence of transfer, with a minority of children scoring 2/2 in the post-test transfer phase.

These findings provide preliminary evidence that counterfactual prompts may be a promising pedagogical tool for supporting CVS. However, these results do not allow us to pinpoint the precise mechanism by which counterfactuals may confer this benefit. We have suggested that

counterfactuals may serve as *imagined interventions*, helping learners to connect their intuitive causal reasoning abilities to the current task. This suggestion is in line with previous work emphasizing the relation between causal and counterfactual reasoning (e.g., Gopnik & Schulz, 2007; Sloman 2005; Gopnik & Walker, 2013).

However, other work suggests that counterfactuals may have a *general* effect on reasoning by activating a "mindset" that is open to alternatives. Previous research shows that prompts to consider alternatives in the form of counterfactuals (Galinsky & Moskowitz, 2000) or multiple explanations (Hirt & Markman, 1995) have wide-reaching effects, with individuals showing generally debiased reasoning across a range of settings. Researchers studying these effects have suggested that counterfactuals activate a *mental simulation mindset* that breaks the reasoner free of a singular viewpoint or hypothesis and incites consideration of alternative, and potentially contrasting possibilities. In ongoing research, we are currently investigating whether children prompted with counterfactuals on one task (e.g., ramps) show improvement on a far-transfer task (e.g., pendulums) to better understand the potential mechanisms by which counterfactual prompts may support performance. In the present study, our counterfactual questions were about the experimental design and specifically pertained to the control-of-variables process. It is an open question whether counterfactual questions about a peripheral or irrelevant feature of the task (e.g., the color of the ball) would scaffold performance. An alternative "mindset" account would predict that counterfactuals should be beneficial regardless of their focus.

Our counterfactual prompts not only focused on the control of variables process, but also specifically invited children to imagine a *confounded test*. An alternate explanation for children's success in the counterfactual condition may therefore be that by engaging children in imagining a confounded test, our prompts led them to recognize that such tests were inconclusive and that they should avoid producing such tests themselves. However, the lack of a relation between children's "can't be sure" responses and their ability to control variables suggests that children did not need to explicitly recognize the inconclusiveness of a confounded test in order to benefit from the process of thinking counterfactually. In other words, the effect of the counterfactual prompts appears to be distinct from the specific response they elicit. This finding aligns with research on children's self-explanation showing that the *process* of generating explanations benefits children's causal reasoning, regardless of the specific explanations they produce (e.g., Walker, Lombrozo, Legare, & Gopnik, 2014).

Another possibility is that our counterfactual prompts drew children's attention to both values of the variable that was held constant (e.g., "Let's imagine that she set Ramp 1 to low" when she had set both ramps to high), whereas the control prompts did not (e.g., "Let's imagine again what happened to the ball on Ramp 1"). This may have made children more likely to consider and control the alternate variable. In a follow-up study, we have adapted our control prompt to highlight both levels of the alternate variable to investigate whether this accounts for children's better performance in the counterfactual condition.

Although we are not yet able to identify the precise mechanism by which counterfactuals confer the benefits observed, these findings connect to a wider body of results that suggest that drawing children's attention to alternatives benefits their scientific inquiry (e.g., Sodian et al, 1991; Engle & Walker, 2018). For instance, children in Sodian et al. (1991) were able to recognize a conclusive test of a hypothesis when presented with two contrasting hypotheses, and, as mentioned above, Engle and Walker (2018) found that counterfactual prompts scaffolded children's ability to detect anomalies during causal learning. These results suggest that thinking of counterfactuals and alternatives may benefit a range of scientific inquiry skills.

Conclusion

Children prompted to think counterfactually showed improvements in their ability to conduct controlled tests of a hypothesis – an ability previous studies have suggested requires direct instruction or intensive scaffolding. These results support theoretical proposals about the role of counterfactuals in scientific reasoning, and suggest that counterfactuals may have educational utility. The prompts used in the current study are short and simple, and could easily be implemented in a range of formal and informal learning contexts.

Acknowledgments

This work was supported by funding from the Social Sciences and Humanities Research Council of Canada, including an Insight Grant to P.A. Ganea, a Postdoctoral Fellowship Award to A. Nyhout, and a Canada Graduate Scholarship-Master's to A. Iannuzziello. We are grateful to the children and families who participated in this research, and to the Ontario Science Centre for facilitating data collection. We thank Hanna Lim for assistance with building stimuli, transcription, and coding, Jayun Bae for acting in videos, and Etri Kocaqi for assistance with coding.

References

Amsel, E., & Brock, S. (1996). The development of evidence evaluation skills. *Cognitive Development*, *11*, 523-550.

Buchsbaum, D., Bridgers, S., Weisberg, D. S., & Gopnik, A. (2012). The power of possibility: Causal learning, counterfactual reasoning, and pretend play. *Philosophical Transactions of the Royal Society B: Biological Sciences, 367*, 2202-2212.

Carey, S. (1985). *Conceptual change in childhood.* Cambridge, MA: MIT Press.

Chen, Z., & Klahr, D. (1999). All other things being equal: Acquisition and transfer of the control of variables strategy. *Child Development*, *70*, 1098-1120.

Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in Cognitive Sciences*, *13*, 148-153.

Engle, J., & Walker, C.M. (2018). Considering alternatives facilitates anomaly detection in preschoolers. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), Proceedings of the 40th Annual Conference of the Cognitive Science Society. Madison, WI: Cognitive Science Society.

Erb, C.D. & Sobel, D.M. (2014). The development of diagnostic reasoning about uncertain events between ages 4-7. *PLOS One, 9(3):* e92285.

Frosch, C. A., McCormack, T., Lagnado, D. A., & Burns, P. (2012). Are Causal Structure and Intervention Judgments Inextricably Linked? A Developmental Study. *Cognitive Science*, *36*, 261–285.

Galinsky, A. D., & Moskowitz, G. B. (2000). Counterfactuals as behavioral primes: Priming the simulation heuristic and consideration of alternatives. *Journal of Experimental Social Psychology, 36*, 384–409.

Gopnik, A., Meltzoff, A. N., & Bryant, P. (1997). *Words, thoughts, and theories*. Cambridge, MA: MIT Press.

Gopnik, A., & Schulz, L. (2007). *Causal learning: Psychology, philosophy, and computation.* New York, NY: Oxford University Press.

Gopnik, A., & Sobel, D.M. (2000). Detecting blickets: How young children use information about novel causal powers in categorization and induction. *Child Development*, *71*, 1205-1222.

Gopnik, A., & Walker, C. M. (2013). Considering counterfactuals: The relationship between causal learning and pretend play. *American Journal of Play*, *6*, 15-28.

Harris, P. L., German, T., & Mills, P. (1996). Children's use of counterfactual thinking in causal reasoning. *Cognition*, *61*, 233-259

Hirt, E.R., & Markman, K.D. (1995). Multiple explanation: A consider-an-alternative strategy for debiasing judgments. *Journal of Personality & Social Psychology*, *69*, 1069-1086.

Keil, F. C. (1992). *Concepts, Kinds, and Cognitive development*. Cambridge, MA: MIT Press.

Klahr, D., Fay, A. L., & Dunbar, K. (1993). Heuristics for scientific experimentation: A developmental study. *Cognitive Psychology*, *25*, 111–146.

Klahr, D., & Nigam, M. (2004). The equivalence of learning paths in early science instruction: Effects of direct instruction and discovery learning. *Psychological Science*, *15*, 661–667.

Klahr, D., Zimmerman, C., & Jirout, J. (2011). Educational interventions to advance children's scientific thinking. *Science*, *333*, 971-975.

Kuhn, D., & Franklin, S. (2006). The second decade: What develops (and how)?. In D. Kuhn & R. Siegler (Eds.), *Cognition, perception, and language. Volume 2 of the Handbook of child psychology* (6th ed.). Hoboken, NJ: Wiley.

Lewis, D. K. (1986). *On the plurality of worlds* (Vol. 322). Oxford: Blackwell.

McCormack, T., Ho, M., Gribben, C., O'Connor, E., & Hoerl, C. (2018). The development of counterfactual reasoning about doubly-determined events. *Cognitive Development*, *45*, 1-9.

National Research Council. (2000). Inquiry and the national science education standards: A guide for teaching and learning. *National Academies Press*.

Nyhout, A., Henke, L., & Ganea, P. A. (2019). Children's counterfactual reasoning about causally overdetermined events. *Child Development*, *9*, 610-622.

Nyhout, A., & Ganea, P. A. (2019). Mature counterfactual reasoning in 4-and 5-year-olds. *Cognition*, *183*, 57-66.

Rafetseder, E., & Perner, J. (2014). Counterfactual reasoning: Sharpening conceptual distinctions in developmental studies. *Child Development Perspectives, 8*, 54–58.

Rafetseder, E., Schwitalla, M., & Perner, J. (2013). Counterfactual reasoning: From childhood to adulthood. *Journal of Experimental Child Psychology*, *114*, 389-404.

Schauble, L. (1996). The development of scientific reasoning in knowledge-rich contexts. *Developmental Psychology*, *32*, 102–119.

Schulz, L. E., & Gopnik, A. (2004). Causal learning across domains. *Developmental Psychology*, *40*, 162-176.

Schulz, L. E., Gopnik, A., & Glymour, C. (2007). Preschool children learn about causal structure from conditional interventions. *Developmental Science*, *10*, 322-332.

Schwichow, M., Croker, S., Zimmerman, C., Hoffler, T., & Hartig, H. (2016). Teaching the control-of-variables strategy: A meta-analysis. *Developmental Review*, 39, 37–63.

Sloman, S. (2005). *Causal models: How people think about the world and its alternatives*. Cambridge, MA: Oxford University Press.

Sodian, B., Zaitchik, D., & Carey, S. (1991). Young children's differentiation of hypothetical beliefs from evidence. *Child Development*, *62*, 753-766.

Walker, C.M. & Gopnik, A. (2013). Causality & Imagination. In Marjorie Taylor (Ed.), *The development of imagination*. Oxford University Press: New York.

Walker, C. M., & Gopnik, A. (2014). Toddlers infer higher-order relational principles in causal learning. *Psychological Science*, *25*, 161-169.

Walker, C. M., Lombrozo, T., Legare, C., & Gopnik, A. (2014). Explanation prompts children to privilege inductively rich properties. *Cognition, 133,* 343-357.

Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. *Developmental Review*, *27*, 172–223.

Learning the Proportional Nature of Probability from Feedback

Shaun O'Grady
shaun.ogrady@berkeley.edu
Department of Psychology

Geoffrey Saxe
saxe@berkeley.edu
Department of Education
University of California Berkeley

Fei Xu
fei_xu@berkeley.edu
Department of Psychology

Abstract

People make decisions based on probabilistic information every day and often use inaccurate, heuristic decision rules. Although a great deal of research has investigated the developmental trajectory of accurate probability judgments, very little research has investigated how the learning process unfolds. In the current study a microgenetic experimental design was deployed to investigate the influence of feedback on children's probabilistic decision making strategies. Seven- to ten-year-old children (N = 50) first performed a computer-based task to assess the type of strategy they use in a probabilistic judgment task. Next, children receive feedback on a series of 24 trials and then perform a post-test consisting of the same computer-based strategy assessment. Findings revealed that some strategies may benefit from feedback more than others. These results suggest that children can learn about the proportional nature of probability from feedback alone and that the amount and type of feedback influence the learning process.

Keywords: probabilistic reasoning; numerical cognition

Introduction

Every day, children and adults are presented with decisions involving probabilistic information and decades of research has shown that their decisions are often influenced by heuristic biases (Kahneman, 2011; Kahneman & Tversky, 1973; Tversky & Kahneman, 1983). Even on tasks involving simple random draws in which a decision maker is asked to choose between two different options with known probabilities both children and adults have been observed using heuristic decision rules (Falk, Yudilevich-Assouline, & Elstein, 2012; O'Grady & Xu, 2019; Pacini & Epstein, 1999). Although these studies have reported a wealth of data on the factors which promote and inhibit biased decision making, and developmental research has reported a large number of experiments on the developmental trajectory of probabilistic reasoning, very little research has investigated the role of outcome feedback (i.e. the result of a random draw) on children's use of heuristics in probabilistic decision making tasks. In the current paper, a microgenetic experimental design is used to assess the influence of feedback on children's decision making strategies in a simple random draw task.

Developmental trajectory of probabilistic reasoning

Developmental research originally devised by Piaget & Inhelder (1975) studied children's choices in a 2-alternative forced-choice (2AFC) random draw task in which they are presented with two groups of marbles, each contain both red and white marbles in different amounts and are asked to choose the group that is best for getting a target color marble. Piaget and Inhelder (1955) found that young children rely on heuristic decision rules, such as 'pick the group with the greatest number of target outcomes' when making these simple probability judgments and argued that these decision biases suggest children have difficulty with part-whole reasoning, often making comparisons of parts (i.e. making a simple comparison about the number of favorable marbles in each group) without taking into account the relation between the part and the whole (i.e. calculating the proportion of favorable to total outcomes). Although decades of research led to incremental improvements in Piagetian methods in this subfield of developmental psychology (Chapman, 1975; Falk, Falk, & Levin, 1980; Fischbein, Pampu, & Mnzat, 1970), the most thorough and recent experiment conducted by Falk et al. (2012) provides the greatest insight into children's probabilistic reasoning abilities.

Falk et al. (2012) devised a 2AFC random draw task involving a series of 24 trials in which 4- to 11-year-old children were presented with a choice between two groups of marbles each containing a number of favored and unfavored marbles. Children were then asked to choose the group with the best chance of yielding their favored color marble from a single random draw. The 24 trials were designed to discern between four of the most common strategies children have been shown to use in similar tasks. These strategies are (1) 'pick the group with more favorable marbles' ('more favorable') or (2) 'pick the group with the least unfavorable marbles' ('less unfavorable') (3) 'pick the group with the largest difference between favorable and unfavorable marbles' ('greater difference') and (4) the formally correct strategy of 'pick the group with the highest proportion of favorable marbles' ('greater proportion'). Findings from a series of experiments revealed that children progress from 1-dimensional strategies in which they focus on either favorable or unfavorable outcomes (strategies 1 & 2) to more complicated, 2-dimensional strategies in which they attend to both favorable and unfavorable outcomes (strategies 3 & 4). Results revealed that children begin to use the formally correct, proportional strategy around 8 years of age.

Children's difficulties with fraction representations of rational number are notorious and the errors are so common that they are often termed the 'whole number bias' (Ni & Zhou, 2005; Siegler, 2016). This bias presents itself in several different ways across many types of tasks and is very similar to the errors children make in part-whole reasoning during 2AFC random draw tasks. For example, when children are asked to choose the greater of two fractions (say, 1/3 or 2/7) they often choose incorrectly based on comparisons of either the numerator or the denominator (in the above example, choosing 2/7 because the numbers are larger).

Teaching Children Probability Concepts

In a review of the research on statistical edcuation, Garfield & Ahlgren (1988) argue that school-age children have difficulty developing an intuitive understanding of fundamental topics in probabiltiy and statistics for three reasons. First, students have difficulty reasoning about rational number and proportions. Second, probability concepts often conflict with students' real-world experience. Finally, Garfield & Ahlgren (1988) argue that students often develop an aversion to statistics and probability because they learn about these concepts at a very abstract and formal level. Previous research has addressed several of these concerns in their attempts to improve children's understanding of probability.

In formal mathematics, probability is represented as a rational number between 0 and 1, and in the 2AFC random draw task these probabilities are computed as proportions of favorable outcomes. Several groups of researchers have attempted to teach children strategies for calculating and reasoning about probability. Fischbein & Gazit (1984) investigated the effect of teaching probability on 10-13 year olds' predictions of probabilistic outcomes using survey questions about the results of rolling two dice. Students in the experimental group received 12 lessons in which they were taught computational strategies and conceptual relationships in probability. Interestingly, the results revealed that while the experimental group outperformed the control group on computational questions (i.e. questions in which children needed to apply a specific algorithm to identify the correct answer), there was no significant difference in performance on conceptual questions (i.e. questions in which children needed to generalize a concept to a novel context). Although Fischbein & Gazit (1984) report the use of a successful intervention, it is possible that the children in the experimental group merely learned the computational algorithms for solving probability problems without changing their prior concepts about part-whole relations in probability.

Using a didactic approach, Castro (1998) taught young high-school students (14-15 year olds) formal probability. This method encouraged the teachers to incorporate student's intuitive understanding of probability into lessons by allowing students to reflect on their experiences. The experimental group demonstrated significantly more improvement from pre-test to post-test on both probability-reasoning and probability calculation tests. An analysis of the amount of children who changed their answers on similar questions from pre-test to post-test revealed that there were more students who switched their answers in the conceptual change group than in the traditional teaching method group. However, since this study included older teens who may have had experience with formal probability, it is impossible to tell if the conceptual change was a result of the teaching method alone or the interaction of prior conceptual understanding and instruction.

In an intervention study, Nunes, Bryant, Evans, Gottardis, & Terlektsi (2014) attempted to teach 10-year-old children about the importance of understanding the sample space when calculating probability. Participants in the experimental group participated in seven, 50-minute lessons on sample space and probability led by a researcher. Another condition received lessons in mathematical problem solving while a control group of children stayed in the classroom and received regular lessons from their teachers. All three groups received four assessments on understanding sample space, a pre-test, and three post-tests given at various points throughout the program. The experimental group outperformed the two control groups on all three post-tests.

Calculating probability based on proportion of outcomes in the sample space can be accomplished through the use of several cultural forms such as absolute number, ratios, fractions, proportions and percentages. The function of calculating probability is not inherent in any one of these forms and the decision to use one form over another entails a complex interaction of social and individual factors as well as aspects of the problem for which the chosen form is recruited. Nunes and colleagues (2014) argue that ratio representations may be better suited for teaching probability to 9-11 year olds because children understand ratios earlier than proportions and most probability problems are based on proportional judgments. However, since ratio judgments only provide part-part comparisons they may prime children to make correspondences between two different quantities rather than integrating the two quantities by using proportions.

Prior knowledge and instructional context

Discordant assumptions about communicative exchanges can be problematic during instruction if a teacher and a learner view the same forms as supporting different functions (Saxe, 2004). Using a quasi-experimental design, Saxe, Gearhart, & Seltzer (1999) investigated the influence of children's prior understanding of fractions and classroom practices on mathematics learning. Children were categorized as either having or not having a rudimentary part-whole understanding of fractions and classrooms were rated on a scale of alignment with reform standards. High alignment was characterized by the degree to which a teacher draws out and expands upon a student's mathematical knowledge as well as the extent to which conceptual issues are highlighted during problem solving tasks. Importantly, classrooms that espouse either self-discovery or procedural memorization would be considered low in alignment with reform policies. Results revealed that high classroom alignment with reform standards predicted greater performance on a post-test requiring a conceptual understanding of fractions and this effect was stronger for children without a rudimentary understanding of fractions. Interestingly, there was no clear relationship between classroom alignment with reform standards and performance on computational problems regardless of students' prior understanding of fractions. With low levels of alignment to reform principles students without a rudimentary understanding had no basis with which to structure their goals and may have relied on their prior conceptual understanding of integers. However, with supportive classroom environments in which teachers

seek to draw-out and build upon a learner's prior knowledge, children can more easily engage with mathematical goals and stand a better chance of learning.

These findings highlight the importance of both prior knowledge and instructional context on a child's ability to learn mathematics. Educators have long understood the importance of providing children with specific feedback based on their prior conceptual knowledge. Indeed, Saxe et al. (1999) found that fraction learning outcomes are a function of a learner's prior understanding (whole number vs rudimentary fraction understanding) and instructional context. Teachers who are able to identify a child's prior conceptual understanding of fractions can construct a learning environment that either confirms accurate conceptualization or scaffolds the learner towards a more thorough conceptualization.

Rationale for the current study

Can children learn to avoid whole number biased choices in probability tasks when they are provided with feedback about the outcomes of their choices? What features of the instructional context allow them to override their non-proportional strategies? We hypothesize that children require consistent feedback on problems which conflict with their prior knowledge and we predict that children who are provided with such feedback will be more likely to reject their incorrect strategy compared to children who are provided with a mix of conflicting and non-conflicting examples.

Methods

Participants

The current experiment was pre-registered (`http://aspredicted.org/blind.php?x=mp6gc9`) with a target sample of 80 children between the ages of 7 and 10 (20 children in each age group: 7-year-olds, 8-year-olds, 9-year-olds, and 10-year-olds), which was determined based on previous research using a similar task (Falk et al., 2012). Currently, data have been collected from N = 50 children (19 7-year-olds, Mean age = 7.5, SD = 0.26; 13 8-year-olds, Mean age = 8.45, SD = 0.23; 9 9-year-olds, Mean age = 9.28, SD = 0.2; and 9 10-year-olds, Mean age = 10.2, SD = 0.2). All fifty children participated in the first session and three children declined to participate in the follow-up session 1 week later (1 7-year-old, 1 8-year-old, and 1 10-year-old).

Material

Images depicting two gumball machines and two groups of green and purple marbles were rendered using Blender (Version 2.78) 3D animation software. Following Falk et al. (2012), each trial image was internally labeled with the trial type designators 'GGGG', 'GGGS', 'SSSG', and 'SSSS' with each letter representing the dimension of comparison and the letter itself relating the correct choice (higher probability of yielding the child's favored color marble) to the incorrect choice (lower probability of yielding the child's favored color). For each target color (i.e. green or purple), two sets of 24 images were created using the same distributions used by Falk et al. (2012) for a total of 96 images.

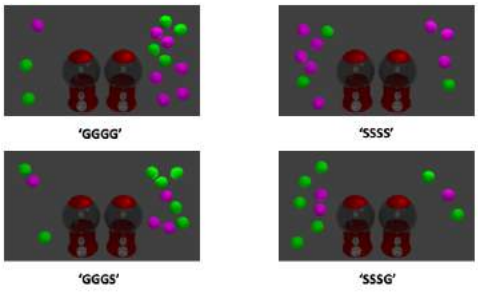


Figure 1: Example images for each of the 4 trial types. In all 4 images, the correct choice for obtaining a purple marble is located on the right side of the image.

Figure 1 presents an example image for each trial type. Note that the correct choice in the figure on the top left (labeled 'GGGG') has a greater amount of favored marbles (1st G), a greater amount of non-favored marbles (2nd G), a greater total of favored and non-favored marbles (3rd G) and a greater difference between favored and non-favored marbles (4th G). In contrast, the correct choice for the image on the top right (labeled 'SSSS') has a smaller amount of marbles in each of these categories compared to the incorrect choice. Children using a strict 'more favorable' strategy would make a correct choice on all 12 'GGGG' and 'GGGS' trials but would choose incorrectly on all 12 'SSSS' and 'SSSG' trials. A child using a strict 'less unfavorable' strategy would make a correct choice on all 12 'SSSS' and 'SSSG' trials but would choose incorrectly on all 12 'GGGG' and 'GGGS' trials. Children using a strict, 'greater difference' would make a correct choice on all 12 'GGGG' and 'SSSG' trials but would choose incorrectly on all 12 'SSSS' and 'GGGS' trials. Finally, a child using the formally correct proportional strategy would choose correctly on all 24 trials.

Procedure

Children were seated approximately 60 cm away from a MacBook Pro laptop (OSX; Screen resolution 1280 x 800) and told they would play a game in which they would try to collect green or purple marbles from one of two different gumball machines. The task consisted of a self-paced game automated using the psychophysics toolbox written for the MatLab programming language (Brainard, 1997; Kleiner, Brainard, & Pelli, 2007; Pelli, 1997). In order to maintain an average testing time of 20 minutes, the experiment was split into 2 testing sessions spaced 1 week apart. Children completed the *assessment phase* during session 1 and then completed the *conflict phase* and *post-test phase* during session 2.

Assessment Phase During the first testing session, the experimenter explained the task and children were prompted to choose their favorite of the two colors, either purple or green.

For each of the 24 images, the computer presented the image at random along with 4 counting prompts, one for each group of marbles (i.e. "How many (green/purple) marbles are on this side (left/right)?"). The child responded by pressing the appropriate number key on the keyboard. An error message was presented if the child chose the wrong number and the game did not progress until the child pressed the correct number key. Counting prompts for each color and side were randomized for each image. Once children completed the counting prompts, they were prompted with the question "Which would you pick to get a (green/purple) marble?". Importantly, the position of the marbles on the screen were randomized to prevent children from choosing based on the positions of their favorite color marble. However, in order to ensure that children did not rely on the placement of the marbles on the screen they were told to "Pretend that the marbles will go into the machines and that the machines will be shaken up so you don't know what's going to come out next."

Following the methods outlined by Falk et al. (2012), the MatLab program discerned which strategy the child used based on their performance on each of the four trial types. After children completed the *assessment phase*, the MatLab program calculated point scores for each strategy based on the choices that the child made. Whichever strategy had the highest point score was deemed to be that child's strategy. Point scores could range from 0 to 24 with 0 indicating no strategy-consistent responses and 24 indicating perfect strategy use. Participants using the 'more favorable' strategy provided about 21 (M = 21.6; SD = 3.5) out of 24 strategy-consistent responses, while those using the 'less unfavorable' strategy provided 16 (M = 16; SD = 1.79) out of 24 strategy-consistent responses, participants using the 'greater difference' strategy provided 20 (M = 20.2; SD = 20.2) out of 24 strategy-consistent responses, and participants using the 'greater proportion' strategy provided 19 (M = 19.78; SD = 19.78) out of 24 strategy-consistent responses.

Conflict Phase Children were semi-randomly assigned to one of two different conditions ensuring that an equal number of children using each strategy were assigned to both conditions. In the 'half-conflict' condition, children viewed all 24 trials, 12 of which conflicted with the child's strategy and 12 of which did not conflict. Children in the 'high-conflict' condition viewed 24 trials that conflicted with their strategy.

In the 'high conflict' condition, feedback trials were assigned as follows. Children designated as using the 'more favorable' strategy viewed 12 'SSSS' trials and 12 'SSSG' trials. Children using the 'less favorable' strategy viewed 12 'GGGG' trials and 12 'GGGS' trials. A child using the 'greater difference' strategy viewed 12 'SSSS' trials and 12 'GGGS' trials while children using the proportional strategy were simply assigned to the 'half-conflict' condition as none of the trials conflicted with their strategy. In all conditions and for all trials, children received feedback in the form of either a favored or unfavored color marble returned in the dispenser of the machine they chose. Importantly, all feedback was provided deterministically, meaning that if a child chose strictly according to their non-proportional strategy in the 'high-conflict' condition, they would receive 24 unfavored marbles and a child in the 'half-conflict' condition would receive 12 favored and 12 unfavored marbles. Children in the 'half-conflict' condition received a mix of confirmatory and dis-confirmatory feedback with respect to their strategy while children in the 'high-conflict' condition received only dis-confirmatory feedback.

Post-test phase After completing the *conflict phase*, each child received the same 24 trials they viewed in the *assessment phase* one week prior in a randomized order. Importantly, the *post-test phase* is an immediate post-test because it occurred directly following the *conflict phase*.

Results

The results of each phase of the experiment are reported separately below along with a brief discussion section. For all three phases, analyses consisted of comparisons of Generalized Linear Regression Models with Mixed effects (GLMMs) using the lme4 package written for the R statistical programming language (Bates, Maechler, Bolker, & Walker, 2015). All models predicted the binary response variable while holding participant ID as a random effect. Nested models were compared using Chi Squared tests for model fits while non-nested models were compared using the Akaike Information Criterion (AIC), a measure of model fit in which models with smaller AICs are preferred over models with higher AICs. For all three phases of the experiment, modeling results revealed no influence of participant gender, favored color, on performance. Model coefficients for GLMMs are reported as log-odds, that is, the log of the odds ratio of correct to incorrect responses.

Assessment Phase Results

In order to investigate the influence of age on strategy use we used a Chi Squared test to assess the independence of age group and strategy. Results of the χ^2 test revealed that older children were significantly more likely to use the correct proportional strategy ($\chi^2(9, n = 50) = 18.11$, $p = .034$). Figure 2 presents the proportions of children using each strategy by age. Note that the two younger age groups (7-year-olds and 8-year-olds) are predominantly relying on the, 'more favorable' strategy whereas children in the two older age groups (9-year-olds and 10-year-olds) have a more equal spread across the four different strategies.

Comparisons of GLMMs revealed that the model with the best fit to the assessment phase data was the model predicting performance from strategy alone ($AIC_{Strategy}$ =1562.96). This model outperformed the null model (AIC_{null} = 1613.55; χ^2= 56.6; df = 3; $p < .001$), as well as the model predicting performance from age (AIC_{Age} =1609.59). More complex models predicting performance from age and strategy ($AIC_{Strat+Age}$ = 1561.79; χ^2= 3.17; df = 1; p = .07) and the interaction of age and strategy ($AIC_{Strat*Age}$ = 1564.95; χ^2= 6.01; df =

4; p = .20) did not perform better than the model for strategy alone. Thus, the simpler model is preferred since it can predict the same amount of variance with fewer model parameters. There was no effect of trial number indicating that children's performance did not improve with time during the *assessment phase*.

Inspection of model coefficients reveals that the log-odds of a correct reponse increased for children using the 'greater difference' ($\beta_{>F-U}$ = 0.53; SE = 0.2; 95% CI [0.13, 0.93]), and 'less unfavorable' strategies ($\beta_{<U}$ = 0.17; SE = 0.18; 95% CI [-0.19, 0.53]), as well as those using formally correct proportional strategy ($\beta_{>F/F+U}$ = 1.49; SE = 0.19; 95% CI [1.11, 1.87]) compared to children using the 'more favorable' ($\beta_{>F(Intercept)}$ = 4.003^{-16}; SE = 0.07; 95% CI [-0.09, 0.2]). However, only the coefficients for 'greater difference' and 'greater proportion' strategies reached statistical significance (Wald test: 'greater difference': $p < .01$; 'greater proportion': $p < .001$).

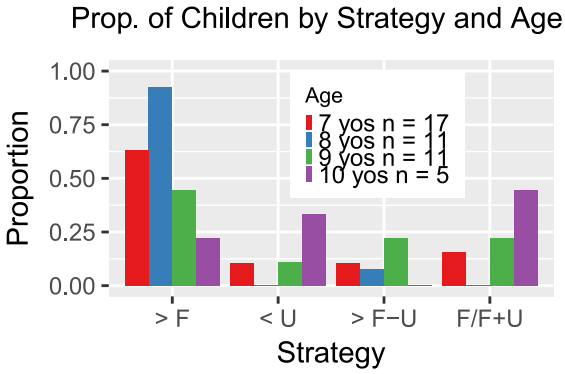


Figure 2: Proportion of children using each strategy by age group. Strategies are designated as follows: '> F': more favorable; < U: less unfavorable ; > F-U: greater difference; > F/F+U: greater proportion.

Assessment Phase Discussion

Results of the current study converge with those of previous reports indicating that children's use of the correct proportional strategy improves with age (Falk et al., 2012; O'Grady & Xu, 2018). Importantly, results of the GLMM comparisons revealed an effect of strategy on performance indicating that children who attended to the number both of favorable and unfavorable marbles in each choice performed better than children who made their choices based on one single dimension (i.e. choosing based solely on the number of favorable or unfavorable marbles).

Conflict Phase Results

Nine children were found to be using the formally correct proportional strategy during the assessment phase the previous week. Since there are no trials that conflict with this strategy, data from these children were excluded from the conflict phase analyses resulting in a sample size of N = 38. Comparisons of GLMMs for this subsample revealed that the model with the best fit to the data predicted performance from the conflict condition and the trial number as well as the interaction between the two variables ($AIC_{Condition*Trial}$ =1143.09). This model outperformed the null model (AIC_{null} = 1613.55; χ^2= 29.74; df = 3; $p < .001$) as well as the simpler models predicting performance from conflict condition ($AIC_{Condition}$ = 1158.39; χ^2= 19.3; df = 2; $p < .001$) and trial number alone (AIC_{Trial} = 1162.96; χ^2= 23.87; df = 2; $p < .001$) and the more complex model accounting for both the conflict condition and trial number without an interaction ($AIC_{Condition+Trial}$ = 1154.51; χ^2= 13.43; df = 1; $p < .001$). There were no significant effects of the three non-proportional strategies, nor was there an interaction between conflict condition.

Inspection of the model coefficients revealed that the log-odds of a correct decreased for children in the 'high-conflict' condition ($\beta_{High-Conflict}$ = -0.13; SE = 0.35; 95% CI [-0.81, 0.55]) compared to children in the 'half-conflict' condition ($\beta_{Half-Conflict}$ = 0.32; SE = 0.25; 95% CI [-0.18, 0.81]), which is not surprising considering that all of the trials in the 'high-conflict' condition conflicted with the children's strategies whereas only 12 of the 24 trials in the 'half-conflict' condition conflicted with the children's strategies. While the model coefficient for trial number was slightly negative (β_{Trial} = -0.01; SE = 0.01; 95% CI [-0.04, 0.02]) indicating a decrease in the log-odds of a correct response, the interaction between trial number and condition revealed that in the 'high-conflict' condition, trial number had a positive effect on the log-odds ($\beta_{Trial*High-Conflict}$ = 0.08; SE = 0.02; 95% CI [0.04, 0.12]). The interaction between trial order and the 'high-conflict' condition was the only model coefficient to reach statistical significance (Wald test: $p < .001$) indicating that performance improved over time in the 'high-conflict' condition suggesting that children in this condition may have learned from feedback on earlier trials.

Conflict Phase Discussion

Results revealed that both conflict condition and trial order had an effect on performance. Importantly, the interaction between conflict condition and trial number produced the greatest positive effect on performance while the coefficient for the 'high-conflict' condition alone had a negative effect on performance. This set of results suggests that children in the 'high-conflict' condition began by choosing according to their strategy but then switched to another strategy after several trials in which they received negative feedback.

Post-Test Phase Results

Of the 9 children who used the correct proportional strategy during the *assessment phase* only one child (a 10-year-old) did not continue to use the correct proportional strategy. Interestingly, this child used the 'more favorable' strategy and reported that they switched to a simpler strategy because "the game was boring and I wanted to finish it faster" suggesting that this child understood the time-accuracy tradeoff among

the various potential strategies. Table 1 presents the number of children using each of strategy in the *post-test phase* ('Post-test' column) based on the child's *assessment phase* strategy ('Assessment' column) and condition ('half-conflict' and 'high-conflict' columns).

Assessment	Post-test	Half-Conflict	High-Conflict
>F	>F	9	3
>F	<U	1	7
>F	>F-U	1	1
>F	>F/F+U	2	4
<U	>F	2	1
<U	<U	0	1
<U	>F-U	0	0
<U	>F/F+U	0	1
>F-U	>F	0	0
>F-U	<U	0	0
>F-U	>F-U	2	0
>F-U	>F/F+U	0	3

Table 1: This table presents the number of children using each strategy listed in the Post-test column during the post-test phase after using the strategy in the Assessment column during the assessment phase.

Comparisons of GLMMs revealed that the model with the best fit to the data predicted *post-test phase* performance based on the interaction between conflict condition and *assessment phase* strategy ($AIC_{Condition*Strategy}$ =1182.6). This model outperformed the null model (AIC_{null} = 1613.55; χ^2= 19.41; df = 5; $p < .001$), the models for conflict condition alone ($AIC_{Condition}$ = 1189.82; χ^2= 15.22; df = 4; $p < .001$) and *assessment phase* strategy alone ($AIC_{Strategy}$ = 1188.48; χ^2= 11.88; df = 3; p = .01) as well as the model for conflict condition and *assessment phase* without any interactions ($AIC_{Condition+Strategy}$ = 1185.81; χ^2= 7.21; df = 2; p = .03).

Inspection of model coefficients revealed that the only model coefficient to reach statistical significance was the interaction between 'greater difference' strategy and the 'high-conflict' condition which increased the log-odds of a correct reponse ($\beta'_{>F-U'*High-Conflict}$ = 1.74; SE = 0.65; 95% CI [0.46, 3.01]; Wald test: $p < 0.01$) compared to the children using the 'more favorable' strategy in the 'half-conflict' condition ($\beta_{Intercept}$ = 0.05; SE = 0.35; 95% CI [-0.64, 0.73]). All remaining coefficients did not reach statistical significance. Figure 3 presents the proportion of correct responses by conflict condition and *assessment phase* strategy.

Post-Test Phase Discussion

The interaction between conflict condition and *assessment phase* strategy indicates that children using the 'greater difference' strategy benefited more from the 'high-conflict' condition compared to children using the other 2 strategies. Although these findings are promising, there were only three children using the 'greater difference' and three children using the 'less unfavorable' assigned to the 'high-conflict' condition thus more data will be needed to make any firm conclusions. However, it is interesting to view the differences between the 'half-conflict' and 'high-conflict' condition for children using the 'more favorable' strategy. Note from table 1 that only 3 of the 15 children using this strategy in the 'high-conflict' condition (20%) continued using their strategy after the feedback condition while 9 of the 13 assigned to the 'half-conflict' condition (69.2%) continued to use the strategy.

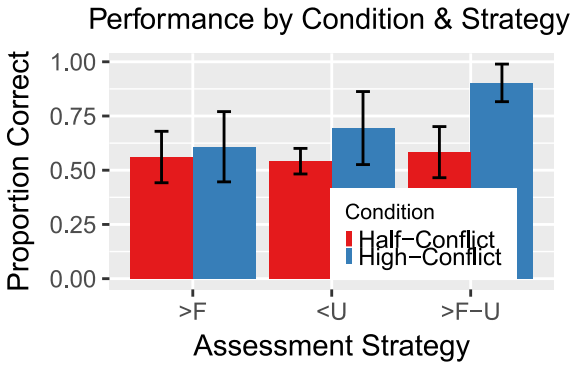


Figure 3: Proportion of correct responses in the post-test phase by conflict condition and assessment phase strategy. Error bars indicate standard deviation.

General Discussion

The current findings provide three important insights into the development of proportional reasoning in probability judgments. First, during the *assessment phase*, children using the correct proportional strategy performed better than children using non-proportional strategies and there was no effect of trial order indicating that children relied on the same strategy throughout the task. These findings provide an important replication of previous research (Falk et al., 2012; O'Grady & Xu, 2018). Second, during the *conflict phase*, results revealed an interaction between trial order and conflict condition indicating that children in the 'high-conflict' condition performed better on later trials compared to earlier trials while this effect was not found in the 'half-conflict' condition. Previous research investigating the influence of feedback in similar decision making tasks have observed the effect of feedback on a single trial (Falk et al., 2012) or presented feedback on a limited number of trials (O'Grady & Xu, 2018). In contrast, the current approach allows for the observation of a comprehensive set of trials allowing for a more thorough understanding of how feedback influences strategy change. Finally, results from the *post-test phase* revealed that while children in the 'high-conflict' condition were more likely to abandon their strategy, children using the 'greater difference' strategy seemed to have gained the most from this feedback. Although more data need to be collected, these preliminary results suggest that strategy-specific feedback can help children overcome 'whole number bias' in probability tasks.

Our findings shed new light on how children learn about probability. In the 'half-conflict' condition, we attempted to mimic the experience a child would gain from actively exploring the environment. In contrast, the 'high-conflict' condition was meant to provide a learning context tailored to the child's prior understanding of proportional relations in probability.

Constructivist theories of cognitive development highlight the interaction between a learner's prior knowledge and new information gained through their own active exploration as well as through socio-cultural processes like education (Piaget & Inhelder, 1975; Vygotsky, 1962). By assessing the child's prior understanding and then presenting examples which conflict with that understanding, the computer program in the 'high-conlfict' condition is acting much like a constructivist teacher, identifying the learner's prior knowledge, providing them with conflicting evidence, and allowing the child to construct a new conceptual understanding.

Why do children benefit from feedback in the 'high-conflict' condition but not from the feedback on the 12 conflicting trials in the 'half-conflict' condition? Children understand the uncertain nature of probability, that is, they understand that they may receieve their non-favored marble even though they chose the 'best' option for getting their favorite color. In the 'half-conflict' condition this prior understanding of uncertainty allows children to continue to believe their inaccurate strategy is correct because the negative feedback can be chalked up to chance. However, children in the 'high-conflict' condition are forced to reconcile their inacurate strategy with the evidence at hand. Thus they must reject their prior conceptualization of probability and construct a more accurate representation. Although the current evidence suggests that children in the 'high-conflict' condition reject their prior conceptualization of probability, future research will be necessary to uncover the new representations that children construct as well as the process through which this transition ocurrs.

Acknowledgements

We would like to thank the families who participated in this study as well Berkeley Early Learning Lab staff for their support. This research was funded by the National Science Foundation Graduate Research Fellowship under Grant No. (DGE 1106400) S.M.O., and an NSF grant to F. Xu (#1640816).

References

Bates, Maechler, Bolker, & Walker. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Brainard. (1997). The psychophysics toolbox. *Spatial Vision*, *10*, 433–436.

Castro, C. S. (1998). Teaching probability for conceptual change la enseanza de la probabilidad por cambio conceptual. *Educational Studies in Mathematics*, *35*(3), 233–254.

Chapman, R. H. (1975). The development of children's understanding of proportions. *Child Development*, *46*, 141–148.

Falk, R., Falk, R., & Levin, I. (1980). A potential for learning probability in young children. *Educational Studies in Mathematics*, *11*(2), 181–204.

Falk, R., Yudilevich-Assouline, P., & Elstein, A. (2012). Children's concept of probability as inferred from their binary choices—revisited. *Educational Studies in Mathematics*, *81*(2), 207–233.

Fischbein, E., & Gazit, A. (1984). Does the teaching of probability improve probabilistic intuitions? *Educational Studies in Mathematics*, *15*(1), 1–24.

Fischbein, E., Pampu, I., & Mnzat, I. (1970). Comparison of ratios and the chance concept in children. *Child Development*, *41*, 377–389.

Garfield, J., & Ahlgren, A. (1988). Difficulties in learning basic concepts in probability and statistics: Implications for research. *Journal for Research in Mathematics Education*, 44–63.

Kahneman, D. (2011). *Thinking fast & slow*. New York, NY: Farrar, Strauss,; Giroux.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*, 237–251.

Kleiner, Brainard, & Pelli. (2007). What's new in psychtoolbox-3? *Perception*, *36*.

Ni, Y., & Zhou, Y.-D. (2005). Teaching and learning fraction and rational numbers: The origins and implications of whole number bias. *Educational Psychologist*, *40*(1), 27–52.

Nunes, T., Bryant, P., Evans, D., Gottardis, L., & Terlektsi, M.-E. (2014). The cognitive demands of understanding the sample space. *ZDM*, *46*(3), 437–448.

O'Grady, S., & Xu, F. (2018). Whole number bias in children's probability judgments. In *Proceedings of the 40th annual conference of the cognitive science society*.

O'Grady, S., & Xu, F. (2019). The development of non-symbolic probability judgments in children the development of non-symbolic probability judgements in children. *Child Development*.

Pacini, R., & Epstein, S. (1999). The interaction of three facets of concrete thinking in a game of chance. *Thinking and Reasoning*, *5*(4), 303–325.

Pelli. (1997). The videotoolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*, 437–442.

Piaget, J., & Inhelder, B. (1975). *The origins of the idea of chance in children*. New York, NY: Norton & Company.

Saxe, G. B. (2004). Practices of quantification from a sociocultural perspective. *Cognitive Developmental Change: Theories, Models and Measurement*, 241–263.

Saxe, G. B., Gearhart, M., & Seltzer, M. (1999). Relations between classroom practices and student learning in the domain of fractions. *Cognition and Instruction*, *17*(1), 1–24.

Siegler, R. S. (2016). Magnitude knowledge: The common core of numerical development. *Developmental Science*, *19*(3), 341–361.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*(4), 293–315.

Vygotsky. (1962). *Thought and language*. Cambridge, MA: The MIT press.

Distinguishing Effects of Executive Functions on Literacy Skills in Adolescents

Teresa M. Ober[1] (tober@gradcenter.cuny.edu)
Patricia J. Brooks[1,2] (pbrooks1@gc.cuny.edu)
Bruce D. Homer[1,3] (bhomer@gc.cuny.edu)
Jan L. Plass[3] (jan.plass@nyu.edu)

[1]Department of Educational Psychology, The Graduate Center, City University of New York
[2]Department of Psychology, College of Staten Island, City University of New York
[3]CREATE Lab, New York University

Abstract

This study investigated direct and indirect effects of executive functions (EF) on reading comprehension in 87 adolescents (*mean age* = 14.0 years, *SD* = 1.5). The operation span task was used to measure the *updating* aspect of working memory, the plus-minus task to measure *task-switching*, and the numerical Stroop task to measure *inhibitory control*. Literacy skills tasks assessed nonword decoding, text recall/inference, and passage comprehension. Regression models indicated that EF measures accounted for significant variance in literacy skills after controlling for age and fluid intelligence. Working memory was associated with passage comprehension, task-switching with nonword decoding, and inhibitory control with nonword decoding as well as text recall/inference. Parallel mediation models tested for indirect effects of EF constructs via decoding and text recall/inference. Working memory showed direct and indirect effects on passage comprehension, the latter mediated by text recall/inference. Task-switching was associated with decoding, but its relation to passage comprehension was not significant. Inhibitory control showed indirect effects on passage comprehension via decoding and text recall/inference. Results indicate overlapping but distinct contributions of EF to literacy skills.

Keywords: reading comprehension, literacy skills, decoding, text recall/inference, executive functions, working memory, task-switching, inhibitory control

Introduction

Reading comprehension is an active process that involves weaving together information contained within a text to construct a coherent, accurate representation of its meaning (Kintsch, 1994). Various theoretical models have proposed that reading comprehension relies on the interplay of literacy subskills, including *decoding* (i.e., mapping orthographic units onto phonological units), *recall* (i.e., activation of previously encountered information), and *inference* (i.e., drawing conclusions to make sense of information). These subskills, in turn, rely on sustained attention and other manifestations of executive functions (EF). The goal of the current study was to explore direct and indirect associations between EF, literacy subskills, and reading comprehension. Our purpose was to shed light on sources of individual differences in reading ability, which in turn inform theoretical models of reading.

The *simple view of reading* identifies decoding and linguistic comprehension as two critical skills supporting readers in constructing meaning from text (Gough & Tunmer, 1986; Hoover & Gough, 1990). Decoding involves the utilization of spelling-to-sound (grapheme-to-phoneme) rules to translate printed text into spoken language. Through decoding, readers are able to sound out words quickly and accurately, and thus gain fluency in recognizing letters in words and words in text.

The *dual route model* (Coltheart, 2006) further distinguishes the processes involved in decoding words. According to this model, word reading occurs either through a lexical route, which involves accessing lexical representations through familiar spelling patterns, or through a non-lexical (phonological) route, which utilizes knowledge of letter-sound associations (i.e., phoneme-to-grapheme correspondence rules) to sound out words. Although the two routes are thought to be separable, readers utilize both routes in parallel, which may place considerable demands on EF.

In addition to decoding, models of the development of reading emphasize the importance of text recall and inference skills (Cain, Oakhill, & Lemmon, 2004; García & Cain, 2014). The *construction-integration model* outlines the process by which readers construct meaning from text (Kintsch & Mangalath, 2011): Readers achieve coherence by organizing information across sentences and linking it with broader contextual and background knowledge (Graesser, Singer, & Trabasso, 1994; Kintsch & Mangalath, 2011). Text representations may encode information verbatim or may encode the gist (Reyna, Corbin, Weldon, & Brainerd, 2016). In constructing such representations, readers rely on recall and inferential processes that bridge information (e.g., to resolve ambiguities, identify pronominal referents, establish causal relations), bring together verbatim and gist representations, and subsequently validate inferences against general knowledge (Singer, Harkness, & Stewart, 1997). Such operations are demanding of cognitive resources, especially working memory (Peng et al., 2018).

EF and the Development of Literacy Skills

EF broadly refers to a constellation of cognitive skills thought to be essential in the planning, monitoring, and control of cognitive processes. According to the unity and diversity framework, EF has three main components: working memory (also referred to as updating), task-switching, and inhibitory control (Miyake et al., 2000). The current study focused on individual differences in these three EF components and how they each influence decoding, text

recall/inference, and reading comprehension in adolescents. As children become more fluent readers capable of recognizing familiar words with automaticity, less cognitive effort needs to be exerted to decode text, thus freeing up cognitive resources to better comprehend and critically understand the meaning behind the text (Kuhn et al., 2010; LaBerge & Samuels, 1974).

Working memory involves maintaining and/or updating information in response to task demands (Baddeley, 2012). As shown in a recent meta-analysis (Follmer, 2018), working memory appears to have a moderate positive association with reading comprehension (r = .38, 95% CI [.34 : .43]). It is less clear whether working memory bears an equally strong relation to decoding skill. In a study with 7- to 8-year-olds, Oakhill, Cain, and Bryant (2003) found that measures of working memory, text integration, and metacognitive monitoring accounted for individual differences in reading comprehension, whereas performance on phoneme deletion, a phonological awareness task, explained variance in word reading. Their findings suggest that working memory may have a limited association with decoding, and a more direct association with reading comprehension.

Task-switching, or the ability to shift between different conceptual representations and rule sets, supports a wide variety of academic tasks including reading (Best, Miller, & Jones, 2009). Meta-analyses have reported a significant, albeit weak, association (r = .21, 95% CI [.11 : .31]) between task-switching and reading achievement in children (Yeniad, Malda, Mesman, van Ijzendoorn, & Pieper, 2013) and a moderate correlation (r = .39, 95% CI [.20 : .56]) between task-switching and reading comprehension in participants ranging from age 6 years to adults (Follmer, 2018). In a study involving 1st and 2nd graders, Cartwright et al. (2017) found that variation in reading comprehension was associated with performance on a color-shape cognitive flexibility task (a measure of task-switching), even after accounting for decoding ability. To date, few studies have examined direct associations between decoding and task-switching, though there is some evidence of a significant, albeit weak, association (Kieffer, Vukovic, & Berry, 2013).

To construct accurate text representations, readers also need to suppress competing sources of information and interpretations that may be concurrently activated (Gernsbacher & Faust, 1991). The mechanism of suppression is thought to stem from inhibitory control processes. An association between reading comprehension and inhibitory control has been reported in various studies with children (e.g., Kieffer et al., 2013), although a recent meta-analysis (Follmer, 2018), spanning ages from 6 years to adults, reported that the association between reading comprehension and inhibitory control was relatively weak (r = .21, 95% CI [.13 : .30]). The strength of this association in decoding is not well established.

Control Variables: Fluid Intelligence and Age

Over childhood and adolescence, reading ability typically improves. This age-related trend likely stems from accumulated experience with oral and written language in the context of formal education (Stanovich, 1986), as well as maturation of linguistic and cognitive abilities, such as improved lexical access (Logan, Schatschneider, & Wagner, 2011) and EF (Christopher et al., 2012). Prior research also suggests that fluid intelligence, i.e., the ability to solve novel reasoning problems, may correlate with specific EF components (Brydges, Reid, Fox, & Anderson, 2012), as well as early literacy skills (Blair & Razza, 2007). However, other studies suggest that individual differences in EF, most notably in working memory, largely account for the contribution of fluid intelligence to literacy skills in children and adolescents (Alloway & Alloway, 2010). In addition, not all EF components appear to be equally correlated with measures of fluid intelligence. Some prior research has reported a strong association between fluid intelligence and working memory (Unsworth, Fukuda, Awh, & Vogel, 2014), but not between fluid intelligence and task-switching or inhibitory control (Friedman et al., 2006). Taken together, previous research suggests the need to control for age and fluid intelligence in efforts to elucidate the unique contribution of EF components to reading skills, including decoding, text recall/inference, and reading comprehension.

Research Objectives

The current study used a battery of assessments to explore relations between components of EF (working memory, task-switching, inhibitory control) and literacy skills (decoding, text recall/inference, and passage comprehension). First, we sought to determine the extent to which the three components of EF were uniquely and directly associated with each literacy skill after controlling for other factors known to be related to reading ability (i.e., fluid intelligence and age). Second, we examined indirect associations between each EF component in relation to passage comprehension as mediated by nonword decoding and text recall/inference. We hypothesized that: (1) some aspects of EF would account for variation in the reading subskills of nonword decoding and text recall/inference; (2) some aspects of EF would account for variance in reading comprehension; and that (3) indirect associations between aspects of EF and reading comprehension would emerge by way of the reading subskills of nonword decoding and text recall/inference.

Method

Participants

Teachers from partnering schools (two middle schools and two high schools in New York City) brought their classes to a university research lab where their students were invited to participate in various computer-based studies including the current study. Only students whose parents had provided written consent were eligible to participate. The sample comprised of 87 students in grades 6 to 12 (49 females, 35 males, and 3 who did not disclose gender), ranging in age from 12 to 17 years (*mean* = 14.0, *SD* = 1.5).

Tasks and Measures

Working Memory. *The operation span task*, a complex span measure shown to correlate with moderately challenging to difficult arithmetic and reading tasks (Unsworth, Heitz, Schrock, & Engle, 2005), was used to assess working memory. Reliability on operation span tasks has been found to range between .70 to .80, depending on scoring methods (Conway et al., 2005), or approximately .77 using split-half reliability coefficient alphas (Kane et al., 2004). In the task used here, participants were instructed to perform simple arithmetic operations (e.g., (3 x 4) + 11 = ?) and indicate whether an answer was correct or incorrect. Between each arithmetic problem, participants were shown a letter to remember. The task presented three blocks of trials, with each trial consisting of an arithmetic problem followed by a letter. At the end of each block, the participant was asked to recall the letters in that block in the order presented. As an index of working memory, we calculated the proportion of correctly ordered letters across the three blocks of trials.

Task-switching. *The plus-minus task* was given as a measure of task-switching (Miyake et al., 2000); reliability of scores on this task has been estimated as approximately .60 using split-half reliability (Del Missier, Mäntylä, & Bruine de Bruin, 2010). In our version of the task, participants were shown three lists of 30 two-digit numbers and asked to perform numerical computations as quickly as possible on each number in the list. For List 1, participants were instructed to add 3 to each two-digit number; for List 2, they were instructed to subtract 3 from each number; for List 3, they were instructed to alternate between adding or subtracting 3 from each number. Standardized mix cost scores (z-scores) were used as an index of task-switching, based on prior studies (e.g., Miyake et al., 2000).

Inhibitory Control. We administered a shortened version of the *numerical Stroop task* (McVay & Kane, 2012) as a measure of inhibitory control. Reliability estimates of the numerical Stroop task indicate sufficient reliability **(**Cronbach α = .71; McVay & Kane, 2012). Our numerical Stroop task presented three blocks of trials in which participants were asked to identify the number of figures shown in an image on the computer screen. In Block 1 (five trials), the participant was shown a series of Xs (ranging from 1 to 9) and instructed to indicate the number of Xs presented (e.g., 5 in response to X X X X X). In Block 2 (five trials), they were shown a series consisting of a repeated digit (ranging from 1 to 9), with the length of the series also varying between 1 and 9 and consistent with the number of digits present (e.g., 4 4 4 4). In Block 3 (five trials), the digit and the number of digits in the series was never the same (e.g., 5 5 5) with the participant instructed to indicate the number of digits while ignoring the digit value. We calculated the number of correct responses in Block 3 as an index of inhibitory control.

Decoding Ability. We used a *nonword decoding task* to assess participants' ability to apply knowledge of grapheme-phoneme correspondences to pronounce letter strings. The nonword decoding task was based on an orally administered task, previously developed for research purposes (Hogan, Catts, & Little, 2005). It used five nonwords that followed phonotactic constraints of standard American English: *bos*, *bune*, *cim*, *gep*, *phoncher*. Participants were shown each nonword along with five options for a phonetically equivalent alternate spelling, with instructions stating, "Select the spelling that most closely matches the pronunciation of the word provided." For the item where the target nonword was *bos*, options included *bose, boz, doz, pose*, and *doze* (correct response is *boz*). Scores were calculated as the proportion of items answered correctly.

Text Recall and Inference. The *component reading processes task* is a multicomponent assessment of the ability to integrate knowledge while comprehending text (Hannon & Daneman, 2001). We used a modified computerized version that assessed participants' ability to recall information and make inferences across statements. Participants were given two three-sentence paragraphs describing relations between three nonwords (nouns), with each sentence relating a pair of nonwords (e.g., *A RILI resembles a DARF but is slower and larger.*) and appearing on a separate line. Participants read the first paragraph and answered four questions, then read the second paragraph and answered four additional questions. Participants were given up to 40 seconds to read each paragraph before being prompted with a set of questions that were presented without the paragraph in view.

Subscores (proportions of correct responses) calculated for each question type (i.e., recall and inference) were highly correlated, $r_p(85) = .49$, $p < .001$, after controlling for age. Subsequently, scores for text recall/inference were computed as the average between the two subscores.

Passage Comprehension. We administered a practice test from the New York State 12th grade English Language Arts Regents Exam (NYSED, 2012). The test presented two passages (one expository, one narrative) of equivalent length (i.e., 38 and 41 sentences; 551 and 559 words). Each passage had an accompanying 7-item multiple-choice test, with four response options per item. Accuracy (percentage correct) was used as the measure of reading passage comprehension.

Fluid Intelligence. A set of *Raven's progressive matrices* (Raven, 2000) was used to assess nonverbal fluid intelligence. The task consisted of five incomplete visual matrices, each with 5 to 8 possible options from which to choose a pattern to complete the matrix. The task has been shown to have robust indicators of reliability, with a test-retest Pearson correlation coefficient of .93 (Burke, 1972). Scores were computed as the proportion of correct responses.

Background Variables. A demographics questionnaire was administered following the research tasks. It included questions about the participant's gender, age, and first language learned (coded as English or not English). These variables were included as possible control variables in preliminary regression models predicting literacy skills.

Procedure

Upon arrival to the lab, students were provided with information about the study. After assenting to participate

they were seated at computer stations to complete the computer-based tasks, administered via Qualtrics software. Students completed the reading passage comprehension test either before or after the computer-based tasks; this was randomized across participants.

Results

Table 1 presents descriptive statistics for the assessments of literacy skills and EF tasks.

Table 1. Descriptive statistics (*N*=87).

Measure	*M* (*SD*)
Raven's Progressive Matrices	61.6% (18.1%)
Executive Functions	
Operation Span Average	47.4% (28.4%)
Plus-Minus Mix Cost (*z*)	0.00 (1.00)
Numerical Stroop	48.5% (38.7%)
Literacy Skills	
Nonword Decoding	52.4% (28.4%)
Text Recall/Inference	52.7% (22.6%)
Reading Passage Comprehension	68.4 % (20.9%)

Preliminary Correlational Analyses

We examined partial correlations (controlling for age) across measures of literacy skills. After adjustment for multiple comparisons (Bonferroni-controlled $\alpha = .0167$), significant correlations were observed between the scores on the passage comprehension test and both nonword decoding, $r_p(85) = .37$, $p < .001$, and text recall/inference, $r_p(85) = .44$, $p < .001$. Nonword decoding and recall/inference were not significantly associated, $r_p(85) = .19$, $p = .084$.

We also examined partial correlations (controlling for age) between measures of EF (operation span for working memory, mix costs on the plus-minus task for task-switching, and numerical Stroop for inhibitory control) and fluid intelligence (Raven's progressive matrices). After adjustment for multiple comparisons (Bonferroni-corrected $\alpha = .0083$), none of the partial correlations were statistically significant, see Table 2. There was a trend towards an association between fluid intelligence and working memory, $r_p(85) = .27$, $p = .012$.

Table 2. *Descriptive Statistics and Age-controlled Partial Correlations for EF Variables (N=87)*

	WM	TS	IC
Working Memory			
Task-switching	–.03		
Inhibitory Control	–.02	.07	
Fluid Intelligence	.27	–.12	.18

WM: Operation Span, TS: Plus-Minus Mix Cost (z-score), IC: Numerical Stroop

Regression Analyses of Reading Subskills

Regression models were used to assess whether EF components accounted for variation in literacy skills.

Nonword Decoding. The overall model was significant, $F(6, 80) = 6.59$, $p < .001$, $R^2 = 33$. Task-switching and inhibitory control were significantly associated with nonword decoding, see Table 3.

Table 3. Multiple regression with nonword decoding as the outcome measure (*N*=87).

Variable	*β*	*SE*	*t*	*p*
Age	.18	.02	†1.68	.098
Fluid Intelligence	.19	.16	†1.81	.074
Recall-Inference	.00	.15	–.03	.974
Working Memory	.08	.11	.73	.468
Task-switching	.27	.03	**2.87	.005
Inhibitory Control	.31	.08	**3.06	.003

***$p < .001$, **$p < .01$, *$p < .05$, †$p < .10$

Component Reading Processes: Recall/Inference. The overall model was significant, $F(6, 80) = 8.81$, $p < .001$, $R^2 = .40$. Age, working memory, and inhibitory were significantly associated with text recall/inference, see Table 4.

Table 4. Multiple regression with text recall/inference as the outcome measure (*N*=87).

Variable	*β*	*SE*	*t*	*p*
Age	.32	.01	**3.22	.002
Fluid Intelligence	.13	.12	1.32	.190
Nonword Decoding	.00	.08	–.03	.974
Working Memory	.21	.08	*2.14	.035
Task-switching	.07	.02	.70	.483
Inhibitory Control	.27	.06	**2.78	.007

***$p < .001$, **$p < .01$, *$p < .05$

Reading Passage Comprehension. The overall model was significant, $F(7, 79) = 10.34$, $p < .001$, $R^2 = .48$. Fluid intelligence, nonword decoding, text recall/inference, and working memory were significantly associated with scores on the reading passage comprehension test, see Table 5.

Table 5. Multiple regression with reading passage comprehension as the outcome measure (*N*=87).

Variable	*β*	*SE*	*t*	*p*
Age	–.03	.01	–.29	.776
Fluid Intelligence	.21	.11	*2.25	.027
Nonword Decoding	.21	.07	*2.11	.038
Recall-Inference	.29	.10	**2.77	.007
Working Memory	.28	.07	**3.01	.004
Task-switching	.08	.02	.98	.332
Inhibitory Control	–.00	.05	–.02	.987

***$p < .001$, **$p < .01$, *$p < .05$

Mediation Analyses

Mediation analyses were run to test whether EF components had indirect associations with reading passage comprehension via nonword or text recall/inference skills. Constraints due to the number of observations and free parameters prevented a single model from being analyzed lest it be under-identified (Kline, 2015). Thus, three separate

parallel mediation models tested for direct and indirect effects of EF measures on reading passage comprehension; see Figure 1 for the analytic model. Note that in the models for each EF construct, age and fluid intelligence were added as covariates associated with passage comprehension.

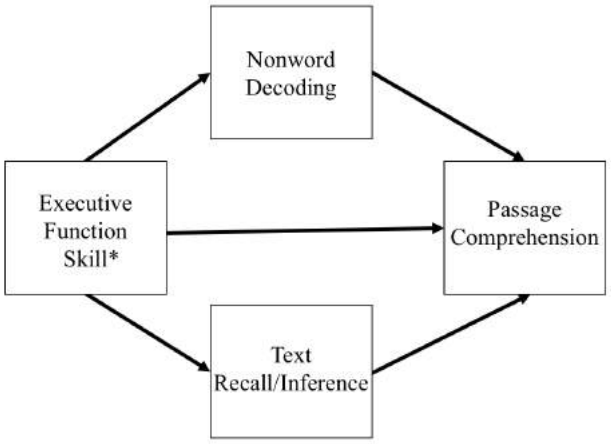


Figure 1. Parallel mediation model showing a direct path from EF skill to reading passage comprehension, and indirect paths through nonword decoding and text recall/inference. Each EF skill was entered as a predictor into one of three separate models with this form.

Indirect Effects of Working Memory. The first mediation analysis confirmed a statistically significant association between working memory and reading passage comprehension, total effect: β = .46, 95% CI [.45, .48], *SE* = .07, z = 4.71, p < .001; direct effect: β = .29, 95% CI [.28, .31], *SE* = .07, z = 2.97, p = .003. Tests of indirect effects indicated significant mediation via text recall/inference, β = .12, 95% CI [.11, .13], *SE* = .03, z = 2.62, p = .009. The indirect effect by way of nonword decoding was not significant, β = .05, 95% CI [.05, .06], *SE* = .02, z = 1.72, p = .087. These results suggest the association between working memory and passage comprehension is direct; however, text recall/inference partially mediates the association.

Indirect Effects of Task-switching. The mediation analysis failed to provide evidence that task-switching ability was associated with passage comprehension, total effect: β = .16, 95% CI [.15, .16], *SE* = .02, z = 1.56, p = .12, direct effect: β = .09, 95% CI [.08, .09], *SE* = .02, z < .01, p = .36.

Indirect Effects of Inhibitory Control. The mediation analysis indicated that inhibitory control was associated with reading passage comprehension, but the effect was indirect; for the total effect: β = .22, 95% CI [.21, .23], *SE* = .05, z = 2.14, p =.032; for the direct effect: β = –.06, 95% CI [–.07, –.04], *SE* = .06, z = –.39, p = .60. Tests of mediation indicated a significant indirect effect of inhibitory control on passage comprehension via nonword decoding, β = .12, 95% CI [.11, .12], *SE* = .03, z =2.41, p = .016, and a significant indirect effect of inhibitory control on passage comprehension via text recall/inference, β = .16, 95% CI [.15, .17], *SE* = .03, z = 2.93, p = .003. These results suggest that inhibitory control influences passage comprehension through its associations with both decoding and text recall/inference abilities.

Discussion

The current study aimed to identify relations between specific EF components (working memory, task-switching, inhibitory control) and literacy skills (nonword decoding, text recall/inference, and reading passage comprehension) in adolescents. Understanding sources of individual differences in literacy skills has implications for developing interventions and refining theoretical models of reading. Such research is urgent given estimates that 1 out of every 10 children in the United States experiences reading difficulties, even among children with average or above average levels of intelligence (National Institutes of Health, 2010).

As a preliminary step in modeling effects of EF on literacy skills, we ran correlational analyses. These indicated a lack of unity across EF measures; hence the EF constructs were treated as separable in subsequent models. After accounting for influences of age and fluid intelligence, regression analyses identified a direct relation between working memory and reading passage comprehension. This result implicating working memory in performance of a complex and integrative reading comprehension task is in line with previous literature (Peng et al., 2018). Working memory also exhibited an indirect association with reading passage comprehension by way of text recall/inference, such that the higher one's operation span, the better one is able to read text fluently and make inferences based on its meaning, and subsequently construct accurate text representations.

Part of the novelty of our findings is in showing that working memory may play a lesser role in lower-level literacy skills, such as nonword decoding, than in higher-level skills, such as text recall/inference processes and reading passage comprehension. Our results corroborate Oakhill et al. (2003) in finding a significant direct association between measures of working memory and reading comprehension, but not between working memory and decoding. However, such an association has been reported by others (Christopher et al., 2012; Kieffer et al., 2013). In light of these mixed findings, a meta-analysis may be warranted to ascertain the relation of working memory to decoding.

Unlike working memory, task-switching was significantly associated only with nonword decoding. This is consistent with prior work that found an association between task-switching and word reading (e.g., Cartwright, 2012), and suggests that the ability to shift attention is instrumental for retrieving and applying letter-sound associations. Although the current study focused only on nonword decoding, we expect task-switching to impact decoding more generally. In relation to the dual-route model (Coltheart, 2006), readers must flexibly alternate between reliance on the lexical and nonlexical routes as they encounter both familiar and unfamiliar words. Within the more transparent French orthography, task-switching has been found to correlate with decoding (Colé, Duncan, & Blaye, 2014), suggesting an association independent of orthographic depth.

Inhibitory control was associated with nonword decoding and with text recall/inference abilities. In contrast to working memory, inhibitory control did not show a direct relation to

passage comprehension. These findings are consistent with a previous large-scale study of adolescents that also found inhibitory control to be associated with decoding ability, but not with reading comprehension (Arrington, Kulesz, Francis, Fletcher, & Barnes, 2014). Thus, as in the current study, the effect of inhibitory control on reading comprehension appeared to be indirect and mediated by decoding ability.

Limitations

The simple view distinguishes decoding ability and linguistic comprehension as factors underlying reading comprehension. However, as we did not assess linguistic comprehension (e.g., receptive vocabulary and grammar) independently of text, it is difficult to apply the current findings to this framework. We also recognize that some cognitive assessments may be poorly suited for individual differences research (Hedge, Powell, & Sumner, 2017); hence future work should not rely on single measures to assess underlying EF constructs (see Denckla, 1994).

Conclusions

Our findings indicate that different components of EF have distinct relations with literacy skills in adolescents, which were evident after accounting for a number of control variables previously shown to influence reading abilities. We did not find evidence in support of unity across EF constructs. Given the complexity inherent to both reading and EF, it is perhaps not surprising that the relation between these cognitive processes is multifaceted. Our findings suggest that problems with a number of different EF skills may underlie reading difficulties in adolescents. Prior research has found a paucity of evidence that EF may be targeted to improve overall academic skills such as reading (Jacob, & Parkinson, 2015), and that it has limited potential in identifying responsiveness to targeted academic skills interventions (Miciak, Cirino, Ahmed, Reid, & Vaughn, 2019). Nevertheless, as the current study indicates, there is evidence of associations between EF and reading skills. Translating these findings into interventions to support reading comprehension will require further work.

References

Alloway, T.P., & Alloway, R.G. (2010). Investigating the predictive roles of working memory and IQ in academic attainment. *Journal of Experimental Child Psychology*, *106*(1), 20–29.

Arrington, C.N., Kulesz, P.A., Francis, D.J., Fletcher, J.M., & Barnes, M.A. (2014). The contribution of attentional control and working memory to reading comprehension and decoding. *Scientific Studies of Reading*, *18*(5), 325–346.

Baddeley, A. (2012). Working memory: Theories, models, and controversies. *Annual Review of Psychology*, *63*, 1–29.

Best, J.R., Miller, P.H., & Jones, L.L. (2009). Executive functions after age 5: Changes and correlates. *Developmental Review*, *29*(3), 180–182.

Blair, C., & Razza, R.P. (2007). Relating effortful control, executive function, and false belief understanding to emerging math and literacy ability in kindergarten. *Child Development, 78*(2), 647–663.

Brydges, C.R., Reid, C.L., Fox, A.M., & Anderson, M. (2012). A unitary executive function predicts intelligence in children. *Intelligence*, *40*(5), 458–469.

Burke, H.R. (1972). Raven's progressive matrices: Validity, reliability, and norms. *The Journal of Psychology*, *82*(2), 253–257.

Cain, K., Oakhill, J., & Lemmon, K. (2004). Individual differences in the inference of word meanings from context: The influence of reading comprehension, vocabulary knowledge, and memory capacity. *Journal of Educational Psychology*, *96*(4), 671–681.

Cartwright, K.B. (2012). Insights from cognitive neuroscience: The importance of executive function for early reading development and education. *Early Education & Development*, *23*(1), 24–36.

Cartwright, K. B., Coppage, E. A., Lane, A. B., Singleton, T., Marshall, T. R., & Bentivegna, C. (2017). Cognitive flexibility deficits in children with specific reading comprehension difficulties. *Contemporary Educational Psychology*, *50*, 33–44.

Christopher, M.E., Miyake, A., Keenan, J.M., Pennington, B., DeFries, J.C., Wadsworth, S.J., ... & Olson, R.K. (2012). Predicting word reading and comprehension with executive function and speed measures across development: A latent variable analysis. *Journal of Experimental Psychology: General*, *141*(3), 470–488.

Colé, P., Duncan, L. G., & Blaye, A. (2014). Cognitive flexibility predicts early reading skills. *Frontiers in Psychology*, *5*, 565.

Coltheart, M. (2006). Dual route and connectionist models of reading: An overview. *London Review of Education*, *4*(1), 5–17.

Conway, A.R., Kane, M.J., Bunting, M.F., Hambrick, D.Z., Wilhelm, O., & Engle, R.W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review*, *12*(5), 769–786.

Del Missier, F., Mäntylä, T., & Bruine de Bruin, W. (2010). Executive functions in decision making: An individual differences approach. *Thinking & Reasoning*, *16*(2), 69–97.

Denckla, M.B. (1994). Measurement of executive function. In G.R Lyon (Ed.), *Frames of reference for the assessment of learning disabilities: New views on measurement issues* (pp. 117–142). Baltimore, MD: P.H. Brookes Publishing.

Follmer, D.J. (2018). Executive function and reading comprehension: A meta-analytic review. *Educational Psychologist*, *53*(1), 42–60.

Friedman, N.P., Miyake, A., Corley, R.P., Young, S.E., DeFries, J.C., & Hewitt, J.K. (2006). Not all executive functions are related to intelligence. *Psychological Science*, *17*(2), 172–179.

García, J.R., & Cain, K. (2014). Decoding and reading comprehension: A meta-analysis to identify which reader and assessment characteristics influence the strength of the

relationship in English. *Review of Educational Research*, *84*(1), 74–111.

Gernsbacher, M.A., & Faust, M.E. (1991). The mechanism of suppression: A component of general comprehension skill. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17*(2), 245–262.

Gough, P.B., & Tunmer, W.E. (1986). Decoding, reading, and reading disability. *Remedial and Special Education*, *7*(1), 6–10.

Graesser, A.C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review*, *101*(3), 371–395.

Hannon, B., & Daneman, M. (2001). A new tool for measuring and understanding individual differences in the component processes of reading comprehension. *Journal of Educational Psychology*, *93*(1), 103–128.

Hedge, C., Powell, G., & Sumner, P. (2017). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, 1–21.

Hogan, T.P., Catts, H.W., & Little, T.D. (2005). The relationship between phonological awareness and reading: Implications for the assessment of phonological awareness. *Language, Speech, and Hearing Services in Schools*, *36*(4), 285–293.

Hoover, W.A., & Gough, P.B. (1990). The simple view of reading. *Reading and Writing, 2*(2), 127–160.

Jacob, R., & Parkinson, J. (2015). The potential for school-based interventions that target executive function to improve academic achievement: A review. *Review of Educational Research*, *85*(4), 512–552.

Kane, M.J., Hambrick, D.Z., Tuholski, S.W., Wilhelm, O., Payne, T.W., & Engle, R.W. (2004). The generality of working memory capacity: A latent-variable approach to verbal and visuospatial memory span and reasoning. *Journal of Experimental Psychology: General*, *133*(2), 189–217.

Kline, R. B. (2015). *Principles and practice of structural equation modeling*. New York, NY: Guilford Publications.

Kieffer, M.J., Vukovic, R.K., & Berry, D. (2013). Roles of attention shifting and inhibitory control in fourth-grade reading comprehension. *Reading Research Quarterly*, *48*(4), 333–348.

Kintsch, W. (1994). Text comprehension, memory, and learning. *American Psychologist*, *49*(4), 294–303.

Kintsch, W., & Mangalath, P. (2011). The construction of meaning. *Topics in Cognitive Science*, *3*(2), 346–370.

Kuhn, M. R., Schwanenflugel, P. J., & Meisinger, E. B. (2010). Aligning theory and assessment of reading fluency: Automaticity, prosody, and definitions of fluency. *Reading Research Quarterly*, *45*(2), 230–251.

LaBerge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology*, *6*(2), 293–323.

Logan, J.A., Schatschneider, C., & Wagner, R.K. (2011). Rapid serial naming and reading ability: The role of lexical access. *Reading and Writing*, *24*(1), 1–25.

McVay, J.C., & Kane, M.J. (2012). Why does working memory capacity predict variation in reading comprehension? On the influence of mind wandering and executive attention. *Journal of Experimental Psychology: General*, *141*(2), 302–332.

Miciak, J., Cirino, P. T., Ahmed, Y., Reid, E., & Vaughn, S. (2019). Executive functions and response to intervention: Identification of students struggling with reading comprehension. *Learning Disability Quarterly*, *42*(1), 17–31.

Miyake, A., Friedman, N.P., Emerson, M.J., Witzki, A.H., Howerter, A., & Wager, T.D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology*, *41*(1), 49–100.

New York State Education Department (2012, June). *12th Grade Comprehensive English Exam*. Retrieved from http://www.nysedregents.org/comprehensiveenglish/612/engl62012-exam_w.pdf.

National Institutes of Health. (2010). *Fact sheet: Reading difficulty and disability*. Retrieved from: https://report.nih.gov/nihfactsheets/ViewFactSheet.aspx?csid=114&key=R.

Oakhill, J. V., Cain, K., & Bryant, P. E. (2003). The dissociation of word reading and text comprehension: Evidence from component skills. *Language and Cognitive Processes*, *18*(4), 443–468.

Peng, P., Barnes, M., Wang, C., Wang, W., Li, S., Swanson, H.L., ... & Tao, S. (2018). A meta-analysis on the relation between reading and working memory. *Psychological Bulletin*, *144*(1), 48–76.

Raven, J. (2000). The Raven's progressive matrices: change and stability over culture and time. *Cognitive Psychology*, *41*(1), 1–48.

Reyna, V.F., Corbin, J.C., Weldon, R.B., & Brainerd, C.J. (2016). How fuzzy-trace theory predicts true and false memories for words, sentences, and narratives. *Journal of Applied Research in Memory and Cognition*, *5*(1), 1–9.

Singer, M., Harkness, D., & Stewart, S.T. (1997). Constructing inferences in expository text comprehension. *Discourse Processes*, *24*(2–3), 199–228.

Stanovich, K.E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly*, *21*(4), 360–407.

Unsworth, N., Fukuda, K., Awh, E., & Vogel, E.K. (2014). Working memory and fluid intelligence: Capacity, attention control, and secondary memory retrieval. *Cognitive Psychology*, *71*, 1–26.

Unsworth, N., Heitz, R.P., Schrock, J.C., & Engle, R.W. (2005). An automated version of the operation span task. *Behavior Research Methods*, *37*(3), 498–505.

Yeniad, N., Malda, M., Mesman, J., van IJzendoorn, M.H., & Pieper, S. (2013). Shifting ability predicts math and reading performance in children: A meta-analytical study. *Learning and Individual Differences*, *23*, 1–9.

Shift of probability weighting by joint and separate evaluations: Analyses of cognitive processes based on behavioral experiment and cognitive modeling

Yutaro Onuki (onuki-yutaro32@g.ecc.u-tokyo.ac.jp)
Graduate School of Arts and Sciences, The University of Tokyo
3-8-1, Komaba, Meguro-ku, Tokyo 153-8902, Japan.

Hidehito Honda (hitohonda.02@gmail.com)
Department of Psychology, Yasuda Women's University
6-13-1, Yasuhigashi, Asaminami-ku, Hiroshima 731-0153, Japan.

Toshihiko Matsuka (matsuka.toshihiko@gmail.com)
Department of Cognitive and Information Science, Chiba University
1-33, Yayoi-cho, Inage-ku, Chiba-shi, Chiba 263-8522, Japan.

Kazuhiro Ueda (ueda@gregorio.c.u-tokyo.ac.jp)
Graduate School of Arts and Sciences, The University of Tokyo
3-8-1, Komaba, Meguro-ku, Tokyo 153-8902, Japan.

Abstract

We examined whether probability weighting in decisions made under risk changed depending on the difference in evaluation methods. In particular, we focused on two methods, joint evaluation (JE) and separate evaluation (SE). We conducted a behavioral experiment and found that participants put more probability weight on small probability when using the SE method than when using JE, and that for large probabilities, the inverse was observed (i.e., participants put more weight in JE). We analyzed these results using a cognitive model and found that participants' subjective value of money does not change owing to differences in evaluation methods. However, beliefs concerning uncertain events shifted depending on evaluation methods, which led to the differences in probability weight. In this paper, we also discuss psychological mechanisms that produce different judgments or evaluations between SE and JE.

Keywords: probability weight; separate evaluation; joint evaluation; computer simulation; cognitive model of decision making

Introduction

It is well known that judgments change greatly depending on the difference in the evaluation methods. In the present study, we focused on one of the most studied topics, the difference between separate evaluation (hereafter, SE) and joint evaluation (hereafter, JE; Hsee, 1996; Hsee, Loewenstein, Blount, & Bazerman, 1999). Hsee (1996) showed that preference reversals by SE and JE occur in several contexts. Imagine people evaluating the worth of the following two dictionaries:

Dictionary A: Number of entries, 10,000
Dictionary B: Number of entries, 20,000 (cover is broken)

When they evaluate dictionaries A and B at the same time (i.e., JE), they may easily spot that there is a difference in the number of entries, and they may be attracted by the number of entries in Dictionary B. Thus, they may evaluate Dictionary B as having a higher price than Dictionary A. However, if people evaluate these dictionaries separately (i.e., SE), they may not notice the difference in the total number of entries (they may feel that either is enough), but they may mind the broken cover of Dictionary B. Thus, they may value Dictionary A as having a higher price than Dictionary B.

We predicted that shifts in evaluations by JE and SE might occur in the evaluation of probabilistic information. Previous studies on decisions under risk have shown that people put unique weights (i.e., non-linear weight) on probabilistic information in making decisions (Kahneman & Tversky, 1979; Tversky & Kahneman, 1992). For example, in decisions under risk, although people tend to be highly sensitive to differences in the end point (e.g., the difference between 0% and 10%, or differences between 90% and 100%), they tend to be less sensitive to differences in the middle degree (e.g., the difference between 30% and 40%). This finding suggests that sensitivity to differences is not constant. Recent studies have also showed that probability weighting is constructed through experimental procedures. In particular, different sets of probabilistic values presented in experimental tasks induce different probability weighting (e.g., Stewart, Reimers, & Harris, 2014; Walasek & Stewart, 2015).

Based on these previous findings, we predicted that differences in evaluation between JE and SE would change the probability weighting. If so, then what differences will be generated between JE and SE? In evaluating a certain probability value, people may refer to their probabilistic beliefs. For example, in evaluating 30% in a probabilistic event, people may refer to their probabilistic beliefs (i.e., how likely is the event to occur) and compare 30% with that belief. If they believe that the event usually occurs with high probability,

Money (Yen)	Prefer Sure Thing	Prefer Gamble
9500	✓	
9000	✓	
8500	✓	
8000	✓	
7500	✓	
7000	✓	
6500	✓	
6000	✓	
5500	✓	
5000	✓	
4500	✓	
4000	✓	
3500	✓	
3000	✓	
2500	✓	
2000	✓	
1500		✓
1000		✓
500		✓

Figure 1. A stimulus for measuring CE. The checks indicate the participant's selection.

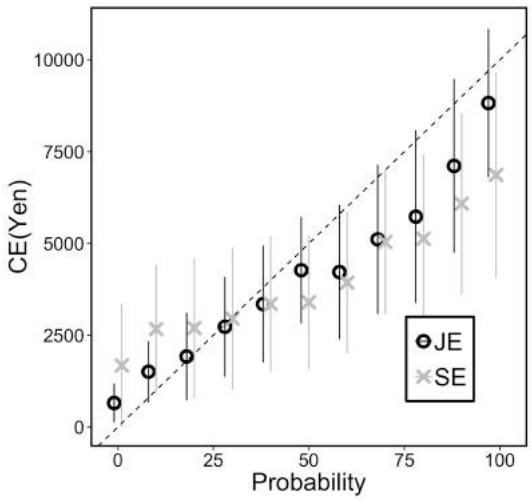


Figure 2. Mean CEs for 11 probabilities in the two groups. Error bars show standard deviation.

they may judge 30% as "not enough." In contrast, if they believe that the event usually occurs with low probability, they may judge 30% as "enough." Then, what is the nature of people's belief about probabilistic events? Stewart, Chater, and Brown (2006) showed that when people communicate probabilistic information using verbal expressions such as "likely" or "impossible," they tend to use highly extreme expressions such as "never" (representing 0%) or "always" (representing 100%). This finding suggests that people tend to easily imagine event occurrences or non-occurrences. In other words, people may refer to "black and white" probabilistic beliefs when evaluating probability.

We predicted that this would be true in evaluations using the SE method, but that it may not be true in evaluations using the JE method. In JE, people are presented with some probabilistic values at the same time, and they can compare these values. Thus, people may refer to probabilistic information in a continuous way. To the best of our knowledge, no previous studies have examined the above issue. In the present study, using a behavioral experiment and cognitive modeling, we examined whether probability weighting would shift depending on differences in the evaluation method between SE and JE. In the following section, we report the results of our behavioral experiment. We then report our analyses based on cognitive modeling.

Behavioral experiment

We examined whether probability weighting would shift due to using different evaluation methods—specifically, JE or SE—in a gambling task.

Method

Participants. We recruited 682 students as participants.

Task, stimulus, and procedure. We followed the method in Gonzalez and Wu (1999) to conduct the following task: Participants were asked to make a choice between a gamble that gets 10,000 yen (around $100) with certain probability p or sure gain. Figure 1 shows an example of the task. For example, participants choose one of two options: 100% chance of winning 5,000yen or a 30% chance of winning 10,000 yen. When they choose to a sure option, the monetary value of the option decreased: "you can get 4,500 yen." Amounts of sure gain ranged from 9,500 yen to 500 yen. As seen in Figure 1, the choice should change from a sure option to a gamble, and in the change point, we can assume that there is an amount of money to which a person is indifferent about getting a sure gain or playing the gamble (i.e., certainty equivalent, hereafter, CE). We assumed that CE was the median of the change point (in Figure 1, CE was assumed to be 1,750 yen). For the probability of the gambles, we set 11 values: 1%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, and 99%.

In the JE group (n = 47), participants were asked to answer the choices for the 11 probabilities. At first, they were instructed to answer the choices for 11 probabilities and then check their choices for each while answering the questions. In the SE group (n = 635), they were presented with one of

Table 1. Results of statistical analyses about the difference in CE between the two groups.

Probability	t-test		Effect size (d)
1	t (99) = 4.00	p = .001	0.80
10	t (106) = 4.23	p < .001	0.82
20	t (109) = 2.45	p = .174	0.47
30	t (118) = 0.68	p = .999	0.13
40	t (95) = 0.01	p = .999	0.00
50	t (93) = 2.58	p = .125	0.53
60	t (111) = 0.80	p = .999	0.15
70	t (101) = 0.19	p = .999	0.04
80	t (98) = 1.28	p = .999	0.26
90	t (99) = 2.13	p = .392	0.42
99	t (101) = 3.99	p = .001	0.79

Note. p-value was adjusted with *Bonferroni's* method.

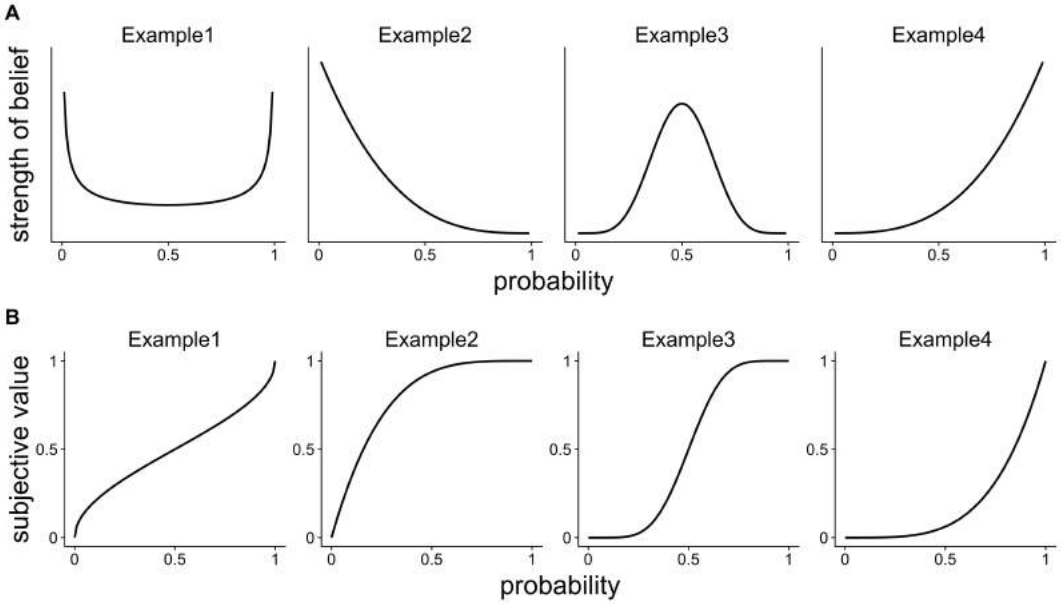


Figure 3. Summaries of DbBS. (a) Probabilistic belief regarding an uncertain event. (b) Subjective value in DbBS. This is represented with the cumulative distribution function (CDF) of the beta distribution.

the 11 probability gambles and answered the choices for the probability.

Results

Figure 2 shows mean CEs for 11 probabilities in the two groups. We found that the CEs differed between the two groups. In particular, in the low-probability range (1–20), the CE was higher for the SE group than in the JE group, suggesting that participants in the SE group applied more probability weight than those in the JE group. However, this trend reversed in the high-probability range (80–99), suggesting that participants in the JE group applied more probability weight than those in the SE group (as to the statistical analyses of CEs, see Table 1).

Taken together, the difference in evaluations between the JE and SE groups induced different probability weighting. In the following sections, we report the analyses of cognitive processes using a cognitive model.

Analyses of cognitive processes based on the cognitive model

Cognitive model of decision making: Decision by Belief Sampling (DbBS)

In this section, we introduce the decision model, called the *decision by belief-sampling model* (hereafter, DbBS; Honda, Matsuka, & Ueda, 2017). This model was proposed based on the *decision by sampling* model (*DbS*; Stewart, Chater, & Brown, 2006; Stewart, 2009). In the DbS model, subjective attribute values are constructed by a series of binary, ordinal comparisons to a sample of attribute values that reflect the immediate decision context and real-world distribution. The subjective value for a target is calculated as follows:

$$r = \frac{R-1}{N-1} \quad (1)$$

where r ($0 \leq r \leq 1$) denotes the subjective value for a target, and R denotes the rank of the target within the decision sample of N items. In this model, if the decision sample differs, r varies in the relationship between R and the decision sample. For example, imagine the subjective value for 60%. When decision samples are 10%, 20%, 30%, 30%, and 70%, the subjective value is r = (5-1)/(6-1) = 0.8. In contrast, in decision samples of 20%, 30%, 70%, 80%, and 90%, the subjective value is r = (3-1)/(6-1) = 0.4. That is, even when the target has the same attribute value, the subjective value varies depending on the decision samples. Previous studies have shown that this model can explain evaluations that vary depending on the samples (e.g., Stewart, Chater, Stott, & Reimers, 2003; Stewart, Reimers, & Harris, 2014).

DbBS is a model representing the subjective evaluation of probability. DbBS has two assumptions. First, the decision maker (DM) refers to the probabilistic belief samples in making decisions, and these samples represent the DM's probabilistic belief of an event's occurrence. For example, imagine the probable success rates of medical procedures for a serious disease and for appendicitis, respectively. Generally, people believe that the probability of success in treating a serious disease is low compared to the probable success of treating something simple, like appendicitis (Honda & Matsuka, 2014). We assume that the DMs refer to belief samples according to their probabilistic beliefs. We represent these beliefs using beta distributions (see the four examples

of DMs' subjective beliefs in Figure 3[a]). Example 1 represents the belief such that an event will occur or not (people refer to event occurrence and nonoccurrence). Likewise, in Examples 2 and 4, the DMs have the belief such that the event will happen with a low or high probability. Example 3 represents the belief that an event has a 50% chance of occurring. Thus, beta distributions can represent extensive kinds of beliefs about uncertain events. As a second assumption, a subjective value for a target is constructed by the comparison between the target value and the belief samples. Figure 3(b) shows subjective values calculated by the DbBS model. Given that beta distributions represent beliefs about uncertain events, subjective values correspond to values in the cumulative distribution functions (CDF) of beta distributions.

Using DbBS, we estimated the beliefs participants had in answering the gambling task in the behavioral experiment. In particular, we focused on the difference in beliefs produced between participants in the JE and SE groups.

Parameter estimation

In the gambling task of the behavioral experiment, when CE is y yen for the gamble that can win 10,000 yen with probability p, we assumed that the following relation:

$$v(y) = v(10000)w(p) \quad (2)$$

where v is a value function, represented with $v(x) = x^{\alpha}$, and $w(p)$ is a subjective weight for probability p. In this study, $w(p)$ is represented by subjective value according to DbBS.

With the above assumptions, we estimated parameters for value function (i.e., α) and two parameters of the beta distribution whose CDF best explains the choice patterns in gambling task.

In the JE group, we estimated the best parameters based on the choice patterns for the 11 probabilities. In this estimation, we conducted a grid search; for α, from 0.04 to 1 with increments of 0.04 (i.e., 25 values); and for each of the two parameters of beta distributions, from 0.01 to 1 with increments of 0.01 (i.e., 100 values). Thus, in total, from 250,000 combinations of parameters, we searched the combinations of parameters, which explained the observed choice pattern best for every participant in the JE group.

For participants in the SE group, it was impossible to estimate their beliefs on uncertainty because they answered choices only for one gamble. Thus, we constructed a hypothetical participant who responded to 11 gambles (i.e., gambles for 11 probabilities), by the "SE" method with the following procedure. CEs for the 11 probabilities were constructed based on the data of the behavioral experiment. In particular, the CE at one probability was randomly sampled from normal distribution. Here, mean and standard deviation were determined by the data of the behavioral experiment (i.e., the data demonstrated in Figure 2). In these random samplings for 11 probabilities, we assumed that the hypothetical participant showed consistent choice patterns such that when $p_1 < p_2$, CE for p_1 (CE_1) and p_2 (CE_2) always satisfied $CE_{p1} \leqq CE_{p2}$. Thus, we estimated the response for 11 gambles using the SE method by the "identical person." With these procedures, we constructed 1,000 hypothetical participants. For the data of the hypothetical participants, we estimate the best parameters for value function and beta distribution using a grid search as we did for the JE group.

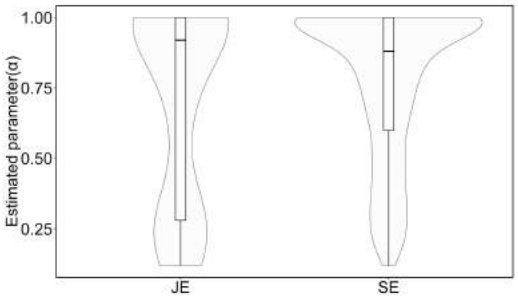


Figure 4. Distribution of estimated parameter for value function (α)

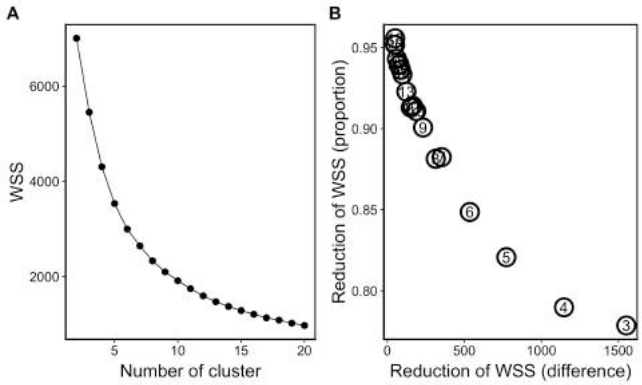


Figure 5. Results of clustering analysis. (a) Scree plot for within-cluster sum of squares (WSS) in K-means clustering. (b) Relationship between reduction of WSS (difference) and that in proportion. The number in the circle (e.g., n) indicates the reductions in WSS when the number of clusters increased from (n-1) to n.

In our parameter estimation, we evaluated the model fit using R^2. In the following analyses, we used the data wherein the model showed a good fit. Here, we set the criterion of "goodness" as $R^2 > 0.5$ (44 out of 47 data in the JE group and 787 out of 1000 data in the SE group satisfied this criterion).

Results of parameter estimations

Value function

Figure 4 shows the distribution of the estimated parameter of α for the JE and SE groups. As shown in the figure, the distributions were similar between the two group, and there was no significant difference ($w = 16328$, $p = .516$, Wilcoxon rank sum test). Thus, this result suggests that the different evaluations did not affect valuation of money.

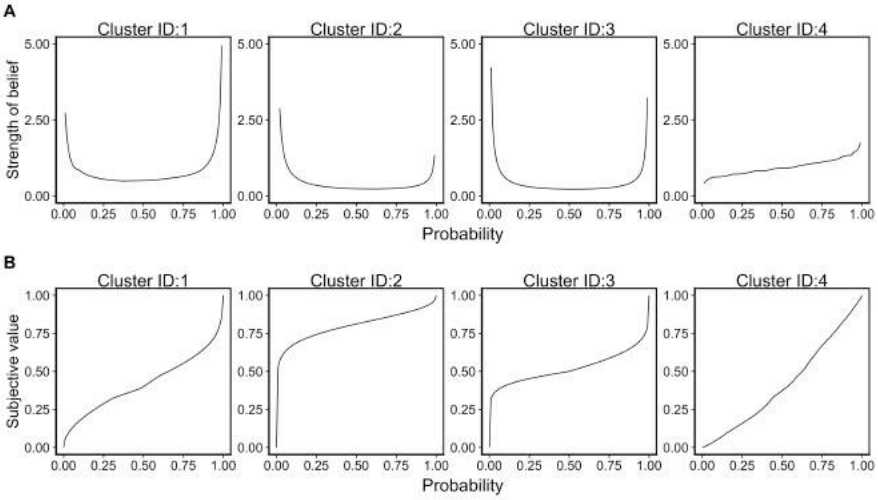


Figure 6. Median of strength of probabilistic belief (A) and subjective value (B) for the four clusters.

Table 2. Proportion of data categorized into the four clusters.

Group	Cluster 1	Cluster 2	Cluster 3	Cluster 4
JE	0.295	0.364	0.068	0.273
SE	0.287	0.159	0.475	0.079

Beliefs on uncertainty

Next, we examined participants' beliefs (i.e., estimated beta distribution) in detail with the following procedure. First, we clustered beliefs (i.e., shape of beta distribution) using probability densities. For the 831 data sets, the patterns of probability densities for 99 probabilities (1%, 2%, 3%,..., 97%, 98%, 99%) were clustered using the K-means method. We determined the number of clusters by considering the tradeoff between parsimony (i.e., as least clusters possible) and informativeness (i.e., as many clusters as required). Here, we calculated the within-cluster sum of squares (WSS) for each cluster and examined reductions in the WSS in terms of the difference and proportion of increasing numbers of clusters. Figure 5 shows the scree plot (a) and the relationship between the reduction in WSS in difference and proportion (b). We adopted four clusters based on their parsimony and informativeness.

We examined features of each cluster: median strengths of belief and median subjective values for 99 probabilities for each cluster. Figure 6 shows these results. The four clusters can be summarized as follows: For the clusters 1, 2, and 3, the probabilistic belief is "black and white" (i.e., deterministic). That suggests that a person refers to "winning" and "losing" gambles. The differences among the three clusters lie in whether a person is more optimistic (i.e., the belief in "winning" is stronger than that for "losing," Cluster 1), more pessimistic (i.e., the belief in "losing" is stronger than that for "winning," Cluster 2), or neutral (i.e., the belief in "losing" is as strong as that for "winning," Cluster 3). Cluster 4 has a different feature: the strength of belief is almost constant, suggesting that a person believes that the probability of winning gamble takes any probability (i.e., referring wide range of probability).

Then, we examined the proportions of data categorized into the four clusters for the two evaluation methods. Table 1 shows those results. Most data were categorized into Clusters 1, 2, or 3, which represented "black and white" belief. Those findings were generally consistent with the previous findings in Stewart et al. (2006) showing that people tend to often use extreme probabilistic expressions representing 0% and 100%. However, the most notable point was the proportion that was categorized into Cluster 4: more data from the JE group were categorized into Cluster 4 than from the SE group (p <.001, Fisher's exact test), suggesting that the participants (though "hypothetical participants") in SE referred to probabilistic information in a continuous way. These findings corroborated our prediction.

Discussion

In this study, we examined whether probability weighting shifts according to which evaluation method, JE or SE, was used in a gambling task. We found that the different evaluation methods induced different weighting. Furthermore, we analyzed our results using a cognitive model. The analyses indicated that differences in probability weighting for JE and SE were derived from a difference in probabilistic beliefs that people refer to in making decisions.

Previous studies have discussed changed preferences based on evaluation methods (JE, SE) but there has been little discussion about the process of making decisions under risk. One reason may be the difficulty of examining decision processes since researchers can obtain only one datum for each participant in an SE group, making model-based analysis highly difficult. In the present study, we proposed a new method to overcome such difficulties by constructing hypothetical participants using behavioral data. We believe that the proposed method makes a substantial contribution

that helps clarify the difference in cognitive processes between JE and SE methods.

Acknowledgments

This study was supported by JSPS KAKENHI Grant Number 18H03501 for the second author and JP16H01725 for the last author.

References

Gonzalez, R., & Wu, G. (1999). On the shape of the probability weighting function. *Cognitive Psychology*, *38*(1), 129–166.

Honda, H., & Matsuka, T. (2014). On the role of rarity information in speakers' choice of frame. *Memory and Cognition, 42*(5), 768–779.

Honda, H., Matsuka, T., & Ueda, K. (2017). Decisions based on verbal probabilities: Decision bias or decision by belief sampling? In G. Gunzelmann, A. Howes, T. Tenbrink, & E. J. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 557–562). Austin, TX: Cognitive Science Society.

Hsee, C. K. (1996). The evaluability hypothesis: An explanation for preference reversals between joint and separate evaluations of alternatives. *Organizational Behavior and Human Decision Processes, 67*(3), 247–257.

Hsee, C. K., Loewenstein, G. F., Blount, S., & Bazerman, M. H. (1999). Preference reversals between joint and separate evaluations of options: A review and theoretical analysis. *Psychological Bulletin*, *125*(5), 576–590.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*, 263–292.

Stewart, N. (2009). Decision by sampling: The role of the decision environment in risky choice. *Quarterly Journal of Experimental Psychology*, *62*(6), 1041–1062.

Stewart, N., Chater, N., & Brown, G. D. A. (2006). Decision by sampling. *Cognitive Psychology, 53*(1), 1–26.

Stewart, N., Chater, N., Stott, H. P., & Reimers, S. (2003). Prospect relativity: How choice options influence decision under risk. *Journal of Experimental Psychology: General, 132*(1), 23–46.

Stewart, N., Reimers, S., & Harris, A. J. L. (2014). On the origin of utility, weighting, and discounting functions: How they get their shapes and how to change their shapes. *Management Science, 61*(3), 687–705.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5(4), 297–323.

Walasek, L., & Stewart, N. (2015). How to make loss aversion disappear and reverse: Tests of the decision by sampling origin of loss aversion. *Journal of Experimental Psychology: General, 144*(1), 7–11.

A proverb is worth a thousand words: Learning to associate images with proverbs

Gözde Özbal[†], **Daniele Pighin**[‡], **Carlo Strapparava**[†]
[†] FBK-Irst - Trento, Italy
[‡] Google Zurich, Switzerland
gozbalde@gmail.com, biondo@google.com, strappa@fbk.eu

Abstract

We describe a system that can associate images with English proverbs. We start from a corpus of proverbs, harvest related images from the web and use this data to train two variants of a convolutional neural network. We then collect a small set of annotations, and use these to combine the outputs of the two networks into a single prediction for each input image. We carry out feature selection experiments on a set of features derived from the images and from the predicted proverbs, and demonstrate that the metaphoricity of the proverbs plays a significant role in classification accuracy. An empirical evaluation with human raters confirms the system's ability to abstract from the raw bits in the images and to learn meaningful, non-trivial associations.

Introduction

Meaningful associations between visual information and short texts are a staple of effective and powerful communication. Instances of this form of communication can be found almost anywhere: on t-shirts, covers of books, records and magazines, social media posts, and ad campaigns, just to name a few. The empirical evidence, in agreement with our common sense and everyday experience, shows that meaningful image-text associations are very good predictors of the success of an online post (Hessel et al., 2017). To add to the value of an image, the caption must convey some information that is not already obvious. For example, consider two possible captions for the image in Figure 1. A purely descriptive caption like (a) is very accurate, but it does not add value to the image. By associating it with a proverb, a caption like (b) radically changes our perception of the image, from a collection of visual elements to an abstract representation of a familiar feeling (i.e., envy).

Recent advances in neural networks and computer vision have made it possible to generate high-quality descriptive captions such as (a) in Figure 1 automatically (Vinyals et al., 2017). Such captions are certainly remarkable from an artificial vision stand point, and very useful when it comes to organizing and accessing large databases of images. However, they do not make an image more memorable or compelling.

In this paper, we focus on the task of producing captions like (b), in which an image is associated with a memorable expression that emphasizes non trivial, suggestive aspects of the image. In particular, we leverage an existing corpus of English proverbs (Özbal et al., 2016) to learn a model that can associate any image to the most appropriate proverb in the repository. The resulting system can have many potential applications, e.g.: suggesting evocative and compelling taglines when posting an image on social media; proposing headlines for news, based on photos of an event; selecting the visual content of ad campaigns so as to evoke specific moods.

Figure 1: Different captions affect our perception of the same image: (a) "A half-barren, half-green field." (b) "The grass is always greener on the other side."

To the best of our knowledge, this is the first attempt to prove that existing models for object recognition can be successfully adapted to associate images to linguistically complex and semantically rich data such as proverbs. We demonstrate that the existing networks have enough capacity to abstract away from the mere graphical content of an image and learn original and surprising associations.

Note that we do not claim that our model can understand the language used in the proverbs. This is a complex problem per se, given the non-literal nature of most proverbs. In addition, from the point of view of our model a proverb is just a class label. Instead, we observe that by using the proverbs to retrieve related images allows the model to learn that some combinations of objects appearing in the pictures are relevant with respect to the meanings commonly attached to the proverbs, also when their meaning is far from literal.

The approach that we propose is simple and scalable, it relies on the availability of large amounts of noisy data and can be tuned using minimal supervision.

Related work

A growing body of literature, including Yamaguchi et al. (2014) and Gelli et al. (2015), has shown that image features do not contribute as much as textual features to the social popularity of multimedia content. In particular, Hessel et al. (2017) study the effect of visual and textual features on the popularity of Internet posts, and conclude that the right combination of visual and textual features plays a very important role. They also note that the cleverness of the accompanying captions can result in a very different response to pictures of very similar subjects, and make a less attractive subject more

popular than a better subject with a less remarkable caption.

Concerning the automatic captioning of images, Hall et al. (2015) propose to automatically generate natural language captions that describe the geographical context of georeferenced photos, such as "Rijksmuseum photographed at 2.15 pm at the corner of Stadhouderskade and Museumstraat near Spiegelgracht in Amsterdam, Netherlands.". Chen et al. (2015) present a large dataset consisting of groups of images observed with the same caption. The associative structure of the data is exploited to retrieve captions for query images. The retrieved captions can be further classified to select the more creative ones. Vinyals et al. (2017) present a generative model based on a deep recurrent architecture that can generate natural sentences describing an image. The model builds on recent advances in machine translation and computer vision. Szegedy et al. (2016) describe Inception-V3, a convolutional neural network that can be used to detect the main objects that appear in an image with very high accuracy.

Pertaining to the association of content with familiar expressions, (Tan et al., 2016) use neural networks to recommend quotes in writing and to make statements more compelling. They point out how computational methods can help writers select the most appropriate quote for a given context from a large repository of alternatives.

Regarding the appropriateness of proverbs as image captions, B. Mieder and Mieder (1977) analyze the reasons behind the common usage of proverbs in advertisement. Proverbs have a "familiar ring" that adds reliability, trustworthyness and a sense of timelessness to a brand or product. More recently, Qing-fang (2004) observes that proverbs are especially suitable for advertisement as they are short and concise, and they are associated with wisdom and moral guidance. To say it in the words of the author, "one proverb may say more than a thousand words".

Associating proverbs to images

In this section, we describe the architecture of a system that, given an image and a set of proverbs, decides whether the image is evocative of one of the proverbs. In particular, we use PROMETHEUS (Özbal et al., 2016) as a proverb repository, but a different set of proverbs or other types of memorable expressions (such as slogans or quotations) could be used in alternative. The resource consists of 1,054 proverbs, grouped into categories (such as "love and hate" or "fate") and annotated with metaphors at the word and sentence level. More than in other genres, such as news, fiction and essays, in proverbs metaphors can resolve a significant amount of the figurative meaning (Faycel, 2012). The richness of proverbs in terms of metaphors and their pervasiveness in all cultures makes them especially suitable for being used as evocative captions (W. Mieder, 1978).

We first use the proverbs to retrieve a large set of noisy data from the web. Then, we use this data to train two convolutional neural networks to associate proverbs to images. The two classifiers use the same architecture, but one is trained to directly associate images to proverbs, while the other builds associations between the objects that it recognizes in the images and the proverbs. Then, we use a small sample of the predictions of the two models to crowd source golden image-proverb associations. Finally, we use the noisy data and the golden labels to combine the output of the two classifiers into a unified model that decides whether it should select the proverb suggested by any of the two classifiers.

Noisy data collection

For each proverb in PROMETHEUS, we used the Flickr API to retrieve a set of candidate images. We included the full text of the proverb as part of the query string, forcing the API to only return images that mention the complete proverb in their title, description or tags. In our experiments we focus on the 98 proverbs for which we could retrieve at least 500 images.

To keep the data set reasonably balanced, we also limit the maximum number of images retrieved for each proverb to 1,000. The resulting data set consists of 83,895 images, each of which is associated with exactly one of 98 distinct proverbs. We then randomly split the data into a training (80,000 images) and a development (3,895 images) set. For the purpose of training and testing the classifiers, we used Flickr API to download 150×150 pixel versions of the images. These are obtained by cropping to a square around the main subject and then scaling to the final size, thus preventing warping or distortions of the elements of the images. As we reckon that color plays an important role with respect to the mood and perceived message of a picture, we did not convert the images to black and white.

Image classification

In this section, we describe the training of two classifiers that, given an image, predict the most likely proverb association. Both classifiers are based on Inception-V3, a convolutional neural network which has been shown to be very accurate in image classification tasks with a large number (1,000) of output classes (Szegedy et al., 2016). For each input image, the model outputs a probability distribution over all the output classes. The predicted label is the class with the highest probability density. For all our experiments, we use the Inception-V3 implementation included in the TensorFlow-Slim image classification model library[1].

Inception from scratch (I-FS) The first model is trained to establish a direct association between the visual clues present in the image and the output proverbs. It is an Inception-V3 network trained *from scratch* (I-FS) on the available training data. We use all the default settings of Slim's Inception implementation and we select the model after 669,923 iterations.

Inception fine-tuned (I-FT) We fine-tune the model starting from the Inception-V3 model[2] trained by Szegedy et al. (2016). This model was trained from the 1.2 million images of the 2012 ImageNet Large Scale Visual Recognition

[1] https://goo.gl/W5ZdQ4
[2] https://goo.gl/nrsdGG


(a)


(b)

Figure 2: Examples of reasonable predictions that differ from the noisy label. (a) Label: "Look before you leap". I-FS: "Rules are made to be broken.". (b) Label: "Beggars can't be choosers.". I-FS and I-FT: "Time and tide wait for no man".

Challenge (ILSVRC-12) (Russakovsky et al., 2015). We refer to the resulting proverb classifier as I-FT, for *Inception fine-tuned*. As the proverb classification task has a different number of output classes from ImageNet (i.e., 98 vs. 1,000), we do not restore the weights of the final layer of the network[3]. In addition, we only allow the weights of the classification layer to be updated during fine-tuning. In doing so, we expect the classifier to retain the object recognition capabilities of the internal layers of the pre-trained model and to establish meaningful association between the target proverb and the dominant objects in an image. Concerning I-FT, we select the model after 1,955,892 iterations[4].

Evaluation of I-FS and I-FT We measured the performance of the two classifiers on the 3,895 images in the development split of the noisy data. I-FT's recall is consistently higher than I-FS's (Recall@1: 0.20 vs. 0.15; Recall@5: 0.39 vs. 0.28). This is an expected result, as the inner layers of I-FT encode classification clues learned from a very large data set. While recall is relatively low for both classifiers, we should consider that each image can possibly evoke more than one proverb, whereas in our data set we only have one label for each image. Therefore, we regard these figures as very conservative lower bounds. For example, Figure 2 shows two images for which the decisions of the classifiers are quite reasonable, yet they do not agree with the noisy label.

It is also important to observe that the two classifiers learn very different models, as exemplified in Table 1. I-FS ans I-FT output a different label in the large majority of the cases (85%), and 27% of the times at least one of the two classifiers can reconstruct the correct association according to the noisy labels. In the next sections, we will explain how we leverage the different "personalities" of the two classifiers and combine them into a unified model that can predict a golden (i.e., human validated) proverb with an accuracy of 74.59%.

[3] https://goo.gl/tfHxzs

[4] We let both I-FS and I-FT learn for ≈1 week. Then, among the last 5 checkpoints, we selected the one having the smallest loss on the training data. Since there is no previous work to compare against, we are not trying to maximize accuracy at all costs. Instead, we aim to demonstrate that our pipeline produces results that are adequate for a range of user facing applications, as those mentioned in the introduction.

Statistics on development data	Count	%
Same prediction	579	14.87
Same prediction, both incorrect	226	5.80
Same predictions, both correct	353	9.06
Different predictions	3,316	85.13
Different predictions, both incorrect	2,604	66.85
I-FS correct, I-FT not correct	276	7.09
I-FT correct, I-FS not correct	436	11.19
I-FT or I-FS prediction correct	1,065	27.34

Table 1: Comparison of I-FS and I-FT. Correct and incorrect counts refer to the noisy development labels.

Gold standard collection

In the previous section, we observed that there is a number of cases in which the output of I-FS or I-FT are more suitable captions for a given image than its noisy label. To quantify this phenomenon, we set-up a crowd-sourced annotation in which we showed the raters an image and four proverbs, and asked the raters to select the most appropriate caption. To maximize the utility of the annotation, we included only the cases in which both models disagree with the noisy label. We decided to crowd-source the annotation of 500 images on the Figure-Eight platform[5].

We first included all the 226 development examples for which the two models predict the same label and the prediction is incorrect (2nd row in Table 1). We refer to these as *Type1* examples. We regard these examples as especially relevant, as we have seen before that the two models do not agree very often. Our hypothesis is that, in many such cases, the models are actually converging to a meaningful interpretation. Then, we added 274 randomly sampled images for which the predictions of the two models differ, and both predictions differ from the noisy label (*Type2*).

For Type1 examples, the raters could choose among: (1) the noisy label, (2) the proverb selected by I-FS and I-FT, and (3 and 4) two random proverbs. For Type2 examples, the raters could choose among: (1) the noisy label, (2) I-FS prediction, (3) I-FT prediction, and (4) a random proverb. In both cases, the random proverbs were selected among the 98 proverbs used to train the models. The raters were instructed to select all the relevant associations, and they also had the option to mark none of the proposed alternatives as relevant.

Due to the inherent subjectivity of the task, we decided to elicit 10 judgments for each image, for a total of 5,000 ratings. The agreement on the ratings, as reported by Figure-Eight, is 64.47%. The aggregated results of the annotation based on majority voting[6] are shown in Table 2. We can see that, overall, raters tend to prefer the decisions of I-FT over the noisy label (27.21% vs. 24.87%), and the noisy label over I-FS (20.70%). It is quite remarkable that I-FT's predictions are rated to be more accurate than the data on

[5] https://www.figure-eight.com/

[6] Even though raters could select multiple options, the majority decision has never included more than one.

	Times selected (%)		
Selected label	Overall	Type1	Type2
Noisy label	24.87	17.85	33.21
Random	3.84	3.69	4.01
None	23.37	17.54	30.29
I-FS	20.70	30.46	9.12
I-FT	27.21	30.46	23.36
I-FS or I-FT	31.39	30.46	32.48

Table 2: Results of the crowd-sourced annotation.

Label	Annotated data	Noisy data	Total
Either	99	353	452
None	312	-	312
I-FS	25	276	301
I-FT	64	436	500
Total	500	1,065	1,565

Table 3: Data distribution of the combined classifier. Note that we only annotated 500 examples out of 2,830 for which both I-FS and I-FT fail to predict the noisy label. As a consequence, 2,330 development examples are not included in this experiment.

which the model has been trained. When the two classifiers make the same decision (Type1), there is a marked preference of the raters for the predicted proverb over the noisy label (30.46% vs. 17.85%), whereas when the two classifiers do not agree (Type2) the raters generally find the noisy label preferable, even though the cases in which either I-FS or I-FT are chosen are almost the same with the noisy label (32.48% vs. 33.21%). Even though I-FS is not as accurate as I-FT to predict the noisy labels, there is a non negligible number of cases in which its decision is considered to be appropriate by the raters, and when the decisions of the two classifiers differ (Type2), I-FS selects a good option in 9.12% of the cases. There are very few cases (3.84% overall) in which a random proverb is preferred to any of the more principled alternatives, whereas there is a very significant number of cases (23.37% overall) in which none of the proposed alternatives, including the noisy label, is considered to be good.

Model combination

In this section, we describe a classifier that, given an image and the output of I-FS and I-FT, classifies the image into one of the following four classes: (a) *I*-FS, if the prediction of I-FS should be selected; (b) *I*-FT, if I-FT should be preferred instead; (c) *None*, for the cases in which neither of the two classifiers predicted an appropriate class; and (d) *Either*, if both the predictions of I-FS and I-FT are appropriate. We introduce the last class *Either* specifically to model the cases in which I-FS and I-FT output the same prediction.

Data set All the annotated examples for which the raters did not select either I-FS or I-FT predictions were mapped to the *None* class. These are all the images annotated as "Noisy label", "None" or "Random". Type1 examples where the prediction of the models was preferred by the raters were mapped to *Either*, whereas Type2 examples where I-FS or I-FT were preferred were mapped to the corresponding label. The distribution of the labels of the annotated data is summarized on the left side of Table 3. By construction, the annotated data contains only cases in which I-FS's and I-FT's predictions differ from the noisy label, and the *None* label is significantly over-represented. In order to come up with a more balanced data set, we also include the non-annotated examples in which either classifier agreed with the noisy label. If both classifiers agree with the noisy label, then we map the example to the *Either* label. If only I-FS (or I-FT) agrees, then we map the example to the I-FS (or I-FT) class. The column labeled "Noisy data" in Table 3 shows the distribution of the data added in this fashion. We regard these examples as highly accurate, given the low chance of random agreement between the noisy label and the classifiers (the output space of I-FS and I-FT consists of 98 proverbs).

Features From each example we extract 12 simple features,which we group into six sets to simplify the feature selection experiments. The set labeled "Base" (***b***) only accounts for the decisions of I-FT and I-FS. To avoid overfitting, we only include the prediction scores, and not the actual predicted classes. The set labeled "Metaphoricity" (***m***) makes use of the proverb-level metaphoricity annotations in PROMETHEUS. The metaphoricity can have one of three values: 0 (literal); 1 (slightly metaphorical); 2 (highly metaphorical). We expect proverbs which are metaphorical to be a good fit for a broader set of images. The feature set "Inception" (***i***) encodes the highest prediction score of the Inception-V3 model for the image. The intuition here is that a high prediction score, regardless of the class, means that the Inception-V3 model is confident that it can recognize a known object in the image. We use this measure as a proxy for the "concreteness" of the image, as a counterpart for the data encoded by *m*. The set "Category similarity" (***cs***) attempts to measure the compatibility between the category of the proverb (e.g., "love and hate" or "fate") and the object recognized in the picture by the Inception-V3 model. We use the DISCO (Kolb, 2009) library together with the provided English word space[7] and encode as feature the maximum cosine similarity between any synonym in the synset predicted by Inception-V3 and any content word in the predicted proverb categories. The feature set "Proverb similarity" (***ps***) is conceptually very similar, but we use the lemmas in the predicted proverb instead of its category. Finally the feature set "Difference" (***d***) encodes the difference in magnitude between the values of the feature in *b* and *m* computed for I-FS and I-FT. These features are meant to help the classifier reason more comparatively about I-FS and I-FT predictions.

Set-up To make the most of the available training data, we evaluate the combination of the two models in a leave-one-out setting, i.e., a cross-fold where the number of folds equals

[7] https://goo.gl/Rc45PW

F1	b	b+ m†‡	b+ d†	b+ ps	b+ i	b+ cs	b,m+ d†‡	b,m+ i	b,m+ ps	b,m+ cs
Macro	53.81	59.90	54.40	54.24	54.06	53.99	60.25	58.00	56.69	56.11
Micro	66.52	68.56	67.09	66.90	66.84	66.77	68.88	67.92	67.22	67.03

Table 4: Feature ablation results for the best learning algorithm. †: Significantly better than *b*. ‡: Significantly better than *b,d*. The difference between *b,m,d* and *b,m* is not significant.

the number of test examples. Please note that none of the images in the test set of the combined classifier is included in the traning of I-FS or I-FT. We compare different groupings of feature sets (always including *b*). As a learning algorithm, we use an SVM with a polynomial kernel of degree 2. We use the implementations provided by SciKit-Learn (Pedregosa et al., 2011). To compare the different feature combinations, we use McNemar's significance test (McNemar, 1947) with a 95% confidence interval ($p < 0.05$).

Results In Table 4 we report the detailed results of the feature inclusion experiments. The set of base features *b* alone achieves a micro F1 measure of 66.52. If we try to add another set of features on top of *b*, only *b,m* and *b,d* achieve a significant improvement, with *b,m* being significantly more accurate than *b,d* (68.56 vs. 67.09). If we try to add another feature set on top of *b,m*, we observe that only *b,m,d* achieves a higher accuracy (i.e., 68.88 vs. 68.56), even though the improvement is not significant. Adding any other feature set yields a negative contribution (micro F1 < 68).

As a further comparison between *b,m* and *b,d,m*, Table 5 shows the difference between the confusion matrices of the two configurations. We can observe that the error distribution of the two models is very similar, with the former being slightly more accurate on the examples labeled I-FS and *Either*, and the latter on *None* and I-FT. Interestingly, both models make very few mistakes on examples labeled *Either*, confirming that the convergence of I-FS and I-FT predictions is a strong signal of the accuracy of the predicted proverb. The error distribution also reflects the fact that I-FT, being a more accurate predictor than I-FS, is more represented in the training data. In fact, there are many more examples labeled I-FS which are predicted as I-FT than the other way round. For the same reason, the model also tends to predict I-FT when the actual label is *None*. All in all, this error analysis suggests that the best way to improve the classifier might be to introduce more data points for the classes *None* and I-FS, which are under-represented in the data (see Table 3).

From all the evidence above, we can conclude that the information about the metaphoricity of the predicted proverb provides very useful clues to the learning algorithm.[8] Contrary to our expectations, the features that account for the similarity between the objects in the pictures and the predicted proverbs (*i*, *ps* and *cs*) do not improve the classification accuracy.

Label	Predicted label: None		I-FS		I-FT		Either	
None	38	(**41**)	33	(37)	151	(148)	90	(86)
I-FS	0	(0)	**167**	(161)	134	(140)	0	(0)
I-FT	3	(1)	69	(61)	428	(**438**)	0	(0)
Either	12	(14)	0	(0)	0	(0)	**440**	(438)

Table 5: Confusion matrices for the combined model with feature groups *b,m* and *b,m.d* (in parentheses).

Finally, in Figure 3 we show 10 examples of system outputs (for the configuration using feature sets *b,m,d*), which we believe are quite representative of what the model has learned. Not all outputs are correct according to the golden labels, and we invite the readers to figure out which examples are correct and which are not before continuing reading (the answer is at the end of the paragraph). Looking at the outputs, we can see that in some cases (e.g., (d) and (i)) the associations are quite literal (hay, detergents). In other cases, the association is less obvious. These are the most interesting cases, in which the predictions showcase the ability of the model to abstract away from concrete objects, or to reproduce the cultural biases observed in the training data . In (a) there is a sense of frugality that is resolved to "every little helps". Concerning (b), in the training data "slow but sure" is very often associated with religious symbols, churches in particular. In (f), the model associates the flooded land with "storm" and the ships with "port". In (g), the model recognized the quietness of situation and the golden tones of the scenes. Concerning (h), a crowded school of fish evokes the association with "first come, first served". According to the golden labels, examples (a) to (e) are classified correctly, whereas the ones from (f) to (j) are incorrect. Nevertheless, for the applications that we have in mind all examples seem appropriate. This fact can be confirmed by restricting the evaluation to the examples annotated by the raters and by considering all the proverbs that have been selected by at least one human rater as good predictions. Under these conditions, the model selects an appropriate proverb in 74.59% of the cases.

Copyright and credits

We are extremely grateful to the authors of the images included in the paper for releasing their images under a permissive licensing scheme or for explicitly allowing us to use their pictures. This section lists all the images used in the paper, including their author, licensing scheme and Flickr URL. All

[8] We have observed the same pattern also using different learning algorithms (RBF, LR), but here we omit these results due to space limitations.

(a) Every little helps.

(b) Slow but sure.

(c) Two heads are better than one.

(d) Make hay while the sun shines.

(e) Like father, like son.

(f) Any port in a storm.

(g) Silence is golden.

(h) First come, first served.

(i) Cleanliness is next to godliness.

(j) Seeing is believing.

Figure 3: Example outputs of the combined model. Five outputs differ from the corresponding noisy label. Can you tell which ones?

the listed URLs were active at the time of submission.

Figure 1. Author: Flickr user "Dano". License: CC BY 2.0[9]. Source: https://www.flickr.com/photos/mukluk/249464230.

Figure 2(a). Author: Flickr user "Gavin Clarke". License: CC BY-NC 2.0[10]. Source: https://flickr.com/photos/70824176@N00/4460439903.

Figure 2(b). Author: Jason Swain. All rights reserved. Used under permission by the author. Source: https://flickr.com/photos/24424426@N00/13058126593.

Figure 3(a). Author: Flickr User "Neil Moralee". License: CC BY-NC-ND 2.0 [11]. Source: https://flickr.com/photos/62586117@N05/21178964709.

Figure 3(b). Author: Flickr User "Cathedrals and Churches". License: CC BY 2.0[9]. Source: https://www.flickr.com/photos/eltb/7246837670/.

Figure 3(c). Author: Flickr User "Peter Trimming". License: CC BY 2.0[9]. Source: https://www.flickr.com/photos/55426027@N03/8730055756.

Figure 3(d). Author: Flickr User "Raymond Barlow". License: CC BY-NC-SA 2.0 [12]. Source: https://flickr.com/photos/62673829@N00/2631618525.

Figure 3(e). © Jay Heymans. All rights reserved. Used under permission by the author. Source: https://www.flickr.com/photos/7830239@N06/12234997804.

Figure 3(f). © Ian Huges. All rights reserved. Used under permission by the author. Source: https://flickr.com/photos/36463157@N08/3818175700.

Figure 3(g). Author: Flickr User "Geraint Rowland". License: CC BY-NC 2.0[10]. Source: https://flickr.com/photos/33909206@N04/23407737789.

Figure 3(h). Author: Flickr user "Steven Harris". License: CC BY-NC 2.0[10]. Source: https://flickr.com/photos/90288178@N00/4060998399.

Figure 3(i). © Melissa Jones. All rights reserved. Used under permission by the author. Source: https://www.flickr.com/photos/msjones166/5511643604.

Figure 3(j). Author: Flickr user "TheoJunior". License: CC BY-NC-SA 2.0[12]. Source: https://flickr.com/photos/88013568@N00/3252673888.

Conclusion and future work

In this paper, we presented a model that can associate images to proverbs. It combines two variants of a high-performance convolutional neural network in a simple voting scheme, it is easily scalable and it requires very minimal supervision. By leveraging high volumes of noisy training data, the model can learn compelling associations at surprising levels of abstraction, such as "Misery loves company." for a sweaty bunch of skaters. To our best knowledge, we are the first ones to

[9] https://creativecommons.org/licenses/by/2.0/

[10] https://creativecommons.org/licenses/by-nc/2.0/

[11] https://creativecommons.org/licenses/by-nc-nd/2.0/

[12] https://creativecommons.org/licenses/by-nc-sa/2.0/

use existing object recognition models to associate images to semantically rich, non-descriptive captions such as proverbs.

Our approach can easily be extended to cover more proverbs as well as other kinds of memorable and familiar expressions, such as slogans, citations or titles of famous works of art that have already been the focus of previous work on creative language generation (Gatti, Özbal, Guerini, Stock, & Strapparava, 2015; Özbal, Pighin, & Strapparava, 2013; Stock, Strapparava, & Valitutti, 2007). We have shown that knowledge about the metaphoricity degree of proverbs plays a significant role with respect to the classification accuracy. While PROMETHEUS already provides this information, this might not be the case for other sources of familiar expressions. On the other hand, it should be possible to automatically assess metaphoricity by leveraging recent state-of-the-art advancements in the field of metaphor detection (Özbal, Strapparava, Tekiroglu, & Pighin, 2016; Veale, Shutova, & Klebanov, 2016). In addition, we would like to generate more captivating captions, by injecting humor into the predicted proverbs through incongruity (Raskin, 1979) or other rhetorical devices. As Veale (2012) suggests, linguistic creativity can be utilized to "re-invent and re-imagine the familiar, so that everything old can be made new again".

References

Chen, J., Kuznetsova, P., Warren, D., & Choi, Y. (2015). Déja image-captions: A corpus of expressive descriptions in repetition. In *Proceedings of NAACL-HLT'15.*

Faycel, D. (2012). Food Metaphors in Tunisian Arabic Proverbs. *Rice Working Papers in Linguistics*, *3*.

Gatti, L., Özbal, G., Guerini, M., Stock, O., & Strapparava, C. (2015). Slogans are not Forever: Adapting Linguistic Expressions to the News. In *Proceedings of IJCAI'15.*

Gelli, F., Uricchio, T., Bertini, M., Del Bimbo, A., & Chang, S.-F. (2015). Image Popularity Prediction in Social Media Using Sentiment and Context Features. In *Proceedings of ICM'15.*

Hall, M. M., Jones, C. B., & Smart, P. (2015). Spatial Natural Language Generation for Location Description in Photo Captions. In *Proceedings of COSIT'15.*

Hessel, J., Lee, L., & Mimno, D. (2017). Cats and Captions vs. Creators and the Clock: Comparing Multimodal Content to Context in Predicting Relative Popularity. In *Proceedings of WWW'17.*

Kolb, P. (2009). Experiments on the Difference between Semantic Similarity and Relatedness. In K. Jokinen & E. Bick (Eds.), *Proceedings of NODALIDA'09.*

McNemar, Q. (1947, Jun 01). Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages. *Psychometrika*, *12*(2), 153–157.

Mieder, B., & Mieder, W. (1977). Tradition and Innovation: Proverbs in Advertising. *The Journal of Popular Culture*, *11*(2), 308–319.

Mieder, W. (1978). Proverbial Slogans are the Name of the Game. *Kentucky Folklore Record*, *24*(2), 49.

Özbal, G., Pighin, D., & Strapparava, C. (2013). BRAINSUP: Brainstorming Support for Creative Sentence Generation. In *Proceedings of ACL'13.*

Özbal, G., Strapparava, C., & Tekiroglu, S. S. (2016). PROMETHEUS: A Corpus of Proverbs Annotated with Metaphors. In *Proceedings of LREC'16.*

Özbal, G., Strapparava, C., Tekiroglu, S. S., & Pighin, D. (2016). Learning to Identify Metaphors from a Corpus of Proverbs. In *Proceedings of EMNLP'16.*

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Qing-fang, X. (2004). The Innovative Use of Proverbs in Advertising English. *Journal of Pla University of Foreign Languages*, *5*, 003.

Raskin, V. (1979). Semantic mechanisms of humor. In *Annual Meeting of the Berkeley Linguistics Society* (Vol. 5, pp. 325–335).

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, *115*(3), 211-252.

Stock, O., Strapparava, C., & Valitutti, A. (2007). Moving Creative Words. *Advances in Brain, Vision, and Artificial Intelligence*, 509–522.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of CVPR'16.*

Tan, J., Wan, X., & Xiao, J. (2016). A Neural Network Approach to Quote Recommendation in Writings. In *Proceedings of CIKM'16* (pp. 65–74).

Veale, T. (2012). *Exploding the Creativity Myth: The Computational Foundations of Linguistic Creativity*.

Veale, T., Shutova, E., & Klebanov, B. B. (2016). Metaphor: A computational perspective. *Synthesis Lectures on Human Language Technologies*, *9*(1), 1–160.

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2017). Show and Tell: Lessons Learned from the 2015 MSCOCO Image Captioning Challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(4), 652–663.

Yamaguchi, K., Berg, T. L., & Ortiz, L. E. (2014). Chic or Social: Visual Popularity Analysis in Online Fashion Networks. In *Proceedings of ICM'14.*

Investigating the exploration-exploitation trade-off in dynamic environments with multiple agents

Denis Omar Palencia
Queen Mary University of London, London, United Kingdom

Magda Osman
Queen Mary University, London, United Kingdom

Abstract

Exploration and Exploitation represent two mutually exclusive goals associated with choices within an environment: search too little and the lack of information will make it difficult to distinguish good options penalizing the agent in the long run (exploiting) or search too much and suffer sub-optimal performance in the short term (exploring). Striking a balance between exploiting and exploring requires the learner to behave optimally in different environments. Managing this trade-off is an important process of our lives but isnt completely understood from a cognitive science perspective. To this end we present the findings from an experiment where the main objective was to examine how much the presence of competition and threats affects both behaviors: the presence of competition directs greater exploration and the presence of threats reduces this behavior, suggesting that learners prioritize their learning behavior in response to the presence of different types of agents in the environment.

Interference in Language Processing Reflects Direct-Access Memory Retrieval: Evidence from Drift-Diffusion Modeling

Dan Parker (dparker@wm.edu)
Department of English, Linguistics Program, College of William & Mary

Adam An (zan@email.wm.edu)
Linguistics Program, College of William & Mary

Abstract

Many studies on memory retrieval in language processing have identified similarity-based interference as a key determinant of comprehension. The broad consensus is that similarity-based interference reflects erroneous retrieval of a non-target item that matches some of the retrieval cues. However, the mechanisms responsible for such effects remain debated. Activation-based models of retrieval (e.g., Lewis & Vasishth, 2005) claim that any differences in processing difficulty due to interference in standard RT measures and judgments reflect differences in the *speed* of retrieval (i.e., the amount of time it takes to retrieve a memory item). But this claim is inconsistent with empirical data showing that retrieval time is constant due to the use of a direct-access procedure (e.g., McElree, 2000, 2006). According to direct-access accounts, differences in judgments or RTs due to interference arise from differences in the *quality* or *availability* of the candidate memory representations, rather than differences in retrieval speed. To adjudicate between these accounts, we employed a novel methodology that combined a high-powered (N = 200) two-alternative forced-choice study on interference effects with drift diffusion modeling to disassociate the effects of retrieval speed and representation quality. Results showed that the presence of a distractor that matched some of the retrieval cues lowered asymptotic accuracy, reflecting an effect of representation quality, but did not affect retrieval speed, consistent with a direct-access procedure. These results suggest that the differences observed in RTs and judgment studies reflect differences in the ease of integrating the retrieved item back into the current processing stream, rather than differences in retrieval speed.

Keywords: language processing; working memory; interference; two-alternative forced-choice task; drift diffusion modeling

Introduction

Successful language comprehension requires the ability to encode complex linguistic representations in memory and accurately access specific pieces of information in those representations to guide further elaboration of the discourse. For example, to relate the verb *play* in (1) with its subject for number agreement and thematic binding, memory retrieval mechanisms must access the encoding of the plural target subject *kids* and ignore featurally-similar items in non-target positions, such as the embedded plural noun *teachers*.

(1) **The kids**$_{pl}$ [that **the teachers**$_{pl}$ watched closely] ***played*** on the slide.

However, many studies have shown that featurally-similar items in non-target positions can interfere with retrieval of the target, impacting judgments of acceptability and reading times (for a review, see Parker, Shvartsman, & Van Dyke, 2017). Such effects are commonly called "similarity-based interference" (Gordon, Hendrick, & Johnson, 2001; Lewis & Vasishth, 2005; Lewis, Vasishth, & Van Dyke, 2006; Van Dyke, 2007; Van Dyke & Johns, 2012; Van Dyke & McElree, 2006, 2011). The goal of the current study is to help identify the source of such effects in language comprehension.

Often, interference from non-target items during retrieval for linguistic dependency formation slows reading times and lowers acceptability. This type of interference is called "inhibitory" interference (see Jäger, Engelmann, & Vasishth, 2017, for a review) and occurs in multiple match configurations where the target and a distractor overlap in some features that are relevant for retrieval, as in (1).

It has also been shown that interference can sometimes *speed up* processing and boost acceptability, resulting in an effect known as "facilitatory interference" or more commonly, "attraction" (Jäger et al., 2017). Attraction arises when the target and distractor are distinct in feature content, but neither is a perfect match to the retrieval cues. Such effects are commonly observed in the processing of subject-verb number agreement. For instance, Wagers and colleagues (2009) examined the comprehension of subject-verb agreement in sentences like (2) using self-paced reading and speeded acceptability judgments. The sentences in (2c-d) are ungrammatical because the plural verb *were* does not agree in number with the head of its subject noun phrase (NP) *key*.

(2) a. The key to the cabinets certainly was rusty …
b. The key to the cabinet certainly was rusty …
c. *The key to the cabinets certainly were rusty …
d. *The key to the cabinet certainly were rusty …

Wagers and colleagues found that in grammatical sentences like (2b), the number marking on the plural attractor *cabinet(s)* did not impact acceptability or RTs after the verb. However, in ungrammatical sentences like (2c), the plural distractor *cabinets* (the "attractor"), which matched the number of the verb *were*, boosted acceptability and facilitated RTs after the verb, relative to the ungrammatical condition with the singular noun *cabinet* (2d). Wagers and colleagues argued that the effects of facilitation and boosted

acceptability of sentences like (2c) were due to erroneous retrieval of the plural attractor. According to their account, retrieval functions as an error-driven repair mechanism that is triggered by the detection of an agreement violation. In (2), the subject NP predicts the number of the verb. When the verb form violates this prediction, as in (2c-d), the parser engages a cue-based retrieval at the verb to recover a number matching noun to license agreement. The attractor *cabinets* in (2c) will sometimes be incorrectly retrieved because it matches the verb in number, easing processing in a way that facilitates reading and boosts overall acceptability. In the grammatical conditions (2a-b), the verb fulfills the number prediction made by the subject NP, and therefore retrieval is not engaged, reducing the likelihood of attraction.

Alternative accounts exist, but many researchers concur that agreement attraction arises due to incorrect memory retrieval (e.g., Dillon, Mishler, Sloggett, & Phillips, 2013; Lago, Shalom, Sigman, Lau, & Phillips, 2015; Phillips, Wagers, & Lau, 2011; Schlueter, Williams, & Lau, 2018; Tanner, Nicol, & Brehm, 2014; Tucker & Almeida, 2017; Tucker, Idrissi, & Almeida, 2015). However, the reason for why incorrect retrieval facilitates RTs is debated and the relationship between RTs and retrieval accuracy remains underspecified.

For example, the prominent activation-based model of memory retrieval (ACT-R) developed by Lewis and Vasishth (Lewis & Vasishth, 2005) claims that the differences in RTs due to facilitatory interference (e.g., 2c vs. 2d) reflect differences in the *speed* of retrieval (i.e., the amount of time it takes to retrieve a memory item). In their model, the strength of an item's activation at the moment of retrieval determines the item's retrieval accuracy and its retrieval speed, such that items with higher activation are more likely to be retrieved and will be retrieved more quickly than items with a lower activation. In sentences that show attraction, like (2c), the plural attractor will have a higher activation than the singular attractor in (2d) because it provides a better match to the cues of the verb, and therefore will have a faster retrieval latency, resulting in faster RTs and boosted acceptability.

The activation-based model has been shown to provide a good fit to a wide range of behavioral data (Parker et al., 2017), but it is inconsistent with empirical evidence showing that retrieval speed is constant (i.e., time invariant) due to the use of a direct-access procedure (Martin & McElree, 2008, 2009, 2011; McElree, 2000; McElree & Dosher, 1989; McElree, Foraker, & Dyer, 2003; Van Dyke & McElree, 2011). According to direct-access accounts, the cues at retrieval make direct contact with the items in memory based on their content, rather than their location, which allows items to be retrieved at a constant speed, regardless of their position or dependency length. Items are differentially activated based on their (partial) match to the cues and the item that is most strongly activated is retrieved for dependency formation. On this view, the differences in RTs in (2c) vs. (2d) reflect differences in the quality (activation strength or availability) of the candidate memory representations, rather than differences in retrieval speed. For instance, the attractors in (2c) and (2d) will be retrieved in equal time, but the plural attractor in (2c) will be integrated into the processing stream more quickly because it provides a better match to the cues, resulting in faster RTs and boosted acceptability.

At present, it is difficult to distinguish between these accounts because the typical measures used to investigate attraction (e.g., reading times and judgments) do not discriminate between effects that arise from differences in retrieval speed and differences in representation quality. Furthermore, the argument for direct-access is based entirely on studies of *inhibitory* interference where distractors slow RTs (see Parker et al., 2017, for a review) and it remains unclear whether facilitatory interference effects like attraction show the same retrieval dynamics as inhibitory interference. These issues are addressed in the present study.

The Present Study

The goal of the present study is to tease apart existing predictions about retrieval speed and representation quality to better understand the source of facilitatory interference effects in language processing. Previously, research on retrieval in sentence processing has relied on the speed-accuracy trade-off (SAT) procedure (Dosher, 1979; Reed, 1973; Wickelgren, 1977) to examine the effects of retrieval speed orthogonally from effects of representation quality. In an SAT task, participants read sentences presented via rapid serial visual presentation (RSVP) and make binary judgments about sentence acceptability at cued intervals, ranging from before the tail of the critical dependency to 3-6 seconds after the dependent constituent is presented. Participants' average performance at these cue times is interpolated into an exponential curve that summarizes the speed-accuracy tradeoff function revealing the time course of retrieval. Importantly, by sampling a range of intervals, independent estimates of retrieval speed and accuracy become available. This method provides a profile of memory retrieval processes that is characterized by three parameters: (i) the **asymptote**, which reflects retrieval accuracy, (ii) the **intercept**, which reflects the time to retrieve an item from memory, and (iii) **rate**, which reflects the speed at which accuracy grows from the intercept to the asymptote. Differences in either the intercept or rate are presented as evidence for differences in retrieval speed, and differences in asymptote are taken to reflect differences in representation quality.

The SAT methodology has been pivotal in arguing that retrieval for sentence processing employs a time-invariant direct-access procedure (e.g., Martin & McElree, 2008, 2009, 2011; McElree, 2000; McElree et al., 2003). For instance, Van Dyke and McElree (2011) found that interference impacts asymptotic accuracy, but not processing speed (SAT intercept and rate parameters). But as noted, existing studies that have used SAT to investigate interference effects have been limited to tests of inhibitory interference. Furthermore, the SAT methodology is time-consuming and resource-intensive (see Chen & Husband, 2018, for discussion).

The current study employed a more efficient alternative methodology, **Drift Diffusion Modeling (DDM)**, which has

also been used to jointly analyze the effects of accuracy and processing speed and model the timing of retrieval (Chen & Husband, 2018; McElree & Dosher, 1989; Ratcliff, 1978; Ratcliff, Smith, Brown, & McKoon, 2016). Importantly, recent research on memory retrieval in sentence processing has shown that DDM yields results that are comparable to the more costly SAT methodology (Chen & Husband, 2018). Based on these results, we extended the DDM methodology to test existing predictions about retrieval speed and representation quality regarding facilitatory interference effects (i.e., activation-based vs. direct-access models of retrieval).

DDM uses data from two-alternative forced choice (2AFC) tasks to generate a conditional cumulative distribution function (CDF) that relates a time *T* to the probability that a correct response is faster than or equal to *T*. Crucially, it relies on four parameters that have been argued to reflect distinct underlying memory retrieval processes in sentence processing (Chen & Husband, 2018):

(i) **τ non-decision time**: encoding and motor response time, including the time to extract the relevant information from memory to make a decision
(ii) **α boundary separation**: the amount of evidence needed to make a decision
(iii) **δ drift rate:** rate of evidence accumulation
(iv) **β response bias**: the bias to respond to a particular alternative

In the current study, we tested a standard agreement attraction paradigm like that in (2) as a hallmark of facilitatory interference in a high-powered (N=200) 2AFC experiment and modeled the data using drift diffusion modeling to distinguish between effects arising from differences in retrieval speed vs. differences in representation quality. Recent research has used DDM to investigate how response biases impact the amount of attraction in sentences like (2c), as measured with the β response bias parameter (Hammerly, Staub, & Dillon, unpublished ms.). However, this work did not test the current predictions about retrieval speed, nor did it explicitly address the question of retrieval time. The present study applies the same methodology, but focuses on the issue of processing dynamics to better understand why interference eases processing in sentences like (2c).

Under both the activation-based and direct-access accounts, facilitatory interference should negatively impact asymptotic accuracy (DDM δ drift rate), such that the sentences that give rise to attraction (2c) should have an overall lower accuracy relative to the other conditions (2a, b, d). Where the accounts differ, however, is in their predictions for processing dynamics. If facilitatory interference arises due to faster memory access, as claimed by activation-based accounts, then we should see a faster intercept (τ non-decision time) for sentences that show attraction (2c). By contrast, if retrieval occurs via direct access, then the intercept parameters should be comparable across conditions.

Method

Participants

Participants were 200 college-age native speakers of English. The large sample size was chosen to ensure high statistical power (i.e., reduce Type II error) and accurate estimation of the DDM parameters. All participants provided informed consent and received credit in an introductory psychology or linguistics course. All participants were naïve to the purpose of the experiment. The experiment lasted approximately 20 mins.

Materials

Experimental materials consisted of 64 sets of 4 items like those shown in Table 1. The high number of item sets was chosen to ensure a stable estimation of the DDM parameters. Experimental conditions consisted of a 2 × 2 factorial design that crossed grammaticality (grammatical/ungrammatical) and attractor number (singular/plural). In all conditions, the subject head noun was modified by a prepositional phrase that contained the attractor. The critical verb was always a full lexical verb in sentence-final position. An adverb created a buffer between the subject and the critical verb to control for processing effects associated with plural nouns (see Wagers et al, 2009). Grammaticality was manipulated by varying the verb number such that it either matched or mismatched the number of the subject head noun. Attractor number was manipulated by varying the number of the attractor such that it either matched or mismatched the verb number.

The 64 target items were distributed across 4 lists in a Latin square design and combined with 66 fillers. Half of the fillers were ungrammatical, yielding an overall grammatical-to-ungrammatical ratio of 1:1. Approximately half of the grammatical fillers involved sentence-final plural verbs in structures similar to the target items and approximately half of the ungrammatical fillers involved sentence-final singular verbs to unconfound grammaticality with verb number in the target items. The remaining fillers involved relative clause structures from an unrelated experiment.

Table 1: Sample set of experimental materials. PL = plural. SG = Singular

Condition	Sentence
Grammatical PL attractor	The tutor for the students often rambles.
Grammatical SG attractor	The tutor for the student often rambles.
Ungrammatical PL attractor	The tutor for the students often ramble.
Ungrammatical SG attractor	The tutor for the student often ramble.

Procedure

Sentences were presented using Ibex (Drummond) one word at a time in the center of the screen in RSVP mode with a stimulus onset asynchrony (SOA) of 300 ms per word and an interstimulus interval (ISI) of 100 ms. Participants were instructed to read each sentence carefully and judge whether each sentence was an acceptable sentence of English. A response screen appeared for 3 s at the end of each sentence during which participants made a 'yes/no' response by button press. If participants waited longer than 3 s to respond, they were given feedback that their response was too slow. The order of presentation was randomized for each participant.

Data Analysis

All data were included in the analyses. A logistic mixed-effects model was fit to the judgment accuracy data and a linear model was fit to the raw response latencies using the *lmerTest* package (Kuznetsova, Brockhoff, & Christensen, 2014) in the *R* software environment (R Development Core Team, 2018), with fixed factors for the experimental manipulations (i.e., grammaticality and attractor number) and their interaction. All models were fit with the maximal random effects structure supported by the data (Barr, Levy, Scheepers, & Tily, 2013). An effect was considered significant if $|t/z| > 2$.

For the DDM analysis, the *RWeiner* package (Wabersich & Vandekerckhove, 2014) was used to fit a Weiner drift diffusion model to each condition for each participant. Parameter values that did not converge were excluded, following Chen and Husband (2018). A linear model was fit to the by-participant parameter fits following the same procedure used in the analysis of the response latencies. All data and code are available via Open Science Framework: https://osf.io/bu2kh/.

Results

Judgments and Response Latencies

Figure 1 shows the percentage of 'yes' responses and latencies (in ms) for the four experimental conditions. Main effects of grammaticality and attractor were observed in the judgments and latencies ($z > |3|$ in all cases). Grammatical sentences were more likely to be accepted and had faster latencies than ungrammatical sentences, and sentences with a plural attractor were more likely to be accepted and had longer latencies than sentences with a singular attractor. Crucially, judgments also showed a significant interaction of grammaticality with attractor number (z = -12.48). Planned pairwise comparisons revealed that this interaction was carried by the ungrammatical conditions: participants were more likely to accept an ungrammatical sentence when a plural attractor was present (z = 12.11). No such effect was observed in the grammatical conditions (z = -1.13). This profile reflects the behavioral signature of agreement attraction (Phillips et al., 2011) and provides an appropriate basis to examine the relationship between retrieval accuracy and retrieval speed using the DDM methodology.

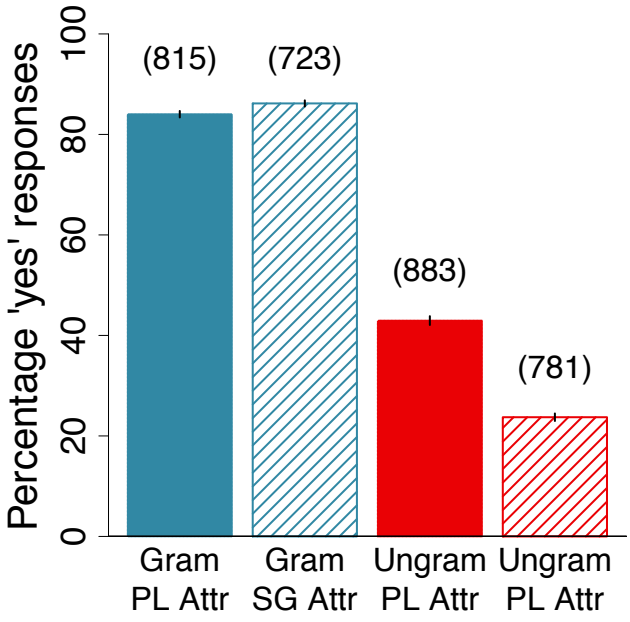


Figure 1: Mean percentage of 'yes' responses and response latencies in (ms) in parentheses by condition. Error bars indicate standard error of the mean. PL = plural, SG = singular.

Drift Diffusion Model (DDM)

Average DDM parameters by condition are shown in Table 2, and the *t*-values for model estimates of effects on DDM parameters are shown in Table 3. Figure 2 shows the cumulative density of accurate responses as a function of response time by condition. DDM revealed an effect of attraction on δ drift rate (asymptotic accuracy), qualified by an interaction between grammaticality and attractor number, such that participants were less accurate in ungrammatical sentences with a plural attractor than in those with a singular attractor. This effect is predicted by both accounts.

With respect to processing dynamics, which is where the accounts diverge, DDM revealed no significant effect of attraction on the processing dynamics reflected in τ non-decision time (intercept). These results suggest that agreement attraction impacts retrieval accuracy but not retrieval speed, consistent with a direct-access model of memory retrieval.

Results also showed a main effect of grammaticality on τ non-decision time (intercept), as grammatical sentences showed faster response latencies than ungrammatical sentences. This effect is unrelated to interference and likely reflects facilitation due to predictive processing in the grammatical conditions (Wagers et al., 2009).

Table 2: DDM parameters by condition.

	τ	α	δ	β
Grammatical PL attractor	0.26	2.03	0.90	0.60
Grammatical SG attractor	0.23	2.18	1.12	0.62
Ungrammatical PL attractor	0.30	1.99	-0.24	0.47
Ungrammatical SG attractor	0.29	2.39	-1.05	0.41

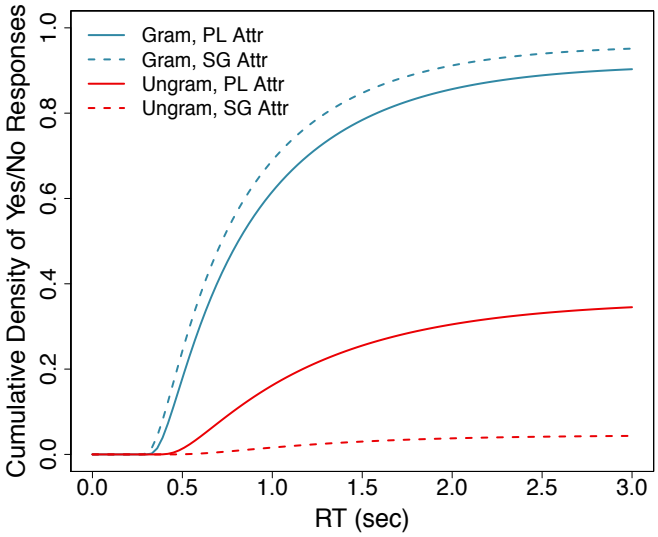


Figure 2: DDM estimations of the cumulative density of "yes" (1) and "no" (0) responsesas a function of response time by condition. PL = plural, SG = singular.

Discussion

The goal of the present study was to distinguish between existing predictions about retrieval speed and retrieval accuracy to better understand the source of interference effects in language processing. On the one hand, activation-based models of retrieval (e.g., Lewis & Vasishth, 2005) claim that differences in processing difficulty due to interference in standard RT measures and judgments reflect differences in the speed of retrieval. On the other hand, direct-access accounts claim that differences in judgments or RTs due to interference arise from differences in the quality of the candidate memory representations, rather than differences in retrieval speed, based on behavioral data showing that retrieval time is constant. To adjudicate between these accounts, we tested for facilitatory interference paradigm in a high-powered 2AFC experiment and modeled the results using DDM to disassociate the effects retrieval speed and representational quality.

Results of the 2AFC task replicated the classic attraction profile, such that ungrammatical sentences with a plural attractor that matched the number of the verb showed boosted acceptability relative to ungrammatical sentences with a singular attractor. Results of the DDM analysis revealed that in the ungrammatical conditions, the presence of a number-matching plural attractor lowered overall asymptotic accuracy, but did not affect retrieval speed.

The lack of an effect on non-decision time is consistent with the predictions of a direct-access procedure. These results suggest that the differences in judgments and RTs observed in agreement attraction studies reflect differences in the ease of integrating the retrieved item back into the current processing stream, rather than differences in retrieval speed.

More specifically, we argue that the quality of the memory representation (described in terms of activation strength) impacts the post-access stage of "binding", rather than the speed of access. In the memory literature, binding refers to the mechanisms by which information in memory is integrated together (Cohen & Eichenbaum, 1993; Hagoort, 2003; van der Velde & de Kamps, 2006), and it has been suggested that the effort required for integration is governed, in part, by the item's representation quality (Budiu & Anderson, 2004). On this view, retrieval of an item that satisfies at least some of the search criteria, such as the number matching attractor in sentences like (2c), will make post-retrieval integration faster compared to integration of an item that does not satisfy the search requirements, such as in (2d), giving rise to facilitatory interference.

More broadly, the current results are consistent with the recent claim that differences in the quality or availability of the information in memory leads to differences in accuracy and that those differences underlie the differences in reaction time studies (Martin & McElree, 2018). The current study extends this conclusion to facilitatory interference, motivating a unified analysis of inhibitory and facilitatory interference as the signature of direct-access retrieval.

Table 3: *t*-values for linear mixed effects model estimates on DDM parameters with 95% Cis in brackets.

	τ	α	δ	β
Grammaticality	**-4.85 [-0.08, 0.31]**	-1.77 [-0.45, 0.02]	**21.92 [1.98, 2.37]**	**11.85 [0.17, 0.24]**
Attractor number	1.17 [-0.00, 0.03]	**-3.64 [-0.62,- 0.18]**	**11.71 [0.67, 0.94]**	**2.55 [0.03, 0.09]**
Interaction	1.04 [-0.01,0.04]	1.88 [-0.01, 0.53]	**-11.47 [-1.20, -0.85]**	**-4.28 [-0.12, -0.04]**

References

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for testing interactions in linear mixed-effects models. *Frontiers in Psychology, 4*, 328.

Budiu, R., & Anderson, J. R. (2004). Interpretation-based processing: A unified theory of semantic sentence comprehension. *Cognitive Science, 28*, 1-44.

Chen, S. Y., & Husband, M. (2018). Comprehending anaphoric presuppositions involves memory retrieval too. Third Volume of Proceedings of the LSA.

Cohen, N. J., & Eichenbaum, H. (1993). *Memory, amnesia, and the hippocampal system*. Cambridge, MA: MIT Press.

Dillon, B., Mishler, A., Sloggett, S., & Phillips, C. (2013). Contrasting intrusion profiles for agreement and anaphora: Experimental and modeling evidence. *Journal of Memory and Lanuage, 69*, 85-103.

Dosher, B. A. (1979). Empirical approaches to information processing: Speed–accuracy tradeoff functions or reaction time. *Acta Psychologica, 43*, 347-359.

Drummond, A. Ibex Farm. Retrieved from http://spellout.net/ibexfarm

Gordon, P. C., Hendrick, R., & Johnson, M. (2001). Memory interference during language processing. *Journal of Experimental Psychology: Learning, Memory and Cognition, 26*, 1411-1423.

Hagoort, P. (2003). How the brain solves the binding problem for language: a neurocomputational model of syntactic processing. *NeuroImage, 20*, 18-29.

Hammerly, C., Staub, A., & Dillon, B. (unpublished ms.). *The grammatical asymmetry in agreement attraction reflects response bias: Experimental and modeling evidence*. https://osf.io/k4xc3/.

Jäger, L. A., Engelmann, F., & Vasishth, S. (2017). Similarity-based interference in sentence comprehension: Literature review and Bayesian meta-analysis. *Journal of Memory and Language, 94*, 305-315.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2014). lmerTest: Tests for random and diexed effects for linear mixed effect models (lmer objects of lme4 package. Retrieved from http://CRAN.R-project.org/package=lmerTest

Lago, S., Shalom, D., Sigman, M., Lau, E., & Phillips, C. (2015). Agreement processes in Spanish comprehension. *Journal of Memory and Language*.

Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science, 29*, 375-419.

Lewis, R. L., Vasishth, S., & Van Dyke, J. A. (2006). Computational principles of working memory in sentence comprehension. *Trends in Cognitive Science, 10*, 447-454.

Martin, A. E., & McElree, B. (2008). A content-addressable pointer mechanism underlies comprehension of verb-phrase ellipsis. *Journal of Memory and Language, 58*, 879-906.

Martin, A. E., & McElree, B. (2009). Memory Operations That Support Language Comprehension: Evidence From Verb-Phrase Ellipsis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*, 1231-1239.

Martin, A. E., & McElree, B. (2011). Direct-access retrieval during sentence comprehension: Evidence from Sluicing. *Journal of Memory and Language, 64*, 327-343.

Martin, A. E., & McElree, B. (2018). Retrieval cues and syntactic ambiguity resolution. *Language, Cognition, and Neuroscience, 33*, 769-783.

McElree, B. (2000). Sentence comprehension is mediated by content-addressable memory structures. *Journal of Psycholinguistic Research, 29*, 155-200.

McElree, B. (2006). Accessing recent events. In B. H. Ross (Ed.), *The psychology of learning and motivation - Adcances in research and theory* (pp. 155-200). San Diego: Academic Press.

McElree, B., & Dosher, B. A. (1989). Serial position and set size in short-term memory: Time course of recognition. *Journal of Experimental Pscyhology, 18*, 346-373.

McElree, B., Foraker, S., & Dyer, L. (2003). Memory structures that subserve sentence comprehension. *Journal of Memory and Language, 48*, 67-91.

Parker, D., Shvartsman, M., & Van Dyke, J. A. (2017). The cue-based retrieval theory of sentence comprehension: New findings and new challenges. In L. Escobar, V. Torrens, & T. Parodi (Eds.), *Language Processing and Disorders* (pp. 121-144). Newcastle: Comabridge Scholars Publishing.

Phillips, C., Wagers, M., & Lau, E. F. (2011). Grammatical illusions and selective fallibility in real-time language comprehension. In J. Runner (Ed.), *Experiments at the Interfaces* (Vol. 37, pp. 147-180). Bingley, UK: Emerald Publications.

R Development Core Team. (2018). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85*, 59-108.

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion Decision Model: Current Issues and History. *Trends in Cognitive Science, 20*, 260-281.

Reed, A. V. (1973). Speed-accuracy trade-off in recognition memory. *Science, 181*, 574-576.

Schlueter, Z., Williams, A., & Lau, E. (2018). Exploring the abstractness of number retrieval cues in the computation of subject-verb agreement in comprehension. *Journal of Memory and Language, 99*, 74-89.

Tanner, D., Nicol, J., & Brehm, L. (2014). The time-course of feature interference in agreement comprehension: Multiple mechanisms and asymmetrical attraction. *Journal of Memory and Language, 76*, 195-215.

Tucker, M. A., & Almeida, D. (2017). The complex structure of agreement errors: Evidence from distributional analyses of Agreement Attraction in Arabic. In A. Lamont & K. Tetzloff (Eds.), *Proceedings of the 47th Meeting of the North East Linguistics Society* (pp. 45-54). Amherst, MA: GLSA.

Tucker, M. A., Idrissi, A., & Almeida, D. (2015). Representing number in the real-time processing of agreement: Self-paced reading evidence from Arabic. *Frontiers in Psychology, 6*.

van der Velde, F., & de Kamps, M. (2006). Neural blackboard architectures of combinatorial structures in cognition. *Behavioral and Brain Sciences, 29*, 1-72.

Van Dyke, J. A. (2007). Interference effects from grammatically unavailable constituents during sentence processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*, 407-430.

Van Dyke, J. A., & Johns, C. L. (2012). Memory Interference as a Determinant of Language Comprehension. *Language and Linguistics Compass, 6*, 193-211.

Van Dyke, J. A., & McElree, B. (2006). Retrieval interference in sentence comprehension. *Journal of Memory and Language, 55*, 157-166.

Van Dyke, J. A., & McElree, B. (2011). Cue-dependent interference in comprehension. *Journal of Memory and Language, 65*, 247-263.

Wabersich, D., & Vandekerckhove, J. (2014). The RWiener Package: an R Package Providing Distribution Functions for the Wiener Diffusion Model. *The R Journal, 6*, 49-56.

Wagers, M. W., Lau, E. F., & Phillips, C. (2009). Agreement attraction in comprehension: Representations and processes. *Journal of Memory and Language, 61*, 206-237.

Wickelgren, W. A. (1977). Speed–accuracy tradeoff and information processing dynamics. *Acta Psychologica, 41*, 67-85.

Interpreting Metaphors in Real-time: Cross-modal Evidence for Exhaustive Access

Iola Patalas (iola.patalas@mail.mcgill.ca)
Department of Psychology, Concordia University
Montreal, Quebec H4B 1R6 CANADA

Roberto G. de Almeida (roberto.dealmeida@concordia.ca)
Department of Psychology, Concordia University
Montreal, Quebec H4B 1R6 CANADA

Abstract

Natural language is replete with figurative expressions like *my lawyer is a shark*, and listeners are expected to intuitively understand the intended, rather than the literal, meaning of such expressions. But what cognitive resources are involved in attaining meaning for such sentences? Most research into metaphor comprehension has employed *offline* reading tasks that provide no insight into the time-course of metaphor processing. In order to investigate the moment-by-moment on-line processes involved in metaphor comprehension, the present study used a naturalistic cross-modal lexical decision paradigm (Swinney, 1979) with novel brief masked target presentations during and after the vehicle word (*shark*). Results obtained from a preliminary sample demonstrated priming of related target words across conditions, but no significant differences between conditions. These results may best be interpreted as supporting an exhaustive-access account of metaphor interpretation, which suggests that literal and metaphorical interpretations are simultaneously accessed during the early stages of metaphor/simile interpretation.

Keywords: metaphors; similes; language comprehension; psycholinguistics; cross-modal lexical decision task; pragmatics

Introduction

How are metaphors interpreted in real time? This question is central to cognitive science because metaphors involve blatantly false statements that are nonetheless easily understood as conveying an ulterior, non-expressed meaning. For instance, upon hearing *my lawyer is a shark,* the listener does not call out the absurdity of the speaker's statement, rather assigning to it an interpretation that supposedly captures the speaker's intended meaning.

Various theories have been proposed to account for how listeners attain meaning for nominal metaphors in the form *X is Y*. Pragmatic theories take metaphors to convey initially a *literal* meaning which works as an invitation for the reader/listener to an interpretation based on the speaker's intended meaning (Davidson, 1978; Grice, 1975; Searle, 1979). While these authors differ in their approach to how metaphor is ultimately understood, they all suggest that the intended meaning can only be understood after an initial rejection of the literal, propositional meaning. In psycholinguistic circles this has been known as the *pragmatic model*, on the assumption that comprehension involves a three-stage process, beginning with the literal interpretation, followed by a rejection of the literal, and a search for the metaphorical meaning.

By contrast, the *direct-access* model suggests that metaphors are immediately comprehensible by the linguistic system and do not involve additional cognitive resources (e.g. Glucksberg & Keysar, 1990; Gibbs, 1994; Wolff & Gentner, 2000). Direct access is obtained either via a mechanism where metaphors are taken as comparisons between categories (e.g. Glucksberg & Keysar, 1990; Glucksberg, 2003), or via mapping of *constituent features* (say, features of *lawyers* and *sharks*), which are stored in the linguistic system as lexical properties of individual words (Wolff & Gentner, 2000). [1]

Although metaphors are pervasive in natural language and cognitive scientists have long debated the nature of their interpretation, to date few empirical studies have investigated the moment-by-moment process of metaphor comprehension using *online* experimental methods such as cross-modal priming with lexical decision (CMLD; e.g., Blasko & Connine, 1993); self-paced reading (e.g., Janus & Bever, 1985); ERP (Pynte, Besson, Robichon & Poli, 1996), and eye-tracking (Ashby, Roncero, de Almeida, & Agauas, 2018). However, support for the *direct-access* view is based primarily on studies involving *offline* tasks, i.e., tasks that require conscious judgment, and are thus not informative regarding what happens as sentences containing metaphors unfold in real time.

For instance, Glucksberg, Gildea, and Bookin (1982) asked participants to read literal and metaphorical sentences and to judge whether they were literally true or literally false. Based on a finding that it took longer for participants to judge statements as false if they had a common metaphorical interpretation (e.g. *jobs are jails*), the authors concluded that a metaphorical meaning is immediately available along with a literal meaning and thus interfered with subjects' classification of metaphorical sentences as

[1] The theories briefly mentioned here certainly do not exhaust the spectrum of metaphor theories. However, we restricted our review to theories concerned with the process of incremental interpretation, while leaving aside theories about the thinking processes that are triggered by or underlie metaphor production and comprehension (e.g., Lakoff & Johnson, 1978).

literally false (Glucksberg et al., 1982). The results obtained by *offline* studies such as Glucksberg and colleagues' (1982) could be equally compatible with the hypothesis that pragmatic processes interfere with literality judgments after the sentence has been fully processed, but before participants register a response.

Studies investigating on-line metaphor processing by measuring event-related potentials (ERP) have demonstrated that figurative targets elicited larger N400 amplitudes than literal targets (e.g. Pynte et al., 1996; Lai, Curran, & Menn, 2009), which suggests that figurative expressions are more difficult to process. This could be due to the detection of an incongruence between literal and intended speaker meaning. Crucially, these studies have not investigated what is accessed at the point at which a figurative expression is first processed.

The purpose of the present study was to investigate the nature of the representations computed during the real-time processing of metaphors and similes containing the same constituents, using an *online* cross-modal priming task. Specifically, we aimed to compare these two types of expressions in real time to elucidate differences in processing that might occur due to their fundamentally metaphorical and literal nature, respectively, and by doing so to shed light on the cognitive mechanisms involved in reaching an understanding of the meaning of such constructions. Our study is unique in that we aimed to study the very earliest moments of metaphor processing—the moment of lexical access—rather than later interpretation processes, and aimed to develop a new, more time-sensitive measure than in previous studies (e.g. Blasko & Connine, 1993).

We sought to compare the moment-by-moment comprehension of nominal metaphors in the form *X is Y* and similes in the form *X is like Y* using a cross-modal lexical decision task (CMLD; Swinney, 1979), and thus limited our study to nominal metaphors which could be compared to similes directly. Though they are traditionally thought to be an alternate form of the simile – a view dating back to Aristotle (trans. 1926) – the key difference between these two forms of expression is the word *like* in a simile, which renders it literally comprehensible. There is evidence that metaphors and similes are produced and understood differently (Ashby et al., 2018; Roncero, de Almeida, Martin, & de Caro, 2016; Roncero, Kennedy & Smyth, 2006) and yield different properties in offline studies (Roncero & de Almeida, 2015). Thus, we chose to use simile sentences as a literal control condition for nominal metaphor sentences due to their nearly identical constituent structure.

In the CMLD task, participants listen to aurally presented sentences for comprehension and are simultaneously presented with a visual target to perform a lexical decision task (i.e., pressing "yes" if the target is a word, "no" otherwise) in which response times (RTs) are collected. The main assumption behind the technique is that RTs to targets reflect the relation between a visual target and a prime word in the sentence (here, the vehicle *Y*). Specifically, recognition of the target word should be facilitated by hearing a related prime word, and thus yield a faster reaction time compared to a target that is semantically unrelated.

This method has two main advantages over other online techniques such as ERPs and offline tasks such as sentence judgments. First, listening to spoken metaphors during an *online* lexical decision task allows for an analysis of metaphor interpretation that is both highly time-sensitive and naturalistic. Using a simple lexical decision task rather than an *offline* judgment task means that participants do not base their responses on a conscious assessment of sentence meaning – indeed, they are not aware that this task is meant to test their comprehension of metaphors at all. Instead, priming for each target should reflect the interpretation of a sentence that is available at the moment visual targets are presented. Second, using similes as literal controls allows for all constituent words besides *like* (including target and vehicle) to remain identical, thus allowing for direct comparisons between literal and figurative interpretations of each topic-vehicle pair.

To our knowledge, the only other metaphor processing study to employ CMLD was that of Blasko and Connine (1993), which employed a substantially different method. In their study participants listened to metaphors and responded to targets presented at the offset of the vehicle. These targets were either (a) metaphor-related, (b) literal-related, or (c) control (unrelated to either the metaphor or the literal interpretation). In the present study, in addition to comparing metaphor to simile, we traced the time-course of interpretation by employing two target presentation points, thus probing for the potential access to literal or metaphorical interpretations over time. In addition, our first probe point was before the offset of the vehicle, during its *recognition point,* to test for the earliest possible position in which a literal or metaphorical interpretation could be obtained. Moreover, unlike Blasko and Connine (1993), our targets were forward- and backward-masked with a series of crosshatches, and presented at a fast rate (80ms) in an attempt to circumvent subjects' potential detection of a relation between prime and target.

Method

Participants

Participants were 37 native English speakers between the ages of 19 and 59 (M = 26.32, SD = 8.07; 26F) with normal or corrected-to-normal vision and hearing who met the following inclusion criteria: (1) They learned English before the age of 5 (M = 1.19, SD = 1.47) and identified it as their native and dominant language; (2) they rated themselves as fluent in speaking, listening, and reading English; (3) they reported no history of hearing or reading disability. Participants who were recruited via Concordia University's online participant pool were compensated with course credit while all other participants were compensated with $10 for one hour of participation. Participants for two pretests are described along with the pretests below.

Materials Experimental materials consisted of 32 sentences containing metaphors/similes in the form *X is (like) Y* and 160 filler sentences. Metaphor/simile sentences were selected from Roncero and de Almeida (2015), which consists of a set of metaphor/simile sentences with accompanying norms. The sentences were chosen on the basis of their high aptness ratings (rated above 6 on a scale of 1 to 10, with 10 being the most apt), but had a broad range of familiarity ratings. The Roncero and de Almeida (2015) norming study asked participants to generate associates/explanatory words for both the simile and metaphor versions of each sentence and for the topic and vehicle words in isolation. For use as our figuratively related targets, we selected explanatory words generated for each metaphor by the highest possible number of participants, which did not appear as associates for the vehicle word in isolation. For our literally related targets, we selected words which were generated as associates of the vehicle word by the highest possible number of participants and which did not appear as explanatory words for the metaphor on the whole.

Exclusion of Automatic Associates To ensure that any potential priming effects were not derived from an 'automatic' association between the vehicle and target words (i.e., due to being frequently paired in speech, like *salt* and *pepper*), we conducted a norming experiment where each vehicle word was read aloud to 12 native speakers of English, who were asked to say out loud the first word that came to mind. Their responses were collected and any word which was named more than twice was excluded from selection as a target for that vehicle word.

The unrelated control words selected to calculate priming effects were chosen according to the following criteria: For each related target word, written frequency was calculated from the Corpus of Contemporary American English (COCA; Davies, 2008), a database of American English texts collected from 1990-2017 including fiction, non-fiction and academic texts. Matched (unrelated) control words were selected to have the same number of letters, same number of syllables, same morphological structure and similar frequency in the COCA database.

Sentence Recording and Target Selection Metaphors/similes were embedded in longer sentences with explanatory contexts which we generated, with the word *because* following each vehicle word to control for interference from explanatory contexts; these sentences also began with generic *proposition-attitude* statements (e.g., *It is hardly a secret that lawyers are sharks, because with few exceptions, lawyers are bloodthirsty and ruthless*). Filler sentences did not repeat the topic or vehicle words of any experimental sentences. Of these, 32 followed a similar sentence structure as experimental sentences, while 128 filler sentences did not syntactically resemble experimental sentences. Visual targets for filler sentences were 64 real English words and 96 'nonsense' strings of letters that did not resemble English words, of varied lengths to reflect the varied lengths of experimental targets. All sentences were read by a female native English speaker and recorded for aural presentation, with natural prosody and reading speed. Special attention was given to matching the prosody and timing of metaphor and simile pairs, to make them nearly identical except for the word *like*.

Recognition Times We employed a gating paradigm to determine the recognition point of each vehicle word, following the procedure developed by Zwitserlood (1989). Recordings of each vehicle word were cut into slices increasing by 50ms each. These were played consecutively to 10 native speakers of English over noise cancelling headphones. Participants were asked to write down what word they thought they were hearing after each slice was presented. Their responses were collected and recognition times for each word were defined as the moment when 80 percent of participants correctly identified the word (with or without pluralization). During the lexical decision task, the early time point was defined as 40ms prior to recognition time, to account for screen refresh rate and the fact that the word could have become recognizable anytime within the 50ms slice participants heard during the gating task. Late time points were defined as 500ms following recognition time to avoid interference from words later in the sentence.

Experimental Design

A total of 16 counterbalanced lists were created following a 2 x 2 x 2 x 2 design. Each topic/vehicle pair was presented in either a metaphor- or simile-containing sentence, along with a figuratively related target, literal target or matched control target, at an early (*recognition*) or late time point. Each block contained two experimental sentences in each condition along with all 160 filler sentences, 20 of which were followed by comprehension questions to ensure participants were attending to aural stimuli. Each participant completed two blocks containing one list each – i.e., each participant heard both the simile and metaphor version of each sentence once in total. The sentences were randomized in order within each block of trials and participants were randomly assigned to each set of lists.

Procedure

Participants were tested on an iMac computer using Psyscope X B57 (Cohen, MacWhinney, Flatt, & Provost, 1993) using a button box. After voluntary consent was obtained, each participant was seated in front of the screen in a dark room, equipped with noise-cancelling headphones, and instructed to attend to both the aurally presented sentences and visual stimuli on the screen. Participants were instructed that their primary task was to identify whether the letters they saw on the screen constituted an English word and to press a button to indicate YES or NO as quickly and accurately as possible, while their secondary task was to answer comprehension questions about the sentences they heard over the headphones.

Each trial consisted of a prompt asking participants to press a button when they were ready for the next trial, followed by an aural presentation of each sentence. Target words appeared in white 20-point Arial font text in capital letters on a black screen for 80ms each, preceded and followed by masks which appeared for 100ms. This brief masked priming procedure was meant to reflect faster and more automatic processes of recognition rather than slower processes of judgment. Masked priming (see Forster, 1999) reflects early processes of lexical recognition which should be uncontaminated by other semantic factors. Each participant was given five randomized practice trials, during which the experimenter answered questions and corrected mistakes.

Data Analysis

Analysis of reaction times (RTs) was restricted to correct trials (i.e., those where participants correctly identified the target as an English word) while incorrect trials were omitted (13% of all data points). As is standard in lexical decision paradigms (Friedmann, Taranto, Shapiro & Swinney, 2008), all reaction times above 2 seconds were discarded prior to data analysis (2% of all data points). Based on a priori decisions, we discarded blocks of trials where participants answered fewer than 70% of comprehension questions correctly.

Results

We performed a linear mixed-effects model regression analysis with subjects and items (vehicles) entered as random effects with random intercepts. Raw RTs were regressed on priming (control/experimental targets), sentence literality (metaphor/simile conditions), target type (figurative/literal) and time-point (early/late), as well as all first order interaction terms. For ease of interpretation, priming effects (RT to control – RT to target) are presented in Figure 1. The full RT model was compared to a null model including only random effects (subject and item), using the Likelihood Ratio Test to determine significance. Our model provided a better fit to the data than the null model ($\chi^2(10) = 25.70$, $p = 0.004$). We derived p-values for all main effects and interactions using the Likelihood Ratio Test to compare the full model to a model excluding the relevant term (see Table 1) and found only one significant main effect of priming.

As predicted, participants took significantly longer to respond to unrelated targets than to related targets ($\chi^2(4) = 22.38$, $p < 0.001$) – overall, RTs to related targets were 40ms faster (*SEM*=23.88). While no other main terms or interaction terms reached significance, the respective means of each condition seemed to show trends which may be worth investigating with a larger sample. Specifically, in the metaphor condition, early priming values were lower for the figurative condition than for the literal condition, but priming for the figurative condition was higher at the later time point. In the simile condition, the reverse was true, with higher priming for literal targets at the late time point. Unexpectedly, the largest priming effect was observed for figuratively related targets at the early time point of the simile condition.

Table 1: Mixed-effects linear model of response times.

Predictor	Estimate	SE	t	95% CI	Null Comparison
Constant	718.19	23.88	30.08	[671.39, 765.00]	
Priming	-39.51	16.10	-2.45	[-71.06, -7.95]	$\chi^2(4)=22.38$, $p<.001$
Time-point	-1.84	16.05	-0.11	[-33.29, 29.61]	$\chi^2(4)=1.67$, $p=.80$
Target type	-16.44	16.15	-1.02	[-48.10, 15.22]	$\chi^2(4)=1.27$, $p=.87$
Sentence literality	-18.04	16.02	-1.13	[-49.44, 13.37]	$\chi^2(4)=2.31$, $p=.68$
Priming x Time-point	1.43	15.81	0.09	[-29.55, 32.40]	$\chi^2(1)=0.0083$, $p=.93$
Priming x Target type	8.09	15.87	0.51	[-23.02, 39.19]	$\chi^2(1)=0.26$, $p=.61$
Priming x Sentence literality	-4.50	15.83	-0.28	[-35.52, 26.52]	$\chi^2(1)=0.08$, $p=.78$
Time-point x Target type	2.01	15.78	0.13	[-28.93, 32.94]	$\chi^2(1)=0.02$, $p=.90$
Time-point x Sentence literality	14.28	15.77	0.91	[-16.64, 45.19]	$\chi^2(1)=0.82$, $p=.36$
Target type x Sentence literality	10.39	15.88	0.65	[-20.74, 41.52]	$\chi^2(1)=0.43$, $p=.51$

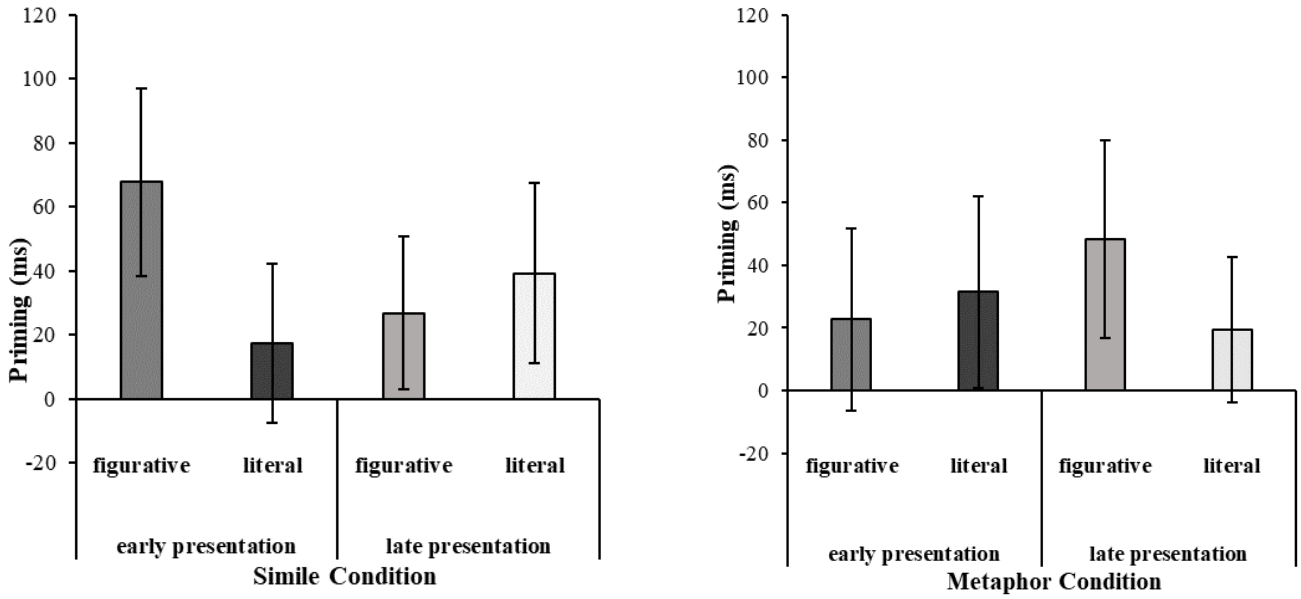


Figure 1: Mean priming effects between unmatched control words and related target words in Metaphor and Simile conditions as a function of time point and literality. Error bars represent SEM

.

Discussion

We used a novel, masked rapid-presentation CMLD task to gain insight into the moment-by-moment processing of metaphors and similes of the form *X is/is like Y*. The results obtained showed significant priming in all conditions and at all time points; and, contrary to our predictions (and those of the *pragmatic* model) no statistically significant differences in priming between conditions were obtained.

These results can be partially accounted for by different models of metaphor interpretation, in particular models that assume some form of exhaustive access. This is so because, contrary to what both *pragmatic* and direct-access models would predict, literally related target words were still primed as much at the later time point (b) as at point (a), suggesting that even after a sentence has been fully processed (and, presumably, understood to have a non-literal intended meaning), literal representations of the vehicle word linger.

We take these results to suggest that metaphor/simile interpretation trigger an *exhaustive* access, an effect also found in some lexical ambiguity resolution studies (e.g., Swinney, 1979). Exhaustive access, in the context of metaphor processing, entails the access to both the literal meaning and a meaning commonly associated to the metaphorical use of the same word. According to Carston (2010), two simultaneous processes contribute to the understanding of metaphorical language—a fast, *online* formation of *ad hoc* concepts linked to the metaphorical vehicle (for example, while the lexical item *shark* may conceptually represent the large predatory fish, it may also represent a concept like *aggressive* or *mean*, especially for highly apt/conventional metaphors such as many of those used in our experiment), and a more nuanced, *offline* process of interpreting the meaning of a metaphorical passage that relies on its literal meaning and the "images" the literal meaning evokes. Thus, according to Carston's (2010) model, the early priming of figuratively related targets presented at recognition point (a) could be a result of *ad hoc* concept formation relating the vehicle word to figurative concepts, while the persistence of priming for literally related targets at point (b) could be explained by the persistent, simultaneous activation of literal (or imagistic) representations. However, it is not clear whether these *ad hoc* concepts—which are obtained from contextually-driven *inferences*—are in fact accessed within the 80 ms window of target processing. In Carston's relevance-theoretical approach, these *ad hoc* concepts would have to be constrained by context. But this would imply a sentence type x target type interaction, which we did not obtain. Alternatively, these *ad hoc* concepts are already associated with the vehicles, such that a rapid *shark→MEAN* access could be obtained similarly to the literally related *shark→BLOOD*.

Our results are also partially compatible with Giora's (2003) graded salience hypothesis and in particular its ancillary hypothesis, *retention*. These hypotheses can be summarized as follows. Effects such as frequency or familiarity lead to a graded representation for meanings or senses of a word. These factors determine a form of exhaustive, but ordered access to meanings in the course of interpretation. For metaphors, this means that the most salient meaning—metaphorical or not—will always be accessed first, or activated more strongly. This theory is, in

large part, context-insensitive; that is, it takes the order of access to be determined by lexical-semantic encoding factors, not determined by context. We only say that our results are partially compatible with this theory because we have not tested specifically for the salience of particular senses.

A third compatible view takes relations between lexical concepts to be established in terms of meaning postulates (see, e.g., de Almeida, 1999; Partee, 1995; and Fodor, 1998). An application of this view to the interpretation of the present results would take vehicle meanings to quickly trigger their related postulates, whether they are related to literal or to figurative meanings. Thus, for example, a meaning postulate would constitute a relation between the meaning of the prime (*shark*) and the meanings of the targets, via postulates with the form such as $(\forall x[P(x)] \rightarrow [Q(x)])_n$. This view requires both $\forall x[SHARK(x)] \rightarrow [MEAN(x)]$ and $\forall x[SHARK(x)] \rightarrow [BLOOD(x)]$ to be postulates related to the meaning of *shark*, with both being equally primed, independent of context, and as a function of lexical-conceptual relations established not *by necessity* (i.e., *analytic*) but as a function of use (i.e., *synthetic*).[2]

Our results cannot currently set these theories apart, nor was this experiment conceived to contrast them directly. Moreover, despite our exhaustive-access effects, tendencies observed in the group means for each condition suggest that there may be differences in priming between conditions. In the metaphor condition, mean priming for figurative targets was higher at time point (b) than at time point (a), and priming for literal targets was higher at time point (a) than (b), although none of these differences reached a threshold of significance, which may suggest that figurative associations of the vehicle word are accessed more easily after *pragmatic* processes have been implemented. Additionally, priming for figurative targets at time point (b) was higher than for literal targets, which may suggest that literal associations with the target word are inhibited once the metaphor has been fully processed and understood.

Conversely, in the simile condition, group means indicated that priming was higher for literal targets at time point (b) than at time point (a); priming for literal targets at time point (b) was also higher than for figurative targets at the same point (b). These results suggest that similes are interpreted as literally true sentences and tend to activate literal meanings once fully processed, as the *pragmatic* model suggests. One unexpected tendency observed in the group means was that figurative targets were primed more at recognition point (a) than literal targets, and primed more at point (a) in the simile condition than the metaphor condition. A possible explanation for this result is that the word *like* in similes could lead participants to anticipate an upcoming vehicle word that is not typically literally related to the topic of the sentence.

The gating paradigm used to determine recognition points tested the moment at which each word is recognized in isolation, but context could bias listeners to correctly identify the word earlier when presented within a sentence. In the context of highly familiar similes such as *time is like money*, the word *like* could in fact trigger an assumption in the listener that the word *money* will follow, due to the frequency with which the simile is used in common language use—and cause the recognition point of the vehicle word to occur earlier than anticipated. In order to test this possibility, additional experiments are being conducted relating the strength of the early figurative priming effect to the familiarity rating of each simile.

A major methodological difference between our study and other psycholinguistic experiments employing cross-modal lexical priming (e.g., Swinney, 1979; Friedmann et al., 2008) was our use of briefly presented masked visual targets. Typically, cross-modal lexical decision tasks employ an unmasked visual target presentation lasting at least 500ms (e.g. Friedmann et al., 2008), which allows for much higher response accuracy. Forster (1999) explained that the use of very rapid masked primes should circumvent conscious thought processes about prime and target words and, instead, reflect unconscious processes of word association. Our use of masked visual targets presented for 80ms combined with presentation times at the recognition point of aurally presented vehicle words followed the rationale that in order to observe unconscious *on-line* access to semantically related concepts during metaphor processing, participants should not be allowed time to consciously consider either visual target or aurally presented vehicle. This created a speed-accuracy trade-off that resulted in a loss of data; however, the data obtained should be reflective of unconscious (*online*) facilitation processes. Experiments with greater statistical power might resolve whether tendencies observed in support of the *pragmatic* model reflect real differences in priming between conditions. Alternatively, we have also considered three views that seem compatible with an exhaustive access of both metaphorical and literal representations. What our present results seem to indicate is that there is no direct access to the contextually-determined, conventional metaphorical interpretation (e.g., Gibbs, 1994) without access to the literal meaning.

In conclusion, we found priming to targets related to both figurative and literal interpretations of metaphor and simile vehicles. The effect was found at both the recognition point (i.e., before the offset of the *vehicle*), and 500ms later. What is surprising is that we obtained priming effects at a fast target presentation time (80ms) under masking conditions, suggesting exhaustive access to literal and nonliteral-related targets before conscious judgments of metaphoricity could be made.

[2] For ease of exposition, we are simplifying the presentation of these meanings postulates, which might involve other predicate-argument relations.

Acknowledgments

This research was supported by a grant from the National Science and Engineering Research Council of Canada (NSERC).

References

Aristotle. (1926). *The art of rhetoric.* New York, NY: G. P. Putnam's Sons.

Ashby, J., Roncero, C., de Almeida, R. G., & Agauas, S. J. (2018). The early processing of metaphors and similes: Evidence from eye movements. *Quarterly Journal of Experimental Psychology, 71*(1), 161-168.

Blasko, D. G., & Connine, C. M. (1993). Effects of familiarity and aptness on metaphor processing. *Journal of Experimental Psychology, 19*(2), 295-308.

Carston, R. (2010). Metaphor: Ad hoc concepts, literal meaning and mental images. *Proceedings of the Aristotelian Society (Hardback), 110*(3pt3), 295-321.

Davidson, D. (1978). What metaphors mean. *Critical Inquiry, 5*(1), 31-47.

Davies, M. (2008-). The Corpus of Contemporary American English (COCA): 560 million words, 1990-present. Available online at https://corpus.byu.edu/coca/

de Almeida. (1999). What do category-specific semantic deficits tell us about the representation of lexical concepts? *Brain and Language, 68,* 241-248.

Fodor, J. A. (1998). *Concepts.* Oxford: Clarendon Press.

Forster, K. I. (1999). The microgenesis of priming effects in lexical access. *Brain and Language, 68*(1), 5-15.

Friedmann, N., Taranto, G., P Shapiro, L., & Swinney, D. (2008). The leaf fell (the leaf): The online processing of unaccusatives. *Linguistic Inquiry, 39*(3), 355-377.

Gibbs, R. (1994). *The Poetics of Mind: Figurative Thought, Language, and Understanding.* New York, NY: Cambridge University Press.

Giora, R. (2003). *On Our Mind: Salience, Context, and Figurative Language.* Oxford: Oxford University Press.

Glucksberg, S. & Keysar, B. (1990). Understanding metaphorical comparisons: Beyond similarity. *Psychological Review, 97*(1), 3-18.

Glucksberg, S. (2003). The psycholinguistics of metaphor. *Trends Cogn Sci, 7*(2), 92-96.

Glucksberg, S., Gildea, P., & B. Bookin, H. (1982). On understanding nonliteral speech: Can people ignore metaphors? *Journal of Verbal Learning and Verbal Behavior, 21*(1), 85-98.

Grice, H. P. (1975). Logic and Conversation. In Cole P., & Morgan, J. (Eds.), *Syntax and Semantics*. New York, NY: Academic Press.

Janus, R., & Bever, T. (1985). Processing of metaphoric language: An investigation of the three-stage model of metaphor comprehension. *Journal of Psycholinguistic Research, 14*(5), 473-487.

Lai, V. T., Curran, T., & Menn, L. (2009). Comprehending conventional and novel metaphors: An ERP study. *Brain Research, 1284*, 145-155.

Lakoff, G., & Johnson, J. (1978). *Metaphors We Live By.* Chicago: University of Chicago Press.

Partee, B. (1995). Lexical Semantics and Compositionality. In Gleitman, L. R., Osherson D. N., & Liberman. M. (Eds.), *An Invitation to Cognitive Science: Language.* Cambridge, MA: The MIT Press.

Pynte, J., Besson, M., Robichon, F.-H., & Poli, J. (1996). The time-course of metaphor comprehension: An event-related potential study. *Brain and Language, 55*(3), 293-316.

Roncero, C., & de Almeida, R. G. (2015). Semantic properties, aptness, familiarity, conventionality, and interpretive diversity scores for 84 metaphors and similes. *Behavior Research Methods, 47*(3), 800-812.

Roncero, C., de Almeida, R. G., Martin, D. C., & de Caro, M. (2016). Aptness predicts metaphor preference in the lab and on the internet. *Metaphor and Symbol, 31*(1), 31-46.

Roncero, C., Kennedy, J. M., & Smyth, R. (2006). Similes on the internet have explanations. *Psychonomic Bulletin & Review, 13*(1), 74-77.

Searle, J. (1979). Metaphor. In A. Ortony (Ed.), *Metaphor and Thought* (pp. 83-111). Cambridge University Press.

Swinney, D. A. (1979). Lexical access during sentence comprehension: (Re)consideration of context effects. *Journal of Verbal Learning and Verbal Behavior, 18*(6), 645-659.

Wolff, P., & Gentner, D. (2000). Evidence for role-neutral initial processing of metaphors. *J Exp Psychol Learn Mem Cogn, 26*(2), 529-541.

Zwitserlood, P. (1989). The locus of the effects of sentential-semantic context in spoken-word processing. *Cognition, 32*(1), 25-64.

Family Resemblance in Unsupervised Categorization: A Dissociation Between Production and Evaluation

John. D. Patterson (jpatter4@binghamton.edu)
Sean Snoddy (ssnoddy1@binghamton.edu)
Kenneth J. Kurtz (kkurtz@binghamton.edu)

Department of Psychology, Binghamton University (SUNY)
Binghamton, NY 13902 USA

Abstract

A plurality of the categories we hold exhibit family resemblance (FR; i.e., many characteristic but few defining features), suggesting FR may occupy a central role in human category formation. However, research in unsupervised learning has shown that when people are asked to sort an array of novel items into categories, they ubiquitously use a unidimensional (UNI) rule – despite the availability of a FR solution. This work suggests that, perhaps, FR similarity is *not* a core tendency in category formation. Here, we question whether the UNI bias is a result of the sorting paradigm. Specifically, we speculate the paradigm conflates two components vital for category formation: production and evaluation. Across three experiments we show that when evaluation is separated from generation – by using a novel forced-choice task that pits different category organizational schemes against one another – people exhibit a FR over UNI preference. The implications of these results are discussed.

Keywords: unsupervised categorization; similarity; family resemblance; unidimensional bias; category construction

Introduction

Understanding the cognitive basis on which we create novel categories in the absence of feedback is foundational to understanding human category learning more broadly. One way to address this question is simply by studying the categories we already hold. Theoretical and behavioral work has shown that the natural categories are described by a family resemblance (FR), or overall similarity, structure wherein members of a category share many characteristic features but share few or no defining features (Rosch & Mervis, 1975, Wittgenstein, 1953). Given the prevalence of FR among natural categories, an intuitive hypothesis is that overall similarity is the preferred or default basis on which we form novel categories.

Under this hypothesis, Medin, Wattenmaker, and Hampson (1987) sought to investigate unsupervised category formation more directly by using a sorting paradigm in which the participant was given an array of novel, multi-dimensional stimuli and asked to sort them into two equal-size categories. Critically, the examples could be sorted based on FR or, alternatively, based on a unidimensional (UNI) rule (e.g., 'red things in one category, blue things in another'). Contrary to their expectations, Medin et al. (1987) demonstrated across several experiments that people overwhelmingly preferred to construct categories described by a UNI rule; they *very seldom* created categories adherent to FR, despite UNI solutions having less within-category, and more between-category, similarities than those based on FR (Medin et al., 1987). Much work has subsequently replicated the strong UNI bias under the full-array sorting task (e.g., Ahn & Medin, 1992; Lassaline & Murphy, 1999; Regehr & Brooks, 1995; Wattenmaker, 1992).

The inconsistency between the UNI bias in the full-array sorting task and the tendency for natural categories to be described by FR has puzzled the field, and much research has been devoted to understanding why such an inconsistency exists. Coarsely, this work has two central themes: feature and task effects. Research on feature effects has shown that, generally speaking, changing the quality or the number of features is ineffective at reversing the UNI bias – and in some cases can exacerbate it (Regehr & Brooks, 1995). This of course comes with one notable exception: prior knowledge. People produce more FR sorting when the features of the FR categories map onto known concepts (e.g., the features of extro- vs. introversion), or novel concepts taught to participants, that explain and relate the features to one another (e.g., Ahn, 1999; Medin et al., 1987). This work is important for understanding category formation; however, we consider it to address a fundamentally different question. Instead of asking, "what is the organizational basis used when forming a completely novel/artificial category," it asks, "given a latent or manifest conceptual/causal basis in the features, do people construct categories by it?"

Research on task effects has shown two critical findings. First, Lassaline and Murphy (1996) showed that an inference task (e.g., 'if it has feature A on dimension 1, what feature is it likely to have on dimension 2') prior to the sorting task increased FR responding. This experiment shows that encoding feature co-prediction is vital for generating FR. However, it leaves open the question of whether people do this kind of feature encoding spontaneously during category formation.

Second, Regehr and Brooks (1995) used a novel Match-to-Standards (MTS) task in which participants sorted each item, one at a time, by matching them to one example item from each FR category; each sorted example covered up the previous example that was sorted into that category and the standards remained visible throughout the task. The MTS task led to much greater FR responding, relative to the full-array sorting task. This is suggestive that FR structure in

natural categories might be an emergent property of many item-item matches, though this connection has never been empirically drawn. Several subsequent studies have followed up on this approach to studying unsupervised categorization, examining factors such as feature separability, time pressure, and working memory load (Milton, Longmore, & Wills, 2008; Milton & Wills, 2004; Wills, Milton, Longmore, Hester, & Robinson, 2013).

However, we have two concerns with the MTS task, as the data currently stand. One, it is based on local item-to-item matching in which only the most recent item sorted into each category is visible (along with the standards); thus, while people produced more FR responding, it is unclear if that responding is attributable to the participant's appreciation of similarity structure generated across examples in each category or if instead FR responding was simply a product of making local matches (without appreciating category-level structure). Two, the task does not measure unsupervised category formation. Given the supervision, in the form of standards from each category, the task is instead a measure of semi-supervised category formation (Patterson & Kurtz, 2018; Vong, Navarro, & Perfors, 2016).

The goal of the current work was to investigate task effects in unsupervised category formation from a novel perspective that aims to address some of the limitations of previous research on task effects. We ask whether UNI similarity is indeed a deep organizational preference in category formation or if, instead, it is a direct product of the standard full-array sorting paradigm (e.g., Medin et al., 1987). We identify three aspects of the sorting task that independently or in conjunction could encourage UNI solutions. First, the task presents a whole array of multi-dimensional stimuli to the participant simultaneously. Intuitively, complexity in both the number of items and number of features might encourage problem simplification in the form of sample or dimensionality reduction. As sample reduction is not an option (participants must include all items in the solution), dimensionality reduction may be utilized.

Second, the goal of the sorting task is to produce two categories; this goal is decidedly intentional and discriminative – i.e., goal is to predict/separate class. Research has shown that intentional and discriminative learning leads to greater rule focus relative to either incidental learning or learning where class-prediction is softened (Levering & Kurtz, 2015; Love, 2002). Third, and critically, the standard sorting task conflates two intuitively essential components for category formation: the *generation* of a candidate category structure and the *evaluation* of that structure relative to possible alternatives. Given the goal of the task is to generate *one* candidate structure, the evaluation of this structure is likely to be inadequate due to insufficient alternative structures with which to compare it to. Furthermore, we expect candidate structures in naturalistic settings are generated not cold (as in the standard paradigm), and not *necessarily* with prior top-down knowledge, but through feature statistics that accrue with incidental experience (Lassaline & Murphy, 1996; Love, 2002). As such, we expect the structure hypotheses generated by participants in the sorting paradigm to be immature.

In the experiments that follow, we introduce a novel Structure Choice Task (SCT) in which two candidate structures are presented side-by-side and the participant chooses which they prefer. The task thus obviates the need to generate structure hypotheses (which we believe are undermined in the standard sorting paradigm) and isolates structure evaluation. If FR is a preferred organizational principle in category formation (and if the UNI bias is a product of the standard sorting paradigm), we should expect people to choose the FR structure more frequently than the UNI one.

We point out that the SCT resolves limitations from previous task effects research in two ways. First, the task does not rely on any pre-task encoding manipulations; participants encode the items/structures however they wish and make a judgment. As such, the preferences produced are spontaneous. Second, the task does not restrict the number of stimuli that are under consideration, as in the MTS task. Because the SCT presents whole categories, organized in two different ways, it should reflect the participant's category-level similarity preference (rather than local matching).

In Experiment 1A, we pit FR against UNI in the SCT and provide first-ever evidence of a spontaneous FR preference in a category-level task that uses knowledge-poor features. In Experiment 1B, we replicate E1A and extend it by comparing SCT results to full-array sorts completed either before or after the SCT; despite replicating the FR preference in the SCT, effectively nobody produced FR sorts – suggesting the standard sorting paradigm encourages UNI solutions. In Experiment 2, we address potential critiques to FR supremacy in the SCT.

Experiment 1A

In a within-subjects design, FR and UNI organizations of the same items were pitted against one another in the SCT. Without having to generate hypotheses, we predicted participants would prefer FR organizations.

Method

Participants 108 undergraduate students at Binghamton University participated.

Materials and Design The stimuli were based on a five-dimensional variant of the abstract FR category structure from Medin et al. (1987); each binary dimension is represented as a pair features (see Figure 1). Each category consisted of a prototype – containing all five characteristic features of the category – and five 'one-off' items that differed from the prototype by a feature that was consistent with the opposite prototype. Five 12-item stimulus sets, from distinct domains (see Figure 1 for prototypes), were

created from this abstract structure.

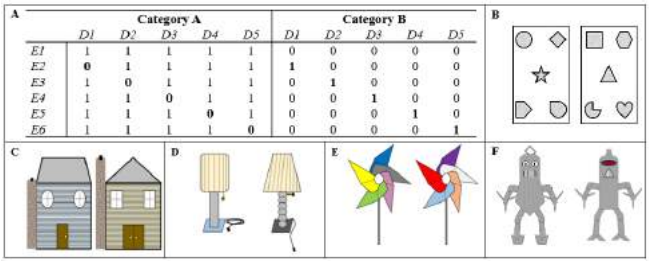


Figure 1: (A) The abstract FR structure [prototypes: row 1]; (B-F) Prototypes from all five stimulus sets: and prototypes of each category structure for all five stimulus sets, respectively: Card, House, Lamp, Pinwheel, Robot.

On each trial of the SCT, two completed sorts from the same stimulus set were presented side-by-side on the computer screen, one a FR solution and the other a UNI solution (see Figure 2 for an example trial). The UNI solution was generated on each trial of the experiment by swapping one-offs of a randomly selected row in the abstract structure, between category. Stimulus sets were presented in a randomized order. The order of items in each category of each completed sort, as well as the side on which FR was shown (left vs. right), were also randomized.

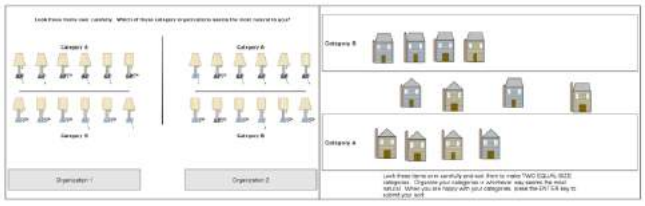


Figure 2. The left panel contains an example trial of the SCT containing a FR (right) and UNI organization (left; base color). The right panel contains an in-progress example of the sort task.

Procedure The SCT was programmed and administered via computer. Immediately prior to the task, participants were given instructions: "In this study you will be presented with a single set of items that is organized in two different ways. An example trial is pictured below. On each trial, carefully look at both organizations and select whichever one seems the most natural to you." The example trial shown with the instructions was identical to Figure 2, except the lamps were replaced with naturalistic ducks and bats – one organization had ducks and bats separated as categories A and B, while the other had them intermixed between categories. After the instructions, participants sequentially performed the SCT on each of the five domains. Participants selected their preference by clicking a button located below the organizations with the mouse. After responding, participants proceeded to the next trial/stimulus set.

Results & Discussion

We obtained a difference score for each participant that reflected their net preference (total FR selections minus total UNI selections). As there were five trials, difference scores could range from -5 (all UNI selections) to 5 (all FR selections). The analysis showed that FR organizations were selected reliably more frequently than UNI ones (the ordinal, non-normal data were subjected to a Wilcoxon signed-rank test: $Mdn = 1$, $Z = 3.787$, $p < .001$; see Figure 3). Supplemental analyses[1] provide a histogram of difference scores (pp. 1) and an analysis of preference by stimulus domain (pp. 2-3).

These results show clearly that when people's structural preferences are assayed – in the absence of being asked to produce candidate structures – they prefer FR. This provides inaugural evidence that FR may be a deep organizational basis that is sought in naïve, unsupervised category formation. Compared to the UNI bias typically seen in the full-array sorting paradigm, these findings represent a massive divergence – suggesting the affinity for UNI organizations in the sorting task is related to shortcomings in the generation of candidate structures. However, these stimuli have never been subjected to a direct comparison between the SCT and the full-array sorting task. It is possible that the stimuli we used are somehow generally more prone to be categorized by FR. This potential critique is addressed in Experiment 1B.

Experiment 1B

In this experiment, we sought to replicate the FR advantage and relate the outcomes of the SCT and the full-array sorting task using the same stimulus sets. To this end, each participant completed both the SCT and the full-array sorting task for all the stimulus sets in a task-blocked format (e.g., SCT for all sets, then sorting task for all sets). The order of the tasks was balanced across participants. We predicted participants would display a profound UNI bias in the sorting task, but that people would readily choose FR in the SCT.

Method

Participants 140 undergraduate students at Binghamton University participated. Participants were randomly assigned to the SCT first (SCT-SORT; $N = 71$) or the sort first (SORT-SCT; $N = 69$) condition.

Materials and Design The materials were identical to Experiment 1A. The same stimuli were used in both the SCT and sort task.

Procedure Both tasks were administered through a computer program. The SCT procedure was identical to that of E1A. In the SORT task, participants were instructed that there were many ways to create two equal-size categories, but their goal was to sort them in a way they thought most natural; an example sort, using the same demo images used

[1] The supplemental analyses can be viewed at this link: https://osf.io/jr2wu/?view_only=de559c73f1ef4b4da3781f4ef680f74f

for the SCT demo, was provided with the instructions. After instructions, participants then sorted each stimulus set in random order. For each stimulus set, the stimuli initialized to a row in the middle of the screen in a random order. Above and below that row were the two category bins. Participants used the mouse to drag and drop items into either category bin, and were able to reclassify the items freely. When finished, participants hit the enter key to submit. Sorts were coded as FR, UNI, or OTHER. OTHER sorts reflect any type of category produced by participants that is not FR or UNI; these sorts lack any interpretable structure. In the SORT-SCT group, participants sorted each of the randomly ordered stimulus sets before then completing the SCT for each set. The SCT-SORT group was the same, but the order of the two tasks was swapped.

Results & Discussion

The primary goal of the experiment was to replicate the FR advantage in the SCT; we use the same SCT analysis as in E1A. The SCT difference scores did not differ as a function of condition (SORT-SCT: *Mdn* = 1; SCT-SORT: *Mdn* = 1; Z = -0.41, p = .682). As such, SCT data from the two conditions were combined. Our analysis using the magnitude of the difference scores showed that participants selected the FR structure reliably more frequently than the UNI one (Wilcoxon signed-rank: *Mdn* = 1, Z = 2.98, p = .003; see Figure 3). Thus, we replicated the effect found in E1A and illustrate that when the task is constrained to the evaluation of candidate structures, FR is preferred over UNI solutions (see supplemental analyses pp. 2-3 for an analysis of preference by stimulus domain).

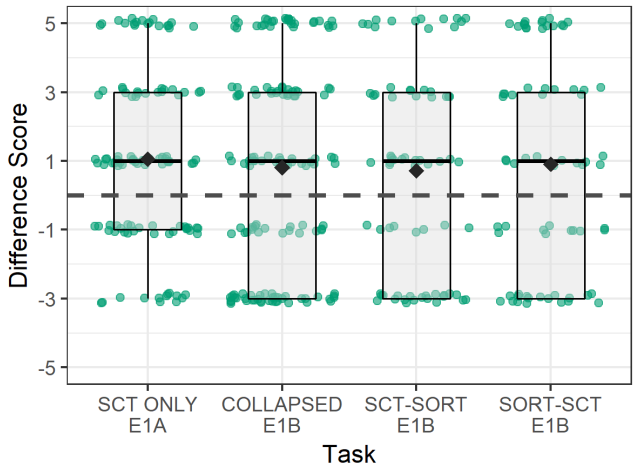


Figure 3. SCT preferences in Experiments 1A (left) and 1B (right three). COLLAPSED reflects an aggregation of both SCT-SORT and SORT-SCT conditions. Green dots reflect participant difference scores. Diamonds show means. Dashed line shows a difference score of 0 (FR = UNI).

The second goal of Experiment 1B was to compare the SCT to the standard full-array sorting paradigm, using the same stimuli as in the SCT. A potential concern from E1A is that the preponderance of FR responding in the SCT is a result of the stimuli, rather than a result of separating evaluation from production. As with the SCT data, we collapsed sort task data across conditions; the conditions did not reliably differ in the number of UNI solutions provided (SORT-SCT: 99%, SCT-SORT: 97%; $\chi^2(1, N = 593) = 0.002$, p = .967), and there were too few alternative solutions generated (FR/OTHER) to compare across order conditions. For the collapsed data, we replicated the prevalent UNI bias; participants produced significantly more UNI sorts (98%) than FR (0.2%) and OTHER sorts (1.8%) [UNI vs. FR: $\chi^2(1, N = 594) = 590.1$, $p < .001$; UNI vs. OTHER: $\chi^2(1, N = 604) = 560.8$, $p < .001$]. Moreover, more OTHER sorts were produced than FR sorts, reflecting the rarity of FR solutions produced in the sorting paradigm [$\chi^2(1, N = 12) = 8.333$, $p = .004$].

Overall, these findings show a successful replication of the FR preference under the SCT. By contrast – and consistent with previous research using the array sorting task – the sorting task led to an overwhelming number of UNI sorts. This highlights two points. First, there were only 12 non-UNI sorts produced; this means the very same people who produced consistent UNI sorts found FR to be more compelling than UNI in the SCT, under the same stimuli. This strongly suggests that the UNI bias in the sorting task arises not because participants evaluate UNI as a superior organizational principle (which the SCT data shows), but instead because some element(s) of the sorting task encourages it. Second, regarding the concern that our stimuli might generally be more prone to FR preference, the strong UNI bias in the full-array sorting task indicates that is not the case.

Experiment 2

The purpose of this experiment is to address an alternative account of the FR advantage observed and replicated in E1. Specifically, we were concerned that participants might have failed to notice the UNI rule and, in the (perceived) absence of a UNI rule, chose FR as a best of bad options (perhaps without appreciating the FR similarity).

In this experiment, we use a three-condition (within-subjects) version of the SCT. The first condition is the same FR vs. UNI choice task as in the previous experiments. However, we introduce two new conditions: UNI vs. OTHER and FR vs. OTHER. Note that the UNI vs. OTHER condition should make UNI a compelling option – provided people do notice the UNI rule in the SCT; a UNI over OTHER preference would thus suggest participants in the previous experiments did notice the UNI rule, but preferred FR. A potential concern about this design is that, by elevating UNI to an 'optimal' choice on those trials, it might invite a demand characteristic (e.g., 'UNI is better option here, maybe they want me to find/select the UNI option elsewhere') or make UNI organizations more appealing than they might otherwise be. The FR vs. OTHER condition was included to minimize these effects, as it serves to balance the number of times FR and UNI were 'optimal' vs.

OTHER groupings. In addition, it allows us to examine if FR is preferred over OTHER organizations and affords us a global measure of preference by making each response type equally probable due to chance (total FR vs. UNI, across all trial types).

Method

Participants 363 undergraduate students from Binghamton University participated.

Materials and Design The materials were the same as in Experiment 1. The design was like E1A, but expanded to include UNI vs. OTHER and FR vs. OTHER trial types, all within-subjects. OTHER organizations were created by taking a FR organization and swapping three randomly-chosen, non-prototype items from one category with three items occupying the same rows in the opposite category, according to the abstract structure shown in Figure 1; these arrangements had no discernible structure. In sum, there were three trial types: 1) FR–UNI, 2) FR–OTHER, and 3) UNI–OTHER. The particular OTHER and UNI groupings that were created for each subject were held constant across trial types within a domain (e.g., if the UNI rule was on the base color for the Lamp set [see Figure 2] in the FR-UNI trial, base color would also be used to form the UNI rule in the UNI-OTHER trial for Lamps). This was done to ensure a consistent comparison across trial types in a stimulus set.

Procedure The procedure was like Experiment 1A. Participants were presented with each stimulus set sequentially, in a random order. For each stimulus set, the participant was presented with each trial type sequentially, in a random order. In each trial type, the participant selected which of the two organizations they preferred. Unlike in the previous experiments, the participants were elicited for an explanation of their choice for each trial type in the first and last stimulus sets.

Results & Discussion

Separate difference scores were calculated for each trial type: UNI minus OTHER, FR minus OTHER, and FR minus UNI. Wilcoxon signed-rank tests were conducted on the magnitudes of the difference scores. Supplemental analyses include a histogram of difference scores (pp. 4) an analysis of preference by stimulus domain (pp. 5-6) and an analysis of FR-UNI trials based on the previous trial (pp. 7).

Our primary concern in this experiment was to determine whether people do in fact detect UNI rules in the SCT. The critical trial type for assessing this was UNI-OTHER; a UNI preference would suggest participants do detect the UNI rule. The analysis of the UNI-OTHER condition yielded a reliable UNI over OTHER preference (Wilcoxon signed-rank: $Mdn = 3$, $Z = 12.487$, $p < .001$). Importantly, the UNI preference provides strong evidence that people do in fact notice UNI rule embedded in UNI organizations and suggests that the FR over UNI preferences observed in E1A and E1B are not derived from participants simply failing to notice the rule.

The same pattern was observed in the FR-OTHER condition; participants selected significantly more FR organizations than OTHER organizations (Wilcoxon signed-rank: $Mdn = 3$, $Z = 13.195$, $p < .001$). As a complement to the UNI-OTHER analysis, this finding shows that people are sensitive to FR as a coherent organizational basis and find it compelling relative to less coherent options.

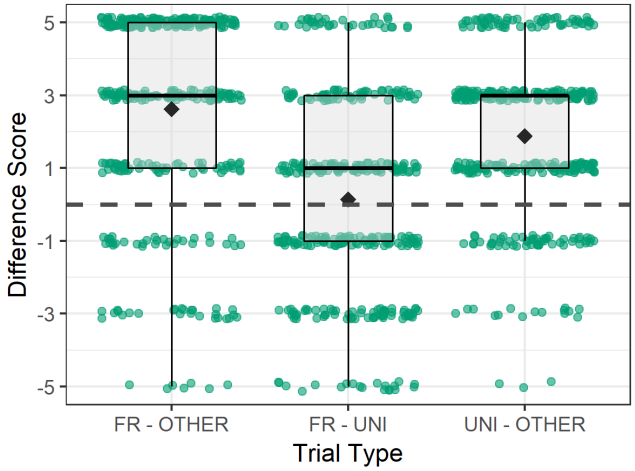


Figure 4. SCT preferences by trial type in Experiment 2. Green dots show each participant's difference score. Diamonds show means. Dashed line shows a difference score of 0. Positive scores in the FR-OTHER and FR-UNI trials reflect a FR preference; positive scores reflect a UNI preference in the UNI-OTHER trials.

Looking to the FR-UNI condition – a replication plus extension (given the novel trial types) – we failed to find the FR preference observed in E1A and E1B. The difference score magnitude did not differ from zero (Wilcoxon signed-rank: $Mdn = 1$, $Z = 0.851$, $p = .395$; see Figure 4). The failure to replicate is curious. One possibility is that the FR preference observed across E1A and E1B is a Type 1 error. However, both of those experiments were well-powered and the effect was replicated. Another possibility is that by introducing: (1) the new trial types; and/or, (2) the verbal explanations for preferences on the first (and last) stimulus set altered participants' behavior in the task. The use of OTHER organizations as a comparator effectively set up both UNI and FR organizations as 'correct' answers to the task. Then, on FR-UNI trials, the participant must decide which 'correct' answer is 'more correct'. Given UNI rules lend themselves more easily to verbal description (e.g., Zeithamova & Maddox, 2006), and given we asked participants for verbal descriptions, participants may have surmised that UNI was the 'more correct' choice and chose it more frequently than in the previous experiments. Upcoming studies will seek to disentangle these possibilities.

Lastly, we consider the global preference measure (all FR minus all UNI, across all trial types within subject). Consistent with the FR preference observed in the previous

two experiments, we found that people chose FR reliably more frequently than UNI organizations (Wilcoxon signed-rank: $Mdn = 1$, $Z = 2.543$, $p = .011$). This suggests that, despite not showing a preference for FR over UNI organizations on FR-UNI trials, participants did demonstrate an overall preference for FR when collapsing across all trial types.

In sum, we provided evidence that the FR over UNI preference observed across E1A and E1B is not attributable to people failing to notice the rule. Moreover, we provided additional evidence that people are sensitive to FR. Although the FR over UNI advantage did not replicate (potentially due to manipulations introduced in the current experiment), the FR over OTHER advantage indicates that people view FR as a meaningful organizing principle for categories.

General Discussion

The widespread UNI bias in unsupervised category formation – and its inconsistency with the FR structure of natural categories – has remained a question mark in the field for decades. In the experiments above, we approach the question from the perspective that the standard sorting paradigm encourages UNI responding by virtue of being an intentional, production-focused task that does not afford the requisite incidental exposure for learners to generate candidate structures as they might otherwise in naturalistic settings. To circumvent these issues, we introduced the SCT – a task that requires only the evaluation of provided candidate structures rather than both generation and evaluation.

Across two high-powered experiments, we observed a FR over UNI preference in the SCT – contesting the prevalent UNI bias under the sorting paradigm. This preference emerged even despite the use of knowledge-poor features (Medin et al., 1987), a full-array format (Regehr & Brooks, 1995), and despite the omission of a pre-task encoding phase (Lassaline & Murphy, 1996). Experiment 1B showed that the FR preference observed in the SCT is not due to the stimuli being generally FR-prone, as evidenced by the sorting task, and showed that people sort according to UNI rules, regardless of their SCT preference. In an extended form of the SCT, we showed in Experiment 2 that the FR preferences observed in the preceding experiments did not arise from a failure to identify the UNI rule in UNI organizations and provided further evidence that people view FR as a meaningful way of organizing categories.

These results are compelling for a number of reasons. First, they suggest that, at one extreme, people prefer FR over UNI structures in category formation (E1A & E1B). At the other extreme, they suggest that people do not have a preference between UNI and FR structures (E2). Regardless of which is true, the data show that people are sensitive to and appreciative of both types of category-level similarity, and this represents a massive departure from the strict UNI bias observed in – to our awareness – every unsupervised category learning study with domain-naïve participants and no additional encoding tasks (though see Pothos & Close, 2008; Pothos et al., 2011 which address multidimensional vs. UNI sorting – though not the specific tension between FR and UNI). As such, these findings present a potential alignment with the basis of similarity that apparently underlies natural categories (Rosch & Mervis, 1975).

Second, these results highlight the importance of task in the behaviors that are produced, afforded, and encouraged. Although we found evidence that people appreciate FR, in E1B we showed that people – many of which displayed a FR preference in the SCT – uniformly produced UNI sort solutions. Thus, by isolating a sub-component of the overall formation task, we found radically different outcomes. This suggests the discrepancy between natural and sort-task categories is tied to the generation of candidate structures and reinforces a need for researchers to examine phenomena with an array of task formats.

We note a few limitations of the present work that motivate future studies. First, though we found and replicated the FR over UNI advantage, this advantage failed to replicate in E2. In future work, we aim to disentangle if and how the additional manipulations are attributable. Second, while our data show that people prefer FR when given candidate structures, our study does not speak to how people might initially generate FR as a candidate structure in the first place. We speculate this might occur through incidental experience (without class prediction) that leads to knowledge of feature statistics, as is hinted at by previous work (Lassaline & Murphy, 1996; Love, 2002). In future work, we intend to assess if a novel, repeated, item-matching task (like MTS, but without supervision) leads to such knowledge that transfers to the sorting task. Finally, the SCT involves an absolute, forced judgment between candidate structures. In future work, we plan to: (1) assess the degree to which people prefer one structure over another using a rating scale to determine if FR is viewed as a compelling way to structure categories as opposed to the better of two poor options; and, (2) afford the option of 'no judgment' to gain greater fidelity in our results.

References

Ahn, W. K. (1999). Effect of causal structure on category construction. *Memory & Cognition*, *27*(6), 1008-1023.

Ahn, W. K., & Medin, D. L. (1992). A two-stage model of category construction. *Cognitive Science*, *16*(1), 81-121.

Lassaline, M. E., & Murphy, G. L. (1996). Induction and category coherence. *Psychonomic Bulletin & Review, 3,* 95-99.

Levering, K. R., & Kurtz, K. J. (2015). Observation versus classification in supervised category learning. *Memory & Cognition*, *43*(2), 266-282.

Love, B. C. (2002). Comparing supervised and unsupervised category learning. *Psychonomic Bulletin & Review, 9(4),* 829-835.

Medin, D. L., Wattenmaker, W.D., & Hampson, S. E. (1987). Family resemblance, conceptual cohesiveness,

and category construction. *Cognitive Psychology, 19,* 242-279.

Milton, F., Longmore, C. A., & Wills, A. J. (2008). Processes of overall similarity sorting in free classification. *Journal of Experimental Psychology: Human Perception and Performance, 34*(3), 676-692.Milton, F., & Wills, A. J. (2004). The influence of stimulus properties on category construction. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 407-415.

Patterson, J.D., & Kurtz, K.J. (2018). Semi-supervised learning: A role for similarity in generalization-based learning of relational categories. In C. Kalish, M. Rau, J. Zhu, T. Rogers (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp 2211-2217). Austin, TX: Cognitive Science Society.

Pothos, E. M., & Close, J. (2008). One or two dimensions in spontaneous classification: A simplicity approach. *Cognition, 107,* 581-602.

Pothos, E. M., Perlman, A., Bailey, T. M., Kurtz, K., Edwards, D. J., Hines, P., & McDonnell, J. V. (2011). Measuring category intuitiveness in unconstrained categorization tasks. *Cognition, 121,* 83-100.

Regehr, G., & Brooks, L. R. (1995). Category organization in free classification: The organizing effect of an array of stimuli. *The Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 347-363.

Rosch, E., & Mervis, C. G. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology, 7,* 573-605.

Wattenmaker, W. D. (1992). Relational properties and memory-based category construction. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18, 1125-1138.

Wills, A. J., Milton, F., Longmore, C. A., Hester, S., & Robinson, J. (2013). Is overall similarity classification less effortful than single-dimension classification?. The *Quarterly Journal of Experimental Psychology, 66*(2), 299-318.

Wittgenstein, L. (1953). Philosophical Investigations. Philosophische Untersuchungen. Oxford, England: Macmillan.

Zeithamova, D., & Maddox, W. T. (2006). Dual-task interference in perceptual category learning. *Memory & Cognition, 34*(2), 387-398.

Subjective Randomness in a Non-cooperative Game

Michael Payton (mjpayton@wisc.edu)
Jeffrey C. Zemla (zemla@wisc.edu)
Joseph L. Austerweil (austerweil@wisc.edu)
Department of Psychology, University of Wisconsin Madison
1202 W. Johnson Street
Madison, WI 53706 USA

Abstract

Rock, Paper, Scissors (RPS) is a competitive game. There are three actions: rock, paper, and scissors. The game's rules are simple: scissors beats paper, rock beats scissors and paper beats rock (all signs stalemate against themselves). Over multiple games with the same opponent, optimal play according to a Nash Equilibrium requires subjects to play with genuine randomness. To examine randomness judgments in the context of competition, we tested subjects with identical sequences in two conditions: one produced from a dice roll, one from someone playing rock, paper, scissors. We compared these findings to models of subjective randomness from Falk and Konold (1997) and from Griffiths and Tenenbaum (2001), which explain assessments of randomness as a function of algorithmic complexity and statistical inference, respectively. In both conditions the models fail to adequately describe subjective randomness judgements of ternary outcomes. We also observe that context influences perceptions of randomness such that some isomorphic sequences produced from intentional play are perceived as less random than dice rolls. We discuss this finding in terms of the relation between patterns and opponent modeling.

Keywords: Randomness, pattern recognition, opponent modeling

Introduction

Humans often detect patterns in everyday life—so much so that they often attribute spurious patterns (generated by a random mechanism) to intentional actions. For example, in the *gambler's fallacy* (Kahneman & Tverksy, 1972), people believe that "a Red is due" after observing several Black rolls from a roulette wheel, despite the rolls being independent from one another. Conversely, according to the *hot hand effect*, if a player has scored several free throws in a row in basketball practice, people believe that player is more likely to score again, despite this not being true empirically (Gilovich, Vallone, & Tversky, 1985). Why are the patterns that people detect sensitive to their context?

Psychologists have approached these phenomena in terms of *subjective randomness*, or the perceived randomness of observations. Previous literature has shown that in some circumstances, people tend to judge sequences as random even when the underlying pattern is systematic. In a classic example, people tend to believe that a sequence of coin flips will have more alternations (e.g., heads followed by tails) and fewer streaks (e.g., several heads in a row) than is likely to occur in a sequence produced by flipping an unbiased coin repeatedly (Falk & Konold, 1997). To date, much of the literature on subjective randomness has focused on binary sequences or grids generated from a truly random mechanism, such as a coin flip, an animate mechanism (Ayton & Fischer, 2004), or a human or other intentional agent (Burns & Corpus, 2004; Caruso, Waytz, & Eply, 2010).

In this article, we compare the subjective randomness of sequences generated from a die roll to comparable sequences generated by a player who is in direct competition with another player in a non-cooperative game: Rock, Paper, Scissors (RPS; also called RoShamBo). We had two hypotheses: (1) sequences generated from a random mechanism (a die roll) would be judged as more random than equivalent sequences generated from a human playing RPS, and (2) "complex" sequences generated from a human playing RPS would be perceived even less random due to opponent modeling in a competitive context.

The rules of RPS are straightforward. Two players simultaneously present one of three hand signs, "rock", "paper", or "scissors". The scoring of the game is also simple: scissors beats paper, rock beats scissors, and paper beats rock (all signs stalemate against themselves). From a game-theoretic perspective, the Nash equilibrium (Nash, 1950) is generating signs uniformly at random. Thus, if people played according to the Nash equilibrium, they would be required to produce truly random sequences. This strategy would prevent any player from gaining advantage over another after playing repeated games over time. However, this is unlikely as people are poor at producing random sequences (Baddeley, 1966, Towse, 1998).

Although RPS may appear to be a simple game, it is actually much more complex than one might first think. For instance, while one might expect that the winners of RPS are determined by luck or chance, there are genuine RPS masters. RPS tournaments have been held throughout the world where experienced RPS players will consistently outplace novices (Hegen, 2004). One might at first think that this is simply due to extraneous factors. For example, perhaps one player produces their sign slightly before the other and the other player uses that information to change their play (note that this is illegal in tournament play). However, it is reported that a player in 2001 was allowed to bring a random number generator to inform his sequence generation. He failed to even make the qualifying rounds in the regional tournament (Hegan, 2004).

Further, there have also been machine RPS competitions (Billings, 1999), where researchers submitted automated RPS agents or "bots" to play against each other. Even when only bots compete against each other (no human players), certain

strategies are far more advantageous than others (Billings, 2000). In fact, a bot playing the Nash equilibrium using a random number generator tends to score poorly in these competitions. The results of human and machine tournaments naturally lead to the question of how opponent modeling of intentionally produced sequences might affect subjective judgments of randomness.

In a recent study of multiple repeated games of RPS between human players, Wang (2014) round that a Nash Equilibrium was never obtained by any subset of the population of participants. Rather, successful players often employ a 'win-stay, lose switch' strategy which is beneficial in identifying patterns in another's strategy, exploiting them and also retaining a fail-safe strategy which prevents repeated loses. This is notable as win-stay lose-switch has also been proposed as an explanation of human category learning (Restle, 1962), and recently has been shown to approximate Bayesian inference in some cases (Bonawitz Denison, Gopnik, & Griffiths, 2014).

Below we outline two cognitive models of subjective randomness and introduce an experiment to test their robustness in explaining ternary sequences in competitive and non-competitive environments. We close by discussing the implications and limitations of the experiment in furthering our understanding of subjective randomness.

Models of Subjective Randomness

People's judgments of randomness notoriously deviate from the prescriptions of formal probability theory in systematic ways. In experimental settings, subjects are less likely to agree that a set of coin flips that come up HHHHHHHHHH are random when compared to a set of flips that came up HTTTHTHHT. Yet, both sets of results are exactly as likely as the other given a fair coin. Even after learning probability theory, it is hard for people to escape the intuition that the latter *feels* more random than the former. How do we explain this intuition?

One popular way to model human deviations from a straightforward probabilistic account is to assume randomness judgments are a function of how difficult it is to encode a sequence or its "complexity". In these models, psychologists try to identify an encoding process or measure of sequence complexity by specifying a theoretically motivated model that is correlated with subjective ratings. Below, we discuss two prominent models from the literature.

Falk and Konold (1997). Building on an intuition from Kahenman and Tversky (1972), Falk and Konold (1997) proposed that people 'chunk' a sequence into smaller subsequences which are easier to encode and remember. The perceived randomness of the sequence is inversely related to the ease with which humans can divide sequences into fewer, more manageable subsequences. To quantify this process, Falk and Konold (1997) developed their model, the Difficulty Predictor (DP), to define the complexity of a sequence to be a function of the number of runs (subsequences with the same outcome) and alternations (subsequences which switch between two outcomes repeatedly). For example, the sequence "XXOXOX" can be described as "X twice, OX twice." Each sequence can be encoded in terms of runs and alternations. The DP of a sequence is the sum of the number of runs and two times the number of alternations. For example, the above sequence would assign one point for a sequence length of one for the first subsequence ("X") and a score of two for a sequence length of two in the second subsequence ("OX"). The DP for this sequence is three.

Only the smallest repeating unit is needed to calculate the score for a subsequence. As such "OXOX" and "OXOXOXOX" are both given a score of two points. Any given sequence can be apportioned many different ways: for example, the sequence "XOXOO" can be described a "XO twice, O once" (DP of 3) or "XOX once, O twice" (DP of 4). DP is the minimal score over possible encodings of a sequence.

This formalization of subjective randomness instantiates the concept of Kolmogorov complexity (Kolmogorov, 1965), which states that the complexity of an object is the length of the shortest program that can be used to generate that object. Previous work in psychology suggests that humans are adept at finding patterns in data and encoding them in a way consistent with Kolmogorov complexity (Chater, 1996, 1999). In fact, Griffiths et al. (2018) showed that DP is a special case of the complexity producing the sequence on a finite state machine with four motifs: all Hs, all Ts, alternating HTs, and alternating THs. The machine is biased to stay in its current motif.

Griffiths and Tenenbaum (2001) Griffiths and Tenenbaum (2001) propose an alternative model of subjective randomness by realizing that randomness judgments are not made in a vacuum: The randomness of a sequence is its relative likelihood of having been generated from a random rather than a *regular* process. Thus,

$$\text{random}(x) = \frac{P(\text{random}|x)}{P(\text{regular}|x)} = \frac{P(x|\text{random})}{P(x|\text{regular})}\frac{P(\text{random})}{P(\text{regular})}$$

Their Bayesian model then differs from the standard normative account, which only considers the likelihood of the sequence assuming a random generating process, or random(x) = P(x|random). They implement a Bayesian model that captures the likelihood of a sequence being generated by a random process compared to a regular (non-random) process. Here, we generalize their model from binary sequences to the ternary sequences that we use in our experiment.

The probability of a ternary sequence x of length N being generated by a random process is:

$$P(x|\text{random}) = (1/3)^N$$

In contrast to a random sequence, we define a regular sequence as one that is generated by a systematic process in which each token in a sequence is generated by a multinomial

process with parameter vector $\vec{\theta}$. $\vec{\theta}$ gives the probability of each type, and so θ_1 would be the probability of the first type. Because the multinomial parameters are unknown for "regular processes", an ideal observer should consider all possible parameter combinations:

$$P(x|\text{regular},\vec{\alpha}) = \int P(x|\vec{\theta})P(\vec{\theta}|\vec{\alpha})d\vec{\theta}$$

where $\vec{\alpha}$ represents the parameters for the prior distribution. We use the Dirichlet distribution due to its conjugacy with the Multinomial distribution. Integrating over all possible values for $\vec{\theta}$, we find:

$$P(x|\text{regular},\vec{\alpha}) = \frac{\Gamma(A)}{\Gamma(\text{N}+\text{A})}\prod_{k=1}^{K}\frac{\Gamma(n_k+\alpha_k)}{\Gamma(\alpha_k)}$$

where $\Gamma(x)$ is the Gamma function evaluated at x, n_k denotes the number of tokens of type k in a sequence and $A = \sum \alpha_k$. Assuming equal prior odds, $P(x|\text{random}) = P(x|\text{regular})$, the complexity of a sequence can then be defined as the log-likelihood that it was generated by a random process, as opposed to a regular process:

$$\text{LR} = \log\frac{P(x|\text{random})}{P(x|\text{regular},\vec{\alpha})}$$

Sequences with a LR greater than zero are more likely to have been generated by a random process, whereas sequences with a LR less than zero are more likely to have been generated by a regular process than a random process. Prior odds can be included in the model to shift the boundary between regular and random to a value other than zero. See Williams and Griffiths (2013) for additional empirical support of this model in capturing human randomness judgments for binary sequences.

Experiment

In previous work and both models, randomness judgments are not made in the context of two intentional human agents directly competing where the result of their competition is based on their joint decisions. Motivated by these considerations, we ask: how does a sequence being generated within a competitive context affect its perceived randomness?

In the present study we examine how well these models explain ternary sequences, rather than binary ones. We also manipulate conditions of how the sequence is assumed to be generated: either by a person playing RPS (competitive) or by the roll of a die (neutral).

Materials and Methods

We collected data from 148 subjects on Amazon Mechanical Turk. We excluded 42 subjects (28%) who had a mean response time of less than 800ms in either condition. This minimum average response time was based on an estimate of how long a subject would need to view the sequence, encode any perceived patterns and make a motor response. The data presented here reflect the remaining 106 subjects (mean age 37.3, 53 male, 52 female, 1 unknown). Each subject saw 100 sequences (sequentially) in each of the two conditions: die (neutral context) and RPS (competitive context).

In the die condition, subjects were told that a friend was playing a board game with a six-sided die that had two blue faces, two yellow faces, and two red faces. On each trial, the subject observed a sequence of seven rolls from that die.

In the RPS condition, subjects were told that they were watching two friends play a game of rock, paper, scissors. On each trial, the subject observed a sequence of seven hand gestures from the game. (See Figure 1.)

Figure 1. (A) An example sequence from the die condition. (B) An example sequence from the RPS condition that is conceptually identical to the die sequence.

There are 2,187 possible ternary sequences of length seven. We assumed that the perceived randomness of individual classes was irrelevant. e.g. that a die sequence BLUE YELLOW YELLOW is perceived as equally random as YELLOW BLUE BLUE. This reduces the pool of sequences to 729. For each subject, sequences were randomly selected without replacement from the 729 possible sequences. Images were assigned randomly to the three types so that all 2,187 sequences were observed.

The two conditions were blocked so that subjects saw 100 trials from one condition, followed by 100 trials from the other condition. The starting condition (RPS or die) was counterbalanced between subjects. The order of trials within a condition was random, but identical for both conditions for each subject (e.g., if a subject saw sequence A from Figure 1 as trial 1, they might see the isomorphic sequence B from Figure 1 as trial 101).

Subjects rated each sequence on a Likert scale from 1 ("Not random at all") to 10 ("Very random"), with midpoint label of "Somewhat random." Following the experiment, subjects completed a brief demographic survey that also included questions about their level of education, experience playing rock paper scissors, and whether they had taken a statistics or probability course.

Results

We began with two hypotheses: (1) sequences generated from a random mechanism (a die roll) would be judged as more random than equivalent sequences generated from a human playing RPS, and (2) high alternation sequences generated by an intentional agent in the context of a game

would be perceived even less random due to opponent modeling in a competitive context.

To test the effect of sequence production on a subject's randomness judgments, we first compared the scores of reported randomness between conditions. We found a significant effect, with sequences produced from a die being considered more random than sequences produced by games of rock, paper, scissors, $M_{die} = 5.78$, $M_{RPS} = 5.47$, $t(105) = 2.93$, $p = .004$. This confirms our first hypothesis that participants perceive outcomes produced by people to be more random than those produced by a die.

Though individual subjects varied greatly in their mean scores of subjective randomness, there is a clear trend towards evaluating sequences from the RPS condition as more random than the sequences produced from the dice condition. See Figure 2.

We expanded this analysis further by examining whether randomness judgments are partially explained by the difficulty of encoding a sequence to memory (as they were in Falk and Konold, 1997). Using response time as a proxy for encoding difficulty, we found that there was a significant effect for reaction time between conditions, $M_{die} = 2372$ms, $M_{RPS} = 3091$ms, $t(105) = 4.45$, $p < .001$. This means subjects took longer to respond to randomness judgements in the RPS condition compared to the dice condition.

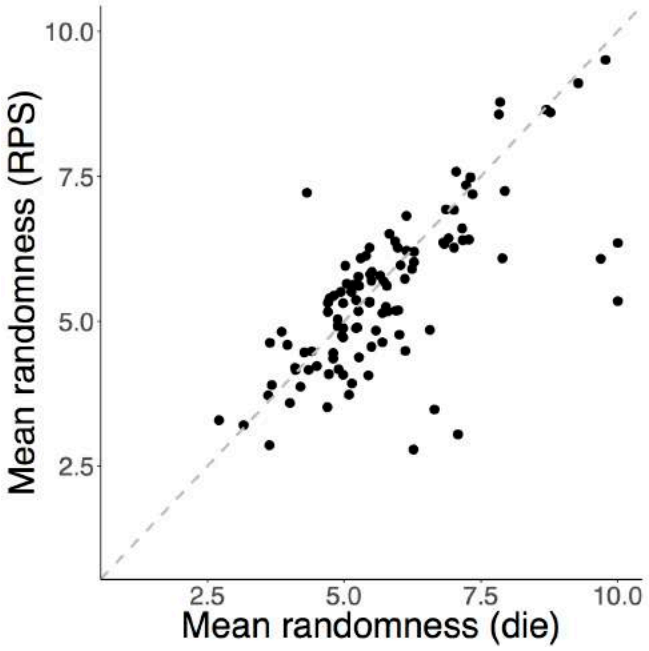


Figure 2. Each point denotes an individual subject's mean randomness judgment for the die condition (x-axis) and RPS condition (y-axis). The identity line is shown for comparison.

We calculated the complexity of each sequence according to three measures: Falk and Konold (1997)'s difficulty predictor (DP), Griffiths and Tenenbaum (2001)'s Likelihood Ratio (LR), and the probability of alternation. Overall, all three measures were highly correlated with subjective randomness judgments (see Figure 3). We found a difference in subjective randomness judgments by condition that was moderated by the complexity of the sequence. Sequences of low complexity (as judged by either DP, LR, or probability of alternation) were judged to be equally non-random regardless of whether the sequence was in the die or RPS condition. However sequences of high complexity were judged to be more random in the die condition compared to the RPS condition. We discuss this further later in the article.

Discussion

Classic studies on subjective randomness have had subjects judge binary (e.g., heads and tails, black and white tiles), or digit sequences. The pattern of performance described by Falk and Konold (1997) is that subjects will overestimate the number of alternations that would need to be present in a truly random sequence. In a sequence of tosses from a fair coin, we expect that a genuinely random sequence has a probability of alternation of 0.5. While subjects studying coin flips might overestimate the number of alterations in a given sequence, a fully alternating sequence would not be seen as random but following a predictable alternating pattern.

This contrasts sharply with the current findings, where a truly random sequence would have a probability of 0.67. Subjects continue to overestimate the number of alterations within a random sequence, but they do so without showing a decline towards less randomness at higher alteration values. This is because using the probability of alternation is not as useful as a measure in the case of ternary sequences. For example, the sequence RPRPRP has the same probability of alternation as the sequence RPSPSR, though the latter appears more random. In a binary sequence, an "alternation" implies what the next item in the sequence will be, but this is not true for ternary sequences. This highlights a limitation of using probability of alternation as a proxy for subjective randomness judgments.

We found that on average, a sequence of die rolls was judged to be more random than an equivalent sequence of rock paper scissors throws. This effect seems to be driven by higher judgments of randomness for high-complexity sequences in the die condition compared to the RPS condition. Currently, no model adequately describes why this difference between conditions might occur, or why the differences between conditions should be primarily observed in high complexity sequences.

There are several potential explanations for these trends. One possibility is that that randomness judgments are primarily influenced by the mechanism that generates that sequence, rather than the sequence itself.

The fact that RPS throws are the product of intentional action, while die outcomes are generated by chance is a promising hypothesis. Caruso, Waytz, and Epley (2010) explored this type of intentional action as a possible explanation for differences between the hot hand effect and the gambler's fallacy. They found that participants who were told to focus on the intentions of a coin tosser were more likely to expect a coin toss streak to continue compared to

participants who were told to focus on the motor actions of the tosser. However participants were never asked explicitly to judge the randomness of sequences and it was not a directly competitive context. Similarly, Ayton and Fischer (2004) tested whether differences in gambler's fallacy and hot hand might be accounted for by animacy in the generation process. Neither of these mechanisms alone explain the results observed in our study where only high complexity sequences appear to show differences in randomness ratings.

A related explanation is that subjects may be reluctant to use the upper end of the randomness scale in the RPS condition because they are explicitly told that the sequences were generated by a human, and their belief that humans cannot (or do not) produce truly random sequences. This interpretation is anticipated by Burns and Corpus (2004) who found that subjects expect streaks to continue if they are generated by a non-random process ie: a human player. Therefore subjects might perceive sequences with high rates of alternation as less likely.

There is some counterevidence to this hypothesis in our results: z-scoring each participants' ratings does not eliminate the lower randomness ratings specific to more complex sequences..

A second hypothesis is that in the context of playing a game of RPS, subjects expect to see more complex sequences. A competent rock, paper, scissors player should try to make each throw as unpredictable as possible in order to beat his or her opponent. Therefore, we should expect a player to generate complex sequences intentionally. Subjects may have judged highly complex sequence as less random in the RPS condition because they believe a player planned that sequence in order to fool their opponent. Somewhat paradoxically, this means that sequences that are descriptively more random are seen as less random, due to the fact that they are unsurprising in the context of the game. This distinction between descriptive complexity and observed complexity has been used to explain, for instance, why descriptively simple lottery results (such as 1-2-3-4-5) are seen as more surprising (Dessalles, 2017).

Thirdly, RPS presents a sequence in a two-player game. This may lead subjects to underestimate randomness by urging them to look more closely for possible subtle patterns in the sequences generated by opponent modeling. In the die condition, each roll is assumed to be independent of the previous roll. But in the RPS condition, each throw may be conditionally dependent not only on the player's previous throw, but also the opponent's previous throw. This naturally leads to a larger hypothesis space from which subjects may be inferring potential patterns. This expansion of the hypothesis space could disproportionately affect more complex sequences and therefore explain the observed differences between low and high complexity sequences.

Hypotheses 2 and 3 operate under the assumption that expectations based on both context and generation method contribute to perceptions of randomness. While previous studies have shown that generation method plays a role, they do not explicitly contrast between a sequence produced in a directly competitive vs. non-competitive contexts. An aim of future research will be to understand how generation process, context and the complexity of a sequence may interact in order to explain the current results.

One limitation of the current study is that subjects may have a biased prior belief that die rolls are more random than RPS sequences, independently of the likelihood of a given sequence. Future studies may explicitly equate these priors. For instance, subjects could be shown two sequences of die rolls and informed that one sequence was generated by a fair die (random) and the other by a weighted die. Identifying the "cheater" in this case depends only on the likelihood, as the experiment can be designed so that the prior probability of each die is equal (0.5).

Another limitation is that our stimuli consist only of sequences of length 7, and each unit in the sequence can only be one of three possible types. It is not clear whether our results extend to longer sequences, and to multinomial sequences beyond three types. We also do not account for perceptual similarity in our stimuli: the die images in our experiment are similar to each other (except for color), and

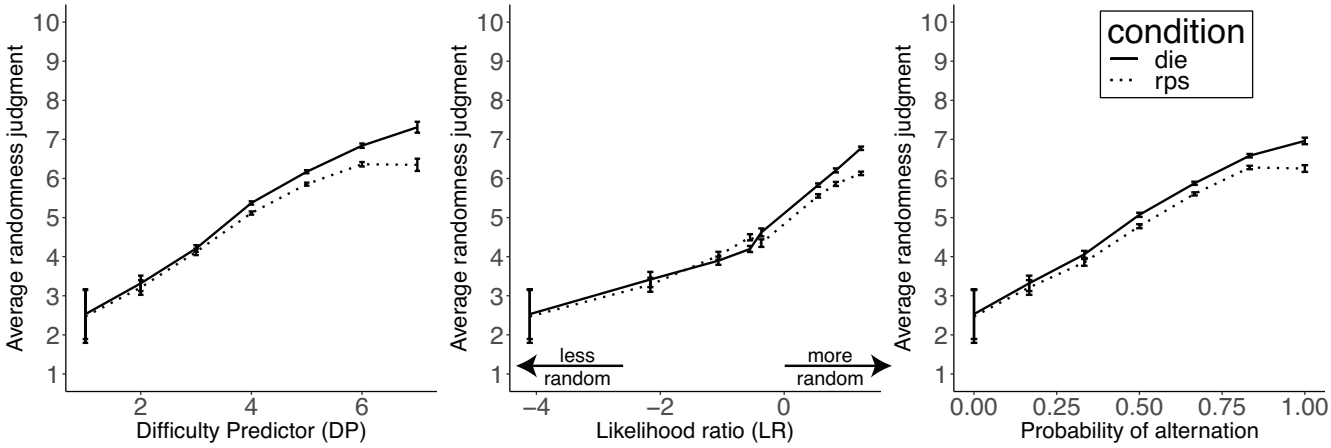


Figure 3. Average subjective randomness ratings were correlated with Falk and Konold (1997)'s Difficulty Predictor (left), Griffiths and Tenenbaum (2001)'s Likelihood Ratio (center) and the probability of alternation (right).

participants may be sensitive to perceptual similarity when assessing sequences.

Humans are notoriously poor at inferring randomness from sequences. This cognitive error seems to be exacerbated in competitive contexts. However, this might just as easily be reframed in a different light: People are more attuned to possible patterns of behavior when they are inspecting it within a competitive context. This may lead them to be less likely to write off certain patterns as 'mere luck' when they might carry valuable adaptive information for future planning and strategizing.

Acknowledgements

We would like to thank Amber Nomani for providing illustrations of the RPS hand signs and 3 anonymous reviewers for their thoughtful comments on an earlier draft of this paper. This work was funded by the Office of the VCRGE at University of Wisconsin-Madison with funding from the WARF

References

Ayton, P., & Fischer, I. (2004). The hot hand fallacy and the gambler's fallacy: Two faces of subjective randomness?. *Memory & cognition*, *32*(8), 1369-1378.

Baddeley, A. D. (1966). The capacity of generating information by randomization. *Quarterly Journal of Experimental Psychology, 18,* 119-129.

Billings, D. (1999). The first international RoShamBo programming competition. *ICGA Journal*, *23*(1), 42-50.

Billings, D. (2000). Thoughts on roshambo. *ICGA Journal*, *23*(1), 3-8.

Bonawitz, E., Denison, S., Gopnik, A., & Griffiths, T. L. (2014). Win-Stay, Lose-Sample: A simple sequential algorithm for approximating Bayesian inference. *Cognitive Psychology, 74,* 35-65.

Burns, B. D., & Corpus, B. (2004). Randomness and inductions from streaks:"Gambler's fallacy" versus" hot hand ". *Psychonomic Bulletin & Review*, *11*(1), 179-184.

Caruso, E. M., Waytz, A., & Epley, N. (2010). The intentional mind and the hot hand: Perceiving intentions makes streaks seem likely to continue. *Cognition*, *116*(1), 149-153.

Chater, N. (1996). Reconciling simplicity and likelihood principles in perceptual organization. *Psychological Review*, *103*(3), 566-581.

Chater, N. (1999). The search for simplicity: A fundamental cognitive principle? *The Quarterly Journal of Experimental Psychology: Section A*, *52*(2), 273-302.

Billings, D. (1999). The first international RoShamBo programming competition. *ICGA Journal*, *23*(1), 42-50.

Dessalles, J. L. (2017, July). Conversational topic connectedness predicted by Simplicity Theory. In *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 1914-1919).

Gilovitch, T., Vallone, R., & Tversky, A. (1985). The hot hand in basketball: On misperception of random sequences. *Cognitive Psychology, 17*, 295-314.

Griffiths, T. L., & Tenenbaum, J. B. (2001). Randomness and coincidences: Reconciling intuition and probability theory. In *Proceedings of the 23rd Annual Conference of the Cognitive Science Society* (pp. 370-375). Edinburgh: University of Edinburgh.

Griffiths, T. L., Daniels, D., Austerweil, J. L., & Tenenbaum, J. B. (2018). Subjective randomness as statistical inference. *Cognitive Psychology, 103,* 85-109.

Hegan, K. (2004). Hand to Hand Combat. *Rolling Stone*, (940), 38-38.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3*, 430-454.

Kolmogorov, A. N. (1965). Three approaches to the quantitative definition of information. *Problems of Information Transmission, 1,* 1-7.

Nash, J. F. (1950). Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences*, *36*(1), 48-49.

Restle, F. (1962). The selection of strategies in cue learning. *Psychological Review, 69*, 329-343.
Towse, J. N., & Neil, D. (1998). Analyzing human random generation behavior: A review of methods used and a computer program for describing performance. *Behavior Research Methods, Instruments, & Computers*, *30*(4), 583-591.

Walker, D., & Walker, G. (2004). *The official rock paper scissors strategy guide*. Simon and Schuster.

Wang, Z., Xu, B., & Zhou, H. J. (2014). Social cycling and conditional responses in the Rock-Paper-Scissors game. *Scientific reports*, *4*, 5830.

Williams, J. J., & Griffiths, T. L. (2013). Why Are People Bad at Detecting Randomness? A Statistical Argument. Journal of Experimental Psychology: Learning, Memory, and Cognition, *39*(5), 1473-1490.

Modelling mental imagery in the ACT-R cognitive architecture

David Peebles (d.peebles@hud.ac.uk)
Department of Psychology, University of Huddersfield
Queensgate, Huddersfield, HD1 3DH, UK

Abstract

I present a novel approach to modelling spatial mental imagery within the ACT-R cognitive architecture. The proposed method augments ACT-R's representation of visual objects to enable the processing of spatial extent and incorporates a set of linear and affine transformation functions to allow the manipulation of internal spatial representations. The assumptions of the modified architecture are then tested by using it to develop models of two classic mental imagery phenomena: the mental scanning study of Kosslyn, Ball, and Reiser (1978) and mental rotation (Shepard & Metzler, 1971). Both models provide very close fits to human response time data.

Keywords: Mental imagery; Mental rotation; Image scanning; ACT-R; Cognitive architectures.

Introduction

Mental imagery plays a crucial role in many aspects of cognition, from problem solving, creativity and scientific discovery to psychological disorders such as post-traumatic stress disorder, social phobia and depression (Kosslyn, Thompson, & Ganis, 2006; Pearson, Deeprose, Wallace-Hadrill, Burnett Heyes, & Holmes, 2013). Mental imagery has also been the subject of one of the longest running and fiercest debates in cognitive science (Kosslyn & Pomerantz, 1977; Pylyshyn, 1973; Anderson, 1978; Tye, 2000) and the nature of the mental representations and processes underlying mental imagery is still a subject of contention.

Two related issues concern the degree to which mental representations bear some structural correspondence to what they represent and whether mental imagery is supported by abstract, amodal propositional representations or depictive representations grounded in perception. In contrast to abstract propositional representations, imagistic visual representations depict rather than describe what they represent and retain the spatial relationships of their referents by having elements with geometric properties organised topographically (Reisberg, 2013).

This debate has been—and continues to be—driven and informed by the various attempts to provide formal computational accounts of mental imagery phenomena (e.g., Glasgow & Papadias, 1992; Kunda, McGreggor, & Goel, 2013; Tabachneck-Schijf, Leonardo, & Simon, 1997; Just & Carpenter, 1985) and the issue of whether imagery requires some form of array based representation or can be accomplished by more abstract, amodal representations and processes.

An early and influential cognitive model that combined pixel array based representations and more abstract representations is the CaMeRa model of expert problem solving with multiple representations (Tabachneck-Schijf et al., 1997). A more recent example is a model of problem solving on the Raven's Progressive Matrices test by Kunda et al. (2013) using 2D arrays of grayscale pixels and associated transformation operations. Using only these representations and processes, the model is able solve between 55% and 63% of Standard Progressive Matrices problems.

Figure 1: Stimulus used by Kosslyn et al. (1978).

Mental imagery in cognitive architectures

In recent years there have been a number of attempts to develop computational accounts of mental imagery from within the assumptions and constraints of *cognitive architectures* (e.g., Rosenbloom, 2012; Wintermute, 2012). Cognitive architectures are theories of the core memory and control structures, learning mechanisms, and perception-action processes required for general intelligence and how they are integrated into a "system of systems" to enable human cognition and autonomous, human-level artificial cognitive agents.

The cognitive architecture with one of the most well developed and comprehensive set of representations for spatial reasoning and visual imagery is Soar (Laird, 2012) and its *Spatial/Visual System* (SVS) (Lathrop, Wintermute, & Laird, 2011; Wintermute, 2012). The SVS system contains two layers of representation: a *visual depictive* layer (a bitmap array representation of space and the topological structure of objects), and a *quantitative spatial* layer (an amodal symbolic/numerical representation of objects and their spatial coordinates, location, rotation and scaling)[1].

SVS also contains operations to transform the continuous information in the quantitative spatial layer into symbolic information that can be used by Soar for reasoning. These pro-

[1] In the current (9.6.0) version of Soar, the visual depictive level has been omitted from SVS.

cesses allow Soar agents to perform mental imagery operations that can manipulate the representations and then extract spatial relationships from the modified states.

Several proposals have been put forward to endow the ACT-R cognitive architecture (Anderson, 2007) with spatial abilities. For example Gunzelmann and Lyon (2007) outlined an extensive proposal for modelling a range of spatial behaviour (including imagery) by augmenting the architecture with a spatial module and several additional buffers and processes for transforming spatial information. These proposals have, as yet, not been implemented however and so it remains to be seen whether the suggested changes would be able to account for human spatial competence.

An alternative approach to providing ACT-R with spatial capacities is the ACT-R/E project to embody ACT-R in robots (Trafton et al., 2013). ACT-R/E incorporates the *Specialized Egocentrically Coordinated Spaces* (SECS) framework (Trafton & Harrison, 2011; Harrison & Schunn, 2002) which adds modules for three aspects of spatial processing: 2D-retinotopic space, configural space for navigation and localisation, and manipulative space for the region that can be grasped by the robot.

Both of these approaches are broad in the sense that they propose extensive changes to the architecture (i.e., new modules and buffers) and seek to endow ACT-R with a wide range of spatial capabilities related to different spaces (Montello, 1993). Neither approach has modelled spatial imagery however. The aim of the study reported here is to fill this gap by developing ACT-R models of human spatial imagery behaviour. The approach adopted here is more limited and focussed than those discussed above in that it does not propose new modules or buffers but seeks to determine whether the phenomena can be accounted for with only minor adjustments to the existing structures and assumptions of ACT-R.

In the following sections I describe the relevant structures and assumptions of ACT-R and the adaptations required to allow the architecture to model spatial imagery. I then test the approach by using it to develop two models of well known mental imagery phenomena: mental scanning and mental rotation. Finally I discuss the implications, strengths and weakness of the approach and consider further applications.

An ACT-R approach to mental imagery

A full description of ACT-R is beyond the scope of this paper and so this description will be limited to the two components most relevant to this work: the *vision* module which allows ACT-R to perceive objects in external task environments and the *imaginal* module, located at the intraparietal sulcus (Borst & Anderson, 2013; Borst, Nijboer, Taatgen, van Rijn, & Anderson, 2015) and which functions as ACT-R's limited capacity working memory store in which information is represented and manipulated during problem solving.

ACT-R's perceptual and motor systems were designed to support interaction with computer interfaces to simulate human participants in psychology experiments and therefore typically works within a screen-based 2D coordinate space. ACT-R's visual module doesn't interact with the computer interface directly but via a *visual icon*, an intermediate symbolic representation of the objects in the visual environment.

When ACT-R's visual attention is directed towards an object in the visual icon, information about the object enters two buffers: a *visual* buffer containing information about the object's features (type, shape, colour etc.), and a *visual-location* buffer representing the object's coordinate location. These two distinct buffers correspond to the dorsal *what* and ventral *where* pathways in human visual processing respectively (Ungerleider & Mishkin, 1982; Milner & Goodale, 1993).

Once information has entered the buffers, it is available for further processing, for example as a cue to retrieve further information from ACT-R's declarative memory module or to create a new problem state representation in the imaginal module. Compared to other modules, the imaginal module has a greater degree of flexibility in that, in addition having standard buffer for creating and holding information, it also has an *imaginal-action* buffer to allow the module to be extended with novel capabilities by enabling arbitrary actions to be performed on information in the imaginal buffer. This feature will be crucial for modelling mental imagery.

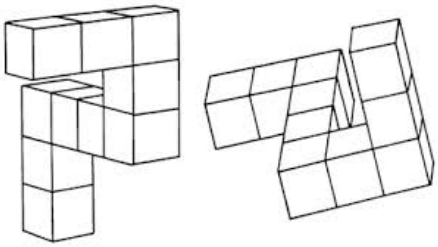


Figure 2: Stimuli used by Shepard and Metzler (1971).

Modifications required to model imagery

Many spatial imagery phenomena involve mental representations of the shape, location, orientation and spatial extent of the imagined objects and a set of processes that are able to transform and compare objects according to these characteristics. While the representational and processing assumptions of ACT-R outlined above impose strict but valuable constraints on methods for modelling mental imagery, in this regard, the discrete symbolic representations of ACT-R's visual module (e.g., shape = 'square') with only one x-y coordinate location for each object are currently inadequate.

In light of this, the approach I adopt augments ACT-R with the addition of a new feature slot in the visual object chunk and a number of functions for spatial processing. The first modification provides ACT-R with additional information regarding the outline shape of environmental objects (in the form of a list of x-y coordinate points). The second provides ACT-R with the ability to perform various imagery operations (e.g., translation, scanning, scaling, zooming, reflection, rotation and composition functions such as intersection, union

and subtraction) using a set of linear and affine transformation functions which act upon the new x-y outline coordinates in the imaginal module via the imaginal-action buffer.

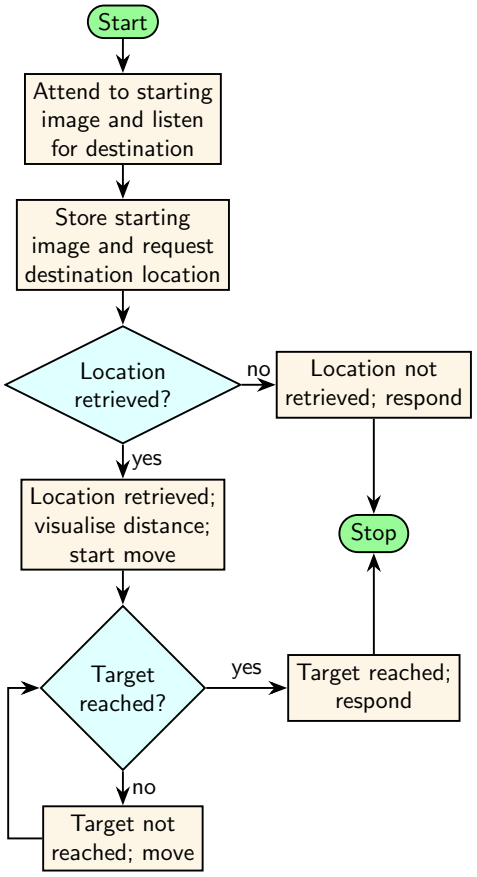


Figure 3: Control structure of the ACT-R model for a trial of the mental scanning experiment. Each rectangle corresponds to one production rule in the model.

Testing the approach

In the remaining sections, the assumptions set out above are tested by using the augmented ACT-R to develop models of two well known mental imagery phenomena: mental scanning and mental rotation[2]. The strategy adopted is one employed by Just and Carpenter (1985) in their model of mental rotation and is similar for both tasks in that the process consists of a series of discrete steps in which the mental image is repeatedly manipulated and then compared to the target image to determine whether they are sufficiently close to stop.

Mental scanning

The first test of the approach is the classic study of mental scanning by Kosslyn et al. (1978) in which people were required to memorise the locations of landmarks on a fictitious map and then imagine travelling between them (see Figure 1).

[2]Both ACT-R models are available to download from GitHub: https://github.com/djpeebles/act-r-imagery-models

On each trial of their experiment participants were asked first to focus on one of the landmarks and then were presented (aurally) with a *destination* word, which may or may not be a landmark. If the given word did name a landmark, participants were required to scan to it and press a button upon reaching it, but if the word was not a landmark, participants simply pressed a second button.

Scanning was performed by imagining a small black speck moving along the shortest straight line from initial to destination landmarks as quickly as possible while still remaining visible. Participants were timed while carrying out the task and analysis of the response times (RTs) revealed a linear relationship between the distance travelled and the time taken to reach the destination.

Modelling the mental scanning task An ACT-R model of the mental scanning task was created consisting of six production rules. The control structure of the model is shown in Figure 3. According to this model, when people hear a destination landmark, they retrieve its location from memory, visualise the distance to be travelled, and then execute a process which incrementally shifts a point from the initial location to the destination by a constant amount. After each movement step, the distance between current and target locations is reviewed to determine whether it is sufficiently short for the process to stop.

The key step involving the new representation and process is represented by a production rule ("Target not reached; move" in Figure 3) which evaluates the distance between the current and target locations and if it is greater than a stopping threshold, uses a translation function to move the current point closer by a fixed amount.

The model assumes that the process of imagining the actual inter-point distance, d_a, is subject to a degree of perceptual error which is a function of d_a, so that visualising greater distances is more errorful. This error, k, is represented by a random value sampled from a logistic distribution with mean 0 and variance $\ln(d_a)$ so that the imagined distance, d_i, is

$$d_i = d_a + bk \tag{1}$$

where b is a scaling parameter.

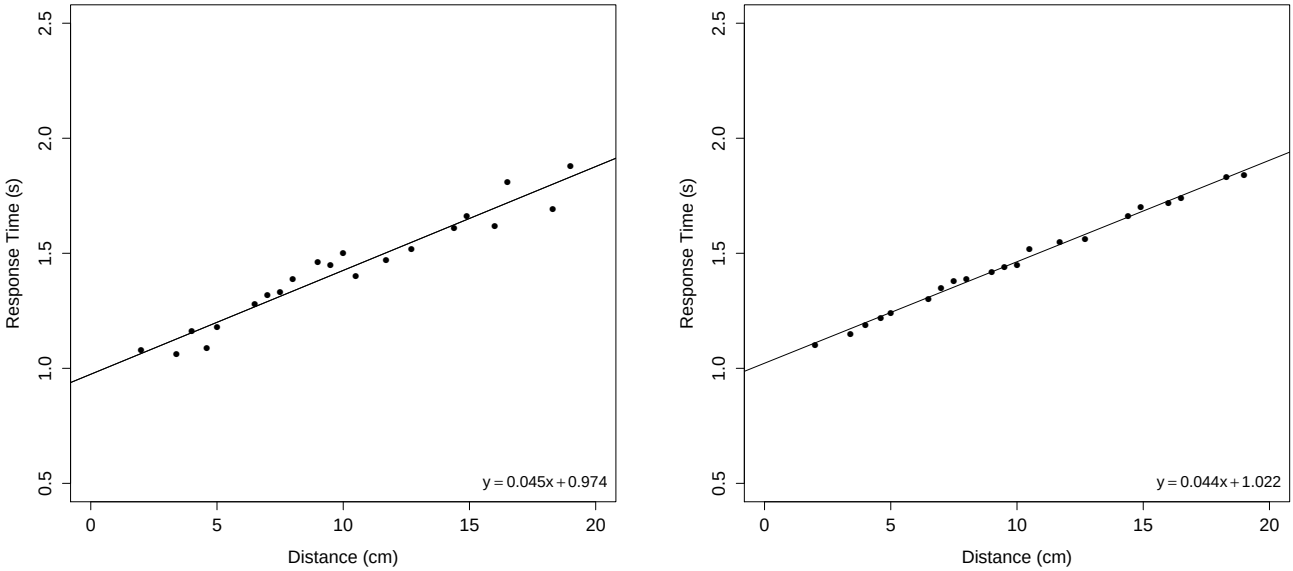
The key determinant of the time taken to traverse the imagined distance is the size of the movement, m, taken at each step and it is assumed that this is related to d_i so that the step size increases with the imagined distance according to

$$m = c \ln(d_i) \tag{2}$$

where c is a scaling parameter.

Finally, it is assumed that the decision to stop is related to the distance to the destination and that this may differ between individuals due to their degree of accuracy or diligence. This distance is represented in the model by a *proximity threshold* parameter, p.

In addition to the three task specific parameters, two ACT-R parameters were also allowed to vary: the *imaginal delay*



(a) Data from Kosslyn et al. (1978).

(b) Data from the ACT-R model

Figure 4: Mean scan time for different distances.

time, t which determines the time cost associated with transforming information in ACT-R's imaginal buffer, and the *latency factor* parameter, F, which modulates the retrieval time for declarative chunks.

To test the model, it was run 50 times (to simulate 50 participants) for all of the 21 distances in the original Kosslyn et al. (1978) study and the mean scan time for each distance computed. Figure 4b shows that the model (with parameters $b = 3$, $c = 18$, $p = 10$, $t = 0.1$ and $F = .75$) provided a close fit to the human data ($R^2 = .97$, RMSD $= 0.07$).

Mental rotation

The second application of the approach is to a mental rotation task, first devised by Shepard and Metzler (1971). In its original form, participants are presented with pairs of similar images, one of which has been rotated around its centre, and then required to decide whether the images are identical or not (see Figure 2). As with the mental scanning task, RT in the mental rotation task increases monotonically with distance—in this case the degree of angular rotation between the images—at approximately 1 second per 60°.

Mental rotation has been studied extensively in a wide variety of different forms and a number of different strategies have been identified (e.g., Khooshabeh, Hegarty, & Shipley, 2013). For this study I model a *holistic* rotation strategy by which mental images (in this case random 2D shapes (Cooper, 1975)) are rotated as single, whole units. This contrasts with a *piecemeal* strategy which subdivides the image and rotates the component pieces separately.

Modelling the mental rotation task An ACT-R model of the mental rotation task was created consisting of five production rules. The control structure of the model is shown in Figure 5. The mental rotation model employs a very similar strategy to the image scanning model in that it performs the task by transforming a current set of coordinate points (in this case by rotation rather than translation) incrementally towards the target, at each step evaluating the remaining distance (i.e., angular displacement) to determine whether or not to stop. As with the scanning model, the key step involving the new representation and process is carried out by a production rule ("Stimuli not aligned; rotate" in Figure 5) which gauges the distance between the current and target images and if it is greater than a stopping threshold, uses a counter-clockwise rotation function to move the current image closer by a fixed amount.

To test the model, it was compared to data from a standard rotation task conducted in Experiment 1 of a recent study conducted by Larsen (2014). The data are taken from a condition in which the target image and a rotated version of the image were presented side by side on a computer screen (the most common form of the task). Ten degrees of rotation were used, from 0 to 180 degrees in increments of 20.

According to the model, when performing the mental rotation task using a holistic strategy, people encode the rotated image, store it in working memory, and then encode the target image. Then, while maintaining visual attention on the target image, people execute a process which incrementally rotates the image counter-clockwise towards the target image by a constant amount (subject to a degree of perceptual er-

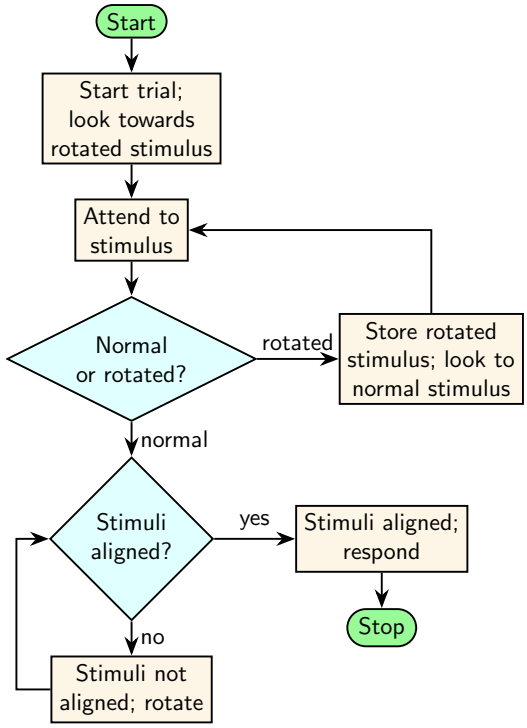


Figure 5: Control structure of the ACT-R model for a trial of the mental rotation experiment. Each rectangle corresponds to one production rule in the model.

ror, represented by a random value sampled from a logistic distribution with mean 0 and variance k).

After each rotation step, the angular disparity between current and target coordinate points is reviewed to determine whether they are sufficiently close for the process to stop. This test is a measure of image similarity in that if the points do not coincide then the rotation process will not stop.

The rotation model shares a number of the same free parameters as the scanning model. As with the scanning model, the rotation model assumes that RT is determined by the size of the rotation increment, m, taken at each step and the proximity threshold, p regulating the stop decision. In the rotation model, the ACT-R *imaginal delay time* parameter, t, was also set to the value of .1s in line with the scanning model.

To test the model, it was run 50 times (to simulate 50 participants) for all of the 10 rotation angles in the original Larsen (2014) study and the mean RT for each distance computed. Figure 6b shows that the model (with parameters $k = 2$, $m = 18$, $p = 10$ and $t = 0.1$) provided a close fit to the human data ($R^2 = .983$, RMSD $= 0.185$).

Discussion

The work described above demonstrates that with only relatively minor modifications and a small number of reasonable assumptions, ACT-R can be applied to develop models of mental imagery phenomena that match human RT data very closely. Crucially, the modifications are restricted to enabling the representation and transformation of shape information but the new representation and processes integrate with the existing control structures of ACT-R so that the behaviour of the model is primarily a result of the strategy encoded in the production rules (which is essentially the same for both tasks) and the information processing assumptions built into the ACT-R's imaginal module.

The architectural parameters used to fit the models are few in number and within acceptable limits. The *imaginal delay time* parameter was set to the same value of .1s for both models but this is shorter than the typical value of this parameter (.2s). The justification for this reduced time is that compared to other tasks that have been used to set this parameter (e.g., algebraic manipulation) the process being carried out in each model (incremental translation or rotation of a representation already in the buffer) is relatively simple and brief.

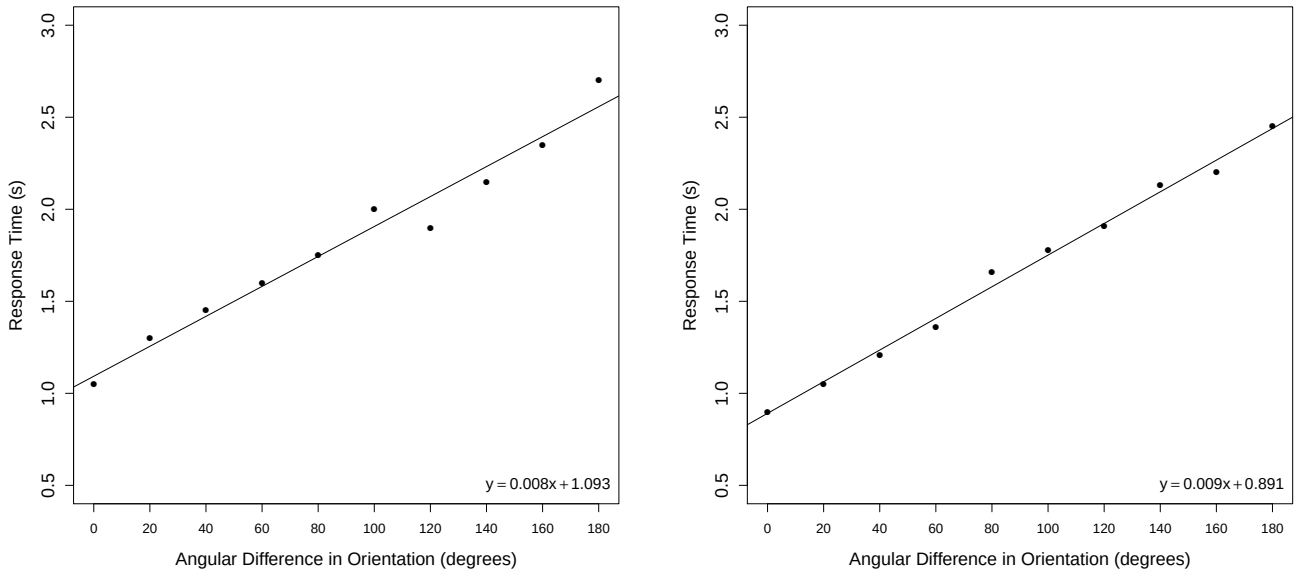
The representation of object spatial extent is not at the level of pixel arrays nor at the level of discrete symbols, but at an intermediate numerical level that abstracts from the pixel level. Similarly, The transformation processes incorporated into the architecture are quantitative in nature and are assumed to belong to the wider set of subsymbolic functions that act upon quantitative information in ACT-R at a level closer to the visual system than the qualitative reasoning processes over symbolic representations.

In this regard, the current work represents a modest step towards answering the question concerning the nature of the representations required to support mental imagery discussed in the introduction. Like many other cognitive architectures, ACT-R is rooted in the classical tradition of cognitive science and the physical symbol system hypothesis (Newell & Simon, 1976) and relies predominantly on amodal symbolic representations and their associated quantitative metadata (Laird, Lebiere, & Rosenbloom, 2017).

As cognitive architectures evolve to capture ever more complex and varied behaviour however, the demand to represent more diverse information formats and computational processes will continue to grow. As this occurs, it will be crucial to investigate the computational capabilities and functional adequacy of alternative representations and processes by modelling tasks that require multiple internal and external representations to provide behavioural evidence for which representations are being used.

There is currently a range of proposals for such representations and processes, several of which were discussed in the introduction. Some advocate some form of bitmap representation to depict the topological structure of objects, while others argue for more abstract representations (or a combination of both). The demands of applying cognitive architectures to more complex, embodied, real world and real time tasks will provide a strong impetus to addressing these questions.

The two behavioural studies modelled here are classics in the literature that have been investigated extensively, and as such they provide a useful initial test of the assumptions. They are relatively simple in nature however (as revealed by



(a) Data from Larsen (2014).

(b) Data from the ACT-R model

Figure 6: Mean response time for different degrees of rotation.

the fact that they can both be modelled by a small number of production rules). A more stringent test of the assumptions is necessary therefore and this will come either from modelling different strategies in the mental rotation task or from different, more challenging tasks, for example the Raven's Progressive Matrices (c.f. Kunda et al., 2013), the *pedestal blocks world* or the *nonholonomic car motion planning* task (Wintermute, 2012) as these require more complex strategies involving a wider range of spatial transformations and will provide richer behavioural data. This is the plan for the next stage of this project.

Acknowledgements

As always, I thank Dan Bothell for his invaluable advice and endless patience.

References

Anderson, J. R. (1978). Arguments concerning representations for mental imagery. *Psychological Review*, *85*(4), 249.

Anderson, J. R. (2007). *How can the human mind occur in the physical universe?* New York, NY: Oxford University Press.

Borst, J. P., & Anderson, J. R. (2013). Using model-based functional MRI to locate working memory updates and declarative memory retrievals in the fronto-parietal network. *Proceedings of the National Academy of Sciences*, *110*(5), 1628–1633.

Borst, J. P., Nijboer, M., Taatgen, N. A., van Rijn, H., & Anderson, J. R. (2015). Using data-driven model-brain mappings to constrain formal models of cognition. *PLoS One*, *10*(3), e0119673.

Cooper, L. A. (1975). Mental rotation of random two-dimensional shapes. *Cognitive Psychology*, *7*(1), 20–43.

Glasgow, J., & Papadias, D. (1992). Computational imagery. *Cognitive Science*, *16*(3), 355–394.

Gunzelmann, G., & Lyon, D. R. (2007). Mechanisms for human spatial competence. In T. Barkowsky, M. Knauff, G. Ligozat, & D. Montello (Eds.), *Spatial Cognition V: Reasoning, Action, Interaction* (pp. 288–307). Springer-Verlag.

Harrison, A. M., & Schunn, C. D. (2002). ACT-R/S: A computational and neurologically inspired model of spatial reasoning. In *Proceedings of the 24th annual meeting of the Cognitive Science Society.*

Just, M. A., & Carpenter, P. A. (1985). Cognitive coordinate systems: Accounts of mental rotation and individual differences in spatial ability. *Psychological Review*, *92*(2), 137–172.

Khooshabeh, P., Hegarty, M., & Shipley, T. F. (2013). Individual differences in mental rotation. *Experimental Psychology*, *60*(3), 164–171.

Kosslyn, S. M., Ball, T. M., & Reiser, B. J. (1978). Visual images preserve metric spatial information: Evidence from studies of image scanning. *Journal of Experimental Psychology: Human Perception and Performance*, *4*(1), 47–60.

Kosslyn, S. M., & Pomerantz, J. R. (1977). Imagery, propositions, and the form of internal representations. *Cognitive Psychology*, *9*(1), 52–76.

Kosslyn, S. M., Thompson, W. L., & Ganis, G. (2006). *The case for mental imagery*. Oxford University Press.

Kunda, M., McGreggor, K., & Goel, A. K. (2013). A computational model for solving problems from the Raven's Progressive Matrices intelligence test using iconic visual representations. *Cognitive Systems Research*, *22*, 47–66.

Laird, J. E. (2012). *The Soar cognitive architecture*. Cambridge, Mass: MIT Press.

Laird, J. E., Lebiere, C., & Rosenbloom, P. S. (2017). A standard model of the mind: Toward a common computational framework across artificial intelligence, cognitive science, neuroscience, and robotics. *AI Magazine*, *38*(4), 13–26.

Larsen, A. (2014). Deconstructing mental rotation. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(3), 1072–1091.

Lathrop, S. D., Wintermute, S., & Laird, J. E. (2011). Exploring the functional advantages of spatial and visual cognition from an architectural perspective. *Topics in Cognitive Science*, *3*(4), 796–818.

Milner, D. A., & Goodale, M. A. (1993). Visual pathways to perception and action. *Progress in Brain Research*, *95*, 317–337.

Montello, D. R. (1993). Scale and multiple psychologies of space. In A. U. Frank & I. Campari (Eds.), *Spatial information theory: A theoretical basis for GIS* (pp. 312–321). Berlin: Springer.

Newell, A., & Simon, H. A. (1976, March). Computer science as empirical inquiry: Symbols and search. *Communications of the ACM*, *19*(3), 113–126.

Pearson, D. G., Deeprose, C., Wallace-Hadrill, S. M. A., Burnett Heyes, S., & Holmes, E. A. (2013). Assessing mental imagery in clinical psychology: A review of imagery measures and a guiding framework. *Clinical Psychology Review*, *33*(1), 1–23.

Pylyshyn, Z. W. (1973). What the mind's eye tells the mind's brain: A critique of mental imagery. *Psychological Bulletin*, *80*(1), 1.

Reisberg, D. (2013). Mental images. In D. Reisberg (Ed.), *The Oxford handbook of cognitive psychology* (pp. 374–388). Oxford University Press. doi: 10.1093/oxfordhb/9780195376746.013.0025

Rosenbloom, P. S. (2012). Extending mental imagery in Sigma. In J. Bach, B. Goertzel, & M. Iklé (Eds.), *International conference on artificial general intelligence* (pp. 272–281). Berlin, Heidelberg: Springer.

Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, *171*(3972), 701–703.

Tabachneck-Schijf, H. J. M., Leonardo, A. M., & Simon, H. A. (1997). CaMeRa: A computational model of multiple representations. *Cognitive Science*, *21*, 305–350.

Trafton, J. G., & Harrison, A. M. (2011). Embodied spatial cognition. *Topics in Cognitive Science*, *3*(4), 686–706.

Trafton, J. G., Hiatt, L., Harrison, A., Tamborello, F., Khemlani, S., & Schultz, A. (2013). ACT-R/E: An embodied cognitive architecture for human-robot interaction. *Journal of Human-Robot Interaction*, *2*(1), 30–55.

Tye, M. (2000). *The imagery debate*. Cambridge, MA: MIT Press.

Ungerleider, L. G., & Mishkin, M. (1982). Two cortical visual systems. In D. J. Ingle, M. A. Goodale, & R. J. W. Mansfield (Eds.), *Analysis of visual behavior* (pp. 549–586). Cambridge, MA: MIT press.

Wintermute, S. (2012). Imagery in cognitive architecture: Representation and control at multiple levels of abstraction. *Cognitive Systems Research*, *19*, 1–29.

Perception of Continuous Movements from Causal Actions

Yujia Peng[1] yjpeng@ucla.edu
Nicholas Ichien[1] ichien@ucla.edu
Hongjing Lu[1,2] hongjing@ucla.edu

[1] Department of Psychology
[2] Department of Statistics
University of California, Los Angeles
Los Angeles, CA 90095 USA

Abstract

We see the world as continuous with smooth movements of objects and people, even though visual inputs can consist of stationary frames. The perceptual construction of smooth movements depends not only on low-level spatiotemporal features but also high-level knowledge. Here, we examined the role of causality in guiding perceptual interpolation of motion in the observation of human actions. We recorded videos of natural human-object interactions. Frame rate was manipulated to yield short and long stimulus-onset-asynchrony (SOA) displays for a short clip in which a catcher prepared to receive a ball. The facing direction of the catcher was either maintained intact to generate a meaningful interaction consistent with causality, or was transformed by a mirror reflection to create a non-causal situation lacking a meaningful interaction. Across three experiments, participants were asked to judge whether the catcher's action showed smooth movements or sudden changes. Participants were more likely to judge the catcher's actions to be continuous in the causal condition than in the non-causal condition, even with long SOA displays. This causal interpolation effect was robust to manipulations of body orientation (i.e. upright versus inverted). These findings indicate that causality in human actions guides interpolation of body movements, thereby completing the history of an observed action despite gaps in the sensory information. Hence, causal knowledge not only makes us see the future, but also fills in information about recent history.

Keywords: causality; causal action; motion interpolation; human action; human interaction

Introduction

In our daily life, we are constantly incorporating new visual information to form a continuous impression of the dynamic world. However, the perceptual construction of smooth movements is not a trivial task, since visual inputs are actually discrete frames or disjointed clips separated by constant eye movements. Flipbooks, for example, exploit our susceptibility to apparent motion (Wertheimer, 1912), where our visual system induces the perception of dynamic scenes from the presentation of static images in rapid succession. Apparent motion offers an illustrative case of the human visual system's tendency to interpolate the paths of perceptual objects over time, and to produce the perception of smooth motion across discrete samples of visual stimuli at different time points. It is well-known that the appearance of smooth motion is determined not only by low-level visual features, such as inter-frame spatial displacement and temporal sampling rate (Braddick, 1974; Burr, Ross & Morrone, 1986), but also by high-level visual knowledge about shapes, objects and events involved in the stimuli (Sigman & Rock, 1974; Braddick, 1980; Shiffrar & Freyd, 1990; 1993; Chen & Scholl, 2016).

In the present paper, we examine whether causal knowledge inherent in human actions influences the extent to which the visual system interpolates body motion. The sense of cause-effect relation can emerge from the irresistible perception of events involving causation, demonstrated by the well-known launching effect between two colliding objects (Michotte, 1946). However, such automatic perception arises not just for physical causation, but also for intentional causation in the social environment. Even as young as 9-month-old, infants perceive objects as "intentional agents" whose states can cause behavioral activities (Crisbra et. al., 1999). Both physical and social causal perceptions are susceptible to the change of spatiotemporal features in dynamic scenes. For example, the perceived causation in the launching event depends on relative speeds of objects in the scene, spatial gaps between those objects, temporal gaps between objects' motions, objects' path lengths (Scholl & Tremoulet, 2000). On the other hand, causal perception can also influence perceptual judgments and memory about spatiotemporal properties in dynamic events.

Previous research has shown that humans rely on their prior knowledge about the causal relation between limb movements and body motions in perceiving human actions (Peng, Thurman, & Lu, 2017), as actions are perceived more natural if visual stimuli are in accordance with causal expectation for human body movements. Causal knowledge has also been shown to elicit false memories of body movements. Strickland and Keil (2011) found that implicit causal connections between agents and objects led to false memories of action frames that were never presented. For example, adults watched videos in which an actor kicked a ball, but the videos omitted the moment in which the actor actually contacted the ball. In a later recall task, participants falsely reported seeing the physical contact when the subsequent footage implied a causal relation between the actor's movements and the motion of the ball. Similarly, Bechlivanidis and Lagnado (2013, 2016) demonstrated that causal knowledge can induce false memories about the temporal order of events. Having a belief that event type A causes event type B made participants more likely to misremember sequences of observed events that violated those causal beliefs (i.e., when an event of type B

temporally preceded an event of type A) than sequences that coincided with their causal belief.

These findings present compelling cases in which causal knowledge plays an influential role in consolidating memories about actions and events. In addition, work on causal binding has shown that causal knowledge biases the perception of time and space (Humphreys & Buehner, 2009, 2010; Buehner, 2012). For example, Buehner and Humphreys (2009) demonstrated that when one event is represented as causing another, the perceived time lapse between the two events appears shorter than when the two events are not causally related. This finding indicates that two causally related events are more likely to trigger the perception of spatiotemporal contiguity.

In the present paper, we test the hypothesis that the perceptual system uses prior knowledge about causal relations in actions to fill in missing information between static frames, yielding the subjective experience of smooth motion in human actions. We recorded videos of human-object interactions in a natural environment (a thrower directing a ball to a catcher). For short clips in which the catcher prepared to receive the ball, the frame rate was manipulated to introduce short and long inter-frame durations, defined as stimulus-onset-asynchrony (SOA). The duration of short SOAs was 33.3 ms/frame; that of long SOAs was 100 ms/frame. For causal actions, the facing direction of the catcher was maintained to generate a meaningful interaction consistent with a causal interpretation. For non-causal actions, the facing direction of the catcher was inverted to disrupt any meaningful interaction and generate an action sequence inconsistent with a causal interpretation. Participants were asked to judge whether the catcher's action showed smooth body movements or sudden changes. If causal knowledge in actions creates a top-down influence on interpolation of discrete pieces of motion information, observers will be more likely to perceive smooth actions when observing causal than non-causal actions. In addition, the predicted effect is expected to be stronger for long-SOA displays in which the visual inputs are sparse, with fewer image frames.

Experiment 1

Experiment 1 was designed to assess how a causal action between an agent and a physical object influences interpolation in the perception of smooth human actions. Causal actions were generated with an agent interacting with a moving object. Non-causal actions were generated with the same agent facing away from the moving object. We hypothesized that in the causal action condition, discretized human actions would be more likely to be perceived as smooth motion sequences.

Method

Participants

Fifty undergraduate students at UCLA (mean age = 21.1; 40 female) participated in the experiment for course credit. All experimental procedures were approved by the UCLA Office for Protection of Human Subjects. All participants had normal or corrected-to-normal vision.

Stimuli

Action videos were filmed in a gym using a camera with a temporal resolution of 30 frames/s. Two pairs of actors (one male pair and one female pair) were filmed. Each pair performed three throwing-catching actions (bounce pass, overhead pass, and chest pass), with each actor being the thrower once and catcher once. Seven video clips were selected as experimental stimuli. Sample video stimuli can be viewed at https://yujiapeng.com/causal-illusion-real.

In Experiment 1, only the catcher and the ball appeared in the video; the thrower was not shown. For each video, a short critical period was selected during which the catcher's arms showed the largest rising momentum during preparation to catch the ball. Each video lasted for 567 ms. There were 10 frames before the critical period, and 1 frame after the critical period. The critical period began when the catcher's arms started to rise, and it ended right before the actor's hands touched the ball. The duration of the critical period was 200 ms. In the long-SOA condition, only the first and the last frame of the catcher's body movements were presented, all the middle frames were omitted. The presentation duration of the first and the last frames were lengthened to cover half of the critical period at 100 ms per frame. In the short-SOA condition, all six frames showing body movements of the catcher were displayed, with the frame duration at 33.3 ms/frame. Note that the duration of the critical period was the same (200 ms) for both long-SOA and short-SOA displays. The movements of the ball were also the same and were kept intact in both long-SOA and short-SOA displays (Figure 1).

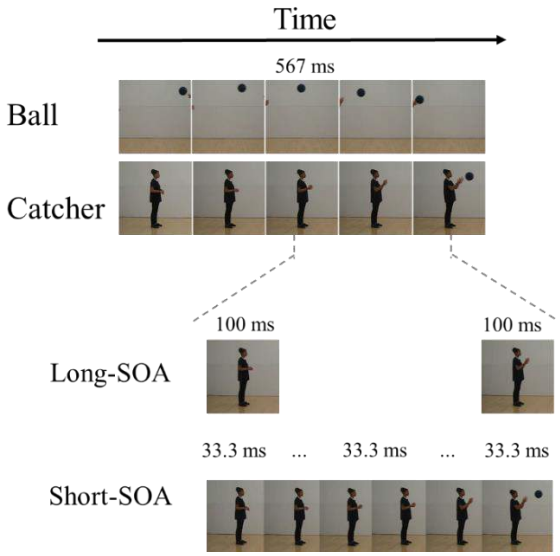


Figure 1. Illustrations of the critical clip in the long-SOA display with two frames (100 ms/frame) with a sudden posture change, and in the short-SOA display with six frames (33 ms/frame).

As shown in Figure 2, the causal condition showed the catcher facing toward the ball as the ball movement causes the catcher to move his or her body in preparation. To generate non-causal actions, image frames were processed using Matlab and Adobe Photoshop to horizontally reverse the facing direction of the catcher. The catcher was flipped horizontally to face away from the ball in the entire video, while keeping the background and the ball movement intact.

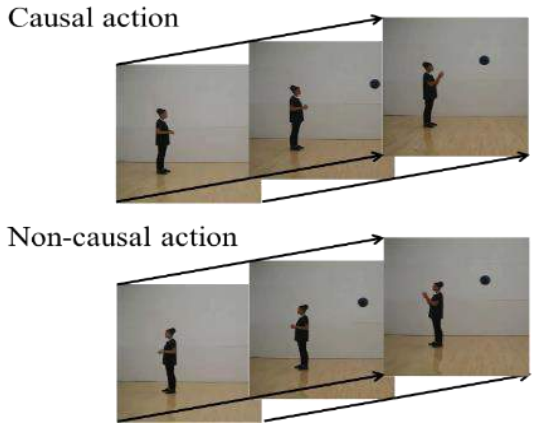


Figure 2. Sample frames of a causal action with the catcher facing towards the ball, and a non-causal action with the catcher facing away from the ball.

Procedure

Participants were seated 35 cm in front of a monitor with a 1024×768 resolution and 60 Hz refresh rate. All the stimuli were generated by MATLAB Psychtoolbox (Brainard, 1997). Participants were instructed, "You will view an actor playing sports (such as passing a basketball) with someone else who is occluded by a whiteboard. The task is to judge whether the catcher actor shows a smooth action or a non-smooth sudden posture change. For a smooth action, the actor smoothly moves from one posture to another. For a non-smooth action, the actor suddenly moves from one posture to another."

On each trial, a white fixation cross was presented at the center of the screen. Participants were asked to focus on the fixation cross throughout the experiment and to use their peripheral vision to see the video without making saccades. The center of the video was presented 13.7 degrees to the left or to the right of the fixation point with a height of 18 degrees. Showing the video in peripheral vision reduced the possibility that observers would track movements of the catcher without paying attention to other parts of the display. Half of the trials presented the video on the left of the fixation and the other half on the right. The catcher actor was always presented on the side relatively farther away from the fixation point. For example, if the video was presented on the right side, the ball flew from left to right and the catcher was located on the right side of the ball. After the video display, participants were asked to press one of two buttons to judge whether the video demonstrated actions with smooth body movements or sudden posture changes.

Participants were first presented with two blocks of practice trials to familiarize them with the task. In the practice blocks, participants saw "correct" on the screen plus a beep after each correct response, and saw "incorrect" without a beep after each incorrect response. Each practice block consisted of eight trials. A separate video was used as the stimulus for the practice block; this video was not presented in the test. In the first block of practice, videos were slowed down to show the entire video with the frame rate of 66.6 ms/frame and to display the critical period for 666 ms. This manipulation was intended to allow participants to become familiar with the experimental setting and to understand the difference between smooth motion and sudden posture changes in body movements. In the second block of practice trials, videos were presented at a frame rate of 33.3 ms/frames, and the duration of the critical period was 200 ms, as in the test session.

The test session followed the practice blocks. Test trials were identical to those in the second practice block with two exceptions: participants received no feedback on test trials, and test trials employed six new videos that were not used in practice blocks. A total of five test blocks were administered, each with 24 trials (causal/non-causal x long-/short SOA x 6 actions). In each block, the presentation order of videos was randomly shuffled. Proportions of responses in judging actions as smooth motion were recorded for each condition.

Results

We first examined the data in Block 1, as performance on subsequent blocks was likely to be affected by increased familiarity with the six videos used in the experiment. We conducted a 2 (SOA: short- vs. long-SOA) by 2 (causality: causal action vs. non-causal action) repeated-measures ANOVA on the proportion of responses judging the catcher's action as smooth motion. As shown in Figure 3a, results revealed a significant main effect of causal action, $F(1,49) = 4.742$, $p = .034$. Specifically, the proportion of "smooth" responses was significantly higher in the causal action condition in the long-SOA condition, in which the catcher faced towards the flying ball than in the non-causal action condition in which the catcher faced away from the ball ($t(49) = 2.243$, $p = .029$). This contrast was not significant in the short-SOA condition ($t(49) = 1.193$, $p = .239$), probably due to much less room of interpolation given the nature of smoothness of short-SOA videos. Note that the smooth motion signal was much weaker in the long-SOA display, since the stimulus included only two static postures with the largest spatial displacements. However, the causal relation between the ball and the body movements of the catcher enhanced interpolation between the two distinct postures, resulting in more misperception of sudden posture changes as smooth body movements. These results indicate that the effect of causality on motion interpolation emerged at the very beginning of the experiment. Not surprisingly, the main effect of the SOA was significant, $F(1,49) = 124.803$, $p < .001$, as short-SOA displays provided stronger motion signals with short inter-frame spatial displacements than did long-SOA

displays. The two-way interaction effect between causality and SOA was not significant, $F(1,49) = .662$, $p = .42$.

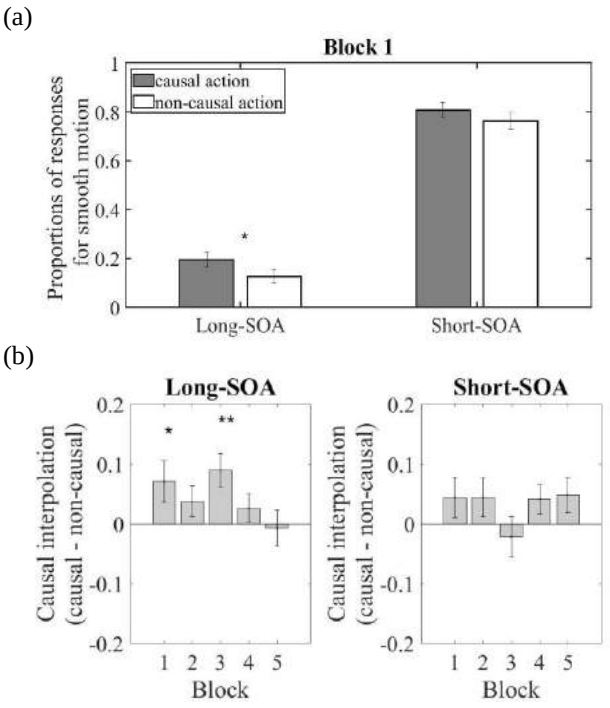


Figure 3. Results of Experiment 1. (a) Proportions of responses in block 1 judging the catcher's action as smooth motion. Asterisks indicate statistically significant differences between conditions (* $p < .05$, ** $p < .01$). (b) The difference between proportions of responses to causal and non-causal actions across 5 blocks in long- or short-SOA displays.

Results of the causal interpolation effect across 5 blocks were presented in Figure 3b. To investigate whether the impact of causal actions on motion interpolation was maintained across blocks despite increased familiarity with the six videos, we conducted a three-way repeated measures ANOVA with blocks as the third factor. We found a significant main effect of causal actions ($F(1,49) = 12.419$, $p = .001$), reflecting a larger proportion of "smooth" responses in the causal condition than non-causal condition. This result suggests that the facilitatory influence of causality on the perception of smooth movements was maintained, even with increased familiarity with the videos. However, this main effect was qualified by a significant three-way interaction ($F(4,196) = 2.815$, $p = .027$), reflecting a complex relation between familiarity and the influence of causal knowledge on the perceptual task. The block variable had a strong impact on responses in the long-SOA displays ($F(4,196) = 4.572$, $p = .001$), but a relatively weaker impact on short-SOA displays, for which the simple main effect of block was not reliable ($F(4,196) = 1.722$, $p = .15$). This pattern was likely the result of close-to-ceiling performance in perceiving smooth motion in the short-SOA displays.

Experiment 2

In Experiment 1, we found evidence that causal interactions between a catcher and the ball facilitated the perception of smooth movements. In Experiment 2, we investigated whether the effect could be generalized from human-object interactions to human-human interactivity. We predict that when the two agents show a causal relation connecting their movements (i.e. one agent throwing and one agent catching), observers will also be more likely to perceive smooth body movements.

Method

Participants

Forty-eight new UCLA students (mean age = 20.48; 33 female) participated in the experiment for course credit. All participants had normal or corrected-to-normal vision.

Stimuli and Procedure

The experiment employed the same basic videos as in Experiment 1, showing two actors pass balls. The stimuli included the body movements of the thrower and the catcher (Figure 4). A white occluder was presented at the center of the video to cover the movements of the ball. Depending on the actual duration of action sequences, the stimuli ranged from 633 ms to 1233 ms. There were 10 frames before the critical period, and 1 frame after the critical period. The duration of the critical period was 200 ms. In the instructions, participants were asked to respond to the movements of the catcher while paying attention to the entire video. The causal manipulation in Experiment 2 was the same as Experiment 1: the facing direction of the catcher was horizontally reversed to generate the non-causal condition. The procedure for Experiment 2 was the same as that for Experiment 1.

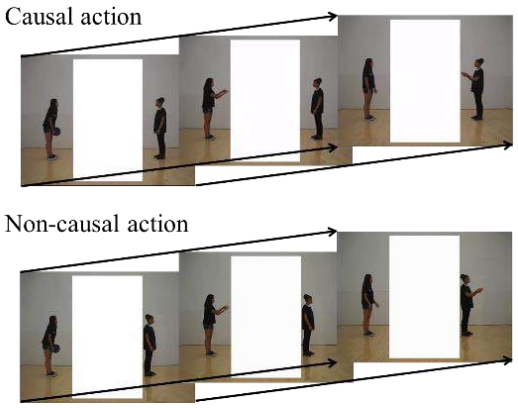


Figure 4. Sample frames of a causal action with the catcher facing towards the thrower, and a non-causal action with the catcher facing away from the thrower.

Results

As shown in Figure 5a, the proportion of smooth responses in Block 1 again revealed a significant main effect of causality ($F(1,47) = 9.874$, $p = .003$). Despite a longer temporal delay between the two actors' actions, the causal relation between the two actors' body movements impacted the visual

experience of the catcher, as perceiving the catcher's movements elicited perception of more smooth and coherent motion. The proportion of smooth responses was significantly greater in the causal action condition compared to the non-causal action condition for the long-SOA condition ($t(47) = 2.887$, $p = .006$), but not for the short-SOA condition ($t(47) = 1.681$, $p = .099$). No interaction effect was found, $F(1,47) = 0.407$, $p = .527$. These results extended the pattern of causal effects observed in Experiment 1.

Results of the causal interpolation effect across 5 blocks were presented in Figure 5b. A three-way repeated measures ANOVA with blocks as the third factor showed a significant main effect of causal actions ($F(1,47) = 6.508$, $p = .014$), with a greater proportion of "smooth" responses in the causal condition than the non-causal condition. There was also a significant main effect of block ($F(4,188) = 5.904$, $p < .001$). Neither the two-way interactions nor the three-way interaction was reliable. In summary, the converging results from the two experiments indicate that the influence of causal action on motion interpolation persisted even with increased familiarity with the videos.

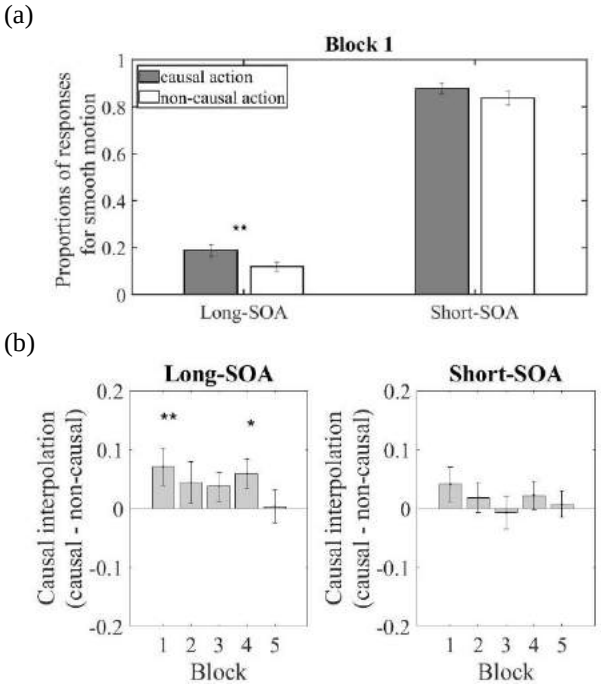


Figure 5. Results of Experiment 2. (a) Proportions of responses in block 1 judging the catcher's action as smooth motion (* $p < .05$, ** $p < .01$). (b) The difference between proportions of responses to causal and non-causal actions across 5 blocks in long- or short-SOA displays.

Experiment 3

Experiment 3 aimed to investigate whether the influence of causal actions on motion interpolation depends on other visual cues. Body orientation is a well-known cue for action recognition (Pavlova & Sokolov, 2000), as observers show worse recognition performance when actions are presented upside-down. If the interpolation effect revealed in the previous two experiments was induced by high-level causal knowledge, then inverting the video would *not* yield a significant difference between upright versus upside-down actions, since both cases preserve the temporal contingency and the causal relation between humans and objects.

Methods

Participants

Fifty-two new UCLA undergraduate students (mean age = 20.0; 43 female) participated in the experiment for course credit. All participants had a normal or corrected-to-normal vision.

Stimuli and Procedure

Experiment 3 used the same stimuli as the causal condition in Experiment 1. On half of the trials, the stimuli used inverted videos, and the other half used intact videos (Figure 6). The task and procedure of Experiment 3 were otherwise the same as in Experiment 1.

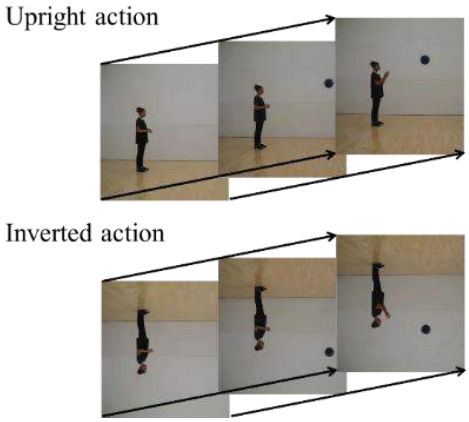


Figure 6. An illustration showing sample frames of an upright and an inverted action in Experiment 3.

Results

We first conducted a 2 (SOA: short- vs. long-SOA) by 2 (orientation: upright vs. inverted) repeated-measures ANOVA on the proportion of responses in Block 1 judging the catcher's action to be smooth motion. As shown in Figure 7a, the main effect of orientation was not significant ($F(1,51) = 2.509$, $p = .119$). The interaction between body orientation and SOA was also not significant ($F(1,51) = 1.525$, $p = .222$). The results from Block 1 suggest that as long as the causal relation is maintained in observed activities, body orientation does not affect the misperception of seeing smooth movements, even when the motion signals were weak (in the long-SOA displays).

Results of the causal interpolation effect across 5 blocks were presented in Figure 7b. To investigate whether the impact of body orientation on motion interpolation changed across blocks with increased familiarity with the six videos, we further conducted a three-way repeated measures ANOVA with blocks as the third factor. This analysis revealed a significant main effect of orientation ($F(1,51) = 5.554$, $p =$

.022). This main effect was largely driven by a significant difference between the upright and inverted conditions in later blocks. For example, in the final block (Block 5), a greater proportion of "smooth" responses was made in the upright conditions than the inverted conditions for the long-SOA condition ($t(51) = 2.139$, $p = .037$). This pattern suggests that the impact of body orientation on visual analysis of actions increased with familiarity of the stimuli.

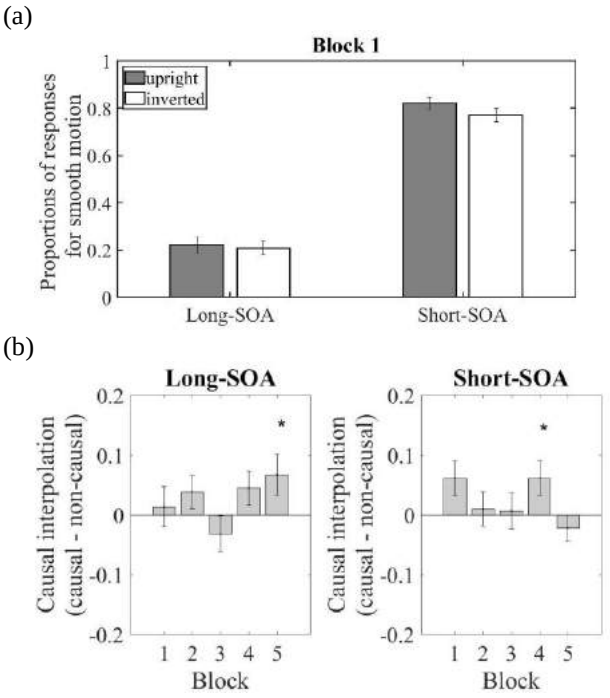


Figure 7: Results of Experiment 3. (a) Proportions of videos in block 1 judged as smooth actions (* $p < .05$, ** $p < .01$). (b) The difference between proportions of responses to causal and non-causal actions across 5 blocks in long- or short-SOA displays.

General Discussion

Apparent motion perception makes it possible to record movements of objects and humans by sampling the motion and displaying the samples as stationary pictures in sequence (e.g., videos, cinema). This study showed that a causal interaction between an agent and a physical object increased the likelihood that people would perceive smooth actions even when the stimuli showed a sudden change in long-SOA displays. This result suggests that causality acts as a temporal "glue" to fill in observers' visual experience by interpolating discrete image frames to produce the perception of smooth, continuous motion. These results extended previous evidence that perception in physical causation helps to fill in important visual information left out from a sequence of events to social causal perception. The representation of an object's implicit causal history has been shown to induce a transformational apparent motion (Tse, Cavanagh, & Nakayama, 1998) of simple objects (Chen & Scholl, 2016), akin to the "causal filling in" effect reported by Strickland and Keil (2011). A "causal filling in" mechanism could have benefitted from evolutionary selection pressure by aiding the continuous perception of animal motions despite occlusion by trees or other obstacles.

Causal knowledge about human body movements may not only help to connect discrete events in the perceptual process, but also may facilitate the process of making inferences and predictions about actions. A causal framework may help the visual system to infer the past. For example, human observers get a vivid feeling of seeing the immediate past of objects or human postures presented in static frames (Kourtzi, 2004). This phenomenon suggests that causal knowledge aids the visual system in inferring and reconstructing the causal history of objects and human actions. On the other hand, as earlier research on motion perception has suggested that the visual system anticipates the positions of simple objects based on their apparent motion trajectory (Freyd & Finke, 1984), more recent research has suggested that similar anticipatory visual processing is also affected by comparatively complex causal knowledge of human actions. For example, Su and Lu (2017) used skeletal biological motion displays and found a flash-lag effect, such that when a briefly-flashed dot was presented physically in perfect alignment with a continuously-moving limb, the flashed dot was perceived to lag behind the position of the moving joint. This finding suggests that the representation of human actions is anticipatory, due to a potential top-down action prediction mechanism. It has also been found that infants as young as five months are able to gaze toward the future direction implied by the static posture of a runner (Shirai & Imura, 2014, 2016), suggesting the early emergence in infancy of an ability to predict dynamic human actions from still pictures.

The present results demonstrated rapid effects of learning across blocks. Experiment 1 showed a significant three-way interaction between block, causality and SOA, suggesting an interaction between the top-down influence of causality and bottom-up perceptual processing of motion stimuli. The top-down influence of causality may be stronger in situations in which uncertainty about the visual input is high, such as when dynamic stimuli are presented in peripheral vision or embedded in noise. The effect may be weakened after repetitive exposures to the stimuli, as perceptual learning may enhance performance for visual tasks. These results are consistent with previous findings that causal perception can change upon repeated exposure of the same stimuli (Rolfs, Dambacher & Cavanagh, 2013).

In conclusion, the current study provides evidence of the important role played by causal knowledge in the perception of smooth motion. Causal relations involving human actions, and their interactions with objects and other agents, have a strong influence on motion perception for body movements. The causal relations involved in actions facilitate visual interpolation of discrete dynamic events to provide a continuous perception of human-involved activities. The top-down influence of knowledge about human actions interacts with bottom-up perceptual processes to enhance the robustness and efficiency in action perception (Lu, Tjan & Liu, 2006; Thurman & Lu, 2014) and intention inference (Shu et. al.,

2018). Causal knowledge not only makes us see the future, but also fills in information about recent history.

Acknowledgments

This research was supported by NSF Grant BCS-1655300. We thank Brian Scholl for helpful comments, Eun Ji Song, Tabitha Safari and Jiming Sheng for helping with filming actions, and Eun Ji Song, Andrew Kwik, Korosh Bahrami for their assistance in data collection.

References

Bechlivanidis, C., & Lagnado, D. A. (2013). Does the "why" tell us the "when"? *Psychological Science*, *24*(8), 1563-1572.

Bechlivanidis, C., & Lagnado, D. A. (2016). Time reordered: Causal perception guides the interpretation of temporal order. *Cognition*, *146*, 58-66.

Braddick, O. (1974). A short-range process in apparent motion. *Vision Research*, *14*(7), 519-527.

Braddick, O. J. (1980). Low-level and high-level processes in apparent motion. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, *290*(1038), 137-151.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, *10*, 433-436.

Buehner, M. J. (2012). Understanding the past, predicting the future: Causation, not intentional action, is the root of temporal binding. *Psychological Science*, *23*(12), 1490-1497.

Buehner, M. J., & Humphreys, G. R. (2009). Causal binding of actions to their effects. *Psychological Science*, *20*(10), 1221-1228.

Buehner, M. J., & Humphreys, G. R. (2010). Causal contraction: Spatial binding in the perception of collision events. *Psychological Science*, *21*(1), 44-48.

Burr, D. C., Ross, J., & Morrone, M. C. (1986). Smooth and sampled motion. *Vision Research*, *26*(4), 643-652.

Chen, Y. C., & Scholl, B. J. (2016). The perception of history: Seeing causal history in static shapes induces illusory motion perception. *Psychological Science*, *27*(6), 923-930.

Csibra, G., Gergely, G., Bíró, S., Koos, O., & Brockbank, M. (1999). Goal attribution without agency cues: the perception of 'pure reason' in infancy. *Cognition*, *72*(3), 237-267.

Freyd, J. J., & Finke, R. A. (1984). Representational momentum. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. *10*(1), 126-132.

Humphreys, G. R., & Buehner, M. J. (2009). Magnitude estimation reveals temporal binding at super-second intervals. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(5), 1542-1549.

Humphreys, G. R., & Buehner, M. J. (2010). Temporal binding of action and effect in interval reproduction. *Experimental Brain Research*, *203*(2), 465-470.

Kourtzi, Z. (2004). But still, it moves. *Trends in Cognitive Sciences*, *8*(2), 47-49.

Lu, H., Tjan, B. S., & Liu, Z. (2006). Shape recognition alters sensitivity in stereoscopic depth discrimination. *Journal of Vision*, *6*(1), 7-7.

Michotte, A. (1946). *La Perception de la Causality*. Louvain: Institut Superieur de Philosophie.

Pavlova, M., & Sokolov, A. (2000). Orientation specificity in biological motion perception. *Perception & Psychophysics*, *62*(5), 889-899.

Peng, Y., Thurman, S., & Lu, H. (2017). Causal action: A fundamental constraint on perception and inference about body movements. *Psychological Science*, *28*(6), 798-807.

Rolfs, M., Dambacher, M., & Cavanagh, P. (2013). Visual adaptation of the perception of causality. *Current Biology*, *23*(3), 250-254.

Scholl, B. J., & Tremoulet, P. (2000). Perceptual causality and animacy. *Trends in Cognitive Sciences*, *4*, 299-309.

Shiffrar, M., & Freyd, J. J. (1990). Apparent motion of the human body. *Psychological Science*, *1*(4), 257-264.

Shiffrar, M., & Freyd, J. J. (1993). Timing and apparent motion path choice with human body photographs. *Psychological Science*, *4*(6), 379-384.

Shirai, N., & Imura, T. (2014). Implied motion perception from a still image in infancy. *Experimental Brain Research*, *232*(10), 3079-3087.

Shirai, N., & Imura, T. (2016). Emergence of the ability to perceive dynamic events from still pictures in human infants. *Scientific Reports*, *6*, 37206.

Shu, T., Peng, Y., Fan, L., Lu, H., & Zhu, S. C. (2018). Perception of Human Interaction Based on Motion Trajectories: From Aerial Videos to Decontextualized Animations. *Topics in cognitive science*, *10*(1), 225-241.

Sigman, E., & Rock, I. (1974). Stroboscopic movement based on perceptual intelligence. *Perception*, 3(1), 9-28.

Strickland, B., & Keil, F. (2011). Event completion: Event based inferences distort memory in a matter of seconds. *Cognition*, *121*(3), 409-415.

Su, J., & Lu, H. (2017). Flash-lag effects in biological motion interact with body orientation and action familiarity. *Vision Research*, *140*, 13-24.

Thurman, S. M., & Lu, H. (2014). Perception of social interactions for spatially scrambled biological motion. *PLoS One*, *9*(11), e112539.

Tse, P., Cavanagh, P., & Nakayama, K. (1998). The role of parsing in high-level motion processing. In T. Watanabe (Ed.), *High-level motion processing: Computational, neurobiological and psychophysical perspectives* (pp. 249–266). Cambridge: MIT Press.

Wertheimer, M. (1912). Experimentelle studien über das sehen von bewegung. *Zeitscrhift für Psychologie*, *61*(1), 161-265.

Age-Related Differences in the Influence of Category Expectations on Episodic Memory in Early Childhood

Kimele Persaud
kimele.persaud@rutgers.edu
Psychology, Rutgers University-Newark

Carla Macias
cm1172@scarletmail.rutgers.edu
Psychology, Rutgers University-Newark

Pernille Hemmer
pernille.hemmer@psych.rutgers.edu
Psychology, Rutgers University-New Brunswick

Elizabeth Bonawitz
lbaraff@gmail.com
Psychology, Rutgers University-Newark

Abstract

Previous research evaluating the influence of category knowledge on memory found that children, like adults, rely on category information to facilitate recall (Duffy, Huttenlocher, & Crawford, 2006). A model that combines category and target information (Integrative) provides a superior fit to preschoolers recall data compared to a category only (Prototype) and target only (Target) model (Macias, Persaud, Hemmer, & Bonawitz, in revision). Utilizing data and computational approaches from Macias et al., (in revision), we explore whether individual and age-related differences persist in the model fits. Results revealed that a greater proportion of preschoolers recall was best fit by the Prototype model and trials where children displayed individuating behaviors, such as spontaneously labeling, were also best fit by the Prototype model. Furthermore, the best fitting model varied by age. This work demonstrates a rich complexity and variation in recall between developmental groups that can be illuminated by computationally evaluating individual differences.

Keywords: Episodic Memory; Children; Computational Models; Category Knowledge; Color

Introduction

Reconstructing events from memory is an important facet of cognition, given that it informs how we perceive, interact with, and reason about the world around us. As with all computational processes, human memory is limited in its capacity and resolution, raising questions of how the mind handles the reconstruction of events from memory. That is, how do we strategically encode information that supports later use, while minimizing effort, error, and large demands on storage? This question is doubly interesting for young children whose memory systems are still developing. Relative to adults, children have comparatively limited cognitive resources (Davinson, Amso, Anderson, & Diamond, 2006; Diamond, 2006; Keresztes, Ngo, Lindenberger, & Newcombe, 2018), and their ability to maintain information in memory becomes compromised when faced with increased cognitive load (e.g., increased inhibition demands). Thus, an important question of development is what cognitive strategies might young learners employ to reduce uncertainty (i.e., noise or error) when retrieving information from memory?

To tackle strategic reconstruction of episodic events, research in adult cognition suggests that adults use prior knowledge and expectations to facilitate retrieval of information from memory. Adults develop prior knowledge and expectations that are well-calibrated to the statistical regularities of the environment (e.g., Griffiths & Tenenbaum, 2006), and use this knowledge to optimally perform on a broad range of cognitive tasks including: categorization (Huttenlocher, Hedges, & Vevea, 2000), reasoning (Oaksford & Chater, 1994), and generalization (Tenenbaum & Griffiths, 2001). In memory, well-calibrated knowledge and expectations for a stimulus category can improve average recall (Huttenlocher, Hedges, & Duncan, 1991; Huttenlocher et al., 2000). For example, Huttenlocher et al. (2000) found that people quickly develop expectations for the underlying categorical distribution of stimulus features, and use this knowledge to fill in noisy and incomplete memories. They demonstrated that responses regressed toward the mean of the overall category, thereby improving average recall.

This relationship between prior knowledge and episodic memory can be captured within a simple Bayesian framework which assumes that prior knowledge and expectations for the environment are optimally combined with noisy episodic content to produce recall of episodic experiences (Hemmer & Steyvers, 2009; Huttenlocher et al., 2000; Persaud & Hemmer, 2014; Steyvers & Dennis, 2006). Bayes rule provides a principled account of how to combine noisy memory representations with prior expectations to calculate the posterior probability for recall.

$$p(\theta|y) \propto p(y|\theta)p(\theta)$$

The posterior probability $p(\theta|y)$ describes how likely a recalled feature θ is, given prior expectations for the recalled feature $p(\theta)$ and noisy memory traces y. In this way, the Bayesian framework makes specific predictions about patterns that are explicitly borne out of the data, namely a regression to the category mean effect. It predicts that recall of stimulus features (e.g., different shades of red) is either over or under-estimated toward the mean of the category.

Recent evidence suggests that children, like adults, adopt a similar process of integrating prior category knowledge with episodic traces to reconstruct events in memory. For example, Duffy et al. (2006) used assumptions of the Category Adjustment model (CAM) (Huttenlocher et al., 1991, 2000) to

evaluate the contribution of category knowledge to memory for object sizes in children. CAM assumes that if category knowledge is integrated in memory, recall would exhibit regression to the mean effects. The model also assumes that the more noisy the episodic information, like memories in children, the stronger recall will regress to the mean. Duffy et al. (2006) found that like adults, children's recall regressed toward the mean of the underlying category distribution. This suggests that on an individual trial, a child might not have remembered the exact studied size, so they might use their learned category knowledge of the most frequently studied object sizes to help reconstruct the true size. They concluded that children use category knowledge to estimate stimulus features from memory.

Similarly, Macias and colleagues (in revision) used a simple episodic memory task, where children were shown shapes paired with different colors and were asked to recall the color-shape pairings. They found that children's recall regressed toward the mean of the seven color categories that were studied, indicating an influence of category knowledge on memory. To further assess episodic memory, they then evaluated the fits of three computational models of memory to explain the data: a Noisy Target model that assumes recall solely mirrors episodic information (i.e., the target color values plus random noise), a Noisy Prototype model that assumes that recall solely mirrors category information (plus noise), and an Integrative model that assumes that recall is an integration of episodic and category information. Quantitative model fits to the aggregate data favored the Integrative model.

These studies of memory in children, taken together, highlight an important role that category knowledge plays in episodic memory at early development (i.e., preschool age) and provide a watershed moment to explore the reconstructive nature of episodic memory at earlier stages. More specifically, this work facilitates the opportunity to perform a critical in-depth analysis of children's recall data to tease apart underlying individual and group-related differences in the reconstructive process. Exploring individual and age related differences is motivated by the Duffy et al. (2006) finding that not only do children rely on category knowledge, but also that memory in younger children exhibited steeper regression to the mean patterns, relative to older children. Recall based solely on category information could also result in steeper regression to the mean, and in turn, might be better fit by the Macias et al., Noisy Prototype ('category only') model. In other words, it could be the case that at the individual subject level, children might differ in the best fitting model, such that those with steeper regression might be better fit by the Noisy Prototype model, while less steep regression might be better captured by an Integrative model.

Furthermore, recall performance in children might not only differ at the individual subject level, but also at the individual trial level, especially if contextual strategies, such as spontaneously labeling study features, are employed to facilitate recall performance. For example, while running their study, Macias et al., observed that participants spontaneously labeled the colors, as they studied them and/or as they recalled them. For example, one older learner (age = 4.64years), stated, "Purple, purple, purple. I got this.", while studying a purple hue value. Counterintuitively, while labeling may boost the learner's ability to remember that an item was observed from a particular category, it may also lead to noisier storage of specific stimuli that deviate from category means, because the label provides a cheaper (albeit potentially less accurate) compression option than storing the details of the original. In this way, this individuating behavior of labeling might impact the reconstruction of events in memory at either the individual subject or trial level. Recent research suggests that labeling can influence recall of continuous color values, such that labeling results in information being lost gradually as opposed to suddenly (see, Donkin, Nosofsky, Gold, & Shiffrin, 2014 for discussion on the role of labeling, sudden death, and gradual decay in memory). To this end, there might be a difference in the best fitting models for children who spontaneously label colors or for specific trials where colors are labeled.

Therefore, the goal of this paper is to assess individual and age related differences in the reconstruction of events from memory in early development. More specifically, we sought to evaluate whether younger and older children employ different strategies to recall episodic events and whether the behavior of spontaneously labeling was better fit by a particular model. We hypothesized that young and older children would differ in their reconstructive processes, such that a different proportion of children from each group would be better fit by the three models. We expected that older children would be better explained by an Integrative model (i.e., combining noisy episodic traces with category knowledge), mirroring the behavior of adults, and younger children would be better explained by a Noisy Prototype model, given the degree of inexactness in their memory traces.

We also hypothesized that the individuating behavior of spontaneous labeling would impact memory reconstruction such that trials where labels were spontaneously provided would be better captured by the Noisy Prototype model. To test our hypotheses, we fit the Noisy Target, Noisy Prototype, and Integrative models to the experimental data from Macias et al., (in revision) at the individual subject level.

We then evaluated the log likelihood scores of the model fits to determine which account most often explained memory performance in younger and older children. In other words, we looked to see which model explained behavior for the greater proportion of children. After, we explored best fitting parameter values that would capture the amount of noise in the recall data for young and older children. A difference in the amount of noise in the data is one potential explanation for age related differences in the best fitting model. Finally we assessed whether labeling behavior affected the proportion of children fit by each of the models.

Three Models of Memory

Noisy Target Model The Noisy Target model assumes that information is stored in episodic memory as noisy traces of studied values (e.g., specific color values). In this way, reconstructed events are just inexact representations of true studied values (and not altered by category knowledge). If children are using the Noisy Target model, we should expect the noise (or error) in recall to be normally distributed around the true studied feature values, with no apparent bias toward a particular recall value. To evaluate this model relative to the data, we calculated the probability of responses given a Gaussian distribution centered on the target value, with noise in memory (we assume the same memory noise value learned from Macias et al.).

Noisy Prototype Model The Noisy Prototype model assumes that information is stored in episodic memory as categorical representations of studied features (e.g., the mean of the category to which the studied value belongs). In other words, under this model, the initial encoding of the representation is simply a pointer to the participant's prototype in that category. Other information about the studied value is not stored. Memory is simply a recall of the prototype – which we define as a sample drawn from this category, assuming a particular distribution, mean, and variance associated with it. To evaluate this model relative to the data, we calculated the probability of responses given a Gaussian distribution centered on the category prototype (i.e., mean) value given by participant ratings in Macias et al., (in revision), with noise on the category also calculated from noise given in a separate study [1].

Integrative Model The Integrative model amalgamates the assumptions of both the Noisy Target and Noisy Prototype models and assumes that recall is an integration of noisy episodic content and prior category knowledge. Under this model, prior category knowledge is used to fill in the gaps when episodic traces are noisy or incomplete. When the category representation is strong, and the memory trace is noisy, recall will resemble the category representation. The probability of responses under the Integrative model are relatively straightforward to calculate, because both the prior and likelihood distributions are Gaussian (which are self-conjugate). Furthermore, there are not specific weights assigned to the contributions of each model – this falls out naturally based on the degree of variance of each target and prototype models. We evaluate this model relative to the data, by calculating the probability of responses given the Gaussian that results from integrating these two Gaussian. Specifically, for the Integrative model, which integrates the Noisy Target and Prototype distributions, the standard solution for the mean and variance is given by,

$$\mu = \frac{1}{\frac{1}{\sigma_t^2} + \frac{n}{\sigma_p^2}}\left(\frac{t}{\sigma_t^2} + \frac{\mu_p}{\sigma_p^2}\right), \sigma = \frac{1}{\frac{1}{\sigma_t^2} + \frac{n}{\sigma_p^2}} \quad (1)$$

where σ_t refers to the memory noise on the target distribution, σ_p refers to the noise on the prototype distribution, t refers to the studied target value, μ_p refers to the mean of the prototype distribution to which the target value belong, and n=1.

Table 1: Frequency of Children Best Fit to Each Model

Model	Count(%)
Integrative	11 (33.33%)
Noisy Target	7 (21.21%)
Noisy Prototype	15 (45.45%)

In what follows, we first briefly explain the experimental methods employed by Macias et al., (in revision), to assess the role of category knowledge in episodic memory in children. We then discuss the results of the model fitting at the individual subject level in general, and age related differences, more specifically.

Experimental Methods and Results

Macias et al., (in revision) conducted two developmental experiments where they examined the relationship between prior color category knowledge and episodic memory in preschoolers (mean age: 54 mos.; range: 43 mos.-73 mos.). In the prior knowledge assessment, participants were presented with 9 color category labels (red, orange, yellow, green, blue, light blue, dark blue, purple, and pink) one at a time on a computer screen, along with a color wheel. The color wheel varied in hue only while luminance and saturation were held constant at 50 and 100 units respectively. Children were asked to point to a location on a color wheel to indicate the color that best represented the label.

In the episodic memory task, 33 participants studied 15 shapes uniquely paired with 15[2] colors, one at a time on a computer screen. At test, participants were presented with a studied shape (filled in white with a black border), along with the color wheel used in the prior knowledge assessment. The task for the participants was to choose along the color wheel to indicate the color they recalled being paired with the presented shape. For complete experimental methodology, refer to the source publication (Macias, et al., in revision).

The results of the memory task revealed a regression to the category mean effect in a majority of the studied color categories such that studied hue values that were greater than the mean of the category were underestimated and studied hue values less than the mean of the category were overestimated. This regression to the mean effect is taken as evidence of an

[1] We also assessed a model in which we sample over variance, but best fit variance matched participant responses on Macias et al.'s prior knowledge task.

[2] One of the study trials was treated as a filler in order to counterbalance presentation order and was therefore removed from the data set prior to running any analyses.

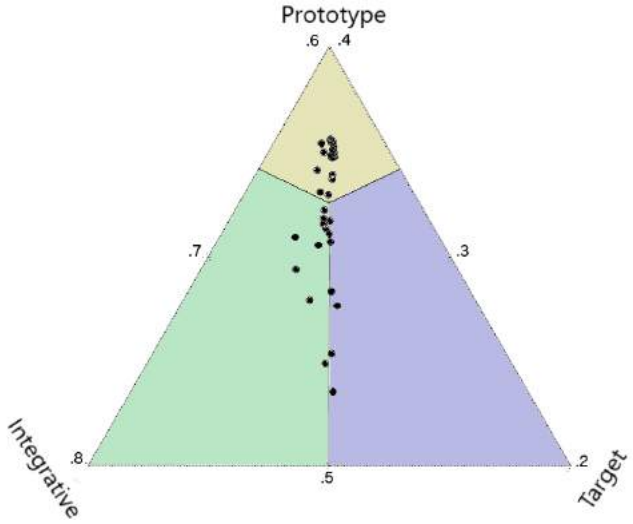


Figure 1: Ternary plot of the proportion of Log probabilities of the Integrative, Noisy Target, and Noisy Prototype models fit to each participant's data. Data points fall within the region of the model where it is best fit. Note that the figure has been zoomed in to the approximate center of the Ternary plot for better visualization of the data.

influence of category knowledge in episodic memory. Macias et al., (in revision) implemented three models and determined that the Integrative model provided the superior fit to the child data on aggregate. Here we fit the three models to individual subject data to assess for age related differences in the best fitting model.

Model Results

We sought to evaluate age and performance related differences between individual subjects and the fits of each model. Here we report the results of the model fits to children overall and then we evaluate the role of age.

Data Preparation for Evaluating Individual Differences

The data were prepared to perform four specific analyses: to evaluate individual differences in the best fitting model across the entire sample of children, to assess age related differences in the proportion of participants best described by each model, to assess additional group differences in the model fitting (e.g., the role of spontaneous labeling), and to evaluate age differences in best fitting model parameter values. We first fit the three models to each subjects' data. As with Macias et al., (in revision), the best fitting model was determined by the model with the largest log-likelihood value. To evaluate group differences, we performed a median split to classify children as younger and older learners (Table 2) and then compared the proportion of younger and older children described by each model. Of the 33 participants in the study, 16 were classified as young and 17 were classified as older. The median age of the total sample was 53 mos. (*sd*=6 mos.).

Table 2: Frequency of Model Fits by Age

Model	Count(%)	
	Young	Older
Integrative	6 (37.50%)	5 (29.41%)
Noisy Target	6 (37.50%)	1 (5.88%)
Noisy Prototype	4 (25.00%)	11 (64.71%)

The median ages for younger and older children were 49 mos. (*sd*=2 mos.) and 56 mos. (*sd*=5 mos.), respectively.

We also sought to evaluate group differences due to spontaneous labeling that was borne out of the experimental task. Of the 16 children classified as younger, 7 produced as least one label and of the 17 older children, 12 produced at least one label. This further suggests that labeling was a consistent strategy employed by children in this task. To evaluate best fitting models based on labeling, we first classified children into two groups: labelers and non-labelers. Labelers referred to learners who provided labels (at either study, test, or both) on more than 50% of trials (*n*=10/33) and non-labelers were all other children tested (*n*=23/33). We chose to use this classification because spontaneously labeling on more than 50% of trials suggests a consistent strategy of the individual to assist in recall.

To evaluate age related differences in the best fitting noise value, we implemented the Integrative model and for each participant, we searched over the space of possible noise values for the value that maximized the likelihood for each participant's data.

Model Fitting Results

Although the Integrative model is the best fitting model at the aggregate data level, it appears that at the individual level a greater proportion of children are better fit by the Noisy Prototype model (*n*=15/33), followed by the Integrative model (*n*=11/33), and then the Noisy Target model (*n*=7/33) (see Table 1). However, as can be seen in Figure 1, although a larger proportion of data points (each representing an individual child) fall towards the prototype apex, these points cluster towards the center (with near equal weight for the Target and Integrative models), suggesting that children who are classified as Prototype fits are nearly equally well fit by the other models. In contrast, for participants that are not best fit by the Prototype model, results skew significantly farther away from the center, suggesting that children who are better fit by other models are much more poorly fit by the Prototype. In light of this result, we next evaluated whether age plays a role in the proportion of children best fit by the models.

Age and Best Fitting Model To evaluate whether the proportion of children best fit by each of the three models was dependent upon age, we used the Freeman-Halton extension of the Fisher's Exact test to compute the (two-tailed) probability of obtaining a distribution of values in a 2(young vs older)x3(Integrative vs Noisy Target vs Noisy Prototype)

contingency table, given the number of observations in each cell. The results revealed that the observed proportion of best fitting models was dependent on age (p=.031). In other words, there was a significant difference in the distribution of best fitting models between the age groups. Young children were evenly split in the number fit by the Integrative (n=6) and Noisy Target (n=6) models, followed closely by the Noisy Prototype model (n=4). Interestingly, however, older children had a different composition. An much larger proportion of older children were better fit by the Noisy Prototype model (n=11), followed by the Integrative model (n=5), and almost not at all described by the Noisy Target model (n=1) (see Table 2).

Age and Best Fitting Noise Parameter Macias et al., (in revision), demonstrated that for aggregated child data, the best fitting model was the Integrative model. To evaluate the fit of the Integrative model to young learners' data, they searched for the best fitting noise parameter value. Comparing this parameter to the best fit for adults revealed a significantly larger noise parameter for the children, suggesting that as children develop the fidelity of their memory gets sharper. Here we searched for the best fitting noise value at the individual subject level to test for age related differences within the preschool population. The goal was to assess whether a difference in the amount of noise between age groups could explain why young and older children were better fit by different models. In conflict with our prediction, there was a weak non-significant *negative* correlation between age and best fitting noise value (r-= -0.17, p=.35). This suggests that a difference in the best fitting model between age group was not a result of a difference in the amount of noise in the data[3]. We return to this point later.

Additional Group Differences and Best Fitting Model Similar to the evaluation of age, we then employed a Fisher's Exact test to evaluate whether the proportion of children best fit by the three models differed between *labelers* and *non-labelers*. To reiterate, we classified labelers as children who spontaneously provided a color label on more than 50% of trials. Figure 2 shows the composition of labelers and non-labelers fit by each model. A Fisher's Exact test yielded, p=.50 , suggesting no difference in the proportion of *labelers* and *non-labelers* best fit by the three models.

Although the difference between the proportion fits was not significant, there appeared to be a trend in which most *labelers* were described by the Noisy Prototype model (60%), while *non-labelers* were more diffused across the three models. Thus, to further evaluate the role of labeling, we separated participants' label trials from the non-labeled trials, creating two new datasets. We fit the three models to the

[3] An alternative explanation is that the sample sizes for young and older children split between each model was insufficient to detect a significant difference. However, the trending direction of the data ran counter to our developmental prediction, suggesting that even if greater power revealed differences, they would be in the unpredicted direction

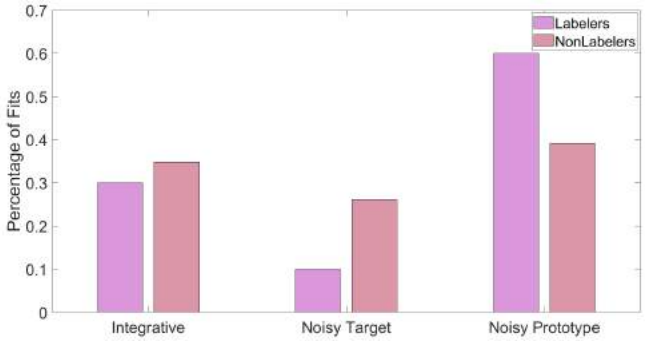


Figure 2: Proportion of Labelers and Non-labelers best fit by each model. Labelers were more likely to be best fit by the Noisy Prototype.

Table 3: Frequency of Model Fits based on Labeled and Non-Labeled Trials

Model	Count(%)	
	Label	Non-label
Integrative	3 (0.10%)	9 (27.27%)
Noisy Target	4 (0.12%)	7 (21.21%)
Noisy Prototype	26 (78.78%)	17 (51.51%)

aggregated label data and the aggregated non-label data. Unsurprisingly, the Integrative model provided the superior fit to both datasets, presumably because the model pays a lower cost for responses that, over the aggregate span between the observed target and category mean. Thus, to better understand the effects of labeling at the trial level, we then fit the three models at the individual subject level, again separating labelled trials from non-labelled trials. For the labelled trials, we found that 3 participants were best fit by the Integrative model, 4 by the Noisy Target model, but the majority of trials (26) were best fit by the Noisy Prototype model. In contrast, for the non-label trials, the distribution was less skewed, with 9 participants were best fit by the Integrative model, 7 by the Noisy Target model, and 17 by the Noisy Prototype model. A Fisher's Exact test revealed a marginally significant difference (p=.054) in the distribution of best fitting models between the labelled trials and non-labelled trials, such that most participants' label trials were best described by the Prototype model, while the non-label trials were slightly more dispersed.

Based on the finding of a difference in model fits between labeled and non-labeled trials, we re-examined the role of labeling on age. We had originally classified whole individuals as either labelers or non-labelers, and found no significant difference by age. Instead, we calculated the proportion of labeled trials provided by younger and older children, to test whether as a group, older children were more likely to provide labels during testing. A Fisher's Exact Probability Test revealed a significant difference in the proportions of la-

beled and non-labeled trials contributed by each age group (p=.002). A larger proportion of labeled trials were generated by older (66%) compared to younger children (34%).

Discussion

Our goal was to evaluate whether age-related differences persist in the strategies young learners use to reconstruct events from memory. Recent work has found that young learners, like adults, adopt the strategy of integrating prior category expectations with noisy episodic traces to reconstruct events from memory (Macias, et al., in revision). This was evidenced by a model that assumes an integration of target and category information (i.e., Integrative model) providing a superior fit to the preschool data. Here we evaluate individual differences in the best fitting strategies. We first fit three models at the individual subject level and found that the larger proportion of children were better fit by the Noisy Prototype model compared to the other models.

In addition, there were marked differences in the proportion of young and older children best fit by each model. While young children were almost evenly split in fit across the three models, surprisingly, older children were most frequently fit by the Prototype model. This result might have been bolstered by the number of trials where older children spontaneously labeled. Recall that a significantly large proportion of labeled trials belonged to older children. In this way, spontaneously labeling during study and test might have induced older children to encode and/or retrieve the prototype of the category they verbally labeled. Thus, older children may have been more likely to adopt a general strategy (labeling) that instead led to less accurate recall of the specific observation. Future work might further explore the role of spontaneous labeling on children's recall performance. For example, it is unclear whether children were still using a labeling strategy on trials where they did not spontaneously label aloud. It is possible that they were silently labeling during the task. It is unlikely that this is the case, given that we found a significant difference in performance between labeled and non-labelled trials in terms of the model fitting. However, this is an empirical for future investigation. For instance, follow up studies could use verbal interference tasks to manipulate children's ability to provide verbal labels during encoding and retrieval to evaluate whether labeling alone encourages the use of the category prototype.

What might explain the finding that the Noisy Prototype model slightly outperformed the Integrative model in terms of best fit at the individual level? First, early memory development is marked by an up-prioritization of category information over nuanced episodic information (Keresztes et al., 2018). Such behavior would equate to encoding a red color value as a prototypical shade of red (e.g., the color of a red apple) as opposed to encoding the specific shade of red studied. Thus, during study, a majority of children may have encoded target information as a pointer to the category from which the target belongs, such as a category representative (i.e., the category mean) as opposed to encoding the exact color value studied.

Alternatively, it could be the case that the use of category knowledge happens at retrieval. After the initial testing phase, the original studied information could have degraded over time and instead of reproducing the degraded information, children reproduced a value closer to the category representative to reduce error or uncertainty. Whether the influence of category knowledge occurs at encoding, retrieval, or both is a question for future research.

A third potential explanation for why a slightly great portion of children were best fit by the Noisy Prototype model might be due to the particular information studied. It should be noted that the study values for each category were selected such that they fell one standard deviation above and below the mean of the category (mean and standard deviations learned from the prior knowledge task). Given that children only studied colors that fell in close proximity of the prototypes, this might have propelled learners to rely on their category expectations, that is, adopting the Prototype strategy. Thus, the finding of a large portion of older children who are better fit by the Noisy Prototype model might be a consequence of the study values falling relatively close to the prototype. Future work might explore whether the model fitting results vary when children are presented with colors that substantially deviate from the prototype (i.e., more than 1 sd).

There were a number of limitations in this study that warrant caution in the interpretation of the results. First, the initial goal of Macias et al., (in revision), was to compare children's episodic memory performance to adults. For this purpose, a sample of 33 child participants was sufficient. However, to evaluate individual and age-related differences, a significantly larger sample of participants is needed to achieve strong statistical power for analysis. Second, the goal of this paper was to assess age related differences. Although a median split of children revealed some clear trends in a difference in model fitting by age, a more diverse age sample of children could provide further insight into differences in memory strategy by age. For instance, we anticipated that older children might rely less on the prototype to facilitate recall (although this might interact with the contrary strategy to label as children get older), but it is possible that the sample of children used here did not contain a wide enough age-range to observe this pattern. To this end, a natural future direction would be to collect more data for the purposes of evaluating age differences.

Despite these limitations, this paper demonstrates clear trends in age related differences in model fitting. Furthermore, we hope to have demonstrated that an approach that applies model fits at the individual level can provide insight into how different cognitive strategies (such as labeling) may color recall.

Acknowledgments

This work has received support from the National Science Foundation Graduate Research Fellowship under Grant Number NSF DGE 0937373 (KP), National Institutes of Health, IMSD Minority Biomedical Research Support Program under grant number 2R25GM096161-07 (CM), National Science Foundation CAREER Grant Number 1453276 (PH), NSF SES-1627971 (EB), and the Jacobs Foundation (EB).

References

Davinson, M. C., Amso, D., Anderson, L. C., & Diamond, A. (2006). Development of cognitive control and executive functions from 4 to 13 years: Evidence from manipulations of memory, inhibition, and task switching. *Neuropsychologia*, *44*, 2037–2078.

Diamond, A. (2006). The early development of executive functions. Lifespan Cognition: Mechanisms of Change. *Lifespan Cognition: Mechanisms of Change*, *210*, 70–95.

Donkin, C., Nosofsky, R., Gold, J., & Shiffrin, R. (2014). Verbal labeling, gradual decay, and sudden death in visual short-term memory. *Psychonomic Bulletin & Review*, *21*, 2–11.

Duffy, S., Huttenlocher, J., & Crawford, E. L. (2006). Developmental Science. *Children use categories to maximize accuracy in estimations*, *9*, 597–603.

Gelman, A., Carlin, J., Stern, H., & Rubin, D. (2003). *Bayesian Data Analysis*. Boca Raton, Florida: Chapman & Hall.

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal Predictions in Everyday Cognition. *Psychological Science*, *17*(9), 767–773.

Hemmer, P., & Steyvers, M. (2009). A Bayesian Account of Reconstructive Memory. *Topics in Cognitive Science*, *1*, 189–202.

Huttenlocher, J., Hedges, L. V., & Duncan, S. (1991). Categories and Particulars: Prototype effects in establishing spatial location. *Psychological Review*, *98*, 352–376.

Huttenlocher, J., Hedges, L. V., & Vevea, J. L. (2000). Why Do Categories Affect Stimulus Judgment ? *Journal of Experimental Psychology*, *129*, 220–241.

Keresztes, A., Ngo, C. T., Lindenberger, W.-B. M., U, & Newcombe, N. S. (2018). Trends in Cognitive Sciences. *Hippocampal maturation drives memory from generalization to specificity*, *22*, 676–686.

Macias, C., Persaud, K., Hemmer, P., & Bonawitz, E. (in revision). Evaluating episodic memory error in preschoolers: Category expectations influence episodic memory for color.

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, *101*, 608–631.

Persaud, K., & Hemmer, P. (2014). The Influence of Knowledge and Expectations for Color on Episodic Memory Knowledge and Expectations for Color. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th annual conference of the cognitive science society* (pp. 1162–1167). Quebec City, Canada.

Persaud, K., & Hemmer, P. (2016). The Dynamics of Fidelity over the Time Course of Long-Term Memory. *Cognitive Psychology*, *88*, 1–21.

Steyvers, G.-T., Mark, & Dennis, S. (2006). Probabilistic inference in human semantic memory. *Trends in Cognitive Sciences*, *10*, 327–334.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and bayesian inference. *Behavioral and Brain Sciences*, *24*, 629–641.

Shared Evidence: It all depends...

Toby D. Pilditch[1,2], Ulrike Hahn[3], and David Lagnado[1]

[1]Department of Experimental Psychology, University College London, 26 Bedford Way, London, WC1H 0AP, UK
[2]University of Oxford, School of Geography and the Environment, South Parks Road, Oxford, OX1 3QY, UK
[3]Department of Psychological Sciences, Birkbeck, University of London, Malet Street, London, WC1E 7HX, UK

Abstract

When reasoning about evidence, we must carefully consider the impact of different structures. For instance, if in the process of evaluating multiple reports, we find they rely on the same, *shared* evidence, then the support proffered by those reports is dependent on that evidence. Critically, normative accounts suggest that such a dependency results in redundant information across reports (reducing evidential support), relative to reports based on distinct items of evidence. In the present work we disentangle the structural and observation-based indicators of this form of dependency. In so doing, we present novel findings that lay reasoners are not only insensitive to shared evidence structures when updating their beliefs, but also that reasoners do not necessarily prefer more diverse sources of evidence. Finally, we replicate prior effects in reasoning under uncertainty, including conservative sequential updating, and difficulty in integrating contradictory reports.

Keywords: evidential reasoning; probabilistic reasoning; dependence; Bayesian Networks; belief updating

Introduction

Over the course of an investigation, you are faced with the weighing up of contradicting reports. Two of your investigators confirm the hypothesis, whilst two disconfirm it. How do you discern which pair may carry more (evidential) weight? One important aspect is what evidence those investigators are relying upon. For instance, if your two confirming investigators are relying on the *same* piece of evidence to inform their reports, whilst the two disconfirming investigators are relying on separate, independent pieces of evidence, then, *ceterus paribus*, the standard intuition is to side with the disconfimers.

This example highlights the traditional understanding of one form of dependency in evidential reasoning. Specifically, the notion of "shared" evidence (Schum & Martin, 1982; Schum, 1994), which is considered to be inferior to reports based on distinct (separate) evidence, i.e. dependence as a form of redundancy (Hogarth, 1989; Schum & Martin, 1982; Soll, 1999).

How such information *should* be integrated is important to a number of areas, from everyday reasoning to investigative domains such as medicine (Eddy, 1982), law (Faigman & Baglioni Jr, 1988; Fenton & Neil, 2012, Fenton, Neil & Lagnado, 2013, Harris & Hahn, 2009; Lagnado, 2011; Pennington & Hastie, 1986; Schum, 1994), risk analysis (Fenton & Neil, 2012), and to the intelligence community (Heuer, 1999). Consequently, failures to account for such dependencies between evidence items – although easing computation (Pearl, 1988; Schum, 1994) – can lead to deleterious overweighting of the support provided by such evidence (e.g., naïve Bayes in medicine – where evidence is assumed to always be independent; Koller & Friedman, 2009; Kononenko, 1993).

The notion of shared-evidence as a form of dependence fits with the correlation-based conceptualisation of dependencies as a form of redundancy in prediction errors (e.g. Soll, 1999). More precisely, when two sources are using the same evidence to inform their reports, vs the same two sources using two different items of evidence, the former case results in an "overlap" of information provided (Schum, 1994). It thus becomes more likely that reports in the former case rely on the *same* information, and the pair of reports therefore carry some redundant information. As a consequence, such correlated reports provide a lesser degree of support for the hypothesis being informed upon.

In the present work, we seek to provide an empirical baseline for lay reasoners judgments regarding this form of dependence. We not only investigate whether belief-updating is in line with the shared-evidence-as-inferior hypothesis proposed in formal work, but whether lay reasoners seek more diverse evidence in their search preferences.

Formalising reasoning about shared evidence

To illustrate what is meant by shared evidence, Fig. 1 below presents a directed acyclic graph (DAG) of an example case. Here there is a hypothesis under investigation (H), three pieces of evidence that inform that hypothesis (E1-3), and four sources (or witnesses) who in turn *report* on said evidence (R1-4). Crucially, the evidence itself remains unobserved, so we are instead trying to infer diagnostically about H (via E1-3) from the reports provided by R1-4, and notably how to judge R1 and R2 (who rely on the same evidence, E1), versus R3 and R4 (who rely on separate evidence, E2 and E3 respectively).

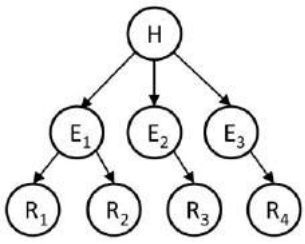


Figure 1. Graphical representation of a hypothesis (H), evidence items that inform upon it (E1-E3), and sources informing their reports upon said evidence (R1-R4).

To understand how reasoners *should* update their beliefs given these observed reports, we use a Bayesian Network (BN) formalism, wherein a DAG is supplemented by conditional probabilities and the use of Bayes theorem, so as to make optimal (i.e. inaccuracy minimization) inferences (Pearl, 1988; 2009). This computational framework for reasoning under uncertainty has been used effectively to model (and shed light on human inferences by comparison) direct dependencies between sources (Pilditch, Hahn & Lagnado, 2018), and dependencies as shared-backgrounds among sources (Madsen, Hahn & Pilditch, 2018), and integration across sources of differing reliabilities (Phillips, Hahn, & Pilditch, 2018).

If we assume, in the above example, that all evidence items are equally diagnostic[1], and all sources are equally reliable[2], then the sole difference-maker between sources is the structural difference entailed by the shared evidence (E1). To best illustrate the impact of shared evidence, we first consider the point at which we have only observed a confirmatory report from R1. Via conditionalization, E1 is now already more likely to be confirmatory than E2 and E3. Given this, if we are to decide whether we want to see a report from R2 (who also relies on E1), or a report from R3, a confirming report from the latter provides *more potential information* regarding H, given that P(E1) – which is already more probable, given the report from R1 – increases less given R2 than P(E2) does, given R3.

Present research We seek to empirically test the degree to which lay reasoners are sensitive to the impact of shared evidence structures on belief updating. More precisely, we use the above formalism to provide an empirical baseline for lay reasoning regarding such dependencies, and notably whether participant probability estimates fit with normative predictions of dependence inferiority. Additionally, we explore two research questions that the formalism allows us to investigate, via the separation of *structural* dependencies from dependencies inferred from (correlated) observations:

First, how do reasoners deal with contradiction across a shared evidence item (as opposed to contradiction across different evidence items)? Recent research that exploits the capacity to tease apart the structural form of a dependency from the dependency inferred from (correlated) observation – as possible in the present work – exposes lay reasoner difficulties in accurately updating (both qualitatively and quantitatively) when an observed contradiction occurs *across* a structural dependency (i.e. information is directly shared from one source to another equally reliable source, yet those sources then disagree; Pilditch, Hahn, & Lagnado, 2018). We predict the same difficulty here.

Second, the present work allows for the investigation into evidence diversity preferences. The computational framework underpinning this work allows for the calculation of the predicted informative value of evidence items, for which we calculate the Kullback-Liebler Divergence (KL-D; a measure of entropy reduction; Kullback & Liebler, 1951)[3].

$$KL(E_j) = \sum P(h_i|e_j) * \log\left(\frac{P(h_i|e_j)}{P(h_i)}\right)$$

where E_j is a set of items of evidence $\{E_1, E_{2...}E_j\}$, e_i the set of possible states of the evidence, $\{e_1, e_2, e_i\}$, and h_i is a set of hypotheses, $\{h_1, h_{2...}h_i\}$. In the present case, we compute the information provided by R2 in reference to the hypothesis (H; given we have already observed R1) when a) R2 also relies on E1, vs b) R2 relies on E2, taking a difference measure between these two values.

As such, in asking lay reasoners for their preference for a forthcoming report to be based on shared evidence (i.e. based on an item of evidence already informed by one report) or new evidence, we may observe whether lay reasoning (if in line with normative expectations) predicates an evidence selection preference for more diverse items.

In sum, the present work uses a BN formalism to disentangle the structural vs observation-based forms of shared evidence dependencies. In so doing we are able to not only establish an empirical baseline of when the two forms agree (and thus whether reasoners fit with standard normative expectations), but also examine how reasoners deal with cases of disagreement (where observations appear uncorrelated, but a structural dependence remains), and use structural relations to determine (diversity-based) evidence preferences.

Method

Participants 200 US participants were recruited and participated online through Amazon Mechanical Turk. Three participants were removed for incomplete data, and 1 for not being a native English speaker. Of the 196 remaining participants, 84 identified as female, and the median age was 34 (*SD* = 9.8). All participants gave informed consent, and were paid for their time (*Mdn* = 8.74 minutes, *SD* = 6.63).

Procedure & Design Participants were presented with a scenario in which *a patient, "RN", may have a disease "MTL"* ("H" in Fig. 1). The participant is placed in the role of a diagnostician, attempting to confirm the above diagnosis. They are informed that the patient has had a number of *cell samples* taken (these can be considered $E_{1\text{-}3}$ in Fig. 1), both independently, and of equal diagnosticity. More precisely, that each *cell sample* may contain a *biomarker*, which has a 90% chance of being due to MTL

[1] I.e. P(E1|H) = P(E2|H) = P(E3|H), and P(E1|¬H) = P(E2|¬H) = P(E3|¬H).

[2] I.e. P(R1|E1) = P(R2|E1) = P(R3|E2) = P(R4|E3), and P(R1|¬E1) = P(R2|¬E1) = P(R3|¬E2) = P(R4|¬E3).

[3] Other information measures exist, such as impact (see Nelson, 2005), information gain (Lindley, 1956), and Bayesian diagnosticity (Good, 1950), though empirical work suggests such measures are highly correlated (Nelson, 2005), and are thus considered interchangeable for the present work.

(hit rate), but also a 10% probability of being a false positive.

Participants are then informed that they are unable to examine the *cell samples* themselves, but must rely on *lab technicians* (R1-4 in Fig. 1), who will independently examine the *cell samples* and provide a *report of whether biomarkers are present or absent*. Crucially, all the lab technicians are indicated as equally reliable, in that they have an 80% chance of detecting and reporting a biomarker (irrespective of whether it is due to MTL), when a biomarker is present (hit rate), and a 20% chance of a false positive.

Lastly, participants were informed that prior to receiving any reports from their lab technicians, given the facts of the case so far, they should assume a prior probability of patient RN having MTL of 50% ("Finally, *prior to getting the reports*, you can assume an initial probability of 50% that patient RN has MTL, based on the facts of the diagnostic process so far… Before you start finding out reports, please answer the following question … What is the probability that **patient RN has MTL**?"). This prior probability was then immediately elicited from participants, for use in individual model fitting (see results section below).

Elicitation Stages Participants then received reports from each of 4 lab technicians in turn (resulting in a total of 4 elicitation stages). Following each new report, participants were asked to provide a new probability estimate of patient RN having MTL – given *everything they now know* (i.e. background + gradually accumulating reports). These probability estimates were the main dependent variable.

Each report statement took the form "*Based on their assessment of* ***cell sample [1/2/3]****, lab tech [1/2/3/4] reports that the* ***biomarker is [present/absent]***."

Crucially, there were two independent, between-subject variables employed, making a 2x2 design. The first of these was the *evidence used by the second lab technician* ("R2Evidence"). Whilst the first lab technician always used cell sample 1, the third cell sample 2, and the fourth cell sample 3, the second lab technician used cell sample 1 in one condition (R2E1), and cell sample 2 in the other (R2E2). This allowed for a) the between-subject comparison of 2 reporters using independent (R2E2) vs shared evidence (R2E1), and b) allowed for the disentanglement of *structure* (i.e. dependency relations) from *order of observations* (i.e. is over/under updating due to the second report relying on shared evidence, or simply because it is the *second* report).

The second between subject factor was the order of positive (biomarker present) and negative (biomarker absent) reports ("RepOrder"). More precisely, either the first lab technicians 1 and 2 gave positive reports (and 3 & 4 gave negative reports; "PosFirst"), or the reverse ("NegFirst"). This general structuring, when taken in conjunction with the R2Evidence factor, allowed for the assessment of the influence of shared evidence when reporters agree about the same evidence (R2E1) or disagree (R2E2 – as lab technicians 2 & 3 will always disagree, yet will share cell sample 2). Additionally, this allows for the further disentanglement of observation *type* from shared evidence (structural) influences. For instance, in R2E1 conditions, the reports from shared evidence (lab technicians 1 & 2) will half the time be positive, and the reports from independent evidence (lab technicians 3 & 4) will half the time be negative, and vice versa. Thus, one may discern the influence of (dis)confirming observations vs structural differences.

Dependent variables Along with the probability estimates elicited at each elicitation stage (0-100% slider, no default)[4], one further qualitative question was asked after the first lab technician provided a report (i.e. elicitation stage 1):

"Given the choice, would you rather *Lab Tech 2* **also independently investigated cell sample 1 for a biomarker, or investigated a different cell sample (cell sample 2)**?" ["Same cell sample (cell sample 1)" / "Different cell sample (cell sample 2)" / "There is no difference."] Forced choice, randomized order of presentation.

The purpose of this question was to assess participant preferences for diversity (independence in this case) in their observations.

Taken together, the probability estimates and evidence preference judgment allow for the assessment of the impact of shared evidence, both in terms of predicted support, and consequent reasoning (and belief-updating), whilst taking into account the influence of observation types and orders.

To concretize the research questions into hypotheses, we predict:

H1. Shared Evidence Structure - Shared evidence will result in estimates of reduced impact of affected reports, in comparison to reports from distinct items of evidence, *ceterus paribus*. Tested via the between subject comparison of the impact of *lab technician 2* at the second elicitation stage when *lab technician 2* does/does not share *cell sample 1* with *lab technician 1* (i.e. R2E1 vs R2E2 conditions).

H2. Contradictions and Dependence – Reasoners will find the integration of two contradicting reporters using the *same* evidence (i.e. *within* a shared evidence item) more difficult than when contradictions are based on separate evidence items. Tested via the comparison to normative expectation at third elicitation stage (when *lab technician 1, lab technician 2,* and *lab technician 3* have reported) in R2E2 condition in comparison to R2E1 condition, where contradiction cuts across (rather than within) evidence items. Such a prediction is informed by previous work that has found lay reasoners struggle with inferences from contradicting reports across a dependency (Pilditch, Hahn, & Lagnado, 2018).

H3. Diversity Preference - Participants (correctly) prefer more diverse evidence (prefer *lab technician 2* to use evidence *cell sample 2*.

i. An additional question of interest is whether diversity preferences will be lower when *lab*

[4] Open text reasoning responses were also collected at the end of each elicitation stage, but for the sake of brevity are not reported here.

technician 1 provides negative evidence (NegFirst condition), than when *lab technician 1* provides positive evidence (PosFirst)?

Fig. 2 below shows the different structural and report order comparisons for the 2x2 design (each cell is a between subject condition), with T1 to T4 within each cell as the within-subject order of evidence. Thus, H1 is investigated by comparing T2 in the top row (when R2 is reliant on the same evidence as R1) with T2 on the bottom row (when R2 is using different evidence). We can then assess H2 by comparing T2 to T3 in the top row (contradiction based on separate items) to the bottom row (contradiction based on shared evidence. H3 is assessed having seen the report at T1 (and is asked prospectively about T2), and H3i. is based on the comparison of responses to the H3 question in left versus right columns of Fig. 2.

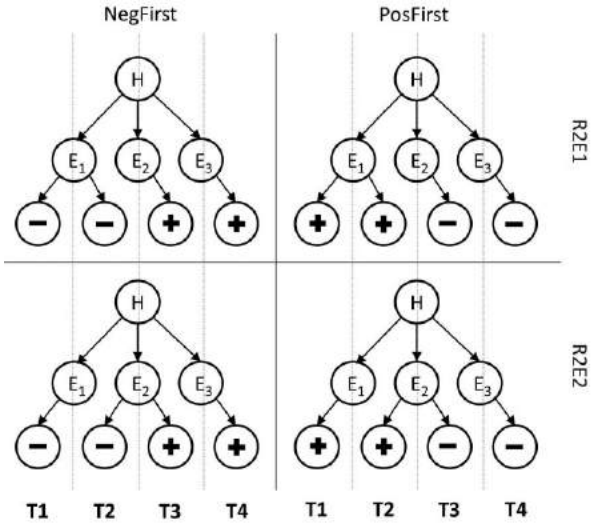


Figure 2. Underlying networks, split by R2Evidence (rows) and RepOrder (columns) conditions. T1 to T4 reflect sequence of reports (within-subjects).

Results

Bayesian statistics were employed throughout[5] using the JASP statistical software (JASP Team, 2018). Using the gRain package in R (Højsgaard, 2012), the elicited priors from each participant were used to individually fit BNs for each participant. Remaining parameters were as specified in the background information presented to participants. The posterior probabilities at each elicitation stage generated from each BIBN model (representing each participant) were used in subsequent comparison analyses.

Probability Estimates

The hypothesis-directed analyses used to unpack a) the influence of when shared evidence is introduced (*H1*), and b) the influence of contradiction within/outside a dependency (*H2*), first employed an RM-ANOVA on participant estimates alone (including between subject factors), so as to determine participant behavior, followed by a further analysis that compared these estimates to BIBN predictions, to determine the "correctness" of this behaviour.

H1. Firstly, to assess *H1*, an RM-ANOVA on participant estimates from T1 to T2, found participants were insensitive to R2Evidence condition overall, $BF_{Inclusion}$ = 0.102, or in interaction with elicitation stage, $BF_{Inclusion}$ = 0.105. This was despite participants updating in light of new evidence *in general*, $BF_{Inclusion}$ > 10000, and whether that evidence was positive or negative, $BF_{Inclusion}$ > 10000. This was further evidenced by the interaction of elicitation stage and RepOrder (participants decreased estimates as negative reports came in, and increased as positive reports came in), $BF_{Inclusion}$ > 10000. Consequently, the model of participant estimates without R2Evidence yielded the strongest fit, BF_M = 63.2, and was decisive overall, BF_{10} > 10000.

Consequently, by subsequent inclusion of the BIBN predictions for each participant (the Observed vs Predicted factor), this insensitivity to shared evidence (i.e. the influence of R2Evidence, was shown to be insufficient relative to (fitted) normative expectation. This was evidenced by a main effect of Observed vs Predicted, $BF_{Inclusion}$ = 5479.38, and critically, strong evidence for the interaction of R2Evidience and Observed vs Predicted (BIBN predictions changed with R2Evidence, whilst participant estimates do not), $BF_{Inclusion}$ = 11.72.

In sum, these analyses revealed participants were insensitive to impact of shared evidence structures, as compared to their fitted Bayesian predictions (*H1*).

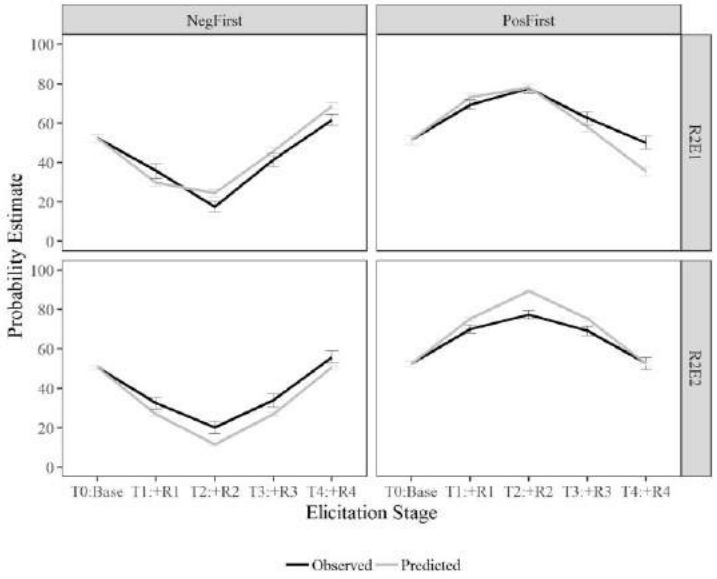


Figure 3. Probability estimates across elicitation stages, split by condition. Error bars reflect standard error.

H2. Secondly, to assess *H2*, the same analytical protocol was used on elicitation stages T2 to T3. This corroborated H1 findings, in that the insensitivity to the influence of

[5] For all analyses, an uninformed prior was used. Wherever possible, sample sizes for a given analysis (*N*), and Bayesian Credibility Intervals (95% CI) are indicated.

R2Evidence (this time via the presence of shared evidence in concert with contradicting reports either outside, R2E1, or within, R2E2, the same shared evidence) persisted both overall, $BF_{Inclusion} = 0.315$, and in interaction with elicitation stage, $BF_{Inclusion} = 0.321$. However, once again participants were sensitive to the introduction of new evidence in general, $BF_{Inclusion} > 10000$, its valence, $BF_{Inclusion} > 10000$, and the interaction of these factors (newly introduced positive reports lead to increased estimates, whilst newly introduced negative reports lead to decreased estimates), $BF_{Inclusion} > 10000$. As with the *H1* analysis, the model of participant estimates without R2Evidence yielded the strongest fit, $BF_M = 20.425$, and was decisive overall, $BF_{10} > 10000$.

To again determine whether this insensitivity was erroneous, BIBN predictions for each participant were included as another within subject factor (Observed vs Predicted). Again, participant estimates were shown to not only be generally insufficient in comparison to BIBN predictions, $BF_{Inclusion} > 10000$, but that this insensitivity extended to shared evidence (R2Evidence x Observed vs Predicted; BIBN estimates change with condition, participant estimates do not), $BF_{Inclusion} > 10000$.

In conclusion, the above analyses corroborate the insensitivity findings of *H1*, extending them to the issue of contradiction (of reports) being based on the same or different evidence items (*H2*).

Taken together, *H1* and *H2* findings suggest participants were insensitive to the impact of shared evidence, both when reporters are corroborating with, and contradicting each other.

Evidence Preference

The BIBN models for each participant, having taken into account the elicited prior for the hypothesis, generated the expected information gained in KL-D, having observed the positive/negative report from the first lab technician, for two models; one in which the second lab technician used the same evidence as the first (E1), and one where the second lab technician used different evidence (E2). The difference in expected information gain between these two models was used to generate a normative preference (based on maximum expected information) for the second lab technician using E1, E2, or them being equivalent ("NoPref").

To assess the observed evidence preferences, a Bayesian binomial test was conducted on observed preferences (dark grey bars of Fig. 4), comparing them to chance responding (0.33). Preferences for the second lab technician to use the *same* evidence as the first lab technician (E1) were found to be at chance level (0.36, 95% CI: [0.293, 0.426]; $N = 196$), $BF_{10} = 0.118$, whilst diversity preferences (second lab technician to use E2) were found to occur decisively above chance (0.51, 95% CI: [0.441, 0.579]; $N = 196$), $BF_{10} > 10000$, lending some support to the diversity preference predicted (*H3*). The frequency of participants opting for "no preference" was decisively below that expected by chance, (0.13, 95% CI: [0.092, 0.187]; $N = 196$), $BF_{10} > 10000$. A Bayesian contingency table revealed these frequencies to not be influenced by whether the first lab technician had made a positive (right-hand facet of Fig. 4) or negative (left-hand facet of Fig. 4) report ($N = 196$), $BF_{10} = 0.045$, speaking against hypothesis *H3i*.

Crucially, participant preferences for the second lab technician to use the *same* evidence as the first lab technician are substantially higher than that predicted by BIBN models (i.e. 0; see light grey bars of Fig. 4). This is corroborated by the decisive deviation in frequencies between observed and predicted preferences ($N = 392$), $BF_{10} > 10000$. Put another way, and contrary to predictions of *H3*, approximately 1/3rd of participants retain an explicit preference for the information-poorer reports that "confirm" (i.e. are based on evidence that has already formed the basis of an observed report), rather than a diversity preference or lack of preference.

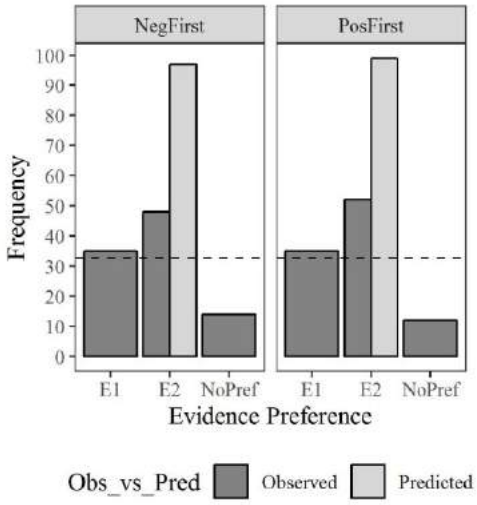


Figure 4. Evidence Preferences, split by condition. Dashed line represents chance level (33%).

Conclusions

When reasoning under uncertainty, an important consideration is the impact of dependencies among evidence items. More precisely, seemingly separate reports, which in fact stem from the same source (or evidence basis), carry redundant information, relative to truly separate reports (based on distinct information). To mistake the former for the latter can lead to overweighting support for a given hypothesis, to deleterious consequences (Dror et al., 2006; Koller & Friedman, 2009).

Here, we show that lay reasoners seem rather insensitive to the impact of this form of dependency and consider the two cases equivalent when estimating degrees of support for a hypothesis. At the same time, our findings corroborate prior research in terms of both a) the consistent under-weighting of introduced evidence (see e.g. Faigman &

Baglioni, 1988; Nance & Morris, 2005), and b) more substantial deviations when having to deal with contradictory reports (Pilditch et al., 2018).

Finally, we present a second novel finding in lay reasoners preferences for further reports based on shared (i.e. previously informed upon) evidence (a "confirmatory" preference) or separate (unseen) evidence (a "diversity" preference). Though the majority of participants conform to a diversity preference in line with maximising expected information, approximately 1/3rd of lay reasoners have a confirmatory preference. While failures to appreciate diversity have been reported before (e.g., Soll, 1999), there are clear preferences for diversity in other inferential contexts (e.g., Rips, 1979; Osherson et al., 1990), even in children (Heit & Hahn, 2001). Hence further work will be required to pinpoint exactly for when, where and why diversity is appreciated and when it is not. It is worth note in this context that where the reliability of the reporting sources is not exactly known (unlike the lab technicians in the present study), diverse evidence is arguably not always normatively superior (see Bovens & Hartmann, 2003). Whether lay reasoners have any understanding of the different circumstances where a diversity advantage does and does not obtain remains to be seen.

References

Bovens, L., & Hartmann, S. (2003). *Bayesian epistemology*. Oxford University Press on Demand.

Dror, I. E., Charlton, D., & Péron, A. E. (2006). Contextual information renders experts vulnerable to making erroneous identifications. *Forensic science international*, 156(1), 74-78.

Eddy, D. M. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In Kahneman, D., Slovic, P., & Tversky, A. (eds.), *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge University Press. pp. 249--267.

Fenton, N., & Neil, M. (2012). *Risk assessment and decision analysis with Bayesian networks*. Crc Press.

Fenton, N., Neil, M., & Lagnado, D. A. (2013). A general structure for legal arguments about evidence using Bayesian networks. *Cognitive science*, *37*(1), 61-102.

Faigman, D. L., & Baglioni Jr, A. J. (1988). Bayes' theorem in the trial process: Instructing jurors on the value of statistical evidence. *Law and Human Behavior*, *12*(1), 1.

Good, I. J. (1950). *Probability and the weighing of evidence*. New York: Griffin.

Harris, A.J.L., & Hahn, U. (2009). Bayesian Rationality in Evaluating Multiple Testimonies: Incorporating the Role of Coherence. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35, 1366–1372.

Heit, E. H. U. (2001). Diversity-Based Reasoning in Children. *Cognitive Psychology*, *43*, 243–273.

Heuer, R. J. (1999). *Psychology of intelligence analysis*. Washington DC: Center for Study of Intelligence.

Højsgaard, S. (2012). Graphical independence networks with the gRain package for R. *Journal of Statistical Software*, *46*(10), 1-26.

Hogarth, R. M. (1989). On combining diagnostic ''forecasts'': Thoughts and some evidence. *International Journal of Forecasting,* 5, 593–597.

JASP Team (2018). JASP (Version 0.9)[Computer software].

Koller, D., & Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. MIT press.

Kononenko, I. (1993). Inductive and Bayesian learning in medical diagnosis. *Applied Artificial Intelligence an International Journal*, *7*(4), 317-337.

Kullback, S., & Liebler, R. A. (1951). Information and sufficiency. *Annals of Mathematical Statistics*, 22, 79–86.

Lagnado, D. A. (2011). Thinking about evidence. In *Proceedings of the British Academy* (Vol. 171, pp. 183-223).

Lindley, D.V. (1956). On a measure of the information provided by an experiment. *Annals of Mathematical Statistics*, 27, 986–1005.

Madsen, J. K., Hahn, U., & Pilditch, T. D. (2018). Partial source dependence and reliability revision: the impact of shared backgrounds. In T.T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 722-727). Austin, TX: Cognitive Science Society.

Nance, D. A., & Morris, S. B. (2005). Juror understanding of DNA evidence: An empirical assessment of presentation formats for trace evidence with a relatively small random-match probability. *The Journal of Legal Studies*, *34*(2), 395-444.

Nelson, J. D. (2005). Finding useful questions: On Bayesian diagnosticity, probability, impact, and information gain. *Psychological Review*, 112(4), 979–999.

Osherson, D. N., Smith, E. E., Wilkie, O., Lopez, A., & Shafir, E. (1990). Category-based induction. *Psychological review*, *97*(2), 185.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems*. San Francisco, CA: Morgan Kaufmann.

Pearl, J. (2009). *Causality. Models, reasoning, and inference*. Second edition. New York: Cambridge University Press.

Pennington, N., & Hastie, R. (1986). Evidence evaluation in complex decision making. *Journal of Personality and Social Psychology*, *51*(2), 242.

Phillips, K., Hahn, U., & Pilditch, T. D. (2018). Evaluating testimony from multiple witnesses: single cue satisficing or integration? In T.T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 2244-2249). Austin, TX: Cognitive Science Society.

Pilditch, T.D., Hahn, U., & Lagnado, D. (2018). Integrating dependent evidence: naïve reasoning in the face of complexity. In T.T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.), *Proceedings of the 40th Annual Conference*

of the Cognitive Science Society (pp. 884-889). Austin, TX: Cognitive Science Society.

Rips, L. J. (1975). Inductive judgments about natural categories. *Journal of Memory and Language*, *14*(6), 665.

Schum, D. A. (1994). *The evidential foundations of probabilistic reasoning*. Northwestern University Press.

Schum, D. A., & Martin, A. W. (1982). Formal and empirical research on cascaded inference in jurisprudence. *Law and Society Review*, 105-151.

Soll, J. B. (1999). Intuitive theories of information: Beliefs about the value of redundancy. *Cognitive Psychology*, *38*(2), 317-346.

Asymmetric belief sensitivity and justification explain the Wells Effect

N. Ángel Pinillos (pinillos@asu.edu)
School of Historical Philosophical and Religious Studies, Arizona State University
Phoenix, AZ, USA

Sara Jaramillo (sdjarami@asu.edu)
School of Social and Behavioral Sciences, Arizona State University
Phoenix, AZ, USA

Zachary Horne (Zachary.Horne@asu.edu)
School of Social and Behavioral Sciences, Arizona State University
Phoenix, AZ, USA

Abstract

Wells (1992) found that jurors are more likely to find a defendant guilty when the evidence against them is 'specific' (that is, when the evidence provides a causal mechanism for how an event occurred) as opposed to being based on base-rate information, or what Wells calls 'general' evidence. Enoch, Spectre, and Fisher (2012) propose that this epistemic difference can be explained by the "sensitivity" of beliefs formed on the basis of these two types of evidence where sensitivity is understood as a counterfactual condition on knowledge judgments. They argue that beliefs are sensitive when formed on the basis of specific evidence, but not when they are formed on the basis of general evidence. In two preregistered experiments, we tested this hypothesis. We replicated an earlier finding that specific, as opposed to general evidence, is more likely to lead to knowledge judgments. Consistent with the hypothesis of Enoch and colleagues, we also found that sensitivity partially mediates the relationship between evidence type and knowledge attributions.

Keywords: Wells effect; counterfactual reasoning; knowledge sensitivity; open science

Introduction

Imagine that you are a juror deliberating on a civil suit against the Blue bus company which is being accused of running over a dog during one of its routes. The only evidence that is presented to you is that a commercial bus killed a dog and that 80% of the commercial buses in the area are run by the Blue bus company. Is the Blue bus company guilty of killing the dog? According to Wells (1992), most people do not think you should convict on this evidence. Now suppose instead that the only evidence presented to you is that a witness reports seeing a Blue bus run over the dog and that this witness is 80% reliable. In this case, is the evidence sufficient for finding the Blue bus company guilty of killing the dog? According to Wells (1992), people are more likely to find a defendant guilty when the second kind of evidence, termed specific evidence, is given than when the first kind of evidence is given (termed "general" evidence). Strikingly, people exhibit this pattern of judgments even though they also perceive the probability of guilt as being the same in both cases. The tendency to trust specific more than general evidence is also reflected in the law. As Enoch and Fisher (2015) report, "Courts and legal scholars often view [general evidence] with suspicion, treating it as inadmissible even when it is probabilistically equivalent to individualized [specific] evidence." The explanation of the Wells effect is therefore not only important to understanding how people reason about evidence but is also pertinent to the law.

What explains the Wells effect? One possibility is that in the first case, the evidence is purely general in the sense that it is not causally connected to the actual killing of the dog: the fact that 80% of the buses are operated by the Blue bus company is, in a sense, independent of the actual killing of the dog. In contrast, the evidence in the second case is specific to the case; there is a causal connection between a witness reporting the Blue Bus company killed the dog and the dog being killed. Consistent with this possible explanation, empirical accounts on the Wells Effect often focus on *counterfactual thinking*—the ability to consider what could have happened but did not happen. For example, Niedermeier, Horowitz, & Kerr (1999) and Sykes and Johnson (1999) account for the reluctance to find the defendant guilty in general cases by appealing to the ease in which the possibility that the defendant is innocent comes to mind. Along these lines, Enoch et al. (2012) proposed that general evidence is considered weaker than specific evidence because beliefs formed on the basis of the former, but not the latter, are sensitive (a term which we explain below). The main goal of this paper is to test this hypothesis.

Many philosophers think that the notion of sensitive belief can explain a wide range of judgments similar to the Wells Effect (Black & Murphy, 2007; DeRose, 1999; Dretske, 1981; Ichikawa, 2011; Nozick, 1981; Roush, 2005). Philosophers typically define a 'sensitive belief' as follows:

> An agent's belief that *P* is 'sensitive' just in case If *P* were false, then the agent would no longer believe *P*.

For example, your belief that you are reading a paper right now is sensitive. This is because if you weren't reading a paper, you would be making coffee, let us suppose, and so would no longer believe you were reading a paper. Sensitive beliefs are counterfactually robust in the sense that they "track" the facts in possible circumstances. This is why

some philosophers have held that if an agent knows *P*, then her belief that *P* has to be sensitive. Let's call this thesis itself 'sensitivity'. Sensitivity is a type of counterfactual condition on knowledge.

Take another example: Consider a person who believes that his beloved pet is healthy, not through good evidence, but because of wishful thinking. We can suppose further that his pet is in fact healthy. Our intuitive reaction is that this agent does not really know his pet is healthy. But why doesn't the agent know this? According to the sensitivity account, the reason is that this agent's belief is not sensitive. For example, although his pet is actually healthy, if the pet were sick, the wishful-thinking agent would continue to think that the pet was healthy. Because his belief is not properly tracking the facts in the world, he doesn't really know his pet is healthy. In contrast, suppose the agent believes his pet is healthy, not through wishful thinking, but because he got a good report from a reliable vet. In this case, his belief is more likely to be sensitive (and hence truth tracking) because if his pet were sick, he would believe that his pet was sick. This is because if his pet were sick, the reliable vet would have told him so.

These two examples demonstrate how the sensitivity hypothesis is supposed to account for our epistemic intuitions. For our purposes, what is important is that sensitivity gives us a counterfactual condition on the perceived strength of one's evidence. We know that counterfactual reasoning also plays a role in a number of other cognitive phenomena. It has been implicated in planning and prediction (Barbey & Sloman, 2007; Epstude & Roese, 2008; Markman, McMullen, & Elizaga, 2008; Roese, 1999; Smallman & Roese, 2009; Tobia, Guo, Schwarze, Boehmer, Gläsher, Finckh, & Sommer, 2014), generating emotions (Alicke, Buckingham, Zell, & Davis, 2008; Brassen, Gamer, Peters, Gluth, & Buchel, 2012; Coricelli & Rustichini, 2010; Davis, Lehman, Wortman, Silver, & Thompson, 1995; Miller, Markman, Wagner, & Hunt, 2013; Pieters & Zeelenberg, 2005; Roese & Olson, 1997, 2007), learning (Byrne, 1997; Epstude & Roese, 2008; Smallman & McCulloch, 2012), as well as in moral and causal reasoning (Halpern & Hitchcock, 2015; Malle, Guglielmo, & Monroe, 2014). Although counterfactual reasoning has been found to play a large role in thinking across many psychological domains, we are not aware of any experimental work *directly* examining how counterfactual judgments, such as those involved in the definition of sensitivity, relate to attributions of knowledge. This is notable because, as discussed, counterfactual reasoning is prominently featured in theories of knowledge attributions in the philosophical literature.

In the present studies, we investigated whether people's judgments about knowledge can be explained by the sensitivity hypothesis. Specifically, we sought to test this sensitivity account as an explanation for the Wells effect, and related effects central to ongoing debates in epistemology (Friedman & Turri, 2014). The original Wells effect concerned judgments about whether a defendant should be found guilty. Following Wells and our discussion so far, our approach is to investigate not determinations of guilt, but attributions of knowledge. We predicted that people differentiate general and specific evidence because of their differential sensitivity—beliefs formed on general evidence are less sensitive than beliefs formed on specific evidence.

General Methods

Analytic Approach To test our hypotheses, we performed Bayesian mixed-effects modeling using the R package `brms` (Bürkner, 2018). We set regularizing priors for all population-level effects in our models, which we detail below. These priors are recommended because they provide conservative effect size estimates and reduce the likelihood of overfitting (Gelman, Lee, & Guo, 2015; McElreath, 2016). Following the recommendations of Liddell & Kruschke (2018), Likert data were modeled with a cumulative probability distribution. The cumulative distribution is recommended for Likert scale data because it assumes that ordered responses represent a continuous latent construct (in this case, the tendency to attribute knowledge to an agent).

Preregistration We preregistered the data collection plan, analyses, and predictions for both experiments. Experimental scripts, analyses, and supplementary online materials are available on the Open Science Framework at https://osf.io/pw7s8/.

Experiment 1

Participants We powered our study to detect a Cohen's d of .2 for a two-condition within-subjects design with 80% power. To this end, 201 participants were recruited through Amazon's Mechanical Turk (43% women, M_{age} = 36 years old). Participants were paid $0.50 for participating in the study. Participants were excluded for missing questions checking their attention (e.g., "Select perhaps knows from the options below"). Our exclusion criteria were determined a priori and were in accordance with our study preregistration. After excluding participants who missed questions checking their attention, 170 participants remained in our sample.

Procedure In Experiment 1, we examined whether the Wells effect can be explained by the sensitivity hypothesis. To test this, we randomly presented participants with six scenarios in a within-subjects design (three topics $\times$ two conditions = six vignettes) adapted from Friedman and Turri (2014). In these scenarios, a protagonist is described as researching a question in which they rely on either general, base-rate information or specific, mechanistic information to draw their conclusion. For example, participants considered the following scenarios:

> *General*: Bob wonders if his spider plant contains the chemical aracnium. He consults a very reliable book on

spider plants. The book says that [98 or 99% randomly presented] of spider plants contain aracnium. So Bob concludes that his spider plant contains aracnium. And he is right: it contains aracnium.

Specific: Joseph wonders if his stinkwood flower will grow yellow stalks. He conducts a very reliable DNA test on the flower. The test shows that it is [98 or 99% randomly presented] likely that the stinkwood flower will grow yellow stalks. So Joseph concludes that his stinkwood flower will grow yellow stalks. And he is right: the flower grows yellow stalks

After reading each scenario, participants were asked on a four-point Likert scale whether, for example, "Bob knows his spider plant contains aracnium" (1 = Definitely does not [know], 4 = Definitely does [know]). After answering this question, participants then answered a question aimed at assessing the sensitivity of the protagonist's belief:

"If Bob's spider plant hadn't contained the chemical aracnium, what would Bob have thought after checking with the book on spider plants?"

Participants then judged whether "Bob would [would not] have thought that his plant contains the chemical aracnium." If participants agreed that Bob would have thought his plant contains aranium regardless, this would mean participants thought that Bob's belief was not sensitive. If they thought that Bob would not have thought that, then this would indicate participants thought that Bob's belief was sensitive.

Predictions We predicted that knowledge judgments would vary based on the condition participants read. Namely, we predicted that participants would be more likely to attribute knowledge to protagonists that formed their belief on the basis of specific rather than general information. In addition, we predicted that participants would think that a protagonist's belief in the General condition is not sensitive (relative to the Specific condition) and this would in turn lead to reduced knowledge attributions.

Results and Discussion

To test our first hypothesis, we performed ordinal mixed-effects modeling. This model regressed knowledge judgments on condition (Reference = General condition) and included two group-level effects: (1) A group-level effect on vignette that allowed for heterogeneity in vignette intercept, (2) a group-level effect on Subject, which allowed for heterogeneity in both the slope and intercept of the condition effect on knowledge judgments. The model is specified below in `brms` syntax:

```
Model 1 <- Knowledge Response ~
Condition + (1|Item) + (1 +
Condition|Subject)
```

Bayesian analyses formulate model parameters as probability distributions wherein the posterior distribution

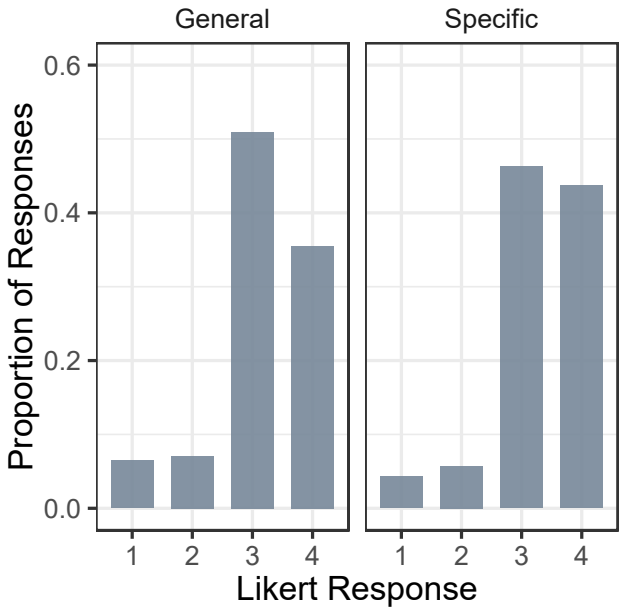


Figure 1: A histogram of the proportion of responses at a given Likert scale point (e.g., 1 = Definitely does not know; 4 = Definitely knows) in the General and Specific conditions in Experiment 1. The figure indicates that participants were less likely to attribute knowledge in the General condition than the Specific condition.

for a parameter θ is computed via the prior and the likelihood of θ. To model the joint probability distribution of participants' knowledge responses, we specified the following priors over the possible effects each parameter could have on the response variable:

Experiment 1 - Model 1 Priors:

$\beta_{Intercept[1]} \sim \mathcal{N}(-1.73, 1)$
$\beta_{Intercept[2]} \sim \mathcal{N}(-.61, 1)$
$\beta_{Intercept[3]} \sim \mathcal{N}(.84, 1)$
$\beta_{Condition} \sim \mathcal{N}(0, .5)$
$\Omega_k \sim LKJ(1)$ where Ω_k is a correlation matrix of group-level parameters
Group-level parameters were distributed as $\mathcal{N}(1, 1)$

We predicted that participants would be more likely to attribute knowledge in the Specific condition than the General condition. This is what we observed, b = 0.72, 95% CI [0.38, 1.05]. Figure 1 indicates that in the General condition, the third Likert scale point [Perhaps knows] was the most probable response. In contrast, in the Specific condition both 3 [Perhaps knows] and 4 [Definitely knows] were probable responses, indicating that participants were more likely to attribute knowledge to a protagonist when the means by which the protagonist formed their belief was specific rather than general, as had been previously found (Friedman & Turri, 2014).

What explains people's tendency to differentially ascribe

knowledge when the probabilities in both situations are perceived as being similar? Our hypothesis is that this effect may be at least partially explained by sensitivity. That is, we predicted that people would think that an agent's beliefs in the Specific condition would be more sensitive than in the General condition. To test this prediction, we first performed logistic mixed-effects modeling. This model regressed participants' judgments of belief sensitivity (i.e., 1 = Protagonist would still believe, 0 = Protagonist would not still believe) on condition.

```
Model 2 <- Counterfactual Response ~ 0 +
intercept + Condition + (1|Item) +
(1 + Condition|Subject)
```

Experiment 1 - Model 2 Priors:

$\beta_0 \sim \mathcal{N}(.75, .25)$
$\beta_{Condition} \sim \mathcal{N}(0, .5)$
$\Omega_k \sim LKJ(1)$
Group-level parameters were distributed as $\mathcal{N}(1, 1)$

Here too, we observed the predicted effect of condition on participants assessments of belief sensitivity, $b = -0.60$, 95% CI [−0.94, −0.25] (see Figure 2).

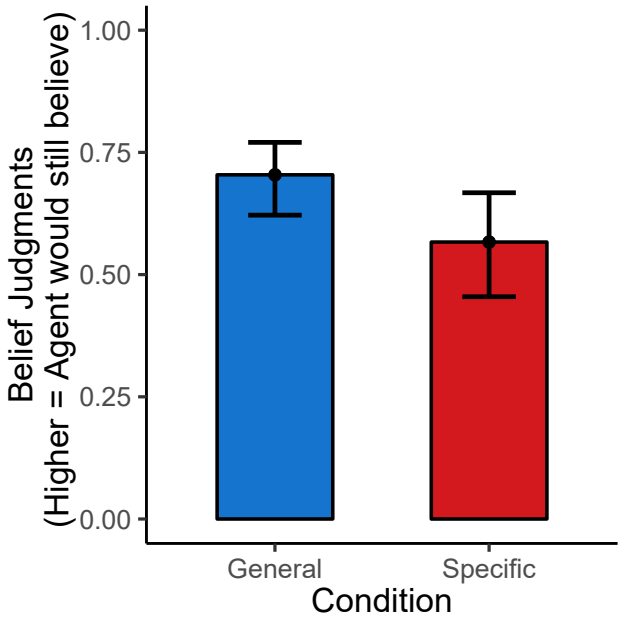


Figure 2: A marginal effects plot of belief sensitivity in the General and Specific conditions in Experiment 1. Error bars represent 95% Credible Intervals (see Cummings, 2005 for interpreting within-subjects 95% CIs).

Our hypothesis was that people tend to differentially attribute knowledge in General and Specific conditions because of the sensitivity of the protagonists' beliefs in these two situations. We observed this effect. We followed up on this observation by performing mixed-effects mediation modeling to test whether sensitivity mediated the relationship between condition and knowledge attributions, as suggested by the partial correlations.

```
Knowledge Model =
bf(Knowledge ~ Condition +
Counterfactual +
(1 + Condition|Subject))

Counterfactual Model =
bf(Counterfactual ~ Condition +
(1 + Condition|Subject))

Mediation Model =
brm(Knowledge Model +
Counterfactual Model)
```

Mediation Model Priors:

All $\beta \sim \mathcal{N}(0, .5)$
$\Omega_k \sim LKJ(1)$
Intercept and group-level parameters were distributed as $t(3,0,10)$

Consistent with our hypothesis, we observed the predicted mediation, *ab* path = .37, 95% CI [.08, .81].

Experiment 2

Experiment 1 provides preliminary evidence for the hypothesis that sensitivity can account for people being less likely to attribute knowledge in general versus specific cases. However, one possibility is that people think the agents in these two situations are not equally justified in drawing their conclusions and, further, that belief sensitivity is correlated with attributions of justification. Although it is often assumed that people are equally justified in believing that *P* in both general and specific cases, this assumption has not been empirically tested. Consequently, we sought to rule out this alternative explanation of the results of Experiment 1 because justification is likely one of the most strongly predictive factors of knowledge attributions. To address this possibility, Experiment 2 directly replicated Experiment 1 but also included a measure of justification.

Participants We powered our study to detect a Cohen's d of .15 for a two-condition within-subjects design with 80% power. We anticipated the effect of counterfactual responses on knowledge attributions would be smaller after accounting for differences in justification between general and specific cases. A total of 344 participants were recruited through Amazon's Mechanical Turk (46% women, M_{age} = 36 years old). Participants were paid $0.50 for participating in the study. After excluding participants who missed questions checking their attention, 288 participants remained in our sample. Our sample size and exclusion criteria were determined a priori and were in accordance with our study preregistration.

Procedure and Predictions The procedure of Experiment 2 was identical to that of Experiment 1. However, in Experiment 2, participants answered additional questions

testing whether they thought the protagonist's beliefs were justified. For example, after reading each vignette participants also judged whether Bob was justified in concluding that his plant contains aracnium. Complete materials can be found in Table S4 in the SOM.

As in Experiment 1, we predicted that knowledge and sensitivity judgments would differ in the General and Specific conditions, and that sensitivity would mediate the relationship between condition and knowledge attributions. However, Experiment 2 also allowed us to test whether participants thought that (1) the protagonists in the Specific condition were more justified in drawing their conclusions and (2) rule out the possibility that justification alone rather than belief sensitivity accounts for the differential attribution of knowledge attributions in the General and Specific conditions.

Results and Discussion

As in Experiment 1, we first examined how condition (General vs Specific) affected knowledge and sensitivity judgments. We performed the same analytic procedure as in Experiment 1. These analyses again revealed that both knowledge and sensitivity were predicted by whether the case participants were reading was general or specific, Knowledge: b = 1.10, 95% CI [0.81, 1.41) and Counterfactual: b = –.52, 95% CI [–0.86, –.18) (see Figures 3 and 4), along with the mediation of knowledge judgments by sensitivity *ab* path = .27, 95% CI [.05, .60). We then tested whether participants differentially attributed justification to the protagonists in the General and Specific conditions. We found that, indeed, participants judged that the protagonist in the Specific condition was more justified in drawing their conclusion than in the General condition, b = .86, 95% CI [0.51, 1.20]. Does justification, rather than sensitivity, account for the difference in knowledge attributions in general and specific cases? To examine this possibility, we performed ordinal mixed-effects modeling regressing knowledge attributions on justification and counterfactual judgments. Justification responses were treated as monotonic effects, following the recommendations of Bürkner and Charpentier (2018):

```
Model 3 <- Knowledge Response ~
Counterfactual + mo(Justification) +
(1|Item) + (1|Subject)
```

Experiment 2 - Model 3 Priors

$\beta_{Intercept[1]} \sim \mathcal{N}(-2,1)$
$\beta_{Intercept[2]} \sim \mathcal{N}(-1.5,1)$
$\beta_{Intercept[3]} \sim \mathcal{N}(4,1)$
$\beta_{Counterfactual} \sim \mathcal{N}(0,.5)$
$\beta_{Justification} \sim \mathcal{N}(2,4)$
Group-level parameters were distributed as $\mathcal{N}(1,1)$

This analysis revealed that sensitivity accounted for unique variance in knowledge attributions over and above attributions of justification, b = –0.51, 95% CI [–0.83, –0.19]. Thus, Experiment 2 provides evidence for two further conclusions: First, knowledge attributions likely

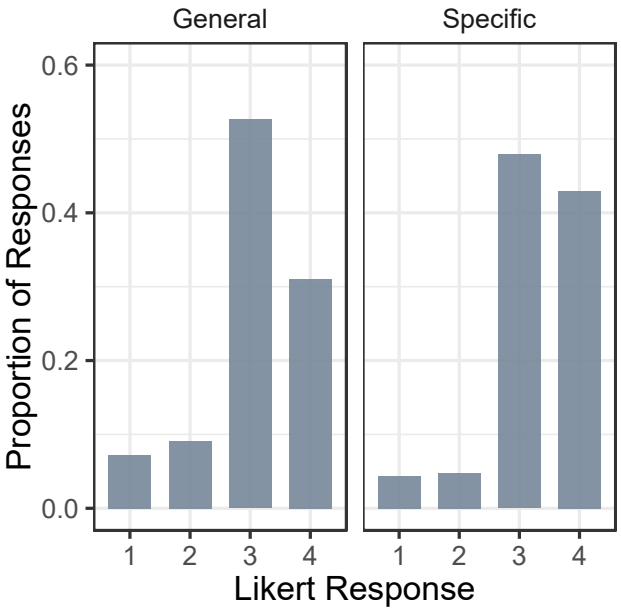


Figure 3: A histogram of the proportion of responses at a given Likert scale point (e.g., 1 = Definitely does not know; 4 = Definitely knows) in the General and Specific conditions in Experiment 2. The figure indicates that participants were less likely to attribute knowledge in the General condition.

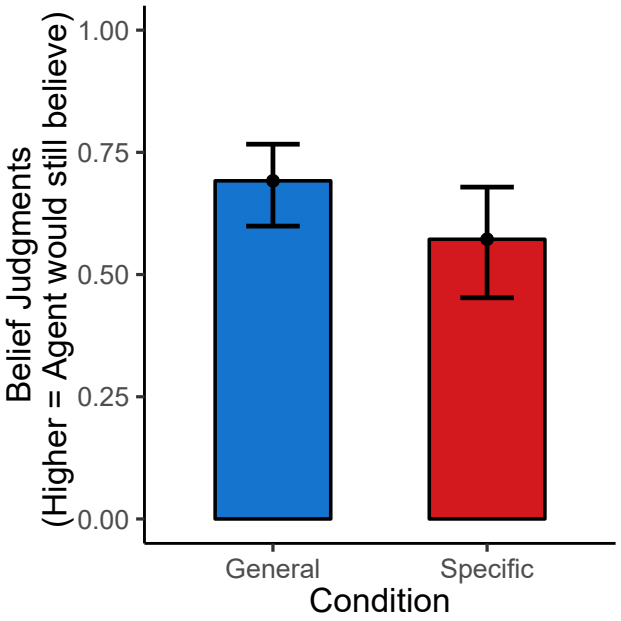


Figure 4: A marginal effects plot of belief sensitivity in the General and Specific conditions in Experiment 2. Error bars represent 95% Credible Intervals.

differ in general and specific cases because of a difference in the perceived justification of agents in these situations. Second, and consistent with our prediction, people judge that beliefs formed on the basis of general evidence are insensitive (relatively) compared to beliefs formed on the basis of specific evidence, an effect that holds over and above the effect of justification.

Discussion

In psychology and the law, researchers have observed that jurors are less likely to trust general as opposed to specific evidence (Wells, 1992). However, this effect appears to extend beyond the courtroom: We found that knowledge attributions were similarly affected by manipulating whether evidence was general or specific, replicating earlier work that third-person knowledge attributions increase when the evidence available to the protagonist is specific as opposed to general. It is notable that this effect was found in cases that have little to do with the law or with witnesses, thus suggesting that the Wells effect is an instance of a more general phenomenon that goes beyond cases of witnesses or eyewitness memory specifically.

What explains people's tendency to distinguish general and specific evidence, given that the probabilities are fixed? We explored the hypothesis that general and specific evidence differ along more lines than their probabilities: namely, we examined whether sensitivity (a counterfactual condition) can account for the tendency to attribute knowledge in specific compared to general cases. Experiment 1 provided evidence that what explains the difference in knowledge judgments (across general and specific cases) is the perceived sensitivity of the belief at issue. In Experiment 2, we considered the alternative hypothesis that justification rather than sensitivity accounts for the difference in people's knowledge attributions in general and specific cases. We found that assessments of sensitivity account for variance in knowledge attributions over and above justification, although justification also differed by condition.

These experiments constitute the first empirical investigation demonstrating a link between knowledge attributions and sensitivity among non-philosophers. Further, this is the first piece of experimental evidence which suggests that a well-known effect in legal decision making—the Wells effect—can be understood in terms of epistemic theories that highlight the link between knowledge attributions and counterfactuals. Still, questions remain about the ability of the sensitivity hypothesis to account for other core epistemic phenomena. Indeed, a number of philosophers, for example, Williamson (2000), Blome-Tillmann (2015), and Hawthorne (2003), have pointed out that the sensitivity hypothesis cannot fully explain a range of intuitions regarding knowledge attributions. Consequently, subsequent experimental work should investigate these proposals to establish the scope and limits of the sensitivity hypothesis.

References

Alicke, M., Buckingham, J., Zell, E., & Davis, T. (2008). Culpable control and counterfactual reasoning in the psychology of blame. *Personality and Social Psychology Bulletin*, *34*, 1371–1381.

Barbey, A., & Sloman, S. (2007). Base-rate respect: From ecological rationality to dual processes. *Behavioral and Brain Sciences*, *30*, 241–254.

Black, T., & Murphy, P. (2007). In Defense of Sensitivity. *Synthese*, *154*, 53–71.

Blome-Tillmann, M. (2015). Sensitivity, Causality, and Statistical Evidence in Courts of Law. *Thought: A Journal of Philosophy*, *4*, 102–112.

Brassen, S., Gamer, M., Peters, J., Gluth, S., & Büchel, C. (2012). Don't look back in anger! Responsiveness to missed chances in successful and nonsuccessful aging. *Science*, *336*, 612–614.

Bürkner, P. (2018). Advanced Bayesian Multilevel Modeling with the R Package brms. *The R Journal*, *10*, 395–411.

Bürkner, P., & Charpentier, E. (2018). Monotonic Effects: A Principled Approach for Including Ordinal Predictors in Regression Models. *PsyArXiv*.

Byrne, R. (1997). Cognitive processes in counterfactual thinking about what might have been. *The psychology of learning and motivation: Advances in research and theory*, *37*, 105–154.

Coricelli, G., & Rushtichini, A. (2010). Counterfactual thinking and emotions: regret and envy learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*, 241–247.

Cumming, G., & Finch, S. (2005). Inference by eye: confidence intervals and how to read pictures of data. *American Psychologist*, *60*, 170.

Davis, C., Lehman, D., Wortman, C., Silver, R., & Thompson, S. (1995). The undoing of traumatic life events. *Personality and Social Psychology*, *21*, 109–124.

DeRose, K. (1999). Contextualism: An Explanation and Defense. *The Blackwell Guide to Epistemology*, 185–203.

Dretske, F. (1981). The Pragmatic Dimension of Knowledge. *Philosophical Studies: An International Journal for Philsophy in the Analytic Tradition*, *40*, 363 – 378.

Enoch, D., & Fisher, T. (2015). Sense and "Sensitivity" Epistemic and Instrumental Approaches to Statistical Evidence. *Stanford Law Review*, *67*, 557.

Enoch, D., Spectre, L., & Fisher, T. (2012). Statistical Evidence, Sensitivity, and the Legal Value of Knowledge. *Philosphy & Public Affair*, *40*, 197 – 224.

Epstude, K., & Roese, N. (2008). The functional theory of counterfactual thinking. *Personality and Social Psychology Review*, *12*, 168 – 192.

Friedman, O., & Turri, J. (2014). Is Probabilistic Evidene a Source of Knowledge? *Cognitive Science*, *39*, 1062 – 1080.

Gelman, A., Lee, D., & Guo, J. (2015). Stan: a probabilistic programming language for Bayesian inference and optimization. *Journal of Educational and Behavioral Statistics*, *40*, 530–543.

Halpern, J., & Hitchcock, C. (2015). Graded Causation and Defaults. *British Journal for the Philosophy of Science*, *66*, 413–457.

Hawthorne, J. (2003). *Knowledge and Lotteries*. New York, NY: Oxford University Press.

Ichikawa, J. (2018). Quantifiers, Knowledge, and Counterfactuals. *Philosophy and Phenomenological Research*, *82*, 287 – 313.

Kruschke, J., & Liddell, T. (2018). The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*, *25*, 178–206.

Malle, B., Guglielmo, S., & Monroe, A. (2016). A Theory of Blame. *Psychological Inquiry*, *25*, 1 – 40.

Markman, K., McMullen, M., & Elizaga, R. (2008). Counterfactual thinking, persistence, and performance: A test of the Reflection and Evaluation Model. *Journal of Experimental Social Psychology*, *44*, 421 – 428.

McElreath, R. (2016). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Boca Raton, FL: CRC Press.

Miller, A., Markman, K., Wagner, M., & Hunt, A. N. (2013). Mental simulation and sexual prejudice reduction: the debiasing role of counterfactual thinking. *Journal of Applied Social Psychology*, *43*, 190–194.

Niedermeier, K., Horowitz, I., & Kerr, N. (1999). Informing Jurors of Their Nullification Power: A Route to a Just Verdict or Judicial Chaos? *Law and Human Behavior*, *23*, 331–351.

Nozick, R. (1981). *Philosophical Explanations*. Cambridge, MA: Harvard University Press.

Pieters, R., & Zeelenberg, M. (2005). On bad decisions and deciding badly: when intention-behavior inconsistency is regrettable. *Organizational Behavior and Human Decision Processes*, *97*, 18–30.

Roese, N., & Olson, J. (1997). Counterfactual Thinking: The Intersection of Affect and Function. *Advances in Experimental Social Psychology*, *29*, 1–59.

Roese, N., & Olson, J. (2012). Learning from yesterday's mistakes to fix tomorrow's problems: when functional counterfactual thinking and psychological distance collide. *European Jouranal of Social Psychology*, *42*, 383–390.

Roush, S. (2005). *Tracking Truth: Knowledge, Evidence, and Science*. New York, NY: Oxford University Press.

Smallman, R., & Roese, N. (2009). Counterfactual thinking facilitates behavioral intentions. *Journal of Experimental Social Psychology*, *45*, 845–852.

Sykes, D., & Johnson, J. (1999). Probabilistic Evidence Versus the Representation of an Event: The Curious Case of Mrs. Prob's Dog. *Basic and Applied Social Psychology*, *21*, 199–212.

Tobia, M., Guo, R., Schwarze, U., Boehmer, W., G(ä)sher, J., Finckh, B., ... Sommer, T. (2014). Neural systems for choice and valuation with counterfactual learning signals. *Neuroimage*, *89*, 57–69.

Wells, G. (1992). Naked Statistical Evidence of Liability: Is Subjective Probability Enough? *Journal of Personality and Social Psychology*, *62*, 739–752.

Williamson, T. (2000). *Knowledge and its Limits*. New York, NY: Oxford University Press.

Egocentric Tendencies in Theory of Mind Reasoning: An Empirical and Computational Analysis

Jan Pöppel (jpoeppel@techfak.uni-bielefeld.de) and **Stefan Kopp (skopp@techfak.uni-bielefeld.de)**
Social Cognitive Systems, CITEC, Bielefeld University
Bielefeld, Germany

Abstract

Humans develop an ability for Theory of Mind (ToM) by the age of six, which enables them to infer another agent's mental state and to differentiate it from one's own. Much evidence suggests that humans can do this in a presumably optimal way and, correspondingly, a Bayesian Theory of Mind (BToM) framework has been shown to match human inferences and attributions. Mostly, this has been investigated with specific, explicit mentalizing tasks. However, other research has shown that humans often deviate from optimal reasoning in various ways. We investigate whether typical BToM models really capture human ToM reasoning in tasks that solicit more intuitive reasoning. We present results of an empirical study where humans deviate from Bayesian optimal reasoning in a ToM task but instead exhibit egocentric tendencies. We also discuss how computational models can better account for such sub-optimal processing.

Keywords: Theory of Mind; Bayesian Modeling; Egocentric Tendencies; Bounded rationality

Introduction

An important ability of humans is to infer and reason about ones own as well as other's mental states such as intentions, (potentially false) beliefs, or emotions (Wellman & Liu, 2004). While the exact development of this so-called Theory of Mind (ToM) (Premack & Woodruff, 1978) is still not clear, there is a consensus that we acquire full ToM abilities around the age of six (Wellman & Liu, 2004). This allows us to make sense of our social environment, to learn more from the actions around us (Jara-Ettinger, Baker, & Tenenbaum, 2012) and to better understand or even manipulate others in cooperative or competitive interactions (Heyes & Frith, 2014).

Because of its importance for social interaction, there is a great interest in endowing artificial systems with similar capabilities. Recently, the most prominent approach has been the Bayesian Theory of Mind (BToM) framework (Baker, Saxe, & Tenenbaum, 2009). Building upon the rational agent assumption (Dennett, 1989) and inverse planning, the BToM framework constructs probabilistic generative models that relate hidden mental states to observable actions. These models can then be inverted to infer mental states from behavior, while accounting for inherent uncertainty. This framework has been shown to make inferences that correlate well with those made by humans in a wide range of different tasks, such as the inference of desires and beliefs (Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017) or preferences (Jern, Lucas, & Kemp, 2017). It is also in line with the Bayesian Brain hypothesis positing that humans incorporate information similar to optimal observers (Knill & Pouget, 2004).

At the same time, humans do not always behave like optimal observers or reasoners. Instead, they exhibit a range of fallacies leading to systematic errors in different types of inference (Haselton, Nettle, & Murray, 2015). This has also been argued to hold for social interaction. For example, Keysar (2007) showed that adults fail to adjust correctly for different perspectives in communication tasks. This trait is often referred to as *egocentric tendency* and refers to the tendency to impute one's own mental perspective on others (Nickerson, 1999). Keysar, Lin, and Barr (2003) present an experiment in which even adults fail a false-belief test, showing that an egocentric tendency is not always effectively suppressed. In other words, we do not appear to always use our ToM capabilities to the fullest extent (cf. (De Weerd, Verbrugge, & Verheij, 2013)). This is often attributed to limited mental resources, such as working memory and processing time. Vul, Goodman, Griffiths, and Tenenbaum (2014) argue that many biases are actually optimal when seen as the result of the number of samples for inference being limited.

It is unclear how those limitations affect ToM reasoning in humans. While the BToM framework has been shown to correlate well with humans' explicit ToM reasoning, it has not been evaluated with regard to humans' intuitive or implicit inferences, i.e. when sophisticated ToM reasoning is not explicitly evoked. Recently, Nakahashi and Yamada (2018) showed that a full inverse planning approach based on the BToM framework overestimates the rationality of humans and that modified inference achieves better correlations with human judgments. We are interested in whether, in an intuitive setting, humans employ different kinds of ToM models as a function of, e.g., computational costs, available resources, or current task demands. We have argued elsewhere that employing different kinds of ToM models for "satisficing mentalizing" can be beneficial for artificial systems, where full Bayesian models often suffer from intractabilities (Pöppel & Kopp, 2018). Here, we study whether humans may also employ different simpler, non-optimal models depending on the given circumstances and, specifically, whether they may fail to realize or account for differences between one's own and another one's mental states. We thus focus on the extent to which humans employ mentalizing in a settings that is more implicit than those used in previous BToM research.

In the remainder of this paper, we present empirical evidence suggesting that humans exhibit different degrees of egocentric tendencies in a simple ToM reasoning task, thus deviating from rational optimal observers usually assumed in previous BToM models. The next section first describes the scenario we are looking at. Then, we present an empirical study we have carried out in this scenario to investigate in-

tuitive human ToM reasoning. After this, we present different computational ToM models, partially based on the BToM framework, and report their correlations with our data.

Scenario

The scenario we chose for our empirical study is the inference of an agent's desire in a navigation task within a 2D maze. The maze has four exits, each of which leading to a distinct destination (denoted Red, Yellow, Blue or Orange). The agent has to find the exit that leads to one specific destination, which we consider to be the agent's desire. In previous work we already gathered behavioral data in the form of trajectories of human participants solving this navigation task in different mazes with differing amounts of information available (Pöppel & Kopp, 2018). Here, we consider the task of an additional observer, who watches the agent move around in the maze and has to infer the agent's desire – a perceptual and cognitive task humans solve frequently in everyday life. According to the ToM scale by Wellman and Liu (2004), this kind of inference is also among the first ones to be mastered by children.

In order to create a need for differentiating between the mental perspectives of the agent and the observer, we employ two conditions: in the first condition, participants acting as agents had full knowledge of the maze, the locations of all exits, and the destinations behind them. That is, they could take an optimal path in order to reach their desired destination. In the second condition, the acting participants knew about the locations of the exits in the maze, but had to discover the corresponding destinations themselves by establishing a line of sight with the exit. Thus participants had to search for the specific exit (one out of four) that leads to their desired destination, resulting in an exploration behavior. This scenario is similar to earlier work on BToM, e.g. (Baker et al., 2017), in that it involves navigating a simple grid-world to achieve a desired outcome with potential uncertainty about the true location of that outcome.

Figure 1 shows an example of the different stimuli that the acting participants received in the two conditions. In the bottom example belonging to the second condition, the exits are marked but covered. Note that in this situation the agent has moved to a position, where it could see the exit thus revealing its corresponding destination (Blue). In the present study, we use recordings of the online navigation behavior in these two conditions and let human participants play the role of the observer. In particular, their task is to identify the desired destination of the observed agent at different points on the recorded trajectory.

Empirical Study: ToM Reasoning in Humans

Humans employ their ToM capabilities rarely to their fullest extent. However, it is still unclear what factors, apart from cognitive load, may influence the extent to which a person employs her ToM capabilities. Previous research has shown that explicit asking for likelihood ratings of all alternatives yields responses predicted by the BToM framework. However, we conjecture that this experimental design inherently evokes explicit reasoning about mental states in the participants, including the full consideration and comparison of all alternatives. This evocation may be part of the reason for the discrepancy between very good fittings in BToM research and findings of suboptimal behavior in other research. In contrast to previous research, we therefore deliberately chose not to ask for likelihood ratings for all possible desires, but instead ask for *soft forced-choice* responses in order to test for a more intuitive and natural ToM reasoning. We call it *soft* because we gave participants the additional option "I do not know".

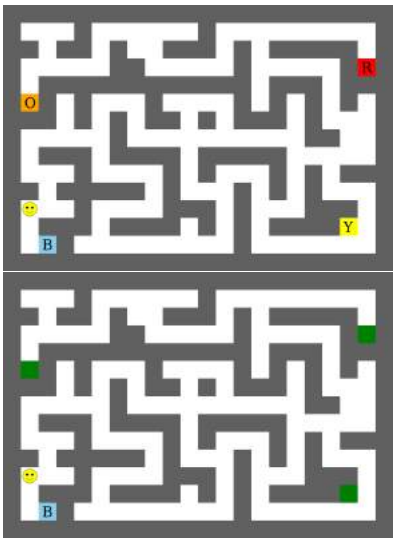


Figure 1: View of the navigating agents. Top: full knowledge about exits and destinations; Bottom: exits only reveal their destination once a line of sight is established.

We also included a second group of participants who were additionally prompted to self-assess their belief about the observed agent's knowledge. We included this group to test the effect of putting an agent's belief into focus of (more explicit) ToM reasoning, thus testing if different task demands influence the employed ToM models.

Participants We recruited two distinct groups of participants (first group 120; second group 65) each via an online platform called "figure-eight" (formerly crowdflower). All participants had a "contributor level three", which is advertised as "Highest Quality: Smallest group of most experienced, highest accuracy contributors". After completion of the study, participants were reimbursed with $0.20 via the figure-eight system.

Stimuli For each of the two conditions mentioned above, we chose two typical trajectory recordings in two different mazes. The four trajectories are shown in figure 2. Participants could see the maze and, importantly, all destinations behind the different exits. That is, they always had full knowl-

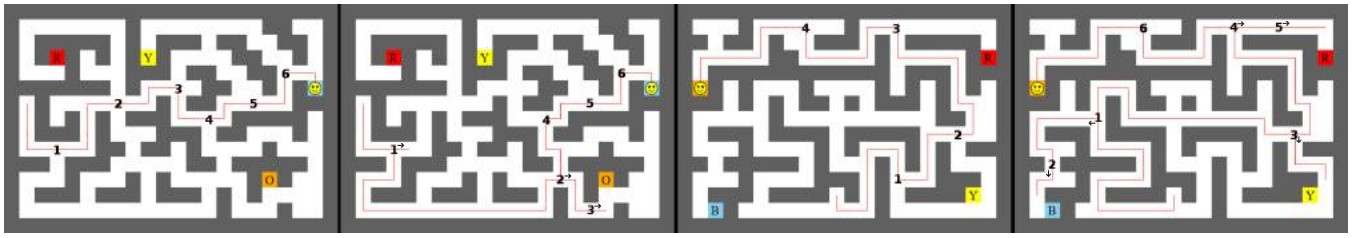


Figure 2: The recordings 1 through 4 used as stimuli, one from each condition (destinations known/unknown to the agent) in two different mazes. Numbers indicate the location of query points at which participants had to give their responses. Small arrows indicate the agent's next action after the query point when ambiguous.

edge about the destinations, while the agents they observed might or might not have had the same knowledge. Participants would see the agent navigating the maze according to the recordings, leaving behind a red trail as in figure 2, so that participants always knew where the agent moved. We chose to replay these recordings with fixed time intervals between two steps in order to remove potential noise within the trajectories (while also eliminating information about speed or possible hesitations of the agents).

Procedure All participants received the same initial instruction that they were going to watch the recordings of four different human players navigating a maze. They were informed that the maze had four exits, each leading to a different destination, and that each player had her own specific destination she had to reach as quickly as possible. We further explained that the four players were part of two different conditions. In one condition, they had full knowledge about the destinations behind the exits, while in the other they had to first discover which exit led to which destination. In order to make this clear, we provided participants with the images in figure 1 alongside the instructions. The instructions read: "*Now you will watch the agents follow their trajectories while you will be able to see which exit corresponds to which destination. At certain points in time you will be asked to tell the agent's desired destination (R,B,Y,O). You may also say that you do not know*." The query points are those shown in figure 2. We deliberately asked for the agent's *desired destination* instead of an exit to focus on the agent's desire instead of the exit locations close to the agent. The second group of participants were further instructed about the additional question regarding the agent's knowledge, which read "*Additionally, you will be asked to specify if you think the agent knows which exit leads to which destination*.".

Upon confirming these instructions, participants got to see the first maze as in figure 2 (without the query point numbers) with the agent at the beginning of its trajectory. After hitting a *Start* button the agent started to move leaving behind the red trail. The playback stopped at each QP and participants were asked to choose one out of the four possible goal destinations, or to signal that they cannot tell otherwise, which we will refer to as *Uncertain* (U) from here on. In order to avoid misinterpretations (such as having asked for current target location only), we instructed participants with: "*Please specify which destination you think the agent wants to reach after leaving the maze*." The second group of participants received an additional question before identifying the agent's destination: "*Do you think the agent you are watching currently knows which exit leads to which destination?*" Participants could respond with either "Yes" or "No". Once the agent reached its destination, participants could proceed to the next recording. In total, each of the 185 participant had to make 22 judgements (taking less than 400s on average for the first group and less than 485s for the second group).

For the first group, we counter-balanced the order in which participants saw the different recordings/mazes. We used a Fisher's exact test on the response frequencies in order to test whether or not the order in which the stimuli was presented had any effect on participants' responses. The test revealed no significant effects of the stimuli ordering for all but one responses (recording 1, QP 6). We thus concluded that the order of presentation of recorded trajectories/mazes did not influence participant's responses. We thus collapse the results of participants in the first group for the analyses in the remainder of this paper. Furthermore, we decided to use only one ordering for the second group in order to simplify the design.

For analyses, we excluded all participant's responses for a particular recording if participants always picked the same destination and if this destination was not the correct one within one recording. We further excluded responses if participants chose to predict a destination after the agent already turned away from it in recordings 2 and 4. We assume that these participants did not really pay attention to the actual trajectories as these are obvious errors. After excluding such participants, we had 110 participants in group 1 and 57 participants in group 2 remaining.

Results Figures 3 and 4 show normalized response frequencies for several interesting query points in the two groups. Note, however, that the reported tendencies also hold for the other recordings and query points.

For the first group of participants who only had to identify the likely destination of the agent, we find a strong bias

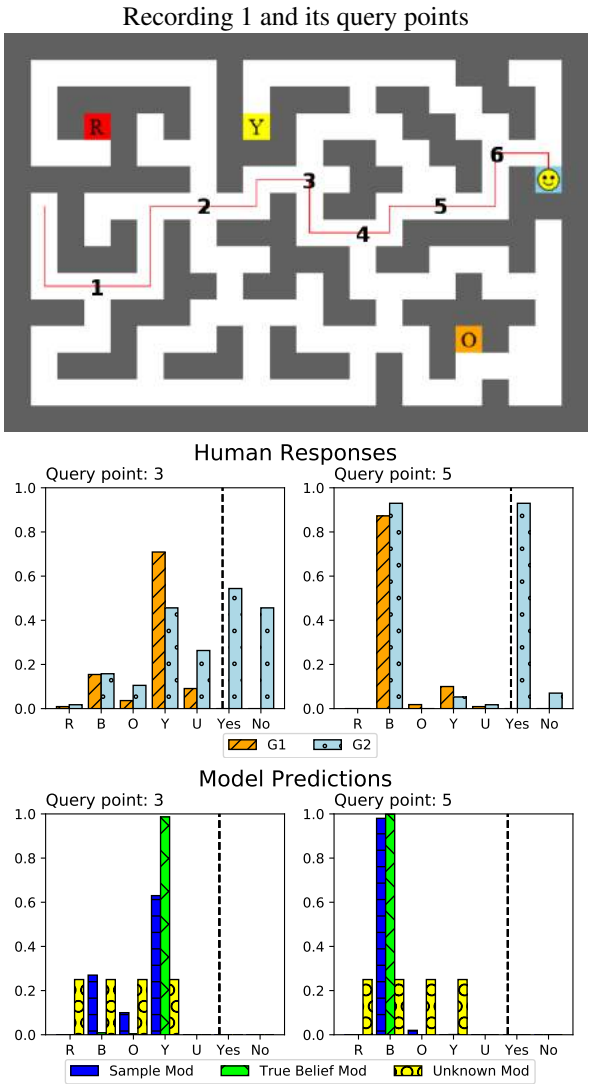


Figure 3: Relative response frequencies by the participants in both groups for Recording 1, query points 3 and 5 and the corresponding predictions made by the models using the likelihood modification

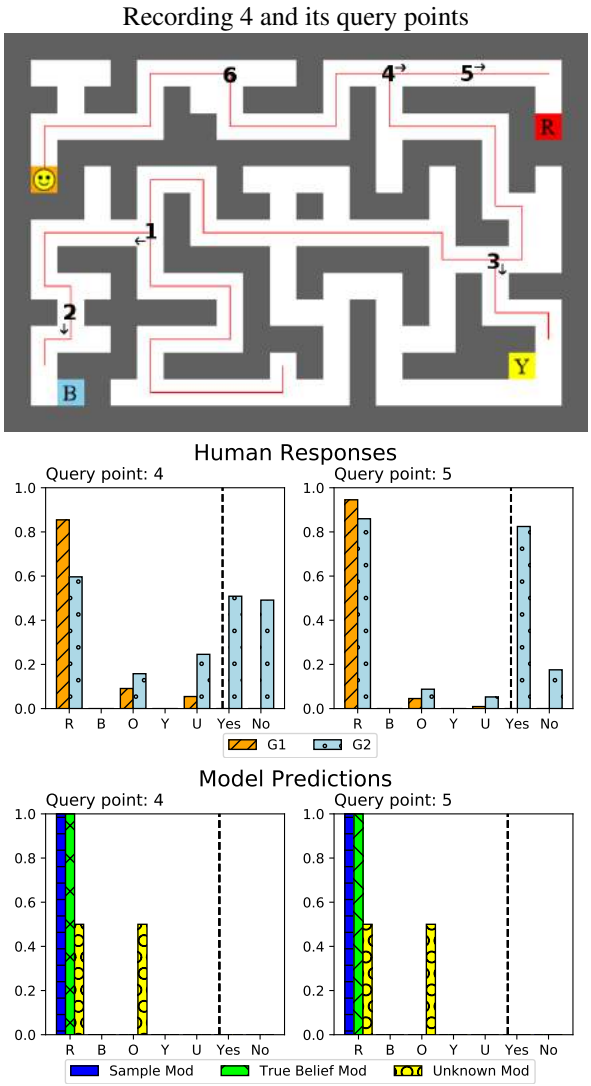


Figure 4: Relative response frequencies by the participants in both groups for Recording 4, query points 4 and 5 and the corresponding predictions made by the models using the likelihood modification.

towards assuming that the agent is seeking the destination behind the closest exit. This also holds true for points at which an agent's behavior was optimally directed to multiple exits (cf. Yellow responses for QP 3 in figures 3). We also find that participants ignore that the agent may have a knowledge state different from their own. For example, in the Red responses at QPs 4 and 5 in figure 4, they had already seen the agent turning away from two exits. Thus they should assume that the agent does not know which exit leads to which destination, even if they see the agent moving towards Red. They thus show an egocentric tendency in their reasoning.

When looking at the second group, i.e. participants that were first asked about the knowledge state of the agent before trying to identify the agent's desire, we find significantly different desire response distributions for 12 of the 22 QPs ($p \ll 0.05$ according to Fisher's exact test). These differences manifest themselves primarily in a difference of the percentage of Uncertain responses, which is significantly higher for the second group (14.9% vs 4.7%, $t = 3.86$ $p < 0.001$), also visible at QP 3 in figure 3 and QP 4 in figure 4. We further find interesting results regarding the preceding question about their belief of the agent's knowledge state: "No" responses (i.e. they believe the agent does not know) increase after an agent turned away from an exit it saw, as expected (e.g. in recording 4 "No" responses are only around 20% at QP 2, but increase to around 45% at QP 3). However, the percentage for "No" never exceeds 55% and quickly decreases again as the agent moves towards any particular exit, as seen for QP 5 in figure 4.

Computational Modeling

These results indicate that our participants in groups 1 and 2 performed their ToM reasoning differently, but neither group appears to make use of their ToM capabilities to their fullest extend. In fact, participants may even employ different strategies at different parts of the recordings. In this section we explore different models for the desire ToM task and present how they correlate with our empirical data.

BToM models The first two models we are considering, which are taken from previous work (Pöppel & Kopp, 2018), follow the general BToM framework and were designed to correspond to the mental states induced by the two conditions described above.

The *True Belief* model assumes that the agent has full knowledge regarding which exit leads to which destination:

$$P(a_{t+1}|\boldsymbol{a}_t) = \sum_{d \in D} P(a_{t+1}|d, b_d^*) P(d|\boldsymbol{a}_t) \tag{1}$$

The *Unknown* model assumes that the agent does not initially know which exit leads to which destination, which is why it needs to consider all possible combinations:

$$P(a_{t+1}|\boldsymbol{a}_t) = \sum_{\substack{d \in D \\ b_d \in B_d}} P(a_{t+1}|d, b_d) P(b_d|\boldsymbol{a}_t) P(d|\boldsymbol{a}_t) \tag{2}$$

With D being the set of desirable destinations and B_d the set of beliefs about which exit leads to which destination. b_d^* is the true belief, i.e. it maps exits to destinations correctly. $\boldsymbol{a}_t = a_1, .., a_t$ is the sequence of past actions, observed up to this point t.

The likelihood $P(a_{t+1}|d, b_d)$ is modeled following the commonly used Boltzmann noisy rationality:

$$P(a_{t+1}|d, b_d) = \frac{exp(\beta U(a_{t+1}, b_d, d))}{\sum_{a_i \in A} exp(\beta U(a_i, b_d, d))} \tag{3}$$

with β specifying the degree of rationality. Low values of β will allow more sub-optimal actions, while a larger β will result in the probability mass to be concentrated on the action with the highest utility $U(a_{t+1}, b_d, d)$ which in this simple scenario can be equated to the remaining distance to the exit leading to d according to belief b_d after executing the action a_{t+1}. The belief b_d is updated when the agent actually sees one of the exits by dismissing any beliefs which do not conform to the evidence.

Simple sampling model As a third model we introduce a model correlating to shallow processing with egocentric tendencies by implementing a very naive sampling approach: At the start of the recording, the model samples one destination from the prior $P(D)$. After observing each action, its likelihood $P(a|d, b_d^*)$ is computed using eq. 3. We keep this sample with the probability of the likelihood. Conversely, we draw a new sample with probability $1 - P(a|d, b_d^*)$ again from the prior $P(D)$, while ensuring not to pick a previously discarded destination. This way, the worse a sample can predict the observed actions the more likely it is to be replaced. Once all destinations have been discarded, we are considering all of them again, as we must have discarded the correct one along the way. The prior $P(G)$ is computed every time we need to draw a sample and depends on the remaining distance between the agent's current position p and the destination:

$$P(d) \propto exp(-\beta dist(p, d)) \tag{4}$$

For our results presented below, we fit β in the range of 0.1 to 3 at 0.1 intervals via a grid search to maximize correlations for each model separately.

Modifications As we are interested in what kind of models are required to model different ToM reasoning strategies employed by humans, we further tested the following modification to the likelihood function (eq. 3) in order to be able to better reflect the biases found in our data. While these modifications may improve the correlations in this case, we note that they may actually decrease correlation with human judgments that employ more thorough ToM reasoning.

To better reflect the bias for the closest exit found in the data, we changed the rationality constant β to a dynamic variable, which decreases with the distance to the exit, effectively dampening the likelihoods for exits that are further away and boosting optimal actions towards closer exits.

$$\beta \propto \alpha exp(-\gamma dist_m(p, d)) \tag{5}$$

where $dist_m$ is the current Manhattan distance between the agent's position p and the considered destination d. In this case α and γ are meta parameters that were fit to maximize correlation with a grid search between 2 and 4 at 0.1 intervals for α and between 0.025 and 0.75 at 0.05 intervals for γ for the results.

Model evaluation We compare our models with our participants' responses both on each recording separately, as well as over all responses. As has been done in previous BToM research (e.g. (Baker et al., 2017)) we considered the correlations between participants' average responses and the models' predicted distributions at each of the different query points. For this we stack the relative response frequencies for the four possible destinations for all QPs within a single recording, resulting in a vector of $4 \times 6 = 24$ elements (16 for recording 3 as there were only 4 QPs). Likewise we stack the destination distributions predicted by our models before computing the Pearson's r correlation. For the sampling model, we generated 100 independent responses and used the resulting normalized frequencies as the model's distribution. We then further stack the vectors for all recordings for the overall comparison, yielding a vector with 88 elements. We are deliberately evaluating in favor of our models in order to consider a best case scenario: All meta-parameters (β, α and γ)

have been fit to maximize the resulting correlation across all recordings. Furthermore, we spread all Uncertain responses across the other alternatives proportional to the model's distribution.

In order to test how well the models match the participants individually, we further had the models create actual predictions and compared these to the responses of each participant. We sampled 100 discrete responses from our models' predictions and computed how often these responses match the participants' responses at each of the different query points. We then averaged these number of matches over all query points for each recording and over all participants to get the average matching performance of our models. Again, in order to evaluate in favor of the different models, we count Uncertain responses as matches.

Tables 1 and 2 summarize the resulting correlations as well as the average number of matching responses (values in brackets) between our models and their modifications with the human responses of the first and second group respectively. Missing correlation values (–) are due zero variance in predictions of the Unknown model in those recordings. Note that Recording 3 contained only 4, instead of 6 query points.

Exemplary model outputs can also be found in figures 3 and 4, which shows the response distributions of the different (modified) models for the same QPs as the human responses.

The first thing to note is that the Unknown model, being the most rational with the least amount of biased assumptions, performs significantly worse than all others. This holds true for both the average correlation, as well as the number of matches. In Recordings 1 and 3 where we cannot compute the correlation due to zero variance, the Unknown model fails to make any predictions, always yielding a uniform distribution, which turns the Unknown model into a random model when comparing response matches. The slightly higher than chance performance of the Unknown model can mostly be attributed to the U responses. Furthermore, we find that the Sampling model correlates best with our human data, with a significant difference to the True Belief model. These results are also reflected by the average number of response matches. All models without the modification, except for the Unknown model correlated significantly more with participants in group 2 than in group 1. With regard to the modifications introduced by eq. 5: The True Belief model can improve its correlation significantly for both groups, while the Sampling model only improves significantly for the first group. Finally, it is noteworthy that the best meta-parameters for the Sampling model differ quite strongly between the two groups. (All significance claims achieved $p < 0.05$ on a t-test using the correlation coefficients after employing a Fisher transformation.)

Discussion and Conclusion

The results reported here suggest that humans can deviate quite strongly from optimal ToM reasoning. The rare use of the U(ncertain) response overall indicates that participants do not always consider the likelihood of all valid alternatives, but rather focus on single alternatives. In particular, they often fail to give U responses even after it became apparent that the agent is not aware of the location of the desired destination. In contrast, optimal reasoning would dictate the use of U responses whenever more than one destination is the most probable, or whenever multiple alternatives have a non-zero probability. Instead, participants show egocentric tendencies by ascribing their own map knowledge to the agent, and moreover a strong bias towards the closest exit as destination. This is also reflected in the decrease of "No" responses in the second group as soon as the agent moves towards any exit: even participants that briefly suppressed this tendency after having observed the agent moving away form a seen exit, tend to discard this evidence again at the next QPs. The results of the second group indicate that posing a question about the mental state of the agent before requesting the desire inference, increases the number of considered ToM alternatives slightly. Still, even participants of the second group that correctly realised that the agent's knowledge state differed from their own, often did not account for it properly when reasoning about the desire of uncertain agents. These findings support the hypothesis that humans may perform ToM reasoning differently. The task to give likelihood ratings for all alternatives (as e.g. in (Baker et al., 2017; Jern et al., 2017)) might evoke more controlled and complex ToM reasoning, suppressing cognitive biases and resulting in good correlations with optimal Bayesian models.

One might object that the observed bias towards the closest exit may stem from interpreting the instructions as "*where do you think the agent is currently going?*". However, the actual instruction was deliberately chosen to prevent this interpretation by stating "*Please specify which destination you think the agent wants to reach after leaving the maze*." While we cannot be certain about the actual interpretation by participants in the online study, we do believe that the biases are more likely to originate from inherent tendencies to use simpler, less demanding mentalizing strategies.

Looking at the correlations with different computational models, we find only comparatively weak correlations of the Unknown model with the empirical data, indicating that participants' responses are quite different from optimal Bayesian reasoning. Instead, the exhibited egocentric tendencies and biases are matched better by the True Belief and Sampling models. The better correlations of the Sampling model compared to the True Belief model can be attributed to the fact that the True Belief model compares all alternative destinations equally, while the Sampling model sticks with the first best guess, which conforms to a closeness bias, as long as it is not invalidated. When introducing likelihood modifications that shift the focus to the closer exits, the True Belief model starts to behave similarly. The lower difference between the correlations with the True Belief models and the Sampling models in Group 2, as compared to in Group 1, also indicates that priming participants with an explicit ToM-related question reduced these biases.

Table 1: Average correlations and number of response matches (in brackets) of models with ratings of Group 1.

Model	Recording 1	Recording 2	Recording 3	Recording 4	Overall
True Belief ($\beta = 0.3$)	0.85 (4.73)	0.95 (2.43)	0.68 (3.84)	0.85 (5.24)	0.85 (4.06)
True Belief Mod ($\alpha = 2.5$,$\gamma = 0.125$)	0.98 (4.81)	0.99 (2.82)	0.89 (4.02)	0.87 (5.23)	0.93 (4.22)
Unknown ($\beta = 1.9$)	– (4.52)	0.30 (2.12)	– (3.41)	0.62 (4.40)	0.40 (3.61)
Unkown Mod ($\alpha = 2.5$,$\gamma = 0.025$)	– (4.87)	0.30 (3.03)	– (4.22)	0.62 (4.95)	0.40 (4.27)
Sampling ($\beta = 1.9$)	0.94 (1.89)	0.96 (1.18)	0.82 (1.85)	0.99 (3.03)	0.94 (1.99)
Sampling Mod ($\alpha = 3.7$,$\gamma = 0.125$)	0.98 (1.94)	0.98 (1.18)	0.96 (1.73)	0.98 (2.95)	0.98 (1.95)

Table 2: Average correlations and number of response matches (in brackets) of models with ratings of Group 2.

Model	Recording 1	Recording 2	Recording 3	Recording 4	Overall
True Belief ($\beta = 0.3$)	0.95 (5.09)	0.96 (4.75)	0.76 (2.75)	0.93 (3.99)	0.91 (4.15)
True Belief Mod ($\alpha = 2.5$,$\gamma = 0.125$)	0.98 (4.95)	0.98 (4.73)	0.94 (2.84)	0.94 (4.21)	0.96 (4.18)
Unknown ($\beta = 1.7$)	– (4.57)	0.34 (4.50)	– (2.42)	0.72 (3.82)	0.45 (3.83)
Unkown Mod ($\alpha = 2.5$,$\gamma = 0.075$)	– (4.98)	0.34 (4.82)	– (3.14)	0.72 (4.41)	0.45 (4.34)
Sampling ($\beta = 0.7$)	0.97 (3.36)	0.97 (2.40)	0.89 (1.33)	0.98 (2.19)	0.96 (2.32)
Sampling Mod ($\alpha = 2.7$,$\gamma = 0.075$)	0.98 (3.23)	0.97 (2.44)	0.99 (1.28)	0.97 (2.32)	0.97 (2.32)

Overall, the actual ToM reasoning of humans appears to be more differentiated than assumed in the BToM literature. Mental reasoning is computationally expensive, especially when considering mental states of others. Unless explicitly triggered, humans appear not to perform a full-blown ToM reasoning but to resort to simpler heuristics instead. Artificial social systems can make use of these findings by adapting to different ToM models employed by their users and assisting when they might overlook important information.

References

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, *1*, 0064.

Baker, C. L., Saxe, R. R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329–349. doi: 10.1016/j.cognition.2009.07.005

Dennett, D. C. (1989). *The intentional stance*. MIT press.

De Weerd, H., Verbrugge, R., & Verheij, B. (2013). How much does it help to know what she knows you know? an agent-based simulation study. *Artificial Intelligence*, *199*, 67–92.

Haselton, M. G., Nettle, D., & Murray, D. R. (2015). The evolution of cognitive bias. *The handbook of evolutionary psychology*, 1–20.

Heyes, C. M., & Frith, C. D. (2014). The cultural evolution of mind reading. *Science*, *344*(6190), 1243091.

Jara-Ettinger, J., Baker, C., & Tenenbaum, J. (2012). Learning what is where from social observations. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 34).

Jern, A., Lucas, C. G., & Kemp, C. (2017). People learn other peoples preferences through inverse decision-making. *Cognition*, *168*, 46 - 64. doi: https://doi.org/10.1016/j.cognition.2017.06.017

Keysar, B. (2007). *Communication and miscommunication: The role of egocentric processes.* Walter de Gruyter.

Keysar, B., Lin, S., & Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition*, *89*(1), 25–41.

Knill, D. C., & Pouget, A. (2004). The bayesian brain: the role of uncertainty in neural coding and computation. *TRENDS in Neurosciences*, *27*(12), 712–719.

Nakahashi, R., & Yamada, S. (2018, July). Modeling human inference of others' intentions in complex situations with plan predictability bias. In J. Z. Chuck Kalish Martina Rau & T. Rogers (Eds.), *Cogsci 2018* (pp. 2147–2152).

Nickerson, R. S. (1999). How we knowand sometimes misjudgewhat others know: Imputing one's own knowledge to others. *Psychological bulletin*, *125*(6), 737.

Pöppel, J., & Kopp, S. (2018). Satisficing models of bayesian theory of mind for explaining behavior of differently uncertain agents: Socially interactive agents track. In *Proceedings of the 17th international conference on autonomous agents and multiagent systems* (pp. 470–478).

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, *1*(4), 515–526.

Vul, E., Goodman, N. D., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? optimal decisions from very few samples. *Cognitive science*, *38*(4), 599–637.

Wellman, H. M., & Liu, D. (2004). Scaling of theory-of-mind tasks. *Child development*, *75*(2), 523–541.

Tracking the wandering mind: Memory, mouse movements and decision making styles

Mariana Rachel Dias da Silva (m.r.diasdasilva@tilburguniversity.edu)

Marie Postma (marie.postma@tilburguniversity.edu)

Tilburg University Cognitive Science and Artificial Intelligence Department,
Warandelaan 2, 5037AB Tilburg, The Netherlands

Abstract

Mind wandering involves internally focused attention and is often conceptualized as the opposite of external attention that is oriented towards the task at hand. Individuals vary according to the amount they mind wander as well as with regards to the pattern of oscillations between mind wandering thoughts and externally directed, focused thought. Assuming that mind wandering is influenced by episodic contents, we explore the proposition that mind wandering frequency is related to the manner in which individuals deal with the contents of episodic memory, as reflected by a maximizing decision making style. Based on previous studies measuring cognitive processes, we assume that mouse trajectories towards a particular response on the screen are continuously updated by time-dependent and temporally-dynamic cognitive processes. As a behavioral methodology, mouse tracking provides potential cues to help predict mind wandering. In our experiment, a total of 274 students completed a decision making questionnaire, episodic and associative memory tests (during which mouse movements were recorded) and a working memory task, during which mind wandering thoughts were assessed. We found certain mouse movement characteristics to be significantly predictive of mind wandering. Also, a maximizing decision making style appeared to be related to a particular type of mind wandering, namely, task-related interference.

Keywords:

mind wandering; episodic memory; mouse-tracking; decision making; maximizing

Introduction

Conscious experience is fluid and dynamic. Mind wandering (MW) involves a flow of thoughts, often from one topic to another, back and forth between the outside external world and internal thoughts and feelings. Where do these thoughts come from? Why do our thoughts wander elsewhere when we are trying to focus on a task? Can we detect whenever a person is mind wandering from their behavior? Over the last two decades, researchers have been investigating these questions empirically in hopes of understanding how we navigate our stream of consciousness and the world around us. In this paper, we aim to shed additional light upon these questions by focusing on behavioral cues to MW and the link between MW and different decision making styles.

Factors influencing MW

During MW, thoughts frequently focus on events that occur in distinct periods in time, either in the past or future, which suggests self-generated mental content to be largely a product of the episodic memory system (Smallwood & Schooler, 2015). Neural accounts of MW demonstrate increased activation in the medial temporal lobe subsystem (Andrews-Hanna, Reidler, Huang, & Buckner, 2010; Ellamil, Dobson, Beeman, & Christoff, 2012), which is associated with episodic retrieval (Klinger, 2013; Mittner, Hawkins, Boekel, & Forstmann, 2016). In addition to being a part of veridical episodic events, details from past experiences can also be recombined in order to construe episodic mental simulations and other mental states that become part of the stream of thought. During MW there is also increased activation in the dorsal medial subsystem, which is associated with social processes, scenarios, meaning and comprehension. The *default variability hypothesis* (Mills, Herrera-Bennett, Faber, & Christoff, 2018) proposes that thoughts ceaselessly move from one topic to the next, with heightened variability over time. The ceaseless flow serves to distinguish different memories while the variability of content serves to provide a time buffer between memories, improving episodic memory efficiency. In addition, heightened variability enables the extraction of commonalities and differences between memories and the eventual development of categorization and category boundaries. Commonalities allow for the creation of meaning while dissimilarities prevent the overlearning of categories. Thus, the default content variability in MW increases the opportunities for interleaved episodic to semantic transformations.

Regular oscillations between engagement with the external environment and engagement in internal thoughts are normative to the human brain functioning (Mills et al., 2018). However, patterns of oscillations are subject to a wide range of individual differences, which vary according to the context (Seli et al., 2018). In particular, low demand contexts (Smallwood & Andrews-Hanna, 2013), less task interest (Unsworth & Mcmillan, 2012) and greater fatigue (Walker & Trick, 2018) are related to more MW, to name a few possible factors. In addition to these factors, we are interested in how a person's episodic memory performance is related to this pattern of oscillations. Previous research has investigated mind wandering during episodic memory tests (Riby, Smallwood, & Gunn, 2008), finding that regardless of the amount of MW reported in a retrospective questionnaire, participants performed equally well on an episodic memory test. However, Event-Related Potentials (ERP) analyses indicated that low MW groups differed from high MW groups in their retrieval strategy. Those who did not mind wander a lot used a pure

recollection strategy[1] for remembering words which they had previously seen before and words which were new. However, those who mind wandered frequently were unable to easily recollect stimuli; to compensate, they used additional monitoring and strategic processes[2] in order to aid episodic remembering.

Decision making styles as indicators of MW

Are there decision making styles that are related to mind wandering? Mind wandering content is dependent on what enters into episodic memory. At the same time, there is variability in the manner in which individuals select and sift through the contents of episodic memory. For example, some individuals tend to become more stuck on particular memories, while others have a greater tendency to quickly navigate from topic to topic. Similarly, there is variability in the manner in which individuals sift through information as they make decisions. When making decisions, individuals must select relevant information to attend to and create meaning out of in order to make a choice (Beach, 1993). Previous research has distinguished between two decision making styles, one which involves a tendency to find the best possible alternative, or *maximizing*, and one which involves a tendency to find the option that is good enough, or *satisficing*. *Maximizers* have more difficulty in making decisions and tend to be less satisfied with their choices, meanwhile *satisficers* tend to have an easier time making decisions and tend to be more satisfied with their choices (Schwartz et al., 2002). Yet, what is it about the nature of *maximizing* and *satisficing* that might be related to mind wandering? We postulate that the rigid quality of a *maximizing* decision making style may be related to greater rumination (Paivandy, Bullock, Reardon, & Kelly, 2008), and in turn manifest as type of MW which involves a tendency to worry about performance on the task at hand (Dias da Silva, Rusz, & Postma-Nilsenová, 2018), namely, task-related interference. We therefore would expect a tendency to *maximize* to be related to more interfering thoughts about performance on a task which inhibit actual performance of the task itself.

Computer mouse movements as indicators of MW

From an embodied cognition perspective, which assumes cognition is evidenced in our bodily behaviors (Barsalou, 2008), as our minds are decoupled from the sensory environment during MW, our minds seem to also disengage from controlling behavioral motor outputs. Consequently, motor performance becomes more automatic or degraded (Franklin, Smallwood, & Schooler, 2011). Initial evidence for this was found in a study by Kam et al. (2012), in which participants were instructed to track a moving ball on a screen with a joystick. Intermittently during the task, participants were asked whether or not they were MW. In trials during which participants were MW, they deviated further from the correct path than in times during which they were focused. Additionally, Arapakis, Lalmas, and Valkanas (2014) found that various mouse movement measures were able to predict engagement− which is often contrasted with MW − in an unsupervisied manner. It is thus plausible that MW in online tasks can be inferred by hand reach movements which are continuously updated by ongoing mental processes (Spivey & Dale, 2006) and become more degraded and automatic during MW (Kam et al., 2012), as attention decouples from the task at hand.

Current Study

The primary goal of the present study is to investigate if episodic memory, decision making style, mouse movements and task interest can predict MW. Previous research consistently indicates a strong negative relationship between MW and task interest. However, little has been done in terms of the relationship between performance on episodic memory tests, mouse movements and MW. To our knowledge, no research so far explored the relationship between mind wandering and decision making styles. Therefore, the guiding questions in this research are: 1) What is the relationship between episodic memory performance, motor output, and mind wandering? 2) How are task interest and decision making styles related to MW? As our aim is to explore the relationship between various measures, we do not propose directional hypotheses.

Methods

Participants and Procedure

In total, 274 participants between 17 and 41 years of age $M = 22.09$), 180 female, performed this experiment and received course credit for their participation. Three participants were excluded due to a procedural error. The study was approved by the university's Institutional Review Board. Before beginning the experiment, participants signed a consent form. Participants then answered questions about their demographics and choice making orientation. Next, they performed episodic and associative memory tests, a working memory test during which mind wandering was measured, and finally, they filled out a questionnaire about their interest in the task (see Figure 1). Note some of the data has been reported in Dias da Silva and Postma-Nilsenová (2019)[3] and, thus, the current data and that data are not from independent samples. Specifically, the MW responses from the current participants are shared with Dias da Silva and Postma-Nilsenová (2019). The purpose of that study was to examine relations between mouse movements and MW probes during an operation span task. In this study, we rather explore relationships between various additional measures from episodic memory tests and decision making styles with the overall MW frequency reported during the OSPAN task.

Materials

Decision Making Decision-making orientation (Schwartz et al., 2002) is an individual difference variable that differ-

[1] as indicated by a larger magnitude of the left-parietal ERP component.

[2] as indicated by larger central negativity effects.

[3] submitted for publication.

entiates people according to how they make decisions. At one extreme, *maximizing* involves a tendency to find the best possible alternative, while at the other extreme, satisficing involves a tendency to find the option that is good enough. Decision-making orientation was assessed by the Maximization Scale (Cronbach $\alpha = .64$), consisting of 13 items assessed on a 7-point Likert scale (1 = completely disagree to 7 = completely agree). Higher scores on the scale reflect a general tendency to *maximize*, while lower scores on the scale reflect a general tendency to *satisfice*.

Episodic Memory: 15-Word List Learning (WLLT) and Recognition Tests (WRT) THE WLLT consisted of free recall of 15 semantically unrelated words (concrete, imaginable nouns), in three trials. Words were selected from SUBTLEX-NL, a database of Dutch word frequencies based on 44 million words from film and television subtitles (Keuleers, Brysbaert, & New, 2010). All words were bisyllabic, had 6 letters, and had a medium frequency ($Range = 2.25 - 3.45$, $M = 2.56$, $Mdn = 2.46$) and a prevalence of above 98%. Each word was presented on the screen for 2 seconds, in a random sequence. Between each set of words, participants performed a 20-second Brown-Peterson distraction task[4], which required them to count backwards from a 3-digit number presented on the screen. During the recall phase, participants were asked to write down the words they could recall. The score was the total number of words reproduced over three trials (0-45). Immediately after the WLLT, participants were shown 30 words (15 distractor words were presented in addition to the ones previously seen) in a random order on a computer screen and were instructed to explicitly recognize whether or not they had seen the word by clicking on yes or no with the computer mouse on the screen. This part was WRT. The score was the sum of true positive and true negative answers (0-30).

Associative Memory: Paired-Associate Learning (PALT) and Recognition Tests (PART) The PALT consisted of cued recall of 12 semantically related word pairs, and 12 semantically unrelated word pairs, constructed in the same format of the 15-word list learning test, with three trials, and a Brown-Peterson distraction task. Words were selected from SUBTLEX-NL (Keuleers et al., 2010). Word length varied from 3 to 8. All words had a prevalence of above 98% and had a medium frequency ($Range : 1.56 - 4.56$, $M = 3.04$, $Mdn = 3.03$). Semantic associations were made according to De Deyne and Storms (2008)'s word association norms, and semantic distance was additionally checked with Snaut (Mandera, Keuleers, & Brysbaert, 2017). Each pair was presented on the screen for 2 seconds. Between each set of 24 pairs, participants performed a 20-second Brown-Peterson distraction task. During the recall phase, participants were asked to write down the target word in response to each cue word which was randomly presented on the screen. The score was the sum of pairs reproduced over three trials (0-72). The PALT was followed immediately by a recognition test (PART), which involved forced choice of the target words of the PALT in response to the presented cue words. In each trial, three distractor words were simultaneously presented on the screen together with the target and cue[5]. Each cue was always presented with 2 semantically related words, and 2 semantically unrelated words. The score was the sum of correct answers (0-24).

Mind Wandering Intermittent thought probes assessing participant's state of mind were presented during a working memory task (Operation Span task, (Conway, Cowan, Bunting, Therriault, & Minkoff, 2002; Mrazek et al., 2012)). MW was calculated as the percentage of thought probes during which participants responded that they were either having task-unrelated thought (TUT) or task-related interference (TRI) (Stawarczyk, Majerus, & D'Argembeau, 2013). Focused attention (FA) was calculated as the percentage of thought probes during which participants were focused on the task.

Task-interest As a last part of the experiment, task interest (TI) was assessed using a 5-point Likert scale with 4 questions (Cronbach $\alpha = .82$): (a) Did you enjoy performing this task? (b) Did you take interest in this task?; (c) Are you interested in performing tasks like this?; and (d) Did you feel pleasant while performing the task? The response categories vary from 1 (not at all) to 5 (very much) (Van Yperen, 2003).

Instrumentation All questionnaires were presented online via Qualtrics. The episodic and associative memory tasks were programmed on OpenSesame 3.1.6 (Mathôt, Schreij, & Theeuwes, 2012). The experiment was run on full screen mode, with a resolution of 1024 by 768 pixels on a Windows 7 operating system. The desktop computer was placed on the table so that participants had enough room to move the mouse without running out of space. Mouse settings were left at their default values (medium acceleration and medium speed). A Dell USB3 Button Scrollwheel Optical Mouse was used to record cursor coordinates during the memory tests. Mouse movements were recorded both in the Recognition parts of the Word List and Paired-Associates tests. In the WRT (Fig. 1a), once participants click on the start button (341 by 85 pixels), two words (a target and a distractor) were displayed to them on the extreme top right and left corners of the screen (192 by 128 pixels). Once participants determined which of the words they had learned in the previous portion of the task, they made a selection with the computer mouse. During the PART (Fig. 1b), once the participants clicked the start button (341 by 85 pixels), they viewed a cue at the center of

[4]The Brown-Peterson distraction task was administered in order to prevent the confounding of episodic memory with short-term memory (as a result of recency effects which occur during learning tests) (Spaan, 2016).

[5]There were always 2 semantically related and two non-semantically related words presented on the screen (i.e.: If the target was semantically related to the cue, one distractor would also be semantically related to the cue and the other two would not).

the screen, along with 1 target and 3 distractors (192 by 128 pixels) distributed along the 4 extreme corners of the screen. Once they determined which word was associated with the cue, they made a selection with the computer mouse.

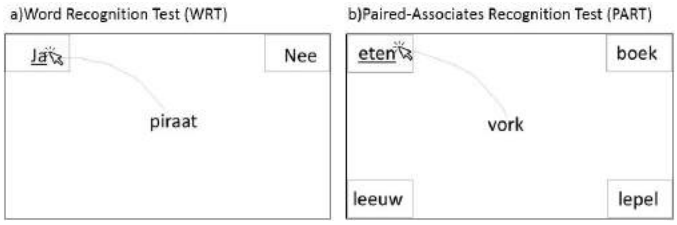


Figure 1: Illustration of the a)WRT and b)PART tests.

Data processing Participants' individual raw data files were merged and read into R version 3.4.1(R Core Team, n.d.). Mouse tracking data were imported and processed using the library mousetrap (Kieslich & Henninger, 2017). Trajectories were recorded from the moment the start button was clicked on, to the moment a target or distractor was selected in both the WRT and the PART tests. All trajectories aligned to a common starting position and were remapped onto one side. Various features such as total distance and maximum velocity[6] were calculated based on the mouse trajectories and aggregated per participant (see Supplementary Information).

Results

Data were analyzed for 271 participants. Descriptives for the WLLT, WRT, PALT, PART memory measures, decision making and task interest questionnaires can be found in Table 1. Ceiling effects were observed in the Word Learning Recognition Test (WRT) and in the Paired Associates Recognition Test (PART), as a majority participants had either perfect or near perfect scores on these tests. Therefore, they were not used further for statistical testing. As mouse movement coordinates were recorded during the tests, they were used for statistical testing instead.

Memory, computer mouse movements, and MW

Correlations In order to examine how both memory recall measures and mouse movements during the recognition tests[7] predict MW during a subsequent task, we first examined which variables were correlated with TUT, TRI, and FA. TUT frequency was found to be significantly correlated with mouse measures in the WRT, namely; maximum x-position ($r(269) = 0.17$, $p = .01$) and total distance travelled ($r(269) = 0.13$, $p = .04$). TRI frequency was positively correlated with various measures in the PART, namely; reaction time ($r(269) = 0.21$, $p < .000$), idle time ($r(269) = 0.18$, $p < .000$), time to maximum deviation towards the alternative response ($r(269) = 0.16$, $p = .01$), time to maximum deviation below the ideal path towards the selected response ($r(269) = 0.13$, $p = .03$), time to maximum deviation from the ideal path overall ($r(269) = 0.15$, $p = .01$), time to maximum acceleration ($r(269) = 0.19$, $p < .000$), time to maximum velocity ($r(269) = 0.18$, $p < .000$), and time to minimum acceleration ($r(269) = 0.19$, $p < .000$). In addition, TRI was negatively correlated with performance on non-semantic items in the paired recall test ($r(269) = -0.13$, $p = .03$). Lastly, FA was inversely correlated with the same measures that were positively correlated with TRI.

Table 1: Descriptives of Task Interest (TI), Maximizing, Word List Learning Test (WLT), Word Recognition Test (WRT), Paired Associates Learning Test (PALT), Paied Associates Recognition Test (PART), frequency of Task-unrelated Thoughts (TUT%), Task-Related Interference (TRI%), and Focused Attention (FA%).

Measure	Mean	SD	95% CI
TI (1-5)	3.31	0.84	3.21 - 3.41
Age	22.07	3.26	21.68 - 22.46
Maximizing (1-7)	4.38	0.70	4.30 - 4.46
WLLT	0.60	0.14	0.59 - 0.62
WRT	0.99	0.03	0.98 - 0.99
PALT	0.75	0.15	0.73 - 0.76
PALT s.	0.86	0.13	0.84 - 0.87
PALT n.s.	0.63	0.20	0.61 - 0.66
PART	0.97	0.09	0.96 - 0.98
PART s.	0.98	0.05	0.98 - 0.99
PART n.s	0.95	0.14	0.93 - 0.97
TUT(%)	8.51	13.46	6.91 - 10.11
TRI(%)	22.58	21.48	20.03 - 25.14
FA(%)	68.91	26.27	65.78 - 72.03

Note: s. = semantic; n.s. = nonsemantic

Dimensionality reduction Pearson's correlations between the mouse-tracking features indicate that some features may be measuring nearly identical underlying constructs (e.g.: *time to reach maximum velocity (WRT)* and *time to reach minimum acceleration (WRT)*, $r = 0.98$). Therefore, PCA (with oblimin rotation) was used to reduce the dimensionality of the data separately for the WRT and PART features, removing any multicollinearity. We used Kaiser's Criterion in order to determine the number of principal components in the WRT and in the PART separately. Five components were used that cumulatively accounted for 29% 57% 70% 77% and 85% of the variance in the mouse-tracking data in the WRT test, respectively. For the PART, 4 components were used that cumulatively accounted for 33% 61% 76% and 86% of the variance in the mouse-tracking data.

Regressions Subsequently, we performed two separate regressions, one with TRI percentage as the dependent variable and one with FA percentage as the dependent variable. As

[6] 28 mouse features for the WRT and 28 for the PART.

[7] Mouse movements were recorded during the WRT and PART.

input for the regressions, we included the PCA components which significantly correlated with MW frequency. Note that no PCA components were significantly correlated with TUT. The second component (temporal) from the PART was significantly correlated with TRI and FA frequency ($r = 0.18$ for TRI and $r = -0.17$ for FA).

Results of the regression indicate that percentage of TRI was significantly predicted by the temporal principal component ($R = .17$, adjusted-$R^2 = 0.03$, $F(1,269) = 8.56$, $p = .004$). Regression coefficients are shown in Table 2.

Table 2: Temporal Principal Component as a predictor of Task-related Interference.

	B	*SE(B)*	*t*	*p*
(Intercept)	22.68	1.29	17.57	$< .000$
PART TC2 (temporal)	3.78	1.29	2.93	.004

Adjusted $R^2 = 0.03$, $p = .004$

Similarly, percentage of FA was also significantly predicted by the temporal principal component ($R = .17$, adjusted-$R^2 = 0.03$, $F(1,269) = 8.23$, $p = .004$). Regression coefficients are shown in Table 3.

Table 3: Temporal Principal Component as a predictor of Focused Attention.

	B	*SE(B)*	*t*	*p*
(Intercept)	68.81	1.58	43.55	$< .000$
PART TC2 (temporal)	-4.54	1.58	-2.87	.004

Adjusted $R^2 = 0.03$, $p = .004$

Task Interest, Decision Making Style, and MW

In order to investigate the relationship between task interest, decision making style, and MW, we observed correlations between the variables. In line with previous findings (Unsworth & Mcmillan, 2012), task interest was positively correlated with FA ($r = 0.18$) and negatively correlated with TUT ($r = -0.24$). Interestingly, and novel to this research, we found that maximizing was positively related to TRI ($r = 0.12$) and negatively correlated to FA ($r = -0.12$). 2).

Discussion

In accordance with previous literature (Unsworth & Mcmillan, 2012), we found that MW is negatively correlated with task interest. Interestingly, we found a *maximizing* decision-making style to be be positively related to TRI and negatively related to FA. Novel to our research, we discovered that TRI percentage is related to more *maximizing*, while FA percentage is related to more *satisficing*. That is, the need to select the best possible option is reflected in the amount of TRI in a

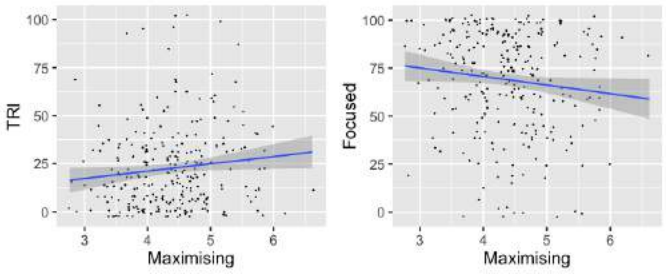


Figure 2: Relationship between TRI and Maximizing ($r = 0.12$) and FA and Maximizing ($r = -0.12$)

task, while satisfaction with selecting the good enough alternative is more related to FA during a task. This relationship has important implications, as it may be that a *maximizing* trait could potentially influence, through rumination, a tendency for having task-related interference, leading to poorer performance in tasks. Although this relationship was not directly tested in this study, future research could investigate the relationship between *maximizing*, rumination, and task-related interference further.

Moreover, we found a negative correlation between performance on non-semantic items in the Paired Recall task and TRI; however, we found no effect of any of the other memory tests on the proportion of TUT or TRI. This may be explained by two reasons. First, it may be that the tests were too easy, as reflected by the particularly high scores on the recognition tests. Second, it is likely that the scores on the episodic and associative memory tests do not accurately represent the aspects of episodic memory that are related to MW, i.e., aspects that are most likely contextually dependent and vary from individual to individual. Finally, if we observe the distribution of MW scores, over half of the participants never reported having TUT during the working memory task. This indicates that the OSPAN task was too engaging and demanding, leaving little room for TUT.

According to the *default variability hypothesis* (Mills et al., 2018), mind wandering serves the purpose of (episodic) memory consolidation. The results found by Riby et al. (2008) demonstrate that performance on episodic memory tests is unaffected by the proportion of mind wandering. However, low mind wanderers used a pure recollection strategy while high mind wanderers used additional monitoring strategies. Thus, it may be that mind wandering about the items in the episodic memory task helped consolidate memories of high mind wanderers during the task. However, something that neither Riby et al. (2008) nor our study did was assess the content of MW thoughts during the task. In order for us to verify the *default variability hypothesis* in the short term, as measured by episodic memory tests, it is also necessary that we consider the contents of mind wandering thoughts. For instance, MW about the items in the episodic memory task versus MW about something completely unrelated would likely have differential effects on memory con-

solidation.

Despite the ceiling effects we found in the recognition tests, we did find that mouse movements recorded during both episodic and associative forced choice recognition tests are related (albeit weakly) to MW in a subsequent working memory task. Therefore, it may be that a greater proportion of MW during a task is related to a general tendency to mind wander and, thereby, be detectable in specific overall motor behaviors beyond the task during which a person is mind wandering. In this study, mouse movements during episodic and associative memory tasks served to predict task-related interference (albeit weakly) and focused attention during a subsequent task. The most important feature in predicting task-related interference and focused attention was a temporal principal component, which contains information about the evolution of trajectories over time. This is consistent with the highly significant correlations that emerged between TRI and the various time-related mouse measures (*RT, idle time, time to reach maximum deviation, time to reach maximum velocity, etc.*) Such features characterize the degree of commitment towards a response during mind wandering, such that negative correlations (with FA) represent quicker and more automatic decisions, while positive correlations (with TRI) represent a delay in the commitment towards a response.

Returning to Riby et al. (2008)'s findings, high mind wanderers differed from low mind wanderers in their use of additional monitoring and strategic processing to compensate for mind wandering. Linking their findings to ours, it may be that monitoring and strategic processing are indicated differently by general mouse movement features according to the type of MW thought. Our findings indicate that some mouse movement features correlated with TUT in one task, and other mouse movements correlated with and predicted TRI and FA in another task. This may be explained by the differences in the two tasks (WRT & PART). The WRT only had 2 alternatives, while the PART had 4 alternatives. Moreover, the tests recruited different parts of memory differentially - the WRT only involved recognition of previously seen words during the WLLT, while the PART required the recruitment of associative memory[8] for remembering associations between words.

Finally, in order to better understand how MW may be related to decision making as well as performance on episodic memory tasks and overall motor behaviors, it would be relevant to assess trait differences in MW in addition to state differences. Moreover, it would be interesting to see if the relationship between trait MW and motor movements generalize to different types of computer mouse-based tasks.

Conclusion

The relationship between episodic memory and MW is a complex one, and it is likely that the episodic and associative memory tests which we used were unable to capture this relationship fully. This may be either due to the tests demands being too low and hence not able to capture individual differences in terms of accuracy, or because the tests did not capture the aspects of episodic memory that vary according to the context and to the individual. Interestingly though, we have found evidence for a relationship between specific computer mouse movements and MW, which warrant further investigation. In particular, future research should see if our findings generalize to unseen data. Lastly, we have found a novel relationship between MW and *maximizing*, in that *maximizing* was related to an increased frequency of TRI and less FA. Our aim in this study was to explore an encompassing model of mind wandering starting from its inputs, determined by what enters into episodic memory and ending with behavioral outputs, which are visible in mouse movement patterns. We believe we have taken a small step for a better understanding of how our minds wander and navigate this world.

[8] In addition, the PART requires inhibition of previously learned semantic associations learned in different contexts.

Supplementary Information

Experiment Materials All materials used in the task are available at https://osf.io/dse3k/

References

Andrews-Hanna, J. R., Reidler, J. S., Huang, C., & Buckner, R. L. (2010). Evidence for the Default Network's Role in Spontaneous Cognition. *J Neuro-physiol*, *104*, 322–335. doi: 10.1152/jn.00830.2009

Arapakis, I., Lalmas, M., & Valkanas, G. (2014). Understanding Within-Content Engagement through Pattern Analysis of Mouse Gestures. In *Cikm.* Shanghai, China.. doi: 10.1145/2661829.2661909

Barsalou, L. W. (2008, 1). Grounded Cognition. *Annual Review of Psychology*, *59*(1), 617–645. doi: 10.1146/annurev.psych.59.103006.093639

Beach, L. R. (1993, 7). Broadening the Definition of Decision Making: The Role of Prechoice Screening of Options. *Psychological Science*, *4*(4), 215–220. doi: 10.1111/j.1467-9280.1993.tb00264.x

Conway, A. R., Cowan, N., Bunting, M. F., Therriault, D. J., & Minkoff, S. R. (2002, 3). A latent variable analysis of working memory capacity, short-term memory capacity, processing speed, and general fluid intelligence. *Intelligence*, *30*(2), 163–183. doi: 10.1016/S0160-2896(01)00096-4

De Deyne, S., & Storms, G. (2008). Word associations: Norms for 1,424 Dutch words in a continuous task. *Behavior Research Methods*, *40*(1), 198–205. doi: 10.3758/BRM.40.1.198

Dias da Silva, M. R., & Postma-Nilsenová, M. (2019). *Mindering Mice, Wandering Minds: Using computer mouse tracking to predict mind wandering.*

Dias da Silva, M. R., Rusz, D., & Postma-Nilsenová, M. (2018, 11). Ruminative minds, wandering minds: Effects of rumination and mind wandering on lexical associations, pitch imitation and eye behaviour. *PLOS ONE*, *13*(11). doi: 10.1371/journal.pone.0207578

Ellamil, M., Dobson, C., Beeman, M., & Christoff, K. (2012, 1). Evaluative and generative modes of thought during the creative process. *NeuroImage*, *59*(2), 1783–1794. doi: 10.1016/j.neuroimage.2011.08.008

Franklin, M. S., Smallwood, J., & Schooler, J. W. (2011, 10). Catching the mind in flight: Using behavioral indices to detect mindless reading in real time. *Psychonomic Bulletin & Review*, *18*(5), 992–997. doi: 10.3758/s13423-011-0109-6

Kam, J. W. Y., Dao, E., Blinn, P., Krigolson, O. E., Boyd, L. A., & Handy, T. C. (2012). Mind wandering and motor control: off-task thinking disrupts the online adjustment of behavior. *Frontiers in Human Neuroscience*, *6*, 329. doi: 10.3389/fnhum.2012.00329

Keuleers, E., Brysbaert, M., & New, B. (2010, 8). SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles. *Behavior Research Methods*, *42*(3), 643–650. doi: 10.3758/BRM.42.3.643

Kieslich, P. J., & Henninger, F. (2017, 10). Mousetrap: An integrated, open-source mouse-tracking package. *Behavior Research Methods*, *49*(5), 1652–1667. doi: 10.3758/s13428-017-0900-z

Klinger, E. (2013). Goal commitments and the content of thoughts and dreams: Basic principles. *Frontiers in Psychology*, *4*(JUL), 1–17. doi: 10.3389/fpsyg.2013.00415

Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, *92*, 57–78. doi: 10.1016/j.jml.2016.04.001

Mathôt, S., Schreij, D., & Theeuwes, J. (2012, 6). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*, *44*(2), 314–324. doi: 10.3758/s13428-011-0168-7

Mills, C., Herrera-Bennett, A., Faber, M., & Christoff, K. (2018). Why the mind wanders: How spontaneous thought's default variability may support episodic efficiency and semantic optimization. In Kieran C.R. Fox & Kalina Christoff (Eds.), *The oxford handbook of spontaneous thought: Mind-wandering, creativity, and dreaming.* Oxford University Press.

Mittner, M., Hawkins, G. E., Boekel, W., & Forstmann, B. U. (2016, 8). A Neural Model of Mind Wandering. *Trends in cognitive sciences*, *20*(8), 570–8. doi: 10.1016/j.tics.2016.06.004

Mrazek, M. D., Smallwood, J., Franklin, M. S., Chin, J. M., Baird, B., & Schooler, J. W. (2012). The role of mind-wandering in measurements of general aptitude. *Journal of Experimental Psychology: General*, *141*(4), 788–798. doi: 10.1037/a0027968

Paivandy, S., Bullock, E. E., Reardon, R. C., & Kelly, F. D. (2008, 11). The Effects of Decision-Making Style and Cognitive Thought Patterns on Negative Career Thoughts. *Journal of Career Assessment*, *16*(4), 474–488. doi: 10.1177/1069072708318904

R Core Team. (n.d.). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing.

Riby, L. M., Smallwood, J., & Gunn, V. P. (2008, 6). Mind Wandering and Retrieval from Episodic Memory: A Pilot Event-Related Potential Study. *Psychological Reports*, *102*(3), 805–818. doi: 10.2466/pr0.102.3.805-818

Schwartz, B., Ward, A., Monterosso, J., Lyubomirsky, S., White, K., & Lehman, D. R. (2002). Maximizing Versus Satisficing: Happiness Is a Matter of Choice. *Journal of Personality and Social Psychology*, *83*(5), 1178–1197. doi: 10.1037/0022-3514.83.5.1178

Seli, P., Kane, M. J., Smallwood, J., Schacter, D. L., Maillet, D., Schooler, J. W., & Smilek, D. (2018, 6). Mind-Wandering as a Natural Kind: A Family-Resemblances View. *Trends in cognitive sciences*, *22*(6), 479–490. doi: 10.1016/j.tics.2018.03.010

Smallwood, J., & Andrews-Hanna, J. (2013). Not all minds that wander are lost: The importance of a balanced perspective on the mind-wandering state. *Frontiers in Psychology*, *4*(AUG), 1–6. doi: 10.3389/fpsyg.2013.00441

Smallwood, J., & Schooler, J. W. (2015). The Science of Mind Wandering: Empirically Navigating the Stream of Consciousness. *Annual Review of Psychology*, *66*(1), 487–518. doi: 10.1146/annurev-psych-010814-015331

Spaan, P. E. (2016). Episodic and semantic memory impairments in (very) early Alzheimers disease: The diagnostic accuracy of paired-associate learning formats. *Cogent Psychology*, *3*(1), 1–25. doi: 10.1080/23311908.2015.1125076

Spivey, M. J., & Dale, R. (2006). Continuous Dynamics in Real-Time Cognition. *Association for Psychological Science*, *15*(5), 207–211. doi: 10.1111/j.1467-8721.2006.00437.x

Stawarczyk, D., Majerus, S., & D'Argembeau, A. (2013). Concern-induced negative affect is associated with the occurrence and content of mind-wandering. *Consciousness and Cognition*, *22*(2), 442–448. doi: 10.1016/j.concog.2013.01.012

Unsworth, N., & Mcmillan, B. D. (2012). Mind Wandering and Reading Comprehension: Examining the Roles of Working Memory Capacity, Interest, Motivation, and Topic Experience. *Journal of Experimental Psychology: © 2012 American Psychological Association Learning, Memory, and Cognition*, *39*(3), 832–842. doi: 10.1037/a0029669

Van Yperen, N. W. (2003). Task Interest and Actual Performance: The Moderating Effects of Assigned and Adopted Purpose Goals. *J Pers Soc Psychol.*, *85*(6), 1006–15. doi: 10.1037/0022-3514.85.6.1006

Walker, H. E., & Trick, L. M. (2018, 11). Mind-wandering while driving: The impact of fatigue, task length, and sustained attention abilities. *Transportation Research Part F: Traffic Psychology and Behaviour*, *59*, 81–97. doi: 10.1016/J.TRF.2018.08.009

Crowdsourcing effective educational interventions

J. Hunter Priniski (priniski@ucla.edu)
Department of Psychology, University of California, Los Angeles
Los Angeles, CA, USA

Zachary Horne (zachary.horne@asu.edu)
School of Social and Behavioral Sciences, Arizona State University
Phoenix, AZ, USA

Abstract

Creating effective educational interventions that correct people's misconceptions is difficult. This has led many researchers to conclude that people do not properly attend to new information in a way that they should. However, even if a scientifically-grounded intervention fails, it is still possible that other interventions would be effective. Yet, it is not practically feasible to systematically explore and test the entire hypothesis space of possible interventions. Here, we examined whether researchers could use online arguments to develop effective educational interventions, in effect, narrowing the intervention hypothesis space. Across two experiments ($N = 1,816$), we found that arguments crowdsourced from Reddit's Change My View were as effective or more effective at changing beliefs than interventions developed by academics and published in top-tier scientific journals. These results suggest that researchers can build on successful crowdsourced arguments to develop effective educational interventions likely to correct people's misconceptions in more naturalistic settings.

Keywords: belief change; crowdsourcing; crowd work

Introduction

It can be difficult to find common ground with people we disagree with. People's beliefs about polarizing issues are often deeply entrenched and evidence that counters these beliefs generally does not lead people to change their minds. This intransigence comes at a cost: Polarization is a growing problem in the United States (Pew Research Center, 2014) and widespread misinformation and misconceptions about, for example, climate change only exacerbate polarization, posing considerable challenges to society. What's more, even in situations when very few people hold a misinformed belief——such as believing that vaccines cause autism——the consequences can still have a widespread negative effect in society; this is evident from the recent resurgence of measles borne from parents refusing to vaccinate their children, citing fears that vaccines cause autism (Center for Disease Control, 2019).

To effectively educate the public, researchers have attempted to confront belief polarization and resistance to evidence by experimentally testing whether educational interventions can induce rational belief updating (e.g., Horne, Powell, Hummel, & Holyoak, 2015; Lai et al., 2014; Nyhan & Reifler, 2015; Nyhan, Reifler, Richey, & Freed, 2014; Turetsky & Sanderson, 2018). Ideally, people would always properly update their beliefs in accordance with the evidence. However, many interventions developed by scientists are ineffective (e.g., Nyhan et al., 2014), leading researchers to conclude that people cannot change their beliefs about issues such as climate change, vaccination, or immigration.

There are several psychological explanations that might explain why educational interventions are often ineffective. First, people interpret evidence to confirm their previously-held beliefs (Klayman, 1995; Nickerson, 1998), and our strongly-held beliefs——such as political and moral beliefs——are deeply rooted in our views of ourselves (e.g., Strohminger & Nichols, 2014; Carney, Jost, Gosling, & Potter, 2008), and thus are particularly resistant to change (Kahan, Peters, Wittlin, Slovic, & Ouellette, 2012). Second, even when people assimilate evidence, they do so imperfectly, requiring much more evidence than seems epistemically warranted (e.g., Priniski & Horne, 2018). Even massive education campaigns seem to yield only minor changes in public opinion and behavior (e.g., Fiore et al., 1990). Together, these results have led many researchers to either conclude that meaningful belief change is, in a practical sense, infeasible or that something other than education and evidence is needed to overcome strongly-held beliefs.

However, when an educational intervention fails to change people's misconceptions, this does not entail that other educational interventions (even similar interventions) would fail as well. It is an empirical question whether an untested intervention would turn out to be efficacious. Indeed, researchers have successfully developed effective educational interventions. For instance, Lewandowsky, Gignac, & Vaughan (2013) found that making people aware of the scientific consensus surrounding climate change using icon arrays positively affected people's beliefs. More recently, researchers have found that educational interventions can change vaccine intentions (Horne et al., 2015), correct mental health misperceptions (Turetsky & Sanderson, 2017), and address implicit racial biases, though these changes may be transient (Lai et al., 2014). However, beyond combing the academic literature, researchers have little to go on in predicting whether a given untested intervention will succeed or fail. Moreover, educational interventions are rarely tested outside of the lab, which allows for the possibility that effective educational interventions developed in the lab will fail to generalize beyond tightly controlled settings (Priniski & Horne, 2018). To complicate matters further, the hypothesis space of

possible interventions is very large (read, infinite). Consequently, it is not feasible for any given lab or even a group of labs to systematically explore the entire hypothesis space of educational interventions to determine whether a possible intervention could change people's beliefs about a given topic. A methodological advance is needed to avoid a protracted search through the intervention hypothesis space.

We propose a new method for developing educational interventions: Using successful persuasive arguments culled from online discussions (for example, from the Reddit forum Change My View). We propose that developing interventions based on existing arguments that have proven to be effective in naturalistic environments provides a compelling starting place for the development of effective educational interventions.

Change My View

Change My View is a popular Reddit forum where users post their views on issues ranging from gun control to opinions about movies. Redditors posting in this community understand that others will attempt to change their view by providing arguments opposing their beliefs (see Table S1 in Supplemental Materials, found at https://osf.io/v54ut/). As one would expect, some arguments are more persuasive than others and thus the variance in argument quality found on the forum provides a naturalistic resource for examining the features of effective arguments.

As a naturalistic data source, Change My View has provided several insights into how belief change occurs outside of the lab. For example, Priniski and Horne (2018) found that arguments containing more statistical language and links to news or scientific articles were more likely to change other users' strongly-held beliefs—evidence can change people's minds. Other researchers have examined the logical qualities of effective arguments on the forum (e.g., use of classical modes of persuasion: ethos, logos, pathos, Hidey, Musi, Hwang, Muresan, & McKeown, 2018). Research on Change My View has extended beyond social psychology. Computer scientists have developed computational models that extract features of argumentation, such as predicting the probability an argument is effective given linguistic features (Tan, Niculae, Danescu-Niculescu-Mizil, & Lee, 2016) or machine classifying "parts" of beliefs most amenable to change (Jo et. al, 2018).

While many researchers have examined the factors that predict belief change among Change My View users, it is unknown whether effective arguments taken from this forum would be equally effective in more controlled contexts or among a population not seeking arguments opposing their beliefs. In fact, there are several reasons why belief revision may look different on Change My View than it does in the lab. These reasons pose concern for the generalizability of effective arguments found on Change My View and need to be experimentally addressed before Change My View can be recommended as a crowdsourcing platform for effective educational interventions.

For one, people who discuss certain topics——and particularly users on Reddit's Change My View——may be more willing to change their minds and consider evidence for an opposing argument. This may not be true for the public at large, limiting the generalizability of these prior findings. Second, people engaged in a debate on a particular topic may be more motivated to deliberate on the topics they're discussing. This fact may make online communities such as Change My View an ideal population to study central rather than peripheral routes to persuasion (Petty & Cacioppo, 1986). However, it may also make online communities unrepresentative of the general population who may not be so ready to entertain evidence that is contrary to their beliefs.

Altogether, controlled laboratory research is necessary to understand if the persuasive tactics deployed online can generalize to other populations and, in turn, serve as a starting place for developing educational interventions.

Present Experiments

In the present experiments, we identified successful arguments on Change My View and performed a head-to-head comparison to interventions reported in academic psychology, public policy, political science, communications, and behavioral economics articles—adopting a methodological approach most analogous to a strategy relied on in clinical trials (e.g., Leuch, et al., 2013). Namely, we compared crowdsourced arguments to academic arguments that have been shown to be somewhat effective at changing people's beliefs (or at a minimum, exert the same task demands on participants). Performing this comparison allowed us to predict whether effective educational interventions can be culled from online communities and used as effective interventions in controlled laboratory settings.

It is worth highlighting how this experimental strategy diverges from comparing the performance of an intervention to an inactive control condition. As opposed to controlling for features of naturalistic interventions to uncover what makes them effective, the paradigm we are proposing first identifies the interventions that yield desirable consequences (e.g., a reduction in misconceptions surrounding structural racism), at which point we can subsequently uncover the mechanisms that realize these positive effects. As a consequence, academic and crowdsourced interventions will differ along many unknown dimensions (including length, the task performed, the information presented, and so on). However, we do have prior evidence (either from empirical studies or from data mined from discussion forums) that signal the efficacy of each of the interventions being compared. Ultimately, researchers aim to develop interventions that can effectively educate the public, making this dimension——efficacy——the most central on which to assess an educational intervention.

With this goal in mind, in Experiment 1 we compared the efficacy of crowdsourced and academic interventions at changing beliefs across four hotly-debated topics. Experiment 2 was an extension of Experiment 1, where we further examined whether crowdsourced arguments would be as effective as academic intervention across four new topics.

Experiment 1

Preregistration The projected sample size, predictions, and analysis scripts were preregistered through Open Science Framework. Experimental scripts, analyses, scales, and Supplemental Materials are available at https://osf.io/v54ut/.

Participants We recruited 916 participants through Amazon's Mechanical Turk to be 80% powered to detect a Cohen's *d* of .1 in a within-subjects design. Of the participants recruited, 816 passed attention checks and were included in the analysis of this study (333 men, 476 women, 4 non-binary, 4 preferred not to say; the median age of participants was 35 years old).

Interventions Participants received four separate interventions that focused on either (a) reducing racist beliefs, (b) increasing support for vaccines, (c) increasing support for gun control, and (d) reducing xenophobic attitudes directed at immigrants. Participants received two crowdsourced interventions and two academic interventions (intervention type: within-subjects) with one intervention for each topic. Therefore, we tested the efficacy of eight interventions in total. Crowdsourced interventions were copied-and-pasted comments that were awarded a "delta" in a Change My View discussion—a signification that the argument changed the view of at least one user on the forum. We selected discussion comments from Change My View as crowdsourced interventions if they met the following three criteria. First, the comment was related to a topic that psychologists have traditionally studied in the lab (e.g., climate change, gun control, xenophobia, etc.). Second, the comment had been awarded a delta. Third, the content of the comments could be developed into an intervention with little-to-no editing, content change, or manipulation. Many comments on Change My View satisfy these criteria and *could* have been empirically tested, but the aim of the present studies is to consider how several representative crowdsourced examples could be developed into effective educational interventions. (Detailed information about the interventions can be found at https://osf.io/v54ut/).

Pretest and Posttest Measures We examined how participants' beliefs about four controversial topics changed as a function of exposure to one of two educational interventions (crowdsourced or academic) for a given topic. Prior to completing the main portion of the study, participants answered four questions assessing their pretest beliefs about each topic. For instance, participants rated their agreement with the assertion, "Gun control in America is ineffective at reducing overall violence and crime", which was taken from a Change My View post (in this case, a post about gun control). After responding to these four assertions, participants proceeded to the intervention and post-test portion of the experiment.

We developed four separate scales to measure people's beliefs about racism, vaccines, gun control, and xenophobia directed at immigrants. Each scale was composed of five items (with two items reverse coded). Items in a topic's posttest scale were created by rewording or expanding on a pretest assertion. For example, an item in the posttest gun control scale stated, "Societies with strict gun control have similar crime rates as societies with little to no gun control." See the Supplemental Materials for more details on pretest and posttest measures.

Procedure The experiment proceeded as follows: First, participants rated their agreement with items measuring their pretest beliefs towards all four topics. Next, participants were randomly assigned either an academic or a crowdsourced intervention for a given topic. After completing this intervention (e.g., after reading information about gun control), participants responded to that topic's posttest scale. After completing the posttest scale for a given topic, participants advanced to a new topic and the procedure was reiterated until they finished reading and responding to questions about all four topics. The ordering and exposure to a given intervention type was counterbalanced and randomized.

Results and Discussion

Analytic Approach To test our hypotheses, we performed Bayesian mixed effects modeling using the R package `brms` (Burkner, 2018). We set regularizing priors for all population-level effects in our models, which we detail below. These priors are recommended because they provide conservative effect size estimates and reduce the likelihood of overfitting (Gelman, Lee, & Guo, 2015; McElreath, 2016). Following the recommendations of Liddell & Kruschke (2018), Likert data were modeled with a cumulative probability distribution. The cumulative distribution is recommended for Likert scale data because it assumes that ordered responses represent a continuous latent construct.

We tested our hypothesis by fitting an ordinal mixed-effects model predicting posttest beliefs based on the interaction between condition (Reference = Academic condition) and topic (Reference = Guns). This model controlled for participants' responses to the pretest statement, which we treated as a monotonic effect. This model included group-level effects of Subject and Topic and allowed for heterogeneity in the slopes of the effects of Condition and Topic on participants' responses. Our model is specified below in `brms` syntax (Bürkner, 2018):

```
Response ~ Condition*Topic + mo(PreTest)
+ (1 + Topic + Condition | Subject)
```

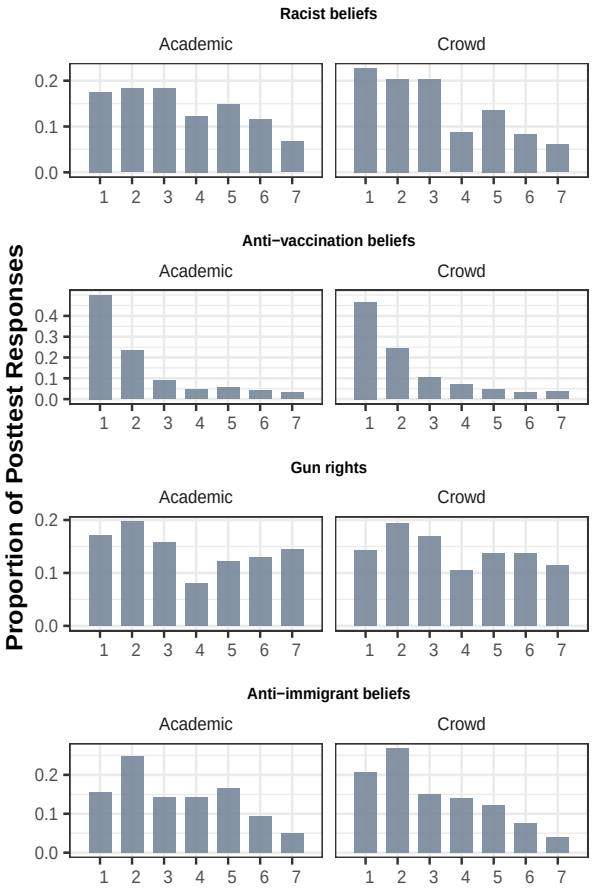


Figure 1: Posttest responses for each intervention tested in Experiment 1 (1 = Strongly disagree; 7 = Strongly agree). Relative effectiveness of a crowdsourced intervention can be seen by comparing the leftward shift of responses across interventions for a topic. Figures S5 through S8 in the Supplemental Materials show posttest responses grouped by pretest response for each intervention tested in tested in Experiment 1.

Bayesian analyses formulate model parameters as probability distributions wherein the posterior distribution for a parameter θ is computed via the prior and likelihood of θ. To model the joint probability distribution of participants' responses, we specified priors over the possible effects each parameter could have on our response variable:

$\beta_{Intercept[1]} \sim \mathcal{N}(2.19, 1)$
$\beta_{Intercept[2]} \sim \mathcal{N}(2.94, 1)$
$\beta_{Intercept[3]} \sim \mathcal{N}(3.17, 1)$
$\beta_{Intercept[4]} \sim \mathcal{N}(3.47, 1)$
$\beta_{Intercept[5]} \sim \mathcal{N}(3.89, 1)$
$\beta_{Intercept[6]} \sim \mathcal{N}(4.59, 1)$
$\beta_{Condition} \sim \mathcal{N}(0, .5)$
$\beta_{Pretest\ Beliefs} \sim \mathcal{N}(4, 2)$
$\beta_{Topics} \sim \mathcal{N}(0, 3)$
$\beta_{Topic \times Condition\ Interactions} \sim \mathcal{N}(0, .5)$
$\Omega_k \sim LKJ(1)$ where Ω_k is a correlation matrix of group-level parameter
Group-level parameters $\sim \mathcal{N}(1, 2)$

These analyses revealed that the crowdsourced interventions countering racist (b = −.58, 95% CI [−.80, −.37]) and anti-immigrant beliefs (b = −.40, 95% CI [−.60, −.18]) were credibly more effective than an academic intervention; interventions on vaccines and gun control were equally effective (see Figure 1). These results suggest that there are arguments being developed in online communities that are comparably effective to interventions behavioral scientists have developed. And considering crowdsourced arguments have the additional virtue of being shown to be effective in a naturalistic setting free from task demands, this may give additional motivation for beginning development of educational interventions on the basis of crowdsourced arguments.

However, given that the present design lacks a completely neutral control condition, it is important to be clear on what these results do not show. First, these results do not demonstrate the true magnitude of the effect of a given intervention. Second, there is a large amount of variance in intervention quality and effectiveness for any intervention type, and there is no reason to think that all crowdsourced arguments will always be as effective or more effective than academic interventions. Rather, one should interpret the results of Experiment 1 as suggesting that crowdsourced arguments can provide a starting place for developing educational interventions and doing so has the additional virtue of giving us a priori reason to think they will generalize to comparatively more naturalistic settings.

Experiment 2

Experiment 2 was a preregistered extension of Experiment 1. The registration for this project can be found at https://osf.io/v54ut/. This experiment followed an identical procedure but tested the efficacy of academic and crowdsourced interventions on four new topics: (a) reducing sexist beliefs, (b) reducing transphobic beliefs, (c) reducing denial in the negative effects of climate change, and (d) reducing favor for capital punishment.

Participants We recruited 900 participants through Amazon's Mechanical Turk to be 80% powered to detect a Cohen's d of .1 in a within-subjects design. Of the participants recruited, 745 passed attention checks and were included in the analysis of this study (325 men, 416 women, 3 non-binary, 1 preferred not to say; the median age of participants was 33 years old).

Results and Discussion

Like Experiment 1, we predicted that crowdsourced interventions would be as effective or more effective than academic interventions for the four new topics. We fit the same ordinal regression model with the same priors as

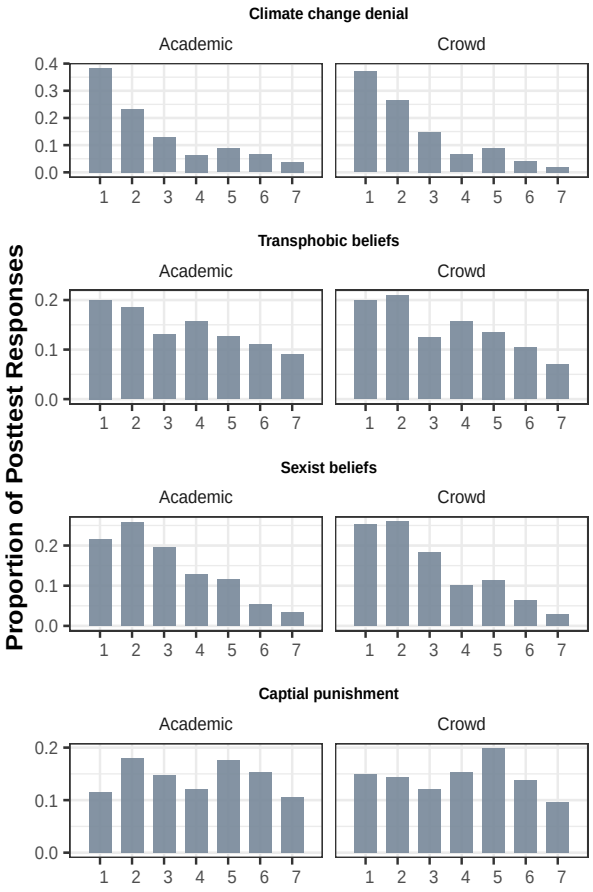


Figure 2: Posttest responses for each intervention tested in Experiment 2 (1 = Strongly disagree; 7 = Strongly agree). Relative effectiveness of a crowdsourced intervention can be seen by comparing the leftward shift of responses across interventions for a topic. Figures S9 through S12 in the Supplemental Materials show posttest responses grouped by pretest response for each intervention tested in tested in Experiment 2.

Experiment 1.

In Experiment 2, we found that crowdsourced interventions were equally effective as academic interventions across three topics; the academic intervention aimed at shifting people's beliefs about climate change was more effective than the crowdsourced intervention, b = .24, 95% CI [.00, .40].

Discussion

People's beliefs about topics like science and morality are stubbornly resistant to new information. Developing educational interventions to correct these beliefs is a difficult task that often results in fruitless outcomes. It is also often unknown whether an intervention that manages to successfully shift beliefs in the lab will be similarly effective in a more naturalistic setting. The present studies suggest that researchers can use crowdsourced arguments to better predict and develop effective educational interventions. Furthermore, crowdsourcing effective arguments can impact the study of belief revision directly by elucidating which types of information are most effective at changing strongly held beliefs: a topic of interest to many researchers studying higher-level cognitive processes. In two experiments, we tested whether arguments crowdsourced from the Reddit forum Change My View could be used to such an end. In Experiments 1 and 2, we compared arguments crowdsourced from Change My View to interventions taken from academic research in psychology, communications, political science, behavioral economics, and public policy. In Experiment 1, we found that across four topics, crowdsourced arguments were as effective or more effective at changing beliefs compared to previously published or tested educational interventions developed by academics. Experiment 2 followed the same procedure, finding that crowdsourced interventions were as effective at changing beliefs in three of four topics. In only one case did an academic intervention perform better at correcting scientific misconceptions than a crowdsourced intervention.

In light of these results, we propose that arguments mined from online communities can be used to develop educational interventions. How might this process work? Consider the results in Experiment 2: We observed that an academic intervention containing an icon array (Lewandowski, et al., 2013) was more persuasive than a similar crowdsourced intervention that did not contain data visualization. This finding is consistent with a large body of research demonstrating that data visualizations can effectively communicate complex information (e.g., Fernandes, Walls, Munson, Hullman, & Kay, 2018). In future research, we propose that researchers could begin to develop an educational intervention by first turning to crowdsourced interventions that appear effective and then extending them based on well-established theoretical considerations. For instance, we found that a crowdsourced intervention about the repercussions of structural racism was much more effective than an academic intervention aimed at shifting people's implicit racial biases (Lai et al., 2014). One possibility, then, is that we could further improve the efficacy of this crowdsourced intervention by augmenting it with compelling visualizations. In this way, researchers would be able to develop interventions that have the twin virtues of demonstrating prior success in naturalistic environments and having strong empirical support from controlled laboratory studies.

However, the present experiments have some clear limitations. By design, both experiments lacked a true control condition, leaving an important question unanswered: Exactly how effective are these interventions at changing beliefs? The present studies compared the *relative* effectiveness of crowdsourced interventions to academic interventions, and didn't demonstrate how effective they are

with respect to a neutral control condition. Future work should compare interventions to a true control condition in order to make explicit how effective a given intervention is at changing beliefs.

Change My View is also not the only place researchers could crowdsource effective arguments; a web application could also assist in mining, for example, Facebook and Twitter for effective arguments. The tool we are proposing could take queries (e.g., topics for an intervention) and return effective arguments filtered by the searched terms. Such a system could allow researchers to not only crowdsource educational interventions more effectively, but also gain an understanding of how arguments are communicated and received among members of online communities.

A cursory look on Reddit, Twitter, and Facebook demonstrates that people naturally engage in (sometimes) persuasive argumentation. Here, we proposed that psychologists can mine this information to efficiently create educational interventions that are more likely to persuade people than the methods researchers currently use—crowdsourced interventions have the advantage of being vetted, so to speak, in naturalistic contexts. Two experiments provide support for this proposal. We observed that crowdsourced arguments were more effective or often as effective as academic interventions aimed at correcting misconceptions about several societally important topics.

References

Bürkner, P. C. (2018). Advanced Bayesian Multilevel Modeling with the R Package brms. *The R Journal*, *10*, 395–411.

Carney, D. R., Jost, J. T., Gosling, S. D., & Potter, J. (2008). The secret lives of liberals and conservatives: Personality profiles, interaction styles, and the things they leave behind. *Political Psychology*, *29*, 807–840.

Center for Disease Control. (2019). Measles cases in the u.s. are highest since measles was eliminated in 2000. Retrieved from cdc.gov.

Fernandes, M., Walls, L., Munson, S., Hullman, J., & Kay, M. (2018). Uncertainty Displays Using Quantile Dotplots or CDFs Improve Transit Decision-Making. In R. Mandryk & M. Hancock (Eds.), *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems.* New York, NY: Association for Computing Machinery.

Fiore, M. C., Novotny, T. E., Pierce, J. P., Giovino, G. A., Hatziandreu, E. J., Newcomb, P. A., … Davis, R. M. (1990). Methods Used to Quit Smoking in the United States: Do Cessation Programs Help? *JAMA*, *263*, 2760–2765.

Flynn, D. J., Nyhan, B., & Reifler, J. (2017). The nature and origins of misperceptions: Understanding false and unsupported beliefs about politics. *Advances in Political Psychology*, *38*, 127–150.

Gelman, A., Lee, D., & Guo, J. (2015). Stan: A Probabilistic Programming Language for Bayesian Inference and Optimization. *Jounral of Educational and Behavioral Statistics*, *40*, 530–543.

Horne, Z., Powell, D., Hummel, J. E., & Holyoak, K. J. (2015). Countering antivaccination attitudes. *Proceedings of the National Academy of Sciences*, *112*, 10321–10324.

Jo, Y., Poddar, S., Jeon, B., Shen, Q., Rose, C., & Neubig, G. (2018). Attentive Interaction Model: Modeling Changes in View in Argumentation. In M. Walker (Ed.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 103 – 116). Stroudsburg, PA: Association for Computational Linguistics.

Kahan, D. M., Peters, E., Wittlin, M., Slovic, P., & Ouellette, L. L. (2012). The polarizing impact of science literacy and numeracy on perceived climate change risks. *Nature Climate Change*, *2*, 732.

Klayman, J. (1995). Varieties of confirmation bias. *Psychology of Learning and Motivation – Advances in Research and Theory*, *32*, 285-418.

Lai, C. K., Marini, M., Lehr, S. A., Cerruti, C., Shin, J. E., Joy-Gaba, J. A., … Nosek, B. A. (2014). Reducing Implicit Racial Preferences: I. A Comparative Investigation of 17 Interventions. *Journal of Experimental Psychology: General*, *143*, 1765-1785.

Leucht, S., Cipriani, A., Spineli, L., Mavridis, D., Orey, D., Richter, F., … Davis, J. M. (2013). Comparative efficacy and tolerability of 15 antipsychotic drugs in schizophrenia: a multiple-treatments meta-analysis. *The Lancet*, *382*, 940.

Lewandowsky, S., Ginac, G. E., & Vaughan, S. (2013). The pivotal role of perceived scientific consensus in acceptance of science. *Nature Climate Change*, *3*, 399 – 404.

Liddell, T. M., & Kruschke, J. (2017). Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology*, *79*, 328–348.

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, *37*, 2098 – 2109.

McElreath, R. (2016). *Statistical rethinking* (1st ed.). Boca Raton : CRC Press/Taylor & Francis Group.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*, 175 – 220.

Nyhan, B., & Reifler, J. (2015). Does correcting myths about the flu vaccine work? an experimental evaluation of the effects of corrective information. *Vaccine*, *33*, 459 – 464.

Nyhan, B., Reifler, J., Richey, S., & Freed, G. L. (2014). Effective messages in vaccine promotion: A randomized trial. *Pediatrics*, *133*, 835 – 842.

Pew Research Center. (2014). Political Polarization in the American Public. Retrieved from people-press.org.

Priniski, J. H., & Horne, Z. (2018). Attitude Change on Reddit's Change My View. In T. T. Rogers, M. Rau,

X. Zhu, & C. W. Kalish (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 2276 – 2281). Austin, TX: Cognitive Science Society.

Strohminger, N., & Nichols, S. (2014). The essential moral self. *Cognition*, *131*, 159 – 171.

Tan, C., Niculae, V., Danescu-Niculescu-Mizil, C., & Lee, L. (2016). Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In J. Bourdeau, J. A. Hendler, & R. N. Nkambou (Eds.), *Proceedings of the 25th International Conference on World Wide Web* (pp. 613 – 624). New York, NY: Association for Computing Machinery.

Turetsky, K. M., & Sanderson, C. A. (2017). Comparing educational interventions: Correcting misperceived norms improves college students' mental health attitudes. *Journal of Applied Social Psychology*, *48*, 46 – 55.

Outcomes Speak Louder than Actions? Testing a Challenge to the Two-Process Model of Moral Judgment

Karolina Prochownik (karolina.prochownik@rub.de)
Faculty of Law, Ruhr University Bochum, Universitätsstrasse 150, 44801 Bochum, Germany
Faculty of Law and Administration, Jagiellonian University, Krakow, Poland

Fiery A. Cushman (cushman@fas.harvard.edu)
Department of Psychology, Harvard University, William James Hall 1480, 33 Kirkland St.
Cambridge, MA 02138, USA

Abstract

Curiously, people assign less punishment to a person who attempts and fails to harm somebody if their intended victim happens to suffer the harm for coincidental reasons. This "blame blocking" effect provides an important evidence in support of the two-process model of moral judgment (Cushman, 2008). Yet, recent proposals suggest that it might be due to an unintended interpretation of the dependent measure in cases of coincidental harm (Prochownik, 2017; also Malle, Guglielmo, & Monroe, 2014). If so, this would deprive the two-process model of an important source of empirical support. We report and discuss results that speak against this alternative account.

Keywords: blame blocking; two-process model; punishment; outcomes; actions; pragmatics

Introduction

Imagine that two runners compete in a championship race. One of the runners is a frequent winner, and so another racer decides to kill him and exclude him from the competition. He mistakenly believes that his rival is fatally allergic to poppy seeds, and so he sprinkles some on his rival's food at the banquet. The champion is not allergic to poppy seeds at all, however, but instead to hazelnuts. What's more, completely by coincidence, the chef happens to have served a hazelnut salad, and the champion dies as a result of consuming it.

The "blame blocking" phenomenon, first reported in the set of studies by Cushman (2008), is that people will tend to reduce blame and punishment assigned to the attempted harmdoer because of the coincidental harm caused by the salad. In other words, if the salad has no hazelnuts and the intended victim survives, the attempted harmdoer is blamed and punished more. This effect is notably large. Specifically, in the study based on the above story Cushman (2008) found that about half as many subjects assigned *no punishment* to the runner where no harm occurred compared with the case in which the rival coincidentally died (p. 374). That is, the coincidental death of the rival made participants twice as likely to let the runner off the hook.

One explanation of this puzzling effect posits two processes of moral judgment that render moral judgments separately on the basis of (1) causal responsibility for harm, or (2) a culpable mental state, such as intent to harm (Cushman, 2008). According to the model, then, when there was no causal input in the story (i.e., no coincidental harm occurs), the "mental state process" dominates and punishment judgments are therefore based on the evaluation of the agent's mental states alone. Because the relevant mental state was severe intentional harm, this tends to result in non-zero levels of punishment. On the other hand, when causal inputs are present (i.e., a coincidental harm occurs) but the runner himself is non-causal, the process of moral judgment predicated on causal responsibility competitively dominates (or "blocks") the evaluation of his mental states. The causal responsibility process assigns no punishment to the runner (who, of course, has no causal responsibility for the harm). Stated more generally, a two-process model of moral judgment can accommodate the pattern of results because it posits competition between a causal process seeking full exculpation (no punishment) and a mental state process seeking full inculpation (punishment) in cases of failed attempts to harm with independently caused harm, while the relative influence of the causal process is minimized in cases of pure failed attempts.

The two-process model is compatible with theories of moral judgments that identify intentional and causal evaluations as primary contributors to blame and punishment (e.g., Alicke, 2000; Alicke & Rose, 2012; Carlsmith & Darley, 2008; Darley & Shultz, 1990; Fincham & Jaspers, 1979; Shultz, Schleifer, & Altman, 1981; Shultz, Wright, & Schleifer, 1986; Weiner, 1995; Guglielmo, Monroe, & Malle, 2009; Malle, Guglielmo, & Monroe, 2014; Piaget, 1932/1965). It departs from most of these theories, however, in the assumption that causal and mental state evaluations proceed separately and compete during moral judgments of blame and punishment, rather than being combined and integrated in a single process.

In addition to the blame blocking phenomenon, some independent evidence provides support for the two-process model. Young, Cushman, Hauser, and Saxe (2007) found neurological signature of conflict for adult judgments of accidental harms in which intentional and causal evaluations point in different directions. Several studies show that punishment judgments are especially strongly influenced by the causal process in ordinary cases of harm (Martin & Cushman, 2015, 2016), and developmental evidence suggests that this pattern is a vestige of an early-emerging "causal" process of moral judgment augmented by a later-emerging

"mental state" process (Cushman, Sheketoff, Wharton, & Carey, 2013).

Here, we consider another explanation of the blame blocking effect—one that depends on assumptions about how people interpret the pragmatics of the dependent measure used to trigger this effect. Specifically, the question "How much prison time does [agent] deserve?" used by Cushman (2008) might be interpreted by participants differently across conditions: as implicitly referring to punishment *for behavior* (how much should the runner be punished for trying to kill his rival with poppy seeds) in the "no harm" condition, but as implicitly referring to punishment *for a harmful outcome* (how much should the runner be punished for the victim's death by the hazelnuts) in the coincidental harm condition (see also Prochownik, 2017). If the agent in two scenarios were evaluated against these very different standards in each case it would explain the blame blocking effect without appeal to two processes of moral judgment. We call this alternative "pragmatics account" because it relies on an assumption that people take a broad context into account when deciding what for to punish others (cf. Prochownik, 2017).[1]

In this paper we examined this alternative hypothesis by conducting two experiments. In Experiment 1, we manipulated the question about punishment to ensure that it is interpreted with a wide scope, encompassing not only what the agent caused (or did not), but also what he intended. Next, in Experiment 2, we used the original dependent measure that was previously used to elicit the blame blocking effect, and then asked participants a series of questions designed to clarify how they understood it.

Collectively, the results of these experiments suggest that unintended interpretations of the dependent measure are not sufficient to explain the full blame blocking effect.

Experiment 1

The goal of Experiment 1 was to test whether a more precise phrasing of the dependent measure would eliminate the previously observed blame blocking effect. In the baseline condition ("unspecified") we left the question identical to previous experiments by Cushman (2008): "In your opinion, how much prison time does *X* deserve?" In the novel condition ("specified") we modified the question so that it more clearly pointed at the agent's total set of behaviors as the target of punishment, thus diminishing the chance that it would be interpreted in terms of outcome alone (following in this respect Prochownik, 2017): "Suppose that X were apprehended by the police and put on trial. Given the complete set of behaviors and facts, in your opinion how much prison time does he deserve?"

The language that we used in the "specified" condition was borrowed from earlier research. In particular, Prochownik (2017) found that people with legal education tended to manifest the blame blocking effect only when the punishment question was unspecified, but the effect disappeared when it was specified, suggesting a key role for pragmatics in this group of respondents. However, Prochownik & Unterhuber (2018) did not replicate this finding in their comparative study including both lay people and legal experts. In Experiment 1 we use the same version of the "specified punishment question" as these researchers, but we focus exclusively on lay people in a well-powered study, and also examine it more systematically (across sixteen scenario contexts instead of just two or three as in these previous studies).

Methods

We tested 20 participants in each of 64 cells of a 2 (harm vs. no harm) x 2 (specified vs. unspecified) x 16 (scenario context) design, for a total sample of 1280. Participants were recruited on MTurk in the US. After consenting to participate in a short study for small compensation ($0.30), they filled an online Qualtrics survey comprised of one scenario, a punishment probe, and demographic questions (age, gender, nationality, exposure to moral philosophy, religiosity, etc.). Participants marked their answers on a scale with 11 anchored options: "None", "1 week", "1 month", "3 months", "6 months", "1 year", "2 years", "4 years", "8 years", "16 years", "32 years".[2]

The total set of 16 scenarios varied along several dimensions. Most notably, half of them involved physical harm (burning, cutting, stabbing, etc.) while the remaining half involved property harm (arson, defacement, etc.). The full text of all study scenarios is available online as Supplementary Materials:
https://osf.io/9w4ke/.

Results

As summarized in Figure 1, we observed the basic blame blocking effect in both the "specified" and "unspecified" conditions. Indeed, if anything, the blame blocking effect was slightly larger in the new "specified" condition. In order to analyze the data more fully we conducted a linear mixed effect analysis. First, we constructed a null model without fixed effects for harm or punishment question type, but including a random effect for scenario. We then found that this model was significantly improved by modelling the harm factor, $\chi^2(3) = 61.49$, $p < .001$. Next, we found that this "harm only" model was not significantly improved by modelling the punishment question type factor $\chi^2(4) = 1.17$, $p = .8826$, or

[1] The importance of pragmatic considerations for participants´ (re)interpretations of research stimuli has been also raised by some recent studies (e.g., Guglielmo & Malle, 2010; Samland & Waldmann, 2016; Wiegmann, Samland, & Waldmann, 2016; Hagan & Rozyman, 2017).

[2] The "unspecified" punishment question was taken from Cushman (2008): Experiment 4. However, the scale differed from the 9-points scale used by Cushman in his study as for the majority of scenario contexts we did not use attempted murders but attempts of less severe crimes (including bodily injuries and damages to property) for which we needed a greater range of less severe sentences. As a result, we could also examine if the previous findings replicate when a different scale of punishment ratings is used.

by modelling both this factor and its interaction with harm χ^2 (9) = 4.1, p = .9047. In summary, then, the best-fitting model included only harm as a factor. In other words, we observe a significant effect for the harm vs. no harm factor, but no significant effect for the specified vs. unspecified factor, or for its interaction with harm.

We next assessed whether there are significant differences between scenarios in the magnitude of the blame blocking effect that they induce by testing whether random intercepts (i.e., an interaction between scenario context and the effect of the "harm" variable) contribute significantly to the model. They do, χ^2 (2) = 20.9, p < .001, indicating meaningful variability between vignettes. We next tested whether the model was improved by adding a fixed effect for "physical" versus "property" harms, but it was not χ^2 (9) = 7.94, p = .54. The precise nature of the relevant differences between scenarios therefore remains an important topic for further research.

Discussion

Experiment 1 shows that the blame blocking effect is not diminished by an alternative phrasing of the dependent measure designed to clarify that punishment could apply to any aspect of an attempted harmdoer's conduct—including, most importantly, the attempted harm.

These results speak against the alternative interpretation of that effect in terms of the pragmatic constraint on the way ordinary people assign punishment, and instead support the two-process model of moral judgment.

However, one limitation to this experiment is that by asking participants to consider the entire event when making their punishment judgments, we cannot completely exclude the possibility that some participants interpreted the question as referring to the outcome alone. If so, it is still possible that people who interpreted the question as referring to the outcome were driving the blame blocking effect. To address this problem we conducted an additional experiment which faithfully replicated the original "runners study" by Cushman (2008) but differed in one important element: participants in the "Harm" condition were presented with an additional question about how they understood the question about punishment (i.e., what they thought the punishment was meant to be for).

Experiment 2

In this experiment we replicated Cushman`s Experiment 4 (2008) but we presented participants in the "Harm" condition with an additional question about how they understood the punishment question after they have responded to it. Specifically, we asked explicitly whether they understood the question "how much punishment does *X* deserve?" to refer to "punishment *for the actual harm*" (e.g., death of a runner) in the coincidental harm condition. Such an interpretation, which is consistent with the pragmatics account, would explain away the purported "blame blocking effect."

We offered participants two alternatives to this interpretation of the question: First, that "punishment" referred only to the attempted harm (e.g., the sprinkling of poppy seeds on a salad with intent to cause an allergic reaction) and, second, that it referred to *both* the attempted

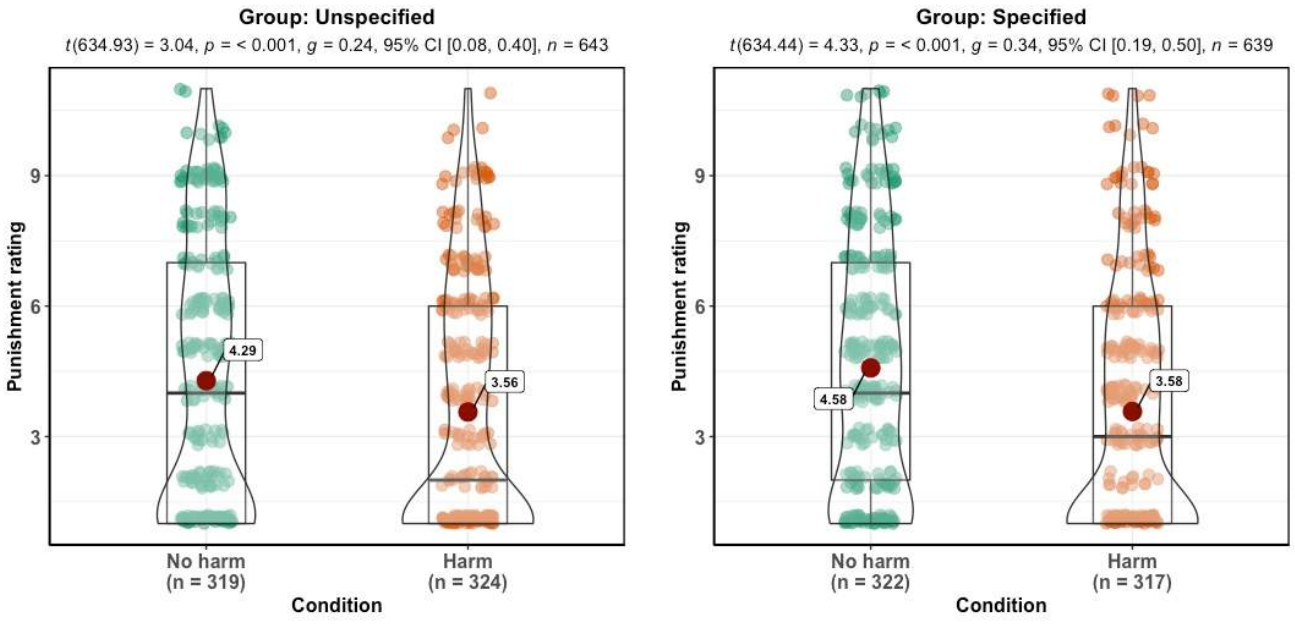


Figure 1: The blame blocking effect was replicated both in the "unspecified" condition, which directly matches the original demonstration (Cushman, 2008) and in the "specified" condition, which was designed to eliminate the alternative explanation of the effect in terms of pragmatics.

and actual harm. The blame blocking model is agnostic with respect to these alternatives—crucially, both of them entail sensitivity to the attempted harm, and thus the null prediction would be equal punishment across the no harm and coincidental harm conditions, both of which involve this attempted harm. The two-process model attempts to explain why participants who interpret the punishment question to *include* this key shared element—the attempted harm—would nevertheless be more likely to fully exonerate the attempted harmdoer in the coincidental harm case.

Study hypotheses, methods of analyses, sample size calculation and exclusion criteria were preregistered (the OSF preregistration document can be viewed at https://osf.io/pf574).

Methods

1007 complete responses were collected via TurkPrime in the US using Qualtrics anonymous link (we intended to recruit 500 participants per each of the study conditions). Participants were payed $0.50 for taking part in the survey.

Participants were asked to imagine that they are in a jury in a case of a defendant named Brown. In following, they were presented with a story of two runners named Brown and Smith competing in a championship race. One group of participants saw the variant of the story where Brown tries to kill Smith by sprinkling the poppy seeds on his food, but no harm results ("No Harm" condition). Another group of participants was presented with the story in which Smith dies because of the hazelnuts in the salad that he is served, completely independently of Brown´s actions ("Harm" condition). After reading the story all participants were asked: "How much prison time does Brown deserve?", and chose between the nine following options: "None", "6 months", "1 year", "2 years", "4 years", "8 years", "16 years", "32 years", "Life" (Cushman, 2008).

On the next page of the survey, participants in the "Harm" condition were presented with the following "Harm Understanding Question":

"On the last screen you were asked to decide how much prison time Brown deserved. Which of the following did you think was meant by that:

1. How much prison time for sprinkling poppy seeds on Smith's food?
2. How much prison time for the death of Smith?
3. How much prison time for both sprinkling poppy seeds on Smith´s food and the death of Smith?"[3]

Participants who chose the third option ("for both") were additionally asked two questions about punishment for the action alone and for the outcome alone to enable the researchers better understand their previous answers: "How much prison time does Brown deserve only for the death of Smith?" and "How much prison time does Brown deserve only for sprinkling poppy seeds on Smith´s food?". Answers to both questions were marked on the same 9-points scale as above.

Finally, all participants were asked two comprehension questions about the story they read: "Why did Brown sprinkle poppy seeds on Smith´s food?" (multi-choice question) and "Did Smith die as a result of Brown sprinkling poppy seeds on his salad?" (two-choice question).[4]

Participants were excluded from the analysis if they answered incorrectly to any of the two comprehension questions (i.e., if to the first question they provided any answer other than "Because he wanted to kill Smith" and/or if they answered "Yes" to the second question). This resulted in 840 responses included in the analysis (N_{Harm} = 420, N_{NoHarm} = 420).

Results

Percentages of different responses to the "Harm Understanding Question" (n = 420, 100%) were as follows**:** 56.4% (n = 237) respondents understood the punishment question as being for sprinkling poppy seeds on Smith´s food, 19.3% (n = 81) as being for the death of Smith, and 24.3% (n = 102) as being for both sprinkling poppy seeds on Smith´s food and the death of Smith.

Consistent with our preregistered plan, and following the key analysis by Cushman (2008), we recoded the responses to the main punishment question to a binary variable with the following values: "No punishment" (all "None" responses) and "Any punishment" (all the responses assigning some punishment from "6 months" to "Life"). Subsequently, to test the main hypothesis we performed two chi-square tests comparing the frequencies of "No punishment" vs. "Any punishment" responses across two study conditions "Harm" and "No Harm": (1) a chi-square test with the overall sample (analysis repeating Cushman, 2008, Experiment 4), and (2) a chi-square test excluding people in the "Harm" condition who in the "Harm Understanding Question" replied that they understood the punishment question as referring to the outcome alone (i.e., the death of Smith).

Overall, 35% people assigned "No punishment" in the "Harm" condition, while in the "No Harm" condition barely half as many (18%) of people did so.[5] The difference was statistically significant, χ^2 (1, N = 840) 31.740, p < .001. Critically, this result held even after excluding participants who indicated that they thought the punishment question referred to punishment for the outcome only: among the remaining participants, 28% assigned "No punishment" in the "Harm" condition comparing to 18% in the "No Harm" condition. The difference was statistically significant, χ^2 (1, n = 759) 11.763, p = .001 (Figure 2).

[3] This and two following questions were omitted in the "No Harm" condition as no death of Smith resulted in this story.

[4] In the first multi-choice comprehension question participants could choose from the following responses: "Because he thought the poppy seeds would make Smith sick for a couple of days", "Because he thought Smith liked poppy seeds"; "Because he wanted to kill Smith", "Because he wanted Smith to go to the bathroom". In the second question they could choose between two options: "Yes" or "No".

[5] Note that in Cushman (2008) the numbers were very similar, with 34.5% participants in the "Harm" and 19.5% in the "No Harm" case deciding not to punish Brown at all (p. 374).

In addition to the main analyses reported above, we also assessed whether the result is driven by remaining participants who understood the punishment question to refer to the attempted harm only, or by those who understood it to refer to both the attempted harm and the outcome. We found that the effect was maintained among those who understood the question to refer to both the attempted harm and the outcome: 43% people decided not to punish in this group comparing to 18% in the no harm group, χ^2 (1, n = 522) 29.801, p < .001. It was not significant, however, among those who understood the question to refer to the attempted harm only; among this group, 22% people assigned "No punishment" in the "Harm" condition, comparing to 18% in the "No Harm" condition, χ^2 (1, n = 657) 1.620, p = .203.[6]

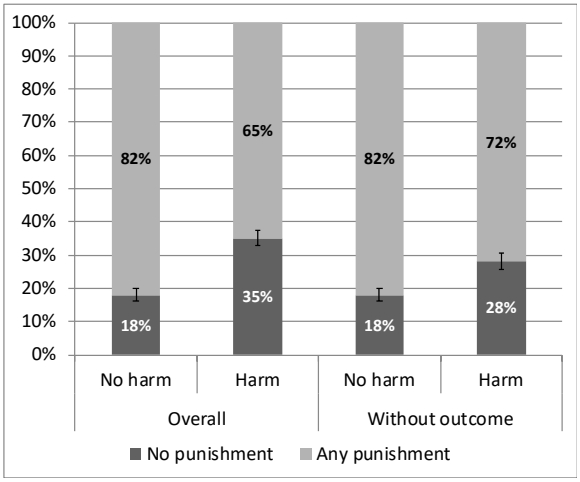


Figure 2: Percentage of punishment responses in different study conditions: overall sample and sample without people who referred to the outcome alone (error bars indicate standard error).

Discussion

Experiment 2 demonstrated that the blame blocking effect persists even after excluding participants who interpreted the punishment question as referring to the outcome alone. This suggests that the "pragmatics account" is insufficient to fully explain the effect.

Notably, however, the effect is driven by participants who say that they interpreted the punishment question to refer to "both" the attempted harm (sprinkling poppy seeds) and the coincidental harm (death by hazelnuts). It is weak, and perhaps entirely absent, among participants who say they interpreted the punishment question instead to refer exclusively to the attempted harm.

On the one hand, this data is consistent with a natural interpretation of the two-process model, according to which some attention to the harm (in the coincidental harm case) is necessary to produce the competitive interaction between the causal process and the mental state process. After all, according to the two-process model, it is precisely the attention paid to (the absence of) causal responsibility for the coincidental harm that competitively blocks assessment of the culpable mental state of the attempted harmdoer in the coincidental harm case.

On the other hand, this data is also consistent with an alternative explanation that we have not yet considered. A variant of this alternative is proposed by Malle, Guglielmo, and Monroe (2014)[7], who argue that blame in the coincidental harm case is the *average* of a high level of blame for the attempted harm and a low level of blame for the coincidental harm, whereas the blame in the no harm case is simply the high level for the attempted harm. In other words, when people interpret the punishment question as "both" about the attempted harm and the coincidental harm, they may therefore assign amount intermediate between these two values.[8]

We are in a good position to evaluate this alternative by analyzing an additional element of our data. Recall that, among people who said they interpreted punishment to refer to "both" the harm and the attempt, we then asked them to assign a specific amount of punishment to just the harm (presumably zero, as the harm was coincidental), and a specific amount of punishment to just the attempt. Thus, we can ask whether the *total* amount of punishment assigned was, on average, lower than the amount of punishment assigned to the *attempt alone*. This would be necessarily true on Malle and colleagues' hypothesis, since they assume that the total amount of punishment will be intermediate between the amount of punishment assigned to each of the two elements individually. Contrary to this prediction, however,

[6] We focused on the binarized results because these were the key analyses to report the blame blocking effect in the original study (Cushman, 2008). However, for the sake of transparency we also report the analyses with the full range of responses. Following the original study we ran Mann Whitney Ranked Sums tests for three groups of participants: (1) in the overall sample participants assigned more punishment in the "No Harm" case than in the "Harm" case (Mdn_H = 3, Mdn_{NH} = 4). The difference was statistically significant, Z(840) = 3.372, p = .001; (2) in the analysis without people who referred to the outcome alone, the difference was marginally statistically significant (Mdn_H = 4, Mdn_{NH} = 4), Z(759) = 1.897, p = .058, (3) in the analysis without people who referred to the outcome alone and both outcome and action, the difference was not statistically significant (Mdn_H = 4, Mdn_{NH} = 4), Z(657) = 0.565, p = .572. Note that for these supplementary analyses Cushman (2008) reported marginally significant results with p = .11 (cf. p. 374).

[7] Malle et al. (2014) apply this reasoning to judgments of blame, but they point out that there exists a similar pattern for judgments of criminal liability (p. 169). Since in the paper we focus on judgments of punishment, we consider their proposal in relation to this class of moral judgments.

[8] This proposal is similar to the account examined above as it assumes that people in different conditions may be judging the perpetrator for two different events. However, while the former would perceive the blame blocking effect as a result of people interpreting the dependent measure in terms of outcome alone, Malle and colleagues` account would explain it in terms of people judging the perpetrator for the conjunction of attempt and outcome.

the mean punishment for the "composite" event (M = 3.57, SD = 3) was not lower than the mean punishment for the attempt alone (M = 3.29, SD = 2.73).[9] Similarly, 43% (n = 44) of these participants assigned "no punishment" to the composite event, while 40% (n = 41) assigned no punishment to the attempt alone (consistent with the principle "no harm, no foul"). This suggests that people did not, for instance, feel that the attempt was punishable and yet assign no punishment for the *composite* event because one cannot be punished at all for something they did do *and something they did not*.

Collectively, these data further speak against pragmatic interpretations of the blame blocking effect. Even among people who say that they judged the coincidental harm case in part by assigning punishment to the attempted harm—and even when asked to make a punishment judgment strictly about that attempted harm—the blame blocking effect persists.

General Discussion

Recent proposals have advanced a potential alternative explanation of the blame blocking effect that does not invoke two independent processes of moral evaluation. According to the "pragmatics alternative" people could have interpreted the pragmatics of punishment question differently across versions of the story with and without harm that were used to trigger this effect in studies by Cushman (2008). In order to address this alternative we conducted two experiments. In Experiment 1 we used two different versions of the punishment question in order to test if the blame blocking would remain after we specify the question as more clearly referring to the total set of the agent´s behaviors as the target of punishment (developing previous research by Prochownik, 2017 and Prochownik & Unterhuber, 2018). The results indicated that the blame blocking effect occurs regardless of the phrasing of the dependent measure. However, a potential limitation of this study was that it did not completely exclude the possibility that some participants could have still interpreted the punishment question (even when specified) as referring to the outcome alone. To address this problem, we conducted another experiment. In Experiment 2 we replicated one of the original studies by Cushman (2008) with one modification: after assigning a specific amount of punishment to the defendant, participants in the "Harm" condition indicated what they thought the punishment was meant to be for (for the attempt, for the outcome or for both the attempt and the outcome). The blame blocking effect was present in the overall sample and also after excluding participants who indicated they thought the punishment was for the outcome alone. Taken together, these two experiments suggest that the blame blocking effect cannot be accounted for in terms of people´s presumed tendency to interpret the punishment question in terms of outcomes rather than actions.

In addition, Experiment 2 speaks against a slightly different proposal by Malle, Guglielmo, and Monroe (2014). According to these researchers, people´s judgements are for the agent´s attempt alone in the "No Harm" case, while they result from the average of the punishment for the attempt and the outcome in the "Harm" case. Yet, in contrast to this prediction, our results suggest that people judge the "composite" event of the attempt and outcome almost the same as they judge the attempt alone. Therefore, the blame blocking effect is not likely to occur due to averaging.

Experiment 1 also recommends some further developments of the two-process model itself. In its original formulation, the model remained open regarding what type of consequences trigger the causal process of moral evaluation, and can eventually lead to the blame blocking effect. Scenarios used by Cushman (2008) featured harms to humans and presented coincidental harms that were roughly the same as the harms intended and attempted by the perpetrators (e.g., the same victim dies, and by similar means to those originally intended). The presence of the blame blocking effect across different scenario contexts in our first experiment suggests that this effect is robust across different types of harms including both severe bodily injuries and gross harms to property, as well as coincidental harms that are somewhat different than originally intended. This suggests a modification to the two-process model such that the blame blocking effect can be triggered by a wide variety of harmful events. However, more research in this direction would help to delineate the scope of the blame blocking phenomenon and the specific conditions under which it occurs.

Finally, although, our experiments suggest that the blame blocking effect cannot be accounted for simply in terms of people interpreting the dependent measure as referring to the outcomes and not the actions (thus outcomes do not speak louder than actions!), future research must test additional possible alternative explanations of blame blocking. Two stand out. First, it might be that people diminish the punishment in the "Harm" case comparing to the "No Harm" case because they think the harmful outcome would have occurred regardless of the agent's attempt to harm (e.g., because Smith would have been killed by the chef anyway people may perceive Brown´s attempted homicide as redundant and release him from responsibility). Second, the "Harm" case is more complex and contains more information than the "No Harm" case that may distract participants (e.g., that Smith ends up being killed by the chef may be an extra element drawing people´s attention away from Brown´s attempted homicide). Finally, in addition to testing these alternatives, the two-process model would benefit from more thorough research on how exactly the two processes of moral analyses operate and interact in everyday moral decision making.

[9] Because the mean might not be well suited to the ordinal scale like the one we used, we also calculated medians and modes for the two punishment questions. The results did not differ, as the medians (2="6 months") and modes (1="No punishment") were the same for both the main punishment rating and the punishment for the attempt alone (n = 102).

Acknowledgments

This work was supported by the research project No. 2014/13/N/HS5/01137 funded by the National Science Centre, Poland.

We would like to thank Alex Wiegmann and four anonymous reviewers for very helpful comments on the previous version of the manuscript.

References

Alicke, M. (2000). Culpable control and the psychology of blame. *Psychological Bulletin*, *126*(4), 556-574.

Alicke, M. D., & Rose, D. (2012). Culpable control and causal deviance. *Social and Personality Psychology Compass, 6*(10), 723-735.

Carlsmith, K., & Darley, J. M. (2008). Psychological aspects of retributive justice. *Advances in Experimental Social Psychology*, 40, 193-235.

Cushman, F. (2008). Crime and punishment: distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, *108*(2), 353-380.

Cushman, F., Sheketoff, R., Wharton, S., & Carey, S. (2013). The development of intent-based moral judgment. *Cognition*, *127*(1), 6-21.

Darley, J. M., & Shultz, T. R. (1990). Moral rules - their content and acquisition. *Annual Review of Psychology, 41*, 525-556.

Fincham, F. D., & Jaspers, J. (1979). Attribution of responsibility to the self and other in children and adults. *Journal of Personality and Social Psychology, 37*(9), 1589–1602.

Guglielmo, S., Monroe, A. E., & Malle, B. F. (2009). At the heart of morality lies folk psychology. *Inquiry, 52*(5), 449-466.

Guglielmo, S., & Malle, B. F. (2010). Can unintended side effects be intentional? Resolving a controversy over intentionality and morality. *Personality and social psychology bulletin*, *36*(12), 1635-1647.

Hagan, J. P., & Royzman, E. (2017). The shadow and the tree: inference and transformation of cognitive content in psychology of moral judgment. In J. F. Bonnefon & B. Trémolière (Eds.), *Moral inferences. Current issues in thinking and reasoning*. Oxford: Routledge.

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. *Psychological Inquiry*, *25*(2)*,* 147-186.

Martin, J. W., & Cushman, F. (2015). To punish or to leave: Distinct cognitive processes underlie partner control and partner choice behaviors. *PloS ONE*, *10*(4).

Martin, J. W., & Cushman, F. (2016). Why we forgive what can't be controlled. *Cognition, 147*, 133-143.

Piaget, J. (1965). *The moral judgment of the child*. New York, NY: Free Press. (Original work published 1932)

Prochownik, K. (2017). Do people with a legal background dually process? The role of causation, intentionality and pragmatic linguistic considerations in judgments of criminal responsibility. In J. Stelmach, B. Brożek & Ł. Kurek (Eds.), *The province of jurisprudence naturalized*. Warsaw: Wolters Kluwer.

Prochownik, K., & Unterhuber, M. (2018). Does the blame blocking effect for assignments of punishment generalize to legal experts? In T.T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.), *Proceedings of the 40th annual conference of cognitive science society* (pp. 2285-2290). Austin, TX: Cognitive Science Society.

Samland, J., & Waldmann, M. R. (2016). How prescriptive norms influence causal inferences. *Cognition*, *156*, 164-176.

Shultz, T. R., Schleifer, M., & Altman, I. (1981). Judgments of causation, responsibility, and punishment in cases of harm-doing. *Canadian Journal of Behavioural Science, 13*(3), 238.

Shultz, T. R., Wright, K., & Schleifer, M. (1986). Assignment of moral responsibility and punishment. *Child Development*, *57*(1), 177-184.

Weiner, B. (1995). *Judgments of responsibility: A foundation for a theory of social conduct*. New York, NY: Guilford Press.

Wiegmann, A., Samland, J., & Waldmann, M. R. (2016). Lying despite telling the truth. *Cognition*, *150*, 37-42.

Young, L., Cushman, F., Hauser, M., & Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. *Proceedings of the National Academy of Sciences, 104*(20), 8235.

A Piecemeal Processing Strategy Model for Causal-Based Categorization

Guillermo Puebla (guillermo.puebla@ed.ac.uk)
Department of Psychology, School of PPLS, The University of Edinburgh
Edinburgh EH8 9JZ, United Kingdom

Sergio E. Chaigneau (sergio.chaigneau@uai.cl)
School of Psychology, Universidad Adolfo Ibáñez
Av. Presidente Errázuriz 3328, Las Condes, Santiago, Chile

Abstract

Over the last 20 years, causal-model theory has produced much knowledge about causal-based categorization. However, persistent violations to the normative causal-model theory are prevalent. In particular, violations to the Markov condition have been repeatedly found. These violations have received different explanations. Here, we develop a model that starts from generally accepted cognitive phenomena (e.g., processing limitations, the relevance of inference in cognitive processing) and assumes that people are not fully causal nor fully associative when performing causal-based categorization, offering a new explanation for Markov violations.

Keywords: causal-based categorization; causal-model theory; causal inference; Markov condition

Introduction

Causal-model theory has taught us much about causal cognition (see Rehder, 2017; Sloman & Lagnado, 2015). However, problematic violations of the theory's predictions persist. The most important violation is that of the Markov condition. This condition states that when the state of a variable's immediate causes is known, then that variable is rendered conditionally independent of all its non-descendants (Pearl, 2000). Illustrative examples of violations are found in Rehder and Burnett (2005), and in Puebla and Chaigneau (2014). There, participants needing to infer the state of an unknown variable used information about other properties, even if those properties were conditionally independent from the unknown variable.

Some authors explain these violations by arguing that they are only apparent, because subjects do not necessarily use the same causal model specified by experimenters (Park & Sloman, 2013). Other authors argue that people may resort to associationist thinking and interpret directed causal links as bidirectional associations (Rehder, 2014). In contrast, here we hypothesize that violations occur because people may combine a partial understanding of causality with underlying similarity-based processing. In what follows, we present a process model of causal categorization, use it to make specific predictions about properties' conceptual weights, and fit it to empirical data.

Informed probabilities influence property-weights

Prior research consistently reports that when people need to infer a central property for categorization (i.e., because they lack information about the state of that property), properties that are causally related to the absent property increase their relevance for categorization proportionately to their informativeness about the missing property (Chaigneau, Barsalou, & Sloman, 2004; Puebla & Chaigneau, 2014; Rehder & Kim, 2009). In the current work we extend these findings to conditions in which the central property's state is explicitly known.

We hypothesize that even if a central property's state is made explicit, there may still remain some uncertainty regarding the property's true state. Thus, other causally related properties may acquire their weight depending on their contribution to decreasing that uncertainty. This idea is discussed in Rehder and Burnett (2005), and preliminary evidence for it can be found in Chaigneau et al. (2004, Exp. 7). In causal-model research, information about a property's inferential contribution is generally provided in the form of probabilities of effects given causes (i.e., p(effect|cause)). Consequently, we assume that when cues indicate that a given property is central, other causally linked conceptual properties acquire their weight as a function of how informative they are of the central property. In particular, in our Exp. 1 we used a causal chain model (A→B→C; with an additional D property which was not causally linked to other variables), and told participants that property C was central (i.e., it gave the category its name), with the expectation that its directly linked properties (i.e., B) would acquire their weight proportional to their p(effect|cause), and that its indirectly linked properties (i.e., A) would be weighted proportional to their probabilistic contribution to the central property's direct causes (i.e., B). Note here that we are assuming that people are intransitive when using causal models to categorize (Johnson & Ahn, 2015).

Making the last property in a causal chain the central property is representative of many categories that are defined by their functions. For artifacts (Carrara & Mingardo, 2013; Chaigneau et al., 2004) and for functionally conceptualized natural kinds (e.g., Barsalou, Sloman, & Chaigneau, 2005; Lombrozo & Rehder, 2012), the goals that they achieve in their normal settings are central for their classification (e.g., an artificial heart is believed to belong to the heart category depending on it being able to pump blood to a greater extent than on it using any particular physical mechanism to achieve that goal).

Cognitive limitations

In typical causal classification experiments participants need to integrate several pieces of information, e.g., information

about the direct causal links and their associated probabilities (p(effect|cause)), the indirect causal links (two-way relations that are mediated by other properties), and also the particulars of the materials provided. Researchers generally assume that people are able to integrate all this information. In fact, as discussed in (Rehder, 2003a), the causal-model theory assumes that people classify entities as category members to the extent that the entity's distribution of properties would be expected from the category's ideal causal model.

In contrast, in our model we hypothesize that people simplify their task by analyzing information in a piecemeal fashion (thus, we call it the Piecemeal Strategy Model or PSM). In particular, we assume that they only evaluate pairs of directly connected properties, and that unconnected properties are considered in isolation (e.g., in the causal chain model A→B→C, with D as an isolated property, subjects would separately evaluate A→B, B→C and D). Regarding the type of computation subjects perform, we assume that they consider each directly connected pair (and each isolated property) in the ideal model presented to them, as a separate prototype with which to compare the particular instances they need to judge. To implement these ideas, we used Nosofsky (1992) Multiplicative Prototype model (MPM). This implies computing a distance, as given by,

$$\delta_{XY} = \sum_{i=X}^{Y} \left(\frac{p_i}{p_X + p_Y} \right) |x_i - M_i| \quad (1)$$

where X and Y are two directly connected properties in the causal model, p_i is the inferential contribution of a property, x_i corresponds to the state of the ith property in the currently considered instance, and M_i corresponds to the ideal state of the ith property in the causal model (i.e., the prototype). Note that the denominator inside the parenthesis allows Eq. 1 to comply with the MPM requirement that the weights (p_i) in the distance computation all add to one. For isolated properties (D in our scenarios), the corresponding distance is defined to be,

$$\delta_D = p_D |x_D - M_D| \quad (2)$$

where p_D is a free parameter estimated from the data, reflecting the inferential weight of the isolated property ($0 \leq p_D \leq 1$), x_D is the state of the D property in the currently considered instance, and M_D corresponds to the ideal state of the D property in the causal model (i.e., the prototype).

Distances cannot be considered by themselves, because they are linear. Similarity, in general, behaves like a generalization gradient (Shepard, 1987). For this reason, distances in Eqs. 1 and 2 need to be transformed into similarities by,

$$S_{XY} = e^{-b(\delta_{XY})} \quad (3)$$

where s_{XY} is similarity, and b is a sensibility parameter that determines the rate at which similarity falls with distance. In our model fitting, we fixed b = 1 (i.e., b was not estimated from the data). To compute the similarity s_D for the isolated D property, δ_{XY} is substituted by δ_D in Eq. 3.

Finally, we assume that the similarities from all the partial models under consideration are averaged to obtain an estimate of the overall similarity of the instance being judged relative to the prototype (i.e., the received causal model) by,

$$S_o = \frac{1}{n} \sum_{i=1}^{n} s_i \quad (4)$$

where so is the overall similarity of the instance being judged, n is the total number of separate pieces of information being considered (A→B, B→C, D), and s_i is the similarity according to Eq. 3. Because the PSM implies considering some properties twice (property B in the causal chain model), for modeling purposes we introduced an adjustment to pi simply by dividing it by 2 to reflect that those properties were being taken into account twice.

In summary, we propose that pairs of features that are causally related (and any features that are causally unrelated) are treated as features in separate multiplicative similarity prototype models, with classification ratings being a function of the averaged similarity of those feature pairs to their corresponding prototypes. A closely related model was proposed and tested by (Rehder, 2003a), but he concluded that the model failed to account for the data. In the Discussion section we will consider possible explanations for why our results suggest a different conclusion.

Experiments

Participants were trained on a causal model representing a given category, until they were able to answer correctly a set of 9 conditional and counterfactual questions. They then received the set of all possible combinations of present and absent properties involved in the causal model and were asked to rate how representative each combination was of the trained category.

Ratings were analyzed using the regression method (Rehder & Hastie, 2001). In this method, participants provide category membership ratings for all possible combinations of m properties in two possible states (present or absent), producing a total of 2^m combinations. For each combination, subjects provided a categorization rating on a 1 to 7 scale. When present and absent properties are coded respectively as 1 and -1 (i.e., effect coding), these values can be entered into individualized regression equations to predict a participant's categorization ratings. Furthermore, 2-way and higher-order interaction terms can be computed by entering the product of the corresponding property coded values as predictors into the equations. The corresponding regression coefficients can then be used as individual data points reflecting, across participants, the contribution of each predictor variable to the ratings.

Subjects were randomly assigned to one of two between-subjects conditions (domain: living things, artifacts) and provided data for two within-subjects conditions (information:

complete, incomplete). In the complete information condition, subjects received descriptions containing information about all properties (A, B, C, and D). In the incomplete information condition, subjects received descriptions lacking information about property C. Because prior research suggests that, in the context of causal classification the incomplete information condition promotes using other properties to infer the state of the unknown property (e.g., Puebla & Chaigneau, 2014), this design allowed us to compare conceptual properties' regression weights across the within-subjects condition. An increase in regression weights in the incomplete information relative to the complete information condition, would show that participants used a given property to infer the state of the unknown property C.

Predictions

The PSM makes the following predictions. Due to the piecemeal strategy, we predicted higher regression coefficients for directly connected properties interaction terms than for not directly connected properties. Furthermore, Eq. 3 predicts the type of interaction that we will find. People will prefer instances where properties X and Y are both in the same state as in the received model (e.g., X = 1, Y = 1), and any deviation (e.g., X = 1, Y = -1) will produce a large decrease in similarity (due to the b parameter). Note that all this means small interaction coefficients (i.e., smaller than main effect coefficients). This contrasts with predictions from the causal-model theory, where people are predicted to produce large interaction terms that are as large as main effects.

The PSM predicts that independent properties in our causal models will not interact. This is the same result that the causal-model theory would lead us to expect (i.e., properties A and C in the causal chain are independent conditional on the state of B). However, the PSM predicts this pattern of interactions, not because people conform to conditional independence principles, but because of the piecemeal simplification strategy. Thus, we expect our data from the complete information condition to only mimic adherence to the causal Markov condition. This should become evident in participants' performance in the incomplete information condition. When comparing regression weights across the information factor, the lack of information about the central C property should produce an increase in the regression weights of the independent properties (A and D in the causal chain) due to those properties being associated to the unknown property C. This is a violation of the Markov condition because only direct causes are normatively relevant to predict the state of the unknown property C. Thus, we predict an apparent adherence to Markov in the complete information condition, and a failure to adhere in the incomplete information condition.

Regarding the main effects, the PSM predicts that properties' conceptual weights will follow their inferentially derived weights (p_i). For the chain model in Exp. 1, we predict that regression coefficients for C will be greater than the average of A and B; D will be smaller than the average of A, B and C; and A will not be different from B.

Experiment 1

Design and Participants Exp. 1 followed a mixed factorial 2 (domain: living things, artifacts) x 2 (information: complete, incomplete) design, with the last being a within-subjects factor. Property D served as an inbuilt control condition for each subject and provided a baseline regression coefficient to which properties in the causal model could be compared. Also, D's interaction with other properties (AD, BD and CD) also provided a baseline for interaction terms' regression coefficients. Subjects (N = 66) were Adolfo Ibáñez University undergraduates (N = 41, males = 16) who participated for course credit, and undergraduate volunteers from other local universities (N = 25, males = 7).

Materials and Procedures The materials were verbal and graphical descriptions of two categories characterized by a chain causal structure. In the living things condition, materials described the structure of a fictional biological cell. In the artifacts condition, materials described the structure of a fictional particle accelerator. Stimuli were presented on screen by means of a locally programed software.

In the learning phase, participants were trained in the causal chain graph. Subjects learned that causes produced their effects with a 0.75 probability. Regarding property D, participants were informed that it occurred in category members with a probability of 0.75. Thus, property D was predictive of the category, but not causally related to the other properties. By keeping property D's probability equal to the conditional probabilities for the other properties, we kept everything other than belonging or not to the causal model constant for property D as compared to properties A, B and C. Importantly, subjects learned that property C gave the category its name (i.e., C was the central property).

In the classification phase, subjects had the causal graph in full view. In the complete information condition, participants received descriptions containing information about all properties either present or absent (16 combinations). In the incomplete information condition, participants received descriptions which lacked information about the state of the central property C (8 combinations). In total, participants classified 24 descriptions, presented in random order. For each description, subjects had to respond whether it was or not a member of the focal category using a 6-point rating scale.

Results Effect coding variables representing 10 variables per subject (4 main factors and 6 interactions, see Fig. 1) were entered as predictors in individualized regression equations with rating as dependent variable. The resulting individualized regression coefficients were submitted to a mixed 2 (domain: living things, artifacts) $\times$ 10 (coefficients) mixed ANOVA. The mixed ANOVA showed there was no effect of the domain factor ($F(1, 64)$=1.37, MS_e=.04, p=.25, ηp^2=.02, power=.21) and it did not interact with the coefficients factor ($F(9, 576)$=1.02, MS_e=.16, p=.39, ηp^2=.02, power=.30). Consequently, we collapsed this factor.

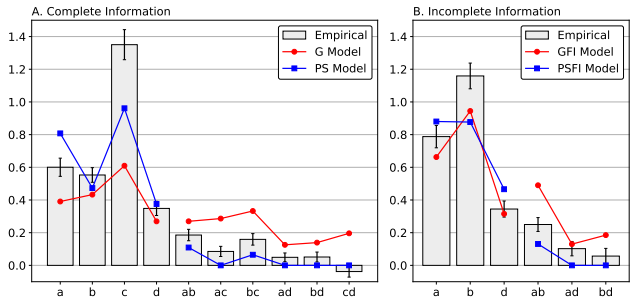


Figure 1: Regression weights for individual features and interactions of Exp. 1. Fits for the PSM (blue) and the GM (red) are superimposed on the data. Error bars are standard errors.

In accordance with our predictions for the main effects, orthogonal planned comparisons showed that the average coefficient for C was significantly greater than the average of coefficients for A and B ($F(1, 65)$=40.19, MS_e=.98, p<.001, ηp^2=.38, power>.99); the average coefficient for D was significantly lower than the average of coefficients for A, B and C ($F(1, 65)$=74.87, MS_e=.21, p<.001, ηp^2=.54, power>.99); and there was no difference between coefficients for A and B (F<1). This pattern of results for main effects suggests that subjects did indeed take information about p(effect|cause) as cue to properties inferential value, and that being in a causal structure increases inferential value beyond that of probabilistically related variables (property D).

Regarding the interaction coefficients, our predictions for the causal chain model were that the average of the AC interaction coefficients would be lower than the average of the AB and BC interaction coefficients (A and C are not directly connected in the causal graph); and that the average of the AB and BC interaction coefficients (directly connected properties) would be greater than the average of the AD, BD, and CD coefficients (i.e., our baseline conditions). As predicted, two non-orthogonal planned comparisons showed that the AC interaction was significantly smaller than the average of the AB and BC interactions ($F(1, 65)$=4.8, MS_e=.10, p=.03, ηp^2=.07, power=.58), and that the average of AB and BC interactions was significantly greater than the average of the AD, BD, and CD interactions ($F(1, 65)$=30.1, MS_e=1.81, p<.001, ηp^2=.32, power>.99).

Note that the low AC coefficient (which in fact was not significantly different from the interactions found for AD, BD and CD; $F(1, 65)$=3.8, MS_e=.07, p=.06, ηp^2=.06, power=.48), could be interpreted as participants complying with the Markov condition. However, analysis of the incomplete information condition reveals a different story. Under this condition, participants did not comply with Markov, using information about the state of property A (the screened-off property) and of property B (C's direct cause) to make inferences about the state of the missing C central property. Paired samples t tests revealed coefficients for properties A and B increased significantly when comparing the complete information condition with the incomplete information condition (respectively, complete information mean=0.60, incomplete information mean=0.79; $t(65)$=3.09, p=.003; complete information mean=0.55, incomplete information mean=1.16; $t(65)$=7.43, p<.001). In contrast, property D did not show evidence of being used to perform inferences about the state of property C (complete information mean=0.3485, incomplete information mean=0.3447; $t(65)$=.07, p=.94). Thus, data supported our hypothesis that subjects' performance in the complete information condition would mimic adherence to Markov.

Model fitting We fit the PSM to the classification ratings of Exp. 1. For comparison, we also fit the generative model (GM) of causal-based categorization (Rehder, 2003a; Rehder & Kim, 2009) (see Fig. 1). In the GM representation, a category k establishes a set of causal mechanisms. Each mechanism relates a feature j with its parent i operating with probability m_{ij} when i is present. Other background causes of j operate collectively with probability b_j. When j's parents operate independently, j's parents and the background causes produce j in members of category k conditional on the state of j's parents with probability,

$$p_k(f_i \mid Pa_k(f_j)) = 1 - (1 - b_j) \prod_{f_i \in Pa_k(f_i)} (1 - m_{ij})^{ind(f_i)} \quad (5)$$

where $ind(i)$ is an indicator variable that evaluates to 1 when i is present and 0 otherwise. The model assumes that root causes are independent of one another and the probability of each is represented with its own parameter c_j. The GM predicts that categorization judgments are a monotonic function of the joint distribution associated with the category's causal model,

$$p_k(f_{k,i}, \ldots, f_{k,N}) = \prod_{j=1 \ldots N} p_k(f_i \mid p_{a_k}(f_i)) \quad (6)$$

Participants ratings were predicted as follows:

$$rating^{PSM}(o_i) = s_k(o_i; p_A, p_B, p_C, p_D)/\beta$$
$$rating^{GM}(o_i) = 6 p_k(o_i; c_A, b_B, b_C, b_D, m_{AB}, m_{BC})^{\gamma}$$

where β and γ are free parameters. We fit both models by searching for the parameter values that minimized the squared difference between the predicted ratings and the empirical ones. In the complete information condition both models achieved a high correlation with the ratings: r_{PSM} = .85, r_{GM} = .90. We used the Akaikes information criterion (AIC[1]) to compare the degree of fit of both models controlling for the different number of parameters. The bigger AIC for the GM (15.2) in comparison to the PSM (12.1) indicates that, in fact, the PSM provides a slightly better characterization of the data

[1] AIC = $ln(SEE/n) + 2(p+1)$ where SSE is the sum of squared error for a participant, n is number of data points fit, and p is the model's number of parameters.

in this condition. The best-fitting parameters for the PSM were: $p_A = 0.325$, $p_B = 0.243$, $p_C = 0.775$, $p_D = 0.238$, $\beta = 0.094$ and for the GM: $c_A = 0.871$, $b_B = 0.802$, $b_C = 0.951$, $b_D = 0.765$, $m_{AB} = 0.558$, $m_{BC} = 0.327$, $\gamma = 0.565$. Note that while both models achieve a similar level of fit to the data, the GM achieves this by assigning values to the causal relation parameters lower than participants were taught during training (0.75).

In the incomplete information condition (see Fig. 2), we adjusted both the PSM and the GM to take into account the unknown state of C. We did this by inferring the probability of C being present given the state of its parent B using the GM equations: $p(E = 1 \mid C = 1) = 1 - (1 - m_{CE})(1 - b_E)$ and $p(E = 1 \mid C = 0) = b_E$. We treated this probability as the state of C and then proceeded as before for both models. In this condition the models achieved a high correlation with the ratings: $r_{PSM} = .89$, $r_{GM} = .92$. Again, we obtained a bigger AIC for the GM (15.0) in comparison to the PSM (12.7). The best-fitting parameters for the PSM were: $p_A = 0.347$, $p_B = 0.297$, $p_D = 0.264$, $b_C = 0.463$, $m_{BC} = 0.392$, $\beta = 0.083$ and for the GM: $c_A = 0.865$, $b_B = 0.884$, $b_C = 0.933$, $b_D = 0.687$, $m_{AB} = 0.552$, $m_{BC} = 0.617$, $\gamma = 0.505$. Note that the causal relation parameter for the relation between B and C was higher in this condition.

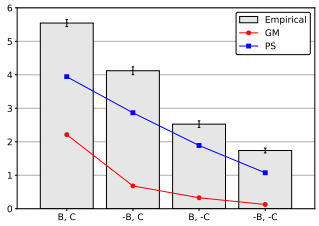


Figure 2: Average ratings for objects with different combinations of states (present or absent) for features B and C in Exp. 1, complete information condition. Fits for the PSM (blue) and the GM (red) are superimposed on the data. Error bars are standard errors.

Two things are noteworthy from these results. First, as shown in Fig. 1, the GM consistently overestimates the magnitude of the coherence effect (i.e., the 2-way interactions), while the PSM shows clearly better fits. Additionally, as shown in Fig. 2, the PSM is better able to predict the consequences of inconsistent information on participants ratings, as compared to the GM. This relates to similarity gradients implied by eq. (3).

Experiment 2

Because results like those of Exp. 1 are difficult to reconcile with causal-model theory, in particular the lack of a coherence effect, Rehder (2017) proposed that small property interactions in results like those of Exp. 1, occur because instructions and materials emphasized a single almost defining property (property C in Exp. 1). Had traditional category labels been used (i.e., a category name, such as "dog"), large interactions would emerge, as expected by causal-model theory. To test Rehder's (2017) hypothesis, in Exp. 2 we used the causal chain model, but subjects were not told that there was a central property that gave the category its name. Instead, an arbitrary category label was provided.

As there should be no inferential processes in this task, we predicted that all properties in the causal model would show about the same weight, and on average they would produce a greater regression weight than the isolated property D. Regarding the interactions, the PSM predicts that, because of the piecemeal strategy, directly connected properties (AB, BC) would exhibit a larger regression weight than the indirectly connected properties (AC), and that the AB and BC terms would show a higher regression weight than the interactions of not connected properties (AD, BD, CD).

Design and Participants Exp. 2's design was identical to that of Exp. 1. Subjects (N = 64) were Adolfo Ibáñez University undergraduates (males = 21) who participated for course credit.

Materials and Procedures Materials were identical to those used in Exp. 1. However, arbitrary names were used to label categories, and no property was described as central or described the category's function. Except for the arbitrary category name, procedures were identical to Exp. 1.

Results Results Individualized regression coefficients were submitted to a mixed 2 (domain: living things, artifacts) × 10 (coefficients) mixed ANOVA (see Fig. 3). The mixed ANOVA showed there was no effect of the domain factor ($F<1$) and it did not interact with the coefficients factor ($F<1$). Consequently, for all subsequent analyses we collapsed this factor.

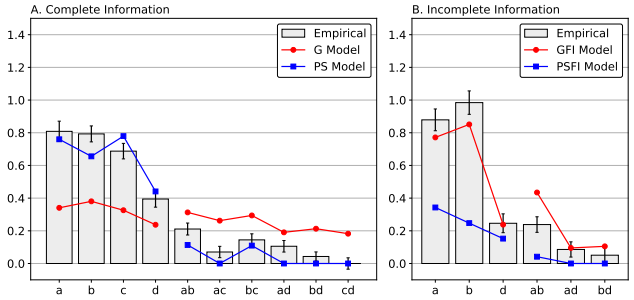


Figure 3: Regression weights for Exp. 2's individual features and interactions. Fits for the PSM (blue) and the GM (red) are superimposed on the data. Error bars are standard errors.

As predicted by the PSM, orthogonal planned comparisons showed that the average coefficient for D was significantly lower than the average of coefficients for A, B and C ($F(1, 63)=44.72$, $MS_e=.19$, $p<.001$, $\eta p^2=.42$, power$>.99$); but that there were no significant differences between C versus A and B ($F(1, 63)=3.46$, $MS_e=.24$, $p=.068$, $\eta p^2=.05$, power$=.45$), or A versus B ($F<1$). This pattern of results for main effects suggests that our subjects did indeed judge all properties to

be about equally central. This result contrasts with Exp. 1, where inference induced differential property weights. However, as in Exp. 1, property D was judged to be less central than properties belonging to the causal model. Again, this shows that participants are sensitive to causal information and are not disregarding it by using a pure associative strategy.

As predicted by the PSM, two non-orthogonal planned comparisons showed that the AC interaction was significantly smaller than the average of the AB and BC interactions ($F(1, 63)=7.02$, $MS_e=.42$, $p=.01$, $\eta p^2=.10$, power=.74), and that the average of AB and BC interactions was significantly greater than the average of the AD, BD, and CD interactions ($F(1, 63)=15.05$, $MS_e=2.52$, $p<.001$, $\eta p^2=.19$, power=.97).

As in Exp. 1, the low AC coefficient suggests that participants are complying with the causal Markov condition. At odds with Exp. 1, participants did not use property A (the screened-off property) to infer the state of property C (complete information mean=0.82, incomplete information mean=0.89; $t(63)=1.2$, $p=.24$), but used property B (Cs direct cause) (complete information mean=0.79, incomplete information mean=1.0; $t(63)=3.95$, $p<.001$). Furthermore, in the incomplete information condition, participants relied less on the isolated property D to make inferences (complete information mean=0.38, incomplete information mean=0.25; $t(63)=2.16$, $p=.04$). These results are broadly consistent with the hypothesis that using an arbitrary category label would promote causal classification.

Model fitting We fit the PSM and the GM to the classification ratings of Exp. 2 in the same ways as in Exp. 1 (Fig. 3). In the complete information condition both models achieved a high correlation with the ratings: $r_{PSM} = 0.83$, $r_{GM} = 0.84$. The bigger AIC for the GM (15.8) in comparison to the PSM (11.9) indicates that the PSM provides a slightly better characterization of the data in this condition. The best-fitting parameters for the PSM were: $p_A = 0.529$, $p_B = 0.455$, $p_C = 0.577$, $p_D = .293$, $\beta = 0.096$ and for the GM: $c_A = 0.921$, $b_B = 0.925$, $b_C = 0.887$, $b_D = 0.801$, $m_{AB} = 0.537$, $m_{BC} = 0.158$, $\gamma = 0.908$. As in Exp. 1, while fits for both models are similar, the GM achieves this by assigning values to the causal relation parameters lower than participants were taught.

For the incomplete information condition, we adjusted the PSM and the GM as in Exp. 1. The models achieved a high correlation with the ratings: $r_{PSM} = 0.88$, $r_{GM} = 0.90$. Again, we obtained a bigger AIC for the GM (14.6) in comparison to the PSM (13.0). The best-fitting parameters for the PSM were: $p_A = 0.316$, $p_B = 0.204$, $p_D = 0.236$, $b_C = 0.527$, $m_{BC} = 0.312$, $\beta = 0.230$ and for the GM: $c_A = 0.889$, $b_B = 0.847$, $b_C = 0.937$, $b_D = 0.640$, $m_{AB} = 0.439$, $m_{BC} = 0.592$, $\gamma = 0.431$. Note that, as in Exp. 1, the causal relation parameter for the relation between B and C was higher in this condition.

Just as for Exp. 1, in Exp. 2 the GM consistently overestimates the magnitude of the coherence effect (i.e., the 2-way interactions in Fig. 3, particularly in Panel A), while the PSM shows clearly better fits. Finally, as shown in Fig. 4, the PSM is better able to predict the consequences of inconsistent information on participants ratings, as compared to the GM.

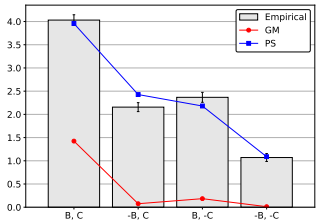


Figure 4: Average ratings for objects with different combinations of states (present or absent) for features B and C in Exp. 2. Fits for the PSM (blue) and the GM (red) are superimposed on the data. Error bars are standard errors.

General Discussion

In our experiments, the PSM was able to predict the pattern of results for main effects and interactions. Importantly, in both experiments the size of interaction coefficients remained low, and were not as high as those of main effects, as predicted by the GM. Furthermore, as would be expected if subjects were using associative mechanisms, the PSM was better able to predict ratings for objects with inconsistent information (Figs. 2 and 4). However, our results were not as clear regarding the mimicking Markov hypothesis. Exp. 1 produced data that is consistent with it, but Exp. 2 did not. In this latter experiment, participants appear to have complied with Markov both in the complete (i.e., low interaction coefficient for conditionally independent features) and in the incomplete information condition (i.e., appropriate screening-off of the conditionally independent distal cause). This pattern of results is consistent with the hypothesis that using an arbitrary category label enhances causal classification (Rehder, 2017). However, as in neither experiment did we obtain coherence effects, evidence for this hypothesis is mixed.

Prior research has found coherence effects in conditions similar to ours (e.g., in Rehder, 2003b, Fig. 4). The question then arises of how to account for these different results. In our experiments, we strove to use procedures as close as possible to those used by other researchers, so we tend to believe that differences do not lie in materials and procedures. Instead, we think it is possible that there are differences in how different populations handle causal information for categorization as well as for other tasks. Recently, using a causal inference task, (Rehder, 2018) found substantial variability in how individuals perform inferences (i.e., a single model was not able to account for the pattern of inferences of all participants, with a substantial minority behaving close to the predictions of an associative model). In a similar vein, we believe that no current model of causal cognition comfortably handles this variability and that future research should look to identify parameters that characterize tasks, individuals and populations in such a way that they are able to account for differences in causal categorization, and causal cognition in general.

References

Barsalou, L. W., Sloman, S. A., & Chaigneau, S. E. (2005). The hipe theory of function. In L. Carlson & E. van der Zee (Eds.), *Representing functional features for language and space: Insights from perception, categorization and development* (pp. 131–147). Oxford, England: Oxford University Press.

Carrara, M., & Mingardo, D. (2013). Artifact categorization. trends and problems. *Review of Philosophy and Psychology*, *3*(4), 351–373.

Chaigneau, S. E., Barsalou, L. W., & Sloman, S. A. (2004). Assessing the causal structure of function. *Journal of Experimental Psychology: General*, *133*(4), 601.

Johnson, S. G., & Ahn, W. (2015). Causal networks or causal islands? the representation of mechanisms and the transitivity of causal judgment. *Cognitive science*, *39*(7), 1468–1503.

Lombrozo, T., & Rehder, B. (2012). Functions in biological kind classification. *Cognitive psychology*, *65*(4), 457–485.

Nosofsky, R. M. (1992). Exemplars, prototypes, and similarity rules. In A. F. Healy, S. M. Kosslyn, & R. M. Shiffrin (Eds.), *From learning theory to connectionist theory: Essays in honor of William K. Estes* (pp. 149–167). Hillsdale, NJ: Erlbaum.

Park, J., & Sloman, S. A. (2013). Mechanistic beliefs determine adherence to the markov property in causal reasoning. *Cognitive Psychology*, *67*(4), 186–216.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. New York, NY, USA: Cambridge University Press.

Puebla, G., & Chaigneau, S. E. (2014). Inference and coherence in causal-based artifact categorization. *Cognition*, *130*(1), 50–65.

Rehder, B. (2003a). Categorization as causal reasoning. *Cognitive Science*, *27*(5), 709–748.

Rehder, B. (2003b). A causal-model theory of conceptual representation and categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(6), 1141.

Rehder, B. (2014). Independence and dependence in human causal reasoning. *Cognitive psychology*, *72*, 54–107.

Rehder, B. (2017). Concepts as causal models: Classification. In M. R. Waldmann (Ed.), *The oxford handbook of causal reasoning* (pp. 347–375). New York, NY, USA: Oxford University Press.

Rehder, B. (2018). Beyond markov: Accounting for independence violations in causal reasoning. *Cognitive psychology*, *103*, 42–84.

Rehder, B., & Burnett, R. C. (2005). Feature inference and the causal structure of categories. *Cognitive Psychology*, *50*, 264-314.

Rehder, B., & Hastie, R. (2001). Causal knowledge and categories: the effects of causal beliefs on categorization, induction, and similarity. *Journal of experimental psychology. General*, *130*(3), 323-60.

Rehder, B., & Kim, S. (2009). Classification as diagnostic reasoning. *Memory & Cognition*, *37 6*, 715-29.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237 4820*, 1317-23.

Sloman, S. A., & Lagnado, D. A. (2015). Causality in thought. *Annual review of psychology*, *66*, 223-47.

Inferring Structured Visual Concepts from Minimal Data

Peng Qian (pqian@mit.edu) Luke Hewitt (lbh@mit.edu)
Joshua B. Tenenbaum (jbt@mit.edu) Roger Levy (rplevy@mit.edu)
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology
43 Vassar Street, Cambridge, MA 02139 USA

Abstract

Humans can learn and reason about abstract concepts quickly, flexibly, and often from very little data. Here, we study how people learn novel concepts within a binary grid domain, and find that even this minimal task nonetheless necessitates the inference of highly structured parts as well as their compositional relationships. Furthermore, by changing the presentation condition of the learning examples, we reveal different approaches involved in learning such visual concepts: given the same images, human generalizations differ between rapid and static presentation conditions. We investigate this difference by developing several computational models that vary in their use of structured primitives and composition. We find that learning in the rapid presentation condition is best described as inference in simple models, while learning in the static presentation condition is best described as inference in a more structured space of graphics programs.

Keywords: Bayesian inference; concept learning; few-shot learning; program induction

Introduction

Human concept learning can involve remarkably fast and flexible abstraction. When we see a bridge or appreciate a sculpture, we not only perceive a set of objects, but also the underlying parts and their relationships. With such intuitive understanding of how the parts make the whole structure, human can productively compose learned primitives, generalize to new kinds of objects, and imagine new scenes.

We wish to study the compositional structure that underlies the richness of human visual concept learning by comparing computationally explicit models with human behavior. Prior work in cognitive psychology has built compositional models to describe human visual concept learning, typically by presupposing relevant, symbolically represented parts as inputs to the model, rather than operating directly on images (Shepard, Hovland, & Jenkins, 1961; Rehder & Hoffman, 2005; Goodman, Tenenbaum, Feldman, & Griffiths, 2008). These models are limited to a small stimulus space generated from the conjunction of the few predetermined features. In contrast, machine vision models successfully perform classification from arbitrary natural images (Krizhevsky, Sutskever, & Hinton, 2012), but recent work has found that these models lack the compositional structure necessary to recapitulate human visual concept learning in specific domains (Lake, Salakhutdinov, & Tenenbaum, 2015).

Here, we add to this literature by introducing a new minimal domain to incorporate both of these necessary ingredients: the inference of primitive parts directly from images, and the discovery of compositional structure that relates them. The domain we choose is 5×5 images with binary pixels. Despite the simplicity of this setup, Figure 1 shows that the visual concepts implied by these images can be complex and compositionally structured. In comparison to existing datasets that also occupy this space, our dataset focuses on occlusion and spatial juxtaposition that makes the basic parts particularly ambiguous, as well as concepts that lack prototypical images.

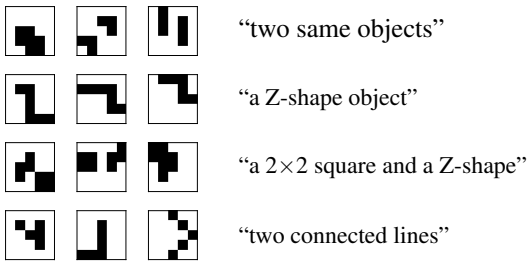
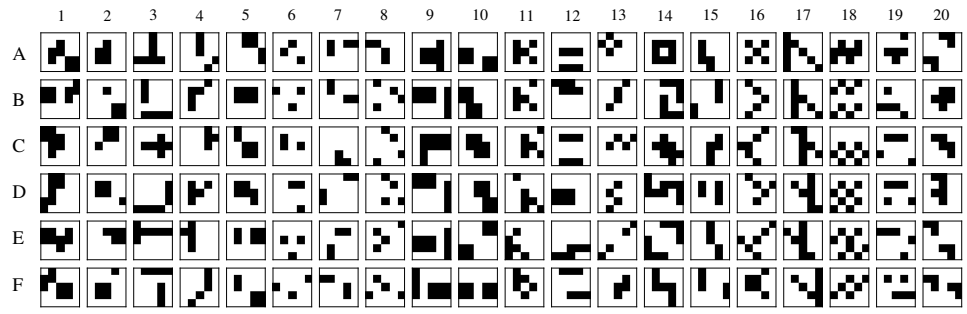


Figure 1: Abstract visual concepts represented by sets of images on a 5×5 binary grid.

Based on these images we develop a few-shot learning task to be presented under either static or rapid viewing conditions. Participants are asked to perform a 9-way classification, for which we compare several computational models that vary in their degree of compositionality and type of structured primitives present. We include a hierarchical Bayesian program learning model, and several additional Bayesian models with alternative primitives. We evaluate these models by quantitatively comparing how the model predictions match human judgments in few-shot generalization. Across the several Bayesian models tested, we find that the ability to jointly infer parts and compose them is critical to explain human generalizations in even this minimal domain, so long as participants are given sufficient time to view the stimulus. However, for rapid viewing conditions, participants' judgements are better explained as inference in a much simpler model with less rich compositional structure.



(a) Set of 20 visual concepts (b) Behavioral task

Figure 2: Visual stimuli and task paradigm used in the behavioral experiments. 60 images (rows A-C) are used for learning concepts and 60 (rows D-F) for testing generalization.

Learning grid concepts

We manually design 20 sets of binary images in the 5×5 grid, covering object occlusion, repetitive structures, and other interesting visual patterns. Each column in Figure 2a is a set of images representing a certain concept. For each concept, three examples (A-C) are designated for learning and a further three (D-F) for testing generalization behaviors. We have 60 different test trials in total.

We use a classification task to compare how humans and models generalize from three examples. The basic task is to learn the underlying concept from the 3 provided examples, and then to select from 9 novel query images the one that most likely displays the same concept. To create each trial, we sample one query image from the same visual concept as the three observed examples, and 8 from distinct other concepts which are drawn uniformly at random (See Fig 2b, 2a col. 8).

To collect human judgements, 216 participants were recruited via Amazon Mechanical Turk to participate in a few-shot classification task, each completing 20 trials: Participants were instructed to observe interesting objects on the visual scenes in the grid world. Subjects were presented with three example images, and then asked to choose one of the new query images that most likely displays the same concept, as is illustrated in Figure 2b.

Each participant was assigned either to the 'rapid' or 'static' viewing condition. In the 'static' condition, subjects could see all three of the example images simultaneously, for as long as required to make a judgement. However, in the 'rapid' condition, subjects instead watched only a video containing the stimuli in quick succession, with an interval between stimulus onsets of 72ms. At the end of the video, a 5×5 grey noise patch was displayed for backward masking.

Bayesian models

Concept learning, from the computational perspective, is fundamentally linked to the generalization problem $P(e'|e_1, e_2, \ldots, e_k)$. Consider a set of k observed examples $e_1, e_2, \ldots, e_k$, and a new observation e'. A concept c naturally plays a role when we factorize the conditional probability $P(e'|e_1, e_2, \ldots, e_k)$ as $\sum_{c \in \mathcal{C}} P(e'|c)P(c|e_1, \ldots, e_k)$.

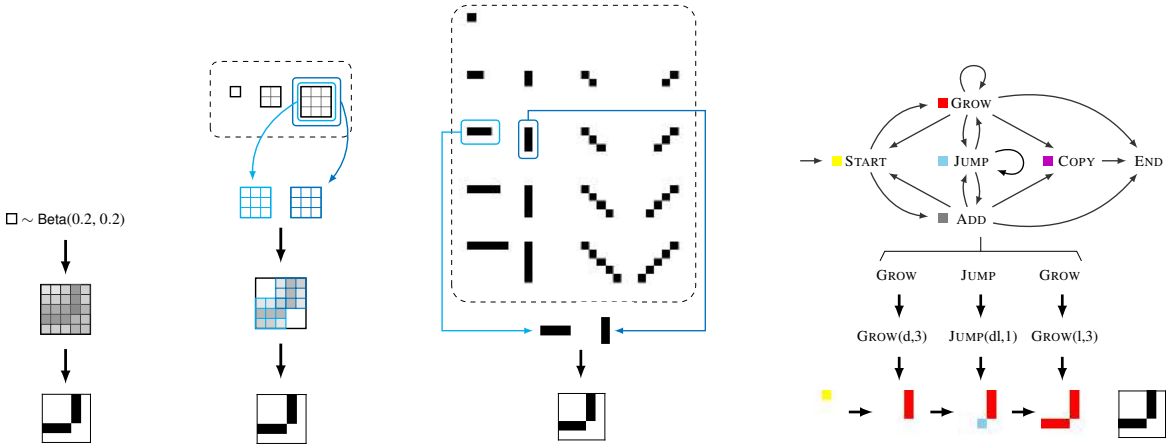
In the Bayesian framework of concept learning, we have the following according to Bayes rule and assuming conditional independence of observations given the concept c:

$$P(c|e_1, \ldots, e_k) = \frac{P(e_1, \ldots, e_k, c)}{\sum_{c \in \mathcal{C}} P(e_1, \ldots, e_k, c)} \propto \prod_{i=1}^{k} P(e_i|c)P(c) \tag{1}$$

The key component is about the structure of $P(e_1, \ldots, e_k, c)$, or more specifically $P(e|c)$ and $P(c)$. Here we construct four different models $P(e_1, \ldots, e_k, c)$ with various assumptions, levels of abstraction, and types of structured representation.

Independent Pixel Model This model assumes that the latent concept $c \in \mathcal{C}$ is a 25-element list of Bernoulli distribution parameters $[p_1, p_2, \ldots, p_{25}]$, each of which corresponds to one of the pixels in the grid and is sampled independently from a prior distribution $\text{Beta}(0.2, 0.2)$. An image instance e is generated from the concept c by sampling the binary state of each pixel in the grid according to its Bernoulli distribution parameter, as is shown in Figure 3a. This model lacks compositionality and complex structured representation, as the primitive available is just a single independent pixel.

Patch Model This model assumes that the latent concept $c \in \mathcal{C}$ is a list of patches drawn from a patch inventory of three different sizes (1×1, 2×2, 3×3), as is shown in Figure 3b. Specifically, c consists of the total number of patches as well as the size of each patches. To generate an image e from the concept c, the model first randomly localizes each patch on the grid and independently samples the Bernoulli distribution parameter for each pixel within the patches. Then an image instance is generated by sampling the binary state of each pixel within each localized patch according to the corresponding Bernoulli distribution parameter. The pixels out of the localized patches will always be turned off. This model has limited compositional structure, as it abstracts an image



(a) Pixel (b) Patch (c) Chunk (d) Program

Figure 3: Generative process of a concept and a image for different models.

as composition of several patches. However, the model lacks explicit structure within a patch, as the Bernoulli distribution parameters of the pixels within a certain patch are not shared at the concept level across multiple generated images.

Chunk Model This model assumes that the latent concept $c \in \mathcal{C}$ is a list of n chunks that are uniformly drawn from an inventory of line primitives (e.g. lines of various sizes and directions), as is shown in Figure 3c. An image e is generated by randomly placing on the grid the list of chunks from the concept c. The locations of the chunks are sampled during the image generation process and not shared at the concept level. While the built-in inventory of basic chunk primitives supports explicit structured representation, the model, however, lacks the mechanism to compose chunks into complex objects during the generative process.

Full Program Model Drawing inspiration from Lake et al. (2015), we design a model in which images are generated from a sequence of consecutive drawing actions. In the full program model, each concept $c \in \mathcal{C}$ takes the form of a probabilistic action program $\{a, \theta\}$, where a refers to the action type and θ refers to the parameters of each action. Once a concept c is generated, a binary image e is sampled from the concept by executing each action in the program step by step. Figure 3d illustrates how an example image is generated from the concept 'GROW(d,3) $\rightarrow$ JUMP(dl,1) $\rightarrow$ GROW(l,3)'.

To generate a concept, namely an action program in this model, the length of the program n is first sampled from an exponential distribution over all the possible program lengths ranging from 1 to 5 ($P(n) \propto \lambda^n$, where $\lambda = 0.9$), with preference to short programs. After that, a sequence of n actions, a, is sampled step by step from the plausible action primitives to construct the template of the program, under the constraints of the action transition grammar specified in Figure3d. For each action, the plausible transitions to other actions are uniformly distributed. The action primitives include GROW (adding pixels in a certain direction), JUMP (skipping over pixels in a certain direction), COPY (making copies of the current drawing trace and placing them randomly on the grid), ADD (generating a square patch of certain size and placing it randomly on the grid.), and START (placing currently generated trace on the grid and initializing a new trace).

After sampling the program template, the parameters θ_i of each action a_i in the program a (e.g. the direction and size of GROW) is uniformly sampled from the plausible values that a certain parameter type can take. There are eight basic values for the direction parameter, u (up), d (down), l (left), r (right), ul (upleft), ur (upright), dl (downleft), and dr (downright). The size parameter can take a number that is smaller or equal to the grid width size for GROW and JUMP actions, and a number less than 3 for COPY action. Both the direction and size parameter can also take a special parameter value '*any*', which refers to randomly sampling one of the basic directions or plausible size values during the image generation process.

Regarding the execution of an action program, the initial empty trace starts at the reference point $(0,0)$ on the temporary canvas. Following the action instructions, we draw pixels or move to other location on the canvas consecutively. The trace generated on the temporary canvas will be placed at a random place on the 5×5 grid once we encounter the end of the program or a START action. It is worth noticing here that the mechanism of composing action traces and starting new traces gives rise to the model's ability of utilizing more relational and object-like compositional structure. Therefore, Bayesian program learning model has more expressive compositionality and explicitly structured representation.

Few-Shot Classification and Generation

In order to evaluate each model against our collected human data, we must perform inference. However, this is computationally challenging to do exactly, and so we perform approx-

	H $_{\text{static}}$	H $_{\text{rapid}}$	M_1	M_2	M_3	M_4
Human $_{\text{static}}$	-	36	51	35	15	10
Human $_{\text{rapid}}$	36	-	36	25	11	10
Program [M_1]	51	36	-	35	15	9
Chunk [M_2]	35	25	35	-	14	6
Patch [M_3]	15	11	15	14	-	8
Pixel [M_4]	10	10	9	6	8	-

Table 1: Proportion of the same choices between model predictions and human judgements for 60 trials.

Model	Static presentation	Rapid presentation
Program	**0.39**	0.49
Chunk	0.43	0.47
Patch	0.52	**0.44**
Pixel	0.78	0.78
Uniform	0.56	0.46

Table 2: Hellinger distance averaged across 60 trials for each model compared to human data under each presentation condition (lower is better)

imate inference using a neural network trained for amortized few-shot classification in each model.

For each model, we train a separate network with a shared architecture, comprising a single convolutional layer and two fully connected layers with 200 hidden units. Each network was trained on model-generated data to produce a distribution of responses for 9-way classification of novel images. Specifically, we generate synthetic training data by sampling 9 concepts from each model's prior, drawing one image from each concept as the query examples, and a further 3 images from one concept as the observed examples. We optimise the network to classify the correct query example given the observed examples.

We then evaluate each of these trained networks on the same stimuli as presented to human subjects. Thus, regarding the behavioral task, each model's inference network is used to select the most likely query image from the 9 options.

For few-shot generation, we approximate the posterior $P(c|e_1,\ldots,e_k)$ using Markov Chain Monte Carlo (MCMC) implemented in WebPPL (Goodman & Stuhlmüller, 2014). Then we are able to produce novel instances from the inferred concepts.

Results

We are particularly interested in how human and the models proposed in this work make generalizations from few examples. We evaluate model predictions with respect to human judgments on 60 trials in the behavioral task, in each presentation condition.

Evaluation results of the models are listed in Table 1. We compute the proportion of choosing the same test images as the top choice for each pair between different models and human judgments. It is shown that the predictions of Bayesian program learning model largely matches the most popular (top 1) choice of human judgments in the static condition.

We compare the probability distribution of model's prediction to the distribution of human judgments over 9 test images for each trial in the experiment. We normalize human judgments to get a distribution of choice over the 9 test items, and similarly calculate $P(e'_i|e_1,e_2,e_3)$ over the 9 test items for each model. For each of 60 trials, we compute the Hellinger distance (Hellinger, 1909) between the distribution of model prediction and human judgement to quantify the distance between human and model responses. The average Hellinger distances are shown in Table 2, highlighting a difference between the two presentation conditions. For static presentations of the stimulus, the highly structured Bayesian program model is by far closest to human judgments in terms of the distribution of the choice in each trials. However, for rapid presentations of the stimulus, the program model suffers from overconfidence while the less structured 'chunk' model provides the best prediction of human judgements. Figure 4 visualizes the distribution of human judgments and models' predictions for several trials.

Regarding the question of what type of compositional structure supports human concept learning, the differences among the proportions of matched choices between human and four Bayesian models of different level of abstraction provide some interesting insights. As is discussed before, these Bayesian models can be summarized briefly with how much abstraction and what level of abstraction is built into the architecture: The pixel model does not have any compositional structures, while the patch model composes a scene by combining several patches. However, neither of these match human judgements well: the compositional ability of the patch model is largely limited due to the lack of explicitly structured primitives in its representation, as the patch only vaguely specify a pattern instead of clearly defining the structure of the pattern. With more structured primitives, the chunk model achieves significantly stronger results.

While the lack of structured representation makes it hard for the patch model to take the advantage of compositional structure in learning concepts, comparison between chunk and program models further suggest that hierarchical compositional structures are important in capturing human few-shot learning of simple visual concepts.

One final advantage of a Bayesian generative model is its generative process. Table 3 lists three of the inferred concepts by Bayesian program learning model, the approximate log posterior probabilities of these concepts, and the posterior samples for several sets of binary images used in the classification experiment. We can see that Bayesian program learning model successfully inferred the program and generated reasonable novel images of the same concept.

Discussion

Our work is an advanced investigation of similarity and generalization, along the line of research of classic Bayesian

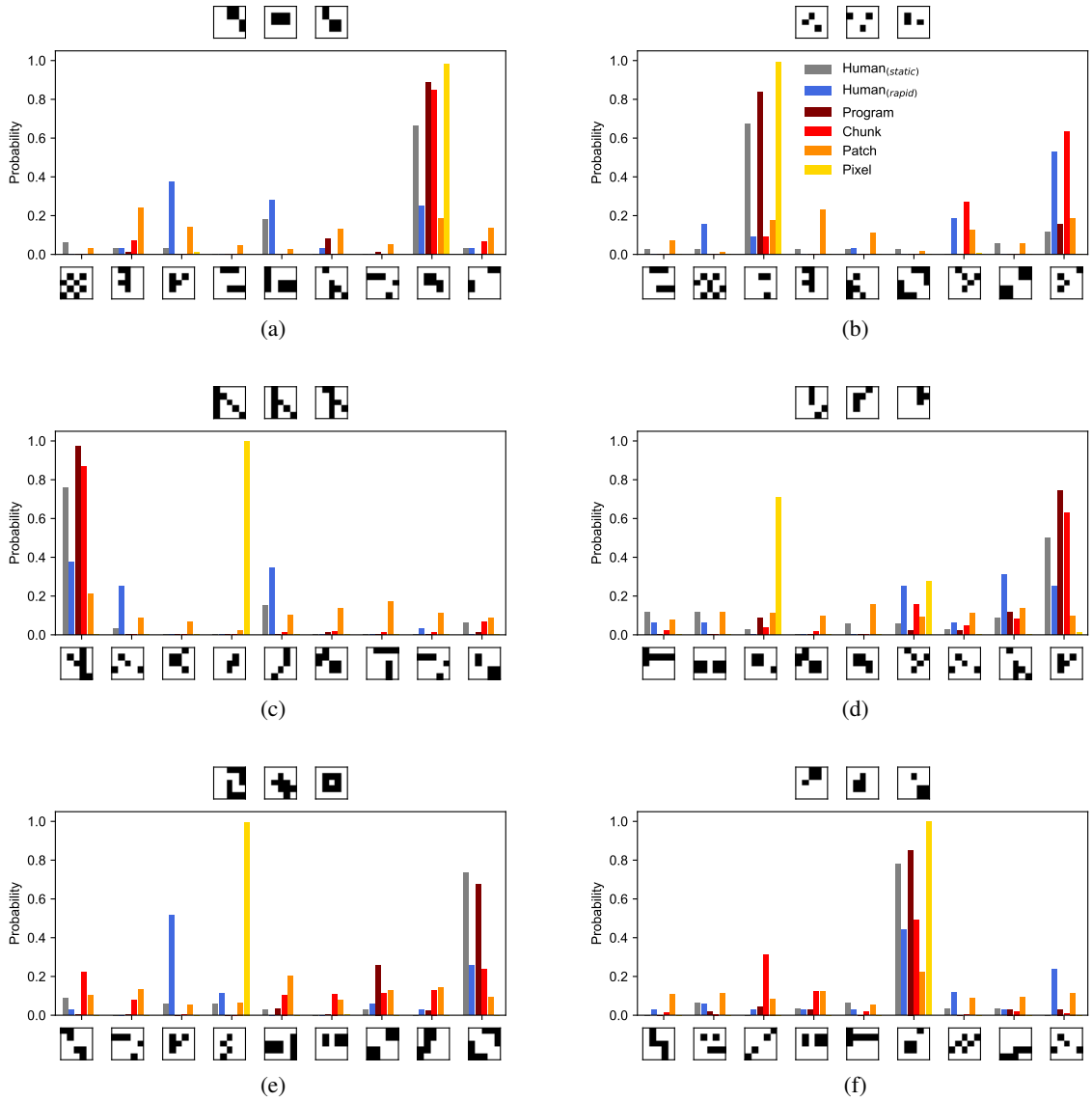


Figure 4: Fine-grained comparison of model responses to human responses.

concept learning (Tenenbaum, 2000; Tenenbaum & Griffiths, 2001; Kemp, Bernstein, & Tenenbaum, 2005; Goodman et al., 2008; Stuhlmuller, Tenenbaum, & Goodman, 2010) in computational cognitive science. We investigated visual concepts with more abstract, relational, compositional, hierarchical and object-like structure.

Compared to previous work (Orbán, Fiser, Aslin, & Lengyel, 2008) that studied learning visual scenes in a grid world composed of simple chunks (i.e. the statistical dependencies are simple associations between adjacent objects), this work explores more complex scenes that allow for more abstract (non-statistical) relations between objects in a scene. Further, objects in the visual scenes might occlude each other, which propose yet another challenge for learners, both model and humans, in identifying the latent structure.

Other important related works are Bayesian program learning of hand-written characters (Lake et al., 2015) and abstract visual concepts (Overlan, Jacobs, & Piantadosi, 2017). Our study introduces a richer grid concept domain, and develops computational account of different levels of abstraction. Although Lake et al. (2015) presents a Bayesian program learning model for few-shot learning of hand-written characters, which are images on a larger grid than what we use here, some interesting differences are worth mentioning here. Human might have a lot of practical experience with hand-written characters in daily life. There could be reasonably good prototype for hand-written characters as they are often standardized for communication purpose. People might rely on inferring a single visual prototype and generalize through similarity matching to the prototypical image. In our case, in contrast, it is hard to infer a single visual prototype for many of our concepts, even though there are only a small number

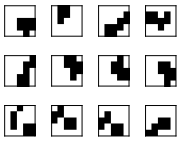
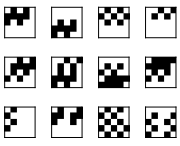
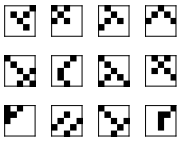
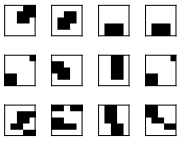
Examples	$\log(P)$	Concepts	Posterior samples
	-1.41	GROW(dl,2) → GROW(u,2) → GROW(ur,2) → START → ADD(2×2)	
	-1.41	GROW(dl,2) → JUMP(d,1) → GROW(ur, 2) → START → ADD(2×2)	
	-7.65	GROW(any,2) → JUMP(u,1) → GROW(ur,2)→ START → ADD(2×2)	
	-1.47	GROW(ur,2) → JUMP(l,2) → GROW(dl,2) → COPY(1)	
	-6.43	GROW(dl,2) → JUMP(r,3) → JUMP(u,1) → GROW(dl,2) → COPY(2)	
	-12.66	GROW(dl,2) → JUMP(r,any) → JUMP(u,1) → GROW(dl,2) → COPY(any)	
	-2.12	GROW(dr,3) → START → GROW(dl,3)	
	-6.28	GROW(dr,any) → START → GROW(ur,3)	
	-8.35	GROW(dl,3) → START → GROW(any,3)	
	-0.42	ADD(2×2) → START → ADD(2×2)	
	-3.07	ADD(2×2) → JUMP(any,any) → ADD(2×2)	
	-23.04	ADD(2×2) → START → GROW(r,2) → COPY(2)	

Table 3: Programs found by MCMC for several test concepts, with corresponding posterior-predictive samples of new images.

of observations to choose and generalize from.

This work also shows that compositionality is not the only important aspect behind human few-shot learning, in line with previous work (Schulz, Tenenbaum, Duvenaud, Speekenbrink, & Gershman, 2016) that demonstrates human's preferences of compositional pattern in function learning domain. The level of abstraction in the representation also plays an important role in building models that can better match the generalization behaviors observed in human concept learning.

We believe that our visual concept learning task contributes to an understanding of how humans learn and reason about novel visual concepts, addressing two questions: (1) what kinds of representation and architecture support flexible inference of underlying abstract structure, and the impressive generalizations that humans achieve from often minimal data? (2) Is this same architecture necessary, and is it sufficient, to explain the kind of rapid inferences humans are able to make given only a short glimpse of a concept? Comparisons among several Bayesian models with different degrees of abstraction demonstrate that, even in this minimal domain, humans can infer concepts with a rich compositional structure, but that the extent of this structure is dependent on the condition of presentation.

Acknowledgments

We would like to thank Maddie Cusimano and members of the MIT Computational Psycholinguistics Lab for their helpful comments on this project.

References

Goodman, N. D., & Stuhlmüller, A. (2014). *The Design and Implementation of Probabilistic Programming Languages.* `http://dippl.org`. (Accessed: 2017-12-17)

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive science*, *32*(1), 108–154.

Hellinger, E. (1909). Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen. *Journal für die reine und angewandte Mathematik*, *136*, 210–271.

Kemp, C., Bernstein, A., & Tenenbaum, J. B. (2005). A generative theory of similarity. In *Proceedings of the 27th annual conference of the cognitive science society* (pp. 1132–1137).

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, *350*(6266), 1332–1338.

Orbán, G., Fiser, J., Aslin, R. N., & Lengyel, M. (2008). Bayesian learning of visual chunks by human observers. *Proceedings of the National Academy of Sciences*, *105*(7), 2745–2750.

Overlan, M. C., Jacobs, R. A., & Piantadosi, S. T. (2017). Learning abstract visual concepts via probabilistic program induction in a language of thought. *Cognition*, *168*, 320–334.

Rehder, B., & Hoffman, A. B. (2005). Thirty-something categorization results explained: selective attention, eyetracking, and models of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(5), 811.

Schulz, E., Tenenbaum, J., Duvenaud, D. K., Speekenbrink, M., & Gershman, S. J. (2016). Probing the compositionality of intuitive functions. In *Advances in neural information processing systems* (pp. 3729–3737).

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological monographs: General and applied*, *75*(13), 1.

Stuhlmuller, A., Tenenbaum, J. B., & Goodman, N. D. (2010). Learning structured generative concepts. In *Proceedings of the cognitive science society* (Vol. 32).

Tenenbaum, J. B. (2000). Rules and similarity in concept learning. In *Advances in neural information processing systems* (pp. 59–65).

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and bayesian inference. *Behavioral and brain sciences*, *24*(4), 629–640.

The Expected Unexpected & Unexpected Unexpected:
How People's Conception of the Unexpected is Not Really That Unexpected

Molly S. Quinn (molly.quinn@ucdconnect.ie)
Kathleen Campbell (kathleen.campbell@ucdconnect.ie)
Mark T. Keane (mark.keane@ucd.ie)

School of Computer Science & VistaMilk SFI Research Centre,
University College Dublin, Belfield, Dublin 4, Ireland

Abstract

The answers people give when asked to "think of the unexpected" for everyday event scenarios appear to be more expected than unexpected. There are *expected unexpected* outcomes that closely adhere to the given information in a scenario, based on familiar disruptions and common plan-failures. There are also *unexpected unexpected* outcomes that are more inventive, that depart from given information, adding new concepts/actions. However, people seem to tend to conceive of the unexpected as the former more than the latter. Study 1 tests these proposals by analysing the object-concepts people mention in their reports of the unexpected and the agreement between their answers. Study 2 shows that object-choices are weakly influenced by recency, that is, the order of sentences in the scenario. The implications of these results for ideas in philosophy, psychology and computing are discussed.

Keywords: expectation; explanation; cognitive; judgments

As we know, there are known knowns; there are things that we know that we know. We also know there are known unknowns; that is to say we know there are some things we do not know. But there are also unknown unknowns, the ones we don't know we don't know.

Donald Rumsfeld, Feb 2002, US Secretary of Defence

1. Introduction

In an uncertain and contingent world, our ability to deal with the unexpected often gives us safe passage through the Siren-like obstacles of everyday life (see e.g., Weiner, 1985a). The Cognitive Sciences have often concerned themselves with how people think about the unexpected. However, most of this research relies on theory-driven definitions of the unexpected (e.g., low probability events), rather than simply asking people to "think of the unexpected" and see how they respond[1]. In the present paper, we report two studies that ask people to generate unexpected events for everyday scenarios and then analyse their responses. As the title of the paper suggests, our main finding is that *people's conception of the unexpected is not really that unexpected at all.*

In Cognitive Psychology, *unexpectedness* is often used as a dependent variable in studies of human thinking and decision making. For example, in reasoning research, the unexpected has often been proposed to elicit counterfactual thinking (Kahneman & Miller, 1986; McEleney & Byrne, 2006). In attribution research, the unexpected has been cast as non-normative behavior in others, that elicits spontaneous causal thinking (Hastie, 1984; Weiner, 1985b). And, surprising events are often defined in terms of their unexpectedness (Meyer et al., 1997; Maguire et al., 2011).

However, most of these studies do not actually *ask* people to report unexpected outcomes; rather they adopt *a priori* operational definitions of *unexpectedness* based on the experimenter's theoretical stance. The unexpected is commonly operationalized as (i) an event rated as having a low subjective probability (Maguire et al., 2011; Teigen & Keren, 2003), or (ii) profiles of people with inconsistent traits (Hastie, 1984), or (iii) events that are simply asserted to be unexpected to the actors in a narrative (McEleney & Byrne, 2006). In contrast, we do not use *a priori* definitions but rather, simply, ask people to tell us what they consider the unexpected to be. This sort of behaviour was observed by Foster & Keane (2015, 2019), in studies on surprise, in the form of familiar surprises ("I am surprised my wallet is missing from my trouser pocket, but I am guessing it was robbed") and unfamiliar surprises ("I am surprise my belt is missing from my trousers but have no idea how that could have happened"; see also Maguire & Keane, 2006).

1.1 Thinking About the Unexpected

Consider the simple task used in the current experiments to elicit unexpected events from people. Imagine being told a story about a woman, called Louise, who is going shopping at her favourite clothes store, in which she draws money from an ATM and heads into town on the bus. Now imagine you are told "Something unexpected occurred. What do you think happened?" One could respond with one of the following unexpected events, saying that:

1) Louise lost the money she drew from the ATM.
2) Louise was delayed in traffic, arrived late and the shop was shut.

However, one could also validly say:

3) The bus stopped at a charity bus-wash and Louise got covered in suds.
4) Louise pulled a gun on the driver and robbed him to raise more money for her shopping spree.

Intuitively, as unexpected outcomes, the first two answers (1-2) are quite conservative and mundane and *less* unexpected

[1] Khemlani et al.'s (2011) Expt. 3 is a notable, but rare, exception though it focusses on the issue of latent scope.

than the latter two responses (3-4) which are more inventive and a lot *more* unexpected. We call the former answers the *expected unexpected* and gloss the latter as the *unexpected unexpected*.

Expected unexpected outcomes tend to maintain the original goal of the story scenario (i.e., shopping) and the stated object-concepts associated with the story's goal; these *goal concepts* tend to be re-used in the unexpected event (i.e., *bus*, *ATM*, *money*, *store*) and few new objects are added (e.g., *traffic*). Furthermore, these events are often "common failures" that are familiar to people; losing one's money or being delayed are common reasons for failed plans and goals.

Unexpected unexpected outcomes, in contrast, may establish new goals for the story scenario (e.g., attending a charity event) and, though *goal objects* may be used (i.e., *bus*, *money*), often "new" object-concepts not present in the original story are introduced (e.g., *guns, suds*). Also, these unexpected events are quite unfamiliar to the scenario: getting involved in a charity bus-wash is not a common everyday event for most people who are going shopping. In previous work on surprise using these everyday scenarios (Foster & Keane, 2015), we noticed that people typically produced *expected unexpected* answers rather than *unexpected unexpected* ones. But, why?

Why do people *minimally* perturb the stated scenario, keeping its goals and goal-concepts in these expected-unexpected events that they seem to prefer to generate? One possibility is that when people are thinking about the unexpected, they are essentially trying to explain how current goals might fail; so, unexpected events tend to describe disruptions to a current plan or undoings of assumed facts that enable current goals. Being delayed in traffic disrupts Louise's shopping plan, undoing the assumption that the bus gets her to the shop on time. Losing one's money is an unexpected event that explains how any shopping-goal might fail. From an adaptive perspective, it makes sense to minimally change the current situation when projecting such unexpected futures. In contrast, more creative unexpected-unexpected events, that depart significantly from the current scenario, may never occur and, therefore, seem not to be considered. In short, the former probably have higher predictive value than the latter. Across the Cognitive Sciences, many researchers have highlighted this minimalist stance in people when they encounter the unexpected.

1.2 The Minimalism of the Unexpected

In Philosophy, when reasoning about inconsistencies (such as new, unexpected facts), it has been repeatedly proposed that any change to stated propositions or prior beliefs should be as minimal as possible; observing the "maxim of minimal mutilation" (Quine, 1992, p.14), or the "principle of conservativism" (Harman, 1986, p.46). Similarly, in considering counterfactual situations (of which unexpected situations could be a subclass), Lewis (1986), taking a possible-world perspective, talks of finding the maximally-similar world to the current one.

In Psychology, related ideas arise in considering the *minimal-mutability* of counterfactual scenarios (Kahneman & Miller, 1986). Also, in the psychology of explanation, several researchers have noted how explanations of the unexpected maintain aspects of the original scenario; they preserve the level-of-abstraction of the original scenario rather than identifying new or more specific information (see e.g., Johnson & Keil, 2014) or they favor explanations with a narrow, latent scope (Khemlani et al., 2011).

In Artificial Intelligence, theories of understanding and explanation directly predict *minimalism* and show how the "expected unexpected" might arise (see Leake, 1991, 1992; Schank, 1986; Schank, Kass & Riesbeck, 1994). David Leake's (1992) computational account of understanding gives the most comprehensive account of what people might be doing when asked to "think of the unexpected" (see also Schank, 1986). Leake argues that people store *explanation patterns* to handle plan failures and anomalies encountered in everyday life. These explanation patterns can be thought of as "script-like" structures (Schank & Abelson, 1977), at varying degrees of abstraction, that can account for difficulties that arise in plans; for instance, in considering how a planned shopping-expedition might be disrupted, a number of standard disruption-events suggest themselves from pre-stored explanation patterns (e.g., that I might be mugged, or that I might lose my money or that I might be delayed). To handle a contingent world, it is proposed that we store these pre-canned explanations and retrieve them to quickly explain unexpected happenings[2]. Although these ideas have been referenced in the psychological literature (e.g., Hastie, 1984), they have not been worked up into a psychological model or specifically tested. Here, we propose an initial psychological account, that we then test this model in two experiments.

1.3 Minimal Retrieval Model

Our psychological account for the generating unexpected events is called the Minimal Retrieval Model (MRM). According to this model, when people are asked for unexpected outcomes to everyday scenarios, they retrieve explanation patterns and adapt them to the situation in hand. Specifically, that people build a *cue frame* using the given information in the scenario (e.g., the goals, actors, actions and objects mentioned) to search memory for suitable explanation patterns. For example, when people are told Louise had the goal of going shopping, took money from the ATM and then went to town, it is assumed that memory is searched for explanation patterns involving shopping-goals, female-shoppers, buses, money, and ATMs. Accordingly, unexpected events such as Louise losing her money, having problems with the ATM, or being delayed will tend to be found in memory and returned as responses, rather than more inventive answers.

Minimal Retrieval Model makes several predictions about the nature of the unexpected outcomes reported by people; specifically, it is predicted that (i) reported unexpected events should tend to use the stated object-concepts in the original

[2] Note, both Leake and Schank maintain that not all situations can be handled by these pre-canned explanation-patterns; explaining the unexpected may sometimes involve much more creative uses of prior knowledge, such as analogical explanations.

scenario because memory will be *cued* with these concepts and the retrieved explanation patterns will instantiate these objects, (ii) goal-related objects will be preferred in reported events, over non-goal objects, (iii) people will tend to agree on the reported unexpected events because they are using familiar plan-failures (i.e., explanation patterns). Note, the first two of these predictions basically propose that minimalism is a side-effect of the retrieval process and the third prediction basically says that answers will be expected-unexpected events rather than unexpected-unexpected ones.

Table 1: Louise-Shopping Story & Answer Categories

Sentence Order Used in Study 1 & 2 (Normal Condition)		
Goal	(S1)	Louise wants to shop at an expensive clothes store.
Non-Goal	(S2)	She is wearing her favourite dress and matching shoes.
Goal Step	(S3)	Louise draws money from the ATM.
Sentence Order Used in Study 2 (Reversed Condition)		
Goal	(S1)	Louise wants to shop at an expensive clothes store.
Goal Step	(S2)	Louise draws money from the ATM.
Non-Goal	(S3)	She is wearing her favourite dress and matching shoes.
Answer Categories for Unexpected Events		
ls_neg_ans1		She has insufficient money to buy
ls_neg_ans2		She has problems with the ATM
ls_neg_ans3		She is robbed or loses money/card/id.
ls_neg_ans4		Clothes issues (dress rips, shoe snaps).
ls_neg_ans5		The shop is closed.
ls_pos_ans1		She finds or ATM gives more money.
ls_pos_ans2		She has more money than she thought.
ls_pos_ans3		Good events involving shoes and dress.
ls_pos_ans4		Sale is on at the shop.
ls_other		e.g.; ATM speaks,gives money to charity.

Preference for Stated Object-Concepts. If memory is being searched with the stated goals and object-concepts given in the scenario then the explanation patterns retrieved should reflect these objects/entities and minimally introduce new objects (e.g., ones that mention *money*, *ATMs*, *buses*). This process thus delivers unexpected events that remain close to the original scenario, with perhaps better predictive value. In Quine's terms, the reported unexpected event will *minimally mutilate* the original scenario. In the present studies, we measure this minimalism by recording the frequency of stated objects versus new objects in the reported unexpected outcomes (excluding references to *Louise* who as the main actor will always tend to be mentioned).

Preference for Goal Objects. Within the preference for stated objects, MRM also predicts that goal-related objects will be preferred over less goal-related objects (which we will call *non-goal objects*). For example, if the scenario mentions that "Louise was wearing her favorite dress and matching shoes" (see Table 1), these objects *dress* and *shoes* are less goal-critical. People need *money* to go shopping but what they wear is less critical to the shopping goal [3]. Although, it is feasible to generate unexpected events from these non-goal objects (e.g., "when Louise got on the bus, everyone was wearing the same dress and shoes"), explanation patterns based on non-goal objects are less likely to be retrieved because they are not goal-critical. In the present studies, we check for this preference by recording the frequency of stated goals-objects versus non-goal objects in the reported unexpected outcomes (obviously, again, excluding references to *Louise* who as the main actor will tend to be mentioned anyway).

Agreement. By definition, explanation patterns are explanations for commonly-occurring disruptions to everyday plan-goal sequences; it makes more sense for the cognitive system to assume that disruptions that happened repeatedly in the past will happen again. As such, they should be familiar to people, they should be expected-unexpected events. Hence, there should be a high level of agreement between people in the unexpected outcomes they propose. This means that most answers should fall into a small set of common answer-categories; for instance, we should see many people using answers that describe "Louise losing her money" or "the shop being shut" (see Table 1). In the present studies, we test this prediction by classifying people's responses into answer-categories and recording the proportion of answers that fall into these categories. In the remainder of this paper, we report two experiments designed to test these predictions. To the best of our knowledge, these tests are new, as are the measures used to assess what people report as the unexpected.

2. Study 1: How Unexpected?

The study presented participants with scenarios describing everyday events such as going shopping, doing exams, going on trips and attending business meetings (adapted from Foster & Keane, 2015). Each story was followed by an instruction to think of the unexpected. The unexpected outcomes reported for each scenario were categorized by three judges in terms of object-concepts (i.e., *goal-objects*, *non-goal-objects*, *both* goal- and non-goal-objects and *neither* of the stated objects) and answer-categories used.

[3] Though, obviously, it could be made goal-critical with additional conditions (e.g., if one said "she wanted to be able to match the clothes she was wearing with those in the shop").

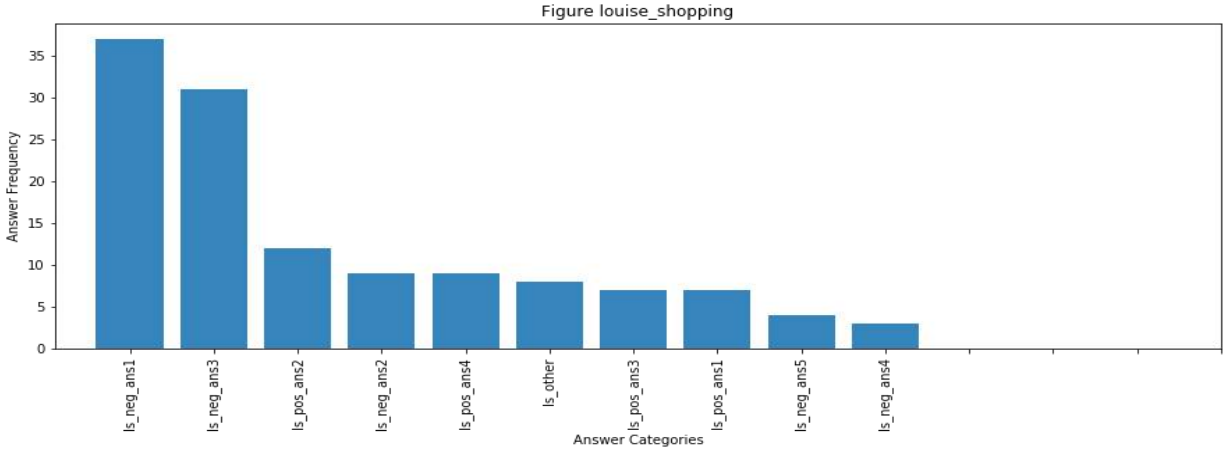


Fig 1. Frequencies for Answer Categories in Louise Story (see Table 1 for Meaning of Answer Code)

2.1 Method

Participants & Design. The study involved 127 participants and was run on the crowdsourcing platform, prolific.com[4]. Participants were native English speakers from Ireland/UK/USA and had not participated in previous studies by the group.

Procedure & Materials. All participants received the same 20 scenarios (randomly re-ordered for each participant), after first being shown two practice materials (these items were not flagged as practice items and not included in the analysis). Participants were presented with a series of web pages, explaining the task and then presented with one scenario after the other, each on its own page. Each material was presented along with the associated instruction to think of unexpected outcomes to the scenario (material lists available on request).

Each scenario was described in three sentences: a setting goal-sentence (S1; Louise going shopping), one giving additional information that was not on the critical path of the plan for the goal (S2; Louise wearing her favourite dress and shoes) and a final one describing some further action taken to achieve the goal (S3: Louise drawing money, see Table 1). Two comprehension questions were asked about the scenario, as a test micro-task, to ensure participants had carefully read and understood the scenario. The instruction to think of the unexpected followed these questions. Participants wrote their responses in a text-field with no upper limit.

Measures & Judgements. In total the study yielded 2,540 responses (127 participants x 20 materials) each of which were judged by three raters (i.e., the three authors) for answer-type and the use of goal-related and non-goal-related objects (the main actor of the scenario was always excluded from this object judgement). Every response was categorized into answer categories specific to the material (e.g., in Louise scenario, answers about losing money, being delayed, the shop being shut). The classification of answers by the three judges revealed high levels of agreement: pairwise comparisons between judgements revealed Cohen's Kappas from K=0.82 to K=0.89. The classification of objects in the answers (as goal, non-goal, both, neither stated object) by the three judges had lower, but acceptable, levels of agreement. Agreement between Judge-1 and Judge-3 was lower (K=0.56) as object classifications were re-defined more tightly for Judge-2 and 3, who agreed more often (K=0.74). The final classifications chosen for all judgements was based on a majority vote from the three judges. Three-way splits (which were rare, N<20) were resolved by discussion.

2.2 Results & Discussion

Overall, the results confirm the predictions made from the Minimal Retrieval Model; the unexpected is really not that unexpected. People tend to (i) stick to the stated objects in the scenario rather than use new objects, (ii) they show a strong preference for given goal-objects over non-goal objects, (iii) they agree on the unexpected events reported, as a few answer-categories cover most responses made.

Preference for Stated Objects. As predicted, people tend to stick to the object-concepts given in the scenario (e.g., the *money*, *buses*, *shoes* of the Louise story), rather than introducing new objects into their answers. Of the 2,540 unexpected outcomes reported by participants, 78% (N=1891) relied on the given objects, while only 22% (N = 649) of answers mentioned none of the stated objects (i.e., "Louise met her best *friend*"). Most unexpected outcomes assert a new relation between the given objects (e.g., "The *ATM* showed *Louise* had more *money"*).

Preference for Goal over Non-Goal Objects. Furthermore, of the 78% (N=1,891) of unexpected outcomes that used the given objects from the scenario, the majority used only goal-objects (80%; N=1,518) with a minority using the non-goal

[4] The original experiment was divided into four conditions that used variants on the main instruction: asking for "something unexpected", "something good and unexpected", "something bad and unexpected", or "what would happen if the goal failed". For brevity, this manipulation is not reported here, as the same pattern of responding is seen across all these four conditions.

objects (14%; N=261) and a few using both stated object-types (6%; N=112). Chi^2 tests performed on frequencies of the four object-types (df=3) for each material were all statistically significant at $p < 0.01$. Also, a by-materials analysis, using Wilcoxen's test, revealed a statistically significant difference in the proportions of goal versus non-goal objects chosen, $z = 3.00$, $p < .001$.

Of course, one could argue that this result is not surprising, as two sentences mentioned goal-objects (S1 and S3) and only one mentions non-goal objects (S2, see e.g. Table 1)[5]. However, even if examine goal and non-goal object choices at the sentence level, the preference for goal-objects remains: on average, goal-objects are chosen from S1 (M=27%) and S3 (M=44%) more often than the non-goal-objects from S2 (M=13%; see Table 2). Chi^2 tests performed on 40 pairwise comparisons of choices, for S1xS2 and S2xS3, found that only 5 comparisons were non-significant (most are $p < .001$). However, it is clear that there is a preference for goal-objects from the final sentence in the scenario (S3 at 44%), suggesting a recency effect, that we explore in Study 2.

Thus far, the evidence suggests that people respond in a minimalist way, sticking close to the original scenario's objects, with a strong preference for stated goal-related objects over stated non-goal objects.

Agreement in Answer Categories. Apart from analysing the object-concepts used in the unexpected outcome, we also categorised responses and noted their frequency of occurrence (e.g. see Table 2 and Fig.1). On average, materials were found to involve 10 answer categories (M=10.68, SD=1.7); Min=6 (*lucy_loan*) and a Max=13 (*robert_essay*; see Table 2). Figure 1 shows a typical distribution of responses across answer-categories for the Louise story; note, the top-3 most-frequently-used answer-categories of 10 categories tend to cover most responses (63%) followed by a long tail of lower-frequencies for other categories. Note, one answer-category was used as a residual one (the *other* category), and it typically also has a low count.

Table 2 shows the percentage of responses that fall into the top-3 most-used answer-categories for each material. In the most extreme case, *lucy_loan*, 89% of responses are covered by the top-3 answer-categories, with the lowest being 49% (for *bill_holiday*). This pattern of responding shows that there are very high levels of agreement between people with respect to the unexpected events they propose. For example, in the *louise_shopping* scenario, 29% of people proposed that Louise had money problems such as spending too much or not having enough money for the clothes (*ls_neg_ans1*), 25% proposed that she lost her money in some way (*ls_neg_ans3)* and 9% said that the ATM told her she had more money than she thought (*ls_pos_ans2).* None of these unexpected events are particularly "unexpected"; they are rather, typical disruptions that occur in everyday plans to achieve mundane goals. They are *expected unexpected* events. Indeed, more inventive answers -- the *unexpected unexpected* -- are quite rare and typically found in the *other* category. For instance, in the Louise story, the *other* category (*ls_other, N=8*) includes responses about (i) Louise deciding to give her money to a charity instead, (ii) Louise being approached by a film director who says she is beautiful and wants to make her a star and, (iii) the wonderful "The ATM opens, and Louise realizes it is a portal to her happiest childhood memory". These sorts of answers are the *unexpected unexpected*, truly unusual possible outcomes but, notably, are rare too.

3. Study 2: The Recently Unexpected?

Study 1 supports the minimalist predictions that people will stick closely to the original scenario, introduce few "new" objects and agree with others when proposing unexpected events for everyday scenarios. However, with respect to the object-concept analyses, the preference for goal-objects (especially, objects from the final sentence, S3) and lack-of-preference for non-goal objects could be due, in part, to a recency effect. That is, maybe people follow on from the last sentence in the story and, hence, use its goal-objects. For example, in the Louise story people do not mention her *shoes* and *dress* (the non-goal objects from S2) but rather follow on from the mention of *ATMs* and *money* (from S3) in proposing their unexpected outcome. If this were true then people's object-choices perhaps hinge less on their goal or non-goal status but more on order of mention. In this study, we put the non-goal sentence last (S3) to check if this changes the object-choices made (see Table 1 for a sample material).

3.1 Method

Participants and Design. The study involved 258 participants on the prolific.com crowdsourcing platform[5,6]. All were native English speakers from Ireland/UK/USA and had not taken part in our previous studies.

Procedure & Materials. All participants received the 20 scenarios used in Study 1, using the same procedure. There were two main conditions of interest: Normal (N = 126) and Reversed conditions (N = 132). Participants in the Normal condition received the same materials as those used in Study 1. Participants in the Reversed condition received variants of these materials, in which the non-goal sentence was moved to the last position in the story (S3; see example in Table 1).

Measures & Judgements. In total the study yielded 5,160 responses (258 participants x 20 materials). Given the very large number of responses in this experiment, we automated the object-judgement process (program and data will be made available on email request). A program, called ObjJudge was developed using the NLTK, Pandas and SciPy python packages to process the answers and identify whether

[5] Also, note the sentences themselves in the original scenarios mentioned equivalent numbers of objects (by-materials, paired t-tests on the object counts in S1, S2 and S3 revealed no differences, all ps > 0.10).

[6] The original experiment had 6 conditions that used (i) the three instructional variants used in Study 1 ("unexpected", "good and unexpected", "bad and unexpected"), crossed with (ii) a variant to think of "bizarre" events. Initial, analyses suggested that these variables do not impact the pattern of responding for object choices and, for brevity, are not reported here.

they mentioned goal-objects, non-goal-objects, both object-types or neither. The program was trained on responses and their respective judgments from Study 1. Using lists of the object-entities (i.e.; object words given in the scenarios and synonyms provided in responses from Study 1), ObjJudge sorts the responses given in Study 2 into goal-object, non-goal object, both, or neither categories.

Stated simply, this program matches object-entities in the response-string against object-lists for each material (including common synonyms that people used in Study 1 answers). With this program, we achieved a high accuracy over all materials comparing its object-judgements against the human-judgments from Study 1 (M=93%; Min=90%, Max = 97%; Cohen's Kappa was K=1). In all other respects, the object-judgement measures were as detailed in Study 1.

3.2 Results & Discussion

The results replicate the findings of Study 1 that people (i) stick to the stated objects in the scenario rather than using new objects, (ii) they show a strong preference for the given goal-objects over non-goal objects. However, it also shows that here is a slight recency effect, in the Reversed condition, where the choice of non-goal objects increased by about 9% relative to the Normal condition.

Preference for Stated Objects. As we saw in Study 1, people tend to stick to the object-concepts given in the scenario (e.g., the *money*, *shoes* of the Louise story), rather than introducing new objects into their answers (e.g., *guns, suds*). Of the 5,160 unexpected outcomes reported by participants, 79% (N=4,098) made use of objects stated in the scenario, while only 21% (N = 1,062) of answers mentioned none of the given objects Most of the unexpected outcomes reported created a new relation between the given objects (e.g., "The *ATM* showed *Louise* had more *money"*).

Preference for Goal over Non-Goal Objects. In a similar vein, of the 4,098 (79% of 5,160) unexpected outcomes that used objects from the original scenario, the majority mentioned only goal-objects (56%; N=2,869) whereas a minority mentioned only non-goal objects (13%; N=688), with some responses mentioning both goal and non-goal objects (10%; N=541). Chi^2 tests performed on frequencies of object-types reported for each material were all significant at $p < 0.05$ (df=3, with corrections for low-valued cells).

However, these analyses collapse across the Normal-Reversed manipulation designed to test for recency. When these conditions are broken out there is a small but statistically-significant increase in the use of the non-goal objects (roughly 9%, with a corresponding reduction in goal-object choices). The following are the relative percentages, for each choice category, $Chi^2(3) = 76.9$, $p < 0.001$:

Cond.	*Goal*	*Non-Goal*	*Both*	*Neither*
Normal	61%	10%	9%	20%
Reversed	51%	16%	12%	21%

In short, when the sentence with the non-goal objects is last in the story, people prefer to use the non-goal objects somewhat more often; showing that they are sensitive, to some degree, to the order in which information is given, though the dominance of goal-object choice still remains.

4. Conclusions

To the best of our knowledge, the current study is the only one simply asking people to "think of the unexpected" when presented with everyday scenarios. Our view is that when people are asked to do this, they tend to recall characteristic explanation patterns that account for common disruptions to everyday plans and goals (e.g., losing a resource, being delayed in executing a plan step). Accordingly, people report unexpected events are not really that unexpected; the *expected unexpected*. This work shows that these are a class of unexpected events -- things that commonly go wrong – that are to be distinguished from "truly unexpected" events (as Foster & Keane, 2015, found for surprising events). These findings should prompt a re-assessment of what we mean by the "unexpected" as a dependent variable in exploring aspects of human thought. It also raises the interesting prospect, that there is a lot more to be discovered about what people conceive the unexpected to be.

Acknowledgements

The first author was supported by a scholarship from the School of Computer Science, University College Dublin, Ireland. Furthermore, this publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) and the Department of Agriculture, Food and Marine on behalf of the Government of Ireland under Grant Number **16/RC/3835.**

References

Foster, M. I., & Keane, M. T. (2015). Why some surprises are more surprising than others: Surprise as a metacognitive sense *of* explanatory difficulty. *Cognitive psychology*, *81*, 74-116.

Foster, M. I., & Keane, M. T. (2019). The Role of Surprise in Learning: Different Surprising Outcomes Affect Memorability Differentially. *Topics in cognitive science*, *11*(1), 75-87.

Harman, G. (1986). *Change in view: Principles of reasoning*. Cambridge, MASS: The MIT Press.

Hastie, R. (1984). Causes and effects of causal attribution. *Journal of Personality and Social Psychology*, *46*(1), 44.

Johnson, S. G., & Keil, F. C. (2014). Causal inference and the hierarchical structure of experience. *JEP: General,, 143(6),* 2223.

Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review, 93*(2), 136-153.

Khemlani, S. S., Sussman, A. B., & Oppenheimer, D. M. (2011). Harry Potter and the sorcerer's scope. *Memory & Cognition*, *39*(3), 527-535.

Leake, D. B. (1991). Goal-based explanation evaluation. *Cognitive Science, 15*(4), 509-545.

Leake, D. B. (1992). *Evaluating explanations: A content theory.* Hillsdale, NJ: Erlbaum

Lewis, D. K. (1986). *On the plurality of worlds.* Oxford: Blackwell.

Maguire, R., & Keane, M. T. (2006). Surprise: Disconfirmed expectations or representation-fit. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society*.
Maguire, R., Maguire, P., & Keane, M. T. (2011). Making sense of surprise. *JEP: LMC, 37(1),* 176-186.
McEleney, A., & Byrne, R. M. (2006). Spontaneous counterfactual thoughts and causal explanations. *Thinking & Reasoning, 12(2*), 235-255.
Meyer, W. U., Reisenzein, R., & Schützwohl, A. (1997). Toward a process analysis of emotions: The case of surprise. *Motivation and Emotion*, *21*(3), 251-274.
Quine, W. V. O. (1992). *Pursuit of truth*. Cambridge, MASS: Harvard University Press.
Schank, R.C. (1986). *Explanation patterns.* London: Psychology Press.
Schank, R. C., & Abelson, R. P. (1977). *Scripts, plans, goals, and understanding*. Hillsdale, NJ: Lawrence Erlbaum.
Schank, R. C., Kass, A., & Riesbeck, C. K. (Eds.) (1994). *Inside case-based explanation.* Hillsdale, NJ: Lawrence Erlbaum.
Teigen, K. H., & Keren, G. (2003). Surprises: Low probabilities or high contrasts? *Cognition, 87(2)*, 55-71.
Weiner, B. (1985a). An attributional theory of achievement motivation and emotion. *Psychological Review, 92,* 548.
Weiner, B. (1985b). "Spontaneous" causal thinking. *Psychological Bulletin, 97(1),* 74-84.

Table 2: Percentages of Responses Made for Various Measures, by Material, in Study 1 (N=127 responses per material)

Material	**% Top-3 Answer Cats.**	**No. of Answer Cats.**	**% Goal Obj.**	**% Non-Goal Obj.**	**% Both Obj.**	**% Neither Obj.**	**% S1 Goal Objs**	**% S2 Non-Goal Objs**	**% S3 Goal Objs**
alan_plane	55%	12	64%	7%	6%	24%	38%	11%	29%
anna_interview	74%	10	71%	6%	9%	15%	33%	12%	43%
belinda_meeting	58%	11	37%	28%	20%	15%	21%	32%	37%
bill_holiday	49%	12	48%	9%	2%	40%	23%	19%	58%
bob_job	81%	8	78%	0%	0%	22%	19%	0%	63%
edith_exam	50%	11	44%	20%	11%	25%	31%	20%	33%
john_party	67%	12	62%	1%	0%	37%	34%	1%	34%
karen_bus	57%	10	81%	2%	0%	17%	31%	1%	54%
mary_food	69%	8	80%	6%	2%	13%	64%	1%	29%
katie_kitten	50%	11	91%	1%	7%	1%	16%	11%	66%
louise_shopping	63%	10	64%	7%	6%	24%	21%	2%	68%
lucy_loan	89%	6	87%	0%	2%	11%	3%	7%	77%
michael_tea	56%	12	34%	35%	6%	24%	9%	38%	30%
peter_college	56%	12	59%	2%	1%	38%	37%	3%	27%
rebecca_swimming	54%	12	46%	6%	6%	43%	36%	8%	25%
robert_essay	54%	13	27%	41%	9%	23%	18%	38%	27%
sally_wine	54%	12	62%	9%	8%	21%	35%	13%	36%
sean_call	65%	10	61%	10%	6%	24%	31%	9%	44%
sam_driving	61%	11	64%	5%	9%	23%	31%	14%	35%
steve_gardening	55%	11	59%	9%	13%	18%	8%	20%	63%

Children's Sentential Complement Use Leads the Theory of Mind Development Period: Evidence from the CHILDES Corpus

Irina Rabkina (irabkina@u.northwestern.edu)
Constantine Nakos (cnakos@u.northwestern.edu)
Kenneth D. Forbus (forbus@northwestern.edu)
Qualitative Reasoning Group, Northwestern University
2233 Tech Drive, Evanston, IL 60208, USA

Abstract

Converging evidence suggests that children's linguistic and theory of mind (ToM) development are linked. Specifically, learning the sentential complement grammatical structure has been shown to play a causal role in the development of some false belief reasoning skills. Here, we extend this line of work to examine this relationship in the wild by means of a corpus analysis of children's speech during the typical period of ToM development. We show that children's use of the sentential complement grammatical structure increases immediately preceding the ToM development period and plateaus shortly thereafter. Furthermore, we find that parents' child-directed speech follows a similar pattern.

Keywords: theory of mind; corpus analysis; sentential complement

Introduction

Most researchers agree that humans' ability to reason about mental states, or their theory of mind (ToM), develops throughout early childhood, with the biggest increases seen during the preschool years, roughly age 3 to 5 (Wellman & Liu, 2004). Other developmental milestones during this time period, such as working memory capacity (Davis & Pratt, 1995), executive control (Perner & Lang, 1999), and language development (de Villiers & Pyers, 1997), have been linked as leading to the apparent improved ToM reasoning ability, either causally or as a side effect. Of these, perhaps the most studied is the role that children's developing language comprehension and production skills play in the development of their ToM (see Milligan, Astington & Dack, 2007).

While some researchers argue that improved language skills merely allow children to express previously-existing ToM concepts (e.g., He, Bolz, & Baillargeon, 2011), it is widely accepted that some interaction between language abilities and performance on ToM tasks exists. In fact, converging evidence suggests that the connection is causal: learning certain linguistic constructions, specifically the sentential complement, is instrumental in children becoming able to perform aspects of ToM reasoning that they were previously unable to perform (de Villiers & Pyers, 1997).

This evidence has taken multiple forms, including (1) a longitudinal study correlating sentential complement use with ToM reasoning ability (de Villiers & Pyers, 2002), (2) training studies that showed children who were trained on sentential complements improved performance on ToM tasks (Lohmann & Tomasello, 2003; Hale & Tager-Flusberg, 2003; Mo et al., 2014), and (3) a computational model of the mechanisms by which children learn ToM from sentential complements (Rabkina, McFate & Forbus, 2018).

Taken together, these studies provide evidence that understanding the sentential complement construction supports ToM development. If this is true, then children's understanding of the sentential complement should precede their ability to pass ToM tests. At a population level, this means that children's use of the sentential complement should begin to increase prior to the ToM development period and plateau by the end. However, prior research has focused on the relationship between children's ToM development and their sentential complement proficiency in a laboratory setting.

Here, we perform a corpus analysis of children's conversational speech (CHILDES; MacWhinney, 2000) to show that the hypothesized pattern exists in the wild. We find that the expected pattern emerges: children's sentential complement use begins just prior to 2 years of age and plateaus around 3 years—just as the ToM development period begins. Furthermore, child-directed speech follows a similar trajectory during the same time period; that is, parents increase their sentential complement use in tandem with their children. These findings support the argument that learning the sentential complement grammatical construction plays an important role in developing ToM reasoning abilities.

We begin with a review of prior work linking ToM development and sentential complement use. We then describe our approach to the corpus analysis and present our findings. We conclude by situating these findings in the context of prior work and outlining steps for future investigation.

Background

A sentence contains a sentential complement if a verb in that sentence takes a full clause as its argument. For example, in the sentence, "Sarah thought the Earth was flat," the clause "the Earth was flat" is an argument to the verb "thought." Crucially, the truth value of the clause is independent of the truth value of the sentence as a whole—the Earth not being flat does not change the fact that Sarah thought it was. De Villiers and colleagues (e.g. de Villiers & Pyers, 1997; de Villiers & de Villiers, 2003) have argued that learning the sentential complement, and the potential

difference in implied truth values between the statement and the embedded clause, is key to ToM development.

Converging evidence supports such a conclusion. In a longitudinal study, de Villiers & Pyers (2002) found a strong correlation between children's performance on a task that measured understanding of sentential complements and their performance on three classic ToM tasks. A hierarchical regression analysis further showed that performance on the understanding of complements task accounted for a significant amount of variance in the ToM tasks, regardless of the order in which variables were presented in the regression. Importantly, this finding was not bidirectional—ToM performance did not predict performance on the sentential complements task.

Intervention studies suggest that the relationship found by de Villiers and Pyers (2002) is causal. Lohmann and Tomasello (2003), Hale and Tager-Flusberg (2003), and Mo et al. (2014) found that sentential complement training leads to improved performance on ToM post-tests in children who failed both sentential complements and ToM pre-tests. Furthermore, Hale and Tager-Flusberg (2003) found that ToM training did not affect performance on sentential complements post-tests, which provides additional evidence that the effect is causal and unidirectional.

Rabkina et al. (2018) proposed a process-level computational model of the effect of sentential complement training on ToM understanding. They argued that, in learning to interpret the sentential complement grammatical structure, children learned a representation that allowed them to separate the truth value of beliefs from reality, analogously to separating the truth value of the sentential complement and the overall statement.

The combination of these studies tells a compelling story of the relationship between ToM development and the sentential complement. However, while the connection has been shown in the laboratory, the story may be different in an everyday setting. Previous work (Koder, 2016) has looked at the developmental trajectory of verbs for reported speech as they appear in children's natural language production in Dutch and German. Others (Gordon & Nair, 2004) have examined more general language use during the ToM development period via corpus analysis. However, to the best of our knowledge, no previous work has addressed the question of sentential complement use in naturally occurring speech.

Here, we perform a corpus analysis of child-directed and child-produced sentential complement use during and immediately preceding the ToM development period. Our results provide further evidence of a link between learning the sentential complement grammatical structure and ToM development.

Approach

If learning the sentential complement grammatical structure bootstraps the development of ToM reasoning skills, then this pattern should hold outside of the laboratory. That is, children's use of the sentential complement in everyday speech should anticipate the developmental trajectory of ToM. Because significant improvements in children's ToM occur between approximately 3 and 5 years of age (Wellman & Liu, 2004), we expect sentential complement use to reach a critical threshold immediately preceding this age range.

To test whether this relationship holds, we performed a corpus analysis of children's use of the sentential complement between 12 and 90 months of age. We also analyzed sentential complement use in child-directed speech (produced by mothers) during the same timeframe.

All data were extracted from the CHILDES project (MacWhinney, 2000), which contains over 130 corpora of child-directed and child-produced speech. A corpus was included in our analysis if it contained speech by a typically developing North American English-speaking child between the ages of 12 months and 90 months. For consistency, only corpora with an available transcript and dependency parse data (Sagae et al., 2007) were included in the analysis. This resulted in a total of 32 corpora, leading to 3982 individual data points[1].

Each corpus included one or more conversations between a child and one or more adults. All conversation transcripts provided the child's age in months and relationship to the adult interlocutor(s) (i.e., mother and/or experimenter).

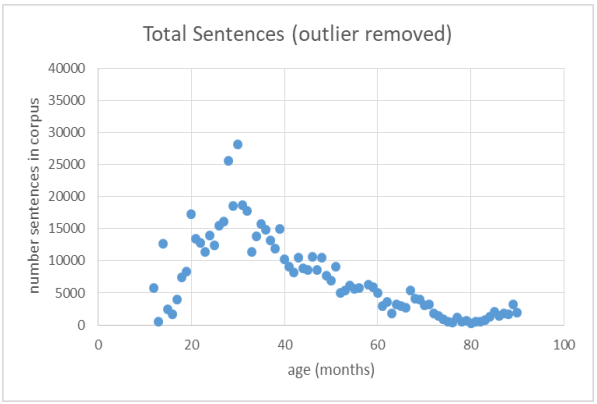
We extracted sentential complements from the children's speech using the "COMP" (finite verb complement) and "XCOMP" (other verb complement) dependency parse tags. Sagae et al. (2007) report overall parse accuracy for children's utterances between 72.7% and 92.3% on varying corpora within CHILDES. Table 1 shows reported precision, recall, and F-score for the "COMP" and "XCOMP" tags in the Eve corpus (Brown, 1973). Overall parse accuracy for the Eve corpus is 92.0%. Note that these analyses include both child and adult utterances.

Because a causal relationship between learning the sentential complement and developing ToM reasoning abilities has been proposed (e.g., de Villier & Pyers, 1997), we expected children's use of the sentential complement to lead their ToM development. To examine this effect, we computed the average number of sentential complements produced per sentence at each age in months. If learning the sentential complement bootstraps ToM reasoning, then children should show an increase in sentential complement use leading into the ToM development period. Moreover, the increase should be specific to this timeframe; that is, children should achieve sentential complement proficiency prior to finishing ToM development.

Table 1: Statistics for COMP and XCOMP tags (Sagae et al., 2007)

	Precision	*Recall*	*F-score*
COMP	0.83	0.86	0.84
XCOMP	0.86	0.87	0.87

[1] For longitudinal studies, a new data point was included for each recorded age in months.



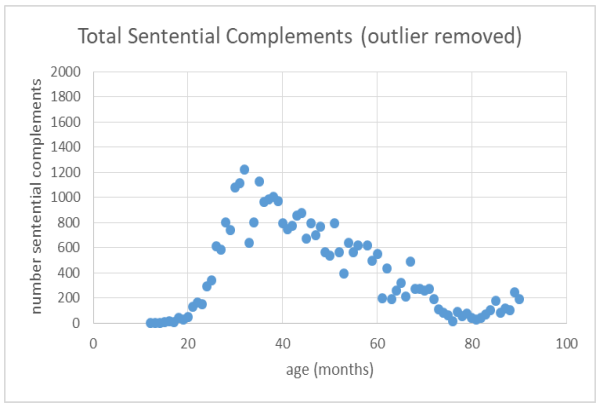


Figure 1: Counts for total sentences (left) and total sentential complements (right) in our corpus at each age in months. Note that one outlier (57 months) was removed from each graph.

Results

Our results indicate a concentrated growth period for children's sentential complement use that begins to plateau at the beginning of the ToM development period, suggesting a causal relationship between the two. Furthermore, this period of increasing sentential complement use coincides with a similar period found in parents' child-directed speech, which suggests a critical role for parents in children's acquisition of this grammatical structure.

Figure 1 shows the total number of sentences in our corpus of child-produced speech at each age in months along with the corresponding counts of sentential complement use. The corpus contains the most data in the range from 25 to 60 months. Note that this is an artifact of the data available and does not necessarily represent an increase in overall speech production during this age range.

Figure 2 shows children's sentential complement production as a proportion of overall sentences produced at a given age. The graph shows a linear increase from approximately 20 months to approximately 40 months of age, with a plateau beginning shortly thereafter. Once this baseline level of sentential complement production is reached, variance visibly increases. However, this variance is likely a byproduct of noise due to lower total sentence counts at later ages (Figure 1).

To determine the period of most concentrated sentential complement development, we isolated the interval with the strongest linear correlation between age and proportion of sentential complements (Figure 3, left). We fixed the starting point at 22 months, the first instance of appreciable sentential complement use (>1%). An endpoint of 38 months produced the strongest correlation, r^2=0.9217, p<0.001. Beginning at 39 months, the distribution plateaus with a slope of approximately 0 (Figure 3, right).

Child-directed adult-produced speech follows a similar pattern (Figure 4). Following a period of linear increase from child's age 12 months to 38 months (r^2=0.8603, p<0.001, Figure 5), sentential complement use peaks and begins to gradually decline. Notably, the absolute

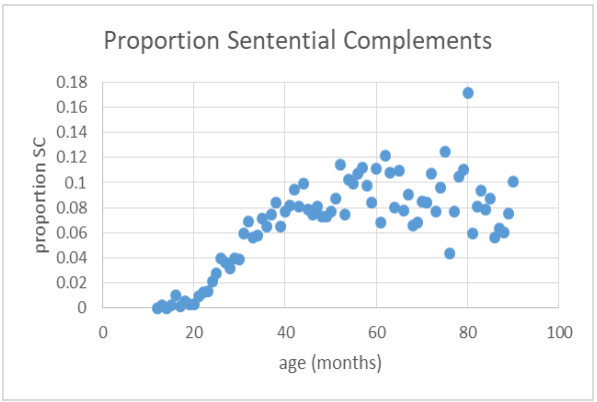


Figure 2: Average number of sentential complements per sentences produced by children at each age in months. No outliers were excluded.

proportion of sentential complements per sentence produced by adults is higher than the proportion produced by children at almost all ages.

As a potential contrast to the sentential complement, we also examined the use of another complex grammatical structure that has been argued to influence ToM acquisition, the relative clause (e.g., Smith, Apperly & White, 2003). However, we found negligible use of the relative clause in both child-produced and child-directed speech. This is consistent with a prior analysis of longitudinal data (Diessel & Tomasello, 2000) which found that children use the relative clause in less than 0.5% of utterances. Absent a direct increase in the use of another such structure in child-produced speech during this period, the sentential complement stands out as the best candidate for a syntactic aid to ToM development.

Discussion

As predicted, children reach a critical threshold of sentential complement use prior to entering the major period of ToM development, typically regarded as 3 to 5 years of age. By

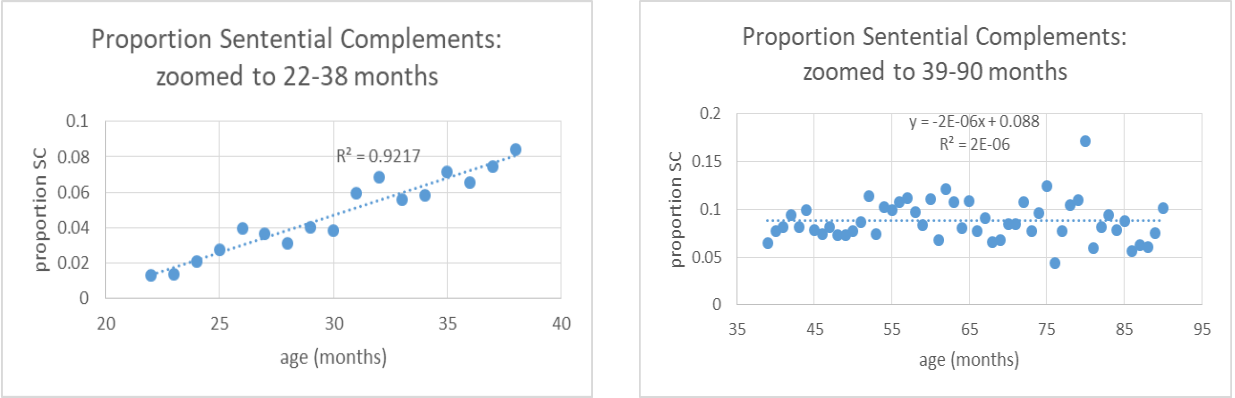


Figure 3: Proportion of sentential complement use by children at each age, zoomed to period of growth (left) and stabilization (right).

36 months children use sentential complements in an average of 6.5% of sentences (Figure 1). Their sentential complement use begins to plateau shortly thereafter, at 38 months and 8.4%.

It is important to note that both the ToM development period and the beginning of the observed plateau in sentential complement use are not hard boundaries. In fact, sentential complement use continues to increase after the onset of the plateau (between 39 and 58 months; r^2=0.3581, p=0.005; Figure 3, right), albeit at a much reduced rate. However, weak correlation and high variance make it difficult to draw firm conclusions about trends within the plateau.

What is clear is that the most concentrated growth occurs before children make significant strides in their ToM development. Previous work has shown that training children to understand the sentential complement leads to improved ToM reasoning skills in a laboratory setting (Lohmann & Tomasello, 2003; Hale & Tager-Flusberg, 2003; Mo et al., 2014). Our results suggest that the same effect occurs outside of the laboratory. Taken together, these findings support the hypothesis that mastery of basic sentential complement use sparks ToM development.

Another finding of note is that child-directed sentential complement use shows a similar pattern of increase to child-produced sentential complement use. Specifically, adult sentential complement use increases from 7.0% at child's 12 months to 16.0% at child's 38 months. This period subsumes the interval of greatest sentential complement development in children and gives way to a period of decline as children's use plateaus. Parents seem to adjust their sentential complement use according to the child's level of proficiency. Moreover, parents' sentential complement use seems to promote sentential complement production in children, as parents consistently overproduce compared to children at a given age.

Several explanations could account for the observed behavior. First, it is possible that parents mirror their children's speech patterns: as the child increases her sentential complement use, so does the parent. Under this hypothesis, other grammatical constructions should follow a similar trajectory. Alternatively, the causality could flow in

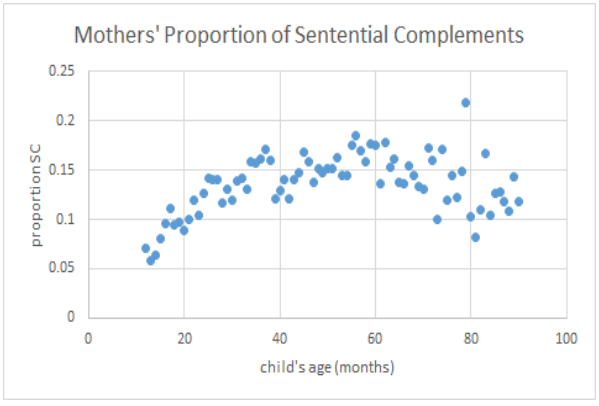


Figure 4: Average number of sentential complements per sentence produced by mothers at child's age in months. No outliers were excluded.

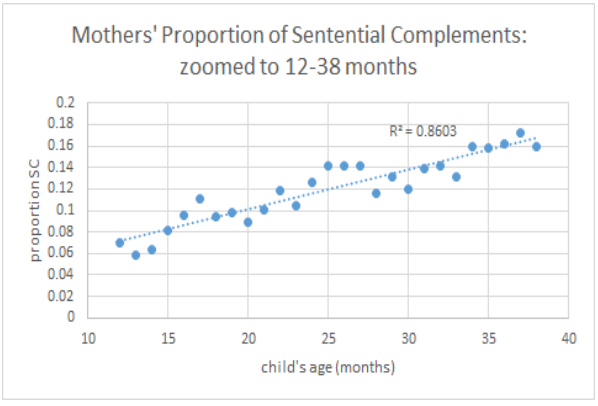


Figure 5: Average number of sentential complements per sentence produced by mothers at child's age in months, zoomed to period of increase.

the opposite direction, with children mirroring their parents. This explanation follows more directly from the present data, since the parents' sentential complement use precedes the children's, but it does not explain why the parents' use increases. Yet another explanation could be a mutual influence effect between children and their parents. As children begin to use the sentential complement, the parents increase their usage of the grammatical form, pacing their children's learning. Identifying the exact relationship at play will require data that can clarify the interaction between children's language use and their parents'.

Overall, our findings paint a picture of parental influence on children's sentential complement development, leading to children's acquisition of ToM. While the corpus analysis is not stand-alone proof of a relationship between sentential complement proficiency and ToM development, it is consistent with prior laboratory evidence of a causal link between the two. This is a step toward showing that such a link exists in the wild.

Limitations

One goal of this paper is to provide evidence in support of the hypothesis that sentential complement acquisition causally drives ToM development. While the evidence presented here supports such a relationship, it is not sufficient to establish causality for two reasons. First, as a correlational study, this can only point to likely interactions and cannot confirm their directionality or factor out potential confounds. Second, our analysis takes the ToM development period as a given and does not examine ToM effects directly.

These limitations mean that our findings cannot be used to draw broad conclusions about the interaction between language and cognition. The observed patterns could arise from effects that contradict the linguistic determinism hypothesis but are not accounted for in the available data. In particular, the lack of explicit ToM performance data means that any conclusions about ToM drawn from this dataset must be based on independently motivated developmental theories. For example, some researchers have found evidence that infants exhibit behaviors consistent with some understanding of ToM (e.g., Baillargeon, Scott & He, 2010). It is unclear how to reconcile such findings with the patterns observed here.

Another caveat to our findings is the potential for noise in the dependency parses we use. Though the analysis in Sagae et al. (2007) shows adequate performance of their parses on the CHILDES dataset (see Approach section for detailed overview), manual inspection showed instances where the dependency parse was inaccurate. It remains to be seen how the overall performance of the parser relates to the specific corpora used in our analysis.

Future Work

This paper considered the relationship between children's sentential complement use and their ToM development. However, evidence exists that a more granular view of the sentential complement might be appropriate. For example, Mo et al. (2014) found that, on ToM post-tests, children trained with sentential complements involving communication verbs outperformed children who were trained with mental state verbs. They note that this may be an artifact of the language used in the study, Mandarin, rather than a more general effect. On the other hand, Hale and Tager-Flusberg (2003) included only communication verbs in their training study of English-speaking children because of the potential confounding factor of the semantics carried by mental state verbs. A deeper analysis of the types of verbs used by children as they learn the sentential complement could shed some light on this question.

Because the effects of sentential complement training on ToM performance have been observed cross-linguistically, it is worth examining whether the patterns found in the present study are consistent across languages as well. Shatz et al. (2003) showed that 3- and 4-year-old speakers of languages with explicit false belief markings outperformed speakers of languages without such markings on some ToM tests. This suggests that other linguistic effects may be at play, and that the sentential complement may not be the sole way ToM is encoded in linguistic structure. For such languages, it is possible that the pattern of sentential complement use found in English may be less strong or entirely nonexistent.

Another question that merits further investigation is the nature of the plateau observed in Figure 2 and Figure 3 (right). A cursory analysis shows a period of continued increase from 39 months to 58 months before a period of mild decrease lasting through the end of the included data. The variance in the available data at this age range precludes a more concrete analysis, but the coincidence of the period of sustained increase in sentential complement use and the period of ToM development points to a tighter connection than can be shown at present.

Current data also does not fully illuminate the relationship between children's sentential complement use and that of their parents. It is curious that the adult-produced speech so closely parallels the patterns observed in children's speech. However, identifying the exact mechanism by which this arises would require paired data to more closely track changes in sentential complement use.

Finally, the questions raised in this paper tie into a broader debate about ToM acquisition as a whole. Although we provide evidence that is consistent with the hypothesis that sentential complement proficiency facilitates ToM development, strict causality has yet to be proven. Further research is required to fully explore this connection.

Acknowledgements

We thank Dedre Gentner, Jongmin Lee, and Jason Wilson for helpful comments and discussions. This research was supported by the Socio-Cognitive Architectures for Adaptable Autonomous Systems Program of the Office of Naval Research, N00014-13-1-0470.

References

Baillargeon, R., Scott, R. M., & He, Z. (2010). False-belief understanding in infants. *Trends in cognitive sciences*, 14(3), 110-118.

Brown, R. (1973). *A first language: The early stages*. Harvard U. Press.

Davis, H. L., & Pratt, C. (1995). The development of children's theory of mind: The working memory explanation. *Australian Journal of Psychology*, 47(1), 25-31.

De Villiers, J. G., & de Villiers, P. A. (2003). Language for thought: Coming to understand false beliefs. *Language in mind: Advances in the study of language and thought*, 335-384.

De Villiers, J., & Pyers, J. (1997). Complementing cognition: The relationship between language and theory of mind. *In Proceedings of the 21st annual Boston University conference on language development* (Vol. 1, p. 136). Cascadilla Press.

De Villiers, J. G., & Pyers, J. E. (2002). Complements to cognition: A longitudinal study of the relationship between complex syntax and false-belief-understanding. *Cognitive Development*, *17*(1), 1037-1060.

Diessel, H., & Tomasello, M. (2000). The development of relative clauses in spontaneous child speech. *Cognitive Linguistics*, 11(1/2), 131-152.

Gordon, A. and Nair, A. (2004) Expressions related to knowledge and belief in children's speech. *Proceedings of the 26th Annual Meeting of the Cognitive Science Society*.

Hale, C. M., & Tager-Flusberg, H. (2003). The influence of language on theory of mind: A training study. *Developmental science*, *6*(3), 346-359.

He, Z., Bolz, M., & Baillargeon, R. (2011). False-belief understanding in 2.5-year-olds: evidence from violation-of-expectation change-of-location and unexpected-contents tasks. *Developmental science*, 14(2), 292-305.

Koder, F. M. (2016). Between direct and indirect speech. Doctoral dissertation.

Lohmann, H., & Tomasello, M. (2003). The role of language in the development of false belief understanding: A training study. *Child development*, *74*(4), 1130-1144.

MacWhinney, B. (2014). *The CHILDES project: Tools for analyzing talk, Volume II: The database*. Psychology Press.

Milligan, K., Astington, J. W., & Dack, L. A. (2007). Language and theory of mind: Meta-analysis of the relation between language ability and false-belief understanding. *Child development*, 78(2), 622-646.

Mo, S., Su, Y., Sabbagh, M. A., & Jiaming, X. (2014). Sentential complements and false belief understanding in Chinese Mandarin-speaking preschoolers: A training study. *Cognitive Development*, *29*, 50-61.

Perner, J., & Lang, B. (1999). Development of theory of mind and executive control. *Trends in cognitive sciences*, 3(9), 337-344.

Rabkina, I., McFate, C., & Forbus, K. D. (2018). Bootstrapping from Language in the Analogical Theory of Mind Model. In *Proceedings of the Fortieth Annual Conference of the Cognitive Science Society*.

Sagae, K., Davis, E., Lavie, A., MacWhinney, B., & Wintner, S. (2007, June). High-accuracy annotation and parsing of CHILDES transcripts. In *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition* (pp. 25-32). Association for Computational Linguistics.

Shatz, M., Diesendruck, G., Martinez-Beck, I., & Akar, D. (2003). The influence of language and socioeconomic status on children's understanding of false belief. *Developmental Psychology*, 39(4), 717.

Smith, M., Apperly, I., & White, V. (2003). False belief reasoning and the acquisition of relative clause sentences. *Child Development*, 74(6), 1709-1719.

Wellman, H. M., & Liu, D. (2004). Scaling of theory-of-mind tasks. *Child development*, 75(2), 523-541.

When Does a Reasoner Respond: Nothing Follows?

Marco Ragni* (ragni@cs.uni-freiburg.de), **Hannah Dames*(damesh@cs.uni-freiburg.de)**,
Daniel Brand (daniel.brand@cognition.uni-freiburg.de), **Nicolas Riesterer (riestern@cs.uni-freiburg.de)**,
Cognitive Computation Lab, Department of Computer Science, Georges-Köhler-Allee 52, 79110 Freiburg, Germany

Abstract

When does a reasoner respond that "no valid conclusion" (NVC) follows in a syllogistic reasoning task? Cognitive theories aim to trace it back to theory specific inference processes. In contrast, systemic theories explain it by depleted cognitive resources among others. This paper investigates possible theories to explain NVC responses in an experiment with 139 participants. Using mixed models we analyze the association of NVC responses with reaction times, the validity as well as the entropy of a syllogism, and how NVC responses change over time. As expected, the number of NVC responses is lower than logically expected, participants respond NVC more often for invalid syllogisms, and the likelihood to respond NVC increases over the time-course of the experiment. Surprisingly, however, only for valid syllogisms, are the entropy and the RTs associated with NVC responses. Consequently, for invalid syllogisms, NVC responses seem to be generated differently as compared to valid ones.

Keywords: Reasoning; NVC; cognitive theories; logic; valid; invalid

Introduction

The psychology of reasoning investigates when and which conclusion is derived from given information. This includes the case when *no conclusion* can be drawn because the information is insufficient or it is too difficult to make an inference. The domain of syllogistic reasoning is probably the best researched domain with most published theories (for an overview see Khemlani & Johnson-Laird, 2012). A *syllogism* consists of two quantified statements. Each statement is formed using one of four quantifiers: All (A), Some (I), Some ... not (O), or None (E). Consider the following syllogistic reasoning problem:

(AA4) All beekeepers are architects.
All beekeepers are chemists.
What, if anything, follows?

The task is to generate a quantified answer using one of the quantifiers A, I, O, E about the two terms architects (A) and chemists (C, in any direction) or to conclude that no logically valid conclusion (*NVC* for short) can be made. Four different arrangements of the terms in the premises, called figures, are possible. The example above, for instance, is a type 4 figure (B-A, B-C). The four quantifiers for each of the two premises times four figures sum up to 64 possible syllogistic problems. Each syllogism can be encoded by a string, describing the quantifiers of the two premises as well as the relation of the used items in a figure. Hence, the syllogism above can be succinctly written as AA4. For the problem (AA4) most of the participants (49% in Khemlani & Johnson-Laird, 2012) infer that *All architects are chemists* and only 16% give one of the logical correct answer that *Some architects are chemists* or that *Some chemists are architects*. However, about 22% of the participants in the metaanalysis (Khemlani & Johnson-Laird, 2012) respond that NVC follows. A *logically valid* problem is one where by applying a logical calculus such as first-order logic allows to infer a conclusion (such as the syllogistic example AA4 above where *Some architects are beekeepers* is one). If this cannot be inferred then it is called *invalid* problem (and the only logical correct answer is NVC). Past research both from a statistical and from a modeling perspective has strongly focused on the case when an inference can be drawn (Oaksford & Chater, 2007; Johnson-Laird, 2006; Costa, Saldanha, Hölldobler, & Ragni, 2017) but less on the case when *no logically valid conclusion* can be inferred. Yet, it is exactly this response that stands out from the rest: Not only is the response NVC a *different class of response*, namely stating that no other conclusion follows, but it is the NVC response, that is the *most frequently observed response* in experiments (Khemlani & Johnson-Laird, 2012). In the current work, we aim to fill the gap of investigations on NVC responses by systematically investigating *when* people respond NVC. In particular, we compare different approaches to explain NVC responses by analyzing experimental data.

When is an NVC response given?

Syllogistic theories have been categorized as heuristic, rule-based, and model-based approaches (Khemlani & Johnson-Laird, 2012): Only few cognitive theories in syllogistic reasoning predict the NVC conclusion at all (e.g., Mental Model Theory, Verbal, Conversion). If a theory does so, it often implies that individuals give NVC as a last-resort, when the inference process yields nothing else (e.g., Mental Logic; Rips, 1994). Most of the heuristic theories do not predict NVC responses, with a rare exception in the case of Conversion and the probabilistic heuristic model (PHM, Oaksford & Chater, 2007, but see Copeland, 2006) that can be extended to predict NVC. The Atmosphere (Woodworth & Sells, 1935) and Matching (Wetherick & Gilhooly, 1995) theories derive only the quantifier in the response from the premise quantifiers. Hence, they do not consider and cannot explain NVC responses. This is remarkable as in the case of syllogisms there are 37 invalid problems (58% of all syllogisms) that would require from a normative logical perspective NVC as the correct response.

In sum, while there are at least twelve cognitive theories about syllogistic reasoning (Khemlani & Johnson-Laird, 2012), there is no explicit cognitive reasoning theory beyond explaining it by a search through the theory specific inference mechanism (e.g., by applying all inference rules or the generation of all models). Beyond explaining NVC by *cognitive reasoning theories*, *systemic theories* can provide alternative accounts emphasizing the role of behavioral response tendencies within experiments. Among others these systemic hypotheses include phenomena such as mental depletion (i.e., with each syllogism the cognitive resources are depleted (e.g., Schmeichel & Vohs, 2018) or cognitive load

*Both authors contributed equally to this manuscript.

(e.g., Sweller, 1994), which lead an individual to stop reasoning as soon as the problem becomes too difficult, or a general aversion to respond NVC (e.g., NVC is interpreted as "giving-up"). These hypotheses including their assumed, underlying cognitive processes are briefly summarized in Table 1. The aim of the current paper is to investigate such cognitive and systemic hypotheses in explaining why a NVC response is given. For the above-mentioned differences in the ability to predict NVC within heuristic theories, we focus on the MMT and mental logic theory in the current paper. Thereby, we aim to provide novel and much needed insights into when participants respond *no valid conclusion*.

Theories, predictions, and hypotheses

We have identified distinctive cognitive theories and systemic hypotheses that can explain when a reasoner responds NVC. In this section, we briefly outline these theories as well as hypotheses and draw implications on five observable patterns of NVC: first, the *response time* (RT) and the *frequency* of NVC response. Furthermore, we investigate the influence of valid and invalid syllogisms (*validity*) and the problem's entropy. The entropy measure (Shannon & Weaver, 1963) has been applied to measure the response diversity of each syllogism (Khemlani & Johnson-Laird, 2012). For each syllogism per study, the authors computed the probability with which each conclusion was drawn and aggregated the probabilities using Shannon's measure. The response diversity demonstrates an uncertainty of reasoners about which conclusion has to be drawn. Last, we analyze how the likelihood to respond NVC changes over the course of 64 syllogisms.

Theory of Mental Models (MMT). The MMT (e.g., Johnson-Laird, 2006) postulates a two-stage process based on the generation of an initial model and a flesh-out process that tests a putative conclusion formed on the initial model by a search for counter-examples (e.g., Bucciarelli & Johnson-Laird, 1999). If the flesh-out process does not yield a conclusion, the reasoner responds NVC. The latest implementation, mReasoner[1], contains a specific parameter that guides the generation of counter-examples. MMT makes the following predictions. *RT of NVC*: On average, the MMT predicts the NVC response at the end of the inference process, hence, responding NVC requires more cognitive steps and thus more time (especially, in the case of multiple model problems, i.e., problems that are invalid). *Frequency of NVC*: The inference process described before, however, can sometimes fail or be stopped early. As a result, not in all cases counter-examples are searched for and putative conclusions are drawn, even in cases where NVC hold. Consequently, less NVC responses are given as required by formal logic. *Entropy and NVC*: In indeterminate cases, the flesh-out process becomes relevant, hence, the more difficult a problem is or the more uncertainty it causes (measured by the entropy), the more NVC responses will be generated. *Time-course of NVC*: The more syllogisms are solved, participants enter more likely the flesh-out process (reasoners become more logical, as it has been recently modeled in mReasoner; Ragni, Riesterer, Khemlani, & Johnson-Laird, 2018). Consequently, participants are more likely to respond NVC for invalid syllogisms over time.

[1] https://mentalmodels.princeton.edu/models/mreasoner/

Theory of Mental Logic (ML). The theory of ML (Rips, 1994) is based on the application of first-order formal inference rules together with the inclusion of Gricean implicature to capture differences between a formal and an everyday language understanding of existential quantifiers. As it is based on formal logic rules, the conclusions are valid and no erroneous results will be predicted. The theory proposes that the erroneous responses generated by human reasoners are due to problems in the recognition, retrieval, or application of the formal rules (Rips, 1994). Following predictions can be derived: *RT of NVC*: ML predicts NVC, if the full application of the inference mechanism does not yield a conclusion. This takes longer than the application of some inference rule in the valid case. *Frequency of NVC*: An NVC response is found in the invalid cases and not in the valid cases. *Entropy and NVC*: A connection has not been reported and so we do not assume a predicted difference. *Time-course of NVC*: The mental logic does not assume a change across time.

Predictions of Mental Depletion. Theories of resource depletion (e.g., Schmeichel & Vohs, 2018) assume that mental activities such as reasoning can deplete cognitive resources. This results in an increase in NVC responses over time due to depletion. This increase appears for valid and invalid syllogisms - due to the depleted cognitive resources. A simple depletion model makes no distinction between logically valid and invalid problems. While combinations with cognitive theory can be thought of, we solely focus on the case where more NVC responses are given over time. Predictions: *RT of NVC*: No effect of NVC-responses on RTs is expected. Depletion processes may result in either generally higher or lower RTs over the course of an experiment, but regardless of an NVC response. *Frequency of NVC*: There are no concrete predictions. *Entropy and NVC*: Entropy has no implications on NVC, but instead generally enhance mental depletion. *Time-course of NVC*: Mental depletion is assumed to strengthen throughout an experiment. Thus, NVC responses should increase across the course of solving the 64 problems, respectively.

Predictions of Early Stoppers. Some syllogisms are more difficult than others and thus require additional cognitive resources. For some syllogisms, reasoners may stop the reasoning process early avoiding the mental effort required by analytic processes by responding NVC. While the application of heuristics would not result in an NVC answer, the early stopping process does (NVC as a last resort). Early stoppers do not necessarily make a distinction between valid and invalid problems as for both types problems with a high entropy exists. Following predictions are derived: *RT of NVC*: An early stopper does not need longer for an NVC response. *Frequency of NVC*: Both valid and invalid problems can be difficult to solve. Therefore, the early stopper hypothesis predicts generally more NVCs as there are logically correct NVC responses. *Entropy and NVC*: Higher entropy resembles a higher uncertainty with the problem at hand, which may lead to more NVC responses the higher the entropy. *Time-course of NVC*: The time-course has no effect.

Predictions of NVC aversion. Logically naive reasoners may interpret responding NVC as "giving up" (similar to the last-resort option as it is assumed in many theories). While participants may

even fear to be regarded as less intelligent or ignorant, they may (at least in the beginning) tend to avoid this answer. The following predictions can be made: *RT of NVC*: NVC aversion leads to higher RTs for NVC responses as the deliberation processes to exclude all other response is time-consuming. *Frequency of NVC*: As NVC is avoided, fewer NVC responses as there are logically correct ones are made. *Entropy and NVC*: It is unclear whether Entropy may have an effect. *Time-course of NVC*: The aversion for NVC may diminish over time due to exposition to invalid syllogisms or because the reasoner learns that some syllogisms do not have a valid response. Hence, NVC responses increase over time.

Hypotheses

The introduced theories and hypotheses differ on predictions for response times, frequency of NVC answers, entropy, and the time-course of NVC. Based on these predictions, we will derive five general hypotheses. The presented cognitive theories and hypotheses do explain an NVC response in one of two ways: by the application of the complete inference mechanism that does not yield any valid conclusion or by a model-based search that yields counter-examples to any putative valid conclusion. This implies, however, that more steps are necessary to infer that nothing follows than to infer that something follows. More cognitive steps, however, require more time. This leads to our first hypothesis: *Hypothesis 1: The RTs significantly increase in trials where a NVC response is given as compared to non-NVC trials.*

Cognitive reasoning theories assume that NVC is a response typically generated after the application of inference rules or through the search through all counter-examples. This process is not necessarily always entered resulting in the miss of NVC responses. Thus: *Hypothesis 2: The number of NVC responses is lower than the number of logically correct NVC responses.*

Since validity is a logical concept, cognitive theories that are closer to logic make a difference between them. Hence, we get as a corollary hypothesis: *Hypothesis 3: The number of NVC responses is lower in the case of valid problems than in the case of invalid problems.*

Moreover, if it is more likely for a reasoner to respond NVC, if there is greater uncertainty operationalized by entropy. *Hypothesis 4: The higher the entropy of a syllogism the higher the likelihood of an NVC response.*

A fifth hypothesis is that across an experiment participants may increasingly respond NVC, which can depend both on cognitive (e.g., MMT) and systemic hypotheses (e.g., mental depletion): *Hypothesis 5: There is an increase in NVC responses across solving more problems.*

The different predictions of the cognitive reasoning theories and systemic hypotheses are summarized in Table 1. In the next section we report experimental data and the analysis.

Table 1: The hypotheses and predictions of the cognitive theories and the systemic factors.

Theories	**Prediction**				
	RT H1	NVC H2	Validity H3	Entropy H4	Time H5
Mental Model	y	y	y	y	y
Mental Logic	y	n	y	n	n
Mental Depletion	n	?	n	n	y
Early Stopper	n	n	n	y	n
NVC aversive	y	y	n	?	y

Explanation of the abbreviation y = the theory predicts yes; ? = the theory does neither predict yes nor no; n = the theory predicts no.

Experiment

Method

The experiment tested 204 participants (125 female and 79 male) on Amazon's Mechanical Turk[2]. They received a nominal fee for their participation. Participants or trials were excluded based on the following criteria: First, in order to identify non-compliers, data from participants that are at or below guessing level were discarded. The cutoff point of 18.8% (n = 64) is calculated as the cumulative binomial probabilities of 1/9 (for 9 possible conclusions) for 64 correct responses. That results in twelve problems correct for the α-value of .05 according to the binomial distribution. Second, trials with exceptionally long response times (RT) were excluded from the analyses: RTs exceeding 10 minutes (n = 1) and RTs deviating more than 3 standard deviations (SDs) from the individual mean RT separated for valid vs invalid syllogisms (n = 147, 1.7% of remaining trials). Last, the first four trials of the experiment were excluded as the four first trials always consisted of the same syllogisms for practice purposes (n = 546). Thus, 139 participants and 8202 observations were included in the following analyses.

Each participant had to select a conclusion from all possible nine response options for all 64 syllogisms (selection task). The order of the problems was randomized for each participant, except that the problems, AA1, AI2, EA3 and IA4, were always presented first in a randomized order, so that participants can familiarize themselves with the experiment. In addition, four single-premise syllogisms (of the four different possible quantifiers) were used as practice trials. Participants received two assertions similar to problem AA4 above. Content was randomly assigned to all 64 syllogisms (thus, valid and invalid problems received the same content with similar premise lengths of the resulting premises). For each set they had to determine which eight possible conclusions logically follow from the assertions by pressing one of the eight keys: 1-4 (the respective quantifier with the conclusion direction A-C) and 7-0 (the respective quantifer with the conclusion direction C-A). If no logical conclusion could be found, participants had to press the space bar. There were eight presentation orders of the conclusion quantifiers to reduce the presentation order effect. Each participant received the same response option order throughout the whole experiment. They could take as much time as they needed, but responses within a second were prohibited.

[2] https://www.mturk.com

Results

The overall percentage of logically correct responses per participant was 38.7% (*SD* = 19.0%), for the 27 syllogisms with valid conclusion(s) 42.1% (*SD* = 15.3%) and for the 37 syllogisms without a valid conclusion (NVC syllogisms) 36.5% (*SD* = 27.1%). On average, for valid syllogisms, participants gave 16.9% NVC responses (*SD* = 16.6%) and 36.5% (*SD* = 27.1%) for invalid syllogisms.

Analysis. Participants' frequency of NVC-responses differed between individuals (*M* = 29.0%, *SD* = 21.4%). In fact, there were a few participants that did not give any ($n = 8$) or less than 10 ($n = 48$) NVC responses. In the following analyses we used (generalized) linear mixed models (short (G)LMM; for an overview see Baayen, Davidson, & Bates, 2008; Judd, Westfall, & Kenny, 2012) as they can handle incomplete and unbalanced data and can account for the multi-level structure of the designs (e.g., multiple measures per participant). GLMMs were analyzed using the lme4 package (Bates, Mächler, Bolker, & Walker, 2015, Version 1.1.19) in the R environment. Models were fit via maximum likelihood (ML). Effect coding was used for all dichotomous fixed effects. Denominator degrees of freedom and *p*-values were estimated via Satterthwaite corrections implemented via lmerTest (Kuznetsova, Brockhoff, & Christensen, 2017, Version 3.0.1). Furthermore, the significance of fixed effect on the model fit was obtained by step-wise removing a fixed effect from the full model and testing whether the exclusion of the variable resulted in a significant loss of the goodness of fit as indicated by likelihood ratio tests and by comparing the Bayesian information criterion (BIC) and the Akaike information criterion (AIC). The reported tables (Table 2 and 3) show the results for the best models.

The analysis of reaction times. For the analysis of RTs, there is currently a debate about whether or not dependent variables should be transformed (Lo & Andrews, 2015). It has been suggested to use GLMMs on the raw RTs to analyze non-normal data that involve random effects (Lo & Andrews, 2015). Here, were use Inverse Gaussian distributions to account for the distinct positive skewed distribution of the continuous, raw RTs (for an overview of this approach, see Lo & Andrews, 2015). However, this approach resulted in a significantly worse fit ($\chi^2 = 156020$, $p < .001$) than the standard logarithmic approach using LMMs (where RTs were logarithmically transformed prior to analyses). As we report only the best-fit models, we therefore only display the LMMs on the logarithmically transformed RTs. However, results were similar both in the transformed and the untransformed analysis. The RTs were analyzed using LMMs with the factors *validity* (invalid = -1 vs. valid = 1), the *"NVC"* response (No NVC = -1, NVC = 1), and the corresponding interaction as fixed factors (1). We implemented the maximal random-effects structure justified by the design (as suggested by Barr, Levy, Scheepers, & Tily, 2013): Participants (including by-participant random slopes for Validity, NVC, and their interaction) and the different syllogism problems were treated as a random factors (2). The trial *"sequence"* (4-64) was added as covariate since it correlated with the NVC response capturing effects due to fatigue or learning (1). All continuous predictor variables were centered and scaled. The full model was specified as follows:

$$log(RT) = NVC * Validity + Sequence \quad (1)$$
$$+ (NVC * Validity | Participant) + (1 | Syllogism) \quad (2)$$

The results of the best-fit model can be taken from Table 2.

Table 2: Fixed-Effect Parameter Statistics for the full/ best-fit Reaction Time model.

Predictors	Estimates	*SE*	*t*	*p*
Intercept	9.43	0.05	176.76	$< .001$
NVC (yes = 1)	-0.02	0.02	-1.12	*.270*
Validity(valid = 1)	0.06	0.02	3.51	*.001*
Sequence	-0.13	0.01	-22.30	$< .001$
NVC:Validity	0.05	0.01	4.80	$< .001$

Hypothesis 1: Other than expected, there was no main effect of NVC on the RT as the RTs did not significantly increase in trials where a NVC response was given as compared to non-NVC trials. However, there was a significant interaction between NVC responses and the validity of the syllogism: the RTs were significantly associated with the occurrence of a NVC responses for valid syllogisms. In trials with NVC responses, the RT increased, but only for valid syllogisms. Any reduction of a parameter (e.g., of the interaction) resulted in a significantly worse model fit as compared to the full model reported. The interaction was also apparent in the mean RTs: For valid syllogisms, the RTs were higher for NVC responses (*M* = 20.17, *SD* = 17.37) as compared to other conclusions (*M* = 16.55, *SD* = 7.64). However, there was no difference for invalid syllogisms (NVC: *M* = 16.42, *SD* = 11.26, Other conclusions: *M* = 16.56, *SD* = 8.30).

The analysis of the likelihood to give a NVC responses. The occurrence of NVC-responses as a bivariate dependent variable was analyzed using GLMMs (NVC response = 1, no NVC response = 0). GLMM estimates were computed with a logit link, binomially distributed residuals using the bobyqa optimizer with 200 000 iterations. Odds ratios (ORs) of the fixed effects coefficients of the full model are reported as effect sizes.

The occurrence of a NVC response was analyzed with the factors Validity (invalid = -1 vs. valid = 1), the Entropy of each syllogism (using the entropy measures computed by Khemlani & Johnson-Laird, 2012), as well as the corresponding interaction and the trial sequence (4-64) as fixed factors (3). We again implemented the maximal random-effects structure justified by the design: Participants (including by-participant random slopes for the factors Validity, NVC, and their interaction) and the syllogism problems (random intercept) were treated as a random factors (4). The entropy variable was centered prior to analysis. The full model of was specified as follows:

$$NVC = Validity * Entropy + Sequence \quad (3)$$
$$+ (Validity * Entropy | Participant) + (1 | Syllogism) \quad (4)$$

Table 3: Fixed-Effect Parameter Statistics for the best-fit NVC model.

Predictors	Estimates	*SE*	*z*	*OR*	*p*
Intercept	1.77	0.19	-9.23	0.17	$<.001$
Validity$_{(valid = 1)}$	-0.77	0.15	-5.20	0.46	$<.001$
Entropy	0.41	0.35	1.15	1.5	*.249*
Sequence	0.18	0.03	5.62	1.19	$<.001$
Validity:Entropy	1.08	0.35	3.05	2.94	$=.002$

Note. OR indicates Odds Ratios.

The results of the best-fit model can be taken from Table 3.

Hypothesis 2. In 53% of the syllogisms a NVC response is the logically conclusion. As hypothesized, in the current experiment, participants gave 28.99% NVC responses (SD = 21.40%) on average which is significantly less than 58% ($V = 382$, $p < .001$; a paired Wilcoxon signed tank test was used due to a deviation from normality). Thus, we can confirm that the number of NVC responses was lower than the number of logically correct invalid syllogisms.

Hypothesis 3. As hypothesized, the occurrence of a NVC response was significantly associated with the validity of the syllogism. NVC responses were more likely to occur for invalid than for valid syllogisms. Excluding this factor from the full model resulted in a significant reduction of the overall fit ($\chi^2 = 189.56$, $p < .001$).

Hypothesis 4. We expected that the higher the entropy of a syllogism was the higher the likelihood of a NVC response would be. Other than hypothesized, there was no significant main effect for entropy on the likelihood to give a NVC response. However, there was a significant interaction between validity and entropy (see Figure 1 for an illustration): Entropy impacted the likelihood to respond NVC, but only for valid syllogisms. Excluding this interaction as well as the entropy factor from the full model resulted in a significant reduction of the overall fit ($\chi^2 = 57.07$, $p < .001$). A post-hoc analysis for the number of NVC responses and entropy also revealed a strong association between entropy and the relative frequency of NVC responses for each syllogisms for valid ($r_\rho = .69$, $p < .001$) but not for invalid syllogisms ($r_\rho = -.27$, $p = .112$).

Hypothesis 5. The effect of the trial sequence on the relative frequency of NVC responses separated for valid and invalid syllogisms is illustrated in Figure 2. The plot highlights that NVC responses do not stay constant over the time-course of the experiment. As expected, there was also a significant main effect of the sequence on the likelihood to give a NVC response in the mixed model (see 3). Excluding the sequence factor from the full model resulted in a significant reduction of the overall fit ($\chi^2 = 30.63$, $p < .001$). Since NVC is a logically sound conclusion only for invalid syllogisms, a simultaneous increase for invalid and decrease for valid syllogisms would indicate a trend towards a support for the theory that reasoners become more logical in the experiment. However, the increase in NVC response probability does not differentiate between valid and invalid syllogisms.

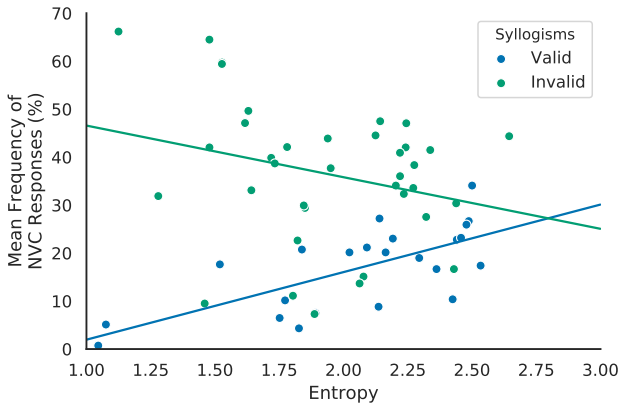


Figure 1: The relationship between the frequency of NVC responses and entropy. Linear regression lines are plotted separately for valid and invalid syllogisms.

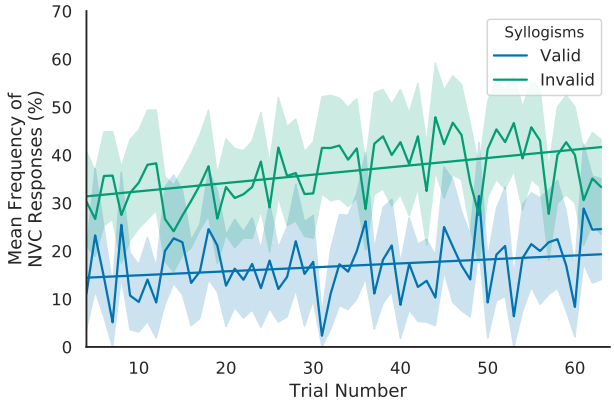


Figure 2: Mean number of NVC responses on valid and invalid syllogisms for the experimental sequence of 64 syllogisms.

An upwards trend can be observed for both. Furthermore, in follow-up analyses the inclusion of an interaction effect for the sequence and the validity of a syllogisms did not result in a fit improvement to the reported best-fit model $\chi^2 = 0.65$, $p = .420$.

Discussion

On which factors does the likelihood to respond NVC depend and how are NVC responses associated with differences in RTs? First, the RTs seem to increase in NVC trials as compared to trials where another conclusion was given - but only (and other than expected) for valid syllogisms. There are various explanations why RTs did not increase for NVC responses in invalid syllogisms. For instance, NVC responses are the "logical" correct response option for invalid syllogisms. Thus, on average, NVC responses for these problems could occur for both logical reasoning and as a consequence of other processes (e.g., guessing, giving-up, etc.). As a consequence, responding NVC may not only be a "last resort" after elaborate reasoning (thus, higher RTs), but also stem from logically correct reasoning. Also, providing any response

other than NVC for invalid syllogisms is logically incorrect and may therefore include deviating processes (possibly leading to prolonged RTs). Therefore, for invalid syllogisms, such effects may mask the effect of NVC responses. Second, while we found a significant main effect of validity on NVC, entropy was associated with NVC responses only for valid syllogisms. The reported interaction between validity and entropy on the frequency of NVC responses is however only logical. As theorized, the frequency of NVC responses seems to be higher for high entropy problems as compared to low entropy problems for valid syllogisms. The opposite relationship observed for invalid problems is logical as naturally for easy invalid syllogisms (reflected in a low entropy), participants should most frequently respond "NVC". The harder an invalid syllogism becomes (possibly reflected by a high entropy), the more the responses spread, and the less often a NVC response is given. Future analyses should investigate whether NVC responses are selected more frequently for high entropy problems in addition to the general benefit or drawback NVC responses receive by a higher variance in responses. Third, as hypothesized, the likelihood to respond NVC increases with the trial sequences during the time-course of the experiment. Surprisingly, this association seems to be apparent for both valid and invalid syllogisms. The effect of trial sequence on NVC responses can thus not be explained by participants becoming more logical. On the contrary, the results point towards other systemic processes taking place during the course of the experiment. Note, that this study used a selection task. It remains an open question how our results relate to tentative studies on generation tasks (for an overview of differences of response formats see Hardman & Payne, 1995). Moreover, variations of the classic syllogism task, such as the countermodel/ "Harry"-task (see Achourioti, Fugard, & Stenning, 2014), would certainly provide additional insights on the questions when participants conclude that "nothing follows" in other test situations.

General Discussion

When does a reasoner respond "nothing follows"? To answer that question we have investigated implications of the mental model (e.g., Johnson-Laird, 2006) and mental logic theory (Rips, 1994) as well as adapted alternative systemic hypothesizes such as the role of mental depletion. First, reasoners seem to take longer when responding NVC only for valid and not for invalid syllogisms. With regard to the proposed theories and systemic hypotheses of interest, this finding poses a challenging novel perspective on NVC responses as this distinction is not yet predicted by cognitive theories: giving a NVC response generally takes longer due to the requirement of more cognitive steps, e.g., by generating all inferences or searching for counterexamples (Khemlani & Johnson-Laird, 2012). So, the time needed to respond "nothing follows" is expected to be independent of the validity of a problem. Moreover, the Early Stopper hypothesis contradicts this empirical finding: An Early Stopper would not need more time for responding NVC. In sum, our assumptions holds true only for valid syllogisms. This raises the question whether invalid and valid syllogisms are processed differently and influenced by other processes such as mental depletion or a NVC aversion. Second, the likelihood to respond NVC increases for both valid and invalid syllogisms over time indicating that these differences cannot be explained by participants becoming more logical within the same experiment. While the Early Stopper hypothesis cannot account for this finding, the results can be well explained by the NVC aversion hypothesis. Participants may have an early aversion to respond NVC. If the NVC response is assigned a meaning of "I give up", participants might need to encounter some of the invalid syllogisms to gain confidence in stating that no conclusion may follow from the premises. It is possible that a reasoner may for instance learn across solving syllogistic problems that for some types of problems a valid conclusion cannot be found. Hence, the reasoner can start to assume that the probability of NVC problems is high (with each such observation). The aversion may however also diminish over time due to depletion or fatigue effects.

What can we conclude regarding our proposed theories and systemic hypotheses based on these findings? We see that cognitive theories seems to be able to provide correct predictions in terms of RTs for NVC responses for valid but does not for invalid syllogisms. The systemic hypotheses proposing an early NVC aversion and a later mental depletion seem to be able to explain why cognitive theories sometimes fail to predict NVC responses correctly: Yet, cognitive theories do not yet take such processes into account. It is noteworthy however, that the systemic hypotheses are unable to explain some of the results found in the present study. Whereas the cognitive theories do at least predict an effect of NVC-responses on the RTs, two of the systemic hypotheses do not necessarily propose higher RTs for such trials. Additionally, one of the systemic hypotheses predicted the main effect of validity on NVC responses.

In summary, with regard to the proposed theories, we see that the cognitive theories seem to be able to provide correct predictions of NVC responses for valid but sometimes not for invalid syllogisms. The strong dependencies on the validity of a syllogism as well as differences over the time-course of an experiment suggest that there are also some other cognitive processes taking place within the individual. The systemic hypotheses can account for some of these effects complementing the cognitive theories. We can conclude that there may indeed be an initial bias against an NVC response, which highly differs between individuals. Hence, more analysis are necessary to analyze the interplay between existing cognitive reasoning theories and possible systemic hypotheses to increase the correct prediction rate of when people answer NVC. Indeed, in parallel to this work, we were able to demonstrate that by attaching heuristic rules for predicting NVC to cognitive models of syllogistic reasoning, their performance can increase up to 20 % on average (Riesterer, Brandt, Dames, & Ragni, in press). Last, the results also highlight that logical correctness need to be used with caution when analyzing syllogistic reasoning data due to the unproportional weight of NVC responses: Such analyses should always consider the validity of the problems.

Acknowledgements

This paper was supported by DFG grants RA 1934/3-1, RA 1934/2-1 and RA 1934/4-1 to MR.

References

Achourioti, T., Fugard, A. J., & Stenning, K. (2014). The empirical study of norms is just what we are missing. *Frontiers in Psychology*, *5*, 1159.

Baayen, R., Davidson, D., & Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390 - 412. (Special Issue: Emerging Data Analysis)

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255 - 278.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Bucciarelli, M., & Johnson-Laird, P. N. (1999). Strategies in syllogistic reasoning. *Cognitive Science: A Multidisciplinary Journal*, *23*(3), 247-303.

Copeland, D. E. (2006). Theories of categorical reasoning and extended syllogisms. *Thinking & Reasoning*, *12*(4), 379–412.

Costa, A., Saldanha, E.-A. D., Hölldobler, S., & Ragni, M. (2017). A computational logic approach to human syllogistic reasoning. In *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (p. 883-888).

Hardman, D. K., & Payne, S. J. (1995). Problem difficulty and response format in syllogistic reasoning. *The Quarterly Journal of Experimental Psychology Section A*, *48*(4), 945–975.

Johnson-Laird, P. N. (2006). *How we reason*. Oxford: University Press.

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, *103*(1), 54-69.

Khemlani, S., & Johnson-Laird, P. N. (2012). Theories of the syllogism: A meta-analysis. *Psychological Bulletin*, *138*(3), 427–57.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26.

Lo, S., & Andrews, S. (2015). To transform or not to transform: using generalized linear mixed models to analyse reaction time data. *Frontiers in Psychology*, *6*, 1171.

Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford: Oxford University Press.

Ragni, M., Riesterer, N., Khemlani, S., & Johnson-Laird, P. (2018). Individuals become more logical without feedback. In T. Rogers, M. Rau, J. Zhu, & C. Kalish (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 1584–1589). Austin, TX: Cognitive Science Society.

Riesterer, N., Brandt, D., Dames, H., & Ragni, M. (in press). Modeling human syllogistic reasoning: The role of no valid conclusion. In A. Goel, C. Seifert, & A. Arbor (Eds.), *Proceedings of the 41th Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Rips, L. J. (1994). *The psychology of proof: Deductive reasoning in human thinking*. Cambridge, MA: The MIT Press.

Schmeichel, B. J., & Vohs, K. D. (2018). Intellectual performance and ego depletion: Role of the self in logical reasoning and other information processing. In *Self-Regulation and Self-Control* (pp. 318–347). Routledge.

Shannon, C. E., & Weaver, W. (1963). The mathematical theory of communication. 1949. *Urbana, IL: University of Illinois Press*.

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction*, *4*(4), 295–312.

Wetherick, N. E., & Gilhooly, K. J. (1995). Atmosphere, matching, and logic in syllogistic reasoning. *Current Psychology*, *14*(3), 169-178.

Woodworth, R. S., & Sells, S. B. (1935). An atmosphere effect in formal syllogistic reasoning. *Journal of Experimental Psychology*, *18*(4), 451–460.

The Design of the Learning Environment Shapes Preschoolers' Causal Inference

Alexandra Rett (arett@ucsd.edu)[1], Elizabeth Bonawitz (elizabeth.bonawitz@rutgers.edu)[2], & Caren Walker (carenwalker@ucsd.edu)[1]

[1] Department of Psychology, UC San Diego, La Jolla, CA 92093

[2] Department of Psychology, Rutgers University, Newark, NJ 07102

Abstract

In the present study, we examine whether the design of the learning environment can impact causal inference in very young children. Specifically, we assess whether the physical features of a novel toy can facilitate children's recognition of an abstract, relational hypothesis (*same-different*) that they typically fail to discover. Three-year-olds were presented with an identical pattern of evidence that was consistent with a relational hypothesis (i.e., pairs of same or different blocks cause a toy to activate) using one of two causal toys. In the *standard* condition, blocks were placed in pairs on top of the toy, while in the *relational* condition, each block was placed inside one of two transparent openings on either side of the toy. The physical design of the latter toy was intended to highlight the relationship between pairs of blocks. Results suggest that even 3-year-olds' causal inferences are sensitive to design, with children in the *relational* condition more likely to infer the abstract relation than those in the *standard* case. These results provide strong evidence that design serves as a constraint on causal inference in early childhood. Findings are discussed in terms of their implications for creating intuitive learning environments for young children.

Keywords: cognitive development; causal inference; relational reasoning; learning environments; design

Introduction

When reasoning about novel causal relationships, learners must select the most likely hypothesis from a range of underdetermined possibilities. For example, to activate a novel appliance, you might consider several possible interventions: the 'on/off' switch might have to be flipped, the reset button on the circuit interrupter might have to be depressed, or perhaps both the switch and the button together activate the device. Depending on your prior belief in the likelihood of each candidate cause and your subsequent observations, you then select the most likely action. If, for example, the switch is in the 'on' position and the appliance does not activate, it provides evidence that it must be activated in conjunction with the interrupter reset. However, one could also imagine a seemingly infinite number of alternative ways the causal system may work. Perhaps the appliance is voice activated, or the buttons need to be pushed in a particular repeating order, or there is an additional hidden switch somewhere else on the device.

To solve the infinite hypothesis search problem, recent work emphasizing the psychological processes underlying inductive inference has proposed that learners likely "sample" from this vast space of hypotheses, based on prior knowledge (e.g., Bonawitz, Denison, Griffiths, & Gopnik, 2014; Ullman, Goodman, & Tenenbaum, 2012; Tenenbaum, Griffiths & Kemp, 2006). Thus, instead of considering all possible hypotheses and weighing each against the observed evidence, learners may only generate a subset of the most likely candidates to evaluate (Bonawitz & Griffiths, 2010). Critically, the specific subset of hypotheses that is generated for a particular learning problem may depend on a variety of factors, including their prior probability, their relevance to the current problem, priming, and so forth (e.g., Dougherty & Hunter 2003; Flin, Slaven & Stewart, 1996; Klein, 1993; Weber et al., 1993; Schunn & Klahr, 1993; Koehler, 1994). In fact, even young children are sensitive to input that constrains the hypotheses they consider, including information about the problem they are trying to solve, how the data were sampled, who generated the evidence, and why (e.g., Buchsbaum, Gopnik, Griffiths, & Shafto, 2011; Butler & Markman, 2012; Gergely, Bekkering & Kiraly, 2002; Walker, Lombrozo, Legare, & Gopnik, 2014).

Accordingly, any input that changes a learner's prior expectations about the most likely causal structure can influence the hypotheses they privilege, and ultimately apply. Here, we consider a specific environmental cue that, to our knowledge, has not yet been examined: the visible *design* of the object itself. If children use information about an object's design to constrain the hypotheses they generate, changes in the *physical features* of the learning context might influence causal learning and discovery. That is, the design of a causal system may serve to increase or decrease the salience of some hypotheses over others.

Effects of Design on Behavior

Although object design has not been specifically examined in the context of causal learning, there are several reasons to expect that the physical features of the learning context may influence children's causal inference. Indeed, nearly all of the objects we interact with include some element of design, and we often use these cues to infer information about an object's function. For example, if a door has no handle, the only way to enter is to push. While this action seems intuitive, the design is intentional. The creator constructed the door so that its physical features would constrain the permissible actions. Norman (1988) includes such constraints as one of several principles of good design, recognizing that design impacts reasoning about object function. A large body of literature has also explored the ways in which subtle environmental influences, or "nudges," have disproportional effects on human choice (Thaler & Sunstein, 2008), impacting hygiene (Holland, Hendriks, & Aarts, 2005), energy use, (Allcott & Mullainathan, 2010), and health (Thorndike, Sonnenberg,

Riis, Barraclough, & Levy, 2012; van Nieuw-Amerongen, Kremers, De Vries, & Kok, 2011), among others.

Other applied research has also begun to examine whether environmental design can change the way we *learn* in select educational contexts. For example, museum designers have used exhibit access, visibility, and object affordances to encourage visitor exploration, engagement, and understanding (e.g., adding a knob to a display suggests that an object can be moved, adding a glass window on the side of a machine encourages visitors to view the internal mechanism; see Allen, 2004; Wineman & Peponis, 2010; Shin, Park & Kim, 2014). Here, we go beyond this past applied work to consider whether similar cues can influence the salience of certain concepts or reasoning strategies in the context of causal learning. That is, we test whether elements of design influence a learner's prior beliefs about the likelihood of a particular causal hypothesis, given some pattern of evidence.

To illustrate how the design of an object might impact a learner's beliefs about its function, we return to our novel appliance. If you are familiar with electronic machines, you might believe that before you can turn something on, you must connect its cord to a power source. Once learned, this general principle can be widely applied to novel cases, even before observing any evidence about how a particular appliance functions. However, now consider a situation in which you are confronted with an appliance that has *two* cords. In this case, your prior belief that a single power cord must be plugged in to turn on the machine seems less probable. You might instead form a hypothesis that the two cords must *both* be connected before the machine will turn on. This demonstrates an even more general assumption that the features of an object are relevant to its function. This sort of abstract causal principle, or "overhypothesis," is a belief about the *kinds* of hypotheses that are most likely to be true (Goodman, 1955; Kemp, Perfors, Tenenbaum, 2007). Based on the learner's prior experience, it might seem unlikely to observe a second power cord that is unnecessary for the machine's operation (without an alternative explanation for the presence of the second cord). In this way, the visible features of an object serve as critical design cues that constrain the hypotheses that are generated about its causal structure (Norman, 1988).

Some existing support for the proposal that an object's design serves to constrain inferences about causal structure can be found in the literature examining human reasoning about artifacts (i.e., human-made objects). That is, both children and adults view features of artifacts as reflective of that object's function and intended use (Keil, 1992; Keleman, 1999; Keleman, Seston, & Saint Georges, 2012). For example, Kelemen and colleagues (2012) showed preschool-aged children two objects that were equally optimal for performing a particular function (i.e., both objects featured a flat surface that could be used to crush popcorn), but one of them had additional salient features that suggested it could *also* be used for an additional purpose (i.e., spikes along the object's handle). When asked which object was designed for the target purpose (crushing popcorn), 3- and 4-year-old children privileged the object with a more efficient design.

Magid and colleagues (2015) also provide evidence that children relate an object's design to its function. The authors argue that young learners represent the abstract criteria for solving a problem, before arriving at a precise solution. These criteria are based on how well a particular hypothesis matches the abstract "form" of the problem to be solved. Specifically, 4 and 5-year-olds mapped the *type* of effect produced (a discrete vs. continuous visual effect) to the *type* of mechanism that produced it (a binary "on/off" switch vs. a dial), providing evidence that children relate the physical structure of an object's causal mechanism to its effect. Additionally, 4- and 5-year-olds have also been shown to map the quantity and diversity of object functions (e.g., making cupcakes vs. making cupcakes *and* wrapping presents) to make inferences about the complexity of the design of its internal mechanism (Ahl & Keil, 2016).

The Current Approach

In the prior work reviewed above, learners made inferences about the design of objects, given information about possible functions. Here, we ask whether children can perform a more challenging task -- whether they will be more likely to generate a particular causal hypothesis, given the object's design. In particular, we present a conceptual case in which 3-year-olds typically fail to discover a relational hypothesis. We then assess whether the object's design influences learning by observing whether subtle changes to the physical structure of the causal system leads to the successful identification of the abstract relational cause.

Specifically, we present 3-year-olds with a relational reasoning problem that they systematically fail at this age (Walker, Bridgers & Gopnik, 2016). In this task, children are introduced to a novel toy that plays music for some objects and not for others (i.e., a "blicket detector," Gopnik & Sobel, 2000). They then observe pairs of blocks being placed on top of the toy. When 3-year-olds are provided with evidence that the toy's activation is caused by the relation between the two blocks in each pair (i.e., whether the blocks are the same or different), rather than by individual object kinds (i.e., blocks of a particular shape and color), they failed to make the correct causal inference at test (see Figure 1).

Notably, younger children (18 to 30-month-olds) successfully infer *same-different* relations in this task, suggesting that later failures are due to a difference in tendency, not a lack of relational competence (Walker et al, 2016; Walker & Gopnik, 2017; Walker, Walker, & Gopnik, under review). In other words, these developmental data provide evidence that older children are capable of inferring such relations, even if they do not spontaneously generate them in most learning scenarios. Critically, this proposal contrasts with decades of research suggesting that preschoolers were simply unable to reason on the basis of

these abstract relations (e.g., Christie & Gentner, 2010; 2014).

Based on these findings, it has been proposed that 3-year-olds' long-documented failure to infer *same-different* relations results from a learned bias in the form of an overhypothesis that privileges the role of individual objects over the relations between them (Walker et al., 2016; Carstensen & Walker, 2017). Walker and colleagues (2016, Exp 3) provide additional support for this idea, demonstrating that prompting children to explain during training trials significantly increases their tendency to endorse the relational hypothesis at test. The authors propose that explanation likely serves as an *internal* constraint on hypothesis search, leading learners to privilege more abstract solutions. This domain therefore provides a promising case study to explore the proposal that an *external* constraint, namely the design of an object, can also influence hypothesis generation in causal learning.

In order to assess whether the tendency to discover the relational hypothesis may be sensitive to constraints imposed by physical design, we made one small modification to the standard causal relational task: Rather than placing pairs of blocks on top of the toy on a single, large platform, the blocks were inserted on either side of the toy, into two transparent openings (see Figure 2). By adding these two intentionally designed openings, a learner who treats object design as relevant to their causal inferences might consider *why* the causal system included these features. These two openings therefore not only draw attention to the presence of two objects, but also suggest a particular affordance: that the machine activates by combining the two. As a result, this may raise the possibility that the *relation between* the blocks—rather than the identity of the blocks themselves—is relevant to the causal structure, leading to the discovery of the relational hypothesis. We return to consider the implications of this particular design choice in the discussion.

Alternatively, it may be the case that the design of the causal system has no effect on 3-year-olds' endorsement of the relational hypothesis. As noted, children at this age repeatedly fail to spontaneously privilege relational information (Christie & Gentner, 2010; 2014; Walker et al., 2016), suggesting a strong prior for hypotheses based on individual object kinds. To correctly infer the relational hypothesis in this case, children must integrate information about the object's design with their prior beliefs about likely causes, taking into account why object design is relevant, *and* weighing this information more heavily than their prior commitment to the object-based hypothesis. That said, if children's failure to infer abstract relations indeed results from a difference in *tendency,* rather than a lack of *competence* (as has been suggested)*, and* they are sensitive to the design of the learning context, then we might reasonably expect them to successfully infer the abstract relational hypothesis, even following such a minor modification to the standard task.

Methods

Participants

A total of 152 3-year-olds participated in the study, with 76 children randomly assigned to either the *standard toy* (M = 41.9 months; 36 female) or *relational toy* (M = 41.6 months; 37 female) conditions. Within each condition, half of the children observed evidence consistent with the *same* relation and half observed evidence consistent with the *different* relation. Sample size satisfies a power analysis with power > .8, given an alpha of .05 and an effect size of .3 (medium). An additional 9 participants were excluded due to experimenter error (3), failure to complete the study (4), parent interference (1), or interference by another child (1). Children were recruited and tested in the lab, at preschools, and at museums. All participants were tested in a quiet, private room with the experimenter.

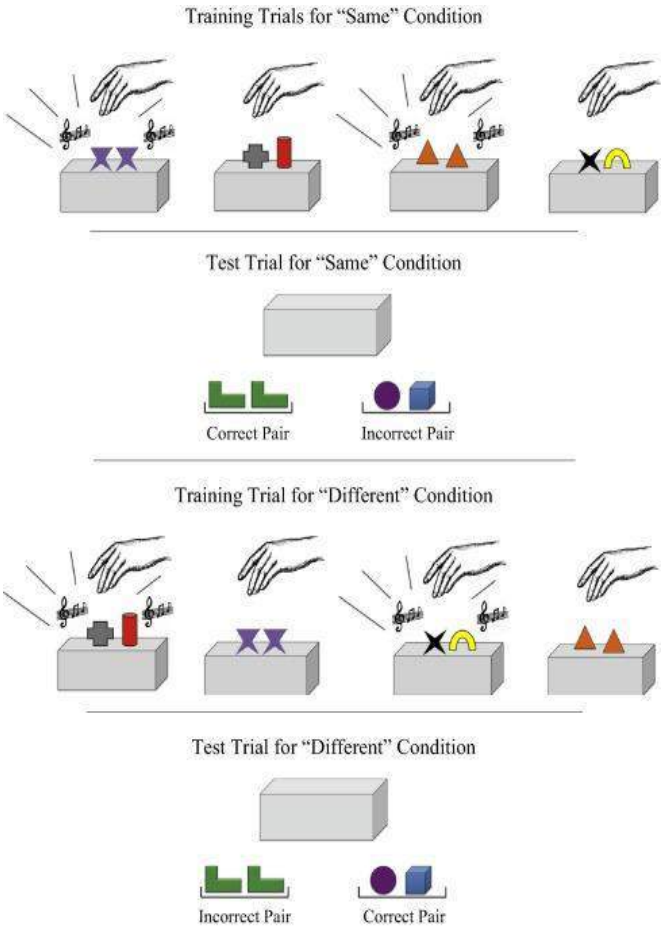


Figure 1. Schematic illustration of evidence presented during training and test trials in the *standard* condition (reprinted from Walker et al., 2016, Exp. 1). Identical pairs and outcomes were presented in the *relational* condition, using the relational toy (see Fig. 2).

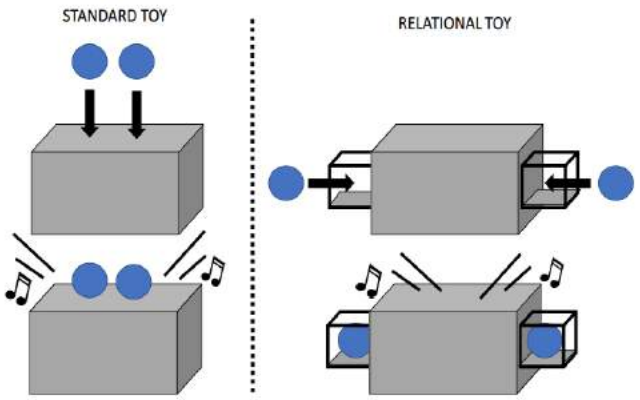


Figure 2. Standard and relational toys

Materials and Procedure

The materials and procedure for the *standard toy* condition replicate those used by Walker et al. (2016, Exp. 1; see Fig. 1). Children were seated at a table across from the experimenter. The experimenter began by placing an opaque cardboard box on the table, saying "This is my toy! Sometimes when I put things on top, the toy will play music, and other times it does not. Should we try some and see how it works?" As in previous research, the toy appeared to activate and play a novel melody in response to certain combinations of blocks. In fact, the experimenter activated a wireless doorbell inside the box by surreptitiously pressing a button.

A total of 4 pairs of *same* and *different* painted wooden blocks (2 pairs of *same* and 2 pairs of *different*) were used during the training trials. After introducing the toy, the experimenter produced two blocks in either the *same* or *different* relation (depending upon the condition), and said, "Let's try!," and put both blocks on top of the toy, simultaneously. The toy played music and the experimenter said, "Music! My toy played music!" The experimenter then picked up the blocks and set them back on the toy, which again played music, saying "Music! These ones made my toy play music!" She then repeated this procedure with a new pair of blocks in the opposite relation. The new pair did not make the toy play music, and the experimenter responded to the first try with, "No music! Do you hear anything? I don't hear anything," and after the second try, said "No music. These ones did not make my toy play music." This pattern was repeated with two additional pairs of blocks, one in each relation. The experimenter always began with a causal pair (identical blocks in the *same* condition and blocks of unique colors and shapes in the *different* condition), and then alternated inert, causal, inert, using novel blocks in each new pair, and randomizing the specific blocks between participants.

After the four training trials, the experimenter said "Now that you've seen how my toy works, I need your help finding the things that will make it play music. I have two choices for you." The experimenter presented the child with two new pairs composed of novel blocks, one "same" pair and one "different" pair. Each pair was presented on a plastic tray, which the experimenter held up, saying, "I have these, and I have these (directing the child's attention to each pair). Only one of these trays has the things that will make my toy play music. Can you point to the tray that has the things that will make it play?" The trays were then placed out of the child's reach, on either side of the toy, with each pair set an equal distance from the child. The order and side of presentation of the correct pair counterbalanced between participants. The experimenter recorded the child's first point or reach, scoring the response as correct (1) if the child chose the test pair (same or different) that corresponded to her training, and incorrect (0) for the opposite pair.

The materials and procedures for the *relational toy* condition were identical to those in the *standard toy* condition with one critical difference: The design of the toy was modified to include two transparent openings located on either side (see Fig. 2). The openings were constructed using clear, 2" x 2" hard plastic boxes. When children observed each of the training trials described above, pairs of blocks were inserted into the two openings (one block on either side), rather than placed on top of the toy. This was the only difference between the two conditions.

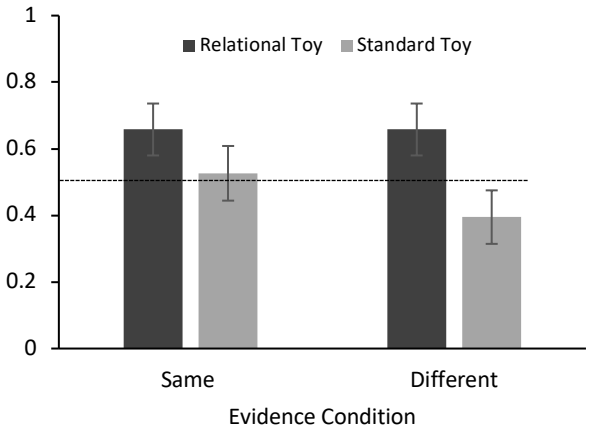


Figure 3. Mean proportion of correct relations by condition. Error bars indicate +/- 1 *SEM*. Chance performance is indicated by the dotted line.

Results

Replicating previous work (Walker et al., 2016), 3-year-old children in the *standard toy* condition responded at chance (46%), p = .57 (two-tailed, exact binomial), with no difference in performance between *same* (53%) and *different* (40%) training trials, p = .35 (two-tailed, Fisher's exact). In contrast, 3-year-olds in the *relational toy* condition succeeded in selecting the test pair that was consistent with their training (66%), p = .008 (two-tailed exact binomial), performing identically in *same* and *different* training trials (p = 1, two-tailed, Fisher's exact). Comparing performance across conditions, children in the

relational toy condition significantly outperformed those in the *standard* condition (p = .022, two-tailed, Fisher's exact) in inferring *same-different* relations.

Discussion

In the current study, we present findings demonstrating that children are indeed sensitive to the physical design of the learning context when reasoning about causal relationships. Although 3-year-olds in the *standard toy* condition failed to recognize the relational hypothesis (replicating prior work), increasing the salience of this hypothesis through the application of a relatively subtle design cue significantly increased their tendency to engage in relational reasoning in this task. In addition to providing evidence for the role of design in constraining causal inference, these data provide additional support for the proposal that children's failure on relational reasoning tasks results from a difference in *tendency*, not a lack of competence (e.g., Walker & Gopnik, 2014; Walker et al., 2016; Carstensen & Walker, 2017; Walker & Gopnik, 2017).

These results are particularly striking given that 3-year-olds have repeatedly failed to spontaneously privilege relational information (Christie & Gentner, 2010; 2014), suggesting a very strong prior to prefer individual object kinds. In order to use the design of the learning context to override this tendency and privilege the relational hypothesis, these very young children had to make a particularly sophisticated inference: They must have noticed this subtle design cue, inferred its relevance to the system's causal structure (i.e., that an object's design is relevant for its function), and weighed this information more heavily than their (strong) prior commitment to the object-based hypothesis.

These surprising findings therefore suggest that relatively minor elements of design can radically change the distribution of a learner's prior expectations, constrain the type of hypotheses that are generated, influence learning outcomes, and even facilitate the early discovery of new causal beliefs. Our results join prior research suggesting that hypothesis generation can be influenced by a variety of cognitive factors (e.g., Dougherty & Hunter 2003; Flin, Slaven & Stewart, 1996; Klein, 1993; Weber et al., 1993; Schunn & Klahr, 1993; Koehler, 1994), prompts (Walker et al., 2014, 2016; Williams & Lombrozo, 2010) and social inferences (Butler & Markman, 2012; Buchsbaun et al., 2011; Gergely, Bekkering, & Kiraly, 2002), and extends this work to include the structure of the learning environment itself. Ongoing work examines whether and how design influences even more entrenched causal beliefs and biases (e.g., in adults; Walker, Rett, & Bonawitz, in prep), and considers how design may interact with other constraints, such as pedagogical cues or prompts to explain. For instance, in some contexts, children privilege an object's visible affordances over an actor's intentional behavior when reasoning about how an artifact is intended to be used (e.g., DiYanni & Keleman, 2008). Future work will explore to what extent learners may be reasoning about the intentions of the designer (as a social agent) when making inferences based on these environmental cues.

There are also open questions surrounding how the particular design modifications used in this experiment influence children's reasoning. One possibility is that the addition of exactly two transparent openings on either side of the toy directly primed the relational hypothesis. Another possibility is that this design cue simply served to disrupt children's initial intuitions about the likely causal mechanism, leading them to consider alternatives more broadly. If so, this may have made it more likely for children to discover the relational hypothesis, albeit indirectly. Future work is needed to address these important questions.

Finally, these results have clear practical implications for early science education, and in particular, the design of formal and informal learning environments intended for children. Our findings dovetail with literature in education pointing to the importance of "mise en place" or setting the stage for learning (Weisberg, Hirsh-Pasek, Golinkoff, & McCandliss, 2014). As demonstrated here, children are sensitive to relatively subtle physical cues in the learning environment when they are engaged in causal reasoning. This simple manipulation led children to consider a relational hypothesis that they typically fail to spontaneously produce. Our findings therefore highlight the importance of careful design when aiming to teach children specific concepts, given that the visible features of objects may increase or decrease the salience of the available evidence, and change the learner's interpretation of their observations. It is impossible to create artifacts without also making specific design choices, so being aware of how these features might be used to facilitate reasoning can have major consequences for learning and instruction. These findings therefore open up new avenues for future work examining how the design of learning environments can be used to support belief revision and guide early learning and discovery.

Acknowledgements

We are grateful to Mike Frank for his early contributions in conceptualizing the methods for this study. We thank Alicia Lunardhi and Emily To for their efforts towards data collection and Nicky Sullivan for facilitating recruitment. We also thank the Fleet Science Center, the Birch Aquarium at Scripps Institute of Oceanography, and the New Children's Museum, as well as all of the participating preschools and families who made this research possible. This research was funded by a Hellman's Fellowship, awarded to C. Walker.

References

Ahl, R. E., & Keil, F. C. (2017). Diverse effects, complex causes: children use information about Machines' functional diversity to infer internal complexity. *Child development, 88*(3), 828-845.

Allcott, H., & Mullainathan, S. (2010). Behavior and energy policy. *Science, 327*(5970), 1204-1205.

Allen, S. (2004). Designs for learning: Studying science museum exhibits that do more than entertain. *Science Education, 88*(S1), S17-S33.

Bonawitz, E., Denison, S., Griffiths, T. L., & Gopnik, A. (2014). Probabilistic models, learning algorithms, and response variability: Sampling in cognitive development. *Trends in cognitive sciences*, *18*(10), 497-500.

Bonawitz, E., & Griffiths, T. (2010). Deconfounding hypothesis generation and evaluation in Bayesian models. In S. Ohlsson & R. Catrambone (Eds.), Proceedings of the 32nd Annual Conference of the Cognitive Science Society (pp. 2260 –2265). Austin, TX: Cognitive Science Society.

Buchsbaum, D., Gopnik, A., Griffiths, T. L., & Shafto, P. (2011). Children's imitation of causal action sequences is influenced by statistical and pedagogical evidence. *Cognition, 120*(3), 331-340.

Butler, L. P., & Markman, E. M. (2012). Preschoolers use intentional and pedagogical cues to guide inductive inferences and exploration. *Child development*, 83, 1416-1428.

Carstensen, A. & Walker, C.M. (2017). The paradox of relational development is not universal: Abstract reasoning develops differently across cultures. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.), *Proceedings of the 39th Annual Conference Cognitive Science Society,* pp. 1721-1726. London, UK: Cognitive Science Society.

Christie, S., & Gentner, D. (2010). Where hypotheses come from: Learning new relations by structural alignment. *Journal of Cognition and Development, 11*(3), 356-373.

Christie, S. & Gentner, D. (2014). Language helps children succeed on a classic analogy task. *Cognitive Science*, 38(2), 383-397.

DiYanni, C., & Kelemen, D. (2008). Using a bad tool with good intention: Young children's imitation of adults' questionable choices. *Journal of experimental child psychology*, *101*(4), 241-261.

Dougherty, M. R., & Hunter, J. E. (2003). Hypothesis generation, probability judgment, and individual differences in working memory capacity. *Acta psychologica*, *113*(3), 263-282

Flin, R., Slaven, G., & Stewart, K. (1996). Emergency decision making in the offshore oil and gas industry. *Human Factors*, *38*(2), 262-277.

Gergely, G., Bekkering, H., & Kiraly, I. (2002). Rational imitation in preverbal infants. *Nature, 415,* 755.

Goodman, N. (1955). Fact, fiction, and forecast. Cambridge, MA: Harvard University Press.

Holland, R. W., Hendriks, M., & Aarts, H. (2005). Smells like clean spirit: Nonconscious effects of scent on cognition and behavior. *Psychological Science, 16*(9), 689-693.

Keil, F. C. (1992). Concepts, kinds, and cognitive development. MIT Press.

Kelemen, D. (1999). Function, goals and intention: Children's teleological reasoning about objects. *Trends in Cognitive Sciences, 3*(12), 461-468.

Kelemen, D., Seston, R., & Saint Georges, L. (2012). The designing mind: Children's reasoning about intended function and artifact structure. *Journal of Cognition and Development, 13*(4), 439-453.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental science, 10*(3), 307-321.

Klein, G.A. (1993). A Recognition-Primed Decision (RPD) Model of Rapid Decision Making. In G. A. Klein, J. Orasanu, R. Calderwood, and C. Zsambok (Eds.), *Decision Making in Action: Models and Methods.* Norwood, NJ: Ablex Publishing Corp., 138–147.

Koehler, D. J. (1994). Hypothesis generation and confidence in judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(2), 461.

Magid, R. W., Sheskin, M., & Schulz, L. E. (2015). Imagination and the generation of new ideas. *Cognitive Development, 34*, 99-110.

Norman, D. (1988). *The design of everyday things*. New York: Basic Books.

Schunn, C.D., & Klahr, D. (1993) Self vs. Other-Generated Hypotheses in Scientific Discovery. In *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society.*

Shin, H., Park, E. J., & Kim, C. J. (2014). Learning affordances: Understanding visitors' learning in science museum environment. In *Topics and Trends in Current Science Education* (pp. 307-320). Springer, Dordrecht.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in cognitive sciences*, *10*(7), 309-318.

Thaler, Richard H., Sunstein, Cass R. (2008) *Nudge: improving decisions about health, wealth, and happiness* New Haven: Yale University Press.

Thorndike, A. N., Sonnenberg, L., Riis, J., Barraclough, S., & Levy, D. E. (2012). A 2-phase labeling and choice architecture intervention to improve healthy food and beverage choices. *American Journal of Public Health, 102*(3), 527-533.

Ullman, T. D., Goodman, N. D., & Tenenbaum, J. B. (2012). Theory learning as stochastic search in the language of thought. *Cognitive Development*, *27*(4), 455-480.

van Nieuw-Amerongen, M. E., Kremers, S. P. J., De Vries, N. K., & Kok, G. (2011). The use of prompts, increased accessibility, visibility, and aesthetics of the stairwell to promote stair use in a university building. *Environment and Behavior, 43*(1), 131-139.

Walker, C. M., Bridgers, S., & Gopnik, A. (2016). The early emergence and puzzling decline of relational reasoning: Effects of knowledge and search on inferring abstract concepts. *Cognition, 156,* 30-40.

Walker, C.M. & Gopnik, A. (2014). Toddlers infer higher-order relational principles in causal learning. *Psychological Science*, 25(1): 161-169.

Walker, C.M., & Gopnik, A. (2017). Discriminating relational and perceptual judgments: Evidence from human toddlers. *Cognition, 166,* 23-37.

Walker, C. M., Lombrozo, T., Legare, C. H., & Gopnik, A. (2014). Explaining prompts children to privilege inductively rich properties. *Cognition*, *133*(2), 343-357.

Walker, C.M., Rett, A., & Bonawitz, E. (in prep). Design drives discovery in causal learning. *Manuscript in preparation.*

Walker, C.M., Walker, J.C., & Gopnik, A. (under review). Toddlers generalize abstract representations of *same* and *different.*

Weber, E.U., Böckenholt, U., Hilton, D.J., Wallace, B. (1993). Determinants of diagnostic hypothesis generation: effects of information, base rates, and experience. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 19, 1151– 1164

Weisberg, D. S., Hirsh-Pasek, K., Golinkoff, R. M., & McCandliss, B. D. (2014). Mise en place: Setting the stage for thought and action. *Trends in Cognitive Sciences, 18*(6), 276-278.

Williams, J. J., & Lombrozo, T. (2010). The role of explanation in discovery and generalization: Evidence from category learning. *Cognitive Science,* 34, 776–806. doi:10.1111/j.1551-6709.2010.01113.x

Wineman, J. D., & Peponis, J. (2010). Constructing spatial meaning: Spatial affordances in museum design. *Environment and Behavior, 42*(1), 86–109.

Distributional semantic representations predict high-level human judgment in seven diverse behavioral domains

Russell Richie, Wanling Zou
Department of Psychology
Sudeep Bhatia
Department of Psychology, Wharton Marketing
University of Pennsylvania
{drrichie,wanlingz,bhatiasu}@sas.upenn.edu

Abstract

The complex judgments we make about the innumerable objects in the world are made on the basis of our representation of those objects. Thus a model of judgment should specify (a) our representation of the many objects in the world, and (b) how we use this knowledge for making judgments. Here we show that word embeddings, vector representations for words derived from statistics of word use in corpora, proxy this knowledge, and that accurate models of judgment can be trained by regressing human judgment ratings (e.g., femininity of traits) directly on word embeddings. This method achieves higher out-of-sample accuracy than a vector similarity-based baseline and compares favorably to human inter-rater reliability. Word embeddings can also identify the concepts most associated with observed judgments, and can thus shed light on the psychological substrates of judgment. Overall, we provide new methods and insights for predicting and understanding high-level human judgment.

Keywords: judgment; semantic memory; machine learning; word embeddings

Introduction

People are constantly perceiving, judging and evaluating entities in the world, on the qualities that these entities possess. They may consider, for example, whether a food item is nutritious, whether a political candidate is competent, whether a consumer brand is exciting, or whether the work of an occupation is significant. Such judgments influence every sphere of life, determining the social, professional, consumer, and health outcomes of individuals, as well as the political and economic makeup of our societies. It is thus of critical importance to cognitive and behavioral scientists to develop predictive and explanatory models of human judgment. To have good empirical coverage and practical utility, such models must apply to naturalistic objects and concepts, i.e., the vast range of entities people encounter every day and have rich knowledge about. They should be able to quantify what people know about these entities, and specify how people map this knowledge onto the diverse array of complex judgments they make on a day-to-day basis.

To date, building such models has been elusive, as it has been difficult to represent the detailed knowledge people have about the millions of entities in the world that they judge. Traditional psychometric methods of formally specifying object knowledge – multidimensional scaling or simply asking people to rate objects on dimensions theorized to be core to a domain – are costly and typically yield sparse representations. Thus, a technique is needed which cheaply delivers rich, high-dimensional knowledge representations for a large number of objects and concepts, which can then be used to model judgments. Fortunately, such a technique can be found in word embeddings, real-valued vector representations of word meaning derived from the statistics of word use in language corpora, such that words that occur in similar linguistic contexts yield similar vectors (see Lenci (2018) for a review). Word embeddings are a useful tool for many practical natural language processing and artificial intelligence applications. However, they also mimic aspects of human semantic cognition: they can be used to predict judgments of word similarity and relatedness, patterns of free word association, strength of semantic priming, and semantic search (Hill, Reichart, & Korhonen, 2015; Hofmann et al., 2018; Hills, Jones, & Todd, 2012; Jones, Kintsch, & Mewhort, 2006). Most relevant, researchers have also found that word embeddings predict certain association-based probability judgments, social judgments, and consumer judgments (Bhatia, 2017, 2018; Caliskan, Bryson, & Narayanan, 2017)

In this paper we show that the structure of knowledge captured by word embeddings can be used to model a very wide range of complex human judgments, including judgments that are not easily captured by association-based measures of vector similarity. More specifically, we find that with some training data in the form of human judgments about a set of words or phrases, it is possible to learn a mapping from these entities word embeddings to the judgment dimension in consideration, and subsequently make accurate predictions for nearly any entity in that domain. In other words, we use word embeddings as feature vectors for supervised machine learning models and predict out-of-sample judgment ratings with high accuracy. We also show that these learnt mappings can be used to identify the concepts that are most related to each judgment, and thus understand the most important psychological factors underlying judgments.

Method

To illustrate the broad applicability of our method, we use study fourteen types of judgment across seven different domains of mental and behavioral life: masculinity and femininity of traits (Bem, 1974), dread and unknowability of potential risk sources (Slovic, 1987), warmth and competence of people (Rosenberg, Nelson, & Vivekananthan, 1968; Cuddy, Fiske, Glick, & Xu, 2002), taste and nutrition of

foods (Raghunathan, Naylor, & Hoyer, 2006), significance and autonomy of occupations (Hackman & Oldham, 1976), sincerity and excitement of consumer brands (Aaker, 1997), and hedonic and utilitarian value of consumer goods (Batra & Ahtola, 1990). The judgment dimensions, items, participant instructions, and various implementation details for this study and for the resulting analysis, have been pre-registered on OSF here and here.

Experimental Details

We recruited 354 participants (mean age = 31.89 years, 46.19% female) through Prolific Academic. We limited our data collection to participants who were from the U.S. and had an approval rate above 80%. Participants were only allowed to participate once, and they were paid $4.40 each. Using a between-subjects design, we randomly assigned each participant to one of the seven judgment domains brands (N = 54), consumer goods (N = 51), traits (N = 46), foods (N = 55), occupations (N = 49), risk sources (N = 49), people (N = 51). These domains were chosen to span a diverse range of cognitive and behavioral sciences. Additional details about the generation of these items and other methodological details can be found on this project's OSF page and especially supplemental information here. After being randomly assigned to one judgment domain, participants were instructed to rate 200 items (e.g., occupations) on two dimensions from -100 (e.g. not at all significant) to 100 (e.g. extremely significant), one item at a time.

Word Embeddings

For our primary analyses, we used a pre-trained word embedding model, word2vec, obtained using the skip-gram technique (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013)[1], applied to a very large dataset of Google News articles. This space has vectors for 3 million words and short phrases, with each vector being defined on 300 dimensions. Although there are other training methods as well as other pre-trained semantic spaces, we base our analysis on the Google News space because of its rich vocabulary, which includes all of the naturalistic entities used in our study (including multiword entities, such as famous people and various consumer brands, which are often absent from other spaces). This pre-trained space has also been shown to accurately capture human ratings on linguistic and semantic judgment tasks (Pereira, Gershman, Ritter, & Botvinick, 2016).

[1]This technique relies on a multilayer feedforward neural network that slides over windows of text in a large corpus, and attempts to predict the words in the periphery of the window, given the word in the center of the window. By learning to predict context words in this way, the weight matrix of the network gradually learns to encode information about the relationships between words, such that semantically related words have similar (weight) vectors. The rows of the weight matrix from the input layer to the hidden layer are precisely the word embeddings we use.

Results

Predictive Accuracy of Mapping Approach

We first evaluated the predictive accuracy of our mapping method for average participant judgments (i.e. averages of the ratings made on each the fourteen judgment dimensions). We tested the ability of a variety of (regularized) regression techniques (ridge and lasso regressions, k-nearest neighbor regression, and support vector regressions with radial basis function, linear, and polynomial kernels), across a range of hyperparameters, to map our word embeddings to judgments in a pre-registered cross-validation exercise (see pre-registration form for more details). A range of models performed well, but we focus here on our best-performing model, a ridge regression with regularization hyperparameter λ set to 10, which achieved an average r-squared of .54 and an average RMSE of 21. Figure 1 shows, for each judgment dimension, scatterplots of actual judgments and predicted judgments, along with Pearson correlation coefficients, for this method. Each predicted judgment in the scatterplot was obtained by leave-one-out cross-validation (LOOCV): we trained our ridge regression model on the vectors for all but one judgment target, and then used the trained model to predict the rating for the left-out judgment target based on the target's vector. As can be seen in Figure 1, our approach was able to predict participant judgments with a high degree of accuracy, with an average correlation rate of .77 across the fourteen judgment dimensions, and all fourteen judgments yielding statistically significant positive correlations (all $p < 10^{-20}$). Our approach can also be applied to individual-level judgments, thereby accommodating participant heterogeneity. We obtain average correlations of .52 for predicted vs. observed judgments, for the individual participants in each of our fourteen tests. These accuracy rates are lower than those obtained on the aggregate level, likely due to the fact that averaging participant ratings reduces variability in data.

Comparison to Model and Human Baselines

We then compared the vector mapping approach with a simpler, baseline approach that relies only on the relative similarity of a judgment target to words denoting high vs. low ends of a particular judgment dimension (Grand, Blank, Pereira, & Fedorenko, 2018). This method works as follows: First, we select words reflective of high and low ends of some judgment dimension. For example, the occupation significance dimension was represented by the words significant, meaningful, important and insignificant, meaningless, unimportant, pointless. Where possible, we chose words used in previous literature to define the dimensions. Then, for each judgment dimension, the average pairwise vector difference between each possible pair of high and low words is computed to obtain a single vector d representing that dimension. Last, to obtain a score for a judgment target entity on that dimension, we compute the dot product between the target entity's embedding x_i and the dimension embedding, $d * x_i$. This method essentially

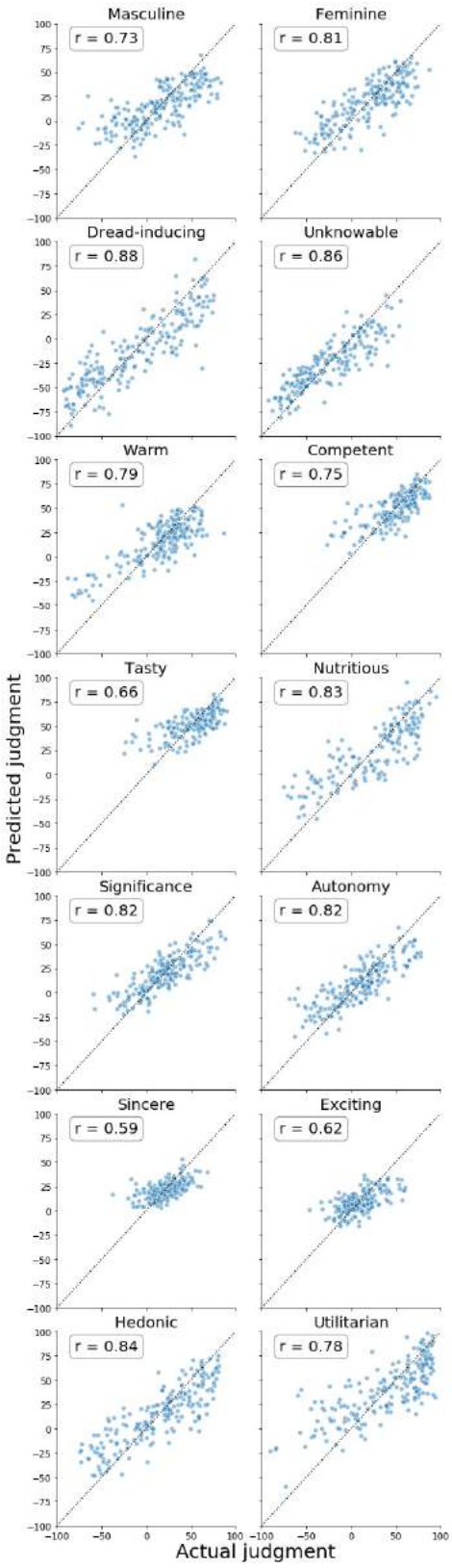


Figure 1: Scatterplots of actual judgments and predicted judgments using leave-one-out cross-validation for each judgment dimension.

computes the similarity of a judgment target (e.g., surgeon) to words high (significant, meaningful, important) relative to words low (insignificant, meaningless, unimportant, pointless) along the dimension of interest. Last, to transform these relative similarities to the range of our human judgment data, we trained OLS models predicting the human judgments from the measures of relative vector similarity, in a leave-one-out cross validation procedure. We found that the average correlation using this method was .30, which is much lower than that obtained using the vector mapping method. Additionally, the similarity method yields significant ($p < .05$) correlations for only eleven out of the fourteen tests. The baseline approach also performs worse on individual-level judgments, for which it generates average correlations of .21. As the baseline approach uses the same distances on the semantic space, for all participants, it cannot substantively accommodate participant heterogeneity (though this approach does allow for different participants to map vector similarities onto responses in different ways).[2]

We also compared the predictive accuracy of our mapping method with human inter-rater reliability, as human inter-rater reliability is often thought to place an upper bound on machine performance (Hill et al., 2015; Grand et al., 2018). To asses models predicting average models, we computed reliability two ways. First, we computed the inter-subject correlation (IS-r, (Grand et al., 2018)), which is the average correlation between one participants ratings and the average of the rest (Hill et al., 2015). This is a commonly used metric in assessing word embeddings' ability to model semantic judgments (e.g., Grand et al., 2018) and is sometimes taken to place an upper bound on machine performance (Pilehvar & Camacho-Collados, 2018). This correlation came out to 0.60, whereas our main model surpassed this with an average correlation of 0.77 across judgments. However, given that our main model is predicting an average judgment rating with word embeddings that more or less constitute the average of human knowledge reflected in word use, it may be more sensible to compare our models' performance to split-half reliability, or the correlation between the average of half the participants with the average of the other half of the participants. Thus, for each judgment dimension, we split participants into two sets, averaged judgment ratings within each set, computed the correlation between the averages, and repeated this process 100 times. The resulting split-half reliability in our judgments averaged across all judgment dimensions is .88, ranging from .69 for taste judgments to .97 for dread-inducing judgments. To assess the individual-level models relative to inter-rater reliability, we again computed reliability two ways. First, we computed the average pairwise correlation between raters (Hill et al., 2015). This correlation

[2]It is perhaps unsurprising that our baseline approach, an unsupervised method, is not as accurate as the mapping method, which is supervised. However, we maintain that this approach is the appropriate baseline to the extent that most previous applications of word embeddings in cognitive science rely on simple relative similarities like our baseline approach does.

came out to 0.34, whereas our individual-level model predictions correlated with actual judgments at an average correlation of 0.53. We can also compare individual-level model accuracy with IS-r rates, since IS-r reflects the ability to predict an individual judgment from the mean of other judgments. As stated above, mean IS-r was .60, somewhat above our average individual-level model accuracy of .53. Overall, for both average- and individual-level judgments, our model performs favorably in comparison to human inter-rater reliability, either exceeding inter-rater reliability or approaching it, depending on choice of inter-rater reliability metric.

Amount of Information Required for Prediction

A natural question for the present work is how much information in the 300-dimensional embeddings is actually required to represent our judgment targets, and hence predict our participants judgments. To this end, we measured predictive accuracy through leave-one-out cross-validation with our primary ridge model ($\lambda = 10$) after reducing the embedding spaces with principal components analysis. Specifically, for each domain, we fit a PCA on the training data design matrix (approximately 199 items, by 300 word2vec dimensions), applied the learned transformation to both the training and held-out data, discarded all but a certain number of initial principal components, and then tested how our ridge model trained on these dimension-reduced matrices predicted the held-out judgment. We emphasize that this approach obtains a *different* reduced space for every domain (cf. retraining word2vec models *for the entire vocabulary* at lower dimensional hidden layers). Figure 2 has predicted vs. actual Pearson correlations for every judgment dimension and number of retained principal components we tested. As can be seen, the 300-dimensional word embeddings can be compressed drastically to $< 10\%$ of their initial dimensionality while preserving predictive performance, with only, on average, a 3-point drop in correlation strength when retaining only the first 25 PCs, and a 7-point drop when retaining only the first 10 PCs. This suggests that, within a domain, the representational space needed to predict the present kinds of judgments is much sparser than the space provided by word2vec. Theoretically, this shows that people may only be evaluating a relative handful of (latent) dimensions when making the kinds of judgments studied here. At the same time, that much of the information relevant to making these judgments is present in the initial principal components further validates previous claims that these 14 dimensions are core dimensions along which we represent objects in these seven domains (Bem, 1974; Slovic, 1987; Rosenberg et al., 1968; Cuddy et al., 2002; Raghunathan et al., 2006; Hackman & Oldham, 1976; Aaker, 1997; Batra & Ahtola, 1990). Practically, these results indicate that future applications of the tested method need not utilize all 300 dimensions, and that successful predictions can be obtained using standard, non-regularized regression methods in the behavioral sciences applied to 10- or 25-dimensional target spaces. What kinds of information the individual principal components represent is an important question for future research, but we believe these dimension-reduced spaces are a step towards more interpretable yet highly predictive models of judgment, as a modeler now has far fewer dimensions (10 to 25, vs 300) to examine or relate to interpretable psychological quantities (by, for example, extracting the words that project onto high and low ends of the principal component's).

Psychological Substrates of Judgment

The ridge regression approach used in most of the above tests involves learning a (regularized) linear mapping from the semantic space to the judgment dimension. The best-fit weights for this mapping have the same dimensionality as the semantic space, and can thus be seen as representing a vector in this space. Judgment items whose vectors project strongly onto the weight vector (typically judgment items whose vectors are highly similar to the weight vector) will be predicted to have the highest judgment ratings. Given this interpretation, we can ask what other objects and concepts (that may not necessarily be judgment targets themselves) project strongly onto the weight vector. Intuitively, these would be the objects and concepts that are most related to the judgment, and may correspond to the judgment-relevant qualities that people evaluate when generating their responses. Thus, we took the 5000 most frequent words in the Corpus of Contemporary American English that were not also judgment targets, and fed their word2vec embeddings through our trained ridge regressions to determine their association with our 14 judgment dimensions. We then computed the difference between a words predicted association with one dimension (e.g., masculinity) and its predicted association with the complementary dimension (e.g., femininity), to find the words most strongly associated with one dimension relative to the other. Figure 3 has word clouds of these words, sized according to the strength of their association with one dimension relative to the other. These word clouds conform with expectations of the bases of these judgments. For example, traits seem to be masculine to the extent they suggest aggression, and feminine to the extent they suggest pro-sociality. A degree of artistry in a job may contribute to perceptions of autonomy, while directly guiding or helping others especially in a medical setting makes for perceptions of significance. Perceived brand sincerity may depend on brand proximity to food, family, and home; perceived brand excitement may depend on brand proximity to science, technology, and the arts.

Discussion

Despite the ubiquity of human judgment, until now we have had limited ability to predict arbitrary human judgments of objects and concepts, as capturing the rich knowledge used to make predictions has been difficult or impossible. Here we demonstrated in a pre-registered study that word embeddings, vector representations for words and concepts based on statistics of language use, proxy this knowledge and can predict 14 diverse judgments across the behavioral sciences with a high degree of accuracy. Our approach to judgment pre-

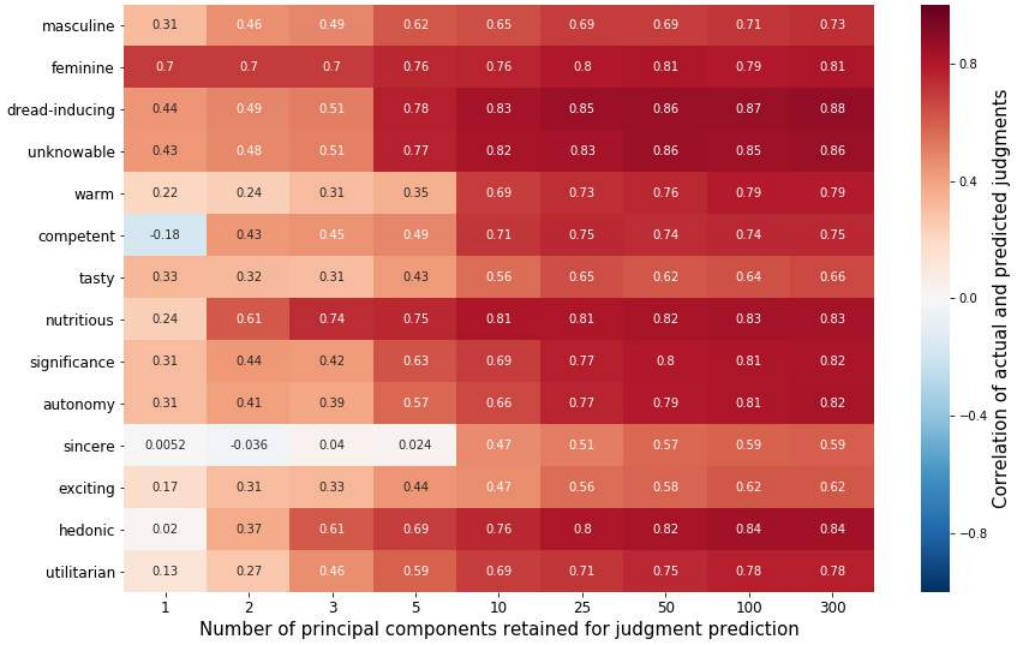


Figure 2: Pearson correlations between predicted and actual judgments for every judgment dimension and varying numbers of retained principal components. Judgment domains (brands, goods, traits, etc.) can be compressed to 5 to 25 principal components while preserving judgment prediction accuracy.

diction learning a (linear) mapping directly from word embeddings to judgment ratings surpassed a similarity-based baseline and compared favorably to human inter-rater reliability. We also showed that, despite our word embedding space (word2vec) being very rich (300 dimensions), predictive accuracy barely dropped when reducing this space to 25, 10, or even fewer dimensions, suggesting that people may only be evaluating a relative handful of pieces of information when making the present kinds of judgments. Finally, we showed that the learned mapping from word embeddings to judgments can also be used to explore the conceptual underpinnings of judgments, by mapping non-judgment target entities onto the judgment dimension.

We view the present approach as a modern extension to classical psychometric approaches used to uncover the underlying representations used for making judgments (Shepard, 1980; Slovic, 1987). However, the present approach offers several advantages over classical techniques. First, the only human data that our approach requires is a (relatively) small number of judgment ratings to train a predictive model. Once a satisfactory model has been trained, no new human psychometric data is required to predict judgments for new entities. Second, word embeddings capture more knowledge about judgment targets than can realistically be collected from human participants, especially when the relevant knowledge used to make a particular judgment is not already theoretically well-understood and thus surveyed from human participants. Capturing a great degree of knowledge leads to the high predictive accuracy we have achieved here, which we suggest may be high enough for applications in downstream behavioral sciences and technologies. For example, marketers could use predicted hedonic and utilitarian values for consumer goods to optimally advertise each of their hundreds or thousands of products, while health policy designers could use predicted risk and food perceptions to guide risk education or nutrition intervention campaigns tailored to individual perceptions.

The present research can be extended in many directions. Besides simply modeling new judgment dimensions for additional domains and entities, one promising avenue is to attempt to model different subpopulations judgments. One way to do this is simply training different regression models for different subpopulations of participants (e.g., Democrats and Republicans), but another is training word embeddings on different corpora more reflective of one population than another (e.g., MSNBC vs. Fox News articles). Under this approach, words and concepts that have somewhat different meanings and associations for different subpopulations, like the word immigrant may for Democrats and Republicans, will be located in different parts of the word embedding spaces for the corresponding representative corpora. Thus, differences in judgments about, say, the warmth and competence of immigrants, elicited from Democrats and Republicans could be predicted from their different word embeddings.

Despite the strength of our approach, it is not without limitations. Cognitive scientists, who are accustomed to inter-

Figure 3: Non-judgment target words with strong association with one judgment relative to its within-domain complement. These suggest potential conceptual underpinnings of judgments.

pretable models, may be most concerned that the dimensions of the most common word embedding techniques including word2vec, which we use here are not themselves interpretable. We attempted to mitigate this problem by using our learnt mappings to predict judgment associations for non-judgment targets, and we suggested that our PCA results were a step towards interpretable models, insofar as they reduced the number of dimensions a modeler would need to examine and relate to psychologically meaningful quantities. Another approach is to train models that predict interpretable psychological qualities that are theorized to subserve different judgments. For example, the unknowability of a potential risk source is theorized to be a composition of its observability, knowledge to the exposed, the delay of their effects, and other specific factors. Thus, one could train a model to predict these quantities from word embeddings, and then train a model predict unknowability from these predicted quantities. It is also worth pointing out that classic psychometric techniques do not always avoid this problem; multi-dimensional scaling is not guaranteed to uncover dimensions corresponding to meaningful psychological qualities. Thus, word embeddings are not always a step down in interpretability relative to other empirical methods of quantifying conceptual knowledge. Finally, cognitive scientists have traditionally focused on interpretable, explanatory models, at the expense of models that make accurate out-of-sample predictions (Yarkoni & Westfall, 2017). Of course, this is undesirable to the extent that we think a good model requires external validity; having statistically significant, interpretable model coefficients is ultimately of limited use if a model cant predict new behavior with any accuracy. Thus, our work can be seen as part of the trend to rebalance the concerns of prediction and explanation in cognitive science.

References

Aaker, J. L. (1997). Dimensions of brand personality. *Journal of Marketing Research*, 347–356.

Batra, R., & Ahtola, O. (1990). Sources of the hedonic and utilitarian measuring attitudes consumer. *Consumer Attitudes*, *2*(2), 159–170.

Bem, S. L. (1974). The measurement of psychological androgyny. *Journal of Consulting and Clinical Psychology*, *42*(2), 155–162.

Bhatia, S. (2017). Associative judgment and vector space semantics. *Psychological Review*, *124*(1), 1–20.

Bhatia, S. (2018). Semantic processes in preferential decision making. *Journal of Experimental Psychology. Learning, Memory, and Cognition*.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186.

Cuddy, A. J., Fiske, S. T., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and

competition. *Journal of Personality and Social Psychology*, *82*(6), 878–902.

Grand, G., Blank, I. A., Pereira, F., & Fedorenko, E. (2018). Semantic projection: recovering human knowledge of multiple, distinct object features from word embeddings. *arXiv preprint arXiv:1802.01241*.

Hackman, J. R., & Oldham, G. R. (1976). Motivation through the design of work: Test of a theory. *Organizational Behavior and Human Performance*, *16*(2), 250–279.

Hill, F., Reichart, R., & Korhonen, A. (2015). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, *41*(4), 665–695.

Hills, T. T., Jones, M. N., & Todd, P. M. (2012). Optimal foraging in semantic memory. *Psychological Review*, *119*(2), 431–440.

Hofmann, M. J., Biemann, C., Westbury, C., Murusidze, M., Conrad, M., & Jacobs, A. M. (2018). Simple co-occurrence statistics reproducibly predict association ratings. *Cognitive Science*, *42*(7), 2287–2312.

Jones, M. N., Kintsch, W., & Mewhort, D. J. (2006). High-dimensional semantic space accounts of priming. *Journal of Memory and Language*, *55*(4), 534–552.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).

Pereira, F., Gershman, S., Ritter, S., & Botvinick, M. (2016). A comparative evaluation of off-the-shelf distributed semantic representations for modelling behavioural data. *Cognitive Neuropsychology*, *33*(3-4), 175–190.

Pilehvar, M. T., & Camacho-Collados, J. (2018). Wic: 10, 000 example pairs for evaluating context-sensitive representations. *CoRR*, *abs/1808.09121*. Retrieved from http://arxiv.org/abs/1808.09121

Raghunathan, R., Naylor, R. W., & Hoyer, W. D. (2006). The unhealthy= tasty intuition and its effects on taste inferences, enjoyment, and choice of food products. *Journal of Marketing*, *70*(4), 170–184.

Rosenberg, S., Nelson, C., & Vivekananthan, P. (1968). A multidimensional approach to the structure of personality impressions. *Journal of Personality and Social Psychology*, *9*(4), 283–294.

Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science*, *210*(4468), 390–398.

Slovic, P. (1987). Perception of risk. *Science*, *236*(4799), 280–285.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122.

Agency Drives Category Structure in Instrumental Events

Lilia Rissman (l.rissman@let.ru.nl)
Center for Language Studies, Erasmusplein 1
Nijmegen, the Netherlands 6525 HT

Asifa Majid (asifa.majid@york.ac.uk)
Department of Psychology
York, UK YO10 5DD

Abstract

Thematic roles such as Agent and Instrument have a long-standing place in theories of event representation. Nonetheless, the structure of these categories has been difficult to determine. We investigated how instrumental events, such as someone slicing bread with a knife, are categorized in English. Speakers described a variety of typical and atypical instrumental events, and we determined the similarity structure of their descriptions using correspondence analysis. We found that events where the instrument is an extension of an intentional agent were most likely to elicit similar language, highlighting the importance of agency in structuring instrumental categories.

Keywords: thematic roles; events; categorization; tools; language production; English

Introduction

Events have event participants – an eating event, for example, involves someone who eats and something that gets eaten. There is extensive evidence that such event participants are represented in terms of abstract event participant categories, sometimes called "thematic roles" (Hafri, Trueswell & Strickland, 2018; Kako, 2006; Lakusta, Spinelli & Garcia, 2017). The category Agent, for example, contains not only the person who eats in an eating event but also the person who cooks in a cooking event and the person who builds in a building event. Thematic roles have been argued to be cross-culturally universal and part of innate knowledge (Carey, 2009; Fillmore, 1968; Strickland, 2016). At the same time, thematic roles have been persistently difficult to define in terms of necessary and sufficient conditions (Cruse, 1973; Dowty, 1991; Levin & Rappaport-Hovav, 2005). For example, the person who sees in a seeing event has fewer agentive properties than the person who eats. The upshot of this prior research is that although humans represent event participants in terms of abstract categories, the structure of these categories is not well understood.

A prominent proposal is that thematic roles have prototype structure (Ackerman & Moore, 2001; Dowty, 1991; Lakoff & Johnson, 1980). Dowty (1991), for example, explains how the arguments of English verbs appear in Subject vs. Object position in terms of Proto-Role properties. The argument with the most Proto-Agent properties (e.g., being sentient, having intention, being a causer) surfaces as Subject, whereas the argument with the most Proto-Patient properties (e.g., undergoing a change of state, being causally affected) surfaces as Object. Ackerman and Moore (2001) argue that being a bounded entity is another Proto-Patient property. Given these properties, the person who sees is a less prototypical Agent than the person who eats because it is sentient but not also a causer. While these proposals have made significant progress in understanding thematic role structure, they are limited in several ways. To fully understand how event participant categories are represented, we first need to investigate a more diverse set of categories beyond Agent and Patient, which have received the most attention. We also need to draw on more diverse forms of evidence, e.g. online psycholinguistic data. The present study achieves both of these goals: we investigate the structure of the English thematic role category Instrument, as in *Marnie sliced the bread with a knife*, using a language production task in which adult speakers described live action videos. We submitted this language description data to correspondence analysis (Greenacre, 2007), allowing us to identify similarity structure within a diverse set of instrumental events.

Within linguistics, thematic roles are often understood to be linguistic objects whose theoretical function is to explain linguistic behavior, such as argument realization. In this paper, we assume that while there may be such domain-specific role representations, there are also domain-general event participant categories that are relevant to both the syntax~semantics interface and non-linguistic event cognition. We take the more conservative position that speakers' descriptions of instrumental events reflect domain-general thematic roles.

Instrument as a Thematic Role

The Instrument role appears frequently in lists of thematic roles, dating back to the ancient Sanskrit grammarian Pāṇini. Like the roles Agent and Patient, Instrument has been characterized as having prototype structure. For example, Luraghi (2001: 388) characterizes a prototypical instrument as "an inanimate manipulable entity which occurs in a controlled state of affairs, where an agent acts intentionally." The prevalence of the Instrument role in linguistic analyses perhaps reflects the importance of tool use for building human culture. In the literature on how tool use differs across human and non-human animals (Plotnik & Clayton, 2015; Seed & Byrne, 2010; Vaesen, 2012), a tool is typically defined as a physical object distinct from the body, that an individual wields intentionally, causing a change in another object or person. We adopt this definition of tool use in the present study. Tools are important because they allow us to

extend the capabilities of our own body, allowing us to solve problems "for which evolution has not provided a rigid morphological or behavioral adaptation" (Seed & Byrne, 2010: R1032).

This definition of tool use does not directly correspond, however, to the event participant categories carved out by human language (Koenig, Mauner, Bienvenue & Conklin, 2008; Lakoff, 1968; Rissman & Rawlins, 2017). English, for example, has two primary morphosyntactic devices for talking about instruments: prepositional *with* (*Remi cut the cake with a knife*) and periphrastic *use* (*Remi used a knife to cut the cake*). In these examples, the knife is an example of a tool. When an object is being used as a tool, both *with* and *use* are appropriate to describe its role. Neither *with* nor *use*, however, is restricted to only the set of tools. *With* is possible for unintentional events (e.g., *Remi tripped and cut her dress with the scissors*). In addition, *use* is possible for instruments that play only a causally indirect role (e.g., *Remi used a stepladder to paint the ceiling*). Both *with* and *use* are also possible for body parts, where no external object extends the reach of the human body (*Remi was eating with her hands*; *Remi was using her hands to eat*). Rissman and Rawlins (2017) ultimately do not use the role Instrument in their analysis of the meanings of *with* and *use*. Thus the boundaries and structure of the Instrument category have been difficult to identify, as with other thematic roles. There is also little empirical evidence that the notion of a tool, as defined above, is a central reference point within this category.

Event Categories and Instruments

Neither *with* nor *use* map onto the category of a tool, and current analyses of the meanings of these words suggest that Instrument is not part of the grammar of English. Nonetheless, there may still be an instrumental category that speakers represent when viewing actual events in the world, and tools may be prototypical members of that category. Events can be construed in multiple ways (DeLancey, 1991). An event of someone pouring orange juice into a glass, for example, can be construed as a caused change of the orange juice from one location to another, or as a caused change to the glass by means of the orange juice. Language provides a window into the construal that is chosen by a speaker at a particular time: the description *Tito poured the orange juice* emphasizes the change of location of the juice. By contrast, *Tito filled the glass with orange juice* emphasizes the change of state of the glass and the causal role of the juice. These two descriptions reflect different ways of construing the event and thus different ways of categorizing the event participants. In this study, we take advantage of this variability to investigate semantic similarity across different types of instrumental participants. To the extent that speakers favor a particular construal of an event, as evidenced through their language, this indicates a dominant way of categorizing the participants in the event. To the extent that speakers use similar language for tools and quasi-tool participants (such as body parts), this suggests that tools and quasi-tools are represented as relatively similar semantically, and may be part of a single event participant category.

We showed adult English speakers videos of tool use as well as seven types of events in which one of the participants shares some but not all of the properties of a tool. These event conditions are displayed in Table 1. For each video, there was a Target participant: we compared linguistic encoding of the Target across all conditions. The Target participants for the example videos are underlined in Table 1. In the No State Change condition, the patient is minimally affected – this contrasts with tool use, where tools bring about a specific

Table 1: Experimental conditions. Target participants are underlined.

	Condition	Description of example video
	Tool	A woman slices a baguette with a knife
Quasi-tool actions	No State Change	A woman hits a box with a pen
	Body Part	A man knocks over a music stand with his hand
	Accidental Agent	A woman sweeps the floor with a broom, accidentally knocking over a bottle
	Causally Indirect	A woman climbs a ladder to open a window
	Locatum	A woman fills a glass with orange juice
	Means of Transit	A trip on Google Maps from Rome to Moscow by plane
	Inanimate Agent	A train rolls down a track, which bumps a red car, which moves a truck
Non-tool actions	Put Theme	A man puts a box on a shelf
	Give Theme	A woman gives a mug to another person

change in an object. In the Body Part condition, the Target is not external to the agent's body. As described above, Accidental Agent events can be described with *with* but not *use*, and Causally Indirect events, where the Target is peripheral to the force exerted on the patient, can be described with *use* but not *with*.

Locatum events are of theoretical interest because some researchers have analyzed such events (e.g., filling a glass with orange juice) in terms of a schema where a substance crosses space, rather than a tool use schema (Jackendoff, 1990). By contrast, Koenig et al. (2008) analyze such events as instrumental, as both a locatum (the orange juice) and a tool are used by an agent to achieve a goal. Similarly, Means of Transit, such as taking a trip by plane, are used by an agent to achieve a goal but are not physically manipulated. Finally, the property of being a causal intermediary has been argued to be essential to instrumentality (Croft, 1991; Talmy, 1976). In Inanimate Agent events, the Target is a causal intermediary but is not manipulated by an animate agent.

We also tested two non-tool-use conditions. In Put Theme events, an agent moved an object to an inanimate location, and in Give Theme events, an agent transferred a physical object to another agent. The Target in both of these events was the theme: themes are intermediary between a source and goal and therefore provide a parallel with tools, which are intermediary between an agent and a patient. Nonetheless, based on prior research on thematic roles (Jackendoff, 1990) we did not expect that participants would use instrumental language to describe the themes in these events.

Method

Participants

43 native speakers of British English participated. An additional four participants were tested but excluded for being native speakers of American English. Participants were tested at Radboud University in the Netherlands and at the University of York in the UK and received either course credit or £5/€5.

Design and Materials

Participants described five videos from each of 10 conditions In Table 1. Each participant saw these 50 videos in a unique random order. The events were live-action videos each lasting 4-5 seconds, with the exception of Means of Transit events. For this condition, we asked participants to describe events in which the mode of transit (e.g., train, bicycle) was construed as a means of getting from one place to another. This construal is difficult to access if participants only see a live-action event of someone riding on a train, for example. We therefore showed a video of someone planning a trip on Google Maps, with a screen capture showing someone typing in a starting point, then a destination, then a means of travel (e.g., walking, driving).

Pilot studies showed that when speakers describe instrumental events, they often omit the instrument from their descriptions (e.g., an event of a man cutting bread with a knife would simply be described as *a man was cutting some bread*). Given this tendency, we highlighted the event participants that we wanted speakers to mention by drawing red circles around them. Circles were drawn around the Target as well as around the agent and patient (or source and goal, as appropriate). A still image of the red circles appeared for two seconds prior to the beginning of the event, as in Figure 1. The circles disappeared as the video began. Means of Transit events did not include red circles.

Figure 1: Initial still image from a video of a woman slicing bread (Tool condition)

Procedure

Participants viewed each of the 50 events on a computer screen and described the events out loud. We gave participants four practice videos to familiarize them with the red circles and Means of Transit events. Speakers were told they could describe the videos in any way they liked, but they needed to mention the three objects in red circles. If a participant failed to mention one of the circled objects during a practice video, they were corrected and given another opportunity to describe the video. Participants were not corrected in the experimental trials. For the Means of Transit events, participants were told that they would see someone planning a trip on Google Maps, and they should describe the trip as if they took it themselves, as if it actually happened. The task itself took about 15 minutes.

Coding

We transcribed speakers' utterances and coded how speakers described the Target in each video (what "term" was used). In syntactic terminology, we coded the lexical item that the Target DP was a complement of. Example terms are shown in (1); these sentences are actual recorded descriptions. The Targets are underlined, terms are noted in boldface and Condition in parentheses. If a speaker described the Target in multiple ways, as in (1f), each of these terms was included. We included all terms to avoid making *a priori* assumptions about which linguistic devices would be relevant for categorizing instrumental events. As we describe below, low-frequency terms were excluded from analysis.

(1) a. The lady smashed the plate **with** a hammer. (Tool)
b. Unfortunately the man **placed** his cup onto the cupcake. (Accidental Agent)
c. A sitting man passes a scarf over to a nearby lady **using** his foot. (Body Part)
d. A woman **used** a toy stick to tap a cat on the head. (No State Change)
e. A lady wrapped the baby **in** the cloth. (Locatum)
f. A man is **holding** a cardboard box. He **lifts** it onto a shelf at about head height and **places** it on that shelf. (Put Theme)

We did not code tense and aspect markings on the verb (e.g. 1b and 1f both included the term *place*). We coded verb and verb-particle constructions as having the same term (e.g., for both *the man held the scarf* and *the man held out the scarf*, the coded term was *hold*).

We excluded trials in which the participant did not mention the Target (e.g., saying *the woman chopped up the carrot* when the Target was the cutting board). We also excluded trials in which the Target was only mentioned as the subject of a clause (e.g., saying *the woman juggled and the ball fell and knocked over the bottle* when the Target was the ball). A total of 4% of all trials were excluded for these reasons.

Results

Descriptive statistics

Across all remaining trials, participants produced 2426 term tokens and 108 term types. Given the high number of term types produced, and the resulting complexity of correspondence analysis models of these data, we focus on only the most frequently produced terms here. We selected the top 16 terms: this was the smallest number of terms needed to ensure that data from all 50 videos were included in the analysis. These top 16 terms constituted 72% of all tokens produced. The 16 most frequent terms were, from most to least frequent: *with, use, put, pick-up, on, take, place, hit, using, knock, drop, in, throw, pass, by* and *into*.

Dimensions of variation

We used correspondence analysis (Greenacre, 2007) to analyze semantic similarity across the descriptions of the 50 videos. We constructed a 16 × 50 matrix in which each cell of the matrix contained a count of how often a particular term was used to describe a particular video. From this high-dimensional space, correspondence analysis extracts dimensions such that the majority of the variance in the data set can be captured using a relatively small number of dimensions. We used the *FactoMineR* package for R (Lê, Josse, & Husson, 2008; R Core Team, 2017). Figure 2 shows the eigenvalues of each of the dimensions in the correspondence analysis, as well as the cumulative variance accounted for with each dimension. Dimensions with higher eigenvalues are more important in interpreting the structure in the semantic space. Drawing on Figure 2, we interpreted the first eight dimensions of the model, which collectively accounted for 86% of the variance.

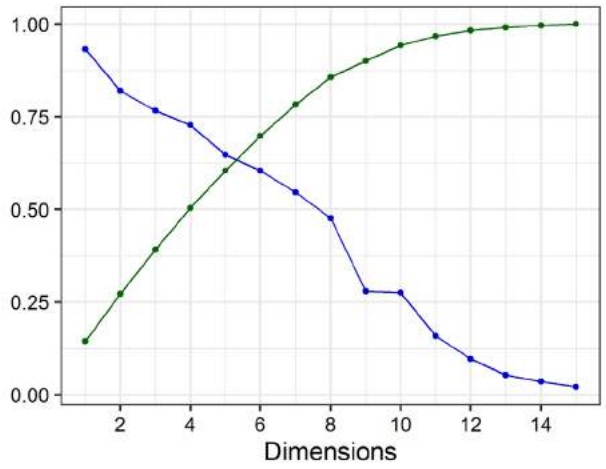


Figure 2: Eigenvalues (blue) and cumulative variance explained (green) for each of the dimensions in the analysis. The y-axis is the same for both values.

We explored how videos in the Tool condition were distinguished in this model from the other conditions. The first dimension distinguished Inanimate Agent videos from the other conditions. The most common terms for Inanimate Agent videos were *knock, hit* and *into*, terms which were rarely used for other videos. The second dimension distinguished videos involving ballistic motion, labeled with the terms *throw* and *drop*, from other videos. These ballistic motion videos came from the Put Theme and Give Theme conditions, as well as the Accidental Agent condition. In one accidental video, for example, a woman tries to juggle three balls but she accidentally drops one of them, knocking over a plastic bottle.

The third dimension grouped Give Theme and Means of Transit videos together, distinguishing them from other videos. The terms distinguished by this third dimension were *take* (e.g., *take a soda can from a woman* but also *take a train to Edinburgh*), *by* (e.g., *go to Paris by car*) and *throw* (e.g., *throw an apple to the man*). The fourth dimension distinguished two conditions from the others, but at opposite ends of the axis: Give Theme videos on one end (labeled by the term *pass*) and Causally Indirect videos on the other (labeled by the term *on*, as in *a woman chops a carrot on a cutting board*). Summarizing the first four dimensions, we see that Inanimate Agent videos are most distinct from Tool videos, followed by Give Theme, Accidental Agent and Means of Transit videos, followed by Causally Indirect videos.

The fifth dimension distinguished the terms *put* and *place* from other terms. These terms were used most often in the Put Theme condition, but also in the Locatum condition (e.g., *place groceries into a basket*) and for one of the Accidental Agent videos, as in (1b). Figure 3 shows a map of the spatial

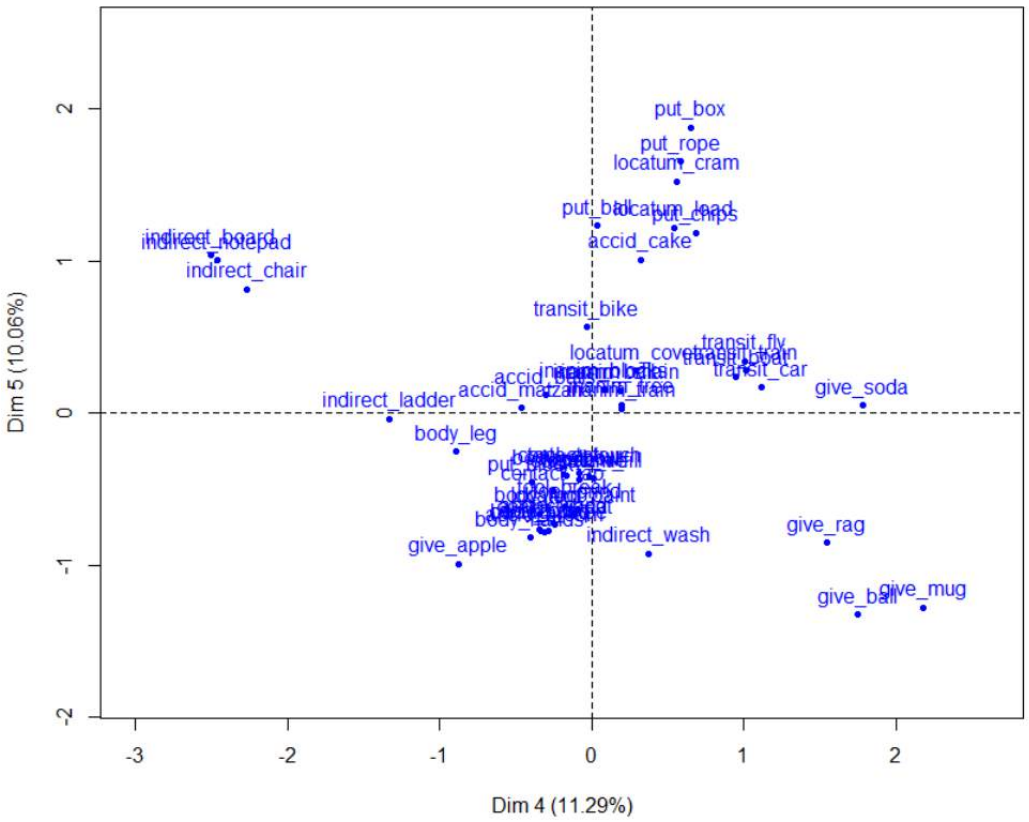


Figure 3: Individual videos plotted on Dimensions 4 and 5 of the correspondence analysis

arrangement of videos as plotted on the fourth and fifth dimensions of the correspondence analysis.

Through five dimensions, all conditions have been distinguished from the Tool condition except for No State Change and Body Part. These two conditions are not, however, distinguished by Dimensions 6-8. Dimension 6 separated transfer events in the Give Theme condition from Means of Transit events. Dimension 7 distinguished a single video in the Causally Indirect condition, where the most frequent term was *in* (e.g., *someone washed spinach in a colander*). Dimension 8 distinguished *throw* from *drop*. The correspondence analysis therefore indicates that No State Change and Body Part events have high semantic similarity to Tool events.

Focus on Tools

We further test this interpretation by analyzing in detail the data from Tool, No State Change and Body Part events taking into consideration the data which was omitted in the above analysis. As described above, 28% of the data was excluded in the correspondence analysis, and these data may reveal that English speakers do in fact categorize the Target in divergent ways across these three events. We calculated how often each term was used in each of these three conditions, as shown in Figure 4. For purposes of visualization, only the 16 most frequent terms are displayed, comprising 93% of all tokens for these three conditions. Black boxes indicate those terms which were not part of the correspondence analysis.

Figure 4 shows that the distribution of terms is similar across Tool, No State Change and Body Part conditions, the most frequent terms being *with, use* and *using*. Smaller differences are also apparent: *pick-up* was relatively common in the Tool and No State Change conditions, but not the Body Part condition. *Over, against* and *elbow* were used for Body Part events but not the other two types of events. Despite these differences, the data in Figure 4 suggest that the similarity across these three conditions observed in the correspondence analysis is not an artefact of 28% of tokens being excluded.

Discussion and Conclusion

In this study, we investigated the structure of thematic roles, focusing on participants that have been classified as Instruments in previous linguistic analysis. We showed live action videos to English speakers, and inferred how participants categorized the events based on the language they used in their descriptions. Correspondence analysis revealed which types of Target participants were described in similar ways to Tools, and which were most distinct. Inanimate Agent events were least similar to Tool events. By contrast, Causally Indirect events were more similar to Tools,

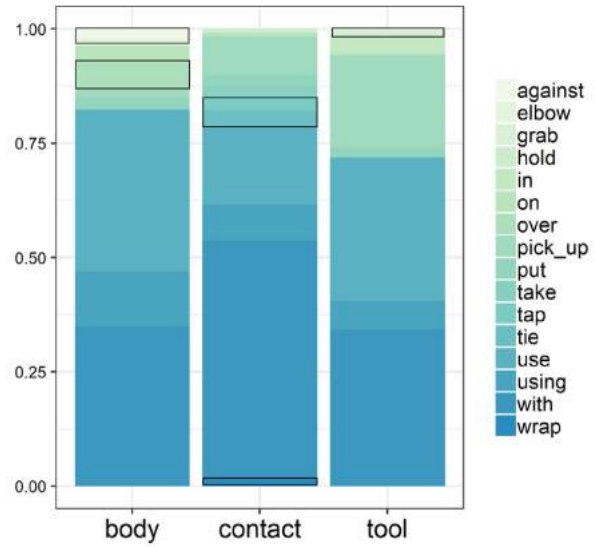


Figure 4: Distribution of the 16 most frequent terms across the Tool, No State Change and Body Part conditions. Black boxes show which terms were excluded from the correspondence analysis

distinguished only by the fourth dimension in the correspondence analysis. Although being a causal intermediary has been argued to be definitive for the Instrument category (Croft, 1991; Talmy, 1976), these results suggest that this property in fact plays a relatively weak role in shaping the categories formed by English speakers.

The results also showed that Put Theme events were more similar to Tools than Give Theme events were to Tools, although neither type of event was predicted to elicit instrumental language. This suggests a relationship between a change of location schema and a tool use schema. In an event of an agent breaking a plate with a hammer, the agent moves the hammer to the location of the plate. And although Tool events were predominantly described with *use* and *with*, not with the locative terms *put* and *place*, Locatum events formed a semantic bridge between Tool events and Put Theme events. Locatum events, such as someone putting a towel over a baby, alternated between locative encodings (e.g., *A woman picked up a towel and placed it onto a toy doll*) and Tool encodings (e.g., *The woman covered the baby with the blanket*). This semantic relationship between Instruments and Themes has been documented cross-linguistically (Bickel, Zakharko, Bierkandt & Witzlack-Makarevich, 2014), but has not been clearly noted for English before.

Previous studies of English have more often emphasized that an Instrument is an extension of an Agent (Rissman & Rawlins, 2017), and we see clear evidence for this relationship in our data. Surprisingly, the terms used for Body Part events were highly similar to the terms used for Tool events. The idea that tools are external to our body, and can therefore extend our reach, is crucial to the role of tools in the development of human culture. *A priori*, we therefore expected that Tool and Body Part events would be categorized in different ways. We did not find a strong distinction between these events, however, suggesting the importance of conceptualizing Instruments as an extension of the Agent. The fact that No State Change events were also similar to Tool events supports this conclusion: the intention and actions of the Agent are more important than the actual outcome. To the extent that tools are prototypical instances of instrumental events, the instruments in Body Part and No State Change events are no less prototypical.

In the video stimuli in this study, we circled the Target participants, in addition to agents, patients, sources and goals, in order to prompt speakers to mention these participants. This likely did affect speakers' construal of the events – in fact, it was our goal to direct speakers to a construal where the Target had high prominence, high enough to be mentioned. We do make the assumption that the descriptions we elicited using these circles would not differ significantly from descriptions where speakers mention the Targets spontaneously, without prompting.

In conclusion, we find that agency plays a prominent role in determining similarity across instrumental events. These conclusions, however, only extend as far as English, and how English speakers conceptualize events. Future research can determine the extent to which similar principles guide categorization in other languages and other cultures.

Acknowledgments

This research was supported by a Radboud Excellence Initiative postdoctoral fellowship awarded to Lilia Rissman, the Radboud University Center for Language Studies, and the University of York Department of Psychology. Thank you to all research participants.

References

Ackerman, F., & Moore, J. (2001). Proto-properties and grammatical encoding. *Stanford Monographs in Linguistics. Stanford: CSLI*.

Bickel, B., Zakharko, T., Bierkandt, L., & Witzlack-Makarevich, A. (2014). Semantic role clustering: An empirical assessment of semantic role types in non-default case assignment. *Studies in Language. International Journal sponsored by the Foundation "Foundations of Language", 38*(3), 485-511.

Carey, S. (2009). *The origin of concepts*. New York, NY US: Oxford University Press.

Croft, W. (1991). *Syntactic categories and grammatical relations: the cognitive organization of information*. Chicago: University of Chicago Press.

Cruse, D. A. (1973). Some thoughts on agentivity. *Journal of Linguistics, 9*(1), 11-23.

DeLancey, S. (1991). Event Construal and Case Role Assignment. In L. Sutton, C. Johnson, & R. Shields (Eds.), *Proceedings of the 17th Annual Meeting of the Berkeley*

Linguistics Society (pp. 338-353). Berkeley, CA: Berkeley Linguistics Society.
Dowty, D. (1991). Thematic proto-roles and argument selection. *Language, 67*(3), 547-619.
Fillmore, C. J. (1968). The case for case. In E. W. Bach & R. T. Harms (Eds.), *Universals in linguistic theory* (pp. 210). New York: Holt, Rinehart and Winston.
Greenacre, M. J. (2007). Correspondence analysis in practice *Interdisciplinary statistics series* (Vol. 2, pp. 280). Boca Raton: Chapman & Hall/CRC.
Hafri, A., Trueswell, J. C., & Strickland, B. (2018). Encoding of event roles from visual scenes is rapid, spontaneous, and interacts with higher-level visual processing. *Cognition, 175*, 36-52.
Jackendoff, R. (1990). *Semantic structures*. Cambridge, MA: MIT Press.
Kako, E. (2006). Thematic role properties of subjects and objects. *Cognition, 101*(1), 1-42.
Koenig, J.-P., Mauner, G., Bienvenue, B., & Conklin, K. (2008). What with? The Anatomy of a (Proto)-Role. *Journal of Semantics, 25*(2), 175-220.
Lakoff, G. (1968). Instrumental Adverbs and the Concept of Deep Structure. *Foundations of Language, 4*(1), 4-29.
Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. Chicago: University of Chicago Press.
Lakusta, L., Spinelli, D., & Garcia, K. (2017). The relationship between pre-verbal event representations and semantic structures: The case of goal and source paths. *Cognition, 164*, 174-187.
Lê, S., Josse, J., & Husson, F. (2008). FactoMineR: An R Package for Multivariate Analysis. *2008, 25*(1), 18.
Levin, B., & Rappaport-Hovav, M. (2005). *Argument realization*. Cambridge, New York: Cambridge University Press.
Luraghi, S. (2001). Some remarks on Instrument, Comitative, and Agent in Indo-European. *STUF - Language Typology and Universals, 54*(4), 385-401.Plotnik, J. M., & Clayton, N. S. (2015). Convergent cognitive evolution across animal taxa: comparisons of chimpanzees, corvids and elephants. In E. Margolis & S. Laurence (Eds.), *The conceptual mind: New directions in the study of concepts*, pp. 29-56.
R Core Team (2017). R: A language and environment for statistical computing. from R Foundation for Statistical Computing https://www.R-project.org/.
Rissman, L., & Rawlins, K. (2017). Ingredients of Instrumental Meaning. *Journal of Semantics, 34*(3), 507-537.
Seed, A., & Byrne, R. (2010). Animal Tool-Use. *Current Biology, 20*(23), R1032-R1039.
Strickland, B. (2016). Language Reflects "Core" Cognition: A New Theory About the Origin of Cross-Linguistic Regularities. *Cognitive Science*.
Talmy, L. (1976). Semantic causative types. In M. Shibatani (Ed.), *The Grammar of Causative Constructions* (Vol. 6, pp. 43-116). New York: Academic Press.
Vaesen, K. (2012). The cognitive bases of human tool use. *Behavioral and Brain Sciences, 35*(04), 203-218.

Auditory Stimuli Disrupt Visual Detection in a Visuospatial Task

Christopher W. Robinson (robinson.777@osu.edu)
Department of Psychology, The Ohio State University at Newark
1179 University Dr., Newark, OH 43056, USA

Dylan Laughery (laughery.12@osu.edu)
Department of Psychology, The Ohio State University at Newark
1179 University Dr., Newark, OH 43056, USA

Abstract

The current study used an eye tracker to examine how auditory input affects the latency of visual fixations and speeded responses on a Serial Response Time Task (SRTT). In Experiment 1, participants viewed a sequence of visual stimuli that appeared in different locations on a computer monitor and the same sequence repeated throughout the experiment. The visual sequence was either presented in silence or paired with uncorrelated sounds (i.e., sounds did not predict visual target location). Participants made more fixations and were more likely to fixate on the visual stimuli when visual sequences were presented in silence than when paired with sounds. Participants in Experiment 2 were presented with the same sequences, but they also had to determine if each visual stimulus was red or blue. The presence of auditory stimuli had no effect on accuracy (red vs. blue), however, there was some evidence that auditory stimuli delayed the latency of first fixations to the visual stimuli and discriminating the images as red or blue was also slower relative to the unimodal visual baseline. While visual stimuli often dominate auditory processing on spatial tasks, the current findings show that auditory stimuli can also slow down visual detection on a task that is better suited for the visual modality. These findings are consistent with a potential mechanism underlying auditory dominance effects, which posits that auditory stimuli may attenuate and/or delay the encoding of visual information.

Keywords: Attention, Multisensory Processing, Auditory Dominance

Introduction

Over the last 40 years, there has been a considerable amount of research examining how individuals process and integrate multisensory information (see Bahrick, Lickliter, & Flom, 2004; Calvert, Spence & Stein, 2004; Robinson & Sloutsky, 2010; Spence, Parise, & Chen, 2012; Stein & Meredith, 1993, for reviews). Much of this research focuses on multisensory integration where information from different sensory modalities is quickly, if not automatically, bound into a multisensory percept in which processing and responding to these multisensory percepts is often faster and more efficient than responding to the unisensory information (Bahrick, Flom, & Lickliter, 2002; Fort, Delpuech, Pernier, & Giard, 2002; Giard & Peronnet, 1999; Miller, 1982). For example, localizing a visual stimulus paired with a sound is often faster than localizing a visual stimulus presented in silence.

However, there are also many situations where multisensory information is arbitrary in nature and information presented to one sensory modality is unrelated to the information presented to the other sensory modality (e.g., listening to music while visually navigating traffic). Under these situations, multisensory presentation can sometimes disrupt encoding, learning, and/or responding, with one sensory modality dominating processing of the other sensory modality. For example, modality dominance research in adults often shows that when auditory and visual stimuli are presented simultaneously, visual input often dominates processing of auditory information (Colavita, 1974; Sinnett, Spence, & Soto-Faraco, 2007; see also Spence et al., 2012, for a review).

There is recent evidence of auditory dominance in adults (Barnhart, Rivera, & Robinson, 2018; Dunifon, Rivera, & Robinson, 2016; Robinson, Moore, & Crook, 2018), however, research pointing to auditory dominance in adult populations typically relies on temporal tasks (e.g., Parker & Robinson, 2018; Robinson & Sloutsky, 2013; Shams et al., 2000; 2002). More specifically, while the auditory modality can sometimes dominate visual processing on temporal tasks, the visual modality typically dominates auditory processing on spatial tasks (Welch & Warren, 1980). These findings suggest that modality dominance effects are flexible in nature and vary as a function of response demands (Robinson, Chandra, & Sinnett, 2016), nature of the task (Welch & Warren, 1980), and signal strength (Alias & Burr, 2004).

Given that auditory dominance effects are less prevalent in the adult literature, the primary goal of the current paper was to focus on these effects. One potential mechanism underlying auditory dominance is that sensory modalities might be competing for attention (Robinson & Sloutsky, 2010; see also Duncan, Martens, & Ward, 1997; Eimer & Driver, 2000; Sinnett et al., 2007; Wickens, 1984, for related discussions). Moreover, because auditory stimuli are often dynamic and transient in nature, it would be adaptive to first allocate attention to this information before it disappears. Attentional resources automatically deployed to the auditory modality might come with a cost - disrupted or delayed visual processing. There is some support for this claim from studies using temporal and recognition tasks (Barnhart et al., 2018; Dunifon et al., 2016; Parker & Robinson, 2018; Robinson et al., 2018, Robinson & Sloutsky, 2013; Shams et al., 2000; 2002), however, a stronger test of this proposed mechanism would be to examine if auditory stimuli also delay visual

processing on a visuospatial task, a task better suited for the visual modality.

A recent study presented adults with a SRTT, which was administered on a touch screen computer (Robinson & Parker, 2016). As in previous research using variations of this paradigm (e.g., Dennis, Howard, & Howard, 2006; Nissen & Bullemer, 1987; Song, Howard, & Howard, 2008), Robinson and Parker (2016) presented visual information to spatially distinct locations, and participants had to quickly respond to this information (i.e., they had to touch each stimulus when it appeared on the touch screen monitor). Unbeknownst to participants, the visual sequences were structured and followed the same sequence throughout the experiment. Motor responses sped up over the time suggesting that, at some level, participants were learning the sequences. More relevant to the current study, motor responses to the visual stimuli were slower when the visual stimuli were paired with uncorrelated sounds (i.e., sounds that did not predict/correspond with location of the visual stimulus).

The current study expands on this research by using variations of a SRTT administered on an eye tracker to examine patterns of visual fixations over time. In both reported experiments, participants were shown two visual sequences of 12 stimuli, and the same sequences repeated throughout the experiment. In one condition, the sequence was presented in silence (unimodal condition) and in the other condition, the visual sequence was paired with sounds that were not correlated with the spatial location of the visual stimuli (cross-modal condition). Participants either counted the number of visual stimuli (Experiment 1) or they responded to each stimulus by quickly making a distinction on whether visual stimulus was red or blue (Experiment 2). If auditory stimuli are disrupting visual detection/encoding, then latency of first fixations to the visual stimuli should also be delayed. However, if auditory stimuli are disrupting later stages of visual processing (e.g., response/decision phase), then auditory interference should only be found in Experiment 2 when participants are making explicit responses to each stimulus.

Experiment 1

Method

Participants Forty undergraduate students (M = 19.41 years, SD = 1.61 years, 22 Females, one person did not disclose gender or age information) from The Ohio State University at Newark participated in the experiment for course credit. Data from 11 other participants were excluded from the study due to technical difficulties such as poor calibration, software crashes, etc.

Apparatus Participants were centrally positioned and seated approximately 65 cm in front of an EyeLink 1000 Plus eye tracker with desktop mount and remote camera. The eye tracker computed eye movements at a rate of 500 Hz, and Experiment Builder 1.10.165 controlled the timing of stimulus presentations. Visual stimuli were presented on a BenQ XL2420 24" 1920 x 1080 monitor and auditory stimuli were presented via Kensington 33137 headphones. Eye tracking data were collected and stored on a Dell Optiplex 7010 computer. Gaze fixation positions, Areas Of Interest (AOIs), and fixations were identified by the EyeLink system and data were exported using Data Viewer. The eye tracker, stimulus presentation computer, and eye tracking computer were stationed in a quiet testing room and a trained experimenter oversaw the entire duration of each participant's study.

Materials and Design Visual stimuli were solid red and blue circles (100 pixels in diameter) and were presented on a white background. Visual stimuli were presented for 700 ms and were presented one right after another with no interstimulus interval. Auditory stimuli were 6 sine waves (500 - 3000 Hz, each stimulus increasing by 500 Hz) and 6 sawtooth waves (250 - 2750 Hz, each increasing by 500 Hz) Auditory stimuli were created in Audacity and were presented via headphones at a comfortable level - approximately 65 dB. Auditory stimuli were presented for 500 ms, and the auditory and visual stimuli shared the same onset.

The experiment consisted of two within-subjects conditions: a unimodal condition and a cross-modal condition. We presented two visual sequences of 12 distinct circle locations that repeated 20 times (see Figure 1 for sequences). In the unimodal condition, the sequence was presented in silence, and in the cross-modal condition, the visual sequence was paired with sounds. The color of the circles in both conditions was random (not correlated with the location of the circle), as were the sounds in the cross-modal condition. The order of the two sequences and the sequence-condition pairings were counterbalanced across participants.

Procedure Participants were told that they would see red and blue circles appear one at a time in different locations across the screen. They were instructed to look at the circles as they appear and respond by pressing a USB button placed in front of them after every 10 circles that they saw. Participants were not told that the circles would appear in the same sequence of 12 locations, however, they were informed that the study was split into two parts, a silent condition and a sound condition. Participants were given a consent form and demographics form to fill out before the study began.

After completing the consent and demographic forms, participants were calibrated on the eye tracker. Drift correction occurred every 50 stimuli (approximately 40 s), and we recalibrated the eye tracker every 100 stimuli (approximately 80 s). When the experiment concluded, the participants were given a three-question survey. On each item, they had to determine if they thought the order of the visual sequence, the order of the visual sequence paired with the sounds, and the order of the auditory sequence, was random or followed a pattern which repeated throughout the experiment. Question order was counterbalanced across participants (e.g., participants who received the unimodal

condition first were first asked about the unimodal sequence and vice versa).

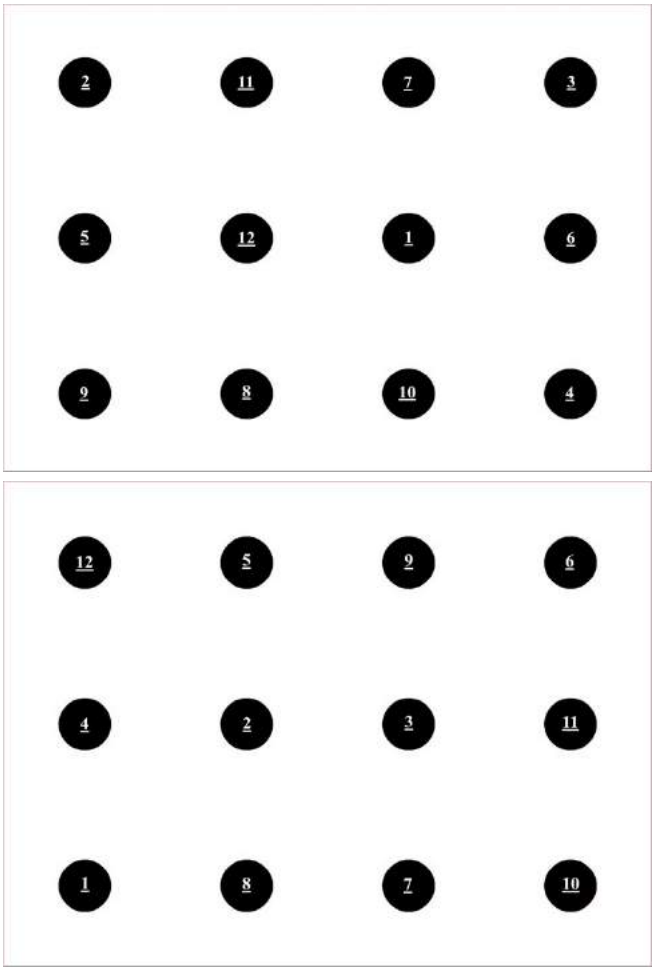


Figure 1: Order of the two visual sequences.

Results

We examined three different eye tracking variables which might be affected by the presence of auditory stimuli. First, we examined the latency of first fixations, which are sometimes slowed down on recognition tasks when visual stimuli are paired with sounds or words (Barnhart et al., 2018; Dunifon et al., 2016). Second, we examined the proportion of stimuli where participants made a fixation to the target location. If auditory stimuli disrupt visual encoding, there should be fewer fixations to the visual targets in the sound condition. Finally, we examined the number of fixations on each trial, however, these predictions are less clear. For example, attention automatically deployed to the auditory modality (or away from the visual modality) could reduce the overall number of fixations or it could make the visual task more challenging and require more fixations before detecting the visual target.

Each participant reported in the final sample completed the unimodal condition and the cross-modal condition, and in each condition, participants were presented with an ordered sequence of 12 visual stimuli, which repeated 20 times. Each sequence of 12 stimuli was considered as a trial, and to reduce noise, we created four blocks by averaging across five trials (60 stimuli). Thus, each condition consisted of four blocks of five trials, with 60 stimuli per block (e.g., Block 1 = first 60 stimuli, Block 2 = 61 - 120, etc.).

Every 700 ms a visual stimulus appeared in one of 12 pre-specified locations on the monitor and we recorded the latency first fixation to the visual stimulus (timestamp of first fixation to AOI - timestamp of stimulus onset). AOIs were created in Data Viewer and were 300 x 300 pixel squares centered around each visual stimulus. We submitted the mean latency of first fixations to a 2 (condition: unimodal, cross-modal) x 4 (block: blocks 1-4) repeated measures ANOVA. Mean latency of first fixations across condition and time ranged from 256 - 261 ms. There were no significant effects and the interaction did not reach significance, *p*s > .31.

We also examined the proportion of stimuli where participants made a fixation to the AOIs. If a participant made a fixation to the location of the target from stimulus onset to stimulus offset, then we coded that stimulus as a 1. If a participant did not make a fixation to the AOI during this time window, then we coded that stimulus as a 0. Proportions of fixations to the AOIs were averaged within each block and we submitted these values to a 2 (condition: unimodal, cross-modal) x 4 (block: blocks 1-4) repeated measures ANOVA. The analyses only revealed a marginally significant effect of condition, F (1, 39) = 3.95, p = .054, η_p^2 = .092, with participants making a higher proportion of fixations to the AOIs in the unimodal condition (M = .94, SE = .01) than in the cross-modal condition (M = .92, SE = .02).

The number of fixations from stimulus onset to stimulus offset (to any location on the monitor) was collected and we submitted these values to a 2 (condition: unimodal, cross-modal) x 4 (block: blocks 1-4) repeated measures ANOVA. The analysis only revealed an effect of condition, F (1, 39) = 5.66, p = .022, η_p^2 = .127, with participants making more fixations in the unimodal condition (M = 2.42, SE = .03) than in the cross-modal condition (M = 2.39, SE = .03).

Finally, at the end of the experiment, participants completed a three-item questionnaire. Random and Patterned responses for the unimodal and cross-modal conditions were analyzed using a McNemar's Chi-square. The McNemar Chi-square was significant (N = 40), p = .049, and one sample binomial tests compared to chance revealed that a majority of the participants thought the unimodal visual sequences were random (M = 68% reported random, p = .04), whereas, only 45% of the participants indicated that the visual sequences paired with sounds were random, which did not differ from chance, p = .64. Forty-five percent of participants also reported that the order of the auditory sequence was random.

In summary, while previous research demonstrated that auditory stimuli can slow down first fixations on recognition tasks (Barnhart et al., 2018; Dunifon et al., 2016) and slow down motor responses on a touch screen SRTT (Robinson & Parker, 2016), the current study found only weak support for

auditory interference. More specifically, participants in the current study were slightly less likely to make a fixation to the visual stimulus when it was paired with a sound and they also made fewer fixations (to any location on the monitor). However, unlike Robinson and Parker (2016), there was no evidence that participants learned the visual sequences. Recall that latency of first fixations to the target locations did not speed up across training, whereas, motor responses sped up in Parker and Robinson (2016). Finally, while a majority of participants thought the unimodal visual sequences were random, participant responses did not differ from chance when sequences were paired with sounds. It is unclear if the uncorrelated sounds increased the perceived structure of visual input or if the sounds simply increased chance responding. However, if the sounds did increase the perceived structure of visual sequences, it did not result in faster or more fixations to the visual targets.

Experiment 2

The primary aim of the Experiment 2 was to further examine possible effects of auditory stimuli on visual sequence learning. Are interference effects restricted to tasks that require an explicit response? To address this aim, we presented participants with structured visual sequences in silence or paired with sounds, however, in contrast to Experiment 1, participants were required to make a response to each visual stimulus (i.e., indicate if the visual target was red or blue). If auditory stimuli interfere with visual processing during the decision/response phase, as opposed to disrupting encoding, then response times should slow down in the cross-modal condition in Experiment 2 while having no negative effect on the latency of first fixations. However, slowed response times and delayed first fixations would be consistent with the claim that auditory stimuli are disrupting visual encoding (Robinson & Sloutsky, 2010a).

Experiment 2 was not originally designed to examine the effects of engagement on sequence learning, however, requiring participants to make an explicit response to each stimulus should make the task more engaging. Thus, it is also possible to examine if poor engagement could account for the lack of learning in Experiment 1. While visual sequence learning on SRTT and statistical learning tasks are often thought to be implicit in nature and not dependent on attention (e.g., Nissen & Bullemer, 1987; Saffran, Newport, Aslin, & Tunick, 1997), it is possible that learning would be more robust if participants were more engaged throughout testing. Requiring participants to indicate if each visual stimulus is red or blue should make the task more engaging, which could result in better learning (i.e., faster response times and/or fixations across time).

Method

Participants, Materials, Design, and Procedure Thirty undergraduate students (M = 20.19 years, SD = 2.51 years, 20 Females) from The Ohio State University at Newark participated in the experiment for course credit. Data from eight other participants were excluded from the study due to technical difficulties, such as software/system crashes, computer lagging, or poor calibrations.

The procedure and design of Experiment 2 were identical to Experiment 1, except that in Experiment 2, a choice response task paradigm was used. Participants were required to make a color distinction with each stimulus by responding with one of two external USB buttons, labeled "RED" and "BLUE" respectively. Participants were instructed to respond as fast and as accurate as possible. The left-right locations of the buttons were counterbalanced across participants.

Results

As in Experiment 1, we examined the latency of first fixations, the proportion of stimuli where participants made a fixation to the visual target, and the number of fixations between stimulus onset and stimulus offset, however, we also examined response times and accuracies on the primary task.

First, as in Experiment 1, we submitted the mean latency of first fixations to a 2 (condition: unimodal, cross-modal) x 4 (block: blocks 1-4) repeated measures ANOVA. The analyses revealed a marginally significant effect of condition, F (1, 29) = 3.62, p = .067, η_p^2 = .111, and a significant time x condition interaction, F (3, 87) = 3.28, p = .025, η_p^2 = .102. While latency of first fixations were numerically faster across all blocks in the unimodal condition, simple effects with Bonferroni adjustments revealed that the difference between unimodal and cross-modal means only reached significance in block 3, p = .012 (see Figure 2).

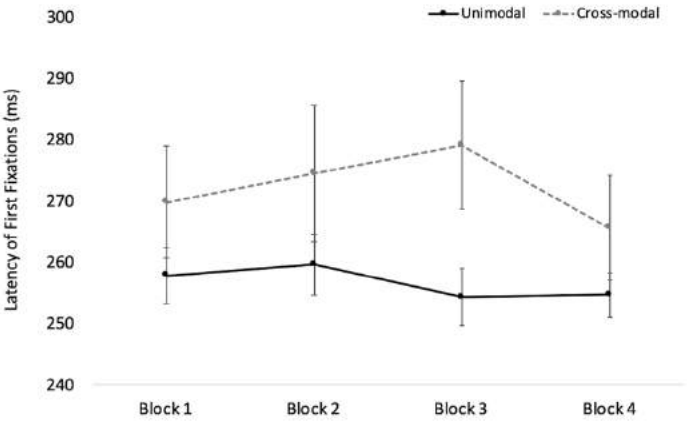


Figure 2. Latency of First Fixations (ms) across condition and time. Error bars denote Standard Errors.

Response times were also submitted to a 2 (condition: unimodal, cross-modal) x 4 (block: blocks 1-4) repeated measures ANOVA. The analysis revealed an effect of condition F (1, 29) = 5.90, p = .022, η_p^2 = .169, with response times being faster in the unimodal condition (M = 584 ms, SE = 19.85) than in the cross-modal condition (M = 624 ms, SE = 28.56). The analysis also revealed an effect of time, F (3, 87) = 16.04, p < .001, η_p^2 = .356. See Figure 2 for response

times across condition and time. Pairwise comparisons with Bonferroni adjustments revealed that mean response times on Block 1 (M = 637 ms, SE = 27.92) were significantly slower than Block 2 (M = 600 ms, SE = 24.18), Block 3 (M = 592 ms, SE = 22.72), and Block 4 (M = 587 ms, SE = 19.48), ps < .001. Blocks 2-4 did not differ, ps > .56. Also note that accuracies (i.e., discriminating red vs. blue stimuli) exceeded .96 across all conditions with no significant effects or interactions, ps > .22.

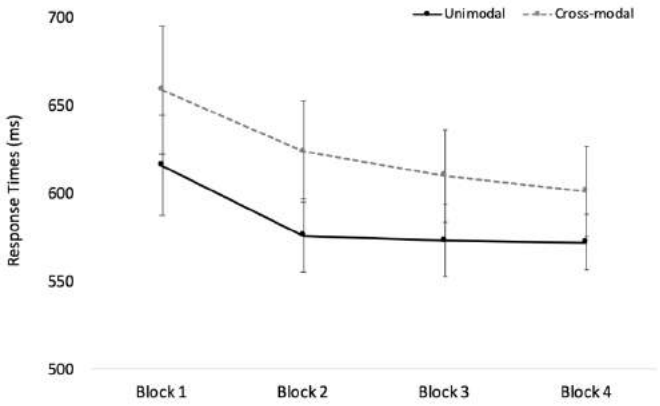


Figure 3. Response times (ms) across condition and time. Error bars denote Standard Errors.

The proportion of stimuli where participants made a fixation to the AOI were submitted to a 2 (condition: unimodal, cross-modal) x 4 (block: blocks 1-4) repeated measures ANOVA. Mean proportion of fixating to the visual stimuli ranged from .93 - .97, and the analysis revealed no significant effects or interactions, ps > .32.

The mean number of fixations between stimulus onset and offset were submitted to a 2 (condition: unimodal, cross-modal) x 4 (block: blocks 1-4) repeated measures ANOVA. The analysis only revealed a marginally significant effect of time, F (3, 87) = 2.62, p = .056, η_p^2 = .083. Pairwise comparisons with Bonferonni adjustments revealed that participants made more fixations on block 1 (M = 2.24, SE = .05) than on block 2 (M = 2.18, SE = .05), p = .023. Block 1 did not differ from block 3 (M = 2.20, SE = .05) or block 4 (M = 2.19, SE = .05), and blocks 2-4 did not differ, ps > .323.

Finally, we also examined responses on the three-item questionnaire, which was administered at the end of the study. Two participants did not complete the questionnaire. Patterned responses for the unimodal visual and visual sequence paired with sounds were analyzed using a McNemar's Chi-square. The McNemar Chi-square was not significant (N = 28), p > .99, Binomial tests compared to chance revealed that 78% of the participants thought the order of the unimodal visual sequence was random, different from chance, p = .004, and 82% of the participants reported that the order of the visual sequence paired with sounds was also random, different from chance, p = .001.

General Discussion

In both reported experiments, participants were shown two visual sequences, and each sequence repeated 20 times over the course of the experiment. One sequence was presented in silence, whereas, the other sequence was paired with sounds, which were not correlated with the location of the visual stimulus. In Experiment 1, participants simply counted the number of visual stimuli, pressed a button after every 10 stimuli, and we examined visual fixations throughout the procedure. Experiment 2 was more engaging, as participants were required to quickly determine if each visual stimulus was red or blue.

Auditory interference effects were found in both experiments. More specifically, in Experiment 1 when participants counted the number of visual stimuli, participants were more inclined to fixate on the visual stimuli in the unimodal condition and also made more overall fixations in the unimodal condition. When participants had to determine if each visual stimulus was red or blue, both latency measures showed some evidence of a slowdown/delay in the cross-modal condition. More specifically, latency of first fixations to the visual stimuli was slower in the cross-modal condition compared to the unimodal baseline, especially in block 3 (see Figure 2). In addition, overall response times were also slower in the cross-modal condition than in the unimodal condition.

The current study contributes to modality dominance research in the following ways. First, most research examining modality dominance in adults often points to visual dominance, with the visual modality dominating auditory processing (Spence, 2009; Spence et al., 2012, for reviews). While the current study did not examine the effects of visual input on auditory processing, the findings provide support for auditory dominance with auditory stimuli slowing down visual fixations and responding. These findings are remarkable given that spatial tasks are typically better suited for the visual modality (Welch & Warren, 1980). Moreover, the current study examined latency of first fixations as well as response times. If sounds were simply interfering during the response/decision phase, then only response times should have been slowed down. Finding evidence that first fixations to the stimuli were also delayed suggests that interference effects are happening early in the course of visual processing (i.e., during the detection phase).

While these findings shed light on the dynamics of multisensory processing, there are some limitations to the current study. First, while response times sped up in Experiment 2, there was no evidence in the eye tracking data that participants were learning the sequences. There are several reasons why learning may have not occurred. First, in both reported experiments, the color of the visual stimuli added noise to the sequences and the sounds in the cross-modal conditions also added additional noise (i.e., participants may have focused on these irrelevant variables and failed to learn the sequences). However, this additional information should not have affected sequence learning if the task is assessing implicit learning. It is also possible that

participants were learning the sequences, but we failed to capture this learning because we primarily focused on participants' responses to visual stimuli and not on their anticipations (fixations before stimulus onset). These possibilities need to be addressed in future research.

In summary, the current study demonstrates that sounds can disrupt visual stimulus detection and response times. These effects have implications on tasks that require processing of multisensory information and shed light on possible mechanisms underlying auditory dominance effects.

References

Alias, D., & Burr., D. (2004). The Ventriloquist Effect Results from Near-Optimal Bimodal Integration. *Current Biology*, *14*, 257 - 262.

Bahrick, L. E., Flom, R., & Lickliter, R. (2002). Intersensory redundancy facilitates discrimination of tempo in 3-month-old infants. *Developmental Psychology*, *41*(4), 352-363.

Bahrick, L. E., Lickliter, R., & Flom, R. (2004). Intersensory redundancy guides the development of selective attention, perception, and cognition in infancy. *Current Directions in Psychological Science*, *13*, 99 –102.

Barnhart, W. R., Rivera, S., & Robinson, C. W. (2018). Different patterns of modality dominance across development. *Acta Psychologica*, *182*, 154-165.

Calvert, G., Spence, C., & Stein, B. (2004). *The handbook of multisensory processe*s. Cambridge, MA: MIT Press.

Colavita, F. B. (1974). Human sensory dominance. *Perception & Psychophysics*, 16, 409-412.

Dennis, N.A., Howard, J.H., & Howard, D.V. (2006). Implicit sequence learning without motor sequencing in young and old adults. *Experimental Brain Research*, *175*, 153–164.

Duncan, J., Martens, S., & Ward, R. (1997). Restricted attentional capacity within but not between sensory modalities. *Nature*, *387*, 808-810.

Dunifon, C. M., Rivera, S., & Robinson, C. W. (2016). Auditory stimuli automatically grab attention: Evidence from eye tracking and attentional manipulations. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(12), 1947-1958.

Eimer, M., & Driver, J. (2000). An event-related brain potential study of cross-modal links in spatial attention between vision and touch. *Psychophysiology*, *37*(05), 697-705.

Fort, A., Delpuech, C., Pernier, J., & Giard, M. H. (2002). Dynamics of cortico-subcortical cross-modal operations involved in audio–visual object recognition in humans. *Cerebral Cortex*, *12*, 1031–1039.

Giard, M.H., & Peronnet, F. (1999). Auditory-visual integration during multimodal object recognition in humans: A behavioral and electrophysiological study. *Journal of Cognitive Neuroscience*, *11*(5), 473-490.

Miller, J. (1982). Divided attention: Evidence for coactivation with redundant signals. *Cognitive Psychology*, *14*, 247-279

Nissen, M.J., Bullemer, P. (1987). Attentional requirements of learning: Evidence from performance measures. *Cognitive Psychology*, *19*, 1–32.

Parker., J.L., & Robinson, C.W. (2018). Changes in multisensory integration across the lifespan. *Psychology and Aging*, *33*(3), 545-558.

Robinson, C.W., Chandra, M., & Sinnett, S. (2016). Existence of competing modality dominances. *Attention, Perception, & Psychophysics*, *78*, 1104-1114.

Robinson, C. W., Moore, R. L., & Crook, T. A. (2018). Bimodal presentation speeds up auditory processing and slows down visual processing. *Frontiers in Psychology*, *9*, 1-10.

Robinson, C.W., & Parker., J.L. (2016). Effects of auditory input on a spatial serial response time task. In Papafragou, A., Grodner, D., Mirman, D., & Trueswell, J.C. (Eds.) *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 2237-2242). Austin, TX: Cognitive Science Society.

Robinson, C. W., & Sloutsky, V. M. (2013). When audition dominates vision: Evidence from cross-modal statistical learning. *Experimental Psychology*, *60*, 113-121.

Robinson, C. W., & Sloutsky, V. M. (2010). Development of cross-modal processing. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*, 135-141.

Saffran, J. R., Newport, E. L., Aslin, R. N., Tunick, R. A., & Barrueco, S. (1997). Incidental language learning: Listening (and learning) out of the corner of your ear. *Psychological Science*, 8, 101–105.

Shams, L., Kamitani, S., Shimojo, S. (2000). Illusions. What you see is what you hear. *Nature*, *408*, 788.

Shams, L., Kamitani, Y., & Shimojo, S. (2002). Visual illusion induced by sound. *Cognitive Brain Research*, *14*(1), 147-152.

Sinnett, S., Spence, C., & Soto-Faraco, S. (2007). Visual dominance and attention: Revisiting the Colavita effect. *Perception & Psychophysics*, 69, 673–686.

Song, S., Howard, J.H., & Howard, D.V. (2008). Perceptual sequence learning in a serial reaction time task. *Experimental Brain Research*, *189*, 145–158.

Spence, C., Parise, C., & Chen, Y. C. (2012). The Colavita visual dominance effect. In M.M. Murray, & M.T. Wallace (Eds.), *The Neural Bases of Multisensory Processes* (pp. 529-556). Boca Raton, FL: CRC Press.

Stein, B. E. & Meredeth, M. A. (1993). *The merging of the senses*. Cambridge, MA. MIT Press.

Welch, R. B., & Warren, D. H. (1980). Immediate perceptual response to intersensory discrepancy. *Psychological Bulletin*, *88*, 638-667.

Wickens, C. D. (1984). Processing resources in attention. In R. Parasuraman & R. Davies (Eds.), *Varieties of attention* (pp. 63–101). New York: Academic Press.

Unknitting the Meshwork: Interactivity, Serendipity and Individual Differences in a Word Production Task

Wendy Ross
(w.ross@kingston.ac.uk)
Department of Psychology, Kingston University
Kingston upon Thames, UNITED KINGDOM KT1 2EE

Frédéric Vallée-Tourangeau
(f.vallee-tourangeau@kingston.ac.uk)
Department of Psychology, Kingston University
Kingston upon Thames, UNITED KINGDOM KT1 2EE

Abstract

Creative ideas emerge from a meshwork of dynamic elements. Resources internal and external to the agent configure a cognitive ecosystem that scaffolds performance. In addition, capitalizing on fortuitous external cues may trigger new ideas. We examined these elements to determine how they come into play during a simple word production task. Participants were video recorded as they generated new words from 7 letter tiles in three different environments (i) high interactivity where the titles could be moved at will (ii) low interactivity where they could not, and (iii) low interactivity where the order of the tiles could be shuffled but once shuffled no additional actions were allowed. Overall, interactivity had a marginally positive impact on performance, while independent measures of participants' verbal fluency were strong predictors of performance in all environments. Based on a detailed coding of the video data, a finer-grained analysis of behaviour in the high interactivity condition revealed that the time participants spent manipulating the tiles was a significant predictor of performance. The video data also allowed us to measure the average latency to the production of a new word after shuffling the letters in the low interactivity condition as an index of how 'lucky' the reset was: Shorter average latencies were a significant predictor of overall word production. These data indicate that interactivity, serendipity, and internal cognitive resources determine problem-solving performance in this task.

Keywords: Creativity; interactivity; serendipity; cognitive eco system.

Introduction

Problem solving is an activity that takes shape in a dynamic meshwork of resources and processes, configured from internal mental resources, embodied actions and environmental affordances. To better appreciate the role of interactivity in problem solving it is important to contrast problem solving performance in task ecologies that differ in the degree to which a problem solver can 'think' through the manipulation of a physical model of the problem. In a low interactivity task environment, the problem solver is decoupled from her immediate environment: She is invited to solve a problem without using her hands to support thinking either through gesture or rearranging the physical elements that configure a model of the problem (such task environments are often the default procedure employed in problem solving research, Vallée-Tourangeau & March, 2019). In other words, problem solving proceeds from mental simulations of possible solutions. In contrast, a high interactivity task environment places no such constraints on her: Participants are presented with physical elements of the problem that can be manipulated to arrive at a solution. In such environments, proto solutions are boundary objects of sorts (Fiore & Wiltshire, 2016) that are physically constructed and perceived, unveiling action affordances and guiding attention in ways that are simply not possible in low interactivity conditions. Creative problem solving in a task ecology that favours interacting with the physical elements of a problem, is driven by three factors: the internal resources of the problem solver, her embodied behaviour and the environmental affordances that unfold dynamically as the physical model of the problem is modified. A full account of these aspects helps better appreciate their transactional nature.

Interactivity in the Word Production Task

The game of Scrabble has been modified to assess whether manipulating the letter array supports word production (see Maglio, Matlock, Raphaely, Chernicky & Kirsh, 1999; Webb & Vallée-Tourangeau, 2009; Vallée-Tourangeau & Wrightman, 2010; Kirsh, 2014; Fleming & Maglio, 2015). In this modified task participants are given 7 letters and invited to generate words. With an open problem of this sort, the dependent measure offers a more nuanced measure of the benefit or otherwise of interactivity. Additionally, letter set difficulty can be manipulated by selecting sets of letters that generate more or fewer words.

There are clear theoretical reasons to suppose that interactivity would benefit solvers in a word production task of this kind. By extending the mental workspace outside of the head, the internal letter representations are reified and are easily manipulated freeing up and scaffolding participants' internal resources (Gavurin, 1967; Webb & Vallée-Tourangeau, 2009; Vallée-Tourangeau & Wrightman, 2010). Furthermore, interactivity allows the solvers to move with less effort through the problem space and even to jump to new places with, at times, unplanned moves (Maglio et al., 1999). Thus, the tiles may either be recruited strategically or, more serendipitously, non-strategic moves may yield lucky combinations of letters.

Empirically, however, the data are less clear than may be imagined. The only study that demonstrates an unequivocal benefit is Flemming and Maglio (2015) where interactivity not only led to an increase in word production but also to rarer (less frequent) words being produced. While Maglio et al. (1999) documented a small overall benefit for interactivity, when this was broken down into the two different letter sets used, interactivity led participants to produce more words with one letter set but fewer words with another, easier, letter

set. With an easy letter set, participants are more capable of generating words without help so the added cost of manipulating tiles may actually slow down word generation. In addition, the serendipitous jumps proposed by Maglio et al. (1999) are less likely to occur when a participant can easily produce words.

Individual Differences

Where the participants have been profiled, the existing data show a clear interaction between the individual resources of the problem solver and the effect of interactivity. Vallée-Tourangeau and Wrightman (2010) found that there was a statistically significant benefit in the high interactivity condition for participants categorised in the low verbal fluency group while the benefit for those in the high verbal fluency group was negligible. This mixed story is echoed by Webb and Vallée-Tourangeau (2009) who manipulated the difficulty of level sets across two groups, children with and without developmental dyslexia. Here the number of words produced by each group depended on the difficulty of the letter set: Interactivity only benefitted the control group with the harder letter set and the children with developmental dyslexia only benefitted from interactivity with the easy set. The evidence to date suggests that interactivity scaffolds the performance of those who have lower verbal fluency or working memory and acts as an additional, reciprocal and non-linear processing loop (Vallée-Tourangeau & Vallée-Tourangeau, 2017) but it appears to confer little benefit when the task is within the capability of the participants either because of their skills or the nature of the letter set employed.

Environmental Affordances

For Maglio et al. (1999) the benefit of high interactivity was in no small part due to the introduction of randomness that seeds intelligent behaviour. Randomness is generated by the external environment and interactivity research explores the way problem solvers both recruit and are entangled with this environment (Ingold, 2017). Kirsh (2014) explicitly examined this role of randomness in the word production task. The participants were invited to take part in a task on a computer with an additional shuffle condition where one click shuffled the letters randomly. He found that the shuffle condition encouraged the production of a significantly higher number of words (M = 18.9) than both the static (M = 16.6) and the interactive (M = 17.7) conditions.

If we consider the constraints in place across Kirsh's three conditions, this becomes a more surprising result. As there were no reported constraints in the high interactivity condition, it theoretically provides the widest range of possible strategies. Participants are not prevented from shuffling the tiles randomly, just such shuffling would have to be self-generated. In practice, it seems unlikely that participants could have fully used the range of possibilities of the high interactivity version. Indeed, the number of shuffles described by Kirsh—the best performing third shuffled once every 3.7 seconds, the worst performing third once every 1.9 seconds—demonstrate the incredibly low cost of shuffling to generate hints in this task environment. In practice, it would be impossible to mimic this strategy with the high interactivity version in the same time.

Just as the skills of the problem solver are important when we consider the ways cognition arises from the interplay between person and external artefacts, so too are the affordances for action offered by the artefacts. Taking the cognitive ecology of this task seriously, requires taking the affordances of the external environment seriously. Rather than making the implicit assumption that the problem solver imposes her will on an inert and indifferent environment, we suggest that the nature of the artefacts selected will determine to some extent the actions undertaken (Steffensen, Vallée-Tourangeau & Vallée-Tourangeau, 2017).

It is our hypothesis that in Kirsh (2014) the low cost of shuffling the tiles with one click on a computer compared to the relatively high cost of moving tiles with a mouse, meant that shuffling functioned as an epistemic action (Kirsh & Maglio, 1994) more closely resembling the actions of a Tetris player who chooses a tretromino drop location based on what she sees *after* multiple physical rotations rather than a true test of luck. Indeed, Kirsh (2014, p. 19) acknowledges this: "the cost in time and mental effort must be sufficiently low that it pays to keep fishing for hints". The benefits of interactivity are only useful when they outweigh the costs of that interactivity (Maglio et al, 1999) and the shuffle condition reported in Kirsh (2014) is incredibly low in cognitive cost. Thus, while environmental randomness was examined it remains to be seen to what extent its benefit resulted from the low cost involved in monitoring the usefulness of a change in the letter array rather than having to take the time or make the mental effort to produce different arrays.

Participant Behaviour

The manipulation of chance by Kirsh (2014) also highlights differences in participants' behaviour. The number of shuffles varied across shufflers. Indeed, the better word generators shuffled "about 50% less" (p.18) than those who produced the fewest words. So, while the shuffle condition produced a higher overall mean of words, when the behaviour of the participants is taken into account, a more nuanced and accurate account of the role of chance is possible. Shuffling did not confer an indiscriminate benefit across all participants.

This difference in the behaviour of the participants is also acknowledged in a footnote in Maglio et al. (1999, footnote 2, p. 330): roughly a third of the participants did not consider it worth using their hands to structure their thoughts in an ostensibly high interactivity environment. This footnote requires us to consider to what extent the participants in this experimental condition could be said to be using interactivity; rather the condition might be more aptly renamed *potential* for high interactivity.

In various experiments investigating the role of interactivity in problem solving (e.g., Vallée-Tourangeau, Sirota & Vallée-Tourangeau, 2016), the low interactivity

condition is invariably tightly controlled, and participants' movements are constrained with them often being requested to lay their hands flat on the work surface. However, there are few controls and rarely any consideration of the manner in which participants recruit resources in a task environment labelled as high interactivity. Only Fleming and Maglio (2015) have taken a closer look at the behaviour of participants in a high interactivity condition. In contrast to Maglio et al. (1999), they suggest that all their participants moved the tiles. However, as the focus of their paper was strategy selection rather than the time spent interacting, the detailed analysis required restricted their coding to the behaviour of 8 participants in the final block with a specific aim of looking for and coding word production strategies. If we are to profile the whole system (Vallée-Tourangeau & Vallée-Tourangeau, 2017) then the level of interaction becomes important each time the participants encounter the tiles as a measure in itself rather than solely as an indication of strategy, especially if the different levels of interactivity designed in these environments do not result in differences in participants' behaviour.

It is further unclear how much participants' behaviour differs as a function of their individual differences. It is plausible that those who do not need the help of the tiles recruit them less. Research on expert Tetris players suggests an inverted U shape relationship of action and expertise with complementary actions decreasing as expertise increases (Destefano, Lindstedt & Gray, 2011). It is not unreasonable to expect a similar relationship in this task.

The Current Experiment

We examined the role of interactivity and chance in a word production task as well as the moderating properties of participants' verbal fluency. Rather than a computer and mouse we employed physical letter tiles. These artefacts more naturally invite interaction in the high interactivity condition and conversely increase the cost of movement in the shuffle condition testing if the benefits of shuffling hold when the nature of the artefacts are taken into consideration. We video recorded participants to undertake a more granular analysis of their behaviour in the high interactivity condition. This allowed us to assess the number of participants in the high interactivity condition who actually chose to move the letter tiles and determine the amount of time they actually interacted with the tiles. In this way, we can begin to disentangle some of the complexities that underlie the reported aggregated means in the high interactivity and the shuffle conditions.

We hypothesised that the increased time and cognitive cost of shuffling would lead to a reduction in the average number of shuffles and so, contrary to Kirsh's findings, we further hypothesised that the high interactivity condition would yield the most words followed by the shuffle condition reflecting the relative cognitive and time costs of each condition. Further, that video data would reveal a range of engagement with the tiles and capture the participants who do not avail themselves of the affordances for creativity in a high interactivity task environment. In line with the data reported for shuffling in Kirsh (2014), we expected an optimum level of interactivity beyond which there would be no further benefit. We hypothesised that verbal fluency would significantly moderate not only the total word count but participant behaviour, that is that high verbal fluency participants might not interact with the letter tiles to the same degree as low verbal fluency participants.

Method

Participants

Forty-two participants took part in the experiment in exchange for course credits. Two participants did not consent to be filmed and still received credit but as their behaviour could not be coded, their data were excluded. This left data for 40 participants (32 females, M_{age} = 25.65, SD = 7.17).

Design

The experiment used a repeated measures design with the order of the three experimental conditions counterbalanced across participants. The three conditions were high interactivity, low interactivity, and low interactivity + shuffle.

Materials and Measures

Three sets of 7 letters (COTFAED, NDRBEOE and TVAERWI) were created with similar average number of possible words of similar frequencies[1] piloted in a prior norming task. In each condition, the participants were given 5 minutes to call out as many words as they could from a set of 7 letter tiles (2cm * 2cm) initially presented in a straight line with the following constraints (i) the words must be at least three letters long and (ii) proper names and acronyms were not allowed. In the high interactivity condition, participants were invited to move the tiles as they saw fit. In the low interactivity condition, they were asked to not move or interact with the tiles in any way. Finally, in the shuffle condition, they were invited "whenever you want to and as many times as you want to" to collect all the tiles up, shake them in closed hands and lay the tiles out again in the new, randomly generated order. In the shuffle condition, when not shuffling the tiles, participants' movements were constrained in the same way as in the low interactivity condition. The dependent variable was the number of words generated by the participants in the three task environments.

Participants were also profiled along three further measures. First, performance on a modified version of the Thurstone (1938) test, which involves writing as many words with the letter S in five minutes and then as many words with the letter C in four minutes, was used to index participants'

[1] Frequencies taken from Zipf scores presented in the SUBTLEX-UK database (Van Heuven, Mandera & Keuleers (2014).

verbal fluency. Second, participants were invited to complete 12 5-letter anagrams taken from a larger set in Webb, Little and Cropper (2016) to assess their skills at anagramming. Finally, their levels of openness to experience was measured using the relevant items from the scale used in Lee and Ashton (2004); although there is little firm evidence on the role of personality traits and luck, this trait has been previously linked to a propensity to experience luck (McCay-Peet, Toms & Kelloway, 2015) and was added as an exploratory measure to assess if participants high on this trait would leverage the luck or otherwise of the shuffle condition.

Procedure

The anagram and verbal fluency tests were used as warm up tasks before the three main experimental conditions. The measure of extraversion was placed at the end of the study producing the following order:

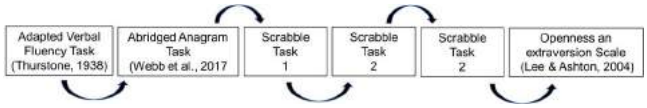


Figure 1: The order of the tasks.

The order of the conditions was counterbalanced across participants as was the set of letter tiles associated with each condition.

Qualitative Coding

In the high interactivity condition, the amount of time participants spent moving the tiles was coded using ELAN. The total time interacting with the tiles was assessed from when a participant touched a tile to when he or she stopped touching it. As there were many moments when a participant touched a tile but did not move it, this was further split into neutral moves (which did nothing to alter the array) and active moves (which changed the array in some way, either deliberate or random). Active moves were considered a reflection of interactivity. In the shuffle condition the number and timing of the shuffles was also recorded in ELAN. The timing of the shuffle was calculated from the moment participants touched the tiles until they had re-laid the array. In some instances, participants generated a word while relaying the tiles after the shuffle and therefore before the end of the full shuffle process; in these cases, the shuffle-new word latencies were negative.

Results

There was broadly similar performance in each experimental condition. Participants produced the highest number of words in the high interactivity condition ($M = 18.4$, $SD = 8.5$). There was virtually no difference between the performance in the low interactivity ($M = 17.0$, $SD = 6.2$) and shuffle ($M = 17.2$, $SD = 6.2$). While there was a slight benefit of interactivity, a one-way repeated measures analysis of variance revealed this to be non-significant, $F(2, 78) = 1.97$, $p = .146$, $\eta p^2 = .048$.

Correlations among measures of verbal fluency, anagram performance, openness and word production in the three experimental conditions are reported in Table 1 (df = 38). As expected, verbal fluency significantly correlated with performance in the high condition, $r = .717$, $p < .001$, the low condition, $r = .734$, $p < .001$ and the shuffle condition, $r = .745$, $p < .001$. Anagram skill also correlated highly with performance in the high, $r = .601$, $p < .001$, low, $r = .679$, $p < .001$ and shuffle conditions, $r = .630$, $p < .001$. There were no significant correlations between the measure of openness to experience and performance in any of the conditions.

Performance in the High Interactivity Condition

The video data enabled us to examine and analyse in finer detail the behaviour of the participants in the high interactivity condition. As we reviewed in the introduction, not all participants avail themselves of the opportunity to interact with the external environment. It is insufficient to analyse group level performance to determine the benefits of interactivity in the absence of a more detailed analysis of individual behaviour and performance. As expected, the video data revealed large differences in the behaviour of the participants in the high interactivity task environment. Active moves constituted 86% of the total time spent touching the tiles. Two participants opted not to interact with the tiles at all. The range of the time spent actually interacting with the letter tiles was 2.9 seconds to 226.9 seconds with a mean time of 106.4 seconds ($SD = 65.10$). The relationship between time interacting and the number of words produced is displayed in Figure 2. As the scatterplot reveals, the longer the participants interacted with the tiles the more words they produced. This relationship was significant, $r(38) = .329$, $p = .038$ and indeed becomes stronger when the effects of anagram skills and verbal fluency are partialled out, $r(38) = .439$, $p = .006$ offering a more direct measure of the impact of interactivity on word production. Contrary to our prediction and the observed shuffling behaviour in Kirsh (2014), interactivity conferred a steady benefit with no tailing off.

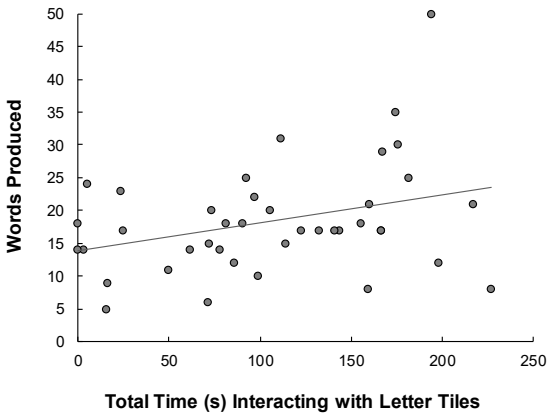


Figure 2: Number of words produced in the high interactivity condition as a function of the time (in seconds) spent interacting with the letter tiles.

Table 1: Descriptive statistics for and correlations among measures of verbal fluency, anagram performance, openness, and word production performance in the three experimental conditions.

	M	*SD*	1	2	3	4	5	6
1. Fluency score	88.30	25.15	-					
2. Anagram total	8.90	2.73	.493 **	-				
3. Openness	102.75	15.25	-.018	-.046	-			
4. High word count	18.35	8.49	.717 **	.601 **	-.230	-		
5. Low word count	17.03	6.59	.734 **	.679 **	-.173	.819 **	-	
6. Shuffle word count	17.23	6.23	.745 **	.630 **	-.019	.848 **	.796 **	-

** $p < .001$ level (2-tailed).

Finally, in contrast to our prediction, the extent to which a participant recruited the letters to aid thinking did not significantly correlate with either verbal fluency, $r(38) = .111$, $p = .481$, or anagram skills, $r(38) = -.032$, $p = .844$.

Performance in the Shuffle Condition

Participants shuffled an average of 1.58 times; there was, however, a wide variation in the number of shuffles. Twenty five percent of the participants opted not to shuffle at all, 17.5% of participants opted to shuffle once, 30% twice and a further 27.5% opted to shuffle 3 times.

As predicted, there was a large time cost to shuffling. The average shuffle took 17.51 seconds ($SD = 3.51$) with the fastest shuffle being 10.22 seconds and the slowest taking 24.64 seconds. Overall, shuffling appeared to be an unhelpful strategy. Participants' word production performance did not differ among those who did not shuffle ($M = 17.81$, $SD = 6.58$), shuffled once ($M = 17.12$, $SD = 5.02$), twice ($M = 16.58$, $SD = 7.35$) or three times ($M = 16.81$, $SD = 6.30$), $F < 1$.

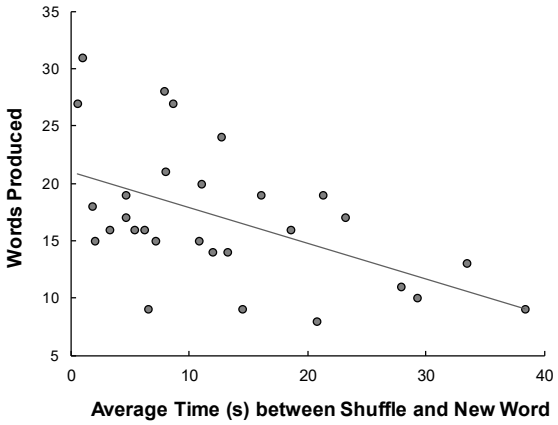


Figure 3: Number of words produced in the shuffle condition as a function of the average time (in seconds) before a word is produced after shuffling the letter tiles.

The effect of the shuffles also varied widely. The average time after the end of shuffling to generate a word was 11.69 seconds ($SD = 13.77$). The minimum time after shuffling to produce a word was -5.05 seconds (producing a word while relaying the tiles after the shuffle) whereas the maximum time after the shuffling had ended to producing another word was over a minute (61.72 seconds). Only one participant did not produce any words after her first shuffle and went on to shuffle again.

It seems likely that a word produced directly after a shuffle has been stimulated by that shuffle whether that shuffle directly yielded the word or whether the act of shuffling and laying out of the tiles stimulates further thought. We thus measured how long after a shuffle a word was produced as a proxy measure of the luckiness of the shuffle – the faster a word was produced the luckier the shuffle. The relationship between this time (averaged out for those participants who had more than one shuffle) and the total number of words produced overall in the shuffle condition is illustrated in Figure 3. The correlation was significant, $r(27) = -.520$, $p = .004$ even when controlling for verbal fluency and anagram skill, $r(27) = .422$, $p = .028$, suggesting that the nature of the array produced by the shuffle and the words it stimulated was important to the overall number of words produced in that condition.

Discussion

This experiment was designed to trace the influence of interactivity, serendipity and verbal fluency in a word production task. These three elements create a dynamic meshwork from which word production skills are enacted. The differences in the mean number of words produced in the three experimental conditions were marginal albeit showing a general trend in line with past findings favouring high interactivity. When performance is viewed through the lens of a condition's mean score with individual variation in behaviour and cognitive skills flattened, the benefits of interactivity in this task are not clearly revealed. While there has been some examination of cognitive profiles which benefit from interactivity, the implicit assumption in previous research has been that there has been no difference in participant behaviour in the high interactivity condition.

However, by subjecting participant behaviour to a finer granularity of analysis, we can start to disentangle how that behaviour affects the numbers of words produced and isolate a purer effect of interactivity. Given that two participants did not interact with the letter tiles at all, it is illogical to attribute

their scores in both the low and high interactivity conditions to different factors (in effect, despite our best efforts to change the task ecology, these participants approached the high and low interactivity conditions in the same manner). Again, those participants who chose not to shuffle essentially participated in an additional low interactivity experimental condition. It is meaningless to assign one score to low and one to shuffle unless we are measuring the effect of experimental instructions.

When the behaviour of the participants is taken into account, there was a significant correlation between the time spent interacting with the letter tiles and the number of words produced in that condition even when controlling for verbal fluency skills. This suggests that interactivity boosts word production only when a participant fully engages in that condition. Measuring participants' behaviour is important and designing a high interactivity task environment does not guarantee that the affordances inherent to a dynamic problem-solving environment will be perceived and exploited to boost performance. Contrary to expectation, there was no relationship between verbal fluency and the time spent interacting. This is in contrast to prior observations on the different effect of interactivity on different individual difference profiles: interacting with the tiles helped everyone.

Further, there was failure to replicate Kirsh's (2014) observation that engineered randomness boosts performance. However, the nature of the current experiment increased the impact of an unlucky shuffle by increasing the time and cognitive cost of shuffling as the materials were moved from a digital to a material environment. This led to a predicted decrease in the number of participants who opted to shuffle and the number of times they shuffled along with a much higher investment in the array produced by the shuffle than that reported by Kirsh. The inherent contingent and transactional nature of luck in this task was partly captured by the latency to first word produced after a random re-arrangement. These average latencies were significant predictors of how many total words would be produced in this otherwise low interactivity environment. It would be interesting to couple the luck and high interactivity manipulation in future research.

The current results suggest that previous research into interactivity may have underestimated its benefit by failing to subject behaviour to a sufficiently granular analysis which can only be done with detailed video coding of behaviour (see also Steffensen, Vallée-Tourangeau, & Vallée-Tourangeau, 2016). A problem solver's trajectory is unique and the interaction with a richer set of environmental resources will trigger more complex behaviours. Thus, it behoves us to take a closer look at what is actually happening in a task environment that fosters interactivity. Interactivity is contingent and messy: its study must take into account the behaviour of the participant and the nature of the materials being used to more accurately capture the factors that drive creative problem solving.

References

ELAN (Version 5.2) [Computer software]. (2018, April 04). Nijmegen: Max Planck Institute for Psycholinguistics. Retrieved from https://tla.mpi.nl/tools/tla-tools/elan/

Destefano, M., Lindstedt, J. K., & Gray, W. D. (2011). Use of complementary actions decreases with expertise. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 2709–2014). Austin, TX: Cognitive Science Society.

Fleming, M., & Maglio, P. P. (2015). How physical interaction helps performance in a Scrabble-like task. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.) *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 716-721). Austin, TX: Cognitive Science Society.

Gavurin, E.I. (1967) Anagram solving and spatial aptitude. *The Journal of Psychology*, *65*, 65-68.

Fiore, S. M., & Wiltshire, T. J. (2016) Technology as teammate: Examining the role of external cognition in support of team cognitive processes. *Frontiers in Psychology, 7*, 1531.

Ingold, T. (2017). On human correspondence. *Journal of the Royal Anthropological Institute*, *23*, 9-27

Kirsh, D. (2014). The importance of chance and interactivity in creativity. *Pragmatics & Cognition*, *22*, 5-26

Lee, K., & Ashton, M. C. (2016). Psychometric properties of the HEXACO-100. *Assessment*, *25*, 543-556

Maglio, P. P., Matlock, T., Raphaely, D., Chernicky, B., & Kirsh, D. (1999). Interactive skills in Scrabble. In M. Hahn, & S. C. Stoness (Eds.), *Proceedings of the 21st Conference of the Cognitive Science Society* (pp. 326-330). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

McCay-Peet, L. Toms, E.G. & Kelloway, E.K. (2015). Examination of relationships among serendipity, the environment, and individual differences. *Information Processing and Management*, *51*, 391-412.

Steffensen, S. V., & Vallée-Tourangeau, F. (2018). An ecological perspective on insight problem solving. In F. Vallée-Tourangeau (Ed.), *Insight: On the origins of new ideas* (pp. 169-190). London: Routledge.

Steffensen, S. V., Vallée-Tourangeau, F., & Vallée-Tourangeau, G. (2016). Cognitive events in a problem-solving task: Qualitative methods for investigating interactivity in the 17 animals problem. *Journal of Cognitive Psychology*, *28*, 79-105.

Thurstone, L.L. (1938). *Primary mental abilities*. Chicago: University of Chicago Press.

Vallée-Tourangeau, F., & March, P. L. (2019). Insight out: Making creativity visible. *Journal of Creative Behavior.*

Vallée-Tourangeau, F., Sirota, M., & Vallée-Tourangeau, G. (2016). Interactivity mitigates the impact of working memory depletion on mental arithmetic performance. *Cognitive Research: Principles and Implications*, *1*, 26.

Vallée-Tourangeau, G. & Vallée-Tourangeau, F. (2017). Cognition beyond the classical information processing

model: cognitive interactivity and the systemic thinking model (SysTM). In Cowley, S. J., & Vallée-Tourangeau, F. (Eds). *Cognition Beyond the Brain*. London: Springer.

Vallée-Tourangeau, F., & Wrightman, M. (2010). Interactive skills and individual differences in a word production task. *AI & Society*, *25*, 433-439.

Webb, M. E., Little, D. R., & Cropper, S. J. (2018). Once more with feeling: Normative data for the aha experience in insight and noninsight problems. *Behavior research methods*, *50*, 2035-2056.

Webb, S., & Vallée-Tourangeau (2009). Interactive word production in dyslexic children. In N. Taatgen, H. van Rijn, J. Nerbonne & L. Schomaker (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society*. (pp. 1436–1441). Austin, TX: Cognitive Science Society

Van Heuven, W., Mandera, P., Keuleers, E., Brysbaert, M. (2014). Subtlex-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology*, *67*, 1176-1190.

Modelling semantics by integrating linguistic, visual and affective information

Armand S. Rotaru (armand.rotaru.14@ucl.ac.uk) **Gabriella Vigliocco (g.vigliocco@ucl.ac.uk)**
Faculty of Brain Sciences, University College London,
WC1H 0DS, London, United Kingdom

Abstract

A number of recent models of semantics combine linguistic information, derived from text corpora, and visual information, derived from image collections, demonstrating that the resulting multimodal models are better than either of their unimodal counterparts, in accounting for behavioural data. However, first, while linguistic models have been extensively tested for their fit to behavioural semantic ratings, this is not the case for visual models which are also far more limited in their coverage. More broadly, empirical work on semantic processing has shown that emotion also plays an important role especially for abstract concepts, however, models integrating emotion along with linguistic and visual information are lacking. Here, we first improve on visual representations by choosing a visual model that best fit semantic data and extending its coverage. Crucially then, we assess whether adding affective representations (obtained from a neural network model designed to predict emojis from co-occurring text) improves model's ability to fit semantic similarity/relatedness judgements from a purely linguistic and linguistic-visual model. We find that adding both visual and affective representations improve performance, with visual representations providing an improvement especially for more concrete words and affective representations improving especially fit for more abstract words.

Keywords: language; vision; emotion; distributional models; multimodal models; similarity/relatedness; concreteness.

Introduction

Despite the success of distributional, linguistic models in accounting for behavioural effects in a variety of semantic tasks, all these models suffer from the *symbol grounding problem* (Harnad, 1990). As a solution to this problem, embodied theories of semantics (e.g., Glenberg, Graesser, & de Vega, 2008) have argued that the sensory-motor representations generated by our experiences with the world play an important role in determining word meaning. Recent computational models of semantics reconcile distributional and embodied theories, by combining linguistic and perceptual (i.e., visual) representations. The fact that language and vision provide complementary sources of information is best illustrated by the finding that multimodal, linguistic-visual models outperform both purely linguistic and purely visual models, in a wide range of tasks (see Bruni, Tran, & Baroni, 2011; 2014; Kiela, Verő, & Clark, 2016).

However, empirical work has shown that semantic representations are not only grounded in sensory-motor experience but also in emotion. A vast literature supports the finding that emotion plays a significant and pervasive role in human cognition (for a review, see Dolan, 2002). Emotion is an important factor in memory (Blaney, 1986; Eich, Macaulay, & Ryan, 1994), and in processing words (e.g., Kousta, Vinson, & Vigliocco, 2009). Kousta et al. (2011) found that a much larger number of abstract than concrete concepts are valenced (have positive or negative emotional associations) and by virtue of being valenced, they are processed faster than neutral matched words. Vigliocco et al. (2014) further showed that because of their greater affective associations, abstract words processing engages the limbic emotional system and Ponari, Norbury, and Vigliocco (2018) showed that emotionally valenced words are learnt earlier and better recognized by children up to 9 years of age. Within a general embodiment framework, the hypothesis is that semantic representations do not only embed sensorimotor properties but also emotional properties. Emotional properties may be especially important for abstract concepts (e.g., *religion*, *society, idea*), however, emotional associations are not limited to abstract words and therefore, we argue, they play a general role in semantic representation.

While many models have integrated linguistic and visual information, only one previous study has considered emotional information along with visual and linguistic information (De Deyne, Navarro, Collell, & Perfors, 2018). De Deyne et al. examined the change in performance for distributional models of semantics, when adding visual and emotional information. They tested the assumption that external language models (i.e., distributional models, trained on word corpora) are relatively poor at representing visual and affective information, in comparison to internal language models (i.e., models based on free association norms). They found that adding visual and emotional information led to little or no improvement for internal language models, but a moderate positive effect for external language models. Here, we develop a quite different multimodal model of semantics that incorporates linguistic, visual and emotional information from corpora of text, images and emoticons, and test the multimodal model against existing datasets of ratings of

semantic similarity/relatedness of words. We use a state-of-the-art emotion model (DeepMoji) and we improve the coverage of the visual model we use. While state-of-the art distributional language models (Pereira et al. 2016) have large coverage of words and have been widely tested for their ability to fit human semantic similarity/relatedness data, this is not the case for visual models. Thus, before being able to develop models that embed linguistic, visual and emotional information, we extend the coverage of existing visual models and carry out their evaluation in order to decide which one to use for our multimodal models. We expect that the integrated model will outperform a purely linguistic, as well as models that combine linguistic-visual and linguistic-emotional information. In addition, we expect that adding visual or emotional representations will especially be beneficial for more concrete concepts whereas emotional information will especially be beneficial for more abstract concepts, in line with the empirical evidence reviewed above (and with initial findings from De Deyne et al, 2018).

Methods

Datasets of behavioural data

We use four datasets of similarity/relatedness ratings to carry out evaluation of the models. The datasets are: SimLex999 (999 pairs of nouns, verbs, and adjectives; Hill, Reichart, & Korhonen, 2015), SimVerb3500 (3500 pairs of verbs; Gerz et al., 2016), MEN (3000 pairs of nouns, verbs, and adjectives; Bruni, Tran, & Baroni, 2014), and SL (7576 pairs of nouns; Silberer & Lapata, 2014). We chose these norms mainly because they are some of the largest datasets currently available, but also because the word pairs they contain cover are very diverse in terms of concreteness and valence, as well as parts of speech. In terms of word pair concreteness, SimLex999 (M = 3.62, SD = 1.07) and SimVerb3500 (M = 3.1, SD = 0.7) cover a broad range of values, whereas MEN (M = 4.4, SD = 0.49) and SL (M = 4.83, SD = 0.14) consist predominantly of concrete words.

Model choice

Language Model. Our language model of choice is GloVe (Pennington, Socher, & Manning, 2014), trained on a corpus of 6 billion words, using 300-dimensional representations. GloVe has been shown to have a performance better than, or equal to, several other state-of-the-art distributional models (Pereira, Gershman, Ritter, & Botvinick, 2016), which makes it one of the best linguistic models available.

Emotion Model. The emotion model that we use is DeepMoji (Felbo et al., 2017), trained on 1.2 billion tweets. This model has been shown to obtain state-of-the-art performance in tasks involving emotion and sentiment analysis, as well as sarcasm detection. DeepMoji is similar to a number of recent approaches, which employ emotional expressions co-occurring with text fragments, such as positive/negative emoticons (Deriu et al., 2016), hashtags (e.g., #anger, #joy; Mohammad, 2012), or mood tags (Mishne, 2005). This model is very different from the one by De Deyne et al. (2018), which was constructed by concatenating valence, arousal, and potency ratings, for men and women separately (i.e., 6 dimensions), from the study by Warriner, Kuperman, and Brysbaert (2013), with valence, arousal, and dominance ratings, from the study by Mohammad (2018). DeepMoji provides better represetations for our purposes than ratings because firstly, a model trained over a corpus of tweets, rather than subjective ratings, makes the emotion model more comparable to the linguistic and visual models, both trained over corpora. Secondly, DeepMoji covers 50,000 words, whereas the combined affective norms cover less than 14,000 words. Finally, the model operates with 256-dimensional vector representations, and is trained to predict the occurrence of 64 types of emoticons, and thus it is able to represent complex patterns of word similarity, driven by richer emotional information than that captured by subjective norms.

Visual Model. To select the best model, we compared five models, based on their performance in predicting subjective similarity/relatedness ratings. The first model (K&B) is the convolutional model employed by Kiela and Bottou (2014; 6144 dimensions), trained on the ESP Game dataset (Von Ahn & Dabbish, 2004), using the mean of the feature vectors per each word. The second, third, and fourth models are AlexNet (Krizhevsky, Sutskever, & Hinton, 2012; 4096 dimensions), GoogLeNet (Szegedy et al., 2015; 1024 dimensions), and VGG-19 (Simonyan & Zisserman, 2014; 4096 dimensions), trained on images obtained from Google Image Search, following the approach used by Kiela, Verő, and Clark (2016). The fifth model uses SIFT descriptors (Lowe, 2004), computed over the NUS-WIDE dataset (Chua et al., 2009; 500 dimensions). The models were tested on similarity/relatedness ratings for 7611 word pairs, covered by all models and obtained by merging the four sets of ratings. Before merging, the scores in each set were linearly rescaled to fall in the interval [0,1], to make them comparable across datasets. The performance of the models was evaluated using the Spearman correlation between the cosine similarity of the model representations, and the similarity/relatedness ratings from the norms. The results are shown in Fig. 1.

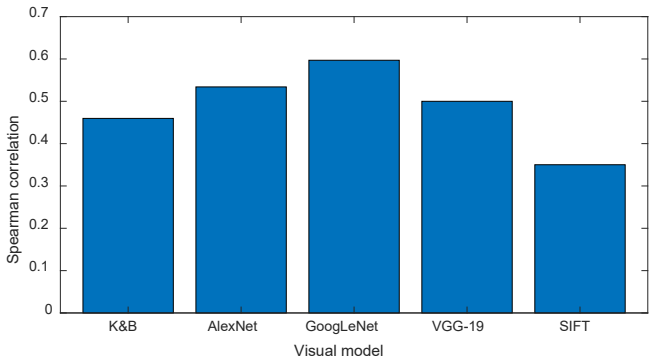


Figure 1. Spearman correlations between model cosine similarities and subjective similarity/relatedness ratings.

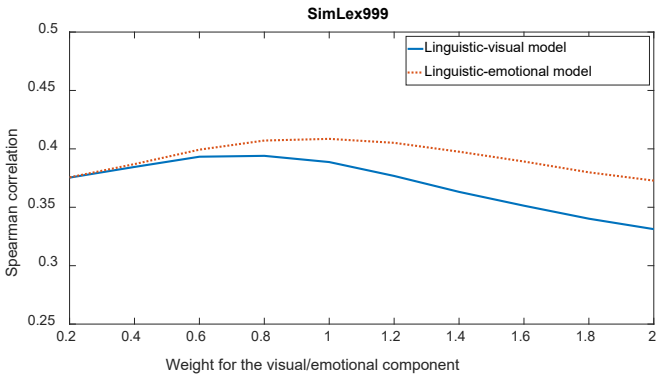
All the correlations are significant[1] ($p < .001$), suggesting that model-based similarities are reliable predictors of subjective similarity/relatedness ratings. Since we want to find the best model, we apply the Fisher Z-Transformation and then run two-tailed Z-tests for all the 10 possible pairings of models. All the differences are significant ($p < .004$), and they reveal that GoogLeNet has the highest performance, followed by Alexnet, VGG-19, K&B, and SIFT. Thus, we use GoogLeNet.

Results

We tested whether linguistic-visual and linguistic-emotional models are indeed better than a purely linguistic one, as well as whether it is the case that linguistic-visual-emotional models are better than linguistic-visual, linguistic-emotional and purely linguistic ones. We also examined whether the models behave differently for concrete and abstract word pairs.

Linguistic-visual and linguistic-emotional models vs purely linguistic model.

To evaluate the change in goodness of fit associated with adding a visual component to the purely linguistic model, we began by normalizing the linguistic and the visual representations to unit length. Next, we concatenated the linguistic representations with the visual ones, assigning a weight of 1 to the linguistic components, and weights from 0.2 to 2, in steps of 0.2, to the visual components. Both here and in our further analyses, we tested various weights, since it was not clear which weight would produce optimal results. Finally, for each of the four similarity/relatedness datasets, we compared the 10 resulting linguistic-visual models with the purely linguistic model, by normalizing the correlations and using two-tailed Z-tests. The same type of analyses were run for the linguistic-emotional models. Results are in Fig. 2.



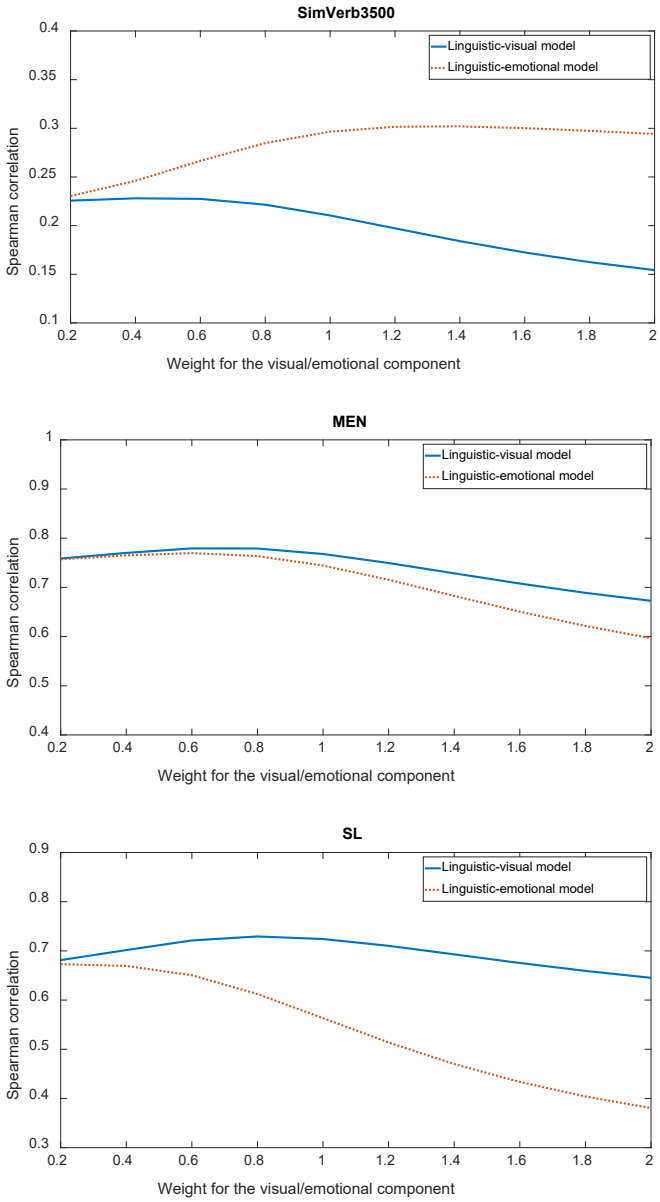


Figure 2. Model performance for the linguistic-visual and linguistic-emotional models. The weights assigned to the visual/emotional component vary from 0.2 to 2, in steps of 0.2

The tests indicate that adding visual information has a significant positive effect only for the SL dataset ($p < .001$), for weights ranging from 0.6 to 1.2, and a significant negative effect for the MEN dataset ($p < .001$), for weights between 1.6 and 2. These results seem to be at odds with previous studies showing that linguistic-visual models always perform slightly better than purely linguistic ones. However, firstly, in almost all the other studies, the authors either weigh the linguistic and visual representations equally, by default (e.g., Kiela, Hill, Korhonen, & Clark, 2014; Silberer, Ferrari, &

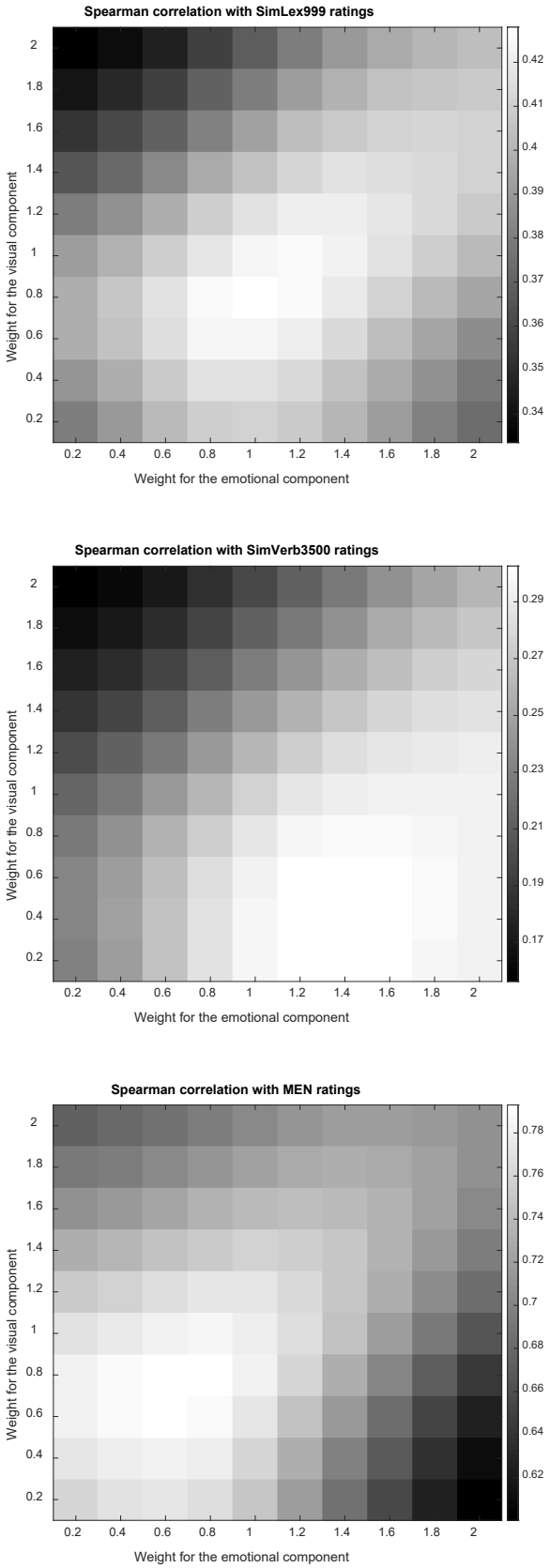
[1] The Bonferroni correction was applied when assessing the statistical significance of all the results presented in this study.

Lapata, 2013), or they only employ the weight that gives the best results for the integration (e.g., Bruni, Tran, & Baroni, 2014; Bruni, Uijlings, et al., 2012), which leaves room for null or detrimental results of linguistic-visual integration, when employing sub-optimal weights. Secondly, we use a linguistic model that is trained over a corpus of 6 billion words, whereas other studies (e.g., Hill & Korhonen, 2014; Kiela & Bottou, 2014; Silberer & Lapata, 2012) typically employ considerably smaller corpora (i.e., containing between 80 and 800 million words). Since smaller corpora lead to a poorer performance of the linguistic model, this leaves more room for a beneficial effect of adding visual information in the other studies, as compared to our study.

Adding emotional information is significantly beneficial only for the SimVerb3500 dataset ($p < .00125$), for weights ranging from 1.2 to 1.6, while it is significantly detrimental for the MEN dataset ($p < .001$), for weights between 1.4 and 2, and for the SL dataset ($p < .001$), for weights between 0.6 and 2. The SimVerb3500 dataset is different from all the others in that it is the only one including only verbs (which are not highly represented in any other dataset). As verbs (words referring to events) are considered to be more abstract, this finding is in line with the view that emotional information is especially important for abstract words (Kousta et al., 2011).

Linguistic-visual-emotional model vs linguistic-visual, linguistic-emotional, and purely linguistic models.

In order to compare the trimodal model with the bimodal and unimodal ones, we again start by normalizing the linguistic, visual, and emotional representations, to unit length. We then construct trimodal models by assigning a weight of 1 to the linguistic components, and weights from 0.2 to 2, in steps of 0.2, to the visual and emotional components, in all pairwise combinations for the last two components. Next, for each dataset, we select the best five and worst five trimodal models, in terms of performance, and compare them to their corresponding linguistic-visual models (i.e., obtained by removing the emotional component), linguistic-emotional models (i.e., obtained by removing the visual component), and purely linguistic model (i.e., obtained by removing both the visual and emotional components). The results are shown in Fig. 3.



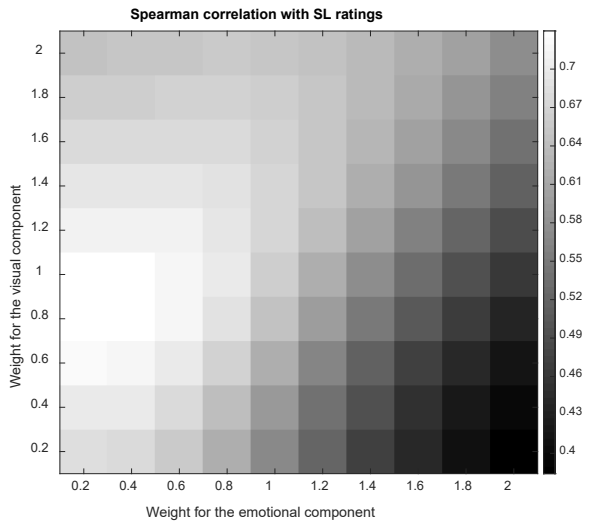


Figure 3. Model performance for the linguistic-visual-emotional model. The weights assigned to the visual/emotional component vary from 0.2 to 2, in steps of 0.2

When comparing the performance of the trimodal models to that of their corresponding linguistic-visual models, the addition of an emotional component has a significant positive effect for the best models on the SimVerb3500 dataset ($p <$ .0016), and a significant negative effect for the worst models on the MEN and SL datasets ($p <$ .001). These results are very similar to those found when comparing the linguistic-emotional models to the purely linguistic one, and might be explained by the fact that verbs, such as those that make up the SimVerb3500 norms, are relatively abstract. In contrast, for concrete nouns, which form the majority of pairs from the MEN and SL norms, emotion should not have a positive effect (the finding of a detrimental effect is unexpected but potentially interesting as may indicate that adding affective information may reduce the separation between different types of words).

The comparison between the trimodal models and their corresponding linguistic-emotional models reveals that including a visual component is significantly beneficial for the best models on the SL dataset ($p <$ .001), but significantly detrimental for two of the worst models on the SimVerb3500 datasets ($p <$ .001). Again, SL consists only of concrete nouns, for which visual information is very salient, while SimVerb3500 consists only of verbs, the semantics of which is likely not to be properly captured in a few tens of images per word, due to its complexity.

Finally, contrasting the trimodal models with the purely linguistic one, we find that bringing in both visual and emotional information significantly increases performance for the best models on the SimVerb3500, MEN, and SL datasets ($p <$ .0016), while it significantly decreases performance for the worst models on the MEN and SL datasets ($p <$ .001). These results are a combination of the partial results regarding the effects of appending visual and emotional components to the purely linguistic and bimodal models, which indicates little overlap between vision and emotional representation.

Comparing the models for concrete and abstract words

In order to test whether visual content is more important for more concrete words, while emotional content for more abstract words, we first combined the SimLex999 and SimVerb3500 datasets, as they cover a broader range of concreteness ratings than MEN and SL. Then, we divided the merged dataset into a low and a high concreteness subset. More specifically, we selected the bottom 25% and the top 25% of pairs, based on the mean concreteness of each word pair covered by the concreteness norms of Brysbaert, Warriner, and Kuperman (2014). We then tested the performance of the emotional and visual models, the two bimodal models, and the trimodal models, setting all the weights set to 1. The results are displayed in Fig. 4.

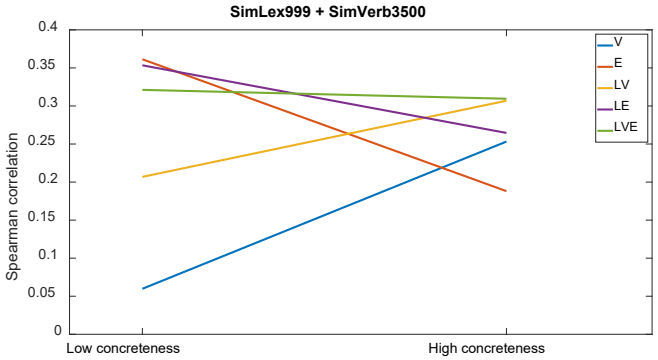


Figure 3. Model performance for low and high concreteness word pairs.

Using one-tailed *Z*-tests, after normalizing the correlations, we found that the performance of the visual model is higher for more concrete pairs, in comparison to the less concrete ones, for the visual ($p <$ .001) and linguistic-visual ($p <$ .01) models. Also, the emotional model has a better performance for the more abstract pairs, as opposed to the less abstract ones. Non-significant results were obtained for the linguistic-emotional and trimodal models. These results seem to suggest that the positive effect of adding visual information should be greatest for datasets consisting mainly of more concrete words, such as MEN and SL, while the beneficial effect of including emotional information should be largest for datasets made up mainly of more abstract words, such as SimLex999 and SimVerb3500.

Discussion

A first goal of this paper was to present an evaluation of visual models in order to identify the model(s) better fitting behavioural semantic data. We found that convolutional neural networks models (i.e., K&B, Alexnet, GoogLeNet, VGG-19) have a better performance than a classical, bag-of-visual-words model (i.e., SIFT), when tested over a large dataset of similarity/relatedness ratings. Among the

convolutional models, GoogLeNet gave the best results, followed by AlexNet, VGG-19, and K&B.

The second, and main goal was to develop models that integrate linguistic, visual and emotional information and to assess their performance against purely linguistic models and models that only include either visual or emotional features. We chose the DeepMoji model for a number of reasons, namely: its state-of-the-art performance in a number of emotional tasks; its distributional nature, since it predicts the occurrence of an emoticon based on its immediate linguistic context; its capacity to use rich emotional information, as it is trained on tweets containing 64 types of emoticons; its high dimensionality, which allows it to encode complex patterns of emotion-based word similarity.

In order to better understand the relative importance of each visual and emotional component, we carried out comparisons in which we parametrically varied the weight of visual and/or emotional information. In this manner, we can see when adding this information leads to better or worse performance. In general, we found that including non-linguistic information has a positive impact. However, first, this impact is modulated by whether the dataset includes predominantly concrete or abstract words. As expected on the basis of previous literature (e.g., Kousta et al., 2011) we see that including visual information is particularly beneficial to more concrete concepts whereas including emotional information is particularly beneficial to more abstract concepts. This is clearly visible when we assess model performance separately for more concrete and abstract words (see Fig 4). It is also clear from the comparison between MEN (only concrete words) and SimVerb3500 (only verbs, hence more abstract): across comparisons, we see that indeed visual information brings more benefit to the former, whereas emotional information brings more benefit to the latter.

Second, the effect is modulated by the weights attributed to the different types of information. While the theoretical interpretation of the differences we found related to weights is not immediate, this finding may have practical value for future modelling.

As mentioned in the introduction, a previous study (De Deyne et al., 2018) also examined the change in performance for distributional models of semantics, when adding experiential (i.e., visual and emotional) information. They found that adding experiential information led to little or no improvement for internal language models, but had a moderate positive effect for external language models. Moreover, they also found that adding visual information had the greatest effect for concrete words, while introducing affective information had the largest impact for abstract words. This finding mirrors our own, when comparing the linguistic-visual and linguistic-emotional models to the purely linguistic model.

However, there are a number of key differences between their approach and ours. Firstly, we avoided the potentially controversial distinction between external and internal language models, focusing on an objective corpus-based approach. Secondly, in a similar vein, we decided to use an emotional model that learns affective information indirectly, by predicting the co-occurrence of emojis and text in a corpus, rather than using emotional representations derived directly from valence, arousal and dominance norms (Mohammad, 2018; Warriner, Kuperman, & Brysbaert, 2013). This also increases the coverage of our model. Finally, since the resulting representations in our model are high-dimensional, they might provide more fine-grained information than representations with only three dimensions.

References

Blaney, P. H. (1986). Affect and memory: A review. *Psychological Bulletin*, *99*(2), 229-246.

Bruni, E., Tran, G. B., & Baroni, M. (2011). Distributional semantics from text and images. In *Proceedings of the GEMS 2011 Workshop on Geometrical Models of Natural Language Semantics* (pp. 22-32).

Bruni, E., Tran, N. K., & Baroni, M. (2014). Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, *49*, 1-47.

Bruni, E., Uijlings, J., Baroni, M., & Sebe, N. (2012). Distributional semantics with eyes: Using image analysis to improve computational representations of word meaning. In *Proceedings of the 20th ACM International Conference on Multimedia* (pp. 1219-1228).

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904–911.

Buechel, S., & Hahn, U. (2016, August). Emotion analysis as a regression problem—dimensional models and their implications on emotion representation and metrical evaluation. In *Proceedings of the 22nd European Conference on Artificial Intelligence* (pp. 1114-1122).

Chua, T. S., Tang, J., Hong, R., Li, H., Luo, Z., & Zheng, Y. (2009). NUS-WIDE: a real-world web image database from National University of Singapore. In *Proceedings of the ACM International Conference on Image and Video Retrieval*.

De Deyne, S., Navarro, D., Collell, G., & Perfors, A. (2018, November 28). Visual and Affective Grounding in Language and Mind. https://doi.org/10.31234/osf.io/q97f8

Deriu, J., Gonzenbach, M., Uzdilli, F., Lucchi, A., Luca, V. D., & Jaggi, M. (2016). Swisscheese at semeval-2016 task 4: Sentiment classification using an ensemble of convolutional neural networks with distant supervision. In *Proceedings of the 10th International Workshop on Semantic Evaluation* (pp. 1124-1128).

Dolan, R. J. (2002). Emotion, cognition, and behavior. *Science*, *298*(5596), 1191-1194.

Eich, E., Macaulay, D., & Ryan, L. (1994). Mood dependent memory for events of the personal past. *Journal of Experimental Psychology: General*, *123*(2), 201-215.

Felbo, B., Mislove, A., Søgaard, A., Rahwan, I., & Lehmann, S. (2017). Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion

and sarcasm. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 1616–1626).

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., & Ruppin, E. (2001). Placing search in context: The concept revisited. In *Proceedings of the 10th International World Wide Web Conference* (pp. 406-414).

Gerz, D., Vulić, I., Hill, F., Reichart, R., & Korhonen, A. (2016). SimVerb-3500: A Large-Scale Evaluation Set of Verb Similarity. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 2173–2182).

Glenberg, A. M., Graesser, A. C., & de Vega, M. (Eds.). (2008). *Symbols and Embodiment: Debates on Meaning and Cognition*. Oxford University Press.

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, *42*(1-3), 335-346.

Hill, F., Reichart, R., & Korhonen, A. (2015). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, *41*(4), 665-695.

Hill, F., & Korhonen, A. (2014). Learning abstract concept embeddings from multi-modal data: Since you probably can't see what I mean. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (pp. 255-265).

Kiela, D., & Bottou, L. (2014). Learning image embeddings using convolutional neural networks for improved multi-modal semantics. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (pp. 36-45).

Kiela, D., Hill, F., Korhonen, A., & Clark, S. (2014). Improving multi-modal representations using image dispersion: Why less is sometimes more. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (pp. 835-841).

Kiela, D., Vero, A. L., & Clark, S. C. (2016). Comparing data sources and architectures for deep visual representation learning in semantics. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 447–456).

Kousta, S. T., Vinson, D. P., & Vigliocco, G. (2009). Emotion words, regardless of polarity, have a processing advantage over neutral words. *Cognition*, *112*(3), 473-481.

Kousta, S. T., Vigliocco, G., Vinson, D. P., Andrews, M., & Del Campo, E. (2011). The representation of abstract words: why emotion matters. *Journal of Experimental Psychology: General*, *140*(1), 14-34.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems* (pp. 1097-1105).

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, *60*(2), 91-110.

Mishne, G. (2005). Experiments with mood classification in blog posts. In *Proceedings of the ACM SIGIR 2005 Workshop on Stylistic Analysis of Text for Information Access* (pp. 321-327).

Mohammad, S. M. (2012, June). #Emotional tweets. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics* (pp. 246-255).

Mohammad, S. (2018, July). Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 174-184).

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Nethods in Natural Language Processing* (pp. 1532-1543).

Pereira, F., Gershman, S., Ritter, S., & Botvinick, M. (2016). A comparative evaluation of off-the-shelf distributed semantic representations for modelling behavioural data. *Cognitive Neuropsychology*, *33*(3-4), 175-190.

Ponari, M., Norbury, C. F., & Vigliocco, G. (2018). Acquisition of abstract concepts is influenced by emotional valence. *Developmental Science*, *21*(2), e12549.

Silberer, C., Ferrari, V., & Lapata, M. (2013). Models of semantic representation with visual attributes. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (pp. 572-582).

Silberer, C., & Lapata, M. (2012). Grounded models of semantic representation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (pp. 1423-1433).

Silberer, C., & Lapata, M. (2014). Learning grounded meaning representations with autoencoders. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (pp. 721-732).

Simonyan, K. & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representations* (pp. 1-14).

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, .V, & Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1-9).

Vigliocco, G., Kousta, S. T., Della Rosa, P. A., Vinson, D. P., Tettamanti, M., Devlin, J. T., & Cappa, S. F. (2014). The neural representation of abstract words: the role of emotion. *Cerebral Cortex*, *24*(7), 1767-1777.

Von Ahn, L., & Dabbish, L. (2004). Labeling images with a computer game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 319-326).

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. Behavior Research Methods, 45, 1191–1207.

Inattentional Blindness in Visual Search

Matt Chapman-Rounds (m.rounds@ed.ac.uk)
Christopher G. Lucas (c.lucas@ed.ac.uk)
Frank Keller (keller@inf.ed.ac.uk)
Institute for Language, Cognition and Computation
School of Informatics, University of Edinburgh
10 Crichton Street, Edinburgh EH8 9AB, UK

Abstract

Models of visual saliency normally belong to one of two camps: models such as Experience Guided Search (E-GS), which emphasize top-down guidance based on task features, and models such as Attention as Information Maximisation (AIM), which emphasize the role of bottom-up saliency. In this paper, we show that E-GS and AIM are structurally similar and can be unified to create a general model of visual search which includes a generic prior over potential non-task related objects. We demonstrate that this model displays inattentional blindness, and that blindness can be modulated by adjusting the relative precisions of several terms within the model. At the same time, our model correctly accounts for a series of classical visual search results.

Keywords: Inattentional Blindness; Conjunction Search; Visual Attention; Bayesian Modelling; Predictive Processing

Introduction

Visual search, where agents search for a target amongst distractors, is an important paradigm in the study of human attention (Wolfe, 1994) (see Figure 1 for an example trial). Inattentional blindness, where unexpected objects fail to capture attention, provides a useful insight into how constraints of processing and access lead to failures in the visual system (Simons, 2000). The literature on the two domains is distinct; in this paper we show that extending a model of visual search by adding an environmental prior produces a model that can reproduce empirical results from both domains.

The motivation for our extension hinges on the idea that the brain, due to the pressures of an ever changing environment, never *solely* models a task; it must always additionally maintain what are effectively generic, non-task-specific prior expectations about possible interesting states of the world. For example, in conjunction search (Nakayama & Silverman, 1986), where participants search for a target amongst distractors, a simple model of the search environment should include both "targets" and "distractors" (the statistics of which are learned during training), and "non-task entities" (which are unrelated to the task), as possible kinds. Ignoring non-task entities allows the brain to attend to (and successfully perform) a task, at the expense of potentially missing useful information about the world.

The contributions of this work are threefold. We demonstrate a successful joint model of visual search and inattentional blindness in which search is driven by saliency, generated using precision-weighted error terms. We show the structural equality of two distinct models of saliency, one top-down, and the other bottom-up. Finally, by constructing a model where both task relevant and task irrelevant stimuli contribute to saliency, we shed light on what it means to perform a task – namely, for an agent to have high confidence in its model of those stimuli that constitute the task, compared with its model of other possible stimuli.

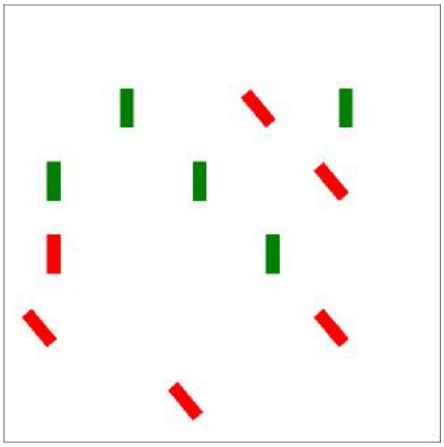


Figure 1: Example trial taken from Task 5 (see Results, below). Task is to find red vertical target amongst green vertical and red titled distractors.

Related Work

Conjunction Search

Empirically, we can distinguish between five forms of guidance in visual search (Wolfe & Horowitz, 2017). The two of interest to this work are bottom-up (where visual properties of aspects of a scene attract more attention than others, Koehler, Guo, Zhang, & Eckstein, 2014), and top-down (where executive control drives attention towards desired targets, Maunsell & Treue, 2006).

The majority of the many models of top-down visual search (Itti, Koch, & Niebur, 1998; Torralba, Oliva, Castelhano, & Henderson, 2006; Navalpakkam & Itti, 2006; Cave, 1999; Choi, Torralba, & Willsky, 2012)[1] share the basic structure of

[1] We cite Itti and Koch's work here, as well as in the section on bottom-up drivers of saliency, because whilst their work focusses on

Guided Search (GS; Wolfe, 1994): primitive visual features are detected across the retina by feature maps, which represent features via a coarse (i.e., highly overlapping) encoding. These feature maps are then passed through a local differencing operator, which enhances local contrasts, and the feature maps are combined using a top-down, task specific weighting to produce a saliency map. A Bayesian treatment of GS called Experience-Guided Search has been proposed by Mozer and Baldwin (2007).

There is also a body of work that focuses specifically on the bottom-up drivers of saliency (Koehler et al., 2014; Itti & Koch, 2001), an example of which is Attention as Information Maximisation (AIM), proposed by Bruce and Tsotsos (2009). These authors argue that the self-information of a location in an image, estimated on its surrounding context, is a good measure of its visual saliency.

Bottom-up models can be thought of as mapping the saliency of a task-neutral environment; but they provide no account of the relationship between this base saliency and the task at hand. Top-down models provide a qualitative account of various phenomena in visual search (see Results for a detailed discussion of relevant phenomena). However, a limitation of these models is that they do not attempt to model tasks as situated in a wider environment containing task-irrelevant stimuli, or competing tasks. A paradigm where modelling the relative saliency of task-relevant and task-irrelevant items becomes important is that of inattentional blindness (IB; Mack & Rock, 1998).

Inattentional Blindness

To model IB in the context of visual search, we use an "additional singleton" approach (see, e.g., Simons, 2000), where an unexpected single item has a distinctive unique feature, and that item is never the target item. There are several factors which have been shown to affect the rate of unexpected object detection when performing a task: increased cognitive load increases blindness (Kreitz, Furley, Memmert, & Simons, 2016), the similarity of the unexpected object to task-relevant objects increases the probability that the unexpected object will capture attention (Most et al., 2001; Simons & Chabris, 1999), as do shared features between task relevant objects and unexpected objects (Koivisto & Revonsuo, 2009).

Models of the causes of inattentional blindness range from claims of inattentional amnesia (we see the object, but fail to report it after the trial, Wolfe, 1999), to arguments that we are blind to objects we do not expect to see (Braun, 2001). More recent accounts have focused on the relationship between bottom-up saliency (which drives transient attentional capture) and a top-down attentional set, which governs whether transient attention is sufficient to generate sustained attention, and subsequent awareness (Most, Scholl, Clifford, & Simons, 2005). In spirit, our approach falls under this latter umbrella, but we show that blindness can be explicitly thought of as a result of a ratio of precisions in a mathematical model that extends the conjunction search literature (and is also able to replicate standard results in that domain).

saliency maps, they assume that these maps are combined according to top-down attentional drivers, which makes them less purely bottom-up than AIM, for example. See the section on bottom-up visual attention, below.

Model

Our starting point is Experience-Guided Search (E-GS; Mozer & Baldwin, 2007), a Bayesian treatment of GS developed to overcome a shortcomings of GS (Wolfe, 1994; Wolfe & Horowitz, 2017), namely that GS produces better than human performance without the addition of noise or regularising constraints on the top-down weighting of features. Mozer and Baldwin's premise is that a location in the visual field is salient if a target is likely to be at that location. They define $P(T_x = 1|\boldsymbol{F}_x)$ as a measure of saliency computed using statistics obtained from recent experience performing the task:

$$P(T_x|\boldsymbol{F}_x,\boldsymbol{\rho}) = \frac{P(T_x)\prod_i P(F_{xi}|T_x,\boldsymbol{\rho}_i)}{\sum_{t=0}^{1} P(T_x = t)\prod_i P(F_{xi}|T_x = t,\boldsymbol{\rho}_i)} \quad (1)$$

Here, $\boldsymbol{F}_x$ is the feature activity vector at retinal location x, T_x is the binary indicator of targethood, $\boldsymbol{\rho}$ parameterises the stimulus environment, and F_{xi} is the feature vector corresponding to feature i.

Whilst we lack the space to give a full treatment here, by assuming a generative model with $F_{xi}|\{T_x = t,\boldsymbol{\rho}\} \sim \text{Binomial}(n,\rho_{it})$, where ρ_{it} is the parameterising spike rate associated with feature i for target and non-target items ($t = 1, t = 0$), in the limit of reasonably large n we can approximate $P(F_{xi}|\ldots)$ as Gaussian, with mean $n\rho_{it}$ and variance $n\rho_{it}(1-\rho_{it})$. This allows the authors to derive a measure of saliency, S_{EGS}, as:

$$S_{EGS} = \sum_i \left[\Lambda_{\rho i0}(f_{xi} - n\rho_{i0})^2 - \Lambda_{\rho i1}(f_{xi} - n\rho_{i1})^2\right] \quad (2)$$

Where $\Lambda_{\rho it}$ denotes the precision (inverse variance) of the model's current estimate of ρ_{it}.[2]

This is a sum of terms, two for each feature i, which capture how surprising the activation corresponding to that feature, f_{xi}, is with respect to the target or not-target cases, the model's beliefs about which are parameterised by ρ_{i1} and ρ_{i0} respectively. The saliency of feature i increases if the observed activation is distant from the mean activity observed in the past in the absence of a target. It decreases if the observed behaviour is distant from the mean activity observed its presence. This surprisal is weighted by observed precisions: high variance features contribute less to saliency.

This remains a strictly task-based model, however. To expand it, we need to consider how the saliency of a feature changes under a generic, non-task specific prior. To do this, we turn to the literature on bottom-up measures of saliency,

[2] A difference between this presentation and that in Mozer and Baldwin is that we have not ignored the scaling n; whilst in E-GS only the relative magnitude of the terms in (2) is relevant, here we do care about how much data the model has seen.

in particular AIM (Bruce & Tsotsos, 2009). In doing so, we note the structural similarity between AIM and E-GS.

The premise of AIM is that those areas of an image that contain the most Shannon self-information are those that contain content of interest. Hence visual saliency is driven by surprise with respect just to visual input. First, "a sparse spatiochromatic basis" is generated in an unsupervised fashion using ICA, such that every image patch can be expressed as a vector of coefficient contributions (if projected back into image space, the coefficients, not incidentally, look a lot like Gabor filters and colour opposition patches[3]).

For each location x we can characterise the content of the local neighbourhood C_x by a vector $\boldsymbol{\alpha}_x$. For each of the i features, the p.d.f of the surround is estimated by making a histogram of all α_i values for every nearby patch. Then α_{xi}'s likelihood $P(\alpha_{xi})$ can be estimated from the histogram and thus its Shannon information content computed by $\log(1/P(\alpha_{xi}))$. Adding the Shannon information from each coefficient in $\boldsymbol{\alpha}_x$ gives us an estimate of the Shannon information contained in patch x, and hence the saliency of that patch:

$$S_{\text{AIM}} = -\sum_i \log P(\alpha_{xi}) \tag{3}$$

We then approximate the histogram $P(\alpha_{xi})$ as Gaussian distributed with mean $\bar{\alpha}_i$ and variance $\sigma^2_{\alpha i}$, which are the statistical mean and unbiased variance computed from the activations of surrounding patches for feature i.

This means we can rewrite (3) as:

$$S_{\text{AIM}} = \sum_i \left[\frac{1}{\sigma^2_{\alpha i}} (\alpha_{xi} - \bar{\alpha}_i)^2 \right] \tag{4}$$

Comparing to (2), we can see that if we make the same assumptions about the form of the likelihood of our incoming data, the measures of saliency used by E-GS and AIM are both sums of precision-weighted errors. The main difference is that AIM learns its model statistics from surrounding, synchronic activations, whereas E-GS learns its statistics diachronically, and with respect to the pertinent categories of a task oriented model.

Our final step is to argue that true saliency is a combination of many such terms, driven by the pressure to balance attention between task-driven stimuli and the world in which a task takes place. We therefore propose the following measure of saliency of location x, $\boldsymbol{S}_x$:

$$\boldsymbol{S}_x = \sum_i \left[-\Lambda_{i,1}(f_i - \mu_{i,1})^2 + \Lambda_{i,0}(f_i - \mu_{i,0})^2 + \Lambda_{i,\alpha}(f_i - \mu_{i,\alpha})^2 \right] \tag{5}$$

Where for clarity we have simplified the learned means to μ, and the learned precisions to Λ, for target, 1, distractors, 0, and non-task foils, α.

[3] AIM uses ICA to find a roughly orthogonal basis which the authors argue can be usefully compared to sparse coding in early visual cortex. E-GS uses the handcrafted sparse basis from GS. Both can summarised as: response activity is computed **in parallel** for multiple features. Activity which is surprising on a feature is salient.

We might think of the third term in (5) as constant: in the absence of any task, this is likely the term that dominates saliency. However, once I have a particular task, then the other terms will contribute to my estimate of the saliency. Rather nicely, we can also see how expertise might play a role: if a task has been repeated many times, then the precisions associated with those task-relevant features will be high, and so will dominate the saliency computation.

It is easy to see how this formulation could give rise to inattentional blindness: if a surprising object is neutral with respect to a task, then whether it affects the overall saliency measure will depend on the relative precision weighting of the first two terms and the third (if it is extremely surprising because it is blue, for example, but our task is clear-cut so the precisions associated with the top-down terms is high, then it still might not be that salient overall – if it is the only blue object we have seen, then its associated precision may be quite low). In a free viewing paradigm, the search task-relevant terms would be absent, the model would collapse to AIM, and unexpected objects would be salient. If the object possesses some task relevant features, then the first and second terms will contribute to its saliency, and it is more likely to be attended.

The leveraging of precision weighted errors to produce different effects can also be related to the predictive coding work of Friston and colleagues (Friston, Adams, Perrint, & Breakspear, 2012), where a free-energy minimising agent passes precision-weighted surprisals up a processing hierarchy, and expectations down. Indeed, Friston has explicitly claimed that (covert) attention can be thought of as precision weighting (Feldman & Friston, 2010), which our simple model certainly aligns with.

Methods

To test our model in a conjunction search paradigm, we simulated image environments of distractor and target objects on a 5×5 grid with a white background (See Figure 1 for an example trial). We represented images both at the object level and the pixel level (see Feature Spaces, below); in either case, at test time a vector valued representation $\boldsymbol{F}_x$ scaled to [0,1] was passed to a learned model, which returned the saliency score for each location x by computing $\mathbb{E}_\rho[\boldsymbol{S}_x]$, where we assume Beta priors over the expected activations, such that $\rho_{i,t} \sim \text{Beta}(\alpha_{i,t},\beta_{i,t})$, $\rho_{i,d} \sim \text{Beta}(\alpha_{i,d},\beta_{i,d})$ and $\rho_{i,nt} \sim \text{Beta}(\alpha_{i,nt},\beta_{i,nt})$. We used a Beta prior as the model assumes f_{xi} are rate activations, and hence fall in [0,1].

$\mathbb{E}_\rho[\boldsymbol{S}_x]$ is the sum of the expected value of each of the terms in (5). For the i^{th} feature of the k^{th} term at location x, this is:

$$\mathbb{E}_{\rho k}\left[\Lambda_{i,k}(f_{xi} - \mu_{i,k})^2\right] = n_k \Big[f_{xi}^2 \frac{(\alpha_{ik}+\beta_{ik}-1)(\alpha_{ik}+\beta_{ik}-2)}{(\alpha_{ik}-1)(\beta_{ik}-1)} - f_{xi} \frac{2(\alpha_{ik}+\beta_{ik}-1)}{\beta_{ik}-1} + \frac{\alpha_{ik}}{\beta_{ik}-1} \Big] \tag{6}$$

In the case of an object-level representation, x corresponds to an object in the image. In the case of the pixel-level representation, we computed saliency for every second pixel, which

gave reasonable results and was less costly than computing for every pixel.

For each task, the posterior beliefs of the model were learned from 100 labelled example trials. The learned model was then used to generate saliency maps for 1000 unlabelled trials, where, for object-level saliency maps, rank order of objects by saliency was taken to be directly proportional to response time (RT).

When pixel-level saliency maps were generated, we explicitly "saccade" through the most salient pixels in order, and introduce inhibition of return, which depresses the saliency S at pixel i at step t according to:

$$\begin{aligned} S_{i,t} &= S_{i,t} - (S_{i,t} \cdot R_{i,t-1}) \\ R_{i,t-1} &= \mathrm{G}(S_{i,t}) + \frac{1}{2} R_{i,t-2} \end{aligned} \tag{7}$$

where $\mathrm{G}(S_{i,t})$ is a Gaussian function, with a standard deviation $1/16$ the size of the image, of the distance of i from the target of the t^{th} saccade. The sequence of response times to any particular object is then taken to be proportional to the value of t when a pixel of that object is first visited.

Feature Spaces

Our saliency measure relies on the assumption that we have access to a sparse, independent feature representation of the visual space; either at the object level or at the pixel level. In principle, both should perform similarly, and so we tested our hypotheses (see Results) against both a variant of Guided Search's handcrafted approach to generating activations from features (Wolfe, 1994), and AIM's unsupervised approach (Bruce & Tsotsos, 2009), which uses ICA to generate a vector of activations from an image patch.

Guided Search has an eight-dimensional feature space: four activations correspond to colour, four to orientation. The four orientation dimensions are given by:

Steep: $\cos(2x)^{0.25}$, $-45 < x < 45$
Shallow: $|\cos(2x)|^{0.25}$, $-90 < x < -45$ and $45 < x < 90$
Left: $|\sin(2x)|^{0.25}$, $-90 < x < 0$
Right: $\sin(2x)^{0.25}$, $0 < x < 90$

The four colour receptors are red, yellow, green, and blue, described as the "quite arbitrary ... third root of triangluar functions" (Wolfe, 1994) that have peaks at positions evenly spaced at their ordinal positions in the spectrum. These activations are then passed through a local differencing operator to yield a bottom-up activation.

For unsupervised extraction of a sparse basis, we sampled 250,000 image patches of size 21×21 from a dataset of natural images (Hodosh & Hockenmaier, 2013), and used JadeICA (Cardoso, 1999), preserving 90% variance to extract an independent basis (27 dimensions were retained). ICA infers the mixing matrix, $\boldsymbol{B}$, between the independent causes and the perceived data (the patches). We then use $\boldsymbol{B}^{-1}$ to produce a vector of activations for any new patch.

Both approaches are claimed to produce activations corresponding to neuronal activity; Wolfe (1994) chose the eight features of Guided Search accordingly, and Bruce and Tsotsos (2009) argue that the roughly orthogonal basis learned by ICA can be usefully compared to sparse coding in early visual cortex. Hence it should be the case that our model produces similar performance from both forms of preprocessing.

Learning

The posteriors are computed using:

$$\begin{aligned} \rho_k | F_{xk} \sim \mathrm{Beta}\Big(&\lambda\alpha_k^0 + (1-\lambda)\Big[\alpha_k + \sum_{x \in X_k} f_{xk}\Big], \\ &\lambda\beta_k^0 + (1-\lambda)\Big[\beta_k + \sum_{x \in X_k} 1 - f_{xk}\Big]\Big) \end{aligned} \tag{8}$$

where X_k denotes the set of points labelled k in the training examples. This, as in Mozer and Baldwin (2007), interpolates between the prior distribution $\sim \mathrm{Beta}(\alpha_k^0, \beta_k^0)$ and the empirical posterior. This interpolation regularises the model's fit to the data, and improves its performance.

For all experiments, $\alpha_{id}^0 = \beta_{jt}^0 = 10$, $\alpha_{it}^0 = \beta_{jd}^0 = 25$, for all i and j, $\alpha_{int}^0 = \beta_{int}^0 = 10$, and $\lambda = 0.3$. These parameter values are mostly taken from Mozer and Baldwin (2007), as there was no reason to change them.

Results

We tested two hypotheses: that our model would reproduce a range of standard effects in visual search, and that our model could reproduce two standard results from the inattentional blindness literature.

Visual Search

To evaluate the performance of the model in the visual search paradigm, we followed Wolfe (1994) and Mozer and Baldwin (2007), and tested our model against six search tasks used to evaluate the original guided search model. These tasks are as follows. All graphs shown are using the eight simple features of guided search. Standard error bars are included.

1. Vertical target among homogeneous distractors (Figure 2): As the angle of the distractors increases from 0–55 degrees (where 0 is vertical), time to target should become constant with respect to the number of distractors (i.e., pop-out occurs).
2. Categorical search (Figure 3): Target among two types of distractors defined with respect to a single feature (angle of orientation). Distractors are 100 degrees apart, and target is 40/60 degrees from the distractors in two cases, but in the third case it is the only near vertical item, allowing pop-out.
3. Target-distractor similarity (Figure 4): Search efficiency for target among heterogeneous distractors. There are two target orientations, and two degrees of target similarity. For each orientation, search should be more efficient when target and distractors are dissimilar.

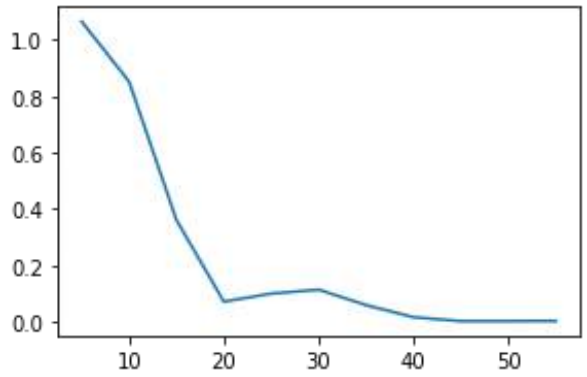


Figure 2: (Task 1) Horizontal; distractor orientation (degrees). Vertical; Gradient of time-to-target against number of distractors. Pop-out clearly occurs at around 20 degrees from the vertical.

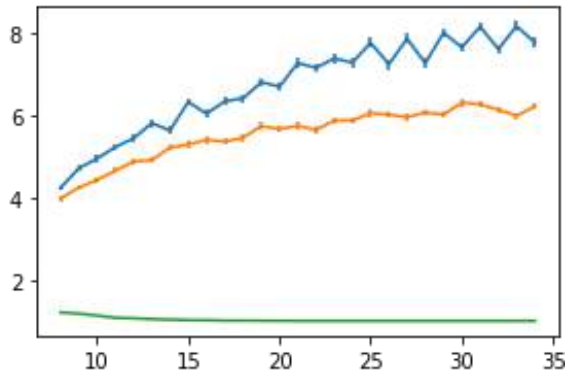


Figure 3: (Task 2) Horizontal; total number of distractors. Vertical; response time/time-to-target (in fixations, t). Blue; target at 10, distractors at -30 and 70 degrees. Orange; target at 20, distractors at -20 and 80 degrees. Green; corresponds to case where distractor is the only near-vertical item. Target at 10, distractors at -50 and 50 degrees.

4. Feature search asymmetry (Figure 5): It is more efficient to find a tilted bar among verticals than a vertical among tilted. This is because tilted items activate features that make them more discriminable; for example in the 8 dimensional feature space described above, one feature activates when presented with vertical objects, but two activate when presented with objects at 20 degrees.

5. Conjunction search – distractor confusability (Figure 6): Red vertical target among green vertical and red tilted distractors. Red tilt can be 90 or 40 degrees: both are inefficient, but should vary in relative difficulty.

6. Distractor ratio effect (Figure 7): Response times for red vertical target amongst red tilted and yellow vertical distractors, as a function of ratio of distractor types. Search should be most efficient in the extremes, where there are a minimum of distractors of one particular type.

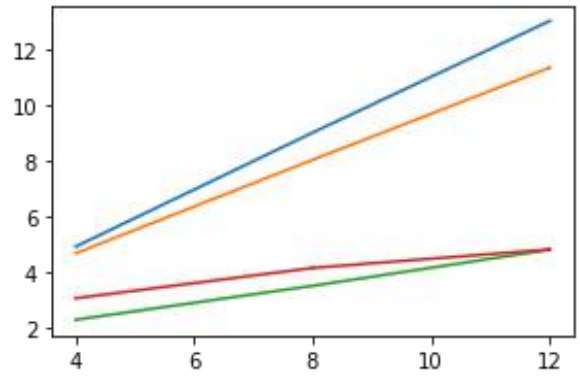


Figure 4: (Task 3) Horizontal; total number of distractors. Vertical; response time/time-to-target (in fixations, t). Blue and Orange; target at 0, distractors at -20 and 20 degrees, and -40 and 40 degrees respectively. Green and Red; target at 20, distractors at 0 and 40 degrees, and -20 and 60 degrees, respectively.

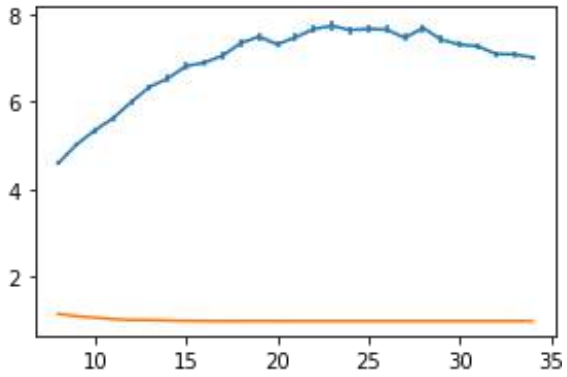


Figure 5: (Task 4) Horizontal; total number of distractors. Vertical; response time/time-to-target (in fixations, t). Blue; Target at 0, distractors at 20, orange; target at 20, distractors at 0.

Inattentional Blindness

We aimed to test two basic results in the inattention blindness literature. First, that performing a task reduces the probability of fixating or reporting unexpected objects, when compared to a task-free control (Simons & Chabris, 1999).

With reference to Equation (5), we assume that $\Lambda_{i,1} \approx \Lambda_{i,0} = \Lambda_T, \forall i$ (i.e., the two task-specific confidences are similar), the relative magnitude of Λ_T to Λ_α should be central to the relationship between performing a task, and corresponding inattentional blindness. This is because if Λ_T is much larger than Λ_α then the task-specific terms dominate the saliency score, and objects which are surprising in features that are not task specific have lower probability of detection.

In free viewing, however, where Λ_α is larger than, or equal to Λ_T (the task does not dominate attention), the context-dependent surprisals should contribute to the overall saliency, and generically unexpected objects (persons in gorilla suits, for example), are more likely to capture attention.

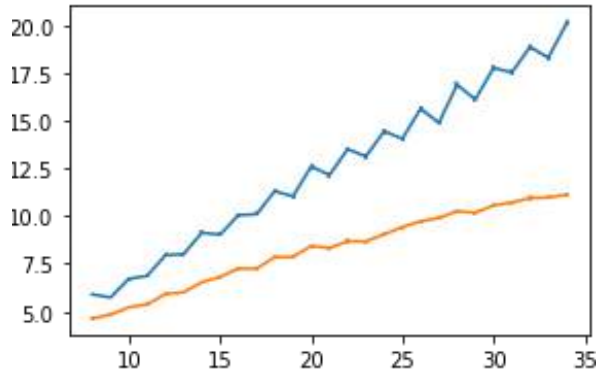


Figure 6: (Task 5) Horizontal; total number of distractors. Vertical; response time/time-to-target (in fixations, t). Red targets at 0 degrees, one set of green distractors at 0 degrees. Blue; second set of red distractors at 40 degrees. Orange; second set of red distractors at 90 degrees.

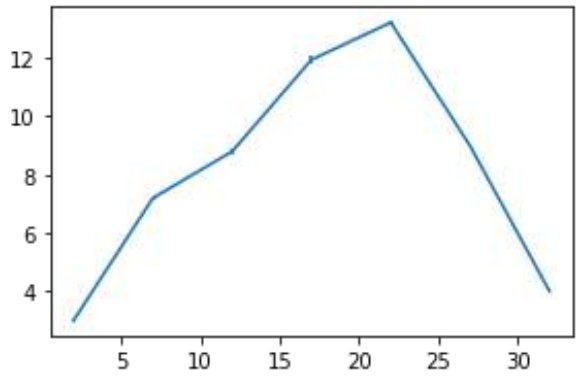


Figure 7: (Task 6) Horizontal; number of red tilted distractors (of a total 35 distractors). Vertical; response time/time-to-target (in fixations, t). Distractors were yellow vertical, and red 60 degrees. Target was red vertical.

To test this we introduced critical trials into our normal experiment. On a critical trial an unexpected (blue, left-leaning) object (see Figure 8a for an example) is also present along with the normal distractors and target. We varied the ratio n_T/n_α between 0.005 and 20. Figures 8b and 8c show a clear transition between the scenario in which the $\boldsymbol{\alpha}$ term dominates the saliency computation – in which the unexpected item pops out amidst the red task-relevant objects – and that in which the $\boldsymbol{T}$ terms dominate – where the same object does not pop out of the target and distractors, even though it is clearly surprising to an outside observer.

Second, we checked that if an unexpected object possesses features that are also task relevant, it is more likely to be fixated or reported (Most et al., 2001). We modified Task 2 (see Visual Search, above), as here the target and distractors are defined with respect to only one feature dimension. For a critical trial with a red target at 10 degrees, and red distractors at 30 and 70 degrees, we added an unexpected blue singleton at −70 or 15 degrees. Average number of fixations to target for the singleton at a task-relevant angle (70 degrees) was 13.99 ± 0.002. For the singleton at a task-irrelevant angle it was 2.0 ± 0.001. This was for a constant 12 distractors, and the ratio n_T/n_α was set to 100. This is quite a substantial difference (probably because the experimental set-up was as simple as possible), but it bears out our hypothesis.

Conclusion and Future Work

A weakness of this work is that as it is intended as a theoretical starting point, our analysis is primarily qualitative, and we have not compared the original predictions of our model to data from human participants. We will focus on these deficiencies in upcoming work via two main avenues.

The first approach is to test human participants to show that modulating the relative model precisions of (i.e., confidences in) targets specifically affects the probability that unexpected objects might be detected. If, for example, participants were initially provided only with a verbal descriptions of a visual target, we would expect probability of inattention to a non-target singleton to increase over the course of several trials, as participants became more confident in the target of their search task. We would also expect probability of inattention to be greater for a comparable task where participants are provided with a visual example of their target.

Our second approach, which lies solely in the conjunction search paradigm, would be to include distractors in a conjunction search task that shared no features with the target. We hypothesise that both overt indications of attention (fixations) and covert indications of attention (average time to target) to these non task-relevant objects would decrease over the course of several trials.

We have made three distinct contributions; we have presented a model of visual search that exhibits inattentional blindness, we have shown the equivalency of AIM and E-GS under certain assumptions, and we have argued that an interpretation of what it is to "perform a task" should be grounded on the relative precisions of parts of the brain's generative model.

We conclude that modelling task-based behaviour as explicitly located in a wider context can bear explanatory fruit.

Acknowledgements

This work was supported in part by the EPSRC Centre for Doctoral Training in Data Science, funded by the UK Engineering and Physical Sciences Research Council (grant EP/L016427/1) and the University of Edinburgh.

References

Braun, J. (2001). Its great but not necessarily about attention. *Psyche*, *7*, 6.

Bruce, N. D., & Tsotsos, J. K. (2009). Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision*, *9*(3), 5–5.

Cardoso, J. (1999). High-order contrasts for independent component analysis. *Neural Computation*, *11(1)*, 157-192.

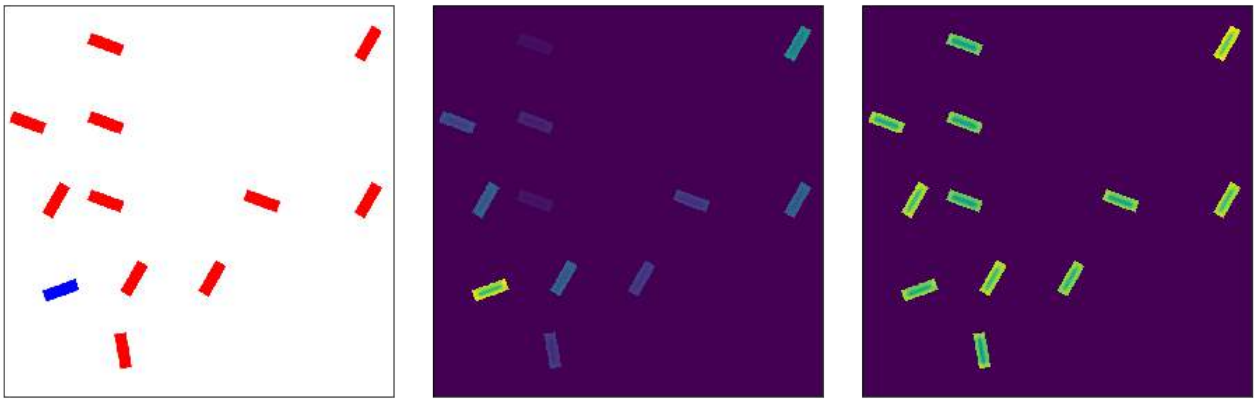


Figure 8: Maps of relative saliency in critical trial in search for red near-vertical target amongst red distractors. Left (a): trial image, Center (b): saliency map when $\boldsymbol{\alpha}$ term dominates the saliency computation, in which the unexpected item pops out amidst the red targets and distractors, Right (c): saliency map when $\boldsymbol{T}$ terms dominate, where the same object does not pop out.

Cave, K. R. (1999). The featuregate model of visual selection. *Psychological Research*, *62*(2), 182–194.

Choi, M. J., Torralba, A., & Willsky, A. S. (2012). Context models and out-of-context objects. *Pattern Recognition Letters*, *33(7)*, 853–862.

Feldman, H., & Friston, K. J. (2010). Attention, uncertainty, and free-energy. *Frontiers in Human Neuroscience*, *4*, 215.

Friston, K., Adams, R. A., Perrint, L., & Breakspear, M. (2012). Perceptions as hypothesis, saccades as experiments. *Frontiers in Psychology*, *151(3)*, 1-20.

Hodosh, M., & Hockenmaier, J. (2013). Sentence-based image description with scalable, explicit models. *CVPR Workshops*, 294-300.

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience*, *2*(3), 194.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, *20*(11), 1254–1259.

Koehler, K., Guo, F., Zhang, S., & Eckstein, M. P. (2014). What do saliency models predict? *Journal of Vision*, *14*(3), 14–14.

Koivisto, M., & Revonsuo, A. (2009). The effects of perceptual load on semantic processing under inattention. *Psychonomic Bulletin & Review*, *16*(5), 864–868.

Kreitz, C., Furley, P., Memmert, D., & Simons, D. J. (2016). The influence of attention set, working memory capacity, and expectations on inattentional blindness. *Perception*, *45*(4), 386-399.

Mack, A., & Rock, I. (1998). *Inattentional blindness*. Cambridge, MA: MIT Press.

Maunsell, J. H., & Treue, S. (2006). Feature-based attention in visual cortex. *Trends in Neurosciences*, *29*(6), 317–322.

Most, S. B., Scholl, B. J., Clifford, E. R., & Simons, D. J. (2005). What you see is what you set: sustained inattentional blindness and the capture of awareness. *Psychological Review*, *112*(1), 217.

Most, S. B., Simons, D. J., Scholl, B. J., Jimenez, R., Clifford, E., & Chabris, C. F. (2001). How not to be seen: The contribution of similarity and selective ignoring to sustained inattentional blindness. *Psychological Science*, *12*(1), 9-17.

Mozer, M., & Baldwin, D. (2007). Experience-guided search: A theory of attentional control. In *NIPS* (p. 1033-1040).

Nakayama, K., & Silverman, G. H. (1986). Serial and parallel processing of visual feature conjunctions. *Nature*, *320*, 264–265.

Navalpakkam, V., & Itti, L. (2006). Top–down attention selection is fine grained. *Journal of Vision*, *6*(11), 4–4.

Simons, D. J. (2000). Attentional capture and inattentional blindness. *Trends in Cognitive Sciences*, *4*(4), 147–155.

Simons, D. J., & Chabris, C. F. (1999). Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception*, *28*(9), 1059–1074.

Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological Review*, *113*(4), 766.

Wolfe, J. (1994). Guided Search 2.0 A revised model of visual search. *Psychonomic Bulletin & Review*, *1*(2), 202–238.

Wolfe, J. (1999). Inattentional amnesia. *Fleeting Memories*, *17(5)*, 71-94.

Wolfe, J., & Horowitz, T. S. (2017). Five factors that guide attention in visual search. *Nature Human Behaviour*, *1*, 0058.

Analysis of review quality by using gaze data during document review

Koki Saito (koki.saito@unisys.co.jp)
Nihon Unisys, Ltd., 1-1-1 Toyosu, Koto-ku, Tokyo, Japan
Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, Ishikawa, Japan

Shohei Hidaka (shhidaka@jaist.ac.jp)
Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, Ishikawa, Japan

Abstract

In software development, deliverables in an upstream process are reviewed to ensure their quality and to reduce error propagation to the downstream process. Methods are available for evaluating the review quality. In this study, we considered the defect detection process in a review of Requirement Definition Documents and tested a potential relationship between the gaze patterns and review quality. Specifically, we analyzed the relationship between the gaze patterns, with a primary focus on the blink rate, in a review of RDDs and detection accuracy. A significant nonlinear correlation between the blink rate and the detection accuracy was observed; moreover, the subsequent regression analysis also verified the blink rate as the best predictor of the review quality, notwithstanding the use of other gaze patterns. This result indicates that the blink rate is a major predictor of a type of review performance.

Keywords: gaze; blink rate; document review; review quality; signal detection theory; machine learning;

Introduction

In software development, it is important to ensure the quality of the specification and design document in the upstream process because they affect the quality of the deliverables in the downstream process. It is five to 200 times more expensive to correct defects in the downstream process than in the upstream process when we correct a low-quality deliverable affected by an ineffective specification (Boehm, 1981). Thus, it is preferable to maximize the defect detection in the upstream process.

In order to remove potential defects, it is common to review a document in an upstream process; moreover, numerous review methods have been used. However, individual differences in the review performance are likely to influence the review quality to a higher degree than the differences among the review methods (Uwano, Nakamura, Monden, & Matsumoto, 2007). A reviewer's performance also depends on the time limit for the task and the degree of the reviewer's concentration. Furthermore, although the defect detection rate based on the items indicated and the review rate are used for quantitatively evaluating the review quality, these indices by themselves are not adequate for accurately evaluating the review quality. First, the defect detection rate, for example, depends both on the quality of the reviewer and the quality of the document reviewed. As a result, we cannot assess whether a low detection rate implies low quality of the reviewer or high quality of the document. Second, these available indicators do not capture the different types of defects, such as simple typos, missing information, ambiguity, and misleading sentences. Accordingly, in this study, we explored a new indicator of the review quality, which characterizes the reviewer's performance and the potential types of defects. As a potential candidate for this indicator, we studied the gaze behavior in the document review task.

Recently, gaze data have been studied in software engineering (SE) to elucidate the cognitive process in various SE tasks such as code review (Sharafi, Shaffer, Sharif, & Gueheneuc, 2015). In SE, there are numerous studies targeting the review of a source code in the downstream process or review of "box and arrow" diagram such as Unified Modeling Language (UML). However, there are few studies on the review of documents in the upstream process (Sharafi, Guéhéneuc, & Soh, 2015). In fields other than SE, there has been studies on reading and understanding of narratives using gaze data (Augereau, Kunze, Fujiyoshi, & Kise, 2016; Campbell & Maglio, 2001; Okoso et al., 2015); however, there are few studies on the review process for detecting defects in a document.

Uwano et al., (2007) have defined the review process: "In the software review, a reviewer reads the document, understands the structure and/or functions of the system, then detects and fixes defects if any." They classified it into the three sub-processes: (1) reading, (2) understanding of the structure, and (3) detection/correction of defects.

Relevant to the three sub-processes above, past literature has reported the three major characteristics of eye blink as follows:

(A) An adult subject typically exhibits 20 eye blinks per minute (Bentivoglio et al., 1997).
(B) A task requiring certain external information such as reading tends to enhance external attention and reduce the number of eye blinks per unit time (Cho, Sheng, Chan, Lee, & Tam, 2000; Karson et al., 1981).
(C) A task requiring internal attention, such as mental arithmetic and association, increases the number of eye blinks per unit time (Cho et al., 2000; Karson et al., 1981).

In light of these observations, we hypothesize that the three sub-processes in the review process are related to the eye gaze patterns as follows: Process (1) is supposed to be associated with observation (B), wherein the rate of eye blinks would be reduced as it requires external information.

Processes (2) and (3) are supposed to be associated with observation (C), wherein the rate of eye blinks would be increased as it requires internal attention. Moreover, we suppose that both effective and ineffective reviewers are largely similar in the sub-process of reading (1); however, they would be different in the sub-processes of understanding (2) and detection (3). More specifically, we suppose that an effective reviewer would utilize more cognitive resources for the two sub-processes (2) and (3) than an ineffective one; as a result, a better reviewer would exhibit a higher rate of eye blinks per time.

Therefore, in this study, we executed an experiment that simulated a review process of a set of Requirement Definition Documents (RDD) and measured the reviewer's gaze patterns during the experiment. Then, we tested our above hypothesis by analyzing the relationship between the eye blinks and the review quality. In this experiment, we prepared a RDD material, to which we introduce defects; moreover, the review quality was defined based on whether these presumed defects were detected or not.

Our analysis of the review quality by using the gaze data revealed that the blinks were the most important component of the gaze data; a significant nonlinear correlation between the blink rate and the detection accuracy was apparent.

Experiment

In the experiment, each of the participants underwent two sessions: the review session and post-review session. In each trial in the review session, they were asked to review one page of the three types of the RDDs and then to mark the sentences with defects. After finishing 11 trials of the review session, they were asked to fill the demographic questionnaire.

Participants

We recruited 19 Japanese adults as the participants (16 male and three female) and the average age of them was 42.2 years (SD = 9.1), with nine of them in their 30s, four in their 40s, and six in their 50s. All of them were system engineer and nine of them had no RDD review experience. All of them had normal (corrected) vision.

Material

The set of original documents used in the review session were based on three types of RDDs that were in actual use at Nihon Unisys, Ltd. Each of the original RDDs was re-arranged such that each document had three pages of summary, three pages of functional requirement, three pages of non-functional requirement, and two additional sample documents—eleven pages in total. They were all in Japanese. On each page of a re-arranged RDD, we introduced a defect that was absent in the original document. In this study, the type of defects was the "omission" of certain necessary piece of information for requirement definition. A part of an original sentence was removed, which made the original definition ambiguous. In order to simulate a natural review process, we did not add more than a defect per page. As a result, we limited the number of sentences, including the one with a defect, to two per page; there were 17 sentences, including those with the defects, in the 11 pages. The demographic questionnaire included questions on age, gender, RDD review experience, document review experience, degree of concentration during review, and degree of comprehension to documents for review.

Procedure and Apparatus

In each trial, one page of the documents to be reviewed was presented on a computer screen; the participant's gaze patterns were measured by an eye tracker device during the document review. The eye tracker used in this experiment was gazepoint GP3HD eye tracker (Figure 1). The participants could spend as much time as they considered necessary for this review process.

After the review of each page, the participants were instructed to mark the sentences to be improved, on a printed document with the reviewed content; they were not informed about the type of defects introduced. This trial was repeated for 11 pages and the order of page was the same for all participants. The participants were not provided any break during the review trial. Moreover, they were instructed to maintain their head still as much as feasible in order to ensure accurate eye tracking.

After the review session, each of the participants was asked to answer the demographic questionnaire.

Figure 1: (left) Experimental situation, (right) eye tracker (set up at the bottom of the monitor)

Results

In the review session, the average, minimum, and maximum review times of the 19 subjects were 21, eight, and 40 min, respectively. In order to exclude the data with large numbers of eye tracking error, we performed the Smirnov–Grubbs test (Grubbs, 1969) to detect pages with valid fixation points less than 60% (Figure 2). Based on this test, we excluded data worth four pages (out of the total 209 pages—11 pages for each of the 19 subjects) from the rest of our analysis. We performed the subsequent analysis on the 205 pages of data with a sufficiently large rate of fixation.

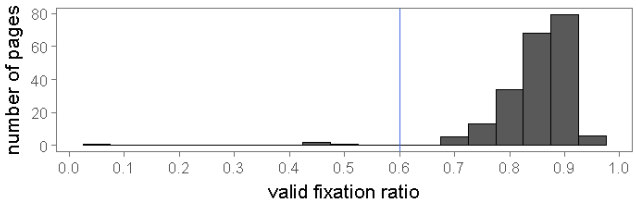


Figure 2: Histogram of valid fixation ratio each page

Review quality

In this study, we defined the correct review report for each unit of document based on the match between the participant's marked sentence and the sentence with a defect. Thus, the defect detection task was formulated as signal detection—the participant report a defect as either being detected or not, given a sentence with defect (signal in the ground truth) or otherwise (noise in the ground truth). We employed the signal detection theory (SDT) (Green & Swets, 1966) and treated the d-prime as an indicator of the accuracy of defect detection or the review quality. In the SDT, the respond bias and the sensitivity (d-prime) are distinguished from the rates of correct rating (the rate of defect detected marked to a sentence with defect) and false alarm (the rate of detect detected marked to a sentence without defect). The d-prime represents the deviation of the signal and noise distribution from the noise distribution as defined by

$$\text{d-prime} = \frac{M_{SN} - M_N}{\sigma_N}, \qquad (1)$$

where M_{SN} is the mean of the signal and noise distribution, M_N is the mean of the noise distribution, and σ_N is the standard deviation of the noise distribution. The d-prime is an indicator of the review quality; it can circumvent the effect of the potential response bias (the behavioral tendency to report detection regardless of the signal).

Analysis

In order to test our hypothesis, we analyzed the relationship between the blink rate and the review quality measured by the d-prime, in Analysis 1. In Analysis 2, we performed a model-based analysis of the relationship between the review quality and the gaze pattern not just the blink rate but also the other types of measurements such as fixation and saccade. The statistical analyses reported here were conducted with the free software R language (R version 3.4.1).

Analysis 1: Is the blink rate related to the review quality?

According to our hypothesis discussed in the introduction, the key sub-process in the review would require internal attention; thus, it would increase the blink rate. In order to verify this relationship between the blink rate and the review quality, the scatter plot of the blink rate and d-prime are shown in Figure 3. The corresponding correlation coefficient and other statistics are listed in Table 1. The maximal information coefficient (MIC) is a correlation coefficient calculated using mutual information; its application is feasible even with a nonlinear relationship. MIC-ρ^2 is an index of nonlinearity, and the maximal asymmetry score (MAS) is an index of non-monotonicity (Reshef et al., 2011). These results are summarized as follows:

- d-primes across the trials of the participants were distributed from - 1 to + 1.
- When d-prime was approximately zero, the blink rate was reduced from the mean blink rate and increased at non-zero d-prime values.
- The Pearson correlation coefficient between the d-prime and the blink rate was significant, although weakly negative.
- Both MIC and MIC-ρ^2 were large, and these together exhibited significant nonlinearity.

From the above facts, it was determined that the relationship between the blink rate and d-prime was a U-shaped or V-shaped nonlinear correlation, in which the blink rate was the smallest for d-primes near zero.

This result, both positive and negative d-primes across the trials, indicates the presence of two distinct groups of participants: One group detected the type of defects incorporated to a few sentences in the experimental manipulation; the other group detected the other type of defect (rather than only random sentences), which were not regulated explicitly in this experiment. Although we incorporated defects to a few sentences in the document, the original document (a RDD in use for some other purpose) is likely to have had certain other type of defects prior to this experimental manipulation. If so, the positive d-prime indicates the sensitivity to the expected type of defects incorporated in this experiment, whereas the negative d-prime indicates the sensitivity to certain unexpected type of defects originally in the RDDs.

Figure 4 shows the relationship between the d-prime and the response to noise in all the trials for each participant. We observed the general trend in these individual differences wherein those who exhibited a negative d-prime tended to detect "noise" as a "signal" (which may be interpreted as a defect by these participants) rather than the signal defined by the pre-experimental manipulation of the RDD. Thus, owing to the ambiguity of definition of the type of defects in the instruction, these participants were likely to detect the other types of defects (which were classified as "noise" by our definition).

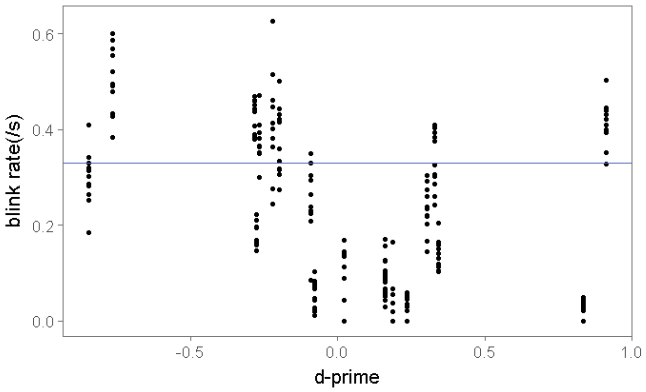


Figure 3: Scatter plot of d-prime and blink rate (blue line: mean blink rate in normal)

Table 1: Correlation coefficient between d-prime and blink rate

		Blink rate
d-prime	Pearson correlation	-.393
	p-value	.000
	MIC	.865
	MIC-ρ^2	.711
	MAS	.450

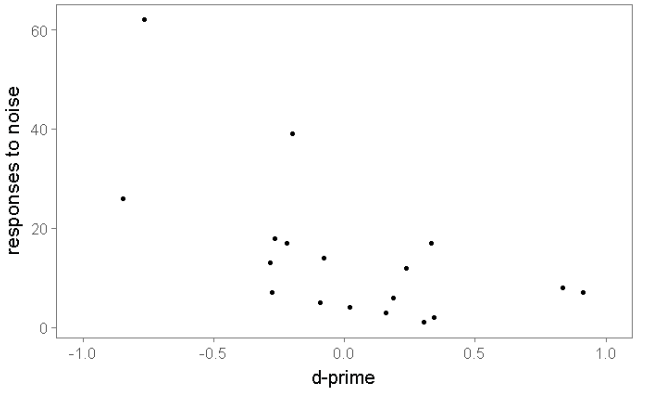


Figure 4: Scatter plot of d-prime and responses to noise

With respect to this interpretation of positive and negative d-primes, both positive and negative (non-zero) d-prime indicate higher sensitivity to certain types of defects; moreover, the blink rate was adequately correlated to the review quality of the potentially mixed types of defects.

This result appears to be evidence supporting for our hypothesis. However, it is likely that this result is caused by a spurious correlation owing to certain other features of gaze patterns, which are also correlated to the blink rate. Accordingly, in Analysis 2, we analyzed the d-primes, the indicator of review quality, with a collection of the other types of gaze features as well as the blink rate. Thereby, we evaluated the significance of the blink rate in the prediction of the d-primes, relative to the other types of gaze features such as fixation and saccade.

Analysis 2: Model-based analysis of review quality

In Analysis2, we constructed a model that predicts the type of detected defects measured by the positivity of the d-prime. Specifically, we employed a machine learning algorithm, random forest (RF) (Breiman, 2001) to predict the d-prime using the blink rate and other gaze patterns as the predictor. First, the set of features were calculated from the gaze patterns. Second, an RF regressor was constructed using the gaze features as a predictor of the d-primes in each trial. Then, we determined which gaze features is more informative for predicting the d-primes.

Extraction of features We extracted a set of 47 gaze features from the four fundamental gaze components: fixation, saccade, blink, and pupil (below). Forty six of these 47 features were originally defined by Bixler & D'Mello (2015); the blink rate was added to the list of features for the purpose of this study.

1. fixation: gazing on a single location
2. saccade: quick eye movement between fixation
3. blink: presence or absence of blink
4. pupil: size of the pupil

For each trial, the gaze pattern was characterized by these 47 features, and it was used to train the RF. As the RF calculated the importance of each feature simultaneously, the feature with low importance (with negligible significance for predicting the d-prime) could be removed.

We employed a sequential forward feature selection procedure using the RF as follows. First, the importance of each feature was calculated by RF using all the gaze patterns. Then, the root mean square error (RMSE) was calculated by RF using the highest importance feature. RMSE is the error from the actual value and is defined as

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(y_k - \hat{y}_k)^2}\,, \qquad (2)$$

where n is the number of pages, y_k is the specified d-prime of the kth page, and $\hat{y}_k$ is the d-prime predicted by the model. Second, RF was executed by adding the feature with the next highest importance, and the RMSE was similarly calculated. Third, the above process was repeated until the RMSE was the smallest, and the features effective for the review quality were extracted.

As a result, we obtained the 15 most important features extracted by RF, which are listed in Table 2. The blink rate was observed to be the most important. This indicates that the blink rate was the feature that was the most predictive of the d-prime.

Table 2: Features and their importance as extracted by RF

Rank	Features	Importance
1	blink rate	2.23
2	kurtosis of sccade duration	2.09
3	fixation duration / saccade duration ratio	1.93
4	number of blinks	1.93
5	max of saccade duration	1.88
6	range of saccade duration	1.80
7	kurtosis of pupil diameter	1.77
8	min of fixation duration	1.56
9	proportion of time spent blinking	1.50
10	mean of saccade duration	1.31
11	standard deviation of saccade duration	1.30
12	skew of saccade duration	1.27
13	proportion of horizontal saccade	1.14
14	median of saccade distance	1.08
15	median of fixation duration	1.01

Review quality prediction model Using the selected 15 features in Table 2, decision tree (DT), support vector regression (SVR), and multiple linear regression (MLR) models were constructed to predict the d-prime.

For validating these regression models, we performed a 10-fold cross validation (random split all trials) using their mean square errors (MAEs) and RMSEs. The MAE is defined as

$$\text{MAE} = \frac{1}{n}\sum_{k=1}^{n}|y_k - \hat{y}_k|. \qquad (3)$$

The results calculated for each algorithm by constructing the review quality prediction model are listed in Table 3. SVM exhibited the lowest MAE and RMSE, whereas the RF exhibited the second lowest ones. To determine how effectively the model predicts the d-prime, we present the scatter plot of the d-prime of the data and the one predicted by SVM in Figure 5 and the corresponding correlation coefficient in Table 4.

To summarize the above results, 15 out of the 47 gaze features are significantly important for predicting the review quality measured by the d-prime. Among these significant features, the blink rate was observed to be the most important. This result of the model-based analysis is consistent with the observation in Analysis 1: The blink rate has a higher predictability than the other types of gaze features; thus, it is unlikely that the relationship between the blink rate and the review quality is the result of a spurious correlation.

Table 3: Review quality prediction model

		RF	DT	SVR	MLR
d-prime	MAE	0.224	0.323	**0.214**	0.283
	RMSE	0.304	0.451	**0.289**	0.361

Table 4: Correlation coefficient between actual and predicted d-prime

		d-prime (actual)
d-prime (predicted)	Pearson correlation	.750
	p-value	.000

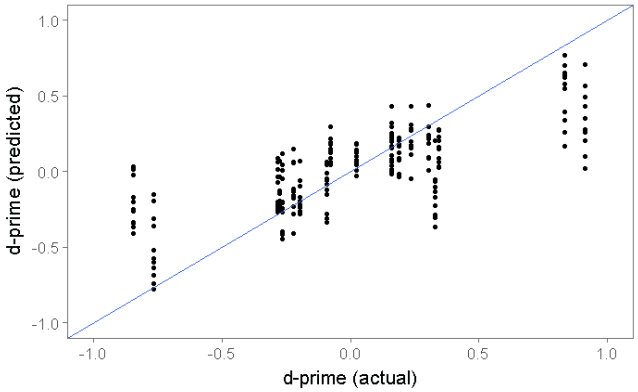


Figure 5: Scatter plot of actual d-prime and predicted d-prime

d-prime positive/negative classification model It is also intriguing whether we can classify the type of defects, which may be reflected as the positivity of the d-primes. Therefore, we next construct a classifier of the positivity of the d-prime by using the selected 15 gaze features. The algorithms adopted in this study were RF and support vector machine (SVM), which had exhibited a high prediction performance of d-prime in the experiment described in the previous section. Unlike the previous model-based analysis, we used the positivity of the d-prime as a class label rather than the d-prime value.

The classification accuracy is the coincidence rate between the predicted class and the specified class defined by

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}, \qquad (4)$$

where TP, TN, FP, and FN are the elements in the confusion matrix presented in Table 5.

Table 5: Confusion matrix

		Actual values	
		positive	negative
Predict values	positive	*TP*	*FP*
	negative	*FN*	*TN*

The classification accuracy for each algorithm is presented in Table 6. This result indicates that a classifier constructed upon the gaze features can predict the two potential types of defect detection with reasonably high accuracy 84%.

Table 6: Accuracy of d-prime positive/negative classification model

	RF	SVM
Accuracy	83.83%	**84.38%**

General Discussion

Blink rate and review quality

In prior study of SE, the gaze data has been used to elucidate cognitive process, however, the fixation and the saccade are often focused on and the blink rate is hardly taken consideration (Sharafi, Shaffer, et al., 2015). And it is also same trend in the study on reading and understanding of narratives (Augereau et al., 2016; Campbell & Maglio, 2001; Okoso et al., 2015). In this study, we focused the blink rate associated with each sub-process in the review and analyzed the relationship between the blink rate and the review quality.

In Analysis 1, we determined a nonlinear relationship between the blink rate and d-prime and that the blink rate was a U-shaped function of the d-prime estimated in each trial. This result is consistent with our hypothesis that the review quality (measured by d-prime) is related to the internal attention (measured by the blink rate). In Analysis 2, we tested the potential possibility that the relationship between the blink rate and d-prime is a spurious correlation owing to other confounding gaze features. We performed the regression analysis on the blink rate as well as the 46 other gaze patterns extracted from fixation, saccade, blink, and pupil. This analysis revealed that the blink rate was the most predictive of the d-prime; moreover, it indicated the blink rate to be a major gaze feature of the degree of review quality.

Limitations

It should be remarked that the result of Analysis 1, both positive and negative d-primes determined, was an indication of the likely presence of two potential groups of subjects detecting different types of defects owing to the ambiguity of the instruction for the review session. Considering this limitation of the experiment, it is feasible to have a few remarkable reviewers who detect both types of defects (the type defined and the other types not adequately defined in this study); such a reviewer may be evaluated near zero d-prime because he/she would detect both "signal" and "noise" according to our definition. Thus, in future works, an improved experimental design should have a list of defects covering most types of defects in the RDD material in order to prevent the problem of multiple types of defects.

Although we could not exhaustively classify all the types of defects using only the blink rate, Analysis 2 revealed that the positivity of the d-prime, indicating whether the detected defect was pre-defined or not, is classifiable with the blink rate and the other gaze features. There were numerous features on saccade duration in the 15 features. In general, these saccadic features capture gaze trajectory, and the saccade duration reveals the time of this trajectory. Thus, this result is likely to indicate the reading style such as reading order; moreover, the speed depends on the type of defects detected.

The set of RDDs used in this study was used for our customer's system development; its quality was supposed to be at least a specified level. However, it was likely that an immature RDD exhibited certain different types of potential defects than the defects introduced in this study. We cannot exclude the possibility that a reviewer's gaze pattern is affected by these mixed types of defects. This fact also necessitates a reconsideration of the experimental design that regulates the types of defects and investigates the relationship between the detection accuracy and the gaze patterns for each targeted defects.

References

Augereau, O., Kunze, K., Fujiyoshi, H., & Kise, K. (2016). Estimation of english skill with a mobile eye tracker. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing Adjunct - UbiComp '16* (pp. 1777–1781).

Bentivoglio, A. R., Bressman, S. B., Cassetta, E., Carretta, D., Tonali, P., & Albanese, A. (1997). Analysis of blink rate patterns in normal subjects. *Movement Disorders*, *12*(6), 1028–1034.

Bixler, R., & D'Mello, S. (2015). Automatic Gaze-Based Detection of Mind Wandering with Metacognitive Awareness. In F. Ricci, K. Bontcheva, O. Conlan, & S. Lawless (Eds.), *User Modeling, Adaptation and Personalization* (pp. 31–43). Cham: Springer International Publishing.

Boehm, B. W. (1981). *Software Engineering Economics* (1st ed.). Upper Saddle River, NJ, USA: Prentice Hall PTR.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32.

Campbell, C. S., & Maglio, P. P. (2001). A Robust Algorithm for Reading Detection. In *Proceedings of the 2001 Workshop on Perceptive User Interfaces* (pp. 1–7). New York, NY, USA: ACM.

Cho, P., Sheng, C., Chan, C., Lee, R., & Tam, J. (2000). Baseline blink rates and the effect of visual task difficulty and position of gaze. *Current Eye Research*, *20*, 64–70.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics. Signal detection theory and psychophysics*. Oxford, England: John Wiley.

Grubbs, F. E. (1969). Procedures for Detecting Outlying Observations in Samples. *Technometrics*, *11*(1), 1–21.

Karson, C. N., Berman, K. F., Donnelly, E. F., Mendelson, W. B., Kleinman, J. E., & Wyatt, R. J. (1981). Speaking, thinking, and blinking. *Psychiatry Research*, *5*(3), 243–246.

Okoso, A., Toyama, T., Kunze, K., Folz, J., Liwicki, M., & Kise, K. (2015). Towards Extraction of Subjective Reading Incomprehension: Analysis of Eye Gaze Features. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems* (pp. 1325–1330). New York, NY, USA: ACM.

Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., … Sabeti, P. C. (2011). Detecting Novel Associations in Large Data Sets. *Science*, *334*(6062), 1518 LP-1524.

Sharafi, Z., Guéhéneuc, Y.-G., & Soh, Z. (2015). A Systematic Literature Review on the Usage of Eye-tracking in Software Engineering. *Elsevier Journal of Software and Information Technology (IST)*.

Sharafi, Z., Shaffer, T., Sharif, B., & Gueheneuc, Y.-G. (2015). Eye-Tracking Metrics in Software Engineering. In *2015 Asia-Pacific Software Engineering Conference (APSEC)* (pp. 96–103).

Uwano, H., Nakamura, M., Monden, A., & Matsumoto, K. (2007). Exploiting Eye Movements for Evaluating Reviewer's Performance in Software Review. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, *90*(10), 2290–2300.

Investigating the role of the visual system in solving the traveling salesperson problem

Zahra Sajedinia (zsajedin@purdue.edu)
Department of Psychological Sciences, 703 Third Street
West Lafayette, IN 47907 USA

Zygmunt Pizlo (zpizlo@uci.edu)
Department of Cognitive Sciences, University of California
Irvine, CA 92697 USA

Sébastien Hélie (shelie@purdue.edu)
Department of Psychological Sciences, 703 Third Street
West Lafayette, IN 47907 USA

Abstract

This article used an empirical experiment and a computational model to test the hypothesis that humans rely on the visual system to solve the traveling salesperson problem (TSP). We tested two consequences of this hypothesis: (1) humans should perform better on Euclidean TSP than not–Euclidean TSP; (2) a model of the visual system should account for performance in Euclidean TSP. Participants were asked to solve Euclidean or not–Euclidean TSP, and a pyramid model of the visual system was used to solve the same tours as the humans. The results show that deviations from the optimal tour were smaller in Euclidean problems than in not–Euclidean problems, and the fit of the pyramid model to human performance was worse on not–Euclidean problems then on Euclidean problems. These results suggest that participants solve Euclidean problems with the visual system, but that other mechanisms are needed to succesfully solve non–visual problems.

Keywords:

Problem Solving; Visual Processing; Traveling Salesperson Problem; Pyramid Model

Introduction

A problem is a situation in which an agent seeks to attain a given goal without knowing how to achieve it. Humans solve problems every day. Example problems include winning at tic–tac–toe or winning a battle, air traffic control, control of an uninhabited vehicle, getting to checkmate in chess, visually–guided navigation, proving a logic theorem, solving math and physics problems, cracking the enigma code, or formulating a new scientific theory. Some problems are more visual, such as planning a tour around a grocery store, while others are more abstract, such as proving a theorem using predicate logic. In this conference article, we focus on the Traveling Salesperson Problem (TSP), a well–known optimization problem. In the TSP, a set of points is presented to participants. Each point represents a city, and the goal is to find the shortest possible route that visits all the cities exactly once, and returning to the starting city. We refer to this route as a TSP tour. The TSP has high relevance since it (1) has an important visual component (i.e., cities or points are spatially laid out on a map) and (2) it has important real–life application in many areas such as logistics, transportation, and shipping.

TSP has been studied extensively by cognitive scientists to reveal the underlying processes in human problem solving (van Rooij et al., 2006; Chronicle et al., 2008; Dry et al., 2006; MacGregor, 2013). One reason that makes the TSP an interesting problem for cognitive scientists is that the problem space of the TSP is very large. Even for solving a 16 city TSP, there are 6×10^{11} possible solutions, which is more than the number of neurons in the human brain (Azevedo et al., 2009). Also, the TSP is proven to be computationally NP–hard, meaning that there is no algorithm that can find an exact optimal solution for the TSP in polynomial time (Pizlo & Stefanov, 2013).

Human working memory can only store and manipulate a few items at a time and cannot make more than a few comparisons at a time (Pizlo & Stefanov, 2013). Yet, even with these severe limitations in memory and processing power, humans are able to solve the TSP near optimally in approximately linear time (MacGregor & Chu, 2011; Pizlo et al., 2006). How can humans with these limitations be able to solve the TSP fast and near optimally? What cognitive systems and processes have evolved to solve the TSP in the human brain?

Goals and Hypotheses

Pizlo and colleagues have argued that the TSP is solved by parallel processes in a pyramid–like hierarchical architecture of the visual system (Graham et al., 2000; Pizlo et al., 2006; Pizlo & Stefanov, 2013). The assumption that humans solve the TSP visually has important implications on the types of problems that can be solved. The human visual system has evolved in Euclidean space, so the visual system likely assumes a Euclidean cost function when solving optimization problems. As a result, performance in optimization problems with not–Euclidean or non–metric cost functions might be impaired.

To test this hypothesis, we designed a TSP experiment where participants solved either a regular (Euclidean) or not–Euclidean TSP. The participant's data was then compared with tours produced by a well–known computational model of the visual system, namely the pyramid model (Adelson et al., 1984; Pizlo et al., 1995). According to our hypothesis,

human participants should perform well in the Euclidean version of the TSP but not in the not–Euclidean version of the TSP. Further, the pyramid model should provide a good account of participant TSP tours in the Euclidean TSP but not in the not–Euclidean TSP. These results would support the hypothesis that participants are solving the regular TSP using the visual system, but not the not–Euclidean TSP. Further, the compensatory mechanisms used to solve the not–Eucliean TSP are not as efficient as the visual system at solving optimization problems.

Method

The first aim of this study was to explore how humans perform in different conditions of the TSP (i.e., Euclidean and not–Euclidean). The second aim was to explore how human performance is compatible with the visual pyramid model. The experiment and model are described in turns.

Participants

Ninety–one Purdue undergraduate students participated in the experiment for course credit. Participants were randomly assigned to one of three conditions: Single–color ($n = 36$), Colored–with–no–switch–cost ($n = 28$), and Colored–with–switch–cost ($n = 27$).

Apparatus and Stimuli

The stimuli were 30 maps each generated by putting 50 randomly scattered cities (points) in a $900px \times 900px$ display. The minimum distance between two cities was set to $50px$ to prevent overlapping points. The resulting set of 30 maps was used to create two different stimulus sets. In the first stimulus set, all cities were colored red. This stimulus set is referred as containing *single–color maps* (See Figure 1a). In the second stimulus set, half of the cities (points) were randomly selected and colored red. The remaining cities (points) were colored blue. This stimulus set is referred as containing *colored maps* (See Figure 1b).

The experiment was run on a regular PC. Stimuli were displayed in a 21–inch monitor (1,920 × 1,080 resolution). Participants responded by clicking on the city (point) that they wanted to visit next using a regular computer mouse. After each mouse click, a dark blue edge was drawn between the last visited city and the city that was clicked in the current trial. The order of the city visited was recorded.

Procedure

Each participant solved all 30 maps in one of three conditions. (1) *Single–color (Euclidean)*: This was a typical TSP experiment. The first stimulus set was used (i.e., single–color maps). Participants were asked to find the shortest TSP tour on each map, one map at a time. The cost between cities was Euclidean (i.e., the distance on the screen). No feedback was provided. (2) *Colored–with–switch–cost (not–Euclidean)*: The second set of stimuli was used (i.e., colored maps). In this condition, the cost between two points was not always Euclidean. Specifically, when travelling from a blue city to a red city (or vice–versa), the calculated distance (cost) was twice the distance on the screen. Otherwise, when travelling between two cities of the same color, the distance was as seen on the screen. Note that this arrangement can break the triangle inequality and make the cost non–metric. (3) *Colored–with–no–switch–cost (control)*: Similar to (2), this condition used the second set of stimuli (i.e., colored maps). However, the distance between two points was always the distance on the screen, so the colors could be ignored. This condition was designed to control for possible grouping effects that could be created by having cities of two different colors. In all conditions the experimenter explained the cost structure to the participants (as described above) and instructed them to find the tour with the smallest cost for each map.

Pyramid Model

A pyramidal architecture refers to multiple representations of the input data, with different representations having different scales and resolutions. In vision, the input data is the retinal image and the first layer is represented by the retinal ganglion cells. Each ganglion cell receives information from a particular region of the retina called the cell's receptive field. Receptive fields of different cells partially overlap. In the second layer of the pyramid, each "parent" cell receives input from several "child" cells. In the third layer, each "grandparent" cell receives input from several of its children. This process continues until a single cell on the top of the pyramid can "see" the entire image. Cells at lower layers can see small parts of the retinal image but they can process the information with high spatial resolution. Cells in higher layers can see larger parts of the retinal image but with lower resolution. More generally, cells in higher layers can handle only some statistical information about their receptive fields. The mean value of some property, like intensity, speed, contrast and so on, is the simplest example.

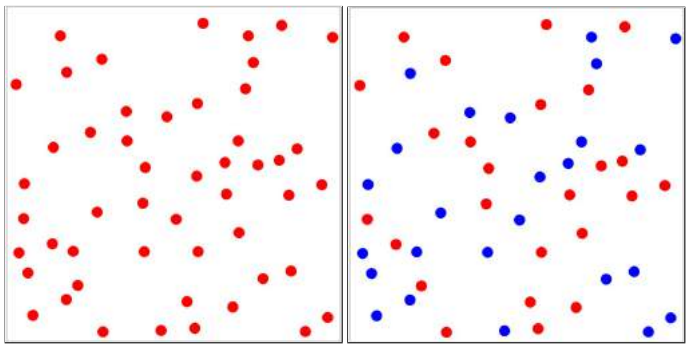
We implemented a Pyramid model adapted from Pizlo et al. (2006). The algorithm is presented in Figure 2. We used Python for our implementation. Inputs to the model were the maps that the participants in the experiment had solved. Because we hypothesized that the visual system evolved in a Euclidean world, the distances between the cities were considered Euclidean in all three conditions. This corresponds to the visual system not being able to process not–Euclidean distances.

Finding Optimal Tours

We used NEOS server of Concorde TSP solver to find the optimal TSP tours for each map. Concorde is one of the best exact TSP solvers currently available. It is available freely for academic use: https://neos-server.org/neos/solvers/co:concorde/TSP.html.

Results

One participant in the Single–color condition had tours that were three standard deviations longer then the condition



(a) Single–color map (b) Colored map

Figure 1: Example maps used in the experiment.

mean. All tours produced by this participants were not included in the following analyses.

Human Performance

Figure 3 shows typical example solutions produced by participants in each condition. As can be seen, the colored–with–switch–cost tour was qualitatively different from those obtained with the single–color and with colored–with–no–switch–cost conditions. Specifically, the not–Euclidean condition included a number of path crossings, which would be suboptimal in Euclidean space (but could be optimal in not–Euclidean space). These crossings were not observed when colors were present without a switch cost.

To quantify the participant performances, the error (i.e., deviation from optimal) was calculated for each map:

$$error_{ji} = \frac{(S_{ji} - O_i)}{O_i} \quad (1)$$

where $error_{ji}$ is the error or participant j on map i, S_{ji} is the length of the tour produced by participant j on map i, and O_i represents the length of the optimal tour for map i .

Table 1 presents the mean error in each condition. As can be seen, the single–color error was 12.6% and the colored–with–no–switch–cost (Euclidean) error was 12.7%, which is almost half of the error observed in the colored–with–switch–cost (not–Euclidean) TSP condition. This shows that participants perform well in Euclidean space but struggle in not–Euclidean space. Also, participants were able to ignore the irrelevant color and the longer tours obtained in the not–Euclidean condition were not caused by a perceptual effect of the city colors. Hence, larger errors for the not–Euclidean condition were not the result of unwanted color grouping effects.

To investigate if the observed differences were statistically significant, we performed Holm–corrected pairwise comparisons t–tests for all three conditions. Error in the not–Euclidean condition significantly differed from error in the single–color $(t(60) = 6.10, p < .0001)$ and error in the color–with–no–switch–cost $(t(53) = 5.63, p < .0001)$ conditions. The two Euclidean conditions did not differ from each other $(t(61) < 1, n.s.)$.

Table 1: Mean participant error in each condition

Condition	Error
Single–color	12.6%
Colored–with–no–switch–cost	12.7%
Colored–with–switch–cost	20.7%

The results show that the errors for the single–color and control conditions were not statistically different. However, the colored–with–switch–cost condition differed from the other two conditions. These statistical differences clearly show that participants' performances were highly dependent on the problem being Euclidean or not–Euclidean, and support the hypothesis that the visual system may assume a Euclidean cost function in solving the TSP.

The performance of the Pyramid model

In Table 2, we compared the Pyramid model generated tours with optimal tours. As can be seen, the error is 14.2% for both Euclidean conditions (single and color), and it increased to 34.5% for the not–Euclidean condition. As expected, the model error was similar to humans in the Euclidean conditions. The RMSD was 4.6% in the single–color condition and 4.5% in the color–with–no–switch–cost condition. However, the model provided a poor fit of human performance in the not–Euclidean condition (RMSD = 15.0%). Assuming that the Pyramid model is an adequate model of human vision, this result suggest that participants solving the Euclidean TSP used the visual system (good model fit), but not the not–Euclidean TSP (poor model fit). Since the participants were doing better than the model in the not–Euclidean condition, this result also suggest that participants may have access to a separate (compensatory) mechanism to attempt to solve not–Euclidean TSP. The pyramid model, in contrast, was purely a model of the visual system and could only deal with Euclidean spaces.

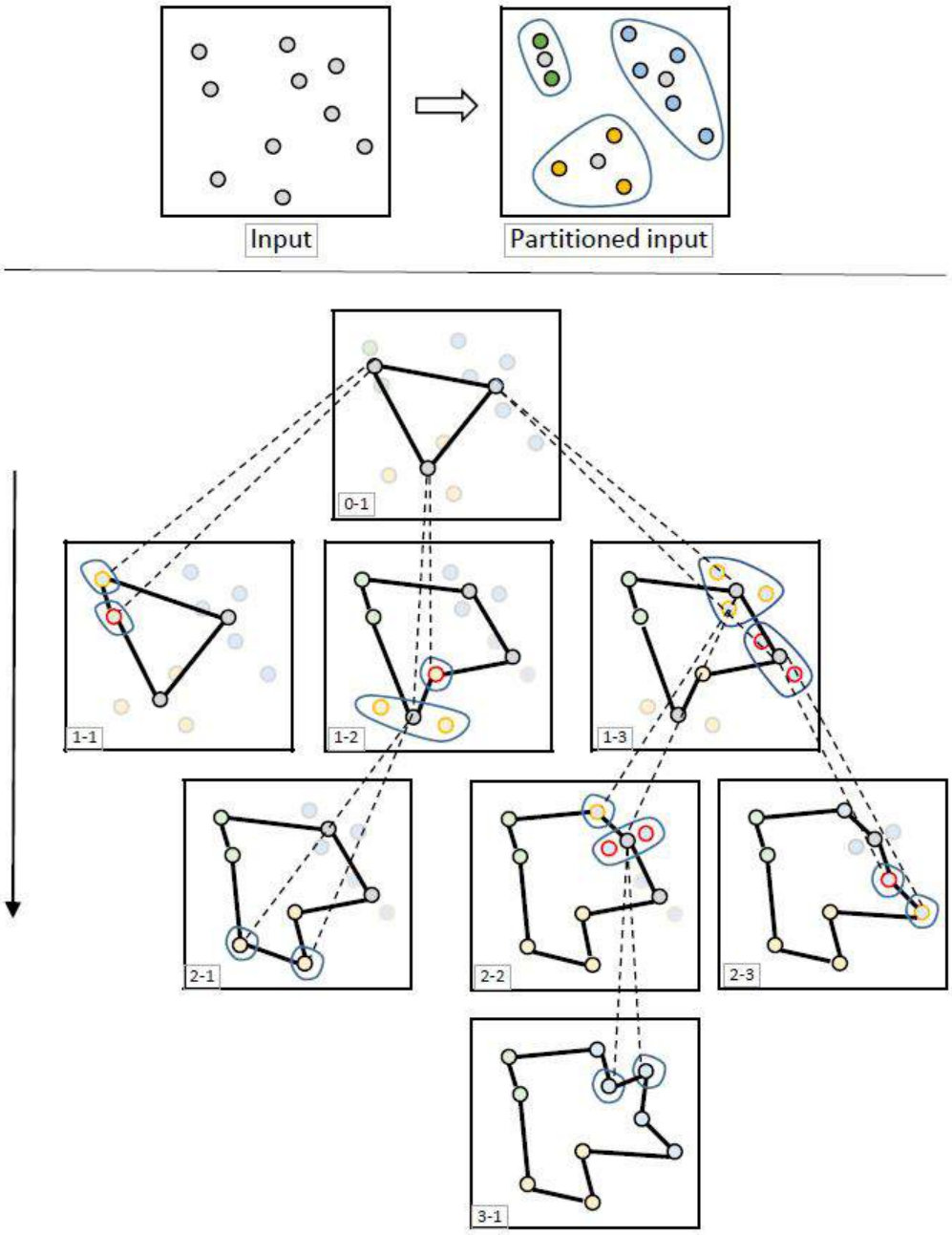
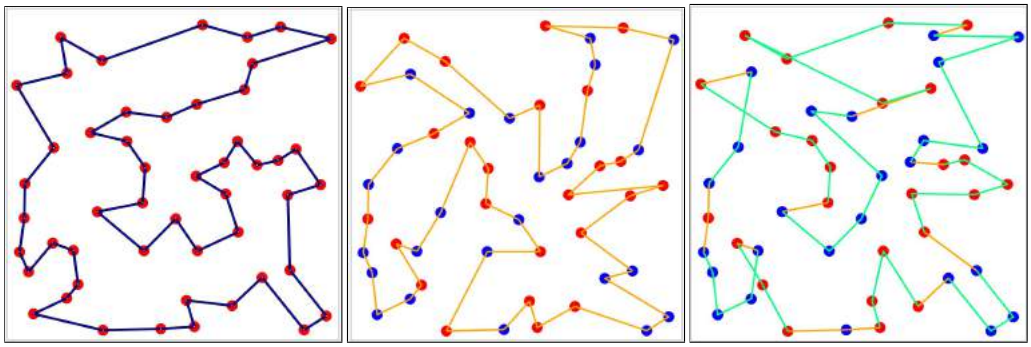


Figure 2: A representation of the Pyramid model. The top row shows the input as it gets partitioned into k clusters (by using a clustering algorithm, such as k–means). In this example, $k = 3$. Next, the pyramid is built. The root of the Pyramid (0–1) is the TSP solution for the centers of the clusters for the partitioned input. The solution for this TSP at the root is trivial because there are only three points and all three points are connected to each other. In the next level of the pyramid (level 1), each cluster is considered separately and recursively repeats the clustering until there is only one point (or city) in each cluster. For example, (1–1) shows the partition of the top–left cluster into k clusters (if the number of points is smaller than k, then $k-1$ is used, here $k = 2$), and a TSP solution for this cluster is found. Since, there were only two points, the solution is trivial, and the two points were connected to each other. Then by brute–force (considering all possibilities), the incoming and outgoing edges are connected to this cluster to obtain the shortest edges. The model then moves to the next cluster (1–2) and repeats the same procedure, until there is no non–visited cluster.



(a) single–color (b) colored–with–no–switch–cost (c) colored–with–switch–cost

Figure 3: Sample tours produced by participants in each condition.

Table 2: The error of the Pyramid model.

Condition	Error
Single–color	14.2%
Colored–with–no–switch–cost	14.2%
Colored–with–switch–cost	34.5%

Discussion

This article used an empirical experiment and a computational model to test the hypothesis that humans solve the TSP by assuming an Euclidean cost function. This assumption follows from the TSP being solved visually, and the visual system having evolved in an Euclidean world. We specifically tested two consequences of this hypothesis, namely that humans would perform better on Euclidean TSP than not–Euclidean TSP and that a model of the visual system could account for performance in Euclidean TSP. Participants were asked to solve the TSP in three conditions, two Euclidean and one not–Euclidean. A pyramid model of the visual system was used to solve the same tours as humans. The results show that the deviations from the optimnal tours were almost twice as small in Euclidean problems than in not–Euclidean problems, and the fit of the pyramid model to human performance was three times worse on not–Euclidean problems than on Euclidean problems.

Relevance for Problem Solving Research

Some problems are visual, like TSP on a Euclidean plane or visual navigation, but other problems may not have an obvious visual representation. Algebra problems, first order logic, and chess are examples. Logic is not visual, but set theory, with Venn diagrams, provides a visual version for at least some logical problems. However, not all problems are amenable to a useful visual representation. In these cases, the massively parallel nature of the visual system is no longer sufficient: problems need to be solved sequentially. One possibility is to use reinforcement learning (Sutton & Barto, 1998). In this framework, the agent is a sequential decision–making system and the environment is another system evaluating the distance between the current problem state and the goal state (Dandurand et al., 2012). In visual cases, the environment could be the visual system with geodesic estimates. In more abstract cases, the environment could be a meta–cognitive system used to evaluate states and rewards. Regardless of how the environment is implemented, actions are selected in each state by using a policy. The policy numerically describes the desirability of each action in each state. The goal of reinforcement learning is to find a policy that maximizes the return, which is the sum of all future rewards, until the problem is solved. However, any sequential system attempting to solve a NP–hard problem, such as the TSP, will quickly be overwhelmed by complexity. This could explain why human participants did better than the pyramid model in not–Euclidean TSP but did not do as well as in the Euclidean problems.

Future Work and Limitations

Future work can be directed in two ways. First, we can further test the theory of the engagement of the visual system in solving the TSP. It can be done by studying whether human performance is compatible with other characteristics of the visual system such as its limited ability to learn. The second direction is proposing a more complete model of human problem solving. Implementing a dual–system model of problem solving, including both a parallel visual module and a sequential decision–making module can be a promising direction. Tentatively, using a reinforcement learning agent for sequential decision–making would allow for learning in problems that cannot be solved visually. This possible dissociation in learning ability for visual and non–visual problems may allow for optimizing the way we represent and solve problems. Future work should be devoted to implementing and testing such a model.

Acknowledgement

This research was supported in part by the National Eye Institute (award #1R01EY024666-01 to PZ) and the National Science Foundation (award #1662230 to SH).

References

Adelson, E. H., Anderson, C. H., Bergen, J. R., Burt, P. J., & Ogden, J. M. (1984). Pyramid methods in image processing. *RCA Engineer*, *29*, 33–41.

Azevedo, F. A., Carvalho, L. R., Grinberg, L. T., Farfel, J. M., Ferretti, R. E., Leite, R. E., ... others (2009). Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. *Journal of Comparative Neurology*, *513*(5), 532–541.

Chronicle, E. P., MacGregor, J. N., Lee, M., Ormerod, T. C., & Hughes, P. (2008). Individual differences in optimization problem solving: Reconciling conflicting results. *The Journal of Problem Solving*, *2*(1), 4.

Dandurand, F., Shultz, T. R., & A, R. (2012). ncluding cognitive biases and distance-based rewards in a connectionist model of complex problem solving. *Neural Networks*, *25*, 41–56.

Dry, M., Lee, M. D., Vickers, D., & Hughes, P. (2006). Human performance on visually presented traveling salesperson problems with varying numbers of nodes. *The Journal of Problem Solving*, *1*(1), 4.

Graham, S. M., Joshi, A., & Pizlo, Z. (2000). The traveling salesman problem: A hierarchical model. *Memory & Cognition*, *28*, 1191–1204.

MacGregor, J. N. (2013). Effects of cluster location on human performance on the traveling salesperson problem. *The Journal of Problem Solving*, *5*(2), 3.

MacGregor, J. N., & Chu, Y. (2011). Human performance on the traveling salesman and related problems: A review. *The Journal of Problem Solving*, *3*(2), 2.

Pizlo, Z., Rosenfeld, A., & Epelboim, J. (1995). An exponential pyramid model of the time course of size processing. *Vision Research*, *35*, 1089–1107.

Pizlo, Z., & Stefanov, E. (2013). Solving large problems with a small working memory. *The Journal of Problem Solving*, *6*(1), 5.

Pizlo, Z., Stefanov, E., Saalweachter, J., Li, Z., Haxhimusa, Y., & Kropatsch, W. G. (2006). Traveling salesman problem: A foveating pyramid model. *The Journal of Problem Solving*, *1*(1), 8.

Sutton, R., & Barto, A. (1998). *Reinforcement Learning: An Introduction*. MIT Press.

van Rooij, I., Schactman, A., Kadlec, H., & Stege, U. (2006). Perceptual or analytical processing? evidence from children's and adult's performance on the euclidean traveling salesperson problem. *The Journal of Problem Solving*, *1*(1), 6.

Learning with an algebra computer tutor: What type of hint is best?

Kyle Sale (kyle.sale@carleton.ca)
Institute of Cognitive Science, Carleton University
Ottawa, Ontario, Canada

Kasia Muldner (kasia.muldner@carleton.ca)
Institute of Cognitive Science, Carleton University
Ottawa, Ontario, Canada

Abstract

While there is substantial evidence showing that assistance provided to students during problem-solving activities influences learning outcomes, it is not yet clear how to best design educational technologies to maximize learning through various types of assistance. One common type of assistance corresponds to hints delivered by an educational technology. To date, however, there is little research on the impact of different types of hints, including high-level hints vs. specific bottom-out hints. Our research takes a step in filling this gap, through an experimental study with an intelligent tutoring system we implemented in the domain of algebra (N = 50). We did not find evidence that the type of hint, high level vs. bottom out, influenced learning, with both types of hints producing similar outcomes. We did, however, find support for the conclusion that the number of hints accessed interacted with the type of hint to influence learning, and specifically, that accessing more hints was correlated with learning but only in the high-level hint condition.

Keywords: Intelligent Tutoring Systems; high-level and bottom-out hints

Introduction

There is established evidence that instructional feedback and assistance, such as hints and explanations during instructional activities, influence student learning (Shute, 2008). An open question, however, is how explicit should this assistance be to facilitate learning?

Prior research suggests that students learn best when they engage in constructive behaviors as compared to ones that are merely active or passive. This is a key prediction made and confirmed by Chi's (2009) ICAP framework that distinguishes levels of student engagement during instructional activities. To illustrate in the context of human tutoring, when a tutor prompts their student with general suggestions and/or questions, this encourages the student to generate substantive contributions, namely domain-related utterances that are positively associated with learning (Chi, Roy, & Hausmann, 2008). As another example, Ferreira, Moore, and Mellish (2007) compared two common strategies human tutors used to respond to student errors and misconceptions, namely *giving-answer assistance* and *prompting-answer* assistance. They found that *giving-answer* type of assistance occurred more often, but *prompting-answer* type of assistance was more effective for learning. Thus, in the context of human tutoring, tutors don't encourage constructive student processing because they provide the answer instead of eliciting it from the student. When students are working on their own without a tutor, they also default to passive strategies. For instance, VanLehn (1991, 1998) showed that when students were given access to worked-out examples during paper and pencil problem-solving activities, they commonly missed learning opportunities because they copied from the examples rather than trying to generate the problem solution without the help of the example.

The findings on learning from human tutoring and related activities have influenced the design of educational technologies, including that of *tutoring systems*. These technologies rely on artificial intelligence techniques to personalize instruction, in some cases approaching the effectiveness of human tutors (Vanlehn, 2011). Based on research that students benefit from active processing and that reduced assistance may promote it, some work has examined the effects of manipulating assistance in computer tutors. For instance, in separate experiments, Borracci, Gauthier, Jennings, Sale, and Muldner (2019) and Lee, Betts, and Anderson (2015) found that students learn better from tutoring systems that provide reduced assistance as compared to high assistance. In these studies, assistance was operationalized through examples that aided problem solving, with the level of similarity between an example and its corresponding problem determining how much assistance the example provided (high similarity resulted in high assistance, low similarity in reduced assistance). We next review tutoring systems that provide assistance through hints.

A common way to integrate hints into tutoring systems is to use a specific progression of assistance, one that starts off general with hints that provide high-level suggestions, but that become more specific as students ask for more help (Arroyo, Mehranian, & Woolf, 2010; Roll, Aleven, McLaren, Ryu, Baker, & Koedinger 2006; Vanlehn, Lynch, Schulze, Shapiro, Shelby, Taylor, Treacy, Weinstein, &

Level 1: *Check your trigonometry* **Level 2:** *If you are trying to calculate the component of a vector along an axis, here is a general formula that will always work: Let qV be the angle as you move counterclockwise from the horizontal to the vector. Let qx be the rotation of the x- axis from the horizontal. (qV and qx appear in the Variables window.) Then: V_x = V*cos(qV-qx) and V_y = V*sin(qV-qx).* **Level 3: (bottom-out hint)** *Replace cos(20°) with sin(20°)*	**Level 1:** *Enter the value of the radius of circle A* **Level 2**: *How can you calculate the value of the radius of circle A given the value of the diameter of circle A*? **Level 3**: *The radius of a circle is half of the diameter* **Level 4: (bottom out hint):** *The radius of circle A 1⁄4 46.5*

Figure 1. The hint progression sequence in two established tutoring systems: Andes (left) and the Cognitive Geometry tutor (right)

Wintersgill, 2005). The final hint in the progression is commonly referred to as a *bottom-out* hint, and this type of hint essentially provides the answer (e.g., the solution step the student needs to produce to make progress in solving the problem). To illustrate, Figure 1 shows an example of such a hint progression from two established tutoring systems: (1) the *Andes* tutor in the domain of physics (Vanlehn et al., 2005) and (2) the *Cognitive Geometry* tutor (Aleven, McLaren, Roll, & Koedinger, 2006).

The rationale behind using a hint progression that starts with high-level hints is to encourage students to be constructive and so generate the answer with minimal assistance from the high-level hint; if students continue asking for help, they are given more specific assistance. While this type of design mirrors what *expert* human tutors do (i.e., start off more general in their assistance and only provide the answer if students are truly stuck), in the context of tutoring systems students often abuse help functionalities (Aleven et al., 2006; Muldner, Burleson, Van de Sande, & VanLehn, 2011; Peters, Arroyo, Burleson, Woolf, & Muldner, 2018), a behavior referred to as *gaming* (Baker, Corbett, Koedinger, & Wagner, 2004). In the context of systems that make hints available, students who "game" tend to quickly and repeatedly ask the tutoring system for a hint, without reading the high-level hints, until they reach the bottom-out hint in the hint progression, at which point they copy the answer the hint provides into the problem they are working on.

Skipping high-level hints in tutoring systems is a well-documented event (e.g., Arroyo et al., 2010, Muldner et al., 2011). How does this behavior impact learning? Some argue that students still learn because they use the bottom-out hints as worked examples, which may promote learning in ways that abstract hints do not (Shih, Koedinger & Scheines, 2011). This conclusion was reached through a data mining analysis. Others have found more mixed findings on the utility of either type of hint. To illustrate, Muldner et al. (2011) used exploratory methods corresponding to Bayesian parameter machine learning to investigate the utility of high-level and bottom-out hints. Specifically, to model learning from hints, a knowledge-tracing Bayesian network was used that included nodes representing student actions, knowledge of domain principles (rules), and hints. The network encoded the probability that students will learn a rule given that they saw a certain type of hint (high-level vs. bottom-out). To obtain those probabilities, machine learning was applied to learn the parameters from data corresponding to students interacting with the Andes tutoring system. The findings showed that neither type of hint was very effective at promoting learning and there was little difference between the two types of hints. Specifically, the probability of a rule being learned was only at about 25% when a hint was used and this value was similar for both bottom-out and high-level hints.

The work cited above used exploratory methods to investigate the utility of different types of hints. The motivation for the present study is that to date there is very little experimental work comparing the effect of different types of hints on student learning. One exception is the study by Chi et al. (2001), albeit this work involves human rather than artificial tutors. Specifically, Chi et al. (2001) manipulated the type of hint human tutors were allowed to give: high-level prompts only vs. detailed hints. The results indicated a lack of a difference in learning between the two conditions, with similar posttest scores. In contrast to this study, our work investigates the effect of different types of hints provided by a computer tutor, as we now describe.

The Present Study

To test how different types of hints influence learning from a tutoring system, we created a computer tutor using the Cognitive Tutor Authoring Tools (CTAT) framework (Aleven, McLaren, Sewall, & Koedinger, 2006). CTAT facilitates the construction of tutoring systems by providing tools that a human author uses to create the tutor interface and specify the tutor's behavior. For the latter, a human author creates a *behavior graph* for each problem that specifies the tutor's behavior for that problem (e.g., what kinds of hints to show, what feedback to provide on solution entries, what to do if a student wants to move on to the next problem).

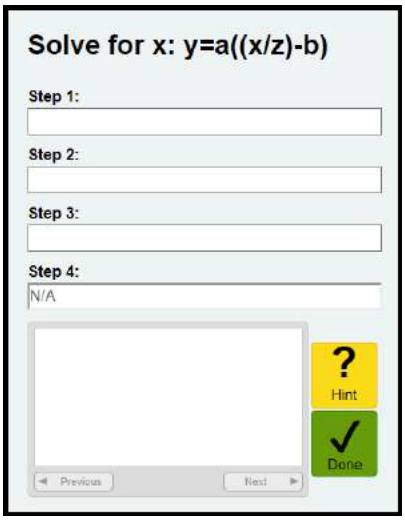


Figure 2. A problem in the algebra tutor

The algebra hint tutor

Our tutor provided students with problems to solve in the target domain of algebra (see Figure 2). The problem format was adopted from prior work and used variables instead of numeric constants (e.g., similar to the approach used by Cooper and Sweller, 1987).

All problems in the tutor required three to four solution steps for the final solution; each step was produced by applying algebraic manipulations (i.e., rules) to the prior step (or specification if the current step was the first one). A single algebraic manipulation corresponded to moving a variable from one side of the equation to the other. For example, given the equation $y = (a+x)b$, a manipulation required to solve the problem for the variable x involves moving the b variable to the other side of the equation, resulting in the equation $y/b = a+x$. Each solution step had its own input box in the tutor's interface that the student could type into. The tutor provided two forms of support: (1) feedback for correctness and (2) on-demand hints.

Feedback for correctness was realized by having the tutor color a student's entry as red (incorrect entry) or green (correct entry) directly after students indicated they were done with the entry by hitting the *return* key. The tutor was flexible in terms of accepting various forms of solutions, e.g., recognized $x = yab$ as equivalent to $x = a *y(b)$. This flexibility was accomplished through functionality we added to the tutor following the algorithm proposed by Shapiro (2005). This algorithm involves using mathematical calculations to check for equivalence without pre-storing all possible versions of a solution, thus saving significant development effort as well as computational cost of evaluating student solutions. To further scaffold the solution entry process, the solution steps had to be entered in the order required by the algebraic process and steps could not be skipped. Once a problem was done (all steps were correctly generated), students clicked the *Done* button (see Figure 1) to move on to the next problem.

Table 1
Examples of hints used in the algebra tutor

Hint Type	Example
Bottom-out Hint	Enter $ya=(x/z)+b$ into the highlighted field.
High-Level Hint (Level 1)	x isn't isolated (alone on one side of the equal sign). So we must reverse the operations acting on the variable(s), starting with the outermost ones (i.e. a in $y=xa$)
High-Level Hint (Level 2)	For this step, you need to move b to the opposite side of the question using addition.

As they were solving problems, students could ask for a hint, done by clicking on the *Hint* button in the interface (see Figure 1). We created two different versions of the tutor: one version provided only *bottom-out* hints and the other provided only *high-level* hints. To design the wording of the hints, we consulted existing tutoring systems as well as online educational sites specific to algebra. To check the wording of the hints was appropriate, we conducted several rounds of pilots.

Bottom out hints Bottom-out hints told students the exact equation they had to enter (see row 1, Table 1), and thus provided high assistance to problem solving. These hints were context specific, meaning that if the student entered a part of the solution and then asked for a hint, the hint would correspond to the next step they had to enter.

High-level hints High-level hints provided reduced assistance because they only prompted the student without giving the answer away. There were two levels of this type of hint: level 1 prompted the student about the next goal they needed to fulfill, but in contrast to a bottom-out hint did not specify exactly how to do that (see row 2, Table 1). If the student wanted further help, they could click the hint button again to access a level 2 hint. This type of hint specified the required operation and the variable that would be moved as a result (see row 3, Table 1).

Like the bottom-out hints, the high-level hints were context specific, and tailored to the student's problem-solving progress. For example, if the next step that had to be entered corresponded to the equation $y-a = (x+b)/c$, the hint would tell the student to move the variable c over to the other side of the equation using multiplication. In instances where two different manipulations were possible, the tutor would pick one at random (students could enter the steps in whichever order they wished). To avoid the hints sounding repetitive, we created several variations of each and the tutor cycled randomly through these variations. We chose to have the high level hints include prompts for both the variable and the operation because both were integral to the solving the problem.

We did not include bottom-out hints in the high-level hint tutor version because we wanted to investigate whether general prompts alone would be sufficient to foster learning.

Each of the two versions of the algebra tutor were populated with the same 12 algebra problems (all required 3-4 steps for their solutions, of the type shown in Figure 2). Both tutor versions logged all student actions in the tutor. We used a basic python script to extract the salient information from the log files (e.g., number of hints, number of errors).

Participants

The participants (N = 50) were undergraduate students at a Canadian University recruited via Sona and compensated with course credit. To be eligible for the study, participants could not have taken or be currently enrolled in any university-level math courses.

Materials

To assess algebra knowledge, we used a paper and pencil algebra pretest and posttest from our prior research that included 11 questions (Borracci et al., 2019). The tests were equivalent (only variable names were changed between them). The tests were scored out of 40, with the points for a given question corresponding to the number of rule applications needed for the question's canonical solution. For instance, if a question required three rule applications for its solution, its point value was three. This scoring method is more sensitive than marking a question as correct or incorrect, given that each question required multiple rule applications.

Several other questionnaires were used in the study to measure personality traits but we do not describe them as we do not include analysis from their data here.

Design

We used a between-subjects design with two conditions: *high-level hints* (participants used the version of the alebra tutor that included only high-level hints) and *bottom-out hints* (participants used the version of the tutor that included only bottom-out hints). As noted above, the problems solved in both conditions were identical, and the only difference between the two conditions was the type of hint available in the tutor.

Procedure

Each session was conducted individually and lasted approximately 90 minutes (the duration varied slightly based upon the amount of time participants spent on the various components). The procedure for the two conditions was the same.

Participants first completed the algebra pretest (they had up to 20 minutes to do so). They then filled in a demographics questionnaire and were assigned to their condition. Participants initially were assigned to a given condition in a round robin fashion; after about 10 participants, we began using a matching procedure based on pretest score with the goal of equalizing pretest scores between the two conditions, while maintaining similar sample size between the two conditions[1]. The experimenter then introduced participants to the algebra tutor, and explained its various features (e.g., that feedback for correctness was provided, and that all solution steps had to be correctly generated for a given problem before moving on to the next problem). Participants were told to treat this part of the study as if it were a homework situation: they had some problems to solve and were doing so to prepare for an upcoming test. Once participants confirmed they understand how to use the tutor, they were given 40 minutes to complete the 12 problems in their respective tutor version. Participants then completed the algebra posttest (20 minutes), and the personality questionnaire (10 minutes).

Results

The analysis is based on 47 participants. We excluded from the analysis three participants who were at ceiling on pretest, i.e., 95% or higher.

Does type of hint influence learning?

The descriptives for the pretest and posttest are in Table 1. Before checking if the type of hints influenced how much students learned from pretest to posttest, we verified there was no significant difference in pretest scores between the two conditions – this was the case (p = .24).

A between-subjects ANCOVA with *pretest* as the covariate, *posttest* as the dependent variable, and *condition* (high-level hints, bottom-out hints) as the independent variable did not find a significant effect of condition, $F(1, 44) = .1$, $p = .75$ and the effect size was very small, $\eta_p^2 < .01$. As shown in Figure 3, the mean posttest scores adjusted by the pretest through the ANCOVA were very similar in the two conditions. There was also no significant effect of condition on performance as measured by the number of errors made during problem solving (we extracted this information from the log files). Specifically, as expected on average participants made more errors with high-level hints, M = 29.2, SD = 20.1, as compared to with bottom-out hints, M = 23.0, SD = 15.7, but this difference was not significant, $t(45) = 1.2$, $p = .25$.

Thus, we did not find evidence that the type of hint provided influenced either learning from the algebra tutor or overall performance. However, it may be the case that that the number of hints participants accessed influenced learning differently depending on the condition. The next analysis investigates this possibility.

[1] The pretests were graded during the experimental session but in a separate room to avoid making participants uncomfortable. To save time, we used a coarser grading scheme than for the present analysis (where each question was assigned one point it was fully correct and zero points otherwise).

Table 1
Descriptive statistics for each condition

	Bottom-out hints (n = 24)		High-level hints (n = 23)	
	M	*SD*	*M*	*SD*
Pretest (/40)	10.8	14.2	15.7	14.3
Posttest (/40)	26.2	11.3	28.1	9.5

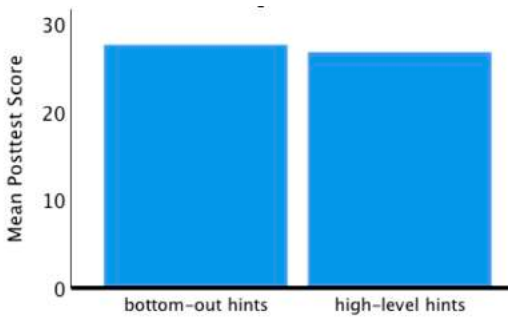


Figure 3. Posttest scores in the two conditions (adjusted by the pretest covariate); posttest was out of 40

What is the relationship between number of hints accessed and learning in each condition?

On average, participants requested more hints in the bottom-out hint condition (M = 15.6, SD = 17.5) than in the high-level condition (M = 13.4, SD = 15.1). This finding is not surprising given that the bottom-out hints facilitated problem solving by telling the students precisely what to do. To get a preliminary view of how the number of hints accessed influenced learning in each condition (operationalized as posttest score – pretest score), we plotted the relationship between these two variables for each condition. As shown in Figure 4, the relationship between the number of hints accessed and learning in the high-level hint condition is positive: the more participants accessed the high-level hints, the more they learned. In contrast, the slope of the line characterizing this relationship in the bottom-out hint condition is almost flat, suggesting there is little association between learning and number of bottom-out hints accessed.

We formalized this analysis by conducting a regression. In preparation, we dummy coded the condition variable so that the bottom-out hint condition was assigned the value 0 and the high-level hint condition the value 1 (the choice of which variable to assign the value 1 is arbitrary and does not impact the results). We proceeded with the regression by entering *posttest* as the outcome variable, and the following four predictors: *pretest*, *condition*, *number-of-hints requested*, and *number-of-hints requested x condition*.

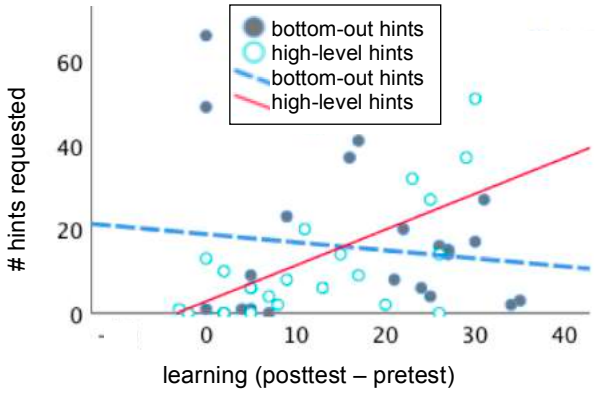


Figure 4. Relationship between number of hints accessed and learning for each condition

The overall model we obtained, shown in Table 2, was significant, $F(4, 42) = 11.3$, $p < .001$, $R^2 = .52$. Of primary interest is the interaction term (i.e., *number-of-hints x condition*), which informs on whether condition influenced the impact of number of hints requested on posttest score. Since the interpretation of the other coefficients is affected by the interaction term (Braumoeller, 2004), which essentially renders them "baseline" slopes (Grace-Martin, 2000), they are not discussed here. The interaction is modest but significant and indicates that overall, the number of hints accessed had a stronger positive relationship with posttest for high-level hints, as compared bottom-out hints. This conclusion is based on the fact that the coefficient for the interaction term is positive, indicating that when students were given high-level hints (recall this was dummy-coded as 1), their posttest score increased by the corresponding amount, controlling for the influence of the other predictors.

Table 2
Linear regression coefficients

Predictors	*B*	*β*	*t*	*p*
# hints x condition	**.34**	**.4**	**2.3**	**.022**
# hints	-.37	-.57	3.6	.001
condition	-5.4	-.26	1.7	.091
pretest	.32	.43	3.2	.003

B = Unstandardized Coefficients
β = Standardized Coefficients

Do high-level hints promote more active processing than bottom-out hints?

High-level hints offer reduced assistance because they don't tell the student the answer directly. Thus, these types of hints should promote more constructive processing on the part of the student. One way to check this is to analyze the amount of time students spent on a solution entry after they saw a high-level hint and compare that to the other

condition in which students only saw bottom-out hints. Students spent longer generating a solution entry after seeing a hint in the high-level hint condition (M = 18.0 sec, SD = 6.1) than after seeing a hint in the bottom-out hint condition (M = 21.2 sec, SD = 7.1). This trend did not reach significance but approached it after controlling for pretest score, $F(2, 34) = 2.8$, $p = .1$, $\eta_p^2 = .08$. While this result is somewhat expected as the high-level hints provided less information, it does open up the possibility that students in the bottom-out condition were not actively processing the contents of the hint before asking the tutor to check their answer (i.e., by pressing the enter button as soon as they finished entering the solution step).

Another way to check if hints are influencing student processing is to analyze how long students waited after entering a solution step (and receiving feedback on it) before pressing the hint button. If there are differences between conditions, this may suggest different levels of processing taking place for each group. Note that the alternative action after entering a solution step is to enter another solution step – here we focus on the subset of actions after a solution entry pertaining to hints only because are interested in conditional effects of hints. When students requested a hint after generating a solution entry, they waited significantly longer to do so in the high-level condition (M = 17.6 sec, SD = 23.9) as compared to the bottom-out condition (M = 6.3 sec, SD = 3.5), F (2, 37) = 5.2, p = .029, η_p^2 = .12 (controlling for pretest does not affect this result). The large standard deviation in the high-level condition implies there is a lot of variability in this condition. To ensure extreme values were not affecting the result, we removed 3 outliers flagged by SPSS and re-ran the analysis. The results remained significant and so the outliers were not influential.

Discussion

The present study investigated the utility of two types of hints in the context of a tutoring system: bottom-out hints that told students exactly what step was needed to proceed with problem solving, versus high-level abstract hints that merely suggested at what was needed to generate the corresponding problem solution step. Thus, the two types of hints provided high vs. reduced assistance to problem solving, respectively. We did not find evidence that either type of hint had a differential impact on learning and in fact the learning outcomes were very similar between the two conditions. While we recognize that conclusions can not be drawn from non-significant results, these findings echo prior experimental results (e.g., Chi et al., 2011). Our findings also echo exploratory studies using machine learning to investigate student learning from different types of hints (Muldner et al., 2011) - this latter work also did not find a difference in learning from the two types of hints.

If high-level hints are not more effective for learning, are they a less *efficient* instructional tool because they take longer to process and thus increase time on task? When we checked total time spent in each condition, we did find a trend that students overall took longer in the high-level hint condition (while this did not reach significance, that may be due to lack of power given the high variability). If high-level hints do not produce more learning than bottom out-hints but are less efficient, then that is an argument for not using them. Our subsequent analysis, however, suggested a more nuanced view of each type of hint's impact, where the number of hints students accessed interacted with the type of hint available to influence learning. It may be that students benefited from both types of hints, but that if they accessed too many bottom-out hints, they failed to learn effectively because they could not resist passively copying from the hints. Prior research in example-based learning found this type of pattern, with students copying indiscriminately from examples (VanLehn, 1998). In contrast to bottom-out hints that promote more passive cognitive processing, high-level hints in general may encourage learning because they promote active processing of the hint content, needed to infer the additional information not provided by the hint. While we did not find strong evidence in this regard, we found some indications: (1) the number of hints accessed was positively associated with learning in the high-level hint condition, and (2) students waited longer in the high-level condition to request a hint, suggesting they were less reliant on assistance provided by the tutoring system and thus more constructive. Promoting constructive processing is generally important, but may be especially challenging to realize when students are interacting with tutoring systems rather than human tutors due to accountability (i.e., students may feel less accountable with technologies than humans), although this conjecture awaits validation through future studies.

A limitation of our study is that we only measured short-term learning. High-level hints require students to process the material, possibly using common-sense or overly general reasoning to infer new rules (Vanlehn, 1991). The benefit of these types of hints may not show up until some time has passed, and so a delayed post-test would be beneficial to include in future studies to measure retention in each condition. Another limitation is the modest sample size, highlighting the need for replication. In general, given the relatively little research on what types of hints best promote learning in tutoring systems, more work is needed to validate and extend our findings.

Acknowledgements

This work was supported with an NSERC Discovery Grant #1507 and a NSERC USRA grant.

References

Aleven, V., McLaren, B., Roll, I., & Koedinger, K. (2006). Toward Meta-cognitive Tutoring: A Model of Help Seeking with a Cognitive Tutor. *International Journal of Artificial Intelligence in Education, 16*(2), 101-128.

Aleven, V., McLaren, B., Sewall, J., van Velsen, M., Popescu, O., Demi, S., ... & Koedinger, K. R. (2016).

Example-tracing tutors: intelligent tutor development for non-programmers. *International Journal of Artificial Intelligence in Education,* 26(1), 224-269.

Arroyo, I., Mehranian, H., Woolf, B. (2010) Effort-based Tutoring: An Empirical Approach to Intelligent Tutoring. In *Proceedings of the 3rd International Conference on Educational Data Mining*, 10 pages.

Baker, R., Corbett, Koedinger, K., & Wagner, A. (2004). Off-Task Behavior in the Cognitive Tutor Classroom: When Students "Game The System". In *Proceedings of the ACM CHI 2004: Computer-Human Interaction (CHI'04)*, 383- 390.

Borracci, G., Gauthier, E., Jennings, J., Sale, K., & Muldner, K. (2019). The Effect of Assistance on Learning and Affect in an Algebra Tutor. *Journal of Educational Computing Research*.

Braumoeller, B. F. (2004). Hypothesis testing and multiplicative interaction terms. *International Organization,* 58, 807–820.

Chi, M. T. H., Siler, S., Jeong, H., Yamauchi, T., & Hausmann, R. (2001). Learning from human tutoring. *Cognitive Science*, 25, 471–533.

Chi, M. T. H., Roy, M., & Hausmann, R. G. (2008). Observing tutorial dialogues collaboratively: Insights about human tutoring effectiveness from vicarious learning. *Cognitive Science*, 32(2), 301–341.

Chi, M. T. H. (2009). Active-constructive-interactive: A conceptual framework for differentiating learning activities. *Topics in Cognitive Science*, 1(1), 73–105.

Cooper, G. & Sweller, J. (1987) Effects of schema acquisition and rule automation on mathematical problem-solving transfer. *Journal of Educational Psychology*, 79(4), 347–362.

Ferreira, A., Moore, J., & Mellish, C. (2007). A study of feedback strategies in foreign language classrooms and tutorials with implications for intelligent computer-assisted language learning systems. *International Journal of Artificial Intelligence in Education, 17*(4), 389-422.

Grace-Martin, K. (2000). Interpreting interactions in regression. Retrieved from http://www.cscu.cornell.edu/ news/statnews/ stnews40.pdf

Lee, H. S., Betts, S., & Anderson, J. R. (2015). Not taking the easy road: When similarity hurts learning. *Memory & Cognition*, 43(6), 939–952.

Muldner, K., Burleson, W., Van de Sande, B., & VanLehn, K. (2011). An analysis of students' gaming behaviors in an intelligent tutoring system: predictors and impacts. *User modeling and user-adapted interaction,* 21(1), 99-135

Peters, C., Arroyo, I., Burleson, W., Woolf, B, & Muldner, K. (2018). Predictors and outcomes of gaming in an intelligent tutoring system. In *Proceedings of the Conference on Intelligent Tutoring Systems,* 366-372.

Roll I., Aleven V., McLaren B., Ryu E., Baker R., & Koedinger K. (2006) The Help Tutor: Does Metacognitive Feedback Improve Students' Help-Seeking Actions, Skills and Learning. In proceedings of *Intelligent Tutoring Systems Conference*, 10 pages.

Shapiro, J. (2005). An Algebra Subsystem for Diagnosing Students' Input in a Physics Tutoring System. *International Journal of Artificial Intelligence in Education,* 15(3), 205-228

Shih, B., Koedinger, K., & Scheines, R. (2011). A response time model for bottom-out hints as worked examples. *Handbook of educational data mining,* 201-212.

Shute, V. (2008). Focus on formative feedback. *Review of educational research*, 78(1), 153-189.

VanLehn, K. (1991). Rule acquisition events in the discovery of problem solving strategies. *Cognitive Science*, 15(1), 1-47.

VanLehn, K. (1998). Analogy events: How examples are used during problem solving. *Cognitive Science,* 22(3), 347-388.

Vanlehn, K., Lynch, C., Schulze, A. Shapiro, Shelby, Taylor, Treacy, Weinstein, & Wintersgill. (2005). The Andes Physics Tutoring System: Lessons Learned. *Int. Journal of Artificial Intelligence in Education*, 15(3),147-204.

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist,* 46(4), 197-221.

Are Cross-Linguistically Frequent Semantic Systems Easier to Learn? The Case of Evidentiality

Dionysia Saratsli (Dsaratsl@Udel.Edu)
Department of Linguistics and Cognitive Science, 125 Main Street
Newark, DE 19702 USA

Stefan Bartell (Sbartell@Udel.Edu)
Department of Linguistics and Cognitive Science, 125 Main Street
Newark, DE 19702 USA

Anna Papafragou (Papafragou@Psych.Udel.Edu)
Department of Psychological and Brain Sciences, 105 The Green
Newark, DE 19702 USA

Abstract

It is often assumed that cross-linguistically more prevalent distinctions are easier to learn (*Typological Prevalence Hypothesis* - TPH). Prior work supports this hypothesis in phonology, morphology and syntax but has not addressed semantics. Using an Artificial Language Learning paradigm, we explore the learnability of semantic distinctions within the domain of evidentiality (i.e. the linguistic encoding of information sources). Our results support the TPH, since the most prevalent evidential system was learned best while the most rare evidentiality system yielded the worst learnability results. Furthermore, our results indicate that, cross-linguistically, indirect information sources seem to be marked preferentially (and acquired more easily) compared to direct sources. We explain this pattern in terms of the pragmatic need to mark indirect, potentially more unreliable sources over direct sources of information.

Keywords: evidentiality; artificial language learning; learnability; semantics; information sources

Learnability and the Typological Prevalence Hypothesis (TPH)

It is often assumed in the literature that linguistic distinctions that are encountered more frequently across different languages share some characteristics that make them easier to learn than others (Jacobson, 1971; Rosch, 1972; Clark, 1976; Pinker, 1984). This idea has been captured effectively by Gentner and Bowerman's (2009, p.467) *Typological Prevalence Hypothesis* (TPH): "All else being equal, within a given domain, the more frequently a given way of categorizing is found in the languages of the world, the more natural it is for human cognizers, hence the easier it will be for children to learn". Gentner and Bowerman (2009) tested this hypothesis within the spatial domain, comparing how English-speaking and Dutch-speaking children acquire their native language's support prepositions. English and Dutch differ in the number of prepositions they use to express spatial support: Dutch utilizes three different prepositions (*op, aan, om*) to express the same meanings that English encodes with the single preposition *on*. Importantly, these two support systems differ in their typological prevalence, with the English preposition system being more typologically common. The TPH therefore predicts that the English preposition system should be more easily learned than the Dutch system. Gentner and Bowerman's results support this prediction. One issue with this conclusion, however, is that the slower acquisition rate could be due to the increased number of subcategories found in Dutch compared to English as opposed to an inherent learnability asymmetry of semantic categories per se. This language asymmetry complicates the interpretation of Gentner and Bowerman's results and hence the evidence in favor of TPH.

In this paper, we offer a new test of TPH using an Artificial Language Learning Paradigm. This type of experimental design often requires participants to learn different versions of a target language that differ minimally from each other in terms of a grammatical or lexical feature (see Folia, Uddén, de Vries, Forkstam, & Petersson, 2010 for a review). Typically, this design includes an initial learning phase in which learners are exposed to the grammar/lexicon of the artificial language, usually with the help of visual stimuli. The learning phase is followed by a test phase in which the extent to which participants learned the linguistic target is assessed. This paradigm offers a unique opportunity to explore the participants' learning process in relation to a specific linguistic feature of interest (Fedzechkina, Newport & Jaeger, 2016). By having participants learn minimally different versions of the same artificial language, one can bypass the role of frequency in the learnability of attested systems in individual languages, such that any learnability pattern that surfaces can be more directly tied to the inherent characteristics of the cross-linguistic distinction that is being explored. Moreover, it is possible to have adults learn the target artificial language which in turn eliminates the possibility that any learnability patterns observed could be due to cognitive-developmental limitations in the learners themselves.

Previous studies using an Artificial Language Learning paradigm have confirmed that cross-linguistically common distinctions are learned more easily than less common ones in the domains of syntax, phonology and morphology

(Newport & Aslin, 2004; Wonnacott, Newport, & Tanenhaus 2008; Merkx, Rastle, & Davis, 2011; Culbertson, 2012; Tabullo, Arismendi, Wainselboim, Primero, Vernis, Segura, Zanutto & Yorio., 2012; Culbertson & Newport, 2015;). Nevertheless, within the domain of semantics (which was the main focus of TPH), this hypothesis remains to be tested systematically. Here we address this open issue. We focus on a semantic domain that is not grammaticalized in English and can be taught to adults within an Artificial Language Learning paradigm without native language interference: the domain of evidentiality, i.e., the linguistic encoding of information source.

Evidentiality and TPH

Languages differ in the way they encode evidentiality: some languages like English make use of lexical means such as verbs (e.g., *see, hear, infer*) or adverbs (e.g., *allegedly, reportedly*) to mark information sources. Other languages use a set of grammatical morphemes to indicate information sources in an utterance. There are three common types of evidential morphemes depending on which information source is marked: Visual (firsthand/perceptual evidence), Inferential (inference based on evidence), and Reportative (hearsay) (Willett, 1988; Papafragou, Li, Choi & Han, 2007; deHaan, 2013b; Aikhenvald, 2018). In the Wanka Quechua examples below, *-mi* in (1) marks the speaker's direct visual experience of the event, *-chr-* in (2) marks an inference drawn by the speaker and *–shi* in (3) marks another person's report about what happened (Aikhenvald, 2004):

(1) Chay-chruu-mi achka wamla-pis walashr-pis alma-ku-lkaa-ña.
this-LOC-DIR.EV many girl-TOO boy-TOO bathe-REEL-IMPF.PL-NARR.PAST.
'Many girls and boys were swimming' (I saw them).

(2) Daañu pawa-shra-si ka-ya-n-chr-ari.
Field finish-PART-EVEN be-IMPF-3-INFR-EMPH.
'It (the field) might be completely destroyed' (I infer).

(3) Ancha-p-shi wa'a-chi-nki wamla-a-ta.
too.much-GEN-REP cry-CAUS-2 girl-1P-ACC.
'You make my daughter cry too much' (they tell me).

Across languages that grammatically mark only one type of information, evidential systems that involve only Reportative morphemes are the most widespread ones; systems that use an indirect morpheme to mark inference or reports are less frequent (Papafragou et al., 2007; deHaan, 2013a; Aikhenvald, 2004, 2018; Ünal & Papafragou, 2018;). Evidential systems that only have Visual morphemes are rare (Aikhenvald, 2018). The reasons for this asymmetry have not been discussed extensively but might be connected to the pragmatic need to mark indirect, probably unreliable sources but not direct/perceptual, and hence more reliable, experience (Dancy, 1985; and discussion below).

Here we used an Artificial Language Learning paradigm to compare the learnability of three evidential systems (see Table 1): 1) a system in which a grammatical morpheme is used only when the speaker has full direct visual access to what happened (*Visual* System), 2) a system where a grammatical morpheme is used only when the speaker infers what happened based on some visual cues (*Inferential* System), and 3) a system in which a grammatical morpheme is used only when the speaker obtains information by another person (*Reportative* System). Based on the typological frequency patterns for evidential systems reviewed earlier, the TPH predicts that the Reportative system should be the most learnable and the Visual system the least learnable (with the Inferential system falling somewhere in-between). The experiment that follows tested these predictions.

Table 1: Evidential Systems.

Evidential System	Speaker's Information Access		
	Visual Perception	Inference	Report
Visual	morpheme		
Inferential		morpheme	
Reportative			morpheme

Experiment

Our experiment consisted of two phases following the general Artificial Language Learning experimental design: a Training Phase and a Testing Phase. In the Training Phase, participants were exposed to one of the three evidentiality systems in Table 1 and had to figure out when the evidential marker was used. In the Testing Phase participants were evaluated on how well they had learned the target evidentiality system through a Production and a Comprehension Task.

Participants. We recruited 101 participants between the ages of 18 and 22. All participants were undergraduate students at the University of Delaware and were enrolled in an Introductory Psychology course that awarded credit for their participation.

Stimuli and Procedure. For the Training Phase, we filmed 21 videos in three versions each, with each version corresponding to a type of information access (Visual Perception, Inference, Report). In each video there were three characters; across videos, they were played by the same three female undergraduate research assistants. The roles of these characters were consistent across the videos: one of the characters (henceforth the "Agent") performed an event using some materials and then put these materials away. The second character accessed the event in one of several ways and would later describe the event (henceforth the "Speaker"). A third character manipulated the Speaker's access to the event (e.g., either allowed the Speaker to watch

Figure 1: Sample screenshots from one Training Phase video shown in 3 versions corresponding to Access types: (A) Visual Perception, (B) Inference, (C) Report. Across Access types the video ended with the Speaker producing a sentence (Panel 5) that either included or omitted an evidential (e.g., "She drawing copiedga", "She drawing copied").

the event or blocked her visual access for the complete duration or part of the event). The setting was identical for all the videos: the Agent and the Speaker were sitting on different sides of a table while the third character stood behind them in full view of the table. Each video was approximately 15 seconds long. At the end, the Speaker turned to the camera and described what happened. At that point, the video stopped and a speech bubble appeared with an artificial language sentence, and stayed there for 7 seconds before the next video began.

Figure 1 shows a sample event in which the Agent copied a drawing (the Speaker is pictured in a blue shirt). In the Visual Perception version (series A), the Speaker had continuous direct visual access to the event (A1 began with occluded access to ensure that the hands-over-eyes would not be an easy-to-detect difference among access types, but the hands are removed from the Speaker's face immediately). In the Inference version (series B), the Speaker had visual access only for the beginning and the end of the event (panels 1 and 4), but her access was blocked for the middle portion (panels 2 and 3); therefore, she could infer what happened from the last stage of the event. In the Report version (series C), the Speaker's visual access was blocked throughout the event (panels 1-3); later (panel 4), the Speaker got a report about what had happened from the third character. All videos ended by displaying the Speaker's artificial-language description of what happened within a speech bubble (panel 5). The artificial language shared the same vocabulary with English (for simplicity's sake) but had a different syntactic structure (Subject-Object-Verb) and lacked function words. A novel verb-final morpheme, *ga*, appeared when appropriate as a marker for evidentiality.

We designed 3 evidential systems to be acquired (Visual, Inferential, Reportative) by having the Speaker describe only one type of Access with an evidentially marked sentence (e,g, *She drawing copiedga*, as in Figure 1) and include no marker for the other two Access types. For instance, for the Visual System, only the sentences in the Visual access versions included *-ga*. Then for each evidential system, we created 3 basic lists for the Training Phase (for a total of 9 lists): each basic list contained 21 videos, with 7 videos per Access type. Across lists, the videos rotated through each Access type. For instance, if the video in Figure 1 was shown in the Visual Perception version for list 1, then the same video was shown in the Inference version for list 2 and the Report version for list 3. The presentation order of the videos was randomized across lists.

We randomly assigned participants to one of 3 conditions depending on the System they were exposed to (n = 34 for the Inferential and Reportative System, and n=33 for the Visual System). Each participant was given one of the 9 stimulus lists. We tested participants in small groups, in a dimly lit, quiet room. Participants were told that they would watch some videos and one character would describe the videos in an "alien language". This language would share some words with English but would be different in several ways and would contain a special marker, *ga*. Their task was to pay attention to when *ga* appeared in order to try and figure out what it meant.

When the Training Phase was over, the Testing Phase began. Participants had to complete both a Production and a Comprehension task. For these tasks, we filmed new videos that were similar to those for the Training Phase (except for some features of the language in the event descriptions – see below).

For the Production task, we used 12 new videos, each filmed in 3 different versions corresponding to the 3 Access types. We arranged these stimuli into 3 basic lists, with each list containing 12 videos, 4 per Access type. As in the Training Phase, the lists were created by rotating each video through the three different Access types. For each basic list, three randomized presentation orders were created, resulting in 9 presentation lists in total. Within each condition, participants were assigned to one of these lists. As mentioned already, the structure of the videos in the Production task was identical to the Training Phase but when the speech bubble appeared at the end, the evidential marker was replaced by a gap next to the verb. Using an answer sheet, participants had

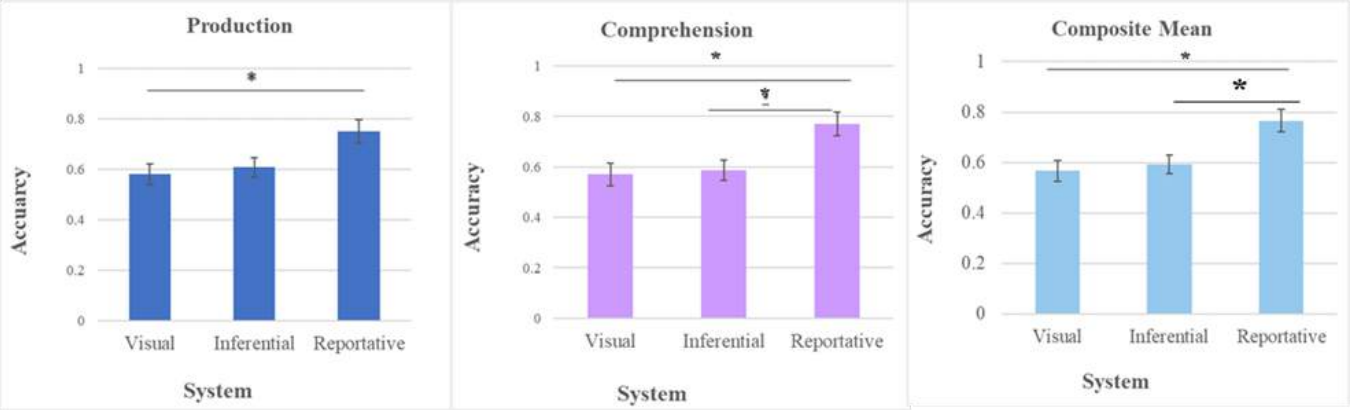


Figure 2. Accuracy Means Across Systems. The composite score represents a combined Production/Comprehension score. Error bars represent ± 1 S.E.

to write down the verb either with or without *ga* depending on whether they thought it was needed to correctly complete the character's phrase.

For the Comprehension task, we used 36 new videos, each filmed in 3 different versions corresponding to the 3 Access types. We arranged these stimuli into 3 basic lists, with each list containing 36 videos (12 per Access type) using the rotation method described above. Similarly to the Production task, for each of these basic lists, 3 lists with a unique randomized presentation order were created (9 lists in total). In half of the videos within each list (and within each Access type), the Speaker erroneously used the marker *ga*: she either failed to use the marker when she should have or used it for the wrong types of Access. In the remaining videos, the use of the marker was correct. Within each condition, participants were assigned to one of the presentation lists. The participants' task was to write 'yes' or 'no' in their response sheet to indicate whether or not they thought the character was using the marker correctly. At the end of the experiment, we asked participants to write down what they thought that the marker *ga* meant and when it was/was not used.

Results

Participants' responses were coded for accuracy. We calculated the accuracy means for each System. In addition, we averaged each participant's Production and Comprehension score yielding a Composite accuracy mean across tasks. We subsequently calculated a Composite Mean per System. The results can be seen in Figure 2.

For the Production task, a one-way ANOVA with System as a factor revealed a main effect of System ($F(2,98)= 4.771$, $p<0.05$). Pairwise comparisons using Bonferroni corrections revealed a significant advantage of the Reportative over the Visual System ($p=.014$) but no significant difference between either the Inferential and the Visual System ($p=1.0$), or the Inferential and the Reported System ($p=.058$).

For the Comprehension task, the same ANOVA revealed a main effect of System ($F(2,98)=6.509$, $p<0.01$). Pairwise comparisons (Bonferroni corrections) showed an advantage of the Reportative System over both the Inferential ($p=0.01$) and the Visual System ($p=.005$). However, there was no statistically significant difference between the Visual and Inferential System ($p=1.0$).

Lastly, a one-way ANOVA conducted on composite Production and Comprehension means revealed an effect of System ($F(2,98)=6.535$, $p<0.01$). Pairwise comparisons (Bonferroni corrections) revealed again a significant advantage of the Reportative System over both the Visual ($p=.004$) and the Inferential System ($p=0.01$).

Participants' answers about the meaning of the marker reflect the results' pattern: out of the 34 participants exposed to the Reportative System, 21 correctly associated the marker with reportative access, specifically alluding to the *speaker's* mental state by mentioning that she "was told" about the event. Of the 33 participants of the Visual system, only 12 associated the marker with speaker's direct visual experience of the event. only Similarly, only 9 out of the 34 participants exposed to the Inferential System correctly associated the marker with the character inferring the action. Across systems, participants that did not identify the correct marker meaning, associated the marker with some type of grammatical distinction (e.g., singular/plural forms, past or completed actions, articles such as *the*/*a*) or associated it with the incorrect type of access. Overall, these responses show that participants associated evidential meanings with the marker, but they did so much more consistently for the Reportative System.

Discussion

Our goal was to test the assumption that the frequency of cross-linguistic semantic patterns is related to the inherent learnability of these patterns, an assumption captured in Gentner and Bowerman's (2009) TPH. Using an Artificial Language Learning paradigm, we set out to compare the learnability of evidential semantic systems, focusing on those that encode a single type of information source (Table 1). The most typologically common evidential system within this group (and also the single most prevalent type of evidential system in general; Aikhenvald, 2018) is the Reportative system in which a marker is used only for the least direct type

of access to information – namely, the cases when the speaker conveys information reported by another person. The least common system is the Visual system in which only direct visual access to an event is marked morphologically. In our study, as predicted by TPH, the Reportative system was learned more easily by our participants compared to the Visual system. Our experiment offers strong evidence for the conclusion that highly frequent semantic distinctions are more learnable than less frequent ones. Furthermore, it adds to previous studies that have studied learnability with the same methodological paradigm within the domains of syntax, phonology and morphology.

Not all aspects of our data are compatible with the predictions of TPH. Specifically, even though exclusive encoding of visual evidentials is rare, and there is a broad preference to mark non-visual/indirect over visual/direct sources cross-linguistically (Aikhenvald, 2018), the Inferential and Visual systems were equally learnable in our data. A possible explanation for this outcome lies with the fact that our Inference videos contained strong visual clues to what happened, bringing this type of information access closer to a direct perceptual experience than to an indirect inference on the speaker's part. This explanation is in line with several findings from a recent study by Ünal, Pinto, Bunger and Papafragou (2016). In that study, when English speakers had to state how they had found out about an event, they stated having seen events that they had experienced in their entirety. However, when they had only seen the beginning and aftermath of an event and had to "fill in" the event from these visual cues, their statements varied. Closer inspection suggested that, when the visual cues were indeterminate, participants consistently stated that they had inferred the event; but when the visual cues were more determinate and highly constrained the inference, participants were equally likely to say that they had seen vs. inferred the event. The authors proposed that there are several varieties of inference, and that stronger, more constrained (and thus more secure) inferences from visual cues might be difficult to distinguish from purely perceptual experience. These varieties of inference had implications for evidential language: Ünal et al. (2016) found that these different types of inference impacted the use of evidential morphology by speakers of Turkish, a language that grammaticalizes evidentiality. Furthermore, inference types had effects on memory: building on classic studies showing that people often have a false memory of having actually experienced events that they have only inferred (Johnson, Hashtroudi, & Lindsay, 1993; Hannigan & Reinitz, 2001; cf. Strickland & Keil, 2011), Ünal et al. (2016) found that, across English and Turkish speakers, such misattributions to perception were more common when inferences were strongly constrained by visual cues and thus harder to distinguish from pure perception. This line of reasoning leads to the prediction that replacing Inference scenarios in our paradigm with less direct cases of inference from visual cues (e.g., footsteps on snow) should allow the learnability difference between the Visual and Inferential systems to emerge.

On a broader level, our results raise questions about the origins of the typological generalizations in the domain of evidentiality. According to the basic observation motivating the present work, across languages, the least formally marked source of information is visual, or direct access (Aikhenvald, 2018, a.o.). Why should this be so? One possibility is that "the tendency to mark direct, or visual, or sensory evidentials less than others may reflect the primacy of vision as an information source" (Aikhenvald, 2018, p.16). Direct perceptual experience of an event is regarded as a very reliable source because it is assumed to correspond to reality (Dancy, 1985). Additionally, developmental research suggests that children draw the connection between seeing and knowing from early on (Pillow, 1989; Pratt & Bryant, 1990; Ozturk & Papafragou, 2016), which highlights the the primacy of visual perception as an information source. Relatedly, indirect sources of information such as inference or reports are deemed more peripheral and less reliable in the sense that the former may be based on incomplete premises while the latter depends on the informant's reliability (Dancy, 1985; Koring & De Mulder, 2014; Papafragou et al., 2007; Matsui & Fitneva, 2009; McCready, 2015; Aikhenvald, 2018; Wiemer, 2018;). This has been found to be true even for languages that do express information access through *perception verbs* and not through obligatory grammatical morphemes (Lesage, Ramlakhan, Toivonen & Wildman, 2015). According to some researchers (Sperber, Clement, Heintz, Mascaro, Mercier, Origgi & Wilson, 2010), human cognition uses epistemic vigilance as a mechanism to avoid unreliable sources and the risk of being misinformed. However, exercising epistemic vigilance could entail an additional processing cost: listeners would have to give up the assumption that the communicative exchange they are engaged in offers truthful, informative contributions and would need to evaluate not only the actual information they receive but also their interlocutor's reliability and intentions. Thus, pragmatic pressures to mark sources of information would affect indirect and probably unreliable sources more than direct/perceptual, and hence more reliable, experience.

If this perspective is on the right track, our results would support a more nuanced version of TPH. Recall that, on Gentner and Bowerman's original proposal, the roots of TPH lie in the cognitive naturalness of the semantic classes that the learner acquires. Here we have proposed a broadened notion of naturalness that also includes pragmatic (and not only conceptual) factors. In our studies, adult learners acquired semantic systems of varying cross-linguistic frequency but both the frequency patterns and the learnability outcomes were pragmatically (not conceptually) motivated.

If the frequency patterns for linguistic evidentiality systems reflect the pragmatic need for information source marking, as we have suggested, a further prediction follows: it might be possible to obtain similar learnability patterns even if we used a non-linguistic marker to encode information source (e.g., a pictorial symbol). We are currently pursuing this possibility in ongoing work.

Acknowledgements

This work was supported by a Leventis Foundation Graduate Educational Grant (Dionysia Saratsli) and NSF Grant #1632849 (Anna Papafragou). We would like to thank the undergraduate research assistant Casey Corallo for her help in collecting and coding data for this project.

References

Aikhenvald, A. (2004). *Evidentiality*. Oxford University Press.

Aikhenvald, A. (2018). Evidentiality: The Framework. In A.Y. Aikhenvald (Ed.), *The Oxford Handbook of Evidentiality*, 1-43. Oxford: Oxford University Press.

Clark, E. (1976). Universal categories: On the semantics of classifiers and children's early word meanings. In A. Juilland (Ed.), *Linguistic studies offered to Joseph Greenberg on the occasion of his sixtieth birthday*, Volume 3: Syntax, 449–462. Saratoga, CA: Anna Libri.

Culbertson, J. (2012). Typological Universals as Reflections of Biased Learning: Evidence from Artificial Language Learning. *Linguistics and Language Compass*, 6(5), 310–329. https://doi.org/10.1002/lnc3.338.

Culbertson, J., & Newport, E. (2015). Harmonic biases in child learners: In support of language universals. *Cognition*, 139, 71–82.

Dancy, J. (1985). *An introduction to contemporary epistemology*. Oxford: Blackwell.

de Haan, F. (2013a). Coding of evidentiality. In M.,Dryer and M.,Haspelmath, (Eds.) *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

de Haan, F. (2013b). Semantic distinctions of evidentiality. In Dryer, M. and Haspelmath, M. (Eds.), *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Fedzechkina, M., & Newport, E. (2016). Miniature artificial language learning as a complement to typological data. *The Usage-Based Study of Language Learning and Multilingualism*, 1–22. Retrieved from https://books.google.com/books?hl=en&lr=&id=88gvDAAAQBAJ&oi=fnd&pg=PA211&ots=I6oc-Bwvl-&sig=SORG26sQudO-ZohoKyOHqtzAUo0.

Folia, V., Uddén, J., de Vries, M. H., Forkstam, C., & Petersson, K. M. (2010). Artificial language learning in adults and children. *Language Learning*, 60, 188–220.

Gentner, D., & Bowerman, M. (2009). Why some spatial semantic categories are harder to learn than others: The typological prevalence hypothesis. In J. Guo et al., (Eds.), *Crosslinguistic approaches to the psychology of language: Research in the tradition of Dan Isaac Slobin*, 465–480. Hillsdale, NJ: Erlbaum.

Hannigan, S. L., & Reinitz, M. T. (2001). A demonstration and comparison of two types of inference-based memory errors. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27, 931–940. http://dx.doi.org/10.1037//0278-7393.27.4.931.

Jakobson, R. (1971). *Studies on child language and aphasia*. The Hague/Paris: Mouton.

Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin*, 114, 3–28. http://dx.doi.org/10.1037/0033- 2909.114.1.3.

Koring, L., & De Mulder, H. (2014). Understanding different sources of information: the acquisition of evidentiality. *Journal of Child Language*, 947–968. https://doi.org/10.1017/S030500091400052X.

Lesage, Claire & Ramlakhan, Nalini & Toivonen, Ida & Wildman, Chris. (2015). The reliability of testimony and perception: connecting epistemology and linguistic evidentiality. *Proceedings from the 37th Annual Meeting of the Cognitive Science Society*.

Matsui, T. & Fitneva, S. A. (2009). Knowing how we know: evidentiality and cognitive development. In S. A. Fitneva & T. Matsui (eds), *Evidentiality: a window into language and cognitive development* (New Directions for Child and Adolescent Development 125) (pp. 1–11). San Francisco: Jossey-Bass.

McCready, E. (2015). *Reliability in pragmatics*. Oxford: Oxford University Press.

Merkx, M., Rastle, K., & Davis, M. H. (2011). The acquisition of morphological knowledge investigated through artificial language learning. *Quarterly Journal of Experimental Psychology*, 64(6), 1200–1220. https://doi.org/10.1080/17470218.2010.538211.

Newport, E., & Aslin, R. (2004). Learning at a distance i. statistical learning of non-adjacent dependencies. *Cognitive Psychology*, 48(2), 127–162.

Ozturk, O., & Papafragou, A. (2016). The Acquisition of Evidentiality and Source Monitoring. *Language Learning and Development*, *12*(2), 199–230. https://doi.org/10.1080/15475441.2015.1024834.

Papafragou, A., Li, P., Choi, Y., & Han, C. (2007). Evidentiality in language and cognition. *Cognition*, 103(2), 253–299. https://doi.org/10.1016/j.cognition.2006.04.001.

Pillow, B. (1989). Early understanding of perception as a source of knowledge. Journal of *Experimental Child Psychology*, *47*, 116–129.

Pinker, S. (1984). *Language learnability and language development.* Cambridge, MA: Harvard University Press.

Pratt, C., & Bryant, P. (1990). Young Children Understand That Looking Leads to Knowing (So Long as They Are Looking into a Single Barrel). *Child Development*, *61*(4), 973-982. doi:10.2307/1130869. Rosch, E. H. (1972). Universals in color naming and memory. *Journal of Experimental Psychology*, 93(1), 10–20.

Rosch, E. H. (1972). Universals in color naming and memory. *Journal of experimental psychology*, *93*(1):10–20.

Sperber, D., Clement, F., Heintz, C., Mascaro, O., Mercier, H. Origgi G. & Wilson, D. (2010). Epistemic vigilance, *Mind & Language*, 25(4), 359–393. https://doi.org/10.1111/j.1468-0017.2010.01394.x.

Strickland, B., & Keil, F. (2011). Event completion: Event based inferences distort memory in a matter of seconds.

Cognition, 121, 409–415. http://dx.doi.org/10.1016/j.cognition.2011.04.007.

Tabullo, Á., Arismendi, M., Wainselboim, A., Primero, G., Vernis, S., Segura, E., & Yorio, A. (2012). On the learnability of frequent and infrequent word orders: An artificial language learning study. *Quarterly Journal of Experimental Psychology*,65(9),1848–1863. https://doi.org/10.1080/17470218.2012.677848.

Ünal, E., Pinto, A., Bunger, A., & Papafragou, A. (2016). Monitoring sources of event memories: A cross-linguistic investigation. *Journal of Memory and Language*, *87*, 157–176. https://doi.org/10.1016/j.jml.2015.10.009

Ünal, E., & Papafragou, A. (2018). Relations between language and cognition: Evidentiality and sources of knowledge. Special issue of *Topics in Cognitive Science* on Lexical Learning, 1-21. https://doi.org/10.1111/tops.12355.

Wiemer, B. (2018). Evidentials and epistemic modality. In A. Aikhenvald (Ed.), *The Oxford Handbook of Evidentiality*. Oxford: Oxford University Press.

Willett, T. (1988). A cross-linguistic survey of the grammaticization of evidentiality. *Studies in language*, *12*(1):51–97.

Wonnacott, E., Brown, H., & Nation, K. (2017). Skewing the evidence: The effect of input structure on child and adult learning of lexically based patterns in an artificial language. *Journal of Memory and Language*, 95, 36–48. https://doi.org/10.1016/j.jml.2017.01.005.

Not All Exceptions Are the Same: Different Memory Demands for Differentiation, Isolation and Odd-ball Exceptions

Abstract

There is an influential body of research arguing that category exceptions have a special status in memory compared to regular category members. However, the memory advantage for category exceptions has typically been demonstrated using one very specific category structure (Differentiation). Here we present a study examining whether the reported memory advantage is specific to this particular structure or whether it can be generalized to other kinds of exceptions (Isolation and Odd-ball). We compare three different types of category exceptions that have varying memory demands due to different levels of feature binding required for accurate categorization. The results suggest that only those exceptions that require binding together multiple features are remembered better than regular, rule-following items. The present work clarifies that the memory advantage for exceptions characterizes certain kinds of exceptions rather than exceptions in general.

Keywords: category exceptions; rule-plus-exception; binding requirement

Introduction

Whales are mammals. Penguins are birds. Tomatoes are fruit. Many categories include items that look different, behave differently, or lack important qualities that define the majority of members of the category. We refer to these items as *exceptions,* because they violate our expectations about the category.

Since the goal of categorization is to encode key aspects about the members of the category, it is reasonable to ask: How are exceptions, items that violate those key aspects, learned and represented?

Memory Advantage for Exceptions

There is an influential body of research arguing that category exceptions have a special status in memory. Palmeri and Nosofsky (1995) demonstrated that exceptions to a category rule are remembered better than the items that follow that rule. This initial finding of a memory advantage for category exceptions is supported by a number of subsequent category learning studies (Sakamoto & Love, 2004, 2006; Davis, Love & Preston, 2012) and found to be in accordance with the predictions of several influential models of category learning: RULEX (Nosofsky, Palmeri & McKinley, 1994) and SUSTAIN (Love, Medin, & Gureckis, 2004).

Work on this topic in the categorization literature was preceded by studies in memory (Von Restorff, 1933) and schema research (e.g. Rojahn & Pettigrew, 1992; Stangor & McMillan, 1992), where an advantage in memory for exception items has long been established. Although the approach (both in terms of methodology and primary research questions) differed between the memory and categorization literatures, the fact that these findings paralleled each other further strengthen the view that there is a general advantage in memory for information that does not fit in with salient knowledge structures.

What Makes (Some) Category Exceptions Memorable?

Although a memory advantage for category exceptions seems to be well established, the nature of the effect is not well understood. One obstacle to understanding what makes category exceptions more memorable is that previous studies focused on one very specific type of category structure. We refer here to this structure as the Differentiation case (see Figure 1).

Although exceptions are not limited to the Differentiation case, the vast majority of influential work on this topic (Palmeri & Nosofsky, 1995; Sakamoto & Love, 2004; Davis, et al., 2012) has studied this kind of exception. One reason the Differentiation structure has received so much attention is that it presented an interesting challenge for models of category learning to explain. And so, researchers often selected structures with the purpose of evaluating or comparing models of categorization, and were not necessarily concerned with representing all types of exceptions.

Thus, it remains unclear whether the reported memory advantage for category exceptions is specific to this particular structure or whether it can be generalized to other kinds of structures. In what follows, we describe how the difference in category structures may affect memory demands.

Different Memory Demands for Differentiation, Isolation and Odd-ball Exceptions

Figure 1 illustrates three different structures of exception items: Differentiation, Isolation and Odd-ball exceptions. All three types of exceptions (a) violate the category rule and (b) are dissimilar to other items in their own category. However, they differ in how much they share with the contrasting category members.

Differentiation exceptions, the most commonly used structure, are highly similar to items of the contrasting category. Not only do they follow the contrasting category rule, but they also share other features with items of the contrasting category. Due to its specific structure, this kind of exception cannot be categorized correctly on the basis of any one individual feature.

Since it is not enough to remember one or even multiple isolated features, Differentiation exception features have to be bound together and committed to memory. This is because each feature of the exception (in our case, the color, the size and the shape) has a competitor in the contrasting

category, thus making only the whole configuration (but not individual features) sufficient. For example, the past tense of the irregular verb teach (taught) is very different from that of similar sounding regular verbs (reach, breach, or preach), but is similar to that of a phonetically different verb (e.g., think).

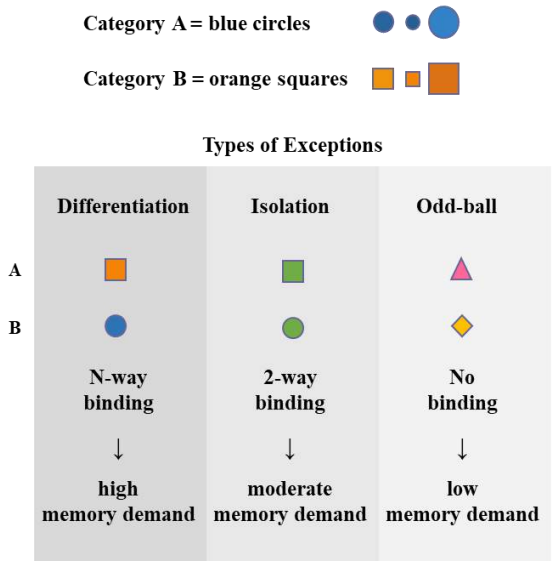


Figure 1: Three different types of exception items for categories of blue circles and orange squares. Exceptions vary in how similar they are to the contrasting category and therefore in the amount of feature binding and memory demands required for accurate categorization.

Although the Differentiation structure is commonly used in experimental studies, this kind of exception is likely quite rare in the real world. Most exceptions found outside of the lab, even the most commonly cited examples (e.g. bats as exceptions to category of mammals), in addition to shared features also have some distinctive features. Figure 1 illustrates two types of such exceptions: Isolation and Odd-ball.

Isolation exceptions follow the contrasting category rule, just like Differentiation exceptions, but, crucially, they also have distinctive features. The distinctive features make Isolation exceptions less similar to items of the contrasting category. As fewer features are shared, less complex binding is required, which reduces memory demands. In the example shown in Figure 1, the green square has the same shape (rule dimension) as members of the contrasting category. However, its unique color (different from both its own and the contrasting category) allows for the categorization problem to be solved based on binding of only two dimensions: color and shape. Analogous to this example, bats are flying creatures (characteristic of the contrasting category of birds) with membranous wings (a distinctive feature), and therefore could be represented as exceptional mammals by binding these two features: flying and membrane wings.

The third kind of exceptions shown in Figure 1 is the Oddball exception. Odd-ball exceptions do not share any features with members of contrasting category. Since all of their features are distinctive, no binding is required, and categorization can be made on the basis of any single feature. Critically, in case of the Odd-ball structure, representation of exceptions can be as simple as representation of regular items. The pink triangle in Figure 1 violates both shape and color of the category of blue circles. However, since there is no overlap on these dimensions with any items in category B, accurate categorization can be based on either its pink color or triangular shape alone. One example of such an oddball is an hourglass as an exceptional member of the category of time-keeping devices.

Present Experiments

Although all three of the different structures presented in Figure 1 represent rule-violating exceptions, they have different memory demands. Since the categorization literature has focused almost exclusively on the Differentiation case, it remains unclear whether in previous studies exceptions were remembered better (a) because they violated a salient knowledge structure (von Restorff, 1933; Hunt & Lamb, 2001; Busey & Tunnicliff, 1999; Nairne, 2006), or (b) because of the additional binding requirement resulting from the high overlap with the contrasting category (Sakamoto & Love, 2006).

In support for the latter possibility, Sakamoto and Love (2006) demonstrated that exceptions that are more similar to the contrasting category (i.e., Differentiation) are remembered better than exceptions that are more distinctive (i.e., Isolation). The stimuli in Sakamoto and Love (2006) were lines that varied in color and size and the focus of this study was on comparing different exceptions to each other. Thus, memory advantage for both kinds of exceptions over regular items was assumed although not directly tested due to the limitations of the stimuli design.

The present experiments were designed to tease apart the roles of (a) violation of a salient knowledge structure (i.e. similarity of an exception to its own category) and (b) differences in binding requirement (i.e. similarity of an exception to the contrasting category).

Three experiments were conducted. Experiment 1 was set as a replication of previous studies that examined recognition memory for Differentiation structure. The category structure completely follows the one reported by Davis, Love, & Preston (2012) and uses the same experimental tasks and procedures. Experiment 2 and Experiment 3 build on Experiment 1 by employing the same experimental design, procedures and materials to examine memory for Isolation (Experiment 2) and Odd-ball (Experiment 3) exceptions.

If the memory advantage for category exceptions results from violation of a salient knowledge structure, exceptions should be remembered better than regular items across the three experiments. However, if the advantage for exceptions

is dependent on differences in binding requirements, Differentiation (and potentially Isolation exceptions) should be remembered better, while there should be no advantage for Odd-ball exceptions.

Experiment 1: Differentiation

Methods

Participants Participants were 38 undergraduate students from a Midwestern university who received course credit for their participation. Two additional participants were excluded due to the failure to finish the experiment.

Materials The stimuli were schematic clown-like faces that varied along four (feature) dimensions. Items were accompanied by two novel category labels: Zuzu and Tati. As it can be seen in Figure 2, the four feature dimensions were hair, eyes, mouth and side whiskers. Side whiskers were selected as a rule dimension and the three other features varied between the two categories.

The category structure Table 1 shows an abstract representation of the category structure used in Experiment 1 (as well as ones used in Experiments 2 and 3). Each of the two categories had three Regular, rule-following items and one Exception.

Rule-following items could be categorized accurately based on the value of a single rule-dimension (side whiskers). The rule dimension was held constant across participants. The other three dimension varied between the categories, with exactly the same combinations of the three features used for constructing items of category A and category B (see Table 1). The Exceptions appeared to belong to the opposing category based on their value on the rule-relevant dimension. Additionally, the two exceptions had the same values on the three other dimensions, and thus could be categorized accurately only based on the representation that captures the combination of the rule and (at least) 2 other features.

Based on the items presented in Table 1 that were used during training, we constructed foils for memory test. The foils had the same feature values as training items, but in novel combinations. There was a total of 8 foils constructed for memory test in Experiment 1.

Table 1: The category structure used in Experiments 1-3

	Category A	Category B
Regular items		
same set across the three experiments	**1** 334 **1** 343 **1** 433	**2** 334 **2** 343 **2** 433
Exceptions		
Exp 1: Differentiation	**2** 444	**1** 444
Exp 2: Isolation	**2** 555	**1** 555
Exp 3: Odd-ball	**8** 888	**9** 999

Note. 1 = rule of category A; 2 = rule of category B; 3 = probabilistic; 4 = probabilistic; 5 = novel non-diagnostic; 8 = unique for exception A; 9 = unique for exception B.

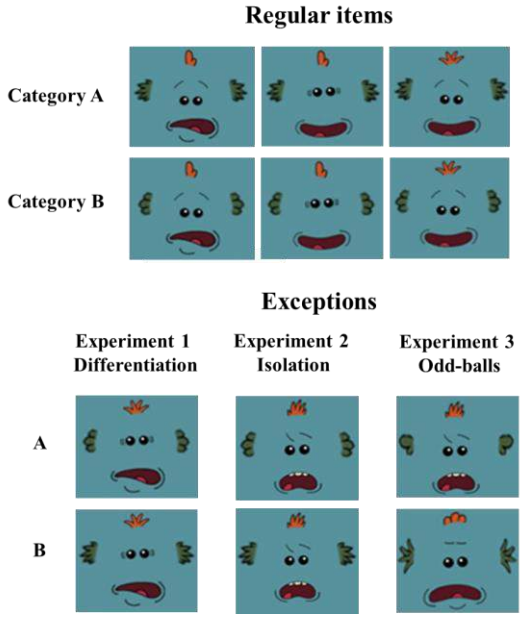


Figure 2: Complete set of stimuli used in the training (Experiments 1–3). Regular items were the same across the experiments.

Procedure

The experiments consisted of three phases: training, memory test and categorization test.

Training During training participants were presented with the exemplars of the two categories and were asked to classify each exemplar.

Following the procedure of previous studies, participants were given explicit instructions indicating the rule feature. They were encouraged to use this feature during categorization and to memorize items that violate this rule (Davis, Love, & Preston, 2012).

Items were presented individually, and corrective feedback was provided after each response. There was a total of 64 trials presented during training, 48 rule-following items and 16 exceptions (i.e. each of the 6 Regular and 2 Exception items presented 8 times in random order).

Memory test Following training, participants were introduced to the memory test. In the memory test, participants saw two items at a time: one training item and one foil (item that had the same features as the training items, but in a novel combination). Their task was to say which of the two items was old (presented during the training). There was no feedback given during memory test. The test had 48 trials presented in a random order.

Categorization test In the categorization test participants were presented with Regular and Exception items they saw during the training. The procedure was exactly the same as in the training session with the only difference being that during the categorization test, participants were not provided with feedback. There were 16 categorization test trials, 8 Regular items and 8 Exceptions.

Results

Figure 3 shows participants' recognition memory and categorization accuracy (panel a).

Participants were less accurate at categorizing Exceptions (M = 0.49, SD = 0.37) than Regular items (M = 0.82, SD = 0.24), $t(37) = 5.06$, $p < 0.001$, $d = 0.82$. However, they had better recognition memory for Exception items (M = 0.63, SD = 0.19) than for Regular, rule-following items (M = 0.47, SD = 0.12), $t(37) = 3.77$, $p < 0.01$, $d = 0.61$.

Both of these results, memory advantage for Exceptions and better categorization accuracy for Regulars, are in accordance with previously reported findings (Palmeri & Nosofsky 1995; Sakamoto & Love, 2004, 2006).

It is important to note that here, as in the previously reported studies, the advantage in memory for rule-violating exceptions results from optimization in memory for Regular items. Since participants categorized Regular items relying on the category rule, the rule feature is the only feature they needed to represent and thus they had no need to remember individual exemplars representing the category. On the other hand, in order to learn Exceptions, they had to bind in memory information about a minimum of three features (the rule and two other features).

Experiment 2: Isolation

The goal of Experiment 2 was to test the robustness of the memory advantage for Exceptions, when rule-violating items are Isolation Exceptions. Isolation Exceptions have lower memory demands than Differentiation Exceptions, but they still require binding of information about the rule and (at least) one more feature.

Methods were identical to Experiment 1, except for the type of Exception participants needed to learn (See Table 1). Foil items for the memory test were designed accordingly to include features of Isolation Exceptions, which resulted in 19 foil items in total. Twenty-three undergraduates from a Midwestern university participated for course credit.

Results

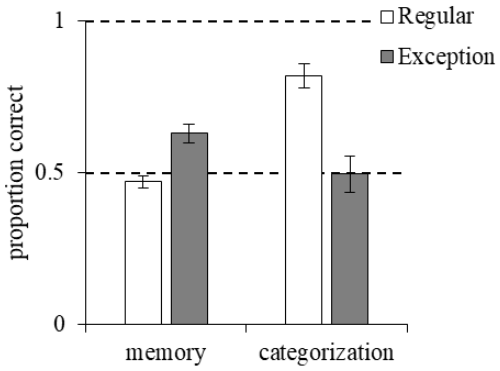
The pattern of results in Experiment 2 closely replicated the one observed in Experiment 1 (Figure 3).

Although participants were more accurate when categorizing Regular (M = 0.85, SD = 0.18) than Exception items (M = 0.65, SD = 0.36), $t(22) = 2.41$, $p < 0.05$, d = 0.50, they remembered Exception items (M = 0.67, SD = 0.26) more accurately than Regular, rule-following items (M = 0.51, SD = 0.09) , $t(22) = 2.91$, $p < 0.01$, $d = 0.61$.

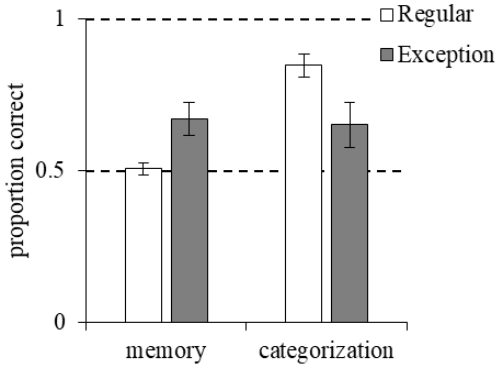
Experiment 3: Odd-ball

The critical difference between Experiment 3 and Experiments 1-2, was that learning of rule-violating items in Experiment 3 did not require forming of a complex binding structure. Both Regular and Exception items could be categorized based on a single, individual feature. Thus, any differences in recognition memory between Regular items and Odd-ball Exceptions could be solely due to effects of rule violation.

a. Experiment 1: Differentiation



b. Experiment 2: Isolation



c. Experiment 3: Odd-ball

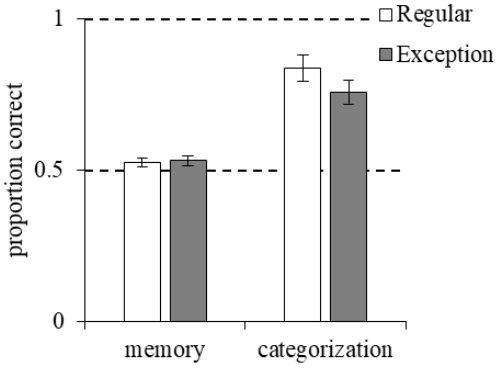


Figure 3: Recognition memory and categorization accuracy across the three experiments. Error bars represent standard errors of mean.

Methods were identical to the ones used in Experiment 1-2, except that Experiment 3 used Odd-ball Exceptions as the rule-violating items (See Table 1). Foil items for the memory test were constructed following the logic of Experiment 1 and 2. There was a total of 21 foil items used for the memory test. Forty undergraduate students from a Midwestern university participated for course credit.

Results

In Experiment 3, participants were equally accurate when categorizing Exceptions (M = 0.76, SD = 0.25) and Regular items (M = 0.84, SD = 0.28), $t(39) = -1.30$, p = .203, d = 0.20.

Critically, we observed no difference in recognition memory between the two item types, $t(39) = 0.18$, p = .858, d = 0.03. Participant had no memory for exemplars of either Regular (M = 0.53, SD = 0.10), or Exception items (M = 0.53, SD = 0.11). One sample t-tests against chance were approaching significance for both Regulars ($t(39) = 1.79$, p = .081) and Exceptions ($t(39) = 1.78$, p = .084).

Discussion

The presented work aimed to clarify whether category exceptions merit a special memory representation because they violate a salient knowledge structure, as it has been previously assumed (e.g., von Restorff, 1933; Hunt & Lamb, 2001; Busey & Tunnicliff, 1999; Nairne, 2006), or is it only those exceptions that have high binding requirements that have the special memory status. This is a critical question, as in the latter case, the special memory status characterizes certain kinds of exceptions rather than exceptions in general. Consequently, the generalizations often present in the categorization literature when discussing exceptions would be unjustified.

The recognition memory comparisons across the three experiments revealed that participants had better recognition memory for exceptions that required binding of two or more features to be accurately categorized. However, when memory demands for regular and exception items were equal, there was no memory advantage for exceptions. In other words, when category structure does not require feature binding for successful categorization and both item types can be classified based on the individual features, exception items are treated as any other regular item. Participants may optimize their memory when learning exceptions in the same manner as they do when they learn regular items, and thus have poor exemplar memory for both regulars and exceptions.

Model Predictions

The results of previous studies on recognition memory for categories with exceptions were found to generally conform to the predictions of RULEX and SUSTAIN (Palmeri & Nosofsky, 1995; Sakamoto & Love, 2004, 2006). Pure exemplar storage models, such as the context model (Medin & Schaffer, 1978), have also been considered and found to be inadequate at simultaneously predicting categorization accuracy and recognition memory for categories with exceptions (Palmeri & Nosofsky, 1995). Our results replicate this failure of exemplar models. Exemplar models have difficultly predicting good categorization, but bad memory, for regulars since categorization relies directly on memory storage. Similarly, they struggle with good memory, but poor categorization, of exceptions. In general, exemplar models would tend to predict that both categorization and memory would be better (or possibly both worse) for exceptions than regulars, but they would not predict opposite patterns for categorization and memory.

RULEX provides good predictions for the patterns that were found in the Differentiation and Isolation structures, correctly predicting better categorization of rule-following items, but poorer memory for those items since they are not represented independently in memory. Memory is predicted to be better for exceptions since they need to be stored individually in memory. It is unclear, though, what RULEX would predict for the Oddball structure.

Versions of RULEX have been formulated for binary-valued discrete dimensions, and continuous-valued dimensions (Nosofsky & Palmeri, 1998), but (to our knowledge) not for discrete dimensions with more than two possible values. While there are straightforward ways in which to extend RULEX to accommodate this type of stimulus structure, there are several alternative formulations that would provide opposite predictions.

Existing versions of RULEX first try simple unidimensional rules. If perfect rules fail, it then attempts unidimensional rules with exceptions stored in memory. If those representations are inadequate, it moves to considering more complex rules. Our Oddball category structure could theoretically be solved with a disjunctive rule on a single dimension (i.e. value 1 *or* 8 on dimension one is Category A, 2 *or* 9 is category B; see Table 1). It is unclear whether RULEX would attempt to use this type of rule prior to or after it attempts to store exceptions in memory (considering that the rule is technically unidimensional but also somewhat complex). If it tries storing exemplars first, it would predict similar behavior as in the Isolation structure (and therefore, fail to predict our results). If it tries the disjunctive rule first, then it would not need to store any exemplars in memory, and could match participants' data well. So, RULEX could predict all of our results in theory, but it depends on exactly how it is formulated to handle this type of stimulus representation.

SUSTAIN (Love, Medin, & Gureckis, 2004), on the other hand, can naturally process the stimulus structures used in our study without needing modification, but its predictions are somewhat less intuitive. Like RULEX, the predictions depend on whether it stores exceptions separately (by creating additional clusters) or together with regulars. In theory it can do either depending on the exact parameter settings and the order in which it encounters the exemplars. To test whether SUSTAIN would create different numbers of clusters for the three different stimulus structure, we

performed simulations of the model. We first fit the model to each participants' training data using maximum-likelihood estimation in order to obtain reasonable parameter estimates. Then we simulated the model using each parameter combination on all three stimulus structures (1000 simulations per parameter combination, per structure).

Results of the simulations generally match the behavioral results: better categorization for regulars than exceptions in all structures, but better memory for exceptions than regulars—except in the Oddball structure where memory performance was roughly equivalent between regulars and exceptions. Additionally, the number of clusters formed was found to be highest for the Differentiation structure (median: 6 clusters; mode: 4 clusters), slightly lower for the Isolation structure (median: 4 clusters; mode: 4 clusters) and lowest for the Oddball structure (median: 3 clusters; mode: 3 clusters). Importantly, that there were typically fewer than 4 clusters in the Oddball structure indicates that exceptions were not represented completely independently of the regulars, which is consistent with worse memory for exceptions compared to the other two structures.

In summary, both RULEX and SUSTAIN can potentially account for our results by representing exceptions separately from regulars in the Differentiation and Isolation structures, but not in the Oddball condition. SUSTAIN produces this pattern as a normal result of its category learning process, while RULEX produces this result under one of several possible instantiations of its decision process. In both models the separate representations of exceptions are consistent with increased feature binding for those items compared to regulars, though they may not have been described in terms of feature binding in previous work.

Conclusions

Taken together, our findings suggest that the previously reported advantage for memory exceptions reflects elevated memory demands of specific kind of exception which does not generalize to other kinds of exceptions.

This work further adds to our understanding of what makes *some* category exceptions more memorable, by focusing on the critical role of competition between the exception and contrasting category members.

Acknowledgments

This information will be added to the final version of this manuscript (in order not to violate double-blind review process).

References

Busey, T. A., & Tunnicliff, J. L. (1999). Accounts of blending, distinctiveness, and typicality in the false recognition of faces. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 25, 1210-1235.

Davis, T., Love, B.C., & Preston, A.R. (2012). Learning the exception to the rule: Model-based fMRI reveals specialized representations for surprising category members. *Cerebral Cortex*, 22, 260-273.

Hunt, R. R., & Lamb, C. A. (2001). What causes the isolation effect? *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 27, 1359-1366.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of human category learning. *Psychological Review*, 111, 309–332.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, 85, 207-238.

Nairne, J. S. (2006). Modeling distinctiveness: Implications for general memory theory. In R. R. Hunt & J. B. Worthen (Eds.), *Distinctiveness and memory*, New York: Oxford University Press.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, 101, 53–79.

Nosofsky, R. M., & Palmeri, T. J. (1998). A rule-plus-exception model for classifying objects in continuous-dimension spaces. *Psychonomic Bulletin & Review*, 5, 345-369.

Palmeri, T. J., & Nosofsky, R. M. (1995). Recognition memory for exceptions to the category rule. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21, 548–568.

Rojahn, K., & Pettigrew, T. F. (1992). Memory for schema-relevant information: A meta-analytic resolution. *British Journal of Social Psychology*, 31(2), 81–109.

Sakamoto, Y., & Love, B. C. (2004). Schematic influences on category learning and recognition memory. *Journal of Experimental Psychology: General*, 133, 534–553.

Sakamoto, Y., & Love, B.C. (2006). Vancouver, Toronto, Montreal, Austin: enhanced oddball memory through differentiation, not isolation. *Psychonomic Bulletin & Review*, 13, 474-9.

Stangor, C., & McMillan, D. (1992). Memory for expectancy-congruent and expectancy-incongruent information: A review of the social and developmental literatures. *Psychological Bulletin*, 111, 42–61.

von Restorff, H. (1933). Analyse von vorgängen im spurenfeld: I. Über die wirkung von bereichsbildungen im spurenfeld [Analysis of processes in the memory trace: I. On the effect of group formations on the memory trace]. *Psychologische Forschung*, 18, 299–342.

Rapid Semantic Integration of Novel Words Following Exposure to Distributional Regularities

Abstract

Our knowledge of words consists of a lexico-semantic network in which different words and their meanings are connected by relations, such as similarity in meaning. This research investigated the integration of new words into lexico-semantic networks. Specifically, we investigated whether new words can rapidly become linked with familiar words given exposure to distributional regularities that are ubiquitous in real-world language input, in which familiar and new words either: (1) directly co-occur in sentences, or (2) never co-occur, but instead share each other's patterns of co-occurrence with another word. We observed that, immediately after sentence reading, familiar words came to be primed not only by new words with which they co-occurred in sentences, but also by new words with which they shared co-occurrence. This finding represents a novel demonstration that new words can be rapidly integrated into lexico-semantic networks from exposure to distributional regularities.

Keywords: word learning; semantic priming; distributional semantics; semantic integration

Introduction

Starting early in development and continuing through adulthood, we amass sizable vocabularies commonly estimated to contain tens of thousands of words (Schmitt & McCarthy, 1997). Beyond the size of the resulting vocabulary, word learning is remarkable both because much of it unfolds merely by encountering words in linguistic contexts without explicit instruction, and because it leads to the formation of an organized lexico-semantic network in which different words and their meanings are linked by relations. For example, our lexico-semantic networks contain links both between words that can be combined to form meaningful utterances (e.g., *eat* and *apple*), and words similar in meaning (e.g., *apple* and *grape*) (Jones, Willits, Dennis, & Jones, 2015). These links are a fundamental facet of our lexico-semantic knowledge that influence behavior even without awareness, reasoning, or recall of relevant information from episodic memory (as is evident from phenomena such as priming). How do the new words we encounter become integrated into our lexico-semantic networks?

The purpose of the present research is to investigate the rapid integration of new words into existing lexico-semantic networks purely on the basis of regularities with they are distributed with other words in linguistic input. As demonstrated by the seminal work of Landauer and Dumais (1997) and many subsequent modeling efforts (Frermann & Lapata, 2015; Huebner & Willits, 2018; Jones & Mewhort, 2007; Rohde, Gonnerman, & Plaut, 2004), sensitivity to distributional regularities may represent a powerful mechanism for building lexico-semantic networks. First, links between words that can be combined to form meaningful utterances such as *eat* and *apple* can be formed from the regularity with which they co-occur in language. Critically, although words similar in meaning such as *apple* and *grape* may not reliably co-occur, links between them can also be formed from the regularity with which they *share* each other's patterns of co-occurrence (e.g., *apple* and *grape* may not reliably co-occur, but do share each other's co-occurrence with *eat*, *juicy*, etc.). These distributional regularities are sufficiently abundant in language that mechanistic models that form representations of words purely on the basis of these regularities capture the majority of links present in human lexico-semantic networks (Jones et al., 2015).

In spite of the extensive evidence from modeling research supporting the potential contributions of sensitivity to distributional regularities, we know little about whether exposure to these regularities actually drives the integration of new words into lexico-semantic networks in human learners. Accordingly, the present research was designed to assess whether adults semantically integrate novel words with familiar words after reading sentences rich in distributional regularities. Specifically, we investigated whether familiar words came to be semantically primed not only by novel words with which they co-occurred, but also by novel words with which they never co-occurred, and instead shared patterns of co-occurrence with another word.

In what follows, we first review existing evidence about human learner's sensitivity to distributional regularities. In this review, we highlight the paucity of prior research that is informative about the role of distributional regularities abundant in language in building human lexico-semantic networks. We then present an experiment designed to illuminate this role.

Human Sensitivity to Distributional Regularities in Language

Extensive evidence from statistical learning research suggests that humans are sensitive to some forms of distributional regularities in some modalities. Specifically, numerous studies have revealed that we are sensitive to the regularity with which items such as speech sounds or shapes co-occur, either simultaneously, sequentially, or separated by some number of other items (Conway & Christiansen, 2005; Fiser & Aslin, 2002; Gomez, 2002; Saffran, Johnson, Aslin, & Newport, 1999).

However, this evidence cannot directly illuminate whether distributional regularities of words in language can drive lexico-semantic integration for two reasons. First, very little statistical learning research conducted to date has investigated whether we form links between items that never occur together, and instead share each other's patterns of co-occurrence with other items (to our knowledge, only one study visual domain, Schapiro, Rogers, Cordova, Turk-Browne, & Botvinick, 2013, is suggestive of this form of learning). However, this process is a critical facet of the potential importance of sensitivity to distributional regularities for building lexico-semantic networks, because many words similar in meaning do not reliably co-occur, and can instead only be linked based on their shared patterns of co-occurrence (Jones et al., 2015). Second, statistical learning research has focused on learning links between items in domains that intentionally do not carry meaning, such as speech sounds, acoustic sounds, and shapes, and tactile stimuli. Because statistical learning phenomena vary even across these studied domains (Conway & Christiansen, 2005), it is unclear whether they generalize to the formation of semantic links between novel and familiar words in language.

To our knowledge, the only evidence relevant to the role of distributional regularities in semantic integration comes from a handful of studies conducted by McNeill (McNeill, 1963, 1966). In these studies, novel words were organized into triads, in which one novel word A co-occurred in sentences with either of two other novel words, B and C. Accordingly, the distributional regularities consisted of both the direct co-occurrence of A- B and A-C, and the shared co-occurrence of B-C (which never actually co-occurred, but both co-occurred with A). By administering a free association task at multiple points during sentence reading in which participants were asked to produce the first novel word that came to mind when prompted with another, McNeill observed that participants first formed links between novel words that directly co-occurred (i.e., A-B and A-C), and then between those that shared co-occurrence (i.e., B-C). This finding provides evidence that people can learn the distributional regularities of words in sentences online, as they are experienced. These regularities therefore represent a viable candidate for drivers of semantic integration. However, these studies were not designed to investigate the semantic integration of novel words into existing lexico-semantic networks, because novel words only ever shared distributional regularities with each other, and not with familiar words. Moreover, the use of a free association task to assess learning leaves open the possibility that these links participants apparently formed were based on retrieving the episodic experiences of reading the sentences from memory, rather than on the formation of automatically-activated semantic links. The role of distributional regularities in lexico-semantic integration therefore formed the focus of the present experiments.

Present Experiments

The present experiments were designed to investigate whether distributional regularities can drive the rapid integration of new words into existing lexico-semantic networks. Specifically, participants read sentences in which were embedded triads of words that consisted of a novel pseudoword (e.g., foobly) that regularly preceded a familiar word (e.g., apple) in some sentences, and another novel pseudoword (e.g., mipp) in other sentences. Accordingly, the sentences contained distributional regularities with which a familiar word (e.g., apple) both directly co-occurred with one novel psuedoword (foobly), and shared this pattern of co-occurrence with another (mipp) (Fig. 1). The sentences otherwise contained no information from which the meanings of the novel pseudowords could be derived. For example, participants might read "My sister loves to see a foobly apple" and "I saw a foobly mipp on vacation".

Immediately following a short session of sentence reading, we then assessed lexico-semantic integration by testing whether the familiar word came to be primed by both the novel pseudoword with which it co-occurred, and the novel pseudoword with which it shared this pattern of co-occurrence. To show both patterns of priming, participants must: (1) Learn the novel word forms, (2) Form links between novel and familiar words that directly co-occur, and (3) Derive links between novel and familiar words that never co-occur, but instead share each other's patterns of co-occurrence.

Method

Participants

Participants were 45 undergraduate students from a Midwestern university who received course credit. An additional five participants were excluded due to failure to complete the experiment.

Stimuli and Design

Training. The training stimuli were two triads of words (1: foobly-apple-mipp; 2: dodish-horse-geck) that each consisted of a pseudoadjective (e.g., foobly) that consistently preceded one familiar noun (e.g., apple) and one pseudonoun (e.g., mipp) in different sentences. Each word pair from these triads (foobly-apple, foobly-mipp, dodish-horse, dodish-geck) was embedded in 10 unique sentence frames, for a total of 40 training sentences. These sentences therefore conveyed both direct co-occurrences between words in the same pair from the same triad, and shared co-occurrences between familiar and pseudonouns from the same triad. The sentences did not convey any other cues to pseudoword meaning (Figure 1).

Test. For testing purposes, we added two new pseudowords (nuppical; boff) and 2 pictures: One of an apple and one of a horse.

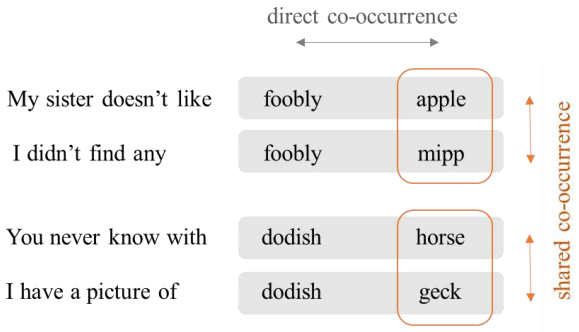


Figure 1: Illustration of training sentence structure.

Using these stimuli, we generated five types of Prime-Target word pairs. Primes were always novel pseudowords, and Targets were always one of the two familiar nouns used during training (apple or horse). First, we generated two types of Related pairs that were consistent with the training triads: Related Direct, in which a pseudoadjective preceded the familiar noun that it had preceded during training (e.g., foobly-apple), and Related Shared, in which a pseudonoun preceded the familiar noun with which it had shared co-occurrence during training (e.g., mipp-apple). Second, we generated corresponding Unrelated Direct and Unrelated Shared pairs in which the Primes from Related pairs were switched, such that they violated the regularities present during training (e.g., foobly-horse). Finally, we generated Neutral pairs, in which the new pseudowords that were only present during Test (nuppical; boff) preceded each familiar noun.

Procedure

The experiment had 2 phases: Training and Test.

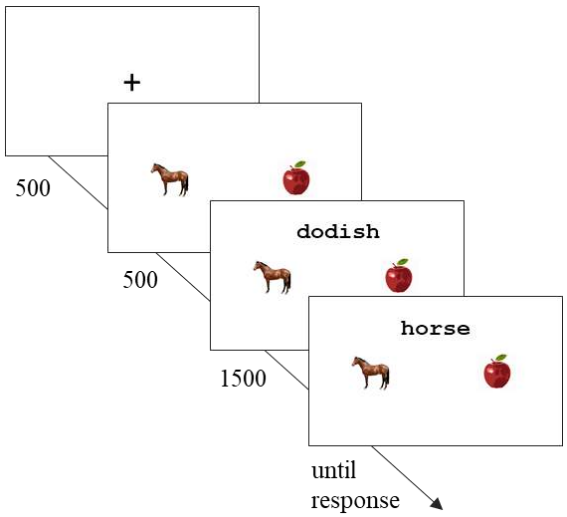


Figure 2: Timing of events during the primed visual search task used in the Test phase.

Training. The Training consisted of three blocks. In each block, participants first read all of the 40 training sentences in a random order at their own pace. To check whether participants were attending to the sentences, three control questions appeared following random sentences in which participants were prompted to type the novel words from the last sentence they had read. The reading component of each block was followed by a free association task in which participants were asked to respond with the first novel (pseudo) word they could think of when prompted with each of the pseudowords from the training sentences. Each of the pseudowords (foobly, dodish, mipp, geck) was presented 3 times in a randomized order.

Test. For the test phase, participants performed a primed visual search task (see Figure 2 for timing of events in trials). At the start of each trial, participants saw a fixation cross followed by two images, one on either side of the screen: A horse, and an apple. Two words (a Prime and Target) were then consecutively presented as text on the top of the screen. Participants' task was to read both words, but choose the image labeled by the second (i.e., Target) word using the mouse. During a practice phase consisting of 8 trials, the two words consisted of Neutral word pairs (i.e., a new pseudoword followed by apple or horse). During the actual task consisting of 144 trials, the two words consisted of Related Direct, Related Shared, Unrelated Direct, Unrelated Shared, and Neutral pairs.

Participants were given an unlimited time to make their responses, but were prompted to respond quickly and were shown a message saying that they were too slow if their response time on a trial was > 800ms.

Results and Discussion

Preliminary analyses: Free association

To test whether participants were attending to the sentences, we analyzed participants' responses on the free association task. Participants responded as instructed by responding with one of the training pseudowords on an average 90.6% of all free association trials. Participants tended to respond with training pseudowords that had directly co-occurred with the prompt pseudoword: 88% of all responses to pseudoadjective prompts were with the noun that followed the pseudoadjective during training, and 77% of responses to pseudonoun prompts were the pseudoadjective that preceded it during training. Only 2.5% of all responses to pseudonouns were based on shared co-occurrence. This confirmed that participants read the sentences and learned the word forms.

Main analyses: Priming[1]

The purpose of the main analyses was to investigate whether the novel pseudowords were semantically

[1] The full pattern of effects on reaction time have also been replicated with two samples (*N*s= 25 and 28) of participants recruited from Amazon Mechanical Turk: Once as an exact replication, and once as a conceptual replication in which

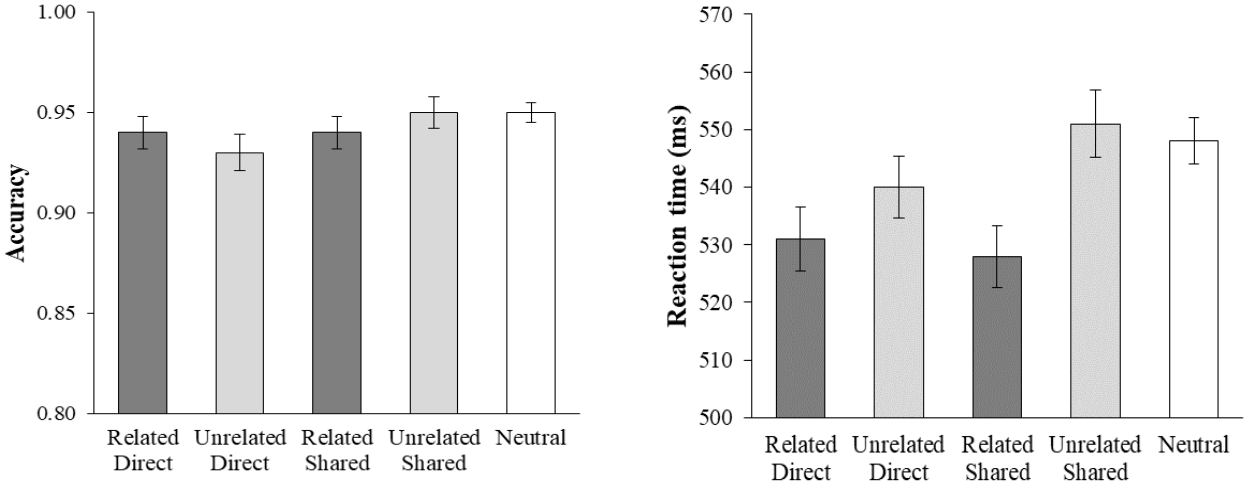


Figure 3: Mean accuracy (left) and reaction times (right) across five conditions. Dark gray bars represent Related (Direct and Shared) conditions, and light gray bars represent Unrelated (Direct and Shared) conditions. The Neutral condition (new pseudoword) is presented in white. Error bars indicate the standard errors of the means.

integrated with familiar words with which they shared distributional regularities (i.e., direct or shared co-occurrence) in training sentences. We accomplished this investigation by measuring whether the novel pseudowords affected the speed and accuracy of processing familiar words in the priming task used during the Test phase. Specifically, we compared the speed and accuracy with which participants identified whether the Target word was apple or horse when it was preceded by a novel pseudoword in the Related Direct, Related Shared, Unrelated Direct, Unrelated Shared, and Neutral conditions. Related pseudowords were expected to facilitate Target word identification, whereas Unrelated pseudowords were expected to inhibit identification. Moreover, these facilitation and inhibition effects may be greater for Direct versus Shared co-occurrences.

Prior to conducting this analysis, we first eliminated data from 8 participants with extremely short reaction times (more than 2/3rds of RTs < 100ms), leaving a sample size of 38 participants. Accuracies and Reaction Times are presented in Figure 3.

Accuracy. We analyzed effects on accuracy using a linear mixed effects regression model in which Relatedness (Related vs Unrelated) and Type (Direct vs Shared) were fixed effects, and Participant was a random effect. This model revealed no effect of either Relatedness or Type on accuracy (Relatedness: B = - 0.004, SE = 0.008, t = - 0.55, p = .59, d = 0.004, Type: B = - 0.012, SE = 0.008, t = - 1.48, p = .15, d = 0.012).

Reaction Time. For analyses of reaction time, we removed data from incorrect trials, and trials with extremely short (<100 ms) and extremely long response latencies (>1500 ms), resulting in removal of 8.1 % of trials.

We then generated a linear mixed-effects model with Relatedness (related; unrelated) and Type (direct; shared) as fixed effect factors and Participants as a random effect. This model revealed no significant effect of Type (neither as a main effect nor in interaction with Relatedness). Thus, Type was excluded from the final model. A log-likelihood ratio test indicated that the best fitting random effects structure included only a random intercept for participants. Thus, the final model included Relatedness as a fixed effect factor and a random intercept for participants. This model revealed a significant effect of Relatedness on reaction times, B = 14.82, SE = 5.10, t = 2.91, p < .01, d = 0.096 (see Brysbaert & Stevens, 2018 for effect size estimate approach). Participants were 14.9 ms faster in related than in unrelated conditions (Figure 3, right panel). The model explained 16% of total variance (R-squared based on Nakagawa & Schielzeth, 2013).

The follow-up analyses compared Related and Unrelated conditions to the Neutral condition. A linear mixed-effects model with Condition (Neutral; Related Direct, Related Shared, Unrelated Direct, Unrelated Shared) as a fixed effect factor and a random intercept for Participants revealed that only Related conditions were significantly different than the Neutral (Related Direct: B = - 16.11, SE = 6.30, t = - 2.56, p = .01, d = .104; Related Shared: B = -16.56, SE = 6.32, t = - 2.62, p < .01, d = .104). There was no significant difference in RT between the Neutral condition and Unrelated conditions. In other words, participants were faster to respond when the Target was preceded by a pseudoword that either directly co-occurred with the Target (Related Direct) or shared the pattern of co-

the pseudoadjectives were changed from foobly/dodish to foobing/doding.

occurrence (Related Shared) than when it was preceded by a new pseudoword that only appeared in the Test and not the Training phase. Primes that were incongruent with the regularities presented during the training (Unrelated Direct, Unrelated Shared) did not affect speed.

General Discussion

The present experiment provides a novel demonstration that new words can be rapidly integrated into existing lexico-semantic networks based on the distributional regularities of words in sentences. Specifically, immediately following a short session of sentence reading, familiar words came to be primed by both novel words with which they co-occurred in sentences, and novel words with which they never co-occurred, but instead shared a pattern of co-occurrence with another novel word. Given that these distributional co-occurrence regularities are ubiquitous in language (Jones et al., 2015), the present results provide evidence that sensitivity to these regularities may represent a critical way in which new words are rapidly integrated into lexico-semantic knowledge.

Implications for Lexico-Semantic Integration

The present findings build upon prior research in two key ways. First, prior evidence about the potentially powerful contributions of distributional regularities to building lexico-semantic networks comes primarily from modeling research. The present findings therefore substantially underline this potential by demonstrating that new words can be added to actual human lexico-semantic networks through mere exposure to distributional regularities.

Second, this evidence also adds to our understanding of how rapidly new words can be integrated into our existing lexico-semantic networks. Specifically, extensive prior research has investigated the lexico-semantic integration of novel words through different kinds of input, such as studying definitions of novel words, or repeatedly observing words co-occurring with images of specific familiar objects (Breitenstein, Zwitserlood, de Vries et al., 2007; Clay, Bowers, Davis, & Hanley, 2007; Dagenbach, Horst, & Carr, 1990; Dobel, Junghöfer, Breitenstein et al., 2010; Tamminen & Gaskell, 2013). Much of this research has suggested that newly learned words are only gradually integrated into existing lexico-semantic networks, following at least one day and up to several weeks of consolidation. In contrast, a handful of recent findings (Borovsky, Elman, & Kutas, 2012; Mestres-Missé, Rodriguez-Fornells, & Münte, 2006; Zhang, Ding, Li, & Yang, 2019) have suggested that lexico-semantic integration of novel words can occur more rapidly when learning is driven by reading sentences in which novel words appear in a position typically occupied by a specific, familiar word (e.g., "It was a windy day, so Peter went to the park to fly his *dax*"). The present findings add to this evidence that novel words can be integrated into existing lexico-semantic networks very rapidly, immediately following an initial learning experience.

Future Directions

The evidence provided by the present experiment highlights a new avenue for future research to investigate *how* distributional regularities foster semantic integration. For example, do direct co-occurrences foster integration more rapidly than shared co-occurrences, or do these processes unfold in parallel? This question could be addressed by measuring integration (e.g., using the priming approach taken in the present experiment) at multiple points throughout training. Moreover, addressing this and related questions could help to generate and arbitrate between different potential mechanistic accounts of distributional regularity-driven semantic integration.

Summary

Throughout our lives, we amass a sizable and interconnected body of knowledge of words and their meanings. The present research highlights how the formation of these lexico-semantic networks may be critically facilitated by the rapid integration of new words via sensitivity to the regularities with which words occur with other words in linguistic input.

References

Borovsky, A., Elman, J. L., & Kutas, M. (2012). Once is enough: N400 indexes semantic integration of novel word meanings from a single exposure in context. *Language Learning and Development, 8*, 278-302.

Breitenstein, C., Zwitserlood, P., de Vries, M. H., Feldhues, C., Knecht, S., & Dobel, C. (2007). Five days versus a lifetime: Intense associative vocabulary training generates lexically integrated words. *Restorative Neurology and Neuroscience, 25*, 493-500.

Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. *Journal of Cognition, 1*, 9.

Clay, F., Bowers, J. S., Davis, C. J., & Hanley, D. A. (2007). Teaching adults new words: the role of practice and consolidation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*, 970-976.

Conway, C. M., & Christiansen, M. H. (2005). Modality-constrained statistical learning of tactile, visual, and auditory sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 24.

Dagenbach, D., Horst, S., & Carr, T. H. (1990). Adding new information to semantic memory: How much learning is enough to produce automatic priming? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*, 581-591.

Dobel, C., Junghöfer, M., Breitenstein, C., Klauke, B., Knecht, S., Pantev, C., & Zwitserlood, P. (2010). New names for known things: on the association of novel word forms with existing semantic information. *Journal of Cognitive Neuroscience, 22*, 1251-1261.

Fiser, J., & Aslin, R. N. (2002). Statistical learning of new visual feature combinations by infants. *Proceedings of the National Academy of Sciences, 99*, 15822-15826.

Frermann, L., & Lapata, M. (2015). Incremental Bayesian Category Learning From Natural Language. *Cognitive Science, 40*, 1333–1381.

Gomez, R. L. (2002). Variability and detection of invariant structure. *Psychological Science, 13*, 431-436.

Huebner, P. A., & Willits, J. A. (2018). Structured semantic knowledge can emerge automatically from predicting word sequences in child-directed speech. *Frontiers in Psychology, 9*.

Jones, M. N., & Mewhort, D. J. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological review, 114*, 1-37.

Jones, M. N., Willits, J., Dennis, S., & Jones, M. (2015). Models of semantic memory. In J. Busemeyer & J. Townsend (Eds.), *Oxford Handbook of Mathematical and Computational Psychology* (pp. 232-254). New York, NY: Oxford University Press.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211.

McNeill, D. (1963). The origin of associations within the same grammatical class. *Journal of Verbal Learning and Verbal Behavior, 2*, 250-262.

McNeill, D. (1966). A study of word association. *Journal of Memory and Language, 5*, 548.

Mestres-Missé, A., Rodriguez-Fornells, A., & Münte, T. F. (2006). Watching the brain during meaning acquisition. *Cerebral Cortex, 17*, 1858-1866.

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution, 4*, 133-142.

Rohde, D. L., Gonnerman, L. M., & Plaut, D. C. (2004). An improved method for deriving word meaning from lexical co-occurrence. *Cognitive Psychology, 7*, 573-605.

Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition, 70*, 27-52.

Schapiro, A. C., Rogers, T. T., Cordova, N. I., Turk-Browne, N. B., & Botvinick, M. M. (2013). Neural representations of events arise from temporal community structure. *Nature Neuroscience, 16*, 486-492.

Schmitt, N., & McCarthy, M. (1997). *Vocabulary: Description, acquisition and pedagogy*: Cambridge University Press.

Tamminen, J., & Gaskell, M. G. (2013). Novel word integration in the mental lexicon: Evidence from unmasked and masked semantic priming. *The Quarterly Journal of Experimental Psychology, 66*, 1001-1025.

Zhang, M., Ding, J., Li, X., & Yang, Y. (2019). The impact of variety of episodic contexts on the integration of novel words into semantic network. *Language, Cognition and Neuroscience, 34*, 214-238.

A Cognitive Model for Understanding the Takeover in Highly Automated Driving Depending on the Objective Complexity of Non-Driving Related Tasks and the Traffic Environment.

Marlene Scharfe (m.scharfe@campus.tu-berlin.de)
Department of Psychology and Ergonomics, Marchstr. 23
10587 Berlin, Germany

Nele Russwinkel (nele.russwinkel@tu-berlin.de)
Department of Psychology and Ergonomics, Marchstr. 23
10587 Berlin, Germany

Abstract

The aim of this study is to refine a cognitive model for the takeover in highly automated driving. The focus lies on the impact of objective complexity on the takeover and resulting outcomes. Complexity consists of various aspects. In this study, objective complexities are divided into the complexity of the non-driving-related task (no-task, listening, playing, reading, searching) and the traffic complexity (relevant vehicles in the driving environment). The impact of a non-driving related tasks' complexity on the takeover is evaluated in empirical data. Following, the cognitive model is run through situations of different traffic complexities and compared to empirical results. The model can account for empirical data in most of the objective complexities. Additionally, model predictions are tested on significant variations in different complexities until the action decision is made. In more complex traffic conditions, the model predicts longer times on different processing steps. Altogether, the model can be used to explain cognitive mechanisms in differently complex traffic situations.

Keywords: highly automated driving; HAD; cognitive modeling; ACT-R; takeover; conditional automation; NDRT; non-driving related tasks; real vehicle study; Objective complexity; traffic complexity; Complexity of NDRT; cognitive model predictions;

Introduction

In the field of Highly Automated Driving, the development of technological innovations is growing rapidly. It is not only necessary to develop working technology, but to understand human cognition, enhance the human-machine interaction (HMI) and improve safety and comfort (Sun et al., 2017). Approaching the next SAE Level of automation (Level 3, conditional automation), where the driver still has to take over the driving task if requested (SAE, 2014), the state of the driver plays an important role. Here, the state is determined as the awareness of the surrounding traffic and necessary action decisions. It depends highly on the situation and its complexity in which the driver has to take over. Different approaches of defining situation complexity exist (Baumann and Krems, 2007; Haerem and Rau, 2007; Schlindwein and Ison, 2004). A key factor concerning the driver is the expectation about the future development of a situation, that is activated when a type of situation occurs (Baumann & Krems, 2007). These types of situations can be distinguished in various ways. They could for example be a traffic situation (congestion, construction zone, intense or low traffic etc.), a type of traffic environment (city, highway etc.), a weather condition or further differentiations.

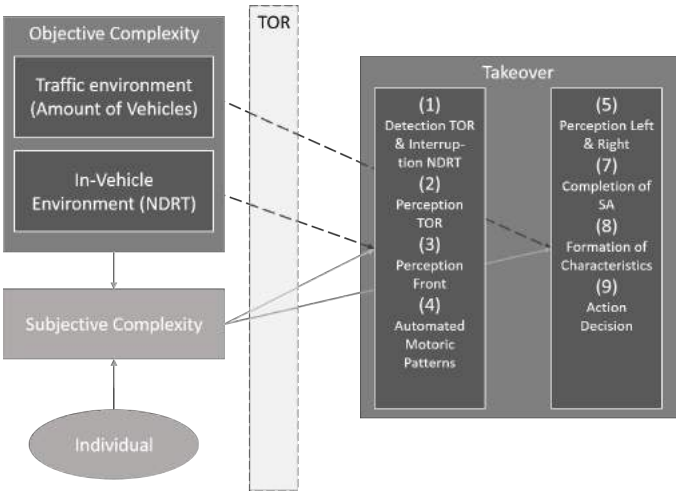


Figure 1: Outline of the Assumed Dependencies, leading to the Approached Hypotheses concerning the Impacts of Objective- and Subjective Complexity on the Takeover Performance. Dark Grey Variables and Interactions are Focus in this Study (Source: own figure).

Due to Schlindwein and Ison, 2004, complexity can be understood as a result of a particular perception of a situation of complexity or resulting from a distinction between expectation and situation development. As we live embedded in situations of complexity, it is important to distinguish between descriptive (objective) and perceived complexity. The perception, that is made by an observer and individually variable, can be determined as perceived complexity (Schlindwein & Ison, 2004). Objective complexity on the other hand describes the complexity a certain traffic situation has. For the first modeling approach presented in this paper, the objective complexity will be the focus of the cognitive model. The impact of the objective complexity on the takeover is analyzed and displayed in the model. According to Paxion, Galy,

and Berthelon (2015), the objective complexity of a situation in driving can vary with road geometry (rectilinear vs. curvilinear), the roadside environments (quantity and variability of traffic signs, variability of scenery) and traffic density (low vs. high). Thus, the role played by objective characteristics is very important (Haerem & Rau, 2007) for the takeover task and will be addressed here. The focus is set on understanding the impact of objective complexity on cognitive mechanisms during a takeover on a highway with varying traffic density in the relevant areas of interest. Thus, an explanation of how the visual perception and the resulting cognitive processing is provided and differences that occur due to different complexities can be displayed. To provide a safe and cognitive adequate takeover, it is necessary to understand which cognitive mechanisms influence different behavior of the driver. Based on such a comprehension of the situation the development of a useful HMI in highly automated driving is possible. It can thus incorporate the current situation and adapt and support the driver accordingly to enable a safe and comfortable takeover.

In this study links between objective complexity and the impact on the takeover are assumed and visualized in (Figure 1). Objective complexity is based on the amount of relevant vehicles in the traffic environment as well as the complexity of the non-driving related task (NDRT) in the in-vehicle environment. The subjective complexity on the other hand is assumed to be influenced by the objective complexity as well as by the individual perception of the objective complexity and management abilities. Both complexity versions should have an impact on cognitive mechanisms and the processing stages during the takeover and the resulting action decision. Nevertheless, as mentioned earlier, in the current context, the focus is set on understanding the impact of objective complexity before approaching subjective complexity. This is important, as the subjective complexity can only be measured, if an understanding about the impact of the objective complexity on the takeover already exists.

In order to perceive different stimuli in a complex environment, awareness of the situation has to be reached and sensory information understood (Plavsic, 2010). In driving, the most important human sense is the visual perception, involving several sub-processes. These are seeing, detection and recognition (Plavsic, 2010). To comprehend the impact of complexity on the takeover in highly automated driving, cognitive processes during a takeover and the influence of objective complexity have to be understood. This can be captured and simulated by a cognitive model. Further resulting behavior can be predicted based on the model.

Cognitive modeling is used to understand more precisely, how complexity emerges and subsequently affects the takeover. Thus, the exploitation of the resources in different complexity combinations can be revealed. For the implementation of the cognitive model, the ACT-R (Adaptive Control of Thought-Rational) cognitive architecture (Anderson et al., 2004) was used. It provides a more accurate representation of human abilities than standard programming languages (Salvucci, Boer, & Liu, 2001). Several cognitive patterns can be modeled and clearly distinguished between the different resources. The architecture provides different modules for each resource that can act simultaneously and interact with each other. Especially the visual module is able to illustrate precisely the above mentioned sub-processes of the visual perception. In conclusion, cognitive modeling is used, as it is a valid and useful method to depict human cognition very detailed with respect to the different resources (visual, haptic, auditory). The ACT-R cognitive architecture is chosen, as it is an architecture that incorporates all relevant mechanisms for the takeover task and enables the modeling of the whole task with respect to the different resources and their interactions. To understand underlying cognitive mechanisms as a function of the objective complexity, the cognitive model is established based on empirical data of a previous study (project KoHAF) and run through different levels of objective complexity. As task performance is reliant on the availability of resources (Kahneman, 1973) and auditory perception uses different resources than visual perception does (multiple resource theory; Wickens, 2008), traffic density has a strong influence on the takeover quality in highly automated driving Radlmayr, Gold, Lorenz, Farid, and Bengler (2014). The developed cognitive model gives an understanding about the underlying cognitive mechanisms. This is necessary for future development of the HMI in highly automated driving. In order to test, whether the model correctly depicts the cognitive processes, the following questions are addressed in the examination of this paper:

- Is the cognitive model able to validly display differences in objective complexity that are found in empirical data?
- Is the cognitive model able to generate predictions that significantly vary with different objective complexities in the traffic environment?

Methods

In this paper, the impact of the complexity of a NDRT and the traffic environment is addressed. As non-driving related tasks (NDRT) play an important role when it comes to taking over the driving task (Radlmayr et al., 2014), the impact of tasks with different complexities is investigated. To validate the cognitive model, data of a previous study (KoHAF) was used. In a first step (Step 1) the influence of NDRT-Complexity on takeover stages in empirical data is evaluated. Further, an ACT-R cognitive model for the takeover task per se, that displays

underlying cognitive mechanisms during a takeover is developed. The model is run through scenarios of different objective complexities and resulting predictions are compared to empirical data (Step 2). The predictions of the model in environments of different complexities are then tested significant differences in prediction times.

Data Acquisition

The data that is evaluated in this paper comes from a previous study in the project KoHAF. As it includes all the relevant information, necessary for the model, it supports the assumptions, that are addressed in this paper. For the realization of a takeover in a real scenario rather than an simulator, a Wizard of Oz vehicle is used. It allows the passenger to drive the vehicle covertly via a hidden control. Thus, a takeover in a real driving environment is possible, simulating a Level 3 automation. Due to that, participants feel like driving an automated vehicle in real traffic (Ko-HAF, 2017) and can engage into secondary tasks during the mock-automation.

The study was held in 2017 in the area of Stuttgart. Overall data of $N = 14$ participants is evaluated. Takeover requests (TOR) after five different NDRTs are covered. A first evaluation of empirical data shows, that NDRTs have a significant influence on the takeover. Due to this, the complexity of the different NDRTs is rated on a ten point likert-scale by three experts based on resource capacities that are needed to solve the tasks. Conditions without NDRT are rated as lowest complex with one point (1P.). A bit more complex, listening to an audiobook (3P.) is valued as it occupies the auditory channel. This is followed by playing Tetris (6P.). Reading a newspaper (7P.) as well as searching something in the back of the vehicle (7P.) is assessed as most complex, each with seven points. Tetris was rated as less complex than reading a newspaper or searching something in the back, as the tablet was mounted to the center console and participants did not need to hold it. Thus, it is assumed as less resource-demanding with regard to the task of taking over. The data was evaluated by two independent raters concerning the different steps of the takeover and the objective complexity of the scenery (amount of visible vehicles on the road and their position). The overall objective complexity thus consists of the scenery and the traffic conditions and of the driving situation. The scenery has a high influence on the objective complexity of a situation (Rommerskirchen, Helmbrecht, & Bengler, 2013), including possible distraction sources from the in-vehicle driver's point of view (e.g. NDRT's).

Cognitive Model

As the most important factor in driving is the visual perception, the focus of the cognitive model to update situation awareness (SA) during the takeover task lies on modeling the perception behavior. Overall longer takeover times are found in a more complex scenery (Radlmayr et al., 2014). This is realized in the model with the focus on visual perception mechanisms of the relevant objects in the traffic environment. The model interacts with a graphical user interface in Lisp. It represents the ego-vehicle on the center lane of a three lane highway. The surrounding traffic is inserted at random, varying between zero and five vehicles in the environment.

Besides visual perception patterns, the cognitive model for the takeover task incorporates motoric and cognitive retrieval patterns. In the following, the steps, that are undertaken until control is regained during a takeover are defined as well as the realization in the cognitive model (Figure 2). While engaging into a secondary task, the driver is alert on whether a takeover request (TOR) appears. This is due to the drivers awareness of situation and task. As soon, as a TOR is detected (0), the NDRT is interrupted (2) and the gaze oriented to the TOR message (1). The model reacts to a stimulus in the visual or aural module, that fits the condition of a TOR message. The meaning of the TOR message is retrieved from the declarative memory and the TOR visually attended, fixated and processed. Then, the visual resource is oriented to the road center and the front lane (near and far area; Salvucci, 2006) is perceived (4). First sensory-motoric patterns (hands to steering wheel, feet to pedals) are automatically applied (3), resting on automated reactions rather than intentionally directed movements. The visual resource further attends and processes the left and right lane (5), storing the status (car or no car) of the attended areas in chunks. In the data, this is followed by the deactivation of automation (6). This is not implemented in the model though, as deactivation modalities vary and there is no common mechanism yet. The model thus completes the perception phase (7), and forms characteristics of current status. The current status of the environment is compared to the task ((8) status-task-mapping) and a decision made based on that (9). Finally, the motoric module performs the selected action ((10) sensory-motoric intervention patterns) that are either to follow, change the lane to the left or right. The vehicle is then stabilized (11). This final step is not explicitly included in the model though.

The cognitive model incorporates these steps and displays the cognitive processes that occur during each one (Figure 2).

Results

Statistical analysis is used, to show, that the cognitive model is able to depict differences that occur due to objective complexity. The two objective complexity measurements (complexity of NDRT and amount of objects in traffic environment) are evaluated separately. The impact of complexity of the NDRT is evaluated in empirical data. NDRT complexities are then scaled and compared

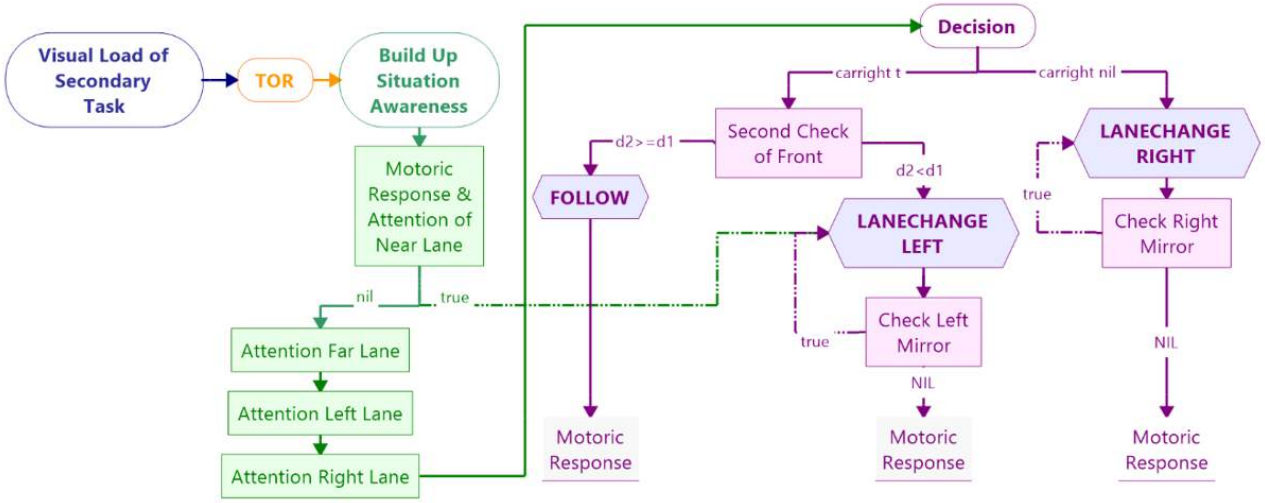


Figure 2: Representation of Main Productions, the Cognitive Model Resolves to Display Cognitive Processes During the Takeover in Highlay Automated Driving (Source: own figure).

to the cognitive model to show, that cognitive model predictions are able to account for empirical data. Based on these results, the model itself is further run through conditions of different traffic complexities and tested on significant variations between those conditions.

The influence of the NDRTs on takeover patterns is tested first using ANOVA for statistical evaluation. Based on that, regression analysis is used to measure the impact of the complexity of the NDRT on the performance of the takeover in empirical data (Step 1). Further, it is examined, whether predictions of the model correlate with the results found in empirical data (Step 2). Finally, model predictions of action decisions are tested on the influence of objective complexity variations (Step 3).

Step 1: Influence of NDRT-Complexity on Takeover Times in Empirical Data

ANOVAs show significant results for the takeover patterns one to four ((1) visual re-orientation and fixation of takeover request (TOR) message, (2) interruption of NDRT, (3) first sensory-motoric patterns, (4) visual orientation to road center). The time until the gaze gaze is directed to the TOR differed statistically significant for the different NDRTs ($F(4,65) = 3.088, p < .05$). The same applies for the time until the NDRT is stopped ($F(4,65) = 4.221, p < .01$), the time until the hands are moved to the steering wheel ($F(4,65) = 12, p < .001$) and the time until the gaze is directed to the road ($F(4,65) = 5.808, p < .001$). Due to this, the impact of complexity on takeover patterns is evaluated, using regression analysis. Based on the regression equation $y = x\beta + \epsilon$, the impact of the Complexity of the NDRT ($CNDRT$) is tested on significance to reject the null hypothesis. Further, the amount of variance that can be explained by the regression (multiple determination coefficient R^2) is evaluated. Regression analysis is tested on normal distribution of residuals, heteroscedasticity, non-linearity and multi-collinearity by plots (Liborius, 2015; Ligges, 2007). Analysis of empirical data ($N = 14$) on $CNDRT$ on the takeover shows significant effects for all takeover processes (Figure 3).

The time until the Gaze is directed to the TOR significantly rises with higher $CNDRT$ ($\beta = .004, p < .01$). The complexity of the NDRT explains 11.3% percent of variance ($R^2 = .113, t(68) = 2.943, p < .01$).

The effect of $CNDRT$ on the time until the NDRT is stopped ($\beta = .0005, p < .001$), explains 16.4% of variance ($R^2 = .164, t(68) = 3.652, p < .001$). Variance in time until the hands are moved to the steering wheel can be explained with 29,07% ($R^2 = .2907, t(68) = 5.297, p < .001$). The time increases significantly ($\beta = 1.47\text{e}-06, p < .001$) with more complex NDRTs. $CNDRT$ also influences the time until the gaze is moved to the road ($\beta = 3.05\text{e}-05, p < .001$). 22.7% of variance can be resolved ($R^2 = .227, t(68) = 4.469, p < .001$). The results show, that the complexity of the NDRT has a significant impact on all four steps of the takeover that were measured empirically (Figure 3). The more complex the NDRT that is performed before the takeover, the longer do drivers need to perform the takeover steps. This shows, that more cognitive occupation during the NDRT occupies relevant resources that need to be freed in order to attend and process objects, that are relevant for the takeover. The more complex a non-driving related task

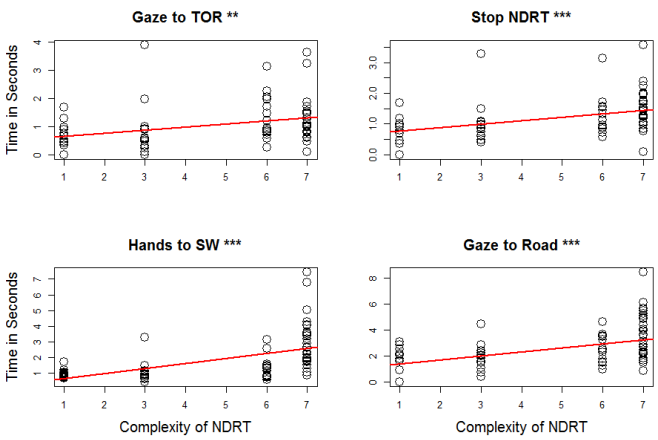


Figure 3: Regressions of the Influence the Complexity of the Non-Driving-Related Task (CNDRT) has on Times of Takeover Patterns (Significance codes: 0 '***' .001 '**' .01 '*' .05 '.' .1; Source: own figure).

is, the longer do drivers need to complete the outlined steps for taking over the driving.

Step 2: Correlation between Model Predictions and Empirical Results of Different Objective Complexities

Further, model predictions under different complex traffic conditions are tested against the results found in empirical data ($N = 14$), that are described above. The comparison of empirical takeover times with different surrounding traffic conditions and model predictions shows, that model predictions correlate with the empirical data for almost all situations of objective complexity (amount of vehicles) significantly (Figure 4). Empirical data (gray) was evaluated for situations with zero to five vehicles in the surrounding traffic. Mean (pink) and median median (red) courses for empirical data are evaluated and median courses correlated with model predictions (green-dotted). For each traffic conditions, model predictions correlate with median values of empirical data. Especially with one, three and four vehicles in the environment, model predictions are in line with empirical data.

Step 3: Test whether Model Predictions of different Complex Traffic Environments show Significant Differences

Finally, predictions of the action decision (9, see section Cognitive Model) of the model are evaluated based on objective complexity measures. In the interaction of the model with different driving situations, it can be shown, that the time for an action decision increases with a more complex driving environment. Overall the model is run through 17 different complexity situations ($N = 17$), varying between zero to five vehicles in the driving environment. The time until an action decision is executed ranges from 1.37s to 4.86s ($M = 1.74$). Regression analysis results in significant regressions for the overall amount of vehicles in the environment ($\beta = 0.04, p < .05$). The parameter resolves 24.91% of variance ($R^2 = .25, t(15) = 2.23, p < .05$). Regarding the vehicle distributon in detail, it can be shown that the amount of vehicles on the right lane has a significant impact on the time until an action decision ($\beta = 0.04, p < .05$). 18.87% of variance ($adj.R^2 = .19, t(14) = 2.3, p < .05$) can be explained. Also the amount of vehicles on the left lane has a small impact on the time until an action decision is made ($\beta = 0.09, p < .1$), explaining 12.13% of variance ($adj.R^2 = .12, t(14) = 1.83, p < .1$). Neither for the vehicle in the front of the ego vehicle a significant impact can be shown. Nor the speed (faster/slower) in relation to the own position has an impact. This shows, that the perception of left and right lane (5), the completion of the perception phase (7) and the formation of characteristics and recognition of the current status (8) need more time in more complex driving environments and lead to a delay of the action decision (9).

Discussion

The results show, that the complexity of the NDRT has a significant impact on the time of takeover patterns in empirical data. It can thus be concluded, that more complex tasks that are done during the automated drive lead to longer takeover times. Portraying the processing patterns that are undergone during the takeover with a cognitive model, similar time trajectories can be shown. This is very important, as results show, that not only the overall time, but also processing steps can be identified and displayed in the model. Further, the model is run through situations with differently complex traffic situations (amount of relevant vehicles). Results show longer times for the processing patterns in more complex environments. The predicted time-lines of the cognitive model are compared to results in empirical data with respect to the traffic complexity. Model predictions correlate with empirically gathered trajectories in differently complex traffic environments. In addition, predictions of the cognitive model are tested on significance in differences between traffic complexities. It can be shown, that the traffic complexity (amount of relevant vehicles) has a significant impact on the time until an action decision is made. These results indicate, that the objective complexity of the NDRT as well as of the traffic situation play an important role concerning processing steps during a takeover in highly automated driving. The takeover behavior as well as the time until an action decision is made, show significant influences of complexity measures (NDRT and traffic environment). Still, the model is slightly faster in the overall performance (0.5 seconds). Since the difference already occurs at the first

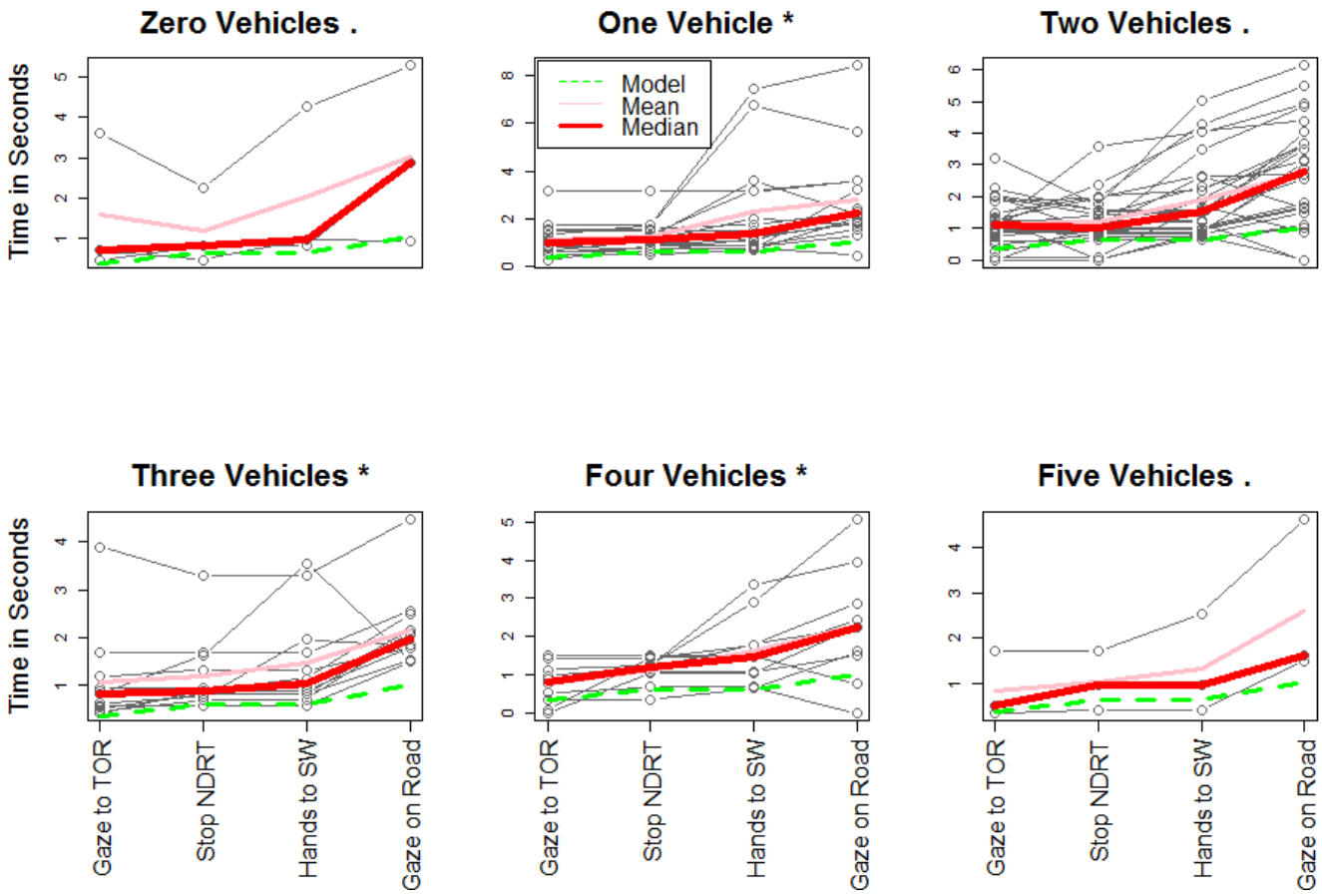


Figure 4: Correlations between Empirical Values and Model Predictions in Different Complex Traffic Situations (Significance codes: 0 '***' .001 '**' .01 '*' .05 '.' .1; Source: own figure).

processing step (Gaze to TOR), it is assumed, that cognitive processes before the gaze is directed to the TOR already have an influence. In the remaining sequence no noticeable time differences are observed. Thus, cognitive processes before the gaze is directed to the TOR have to be included into the model. Further, more aspects of the objective complexity have to be incorporated (e.g. notifications in the HMI, relevance of colors). It is though necessary to investigate on complexity measures concerning the takeover and incorporate further aspects of objective complexity. For an efficient development of interaction devices and estimates in highly automated driving cognitive models are important. They uncover underlying processes and should guide the development of highly automated driving. In this study, empirical data was collected in real traffic. The advantage of this is the creation of a more realistic scenario. However, traffic situations were not controllable and action decision patterns could hence not be evaluated. A simulator study in which the traffic conditions at the moment of the takeover request are controllable will thus be executed. This enables the collection of action decision parameters. The action decision is unequal to the action execution, as the decision may take place before the execution is possible due to the traffic environment. Thus, model predictions of the action decision in different complex situations can be validated by empirical data. In further investigations it will also be important to focus on subjective complexity in addition to objective complexity measures to include the individual into predictions. This is a very important factor, as only the consideration of individual differences enables a suitable, adaptable and safe development of the human machine interface. In order to focus on subjective complexity measures validly, it is though necessary to completely understand and control the objective complexity to separately carry out result analysis for subjective complexity measures.

Conclusion

Results of this study provide a first understanding of the impact of objective complexity on the takeover task. In a next step, action decision mechanisms in dependance of the objective complexity will be gathered. These will

be incorporated to further investigate in the subjective complexity of participants during a takeover. Additionally, steps that are undertaken during the takeover will be differentiated more detailed. Patterns like action decision, action execution and the quality of the takeover and of the action execution should be included. Later, subjective complexity measures will be addressed, to additionally select model predictions based on the individual.

Acknowledgements

I wish to acknowledge the help provided by my supervisors Kathrin Zeeb and Michael Schulz at Robert Bosch GmbH and the public promoted projects PAKoS and Ko-HAF.

References

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological review*, *111*(4), 1036.

Baumann, M., & Krems, J. F. (2007). Situation awareness and driving: A cognitive model. *Modelling driver behaviour in automotive environments*, 253–265.

Haerem, T., & Rau, D. (2007). The influence of degree of expertise and objective task complexity on perceived task complexity and performance. *Journal of Applied Psychology*, *92*(5), 1320.

Ko-HAF. (2017). Ko-haf - wizard-of-oz-konzept. YouTube. Retrieved from https://www.youtube.com/watch?v=4mm3xaBfQZc

Kahneman, D. (1973). *Attention and effort*. Citeseer.

Paxion, J., Galy, E., & Berthelon, C. (2015). Overload depending on driving experience and situation complexity: Which strategies faced with a pedestrian crossing? *Applied ergonomics*, *51*, 343–349.

Plavsic, M. (2010). *Analysis and modeling of driver behavior for assistance systems at road intersections* (Doctoral dissertation, Technische Universität München).

Radlmayr, J., Gold, C., Lorenz, L., Farid, M., & Bengler, K. (2014). How traffic situations and non-driving related tasks affect the take-over quality in highly automated driving. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 58, *1*, pp. 2063–2067). Sage Publications Sage CA: Los Angeles, CA.

Rommerskirchen, C., Helmbrecht, M., & Bengler, K. (2013). Increasing complexity of driving situations and its impact on an adas for anticipatory assistance for the reduction of fuel consumption. In *Intelligent vehicles symposium (iv), 2013 ieee* (pp. 573–578). IEEE.

SAE, T. (2014). *Surface vehicle information report. taxonomy and definitions for terms related to on-road motor vehicle automated driving systems.* SAE International.

Salvucci, D. (2006). Modeling driver behavior in a cognitive architecture. *Human factors*, *48*(2), 362–380.

Salvucci, D., Boer, E., & Liu, A. (2001). Toward an integrated model of driver behavior in cognitive architecture. *Transportation Research Record: Journal of the Transportation Research Board*, (1779), 9–16.

Schlindwein, S. L., & Ison, R. (2004). Human knowing and perceived complexity: Implications for systems practice. *Emergence: Complexity and Organization*, *6*(3), 27–32.

Sun, B., Deng, W., Wu, J., Li, Y., Zhu, B., & Wu, L. (2017). Research on the classification and identification of driver's driving style. In *Computational intelligence and design (iscid), 2017 10th international symposium on* (Vol. 1, pp. 28–32). IEEE.

Wickens, C. D. (2008). Multiple resources and mental workload. *Human factors*, *50*(3), 449–455.

Technology-Based Cognitive Enrichment for Animals in Zoos: A Case Study and Lessons Learned

Benjamin Scheer (benjamin.j.scheer@vanderbilt.edu)[1]
Fidel Cano Renteria (fidel1@mit.edu)[2]
Maithilee Kunda (mkunda@vanderbilt.edu)[1]
[1]Electrical Engineering and Computer Science, Vanderbilt University, Nashville, TN, USA
[2]Mathematics, Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA

Abstract

Cognitive enrichment for captive animals is the idea that cognitive stimulation can improve animal welfare. In zoos, cognitive enrichment not only helps the animals themselves but also contributes to zoo missions of educating the public, supporting research, and more. Technology-based cognitive enrichment tools are increasingly popular for a variety of reasons, though they also present unique challenges for design and deployment. In this paper, we present a short review of technology-based cognitive enrichment programs in zoo settings, and then describe the design and development process we used to create a new, touchscreen-based enrichment app for a group of orangutans at Zoo Atlanta. We discuss initial observations about the orangutans' use of this app, as well as lessons learned by our research team.

Keywords: animal-computer interaction (ACI); comparative cognition; interactive technology; user-centered design.

Figure 1: An orangutan uses the video activity in our app on a touchscreen at Zoo Atlanta.

Introduction

Zoos worldwide, which welcome nearly 200 million visitors annually, provide important scientific, economic, and educational benefits to the public through their captive animal care and management programs (*Zoo & Aquarium Statistics*, 2018). These programs are increasingly paying attention to the cognitive health of captive animals, in addition to their physical health, as a critical dimension of animal well-being.

In nature, animals experience many cognitive challenges requiring attention, search, memory, etc., often in situations directly related to their survival. Examples include foraging, reasoning about social dominance relationships, detecting predators/prey, and so on. Many animals in captivity do not experience the same kinds of challenges, since they are guaranteed food, water, territory, safety, and a social group.

Cognitive enrichment for captive animals is the idea that cognitive stimulation can improve animal welfare (Meehan & Mench, 2007), not just in zoos but in other settings as well, such as farm/livestock facilities (Manteuffel, Langbein, & Puppe, 2009). In zoos, cognitive enrichment programs can provide additional benefits beyond those for the animals themselves. Zoos generally have a mission of educating the public about animals, and cognitive enrichment can be both a point of connection with visitors as well as a topic for informal science education, for instance to engage the public around issues of cognitive health, the role of play and challenge in mental development, and more. In addition, zoos are often sites for important research in comparative psychology and anthropology, and cognitive enrichment both helps to maintain a healthy animal population for researchers and also provides platforms for conducting cognitive research.

Cognitive enrichment approaches that use technology are becoming increasingly popular in zoos, and have both advantages and disadvantages. Technology-based enrichment programs often require significant up-front investments in hardware acquisition and software development, especially when compared to non-technology-based enrichment tools like physical toys. However, if designed properly, these technologies can be reusable and extensible for continued enrichment activities, and furthermore, hardware costs for consumer-grade devices are continually decreasing. On the other hand, hardware can often be damaged by animals, technology-based activities may not be "realistic" to animals, and, just as with people, there may be harmful effects from animals spending too much time using screens.

Parallels can be drawn between studying technology usability for animals and for humans. The emerging field of animal-computer interaction (ACI), like human-computer interaction (HCI), emphasizes the development of systematic, user-centered design and evaluation practices for interactive technology applications (Mancini, 2011).

In this paper, we briefly review technology-based cognitive enrichment in zoos, and then describe the design process we used to create a new, touchscreen-based enrichment app for a group of orangutans at Zoo Atlanta. We discuss initial observations about the orangutans' use of this app, as well as lessons learned by our research team.

Cognitive Enrichment Using Touchscreens

In this section, we review and discuss a sampling of previous studies that used technology-based approaches for cognitive enrichment for captive zoo animals, with a focus on touchscreen-based applications.

Cognitive Enrichment Versus Cognitive Research

Many interactive technologies used with captive animals involve applications designed for research purposes. For example, many of the orangutans in our case study in Zoo Atlanta already have experience using touchscreens through cognitive psychology experiments that have studied capabilities like conspecific face recognition, working memory, etc. Often, the goal of a research application is to answer a specific cognitive science question, but the goal of cognitive enrichment is to provide enrichment. Secondarily, enrichment applications may, of course, also provide data that are interesting for cognitive research, but that is not the primary goal.

Thus, while the design of cognitive enrichment applications may be motivated by cognitive observations about a species, the applications themselves may be more open-ended or complex than those developed for research. Training animals to perform a task may also be less important in enrichment, as tasks are often designed to draw upon an animal's intrinsic curiosity and motivation. In this vein, cognitive enrichment applications do not always require the use of food rewards to motivate animals (though some do). Instead, enrichment can provide animals with entertainment, challenges, and a sense of control over their environment.

Research on the effectiveness of cognitive enrichment applications is critical (Weed & ONeill-Wagner, 2015), which brings up questions about how to measure the degree to which an animal is "enriched" by engaging in a set of activities. Usage or participation in the activity is one measure, but generally, the goal for enrichment is to provide benefits of a more holistic kind. Studies have used both qualitative and quantitative measures of observed animal behaviors to estimate different aspects of the overall "well-being" of an animal, such as, for example, reducing stereotypies or aggression, or increasing play behaviors, exploration, or social interaction (Alligood & Leighty, 2015).

Enrichment and Animal Welfare

Negative effects of typical zoo animal environments can stem from limitations in physical space, diversity of activities, lack of problem solving challenges, and even the withdrawal of rewards from previously entrenched reward-based activities. Minor mental challenges like puzzle-solving can therefore be positive for animal welfare, especially if the challenges are at an appropriate difficulty level. Benefits can include reduced anxiety, increased learning abilities, improved physical condition, more resistant immune systems, faster recovery from illness, and less fearfulness in new scenarios.

Although enrichment benefits animals the most when they are provided it from birth, it appears to have measurable benefits even if introduced later (Millar, 2013). Given the limited flexibility of many enrichment strategies, their utility is often limited by a lack of challenge or the inability to provide a lasting sense of control for animals, making them prone to habituation or frustration. Thus, what is needed is not just the introduction of one-off, ad hoc enrichment activities, but rather the development of flexible tools that support continually evolving, diverse enrichment programs.

Touchscreens and other technology-based applications have often been examined with respect to their effects on animal welfare outcomes. In some studies, technology-based enrichment approaches were found to have some aversive effects on animals (Ritvo & MacDonald, 2016; Tarou, Kuhar, Adcock, Bloomsmith, & Maple, 2004; Elder & Menzel, 2001). However, many other studies have found beneficial effects such as reduced negative behaviors like frustration, and more (see examples in Table 1).

Touchscreen applications have several practical advantages. Because digital enrichment can be dynamic and flexible, it can provide a breadth of activities and be customized to the needs of individual animals or groups of animals (Kim-McCormack, Smith, & Behie, 2016; Boostrom, 2013). This makes this technology more resistant to habituation when compared to traditional enrichment. This form of enrichment may also require little-to-no training for the animals. Touchscreen applications can also be useful in situations when animals are unable to be on exhibit because of inclement weather, injury, or group management.

Designing Applications for Non-Human Animals

When considering the design of an application for a different species, the consensus has been to begin with a user-centered approach (Wirman, 2013; Kim-McCormack et al., 2016; Wirman, 2014; Boostrom, 2013; Péron et al., 2012; Dolins, Schweller, & Milne, 2017). In the case of animal users who cannot directly provide input or feedback, zookeepers and other domain experts are critical resources for informing the design of new applications.

Studies have shown that applications with auditory and visual components along with frequent opportunities for touch interaction have been seen to have the highest interaction times overall, and in terms of content in the application, content displaying photorealistic images has been preferred over 2D graphics (Boostrom, 2013; Wirman, 2014). Having an application that provides immediate response to physical action can be rewarding because of the sense of control (Kim-McCormack et al., 2016). When creating graphics that encourage interaction from primates, designers of applications for primates have heavily focused on small details, such as thickness of borders around content, decisions about colors that will stand out against a background, the sizes of graphics, and more (Péron et al., 2012; Dolins et al., 2017; Wirman, 2013, 2014). One such program that focused on graphics presented to four chimpanzees a simple training regimen with thick, wide green borders along the four sides of a square against a black background (Dolins et al., 2017).

Designers have also considered the touchscreen itself, in-

Table 1: A sampling of the literature on technology-based cognitive enrichment for captive animals.

Reference	Species	# Individuals	Technology	Frequency/Duration	General Findings
Boostrom, 2013	Orangutans	16	iPad	5 min sessions at least twice per month per individual, over a span of 6 months.	There was varying interest in the iPad among the groups, with all groups showing a preference for brightly colored applications that also provided auditory stimulation.
Elder & Menzel, 2001	Orangutans	1	Computer with joystick	33 test days over a total study period of 90 days.	Extended periods of delay between trials induced signs of frustration, but stress was not induced by task performance.
Gray et al., 2018	Gorillas	7	Objects with IoT	Two 60-minute sessions.	Technology can help us learn about and tailor playful experiences for gorillas.
Martin & Shumaker, 2018	Orangutans	12	Touchscreen	Single 20 min session.	The versatility and programmability of computers tasks makes them an ideal platform for achieving functionally naturalistic outcomes for great apes.
Millar, 2013	Pigeons, dogs	16, 58	iPad	10-min sessions for 10 days (pigeons). Various number of 10-min sessions (dogs).	Both cognitive and physical enrichment were found to reduce agonistic behaviour and increase alertness.
Mueller-Paul et al., 2014	Tortoises	4	Touchscreen	Tested five days a week from 9 am to 5 pm.	Red-footed tortoises could operate a touchscreen and solve a spatial task.
Perdue et al., 2012	Orangutans	4	Touchscreen	Random 30-minute observation periods.	Touchscreen technology had no negative effects on the animals.
Péron et al., 2012	Parrots	3	Touchscreen	Always available; each piece of music lasted 90 seconds when played.	Music can be used as an environmental enrichment for captive parrots, and musical preferences seemed to be influenced by personality.
Ritvo & MacDonald, 2016	Orangutans	3	Touchscreen	30-60 min sessions, once per day for 3-4 days per week.	Musical stimuli were not reinforcing; use of music as enrichment may be more aversive than enriching for some species.
Scheel, 2018	Orangutans	11	Touchscreen	There were 10 random observations sessions.	Overall, the orangutans appeared to have enjoyed the touchscreen.
Schmitt, 2018	Gorillas, chimps, orangutans	5, 4, 4	Touchscreen	Available about 45 minutes per day, 3 to 5 times per week.	The ZACI system proved to be highly applicable for work with zoo-housed primates.
Tarou et al., 2004, Mallavarapu et al., 2013	Orangutans	8	Computer with joystick	1 hour sessions, for a total of 240 hours.	Behavioral changes associated with the computer included increases in aggressive behavior and more. The lack of habituation by frequent users indicates that computer-assisted tasks may be useful environmental enrichment for orangutans.
Wirman, 2012, 2013, 2014	Orangutans	2	Tablet-based touchscreen	Random, short durations of play with the touchscreen.	Digital play allows a form of communication that eliminates some obstacles and creates new ways for togetherness in play.

cluding viewing angle, software and hardware specs, and input mechanisms (Wirman, 2014), as well as expected physical usage. For example, orangutans often sit in an upright position to use a touchscreen, similar to human behavior.

Just as with the design of technologies for people, design for non-human animals requires significant creativity and imagination on the designers' part. Furthermore, non-human animals may interact with applications in unexpected ways, despite a designer's best-laid plans. Thus, early prototyping and "user testing" is key.

Our Case Study

We developed an enrichment application intended for use by the eleven orangutans (age 3-48 years) at Zoo Atlanta in Atlanta, GA, USA.

The zoo that our team worked with has an existing, somewhat unique technology installation in one of their open-air orangutan enclosures: an artificial "tree" (fiberglass, etc.) that has a touch-screen monitor built into one face, and houses a desktop computer inside (see Figure 2). The intent of this installation was to provide a platform for cognitive enrichment for the orangutans that could be used in a relatively unstructured way. The tree had previously been loaded with applications designed for comparative psychology research.

In conversations with our team, zookeepers stated their desire for a new "app" for the tree that would: 1) be easy enough for the orangutans to use without oversight from staff members; and 2) be engaging enough for the orangutans that they might choose to use it without extrinsic rewards (e.g., food).

In addition to these criteria, we added two more from the software development side: 3) be easily extensible by zookeepers to add/modify content to individual app activities; and 4) be modular and thus easily extensible by future developers to add/modify individual activities within the app.

Figure 2: "Learning Tree" touchscreen installation.

App Design and Development

Through many discussions with four zookeepers over a period of two and a half years, we designed a cognitive enrichment application that consisted of an "activity chooser" home screen, from which the orangutans would be able to select individual activities to engage in. Initially, we designed three modular activities to populate the system: 1) a video player activity, 2) a visual puzzle activity, and 3) a musical instruments activity.

A primary concern throughout the design process was aiming to ensure that the interface and individual activities would be simple enough for use by the orangutans. Often, such enrichment app designs overestimate the level of complexity that animals can understand in terms of interface and task design. In conversations with other zoos that attempted similar technology-based enrichment efforts with non-human primates, we learned that **simplicity** of the interface and **familiarity** of the elements presented to animals were both important design factors to keep in mind (McAuliffe, 2017).

Home screen. As illustrated in Figure 3 (left), the application's home screen holds an array of orangutan images, and a vertical green home button. Contrary to many common formats of reward-based applications, where the orangutans have to figure out where to press or what to do in order to get food, this design seeks to draw the animal's attention by showing them images they will recognize: familiar orangutans from their own social group. Research has shown that orangutans, while not among the most social of non-human primates, still do show fairly robust conspecific (e.g. within-species) visual recognition of familiar faces (Hanazuka et al., 2013; Talbot et al., 2015). Each cell with an orangutan image leads to one of three different activities.

The home button appears on all screens of the app and will always return the orangutan to the original home screen. The button appearance was not designed arbitrarily; the green gradient was featured in the home buttons of other reward-based applications that the orangutans had already been using. As a result, the familiar pattern attempts to give the orangutans visual clues about the button's function.

Video player activity. The first of the activities is illustrated in Figure 3 (right side, top row). Pressing any of four brightly colored boxes triggers a short video clip (15-30 seconds) of one of the zoo's orangutans, taken from the zoo's existing store of videos. The layout was designed to be simple and visually distinguishable from the other screens. Again, zookeepers thought that showing videos of individuals familiar to the orangutans would be engaging.

Visual puzzle activity. The second activity is the simple visual puzzle illustrated in Figure 3 (right side, middle row). We wanted to create an activity a bit more challenging than the passive-viewing video player activity, but also simple enough to be solvable by most of the orangutans fairly quickly, especially in their initial exposure to the app.

Thus, we created a design in which puzzle "pieces" of an image are shown around the perimeter of the screen, with a target grid in the middle. Pressing any of the puzzle pieces prompts the piece to move on its own to the correct grid location, greatly reducing the difficulty of the task while also providing some visual interest. Once all four pieces have been pressed and are in position, the completed image then plays as a video (puzzle images are taken from the first frames of video clips), providing a type of "visual reward" for completing the puzzle. As with the previous screens, images were chosen from the zoo's stock of photographs of their own orangutans, to facilitate interest through familiarity.

Musical instruments activity. Finally, as illustrated in Figure 3 (right side, bottom row), we created an activity that displays an array of eight different musical instruments which play an audio clip (1-30 sec) of their corresponding sound when pressed. While the recording is playing, the selected instrument icon also oscillates in place to emphasize the connection between the button and the sound.

Previously cited research suggests that musical stimuli is not reinforcing as orangutan enrichment, but we believed that research on musical enrichment is minimal enough to explore further. Additional cited research proposes that orangutans tend to be more interested in photorealistic images than graphics, but through conversations with zoo researchers, we concluded that this information is not as essential for images of instruments as it is for images and videos of orangutans contained in the other activities.

Modularity of design. Though our design choices primarily strive to achieve simplicity of use, it is also important to keep activities somewhat novel when designing for orangutans in order to prevent boredom and frustration (Wirman, 2013). Therefore, one important aspect of modular design in our app relates to the video content used for the video player activity and for the visual puzzle activity. Videos are drawn from a specific folder, and zookeepers can easily change the available videos by adding or removing video files from the library. In addition, the image/video used for

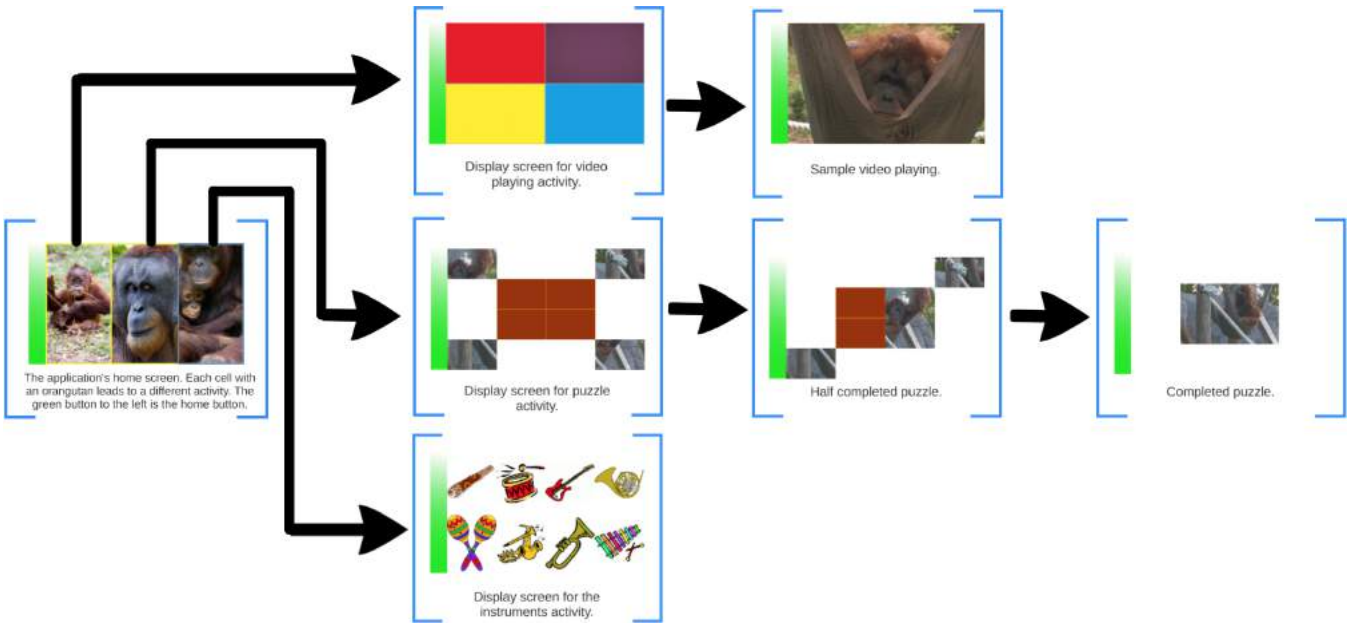


Figure 3: Flow diagram of the touchscreen-based application we designed to provide cognitive enrichment for orangutans.

the puzzle activity is randomly chosen from this store, to add some novelty to that activity over repeated sessions.

Because some orangutans become familiar with touchscreen activities faster than others, the goal was to design activities that were, on average, challenging, but not frustrating. Thus, an expectation was that some orangutans would learn activities more quickly than others, and that many orangutans would continue to perceive activities as novel for a substantial amount of time.

One other modular design choice is that each individual activity resides in its own "container." Thus, activities can be added or swapped in a relatively straightforward fashion. In conversations with zookeepers, one common issue with enrichment apps seems to be the lack of ease of extensibility, especially given that it is often difficult to access software developers to work on extensions or modifications.

Finally, log files are saved from each session and hold a timestamped record of every activity performed in the app. The information in these logs is valuable for understanding orangutan usage patterns, and perhaps inferring measures of orangutan amusement and satisfaction, as potentially important components of overall cognitive enrichment.

Ideally, additional data would be collected to establish a durable record of which individual orangutans were using the app at various times. For example, we discussed with zookeepers the potential value of having a webcam-like setup that would record video of the orangutan user every time the app was activated. Such a video would provide not only identifying information about users but potentially also information on the user's affect and engagement levels. However, due to logistical constraints, we were not able to deploy such a setup. As a result, our log files record usage but not which individual or individuals were using the app.

Observations and Lessons Learned

Initial deployment of the app with the group of orangutans at the zoo seems to show many positive signs. Based on our team's qualitative observations of the orangutans, they seemed curious and interested in what was happening on the touch screen, often rushing over in a group to see what was happening when zookeepers opened the app on the learning tree computer. Several individual orangutans were also observed at various times interacting with the app for moderately lengthy durations.

Figure 4 shows ten examples of orangutan interactions with the app, as recorded in the system logs. These ten sessions, shown on the y-axis in no particular order, were chosen from the full set of log data to show a sampling of interaction patterns that were observed. The x-axis shows time across a duration of about 24 minutes, measured from the first touch screen press that was recorded by the app during a given session.

Clearly, there is a lot of variability in usage. Sometimes (e.g., logs 2, 8, and 9), there is some initial activity that quickly tails off within a minute or two. Other sessions show much more sustained activity. The orangutans seem to have accessed each of the app's activities more than once, though the extent to which they are purposefully navigating through the app, versus just pushing various buttons, is an open question. More detailed analyses of such log files will be an important part of our future work.

In addition to the log data, we also discuss, in qualitative terms, two episodes of interaction that were particularly noteworthy. First, one of the most interesting moments occurred when Madu, a female orangutan in her 30s, was viewing videos through the video player activity. (The image in Fig-

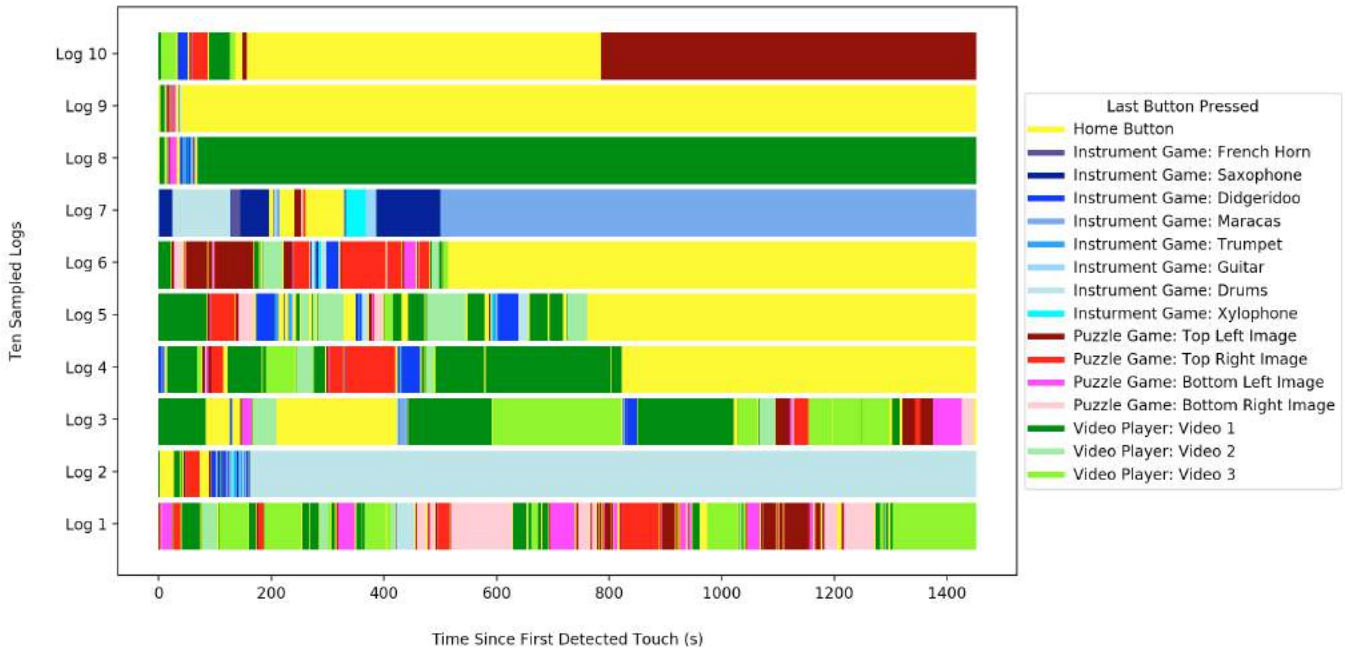


Figure 4: Sampling of app log data from 10 different interaction sessions (y-axis) across time, as measured from first touch screen press (x-axis).

ure 1 is a still from this episode.) She encounters a video of Alan, an older male orangutan who had passed away nearly four years earlier. Madu is visibly transfixed for 20 long seconds, and then tries to interact further by touching his figure in the video several times. Of course, we do not know for sure what was going on in Madu's mind at the time, but at least to the zookeepers who know her best, it seemed like she was remembering her old habitat neighbor. Note that we did not include videos of Alan on purpose; we were simply pulling from the zoo's stock of orangutan videos. However, this incident does seem to support the idea that familiar stimuli can be uniquely engaging to orangutans, even (or perhaps especially) in the case of dearly departed old friends.

Another time, one of the orangutans was interacting with the app and seemed to be searching for a food reward, looking up at the feeder mechanism on the tree (used in other, reward-based applications), tapping on the tree, and even banging on the touchscreen with a fist. The orangutan appeared to be quite frustrated at the absence of a food reward! Despite this episode, however, several of the orangutans at other times did find the app interesting enough to be worth engagement, even without food rewards. This raises interesting questions about the longer-term impacts on learning and motivation of using food rewards to stimulate certain behaviors.

We conclude with three takeaway lessons from our case study of cognitive enrichment for captive orangutans.

Familiarity of stimuli. The use of photos and videos of familiar orangutans did seem to support interest and engagement in the orangutans using our app. Further evaluations of this principle would be extremely valuable for cognitive enrichment programs in general, including at other zoos, with other species, and in a variety of enrichment activities.

Modularity in design. The modular design that we implemented, especially in terms of enabling zookeepers to easily swap out image/video content without requiring any programming skills, seems to be a promising approach to enable technology-based enrichment activities to be modifiable. Studies of orangutan engagement over time will be informative, and we expect that the ability to regularly change app content will keep the novelty factor up.

Engagement with the public. There is continued debate on the pros and cons of adding technology to the daily experiences of zoo animals, not only with respect to the animals themselves but also with respect to the public education missions that most zoos have. Is it really teaching the public the right ideas about wildlife to show orangutans playing "video games?" There is no easy answer to this question, but Zoo Atlanta has aimed to strike a balance by establishing a "naturalistic" setting for their technology-based enrichment—the learning tree shown in Figure 2.

These applications can also serve to teach the public about the cognitive health of captive animals, and to showcase joint efforts by zookeepers and researchers to ensure that animals have a stimulating cognitive environment. The "coolness" factor of technology-based interventions may provide positive benefits for engaging the public (Perdue et al., 2012); for example, the video of the Madu-Alan episode described above was viewed on the zoo's social media page over 15 thousand times. In addition to media influence, zoo visitors can participate in weekly zookeeper-led showcases of

the orangutans using the touchscreen in their exhibit, through which they have the opportunity to ask questions about orangutan behavior with the touchscreen and view real-time touchscreen interactions on an additional screen display that is located just outside the enclosure.

In summary, while non-technological cognitive enrichment activities also have their place, we expect that technology-based interventions will continue to provide valuable contributions for the study and care of captive zoo animals, as well as for basic research in cognitive science.

Acknowledgment

This research study was approved by the Zoo Atlanta Scientific Review Committee. Many thanks to Laura Mayo for her original suggestions on creating this enrichment application, and also to Liam Kelly, Avery Twitchell-Heyne, Jodi Carrigan, Marieke Gartner, and others at Zoo Atlanta for their contributions to this work.

Thanks also to the following individuals who worked on initial app design and implementation at the Georgia Institute of Technology in 2014-2015: Benjamin Seco, Katie Delisle, Brett Garcia, Samantha Kassem, Brandon Levester, Shannon Nguyen, Jee Kong, Andrew Scheinbach, George Zheng, and Weng Ling Chen.

References

Alligood, C., & Leighty, K. (2015). Putting the e in spider: Evolving trends in the evaluation of environmental enrichment efficacy in zoological settings. *Animal Behavior and Cognition*, *2*(3), 200–217.

Boostrom, H. (2013). *Problem-solving with orangutans and chimpanzees: Using the iPad to provide novel enrichment opportunities.* M.S. Thesis, Texas A&M University.

Dolins, F. L., Schweller, K., & Milne, S. (2017). Technology advancing the study of animal cognition: using virtual reality to present virtually simulated environments to investigate nonhuman primate spatial cognition. *Current zoology*, *63*(1), 97–108.

Elder, C. M., & Menzel, C. R. (2001). Dissociation of cortisol and behavioral indicators of stress in an orangutan during a computerized task. *Primates*, *42*(4), 345–357.

Gray, S., Bennett, P., Burgess, K., Cater, K., & Clark, F. (2018). Gorilla game lab: Exploring modularity, tangibility and playful engagement in cognitive enrichment design. In *5th international conf. animal computer interaction.*

Hanazuka, Y., Shimahara, N., Tokuda, Y., & Midorikawa, A. (2013). Orangutans (pongo pygmaeus) remember old acquaintances. *PloS one*, *8*(12), e82073.

Kim-McCormack, N., Smith, C., & Behie, A. (2016). Is interactive technology a relevant and effective enrichment for captive great apes? *Applied animal behaviour science*, *185*, 1–8.

Mallavarapu, S., Bloomsmith, M., Kuhar, C., & Maple, T. (2013). Using multiple joystick systems in computerised enrichment for captive orangutans. *Animal Welfare*, *22*(3), 401–409.

Mancini, C. (2011). Animal-computer interaction: a manifesto. *Interactions*, *18*(4), 69–73.

Manteuffel, G., Langbein, J., & Puppe, B. (2009). From operant learning to cognitive enrichment in farm animal housing: bases and applicability. *Animal Welfare*, *18*(1), 87–95.

Martin, C., & Shumaker, R. (2018). Computer tasks for great apes promote functional naturalism in a zoo setting. In *5th international conference on animal computer interaction.*

McAuliffe, J. (2017). *Touchscreen-based enrichment for chimpanzees at the Houston Zoo.* Private communication.

Meehan, C., & Mench, J. (2007). The challenge of challenge: can problem solving opportunities enhance animal welfare? *Applied Animal Behaviour Sci.*, *102*, 246–261.

Millar, L. (2013). *Improving captive animal welfare through the application of cognitive enrichment.* PhD Thesis, University of Exeter.

Mueller-Paul, J., Wilkinson, A., Aust, U., Steurer, M., Hall, G., & Huber, L. (2014). Touchscreen performance and knowledge transfer in the red-footed tortoise (chelonoidis carbonaria). *Behavioural processes*, *106*, 187–192.

Perdue, B., Clay, A., Gaalema, D., Maple, T., & Stoinski, T. (2012). Technology at the zoo: the influence of a touchscreen computer on orangutans and zoo visitors. *Zoo Biology*, *31*(1), 27–39.

Péron, F., Hoummady, S., Mauny, N., & Bovet, D. (2012). Touch screen device and music as enrichments to captive housing conditions of african grey parrots. *J. Veterinary Behavior-Clinical Applications and Research*, *7*(6), e13.

Ritvo, S. E., & MacDonald, S. E. (2016). Music as enrichment for sumatran orangutans (pongo abelii). *Journal of Zoo and Aquarium Research*, *4*(3), 156-163.

Scheel, B. (2018). Designing digital enrichment for orangutans. In *5th int. conf. animal computer interaction.*

Schmitt, V. (2018). Implementing new portable touchscreen-setups to enhance cognitive research and enrich zoo-housed animals. *bioRxiv*. doi: 10.1101/316042

Talbot, C., Mayo, L., Stoinski, T., & Brosnan, S. (2015). Face discriminations by orangutans vary as a function of familiarity. *Evolutionary Psychological Science*, *1*(3), 172–182.

Tarou, L., Kuhar, C., Adcock, D., Bloomsmith, M., & Maple, T. (2004). Computer-assisted enrichment for zoo-housed orangutans. *Animal Welfare*, *13*(4), 445–453.

Weed, J., & ONeill-Wagner, P. (2015). Animal behaviour research findings facilitate comprehensive captive animal care: The birth of behavioral management. In *Environmental enrichment for nonhuman primates resource guide.* USDA Animal Welfare Information Center.

Wirman, H. (2012). *A touch screen as encountered by an orangutan.* Poster abstract, Minding Animals Conference.

Wirman, H. (2013). Orangutan play on and beyond a touchscreen. In *Proc. 19th int. symposium of electronic art.*

Wirman, H. (2014). Games for/with strangers: Captive orangutan touch screen play. *Antennae*, *30*, 105-115.

Zoo & aquarium statistics. (2018). https://www.aza.org/zoo-and-aquarium-statistics. Assoc. Zoos & Aquariums.

Capturing Intra-and Inter-Brain Dynamics with Recurrence Quantification Analysis

Rebecca Scheurich (rebecca.scheurich@mail.mcgill.ca)
McGill University, Department of Psychology
Montreal, QC, Canada

Alexander P. Demos (ademos@uic.edu)
University of Illinois at Chicago, Department of Psychology
Chicago, IL, USA

Anna Zamm (zamma@ceu.edu)
McGill University, Department of Psychology
Montreal, QC, Canada
Central European University, Department of Cognitive Science
Budapest, Hungary

Brian Mathias (bmathias@cbs.mpg.de)
McGill University, Department of Psychology
Montreal, QC, Canada
Max Planck Institute for Human Cognitive and Brain Sciences
Leipzig, Germany

Caroline Palmer (caroline.palmer@mcgill.ca)
McGill University, Department of Psychology
Montreal, QC, Canada

Abstract

We investigated the application of non-linear analysis techniques for capturing stability of neural oscillatory activity within and across brains. Recurrence Quantification Analysis (RQA), a technique that has been applied to detect stability and flexibility of motor performance, was extended to observe and quantify changes in patterns of non-linear neural activity. Participants synchronized their finger-tapping with a confederate partner who tapped at two different rhythms while neural activity was recorded from both partners using electroencephalography (EEG). Auto-recurrence (intra-brain) and cross-recurrence (inter-brain) of EEG activity were able to distinguish differences across tapping rhythms in stability of neural oscillatory activity. We also demonstrated the efficacy of RQA to capture how both period and phase changes in neural dynamics evolve over time.

Keywords: joint action; neural dynamics; electroencephalography; recurrence quantification analysis

Introduction

Researchers have become increasingly interested in capturing complex oscillatory signals common to human behaviors, and which often show non-linearities that evolve over time. This can be seen in individual motor behaviors like postural sway and finger-tapping (Schmit, Regis & Riley, 2005; Schmit, Riley, Dalvi, Sahay, Shear, Shockley, & Pun, 2006; Scheurich, Zamm, & Palmer, 2018), and in joint motor behaviors like conversational speech and music performance (Dale & Spivey, 2006; Demos, Chaffin, & Kant, 2014). One way in which these complex signals can be represented is through Recurrence Plots (RPs), which display the points in time at which an individual returns to previous behavioral states (i.e., self-similarity), or the points in time at which two individuals visit the same behavioral state (i.e., similarity between individuals; Eckmann, Kamphorst, & Ruelle, 1987). RPs are useful tools for observing transitions between states in a system and can be further quantified using Recurrence Quantification Analysis (RQA). These quantifications give insights into the behavioral dynamics of one or more systems over time through measures such as recurrence rate: how often a system returns to previous states or two systems visit the same state; and mean diagonal line length: the time over which one or more systems are stable (Marwan, Romano, Thiel, & Kurths, 2007; Marwan & Webber, 2015). One advantage of RQA is that it can be applied both within individuals during solo tasks (i.e., auto-recurrence) and between individuals during joint tasks (i.e., cross-recurrence; Marwan, Romano, Thiel, & Kurths, 2007; Marwan & Webber, 2015). Thus, these tools have been useful for characterizing dynamics of motor behaviors over time both within and across individuals during a variety of solo and joint behaviors (e.g., Schmit, Regis, & Riley, 2005; Schmit, Riley, Dalvi, Sahay, Shear, Shockley, & Pun, 2006; Romero, Fitzpatrick, Schmidt, & Richardson, 2016; Demos & Chaffin, 2017; Scheurich, Zamm, & Palmer, 2018).

Complex oscillatory signals are not unique to behavior, but are also observed in human brain activity. This can be seen, for example, in the oscillatory neural activity that underlies rhythmic auditory-motor behaviors (e.g., Nozaradan, Zerouali, Peretz, & Mouraux, 2013; Nozaradan, 2014; Morillon & Baillet, 2017; Zamm, Debener, Bauer, Bleichner, Demos, & Palmer, 2018). However, common methods for examining oscillatory neural activity supporting these kinds of behaviors often do not measure dynamics over time, but instead assume stationarity of the signal. RQA has been applied to oscillatory neural activity, as measured through electroencephalography (EEG), in a limited scope. This has been primarily in clinical settings, in which outcomes such as recurrence rate and mean diagonal line length, which provide information about the stability of neural activity, have been used successfully to classify periods of epileptics' EEG activity as normal, pre-ictal, and

ictal activity (Acharya, Sree, Chattopadhyay, Yu, & Ang, 2011). Furthermore, RQA outcomes have been applied for monitoring consciousness of patients undergoing anesthesia (Becker, Schneider, Eder, Ranft, Kochs, Zieglgänsberger, & Dodt, 2010). In addition to its clinical applications, researchers have proposed RQA as a method for studying event-related potentials (ERPs). Although traditional methods of studying ERPs require averaging over many trials to obtain a clear waveform, RQA allows for the use of single trials to identify changes in ERP components, as demonstrated in an auditory perception experiment using the auditory oddball paradigm (Marwan & Meinke, 2004). No research, to our knowledge, has yet applied RQA to capture oscillatory neural activity that distinguishes different rhythmic auditory-motor behaviors.

The current study applies RQA to capture the dynamics of oscillatory neural activity during a 2-person rhythmic tapping task. Participants tapped at two different rhythms with a confederate partner while EEG was recorded from each partner. In one rhythm condition, the confederate tapped at twice the frequency of the participant. In the second rhythm condition, the confederate tapped at half the frequency of the participant. The neural activity at the participant's (constant) tapping frequency was compared across rhythm conditions. Only activity at the constant frequency was examined to identify changes in oscillatory neural activity related to changes in tapping ratios between partners as opposed to changes in absolute frequency. Auto- (intra-brain) and cross-recurrence (inter-brain) analyses of EEG activity were expected to reveal greater stability of oscillatory neural activity when the participants' tapping frequency was the dominant frequency (i.e., more auditory feedback at that frequency).

Methods

Participants

Data from eight adult musicians aged 18-30 years old with at least 6 years of private music instruction on an instrument other than percussion were taken from a larger study. Their duet tapping trials met a performance cut-off of at least 75% error-free trials (i.e., no missed taps) for each condition in which partners performed live together. Other conditions included in the larger study in which participants performed with pre-recordings of their partner were not examined in the current paper. A single confederate experimenter (more than 6 years of piano instruction) tapped with each participant to maintain consistent timing properties of live and pre-recorded conditions as well as social presence across participants. All participants and the confederate were right-handed and had normal hearing (< 30 dB HL threshold, 125 – 750 Hz) as determined by an audiometry screening. Participants and the confederate reported no current psychiatric or neurological conditions and were not taking medication affecting the central nervous system at the time of testing.

Equipment and Materials

Participants' hearing was assessed with a Maico MA40 audiometer. Participants tapped on a Roland A500s MIDI keyboard and the confederate tapped on a Yamaha PSR 500m MIDI keyboard. Auditory feedback was delivered in a sine tone timbre generated by a Roland Sound Canvas, amplified to a comfortable listening level using a Behringer Headphone Amplifier, through EEG-compatible earphones (Etymotic ER-1, Etymotic Research Inc.). Participants' auditory feedback was presented at pitch G4 (392.00 Hz), and the confederate's auditory feedback at pitch E5 (659.25 Hz). MIDI data were collected using FTAP software (Finney, 2001). FTAP was modified to integrate Lab Streaming Layer (LSL; Kothe, 2014) similar to Zamm, Palmer, Bauer, Bleichner, Demos, & Debener, 2017. This modification allowed for keystroke, metronome, and time triggers from FTAP on a Dell computer running Linux to be sent over the local area network and received by a second Dell computer running Windows 7, where LSL synchronized the keystroke and EEG data collection from both partners (Zamm et al., 2017).

EEG Data Recording

EEG data were recorded from each partner at a 512 Hz sampling rate via two separate but synchronized 64-channel BioSemi Active-Two systems (BioSemi, Inc.). Electrodes were positioned according to the 10-20 system. Data were recorded using a common mode sense (CMS) active electrode and driven right leg (DRL) passive electrode which formed the reference (http://www.biosemi.com/faq/cms&drl.htm). External electrodes were placed above and below the right eye to detect eyeblinks, on the outer corner of each eye to detect lateral eye movements, and on the mastoids to detect muscle artefacts.

Stimulus Materials and Design

Each stimulus was constructed of an approximately 40-second series of taps generated by the Participant and Confederate. Each pair (Participant and Confederate) completed the joint tapping tasks in a within-subjects design with 2 rhythm conditions: 1-2 (Confederate-Participant) and 4-2 (Confederate-Participant). In the 1-2 condition, the confederate tapped at half the rate (~0.95 Hz) of the Participant (~1.89 Hz). In the 4-2 condition, the Confederate tapped at twice the rate (~3.78 Hz) of the Participant (~1.89 Hz). Thus, the Participants' tapping frequency was constant across conditions. Each pair completed one practice trial and 12 experimental trials in each rhythm condition. Rhythm was blocked within pair, and blocks were counterbalanced across pairs. The dependent variables were auto- (intra-brain) and cross-recurrence (inter-brain) outcomes of Recurrence Rate, describing how much of the RP is occupied by recurrent points (how often a single system returns to previous states in auto-recurrence, or two systems visit similar states in cross-recurrence), and Meanline,

describing the average diagonal line length (the mathematical stability of the system(s); see **RQA Application to EEG**).

Procedure

After giving informed consent upon arrival to the lab, participants completed an audiometry screening. Then both the participant and the confederate were outfitted with EEG caps and electrodes. The participant and confederate were taken to the testing room where the confederate was introduced to participants as an experimenter who served as the partner in each pair to maintain consistency of interactions across pairs. The participant and the confederate were seated at two separate keyboards across from one another with a barrier placed between the keyboards such that the partners could only see one another above the shoulder.

The participant and confederate then completed the two tapping tasks together at the two different rhythmic ratios. They were instructed to tap with the index finger of their right hands on a single key of the keyboard while minimizing eyeblinks and eye movements. The participant and confederate were first presented with separate recorded examples of each tapping part in isolation, and then they were presented with a recorded example of how the two parts sounded together. After listening to the examples, the participant and confederate were instructed that they would hear a four-beat metronome cue sounded at the participant's prescribed rate at the beginning of each trial, and they were presented with a recorded example of how their parts sounded together with the metronome cue. The participants were instructed that they should synchronize with the confederate's tapping while maintaining the rate cued by the metronome, and the confederate was instructed to maintain a steady pulse. After completing a practice trial, pairs completed 12 experimental trials. This procedure was repeated for each rhythm condition. After completion of the tasks, participants were debriefed and received a small compensation. The whole experiment lasted approximately three hours.

EEG Preprocessing

EEG data were preprocessed in EEGLAB (Delorme & Makeig, 2004). Data were first prepared for artefact correction with Independent Component Analysis (ICA), using a procedure adapted from Zamm et al. (2017). Data were concatenated across all trials in all experimental tapping tasks, and re-referenced to the common average across electrodes. Electrodes reflecting poor signal quality were identified by visually inspecting electrode distributions of deviations from mean activity for each subject. Electrodes with very large deviations from mean activity were identified as noisy, and electrodes with no deviation from mean activity were identified as flat. These electrodes were removed, and data were subsequently filtered between 1 Hz and 40 Hz using a Hanning windowed sinc FIR filter (high and low pass filter order = 1000). Filtered data were then segmented into 1-second epochs, pruned for non-stereotypical artefacts, and submitted to extended infomax ICA. ICA components representing eyeblinks and lateral eye movements were visually identified and removed from the unfiltered data. After removing bad components, previously rejected electrodes with poor signal quality were spherically interpolated.

RQA Application to EEG

Power Spectral Density (PSD) estimates of ICA-corrected EEG activity were then computed similar to Zamm et al. (2017). PSD gives the amount of power present in the EEG signal at component frequencies. Preprocessed EEG data were high then low pass filtered using a Hanning windowed sinc FIR filter (high pass filter order = 1000, cutoff = 0.1 Hz; low pass filter order = 1000, cutoff = 20 Hz) and segmented into 3 10.56-second epochs (to control for tapping frequency drift). PSD was estimated for each electrode and epoch, and then was log-transformed before averaging across epochs and then trials. The electrode with maximal power on average across conditions, tapping frequencies, and participants was identified as electrode C1 (central and left-lateralized). This electrode is commonly identified as showing maximal activity in auditory-motor behaviors (e.g., Nozaradan, Zerouali, Peretz, & Mouraux, 2013; Nozaradan, 2014). Data from this electrode were used as input to auto- and cross-recurrence analyses.

ICA-corrected data from electrode C1 for participants and the confederate were then prepared for auto- and cross-recurrence analyses. First, the data were filtered at the participants' observed tapping frequencies. The filter frequency cutoffs were tailored per participant and confederate pair and rhythm condition to account for any deviations in expected tapping frequency. The data were high then low pass filtered using a Hanning windowed sinc FIR filter (high and low pass filter orders = 1000) with cutoff frequencies ± 2 standard deviations around the observed participant tapping frequency. Data were then segmented into 3 10.56-second epochs (for computational tractability) and *z*-scored per epoch.

Auto- and cross-recurrence analyses were run using the Cross Recurrence Plot Toolbox (Marwan, Romano, Thiel, & Kurths, 2007). Optimal auto- and cross-recurrence parameters were determined per epoch; final selected parameters were determined by examining the distribution of parameters across epochs. The optimal delay parameter was determined by computing Average Mutual Information (AMI). AMI gives the amount of information a time series shares with itself at different time delays, with the delays at which it shares least information with itself being optimal for RQA. The first delay at which shared information of the C1 time series with itself reached a minima was selected (selected delay = 68 samples, corresponding to 1/4 cycle of the participant tapping frequency). The optimal number of embedding dimensions was determined by computing False Nearest Neighbors (FNN). FNN gives the amount of false neighbors in phase space as a function of the number of

embedding dimensions (copies of the time series at the specified delay). The number of embedding dimensions at which number of false nearest neighbors was minimized and adding more dimensions no longer reduced number of false nearest neighbors was selected (selected embedding dimensions = 4). Finally, the maximum phase space diameter, corresponding to the standard deviation of the time series, was computed using the selected delay and embedding dimensions. The optimal threshold for which points in phase space are considered recurrent was determined by computing 10% of this value (selected threshold = 0.49; Schinkel, Dimigen, & Marwan, 2008). For auto-recurrence, the Thieler window, minimum diagonal line length, and minimum vertical line length were set to 34 samples (corresponding to 1/8 cycle of the participant tapping frequency). For cross-recurrence, the Thieler window was set to 0 samples and the minimum diagonal and vertical line lengths were set to 34 samples.

Results

Auto-recurrence Outcomes

We first investigated how auto-recurrence (intra-brain) outcomes changed with Rhythm, and whether these patterns held or changed across Partners within each pair. Separate two-way ANOVAs were run on Recurrence Rate and Meanline with Rhythm (1-2 and 4-2) and Partner (Participant and Confederate) as factors and pair as random variable. Results are summarized in Table 1 and sample RPs are shown in Figure 1. There was a significant main effect of Rhythm on Recurrence Rate: Recurrence Rate was higher for the 1-2 Rhythm (in which the participant tapped at twice the rate of the confederate) than for the 4-2 Rhythm. There was no significant main effect of Partner, $F(1, 7) = 0.012$, $p = 0.92$, or significant interaction between Rhythm and Partner, $F(1,7) = 0.415$, $p = 0.54$, on Recurrence Rate. There was also a significant main effect of Rhythm on Meanline: Meanline was higher for the 1-2 Rhythm than for the 4-2 Rhythm. Again, there was no significant main effect of Partner, $F(1,7) = 0.017$, $p = 0.90$, or significant interaction between Rhythm and Partner, $F(1, 7) = 0.582$, $p = 0.47$, on Meanline. These effects were replicated with mixed models in which random effects of Partner and Rhythm were allowed to vary as a function of the pair.

To ensure that the main effect of Rhythm on Meanline was not a function of differences in Recurrence Rate across Rhythms, we also examined the outcome of Meanline when Recurrence Rate was fixed to 10% across Rhythms during the process of computing the RQA. A two-way ANOVA was run on Meanline with Rhythm and Partner as factors and pair as random variable. The main effect of Meanline held when Recurrence Rate was fixed across Rhythms, $F(1, 7) = 17.577$, $p = 0.004$. Meanline was higher for the 1-2 Rhythm than for the 4-2 Rhythm. There was no significant main effect of Partner, $F(1, 7) = 0.001$, $p = 0.97$, or significant interaction between Rhythm and Partner, $F(1, 7) = 0.579$, $p = 0.47$.

Table 1: Auto-recurrence main effects of Rhythm.

Outcome	1-2	4-2	F	η^2	p
Recurrence Rate	3.06%	2.59%	23.03	0.79	0.002
Meanline	136.26	126.44	20.32	0.77	0.003

Figure 1 shows RPs for an example epoch from one participant for each Rhythm. As can be seen in these examples, there are more recurrent points and longer diagonal lines in the 1-2 RP (when the participant's tapping frequency is the dominant performance frequency) than the 4-2 RP. The white space between the diagonal lines on each plot corresponds approximately to the participant tapping frequency (1.89 Hz or approximately 271 samples).

Cross-recurrence Outcomes

Separate one-way ANOVAs were conducted on the same outcome measures (Recurrence Rate and Meanline) from cross-recurrence quantification analysis with Rhythm as factor and pair as random variable. Results are summarized in Table 2 and sample RPs are shown in Figure 2. There was a significant main effect of Rhythm on Recurrence Rate: Recurrence Rate was higher for the 1-2 Rhythm than for the 4-2 Rhythm. There was also a significant main effect of Rhythm on Meanline: Meanline was higher for the 1-2 Rhythm than for the 4-2 Rhythm. These effects were replicated with mixed models in which random effects of Rhythm were allowed to vary as a function of the pair.

To again ensure that the main effect of Rhythm on Meanline was not a function of differences in Recurrence Rate across Rhythms, we also examined the outcome of Meanline when Recurrence Rate was fixed to 10% across Rhythms during the process of computing the RQA. A one-way ANOVA was run on Meanline with Rhythm as factor and pair as random variable. The main effect of Meanline held when Recurrence Rate was fixed across Rhythms, $F(1, 7) = 14.264$, $p = 0.007$. Again, Meanline was higher for the 1-2 Rhythm than for the 4-2 Rhythm.

Table 2: Cross-recurrence main effects of Rhythm.

Outcome	1-2	4-2	F	η^2	p
Recurrence Rate	2.93%	2.53%	16.84	0.74	0.005
Meanline	131.22	122.78	16.81	0.74	0.005

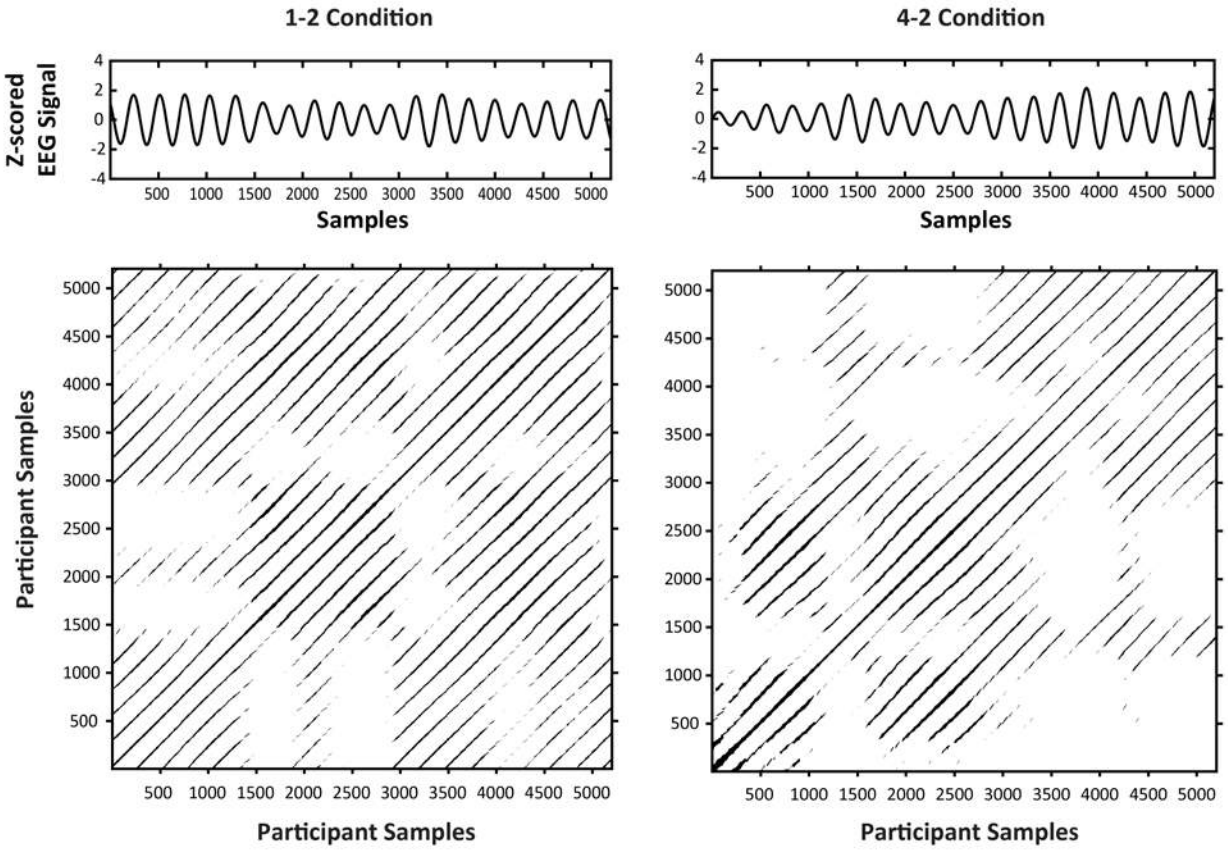


Figure 1: Time series and RPs with samples as a unit of time for one epoch from one participant for Rhythms 1-2 and 4-2. The time series shows the *z*-scored preprocessed signal from electrode C1.

Figure 2 shows example cross-recurrence plots (CRPs) for a single epoch from one pair for Rhythms 1-2 and 4-2 for the same trials shown in Figure 1. As can be seen in these examples, the 1-2 CRP is more densely occupied by recurrent points than the 4-2 CRP; these points also form longer diagonal lines than those in the 4-2 CRP. This indicates that the two signals overlap more often and for longer periods in phase space during the 1-2 Rhythm than the 4-2 Rhythm, indicating greater inter-brain stability. Furthermore, the white space between diagonal lines indicates the period at which the two neural signals recur with one another, and this period corresponds approximately to the participant tapping frequency (1.89 Hz or approximately 271 samples). Phase shifts between the two signals over time can also be observed by the degree of curvature in the diagonal lines in each CRP.

Discussion

The current experiment examined the application of RQA to neurophysiological data collected during a rhythmic tapping task between partners. Both auto- and cross-recurrence measures were sensitive to changes in stability of neural oscillations across tasks. Stability of neural oscillations at the participant tapping frequency was greater both within and across brains, as shown by larger recurrence rate and meanline outcomes from auto- and cross-recurrence, respectively, when there was more auditory feedback for both partners at the participants' tapping frequency.

We showed intra- and inter-brain recurrence that corresponded approximately to the participant tapping frequency. We also showed phase shifts in time as observed by the degree of curvature of the diagonal lines. Future work can further examine the time delay in recurrent points between two signals using quantifications such as the diagonal recurrence profile (e.g., Richardson & Dale, 2005; Dale, Kirkham, & Richardson, 2011), and subsequently relate this to behavioral performance. In contrast to other inter-brain metrics such as phase coherence, one advantage of cross-recurrence is the ability to show and subsequently quantify inter-brain dynamics when neural signals occupy the same phase space.

One limitation of the current experiment is that we only examined neural activity filtered at the participant tapping frequency. Future work can extend this technique to look at other stimulus frequencies to further examine the time evolution of neural dynamics in a joint motor task. We were also limited in our analyses by a small sample size. With more pairs, it could be possible to apply more sophisticated analysis methods to RQA outcomes such as an Actor-Partner Interdependence Model to examine how partners influence one another (Kenny, Kashy, & Cook, 2006). We also used PSD estimates for selecting a single electrode whose data were used for auto- and cross-recurrence

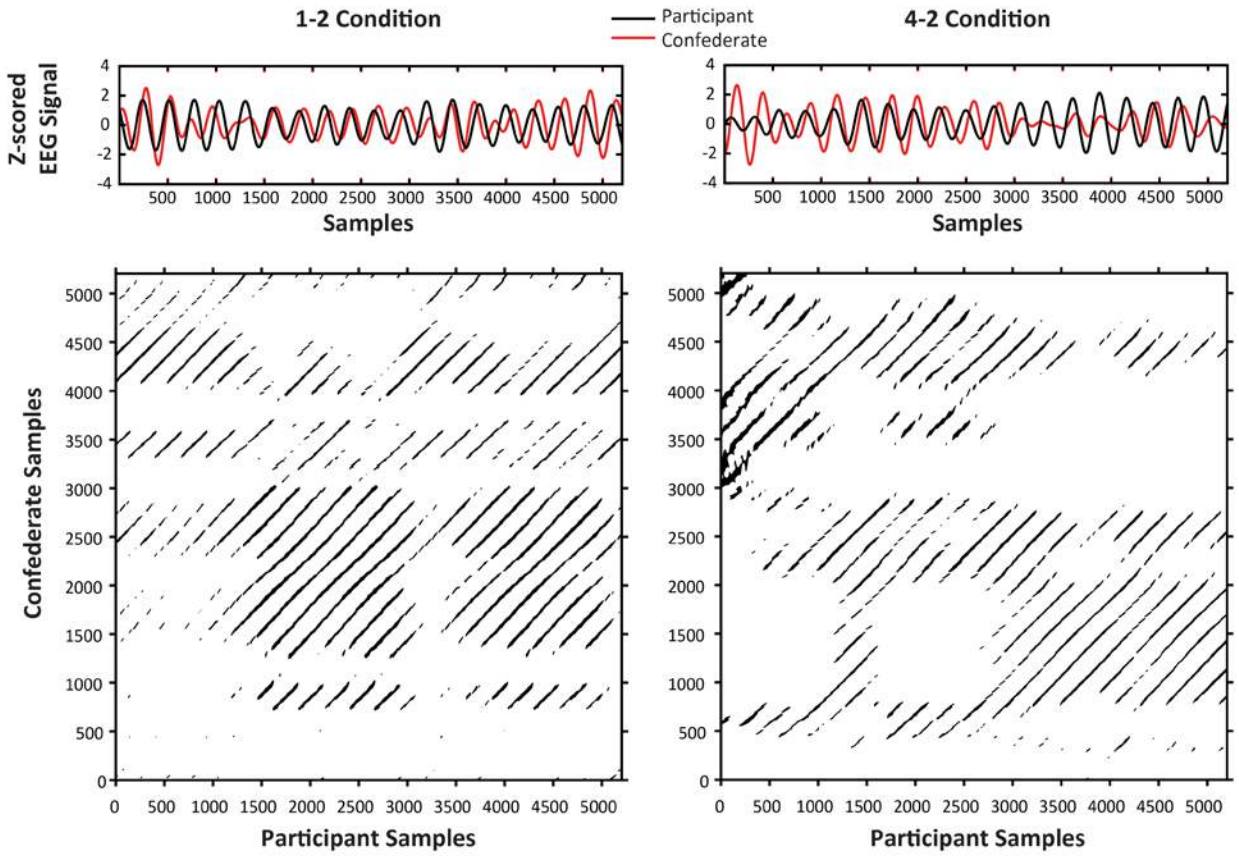


Figure 2: Time series and CRPs with samples as a unit of time for one epoch from one pair for Rhythms 1-2 and 4-2. Time series show the *z*-scored preprocessed signal from electrode C1 for the participant (in black) and the confederate (in red).

analyses. Future work can also extend this technique to identify regions of interest (i.e., multiple EEG electrodes) on which Multidimensional Recurrence Quantification Analysis (MdRQA) could potentially be applied (Wallot, Roepstorff, & Mønster, 2016).

In sum, recurrence quantification techniques were sensitive to changes in the dynamics of oscillatory neural activity that occurred during a joint rhythmic task. This is the first demonstration, to our knowledge, of RQA techniques to show consistent intra- and inter-brain differences in a joint auditory-motor task. These findings suggest that the sensitivity of RQA to stability of oscillatory neural activity might lend the technique to more fine-grained characterization of non-linearities in neural dynamics in a variety of behaviors and participant populations.

References

Acharya, U. R., Sree, S. V., Chattopadhyay, S., Yu, W., & Ang, P. C. A. (2011). Application of recurrence quantification analysis for the automated identification of epileptic EEG signals. *International Journal of Neural Systems, 21*(3), 199-211.

Becker, K., Schneider, G., Eder, M., Ranft, A., Kochs, E. F., Zieglgänsberger, W., & Dodt, H. U. (2010). Anaesthesia monitoring by recurrence quantification analysis of EEG data. *PLoS One, 5*(1), e8876.

Dale, R., Kirkham, N. Z., & Richardson, D. C. (2011). The dynamics of reference and shared visual attention. *Frontiers in Psychology, 2*, 355.

Dale, R., & Spivey, M. J. (2006). Unraveling the dyad: Using recurrence analysis to explore patterns of syntactic coordination between children and caregivers in conversation. *Language Learning, 56*(3), 391-430.

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods, 134*(1), 9-21.

Demos, A. P., & Chaffin, R. (2017). Removing obstacles to the analysis of movement in musical performance: Recurrence, mixed models, and surrogates. In M. Lesaffre, P-J. Maes, M. Leman (Eds) *The Routledge Companion to Embodied Music Interaction.* New York, NY: Routledge.

Demos, A. P., Chaffin, R., & Kant, V. (2014). Toward a dynamical theory of body movement in musical performance. *Frontiers in Psychology, 5*, 477.

Eckmann, J. P., Kamphorst, S. O., & Ruelle, D. (1987). Recurrence plots of dynamical systems. *Europhysics Letters, 4*(9), 973-977.

Finney, S. A. (2001). FTAP: A Linux-based program for tapping and music experiments. *Behaviour Research Methods, Instruments, and Computers, 33*(1), 65-72.

Kenny, D. A., Kashy, D. A., & Cook, W. L. (2006). *Methodology in the social sciences (David A. Kenny, Series Editor). Dyadic data analysis.* New York, NY: Guilford Press.

Kothe, C. (2014). Lab streaming layer (LSL). *https://github. com/sccn/labstreaminglayer*.

Marwan, N., & Meinke, A. (2004). Extended recurrence plot analysis and its application to ERP data. *International Journal of Bifurcation and Chaos, 14*(2), 761-771.

Marwan, N., Romano, M. C., Thiel, M., & Kurths, J. (2007). Recurrence plots for the analysis of complex systems. *Physics Reports, 438*(5-6), 237-329.

Marwan, N., & Webber, C. L. (2015). Mathematical and computational foundations of recurrence quantifications. In C. Webber, Jr., N. Marwan (Eds) *Recurrence Quantification Analysis: Theory and Best Practices* (pp. 3-43). Cham, Switzerland: Springer International Publishing.

Morillon, B., & Baillet, S. (2017). Motor origin of temporal predictions in auditory attention. *Proceedings of the National Academy of Sciences, 114*(42), E8913-E8921.

Nozaradan, S. (2014). Exploring how musical rhythm entrains brain activity with electroencephalogram frequency-tagging. *Philosophical Transactions of the Royal Society B: Biological Sciences, 369*(1658), 20130393.

Nozaradan, S., Zerouali, Y., Peretz, I., & Mouraux, A. (2013). Capturing with EEG the neural entrainment and coupling underlying sensorimotor synchronization to the beat. *Cerebral Cortex, 25*(3), 736-747.

Richardson, D. C., & Dale, R. (2005). Looking to understand: The coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension. *Cognitive Science, 29*(6), 1045-1060.

Romero, V., Fitzpatrick, P., Schmidt, R. C., & Richardson, M. J. (2016). Using cross-recurrence quantification analysis to understand social motor coordination in children with autism spectrum disorder. In: Webber, Jr. C., Ioana, C., Marwan, N. (eds) *Recurrence Plots and Their Quantifications: Expanding Horizons.* Springer Proceedings in Physics, vol 180. Springer, Cham.

Scheurich, R., Zamm, A., & Palmer, C. (2018). Tapping into rate flexibility: Musical training facilitates synchronization around spontaneous production rates. *Frontiers in Psychology, 9,* 458.

Schinkel, S., Dimigen, O., & Marwan, N. (2008). Selection of recurrence threshold for signal detection. *The European Physical Journal Special Topics, 164*(1), 45-53.

Schmit, J. M., Regis, D. I., & Riley, M. A. (2005). Dynamic patterns of postural sway in ballet dancers and track athletes. *Experimental Brain Research, 163*(3), 370-378.

Schmit, J. M., Riley, M. A., Dalvi, A., Sahay, A., Shear, P. K., Shockley, K. D., & Pun, R. Y. K. (2006). Deterministic center of pressure patterns characterize postural instability in Parkinson's disease. *Experimental Brain Research, 168*(3), 357-367.

Wallot, S., Roepstorff, A., & Mønster, D. (2016). Multidimensional Recurrence Quantification Analysis (MdRQA) for the analysis of multidimensional time-series: A software implementation in MATLAB and its application to group-level data in joint action. *Frontiers in Psychology, 7,* 1835.

Zamm, A., Debener, S., Bauer, A-K. R., Bleichner, M. G., Demos, A. P., & Palmer, C. (2018). Amplitude envelope correlations measure synchronous cortical oscillations in performing musicians. *Annals of the New York Academy of Sciences.*

Zamm, A., Palmer, C., Bauer, A-K. R., Bleichner, M. G., Demos, A. P., & Debener, S. (2017). Synchronizing MIDI and wireless EEG measurements during natural piano performance. *Brain Research.*

Big, hot, or bright? Integrating cues to perceive home energy use

Eleanor B. Schille-Hudson[1] (erbrower@iu.edu)**, Tyler Margehtis**[1,2] (tyler.marghetis@gmail.com)**, Deidra Miniard**[2] (deidraminiard@gmail.com)**, David Landy**[1,3] (dhlandy@gmail.com)**, Shahzeen Z. Attari**[2] (sattari@indiana.edu)

[1]Department of Psychological and Brain Sciences, Indiana University; [2]O'Neill School of Public and Environmental Affairs, Indiana University; [3]Netflix

Abstract

Despite constantly using energy and having extensive interactions with household appliances, people consistently mis-estimate the amount of energy that is used by home appliances. This poses major problems for conservation efforts, while also presenting an interesting case study in human perception. Since many forms of energy used are not directly perceptible, and since the *amount* of energy that is being used by an appliance is often difficult to infer from appearances alone, people often rely on cues. Some of these cues are more reliable than others and previous literature has investigated which of these cues people rely on. However, past literature has always studied these proximal cues in isolation—despite the fact that, during real-world perception, people are always integrating a variety of cues. Here, we investigate how people rely on a variety of cues, and how individual differences in the reliance on those cues predicts the ability to estimate home energy use.

Keywords: energy; perception; estimation; home appliances; multi-dimensional scaling

Introduction

Despite its importance in the face of catastrophic climate change, energy and energy use are not well or widely understood by the public. For many home appliances, we have only indirect access to the appliances' energy use and energy units are difficult to understand. However, people frequently make choices as energy consumers: When should I turn off the lights? For how long should I take a hot shower? To what temperature should I set the thermostat? These daily decisions all depend on a perception of energy use. Given people's poor understanding of energy use and the difficulty of perceiving energy use, how do people make these daily decisions about using energy?

In some contexts, people have access to explicit information about appliances' energy use. For instance, some smart meters are digital devices that indicate, in real time, how much energy is being used by an appliance; other appliances may have labels indicating their average energy use (e.g., "Energy Star" labels on efficient appliances). However, explicit information about energy use is the exception, not the rule.

In the absence of direct, explicit information about energy use, people must rely on indirect indices of energy use. (For reference on similar work done in the HCI community see He, Greenberg, and Huang, 2010 and Heller, Konstantinos, Borchers, 2013.) Vacuums are noisy. Lightbulbs are luminous and sometimes hot. It is these observable features that are typically available to individuals when they are making decisions about their energy use. Some of these cues, however, are more reliable than others. For instance, generating and extracting heat requires a lot of energy; mechanical movement, while perhaps more perceptually salient, can be accomplished with far less energy. Good judgements and decisions about energy use, therefore, requires a good sense of which proximal cues to rely on, and which to ignore. Understanding and improving these judgments can translate to energy conservation, as illustrated by the conservation benefits of in-home smart devices that give real-time feedback on energy-use (Darby, 2006; Delmas, Fischlein, & Asensio, 2013), although these energy technologies may be years away from becoming mainstream.

Past work on situated perception and decision making has advocated for similar approaches to understanding how people make judgments about entities that cannot be perceived directly. For instance, Brunswick (1956) proposed a "lens model" of perception, in which people must integrate across proximal cues in order to decide whether some target entity or property exists in the world; on this account, learning to perceive correctly involves learning how best to weight these different cues, so that more reliable cues (i.e., those that most often co-occur with the target phenomenon) are weighted more. A similar perspective has been advocated by researchers in the Judgement and Decision Making world, who have argued that, for many difficult decisions, people deploy 'replacement heuristics' — relying on some simpler or more easily perceived property or feature to make decisions about some target phenomenon that is more complex or difficult to perceive (Kahneman & Frederick, 2002). On all these approaches, understanding how people make complex perceptual judgments about 'invisible' entities, such as energy use, requires understanding the proximal cues or features they are relying on.

A number of past studies have tried to do exactly that. Previously, in the energy literature, different replacement heuristics have been studied. Past work has suggested that novices base their estimates of home energy use on perceptions of appliances' size (Cowen & Gatersleben, 2017), frequency of use (Schley & DeKay, 2015), effect on temperature (heating or cooling) (Attari, DeKay, Davidson, & de Bruin, 2010), and type of appliance (Lesic, Bruin, Davis, Krishnamurti, & Azevedo, 2018). But these past studies have focused on a single dimension

of experience (e.g., size), in isolation from the many other features which that dimension may be correlated (e.g., frequency of use). As a result, we still do not know how people weight the range of features to which they have access, or whether there is one or a subset of features that are driving most of people's energy estimates.

Moreover, all these approaches share the prediction that *better* judgements will involve better weighting of proximal cues. How do individual differences in weighting these features relate to individual differences in estimation ability?

Here, we attempt to answer these three outstanding questions: Which features are people relying on to make energy estimates? How do individual differences in cue-weighting relate to estimation skill? And how can we capture people's feature representation of appliances in a way that accounts for correlations among features?

In the following studies, we first surveyed participants for the most important or relevant features of energy in home appliances. We then took the most frequently cited features and used them to create feature rating scales for participants to rate multiple home appliances along. A multiple regression was performed on a few theoretically-driven features to determine how they competed with one another. Multi-dimensional scaling (MDS) was performed on all the features to capture the structure in how people perceive appliances and their energy use.

By performing these analyses on multiple features at once, we can establish which features matter *most* in the larger context of available appliance features. We also hope to paint a more clear and nuanced — and thus complete — picture of how these features are combined with one another. MDS affords us a look at categories of appliances that emerge and have implications for why some categories matter. These targeted analyses in concert with the larger picture of appliance feature perception, will hopefully inform future projects on how to help people better understand and use energy (Marghetis, Attari, and Landy, under review).

Methods

Participants

We recruited adults (N = 299) from the United States through Amazon Mechanical Turk, an online labor market that has been used previously for online studies. Each subject participated in return for $5. Only the data from those participants who completed the entire study were analyzed ($N = 261$). We also removed participants who repeated the exact same response for their estimates of all appliances (n = 1), giving us a final sample of N = 260.

Feature Selection

Participants rated features that were selected based on a previous study with different participants ($N = 17$) in which people were asked to list all features that they would use to estimate an appliance's energy use. On the basis of these free response features, we compiled a list of features that were most frequently cited and most widely applicable to our list of home appliances ($N = 13$, see Appendix).

Procedure

Participants first completed a feature rating task, in which they were presented with typical home appliances ($N = 36$) and asked to judge each appliance in terms of a set of perceptual or experiential features (e.g., brightness, loudness). They were first instructed "For each question, [to] please imagine a typical version of that appliance *while it is in use* and answer accordingly." The survey was organized by feature. For each feature, e.g. "How loud is each appliance?", participants were given a Likert scale from 1-10 as well as a Not Applicable box for each appliance. Both appliances and features were presented in a random order. Participants supplied ratings for the following features: how frequently the appliance is used, how big the appliance is, how long the appliance is used, how much light the appliance produces, how much the appliance heats itself/its environment, how much sound it makes, how much water it uses, how much it cools itself/its environment, how big its motor is, how much it heats water, how complex its software is, how complex its internal electronic components are, how complex its internal mechanical components are, how much movement it generates in itself/environment. Each participant rated each appliance along each feature dimension, totaling 36 x 13 ratings for each participant.

After the feature rating task, participants were asked to make energy estimates for each appliance. They were given a point of reference: "A 100-watt incandescent light bulb uses 100 units of energy in one hour." Then they were asked to make an estimate for each appliance, "How many units of energy do you think each of the following devices typically uses in one hour?" Appliances were presented in a random order. This task has been used in prior studies to investigate and elicit accuracy in energy perceptions (e.g., Attari et al., 2010).

Analysis

A multiple regressions analysis was run on features that have been identified in past research as important for energy estimation use (Cowen & Gatersleben, 2017; Schley & DeKay, 2015; Marghetis, Attari, and Landy, under review), namely: size, how "electronic" the appliance is, frequency of use, and how much the appliance changes the temperature (i.e., the maximum of the heating and cooling ratings). Feature ratings were z-scored across participants. In a mixed effects model, there were fixed slopes for the interaction of features and feature ratings of every participant, random intercepts on every appliance, and random slopes on feature ratings by participant. The random slopes for every participant's ratings were extracted and used to investigate individual differences in energy estimating accuracy.

Results

What proximal cues do people use to estimate appliances' energy use?

We first zoomed in on those features that have been identified, in past literature, as playing a role in novice's judgements of home energy use. These included how frequently the appliance is used, how "electronic" the appliance is, how much the appliance changes the temperature (the max of the 'heat' and 'cool' ratings), and how large the appliance is. Using a linear mixed effects model, we predicted participants' energy estimates (log transformed) using these four features, with random intercepts and slopes for participants, and random intercepts for appliances. Feature ratings were z-scored within each participant. See Figure 1 for coefficient estimates of reliance on these proximal cues.

Participants' estimates of appliances' energy use were driven almost entirely by how large they judged the appliance to be (b = 0.10 ± 0.01 SEM, p<.001). Most variance in estimates is accounted for by differences in *size*. By contrast, people's judgments of how much the appliance changed the temperature and of how "electronic" an appliance was also had much smaller relations to their energy estimates (b = 0.04 ± 0.01 SEM, p<.001, b = 0.05 ± 0.01 SEM, p<.001). Critically, we found no relation between judgments of how often an appliance is used and estimates of how much energy it uses — despite past work that has argued that frequency-of-use is used as a 'replacement heuristic' for energy estimation (Schley & DeKay, 2015). Note that people's estimates of energy use were explained primarily by judgments of the appliance's size rather than by how much the appliance changed the temperature, even though heat is a more reliable cue to energy use, because heating and cooling use a lot of energy.

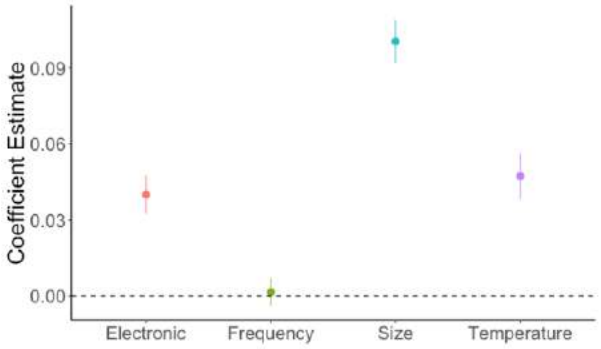
Individual differences in the use of proximal cues to estimate home energy use

We next investigated individual differences in the features that were associated with energy estimates — that is, we asked whether some people relied more on some proximal cues (e.g., size) than on others (e.g., temperature change).

	Size	Electronic	Frequency **of use**	Tempe**rature** Change
Size	1.00	0.183	0.066	0.215
Elect.		1.00	0.103	-0.131
Freq.			1.00	0.023
Temp.				1.00

Table 1: Correlation matrix of key features

Figure 1: Reliance on proximal cues to estimate energy use. Points indicate coefficient estimates from a mixed-effects model of energy estimates. Error lines indicate standard errors.

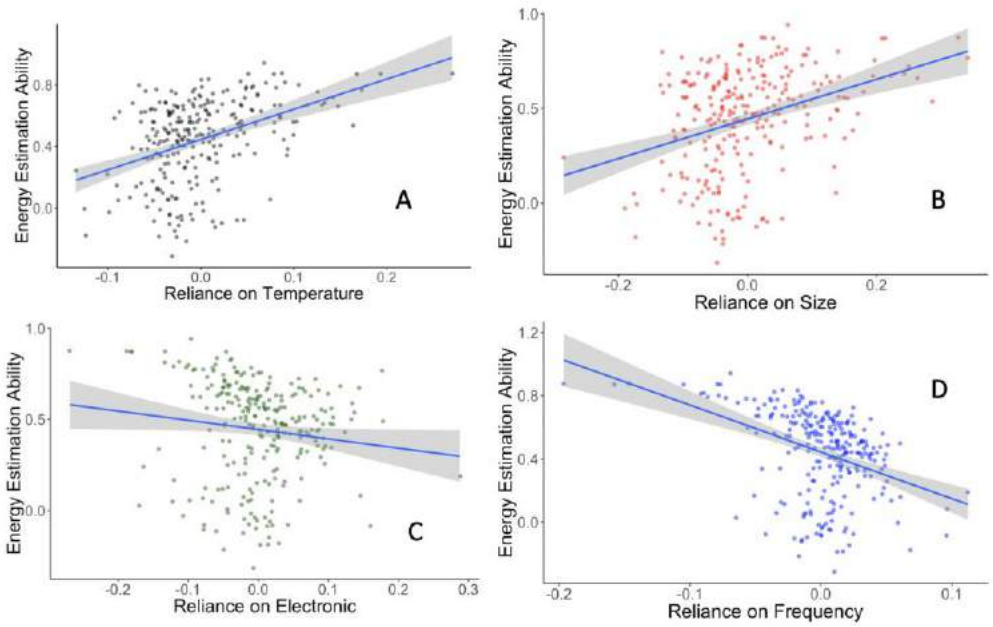


To capture these individual differences, we used the random by-participant slopes from our mixed effects model of energy estimates; for each participant, therefore, we had four random slopes, one for each feature (size, frequency-of-use, temperature change, and electronic-ness). Positive values of these random slopes indicate that a participant relies on that feature more than the group average; negative values indicate that they rely on that feature less than average.

In general, there was considerable variability in how strongly these features were associated with individuals' energy estimates (Fig. 2, panels A, B, C, and D). Some individuals' energy estimates were explained primarily by their judgments of the frequency of an appliance's use, despite the fact that frequency of use is a poor proxy for energy use. Others, however, appeared to ignore frequency and instead relied on temperature change, a reliable cue to energy use. Indeed, participants who relied more on temperature change tended to rely less on frequency of use (R = -0.60). Size and temperature change, both fairly good proxies for energy use, were highly correlated (R = 0.95), suggesting that people who use one feature to evaluate appliances' energy use are also likely to use the other.

All this together suggests that individual difference in the reliance on proximal cues might be associated with variability in how *good* people were at estimate home energy use. To quantify individual differences in estimation ability, we calculated, for each individual, the correlation between their estimates and the true energy used by each appliance. As predicted, participants who relied more on how much an appliance changed the temperature were also, overall, significantly better at estimating home energy use (b = 1.97 ± 0.27 SEM, p<.001); the same held for participants who relied more on the appliance's size, though to a lesser degree. Indeed, past work has found that lay people reliably underestimate the energy used by large appliances that heat or cool (Attari et al, 2010); here, our results suggest that there may be important variability in people's sensitivity to appliances' size and temperature change (Fig. 2A, 2B). By contrast, participants

Figure 2: Energy estimation ability as predicted by reliance on select features



who relied more on electronic-ness and frequency-of-use were overall worse at estimating home energy use (b = -0.51 ± 0.24 SEM, p<.05, b = -2.96 ± 0.43 SEM, p<.001). We also ran a correlation on the participants' reliance on each of these four features (Table 1). We found reliance on frequency of use and electronic-ness to be positively correlated, while frequency of use and temperature change were negatively correlated.

Characterizing the complex structure of the full appliance space

Finally, we combined ratings of all thirteen features (e.g., size, brightness, movement, etc.) to characterize lay perception of home appliances. To do so, we used multi-dimensional scaling (MDS). This technique takes the similarity between paired appliances and uses that to generate a reduced dimensional representation that captures how similar or different appliances are to each other. This approach gets at the rich structure that exists in how people perceive appliances as varying along multiple dimensions, many of which covary with each other. This approach is also necessary, because when dimensions are treated as independent, classic approaches like multiple regression do not account for collinearity of dimensions.

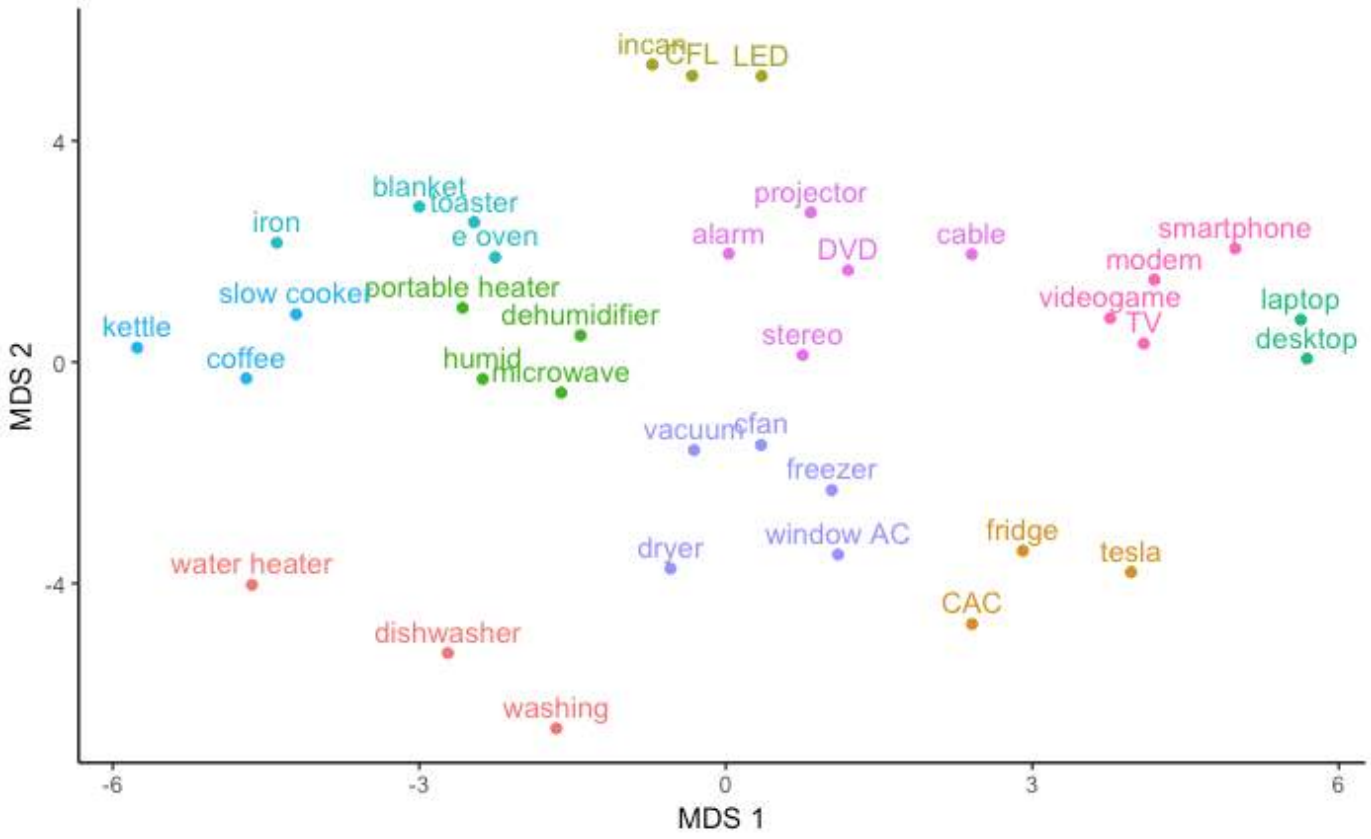
The two-dimensional MDS solution is illustrated in Figure 3. Note the rich structure that emerges bottom-up from this approach, with some appliances clumping together into meaningful groups, with related appliances clustering together into meaningful categories. We used k-means clustering (k=8) to capture these categories (Fig. 2). For example, all the light-bulb appliances (i.e. incandescent lightbulbs, Compact Fluorescent Light bulb, and LED bulb) group together because people rated those appliances very similarly.

While this MDS solution can characterize people's mental representations of appliances, it is blind to people's estimates of the appliances' energy use. However, when we regressed the MDS dimensions onto estimation ability, we found both MDS axes were related significantly to energy estimates increase (dimension 1: b = 146.88 ± 67.5 SEM, p<.05; dimension 2: b = -254.68 ± 104.0 SEM, p<.05). This was true despite the fact that these MDS dimensions combine multiple experiential features in complex, non-linear ways. Thus, lay people have structured perceptions of appliances, and these perceptions seem to relate systematically to their perceptions — and misperceptions — of their energy use. Future work should try to leverage this to improve energy decisions and behaviors.

Discussion

We began by asking how it is that people are able to estimate the energy used by appliances, when that energy use is often hidden. We found that estimates of appliances' size accounted for most of the variance in people's energy estimates. People relied, to a lesser extent on temperature change and how "electronic" an appliance, but they did not rely on frequency of use as a cue. Previous literature has claimed that all these features matter. Our results put those findings in a new light because we found that size is the primary driver of energy estimates. Since these replacement cues correlate, previous findings such as 'people use frequency of use as a replacement

Figure 3: Two-dimensional MDS solution for home appliance space



heuristic' might indicate that people tend to use bigger appliances more often. Interestingly, people relied more on size than heat, despite heat being a better indication of energy use. Heating (and cooling) both take a lot of energy but are perhaps not as obvious to people because the energy used to heat (and cool) are often used to achieve homeostasis. Your heating bill is high in the winter because so much energy has to be exerted to maintain your home at a constant temperature.

When we examined individual differences in the reliance on these cues, we found that the degree to which people relied on certain features predicted how good their energy estimates were. People who relied more on temperature change had better energy estimates than people who relied more on size, or any of the other theory-driven features used in our model. The more participants relied on how "electronic" an appliance was or on frequency of use, the worse their energy estimation ability was. When we ran a correlation on individual differences of reliance, we found that reliance on frequency is negatively correlated with reliance on temperature change. We also found that reliance on frequency is positively correlated with reliance on electronic-ness. This suggests that teaching people to use these more reliable cues may have benefits for energy judgments and decisions (Marghetis et al., under review).

Using multi-dimensional scaling, we also sought to characterize the public's mental representation of home appliances. This bottom-up approach found significant structure in people's perceptions of appliances; moreover, this two-dimensional representation was related systematically to people's energy estimates. In Fig. 3, the upper-left quadrant of the graph seems to include all the appliances that heat water, while the lower-left quadrant includes the appliances that heat without water. This suggests that this two-dimensional MDS solution has picked out *heat* as a notable component of one of its major axes. The appliances near the top of Fig. 3 are quite small and increase in size as you go down the MDS 2 axis, suggesting that this MDS solution has picked out size as a major component of its other axis. It is quite notable that even just a two-dimensional solution has, in a bottom-up way, picked out the two most useful and frequently used replacement heuristics. The clustering as shown in Fig. 3, also created through the bottom-up k-means algorithm, is quite remarkable as well. Kitchen appliances that heat water have clustered together on the left (blue); devices that are electronic or involved in entertainment have clustered together on the right (pink and green); appliances that heat or cool and move air around have also clustered together in the middle of the figure (purple). These clusters suggest that this MDS solution is a fruitful way to access the internal structure of people's complex perceptions.

Conclusion

We set out to answer three main questions. The first was 'Which features are people relying on to make energy estimates?' The answer to this is not simple. Our MDS solution shows that people rely on a complex and correlated set of proximal features. However, when comparing a smaller set of theoretically important features, size far outstrips any of them. Among the features we examined, people seem to rely most on size, even though it is not the best indicator of energy use. The best indicator of energy use was heat or temperature change.

We also set out to answer how individual differences in cue-weighting relate to estimation skill. Fig. 2A shows that as people rely on heat as a cue, their estimation skill improves. This is true to a lesser extent of size as well (Fig. 2B). As people rely on how electronic an appliance is, or how frequently it is used, their estimation skill decreases (Figs. 2C, 2D).

Finally, we set out to capture people's feature representation of appliances in a way that accounts for correlations among features. With an MDS solution, we found that meaningful clusters of appliances emerge, even from bottom-up clustering methods, and that the dimensions of this representation were related systematically to estimates of energy use.

This study speaks to previous energy literature that has attempted to identify the most predictive cue of people's energy estimates. By looking at several cues at once while accounting for correlations, we can say with confidence that despite the many, many features to choose from, the size of an appliance matters to people.

People do rely on the superficial cues about energy that they have access to. It is important to understand which of these people most rely on, so that we can more deeply understand how people understand and choose to use energy. Good energy choices can be encouraged in a variety of way, including but not limited to top-down policies, market-based incentives, extensive educational programs, home energy audits, and new home technologies. For example, in-home smart devices that give real-time feedback on energy-use can encourage energy conservation (Darby, 2006; Delmas et al., 2013). But implementing effective climate policies is politically difficult (Dietz, Ostrom, & Stern, 2003), home audits require time and resources that make scaling up nearly impossible, and new in-home energy technologies may be years away from mainstream use.

By understanding, and eventually changing either the cues people have access to, or their perceptions, we hope to encourage better ways of communicating energy information and making possible good and widely usable energy consumption habits.

References

Attari, S. Z., DeKay, M. L., Davidson, C. I., & de Bruin, W. B. (2010). Public perceptions of energy consumption and savings. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(37), 16054–16059.

Brunswik, E. (1956). Perception and the Representative Design of Psychological Experiments. University of California Press.

Cowen, L., & Gatersleben, B. (2017). Testing for the size heuristic in householders' perceptions of energy consumption. Journal of Environmental Psychology, 54, 103–115.

Darby, S. The effectiveness of feedback on energy consumption: A Review for DEFRA of the Literature on Metering, Billing and direct Displays. 24 (2006).

Delmas, M. A., Fischlein, M. & Asensio, O. I. Information strategies and energy conservation behavior: A meta-analysis of experimental studies from 1975 to 2012. *Energy Policy* **61**, 729–739 (2013).

Dietz, T., Ostrom, E. & Stern, P. C. The struggle to govern the commons. *Science* **302**, 1907–1912 (2003).

He, H. A., Greenberg, S., and Huang, E. M. One size does not fit all: applying the transtheoretical model to energy feedback technology design. In Proc. CHI '10, ACM (2010), 927–936.

Heller, Konstantinos, Borchers. Counter Entropy: Visualizing Power Consumption in an Energy+ House. In CHI '13: Extended Abstracts of the 2013 ACM annual conference on Human Factors in Computing Systems, April 2013.

Kahneman, D., & Frederick, S. (2002). Representativeness Revisited: Attribute Substitution in Intuitive Judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), Heuristics and Biases (1st ed., pp. 49–81). Cambridge University Press.

Lesic, V., Bruin, W. B. de, Davis, M. C., Krishnamurti, T., & Azevedo, I. M. L. (2018). Consumers' perceptions of energy use and energy savings: A literature review. Environmental Research Letters, 13(3), 033004.

Marghetis, T., Attari, S., and Landy, D., (Under Review). Correcting misperceptions of home energy use.

Schley, D. R., & DeKay, M. L. (2015). Cognitive accessibility in judgments of household energy consumption. Journal of Environmental Psychology, 43, 30–41.

Acknowledgments: This work is supported by NSF grant SES-1658804 from Decision, Risk and Management Sciences.

Appendix: Features

1. How **big** is each appliance?
2. How **long** is each appliance typically used?
3. How much **light** does each appliance produce?
4. How much does each appliance **heat** itself or its environment?
5. How **loud** is each appliance?
6. How much **water** does each appliance use?
7. How much does each appliance **cool** itself or its environment?
8. How big is the **motor** of each appliance?
9. How much does each appliance **heat water**?
10. How complex is the **software** each appliance runs?
11. How **electronic** is each appliance?
12. How **mechanical** is each appliance?
13. How much does each appliance **move** itself or its environment?
14. How **frequently** do you use each appliance?

Exploring the space of human exploration using Entropy Mastermind

Eric Schulz[1,2,*] **(ericschulz@fas.harvard.edu),**
Lara Bertram[2,3,*], **Matthias Hofer**[2,4], & **Jonathan D. Nelson**[2,3]

[1]Department of Psychology, Harvard University, Cambridge, Massachusetts, USA
[2]ABC / iSearch Group, Max Planck Institute for Human Development, Berlin, Germany
[3]School of Psychology, University of Surrey, Guildford, UK
[4]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
[*] Contributed equally to this work.

Abstract

What drives people's exploration in complex scenarios where they have to actively acquire information? How do people adapt their selection of queries to the environment? We explore these questions using Entropy Mastermind, a novel variant of the Mastermind code-breaking game, in which participants have to guess a secret code by making useful queries. Participants solved games more efficiently if the entropy of the game environment was low; moreover, people adapted their initial queries to the scenario they were in. We also investigated whether it would be possible to predict participants' queries within the generalized Sharma-Mittal information-theoretic framework. Although predicting individual queries was difficult, the modeling framework offered important insights on human behavior. Entropy Mastermind opens up rich possibilities for modeling and behavioral research.

Keywords: Curiosity; Active Learning; Exploration; Entropy

Introduction

Humans are curious animals. From learning how to speak to launching rockets into space, exploration drives mankind's progress small and large. Human exploration has been studied in self-directed learning paradigms in adults and children, in domains including causal learning (Bramley, Dayan, Griffiths, & Lagnado, 2017), categorization (Meder & Nelson, 2012), control (Osman & Speekenbrink, 2012), and explore-exploit tasks (Wu, Schulz, Speekenbrink, Nelson, & Meder, 2018). Some experiments have used games including Battleship (Gureckis & Markant, 2009) and 20 questions (Nelson, Divjak, Gudmundsdottir, Martignon, & Meder, 2014). Self-directed learning can lead to improved performance (Gureckis & Markant, 2012; Markant, Ruggeri, Gureckis, & Xu, 2016). For instance, participants actively intervening on a causal system made better inferences about the underlying causal structure than subjects who received identical information passively (Lagnado & Sloman, 2004).

Recent conceptual work (Coenen, Nelson, & Gureckis, 2018; Gureckis & Markant, 2012; Schulz & Gershman, 2019) is underpinned by the assumption that behavior is goal-directed and that people select observations based on a metric of usefulness (Settles, 2009). What metric best predicts how people evaluate the usefulness of possible queries? Past work has focused on the expected reduction of uncertainty, the extent of predictions' improvement, or the maximization of future rewards (Nelson, 2005). One study optimized experimental materials to maximally distinguish between different measures in an experience-based probabilistic classification task (Nelson, McKenzie, Cottrell, & Sejnowski, 2010). Results showed that participants were better described by probability gain than by information gain or other measures.

Markant and Gureckis (2012) tested whether participants maximize payoffs or information gain in a game of "battleships" (Gureckis & Markant, 2009), where each query cost money and an attempt to maximize utility would lead to different queries than information-gain based strategies. Surprisingly, participants' sampling behavior was nonetheless best matched by information gain. The authors argued that using information gain would lead to more knowledge about the underlying structure and therefore can be an effective strategy, no matter what the final task will be. Similar results have been obtained in an active causal learning task (Bramley, Lagnado, & Speekenbrink, 2015).

Exploiting the characteristics of the Entropy Mastermind game, we investigate people's sensitivity to the information structure of their environment (mathematical entropy or psychological uncertainty) and adaptive strategy selection when facing different levels of probabilistic uncertainty. In particular, we focus on what information metrics best predict how people evaluate the usefulness of possible queries, and on what initial-guess strategies people use.

A quintessential game of exploration

In the Mastermind code-breaking game, both information search and exploitation are essential for breaking the code. Thus, Mastermind offers a potential platform for bringing together pure information models (like expected information gain) and reinforcement learning models. In the classic two-player version of the game one player generates a secret colour code (e.g. blue, red, green) and the other player has to guess the secret code by repeatedly testing codes (making queries) and receiving feedback about the correctness of items in the guessed code. Although Mastermind has been extensively studied in computer science (for references see Berghman, Goossens, & Leus, 2009; Knuth, 1976), comparatively less work has been done in cognitive science (but see Laughlin, Lange, & Adamopoulos, 1982; Zhao, van de Pol, Raijmakers, & Szymanik, 2018).

We introduce the game "Entropy Mastermind" for studying exploration-driven problem solving and uncertainty reduction (Fig. 1). Key attributes of Entropy Mastermind, which distinguishes it from the classic game, are that Entropy Mastermind is a single-player app-based game in which hidden codes are drawn from known, and typically nonuniform, probability distributions. The probability distribution from which the hidden fruit code is drawn is depicted as a "fruit bowl" icon

array. The player is informed that the items are mixed before each draw, and drawn with replacement to form the hidden fruit code. Thus, Entropy Mastermind makes it possible to research how the level of entropy affects people's strategies and efficiency in game play.

As a first step toward modeling behavior in a probabilistic framework, we use a model that values both maximizing the probability of a correct query and a curiosity bonus, similar to recent work on human reinforcement learning (Schulz, Konstantinidis, & Speekenbrink, 2018; Wu et al., 2018). The curiosity bonus can be defined as information gain in the space of possible hypotheses (hidden codes). Whereas information gain has traditionally been thought of as reduction in Shannon entropy, any entropy metric could be used. We use the Sharma-Mittal space of entropy measures (Sharma & Mittal, 1977), which provides a framework within which many different kinds of entropy measures arise. According to the setting of two parameters, known as the *order* and *degree*, this entropy space can recover Shannon entropy, Bayes's error, and entropies from the Arimoto, Rényi and Tsallis families of entropy measures, among others (Crupi, Nelson, Meder, Cevolani, & Tentori, 2018). One of our research questions is whether Entropy Mastermind can help identify which model of uncertainty best predicts exploratory behavior.

In what follows, we formally define the Sharma-Mittal space as a unifying framework for information gain measures. We then report a preliminary study assessing and modeling human behavior in Fruit Salad Mastermind, a version of Entropy Mastermind in which the code jar is a fruit bowl and items are different kinds of fruits. First results show that participants adapted their queries to the level of entropy in the environment, solving games in less entropic environments more efficiently than in more entropic environments. Thus, basic assumptions for using Entropy Mastermind as a model of an information environment varying in entropy were met. Both the exploration and exploitation parts of the model were important to account for human behavior. However, distinguishing between different parts of the Sharma-Mittal space turned out to be difficult. Future research could work towards designing tasks that are optimized for the purpose of discriminating among specific entropy models.

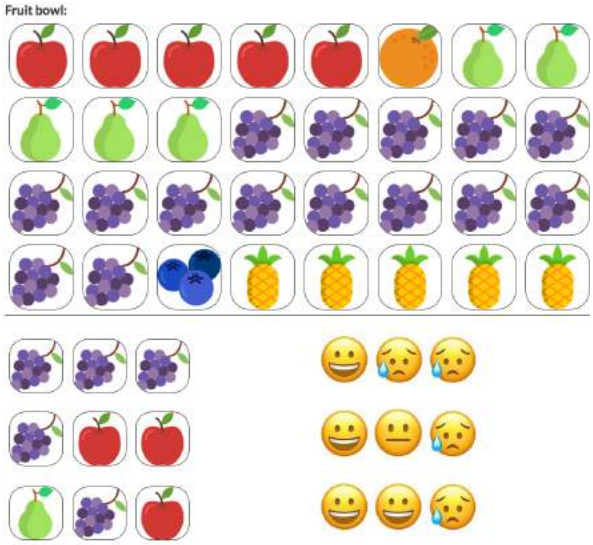


Figure 1: **Fruit Salad Mastermind: High Entropy Condition. Top:** Icon array presenting an example fruit bowl that generated the secret code. Probability distributions follow one of four entropy recipes, resulting in low, medium low, medium high and high entropy levels. Fruit types are apples, oranges, blueberries, grapes, pears, and pineapples for all possible versions of the fruit bowl. Codes are generated by randomly sampling fruits with replacement. Duplicates are allowed, so it is possible that the same fruit could appear in all positions of the hidden code. Players have to guess which fruit is in which position of the three slots of the secret code, by clicking on the position they want to change. Each position is initially blank; clicking cycles through the possible fruits. Once participants are satisfied with the proposed code, they can click on a "Check" button (not shown), and then receive feedback. **Bottom:** History of game play illustrating feedback. In the first guess, the player guessed 3 grape items. The feedback (one smiling face followed by two frowning faces) conveys that exactly one of the items is correct in type of fruit and in location. However, the player does not know which of their guesses is correct. There is no correspondence between the position of the guess and the position of the feedback: happy faces always come first, then neutral and lastly frowning faces. In the second guess, the player tested grape in the first position, and apple in each of the other two positions. The feedback (smiling face, neutral face, frowning face) indicates that one of the items is the correct type of fruit in the correct location, another item is in the code but needs to be moved to a new location, and another item is not in the code at all. As before, the guesser has to figure out which feedback face corresponds to which item in the code. The third guess of pear, grape, apple obtains two smiling faces and one frowning face. At this point the guesser can infer that the middle position is grape, and the final position is apple; the guesser must still figure out the first item.

Mapping the space of exploration

In Mastermind both *learning* about the true code and *guessing* the true code are important. To make this intuitive, suppose that there are two possible codes, given everything that has been learned to date, and that one of these codes has 90% probability of being the correct code. The same information, namely which code is correct, will be gleaned from testing either code; thus, the queries have equal value irrespective of which model of information gain is used. But clearly it is sensible to test the code that has 90% probability of being correct, thus having 90%, rather than 10%, probability of ending the game after the next query. We implement this idea via a softmax response rule on a value function which is based on the probability of each query being the correct code in the immediate time step, as well as a curiosity-driven exploration bonus[1]:

$$P(\text{action} = a_i) \propto P(\text{success}|a_i) + \beta \times \text{curiosity bonus}(a_i) \quad (1)$$

How promising a code seems is determined by its current probability of being correct $P(success|a_i)$. This probability is always the same given a specific history of queries and feedback. The curiosity bonus(a_i) is weighted by a free parameter

[1]Note that the parts of Eq. 1 are additive. Thus, even a query that has a probability of 0 of being the true code can still be chosen if it offers enough informational value.

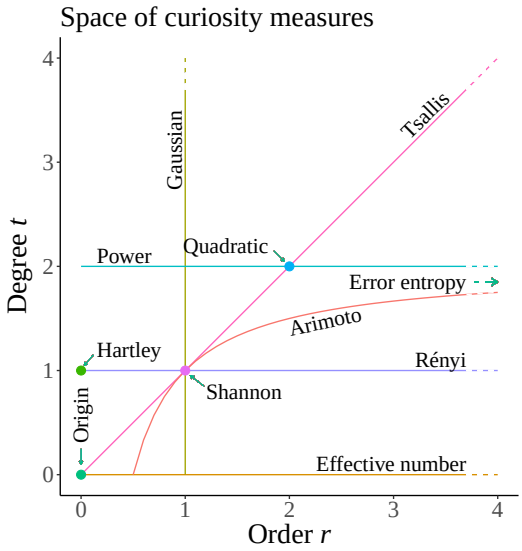


Figure 2: **Sharma-Mittal space.** The Sharma-Mittal family of entropy measures is represented in a Cartesian quadrant with values of the order parameter r and of the degree parameter t. The order parameter captures how much minor hypotheses are disregarded (e.g. that grapes may be contained in the code when the fruit bowl contains only a small proportion of grapes) and the degree parameter captures how prominent the goal of getting as close as possible to the state of certainty is (i.e. how much one strives to falsify existing hypotheses, e.g. that grapes are contained at all in the code). Each point in the quadrant corresponds to a specific entropy measure, each line corresponds to a distinct one-parameter generalized entropy function. Several special cases are highlighted.

β and can be defined as how much an action promises to reduce uncertainty over the space of possible hypotheses (i.e., how much it reduces uncertainty about possible codes).

The uncertainty in a discrete random variable $K = k_1, k_2, \ldots k_n$ can be measured by its entropy. We use the generalized Sharma-Mittal space of entropy measures, that unifies multiple past proposals (Crupi et al., 2018), and can be defined as:

$$\text{entropy}(K) = \frac{1}{t-1}\left[1 - \left(\sum_{i=1}^{n} P(k_i)^r\right)^{\frac{t-1}{r-1}}\right], \qquad (2)$$

where r is the *order* and t the *degree* of the entropy measure. Note that limits, which exist, are used for points where the above equation is undefined. Although the above equation may not be immediately intuitive, there are a number of ways to build understanding about this space. All of the Sharma-Mittal entropy measures can be thought of as quantifying the average surprise that would be experienced if the value of the random variable K was learned. In the case of Mastermind, this is the average surprise that would be experienced if one were to immediately learn the true hidden code.

The degree parameter t governs which kind of surprise is averaged. If $t = 1$, then $surprise(k_i) = ln(1/P(k_i))$, as in Shannon and all of the Rényi entropies. If $t = 2$, then $surprise(k_i) = 1 - P(k_i)$, as in the cases of Quadratic entropy and Bayes's error. If $t > 1$, a test is more useful if it is conclusive than if it is not. If $t < 1$, a test is always less useful if it is conclusive than if it is not. The order parameter r determines what kind of averaging function is used. It can be thought of as an index of the imbalance of the entropy function, which indicates how much the entropy measure discounts minor (low probability) hypotheses. For example, when $r = 0$, entropy becomes an increasing function of the mere number of the possible options. When r goes to infinity, entropy becomes a decreasing function of the probability of a single most likely hypothesis. For further discussion and examples see (Crupi et al., 2018).

Several special cases exist within the Sharma-Mittal space, as Figure 2 illustrates. For example, Shannon entropy is the result of setting $r = t = 1$, and probability gain (also called error entropy) is the result of setting $t = 2$ and letting $r \to \infty$. One of the goals of the present research is to investigate whether people's striving for information (the curiosity goal) can be represented well as a generalized information gain metric, where information is defined as the expected reduction in one of the Sharma-Mittal entropy functions over the probability distribution of the possible codes.

Methods

Participants and Design Forty-seven first-year undergraduate students (38 female, M_{age}=19.04; SD=1.04; range: 18 to 23) at University of Surrey participated in our study as part of a cognitive psychology class. Participants gave informed consent in accordance with the University's procedures and the Helsinki Declaration. They were introduced to the rules and interface of the game and completed a pretest. Participants then played Fruit Salad Mastermind, spending an average of 10.5 minutes on the task.

Materials and Procedure Participants were required to correctly answer four comprehension questions before game play began. These questions tested participants' understanding of the goal of the game and the interpretation of feedback (i.e. making sure that they understood that the position of the faces did not correspond to the position of items in the entered code). Participants were instructed to figure out the secret code using as few guesses as possible. Since the experiment was self-paced, the number of rounds played varied between participants.

Entropy conditions In each game, one of the four entropy conditions was chosen at random and the six fruits were randomly assigned to the six proportions of that condition. The resulting generating "fruit bowl" was presented to participants as an icon array above the current game. A "hidden fruit code" was generated from that distribution. In the *very high entropy* condition, the secret code was sampled based on the proportions $(5,5,5,5,6,6)$. This means, for example, that there could be 5 pineapples, 5 apples, 5 pears, 5 blueberries, 6 grapes, and 6 oranges, out of a total of 32 items, from which three fruits were sampled with replacement to generate the secret code. In the *high entropy* condition, the secret code

was sampled based on the proportions $(1,1,5,5,5,15)$. In the *low entropy* condition, the secret code was sampled based on the proportions $(1,1,1,4,4,21)$. Finally, in the *very low entropy* condition, the secret code was sampled based on the proportions $(1,1,1,1,1,27)$.

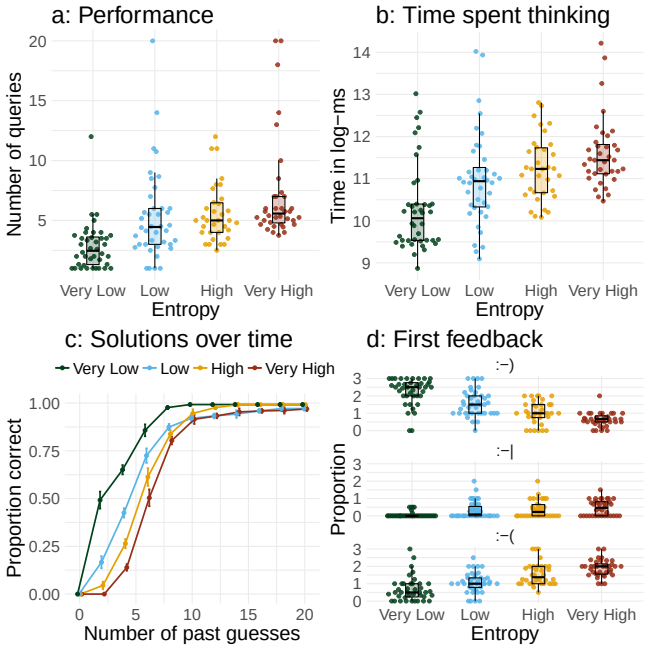


Figure 3: **Behavioral results. a:** Number of queries required to solve a game by entropy condition (ordered from lowest to highest). **b:** Time spent thinking (measured in log-ms per guess) by entropy condition (ordered from lowest to highest). **c:** Proportion of correct guesses in dependency of number of past guesses by entropy condition. **d:** Mean proportional feedback after first guess by entropy condition. Points represent mean per participant. Error bars indicate the standard error of the mean.

Behavioral results

We analyzed behavioral results using both frequentist and Bayesian statistics. For testing hypotheses regarding the behavioral data and the model comparison, we used the default two-sided Bayesian t-test for independent samples with a Jeffreys-Zellner-Siow prior with its scale set to $\sqrt{2}/2$ (Rouder, Speckman, Sun, Morey, & Iverson, 2009).

We first analyzed the number of required guesses to solve a game as a function of the entropy condition (Fig. 3a). This revealed a positive average rank correlation between how much entropy a condition contained and the number of queries participants required to solve a game (Kendall's $\tau = 0.48$, $t(46) = 12.44$, $d = 1.81$, $BF > 100$). More specifically, participants required fewer queries on average for the very low entropy games as compared to low entropy games ($t(46) = -5.69$, $p < .001$, $d = 0.83$, $BF > 100$). They also required fewer queries for the low entropy games than for the high entropy games ($t(46) = -3.16$, $p = .002$, $d = 0.46$, $BF = 11.8$). Finally, participants needed fewer queries for the high entropy games than for the very high entropy games ($t(46) = -3.96$, $p < .001$, $d = 0.58$, $BF = 97.2$).

Next, we analyzed how much time participants spent thinking to enter a guess by entropy condition (Fig. 3b). Thus, we assessed their mean time to submit a query measured in log-milliseconds. There was a positive average rank-correlation between a game's entropy and participants' average time spent thinking, Kendall's $\tau = 0.48$, $t(46) = 12.44$, $d = 1.68$, $BF > 100$. More specifically, participants spent less time thinking during the very low entropy games than during the low entropy games ($t(46) = -4.07$, $p < .001$, $d = 0.59$, $BF = 97.2$). They also spent less time thinking in the low entropy than in the high entropy games ($t(46) = -3.68$, $p < .001$, $d = 0.54$, $BF = 45.5$). Finally, they spent less time in the high entropy than in the very high entropy games ($t(46) = -4.05$, $p < .001$, $d = 0.59$, $B > 100$).

We also analyzed the proportion of solved games as a function of the number of past guesses, again comparing the different entropy conditions (Fig. 3c). We thus estimated a Bayesian logistic regression of number of past guesses onto the proportion of correct guesses for each of the entropy conditions, using Metropolis-Hastings Markov chain Monte Carlo sampling (implemented in `MCMCpack`, Martin, Quinn, Park, & Park, 2018). The resulting posterior estimate for the effect of number of past guesses onto the probability of guessing correctly was smallest for the very high entropy condition ($\hat{\beta} = 0.15$, 95%HDI=[0.14, 0.16]). The same estimate was higher for the high entropy condition ($\hat{\beta} = 0.19$, 95%HDI=[0.18, 0.20]), which did not differ meaningfully from the low entropy condition ($\hat{\beta} = 0.18$, 95%HDI=[0.17, 0.20]). The very low entropy condition showed the highest estimated effect ($\hat{\beta} = 0.30$, 95%HDI=[0.28, 0.33]). Thus, participants' solution rates differed meaningfully between entropy conditions, with lower entropy leading to faster rates.

In our last behavioral analysis, we looked at the very first query participants submitted as well as the feedback they received for that query (Fig. 3d). The number of smiling faces received on the very first guess was negatively rank-correlated with entropy condition, $\tau = -0.51$, $t(41) = -9.80$, $p < .001$, $d = 1.51$, $BF > 100$, whereas the number of frowning faces showed a positive rank-correlation, $\tau = 0.30$, $t(30) = 6.00$, $p < .001$, $d = 1.06$, $BF > 100$. Interestingly, participants adapted their first queries to the entropy condition, leading to a positive rank correlation between the set size of their first query (the number of unique kinds of fruit contained in the query) and the entropy of the generating distribution, $\tau = 0.40$, $t(46) = 9.00$, $p < .001$, $d = 1.31$, $BF > 100$. Put differently, if the generating distribution was higher in entropy, then participants tested a larger number of different fruits as part of their first query.

Computational modeling

We now turn to a model-based analysis of participants' exploration strategies. For this, we first need a formal account of intelligent Mastermind play. Logically, all combinations that are still consistent in round i based on the feedback received so far are part of a feasible set $\mathcal{F}_i$. Note that in Entropy Mastermind, not only the feasible codes but also their prob-

abilities (which are not in general equal) are relevant. Code combinations ruled out by prior feedback have zero probability. The remaining items' probability mass is proportional to the probability of obtaining the item via sampling from the code jar. The effective size of the feasible set is the total number of all non-zero probability codes left in the set. Let the probability that c_i is the hidden code given the current feasible set be denoted $P(c_i)$. The feasible set is guaranteed to shrink after each round unless a guess c_i is repeated. A general playing strategy consists of (i) identifying the set of feasible combinations $\mathcal{F}_i$ (with $\mathcal{F}_0 = \mathcal{E}$), where prior feedback is used to determine which combinations are still viable; and (ii) picking a combination c_i for the next guess. Let us denote the informational usefulness of playing combination c in the current round with $u(c)$, with

$$u(c) = \text{entropy}(\mathcal{F}_i) - \sum_{r}^{\mathcal{R}} P(f) \cdot \text{entropy}(\hat{\mathcal{F}}_{c,f}), \qquad (3)$$

i.e. the difference in entropy (under a particular Sharma Mittal entropy measure with specified order and degree) between the current feasible set and the expected entropy when playing code c. To compute expected entropy, for each possible feedback $f \in \mathcal{R}$, we compute the product of the probability of receiving that feedback $P(f)$ times the entropy of the updated feasible set $\hat{\mathcal{F}}_{c,r}$ when playing combination c and receiving feedback r. To compute $P(f)$ for a given c, we look at all the combinations $c_j \in \mathcal{F}_i$, that lead to feedback f. To this end, we define a feedback function $h(c,c_j) = f$ that returns the feedback f obtained from checking code c against code c_j. The probability of feedback f for code c can then be calculated as follows:

$$P(f) = \frac{\sum_{c_j}^{\mathcal{F}} P(c_j) \cdot \mathbb{1}_{h(c,c_j)=f}}{\sum_{c_j}^{\mathcal{F}} \sum_{c_k}^{\mathcal{F}} P(c_k) \cdot \mathbb{1}_{h(c_j,c_k)=f}}.$$

The indicator function $\mathbb{1}_{h(c,c_j)=f}$ ensures that we only sum over codes c_j that generate the required feedback f. The probability of any combination of fruits $c = m_1 m_2 ... m_n$ can be computed as

$$P(c = m_1 m_2 ... m_n) = P(m_1) \cdot P(m_2) \cdot ... \cdot P(m_n) \qquad (4)$$

where each $P(m)$ represents the probability of sampling the corresponding fruit item from the fruit jar. The other term of Equation 3, entropy$(\hat{\mathcal{F}}_{c,f})$, requires us to compute hypothetical feasible sets $\hat{\mathcal{F}}_{c,r}$. Given the current feasible set $\mathcal{F}_i$, a combination c we want to evaluate, and hypothetical feedback f, we need to exclude all combinations $c_j \in \mathcal{F}_i$ for which $h(c,c_j) \neq f$; that is, all combinations c_j that are not consistent with obtaining feedback f.

Lastly, one has to assign a utility to a feasible set $\mathcal{F}$. For this, we use the Sharma-Mittal entropy framework to compute the entropy of a probability distribution defined over set $\mathcal{F}$, $P_{\mathcal{F}}(c)$. For each combination $c \in \mathcal{F}$

$$P_{\mathcal{F}}(c) = \frac{P(c)}{\sum_{c_j}^{\mathcal{F}} P(c_j)},$$

where the nominator $P(c)$ is computed according to Equation 4 and the denominator is a normalization term.

We assess how well the combination of an entropy-based exploration bonus and the probability of making a correct guess describes players' guesses over time. For this, we analyzed the last five games of the 34 participants who played at least five games in total. We restricted our analysis to the last five games as our goal was to study strategies used rather than early learning. Next, we calculated the expected information gain for all of the $6 \times 6 \times 6$ possible fruit combinations that a participant could enter on every trial for each participant, given the participant-specific history of queries in a game. We calculated this information gain for every combination of order $r = [1/16, 1/8, 1/4, 1/2, 1, 2, 4, 8, 16, 32, 64]$ and degree $t = [1/16, 1/8, 1/4, 1/2, 1, 2, 4, 8, 16, 32, 64]$, i.e. 121 models per participant in total. We then combined the probability of a guess being correct with the information gain assessed by the specific entropy measure following Equation 2 to arrive at a value of an action's usefulness $V(a_t)$, which we put in a softmax function to calculate choice probabilities:

$$P(x) = \frac{\exp(V(a_t(\mathbf{x}))/\tau)}{\sum_{j=1}^{N} \exp(V(a_t(\mathbf{x}))/\tau)} \qquad (5)$$

where τ is a free temperature parameter. We followed previous work (Wu et al., 2018; Parpart, Schulz, Speekenbrink, & Love, 2017) and calculated each model's AIC$(\mathcal{M}) = -2\log(L(\mathcal{M})) + 2k$ and standardized it using a pseudo-R^2 measure as an indicator for goodness of fit, comparing each model $\mathcal{M}_k$ to a random model: $\mathcal{M}_{\text{rand}}$, $R^2 = 1 - \text{AIC}(\mathcal{M}_k)/\text{AIC}(\mathcal{M}_{\text{rand}})$.

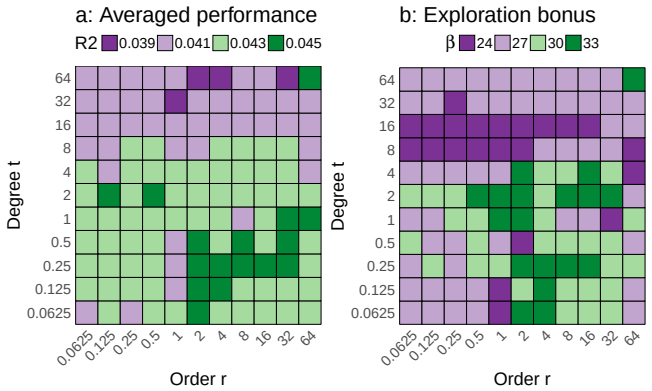


Figure 4: **Modeling results. a:** Averaged r^2 for different Sharma-Mittal parameters. **b:** Estimated exploration bonus β for different Sharma-Mittal parameters.

The results of this analysis revealed a mean pseudo-R^2 of 0.041 over all orders and degrees, which was low but significantly better than chance ($t(33) = 20.52$, $p < 0.001$ $d = 1.86$, $BF > 100$). Moreover, the estimated median temperature parameter was $\tau = 1.02$, indicating a relatively wide spread of

predictions. There was a significant negative rank-correlation between the degree parameter and model fit, $\tau = -0.37$, $z = -5.84$, $p < .001$, $BF > 100$, whereas this correlation was not significant for the order parameter, $\tau = 0.04$, $z = 0.60$, $p = .54$, $BF = 0.3$. Thus, even though entropies with smaller degree parameters seemed to generally work better at modeling participants' queries, there was no meaningful effect of the different order parameters.

The range of pseudo-R^2 values, $0.038 - 0.045$, also shows that most of the entropy measures led to similar performance. We also assessed the magnitude of the estimated exploration bonus β (Fig. 4b), which had a mean of $\hat{\beta} = 27.81$, and therefore differed significantly from 0, $t(33) = 115.47$, $p < .001$, $d = 10.9$, $BF > 100$. This means that the final model of participants' game play had to incorporate both a code's probability of being correct as well as its potential information gain. Interestingly, areas of the Sharma-Mittal space with higher r^2 also tended to have higher β estimates.

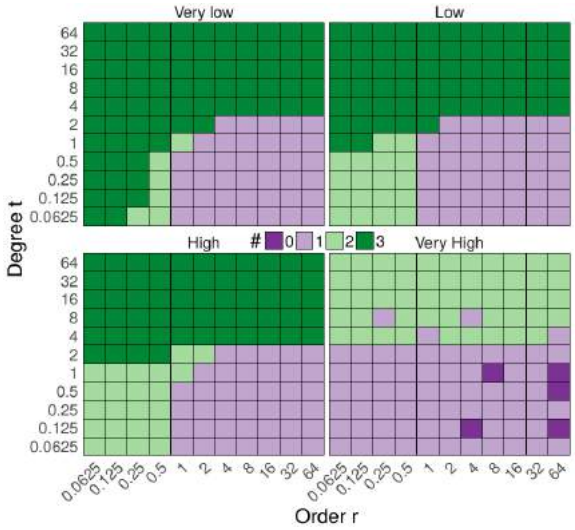


Figure 5: **Number of times the most likely fruit was chosen in the first query by simulated entropy models across entropy conditions.**

Finally, we compared how often participants put the most likely fruit into their first query with how often simulated models of different order and degree parameters chose the same fruit in their first query, for each entropy condition (see Fig. 5). The higher degree models chose the most likely fruit more often than people did. Specifically, participants put on average 2.14 of the most likely fruit in their first query in the very low entropy condition, 1.60 in the low entropy condition, 1.26 in the high entropy condition and 0.48 in the very high entropy condition. This analysis therefore corroborated our previous finding that the lower degree entropies better matched participants' queries. In relation to previous work modeling behavior with the Sharma-Mittal framework, Entropy Mastermind appears to be more similar to experience-based than to description-based probabilistic classification tasks (see Crupi et al., 2018, Fig. 7).

Discussion and conclusion

We introduced Entropy Mastermind as a game for researching human curiosity and exploration in complex environments. More specifically, we suggest this game as a paradigm for the study of how people select queries to reduce uncertainty under different levels of initial entropy. The complexity of the game resembles aspects of scientific inference (Strom & Barolo, 2011) and life. For instance, in life and in science, it can be a challenge to fully assimilate feedback that we get when we make queries. Entropy Mastermind thus complements existing games, such as Battleship (Gureckis & Markant, 2009), 20-questions (Nelson et al., 2014), or explore-exploit (Wu et al., 2018) tasks.

We found that participants required fewer queries, spent less time thinking about queries and showed faster learning rates if the distribution generating the secret code had lower entropy. They also adapted their queries to the code-generating distribution, and did so in sensible ways. In particular, many of the informational models (Figure 5) used greater proportions of the most-probable fruit in the first guess in lower-entropy conditions; participants also followed this pattern. Thus, one may conclude that people are generally sensitive to different levels of entropy, which is a prerequisite for a research agenda modeling human exploratory behavior within the Sharma-Mittal space.

Our modeling results paralleled earlier findings from other tasks (Crupi et al., 2018) suggesting that it is easier to identify the value of the degree parameter than of the order parameter in the Sharma-Mittal space. Interestingly, to identify the order parameter a different type of question could be asked, translating higher entropy into difficulty of game play in the sense of the number of queries required to guess the secret code (for the underlying mathematical result see Crupi et al., 2018). Participants could be directly asked which of two code jars would be harder to play Mastermind with (Figure 6).

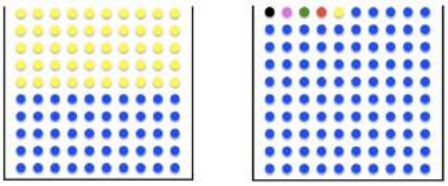


Figure 6: **Identifying the order parameter.** Which distribution is harder for playing Entropy Mastermind? Shannon entropy (order=1) deems the 50:50 distribution higher entropy, but lower-order entropies deem the 95:1:1:1:1:1 distribution higher entropy.

The general predictive performance of many models was relatively similar and rather low. This might be due to the overall complexity of choices, since there were 216 possible options on every trial, making it difficult to compare among candidate models (also see Parpart et al., 2017).

The difficulty of modeling could also be due to participants using cognitive shortcuts, as has been observed in other domains of active learning (Bramley et al., 2017). Furthermore, it is unlikely that participants evaluate the usefulness of all possible queries at each time point. Instead,

they might approximate a query's usefulness by sampling and reusing past hypotheses, as has been shown in other domains of human reasoning and hypothesis evaluation (Dasgupta, Schulz, & Gershman, 2017; Dasgupta, Schulz, Goodman, & Gershman, 2018; Lieder, Griffiths, & Hsu, 2018). Future studies should therefore investigate both heuristic strategies (Gigerenzer & Gaissmaier, 2011) and boundedly rational approaches (Griffiths, Lieder, & Goodman, 2015). Adaptive experimental designs (Cavagnaro, Myung, Pitt, & Kujala, 2010) could also be used to maximally discriminate among models.

Summing up, we propose Entropy Mastermind as a promising paradigm for investigating human exploration behavior in complex hypothesis testing scenarios. In related research we are assessing whether Entropy Mastermind can be used as an educational tool for primary or secondary school students, and for studying the effects of emotional states on strategies used and information search efficiency. Although our current modeling framework did not fully map out the space of exploration behavior, we believe that combining the Sharma-Mittal space of entropy measures with an enjoyable game rich in scientific history can further inform our theories of self-directed learning. We will keep exploring.

Acknowledgments

ES is supported by the Harvard Data Science Initiative. This work was supported by grant NE 1713/2 to JDN from the Deutsche Forschungsgemeinschaft as part of the priority program New Frameworks of Rationality (SPP 1516). We thank Eloisa Bentivegna, Neil Bramley, Vincenzo Crupi, Florian Ellsaesser, Alberto Feduzi, Flavia Filimon, George Kachergis, Laura Martignon, Bjoern Meder, Elif Oezel, Anselm Rothe, Azzurra Ruggeri, Katya Tentori, John Wong and the iSearch research group for helpful comments and ideas.

References

Berghman, L., Goossens, D., & Leus, R. (2009). Efficient solutions for mastermind using genetic algorithms. *Computers & Operations Research*, *36*, 1880–1885.

Bramley, N. R., Dayan, P., Griffiths, T. L., & Lagnado, D. A. (2017). Formalizing Neuraths ship: Approximate algorithms for online causal learning. *Psychological Review*, *124*, 301–338.

Bramley, N. R., Lagnado, D. A., & Speekenbrink, M. (2015). Conservative forgetful scholars: How people learn causal structure through sequences of interventions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*, 708–731.

Cavagnaro, D. R., Myung, J. I., Pitt, M. A., & Kujala, J. V. (2010). Adaptive design optimization: A mutual information-based approach to model discrimination in cognitive science. *Neural Computation*, *22*, 887–905.

Coenen, A., Nelson, J. D., & Gureckis, T. M. (2018). Asking the right questions about the psychology of human inquiry: Nine open challenges. *Psychonomic Bulletin & Review*, 1–41.

Crupi, V., Nelson, J. D., Meder, B., Cevolani, G., & Tentori, K. (2018). Generalized information theory meets human cognition: Introducing a unified framework to model uncertainty and information search. *Cognitive Science*, *42*, 1410–1456.

Dasgupta, I., Schulz, E., & Gershman, S. J. (2017). Where do hypotheses come from? *Cognitive Psychology*, *96*, 1–25.

Dasgupta, I., Schulz, E., Goodman, N. D., & Gershman, S. J. (2018). Remembrance of inferences past: Amortization in human hypothesis generation. *Cognition*, *178*, 67–81.

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology*, *62*, 451–482.

Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in Cognitive Science*, *7*, 217–229.

Gureckis, T. M., & Markant, D. B. (2009). Active learning strategies in a spatial concept learning game. In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st annual conference of the cognitive science society* (pp. 3145–3150).

Gureckis, T. M., & Markant, D. B. (2012). Self-directed learning: A cognitive and computational perspective. *Perspectives on Psychological Science*, *7*, 464–481.

Knuth, D. E. (1976). The computer as master mind. *Journal of Recreational Mathematics*, *9*, 1–6.

Lagnado, D. A., & Sloman, S. (2004). The advantage of timely intervention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 856–876.

Laughlin, P. R., Lange, R., & Adamopoulos, J. (1982). Selection strategies for "mastermind" problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *8*(5), 475.

Lieder, F., Griffiths, T. L., & Hsu, M. (2018). Overrepresentation of extreme events in decision making reflects rational use of cognitive resources. *Psychological review*, *125*(1), 1.

Markant, D. B., & Gureckis, T. M. (2012). Does the utility of information influence sampling behavior? In P. D. Miyake N. & R. Cooper (Eds.), *Proceedings of the 34th annual conference of the cognitive science society* (pp. 719–724).

Markant, D. B., Ruggeri, A., Gureckis, T. M., & Xu, F. (2016, sep). Enhanced memory as a common effect of active learning. *Mind, Brain, and Education*, *10*, 142–152.

Martin, A. D., Quinn, K. M., Park, J. H., & Park, M. J. H. (2018). *Package mcmcpack.*

Meder, B., & Nelson, J. D. (2012). Information search with situation-specific reward functions. *Judgment and Decision Making*, *7*, 119–148.

Nelson, J. D. (2005). Finding useful questions: On Bayesian diagnosticity, probability, impact, and information gain. *Psychological Review*, *112*, 979–999.

Nelson, J. D., Divjak, B., Gudmundsdottir, G., Martignon, L. F., & Meder, B. (2014). Children's sequential information search is sensitive to environmental probabilities. *Cognition*, *130*, 74–80.

Nelson, J. D., McKenzie, C. R., Cottrell, G. W., & Sejnowski, T. J. (2010). Experience matters: Information acquisition optimizes probability gain. *Psychological Science*, *21*, 960–969.

Osman, M., & Speekenbrink, M. (2012). Prediction and control in a dynamic environment. *Frontiers in Psychology*, *3*, 68.

Parpart, P., Schulz, E., Speekenbrink, M., & Love, B. C. (2017). Active learning reveals underlying decision strategies. *bioRxiv*, 239558.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237.

Schulz, E., & Gershman, S. J. (2019). The algorithmic architecture of exploration in the human brain. *Current Opinion in Neurobiology*, *55*, 7–14.

Schulz, E., Konstantinidis, E., & Speekenbrink, M. (2018). Putting bandits into context: How function learning supports decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*, 927–943.

Settles, B. (2009). *Active learning literature survey* (Computer Sciences Technical Report No. 1648). University of Wisconsin–Madison.

Sharma, B. D., & Mittal, D. P. (1977). New non-additive measures of relative information. *Journal of Combinatorics Information & System Sciences*, *2*, 122–132.

Strom, A. R., & Barolo, S. (2011). Using the game of mastermind to teach, practice, and discuss scientific reasoning skills. *PLoS Biology*, *9*(1), e1000578.

Wu, C. M., Schulz, E., Speekenbrink, M., Nelson, J. D., & Meder, B. (2018). Generalization guides human exploration in vast decision spaces. *Nature Human Behaviour*, *2*, 915–924.

Zhao, B., van de Pol, I., Raijmakers, M., & Szymanik, J. (2018). Predicting cognitive difficulty of the deductive mastermind game with dynamic epistemic logic models. In C. Kalish, M. Rau, J. Zhu, & T. Rogers (Eds.), *Cogsci 2018* (pp. 2789–2794).

Speaker-specific adaptation to variable use of uncertainty expressions

Sebastian Schuster and **Judith Degen**
{sebschu, jdegen}@stanford.edu
Department of Linguistics, Stanford University
Stanford, CA 94305, USA

Abstract

Speakers exhibit variability in their choice between uncertainty expressions such as *might* and *probably*. Recent work has found that listeners cope with such variability by updating their expectations about how a specific speaker uses uncertainty expressions when interacting with a *single speaker*. However, it is still unclear to what extent listeners form speaker-specific expectations for *multiple speakers* and to what extent listeners are adapting to a situation independent of the speakers. Here, we take a first step towards answering these questions. In Experiment 1, listeners formed speaker-specific expectations after being exposed to two speakers whose use of uncertainty expressions differed. In Experiment 2, listeners who were exposed to two speakers with identical use of uncertainty expressions formed considerably stronger expectations than in Experiment 1. This suggests that listeners form both speaker-specific and situation-specific expectations. We discuss the implications of these results for theories of adaptation.

Keywords: psycholinguistics; semantics; pragmatics; adaptation; uncertainty expressions

Introduction

Speakers exhibit considerable production variability at all levels of linguistic representation (e.g., Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967; Weiner & Labov, 1983; Finegan & Biber, 2001). This includes variation in lexical choice to describe a world state. For example, Yildirim, Degen, Tanenhaus, and Jaeger (2016) found that when asked to describe a scene with a candy bowl in which approximately half of the candies were green and half of the candies were blue, some participants judged "Some of the candies are green" to be the more appropriate utterance to describe the scene than "Many of the candies are green", while others displayed the opposite pattern.

Schuster and Degen (2018) found that participants exhibit similar production variability when describing an event with an objective event probability of 60%: Some participants judged the event to be best described with a sentence containing the uncertainty expression *might* ("You might get a blue gumball") whereas others judged a sentence with *probably* ("You'll probably get a blue gumball") more appropriate.

Such variability poses a challenge to a listener who aims to know what the world is like that the speaker is describing. When confronted with two speakers who use the same expression to convey different states of the world or who use different expressions to convey the same state of the world, listeners are doomed to draw the wrong inferences about the actual state of the world unless they track how individual speakers use language. Recent work suggests that listeners deal with this kind of variability by adapting to it (e.g., Norris, McQueen, & Cutler, 2003; Kraljic & Samuel, 2007; Bradlow & Bent, 2008; Kamide, 2012; Kleinschmidt & Jaeger, 2015; Fine & Jaeger, 2016; Roettger & Franke, submitted) and that in interaction, they learn how speakers choose among alternative utterances. In the domain of quantifiers, Yildirim et al. (2016) showed that listeners update their expectations about how a specific speaker uses the quantifiers *some* or *many* after being briefly exposed to a specific speaker. In line with their results, Schuster and Degen (2018) found that listeners update their expectations of how a specific speaker uses the uncertainty expressions *might* and *probably* to describe different event probabilities after a brief exposure phase. Participants who were exposed to a "*confident*" speaker, who used *probably* to describe the 60% probability event, expected the use of *probably* with a wider range of probabilities; participants who were exposed to a "*cautious*" speaker, who used *might* to describe the 60% probability event, expected the use of *might* with a wider range of probabilities.

The processes that lead listeners to update their expectations during semantic adaptation are poorly understood. In particular, it remains a largely open question to what extent listeners form speaker-specific expectations when interacting with multiple speakers. Some evidence for speaker-specific adaptation comes from the referring expressions literature. Metzing and Brennan (2003) found that participants exhibited a slowdown in resolving referring expressions when a confederate started referring to an object with a new expression after establishing a conceptual pact, but did not find such a slowdown when a new confederate was using a different referring expression than the original confederate.

Most closely related to our work, Yildirim et al. (2016) found that listeners form speaker-specific production expectations after being exposed to two speakers who used different quantifiers to describe a scene with a candy bowl in which half of the candies were green. While this suggests that listeners should also form speaker-specific expectations about the use of uncertainty expressions, there is evidence from other linguistic domains that speaker-specific adaptation is limited to specific items. For example, Kraljic and Samuel (2007) found that listeners adjust their phonemic representations for the fricatives /s/ and /sh/ to multiple speakers whereas lis-

teners adjusted their phonetic representations for stop consonants such as /d/ and /t/ only to the most recent conversational partner. It could therefore be that speaker-specific adaptation in other linguistic domains is also limited to specific items and that listeners do not form speaker-specific expectations for the use of uncertainty expressions.

Further, Yildirim et al. (2016) observed that the adaptation effect was considerably smaller when they exposed participants to two speakers with opposing biases as compared to only exposing participants to one speaker and comparing the adaptation effect between groups. There seem to be two likely explanations for this observation. First, it could be that due to memory limitations, listeners were unable to keep track of the exact statistics of each speaker's utterances. Since everything about the context except the speaker identity stayed constant throughout the experiment, it could be that listeners had difficulty separating their experiences with the two speakers in memory (see Horton and Gerrig (2005) for a similar account of memory limitations affecting audience design). Second, it could be that listeners were tracking the statistics of the individual speakers as well as the overall statistics in the experimental situation and their post-exposure expectations were a combination of their speaker-specific expectations as well as their expectations about the situation.

In this work, we build on the recent work by Schuster and Degen (2018) on adaptation to variable use of uncertainty expressions and take a first step towards investigating the nature of semantic adaptation in response to multiple speakers. In particular, we aim to answer the following two questions:

1. Do listeners form speaker-specific production expectations when they are exposed to speakers whose use of uncertainty expressions differ?

2. Do listeners form situation-specific production expectations independent of speaker identity?

In Experiment 1, we address question 1 by exposing listeners to two speakers whose use of uncertainty expressions differs. In Experiment 2, we expose listeners to two speakers whose use of uncertainty expressions is the same. We compare adaptation effect sizes across experiments to address question 2.

Experimental paradigm

In both of our experiments, we build upon the semantic adaptation paradigm used in Schuster and Degen (2018), which we briefly review here. This paradigm is a classic exposure-and-test paradigm which has been used to study adaptation across several linguistic domains (e.g., Norris et al., 2003; Kleinschmidt & Jaeger, 2015; Yildirim et al., 2016). As shown in Figure 1, each trial shows an adult sitting behind a table with a gumball machine on it. The gumball machine is filled with orange and blue gumballs. Next to the table, there is a child who is requesting a blue or an orange gumball with the utterance "I want a blue/an orange one". Participants are told that the gumball machine is too high up for the child to see and that only the adult can see the contents of the gumball machine.

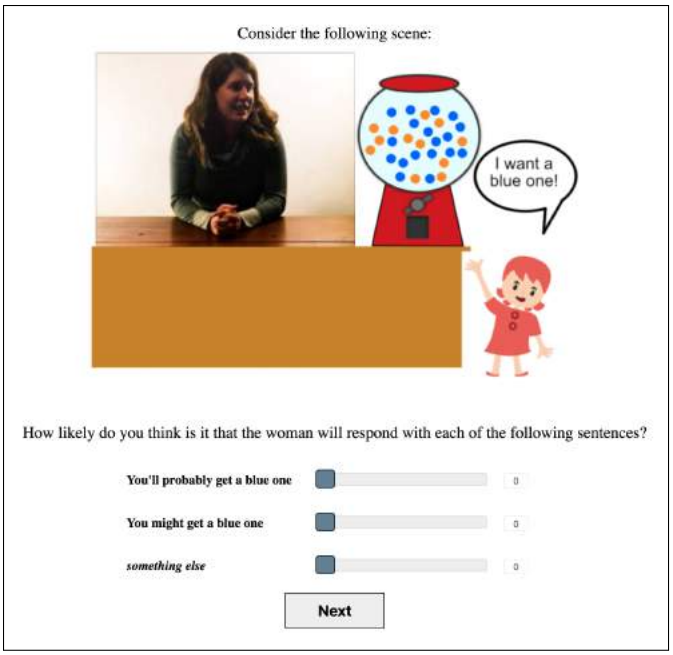


Figure 1: Example post-exposure test trial. On exposure trials the rating scales were absent, and the image of a speaker was replaced by a video of a speaker producing an utterance.

On each exposure trial, participants watch a short video clip in which the adult responds to the child with an utterance like "You might get a blue one". Across trials, the proportion of gumballs as well as the response by the adult vary.

On each test trial (Fig. 1), participants are shown a static scene in which they only see a picture of the speaker from the exposure trials. On these trials, participants are asked to provide ratings of how likely they think it is that the speaker would use the two provided utterances or some other utterance. Across trials, the proportion of blue and orange gumballs as well as the color of the gumball that the child is requesting (the target color) varies.

Experiment 1: Different speaker types

In Experiment 1, we exposed participants to two different speakers who use the uncertainty expressions *might* and *probably* differently. The primary purpose of this experiment was to test whether listeners form speaker-specific utterance choice expectations. Procedure, materials, analyses and exclusions were pre-registered on OSF (https://osf.io/qnspg).

Methods

Participants We recruited 104 participants on Amazon Mechanical Turk. Participants had to have a US-based IP address and a minimal approval rating of 95%, and they were paid $4.75 (approximately $12–$15/hr).

	MIGHT		PROBABLY		BARE	
	n	*p*	*n*	*p*	*n*	*p*
cautious	**10**	**60%**	5	90%	5	100%
confident	5	25%	**10**	**60%**	5	100%

Table 1: Number of exposure trials (n) per utterance (MIGHT, PROBABLY, BARE) and associated proportion of target gumballs (p) in the cautious vs. confident speaker block. Critical trials bolded.

Materials and procedure In the first part of the experiment, participants saw 40 exposure trials in two blocks. As mentioned above, each trial showed a child requesting a blue or orange gumball, a gumball machine with blue and orange gumballs, and a video of an adult male or female speaker. The speaker always produced one of the following six utterances:

- You'll get a blue/orange one (BARE)
- You might get a blue/orange one (MIGHT)
- You'll probably get a blue/orange one (PROBABLY)

The number of trials with each of these utterances as well as the gumball proportions varied across the two blocks (see Table 1 for an overview). Filler trials with the bare form were included to provide evidence that the speaker is generally cooperative. One of the blocks always showed a female speaker and the other block always showed a male speaker. Both speakers were from the East Coast and native speakers of American English. The order of blocks and the speaker assignment to blocks was counterbalanced across participants.

Participants were instructed to watch what the speaker had to say to the child. The video started automatically after a 400ms delay and participants had the option to replay the video as often as they wanted. To advance, participants had to press a button which was disabled until the video had ended.

After the two exposure blocks, participants went through two test blocks. In each of the blocks they saw a picture of one of the two speakers with a gumball machine next to it, and again, a child requesting a blue or an orange gumball. On each trial, participants were asked how likely they thought it was that the adult would respond with MIGHT, PROBABLY or a blanket *something else* option. Participants indicated their expectations by distributing 100 points across these three options using sliders. In each block, participants provided ratings for scenes with 9 different gumball machines ranging from 0% to 100% blue gumballs. For each machine, participants provided four ratings in total, resulting in 36 trials per block. The order of blocks was counterbalanced such that half of the participants saw them in the same order as the exposure blocks whereas the other half saw them in opposite order.

Attention checks To verify that participants were paying attention to the video and the scenes, we included 14 attention checks: after 14 of the exposure trials, participants were shown two different gumball machines and were asked to choose the one they saw on the previous trial.

Exclusions We excluded participants who provided correct responses to fewer than 11 attention checks. Based on this criterion, we excluded 31 participants. We further excluded participants whose utterance ratings for the different event probabilities strongly correlated ($R^2 > 0.75$) with their mean utterance ratings across all event probabilities. This suggests that they provided approximately the same ratings independent of the observed scenes and indicates that they did not pay attention. This led to one additional exclusion. None of the results discussed below depend on these exclusions.

Analysis and predictions Intuitively, a more confident speaker uses PROBABLY for a larger and MIGHT for a smaller range of gumball proportions than a more cautious speaker. Therefore, if participants track these different uses, we expect their ratings of what they think a specific speaker is likely to say to depend on how that speaker used uncertainty expressions during the exposure phase. Following Yildirim et al. (2016) and Schuster and Degen (2018), we quantify this prediction by fitting a spline with four knots for each expression and each participant and computing the area under the curve (AUC) for the splines corresponding to each expression, block and participant. The area under the curve is proportional to how highly and for how large of a range of gumball proportions participants rate an utterance, so if an utterance is rated highly for a larger range of gumball proportions, the AUC will also be larger. We therefore test whether listeners update their expectations by computing the difference between the AUC of the spline for MIGHT and of the spline for PROBABLY for each test block for each participant.

Based on the results of the adaptation experiment with multiple speakers by Yildirim et al. (2016), we expect speaker-specific adaptation effects. We thus predict that the mean AUC difference will be bigger for the *cautious* speaker test blocks than for the *confident* speaker test blocks.

As a secondary analysis, we also investigate whether the order of exposure blocks (*confident* or *cautious* first), the assignment of speaker to speaker type (whether the male speaker was the *cautious* speaker or vice versa), or the order of the test blocks (same as exposure or reverse) has an effect on adaptation. We do not expect any of these factors to have an effect on adaptation.

Results and discussion

Figure 2 shows the mean utterance ratings of participants grouped by the two post-exposure test blocks. As this plot shows, participants expected the *confident* speaker to be more likely to use *probably* for lower event probabilities than the *cautious* speaker. This is also reflected in the AUC differences between the splines for MIGHT and of the splines for PROBABLY: As predicted, this difference was greater for the *cautious* speaker ratings than for the *confident* speaker ratings ($t(142) = 2.92, p < 0.01$).

For our secondary analysis, we fit a linear regression model to predict the AUC difference with speaker type, exposure

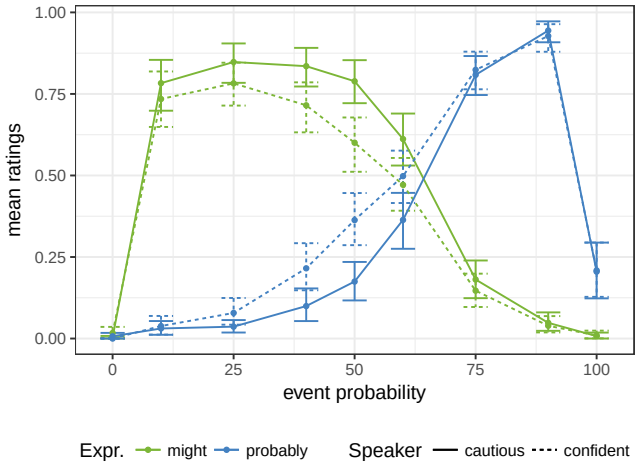


Figure 2: Mean utterance ratings for scenes with different event probabilities in Experiment 1. Error bars indicate bootstrapped 95% confidence intervals.

block order, speaker assignment, and test block order as predictors. Only speaker type is a significant predictor in this model (exposure block order: $\beta = 5.72$, $t(139) = 1.30$, *n.s.*; speaker assignment: $\beta = 1.21$, $t(139) = 0.28$, *n.s.*; test block order: $\beta = 2.28$, $t(139) = 0.52$, *n.s.*). Further, a model that includes these four predictors does not explain significantly more variance than a model that only includes speaker type as a predictor ($F(3,139) = 0.67$, *n.s.*).

The results of this experiment suggest that listeners form speaker-specific expectations of how different speakers use uncertainty expressions after brief exposure. At the same time, the results provide concrete evidence against two other accounts. First, they provide evidence against an account according to which participants only adapt to the experimental situation: If participants had only updated their expectations of what a generic speaker would say in the scenes presented in the experiment, we would not have expected to see differences in ratings between speakers. Second, they also provide evidence against a pure priming account according to which listeners update their expectations to the most recent exposure. Note that the adaptation effect was independent of the order of presentation and the order of test blocks. If participants had been primed by the most recent exposure speaker, we would have expected that participants' post-exposure ratings were primarily influenced by the behavior of the second exposure speaker.

The results of this experiment also replicate the finding by Yildirim et al. (2016) of differing effect sizes between the single-speaker and two-speaker experiments: The adaptation effect was considerably smaller in this two-speaker experiment (Cohen's d: 0.486) than in the single-speaker adaptation experiment by Schuster and Degen (2018) (Cohen's d: 1.263).

As suggested by a reviewer, one reason for the smaller effect size in the two-speaker experiment could be some form of self-priming and that participants' responses in the first test block influenced their responses in the second block. We evaluated this hypothesis in a post-hoc analysis of the responses from the first test block. We compared the responses of participants who were first tested on the *cautious* speaker to the responses of participants who were first tested on the *confident* speaker. If responses in the first test block influenced responses in the second test block, we would expect a larger effect size if we only consider the data from the first block. We did indeed find a larger effect size in the first block (Cohen's d: 0.723), which suggests that participants exhibited some form of self-priming.

However, even if we only consider the first block of responses, the adaptation effect remains smaller in the two-speaker experiment (Cohen's d: 0.723) than in the one-speaker experiment (Cohen's d: 1.263). This could be either a result of memory limitations or a result of listeners jointly tracking the statistics of each speaker as well as of the overall experimental situation (situation-specific statistics). We further investigate these possibilities in the next experiment.

Experiment 2: Identical speaker types

In Exp. 1, we found that the adaptation effect was smaller than it was in the single-speaker version of the experiment, which could have either been a result of memory limitations or joint speaker-specific and situation-specific adaptation. In this experiment, we investigate whether there is evidence for one of these two accounts. We exposed listeners to two speakers of the same type.[1] If the smaller effect in Exp. 1 was caused by listeners' inability to separate their experiences with the two speakers in memory, i.e, some experiences might have been attributed to the incorrect speaker, we would expect the adaptation effect in this experiment to be on average the same as in the one-speaker experiment. This is because even if listeners cannot perfectly separate their experiences with each speaker, they would on average still have the same number of experiences with each of the two speakers as listeners had with the one speaker in the single-speaker experiment. If, on the other hand, the smaller effect in the previous experiment was a result of listeners jointly tracking speaker-specific and situation-specific statistics, we would expect the adaptation effect to be larger here than in the single-speaker experiment. This is based on the assumption that more exposures lead to a larger adaptation effect and thus listeners' should adapt more to the situation if they are exposed to two

[1] In the spirit of open science, we note that the data from this experiment comes from a faulty version of Experiment 1. A scripting error led to participants always being exposed to the same speaker type instead of two different speaker types. Because of this error, the pre-registered analysis (https://osf.io/3cw79) deviates from the analysis that we report here. The reported analyses here are the only additional analyses we performed on the data. The reason for not discarding the data from this experiment but rather including it here is that it provides an informative data point for the question of whether listeners track situation-specific expectations.

speakers and hence also twice the number of interactions.

Methods

Participants We recruited 104 participants on Amazon Mechanical Turk. Participants had to have a US-based IP address and a minimal approval rating of 95%, and they were paid $5 (approximately $12–$15/hr).

Materials and procedure The materials and procedures were the same as in Exp. 1 except for the following two modifications. First, the speaker types for each participant were identical across the two exposure blocks: both speakers were either *confident* or *cautious* speakers. Second, the number of trials with PROBABLY and the number of trials with MIGHT were the same (10 trials per utterance and block) whereas in Experiment 1, the *confident* speaker produced only 5 instances of MIGHT and the *cautious* speaker produced only 5 instances of PROBABLY.[2] Assignment of speaker types was counterbalanced across participants, which means this experiment had a between-subjects manipulation.

As in Experiment 1, we excluded participants who provided correct responses to less than 11 of the attention checks as well as participants who seemed to provide random responses as defined above. In total, we excluded 11 participants because of the attention check criterion and 1 more participant because of random responses.

Analysis and predictions As the primary analysis, we compare the AUC differences between the splines for MIGHT and of the splines for PROBABLY between participants in the two conditions. Analogous to Experiment 1, we predict that the mean AUC difference will be bigger in the *cautious* speaker condition than in the *confident* speaker condition.

We again also investigate whether the assignment of speaker to speaker type or the order of the test blocks have an effect on the AUC difference. We do not expect either of these factors to affect adaptation.

Lastly, we compute the effect size measured by Cohen's d. As explained above, we expect the effect size either to be the same as in the single-speaker experiment or to be larger.

Results and discussion

Figure 3 shows the mean utterance ratings of participants for the two conditions. We again observe listener adaptation, resulting in a greater AUC difference in the *cautious* speaker condition than in the *confident* speaker condition ($t(89) = 8.01, p < 0.001$). Further, no factors other than speaker type are significant predictors of the AUC difference (speaker assignment: $\beta = -1.32$, $t(87) = -0.398$, *n.s.*; test block order: $\beta = 4.28$, $t(139) = 1.30$, *n.s.*).

Lastly, the effect size (Cohen's d: 1.68) was larger in this experiment than in Experiment 1 and the single-speaker experiment by Schuster and Degen (2018). While it would be premature to definitively conclude from these three experiments that listeners' expectations are jointly influenced by individual speaker's productions as well as all the productions in the experiment, our results point in this direction.

[2]The reason for the second modification is the above mentioned scripting error. See below for a discussion of potential implications.

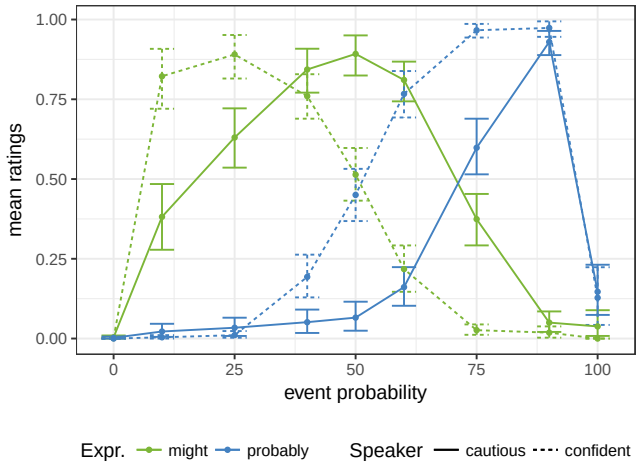


Figure 3: Mean utterance ratings for scenes with different event probabilities in Experiment 2. Error bars indicate bootstrapped 95% confidence intervals.

There is a potential confound in this experiment because participants saw 5 additional filler trials during each exposure block which could have led to the larger effect size as compared to the single-speaker experiment. However, this explanation seems unlikely considering previous work.[3]

General discussion and conclusion

In two experiments, we found that listeners form speaker-specific production expectations of uncertainty expressions after brief exposure to two speakers. This shows that the results by Yildirim et al. (2016) also extend to lexical items other than quantifiers.

At the same time, however, we found that the adaptation effect size varied depending on whether the two speakers had the same or divergent bias during the exposure phase. When listeners were exposed to two different speaker types, the adaptation effect was smaller and their expectations seemed to have been shaped by their experiences with the two speakers as well as all the experiences encountered in the experiment. When both speakers behaved the same, on the other hand, the adaptation effect was much more pronounced and even greater than in the single-speaker experiment from previous work.

[3]Yildirim et al. (2016) used a very similar paradigm to study semantic adaptation to the use of the quantifiers *some* and *many*. Analagous to our *confident* and *cautious* speakers, they had a *some-biased* and a *many-biased* speaker. They report two versions of their experiment: one in which there were no filler trials with the other quantifier and another version in which there was a balanced number of exposure trials with both quantifiers in both conditions. They found that the adaptation effect was smaller when there were more filler trials, so we would expect that if the additional fillers affected the size of the adaptation effect, the effect would be even larger had we not presented the extra fillers to participants.

One likely explanation for these observations is that apart from tracking speaker-specific statistics, listeners also track the situation-specific statistics of all interactions in the experiment and their expectations are guided by both of these factors. In the case of speakers with different uses of uncertainty expressions, speaker-specific adaptation is attenuated since the overall statistics guide listeners towards an "average" speaker whose use falls somewhere in between the *cautious* and the *confident* speaker. When listeners are exposed to two speakers of the same type, on the other hand, the situation-specific statistics reinforce the speaker-specific statistics and hence listeners adapt more to the two speakers.

An account based on "faulty" memory, according to which listeners have trouble keeping the speaker-specific experiences separate, does not predict the larger adaptation effect when listeners are exposed to two speakers of the same type. If every experience is encoded as an episode in memory but some with the incorrect speaker information, on average, the number of experiences with each speaker should still be the same as in the one-speaker condition and therefore it is unclear why listeners adapt more in the two-speaker experiment than in the one-speaker experiment.

Our findings also have implications for current models of semantic adaptation. Following the recent successes in modeling phonetic adaptation as an instance of Bayesian belief updating (Kleinschmidt & Jaeger, 2015), Schuster and Degen (2018) propose a computational model of semantic adaptation. According to this model, when interacting with a speaker Sp, listeners update their beliefs about a set of speaker-specific parameters Θ_{Sp}, which govern the speaker's lexicon and preferences.[4] Their model predicted the results of the single-speaker experiment well, but without modifications, it does not predict the differences in effect size.

We consider two promising extensions of this model. First, the model could be cast as a hierarchical model. Hierarchical models have been argued to explain many cognitive and perceptual phenomena (see e.g., Clark, 2013, for a review), including phonetic adaptation (Kleinschmidt, 2019), and also seem applicable here. In a hierarchical version of the adaptation model, we would assume that the speaker-specific parameters Θ_{Sp} are not only shaped by the listener's prior beliefs and the observed interactions by a speaker Sp but rather also depend on a distribution reflecting the situation-specific expectations. Figure 4 shows a sketch of a potential hierarchical model. Such a model would explain the differences in effect size: When listeners are exposed to different speaker types, the situation-specific parameter distribution would be influenced by two speaker types that essentially cancel each other out, which in turn would lead to less extreme speaker-specific distributions. On the other hand, when both of the speakers are of the same type, the situation-specific parameter distribution would be more strongly shifted towards the observed distributions which in turn would lead to more extreme speaker-specific distributions.

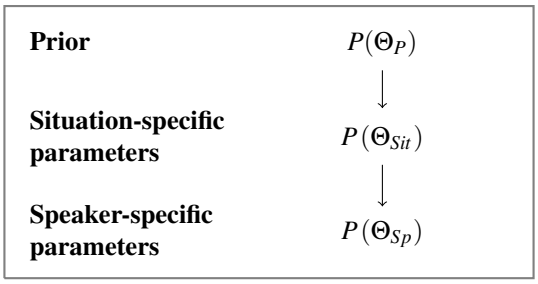


Figure 4: Hierarchical model of semantic adaptation. Situation-specific parameters $P(\Theta_{Sit})$ depend on prior beliefs $P(\Theta_P)$ and speaker-specific parameters $P(\Theta_{Sp})$ depend on the situation-specific parameters.

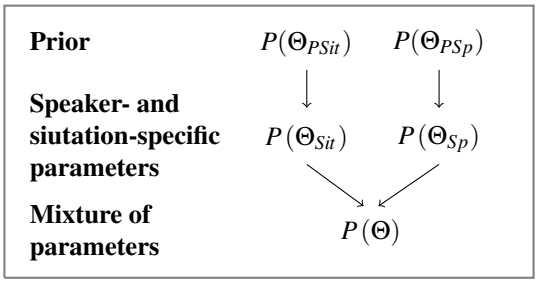


Figure 5: Mixture model of semantic adaptation. Overall production parameters $P(\Theta)$ are a weighted combination of situation-specific parameters $P(\Theta_{Sit})$ and speaker-specific parameters $P(\Theta_{Sp})$.

A second possibility would be to cast the model as a mixture model in which overall production parameters are a weighted combination of situation-specific and speaker-specific parameters (and potentially other factors). Figure 5 shows a sketch of a potential mixture model. According to such a model, listeners would form both situation-specific and speaker-specific expectations as a result of adaptation and then combine these expectations to their overall expectations. Such a model would also predict the smaller effect size in Experiment 1 since it would predict that the overall production expectations are influenced by the speaker-specific statistics as well as the situation-specific statistics and the latter drive the production expectations to be more similar to an "average" speaker. When listeners are exposed to two identical speakers, on the other hand, the situation-specific expectations (which are in line with the speaker type of both exposure speakers) would reinforce the speaker-specific expectations and therefore lead to a larger adaptation effect. Future experimental work should adjudicate between the hierarchical and the mixture model account.

In conclusion, we presented new experimental results from the domain of uncertainty expressions which suggest that speaker-specific semantic adaptation is a product of forming speaker-specific expectations and forming expectations about

[4] See also Hawkins, Frank, and Goodman (2017) for a similar model of the formation of conceptual pacts.

the situation independent of the speaker. These results raise a number of interesting questions, most pressingly regarding transfer effects to novel speakers, which have been observed in other linguistic domains (e.g., Bradlow & Bent, 2008; Xie, Earle, & Myers, 2018). In our experiments, the exposure and test speakers did not differ. This raises the question about whether and to what extent updated expectations transfer to novel speakers whose similarity to the exposure speaker(s) varies. Both models sketched above lend themselves well to capturing such transfer effects. In addition, participants saw very similar visual scenes on each trial. Another potential direction would be to study the extent of speaker-specific adaptation when listeners encounter more novel scenes during the test phase to investigate to what extent listeners form speaker-specific expectations independent of other contextual factors. Answering these questions will help disentangle the different adaptation processes and give us a better understanding of how listeners infer meanings in context.

References

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, *106*(2), 707–729.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181–204.

Fine, A. B., & Jaeger, T. F. (2016). The role of verb repetition in cumulative structural priming in comprehension. *Journal of Experimental Psychology: Learning Memory and Cognition*, *42*(9), 1362–1376.

Finegan, E., & Biber, D. (2001). Register variation and social dialect variation: The register axiom. In *Style and Sociolinguistic Variation* (pp. 235–267).

Hawkins, R. X. D., Frank, M. C., & Goodman, N. D. (2017). Convention-formation in iterated reference games. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society (CogSci 2017)* (pp. 482–487).

Horton, W. S., & Gerrig, R. J. (2005). The impact of memory demands on audience design during language production. *Cognition*, *96*(2), 127–142.

Kamide, Y. (2012). Learning individual talkers' structural preferences. *Cognition*, *124*(1), 66–71.

Kleinschmidt, D. F. (2019). Structure in talker variability: How much is there and how much can it help? *Language, Cognition and Neuroscience*, *34*(1), 43–68.

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, *122*(2), 148–203.

Kraljic, T., & Samuel, A. G. (2007). Perceptual adjustments to multiple speakers. *Journal of Memory and Language*, *56*(1), 1–15.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*(6), 431–461.

Metzing, C., & Brennan, S. E. (2003). When conceptual pacts are broken: Partner-specific effects on the comprehension of referring expressions. *Journal of Memory and Language*, *49*(2), 201–213.

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *47*(2), 204–238.

Roettger, T. B., & Franke, M. (submitted). Evidential strength of intonational cues and rational adaptation to (un-)reliable intonation. *Cognitive Science*.

Schuster, S., & Degen, J. (2018). Adaptation to variable use of uncertainty expressions. In *Proceedings of Architectures and Mechanisms for Language Processing 2018 (AMLaP 2018)*.

Weiner, E. J., & Labov, W. (1983). Constraints on the agentless passive. *Journal of Linguistics*, *19*(1), 29–58.

Xie, X., Earle, F. S., & Myers, E. B. (2018). Sleep facilitates generalisation of accent adaptation to a new talker. *Language, Cognition and Neuroscience*, *33*(2), 196-210.

Yildirim, I., Degen, J., Tanenhaus, M. K., & Jaeger, T. F. (2016). Talker-specificity and adaptation in quantifier interpretation. *Journal of Memory and Language*, *87*, 128–143.

How does a doll play affect socio-emotional development in children?: Evidence from behavioral and neuroimaging measures

Kazuki Sekine[1] (kazuki.sekine@keio.jp)
Eriko Yamamoto[1] (eyamamoto@keio.jp)
Saeka Miyahara[1] (m1yahara@keio.jp)
Yasuyo Minagawa[1] (minagawa@flet.keio.ac.jp)

[1] Keio University, Faculty of Letter, Department of Psychology, Tokyo JAPAN

Abstract

Mentalization is an important ability to acquire for children, as it allows humans to understand the mental state of others or oneself, that underlies overt behavior (Fonagy & Target, 1996). In the current study we examined the relationship between development of mentalization ability in children and their experience of playing with a doll by observing child-mother interaction and by using functional near-infrared spectroscopy (fNIRS). 44 dyads of children aged 2 to 3 and their mothers were divided into two groups (high and low) depending on the frequency of doll-play experience. We examined mother-speech interaction during the doll play. We also used fNIRS system to measure cerebral hemodynamic activation in the frontal and temporal regions during the observation of video clips showing hindering and helping behaviors. The results showed that a mother's proxy talk was related to a child's doll directed speech in the high group, but not in the low group. fNRIS data showed that cerebral activation in the helping condition was more increased in the low group than the high group. This suggests that doll-play experience facilitates the development of mentalization, which enables children to be aware of and understand other's psychological states.

Keywords: doll play; social understanding; mentalizing, young children; fNIRS.

This study investigated the relationship between children's experience in playing with dolls and the development of mentalization by using behavioral and neuroimaging measures. Mentalization refers to an ability to speculate and to understand other's psychological states (e.g., needs, desires, feelings, beliefs, goals, and reasons) based on their behavior (Fonagy, Gergely, & Target, 2007). Development of mentalization is important for children, as they need to interact with others by assuming other's mental sates in their socialization process. Fonagy and Target (1996) have suggested that play provides an intermediate area for the acquisition of symbolic thinking which is crucial for mentalization. Given this, it is important to see the relationship between play and the development of mentalization.

Children around age 1 begin to play by using an object as if it were something else or by pretending as if he or she was doing an actual action without the visible object. This kind of play is called "pretend play'. Pretend play is defined as a play expressing internal images by using actions, words, or objects, such as pretending to drink water by moving an empty cup to her mouth or feeding a doll by moving an empty toy fork to a doll's mouth (Lillard, Lerner, Hopkins, Dore, Smith, & Palmquist, 2013). Research has investigated pretend play because it indicates the emergence of mental representation in children in the sense that they enact an event or represent an invisible object by using their own body or different objects during play.
Pretend play normally peaks around preschool years when children start interacting with other children and gain access to more toys and resources for play (Lindsey & Colwell, 2013). Pretend play during preschool age is particularly important as it is related to the development of language (e.g., Orr & Geva, 2015), executive function (e.g., Carlson, White, & Davis-Unger, 2014), and social understanding including theory of mind (e.g., Lillard & Kavanaugh, 2014). Theory of mind refers to the ability to attribute mental states to others in order to understand and predict social behavior. The difference between mentalization and theory of mind is that mentalization mainly concerns the reflection of affective mental states, whereas, theory of mind focuses on epistemic states such as beliefs, intentions and persuasions (Wyle, 2014).
However few studies have shown the relationship between children's experience in pretend play and the development of mentalization. The current study addressed this issue.

Sachet and Mottweiler (2013) emphasized the distinction between two types of pretend play; Role-play and Object Substitution. *Role play* refers to pretend play that involves the mental representation of social or interpersonal content (e.g., pretending that a doll likes to eat sweets), whereas *object substitution* refers to pretend play that involves the mental representation of nonsocial content (e.g., pretending that a block is a chocolate). Both types of pretend play can provide opportunities for children to practice social skills or events happenings in the real world. Role-play has a special significance in the development of social understanding, because it provides opportunities for simulating social interaction (Harris, 2000). In fact, this was demonstrated by Wolf, Rygh, and Altshuler (1984). They visited children's houses from ages 1 to 7 and recorded their play with replica toys. They found that by the age of four, children can ascribe complicated psychological states including perceptions, sensations, emotions, and thinking to figures with which they are playing (Wolf et al., 1984). However there are three limitations in the previous studies on pretend play.

First, little research has been conducted to address how doll play affects social development such as mentalization, sympathy or prosocial behavior. Most studies have focused on object substitution and how it is related to social or cognitive development. Given that doll play provides opportunities for simulating social interaction (Harris, 2000), playing with a doll may foster children's social understanding. Brownell, Svetlova, Anderson, Nichols, and Drummond (2013) observed an interaction between toddlers and caregivers while reading a picture book in relation to toddlers' prosocial behavior. They found that children who helped and shared more tended to have parents who more often asked them to label and explain the emotions depicted in the books. This result suggests that caregiver's inputs that direct children's attention to inner thoughts or feelings of themselves or others assist the development of children's social understanding including mentalization.

Second, it is not clear how children's ability to ascribe the psychological states to dolls develops up until 4 years old. Lillard (2017) suggested that parent's input in pretend play is a crucial factor to develop children's social understanding because children need to learn how to pretend by properly interpreting social signals that parents send (e.g., strong eye contact or smile) as a cue of pretend play. Thus, it is worth investigating both children's and parent's behaviors during doll play to see how it affects the development of social understanding by the age of 4 years old.

Third, as Lillard (1993) pointed out, pretend play has been mostly analyzed by behavioral measures. There is no neuroimaging work on pretend play in children, although there are a few that have been done with adults (German et al., 2004; Whitehead et al., 2009). To see whether an experience in doll play affects the development of mentalization, the present study used functional near-infrared spectroscopy (fNIRS). Compared with other neuroimaging techniques, fNIRS imposes less physical constraints on the participant and it is relatively unaffected by motion artifact. Thus it can be applied in a natural setting even in young children (Nagamitsu, Yamashita, Tanaka, & Matsuishi, 2012). Previous studies have shown that medial prefrontal cortex (mPFC) and temporoparietal junction (TPJ) are involved in the mentalization process (Frith & Frith, 2006; Minagawa Xu, & Morimoto, 2018). Particularly mPFC is responsive when making social judgments about dissimilar others (Mitchell et al., 2005), whereas TPJ is activated more in response to theory of mind tasks (Mahy, Moses, & Pfeifer, 2014). Thus, if a doll-play experience facilitates the development of mentalization ability in children, these brain regions would be activated more in children having more experience in doll play than those who have less experience.

To address these three limitations, the current study aimed to reveal the relationship between doll-play experience and development of mentalization in children aged 2 to 3 by observing mother-child interaction and measuring fNIRS. We predicated that fNIRS data would show that brain areas involving the mentalizing process would be more activated in children having more doll-play than children having less experience when they see someone's helping/hindering behavior. We also predicted that mother-child interaction would be qualitatively different depending on children's doll-play experience.

Methods

Participants 44 female children aged 2 to 3 and their mothers participated in this study. They were divided into two groups in terms of frequency of doll play; high and low group. The playtime with a doll was taken via a questionnaire for mothers before the experiment. Each group included 22 children. Children in the High group play with a doll more than one hour per week, and children in the low group play with a doll fewer than 20 minutes per week. The mean age in months and the standard deviation for each group were as follows; Low group, M = 35.4, SD = 2.5, and High group M = 36.7, SD = 3.3. There was no significant difference in the average age between the two groups. All the participants were native monolingual Japanese speakers from middle-class families, and the children attended nursery schools in Tokyo, Japan.

Material and Apparatus The test consisted of two sessions; a doll play session and a fNIRS session. In the doll play session, the child-mother dyad participated in a 7-minute doll play session. We encouraged the child-mother dyad to play with a set of toys including a doll and replica of house items, as shown in the left panel in Figure 1. Experimenters recorded the child-mother interactions but did not participate in their play.

For fNIRS session, we created audiovisual video clips. Each video clip lasted 17sec (30 fps), presenting two girls making an event as shown in Figure 2. In total, we created nine clips (stories). After presenting a still picture for 0.5 second, each clip starts with the introduction phase (8.5sec) where one girl (girl A) is in need and the other girl (girl B) notices the trouble (e.g., girl A is looking for a pencil, and girl B notices it). Then each story ends with the ending phase (8sec), which has two types of endings; hindering vs. helping ending. In the hindering condition, girl B obstructs girl A's need (e.g., throwing the pencil away), whereas in the helping condition, girl B assists girl A's need (e.g., passing the pencil to girl A).

Figure 1. Toys used in doll play session (left panel) and a screenshot of data coding with ELAN (right panel).

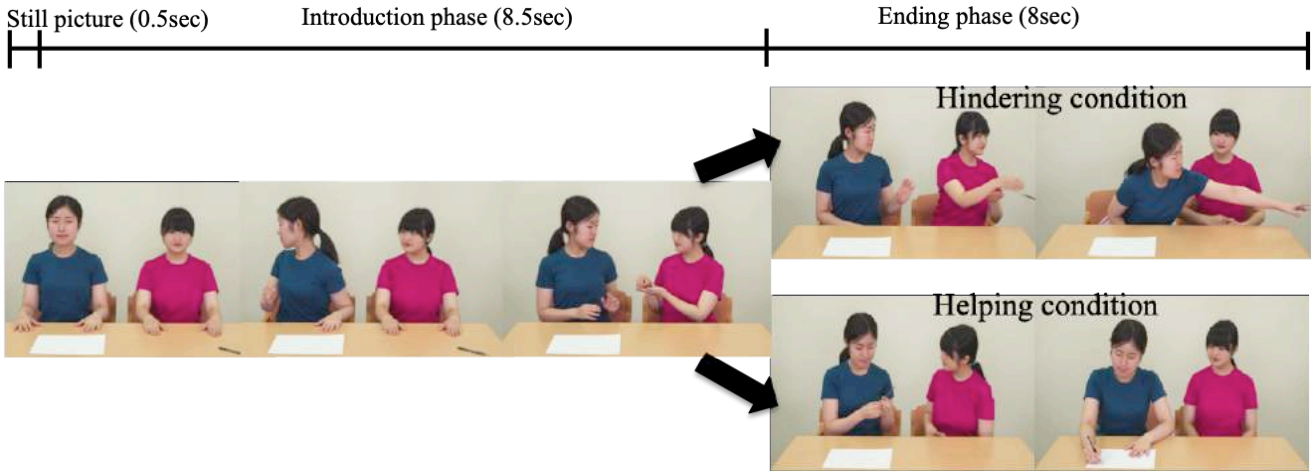


Figure 2. An example of video audio stimulus and the time line for fNIRS experiment. Each clip has two different endings; Hindering (top) and helping ending (bottom) in the Ending phase.

Procedure

The experiments were conducted in a test room with a sound-attenuated cabin for fNIRS in Keio University in Tokyo, Japan. First, each child-mother pair was instructed to play with a doll and some house items, shown in Figure 1, for seven minutes. They were encouraged to play in the same way as they do in their home. After the doll play session, they had a short break and moved to a booth with the fNIRS device. In the booth, mothers were instructed to hold their children on their laps and not interact with them during the fNIRS data collection. Children were instructed to watch the video clips on the monitor in front of them. The whole session was recorded using the mini-DV camcorder on a tripod. We counterbalanced the number of ending types that children were presented with, the locations of girl A and B, and the role of girls (e.g., one child watched the girl A helping in the pencil story, but other child watched the girl B helping in same story).

Analysis

Doll play data All narratives were verbatim transcribed. From the transcriptions, the mean number of utterances was then calculated. In this study an utterance was defined as a breath group. A breath group refers to a stretch of speech between two interword pauses, lasting 200ms or longer. We excluded an utterance from this study when it consisted of only fillers or meaningless exclamations such as "ah" or "um".

We counted the number of the following three speech types (*Desires, Emotion labels,* and *other internal state talk*) by using coding software ELAN (Lausberg & Sloetjes, 2009) (right panel in Figure 1). These categories were borrowed from a study by Brownell et al. (2013). *Desires* are references to wanting, or needing something concrete such as "he wants to eat an apple" or "she needs to go to bed". *Emotion labels* are defined as an utterance naming emotional feelings or behaviors without expansion or elaboration such as "the doll is happy" or "she likes vegetables". Other internal state talk is references to other internal states that are not affect- or mental state-related (e.g., physiological states) such as "she is thirsty" or "the doll is tired".

To see the psychological distance between participants and the doll, we also counted the number of instances of *proxy talk* and of *doll directed speech*. Proxy talk refers to an utterance when the speaker says something from doll's perspective, just like a ventriloquist, as if she or he is the doll (e.g., "oh I am so hungry, can you make a meal for me?"). In addition to the content of the utterance, if the pitch of the speaker's voice heightens higher than their usual pitch and/or he or she produced the utterance while operating the doll, it was counted as proxy talk. Doll directed speech refers to an utterance that directly addresses to the doll (e.g., "I will cook something for you").

fNIRS data We measured changes in concentrations of oxy-hemoglobin (oxy-Hb) and deoxy-hemoglobin (deoxy-Hb) in the frontal and temporal regions during the observation of video clips, using the NIRS system (ETG-7000, Hitachi). The NIRS system measures temporal changes in concentrations of oxy-Hb and deoxy-Hb in the cerebral cortex resulting from an increase in local cerebral blood flow by emitting and detecting two wavelengths of near-infrared light (780 nm and 830 nm). We used a 2 x 11 optode array, containing 27 measurement channels. The center optode in low row was placed on Fpz in the international 10-20 system to cover the frontal and temporal regions (Figure 5). The distance between each emitter and the corresponding detector was set at 2.5 cm.

Data was preprocessed using a platform for optical topography analysis tools (POTATo) developed by Research and Development Group, Hitachi, Ltd, within MATLAB2012 (Mathworks, Natwick, MA, USA). Pulse-related signal changes for head motion and overall trends were eliminated by high-pass (0.02 HZ) and low-pass (1 Hz) filtering. We defined 3.5 seconds before the onset of the ending phase as a baseline period and compared the relative change in oxy-Hb during a time analysis window (between 5 s and 8 s after the onset of the ending phrase) with the baseline period using a t-test. It is controversial which chromophore, namely oxy-Hb or deoxy-Hb best represents

BOLD (blood oxygen level dependent) signal. However, oxy-Hb has been dominantly used in previous fNIRS studies (Lloyd-Fox et al., 2010) and it has been pointed out that signal-to-noise ratio is higher for oxy-Hb rather than deoxy-Hb (Strangman et al. 2002). Thus, we decided to analyse only oxy-HB in this study.

Reliability

The first author coded the entire data set. To ensure the reliability of the gesture coding, about 50% of the data was re-analysed by a trained and independent native Japanese-speaking student. Ten children and mothers from each group (40 participants in total) were randomly selected and re-coded by the second coder. Point-to-point percentage agreement was calculated. The two coders agreed on the number of utterances 98% of the time for children, 97% of the time for mothers in the low group, and 98% of the time for children, 98% of the time for mothers in the high group. We calculated the percentage agreement for each speech category by collapsing groups. The two coders agreed on the number of doll directed speech 90% of the time in the low group and 93 % of the time in the high group, and on the number of proxy talk 93% of the time in the low group and 88% of the time in the high group. The Cohen's kappa statistic was used to assess inter-rater reliability for coding with more than two categories. Agreements between the two independent coders were overall high; for the low group, for desire speech kappa=.91; for emotion label kappa=.89; for other internal state talk kappa=.93, and for the high group, for desire speech kappa=.94; for emotion label kappa=.96; for other internal state talk kappa=.95. Any coding disagreements were resolved through discussion and subsequent consensus.

Results

1.1. Doll play analysis

The number of utterances We first calculated the number of utterances that children and mothers produced for each group. Children produced 46.3 (SD = 19.6) utterances in the low group, and 57.2 (SD= 26.9) in the high group. Mothers produced 121.3 utterances (SD = 26.3) in the low group, and 115.5 (SD = 26.8) in the high group. We conducted independent-sample t-tests and did not find a significant difference between the two groups, $t(42) = 1.53$, $p = 0.13$, for children, and $t(42) = 0.73$, $p = 0.47$, for mothers. This result showed that children and mothers in both groups produced same amount of utterances during the doll play.

Proxy talk and doll directed speech We first counted the number of instances of proxy talk and doll directed speech that were produced by children and mothers during the session. Then we divided them by the total number of utterances for each participant to calculate the proportion. The proportions of proxy talk were 0.00 (0.01) in children in the low group, 0.02 (0.02) in children in the high group, 0.13 (0.11) in mothers in the low group, and 0.37 (0.19) in mothers in the high group. To see whether there was a difference in the proportion of proxy talk between the low and high groups, after arcsine transformation of the proportion data, independent-sample t-tests were conducted for children and mothers. A significant difference was found in children, $t(42) = 3.39$, $p < .01$, $d = 1.02$, and in mothers, $t(42) = 5.00$, $p < 0.01$, $d = 1.51$. This result indicated that mothers and children in the high group produced proxy talk more frequently than those in the low group. The proportions of doll directed speech were 0.06 (0.09) in children in the low group, 0.23 (0.17) in children in the high group, 0.00 (0.00) in mothers in the low group, and 0.00 (0.01) in mothers in the high group. Independent-sample t-tests were conducted, and a significant difference was found only in children, $t(42) = 3.39$, $p < .01$, $d = 1.59$. This result indicated that children in the high group produced doll directed speech more frequently than children in the low group.

Correlation between proxy talk and doll directed speech

To see whether there was a relationship between mother's and children's statements, Pearson's correlation coefficient was calculated between the proportion of proxy talk and of doll directed speech made by mothers and children for each group. The result revealed that in the low group, there are no correlations between them, but in the high group, there is a significant correlation between mother's proxy talk and children's, $r = .69$, $p < .001$ (two-tailed). This indicated that when mothers produce proxy talk, their children produce doll directed speech during their play in the high group.

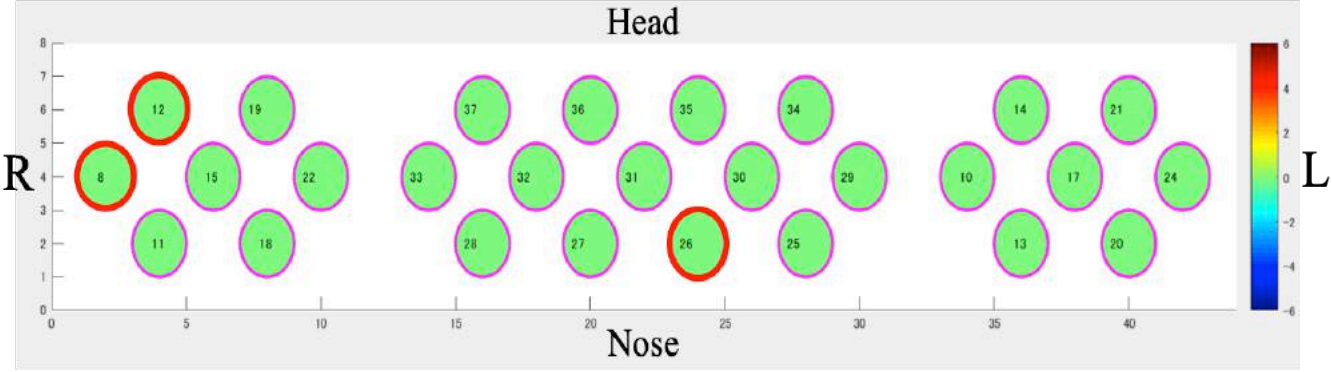


Figure 5. Arrangement of the near-infrared spectroscopy (NIRS) channels. The channels with a red circle indicate the channels showing the significant difference between the low and the high groups.

Proportion of speech type in children and mothers Proportions of each speech type were calculated by dividing the number of each speech type by the total number of utterances for children (Figure 3) and mothers (Figure 4). After arcsine transformation of the proportion data, paired t-tests were conducted for each speech type between two groups in children and mothers. A significant difference was found only in Emotion labels in adults, $t(42) = 2.43$, $p < .05$, $d = 0.73$. This indicates that mothers in the high group produced Emotion labels more frequently than those in low group.

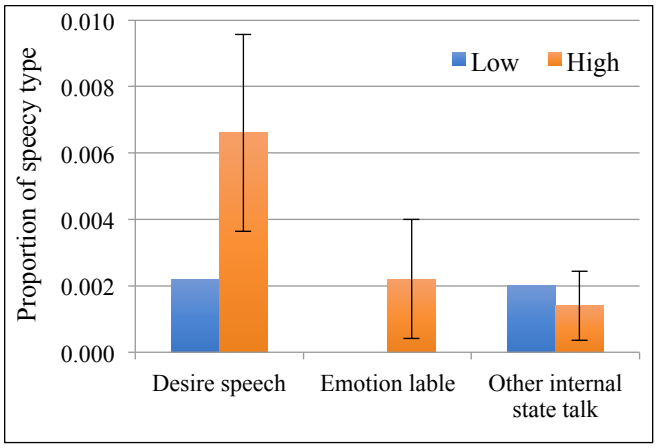


Figure 3. Proportion of each speech type in children

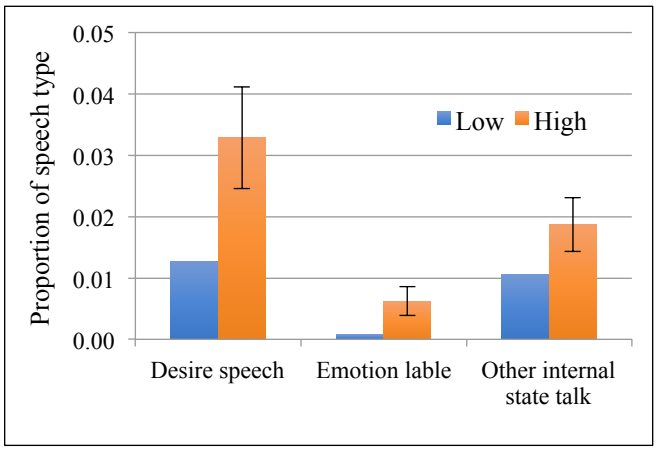


Figure 4. Proportion of each speech type in adults

2. 1. fNIRS analysis

We conducted paired t-tests on the difference in oxy-Hb change during the ending phase between two groups (the high vs. low group) for each channel, and for each condition (hindering and helping condition)

The result showed that there were no significant differences between two groups in the hindering condition, but the oxy-Hb concentration in the helping condition was significantly more increased in the low group than the high group for the measurement channel 8, $t(27) = 3.05$, $p < .01$,

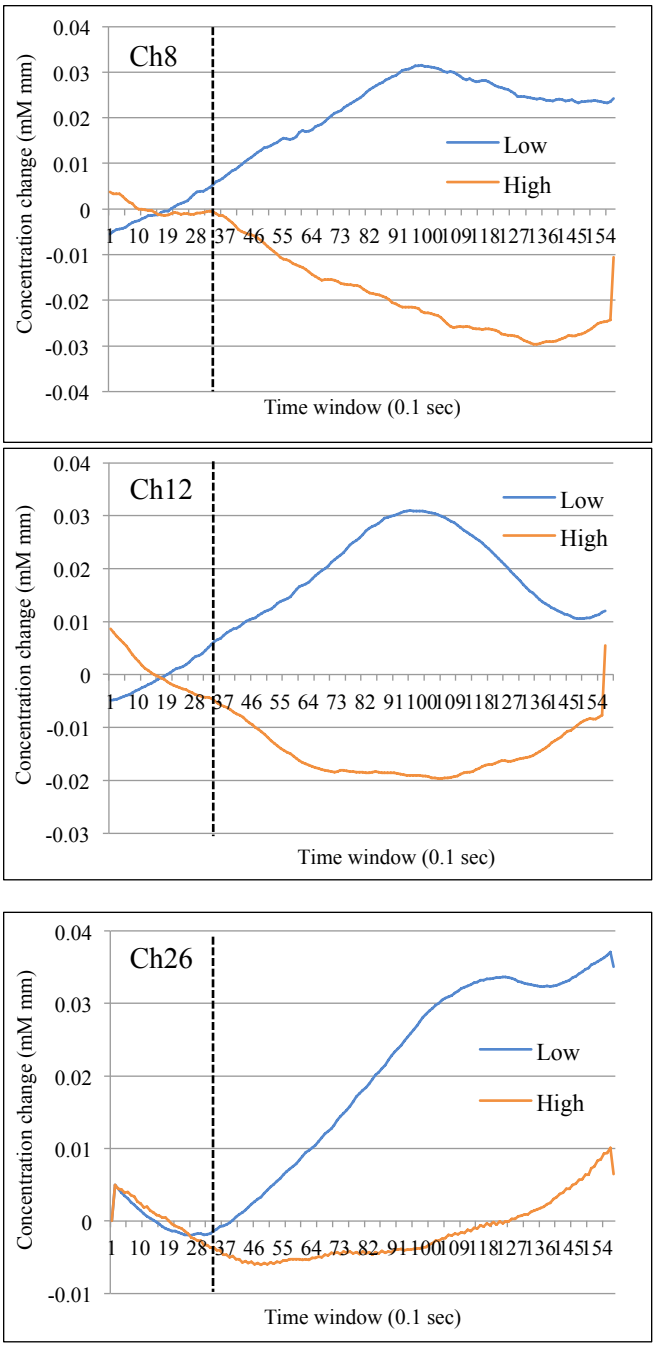


Figure 6. The time course of signal changes in oxy-Hb at TPJ (Ch8 and 12) and mPFC (Ch26) channels during the ending phase. The x-axis represents time units of 0.1 second. The dot line indicates the onset of the ending phase. We analysed the last 3 seconds of the ending phase, which is time windows from 85 to 115 on x-axis.

$d = 1.12$, the channel 12, $t(29)=3.58$, $p < .001$, $d = 1.29$, the channel 26, $t(30)= 3.28$, $p < .01$, $d = 1.16$. The channel 8 and 12 cover TPJ (temporoparietal junction) region, and the channel 26 covers mPFC (medial prefrontal cortex) region.

Taking TPJ's and mPFC's functional roles into account, these results suggested that children in the low group are more sensitive to other's helping behaviors and feelings than those in the high group. We conducted the correlation analysis between behavioral data (child/mother speech) and the three channels showing the significant differences between groups. However, any statically significant correlations between them were not found.

Discussion

The current study examined the relationship between doll-play experience and development of mentalization in children aged 2 to 3 by observing mother-child interaction and measuring brain activation by fNIRS. We found three main results. The first finding is that mother's talk differs depending on children's experience in playing with dolls. Mothers who have children with more experience in doll-play tended to produce more proxy talk and emotion labels during doll-play than mothers in the low doll-play experience group. The second finding is that children who have more experience in doll-play produced doll directed speech more frequently than children who have less experience in doll-play.

These findings indicate that mentioning a doll's internal feelings or talking to children by using the doll's voice direct children's attention to a doll's inner psychological states, which may lead children to the development of mentalization. In turn, children tend to talk to the doll as if the doll is an animate entity by using doll directed speech. This interpretation is consistent with previous research showing that mother's inputs are important to develop children's social understanding (e.g., Brownell et al., 2013; Lillard. 2017; Nakamachi, 2015). For example, Nakamachi (2015) found that mothers' pretend behaviors when toddlers were at 18 months predicted toddlers' understanding of a stranger's pretense 6 months later. Our data added new insight to this line of research. That is, as the result of correlation analysis shows, mother's proxy talk was significantly correlated to children's doll directed speech in the high doll-play experience group. Although it is difficult to determine the cause-effect relationship from our data set, we can speculate that mother's proxy talk makes children aware of a doll's inner feelings. In turn, children address their talk to the doll. This caregiver-child interaction through a doll may lead children to facilitate the development of mentalization.

The third finding is that in the measurement channel above TPJ (temporoparietal junction) and mPFC (medial prefrontal cortex) regions, the oxy-Hb concentration in the helping condition was more increased in the low group than the high group. This finding tells us that children in the low group are more sensitive to other's helping behaviors and feelings than the high group. It can be interpreted that children in the high group have seen a variety of helping scenes in doll-play. In contrast, children in the low group may not be as familiar. As the helping condition requires children in the low group to mentalize others feelings, the oxy-Hb concentration was more increased in TPJ and mPFC than children in the high group.

Contrary to our expectation, we did not find any difference in the oxy-Hb concentration in the hindering condition. This may be because it is too hard for children aged 2 to 3 to understand the situation in video clips as a hindering situation. Also, unlike the helping condition, the stories of the hindering condition do not have clear ending in the sense that the issue of the hindered person still remains. Thus, different time windows to measures brain activity may need for hindering and helping conditions respectively.

In conclusion, the current study showed that the experience in playing with a doll is related to the development of mentalization, and that maternal inputs toward her child and child's response toward a doll play important roles in the development of mentalization.

There are three directions of future studies. First, the current study used only audiovisual stimuli presenting helping/hindering behaviors to see the relationship between mentalization and doll play. But role-play is comprised of different elements such as verbal and nonverbal interaction, theory of mind, mentalization, sharing and reading intention, and object manipulation (Lillard, 2017). Thus, as a future task, it is important to examine whether and how doll play affects other social, linguistic, or cognitive process by using other stimuli that are sensitive to those domains. The second future task would be conducting longitudinal studies. Sachet and Mottweiler (2013) pointed out that it has remained unclear whether engaging in role-play enhances children's social understanding or the other way around. To make the cause-effect relationship between doll-play and social understanding clear, we need to longitudinally examine the developmental path of social understanding including mentalization and how caregiver's input and other environmental resources affect children's behaviors. Finally, given that play is a culturally constructed activity (Gaskins, 2013), it is important to examine whether findings in the current study holds true for populations with other demographics (e.g., different social-economic status, people with non-Japanese backgrounds, mother-son or father-child dyads).

Acknowledgments

This work was supported by contract research grant from INFER Co., Ltd. and Japan Science and Technology Agency CREST (JP- MJCR14E2).

References

Brownell, C. A., Svetlova, M., Anderson, R., Nichols, S.R., & Drummond, J. (2013). Socialization of early prosocial behavior: Parents' talk about emotions is associated with sharing and helping in toddlers. *Infancy*, 18(1), 91-119.

Carlson, S. M., White, R. E., & Davis-Unger, A. (2014). Evidence for a relation between executive function and pretense representation in preschool children. *Cognitive Development,* 29, 1-16.

Fonagy, P., Gergely, G., & Target, M. (2007). The parent-infant dyad and the construction of the subjective self. *Journal of Child Psychology and Psychiatry*, 48, 288-328.

Fonagy, P., & Target, M. (1996). Playing with reality I. The *International journal of psycho-analysis*, 77(2), 217.

Frith, C. D. & Frith, U. (2006). The neural basis of mentalizing. *Neuron*, 50, 531-534.

Gaskins, S. (2013). Pretend play as culturally constructed activity. In M. Taylor (Ed.), *The Oxford handbook of the development of imagination* (pp. 223–247).

German, T.P., Niehaus, J.L., Roarty, M.P., Giesbrech, B., & Miller, M.B. (2004). Neural correlates of detecting pretense: automatic engagement of the intentional stance under covert conditions. *Journal of Cognitive Neuroscience*, 16, 1805-1817.

Harris, P. (2000). *The work of the imagination*. Oxford, UK: Blackwell.

Lausberg, H., & Sloetjes, H. (2009). Coding gestural behavior with the NEUROGES-ELAN system. Behavior Research Methods, 4(3), 841–849.

Lillard, A. S. (2013). Fictional World, the Neuroscience of the Imagination, and Childhood Education. In M. Taylor (Ed.) in *The Oxford Handbook of the Development of Imagination,*. Oxford University Press (pp. 137-160).

Lillard, A. S. (2017). Why Do the Children (Pretend) Play? *Trends in Cognitive Sciences,* 21(11), 826–834.

Lindsey, E., & Colwell, M. (2013). Pretend and phyiscal play: links to preschoolers' affective social competence. *Merrill-Palmer Quarterly*, 59(3), 330-360.

Lillard, A. S., & Kavanaugh, R. D. (2014). The Contribution of Symbolic Skills to the Development of an Explicit Theory of Mind. *Child Development*, 85(4), 1535–1551.

Lillard, A. S., Lerner, M. D., Hopkins, E. J., Dore, R. A., Smith, E. D., & Palmquist, C. M. (2013). The impact of pretend play on children's development: A review of the evidence. *Psychological Bulletin*, 139(1), 1–34.

Lloyd-Fox, S., Blasi, A., & Elwell, C. E. (2010). Illuminating the developing brain: The past, present and future of functional near infrared spectroscopy. *Neuroscience and Biobehavioral Reviews,* 34, 269-284.

Mahy, C. E. V., Moses, L. J., & Pfeifer, J. H. (2014). How and where: Theory-of-mind in brain. *Developmental Cognitive Neuroscience*, 9, 68-81.

Minagawa, Y., Xu, M., & Morimoto, S. (2018). Toward interactive social neuroscience: Neuroimaging real-world interactions in various populations. Japanese Psychological Research, 60(4), 196-224.

Mitchell, J.P., Macrae, C.N., Banaji, M.R., (2006). Dissociable medial prefrontal contributions to judgments of similar and dissimilar others. *Neuron,* 50, 1–9.

Nagamitsu, S., Yamashita, Y., Tanaka, H., & Matsuishi, T. (2010). Functional near-infrared spectroscopy studies in children. BioPsychoSocial Medicine, 6(7), 1-7.

Nakamichi, N. (2015). Maternal behavior modifications during pretense and their long-term effects on toddlers' understanding of pretense. *Journal of Cognitive Developmebt*. 16, 541-558

Orr, E. & Geva, R. (2015). Symbolic play and language development. *Infant Behavior and Development*, 38, 147-161.

Sachet, A. B., & Mottweiler, C. M. (2013). The Distinction Between Role-Play and Object Substitution in Pretend Play in *The Oxford Handbook of the Development of Imagination,* M. Taylor (Ed.). Oxford University Press.

Strangman, G., Franchescini, M.A., Boas, D., (2003). Factors affecting the accuracy of near-infrared spectroscopy concentration calculations for focal changes in oxygenation parameters. *NeuroImage,* 18, 865-879.

Whiteheat, C., Marchant, J.L., Craik, D., & Frith, C.D. (2009). Neural correlates of observing pretend play in which one object is represented as another. *Social Cognitive and Affective Neuroscience*, 4, 369-378.

Wolf, D. P., Rygh, J., & Altshuler, J. (1984). Agency and experience: Actions and states in play narratives. In I. Bretherton (Ed.), *Symbolic play: The development of social understanding* (pp.195-217). Orlando, FL: Academic Press.

Wyle, A. (2014). Mentalization and theory of mind. *Prax Kinderpsychol Kinderpsychiatr,* 63(9), 730-737.

Introducing quantitative cognitive analysis: ubiquitous reproduction, cognitive diversity and creativity

Cameron Shackell (cw.shackell@qut.edu.au)
Queensland University of Technology, Brisbane, Australia

Peter Bruza (p.bruza@qut.edu.au)
Queensland University of Technology, Brisbane, Australia

Abstract

The rise of ubiquitous computing has cemented ubiquitous reproduction (UR) as a defining feature of contemporary human environments. UR is most obvious on our televisions and smartphones but has homogenised most material aspects of our lives. Emerging technologies such as 3D printing and robotics will ensure that this trend intensifies. UR is an issue of global scale that is relatively intractable to qualitative treatment. This paper introduces a novel *quantitative* approach to cognitive science and to analysis of UR. The approach uses the finiteness of cognition to establish a minimal ontology with which to model cognitive diversity under UR. It demonstrates that, despite widespread valorisation of diversity, cognitive diversity must be declining at a global level. The implications of this for creativity are that the arc for creative impact is growing shorter as the need to be immediately intelligible promotes the formulaic at the expense of the interpretable.

Keywords: ubiquitous computing; ubiquitous reproduction; cognitive diversity; creativity; intelligibility

Introduction

Ubiquitous reproduction (UR), a feature of contemporary society accelerating under ubiquitous computing, has brought an unprecedented rise in the homogeneity of human environments. Our attention is increasingly occupied by images and sounds reproduced synchronously and asynchronously in millions of widely dispersed locations – on mega-screens that tower over us in cityscapes (as in Figure 1); on televisions and monitors in our homes; and on smart devices in our pockets and on our wrists. Within the cocoon of our digital habits we are now as likely to be glued to our favourite online resources and entertainments walking through a Bangkok market as a Finnish airport.

The material effects of UR are far from straightforward or short term. The digital reproduction of images and sounds has provided the scaffold for broader standardization of our physical world. Human environments are now measured, planned, designed, manufactured, distributed, and assessed with digital assistance. Our experience, derived from objects on computer screens in environments of computational origin, is becoming more and more homogenous. Everything from our first appearance *in utero* on ultrasound screens, to the digital curricula of the schools we attend, to the 3D printed artefacts we use, to the temperature and humidity of the air we breathe is melding into a common background.

Figure 1. Ubiquitous reproduction (UR) is occurring in many forms and on many scales in human environments.

The identical representations information technology makes possible are certainly a boon for productivity. They are also often touted as a godsend for creativity. Indeed, the ease with which digital amateurs can create and disseminate images and sounds has progressed to the level of "deep fakes" that threaten to undermine, as Baudrillard (1981/1994) foresaw, trust in reality itself. Thanks to UR there have appeared many new and popular activities to stock the digital repository. Entire new genres such as emoji, gifs and memes have emerged, and as technology progresses no doubt these will be joined by other technologically defined creative – if similarly pastiche-based – categories.

Alloyed to claims that UR promotes creativity is the promise (often promoted in marketing of new technology) that UR is a breakthrough for cognitive diversity. The argument is that UR allows us to learn about (or even virtually experience) other viewpoints, expand our cognitive degrees of freedom, and so overcome the ignorance that engenders bigotry. This meshes well with the conjointly valorised idea that creativity necessarily involves an increase in cognitive diversity. Afterall, if nothing new and hence expansive of diversity appears, how can creativity have taken place? In terms of cognition under UR, however, a worthy question is whether rising environmental homogeneity carries broader, unrecognized structural contractions actually inimical to *global* cognitive diversity. This paper introduces a quantitative view, based simply on the finiteness of cognition, that localized and anecdotal creative benefits of UR disguise a broader pattern of

reduced cognitive diversity and an inflection point in what is possible in, or even meant by, creativity.

Quantitative cognitive analysis

A common difficulty in cognitive science, and one perhaps retarding analysis of UR at present, is achieving ontological agreement. Cognitive science often ignores thorny philosophical dilemmas to concern itself almost entirely with *qualitative* aspects of cognition, which are typically treated as self-evident. It is, however, specious to claim that we know or can infer what a particular individual or group of individuals is thinking, or that any symbolic representation of cognition is meaningful outside symbolic systems, which rely themselves, after all, upon cognition. Well over sixty years ago, Quine (1951), as part of his critique of "modern empiricism", pointed out the circularity of assuming cognitive synonymy. Yet such assumptions continue to underpin psychology and cognitive science and are rarely challenged.

Historically, however, qualitative enquiry is not the sole – or even foundational – charter of either psychology or cognitive science. James (1890/2012, p. 9), for example, defined psychology as "the science of *finite* individual minds" [emphasis added]. This underexplored distinction has been examined recently by Shackell (2018, 2019, in press) in an attempt to bring clarity to information age challenges in semiotics. The bootstrapping move is to first treat questions of cognition *quantitatively* in order to derive a minimally committed and hence maximally surefooted ontology of cognition. For the present analysis, this "quantitative cognitive science" approach can be summarized in three axioms that can be confirmed from common experience:

1. Cognition is finite (i.e., we do not know everything; what we never think we never know)
2. Cognition can be similar or at least closely related (e.g., communication is possible)
3. Over time, common environments produce similarly structured cognition and behaviour in a population (e.g., many people in Paris speak French)

Most crucially, from the first of these axioms the construct of a global human cognitive field can be derived, which is simply a space-time concept of cognition occurring at a species level. This simple construct, shown in Figure 2, is the blank slate for quantitative analysis of cognition. Further explicit ontological commitments can be carefully introduced to examine various phenomena, among which the rise and role of UR is our present focus.

An attentional definition of environment

Environments that humans create and customise for themselves are more complicated than their appearance at any one moment in time suggests. Human environments are cognitive, defined just as much by habits of attention as possible targets of attention. Habits of attention, in turn, are shaped by perceptual processes over time – largely by what changes or modulates in an environment. Many people who live in sight of some remarkable wonder such as the Grand Canyon, for example, may nonetheless currently devote much of their time to Facebook or Twitter.

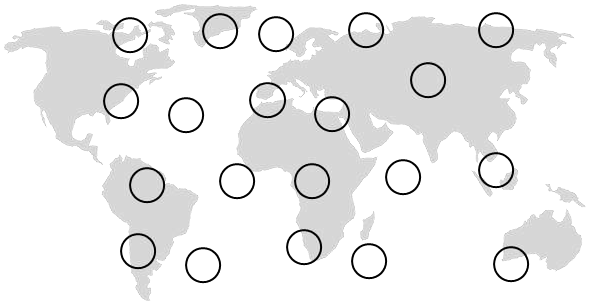


Figure 2. The cognitive field, a minimal construct facilitating careful *quantitative* analysis of cognition. Each circle represents a single agent's thought at a point in space and time. Adapted from Shackell (2018).

Our experience is a complex function of the attended and unattended stimuli we inhabit. UR brings determinative constraints to this function in ways that are difficult to perceive. An initial reaction to email in the 1990s, for example, may have been that it brings some incidental, standardising reproduction to our experience but ultimately is a source of extremely varied stimuli (e.g., words and images). However, if we examine the email of a large number of people today, will we find a rich and open-ended diversity of images, words, and most importantly, habits of interpretation? More likely we will find a quite clustered, reducible set of generic artefacts such as advertising, jokes, school news and so on. In fact, spam filters rely upon the very fact that email messages at a global level are not very diverse.

The issue of material homogeneity interacting with attentional proclivities to determine cognition is a complex one. For example, a hotel room containing a large television that is switched *off* has a very high level of material homogeneity relative to other hotel rooms but nonetheless allows many degrees of freedom for cognition. An agent sitting in that room may be thinking about a passing car, the colour of the carpet, a memory from childhood, rice salad recipes et cetera, each with a low degree of predictability. In contrast, the same hotel room with the large television switched *on* is less materially homogenous – the light and sound emitting from the television are dynamic and change the environment constantly. Cognitively, however, it is *less* diverse as a large proportion of people in such a situation will be on very similar trajectories promoted by the modulating television (e.g., experiencing an episode of *Seinfeld*). Such attentional and focus dynamics have a historical or formative component and are in a sense nested within one another (the homogenous, generic hotel room nests the seemingly diverse television output, which is itself derived from a restricted content set e.g., *Seinfeld* episodes). Cognition, therefore, can be homogenised by dynamic as

much as static stimuli according to prevailing attentional habits.

Pre-UR environments

Natural, pre-modern environments provide a baseline (or at least a point of comparison) for homogeneity. UR is limited in nature. No two mountains are the same. Nor any two rivers or places in them. Moreover, in nature there is little opportunity to view things habitually from the same perspective, or to view the same event repeatedly[1]. Importantly, there is a close linkage in pre-modern environments between stimulus and response. In pre-modern environments, if you saw a tiger you would think to run. Today you will likely just turn your head away and dismiss the vision as an advertisement for sneakers or a charity. Even in the nineteenth century, despite the rise of newspapers, museums and public libraries, it was relatively rare for large numbers of human beings to have encountered identical objects. Human cognition in the past likely exhibited a high level of idiosyncratic abstraction derived from variant stimuli. One person's concept of a mountain or a steam engine may have been much different than another's without ever causing economic or social friction. In other words, economic and social functions were performed despite quite a high degree of cognitive diversity.

Formalising cognitive diversity

To derive a formal model of cognitive diversity based on the axioms of finite cognition, we can begin by formalising the cognition of a population with *n* members over a chosen time period. Let *P* be the set of *n* contemporaneous thought sequences *s* in the population over the period:

$$P = \{s_1, s_2, s_3 \dots s_n\}$$

C can be defined as the set of distinct thoughts of *s*, a member of *P*, over the period.

The total cognitive diversity of *P* can therefore be expressed as the union of all *C*:

$$\bigcup_{1}^{n} C$$

Conversely, the total cognitive commonality of *P* can be written as the intersection of all *C*:

$$\bigcap_{1}^{n} C$$

A more meaningful measure of cognitive diversity for *P*, however, must include an awareness of the distribution of similar and different cognitive states among agents. A number of metrics are available in statistics for comparing set similarity. The Jaccard index, for example, is the size of the intersection of two sets divided by the size of the union. For *P*, a useful metric of similarity is the mean of the pairwise Jaccard indices of all *C*, which we can call that population's cognitive similarity *z*:

$$z = \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{s_i \cap s_j}{s_i \cup s_j}}{\left| \binom{n}{2} \right|}$$

A *z* value of 1 would indicate a complete lack of cognitive diversity while a value of 0 would indicate complete cognitive diversity. While *z* can perhaps never be directly measured except in some future dystopia, it does give us a formal tool with which to reason about certain situations in which diversity is at issue, and more broadly the effects of phenomena such as UR.

Modelling cognition in increasingly homogenous environments

It would be difficult to sustain the argument that human beings living in environments that are increasing similar will not tend to think in increasingly similar or at least related ways. Even if thoughts do not circulate in an epidemiological manner, disparate reactions to common artefacts must lead to thoughts falling into finite patterns and hence concentrating within generic categories. For example, while everyone may not have positive emotions around the massively reproduced images of the last FIFA World Cup, large numbers of people will have some *species* of reaction such as disappointment, outrage, respect, indifference et cetera. Moreover, these reactions will be patterned in very broad ways, with people in the winning country, France, more likely to exhibit one of the more positive cognitive states. Such "made for television" events are often lauded, and indeed sought after, for bringing the world together and creating *connection*.

A connected world or a homogenous world?

In analyses of technological change, a common assertion is that cognition is changing because individuals are now highly "connected". This idea of connection, however, despite having the appearance of explanatory power and finality, bears some deconstruction. If we probe a little deeper into the material nature of increased connectivity, we find that UR is the enabling mechanism. Connection is possible because a message on a device at one location can be reproduced at another. More subtly still, reproduction of any image at different locations creates a connection *potential* between individuals by synchronising their experience to some extent – that is, by honing their ability to receive a related image later.

[1] This point is illustrated by the debate that raged for millennia as to whether, and at what point, all four of a horse's hooves leave the ground while galloping. The debate was resolved by Muybridge's 1878 *Sallie Gardner at a Gallop* photographs.

When we draw connections as an edge on a social network graph (as is routinely done – see, for example, Baronchelli, Ferrer-i-Cancho, Pastor-Satorras, Chater, and Christiansen, 2013), we are abstracting a very complex structure of UR into a simple metaphor. In specific *qualitative* analysis this reduction is often not afforded enough scrutiny: the complexity of a single connection is enormous and drags with it an implicit micro-mechanics that has never really been made clear. In *quantitative* analysis, however, such edges can be given a very precise meaning at the systemic level as environmental commonalities occasioning synchronisation of thought. Edges with such a meaning can be assigned probabilities based on environmental commonalities and attentional factors (as indeed marketing and advertising already do in some situations e.g., Allenby and Rossi, 1998).

Diversity in the cognitive field under rising UR

Using the construct of the cognitive field and edges introduced above we can model the effects of UR by assigning discrete *values of difference* to cognition – that is, by marking cognitive states as different or similar without claiming to know anything qualitative about them. In Figure 3, the colour of circles in the field indicates the difference or similarity of cognition. The edges are indicators of common experience produced by UR. As per the discussion above, the edges do not necessarily "spread" a cognitive state but rather increase the tendency of other agents to assume some complementary state.

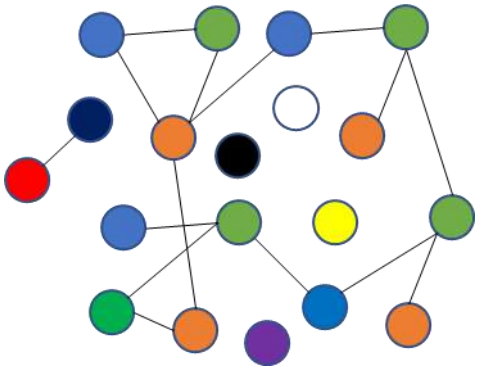


Figure 3. A cognitive field under low UR. Different colours represent different cognitive states. Edges represent common experiences facilitated by UR.

Figure 4 depicts a cognitive field under greater UR which facilitates more common cognition and hence less cognitive diversity. Notice that the "connections" between agents (the products of UR) are greater and hence the number of distinct states is lower than in Figure 3. The result is a move towards what is known as a "small world" graph.

Figures 3 and 4 show that if we stipulate that increasingly homogenous stimuli tend to produce less diverse cognition, then under global UR we can assume a falling cognitive diversity in human populations. The next question we may ask is why such a movement is underway?

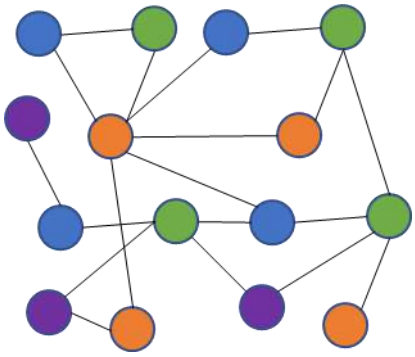


Figure 4. A cognitive field in an environment with a high level of UR. Greater "connection" leads to lower diversity.

The teleology of falling cognitive diversity

Falling cognitive diversity has one obvious cause: economics. As new methods for production and distribution of material goods and information evolve, these are quickly disseminated and adopted around the world. For example, producers will not continue to smelt iron in an inefficient and idiosyncratic way if a better method is obviously available. The adoption of the new smelting method, however, will require remote peoples to synchronise some of their cognition with others already using the method. This will also likely bring larger flow-on material changes: altered city locations and landscapes with smokestacks of a certain shape, new jobs with similar duty descriptions, and, most broadly, societal changes to do with increased availability of iron. The pace of economic change in regard to information technology is many times greater. Such changes are economically optimal but operate by decreasing cognitive diversity globally.

A counterintuitive view of diversity

With the current valorisation of diversity in race, gender, politics and religion, the conversation around cognitive diversity – which if one believes in the mediation of reality by cognition is the root of *all* diversity – turns quite counterintuitive. To anyone with a positive notion of truth or a commitment to democratic philosophy as espoused, for example, by Dewey (1916/2012), diversity is a fundamental value and a cornerstone of contemporary UR-dominated society. In fact, however, in terms of the analysis of diversity presented above, the rise of discourses about diversity – as for all *global* discourses – must be viewed as a symptom of *decreasing* cognitive diversity. Whereas a broad and idiosyncratic range of cognition around diversity once obtained, UR has brought global templates for cognition to every agent. Recently, for example, the #MeToo movement set the issue of gender into a certain polarised structure around the world. Whether this one phenomenon has reduced or increased cognitive diversity is

a moot, *qualitative* point. Taken as a whole, however, discourses that dominate globally via UR must have the *systemic* effect of reducing cognitive diversity overall.

Although it will not be traced here, it would seem possible to reconcile the teleology of declining cognitive diversity with the rise in global discourses *about* diversity. Put simply, the drive to economic and social optimality gives rise to discourses that efficiently allocate – or at least *consume* – cognition in support of it.

Intelligibility

The counterintuitive result above is that while technology is providing information in seeming abundance via UR, this leads to increased homogeneity of environments which must lead to a decrease in cognitive diversity. This decrease is disguised by a perceived increase in intelligibility whereby we expect, and have patience only for, stimuli that fit immediately into our cognition. The rise of the "random" as a term in popular culture might be regarded as a symptom of this increase in intelligibility: what is not immediately recognizable and intelligible is pushed from cognition as "random" as the mind seeks to navigate only those states on the homogenous, highly connected network. The term "stranger", for example, has largely been replaced by "random guy/girl" in many idiolects. In such a context we must re-examine what creativity – once capable of generational, revolutionary effect – now means.

Schematisation

In work that is easily related to the rising trend to fast intelligibility, Stiegler (1994/1998) has criticised the role of technology in "schematicising" cognition – that is, patterning thought into generalisable routines, or, as Quine, his forerunner, defined it: "positing sharp boundaries where none can be drawn" (Quine, 1990, p.12). Schematisation leads to shortcuts in thought for activities as diverse as recognizing villains in a movie by their smoking habits; interacting with checkout staff in ways learned from vending machines; or calling one's memory one's "hard drive".

Via schematisation, UR impacts social relations by the spread of common experience. To navigate the common environment, one must learn and acquiesce to routines of thought and action or be nudged[1] into line with other members by shared norms. The paradox of diversity applies: if one wishes to increase diversity in UR contexts one must commit acts of rebellion which will only be noticed if they in fact fit current schemas. As a rebel against UR one is in danger of creating a rebellious movement that can only thrive on the commonality supplied by technology, which under UR will quickly normalize it.

A relevant metaphor for cognition in homogenous environments is the (integrally related, embodying) adaptive development of our bodies. When living in natural environments full of uneven and undulating surfaces, we can attain almost any position. We will of course begin to wear pathways, but these evolve with our activities and are not fully determinative. Our physiology adapts so that our feet retain degrees of freedom, develop callous from certain movements over rocks or sand, and our awareness of terrain is of a certain fluid kind. Consider, in contrast, a human built environment in which surfaces are generally flat and even and any obstacles are essentially vertical (such as the side of a building or house). Our movement becomes limited in absolute ways. If we wish to go to a certain place there are hard restrictions on what paths are possible. Our feet will adapt to walk on flat surfaces; our awareness of terrain will be of a more binary kind; and we will inhabit an area having experienced only a small fraction of its terrain or viewpoints (not many of us have seen inside all the houses or apartments within 100 metres of our own). The effects of UR on cognition are of a similar, schematicising kind, which has profound ramifications for creativity.

Creativity under declining cognitive diversity

Under the axioms of finite cognition introduced at the start of this paper, a very straightforward quantitative definition of creativity is possible. Creativity is the mechanism by which one agent induces, or more romantically *inspires*, new cognitive states in another agent. UR under this definition has obviously increased the creative potential of each individual enormously. Each agent has the means to offer images and sounds via telecommunications to billions of others and to create new cognitive possibilities for them. This is in stark contrast to the world prior to UR, when it was difficult to affect large numbers of people even over long periods. It may, for example, have taken centuries for any significant number of people to have heard of, or formed a view about, an enormous public creative work such as Chartres cathedral.

We must pause to reflect, however, that, despite the new possibilities technology introduces, the total amount of cognition remains relatively constant. There are therefore two types of creativity that fit our quantitative definition. These roughly parallel Boden's (1990) psychological or "p-creativity", and historical or "h-creativity" but are worth reframing for the quantitative approach. Firstly, the new cognition need not necessarily be new in a global sense – only to one or more individuals. Creativity, therefore, involves, most minimally, a local increase in cognitive diversity that does not increase overall diversity. We might call this "zero gain" creativity and note that it tends to increase the *z* metric proposed above (makes cognition more similar). At the other extreme, creativity may involve provoking a completely new cognition never before attained by any agent (for example, Archimedes' eureka moment). We can call this "global increase" creativity and note that – at least initially – it tends to reduce *z* (makes cognition less similar). UR, by this distinction, overwhelmingly provokes a disproportionate amount of zero gain creativity.

[1] It is perhaps no coincidence that, in recent years, governments have formally embraced the notion of using UR to shape behaviour using techniques such as "nudging" (Thaler, 2009).

Diffusion and interpretation

A surfeit of zero gain creativity under UR has reduced the half-life of creative activity to very low levels. Anything new is quickly disseminated throughout the cognitive field. Novel thought is under pressure from (and likely to be displaced by) the next low gain stimulus. Moreover, the rapidity of dissemination discourages prolonged or novel interpretation of reproductions. Interpretation must be relatively shallow: the stimulus, as noted, must be immediately intelligible or will be simply ignored by the majority of receivers. It would seem absurd in the current context to spend years in careful interpretation of a single work of art to achieve something novel, but such activity was common and valorised in centuries past (as the long traditions of exegesis and hermeneutics attest).

In terms of the quantitative definition of creativity, under UR there is much creativity. Afterall, UR provokes new cognitive states in unprecedented numbers of individuals. The ubiquity and speed of that creative reception, however, is not growing the broader diversity of cognition as rapidly as the pre-UR age, which *ipso facto* lacked the apparatus of cognitive synchronisation.

Creativity, bending with the decrease in cognitive diversity, is becoming a short rather than long term possibility. Creative activity in homogenous environments is under pressure to be continuous and schematic or risk exclusion as "random". This leads to the increasing dominance of formulaic creativity. In Figure 5, for example, a piece of graffiti attributed to the artist Banksy combines simple images and colours. The placement of the graffiti in a drab urban context (not shown) draws viewers to its intelligibility and achieves – almost formulaically – a flash of creativity while also providing a ready-made meme for UR.

Should we limit environmental homogeneity?

Future historians may refer to our era not as The Information Age but as The Great Synchronisation. There exists a danger that the growing ubiquity of human interaction with technology and the homogenizing reproduction it enables may lead to restricted "closed" paradigms that we are not in control of – paradigms that are instead defined by the affordances and economics of the technology itself. The end result may be a counterintuitive and potentially pernicious reduction in cognitive diversity occasioning a new sterile aesthetics – an air-conditioned Dark Age in which there are no wrong clocks. Creativity expansive of human thought ("global gain") is at risk from creativity that is merely distributive ("zero gain"). We must beware that what is not instantly intelligible is not denied a place in the panoply of human cognition. A possible remedy that warrants further formalisation and research is the measurement and control of environmental homogeneity – something which must become a recognised parameter of our tolerance for ubiquitous computing.

Figure 5. Graffiti art attributed to Banksy known as *Girl with Balloon*. The image can be seen as an example of the trend to intelligible, formulaic, meme-friendly creativity.

Conclusion

This paper examined the paradox of cognitive diversity as a lauded societal value in the increasingly homogenous environments created by ubiquitous computing and the ubiquitous reproduction it allows. If we value diversity in any form, we must value cognitive diversity, for by definition all diverse reality springs from diverse cognition. The paradox is that ultimately our drive to communicate using reproductive digital means cannot be other than a force for reducing cognitive diversity to some optimally oscillating set.

The practical benefits of reduced cognitive diversity in the relation of humanity to its material needs – thriving in material terms with all resource exploitation and population itself optimized to maximum carrying rate – is unquestionable. Inefficient cognition leads to waste and error. We should question, however, whether we are ready to abandon some long-held commitments to our destiny as a species in order to embrace these benefits. For if creativity involves producing or inducing *new* types of cognition there can be only localized, short term creativity in a system that distributes stimuli and displaces existing diversity with superlative efficiency. We may in the process be condemning those who come after to lives of robotic absurdity, making them martyrs to our vainglorious and infinite conception of our very finite selves.

Acknowledgments

Figures 1, 2 and 5 contain images from pixabay.com used under their proprietary Pixabay license, which allows for free commercial and non-commercial use without attribution.

References

Allenby, G. M., & Rossi, P. E. (1998). Marketing models of consumer heterogeneity. *Journal of econometrics, 89*(1-2), 57-78.

Baudrillard, J. (1994). *Simulacra and simulation*. Trans. S. Glaser. Original work published 1981. Ann Arbor, MI: University of Michigan Press.

Baronchelli, A., Ferrer-i-Cancho, R., Pastor-Satorras, R., Chater, N., & Christiansen, M. H. (2013). Networks in Cognitive Science. *Trends in Cognitive Sciences, 17*(7), 348-360. doi:https://doi.org/10.1016/j.tics.2013.04.010

Boden, M. (1990). *The creative mind : myths and mechanisms*. London: Weidenfeld and Nicolson.

Dewey, J. (2012). *Democracy and Education*. Original work published 1916. Newburyport, MA: Dover Publications.

James, W. (2012). *The Principles of Psychology, Vol. 1*. Original work published 1890. Newburyport: Dover Publications.

Quine, W. V. O. (1951). Two dogmas of empiricism. *The philosophical review*, *60*(1), 20-43.

Quine, W. V. O. (1990). *Pursuit of truth*. Cambridge, MA: Harvard University Press.

Shackell, C. (2018). Finite cognition and finite semiosis: a new perspective on semiotics for the information age. *Semiotica*, 2018(222), 225-240 doi: https://doi.org/10.1515/sem-2018-0020

Shackell, C. (2019). Finite semiotics: recovery functions, semioformation and the hyperreal. *Semiotica*, 2019(227), 211–226. doi:10.1515/sem-2016-0153

Shackell, C. (in press). Finite semiotics: cognitive sets, semiotic vectors, and semiosic oscillation. *Semiotica*.

Stiegler, B. (1998). *Technics and time: The fault of Epimetheus* (Vol. 1). Trans. R. Bearsworth and G. Collins. Original work published 1994. Stanford, CA: Stanford University Press.

Thaler, R. (2009). *Nudge : improving decisions about health, wealth, and happiness*. New York: Penguin Books.

Symmetry: Low-level visual feature or abstract relation?

Ruxue Shao (ruxueshao@u.northwestern.edu)
Dedre Gentner (gentner@northwestern.edu)
Northwestern University
Department of Psychology, 2029 Sheridan Road
Evanston, IL 60208, USA

Abstract

We traced the development of sensitivity to symmetric relational patterns by creating a symmetry match-to-sample task. Children saw a symmetric standard made up of two shapes and choose between two novel alternatives: a symmetric pair and an asymmetric pair. We found that young children chose randomly between the two alternatives. Children were not reliably above chance until 8-to 9 years of age. In a second study, we found that young children could succeed in making symmetric relational matches if the triads were designed to invite informative comparisons. These findings show that relational insight of symmetry develops relatively late. However, as with other relations, comparison processes can promote sensitivity to the symmetry relation.

Keywords: symmetry; relational processing; comparison and contrast

Introduction

The acquisition and use of relational concepts are critical to higher-order cognition, and to learning in complex domains. Symmetry is arguably one of the most basic and ubiquitous relations in nature, evident in structures as small as molecules and as large as blue whales. Non-human animals are thought to show a preference for symmetrical over asymmetrical bodily features when choosing a mate, and there is evidence that humans rate symmetrical faces as more attractive than non-symmetrical ones (e.g., Grammer & Thornhill, 1994; Møller, & Thornhill, 1998). Based on these patterns, some researchers have suggested that sensitivity to symmetry may be biologically endowed (e.g. Grammer & Thornhill, 1994).

Evidence in favor of this claim comes from three lines of research. First, symmetry is easily processed by the human visual system (e.g., Wagemans, 1997). Researchers have suggested that symmetry detection is an automatic process that is rapid and robust to noise (Carmody, Nodine, & Locher, 1977; Royer, 1981). Symmetry processing is also thought to be a fundamental component of perceptual organization, playing a crucial role in object representation (e.g., Driver, Baylis, & Rafal, 1992; Marshall & Halligan, 1994).

Second, symmetry processing is widespread across species. Dolphins, pigeons, bamboo sharks, and bees are all capable of learning to discriminate between symmetric and asymmetric objects (Delius & Nowak, 1982, Giurfa, Eichmann, & Menzel, 1996, Schluessel et al., 2014, von Fersen et al., 1992).

A third point is that sensitivity to symmetry is early to emerge in human infants. Human children are sensitive to symmetry from infancy, although vertical symmetry is typically more readily perceived than horizontal symmetry. For example, using a habituation-dishabituation paradigm, Fisher, Ferdinandsen, and Bornstein (1981) found that 4-month-olds discriminated vertically symmetric single objects from those that were horizontally symmetric or asymmetric, but did not discriminate between horizontally symmetric and asymmetric objects. Other researchers have found converging results with older children (Bornstein and Stiles-Davis, 1984; Chipman & Mendelson, 1979).

The findings reported above have all focused on within-object symmetry. Taken together, they suggest that within-object symmetry may be a low-level visual feature that is universally detected. However, symmetry is not confined to single objects—many scientific discoveries emerge from detecting symmetrical patterns between objects or events (e.g., Gross, 1996). We want to raise the possibility that discriminating within-object symmetry is quite different from detecting symmetry between two or more distinct objects; the latter requires symmetry to be construed as a relation while the former does not. Although previous research on symmetry processing has revealed much on how humans and non-human animals perceive symmetry within a single object (see Cattaneo, 2017; Giannouli, 2013; Treder, 2010; Wagemans, 1997 for reviews), comparatively little is known about the development of the ability to recognize and match symmetry between objects. This paper aims to shed light on the development of children's insight of the between-object symmetry relation.

Is Symmetry the Basis for *Same/Different* Detection?

A secondary motivation for examining children's ability to detect and match symmetry relations is to explore how symmetry pertains to other fundamental relational concepts, such as *same* and *different*.

If between-object symmetry is fluently processed, as a low-level visual feature, even by very young children, it is possible that symmetry detection may inflate children's performance on *same*/*different* relational tasks. In an insightful analysis, Hochmann and colleagues (2017) discussed this possibility. They pointed out that in many *same*/*different* relational tasks, *same* pairs are also symmetrical (e.g., [O,O]), whereas *different* pairs are asymmetrical (e.g., [O,X]). Thus, participants could potentially pass such tasks by responding to symmetry.

Walker and Gopnik (2017) reported evidence that runs against this contention. Using a relational causal paradigm (the "Blicket Detector"), they found that 18-to 30-month-olds

could learn to discriminate between *different* pairs (e.g., [A, B]) and *same* pairs (e.g., [Ↄ, C]). Note that, as discussed above, the *different* pair is asymmetrical and the *same* pair is symmetrical, leaving open the question of whether the children were relying on symmetry rather than sameness. However, when the objects were fused together to form either a single symmetrical object (made from two identical objects) or a single asymmetrical object (made from two different objects), the toddlers failed to learn the discrimination. These findings suggest that within-object symmetry is not the basis for the children's performance on this *same/different* relational task. However, it does not address whether between-object symmetry detection influences *same*/*different* detection.

Can Children Detect Symmetry Between Objects?

One study that explicitly examined whether children can detect symmetry between objects was done by Kotovsky and Gentner (1996). They presented 4-, 6-, and 8-year-olds with a relational matching task in which children were given a standard composed of three figures and had to choose which of two alternatives was more like the standard. One of the alternatives matched the standard's relation and the other had the same objects in a nonmatching configuration (see Figure 1). Within each trial, the two alternatives included the same objects. Children were given a random mixture of four trial types that differed across dimension and polarity.

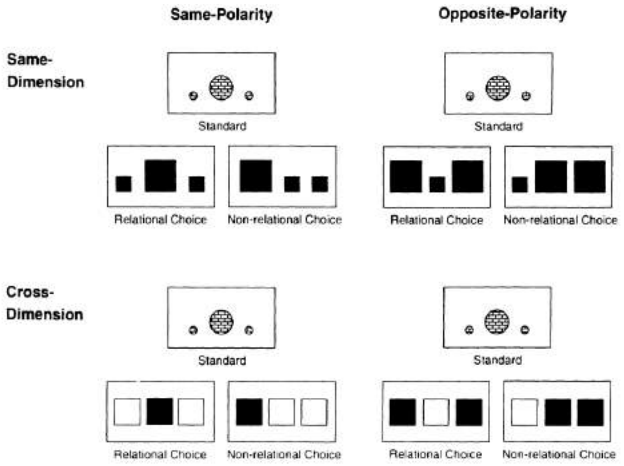


Figure 1: Schematics of stimuli used in Kotovsky and Gentner (1996).

The 6-and 8-year-olds performed well on this task. In contrast, 4-year-olds performed above chance only on trials where the correct alternative and the standard shared a concrete relation—same dimension and same polarity (Figure 1 top left panel). In these trials, the standard and the correct relational alternative share an overall shape (a low-high-low or an inverted V pattern), so it is not clear whether the 4-year-olds were attending to the relational pattern that defines symmetry or were instead simply responding to the common low-high-low shape.

The Kotovsky and Gentner (1996) study provides evidence that, at least by 6 years of age, children can perceive symmetry as a relation between objects, as well as make relational matches based on symmetry. Previous research has shown that children can make abstract *same/different* relational matches by 4 years of age without practice (e.g. Christie & Gentner, 2014), whereas the 4-year-olds in Kotovsky and Gentner's study were only able to make concrete symmetry matches. Further, we cannot confidently extrapolate from Kotovsky and Gentner's findings with 6- and 8-year-olds to the case of symmetric pairs like [Ↄ, C], because the figures in Kotovsky and Gentner's study (1996) all involved three objects. Although these are more complex than two-object figures, it could be that the larger patterns are easier to perceive.

In the current work, we trace the trajectory of children's ability to perceive and match symmetry in a task analogous to a classic *same/different* relational matching task in order to facilitate comparison of the developmental trajectories of these two relations. If we find that between-object symmetry matching is mastered earlier than same-different matching, this will leave open the possibility that same-different judgments could be drawing on symmetry perception.

Current Studies

The current work aims to (1) trace the development of human children's ability to detect and use the symmetry relation; and (2) investigate the learning processes by which children gain insight into the symmetry relation.

To do so, we created a Symmetry-Match-to-Sample task (SMTS) by analogy with the Relational-Match-to-Sample (RMTS) task (Christie & Gentner, 2014; Hochmann et al., 2017; Premack, 1983, Thompson, Oden, & Boysen, 1997). The RMTS task assesses understanding of *same* and *different* relations. For example, to assess the ability to match the *same* relation, the RMTS triad is AA (standard), BB & CD (alternatives). It is designed so that there is only one viable similarity match—the relational match based on the *same* relation. Analogously, in the SMTS task, children are shown a symmetric standard and asked to choose which is more similar: another symmetric pair, or an asymmetric pair. The standard and alternatives did not share any common objects, so there was only one viable choice (See Figure 2a).

Experiment 1

Methods

Participants One hundred 3-to 9-year-olds participated in this study: 19 3-year-olds (M = 42.8 months, SD = 2.3 months, 11 females), 21 4-year-olds (M = 53.6 months, SD = 3.4 months, 11 females), 20 5-year-olds (M = 68.4 months, SD = 1.6 months, 10 females), 20 6-year-olds (M = 80.4 months, SD = 1.8 months, 11 females), and 20 8- to 9-year-olds (M = 105.8 months, SD = 7.5 months, 9 females). An additional 11 children were tested but excluded from the final analysis, one child due to experimental error and ten children

due to failing to pass the catch trials described below (one 4-year-old and nine 3-year-olds). The racial and economic composition of the sample reflected those of the local population (majority European-American, middle- and upper-middle-class). All children were recruited from the greater Chicago area and received a small gift for their participation.

Materials and Procedure Children completed a Symmetry-Match-to-Sample (SMTS) task. The SMTS included eight test trials and three catch trials. Each trial was composed of a standard card and two alternative match cards (see Figure 2). The child was asked to choose the alternative that was most like the standard. In all test trials, the standard and correct match both depicted two identical shapes that were symmetric around the vertical axis; the incorrect match card showed two shapes that were in an asymmetric configuration (Figure 2a). Within a triad, each card was made up of unique shapes and colors.

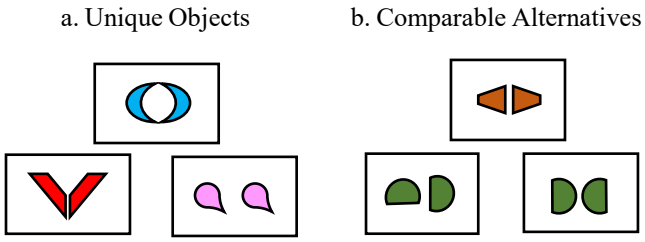


Figure 2: a. Sample test trial from Exp. 1: SMTS; b. Sample test trial from Exp. 2: SMTS with Comparable Alternatives.

After the test trials, there were three catch trials to determine whether the participants understood the task. These catch trials were literal similarity matches that did not require the child to judge relational similarity. For example, on one of the catch trials, children saw a red fish as the standard and had to choose between a blue fish (correct match) and a yellow cup. Children who failed any of the catch trials were not included in the analysis (n = 10).

Children were tested individually by an experimenter in a quiet room in the child's school or in a research laboratory. On each trial, the experimenter first presented the standard card and asked, "Do you see this one?" Then she placed the two alternative cards below the standard (as in Figure 1) and asked "Do you see these two? Which one of these two is more like this one?" Left/right placement of the alternatives was counterbalanced and no more than two subsequent trials had the correct match on the same side. Children were not given corrective feedback; only general encouragement (e.g., "You were so fast!", "Alright!") was provided.

Results

We measured the mean proportion of relational matches participants made in the eight test trials of the SMTS task. A one-way ANOVA revealed no difference in performance across the age groups, $F(4,95) = 1.16$, $p = .33$, $\eta 2 = 0.05$. When we compared the means of each age group to chance (50%), we found that only the 8- to 9-year-olds ($M = 0.69$, $SD = 0.27$) selected relational matches significantly more than chance, $t(19) = 3.17$, $p = .005$. The younger groups scored at chance (6-year-olds [$M = 0.57$, $SD = 0.29$]; 5-year-olds [$M = 0.59$, $SD = 0.28$]; 4-year-olds [$M = 0.61$, $SD = 0.26$]; 3-year- olds [$M = 0.53$, $SD = 0.15$]; all ps > .05]. Figure 3 shows the mean percentage of correct responses in each age group.

To assess whether learning occurred across trials despite the absence of feedback, we compared the proportion of correct matches that children made in the first three trials with that of the last three. There were no differences between the

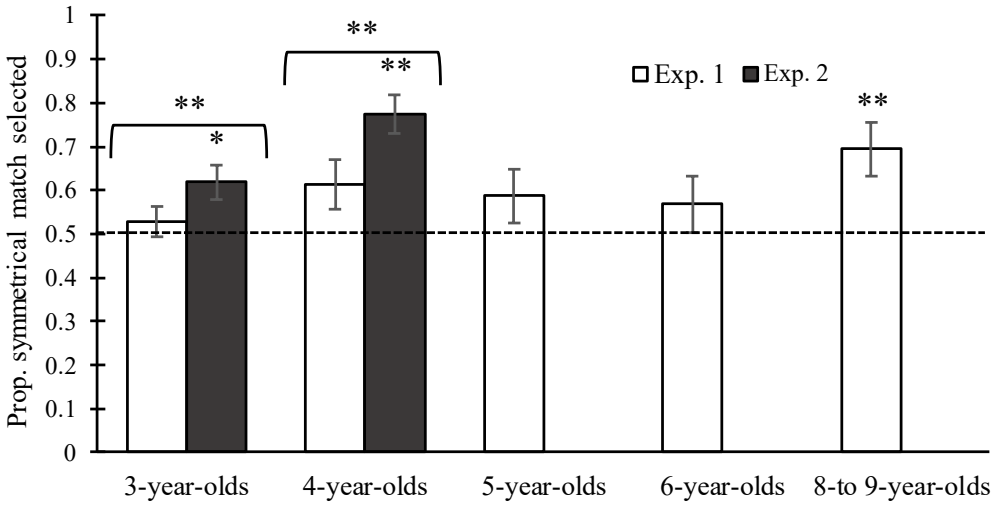


Figure 3: Mean proportion of symmetrical matches selected by children in Experiment 1: SMTS and Experiment 2: SMTS with Comparable Alternatives. Error bars depict standard error. * p < .05; ** p < .01

two in any age group, all *p*s > .13. Thus, performance did not improve across the eight test trials.

Discussion

The SMTS task was surprisingly challenging for children. Children who were six years of age or younger performed at chance rates. Even the 8- to 9-year-olds, who chose the relational match at significantly above chance rates, were only correct 69% of the time. It is unlikely that the younger age groups' poor performance was due to a failure to understand the task, since all participants were correct on the catch trials.

Why did children perform so poorly on the SMTS task compared to the findings in Kotovsky and Gentner? Kotovsky and Gentner (1996) found that 6-and 8-year-olds, and even 4-year-olds to a lesser extent, were able to make relational matches based on between-object symmetry. We propose that a crucial difference between our Experiment 1 and the Kotovsky and Gentner study was how much the experimental design scaffolded children's detection of the target relation.

Research has shown that an effective way to promote relational reasoning is by decreasing the salience of individual objects (Gentner & Rattermann, 1991; Goldstone & Son, 2005; Kaminski, Sloutsky, & Heckler, 2008). In similarity tasks, young children and novices tend to focus on objects rather than relations, and this can impede their relational processing (Gentner, 1988; Gentner & Toupin, 1986; Richland, Morrison, & Holyoak, 2006). In many studies, object salience has been reduced by using simple and uniform objects (Gentner & Rattermann, 1981; Mix, 2008). Kotovsky and Gentner (1996) further reduced object salience by presenting children with triads in which the two alternatives shared the same objects and differed only in the relation between the objects.

We hypothesize that using comparable alternatives promoted children's symmetry matching for two reasons: (1) using the same objects in both alternatives invites spontaneous comparison between them, and this may call attention to the key relational difference—that one is symmetric and the other is not; and (2) using the same objects in the two alternative pairs allows the child to discount object matches in making their choice and focus instead on any relations they may have perceived. In Experiment 2, we test this hypothesis by presenting children with a version of the SMTS that utilized comparable alternatives. We focused on the two younger age groups—the 3-and 4-year-olds.

Experiment 2: Comparable Alternatives

Methods

Participants Twenty 3-year-olds (M = 44.7 months, SD = 1.8 months, 9 females) and twenty 4-year-olds (M = 53.1 months, SD = 3.2 months, 11 females) were recruited for this experiment using the same methods as Experiment 1. Six additional children (four 3-year-olds) participated in the study but were excluded from analysis due to failing at least one of the catch trials.

Materials and Procedure As in Experiment 1, we created a relational matching task based on the symmetry relation. However, we modified the alternatives so that the two alternatives in a given trial consisted of the same objects, one in a symmetric configuration and the other in a non-symmetric configuration (see Figure 2b). The catch trials and procedure were as in Experiment 1.

Results

The mean proportions of relational matches are shown in Figure 3. Two-tailed one sample t-tests revealed that both 3-year-olds (M = 0.62, SD = 0.18) and 4-year-olds (M = 0.78, SD = 0.20) performed significantly better than chance, $t(19) = 2.97$, $p = .008$, and $t(19) = 6.24$, $p < .001$, respectively. However, the 4-year-olds made a significantly higher proportion of relational matches than the 3-year-olds, $t(38) = 2.63$, $p = .01$. Children in both age groups performed equally well on the first three and last three trials (all *p*s > .05).

We next compared the performances of the 3-and 4-year-olds in the current experiment (Comparable Alternatives condition) and those in Experiment 1. A two-way ANOVA revealed a significant main effect of age (3-year-olds vs. 4-year-olds, $F(1,76) = 7.19$, $p = .009$) and a significant main effect of condition (Experiment 1 vs. Experiment 2, $F(1,76) = 7.88$, $p = .006$). The interaction between age and condition was not significant. In both experiments, 4-year-olds performed better than 3-year-olds. Both age groups performed better in Experiment 2 than Experiment 1.

Discussion

Consistent with our hypothesis, 3-and 4-year-olds performed well on the SMTS when presented with alternatives that were composed of the same objects, but in different relational configurations. Both age groups performed significantly better in Experiment 2 than in Experiment 1. In addition, both 3-and 4-year-olds chose the symmetric match at above chance rates, whereas only the 8-and 9-year-olds in Experiment 1 were able to do so.

The two alternatives in Experiment 2 were extremely similar— the same object was used to form the object pairs on both alternative cards, with the only difference being the symmetric or asymmetric configuration between the objects. As noted above, we hypothesized that this would have two advantages: first, common objects can invite comparison between the alternatives, and this may lead to noticing that the relational patterns differ; and, second, when the same objects are used in both alternatives, children should be less likely to rely on object matches to discriminate between them, thus inviting attention to the previously less salient relational information (e.g., Mix 2008).

Consistent with this prediction, children performed markedly better in Experiment 2 than in Experiment 1. When the relation depicted by each alternative card was more salient, the process of detecting and matching these relations

(comparison and contrast) seemed to be more fluent. Thus, 3- and 4-year-olds who previously were not able to pass the SMTS were able to do so when presented with comparable alternatives.

General Discussion

Across two experiments, we explored children's ability to perceive and match symmetry between two objects using a Symmetry-Match-to-Sample (SMTS) task. In Experiment 1, we found a long developmental course for between-object symmetry matching: children did not pass the task until after 6 years of age.

In Experiment 2, we explored the method of Comparable Alternatives in facilitating children's relational insight. We presented children with a matching task in which the two alternatives were composed of the same objects. With Comparable Alternatives, even 3-and 4-year-olds chose the symmetric match at significantly above chance rates. We propose that there were two reasons for this improvement. First, the common objects promoted online comparison between the two alternatives, setting the stage for children to discover the crucial difference between them—whether or not the two objects in each alternative were symmetrical to each other. Second, using common objects in the two alternatives signaled to the children that object similarity could not be the basis for matching, thus allowing them to shift their attention to relations.

Symmetry does not inform *Same/Different* Detection

The present findings provide evidence against the claim that symmetry detection informs *same/different* detection. Researchers have consistently found that 4-and 5-year-olds can pass the standard Relational-Match-to-Sample (RMTS; with unique objects) task without any prior training, corrective feedback, or linguistic assistance (Christie & Gentner, 2014; Hochmann et al., 2017, Hoyos, Shao, & Gentner, 2016). However, children do not pass a similar Symmetry-Match-to-Sample (SMTS) task (Experiment 1) until after 6 years of age. Taken together, these findings suggest that children are not passing the RMTS by responding to symmetry. In fact, between-object symmetry matching appears to emerge later than *same*/*different* matching.

Bootstrapping relational insight

Because of the importance of comparison in acquiring relational insight, a number of techniques have been explored for promoting relational comparison. One such technique is Progressive Alignment—the phenomenon whereby carrying out relatively concrete and easy-to-align matches promotes subsequent ability to match less surface-similar, more challenging pairs that instantiate the same relation (e.g., Gentner, Loewenstein, & Hung, 2007; Haryu, Imai & Okada, Kotovsky & Gentner, 1996; Loewenstein & Gentner, 2001).

In Kotovsky and Gentner's (1996) initial study, 4-year-olds only succeeded on trials that involved concrete matches, suggesting that they did not have an abstract representation of the symmetry relation. In a follow-up study, Kotovsky and Gentner (1996) presented a new group of 4-year-olds with the same trials as before, but in an order designed to promote progressive alignment. Children were first shown a block of concrete (within-dimension) trials and then progressed on to more abstract (across-dimension) trials. The 4-year-olds showed a gain in performance on the abstract trials.

The technique used in Experiment 2—Comparable Alternatives—is another way to scaffold children's relational insight. Here, the two alternatives share the same objects but instantiate different relations, only one of which matches the standard. This design not only promotes comparison between the two alternatives, potentially highlighting the relational difference, but also de-emphasizes the role of objects, signaling that objects are not the basis for matching.

To our knowledge, the current study is the first to explicitly investigate whether the use of comparable alternatives is a way to promote relational insight. Prior studies have used alternatives that share common objects in relational matching tasks, but have not investigated whether this procedure promotes relational insight than using standard dissimilar alternatives (e.g., Kotovsky & Gentner, 1996; Mix, 2008). We are currently investigating whether presenting children with relatively easy comparable alternatives trials could serve to bootstrap later performance on the more abstract SMTS—analogous to progressive alignment.

Within-Object Versus Between-Object Symmetry

In this paper, we focused on a relatively overlooked aspect of children's symmetry development—the ability to detect and match between-object symmetry. Our findings contrast with a large body of research on within-object symmetry detection that has viewed symmetry as a low-level visual feature. We found evidence suggesting that between-object symmetry—at least for 3- and 4-year-olds—can be perceived and processed as a relation. As with other relations, children's initial relational representations may be quite concrete (Gentner, 2010); but further experience—notably experience in comparing examples (and nonexamples) of the relation—can lead to more abstract, portable representations.

This leads to the question of whether the representations and mechanisms that support processing within-object symmetry are the same as those that support processing between-object symmetry. For example as noted above, there is evidence that many animals can detect within-object symmetry. Can the same species detect between-object symmetry, and can they construe symmetry as an abstract relation?

Although we do not have the answers to these questions, we propose that researchers may take inspiration from the existing rich literature on *same/different* processing. Premack (1983), among others, has proposed species graded differences in relational reasoning ability (see also Gentner, 2003; 2010; and Penn, Povinelli & Holyoak, 2008). A substantial body of empirical findings supports this proposal. For example, there are more species that can learn to

discriminate between *same* and *different* pairs (e.g., rhesus macaques [Katz, Wright, & Bachevalier, 1984] than species that can learn to make relational matches based on *same*/*different* pairs, hence passing the Relational-Match-to-Sample task (e.g., chimpanzees [Premack, 1983; Thompson, Oden, & Boysen, 1997]; and hooded crows [Smirnova, Zorina, Obozova, & Wasserman, 2015]). Does a similar distinction hold for symmetry? If so, we would expect to see a gradient between species that can detect symmetry and those that can pass a Symmetry-Match-to-Sample task, as investigated here.

Conclusions

The present work provides an initial exploration of the development of insight into the symmetry relation. Using a Symmetry-Match-to Sample task, we found that the ability to process relational matches based on symmetry emerges relatively late in development. However, as with other relations, insight into the symmetry relation can be scaffolded through comparison processes. The present work also explores a novel way of promoting relational insight—using comparable alternatives that share objects but not relations. These findings underline the importance of comparison in supporting children's understanding of symmetry and other relations.

Acknowledgements

This work was supported by the Office of Naval Research, ONR grant N00014-16-1-2613 to Dedre Gentner. We thank the children and families who participated in this research, members of the Cognition and Language Lab for help with data collection, and Christian Hoyos for his insightful comments.

References

Bornstein, M. H., & Stiles-Davis, J. (1984). Discrimination and memory for symmetry in young children. *Developmental Psychology, 20*(4), 637.

Carmody, D. P., Nodine, C. F., & Locher, P. J. (1977). Global detection of symmetry. Perceptual and motor skills, 45(3_suppl), 1267-1273.

Cattaneo, Z. (2017). The neural basis of mirror symmetry detection: a review. *Journal of Cognitive Psychology*, *29*(3), 259-268.

Chipman, S. F., & Mendelson, M. J. (1979). Influence of six types of visual structure on complexity judgments in children and adults. *Journal of Experimental Psychology: Human Perception and Performance, 5*(2), 365.

Christie, S., & Gentner, D. (2014). Language helps children succeed on a classic analogy task. *Cognitive Science, 38*(2), 383-397.

Delius, J. D., & Nowak, B. (1982). Visual symmetry recognition by pigeons. *Psychological Research, 44*(3), 199-212.

Driver, J., Baylis, G.C., & Rafal, R. D. (1992). Preserved figure-ground segregation and symmetry perception in visual neglect. *Nature, 360*(6399), 73-75.

Fisher, C. B., Ferdinandsen, K., & Bornstein, M. H. (1981). The role of symmetry in infant form discrimination. *Child Development,52*(2), 457-462.

Gentner, D. (1988). Metaphor as structure mapping: The relational shift. *Child Development, 59*(1), 47-59.

Gentner, D. (2003). Why we're so smart. In D. Gentner and S. Goldin-Meadow (Eds.), *Language in mind: Advances in the study of language and thought* (pp.195-235). Cambridge, MA: MIT Press.

Gentner, D. (2010). Bootstrapping the mind: Analogical processes and symbol systems. Cognitive Science, 34(5), 752-775.

Gentner, D., Loewenstein, J., & Hung, B. (2007). Comparison facilitates children's learning of names for parts. *Journal of Cognition and Development, 8*(3), 285-307.

Gentner, D., & Rattermann, M. J. (1991). Language and the career of similarity . In S. A. Gelman &. J. P. Byrnes (Eds.), *Perspectives on thought and language: Interrelations in development* (pp. 225-277). London: Cambridge University Press.

Gentner, D., & Toupin, C. (1986). Systematicity and surface similarity in the development of analogy. *Cognitive Science, 10*(3), 277-300.

Giannouli, V. (2013). Visual symmetry perception. *Encephalos, 50*, 31-42.

Giurfa, M., Eichmann, B., & Menzel, R. (1996). Symmetry perception in an insect. *Nature, 382*(6590), 458.

Goldstone, R. L., & Son, J. Y. (2005). The transfer of scientific principles using concrete and idealized simulations. *The Journal of the Learning Sciences, 14*(1), 69-110.

Grammer, K., & Thornhill, R. (1994). Human (Homo sapiens) facial attractiveness and sexual selection: the role of symmetry and averageness. *Journal of Comparative Psychology, 108*(3), 233.

Gross, D. J. (1996). The role of symmetry in fundamental physics. *Proceedings of the National Academy of Sciences*, *93*(25), 14256-14259.

Haryu, E., Imai, M., & Okada, H.(2011) Object similarity bootstraps young children to action-based verb extensions. *Child Development*, 82(2). 674–686.

Hochmann, J. R., Tuerk, A. S., Sanborn, S., Zhu, R., Long, R., Dempster, M., & Carey, S. (2017). Children's representation of abstract relations in relational/array match-to-sample tasks. *Cognitive Psychology*, *99*, 17-43.

Hoyos, C., Shao, R., & Gentner, D. (2016). The paradox of relational development: Could language learning be (temporarily) harmful? D. Grodner, D. Mirman, A. Papafragou, J. Trueswell, J. Novick, S. Arunachalam, S. Christie, & C. Norris (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Kaminski, J. A., Sloutsky, V. M., & Heckler, A. F. (2008). The advantage of abstract examples in learning math. S*cience, 320*(5875), 454-455.

Katz, J. S., Wright, A. A., & Bachevalier, J. (2002). Mechanisms of same-different abstract-concept learning by rhesus monkeys (Macaca mulatta). *Journal of Experimental Psychology: Animal Behavior Processes, 2*8(4), 358.

Kotovsky, L., & Gentner, D. (1996). Comparison and categorization in the development of relational similarity. *Child Development, 67*(6), 2797-2822.

Loewenstein, J., & Gentner, D. (2001). Spatial mapping in preschoolers: Close comparisons facilitate far mappings. *Journal of Cognition and Development, 2*, 189-219.

Marshall, J.C., & Halligan, P.W. (1994). The Yin and the Yang of visuo-spatial neglect: A case study. *Neuropsychologia, 32*(9), 1037–1057.

Mix, K. S. (2008). Children's equivalence judgments: Crossmapping effects. *Cognitive Development, 23*(1), 191-203.

Møller, A. P., & Thornhill, R. (1998). Bilateral symmetry and sexual selection: a meta-analysis. *The American Naturalist, 151*(2), 174-192.

Penn, D. C., Holyoak, K. J., & Povinelli, D. J. (2008). Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. *Behavioral and Brain Sciences*, *31*(2), 109-130.

Pepperberg, I. M. (1987). Acquisition of the same/different concept by an African Grey parrot (Psittacus erithacus): Learning with respect to categories of color, shape, and material. *Animal Learning and Behavior, 15*(4), 423-432.

Premack, D. (1983). The codes of man and beasts. *Behavioral and Brain Sciences, 6(*1), 125-136.

Richland, L. E., Morrison, R. G., & Holyoak, K. J. (2006). Children's development of analogical reasoning: Insights from scene analogy problems. *Journal of Experimental Child Psychology, 94*(3), 249-273.

Royer, F. L. (1981). Detection of symmetry. *Journal of Experimental Psychology: Human Perception and Performance*, 7(6), 1186.

Schluessel, V., Beil, O., Weber, T., & Bleckmann, H. (2014). Symmetry perception in bamboo sharks (Chiloscyllium griseum) and Malawi cichlids (Pseudotropheus sp.). *Animal Cognition, 17*(5), 1187-1205.

Smirnova, A., Zorina, Z., Obozova, T., & Wasserman, E. (2015). Crows spontaneously exhibit analogical reasoning. *Current Biology, 25*(2), 256-260.

Thompson, R. K. R., Oden, D. L., & Boysen, S. T. (1997). Language-naive chimpanzees (Pan troglodytes) judge relations between relations in a conceptual matching-to-sample task. *Journal of Experimental Psychology: Animal Behavior Processes, 23(1),* 31-43.

Treder, M. S. (2010). Behind the looking-glass: A review on human symmetry perception. *Symmetry, 2*(3), 1510-1543.

von Fersen L., Manos C.S., Goldowsky B., Roitblat H. (1992) Dolphin Detection and Conceptualization of Symmetry. In: Thomas J.A., Kastelein R.A., Supin A.Y. (eds) *Marine Mammal Sensory Systems*. Springer, Boston, MA

Wagemans, J. (1997). Characteristics and models of human symmetry detection. *Trends in Cognitive Sciences, 1*(9), 346-352.

Walker, C. M., & Gopnik, A. (2017). Discriminating relational and perceptual judgments: Evidence from human toddlers. *Cognition*, *166*, 23-27.

Is an over-polite compliment worse than an impolite insult?: Pragmatic effects of non-normative politeness in Korean

Hagyeong Shin (hshin@sdsu.edu), Gabriel Doyle (gdoyle@sdsu.edu)
Department of Linguistics, San Diego State University
5500 Campanile Drive, San Diego, CA USA 92182

Abstract

Honorifics in Korean appear as verbal inflections and have been considered as markers of politeness. This study investigates the pragmatic effects of honorifics, and suggests that honorifics can contribute to the semantic interpretation of verb phrases in complex ways. Native Korean speakers reported different inferred meanings of "did very well" and "did very poorly" based on the normative or non-normative honorific forms. We found significant effects of non-normative honorifics in positive assessments: over-polite honorifics brought negative interpretations. This suggests that pragmatic listeners interpret utterances based on the interaction between literal meanings, honorifics, and the normativity of the honorifics within a relationship context, to obtain an estimate of the speaker's intended meaning. This is inconsistent with the previous explanations of honorific usage as discernment or volitional politeness. We suggest that non-literal meaning inferences reflect listeners treating the honorifics as signals to potential communicative goals.

Keywords: pragmatics; semantics; politeness; honorifics; pragmatic inference; Korean

Introduction

Languages have many ways of expressing politeness. Some languages explicitly mark politeness with *honorifics*: grammaticalized or lexicalized forms for politeness. Honorifics are prevalent in languages such as Japanese, Javanese, Hindi, and the subject of this investigation: Korean. Because appropriate honorifics depend on the speaker-listener relationship, they primarily function as a reflection of social norms (*discernment politeness*). However, speakers may strategically deviate from the normative form in certain contexts, such as when making requests (*volitional politeness*, Hill et al., 1986).

In this paper, we investigate whether such deviations integrate more generally into the pragmatic inference process that listeners undertake when interpreting a message. Specifically, we look at whether a speaker's choice of honorific forms influences how a listener assesses the speaker's true opinion. We carry out judgment experiments to compile data on Korean listeners' interpretations. Our result shows that the inferred meaning of the message changes with honorifics in a complex manner that cannot be adequately explained by either strictly normative or strictly strategic use of honorifics. Instead, honorifics could be used as pragmatic signals to the meaning depending on the context.

Overall, this suggests that despite grammaticalized forms seeming to be low in semantic content, they can still significantly influence the inferred meaning of the message. We argue that a full understanding of honorific use will require their incorporation into frameworks of pragmatic inference, such as the Rational Speech Act framework (RSA, Frank & Goodman, 2012; Goodman & Stuhlmüller, 2013).

Table 1: Honorific inflections of the past tense of "do" (*ha-*) in Korean. Honorification decreases from top to bottom and left to right. *-ess* is past tense suffix.

Speech Level	Honorific Suffix *-sy*	
	Present	Absent
DEF (deferential)	ha-**sy**-ess-*supnita*	ha-ess-*supnita*
POL (polite)	ha-**sy**-ess-e-*yo*	ha-ess-e-*yo*
INT (intimate)	ha-**sy**-ess-*e*	ha-ess-*e*
PLN (plain)	ha-**sy**-ess-*ta*	ha-ess-*ta*

Honorifics in Korean

Honorifics in Korean have two main realizations: honorific lexical items and honorific inflections. This study focuses on verbal honorific inflections, specifically speech levels and the *-sy* suffix. Table 1 demonstrates some of the honorific inflections that are available for the verb "do" (*ha-*). The speech level appears at the end of the verb phrase, and reflects the relationship between the speaker and the addressee. Four levels presented in Table 1 are tested in this study, and their perceived honorification decreases from top to bottom. The presence of honorific suffix *-sy* increases the honorification toward a subject of a sentence or a referent of the verb. This study examines cases where the subject of the sentence is the addressee, thus both the suffix and the speech level refer toward the listener. Honorific inflections, therefore, can be generally defined as stylistic features reflecting the speaker and listener's position within a social hierarchy, not the truth-conditional meanings (Sohn, 1999).

In colloquial Korean, speakers must choose some level of honorific inflection to form valid verb phrases; there is no default form or level. In most cases, the appropriate honorifics can be determined by the speaker-listener relationship, as honorifics were mentioned as relationship-acknowledging devices (Matsumoto, 1988). Honorifics are grammaticalized and conventionalized in relation to the speaker-listener relationship. Speakers using appropriate honorific forms assigned by the relationship context will therefore stay aligned with the normative use of honorifics. This type of honorific use can be summarized as *discernment politeness* (Hill et al., 1986; Ide, 1989; Koo, 1995).

Besides the normative use of honorifics, they can also be used more strategically. Politeness Theory (Brown & Levinson, 1987) has explained strategic honorific use through *negative politeness*, a politeness strategy for minimizing threats to the listener's *negative face*—the desire not to be imposed

upon.[1] This form of politeness is distinguished from *positive politeness*, a strategy used to minimize threats to *positive face*—the desire to be liked or approved. In the Politeness Theory perspective, speakers use honorifics largely to mitigate the potential face threats existing in the utterance. This type of honorific use can be summarized as *volitional politeness* (Hill et al., 1986; Ide, 1989; Koo, 1995)

These explanations for honorifics' uses are well-supported, but such general politeness strategies may represent only a subset of how honorifics are actually used. We argue that deviations from normative politeness levels can function as a pragmatic signal to the listener about the intended meaning of an utterance. We suggest that honorific use ties to a more general pragmatic behavior than previously described, providing pragmatic information beyond mitigating face-threats and potentially signaling a speaker's communicative goals.

Hypotheses

Based on the above discussion, we consider three hypotheses for the potential effects of honorifics on pragmatic inference. These span from a null pragmatic effect (if honorifics mainly express the speaker-listener relationship) to a monotonic relationship between inferred meaning and levels of honorifics (if honorifics mainly manage face-threat) to a complex relationship between honorifics and inferred meaning (if honorifics provide cues about the speaker's communicative goals).

To test this, we examined listeners' inferences of values for scalars: speakers' statements that a listener had done "well" or "poorly" on a test. We first described the speaker-listener relationship, then provided assessment sentences with eight honorific inflections from Table 1, and asked participants to estimate the exam score based on the assessment. More details are in the next section, but our hypotheses and the predictions they make follow.

> **Hypothesis 1:** Honorifics are primarily about *discernment politeness*. **Changes in levels of honorifics will have no significant effects on pragmatic interpretation of scalars.**

Under this hypothesis, the speaker-listener relationship determines the appropriate honorifics, and forms that deviate from the normative standard would be similar to errors of subject-verb agreement—they could affect the perceived acceptability of a sentence, but not the meaning. If Hypothesis 1 is correct, we should not see differences in the listener's interpretations depending on the honorific forms used within the relationship context. This hypothesis is consistent with traditional analyses of the Korean honorifics, as in Sohn (1999).

> **Hypothesis 2:** Honorifics primarily serve to mitigate face threat through *volitional politeness*. **As the utterance becomes more honorific, inferred values of scalars will be monotonically decreased**.

Under this hypothesis, speakers would use higher levels of honorifics to offset the negativity of an honest assessment. Therefore, we can expect to see a monotonic decrease in listeners' inferred values of scalars as the honorific level increases, with the "poorly" condition possibly showing a larger effect due to the more explicit face-threatening assessment. Being over-polite or being under-polite (relative to normative forms) should show opposite effects on the inferred meaning. This hypothesis is similar to the threat-management account of Politeness Theory (Brown & Levinson, 1987), or the social utility addition (Yoon et al., 2016) to the RSA framework explaining listeners' discounting of compliments when they thought the speaker was being polite.

> **Hypothesis 3:** Honorifics, in addition to their discernment or volitional use, also can signal cues that influence the listener's interpretations in complex ways. **Effects of honorific levels will differ by the relationship context and the literal meanings of the utterance.**

Under this hypothesis, there will be a significant but non-monotonic effect of honorific levels. Unlike Hypothesis 2, here we do not necessarily expect under- versus over-polite messages (again, relative to normative forms) to have different effects on the listener's inference. Instead, deviations from normative honorifics could signal that the speaker is indicating different meanings or goals, for example, being hyperbolic or sarcastic. This hypothesis is similar to the QUD (Question Under Discussion) addition (Kao & Goodman, 2015) to the RSA explaining ironic interpretations.[2]

Experiment 1: Literal interpretations

Method

Design The purpose of Experiment 1 was to establish the literal baseline interpretations of the phrases "did very well" and "did very poorly". Each question in the experiment started with a vignette describing a conversation and a relationship context: a speaker is asked to tell a listener how the listener did on an exam, when the listener does not know of his own exam score. The speaker's assessments of the listener's exam score were then presented. Each participant rated 8 assessment sentences: 2 valences (positive, negative) in 4 relationship settings. Participants were asked to guess the listener's exam score in a number between 0 and 100.

Relationship settings were explicitly stated. In the Friend-Friend setting, the speaker and the listener were defined as friends who were in the same year at college. In the Upperclass-Underclass setting, the speaker was a student senior than the listener. In the Professor-Student setting, the speaker was a professor and the listener a student. In the Underclass-Upperclass setting, the speaker was a student junior than the listener. These four settings were chosen to have normative honorifics that allowed for a range of under-

[1] Despite the term, *negative politeness* is still a way of being polite; it is the "do no harm" counterpart to the "do good" sense of *positive politeness*.

[2] More details on the hypotheses can be found in the Open Science Foundation preregistration page: `http://osf.io/s8nfu/register/5771ca429ad5a1020de2872e`.

and over-polite forms by varying the honorific forms. In all settings, both the speaker and the listener had male Korean names to keep gender differences from influencing the result.

Stimuli Participants saw the speaker's description of the listener's score presented as indirect quotes (i.e., [Speaker] said [Listener] did very well/poorly on the exam), so that participants would not see what honorific inflections the speaker used and thus would respond with their baseline inference in the absence of honorifics.

Participants Experiment 1 was posted on the online crowdsourcing website Dooit Survey (`http://www.dooit.co.kr`) based in South Korea. A total of 67 adult native Korean speakers completed the experiment for a small cash-value reward.

Result

Baseline scores In Experiment 1, literal interpretations of positive and negative phrases "did very well" and "did very poorly" were measured within each relationship setting. The mean of the scores in each condition were then treated as baseline scores representing literal interpretations in further analyses, since they represented the participants' estimates in the absence of honorifics. Baseline scores in each setting and condition are presented by horizontal dashed line in Figure 1. Participants reported mean baseline scores of 85.60 for the positive and 47.92 for the negative phrases. There was no significant differences according to t-tests between the relationship settings within the positive or negative valence, suggesting that participants viewed all four relationship settings having similar expected literal meanings.

Experiment 2: Inferences from honorific use

Design Experiment 2 followed the same basic idea of Experiment 1, but participants were asked to infer scores based on direct quotes, with honorific inflections. Deviations between the literal baselines from Experiment 1 and the inferences in Experiment 2 should therefore reflect pragmatic interpretations guided by the honorifics. Each participant rated a total of 16 sentences: 2 valences (positive, negative), each with 8 honorific inflections (4 speech levels $\times$ *-sy* present/ absent), in one of the 4 relationship settings (Friend-Friend, Upperclass-Underclass, Professor-Student, Underclass-Upperclass). After presenting the relationship context and the speaker's assessment, participants were again asked to infer the listeners exam score with a number between 0 and 100.

Stimuli Each assessment sentence was presented as a direct quote (i.e., [Speaker] said the following sentence: "[Address of the listener] did very well."). The presence of a direct quote meant that the sentence included one of the eight honorific inflections from Table 1, and therefore could influence participants' inferences accordingly. A sample vocative address of the listener by the speaker was included in these sentences to reinforce the normative honorifics for each relationship. In the Friend-Friend and Upperclass-Underclass setting, where the speaker was in an equal or higher position to the listener, the speaker addressed the listener with a plain "you". In the Underclass-Upperclass setting, the honorific addressee term *senbay-nim* was used. In the Professor-Student setting, no addressee term was presented, because the speaker is on a much higher social rank than the listener and could in principle use any of the honorific inflections. Below shows the assessment sentences given in the Friend-Friend setting, with *-sy* and the deferential speech level[3].

neo	*cham*	*cal*	*ha-sy-ess-supnita*
You	very	well	do.AH.PST.DEF

Positive: "You did very well."

neo	*cham*	*mos*	*ha-sy-ess-supnita*
You	very	poorly	do.AH.PST.DEF

Negative: "You did very poorly."

Participants Experiment 2 was also posted on Dooit survey. Unlike Experiment 1, we asked each participant to answer for only one relationship type, to avoid any confusion about the speaker-listener relationships. 81 adult Korean participants were collected in total, with 20 participants in three settings and 21 participants in the Underclass-Upperclass setting. The participants in Experiment 1 and 2 were recruited separately.

Results

Baseline scores vs. Normative scores The baseline scores from Experiment 1 were then compared against the "normative" scores from Experiment 2. These normative scores are the mean of the inferred scores in each setting obtained from sentences with normative honorific forms established for that setting. Normative scores are presented by black points in Figure 1. For example, in the Friend-Friend setting, normative scores were calculated from the scores reported on sentences with intimate or plain speech level without an honorific suffix *-sy* (INT, PLN, black triangles in Figure 1). The Professor-Student setting did not include a specific addressee term to establish normative forms, because of the high social position the speaker was in. To calculate the baseline scores, we considered deferential and polite speech levels used as teacher's classroom register (Sohn, 1999). In Figure 1, normative forms (black points) aligned closely to the baseline scores (dashed line), showing that participants treated utterances with normative honorific forms similarly to the literal interpretations. This is confirmed by the regression below.

Baseline scores vs. Non-normative scores Scores reported on non-normative honorific forms are presented by the red points in Figure 1. In the Friend-Friend setting, participants reported the baseline score of 82.74 for the positive condition and 46.68 for the negative. When positive assessments appeared with non-normative *-sy* (red circles), regardless of the following speech levels, participants reported scores far lower than the baseline score (*sy*+DEF: 54.25, -28.49 score difference from the baseline score; *sy*+POL: 50.10, -32.64; *sy*+INT: 46.00, -36.74; *sy*+PLN: 56.25, -26.49)[4]. This is

[3] AH: Addressee honorification, PST: Past tense suffix, DEF: Deferential speech level

[4] We test for significance on these values in the following regression model.

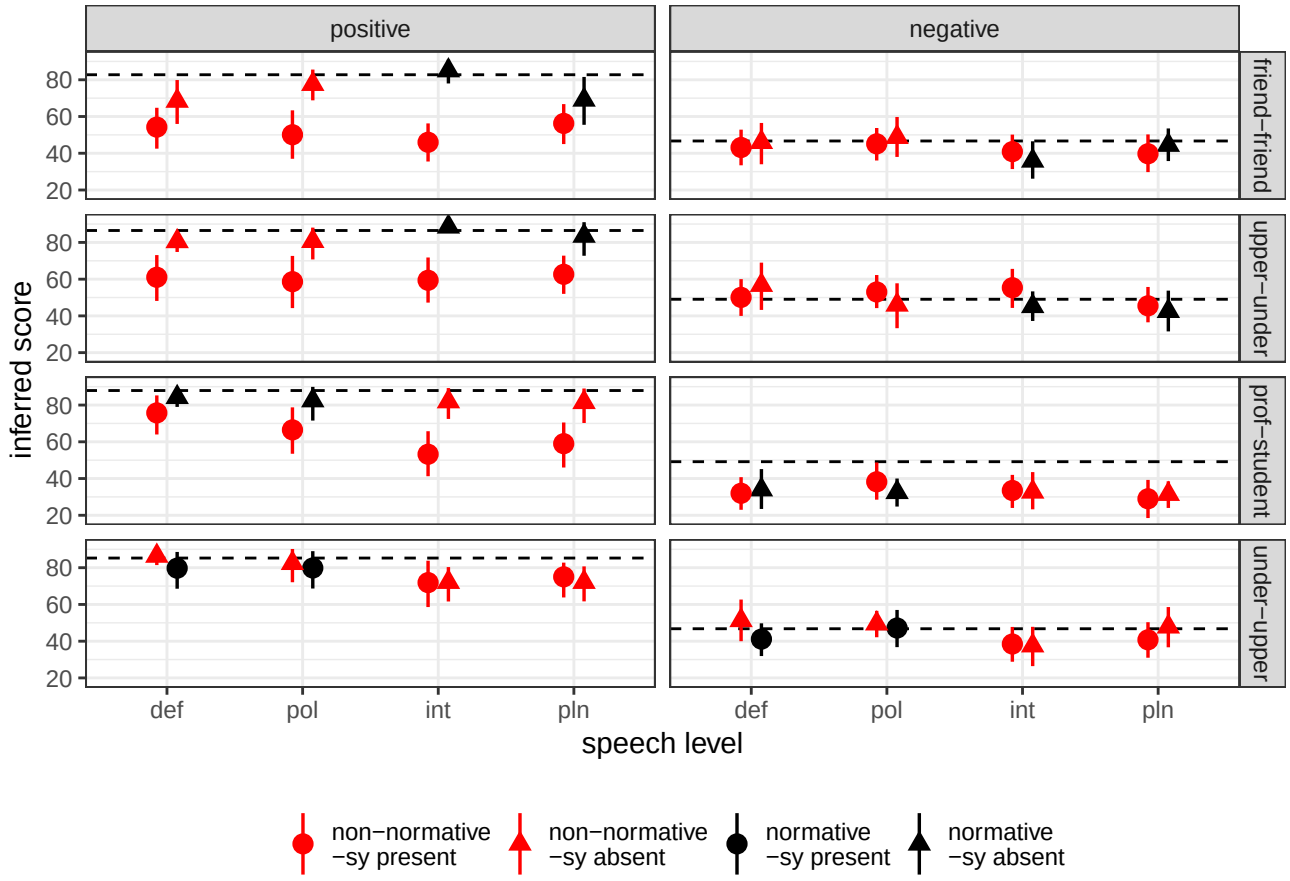


Figure 1: Participants inferred exam scores under valence (positive, negative), honorific suffix *-sy* (*-sy* present, *-sy* absent), and speech level (DEF, POL, INT, PLN) conditions in each relationship settings (friend-friend, upper-under, prof-student, under-upper). Circles and triangles indicate the *-sy* suffix being present or absent. The colors show the normativity of the forms, red being non-normative and black being normative. The dashed line in each condition shows baseline scores (from Expt 1). Vertical lines in each score point show 95% confidence interval calculated from 5000 bootstrap samples.

our first piece of evidence that over-polite forms can induce large pragmatic effects that substantially reduce the estimates of the test scores.

In the Upperclass-Underclass setting (Upperclass speaker), participants reported the baseline score of 86.49 for the positive condition and 49.05 for the negative. Normative forms were defined as intimate or plaivel without the *-sy* suffix (INT, PLN, black triangles). When positive utterances appeared with the *-sy* suffix (red circles), participants reported scores below the baseline score across the speech levels (*sy*+DEF: 61.05, −25.44; sy+POL: 58.65, −27.84; *sy*+INT: 59.40, −27.09; *sy*+PLN: 62.65, −23.84). Again, over-polite forms caused participants' pragmatic inferences to substantially drop.

Professor-Student setting showed the similar result. The baseline score was 87.91 for the positive condition and 49.17 for the negative. Normative forms in this setting were the deferential or polite speech level without the *-sy* suffix (DEF, POL, black triangles). Participants reported lower scores when the professor's positive feedback were given with non-normative *-sy* (*sy*+DEF: 75.75, −12.60; *sy*+POL: 66.50, −21.41; *sy*+INT: 53.25, −34.66; *sy*+PLN: 59.00, −28.91). Under-polite forms (the intimate/plain speech levels) do not show any increase over the normative forms, as Hypothesis 2 would have predicted.

In the Underclass-Upperclass setting, the baseline score was 85.26 for the positive condition and 46.76 for the negative. The normative forms were defined as deferential or polite speech level with the *-sy* suffix (*sy*+DEF, *sy*+POL, black circles). This relationship setting showed the least amount of score variance among all settings. One explanation could be that non-normative forms in this setting produced outright socially unacceptable sentences. Not coincidentally, this is the one setting where the speaker is of a lower social standing than the listener. Since a lower-standing speaker speaking in under-polite forms violates social norms in a great degree, participants might have been confused with those sentences and have failed to properly reason on the meaning.

Table 2: Estimated effect sizes in the linear regression with random by-participant intercepts. Default values for Valence, Speech Level and Setting are negative, deferential and Underclass-Upperclass, respectively.

	β
(Intercept)	-0.29
Valence	
Positive	3.15
Speech Level	
Intimate	-4.40
Plain	-4.10
Polite	0.77
***-sy* suffix**	
Present	-0.63
Setting	
Friend-Friend	-1.14
Professor-Student	-13.39 ***
Upper-Under	3.12
Valence × Speech Level	
Positive × Intimate	0.27
Positive × Plain	0.12
Positive × Polite	-2.30
Valence × *-sy* suffix	
positive × *-sy* present	-15.88 ***
Valence × Setting	
Positive × Friend-Friend	-10.46 **
Positive × Professor-Student	6.67
Positive × Upper-Under	-9.62 **

$^{***}p < 0.001, ^{**}p < 0.01, ^{*}p < 0.05$

Overall, the presence of the over-polite and non-normative honorific *-sy* suffix when talking to an equal or lower-standing listener signaled participants that the positive feedback could not be taken literally, and participants substantially reduced their estimates of the test scores. This is strongly at odds with Hypothesis 1. At the same time, non-normative scores (red points) were mostly at or below the normative scores (black points) in the positive valence cases, regardless of the non-normative forms being over- or under-polite. This argues against both Hypothesis 1 and Hypothesis 2.

Linear mixed-effect model

We fit a linear mixed-effect model predicting the score differences (inferential score from Expt 2 minus the baseline score from Expt 1) with random by-participant intercepts. The fixed-effect variables were Valence × Speech level, Valence × *-sy* suffix, Valence × Setting. Regression coefficients (β) are reported in Table 2. The default values were set to the scores obtained in the Under-Upper setting, from negative sentences with no *-sy* suffix and the deferential speech level.

Starting from the top of Table 2, the intercept term confirmed that there were no significant differences between inferred scores and baseline scores in the default setting of the model. Among the main effects of the Setting, only the Professor-Student setting showed significantly lowered scores. This was because the default condition (no *-sy* suffix, deferential speech level) of the model was a normative form in the setting, thus brought literal (thus, more negative) meaning. Overall, the main effects were largely small and non-significant, with the exception of one setting. This confirmed our initial anticipation that honorifics' meanings should be considered relative to the context, including the literal meanings of the message and the speaker-listener relationship.

Moving onto the interaction terms, Valence × *-sy* suffix had a large effect, lowering the inferred scores by 15.88 points from the baseline ($p < 0.001$). This shows that positive sentences with the *-sy* suffix in the three settings where the speaker was at least equal in social standing to the listener (Friend-Friend, Professor-Student, Upperclass-Underclass) showed lower scores than in the one where the speaker was of higher social standing. Those three settings shared the normativity context that the *-sy* suffix was an over-polite form, and the inferred scores dropped as a result.

Valence × Setting term also reflected the result shown in Figure 1. Friend-Friend and Upperclass-Underclass showed lowered scores in positive sentences with deferential speech level (this was default setting of the model, which was a non-normative form in the relationship). The Professor-Student setting did not show significant differences from Underclass-Upperclass setting, again because the deferential speech level was the normative form in the relationship. This showed that the Friend-Friend and Upperclass-Underclass setting behaved similarly, as higher honorifics became non-normative and over-polite forms. We could see that Professor-Student and Underclass-Upperclass behaved similarly as well. These two settings shared deferential and polite speech levels (higher honorifics) as normative forms.

In both the numeric values in Figure 1 and the regression coefficients in Table 2, we see a few patterns. First, participants' inferred scores varied substantially based on the honorifics, contrary to Hypothesis 1, which considered the honorifics to primarily serve as an agreement to the relationship. This suggests that listeners assume that speakers have made volitional choices in their honorific inflections when they deviate from normativity. Second, looking at the regression model, we see strong evidence of an effect of the *-sy* suffix in non-normative context, generally lowering the score differences (inferred score − baseline score). But there is no consistent effect with the speech levels. This runs counter to the expectation of Hypothesis 2; there is no monotonic relationship between the honorific levels and the inferred scores. Instead, we see a complex pattern that is driven largely by the normativity of the forms, rather than their relative politeness. In the next section, we discuss the implications of these results and sketch a possible explanation for the phenomenon.

Discussion

The experiment results showed clear differences on listeners' interpretations, depending on the honorific inflections, the speaker-listener relationship, and the valence of the verb phrases. Contrary to Hypothesis 1, we saw significant effects of honorifics on the pragmatically-inferred values even within a given relationship setting. Contrary to Hypothesis 2, we did not see a monotonic effect of politeness levels; deviation from the normative form generally decreased (or maintained) the inferred value of positive verb phrases regardless of whether the non-normative form was more or less polite than expected. Furthermore, in the negative valence, normative and non-normative forms showed no significant difference in their interpretation. These results suggest that while there appears to be a volitional component to the speaker's choice of honorific forms, the choice extends beyond straightforward face-threat mitigation.

Hypothesis 3 therefore appears to be the best fit to our data, though in some sense it reflects a less specific explanation at present. How can we further expand our hypothesis to explain the observed variance in pragmatic inferences? A promising direction is to build on the Rational Speech Act (RSA, Frank & Goodman, 2012; Goodman & Stuhlmüller, 2013) framework. The core idea of the RSA is that speakers and listeners each know that the other is trying to communicate in an efficient, rather than a literal, manner. The listener considers the speaker's choice of utterance as a rational decision over the set of available alternatives. The speaker prefers utterances that maximize the expected conversational utility (such as maximizing the listener's probability of inferring the intended message). The listener then uses a recursive inference process to determine the most likely meaning.

Two particular extensions of RSA inference may be relevant for unravelling the inferences that result from honorifics. Yoon et al. (2016) proposed that the speaker not only has a desire to provide epistemic utility in their utterance (giving the listener an accurate representation of the world) but also social utility (such as minimizing listener's face-threats). Honorifics, especially in the Politeness Theory (Brown & Levinson, 1987) framework, can supply social utility alongside the epistemic utility of the message itself. This fits with, for instance, speakers' selective use of honorifics when making requests. In our data, however, we see that the same honorific forms lead to substantially different inferences based on the setting. Even if we view social utility relative to the speaker-listener relationship, with over-polite forms adding social utility and under-polite forms reducing it, this would still not be sufficient to explain the variation in Figure 1.

By combining the idea of social utility with a goal- or QUD-based approach (Kao & Goodman, 2015) in RSA, though, we may be able to capture the pragmatic effects of honorifics. A Goal/QUD framework says that when a speaker produces a message that seems to violate the listener's expectations, the listener may instead interpret the message with a different goal in mind. For example, if a speaker complains that they paid an unbelievably high cost for some object, the listener may infer that the speaker's epistemic utility is not coming from conveying the literal cost but rather an affective interpretation of the cost (i.e., hyperbole).

Building on these extensions to the RSA framework, we suggest that Korean honorifics may be modelled as an interaction between the relationship context r, shared knowledge of normative honorifics k, an intended meaning s, and a goal g. The speaker's choice of utterance can be broken down into the semantic content of the word stem c and the honorific inflections m:

$$P_{speaker}(c, m|s, r, k, g) \tag{1}$$

This expresses the idea that a speaker chooses c and m jointly to deliver their intended meaning s, conditioned on the relationship context r and normativity k for the honorifics, as well as their communicative goal or QUD g. If we assume that the listener has no uncertainty about the relationship r or normativity k, we can express the listener's inference process as Bayesian inference, marginalized over the potential goals of the speaker:[5]

$$P_{listener}(s|c, m, r, k) \propto \sum_{g} P_{speaker}(c, m|s, r, k, g)P(s)P(g) \tag{2}$$

This joint distribution over c and m gives the model the flexibility to capture the complex patterns in our results in a way that a basic social utility term alone cannot. Being overly polite may come from the speaker signalling their ironic intentions by violating normative expectations of the honorific inflections. When such a deviation from the norms is slight, or consistent with a goal of mitigating face-threat, the listener merely tweaks their interpretation. When the deviation from the honorific norms is large (as when a student is overly polite to their friend, or the professor talks to the student with the honorific *-sy* suffix), the listener assumes the speaker's goal has changed. In this way, honorifics signal cues to the meaning similar to the inferred product prices in Kao et al. (2014); a small deviation from expectations retains an approximately literal interpretation, while a large deviation triggers an ironic interpretation. This argument could be verified by a follow-up experiment measuring inferred goals.

Conclusion

We have examined honorific inflections and their effect on pragmatic inference. Contrary to discernment or volitional politeness accounts, we find complex interactions between honorifics and a listener's pragmatic interpretation. We propose that this result may be explained with an extended RSA framework with jointly-distributed content and honorifics that can both provide social utility and serve to signal a speaker's communicative goals.

[5] Of course, the listener may want to update their belief about their relationship with the speaker based on the speaker's choice of honorifics! If so, the listener could marginalize over r and k.

Acknowledgements

We thank three anonymous reviewers for helpful comments and suggestions; all mistakes remain ours.

References

Brown, P., & Levinson, S. C. (1987). *Politeness: Some universals in language usage*. Cambridge University Press.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998.

Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and Implicature: Modeling Language Understanding as Social Cognition. *Topics in Cognitive Science*, *5*(1), 173–184.

Hill, B., Ide, S., Ikuta, S., Kawasaki, A., & Ogino, T. (1986). Universals of linguistic politeness: Quantitative evidence from Japanese and American English. *Journal of pragmatics*, *10*(3), 347–371.

Ide, S. (1989). Formal forms and discernment: Two neglected aspects of universals of linguistic politeness. *Multilingua-journal of cross-cultural and interlanguage communication*, *8*(2-3), 223–248.

Kao, J. T., & Goodman, N. D. (2015). Let's talk (ironically) about the weather: Modeling verbal irony. In *Proceedings of the 36th conference of the cognitive science society* (pp. 1051–1056).

Kao, J. T., Wu, J. Y., Bergen, L., & Goodman, N. D. (2014). Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences*, *111*(33), 12002–12007.

Koo, J. S. (1995). *Politeness Theory: Universality and Specificity*. Unpublished doctoral dissertation, Harvard university.

Matsumoto, Y. (1988). Reexamination of the universality of face: Politeness phenomena in Japanese. *Journal of Pragmatics*.

Sohn, H. M. (1999). *The Korean Language*. Cambridge: Cambridge University Press.

Yoon, E. J., Tessler, M. H., Goodman, N. D., & Frank, M. C. (2016). Talking with tact: Polite language as a balance between kindness and informativity. In *Proceedings of the 38th annual conference of the cognitive science society* (pp. 2771–2776).

Impact of Explicit Failure and Success-driven Preparatory Activities on Learning

Tanmay Sinha[1], **Manu Kapur**[1], **Robert West**[2], **Michele Catasta**[3], **Matthias Hauswirth**[4], **Dragan Trninic**[1]
[1] ETH Zürich, Switzerland, [2] EPFL Lausanne, Switzerland, [3] Stanford University, USA, [4] University of Lugano, Switzerland

Abstract

Unscaffolded problem-solving before receiving instruction can give students opportunities to entertain their exploratory hypotheses at the expense of experiencing initial failures. Prior literature has argued for the efficacy of such Productive Failure (PF) activities in preparing students to "see" like an expert. Despite growing understanding of the socio-cognitive mechanisms that affect learning from PF, the necessity of success or failure in initial problem-solving attempts is still unclear. Consequently, we do not know yet whether some ways of succeeding or failing are more efficacious than others. Here, we report empirical evidence from a recently concluded classroom PF intervention (N=221), where we designed scaffolds to explicitly push student problem-solving towards success via structuring, but also radically, towards failure via problematizing. Our rationale for explicit failure scaffolding was rooted in facilitating problem-space exploration. We subsequently compared the differential preparatory effects of success-driven and failure-driven problem-solving on learning from subsequent instruction. Results suggested explicit failure scaffolding during initial problem-solving to have a higher impact on conceptual understanding, compared to explicit success scaffolding. This trend was more salient for the task topic with greater difficulty.

Keywords: Classroom Study; Productive Failure; Scaffolding

Introduction

Substantial research has demonstrated the efficacy of learning approaches where problem-solving as a preparatory activity precedes instruction (PS-I). PS-I includes (i) an initial problem-solving phase where students explore solutions to complex problems based on concepts they haven't formally learnt yet, and (ii) a subsequent explicit instruction phase where a coach introduces formalisms of the targeted concepts along with the canonical solution. Research suggests that PS-I is an effective learning design that improves student's conceptual understanding and positively impacts how well they transfer their knowledge to novel problem-solving contexts (Loibl, Roll, & Rummel, 2017).

A particular variant of the PS-I design that embodies learning from failure is Productive Failure (PF) (Kapur & Bielaczyc, 2012). PF comprises rich problem design that affords multiple representations and solution methods (RSMs), and follow-up instruction that compares and contrasts student-generated solutions with the canonical one. The positive benefits of approaches implemented based on the PS-I design (e.g., PF, Invent with Contrasting Cases (Schwartz & Martin, 2004)) have been attributed to different cognitive mechanisms. These include intentional activation of relevant prior knowledge, enhancement of students' awareness of the problem situation and own knowledge gaps, focused attention on search for deeper patterns rather than surface characteristics, and effortful retrieval to resolve incongruity. Some posited socio-emotional mechanisms include increased motivation to learn targeted concepts and elicitation of curiosity (Kapur & Bielaczyc, 2012; Loibl et al., 2017).

Research Gap

Despite PS-I designs often working better compared to traditional instructional approaches (usually direct instruction) on the acquisition of conceptual knowledge and/or transfer, there is a considerable variation in effect sizes (Cohen's d = 1.12$\pm$ 0.54) (Loibl et al., 2017). This has spurred lines of inquiry into systematically analyzing reasons for failure of PS-I approaches (Sinha & Kapur, 2019), and developing ways to improve overall effectiveness of the learning design. One prominent area of focus has been the initial problem-solving phase. Here, research has started to investigate the impact of scaffolding student solutions on fostering conceptually sound and transferable learning (Kapur, 2011; Loibl & Rummel, 2014). Despite growing research in the PS-I design space, we don't have conclusive evidence yet.

Templates of successful problem-solving usually aim at pro-active error elimination, and directing student's attention to the task by providing immediate feedback. Such instruction has the advantage of helping students perform the correct procedure. However, this may not always imply that students engage in optimal reasoning or acquire high depth of understanding of domain principles. Evidence favoring success-driven (SD) learning in PS-I suggests the presence of an association between successful problem-solving during the problem-solving phase and learning from instruction (e.g., Chin, Chi, and Schwartz (2016); Schwartz, Chase, Oppezzo, and Chin (2011); Loibl and Rummel (2014); Schalk, Schumacher, Barth, and Stern (2017); Chase and Klahr (2017)). However, attempts to scaffold such success, both cognitively (e.g., Kapur (2011); Loibl and Rummel (2014)) and metacognitively (e.g., Holmes, Day, Park, Bonn, and Roll (2014); Roll et al. (2018)), have been largely unsuccessful.

Templates of exploratory or unsuccessful problem-solving, on the other hand, hold the view that acquisition of solution schema is not the solitary goal of learning through problem-solving (Schwartz & Martin, 2004; Kapur & Bielaczyc, 2012). It is equally important to develop the cognitive and socio-emotional prerequisites to prepare novice students to see like an expert. Therefore, one should provide opportunities that help students develop awareness and appreciation for what is known and not known. Instructional attempts that increase chances of failure during problem-solving have the advantage of stimulating student's initiative in gaining knowledge. However, students may not spontaneously come back to the right track if an incorrect problem representation is invoked and they continue to work on it. Evidence disfavoring SD learning in PS-I suggests that a lack of success when the problem-solving phase is implicitly scaffolded (e.g. Aleven et al. (2017); Roelle and Berthold (2016); Mazziotti, Rummel, and Deiglmayr (2016)) or left unscaffolded (e.g., Kapur

and Bielaczyc (2012)) does not harm learning. Providing no explicit cognitive or metacognitive support is imperative in view of giving students complete agency in solution generation. A consequential side-effect is that the likelihood of experiencing failures increases.

However, there is no PS-I research that looks at explicitly scaffolding problem-solving phase towards failure. This sets up the guiding question of whether and to what extent is success or failure during initial problem-solving necessary for learning from PS-I. How does increasing likelihood of students experiencing success or failure differentially prepare them to learn from the instruction at a deeper conceptual level? Are some ways of succeeding or failing more efficacious than others? To answer these questions, we design SD or failure-driven (FD) scaffolds for the problem-solving phase, as inputs into a classroom PS-I intervention. Evidence for impact of these scaffolds on learning from PF is discussed.

Method

Participants and Task Domain

We conducted a classroom PS-I intervention with N=221 students in an introductory data science course offered at a large public university in Switzerland. Based on data from a previous course iteration, two topics Spurious Correlation (SC) and Anscombe's Quartet (AQ) were chosen to develop learning materials. Problem-solving based on these topics had demonstrated different initial failure rates, and different levels of improvement after students were presented with clues pointing them to the correct answer (SC task, 40% $\rightarrow$ 23%; AQ task, 81% $\rightarrow$ 38%). The SC learning goal was to help students tease apart the difference between strong versus meaningful relationships among dataset variables. The AQ learning goal was to help students understand the complementary importance of numerical and graphical representations in reasoning with data. Students worked individually in an online problem-solving environment (Python Jupyter notebook) that was dynamically executable, and helped in offloading procedural or syntactical aspects of the computation required (for task details, see www.tinyurl.com/CogSci2019Tasks).

Experimental Design and Scaffolding Rationale

A mixed experimental design was followed. Scaffolding in initial problem-solving (SD, FD) was the between-subject variable, and problem-solving topic (SC, AQ) was the within-subject variable. Students were randomly assigned to experimental conditions, and ordering of problem-solving topics was counterbalanced within each condition. We had two conditions representative of SD scaffolding with varying degrees of specificity, and two conditions representative of FD scaffolding with varying levels of suboptimality. For all four conditions, the instruction phase was kept constant. Student solutions were compared and contrasted with the canonical one.

The rationale for the concrete design of scaffolding in our research was inspired by mechanisms of structuring and problematizing student work (Reiser, 2004). Structuring scaffolds reduce degrees of freedom to lower task complexity, help students maintain direction, and make problem-solving tractable. Problematizing scaffolds increase degrees of freedom to challenge student's current understanding, and highlight discrepancies between what they might generate and critical/canonical task features. We chose an initial set of structuring and problematizing scaffolds in line with keeping the generative characteristics of the problem-solving phase intact, as well as explicitly increasing success or failure likelihood as the intended design rationale.

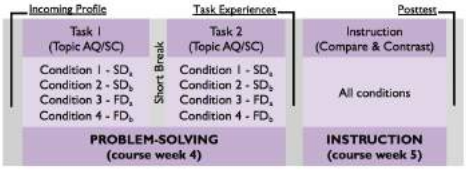


Figure 1: Experimental Design. SD_a, SD_b, FD_a, FD_b are two instantiations of success-driven (SD) and failure-driven (FD) scaffolding in the problem-solving phase respectively.

Structuring scaffolds included a combination of prompts, hints and bottom-out hints for different task topics. Prompts point students to the problem conditions that should likely remind them of the knowledge component's relevance. Very little information is divulged, thus encouraging students to do most of the thinking themselves. Our design of hints incorporates the idea of teaching students the knowledge component that is actually relevant in the current problem-solving context (what to do but not how). Finally, bottom-out hints tell students precise (and potentially optimal) ways of moving ahead in the problem-solving task. Such a scaffolding sequence mimics the behavior of expert human tutors, and is almost universally used in tutoring systems (VanLehn, 2011).

Problematizing scaffolds included asking students to explicitly generate suboptimal RSMs to facilitate problem-space exploration in a more comprehensive manner, rather than following an isolated solution path. In essence, students are led towards questionable decision-making by being asked to consider a subset of conceptual domain factors (that don't lead to the canonical solution), and reason with those partially-gained insights. No former PS-I work has looked directly into such "explicit" failure scaffolding. However, one could view the classic PF design as providing "implicit" opportunities for students to create suboptimal RSMs (Kapur & Bielaczyc, 2012). This is because rich problem design "inherently" affords multiple RSM generation, and targets concepts students haven't learnt yet. Work on preparatory benefits of vicarious failure activities before receiving instruction suggests the evaluation of suboptimal or failed RSMs generated by others as a significant predictor of learning (Kapur, 2014).

As a concrete example, when reasoning about the relationship between two variables, a prompt would give students general information about statistical dependence between variables, a hint would provide explanation of the exact phenomena under consideration (e.g., SC, AQ), and a bottom-out hint might ask for reasoning with a scatterplot (optimal graphical representation). Alternatively, reasoning with a 2-

Table 1: Examples of constructive reasoning coding applied for the analyses of posttest reasoning/code

Category	Sub-category	Sub-sub-category	Examples from data
Non-mathematical elaboration	Graphical	Complete	*Thinking for a good distribution that fit with this theory we can imagine a bar in the middle and nothing around. That means that all the people have the same degree of wealth. Looking at the plots we can already see that the distribution that seems what we have imagined is the normal distribution for the scenario A. We can also look at the standard deviation that confirm this reasoning*
		Not Complete	*Taking into account histogram with 50 bins, a better idea of distribution of wealth between citizens is given*
	Numerical	Complete	*By using a histplot, we see that for B there is no middle class, only rich and poor people =>not socialist.*
		Not Complete	*I add the values of each person and I divide by the number of person to find if the money is well distributed*
Mathematical elaboration	Graphical	Complete	*Datasets are almost identical specially in descriptive statistics but when we see plot of wealth distribution we can see that in B, there are more people with less wealth distribution specially after median and with similar reasoning we can say that as C is upper than B and A in most cases, it is the worst*
		Not Complete	*Linecharts show that C has the most wealth in the middle*
	Numerical	Complete	*Using the variance of each set, we can see that the values of dfA are much more centered around the mean (and thus a more egalitarian society). Followed by C then B*
		Not Complete	*comparing the median values of the different datasets*

D or 1-D histogram are examples of suboptimal RSMs. Here, information is lost because of binning and/or the lack of directly perceivable information about co-variation in the data.

Analytical Procedures

Due to dropout at various stages of the study (12%-57%), we applied standard multiple imputation (MI) procedures (n=5) to fill missing dataset values (Van Buuren, 2018). Discarding missing data may result in the complete cases being no longer representative of the target population, and consequently, estimates derived from them being subject to non-response bias. MI accounts for the process that created the missing data, and preserves uncertainty among relations in the data. Logistic regression and its variants (multinomial, ordered) were used for binary, nominal (>2 categories) and ordinal data respectively. Predictive mean matching was used to impute numeric data. Density plots of observed and imputed values were visually inspected for validity. Non-parametric statistics were used to see differences in ordinal posttest scores (e.g., Kruskal-Wallis tests, follow up Dunn tests). Multiple comparisons were adjusted using the Benjamini-Hochberg method. For non-significant results ($p > 0.05$), equivalence tests were performed to provide evidence for absence of a meaningful effect (Lakens, Scheel, & Isager, 2018). Here, the smallest effect size of interest was set within Cohen's d bounds of $\pm$ 0.2.

We also developed a coding scheme (Krippendorf's $\alpha >$ 0.7) for qualitative analyses of student's posttest reasoning and code, based on prior work (Chi, 2009; Kapur & Kinzer, 2009). First, we identified if reasoning was constructive (meaningful elaborations that went beyond what was presented). If yes, we identified if the elaborations were non-mathematical or mathematical. The former refers to elaborations that explain inferences leading up to the results, while the latter refers to elaborations that explicitly mention mathematical formalisms in words and/or in the code and base solution inferences on these formalisms. Next, for each kind of elaboration, we further checked if the elaborations comprised one or more graphical/numerical representation(s), meaning graphs, plots or other quantitative indices. Finally, we checked if these representation(s) were complete. Non-mathematical elaborations were coded as complete if all variables were set in relation to each other, and the result could be clearly derived from the elaboration. No information was missing and the connection between evidence and claim was fully explained using reasoning. Mathematical elaborations were coded as complete if all necessary methods in order to derive results were mentioned in words and/or presented in the code. Table 1 provides examples from the data.

Measures

Before the problem-solving phase, we collected student's prior knowledge using high school math scores as a proxy. No explicit pretest was conducted to prevent redundancy with the problem-solving phase. Based on prior literature on inter-individual factors that characterize heterogeneity in student's approach to FD and SD learning, we also included questionnaires assessing incoming profile variables like effort regulation (Pintrich et al., 1991), self-esteem (Jones, 1973), learning goal orientation (LGO) (Dweck, 1992) and attitude towards mistakes (ATM) (Leighton, Tang, & Guo, 2015). Effort regulation reflects a commitment to completing one's goals despite difficulties. High self-esteem triggers positive attributional style towards success and failure. An LGO disposition affects whether students view failures as learning opportunities. Finally, ATM, which includes the utility of making mistakes and induced affective reactions, enhances or impedes receptivity to failures. After the problem-solving phase, students answered task experience questionnaires, in line with PS-I preparatory mechanisms (Loibl et al., 2017).

These experiences included perceived awareness of knowledge gaps at the current moment (Glogger-Frey, Gaus, & Renkl, 2017), state curiosity about task actions and what they would like to know (Naylor, 1981), germane and extraneous cognitive load induced by problem-solving (Leppink, Paas, Van Gog, van Der Vleuten, & Van Merrienboer, 2014), and the experienced cognitive dissonance. Cognitive dissonance, defined as a state of discomfort associated with detection of conflicting concepts (Levin, Harriott, Paul, Zhang, & Adams, 2013), has not been studied in prior PS-I work because of lack of work on problematizing. Consistency of both incoming profile and task experience questionnaires was good for our dataset (McDonald's $\omega >$0.7). After the instruction phase, students solved an isomorphic and a non-isomorphic conceptual understanding posttest for each of the two task topics.

Results

Variable-centered Approach

We first performed variable-centered analyses to look at overall patterns of the impact of SD and FD preparatory activities on conceptual understanding in PF (figure 2).

Task topic SC For the SC topic, we found a significant omnibus effect of the experimental grouping on the

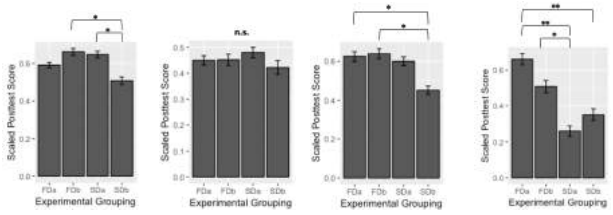


Figure 2: Scaled posttest scores with inferential error bars (L to R: SC non-isomorphic, SC isomorphic, AQ non-isomorphic, AQ isomorphic). Significant differences marked.

non-isomorphic conceptual understanding posttest ($\chi^2(3) = 11.73$, $p = 0.008$, Cohen's $d = 0.409$). The FD condition was better than the SD condition that offered the more-specific clue, in this case a hint describing the SC phenomena. However, the FD conditions were equivalent to the SD condition that offered the less-specific clue, in this case a prompt describing what statistical dependence among variables is. The two FD conditions here asked students to generate/reason with a correlation table and scatterplot matrix respectively, both of which reflect suboptimal numerical and graphical representations respectively. This is because they don't fully allow inferences on the nature of relationships (strength and/or meaningfulness) between dataset variables. No significant omnibus difference in scores on the isomorphic conceptual understanding posttest was observed across the four experimental conditions ($\chi^2(3) = 1.37$, $p = 0.712$, Cohen's $d = 0.174$). Equivalence testing suggested that the observed effect was neither statistically different from zero nor statistically equivalent, indicating insufficient data to draw conclusions.

Qualitative analysis suggested that for the non-isomorphic conceptual understanding posttest, the trend mirrored posttest scores. Students in the FD conditions had higher percentage of complete mathematical (32.1%, 38.7% >> 27.3%) and non-mathematical elaborations (44.7%, 56.7% >> 27.3%), compared to the SD condition with the more specific clue. Additionally, completeness of reasoning was almost identical between the FD condition and the SD condition with the less specific clue. However, for the isomorphic conceptual understanding posttest, student reasoning was often dominant in either complete mathematical or complete non-mathematical elaborations across the experimental conditions. The SD conditions had comparatively higher percentage of the former (38.9%, 45.8% >> 32.4%, 25%), while the FD conditions had comparatively higher percentage of the latter (33.3%, 48.8% >> 43.3%, 23.6%). This might be one reason why we saw no posttest score differences.

Task topic AQ For the AQ topic, we found significant omnibus effects of the experimental grouping on both the non-isomorphic ($\chi^2(3) = 10.84$, $p = 0.012$, Cohen's $d = 0.387$) and isomorphic ($\chi^2(3) = 20.16$, $p = 0.0001$, Cohen's $d = 0.586$) conceptual understanding posttest. Follow up pairwise comparisons suggested that scores for students in FD condition were greater than those in the SD condition with the more specific clue, in this case a bottom-out hint asking for scatterplot generation. However, the difference did not reach significance when comparing the FD condition and SD condition with the less specific clue, in this case a hint describing the AQ phenomena. The two FD conditions here asked students to generate/reason with a 2-D and 1-D histogram respectively. Both reflect suboptimal graphical representations.

We separated the coding of numerical and graphical representations to assess their independent usage in student reasoning. Qualitative analysis suggested that for the non-isomorphic conceptual understanding posttest involving graphical representations, students in the FD conditions had higher percentage of complete mathematical (72.2%, 68.7% >> 33.3%, 40%) and non-mathematical elaborations (44.4%, 37.5% >> 26.6%, 40%), compared to the SD conditions. This also held true for complete mathematical (27.7%, 37.5% >> 20%, 0%) and non-mathematical elaborations (27.7%, 31.2% >> 13.3%, 0%) involving numerical representations. For the isomorphic conceptual understanding posttest, a similar trend held for elaborations involving graphical representations. We did not see clear trends in qualitative differences in student reasoning for elaborations involving numerical representations, the less straightforward (and dominant) approach for this isomorphic question. Taken together, despite no posttest score differences between students who received FD scaffolds and the less-specific SD scaffold, there were salient differences in reasoning quality.

Person-centered Approach

We performed complementary person-centered analyses to go beyond an average FD or SD learning pattern (figure 3). The rationale here was to factor in the interactions among incoming student characteristics, in order to understand the impact of this heterogeneity on learning. We used latent profile analysis to first cluster students based on incoming profile variables like prior knowledge, effort regulation, learning goal orientation, self-esteem and attitude towards mistakes. This approach provides an elegant way to discover subgroups by "simultaneously" considering interactions among "more than one" incoming cognitive and motivational student characteristic. Non-parametric multivariate finite mixture models were used (Hickendorff, Edelsbrunner, McMullen, Schneider, & Trezise, 2017). A two-cluster solution (figure 4) reflected parsimonious fit to the data (based on model fit ($loglik = 530.41$), mixture distributions and visual inspection of mixture density plots when fitting more than two clusters).

Cluster assignments for students into these homogeneous subgroups were based on posterior probability distributions. These cluster assignments allowed us to then use this information for studying interaction effects (reported below). Statistically, we found one of these clusters (henceforth, Cluster_{high}) to have significantly higher scores on all of these incoming characteristics, compared to the other cluster (Cluster_{low}). Cluster_{low} reported higher extraneous cognitive load than Cluster_{high} ($W = 7319.5$, $p = 0.001$) after problem-solving. All other task experiences were statistically similar.

Task topic SC/AQ Not surprisingly, we did find that students in Cluster_{high} scored significantly higher than those in

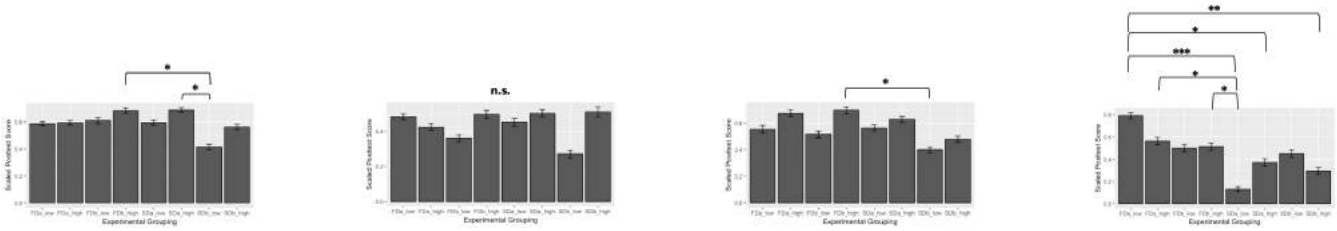


Figure 3: Scaled posttest scores with inferential error bars (L to R: SC non-isomorphic, SC isomorphic, AQ non-isomorphic, AQ isomorphic). Significant differences marked. *Low* and *High* represent students from Cluster$_{low/high}$ within a condition.

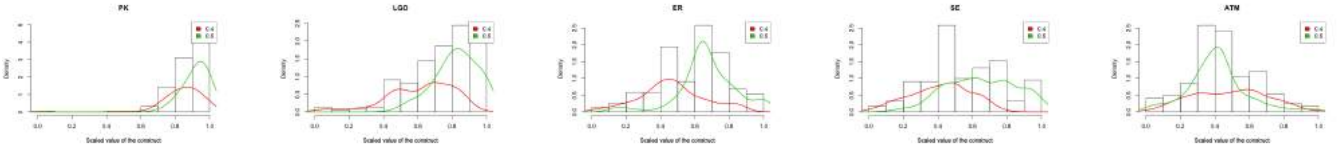


Figure 4: Mixture density distributions when clustering students based on incoming cognitive and motivational characteristics (L to R: Prior knowledge, Learning goal orientation, Effort regulation, Self-esteem, Attitude towards mistakes)

Cluster$_{low}$ on the non-isomorphic conceptual understanding posttest. This trend held for both SC ($W = 4885.5$, $p = 0.04$) and AQ task topics ($W = 4811.5$, $p = 0.02$). On the other hand, both Cluster$_{high}$ and Cluster$_{low}$ performed equally well on posttests scores for the isomorphic conceptual understanding question. Equivalence testing results were inconclusive.

Interaction We finally looked at impact of the interaction between experimental grouping and incoming student profile on posttest. Results suggested that the omnibus trend for difference in non-isomorphic conceptual understanding posttest still held for the SC topic ($\chi^2(7) = 17.16$, $p = 0.016$, Cohen's $d = 0.448$). Descriptively, students in the FD sub-groupings outperformed those in the SD sub-groupings. As before, the omnibus effect was still not significant for the isomorphic conceptual understanding posttest ($\chi^2(7) = 11.5$, $p = 0.118$, Cohen's $d = 0.294$). Equivalence testing showed 24/28 pairwise comparisons to be inconclusive.

For the more difficult topic AQ, again, as before, exposure to failure-driven scaffolds benefited students on both the non-isomorphic ($\chi^2(7) = 16.91$, $p = 0.017$, Cohen's $d = 0.442$) and isomorphic ($\chi^2(7) = 27.16$, $p = 0.0003$, Cohen's $d = 0.647$) conceptual understanding posttest. Descriptive trends for students in FD sub-groupings scoring higher than their counterparts in the SD sub-groupings still held. This also suggests that perhaps task difficulty and the extent to which student reasoning requires manipulation and integration of multiple representations, might be an important factor when looking at the relative efficacy of FD and SD scaffolds.

Underlying Mechanisms

We computed partial correlations between student's task experiences during the problem-solving phase and their posttest scores, controlling for experimental grouping (SD, FD) and incoming student profile (Cluster$_{high/low}$). For the more difficult topic (AQ), we saw positive associations of both isomorphic and non-isomorphic posttest scores with awareness of knowledge gaps ($\rho = 0.112^{+}$, 0.172*) and germane cognitive load ($\rho = 0.114^{+}$, 0.120^{+}). The correlation between these task experiences and posttest scores was not significant for the easier topic (SC). Experiencing higher state curiosity ($\rho = 0.158$*, 0.184**) and cognitive dissonance ($\rho =$ 0.193**, 0.187**) was positively associated with only with non-isomorphic posttests, however for both SC and AQ topics. Finally, experiencing greater extraneous cognitive load was negatively associated with posttest scores for both SC ($\rho =$ -0.243**, -0.236**) and AQ ($\rho =$ -0.185**, n.s.) topics.

Manipulation Check and Design Implications

Students in every experimental condition had the opportunity to make two solution attempts (prior/post exposure to the scaffold) during the problem-solving phase. This design allowed us to assess the percentage of students who improved/degraded their solution across these two time points within the initial problem-solving. We computed a summary index S (ranging from -100 to 100) for each condition and task topic, by subtracting (i)ΔD, the percentage of students who degraded (got the right answer pre-scaffold, but wrong answer post-scaffold), from, (ii)ΔI, the percentage that improved (got the wrong answer pre-scaffold, but right answer post-scaffold). For the two FD conditions, we found S to be highly negative ($\Delta D > \Delta I$) for the more difficult task topic AQ (-72%, -47%), suggesting that the problematizing scaffold indeed pushed students towards explicit failure. For the easier task topic SC (-51%, -34%), S was still negative but comparatively lower in absolute terms.

Interestingly, for the two SD conditions, S was not positive or ΔI was $\not>$ than ΔD (as one might intuitively expect). Overall, despite S being lower in absolute terms compared to the FD conditions, it was still negative for both the AQ (-52%, -54%) and SC (-40%, -6%) task topics. This suggests that although explicit structuring prior to instruction led to greater net solution accuracy (compared to explicit problematizing), it was still not enough to push majority of student solutions to match the canonical answer. Taken together, these analyses show that students may not necessarily be prepared to receive explicit structuring during initial exploration, especially for difficult topics. It also opens up questions about re-calibrating the specificity of structuring scaffolds so that $\Delta I > \Delta D$.

Discussion and Conclusion

Table 2: posttest differences across experimental grouping

	Non-isomorphic conceptual understanding posttest	Isomorphic conceptual understanding posttest
Topic SC (Variable-centered)	$\chi^2(3) = 11.73$, $p = 0.008$ Cohen's $d = 0.409$	$\chi^2(3) = 1.37$, $p = 0.712$ Cohen's $d = 0.174$
Topic SC (Person-centered)	$\chi^2(7) = 17.16$, $p = 0.016$ Cohen's $d = 0.448$	$\chi^2(7) = 11.5$, $p = 0.118$ Cohen's $d = 0.294$
Topic AQ (Variable-centered)	$\chi^2(3) = 10.84$, $p = 0.012$ Cohen's $d = 0.387$	$\chi^2(3) = 20.16$, $p = 0.0001$ Cohen's $d = 0.586$
Topic AQ (Person-centered)	$\chi^2(7) = 16.91$, $p = 0.017$ Cohen's $d = 0.442$	$\chi^2(7) = 27.16$, $p = 0.0003$ Cohen's $d = 0.647$

To summarize, our results indicate the efficacy of FD over SD preparatory activities on student's conceptual understanding. We go beyond prior PS-I work by performing stringent comparisons between explicit ways of pushing students towards success and failure in problem-solving prior to instruction, and investigating their impact on learning. Overall, we found a significant main effect for experimental grouping on the non-isomorphic conceptual understanding posttest, with the FD conditions outperforming the SD condition with the more-specific clue, but not the SD condition with the less-specific clue. Posttest score similarity between the latter comparison indicates that FD and SD approaches might potentially offer two distinct but effective paths to learning. Nudging students to make them realize by themselves the extent to which their activated knowledge is (ir)relevant for solving the problem (we can have both SD and FD ways towards this end), is better than directing their activation of relevant prior knowledge (via a highly specific SD scaffold).

However, we also found that a comparatively higher percentage of students who received FD scaffolds demonstrated reasoning with complete mathematical or non-mathematical elaborations, indicating better quality of reasoning than students in the SD conditions. This result supports the idea that focusing on the pragmatic goal of performing the correct procedure (in presence of SD scaffolding) without appropriately articulating understanding (non-reflective work) can lead to fragile conceptual gains (Jonassen, 2010). We also found a significant main effect for the incoming student profile on the non-isomorphic conceptual understanding posttest, with students having high self-reported scores significantly performing better. There was no evidence for an interaction effect. Exposure to FD scaffolds had a greater impact on posttest scores for the more difficult topic (AQ). Finally, we found mechanistic task experiences to be positively associated with posttest scores (stronger associations for AQ task topic and for non-isomorphic posttests), controlling for experimental grouping and incoming profile.

What might explain the superiority of problematizing scaffolds over structuring scaffolds in the PS-I design? Although scaffolding for success might push for speed/accuracy to facilitate fluency in knowledge application for one form of independent performance (Schwartz, Sears, & Chang, 2007), both posttest scores and qualitative analysis of reasoning suggest that it does not guarantee improved conceptual understanding. Correct performance of a procedure scaffolded via structuring might stem from the lack of awareness and appreciation of long-term sub-optimality of a solution that works reasonably well in the short-term (Schwartz, Chase, & Bransford, 2012). The resulting quick/easy success may be insufficiently disruptive to challenge existing thought processes, and induce inattention when learning from instruction.

Existing meta-analysis of PS-I literature (Loibl et al., 2017) also suggests that students need to be made aware of the limitations to their knowledge (knowledge gaps). Further, we must instill in them a strong desire to know more about the canonical solution to fill these knowledge gaps. Finally, the learning design needs to facilitate understanding of which solutions don't work and why. In line with these vital pre-instructional goals, the suboptimal RSM generation strategy triggers "effortful activation" of prior knowledge conceptually relevant to the targeted learning concept.

By exposing students to additional exploration of the problem-space structure that doesn't necessarily lead to the canonical solution, suboptimal RSM generation provides support for meaningful variation in reasoning (Soderstrom & Bjork, 2015), which aids in improved conceptual understanding. Further, the uncertainty induced about consequences of partially-gained insights during solution revision is likely to trigger momentary curiosity driven by student's problem-solving experiences. One's own failed attempt is also likely to better prepare students for acquisition of negative knowledge regarding applicability conditions of solution strategies during instruction. Finally, at a methodological level, we see an improvement in effect sizes compared to a traditional variable-centered approach for both task topics (table 2). Complementary person-centered analyses provide a more accurate assessment of the impact of our PF intervention, since they factor in the differential benefits arising due to individual differences in SD and FD learning.

The scaffolding implemented in this work can be embedded into metacognitive tutors (Joyner & Goel, 2015) that deploy computer agents to imitate functional roles of teachers - "guides" to offer structuring, and "critiques" to problematize exploration. Limitations of this work stem primarily from the classroom time constraints. This was reflected, for e.g., in choice of datasets we used. For future work, we will design rich(er) datasets (that allow greater scope of inferences). The allocated time budget also led us to design one-step SD or FD scaffolds, and collect single task experience questionnaire after students finished solving problems on both topics (SC, AQ). Finally, optional university attendance resulted in considerable student dropout over the two study weeks, despite our efforts to mitigate this threat via participation reminders.

References

Aleven, V., Connolly, H., Popescu, O., Marks, J., Lamnina, M., & Chase, C. (2017). An adaptive coach for invention activities. In *International conference on artificial intelligence in education* (pp. 3–14).

Chase, C. C., & Klahr, D. (2017). Invention versus direct instruction: for some content, itsa tie. *Journal of Science Education and Technology*, *26*(6), 582–596.

Chi, M. T. (2009). Active-constructive-interactive: A conceptual framework for differentiating learning activities. *Topics in cognitive science*, *1*(1), 73–105.

Chin, D. B., Chi, M., & Schwartz, D. L. (2016). A comparison of two methods of active learning in physics: inventing a general solution versus compare and contrast. *Instructional Science*, *44*(2), 177–195.

Dweck, C. S. (1992). Article commentary: The study of goals in psychology. *Psychological Science*, *3*(3), 165–167.

Glogger-Frey, I., Gaus, K., & Renkl, A. (2017). Learning from direct instruction: Best prepared by several self-regulated or guided invention activities? *Learning and Instruction*, *51*, 26–35.

Hickendorff, M., Edelsbrunner, P. A., McMullen, J., Schneider, M., & Trezise, K. (2017). Informative tools for characterizing individual differences in learning: Latent class, latent profile, and latent transition analysis. *Learning and Individual Differences*.

Holmes, N. G., Day, J., Park, A. H., Bonn, D., & Roll, I. (2014). Making the failure more productive: scaffolding the invention process to improve inquiry behaviors and outcomes in invention activities. *Instructional Science*, *42*(4).

Jonassen, D. H. (2010). *Learning to solve problems: A handbook for designing problem-solving learning environments*. Routledge.

Jones, S. C. (1973). Self-and interpersonal evaluations: esteem theories versus consistency theories. *Psychological bulletin*, *79*(3), 185.

Joyner, D. A., & Goel, A. K. (2015). Organizing metacognitive tutoring around functional roles of teachers. In *Cogsci.*

Kapur, M. (2011). A further study of productive failure in mathematical problem solving: Unpacking the design components. *Instructional Science*, *39*(4), 561–579.

Kapur, M. (2014). Comparing learning from productive failure and vicarious failure. *Journal of the Learning Sciences*, *23*(4), 651–677.

Kapur, M., & Bielaczyc, K. (2012). Designing for productive failure. *Journal of the Learning Sciences*, *21*(1), 45–83.

Kapur, M., & Kinzer, C. K. (2009). Productive failure in cscl groups. *International Journal of Computer-Supported Collaborative Learning*, *4*(1), 21–46.

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, 2515245918770963.

Leighton, J. P., Tang, W., & Guo, Q. (2015). *Developing and validating the attitudes towards mistakes inventory (atmi): A self-report measure.*

Leppink, J., Paas, F., Van Gog, T., van Der Vleuten, C. P., & Van Merrienboer, J. J. (2014). Effects of pairs of problems and examples on task performance and different types of cognitive load. *Learning and Instruction*, *30*, 32–42.

Levin, D. T., Harriott, C., Paul, N. A., Zhang, T., & Adams, J. A. (2013). Cognitive dissonance as a measure of reactions to human-robot interaction. *Journal of Human-Robot Interaction*, *2*(3), 3–17.

Loibl, K., Roll, I., & Rummel, N. (2017). Towards a theory of when and how problem solving followed by instruction supports learning. *Educational Psychology Review*, *29*(4).

Loibl, K., & Rummel, N. (2014). The impact of guidance during problem-solving prior to instruction on students inventions and learning outcomes. *Instructional Science*, *42*(3).

Mazziotti, C., Rummel, N., & Deiglmayr, A. (2016). Comparing students solutions when learning collaboratively or individually within productive failure. Singapore: International Society of the Learning Sciences.

Naylor, F. D. (1981). A state-trait curiosity inventory. *Australian Psychologist*, *16*(2), 172–183.

Pintrich, P. R., et al. (1991). A manual for the use of the motivated strategies for learning questionnaire (mslq).

Reiser, B. J. (2004). Scaffolding complex learning: The mechanisms of structuring and problematizing student work. *The Journal of the Learning sciences*, *13*(3).

Roelle, J., & Berthold, K. (2016). Effects of comparing contrasting cases and inventing on learning from subsequent instructional explanations. *Instructional Science*, *44*(2).

Roll, I., Butler, D., Yee, N., Welsh, A., Perez, S., Briseno, A., ... Bonn, D. (2018). Understanding the impact of guiding inquiry: The relationship between directive support, student attributes, and transfer of knowledge, attitudes, and behaviours in inquiry learning. *Instructional Science*.

Schalk, L., Schumacher, R., Barth, A., & Stern, E. (2017). When problem-solving followed by instruction is superior to the traditional tell-and-practice sequence. *Journal of Educational Psychology*.

Schwartz, D. L., Chase, C. C., & Bransford, J. D. (2012). Resisting overzealous transfer: Coordinating previously successful routines with needs for new learning. *Educational Psychologist*, *47*(3), 204–214.

Schwartz, D. L., Chase, C. C., Oppezzo, M. A., & Chin, D. B. (2011). Practicing versus inventing with contrasting cases: The effects of telling first on learning and transfer. *Journal of Educational Psychology*, *103*(4), 759.

Schwartz, D. L., & Martin, T. (2004). Inventing to prepare for future learning: The hidden efficiency of encouraging original student production in statistics instruction. *Cognition and Instruction*, *22*(2), 129–184.

Schwartz, D. L., Sears, D., & Chang, J. (2007). Reconsidering prior knowledge. *Thinking with data*, 319–344.

Sinha, T., & Kapur, M. (2019). When productive failure fails. In *Proceedings of the annual meeting of the cognitive science society.*

Soderstrom, N. C., & Bjork, R. A. (2015). Learning versus performance: An integrative review. *Perspectives on Psychological Science*, *10*(2), 176–199.

Van Buuren, S. (2018). *Flexible imputation of missing data*. Chapman and Hall/CRC.

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, *46*(4), 197–221.

When Productive Failure Fails

Tanmay Sinha, Manu Kapur
ETH Zürich, Switzerland

Abstract

Productive Failure (PF) is a learning design that intentionally designs for and uses failure in preparatory problem-solving for learning. Over the past decade, there has been growing evidence supporting the effectiveness of learning from PF. The purpose of this paper, however, is to critically examine evidence for when PF fails. We analyze 95 experimental comparisons from 57 studies reported in 44 articles into the extent to which they conform to PF design criteria. These criteria, as outlined in the original PF work, span the problem-solving activity, the participation structures, and the social surround. Results suggest lack of design fidelity as a critical factor for when PF fails to outperform alternative instructional approaches on conceptual knowledge and/or transfer.

Keywords: Direct Instruction; Productive Failure; Scaffolding

Introduction

The past decade has seen a growing body of evidence for the efficacy of Productive Failure (PF) for developing conceptual knowledge and transfer (for a review, see Kapur (2016); Loibl, Roll, and Rummel (2017)). PF comprises an initial problem-solving phase where learners generate and explore representations and solution methods (RSMs) to complex problems based on concepts they have not formally learnt yet, followed by an instruction phase where an expert or a teacher builds upon student-generated solutions to teach them the targeted concepts. According to PF, generating solutions to novel problems prior to instruction can help students learn better from the instruction, even if students fail to generate the correct solution in the problem-solving phase (Kapur, 2016). Thus conceived, PF can be seen as a subset of a general class of designs where problem-solving precedes instruction (or PS-I). It must be noted that not all PS-I designs are PF, but only those in which students generate multiple solutions but fail to generate the correct one.

In experimental comparisons, PF is typically compared with a design where students are initially given instruction on the targeted concepts, followed by problem-solving practice. Loibl et al. (2017) referred to this design as an Instruction-followed-by-Problem-Solving (I-PS) design. Findings in support of PF suggest that both PF and I-PS are similar in the development of procedural knowledge, but PF significantly outperforms I-PS in conceptual understanding and transfer (Kapur, 2016). Evidence comes not only from quasi-experimental studies conducted in the real ecologies of classrooms (e.g., Kapur (2012); Kapur and Toh (2013); Schwartz and Bransford (1998); Schwartz and Martin (2004)), but also from controlled experimental studies (e.g., M. S. DeCaro and Rittle-Johnson (2012); Kapur (2014); Loibl and Rummel (2014a); Roll, Aleven, and Koedinger (2011); Schmidt and Bjork (1992); Schwartz, Chase, Oppezzo, and Chin (2011)).

Although we now have substantial empirical evidence for when PF succeeds (Loibl et al., 2017), we argue it is equally important, if not more, to examine evidence when PF fails and delineate boundary conditions for how, when and why PF works. By success of PF, here we mean experimental comparisons in which PF significantly outperforms alternative instructional approaches (usually instruction followed by problem-solving (I-PS), but also scaffolded problem-solving followed by instruction (+PS-I), or a different preparatory activity followed by instruction (!PS-I)[1]). By failure of PF, here we mean experimental comparisons between PF and I-PS, PF and +PS-I, PF and !PS-I, where I-PS, +PS-I, !PS-I conditions significantly outperform PF on measures of either conceptual understanding or transfer.

At the same time, we also examine experimental comparisons with null results, that is, when there was no significant difference between PF and these three alternate experimental conditions. Although attribution of null effects to causal factors is not always straightforward, examining null effects may nevertheless shed light on the critical factors that confluence efficacy of PF. Bridging the gap between instructional decision-making and the science of learning from failure necessitates prescribing conditions under which positive or negative failure effects emerge and how to foster them.

Search Criteria

Our search process and the criteria for including and excluding comparisons for this analysis included articles in the Google Scholar databases that (i) cited either of the two seminal PF articles (Kapur, 2008; Kapur & Bielaczyc, 2012), and those that cited other key follow-up PF articles (Kapur, 2014, 2015, 2016), and (ii) reported experimental or quasi-experimental comparison between PF and I-PS, or between PF and +PS-I, or between PF and !PS-I; and (iii) assessed conceptual knowledge and/or transfer. Criteria i resulted in close to 700 articles as of 29th June 2018. Of these, 44 articles met criteria ii and iii. These 44 articles reported 57 studies and comprised 95 experimental comparisons[2]. Table 1 presents a breakdown of their demographic characteristics, with majority of the studies spanning Europe, North America and Asia, and covering mathematics concepts for 6th-10th graders. We also see evidence for PS-I work gradually expanding to different student populations at the post-graduate and professional levels within other STEM domains like physics, chemistry, biology, as well as within non-STEM domains like psychology and medicine.

Using a two-phase workflow, we now report key findings synthesized from these experimental comparisons. The first phase comprised a fidelity check for examining conformity of

[1] Exclamation (!) denotes [(NOT) Problem-solving], e.g., [Reading worked examples], [Problem posing], [Explanation generation]

[2] https://tinyurl.com/WhenPFfails

Table 1: Demographic characteristics of articles included in the review (Number of comparisons = 95)

	Variable of Interest	# of Comparisons (%)
1. Geographical distribution	Europe (Germany, Switzerland, UK)	30 (31.6%)
	North America (USA, Canada)	31 (32.6%)
	Asia (Singapore, Taiwan, India, Hong Kong, Saudi Arabia)	27 (28.4%)
	Australia	7 (7.4%)
2. Learner grade	6th - 10th graders	59 (62.1%)
	2nd - 5th graders	17 (17.9%)
	Undergraduates	16 (16.9%)
	Others (Postgraduates, Professionals)	3 (3.1%)
3. Targeted concept	Math (equivalence, geometry, fractions, variance, linear functions, central tendencies, least squares fitting, weighted averages, z-scores, statistics process control)	63 (66.3%)
	Physics (average speed, density, collision, electricity, mechanics)	16 (17%)
	Medical (dental hygiene, dental surgery)	4 (4.2%)
	Chemistry (solutions, atomic structure)	3 (3.1%)
	Psychology (memory)	2 (2.1%)
	Domain general skill (control of variables strategy)	2 (2.1%)
	Biology (genetics)	2 (2.1%)

PS-I implementations to PF design criteria (for detailed criteria definition, refer Kapur and Bielaczyc (2012)). A detailed breakdown of these PF fidelity check criteria for the current analyses is shown in table 2. Looking vertically across the table (from comparisons with positive results for PF to those with null and negative results for PF), the decrease in fidelity along many of the PF design criteria is striking. This suggests that our evidence base comprises a mixture of the original PF design as well as its low-fidelity versions. In the second phase, we explored additional reasons that could not be convincingly explained by fidelity check parameters alone. The rest of the article focuses on 44 of these 73 comparisons, 54.6% of which had significant negative ($p<0.05$) or null results ($p>0.05$) with I-PS as the comparison condition, 25% of which had negative or null results with +PS-I as the comparison condition, and remaining 20.4% of which had negative or null results with !PS-I as the comparison condition.

Negative Results for PF (compared to I-PS)

PF fidelity check revealed that most of the 7 experimental comparisons in this cluster (Loehr, Fyfe, & Rittle-Johnson, 2014; D. A. DeCaro, DeCaro, & Rittle-Johnson, 2015; Schalk, Schumacher, Barth, & Stern, 2017; Marei, Donkers, Al-Eraky, & van Merrienboer, 2017) considered affective draw of the problem (5/7), and provided evidence for multiple RSM generation during the initial problem-solving phase (5/7). However, what is striking is that in none of the comparisons did follow-up instruction build on failed or suboptimal learner generated solutions, or include group work as the participation structure. Since such consolidation and knowledge assembly is often a key component of PF (Kapur & Bielaczyc, 2012), we would not necessarily expect these low fidelity PF implementations to be better than I-PS comparison conditions. Other salient factors influencing results from these comparisons are described below.

First, learners with high performance orientation, who primarily seek to demonstrate ability, may view challenging task situations as a threat to this goal and withdraw their effort. Such learners are less likely than those with a learning-goal orientation disposition to focus on viewing failures as opportunities to learn, processing negative feedback as ways to improve performance, and experiencing positive emotions following failure (Dweck, 1992; Tulis & Ainley, 2011). Thus, there is no reason to believe that challenging exploratory problem-solving phase of PF might benefit them more so than an instruction-first approach (D. A. DeCaro et al., 2015).

Second, the presence of additional problem-solving practice following the PS-I routine allows learners to use the taught information immediately and integrate it with prior knowledge. Thus, PF can be expected to fail when the overall learning design lacks this practice activity, or, when the overall learning design includes this activity, but such an activity invokes application of procedural knowledge to solve problems and correct errors to a greater extent, rather than influencing processing and development of conceptual knowledge. Empirical evidence suggests that these negative effects were mitigated to some extent in a follow-up study (although not fully) when learners self-checked initial solutions immediately after instruction (Loehr et al., 2014).

Third, implementation-level details of preparatory problem-solving activities are important. PF can be expected to fail when the problem-solving phase comprises too loosely anchored instruction (e.g., an idealized contrasting case that represents a principle in an abstract and generic fashion, followed by self-explanation prompts). PF can, however also fail with relatively more anchored instruction (e.g., a grounded contrasting case that situates a principle in a specific context but also potentially contains (irr)relevant details, followed by self-explanation prompts).

In Schalk et al. (2017) for instance, idealized contrasting cases were operationalized by providing no labels for the axes of coordinate systems when introducing the concept of linear slopes in mathematics, while grounded cases had axes labeled with meaningful concepts (e.g., filling level in a rain barrel on the y-axis, and time in hours on the x-axis). Schalk et al. (2017) conjecture that although self-explanation prompts can help learners to abstract from the context provided in the grounded cases (Chi, De Leeuw, Chiu, & LaVancher, 1994), contextual details from the learning materials are likely to be preserved in the encoded knowledge representation (De Bock, Deprez, Van Dooren, Roelens, & Verschaffel, 2011). This can hamper transfer.

The detrimental effect of grounded cases might exist even if self-explanation prompts in the problem-solving phase are

Table 2: PF fidelity check criteria for the PS-I design, with 60 I-PS and 13 +PS-I and 22 !PS-I experimental comparisons. Results separated by positive, null and negative effects for PF. Table values show number (percentage) of comparisons. We describe an analyses of experimental comparisons with null and negative effects for PF in this paper.

Comparison condition	Effects for PF	Problems affording multiple RSMs	Evidence for multiple RSM generation	Affective draw of the problem	Group work as the participation structure	Building on learner solutions in Instruction
1. I-PS	Positive	36 (100%)	29 (80.5%)	32 (88.9%)	25 (69.4%)	23 (63.8%)
	Null	17 (100%)	8 (47%)	13 (76.4%)	9 (52.9%)	6 (35.3%)
	Negative	7 (100%)	5 (71.4%)	5 (71.4%)	0 (0%)	0 (0%)
2. +PS-I	Positive	2 (100%)	2 (100%)	2 (100%)	2 (100%)	2 (100%)
	Null	7 (100%)	7 (100%)	7 (100%)	3 (42.8%)	1 (14.2%)
	Negative	4 (100%)	2 (50%)	4 (100%)	2 (50%)	2 (50%)
3. !PS-I	Positive	11 (100%)	6 (54.5%)	11 (100%)	5 (45.4%)	4 (36.3%)
	Null	5 (71.4%)	2 (28.5%)	5 (71.4%)	2 (28.5%)	2 (28.5%)
	Negative	4 (100%)	3 (75%)	4 (100%)	1 (25%)	2 (50%)

replaced by explicit invention prompts. From an instructivist point of view, the need to come up with unifying functional relations already makes the invention prompt inherently challenging. Addition of grounded cases can further overburden learners with unnecessary details. Experiencing increased extraneous load can negatively affect invention quality and subsequently transfer, placing learners in the PF condition at a disadvantage. More work is needed, however to understand relative efficacy of concrete or abstract preparatory activities.

Null Results for PF (compared to I-PS)

PF fidelity check revealed that most of the 17 experimental comparisons in this cluster (Schwartz & Martin, 2004; Belenky & Nokes-Malach, 2012; Matlen & Klahr, 2013; Loehr et al., 2014; Loibl & Rummel, 2014b; Fyfe, DeCaro, & Rittle-Johnson, 2014; D. A. DeCaro et al., 2015; Hsu, Kalyuga, & Sweller, 2015; Mazziotti, Loibl, & Rummel, 2015; Chase & Klahr, 2017; Tam, 2017; Marei et al., 2017; Newman & DeCaro, 2018) considered affective draw of the problem (13/17), about half of them provided evidence for multiple RSM generation during the initial problem-solving phase (8/17), while about one third of the comparisons included follow-up instruction building on learner generated solutions (6/17). This suggests moderate conformity to the PF design criteria, and calls for a nuanced understanding of the results.

While young learners (e.g., 2nd - 5th graders) may have insufficient prior knowledge about cognitive and metacognitive learning strategies to generate RSMs on their own (Mazziotti et al., 2015), adult learners with very high incoming mastery-approach orientation are likely to transfer regardless of the type of instruction. This is because the inventing activity in and of itself provides motivational impetus to learn the targeted concepts (Belenky & Nokes-Malach, 2012). These null results suggest that learners with such incoming cognitive or motivational profiles may not necessarily benefit from PF.

The nature of problem-solving task is an important factor as well. Tasks with high element interactivity (Sweller, 1988) have high expected error rate. As Loibl and Leuders (2018) suggest, revision of mental models following instruction for such tasks is contingent on whether or not learners spontaneously elaborate on erroneous solutions generated during initial problem-solving. As long as learners are prompted to explicitly compare and contrast their suboptimal solutions with the canonical solutions, they are likely to integrate negative knowledge in their repertoire of future problem-solving strategies. Consequently, there is no reason to suppose that such learners will benefit from problem-solving first (Hsu et al., 2015; Loibl & Leuders, 2018). While the sole impact of solution generation on the efficacy of PF is not yet clear, what is clearer is that the form of instruction matters (Loibl & Rummel, 2014b). Without instruction that compares and contrasts learner solutions with a canonical solution, PF can be expected to fail. Further, impact of the ordering of such instruction (before or after problem-solving) is less clear.

PF can also be expected to fail when the task provides no explicit feedback regarding what problem-solving actions are actually failures. Consequently, learners might not be in a position to use their awareness of knowledge gaps to consolidate information during the instruction phase (Matlen & Klahr, 2013). Finally, as Chase and Klahr (2017) suggest, when learning domain-general skills, the problem-solving phase in and of itself is less likely to provide implicit feedback about what goals to adopt during the inquiry process (that strongly impacts learning). For instance, learner's goals in pursuing inquiry might be scientific (finding out whether a variable impacts an outcome) or engineering-oriented (guarantying some desired outcome). In such situations, aligning learner's goals to a scientific one takes precedence over the relative ordering of the instruction phase in which this might happen.

Shifting focus to learner solutions, a key recurring factor for failure of PF is lack of evidence for learning to learn, i.e., spontaneous internalization of skills needed for application of domain-knowledge in novel situations. Gaining knowledge of how to perform a correct procedure after the consolidation phase of PF does not necessarily mean gaining high depth of understanding of the domain principle (Vollmeyer, Burns, & Holyoak, 1996; Schwartz, Chase, & Bransford, 2012; Soderstrom & Bjork, 2015). Self-regulated reasoning strategies (e.g., solution evaluation, unprompted self-explanation) require sufficient practice opportunities to get internalized (Schwartz & Martin, 2004; Tam, 2017). Finally, with respect to the overall learning design, PF is expected to fail or produce comparable effects to an I-PS design when the pretest targets concepts similar to the invention activity. Engaging learners in important exploratory learning processes such as prior knowledge activation, attention to knowledge gaps etc create redundancy with initial problem-solving phase of the PS-I setting, thus diluting effects (Newman & DeCaro, 2018).

Negative/Null Results for PF (compared to +PS-I)

PF fidelity check revealed that all the 11 experimental comparisons in this cluster (Kapur & Bielczyc, 2011; Holmes, Day, Park, Bonn, & Roll, 2014; Kim, Pathak, Jacobson, Zhang, & Gobert, 2015; Roelle & Berthold, 2016; Kuo & Wieman, 2016; Loibl & Leuders, 2018) considered affective draw of the problem, most of them provided evidence for multiple RSM generation during the initial problem-solving phase (9/11). However, about half of the comparisons used group work as the participation structure (5/11), and even fewer included follow-up instruction building on learner generated solutions (3/11). This suggests moderate conformity to the PF design criteria.

Evidence suggests that the extent to which activated prior knowledge is conceptually related to the targeted learning concept affects whether the failure resulting from it is productive. This can impact whether and when PF outperforms a scaffolded PS-I condition. If learners are scaffolded to detect high number of relevant similarities and differences in the contrasting cases during an initial problem-solving phase, this can lead them to focused elaboration/explanation regarding deep features of the problem after the instruction phase, resulting in improved conceptual understanding (Roelle & Berthold, 2016). Goal specificity research also suggests that the benefits of preparatory activities with low to medium goal specificity (as in the PS-I design) are contingent on affording opportunities for relevant prior knowledge activation, e.g., by guiding learners towards strategies that facilitate reasoning with the deep problem structure (Vollmeyer et al., 1996), or, by illustrating desirable sub-goals along a solution path that requires learners to focus on relevant task relationships (Miller, Lehman, & Koedinger, 1999). We describe such forms of scaffolded problem-solving in more detail below.

In the study by (Vollmeyer et al., 1996) for instance, explicit instruction in a systematic strategy (varying a single factor while holding other factors constant at zero) during the initial exploratory task fostered acquisition of the casual structure of a biological system. This was based on the premise that despite the presence of a nonspecific goal during the exploratory task, learners might not spontaneously make full use of effective rule-induction strategies. In the study by (Miller et al., 1999) where learners had to work in an exploratory micro-world to understand interactions of electrically charged particles, specializing the learning goal assisted learners in activating relevant prior knowledge. Illustrating a particular path and asking learners to arrange charged particles so that the moving charges would follow the illustrated path as closely as possible achieved this.

Richland and Simms (2015), more generally, have documented the importance of scaffolding exploratory problem-solving through a series of studies on induction within (non-) STEM domains. They emphasize explicit support in noticing the relevance of relational thinking, providing adequate processing resources to mentally hold and manipulate relations, and facilitating recognition of both similarities and differences when drawing analogies between systems of relationships. This is because learners may not spontaneously search for a common deep structure across problem instances.

Similar findings have been echoed in prior PS-I work (Schwartz et al., 2011; Kapur, 2015), which suggest that the benefits of prior knowledge activation such as noticing inconsistencies across multiple problem instances, encoding critical features from instruction etc are contingent on relevance of the activation. For instance, in an invention with contrasting cases study on the topic of density (Schwartz et al., 2011), students who recalled the deep structure of ratio from their invention activity were the ones who ultimately benefited from activating their prior knowledge on assessments of transfer. Scaffolding initial problem-solving as part of the PS-I design might then be one means to help learners activate relevant prior knowledge before receiving instruction.

Prior research on the mechanisms of errorful generation suggests that benefits are more likely when learners generate information semantically related to relevant task concepts and/or when subsequent feedback is related to these concepts (Clark, 2016). For e.g., in word-pair generation tasks, generations based on word stems or rhyming are unlikely to produce as much semantic activation, and do not show the beneficial effects of generation. Conceptual processing (guesses) afforded by error generation facilitate richer memory trace through ordered relations between errors and targets (leading to better recall and problem-solving performance), compared to, non-conceptual processing (lexical guesses) that is more likely to create retrieval noise without effortful semantic elaboration on part of the learner (Cyr & Anderson, 2015). Taken together, we can say that in absence of spontaneous task reasoning with relevant induction criteria (that can potentially be scaffolded within a +PS-I design), PF can fail.

However sometimes, even if task reasoning comprises relevant induction criteria, PF can be expected to fail if such task reasoning is then followed explicit instructions to come up with a unifying functional relation (how variables interact to produce a single quantitative result). Finding a very high number of similarities and differences in the contrasting cases (as part of initial task reasoning) can actually hurt posttest performance. Inventing can be expected to decrease learner's willingness to deeply process subsequent instruction because of clinging on to these self-generated suboptimal inventions (Johnson & Seifert, 1994), and valuing self-made products highly (Norton, Mochon, & Ariely, 2012). This can result in failure to recognize deficiency in problem-solving performance. Often, learner inventions fail to consider all factors necessary for developing the canonical solution, but focus only on subsets of these contrasting cases. In +PS-I work by Roelle and Berthold (2016), such detrimental effects increased as a function of the number of detected similarities and differences for which learners had generated inventions.

PF can fail if the delay caused in reaching an appropriate solution makes learners less interested and less self-efficient. As Glogger-Frey, Gaus, and Renkl (2017) found in their

work, this invoked feelings of knowledge insufficiency during preparation and consequently low confidence. With repeated failures, it becomes harder to perceive the value of engaging in good inquiry behaviors during the problem-solving phase because of lowered expectations and increased self-doubt (Ilgen & Hamstra, 1972), acceptance of absence of control (Mikulincer, 2013), susceptibility to demotivation and negative emotions like stress (LePine, LePine, & Jackson, 2004), and increased stability of future failure expectancies (Weiner, 1974). In +PS-I research conducted by Lee (2017) in physics, task failure in the form of circuit explosion (entire electrical circuit goes up in flames and a restart is required) was found to be negatively related to learning outcomes, perhaps because learners were not able to meaningfully grapple with the task complexity and lacked understanding of basic task elements. Prompts for metacognitive reflection did not help learners address these recurring failures.

Further, the temporal distance between the problem-solving and instruction phase matters. PF can be expected to fail if the instruction phase is temporally detached from all the conceptual exploration and reflection, compared to multiple smaller cycles of problem-solving and instruction happening closely together (Kim et al., 2015). The latter offers differentiated and redundant scaffolding opportunities (Tabak, 2004) to address the magnitude/diversity of knowledge assembly that learners need for understanding different conceptual task elements during the consolidation phase. Finally, PF can be expected to perform as well as +PS-I when cognitive support offered in the initial problem-solving phase is focused on principle-based guidance (covering definitions, conditions of applicability, relevant equations etc), as opposed to, being focused on clarifications and hints regarding correct solution steps, accuracy feedback etc. When learners have no or little relevant prior knowledge related to the target learning content, providing principle-based guidance during their initial problem-solving reduces extraneous cognitive load and in turn facilitates attention to critical task concepts.

Negative/Null Results for PF (compared to !PS-I)

PF fidelity check revealed that most of the 11 experimental comparisons in this cluster (Aleven, Koedinger, & Roll, 2009; Roll et al., 2011; Glogger-Frey, Fleischer, Grüny, Kappich, & Renkl, 2015; Kapur, 2015; Likourezos & Kalyuga, 2017; Newman & DeCaro, 2018) considered affective draw of the problem (9/11). However, about only half of these comparisons provided evidence for multiple RSM generation during the initial problem-solving phase (5/11). Further, only one third comparisons included follow-up instruction building on learner generated solutions (4/11) and used group work as the participation structure (3/11). This suggests low conformity to the PF design criteria. Comparison of such low fidelity versions of PF with !PS-I implementations indicates relatively lower extraneous load in !PS-I conditions as a key factor for the pattern of results. The !PS-I conditions usually include worked example followed by instruction, but sometimes also preparatory activities such as evaluating pre-designed solutions, problem-posing, reading/summarizing text etc followed by instruction.

One way to interpret the null results across these comparisons is by considering the relative contribution of different instructional activities and the socio-cognitive processes they trigger. As Kalyuga and Singh (2016) suggest, high(er) extraneous load for the PS-I condition is compensated by increase in intrinsic load (because of the diversity of instructional goals in the problem-solving phase such as prior knowledge activation, deep feature identification etc, as opposed to a solitary goal of solution schema acquisition). Also, PF learners experience motivational effects (acceptance of challenge, resolving conflict etc) that are different from those experienced by learners in a !PS-I condition (belief of success probability etc). Thus, one might conjecture the relative efficacy of PF over !PS-I implementations to depend on the balance between extraneous load and intrinsic load triggered by sequences of instructional tasks (that individually achieve different sub-goals). More research is needed along these lines.

Summary and Conclusion

We articulated factors representative of learner's situatedness relative to their problem-solving experiences to examine boundary conditions for failure of PF. PF (or more generally, PS-I) was compared with three alternate experimental conditions, (i) I-PS (instruction followed by problem-solving), (ii) +PS-I (scaffolded problem-solving followed by instruction), (iii) !PS-I (preparatory activity other than problem-solving followed by instruction). To summarize, our current analyses suggested low design fidelity (weak conformity to PF design criteria) as the starting point for when PF fails. However, deeper exploration into experimental comparisons with negative and null results for PF highlighted four important factors.

First, incoming cognitive and motivational characteristics (e.g., mastery orientation, self-regulation skills, inquiry goals) influence whether learners can be expected to benefit from PF. Second, nature of the problem-solving task (e.g., task difficulty/calibration to prior knowledge, triggered socio-cognitive processes, domain specificity, implicit task feedback) sheds further light into when PF can be expected to fail. Prior knowledge activation is a key cognitive mechanism that explains the beneficial effects of problem-solving based preparatory activities within the learning design of PS-I (Loibl et al., 2017). The boundary conditions explored in this work open up new research opportunities for developing variants of PF, or combining PF with other cognitively activating instructional methods (Hofer, Schumacher, Rubin, & Stern, 2018) for achieving stronger and more sustainable results. Such methods, which focus on learner's naïve concepts and beliefs as the starting point for knowledge construction and reorganization (Schneider & Stern, 2010) can include self-explanations, metacognitive questioning etc.

Third, learner solutions during the problem-solving phase (e.g., usage of relevant induction criteria, evidence for internalization, behavior rigidity) and the extent to which they

are scaffolded impacts learning from PF. Finally, nuances related to the overall PS-I learning design (e.g., redundancy of pretest, anchoring of initial problem-solving tasks, feedback in instruction phase, additional practice activities after instruction) matter for efficacy of PF activities over alternate designs. Although not exhaustive, these factors synthesized from studies around PF (and more broadly the PS-I literature) provide evidence-driven rationale for more careful design/labeling of future implementations. We hope this will spur lines of inquiry (e.g., see Sinha et al. (2019)) that design for balancing the incommensurable goals of learning versus performance (Soderstrom & Bjork, 2015), given the differential relationship of failure to these goals (Kapur, 2016).

References

Aleven, V., Koedinger, K., & Roll, I. (2009). Helping students know further–increasing the flexibility of students knowledge using symbolic invention tasks. In *Proceedings of the annual meeting of the cognitive science society.*

Belenky, D. M., & Nokes-Malach, T. J. (2012). Motivation and transfer: The role of mastery-approach goals in preparation for future learning. *Journal of the Learning Sciences*, *21*(3), 399–432.

Chase, C. C., & Klahr, D. (2017). Invention versus direct instruction: for some content, itsa tie. *Journal of Science Education and Technology*, *26*(6), 582–596.

Chi, M. T., De Leeuw, N., Chiu, M.-H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive science*, *18*(3), 439–477.

Clark, C. M. (2016). *When and why does learning profit from the introduction of errors?* Unpublished doctoral dissertation, University of California, Los Angeles.

Cyr, A.-A., & Anderson, N. D. (2015). Mistakes as stepping stones: Effects of errors on episodic memory among younger and older adults. *Journal of experimental psychology: learning, memory, and cognition*, *41*(3), 841.

De Bock, D., Deprez, J., Van Dooren, W., Roelens, M., & Verschaffel, L. (2011). Abstract or concrete examples in learning mathematics? a replication and elaboration of kaminski, sloutsky, and heckler's study. *Journal for research in Mathematics Education*, *42*(2), 109–126.

DeCaro, D. A., DeCaro, M. S., & Rittle-Johnson, B. (2015). Achievement motivation and knowledge development during exploratory learning. *Learning and Individual Differences*, *37*, 13–26.

DeCaro, M. S., & Rittle-Johnson, B. (2012). Exploring mathematics problems prepares children to learn from instruction. *Journal of experimental child psychology*, *113*(4).

Dweck, C. S. (1992). Article commentary: The study of goals in psychology. *Psychological Science*, *3*(3), 165–167.

Fyfe, E. R., DeCaro, M. S., & Rittle-Johnson, B. (2014). An alternative time for telling: When conceptual instruction prior to problem solving improves mathematical knowledge. *British Journal of Educational Psychology*, *84*(3).

Glogger-Frey, I., Fleischer, C., Grüny, L., Kappich, J., & Renkl, A. (2015). Inventing a solution and studying a worked solution prepare differently for learning from direct instruction. *Learning and Instruction*, *39*, 72–87.

Glogger-Frey, I., Gaus, K., & Renkl, A. (2017). Learning from direct instruction: Best prepared by several self-regulated or guided invention activities? *Learning and Instruction*, *51*, 26–35.

Hofer, S. I., Schumacher, R., Rubin, H., & Stern, E. (2018). Enhancing physics learning with cognitively activating instruction: A quasi-experimental classroom intervention study. *Journal of Educational Psychology*.

Holmes, N. G., Day, J., Park, A. H., Bonn, D., & Roll, I. (2014). Making the failure more productive: scaffolding the invention process to improve inquiry behaviors and outcomes in invention activities. *Instructional Science*, *42*(4).

Hsu, C.-Y., Kalyuga, S., & Sweller, J. (2015). When should guidance be presented in physics instruction? *Archives of Scientific Psychology*, *3*(1), 37.

Ilgen, D. R., & Hamstra, B. W. (1972). Performance satisfaction as a function of the difference between expected and reported performance at five levels of reported performance. *Organizational Behavior and Human Performance*, *7*(3), 359–370.

Johnson, H. M., & Seifert, C. M. (1994). Sources of the continued influence effect: When misinformation in memory affects later inferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(6), 1420.

Kalyuga, S., & Singh, A.-M. (2016). Rethinking the boundaries of cognitive load theory in complex learning. *Educational Psychology Review*, *28*(4), 831–852.

Kapur, M. (2008). Productive failure. *Cognition and instruction*, *26*(3), 379–424.

Kapur, M. (2012). Productive failure in learning the concept of variance. *Instructional Science*, *40*(4), 651–672.

Kapur, M. (2014). Productive failure in learning math. *Cognitive Science*, *38*(5), 1008–1022.

Kapur, M. (2015). The preparatory effects of problem solving versus problem posing on learning from instruction. *Learning and instruction*, *39*, 23–31.

Kapur, M. (2016). Examining productive failure, productive success, unproductive failure, and unproductive success in learning. *Educational Psychologist*, *51*(2), 289–299.

Kapur, M., & Bielaczyc, K. (2012). Designing for productive failure. *Journal of the Learning Sciences*, *21*(1), 45–83.

Kapur, M., & Bielczyc, K. (2011). Classroom-based experiments in productive failure. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 33).

Kapur, M., & Toh, P. L. L. (2013). Productive failure: From an experimental effect to a learning design. *Educational design research–Part B: Illustrative cases*, 341–355.

Kim, B., Pathak, S. A., Jacobson, M. J., Zhang, B., & Gobert, J. D. (2015). Cycles of exploration, reflection, and consolidation in model-based learning of genetics. *Journal of Science Education and Technology*, *24*(6), 789–802.

Kuo, E., & Wieman, C. E. (2016). Toward instructional design principles: Inducing faradays law with contrasting

cases. *Physical Review Physics Education Research*, *12*(1), 010128.

Lee, A. (2017). *Productive responses to failure for future learning*. Columbia University.

LePine, J. A., LePine, M. A., & Jackson, C. L. (2004). Challenge and hindrance stress: relationships with exhaustion, motivation to learn, and learning performance. *Journal of Applied Psychology*, *89*(5), 883.

Likourezos, V., & Kalyuga, S. (2017). Instruction-first and problem-solving-first approaches: alternative pathways to learning complex tasks. *Instructional Science*, *45*(2).

Loehr, A. M., Fyfe, E. R., & Rittle-Johnson, B. (2014). Wait for it... delaying instruction improves mathematics problem solving: A classroom study. *The Journal of Problem Solving*, *7*(1), 5.

Loibl, K., & Leuders, T. (2018). Errors during exploration and consolidation - the effectiveness of productive failure as sequentially guided discovery learning. *Journal für Mathematik-Didaktik*, *39*(1), 69–96.

Loibl, K., Roll, I., & Rummel, N. (2017). Towards a theory of when and how problem solving followed by instruction supports learning. *Educational Psychology Review*, *29*(4).

Loibl, K., & Rummel, N. (2014a). The impact of guidance during problem-solving prior to instruction on students inventions and learning outcomes. *Instructional Science*, *42*(3), 305–326.

Loibl, K., & Rummel, N. (2014b). Knowing what you don't know makes failure productive. *Learning and Instruction*, *34*, 74–85.

Marei, H. F., Donkers, J., Al-Eraky, M. M., & van Merrienboer, J. J. (2017). The effectiveness of sequencing virtual patients with lectures in a deductive or inductive learning approach. *Medical teacher*, *39*(12), 1268–1274.

Matlen, B. J., & Klahr, D. (2013). Sequential effects of high and low instructional guidance on childrens acquisition of experimentation skills: Is it all in the timing? *Instructional Science*, *41*(3), 621–634.

Mazziotti, C., Loibl, K., & Rummel, N. (2015). Collaborative or individual learning within productive failure: Does the social form of learning make a difference? International Society of the Learning Sciences, Inc.[ISLS].

Mikulincer, M. (2013). *Human learned helplessness: A coping perspective*. Springer Science & Business Media.

Miller, C. S., Lehman, J. F., & Koedinger, K. R. (1999). Goals and learning in microworlds. *Cognitive Science*, *23*(3).

Newman, P., & DeCaro, M. (2018). How much support is optimal during exploratory learning? In *Proceedings of the 40th annual conference of the cognitive science society.*

Norton, M. I., Mochon, D., & Ariely, D. (2012). The ikea effect: When labor leads to love. *Journal of consumer psychology*, *22*(3), 453–460.

Richland, L. E., & Simms, N. (2015). Analogy, higher order thinking, and education. *Wiley Interdisciplinary Reviews: Cognitive Science*, *6*(2), 177–192.

Roelle, J., & Berthold, K. (2016). Effects of comparing contrasting cases and inventing on learning from subsequent instructional explanations. *Instructional Science*, *44*(2).

Roll, I., Aleven, V., & Koedinger, K. (2011). Outcomes and mechanisms of transfer in invention activities. In *Proceedings of the annual meeting of the cognitive science society.*

Schalk, L., Schumacher, R., Barth, A., & Stern, E. (2017). When problem-solving followed by instruction is superior to the traditional tell-and-practice sequence. *Journal of Educational Psychology*.

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological science*, *3*(4), 207–218.

Schneider, M., & Stern, E. (2010). The cognitive perspective on learning: Ten cornerstone findings. *The nature of learning: Using research to inspire practice*, 69–90.

Schwartz, D. L., & Bransford, J. D. (1998). A time for telling. *Cognition and instruction*, *16*(4), 475–5223.

Schwartz, D. L., Chase, C. C., & Bransford, J. D. (2012). Resisting overzealous transfer: Coordinating previously successful routines with needs for new learning. *Educational Psychologist*, *47*(3), 204–214.

Schwartz, D. L., Chase, C. C., Oppezzo, M. A., & Chin, D. B. (2011). Practicing versus inventing with contrasting cases: The effects of telling first on learning and transfer. *Journal of Educational Psychology*, *103*(4), 759.

Schwartz, D. L., & Martin, T. (2004). Inventing to prepare for future learning: The hidden efficiency of encouraging original student production in statistics instruction. *Cognition and Instruction*, *22*(2), 129–184.

Sinha, T., Kapur, M., West, R., Catasta, M., Hauswirth, M., & Trninic, D. (2019). Impact of explicit failure and success-driven preparatory activities on learning. In *Proceedings of the annual meeting of the cognitive science society.*

Soderstrom, N. C., & Bjork, R. A. (2015). Learning versus performance: An integrative review. *Perspectives on Psychological Science*, *10*(2), 176–199.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive science*, *12*(2), 257–285.

Tabak, I. (2004). Synergy: A complement to emerging patterns of distributed scaffolding. *The journal of the Learning Sciences*, *13*(3), 305–335.

Tam, K. (2017). Examining productive failure instruction in dental ethics.

Tulis, M., & Ainley, M. (2011). Interest, enjoyment and pride after failure experiences? predictors of students state-emotions after success and failure during learning in mathematics. *Educational Psychology*, *31*(7), 779–807.

Vollmeyer, R., Burns, B. D., & Holyoak, K. J. (1996). The impact of goal specificity on strategy use and the acquisition of problem structure. *Cognitive Science*, *20*(1).

Weiner, B. (1974). *Achievement motivation and attribution theory*. General Learning Press.

Complex exploration dynamics from simple heuristics in a collective learning environment

Sabina J. Sloman (SSLOMAN@Andrew.Cmu.Edu)
Department of Social and Decision Sciences, Carnegie Mellon University, 5000 Forbes Avenue
Pittsburgh, PA 15213 USA

Robert L. Goldstone (RGOLDSTO@Indiana.Edu)
Department of Psychological and Brain Sciences and Program in Cognitive Science, Indiana University, 1101 E. 10th St.
Bloomington, IN 47405 USA

Cleotilde Gonzalez (COTY@Cmu.Edu)
Dynamic Decision Making Laboratory, Department of Social and Decision Sciences, Carnegie Mellon University, 5000 Forbes Avenue
Pittsburgh, PA 15213 USA

Abstract

Effective problem solving requires both *exploration* and *exploitation*. We analyze data from a group problem-solving task to gain insight into how people use information from past experiences and from others to achieve explore-exploit trade-offs in complex environments. The behavior we observe is consistent with the use of simple, reinforcement-based heuristics. Participants increase exploration immediately after experiencing a low payoff, and decrease exploration immediately after experiencing a high or improved payoff. We suggest that whether an outcome is perceived as "high" or "low" is a dynamic function of the outcome information available to participants. The degree to which the distribution of observed information reflects the true range of possible outcomes plays an important role in determining whether or not this heuristic is adaptive in a given environment.

Keywords: exploration; exploitation; networks; social learning

Introduction

Search—a dynamic maximization problem where outcomes depend on the agent's location in the problem space—is a fundamental part of our cognitive experience (Hills et al., 2015). When in a new city, we sample from different restaurants in order to find the best places to eat (Mehlhorn et al., 2015). When coming up with a new idea for a research project, the amount of intellectual and social "reward" we expect to experience is a function of whether the point in conceptual space we're interested in is novel and appreciated by others.

Effective search requires both *exploration*, or sampling from the space of outcomes to gain information about what's available, and *exploitation*, or taking advantage of the information available and resampling from places known to produce good outcomes. Should the traveller stick with the first decent restaurant she finds, or keep exploring her options? Should the scientist stick with her current line of work, or branch out into unchartered intellectual territory?

We analyze data from a group problem-solving task to gain insight into how participants use information from past experiences and from others to achieve explore-exploit trade-offs in rugged, networked environments. When the world is uncertain, complex and interconnected, the optimal trade-off between exploration and exploitation depends on the degree of complexity, on the extent of interconnectedness—and on the strategies individuals adopt to process and act on the information they encounter (Barkoczi, Analytis, & Wu, 2016; Barkoczi & Galesic, 2016; Toyokawa, Whalen, & Laland, 2019). In some cases, we may adapt our exploration level to the environment we're in, even when the shape of the environment is unknown to us (Mason & Watts, 2012).

We add to existing work that has looked at behavioral patterns of exploration in different environments, and examine the mechanisms that lead to the individual- and group-level patterns we observe. Our contributions are both methodological and theoretical. From a methodological perspective, we specify a generalization gradient and propose it as a useful measure of both individual- and group-level exploitation in smooth search spaces. From a theoretical perspective, we document exploration patterns and systematic behavioral responses to outcome information. We find that context-dependent explore-exploit trade-offs emerge even when participants are not told what kind of environment they're in, and speculate that differences in exploration patterns can be explained by differences in the outcome information available to participants.

Methods

Experimental paradigm

We analyzed data from the group search task designed and implemented by Mason, Jones, and Goldstone (2008). Each participant guessed numbers between 0 and 100 and a computer revealed to them how many points were obtained from the guess by consulting a hidden *fitness function*[1] that translates a guess into a number of points. Random noise was added to these points so that repeated sampling was necessary to accurately determine the underlying function relating guesses to scores. On each trial, a group of participants was assigned to one of several conditions (discussed below). Trials consisted of 15 rounds, over which each member of the

[1] We will use the terms "fitness function" and "fitness landscape" interchangeably.

group tried to maximize their total number of earned points. Importantly, on each round, participants got feedback not only on how well their own guess fared, but also had access to information about the actions and outcomes of their neighbors.

Two aspects of the environment were experimentally manipulated: the social network (the network topology that determines who counts as neighbors) and the complexity of the task (the shape of the fitness function that converted guesses to earned points). These are discussed in the sections below.

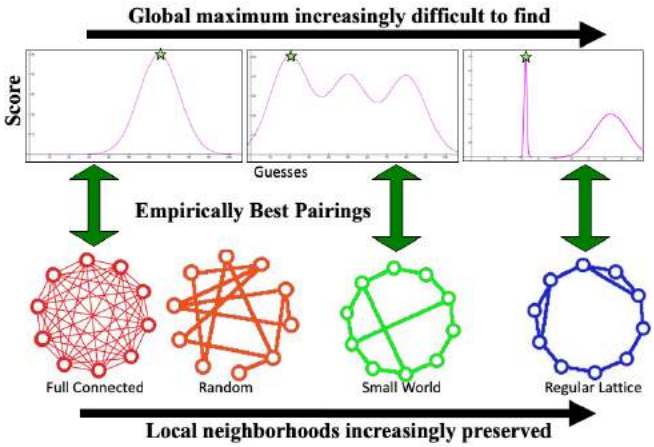


Figure 1: The network structures and fitness functions used in Mason et al. (2008). Reproduced from Goldstone et al. (2013) with permission of the authors.

Social network structure Neighborhoods of participants were created to reflect *random*, *regular lattice*, *small world*, or *fully connected* networks. Examples of the graph topologies for groups of 10 participants are shown in Figure 1. In the random graph, connections are randomly created under the constraint that the resulting graph is connected. Participants in random graphs tend to be connected to others via relatively short paths.

The regular lattice configures a group with an inherent spatial ordering such that people are connected to each other if and only if they are close to one other. The regular lattice also captures the notion of social "cliques": If there is no short path from A to Z, then there will be no direct connection from any of A's neighbors to any of Z's neighbors. The paths connecting people are much longer, on average, in lattice than in random graphs.

"Small world networks" have both cliques and a short average path length (Watts & Strogatz, 1998). From an information processing perspective, small-world networks are attractive because the spatial structure of the networks allows information search to proceed systematically, and the short-cut paths allow the search to proceed quickly (Kleinberg, 2000).

A fourth network, a fully connected graph, allowed every participant to see the guesses and outcomes of every other participant.

	Full	Small world	Random	Lattice	Total
Unimodal	11	11	19	11	**52**
Trimodal	9	12	20	11	**52**
Needle	28	27	18	28	**101**
Total	**48**	**50**	**57**	**50**	**205**

Table 1: Number of trials of each condition in our data.

Environmental complexity Three hidden fitness functions for converting guessed numbers to points were tested across two experiments. The *unimodal* function has a single best solution that can eventually be found with a hill-climbing method. The *trimodal* function increased the difficulty of the search by introducing local maxima. A local maximum is a solution that is better than all of its immediate neighboring solutions, yet is not the best solution possible. Thus, a simple hill-climbing search might not find the best possible solution. Finally, the *needle* function has one very broad local maximum, and one hard-to-find global maximum.[2] The height and variance of the global maximum in the unimodal conditions, global maximum in the trimodal conditions, and local maximum in the needle conditions are all equal (the height of these peaks is 50, while the height of the needle's global maximum is 70).

After excluding some trials due to apparently incomplete data, we used 205 trials in total for our analyses. The number of trials in each conditions is reported in Table 1. The number of players in each trial ranged from 5 to 19, with a mean of 11.89 ($SD = 4.05$).[3]

Measuring exploration

To measure the degree to which a sequence of guesses exploited a location of the search space (or, conversely, didn't explore the space), we developed a similarity metric (hereafter referred to as *similarity*) that captures the average degree of closeness of all pairwise combinations of the elements of a set of guesses along a generalization gradient[4] adapted to the specific problem space:

$$similarity(G_i, G_j) = e^{-(\frac{G_i - G_j}{c})^2}$$

[2]Mason et al. (2008) collected data on variations of the needle function in two separate experiments. In our analyses, when referring to the needle conditions we pooled data from the two experiments.

[3]Each group of participants was assigned to several conditions in sequence (for more details on the experimental procedure, see Mason et al. (2008)). We consider a "trial" to be uniquely specified by a combination of a group and a condition. In other words, if a group completed the task in n conditions, this is recorded in Table 1 as n distinct observations.

[4]A *generalization gradient* is a function that transforms distance in some space to distance in another—usually more psychologically interesting—space.

where $c = .07$ was set to reflect the variance of the global maxima on the unimodal and trimodal landscapes, and the local maxima on the needle landscapes. The total average similarity of a group of guesses G is

$$similarity(G) = \frac{\sum_{i,j} similarity(G_i, G_j) - n}{n^2 - n}$$

where $n = |G|$. We use $1 - similarity(G)$ as our measure of the degree to which G spans—or explores—the problem space.

While other measures, such as variance or the average volatility measure developed by Mason et al. (2008), capture the average distance between guesses, they do not directly capture the idea of the extent to which a set of guesses spans the problem space. Consider a participant A who alternates between guessing 0 and 100, and a participant B who guesses a number at every multiple of 10. We'd like to say that B is the better explorer, because their guesses are spread across the landscape. However, the variance and average volatility of participant A's guesses are much higher than the variance and volatility of B's guesses. By taking the average of *all* pairwise combinations of guesses, our similarity metric captures the *spread* of guesses, rather than simply the extent of their range.

In addition, our metric captures the intuition that similarity drops off steeply with the distance between two nearby solutions, but quickly flattens out (see Figure 2). The choices to jump 99 or 100 units away from where one is are considered effectively identical, while the choices to jump 0 or 1 unit away are much less similar.[5]

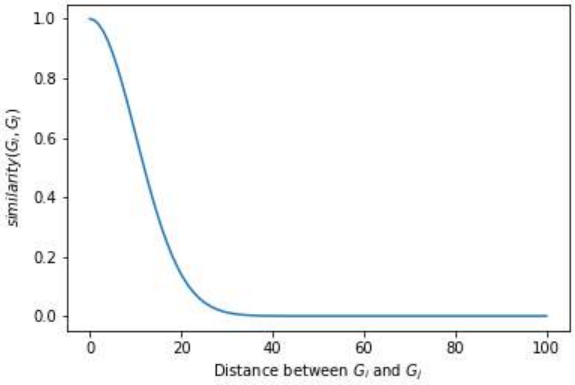


Figure 2: The shape of the generalization gradient underlying the *similarity* metric.

Some evidence suggests that a gradient of this form is a good approximation of how people make inferences about unseen locations in spatial search tasks (Wu, Schulz, Speekenbrink, Nelson, & Meder, 2018). However, here we invoke *similarity* only to operationalize the degree of "exploratoryness" of a set of guesses, not to model participants' inferences.

[5]The Gaussian shape of the fitness functions is compatible with the Gaussian similarity drop-off gradient we used. While this captures many of the same intuitions, it differs from the well-known exponential similarity function (Shepard, 1987). All our results are robust to the use of an exponential similarity function.

Exploration patterns

Individual exploration

Figure 3 shows the heterogeneity in the amount of exploration between participants. Higher density on the right side of the histograms indicates that participants in that condition tended to distribute their guesses more evenly across the problem space.

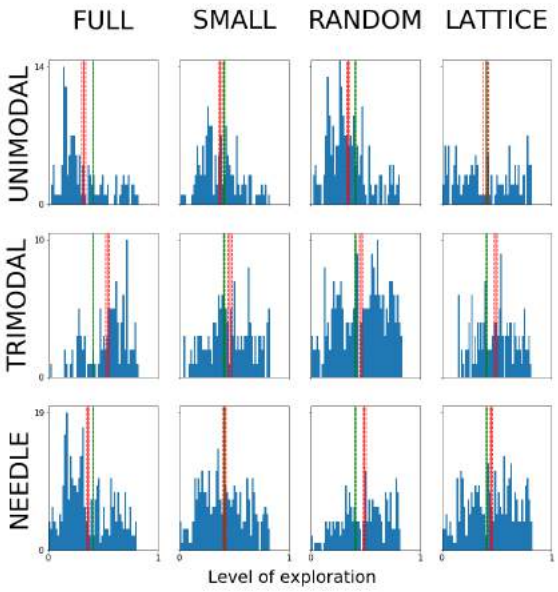


Figure 3: Histogram of participant-level exploration levels. A participant i who makes a sequence of guesses G_i has an exploration level equal to $1 - similarity(G_i)$. The green lines indicate the global mean and standard error of exploration levels across participants in all conditions (.408 (SE = .004)). The red lines indicate the mean and standard error of exploration levels across all participants *within* the respective condition. If a person participated in several conditions, they are treated as a separate participant in each condition.

Individuals tend to explore less than average on the unimodal landscapes, which were explicitly constructed so that their global maxima were easy to find. When the best solution can be found with very little exploration, more extensive search may just lead to foregone payoffs rather than valuable information. We tested this intuition by looking at the correlation between participants' exploration levels and their average payoffs. As expected, this correlation is much lower on the unimodal landscapes than on the more complex landscapes.

Exploration levels tend to be lower than average in one other condition: the fully-connected network on the needle landscape. Mason et al. (2008) found that when confronted with the difficulty of the needle landscape, participants tended to do better when in the sparsely-connected lattice network (see Figure 1). They speculated that this was because distributing social information hindered bandwagoning, or collective convergence on the tempting local maximum. Our results corroborate this speculation: While participants in the fully-connected networks explore the needle landscape less than average, the mean exploration level in the lattice networks is higher than the global average.

Collective exploration

The similarity metric allows us to calculate the "exploratoryness" of an arbitrary sequence of guesses. In particular, we can also use it to measure *group-level*, or collective, exploration.

Figure 4 shows the evolution of collective exploration over rounds, alongside the proportion of participants who were within one standard deviation of the global maximum on each round. Collective exploration declines quickly on the unimodal and needle landscapes. While this coincides with more participants finding the global maximum on the unimodal landscape, the proportion of participants who find the global maximum in the needle condition remains relatively low. These patterns reflect dynamics analogous to the individual-level patterns we discussed in the previous section: The group explores less when there is a salient local maximum, and especially so when outcome information is rapidly broadcast throughout the network.

The consequences of early exploration

Our explanations for many of the results in the previous sections depend on our assumption that exploration is more important in some cases than in others. In some environments, low exploration may cause high payoffs; quick convergence on promising areas of the landscape may cause the average payoff to rise. In others, maintaining a high amount of exploration and broadly surveying the problem space could lead to subsequently higher payoffs. This section further unpacks the sequentially contingent relationship between exploration and expected reward in the different conditions.

Figure 5 plots the cross-correlations between average payoffs and the collective exploration level within a round. It's unsurprising that all the correlations are below zero; as shown Figure 4, collective exploration subsides while payoffs increase over time. More informative for our purposes is the *difference* between the correlation of early exploration with later payoffs, and the correlation of early payoffs with later exploration. An interpretation that exploration causes higher downstream payoffs would require that the former be higher than the latter. The insets of the plots shows this difference for each condition. When the blue line is above zero, this indicates that exploration now is more highly correlated with payoffs later, than payoffs now are with exploration later.

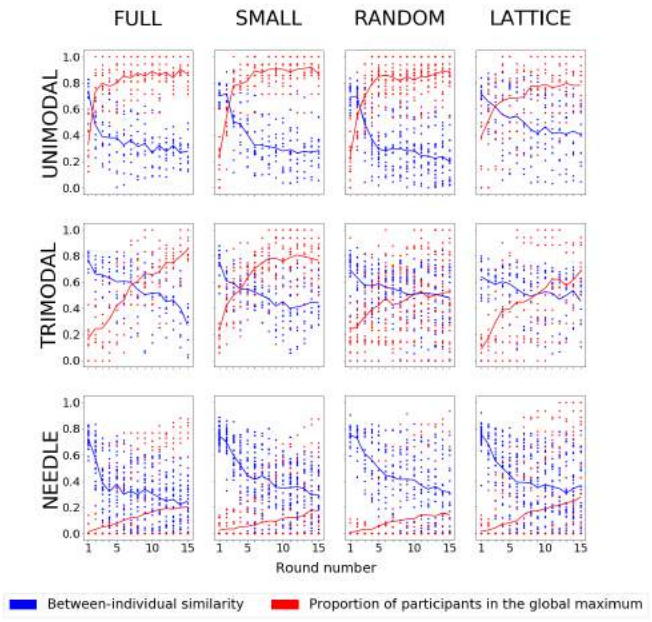


Figure 4: Collective exploration levels (blue) and proportion of players within one standard deviation of the global maximum (red) over rounds. Each dot represents one trial. On round t, group k has an exploration level equal to $1 - similarity(G_{k,t})$ where $G_{k,t}$ is the set of all guesses the members of group k made on that round. The blue line plots the mean exploration level across trials, and the red line plots the proportion of all players across trials who were within one standard deviation of the fitness function's global maximum.

The positive trend in the inset is most consistent across the needle and regular lattice conditions. When connectivity is low and finding the global maximum is especially difficult, group-level exploration leads to higher downstream payoffs. While this pattern is also consistent with an account that higher payoffs cause quicker collective convergence, the "exploration leads to downstream payoffs" account has the advantage that it predicts the particularly strong positive trend for the needle landscape, which is explicitly designed so that the global maximum is hard to find without considerable exploration.

WSLS: Win-shift-less, lose-shift-more

Win-stay, lose-shift (WSLS) is a heuristic applicable to search tasks by adaptive biological and artificial systems. The rule is simple: When you're successful, stay close to where you currently are. When you're unsuccessful, move further away (Bonawitz, Denison, Gopnik, & Griffiths, 2014; Nowak & Sigmund, 1993; Robbins, 1952).

WSLS is usually applied in contexts with discrete binary outcomes that can be easily dichotomized into wins and losses. However, the problem space facing the currently considered participants, like many real-world problem spaces, is both smooth—similarity of actions predicts similarity of

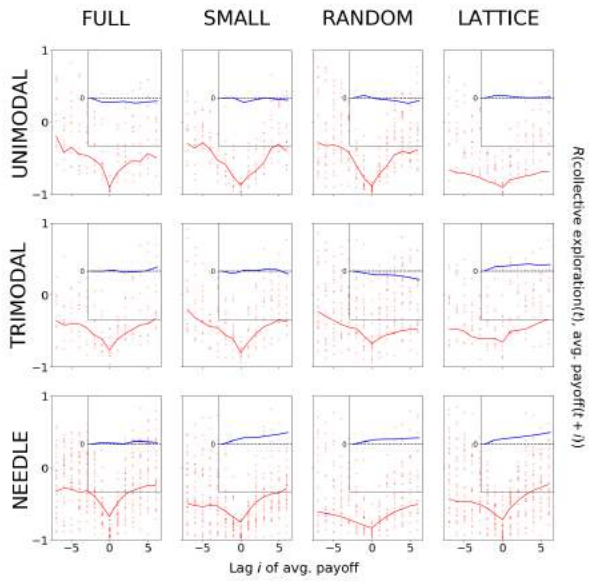


Figure 5: Cross-correlations between group exploration level and payoffs. On round t, group k has an exploration level equal to $1 - similarity(G_{k,t})$ where $G_{k,t}$ is the set of all guesses the members of group k made on that round. Payoff observations are the average payoff participants experienced on round t. A lag of i on the x-axis indicates the correlation between group exploration level at time t and average payoffs at time $t+i$. Each dot corresponds to the correlation using the data from one trial. The red lines plot the correlations across trials. Insets show the difference between the cross-correlation coefficients at lag i and lag $-i$ for $0 \leq i \leq 7$.

outcomes—and continuous. In this section, we show that participant behavior is consistent with a generalization of WSLS: The degree to which participants stray from promising locations varies with both the absolute and relative amount of reward they've experienced there. Participants shift less when they win, and shift more when they lose.

In many situations, WSLS or close variants can lead to approximately optimal search behavior on intractable problem spaces (Bonawitz et al., 2014; Robbins, 1952). To the participants facing the task at hand, the range of possible outcomes is unknown. We suggest that they dynamically incorporate outcome information into their understanding of what's a "win" and what's a "loss". When good outcomes are easy to find, the outcome information participants accumulate accurately reflects the range of attainable payoffs. In these cases, the application of WSLS-like rules may lead to adaptive explore-exploit trade-offs. But when the best outcomes are difficult to find, participants do not get full outcome information about the range of possible payoffs. They fail to appropriately calibrate their "shift-more" and "shift-less" responses. On the needle landscape, WSLS-like rules may lead participants to prematurely converge on the local maximum. In short, when good outcomes are hard to find, information flow is reduced, and individuals cannot appropriately tune their behavior to the relevant search space, resulting in suboptimal individual- and group-level outcomes.

Absolute "wins": Responses to high payoffs

Figure 6 shows how the similarities between participants' preceding guesses (blue) and subsequent guesses (red) covary with the payoffs they experience. Recall that the similarity of two guesses is a measure of the closeness of the guesses. If a participant's guesses on round t and round $t+1$ are more similar than their guesses on round t and round $t-1$, we say they are *exploiting* more on round $t+1$ than on round t.

In all conditions, there is some payoff value above which participants tend to exploit more than explore. The blue vertical lines mark the normalized payoff values where the trend in participants' future level of convergence dips below their past level of convergence—participants begin to shift more (explore). The red vertical lines indicate payoff values where the reverse switch occurs—participants begin to shift less (exploit). In general, participants shift more when payoffs are low, and shift less when payoffs are high.

Where this switch occurs varies by landscape. We speculate that these differences are a direct effect of differences in the outcome information available to participants, and how they adjust their beliefs about the range of possible payoffs based on their observations (Parducci, 1965). In the trimodal conditions, the "switch point" is shifted to the right (participants wait for relatively high payoffs before they begin to settle down), but so is the density of payoff observations. As shown in Figure 1, payoffs on the trimodal landscape remain relatively high even when participants stray from the global maximum. When they observe that locations that lead to "wins" are distributed widely across the landscape, participants are more reluctant to settle down.

By contrast, in the needle conditions, both the switch point and the bulk of the density is shifted towards the left side of the plot. Few participants stumble upon the narrow global maximum, and most experienced payoffs are a smaller proportion of the highest possible payoff. The emergent patterns resemble those in the unimodal conditions because most participants do not have outcome information to suggest that they *are not* on a well-behaved landscape with a similar payoff distribution.

Relative "wins": Responses to improving payoffs

Figure 7 shows how the similarity of participants' guesses changes as a function of the difference between their most recently experienced payoffs. Points to the right of the y-axis correspond to instances where a player had just experienced an immediate increase in payoff. Points above the x-axis correspond to instances where the player's round-to-round exploration level decreased. In general, immediate gains lead to convergence, and losses lead to continued exploration.

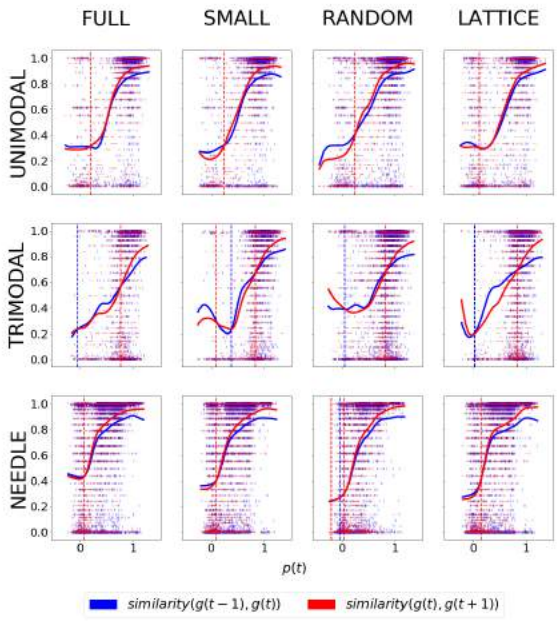


Figure 6: Experienced payoffs against similarity of guesses. $p(t)$ denotes the payoff a player experienced at time t (normalized by the height of the global maximum in each condition), and $similarity(g(t), g(t'))$ denotes the similarity between a guess made at time t and a guess made at time t'. Each dash corresponds to one experienced outcome. Solid lines show the estimated Gaussian kernel regressions. Vertical lines mark shifts between exploitation and exploration (see text).

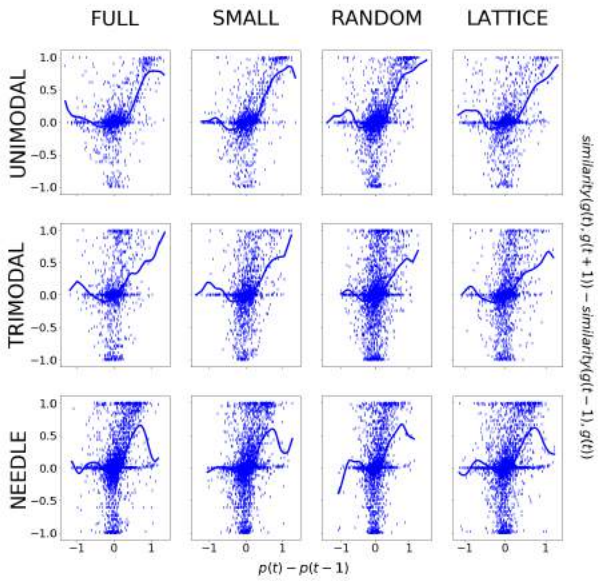


Figure 7: *Differences in* experienced payoffs against *differences in* similarity of guesses. Notation is the same as in Figure 6. Each dash corresponds to one experienced outcome. Solid lines show the estimated Gaussian kernel regressions.

Note the inverted U-shaped trend recovered by the kernel regression across the needle landscapes. Participants who have just experienced an exceptionally large improvement tend to shift more than those who have just experienced a moderate improvement. This is consistent with our understanding of WSLS as a dynamic process: Participants who stumble upon the global maximum dynamically adjust their understanding of the range of possible payoffs, and are less willing than before to settle with what they have.

Discussion

We argued that the behavioral patterns we observe are consistent with the application of a dynamic, continuous variant of win-stay, lose-shift. While participants tend to "stay" in areas where they've experienced both high and improving payoffs, they use information from themselves and others to adapt their willingness to "stay" and "shift" to their environment.

One phenomenon we have only briefly addressed is Mason et al. (2008)'s finding that participants in the lattice network were more likely than participants on other networks to find the needle landscapes' global maxima. Our central claim is that reduced information flow can lead to suboptimal outcomes when participants do not have full information about the range of possible payoff values. Why would the network that restricted information flow the most perform the best when the search task is especially hard?

Visual inspection of Figure 6 suggests that the payoff values at which participants switch from exploration to exploitation do not vary much as a function of the network structure, but depend more on the underlying fitness function. The lattice network's structural restriction of information flow could mean that it takes participants even longer to reach their threshold value or "switch point". Participants may search longer for "wins", resulting in more exploration where it matters the most.

While our analyses were motivated by our desire to understand the relationship between individual- and group-level exploration dynamics, we did not assume that participants make this trade-off explicitly. Rather, we suggested that participants may be using simple heuristics from which an apparent trade-off emerges. We adopted an information processing framework (Oppenheimer & Kelso, 2015): The environment affects behavior and outcomes via its effect on the information group members receive and broadcast to others. By analyzing participant behavior through the lens of *information flow*, we can come closer to understanding what determines the search conditions under which we do well, and the conditions under which we could do much better.

Acknowledgements

Cleotilde Gonzalez was supported by the Army Research Office, Network Science Program, Award Num-

ber:W911NF1710431.

Author note

A limited version of this work will appear as an abstract in the proceedings for ACM Collective Intelligence 2019.

References

Barkoczi, D., Analytis, P., & Wu, C. M. (2016). Collective search on rugged landscapes: A cross-environmental analysis. In *Proceedings of the 38th annual conference of the cognitive science society.*

Barkoczi, D., & Galesic, M. (2016). Social learning strategies modify the effect of network structure on group performance. *Nature Communications*, *7*.

Bonawitz, E., Denison, E., Gopnik, A., & Griffiths, T. L. (2014). Win-stay, lose-sample: A simple sequential algorithm for approximating bayesian inference [Journal Article]. *Cognitive Psychology*. doi: 10.1016/j.cogpsych.2014.06.003

Goldstone, R., Wisdom, T., Roberts, M., & Frey, S. (2013). Learning along with others. *Psychology of Learning and Motivation*, *58*, 1–45.

Hills, T. T., Todd, P. M., Lazer, D., Redish, A. D., Couzin, I. D., & the Cognitive Search Research Group. (2015). Exploration versus exploitation in space, mind, and society. *Trends in Cognitive Sciences*, *19*(1), 46–54.

Kleinberg, J. (2000). Navigation in a small world. *Nature*, *406*.

Mason, W., Jones, A., & Goldstone, R. (2008). Propagation of innovations in networked groups. *Journal of Experimental Psychology: General*, *137*, 422–433.

Mason, W., & Watts, D. (2012). Collaborative learning in networks. *Proceedings of the National Academy of Sciences*, *109*(3), 764–769.

Mehlhorn, K., Newell, B. R., Lee, M., Morgan, K., Braithwaite, V. A., Hausmann, D., ... Gonzalez, C. (2015). Unpacking the explorationexploitation tradeoff: A synthesis of human and animal literatures [Journal Article]. *Decision*. doi: 10.1037/dec0000033

Nowak, M., & Sigmund, K. (1993). A strategy of win-stay, lose-shift that outperforms tit-for-tat in the prisoner's dilemma game. *Nature*, *364*.

Oppenheimer, D. M., & Kelso, E. (2015). Information processing as a paradigm for decision making [Journal Article]. *The Annual Review of Psychology*, *66*, 277-294. doi: 10.1146/annurev-psych-010814-015148

Parducci, A. (1965). Category judgment: A range-frequency model. *Psychological Review*, *72*(6), 407-418.

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, *58*, 527-535.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science [Journal Article]. *Science*, *237*(4820), 1317-1323.

Toyokawa, W., Whalen, A., & Laland, K. N. (2019). Social learning strategies regulate the wisdom and madness of interactive crowds. *Nature Human Behavior*, *3*, 183-193.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of "small-world" networks. *Nature*, *393*.

Wu, C., Schulz, E., Speekenbrink, M., Nelson, J., & Meder, B. (2018). Generalization guides human exploration in vast decision spaces. *Nature Human Behavior*, *2*, 915-924.

Contextual Determinants of Adjective Order: Beyond *Itsy Bitsy Teeny Weeny Yellow Polka Dot Bikini*

Anastasia Smirnova (smirnov@sfsu.edu)
Department of English Language and Literature, 1600 Holloway Ave
San Francisco, CA 94132 USA

Ricardo Romero Sanchez (rricardo@mail.sfsu.edu)
Department of English Language and Literature, 1600 Holloway Ave
San Francisco, CA 94132 USA

Alexander Lenarsky (alenarsk@mail.sfsu.edu)
Department of English Language and Literature, 1600 Holloway Ave
San Francisco, CA 94132 USA

Abstract

Previous research on adjective ordering in linguistics and psychology has focused primarily on the unmarked or default order of adjectives, as in *large blue car*. Inverted word order, as in *blue large car*, which violates the proposed semantic constraints on adjective placement, received relatively little attention. In two studies we show that the inverted order is not as limited in scope as previous researchers have argued. We propose that the inverted word order reflects the *subjective distance* principle: the attribute that is psychologically closer to the speaker is mentioned first. Our explanation draws on research on word order in binomials, thus connecting two previously unrelated research traditions on word order in linguistics and cognitive psychology.

Keywords: adjective ordering; binomials; context-dependency; semantics; pragmatics; subjective distance

Introduction

Why does *Itsy bitsy teeny weeny yellow polka dot bikini* sound so good to the ear? One possible factor is the choice of adjectives and their artful arrangement. What factors determine the 'right' order of adjectival modifiers in a phrase has been a topic of active inquiry in linguistics and psychology (Cinque, 2014; Danks & Glucksberg, 1971; Kotowski & Härtl, 2019; Martin, 1969; Scontras, Degen, & Goodman, 2017; Truswell, 2009; Wulff, 2003). According to the semantic approach, the order of adjectival modifiers is dependent on their semantic class, such as e.g. Color and Size. The underlying assumption is that semantic classes form a scale with respect to some psychological property, such as subjectivity. If class Size precedes class Color on the subjectivity scale, we expect that adjectives denoting size (e.g., *teeny weeny*) will precede adjectives denoting color (e.g., *yellow*)(cf. Dixon, 1982; Hetzron, 1978; Whorf, 1945).

Most of the research in the semantic tradition aims to explain the unmarked or default adjective order, as in *large blue car*. While many authors acknowledge that in some communicative situations the default, semantically determined word order can be overridden, the mechanisms that give rise to the inverted word order, as in *blue large car*, have received relatively little attention in the literature. One exception is a series of studies by Danks and co-authors in the early 70s (Danks & Glucksberg, 1971; Danks & Schwenk, 1972). These authors advocate a pragmatic approach and propose that the order of adjectives depends on how well they differentiate among salient contextual alternatives: the most discriminative adjective is mentioned first. For example, in a context in which two large cars, one red and one blue, are equally salient, color would be more discriminative than size, and would give rise to the inverted word order: *blue large car*. In this approach the communicative goals of conversation participants rather than semantic classes of adjectives and their properties determine adjective ordering.

One of the limitations of the pragmatic approach proposed by Danks and co-authors is that it only applies to cases in which the set of potential discourse referents and their attributes (the two cars in the example above) have already been established. Our two experimental studies demonstrate that the inverted adjective ordering is also attested in contexts without previously established referents. Such cases cannot be explained by reference to discriminative attributes, because there are no alternatives that need to be differentiated. Our explanation of the inverted word order in such contexts is based on research on flexible word order in binomials, i.e. constructions with two conjoined nouns or adjectives, as in *Democrats and Republicans* (Iliev & Smirnova, 2016; see also Cooper & Ross, 1975). Specifically, we propose that the inverted adjective order reflects the same psychological principle that was proposed to explain word order in binomials – the *subjective distance principle*. According to this principle, the attribute that is psychologically closer to the speaker is mentioned first. Our paper offers a principled explanation for inverted word order and uncovers parallels between two previously unrelated domains of research: word order in binomials and word order in adjectival modifiers.

Two Approaches to Adjective Ordering

The Semantic Approach

According to the semantic approach, the meaning of adjectives is responsible for the observed word order (Martin, 1969; Scontras et al., 2017). Several semantic features have been proposed in the literature as determinants of adjective ordering, including definiteness of denotation, inherentness (also referred to as intrinsicalness, substantiveness, and closeness to the noun in meaning), and subjectivity.[1] An adjective's definiteness of denotation is determined both by the number of nouns it can modify and by the number of senses an adjective may have in different contexts (Martin, 1969; Wulff, 2003). Adjectives which are relatively limited in the number of nouns they may modify, and whose denotations remain constant across varying contexts, are said to be more definite, and are positioned nearer to the noun (Martin, 1969). Thus, *old embroidered pillow* sounds more natural than *embroidered old pillow*, because the more definite adjective, *embroidered*, is positioned closer to the noun.

In one of the first psychological studies on adjective ordering, Martin (1969) contrasted different semantic features of adjectives and found that definiteness of denotation is the most reliable predictor of an adjective's position in a phrase (cf. Wulff, 2003). Martin proposed that the definiteness of denotation is directly related to an adjective's accessibility, which in turn determines the placement of adjectives. Definite adjectives are accessed faster than indefinite ones, because their meaning does not vary depending on the noun they modify. Therefore, less time is needed to scan the meaning of the definite adjective and to establish the relation between the property denoted by this adjective and the meaning of the modified noun. In this approach, the adjectives that are accessed faster – definite adjectives – appear closer to the noun. In production, the adjectives that were accessed first come last in a sequence of adjectives.

An adjective's inherentness is defined as the extent to which its intrinsic, essential properties reflect the properties of the noun it modifies. Adjectives denoting color, material, breed, nationality, and function are considered to be more inherent than adjectives denoting size, shape or similarly non-intrinsic qualities (Martin, 1969). More inherent adjectives are positioned closer to the noun. The scale in (1) ranks different semantic classes in terms of their inherentness.

(1) Personal judgment < Size < Color < Place of origin **Noun**

less inherent ⟷ more inherent

[1] Other possible determinants of adjective ordering are morphological weight, as well as various semantic considerations, including whether the adjectives refer to temporary or permanent properties, whether adjectives are subsective or intersective (Truswell, 2009), etc. See Scontras et al. (2017) and Kotowski and Härtl (2019) for evaluation of these models' predictions.

According to the scale in (1), based on the adjective classification in Danks and Glucksberg (1971), adjectives denoting size will be placed before adjectives denoting color, because the latter are more inherent and should be positioned closer to the nominal head (Noun).

Besides definiteness of denotation and inherentness, which dominated research on adjectives in the 60s and 70s, another semantic feature – subjectivity – came to the attention of psychologists and linguists (Hetzron, 1978; Quirk et al., 1985; Scontras et al., 2017). According to Quirk et al. (1985: 1341), there is "one principle accounting for all premodifiers: a subjective / objective polarity. That is, modifiers relating to properties which are […] visually observable, and objectively recognizable or assessible, will tend to be placed nearer to the head and be preceded by modifiers concerned with what is relatively a matter of opinion, imposed on the head by the observer, not visually observed, and only subjectively assessible."

Scontras et al. (2017) conducted the first experimental studies probing the effect of subjectivity on adjective ordering. The authors first measured ordering preference by eliciting naturalness ratings for adjective-adjective-noun sequences. Experimental participants were asked to indicate which ordering sounded more natural, e.g. *big blue box* or *blue big box*. Once ordering preferences were established, the authors tested adjective subjectivity by asking participants to indicate the subjectivity of the same adjectives on a sliding scale, ranging from "complete subjectivity" to "complete objectivity". Results indicated that adjective subjectivity indeed predicts ordering preference, where the higher an adjective's subjectivity is, the farther it is placed from the noun it modifies. The authors concluded that subjectivity alone was more reliably predictive of adjective ordering than other semantic features, including the adjectives' inherentness.

A common feature of all semantic accounts is that they are designed to explain the unmarked / default adjective ordering. While most authors working in this tradition acknowledge that the unmarked order can be violated, little is known about the mechanisms that regulate the ordering of adjectives in such cases. The exception is a series of work by Danks and colleagues, to which we turn next.

The Pragmatic Approach

According to the pragmatic approach, communicative goals of conversation participants affect how adjectives are positioned in a noun phrase (Danks & Glucksberg, 1971; Danks & Schwenk, 1972). Specifically, more discriminative or informative adjectives tend to be mentioned first. In different communicative contexts, different features will be more informative, which explains variation in adjective ordering. In the experimental paradigm employed by Danks and co-authors, participants were presented with a set of objects, and were asked to choose the description of the target object (or referent) that would help the listener to distinguish it from non-referents. For example, if the set of objects consists of two cars, which have the same size,

large, but differ in color, one is blue and another one is red, the color is the most informative feature, and is predicted to be mentioned first. Thus, if the target object in question is the blue car, the participants are expected to describe it as *the blue large car*, mentioning the color attribute first, and violating the default word order (e.g., *the large blue car*), where size precedes color, as predicted by (1). The results of experimental studies confirmed this prediction in both comprehension and production tasks. Danks and Schwenk (1972) found that when color is the discriminative feature, it is mentioned first in 57 % of the cases. In the control condition, the normal word order, i.e. size before color, was preferred in 85% of cases.

Danks and co-authors argue that the pragmatic rule is more general and that the semantic rule based on inherentness, which explains the default adjective ordering, is in fact "the most *frequent* case of the more general pragmatic rule" (Danks & Glucksberg, 1971). That is, since more intrinsic adjectives tend to be less informative, they are less likely to discriminate between referents and non-referents, and, therefore, are less likely to appear first in a sequence of adjectives.

The pragmatic approach proposed by Danks and co-authors was criticized by the advocates of the semantic approach. For example, Martin and Ferb (1973) observe that the unmarked and marked adjective orders have different phonological and syntactic properties. The unmarked word order is characterized by constant stress on all adjectives (or by increasing stress) and by the lack of juncture (pause) between the adjectives, e.g. *large blue car*. Syntactically these phrases are argued to have a flat, multiple-branching structure. On the other hand, the contextually-determined order shows contrastive stress on the discriminating adjective, and a juncture, e.g. *BLUE, large car* Syntactically these constructions have a right-branching structure (see Kotowski & Härtl, 2019; Scott 2002; Sproat and Shih, 1988 for discussion). Since the unmarked and marked structures have different properties, they cannot be accounted for by the same rule, i.e. the general pragmatic principle proposed by Danks and co-authors.

Martin and Ferb (1973) and Richards (1975) further argue that while communicative demands can sometimes trigger the inverted word order observed in Danks and Schwenk's (1972) experiments (*BLUE large car*), the same effect can be achieved by preserving the unmarked order but stressing the informative adjective, as in *large BLUE car*. Richards (1975) argues that in the paradigm adopted by Danks and Schwenk (1972), the color adjective must be stressed to produce preference for the inverted word order. Another weakness of Danks and Schwenk's studies is that they only take into consideration two classes of adjectives: color and size. Based on these observations, Richards (1975: 213) concludes that "the speakers are reluctant to give up their a priori preference for normal order and will do so only under highly specialized circumstances." From this perspective, the inverted order is seen as an optional, limited in scope phenomenon, which is peripheral to the study of adjective ordering in general.

While some of the criticism against the pragmatic approach might be justified, neither Martin and Ferb (1973), nor a more recent study by Scontras et al. (2017) offer a principled explanation of contextually-induced order. In what follows, we (i) present the results of two experimental studies which show that the inverted order of adjectives is more common and appears in a larger number of contexts than what was previously assumed, and (ii) propose that some cases of the inverted order can be explained by the *subjective distance* principle, which was proposed to explain word order in binomials (Iliev & Smirnova, 2016).[2] In the next section we compare the two phenomena and formulate our hypothesis about the effect of the subjective distance principle on inverted adjective ordering.

The Subjective Distance Principle in Binomials

Binomials are constructions with two conjoined elements belonging to the same lexical class, such as *Democrats and Republicans* (two nouns are conjoined) or *good and bad* (two adjectives are conjoined). While research on adjective ordering and word order in binomials has developed independently, there are surprising parallels between the two phenomena. First, word order in both domains is rather flexible, unlike word order in English in general. For example, while reversing the position of the subject and the verb results in purely ungrammatical constructions (**Slept John*), adjectives and binomials show more flexibility, despite the fact that there is often a clearly preferred word order. Thus, while the binomial *men and women* is more frequent, *women and men* is also possible (Iliev & Smirnova, 2016).[3] Similarly, in the domain of adjectives, *large blue car* sounds more natural than *blue large car*, but the latter is also possible.

Second, in both domains phonological factors might affect word order to some extent. For example, in binomials and adjectives, word length and the number of syllables appear to affect word order: the shorter word and the word with a lesser number of syllables tends to be mentioned first. This explains *bread and butter* and *boots and saddles* in binomials (Cooper & Ross, 1975: 79), and the order of adjectives in *the long intelligent book* (Wulff, 2003). Importantly, however, the phonological rule explains some of the data, but reference to semantic and pragmatic constraints, which in turn are seen as a reflection of deeper psychological principles, is needed in both domains.

One explanation for the word order in binomials is the *subjective distance* principle proposed by Iliev and Smirnova (2016). According to this principle, the attributes that are psychologically closer to the speaker – more

[2] Not to be confused with subjectivity in Scontras et al. (2017), discussed in the previous section.

[3] Binomials with relatively flexible word order, such as *men and women*, should be distinguished from the so-called freezes, where the order is fixed, as in *here and there* (cf. the ungrammatical **there and here*).

desirable, more familiar, or closer to the identity of the speaker more generally – will tend to be mentioned first in binomials (cf. Cooper & Ross, 1975). A series of studies confirmed this prediction in the domain of consumer preferences, political orientation, religion, gender, race, and geographic locations. For example, the analysis of the corpus of senate speeches showed that in the domain of political orientation, liberals are more likely to use *Democrats and Republicans*, thus mentioning their own political party first, while conservatives prefer the reversed word order: *Republicans and Democrats*. In another study, Iliev and Smirnova (2016) analyzed the distribution of gender words in binomials, looking at the literary work of more than 6000 authors. They found that female authors, when compared to male authors, tended to mention words referring to females first, as in *sister and brother*, *women and men*, and *daughter and son*. The distribution of gender words in binomials is particularly illuminating as it shows how the subjective properties of the speakers can override the default or more common word order, such as *men and women*.

We hypothesize that the subjective distance principle can also explain some cases of inverted word order in adjectival sequences. Specifically, we predict that the attribute that refers to a more desirable property according to the speaker would be mentioned first. The two studies below test this hypothesis for written and spoken modality.

Experimental Studies

Study 1: Adjective Order in Written Language

Participants Twenty-one participants were recruited from Amazon Mechanical Turk web service. All participants indicated that they were native speakers of English. The average age was 38 years old (the youngest 19, and the oldest 69). 48% were male, and 52% were female. The participants were compensated for their participation.

Stimuli Each stimulus consisted of two adjectives followed by a noun. All nouns referred to common objects: shoes, table, scarf, bike, watch, cat, and restaurant. The adjectives within the same nominal phrase belonged to different semantic classes, e.g. color and material in the case of *brown suede shoes*. We intentionally avoided modifiers belonging to the same semantic class within a query, since it has been observed that members the same semantic class are not ordered with respect to each other. For example, both *clever brave man* and *brave clever man* are possible, where *brave* and *clever* belong to the same semantic class – human propensity (Dixon, 1982). Moreover, unlike Danks and Schwenk (1972), who used only color and size adjectives, we included adjectives belonged to different semantic classes, including color, material, size, origin, and composition.

Design and Procedure At the beginning of the study, the participants read a short story introducing the main protagonist, Jim. Jim was looking for an object or place online, and needed help formulating his search queries.

Next, participants saw 7 questions, each dedicated to a particular item that Jim was looking for. The seven items were Shoes, Table, Scarf, Bike, Watch, Cat, and Restaurant. Each item had two attributes, e.g. color and material for shoes. For each item, there were two conditions. In one condition (Condition A), the context of the story specified that one attribute was more important than another. In another condition (Condition B), the importance of the attributes was reversed. For example, in condition A for Shoes, the participants learned that the color (brown) is very important to Jim, but material (suede) is less important. In condition B for the same item, the material (suede) was very important and the color (brown) was negotiable. (See the Appendix for the exact formulations.)

The participants then saw two alternative formulations of a query. Each formulation mentioned the two attributes but in a different order, e.g. *brown suede shoes* and *suede brown shoes*. The participants were asked to choose the formulation that is more appropriate given the context. Each participant saw only one condition per item (between-subject design). The conditions and the choice of the order in which two alternative queries were presented were randomized. Table 1 shows the list of stimuli and the two alternative formulations for each query.

Table 1: List of stimuli and the default word order predicted by semantic theories.

Items	Two alternative formulations of a search query	Default order
Shoes	brown suede shoes suede brown shoes	✓
Table	large oak table oak large table	✓
Scarf	long wool scarf wool long scarf	✓
Bike	red aluminum bike aluminum red bike	✓
Watch	silver quartz watch quartz silver watch	✓
Cat	short-haired white cat white short-haired cat	✓
Restaurant	Indian vegetarian restaurant vegetarian Indian restaurant	✓

Results To analyze whether the order of adjectives depended on the importance of a particular attribute to the speaker, we used the following coding scheme: When the participants chose the query in which the most important attribute in a given context was mentioned first, their answer was coded as 1. The answer in which the less important attribute was mentioned first was coded as 0. For example, if the context specified that the color of the shoes was more

important than their material, and the participant chose the query in which the color preceded the material (*brown suede shoes*), the answer was coded as 1. If in the same context the participants chose the reverse order (*suede brown shoes*), the answer was coded as 0.

If adjective ordering is not dependent on the subjective importance of the attribute, and the same (default) word order is preferred across different conditions, then the participants' answers will be at the chance level. Specifically, if a participant in Condition A chose *brown suede shoes*, her answer is coded as 1; and if the participant in the B condition chose the same query, her answer is coded as 0. The mean of the two answers is 0.5. If, however, adjective ordering is affected by the subjective importance of the attribute, then the answers for each condition will be higher than the chance level.

Collapsing across items, there was a strong tendency for mentioning the most important attribute first (m=.80, SD=.21). The choices were significantly higher than the chance level, which was .5 ($t(20)$=6.63, $p < .001$, one-tailed). The results are shown in Figure 1. These results support our hypothesis that adjective order is dependent on the subjective preferences of the speaker.[4]

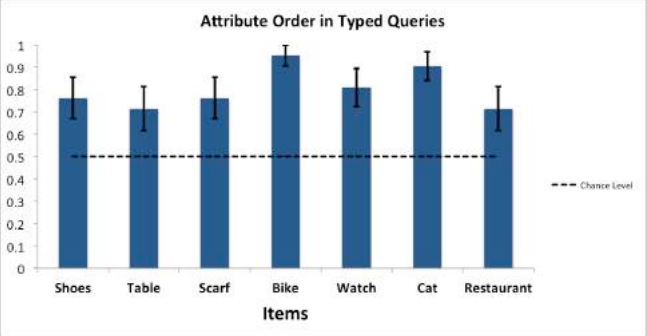


Figure 1: The proportion of times word order preferences are driven by the subjective importance of attributes in written language (typed queries). Higher numbers on the y-axis show greater association between adjectival order and the subjective distance. Values at the chance level would show that participants disregard subjective importance and chose the same word order in both conditions. Error bars represent +/-1SE.

[4] A reviewer raised the point that the inclusion of the congruent condition (canonical order and importance) is not informative. In our design, the congruent condition serves as a control for the incongruent condition. It might be the case that the canonical order expected by the researcher is incorrect, or that there is a substantial variance in the preference for canonical order among subjects. By averaging across the congruent and incongruent choices we control for that risk, so that the deviation of mean choices higher than .5 could safely be interpreted as importance preference, and mean choices lower than .5 would indicate reversed importance preference.

While the results from Study 1 provide support for the hypothesis that the subjective preferences of the speakers affect word order of adjectives, they are limited to a particular modality – written language. In Study 2 we test whether the same principle holds for spoken language. This question becomes particularly important in light of the criticism of the early pragmatic approaches about the role of intonation.

Study 2: Adjective Order in Spoken Language

Participants Thirty participants were recruited from Amazon Mechanical Turk web service. All participants indicated that they were native speakers of English. The average age was 32 years old (the youngest 23, and the oldest 53). 67% were male, and 33% were female. The participants were compensated for their participation.

Stimuli We used the same adjectives and nouns as in Study 1. Unlike Study 1, all stimuli were presented in audio format. The stimuli were read by a male native speaker of English. Each attribute within a query was read with even intonation, and there were no contrastive stress or juncture between attributes. This design intentionally separates intonation from word order, and thus can help us to assess the criticism that the inverted word order alone is not sufficient to convey the importance of the attribute in a given context (Richards, 1975).

Design and Procedure The study had the same design as Study 1, except that this time participants had to click on a button to hear a search query. As in Study 1, the order of the conditions and the order of the stimuli were randomized. Each participant saw only one condition per item.

Results We used the same coding scheme as in Study 1: all answers in which the order of the attributes matched the context, i.e. the most important attribute in a given context was mentioned first, were coded as 1. The answers in which the most important attribute was mentioned second were coded as 0. As in Study 1, we found strong preference for the most important attribute to be mentioned first (m=.68, SD=.35). The answers differed significantly from the chance level ($t(29)$=2.77, p=.004). The results of Study 2 are shown in Figure 2.

The results of Study 2 confirmed our findings in Study 1. We controlled for intonation and prosodic features and found that word order of adjectival modifiers reflects subjective preferences of the speaker, which is seen as manifestation of the subjective distance principle. The subjective distance principle extends to both written and spoken domains.

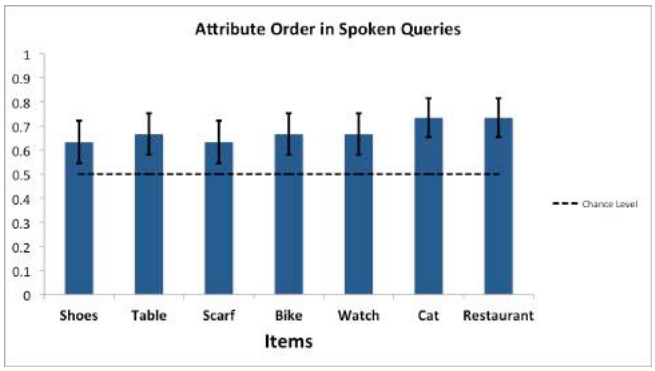


Figure 2: The proportion of times word order preferences are driven by the subjective importance of attributes in spoken stimuli. Error bars represent +/-1SE.

General Discussion

Our paper makes empirical and theoretical contributions to research on adjectival word order, and inverted order, specifically. First, our study shows that inverted word order is not limited to contexts with previously established salient referents, as in the original studies by Danks and co-authors. Second, we demonstrate that the inverted word order is manifested in both spoken and written domains. Our experimental design in Study 2 divorces intonation from word order, and we find that word order alone is meaningful and can convey the value of a particular attribute to the speaker, contra Richards (1975). Third, we propose that the inverted word order can be accounted for by the same psychological principle that explains word order in binomials. If a particular attribute, e.g. material, is more important to the speaker than color, this attribute would be mentioned first and would be positioned further away from the noun. Our explanation connects two previously unrelated research domains: binomials and adjectival modifiers.

One important question raised by a reviewer pertains to the applicability of the subjective distance principle to languages with post-nominal adjectives. It is worth to point out that the default ordering preferences based on a semantic principle are reversed in such languages. Specifically, the adjectives that tend to be mentioned first in languages with pre-nominal modifiers are usually mentioned last in languages with post-nominal modifiers. Despite the differences in word order, the distance between the head noun and the adjectival modifier remains more or less the same (Hetzron 1989; Scontras et al. 2017). Whether the subjective distance principle is also reversed in languages with post-nominal adjectives, is a question for future research.

Unlike Danks and his co-authors, we do not assume that the default and inverted word order should be explained by the same principle. It is plausible that the default word order can be explained with the semantic principle, such as adjective's subjectivity, as Scontras et al. (2017) argue. On the other hand, the inverted word order, at least in some cases, can be explained by the subjective distance principle, and the importance of a particular attribute to the speaker, specifically, as we show here. That the default and inverted word orders are explained by different principles is not surprising and is consistent with a more general observation in the literature that one principle, phonological, semantic, or pragmatic, is not sufficient to explain word order phenomena (Benor & Levy, 2006; Cooper & Ross, 1975 on binomials, Wulff, 2003 on adjective ordering).

Acknowledgments

This research was supported in part by SF State Research and Scholarly Activity Fund awarded to Anastasia Smirnova.

Appendix: Stimuli

Study 1: General Instructions

Jim has just moved to a new city and is now in the process of settling down. He looks to buy several items online, and he also plans to use online information to find certain places in his new hometown. However, Jim is not sure how exactly to formulate his queries, and he needs your help deciding which query would be more effective. In what follows, you will see the description of the items that Jim is looking for. You need to help him choose which of two alternative queries he should use.

Specific Instructions: Shoes – Condition A

Jim is looking for a pair of shoes. He would prefer a pair that is made of suede and is brown. He is firm about the material – he wants suede and not leather – but he can compromise on the color. If he finds a pair he likes, and it's in black instead of brown, he might still take it. If he can enter only one query in the search box, which query should he choose? (The participants were then presented with two alternative formulations of a query).

Specific Instructions: Shoes – Condition B

Jim is looking for a pair of shoes. He would prefer a pair that is made of suede and is brown. He is firm about the color – he wants brown and not black shoes – but he can compromise on the material. If he finds a pair he likes, and it's in leather instead of suede, he might still take it. If he can enter only one query in the search box, which query should he choose? (The participants were then presented with two alternative formulations of a query).

References

Benor, S., & Levy, R. (2006). The chicken or the egg? A probabilistic analysis of English binomials. *Language*, *82*, 233-278.

Cinque, G. (2014). The semantic classification of adjectives: A view from syntax. *Studies in Chinese Linguistics, 35*, 1-30.

Cooper, W., & Ross, J. (1975). World order. In R. E. Grossman, L. J. San, & T. J. Vance (Eds.), *Papers from the parasession on functionalism*. Chicago: CLS.

Danks, J. H., & Glucksberg, S. (1971). Psychological scaling of adjective orders. *Journal of Verbal Learning and Verbal Behavior*, *10*, 63-67.

Danks, J. H., & Schwenk, M. A. (1972). Prenominal adjective order and communication context. *Journal of Verbal Learning and Verbal Behavior*, *11*, 183-187.

Dixon, R. M. W. (1982). *Where have all the adjectives gone?* Berlin: Mouton.

Hetzron, R. (1978). On the relative order of adjectives. In H. Seller (Ed.), *Language universals*. Tübingen: Narr.

Iliev, R., & Smirnova, A. (2016). Revealing word order: Using serial position in binomials to predict properties of the speaker. *Journal of Psycholinguistic Research*, *45*, 205-235.

Kotowski, S., & Härtl, H. (2019). How real are adjective order constraints? Multiple prenominal adjectives at the grammatical interfaces. *Linguistics*, *57*(2), 395-427.

Martin, J. E. (1969). Semantic determinants of preferred adjective order. *Journal of Verbal Learning and Verbal Behavior*, *8*, 697-704.

Martin, J. E., & Ferb, T. E. (1973). Contextual factors in preferred adjective ordering. *Lingua*, *32*, 75-81.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive grammar of the English language*. New York: Longman.

Richards, M. M. (1975). The pragmatic communication rule of adjective ordering: A critique. *The American Journal of Psychology*, *88*, 201-215.

Scontras, G., Degen, J., & Goodman, N. D. (2017). Subjectivity predicts adjective ordering preferences. *Open Mind*, *1*, 53-66.

Scott, G.-J. (2002). Stacked adjectival modification and the structure of nominal phrases. In G. Cinque (Ed.), *The cartography of syntactic structures, Volume 1: Functional structure in the DP and IP* (pp. 91-120). Oxford, UK: Oxford University Press.

Sproat, R., & Shih, C. (1991). The cross-linguistic distribution of adjective ordering restrictions. In C. Georgopoulos & R. Ishihara (Eds.), *Interdisciplinary approaches to language: Essays in honor of S.-Y. Kuroda* (pp. 565-593). Dordrecht, Netherlands: Kluwer Academic.

Truswell, R. (2009). Attributive adjectives and nominal templates. *Linguistic Inquiry*, *40*, 525-533.

Wulff, S. (2003). A multifactorial corpus analysis of adjective order in English. *International Journal of Corpus Linguistics*, *8*, 245-282.

It's Alive! Animate Sources Produce Mnemonic Benefits

Sean Snoddy (ssnoddy1@binghamton.edu)
Daniel C. Silliman (dsillim1@binghamton.edu)
Joseph C. Wilson (jwilso10@binghamton.edu)
Kenneth. J. Houghton (khought2@binghamton.edu)
Deanne L. Westerman (wester@binghamton.edu)

Department of Psychology, Binghamton University (SUNY)
Binghamton, NY 13902 USA

Abstract

The mnemonic benefits of animate (e.g., Tiger) over inanimate (e.g., Table) stimuli have been demonstrated across several different memory paradigms. Given the ubiquity of inanimate, computer-generated voices we investigated if the animacy of a presentation source confers mnemonic benefits. We asked: is information delivered by a human voice better remembered than information presented by a computer-generated voice? Word-lists were presented auditorily by either a human or a computer-generated voice and memory was measured using a free recall assessment. In Experiment 1, words presented in a human voice were better remembered than words presented in a computer voice. Experiment 2 demonstrated that beliefs about the animacy of a computer-generated voice were not sufficient for any benefits to accrue, suggesting a possible boundary condition for the effect. Both experiments replicated the mnemonic benefits of animate words and demonstrated further extensions of the effect to spoken word presentation.

Keywords: Animacy; Recall; Memory

Introduction

Evolutionary psychologists have long argued that our minds have been adapted through the forces of natural selection (Cosmides & Tooby, 1994). Extending this evolutionary logic, it is further argued that our memory system has been adapted to serve the purposes of surviving in our distant ancestral environments. A recent example of this work would be the effect of "survival processing" by which mnemonic benefits are observed for stimuli experienced in evolutionarily salient contexts (Nairne, Thompson, & Pandeirada, 2007). Another example of adaptive memory is the finding of superior memory for animate compared to inanimate stimuli (Bonin, Gelin, & Bugaiska, 2014; Nairne, VanArsdall, Pandeirada, Cogdill, & LeBreton, 2013; VanArsdall, Nairne, Pandeirada, & Cogdill, 2015).

The *animacy effect* (henceforth *item-animacy*) has been observed in several memory paradigms such as free recall (Nairne et al., 2013), paired-associate recall (VanArsdall et al., 2015), and recognition (Bonin et al., 2014). Nairne and colleagues (2013) posit that our memories would be better attuned to animate entities in the environment for several evolutionary reasons. These include the special threat that living entities can pose, the sustenance that they can provide, and their broad social utility given that interactions with other animate entities (e.g. humans) were crucial for survival and reproduction.

It is this last reason relating to human sociality that drives the current investigation. The central question considered here is: does the animacy of the source of information matter for memory performance? In our modern computer-age, we are constantly interacting with voices generated by computers. How does the perceived humanness of such voices affect our cognition? Could it be that information delivered by Siri would be remembered differently than information provided by an actual person? It's possible that the findings regarding the animacy effect might bear on such questions. To the extent that such computer voices are perceived as inanimate (or at least less animate), there is a possibility that our memories might be worse for the information produced by a computer voice.

This ostensible *source-animacy* effect might emerge due to possible animacy contamination mechanisms (Cogdill, Nairne, & Pandeirada, 2016; as cited in Nairne, VanArsdall, & Cogdill, 2017). For example, Nairne and colleagues (2017) had participants read sentences in which two objects come in contact with each other. Target inanimate words in these sentences are "touched" by either animate ("The mouse is touching the sled.") or inanimate ("The lamp is touching the bottle.") stimuli. They found superior recall performance for inanimate target words when they were "touched" by the animate stimuli as compared to inanimate ones. Nairne and colleagues suggested that the "law of contamination" (Rozin, Millman, & Nemeroff, 1986) may account for such effects, with the property of animacy being conferred contagiously to inanimate words. Therefore, it may be the case that words spoken by the human voice are "contaminated" by the animacy of the voice, thus conferring a benefit for their recall.

Another account points to the importance of the voice itself. The quality of humanness in auditory perception might be especially well-processed. Evidence suggests that, from infancy, there is a predilection for human speech over non-speech analogues (Vouloumanos & Werker, 2007). There is also precedent in the music literature regarding the importance of human vocality for memory. For example, melodies sung by humans are better remembered than instrumental melodies (Weiss, Trehub, & Schellenberg, 2012). The authors proposed that we are especially attuned to human timbres because of their biological significance.

Due to the paucity of research on the subject and the implications for our interactions with machines in daily life, the current study was undertaken. The present experiments employed a free recall test on stimuli delivered through the auditory modality. The central manipulation involved the animacy of the voice delivering the word lists to be recalled (human vs. computer-generated). There were three main objectives: 1) To examine the influence of animate vs. inanimate sources on recall, 2) To provide a direct reproduction of the standard item-animacy effect in an auditory modality with a free recall assessment (see Aslan & John, 2016 for a paired-associate animacy effect using the auditory modality; see Stori, Zaar, Cooke, & Mattys, 2018 for a recognition memory assessment), and 3) To explore whether there would be an interaction between item-animacy and source-animacy. Following the evolutionary reasoning of Nairne and colleagues (2013), superior recall should be evidenced for words delivered by the human (animate) voice.

Experiment 1

The aim of Experiment 1 was to extend the classical item-animacy effect to an auditory source paradigm. Past research exploring animacy has typically consisted of the visual presentation of word lists that included animate and inanimate words (Nairne et al., 2013).The key departure from many past studies is that these lists are presented aurally through two different voices to manipulate source-animacy (cf. Aslan & John, 2016) along with a free recall assessment (cf. Stori et al., 2018). One of these voices was human and the other was computer-generated. Based on the animacy and evolutionary literature, memory should be superior for those words presented by the human-voiced (animate) compared to the computer-voiced (inanimate) source. Furthermore, this paradigm should replicate the standard item-animacy effect.

Method

Participants Binghamton University undergraduates (N = 51) participated in this study. An additional participant did not complete the entire experiment and was excluded.

Materials and Design Thirty-six English words were used in this experiment (18 animate, 18 inanimate). Thirty-two of these words (17 animate, 15 inanimate) were selected from word lists used in Nairne and colleagues (2013) and VanArsdall and colleagues (2015). An additional 1 animate and 3 inanimate words were obtained from the MRC database (Wilson, 1988) to supplement the lists. Following Nairne and colleagues (2013), all words were concrete nouns and matched on several dimensions: age of acquisition (19 words were missing data), number of letters, familiarity, imageability, concreteness, Kučera and Francis written frequency and number of categories, and mean Colorado meaningfulness.

Two versions of each word were recorded using version 2.1.3 of Audacity® (Audacity, 2014). The human spoken words were recorded by an experimenter that read each word aloud into the built-in microphone of an Apple Macbook laptop computer. A second Macbook computer was used to recreate the same set of words voiced by a computer using the voice-over accessibility function that comes standard with Apple computers and recorded via the built-in microphone of the first Macbook. The result of each recording was a continuous WAV file for each human- and computer-voiced word list. These continuous files were edited into discrete WAV files in Audacity for all of the words in both human- and computer-voiced presentations. All words were adjusted to have comparable volumes in both the human- and computer-voiced conditions (range: 9-15 dB). As a pilot test for clarity in the presentation of the words, a research assistant listened to both human- and computer-voiced presentations of all words and wrote them down. All words used in the present study were clearly perceptible to the research assistant, however, an additional four words were unclear and instead used as buffer words.

Two lists of intermixed animate and inanimate words were used for each experiment session and randomly assigned to either human- or computer-voiced conditions. Words were assigned to each list such that both human- and computer-voiced conditions were balanced on the aforementioned item-level variables. Two fixed buffer words were presented at the beginning of the first list and the end of the second list. Recall for these words were not coded nor included in the final analyses. PsychoPy psychophysics software version 1.8.3 (Peirce, 2007) was used to randomly select the word-to-list assignment, and to present each list in a randomly determined order.

Recall packets were printed on paper and included a maze (distractor task), a blank recall sheet, and a four-item questionnaire to assess the clarity and pleasantness of each word list using 7-point Likert scales.

Procedure For each session, between one and five participants were seated in a quiet room and told that they were participating in a memory experiment. The experiment was displayed on a 48 in. LCD television. Participants were presented with instructions both verbally and on-screen. They were instructed to face forward during presentation of the word lists and to focus on a black fixation cross on a white background. The words were not presented visually on the screen, but the display helped ensure that all participants were attending to the list presentation. Additionally, they were instructed to listen carefully and to expect a recall test later in the experiment.

Participants were then presented with the two word lists via speakers on the television. Each list was either presented using the human- or computer-voiced recordings in their entirety, with the alternative list being subsequently presented. Counterbalancing, which occurred across sessions, determined what source they heard first (human- or computer-voiced). The presentation of the first and the second list was separated by a 30 second break. Following the presentation of the second list, participants were handed a recall packet and directed to begin the maze distractor task. After one minute, participants were instructed to flip the page

and recall as many words as they could from both word lists. Participants were given an unlimited amount of time for recall but were told that they could turn the page if they could not remember any more words. On the final page of the packet, participants were asked to indicate the level of clarity and pleasantness of each source-animacy condition.

The following criteria were used to score a participant response as a correct recollection: correctly spelled target words (e.g., 'rabbit'); incorrectly spelled, but closely approximated target words (e.g., 'rabit'); different forms (i.e., tense, plurality) of target words (e.g., 'rabbits'). Responses that were confusable with a non-target word (e.g., 'rabid') were not counted as a correct recollection.

Results and Discussion

There were two main predictions about recall. First, that the animacy effect would be replicated with an auditory presentation of word lists and free recall assessment. Specifically, recall performance would be higher for animate words than inanimate words. Second, it was predicted that presentations voiced by a human should lead to better recall than presentations from a computer. A repeated measures ANOVA that contained item-animacy and source-animacy tested these predictions. There was a main effect of item-animacy such that animate words (M = .298, SD = .180) were recalled at a higher rate than inanimate words (M = .192, SD = .159), $F(1, 49) = 13.557$, $p < .001$, $\eta_p^2 = .217$. There was also an effect of source-animacy, where human-voiced words (M = .288, SD = .158) were recalled at a higher rate than computer-voiced words (M = .203, SD = .186), $F(1, 49) = 21.401$, $p < .001$, $\eta_p^2 = .304$. There was no significant interaction between item- and source-animacy ($F < 1$). (See Figure 1).

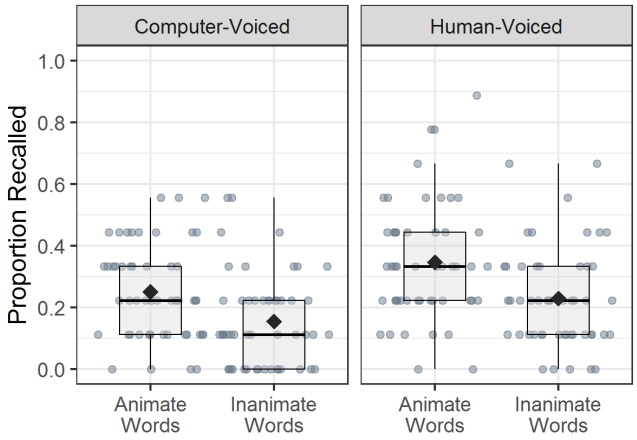


Figure 1: Proportion of words recalled in Experiment 1. The left panel presents both animate and inanimate words in the computer-voiced source animacy condition while the right panel reflects the human-voiced condition. Each point represents a participant's proportion of words recalled. Diamonds represent the overall mean for each condition.

Cumulative-link regression models were used to assess if human-voiced words were perceived as clearer and more pleasant than computer-voiced words. Each model predicted the rating of interest with source-animacy, the presentation order of human- and computer-voiced sources, and their interaction. Human-voiced words were rated as clearer than computer-voiced words (β = 1.873, SE = 0.525, Wald Z = 3.569, $p < .001$). There was no significant difference in ratings based on the order in which human- and computer-voiced sources were presented (β = -0.585, SE = 0.513, Wald Z = -1.140, p = .254) and no significant interaction (β = 0. 432, SE = 0. 713, Wald Z = 0.606, p = .544). Likewise, the human-voiced source received significantly higher pleasantness ratings than the computer-voiced source (β = 2.399, SE = 0.548, Wald Z = 4.378, $p < .001$). Again, there were no significant differences in ratings based on the order in which human- and computer-voiced sources were presented (β = 0.852, SE = 0.537, Wald Z = 1.587, p = .112), and no significant interaction (β = -0.871, SE = 0.727, Wald Z = -1.198, p = .321). (See Figure 2).

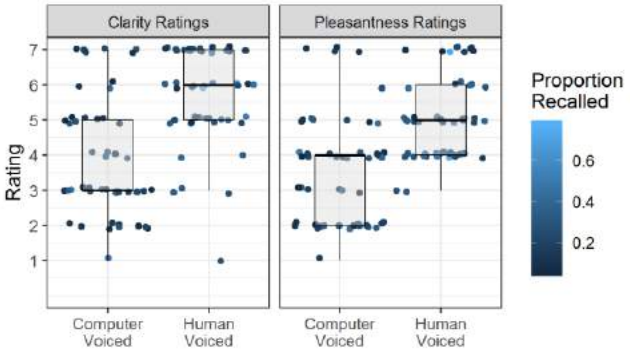


Figure 2: Clarity (left) and pleasantness (right) ratings for each source-animacy condition. Each point represents an individual participants' rating. The shading of each point reflects that participant's proportion of successfully recalled words within each condition. While both clarity and pleasantness ratings differed between source-animacy conditions there was no relationship between ratings and recall.

Mixed-effects logistic regression models that predicted recall success of each word tested if a series of control variables could account for either animacy effect. The baseline model included participants as random intercepts, source-animacy as random slopes, and source-animacy, item-animacy, and their interaction as fixed effects. Control variables were individually entered into the baseline model as an additional fixed effect. Clarity ratings (β = -0.075, SE = 0.050, Wald Z = -1.507, p = .132), pleasantness ratings (β = -0.058, SE = 0.055, Wald Z = -1.07, p = .287), the list participants received (β = 0 .035, SE = 0.171, Wald Z = 0.203, p = .839), and counterbalance order of source-animacy conditions (β = -0.105, SE = 0.167, Wald Z = -0.630, p = .529) were not predictive of recall successes. None of these variables altered the significance of item- and source-

animacy effects, or significantly improved the model's ability to account for variance in recall (all *p*s > .14).

Experiment 1 replicated and extended the standard item-animacy effect by demonstrating that animate words were better recalled than inanimate words when using auditorily presented stimuli. Animacy effects were not just observed for items, but also for the sources that presented items. This is reflected by the source-animacy effect: words presented by an animate, human voice were better remembered than items presented by an inanimate, computer-generated voice. The human-voiced source was rated as clearer and more pleasant than the computer-voiced source, however, follow-up analyses revealed that differences in ratings between the presentation source conditions could not account for the source-animacy effect. Taken together, these results provide evidence of the systemic effects of animacy on human memory.

Experiment 2

A potential limitation of the previous experiment was that there may have been differences between the human and computer voice that were not controlled for and that are not related to animacy or evolutionary mechanisms. One such difference was the human-voiced source being rated as clearer than the computer-voiced source (although clarity was not found to be predictive of recall success). The difficulty in controlling human and computer voices across relevant dimensions such as familiarity, tonality, and articulation (which may all contribute to clarity) raised the question of whether the source-animacy effect is contingent upon these differences (i.e. it is due to intrinsic qualities of the human voice) or participants' beliefs about the animacy of the source. To address this question, Experiment 2 circumvented the issue of auditory differences entirely. Instead of two different voices, the words in Experiment 2 are all delivered by one computer-generated voice. While the source (i.e. the voice) was held constant for both conditions, the belief regarding the animacy of the source was manipulated between conditions. Those in the stated-computer condition were told that the voice is computer-generated, while those in the stated-human condition were told that the voice is human.

Instead of serving as a direct replication of Experiment 1, the present experiment tested two hypotheses about the source-animacy effect. The belief-based hypothesis states that the source-animacy effect is determined by participants' belief about animacy independent of the auditory signal. This hypothesis predicts that when participants are presented with a computer-voiced source and their belief in the animacy of the source is manipulated, a source-animacy effect will be observed between the stated-human and stated-computer conditions. The intrinsic qualities hypothesis states that the source-animacy effect is determined by intrinsic qualities of the source. This hypothesis predicts that when presented with a computer-voiced source, no differences between the stated-human and stated-computer conditions will emerge as they are listening to the same computer-generated auditory signal.

Method

Participants Binghamton University undergraduates (N = 95) participated in this experiment. Two additional participants were dropped due to technical problems.

Materials and Design The word stimuli were the same as those used in Experiment 1 except that the buffer words were omitted. The audio was produced using Natural Readers online software, a text-to-speech tool[1]. All words recorded for this experiment were produced using a single computer voice from this software, which resembled a British-accented male. The procedure used to convert each word into an audio file was identical to Experiment 1, except that a human voice was not also recorded.

The WAV files for each word were presented through PsychoPy software in a random order to each participant. A between-subjects presentation of the words was used, such that all 36 words were presented to each participant through the one computer voice—the only difference being whether the participant was told that the voice was human or a computer program. In this way, all participants heard the same audio, ensuring that there were no aural or linguistic confounds between the animate and inanimate conditions.

Procedure Participants were randomly assigned to either the stated-computer or stated-human condition. In the stated-computer condition, the participants were told that each word was produced by a computer and in the stated-human condition, they were told that the words were produced by a human.

Each participant was brought individually into a room and told that they would be participating in an experiment that would require them to judge the clarity of a series of words that were to be used as part of a later experiment, which they would *not* be participating in on that day. These clarity judgments served as an incidental encoding task that was followed by a surprise free recall test that immediately followed the clarity judgment task. Participants were provided with closed ear headphones to listen to the words.

Each word of the study list was presented aurally through the headphones in a randomized order across the 36 trials. During each trial, a fixation cross appeared on the screen to focus their attention while the words were presented. A clarity rating scale replaced the fixation cross at the onset of each word. The participant would render their clarity rating on a 5-point Likert scale, with the wording being different according to the condition they were in ("Please rate how clear this human/computer produced word is", 1= not at all clear, 5 = extremely clear). Selecting a rating on the scale would begin the next trial (i.e. the following word).

[1] Navigate to https://www.naturalreaders.com/online/ to access the text-to-speech tool. The voice used was Peter at -1 speed.

At the end of the clarity judgment phase, participants were asked to recall as many words as possible from the list they just heard. They were given an unlimited amount of time to type their responses into an array of boxes that appeared on the screen. Once they completed this recall session, participants were asked how much they believed in the story they were told in the beginning of the experiment as a manipulation check. Those in the stated-computer condition were asked the extent to which they believed the voice they heard came from a computer, while those in the stated-human condition were asked how much they believed the voice to be from a human. Participants were probed about their beliefs on a 5-point scale. The criteria for a correct recollection were the same as in Experiment 1.

Results and Discussion

Data were first analyzed using a two-way ANOVA, with item-animacy as a within-subjects factor and source-animacy as a between-subjects factor. In line with our predictions, the standard animacy effect was replicated in this analysis as a main effect for item-animacy, $F(1, 93) = 52.008$, $p < .001$, $\eta_p^2 = .359$, with a greater proportion of animate words ($M = 0.25$, $SD = .10$) recalled than inanimate words ($M = 0.16$, $SD = .10$). However, there was no main effect found for source-animacy ($F < 1$). No interaction was found between item- and source-animacy ($F < 1$). (See Figure 3). Given the lack of an effect of source-animacy and the intrinsic qualities hypothesis's prediction of a null result, the stated-computer ($M = .202$, $SD = .129$) and stated-human ($M = .216$, $SD = .107$) conditions were analyzed with a Bayesian independent samples t-test. The Bayes Factor indicated substantial support (Jefferies, 1961) for the null hypothesis, (i.e., no differences between conditions), $BF_{01} = 4.425$.

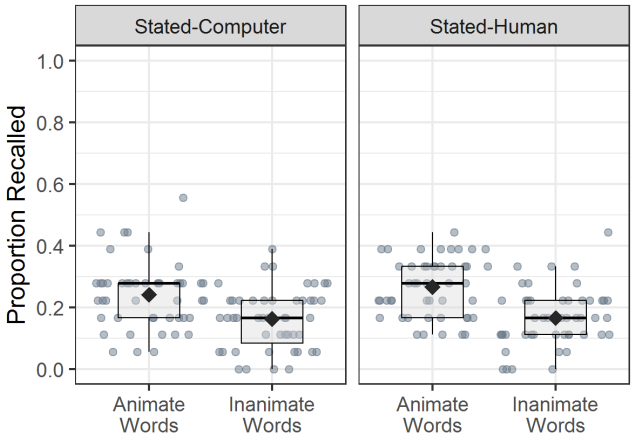


Figure 3: Proportion of words recalled in Experiment 2. The left panel presents both animate and inanimate words in the stated-computer source animacy condition while the right panel reflects the stated-human condition. Each point represents a participant's proportion of words recalled. Diamonds represent the overall mean for each condition.

Cumulative-link regression was used to test if clarity ratings differed as a function of both item- and source-animacy. Animate words were judged as clearer than inanimate words ($\beta = 0.338$, $SE = 0.063$, Wald $Z = 5.363$, $p < .001$). Despite both source conditions receiving identical stimuli, there was a significant difference in perceived clarity across the two source conditions, such that participants in the stated-computer condition judged the words they heard as clearer than those in the stated-human condition ($\beta = 0.297$, $SE = 0.063$, Wald $Z = 4.714$, $p < .001$). Regarding the manipulation check, the analyses revealed no significant differences between source-animacy conditions in the extent that participants believed the cover story ($\beta = 0$, $SE = 0.2637$, Wald $Z = 0$, $p = .999$). Participants who were told that the items were produced by a computer accepted this story to a similar degree as those who were told the voice was human. The median belief across conditions ($Mdn = 3$) suggests a moderate belief in the manipulation, with perhaps some degree of uncertainty.

Mixed-effects logistic regression models that predicted recall success of each word were used to test if any control variables could account for the observed item-animacy effect. The baseline model included participants as random intercepts and source-animacy, item-animacy, and their interaction as fixed effects. Source-animacy was not allowed to vary as random slopes, as in Experiment 1, because it was not a significant predictor of recall and did not alter the subsequent pattern of results. Control variables were individually entered into the baseline model as a fixed effect. Clarity ratings for each item ($\beta = 0.268$, $SE = 0.046$, Wald $Z = 5.794$, $p < .001$) were a significant predictor of recall success, such that recall was more likely for items with higher clarity ratings. While including clarity ratings into the model did not alter the observed item-animacy effect, the model did account for significantly more variance in recall than the baseline model, $\chi^2(1, N = 1) = 36.615$, $p < .001$. Participants' belief in the cover story ($\beta = 0.063$, $SE = 0.05$, Wald $Z = 1.233$, $p = .217$) was not a significant predictor of recall, did not alter the significance of the item-animacy effect, and did not significantly improved the model's ability to account for variance in recall, $\chi^2(1, N = 1) = 1.506$, $p = .22$.

The present experiment failed to find evidence of a source-animacy effect. It is important to note that participants were not actually exposed to an animate source, and instead those in the stated-human condition were told an inanimate source was animate. This result provides support for the hypothesis that some intrinsic qualities of the auditory signal are necessary for a source-animacy effect to accrue and suggests a boundary condition for the source-animacy effect: beliefs about the animacy of sources alone do not confer mnemonic benefits. This appears congruent with the evolutionary argument that the human voice has a special status in information processing, which may have been selected for by similar evolutionary forces that gave rise to the item-animacy effect. The present experiment provided an additional replication of the item-animacy effect within both an auditory presentation modality and an incidental encoding task.

General Discussion

The key finding of Experiment 1 was that words presented by the human-voiced source were better remembered than words presented by the computer-voiced source. This novel result suggests that the animacy of the source, and not only of the word presented, influences recall. The item-animacy effect was also replicated in an auditory modality. While prior work has explored auditory presentation of nonwords paired with animate or inanimate characteristics (Aslan & John, 2016) or auditory presentation of items followed by a recognition memory test (Stori et al., 2018), the present extension of the item-animacy effect demonstrated that it can also be observed with auditorily presented words and a free recall assessment, which is consistent with the evolutionary explanation of the animacy effect (Bonin et al., 2014; Nairne et al., 2013), as human speech emerged before written communication.

Experiment 1 demonstrated that animacy not only influences the memorability of items, but also the memorability of items presented by an animate source. One possible explanation of this source-animacy effect may be a contagion mechanism (Rozin et al., 1986), where the animacy of the source confers a mnemonic benefit to the information presented by it through association. A second possible explanation is that the human voice holds a special status in memory (Weiss et al., 2012), which may have been conferred through natural selection and may possibly extend to other animate sources. While the present experiments were not intended to disambiguate between these two explanations, future work should attempt to uncover its underlying mechanism.

Though not a direct replication of Experiment 1, Experiment 2 also examined source-animacy using an auditory modality. This experiment explored whether the mnemonic benefit of animate sources is determined by belief about animacy independent of the auditory signal or if it is determined by intrinsic qualities of the auditory signal itself. To this end, participants were presented with a single voice and their belief about whether it was from a human-voiced or computer-voiced source was manipulated. No differences in recall were found under these conditions, which provides support for the intrinsic qualities hypothesis: the human voice may be necessary for the source-animacy effect to emerge and that belief about the source's animacy is not sufficient for the effect to emerge. The necessity of the human voice may arise from either perceptual expertise with human voices or a particular biological significance. Under this hypothesis, the computer-generated voice in Experiment 2 would be treated fundamentally differently than a human voice regardless of what participants are told, or believe, about the source. One possible alternative explanation to this is that the suggestion was not strong enough for participants in the stated-human condition to treat the computer-generated voice in the same way they would a human voice. A stronger suggestion could be provided to increase belief in the manipulation and possibly give rise to a source-animacy effect. With this alternative explanation in mind, future research is warranted to further disambiguate these possible accounts.

Both experiments included additional analyses to mitigate possible alternative explanations. While pleasantness and clarity differed between source-animacy conditions, they were not related to recall performance and could not explain either of the observed item- or source-animacy effects. The divergence in clarity ratings between source-animacy conditions were not related to recall, which is theoretically interesting. There is some research suggesting a desirable difficulty effect in memory such that difficult-to-perceive words are better remembered (Rosner, Davis, & Milliken, 2015). Besken and Mulligan (2014) provided evidence supporting the benefits of desirable difficulties by demonstrating that aurally-distorted words were better remembered on a free recall assessment than non-distorted words. Experiment 1 results showed, however, that although the computer-voiced words were judged as less clear, they were not better recalled, inconsistent with a desirable difficulty effect. It is possible that the source-animacy effect overwhelmed any benefits of perceptual dis-fluency.

The results of Experiment 2 further complicate the role of perceptual clarity. Participants in the stated-computer condition rated the words they heard as significantly clearer than those in the stated-human condition. This is despite that the voices used were identical in both conditions. Though the results of Experiment 2 suggest that while beliefs might play a negligible role in a possible source-animacy effect, they may influence how people judge perceptual clarity. Whatever the case, the results suggest that clarity differences between the voices cannot account for the mnemonic benefit of human-spoken words.

Despite the noteworthy finding of the source-animacy effect and the replication of the item-animacy effect in Experiment 1, it is necessary to consider some important limitations. First, the materials were recorded using a limited number of voices. In order to ensure that these findings are generalizable, future studies must use a wider variety of voices, both computer-generated and human. Second, Experiment 1 rested on an experimenter-defined source-animacy manipulation without testing whether participants viewed the human-voiced source as more animate than the computer-voiced source. Third, while the stimuli were tested for perceptibility by a single research assistant, it is possible that participants may have had more difficulty in perceiving each word. Future work would benefit from more robust norming of the animacy and perceptibility of the auditory sources. Fourth, as Experiment 2 was a between-subject manipulation, participants may have anchored their clarity ratings differently based on whether they were told it was from a computer or human, which may have biased the clarity ratings and obscured a potential relationship between clarity and source-animacy condition. To address this concern, future work should provide a fixed reference for participants to evaluate clarity with respect to. Finally, though clarity and pleasantness were found not to affect the main findings, there may have been other potentially nontrivial differences in vocal variables (e.g. tempo, pitch) that were not recorded or analyzed in the present experiments.

References

Aslan, A., John, T. (2016). The development of adaptive memory: Young children show enhanced retention of animacy-related information. *Journal of Experimental Child Psychology, 152,* 343-350.

Audacity [Computer Software]. (2014). Retrieved from https://www.audacityteam.org/

Besken, M., & Mulligan, N. W. (2014). Perceptual fluency, auditory generation, and metamemory: analyzing the perceptual fluency hypothesis in the auditory modality. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40(2),* 429.

Bonin, P., Gelin, M., & Bugaiska, A. (2014). Animates are better remembered than inanimates: Further evidence from word and picture stimuli. *Memory & Cognition, 42(3)*, 370-382.

Cogdill, M., Nairne, J., & Pandeirada, J. (2016). Enhanced retention for objects touched by agents. Poster presented at the 28th Annual Convention of the Association for Psychological Science, Chicago, IL.

Cosmides, L., & Tooby, J. (1994). Origins of domain specificity: The evolution of functional organization. In L. Hirschfeld, S. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture* (pp. 85-116). Cambridge, England: Cambridge University Press.

Jeffreys, H. (1961). *Theory of probability* (3rd Ed.). Oxford, UK: Oxford University Press.

Nairne, J. S., Thompson, S. R., & Pandeirada, J. N. (2007). Adaptive memory: survival processing enhances retention. Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(2),* 263-373.

Nairne, J. S., VanArsdall, J. E., Cogdill, M., (2017). Remembering the living: Episodic memory is tuned to animacy. *Current Directions in Psychological Science, 26(1),* 22-27.

Nairne, J. S., VanArsdall, J. E., Pandeirada, J. N., Cogdill, M., & LeBreton, J. M. (2013). Adaptive memory: The mnemonic value of animacy. *Psychological Science, 24(10),* 2099-2105.

Peirce, J. W. (2007). PsychoPy—psychophysics software in Python. *Journal of Neuroscience Methods, 162(1),* 8-13.

Rosner, T. M., Davis, H., & Milliken, B. (2015). Perceptual blurring and recognition memory: A desirable difficulty effect revealed. *Acta Psychologica, 160,* 11-22.

Rozin, P., Millman, L., & Nemeroff, C. (1986). Operation of the laws of sympathetic magic in disgust and other domains. *Journal of Personality and Social Psychology, 50(4),* 703-712.

Stori, D., Zaar, J., Cooke, M., & Mattys, S. L. (2018). Sound specificity effects in spoken word recognition: The effect of integrality between words and sounds. *Attention, Perception, & Psychophysics, 80,* 222-241.

VanArsdall, J. E., Nairne, J. S., Pandeirada, J. N., & Cogdill, M. (2015). Adaptive memory: Animacy effects persist in paired-associate learning. *Memory, 23(5),* 657-663.

Vouloumanos, A., & Werker, J. F. (2007). Listening to language at birth: Evidence for a bias for speech in neonates. *Developmental Science, 10(2),* 159-164.

Weiss, M. W., Trehub, S. E., & Schellenberg, E. G. (2012). Something in the way she sings: Enhanced memory for vocal melodies. *Psychological Science, 23(10),* 1074-1078.

Wilson, M. (1988). MRC Psycholinguistic Database: Machine-usable dictionary, version 2.00. *Behavior Research Methods, Instruments, & Computers, 20(1)*, 6-10.

The Director Task Fails to Differentiate Young Adult Theory of Mind Abilities: An IRT Analysis

Mikhail Sokolov (MishaSokolov@cmail.carleton.ca)
John Logan (JohnLogan@cunet.carleton.ca)
Department of Psychology, Carleton University
1125 Colonel By Drive Ottawa, ON K1S5B6 Canada

Abstract

The goal of the present study was to demonstrate the potential application of Item Response Theory (IRT) outside its traditional use in assessing questionnaires by applying it to data from behavioural task. We did this by validating a perspective taking task called the Director Task used to assess Theory of Mind (ToM) abilities in young adults. IRT and convergent validity analyses indicated that, contrary to our hypotheses, the Director Task had an unduly narrow range of responding for measuring ToM. Furthermore, the Director Task did not correlate with other established measures of ToM. Our results suggest that the task should be used with caution when assessing a young adult population. Furthermore, since convergent validity was not established, it is uncertain what specifically the task measures. Overall, we show how IRT may serve as a useful tool in evaluating behavioural measures.

Keywords: Theory of Mind, Item Response Theory, Director Task

Introduction

Item Response Theory is an approach to assessing the psychometric properties of measures designed to measure psychological constructs such as attitudes. Modern test construction methodology suggests that simply having a range of scores on a measure is not a sufficient determinant of the psychometric properties of a test. In the current research article, we extend the use of Item Response Theory (IRT) methodology from its traditional application of evaluating personality scales and achievement to validate the effectiveness of a behavioural task, specifically, a Theory of Mind task called the Director Task.

IRT provides sample invariant information for each item at varying levels of the underlying traits or ability (Embretson & Reise, 2000; Thissen & Wainer, 2001). The simplest IRT model is the dichotomous Rasch (1960) model which is applied to tests, or other tasks, where each trial can be classified as correct or incorrect. The Rasch model allows us to calculate the difficulty of each item (1PL), its discriminatory power (2PL), as well as account for the effect of guessing (3PL). By calculating the probability of answering each question correctly based on assumed trait levels, IRT can supply researchers with information about the suitability of individual test items, as well as the test in general. IRT provides a number of advantages over classical test construction methods, such as allowing for identification of sensitivity and difficulty of individual items (Embretson, 1996; Hambleton & Swaminathan, 2013). Most crucially, IRT allows researchers to empirically assess the suitability of the test at varying levels of the trait of interest. This information allows researchers to determine the effective range of discrimination for the tool.

IRT models make four major assumptions: unidimensionality, local independence, monotonicity, and a normally distributed latent trait. Unidimensionality of the trait and local independence are generally assumed to coexist. Unidimensionality is the assumption that there is only one latent trait being measured, whereas local independence is the assumption that each response is independent and only conditional on the latent trait. Monotonicity is the assumption that as the latent trait increases, so does the probability of correctly responding to each trial. Finally, the assumption of a normally distributed latent trait is common to many parametric tests used in psychology research. To our knowledge, IRT has never been applied to data from a behavioural task. However, there are, in principle, no conceptual restrictions that would restrict the use of IRT for the assessment of a behavioural measure.

Theory of Mind (ToM) is a cognitive ability that allows individuals to mentalize about other's minds (Heider, 1958). ToM is believed to be an important component of empathy which, along with emotion empathy, allows individuals to accurately recognize and understand other's emotional states (Smith, 2006). Disruptions in ToM abilities can lead to impairments in adult functioning where individuals are less able to interpret the beliefs and intentions of others (Perner, Frith, Leslie, & Leekam, 1989). Theory of Mind deficiencies are closely associated with Autism Spectrum Disorders (Baron-Cohen, Leslie, & Frith, 1985).

Unlike emotion perception, which is largely an inborn ability and therefore, develops extremely early (Grossmann, 2010), ToM abilities continue to develop beyond childhood. For example, infants can discriminate emotional faces at 3.5 months (Montague & Walker-Andrews, 2002), or possibly earlier, and at 6.5 months are able to differentiate between emotional postures of adults (Zieber, Kangas, Hock, & Bhatt, 2014a, 2014b). In contrast, ToM skills develop much later in life (Calero, Salles, Semelman, & Sigman, 2013; Frith & Frith, 2001). ToM development is even believed to stretch into early adulthood (Dumontheil, Apperly, & Blakemore, 2010), as evidenced by the continued neurodevelopment of brain regions responsible for ToM such as the medial frontal gyrus, the anterior paracingulate, and the right temporoparietal junction (Kana, Keller, Cherkassky, Minshew, & Just, 2009) into late adolescence and early adulthood (Shaw et al., 2008).

From a practical point of view, the assessment of ToM abilities poses a particular difficulty for clinicians and researchers. Many tasks that measure the development of ToM abilities, such as the presence of false beliefs or perspective taking, have ceiling effects since these abilities are well developed by the age of 5 (Wellman, Cross, & Watson, 2001). Other measures, such as the Reading the Mind in the Eyes Task (Baron-Cohen, Wheelwright, Hill, Raste, & Plumb, 2001), are confounded by the emotion perception aspect of the task. However, some perspective taking tasks, such as the Director Task (Keysar, Barr, Balin, & Brauner, 2000) have been shown to discriminate ToM abilities later into adolescence, and even early adulthood (Keysar, Lin, & Barr, 2003).

The Director Task is a perspective taking task where the participant is instructed to follow directions of a confederate who has a different view of a 4 x 4 grid. The grid contains various items that the participant must manipulate based on the director's instructions. Some of the grids are closed to the view of the director, but not the participant (see Figure 1). During the experimental trials of the task, the director gives an ambiguous instruction to the participant to move an item (e.g.: "Move the bottom block"). In this example, there would be two distractor blocks, one of which is the lower most from an egocentric perspective, but is closed off (i.e., unable to be seen) from the view of the director, and therefore is not the target. If participants select the lower-most block that is visible to them, they would not have taken the director's perspective into account, and would thus commit an error.

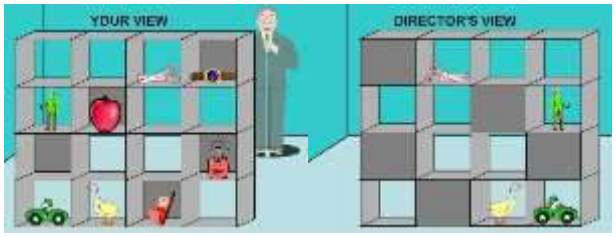


Figure 1. Instruction examples given to participants to demonstrate the director's perspective.

In previous studies, the Director Task showed that even adults have a natural tendency for the egocentric perspective (Keysar et al., 2003), and that the task reliably differentiates between youth and young adults (Dumontheil et al., 2010). These findings indicate that the Director Task may be a useful tool to differentiate between Theory of Mind abilities within the young adult/ adult population. If it is true that the task can reliably differentiate between young adults on ToM abilities, this would allow for the study of ToM perspective taking using convenience samples, making ToM research more accessible.

Present Study

The purpose of the present study was to assess the Director Task using IRT. Specifically, would the Director Task prove suitable for use with the young adult population as a tool for discriminating between individuals who are low and those who are high in Theory of Mind abilities? We hypothesize that a modified, computer based, version of the Director Task would allow for the discrimination across a sample of young adults on the basis of ToM abilities, and the results would show convergent reliability with more established measures of ToM. Although the Director task has already been shown to differentiate between age groups (Dumontheil et al., 2010), this finding does not automatically extend to within group differentiation.

With regard to convergent validity, two established ToM tasks were selected, the Reading the Mind in the Eyes Task ("Eyes Task") (Baron-Cohen et al., 2001) and the 40-item Empathy Quotient (EQ 40) (Baron-Cohen & Wheelwright, 2004). Although these tasks are sufficiently different from the Director Task, we predicted that a weak, but significant positive correlation would be observed between these tasks and the Dirtector Task.

Method

Participants

94 Carleton University undergraduate students (20 male) with a mean age of 19.8 (SD = 4.3) volunteered to participate in exchange for course credit. All participants self-identified as right-handed.

Measures

As part of a larger study participants completed the Eyes Task (Baron-Cohen et al., 2001), as well as the 40 item Empathy Quotient (Baron-Cohen & Wheelwright, 2004). The Director Task (Keysar et al., 2000) used was kindly provided by Dumontheil et al. (2010) and modified for use with PsychoPy software (Peirce, 2007). The Director Task was modified to exclude the non-director items, allowing a doubling of the number of Director trials to 95. Altogether, 16 trials were experimental trials, 16 trials were control trials, and the rest of the trials were filler trials. If our hypothesis is correct, by increasing the number of experimental trials, a grater range of scores will be observed, and with it, a finer discrimination of individuals along the latent trait associated with ToM.

Procedure

After providing informed consent, participants were tested individually in a sound-attenuated booth. Instructions, stimuli, and questionnaires were presented on a PC using PsychoPy software (Peirce, 2007). The task was presented to participants as a static image with verbal instructions played over computer speakers. For each trial the target item was overlaid by a 3 cm^2 invisible square which would record mouse button presses. All mouse presses outside of the target square were scored as incorrect; trials with no mouse button presses were discarded. Each Director Task maximum trial length was set to 5 seconds from the onset of audio instructions. Trials in the Director Task were presented to participants in a predetermined order. Next, participants completed the Eyes Task and the EQ 40 task.

Trials in these latter tasks were randomized. The study required approximately 45 minutes to complete. Participants were debriefed as to the purpose of the experiment after they completed the EQ 40 task.

Results

Responses were tallied and scored using custom Visual Basic scripts. Outliers were identified based on deviations from predicted Mahalanobis distance using the R package "careless" (Yentes & Wilhelm, 2018). One case was identified as unusual and removed leaving 93 participants (see Figure 2). Descriptive statistics are presented in Table 1. Scores from the Director Task appeared to take on a bi-modal distribution, with upper and lower scores trending towards extremes (see Figure 3).

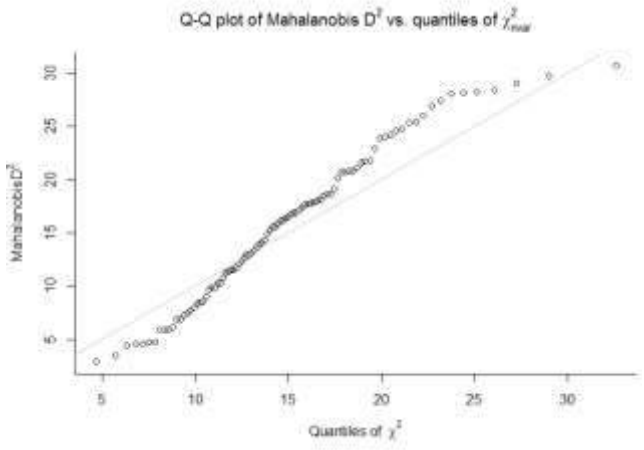


Figure 2. q-q plot of actual vs predicted Mahalanobis distance

Table 1. Overall descriptive statistics for each measure.

	M	SD
EQ 40 Score	68.21	7.80
Eyes Task	26.80%	5.39%
Director Task	53.14%	35.74%

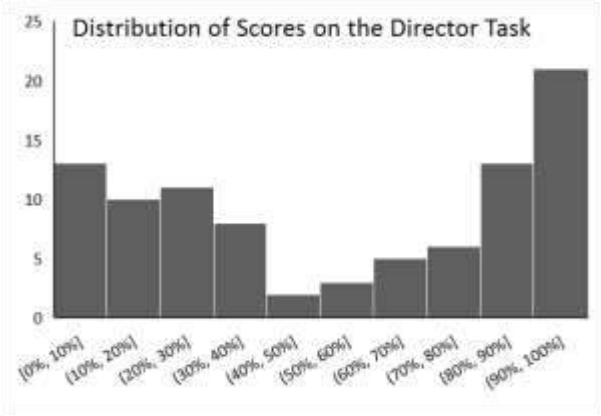


Figure 3. Histogram showing the distribution of accuracy scores on the Director Task

A paired samples t-test showed a significant decrease in accuracy when comparing the control trials with the experimental trials (t (92) = -9.05, $p < .01$) but not reaction times (t (92) = -1.14, *ns*). This suggests that performance on the task deteriorated as expected due to the increased difficulty of the experimental trials compared to the control trials.

IRT

The IRT analysis was performed using the ltm (Rizopoulos, 2006) package in the R environment (R Core Team, 2013). A constrained One-Parameter Logistic Model (1PL) and unconstrained Two-Parameter Logistic Model (2PL) dichotomous models was run to determine which created a better fit. The constrained model assumes that each item on the unidimensional scale is equally good at discriminating between individuals with varying trait levels whereas the unconstrained model does not make this assumption. Since the two models are nested, a χ^2 difference test was performed to assess model fit.

Significant model fit improvement was observed when the model was unrestricted from constrained to the unconstrained discrimination parameters (χ^2 (14) = 27.97, p = 0.014). As such, a 2PL model was selected for the analysis of the Director Task. A 3PL model was not used because the Director Task is not strictly a forced choice multiple choice test, and therefore it is improbable that participants would attempt to randomly select their answers.

Results of the individual item difficulty and discrimination, under the 2PL model, are presented in Table 2. Figure 4 contains the Item Information Curves (IIC) and Figure 5 Shows Total Test Information Function relative to Standard Error of measurement. Standard errors were estimated using the delta method.

The results from the model suggest that, congruent with our hypothesis, all the experimental trials of the Director Task have good discriminatory power. However, contrary to our hypothesis, the difficulty of the items appears quite low with only half of the items showing a significant deviation from 0.

Table 2. Difficulty and discrimination of the experimental items of the Director Task

Trial	Difficulty (*b*)	Discrimination (*a*)
4	-0.33	1.40**
14	-0.03	1.16**
20	-0.18	2.43**
26	-0.06	2.52**
30	-0.27*	2.66**
36	-0.34**	3.10**
40	-0.07	4.00**
49	-0.026	3.62**
59	-0.61**	2.22**

65	-0.40**	3.32**
70	0.17	2.25**
74	-0.67**	1.66**
78	-0.36**	3.72**
84	-0.18	2.21**
88	-0.44**	2.38**

Note: * $p < .05$; ** $p < .01$

The IIC plot visually confirms that, although the information content of many trials is very high, the range of ToM ability that they represent is poor.

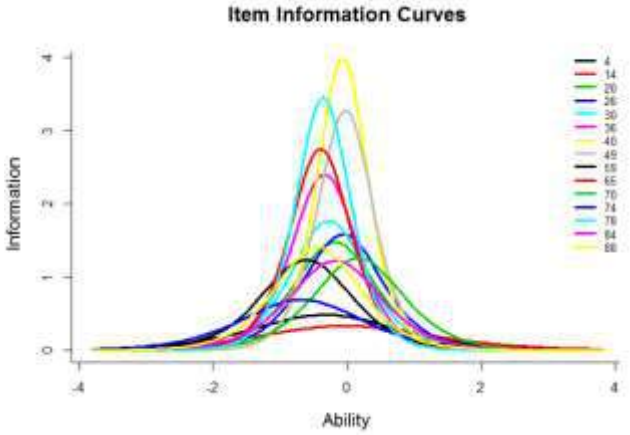


Figure 4. Item Information Curves for the experimental items of the Director Task

Finally, Figure 5 shows that the information content of the Director Task as a whole is very large, with an area under the curve of 38.66. However, 55% (20.75) of this information content falls within 0.5 standard deviations of the mean, and 82.5% (31.89) within 1 standard deviation. This once again reaffirms that the Director Task is poor at discriminating between individuals of different ToM abilities.

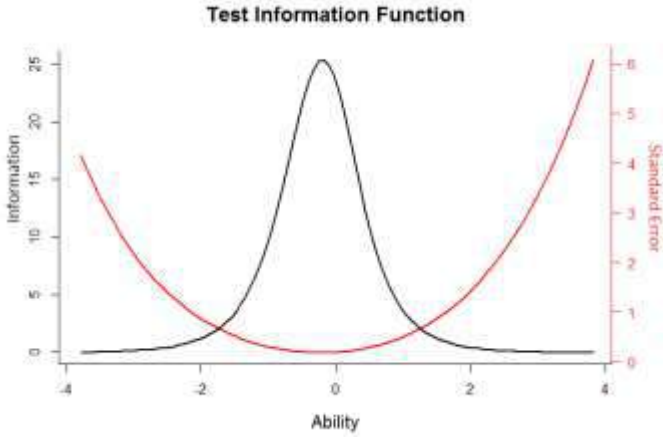


Figure 5. Total Test Information Function relative to Standard Error of Measurement.

Convergent Validity

Convergent validity for the Director Task were assessed using a self-report measure of ToM, the EQ 40, as well as a behavioural discrimination task, the Eyes Task. The results are presented in Table 3.

Table 3. Correlation matrix for the Director and other convergent validity tasks

	1	2
EQ 40	-	
The Eyes Task	-0.102	-
The Director Task	0.065	0.114

Contrary to our hypothesis, we did not find any significant correlations between the Director Task, or any of the other two popular tasks for assessing ToM abilities.

Discussion

Our findings did not support the hypothesis that the Director Task is good at discriminating between Theory of Mind abilities in a sample of young adults. Our findings are surprising in light of previous findings with the same (Dumontheil et al., 2010) or similar (Keysar et al., 2003) tasks allowing for discrimination in the young adult population.

Our sample showed significant variability in the range of scores on this task, which under normal circumstances would be an encouraging finding. However, IRT analysis showed that despite strong information content of the individual trials (discrimination), the Director Task does not measure well different levels of the ToM trait (difficulty). We interpret these findings as a strong indication that the Director Task is able to differentiate participants as either good or bad at TOM abilities, with little useful information beyond that. This interpretation is supported by both the poor difficulty gradient of the trials, as well as the tendency for participant scores to conform to a bimodal distribution.

Beyond the poor psychometric properties of the task, we failed to observe convergent validity between the Director Task and other established ToM tasks. This finding brings into question what trait or state the Director Task actually is measuring. One possible explanation for the lack of relationship between the three ToM Tasks examined in this study is that there is a sufficiently large distinction between the perspective taking ToM component, and emotion perception ToM component. However, this would not explain the lack of relationship between the Eyes Task and the EQ scores. Another possible explanation is that the Director Task is measuring some other quality, such as selective-attention to the task (Rubio-Fernández, 2017).

Regardless, we would caution researchers using the Director Task in its present form. Specifically, the task suffers from overly homogenous difficulty of trials. Nonetheless, there is potential for a modified version of this

task to be more successful. If the task is modified such that there is a greater range of experimental trial difficulty, with some being more difficult, while others being easier, the likely utility of the task will greatly improve. Finally, it is possible that by assessing other behavioural measures beyond the accuracy of answers, such as mouse-tracking or eye-tracking (Symeonidou, Dumontheil, Chow, & Breheny, 2016) we could use the extra sources of information to supplement our inferences about participants' ToM abilities.

Regarding the more general goal of extending IRT to assess the results of a behavioural task by validating the Director Task, the present results suggest that IRT can provide useful information about the relationship between participants' responses and the construction of tasks. IRT is often associated with pen and paper test construction, however, the underlying probability models are agnostic to the source of the data. With many available statistical packages, and a well developed literature, IRT is easily accessible to all researchers. We encourage the use of IRT as a readily available tool to aid the validation of measures.

Conclusion

In this study we used Item Response Theory to validate the Director Task (Keysar et al., 2000) as a tool in studying Theory of Mind abilities in young adults. Contrary to our hypotheses, we found that the task performed poorly in discriminating between levels of the latent trait. Furthermore, a convergent validity measure brought into question what latent trait is being measured using the Director Task. Overall, the present study provided a novel demonstration of how an Item Response Theory analysis can be profitably extended to assess behavioural measures.

References

Baron-Cohen, S., Leslie, A.M., & Frith, U. (1985). Does the autistic child have a "theory of mind" ? *Cognition, 21*(1), 37-46. doi: 10.1016/0010-0277(85)90022-8

Baron-Cohen, S., & Wheelwright, S. (2004). The empathy quotient: An investigation of adults with asperger syndrome or high functioning autism, and normal sex differences. *Journal of autism and developmental disorders, 34*(2), 163-175. doi: 10.1023/b:jadd.0000022607.19833.00

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The "reading the mind in the eyes" test revised version: A study with normal adults, and adults with asperger syndrome or high-functioning autism. *The Journal of Child Psychology and Psychiatry and Allied Disciplines, 42*(2), 241-251. doi: 10.1017/S0021963001006643

Calero, C.I., Salles, A., Semelman, M., & Sigman, M. (2013). Age and gender dependent development of theory of mind in 6-to 8-years old children. *FRONTIERS IN HUMAN NEUROSCIENCE, 7*(May), 281. doi: 10.3389/fnhum.2013.00281

Dumontheil, I., Apperly, I.A., & Blakemore, S.-J. (2010). Online usage of theory of mind continues to develop in late adolescence. *Developmental Science, 13*(2), 331-338. doi: 10.1111/j.1467-7687.2009.00888.x

Embretson, S.E. (1996). The new rules of measurement. *Psychological Assessment, 8*(4), 341.

Embretson, S.E., & Reise, S.P. (2000). *Item response theory for psychologists*. Mahwah, NJ: L. Erlbaum Associates.

Frith, U., & Frith, C. (2001). The biological basis of social interaction. *Current Directions in Psychological Science, 10*(5), 151-155. doi: 10.1111/1467-8721.00137

Grossmann, T. (2010). The development of emotion perception in face and voice during infancy. *Restorative Neurology and Neuroscience, 28*(2), 219-236. doi: 10.3233/RNN-2010-0499

Hambleton, R.K., & Swaminathan, H. (2013). *Item response theory: Principles and applications*: Springer Science & Business Media.

Heider, F. (1958). *The psychology of interpersonal relations*. New York Wiley.

Kana, R.K., Keller, T.A., Cherkassky, V.L., Minshew, N.J., & Just, M.A. (2009). Atypical frontal-posterior synchronization of theory of mind regions in autism during mental state attribution. *Social Neuroscience, 4*(2), 135-152. doi: 10.1080/17470910802198510

Keysar, B., Barr, D.J., Balin, J.A., & Brauner, J.S. (2000). Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science, 11*(1), 32-38. doi: 10.1111/1467-9280.00211

Keysar, B., Lin, S., & Barr, D.J. (2003). Limits on theory of mind use in adults. *Cognition, 89*(1), 25-41.

Montague, D.P.F., & Walker-Andrews, A.S. (2002). Mothers, fathers, and infants: The role of person familiarity and parental involvement in infants perception of emotion expressions. *Child development, 73*(5), 1339-1352. doi: 10.1111/1467-8624.00475

Perner, J., Frith, U., Leslie, A.M., & Leekam, S.R. (1989). Exploration of the autistic child's theory of mind: Knowledge, belief, and communication. *Child development*, 689-700.

R Core Team. (2013). R: A language and environment for statistical computing.

Rasch, G. (1960). Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests.

Rizopoulos, D. (2006). Ltm: An r package for latent variable modeling and item response theory analyses. *Journal of Statistical Software, 17*(5).

Rubio-Fernández, P. (2017). The director task: A test of theory-of-mind use or selective attention?

Psychonomic Bulletin & Review, 24(4), 1121-1128. doi: 10.3758/s13423-016-1190-7

Shaw, P., Kabani, N.J., Lerch, J.P., Eckstrand, K., Lenroot, R., Gogtay, N., . . . Rapoport, J.L. (2008). Neurodevelopmental trajectories of the human cerebral cortex. *Journal of Neuroscience, 28*(14), 3586-3594.

Smith, A. (2006). Cognitive empathy and emotional empathy in human behavior and evolution. *The Psychological Record, 56*(1), 3-21.

Symeonidou, I., Dumontheil, I., Chow, W.-Y., & Breheny, R. (2016). Development of online use of theory of mind during adolescence: An eye-tracking study. *Journal of Experimental Child Psychology, 149*, 81-97. doi: 10.1016/j.jecp.2015.11.007

Thissen, D., & Wainer, H. (2001). *Test scoring*: Routledge.

Wellman, H.M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child development, 72*(3), 655-684.

Yentes, R.D., & Wilhelm, F. (2018). Careless: Procedures for computing indices of careless responding (Version 1.1.3): R package

Zieber, N., Kangas, A., Hock, A., & Bhatt, R.S. (2014a). The development of intermodal emotion perception from bodies and voices. *Journal of Experimental Child Psychology, 126*, 68-79. doi: 10.1016/j.jecp.2014.03.005

Zieber, N., Kangas, A., Hock, A., & Bhatt, R.S. (2014b). Infants' perception of emotion from body movements. *Child development, 85*(2), 675-684. doi: 10.1111/cdev.12134

Processing of affirmation and negation in contexts with unique and multiple alternatives: Evidence from event-related potentials

Maria Spychalska[1,2] **(mspychal@uni-koeln.de)** & **Viviana Haase**[2] **(viviana.haase@rub.de)**

Jarmo Kontinen[2] **(kontinen.jarmo@gmail.com)**

Markus Werning[2] **(markus.werning@rub.de)**

1. Department of German Linguistics, University of Cologne, Germany
2. Institute of Philosophy II, Ruhr University Bochum, Germany

Abstract

We employ a scenario-sentence-verification paradigm to investigate the role of scenario-given alternatives for the processing of affirmative and negative sentences. We show that for both affirmative and negative sentences the N400 amplitude is larger if the context model provides multiple alternatives for a true sentence continuation relative to the case when it provides only a unique referent. Additionally, we observe a late positivity effect for negative relative to affirmative sentences, independent of the context model.

Keywords: Negation; alternatives; N400; P600

Introduction

Negation occurs in every human language and is essential for communication. Nevertheless, negative sentences seem to require elevated processing resources when compared to affirmative ones and negation has posed a challenge to psycholinguistic theories. Language comprehension is generally considered to happen incrementally, i.e. as the linguistic input unfolds in real time; yet, empirical evidence suggests that negation is not compatible with this view, and it is argued that negation may be integrated into sentence meaning only at a later stage of the comprehension process. Furthermore, language processing is considered to be predictive, meaning that we not only process incoming linguistic information but also anticipate upcoming content (cf. DeLong et al., 2005; Pickering & Garrod, 2007; Van Petten & Luka, 2012; Kuperberg & Jaeger, 2015). However, despite of the large amount of experimental literature on prediction, it remains an open question to what extent and at which processing stage the number of contextually available alternatives for an upcoming word modulates the prediction of that word during sentence comprehension. Furthermore, as it is unknown at which stage negation becomes a part of compositional meaning, it is also not clear whether and how negation modulates predictive processing.

In the psycholinguistic literature emphasis has been put on providing a model of negation processing and experimental work has focused on investigating the cognitive costs related to the processing of sentences with negation. Since negative sentences are structurally more complex, they are expected to involve more cognitive resources than their affirmative counterparts. From a semantic point of view, the potential need to suppress positive information is also likely to result in an increase of processing costs. Early on it has been shown that negative sentences are associated with higher error rates as well as longer response and reading times than affirmative sentences (Just & Carpenter, 1971; Clark & Chase, 1972; Lüdtke & Kaup, 2006; Dale & Duran, 2011).

Early electroencephalography (EEG) studies on negation processing further suggested that the integration of negation into the compositional sentence meaning is delayed. For instance, many studies (Fischler et al., 1983; Dudschig et al., 2019) found no effect of negation on the N400 event-related potential (ERP) component. The N400 is a negative shift in the ERP waveform, of a latency between 200 and 600 ms post-stimulus onset and maximal over centro-parietal scalp sites. Its amplitude tends to be larger for words that are (semantically) less expected given the background context, or world-knowledge, as well as for words of lower corpus-based frequency (see Kutas & Federmeier (2011) for an overview). The N400 has furthermore been shown to be inversely correlated with the triggering word's cloze probability, i.e. the percentage of individuals who would continue a given sentence fragment with that word (Federmeier et al., 2007). Various theories interpret the N400 as a marker of (i) lexical retrieval (Brouwer & Hoeks, 2013), (ii) integration into the prior context (Hagoort et al., 2004), (iii) predictive preactivation (DeLong et al., 2005), or even (iv) meaning-related probabilistic prediction (Lau et al., 2013; Kuperberg & Jaeger, 2015; Rabovsky et al., 2018). According to a recent account (Rabovsky et al., 2018), the N400 reflects meaning-related prediction error, where prediction is understood in a non-intentional sense, as an implicit state of the system that is tuned to anticipate the upcoming input in a graded manner.

In the first ERP study on negation by Fischler et al. (1983), the N400 was only modulated by the lexical-semantic relation of the predicate to the main noun and it was larger for the sentence-final predicates when this relation was weak, as for instance in *A robin is/is not a truck* relative to the case when the relation was strong, as in *A robin is/is not a bird*, independently of the presence of negation in a sentence, which reversed the sentence truth-value. This result is usually interpreted as evidence that negation is not processed incrementally and thus is not immediately integrated into the compositional sentence meaning. This interpretation was further sup-

ported by an EEG-study employing a sentence-picture verification paradigm that revealed a delayed integration of the negation in the sentence meaning (Lüdtke et al., 2008). In their experiment, affirmative and negative sentences, such as *In front of the tower there is a/no ghost*, were followed by matching or mismatching pictures, e.g. a tower with a lion or a tower with a ghost, after a short (250*ms*) or a long (1500*ms*) delay. Note that in the case of the affirmative sentences, the matching pictures are explicitly mentioned and thus primed by the sentences, but the negative sentences primed the mismatching pictures. In the short-delay condition, the N400 ERPs reflected a priming effect, namely, for the affirmative sentences the mismatching pictures were associated with a larger N400 than the matching pictures, whereas for the negative sentences the effect was opposite. An effect of negation was only reflected by a late positivity effect that was identified as the P600 effect. In contrast, for the long-delay condition, main effects of truth-value and negation in addition to the priming effect were already observed in the N400 time-window. This result was taken as evidence that integrating negation into the sentence meaning required additional time after the sentence was read. The P600 effect observed in response to the use of negation is an especially noteworthy result. The P600 is a slow, late (around 500-800 ms post-onset) positive shift in the ERP waveform that is maximal over posterior scalp sites. It is often observed for structural violations, grammatical errors or syntactically more complex sentences (Hagoort et al., 1993) but also for some pragmatic and semantic anomalies (Kuperberg et al., 2003). It has been argued to reflect combinatorial aspects of linguistic processing (Kuperberg, 2007) or even semantic integration mechanisms (Brouwer et al., 2012). In the case of negative sentences it may be taken as a marker of the increased processing demands related to integrating the negation into sentence meaning.

However, the comprehension of negated sentences is facilitated if they are embedded into context. Nieuwland & Kuperberg (2008) showed that pragmatically licensed negative sentences such as for example *With proper equipment, scuba diving isn't very safe/dangerous...* did not lead to elevated N400-components for true compared to false sentences. Instead, without pragmatic licensing, e.g. *Bulletproof vests aren't very safe/dangerous...* the true negated sentences led to higher N400s than the false sentences, in line with Fischler et al. (1983). Tian et al. (2016) furthermore showed that for cleft-structures which narrow the scope of the negation and therefore the potential alternatives such as in, e.g. *It is John who hasn't ironed his shirt*, incremental comprehension of negated sentences is facilitated as well.

The role of alternatives for the processing of both negative and affirmative sentences was directly studied in an eye-tracking experiment by Orenes et al. (2014). They investigated whether the presence of multiple alternatives in the context has an effect on the processing of negative and affirmative sentences. An auditory context sentence was introduced that either indicated that all objects are possible choices (*multary* condition: *The figure could be red, green, blue, or yellow*), or restricted the choice to only two objects (*binary* condition: *The figure could be red or green*). Then, the visual context appeared, which always included four different objects, e.g. circles of different colors. The target sentence *The figure is red/not red* was presented auditorily while the four figures were shown on the screen and eye-movements were monitored. For affirmative target sentences subjects fixated the target object (*red circle*) in both context conditions. In contrast, for negative sentences, subjects fixated the target object (*green circle*) only in the *binary* condition, whereas in the *multary* condition they fixated the object that had the mentioned, negated feature (*red circle*). The interpretation of these results is problematic, since the affirmative and negative conditions were not logically comparable. Whereas affirmative sentences directly mentioned the target object, for negative ones the identification of the target was only possible in the *binary* condition, where one could infer the color of the target object from the pair of the context and the target sentence. In the *multary* condition the target object was not identifiable: The intended figure that was described as *not red*, could be blue, yellow, or green.

It is not surprising that in the case when the referent cannot be identified, the processing of a sentence differs relative to the case when the referent can be uniquely established in the context. However, the question arises whether this effect has anything specifically to do with negation. In our current experiment we aimed at directly comparing the processing of affirmative and negative sentences in situations when the context scenario provides multiple or unique potential true sentence continuations. Suppose that Julia is dealt three cards (*ace*, *king*, *queen*) and the game is to choose some of these cards, while rejecting the rest of them at the same time. If Julia selects one card (e.g. *ace*), one can describe her choice by saying that *Julia selected the ace*. There is only one card that can be mentioned in a true affirmative sentence. Additionally, about each of the remaining two cards one can say that it was not selected, e.g. *Julia did not select the king/queen*. In this case there are two potential true sentence continuations. The situation is exactly opposite if Julia had selected two cards, while rejecting only one. By manipulating a game situation of this type we can create unique and multiple contexts equivalent for negative and affirmative descriptions. It has previously been suggested that the N400 amplitude elicited by a given word seems to be directly dependent on the number of contextually given alternatives to this word. Spychalska et al. (2016) showed that, for sentences such as *Some pictures contain X*, if the context scenario provided additional objects *Y* that could be mentioned instead of *X*, the N400 was larger when *Y* would complete a true sentence (thus was a true alternative to *X*) relative to the case when *Y* would complete a false sentence (and was not a true alternative to *X*). Since the N400 is known to inversely correlate with the cloze probability of the triggering word, one can hypothesize that

if the context scenario provides alternative referents for a true sentence condition, the N400 should be larger relative to the case when the context allows to uniquely predict the referent. By contrasting negative and affirmative sentences in logically equivalent conditions, the design allows us to directly measure the effect of negation on the processing of a referent in the context with multiple alternatives.

Method

The experimental design used a scenario-sentence verification paradigm. Participants were informed that they follow a player's moves in a game. In each target trial the player (introduced in the form of a clipart-like image) was dealt three cards, each depicting a different object, which were presented on the screen. Then, the player selected or rejected one or two of the cards. Subjects were informed that this action leads to an exhaustive divide of the set of cards. Selection was marked by framing the selected cards green, which implied that the unframed cards are rejected. Rejection was marked by framing the rejected cards red, which implied that the unframed cards are selected. After the cards were marked, the scene disappeared and a sentence (in German) was presented phrase-by-phrase. At the end of the trial, participants were asked whether the sentence is a true description of the action taken by the player. The target sentences were either of the form 1a (*affirmative* conditions) or of form 1b (*negative* conditions), where *X* is a proper name referring to the player and *Y* denotes the critical noun.

(1) a. *X hat den/die/das Y ausgewählt.*
X has chosen Y.
b. *X hat nicht den/die/das Y ausgewählt.*
X has not chosen Y.

In each target scenario there was only one object of the given type, thus a definite article was used in the sentence. All objects presented in a given trial were of the same grammatical gender to rule out that the noun in the sentence could be predicted based on the article. We ran two experiments: In **Experiment I**, all target sentences referred to one of the **unframed** objects. Thus, negative sentences followed scenarios with green frames, whereas affirmative sentences followed scenarios with red frames. In this way, both the affirmative and negative conditions required the participant to make an inference about the unmarked cards from the information provided visually (i.e. the marked set of cards). In the **unique** conditions, two out of three cards were framed, which left only one possible and therefore unique referent (unframed picture) to be named in the target sentence, whereas in the **multiple** conditions only one card was framed leaving two and hence multiple possible referents (unframed pictures) to be potentially named in the target sentence. In **Experiment II**, all target sentences referred to one of the **framed** cards. Thus, negative sentences followed scenarios with red frames, whereas affirmative sentences followed scenarios with green frames. In this experiment both negative and affirmative target sentences directly described the player's action and did not involve any inferential step. Both experiments used a

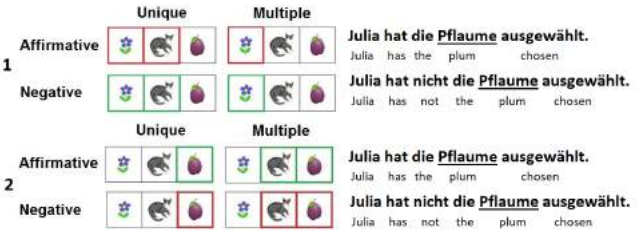


Table 1: Example for the conditions in Experiment 1 (top) and Experiment 2 (bottom)

2x2 design with the factors (i) *Alternatives* (unique/multiple); and (ii) *Polarity* (affirmative/negative), resulting in four conditions as shown in Table 1. [1]

Our main hypothesis was that the N400 recorded for the nouns referring to the target objects should be larger for the *multiple* relative to the *unique* context. This effect was in principle expected both for negative and affirmative sentences; however, we also hypothesized a possible interaction effect between *Polarity* and *Alternatives*, indicating that the processing of negative and affirmative sentences involves non-overlapping processes especially in the context where multiple alternatives are available. Furthermore, based on prior studies (Lüdtke et al., 2008), we hypothesized that the main effect of negation may occur in the late (P600) time-window.

Materials We created a list of 240 German nouns that are depictable and concrete. They were all mono- or bi-syllabic, moderately frequent[2], had a length between three to nine characters and were used in their singular form. The words were combined into 240 unique triples $\langle N_1, N_2, N_3 \rangle$, i.e. each word was used with each other word only once. These triples were used to generate scenarios presenting three different objects, assigned pseudo-randomly to experimental conditions, in such a way that each word was a critical noun only once. This resulted in 60 trials per condition, 240 target trials in total. To balance out the overall truth-value ratio and to make the material more variable we added 240 filler trials: 200 false and 40 true[3], based on a list of 84 new nouns and with affirmative and negative sentences evenly distributed. To rule out the possibility of creating expectations for negative or affirmative sentences based on the color of the frames, the framing of the filler's pictures exploited all possible options cross-balanced, so that affirmative/negative filler sentences followed red or green frames the same number of times.

Procedure Upon arrival participants signed an informed consent of participation. They were given a written instruction including few examples and completed an exercise ses-

[1] The framing was realized as a between-subject factor in order to have a sufficient number of trials per condition (60 in the current design) without inflating the length of the experiment. In addition, false filler sentences were needed to balance the materials (see below).

[2] The logarithmic frequency of all stimuli words was controlled with the use of Leipzig Wortschatz http://wortschatz.uni-leipzig.de/

[3] Overall, 41.6% of the trials in the experiment were false.

sion consisting of eight trials. Feedback was provided for the exercise to make sure that participants understood the task, especially the meaning of the framing and the exhaustive divide of the set of cards. The experiment comprised of eight blocks lasting approximately 10 minutes, with optional breaks in between.[4] The whole experiment lasted approximately 90 minutes. Each trial started with the presentation of three pictures in the center of the screen. The pictures were first shown without any frame. Subsequently, one or two of the objects were framed green or red, followed by the sentence that was displayed phrase-by-phrase on the screen (see Figure 1). The main ERP trigger was the noun referring to (one of) the selected or rejected target object(s). After each sentence, subjects had to respond to a truth-value judgment task by pressing a left- or a right-hand button on a response pad. The buttons were assigned pseudo-randomly by displaying "TRUE" and "FALSE" on the screen sides.

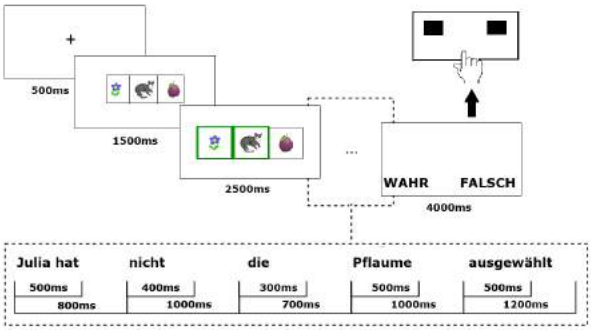


Figure 1: The time course of an example trial with presentation times. The main ERP trigger is the word *Pflaume* (*plum*).

EEG recording and data preprocessing The EEG was recorded with a BrainAmp acticap 64 channel recording system. Electrode position AFz was used as ground and FCz as physical reference. The electro-oculogram (EOG) was measured with four electrodes (PO09, PO10, FT9, FT10), which were reprogrammed and placed above and below the right eye and on both temples. Electrode impedances were kept below 5kΩ. The EEG was recorded with a sampling rate of 500Hz, a 10 sec low cut-off filter and an anti-aliasing hardware filter. The EEG-data was analyzed in the BrainVision Analyzer 2.0 software. We applied a 0.1-30 Hz off-line bandpass filter. All trials with an absolute amplitude difference over $200\mu V/200ms$ or with an activity lower than $0.5\mu V$ in intervals of at least $100ms$ were automatically rejected. The maximal voltage step that was allowed was $50\mu V/200ms$. Eyeblinks were corrected by means of an independent component analysis. The data was re-referenced to the averaged mastoids (TP9, TP10). The baseline-correction was done based on the $200ms$ pre-onset interval of the stimulus. Segments with any remaining physical artifacts (lower than -$90\mu V$ or higher than $90\mu V$) were removed before averaging. At least 50% of segments in each condition were preserved.

[4]For each block a player was introduced. In total, four different players (two female) appeared during one experiment, each assigned to two of the eight blocks.

Statistical analysis of the ERPs following the onset of the critical noun For the analysis of the ERPs we used a repeated measures ANOVA with the factors *Polarity* (negative/affirmative), *Alternatives* (multiple/unique) and *Region* (anterior/posterior).[5] The anterior ROI covered frontal, anterior frontal, frontal parietal, frontal temporal as well as frontal central regions of both hemispheres. The posterior ROI reached across temporal posterior, central posterior, posterior, posterior occipital and occipital regions of both hemispheres. The electrodes from the horizontal midline (central electrodes) were analyzed separately. We analyzed the averaged subjects' ERPs in two time-windows: 250-550 ms post-stimulus onset for the N400 effect and 550-850 ms post-onset for the P600. The assumptions of parametric data (e.g. normal distribution) were met.

Results

Twenty-five volunteers participated in Experiment 1 (nine male; mean age 25.2 ($SD = 4.42$, range $18-33$). In Experiment II we measured twenty-five new (not participating in Experiment I) volunteers (nine male; mean age 23.88 years (SD 3.94), age range 18-33). We excluded one subject per experiment from the analyses due to excessive artifacts in the EEG-data.

Behavioral results. Accuracy was at ceiling level in all conditions, indicating that the task was not too difficult for the subjects (see Table 2). Although minor differences are observed across conditions, due to space limitations, the statistical analysis of the behavioral data is omitted in the paper.

		Unique	Multiple
Experiment I	**Affirmative**	97.22(3.25)	94.17(7.17)
	Negative	95.63(5.83)	95.42(4.12)
Experiment II	**Affirmative**	97.01(2.82)	95.26(5.00)
	Negative	95.34(4.14)	95.27(4.74)

Table 2: Mean accuracy in Experiment I and II, in percentages, and the standard deviation ($\mu(\sigma)$) for all conditions

Polarity independent modulation of the N400 by alternatives. The visual inspection of the grand averages revealed that critical nouns in the multiple conditions elicited clearly larger N400 ERPs than critical nouns in the unique conditions for both sentence polarities and in both experiments. The ANOVA for the time-window **250-550 ms** for **Experiment I**, revealed a main effect of *Alternatives* ($F(1,23) = 30.040$, $p<.001$, $\eta^2 = .566$), with the mean difference between the multiple and unique conditions of $-2.157\mu V$ ($M_{mult} = -.59\mu V$, $M_{unq} = 1.567\mu V$). There was also a main effect of *Polarity* ($F(1,23) = 10.854$, $p = .003$, $\eta^2 = .321$), namely, the negative sentences showed more positive ERPs compared to the affirmative sentences ($\Delta_{Neg,Aff} = .919\mu V$),

[5]The regions were chosen based on the visual inspection of the effect's topography that suggested clear anterior-posterior differences, but no clear lateralization differences. Since we had no specific hypotheses regarding potential lateralization effects, we decided to include only AP as a factor in order to keep the analysis more transparent.

as well as an effect of *Region* ($F(1,23) = 67.535$, $p<.001$, $\eta^2 = .746$, $\Delta_{Post,Front} = 3.411\mu V$). The interaction *Alternatives*×*Region* was significant ($F(1,23) = 11.401$, $p = .003$. $\eta^2 = .331$), which can be attributed to a larger *multiple vs. unique* N400 effect in the posterior ($\Delta_{Mult,Unq} = -2.558\mu V$) relative to the frontal regions ($\Delta_{Mult,Unq} = -1.756\mu V$). Given that the three-way *Polarity*×*Alternatives*×*Region* interaction was also significant ($F(1,23) = 9.712$, $p = .005$, $\eta^2 = .297$), we broke down this interaction by *Region*. For the frontal region, there was a significant effect of *Alternatives* ($F(1,23) = 18.357$, $p < .001$, $\eta^2 = .444$, $\Delta_{Mult,Unq} = -1.756\mu V$), *Polarity* ($F(1,23) = 16.947$, $p < .001$, $\eta^2 = .424$, $\Delta_{Neg,Aff} = 1.232\mu V$) and the *Polarity*×*Alternatives* interaction ($F(1,23) = 5.288$, $p = .031$, $\eta^2 = .187$): The *unique vs. multiple* effect was larger for the affirmative sentences $\Delta_{Mult,Unq} = -2.4\mu V$ than for the negative sentences $\Delta_{Mult,Unq} = -1.112\mu V$. However, for the posterior region only the main effect of *Alternatives* was significant ($F(1,23) = 38.511$, $p < .001$, $\eta^2 = .626$, $\Delta_{Mult,Unq} = -2.557\mu V$). For the midline electrodes the effects were similar, i.e. there was a significant effect of *Alternatives* ($F(1,23) = 33.788$, $p < .001$, $\eta^2 = .595$, $\Delta_{Mult,Unq} = -2.473\mu V$), as well as of *Polarity* ($F(1,23) = 8.58$, $p = .008$, $\eta^2 = .272$, $\Delta_{Neg,Aff} = .943\mu V$), but no interaction.

The results of **Experiment II** were in line with the first experiment. There was a main effect of *Alternatives* ($F(1,23) = 21.045$, $p<.001$, $\eta^2 = .478$), i.e. the multiple conditions showed larger negativity than the unique conditions ($\Delta_{Mult,Unq} = -.972\mu V$), as well as *Region* ($F(1,23) = 36.042$, $p<.001$, $\eta^2 = .610$, $\Delta_{Post,Front} = 3.109\mu V$). No main effect of *Polarity* was observed. Unlike in the first experiment, there were no significant interactions. For the midline electrodes only a main effect of *Alternatives* was found ($F(1,23) = 24.920$, $p < .001$, $\eta^2 = .520$, $\Delta_{Mult,Unq} = -1.160\mu V$).

Alternatives independent Late Positivity for negated sentences. The visual inspection of grand averages revealed a late positivity effect for the negative relative to affirmative conditions, that was apparent for both alternatives conditions and both experiments. The analysis in the late time-window **550-850 ms** for **Experiment I** revealed a main effect of *Polarity* ($F(1,23) = 25.714$, $p<.001$, $\eta^2 = .528$), driven by the negative sentences showing more positive average amplitudes than affirmative sentences ($\Delta_{Neg,Aff} = 1.177\mu V$), as well as a main effect of *Region* ($F(1,23) = 108.986$, $p<.001$, $\eta^2 = .826$, $\Delta_{Post,Front} = 2.58\mu V$). No effect of *Alternatives* was observed; however, there was a significant *Alternatives*×*Region* interaction ($F(1,23) = 19.308$, $p < .001$, $\eta^2 = .456$): The mean amplitude difference between multiple and unique conditions was more negative in the frontal ($\Delta_{Mult,Unq} = -1.028\mu V$) than in the posterior region ($\Delta_{Mult,Unq} = .172\mu V$). There was also a significant three-way interaction *Polarity*×*Alternatives*×*Region* ($F(1,23) = 9.430$, $p = .005$, $\eta^2 = .291$), which we broke down by *Region*. In the frontal region, we found a significant effect of *Polarity* ($F(1,23) = 16.387$, $p < .001$, $\eta^2 = .416$, $\Delta_{Neg,Aff} = 1.192\mu V$), *Alternatives* ($F(1,23) = 11.088$, $p = .003$, $\eta^2 = .325$, $\Delta_{Mult,Unq} = -1.028\mu V$), as well as significant *Polarity*×*Alternatives* interaction ($F(1,23) = 6.867$, $p = .015$, $\eta^2 = .230$): the mean amplitude difference between the negative and affirmative sentences was larger for the multiple ($\Delta_{Neg,Aff} = 1.894\mu V$) than for the unique ($\Delta_{Neg,Aff} = 0.49\mu V$) condition. In the posterior region only the effect of *Polarity* was significant ($F(1,23) = 23.064$, $p < .001$, $\eta^2 = .501$, $\Delta_{Neg,Aff} = 1.161\mu V$). For the midline electrodes, we found an effect of *Polarity* ($F(1,23) = 29.341$, $p < .001$, $\eta^2 = .561$, $\Delta_{Neg,Aff} = 1.32\mu V$), but no effect of *Alternatives*, and no *Polarity*×*Alternatives* interaction.

The results for **Experiment II** were again generally consistent with **Experiment I**. There was a main effect of *Polarity* ($F(1,23) = 15.269, p = .001$, $\eta^2 = .399$) driven by the negative sentences showing more positive ERPs than the affirmative sentences ($\Delta_{Neg,Aff} = .883\mu V$), and a main effect of *Region* ($F(1,23) = 45.121$, $p<.001$, $\eta^2 = .662$, $\Delta_{Post,Front} = 2.234\mu V$), but there was no main effect of *Alternatives*. The interaction *Polarity*×*Region* was significant ($F(1,23) = 9.915$, $p = .004$, $\eta^2 = .301$): The difference between negative and affirmative conditions was larger in the posterior ($\Delta_{Neg,Aff} = 1.373\mu V$) than in the frontal regions ($\Delta_{Post,Front} = .393\mu V$). Additionally, and similar to the first experiment, the interaction *Alternatives*×*Region* was significant ($F(1,23) = 10.739$, $p = .003$, $\eta^2 = .318$, frontal $\Delta_{Mult,Unq} = -.214\mu V$ and posterior $\Delta_{Mult,Unq} = .688\mu V$). The interaction *Polarity*×*Alternatives* was not significant and neither was the three-way interaction. For the midline electrodes, again we only found an effect of *Polarity* ($F(1,23) = 17.516$, $p < .001$, $\eta^2 = .432$, $\Delta_{Neg,Aff} = 1.066\mu V$).

Comparison across experiments. Although both experiments showed similar main effects, some differences between the two framing variants were observed, in particular, some interactions showed significant in one experiment and not in the other. As the experiments were otherwise identical, as a meta-analysis we conducted a full-factorial ANOVA with *Polarity*, *Alternatives* and *Region* as within-subject factors and *Experiment* as a between-subject factor.

This analysis in the early time window **250-550 ms**, showed no main effect of *Experiment*; however, the interaction *Alternatives* × *Experiment* was significant ($F(1,46) = 7.029$, $p = .011$, $\eta^2 = .133$): the multiple vs. unique N400 effect was larger in Experiment I than in Experiment II. There was also a significant *Polarity* × *Region* × *Experiment* interaction ($F(1,46) = 5.311$, $p = .026$, $\eta^2 = .104$).[6] For the midline electrodes there was again no effect of *Experiment*, but the *Alternatives*×*Experiment* interaction was significant ($F(1,46) = 7.338$, $p = .009$, $\eta^2 = .138$).

No main effect of Experiment was found for the time-window of **550-850 ms**, but the interaction between *Alternatives* and *Experiment* was significant both in the main analysis

[6] See the experiment-specific analyses reported above for the relevant mean differences.

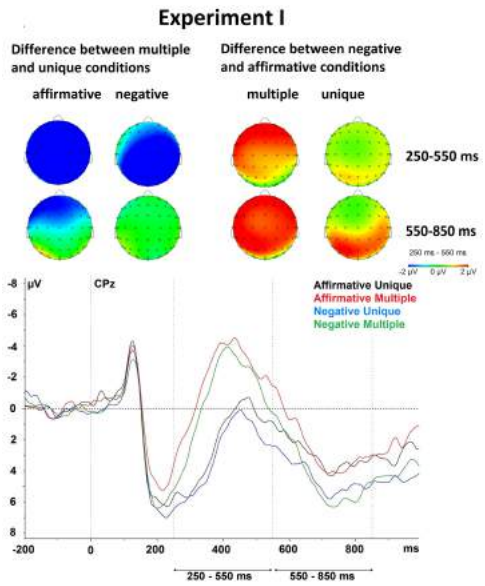


Figure 2: The comparison of grand averages at the critical word in all four conditions at the electrode CPz and the topographical distribution of the effects in Experiment I.

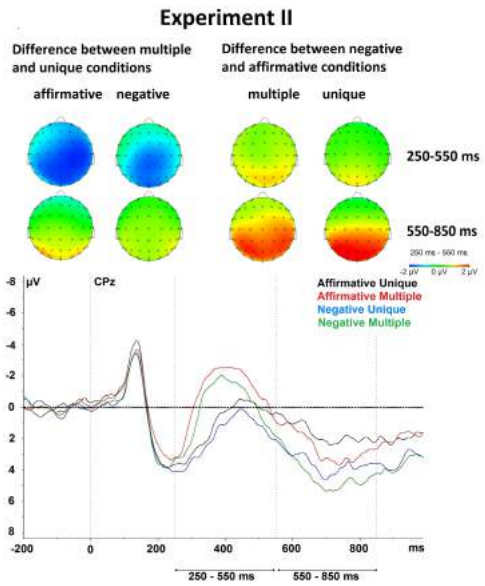


Figure 3: The comparison of grand averages at the critical word in all four conditions at the electrode CPz and topographical distribution of the effects in Experiment II.

($F(1,46) = 4.808$, $p = .033$, $\eta^2 = .095$, $\Delta_{Mult,Unq} = -.428\mu V$ in Experiment 1, and $\Delta_{Mult,Unq} = .237\mu V$ in Experiment 2), as well as for the midline electrodes ($F(1,23) = 4.603$, $p = 0.037$, $\eta^2 = .091$, $\Delta_{Mult,Unq} = -.443\mu V$ in Experiment 1 and $\Delta_{Mult,Unq} = .305\mu V$ in Experiment 2).

Discussion

In two experiments we compared the processing of affirmative and negative sentences in two contexts: (i) a unique context, that allows to make a specific prediction of the critical noun to be mentioned in a true sentence, (ii) a multiple context, where two alternatives can potentially be mentioned in a true sentence. In our design the affirmative and negative conditions were fully comparable: In both cases, sentence verification either required inferring the status of the unframed cards from the status of the marked set of cards (Experiment 1), or no inference was involved, since the sentence directly referred to the marked cards (Experiment 2).

It is generally accepted that the N400 reflects meaning-related expectancy of the stimulus. What this means precisely remains debated and the theories vary between taking the N400 to be a marker of lexical retrieval, lexical predictive preactivation or even meaning-related probabilistic prediction. We hypothesized that the presence of multiple alternatives, where the processor cannot uniquely predict the referent, should lead to a higher N400 relative to the case of a unique referent. This hypothesis was supported, as we observed a larger N400 effect for multiple vs. unique conditions independent of the sentence polarity. Although it is well-established that the N400 is correlated with expectancy and cloze probability, no prior experiments focused directly on the relation between the N400 and the availability of equally predictable alternatives in the scenario. Furthermore, our study is the first that directly compares how the scenario-based cloze probability of the upcoming word affects the processing of that word in affirmative and negative sentences. As the main result we show that the effect of alternatives is similar for both sentence polarities, thus, the possibility of predicting a unique true sentence continuation facilitates the processing also under the scope of negation.

In addition we also showed that, for both alternatives conditions, the presence of the negative particle led to a late positivity effect at the sentence critical noun occurring after the negation, i.e. when the noun's expectancy was related to the prior use of negation in the sentence. This effect forms a clear and large P600 effect in both experiments, although in the first experiment some modulation is already observed in the early time-window. Under the assumption that the P600 amplitude reflects integration of the lexical information into the semantic representation of the sentence (see Brouwer & Hoeks, 2013), this effect may be taken to indicate that in the case of negative sentences the construction of the semantic representation was more effortful.

Although the main pattern of effects was the same in both experiments, the type of framing made a significant contribution to the size of the effects, namely, the N400 effect for multiple vs. unique alternatives turned out to be significantly larger in Experiment 1 than in Experiment 2, as shown by the *Alternatives*Experiment* interaction. Furthermore, only in Experiment I we observe an interaction between *Polarity* and *Alternatives* by *Region* in both time-windows, specifically, in the posterior region the effect of alternatives was similar for both sentence polarities, but in the frontal region it was larger for affirmative sentences. This interaction result indicates that, in Experiment 1, affirmative and negative sentences possibly engaged slightly different processes in the

two alternatives conditions. These between-experiment differences may be explained in terms of different task demands. Although from a logical perspective the two tasks were equivalent, cognitively they differ in an important manner. In Experiment 1, all target sentences referred to the unframed objects, whose status (chosen vs. unchosen) could only be inferentially determined based on the status of the framed objects and the assumed exhaustivity of the set divide. Thus, to determine the status of the unframed objects one had to include the so-called closed world assumption, which basically means that what is not known to be true is false. Given this assumption, in the negative condition, one could infer that if A was chosen (framed green), then B & C were not (or if A & B were chosen, then C was not). Similarly, in the affirmative condition, one could infer that if A was not chosen (framed red), then B & C were chosen (or if A & B were not chosen, hence C was). This inferential process is slightly different in the two conditions as it either goes from a negative premise to a positive conclusion, or the other way round. In contrast, in Experiment 2, all target sentences mentioned the framed, highlighted objects and hence there was no need to reason about the status of the remaining cards.

In sum, we showed that if the scenario allows to uniquely predict the upcoming word in a true sentence continuation, the processing of that word is significantly facilitated both in the case of affirmative and negative sentences, which is observed in the form of a reduced N400 component for the uniquely predicted words. The (higher) cognitive cost of processing the negative particle is observed in the form of a late positivity effect. Finally, the task demands, i.e. whether the status of the referent is directly marked or has to be inferentially determined, make a significant contribution to the size of the effects and appear to differently affect the negative and affirmative sentences. Further research should explore how the N400 is modulated by a larger number of alternatives provided and how the effect depends on the sentence truth-value.

Acknowledgments

This collaborative research was funded by *Stiftung Mercator* within the project *Structure of Memory* as well as by the German Research Foundation, DFG, within the priority program XPrag.de.

References

Brouwer, H., Fitz, H., & Hoeks, J. (2012). Getting real about semantic illusions: Rethinking the functional role of the p600 in language comprehension. *Brain Res. 2012 Mar 29;1446:127-43. doi: 10.1016/j.brainres.2012.01.055.*, *29*(1446), 127-43.

Brouwer, H., & Hoeks, C. (2013). A time and place for language comprehension: mapping the N400 and the P600 to a minimal cortical network. *Frontiers of Human Neuroscience*, *7*(758).

Clark, H. H., & Chase, W. G. (1972). On the process of comparing sentences against pictures. *Cognitive Psychology*, *3*(3), 472-517.

Dale, R., & Duran, N. (2011). The cognitive dynamics of negated sentence verification. *Cognitive Science*, *35*(5), 983-996.

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, *8*(8), 1117-1121.

Dudschig, C., Mackenzie, I. G., Maienborn, C., Kaup, B., & Leuthold, H. (2019). Negation and the n400: investigating temporal aspects of negation integration using semantic and world-knowledge violations. *Language, Cognition and Neuroscience*, *34*(3), 309-319.

Federmeier, K., Wlotko, E., De Ochoa-Dewald, E., & Kutas, M. (2007). Multiple effects of sentential constraint on word processing. *Brain Research*, *18*(1146), 75-84.

Fischler, I., Bloom, P., Childers, D., Roucos, S., & Perry, N. (1983). Brain potentials related to stages of sentence verification. *Psychophysiology*, *20*(4), 400-409.

Hagoort, P., Brown, C. M., & Groothusen, J. (1993). The syntactic positive shift (SPS) as an ERP measure of syntactic processing. *Language and Cognitive Processes*, *8*(4), 439-483. (Doi:10.1080/01690969308407585) doi: 10.1080/01690969308407585

Hagoort, P., Hald, L. A., Bastiaansen, M. C. M., & Petersson, K. M. (2004). Integration of word meaning and world knowledge in language comprehension. *Science*, *304*(5669), 438-441.

Just, M. A., & Carpenter, P. A. (1971). Comprehension of negation with quantification. *Journal of Verbal Learning and Verbal Behavior*, *10*(3), 244-253.

Kuperberg, G. R. (2007). Neural mechanisms of language comprehension: Challenges to syntax. *Brain Research*, *1146*, 23-49.

Kuperberg, G. R., & Jaeger, T. F. (2015). What do we mean by prediction in language comprehension? *Language, Cognition, and Neuroscience*, *31*(1), 32-59.

Kuperberg, G. R., Sitnikova, T., Caplan, D., & Holcomb, P. J. (2003). Electrophisiological distinction in processing conceptual relationship within simple sentences. *Cognitive Brain Research*, *17*(1), 117-129.

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event related brain potential (ERP). *Annual Review of Psychology*, *62*, 621-647.

Lau, E., Holcomb, P., & Kuperberg, G. (2013). Dissociating N400 effects of prediction from association in single word contexts. *Journal of Cognitive Neuroscience*, *25*(3), 484-502.

Lüdtke, J., Friedrich, C. K., De Filippi, M., & Kaup, B. (2008). Event-related potential correlates of negation in a sentence-picture verification paradigm. *Journal of Cognitive Neuroscience*, *20*(8), 1355-1370.

Lüdtke, J., & Kaup, B. (2006). Context effects when reading negative and affirmative sentences. In R. Sun (Ed.), *Proceedings of the 28th annual conference of the cognitive science society* (p. 1735-1740). Lawrence Erlbaum.

Nieuwland, M. S., & Kuperberg, G. R. (2008). When the truth isn't too hard to handle: An event-related potential study on the pragmatics of negation. *Psychological Science*, *19*(12), 1213-1218.

Orenes, I., Beltran, D., & Santamaria, C. (2014). How negation is understood: Evidence from the visual world paradigm. *Journal of Memory and Language*, *74*, 36-45.

Pickering, M., & Garrod, S. (2007). Do people use language production to make predictions during comprehension? *Trends in Cognitive Sciences*, *11*, 105110.

Rabovsky, M., Hansen, S. S., & McClelland, J. L. (2018). Modelling the N400 brain potential as change in a probabilistic representation of meaning. *Nature Human Behaviour*, *2*(693-705).

Spychalska, M., Kontinen, J., & Werning, M. (2016). Investigating scalar implicatures in a truth-value judgment task: Evidence from event-related brain potentials. *Language, Cognition and Neuroscience*, *31*(6), 817840.

Tian, Y., Ferguson, H., & Breheny, R. (2016). Processing negation without context - why and when we represent the positive argument. *Language, Cognition and Neuroscience*, *31*(5), 683-698.

Van Petten, C., & Luka, B. J. (2012). Prediction during language comprehension: Benefits, costs, and ERP components. *International Journal of Psychophysiology*, *83*(2), 176-190.

Evidence for effort prediction in perceptual decisions

Nisheeth Srivastava
Dept of Computer Science, IIT Kanpur
India

Abstract

The classic drift diffusion model of the 2AFC choice process assumes that observers select evidence accumulation thresholds to optimize some desired level of accuracy across the experiment. We argue that it is more ecologically natural to assume that decision-makers set this threshold adaptively, using information from recent trials to adjust it for upcoming ones. To test this hypothesis, we designed and conducted a pair of random dot motion discrimination experiment where the coherence parameter that controls task difficulty varies across trials in a predictable manner. To analyze data from these experiments, we also designed a hierarchical drift diffusion model that allows decision-makers to adapt their evidence threshold based on the trend of difficulty of previous trials. Our results suggest that observers rationally integrate cross-trial information about trial difficulty into perceptual decision-making by adjusting their internal evidence thresholds. We briefly discuss the implications of the existence of such trial-level effort inference on contemporary models of the choice process.

Keywords: drift diffusion model; ideal observer model; Bayesian modelling; cognitive effort; rational inference

Introduction

The drift diffusion model (DDM) is a very successful sequential sampling model of the choice process (Ratcliff & McKoon, 2008). Particularly when applied to perceptual decision-making tasks, where the stream of evidence is transparent to the experimenter, this model has shown excellent fits to choice and response time data from a wide variety of experimental paradigms, even generalizing across organisms (Brunton, Botvinick, & Brody, 2013). While it shares several components, including parallel accumulation and race-to-a-threshold with other competing paradigms such as leaky competing accumulation (Usher & McClelland, 2001) and decision field theory (Busemeyer & Townsend, 1993), its stochastic specification of important components of the choice process - rate of accumulation of evidence, response bias, and variability in the evidence accumulation rate - gives it excellent flexibility and interpretability in modelling the summary statistics seen in perceptual decision-making experiments.

While on the one hand, its mathematical construction makes the DDM an excellent *descriptive* model of the choice process, it simultaneously makes it challenging to associate its optimality criterion with the goals and costs faced by real decision-makers. Specifically, the drift diffusion model is known to implement the sequential probability ratio test (SPRT) (Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006), which is statistically optimal in the sense that for any choice of the decision threshold, using the DDM criterion for choice will yield the highest possible accuracy (Wald & Wolfowitz, 1948).

The relationship between the SPRT and DDM imbues it with a normative sense of optimality - observers are being statistically optimal in the SPRT-sense if we show that the DDM model fits their behavior well. But the underlying assumptions of SPRT - perfect evidence integration, approximately linear evidence accumulation rates, race to a fixed evidence threshold - are not good fits for the information and processing limitations that organisms face in real-world decision-making scenarios. In recent years, objections to these premises have been raised on both computational and empirical grounds. Deneve has documented how the conventional drift diffusion paradigm fails to accommodate situations where sensory inputs are unreliable (Deneve, 2012). Thura and colleagues have shown how reaction time distributions in perceptual decision-making tasks may be better described by evidence accumulation terminated by breaching a time-collapsing threshold responsive to an increasing 'urgency' signal than classic accumulation to a fixed threshold (Thura, Beauregard-Racine, Fradet, & Cisek, 2012). Glaze et al have demonstrated how perfect evidence integration - a fundamental assumption of drift diffusion models - is sub-optimal in the face of unsignalled context shifts in the decision-making environment (Glaze, Kable, & Gold, 2015). Thus, using DDM as a normative baseline, research is increasingly focusing on identifying aspects of the environment that constrain real-world decision-making.

Threshold adaptation as effort inference

We propose to revise the premise that decision-makers accumulate evidence up to a fixed threshold. Whereas such proposals have been made previously (Thura et al., 2012), they have focused on incorporating the opportunity cost of continued sampling in the form of a threshold that decreases over time (Drugowitsch, Moreno-Bote, Churchland, Shadlen, & Pouget, 2012). We focus on a different aspect of the threshold determination process - that decision-makers are likely to use information from previous trials to set decision thresholds for upcoming trials. Here again, it is well-documented that patterns of responding can introduce response biases in experiments (Ratcliff & McKoon, 2008). The drift diffusion model allows such biases to be modeled explicitly using changes in diffusion start-point parameters.

We consider a different normative possibility: we propose to model the threshold parameter used in drift diffusion modeling as a proxy for the amount of effort the observer believes is necessary for adequate performance, and we intend to investigate whether human observers can infers the effort needed for upcoming trials using effort observations in recent trials. Grounding this hypothesis in a perceptual decision-

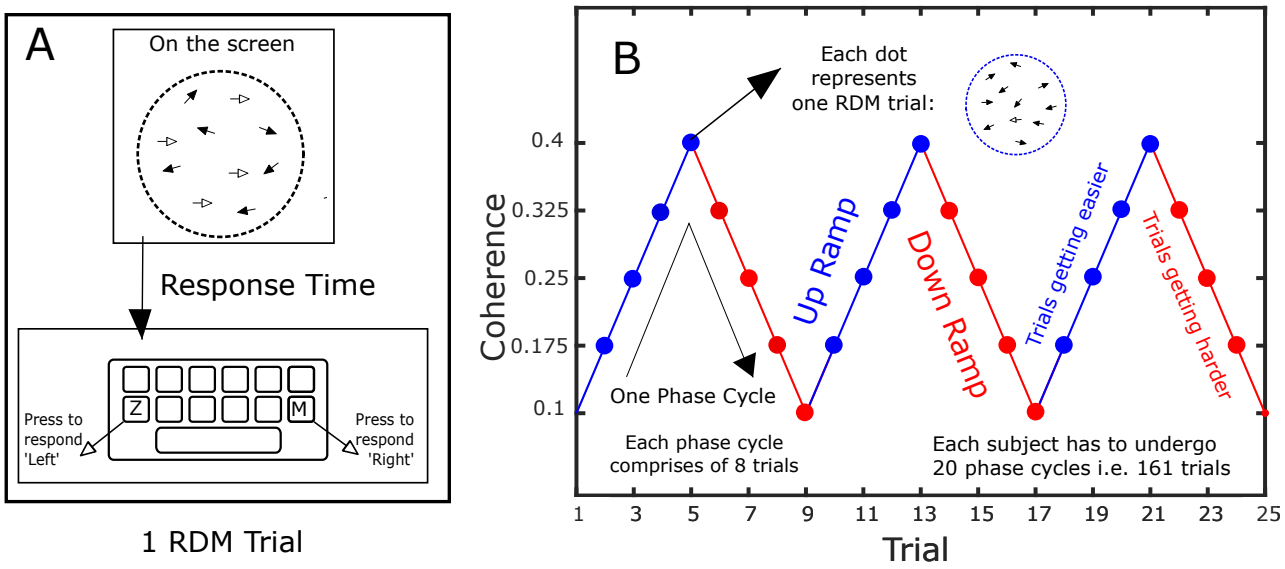


Figure 1: Schematic illustrating the experiment design. (A) On each experiment trial, participants saw a set of dots in Brownian motion, with a horizontal drift added to some fraction of the dots. Participants had to discriminate motion direction using key presses on a computer keyboard, and were incentivized to emphasize accuracy. (B) The sequence of trials each participant saw possessed a higher-order structure, with the difficulty of successive trials increasing and decreasing in a cyclic manner.

making task, we model behavior on this task using a hierarchical ideal Bayesian observer that performs 2AFC random dot motion (RDM) discrimination. The lower level of this hierarchical model simulates individual RDM trials using a classic drift diffusion setup. The higher level of this model uses a reinforcement learning-inspired controller to set appropriate values of the evidence threshold for each trial.

To test this hypothesis, we designed a specific variation of the standard RDM task. In the standard task, trial difficulty is either blocked or randomized across trials. We instead designed a sequence of trial presentation that introduced a predictable trend in the coherence parameter across trials. If people are adaptively tracking the amount of effort they are having to expend on individual trials, we expect such inference to inform their effort allocation on upcoming trials. A hierarchical extension of the drift diffusion model, with a top-down controller setting the evidence threshold adaptively across trials, would potentially fit choice and RT data gathered from such an experiment design better than a simple DDM that assumes a fixed evidence threshold.

Experiment 1

RDM with higher-order structure

Participants saw a screen with moving dots designed according to the following algorithm. Random motion of the dots was provided by allowing Brownian motion in the vertical direction, i.e. all the dots drifted vertically about their mean position by a distance chosen from a normal distribution. Horizontal motion was either randomly selected from a bidirectional (left/right) uniform distribution (for incoherent dots) or from a unidirectional uniform distribution (for coherent dots). The selection of dots as coherent or incoherent was determined at the beginning of each experimental trial using Bernoulli trials controlled by the coherence parameter c.

As in all RDM discrimination experiments, the critical manipulation of task difficulty was governed entirely by the coherence parameter c. We selected a range of values of the coherence parameter by running a calibration pilot with 5 participants, performing 280 trials of the discrimination task under accuracy emphasis. We picked a range of coherence values that permitted 65% accuracy at the low end of the range and 95% accuracy at the high end of the range.

Participants had to indicate the direction of motion of the overall dot pattern on each trial, as illustrated in Figure 1A. They were allowed to take as much time as they wanted to respond to each trial, and as much time as they wanted to rest between the trials. Each correct response fetched points. The scoring system was such that a correct response fetches 10 points; the score of each correct response doubled on responding correctly to three successive trials, and reset to 10 points in case the streak was broken. We further encouraged accuracy emphasis by promising a reward to the highest scorer.

The specific higher-order structure introduced in our experiment was a cyclic shift across the 5 specific values of the coherence parameter used in the experiment $\{0.1, 0.175, 0.25, 0.325, 0.4\}$. For example, a participant starting the experiment with a trial with coherence 0.1 would next see a trial with coherence 0.175, then one with 0.25, up to the maximum coherence level of 0.4, beyond which the coherence would begin dropping down to 0.325, then to 0.25 etc. Each such phase cycle from one coherence value through the other 4 and back to the original takes 8 trials, given we use 5 unique coherence values. Participants completed 20 such phase cy-

cles for a total of 161 trials per participant, as illustrated in Figure 1B.

Task

The task was administered via a web-based interface. Participants indicated responses with keyboard button presses, and were allowed to take as long as they liked before pressing the space bar to begin the next trial. Distance from the screen was not fixed, but the display size was selected such that the display was well within the foveal range (20 degrees visual angle) of normally sighted observers.

Participants

52 undergraduate and postgraduate students participated in the experiment for course credit (4 female, mean age 20.5 ± 1.57 SD). All participants had normal or normal-corrected vision. Since the experiment was conducted without personal supervision, some participants showed significant guessing behavior. Post-task, we excluded the data for 18 participants who had less than 85% accuracy on the highest coherence trials.

Results

We expected that an observer tracking the cross-trial variations in difficulty would end up tracking the repeated ramp-like movement of the coherence parameter, and take longer on an upcoming trial if the cross-trial coherence was trending downward (i.e. the trials were becoming more difficult) than if it was trending upward (i.e. the trials were becoming easier).

However, even observers insensitive to cross trial information are expected to show the same pattern globally in our data, because upcoming trials aren't just expected to be easier/harder on up/down ramps, they actually are easier/harder too. Therefore, the critical test for whether information from previous trials are affecting the current trial is to see whether trials of the same difficulty (coherence) level have longer RTs when they occur within a down coherence ramp (sequence of decreasing coherence, increasing trial difficulty) than when they occur within an up coherence ramp (sequence of increasing coherence, decreasing trial difficulty).

In Figure 2, we plot RTs stratified by coherence level of the immediate trial for the three intermediate coherence values in our experiment across all participants, separating out trials occurring during up and down ramps. For each of the three coherence levels, the difference between down ramp RTs and up ramp RTs is directionally in the predicted direction. While the data is noisy, the large sample size of our experiment (34 participants $\times$ 20 cycles = 680 data points per coherence level per ramp) allows us to assert statistical significance at Bonferroni-corrected $p < 0.01$ for two of the levels (0.175 and 0.325) in two-sample t-tests. The third comparison (for coherence level 0.25) does not yield a statistically significant difference in such a test.

These summative statistical results, while conceptually congruent with our hypothesis, are inadequate to draw strong inferences. The large standard deviations evident in Figure 2 reflect considerable inter-participant heterogeneity in response times. We therefore sought to test our hypothesis at a trial-by-trial level by developing a generative model of the information accumulation process involved in a 2AFC perceptual learning task that accommodates such adaptive control over the evidence threshold, and comparing its ability to explain our data vis-a-vis a fixed threshold evidence accumulation model.

A hierarchical drift diffusion model

The DDM, in its standard form is a Wiener diffusion process with drift,

$$dy = vdt + sdW, \tag{1}$$

where y is the diffusion state, v is the drift, s determines the amount of diffusion and dW represents the standard Wiener process. The model accumulates normally distributed pieces of evidence for either alternative until a bound on the cumulative evidence is crossed, and then emits the winning option as the choice.

We developed a hierarchical model of the choice process using a recently proposed Bayesian version of the DDM (Bitzer, Park, Blankenburg, & Kiebel, 2014). This takes the form of a sequential Bayesian update model that maps noisy stimuli observations to latent Gaussian representations

$$x_t \sim N(\mu_i, \sigma^2), \tag{2}$$

constructs a generative model of the likelihood of seeing certain latent feature values for each stimulus alternative,

$$p(x_t|A_i) = N(\hat{\mu}, \hat{\sigma}^2), \tag{3}$$

and updates beliefs about the correctness of a decision alternative given these noisy observations

$$p(A_i|X_{1:t}) = \frac{p(x_t|A_i)p(A_i|X_{1_t-1})}{\sum_{j=1}^{M} p(x_t|A_i)p(A_j|X_{1_t-1})}, \tag{4}$$

where x_t are noisy observations from stimuli belonging to category i, with true prototypical values μ_i and measurement noise σ, estimated prototypical values $\hat{\mu}$ and internal generative variability $\hat{\sigma}$, A_i as possible alternatives and M as the number of considered alternatives. Bitzer et al show that this intuitive ideal observer model is formally identical to the drift diffusion model given certain assumptions about the relationship between parameters of the model.

We augmented this model with a metalearner that estimates the expected sampling effort needed for upcoming trials based on effort allocation on previous trials. We assume a very simple model for this metalearner. It simply updates the effort estimate λ as

$$\lambda_t = \lambda_{t-1} + \gamma\Delta_t, \tag{5}$$

where

$$\Delta_t = z(RT_{t-1}) - z(RT_{t-2}), \tag{6}$$

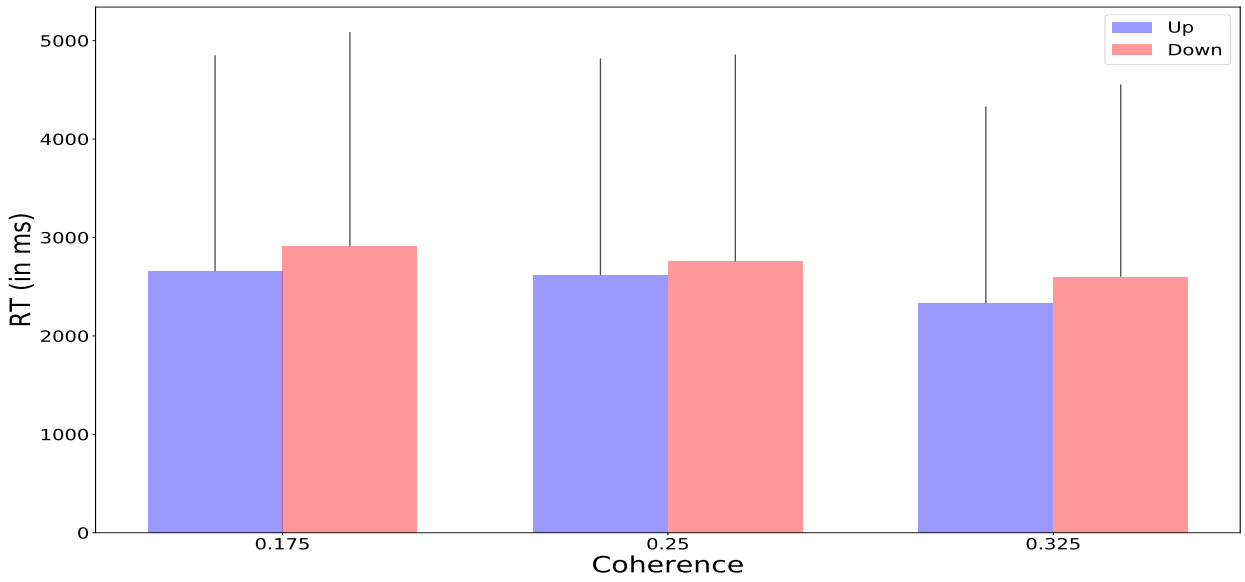


Figure 2: RT at intermediate coherence levels while the cyclic higher-order coherence trend is going up (blue bars) and down (red bars). Higher RTs seen for trials of equal difficulty, when coherence is trending downwards, i.e. trials are getting harder, is evidence for adaptive changes in the evidence threshold responsive to past trials. Error bars represent 1 S.D.

γ is a free scaling parameter, $z(RT)$ is the normalized z-score of RT at time t with respect to the RT distribution and λ_t serves as the threshold for the DDM for the t^{th} trial. This model is not meant to be comprehensive. We have designed it purely to simulate our expectation of the role of predictable up and down changes in dot motion coherence on observer behavior. We expect that observers will be sensitized to these trends and extrapolate from them to set decision thresholds for upcoming trials. Increasing effort on recent trials should yield a larger threshold for the upcoming trial and vice versa. Normalization is used to induce a natural scale on the size of the change in the threshold; the RT distribution is admittedly non-Gaussian so this assumption could be further refined in future work. Also, to avoid over-fitting, we have used the global RT distribution to normalize the RTs, whereas a more realistic model may use sequential summary statistics within participants. Indeed, a filtering-based model might capture the basic intuition of the metalearner more elegantly, but we wished to compare the augmented model with a complicated baseline using only choice and RT data, necessitating parsimony in parameter extension. The version of the meta-learner we have proposed has just one additional free parameter beyond the baseline.

As in (Bitzer et al., 2014), we reduced the set of estimated parameters of the Bayesian model from seven to three by assuming equal amount of drift for both stimuli. In practice we did this by setting $\mu = \hat{\mu} = \pm 1$ for the 2AFC case. Parameter fitting for the 3 parameters to be estimated $\theta = \{\sigma, \hat{\sigma}, t_{nd}\}$ also followed the procedure outlined in (Bitzer et al., 2014). We defined the log likelihood of the data given all parameters as

$$\begin{aligned}\log p(\text{Acc}, \text{RT}|\theta) &= \log p(\text{Acc}|\theta) + \log p(RT|\theta) \\ &\propto -w_{\text{acc}}\ (\text{Acc} - \text{Acc}(\theta))^2 \\ &- \sum_{e=0}^{1}\sum_{i=1}^{7} w_{q_{e,i}}\ (q_{e,i} - q_{e,i}(\theta)^2,\end{aligned}$$

where $q_{e,i}$ is the i^{th} of seven quantiles $(0.02, 0.05, 0.1, 0.3, 0.5, 0.7, 0.9)$ for either correct or error responses as indicated by e.

To evaluate the log likelihood function, again following the procedure in (Bitzer et al., 2014), we simulated our experiment with the Bayesian observer model for different parameter values, averaging the accuracy and RT quantiles obtained across 30 model runs per parameter tuple θ and setting the likelihood weights w to the inverse variance over these repetitions. We scaled each iteration in our simulation to 125 ms. However, we found the MCMC approach advised in (Bitzer et al., 2014) to be too slow (on the order of days) for fitting our hierarchical model that used different threshold values for each trial. Therefore, we used a two-stage grid-search of the parameter space (logarithmic exploration in one stage followed by linear refinement in the second) to optimize the negative log likelihood. In practice, we found that the grid search yields comparable mean parameter values for the baseline DDM model as the MCMC procedure implemented in (Bitzer et al., 2014).

Since we don't use MCMC to obtain a posterior distribution over the parameters, it is useful to average over multiple runs of the likelihood computation at the optimal parameter values to obtain representative likelihood values. After

Model	AIC	BIC
Simple DDM	16.5 (2.7)	36.3 (2.7)
Hierarchical DDM	**8.9 (1.5)**	**28.5 (1.5)**

Table 1: Model comparison. Standard deviation across 20 model runs are given in parentheses.

Block	1	2	3	4
Δ BIC	12.4	-3.0	-0.07	-0.78

Table 2: Model comparison across four sequential blocks of 40 trials each from all participants. Block 1 contains the first 40 trials from each participant, etc.

finding optimal parameters via grid search for both models, we calculate model likelihood as the average likelihood obtained from 20 runs of the model for the optimal parameter values. The results from our model comparison using these average likelihood values are presented in Table 1. ΔBIC measures difference between baseline DDM and hierarchical DDM BIC. Positive values support the hierarchical DDM, negative values support the baseline model. ΔBIC values with magnitudes smaller than 2 imply insignificant differences between models; values larger than 10 constitute very strong support for a model. Using this measure, the hierarchical DDM is clearly preferable to the simple DDM, across data from all 34 participants (ΔBIC $= 7.8$ from Table 1).

We additionally ran a block-wise analysis, dividing each participant's trials into 4 sequential blocks of 40 trials and calculating model complexity statistics on the likelihoods emitted by the model for the best fitting parameter values of the overall model ($\sigma = 11, \hat{\sigma} = 8, t_{nd} = 250ms$). We anticipated that any evidence of gradual adoption or relinquishment of threshold metareasoning would show up in the relative model complexity tracked across these four blocks.

As the results in Table 2 show, the hierarchical model is heavily preferred over the simpler model during the first quarter of trials, measured across all participants. For later trials, both models are evenly matched, with the simpler model slightly preferred.

Experiment 2

The block-wise analysis of our data revealed an interesting property: participants behaved as if they were tracking higher-order structure at the beginning of the experiment, but appeared to switch away to behaving more like conventional DDM decision-makers later on. We thought this was because participants tried to use the higher-order structure between trials, but then shifted away from it, since doing so does not offer any material advantage. To falsify such an explanation, we conducted a followup experiment where tracking the higher-order structure would confer a material advantage.

RDM with helpful higher-order structure

This experiment design was equivalent to the first one, with an alteration only in score-keeping. Recall that the first experiment used scoring with a multiplicative boost for maintaining accuracy streaks. Score per correct response would double for every three consecutive correct responses, and reset to the default value on each error.

For this second experiment, we transformed the incentive system into an optional 'auto boost' mechanism, such that accurately responding on three consecutive trials would fill up a booster bar, and optionally selecting the booster bar would allow a participant to *buy* out any one trial of their choice. The bought out trial would be assumed to have been answered correctly, would not be displayed on screen, and would not count towards win streak counts.

Participants

31 undergraduate students (age = 20.6 ± 1.3 years, 8F) volunteered to participate in this experiment; none having participated in the first one. These participants performed the experiment under supervision using the same web-based interface as before. All other experimental protocols were identical to the first experiment. Post-task, we excluded the data for 5 participants who had less than 85% accuracy on the highest coherence trials.

Results

The primary difference between the two experimental tasks was that tracking higher-order structure could not help participants in the first one, but potentially could in the second one. Specifically, in this second experiment participants could hold on to filled up booster bars for what they *predicted* to be the hardest trials, considerably reducing their overall error and effort.

Block	1	2	3	4
Δ BIC	9.7	8.4	8.9	3.2

Table 3: Model comparison across four sequential blocks of 40 trials each (counting bought out trials) from all participants for Experiment 2. Block 1 contains the first 40 trials from each participant, etc.

A direct measure of whether participants did deploy such a strategy in this task is the relative distribution of trials on which participants used boosts across coherence levels. Random use of boosts would indicate no use of the optimal strategy outlined above, whereas exclusive concentration of boost use in the lowest coherence level would indicate perfect adherence to this strategy.

Empirically, we found that 45.5% of all boosts were used for the lowest coherence trials (chance 12.5%, $p < 0.005$), and 78.6% of all boosts were concentrated within the two lowest coherence levels (chance 37.5%), suggesting strong utilization of the optimal strategy.

As a secondary measure of strategy adherence, we fit both classic and hierarchical DDM to these participants' data following the same methodology as for the first experiment. Boosted trials were assigned correct responses and subject-specific mean RT for the corresponding coherence level during the simulations. For lack of space, we only present the difference in BIC values betwe en hierarchical and classic DDM fits for this data. Contrasting these ΔBIC results tabulated in Table 3 with those from the original experiment (see Table 2) strongly suggest that the manipulation of the incentive system does affect participants' behavior on the task in the predicted direction.

Discussion

With increases in computational power and experimental methods, behavioral researchers are increasingly able to track behavior with greater granularity, which makes hypotheses about intermediate computations underlying behavior tractable to investigation. This development is making the study of algorithmic aspects of the decision-making process increasingly more feasible. Not only does such research characterize biological observers' decision-making *processes*, it also provides constraints on the nature of the cost functions that computational theorists can reasonably set up for decision-making agents.

As part of this paradigm shift, recent work has begun to question the classic drift diffusion model's assumption of a fixed evidence threshold in recent years, basing these arguments on the temporal opportunity costs of continued sampling (Thura et al., 2012). In this work, we asked the same question from a different standpoint. We asked whether observers might be sensitive to higher-order statistics in decision-making tasks and adaptively adjust evidence thresholds on upcoming trials to use them efficiently.

To see if this can happen, we created a novel variation of the classic random dot motion discrimination task, introducing an up-and-down ramp in difficulty across trials (Gold & Shadlen, 2000). We predicted that observers would be sensitive to this variation. We designed an extension of the drift diffusion model that incorporates metacognitive adaptation of the evidence threshold based on the trend of difficulty of recent trials, and found that it offers a better explanation of participants' behavior in our experiment than a simple drift diffusion model. A followup experiment further demonstrated that participants' tracking of higher-order structure in this task was intentional - they shifted away from tracking when it offered no advantage and continued tracking when it did.

Our results support a shift in interpretation of the evidence threshold from its SPRT-driven association with *accuracy*, towards a more general view of it as an effort parameter influenced by a variety of information sources. Such an interpretation also makes it easier to generate normative accounts of decisions from memory using DDM-like models, building upon its descriptive success in modeling retrieval success and RT distributions in this domain (Krajbich & Rangel, 2011). Unlike in perceptual decisions where the evidence presentation rates are fixed, and decisions receive immediate feedback, decisions from memory are made with evidence streams of unknown provenance and without feedback. The empirical success of DDM in explaining data from such experiments warrants a broader interpretation of the normative principles of the framework, along lines proposed in this work.

References

Bitzer, S., Park, H., Blankenburg, F., & Kiebel, S. J. (2014). Perceptual decision making: drift-diffusion model is equivalent to a bayesian model. *Frontiers in human neuroscience*, *8*, 102.

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological review*, *113*(4), 700.

Brunton, B. W., Botvinick, M. M., & Brody, C. D. (2013). Rats and humans can optimally accumulate evidence for decision-making. *Science*, *340*(6128), 95–98.

Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: a dynamic-cognitive approach to decision making in an uncertain environment. *Psychological review*, *100*(3), 432.

Deneve, S. (2012). Making decisions with unknown sensory reliability. *Frontiers in neuroscience*, *6*, 75.

Drugowitsch, J., Moreno-Bote, R., Churchland, A. K., Shadlen, M. N., & Pouget, A. (2012). The cost of accumulating evidence in perceptual decision making. *Journal of Neuroscience*, *32*(11), 3612–3628.

Glaze, C. M., Kable, J. W., & Gold, J. I. (2015). Normative evidence accumulation in unpredictable environments. *Elife*, *4*.

Gold, J. I., & Shadlen, M. N. (2000). Representation of a perceptual decision in developing oculomotor commands. *Nature*, *404*(6776), 390.

Krajbich, I., & Rangel, A. (2011). Multialternative drift-diffusion model predicts the relationship between visual fixations and choice in value-based decisions. *Proceedings of the National Academy of Sciences*, *108*(33), 13852–13857.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: theory and data for two-choice decision tasks. *Neural computation*, *20*(4), 873–922.

Thura, D., Beauregard-Racine, J., Fradet, C.-W., & Cisek, P. (2012). Decision making by urgency gating: theory and experimental support. *Journal of neurophysiology*, *108*(11), 2912–2930.

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: the leaky, competing accumulator model. *Psychological review*, *108*(3), 550.

Wald, A., & Wolfowitz, J. (1948). Optimum character of the sequential probability ratio test. *The Annals of Mathematical Statistics*, 326–339.

Decision-makers minimize regret when calculating regret is easy

Nisheeth Srivastava
Department of Computer Science, IIT Kanpur
India

Abstract

This paper provides empirical evidence that human decision-makers use prospective regret minimization as their dominant decision strategy when regret calculations are cognitively easier to perform, and use expected utility maximization when they aren't. We designed a simple decision problem wherein utility maximization and expected regret minimization yield distinctly difference choices, and manipulated the cognitive effort involved in making regret calculations across respondent samples to arrive at our results. While previous research has associated ecological considerations like sense of responsibility and familiarity with this difference, we show that, at least in experimental settings, cognitive calculability in regret space appears to predominantly drive this difference. We also show that this preference for regret minimization can be countermanded by changing the distribution of options presented to the respondent, posing a challenge to simple sequential accounts of strategy selection learning which sequence strategy selection and application in order.

Keywords: decision-making; cognitive heuristics; cognitive effort; regret minimization; utility maximization

Introduction

Regret is an important variable in humans' decision-making. Empirical investigations spanning psychology (Zeelenberg 1999; Connolly & Reb, 2005), neuroscience (Coricelli et al., 2005) and economics (Loomes & Sugden, 1982; Sarver, 2008) have demonstrated that in several decision contexts, humans behave as if they are trying to minimize *prospective* regret, rather than minimize *prospective* expected utility.

This distinction is of great significance for choice models that wish to track consequential human decisions. For instance, Chorus and colleagues have published a series of papers showing that a discrete choice model designed assuming regret minimization as the underlying choice strategy outperforms conventional random utility models (RUM) style discrete choice models in predicting future travel demand (Thiene, Boeri & Chorus, 2012).

At the same time, conventional RUM models, assuming implicit utility maximization have proved their value in modeling human choices in a large array of applications (Small & Rosen, 1981), suggesting that utility maximization is a useful approximation for peoples' intentions in such situations. Consequently, it is important to attempt to characterize situations wherein decision-makers are likely to prefer either of these decision-making strategies. Zeelenberg & Pieters (2007) have suggested, on theoretical grounds, that regret-minimization is more likely to be used:

(a) when choices are perceived to be important and difficult,

(b) when the decision-maker expects to be held accountable for their choice and

(c) when the decision-maker anticipates receiving feedback about options in the near future.

There is also some empirical evidence supporting the basic premise that domain unfamiliarity may drive the use of regret minimization strategies, a mechanism that is substantially congruent with the theoretical factors identified by Zeelenberg & Pieters (2007). Boeri, Scarpa & Chorus (2014) have showed using discrete choice modeling on a transport choice dataset that the behavior of respondents unfamiliar with the choice context was better explained by regret minimizing models.

A common thread between such theoretical and empirical observations is the notion that regret is explicitly *calculated* by the respondent (Zeelenberg & Pieters, 2007). It is because of this commitment to explicit psychological calculation that the role of prospective feedback and accountability etc. become important in predicting the use of regret minimization as a strategy. Since regret is arrived at via comparison with alternative outcomes, no possibility of feedback would imply no possibility of experiencing regret, which could shift respondents' behaviors towards other strategies.

This commitment to explicit psychological calculation differentiates regret from utility, for which no such commitments are necessary. It is common to observer proposals suggesting direct reward encoding in human observers' brains (Padoa-Schioppa & Assad, 2006). At a minimum, the idea that utilities may be constructed is not yet consensual in the corresponding literature at the interface between psychology and economics (Slovic, 1995; Srivastava & Schrater, 2015).

The centrality of explicit calculation for regret is the focus of the work we report in this paper, wherein we sought to characterize the effect of cognitive ease of calculation of regret on decision-makers' meta-decision to use it as a choice strategy. Our hypothesis was that observers would switch away from use of a regret minimization strategy as the cognitive costs of calculating regret increased. To test this hypothesis, we designed a simple choice task wherein expected regret minimization and expected utility maximization yield clearly divergent choice behaviors, and manipulated the choice stimuli to make explicit comparison of items in regret space easier or harder.

We obtained empirical results substantially supporting our hypothesis. Specifically, we found that participants preferred regret minimizing choices when the choice set was a set of monetary labels, but preferred utility maximizing choices when it was a set of product photos, albeit

associated with money labels. A chronometric assessment of difficulty in judging valuation differences between stimuli of the same category was used to establish that regret calculations for the former stimuli category were relatively easier than for the latter. Challenging simple sequential accounts of strategy selection in decision-making, a final experiment demonstrated that decision-makers' stimulus-category specific bias could be countermanded by changes in the distribution of stimulus valence at the time of presentation. We conclude with a discussion contextualizing our findings within existing formal accounts of strategy selection in decision-making.

Discriminating between choice strategies

Some econometric research in the past has sought to discriminate between the use of utility maximization and regret minimization strategies by fitting different varieties of discrete choice models to data (Thiene, Boeri & Chorus, 2012). However, such models have several free parameters and idiosyncrasies in estimation procedures, and their result interpretations are frequently susceptible to validity challenges. To avoid such complications, we sought to design a simple experimental task in which utility maximization and regret minimization would predict clearly divergent choices.

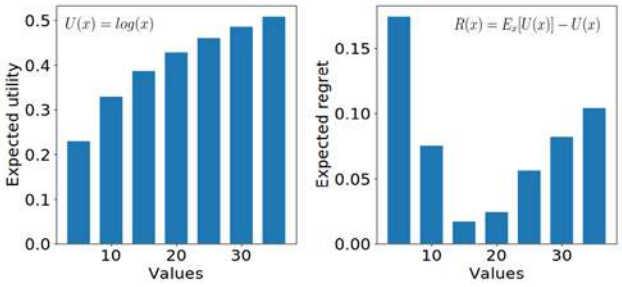


Figure 1: Expected utility (left) and expected regret (right) for nominal x values plotted on the x-axis. A logarithmic form is assumed for the utility function.

This took the form of a choice problem where respondents are told that the correct choice is one of N positive integer-valued options, that each of the options has an equal chance of winning, and that if they guess the correct option, they get the amount of money, or a product of equivalent cost, indicated by the integer value indexing that option[1].

Our interest was to contrast the relative performance of utility maximization and regret minimization strategies in this setup. Formally, given a set of alternatives X, and some estimate or direct measurement of the utility of alternatives a utility maximizer would select according to the choice rule

$$\arg\max_x U(x) \quad (1)$$

It is trivial to see that the expected utility maximizing choice in this problem is to always select the option with the highest integer value.

A regret minimizer, on the other hand would calculate the potential *regret* for choosing each one of the outcomes

$$R(x) = |U^* - U(x)| \quad (2)$$

where U* is some counterfactual comparative benchmark utility, and then use the choice rule

$$\arg\max_x R(x) \quad (3)$$

The choice of benchmark utility differentiates regret calculations into different categories. Minimax regret computations take the benchmark utility to be the utility from the best possible outcome (Savage, 1951), and is commonly used in game-theoretic settings to model behavior. Such a criterion is reasonable for when the decision-maker is expected to know the correct option, a common premise in game-theoretic settings. For decision-makers operating with little domain knowledge, average or expected utility is frequently selected as the benchmark utility, as is common in reinforcement learning settings (Kaelbling, 1996). Since our task falls in the latter category, we use expected utility to perform our regret calculations.

Assuming a linear relationship between utility and regret as defined in Equation (3), we see that the regret minimizing choice in this problem is to pick the option in the middle of the range of available options, calculating U(x) as the *prospective* utility of x should it win and treating U(.) as a logarithmic map of x, a classic micro-economic assumption. This pattern is, in fact, inevitable since the benchmark expected utility occurs in the middle of the value range given equi-probable outcomes and draws the regret minimum towards itself. Figure 1 illustrates this intuition quantitatively, showing that prospective regret is lowest when selecting in the middle of the range.

Thus, this simple decision problem potentially gives us a straightforward way of empirically differentiating the use of utility maximizing versus regret minimizing strategies. Assuming even spacing of choice set options, respondents selecting options towards the extreme large values of the offered range are expected utility maximizers, while respondents selecting options in the middle of the offered range are expected regret minimizers.

Given this premise, we next designed a simple experiment to test it. We designed two sets of choice stimuli, one for which regret calculation should be easy, and one for which it should be hard, and asked two different set of respondents to choose between them using the paradigm described above.

[1] The inspiration for this problem design is drawn from an unpublished draft by Oleg Urminsky & Adele Yang, which in turn derived this problem from a common radio station contest - the jackpot guessing game.

Experiment: easy money and hard pens

Our basic prediction is that decision-makers prefer a regret minimization strategy for option sets wherein comparing the value of options is relatively easy, and prefer utility maximization (or other strategies) when such comparisons are hard. To test this, we designed a between participants' experiment, with one cohort making decisions using a stimuli set that permits *easy* regret calculations and the other using a stimuli set that does not. As a precursor to this, we ran another study to quantitatively identify which stimuli categories are, respectively, easy and hard for respondents to calculate regret.

Precursor study

Regardless of whether comparisons are utility function-wise, feature-wise or heuristic-based, it appears natural that the presence of more features should make regret calculations harder. Therefore, we designed choice option sets to have either just one feature (a money amount) or multiple features (money amounts plus other features), corresponding to easy and hard regret calculation settings.

Design. Specifically, we selected two categories of stimuli to test for relative difficulty vis-à-vis a baseline of simple numerical comparisons. These were

(A) two-digit money amounts, and

(B) images of pens, presented alongside their actual market price.

Each participant completed two blocks of 35 trials each for either category of stimulus, with the block presentation order (ABAB/BABA) counter-balanced across participants. Within a block all participants saw a stream of 36 stimuli from a single category (ITI = 500ms), and had to successively respond to the cue, "Is this one much better or worse than the last one?" prompting 1-back comparisons with the stimulus currently on the screen. The sequence of stimuli presentation was pseudo-randomly generated in each category using sampling with replacement from a set of 7 unique stimuli (described in the main experiment for both categories), with the constraint that the new stimulus had to be different from the previous two stimuli in the sequence. "Yes" and "no" responses were coded to the "left" and "right" arrows of a regular QWERTY keyboard. Responses were disabled for the first stimulus in each block since it had no valid comparison. Participants were asked to take as much time as needed to respond, and the trial number within the block was shown alongside the total number of trials in the block on the screen.

Before these four stimuli-specific blocks were presented, participants' response time baselines for numeric distance calculations were established by presenting them with a stream of 36 three digit numbers (ITI = 500ms) sampled from a uniform distribution on [10,99], successively asking the question, "Is this number much larger or smaller than the last one?" The presentation and response interface used for this block was identical to the one used for the subsequent stimuli-based blocks.

Sample and analysis. For this precursor study, we recruited 10 volunteers (2F, age = 24 +/- 2.3 years, 0 left-handed) using convenience sampling.

The regret calculation conditions (easy vs. hard), in the form of different stimuli sets, were empirically validated on the premise that the critical step in regret calculation is the utilitarian comparison of the outcome received with an alternative. Adopting a mental chronometric approach, the relative time taken in performing this calculation for different categories of stimuli was used to operationalize our sense of relative *difficulty* of regret calculations. For all our calculations we report below, we excluded outlier RTs (> 2S.D. from category mean). These constituted 1.5% of all trials (21 out of 1400 total trials), but occurred primarily in the *pens* category trials. The exclusion of these outliers in fact deflates the size of primary result we report below. Therefore, we do not report results including them.

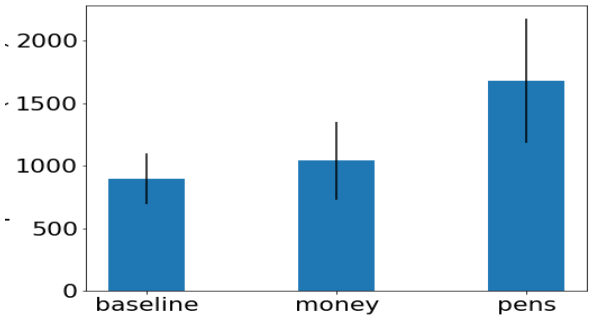


Figure 2: Mean Response times for pair-wise difference judgments within different categories of stimuli for all 10 participants of the precursor study. Errors bars represent +\- 1 S.D.

Results. Figure 2 displays average response times category-wise, combining trials across participants and category blocks. Clearly, respondents found the monetary comparisons of value approximately as easy as numeric comparisons of magnitude (Cohen's d = 0.40), demonstrating the intuitive mapping of number to value in the monetary domain. Equally clearly, respondents took longer to respond to comparisons involving images of pens alongside their prices (Cohen's d = 1.39), implicating multi-dimensional considerations in estimating the value of these objects.

Thus the precursor study objectively established that respondents take longer to assess whether two pens offered at different price points are significantly different from each other than to assess this for just two money amounts. Granted the chronometric assumption that RT predicts task difficulty, this result validates our consideration of choice stimuli drawn from the former category as *harder* than the latter. This distinction, in turn, permits the design and conduct of our main experiment.

Main study

Design. Volunteers for the main experiment were recruited from the general university population. However, participants from our precursor study were excluded. All consenting volunteers were randomly assigned to *easy* (N = 54, age 20.4 +/- 1.4 years, 31 F) and *hard* (N = 53, age 19.8 +/- 1.0 years, 25 F) regret calculation conditions respectively.

Both sets of respondents participated in the experiment in a classroom setting separated spatially from each other, transmitting their responses via text messages. The *easy* group respondents were presented with the following instructions, "Consider this hypothetical scenario. I have a bowl of seven paper tokens, each one with one of the first seven multiples of five written on them. Every number is written on at least one token, and no token has more than one number on it. At the end of the class, I will draw a token and whoever can text me (response number) the number on the token I will draw will win the amount of money written on that token."

Figure 3: Choice stimuli presented to respondents in the *hard* condition. Numbers in parentheses represent pen codes. Money amounts are true prices of the corresponding pens. Pens are arranged in randomized order with respect to money amounts to prevent positional bias in responses

The *hard* group respondents were presented with the visual display shown in Figure 3 accompanied by the instructions, "Consider this hypothetical scenario. I have a bowl of seven paper tokens, each one with a number between 1 and 7 written on it. Every number is written on at least one token, and no token has more than one number on it. At the end of the class, I will draw a token and whoever can text me (response number) the number on the token I will draw will win a pen of the type listed under that number on this display."

Analysis and results. Figure 3 summarizes the responses from both groups of respondents as a histogram of the number of respondents that selected each response option. The difference between the response patterns is visually apparent in the modes of the two distributions in Figure 3, and a two-sample T-test of the individual responses from the two cohorts also indicates a significant difference (t_{105} = 2.18, p = 0.03). An effect size calculation yielded a Cohen's *d* of 0.41, again consistent with a significant difference between the two response patterns.

A comparison with the predictions from Figure 1 clearly suggests that respondents from the *easy* group, who were significantly biased towards responding in the middle of the proffered range, were likely using a regret minimization strategy, whereas respondents from the *hard* group, who preferred the pricier pens, were likely using a utility maximization strategy.

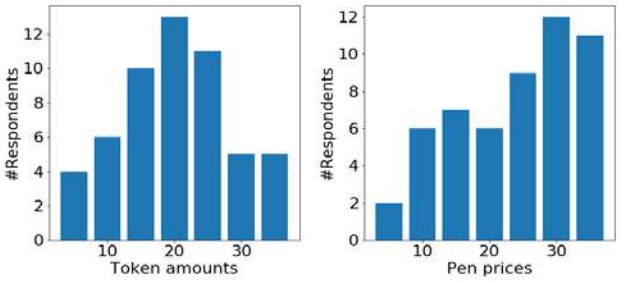


Figure 4: Histogram of respondents' selections for choices where regret calculation is designed to be (left) easy and (right) hard.

This finding is not easily explicable by alternative hypotheses. Previous theoretical proposals have suggested that respondents prefer to decide based on prospective regret when choices are difficult or consequential (Zeelenberg & Pieters, 2007). If anything, it appears intuitive that choosing in pen space is more difficult than choosing in money space. Results from our precursor study establish, at the very least, that estimating value differences between pens in our display is harder than estimating value differences between money amounts. If the pens are harder to choose from, then Zeelenberg & Pieters (2007) would predict the opposite pattern of results than what is seen. Similarly, arguments explaining regret minimization being preferred in unfamiliar domains should also predict it being used when selecting between pens than between money amounts, since choosing between money amounts is unlikely to be more unfamiliar than choosing between idiosyncratic stimuli like pens. Thus, this result appears to clearly favor an ease of calculation explanation for preferring a prospective regret minimization strategy.

Input or enabler?

While the difference in responding elicited by our manipulation does suggest a role for the ease of regret calculation entering into respondents' decision about which strategy to use, it does not clarify how this variable enters this reasoning.

We conducted a variant of the original experiment to differentiate between two potential roles for this cognitive effort variable: (i) as an input to hierarchical decision

process, where the strategy is selected first, then implemented, followed further by assimilation of feedback, or (ii) as a mechanistic enabler, in the sense that quicker regret calculation makes results from the use of a regret minimization strategy available sooner to participants, and hence more likely to be used.

As we discuss further below, the first possibility would fit this cognitive effort variable within formal hierarchical models of strategy selection and learning, such as Rieskamp & Otto's influential SSL model (Rieskamp & Otto, 2006). The latter would be more compatible with heuristic accounts of the effect of availability and accessibility on decision-making (Carroll, 1978), which are yet to be successfully formalized to the same extent (Gigerenzer & Gaissmaier, 2011).

Design. If strategy selection precedes outcome evaluation, then we would expect changes in the range of outcomes used for our decision problem to not affect the choice of decision strategy. Conversely, if changing the range of outcomes for the decision problem reveals differences in the pattern of responding, it is clear that some aspects of outcome evaluation must precede the decision of which strategy to use.

To test this hypothesis, we again used a between-subjects design. The decision problem and setup was identical to the one used in the *easy* condition of the main experiment, with two different groups of respondents making choices using two different sets of money amounts. The first set used the same stimuli as the original experiment. The second set used the stimulus set {5,10,15,20,25,30,**105**}, replacing the largest stimulus in the original set with a much larger value. All participants received the same instructions as in the main experiment's *easy* condition in a classroom setting, and transmitted their selections using text messaging as before.

Sample. Volunteers for the experiment were recruited from the general university population. We screened the recruited sample for previous participation in either our precursor study or the main experiment. A total of 90 participants (23F, Age = 20.3 +\- 1.8 years) were finally selected for participation in the experiment, and randomly assigned to two equally-sized groups for this study.

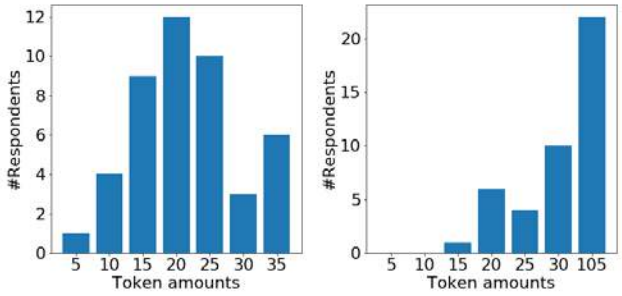


Figure 5: Histogram of responses for groups responding to (left) original stimuli set and (right) changed stimuli set

Analysis and results. As is evident from Figure 5, the response patterns in both groups were starkly different. A two sample T-test for the individual responses returned strongly statistically significant $t_{88} = 6.00$, $p < 10\text{-}6$ and an effect size calculation yielded a large effect (Cohen's d = 1.08). Further, the response pattern elicited from the 45 participants who were presented with the same stimuli as in the *easy* condition in the main experiment were not statistically different from the 54 participants' responses obtained during the former experiment (two sample t-test p = 0.40), suggesting that the original result was robust.

This result shows that changing the set of choice options by adding an extremely valuable alternative makes respondents substantially more likely to prefer the expected utility maximizing strategy, suggesting that a sequential view of strategy selection followed by application cannot faithfully reflect how participants use information available from the choice set before making their decision. Thus, the present evidence suggests that the expected different costs of regret computation for different stimuli sets does not enter explicitly into participants' strategy-selection calculations, but rather enables regret-based determinations to be emitted preferentially by virtue of being generated quicker, in line with the bag of heuristics view of decision-making strategies (Gigerenzer & Gaissmaier, 2011). However, we discuss below how our findings could potentially be reconciled with a hierarchical view of strategy selection further below.

Discussion

Summary of results. In this paper, we have proposed a novel characterization of when human decision-makers are likely to prefer minimizing prospective regret over alternative decision-making strategies like expected utility maximization. Our proposal is that decision-makers prefer to minimize regret when the cognitive cost of calculating regret is low, and switch to alternative decision strategies when this cost is high.

To test this hypothesis, we designed a simple decision problem which permits a clear empirical differentiation between the use of either of these two decision-making strategies. We conducted a chronometric assessment of two stimuli sets, for which relative value judgments had distinctly different difficulty levels. Although other measures of effort have been proposed in the literature, drawing upon information-theoretic considerations (Huber, 1980), the validity of these measures is ultimately assessed using reaction time data (Johnson & Payne, 1985). Thus, while alternative operationalizations of effort are certainly possible, our response time-based definition appears reasonable.

Using this observation to establish the relative difficulty of regret calculations using options selected from these two stimuli sets, we asked two separate groups of participants to make decisions that were formally identical, except for the stimuli identity difference. We found that the pattern of responses for choices made using stimuli that were hard to

evaluate comparatively was more consistent with the use of an expected utility maximization strategy, whereas for stimuli that were easier to compare, the pattern of responses was more consistent with the use of an expected regret minimization strategy. It is, of course, impossible to verify that these were the only two strategies possible for participants to use in the choice problem. Ad hoc heuristic approaches such as 'bias towards the middle of the range' etc could, in principle, be potentially confounded with the regret minimizing predictions for this choice problem. Such ad hoc proposals, however, are not parsimonious, in the sense that they fail to explain the shift to expected utility maximization for the same choice problem using different stimuli, whereas the cost of calculation explanation does.

A more significant question is how well the result demonstrated in this somewhat arbitrary choice problem generalize to richer experimental settings and real-world decisions. We consider this an important consideration for future work.

Related work. There is a large literature on strategy selection, anchored in contemporary times by Rieskamp & Otto's powerful SSL theory (Rieskamp & Otto, 2006). The basic outline of this theory is that observers select a strategy to tackle each instance of a decision problem stochastically, guided by their preference for each of the possible strategies. This strategy-preference in SSL has three components, (i) the maximum reward possible in a trial, (ii) an initial strategy-specific preference, and (iii) a learning-based association of strategy to the choice problem, based on the long-run trend of the use of that strategy resulting in a higher reward. The basic intuition underpinning SSL is that observers adapt to choice contexts by gradually learning to prefer strategies that prove more rewarding in them. Notably, Rieskamp & Otto (2006) explicitly consider the possibility that the cognitive costs of applying a strategy may enter observers' calculations for strategy preference. However, how such strategy-specific costs would enter their model's calculation has remained an open question.

The results in this paper provide useful constraints on the potential development of such a cost-sensitive model. Our main experiment strongly suggests a role for cognitive cost of applying a strategy in determining observers' preference for it. A naïve approach might be to subtract some notional *cost* of calculation from the reward term in the SSL prior on strategy preference. However, our follow-up experiment demonstrates that strategy preference can be affected by complex informational aspects of the choice problem, such as the distribution of options in value space.

Such a complex interaction does not appear to be possible in the baseline two-step algorithmic specification of SSL, wherein first the strategy is selected based on existing strategy preferences, and then information from the current trial updates the strategy preferences. We conjecture that a drift-diffusion based (Ratcliff & McKoon, 2008) extension of the SSL model, wherein the evidence for the utility of options accumulates competitively and becomes available to assist in strategy evaluation depending on how soon this competition terminates, could accommodate our results.

References

Boeri, M., Scarpa, R., & Chorus, C. G. (2014). Stated choices and benefit estimates in the context of traffic calming schemes: Utility maximization, regret minimization, or both?. *Transportation research part A: policy and practice*, *61*, 121-135.

Carroll, J. S. (1978). The effect of imagining an event on expectations for the event: An interpretation in terms of the availability heuristic. *Journal of experimental social psychology*, *14*(1), 88-96.

Connolly, T., & Reb, J. (2005). Regret in cancer-related decisions. *Health Psychology*, *24*(4S), S29.

Coricelli, G., Critchley, H. D., Joffily, M., O'Doherty, J. P., Sirigu, A., & Dolan, R. J. (2005). Regret and its avoidance: a neuroimaging study of choice behavior. *Nature neuroscience*, *8*(9), 1255.

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual review of psychology*, *62*, 451-482.

Huber, O. (1980). The influence of some task variables on cognitive operations in an information-processing decision model. *Acta Psychologica*, *45*(1-3), 187-196.

Johnson, E. J., & Payne, J. W. (1985). Effort and accuracy in choice. *Management science*, *31*(4), 395-414.

Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of artificial intelligence research*, *4*, 237-285.

Loomes, G., & Sugden, R. (1982). Regret theory: An alternative theory of rational choice under uncertainty. *The economic journal*, *92*(368), 805-824.

Padoa-Schioppa, C., & Assad, J. A. (2006). Neurons in the orbitofrontal cortex encode economic value. *Nature*, *441*(7090), 223.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: theory and data for two-choice decision tasks. *Neural computation*, *20*(4), 873-922.

Rieskamp, J., & Otto, P. E. (2006). SSL: a theory of how people learn to select strategies. *Journal of Experimental Psychology: General*, *135*(2), 207.

Sarver, T. (2008). Anticipating regret: Why fewer options may be better. *Econometrica*, *76*(2), 263-305.

Savage, L. J. (1951). The theory of statistical decision. *Journal of the American Statistical association*, *46*(253), 55-67.

Slovic, P. (1995). The construction of preference. *American psychologist*, *50*(5), 364.

Small, K. A., & Rosen, H. S. (1981). Applied welfare economics with discrete choice models. *Econometrica: Journal of the Econometric Society*, 105-130.

Srivastava, N., & Schrater, P. (2015). Learning what to want: context-sensitive preference learning. *PloS One*, *10*(10), e0141129.

Thiene, M., Boeri, M., & Chorus, C. G. (2012). Random regret minimization: exploration of a new choice model for environmental and resource economics. *Environmental and resource economics*, *51*(3), 413-429.

Zeelenberg, M. (1999). Anticipated regret, expected feedback and behavioral decision making. *Journal of behavioral decision making*, *12*(2), 93-106.

Zeelenberg, M., & Pieters, R. (2007). A theory of regret regulation 1.0. *Journal of Consumer psychology*, *17*(1), 3-18.

To Teach Better, Learn First

Oana Stanciu (Stanciu_Oana@phd.ceu.edu)[1]

Máté Lengyel (m.lengyel@eng.cam.ac.uk)[1,2]

József Fiser (FiserJ@ceu.edu)[1]

[1]Department of Cognitive Science, Central European University
7 Oktober 6 utca, Budapest 1051, Hungary
[2]Computational and Biological Learning Lab, Department of Engineering, University of Cambridge
Trumpington street, Cambridge CB2 1PZ, UK

Abstract

There has been little cross-fertilization between research on active learning and teaching, despite extensive conceptual similarities. The current study aims to bridge the gap by showing that engaging in active learning can influence subsequent teaching performance. In a one-dimensional boundary teaching task, participants who first took the role of an active learner went on to become better teachers than participants who did not. In order to disentangle the effect of active selection of samples from their information content, the performance of active learners was compared to that of yoked passive learners. While prior passive learning also significantly boosted teaching performance, it did so to a lesser extent. However, in paired comparisons, teachers with active learning experience did not differ significantly from their yoked-passive learning counterparts. Based on the current results we cannot argue for a teaching benefit specific to active learning as opposed to a more general improvement caused by experiencing the task from the learner's perspective. However, we suggest that this is a promising line of inquiry using more complex learning and teaching tasks.

Keywords: teaching; active learning; evidence selection

Introduction

Perhaps the most enduring debate in the education literature, as well as around kindergartens and classrooms, concerns the virtues of exploratory play in contrast to the canonical, largely passive mode of teacher-led instruction (Bruner, 1961; Mayer, 2004). The discussion has been naturally phrased in terms of the relative benefits and disadvantages that the learner incurs when learning from self-guided discovery compared to direct instruction. However, the complementary, and equally important, link between efficient self-guided learning and good teaching has remained largely unexplored.

The common thread running between teaching and active learning is easy to identify when comparing their formal descriptions. Recent rational-agent models have conceptualized teaching as a recursive process in which the teacher and the learner reason about each other. Specifically, the teacher selects training samples for the learner such that, given the learner's prior knowledge and inference making mechanisms, these samples would lead the learner to the desired conclusion efficiently, i.e. by requiring the smallest number of samples (Shafto, Goodman, & Griffiths, 2014). Conversely, the learner interprets the observed samples assuming they were generated by this pedagogical process (as opposed to randomly). Similarly, an ideal active learner will also sample the environment strategically. However, they will do so by directing their information gathering (e.g. by moving their eyes to explore a visual scene or choosing interventions on the environment) in order to maximize their expected information gain (Yang, Wolpert, & Lengyel, 2018). There are two ways in which active learning can be advantageous. First, observations collected in a strategic way will be more informative for any learner (not just the one sampling information); for instance, by avoiding irrelevant or redundant evidence. Second, and more importantly, there is an added advantage specific to the active learner stemming from the fact that they sample information in light of their prior knowledge and the hypotheses that they wish to test. This effect was demonstrated in experiments in which the data selected by an active learner was also presented to a yoked "passive" learner, and, despite the observations being matched, active learners performed better than their yoked passive counterparts (Markant & Gureckis, 2014).

Thus, both being a good teacher and a good active learner rest on the same general ability to evaluate the potential value of a new piece of evidence relative to a current state of knowledge and a task. Nonetheless, there are important differences. First, teaching brings the added complexity of selecting data for the use of another agent, who might differ widely from the teacher in their state of knowledge and inference making. In line with this, Bass, Shafto, and Gopnik (2017) have linked Theory of Mind (ToM) development to children's pedagogical sampling ability. Second, the active learner does not have access to the target hypothesis, and thus can only select data that minimize uncertainty. However, Yang, Vong, Yu, and Shafto (2019) recently proposed a reconceptualization of active learning as self-teaching by envisioning a learner who simulates an uninformed teacher whose task is limited to providing queries. In this framework, the self-teacher does not optimize for expected information gain, although this will often be the collateral result. Thus, despite differences, it is still feasible to think about teaching and active learning as two highly related cognitive processes.

Given the computational similarity of teaching and active learning, is it possible that they are also integrated through linked processes in human behavior? In other words, would it be possible to hone teaching skills through active learning?

Figure 1: Example image array from the teaching task. In this trial, food items were sorted from left to right in ascending order of their vitamin B content. The black vertical bar represents the daily recommended dose of vitamin B, which is the boundary the participant had to teach. In this case, the participant clicked on the two images closest to the boundary, which were automatically labelled.

Intuitively, taking the perspective of the learner prior to teaching should be a useful experience. It could allow the teacher to better understand, even if implicitly, how a learner would make inferences to solve the task at hand based on the data provided, which in turn would help refine the data selection process.

Taking this reasoning one step further, having the experience of being an active learner prior to teaching should generate robust insights about how to select good examples for teaching in similar tasks. Additionally, if both tasks rely on a core ability to sample environmental data efficiently, the transfer could occur automatically during learning, without the knowledge or expectation that the acquired information will need to be used for teaching in the future. Furthermore, active learning should improve teaching performance beyond passive learning (even when the same information content is acquired) if the active selection of data was the crucial driver of the learning effect, rather than the benefit of familiarity with the teaching task or taking the perspective of a learner.

Experiment

In order to test the hypothesis that active learning improves teaching performance, we designed a simple task in which participants were required to both learn a one-dimensional categorization boundary, and teach it, in counterbalanced order. Thus, there were two independent groups of participants in our design, those who learned actively first and then taught, and those who first taught and then performed active learning. In addition, to probe whether the effect learning on teaching performance was specific to active learning, a yoked control group performed the same teaching task after learning passively from watching the active learners labeled queries.

Method

Participants Eighty-eight participants (54 female, M_{age} = 24 years, range = 18 - 42 years old) were recruited from the local population through the university online research participation system and the student union. Ethical approval was obtained from the United Ethical Review Committee for Research in Psychology (EPKEB) in Hungary.

Tasks All tasks (active learning, passive learning, and teaching) consisted of three trials. In each trial, participants were shown eight images in a horizontal array such as the one displayed in Figure 1. Participants were told that the images were sorted left-to-right according to a given "key" feature. For instance, animals were sorted according to their speed relative to body size or the average amount of time they sleep, or foods were sorted by their carbon footprint or their vitamin content. Images belonged to one of two categories (which were clearly marked at the extremes of the image array) according to whether their key feature was below or above a "boundary" (threshold value) which lied between two adjacent images (i.e. at one of seven possible locations). Unknown to the participants, the true boundary locations which dictated the category membership of the images were uniformly sampled in each trial from all the possible locations. [1]

The categories used for the learning and teaching tasks were randomly selected for each participant. Images and category cover stories were only presented once throughout the entire experiment.

In **active learning trials**, participants first saw the image array alongside the description of the categories and the boundary, following which they were told that their task was to find the boundary by querying two images. An image could be queried by clicking on it, which immediately revealed its category membership through the color of the frame drawn around it. After the second query, participants were asked to pinpoint where they thought the boundary was located, again by clicking on one of the seven possible boundaries. Participants received feedback on whether they were correct, un-

[1] Participants were provided with a description of a seemingly objective classification boundary (e.g. that slow and fast animals were separated by the speed of the average human scaled by size). These descriptions were intentionally chosen such that the participants were unlikely to have any strong priors about the location of the boundary. Knowing the participants' prior was essential because it determined the optimal query choice in active learning. The six categories and boundary descriptions used in the experiment were chosen based on a pilot in which participants were asked to select the boundary location by relying merely on their prior knowledge. The distribution of chosen boundaries (across participants) for the items included in the current experiment was not significantly different from the uniform distribution.

lucky (they selected a boundary compatible with the labelled images that was not the true boundary) or incorrect (selected an incompatible boundary).

The **passive learning trials** had the same structure, except that the labels of two images were sequentially revealed to the participants before they had to make their decision about the location of the boundary. Crucially, for each passive learning participant, the images labelled corresponded to the queries of a previous active learning participant.

In **teaching trials**, participants were shown the boundary separating the two categories and were asked to teach it to another participant who they were told would take part in the experiment at a later time. It was made explicit that the other participant would be presented with the same set of sorted images. The participant only needed to click on an image to mark it as an example, and it was automatically labelled. Mirroring the learning tasks, participants were only allowed to provide two examples, which is the number of examples sufficient to fully specify the correct boundary. Intuitively, selecting two adjacent images with different labels is sufficient to identify the boundary in this task.

Materials All the images were selected from the MultiPic databank of standardized color drawings of concrete concepts (Duabeitia et al., 2018).

Procedure Participants were pseudo-randomly assigned to one of three groups: active learning followed by teaching (N = 29), passive learning followed by teaching (N = 29), and teaching followed by active learning (N = 30). The experiment was presented on a 27inch screen in a quiet room and lasted for an average of 20 minutes (unspeeded). Following the experiment, participants completed an open-ended questionnaire about the strategies that they used to solve the tasks.

Quantifying performance Teaching performance was measured by the information gain, $\mathrm{IG}_{\mathrm{teach}}$, which is the amount of entropy by which the teacher reduced the imagined learner's prior entropy $\mathbb{H}(b)$ by labelling two images:

$$\mathrm{IG}_{\mathrm{teach}} = \mathbb{H}(b) - \mathbb{H}(b|s_1, s_2, l_1, l_2)$$

where s, l, and b respectively denote image stimuli, category labels, and potential boundary locations. $\mathbb{H}$ is the Shannon entropy over the possible hypotheses, the prior entropy is $\mathbb{H}(b) = -\sum_{b\in\mathcal{B}} \mathrm{P}(b)\log_2 \frac{1}{\mathrm{P}(b)}$, where $\mathrm{P}(b)$, the learner's prior over the boundary locations, is assumed to be uniform. The optimal teaching strategy is to label the examples immediately preceding and following the boundary as this will eliminate any uncertainty about the location of the boundary, thus reducing all of the original entropy. On the other hand, selecting an example set that will leave the learner uncertain about the true hypothesis because many potential boundaries compatible with the example set will translate into a lower information gain.

Using the observed information gain to evaluate active learning performance would introduce arbitrariness since it cannot distinguish a learner's well-planned query from a lucky one. An ideal learner should choose a query in light of their uncertainty about the labels that will be observed. First, learners should compute the expected information gain of the queries by weighing the posterior entropy by the probability of observing the given labels for the query made and then choose the query that maximizes the expected gain. Therefore, $\mathrm{EIG}_{\mathrm{learn}}$, the sum of the expected information gain of the first and second queries, was used instead of observed information gain. The expected information gain of the first query is:

$$\mathrm{EIG}_{\mathrm{learn}}(s_1) = \mathbb{H}(b) - \sum_{l_1\in\mathcal{L}} \mathbb{H}(b|s_1, l_1)\cdot \sum_{b\in\mathcal{B}} \mathrm{P}(l_1|s_1, b)\,\mathrm{P}(b)$$

After observing the first label, the prior over the boundary locations is updated, and the expected information gain is computed again relative to the entropy remaining after the first labelled sample:

$$\mathrm{EIG}_{\mathrm{learn}}(s_2|s_1) =$$
$$= \mathbb{H}(b|s_1, l_1) - \sum_{l_2\in\mathcal{L}} \mathbb{H}(b|s_2, l_2, s_1, l_1)\cdot \sum_{b\in\mathcal{B}} \mathrm{P}(l_2|s_2, s_1, l_1, b)\,\mathrm{P}(b)$$

Unless otherwise specified, statistical analyses of participants' responses were performed based on the average measures of $\mathrm{IG}_{\mathrm{teach}}$ and $\mathrm{EIG}_{\mathrm{learn}}$ in the three trials of each task.

Decisions about the boundary location In learning trials, after observing two labelled stimuli, participants marked the location of the categorization boundary. Their choice could be assessed based on whether or not the selected boundary was compatible with the labelled images they had seen. However, simply using the proportion of compatible answers (across the three trials) to assess their performance ignores the fact that trials differed in the number of remaining compatible boundaries. To control for this and characterize performance appropriately, we fitted a model that captured the intuition that participants behaved optimally and selected (randomly) from among the remaining compatible boundary locations in some r fraction of trials, while in the rest of the trials they "lapsed" and selected a boundary randomly among all locations:

$$\mathrm{P}(\text{choice} = b_i|s_1, s_2, l_1, l_2) =$$
$$= r\cdot \mathbb{1}\{b_i \in \mathcal{B}^{(i)}_{\text{compatible}}\}\cdot \frac{1}{\left|\mathcal{B}^{(i)}_{\text{compatible}}\right|} + (1-r)\cdot \frac{1}{|\mathcal{B}|}$$

Thus, $r = 1$ indicates optimal behavior, while $r = 0$ indicates chance performance. We estimated r for each participant by maximum likelihood (under the assumption that trials were *i.i.d*).

Data analysis Predictions were tested using planned independent t-tests to compare the teaching information gain in the teaching first and learning first conditions. Paired comparisons were used for the two groups who experienced being learners first, the active learners and passive learners.

Post-hoc analyses were conducted to ensure that variables extraneous to the predictions did not have a meaningful impact on performance or modulate the reported effects. The design of the experiment lends itself naturally to mixed model analysis, since it allows fitting trial level data (without aggregation) and can describe variation arising from the experimental design. Starting from a baseline fixed effects only model with the experimental condition as a predictor of teaching performance, we sequentially fitted and compared models using two additional fixed factors, learning performance and trial number (and their interactions with the condition), as well as random intercepts for participant and trial identity (i.e. dimension used for classification of the objects). Fixed effects were tested using log-likelihood ratio tests for nested models with the same random effects structure. Non-nested models were compared using the Bayesian Information Criterion (BIC) and Aikaike Information Criterion (AIC). Similarly, random effects (fitted via maximum likelihood) were tested using log-likelihood ratio tests while keeping the fixed effects model identical. Given that the mixed-effects analysis confirmed the results of the planned comparisons on the aggregated trial data, we will focus on these comparisons in the Results section for brevity and clarity.

Results

Descriptives Despite the surface level simplicity of the teaching task, a large proportion of participants ($\approx$ 60 %) did not perform it optimally (i.e. did not choose the two images on either side of the boundary as the teaching samples). However, prior active learning made it easier to gain insight into the optimal solution for teaching. More than half of active learners, 17 out of 29 participants, performed at ceiling level by consistently providing example sets compatible with only one categorization boundary. In contrast, only 11 of 29 participants in the yoked passive learning group, and 7 out of 30 of the participants who did not complete a learning task before teaching managed to select the optimal example sets.

Teaching performance across conditions As predicted, participants who were active learners before being teachers outperformed those who started directly with teaching, on average providing .63 bits, 95% CI [.22, 1.05], of additional information to their (fictitious) learners (see Figure 2). The group difference was highly significant in an independent t-test, $t(57) = 3.04$, $p = .01$, Bayes Factor (BF) [2] = 10.81 in favor of the alternative hypothesis.

Learning passively before teaching conferred a smaller, but still significant, benefit relative to foregoing learning. Passive learning increased teaching information gain by an average of .45 bits, 95% CI [.05, .85], $t(57) = 2.26$, $p = .03$, $BF = 2.16$ in favor of the alternative.

While we found strong evidence in support of the differences between the groups completing the learning and teaching tasks in different orders, a possible concern was that these differences were not induced by the experimental manipulation *per se*. Specifically, if there are prior differences in learning performance favoring the group that completed the active learning task first, and learning performance is correlated with teaching performance, then the condition effect could be just an artifact. In order to eliminate this possibility, a regression was performed on teaching performance with both the group (active learning before / after teaching[3]), learning performance, and their interaction as predictors. The group difference remained significant, $\beta = .62$, $p = .01$, when controlling for expected information gain in learning, which was not a significant predictor of teaching ability, $\beta = .08$, $p = .81$, nor did it interact with the group effect, $\beta = .68$, $p = .3$. Figure 2 shows, for each condition, the estimated (non-significant) slopes for information gain from teaching as predicted by ex-

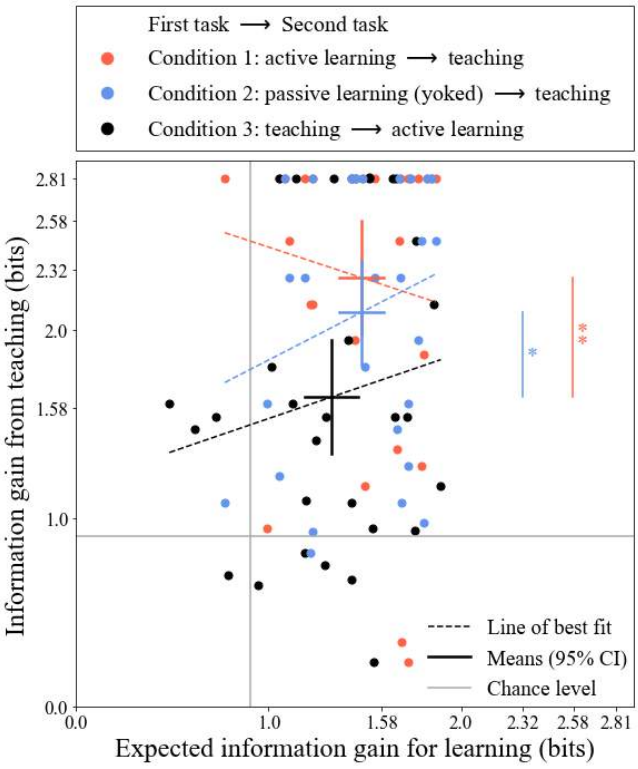


Figure 2: Teaching and learning performance across the three conditions. Each dot represents the information gain for one participant, averaged across the three trials of each task. Crosses represent the 95% confidence intervals for the group means. Dotted lines represent the expected mean information gain from teaching as a function of expected information gain. The maximum information gain for the task is 2.81 bits. The asterisks mark significance levels in independent t-tests (* $p < .05$, ** $p < .01$).

[2]Bayes Factors were calculated for a null model that assumes a zero standardized difference between groups, and a Cauchy alternative with a prior scaled to an effect size of .7, following Rouder, Speckman, Sun, Morey, and Iverson (2009).

[3]The same pattern of results was found for the difference between the group learning passively and then teaching, and the one teaching before active learning.

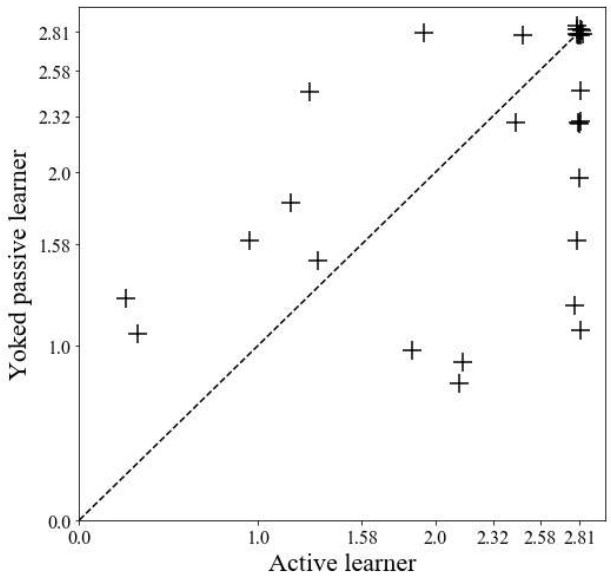


Figure 3: Teaching performance for the active-passive learning dyads. Each dot represents the information gain from teaching for one dyad. In dyads situated under the diagonal identity line, the active learner was the better teacher. A small Gaussian scatter was applied to make overlapping dots visible.

pected information gain for learning. Coupled with the fact that the difference in active learning performance between the two groups was not significant, $t(57) = 1.77$, $p = .08$, $BF = 1.28$ in favor of the null hypothesis, this suggests that the effect of the manipulation was not mediated by prior differences in active learning performance. To investigate this issue further, the two groups were repeatedly resampled with replacement such that the learning performance between groups could be matched and fixed at different levels. Comparing the teaching performance across these resampled groups confirmed the advantage of those who completed the learning tasks prior to the teaching task (the 95% CI of the mean of the resampled groups' differences did not include a null effect).

The second prediction of the study was that active learners would gain a larger benefit from learning before teaching than the yoked passive controls. Active learners fared on average only slightly better in the teaching task than their passive learning counterparts who were shown the same labelled data, with an average difference of .18 bits, 95% CI [-.11,.47]. The dyads' performance is illustrated in Figure 3. The difference was not significant in a paired t-test, $t(28) = 1.29$, $p = .21$, $BF = 2.39$ in favor of the null. It should be noted though that the paired comparison was underpowered (post-hoc power = .24) given the magnitude of the effect size observed.

While there was no significant difference in teaching performance in the planned, marginal comparison between the dyads, Figure 2 suggests that differences may potentially be present conditional on learning performance. There was no interaction between the three-level condition and learning performance, however, this analysis does not account for the dependence in the active learning and passive learning dyad data. As pairs of active and yoked passive learners had, by design, the same expected learning information gain, we regressed the within dyad difference in teaching performance against learning expected information gain. Learning performance was not a significant regressor of the difference in teaching, $\beta = -.86$, $p = .07$. On the one hand, the predicted within-dyad difference, conditioned on low values of learning performance, was significant (see Figure 4). For instance, the predicted within-dyad teaching difference was .60 bits, p=.03, at a one bit expected learning entropy. On the other hand, there was no discernible difference for dyads with high expected information gain. While this is not a strong result, given the low number of dyads and the small effect, it might suggest a potential modulation of the relative benefit of active learning.

Mixed effects analysis The best-fitting model contained the condition, $F(2,85) = 4.30$, $p = .02$, and trial number, $F(2,174) = 6.93$, $p = .01$, as fixed effects, alongside a participant level random intercept ($SD = .70$). The addition of the random intercept was judged meaningful based on the magnitude of the variance at the participant level (SD =.70). It also led to a reduction in BIC, from 796.6 for the fixed effects only model to 749.7.

The previous results regarding the condition effect hold, with a significant estimated difference of .63 bits, $se = .22$, $t(85) = 2.94$, $p = .01$, between the active learning first and teaching first conditions. Similarly, no significant difference was found between active and passive learners, estimated difference of .32 bits, $se = .22$, $t(85) = 2.94$, $p = .01$. Additionally, teaching performance improved from the first to the third trial by an estimated .38 bits, $se = .11$, $t(174) = 3.30$, $p = .01$. However, performance improvement from the first to the second trial was not significant, .02 bits, $se = .11$, $t(174) = .17$, p= .87.

Decisions about the boundary location In the active learning first condition, the mean of the best-fit individual r values was .79 (SD = .35), whereas for those completing the active learning following teaching it was lower, .58 (SD = .42). Yoked controls has the smallest average r, .51 (SD = .38). Active learners made better inferences about the boundary location than their matched controls as the average within-dyad difference in estimated probability r was .28, $t(28) = 2.99$, $p = .01$, $BF = 7.29$. The order of the active learning task led to marginally significant differences in an independent t-test, $t(57) = 2$, $p = .05$, $BF = 1.37$ in favour of the alternative.

The difference in r within active-passive learning pairs did not correlate significantly with differences in teaching performance, $r(26) = -.28$, $p = .13$.

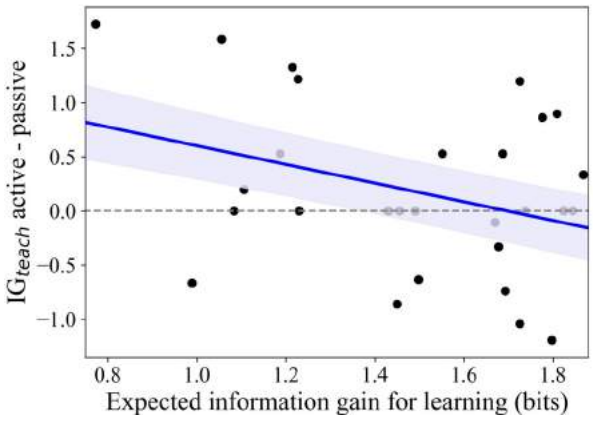


Figure 4: The difference in teaching information gain within dyads of active and passive learners as a function of the expected information gain for (active) learning. The fitted OLS regression line is shown alongside its 95% confidence bound.

Discussion

It has been proposed that humans have a likely innate sensitivity to pedagogical guidance (Csibra & Gergely, 2009) and a propensity for teaching others. From a normative standpoint, the prevalence of teaching in social groups is to be expected given that learning from others who are knowledgeable, well-intentioned and attuned to the learner is more efficient than self-guided learning. Experimental evidence is also accumulating to suggest that, at least in constrained laboratory settings, the behavior of human teachers matches the predictions of normative models (Shafto et al., 2014). However, while we know that humans are effective and keen teachers, we don't know much about the underlying abilities enabling teaching and how it relates to performance in other tasks, specifically active learning.

In the current study we observed an improvement in teaching performance for participants who engaged in active learning prior to teaching. Three active learning trials, using different stimulus sets than those used for teaching, were sufficient for the majority of participants to gain insight into the optimal solution of the teaching problem on the first attempt. Furthermore, they were able to draw on their experience as learners even though at the time of learning they had not been aware that the teaching task would follow.

The poor performance of participants with no learning experience resonates with previous findings of Khan, Zhu, and Mutlu(2017), who used a boundary teaching task as well, but did not constrain the example set size by their design. It seems that simply asking teachers to generate the minimally sufficient number of examples for optimal teaching was not enough to solicit the optimal solution.

The fact that the active learning benefit, relative to teaching first, was not modulated by the initial active learning performance suggests that active learning can improve teaching across the board, for poor and good active learners alike. However, prior active learning performance may play a role in differentiating teachers in a more complex teaching scenarios. Indeed, the surprising lack of a significant correlation between active learning performance and subsequent teaching performance can be explained by ceiling effects.

The impact of passive learning on teaching, relative to the baseline teaching first group, was smaller than that observed for active learning. However, we did not find a significant effect in the matched comparison between active and yoked passive learners. It is important to note here that the current task can be thought of as an insight problem, which means that there was less scope for observing gradual differences in performance. Further, once insight was achieved in the learning task, the solution was easy to verbalize, allowing the optimal strategy to be explicitly transferred to the teaching task.

On the other hand, for poor performing learning dyads, we observed a difference in the predicted direction. This suggests that in a more complex and ecological task in which the learning is more gradual, and the optimal solution is explicitly unknown to participants, active and yoked passive learners are likely to diverge more in terms of teaching performance. This would provide evidence for a more automatic, implicit link between active learning and teaching. In such a future teaching task it would also be interesting to examine whether the differences between active and passive learners, matched for information content, are moderated by the quality of the queries they both observe. Specifically, it should be tested whether the negative linear trend we observed generalizes to non-insight tasks.

Lastly, it is surprising that those who performed the teaching task prior to the active learning task did not differ in their expected information gain in the learning task, and, if anything, performed poorer than their counterparts who started by active learning. This resonates with previous experimental evidence from the developmental literature that has also highlighted more subtle ways in which being taught can hinder learning, for instance by limiting subsequent exploration (Bonawitz et al., 2011). It is an intriguing idea that, perhaps, not just the experience of being taught, but also teaching itself, can have an effect on exploration. Alternatively, if we assume that the teaching task is more cognitively demanding as it has a meta-cognitive component engaged in reasoning about the learner's knowledge and inference making, results can be explained by the known effect that an easier-to-harder progression of tasks is beneficial for learning, while the opposite order does not provide an appropriate stepping stone for active learning. On the other hand, Yang et al. argue that active learning can be re-formalized to also include a meta-cognitive aspect, reasoning that is applied reflexively to one's own reasoning.

To conclude, active learning proved to be a reliable intervention to improve teaching performance. It is important to investigate if the effect of active learning generalizes to more

complex and more ecologically valid tasks, or even between different learning and teaching tasks. If it does, it will open the way for quantitative inquires about whether successful teaching benefits from the ability of taking the perspective of an active learner and as such can be improved by prior active learning.

Acknowledgements This research has received funding from the European Research Council (ERC) Consolidator Grant (ERC-2016-COG/726090) awarded to Máté Lengyel.

References

Bass, I., Shafto, P., & Gopnik, A. (2017). I know what you need to know: Childrens developing theory of mind and pedagogical evidence selection. In *Proceedings of the 39th annual conference of the cognitive science society* (p. 6).

Bonawitz, E., Shafto, P., Gweon, H., Goodman, N. D., Spelke, E., & Schulz, L. (2011). The double-edged sword of pedagogy: Instruction limits spontaneous exploration and discovery. *Cognition*, *120*(3), 322–330.

Bruner, S., J. (1961). The act of discovery. *Harvard Educational Review*, *31*, 21-32.

Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in Cognitive Sciences*, *13*(4), 148–153.

Duabeitia, J. A., Crepaldi, D., Meyer, A. S., New, B., Pliatsikas, C., Smolka, E., & Brysbaert, M. (2018). MultiPic: A standardized set of 750 drawings with norms for six European languages. *Quarterly Journal of Experimental Psychology*, *71*(4), 808–816.

Khan, F., Zhu, X., & Mutlu, B. (2017). How do humans teach: On curriculum learning and teaching dimension. *Advances in Neural Information Processing Systems 30 (NIPS 2017)*.

Markant, D. B., & Gureckis, T. M. (2014). Is it better to select or to receive? Learning via active and passive hypothesis testing. *Journal of Experimental Psychology: General*, *143*(1), 94–122.

Mayer, R. E. (2004). Should there be a three-strikes rule against pure discovery learning? *American Psychologist*, *59*(1), 14-19.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237.

Shafto, P., Goodman, N. D., & Griffiths, T. L. (2014). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive Psychology*, *71*, 55–89.

Yang, S. C.-H., & Shafto, P. (2017). Teaching Versus Active Learning: A Computational Analysis of Conditions that Affect Learning. In *Proceedings of the 39th annual conference of the cognitive science society.*

Yang, S. C.-H., Vong, W. K., Yu, Y., & Shafto, P. (2019). A unifying computational framework for teaching and active learning. *Topics in Cognitive Science*.

Yang, S. C.-H., Wolpert, D. M., & Lengyel, M. (2018). Theoretical perspectives on active sensing. *Current Opinions in Behavioural Science*, *11*, 100–108.

Children's Generalization of Novel Object Names in Comparison Contexts: An eye tracking analysis

Ella Stansbury, Arnaud Witt and Jean-Pierre Thibaut
(ellastansbury@gmail.com; arnaud.witt@u-bourgogne.fr; jean-pierre.thibaut@u-bourgogne.fr)
LEAD-CNRS, UMR 5022, Université de Bourgogne Franche-Comté,
Pôle AAFE – Esplanade Erasme, 21065 Dijon, France

Abstract

A common result is that comparison settings (i.e., several stimuli introduced simultaneously) favor conceptualization and generalization. In a comparison setting, we manipulated the semantic distance between the two training items (e.g., two bracelets versus a bracelet and a watch), and the semantic distance between the training items and the test items (e.g., a pendant versus a bow tie). We tested 5- and 8-year-old children's generalization of novel names for objects. This study is the first one to study the temporal dynamics of comparison in a generalization task with eye-tracking data. The eye movement data revealed clear patterns of exploration in which participants first focused on the training items and compared them with each of the choice options. We also compared the search profiles for correct answers and errors. The results show that participants first found commonalities in the learning items, which they compared with each items in the solution set. This pattern is consistent with an alignment view of generalization.

Key words: comparisons; conceptual distance; generalization; strategies; eye tracking measures

Introduction

Children usually learn the reference of novel words with a limited number of stimuli which are associated with these words. Which learning stimuli lead to accurate generalizations and which mode of presentation would be optimal to achieve this goal are crucial issues for concept learning. A large set of recent studies have shown that comparison settings lead to better generalization results than no-comparison learning conditions. In the latter case, young children tend to generalize novel words to objects that are perceptually similar to the learning items rather than to conceptually related ones (Imai, Gentner, & Uchida, 1994). By contrast, comparison settings favor conceptually based generalizations because they enable children to neglect irrelevant perceptual dimensions and highlight non-obvious properties that need to be identified to choose a taxonomic match (e.g.,(Augier & Thibaut, 2013; Gentner & Namy, 1999; Namy & Gentner, 2002).

However, still little is known of the solving strategies used to process comparison settings and generalize novel words, or of the steps that lead to generalization. In the present study we use eye tracking data to identify these strategies and get a better understanding of the cognitive processes that undergo comparison and generalization during learning.

Comparison and generalization

A large body of research demonstrates the benefits of comparison settings for learning novel object names (e.g., (Graham, Namy, Gentner, & Meagher, 2010), adjectives (e.g., Waxman & Klibanoff, 2000), action verbs (e.g., (Childers & Paik, 2009), objects (Thibaut, 1991; 1995) relational nouns (Gentner, Anggoro, & Klibanoff, 2011; Thibaut & Witt, 2015; see (Alfieri, Nokes-Malach, & Schunn, 2013). For example, Gentner and Namy (1999) presented 4-year olds with familiar objects with an imaginary name and asked them to extend the name. Children had to choose between two pictures, a taxonomic match and a perceptual match. Results showed that children preferred the perceptual match when they had only seen one object during the learning phase but preferred the taxonomic match when they had the opportunity to compare two objects with the same name, introduced simultaneously during the learning phase. The conditions under which comparisons lead to better learning and generalization have received much attention in recent years.

One crucial point is that comparisons generate cognitive costs (e.g., Richland, Morrison, & Holyoak, 2006; Thibaut, French, & Vezneva, 2010b) in the field of analogical reasoning). The hypothesis is that comparing multiple items, and choosing a match, while neglecting irrelevant dimensions including salient dimensions such as perceptual similarities may generate cognitive coasts because of the inhibition, decision making and flexibility involved in the task. For example, Augier and Thibaut (2013) studied conceptualization of unfamiliar objects in a comparison paradigm and manipulated the number of exemplars shown during the comparison phase. They tested 4- and 6-year olds and compared a no comparison condition, a 2-item comparison condition and a 4-item comparison condition. Interestingly, all children benefited from the comparison conditions compared to the no-comparison conditions. However only older children benefited from the four-item comparisons compared to the two-item comparisons. This suggests that cognitive control is necessary to succeed the task as suggested by contributions in numerous domains involving comparisons and integration of multiple information (see (Wiebe & Karbach, 2018).

The semantic distance between the compared items might contribute to increase the cognitive costs of comparison. For example, Green, Kraemer, Fugelsang, Gray, and Dunbar (2010) have shown that analogies based on distant domains were more difficult than equivalent analogies connecting closer domains because distant analogies involved more creativity, which was related to the central role of the prefrontal cortex in cognitive control. In children, Thibaut, French, and Vezneva, (2010a) have shown that semantic analogies based on weakly associated

relations are more difficult than those based on strongly associated relations. The authors interpreted this result in terms of the necessity to inhibit strongly associated but irrelevant items in the context at hand or in terms of the necessity to generate new candidate relations, which requires cognitive flexibility in the case of distant semantic domains.

In this cognitive control framework, it is argued that aligning semantically distant training items might involve deeper conceptual encoding. Indeed, for semantically close items, perceptual similarities are aligned with conceptual similarities (e.g. two apples) whereas for semantically more distant items alignable perceptual similarities are less synonymous of conceptually alignable similarities: aligning surface similarities does not entail an alignment of conceptual similarities or surface similarities are less correlated with conceptual similarities (e.g. a bracelet might look like a watch, but the nature of a watch is strongly connected with a devise giving the time, which can have a low saliency). On the other hand, if alignable perceptual similarities are well correlated with deep similarities for close learning items, the fact that these deeper similarities are embedded in perceptual similarities might prevent them from being easily aligned with conceptually important features when the generalization items are perceptually dissimilar. In that case, generating conceptual similarities might be difficult because the conceptual space cannot be grounded on perception and thus requires more extensive conceptual analysis.

Exploring children's strategies with eye tracking movements in a learning-generalization task

The present study's aim is to analyze the temporal dynamics of a comparison task, from the study of learning items to the selection of a candidate generalization stimulus, which, to the best of our knowledge has never been done. We will use materials by Thibaut and Witt (2017). They manipulated the semantic distance between the learning items (e.g., two bracelets versus a bracelet and a watch), and the semantic distance between the learning items and the generalization items (e.g., a jewel, near distance, versus a bow tie, far distance), and analyzed which combination of conditions would lead to more taxonomic choices. Four-year-old children made less taxonomic choices in the far generalization condition than the close generalization condition whatever the learning distance, whereas only six-year-old children got better results in the far learning distance, a condition in which participants had to coordinate information coming from very different domains. In the above cognitive control context, the authors argued that, as executive functions develop, children are able to compare stimuli from remote conceptual spaces more systematically. Indeed, the common features between two items may be found more easily with semantically close items than with semantically distant items. In the latter case, these features might be less salient and require more comparisons to be noticed. Also, in a broader conceptual space, the set of irrelevant properties to inhibit is likely to be larger than in a close domain.

Recent eye-tracking research on analogical reasoning tasks (another generalization task) have shown that during development younger children's solving strategies differ from older children's and adults' strategies (J.-P. Thibaut & French, 2016). They confronted two main hypotheses to the data, the projection first and the alignment first strategies. Projection-first refers to an initial analysis of the learning domain, in search of a relation connecting A and B. Once a relation is found it is projected on the generalization domain (which generalization item goes with C with the same relation). The alignment-first strategy refers to the alignment of equivalent stimuli (i.e., that play the same role) in the learning and the generalization domains (A with C, and B with D in a A:B::C:D proportional analogy). The authors showed that adults and children followed different search strategies.

In the type of comparison task we use, successful learning requires the learning items (L1 and L2) to be compared and conceptually aligned. Generalization requires switches (witnessing comparisons) between L1-L2 and the Ta(xonomic) target. How one reaches the taxonomic solution (or fails to) will be reflected in the transitions between L1-L2 and the set of the available options (taxonomic, thematic, and perceptual). The set of transitions and the time spent on each item will illustrate the search-construction of a solution.

Among the potential strategies, the projection-first strategy predicts early L1-L2 transitions (finding commonalities between L1 and L2), followed by comparisons between the three generalization options in terms of the features they actually share (either thematic, or perceptual, or taxonomic) with the common feature they have discovered for L1 and L2. The alignment-first hypothesis predicts early comparisons between learning items but also between the learning items and the generalization items, in order to find conceptually analogous items in the transfer set. One additional prediction is that participants compare each learning stimulus with each of the options

Another strategy contrast exists between constructive matching or response elimination (Bethell-Fox, Lohman, & Snow, 1984). Constructive matching predicts early L1-L2 comparisons followed by comparisons between generalization items that may reveal a careful construction of a solution and the application of the solution to the generalization set. This hypothesis makes similar predictions to those from the projection-first hypothesis. Response elimination predicts L1-L2 comparisons followed by back and forth switches between L1-L2 items and generalization options that may reveal a systematic response elimination strategy until a final choice is made. A difference between alignment first and response elimination, is that alignment-first predicts a progressive convergence towards the solution whereas the response elimination predicts no systematic search pattern.

The present study's main goal is to describe the strategies used by children to generalize correctly in a comparison setting, by analyzing eye movement data from two groups of children (5- and 8-year olds). We selected these two age

groups because previous research has shown that participants eye-tracking methods can be used with complex tasks with 5-year olds. Also, (J.-P. Thibaut et al., 2010b) showed that five-year olds might adapt their search strategy to the difficulty of the task in a less systematical way than 8-year olds. Thus, a priori, these two age groups were good candidates for studying strategy differences (if any exist). Also Thibaut and French (2016) showed that reliable results could be obtained with 5-year olds in an eye-tracking task.

We will confront our data with the strategies mentioned above. One hypothesis is that age matters: younger participants should use the response elimination strategy more often than the older group because it is cognitively less demanding: participants compare each transfer with the learning items, one by one, rather than store the found dimensions in working memory and compare all the transfer items in a row. They should also produce less systematic search patterns. For example, correct trials should start with L1L2 transitions less often for young children. There should also be differences depending on the difficulty of the task: easier conditions should elicit less transitions than difficult ones. Far generalization should be more difficult and should result in a larger proportion of comparisons between the options compared to the learning items.

Of particular interest are the differences, if any, between correct and error trials. Do strategy differences between errors and correct trials appear at the onset of the trial (thus, with significant differences in the first slices) or do they result from a wrong decision at the end of the trial (i.e. differences in the last slice), once all the options have been considered.

Methods

Participants 109 French speaking children were tested individually in a quiet room at their school. Two age groups were tested, five year olds, and eight year olds. Forty-nine younger children were recruited (mean age = 5;3; range: 4;11 to 5;9), and 60 children for the older age group (mean age = 8;4; range: 7;11 to 9;4). Informed consent was obtained from their school and their parents.

Materials Fourteen experimental sets of pictures were built, plus three warm-up trials. Each set was associated with a category *(e.g., clothing, food, tools, accessories, animals)*, and was composed of 7 pictures. Two learning objects, either from the same basic level category (close learning, L1 and $L2_C$) or from the same superordinate category (far learning, L1 and $L2_F$) (see Figure 1). The test pictures subsets were composed of three pictures: a taxonomically related generalization object (Ta), either near Ta_N, or distant, Ta_D, see Figure 1), an object perceptually similar to the initial learning object (P) and an object thematically related to the category (Th) (see Figure 1). This design worked as follows. For each object category (e.g., clothing accessories), the close learning objects (L1, $L2_C$) were composed of perceptually and semantically close items (e.g., a bracelet - a curb chain), while the far pairs (L1, $L2_F$) were composed of perceptually similar but conceptually more distant items (e.g., a bracelet – a watch). The three test pictures consisted of three objects in both the near and the distant generalization conditions. The perceptual match (P) was perceptually similar but semantically unrelated to the two training items (e.g., a tire in our bracelet case), the taxonomic choice (Ta) was perceptually dissimilar but taxonomically related to the learning objects and a thematically related object that was not perceptually related but thematically related (Th, e.g., a hand). Depending on the generalization condition, near or distant, the taxonomic choice was semantically near (Ta_N) or more distant (Ta_D) to the learning items (e.g., a jewel pendant in the near generalization case, or a bow tie in the distant generalization case). See Figure 1 for the "clothing accessories" category. Thus, a trial was composed of 5 pictures, L1, L2 ($L2_C$ or $L2_F$) , Th, Ta (Ta_N or Ta_D) and P, resulting in four possibilities (Close learning - Near or Distant generalization; Far learning - Near of Distant generalization.

Independent similarity ratings were obtained from fifty-four university undergraduate students. They are described in Thibaut and Witt (2017). They revealed that the close learning objects in a pair were conceptually closer one to the other than the objects composing the far learning pairs ($p < .01$, see Thibaut & Witt, 2017, for details) and that close generalization stimuli were semantically more similar to the two learning stimuli than far generalization stimuli were (see Thibaut & Witt for details $p < .01$). The same is true for perceptual similarity ratings which also revealed that the perceptual choices were more perceptually similar to the learning material than the objects used to instantiate taxonomic choices, ($p < .01$). For example, for the accessories category, a jewel pendant (near generalization object) or a bow tie (distant generalization object)

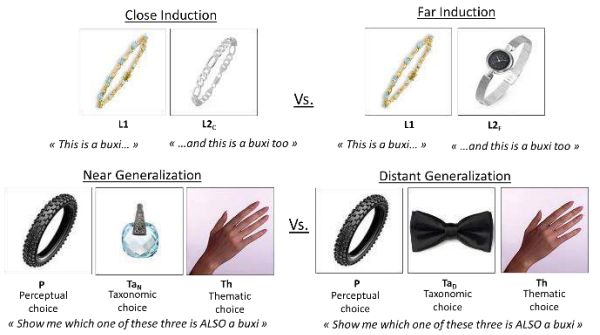


Figure 1: Example of a stimulus set and instructions adapted for the fourteen experimental conditions resulting from crossing Learning distance (Close vs. Far comparison) and Generalization distance (Near vs. Distant generalization) factors.

We forged 14 different bisyllabic labels (pseudo-words) which are, as shown by Gathercole and Baddeley (1993), easier to remember than monosyllabic pseudo-words (e.g., buxi, dajo, zatu, xanto, vira). Syllables were of the CV type which is the dominant word structure in French (from Lexique.org, New, Pallier, Brysbaert, & Ferrand, 2004).

The pictures were displayed on a Tobii T120 eye-tracker device with a 1024x768 screen resolution. The five pictures

of a trial were displayed simultaneously until the answer was chosen. Between each trial a standard fixation cross was shown for 3 seconds. Each experimental session started with a standard calibration phase, after the three warm-up trials. The experiment was run with E-prime®.

The five areas of interest (AOI, L1, L2 , Ta, Th or P) of a trial had a size of 500 by 500 pixels regardless of the object size inside the frame. The frame was chosen as the AOI's outline instead of the picture's outline, to standardize the AOI size.

Procedure The learning pair was displayed at the top of the screen and the test objects at the bottom. First the experimenter introduced the experiment as a game, using the following instructions. "Hello, we are going to play together, and we are going to play with a bear called Sammy. Look, this is Sammy, he lives far away from here and speaks a different language, we are going to learn his language" Then the children saw all three warm-up sets, with the trial instruction, which were followed by eye-tracking calibration. The experimenter then showed the fourteen trials, with the following instructions: "See Sammy's mummy says this is a buxi. And this is a buxi too. Sammy must find another buxi. Can you show which one is also a buxi, to help Sammy? Can you point to the other buxi?". Children chose one of the three test objects by pointing to it on the screen and the experimenter selected it with the mouse.

The presentation order of the fourteen experimental trials, the learning pair objects' position and the generalization objects' position were randomly assigned by the program (e.g., L1 L2, left right or right left, on the top of the screen; Th Ta P, Ta Th P for generalization objects). The names were assigned randomly to each trial. Participants were supposed to know the items. Indeed, these items were calibrated for knowledge by Thibaut and Witt (2017) and, in their experiment, were used with younger children. In their case, the percentage of unknown items was very low.

Design: Five and eight-year-old children were compared. Children were randomly assigned to one of the two experimental conditions (*close* comparison, 55 children or *far* comparison, 54 children). Age was crossed with Learning distance (*close* vs. *far* comparison, between-subject factor) and Generalization distance (*near* vs. *distant*, within-subject factor).

Results

Our first point of interest was the strategies used by the participants to compare and generalize the novel word to the taxonomic item.

Performance data We ran a three-way ANOVA on the percentage of correct taxonomic answers with Age (5, *8 years*), Learning distance (*close*, *far*) as a between factor, Generalization distance (*Near*, *Distant*) as a within factor. This ANOVA revealed a significant main effect of Age $F(1,101) = 29.41$, $p < .01$, $\eta_P^2 = .23$ (5-y-o., $M = 70.91\%$; $SD = 3.31$; 8-y-o, $M = 44.37\%$; $SD = 3.61$). The main effect of the Generalization Distance was significant, $F(1,101) = 31.04$, $p < .01$, $\eta_P^2 = .24$. Age and Generalization Distance interacted, $F(1,102) = 6.61$, $p < .05$, $\eta_P^2 = .06$ (Figure 2). A posteriori Tukey analyses showed that both generalization levels did not differ significantly in the younger group ($p = .18$ $M_{Near} = 45.9\%$ $M_{Distant} = 38.9\%$) whereas 8-year olds had better results for near generalization stimuli ($p < .001$ $M_{Near} = 76.1\%$; $M_{Distant} = 58.6\%$). Both age groups answered significantly above chance (5-year olds, $p < .001$; 8-year olds, $p < .001$).

A one sample t-test revealed that the majority of errors were perceptual matches, (5-year olds: $t = 5.18$ $p < .001$ $M_P = 5.7$ $M_{Th} = 2.42$; 8-year olds: t=2.81 $p < .01$ $M_P = 3.12$ $M_{Th} = 1.49$)

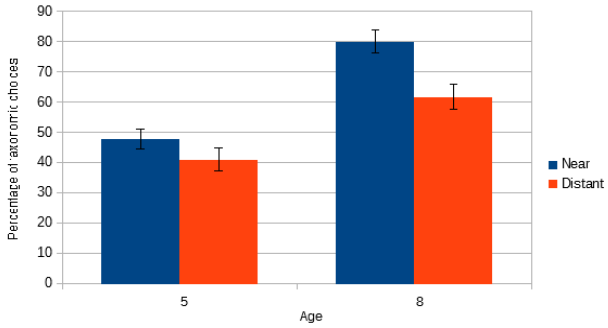


Figure 2: Percentage of taxonomic choices as a function of Age (5 and 8 years) and Generalization Distance (Near, Distant). Error bars are SEM.

We ran the same 3-way ANOVA on reaction times for the items that were correctly answered (see Thibaut & French, 2016). This ANOVA revealed the effect of Age, $F(1,101) = 13.35$, $p < .01$, $\eta_P^2 = .12$, the older children made faster choices ($M = 9539.83$ ms) than the younger children ($M = 11\,618$ ms). Age interacted with Learning distance $F(1,101) = 4.01$, $p < .05$, $\eta_P^2 = .04$ (Figure 3), and with Generalization distance $F(1,101) = 9.62$, $p < .01$, $\eta_P^2 = .09$ (Figure 4). As shown by Figures 3 and 4, the interactions resulted from an opposite pattern in the two age groups, longer RTs in the close and near conditions for the

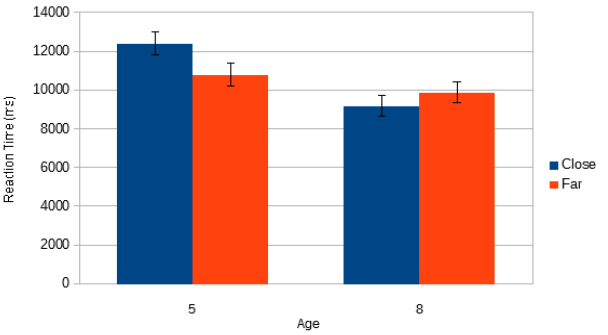


Figure 3: Reaction times (in ms) as a function of Age (5 or 8 years) and Learning Distance (Close or Far). Error bars are SEM.

younger group, and the opposite in the older group.

One interpretation of this pattern of results is that 8-year olds had a high level of performance in all conditions, but that the distant generalization condition was more difficult than the near generalization condition. The higher RTs reflect this higher level of difficulty. For younger children,

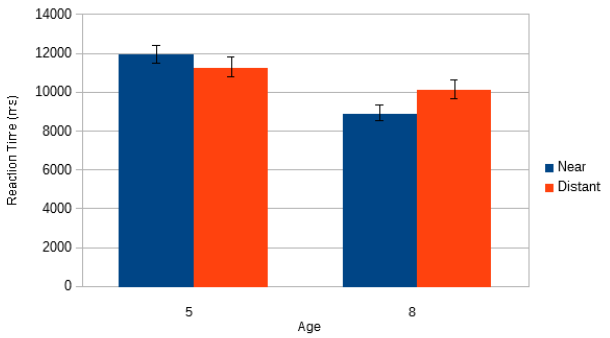


Figure 4: Reaction times (in ms) as a function of Age (5 or 8 years) and Generalization Distance (Near or Distant). Error bars are SEM.

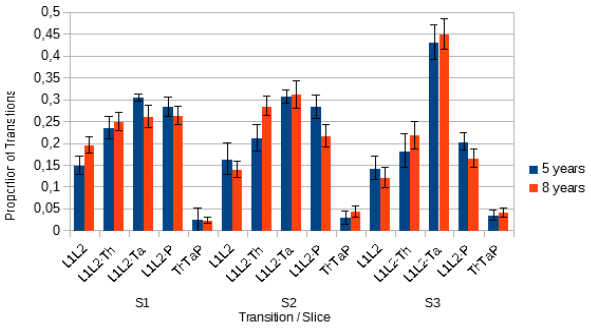


Figure 5: Proportion of transitions as a function of the Slice (S1, S2 or S3) and the Transition type (L1L2, L1L2-Th, L1L12-Ta, L1L2-P, ThTaP) for correct trials. Note: L1L2 are transitions between Learning1 and Learning2; L1L2-Th, between L1 or L2 and Thematic; L1L2-Ta, between L1 or L2 and Taxonomic; L1L2-P, between L1 or L2 and Perceptual; ThTaP, between Th, Ta and P) (Error bars are SEM)

the level of performance was close to chance, and lower RTs might reflect a tendency to answer quickly when the answer was difficult to find, or was less obvious, resulting in shorter RTs

Eight-year-olds had a higher level of performance in both conditions compared to 5-year-olds. Younger children's RTs are lower than the 8 year-old's RTs. However, the younger group does not significantly differ from chance. Chance performance might reflect a tendency to answer too quickly whereas the 8-year-olds RTs are likely to reflect the time necessary for the children to perform a more systematic analysis of a trial before giving an answer.

Eye tracking analyses on transitions (saccades)

The design of the analysis is complex. Since we focus on the temporal dynamic of the search for a solution, interactions involving transitions and time slices are central. A transition (or switch) was defined as a saccade between two stimuli. Each trial was decomposed into 3 time slices (S1-beginning, S2-middle, S3-end) of equal size. We ran a five-way analysis of variance (ANOVA) on the proportions of transitions for *correct* answers, with Age (5 and *8 years*), Learning Distance (*Close*, *Far*) as between factors, Generalization distance (*Near*, *Distant*), slice (S1, S2, S3), and Transition type (L1L2, L1L2-Ta, L1L2-P, L1L2-Th, Ta-P-Th) as within factors. There was a main effect of the Transition type factor, $F(8,640) = 111$, $p < .01$, $\eta_P^2 = .58$. Transition type and Slice interacted, $F(8,640) = 22$, $p < .01$, $\eta_P^2 = .21$. The ANOVA revealed two three-way interactions. The most interesting was the interaction between Slice, Transition type, and Age: $F(8,640) = 2.19$, $p < .05$, $\eta_P^2 = .03$ (see Figure 5). Slice, Transition type, Learning distance also interacted: $F(8,640) = 2.41$, $p < .05$, $\eta_P^2 = .03$, an interaction that we will not analyze here.

Figure 5 shows that all the transition types appeared in the first slice, at the same level, except transitions Th-Ta-P (i.e., between Thematic, Taxonomic and Perceptual generalization items) which are virtually absent in the threeslices. This absence of between-solution transitions is important because it shows that the alignment hypothesis is confirmed (ie. back and forth transitions between learning and generalization items). Second, overall, the general search profile was similar in both age groups. They compared L1 with L2 and each option with L1 and L2 in the first slice and then progressively converged on the correct solution. The large proportion of saccades between L1-L2 and each option is a bit unexpected, since we expected more L1-L2 transitions than any other type. However, it might mean either that from the onset of the trial participants actually looked at all the options at the same rate or that participants looked at L1 and L2 first, and very slightly later transitioned between generalization options and L1-L2 during the first time slice.

In order to disentangle these two possibilities, we ran an ANOVA on the fixation times towards the five AOIs in the first time slice, for correct trials. The analysis revealed a significant effect of AOI, $F(4, 388) = 28.864$, $p < .0001$, $\eta_P^2 = .22$, M = L1 = 27%, L2 = 27%, Th = 15 %, Ta = 15,5%, P = 16%. Tukey HSD revealed that L1 and L2 looking times were significantly larger than the other three AOIs. These results show that children gazed more at L1 and L2 than at the other stimuli at the beginning of the trial, but switched to the options quite early in the trial.

Correct answers and errors profile A last analysis compared the search profiles for correct answers and errors in the younger group only (5-year olds), because the number of errors was low for 8-year olds. An error was either a thematic or a perceptual choice. Two options are possible. First, errors and correct answers have similar search profiles: errors would be the result of a correct search, but followed by a wrong decision. Second, errors might result from different search patterns, which would differ from the onset of the trial. We ran a five-way ANOVA with Learning distance (Close, Far) as a between factor, and Accuracy (Correct, False), Generalization Distance (Near, Distant), Slice (S1, S2, S3), Transition type (L1L2, L1L2-Th, L1L12-Ta, L1L2-P) as within factors. Time slices were defined as the 1st, 2nd, and 3rd thirds of a trial. We excluded the transition Th-Ta-P from the analysis, because its frequency was close to 0. The ANOVA revealed a main effect of Transition type $F(6,162) = 19$, $p < .01$, $\eta_P^2 = .41$; an interaction between Accuracy and Transition type $F(6,162) = 6.64$, $p < .01$, $\eta_P^2 = .19$; an interaction between Generalization distance, Slice and Learning distance,

$F(6,162) = 3.65$, $p < .05$, $\eta_p^2 = .12$. The main result was the interaction between Accuracy, Slice and Transition type: $F(6,162) = 6.70$, $p < .0001$, $\eta_p^2 = .19$ (Figure 6). It shows that the main difference between errors and correct answers takes place during the third slice of the trial, participants focusing on the selected option, error or correct. Note that there were two peaks for errors, on Th and P. This can be related to the predominance of perceptual errors, in the 3rd slice. This suggests that participants studied both incorrect options but, eventually, went for the most salient one. Another interesting feature of this interaction is that the first two slices of the error trials had a flatter pattern than the

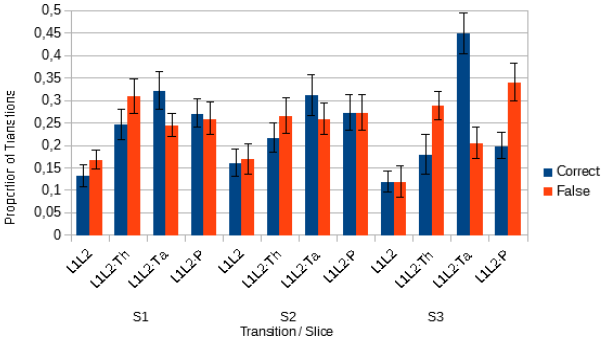


Figure 5: Proportion of transitions as a function of Accuracy (Correct, False), Transition type (L1L2, L1L2-Th, L1L12-Ta, L1L2-P) and Slice (S1, S2, S3). Error bars are SEM

correct answers. This might suggest that errors take hold in the 1st and 2nd slice, that is earlier than at the decisional stage. A priori contrasts between correct answers and errors revealed significantly more L1L2-Th transitions in error patterns rather than in correct trials in slice 1 and significantly more L1L2-Th and L1L2-P for error than for correct trials in slice 3, and significantly more L1L2-Ta in correct than in errors.

Discussion

First, we assessed which learning and generalization conditions would give the best generalization results as a function of conceptual distance between learning items and between learning items and generalization items. Second, we characterized the temporal dynamics of a solution, as a function of age and learning and generalization conditions with eye tracking movements. Generalization was better for near items than for distant items, with a larger difference in older children. This was confirmed by the higher RTs in the distant generalization case, in the 8-year-old group (younger participants results are difficult to interpret since they were at chance). These results might seem straightforward, at first glance. However, we predicted that the difference between near and distant trials should decrease for older children, because they should have a deeper conceptual understanding in the far learning case. Older children's performance significantly differed from chance in the two generalization conditions whereas younger children were at chance in the four conditions. This pattern of results suggests that some of the younger children encountered difficulties to integrate the information resulting from the comparisons, but does not mean that younger children answered randomly, as shown by the difference between error-correct gaze patterns.

The other main contribution, the eye-tracking analysis, revealed a consistent pattern of results across ages. All the comparisons were between learning items (L1-L2) and each of the three types of options, with virtually no comparison of the three options (i.e., no Th-Ta-P transitions). In the 1st slice, the remaining four transitions were equally distributed. However, the looking times on each of the five AOIs in the 1st slice showed that both age groups spent significantly more time on the training items than on the three generalization items, which is consistent with an initial search of commonalities between the learning items before considering the options.

Overall, in terms of the compared solving strategies, the results are consistent with an alignment view rather than with a projection view: participants first compare the learning stimuli, that is align each one with the other. Then comparisons between the learning pair and each option show that participants align commonalities found in L1-L2 with each of the options. A projection interpretation would be compatible with a high number of Th-Ta-P transitions, with participants comparing the three solution options one with the other in terms of the commonalities initially extracted from L1-L2, which occurred very rarely. In a similar way, results are not compatible with a constructive matching strategy. Indeed, as Figure 4 and 5 show, and the discussion above suggests, participants keep on looking at L1-L2 during the entire trial, while testing each option, the latter occurring very early.

There was no interaction between generalization distance and the transitions and slices. The significant interaction between learning distance, transitions and slices, though significant had a small effect size, and seemed to result from small differences, essentially in slice 1. This seems to suggest that generalization distance did not affect the search strategy in a systematic way.

The anatomy of errors Do search patterns for correct answers differ from those for errors? Much of the triple interaction between accuracy, slice and transitions seems to be explained by the distribution of taxonomic, thematic and perceptual choices in correct trials and errors in the 3rd slice (see Figure 5). This pattern would mostly reflect decisional processes at the end of the trial rather than early differences. However, the flatter profile in slice 1 for errors together with the significant difference between errors and correct trials for thematic answers suggest that errors might be prepared early on. This would be consistent with Thibaut and French (2016) who, in their eye-tracking study of analogical reasoning, showed that children's errors differed from correct answers in significant respects even at the onset of the trials.

In sum, younger children had difficulties across conditions whereas the older group could reliably extract the relation in most conditions, especially in the close generalization cases. The eye-tracking measures revealed similar search patterns in both groups of children, with early transitions between L1 and L2 and L1-L2 towards each solution option. Errors seemed to result from an incorrect decision but seemed to be prepared early on, maybe by a less systematic analysis of the taxonomic choice. A more extensive analysis of AOIs looking times and of the order

of the initial gazes should give us a more refined picture of early search steps. We might also analyze the response evaluation processes, for example with an analysis of the distribution of the backward transitions from the options towards L1-L2 separately. These transitions might reflect evaluation of participants' choices.

Acknowledgments

The authors wish to thank the Conseil Regional of the Bourgogne Franche-Comté (PARI program) and the FEDER, for their financial support to this project.

References

Alfieri, L., Nokes-Malach, T. J., & Schunn, C. D. (2013). Learning Through Case Comparisons: A Meta-Analytic Review. *Educational Psychologist*, *48*(2), 87-113. https://doi.org/10.1080/00461520.2013.775712

Augier, L., & Thibaut, J.-P. (2013). The benefits and costs of comparisons in a novel object categorization task: interactions with development. *Psychonomic bulletin & review*, *20*(6), 1126-1132.

Bethell-Fox, C. E., Lohman, D. F., & Snow, R. E. (1984). Adaptive reasoning: Componential and eye movement analysis of geometric analogy performance. *Intelligence*, *8*(3), 205-238. https://doi.org/10.1016/0160-2896(84)90009-6

Childers, J. B., & Paik, J. H. (2009). Korean- and English-speaking children use cross-situational information to learn novel predicate terms. *Journal of Child Language*, *36*(01), 201-224.

Gentner, D., Anggoro, F. K., & Klibanoff, R. S. (2011). Structure Mapping and Relational Language Support Children's Learning of Relational Categories: Structure Mapping and Relational Language. *Child Development*, *82*(4), 1173-1188. https://doi.org/10.1111/j.1467-8624.2011.01599.x

Gentner, D., & Namy, L. L. (1999). Comparison in the Development of Categories. *Cognitive Development*, *14*(4), 487-513. https://doi.org/10.1016/S0885-2014(99)00016-7

Graham, S. A., Namy, L. L., Gentner, D., & Meagher, K. (2010). The role of comparison in preschoolers' novel object categorization. *Journal of Experimental Child Psychology*, *107*(3), 280-290.

Green, A. E., Kraemer, D. J. M., Fugelsang, J. A., Gray, J. R., & Dunbar, K. N. (2010). Connecting Long Distance: Semantic Distance in Analogical Reasoning Modulates Frontopolar Cortex Activity. *Cerebral Cortex*, *20*(1), 70-76. https://doi.org/10.1093/cercor/bhp081

Imai, M., Gentner, D., & Uchida, N. (1994). Children's theories of word meaning: The role of shape similarity in early acquisition. *Cognitive Development*, *9*(1), 45-75.

Namy, L. L., & Gentner, D. (2002). Making a silk purse out of two sow's ears: Young children's use of comparison in category learning. *Journal of Experimental Psychology: General*, *131*(1), 5-15. https://doi.org/10.1037/0096-3445.131.1.5

Richland, L. E., Morrison, R. G., & Holyoak, K. J. (2006). Children's development of analogical reasoning: Insights from scene analogy problems. *Journal of Experimental Child Psychology*, *94*(3), 249-273. https://doi.org/10.1016/j.jecp.2006.02.002

Thibaut, J. P. (1991). Récurrence et variations des attributs dans la formation des concepts. *Unpublished doctoral thesis, University of Liège, Liège*.

Thibaut, J.-P. (1995). The abstraction of relevant features by children and adults: The case of visual stimuli. *Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society*, *17*, 194. Psychology Press.

Thibaut, J.-P., & French, R. M. (2016). Analogical reasoning, control and executive functions: A developmental investigation with eye-tracking. *Cognitive Development*, *38*, 10-26. https://doi.org/10.1016/j.cogdev.2015.12.002

Thibaut, J.-P., French, R., & Vezneva, M. (2010a). Cognitive load and semantic analogies: Searching semantic space. *Psychonomic Bulletin & Review*, *17*(4), 569-574. https://doi.org/10.3758/PBR.17.4.569

Thibaut, J.-P., French, R., & Vezneva, M. (2010b). The development of analogy making in children: Cognitive load and executive functions. *Journal of Experimental Child Psychology*, *106*(1), 1-19. https://doi.org/10.1016/j.jecp.2010.01.001

Thibaut, J.-P., Stansbury, E., & Witt, A. (2018). Generalization of novel names for relations in comparison settings: the role of conceptual distance during learning and at test. *Livre/Conférence Proceedings of the 40th Annual Meeting of the Cognitive Science Society*, 1114–1119.

Thibaut, J.-P., & Witt, A. (2015). Young children's learning of relational categories: multiple comparisons and their cognitive constraints. *Frontiers in psychology*, *6*, 643.

Waxman, S. R., & Klibanoff, R. S. (2000). The role of comparison in the extension of novel adjectives. *Developmental Psychology*, *36*(5), 571-581. https://doi.org/10.1037/0012-1649.36.5.571

Wiebe, S. A., & Karbach, J. (2018). *Executive Function | Development Across the Life Span*. New-York, Routledge.

Using eye gaze data to examine the flexibility of resource allocation in visual working memory

Edmond Stewart (e.stewart@unsw.edu.au)
Chris Donkin (christopher.donkin@gmail.com)
Mike Le Pelley (m.lepelley@unsw.edu.au)
School of Psychology, University of New South Wales

Abstract

Computational models of visual working memory (VWM) generally fall into two categories: slots-based models and resources-based models. Slots-based models theorise that the capacity of memory is defined by a finite number of items. Each slot can only contain one item and once an item is in memory it is remembered accurately. If an item is not in memory, however, there is no memory of the item at all. By contrast, resources-based models claim that all items, rather than just a few enter memory. However, unlike the slots model they are not necessarily remembered accurately. On the surface, these models appear to make distinct predictions. However, as these models have been developed and expanded to capture empirical data, they have begun to mimic each other. Further complicating matters, Donkin, Kary, Tahir and Taylor (2016) proposed that observers were capable of using either slot- or resource-based encoding strategies. In the current experiment, we aimed to test the claim that observers adapt their encoding strategies depending on the task environment by observing how participants move their eyes in a VWM experiment. We ran participants on a standard colour recall task (Zhang and Luck, 2008) while tracking their eye movements. All participants were asked to remember either 3 or 6 items in a given trial, and we manipulated whether the number of items was held constant for a block of trials, or varied randomly. We expected to see participants use more resource-like encoding when the number of items to remember was predictable. Contrary to these expectations, we observed no difference between blocked and unblocked conditions. Further, the eye gaze data was only very weakly related to behaviour in the task. We conclude that caution should be taken in interpreting eye gaze data in VWM experiments.

Keywords: visual working memory; eye gaze; hierarchical modelling

Introduction

In recent years, there have been a number of attempts to describe visual working memory (VWM) using computational models. These models attempt to address fundamental questions such as whether VWM has a strict capacity limit and how likely a stimulus is to be remembered. Broadly speaking, these models fall into two categories: slots-based and resources-based models.

Slots-based models propose that memory functions like a finite set of slots with each slot able to hold one item. The slots-based model proposed by Luck and Vogel (1997) is the prototypical account of this type. If an item is in a slot then it will be remembered. Critically, this account states that if an item is in a memory slot it will be remembered with a very high precision. If it is not in memory, no information is retained about the item. Therefore, if asked to recall an item that is not in memory, a slots-based account assumes that person – having no information about the item – will be forced to guess. Zhang and Luck (2008) expanded on this basic model to create their slots plus averaging model that makes the additional assumption that when the observer has more slots than items to remember, then items are stored in multiple slots. The information in multiple slots can then be combined to produce a more accurate response, thus leading to better performance when set sizes are small.

The resources-based model, on the other hand, conceptualises memory as being more flexible than does the slots model. Memory is described as a resource that is allocated to different items. This memory resource determines the quality of the memories. The more memory resource an item is allocated, the more precise the memory. According to the standard resources model (Frick, 1988) memory is divided equally between all items in the display. Since the amount of memory resource is constant, the more objects there are in a display, the less memory each item is allocated. Unlike the slots model, all items are remembered however, they are remembered with less accuracy as the number of items increases. Beyond the standard model, a variety of resources-based models exist that allow resources to be distributed more flexibly, or models that favour selecting a few items to focus most resources on (Alvarez & Cavanagh, 2004; Bays & Husain, 2008).

The mimicry problem

Zhang and Luck (2008) compared their slots-plus-averaging model to a slots model and a resources model and found that their model provided a better account of the data in a colour recall task. They concluded that the slots-plus-averaging model provided a favourable account of their VWM data. However, this model was challenged by Van den Berg et al. (2012), who developed the resources-based, variable-precision model. Unlike the standard resources model, the variable-precision model assumed that memory resources could be distributed unequally between items in memory. Van den Berg et al. (2012) compared the resources, slots-plus-averaging and variable-precision models and found that the variable-precision model had a better account of the data. Furthermore, in a large scale study, Van den Berg, Awh, and Ma (2014) tested a host of computational models of VWM against the variable-precision model. Using data from multiple experiments across multiple sites it was found

that versions of the variable-precision model tended to provide the best account of the data.

Due to its flexible allocation of memory, the variable-precision model can produce many memory states that closely resemble what is predicted by a slot based model. For example, it is possible that in a 6 items display the mnemonic resource could be allocated equally between four items with no memory resource allocated to the other two. The result of such a memory state would be that four items are remembered with high precision and two are remembered with very low precision. If these very low precision memories are probed, the predictions of the model are indistinguishable from guessing. Such a memory state appears very much like a slots model. On the one hand, such overlap between models is problematic, due to model mimicry. Despite making the same predictions, the interpretations from the slots and resource models are very different. The variable-precision model states that "random" responses are caused by extremely low precision memories, while a slots model says that such responses are not based on memory. At our current level of understanding it is not possible to distinguish between these models. On the other hand, it could be that the mimicry between models represents what is actually shown in individuals. That is, perhaps observers do alternate between slot- and resource-like encoding of VWM displays.

A slot and resource model of encoding in VWM

A recent finding from Donkin, Kary, Tahir, and Taylor (2016) suggested that participants may be able to change their memory "strategy" in VWM tasks. Specifically, they argued that if people know how many items they will be required to remember, they are more likely to use a resource-like encoding, attempting to remember information about all items in the display, compared to if they don't know the set size of the next trial.

In their study, Donkin et al. (2016) analysed data from old and two new experiments with a model that used a mixture of slots-based and resources-based memory processes. The experiments were change detection experiments in which participants were tasked with recalling 2, 4, 6 or 8 items. In one experiment, set size varied from trial to trial, with an equal number of each size in each block of trials (the 'unblocked' condition). In the other experiment, set size was constant within each of four one-hour sessions (the 'blocked' condition). Compared to the experiments with unblocked set size, participants in the blocked experiment appeared more likely to use resource-like encoding (Figure 1). By contrast, participants in the unblocked condition were better accounted for by a slot-like encoding.

The authors suggest that VWM may be more flexibly applied than previously thought. It is possible that the task environment affects how people apply their memory. Perhaps if people know the number of items presented in a trial they will attempt to remember all items instead of focusing on a few. This would increase the chance of an item being in memory, but lower precision on blocked trials relative to unblocked trials and thus following a more resources-like pattern. While the behavioural data in a change detection task appear consistent with this suggestion (when analysed with these particular models), such a claim warrants more evidence to its support.

Here, we present data from a continuous production task in which participants were presented with items in either a blocked or unblocked conditions. To replicate the general results in Donkin et al. (2016), we expect that the number of items remembered should increase in the blocked compared to the unblocked condition (with a corresponding decrease in the precision of memory). As a further test of this prediction, we also use eye tracking to see whether eye movements differ between blocked and unblocked trials, thus suggesting endogenous attention is able to change the strategy used in VWM. We expect that participants in the blocked condition would move their eyes to more items in the display, presumably spending less time fixating on any given item.

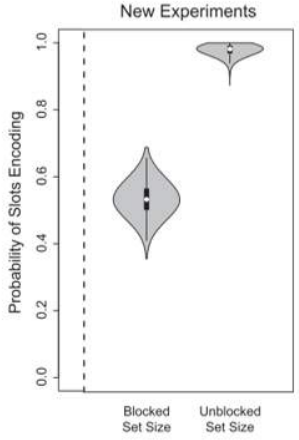


Figure 1: Results from Donkin et al. (2016) depicting the likelihood that participants used slots-like compared to resources-like encoding in the blocked and unblocked (new) experiments.

The current experiment

This experiment aimed to examine Donkin et al.'s (2016) claims that participants may be able to change their memory strategy depending on task environment. We wanted to determine whether 1) flexible memory allocation could be seen in a continuous report task, and 2) eye gaze could provide evidence of a change in memory strategy.

The task used was an adaptation of the colour recall task used by Zhang and Luck (2008), with the addition of eye tracking as well as a between-subjects condition of blocked or unblocked trials. The colour stimuli used in the standard production experiments are very simple to encode (Eng, Chen, & Jiang, 2005). As a result, a participant may be able to encode items quickly. We thought that more complex stimuli would encourage longer fixations and thus provide more data to assist our analysis. While more complex stimuli were desirable, it was also necessary to have stimuli that could be reproduced from a continuous range (the key benefit of colour stimuli). To this end, we used a "ring" set of stimuli. Shown in Figure 2a, these stimuli consisted of a coloured ring

with a "bead" placed randomly on the ring's circumference. Participants were asked to place the bead on the ring as it appeared during study (Figure 2b). The stimuli were presented in set sizes of N = 3 or 6 for 1000ms.

By introducing eye gaze to this task, we hope to observe some differences in attention between our blocked and unblocked conditions. In tasks in which participants can move their eyes freely – such as a visual search task – there is little evidence that participants utilise peripheral attention (Findlay & Gilchrist, 2001; Rayner, 2009). As such, where participants fixate their gaze provides a proxy for what they are attending. If participants are able to change their memory strategy it seems likely that there would be differences in attention allocation as well. In order to see every item on a trial, a participant must move their eyes faster for a 6-item trial than for a 3-item trial. In the blocked condition the participant knows the set size of the next item. With this knowledge, it is possible that they prepare to move their eyes more quickly in the 6-item blocks. In the unblocked condition, participants are unsure of the set size on the next trial. While they might encode set sizes of 3 fairly easily, without additional preparation they may not able to see every item when the set size is 6.

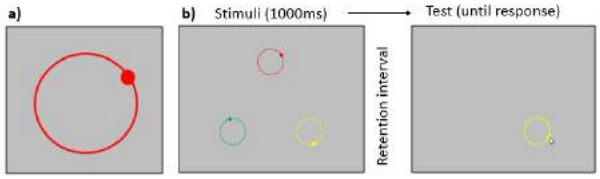


Figure 2: a) An example of the stimuli used. Stimuli varied in colour and location of the bead. b) The trial sequence. 3 or 6 stimuli were presented on a grey background for 1000ms followed by a retention interval (mask then blank screen) of 700ms. Participants were presented with a ring at the study location and were asked to place the bead on the ring as it appeared during study.

It was predicted that 1) similar to Donkin et al.'s (2016) results, we would find an increased probability resources-like encoding in the blocked condition of this experiment. This would be measured by higher chances on an item being in memory and lower precisions when compared to an unblocked condition. 2) We predicted we would see eye gaze data that supported more resources like encoding in the blocked condition. Specifically, more fixations but lower fixation durations compared to the unblocked condition.

Method

Participants 40 participants were recruited from the UNSW sign-up system SONA to complete a single one-hour session. Participants received $15 in exchange for participating.

Apparatus A Tobii TX300 eye-tracker, with 300 Hz temporal and 0.15° spatial resolution, mounted on a 23-inch widescreen monitor (1920 x 1080 resolution, refresh rate 60 Hz) was used. Participants' heads were positioned in a chinrest 60 cm from the screen.

Stimuli The stimuli (Figure 2a) were coloured rings with a filled circle placed on the ring's circumference. They could be one of eight distinct colours (red, yellow, green, cyan, blue, magenta, brown or salmon pink) and were presented on a grey background. All rings had a fixed diameter of 120 pixels (visual angle = 3.03°) and a thickness of 2 pixels (visual angle = 0.05°). Beads had a fixed diameter of 20 pixels (visual angle = 0.51°). Stimuli were presented randomly around the circumference of an invisible circle at the centre of the screen (diameter 600 pixels, visual angle = 15.1°). The angle between items was equal. Beads were randomly placed on the ring for each item, on each trial. The target was indicated by presenting just the ring of the stimulus in the location it had appeared in during study.

Design This recall task followed Zhang and Luck's (2008) design. 420 trials were divided into 14 blocks of 30 trials. Either N = 3 or 6 items were presented on each trial. In the unblocked condition, presentation was randomised per block with an equal number of 3 or 6 item displays per block. In the blocked condition, the first half of the experiment consisted of trials of all one set size and the second half consisted of only the other set size (counter balanced between participants).

Procedure A fixation cross was presented for 500ms at the start of each trial, followed by a blank screen for 400ms. The study array of N rings were presented for 1000ms. This was followed by a mask for 200ms then a blank screen for 500ms. The participant was then presented with a ring they saw at study (same location and colour) and was asked to place a bead on the ring where it appeared during the trial. The participant indicated where they believed the bead was with the mouse and confirmed their selection with the spacebar. Participants received feedback on their selection for 1000ms. Their deviation from the correct bead location was given in degrees alongside verbal feedback ("OUTSTANDING!" for deviations less than 10°, "Very good!" between 10° and 20°, "Good" between 20° and 35°, "OK" between 35° and 45° and deviations greater than 45° were labelled "Poor"). Figure 2b depicts the trial sequence. After each block, participants were given a break for a minimum of 20s before continuing.

Model procedure We used a model to allow us to compare the probability of an item being in memory (P_m) and the precision of memories (Prec) between the blocked and unblocked conditions. The model was a Bayesian hierarchical version of the Zhang and Luck (2008) mixture model (Oberauer, Stoneking, Wabersich, & Lin, 2017). The model assumed that the deviation between given response and the correct response either came from memory or from a separate guessing process. Responses based on memory were associated with Von Mises distributions with a mean that was centred on the correct response and a precision that varied depending on condition (blocked and unblocked), set size (3 and 6) and individual participant. Responses based on guessing were uniformly distributed around the circle for all conditions and all participants. The model allocated responses to either memory or guessing process by taking a value from a Bernoulli distribution with a probability of using

memory equal to P_m. The parameters P_m, like Prec, also varied with condition, set size and individual participant. Thus, four values were estimated for each participant, P_m and Prec for set sizes 3 and 6 (remembering that blocked and unblocked conditions are between subjects). Rather than estimating parameters separately for each participant, we instead constrained individual-participant level parameters such that they came from their own population-level Normal distributions (i.e., one for each parameter in each set size and blocked/unblocked condition). We focus our analysis on the population-level posterior distributions of P_m and Prec across the four conditions of our experiment.

Results

Prior to analysis, trials with no eye gaze data collected were removed (544 trials or 3.24% of trial data). Trials with more than 10 fixations during the presentation window were also removed (964 trials, 5.74% of the data) as were trials where the average fixation duration was less than 125ms (2703 trials, 16.09% of the data).

Behavioural results For each trial, the deviation between the participant's answer and the true bead location was recorded. Since the range of answers varied around the circumference of the circle, the deviation was expressed in radians (π radians = 180 degrees). Figure 3 shows the frequency distribution of deviations for set sizes 3 and 6 (green and red lines respectively) for unblocked and blocked set sizes. Both conditions displayed the typical response pattern for this task (e.g. Zhang and Luck, 2008) with most responses clustered around the correct response for both set sizes but with more accurate responses for set size 3.

Modelling results Figure 4 shows plots of the population-level posterior distribution for P_m and Prec across condition and set size. There was no visible difference in P_m values for set size 3 between the blocked and unblocked conditions. There was a slight indication of a difference between the unblocked and blocked conditions for set size 6, with smaller P_m values in the blocked compared to the unblocked condition. Note that this pattern is the opposite of what we expected. Prec values appear to differ across set size, with higher precision in set size 3 compared to 6. However, there was no observable difference between the blocked and unblocked conditions.

The differences in Pm and Prec values between conditions for set size 6 only are presented in Figure 5. The difference between the posterior distributions for Prec centres on zero, suggesting no difference between conditions in precision for set size 6. The plot of P_m difference shows higher values for P_m in the unblocked condition compared to the blocked condition. However, this difference is small. Since an appreciable mass of the posterior distribution surrounds zero, there is little evidence of a difference between the conditions.

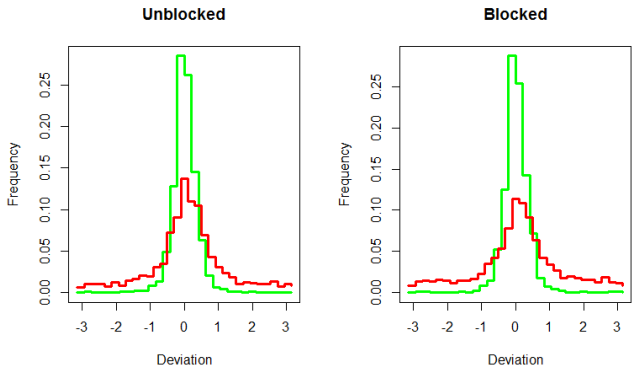


Figure 3: The frequency of responses by deviation from actual bead location for the unblocked and blocked conditions. The green line represents set size 3 the red line represents set size 6.

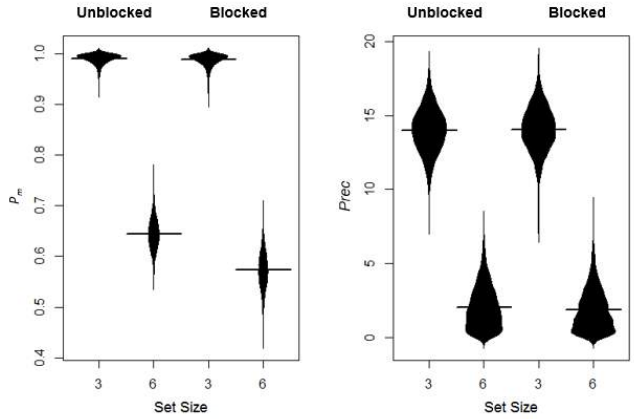


Figure 4: Posterior distribution for Pm and Prec parameters in blocked and unblocked conditions for both set size 3 and 6. Horizontal lines show the mean of the distributions.

Eye gaze results We now compare unblocked and blocked conditions using the average fixation duration per trial and the average number of fixations for each set size. The mean values of each measure in each condition are plotted in Figure 6. On average, the unblocked condition had more fixations and less fixation duration compared to the blocked condition. Again, the qualitative pattern, if present, is in the opposite direction of what was expected.

An analysis of variance (ANOVA) on average number of fixations and average fixation duration yielded no significant effect of condition ($F(1,37) = 1.441$, $p = 0.238$; $F(1,37) = 1.443$, $p = 0.237$ respectively) or set size ($F(1,37) = 0.337$, $p = 0.543$; $F(1,37) = 0.170$, $p = 0.682$ respectively) on either measure. We find no strong evidence for a difference between the average number of fixations or average fixation duration between trials of different set size or condition.

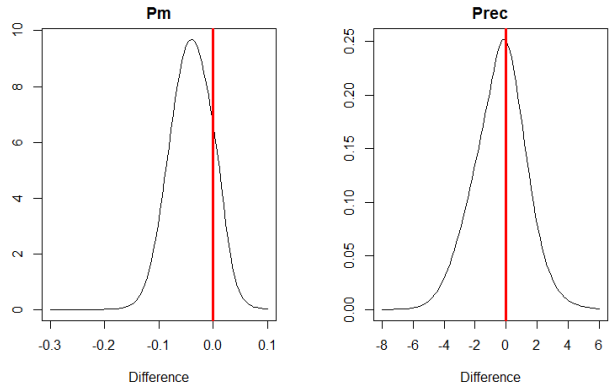


Figure 5: Difference in the posterior distributions of Pm and Prec between the unblocked and blocked conditions (set size 6 only).

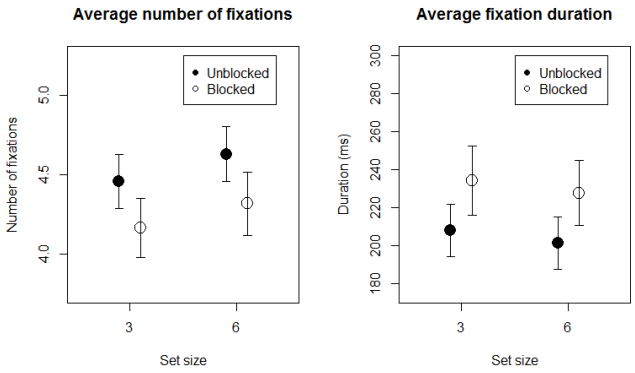


Figure 6: Average number of fixations and average fixation duration per trial for blocked and unblocked conditions for set sizes 3 and 6. Error bars indicate the standard deviations.

Discussion

The behavioural results for both the blocked and unblocked conditions were similar. The modelling results indicated very little difference in the memory strategies between conditions. For each condition, similar values for the probability of an item being in memory and for precision were found. Similarly, there was little difference in eye gaze patterns. There was a suggestion that there was a higher probability of an item being in memory in the unblocked condition. However, there were more fixations with lower durations in this condition as well. This trend is counter to our prediction, that the blocked condition would have a higher probability of items in memory, more fixations and lower average fixation duration.

Overall, these results are not consistent with what we expected based on Donkin et al.'s (2016) finding that memory can be flexibly allocated based on task environment. There were a number of differences in the experimental design between the Donkin et al. experiments and those reported here. The largest difference seems to be that here we used a continuous production task. It may be that participants are less able or willing to adapt their mnemonic resources in production tasks. On the face of it, production tasks require a more precise response than in a recognition/change detection task (in which there are only two responses). In the change detection experiments reported in Donkin et al., it was the blocked condition that was unlike previous experiments. It may have been that participants in our blocked condition did not spread their resources more diffusely in an attempt to remember more items because of the resultant cost to the precision of their memories. Future experiments could encourage participants to accept more error in their response, giving positive feedback whenever a response falls within a particular region around the correct response. Perhaps participants would adjust their mnemonic allocation in blocked conditions (where the number of items to remember is predictable) in such lenient environments. That said, such an explanation is obviously post-hoc, and so we do interpret this data as problematic for a model of VWM that proposes that mnemonic allocation is flexible and under strategic control.

In future work, we aim to connect the eye gaze data and the behaviour of individuals on individual trials. We have conducted preliminary analyses in which we see a weak correlation between fixation duration and the deviation between the correct response and the response given by the participant. We also see that whether an item was fixated during study is a weak predictor of deviation accuracy. These results were much weaker than we had anticipated, and so we will follow up these analyses with more refined methods. In particular, we will use summary statistics from eye gaze data as predictors for the parameters of the Zhang and Luck (2008) mixture model. For example, we might expect that an item not fixated during study would be more likely to come from a guessing process in the mixture model. We would also expect the fixation duration to affect the precision parameter of the memory process in the model. We have carried out versions of these analyses that we are not yet confident enough to report here, but were again very surprised by the lack of relationship between the eye gaze data and the behaviour of participants in the task.

Some of the flaws in the current design need to be addressed to convincingly link eye gaze and memory in this task. For example, one of the problems with the eye gaze data is perhaps that there is not enough distinction in where people are looking (their fixation locations) and their fixation durations. In this task, we suspect it is possible for participants to encode more than one stimulus in a single fixation as these relatively simple stimuli can be encoded quickly. We anticipate that either spatially separating items or more complex stimuli would therefore help distinguish which items a participant has looked at and thus attended and encoded.

Conclusions

Given participants did not move their eyes as much as anticipated, this seems to have impacted the collection of eye gaze information. In turn, the value of using eye gaze as our proxy for attention was thus diminished. As a result, we did not observe the difference in memory strategy between

unblocked and blocked conditions as seen by Donkin and colleagues (2016).

Logically, vision must be helpful in encoding visual items into memory. The lack of a connection between memory and eye gaze in this study is likely due to methodological reasons. As mentioned, it might be necessary to make items more complex or make the display array more separated. However, to what extent alterations need to be made in order to observe an effect of eye gaze on memory remains to be seen. Future experiments could include gaze contingent presentations. Such a paradigm could require participants to fixate on a stimuli for a set period of time within a study array. As a result, there would be more certainty in what participants have looked at and perhaps encoded.

Presently, the current experiment serves as a caution to those interested in investigating VWM tasks using eye gaze.

Acknowledgments

This research was funded by the Australian Research Council (project number DP DP170101684)

References

Alvarez, G. A., & Cavanagh, P. (2004). The capacity of visual short-term memory is set both by visual information load and by number of objects. *Psychological science, 15*(2), 106-111.

Bays, P. M., & Husain, M. (2008). Dynamic shifts of limited working memory resources in human vision. *Science, 321*(5890), 851-854.

Donkin, C., Kary, A., Tahir, F., & Taylor, R. (2016). Resources masquerading as slots: Flexible allocation of visual working memory. *Cognitive psychology, 85*, 30-42.

Eng, H. Y., Chen, D., & Jiang, Y. (2005). Visual working memory for simple and complex visual stimuli. *Psychonomic Bulletin & Review, 12*(6), 1127-1133.

Findlay, J. M., & Gilchrist, I. D. (2001). Visual attention: The active vision perspective. In *Vision and attention* (pp. 83-103): Springer.

Frick, R. W. (1988). Issues of representation and limited capacity in the visuospatial sketchpad. *British Journal of Psychology, 79*(3), 289-308.

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature, 390*(6657), 279.

Oberauer, K., Stoneking, C., Wabersich, D., & Lin, H.-Y. (2017). Hierarchical Bayesian measurement models for continuous reproduction of visual features from working memory. *Journal of Vision, 17,* 1-27.

Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *The quarterly journal of experimental psychology, 62*(8), 1457-1506.

Van den Berg, R., Awh, E., & Ma, W. J. (2014). Factorial comparison of working memory models. *Psychological Review, 121*(1), 124.

Van den Berg, R., Shin, H., Chou, W.-C., George, R., & Ma, W. J. (2012). Variability in encoding precision accounts for visual short-term memory limitations. *Proceedings of the National Academy of Sciences, 109*(22), 8780-8785.

Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature, 453*(7192), 233-235. doi:10.1038/nature06860

Correction of Manipulated Responses in the Choice Blindness Paradigm: What are the Predictors?

Thomas Strandberg (thomas.strandberg@lucs.lu.se)
Lars Hall (lars.hall@lucs.lu.se)
Petter Johansson (petter.johansson@lucs.lu.se)
Lund University Cognitive Science, Lund University, Box 192, S-221 00 Lund, Sweden
Fredrik Björklund (fredrik.bjorklund@psy.lu.se)
Department of Psychology, Lund University, Box 213, S-221 00 Lund, Sweden
Philip Pärnamets (philip.parnamets@ki.se)
Department of Clinical Neuroscience, Div. of Psychology, Karolinska Institute
Solnavägen 1, S-17177 Solna, Sweden

Abstract

Choice blindness is a cognitive phenomenon describing that when people receive false feedback about a choice they just made, they often accept the outcome as their own. Little is known about what predisposes people to correct manipulations they are subjected to in choice blindness studies. In this study, 118 participants answered a political attitude survey and were then asked to explain some of their responses out of which three had been manipulated to indicate an opposite position. Just over half (58.4%) of the manipulations were corrected. We measured extremity, centrality and commitment for each attitude, and one week prior to the experiment we assessed participants' preference for consistency, need for cognition and political awareness. Only extremity was able to predict correction. The results highlight the elusiveness of choice blindness and speak against dissonance and lack of motivation to engage in cognitively demanding tasks as explanations why the effect occurs.

Keywords: choice blindness; attitude change; attitude strength; need for cognition; preference for consistency; political awareness.

Introduction

Choice blindness (CB) is a cognitive phenomenon indicating a dissociation between making a choice and its later justification. It highlights the limitations of our introspective capacity when reasoning about past choices. CB occurs when people receive false feedback about a choice they just made accepting the outcome as their own and reporting seemingly introspective (albeit confabulated) reasons for having made that choice (see Johansson et al., 2005 for details). CB has been reported for many domains and modalities, ranging from taste and smell preferences (Hall, Johansson, Tärning, Sikström & Deutgen, 2010) to eye-witness testimony (Cochran, Greenspan, Bogart & Loftus, 2018), and has been shown to affect both later memories and preferences (e.g. Strandberg, Sivén, Hall, Johansson & Pärnamets, 2018; Pärnamets, Hall & Johansson, 2015; Johansson, Hall, Tärning, Sikström & Chater, 2014). CB has also been applied to the study of attitudes and attitude change, an area of research where deliberation and introspection are often seen as important ingredients. In Hall, Johansson and Strandberg (2012) about 60% of manipulations to a survey on moral dilemmas were accepted by the participants' as being their own attitudes. Hall et al., (2013) reported similar findings for salient political issues in the run up for a Swedish general election. In that study participants not only changed their attitudes on political issues, but their actual voting intention was also affected in the direction of the false feedback. Notably, Strandberg and colleagues (2018) found that when participants accepted the manipulations to political attitudes, these shifted congruently with the false feedback when re-elicited one week later. Although CB is ubiquitous, and undeniably relevant for the study of attitudes and decisions, little is known about what factors that predisposes people to correct the manipulated responses. So far, only a few studies have attempted to establish CB mediators, and thereby link the effect to other psychological constructs (e.g. Strandberg et al., 2018). However, no studies have focused purely on why people correct the false feedback. In this study, we aim to explore several factors that we have identified as meaningful for understanding why correction in the CB paradigm occurs, particularly in the domain of attitudes.

Subjective experience of attitude strength

One possible key to CB susceptibility could be in the relationship between the individual and the attitude itself. This is supported by the literature describing strong attitudes as "resistant to change, persuasion, and contextual influence" and weak attitudes as "unpredictable, malleable, and created in the moment" (Krosnick & Petty, 1995). Given this definition, it seems reasonable that correction of manipulations to attitudes should correlate with attitude strength. Here we tested three self-report measures adopted from Bassili's (1996) seminal work on attitude strength: extremity, centrality and commitment. Extremity directly estimates how strongly a person agrees with an issue on a bipolar scale. Extremity, which is basically just the response to the survey item, is what Bassili calls an operative measure based on first order cognitive processing. Extremity is operative because, for example, the

experienced valence of the extremity could be directly retrieved from memory and not the product of inference. Centrality and commitment, on the other hand, are so called meta-attitudes. These are second order impressions of attitudes that rely on people to report on psychological properties not necessarily represented in long-term memory. As such, meta-attitudes are often inferred from sources more or less relevant to the strength of which the attitude is held. Centrality is described as tapping into the importance of an attitude and how it relates to personal values. Studies show that central attitudes are often more memorable and resistant to persuasion and contextual influence compared to peripheral attitudes (Holland, 2003; Pomerantz et al., 1995). Commitment is described as tapping into the confidence in an attitude: the conviction that the attitude is correct and valid. Commitment has been shown to moderate self-perception and contextual influence in attitudes (Holland, 2003; Pomerantz et al., 1995). Since these measures are meant to capture attitude strength – with strong attitudes being defined by their "resistance to change, persuasion, and contextual influence" – they should also correlate with correction of CB manipulations.

Variation in cognitive style

Another possibility is that aspects of the CB task might be experienced as rather cognitively demanding, such that some individuals may be more susceptible to CB than others due to being less motivated to perform them. Previous studies have shown that individuals with a larger set of general analytic skill are more prone to correct the manipulations (Strandberg et al., 2018). Hence, measures capturing peoples' motivation to engage in cognitively demanding task, such as the *Need for Cognition* (NC; Cacioppo, Petty & Kao, 1984; Cacioppo, Petty, Feinstein & Jarvis, 1996) might also correlate with correction. NC is commonly used in attitude change research, where studies have shown that people with high NC tend to form attitudes that are more resistant to persuasion compared to people with low NC (Haugtvedt & Petty, 1992). CB could also be affected by a consistency motive, which is the case for dissonance phenomena such as cognitive dissonance, cognitive balance, foot-in-the-door etc. These phenomena show that people often change either their behavior or their attitudes to appear consistent (cf. Festinger, 1957). One measure for estimating peoples' need to have consistent cognitions is *Preference for Consistency* (PFC; Cialdini, Trost & Newsom, 1995). Further, PFC has also been shown to predict if people change their attitudes due to social pressure or external demand (Bator & Cialdini, 2006). Thus, if CB share properties with cognitive dissonance phenomena; or if participants accept manipulations due to demand from the experimental situation, correction may correlate with the PFC score.

Variation in political awareness

We would also like to consider variation in political awareness, since much research in political science highlights political awareness as one of the most important factors when forming strong and resilient political attitudes (Zaller, 1992). Interestingly, recent CB studies involving political attitudes have yielded mixed results. In Hall et al. (2012) politically involved participants were more likely to correct the manipulations, and this was not found in Strandberg et al. (2018). However, since political awareness is supposed to determine how people select, interpret and internalize political information (Sidanius, 1988; Lusk & Judd, 1988) we continue to explore the relationship between various measures of political awareness and participants' behavior in a CB study involving political issues.

Thus, we set out to test if susceptibility to correct manipulated responses in CB could be predicted by any of the attitude strength measures, variation in cognitive style, or political awareness described above.

Method

Participants

A total of 128 (70 female) participants, with ages ranging from 18 to 64 years (M = 23.5, SD = 16.8), were recruited to answer a political survey. Sample size was predetermined based on previous CB studies (e.g. Johansson et al. 2005). Ten participants were excluded due to malfunctions with the experimental equipment. Thus, 118 participants remained for the final analysis. The participants were recruited through posters and flyers distributed at the university campuses of Lund and Malmö and compensated with a cinema voucher. At the start of the experiment, we described the general purpose of the study, but without telling the participants that some of their answers would be manipulated. Participants were informed that they could quit the experiment at any time, request their data to be erased, and still receive the cinema voucher. Participants were fully debriefed at the end of the experiment, before consenting to their anonymized data to be used by signing a consent form. All but six participants allowed their interviews to be recorded (leaving a total of 112 verbal recordings to be analyzed). The study was approved by the Lund University Ethics board, D.nr. 2008–2435.

Materials and design

Pre-test One week before the main experiment, participants completed an online questionnaire assessing their demographics, political awareness, PFC and NC. PFC was assessed using the abbreviated 9-item version (Cialdini et al., 1995) with scales ranging from 1 (low consistency) to 9 (high consistency). The PFC questionnaire assessed the participants' internal and external consistency and included items such as: "It is important to me that my actions are consistent with my beliefs". For NC, we used the 18-item version (Cacioppo, Petty & Kao, 1984) with scales ranging from 1 to 9 where a nine gave four points and a one subtracted four points (five gave zero points, and so on).

The NC questionnaire assessed the participants' attitudes towards effortful thinking, and contained items such as: "I usually end up deliberating about issues even when they do not affect me personally". Further, political awareness was established by assessing the participants' political interest with a scale ranging from extremely uninterested (1) to extremely interested (9), and whether they were involved in any political party or organization (yes/no). Visit *https://osf.io/zsy47/* for a list of all measures and items.

Main experiment After the pre-test, participants scheduled to partake in the main experiment being held one week later. It consisted of a questionnaire running on a tablet with a touch-based interface that the participants interacted with using a tablet pen. The experiment consisted of two parts: (1) responding to political issues, (2) explaining the responses, and ended with a full debriefing.

Procedure

Part 1 – responding to political issues During the first part, participants responded to 12 sets of political issues with each set containing a political statement and corresponding six meta-attitudes; three centrality, such as "how important is this issue to you?", and three commitment, such as "how confident are you about your attitude towards this issue?" (visit *https://osf.io/zsy47/* to see all centrality and commitment items). The political issues were selected together with leading political scientists, and represented 12 of the most salient and important issues in Sweden at the time of the study (Table 1). As such, we believe that the vast majority of our participants were familiar with them. This was also confirmed by the verbal reports: most participants were able to intelligibly and knowingly discuss the various issues. Below each item were visual analog scales with endpoints at 0 and 100 (completely disagree to completely agree for the political statements and for example extremely unimportant to extremely important for the centrality item "importance"). The participants were instructed respond to each item by drawing a mark using the pen. They could change their responses as many times as they wanted by clicking a change icon located to the left of each scale, as well as toggle freely between the 12 sets of issues. The participants were left to complete the questionnaire at own, and told to inform the experimenter when finished.

False feedback and correction When going over and explaining the responses, participants had received false feedback on three of the six trials. Trials 2, 4 and 6 had been manipulated by the tablet application to indicate a position opposite to the original (Figure 1). Trials 1, 3 and 5 were non-manipulated controls. The manipulation had two rules: move the participants' rating across the midline of the scale (with a minimum of 5 mm from the middle, i.e. ratings 45 or 55), and then randomly positioned on the opposite axis. If participants in any way indicated that their responses did not correspond with their views, or indicated that something was wrong, the experimenter would tell them that they could change their response if they wanted to, after which they could base their explanation on that response instead. Correction was operationalized when change was clicked and a new response drawn.

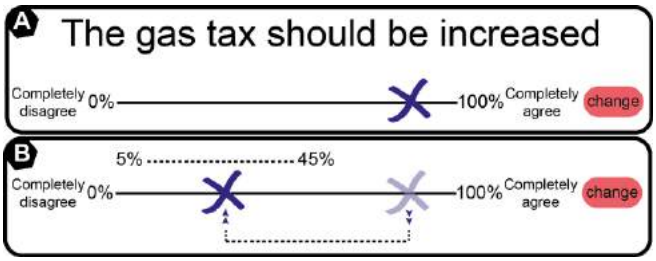


Figure 1: To respond, participants drew an X on a scale going from completely disagree to completely agree (A). On manipulated trials, participant's X was surreptitiously moved from one side of the scale and then randomly placed on the other side (B). Participants could change their X as many times as they wanted by clicking 'change' (A-B).

Table 1: The political issue statements.

1.	The gas tax should be increased
2.	A wealth tax should be reinstated
3.	The labor taxes should be lowered
4.	The monarchy should be abolished
5.	The government should run all elementary schools
6.	The punishment for violent crimes should be stricter
7.	The subsidized service for homework assistance should be abolished
8.	High schools should offer more applied and fewer theoretical courses
9.	Women should be recruited to company boards through affirmative action
10.	Private health care companies should be allowed to make profits in the welfare sector
11.	Copyright protected material from internet should be free to download for personal use
12.	The government should be allowed to monitor telephone conversations and internet traffic

Analysis

Consistent with Bassili (1996) extremity was calculated by taking the absolute value of the deviation between a rating on the 100 point scale and the midpoint. All other variables are reported using their averages. Since attitude extremity, and the difference between the original rating and the manipulated rating, labeled 'manipulation length', are core features in CB studies using rating scales; we first tested how well these would predict correction. In our dataset, extremity and manipulation length were highly correlated, r = .73, $t_{(333)}$ = 19.6, p = 2.2*10^{-16}. To address this we performed our analyses using decorrelated variables by transforming manipulation length to be the distance on the scale the manipulated attitude was moved *beyond* the midpoint. The resulting variables were independent, r = -.028, $t_{(333)}$ = -0.52, p = .61. We then used these two variables to fit a baseline for the other predictor variables (i.e. meta-attitudes and cognitive style). We analyzed our data using mixed regression models including by participant varying intercepts and slopes. Models were estimated in a Bayesian framework using the brms package in R (Bürkner, 2016). Weakly regularizing priors were used for all parameters.

Results

On average participants were moderately interested in politics (M = 6.0, SD = 2.1) and about one fifth identified as politically involved (M = 22.9, SD = 42.2). As we can see in Table 2, extremity, centrality and commitment was rated fairly strong, averaging between 60 to 65 points of 100. The PFC score in our sample was similar to the 48.9 (SD = 10.7) that Cialdini et al. (1995) reported, and the NC score was similar to that reported in a recent meta-analysis of the NC scale (M = 33.2, SD = 10.2 (de Holanda & Wolf, 2018)).

Table 2: Means and SD for the main predictor variables.

Predictor	Mean	SD
Extremity	29.2	13.9
Centrality	63.9	18.2
Commitment	65.0	20.1
NC	29.1	17.8
PFC	44.8	12.6

False feedback correction

Participants corrected 58.4% of the total 347 manipulations. Each participant was exposed to three manipulations and the average correction rate was 1.66 (SD = 0.98), with 15 participants accepting all manipulations and 27 participants correcting all. After correcting a manipulation participants were instructed to replace it with a new response. This corrected rating was on average placed within 9.43 points (SD = 11.7) of their original rating; or -4.45 points (SD = 14.4) when taking the direction of the corrected rating into account (defining a weakened new rating as a negative quantity and a strengthened new rating as a positive quantity). As in previous CB studies, correction did not vary as a function of sex, gender, age, or political party.

Predictors of correction

To test for predictors of correction we conducted mixed-effects logistic regression analyses using standardized variables. We first fit a baseline model consisting of extremity and manipulation length. This model (LOO = 402.77, SE = 14.86) indicated a large effect of extremity on correction (β = 1.77, SD = 0.32, *95% CI* = [1.17, 2.43], BF_{10} > 1.0*10^5), but only a smaller, uncertain effect of manipulation length (β = 0.53, SD = 0.28, *95% CI* = [-0.0043, 1.09], BF_{10} = 1.67), with the intercept estimated as β = 0.46 (SD = 0.19, *95% CI* = [0.10, 0.84]). See Figure 2 for the marginal posterior predictions of the attitude extremity and manipulation length.

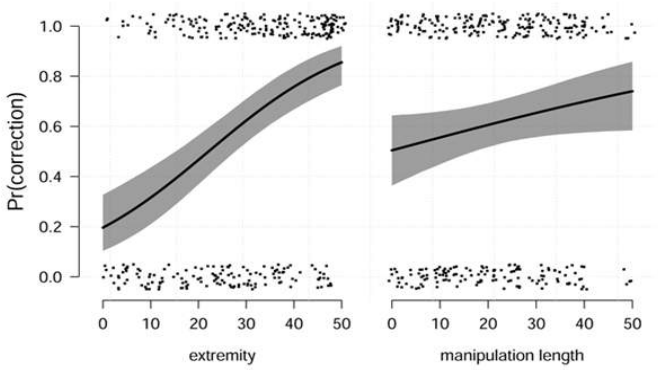


Figure 2: Marginal posterior predictions from the baseline model. Predictions assume other variable held at its average value (0 for standardized predictors). X-axes renormalized to increase interpretability. Shaded regions indicate 95% posterior intervals.

We next fit a full model with all our candidate predictors: extremity, centrality, commitment, preference for consistency (PFC), need for cognition (NC), political involvement, political interest and manipulation length (LOO = 401.65, SE = 17.03). The estimated coefficients, their credible intervals and associated Bayes Factors can be found in Table 3. Marginal posterior predictions are depicted in Figure 3. Notably, when comparing the baseline and full model using LOO we found that the baseline model and the full model did not differ, with a difference of 1.12 (SE = 6.23), this is also mirrored in the estimates where there is little evidence that any of the added predictors are particularly successful at estimating correction.

Predictors of correction types

On an exploratory note, we tried to better capture participants' subjective experience of correcting a manipulation. We conducted a simple classification of the reasons participants reported for wanting to correct.

Table 3: Estimates and Bayes Factors from the full model.

Predictor	Est (β)	SD	95% CI	BF_{10}
(Intercept)	0.51	0.21	[0.12, 0,94]	-
Extremity	1.34	0.38	[0.06, 2.10]	333.69
Centrality	0.44	0.39	[-0.31, 1.23]	0.71
Commitment	0.69	0.39	[-0.06, 1.46]	1.78
NC	-0.07	0.40	[-0.86, 0.72]	0.40
PFC	-0.12	0.37	[-0.85, 0.59]	0.38
Pol.Involvement	0.71	0.47	[-0.21, 1.62]	1.45
Pol.Interest	-0.22	0.43	[-1.06, 0.65]	0.49
Manip.Length	0.59	0.31	[-0.0013, 1.20]	1.94

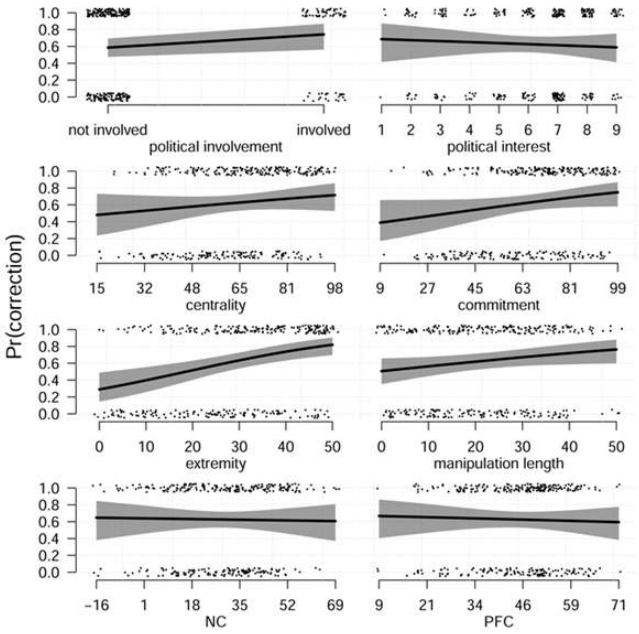


Figure 3: Marginal posterior predictions from the full model presented with the same properties as Figure 2.

One independent rater listened to all the 112 recorded interviews and coded the different reasons participants gave when correcting a manipulation. We identified three distinct types distributed evenly among the corrections: *internal attribution* (36.8%), when participants claimed to have misinterpreted the question, the scale, or something in the task; *external attribution* (33.9%), when participants blamed the experimental equipment; and *change* (29.2%), when participants felt they had spontaneously changed their minds about the issue. Only for a few trials did participants report suspicion that their responses had been manipulated; these were categorized as external attribution. A second rater then classified a subset of 40 interviews; the raters agreed on 90% of the classifications. To test for determinants of the correction types we conducted a hierarchical multinomial logistic regression analysis using correction type as dependent variable and the predictors used in previous analyses. Since we were mainly interested in whether people attributed the wish to correct internally or externally, the change category was used as the reference level in the analysis. Consistent with previous findings, most variables were unable to predict whether participants would attribute correction internally (e.g. feeling that they had made a mistake) or externally (e.g. blaming the experimental equipment). However, we did find that the larger absolute difference between the original response and the manipulated response the more likely participants were to attribute correction internally (β = 2.40, *SD* = 0.57, *95% CI* = [1.03, 3.57], BF_{10} = 1017.52). We also found a small negative effect of political involvement, meaning that participants that were uninvolved politically were more likely to attribute correction externally (β = -1.15, *SD* = 0.77, *95% CI* = [-2.63, 0.37], BF_{10} = 2.36). However, the effect size of this latter finding was very small, but could potentially be a subject for future research.

Discussion

To summarize, we first assessed participants' preference for consistency, need for cognition, and political awareness; and one week later measured attitude extremity, centrality and commitment on a questionnaire containing 12 political issues. Participants were then asked to explain their responses to six of these issues out of which three had been manipulated to indicate the opposite position using the Choice Blindness Paradigm. Just over half of the manipulations were corrected by the participants, meaning that the remaining was accepted by the participants as being their own attitudes. This is similar to previous CB studies on political attitudes (Strandberg et al. 2018; Hall et al. 2012).

Attitude strength

In this study we were particularly interested in testing potential underlying factors that predisposes participants to correct the manipulations. We found that correction was mainly predicted by attitude extremity; meaning that the stronger participants agreed with an issue on the bipolar scale, the more likely they were to correct it. That attitude extremity correlates with correction is also in line with previous CB research (Strandberg et al. 2018; Hall et al. 2012; 2013) and corresponds with for example Bassili's (1996) findings on the relationship between extremity and attitude stability. However, surprisingly, the two meta-attitudes centrality and commitment did not contribute to the correction prediction. One possible explanation to this could be that operative measures of attitude strength, such as extremity, are more relevant to the task compared to second order impressions such as centrality and commitment. Bassili (1996) suggested that extremity is closely associated with the cognitive processing involved in attitude formation and retrieval which is two main components in a CB task. Centrality and commitment on the other hand rather tap into more abstract concepts of the attitude structure (Holland, 2003) not necessarily relevant for scrutinizing one's own survey responses. It could also be that higher extremity is the product of deeper and more involved elaboration,

making those responses more salient and memorable (Petty & Cacioppo, 1986). These results highlight the difficulties in assuming an attitude's strength and stability based on seemingly relevant self-report measures.

Individual difference and cognitive style

The two measures of cognitive style, preference for consistency and need for cognition, were also not able to predict correction.

Preference for consistency In the case of PFC (Cialdini, Trost & Newsom, 1995), we interpret this as an indicator that the correction of CB manipulations is not based on consistency motives or social influence. Further, PFC is mainly about people self-monitoring and being aware about their own consistency; whereas CB corrections tend to occur outside of the participants' awareness. This could be seen in the reasons people reported when wanting to correct: they were almost exclusively about having made a mistake, detected a glitch in the survey application, or having spontaneously changed their minds. Importantly this result also distinguishes CB from cognitive dissonance (Festinger, 1957) and other consistency phenomena that are typically highly correlated with PFC. This is useful when discussing CB and its consequences in a larger theoretical context.

Need for cognition NC (Cacioppo, Petty & Kao, 1984) is often used in social psychology research for its supposed implications to people's attitudes, judgments and decisions. In this literature, NC is described as associated to peoples' tendency to process information and form elaborated and coherent attitudes. Because of this, attitudes of individuals high in NC should be more resilient to change, persuasion, and context effects (e.g. Haugtvedt & Petty, 1992). This is not what we found in this study. However, while individuals high in NC tend to be more resistant to various biases, previous research argue that even these individuals can be influenced if the bias is very subtle (Cacioppo, Petty, Feinstein & Jarvis, 1996). The subtlety factor might help explain why NC and CB correction did not correlate. Further, people with low NC can perform at a comparable level to those with high NC given enough external motivators. One such motivator could be the perception of what participants believe to be their own survey response.

Political awareness The two political awareness measures (political interest and involvement) also did not correlate with correction. While there is nothing uniquely special to *political* awareness per se, the awareness part addresses a domain specific aspect that could determine the participants' understanding, knowledge, and vested interest about the current CB theme (Zaller, 1992). For example, one previous CB study did find that political involvement correlated with correction (Hall, Johansson & Strandberg, 2012), and in this study we found a tendency (albeit small) that politically involved participants were more likely to attribute the correction externally (e.g. believing that there was some error with the equipment). This tendency at least indicates that politically involved participants experienced the false feedback differently from the uninvolved. It could simply be that politically involved individuals have stronger convictions in the politically attitudes; so when they notice a discrepancy between their original and present response, their main explanation is that software application malfunctioned.

Limitations and future studies

The main limitation of this study was the small number of participants. While we only found a relationship between correction and attitude extremity, the lack of relationship between the other variables might at least be partially explained by the small sample size. Thus, one interesting avenue of future research would be to more systematically, and with more participants, test how a variety of attitude, personality, and performance measures affect correction rates and correction types. This would also allow us to examine subgroups within our sample; for example: what is it that makes some participants correct all the manipulations and some accept all? Importantly, while we found no relationship between correction and any of the two motivated cognition measures (NC and PFC), other more performance based variables might be relevant to CB and worth exploring. For example, in Strandberg et al. (2018) the Cognitive Reflection Test (CRT) correlated with correction, with participants having higher CRT score also being more likely to correct the manipulations. CRT is a performance based cognitive processing measure that captures peoples' ability to use reflective and deliberative thinking instead of gut feelings (Frederick, 2005). Thus, future research could try to link CB to performance based measures that taps into working memory, attention, or perhaps factual knowledge. Another potential shortcoming of this study was that the majority of the participants were students. Although we have no reason to believe, given previous studies, that a phenomenon such as CB would drastically differ between different demographics, it is always important to establish whether the experimental findings generalize across the public. However, similar levels of correction have been found in experiments with a more diverse and representative sample (Strandberg, Olsson, Hall, Woods & Johansson, in preparation).

Conclusion

Choice blindness is a cognitive phenomenon powerful enough to influence peoples' opinions and reasoning in important political issues. Still, it is difficult to pinpoint what disposes people to accept or correct the manipulations. It seems that the CB manipulation is so surreptitious that it sometimes flies under the radar even for people with strong convictions and motivations to engage in political reasoning. This study contributes to the understanding of CB, serving as both a backdrop for future research, and an important piece of a broader theoretical puzzle.

Acknowledgments

We would like to thank Anders Lindén at AndVision for implementing the experimental design to a tablet interface, Henrik Ekengren Oscarsson at the University of Gothenburg for helping to select the political issues, and Uno Otterstedts Fund (RFh2014-0390) for financing the movie tickets.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1-48.

Bator, R. J., & Cialdini, R. B. (2006). The nature of consistency motivation: Consistency, aconsistency, and anticonsistency in a dissonance paradigm. *Social Influence, 1*, 208–233.

Bassili, J. N. (1996). Meta-judgmental versus operative indexes of psychological attributes: The case of measures of attitude strength. *Journal of Personality and Social Psychology, 71*(4), 637-653.

Bürkner, P. C. (2016). brms: Bayesian regression models using Stan. *R package version 0.10. 0*.

Cacioppo, J. T., Petty, R. E., & Kao, C. F. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment, 48*(3), 306-307.

Cacioppo, J. T., Petty, R. E., Feinstein, J. A., & Jarvis, W. B. G. (1996). Dispositional differences in cognitive motivation: The life and times of individuals varying in need for cognition. *Psychological Bulletin, 119*, 197-253.

Cialdini, R. B., Trost, M. R., & Newsom, J. T. (1995). Preference for consistency: The development of a valid measure and the discovery of surprising behavioral implications. *Journal of Personality and Social Psychology, 69*(2), 318-328.

Cochran, K. J., Greenspan, R. L., Bogart, D. F., & Loftus, E. F. (2018). (Choice) blind justice: Legal implications of the choice blindness phenomenon. *University of California, Irvine Law Review 8*, 85.

Festinger, L. (1957). *A Theory of Cognitive Dissonance*. California: Stanford University Press.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives, 19*(4), 25-42.

Hall, L., Johansson, P., Tärning, B., Sikström, S., & Deutgen, T. (2010) Magic at the marketplace: Choice blindness for the taste of jam and the smell of tea. *Cognition, 117*, 54–61.

Hall, L., Johansson, P., & Strandberg, T. (2012). Lifting the veil of morality: Choice blindness and attitude reversals on a self-transforming survey. *PLoS One, 7*(9), e45457.

Hall, L., Strandberg, T., Pärnamets, P., Lind, A., Tärning, B., & Johansson, P. (2013). How the polls can be both spot on and dead wrong: Using choice blindness to shift political attitudes and voter intentions. *PLoS One, 8*(4), e60554.

Haugtvedt, C. P., & Petty, R. E. (1992). Personality and persuasion: need for cognition moderates the persistence and resistance of attitude change. *Journal of Personality and Social Psychology, 63*(2), 308-319.

de Holanda, G. L., P., P. H., & Wolf, L. J. (2018). The Very Efficient Assessment of Need for Cognition: Developing a Six-Item Version. *Assessment*.

Holland, R. W. (2003). *On the structure and consequences of attitude strength.* Dissertation Thesis, Radboud University, Nijmegen.

Johansson, P., Hall, L., Sikström, S., & Olsson, A. (2005). Failure to detect mismatches between intention and outcome in a simple decision task. *Science, 310*, 116–119.

Johansson, P., Hall, L., Tärning, B., Sikström, S., & Chater, N. (2014). Choice blindness and preference change. *Journal of Behavioral Decision Making, 27*(3), 281–289.

Krosnick, J. A., & Petty, R. E. (1995). Attitude strength: An overview. In R. E. Petty & J. A. Krosnick (Eds.), *Ohio State University series on attitudes and persuasion, Vol. 4*. Attitude strength: Antecedents and consequences (pp. 1–24). Hillsdale, NJ: Lawrence Erlbaum Associates.

Lusk, C. M., & Judd, C. M. (1988). Political expertise and the structural mediators of candidate evaluations. *Journal of Experimental Social Psychology, 24*(2), 105-126.

Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. *Advances in Experimental Social Psychology, 19*, 123–205.

Pomerantz, E. M., Chaiken, S., & Tordesillas, R. S. (1995). Attitude strength and resistance processes. *Journal of Personality and Social Psychology, 69*(3), 408–419.

Pärnamets, P., Hall, L., & Johansson, P. (2015). Memory distortions resulting from a choice blindness task. In: *Proceedings of the 37th Annual Conference of the Cognitive Science Society*.

Sidanius, J. (1988). Political sophistication and political deviance: A structural equation examination of context theory. *Journal of Personality and Social Psychology, 55*(1), 37-51.

Strandberg, T., Sivén, D., Hall., L., Johansson, P., & Pärnamets, P. (2018) False beliefs and confabulation can lead to lasting changes in political attitudes. *Journal of Experimental Psychology: General, 147*(9), 1382-1399.

Strandberg, T., Olson, J.A., Hall, L., Woods, A.T., Johansson, P. (manuscript in preparation).

Zaller, J. R. (1992). *The nature and origins of mass opinion*. New York: Cambridge University Press.

It's not the treasure, it's the hunt: Children are more explorative on an explore/exploit task than adults

Emily Sumner (sumnere@uci.edu)
Mark Steyvers (mark.steyvers@uci.edu)
Barbara W. Sarnecka (sarnecka@uci.edu)

Department of Cognitive Sciences, University of California, Irvine,
Social and Behavioral Sciences Gateway, Irvine, California, 92697

Abstract

The current study investigates how children act on a standard explore-exploit bandit task relative to adults. In Experiment 1, we used child-friendly versions of the bandit task and found that children did not play in a way that maximized payout. However, children were able to identify the machines that had the highest level of payout and overwhelmingly preferred it. We also show that children's exploration is not random. For example, children selected the bandits from left to right multiple times. In Experiment 2, we had adults complete the task in Experiment 1 with different sets of instructions. When told to maximize learning, adults explored the task in much the same way that children did. Together, these results suggest that children are more interested in exploring than exploiting, and a potential explanation for this is that children are trying to learn as much about the environment as they can.

Keywords: cognitive development; explore-exploit; decision making

Introduction

Imagine that it is your second day at a new job. You are standing at the coffee cart outside your office building, considering the unfamiliar menu. Yesterday you had a cappuccino and enjoyed it; today you must decide whether to get the cappuccino again like yesterday, or try the matcha green tea latte, which you might not like. This is known as an explore/exploit problem, because you must choose between exploiting a familiar option (the cappuccino) or exploiting a new one (the matcha latte). Such problems arise all the time: Do you buy the same brand of hiking boots you just wore out, or try a new style? Make the same old macaroni and cheese that you know your kids will eat, or gamble on pasta puttanesca? Re-watch that Netflix movie that you enjoyed before, or try a new one?

Researchers studying the explore/exploit problem in adults have traditionally defined a good decision as one that maximizes payout and minimizes cost. The problem is commonly operationalized in bandit tasks named after the 'one-armed bandits' (i.e., slot machines) found in casinos. In a bandit task, participants decide between two or more bandits, each of which has an unknown rate of reward. The goal of the task is to maximize return by using a combination of exploration and exploitation. Formally, the optimal strategy is to explore the different bandits just long enough to learn which one pays out best, and then switch to exploiting that one (Mehlhorn et al., 2015). Indeed, that's what most adults do. The explore-exploit problem has been widely investigated in many different contexts, including reinforcement learning (Daw, O'doherty, Dayan, Seymour, & Dolan, 2006; Wilson, Geana, White, Ludvig, & Cohen, 2014), psychiatric populations (Addicott, Pearson, Sweitzer, Barack, & Platt, 2017), and animal behavior (Beachly, Stephens, & Toyer, 1995; Chen et al., 2016; Snell-Rood, Davidowitz, & Papaj, 2011). The present studies asked, What about children?

Very few studies have investigated how children approach explore-exploit tasks, and if they approach these tasks with similar strategies as adults do. A large number of studies suggests that children, and even infants, possess intuitive statistics and a basic understanding of probability (Xu & Kushnir, 2013). Further, there is substantial evidence suggesting that children are better at maximizing statistics in scenarios whereas adults are drawn towards probability matching (Derks & Paclisanu, 1967; Hudson Kam & Newport, 2005). Taken together, this suggests that children would indeed maximize reward on this type of task, perhaps at an even faster rate than adults would. However, Blanco & Sloutsky (2018) had children complete a 4-armed bandit task on a tablet. They found that children do not maximize payout as adults typically do. Instead, they visit each bandit equally across 100 trials.

The key question in this study is that in a bandit task, in its most simplistic form, will children follow similar strategies to adults and attempt to maximize payout. If we find that children do not maximize payout, what are the reasons for their suboptimal performance? Are children viewing the task in the same way as adults?

In Experiment 1, we conduct a simplified version of the bandit task with 159 children. Critically, we designed this task to be simplistic and minimize the memory constraints which other bandit tasks possess. In Experiment 2, we conducted the same bandit task used in

Experiment 1 with adults and give different motives: to learn or to win.

Experiment 1

One explanation for children's over-exploring in Blanco & Sloutsky (2018) could be that they simply cannot figure out which bandit pays out better. In this task, we made it easy for children to see the payouts. To do this, we showed children all of the previous results from each bandit.

We also changed the task structure so that there were three machines: One that paid off every time, one that paid off half the time, and one that never paid off.

Methods

Participants. We tested 159 children between the ages of 3 and 9 (mean 5 years, 7 months; range 2 years, 11 months to 8 years, 11 months). Of those, 69 were girls, 90 were boys. Children were recruited from science museums and preschools in an urban area. An additional 11 participants were tested but excluded because they did not answer both control questions correctly. Participants were given a small toy upon sign-up (e.g., a plastic slinky).

Procedure. Participants were presented with three "Mystery Machines," each with a different proportion of winning (green) and losing (red) balls (see Figure 1). One box dispensed only winning balls, one dispensed only losing balls, and one alternated between winning and losing. The machines associated with different payoffs were counterbalanced across participants. Green balls contained a sticker that the child was allowed to take home; red balls contained no sticker.

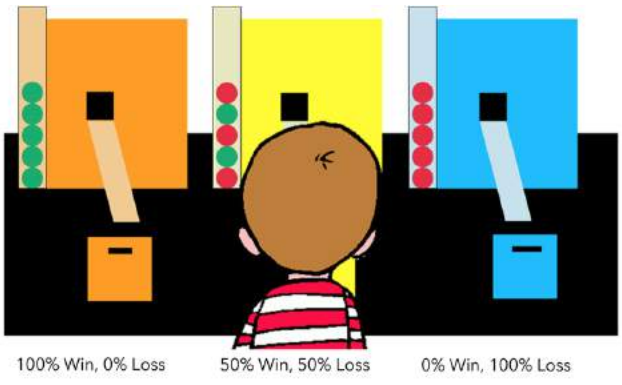


Figure 1: An illustration of the machines at the end of the study for a participant who chooses each machine equally. Green balls have stickers inside, red balls are empty. The tubes are empty at the beginning of the task.

Children were told, "These are Mystery Machines. When you put a coin in the orange box, a ball rolls down the orange slide. When you put a coin in the blue box, a ball rolls down the blue slide. When you put a coin in the yellow box, a ball rolls down the yellow slide. Balls can be green or red. A green ball, like this one, [shows green ball] has a prize that you can take home inside. See? [Opens ball and shows sticker] A red ball, like this one, [shows red ball] has no sticker in it. See? [Opens ball and shows inside.] Some machines have more green balls, some machines have more red balls."

Afterward, children were asked whether they would win any stickers if they received one green ball and one red ball. Those who were unable to correctly answer this question were excluded. Children were then given fifteen coins to put in the coin slots corresponding to the machines of their choosing. When the child put a coin in one of the three coin slots, a machine sound played, and a ball rolled down the slide of the machine corresponding to the coin slot that the child chose. Before the experiment started, a machine operator hid underneath the table at which the children were tested and crawled out from under the table once the child was seated on the other side of the table and thus could not see this person.

The machine operator used a noise maker to make the machine noise and rolled a ball down the slide each time the participant made a selection by dropping a coin in a coin slot. Empty balls were placed in the tube corresponding to the machine that they came from, in order to reduce the number of things that the participant needs to keep track of. This clearly showed the distribution of wins and losses that the participant encountered from each machine (See Figure 3).

After participants completed the task, they were asked the following questions: 1) Which machine was your favorite? 2) Why was the "xxx" colored machine your favorite? 3) Which machine has the most green balls? And 4) How do the machines work?

Results

Few children maximize, many explore each box equally. Children on average pick the winning box 44.72% of the time, a value a Wilcoxon-signed rank test finds as significantly above chance ($V = 6338.5$, $p < 2.2e\text{-}16$). However, this value is still significantly below what maximizing would look like ($V = 297$, $p < 2.2e\text{-}16$). Figure 2B shows each individual child's choice.

Children's behavior is neither random nor optimal. To better understand children's behavior, we simulated two different strategies: a random strategy where the boxes were chosen from a uniform distribution and the optimal policy (from Steyvers, Lee, & Wagenmakers, 2009). We sampled 1000 instances from each strategy so we could compare how children acted to these datasets (See Figure 2 and Figure 3).

Children often explored the machine in a Left to Right pattern as if they were reading a book (see Figure 4) and therefore deviated from a pure random strategy of picking a

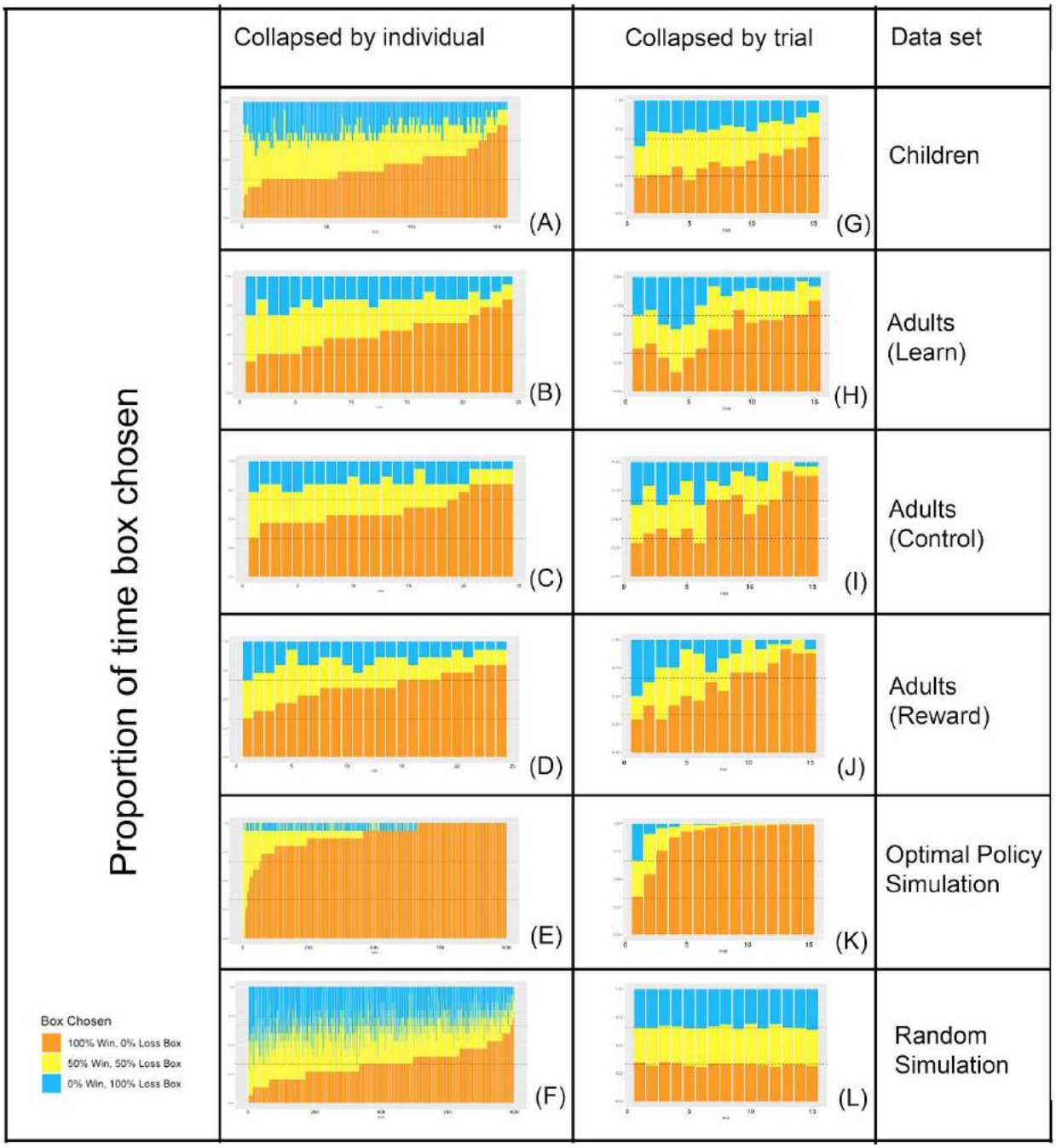


Figure 2: The proportion of choices for each box. The dotted red line indicates chance (.3333 and .6666). Orange indicates the proportion of time choosing the 100% win box, yellow indicates the proportion of time choosing the 50% win box, and blue indicates the proportion of time choosing the 0% win box. Adults in the control, reward, and the optimal policy data chose the winning box significantly more than children. Child data was collected as part of Experiment 1. The adult data were collected as part of Experiment 2. Optimal Policy data was simulated following the optimal policy in Steyvers, Lee, & Wagenmakers (2009). (A-F) Each line is a participant or simulated data point from a particular group. The red dotted line indicates chance (.3333) and (.666). Participants are ordered in increasing levels of picking the winning box. (G-H) Each line represents the proportion of participants that choose each box on a given trial.

machine at chance regardless of the previous machine chosen. For example, they would choose the orange machine, the yellow machine, and then the blue machine. Some children continued with this strategy throughout the entire task. A chi-square test of independence was performed to examine the relationship between the frequency of children going from left to right, and the frequency of this occurring in our randomly generated data set. This relationship was significant. $X^2(5, N = 1159) = 105.52$, $p = < 2.2e\text{-}16$. This suggests that children moving from left to right between boxes would not be plausible under random chance. Children could be acting in this way to strategically explore their environment, and are not necessarily acting purely randomly.

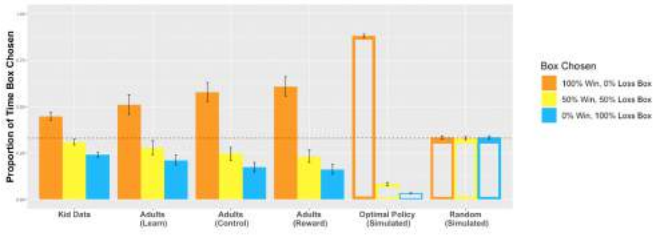


Figure 3: The proportion of times choosing each box. The dotted red line indicates chance (.3333). The error bars indicate 95% confidence intervals. Adults in the control, reward, and the optimal policy data chose the winning box significantly more than children. Child data was collected as part of Experiment 1. The adult data were collected as part of Experiment 2. Optimal Policy data was simulated following the optimal policy in Steyvers, Lee, & Wagenmakers (2009).

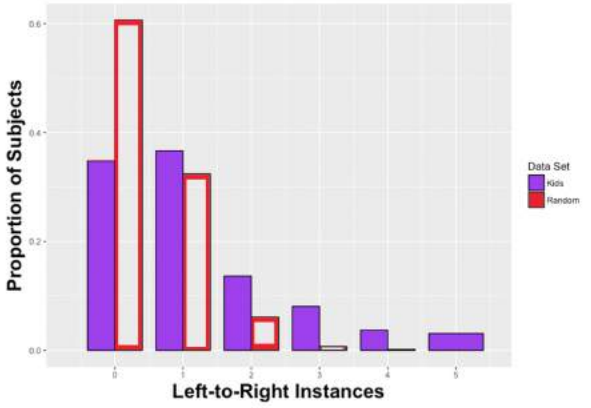


Figure 4: Histogram comparing the Left-to-Right instances made by children (purple columns) and the data set we generated where the choice of machine was random (red-outlined columns).

Children prefer the 100% win box. 128 children (83%) said their favorite box was the one that had 100% winning balls. 15 children (9.7%) prefer the box which alternated between winning and losing balls, 8 children (5.2%) preferred the box that always dispensed losing balls, 5 children (3.2%) said they liked all of the boxes, 2 children (1.3%) said they liked two of the boxes, and 1 child (.06%) said they didn't like any of the boxes (see Figure 5). We performed a proportion test to see if this number was significantly above chance (.3333). We found that this was unlikely due to chance $X^2(1, N = 159) = 159.2$, $p < 2.2e\text{-}16$. Of the 128 children who said their favorite box was the one that had the highest payout, we asked the children why the box they chose was their favorite, 105 children (97.2%) said it was because it gave the most green balls. The other explanations included that they liked the color of the box, or they gave an unrelated answer such as, "because rainbow."

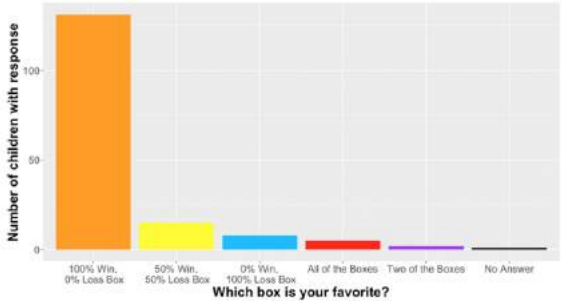


Figure 5: Children's responses to, "Which box is your favorite?"

Age is not related to the proportion of max choices. We looked for a correlation between age and number of max choices made. We found that age and number of times choosing the max box were not correlated. We found evidence in favor of the null ($r(159) = .086$, $BF_{01} = 5.17$).

Experiment 2

In Experiment 1, most children did not maximize stickers. However, at the end of the task, 83% of the participants stated that their favorite machine was the one that always won. This suggests that perhaps children knew, and preferred the winning machine, but chose to explore anyways. Additionally, we found that children were making their choices in a sequential pattern that would not be predicted by chance. While it is plausible that children have not learned the distributions yet and that is why they are exploring more, in Experiment 1, children were able to identify the machine or side that had the highest payout. Children might be approaching this task with a different motive than adults.

As adults, we have more experiences with gaming and decision making. We have learned that often times, the best thing to do is maximize reward. Maybe children, who have

less experience with decision-making tasks of this nature, could just be trying to pick up as much information about the environment as possible. This would mean exploring even when you are fairly confident about the stability of an environment. Perhaps this difference in goals is what can account for differences in behavior.

To test this hypothesis, we had adults complete the protocol outlined in Experiment 1 with one of three sets of instructions: a control scenario, a reward scenario, and a learning scenario. In the control scenario, adults were told exactly what the children were told. In the reward scenario, adults were told that they would be evaluated on how many stickers they won. In the learning scenario, adults were told they would be evaluated based on how well they learned the different distributions of the three machines.

Methods

Participants. We recruited and tested 72 adults from the University of California, Irvine SONA system. Our participants were primarily female (55 females, 17 males) and between the ages of 18-21 (57). 15 other participants were between the ages of 22-30. Participants were compensated with course credit.

Procedure. We followed the procedure outlined in Experiment 1. Adults were randomly assigned to one of three scenarios: Control, Reward, or Learning. In the Control scenario, adults were told exactly what the children were told in Experiment 1. In the Reward scenario, adults were told that they would be evaluated on how many stickers they won. In the Learning scenario, adults were told they would be evaluated based on how well they learned the different distributions of the three machines.

Results

Comparing adult data to children's data Adults in the Control & Reward conditions picked the winning box significantly more than the children and the adults in the Learn condition (See Figure 2, Figure 3, & Table 1).

Using an ANOVA, we found that there was a significant relationship between the instruction condition and the number of times the winning box was chosen ($F(3,229) = 15.41$, $p < .001$, $\eta = .168$). When comparing the children data, adult data, random strategy, and optimal policy, using an ANOVA there is an even larger effect and a significant relationship between condition and number of times the winning box was chosen ($F(5, 2227) = 1570$ $p < .001$, $\eta = .779$). We found that children picked the winning box significantly less than adults in the Control condition & Reward condition, but not the Learning condition.

Table 1: Pairwise Comparison using t-tests with pooled standard deviation. t-value (p-value)

	Control	Children	Learn
Children	-4.274 (1.0e-04)	---	---
Learn	-1.652 (0.197)	2.09 (0.109)	---
Reward	10.519 (0.449)	5.273 (8.7e-07)	2.409 (0.064)

Children switch between boxes more than adults do. A way of quantifying exploration is through looking at the proportion of switch trials participants made. If a participant is exploring, they would have a high proportion of switch trials. A participant who never makes the same choice twice in a row would have a switch proportion of 1. A participant who always chooses the same box would have a switch proportion of 0. On average, child participants had a switch proportion of .8051, whereas adults had an average switch proportion of 0.5962. Conditions that adults were in did not influence their switching behavior. A Wilcoxon rank sum test with continuity correction showed that children switched significantly more than adults did ($W = 2400$, $p = 1.527e-12$).

Trial level analysis. When looking at the data at the trial level (See Fig 2G-L), children and adults do choose the winning box more frequently near the end of the task. We did a logistic regression looking at trials as a predictor of choosing the winning box. For children, we found that trial number was an indicator of choosing the winning box ($\beta = 0.094$, $p < 2e-16$). However, for adults, the coefficient was higher ($\beta = 0.181$, $p < 2e-16$). This shows that while children are more likely to chose the maximal option as the task goes on, the change at a much slower rate than adults.

Discussion

In this paper, we presented two bandit experiments had no memory constraints. The first was with children, who did not play in a way that maximized payout and explored more than would be optimal. Uniquely, our paper showed that children were able to identify the machines that had the highest level of payout and overwhelmingly preferred the bandit with the highest payout. We also show that children's exploration is not random. For example, children moved across the bandits from left to right over and over again, as if they were reading a book.

In Experiment 2, we instructed adults to either maximize payout or learn the distributions of the 3-armed-bandit task. Experiment 2 showed that adults maximized payout, but when they are asked to maximize learning, they explore more -- like children.

Together, these results suggest that children are more interested in exploring than exploiting, and a potential explanation for this is that children are trying to learn as much about the environment as they can. There are several possible explanations for our findings.

Explanation 1: Children don't maximize payout because they don't know which machine pays out the best. One potential explanation for our data is that children spend longer than adults exploring the environment of the game because it takes children longer to figure out which machine has the best payout. This is plausible, given that children have much poorer working memory than adults do (Gathercole, Pickering, Ambridge, & Wearing, 2004), as well as less experience with this type of task.

But the children in our study did know which machine gave the most stickers -- at least, they knew this by the time they finished playing. And children overwhelmingly preferred the machine with the best payout.

Moreover, despite their poorer working memory, there are some domains of learning where children come to the correct answer faster than adults do. In language learning, for example, 6-year-old children outperform adults by maximizing probability whereas adults tend to match probability (Hudson Kam & Newport, 2005). Children also outperform adults in a simple probability guessing game (Derks & Paclisanu, 1967).

Explanation 2: Children don't maximize payout because they would rather explore the game environment. In any explore/exploit task, participants must explore to find the resources before they can shift to exploiting those resources. But it is possible that the shift from exploring to exploiting is not only seen over the course of an individual task, but over many timescales. Just as all participants explore in the early stages of a task, perhaps people explore more (in general, across domains) in the early years of life -- that is, in childhood. the time scale of one human life. Perhaps children are more 'exploring' than adults in general, meaning that they seek information about the environment in a broad sense rather than in just the narrow sense needed to maximize immediate payout (in this case, stickers).

According to this explanation, children sacrifice payout in order to get more information. But presumably, if children could get that information, either way, they would still want to maximize payout. And indeed, in bandit tasks where children are given all of the information that they would have gained from each of the different choices, they do maximize payout (Plate, Fulvio, Shutts, Green, & Pollak, 2018; Starling, Reeder, & Aslin, 2018).

We hypothesize that the optimal time to shift from exploring to exploiting depends on (A) how well you know the environment, and (B) how likely it is that the environment will change. When you don't know the environment well and/or the environment is likely to change, then more exploring is beneficial because it provides more information about all aspects of the environment (addressing Problem A) and it provides information that may be helpful if something in the environment changes (Problem B.) We hypothesize that children typically are in a situation where (A) is low and (B) is high, so they naturally explore, whereas adults know the environment better and also have fewer years left ahead of them, meaning that the amount of change they must prepare for is lower.

Our results are consistent with the idea that children develop flexible knowledge through exploration and broader search (Gopnik et al., 2017). From a child's point of view, the world is constantly changing. It makes sense to prioritize gathering data rather than maximizing immediate payouts. As our Experiment 2 showed, adults do the same when they are told to focus on learning, rather than on immediate rewards.

Acknowledgments

We thank the University of California, Irvine Graduate Division and Undergraduate Research Opportunity Program for supporting this project. We thank the Discovery Cube of Orange County & the Montessori Schools of Irvine for allowing us to collect data; Chelsea Parlett Pelleritti & Mac Strelioff for useful discussion; Fatima Pineda, Tina Singh, Mikaya Hand, Kelly Fogarty, Kelsy Chou, Julissa Navas, Hasmik Mehrabyan, Amanda Jamison, and Eden Harder for assistance with data collection.

References

Addicott, M. A., Pearson, J. M., Sweitzer, M. M., Barack, D. L., & Platt, M. L. (2017). A primer on foraging and the explore/exploit trade-off for psychiatry research. Neuropsychopharmacology, 42(10).

Beachly, W. M., Stephens, D. W., & Toyer, K. B. (1995). On the economics of sit-and-wait foraging: site selection and assessment. Behavioral Ecology, 6(3), 258–268.

Chen, W., Koide, R. T., Adams, T. S., DeForest, J. L., Cheng, L., & Eissenstat, D. M. (2016). Root morphology and mycorrhizal symbioses together shape nutrient foraging strategies of temperate trees. Proceedings of the National Academy of Sciences, 113(31), 8741–8746.

Daw, N. D., O'doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. Nature, 441(7095), 876.

Derks, P. L., & Paclisanu, M. I. (1967). Simple strategies in binary prediction by children and adults. Journal of Experimental Psychology, 73(2), 278–285. http://dx.doi.org/10.1037/h0024137

Gathercole, S. E., Pickering, S. J., Ambridge, B., & Wearing, H. (2004). The structure of working memory from 4 to 15 years of age. Developmental Psychology, 40(2), 177.

Gopnik, A., O'Grady, S., Lucas, C. G., Griffiths, T. L., Wente, A., Bridgers, S., … Dahl, R. E. (2017). Changes in cognitive flexibility and hypothesis search across human life history from childhood to adolescence to adulthood. Proceedings of the National Academy of Sciences, 114(30), 7892–7899. https://doi.org/10.1073/pnas.1700811114

Hudson Kam, C. L., & Newport, E. L. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. Language Learning and Development, 1(2), 151–195.

Mehlhorn, K., Newell, B. R., Todd, P. M., Lee, M. D., Morgan, K., Braithwaite, V. A., … Gonzalez, C. (2015). Unpacking the exploration–exploitation tradeoff: A synthesis of human and animal literatures. Decision, 2(3), 191.

Plate, R. C., Fulvio, J. M., Shutts, K., Green, C. S., & Pollak, S. D. (2018). Probability Learning: Changes in Behavior Across Time and Development. Child Development, 89(1), 205–218. https://doi.org/10.1111/cdev.12718

Snell-Rood, E. C., Davidowitz, G., & Papaj, D. R. (2011). Reproductive tradeoffs of learning in a butterfly. Behavioral Ecology, 22(2), 291–302.

Starling, S. J., Reeder, P. A., & Aslin, R. N. (2018). Probability learning in an uncertain world: How children adjust to changing contingencies. Cognitive Development, 48, 105–116.

Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A., & Cohen, J. D. (2014). Humans use directed and random exploration to solve the explore–exploit dilemma. Journal of Experimental Psychology: General, 143(6), 2074.

Xu, F., & Kushnir, T. (2013). Infants are rational constructivist learners. Current Directions in Psychological Science, 22(1), 28–32.

Slang Generation as Categorization

Zhewei Sun (zheweisun@cs.toronto.edu)
Department of Computer Science
University of Toronto

Richard Zemel (zemel@cs.toronto.edu)
Department of Computer Science
University of Toronto
Vector Institute

Yang Xu (yangxu@cs.toronto.edu)
Department of Computer Science
Cognitive Science Program
University of Toronto

Abstract

Slang is a common device for expressivity in natural language. While slang has been studied extensively as a social phenomenon, its cognitive bases are not well understood. We formulate the processes of slang generation as a categorization problem. We explore a set of cognitive models of categorization that recommend slang words based on intended referents of the speaker beyond the existing senses of words. We test these models against a large repertoire of slang sense definitions from the Online Slang Dictionary and show that the categorization models predict slang word choices substantially better than chance, without explicit consideration of external social factors. We also show that words similar in existing senses tend to extend to similar novel slang senses, reflecting a process of parallel semantic change. Our work helps to ground theories of slang in cognitive models of categorization and provides the potential for machine processing of informal natural language.

Keywords: informal language; slang; generative model; categorization; language and cognition

Introduction

Slangs—a representative form of informal language—are ubiquitous in natural language, making up approximately 52% of words in all English books written in the past two centuries (Michel et al., 2011). Slang is a common device for enhancing expressivity in human language, allowing us to express a multitude of ideas beyond the standard lexicon. Slang also adds stylistic richness to language, often allowing the identification of social groups (Millhauser, 1952). Although slangs are prevalent and accountable for language expressivity, the cognitive processes that give rise to slangs are not well understood.

Previous work has characterized slang as a social phenomenon. For instance, Labov (1972, 2006) studied how informal language emerges as a result of differing ethnicity and social-economic status. More recent work has also suggested how slang might be influenced by multiple social factors including ethnicity (Blodgett, Green, & O'Connor, 2016), gender (Bamman, Eisenstein, & Schnoebelen, 2014), and geography (Eisenstein, O'Connor, Smith, & Xing, 2010). Although it is undeniable that slang is a social phenomenon, recent work on social media analysis has suggested that slangs are more likely to catch on if they are also linguistically appropriate (Stewart & Eisenstein, 2018). We extend these work by exploring the bases of slang from a cognitive perspective, complementary to the social factors that could influence slang formation.

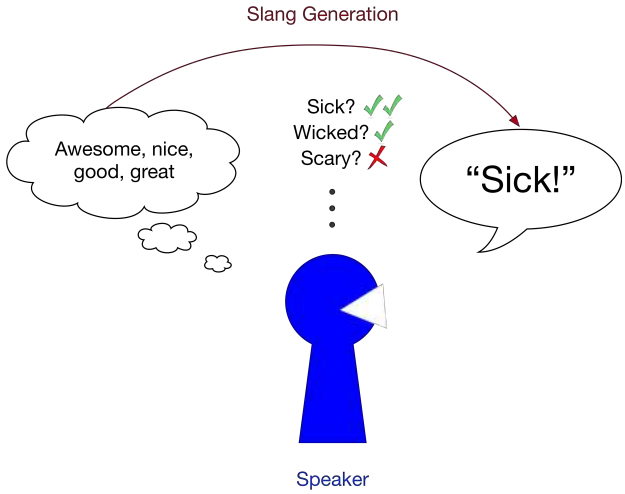


Figure 1: Illustration of the slang generation problem.

Recent work in cognitive science has explored related topics in the context of non-literal language, particularly the comprehension of metaphors (Kao, Wu, Bergen, & Goodman, 2014; Kao, Bergen, & Goodman, 2014). While slangs can often emerge from metaphorical relations, there exist many cases suggesting otherwise. For example, the slang word *sick* has the existing sense "ill" while its slang sense refers to "awesomeness". In this case, the link between the slang and existing senses are not metaphorical, but instead accounts to a polarity shift in sentiment from the existing sense.

Here we consider the general problem of slang generation by asking what cognitive processes can give rise to slang word choices for novel senses. Specifically, given a new intended slang referent one wishes to convey, how does the speaker choose an appropriate word for expressing that sense? Figure 1 illustrates this problem of *slang generation*.

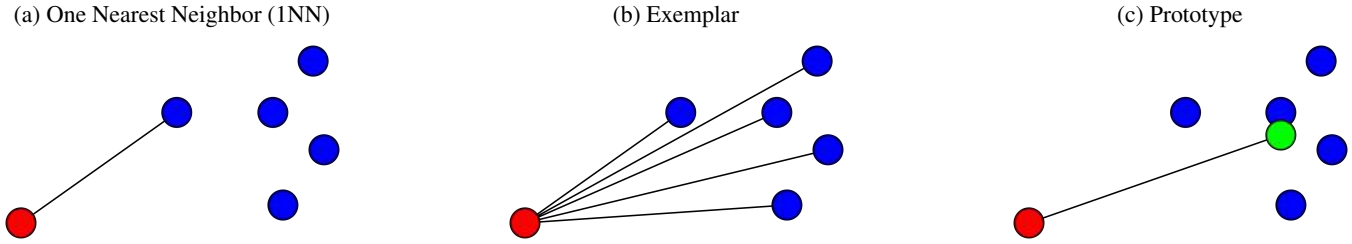


Figure 2: Illustration of categorization models for slang generation. Red (bottom-left) dot denotes novel slang sense. Blue dots denote existing senses of a candidate word. Green dot denotes prototype (or mean) of the existing senses.

Given a slang sense such as "awesome/nice", we wish to predict the word choice made by the speaker among possible alternative candidate words. In the illustrated case, the target word *sick* might be chosen if its existing senses relate to the novel slang sense, and words similar to the target word *sick* such as *wicked* might also have a good chance of being chosen. We formalize these intuitive notions of slang generation in terms of lexical choice via categorization, where we consider each candidate word as a category of existing word sense definitions. For this study, we focus on the problem of slang generation from words that are part of the existing lexicon, so we do not consider out-of-vocabulary or novel word forms for slang (e.g., Kulkarni & Wang, 2018).

We explore slang generation based on two key ideas from recent work on lexical semantic change, particularly historical word sense extension: 1) Words that bear closely related senses to a novel sense are likely to be extended to express that novel sense, a process known as semantic chaining (Lakoff, 1987; Malt, Sloman, Gennari, Shi, & Wang, 1999; Ramiro, Srinivasan, Malt, & Xu, 2018); 2) Words that begin with similar senses tend to extend to similar novel senses, a process also known as the law of parallel semantic change (Lehrer, 1985; Xu & Kemp, 2015). We formalize these ideas along with classic proposals of categorization in a simple computational framework and test them against a large online dictionary of slang.

To preview our findings, we show that cognitive models of categorization predict slang word choices substantially better than chance, and these models can be enriched by a mechanism of collaborative filtering that accounts for parallel semantic change.

Computational formulation

Models of categorization

We formulate slang generation as a categorization problem. Given a set of candidate words as categories $\{w_1, w_2, \cdots, w_N\}$ with sets of existing senses as exemplars $\{\mathcal{E}_1, \mathcal{E}_2, \cdots, \mathcal{E}_N\}$ associated with those words, we wish to find the word w_s that is most appropriate for expressing a novel slang sense s, where we represent word senses by embedding their dictionary definitions into a high-dimensional vector space (see details in the next section). For a given slang sense s, a categorization model specifies a distribution over the space of candidate words based on similarities between s and existing senses of the candidate word w_j in $\mathcal{E}_j$.

We recommend a slang word choice based on the probability distribution $p(w_j|s)$ via Bayes' rule:

$$p(w_j|s) \propto p(s|w_j)p(w_j) \tag{1}$$

Here $p(s|w_j)$ is the likelihood of the novel slang sense s given the word w_j or equivalently the collective set of its existing senses E_j, and $p(w_j)$ is the prior on the candidate word. Because we constrained our analyses to words with slang senses, we used a uniform prior on the set of candidate words. We thus estimate $p(w_j|s)$ using the maximum likelihood formulation:

$$p(w_j|s) \propto p(s|w_j) = p(s|E_j) \tag{2}$$

We specify the likelihood by considering similarity relations between existing senses of the word w_j in E_j and the slang sense s. Given a set of existing senses $E_j = \{e_1, e_2, \cdots e_M\}$, we compute its similarity with the slang sense s by considering how individual exemplars in E_j are similar to s:

$$p(s|E_j) = f(s, E_j) = f(\{sim(s, e_i); e_i \in E_j\}) \tag{3}$$

We consider the specific forms of the similarity function based on three existing models of categorization: One Nearest Neighbor (1NN), Exemplar, and Prototype. We illustrate these models in Figure 2.

One Nearest Neighbor (1NN) model. Motivated by work on semantic chaining (Ramiro et al., 2018), this model predicts that a novel word sense is attached to an existing sense of a word that is closest in semantic space. We test this hypothesis in slang generation by postulating that a novel slang sense would be attached to the most similar existing sense of a word:

$$f(s, E_j) = \max_{e_i \in E_j} sim(s, e_i) \tag{4}$$

Exemplar model. Motivated by the exemplar theory (Nosofsky, 1986), this model evaluates similarities between the novel sense s and all existing senses of a word. Here we postulate that slang choice depends on the aggregated similarities of existing senses of a word to the slang sense:

$$f(s,E_j) = \sum_{e_i \in E_j} sim(s,e_i) \quad (5)$$

Prototype model. Motivated by the prototype theory (Rosch, 1975), this model predicts that category membership is established by similarity between the slang sense and a representative or prototypical existing sense:

$$f(s,E_j) = sim(s,E_j^{prototype}) \quad (6)$$

Because we do not have an accurate estimate of sense frequencies, we consider the simple version of this model where the prototypical sense is taken as the average of the existing senses, i.e., by assuming senses are equally frequent:

$$E_j^{prototype} = \frac{1}{M} \sum_{e_i \in E_j} e_i \quad (7)$$

Where M is the set size of E_j.

Similarity. To estimate individual similarities between s and e_i, we consider vector-based embeddings that transform word sense definitions into a high-dimensional vector space. We then compute the similarity as follows:

$$sim(s,e_i) = \exp(-\frac{d(s,e_i)^2}{h_s}) \quad (8)$$

Here $d(s,e_i)$ is the Euclidean distance between the vector representations of senses and h_s is a parameter controlling the degree of sense specificity that we fit to data.

Collaborative filtering

We consider an enriched version of the categorization models by taking into account parallel semantic change, cast as a variant form of collaborative filtering (Goldberg, Nichols, Oki, & Terry, 1992) that is commonly used in recommendation systems. The rationale is that words similar in existing senses may extend to label similar novel slang senses. For example, *massive* and *stellar* both refer to *large* in their existing senses, but both of them can refer to *impressiveness* in the slang context. We capture parallel semantic change by considering the influence of neighboring words to candidate words w_j's by nested likelihoods:

$$p(w_j|s) \propto \sum_{w' \in \mathcal{L}(w_j)} p(w_j,w'|s) = \sum_{w' \in \mathcal{L}(w_j)} p(w_j|w')p(w'|s) \quad (9)$$

Here $\mathcal{L}(w_j)$ indicates a small neighborhood around the word w_j in word embedding space. We estimate $p(w_j|w')$ by computing similarity between w_j and its neighboring words:

$$p(w_j|w') \propto sim(w_j,w') = \exp(-\frac{d(w_j,w')^2}{h_w}) \quad (10)$$

For the word itself, $sim(w_j,w_j) = 1$. h_w is a free parameter that controls the strength of influence from the neighbors. This nested model estimates $p(w'|s)$ using the same likelihood functions described in the previous section. The resulting collaborative filtering model effectively provides a weighted average of the likelihoods corresponding to words in the neighborhood $\mathcal{L}(w_j)$.

Materials and methods

We collected lexical data from the freely available Online Slang Dictionary (OSD; `http://onlineslangdictionary.com`) and WordNet (Miller, 1998) for novel slang and existing word sense definitions respectively. In OSD, we considered all available slang word forms with at least one available example usage. We removed words that do not exist in WordNet and extracted all word-definition pairs from the remaining words, resulting in 4,805 slang definitions from 2,357 distinct slang words. We also extracted existing definitions from WordNet by first querying the slang word and then extracting definition sentences from all retrieved *synsets*, resulting in 11,780 existing definitions. On Average, each candidate word in our dataset has 2.00 slang definitions (SD: 1.74) and 5.54 existing definitions (SD: 6.82).

We excluded acronyms because they do not extend to new senses. We removed all slang definitions containing the word 'acronym' and words that have fully capitalized spellings. Finally, we excluded slang definitions that are already part of WordNet by performing two pre-processing steps: 1) Remove a slang definition if one of the corresponding existing definitions in WordNet has at least 50% overlap in the set of content words. 2) Remove WordNet definitions that contain the token 'slang' and remove slang words that no longer have corresponding WordNet definitions. We performed a manual sanity check on 100 randomly sampled slang definitions and only 6 of them have close definitions in WordNet. After preprocessing, there are $N = 4{,}256$ slang definitions from $V = 2{,}128$ slang words. We used these words as the vocabulary for candidate slang words. We partitioned the data of sense definitions by randomly splitting into a 90% training set and a 10% test set for model evaluation.

To represent the sense definitions in a vector space, we used distributed word embeddings from fastText (Bojanowski, Grave, Joulin, & Mikolov, 2017) pretrained with subword information on 600 billion tokens from Common Crawl (`http://commoncrawl.org`). To obtain a fixed dimensional representation for the definition sentence, we take the average word embedding of all content words within the definition sentence (Landauer, Laham, & Rehder, 1997). The average pooling scheme has been shown to be a competitive sentence encoder in machine learning literature (Wieting & Kiela, 2019) and has consistently achieved better results in our experiments compared to pre-trained deep sentence encoders. We apply the same encoding method to both existing and slang definitions with no distinction. We estimated the free model parameters (h_s, h_w) using L-BFGS-B (Byrd, Lu, Nocedal, & Zhu, 1995), a quasi-newton method for bound constrained optimization, to minimize negative

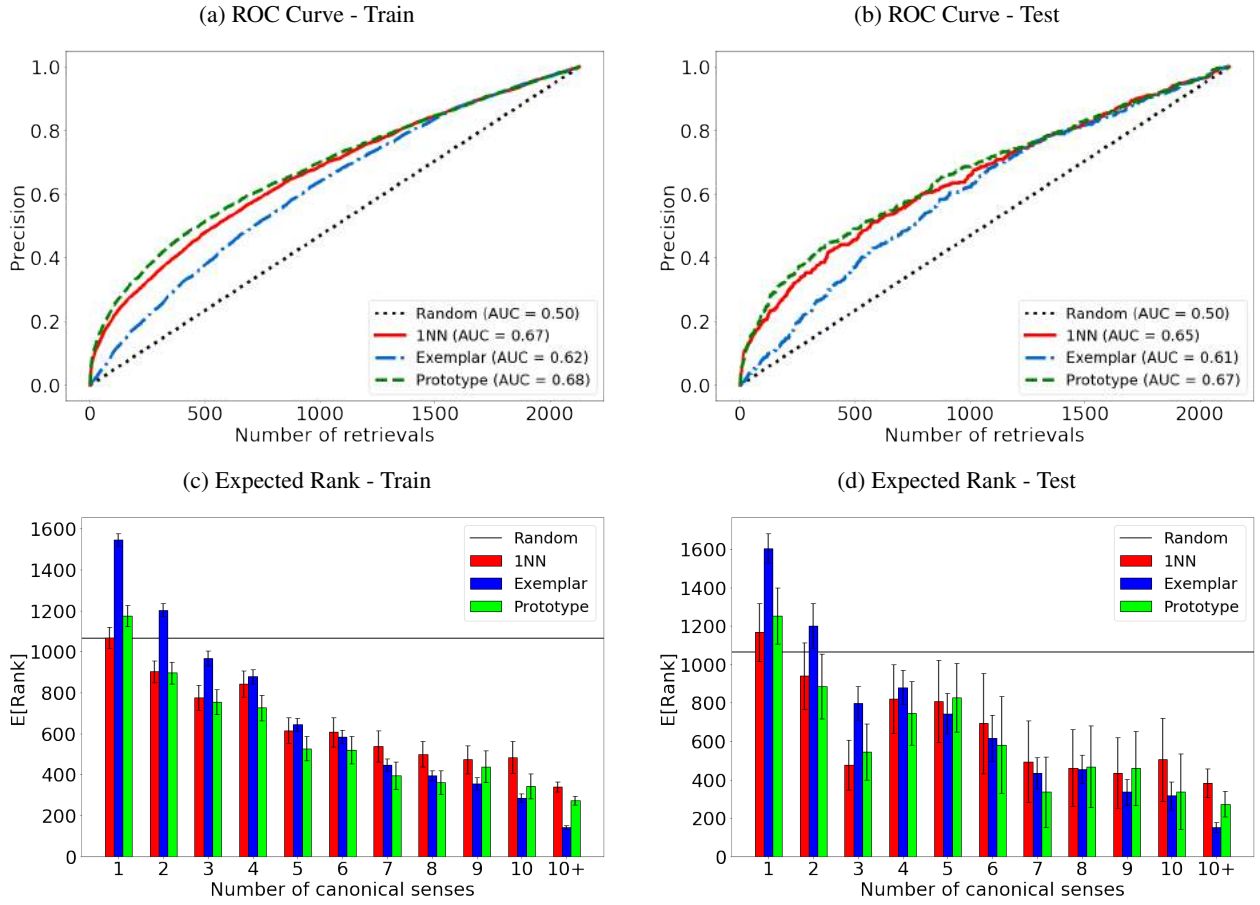


Figure 3: Top row: ROC-type curve for rank retrieval. Bottom Row: Expected Rank with respect to the number of existing senses. Ranks are computed amongst all candidate words. Whiskers denote 95% confidence intervals.

log-likelihood of the posterior:

$$\min(-\log \mathcal{L}) = \min(-\sum_{s} \log p(w_s|s)) \tag{11}$$

Here w_s is the ground truth word corresponding to the slang sense s. We estimate the free parameters on the training set while keeping them fixed in testing. For all analyzed models, we set the initial h values to 1 with bounds $[10^{-2}, 10^{2}]$. For the collaborative filtering models, both free parameters were jointly optimized.

Results

We evaluate our approach by first examining prediction of slang word choices from the three categorization models: 1NN, Exemplar, and Prototype. We then examine how collaborative filtering influences these basic categorization models on the same predictive task.

Evaluation of models of categorization

We assessed our models by ranking all candidate words according to the posterior distribution $p(w_j|s)$ from the categorization models that we described. For each slang sense definition s in the dataset, we assigned a rank to all candidate words in the vocabulary for a given model.

We first present receiver-operater curves (ROC) of model accuracy: How probable is each model to predict the correct target slang word in the first n guesses? We computed the standard Area-Under-Curve (AUC) statistics to compare cumulative precision of the models. The top row of Figure 3 shows both the ROC curves and AUC statistics of the three categorization models. All three models perform substantially better than chance. In particular, 1NN and Prototype perform better than exemplar on average in both training and testing data, which suggests that slangs are unlikely to be generated based on aggregate similarities between the existing senses and the slang sense.

Differing from previous findings on historical word sense extension where the 1NN model outperforms Prototype (Ramiro et al., 2018), we observed no substantial difference between the two models in predicting slang choices. We also considered a k-nearest-neighbor extension of the 1NN model, but we did not find any improvement in performance. We observed little difference between training and testing performances from all models, which suggests that the models did not overfit to free parameters.

For the same set of models, we also computed the expected rank of the ground-truth target words over all slang defini-

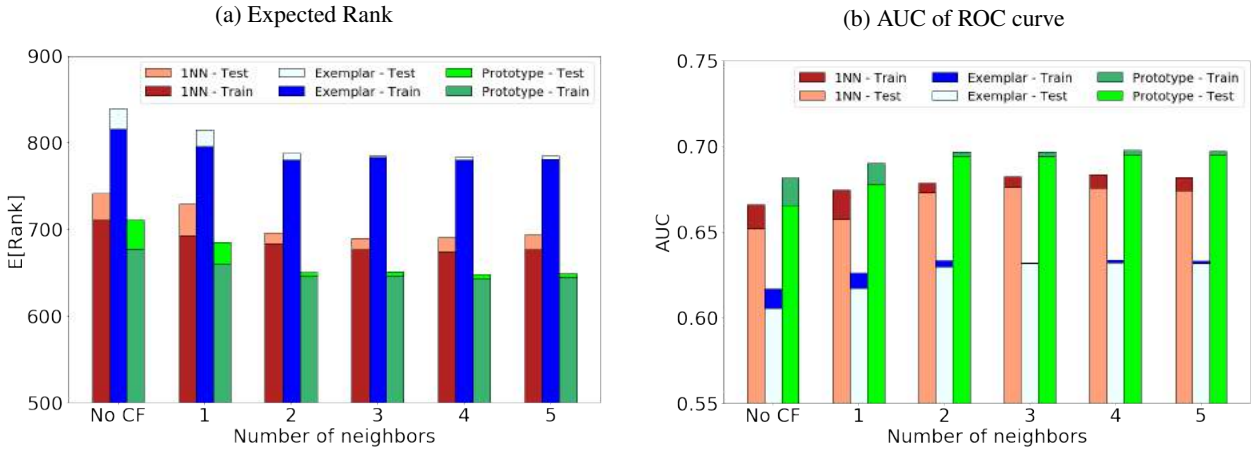


Figure 4: Summary statistics of collaborative filtering models. a): Expected Rank, b): Area-Under-Curve of ROC curves (AUC)

Table 1: Expected ranks from the categorization models.

Model	E[Rank] - Train	E[Rank] - Test
Random	1064.0	1064.0
1NN	710.89	741.29
Exemplar	815.54	839.71
Prototype	677.44	711.30

tions, based on both training and testing data. A lower expected rank indicates better predictive power. Table 1 summarizes the results. We observed similar findings with results based on AUC: All three models perform better than chance, while 1NN and Prototype models both perform better than the Exemplar model. Although these models perform above chance, the predicted expected ranks are quite high. Although some predicted words differ from the ground truth word, they may still be valid candidates for slang given sufficient social popularity. How to improve and better evaluate these model predictions will be topics of future research.

The bottom row of Figure 3 visualize the expected ranks via binning the slang definitions by degree of polysemy of their respective ground-truth candidate words w_s. We observed that all three categorization models generally perform better on more polysemous words. In particular, all three models perform better than chance when the target word has at least three existing senses. This behavior is the most prominent on the Exemplar model. Although the Exemplar model performs worse than the other two models on average, it tends to perform better on highly polysemous words. However, the Exemplar model has a natural tendency to favor those words by construction because it computes a sum of similarities instead of averaging. Both 1NN and Prototype also perform better as the number of existing senses increases. With more existing senses, it is more likely for one of them to have a close match with the slang sense, thus the improvement on 1NN. The prototypical senses would also become more accurate due to a larger sample for estimation. Compared to 1NN, the Prototype model performs slightly worse when the target word has few senses, but it outperforms 1NN as the degree of polysemy increases.

In sum, these results show that slang word choices are predictable without considering external social factors and provide evidence that simple models of categorization can capture non-arbitrariness in the generative processes of slang.

We provide examples of model success and failure in Table 2. In the *wicked* example, our models captured polarity shift in slang generation, indicated by low expected ranks from all models. The second example shows how our model can have limited predictability when the slang and existing senses are cognitively distant. In both examples, the Exemplar model consistently gave low ranks to candidate words *broken*, *play*, and *cut* because they are some of the most polysemous words in our vocabulary with more than 50 existing senses each.

Evaluation of collaborative filtering

We next examined the influence of collaborative filtering on each of the three categorization models. For each model, we considered variants of these models with up to five neighboring words.

Figure 4 summarizes the results. All collaboratively filtered models achieve better AUC and expected rank on both the training set and testing set compared to their respective basic categorization models. The improvement is most prominent on the test set, lowering expected rank by more than 50 and improving AUC by over two percent for all three models. In particular, collaborative filtering improved model prediction most substantially when two closest neighboring words were considered. Consideration of more neighbors did not improve model prediction further, suggesting that information about slang word choice is sufficiently encapsulated in a small set of neighboring words.

Table 3 illustrates collaborative filtering with two examples. In both cases, the basic categorization models perform poorly because existing senses of the ground-truth words do not have strong similarity with the slang senses. The neighboring words however, contain senses that are more rele-

Ground truth target word [w]:	*wicked*
Slang sense in OSD [s]:	impressive.
Corresponding WordNet senses [E]:	(1) morally bad in principle or practice; (2) having committed unrighteous acts; (3) intensely or extremely bad or unpleasant in degree or quality; (4) naughtily or annoyingly playful; (5) highly offensive; arousing aversion or disgust.
Model expected rankings [E(Rank)]:	(1NN): 93/2128; (Exemplar): 369/2128; (Prototype): 33/2128
Top ranked words:	(1NN): *bonzer, spot, point, tall, grand*; (Exemplar): *broken, play, cut, point, heavy*; (Prototype): *bonzer, good, tall, grand, hot*
Ground truth target word [w]:	*breezy*
Slang sense in OSD [s]:	an unimportant girlfriend or girlfriend on the side.
Corresponding WordNet senses [E]:	(1) fresh and animated; (2) abounding in or exposed to the wind or breezes.
Model expected rankings [E(Rank)]:	(1NN): 1977/2128; (Exemplar): 1829/2128; (Prototype): 1762/2128
Top ranked words:	(1NN): *man, buddy, pal, beard, associate*; (Exemplar): *broken, play, cut, run, line*; (Prototype): *front, mate, face, joker, associate*

Table 2: Examples of model success and failure.

Ground truth target word [w]:	*icky*
Slang sense in OSD [s]:	gross, unappealing.
Corresponding WordNet senses [E]:	(1) very bad; (2) soft and sticky.
5 neighboring words used in collaborative filtering [$\mathcal{L}(w)$]:	*yucky, nasty, stinky, freaky, dirty*
Ground truth target word [w]:	*scary*
Slang sense in OSD [s]:	ugly, weird.
Corresponding WordNet senses [E]:	provoking fear terror.
5 neighboring words used in collaborative filtering [$\mathcal{L}(w)$]:	*freaky, crazy, nightmare, awesome, stupid*

Table 3: Examples that illustrate how collaborative filtering helps predicting slang word choice.

vant to the probe slang sense, hence informing the model better about the ground-truth words. We also observed that the neighboring words used in collaborative filtering have strong semantic correlations, which explains the diminishing effect in performance when introducing additional neighboring words.

Conclusion

We have presented slang generation as a categorization problem. Our formulation relies on few free parameters and sheds light on the cognitive processes that give rise to slang word choice. Although the full slang generation processes are beyond the models we have explored, our framework was able to capture substantial predictability without explicitly modeling external social variables. Furthermore, we incorporated parallel semantic change in slang generation using collaborative filtering and found that it improves slang prediction beyond the basic categorization models. Future work should explore richer semantic representations of slang and extend the current framework to novel slang word forms.

Acknowledgments

We thank members of the Language, Cognition, and Computation (LCC) Group at the University of Toronto for their thoughtful feedback, particularly Suzanne Stevenson, Barend Beekhuizen, and Renato Ferreira Pinto Junior. This research is supported by an NSERC DG grant and a Connaught New Researcher Award to YX.

References

Bamman, D., Eisenstein, J., & Schnoebelen, T. (2014). Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, *18*, 135–160.

Blodgett, S. L., Green, L., & O'Connor, B. (2016). Demographic dialectal variation in social media: A case study of african-american english. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 1119–1130). Association for Computational Linguistics.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146.

Byrd, R., Lu, P., Nocedal, J., & Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, *16*, 1190-1208.

Eisenstein, J., O'Connor, B., Smith, N. A., & Xing, E. P. (2010). A latent variable model for geographic lexical variation. In *Proceedings of the 2010 conference on empirical methods in natural language processing* (pp. 1277–1287). Association for Computational Linguistics.

Goldberg, D., Nichols, D., Oki, B. M., & Terry, D. (1992). Using collaborative filtering to weave an information tapestry. *Commun. ACM*, *35*, 61–70.

Kao, J. T., Bergen, L., & Goodman, N. D. (2014). Formalizing the pragmatics of metaphor understanding. In *Proceedings of the 36th annual conference of the cognitive science society* (pp. 719–724). Cognitive Science Society.

Kao, J. T., Wu, J. Y., Bergen, L., & Goodman, N. D. (2014). Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences*, *111*, 12002–12007.

Kulkarni, V., & Wang, W. Y. (2018). Simple models for word

formation in slang. In *Proceedings of the 2018 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 1424–1434). ACL.

Labov, W. (1972). *Language in the inner city: Studies in the black english vernacular*. University of Pennsylvania Press.

Labov, W. (2006). *The social stratification of english in new york city*. Cambridge University Press.

Lakoff, G. (1987). *Women, fire, and dangerous things: what categories reveal about the mind*. University of Chicago Press.

Landauer, T., Laham, D., & Rehder, R. (1997). How well can passage meaning be derived without using word order? a comparison of latent semantic analysis and humans. In *Proceedings of the 19th annual conference of the cognitive science society* (pp. 412–417). Cognitive Science Society.

Lehrer, A. (1985). The influence of semantic fields on semantic change. *Historical Semantics: Historical Word Formation*, *29*, 283–296.

Malt, B. C., Sloman, S. A., Gennari, S., Shi, M., & Wang, Y. (1999). Knowing versus naming: Similarity and the linguistic categorization of artifacts. *Journal of Memory and Language*, *40*, 230–262.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., ... Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, *331*, 176–182.

Miller, G. (1998). *Wordnet: An electronic lexical database*. MIT press.

Millhauser, M. (1952). The case against slang. *The English Journal*, *41*, 306–309.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39–57.

Ramiro, C., Srinivasan, M., Malt, B. C., & Xu, Y. (2018). Algorithms in the historical emergence of word senses. *Proceedings of the National Academy of Sciences*, *115*, 2323–2328.

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, *104*, 192–233.

Stewart, I., & Eisenstein, J. (2018). Making "fetch" happen: The influence of social and linguistic context on nonstandard word growth and decline. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 4360–4370). Association for Computational Linguistics.

Wieting, J., & Kiela, D. (2019). No training required: Exploring random encoders for sentence classification. In *International conference on learning representations.*

Xu, Y., & Kemp, C. (2015). A computational evaluation of two laws of semantic change. In *Proceedings of the 37th annual conference of the cognitive science society* (pp. 2703–2708). Cognitive Science Society.

A generalization becomes suppressed over time in the context of exceptions

Karina Tachihara (tachihara@princeton.edu)
Department of Psychology, Princeton University, Princeton NJ 08544 USA

Kenneth A. Norman (knorman@princeton.edu)
Department of Psychology, Princeton University, Princeton NJ 08544 USA

Nicholas Turk-Browne (nicholas.turk-browne@yale.edu)
Department of Psychology, Yale University, New Haven, CT 06520 USA

Adele E. Goldberg (adele@princeton.edu)
Department of Psychology, Princeton University, Princeton NJ 08544 USA

Abstract

There has been a great deal of interest in how generalizations and exceptions are learned, with particular interest in how speakers learn to avoid overgeneralizations. Do overgeneralizations disappear only because exceptions become more strongly represented or does the generalization itself become suppressed? Novel labels were constructed by combining 56 syllables with one of two prefixes, and each label was assigned a unique image. Most labels with the first prefix were paired with images from a generalization category, whereas exceptional labels were paired with images from a different semantic category. All labels with the second prefix appeared with a third category ("baseline"). Participants used a computer mouse to choose one of two images for each label. Mouse-tracking results show that the generalization itself became suppressed over time in the context of exceptional labels. A post-test demonstrated that exceptions were learned with item-specific precision.

Keywords: language acquisition, generalization, exceptions, overgeneralization, mouse-tracking

Introduction

In order to speak a language fluently, it is critical to learn subclasses of exceptions within otherwise broad generalizations. For instance, in Spanish, words ending *–a* are generally grammatically feminine, but roughly half of the words that end in *–ma* are masculine (e.g., *el drama*). The present work investigates how these sorts of generalizations and exceptional subclasses interact with one another during the learning process. In particular, we investigate whether competition between a generalization and a subclass of exceptions persists to the same degree throughout learning.

Competition between generalizations and exceptions is widely recognized to affect language processing (Bates & MacWhinney 1987; Christiansen & Chater 1999; McClelland, & Rumelhart 1986; Goldberg 2019). However, less attention has been focused on how the process of learning exceptions might affect memory for the generalization. One possibility is that the generalization and exceptions are represented independently, and learning the exceptions has no effect on memory for the generalization. According to this perspective, the generalization and exceptions may operate in parallel and race to provide the correct form during production (Pinker 1999), or they may operate as sequential rules (Yang 2016). Both of these proposals are consistent with the idea that speakers learn to avoid overgeneralizations because exceptions become more strongly represented. No change in the representation of the generalization is required.

A third possibility we investigate here is that the generalization becomes suppressed in the context of exceptions. Support for this hypothesis comes from the literature on how competition between memories drives learning. Numerous studies have found that, when memories (semantic or episodic) compete, the "losing" memories (i.e., memories that are partially activated, but less than the memory that is fully retrieved) become harder to subsequently access, compared to memories that do not undergo competition. (Anderson et al., 1994; Anderson et al., 2000; Bäuml, 1998; Bäuml 2002; Johnson & Anderson, 2004; Levy et al., 2007; Murayama et al., 2014; Lewis-Peacock & Norman 2014; Kim et al., 2014).

For example, in Anderson et al. (1994), participants memorized a set of word pairs, some of which shared a semantic category (*fruit: orange; fruit: apple*) while other items were part of an unrelated category (*tool: hammer*). During the retrieval practice phase, participants were given a semantic cue and asked to recall a subset of the items (*fruit: ap___*). Note that the semantic category fruit can be expected to activate *orange*, but *orange* would lose in competition to *apple* because it is inconsistent with the partial cue "*ap___*".

That is, the cue ensures that *apple* wins in a competition with *orange* (and other prototypical fruits). At the final test phase, unsurprisingly, participants recalled practiced items (*apple*) best. Critically, items in the same category which were not themselves practiced (e.g., *fruit: orange*), had a lower recall rate than unrelated baseline items (*tool: hammer*), an effect known as retrieval-induced forgetting (RIF).

Anderson et al., (2000) emphasized the role of competition during retrieval in RIF. They found that simply repeating an item (e.g., *apple*) without the semantic cue (*fruit: ap____*) that could be expected to partially activate competitors such as *orange,* did not result in the subsequent suppression of *orange*. In this case where there was no competition-inducing cue, the repeated item (i.e., *apple)* was strengthened but the other word from the same category (i.e., *orange*) was not less likely to be recalled than words from other categories (like, *hammer*). These results demonstrate that it is not merely the strengthening of the more activated memory that resolves the competition. Rather, competition also leads to suppression of the less activated memory.

In the domain of language learning, we hypothesize that exceptions serve to delimit the domain of a generalization, suppressing its activation and carving out a space of their own so that the generalization and exceptions become more differentiated over the course of learning. The alternative hypothesis is that exception learning is the strengthening of the exception alone, with no change to the generalization. We aim to evaluate these hypotheses by exposing participants to a mini-artificial language that contained a generalization and a subclass of exceptions. We then used a mouse-tracking design, as it provides a sensitive way to detect competition between two alternatives in a forced choice task.

The mini-artificial language consisted of two prefixes and 56 syllables and images. One prefix appeared with 40 syllables paired with images of one semantic category (the generalization) and 8 other syllables paired with a second semantic category (the exceptions). The second prefix consistently appeared with 8 instances of a third semantic category and served as a baseline. For example, as presented in Figure 1, a subset of participants witnessed the prefix, *abber,* paired with 40 unique syllables and unique faces and 8 different syllables and unique scenes. The other prefix, *belling*, was then paired with 8 unique syllables and unique objects. The combination of semantic category (faces, scenes, objects) and prefix (*abber, belling*) was counterbalanced across participants, and additionally, the pairing of each syllable and image was randomized for each participant. However, for ease of description, we refer to the assignment of categories and prefixes represented in Figure 1 throughout the paper.

Participants were first exposed to all 56 <prefix+ syllable> pairs (hereafter, labels) and images. In the main task, participants heard each label and decided which of two images on the screen matched that label (vs. the other "lure" image) by using a computer mouse to move from the bottom of the screen to the chosen image (Spivey, Grosjean, & Knoblich, 2005; Spivey & Dale, 2006). These mouse-tracking trials were repeated over 8 blocks in order to investigate learning over time. Only correct trials are included in the main analysis. But the dependent measure used was deviation toward the distractor image (the "lure"), weighted by time, which captures the degree to which participants were lured by the *incorrect* category (Figure 2).

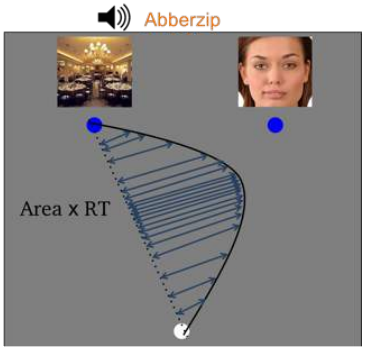


Figure 2: Example mouse-tracking trajectory sampled every 30 ms to determine the strength of the lure category (here, the face image).

Specifically, the distance between the cursor's position and a straight line to the correct response was measured at 30 millisecond intervals. To the extent that participants drew a relatively straight line from the start to the correct target, the deviation measure was low, indicating that the lure was not active in their minds. On the other hand, if participants drew an arc that trended toward the lure, we can conclude that the lure was activated by the label to some degree

Since our interest was in the relationship between a generalization and a subclass of exceptions, it might be tempting to focus on trials that included both an image from the generalization category and an image from the exception category. However, it would be impossible to determine in that case whether the trajectory was due to being lured by one image or by avoidance of the other. Specifically, an overgeneralization may be captured by a strong pull towards the generalization lure image or a lack of pull towards the correct exception image.

Therefore, in order to investigate how generalization activation changes without contamination from the lure of an exception image, a second trial type was introduced, "Scrambled-Image" trials (Figure 3). On these trials, participants were told to always select the scrambled image, regardless of what label was heard or which other image was available. For these

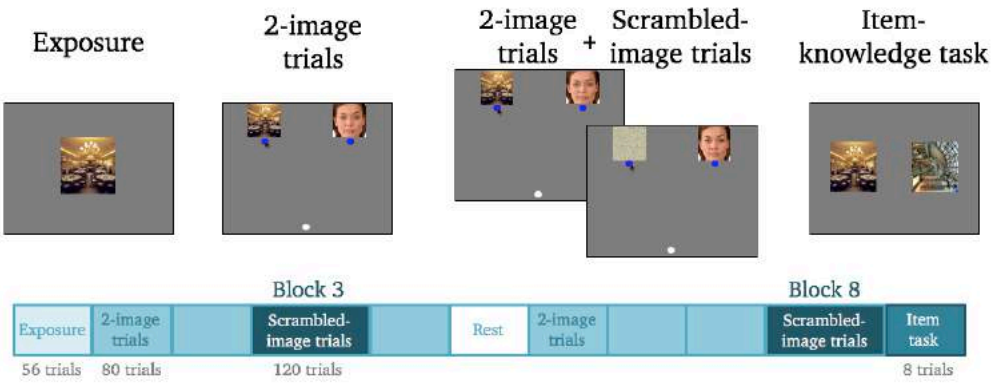


Figure 3: Exposure phase, followed by 8 blocks of 2AFC trials. Blocks 3 and 8 also contained scrambled-image trials in order to measure the strength of lures as directly as possible. Lastly, there was an item-knowledge task.

scrambled-image trials, the trajectory indicated how much participants activated the lure image category without the confound of preference toward the chosen image, since this was held constant across all scrambled-image trials. In order to avoid over-learning of the scrambled-image task (i.e., participants becoming good at going straight to the scrambled image without consideration of the lure), this type of trial was only included in blocks 3 and 8.

Blocks with two intact-image trials (blocks 1,2, & 4-7) each had 80 trials per block. Blocks 3 and 8 had two intact-image trials intermixed with scrambled-image trials for a total of 120 trials (Figure 3). Thus, in total, there were 720 trials, 11.11% of which were scrambled-image trials. Participants were not given any indication of the block structure of the task, except that they were given a rest after block 4, half way through the experiment.

By comparing performance on scrambled-image trials in blocks 3 and 8, the activation of lures over the course of learning can be detected. See Figure 3 for experimental design.

The scrambled-image trials, along with the mouse-tracking measure, allowed us to home in on the activation of a particular category for a particular label and how it changed over the course of learning. This enabled us to test the following hypothesis: a generalization becomes suppressed over time in the context of exception labels.

Method

The sample size and the main analysis were preregistered on Aspredicted.org, prior to data collection.

Participants

42 undergraduate students from Princeton University were compensated with course credit and up to an additional $5, depending on task performance.

Stimuli

The 2 prefixes (*abber* and *belling*) and 56 syllables (e.g., *zip, ber,* and *za)* were all phonotactically regular. The labels (prefix + syllable) were presented auditorily without pauses between the prefix and the syllable, and each lasted approximately 800 ms. Each scrambled image was created by scrambling the pixel locations of the lure image used in the same trial.

Procedure

Participants were given general instructions at the beginning as well as 6 practice trials for the 2AFC mouse-tracking task. They were told to pay attention to the pairing of the labels and images, but were not told about the structure of the stimuli (i.e., that the labels were a combination of a prefix and a syllable, nor the general distribution of categories). They were instructed to make their choices as quickly as possible and to move the cursor as directly to the target as possible while trying to avoid errors. The entire experiment lasted 1.5-2 hours, including a short rest period.

Initial exposure phase: Each label-image pair was presented once, for a total of 56 trials (40 generalization items + 8 exception items + 8 baseline items), with order of presentation randomized for each participant.

Mouse tracking task: Blocks 1-8: Participants were instructed to choose the image that matched the label they heard, except on scrambled-image trials in which they were instructed to always choose the scrambled image. For the intact-image trials, the two images always came from different categories, so participants could perform at ceiling by recognizing which category each label belonged to, without necessarily learning which face, scene or object each label corresponded to. For the scrambled-image trials, one of the images was created by scrambling the pixel

location of the other intact image. All other procedures were equal between intact image trials and scrambled-image trials.

The label was played through headphones. Once it was finished, participants could click the white button at the bottom of the screen, causing 2 images to be displayed. Participants then moved their cursor to the image that was associated with the label and clicked on the blue button underneath that image. In order to encourage participants to respond as quickly as possible, a score appeared on the center of the screen, calculated according to the trajectory of the mouse and speed of response. When the score was displayed, the incorrect image would disappear, leaving the correct image only. If participants had chosen incorrectly, they had to move their cursor to the correct image and click, before continuing to the next trial. After block 4, participants were given a mandatory 5-10-minute break before continuing with block 5.

Item-knowledge task: After the 8 blocks of the 2AFC task, participants performed a short task designed to test whether they had incidentally learned to associate particular exception labels with particular images within that category. The 2 images in this task were both instances of the exception category (e.g., 2 different scenes), requiring participants to identify the item-specific association of label and image. Participants were unaware they would be tested on item-specific knowledge for this task.

Results

All 42 participants exceeded the preregistered threshold of 75% accuracy on the mouse-tracking task ($M = 87.14$, $SD = 0.0085$), and none were excluded ($N = 42$). 3.25% of all trials were excluded because participants took > 2 seconds to click the start button or > 5 seconds to make a choice between images.

Accuracy on intact image trials

For trials in which participants decided which one of the two intact images matched the label they had heard, we can look at their accuracy against chance (50%) to see how well they knew the label-image pairings. Participants were above chance on all trial types in the first block after exposure ($t = 22.95$, $p < 0.0001$, $M = 0.89$), except for exception-label trials. On exception-label trials, participants heard an exception label and had to choose between an image from the exception category (e.g., scene, the correct choice) and the generalization category (e.g., face, the incorrect choice). Initially, accuracy on exception-label trials was significantly below chance (block 1), indicating that participants were systematically choosing the generalization image ($t = -2.13$, $p = 0.039$, $M = 0.43$). Accuracy for exception-label trials quickly rose, however, becoming significantly above chance in block 2 ($t = 2.43$, $p = 0.020$, $M = 0.59$). By block 8, accuracy for exception-label trials was as high as that for other trial types ($t = 1.30$, $p = 0.20$, $M = 0.93$ for exception trials, $M = 0.96$ for other trials).

Trajectory toward lure

The dependent measure for each trial was the area underneath the trajectory weighted by reaction time (area x RT). To calculate the area x RT, we compared the trajectory against the most direct, straight line connecting the starting point and the end point. The starting point was the position the participant had to click at the start of the trial and the end point was where the participant clicked when they made a choice (one of two blue circles). We measured points on the actual trajectory every 30 ms and calculated the distance between each of these points from the straight line. The sum of these distances is the area x RT. Note that the farther participants moved their cursor away from the straight line and the longer it stayed there, the higher the area x RT was. We had preregistered the dependent measure to be the maximum distance from the straight trajectory, and the result of the preregistered main analysis does not qualitatively differ when the maximum distance is used. However, after preregistering, we decided that area x RT was more appropriate and sensitive, allowing us to take both speed and deviation into account.

We report the results from the trajectory of the scrambled-image trials because it is the most direct measure of the activation of a category (i.e., the lure image category) given a label. For all analyses we used a maximal multilevel model with trial type or an interaction of trial type and block as the fixed effects and random intercepts and slopes for subjects and items where convergence would allow (Barr, Levy, Scheepers, & Tily 2013), using the lmerTest library (R Development Core Team 2008).

First, to confirm that highly activated lure images would indeed yield greater deviation and thus higher area x RT measures, we compared trials in which the label matched the lure image (e.g., the label was paired during the training phase with a specific scene and the lure image is that specific scene) and trials in which the label did not match the lure image or its category (e.g., the label was paired with a scene and the lure image is a face). As expected, we found that matched trials had a higher area x RT than unmatched trials ($\beta = -0.69$, $t = -11.06$, $p < 0.0001$).

Recall our hypothesis: that generalization activation (e.g., face image activation) would decrease over time for exception items (e.g., for labels that are paired with scenes). Thus, the critical preregistered comparison

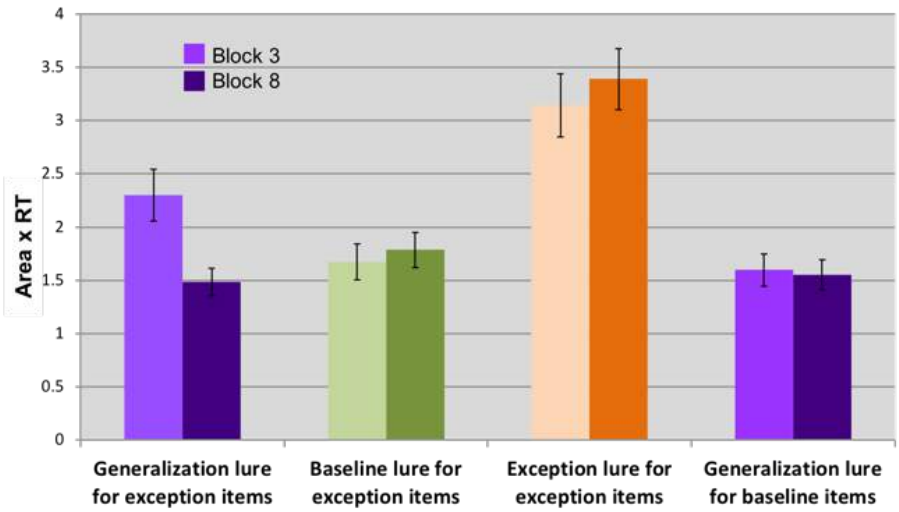


Figure 4: Deviation measure toward lure in block 3 (lighter color) and block 8 (darker shade), for exception-label trials with generalization image lures (left purple), exception-label trials with baseline lures (green), exception-label trials with exception lures (orange), and baseline-label trials with generalization lures (right purple). The correct choice in all cases was a scrambled image.

was the change of activation of the lure from block 3 to block 8 on trials when an exception-label was heard. We compared trials in which the lure image was the generalization (e.g., the label was paired with scene and the lure was a face image) against the baseline (e.g., the label was paired with a scene and the lure was an object image). The model found a significant interaction of trial type and block (β = 0.23, t = 2.66, p = 0.010). In other words, for exception items, the generalization activation became suppressed over time, more so than did baseline activation. In Figure 4, the far-most left panel shows the key generalization suppression from block 3 (light purple) to block 8 (dark purple). There is no suppression over blocks for baseline activation (green). Thus, generalization suppression cannot be attributed to general improvement over time or to general improvement on scrambled image trials.

Another critical part of the hypothesis is that the generalization was suppressed due to competition from learned exceptions. For exceptions to compete with the generalization, exceptions must be activated to some degree. In other words, exception-labels must be identified as exceptions and activate the correct exceptional category (scene) for competition to occur. Results additionally provide evidence that, as early as block 3, participants had learned which labels were exceptional. In particular, when an exception-label was heard, the matched exception image (scene) exerted a strong pull away from the scrambled image (third panel, light orange bar). In fact, the area x RT for matched exceptional images was higher than that for generalization images (face images) at block 3 (β = 0.42, t = 2.10, p =0.038). This means that the generalization suppression we observe occurred after participants had already learned the exceptions to some degree.

An alternative explanation for generalization suppression over time for exception-label trials could be that the generalization (e.g., face images) became less of a lure across the board, for exception items as well as baseline items. To investigate this possibility, we compared area x RT towards generalization lure images for exception-label trials against baseline-label trials. If (as hypothesized) generalization suppression is unique to exception labels (because of the competition from sharing a prefix), there should be no generalization suppression for baseline labels (where a prefix was never shared, and thus no competition took place). We again found a significant interaction of trial type and block (β = 0.19, t = 2.37, p = 0.018). In other words, generalization suppression over time was evident only in the context of exception items, not baseline items. In Figure 4, the far-most right panel shows no change of generalization activation over time for baseline items (purple bars). This also rules out the possibility that generalization suppression for exception items was specific to an image category (e.g., generally disliking faces over time).

Item-knowledge task

Despite high accuracy in the main task being achievable based purely on recognition that certain labels were exceptional (i.e., were associated with the non-dominant category for the prefix), the final task demonstrated that participants nevertheless learned with near-ceiling level accuracy which specific scene was paired with which specific label (M = .9494; t = 29.19, p < 0.0001).

Discussion and Conclusion

This experiment assessed how the activation of a generalization changed in the context of exceptions

over the course of learning. By using mouse-tracking to measure lure activation, we were able to isolate the activation of the generalization from the activation of the exception for a given label. Results demonstrate that the competing probabilistic generalization was a strong lure for the exceptions early on, but the generalization became suppressed over time in the context of exceptions. That is, the suppression of the generalization over time was evident only for exception labels. Because accuracy on exceptions was already high and exception lures were already even stronger lures early on, we suggest that the suppression was caused by competition from learned exceptions. Results of a post-test demonstrated that exceptions were learned with item-specific precision even though ceiling performance was possible by reliance on category membership only.

Importantly, our claim is not that learning eliminated all competition between generalizations and exceptions in this study. We know that comprehension is incremental, so we fully expect that listeners activated multiple options that were consistent with the input they witnessed until the point of disambiguation (Jurafsky, 1996; McQueen, 2007; Rayner & Clifton, 2009; Swinney, Prather, & Love, 2000); hearing *abber* should trigger a competition between exceptional items (e.g., *abber zip*) and other items that begin with the same prefix (e.g., *abber fep*), even after learning takes place. Rather, our main hypothesis pertains to what happens after the disambiguating syllable (*zip*) is heard: Would learning of exceptions affect the activation of the generalization, specifically in the case of exceptions like *abber zip*? We found that it did: the generalization became a less powerful lure as exceptions became more easily identified.

This work was motivated, in part, by the effects of competition on memory observed during studies of *retrieval-induced forgetting* (RIF). Consistent with RIF findings in the memory literature, the linguistic generalization became suppressed ("forgotten") in the context of exceptions. At the same time, it is important to point out a key difference between our study and the way RIF is usually tested. Most RIF studies look at final recall to measure memory performance. Our study, on the other hand, considered the change in activation over the course of learning. This difference led us to use a different baseline for determining whether suppression occurred. In standard RIF studies, suppression is measured by how much lower the memory for competing items is, compared to baseline items which had not been in competition with the practiced items. In our studies, suppression was measured by how much lower the activation for the generalization became over time. We found that generalization activation significantly decreased over time for exception items, much more than it did for baseline items. However, we did not find that activation levels of the generalization fell below baseline activation; as such, we did not find RIF in the classic sense. Nonetheless, our results are consistent with the idea that competition leads to suppression of the less activated memory.

We selected the "prefix plus syllable" structure for the labels to allow for prediction of the category given the prefix. The prefix plays the role of a classifier (i.e., a grammatical element that selects for nouns of certain semantic categories; Dixon 1986) As noted in the introduction, linguistic categories, including classifier categories, often have subclasses of exceptions, as in the present experiment. However, the finding in this study is not, in principle, specific to words. For example, similar competitive mechanisms may serve to suppress grammatical overgeneralizations through what has been called *statistical preemption* (e.g., Goldberg, 2019; Perek & Goldberg 2017; Robenalt & Goldberg, 2015). Future work will build on these results to explore how generalizations and exceptions compete in other domains, how the underlying neural representations change, and how these competition-driven changes relate to behavior.

References

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Remembering can cause forgetting - Retrieval Dynamics in long-term-memory. *Journal of Experimental Psychology-Learning Memory and Cognition*, *20*(5), 1063–1087.

Anderson, M. C., Bjork, E. L., & Bjork, R. A. (2000). Retrieval-induced forgetting: Evidence for a recall-specific mechanism. *Psychonomic Bulletin and Review*, *7*(3), 522–530.

Bates, E., MacWhinney, B., & MacWhinney, B. (1987). Competition, variation, and language learning. *Mechanisms of language acquisition*, 157-193.

Bäuml, K. (1998). Strong items get suppressed, weak items do not. *Psychonomic Bulletin and Review*, *5*(3), 459–463.

Bäuml, K.-H. (2002). Semantic generation can cause episodic forgetting. *Psychological Science*, *13*(4), 356–360.

Christiansen, M. H., & Chater, N. (1999). Connectionist natural language processing: The state of the art. *Cognitive science*, *23*(4), 417-437.

Dahan, D., Magnuson, J. S., Tanenhaus, M. K., & Hogan, E. M. (2001). Subcategorical mismatches and the time course of lexical access: Evidence for lexical competition. *Language and Cognitive Processes*, *16*(5-6), 507-534.

Dixon, R. M. (1986). Noun classes and noun classification in typological perspective. *Noun classes and categorization*, 105-112.

Goldberg, A.E. (2019). *Explain me this: creativity, competition and the partial productivity of constructions.* Princeton: Princeton University Press.

Grosjean, F. (1980). Spoken word recognition processes and the gating paradigm. *Perception & psychophysics*, *28*(4), 267-283.

Johnson, S. K., & Anderson, M. C. (2004). The role of inhibitory control in forgetting semantic knowledge. *Psychological Science*, *15*(7), 448–453.

Jurafsky, D. (1996). A probabilistic model of lexical and Syntactic Access and Disambiguation. *Cog. Sci. 20*(2), 137–194.

Kim, G., Lewis-Peacock, J. A., Norman, K. A., & Turk-Browne, N. B. (2014). Pruning of memories by context-based prediction error. *Proceedings of the National Academy of Sciences*, *111*(24), 8997–9002.

Levy, B. J., McVeigh, N. D., Marful, A., & Anderson, M. C. (2007). Inhibiting your native language: The role of retrieval-induced forgetting during second-language Acquisition. *Psychological Science*, *18*(1), 29–34.

Lewis-Peacock, J. A., & Norman, K. A. (2014). Competition between items in working memory leads to forgetting. *Nature Communications*, *5*(5768), 1–10.

McClelland, J. L., Rumelhart, D. E., & PDP Research Group. (1986). Parallel distributed processing. *Explorations in the Microstructure of Cognition*, *2*, 216-271.

McQueen, J. M. (2007). Eight questions about spoken-word recognition. *The Oxford Handbook of Psycholinguistics*, 37–53.

Murayama, K., Miyatsu, T., Buchli, D. R., & Storm, B. C. (2014). Forgetting as a consequence of retrieval : A meta-analytic review of retrieval-induced forgetting. *Psychological Bulletin*, *140*(August), 1383–1409.

Perek, F., & Goldberg, A. E. (2017). Linguistic generalization on the basis of function and constraints on the basis of statistical preemption. *Cognition*, *168*, 276-293.

Pinker, S. (1999). *Words and rules: The ingredients of language*. Basic Books.

Rayner, K., & Clifton, C. (2009). Language processing in reading and speech perception is fast and incremental. *Bio. Psych.*, *80*(1), 4–9.

Robenalt, C., & Goldberg, A. E. (2015). Judgment evidence for statistical preemption: It is relatively better to vanish than to disappear a rabbit, but a lifeguard can equally well backstroke or swim children to shore. *Cognitive Linguistics*, *26*(3), 467-503.

Spivey, M. J., & Dale, R. (2006). Continuous dynamics in real-time cognition. *Current Directions in Psych. Sci.*, *15*(5), 207–211.

Spivey, M. J., Grosjean, M., & Knoblich, G. (2005). Continuous attraction toward phonological competitors. *PNAS*, *102*(29), 10393–10398.

Swinney, D., Prather, P., & Love, T. (2000). The time course of Lexical Access and the role of context. In *Language and the Brain: Representation and processing* (pp. 273–292). New York: Academic Press.

Yang, C. (2016). *The price of linguistic productivity: How children learn to break the rules of language*. MIT Press.

Redundant morphological marking facilitates children's learning of a novel construction

Shira Tal (shira.tal1@mail.huji.ac.il)
Department of Cognitive Science, The Hebrew University of Jerusalem,
Mount Scopus, Jerusalem 91905, Israel

Inbal Arnon (inbal.arnon@mail.huji.ac.il)
Department of Psychology, The Hebrew University of Jerusalem,
Mount Scopus, Jerusalem 91905, Israel

Abstract

The presence of redundant marking in languages raises interesting questions about the balance of different pressures in language learning and use. Speakers tend to avoid redundant elements in production: omitting (or reducing) more predictable elements. At the same time, languages maintain different types of redundant marking, such as encoding thematic assignment by both word order and case marking. Why is redundancy found in languages even though speakers seem to avoid it? Here, we propose that redundant cues can facilitate learning. We test this hypothesis in an artificial language learning study with children, where either word order alone or both word order and case marking serve as cues for thematic assignment in a novel construction. Results show that children learned the redundant language better despite having to learn an additional morpheme. We discuss implications for the effect of different cognitive pressures on language change.

Keywords: redundancy; artificial language learning; language acquisition; language evolution

Introduction

It has long been claimed that languages exhibit an optimal trade-off between two competing pressures: minimizing effort and maximizing understandability (Givón, 1991; Haspelmath, 2008; Jäger, 2007; Jaeger & Buz, 2017; Piantadosi, Tily, & Gibson, 2012; Zipf, 1949). Under this view, redundant cues will be dispreferred in language use (because they increase effort without increasing understandability), eventually leading to their reduction in language structure (Fedzechkina, Newport, & Jaeger, 2016; Gibson, Piantadosi, et al., 2013; Givón, 1991; Jaeger, 2013). In line with this view, speakers seem to avoid redundant marking in production: more predictable messages are more likely to be omitted or reduced (Aylett & Turk, 2004; Cohen Priva, 2015; Frank & Jaeger, 2008; Jaeger, 2010; Kurumada & Jaeger, 2015; Levy & Jaeger, 2007). For example, speakers tend to omit optional case marking when the meaning it encodes is more predictable from context (Kurumada & Jaeger, 2015; Lee & Kim, 2012).

At the same time, redundant marking is attested, in different forms, in multiple language systems. For instance, a number of typologically diverse languages are documented as redundantly marking a single meaning using multiple morphological markers (defined as *multiple exponence*, Caballero & Harris, 2012; Harris, 2017). In Choguita Rarámuri, for example, words containing an inner derivational marker for causatives and applicatives can have a second, optional marker suffixed to the noun (Caballero & Kapatsinski, 2015). Languages can also redundantly mark the same grammatical information by more than one means: one such example is the encoding of thematic assignment (who is doing what to whom) by both word order and case marking (e.g., Icelandic, see Siewierska, 1998). That is, morpho-syntactic redundancy—where two linguistic cues are used to mark the same function—is found across language systems.

How can we reconcile the presence of redundancy in language structure with speakers' tendency to avoid it in production? One way is to examine the possible functions of redundancy: why does it come about in the first place and what advantage can it confer? Here, we propose that the answer may lie in the impact of redundancy on learning. In particular, we suggest that (a) redundancy can facilitate learning under certain conditions, and (b) if this is so, speakers may increase (or maintain) the use of redundant cues when conversing with learners, supporting their continued presence in language. This proposal is compatible with the principles of efficient communication: the balance between effort and understandability can change depending on the comprehension ease or difficulty within a conversation (Gibson, Bergen, & Piantadosi, 2013; Kurumada & Jaeger, 2015; Levy & Jaeger, 2007). Speakers may allow (or even prefer) more redundancy when the listener is seen as having more difficulty in comprehension, as in the case of learners. The two predictions—the facilitative effect of redundancy on learning and its increased use with learners—are related, but theoretically independent. In the present study, we focus on the first prediction: Does redundancy facilitate language learning? Our focus here is on morpho-syntactic redundancy: cases in which different morpho-syntactic cues encode the same information. In line with previous work, we treat the omission and reduction of linguistic material as reducing redundancy (Aylett & Turk, 2004; Jaeger, 2010; Kurumada & Jaeger, 2015).

The advantage of multiple cues in learning has been demonstrated for different domains, such as vision (Sloutsky & Robinson, 2013) and category formation

(Yoshida & Smith, 2011). Recent computational evidence suggests that multiple cues can also facilitate language learning (Monaghan, 2017). Monaghan (2017) examined learning new mappings between forms and meanings with multiple cues. The cues were probabilistic in the learning phase (appeared only some of the time), and were absent during testing (where only the labels appeared). In line with our prediction, the computational model showed that words were learned better from multiple cues (pointing, prosody, and distributional cues) compared to single cues. Importantly for the present research question, the multiple cues used in these studies involved the combination of linguistic and non-linguistic cues (e.g., pointing) where the non-linguistic cues did not have to be learned in and of themselves (they did not carry information beyond increasing attention to the label). Here, we go beyond this work to ask whether redundant *linguistic* cues can also facilitate language learning. If redundant morphological marking is facilitative, we should see improved learning despite the added complexity of having to learn an additional cue. We focus on the morpho-syntactic redundancy of word order and case marking and use the transitive construction as a test case.

The Transitive Construction

Languages use different cues to indicate who-did-what-to-whom in the transitive construction. Two prominent cues are word order and case marking[1]: looking at learning when both cues are used lets us examine the possible advantage of redundant marking. The contribution of different cues to sentence interpretation has been studied extensively within The Competition Model (MacWhinney, 1987). Stemming from this theoretical framework, studies in various languages have tested how children utilize these two cues to comprehend transitive constructions. Dittmar, Abbot-Smith, Lieven and Tomasello (2008) examined the relative reliance of German-speaking toddlers on word order and case marking. They found that 2;6-year-olds could comprehend transitive sentences only when there was redundant marking of agent and patient (both cues were used). These findings were replicated across several languages (Cantonese: Chan, Lieven, & Tomasello, 2009; Japanese: Matsuo, Kita, Shinya, Wood, & Naigles, 2012; Warlpiri: O'Shannessy, 2010), and are in line with the predictions of The Competition Model, according to which a convergence of cues should facilitate comprehension of thematic assignment (Bates, McNew, MacWhinney, Devescovi, & Smith, 1982; Bates & MacWhinney, 1989; Ibbotson & Tomasello, 2009). However, in many of these cases, the redundant form is also the prototypical and the most frequent form in child-directed speech (Dittmar et al., 2008; Ibbotson & Tomasello, 2009). Therefore, it is not clear whether comprehension was facilitated because of the redundant cues, or because of the greater frequency of the prototypical structures (which happened to also have redundant cues). These explanations are hard to tease apart using natural language data, since individual cues are often correlated with one another (Ibbotson & Tomasello, 2008), and confounded with frequency.

[1] There is typological and historical debate about the relation between those two cues (with many languages showing a trade-off between the two), we return to this in the discussion.

The Current Study

In the current study, we use an artificial language to assess the impact of redundant morpho-syntactic cues on children's learning of a novel language. We compare the learnability of transitive constructions in two artificial languages: one with fixed OSV word order and no additional cues to thematic assignment (the non-redundant language), and the other with the same fixed OSV word order but with additional redundant case marking on objects (the redundant language). We used OSV word order because it differs from the dominant word order of Hebrew (SVO) - the language of the children in our study. Following exposure, we tested both comprehension and production of sentences in the novel language. Although the language without the case marking is simpler, in the sense of having fewer elements to learn, we predict that the redundant language will lead to better comprehension of thematic assignment because it contains redundant cues to indicate who-did-what-to-whom. Our prediction about the effect of redundancy on prediction is less clear-cut. On the one hand, if redundant markers indeed facilitate learning in general, this could aid production as well. Alternatively, the need to produce an additional element could make the sentences harder to produce. Such a dissociation between comprehension and production pressures is documented in other linguistic domains (e.g., Harmon & Kapatsinski, 2017).

Method

Participants

60 children participated in the experiment (age range: 7;0-9;0y, mean age: 7.10y, 41 boys and 19 girls). All children were visitors at the Bloomfield Science Museum in Jerusalem. They were recruited for this study as part of their

Figure 1: A trial example in the Sentence comprehension test phase.

Table 1: regression model for comprehension scores

	Estimate	Std. Error	z -value	p-value
(Intercept)	1.80068	0.21181	8.501	**<0.0001 *****
Condition (R-language)	1.03624	0.20390	5.082	**<0.0001 *****
Trial number	0.03324	0.03046	1.091	0.275
Age	-0.36756	0.32724	-1.123	0.261

visit to the Israeli Living Lab in exchange for a small reward. Parental consent was obtained for all children. All children were native Hebrew speakers, and none of them had known language or learning disabilities.

Materials

In both language conditions, participants were exposed to the same lexicon, which was composed of 6 semi-artificial Hebrew nouns (Hebrew nouns with nonce suffixes) and two Hebrew verbs. All nouns corresponded to masculine human characters, that were differentiated by their profession (e.g., clown, chef). The verbs were the Hebrew translations of "kick" and "touch". The constituent order of the language in both conditions was the non-Hebrew like OSV. In the redundant language (henceforth R-language) a nonce case marking ("patz") followed all objects, while in the control language (C-language) there was no such case marking. This cue was also non-native-like: Hebrew doesn't have post-nominal case-marking on objects. Participants saw and described the exact same drawings in both conditions[2].

Procedure

Participants were told they were going to meet some aliens who "say things differently from us" and that they would learn to speak like these aliens. Children were randomally assignmed to one of the two language conditions. Children sat with headphones in front of a computer next to a research assitant that provided them with verbal instructions. They saw drawings and heard recorded descriptions of these drawings in the alien-language (concatenated from recordings of the individual words spoken by a female Hebrew speaker). The experiment had several stages. First, a noun exposure phase, in which children saw each character, heard its name in the alien language, and had to repeat each name outloud (6 trials, one per noun). In both conditions, only the noun label was presented (without case marking). This was followed by a noun comprehension test (12 trials, two per noun) where participants saw two drawings, heard one label and had to match the label to the correct drawing. Feedback was provided after each trial. The following phase was sentence exposure (12 trials) where children saw a drawing of a transitive action (involving two of the characters, all characters could appear as agents and patients), heard a transitive sentence, and had to repeat it. The position of the agent and the patient in the drawing (left vs. right) was counterbalanced. The next stage was a sentence comprehension test (12 trials) where children saw two drawings of events, heard a sentence, and had to match the sentence to the correct drawing (see Figure 1). All the sentences here involved previously unheard combinations of agents and patients. The children had to use the mouse to choose the matching drawing. No feedback was given. The next phase was sentence production (12 trials) where chlidren saw previously unseen drawings of a transitive actions and had to describe them in the alien-language. Children's descriptions were recorded. Children in the R-language condition had one additional sentence forced-choice phase (12 trials) where they saw a previously unseen drawing, heard two descriptions of it, and had to choose the correct one. One option had case marking (like the sentences they heard before) and one was without case marking (as in the C-language). Children had to say which was a better way to describe the drawing by pressing on "1" or "2", corresponding to the order in which the options were presented. This phase was added to ensure that children in the R-language condition noticed the case marking cue.

Results

Comprehension

Children successfully learned the language (better than chance) in both conditions (C-language: M=65%, SD= 26%, t-test (29) = 3.1, p=0.004; R-language: M=91%, SD= 12%, t-test (29) = 26.8, p<0.0001). We used a mixed-effect logistic regression model to examine the effect of language condition on sentence comprehension (using the glmer function in R software, Bates, Maechler, Bolker, & Walker, 2015), and the maximum random effect structure justified by the data that converged, Barr, Levy, Scheepers, & Tily, 2013). The dependent variable was accuracy on each trial (as a binary variable). The model included fixed effects for condition (R-language vs. C-language, effect coded), age and trial number as centered continuous factors, and random intercepts for participants (see Table 1 for full model). As predicted, children showed better learning in the R-language condition (91% vs. 65%, β=1.04, SE=0.2, p<0.0001, Figure 2). Importantly, the difference in sentence comprehension

[2] The drawings were drawn by Sara Rolando from the University of Edinburgh, courtesy of Kenny Smith and Jennifer Culbertson

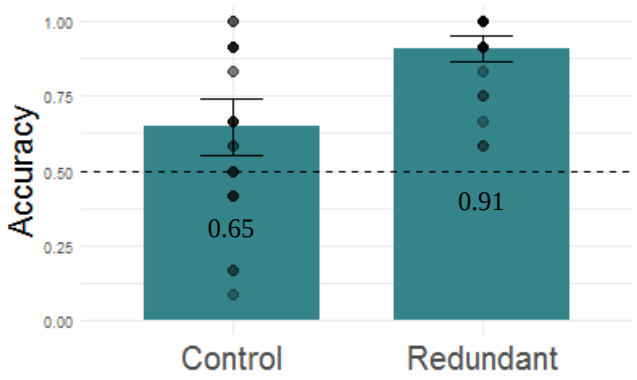


Figure 2: Accuracy scores by language condition. The dashed line indicates the chance level; error bars indicate confidence intervals; individual points indicate by-participant means.

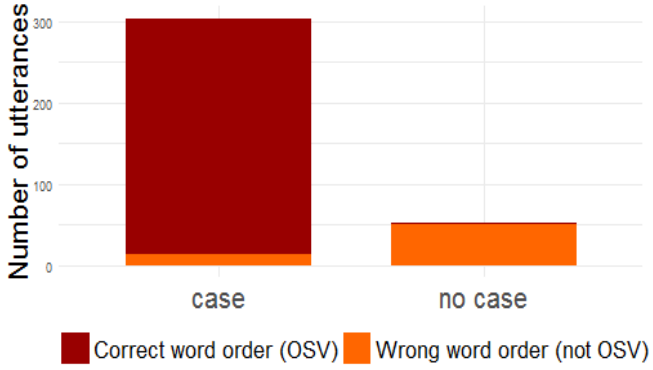


Figure 3: Production of correct word order (OSV) by production of case marking for children in the R-language condition.

did not stem from differential learning of the lexicon, both groups learned the nouns equally well (99% vs. 97%, t(58)=0.88, *p*=0.38).

Production

Children's productions were transcribed and coded for word order and vocabulary accuracy. Both measures were binary, and were scored by a research assistant blind to the condition and the experimental hypothesis. We used a mixed-effect logistic regression model to examine the effect of language condition on these production measures. The dependent variable was word-order accuracy on each trial (as a binary variable). The model included fixed effects for condition (R-language vs. C-language, effect coded), age and trial number as centered continuous factors, and random intercepts for participants. We found no significant difference in word order accuracy between the two conditions, although the trend was in favor of the R-language (82% vs. 69%, β=0.42, SE=0.86, *p*>0.6). We used the exact same model with lexical errors as the predicted variable and found no effect of condition here as well (0.1% vs. 0.07%, β=0.2, SE=0.18, *p*>0.2). Production did not seem to be facilitated in the R-language condition.

To further look at the possible facilitative effect of the redundant case marking, we looked only at the productions of children in the R-language condition (since only learners of the R-language had the potential to use both cues in production). We noticed several interesting patterns. First, children produced case marking in the majority of their productions (M=85%, SD=35%), indicating that they treated the cue as an inherent part of the system, despite its redundancy and the additional effort involved in producing it. Second, children produced the correct word order in most of the sentences (M=82%, SD=32%), indicating they managed to learn the non-Hebrew order. Interestingly, word order accuracy was impacted by the production of case marking: word order was more accurate when case marking was produced (291 correct utterances vs. 13 incorrect utterances) and was less accurate when it was not produced (only 1 correct utterance vs. 51 incorrect ones). To quantify this effect, we ran an additional model in which word-order accuracy on each trial was the dependent variable. The model included fixed effects for case marking on each trial (presence or absence, effect coded), age and trial number as centered continuous factors, and random intercepts for participants (see Table 2 for full model). In line with our hypothesis, case-marking had a significant effect on word order accuracy β=10.35, SE=2.7, *p*=0.0001, Figure 3).

Finally, we looked at participants' responses in the forced-choice part to sentences with and without case marking (note that only participants that learned the R-language had this additional part). Participants generally preferred the sentences with the case marking (M=88%, SD=19%), indicating, again, that they noticed the cue and learned it as part of their language.

Table 2: regression model for production scores in the R-language condition

	Estimate	**Std. Error**	**z -value**	**p-value**
(Intercept)	0.0826	1.8311	0.045	0.964019
Case	10.3504	2.7000	3.833	**0.0001 *****
Age	0.8787	2.5566	0.344	0.731079
Trial number	1.0889	0.7204	1.512	0.130652

Discussion

The presence of redundancy in languages is puzzling: if speakers are driven by a bias for efficient communication (Aylett & Turk, 2004; Cohen Priva, 2015; Jaeger, 2010; Kurumada & Jaeger, 2015; Levy & Jaeger, 2007), why do languages use more than one cue to convey the same information? Our study set out to test the prediction that redundant marking could be facilitative in learning circumstances. Our results show that having a redundant morpheme facilitates children's learning of thematic assignments. Although the R-language was more complex than the C-language, since it contained an additional cue to attend to and learn, children learned a non-native like word order better in this condition. The redundant morpheme benefited not only comprehension, but also production of the correct word order: despite the additional effort involved in producing it, word order was more accurate when case marking was produced. Taken together, these findings suggest that redundancy can be functional in learning circumstances. This is in line with previous work on the effect of learning from multiple cues (Monaghan, 2017; Sloutsky & Robinson, 2013; Yoshida & Smith, 2011). It further suggests that redundant cues can help language learning even when these cues need to be learned themselves.

These findings are of relevance for a recent influential proposal about the impact of different kinds of learners on the morphological complexity of a language. The Linguistic Niche hypothesis (Lupyan & Dale, 2010) proposes a causal link between the proportion of L2 speakers in a community and the degree of morphological complexity of the language. The prediction is that languages with more L2 speakers will have less complex morphology, a prediction that is supported by a large-scale study of over 2000 languages. Importantly, the proposed mechanism rests on the assumption that children and adults differ in the impact of redundancy on learning: whereas child learners benefit from redundant cues (leading to their existence in language), adult learners do not (leading to their simplification). While intuitively appealing, there is no direct evidence that children and adults differ in their response to redundant cues in learning. Building on the current findings, we are currently running a version of this study on adults to see if they indeed differ from children in the impact of redundant cues on learning. The Linguistic Niche hypothesis predicts that adult learning should be less facilitated by redundant cues. On the other hand, adult learners benefit from repetition (e.g., Jensen & Vinther, 2003; Onnis, Waterfall, & Edelman, 2008), and therefore may benefit from redundant cues as well.

Our findings document an effect of morpho-syntactic redundancy on learning, but they need to be extended in several ways. First, children in the R-language were exposed to the redundant morphome in both exposure and test. We are currently running additional versions of the study to better understand the impact of the redundant cue on exposure/testing. Second, we want to know if similar faciliatitve effects can be found for other linguistic domains beyond the learning of thematic assigment. Finally, the current study did not examine the prediction that redundancy is found more when conversing with learners. In an additional line of work, we investigate whether learning interactions are in fact characterized by more redundant marking, and whether this, in combination with their facilitative role can give rise to patterns of redundant marking in language.

Finally, the current findings are informative for our understanding of how languages are shaped by different cognitive pressures. Although in many languages the input children hear contains multiple cues for thematic assignment (Dittmar et al., 2008; Ibbotson & Tomasello, 2009), different typological studies suggest languages tend to trade off between these cues. In particular, languages that rely on word order to encode thematic assignment often lack productive case marking (Blake, 2001; Koplenig, Meyer, Wolfer, & Mu, 2017; Siewierska, 1998). Furthermore, several historical studies document this trade-off overtime in some Latin languages (e.g., Old English, Marchand, 1951, though see Detges, 2009; Pintzuk, 2002 for challenges to this claim). Recent experimental work suggests that this trade-off reflects speakers' bias for efficient communication (Fedzechkina et al., 2016; Roberts & Fedzechkina, 2018): When participants learned a novel language with fixed word order and optional case marking, they tended to decrease the use of case marking relative to their input. On the surface, these findings seem to contrast with our own: learners reduced the use of a redundant morpheme. However, this highlights the differential impact various pressures can have on language. First, case marking was deterministic in our design (case marking in the R-language was present on 100% of the objects), therefore, it is likely that participants were trying to faithfully reproduce it (see discussion in Fedzechkina et al., 2016). More importantly, we saw facilitation in comprehension: the impact of communicative and learnability pressures may differ for production and comprehension (e.g., Harmon & Kapatsinski, 2017). While redundancy may facilitate comprehension, it is costly in production (Zipf, 1949). These competing pressures (ease of production vs. understanding) may be weighted differently depending on the conversational situation: learning circumstances (or conversing with learners) may benefit from redundancy while other situations will not, leading to the observed trade-off between case-marking and word order seen in languages.

In sum, we have shown that children learn thematic assignment better from a language that has both word order and case marking, despite having to learn an additional morpheme. The present study serves as an important first step for understanding the functionality of having both cues.

References

Aylett, M., & Turk, A. (2004). The Smooth Signal Hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration

in spontaneous speech. *Language and Speech*, *47*(1), 31–56.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing : Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. Journal of Statistical Software, 67,1–48.

Bates, E., & Macwhinney, B. (1989). Functionalism and the Competition Model. In B. Macwhinney & E. Bates (Eds.), The crosslinguistic study of sentence processing (pp. 3–76). New York: Cambridge University Press.

Bates, E., McNew, S., MacWhinney, B., Devescovi, A., & Smith, S. (1982). Functional constraints on sentence processing: A cross-linguistic study. Cognition, 11(3), 245–299.

Blake, B. J. (2001). *Case*. Cambridge: Cambridge University Press.

Caballero, G., & Harris, A. C. (2012). A working typology of multiple exponence. In F. Kiefer, M. Ladányi, & P. Siptár (Eds.), *Current issues in morphological theory: (Ir)regularity, analogy and frequency. Selected papers from the 14th International Morphology Meeting, Budapest* (pp. 163–188). Amsterdam: John Benjamins Publishing.

Caballero, G., & Kapatsinski, V. (2015). Perceptual functionality of morphological redundancy in Choguita Rarámuri (Tarahumara). *Language, Cognition and Neuroscience*, *30*(9), 1134–1143.

Chan, A., Lieven, E., & Tomasello, M. (2009). Children's understanding of the agent-patient relations in the transitive construction: Cross-linguistic comparisons between Cantonese, German, and English. *Cognitive Linguistics*, *20*(2), 267–300.

Cohen Priva, U. (2015). Informativity affects consonant duration and deletion rates. *Laboratory Phonology*, *6*(2), 243–278.

Detges, U. (2009). How useful is case morphology? The loss of the Old French two-case system within a theory of preferred argument structure. In J. Barðdal & S. Chelliah (Eds.), *The role of semantic, pragmatic, and discourse factors in the development of case* (pp. 93–120). Amsterdam: John Benjamins Publishing.

Dittmar, M., Abbot-Smith, K., Lieven, E., & Tomasello, M. (2008). German Children's Comprehension of Word Order and Case Marking in Causative Sentences. *Child Development*, *79*(4), 1152–1167.

Fedzechkina, M., Newport, E. L., & Jaeger, T. F. (2016). Balancing Effort and Information Transmission During Language Acquisition: Evidence From Word Order and Case Marking. *Cognitive Science*, (March), 1–31.

Frank, A. F., & Jaeger, T. F. (2008). Speaking Rationally : Uniform Information Density as an Optimal Strategy for Language Production. *The 30th Annual Meeting of the Cognitive Science Society (CogSci08)*, 939--944.

Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, *110*, 8051–8056.

Gibson, E., Piantadosi, S. T., Brink, K., Bergen, L., Lim, E., & Saxe, R. (2013). A Noisy-Channel Account of Crosslinguistic Word-Order Variation. *Psychological Science*, *24*(7), 1079–1088.

Givón, T. (1991). Markedness in Grammar : Distributional , Communicative and Cognitive Correlates of Syntactic Structure. *Studies in Language*, *15*(2), 335–370.

Harmon, Z., & Kapatsinski, V. (2017). Putting old tools to novel uses: The role of form accessibility in semantic extension. *Cognitive Psychology*, *98*, 22–44.

Harris, A. C. (2017). *Multiple Exponence*. New York: Oxford University Press.

Haspelmath, M. (2008). Frequency vs . iconicity in explaining grammatical asymmetries. *Cognitive Linguistics*, (19), 1–33.

Ibbotson, P., & Tomasello, M. (2009). Prototype constructions in early language acquisition. *Language and Cognition*, *1*, 59–85.

Jaeger, T. F. (2010). Redundancy and reduction : Speakers manage syntactic information density. *Cognitive Psychology*, *61*(1), 23–62.

Jaeger, T. F. (2013). Production preferences cannot be understood without reference to communication. *Frontiers in Psychology*, *4*(April), 1–4.

Jaeger, T. F., & Buz, E. (2017). Signal Reduction and Linguistic Encoding. In E. Fernández & H. Cairns (Eds.), *The Handbook of Psycholinguistics* (pp. 38–81). Hoboken: John Wiley & Sons.

Jäger, G. (2007). Evolutionary Game Theory and Typology: A Case Study. *Language*, *83*(1), 74–109.

Jensen, E. D., & Vinther, T. (2003). Exact Repetition as Input Enhancement in Second Language Acquisition. *Language Learning*, *53*(3), 373–428.

Koplenig, A., Meyer, P., Wolfer, S., & Mu, C. (2017). The statistical trade-off between word order and word structure – Large-scale evidence for the principle of least effort. *PLoS ONE*, *12(3)*.

Kurumada, C., & Jaeger, T. F. (2015). Communicative efficiency in language production : Optional case-marking in Japanese. *Journal of Memory and Language*, *83*, 152–178.

Lee, H., & Kim, N. (2012). Non-Canonical Word Order and Subject-Object Asymmetry in Korean Case Ellipsis. In *Proceedings of the 19th International Conference on Head-Driven Phrase Structure Grammar Chungnam* (pp. 427–442).

Levy, R., & Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. In B. Schlökopf, J. Platt, & T. Hoffman (Eds.), *Advances in neural information processing systems (NIPS) 19* (pp. 849–856). Cambridge: MIT Press.

Lupyan, G., & Dale, R. (2010). Language Structure Is Partly Determined by Social Structure. *PLoS ONE*, *5*(1).

MacWhinney, B. (1987). The Competition Model. In B.

MacWhinney (Ed.), Mechanisms of language acquisition (pp. 249–308). Hillsdale, NJ: Lawrence Erlbaum.

Marchand, H. (1951). The syntactical change from inflectional to word order system and some effects of this change on the relation verb-object in English. *Anglia*, *70*, 70–89.

Matsuo, A., Kita, S., Shinya, Y., Wood, G. C., & Naigles, L. (2012). Japanese two-year-olds use morphosyntax to learn novel verb meanings. *Journal of Child Language*, *39*(3), 637–663.

Monaghan, P. (2017). Canalization of Language Structure From Environmental Constraints : A Computational Model of Word Learning From Multiple Cues. *Topics in Cognitive Science*, *9*, 21–34.

O'Shannessy, C. (2010). Competition between word order and case-marking in interpreting grammatical relations : a case study in multilingual acquisition, *38*(2011), 763–792.

Onnis, L., Waterfall, H. R., & Edelman, S. (2008). Learn locally , act globally : Learning language from variation set cues. *Cognition*, *109*(3), 423–430.

Piantadosi, S. T., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, *122*(3), 280–291.

Pintzuk, S. (2002). Morphological case and word order in Old English. *Language Sciences*, *24*, 381–395.

Roberts, G., & Fedzechkina, M. (2018). Social biases modulate the loss of redundant forms in the cultural evolution of language. *Cognition*, *171*, 194–201.

Siewierska, A. (1998). Variation in major constituent order : a global and a European perspective *. In A. Siewierska (Ed.), *Constituent Order in the Languages of Europe* (pp. 475–552). Berlin: Mouton De Gruyter.

Sloutsky, V. M., & Robinson, C. W. (2013). Redundancy Matters: Flexible Learning of Multiple Contingencies in Infants. *Cognition*, *126*(2), 156–164.

Yoshida, H., & Smith, L. B. (2011). Linguistic Cues Enhance the Linguistic Cues of Perceptual Learning. *Psychological Science*, *16*(2), 90–95.

Zipf, G. K. (1949). *Human behavior and the principle of least effort: An introduction to human ecology*. Cambridge: Addison-Wesley Press.

Bayesian Inference of Social Norms as Shared Constraints on Behavior

Zhi-Xuan Tan (xuan@mit.edu) and **Desmond C. Ong (desmond.c.ong@gmail.com)**
A*STAR Artificial Intelligence Initiative, Agency for Science, Technology and Research
1 Fusionopolis Way, Singapore 138632

Abstract

People act upon their desires, but often, also act in adherence to implicit social norms. How do people infer these unstated social norms from others' behavior, especially in novel social contexts? We propose that laypeople have intuitive theories of social norms as behavioral constraints shared *across* different agents in the same social context. We formalize inference of norms using a Bayesian Theory of Mind approach, and show that this computational approach provides excellent predictions of how people infer norms in two scenarios. Our results suggest that people separate the influence of norms and individual desires on others' actions, and have implications for modelling generalizations of hidden causes of behavior.

Keywords: Social Norms; Social Cognition; Bayesian Theory of Mind; Intuitive Theories

Introduction

Imagine entering a cafeteria in a foreign country that you know little about. There are but a handful of individuals in the cafeteria; you notice a tray-return receptacle at the far end, but you do not notice any signage on the walls detailing the expectations governing tray returns. If you observe someone carry their tray to the far end in order to return it, what inferences might you make? Does that person *like* returning trays, incurring the cost of walking across the room to do so? Or, is there an implicit social norm at play? A second person leaves without returning their tray. What might you now infer about them or about the potential social norm? Lastly, a third person approaches the second and loudly chastises them for not returning their tray, what would you then infer about everybody's preferences and the social factors at play?

Social norms are ubiquitous features of human societies, and as the example above suggests, we are able to rapidly infer their presence in novel situations. Children as young as three (Schmidt, Butler, Heinz, & Tomasello, 2016) demonstrate the ability to learn and generalize these rules of social behavior (Rakoczy & Schmidt, 2013). Not surprisingly, this ability continues into adulthood, allowing us, for example, to travel to new countries and then learn through observing others whether one is obliged to tip at restaurants, what the appropriate manner of greeting is, or what topics of conversation are considered impolite. In other words, we seem to possess not just an intuitive Theory of Mind (ToM) which allows us to infer the beliefs and desires of individuals, but also a corresponding ability to efficiently make inferences about shared norms that drive behavior *across* individuals. Furthermore, we appear naturally capable of disentangling the influence of social norms and individual desires: when we see someone picking up trash on the sidewalk, we infer that this is more likely due to an obligation to keep the streets clean rather than enjoyment of the act itself.

Despite (or perhaps because of) its ubiquity, how people *infer* social normativity is relatively understudied. The philosophical literature on social norms has generally focused on characterizing the precise nature of such norms—whether they are best understood as social practices, preferences conditioned upon shared expectations of behavior, or commonly-held normative attitudes (Bicchieri, 2005; Brennan, Eriksson, Goodin, & Southwood, 2013). Across philosophy, economics and psychology, there has also been an emphasis upon understanding the conditions and mechanisms for the emergence of norms (Hawkins, Goodman, & Goldstone, 2018)—whether they arise, for example, as Nash equilibria (Axelrod, 1986; Young, 2015), correlated equilibria (Gintis, 2010), or through maximization of cultural values (Bölöni, Bhatia, Khan, Streater, & Fiore, 2018). Other studies investigate how norms influence decision making (Chang & Sanfey, 2011), and how they are enforced (Fehr & Fischbacher, 2004). However, apart from a few simulation-based studies (Savarimuthu, Cranefield, Purvis, & Purvis, 2010; Cranefield, Meneguzzi, Oren, & Savarimuthu, 2016), research into how social norms are *inferred* remains scarce.

How then to explain our ability to infer social norms? In recent years, Bayesian models of cognition have begun to establish a computational basis for how people make social inferences. The Bayesian Theory of Mind (BToM) approach is perhaps the most prominent example, allowing researchers to formalize how people make graded judgments about unobservable mental states by observing the actions of others (Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017). This approach to social cognition has been extended to model reasoning about others' emotions (Ong, Zaki, & Goodman, 2015), inferring others' beliefs and desires from observed actions and emotional expressions (Wu, Baker, Tenenbaum, & Schulz, 2018), reasoning about how others balance costs and rewards in deciding how to act (Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016), learning how people value the welfare of others (Kleiman-Weiner, Saxe, & Tenenbaum, 2017), and inferring the presence of co-operation or competition from the behavior of multiple individuals (Shum, Kleiman-Weiner,

Littman, & Tenenbaum, 2019). Related work modelling human concept learning as Bayesian rule inference (Goodman, Tenenbaum, Feldman, & Griffiths, 2008) has also been used to develop theories of why people tend to learn act-based moral rules rather than outcome-based ones (Nichols, Kumar, Lopez, Ayars, & Chan, 2016).

We build upon this tradition of computational cognitive modelling, and hypothesize that people intuitively understand social norms as factors of behavioral influence which are shared across agents in a particular social context. These shared norms influence behavior alongside the individual desires of agents, and can generally be understood as injunctions or constraints on behavior, i.e., they prescribe, recommend, or prohibit certain kinds of actions[1]. We propose that people include these norms in their lay theories of social behavior, and we model these theories as Bayesian networks which include both norms and desires as possible causes of action. Judgments about the presence of a norm can thus be modelled using Bayesian inference conditioned upon observed actions, which can be made alongside desire inferences. Furthermore, since social norms are shared, inferences about them can be made from observations of multiple agents, unlike those for desires. We discuss several of these models, each of which captures a plausible intuitive theory of how norms influence both desires and actions. We then describe an experiment to test which model provides the best explanation of lay people's judgments in two social scenarios.

Computational Models

In order to study how people make inferences about social norms given observations of behavior, we choose to model situations where norm-driven actions are likely to be salient. In such situations, agents can take the role of **actors**, who are in the position to comply with a potentially applicable norm, or they can take the role of **judges**, who are in the position to enforce that norm after observing non-compliance. For simplicity, we restrict ourselves to the smallest multi-agent setting, with only one actor and one judge. The actor takes an action A_1, which corresponds to compliance or non-compliance with a potential norm. We also assume that the actor has some latent (binary) desire D_1 over action A_1 and its associated outcome. If the actor decides not to comply with the norm, the judge may take an action A_2, which corresponds to enforcement or non-enforcement of the potential norm. The judge also has a (binary) desire D_2 over the space of outcomes of A_1 (note however that D_2 is conditionally independent of A_1, since the judge's desires exist whether or not A_1 is taken). D_2 influences the enforcement action A_2 because A_2 can rectify the outcome produced by A_1. We denote the norm by N, which either exists ($N = 1$) or does not ($N = 0$) in the modelled situation.

[1]Here we are not interested in modelling descriptive norms—statistical regularities—since they can be directly learned through observation, though it is an interesting but separate research question as to how people might infer injunctive norms from descriptive ones.

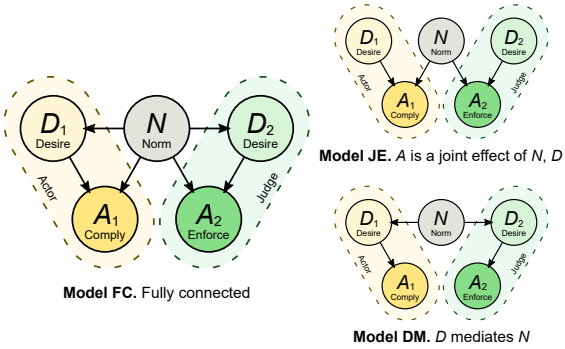


Figure 1: Possible intuitive causal models of how social norms influence behavior. We consider two agents, an Actor (on the left of each model, with nodes in yellow) and a Judge (on the right, with nodes in green). In Model FC, norms influence desires, and both norms and desires jointly influence actions. In Model JE, desires are independent of norms, while in Model DM, desires are the sole mediators of norms upon behavior.

We consider three plausible models of laypeople's intuitive theories of norm-driven behavior (Fig. 1). In the **Fully Connected (FC)** model, we posit that social norms influence the desires of each agent, reflecting the idea that norms can be (partly) internalized, such that norm compliance becomes part of one's desire. We also assume that norms and desires jointly affect an agent's actions. This corresponds to the notion that whether or not a norm is internalized, it continues to directly influence actions, whether by imposing an expected social cost to non-compliance, or simply by having a normative force that is separate of individual desire.

In our second candidate model, the **Joint Effect (JE)** model, desires and norms still jointly influence actions, but agent desires are independent of norms. That is, lay people's notions of what agents 'want' to do are separate from what they 'should' do, mapping roughly to the Kantian distinction between desire and duty. Lastly, the **Desire Mediation (DM)** model assumes that norms do not influence actions directly, but only when mediated by desires. This assumption corresponds to the Humean notion that desires are the sole motives for action—that 'shoulds' cannot influence action unless they also become 'wants'—a notion which might plausibly feature in lay intuitions about norms. In addition to these three 'complete' models, we also tested two lesioned models: a desire-only model (**D-only**) and a norm-only model (**N-only**).

It is worth emphasizing that in proposing these models, we are *not* attempting to give a rigorous philosophical account of the relationship between norms, desires, and actions, nor are we attempting to argue that social norms cannot be explained in terms of desires, or for that matter, other cognitive variables such as shared beliefs or expectations. Rather, our intention is to propose models of how people *intuitively* understand norms and norm-driven behavior.

Experiment

We conducted an experiment through Amazon's Mechanical Turk (AMT) to elicit lay people's likelihood judgments about norms, desires and actions in two different social scenarios: one involving an obligative norm—the norm that people should return their trays after eating—and another involving a prohibitive norm—the norm that people should not litter. Given widespread intuitive acceptance of the act-omission distinction (Kahneman & Sunstein, 2005), we expected that people might also respond differently to obligations and prohibitions. We also chose common, but not universal, social norms, so that we could better observe how people adjusted their certainty about the existence of each norm.

Methods

We provide a sample of our experiment, our data, and analysis code at `https://github.com/ztangent/norms-cogsci19`.

Scenario Structure. Both scenarios were identical in structure—participants were introduced to an actor in a position to comply with a potentially applicable norm, and asked to make various likelihood judgments. They were then shown the actor not complying with the potential norm, introduced to a judge who (unknown to the actor) had observed the actor's action, and made another round of judgments.

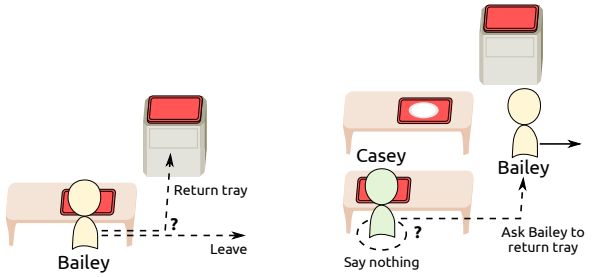
Experimental Conditions. To measure how well our proposed models predict lay people's inferences, as well as determine which model best captures intuitions about norms, we divided each each scenario into five conditions, each querying for different sets of likelihood judgments:

A. Norm and desire priors, and desires given norms:
   $P(N)$, $P(D_1)$, $P(D_2)$, $P(D_1|N)$, $P(D_2|N)$.
B. Actions conditioned on desires only: $P(A_1|D_1)$, $P(A_2|D_2)$.
C. Actions conditioned on norms only: $P(A_1|N)$, $P(A_2|N)$.
D. Actions conditioned on both: $P(A_1|D_1,N)$, $P(A_2|D_2,N)$.
E. Norm and desire posteriors:
   $P(D_1|A_1)$, $P(D_2|A_1,A_2)$, $P(N|A_1)$, $P(N|A_1,A_2)$

Data from conditions A–D were used both to calibrate the models and to investigate people's intuitions about the relationship between norms, desires and actions, e.g., by comparing $P(D_1|N)$ (condition B) to $P(D_1)$ (condition A) to see if people judge desires to be dependent upon norms. After calibrating the models, data from condition E was compared against the models' posterior inferences to see if they predicted participants' inferences about norms and desires.

Participants. We recruited 200 US participants (mean age 35.6, SD 11.2; 104 male, 95 female, 1 unreported) via AMT, restricting to those with a HIT approval rate of 99% and above. All participants went through both scenarios in random order, and each participant was randomly assigned to a different condition within each scenario (i.e. assigned conditions for Scenario 1 and 2 were independent). For Scenario 1, conditions 1A through 1E, sample sizes were n =51, 24, 25, 51 and 49 respectively. For Scenario 2, conditions 2A through 2E, sample sizes were n =49, 25, 25, 50 and 51 respectively. Assuming a large effect size (Cohen's $f = 0.5$), these sample sizes give $> 93\%$ power for one-way ANOVA between sub-conditions at the 5% significance level (e.g. comparing $P(A_1|N=1)$ and $P(A_1|N=0)$ in condition C).

Scenario 1: An Obligative Norm



(a) Actor's choices (b) Judge's choices

Figure 2: Returning one's tray as an obligative social norm.

Details. Participants were presented with a vignette with two phases, with text descriptions accompanied by illustrations. In the first phase (Figure 2(a)), Bailey, the actor, has finished a meal served on a tray in a restaurant. The restaurant has a tray return station, and Bailey can choose to either return ($A_1 = 1$) or leave ($A_1 = 0$) the tray. Since the norm at play was obligative ('People should return their trays after eating.'), $A_1 = 1$ corresponds to compliance with the norm (if it exists, i.e, if $N = 1$). In the second phase, participants are told that Bailey decides to leave the tray, and are introduced to Casey, who, unknown to Bailey, has watched this occur (Figure 2(b)). Casey now has the option of either asking Bailey to return the tray ($A_2 = 1$) or saying nothing ($A_2 = 0$), where $A_2 = 1$ corresponds to enforcement of the potential norm.

Questions were presented after *each* phase was introduced, asking participants to make judgments depending on the condition they were assigned. When queried for priors (condition A), participants were directly asked how likely they thought a certain state of affairs was true (e.g. "How likely do you think Casey *wants* the tray to be returned?" for $P(D_2)$, "How likely do you think the following norm exists?" for $P(N)$). When queried for conditional likelihoods, including the posteriors, participants were first asked to suppose a certain state of affairs (e.g. "If you saw Casey ask Bailey to return the tray," for $P(\cdot|A_1=0,A_2=1)$ or "Suppose that Bailey does not want to return the tray." for $P(\cdot|D_1=0)$), and then asked to give likelihood judgments given those suppositions. To make these counterfactuals more concrete in the case of the posterior inferences (condition E), we also provided corresponding illustrations of the counterfactual actions.

Results. To determine if participants intuitively judged desires to be independent of norms, we first analyzed the data from condition A to see if $P(D_i)$, $P(D_i|N=1)$ and $P(D_i|N=0)$ ($i \in \{1,2\}$) exhibited significant differences. As can be

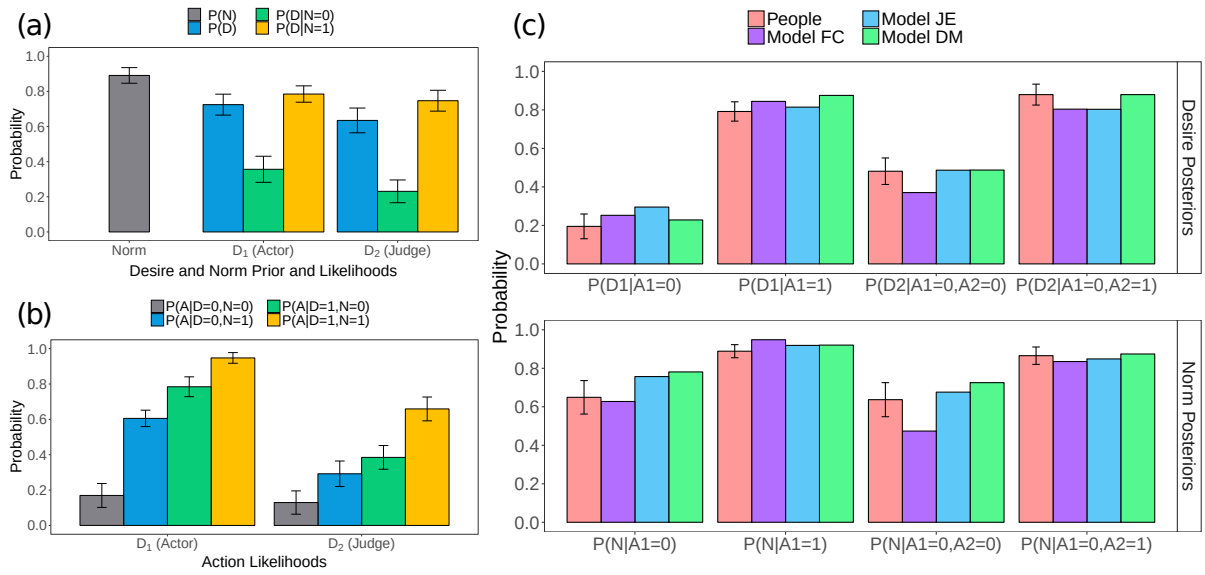


Figure 3: Scenario 1 Results. (a) Empirical norm and desire judgments from condition 1A. (b) Empirical action likelihoods, from condition 1D. (c) Comparing empirical posteriors from condition 1E with posterior judgments from our proposed models.

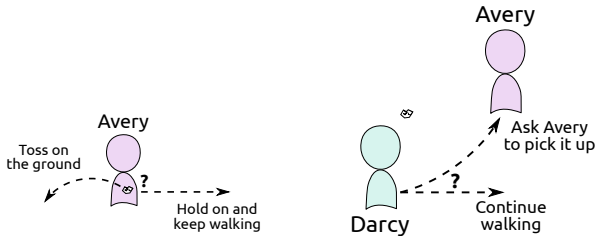
seen in 3(a), this was indeed the case, with one-way ANOVA giving $F(2,150) = 55.13$, $p < 0.001$, for the actor's conditional and prior desires (D_1), and $F(2,150) = 66.95$, $p < 0.001$, for the judge's (D_2) conditional and prior desires. This provides evidence against Model JE, which assumes that desires are independent of norms.

Next, to determine if desires and norms jointly influence behavior, we analyzed the conditional action likelihoods from condition D. As Figure 3(b) shows, given a fixed value of desire, there were significant differences between action likelihoods when the norm was either absent or present (all ts > 5.77, all ps < 0.001, df$= 50$, paired test). That is, regardless of whether the agent *wanted* to act, the presence of the tray-return norm ($N = 1$) led people to judge both norm compliance ($A_1 = 1$) and norm enforcement ($A_2 = 1$) as more likely. This provides evidence against Model DM, which assumes norms have no direct effect on actions. (For brevity, we omit comparisons between $P(A_i|D_i,N)$, $P(A_i|D_i)$ and $P(A_i|N)$, $i \in \{1,2\}$ using the data from conditions B and C, but these display significant differences as well.)

Finally, we computed the desire and norm posteriors under the FC, JE and DM models, then compared them against participants' posterior judgements, as shown in Figure 3(c). All three models displayed high correlation with the empirical data (FC: $r = 0.944$, JE: $r = 0.974$, DM: $r = 0.981$). The correlations of the lesioned models were worse by comparison (D-only: $r = 0.944$, N-only: $r = 0.384$). Both norm and desire posteriors increased when compliance ($A_1 = 1$) or enforcement ($A_2 = 1$) were observed, and decreased otherwise. The three models also captured people's ability to integrate information across multiple agents to infer the presence of norms—when non-compliance by the actor ($A_1 = 0$) is observed, the likelihood of the norm's existence decreases ($P(N) > P(N|A_1 = 0)$), but when enforcement by the judge (A_2) is subsequently observed, the likelihood of the norm increases again ($P(N|A_1 = 0) < P(N|A_1 = 0, A_2 = 1)$).

Despite the desire and action likelihoods providing strong evidence against the JE and DM models, these models were surprisingly more correlated with participants' posterior judgements than the FC model. The results for Scenario 1 are thus hard to interpret conclusively. Plausibly, this was due the high degree of inter-subject variance in likelihood ratings, suggesting that people's intuitive models of social normativity have substantial heterogeneity.

Scenario 2: A Prohibitive Norm



(a) Actor's choices (b) Judge's choices

Figure 4: Not littering as a prohibitive social norm.

Details. As in Scenario 1, participants were presented with a two-phase vignette. In the first phase (Figure 4(a)), Avery, the actor, is walking along a city street while holding on to some crumpled paper. Avery can choose to either toss the paper ($A_1 = 1$) or continue holding on ($A_1 = 0$). Since the norm at play was prohibitive ('People should not discard their belongings on the ground.'), $A_1 = 1$ corresponds to violation of

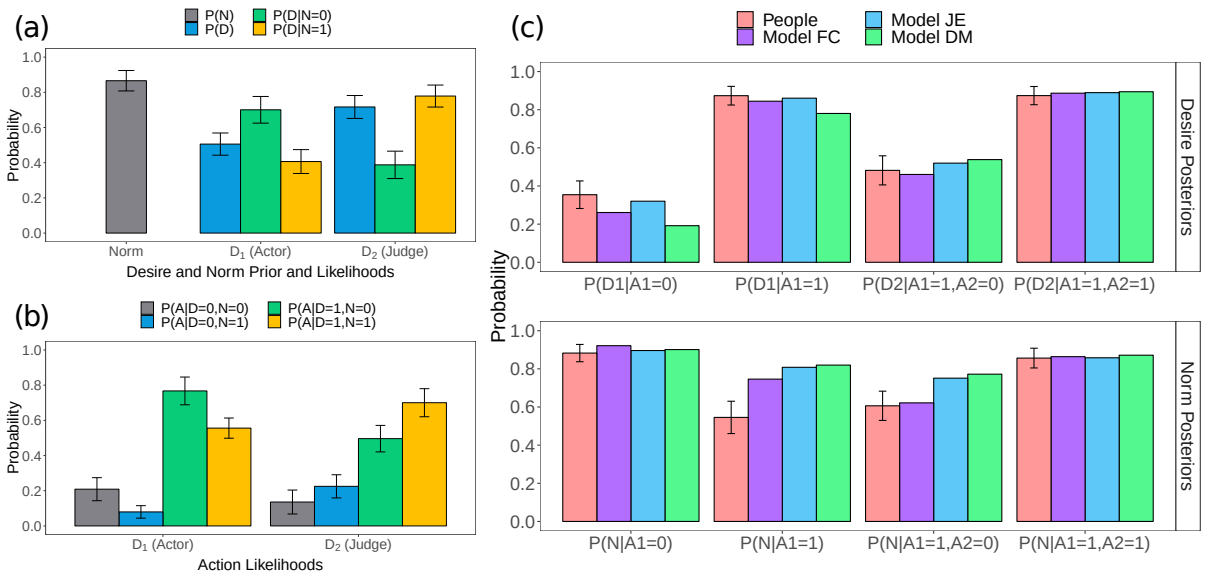


Figure 5: Scenario 2 Results. Figure is similar to Fig. 3 except for the fact that Scenario 2 involved a prohibitive norm, where $A_1 = 1$ (littering) contravenes the norm. This explains the different patterns for the likelihood of Actor variables (D_1, A_1).

the norm, unlike in Scenario 1. In the second phase, participants are told that Avery decides to toss the crumpled paper on the ground, and are introduced to Darcy, who, unknown to Avery, has watched this occur (Figure 4(b)). Darcy now has the option of either asking Avery to pick up the crumpled paper ($A_2 = 1$) or saying nothing ($A_2 = 0$), where again $A_2 = 1$ corresponds to enforcement of the potential norm. After each phase was introduced, participants were asked to make likelihood judgments depending on the condition they were assigned, using the same question formats as in Scenario 1.

Results. First, we analyzed the data from condition A to see if participants judged desires to be norm-dependent. As 5(a) shows, this was once more the case, with one-way ANOVA giving $F(2,144) = 18.02$, $p < 0.001$, for the actor's conditional and prior desires (D_1), and $F(2,144) = 35.86$, $p < 0.001$, for the judge's (D_2) conditional and prior desires. We then analyzed the conditional action likelihoods from condition D, and found similarly that there were significant differences when the norm was either absent or present (all ts > 3.10, all ps < 0.005, df$=49$, paired test). To be clear, the influence of the norm here was in the *opposite* direction—$N = 1$ led to littering $(A_1 = 1)$ being less likely.

Lastly, we computed the desire and norm posteriors under the various models and compared against the empirical data, as shown in Figure 5(c). Compared to Scenario 1, more pronounced differences could be observed between models. In particular, Model DM significantly over-estimates the norm's likelihood when a norm-violating or non-enforcing action is taken (see Figure 5(c) $P(N|A_1 = 1)$, $P(N|A_1 = 1, A_2 = 0)$), because it attributes the cause of the action primarily to the desire not to comply with or enforce the norm. In contrast, Model FC better predicts that the norm's likelihood should decrease when norm violation is observed. This is because there are multiple causal pathways that lead from the norm to the actions in Model FC—norm-violating actions directly imply that the norm is unlikely to exist, but they also imply that the desire for the norm-violating action exists, which indirectly implies the non-existence of the norm. Model JE over-estimates the norm's likelihood slightly less than Model DM, but still more than Model FC, because it has only one causal pathway from the norm to the action.

Nonetheless, all three models still correlated highly with the data (FC: $r = 0.934$, JE: $r = 0.887$, DM: $r = 0.828$), and the lesioned models again performed worse (D-only: $r = 0.772$, N-only: $r = 0.461$). Both norm and desire likelihoods increased when compliance $(A_1 = 0)$ or enforcement $(A_2 = 1)$ were observed, and decreased otherwise. We similarly observed that people integrated information from multiple agents: the likelihood of the norm decreases after observing norm violation $(P(N|A_1 = 1) < P(N))$, but increases again after subsequently observing enforcement $(P(N|A_1 = 1, A_2 = 1) > P(N|A_1 = 1))$.

Unlike in Scenario 1, Model FC displayed the highest degree of correlation out of our three proposed models. This, combined with the analysis showing that both actions and desires are directly norm-dependent, provide strong evidence for model FC over the other two models. One reason this might have been the case for Scenario 2 is that a prohibitive norm tends to *conflict* with the direction of desire—people are more likely to act how they believe they *should*, whatever they happen to *want* for themselves. This would disfavor Model DM (hence its over-estimation of the norm's likelihood) but favor Model FC, because it betters captures the restraining effect of norms on both desires and actions.

General Discussion

We experimentally investigated how laypeople infer norms from behavior, and showed that a Bayesian model provides excellent predictions of people's posterior inferences of both obligative and prohibitive norms. We tested several plausible theories, and found strong evidence that people understand norms to directly influence both desires and actions. This suggested that the JE and DM models should be ruled out, leaving the FC model. While our analysis of the model's posterior inferences in Scenario 1 did not unambiguously support this conclusion, the corresponding analysis for Scenario 2 did so, with the FC model showing the highest correlation when averaged across both scenarios (FC: $r = 0.939$, JE: $r = 0.931$, DM: $r = 0.905$). Furthermore, comparison with lesioned models showed that accurate inferences cannot be made by omitting either desires on norms.

Our results lend support to our hypothesis that people understand social norms as behavioral constraints shared *across* agents, in contrast to preferences that are idiosyncratic to individual agents. In this way, people are able to observe different actions made by individuals in different roles, integrating that information and allowing them to rapidly make inferences about the presence of social norms in a given context. This ability is highly useful, for it allows us navigate unfamiliar social environments without deducing the preferences of every stranger about how one should act.

Of course, not all environments are unfamiliar—people spend their whole lives with a familiar set of norms. Thus, one might expect a person to bring strong expectations to bear when making inferences about norms in a new, but familiar, situation. Indeed, this was the case for many participants in our experiment—while we constructed scenarios with common but not universal norms, participants were often *certain* that the norm in question was present, even before observing any actions. These priors made it harder for our experiment to detect whether people deemed a norm *more* likely to exist after observing norm compliance, but made it easier to detect when people deemed a norm less likely to exist after observing a norm violation. Future experiments should introduce participants to a more alien environment where they have no sense of what the norms might be, and see if they can infer the presence of a norm through a few observations.

Participants not only came in with varied priors, but also varied likelihood judgements, with some participants giving more weight to the influence of norms than others. This heterogeneity may help explain why our results for Scenario 1 were not conclusive—participants' internal models might have diverse parameters, and some might even have different model structures altogether. As such, while the results clearly showed that the models predicted average judgments with high correlation, the ability to distinguish the exact model type may have been lost. Still, it is interesting that even with such diversity, people are still able to rapidly learn and converge upon the same set of norms. How exactly this convergence occurs is a topic worth exploring further.

In conclusion, we have demonstrated a principled, computational framework for how people infer the shared drivers of behavior that we call social norms. In addition, our results give us insight into people's intuitive theories of norms as influencing *both* our desires and our actions. This builds upon previously studied models that infer the beliefs, desires, and intentions of single agents, extending their inferential capacity to large groups of agents constrained by shared context. By laying the groundwork for how we make these inferences, our work elucidates one way in which we make sense of the full richness of social life—and how we *ought* to live it.

Acknowledgments

This work was supported by the A*STAR Human-Centric Artificial Intelligence Programme (SERC SSF Project No. A1718g0048).

References

Axelrod, R. (1986). An evolutionary approach to norms. *American Political Science Review*, *80*(4), 1095–1111.

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, *1*(4), 0064.

Bicchieri, C. (2005). *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.

Bölöni, L., Bhatia, T. S., Khan, S. A., Streater, J., & Fiore, S. M. (2018). Towards a computational model of social norms. *PloS one*, *13*(4), e0195331.

Brennan, G., Eriksson, L., Goodin, R. E., & Southwood, N. (2013). *Explaining norms*. Oxford University Press.

Chang, L. J., & Sanfey, A. G. (2011). Great expectations: neural computations underlying the use of social norms in decision-making. *Social Cognitive and Affective Neuroscience*, *8*(3), 277–284.

Cranefield, S., Meneguzzi, F., Oren, N., & Savarimuthu, B. T. R. (2016). A bayesian approach to norm identification. In *Proceedings of the Twenty-second European Conference on Artificial Intelligence* (pp. 622–629).

Fehr, E., & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior*, *25*(2), 63–87.

Gintis, H. (2010). Social norms as choreography. *Politics, Philosophy & Economics*, *9*(3), 251–264.

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, *32*(1), 108–154.

Hawkins, R. X. D., Goodman, N. D., & Goldstone, R. L. (2018). The emergence of social norms and conventions. *Trends in Cognitive Sciences*.

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in Cognitive Sciences*, *20*(8), 589–604.

Kahneman, D., & Sunstein, C. R. (2005). Cognitive psychology of moral intuitions. In *Neurobiology of Human Values* (pp. 91–105). Springer.

Kleiman-Weiner, M., Saxe, R., & Tenenbaum, J. B. (2017). Learning a commonsense moral theory. *Cognition*, *167*, 107–123.

Nichols, S., Kumar, S., Lopez, T., Ayars, A., & Chan, H.-Y. (2016). Rational learners and moral rules. *Mind & Language*, *31*(5), 530–554.

Ong, D. C., Zaki, J., & Goodman, N. D. (2015). Affective cognition: Exploring lay theories of emotion. *Cognition*, *143*, 141–162.

Rakoczy, H., & Schmidt, M. F. (2013). The early ontogeny of social norms. *Child Development Perspectives*, *7*(1), 17–21.

Savarimuthu, B. T. R., Cranefield, S., Purvis, M. A., & Purvis, M. K. (2010). Obligation norm identification in agent societies. *Journal of Artificial Societies and Social Simulation*, *13*(4), 3.

Schmidt, M. F., Butler, L. P., Heinz, J., & Tomasello, M. (2016). Young children see a single action and infer a social norm: Promiscuous normativity in 3-year-olds. *Psychological Science*, *27*(10), 1360–1370.

Shum, M., Kleiman-Weiner, M., Littman, M. L., & Tenenbaum, J. B. (2019). Theory of minds: Understanding behavior in groups through inverse planning. In *AAAI 2019*.

Wu, Y., Baker, C. L., Tenenbaum, J. B., & Schulz, L. E. (2018). Rational inference of beliefs and desires from emotional expressions. *Cognitive Science*, *42*(3), 850–884.

Young, H. P. (2015). The evolution of social norms. *Annual Review of Economics*, *7*, 359–87.

Utilizing eye-tracking to explain variation in response to inconsistent message on belief change in false rumor

Yuko Tanaka (tanaka.yuko@nitech.ac.jp)
Graduate School of Engineering, Nagoya Institute of Technology
Gokiso-cho, Showa-ku, Nagoya, Aichi 466-8555 Japan

Miwa Inuzuka (minuzuka@u-gakugei.ac.jp)
Department of Education, Tokyo Gakugei University
4-1-1 Nukui-kita-machi, Koganei, Tokyo, 184-8501 Japan

Rumi Hirayama (r.hirayama@daion.ac.jp)
Osaka Junior College of Music
1-1-8 Shonaisaiwaimachi, Toyonaka, Osaka, 561-8555 Japan

Abstract

Exposure to Inconsistent message has been demonstrated as a useful method to alleviate belief in false rumor. However, the data from previous research included unexplained variation in response to inconsistent message. Existing research also included the possibility that participants skipped out on reading and therefore they were not exposed to a message. We used an eye tracker to eliminate the possibility. Eye tracking data revealed that participants not only did not skip but they paid more visual attention to inconsistent messages comparing with consistent messages. Despite the overall effectiveness of inconsistent message, some responses showed continued belief in rumors even after the exposure. Eye-tracking analyses demonstrated that when participants had positive pre-belief for a rumor, more visual attention to inconsistent message predicted strengthened the belief. We discuss when exposure to inconsistent message does not work well as a way for harnessing belief in false rumor.

Keywords: rumor; belief change; eye tracking; social media

Introduction

The recent exponential growth of social media, such as Twitter and Facebook, has been affecting how rumors spread. Once a rumor is posted on social media, it can be shared widely in a very short time. In this sense, social media can be a salient rumor mill. Additionally, while prior eras included the spread of rumors by word of mouth, online rumors never go away completely online. Accumulation of rumors can increase the risk of misunderstanding, miscommunication, and potential social problems.

Several researchers have provided definitions for rumors. One popular definition is, "public communications that are infused with private hypotheses about how the world works" (Rosnow, 1991). Although other definitions emphasize aspects that include circulation in contexts of ambiguity, danger, or potential threat (e.g., DiFonzo & Bordia, 2007), we should not ignore the risk of spreading rumors within normal everyday interactions. As misinformation, propaganda, and "fake news" are diffused every day under the semblance of rumors, the risk for rumor belief should be acknowledged more broadly.

How we handle rumors in the digital age? Past studies have demonstrated that exposure to inconsistent messages, including denial, rebuttal, and criticism, was effective to mitigate belief in various types of rumors, such as an alleged misdemeanor (Koller, 1993), organizational rumors (DiFonzo & Bordia, 2000), disaster related rumors (Tanaka, Sakamoto, & Matsuka, 2013), ill effects of smoking (Iyer & Debevec, 1991), and a computer virus (Bordia, DiFonzo, Haines, & Chaseling, 2005). Bordia et al. (2000) revealed that exposure to a denial message mitigated belief in a rumor. In reality, recent studies have reported that while many people spread rumors through social media, others try to stop the spread of false rumors by posting inconsistent messages(Mendoza & Poblete, 2010; Starbird et al., 2016).

Despite the overall effectiveness of inconsistent message exposure, previous studies have shown the variation in the effect. For example, an experiment reported by Bordia et al. (2005) showed that an average belief in a rumor was reduced from 5.10 to 3.60 (1 = not at all believable to 7 = totally believable) after inconsistent message exposure. Although this is a significant belief reduction the result indicates that participants evaluate the rumor as moderately believable even after exposure to an inconsistent message. The same patterns emerged in another study by Tanaka et al. (2013) where some participants did not change their belief in rumor after they were exposed to an inconsistent message and even decided to spread the rumor. To consider the practical application of rumor control to real life, it is necessary to understand the responses of continual believing rumor despite inconsistent messages.

The present study examined the mixed effectiveness of inconsistent message exposure on belief change in rumors. We measured eye movements of participants when they were exposed to an inconsistent message with a rumor. The reason for this is to check whether participants are genuinely exposed to an inconsistent message. Even if a participant was asked to read an inconsistent message on a traditional

questionnaire or a computer screen, the possibility that he or she skipped reading the message still remains. Recent eye-tracking research demonstrated that people tend to skip reading posts on social media when they look uninteresting (Bode, Vraga, & Troller-Renfree, 2017). Thus, to examine the effect of inconsistent message exposure, it is important to eliminate the possibility of participants skipping out on reading and not actually being exposed to the message.

We focus on fixation duration and fixation frequency as eye movements. Fixation is defined as the periods when an eye is close to immobile and distinguished from rapid movement termed saccades (Rayner, Pollatsek, Ashby, & Clifton, 2012, p.91). A reader extracts printed visual information during fixation. Fixation duration tends to be longer when text becomes conceptually difficult (Rayner, 1998) and unpredictable (Ehrlich & Rayner, 1981; Rayner, Slattery, Drieghe, & Liversedge, 2011).

Based on earlier research, we hypothesize that exposure to an inconsistent message with a targeted rumor will cause belief change in the rumor by reducing pre-belief when compared with the control response to a targeted filler after consistent message exposure (Hypothesis 1). If an effect was found, eye movement record can help determine whether it was caused by exposure to, or skipping an inconsistent message. We expect that eye movement record will show that participants were exposed to an inconsistent message because belief change would not occur if they are skipping the inconsistent message (Hypothesis 2). As for eye movements, we expect that fixation duration will be longer when participants are reading an inconsistent message because the inconsistent message is not predictable based on prior experience of reading a rumor (Hypothesis 3). Predictability effect was observed not only in alphabetic scripts such as English but also in a logographic script (Rayner, Juhasz, & Yan, 2005), therefore, it is reasonable to apply this hypothesis to Japanese which uses logographic characters. This kind of research has implications for understanding what makes people stop believing or keep believing rumors.

Method

Participants

The participants were 46 college undergraduate and graduate students in Japan (32 males, 14 females, M_{age} = 20.8, SD_{age} = 1.89). They received 1,000 Japanese yen (about 9.00 USD) for their participation in an approximately 50-minute session. They all reported having Japanese as their native language and 72% reported attending a psychology class.

Materials

Stimuli For rumor tweets, 12 false rumors related to popular psychology topics were selected from the Japanese translation of Lilienfeld, Lynn, Ruscio, & Beyerstein (2010), including topics such as "*Subliminal messages can persuade people to purchase products*" and "*People use only 10 % of their brain power*" (see Appendix for stimulus materials). All rumors were written in Japanese horizontally and the number of characters was controlled to fall within the range of 46 to 48. Each rumor was transformed into a Twitter PNG image tweet. The user name associated with each tweet was randomly generated.

For each rumor, an inconsistent message with the rumor was developed based on the criticisms against the rumors (Lilienfeld et al., 2010). An inconsistent message was operationally defined as a message including inconsistent or contradictory information against a target rumor. For example, an inconsistent message for the rumor regarding "*subliminal messages*" mentioned above was "*An analysis of research from a Canadian television station revealed that the subliminal message 'please telephone us right away' was aired 352 times. However, there was no increase in the incoming telephone calls. Likewise, people cannot be made to buy things in this manner"*. For each of the 12 false rumors, an inconsistent tweet was developed. The number of characters was controlled to fall within the range of 70 to 75.

In order to prevent the participants learning the characteristics of the inconsistent message stimuli and acting strategically, filler stimuli were added. Twelve tweets were created from psychological knowledge based on textbooks and made into filler stimuli. For each filler, a consistent message was developed. There were no significant differences in the number of characters between rumor and filler tweets, and inconsistent and consistent messages.

Figure 1. A slide image on the eye-tracking computer. It presents a set of a rumor tweet (target: upper left) and its inconsistent message (bottom-left).

Apparatus

Eye movements were recorded only in the following inconsistent message exposure phase by a Tobii Pro Nano, which samples eye position at 60 Hz. All images were presented on a 17.3-inch display with a screen resolution of 1920 × 1080 pixels. Participants were seated ~60cm from the display. For each participant, the system was calibrated before the experiment using a set of 5 calibration points covering the whole screen area. Informed written consents

from participants were obtained. A 23.8-inch display was used except in the eye-tracking phase.

Procedure

Each participant was tested individually. Participants were told that the experiment concerned understanding students' knowledge about psychology, and it was not revealed that the research was interested in false rumors and the inconsistent message exposure until the debriefing period at the end of the experiment. The experiment was administered in the following order.

1. Pre-belief measurement The rumor tweets and the filler tweets were presented one at a time on a computer screen. Presentation of the stimuli was randomized for each participant. Participants were not informed that some stimuli were false. They were asked to answer the following three questions about each tweet: (1) Familiarity – How much do you know about this information? (*Well, Slightly, Not at all*); (2) Accuracy – How accurate do you think this information is? (1 *Not at all*, 5 *Highly accurate*); (3) Importance – How important do you think this information is? (1 *Not at all*, 5 *Highly important*).

2. Inconsistent/consistent message exposure Inconsistent message was a message including inconsistent or contradicting information. For example, the inconsistent message for the "*subliminal messages*" rumor refers to research which showed that the subliminal effect was not observed. On the other hand, consistent messages for fillers mentioned supportive examples and did not include any inconsistent or contradict information (see Appendix). After a five-point calibration for each participant using Tobii Pro Lab software (Tobii Technology), the 12 sets of a rumor tweet and an inconsistent message were presented one at a time mixing with the 12 sets of a filler tweet and a consistent message. The order of presenting the 24 sets was counterbalanced. Figure 1 shows a slide which presents one of the sets. Participants were instructed that each message was referring to the message of the target tweet. They were asked to read each set of tweets silently at their own pace and to judge the message interesting or not interesting. They were required to press 4 on the numeric keypad if the message was not interesting and 6 if the message was interesting.

3. Post-belief measurement The same set of target tweets from the pre-belief session were shown a second time. Participants evaluated accuracy and importance for each tweet.

After completing all tasks, participants were debriefed as to the purpose of the study. It was emphasized that some tweets were false. Participants provided another informed consent.

Data analyses

Eye-movement type and eye coordinates were recorded per millisecond (ms) with Tobii Pro Lab software throughout the message exposure phase. Direct visual attention (fixated) was extracted from the raw eye-tracking data using a minimum fixation duration of 100ms. To identify visual attention per tweet, we calculated fixation duration for target and message regions respectively. For each region, fixation frequency and total fixation duration were calculated. As the number of characters in each tweet was different between target tweets and messages, fixation duration rate was calculated per tweet, for both target tweet and message, by dividing the total fixation duration by the number of characters. We also calculated fixation frequency rate by dividing the total frequency of each tweet by the number of characters.

Results

Table 1 shows the means and standard deviations for pre-belief and post-belief. For pre-belief, there was a significant difference in accuracy perception between rumor and filler tweets. The accuracy perception for rumor tweets was lower than filler tweets, whereas there were no significant differences in importance between rumor and filler tweets.

Table 1. Pre- and post-beliefs for target tweets

Target Tweet	Accuracy		Importance	
	Pre-belief	Post-belief	Pre-belief	Post-belief
Rumor	3.14 (0.47)	2.67 (0.50)	3.52 (0.49)	3.23 (0.57)
Filler	3.53 (0.44)	3.46 (0.46)	3.56 (0.47)	3.57 (0.51)

Note. The numbers in parentheses are standard deviations.

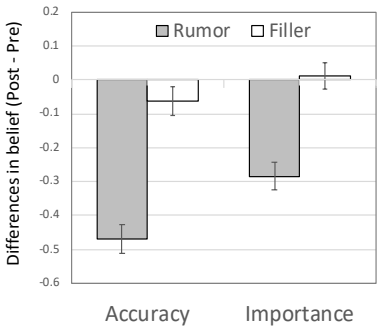


Figure 2. Means and standard errors for belief change after message exposure. Participants were exposed to inconsistent message for rumor and to consistent message for filler.

To test the effects of inconsistent message exposure, belief change after inconsistent message exposure was compared with the responses after the filler consistent message exposure. An analysis of variance (ANOVA) test with message type (inconsistent vs. consistent) was

conducted on belief change in accuracy of target tweets. The main effect of message type was significant, F (1, 45) = 24.28, $p < .001$, $\eta^2_G = .22$. The belief change in accuracy for rumor tweets exposure was bigger (M = -0.47, SD = 0.98) than that for filler tweets (M = -0.06, SD = 0.99) (Figure 2). We also performed a one-way ANOVA with message type on belief change in importance. Result showed a significant effect of message type, F (1, 45) = 12.26, $p < .005$, $\eta^2_G = .11$. The belief change in importance for rumor tweets exposure was bigger (M = -0.29, SD = 1.00) than that for filler tweets (M = 0.01, SD = 0.94) (Figure 2).

To examine eye movement, we performed a one-way ANOVA with message type on fixation duration rate on target tweets and messages, respectively. There were no significant differences in fixation duration rates between rumor and filler target tweets. However, the result of a one-way ANOVA on fixation duration rates on the message region showed that the main effect of message type was significant, F (1, 45) = 38.77, $p < .001$, $\eta^2_G = .04$. The fixation duration per character on inconsistent message (M = 85.76, SD = 64.14) was longer than consistent message (M = 71.30, SD = 50.40)(Figure 3, left figure). The same pattern emerged for fixation frequency rate. There was no significant difference between rumor and filler target tweets, whereas the result of a one-way ANOVA revealed a significant main effect of message type, F (1, 45) = 33.67, $p < .001$, $\eta^2_G = .04$. The fixation frequency rate on inconsistent message (M = 0.34, SD = 0.21) was higher than consistent message (M = 0.29, SD = 0.18) (Figure 3, right figure).

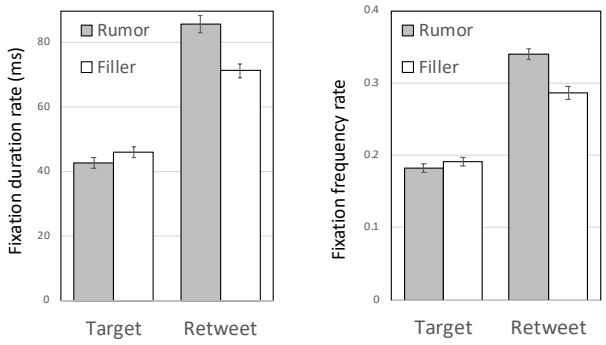


Figure 3. Means and standard errors of fixation duration rate and fixation frequency rate (per character).

To examine the relationship between eye-tracking data and belief, we used 'lme4' package (Bates, Maechler, Bolker, & Walker, 2015) in R (R Core Team, 2012). As fixed effects, we entered tweet type (inconsistent vs. consistent), pre-belief on accuracy and importance into the model. As random effects, subjects and multiple stimuli for each tweet type were added into the model.

We constructed a generalized linear mixed model (GLMM) of fixation duration rate. The inconsistent message affected fixation duration rate ($\chi^2(1) = 31.36$, $p < .001$), increasing it by about 14.5 ms ± 2.56 (standard errors)[95% CI: 9.44, 19.48]. Pre-belief and post-belief in a target tweet were not related to fixation duration rate on the message. As for fixation frequency rate, a GLMM revealed the same pattern. The inconsistent message affected it ($\chi^2(1) = 34.15$, $p < .001$), increasing it by about 0.05 ± 0.01 (standard errors) [95% CI: 0.04, 0.07]. There were no other factors related to fixation frequency rate.

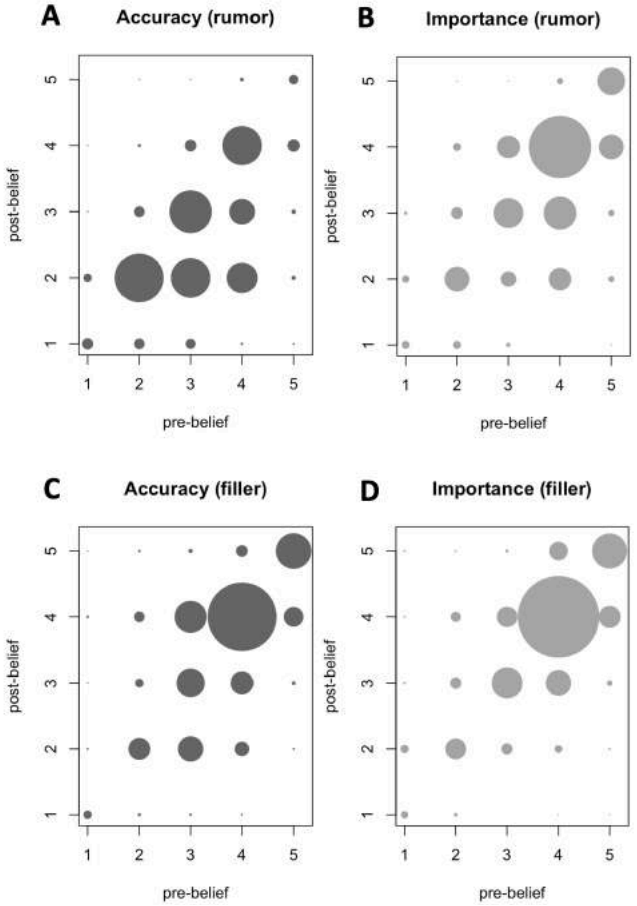


Figure 4. Belief change after exposure to inconsistent message associated with rumor tweets. The circle size represents the number of data points. Belief changes on rumor tweets in accuracy (A) and importance (B) after inconsistent message exposure. Belief changes on filler tweets in accuracy (C) and importance (D).

Figure 4 shows the distributions of the relationship between pre-belief and post-belief in rumors and fillers. Comparing with pre-belief in fillers that tended to be stable after consistent message exposure (Fig.C and D), pre-belief in rumors tended to be weakened after inconsistent message exposure (Figure A and B). However, the distributions of rumor tweets include not a few responses that did not change their belief and kept believing rumors after inconsistent messages.We examined predictor of the

difference in responses after inconsistent message exposure. As our focus here was especially on the belief change in the responses which had positive pre-belief (rated 3, 4, or 5) for rumor tweets, other responses which had negative pre-belief (rated 1 or 2) were excluded from the following analysis. This is an attempt to avoid a floor effect. That is, there is very little or no room to show a decrease in belief after inconsistent message exposure for the responses which had negative belief towards the bottom of the graph. Among all 552 responses associated with rumors (46 participants × 12 rumor tweets), 396 (71.7%) and 439 (79.5%) responses were analyzed in terms of accuracy and importance, respectively.

We constructed a GLMM to predict belief change (pre – post) in accuracy on the rumor with the responses with positive pre-belief. Fixation duration rate on inconsistent message increased belief change of accuracy positively ($\chi^2(1) = 4.91$, $p = .03$), that is, strengthening pre-belief of accuracy by about 0.002± 0.001 (standard errors) [95% CI: 0.0002, 0.003]. Fixation frequency rate on inconsistent tweet also affected belief change in accuracy ($\chi^2(1) = 3.85$, $p = .05$), strengthening pre-belief of accuracy by about 0.47± 0.23 (standard errors) [95% CI: 0.005, 0.92]. A GLMM to predict belief change of importance for rumor tweet revealed no significant relationship between eye movements and belief change.

Discussion

The present study investigated the effects of inconsistent message exposure on belief change in rumor and relationship between the belief change and eye movement.

First, we examined whether inconsistent message exposure changes the belief in rumor target message. Results showed that inconsistent message exposure tends to reduce pre-belief associated with rumors. Both perceived accuracy and importance associated with rumors were significantly reduced after inconsistent message exposure, whereas pre-belief of filler target did not significantly change after consistent message exposure. Thus, Hypothesis 1 was supported. These results support previous findings on the exposure of readers to denial messages being helpful to mitigate their false belief in rumors (Bordia et al., 2005; Bordia, DiFonzo, & Schulz, 2000; Koller, 1993).

Eye tracking data demonstrated that participants paid more visual attention to inconsistent messages associated with rumors than consistent message associated with fillers. Both fixation duration rate and fixation frequency rate associated with inconsistent messages were higher than consistent messages. This result provides support for Hypothesis 2, eliminating the possibility that participants skipped reading an inconsistent message. This result indicates that the inconsistency of the message attracted visual attention. These results provide support for the literature indicating that fixation duration becomes longer when text becomes more unpredictable (Rayner et al., 2011) and when it includes inconsistency (Rayner, Chace, Slattery, & Ashby, 2006). When people encounter inconsistent message, they need to consider the relationship between pre-belief and the inconsistent message and to update the pre-belief if needed. This cognitive procedure could result in longer visual attention. Taken together with a rumor study which demonstrated that people tend to spread false rumors because of novelty (Vosoughi, Roy, Aral, 2018), one explanation is that inconsistent messages were unpredictable and novel, thus, resulting in a relatively decrease in the novelty of rumors. This explanation is corroborated by the result that there were no significant differences in visual attention to the target tweets between rumor and filler. Participants have read the target tweets prior to eye measurement, that is, both types of target tweets were predictable. This prior experience resulted in no significant differences in eye movements between rumor and filler.

Next, we focused on the variation in the effect of inconsistent message exposure. Although the exposure to inconsistent message tends to devalue the accuracy and importance of rumor, the distribution of the relationship between pre- and post-belief in rumor showed that some responses showed a continued belief in rumors even after the exposure to inconsistent messages. Further examination focusing on the belief change of the responses with positive pre-belief in rumor tweets demonstrated that the belief change of accuracy was predicted by eye movement. Longer fixation duration and higher fixation frequency on inconsistent message predicted that the accuracy of rumors would be strengthened. These results can be interpreted in line with the previous findings (Espino, Santamaria, & Garcia-Madruga, 2000; Masson, 1983; Rayner et al., 2006): that the difficulty of text can lead a longer reading duration. Our findings indicate that the effect of inconsistent message exposure became limited for the participants having positive pre-belief in a rumor when they did not fully comprehend the inconsistent message.

There are some limitations in the current study. This study did not measure the level of comprehension of inconsistent messages. It is unclear whether longer fixation was related directly to low comprehension. Additionally, fixation predicted belief change in accuracy but it was not related to belief change in importance. Further research is needed to clarify these relationships.

In conclusion, the current study demonstrated the overall effect of exposure to inconsistent messages to reduce false belief in rumors, supporting previous research on rumor control. Our findings demonstrated the relationship between eye movement and belief change after inconsistent message exposure. The effectiveness of inconsistent message exposure was limited when the inconsistent message was difficult to process, resulting in as slightly strengthened pre-belief.

Acknowledgments

This research was supported by JSPS KAKENHI Grant Number 18K12010. The authors thank the anonymous reviewers for their insightful comments and suggestions.

References

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme. *Journal of Statistical Software*, *67*(1), 1–48.

Bode, L., Vraga, E. K., & Troller-Renfree, S. (2017). Skipping politics: Measuring avoidance of political content in social media. *Research and Politics*, *4*, 1–7.

Bordia, P., DiFonzo, N., Haines, H., & Chaseling, E. (2005). Rumors denials as persuasive messages: Effects of personal relevance, source, and message characteristics. *Journal of Applied Social Psychology*, *35*(6), 1301–1331.

Bordia, P., DiFonzo, N., & Schulz, C. A. (2000). Source characteristics in denying rumors of organizational closure: Honesty is the best policy. *Journal of Applied Social Psychology*, *30*(11), 2309–2321.

Difonzo, N., & Bordia, P. (2000). How top PR Professionals handle hearsay: Corporate rumors, their effects, and strategies to manage them. *Public Relations Review*, *26*(2), 173–190.

DiFonzo, N., & Bordia, P. (2007). Rumor Psychology: Social and Organizational Approaches. Washington, DC: American Psychological Association.

Ehrlich, S. F., & Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior*, *20*(6), 641–655.

Espino, O., Santamaria, C., & Garcia-Madruga, J. A. (2000). Figure and difficulty in syllogistic reasoning. *Cahiers de Psychologie Cognitive/Current Psychology of Cognition*, *19*(4), 417–428.

Iyer, E. S., & Debevec, K. (1991). Origin of rumor and tone of message in rumor quelling strategies. *Psychology and Marketing*, *8*(3), 161–175.

Koller, M. (1993). Rebutting accusations: When does it work, when does it fail? *European Journal of Social Psychology*, *23*(4), 373–389.

Lilienfeld, S. O., Lynn, S. J., Ruscio, J., & Beyerstein, B. L. (2010). *50 Great myths of popular psychology: Shattering widespread misconceptions about human behavior*. Oxford, UK: Wiley-Blackwell.

Masson, M. E. J. (1983). Conceptual processing of text during skimming and rapid sequential reading. *Memory & Cognition, 11*(3), 262-274.

Mendoza, M., & Poblete, B. (2010). Twitter under crisis: Can we trust what we RT? *Proceedings of the First Workshop on Social Media Analytics* (pp.71–79).

R Core Team. (2012). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. Vienna, Austria.

Rayner, K. (1998). Eye movements in reading and Information Processing : 20 Years of Research. *Psychological Bulletin*, *124*(3), 372–422.

Rayner, K., Chace, K. H., Slattery, T. J., & Ashby, J. (2006). Eye movements as reflections of comprehension processes in reading. *Scientific Studies of Reading*, *10*(3), 241–255.

Rayner, K., Juhasz, B. J., & Yan, G. (2005). The effect of word predictability on the eye movements of Chinese readers. *Psychonomic Bulletin & Review*, *12*(6), 1089–1093.

Rayner, K., Pollatsek, A., Ashby, J., & Clifton, C. (2012). *Psychology of reading*. New York, NY: Psychology Press.

Rayner, K., Slattery, T. J., Drieghe, D., & Liversedge, S. P. (2011). Eye movements and word skipping during reading: effects of word length and predictability. *Journal of Experimental Psychology. Human Perception and Performance*, *37*(2), 514–28.

Rosnow, R. L. (1991). Inside rumor: A personal journey. *American Psychologist*, *46*(5), 484–496.

Starbird, K., Spiro, E., Edwards, I., Zhou, K., Maddock, J., & Narasimhan, S. (2016). Could this be true?: I think so! Expressed uncertainty in online rumoring. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems,* 360–371.

Tanaka, Y., Sakamoto, Y., & Matsuka, T. (2013). Toward a social-technological system that inactivates false rumors through the critical thinking of crowds. *Proceedings of the 46th Hawaii International Conference on System Sciences,* 649–658.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. Science, 359, 1146–1151.

Appendix

Example of stimulus set (translated from Japanese into English). The numbers in brackets under each rumor or filler are means and standard deviations for pre- and post-belief of accuracy.

Rumor #1 The brain weighs approximately 1.2-1.5 kg, but in actual fact, most of us use only 10% of that. [M_{pre} = 3.41 (0.96), M_{post} = 2.85 (1.19)]

Inconsistent message for Rumor #1 When even a part of the brain is damaged by an accident or illness, it affects physical exercise, perception, language, and thought. These kinds of effects would be strange if 90% of the brain was not being used.

Rumor #2 People can be broadly divided into left-brained and right-brained, where left-brained people are logical and

analytical, while right-brained people are artistic. [M_{pre} = 3.30 (1.05), M_{post} = 3.07 (1.10)]

Inconsistent message for Rumor #2 Language is necessary for both logical and artistic activities, but both sides of the brain are working when language is used. The left side of the brain is superior in producing grammar and vocabulary, and the right side of the brain, responsible for intonation.

Rumor #3 Because one's personality appears in their handwriting, experts can understand a person's character by judging their letters and lines. [M_{pre} = 3.35 (1.06), M_{post} = 2.63 (1.01)]

Inconsistent message for Rumor #3 Handwritten application documents include information like one's work experience and criminal record in addition to just handwriting. When these indirect clues were regulated, handwriting analysts' predictive abilities were at coincidental levels.

Filler #1 Comics are funnier read when holding a pencil between your teeth so as not to touch your lips than when holding a pencil between puckered lips. [M_{pre} = 2.70 (1.11), M_{post} = 3.15 (1.23)]

Consistent message for Filler #1 When you hold a pencil in your teeth without touching your lips, your mouth spreads from side to side and makes an expression like when laughing. The laughing expression influences the way you read or perceive the comics.

Filler #2 When people repeatedly experience that their situation does not improve whether they work hard or resist, they learn the feeling of powerlessness. [M_{pre} = 3.96 (1.07), M_{post} = 3.96 (0.99)]

Consistent message for Filler #2 People must be tormented by feelings of powerlessness in companies where they are scolded "not to do whatever they feel like" if they independently think and take action, but scolded "not to be passive and to think for themselves" if they wait for instructions.

Filler #3 As in the case of things studied at home being easier to remember at home than in the classroom, circumstances influence memory. [M_{pre} = 4.07 (0.93), M_{post} = 3.80 (1.07)]

Consistent message for Filler #3 I have heard the same kind of thing about feelings — it is apparently easier to remember sad events when feeling sad and easier to remember happy events when feeling happy.

Predicting the Appreciation of Multimodal Advertisements

Serra Sinem Tekiroğlu (serrasinem@gmail.com)
Carlo Strapparava (strappa@fbk.eu)
Gözde Özbal (gozbalde@gmail.com)
FBK - irst, Trento, Italy

Abstract

Creativity is an essential factor in successful advertising where catchy and memorable media is produced to persuade the audience. The creative elements in the visual design and in the slogan of an advertisement elevate the overall appeal providing a perceptually grounded attractive message. In this study, we propose the exploitation of creativity cues in textual and visual information for the appreciation prediction of multimodal advertising prints. Moreover, as a novel dimension space of multimodality, we propose using the human sense (i.e., sight, hearing, taste, and smell) information embedded in the language. Our findings show that sensorial information is an invaluable indication of whether the advertisement is appreciated or not. Furthermore, combining linguistic and visual models significantly improves the unimodal appreciation detection performances.

Keywords: advertising creativity; human senses; multimodal creativity

Introduction

Creativity in advertising is an entangled, multi-dimensional phenomenon that reflects the complex structure of human creativity. A catchy and memorable advertisement is coherent and captivating. It is carefully designed with a diverse range of approaches including the ways of visualizing concepts, the use of rhetorical devices, such as exaggeration, paradox, metaphor and analogy, and taking advantage of shock tactics and humour (Pricken, 2008). In case of the advertising prints, visual and textual contents are designed to have a complementary and coordinated meaning. Advertising makes use of sensory and linguistic sensorial information heavily in order to reach the customers and persuade them. Elder and Krishna (2009) propose that multi-sensory ads induce higher taste perceptions than ads focusing on taste alone. They also state that using multiple senses in slogan increases the positive thought about the advertised food product. As a way of improving advertising communication, Percy (1982) suggests the use of concrete and high imageary words and concepts to stimulate better recall, better comprehension of the advertised message leading to an easier and more accurate understanding of the ad. Another creativity infusion strategy in ad production is using sensory words especially generating linguistic synaesthesia as an imagination boosting tool (Pricken, 2008). The slogans '*The taste of a paradise*' (Bounty bar commercial), where the sense of *sight* is combined with *taste*, and '*Hear the big picture*'(CBC Radio One commercial), where *sight* and *hearing* are merged, can be considered as the examples of linguistic synaesthesia.

As a topic being on the rise in computational linguistics, multimodality is mostly exploited by adding other modalities on top of the linguistic models to perceptually ground the current tasks. For instance, semantic representations benefit from the reinforcement of linguistic modality with visual (Bruni, Tran, & Baroni, 2014) and auditory (Kiela & Clark, 2015) modalities. In the same manner, we propose devising visual modality in collaboration with linguistic modality in the appreciation detection task. To our knowledge, this is the first study aiming to identify multimodal appreciation in a computational manner. Moreover, the multimodality of the dataset stands out amongst the others since the linguistic channel of an ad is complementary to the visual channel instead of being a scene description or an image label. This study focuses on the appreciation of the advertising print by the advertising professionals and communities instead of the appreciation by the audience/customer of the advertised product.

In this paper, we investigate the appreciation level of multimodal advertising prints focusing on the creativity cues in the slogan and in the corresponding image. We use a set of fundamentally creative artworks; an advertising dataset which is composed of 4265 images and corresponding slogans. The objective of this paper is twofold: i) to capture the potency of sensorial dimension of semantics as a creativity cue in the language along with various creative properties both in visual and linguistic modalities, ii) to develop a multimodal appreciation detection model. We utilize a random forest model trained on a dense feature set extracted from the slogans, ad categories and product types for the linguistic modality. For the visual modality, we employ a fine-tuned convolutional neural network model and a random forest model trained on the observable features of the images to determine whether the appreciated images display common visual characteristics and whether these characteristics have a distinctive effect on the overall appreciation of a multimodal ad.

Related Work

Considering that the essential focus of this study is computational creativity and multimodality, we summarize the most relevant studies conducted on these topics. Elgammal and Saleh (2015) quantify the creativity in paintings within the context of historical creativity where creative paintings adequately differ from the antecedent paintings and influence the subsequent. They present a computational framework that is

based on a creativity implication network.

Regarding the linguistic creativity, Özbal, Pighin, and Strapparava (2013) present a creative sentence generation framework, BRAINSUP, on which several semantic aspects of the output sentence can be calibrated. The syntactic information and a huge solution space are utilized to produce catchy, memorable and successful sentences. Kuznetsova, Chen, and Choi (2013) focus on identifying creativity in lexical compositions. They consider two computational strategies, first investigating the information theoretic measures and the connotation of words to find the correlates of perceived creativity and then employing supervised learning with distributional semantic vectors. Alnajjar, Kundi, Toivonen, et al. (2018) propose a methodology to automatically create slogans for a target concept and its adjectival property by first generating metaphors, based on a metaphor interpretation model. They produce a semantic space with the generated metaphors and use the semantic space to fill the slogan skeletons extracted from the existing slogans. They evaluate the slogans through crowd-sourcing with respect to the relatedness of the slogan to the concept and property, the correctness of the language, the metaphoricity, the catchiness, attractiveness and memorability, and the overall appropriation of the expression as a slogan.

Concerning the multimodality, Bruni, Boleda, Baroni, and Tran (2012) analyze the affect of different types of visual features such as SIFT and LAB on semantic relatedness task, and present a comparison of unimodal and multimodal models. Sartori et al. (2015) experiment on a complementary multimodal dataset similar to ours. They explore the influence of the metadata (i.e., titles, description and artists statement) of an abstract painting for the computational sentiment detection task. For the combination of modalities, they propose a novel joint flexible Schatten p-norm model exploiting the common patterns shared across visual and textual information. Shutova, Kiela, and Maillard (2016) exploit visual modality to improve the metaphor detection performance while Zadeh, Chen, Poria, Cambria, and Morency (2017) apply multimodal input to sentiment analysis.

Creativity in Advertising Prints

The creativity elements and dimensions in advertising have been investigated thoroughly. Ang and Low (2000) explore the influence of dimensions of creativity such as novelty (expectancy), meaningfulness (relevancy), and emotion (valence of feelings) to the effectiveness of the advertisement. While novelty could be identified as the unexpectedness and out-of-box degree of an advertisement, meaningfulness is the relevancy of the advertisement to the message aimed to be conveyed. The third dimension, emotional content, focuses on the feelings awakened in the audience. These three dimensions should manifest themselves in a creative advertising media. Smith, MacKenzie, Yang, Buchholz, and Darley (2007), on the other hand, elaborate on the divergence, which is the encapsulation of novel, different, or unusual elements, in ads proposing that the most significant characteristic of creative ads is their divergence. In addition to the above-mentioned general dimensions of advertising creativity, we specifically focus on the sensorial elements and their effect in the objective creativity level. Sensorial language makes use of multiple senses to induce higher taste perceptions (Elder & Krishna, 2009). Multiple senses in the advertising text trigger the positive thinking in the audience. Using highly imageable, in other words highly sensory words, helps to convey the advertised message better and easily. Finally, linguistic synaesthetia is a specific but a very significant way of effective advertising. Furthermore, for visual creativity in advertising, we focus on capturing the divergence factors through a transfer learning mechanism built on top of a deep learning image classification model and artistic values by taking advantage of the observable visual features in the image.

Multimodal Advertising Creativity Dataset

To investigate the appreciation of a multimodal advertising print, we first need to identify a dataset that reflects relatively upper and lower levels of appreciation from human subjects. To this respect, we chose AdsOfTheWorld[1], which has a wide range of coverage of ads considering its characteristic of being a social network that aims to inspire the advertising professionals. The members of the website can share their advertisement artwork, rate and discuss the ads created by others. The published advertising prints are diverse in terms of the level of creativity and ratings such that some ads are award-winning while some are highly disfavored by the community. We collected the ad images, their slogans and meta-data from AdsOfTheWorld[2]. The meta-data of an ad includes the average user rating, which is an integer within the range from 1 to 10, the number of raters, brand name and category.

While constituting the appreciated and unappreciated classes, namely *AP* and *UNAP*, we considered the opposite endings of the rate scale to distinguish the appreciation levels of advertisements as much as possible. We also paid regard to the number of instances that we obtained after filtering in order to have sufficient data for generalization of the *AP* and *UNAP* classes for training a classifier. In order to avoid feeding noise to our models, we empirically determined a minimum number of votes to postulate an average rating as reliable. To this end,we incorporated an ad into our final dataset if it is voted by at least 20 users and if it has an average rating in the range from 1 to 4 for *UNAP* class, or in the range from 7 to 10 for *AP* class. Finally, we eliminated the improper image styles, such as photographs of the billboards or images containing only textual content, from the dataset so that we can guarantee each image and its respective slogan contribute to the targeted message. From the final dataset that contains 4265 images-slogans, we sampled 3265 instances for training, 100 instances for development and 900 instances for testing. While the development and test sets are perfectly balanced for both classes, the training set includes

[1] http://adsoftheworld.com

[2] AdsOfTheWorld has recently changed its interface, no more showing the user ratings.

1470 *UNAP* and 1795 *AP* instances. During the sampling, we paid attention to putting the ads from the same brand into the same set since a slogan for a brand can be paired with various visual designs leading to more than one instance with the same linguistic input. As an additional meta-data, we collected the type of the products since category labels are considerably high-level. For instance, the category *House, Garden* includes a great variety of product types, such as furniture, laundry detergent, or insect killer. Utilizing product type and category labels as reference points allows us to appraise the meaningfulness of a slogan which affects its appreciation level notably. Advertisement 1[3] and Advertisement 2[4] exemplify highly appreciated and highly unappreciated samples in the final dataset, respectively. Although these advertising images seem to be very similar at the first glance with an object in the middle of the frame and in front of a blurred background, the subtle and creative details in the pictures, such as perplexing design of an octopus and sailboat made of pages of a book in Advertising 1 aims to immediately draw the attention of the audience. It holds an average rating of 10, which is the highest appreciation score, and is rated by 23 users. On the other hand, the obvious irrelevance of the main object in the image to the advertised message is a sign of an unappreciated design. Advertisement 2 with the slogan "Can't sleep?" promotes a tea brand that helps with sleeping problems by using a clearly irrelevant main object in the image. It has an average rating of 1 and its unappreciated label is trusted considering that it is rated by 171 users.

Appreciation Prediction Experiments

We design the appreciation prediction experiment of multimodal advertising prints exploiting the creativity dimension cues in the slogan and in the corresponding image, considering the studies done on the dimensions of creativity (Smith et al., 2007; Ang & Low, 2000; Elder & Krishna, 2009). To be more precise, we intend to capture surprisal, novel, meaningful, emotional, unusual and perceptual properties in an advertising slogan. Moreover, we aim to extract artistic components in the visual elements along with the latent visual descriptions and patterns.

Appreciation Detection on Slogans

For the textual model, we hypothesize that the creativity elements in the ad slogan can be mapped to features that are useful to detect the appreciation of an advertisement.

Surprisal (Self Information), as contributing to novelty/expectancy (Ang & Low, 2000) and surprisal(Smith et al., 2007) dimensions of creativity, can be interpreted as the information load of a specific outcome of an event. We calculate the self-information s of a bigram B by $s(B) = -log(p(B))$ exploiting the conditional probability distribution of bigram model trained on the corpus. We obtain the slogan self information as the average s of the bigrams extracted from the sentence.

Domain Relatedness features for the slogan are generated to address the meaningfulness (relevance) dimension (Ang & Low, 2000) of creativity. We expect that a meaningful slogan could contain words that are mapped to the same semantic domain with the product type and product category. On the other hand, a surprising effect could be achieved by injecting words from different domains. The ads dataset contains 24 categories, such as fashion or food. We obtain the domain information for each category as a noun, from WordNet Domains (Magnini, Strapparava, Pezzulo, & Gliozzo, 2002). Similarly, for each lemma-POS in the slogan and for the product type, we collected the related domains. The categories, product types and lemma-POS pairs are associated with the first sense from WordNet. In addition, we exploit a smaller set of domains that is constructed by normalization with respect to the middle level of WordNet hierarchy. The normalization of the domains provides a higher level of abstraction (Ozbal, Strapparava, Tekiroğlu, & Pighin, 2016) and could allow us to capture whether indirect concepts or ideas are employed for expressing the targeted message.

Semantic Similarity features also allow us to capture the meaningfulness dimension (Ang & Low, 2000) of ad creativity. We exploit ad category and the product type to calculate similarity scores with respect to the lemmas in the sentence. We employ 300 dimensional word representation vectors from GloVe (Pennington, Socher, & Manning, 2014) pretrained embeddings trained on Wikipedia 2014 articles and English GigaWord 5 (LDC). The similarity scores between a category/product type and a lemma are obtained by calculating the cosine similarity of their embedding vectors. The average score of a slogan is encoded as a real valued feature for category and another for product type.

Emotion as a creativity dimension focuses on the feelings awakened in the audience(Ang & Low, 2000). We also generated the emotion features as suggested by Özbal et al. (2013).

Sentiment scores are estimated and used as a part of the emotion(Ang & Low, 2000) dimension features. A word with a highly negative or positive sentiment can induce a positive or negative feeling and might alter the effectiveness and appreciation of the sentence. For instance, an environmental awareness slogan would intend to evoke negative sentiment intensifying the feeling of danger in order to be more striking. Thus, we determined the highest values of positive and negative sentiments in the slogan by checking each lemma-POS and encode them as real valued features. We use the sentiment scores of SentiWordNet (Esuli & Sebastiani, 2007).

Unusual Words contribute to the creativity as the unusual elements dimension suggested by Smith et al. (2007), also as a suprisal factor. We generated unusual words features following the study by Özbal et al. (2013).

Variety can be mapped to the flexibility dimension (Smith et al., 2007). We employ variety scores to detect whether creative and appreciated language displays a particularly different word variety than a less-creative and unappreciated lan-

[3] http://www.adsoftheworld.com/media/print/anagram_sea

[4] http://www.adsoftheworld.com/media/print/gryphon_slippers

guage in a similar way with Özbal et al. (2013).

Phonetic scores can be considered as contributing to the artistic value dimension (Elgammal & Saleh, 2015; Smith et al., 2007; Fichner-Rathus, 2011). The exploitation of phonetic features in creative and persuasive sentence analysis has been deeply explored by Özbal et al. (2013). Following them, we explore the alliteration, rhyme and plosive scores generated by using the HLT Phonetic Scorer[5].

Sensorial features are created regarding the sensory dimension of ad creativity (Elder & Krishna, 2009). A slogan aims to trigger a sensory activation in the mind of the audience. For instance, to evoke the sense of taste for an ad in the food category, certain sensorial information, such as the 'warmness' of a soup or the 'sweet aroma' of a cake, should be transmitted through the language. To identify the sensorial load of the sentences, we obtain the word-sense associations from Sensicon (Tekiroğlu, Özbal, & Strapparava, 2014) and Voted Norms, which we generated as a new set of sensory modality association norms through a voting mechanism and labeling the words with the senses that receive the majority of the votes from 4 different sensorial lexicons (Lievers & Winter, 2017; Tekiroğlu et al., 2014; Lynott & Connell, 2009, 2013; Winter, 2016). Sensicon embodies 22,684 and Voted Norms Lexicon includes 3890 English lemmas together with their part-of-speech (POS) information that have been linked to one or more of the five senses. For each sensory modality, we encode the average sensorial associations of the lemma-pos tuples in the slogan. In addition, we explore how the sensorial trait of a product interacts with the sensorial information in the slogan. Therefore, we create a binary feature indicating whether the sensorial modalities with the highest value of the product type and the slogan are identical. We also add the sensorial association relation of the product type and the slogan as a feature set by taking the mean of the slogan associations and product associations with respect to Sensicon and the Voted Norms. As another hypothesis, we expect that sensory experience ratings (Juhasz, Yap, Dicke, Taylor, & Gullick, 2011) can provide a second channel of sensorial information since *SER* resource estimates the sensory experience triggered in human mind instead of the sensorial information that one word carries. We extracted sensory experience ratings by averaging the SER values of the words in the slogan. Based on the category and sensorial modality correlations provided by Tekiroğlu et al. (2014), we propose a set of sensorial features encoding whether the sensorial information in the slogan conforms to the predetermined sensorial structure of its category.

We generated category conformity scores utilizing Voted Norms. For each sense, we set a binary flag indicating if the average association value of the slogan and the sensorial value of the category are both positive. As an example, the feature set of an ad from the food category contains the binary features *conforms_taste=1* and *conforms_hearing=1* if the average sensorial association value of the slogan for these modalities are over 0.0. In addition, we checked the sensorial association peak of the slogan which shows the modality of the highest sensorial association among the lemma-POS tuples in the slogan. We created a binary feature if the peak modality of the slogan and its category are identical. Contrary to our hypothetical assumption, the peak sensorial conformity is observed to be an indicator of a lower level of appreciation (Mann-Whitney $p < 0.001$) in the training set. A possible explanation for this can be that the unexpected sensorial elevation contributes to the appreciation level of a slogan and a less-appreciated slogan is associated to the senses in a more conventional manner. For instance, a less-appreciated toothpaste slogan "For brighter smiles"[6], which is from the *health category*, has the sensorial peak conformity since both the sensorial peak, i.e. *brighter*, and the ad category are associated with the sense of sight. Using an overly well-known effect of the product to describe a stereotypical metonymic replacement, i.e. *brighter smile* for *whiter teeth*, might be one of the causes of a lower level of appreciation of the ad.

Model	# Feat	Training F1	Testing F1
L	62	0.573	0.577
$L \setminus Sensorial$	34	$*$0.542	$\star$0.496
$L \setminus V \cup U \cup SI$	62	0.585	0.558
$L \setminus Similarity$	61	0.573	0.560
$L \setminus Domain$	55	0.580	$*$0.548
$L \setminus Phonetic$	59	0.573	0.561
$L \setminus Emotion \cup Sentiment$	45	0.568	$^{\#}$0.552

Table 1: The linguistic modality ablation study results. $\star$ denotes $p < 0.001$, $*$ denotes $p < 0.01$, $^{\#}$ denotes $p < 0.05$ for the McNemar significance test between L and ablated models.

Linguistic Experiment Results

We investigated the performances of linguistic features with a classification task employing Random Forest algorithm implemented within the *scikit-learn* package. To fine tune the hyper-parameters of the classifier, we perform a grid search over the number of the generated trees (between 100 and 500, with a step size of 100), the maximum depth of the tree (as [5, 10, 20]) using 10-fold cross validation on the training data. To guarantee the same slogan being only in the training folds or only in the validation fold, we divided the training set into 10 folds by taking into consideration the brand information. Since the training data is unbalanced, we selected the best model by using the weighted average of F1 values.

The results of the full model with all the implemented features and the ablation study are summarized in Table 1. The first row labeled 'L' shows the micro F1 scores for the cross validation and test phases using all the linguistic features. Each row in the rest of the table shows the ablation of the indicated feature. We marked statistical significance in terms of the drop of the performance during ablation in comparison to all features L according to McNemar's test.

In the linguistic experiment, we found out that all the fea-

[5] hlt-nlp.fbk.eu/technologies/hlt-phonetic-scorer

[6] http://www.adsoftheworld.com/media/print/colgate_hide_and_seek

tures contribute to the performance of the final linguistic model even if they cause a slight increase in the F1 scores. By utilizing all the features, we obtain an average training cross-validation F1 score of 0.573 and testing F1 score of 0.577. The linguistic model without *Sensorial Information* yields an F1 test score of 0.496 that is significantly lower ($p < 0.001$) than the model *L*. We obtain a significantly lower ($p < 0.01$) F1 score of 0.548 on the absence of *Domain* features in the linguistic feature set. Removing the *Emotion* and *Sentiment* features from the model decreases the score down to 0.552 on the test set causing a statistically significant loss of performance ($p < 0.05$). The contributions of the strongest features point out that the relevance of a slogan to the product category and type, the positive or the negative feeling that a slogan induces and most importantly the sensorial structure of the slogan and its sensory impact in the audience are indeed essential for a creative and *AP* slogan which is in line with the creativity dimension analysis.

Appreciation Detection on Images

The message of an advertising print is conveyed through both linguistic and visual channels. In this experiment,we utilize the raw sensory input in the form of embedded representations of the image and visual surface features.

Transfer Learning (CNN) Deep learning approaches are proven to be successful in multimodality tasks yielding the state of the art performances on Computer Vision studies such as image classification (Krizhevsky et al., 2012) or object detection (Ren et al., 2015). Considering the promising strength of the convolutional neural networks in image recognition, we hypothesize that certain characteristics of an image, such as objects and patterns, can tamper with its appreciation level as a creative artwork. For instance, marketing images mostly encode cultural and historical stereotypes of masculinity and femininity in order to invoke the feeling of gender identity in the customers (Schroeder & Zwick, 2004). We conjecture that such patterns can be utilized to predict the appreciation level of an ad image if we can capture them automatically.

Since our dataset is not large enough to train a deep network from scratch, we employ transfer learning where we fine-tune Inception V3 image recognition model(Szegedy et al., 2016) as an appreciation predictor. Inception V3 is a deep convolutional neural network that significantly improves the state of the art ILSVRC 2012 1000-class ImageNet classification benchmark. It is trained using stochastic gradient on Tensorflow. Although, the object classification on ImageNet and the appreciation classification task on carefully designed ads are fundamentally dissimilar, Yosinski et al. (2014) state that transferring features from a distant task is still better than randomly initialized variables. Therefore, it would be feasible to boost a network by transferring deeply trained features to overcome the scarcity of the advertising data.

While conducting a transfer learning on Inception V3, first, we only retrained the top 2 layers, labeled as Inception-V3/Logits and Inception-V3/AuxLogits, and kept the earlier layers frozen. In this phase, we obtained a checkpoint after 1000 steps. We exploited Tensorflow Slim[7] implementation to train the new layers and we set the *learning rate* to 0.01. The last layer of the network is a softmax layer that provides posterior probabilities as normalized prediction values for *AP* and *UNAP* classes. In the second phase, we fine-tune all trainable weights in the whole network only for 500 steps and with a learning rate of 0.001. We keep the learning rate small in order to protect the powerful weights of the original Inception V3 model from changing too quickly and loosing their representation ability.

Observable Visual Features (OVF) Together with the implicit properties and patterns belonging to *AP* and *UNAP* classes, we also utilize the explicit elements such as lines and their properties found in the images. In addition, we seek the impact of the color information per se in the creativity detection by encoding the dominant colors as another feature. These features are mostly related to the artistic dimension of the creativity. We extract top 10 dominant colors in the images by k-means clustering on the color values of the pixels and map the center values of the clusters to 16 colors. In addition, we extracted lines in an image through Hough Transform (Duda & Hart, 1972). This feature set contains the normalized length of the longest line and average line length; 3 binary flags indicating whether the longest line is horizontal, vertical or diagonal. We also encoded an interpretation of the "Rule of Thirds", which is a well-known rule of photographic composition[8] and mainly states that the center of interest in images should be on the intersection points or along the lines when an image is divided into 9 equal sections by 2 horizontal and 2 vertical lines. We generated a binary feature indicating if the longest line in the image starts from an outer area and crosses over only 2 sections horizontally and/or 2 sections vertically. Our intuition is that cutting the continuum of the line close to "Rule of Thirds" interest areas can guide the eye of the viewer to the center of focus and contribute to the message and aesthetic value of the image.

Visual Modality Experiment Results

We trained the CNN models on 3265 images from the training set. Using the trained models, we performed the testing on 900 images from the test set. At the end of the test phase, we obtained *AP* and *UNAP* scores for each test image. We employed a straightforward decision process with a 0.5 cut-off where the labeling is conducted by finding the higher value among the *AP* and *UNAP* scores. The performances of the CNN models are summarized in Table 2. The validation accuracy shown in the table is calculated by evaluating the model over the randomly sampled 265 images as the validation set during the training phase. We observe that fine-tuning all the network after the retraining the last 2 layers provide a clear boost to the classifier performance of CNN-1K. The F1 score on the test set significantly ($p < 0.05$) increases from 0.561 to 0.596.

[7] http://github.com/tensorflow/models/tree/master/slim

[8] http://en.wikipedia.org/wiki/Rule_of_thirds

Visual Model	Validation Acc	Testing F1
CNN-1K	0.608	0.561
CNN-1K+500	0.626	$^{\#}$0.596

Table 2: The CNN visual modality experiment results. $^{\#}$ denotes $p < 0.05$ for the McNemar significance test between the models.

Model	#Features	Training F1	Testing F1
OVF	22	0.547	0.560
Colors	17	$\star$0.517	$*$.0.515
Lines	5	0.533	$^{\#}$ 0.508

Table 3: The OVF model visual modality experiment results. $\star$ denotes $p < 0.001$, $*$ denotes $p < 0.01$, $^{\#}$ denotes $p < 0.05$ for the McNemar significance test between *OVF* and Colors or Lines.

For the *OVF* model, we performed the same training strategy that we employed in the linguistic experiment. The model yields relatively poor training (F1:0.547) and testing (F1:0.560) results as they are shown in Table 3. During the ablation study, we observed that a significant performance change occurred when we removed the *Lines* (F1:0.517, $p < 0.001$) for the training cross-validation results. Through the analysis on testing results, we found out that both *Colors* (F1:0.515, $pval < 0.05$) and *Lines* (F1:0.508, $p < 0.01$) contribute to the final model significantly. Although the results of the whole *OVF* model suggest that these aesthetic features are indeed indicating factors of image creativity, we can imply that the overall visual appreciation of an ad is affected by more subtle properties to be discovered than the aesthetic features that we implemented.

Multimodal Fusion

We embraced the late fusion (score level) strategy (Kiela & Clark, 2015) to obtain the multimodal appreciation score. To combine the scores from each model for a class by soft voting approach, we employed Equation 1 where s_n denotes the appreciation score for the ad x by the model m_k and α_n denotes the weight for the model m_k.

$$ms(x) = \sum_{n:m_k} \alpha_n \times s_n \quad (1)$$

We obtained the peak points for α values by running a grid search on the multimodal fusion of model outputs for development set. In this search, all α values are positive and their sum equals to 1.0. After calculating the multimodal appreciation scores, namely *ms*, we labeled the instance *a* by finding the maximum value among the class scores. We evaluate the multimodal experiment results by averaging (Equal α) and soft voting in terms of the F1 scores on the test set and we present the results in Table 4. The α values that we employ during the fusion are shown in the last column. While calculating the multimodal fusion results, we employ the unimodal models; *L*, *OVF* and *CNN-1K+500*. We chose to use the model *CNN-1K+500* since it yields the highest validation accuracy and has a significant improvement for the testing in comparison to its predecessor. Regarding the uni-modal results of linguistic and visual models, the lowest performance is obtained by using OVF while the highest F1 score is yielded by the CNN model. As shown in Table 4, the *ALL* model significantly outperforms the linguistic and observable visual features models. *ALL* model surpasses *CNN*, which is the best unimodal model, by increasing the performance from 0.596 to 0.625 for the soft fusion and to 0.620 for the equal α fusion. This outcome can be considered as conforming with our initial anticipation that the different modalities play complementary roles in expressing the creativity and appeal of an advertising print. The highest contributor of the complete model *ALL* is the *CNN* model and when we remove it from the fusion, the Equal α F1 score drops to 0.587.

Model	Eq. α F1	Soft α F1	α values
ALL	0.620	0.625	L:0.18,C:0.24,O:0.58
$L \cup OVF$	0.587	$*$0.573	L:0.11, O:0.89
$L \cup CNN$	0.605	0.606	L:0.74, C:0.26
$OVF \cup CNN$	0.618	0.612	C:0.25, O:0.75
CNN	0.596	0.596	C:1.0
OVF	$*$ 0.560	$\star$0.560	O:1.0
L	$^{\#}$0.577	$^{\#}$0.577	L:1.0

Table 4: Multimodal fusion results and comparisons to the uni-modal experiments. $\star$ denotes $p < 0.001$, $*$ denotes $p < 0.01$, $^{\#}$ denotes $p < 0.05$ for the McNemar significance test between *L* and ablated models.

Discussion and Conclusion

For the example in `www.adsoftheworld.com/media/print/act_tv_numbers_insects_vs_frog`, which is an *AP* sample resolved by the model *ALL* but not by the visual models, although the visual channel is highly expressive too, the lack of straight lines and dull color palette decreases the prediction performance of *OVF* model. *CNN* model also mislabels the image with a very low confidence since it possibly fails to recognize the peculiar focus elements. Therefore, a wider range of training samples for advertising images would be necessary to identify the style marks of creative and appreciated compositions.

Our quantitative results show that the sensorial structure of the relation between the slogan, and the product category/type is a strong indicator of the creativity appreciation level. When we analyze the feature importance of the final model *L*, we detected that especially the sensorial relation features between the product type and slogan become prominent among the implemented sensorial features. To better illustrate the contribution of the sensorial information to the final model, we fused the visual models with the model $L \setminus Sensorial$ in which we removed all sensorial features from the linguistic model. The Equal-α F1 score of the fusion model decreases to 0.594 without the sensorial features. In the ad `www.adsoftheworld.com/media/print/febreze_french_fries`, we show an *AP* test sample resolved by the contribution of sensorial features. In fact, the example epitomizes the usage of the olfactory disadvantage of the language as a creativity inducing tool. We believe

that smell related words, such as the word "odor" in the example, possibly contributes to the surprisal dimension of the creative and *AP* advertising since olfactory words tend to be less expected by the audience. Indeed, our analysis on the training set reveals that the smell association of the words are inclined to be higher in the *AP* samples in comparison to the *UNAP* samples (Mann-Whitney $p < 0.001$). On the other hand, *taste* association tends to denote the opposite behaviour in our training set (Mann-Whitney $p < 0.001$) while we cannot observe any significant difference for the other senses w.r.t. the Sensicon association values.

Although the automatic assessment of the appreciation level of advertising is a substantially compelling challenge, our findings suggest that sensorial information along with the other linguistic, semantic, cognitive and, finally visual aspects establish a starting point to tackle its complexity.

References

Alnajjar, K., Kundi, H., Toivonen, H., et al. (2018). Talent, skill and support. In *Proceedings of the ninth international conference on computational creativity, salamanca, spain, june 25-29, 2018.*

Ang, S. H., & Low, S. Y. (2000). Exploring the dimensions of ad creativity. *Psychology & Marketing*, *17*(10), 835–854.

Bruni, E., Boleda, G., Baroni, M., & Tran, N.-K. (2012). Distributional semantics in technicolor. In *Proceedings of ACL 2012*.

Bruni, E., Tran, N.-K., & Baroni, M. (2014). Multimodal distributional semantics. *J. Artif. Intell. Res.(JAIR)*, *49*.

Duda, R. O., & Hart, P. E. (1972). Use of the hough transformation to detect lines and curves in pictures. *Communications of the ACM*, *15*(1), 11–15.

Elder, R. S., & Krishna, A. (2009). The effects of advertising copy on sensory thoughts and perceived taste. *Journal of consumer research*, *36*(5), 748–756.

Elgammal, A., & Saleh, B. (2015). Quantifying creativity in art networks. In *Proceedings of ICCC 2015.*

Esuli, A., & Sebastiani, F. (2007). Sentiwordnet: A high-coverage lexical resource for opinion mining. *Evaluation*.

Fichner-Rathus, L. (2011). *Foundations of art and design: An enhanced media edition*. Cengage Learning.

Juhasz, B. J., Yap, M. J., Dicke, J., Taylor, S. C., & Gullick, M. M. (2011). Tangible words are recognized faster: The grounding of meaning in sensory and perceptual systems. *The Quarterly Journal of Experimental Psychology*, *64*(9).

Kiela, D., & Clark, S. (2015). Multi-and cross-modal semantics beyond vision: Grounding in auditory perception. In *Proceedings of EMNLP 2015.*

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural inf. processing systems.*

Kuznetsova, P., Chen, J., & Choi, Y. (2013). Understanding and quantifying creativity in lexical composition. In *Proceedings of EMNLP 2013.*

Lievers, F. S., & Winter, B. (2017). Sensory language across lexical categories. *Lingua*.

Lynott, D., & Connell, L. (2009). Modality exclusivity norms for 423 object properties. *Behavior Res. Methods*, *41*(2).

Lynott, D., & Connell, L. (2013). Modality exclusivity norms for 400 nouns: The relationship between perceptual experience and surface word form. *Behav. Res. Methods*, *45*(2).

Magnini, B., Strapparava, C., Pezzulo, G., & Gliozzo, A. (2002). The Role of Domain Information in Word Sense Disambiguation. *Natural Language Engineering*, *8*(4).

Özbal, G., Pighin, D., & Strapparava, C. (2013). Brainsup: Brainstorming support for creative sentence generation. In *Proceedings of ACL 2013.*

Ozbal, G., Strapparava, C., Tekiroğlu, S. S., & Pighin, D. (2016). Learning to identify metaphors from a corpus of proverbs. In *Proceedings of EMNLP 2016.*

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of EMNLP 2014.*

Percy, L. (1982). Psycholinguistic guidelines for advertising copy. *ACR North American Advances*.

Pricken, M. (2008). *Creative advertising ideas and techniques in the world's best campaigns*. Thames&Hudson.

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In *Adv. in neural inf. processing systems.*

Sartori, A., Yan, Y., Ozbal, G., Salah, A., Salah, A., & Sebe, N. (2015). Looking at Mondrian's victory boogie-woogie: What do I feel? In *Proceedings of IJCAI 2015.*

Schroeder, J. E., & Zwick, D. (2004). Mirrors of masculinity: Representation and identity in advertising images. *Consumption Markets & Culture*, *7*(1), 21–52.

Shutova, E., Kiela, D., & Maillard, J. (2016). Black holes and white rabbits: Metaphor identification with visual features.

Smith, R. E., MacKenzie, S. B., Yang, X., Buchholz, L. M., & Darley, W. K. (2007). Modeling the determinants and effects of creativity in advertising. *Marketing science*, *26*(6).

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the CVPR 2016.*

Tekiroğlu, S. S., Özbal, G., & Strapparava, C. (2014). Sensicon: An automatically constructed sensorial lexicon. In *Proceedings of EMNLP 2014.* Doha, Qatar.

Winter, B. (2016). *The sensory structure of the english lexicon*. Unpublished doctoral dissertation, University of California.

Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks? In *Advances in neural information processing systems.*

Zadeh, A., Chen, M., Poria, S., Cambria, E., & Morency, L.-P. (2017). Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250.*

Speaking but not Gesturing Predicts Motion Event Memory Within and Across Languages

Marlijn ter Bekke (Marlijn.terBekke@mpi.nl)
Radboud University, Nijmegen, The Netherlands
Wundtlaan 1, 6525XD Nijmegen, The Netherlands

Aslı Özyürek (Asli.Ozyurek@mpi.nl)
Center for Language Studies & Donders Center for Cognition, Radboud University, Nijmegen, The Netherlands
Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
Wundtlaan 1, 6525XD Nijmegen, The Netherlands

Ercenur Ünal (Ercenur.Unal@ozyegin.edu.tr)
Center for Language Studies, Radboud University, Nijmegen, The Netherlands
Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
Özyeğin University, Istanbul, Turkey
Nişantepe Mahallesi Orman Sokak 34794 Çekmeköy, Istanbul, Turkey

Abstract

In everyday life, people see, describe and remember motion events. We tested whether the type of motion event information (path or manner) encoded in speech and gesture predicts which information is remembered and if this varies across speakers of typologically different languages. We focus on intransitive motion events (e.g., a woman running to a tree) that are described differently in speech and co-speech gesture across languages, based on how these languages typologically encode manner and path information (Kita & Özyürek, 2003; Talmy, 1985). Speakers of Dutch (n = 19) and Turkish (n = 22) watched and described motion events. With a surprise (i.e. unexpected) recognition memory task, memory for manner and path components of these events was measured. Neither Dutch nor Turkish speakers' memory for manner went above chance levels. However, we found a positive relation between path speech and path change detection: participants who described the path during encoding were more accurate at detecting changes to the path of an event during the memory task. In addition, the relation between path speech and path memory changed with native language: for Dutch speakers encoding path in speech was related to improved path memory, but for Turkish speakers no such relation existed. For both languages, co-speech gesture did not predict memory speakers. We discuss the implications of these findings for our understanding of the relations between speech, gesture, type of encoding in language and memory.

Keywords: Motion events; Memory; Cross-linguistic differences; Co-speech gesture

Introduction

People frequently perceive, remember and communicate about events. The relations between these different cognitive processes are not well-understood. In this study, we ask whether the way a visually perceived event is described relates to how it is remembered. How exactly an event is described, varies across typologically different languages. In addition, within languages there is also variation: two speakers of the same language may perceive the same event, but describe it differently. Importantly, in describing events people not only use speech but also co-speech gestures that describe main components of events. These gestures also vary both across and within languages. How does the way one speaks and gestures about events predict one's memory for various aspects of events?

Many of the events people see in their daily lives involve motion, because the world around us is constantly moving. Two crucial components of motion events are the manner of motion (e.g., running) and the path that the motion follows (e.g., to the tree). Whether people mention the manner or path during a motion event description is strongly affected by the language they speak. Verb-framed languages (e.g., Turkish, Greek, Spanish) typically encode path in the main verb and can optionally add manner of motion, for example in subordinate verbs or in adverbial phrases (see example sentence (1) from Turkish below; Talmy, 2000). By contrast, satellite-framed languages (e.g., Dutch, English, Russian) typically encode manner in the main verb and path in a variety of other structures, such as prepositional phrases (see example sentence (2) from Dutch below). A crucial difference between verb-framed and satellite-framed languages is that speakers of satellite-framed languages typically mention both path and manner information, while speakers of verb-framed languages regularly omit manner information (Slobin, 2003).

(1)

Kadın	*(koş-arak)*	*ağac-a*	*yaklaş-ıyor*
Woman	(run-Connective)	tree-Dative	approach-Present
Noun phrase	(Verb)	Noun phrase	Verb
Figure	(Manner)	Ground	Path

(2)

De vrouw	*rent*	*naar*	*de boom*
The woman	runs	to	the tree
Noun phrase	Verb	Preposition	Noun phrase
Figure	Manner	Path	Ground

If speakers of different languages describe the same motion event differently, do they also remember the event differently? Prior work found no cross-linguistic differences in how speakers of verb-framed and satellite-framed languages remember manner and path (Engemann et al., 2015; Gennari et al., 2002; Papafragou et al., 2002; Papafragou, Hulbert, & Trueswell, 2008, but see Filipović, 2011 for differences using complex motion events). However, these studies simply compared speakers of verb-framed and satellite-framed languages at the group level, without considering the variation within languages in terms of which motion event information is described. It remains unknown whether which information a speaker mentions in a particular motion event description may predict their later memory for that information, regardless of their native language. For example, if a speaker described the path of a motion event, do they remember that path better? In addition, how these specific descriptions might interact with native language to predict memory also remains unclear. For example, does describing path have a different effect on path memory for speakers of verb-framed languages compared to speakers of satellite-framed languages?

It is plausible that the information encoded in linguistic descriptions predicts memory performance for two reasons. First, it could be that the description is a window into the mental representation of the event: if a speaker describes the path, this might indicate that the speaker has mentally represented the path of the event. Therefore, the speaker may be more likely to remember the path (Papafragou et al., 2002). Second, it could be that the verbal description functions as an additional format in which the event is encoded in memory. This way, the description itself might be remembered and thus aid memory for the components encoded in the description (Papafragou et al., 2002). Indeed, it appears that what exactly is said in a motion event description is important for memory: speakers who described a path of motion later remembered this path better (Billman, Swilley, & Krych, 2000).

When investigating the link between descriptions and memory, it is important to keep in mind that language is multimodal (Vigliocco, Perniss, & Vinson, 2014). In fact, descriptions of events are often accompanied by iconic co-speech gestures. For example, while saying "The woman ran to the tree", a speaker might wiggle one's index and middle fingers in an inverted V-shape across space from left to right. Co-speech gestures can represent path, manner, or both in one gesture (Figure 1). Importantly, co-speech gestures accompanying motion event descriptions differ both across and within languages. In terms of cross-linguistic differences, the *form* of motion event co-speech gestures differs between speakers of verb-framed and satellite-framed languages (Kita & Özyürek, 2003). However, it is yet unknown whether there are cross-linguistic differences between speakers of verb-framed and satellite-framed languages in terms of *how often* they gesture about path and manner, and whether this relates to their memory for path and manner. In addition, co-speech gesture production also differs within languages. Within speakers of a language, one element of motion might be gestured more often than another element for different events. Therefore, both speech and co-speech gesture need to be taken into account to see how differences within and across languages in motion event descriptions relate to motion event memory.

Figure 1: Gestures can represent only path (A), only manner (B) or both manner and path (C)

Indeed, prior work shows that gestures are related to event memory. For example, producing co-speech gestures when describing motion and action events leads to better memory for these events (Cook, Yip, & Goldin-Meadow, 2010). In addition, the specific action event information conveyed in gesture predicts the information later remembered (Koranda & MacDonald, 2015). These results are in line with research on the enactment effect, which shows that reading descriptions of action events and performing these actions leads to better memory for the descriptions that does only reading (for review, see e.g., Cohen, 1989). The involvement of the motor system could lead to richer memory representations, or to stronger memory representations (Madan & Singhal, 2012). These studies point to the importance of taking co-speech gestures into account when investigating the relation between motion event descriptions and memory.

The Present Study

The main aim of the present study was to investigate whether the speech and co-speech gestures that speakers use to describe motion events predict their memory, and whether cross-linguistic differences in speech and gesture lead to cross-linguistic differences in memory. To test these questions, Dutch and Turkish speakers watched and described motion events, after which their surprise

recognition memory for manner and path was tested. We had the following predictions:

(a) In general, we expected that encoding a motion event component in speech would predict better memory for that component.
(b) Similarly, we expected that encoding a motion event component in gesture would predict better memory for that component.
(c) Cross-linguistically, we expected Dutch speakers to encode manner more often in speech and gesture than Turkish speakers, due to the optional encoding of manner in Turkish. As a result, we expected Dutch speakers to have better memory for manner.

Method

Participants. Data were collected from 19 adult native speakers of Dutch (15 females, M_{age} = 23) and 22 adult native speakers of Turkish (16 females, M_{age} = 21). Dutch speakers received monetary compensation for their participation. Turkish speakers received course credit for their participation.

Materials. Target events presented in the study phase consisted of 16 silent video clips that depicted a female actor moving with respect to a landmark object along a particular path with a particular manner (e.g., a woman hopped to a cactus). Each clip was 2500ms long. Each clip was created by combining four spontaneous manners of motion (run, hop, twirl, tiptoe) with four motion paths (to, into, from, out of). Sixteen additional video clips of transitive events served as fillers (e.g., a woman biting an apple).

In the memory phase, half of the events had a change to either the manner (e.g., a woman tiptoed instead of hopped to a cactus) or the path (e.g., a woman hopped from instead of to a cactus) of motion (Figure 2). The other half of the events remained the same. Of the 15 filler events, half remained the same and half involved an object change (e.g., a woman biting a banana).

Procedure. Each participant was tested in a quiet room at their university campus in their native language by a native speaker together with a confederate who served as an addressee.

In the study phase, participants saw 16 target and 16 filler events. Each trial started with a fixation screen of 1000ms, followed by the event shown for 2500ms. Then a gray screen appeared, during which participants described "what happened in the video" to the addressee. Participants' speech and gestures were videotaped for later coding. The memory task was presented immediately after the study phase. The memory task was a surprise for the participants, because this way the prospect of the memory task could not affect the production results. During the memory task, participants saw another set of events and for each event indicated whether they had seen this exact video before by pressing a button. In both study and memory phases, each participant saw the events in different randomized order.

Figure 2: Example of a manner change (hop became tiptoe; left panel) and a path change (to became from; right panel)

Coding. Descriptions of target events were coded for the presence of path and manner information in speech and gesture using ELAN software (Lausberg & Sloetjes, 2009) by a native speaker of the relevant language. In speech, manner information was coded as present if how the motion was performed was encoded with a manner verb (e.g., rennen; running – mostly in Dutch) or a manner verb subordinated to a path verb via a connective (e.g., koşarak; run-Connective – mostly in Turkish). Path information was coded as present if the change of location with respect to something was encoded with prepositions or spatial/directional nouns (e.g., naar (to), içine (inside)) or path verbs (e.g., gir (enter), yaklaş (approach)).

In gesture, manner information was coded as present if speakers produced a gesture representing the motion in a non-linear way. Gestures could represent the manner from a third person perspective (e.g., for twirling, a manner gesture could involve the index finger turning in circles) or could be an enactment of the figure's posture during the movement (e.g., for running, a manner gesture could involve moving the arms up and down). Path information was coded as present if speakers deliberately traced the change of location with a body part chosen to represent the figure. Path gestures could trace the change of location in the lateral axis (either with a correct or incorrect direction) or in the sagittal axis (moving towards or away from the body). Points to the location of the landmark were not coded as path gestures. Gestures could either include one motion element (manner-only or path-only) or a combination of both elements.

Results

Data were analyzed with generalized binomial linear mixed effects modelling (glmer) with crossed random intercepts for Subjects and Items using *lme4* package (Bates et al., 2015) in *R* (R Core Team, 2018). This mixed effects approach allowed us to take into account the random variability that is due to having different participants and different items.

Speech and gesture production

First, we tested whether there were cross-linguistic differences in how often path and manner components of motion events were mentioned in speech (Figure 3). We excluded three trials (two Dutch) in which the addressee talked and affected the speaker's speech production. A glmer model that tested the effects of Language (Turkish, Dutch) and Component (Path, Manner) on binary values for mention in speech (0 = no, 1 = yes) at the item level revealed only a main effect of Component (β = 3.41, *SE* = 1.50, *z* = 2.28, *p* = .02). Speakers mentioned Manner (*M* = 0.97) more often than Path (*M* = 0.72). No other effects or interactions were significant. Furthermore, the proportion of mention of path and manner components in speech by Turkish and Dutch speakers were similar per specific types of path or manner (Table 1).

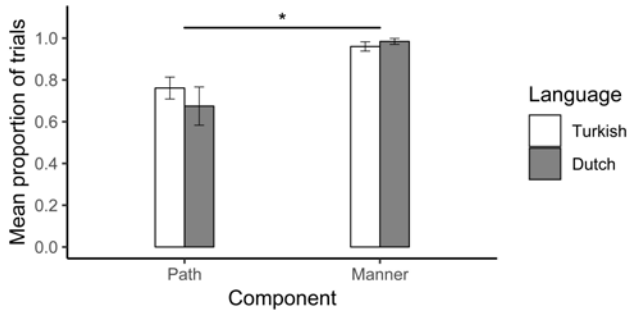


Figure 3: Event components encoded in speech. Error bars represent the standard error around the mean proportion of trials in which a component is mentioned per participant.

Table 1: Proportions of event components encoded in speech for each manner and path type, separated by language.

		Language	
		Turkish	Dutch
Manner	hop	0.99	0.98
	run	0.99	1.00
	tiptoe	0.91	1.00
	twirl	0.95	0.96
Path	to	0.67	0.56
	from	0.61	0.51
	into	0.95	0.90
	out of	0.80	0.73

Next, we tested whether there were cross-linguistic differences in how often speakers gestured about path and manner components while describing motion events (Figure 4). We excluded the same three trials that were excluded from the speech data analyses. A glmer model that tested the effects of Language (Turkish, Dutch), Component (Path, Manner) and their interaction on binary values for whether a component was encoded in gesture in an event description (0 = no, 1 = yes) revealed only a main effect of Language (β = -1.59, *SE* = 0.49, *z* = -3.24, *p* < .01). Turkish speakers (*M* = 0.48) gestured more often about both elements than Dutch speakers (*M* = 0.28). No other effects or interactions were significant. These patterns were replicated in a follow-up analysis that selected only the trials in which speakers gestured, and thus eliminated the possibility that differences in gesture rates hide cross-linguistic differences in what speakers of Dutch and Turkish prefer to gesture about.

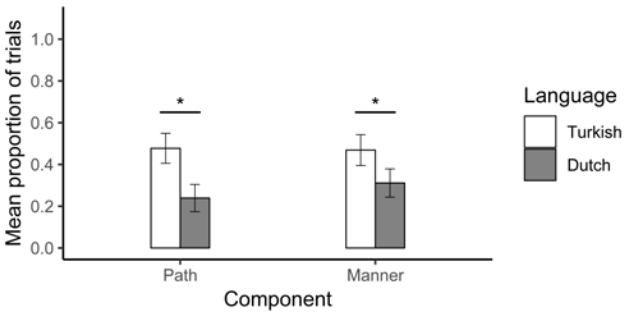


Figure 4: Event components encoded in gesture. Error bars represent the standard error around the mean proportion of trials in which a component is gestured per participant.

Memory performance

Beginning with filler events, Dutch (*M* = 0.99) and Turkish (*M* = 0.95) speakers had similar memory accuracy, indicating that the language groups were comparable in general memory performance. Furthermore, collapsed across language groups, memory for No change items (*M* = 0.78, *SD* = 0.15, *t*(40) = 11.98, *p* < 0.001) and Path changes (*M* = 0.68, *SD* = 0.26, *t*(40) = 4.29, *p* < 0.001) were significantly higher than chance level. However, memory for Manner changes (*M* = 0.40, *SD* = 0.26, *t*(40) = -2.39, *p* = 0.99) did not differ from chance level. This suggests that the participants may have simply been guessing when there was a Manner change. In addition, looking at the distribution of Manner change detection accuracy, it was clear that almost all participants had poor manner memory. It was thus not the case that some participants' memory was very poor, while other participants' memory was good. Therefore, we did not further attempt to predict manner memory using speech, gesture and language, because we did not want to predict guessing behavior.

For predicting path memory, path mentions in speech that only used unspecific verbs (e.g., to advance) that do not indicate or imply the spatial relation between the figure and the landmark were analyzed together with no mention trials and were contrasted to path mentions with prepositions, spatial/directional nouns or path verbs. Because these unspecific path verbs could be used regardless of the trajectory of motion we reasoned that they would not aid

memory. Following a similar reasoning for gestures, we analyzed path gestures in the sagittal axis together with no gesture trials and contrasted them to path gestures in the lateral axis with the correct direction. Path gestures in the lateral axis with the incorrect direction were excluded from the analyses because they might even hinder memory.

A glmer model tested the effects of Path in speech (0 = no mention, 1 = mention), Path in gesture (0 = no gesture, 1 = path gesture), Language (Turkish, Dutch) and Condition (No change, Path change) on binary values for whether an item was remembered (0 = no, 1 = yes). The best-fitting model revealed a main effect of Condition as well as an interaction between Condition and Path in speech (β = 1.26, SE = 0.51, z = 2.46, p = .01): for No change items, speakers had similar accuracy regardless of whether Path was mentioned in speech; for Path changes accuracy was higher if Path was mentioned in speech than if it was not. There was also an interaction between Path in speech and Language (β = 1.57, SE = 0.58, z = 2.71, p < .01): Dutch and Turkish speakers had similar accuracy when they did not mention Path in speech, but Dutch speakers had higher accuracy than Turkish speakers when they mentioned Path in speech (Figure 5). No other main effects or interactions were significant. Notably, there were no effects or interactions involving the factor Path in gesture. Thus, contrary to our expectations, gesturing about path did not predict better memory for path of motion. We turn to the significance of these findings below.

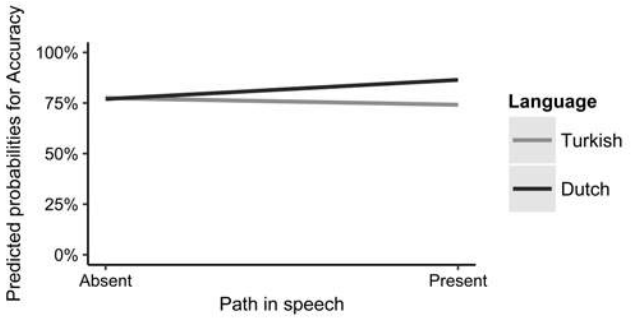


Figure 5: Interaction between Path in speech and Language for path memory accuracy, as predicted by the glmer model

Discussion

We tested whether the speech and gesture used to describe a particular motion event predicts memory for that motion event, looking at variation across and within languages. Our study has five key findings. First, speakers of Turkish did not omit the manner more often than speakers of Dutch. Second, speakers of both Dutch and Turkish had chance level memory for manner of motion. Third, speakers who mentioned path in their speech were later more accurate at detecting changes to this path. Fourth, path mention in speech was positively related to path memory for Dutch speakers, but not for Turkish speakers. Finally, we found that speaking but not gesturing predicts memory for path information.

Regarding the production results, we did not replicate the classic typological finding that speakers of verb-framed languages omit the manner more often than speakers of satellite-framed languages (Slobin, 2003). Instead, we found that speakers of both Dutch and Turkish almost always mentioned the manner of motion. A possible explanation can be found in the stimuli used in the present study. In an attempt to increase manner memory, we used manners that were rather salient (tiptoe, twirl, hop, run). It is plausible that because these manners were so salient, speakers of Turkish deemed it important to mention them. This interpretation is in accordance with the finding that speakers of Greek, a verb-framed language, mention the manner of motion much more often when it is not inferable for the listener compared to when it is inferable (Papafragou, Massey, & Gleitman, 2006). Although the cross-linguistic difference in manner omission has been reported many times, our findings show that within-language encoding flexibility makes it possible that under certain conditions (e.g., for some events), such cross-linguistic differences can be diminished.

Our study was the first to directly compare memory for manners and paths, where the path and manner changes did not involve object changes, but manner and path changes for intransitive events (unlike e.g., Bunger, Trueswell, & Papafragou, 2012 investigating instrumental motion). The finding that path is remembered better than manner is in accordance with a previously reported developmental path bias in terms of categorization (Konishi et al, 2016; Pruden et al., 2012, 2013). It is possible that path is remembered better than manner because it is related more to intentionality or goal-directedness of the motion (Pourcel, 2004). Such a relation between intentionality and memory of motion is also found when comparing memory for goal paths (e.g., to) versus source paths (e.g., from). Goals are remembered better than sources, possibly because they are more informative about the figure's intentions (Lakusta & Landau, 2012; Papafragou, 2010). Notably, this goal-source asymmetry exists only for animate figures, who can have intentions (Lakusta & Landau, 2012).

Interestingly, while speakers were not successful at remembering the manner, they did almost always describe the manner. This dissociation indicates that in terms of manner, there is no strong correspondence between speech and memory. However, this overall comparison is based on data that is averaged across different participants, items, and languages. It is still possible that when these factors are taken into account, one might find a subtle relation between mentioning manner in speech and remembering manner. In future research, this can be tested if manner memory accuracy is increased to above chance level. Nevertheless, this overall dissociation between manner mention in speech and manner memory is still quite striking. This suggests that there are at least partly different criteria for which motion event information is important to describe to another person and for which motion event information is important to remember.

For describing motion events to another person, manner of motion may be important when it is salient and not inferable. By contrast, for remembering motion events, path of motion may be important because it relates to the intentions of the figure.

In terms of the relation between descriptions and memory, we found that speakers who described a path in speech were more accurate at detecting changes to that path. This is consistent with a previous finding that speaking about path predicts better memory for path (Billman et al., 2000). It is also consistent with prior findings from other domains, demonstrating relations between how speakers describe and remember visual stimuli (e.g., eye-witness memory, Marsh, Tversky, & Hutson, 2005; picture recognition, Zormpa et al., 2018). Whether this relation between path speech and path change detection is causal is a question for further research.

In addition, the relation between path speech and path memory differed cross-linguistically: for Dutch speakers only, speaking about path predicted better memory for path. For Turkish speakers, path memory was similar regardless of whether path had been mentioned. This result might be due to cross-linguistic differences in how path was mentioned. For example, while Dutch speakers mentioned path in prepositions, Turkish speakers mentioned path mainly in verbs. Perhaps these are differentially related to memory. Another cross-linguistic encoding difference is that while Dutch speakers almost always used path prepositions that indicate the spatial relation between the figure and the landmark, Turkish speakers sometimes used unspecific verbs (e.g., to advance) to describe the path. Thus, if a Turkish speaker wants to mention path, specifically mentioning the relation to the landmark is optional. This greater optionality may have resulted in a weaker link between linguistically encoding the relation to the landmark in speech and remembering it. Further research is necessary to investigate these speculations. Either way, this interaction indicates that when linking typological differences to cognition, it is important to move from studying main effects of native language to investigating more subtle interactions of native language and descriptions.

Finally, we found no relation between co-speech gesture and memory. Importantly, path gestures typically co-occur with path speech. Therefore, this lack of a relation between gesture and memory can be interpreted to mean that path memory is equally accurate for speakers who speak and gesture about path, compared to speakers who only speak about path. The lack of a relation between path gesture and memory was surprising, given that previous research has shown a link between gesture production and event memory (Cook et al., 2010; Koranda & MacDonald, 2015). However, these studies differ from ours in one important respect: while we used motion events only, they either collapsed motion events with actions (Cook et al., 2010) or used actions only (Koranda & MacDonald, 2015). Perhaps the different memory results can be attributed to the differences between iconic co-speech gestures that describe actions versus gestures that describe paths of motion events. For example, action gestures might involve motor simulation more strongly than tracing path gestures (Hostetter & Alibali, 2008).

Either way, it appears that for the path of motion events, speech but not co-speech gesture predicts memory. There are two potential explanations of this finding. One possibility is that speech planning affects attention more than does co-speech gesture planning. Speech planning affects attention: while watching motion events to prepare for description, people look at the events in such a way that they can describe it later (Bunger et al., 2012; Flecken et al., 2015; Flecken, von Stutterheim, & Carroll, 2014; Papafragou et al., 2008; Trueswell & Papafragou, 2010). By contrast, because gestures do not follow such a strict system, their planning might have less of an impact on attention, and in turn have less of an impact on memory. Another possible explanation for why speech but not gesture predicts memory concerns the nature of speech and gesture representations. While speech is categorical and relies on discrete units, gesture is analogue and allows information to be conveyed imagistically (Cook, Yip, & Goldin-Meadow, 2012). Therefore, the verbal representation is an easier, more simplified version of the real event, compared to the gestural representation, and thus might be more useful as a memory cue.

In conclusion, the present study reveals differential contributions of speech and gesture in predicting motion event memory. Our findings underline that the relation between language and event memory is intricate and is influenced by subtle variations in how motion events are described within and across speakers of different languages.

Acknowledgments

We acknowledge support from an NWO-VICI grant awarded to A.Ö.

References

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Billman, D., Swilley, A., & Krych, M. (2000). Path and manner priming: Verb production and event recognition. *Proceedings of the 22nd Annual Conference of the Cognitive Science Society* (pp. 615-620). Hillsdale, NJ: Lawrence Erlbaum Associates.

Bunger, A., Trueswell, J. C., & Papafragou, A. (2012). The relation between event apprehension and utterance formulation in children: Evidence from linguistic omissions. *Cognition*, *122*(2), 135-149.

Cohen, R. L. (1989). Memory for action events: The power of enactment. *Educational Psychology Review*, *1*(1), 57-80.

Cook, S. W., Yip, T. K., & Goldin-Meadow, S. (2010). Gesturing makes memories that last. *Journal of Memory and Language*, *63*(4), 465-475.

Engemann, H., Hendriks, H., Hickmann, M., Soroli, E., & Vincent, C. (2015). How language impacts memory of motion events in English and French. *Cognitive Processing*, *16*(1), 209-213.

Filipović, L. (2011). Speaking and remembering in one or two languages: Bilingual vs. monolingual lexicalization and memory for motion events. *International Journal of Bilingualism*, *15*(4), 466-485.

Flecken, M., Carroll, M., Weimar, K., & von Stutterheim, C. (2015). Driving along the road or heading for the village? Conceptual differences underlying motion event encoding in French, German, and French–German L2 users. *The Modern Language Journal*, *99*(S1), 100-122.

Flecken, M., von Stutterheim, C., & Carroll, M. (2014). Grammatical aspect influences motion event perception: Findings from a cross-linguistic non-verbal recognition task. *Language and Cognition*, *6*(1), 45-78.

Gennari, S. P., Sloman, S. A., Malt, B. C., & Fitch, W. T. (2002). Motion events in language and cognition. *Cognition*, *83*(1), 49-79.

Hostetter, A. B., & Alibali, M. W. (2008). Visible embodiment: Gestures as simulated action. *Psychonomic Bulletin & Review*, *15*(3), 495-514.

Kita, S., & Özyürek, A. (2003). What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: Evidence for an interface representation of spatial thinking and speaking. *Journal of Memory and Language*, *48*(1), 16-32.

Konishi, H., Pruden, S. M., Golinkoff, R. M., & Hirsh-Pasek, K. (2016). Categorization of dynamic realistic motion events: Infants form categories of path before manner. *Journal of Experimental Child Psychology*, *152*, 54-70.

Koranda, M., & MacDonald, M. C. (2015). Language and gesture descriptions affect memory: A nonverbal overshadowing effect. *Proceedings of the 37th Annual Meeting of the Cognitive Science Society* (pp. 1183-1188). Austin, TX: Cognitive Science Society.

Lakusta, L., & Landau, B. (2012). Language and memory for motion events: Origins of the asymmetry between source and goal paths. *Cognitive Science*, *36*(3), 517-544.

Lausberg, H., & Sloetjes, H. (2009). Coding gestural behavior with the NEUROGES-ELAN system. *Behavior Research Methods*, *41*(3), 841-849.

Madan, C. R., & Singhal, A. (2012). Using actions to enhance memory: Effects of enactment, gestures, and exercise on human memory. *Frontiers in Psychology*, *3*, 507.

Marsh, E. J., Tversky, B., & Hutson, M. (2005). How eyewitnesses talk about events: Implications for memory. *Applied Cognitive Psychology*, *19*(5), 531-544.

Papafragou, A. (2010). Source-goal asymmetries in motion representation: Implications for language production and comprehension. *Cognitive Science*, *34*(6), 1064-1092.

Papafragou, A., Hulbert, J., & Trueswell, J. (2008). Does language guide event perception? Evidence from eye movements. *Cognition*, *108*(1), 155-184.

Papafragou, A., Massey, C., & Gleitman, L. (2002). Shake, rattle, 'n' roll: The representation of motion in language and cognition. *Cognition*, *84*(2), 189-219.

Papafragou, A., Massey, C., & Gleitman, L. (2006). When English proposes what Greek presupposes: The cross-linguistic encoding of motion events. *Cognition*, *98*(3), 75-87.

Pourcel, S. (2004). What makes path of motion salient?. *Proceedings of the 30th Annual Meeting of the Berkeley Linguistics Society* (pp. 505-516). Ann Arbor, MI: Sheridan Books.

Pruden, S. M., Göksun, T., Roseberry, S., Hirsh-Pasek, K., & Golinkoff, R. M. (2012). Find your manners: How do infants detect the invariant manner of motion in dynamic events?. *Child Development*, *83*(3), 977-991.

Pruden, S. M., Roseberry, S., Göksun, T., Hirsh-Pasek, K., & Golinkoff, R. M. (2013). Infant categorization of path relations during dynamic events. *Child Development*, *84*(1), 331-345.

R Core Team. (2018). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundations for Statistical Computing. Available online at: https://www.R-project.org/

Slobin, D. I. (2003). Language and thought online: Cognitive consequences of linguistic relativity. In D. Gentner and S. Goldin-Meadow (Eds.), *Language in mind*. Cambridge, MA: MIT Press.

Talmy, L. (1985). Lexicalization patterns: Semantic structure in lexical forms. In T. Shopen (Ed.), *Language typology and syntactic description*. New York, NY: Cambridge University Press.

Talmy, L. (2000). *Toward a cognitive semantics*. Cambridge, MA: MIT Press.

Trueswell, J. C., & Papafragou, A. (2010). Perceiving and remembering events cross-linguistically: Evidence from dual-task paradigms. *Journal of Memory and Language*, *63*(1), 64-82.

Vigliocco, G., Perniss, P., & Vinson, D. (2014). Language as a multimodal phenomenon: Implications for language learning, processing and evolution. *Phil. Trans. R. Soc. B*, *369*(1651), 20130292.

Zormpa, E., Brehm, L., Hoedemaker, R. S., & Meyer, A. S. (2018). The production effect and the generation effect improve memory in picture naming. *Memory*, *27*(3), 340-352.

Sequential diagnostic reasoning with independent causes

Marko Tešić (mtesic02@mail.bbk.ac.uk) & Ulrike Hahn (u.hahn@bbk.ac.uk)

Department of Psychological Sciences
Birkbeck, University of London
Malet Street, London, WC1E 7HX, UK

Abstract

In real world contexts of reasoning about evidence, that evidence frequently arrives sequentially. Moreover, we often cannot anticipate in advance what kinds of evidence we will eventually encounter. This raises the question of what we do to our existing models when we encounter new variables to consider. The standard normative framework for probabilistic reasoning yields the same ultimate outcome whether multiple pieces of evidence are acquired in sequence or all at once, and it is insensitive to the order in which that evidence is acquired. This equivalence, however, holds only if all potential evidence is incorporated in a single model from the outset. Hence little is known about what happens when evidence sets are expanded incrementally. Here, we examine this contrast formally and report the results of the first study, to date, that examines how people navigate such expansions.

Keywords: sequential diagnostic reasoning; sequential causal structure learning; causal Bayesian networks; order effects.

Introduction

Tom wakes up one morning and notices a rash on his skin. He does not think the rash is a big deal, but after a couple of days the rash is still present so he decides to see a doctor. Before he visits a doctor he thinks that the rash is either caused by a bacterial or a viral infection or, perhaps, both. The doctor agrees with him that the rash could be caused by a bacterial and/or a viral infection. However, she additionally informs Tom that he also has a swelling he didn't notice, which can also be caused by a bacterial and/or a viral infection. Furthermore, she tells him that either type of infection is more likely to cause the swelling than the rash. How do (should) Tom and the doctor revise their beliefs about multiple independent causes given multiple pieces of evidence of different diagnosticity?

From a normative standpoint, many would argue that the answer is encoded in the causal Bayesian networks (CBNs): directed acyclic graphs with nodes representing variables (causes and effects) and arrows representing probabilistic and causal relations between the nodes (Pearl, 2009, 1988; Neapolitan, 2003). Here one would build a 4-node CBN with 2 common effects and 2 independent causes.[1] For instance, the CBN in Figure 1 would model the situation we described above: C_1 = viral infection, C_2 = bacterial infection, E_1 = rash, and E_2 = swelling.

To fully parameterize CBN from Figure 1, one needs to specify the following probabilities:

$$P_1(\mathrm{C}_1) = c_1 \quad , \quad P_1(\mathrm{C}_2) = c_2$$
$$P_1(\mathrm{E}_1 \mid \mathrm{C}_1, \mathrm{C}_2) = \alpha_1 \quad , \quad P_1(\mathrm{E}_1 \mid \mathrm{C}_1, \neg\mathrm{C}_2) = \beta_1$$

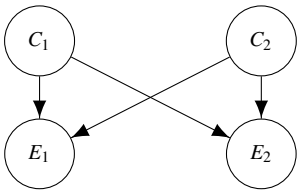


Figure 1: CBN with 2 independent causes and 2 common effects.

$$\begin{aligned} P_1(\mathrm{E}_1 \mid \neg\mathrm{C}_1, \mathrm{C}_2) = \gamma_1 \quad , \quad P_1(\mathrm{E}_1 \mid \neg\mathrm{C}_1, \neg\mathrm{C}_2) = \delta_1 \\ P_1(\mathrm{E}_2 \mid \mathrm{C}_1, \mathrm{C}_2) = \alpha_2 \quad , \quad P_1(\mathrm{E}_2 \mid \mathrm{C}_1, \neg\mathrm{C}_2) = \beta_2 \\ P_1(\mathrm{E}_2 \mid \neg\mathrm{C}_1, \mathrm{C}_2) = \gamma_2 \quad , \quad P_1(\mathrm{E}_2 \mid \neg\mathrm{C}_1, \neg\mathrm{C}_2) = \delta_2 \end{aligned} \tag{1}$$

$P_1(\mathrm{C}_1)$ and $P_1(\mathrm{C}_2)$ are usually referred to as the prior probability of the two causes and the remaining probabilities as being part of the conditional probabilities tables (CPTs) for the two effects. The doctor then could use this CBN to update her beliefs about the probability that Tom has a viral infection after learning that Tom has a rash by calculating $P_1(\mathrm{C}_1 \mid \mathrm{E}_1)$. After additionally learning that Tom also has swelling the doctor could further update her probability of Tom having a viral infection by calculating $P_1(\mathrm{C}_1 \mid \mathrm{E}_1, \mathrm{E}_2)$ (similarly for the bacterial infection).

However, it is somewhat accidental that Tom first noticed the rash and not the swelling. He could have plausibly first seen the swelling and gone to the doctor and then noticed the rash. Would the CBN calculation be different in this scenario? It depends. If the rash and the swelling are not equally diagnostic of the two causes as is suggested by the example, then it is possible that $P_1(\mathrm{C}_1 \mid \mathrm{E}_1) \neq P_1(\mathrm{C}_1 \mid \mathrm{E}_2)$, in which case the doctor's degrees of belief about a viral infection after first learning that Tom has swelling would not be equal to those where she first leaned about the rash. However, after learning the second effect the order in which the effect appear no longer matters: that is, $P_1(\mathrm{C}_1 \mid \mathrm{E}_1, \mathrm{E}_2)$ is always equal to $P_1(\mathrm{C}_1 \mid \mathrm{E}_2, \mathrm{E}_1)$.

It is then empirically interesting to investigate whether people are sensitive to these different orders of effects and whether they update the causes differently depending on the order in which the effects appear. Studies on sequential di-

[1] Hayes, Hawkins, Newell, Pasqualino, and Rehder (2014) have used a dynamic CBN to model these kinds of situations. However, in this paper we employ static CBNs as there are no significant differences in the formalism in this case.

agnostic reasoning have sought to tackle exactly these issues (see Meder & Mayrhofer, 2017b; Hogarth & Einhorn, 1992). They presented participants with a sequence of effects and asked them to reason from multiple effects to causes either with each effect they learned (step-by-step procedure) or after they learned about the whole sequence of effects (end-of-sequence procedure) (see Hogarth & Einhorn, 1992; Rebitschek, Bocklisch, Scholz, Krems, & Jahn, 2015). Their studies were primarily interested in investigating primacy effects (most of the evidential weight is given to the first piece of evidence) and recency effects (most of the evidential weight is given to the most recent pieces of evidence). Meder and Mayrhofer (2017a) investigated sequential diagnostic reasoning by providing participants with verbal information regarding the strengths of the causes instead of a more quantitative information (like the CPTs) and found that participants are remarkably accurate in their judgements. However, all these studies investigated only situations where the causes were mutually exclusive and exhaustive causes (which would be modeled as one node for all causes). Hayes et al. (2014) investigated a scenario where two symptoms could be produced by two independent causes. However, in their study both effects had exactly the same diagnosticity (i.e. the same CPT) and for that reason there are no order effects, i.e. it does not matter whether we learn first E_1 or E_2, $P_1(\mathrm{C}_1 \mid \mathrm{E}_1) = P_1(\mathrm{C}_1 \mid \mathrm{E}_2)$.

One of the goals of this paper is to empirically investigate people's ability to reason diagnostically from multiple effects with different diagnosticities (CPTs) to multiple independent causes. More specifically, we aim to test how people's judgements compare to the normative answer from CBNs such as the one in Figure 1 by manipulating the way in which multiple pieces of the evidence of different diagnosticity are presented (in a particular order or at the same time) and the way judgements about the causes are elicited from the participants (step-by-step (SbS) or all-at-once (AaO)).

Another interesting issue emerges when reasoning with independent causes. Not only can we learn the evidence sequentially, but we can sequentially learn about new variables that may influence our beliefs about the causes. In technical parlance, we may need to expand the algebra. Consider Tom from our example. Initially Tom only knew about his rash and, based on that knowledge, he updated his probabilities of the two causes. Unlike the doctor, Tom did not even know that the two types of infection could also cause swelling. It is only after he visited his doctor that he learned about the another potential effect and the occurrence of that effect. At the time he only knew about the rash he updated the probabilities of the two causes on the basis of a CBN model with only three nodes: two independent causes and one common effect while the doctor always had in mind the CBN from Figure 1. Despite operating with two different CBNs, both Tom and the doctor would arrive at the same probabilities (assuming the same priors and CPTs for the effect) at this first step. The next step is, however, crucial. After learning about the swelling, the doctor would simply learn the new piece of evidence and update the probabilities of the causes based on the CBNs from Figure 1. Tom, by contrast, might do one of two things: (1) forget about his original 3-node network and create a new 4-node one like the one in Figure 1 in which case he would arrive at the same estimates as the doctor; or (2) take his (and doctors) previous estimates of the two causes based on only one piece of evidence and take them as new priors in his new 3-node network with the second piece of evidence as a common effect (see Figure 2). In the latter case he would be 'splitting' the CBNs from Figure 1 into two 3-node networks.

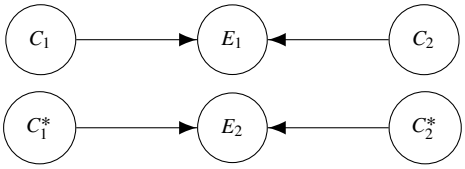


Figure 2: 'Split' CBN from E_1 to E_2

$$
\begin{aligned}
P_1(\mathrm{C}_1) = e_1 \quad &, \quad P_1(\mathrm{C}_2) = e_2 \\
P_2(\mathrm{C}_1^*) = P_1(\mathrm{C}_1 \mid \mathrm{E}_1) \quad &, \quad P_2(\mathrm{C}_2^*) = P_1(\mathrm{C}_2 \mid \mathrm{E}_1) \\
P_1(\mathrm{E}_1 \mid \mathrm{C}_1, \mathrm{C}_2) = \alpha_1 \quad &, \quad P_1(\mathrm{E}_1 \mid \mathrm{C}_1, \neg\mathrm{C}_2) = \beta_1 \\
P_1(\mathrm{E}_1 \mid \neg\mathrm{C}_1, \mathrm{C}_2) = \gamma_1 \quad &, \quad P_1(\mathrm{E}_1 \mid \neg\mathrm{C}_1, \neg\mathrm{C}_2) = \delta_1 \\
P_2(\mathrm{E}_2 \mid \mathrm{C}_1^*, \mathrm{C}_2^*) = \alpha_2 \quad &, \quad P_2(\mathrm{E}_2 \mid \mathrm{C}_1^*, \neg\mathrm{C}_2^*) = \beta_2 \\
P_2(\mathrm{E}_2 \mid \neg\mathrm{C}_1^*, \mathrm{C}_2^*) = \gamma_2 \quad &, \quad P_2(\mathrm{E}_2 \mid \neg\mathrm{C}_1^*, \neg\mathrm{C}_2^*) = \delta_2
\end{aligned}
\tag{2}
$$

Eq. (2) specify the priors and the CPTs of the two networks. Although one might intuitively think that Tom will arrive at the same probabilities as the doctor even in the case where he models the situation as in Figure 2, that turns out to be true only under very specific conditions, some of which may violate common assumptions in causal Bayesian reasoning (see Appendix A). Less technically, this is because once one learns evidence (E_1) and updates the probabilities of the two causes (C_1 and C_2) in a common-effect CBN, the two previously independent causes become dependent: although $P_1(\mathrm{C}_1 \mid \mathrm{C}_2) = P_1(\mathrm{C}_1)$, generally $P_1(\mathrm{C}_1 \mid \mathrm{C}_2, \mathrm{E}_1) \neq P_1(\mathrm{C}_1 \mid \mathrm{E}_1)$. This dependency is preserved in the full CBN network in Figure 1 even *before* one learns the second piece of evidence (E_2) and again updates the probabilities of the two causes. However, in the lower 3-node CBN in Figure 2 the dependency is lost since it is assumed that C_1^* and C_2^* are independent *before* observing E_2. Therefore, the final probability estimates of the two causes, i.e. their estimates *after* learning both pieces of evidence, will most likely diverge on the two different modeling strategies. More specifically, the final estimates of the two causes will always be higher according to the 'split' CBN in Figure 2 than those according to the full one in Figure 1 precisely because the full one accounts for the above-mentioned dependency and the 'split' one does not. Moreover, when the diagnosticity of the two pieces of evidence is different (as is the case in this study), the height of the final estimates in the 'split' CBN will depend on the order the evidence is observed: learning E_1 then E_2 will result

in the final estimates of the causes that are different to those that result from learning E_2 then E_1 (as previously mentioned, whether we learn E_1 first then E_2 or vice versa does not affect the final probability estimates of the causes in the full CBN). It is also worth pointing out that this divergence only happens when the causes are independent. If the causes are mutually exclusive and exhaustive, one can safely 'split' the full network into multiple ones without worrying about ending up with different estimates (see Appendix B).

To the best of our knowledge no study has yet investigated sequential diagnostic reasoning with sequentially learning the algebra. In the literature mentioned above participants were presented with all the variables and the causal/probabilistic information related to them before they started making judgements about the causes. Even in such contexts, it is worth looking at order effects because it has long been recognized that order effects may be particularly diagnostic with respect to the processes underlying the formation of a judgment. Specifically, there is a long literature concerned with order effects in contexts such as impression formation (Anderson, 1965) or numerical estimation (Jacowitz & Kahneman, 1995). However, our concerns in this paper go beyond this. We are interested in examining how reasoners fare in probabilistic reasoning contexts where they are faced with entirely new variables. This issue has, to the best of our knowledge, not been explored. In many scientific and everyday situations we must make judgements about potential causes given effects without being aware of other potential effects that could also inform our judgements. The main aim of this study was to examine how people reason with multiple pieces of evidence when they successively learn not just that some piece of evidence obtains, but also that there is another potential piece of evidence not known before. We compared participants' estimates to both the full network's predictions (Figure 1) and the 'split' networks' predictions (Figure 2).

Experiment overview

In the present experiment we investigated the influence of manipulating algebra and evidence learning on probabilistic judgements of the two independent causes. Participants were prompted to reason with either the full 4-node model (Figure 1) from the outset or they learned in stages that there is another possible effect of the two causes. Further, participants either observed the effects in one of the two sequences or they observed both effects at once. The prior probabilities of the cases and CPTs of the effects were the same in all conditions: $P(\mathrm{C}_1) = P(\mathrm{C}_2) = 0.15$; $P(\mathrm{E}_1 \mid \mathrm{C}_1, \mathrm{C}_2) = 0.99$, $P(\mathrm{E}_1 \mid \mathrm{C}_1, \neg\mathrm{C}_2) = P(\mathrm{E}_1 \mid \neg\mathrm{C}_1, \mathrm{C}_2) = 0.7$, $P(\mathrm{E}_1 \mid \neg\mathrm{C}_1, \neg\mathrm{C}_2) = 0$; $P(\mathrm{E}_2 \mid \mathrm{C}_1, \mathrm{C}_2) = 0.6$, $P(\mathrm{E}_2 \mid \mathrm{C}_1, \neg\mathrm{C}_2) = P(\mathrm{E}_2 \mid \neg\mathrm{C}_1, \mathrm{C}_2) = 0.2$, $P(\mathrm{E}_2 \mid \neg\mathrm{C}_1, \neg\mathrm{C}_2) = 0$. For simplicity the priors of the causes were the same and the CPTs of the effects reflected different diagnosticities of the two effects.

Methods

Participants and design

A total of 271 participants ($N_{\mathrm{MALE}} = 101$, $M_{\mathrm{AGE}} = 32.1$ years; one participant identified as neither male nor female) were recruited from Prolific Academic (`www.prolific.ac`). All participants were native English speakers who gave informed consent and were paid £1.25 for partaking in the present study, which took on average 13.9 minutes to complete. Participants were randomly assigned to one of the 2 (algebra: full or sequential) × 3 (evidence learning: all-at-once (AaO), step-by-step from E_1 to E_2 (SbS1), or step-by-step from E_2 to E_1 (SbS2)) = 6 between-participants groups (one group with 44 participants, 3 groups with 45 participants, and 2 groups with 46 participants).

Materials

All participants were given the same cover story wherein rain (C_1) and a lawn sprinkler (C_2) (two binary and independent variables) could cause a wet lawn (E_1) and/or a wet exterior house wall (E_2) (a version of the cover story can be found in Pearl, 1988). The participants in AaO condition completed an online inference questionnaire comprising of 10 comprehension questions (2 about the priors of the causes and 8 about the CPTs) and 2 test questions (one relating to $P(\mathrm{C}_1 \mid \mathrm{E}_1, \mathrm{E}_2)$ and one to $P(\mathrm{C}_2 \mid \mathrm{E}_1, \mathrm{E}_2)$). Everyone else completed the same 10 comprehension questions and 4 test questions (relating to $P(\mathrm{C}_1 \mid \mathrm{E}_\mathrm{i})$, $P(\mathrm{C}_2 \mid \mathrm{E}_\mathrm{i})$, $P(\mathrm{C}_1 \mid \mathrm{E}_\mathrm{i}, \mathrm{E}_\mathrm{j})$, and $P(\mathrm{C}_2 \mid \mathrm{E}_\mathrm{i}, \mathrm{E}_\mathrm{j})$).

Procedure

In the full algebra condition, the participants were initially presented with a causal cover story (both in a textual and a visual form) which explained the relations between variables and probabilistic information relating to the priors of both causes (priors were textually communicated as a percentage chance). They were then asked 2 priors comprehension questions. Following that, participants were told the CPTs of the two pieces of evidence (also textually communicated as a percentage chance) and subsequently asked 8 comprehension questions regarding the CPTs (in a random order). After completing the comprehension questions, participants in the AaO condition learned that both pieces of evidence occurred and were prompted to answer 2 test questions (one for each cause) presented in the same order. Participants in the SbS conditions first learned about one piece of evidence and answered 2 test questions relating to the 2 causes and then learned that the second piece of evidence occurred and asked final 2 questions. When answering the test questions participants were reminded of the priors of the causes and the CPTs of each piece of evidence, as well as their previous estimates of the two causes (in the SbS conditions).

Participants in the sequential algebra condition were initially told a cover story (both in a textual and a visual form) including only two causes and one effect. As in the full algebra condition, they were told the priors of the causes (percentage chance) and asked 2 priors comprehension questions.

In contrast to the full algebra contention, they were then told CPTs (percentage chance) regarding only one piece of evidence and completed 4 comprehension questions related to CPTs (in both the AaO and the SbS conditions). This was followed by 2 test questions relating to the probability of the causes given that one piece of evidence was observed (only in the SbS conditions). Participants then additionally learned that there is another piece of evidence potentially relevant to the probability estimates of the two causes. They learned the CPTs for the second piece of evidence and completed 4 comprehension questions followed by 2 test questions prompting them to estimate their confidence in the causes happening given the additional piece of evidence obtained. Again, participants were reminded of the priors of the causes, CPTs (but only for the current piece of evidence), and their previous estimates of the two causes (in the SbS conditions). In the AaO, after completing the first 4 comprehension questions participants were not told that the evidence obtained. Rather, they went on to learn that there is another potentially relevant piece of evidence, completed additional 4 comprehension questions, and subsequently told that both pieces of evidence obtained. After that, participants were reminded of the priors, CPTs (for the both pieces of evidence) and completed 2 test questions regarding the probabilities of the two causes.

In all conditions the test questions prompted participants to provide percentage confidence (0–100%) of C_i given one or two effects. For example, after learning that E_1 occurred, they were asked (in SbS1 condition) a diagnostic reasoning questions: "How confident are you that it **rained** overnight now that you know that the lawn is wet?" After additionally learning E_2 occurred they were asked: "How confident are you that it **rained** overnight now that you know that *both* the lawn *and* the house wall are wet?" (the full algebra condition) or "How confident are you that it **rained** overnight now that you know that the house wall is also wet?" (the sequential algebra condition). All participants provided explanations for each answer to the test questions.

Results

All the participants' responses to the test questions are plotted in Figure 3. To test the effect of the algebra and the evidence learning conditions on participants estimates on the test questions, we built a linear mixed effects model using the lme4 package (Bates, Mächler, Bolker, & Walker, 2014). The model had two fixed effects, Algebra and Evidence learning, with a random intercept for each participant (there was no random slope for participant since algebra and evidence learning conditions vary between participants). We found a main effect of Evidence learning but no main effect of Algebra (see Table 1). We also found no interaction between Algebra or Evidence learning. However, likelihood ratio tests showed that including the predictors in the model does not improve model fit compared to just having an intercept as a predictor ($\chi^2(3) = 6.11,\ p = 0.11$). That is, the data grand mean fits the data no worse than the model which includes both predictors.

Table 1: Linear mixed effect model results

A=Algebra; EL=Evidence learning

	Estimate	95% CI	t-value	p
A	-6.51	[-17.76, 4.73]	-1.13	0.26
EL	-0.53	[-1.03, -0.03]	-2.1	0.04*
A$\times$ EL	3.28	[-17.76, 4.73]	1.29	0.2

A finer grained analyses on the data within each group showed a significant difference between $P(C_1 \mid E_i)$ and $P(C_1 \mid E_i, E_j)$ in the full algebra SbS1 condition ($t(44) = -4.04, p = 0.0002$); in the full algebra SbS2 condition both between $P(C_1 \mid E_i)$ and $P(C_1 \mid E_i, E_j)$ ($t(45) = -4.87 p < 0.0001$) and $P(C_2 \mid E_i)$ and $P(C_2 \mid E_i, E_j)$ ($t(45) = -2.98, p = 0.005$); as well as in the sequential algebra SbS2 condition between $P(C_1 \mid E_i)$ and $P(C_1 \mid E_i, E_j)$ ($t(45) = -5.57, p < 0.0001$) and between $P(C_2 \mid E_i)$ and $P(C_2 \mid E_i, E_j)$ ($t(45) = -6.13, p < 0.0001$). No significant differences in the sequential SbS1 condition.

Further analyses showed that none of the $P(C_1 \mid E_i, E_j)$ and $P(C_2 \mid E_i, E_j)$ are significantly different across the levels of the evidential learning condition whereas some $P(C_2 \mid E_i)$ are: in the full algebra condition $P(C_2 \mid E_i)$ in SbS2 and SbS1 are statistically different, $t(89) = -2.09, p = 0.04$, as well as $P(C_2 \mid E_i)$ in the sequential algebra condition SbS2 and SbS1 $t(88.5) = -2.51, p = 0.014$, with those in SbS1 having higher means. Combining these results from those above regarding participants estimates withing each group suggests that (i) people are sensitive to the different orders the pieces of evidence of different diagonsticity were presented and (ii) that their estimates go against both the full CBN and the 'split' CBNs (qualitative) predictions since the differences $P(C_1 \mid E_i, E_j) - P(C_1 \mid E_i)$ and $P(C_2 \mid E_i, E_j) - P(C_2 \mid E_i)$ are larger in SbS2 condition than in SbS1 condition whereas the full CBN and the 'split' CBN predict exactly the opposite (see Figure 3).

A closer look at the data distributions in Figure 3 reveals the driving force of the results; namely, that participants' responses are highly clustered. Three clustering points ('20%', '60%', and '70%') seem to correspond to the probability values one finds in the CPTs for the effects. One clustering point corresponds to the priors of the causes ('15%'). The largest clustering point seems to be around the '50%' mark. Table 3 shows a frequency of responses around ($\pm 2\%$) the clustering points. The data captured in Table 3 amounts to $\approx 67\%$ of all data.

Finally, to assess the fit of each model to the data, we calculated mean squared errors (MSEs) for each model across the two algebra conditions.[2] Given the above-mentioned cluster-

[2]Note that the 'split' CBN does not have a unique prediction for AaO condition (see Figure 3). In calculating the MSE for that model we included the prediction that has the lower MSE.

ing around particularly the '50%' mark, we additionally calculated the MSEs for a simple model that included the correct priors (same as in both the full CBN and the 'split' CBN modeling), but has 50% as a response to all test questions. The results are presented in Table 2.

Table 2: MSEs for the full CBN, 'split' CBN, and '50%' model in the full and sequential algebra conditions

	Full algebra	Sequential algebra
Full CBN	621.18	536.94
'Split' CBN	778.93	701.97
'50%' model	573.73	496.65

The best fitting model of the three was the simple '50%' model, further confirming the clustering effect around the '50%' mark and the results of the linear mixed effect model. The full CBN model was a better fit than the 'split' CBN model of both the full algebra condition data and sequential algebra condition data. All three models fit better the sequential algebra condition data than the full algebra condition data suggesting a difference between the two conditions. However, according to the linear mixed effect model that difference is not statistically significant.

Discussion

The general goal of the paper was twofold. First, we sought to explore new avenues in sequential diagnostic reasoning by investigating peoples causal judgements with multiple independent causes and multiple pieces of evidence of different diagnosticity. To this effect we found that people are sensitive to the order of presentation of the different pieces of evidence. However, although there was a trend in increasing the probabilities of the causes after finding out that the second piece of evidence obtained (in accordance with both the full and the 'split' CBN model), the (qualitative) predictions of both models regarding the amount of increase in each order go against the participants' mean estimates.

Second, we introduced the issue of the novel variables in sequential reasoning and the practical challenge it presents. In the first empirical study on this issue, we found that people update almost identically when they are presented with the full algebra and when the algebra is expanded sequentially. In principle, this lack of difference could mean either that people are very good at this expansion, or that they inappropriately treat the full model in a sequential, local fashion. The MSE analysis showed that the full CBN model is a better fit than the 'split' CBN model across board supporting the latter option. However, the significant clustering in our data and the fact that the '50%' model fitted the data better then either the full or the 'split' CBN model suggest that participants employed different strategies in answering our test questions. Some of these seem indicative of well-established errors in human causal/probabilistic reasoning such as 'the inversion fallacy' where people confuse $P(\mathrm{A} \mid \neg \mathrm{B})$ with $P(\mathrm{B} \mid \mathrm{A})$ (Nance & Morris, 2002) or more recently identified errors such as 'the zero-sum fallacy' where people treat evidence as a zero-sum game in which alternative independent hypotheses compete for evidential support which may lead to splitting the probability space between the hypotheses (Pilditch, Fenton, & Lagnado, 2019). The prevalence of such errors may mask other differences that would emerge across those contexts. In particular, systematic differences may yet be found in more naturalistic scenarios where there are no explicit numbers for people to hold on to. This should be pursued in future work.

Acknowledgments

This research has been partly supported by the Humboldt Foundation's "Anneliese Meier Research Award" to Ulrike Hahn.

References

Anderson, N. H. (1965). Primacy effects in personality impression formation using a generalized order effect paradigm. *Journal of personality and social psychology*, *2*(1), 1–9.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

Hayes, B. K., Hawkins, G. E., Newell, B. R., Pasqualino, M., & Rehder, B. (2014). The role of causal models in multiple judgments under uncertainty. *Cognition*, *133*(3), 611–620.

Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive psychology*, *24*(1), 1–55.

Jacowitz, K. E., & Kahneman, D. (1995). Measures of anchoring in estimation tasks. *Personality and Social Psychology Bulletin*, *21*(11), 1161–1166.

Meder, B., & Mayrhofer, R. (2017a). Diagnostic causal reasoning with verbal information. *Cognitive psychology*, *96*, 54–84.

Meder, B., & Mayrhofer, R. (2017b). Diagnostic reasoning. In M. R. Waldmann (Ed.), *Oxford handbook of causal reasoning* (pp. 433–458). Oxford University Press New York.

Nance, D. A., & Morris, S. B. (2002). An empirical assessment of presentation formats for trace evidence with a relatively large and quantifiable random match probability. *Jurimetrics*, *42*, 403–448.

Neapolitan, R. E. (2003). *Learning bayesian networks*. Pearson Prentice Hall Upper Saddle River, NJ.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Francisco, CA: Morgan Kauffman.

Pearl, J. (2009). *Causality*. Cambridge university press.

Pilditch, T. D., Fenton, N., & Lagnado, D. (2019). The zero-sum fallacy in evidence evaluation. *Psychological Science*. Retrieved from https://doi.org/10.1177/0956797618818484

Rebitschek, F. G., Bocklisch, F., Scholz, A., Krems, J. F., & Jahn, G. (2015). Biased processing of ambiguous symptoms favors the initially leading hypothesis in sequential diagnostic reasoning. *Experimental psychology*, *62*(5), 287–305.

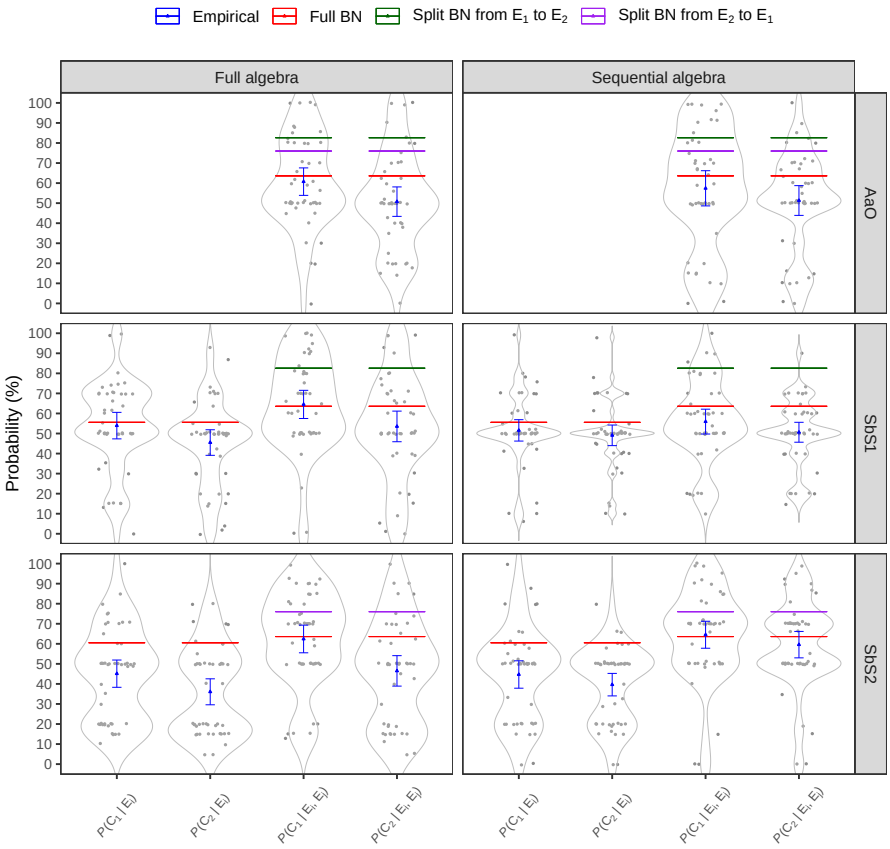


Figure 3: Distributions of participants' responses to the test questions. Error bars are 95% confidence intervals.

Appendix A

We adopt the following convention: $\overline{a} = 1 - a$.

Theorem 1. *$P_1(\mathrm{C}_1 \mid \mathrm{E}_1, \mathrm{E}_2) = P_2(\mathrm{C}_1^* \mid \mathrm{E}_2)$ if and only if (i) $\alpha_1\,\delta_1 = \beta_1\,\gamma_1$ or (ii) $\alpha_2 = \beta_2$ and $\gamma_2 = \delta_2$.*

Proof.

$$
\begin{aligned}
P_1(\mathrm{C}_1 \mid \mathrm{E}_1, \mathrm{E}_2) &= \frac{P_1(\mathrm{C}_1, \mathrm{E}_1, \mathrm{E}_2)}{P_1(\mathrm{E}_1, \mathrm{E}_2)} \\
&= \frac{P_1(\mathrm{C}_1) \sum_{C_2} P_1(\mathrm{E}_1 \mid \mathrm{C}_1, C_2)\, P_1(\mathrm{E}_2 \mid \mathrm{C}_1, C_2)\, P_1(C_2)}{\sum_{C_1, C_2} P_1(\mathrm{E}_1 \mid C_1, C_2)\, P_1(\mathrm{E}_2 \mid C_1, C_2)\, P_1(C_1)\, P_1(C_2)} \\
&= \frac{A_1}{A_1 + A_2} \\
A_1 &:= c_1\,(\alpha_2\,\alpha_1\,c_2 + \beta_2\,\beta_1\,\overline{c_2}) \\
A_2 &:= \overline{c_1}\,(\gamma_2\,\gamma_1\,c_2 + \delta_2\,\delta_1\,\overline{c_2})
\end{aligned}
$$

$$
\begin{aligned}
P_2(\mathrm{C}_1^* \mid \mathrm{E}_2) &= \frac{P_2(\mathrm{C}_1^*, \mathrm{E}_2)}{P_2(\mathrm{E}_2)} \\
&= \frac{P_2(\mathrm{C}_1^*) \sum_{C_2} P_2(\mathrm{E}_2 \mid \mathrm{C}_1^*, C_2^*)\, P_1(C_2^*)}{\sum_{C_1^*, C_2^*} P_2(\mathrm{E}_2 \mid \mathrm{C}_1^*, C_2^*)\, P_2(\mathrm{C}_1^*) P_2(C_2^*)} \\
&= \frac{P_1(\mathrm{C}_1 \mid \mathrm{E}_1) \sum_{C_2} P_1(\mathrm{E}_2 \mid \mathrm{C}_1, C_2)\, P_1(C_2 \mid \mathrm{E}_1)}{\sum_{C_1, C_2} P_1(\mathrm{E}_2 \mid C_1, C_2)\, P_1(C_1 \mid \mathrm{E}_1) P_1(C_2 \mid \mathrm{E}_1)} \\
&= \frac{B_1}{B_1 + B_2}
\end{aligned}
$$

$$
\begin{aligned}
B_1 &:= c_1\,(\alpha_1\,c_2 + \beta_1\,\overline{c_2}) \cdot \\
&\quad \cdot [\alpha_2\,c_2\,(\alpha_1\,c_1 + \gamma_1\,\overline{c_1}) + \beta_2\,\overline{c_2}\,(\beta_1\,c_1 + \delta_1\,\overline{c_1})] \\
B_2 &:= \overline{c_1}\,(\gamma_1\,c_2 + \delta_1\,\overline{c_2}) \cdot \\
&\quad \cdot [\gamma_2\,c_2\,(\alpha_1\,c_1 + \gamma_1\,\overline{c_1}) + \delta_2\,\overline{c_2}\,(\beta_1\,c_1 + \delta_1\,\overline{c_1})]
\end{aligned}
$$

Let $\Delta_1 := P_1(\mathrm{C}_1 \mid \mathrm{E}_1, \mathrm{E}_2) - P_2(\mathrm{C}_1^* \mid \mathrm{E}_2)$. Then

$$
\begin{aligned}
\Delta_1 &= \frac{A_1\,(B_1 + B_2) - B_1\,(A_1 + A_2)}{(A_1 + A_2)\,(B_1 + B_2)} \\
&= \frac{A_1\,B_1 + A_1\,B_2 - A_1\,B_1 - A_2\,B_1}{P_1(\mathrm{E}_1, \mathrm{E}_2)\,P_2(\mathrm{E}_2)} = \frac{A_1\,B_2 - A_2\,B_1}{P_1(\mathrm{E}_1, \mathrm{E}_2)\,P_2(\mathrm{E}_2)} \\
&= \frac{c_1\,\overline{c_1}\,c_2\,\overline{c_2}\,(\alpha_1\,\delta_1 - \beta_1\,\gamma_1)\,[G_1 + G_2]}{P_1(\mathrm{E}_1, \mathrm{E}_2)\,P_2(\mathrm{E}_2)}. \\
G_1 &:= (\gamma_2 - \delta_2)\,c_1\,(\alpha_2\,\alpha_1\,c_2 + \beta_2\,\beta_1\,\overline{c_2}) \\
G_2 &:= (\alpha_2 - \beta_2)\,\overline{c_1}\,(\gamma_2\,\gamma_1\,c_2 + \delta_2\,\delta_1\,\overline{c_2})
\end{aligned}
$$

■

Using a similar proof strategy one can show that: (a) $P_1(\mathrm{C}_2 \mid \mathrm{E}_1, \mathrm{E}_2) = P_2(\mathrm{C}_2^* \mid \mathrm{E}_2)$ if and only if $\alpha_1\,\delta_1 = \beta_1\,\gamma_1$ or (ii) $\alpha_2 = \gamma_2$ and $\beta_2 = \delta_2$; (b) $P_1(\mathrm{C}_1 \mid \mathrm{E}_1, \mathrm{E}_2) = P_3(\mathrm{C}_1^* \mid \mathrm{E}_1)$ if and only if (i) $\alpha_2\,\delta_2 = \beta_2\,\gamma_2$ or (ii) $\alpha_1 = \beta_1$ and $\gamma_1 = \delta_1$; and (c) $P_1(\mathrm{C}_2 \mid \mathrm{E}_1, \mathrm{E}_2) = P_3(\mathrm{C}_2^* \mid \mathrm{E}_1)$ if and only if (i) $\alpha_2\,\delta_2 = \beta_2\,\gamma_2$ or (ii) $\alpha_1 = \gamma_1$ and $\beta_1 = \delta_1$ (proofs omitted).

It follows then that $P_1(\mathrm{C}_1 \mid \mathrm{E}_1, \mathrm{E}_2) = P_2(\mathrm{C}_1^* \mid \mathrm{E}_2) = \mathrm{P}_3(\mathrm{C}_1^* \mid \mathrm{E}_1)$ if (1) $\alpha_1\,\delta_1 = \beta_1\,\gamma_1$ and $\alpha_2\,\delta_2 = \beta_2\,\gamma_2$, or (2) $\alpha_1 =$

Table 3: Frequency of participants' reposes around five focal points

	Full algebra				Sequential algebra			
	$P(\mathrm{C}_1 \mid \mathrm{Ei})$	$P(\mathrm{C}_2 \mid \mathrm{Ej})$	$P(\mathrm{C}_1 \mid \mathrm{Ei},\mathrm{Ej})$	$P(\mathrm{C}_2 \mid \mathrm{Ei},\mathrm{Ej})$	$P(\mathrm{C}_1 \mid \mathrm{Ei})$	$P(\mathrm{C}_2 \mid \mathrm{Ej})$	$P(\mathrm{C}_1 \mid \mathrm{Ei},\mathrm{Ej})$	$P(\mathrm{C}_2 \mid \mathrm{Ei},\mathrm{Ej})$
AaO								
'15%'			0	2			3	3
'20%'			2	6			2	0
'50%'			14	13			11	14
'60%'			4	3			2	5
'70%'			3	3			5	5
SbS1								
'15%'	5	3	0	1	1	2	0	1
'20%'	0	3	0	2	0	0	5	5
'50%'	15	19	14	14	23	22	13	18
'60%'	3	0	4	3	3	1	9	7
'70%'	8	4	5	4	5	6	3	6
SbS2								
'15%'	4	9	4	7	3	4	1	1
'20%'	10	11	2	4	9	9	0	1
'50%'	15	12	12	13	17	20	13	17
'60%'	2	3	3	3	5	4	1	0
'70%'	5	3	10	4	0	0	15	12

β_1 and $\gamma_1 = \delta_1$, or (3) $\alpha_2 = \beta_2$ and $\gamma_2 = \delta_2$. Similarly, $P_1(\mathrm{C}_2 \mid \mathrm{E}_1,\mathrm{E}_2) = P_2(\mathrm{C}_2^* \mid \mathrm{E}_2) = \mathrm{P}_3(\mathrm{C}_2^* \mid \mathrm{E}_1)$ if (1) $\alpha_1\delta_1 = \beta_1\gamma_1$ and $\alpha_2\delta_2 = \beta_2\gamma_2$, or (2) $\alpha_1 = \gamma_1$ and $\beta_1 = \delta_1$, or (3) $\alpha_2 = \gamma_2$ and $\beta_2 = \delta_2$. Therefore, the order is not important and one can decompose a full CBN in smaller ones while preserving the same probability distributions if (1) $\alpha_1\delta_1 = \beta_1\gamma_1$ and $\alpha_2\delta_2 = \beta_2\gamma_2$; or (2) $\alpha_1 = \beta_1$, $\gamma_1 = \delta_1$, $\alpha_2 = \gamma_2$, and $\beta_2 = \delta_2$; or (3) $\alpha_2 = \beta_2$, $\gamma_2 = \delta_2$, $\alpha_1 = \gamma_1$, and $\beta_1 = \delta_1$; or (4) $\alpha_1 = \beta_1 = \gamma_1 = \delta_1$; or (5) $\alpha_2 = \beta_2 = \gamma_2 = \delta_2$. (4) and (5) make E_1 and E_2 respectively fully undiagnostic with respect to C_1 and C_2, which violates the faithfulness condition (see Neapolitan, 2003). (1) implies that C_1 and C_2 are conditionally independent given E_1 and that they are also conditionally independent given E_2, that is, learning E_1 makes C_1 and C_2 independent and learning E_1 makes C_1 and C_2 independent. (2) and (3) both entail (1) and are more specific versions of (1).

Appendix B

Here we show that there are no order effects when the causes are mutually exclusive and exhaustive, i.e. when $P(C_1, C_2) = 0$ and $P(C_1) + P(C_2) = 1$. We model mutually exclusive and exhaustive causes with one node, C, that has two values: C_1 and C_2.

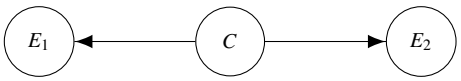


Figure 4: CBN with mutually exclusive and exhaustive causes

$$\begin{aligned} P_4(C = \mathrm{C}_1) = c \quad &, \quad P_4(C = \mathrm{C}_2) = \overline{c} \\ P_4(\mathrm{E}_1 \mid \mathrm{C}_1) = \alpha_1 \quad &, \quad P_4(\mathrm{E}_1 \mid \mathrm{C}_2) = \beta_1 \\ P_4(\mathrm{E}_2 \mid \mathrm{C}_1) = \alpha_2 \quad &, \quad P_4(\mathrm{E}_2 \mid \mathrm{C}_2) = \beta_2 \end{aligned} \tag{3}$$

Splitting the CBN from Figure 4 we get two CBNs:

$$P_5(C = \mathrm{C}_1) = c \quad , \quad P_5(C = \mathrm{C}_2) = \overline{c}$$

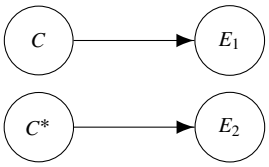


Figure 5: 'Split' CBN from E_1 to E_2

$$\begin{aligned} P_5(C^* = \mathrm{C}_1^*) = P_4(\mathrm{C}_1 \mid \mathrm{E}_1) \quad &, \quad P_5(C^* = \mathrm{C}_2^*) = P_4(\mathrm{C}_2 \mid \mathrm{E}_1) \\ P_5(\mathrm{E}_1 \mid \mathrm{C}_1) = \alpha_1 \quad &, \quad P_5(\mathrm{E}_1 \mid \mathrm{C}_2) = \beta_1 \\ P_5(\mathrm{E}_2 \mid \mathrm{C}_1^*) = \alpha_2 \quad &, \quad P_5(\mathrm{E}_2 \mid \mathrm{C}_2^*) = \beta_2 \end{aligned} \tag{4}$$

Theorem 2. *$P_4(\mathrm{C}_1 \mid \mathrm{E}_1,\mathrm{E}_2) = P_5(\mathrm{C}_1^* \mid \mathrm{E}_2)$ when $P_{4,5}(C_1^{(*)}, C_2^{(*)}) = 0$ and $P_{4,5}(C_1^{(*)}) + P_{4,5}(C_2^{(*)}) = 1$.*

Proof.

$$\begin{aligned} P_4(\mathrm{C}_1 \mid \mathrm{E}_1,\mathrm{E}_2) &= \frac{P_4(\mathrm{C}_1)\,P_4(\mathrm{E}_1 \mid \mathrm{C}_1)\,P_4(\mathrm{E}_2 \mid \mathrm{C}_1)}{\sum_C P_4(C)\,P_4(\mathrm{E}_1 \mid C)\,P_4(\mathrm{E}_2 \mid C)} \\ &= \frac{c\,\alpha_1\,\alpha_2}{c\,\alpha_1\,\alpha_2 + \overline{c}\,\beta_1\,\beta_2} \\ P_5(\mathrm{C}_1^* \mid \mathrm{E}_2) &= \frac{P_5(\mathrm{C}_1^*)\,P_5(\mathrm{E}_2 \mid \mathrm{C}_1^*)}{\sum_{C^*} P_5(C^*)\,P_5(\mathrm{E}_2 \mid C^*)} \\ &= \frac{P_4(\mathrm{C} \mid \mathrm{E}_1)\,P_4(\mathrm{E}_2 \mid \mathrm{C})}{\sum_C P_4(C \mid \mathrm{E}_1)\,P_4(\mathrm{E}_2 \mid C)} \\ &= \frac{J\,\alpha_2}{J\,\alpha_2 + (1-J)\,\beta_2} \\ J &:= \frac{c\,\alpha_1}{c\,\alpha_1 + \overline{c}\,\beta_1} \end{aligned}$$

Let $\Delta_2 := P_4(\mathrm{C}_1 \mid \mathrm{E}_1,\mathrm{E}_2) - P_5(\mathrm{C}_1^* \mid \mathrm{E}_2)$. Then

$$\Delta_2 = \frac{c\,\alpha_1\,\alpha_2\,\beta_2\left[1 - \frac{c\,\alpha_1 + \overline{c}\,\beta_1}{c\,\alpha_1 + \overline{c}\,\beta_1}\right]}{(c\,\alpha_1\,\alpha_2 + \overline{c}\,\beta_1\,\beta_2)(J\,\alpha_2 + (1-J)\,\beta_2)} = 0$$

■

Since $P_4(\mathrm{C}_2 \mid \mathrm{E}_1,\mathrm{E}_2) = 1 - P_4(\mathrm{C}_1 \mid \mathrm{E}_1,\mathrm{E}_2)$ and $P_5(\mathrm{C}_2^* \mid \mathrm{E}_2) = 1 - \mathrm{P}_5(\mathrm{C}_1^* \mid \mathrm{E}_2)$, then given Theorem 2 it also true that $P_4(\mathrm{C}_2 \mid \mathrm{E}_1,\mathrm{E}_2) = P_5(\mathrm{C}_2^* \mid \mathrm{E}_2)$. Similarly we get that $P_4(\mathrm{C}_1 \mid \mathrm{E}_1,\mathrm{E}_2) - P_6(\mathrm{C}_1^* \mid \mathrm{E}_1) = 0$ and $P_4(\mathrm{C}_2 \mid \mathrm{E}_1,\mathrm{E}_2) - P_6(\mathrm{C}_2^* \mid \mathrm{E}_1) = 0$ (proofs omitted).

Incremental understanding of conjunctive generic sentences

Michael Henry Tessler, Karen Gu, and Roger Levy
{tessler, karengu, rplevy}@mit.edu
Department of Brain and Cognitive Sciences, MIT
Cambridge, MA 02139 USA

Abstract

Generic statements convey generalizations about categories, but how generic predications combine is unclear. "Elephants live in Africa and Asia" does not mean that individual elephants live on both continents. In addition, such conjunctive generics pose interesting questions for theories of incremental processing because the meaning of the sentence can change part-way through: "Elephants live in Africa" would imply most or all do, but "Africa and Asia" implies some live in each. We extend a recently proposed computational model of generic language understanding with an incremental processing mechanism that can begin to interpret an utterance before a speaker has finished their sentence. This model makes novel predictions about partial interpretations of conjunctive generic sentences, which we test in two behavioral experiments. The results support a strong view of incrementality, wherein listeners continuously update their beliefs based on expectations about where a speaker will go next with their utterance.

Keywords: semantics; pragmatics; incremental processing; generics; psycholinguistics

Introduction

Much of what we come to learn about the world comes not from direct experience but from knowledge conveyed to us from others, often in the form of linguistic utterances. "Elephants eat 300 pounds of a food in a day" succinctly conveys information extending beyond any particular moment in time or space: It could apply to any elephant, on any day of the week. Utterances that communicate generalizations are called *generic* utterances (Carlson, 1977; Carlson & Pelletier, 1995), and they are the foremost case study of rich, abstract knowledge conveyed in simple utterances (Gelman, 2009).

Generics are rife with philosophical puzzles that make it difficult to develop a unified, formal theory of their meaning (for useful reviews: Carlson & Pelletier, 1995; Nickel, 2016). One largely understudied puzzle concerns how generic predications combine. Consider the null hypothesis that generics convey information about the percentage of the category with the property—the *prevalence*—in a way analogous to how majority quantifiers (e.g., *most*, *all*) work (e.g., "Most elephants eat 300 pounds of food in day"). How can such an account treat a generic involving a conjunctive predication like "Elephants live in Africa and Asia"? No elephant actually lives on both continents; instead, the sentence should be understood as *(generically) elephants live in Africa and (generically) elephants live in Asia*, but this is impossible if each individual generic sentence means that the majority holds the property (i.e., it is impossible for more than half of elephants to live in Africa and more than half of elephants to live in Asia; Nickel, 2008). The prevalence implied by a generic involving a conjunctive, mutually-exclusive predicate seems more lax than if only one of conjuncts were mentioned: If a speaker said "Elephants live in Africa", you might think they all do.

The puzzle of understanding conjunctive generic sentences deepens when one considers that linguistic input is processed incrementally (e.g., Altmann & Kamide, 1999): Listeners ubiquitously form expectations about the intended meaning of a sentence before the speaker finishes it (e.g., Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). For conjunctive generics about mutually exclusive properties, strongly incremental language understanding might produce non-monotonic belief updates: after the sentence prefix "Elephants live in Africa…", a comprehender might infer a higher prevalence than after hearing the sentence completion "…and Asia". If such non-monotonic updates occur, what types of linguistic input trigger them?

In this paper we show that a recently proposed model of generic language can accommodate these complex inferential patterns and we empirically test two predictions about generic interpretation that address these puzzles. The model of Tessler and Goodman (2019) treats generics as a kind of vague quantifier: interpretation of a generic depends on prior beliefs about properties. First, we show how when properties are likely to be mutually incompatible (as in *live in Africa and Asia*), listeners infer lower prevalences of each property following a conjunctive generic sentence than when the properties are compatible. Second, we show how the above model, when integrated with expectation-based probabilistic theories of syntactic processing (Hale, 2001; Levy, 2008), predicts that comprehenders update their beliefs about property prevalence not just when encountering a second, conjoined property, but immediately upon encountering evidence that a second, conjoined property is likely to be forthcoming. We test these predictions in two behavioral experiments that probe listeners' understanding of conjunctive generic sentences at different points mid-sentence, analogous to gating paradigms in psycholinguistics (Grosjean, 1980). Our empirical data confirm both predictions, suggesting that generic language interpretation interacts jointly with world knowledge and strongly incremental syntactic processing according to principles of probabilistic inference under uncertainty.

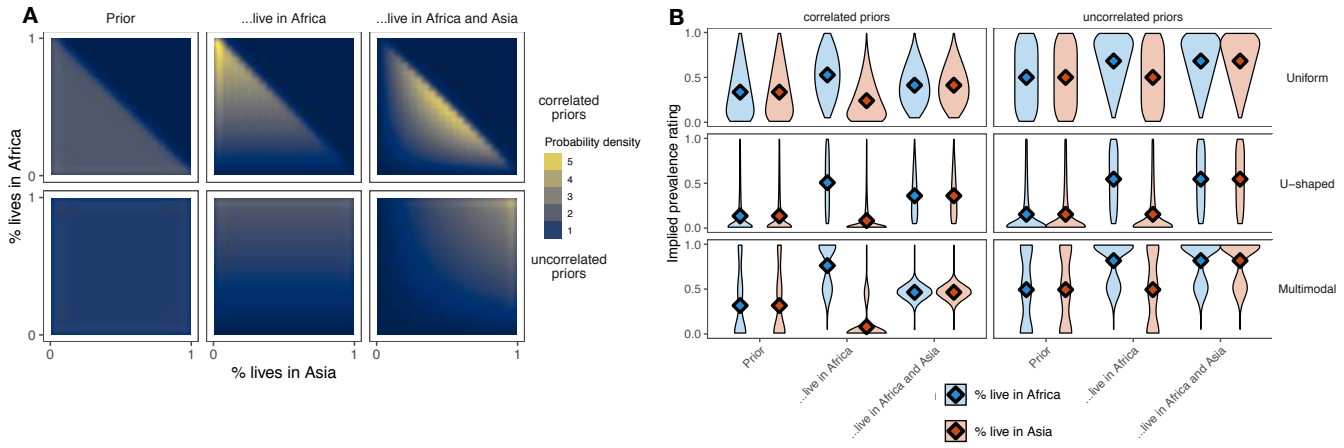


Figure 1: Model's sequential interpretation of "Elephants live in Africa... and Asia." A: Correlated priors reflected in the joint probability distribution over two features result in a mutual exclusivity inference. When the model hears only "... live in Africa", it believes that probably all live in Africa (middle facet); when it hears they live in Asia as well, the model non-monotonically updates its beliefs about how many live in Africa. B: The mutual exclusivity inference holds for priors of different shapes and never holds if the prior knowledge about the two features is uncorrelated. Points show means of distributions. U-shaped priors were Beta(0.1, 1); multimodal priors were an equal mixture of a Beta(1, 100) and a Beta(25, 1). Correlated priors were created by adding an additional factor that decreased the probability of a prevalence-level if the sum of the prevalence of the two features exceeded 100%.

Computational Model

We extend a model for interpreting generics to incorporate an incremental processing mechanism that allows a listener to understand partial utterances. The original model of Tessler and Goodman (2019) interprets a generic utterance predicating a property of a category ("Elephants eat 300 pounds of food in a day") as meaning that the prevalence (or probability) x of the property given the category—P(eats 300 lb. of food in a day| is an elephant)—is greater than an *a priori* uncertain threshold θ. The literal meaning of the generic—an uncertain threshold function, with uniform uncertainty over the threshold $P(\theta)$—combines with a listener's prior knowledge about the prevalence of the feature $P(x)$ within a relevant set of alternative categories (e.g., other animals) to compute a posterior distribution over prevalence x:

$$P(x \mid u) = \int_{\theta} P(x, \theta \mid u) d\theta \propto P(x) \cdot P(\theta) \cdot \delta_{[\![u]\!](x,\theta)} \quad (1)$$

where $\delta_{[\![u]\!](x,\theta)}$ is the Kronecker delta function assigning a value of 1 for utterances that are literally true (in the case of a generic: where $x > \theta$) and 0 for utterances that are false.

To interpret a generic with a conjunctive predicate such as "Elephants live in Africa and Asia", we assume the semantic representation contains a conjunction of two generic statements: $[Gen(\text{elephant})(\text{live\_in\_Africa})] \wedge [Gen(\text{elephant})(\text{live\_in\_Asia})]$, where the *Gen* operator acts according to the belief-updating rule of Eq. 1 (see Nickel (2008) for supporting arguments of this semantic parse). A listener starts with a joint prior over the prevalence of the two properties (we denote variables associated with *living in Africa* with subscript r and *Asia* with s): $P(\mathbf{x}) = P(x_r, x_s)$, which is incrementally updated with each successive generic. The model can then interpret multiple generics in succession, using the posterior distribution over prevalence $P(\mathbf{x} \mid u)$ (Eq. 1) as the prior for the next utterance.

$$P(\mathbf{x} \mid u_r, u_s) \propto \int_{\theta_s} \int_{\theta_r} P(\mathbf{x}, \theta \mid u_r) \cdot \delta_{[\![u_s]\!](x_s,\theta_s)} d\theta_r d\theta_s \quad (2)$$

where $P(\mathbf{x}, \theta \mid u_r)$ is the posterior that results from hearing "Elephants live in Africa" given by Eq. 1.

The predictions for a sequential understanding of "Elephants live in Africa and Asia" are shown in Fig. 1. Upon hearing the first part of the utterance, the model believes that almost all elephants live in Africa (simulations assuming a uniform prior shown in Fig. 1A). What happens next depends upon the correlational structure of the prevalence prior: If the listener has prior knowledge suggesting the properties (living in Africa, living in Asia) are mutually exclusive (Fig. 1A top), they interpret the next part of the utterance ("...and Asia") as indicating that some (perhaps half) of elephants live in Africa and other ones live in Asia. Without this correlation in the prior, the model ends up believing that most or all elephants live both in Africa and in Asia (Fig. 1A bottom). These inferences are robust to a variety of different prevalence prior distributions, so long as the prior has the necessary correlational structure (Fig. 1B shows predictions assuming a uniform, U-shape Beta, and mixture-of-Beta distributions).

When processing a conjunctive phrase, listeners may form expectations about the complete utterance even before the sentence is over. For example, when a speaker reaches the word *Africa* in "Elephants live in *Africa*", she has many syntactically distinct options available to her to complete the sen-

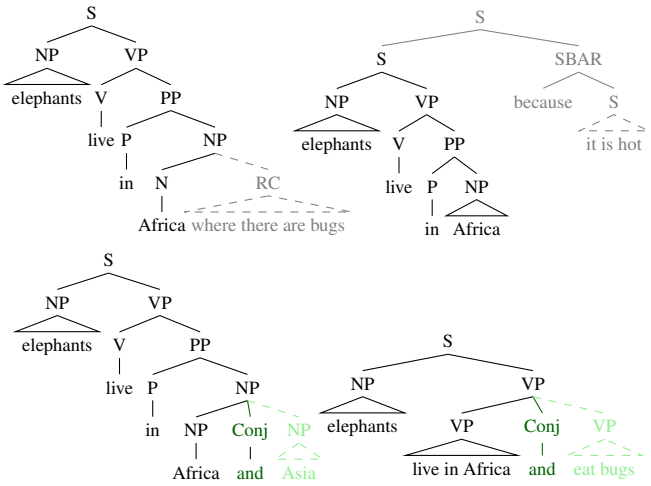


Figure 2: Incremental parse trees and syntactic expectations for upcoming conjunct properties in generic predication. The string prefix "Elephants live in Africa…" is compatible with a variety of continuations, including the four listed above. The next word, "and", rules out the continuations in the top row (depicted in gray) and sharpens expectations around a conjunct at potentially different structural levels (light green). Probabilistic renormalization implies that an upcoming conjunct mutually exclusive with the first conjunct becomes more likely when "and" is encountered, driving the strong incremental predictions depicted in Fig. 3.

tence (Fig. 2 shows four possibilities). One such possibility is that she continues with a NP-coordination that includes a mutually exclusive property (e.g., *and Asia*; bottom-left tree). When the listener encounters the word *and* in "Elephants live in Africa *and*", he knows he is entering into a coordination and the relative probability of a forthcoming mutually exclusive property increases. Such a continuation would yield a different inference about the prevalence of elephants in Africa than a continuation with a non-mutually exclusive property (e.g., with a verb phase such as "and eat bugs").[1] If listeners parse and interpret utterances incrementally at the level of individual words, then we would expect their inferences about the prevalence of elephants in Africa to represent a mixture of the inferences derived from different possible continuations, which can be represented by conditional probabilities of the full utterance u' given the sentence fragment heard f:

$$P(x \mid f) = \sum_{u'} P(x \mid u') P(u' \mid f) \tag{3}$$

If, however, listeners do not derive incremental interpretations at each moment, but instead wait for meaningful pieces of an utterance (e.g., content words like *Asia*) to compute interpretations, then we would not expect such an intermediate degree of interpretation: "Elephants live in Africa and…" should mean the same thing as "Elephants live in Africa…" (Fig. 3). We test this prediction in Expt. 2 using a gating paradigm in the spirit of Grosjean (1980).

[1] For illustrative purposes, we assume a correlation between NP vs. VP coordination and mutually exclusive vs. non-mutually exclusive predicates. Of course, it is possible to continue with a verb phrase about a mutually exclusive property such as "…live in Africa and live in Asia" as well as continue with a noun phrase about a non-mutually exclusive property (e.g., "…eat figs and nuts"). The crucial fact is that the probability of a forthcoming mutually-exclusive property increases when the comprehender encounters the word *and*.

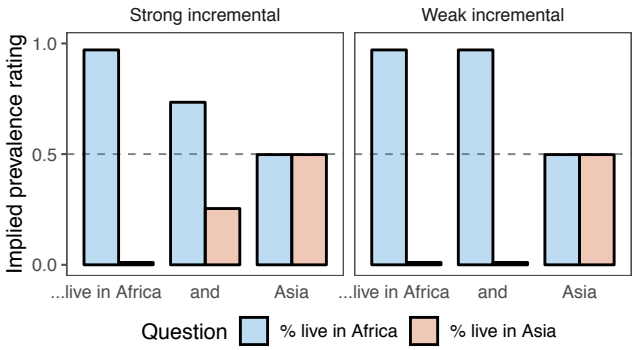


Figure 3: A model that incorporates syntactic expectations at the level of individual words (*strong incremental*) predicts intermediate mutual-exclusivity inferences part-way through the conjunction (at "and"), whereas a model that waits for content-words (*weak incremental*) does not show a difference in expected prevalence at the word "and."

Experiments

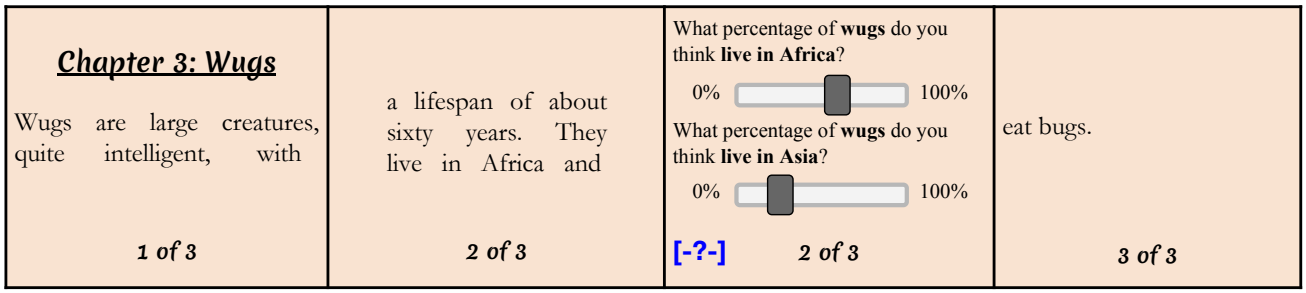
We design two experiments to test the mutual exclusivity (ME) and incremental predictions. Expt. 1 tests the ME prediction that "Elephants live in Africa and Asia" means roughly that half live in Africa and half live in Asia; this experiment also serves to validate the gating procedure we employ in the second experiment. Expt. 2 is a pre-registered study that uses the gating paradigm to test the fine-grained incremental predictions of the model. The experiments and a full list of materials can be viewed at `tinyurl.com/elephants-cogsci`.

Experiment 1: Mutual exclusivity inference

Participants We recruited 27 participants through Amazon's Mechanical Turk. Participants were restricted to those with verified U.S. IP addresses and at least a 95% work approval rating. The study took about 10 minutes and participants were compensated $1.50.

Materials Participants read a storybook with chapters about creatures on a faraway planet. Each chapter contained a short paragraph presented across 2–4 screens, with a button to "turn the page" (Fig. 4A). A chapter introduced one or a few novel categories (e.g., *wugs*) and semi-novel properties (e.g., *live on the continent of Caro*). Critical trial chapters ended with a generic sentence about conjunctive properties, which differed only in whether the second property was mutually exclusive with the first (*conjunct type*): "Glippets live on the continent of Caro and *on the continent of Este* (ME) /

A



B

			[-?-]
		Experiment 1	
INTERRUPTED	**ME**	[Wugs…] [Wugs live in Africa] **[-?-]** *[and eat bugs.]*	1. % AFRICA 2. % ASIA
INTERRUPTED	**NME**	[Wugs…] [Wugs live in Africa] **[-?-]** *[and drink water.]*	1. % AFRICA 2. % EAT BUGS
UNINTERRUPTED	**ME**	[Wugs…] [Wugs live in Africa and Asia.] **[-?-]**	1. % AFRICA 2. % ASIA
UNINTERRUPTED	**NME**	[Wugs…] [Wugs live in Africa and eat bugs.] **[-?-]**	1. % AFRICA 2. % EAT BUGS
		Experiment 2	
INTERRUPTED	**A**	[Wugs…] [Wugs live in Africa] **[-?-]** *[and eat bugs.]*	1. % AFRICA 2. % OTHER CONTINENT
INTERRUPTED	**A&**	[Wugs…] [Wugs live in Africa and] **[-?-]** *[eat bugs.]*	1. % AFRICA 2. % OTHER CONTINENT
INTERRUPTED	**A&B**	[Wugs…] [Wugs live in Africa and Asia] **[-?-]** *[which are warm.]*	1. % AFRICA 2. % OTHER CONTINENT
UNINTERRUPTED	**A&B**	[Wugs…] [Wugs live in Africa and Asia.] **[-?-]**	1. % AFRICA 2. % OTHER CONTINENT

Figure 4: Overview of experiments. A: Example book chapter from Expt. 2, depicting the *Interrupted A&* condition. "Africa and Asia" property is shown for illustration; actual stimuli used novel names for properties ("Caro and Este"). B: Overview of conditions for Expts. 1 and 2. [-?-] denotes point in the sentence at which the question appeared. Highlighting shows which properties were mentioned before the question, and what was asked about. See main text for full description of conditions.

enjoy the sunshine there (not ME: NME)."

The earlier content of the chapter supported the mutually exclusive interpretation of the properties when the properties were not *ipso facto* mutually exclusive. For example:

> Krens are a tribe of the aliens that live on the continent of Benli. Animals like stups, four-legged creatures with large antlers, are a resource for the Krens. Stups roam all over the windy highlands of Benli, far from the oceans. Krens are stup-herders and (*fishermen / incorporate stups into their religion*).

Conjunct type (ME vs. NME) was manipulated within participants and items. There were 14 filler chapters with content similar to the critical chapters but using explicit quantifiers (*most*, *all*, *none*) to describe the properties of categories.

Procedure Participants were told they would be reading a storybook with a question in each chapter. Questions were all of the same type, an *implied prevalence* question (Gelman, Star, & Flukes, 2002; Cimpian, Brandone, & Gelman, 2010): "What percentage of Ks do you think F?", where K represents a category and F a feature. Responses were recorded using a slider with endpoints labeled 0% and 100%, with the exact value selected appearing above the slider. Participants were familiarized with the response variable in a practice trial, where they were asked to report how many *dogs bark*, *birds are male*, *cats get cancer*, and *lions lay eggs*. These questions encouraged participants to use the full range of the response scale as well as served as a comprehension check.

In each critical chapter of the storybook, two questions appeared either at the end of the chapter (*Uninterrupted* conditions) or interrupting the chapter right before the final page (*Interrupted* conditions; Fig. 4). In the Interrupted conditions, the question came in the middle of a conjunctive generic sentence, but before the conjunction so the reader was unaware the sentence would continue with a conjunction. The questions asked about the mentioned property (e.g., *Africa*) and either a mutually exclusive property (*Asia*; *ME* conditions) or a nonmutually exclusive property (e.g., *eats bugs*; *NME* conditions); the chapter then concluded with a conjunction about an unmentioned, nonmutually exclusive property (Fig. 4B), so as to not give the impression that the participant was being tricked by being asked about a property that we would eventually reveal. In the critical trials, the second question was asked about the second property mentioned (ME vs. NME). Filler trials asked about two properties described in the chapter using quantifiers (i.e., *all*, *most*, or *none*). The order in which the two questions appeared on the screen was randomized on each trial.

Each participant read a total of 21 chapters, which included 8 ME conjunctions, 4 NME conjunctions, and 6 quantifier fillers; for each of these categories, equal numbers of interrupted and uninterrupted were used. The experiment began with a chapter with no questions and 2 filler trials; the remaining trials were presented in a random order such that no two critical trials were presented back-to-back. Subjectively, the task is very difficult as participants learn about many different animals with lots of new names; in practice, however, participants only need to recall information from the previously en-

countered sentence to answer the trial questions. Following the storybook, participants completed a memory check where they had to select all the facts they had learned from a list of 10 (5 real, 5 distractor); in addition, participants were asked to explain what the experiment was about in broad terms.

Results 11 participants were excluded for failing to respond accurately to all of the practice trials or failing to respond accurately to at least 7 of the 10 memory check prompts (same exclusion criteria for Expt. 2). We describe the results using the running example of "Elephants live in Africa and Asia", but the experimental stimuli used novel categories and relatively novel properties.

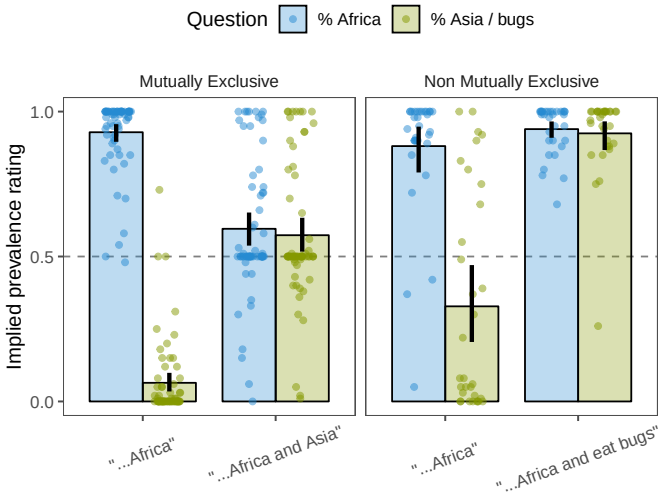


Figure 5: Experiment 1 results. Participants rate prevalence for mentioned property (% live in Africa) and either the mutually exclusive property (left facet) or non-mutually exclusive property (right facet), mid sentence ("Africa") or after the sentence finishes ("Africa and X"). Error-bars denote bootstrapped 95% confidence intervals.

The results are visually apparent in Fig. 5. Reading only "Elephants live in Africa..." led listeners to believe, on average, that all elephants lived in Africa and that none lived in Asia, whereas if the sentence finished "...and Asia", listeners inferred that roughly half live in Africa and half live in Asia. This inference is not categorical, however; there are a number of responses to ME conjunctive generics wherein participants infer that all or almost all have both properties (recall that many of our items are unfamiliar properties). A different pattern was observed for the NME properties, where hearing about the second property ("eat bugs") only increased participants' degree of belief in each property applying. Further, when answering about unmentioned properties, participants rated the prevalence of ME properties close to 0% whereas NME properties were rated as somewhat prevalent (green bars, "Africa"). The results replicate intuitions about how "Elephants live in Africa" should be interpreted in a context where the sentence is interrupted. The comparison with the non-mutually exclusive condition shows that the results cannot be attributed to the very act of being asked about two properties mentioned in a conjunctive generic sentence.

Experiment 2: Strong incrementality

In Expt. 1, we demonstrated that the mutually exclusive inference effects can be measured using a gating paradigm wherein participants are queried for their beliefs part-way through a sentence. Here, we exploit this paradigm to test the strong incremental processing predictions of the model, where syntactic expectations can modulate the interpretations of generic sentences in a fine-grained manner. Sample size, participant exclusion criteria, and analyses for this experiment were pre-registered on OSF `osf.io/pjt9c`.

Participants We recruited 108 participants through Amazon's Mechanical Turk. Participants were restricted to those with verified U.S. IP addresses and at least a 95% work approval rating. The study took about 10 minutes and participants were compensated $1.50.

Materials and procedure The materials and procedure followed those of Expt. 1 with the following exceptions. We modified the critical conjunctive generics to primarily involve the conjunction of two noun phrases (e.g., *ascribe to Caboo-ism and Daithism*) in order to strength the correlation between the NP-conjunction and mutual exclusivity.[2] The fillers were modified to introduce page breaks immediately before and immediately after conjunctions ("and") in order to raise participants' expectations that a sentence might be broken at a conjunction. We used additional examples of the Uninterrupted ME condition of Expt. 1 ("live in Africa and Asia.") as fillers to raise participants' expectations about ME continuations. Half of the filler trials had page breaks immediately before the "and" and half immediately after.

On critical trials, questions always interrupted the chapter right before the last page. On the question screen, the page number of the penultimate page remained on the screen to provide an additional cue that the chapter was not complete (Fig. 4A). There were three conditions corresponding to the point in the sentence at which the page break and prevalence questions occurred: "Elephants live in Africa__ and__ Asia__" (where __ denotes the page-break). In the two conditions where participants did not see the full conjunctive property before the question (INT A and A&), the sentence continued with a non-mutually exclusive property (e.g., *eat bugs*).

Finally, we changed the question about the second property (% live in Asia) to ask about "some other X", where X was the kind of property that was mentioned in the first conjunct (e.g., *live on some other continent*). This change was introduced to raise the plausibility that a second, ME property was possible while not naming one explicitly, which would be pragmatically odd given that the property is unmentioned in the INT A and A& conditions. Participants saw 18 chapters. The story started with a chapter with no questions, then participants saw 4 fillers: 2 quantifiers with interrupting questions

[2]Of the 13 items in this experiment, 9 of them were NP-coordinated (the others used coordination of prepositional phrases and adjectives). In both experiments, critical conjunctive generics always involved conjunctions of the same syntactic types (e.g., *ascribe to the Caboo religion and the Daith religion*).

and 2 uninterrupted ME fillers, in a random order. Finally, participants saw 2 of each kind of critical trial with 4 uninterrupted ME fillers and 3 quantifier fillers interleaved to avoid back-to-back critical trials.

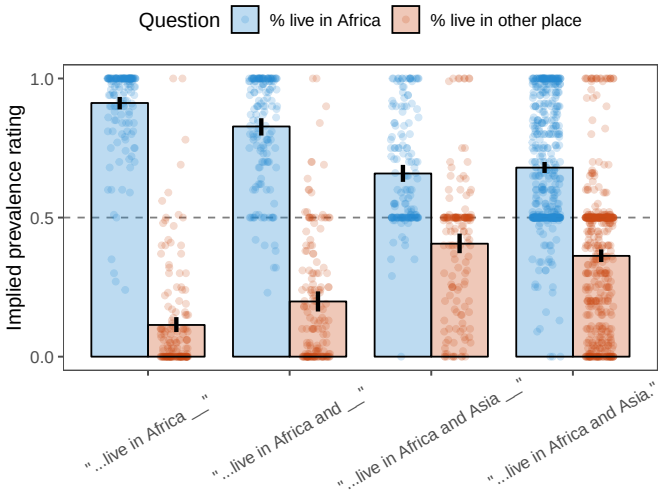


Figure 6: Experiment 2 results. Participants are interrupted at various stages of the sentence (after *Africa*, *and*, or *Asia*) to be asked about the prevalence of *living in Africa* and *living in some other place*, or asked at the end of the sentence (rightmost bars). When participants are interrupted before the second conjunct (*Asia*), the sentence continues with a non-mutually exclusive property. Error-bars denote bootstrapped 95% confidence intervals.

Results 28 participants were excluded for failing at least one of the two attention checks.[3] To test our main prediction, we constructed a Bayesian mixed-effects regression model predicting implied prevalence ratings for the first conjunct (e.g., % Africa) as a function of the point in the sentence in which participants were queried. We included by-item and by-participant random intercepts and slopes.[4] The regression model was created in Stan (http://mc-stan.org/) accessed with the brms package using default priors (Bürkner, 2017).

Replicating the findings of Expt. 1, when participants only read that "Elephants live in Africa", they tended to infer that almost all lived in Africa. When they read that "Africa and Asia", they tended to infer that the distribution was close to 50%-50%. Finally, as predicted by a strong version of incremental processing, participants began to anticipate a mutually-exclusive conjunct when only the word "and" was mentioned, as evidenced by their implied prevalence ratings being substantially less for the "..live in Africa and__" condition than the "live in Africa" condition (posterior mean estimate and 95% credible interval: $\beta = -0.08$ (-0.13, -0.04)). In addition, these ratings were substantially higher than when the full conjunctive predicate was present "..live in Africa and Asia" ($\beta = 0.17$ (0.12, 0.23); Fig. 6). Thus, we find that participants' implied prevalence ratings of how many elephants live in Africa monotonically decreased as a function of how many words of the conjunctive predicate they were allowed to see. These results suggest that listeners begin to draw pragmatic interpretations of generics before the end of the sentence and even in the absence of additional content words.

It is notable that in the "Africa and Asia" conditions, participants on average infer greater than 50% prevalence for *Africa* and lower than 50% for *Asia*, a departure from the results of Expt. 1. This deviation may be due to participants forgetting what they have read and/or not making the inference that the second conjunct stands in a subset relation to the category in the second question (e.g., that *Asia* is a kind of "some other continent"). Explicitly asking about the conjuncts alleviates memory demands by allowing participants to merely recognize, rather than recall, that they have seen this conjunct mentioned. Asking about *some other continent* (Expt. 2) requires participants' to recall the second conjunct and could lead to lower prevalence ratings in response to this question.

Discussion

Generic sentences exhibit extreme sensitivity to context that make it difficult to precisely define what a single generic conveys. "Elephants live in Africa and Asia" means neither that *most elephants live in (both) Africa and Asia* nor that *most elephants live in Africa, and most live in Asia*. Here, we empirically measured interpretations of generics about conjunctive predicates, building on the observation of Nickel (2008) of the range of troubling examples for quantificational views of generics. Notably, the uncertain threshold model of Tessler and Goodman (2019) accounts for such conjunctive generics seamlessly: An underspecified threshold can be updated as more information comes in and is sensitive to prior beliefs regarding compatibility of the conjunct properties.

We extended that model to include syntactic expectations and found evidence for the strongest form of incremental syntactic processing, wherein beliefs are continually updated based on expectations of how a sentence will continue. The fact that generic language understanding can be modulated simultaneously by correlations in background knowledge and by syntactic expectations calls for a tighter coupling between models of syntactic processing (Levy, 2008), pragmatic language understanding (Goodman & Frank, 2016), and intuitive theories (Tenenbaum, Kemp, Griffiths, & Goodman, 2011).

It remains an open question, however, how specific the effects observed in this paper are to generics rather than to quantification more generally. For example, it appears that, in some contexts, one can use *most* to convey similar mutually exclusive conjunctions: "Elephants are the largest land animal on Earth and are one of the gentlest creatures. Most live in Africa and Asia but are brought to other places for the entertainment of humans."[5] Further work is needed to determine the felicity and interpretation of such utterances.

Data, code, and links to experiments are available at
`https://github.com/mhtess/elephants`

[3] 6 failed slider check; 9 failed memory check; 13 failed both.

[4] model: $\texttt{rating} \sim \texttt{cond} + (1 + \texttt{cond} \mid \texttt{subj}) + (1 + \texttt{cond} \mid \texttt{item})$

[5] Example from theodysseyonline.com/want-to-ride-an-elephant

Acknowledgments

This work was supported by the National Science Foundation Grant #1456081, Elemental Cognition, and the MIT Sensetime Alliance and Quest for Intelligence.

References

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*, 247–264. doi: 10.1016/S0010-0277(99)00059-1

Bürkner, P.-C. (2017). brms: An r package for bayesian multilevel models using stan. *Journal of Statistical Software*, *80*(1), 1–28.

Carlson, G. N. (1977). *Reference to kinds in English*. Unpublished doctoral dissertation, University of Massachusetts, Amherst.

Carlson, G. N., & Pelletier, F. J. (1995). *The Generic Book*. Chicago, IL: Chicago University Press.

Cimpian, A., Brandone, A. C., & Gelman, S. A. (2010). Generic statements require little evidence for acceptance but have powerful implications. *Cognitive Science*, *34*(8), 1452–1482.

Gelman, S. A. (2009). Learning from others: Children's construction of concepts. *Annual review of psychology*, *60*, 115–140. doi: 10.1146/annurev.psych.59.103006.093659.LEARNING

Gelman, S. A., Star, J. R., & Flukes, J. E. (2002). Children's Use of Generics in Inductive Inferences. *Journal of Cognition and Development*, *3*(2), 179–199.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*.

Grosjean, F. (1980). Spoken word recognition processes and the gating paradigm. *Perception & Psychophysics*, *28*(4), 267–283. doi: 10.3758/BF03204386

Hale, J. (2001, 2–7 June). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the north american chapter of the Association for Computational Linguistics* (pp. 159–166). Pittsburgh, Pennsylvania. Retrieved from `http://acl.ldc.upenn.edu/N/N01/N01-1021.pdf`

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*, 1126–1177. doi: 10.1016/j.cognition.2007.05.006

Nickel, B. (2008). Generics and the ways of normality. *Linguistics and Philosophy*, *31*(6), 629–648. doi: 10.1007/s10988-008-9049-7

Nickel, B. (2016). *Between logic and the world*. Oxford: Oxford University Press.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*(5217), 1632–1634.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*(6022), 1279–1285.

Tessler, M. H., & Goodman, N. D. (2019). The language of generalization. *Psychological Review*, *126*(3), 395–436.

Using Big Data to Understand Memory and Future Thinking

Robert Thorstad & Phillip Wolff
{rthorst, pwolff} @ emory.edu
Emory University Department of Psychology

Abstract

Imagining the future and remembering the past both involve mental time travel. This commonality could indicate shared mental processes, as held by the Constructive Episodic Simulation Hypothesis (Schacter & Addis, 2008), or else interactive processes that complement one another, a possibility we call the Complementarity Hypothesis. According to the Complementarity Hypothesis, future thoughts are constructed from schemas making them episodically poor, whereas past thoughts are constructed from schemas and direct retrieval of memory traces, making them relatively episodically rich. We tested these hypotheses using machine learning to data mine mental operations in language, much as a geologist can recover physical processes from the geological record. People's natural, unprompted talk on web blogs was automatically analyzed for past, present, and future references using a temporal orientation classifier. In Study 1, we found that perceptual details were mentioned more often in past than future talk, implying greater use of episodic processing in past than future thinking. In Study 2, a neural network using schemas generated from Latent Dirichlet Allocation better predicted the content of references to the future than the past, implying that constructive processes are more common in future than past thinking. In Study 3, we used the results from the two prior studies to construct an episodic-by-constructive process space. We adapted techniques from fMRI analysis to analyze this space for clusters of activity, as if the frequency of past and future thinking were BOLD responses in cortical space. We found that past and future thinking occupy highly separable regions of processing space, supporting the Complementarity Hypothesis.

Keywords: Prospection; Memory; Future Thinking; Big Data; Naturally Occurring Datasets

Introduction

Memory is not just used to remember the past. It also helps people predict and plan for the future (Schacter & Addis, 2007; Klein, Robertson, & Delton, 2010). At a minimum, then, the cognitive process used to think about the future must be able to connect with those used to remember the past. Such a connection would be facilitated by overlap in the processes used to think about the future and past. According to the Constructive Episodic Simulations Hypothesis (Schacter & Addis, 2008) the overlap in these processes is considerable. An alternative possibility is that the thought processes used to think about the past and the future are largely unique and non-overlapping, but connect with each other in manner that complements the other. We will refer to this later possibility as the Complementarity Hypothesis. In this research, we seek to test between these two competing proposals using information afforded by machine learning and big data analytics.

The idea that thinking about the future and the past might involve similar kinds of process has received significant empirical support. Viard et al (2011) found that past and future thinking engage several common brain regions including the hippocampus, precuneus, prefrontal cortex, and posterior cingulate cortex. Addis, Wong, & Schacter (2007) found that past and future thinking both engage the left hippocampus, a region known to be involved in episodic memory. Meta-analyses suggest that the overlap between past and future thinking is robust and involves a broad set of regions in the brain's default network (Benoit & Schacter, 2015; Spreng, Mar, & Kim, 2009).

The evidence for common processes is not, however, uniform. Irish, Addis, Hodges, and Piguet (2012) found that conceptual knowledge impairments in semantic dementia were more severe in future thinking than past thinking. Craver, Kwan, Seindam, and Rosenbaum (2014) found that people who lost the ability to remember the past due to hippocampal amnesia often retained some ability to think about the future. Such patients make normal future-oriented decisions in delay discounting and score normally on surveys of future orientation. Findings such as these suggest that past and future thinking may rely on different cognitive processes.

The conflicting findings from past research are associated with different kinds of methodology. Studies supporting shared process have been those using brain imaging, while those indicating differences have been based on neuropsychological research investigating the effects of brain damage (although see Klein, Loftus, & Kihlstrom, 2002 for neuropsychological evidence for similar processing). Both kinds of research have their limitations. One of the challenges in neuroimaging work is the problem of how to elicit thoughts about the past and the future without bias to the results. Typically, temporal thoughts are elicited by explicit instructions to do so. The problem is that these instructions may alter the cognitive processes involved. For example, to image the future, participants are often instructed to imagine specific events that are highly likely to occur (e.g., Addis, Wong, & Schacter, 2007). These instructions might bias people to use their memory of the past to imagine future events because it requests that they offer specific details, a process that may not necessarily be associated with future thinking. Neuropsychological research investigating brain damage is limited by the (fortunately) relatively small numbers of participants. Most importantly, the research using both kinds of methodology has focused on people's ability to remember or imagine scenes with significant perceptual detail, but not all thoughts about the future and past are necessarily high in episodic detail. Certain thoughts about the future and past might be driven by abstract conceptual knowledge, possibly by schemas. Some research has investigated the role of cultural life scripts on future thinking (Bernsten & Bohn, 2010), but life scripts are only a small

portion of people's abstract conceptual knowledge. In sum, prior research has been limited in its ability to study the potential impact of abstract knowledge and schemas on people's thoughts about the past and the future for lack of an inventory of the generic abstract knowledge structures that people are likely to possess.

The limitations of prior work can be addressed using big data methods. Big data methods involve mining large-scale naturally occurring behavior to provide insight into human cognition (Goldstone & Lupyan, 2016; Thorstad & Wolff, 2018a). In the case of mental time travel, people talk regularly about the past, such as what they did yesterday, and the future, such as what they plan to do tomorrow. This talk can be mined to understand the cognitive processes of memory and future thinking. These big data methods address some of the challenges of prior work. Big data methods avoid the explicit prompting in prior work by studying natural, unprompted talk about time. Big data methods also allow investigation of a much wider set of conceptual knowledge by learning the relevant concepts from the data.

Here, we examine people's temporal talk in a large web blog corpus. This corpus is ideal for studying mental time travel because people write without prompting about topics of their choosing. Once the sentences in the corpus are analyzed for their temporal orientation, we can investigate the cognitive processes associated with this talk to test between the Constructive Episodic Simulation and Complementarity hypotheses.

Study 1: What is the Content of Past and Future Thinking?

The view that past and future thinking share common cognitive processes makes a strong prediction about the content of people's temporal talk. Past and future thinking have been argued to rely on shared episodic processes (Schacter & Addis, 2008), and these episodic processes have characteristic types of representation that can be identified in text. Episodic thoughts are highly concrete and perceptual, with episodic future thinking typically described as a kind of pre-experiencing (Atance & O'Neil, 2001) or simulation (Schacter & Addis, 2008). Episodic thoughts also involve a spatial location (Tulving, 1993), as also reflected in work using spatial relations as a marker of episodic future thinking (Russell, Alexis, & Clayton, 2010). We measured these episodic representations in people's talk about the past, present, and future, based on psychometric dictionaries. If past and future thinking rely on common episodic processes as predicted by the Constructive Episodic Simulation Hypothesis, then we should observe similar amounts of episodic processing in past and future talk. By contrast, if past and future thinking rely on different processes as predicted by the Complementarity Hypothesis, then we should observe more episodic processing in talk about the past than the future. Such a pattern could occur if thoughts about the future are more constructed than thoughts about the past.

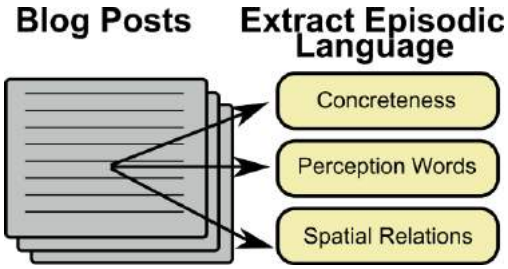


Fig. 1. Analyzing Episodic Language in Blog Posts. We extracted three episodic language indicators from a corpus of blog posts: concreteness, the amount of perceptual words, and the presence of spatial relation words.

Methods

All procedures were approved by the Emory University IRB.

Materials The analyses used the Blog Authorship Corpus (Schler, Koppel, Argamon, & Pennebaker, 2006). The corpus is demographically diverse, including 19,320 bloggers (50% female) from 40 different occupational categories and a wide range of ages (13-17y: N=8,240, 23-27y: N=8,086, 30-47y: N=2,994)**.**

Procedures Several preprocessing steps were taken to clean the corpus. Special characters, emoticons and URLs were removed. Misspellings were automatically corrected using a dictionary from Han, Cook, & Baldwin (2012). Extremely short posts with less than 10 words were dropped. Non-English sentences were removed using the Python library langdetect.

We extracted temporal talk from the corpus by automatically classifying the sentences using a temporal orientation classifier. As a first step, the sentences in the corpus were syntactically parsed using the Stanford Parser (Chen & Manning, 2014). These parses could then be used to determine temporal orientation using a set of 121 syntactic and lexical rules written in the regular expression-like language Tregex (Levy & Andrew, 2006). References to the past were flagged using rules like "VP>VG>have" and references to the future by rules like "MD>will" (Copley & Wolff, in prep.)

Before running the classifier on the corpus, the performance of the classifier was verified in a separate rating study where we recruited 30 human raters via Amazon Mechanical Turk. We obtained 3 ratings for each of 1,000 randomly drawn sentences from the blog corpus (100 ratings/participants), as to whether the sentences referred to the past, present, future, atemporal, or unintelligible. Participant quality was ensured using unmarked attention checks and by requiring participants to have completed 100 previous MTurk tasks with 95% approval rating. We found that the performance of the classifier, as indicated by the *F* statistic (Raschka, 2015), $F = 0.61$, where chance = 0.33, approached human-level accuracy, $F = 0.67$. We also compared the classifier to other classifiers based on the Linguistic Inquiry and Word Count psychometric dictionary (Pennebaker et al, 2015), a decision-tree model based on a variety of language features (Schwartz et al, 2015), and a

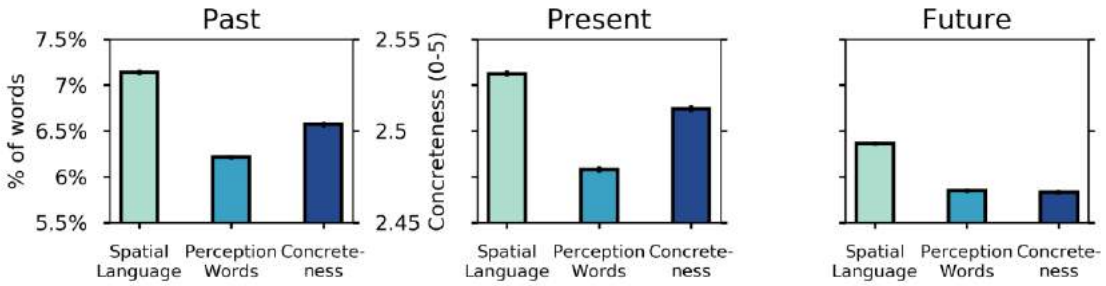


Fig. 2. Amount of episodic processing (+/- 95% CI) in sentences about past, present, and future.

regular-expression pattern-based temporal classifier known as SUTime (Chang & Manning, 2012). Our temporal orientation classifier outperformed these other temporal orientation classifiers, specifically SUTime, $F = 0.25$, decision trees, $F = 0.33$, and the Linguistic Inquiry and Word Count psychometric dictionary, $F = 0.56$.

Once applied to the blog corpus, the classifier was able to identify which sentences referred to the past (2,134,357 sentences, 39.5%), present (1,428,626 sentences, 26.5%), or future (1,834,206 sentences, 34.0%).

As shown in Fig. 1, we measured the episodic processing in each sentence in the blog corpus using three measures. We analyzed concrete language based on averaging the concreteness of the words in each sentence using concreteness ratings of 40,000 English lemmas from Byrsbaert et al (2014). We analyzed the perceptual and spatial language in each sentence by **c**alculating the proportion of words in the sentence matching predefined lists from the Linguistic Inquiry and Word Count psychometric dictionary (Pennebaker et al**.**, 2015).

Results and Discussion

We found that past thoughts involved more of all three types of episodic representations than future thoughts. In all three cases, past thoughts were more similar to present thoughts, which do not require mental time travel, than to future thoughts. As shown in Fig. 2, references to the past were rated as more spatial, $t_{(18,808)} = 48.34$, $p < 0.001$, perceptual, $t_{(18,808)} = 23.27$, $p < 0.001$, and concrete, $t_{(18,806)} = 46.65$, $p < 0.001$, than references to the future. As also shown in Fig. 2, references to the past are as perceptually rich as references to the present. Together, the results suggest that past thinking is more episodic than future thinking, a result that is fully consistent with the Complementarity Hypothesis.

Study 2: What Processes are used for Future Thinking?

Study 1 suggests that past thoughts are more episodic than future thoughts. These results raise the question of what processes are used to think about the future. An intuitive possibility is that because the past has happened but the future has not, future thoughts may be more constructed than past thoughts. This construction could be performed by relying on stored knowledge structures known as schemas. While the possibility that future thoughts rely more on schemas is intuitive, it is broadly agreed that memory also relies on schemas (Bransford & Johnson, 1972), and so one could also predict that past and future thoughts rely equally on schemas. In Study 2, we therefore asked whether future thoughts rely more on schemas than past thoughts.

There are two challenges to quantifying the influence of schemas on temporal thoughts. First, it is difficult to know in advance which schemas people use to mentally time travel. It seems likely that the most important schemas may be used in everyday talk. With a large enough sample of everyday talk, it should be possible, then, to extract these schemas. To do this, we analyzed 1 month of posts from the social media website Reddit (307 million words). We extracted the 500 most common schemas using a machine learning model known as a topic model (Blei, Ng, & Jordan, 2003). As shown in Fig. 3, a topic model works by inferring the latent topics that organize people's choices of which words to write in certain documents, or Reddit posts. These topics can thought of as probability distributions over words. While these topics do not share every feature of schemas (for example they are not hierarchical), they share some of the essential features, such as the fact that the important words represent slots that can be filled by words, which are conceptually similar, but not necessarily semantically related.

The second important challenge is that the mere presence of a schema does not necessarily imply a cognitive process. It is necessary to ask whether an author used a schema to guide their writing or merely invoked the schema incidentally. To make this leap from describing schemas to cognitive processes, we capitalized on a key cognitive function of schemas: schemas are thought to fill in missing information. For example, if a person goes to a restaurant, they can use their schemas to know that there will be a waiter even before they have seen a waiter. We created an analogue of this prediction in text by asking whether, if only a part of a sentence is provided, the rest of the sentence can be filled in based on knowledge of the schema. We did this by training a neural network to use the schemas evident in people's blog posts to predict the words they wrote next. We performed this prediction separately for sentences about the past, present, and future, thus allowing us to investigate whether schemas are more involved in filling in missing information for the past, present, or future.

If past and future thinking rely on common cognitive processes as predicted by the Constructive Episodic Simulation Hypothesis, then we would expect schemas to be

equally useful for predicting the content of people's past and future talk. By contrast, if past and future thinking rely on different cognitive processes as predicted by the Complementarity Hypothesis, then we would expect schemas to be more useful for predicting the content of people's past talk than future talk.

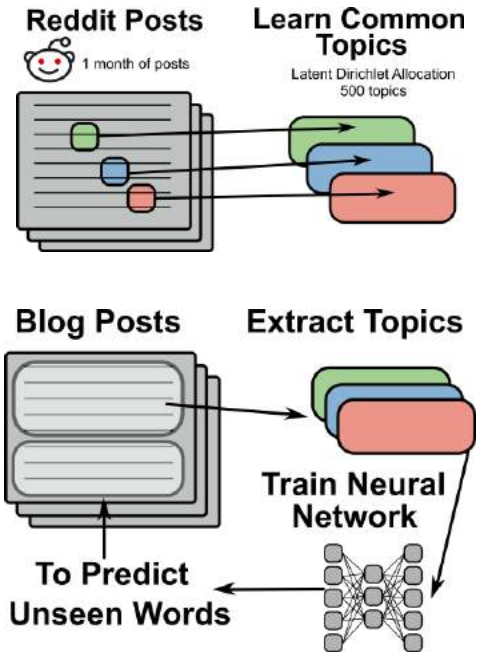


Fig. 3. Learning and Identifying Schemas in Temporal References. (Top) We identified the most prevalent schemas in a large social media corpus using Latent Dirichlet Allocation, which learns the 500 most common topics across many social media posts. (Bottom) For a particular blog post, we identified the schemas implicit in the post, and then trained a neural network to use those schemas to predict words in the unseen last sentence of the post. We conducted this prediction separately for sentences referring to the past, present, and future, allowing us to ask whether schemas were more useful for filling information for particular kinds of temporal references.

Methods

Schema Identification As shown in the top row of Fig. 3, we identified common schemas in a large corpus based on every post to the social media website Reddit in the month of January 2017 (307 million words). As shown in the top row of Fig. 3, we identified schemas in the posts by training a type of topic model known as Latent Dirichlet Allocation (Blei, Ng, & Jordan, 2003). The model was trained using the Python library *gensim* with the parameters α = 0.002, η = 0.002, *number of topics* = 500, using 100 training iterations.

Using Schemas to Fill in Information As shown in the bottom row of Fig. 3, for every post in the blog corpus, we used the LDA model to identify the schemas in the post based on every sentence except the last sentence in the post. Next, we created a dataset where the input was the schema of the post, and the output was a randomly selected word from the unseen last sentence in the post, restricting to the 5,000 most common words in the corpus. We then trained a neural network model to use the schema to predict the unseen word (out of 5,000 possible words). The model had a relatively simple architecture, with a single hidden layer with 500 units and a *relu* activation function, and was trained to minimize cross-entropy loss with *Adam* optimization, based on 25,000 training batches with a minibatch size of 100. The model was evaluated using unseen test data (10%). As described in the main text, we also trained a scrambled version of the model using the same procedure but randomly assigning words to Reddit posts.

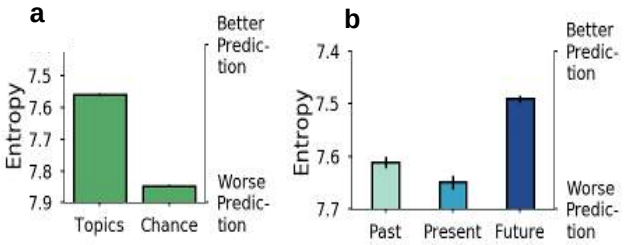


Fig. 4. Schema Usage. (A) The schemas learned by our model could predicted unseen words better than random schemas, replicating a key property of schemas. (B) These schemas were more useful for filling in unseen words for future references than past or present references.

Table 1. Schemas learned by the topic model.

Feelings	Reading	Studying	Food	Sounds	Health-care
Feeling	Read	Study	Food	Hear	Care
Feels	Reading	Subject	Eat	Sound	Health
Felt	Book	Passed	Eating	Sounds	Insuran-ce
Pain	Books	Studying	Healthy	Audio	Letter
Worse	Library	Studies	Diet	Hearing	Legal
Bad	Stuff	Exam	Dinner	Noise	Medical

Results and Discussion

We found that our model learned semantically coherent schemas from social media. We also found that future thoughts drew more on these schemas than did past thoughts, consistent with the Complementarity Hypothesis.

We first asked whether our topic model learned semantically coherent schemas. Several of the schemas are shown in Table 1. The schemas are highly coherent on visual inspection. For example, the model learned a schema about feelings including the words *feeling*, *feels*, *felt*, *pain*, *worse*, and *bad*. We quantified this semantic coherence by training a second model but ablating the semantic information by randomly assigning words to documents in the Reddit corpus. We asked human raters to judge which model generated more semantically coherent topics. Raters judged the topics from the real model as more semantically coherent than the

semantically ablated model, $t_{(22)} = 11.68$, $p < 0.001$, a difference that was observed in every individual rater (23/23 raters).

We next asked whether these schemas fill in missing information in a sentence. We trained a neural network to predict the words in people's talk in the blog corpus based on either the real schemas, or based on ablating schema knowledge using randomly generated schemas. As shown in Fig. 4A, we found that the model based on real schemas outperformed the model based on scrambled schemas, $t_{(220,174)} = 174.26$, $p < 0.001$, suggesting that these schemas do indeed fill in missing information.

Finally, we asked whether these schemas fill in more information for references to the past, present, or future. We found that past references drew on schemas, evidenced by increased prediction performance for past relative to present thoughts, $t_{(110,952)} = 4.60$, $p < 0.001$ (Fig. 4B). However, we found that future references drew more on schemas than did past references, $t_{(175,248)} = 18.53$, $p < 0.001$ (Fig. 4B). This increased prediction for future thoughts relative to past thoughts suggest that thoughts about the future rely more on schemas than thought about the past. This result is consistent with the predictions of the Complementarity Hypothesis.

Study 3: Are these Findings the Result of Different Processes?

Studies 1-2 suggest differences in the cognitive processes used for past and future thinking. However, these results are also open to an alternative interpretation, which we may call the Difference-in-Amount view. On this account, past and future thinking rely on the same basic cognitive processes, but to different extents.

Testing between the Difference-in-Amount and Complementarity Hypothesis requires an analysis for determining whether two operations reflect different underlying cognitive processes. Our key idea is that such a procedure exists in cognitive neuroscience, and can be adapted to big data. In fMRI studies, it is widely accepted that there are different cognitive processes if the two processes activate non-overlapping patterns of voxels in the brain. Indeed, the spatial overlap between past and future thinking in the brain has been taken as evidence for common processing. While this analysis is based on a brain space, a similar logic should hold for operations projected into what we will call a cognitive process space. As shown in Fig. 5, a process space can be created by projecting the candidate operations into a space composed by two or more cognitive processes. Evidence for a single process would be largely overlapping representations in process space (Fig. 5A), while evidence for multiple processes would be largely non-overlapping representations in process space (Fig. 5B). To evaluate the Difference-in-Amount and Complementarity Hypotheses, we pooled the data from Studies 1 and 2 to create a cognitive process space defined by constructive and episodic processing. We projected both past and future thinking into the process space, and quantified the amount of overlap between the processes. We asked whether this overlap was better explained by the Difference-in-Amount view or the Complementarity view.

Methods

Materials We combined the data from Studies 1-2. We created an aggregate measure of episodic processing by separately *z*-scoring the concreteness, spatial relation, and perceptual measures and then averaging the resulting *z*-scores for each sentence.

Process Space Creation We created a 10x10 cognitive process space using the episodic and constructive processing scores. For each measure we calculated 10 deciles. For example the bottom-left corner represents 0-10% episodic processing and 0-10% constructive processing. We then calculated the proportion of past and future thoughts falling in each region of process space. We stored the difference score (future - past) for each region and retained only scores larger than 0.25 in magnitude to avoid false positives.

Cluster Permutation Test We next tested how large a cluster would be obtained in the process space by chance. We did this by creating 10,000 permutations of the data by shuffling the past and future labels. In each permutation we repeated the cognitive process analysis and stored the size of the largest cluster, again retaining only scores large than 0.25 in magnitude. We used these cluster sizes to create a chance distribution (Fig. 5D, green distribution).

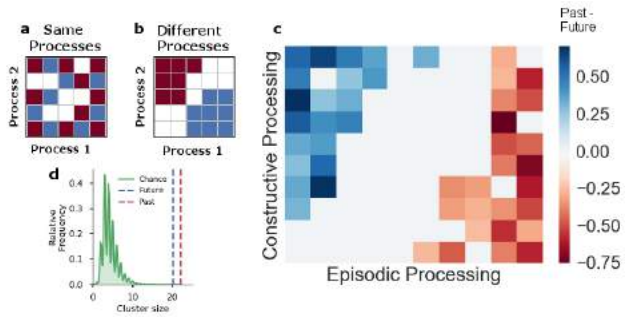


Fig. 5. Cognitive Process Space. (a-b) Hypothetical results in cognitive process space that would indicate reliance on the same or different processes. (c) Past thinking (blue) and future thinking (red) projected into cognitive process space. (d) Chance distribution of clusters in process space (green) compared to observed cluster sizes (vertical dotted lines).

Results and Discussion

We found that past and future thinking occupied largely non-overlapping clusters of the process space, supporting the predictions of the Complementarity hypothesis.

As shown in Fig. 5C, when past and future thinking were projected into cognitive process space, they occupied largely non-overlapping regions of the space. As shown in Fig. 5D, we quantified whether this pattern would be expected due to chance. We did this by creating 10,000 random permutations of the data, and recording the largest cluster size observed in each permutation (Fig. 5D, green distribution). We found that both the past thinking cluster (red dotted line) and the future thinking cluster (blue dotted line) were larger than those

observed in any of the 10,000 permutations, suggesting dissociable cognitive processes that would not be likely observed due to chance (e.g. $p < 0.0001$). This result suggests that past and future thinking rely on different cognitive processes, consistent with the Complementarity Hypothesis.

General Discussion

There is growing consensus that memory is not just for remembering the past, but also for imagining the future. Here, we considered a strong version of this idea that past and future thinking could rely on largely similar cognitive processes. In a series of three studies based on people's natural talk about time, we found support for the alternative hypothesis that past and future thinking rely on different cognitive processes. In Study 1, we found that thoughts about the past were more episodic than thoughts about the future, as revealed by the increased presence of concrete words, perceptual words, and spatial relation words. In Study 2, we found that thoughts about the past were less constructed than thoughts about the future, as revealed by the decreased ability of a machine learning model to use the topics of people's writing to predict the contents of future references compared to past references. Finally, in Study 3 we found that these findings were better explained by differences in cognitive processing than by a Difference-in-Amount view, a conclusion supported by projecting the data into cognitive process space.

While we believe that the schemas learned by our model are quite general, a limitation of the current analysis is that the schemas are only derived from a single social media corpus. The social media corpus spans a broad range of topics and covers millions of posts, but it may be limited in some ways; for example, social media users may be younger than or more likely to be male than the general population (Duggan & Brener, 2013). Future work should address whether similar schemas would be discovered in other kinds of corpora.

Beyond future thinking, our results have implications for the role of big data in psychology. It has previously been shown that big data can predict many psychological traits, including personality (Youyou, Kosinski, & Stillwell, 2015), mental illness (Thorstad & Wolff, 2018b), and decision-making (Thorstad & Wolff, 2018a). However, psychologists are often interested in going beyond prediction to make inferences about the underlying cognitive processes. It is not obvious that cognitive processes are recoverable from big data, since in written text the cognitive processes that generated the text have already occurred. Our findings suggest that big data can in fact recover cognitive processes, in two ways. First, big data can be used to look for characteristic representations of a cognitive process, such as the episodic language markers in Study 1. Second, big data can be used to train a model to mimic the cognitive process used to generate the text, as in the schema-based prediction model in Study 2. Both of these techniques suggest that big data may be useful not just for predicting human psychology, but also for understanding cognitive processes, a kind of data mining the mind.

References

Addis, D., Wong, A., & Schacter, D. (2007). Remembering the past and imagining the future. *Neuropsychologia*, *45*(7), 1363-1377.

Atance, C. & O'Neill, D. (2001). Episodic future thinking. *Trends in cog. sci.*, *5*(12), 533-539.

Benoit, R. & Schacter, D. (2015). Specifying the core network supporting episodic simulation and episodic memory. *Neuropsychologia*, *75*, 450-457.

Bernsten, D. & Bohn, A. (2010). Remembering the forecasting: the relation. *Memory & Cog.*, *38*(3), 265-278.

Boyer, P. (2008). Evolutionary economics of mental time travel. *Trends in Cog. Sci.*, *12*(6), 219-224.

Bransford, J. & Johnson, M. (1972). Contextual prerequisites for understanding: some investigations of comprehension and recall. *Journal verbal learn. & behav.*, *11*(6), 717-726.

Brysbaert, M., Warring, A., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior res. meth.*, *46*(3), 904-911.

Chang, A. & Manning, C. (2012). Sutime: a library for recognizing and normalizing time expressions. In *Proceedings of LREC 2012*.

Chen, D. & Manning, C. (2014). A fast and accurate dependency parser using neural networks. In *Proceedings of EMNLP 2014*.

Craver, C., Kwan, D., Steindam, C., & Rosenbaum, R. (2014). Individuals with episodic amnesia are not stuck in time. *Neuropsychologia*, *57*, 191-195.

Dugan, M. & Brenner, J. (2013). The demographics of social media users – 2012. *PEW Research Center*.

Goldstone, R. & Lupyan, G. (2016). Discovering psychological principles by mining naturally occurring datasets. *Topics in cog. sci.*, *8*(3), 548-568.

Irish, M., Addis, D., Hodges, J., & Piguet, O. (2012). Considering the role of semantic memory in episodic future thinking. *Brain*, *135*(7), 2178-2191.

Klein, S., Loftus, E., & Kihlstrom, J. (2002). Memory and temporal experience: the effects of episodic memory loss on a patient's ability to remember the past and imagine the future. *Social Cognition*, *20*(5), 353-379.

Klein, S., Robertson, T., & Delton, A. (2010). Facing the future: memory as an evolved system for planning future acts. *Memory & cognition*, *38*(1), 13-22.

Levy, R. & Andrew, G. (2006). Tregex and Tsurgeon: tools for querying and manipulating tree data structures. In *Proceedings of LREC 2006*.

Pennebaker, J., Boyd, R., Jordan, K., & Blackburn, K. (2015). The development and psychometric properties of LIWC 2015.

Raschka, S. (2015). *Python Machine Learning*. Packt Publishing.

Russell, J., Alexis, D., & Clayton, N. (2010). Episodic future thinking in 3- to 5-year-old children. *Cognition*, *114*(1), 56-71.

Schacter, D., & Addis, D. (2007). The ghosts of past and future. *Nature*, *445*(7123), 27.

Schacter, D., & Addis, D. (2008). Episodic simulation of future events. *Annals of the New York Academy of Sciences*, *1124*(1), 39-60.

Schler, J., Koppel, M., Argamon, S., & Pennebaker, J. (2006). Effects of age and gender on blogging. *AAAI*, *6*, 1919-205.

Schwartz, H. et al (2015). Extracting human temporal orientation from Facebook language. In *Proceedings of NAACL 2015*.

Spreng, R., Mar, R., & Kim, A. (2009). The common neural basis of autobiographical memory, prospection, navigation, theory of mind, and the default mode. *Journal of Cog. Neur.*, *21*(3), 489-510.

Thorstad, R. & Wolff, P. (2018a). A big data analysis of the relationship between future thinking and decision-making. *PNAS*, *115*(8), E1740-E1748.

Thorstad, R. & Wolff, P. (2018b). Using big data methods to identify conceptual frameworks. In T.T Rogers et al (Eds.) Proceedings of the 40th Annual Meeting of the Cog. Sci. Soc. Austin, TX: Cog. Sci. Society.

Tulving, E. (1993). What is episodic memory? *Current directions in psyc. sci.*, *2*(3), 67-70.

Viard, A. et al (2011). Mental time travel into the past and future in healthy aged adults. *Brain & Cog.*, *75*(1), 1-9.

Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *PNAS*, *112*(4), 1036-1040.

Children's causal inferences about past vs. future events

Katharine A. Tillman
Department of Psychology
The University of Texas at Austin
ktillman@utexas.edu

Caren M. Walker
Department of Psychology
University of California, San Diego
carenwalker@ucsd.edu

Abstract

Causal and temporal reasoning are fundamentally linked, but few studies have directly examined how the ability to make causal inferences about the past vs. the future develops. We used a counterfactual reasoning task to explore 4- to 6-year-old children's understanding of the causal relationships among past, present, and future events. Like adults, even 4-year-olds judged that future, but not past, events could be altered by interventions in the present. This early sensitivity to the causal asymmetry between the past and future became more pronounced with age. We also found that children and adults selectively and appropriately use evidence about the present to make inferences about past events. Implications for theoretical accounts of the development of causal reasoning and abstract concepts of time are discussed.

Keywords: cognitive development; temporal cognition; causal inference; counterfactual reasoning

Introduction

You can change the future, but you can't change the past. This fundamental distinction between the past and future is central to an abstract, linear concept of time, and has profound effects on adults' everyday behavior. Although philosophers and physicists have argued about the ultimate reality of the past/future asymmetry, many of us find it difficult, if not impossible, to conceive of a world without it. Is the past/future distinction a "built in" feature of human cognition? If it isn't, when and how does it develop? While we know that children's reasoning about both temporal and causal relationships improves during the preschool years, few studies have directly explored the relationship between children's reasoning about causality and their knowledge of the ontological distinction between past and future (see McCormack & Hoerl, 2017). Here, we use a counterfactual[1] reasoning task to explore how children use information about present events to make causal inferences about the past and future.

Despite the central role of the past/future distinction in expliciting temporal reasoning, it remains unclear to what extent young children possess this understanding. While it is difficult to test this in preverbal infants, researchers have looked to children's earliest production of temporal language for clues. The past-tense verb marking *-ed* is one of the first grammatical inflections English-speaking children produce, usually at or before age 2 (Brown, 1973), which has been taken as evidence that understanding of the past/future status of events relative to the present develops early. However, there is debate in the language acquisition literature over how accurate and generalizable children's early uses of tense are, and particularly whether they may indicate perfective aspect, rather than event time (Anderson & Shirai, 1996). Deictic time words like "tomorrow" and "yesterday" are also early to appear in the child's lexicon, though children don't use them reliably for several years (e.g., Tillman et al, 2017). Nonetheless, while these studies suggest early onset and prolonged development of past-future reasoning, it remains possible that children's understanding the causal asymmetry of the past and the future develops prior to the ability to express these differences in language.

When do children understand how causality operates over time? Suggesting that even infants intuitively understand the relationship between temporal order and causality, 4-month-olds look longer when presented with impossible causal chains of events, including those with apparent breaks in temporal continuity (Cohen et al., 1998). Nevertheless, recognizing the temporal-causal structure of a simple event, like one ball striking another, does not imply that infants have a concept of the past or the future. Later in development, when asked which of two possible events caused another event to happen, 3-year-olds chose the prior rather than the subsequent one (Bullock and Gelman, 1979). Four-year-olds recognize that past, but not future, events determine present mental states and states of the world (Busby Grant & Suddendorf, 2010). However, 4-year-olds struggle to use information about the relative ordering of multiple past events to make inferences about the present, and fail to solve temporal reasoning tasks in which the order they receive information about events doesn't match the order in which those events occurred (e.g., McCormack & Hoerl, 2007), suggesting that young children lack flexible temporal perspective-taking.

[1]The precise definition of counterfactual reasoning, and thus the age at which children are first capable of it, is the subject of much debate (e.g., Beck, 2016; Weisberg & Gopnik, 2013). In the present paper, we take a broad view of counterfactual reasoning, which incompasses hypothetical questions about the past, present, and future, as well as conditionals.

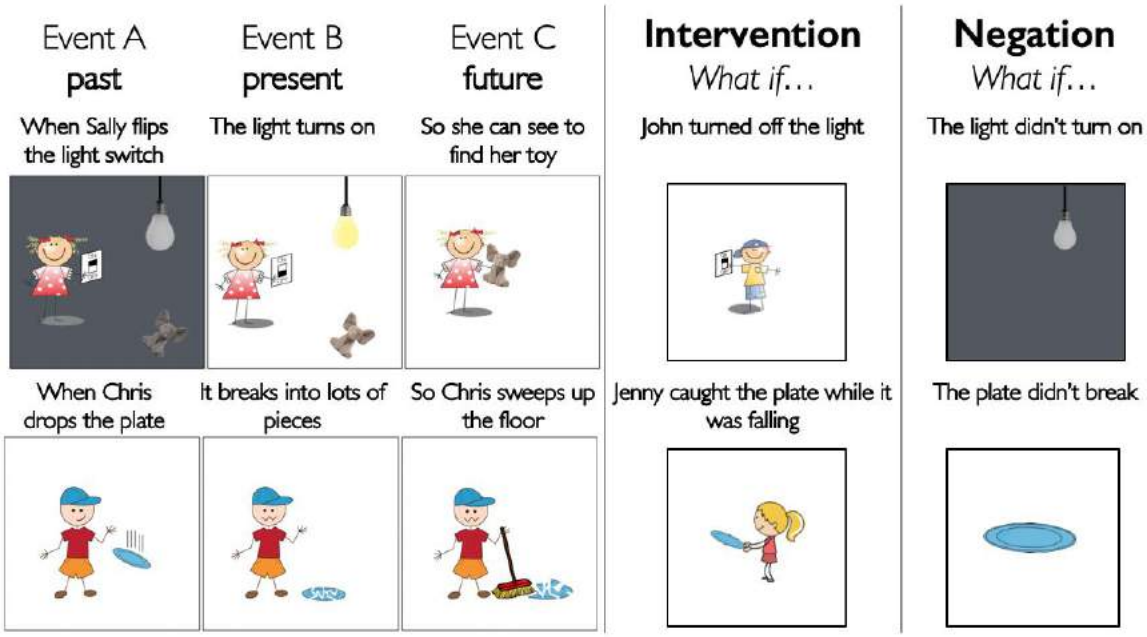


Figure 1: Example storyboards and intervention/negation cards. The experimenter (E) recited the story on the front of the card, then flipped it to reveal three empty boxes. At test, E placed either an intervention or negation card in the center square (B) before posing counterfactual questions about its effects on past (A) and future (C) events.

Critically, none of these prior studies directly test whether children understand that the future is alterable in the present, but the past is not. Instead, a separate literature has explored the development of children's causal reasoning skills (see Gopnik & Wellman, 2012 for a review), and, despite the related subject matter, this literature has developed largely independently from the work on the development of temporal cognition discussed above. Within the causal reasoning literature, some accounts suggest that causal relationships are defined in terms of their counterfactual dependency. That is, if event A causes event B, then an intervention on A will lead to a change in B (Pearl, 2000). Relevant to this, researchers have examined children's understanding of this link between causal and counterfactual reasoning. For example, in early work, Harris, German, and Mills (1996) conducted a series of studies in which 3- and 4-year-olds were presented with scenarios in which they were asked to reason about a short causal sequence (e.g., a character walks across the floor with muddy boots [A], making a mess [B]). When asked conditional questions about what would have happened had A not occurred (e.g., "What if Carol had taken her shoes off? Would the floor be dirty?"), children made accurate judgements about effects on B.

If children have a unidirectional view of causality, given a 3-step causal chain of events, A → B → C , they should judge that an intervention at B can alter future event C, as previous studies have found. Importantly, however, they should also judge that the intervention will *not* alter past event A (Sloman & Lagnado, 2005).

Consider the following scenario (Figure 1): *When Sally flips the lightswitch, then the light turns on, so she can see to find her toy.* If told that another character, John, turned off the light[2] (at B), adults may reasonably predict that Sally will no longer be able to see (at C). However, they should not infer that Sally never turned the light on in the first place (at A), because John's actions at time B can't change what happened in the past, at time A. Here we test whether 4- to 6-year-old children make the same inferences.

Retrospective reasoning

Despite understanding time itself to be linear, under some circumstances, adults use information about the present to reason "back in time" and make inferences about what already happened in the past. For example, if an expected event does not occur—and no other explanation for this is

[2] Note that under some definitions of counterfactual reasoning, our use of the simple past tense here, rather than pluperfect subjunctives such as "What if John *had turned off* the light?" or (in the negation condition) "What if the light *hadn't turned on*?", indicates that these statements are hypothetical rather than counterfactual *per se* (see Lucas & Kemp, 2015). We chose simpler language primarily to make the task more comprehensible to young children, but are currently exploring whether this tense modification impacts performance on our task.

given—an adult might reasonably infer that the event's usual cause must not have occurred. For example, when simply told that the light *didn't* turn on (at B; see footnote 2), an adult might indeed conclude that Sally never flipped the lightswitch (at A). Here, we presented children and adults with stories involving 3-step causal chains, and then asked them to consider scenarios in which the second step (B) was different. Importantly, we asked both about the effects of the "present" change on the future (C) and on the past (A).

Given prior work showing that adults generate different causal predictions following passive observation (e.g., observing that B did not occur) than they do following interventions (e.g. acting on B to prevent it from occurring; Sloman & Lagnado, 2005; Waldmann and Hagmeyer, 2005), we also varied this feature in the current study. In the *intervention* condition, an external agent (e.g., another character) caused the "present" change. We hypothesized that participants with a linear concept of time would judge that the future event would also change, but not the past event. In the *negation* condition, however, no explanation for the present change was given. Here we hypothesized, again, that participants with a linear concept of time would judge that the future would change. If participants also engage in retrospective causal reasoning, we hypothesized that, unlike in the intervention condition, they would also systematically judge that the past event (A) had changed (e.g., Sloman & Lagnado, 2005). In contrast, if participants do not reason retrospectively, we predicted that they would perform similarly in the two conditions.

Method

Participants

A total of 258 subjects participated, including 65 4-year-olds (M_{age} = 4.5 years, range = 4.0-5.0 years), 70 5-year-olds (M_{age} = 5.5 years, range = 5.0-6.0 years), 63 6-year-olds (M_{age} = 6.4 years, range = 6.0-7.0 years) and 60 adult controls (M_{age} = 21.6 years, range = 18.2-31.1 years). Participants were pseudo-randomly assigned to either the Intervention or Negation condition. An additional 43 children participated, but were not included in analyses due to being outside the target age range (n = 3), experimenter or technical error (n = 5), failure to complete the task (n = 4), developmental delay (n = 2), insufficient fluency in English (n = 1), incomplete age information (n = 2), or failing more than one control trial, as described below (n = 26).

Materials

Study materials included eight 3-panel storyboards illustrating sequences of events from left-to-right. Two examples are shown in Fig 1. Each panel was 2.8 in. × 2.8 in. Single images corresponding to event B in each story were also used in testing, which represented either identical pictures (control stories), interventions, or negations, depending on condition. Each individual image was square with a black outline, and on the reverse side of each storyboard were three empty black squares positioned like the filled images on the front of the card.

Procedure

Children were tested one-on-one, in a quiet room with the experimenter. The experimenter began the session by placing the first storyboard in front of the participant, saying "I'm going to tell you some stories. There are three things that happen in each story, see?" She then pointed to each image in the story while reciting the corresponding part of the narrative, in this case, "When [A] Julie opens the door, then [B] her dog runs outside, so [C] he smashes up all the flowers in the garden".

The experimenter then flipped over the storyboard, revealing the 3 empty boxes, and initiated a demonstration control trial. While placing a duplicate of the center image from the front of the card in the empty center square, she asked the child "tell me, *just like in that story*, if [the dog ran outside]...," and then pointed to the empty third [event C; future] box while completing the question with a forced choice: "will [he smash all the flowers in the garden] or not [smash the flowers in the garden]?" After receiving a verbal response from the participant, the experimenter repeated the procedure again, instead pointing to the first [event A; past] box and asking, "did [Julie open the door] or not [open the door]?"

Next, the experimenter flipped the card back over, repeated the original story, and explained that she would now be asking the participant to think about what would happen in the story if something had been *different*. In this demonstration critical trial, the experimenter placed a modified image B in the empty center square on the back of the card. This image showed either the **intervention, "**What if the dog were on a leash and couldn't get out?", or the **negation,** "What if the dog *didn't* run outside?", according to condition. The test concluded with the past and future test questions, as above. No feedback was given on either the control or critical trials using the demonstration story.

After this demonstration phase, the experimenter told the participant that she would tell some other stories, sometimes asking if things had been the same (control stories), and sometimes asking if things had been different (critical stories), and sometimes asking about the first part or the story, and sometimes about the last part. The remainder of the task included 7 new stories, 5 of which were used on critical trials (see examples in Fig. 1) and 2 were control stories. The ordering of the stories (other than the demonstration story), the past and future questions about each story, and the positive and negative response options in each question were counterbalanced across participants. The third and sixth stories were always control stories.

Procedures used in the intervention and negation conditions were identical, apart from the different counterfactual questions and corresponding images used during test.

Data from children who responded incorrectly more than one control trials, i.e., by denying that an event from the story they had just heard had occurred in that story, were excluded due to suspected incomprehension of the task. This exclusion criterion was particularly important because the predicted "adult-like" response pattern in the negation condition was one in which the participant judges that *none* of the events in either past or future critical trials had occurred. We therefore wanted to minimize the chances of potentially confusing a "no bias" in children who did not comprehend the task at all with adult-like conditional reasoning.

Coding

During testing, the experimenter recorded whether the participant affirmed or denied that each past or future event would occur. Yes responses were coded as 1, no responses as 0. These were later reverse-coded as described below. Participants who answered more than one of the four control questions incorrectly (i.e., by responding that the event did not occur; $n = 26$) were excluded from further analysis. Data from the demonstration story were not included in analysis. All analyses were conducted in R, using the *lme4* package for mixed-effects modeling.

Results

We began with two primary questions about our dataset: (1) Do participants differentiate past from future in the intervention condition? and (2) Do participants ever reason retrospectively (i.e., "back in time") when answering questions about the past in the negation condition?

Before addressing these, we asked whether children's performance differed between the two conditions. Because our DV was a binary choice (either an event would occur or not), we conducted a mixed-effects logistic regression. For ease of exposition, the data were reverse-coded, such that answers indicating that events would *not* occur in the counterfactual scenario were considered "changes" (1), while answers indicating that events would still occur were considered non-changes (0). We modeled the likelihood that a child[3] would say an event changed as a function of their age (continuous; between-subjects), condition (intervention vs. negation; between-subjects), and event time (past vs. future; within-subjects). We also included an interaction between event time and condition in this model, and random intercepts for subjects and stories. Results of this analysis revealed significant main effects of age ($\beta = 0.5$, $p = 0.004$) and event time ($\beta = -3.4$, $p < 0.001$) as well as a significant interaction between event time and condition ($\beta = 2.3$, $p < 0.001$). Given the evidence that children's behavior differed between conditions, we proceeded to analyze the data from the two conditions separately.

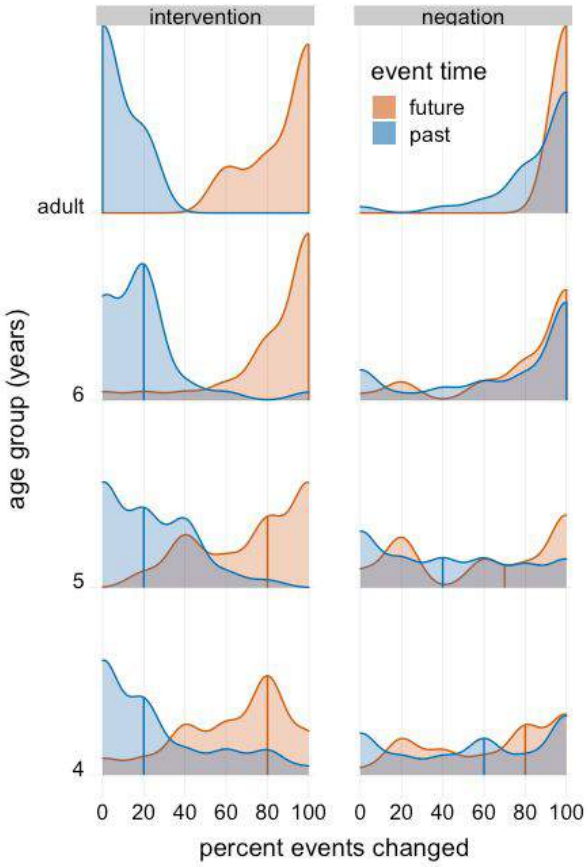


Figure 2: Distributions of responses to past (blue) and future (red) counterfactual questions, in the intervention (left) and negation (right) conditions. Height of shaded areas indicates the density of responses at each level of consistency, e.g., 80% = 4 of 5 events changed. Vertical lines = medians. Density calculation bandwidth = 8.

Intervention condition

Our goal in the intervention condition was to test whether participants differentiate the effects of present interventions on past vs. future events. In other words, do they know that you can change the future, but not the past? The distributions of responses to past- and future- questions, i.e., the percentage of target events that changed, for each age group are shown in the left column of Figure 2, with medians represented by vertical lines. As expected, and in line with prior work, adults strongly distinguished past from future: the median percentage of past events they said would change as a result of the intervention was 0%, 95% CI [0%-0%], while the median percentage of future events that would change was 100% [80%-100%].

[3] Adult controls were not included in this analysis.

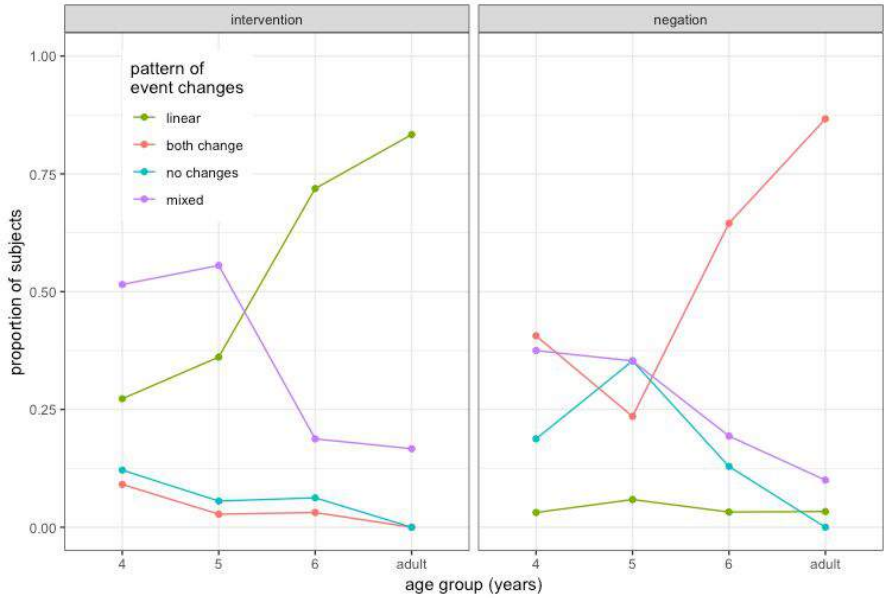


Figure 3: Proportion of subjects in each age group who demonstrated each of 4 response patterns across stories, in the intervention (top) and negation (bottom) conditions. "Linear" (green) = future events consistently judged to change, past events to stay the same. "Both change" (red) = both past and future events judged to change. No-changes (blue) = both past and future events were judged to stay the same. Mixed (purple) = inconsistent responses.

Considering the developmental data, a logistic regression of the children's likelihood of saying that an event changed revealed significant effects of age ($\beta = 0.7$, $p < 0.001$) and event time ($\beta = 2.5$, $p = 0.03$) as well as a significant interaction ($\beta = -1.1$, $p < 0.001$). As shown in Fig. 2, children were more likely to judge that interventions would change the future than the past, and this effect increased with age. Wilcoxon signed-rank tests confirmed that the past vs. future effect was significant even in 4-year-olds, who reported that 80% [60%-80%] of future events changed, but only 20% [0%-20%] of past events did ($W = 400$, $p < 0.001$).

Interestingly, the three groups of children did not differ in their likelihood of judging that past events changed (Wilcoxon rank-sum tests, 4's vs 6's, $W = 567$, $p = 0.6$), though even 6-year-olds were significantly more likely to do so than were adults ($W = 306$, $p = 0.006$). On future questions, 4- and 5-year-olds were significantly less likely to say that events changed than were 6-year-olds and adults (5's vs 6's, $W = 731$, $p = 0.04$), though neither of these pairs differed (4's vs 5's $W = 456$, $p = 0.09$; 6's vs adults $W = 489$, $p = 0.9$).

In addition to overall performance on past vs future questions, we were interested in the patterns of responses provided by individual subjects. For instance, did children who said the past wouldn't change also say the future *would* change, as a linear model of time would predict? For the purpose of this analysis, we operationalized a "linear" pattern as one in which the participant judged that at least 4 out of 5 future events would change after intervention, *and* that at least 4 of the 5 past events would not. As shown in Figure 3 (top panel, green line), we found that 83% [65%-94%] of adults conformed to this pattern, as did 72% [53%-86%] of 6-year-olds, 36% [21%-54%] of 5-year-olds, and 27% [13%-46%] of 4-year-olds. Subjects who didn't follow a linear pattern typically reported fewer than 4 changes to future questions, resulting in a mixed pattern. Patterns in which either both events or neither event changed were rare in all age groups.

Negation condition

In the negation condition we assessed whether participants would reason retrospectively, making the inference that an observed, unexplained change in the present was caused by a prior change in the past. As shown in Fig 2 (right column), we found strong retrospective reasoning in adults: when simply told that the present event "didn't" occur, they judged that the future effect would not occur on a median of 100%, 95% CI [100%-100%], of trials, and, in contrast to the intervention condition, that the past had changed on 100% [80%-100%] of trials, in line with prior adult work (e.g., Waldmann & Hagmeyer, 2005).

Next we considered the developmental data. A logistic regression model of the children's data in the negation condition, with the same effects structure as the one used in the intervention condition, revealed only a main effect of age ($\beta = 1.1$, $p = 0.04$). Older children were more likely than younger children to judge that events changed. However, unlike in the intervention condition, there was no significant effect of event time ($\beta = 0.35$, $p = 0.8$), and no interaction ($\beta = -0.29$, $p = 0.24$). In other words, we did not detect evidence that children were treating past and future events

differently in this condition. Examining the age effects further, we found that 5-year-olds were less likely to judge that past events had changed than 6-year-olds and adults (5's vs 6's, W = 302, p = 0.002), but no other age-group comparisons reached significance. On future questions, 4- and 5-year-olds could not be distinguished, but 6-year-olds were significantly more likely to say that future events changed than were 5-year-olds (W = 677, p = 0.04), and adults were more likely to do so than were 6-year-olds (W = 660, $p < 0.001$).

Importantly, retrospective reasoning about the implications of the present on the past, when combined with knowledge of how the present influences the future (i.e., future "prospective" reasoning), predicts not only that the past and future will *not* be differentiated, as we found above, but also that both will be judged to have changed. This pattern was less common across the 4- and 5-year-old groups, as can be seen in the flatter distributions (broader confidence intervals) in Fig 2. For example, the median percentage of past events that 4-year-olds judged to have changed was 60% [40%-80%], and for 5-year-olds was 40% [20%-60%]. In contrast, the median percentage for both 6-year-olds and adults was 100%. In our individual-subjects analysis, a consistent retrospective/prospective reasoning pattern was operationalized as one in which at least 4 or 5 past events and 4 of 5 future events changed. As shown in Fig 3, we found that 41% [24%-59%] of 4-year-olds, 24% [11%-41%] of 5-year-olds, 65% [45%-81%] of 6-year-olds, and 87% [69%-96%] of adults displayed this pattern, while linear response patterns were very rare in this condition. Interestingly, among children who did not show this pattern, particularly 5-year-olds (who were surprisingly less adult-like than 4-year-olds), a larger proportion said that *neither* event changed (Fig. 3, blue lines) than we observed in the intervention condition. We discuss this further below.

Discussion

In the current study, we explored the development of children's reasoning about causal relationships among events in the past, present, and future. We discovered that children as young as 4 already distinguish the past and future: they are more likely to judge that an intervention in the present will change a future event than a past one. To our knowledge, this is the strongest evidence to date that pre-school children appreciate the causal asymmetry between the past and future (see McCormack & Hoerl, 2017). Moreover, children treated counterfactual scenarios with an explicit causal agent differently from those in which the cause must be inferred. In the latter case, children did not distinguish past from future, and by age 6, 65% of children consistently demonstrated both prospective and retrospective causal reasoning.

To test their reasoning about past and future events, we told participants 3-step stories, and then asked them to consider counterfactual cases in which an outside agent disrupted the middle step. We found that even 4-year-olds very rarely judged that the past event would retroactively change. This finding extends previous literature showing that 4-year-olds understand that past (but not future) events can cause present ones (e.g., Busby & Suddendorf, 2010), and suggests that the understanding that time is irreversible is strong and early-developing. However, despite the high overall rate of denials that the past would change, we only found a consistently "linear" response pattern in about a quarter of 4-year-olds. This was because, compared to older children, younger children were less likely to judge that *future* events would change after intervention.

Finding more adult-like behavior from children on past than future trials is somewhat surprising in light of previous studies showing that 3-year-olds are capable of prospective ("forward") conditional reasoning (e.g., Harris et al., 1996). In fact, it has been proposed that future conditionals are easier than past counterfactuals for children, because they do not require them to hold both the real world, i.e., how things actually occurred, and the possible world, i.e., how things could have been otherwise, in mind simultaneously (e.g., Beck, 2016; Beck & Riggs, 2014; Raefsteder et al. 2010). Perhaps, however, children in our task were more variable in their predictions about the future than the past simply because the future is intrinsically more open-ended. In linear time, a given intervention may or may not be effective at generating a particular outcome, but will *never* change what has already occurred.

Although children consistently denied that the past would change in the intervention condition, their judgments about past events were not rigid. In the negation condition, children's responses to past questions were more mixed, and like adults, they were more likely to say that the past *had* changed than that it hadn't. Given the minimal changes to the task across conditions, our finding that children are already sensitive to the precise nature of the conditional statement, and to the increased ambiguity of negations relative to explicit interventions (with respect to the past) is striking. Given children's high performance in the intervention condition, and the lack of a bell curve centered around random responding in the negation condition, we do not believe these results can be attributed to confusion about the nature of the task. Instead, these findings may suggest that some children (but not others) are already able to reason backward in time.

One intriguing possibility is that children who perform like adults in the negation condition are deploying what the adult counterfactual reasoning literature has termed "*backtracking*" (i.e., engaging in a special type of counterfactual reasoning that involves inference about upstream causal variables; Gerstenberg, Bechlivanidis, & Lagnado, 2013; Rips, 2010; Rips & Edwards, 2013; Sloman & Lagnado, 2005). Although there has been substantial debate over what types of counterfactuals lead to backtracking inferences in adults (e.g., Han et al., 2014),

these investigations have not yet been extended to children. The tendency to engage in backtracking (or not) has important implications for interventionist accounts for causal reasoning. Specifically, because evaluating the effects of an intervention on a given variable requires "cutting off" that variable from its upstream causal antecedents, backtracking should not be possible (Pearl, 2000; see Lucas & Kemp, 2015). While our negation condition is similar to certain backtracking tasks previously used with adults (e.g., Han et al., 2014; Lucas & Kemp, 2015), given the methodological differences (e.g., using child-friendly events that operate over time; presenting interventions/negations in past vs. pluperfect tense), additional work will be required to explore this potential developmental link.

In sum, the current study brings together the literatures on the development of causal reasoning and temporal cognition, by leveraging a counterfactual reasoning task to explore children's understanding of the past and the future. We found that children are able to recognize the causal asymmetry between past and future prior to the age of 4, reflecting the early development of a linear view of time.

Interestingly, it has been hypothesized that counterfactual reasoning itself may hinge on the development of an abstract, event-independent concept of time (McCormack & Hoerl, 2017). To consider different possible worlds, one must separate the time-point at which an event occurred from the event itself. Linear time thus provides a framework in which events can be organized and even mentally "switched out," so that their causal consequences can be considered. By studying these phenomena in tandem, future studies may uncover new insights about how time and causality are mentally represented.

References

Anderson, S. W. & Shirai, Y. (1996). The primacy of aspect in first and second language acquisition: the pidgin-creole connection. In W. C. Ritchie & T. K. Bhatia (eds), Handbook of Second Language Acquisition. London: Academic Press.

Beck, S. R. (2016). Counterfactuals matter: A reply to Weisberg & Gopnik. *Cognitive science*, *40*(1), 260-261.

Beck, S. R., & Riggs, K. J. (2014). Developing thoughts about what might have been. *Child development perspectives*, *8*(3), 175-179.

Brown, R. (1973). A first language: The early stages. Harvard U. Press.

Bullock, M., & Gelman, R. (1979). Preschool children's assumptions about cause and effect: Temporal ordering. *Child Development*, 89-96.

Busby, J. G., & Suddendorf, T. (2010). Young children's ability to distinguish past and future changes in physical and mental states. *British Journal of Developmental Psychology*, *28*(4), 853-870.

Cohen, L. B., & Amsel, G. (1998). Precursors to infants' perception of the causality of a simple event. *Infant behavior and development*, *21*(4), 713-731.

Gerstenberg, T., Bechlivanidis, C., & Lagnado, D. A. (2013). Back on track: Backtracking in counterfactual reasoning. In M. P. Knauff, M., N. Sebanz & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Gopnik, A., & Wellman, H. M. (2012). Reconstructing constructivism: Causal models, Bayesian learning mechanisms, and the theory theory. *Psychological bulletin*, *138*(6), 1085.

Han, J. H., Jimenez-Leal, W., & Sloman, S. (2014). Conditions for backtracking with counterfactual conditionals. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 36, No. 36).

Harris, P. L., German, T., & Mills, P. (1996). Children's use of counterfactual thinking in causal reasoning. *Cognition*, *61*(3), 233-259.Lucas, C. G. & Kemp, C. (2015). An improved probabilistic account of counterfactual reasoning. *Psychological Review, 122*(4), 700-734.

McCormack, T., & Hoerl, C. (2007). Young children's reasoning about the order of past events. *Journal of experimental child psychology*, *98*(3), 168-183.

McCormack, T., & Hoerl, C. (2017). The development of temporal concepts: Learning to locate events in time. *Timing & Time Perception*, *5*(3-4), 297-327.

Pearl, J. (2000). Causality: models, reasoning, and inference. Cambridge University Press, Cambridge, UK.

Rafetseder, E., Cristi-Vargas, R., & Perner, J. (2010). Counterfactual reasoning: Developing a sense of "nearest possible world". *Child development*, *81*(1), 376-389.

Rips, L. J. (2010). Two Causal Theories of Counterfactual Conditionals. *Cogn. Sci., 34*(2), 175-221. Sloman SA, Lagnado DA. (2005). Do we "do"? *Cogn. Sci.* 29(1):5–39

Rips, L. J., & Edwards, B. J. (2013). Inference and explanation in counterfactual reasoning. *Cognitive Science*, *37*(6), 1107-1135.

Sloman, S., & Lagnado, D. A. (2005). Do people 'do'. *Cogn. Sci.*, *29*, 5-39.

Sloman, S. A., & Lagnado, D. (2015). Causality in thought. *Annual review of psychology*, *66*, 223-247.

Tillman, K. A., Marghetis, T, Barner, D., & Srinivasan, M. (2017). Today is tomorrow's yesterday: Children's acquisition of deictic time words. Cognitive Psychology, 92, 87-100.

Waldmann , M. R. , & Hagmayer , Y. (2005). Seeing versus doing: Two modes of accessing causal knowledge . *Journal of Experimental Psychology: Learning, Motivation, and Cognition, 31* , 216–227.

Weisberg, D. S., & Gopnik, A. (2013). Pretense, counterfactuals, and Bayesian causal models: Why what is not real really matters. *Cognitive science*, *37*(7), 1368-1381.

Explanation Versus Prediction: Statistical Differences in Detecting Fraudulent Events Do Not Necessarily Have Predictive Power

Angelica M. Tinga (A.M.Tinga@uvt.nl)

Welmoed Kuperus (Welmoed.Kuperus@gmail.com)

Maira B. Carvalho (M.BrandaoCarvalho@uvt.nl)

Max M. Louwerse (M.M.Louwerse@uvt.nl)

Tilburg University, Department of Cognitive Science and Artificial Intelligence, Tilburg, The Netherlands

Abstract

A large body of research in the cognitive sciences relies on examining statistical differences. While the approach of examining differences can aid in explaining behavior, it does not necessarily mean that these differences have predictive power. Yet, understanding behavior both involves explaining and predicting behavior. As a point in case, the current study used a naturalistic email dataset to examine statistical differences and predictive power in fraudulent activities. Differences between 1st and 3rd person pronoun use in liars and people telling the truth are widely reported in the literature. The current study aimed to test for the effect of fraudulent events on pronoun use in emails using the Enron corpus and additionally applied a machine learning approach to estimate whether pronoun use predicts fraud. While the ratio between 1st and 3rd person pronoun use was related to fraud, this construct did not have predictive power. The current study highlights an important conclusion for the cognitive sciences: The importance of not only testing for differences, but of also applying predictive models. In this way it can be determined whether effects of a construct on an outcome can also predict the outcome.

Keywords: corpus linguistics; machine learning; deception; pronouns

Introduction

Many studies in the cognitive sciences rely on examining statistical differences. This approach provides us with important knowledge about differences in for example behavior between extroverts and introverts (Lu & Hsiao, 2010), clinical populations and non-clinical ones (Garnefski et al., 2002) and males and females (Bleidorn et al., 2016). While examining differences can aid in explaining behavior, it does not necessarily mean these differences have predictive power. Understanding behavior both involves explaining and predicting behavior (Rosenberg et al., 2018). A model focused on explanation could be appealing theoretically, but could be very limited in predicting actual human behavior (Yarkoni & Westfall, 2017).

One field of study in which differences have been widely examined is that of deception, in which comparisons are made between when people are lying and when they are telling the truth (DePaulo et al., 2003). Lying is cognitively more complex than telling the truth. To make a lie convincing we have to exert a lot of cognitive control, which might paradoxically be reflected in cues that betray our deception (Zuckerman et al., 1981), both verbally and non-verbally (DePaulo et al., 2003).

Several studies have examined these cues to deception using experimental manipulations, for example by asking participants to lie or to tell the truth, testing for a statistical difference between the manipulations. These studies demonstrated that there is a difference between liars and people that tell the truth: Liars provide fewer details and tell fewer compelling stories, as they are uncertain and less engaged (DePaulo et al., 2003). Liars apparently try to distance themselves from the content of the communication, with content increasing in abstractness (Louwerse et al., 2010). Abstractness in communication may be reflected in pronoun use (Hancock et al., 2008; Humpherys et al., 2011; Louwerse et al., 2010; Newman et al., 2003), with a decrease in self-references and an increase in other-references reflecting increasing abstractness. Even though experienced liars may be avoiding tainted words that reveal their intentions, pronoun use is outside of conscious control of speakers and writers and therefore a useful measure to determine whether statements are truthful or not (Pennebaker, 2011).

Newman et al. (2003) examined 1st person pronouns (self-references) and 3rd person pronouns (other-references) when participants were instructed to produce a story on abortion which matched their opinion or not. They demonstrated that participants who wrote a story they did not agree with used fewer 1st person singular and fewer 3rd person pronouns than participants that agreed with their story. Similarly, Hancock et al. (2008) asked participants to either write a truthful or untruthful story on several different topics. The participants that were untruthful used fewer 1st person pronouns and more 3rd person pronouns than truthful participants. A meta-analysis on 116 studies on lying and deceptive cues by DePaulo et al. (2003) also demonstrated that there is an effect of being truthful on pronoun use, with fewer self-references and more other-references showing up in liars. However,

Louwerse et al. (2010) found that fraudulent events were associated mostly with an increase in 1st person pronouns.

In sum, statistical differences between pronoun use in liars and people that are truthful have been established in the existing literature, although the direction of these effects varied across studies. Perhaps this is not entirely surprising, as the context and the ecological validity of these studies also differ. Most studies on pronoun use and deception induced lying with an experimental manipulation, by for instance asking participants to write about an opinion opposite to what they truly belief. These cases are considered to be deception (Newman et al., 2003), except that there is no consequence to participants' 'lying'. Such laboratory studies provide excellent insights in linguistic deceptive cues but lack ecological validity.

To use a case where the stakes of deception were higher than a manipulated laboratory setting, Louwerse et al. (2010), used an email dataset (Klimt & Yang, 2004), which contained 517,431 emails from about 150 Enron executives and employees from 1999 to 2001. The Enron Cooperation was one of the world's leading gas, electricity, and communication companies and is most famous for the elaborate and systematic way in which accounting fraud spread throughout the organization, which led to declaration of bankruptcy in 2001. The advantage of using this corpus is that, besides its ecological validity, it covers a relatively large time span and it has detailed information available on the company and its fraudulent activities (Diesner et al., 2005). The disadvantage of using a naturalistic corpus, however, is that it is very difficult to determine which emails actually contain deception and which ones do not. Louwerse et al. (2010) operationalized deception by identifying the periods during which fraudulent events took place, capitalizing on the sheer number of emails in these different time frames.

Although statistical differences show up between liars and non-liars in pronoun use and although this difference is theoretically making sense, it is not clear whether pronoun use allows for predicting deception, and if so, to what extent. The current study uses Louwerse et al. (2010) as an illustration. We used the Enron email dataset, but rather than only investigating whether there is an effect of fraudulent emails on linguistic variables as in Louwerse et al. (2010), we additionally applied a machine learning approach to estimate whether linguistic variables also predicted fraud. Moreover, rather than taking a large number of linguistic variables, we applied the principle of parsimony and only focused on pronoun use.

Table 1: Overview of events within the Enron Cooperation from 2000-2001. Marked events are considered fraudulent. Adapted from Louwerse et al. (2010) p. 964.

Event	Description of event	Date (month-year)
- Layoffs	Employees within Enron Corporation were laid off.	12-01
- CEO	Indicating involvement of the CEO within any coded event.	3-00, 08-00, 11-00, 01-01, 04-01, 08-01, 10-01, 11-01
- Fraudulent paperwork filled signed	Filing and/or signing of fraudulent paperwork (by the CEO or COO).	03-00, 08-00
- Fraudulent comments	Enron made fraudulent comments, to the employees and/or investors.	01-01, 09-01
- Discussion of ethics	A discussion of ethics occurred between Enron executives or between the CEO and employees.	07-00, 03-01, 05-01, 08-01, 09-01, 10-01
- Selling Enron shares	Selling of Enron stock by high-level executives occurs.	11-00, 05-01, 06-01, 07-01, 08-01, 09-01
- Rolling blackouts initiated	Intentional initiation of rolling blackouts in California.	01-01
- Meetings with national political figures	High-level Enron executives met with national political figures including the Secr. of the Treasury and the Secr. of Commerce.	02-01, 03-01, 04-01, 08-01, 10-01, 11-01
- Financial support of political candidate	High-level Enron executives (CEO & President) provided financial support for a newly elected national political figure.	01-01
- Profit announced	Profits were announced for the quarter.	04-01
- Loss announced	Losses were announced for the quarter.	10-01
- SEC inquiry developments	Beginning of the SEC inquiry and the point at which the SEC inquiry became a formal investigation.	10-01
- Shredding occurs	Shredding of Enron documents in Enron and/or Arthur Andersen accounting firm.	10-01
Shredding stopped	Shredding of Enron documents stopped in Enron and/or Arthur Andersen.	10-01, 11-01
- Fraud announced	Enron admitted to having overstated the company's profits.	11-01
- Bankruptcy filed	Bankruptcy was filed.	12-01

Methods

Selection and Classification of Data

Only emails sent by Enron employees were selected to filter the data from noise, such as advertisements, promotions, and other junk mail. Accordingly, we excluded duplicate emails and emails from other organizations (number of emails excluded in this step: 486,272). Next, we used the Interquartile Range rule for outlier removal: emails that had length above 1.5 times the Interquartile Range were excluded (number of emails excluded in this step: 2,157). This was necessary as some emails included the quoted replies from previous emails, thus providing redundancy. Finally, since our objective was to explore pronoun use, specifically the relationship between the use of different types of pronouns, we excluded emails that did not have at least one 1st person pronoun and one 3rd person pronoun (number of emails excluded in this step: 19,523). In summary, out of the 517,431 emails in the Enron dataset, 9,479 emails (1.83%) were included for further analyses.

Based on Louwerse et al. (2010), 16 types of events within the Enron Corporation from 2000-2001 were identified based on the timeline of the Enron case (Table 1). The event types 'fraudulent paperwork filed signed', 'fraudulent comments', 'rolling blackouts initiated' and 'shredding of documents' were identified as clearly fraudulent. Additionally, as Enron admitted to having overstated the company's profit, the events of 'profit announced' and 'loss announced' were also considered fraudulent. All events considered fraudulent are marked in gray in Table 1.

The dataset primarily consisted of emails from higher executives, increasing the probability that the content of the emails involved decision-making processes related to the fraudulent events. Emails sent during those activities that were sent in periods of fraudulent events were labeled as fraudulent. All other emails were labeled as non-fraudulent. Obviously, this is an overgeneralization, but a useful one given the illustrative purposes of the current study examining significant differences and predictive power.

A total of 28.1% (N = 2,664) of the 9,479 included emails was classified as being related to fraudulent events (compared to 71.9% [N = 6,815] being not related to fraudulent events). An overview of the distribution of normalized 1st and 3rd person pronouns in fraudulent and non-fraudulent emails is depicted in Figure 1.

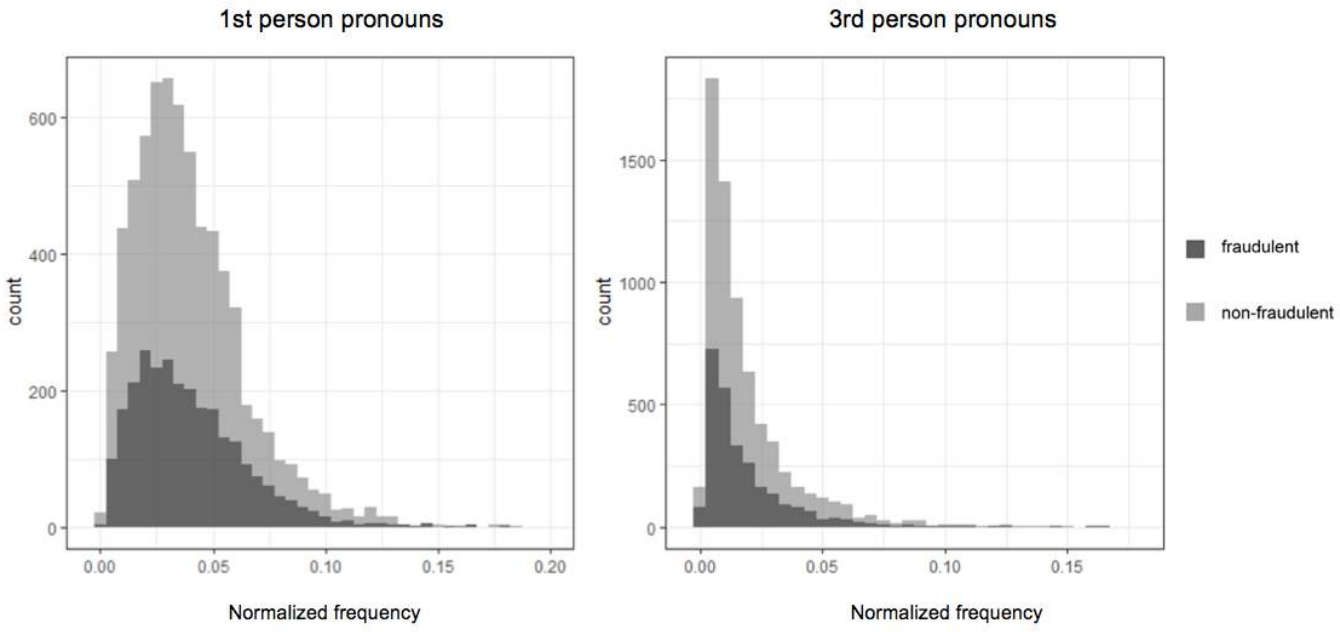


Figure 1: Distribution of normalized 1st and 3rd person pronouns in fraudulent (dark gray) and non-fraudulent (light gray) emails.

Feature Extraction

The number of occurrences of 1st person and 3rd person pronouns were computed for each of the 9,479 emails (see Table 2 for an overview of included pronouns). These occurrences were then normalized by the number of word tokens in each email.

Table 2: Included 1st and 3rd person pronouns.

Type of pronouns	Included pronouns
1st person	*I, me, my, mine, myself, we, us, our, ours, ourselves*
3rd person	*he, she, him, her, his, hers, himself, herself, they, them, their, theirs, themselves*

Data Analysis

Relationship Pronoun Use and Fraudulent Events The relationship between 1st person and 3rd person pronoun use frequency and fraudulent events was examined by computing two logistic regression models.

In Model 1, we used the (normalized) frequency of 1st person pronouns and the (normalized) frequency of 3rd person pronouns as independent variables, and the class label (fraudulent/non-fraudulent) as dependent variable. In Model 2, we computed whether the ratio between 1st person and 3rd person pronoun use had any relationship to the class label. Both models were fitted using a maximum likelihood estimator, using the Python package StatsModels.

Predicting Fraudulent Events Through Pronoun Use In order to predict whether an email was related to fraudulent events, we used two logistic regression classifiers. The classifiers were fitted using the same features as the logistic regression models; specifically, Classifier 1 was trained on the (normalized) frequency of 1st person pronouns and (normalized) frequency of 3rd person pronouns, while Classifier 2 was trained on the ratio between 1st person and 3rd person pronouns. In order to deal with imbalanced data in the training phase, class weight was set to "balanced". In this way, the classifier penalizes mistakes in each class with a weight inversely proportional to the frequency of that class, in order to avoid favoring only the overrepresented class. Both classifiers were evaluated on accuracy, precision, recall, and F1. We also plotted the ROC curve to facilitate the visualization of the relationship between precision and recall. The performance scores were calculated using 10 x stratified 10-fold cross validation, and we report the mean value of all 100 individual scores. For the implementation, we used the LogisticRegression class from the Python library Scikit-learn, with all default parameters (except for class_weight, set to "balanced" to deal with the imbalance over the classes).

Results and Discussion

Relationship Pronoun Use and Fraudulent Events Model 1, which uses normalized 1st and 3rd person pronouns as separate independent variables, did not show a significant relationship between pronoun use and fraudulent events, p = .108 (Table 3).

Table 3: Logistic regression results for Model 1 (normalized 1st and 3rd person pronoun frequency).

			p
Model likelihood			.108
Variable	β	S.E.	p
Intercept	-0.95	0.04	<.001
1st person pronouns	1.38	0.97	.155
3rd person pronouns	-2.61	1.34	.051

Table 4: Logistic regression results for Model 2 (ratio between 1st and 3rd person pronoun frequency).

			p
Model likelihood			.023
Variable	β	S.E.	p
Intercept	-0.99	0.03	<.001
1st/3rd person pronouns	0.01	0.01	.022

Model 2, which uses the ratio between 1st and 3rd person pronouns, did show a significant relationship between pronoun use and fraudulent events, p = .023 (Table 4). The results demonstrated that 1st and 3rd person pronoun use were not individually related to fraudulent events, but that the ratio between the two was. The average ratio for emails that were and were not related to fraudulent events was 3.93 and 3.74 respectively, demonstrating that during times of fraudulent events the use of 1st person pronouns relative to the use of 3rd person pronouns increased. This relationship conflicts with the notion that people try to distance themselves from the information they are conveying when they are being untruthful by increasing abstractness by reducing self-references and increasing other-references in their communication. Yet, these findings are in line with the study of Louwerse et al. (2010) which also did not find support for pronouns reflecting increased abstractness during fraud, but did find support for abstractness in verbs.

As it is important not only to examine the relationship between a construct and an outcome, but also to examine whether the construct has predictive power, we also report the results of the logistic regression classifiers, to predict whether or not an email is related to fraudulent events based on 1st and 3rd person pronoun use.

Table 5: Average results from the 10 x 10-fold cross validation for Classifier 1 (normalized 1st and 3rd person pronoun frequency).

	Accuracy	Precision	Recall	F1
Predicting fraud	48.24%	28.40%	55.39%	37.51%
Predicting non-fraud	48.24%	72.29%	45.44%	55.71%

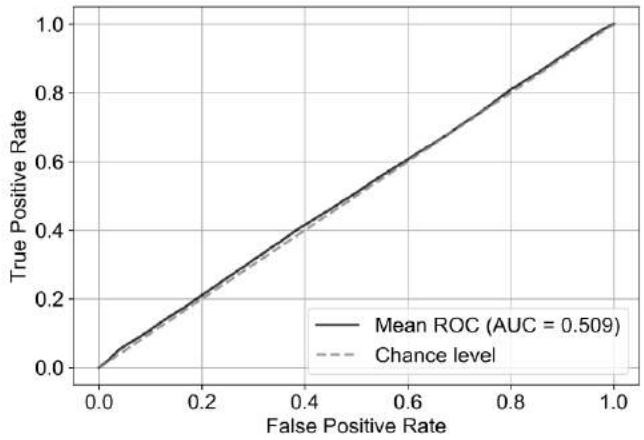


Figure 2: ROC curve for Classifier 1 (normalized 1st and 3rd person pronoun frequency).

Predicting Fraudulent Events Through Pronoun Use The evaluation scores from the 10 x 10-fold cross validation of Classifier 1 (trained on the normalized frequency of 1st person and 3rd person pronouns) are presented in Table 5 and the model's ROC curve is depicted in Figure 2.

As can be seen in Table 5 and in Figure 2, Classifier 1 did not perform above chance level (accuracy = 48.24%). F1 scores were also relatively low, reaching a maximum of 55.71% for predicting non-fraud. Precision scores were considerably higher for predicting non-fraud than for predicting fraud, indicating that the model favored the more common class. The evaluation scores and the ROC curve thus demonstrated the classifier based on the normalized individual frequencies has limited predictive power. This finding is in line with the absence of a significant relationship between these individual frequencies and fraudulent events.

The evaluation scores from the 10 x 10-fold cross validation for Classifier 2 (ratio between 1st and 3rd person pronoun use) are presented in Table 6 and the model's ROC curve is depicted in Figure 3. As can be seen in Table 6 and in Figure 3, the model using the ratio between 1st and 3rd person pronouns performed slightly above chance level (accuracy = 57.37%). F1 scores were again relatively low, reaching a maximum of 68.80%. As was the case for the classifier using the normalized individual frequencies, precision scores were a lot higher for the most common class. Considering all evaluation scores and the ROC curve, the model using the ratio between different pronouns also had limited predictive power.

Table 6: Average results from the 10 x 10-fold cross validation for Classifier 2 (ratio between 1st and 3rd person pronoun frequency).

	Accuracy	Precision	Recall	F1
Predicting fraud	57.37%	29.35%	36.79%	32.64%
Predicting non-fraud	57.37%	72.59%	65.41%	68.80%

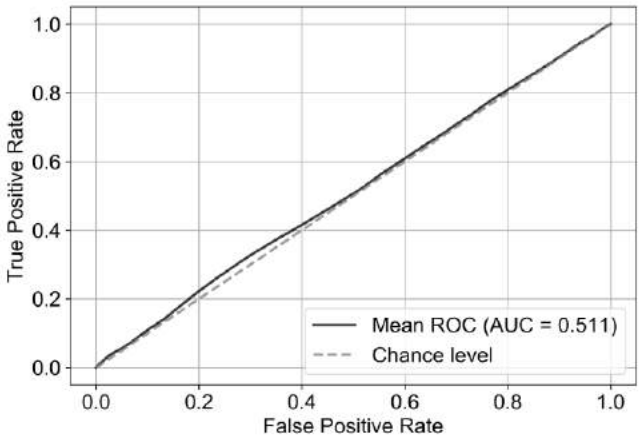


Figure 3: ROC curve for Classifier 2 (ratio between 1st and 3rd person pronoun frequency).

The mean ROC AUC of 0.511 indicates that this classifier is not able to classify deception any better than chance level. In conclusion, even though there was a significant relationship between pronoun use ratio and fraud, this construct had relatively little predictive power.

General Discussion

In the cognitive sciences many studies focus on examining statistical differences. This approach in which differences are examined provides us with valuable insights to explain behavior. Yet, it does not necessarily mean that these differences have predictive power. Understanding behavior both involves explaining and predicting behavior.

As a point in case, the current study examined the relationship between 1st and 3rd person pronoun use in emails sent by Enron employees and fraudulent activities. Additionally, we attempted to predict fraudulent events using 1st and 3rd person pronoun use in the emails.

Previous research demonstrated statistical differences between pronoun use in liars and people that are truthful, but the direction of the effects varied across studies. These studies generally examined the separate effects of 1st and 3rd person pronoun use. The current study demonstrated that the ratio between 1st and 3rd person pronouns was related to fraudulent events. This relationship indicated that the use of 1st person pronouns relative to 3rd person pronouns increased during times of fraudulent activities.

Differences in pronoun use between fraudulent and non-fraudulent communication do not necessarily imply that this construct has any predictive power. In our attempt to predict fraudulent events through pronoun use in emails, classification models scored relatively low on all evaluation measures. These models are therefore limited in their predictive power, not being able to classify deception any better than chance level.

The finding of the current study that differences in 1st and 3rd person pronoun use had no predictive power warrants reported differences in pronoun use between truthful and deceptive communication to not be interpreted as providing a meaningful tool for predicting fraud.

Possibly, classification models were limited in their predictive power in the current study due to the way in which the data were labeled. Whether an email was considered fraudulent or not was based on a general timeline, which might add extra noise to the data. One cannot be sure about the number of emails containing deception and the amount of emails containing no deception that was correctly labeled. However, this issue is inherent in using a naturalistic dataset. The fact remains that it is of utmost importance when one wants to gain a comprehensive insight to not only examine constructs in the laboratory, but also in settings that are of higher ecological validity.

Although effects of deception on 1st and 3rd person pronoun use in communication are widely reported and have been demonstrated in the current study, this construct seems to lack in predictive power. The current study highlights an important conclusion for the cognitive sciences: The importance of not only testing for differences, but of also applying predictive models in order to determine whether effects of a construct on an outcome are also meaningful in predicting the outcome.

References

Bleidorn, W., Arslan, R. C., Denissen, J. J., Rentfrow, P. J., Gebauer, J. E., Potter, J., & Gosling, S. D. (2016). Age and gender differences in self-esteem—A cross-cultural window. *Journal of Personality and Social Psychology*, *111*(3), 396. DOI: 10.1037%2Fpspp0000078.

Campbell, W. M., Campbell, J. P., Reynolds, D. A., Singer, E., & Torres-Carrasquillo, P. A. (2006). Support vector machines for speaker and language recognition. *Computer Speech & Language*, *20*(2), 210-229. DOI: 10.1016/j.csl.2005.06.003.

Cohen, W.W. (2015). *Enron Email Dataset*. Retrieved by https://www.cs.cmu.edu/~enron on 02/01/2017.

DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. *Psychological Bulletin, 129*, 74-118. DOI: 10.1037/0033-2909.129.1.74.

Diesner, J., Frantz, T., Carley, K.M. (2005). Communication networks from the Enron email corpus "It's always about the people. Enron is no different". *Computational and Mathematical Organization Theory, 11*, 201-228. DOI: 10.1007/s10588-005-5377-0.

Garnefski, N., Van Den Kommer, T., Kraaij, V., Teerds, J., Legerstee, J., & Onstein, E. (2002). The relationship between cognitive emotion regulation strategies and emotional problems: comparison between a clinical and a non-clinical sample. *European Journal of Personality*, *16*(5), 403-420. DOI: 10.1002/per.458.

Hancock, J.T., Curry, L., Goorha, S., & Woodworth, M.T. (2008). On lying and being lied to: A linguistic analysis of deception in computer-mediated communication. *Discourse Processes, 45*, 1-23. DOI: 10.1080/01638530701739181.

Humpherys, S. L., Mott, K. C., Burns, M. B., Burgoon, J. K., & Felix, W. F. (2011). Identification of fraudulent financial statements using linguistic credibility analysis. *Decision Support Systems, 50*(3), 585-594. DOI: 10.1016/j.dss.2010.08.009.

Klimt, B. & Yang, Y. (2004). The Enron corpus: A new dataset for email classification research. *Proceedings of the Fifteenth European Conference on Machine Learning*, pp. 217–225. DOI: 10.1007/978-3-540-30115-8_22.

Louwerse, M., Lin, K. I., Drescher, A., & Semin, G. (2010). Linguistic cues predict fraudulent events in a corporate social network. *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, 961-966.

Lu, H. P., & Hsiao, K. L. (2010). The influence of extro/introversion on the intention to pay for social networking sites. *Information & Management, 47*(3), 150-157. DOI: 10.1016/j.im.2010.01.003.

Newman, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J. N. (2003). Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin, 29*, 665–675. DOI: 10.1177/0146167203029005010.

Ott, M., Choi, Y., Cardie, C., & Hancock, J. T. (2011). Finding deceptive opinion spam by any stretch of the imagination. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 309-319.

Pennebaker, J. W. (2011). *The secret life of pronouns: How our words reflect who we are.* New York: Bloomsbury. DOI: 10.1016/S0262-4079(11)62167-2.

Rosenberg, M. D., Casey, B. J., & Holmes, A. J. (2018). Predicting complements explanation in understanding the developing brain. *Nature Communications, 9*(1), 589.

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data mining: practical machine learning tools and techniques*. Burlington, Mass.: Morgan Kaufmann. DOI: 10.1145/507338.507355.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100-1122. DOI: 10.1177/1745691617693393.

Zuckerman, M., DePaulo, B. M., & Rosenthal, R. (1981). Verbal and nonverbal communication of deception. *Advances in Experimental Social Psychology, 14*, 1-59. DOI: 10.1016/S0065-2601(08)60369-X.

Applying the Visual World Paradigm in the Investigation of Preschoolers' Online Reference Processing in a Continuous Discourse

Abigail Toth[1,2] **(a.g.toth@rug.nl)**
[1] Department of Linguistics, University of Alberta
4-32 Assiniboia Hall, Edmonton, AB, T6G 2E7, Canada
[2] Department of Artificial Intelligence, University of Groningen
Nijenborgh 9, 9747 AG Groningen, The Netherlands

Monique Charest (mcharest@ualberta.ca)
Department of Communication Sciences and Disorders, University of Alberta
2-70 Corbett Hall, Edmonton, AB, T6G 2G4, Canada

Jacolien van Rij (j.c.van.rij@rug.nl)
Department of Artificial Intelligence, University of Groningen
Nijenborgh 9, 9747 AG Groningen, The Netherlands

Juhani Järvikivi (jarvikivi@ualberta.ca)
Department of Linguistics, University of Alberta
4-32 Assiniboia Hall, Edmonton, AB, T6G 2E7, Canada

Abstract

Using a novel adaptation of the visual world eye-tracking paradigm we investigated children's and adults' online processing of reference in a naturalistic language context. Participants listened to a 5-minute long storybook while wearing eye-tracking glasses. The gaze data were analyzed relative to the onset of referring expressions (i.e., full noun phrases (NPs) and pronouns) that were mentioned throughout the story. We found that following the mention of a referring expression there was an increase in the proportion of looks to the intended referent for both children and adults. However, this effect was only found early on in the story. As the story progressed, the likelihood that participants directed their eye gaze towards the intended referent decreased. We also found differences in the eye gaze patterns between NPs and pronouns, as well as between children and adults. Overall these findings demonstrate that the mapping between linguistic input and corresponding eye movements is heavily influenced by discourse context.

Keywords: visual world paradigm; eye-tracking; reference processing; discourse

Introduction

During spoken communication, we use different types of referring expressions in order to specify people, places and things. These include both full noun phrases (NPs) (e.g., 'Sarah', 'the bear') and pronouns (e.g., 'she', 'it'). In order for communication to be successful, speakers must choose appropriate referring expressions and listeners must rapidly map those referring expressions onto the intended referents. One method that has been used to investigate the online comprehension of reference is the visual world eye-tracking paradigm (VWP) (Cooper, 1974; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). In the VWP, an individual's eye movements are monitored as they receive spoken language input and view a visual scene. The eye gaze response relative to the spoken language input is taken to reflect underlying processes involved in online language comprehension.

In a seminal paper, Cooper (1974) found that when people were simultaneously presented with spoken language input and a visual scene, they naturally directed their eye gaze towards entities in the visual scene that are semantically related to the language being heard. For example, when participants heard phrases such as 'suddenly I noticed a hungry lion' they fixated on a lion present in the visual scene ~200 milliseconds (ms) after hearing 'lion'. Cooper proposed that the relationship between spoken language input and eye gaze fixations could be viewed as an active online process, and as such could be used to investigate online language processing. Tanenhaus and colleagues (1995) further investigated the influence of the visual scene on spoken language comprehension. They recorded participants' eye movements as participants followed spoken instructions and manipulated real objects visible in front of them. They found that participants fixated on target objects ~250 ms after hearing the word that uniquely identified the target. For example, when participants heard instructions such as 'touch the starred yellow square', they fixated on the target 250 ms after hearing 'starred' when there was only one starred object. However, if participants were given the same instruction and there were two starred objects, they did not fixate on the target until 250 ms after hearing 'yellow', highlighting the high temporal resolution between linguistic input and corresponding eye

movements. Both these studies were instrumental in the development of the VWP as a tool for investigating online language processing.

The mapping between linguistic input and corresponding eye movements has also been utilized to investigate language processing that relies on inference, as is the case in online pronoun resolution (e.g., Arnold, Eisenband, Brown-Schmidt, & Trueswell, 2000; Järvikivi, Van Gompel, Hyönä, & Bertram, 2005). Typically in these studies, participants view scenes with two (or more) referents while listening to passages that contain an ambiguous pronoun. Because pronouns do not have a fixed meaning, there is not a direct association between the linguistic input (i.e., 'he') and a corresponding referent in the visual scene. Thus, listeners must infer which referent the pronoun refers to and eye gaze is taken to reflect this process. For example, VWP studies have reported that following the mention of an ambiguous third person singular pronoun (i.e., 'he'), there is an increase in the proportion of looks to the grammatical subject of the preceding sentence/clause (e.g., Järvikivi, et al., 2005; Kaiser & Trueswell, 2008). This increase in the proportion of looks has been taken to suggest that participants interpreted the pronoun as co-referring with the preceding subject, providing further evidence for what is known as the subject bias (e.g., Crawley, Stevenson, & Kleinman, 1990; Frederiksen, 1981).

In addition to high temporal resolution, another advantage of the VWP is that it does not require participants to read or carry out demanding tasks, and therefore can be used to investigate online language processing in young children. This allows for direct comparisons between children and adults without the potential confounds introduced by response requirements. For example, VWP studies have reported that children as young as 2.5 to 4 years old appear to be sensitive to the subject bias (e.g., Hartshorne, Nappa, & Snedeker, 2015, for an overview; Järvikivi, Pyykkönen-Klauck, Schimke, & Hemforth, 2014; Pyykkönen, Matthews, & Järvikivi 2010; Song & Fisher 2005; 2007). However, the increase in the proportion of looks to the grammatical subject usually did not occur until relatively late (i.e., 1200 ms) after the pronoun onset, suggesting there is still a difference between children and adults.

Another advantage that has been attributed to the VWP is that it can be used to investigate language processing under relatively realistic conditions. This is primarily because the comprehension processes can proceed uninterrupted by response requirements. However, the majority of previous VWP studies have used carefully designed tasks that often encourage participants to carefully look at the visual scene. Furthermore, participants were usually presented with a series of isolated experimental items, where each item was no more than 2-3 sentences, and thus lacked any sort of rich context. This means that each item introduced a new situation or topic, for which participants had no context. In sum, previous applications of the VWP may not accurately reflect naturalistic language processing.

To date, only a single study has used the VWP to investigate reference processing in a continuous discourse. Engelen, Bouwmeester, de Brain and Zwaan (2014) had children listen to a 7-minute long story while viewing a display containing four black and white animal line drawings. They analyzed eye gaze data for both full NPs (e.g., 'rabbit') and pronouns (e.g., 'he') and found that following the onset of a referring expression there was an increase in proportion of looks to the target. However, they also found that overall target fixations (following an NP or pronoun) decreased as the story unfolded over time. However, given that participants viewed the same simple display for the entire 7-minutes, it is possible that the overall decrease was an artifact of fatigue or boredom. To date no study has used the VWP to investigate reference processing in a context where both the language input and visual scene reflect a natural language setting.

Present Study

The present study applied a novel adaptation of the visual world eye-tracking paradigm (VWP) in order to explore online reference processing in a naturalistic language setting. Children and adults listened to a 5-minute story containing multiple animal characters while wearing eye-tracking glasses (ETG). We opted to use ETG over a more traditional table-mounted eye tracker because we wanted to keep the language processing context as naturalistic as possible. The ETG are akin to normal reading glasses and allow for participants to move more freely throughout the duration of the experiment. We analyzed eye gaze patterns with respect to the onset of referring expressions (full NPs and pronouns) mentioned throughout the story. Given the novelty of the methodological application it was important to be able to compare the eye gaze patterns between the two types of referring expressions, as well as between children and adults. Our primary goal was to explore language mediated eye movements outside the context of a carefully designed VWP task. We were interested in what eye gaze patterns could tell us about the processing of continuous discourse in a naturalistic language setting.

Method

Participants

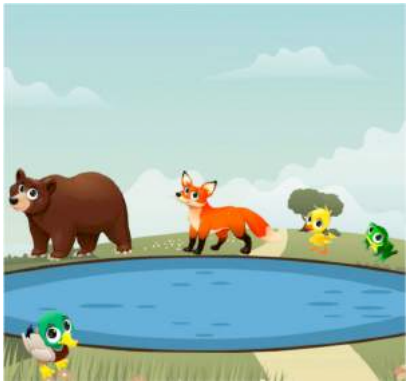
Thirty-five native English-speaking children recruited from preschools and daycares in Edmonton, Alberta, participated in the study. Written parental consent was obtained prior to participation and children received stickers and a t-shirt in exchange for their participation. Despite the fact that the ETG were supposed to be child-friendly, during the analysis it became evident that there

was large amount of gaze loss across participants. This was because the angle between the cameras on the ETG and children's pupils was too large, meaning that the ETG were too big to accurately keep track of eye gaze for some children. This resulted in 17 children being excluded from the analysis. An additional 3 children were excluded because they did not fit the age range (> 6 years old). This resulted in 15 children (7 female; mean age = 4.8 years; range 4.2-5.6) being included in the final analysis. All children had normal vision and hearing based on parental-report.

Sixteen native English speaking adults also participated in the study to serve as a control group. All adults were undergraduate students from the University of Alberta and received partial course credit for their participation. Written consent was obtained prior to participation. Four adults were excluded from the analysis due to technical issues with the ETG. This resulted in 12 adults (10 female; mean age = 20 years; range 18.2-22) being included in the final analysis. All adults had normal vision and hearing based on self-report.

Materials

A 22-page electronic storybook was constructed to be similar in style to an everyday storybook that would be read to children. The story was about a group of animal friends helping a duckling find his father. It contained multiple referring expressions in the linguistic discourse, with corresponding referents in the illustrations. The story began with a single character and after every 3-5 pages a new character was introduced so that it ended with 5 characters in total. All characters were referred to using the masculine pronoun 'he' to ensure ambiguity. The storybook illustrations were created using images from freepik.com. The audio was recorded by a female native speaker of English in a sound-attenuated booth. The illustrations and audio were then pieced together into a 5 minute and 26 second long .mp4 video, where the pages flipped as if it were a real book. An example illustration and associated dialogue can be seen Figure 1 below.



'But before anyone could start looking, Duckling spotted Daddy Duck across the pond! He flapped his wings with excitement. Daddy Duck looked up and saw Duckling. He sighed with relief and started swimming across the pond.'

Figure 1: Example illustration and associated dialogue

Critical Items

Thirty-six full NPs (i.e., character names) and 10 ambiguous pronouns (i.e., he) were selected as experimental items. These items were selected with the criteria that they did not overlap with other referring expressions, meaning that another referring expression could not occur within the ~1200 ms window following their onset. Furthermore, pronouns had to follow a clause where both the grammatical subject and object were animal characters. Because this was a natural story, there was variation in the input that both preceded and followed the critical pronouns. However, it should be noted that at pronoun onset (and for a period of time afterwards), all critical pronouns were ambiguous. Two examples can be seen below and the full set can be found in the supplementary materials[1].

1) Fox thanked Bear. He wanted to play a different game.
2) Bear told Fox to go and hide. He started counting to five.

Procedure

All children were tested individually at their preschool or daycare. The children sat approximately 50 cm in front of a Lenovo laptop and were first familiarized to the animal characters by being shown each animal individually and then being asked to name the animal. In the event that the child misnamed the animal they were corrected. The children were then told they would listen to a short story about the animals while wearing special ETG. The ETG were placed on the child's head and secured using an adjustable strap. The children completed a 3-point calibration and then listened to the electronic storybook. The eye gaze data were collected with SensoMotoric Instruments (SMI) ETG wireless 2 eye-tracking glasses, which included a built-in high-definition scene camera that recorded all audio and video. Registration was binocular with a sampling rate of 60 Hz (16.6ms/frame). After listening to the storybook children were asked a series of five comprehension questions in order to ensure that they had been paying attention. Children had to answer at least four questions correctly in order to be included in the analysis. All children met this criterion.

All adults were tested at The Centre for Comparative Psycholinguistics at the University of Alberta following a similar procedure.

Gaze coding

The gaze data were coded frame-by-frame using Noldus ObserverXT software (Noldus Information Technology, 2012). The areas of interest were each of the five animal characters. Eye gaze that fell outside of interest areas (IAs) was coded as 'elsewhere' or in the case of trackloss was coded as 'NA'. This resulted in a data frame in which each IA had a separate column and every row represented a single frame (~16.7 ms of time). For each frame the IA columns either had a value of 0 (gaze not within IA) or 1

[1]The supplementary materials can be found at: https://git.lwp.rug.nl/a.g.toth/VWP- discourse

(gaze within IA). The gaze data were then binned into 5-frame time bins each representing ~83 ms so that the values in the IA columns could have values between 0 and 5.We were interested in analyzing looks to the target referent (versus looks elsewhere), which in the case of pronouns was coded as the subject of the preceding clause. The time window we were interested in was 2 bins before the referring expression onset up until 14 bins after the referring expression onset (~1415 ms in total).

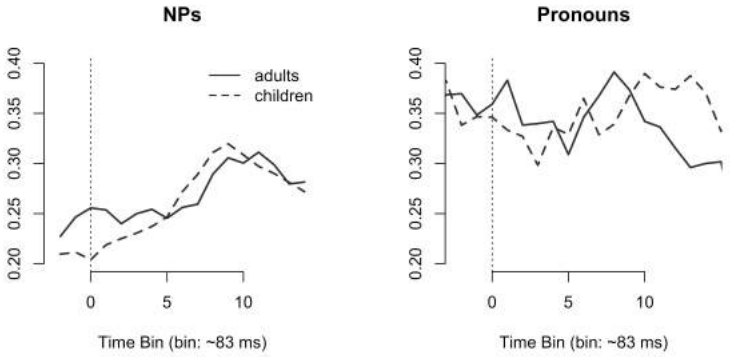


Figure 2: Average proportion of target looks across the time bin analysis window (~1415 ms)

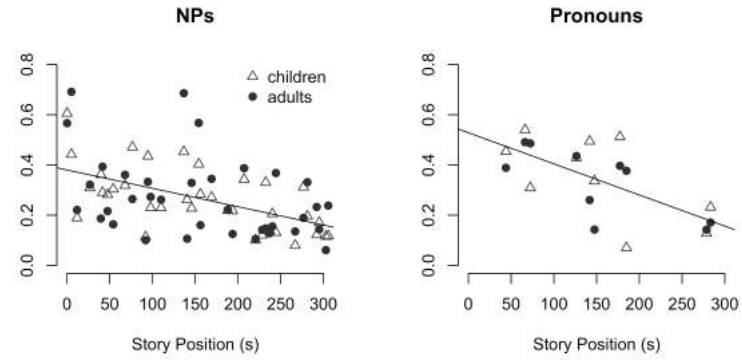


Figure 3: Average proportion of target looks across story duration

Results

Average proportion of looks

Figure 2 shows the average proportion of looks to the target referent across *Time Bin,* where the zero line indicates the referring expression onset and each time bin represents ~83 ms of time. The proportion of looks were averaged over age group (children versus adults separately) and items (NPs versus pronouns separately). For the NPs (left panel) you can see that the proportion of looks to the target increased following referring expression onset and more so for the children (dashed line) compared to adults (solid line). For the pronouns (right panel) the relationship is less clear. However, it should be noted that there were many fewer pronouns than NPs, which can also be seen in the relative smoothness of the lines. Figure 3 shows the average proportion of looks to the target across the duration of the story (*Story Position*), averaged over age group and time bin. For both NPs (left panel) and pronouns (right panel), as well as children (triangles) and adults (circles) there is a clear downward linear trend, meaning that the overall proportion of looks to the target decreased as the story progressed over time.

Analysis

The gaze data were analyzed in R (version 3.1.2; R Core Team 2014) using Generalized Additive Mixed Modeling (GAMM, Wood 2006, mgcv R-package). GAMM is a nonlinear regression method that allows for the modeling of both linear and nonlinear random effects. We opted to use GAMMs over more standardized linear modeling because GAMMs are specifically designed to model nonlinear data and like most time series data, eye-tracking data is almost always nonlinear (Porretta, Kyröläinen, van Rij, & Järvikivi, 2018). The nonlinear relationship between the dependent variable and the predictors is modeled as a smooth function, which is a weighted sum of a set of base functions that each have a different shape. Using logistic GAMMs, we analyzed the counts (looks to the target vs. looks elsewhere) for each time bin (see Porretta et al. for discussion of binomial GAMMs for eye tracking data).

To determine the best-fitting model we did not perform a model comparison procedure, as model comparisons are not very reliable for binomial data (Wood, 2017). Instead, we included binary predictors (which model the difference between conditions) so that we could use summary statistics provided by the mgcv package to determine the significance of the smooth terms. In addition, we used visualization methods to interpret and verify the contribution of the smooth terms (cf van Rij, Hollebrandse, & Hendriks, 2016; van Rij, Hendriks, van Rijn, Baayen, & Wood, in press).

GAMM model of target looks

In order to investigate the eye gaze patterns between the four experimental conditions (adult NPs, child NPs, adult pronouns and child pronouns), we created three binary predictors. *IsChild*, which models the difference between adults (reference level) and children, *IsPronoun*, which models the difference between noun phrases (reference level) and pronouns, and *IsChildPronoun*, which models the additive interaction effect between *IsChild* and *IsPronoun*. We then included the predictor *Time Bin,* in order to analyze looks to the target referent across the time bin analysis window (where time bin 0 was the referring expression onset). We also included the predictor *Story Pos* (i.e., how far into the story the referring expression occurred), in order to test whether looks to the target referent changed as the story progressed. All predictors were allowed to interact. The smooth functions (s()) model the nonlinear regression lines for *Time Bin* and *Story Pos* interacting with the four experimental conditions. The nonlinear tensor product interactions (ti()) model the nonlinear interaction surface between *Time Bin* and *Story Pos* and the four experimental conditions, allowing us to

investigate whether the gaze patterns relative to hearing the referring expression change over the course of the story. To account for individual variation between participants, we included nonlinear random by-Subject factor smooths for *Time Bin* and *Story Pos*, as well as a random intercept for Event (unique Subject-Item combination). To account for autocorrelation in the residuals, an AR1 model was included by specifying the rho parameter and starting point for each time series (cf. Baayen, van Rij, de Cat, & Wood, 2018; van Rij et al., in press). The full analysis can be found in the supplementary materials and the final model summary is presented in Table 1.

Table 1: Summary of the partial effects in GAMM fitted to count data (looks to target vs. looks elsewhere)

A. parametric coefficients	Estimate	Std. Error	z value	Pr (>|z|)
(Intercept)	-1.8000	0.2809	-6.409	1.47e-10 ***
B. smooth terms	EDF	Ref.df	Chi.sq	p-value
s(Time Bin)	4.156	4.626	7.369	0.12902
s(Time Bin):IsChild	4.436	4.834	3.067	0.68953
s(Time Bin):IsPronoun	6.134	7.426	19.762	0.00653 **
s(Time Bin):IsChildPronoun	8.908	9.618	85.969	4.87e-14 ***
s(Story Pos)	1.001	1.001	7.169	0.00743 **
s(Story Pos):IsChild	1.000	1.000	0.410	0.43476
s(Story Pos):IsPronoun	2.599	2.650	7.079	0.14016
s(Story Pos):IsChildPronoun	1.005	1.006	1.066	0.45868
ti(Time Bin, Story Pos)	14.429	15.616	130.820	< 2e-16 ***
ti(Time Bin, Story Pos):IsChild	14.153	15.493	126.287	< 2e-16 ***
ti(Time Bin, Story Pos):IsPronoun	14.677	15.613	182.723	< 2e-16 ***
ti(Time Bin, Story Pos):IsChildPronoun	14.448	15.444	153.732	< 2e-16 ***
s(Time Bin, Subject)	174.863	241.000	2946.556	< 2e-16 ***
s(Story Pos, Subject)	38.477	241.000	10609.166	0.00118 **
s(Event)	986.692	1218.000	6713.175	< 2e-16 ***

For the reference level (adult NPs) there was a significant nonlinear interaction between *Time Bin* and *Story Pos* (Chi.sq(14.429)=130.82; p<.001), meaning that target looks relative to hearing an NP changed as the story progressed over time. The interaction between *Time Bin* and *Story Pos* was also significant for each binary predictor: *IsChild* (Chi.sq(14.153)=126.29; p<.001), *IsPronoun* (Chi.sq(14.677)=182.72; p<.001) and *IsChildPronoun* (Chi.sq(14.448)=153.73; p<.001). Thus, we can conclude that all four experimental conditions have unique interaction surfaces. In order to interpret the interactions, we must use visualization. The contour plots in Figure 4 show how target looks relative to hearing an NP changed as the story progressed over time for both adults and children. The plots can be read like a topographic map with peaks and valleys, where pink indicates more looks to the target and green indicates more looks elsewhere. Both adults' and children's target looks increased after hearing an NP; however, this likelihood decreased in a nonlinear fashion as the story progressed over time. For example, approximately 30 seconds into the story (y-axis), it can be seen that the color changes from green to pink, moving from left to right (x-axis), indicating that target looks began increasing around time bin 3 (~250 ms after NP onset). However, 250 seconds into the story (y-axis), it can be seen that there is relatively solid green, moving from left to right (x-axis), indicating that there was almost no effect of Time Bin later on in the story. It can also be seen that the peak is steeper for children as compared to adults. The contour plots in Figure 5 show how target looks relative to hearing a pronoun changed as the story progressed over time for both adults and children. The white bands indicate the places throughout the story where there are no data, and thus should not be taken into consideration when interpreting the interaction surface. Similar to the NPs, both adults' and children's target looks increased after hearing a pronoun, but again the likelihood decreased in a nonlinear fashion as the story progressed over time. Reflected by the overall color becoming greener as you move from the bottom to the top of the plots (y-axis). It also appears that earlier on in the story (~30 seconds on the y-axis) the increase in target looks happens sooner for adults than it does for children; around time bins 3 (~250 ms) and 8 (~667 ms) respectively. However, because there were a lot less data for the pronouns we need to be careful to avoid over-interpretation.

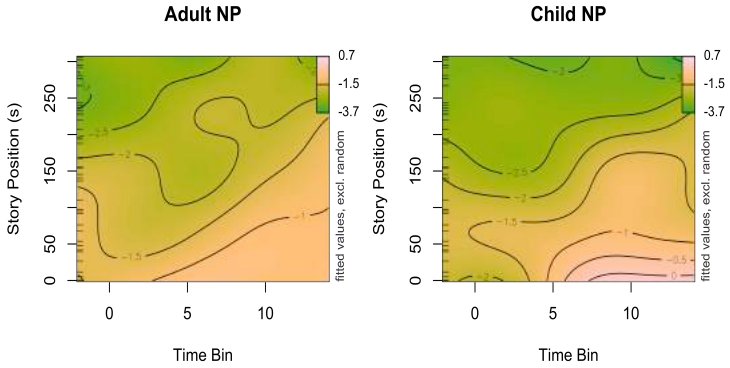


Figure 4: NP contour plots of the interaction between *Time Bin* and *Story Position* for adults and children. Pink indicates more target looks and dark green indicates fewer target looks.

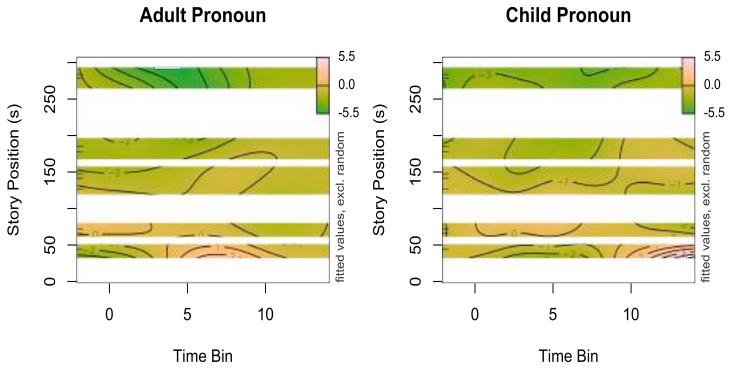


Figure 5: Pronoun contour plots of the interaction between *Time Bin* and *Story Position* for adults and children. Pink indicates more target looks and dark green indicates fewer target looks.

Discussion

The current study applied a novel adaptation of the visual world eye-tracking paradigm (VWP) in order to investigate the online comprehension of reference in a naturalistic language setting.

Overall, we found that eye gaze patterns relative to the onset of referring expressions (full NPs and pronouns) were largely influenced by when in the story the referring expressions occurred. When participants (both children and adults) heard a referring expression earlier on in the story, there was an increase in looks to the target referent. However, when participants heard a referring expression later on in the story there was no increase in looks to the target referent. This finding is in line with that of Engelen et al. (2014), who also found that overall target fixations following a referring expression decreased across a continuous discourse. They suggested that the visual scene may be particularly useful when first building a mental representation of the discourse, such that listeners search for appropriate referents in the visual scene, which results in eye movements being closely time-locked with the unfolding linguistic input. However, once a mental representation is well established, the visual scene does not provide any additional information and therefore eye movements are not closely time-locked. The argument for the visual scene not providing any additional information may be particularly relevant in the case of Engelen et al. (2014), given that the same visual scene was on display for their entire 7-minute discourse. As such, we originally proposed that their findings may be an artifact of fatigue or boredom. Interestingly, we found the same pattern despite using 22 different visual scenes throughout a 5-minute discourse. Based on the similarity of findings, we no longer believe that it is the type of visual scene (simple versus more complex) that causes the downward trend, but more likely a difference in the role that the visual scene plays throughout continuous discourse processing.

Although the visual context in the present study differed from that of Engelen and colleagues (2014), in both studies only a single mental representation of the linguistic discourse was required. This differs from more traditional VWP studies, in which items are presented in isolation and therefore participants must build a new mental representation for each item. In these studies, the eye gaze patterns associated with each item reflect the same type of processing that is, trying to understand who is doing what to whom (i.e., constructing a mental representation). But, because there is no additional linguistic context, the visual scene becomes particularly important for extracting information. This results in a close time-locking between linguistic input (i.e., the mention of a referring expression) and corresponding eye movements in the visual scene (i.e., looking at the referent almost immediately after it being mentioned). In our study, we also saw this for items that occurred earlier on in the discourse. So why did we not see it for items that occurred later on? Perhaps it is simply because later on in the discourse participants already know who the referents are and generally what is going on. In other words, participants already have an established mental representation of the discourse. Therefore, they do not rely on the visual scene for information to the same extent as they do at the beginning of the discourse (or in the more traditional VWP experiments that use thematically mutually unrelated 1-3 sentence stimuli). This results in eye movements not being closely time-locked with the linguistic input in the same way that they are at the beginning of the discourse and in more traditional VWP studies. It is not that eye gaze patterns later on in the discourse are arbitrary (or unmeaningful), but rather that they may reflect a type of processing for which the timing between the linguistic input and the corresponding eye movements and gaze location is not yet well understood. One possibility is that, as the discourse status of a referent changes due to repeated mentions and due to the continuous story context, participants engage in inspecting other aspects of the visual scene to refine their discourse representation, instead of looking at the referent each time it is mentioned.

In addition to looking at overall eye gaze patterns, we also compared the eye gaze patterns between full NPs and pronouns, as well as between children and adults. Following the mention of an NP, we found that earlier on in the story there was a greater increase in the proportion of looks to the target referent for children compared to adults. However, as the story unfolded children became less likely to fixate on the target compared to adults (i.e., there was a stronger effect of story position for children). Following the mention of a pronoun, we found that earlier on in the story children and adults showed a similar increase in the proportion of looks to the target referent, but this happened sooner for adults than children (~250 versus ~667 ms after pronoun onset, respectively). However, given that the dataset was relatively limited, these findings are preliminary and invite further research.

Given the novelty of the present study there were several challenges, the primary one being the technical issues with the eye-tracking glasses, which resulted in >50% of the children being excluded from the analysis. Furthermore, we ended up having a lot fewer referring expressions to analyze than we would have liked (especially in the case of pronouns). This was because our primary goal was to keep the story as natural-sounding as possible and including an excessive amount of referring expressions would have been counterproductive to this goal. Together, these two factors resulted in there being a relatively small dataset, which always runs the risk of lacking statistical power. Nonetheless, the findings from the current study build upon those reported by Engelen et al. (2014) and provide convincing evidence that the relationship between linguistic input and gaze behavior is heavily influenced by context. They further suggest that this relationship is affected by the discourse status of the referent, which changes over the course of a normal continuous story. Furthermore, the findings demonstrate the importance of investigating language processing under naturalistic conditions. Future research is needed to better understand the link between linguistic input and corresponding eye movements.

Acknowledgments

We would like to acknowledge the Social Sciences and Humanities Research Council of Canada (SSHRC), as well Words in the World (a SSHRC partnered research training initiative) for funding this research. We would also like to thank Kaleigh and Romy for all their help with the creation and recording of the experimental stimuli.

References

Arnold, J. E., Eisenband, J. G., Brown-Schmidt, S., & Trueswell, J. C. (2000). The immediate use of gender information: Eye-tracking evidence of the time-course of pronoun resolution. Cognition, 76, B13–B26.

Baayen, R. H., van Rij, J., de Cat, C., & Wood, S. (2018). Autocorrelated errors in experimental data in the language sciences: Some solutions offered by Generalized Additive Mixed Models. In D. Speelman, K. Heylen, & D. Geeraerts (Eds.), Mixed-Effects Regression Models in Linguistics (pp. 49-69). (Quantitative Methods in the Humanities and Social Sciences). Springer International Publishing AG.

Cooper, R. M. (1974). The Control of Eye Fixation by the Meaning of Spoken Language: A New Methodology for the Real-Time Investigation of Speech Perception, Memory, and Language Processing. Cognitive Psychology, 684-107.

Crawley, R., Stevenson, R., & Kleinman, D. (1990). The use of heuristic strategies in the interpretation of pronouns. Journal of Psycholinguistic Research, 4, 245–264.

Engelen, J. A., Bouwmeester, S., de Bruin, A. B., & Zwaan, R. A. (2014). Eye movements reveal differences in children's referential processing during narrative comprehension. Journal Of Experimental Child Psychology, 11857-77.

Hartshorne, J. K., Nappa, R., & Snedeker, J. (2015). Development of the first-mention bias. Journal of child language, 42(2), 423-446.

Järvikivi, J., Pyykkönen-Klauck, P., Schimke, S., Colonna, S., & Hemforth, B. (2014). Information structure cues for 4-year-olds and adults: Tracking eye movements to visually presented anaphoric referents. Language, Cognition And Neuroscience, 29(7), 877-892.

Järvikivi, J., van Gompel, R. P. G., Hyönä, J., & Bertram, R. (2005). Ambiguous pronoun resolution: Contrasting the first-mention and subject-preference accounts. Psychological Science, 16(4), 260-264.

Kaiser, E., & Trueswell, J. C. (2008). Interpreting Pronouns and Demonstratives in Finnish: Evidence for a Form-Specific Approach to Reference Resolution. Language And Cognitive Processes, 23(5), 709-748.

Noldus Information Technology. (2012). The Observer XT reference manual 11.0. Wageningen, the Netherlands: Author.

Porretta, V., Kyröläinen, A., van Rij, J., & Järvikivi, J. (2018). Visual world paradigm data: From preprocessing to nonlinear time-course analysis. In Czarnowski I, Howlett R and Jain L (eds.), Intelligent Decision Technologies 2017, number 73 series Smart Innovation, Systems and Technologies, pp. 268–277.

Pyykkönen, P., Matthews, D., & Järvikivi, J. (2010).Three-year-olds are sensitive to semantic prominence during online language comprehension: A visual world study of pronoun resolution. Language and Cognitive Processes, 25, 115-129.

Song, H., & Fisher, C. (2005). Who's ''she''? Discourse prominence influences preschoolers' comprehension of pronouns. Journal of Memory & Language, 52(1), 29-57.

Song, H., & Fisher, C. (2007). Discourse prominence effects on 2.5- year-old children's interpretation of pronouns. Lingua, 117, 1959- 1987.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of Visual and Linguistic Information in Spoken Language Comprehension. Science, (5217), 1632.

van Rij, J., Hollebrandse, B., Hendriks, P., (2016). Children's Eye Gaze Reveals their Use of Discourse Context in Object Pronoun Resolution. In: Holler A. Glauw C. Suckow K. (eds.) *Empirical Perspectives on Anaphora Resolution*. Berlin: Mouton de Gruyter.

van Rij, J., Wieling, M., Baayen, R.H., & van Rijn, H. (2015). itsadug: interpreting time series and autocorrelated data using GAMMs.

van Rij, J., Hendriks, P., van Rijn, H., Baayen, R.H., & Wood, S.N. (Accepted for publication in Trends in Hearing Science). Analyzing the time course of pupillometric data.

Wood, S. (2017). Generalized Additive Models. New York: Chapman and Hall/CRC.

Wood, S. N. (2006). Generalized additive models: an introduction with R (Vol. 66). Boca Raton: Chapman & Hall/CRC Press.

Top-down information is more important in noisy situations: Exploring the role of pragmatic, semantic, and syntactic information in language processing

Fabio Trecca (fabio@cc.au.dk)
School of Communication and Culture, Aarhus University, 8000 Aarhus, Denmark

Kristian Tylén (kristian@cc.au.dk)
Riccardo Fusaroli (fusaroli@cas.au.dk)
School of Communication and Culture & Interacting Minds Centre, Aarhus University, 8000 Aarhus, Denmark

Christer Johansson (christer.johansson@uib.no)
Department of Linguistics, Literary and Aesthetic Studies, University of Bergen, 5020 Bergen, Norway

Morten H. Christiansen (christiansen@cornell.edu)
Department of Psychology, Cornell University, Ithaca, NY 14853
School of Communication and Culture & Interacting Minds Centre, Aarhus University, 8000 Aarhus, Denmark

Abstract

Language processing depends on the integration of bottom-up information with top-down cues from several different sources—primarily our knowledge of the real world, of discourse contexts, and of how language works. Previous studies have shown that factors pertaining to both the sender and the receiver of the message affect the relative weighting of such information. Here, we suggest another factor that may change our processing strategies: perceptual noise. We hypothesize that listeners weight different sources of top-down information more in situations of perceptual noise than in noise-free situations. Using a sentence-picture matching experiment with four forced-choice alternatives, we show that degrading the speech input with noise compels the listeners to rely more on top-down information in processing. We discuss our results in light of previous findings in the literature, highlighting the need for a unified model of spoken language comprehension in different ecologically valid situations, including under noisy conditions.

Keywords: sentence processing; perceptual noise; pragmatic context; real-world semantics; rational inference.

Introduction

Language processing is based on the integration of bottom-up and top-down information (Marslen-Wilson, 1987; McClelland & Elman, 1986). As we process language, the incoming input is integrated with our existing knowledge—of the local discourse contexts, of the world, and of language—and creates a frame of reference for what comes next (Ferreira & Chantavarin, 2018). This integration happens rapidly (Christiansen & Chater, 2016) and entails that the available evidence must be promptly weighted against prior information, in an effort to determine the likelihood of different specific interpretations of the perceived input (e.g., Gibson, Bergen, & Piantadosi, 2013; Levy, 2008). Success in processing is therefore dependent on the availability of reliable (probabilistic) cues to correct sentence interpretation (Martin, 2016).

At least three sources of information seem to concurrently constrain this inferential process (Venhuizen, Crocker, & Brouwer, 2019). At a local level, the syntactic structure of the language input affects the interpretation of the content of a given linguistic input. An example hereof is that the meaning of syntactically complex sentences is more likely to be misconstrued than that of their less complex counterparts: for instance, listeners more often fail to identify semantic roles in passive sentences than in active sentences (Ferreira, 2003). It has also been shown that listeners tend to take the content of semantically implausible sentences at face value when their syntactic structure is relatively straightforward (e.g., prepositional datives: *The mother gave the daughter to the candle*), but prefer more semantically plausible interpretations when the syntactic structure of the sentences is more complex (e.g., the double-object dative *The mother gave the candle the daughter* is misread as *The mother gave the candle to the daughter*)—even if the semantic content of the two sentences is identical (Gibson et al., 2013).

Lexical-semantic information rooted in our 'real-world' knowledge also points toward specific interpretations of the linguistic input and can even overrule syntactic information (see e.g., MacDonald, Pearlmutter, & Seidenberg, 1994). Semantic properties of the constituents of a sentence, such as animacy, have been shown to affect the inferential process: for instance, listeners tend to interpret animate characters as agents in *who-did-what-to-whom* sentences, independently of syntax (e.g., Larsen & Johansson, 2008; Szewczyk & Schriefers, 2011). This animate-agency bias is consistent with the suggestion that our semantic knowledge may largely originate from sensorimotor representations (see e.g., *situation model* theories of sentence processing; e.g., Zwaan, 2016), which drives listeners toward interpretations of the input that fit with their knowledge of state of affairs in the real world (e.g., Fillenbaum, 1974).

Lastly, the broader discourse context in which a given linguistic input is embedded can affect (and even overrule) our interpretation of semantic and syntactic cues.

Referential/pragmatic contexts and lexical semantics seem to have an additive influence on processing, with (linguistic and extralinguistic) contextual information playing a central role in disambiguating syntactical ambiguities (e.g., the sentence *put the apple on the napkin in the box*, in which the listener can disambiguate whether *on the napkin* modifies *the apple* or *in the box* only by relying on the informativeness of, e.g., elements in the visual world; Snedeker & Trueswell, 2004; see also Spivey, Tanenhaus, Eberhard, & Sedivy, 2002). Pragmatic/contextual expectations can even override our semantic preference for animate agents, for instance through the introduction of a discourse context in which an inanimate object is presented as the agent: Nieuwland and Van Berkum (2006) showed that animacy violations (e.g., *The peanut was in love*), which normally elicit clear N400 effects in ERP experiments, do not do so when the sentences are presented in a context that justifies the violation (e.g., *A woman saw a dancing peanut who had a big smile on his face.* […] *The peanut was in love*). In these semantically implausible contexts, the more canonical sentences (e.g., *The peanut was salted*) suddenly become the violation to the pragmatic/contextual expectations.

All three information sources—pragmatic/contextual information, real-world semantics, and syntax—converge ideally to determine one unequivocal interpretation of the input (cf. Bates & MacWhinney, 1989). However, the relative weighting of each of these information sources in different processing situations seems to be affected by properties of the language input, as well as of the language users. For instance, Dąbrowska and Street (2006) showed that demographic factors such as years of formal education predicted the listeners' ability to interpret semantically implausible sentences when these were presented in passive constructions (e.g., *The soldier was protected by the boy*). Less educated listeners tended to disregard syntactic cues and focus more on semantic and pragmatic/contextual cues (e.g., interpreting the sentence as the more plausible *The soldier protected the boy*). Similar observations have been made in relation to language spoken by non-native speakers: for instance, Gibson et al. (2017) showed that English speakers were more likely to accept literal interpretations of semantically implausible sentences, if these were produced by native English speakers, than if the speakers talked with a foreign accent (thus giving foreigners the benefit of the doubt). Likewise, both children and adults have been shown to adjust their weighting of cues based on the apparent reliability of cues in the input, for instance by being more willing to accept implausible sentences from speakers who previously have produced more implausible utterances (Yurovsky, Case, & Frank, 2017; see also Gibson et al., 2013).

In this study, we suggest that factors pertaining to the communicative environment—e.g., the presence of perceptual noise—are also likely to affect the dynamic weighting of different information sources. The aim of the present study is therefore two-fold: First, we devise a novel experimental paradigm that allows us to individuate and access the relative weight given to different sources of information (pragmatic context, semantics, and syntax) in language processing. Second, we investigate how these weights are dynamically shifted relative to each other as a function of extra-linguistic conditions that can hinder speech communication—in this case, acoustic noise in the speech signal.

Language processing in the real world is prone to be affected by noise (Shannon, 1948): conversations in crowded places or phone calls with bad reception are but a few examples of how noise commonly affects language use in everyday situations (see Mattys, Davis, Bradlow, & Scott, 2012). In these situations, listeners have been shown to devote more cognitive effort to compensate for the reduced informativeness of the signal (Peelle, 2018). Here, we propose that, in order to compensate for less informative bottom-up input, listeners dynamically shift how they weight different information sources: in situations of noise, we are more likely to rely less on bottom-up information and implicitly adopt a more top-down-guided processing style. To test this hypothesis, we used a simple sentence-picture matching task to probe for comprehension. Participants listened to eight short stories; after each story, the participants were presented with four pictures in a four-alternative forced-choice (4AFC) test and instructed to select the picture that matched the central event of the story. In each 4AFC test, only one picture matched the actual language input; the three remaining pictures corresponded to different potential misinterpretations of the language input, and they were specifically designed to reveal processing biases driven by one or more of the three information sources under scrutiny. Half of the participants listened to the short stories in a baseline condition without noise; the other half was presented with the same stories under conditions of perceptual noise.

Method

Participants

167 native Norwegian-speaking (56% female; age: $M = 23.4$, $SD = 3.03$), right-handed undergraduate and graduate students from the University of Bergen (Bergen, Norway) participated in exchange for monetary compensation. Participants were pre-screened for previous or current neurological and/or psychiatric diagnoses, dyslexia, and hearing impairments. The participants were randomly assigned to two experimental conditions: Noise and No-noise ($N_{noise} = 89$, $N_{no\text{-}noise} = 78$).

Materials

Speech stimuli The language stimuli were eight aurally-presented short stories. All stories had an identical narrative structure consisting of four sentences, as in the following example (approximate translation from Norwegian):

S1: The boy walked into the pet store.
S2: His younger sister had been wanting a goldfish for a long time, and now it was time for her to get one.

S3: Everybody thought it was adorable that the boy bought a goldfish for his sister.
S4: As expected, his sister was very happy.

S1 and S2 provided the pragmatic context of the story; S3 was the target sentence and contained the central event of the story (underlined in the example), which was to be matched to the relevant image; and S4 served as a wrap-up sentence. All stories comprised three characters: an agent (e.g., the boy), an object (e.g., the goldfish), and a recipient (e.g., the sister). By switching roles between agent and object, we created different versions of each story, in which both the pragmatic context (S1+S2) and the central event of the story (S3) could be either plausible or implausible in relation to real-world semantics (e.g., S1: *the boy walked into the pet store* vs. *the goldfish walked into the pet store*; S3: *[…] the boy bought a goldfish for his sister* vs. *the goldfish bought a boy for its sister*). Additionally, we manipulated the markedness of the syntactic structure of the target sentence in S3, so that the main event was expressed either using a prepositional dative (unmarked, e.g., *the boy bought a goldfish for his sister*) or a double object construction (marked, e.g., *the boy bought his sister a goldfish*). Together, these 2×2×2 manipulations (pragmatic context semantics × central event semantics × syntactic markedness) resulted in eight possible versions of each story, as shown in Table 1. Participants were tested on all eight story structures. Each story structure-type was randomly assigned to a specific story-token for each participant, so that participants only heard one version of each of the eight stories (e.g., Participant 1 heard Story 1 version A, Story 2 version B, etc.; Participant 2 heard Story 1 version B, Story 2 version C, etc.). The eight stories were interspersed with eight stories from another experiment (with an identical procedure), which served as filler trials.

Table 1: The eight possible narrative structures of Story 1

	S1+S2: Plausible	S1+S2: Implausible	
S3: Unmarked syntax	Story 1a	Story 1b	S3: Plausible
	Story 1c	Story 1d	S3: Implausible
S3: Marked syntax	Story 1e	Story 1f	S3: Plausible
	Story 1g	Story 1h	S3: Implausible

The 64 sound files (8 stories × 8 story structures) were recorded in a soundproof booth by a male native speaker of Norwegian from the Stavanger area, using an Audio-Technica AT2020 Cardioid Condenser USB microphone and Audacity version 2.2.2 for Mac. For the participants in the Noise group, Brownian noise with a signal-to-noise ratio of –19 was added to the sound files using the MixSpeechNoise function from the *praat-semiauto-master* package (https://github.com/drammock/praat-semiauto) in Praat version 6.0.31 (Boersma, 2001).

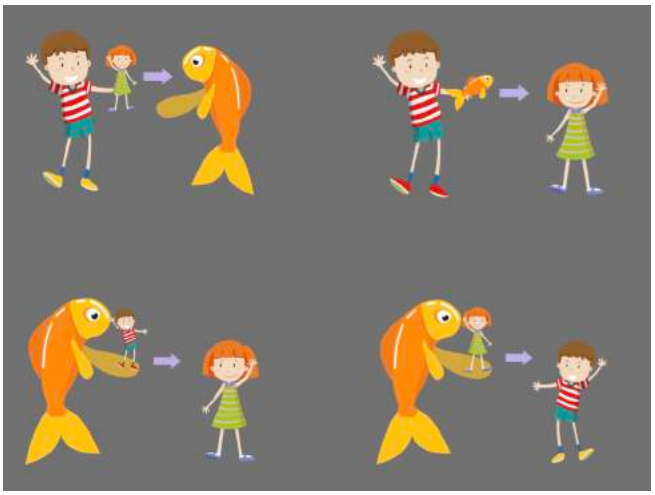


Fig. 1. The visual stimuli in the 4AFC test.

Visual stimuli For each story, four digital color images depicted the three story characters in four different agent-object-recipient relations to each other (Fig. 1). Each image featured an arrow intended to make the direction of the action (e.g., who gave what to whom) more explicit. For each version of each story, only one image corresponded to the central event described in the story and was therefore the correct choice. For instance, the correct match for the target sentence (S3) *the boy bought a goldfish for his sister* would be the top-right image in Fig. 1. The three remaining pictures were foils corresponding to possible misinterpretations of the narrative. These foils were designed to depict misinterpretations that were likely to be elicited by three different processing biases:

(i) *Pragmatic context bias*: an incorrect interpretation of the target sentence driven by the expectations set in the pragmatic context of the story (S1+S2). For instance, given the following pragmatic context: *The goldfish walked into the pet store. His younger sister had been wanting a boy for a long time, and now it was time for her to get one*, and the following target sentence: *The boy bought a goldfish for his sister*, a pragmatic-context bias would be indicated by the participant picking the bottom-left image in Fig. 1, instead of the correct picture match (the top-right image);
(ii) *Real-world semantics bias*: an incorrect interpretation of the narrative in which the target sentence is misinterpreted to match what is plausible in the real world. For instance, given the target sentence *The goldfish bought a boy for his sister*, choosing the top-right image in Fig. 1 (instead of the correct bottom-left image) would indicate a real-world semantic plausibility bias;
(iii) *Syntactic bias*: an incorrect interpretation of the narrative in which marked target-sentence syntax is misinterpreted as unmarked syntax (e.g., the double object construction is misread as prepositional object one), or vice versa. For instance, misinterpreting the target sentence *The boy*

bought the sister the goldfish as *The boy bought the sister for the goldfish* (through the accidental insertion of the preposition *for*) would result in the participant mistakenly clicking on the incorrect top-left image, instead of the correct top-right image.

Given the different narrative structure of each story, a one-to-one mapping between the three picture foils and the three processing biases under scrutiny was not achievable in every trial. However, we estimated that the chances of identifying the three biases in incorrect choices would be equally high when looking across all trials from each participant.

Procedure

Participants sat in front of a computer screen and wore headphones for the entire procedure. Responses in the 4AFC tests were given with a mouse click. Instructions were presented on screen in Norwegian Bokmål and were identical for all participants; however, the participants in the Noise group were advised orally about the presence of noise in the stimuli. The experiment was programmed in PsychoPy2 version 1.90.3 (Peirce & MacAskill, 2018) and began with a practice story (with plausible pragmatic context, plausible target-sentence semantics, and unmarked target-sentence syntax) intended to familiarize the participants with the procedure. After familiarization, the eight stories were presented in fully randomized order. Each story was introduced by a 3 s countdown on screen, after which the sound file was played and a drawing of the three characters of the story were shown on screen (order of presentation for the three characters was fully randomized across participants). After the end of the story, four pictures were presented at the four corners of the screen (as shown in Fig. 1), and the participants were instructed to click at the picture corresponding to what they thought to be the main event in the story. Mouse cursor position was reset at the center of the screen for each 4AFC test.

Data analysis

Accuracy and response time (RT) data were recorded by the experiment script. All possible types of incorrect responses were manually coded as being either due to a pragmatic context bias, a real-world semantics bias, a syntactic bias, or to a combination of two or more biases (for cases in which the incorrect choices were likely to be due to multiple biases). Data pre-processing and statistical analyses were run using R version 3.5.0 (R Core Team, 2018) in RStudio 1.2.1186. Linear mixed-effects models were run using the package lme4 version 1.1-19 (Bates, Maechler, Bolker, & Walker, 2015) and lmerTest 3.0-1 (Kuznetsova, Brockhoff, & Christensen, 2017). All accuracy (correct vs. incorrect) models were logistic mixed-effects models fit through maximum likelihood (Laplace Approximation) with a BOBYQA-optimizer. In addition to accuracy, we analyzed RTs for accurate answers using linear mixed-effects models with log-rescaled outcome variable. All models included random intercepts for subjects and items (random slopes were omitted for model convergence reasons). In the case of null results, we ran Bayes Factor analyses to get indication of whether there was evidence in favor of the null hypothesis, using the brms package (Bürkner, 2017) in R. All Bayesian models had weakly conservative priors for intercept (normal[μ=0, σ=1]), beta estimates (normal[μ=0, σ=1]), SDs of random effects (normal[μ=0, σ=.2]), as well as for correlation coefficients in interaction models (lkj[η=5]).

Results

Accuracy and RTs

To map the relative weight of pragmatic, semantic, and syntactic information sources in noisy and noise-free conditions, we looked at accuracy, response time (RT), and rate and types of errors. For both the No-noise group and the Noise group, overall accuracy on the 4AFC test was high. The average proportion of trials in which participants clicked on the correct picture was 0.78 (within-subject SD = 0.25) in the No-noise group, and 0.69 (within-subject SD = 0.21) in the Noise group. This difference was statistically significant (*Correct* ~ *Noise* + ε: β = –0.92, *SD* = 0.41, z = –2.25, p = .024), suggesting an overall detrimental effect of perceptual noise on comprehension. No statistically significant difference in RTs was found across conditions (*RTs* ~ *Noise* + ε: β = 0.38, *SE* = 0.69, t = 0.55, p = .58). We found no cumulative main effect of semantic plausibility and syntactic markedness on accuracy (*Correct* ~ *Plausibility/Markedness* + ε: β = –0.53, *SD* = 0.14, z = –0.38, p = .7) and RTs (*RT* ~ *Plausibility/Markedness* + ε: β = 0.01, *SE* = 0.32, t = 0.45, p = .65). A Bayes Factor analysis indicated substantial evidence for the null hypothesis (BF = 28.51, Post.Prob. = 0.97), suggesting that the concurrence of semantic implausibility and syntactic markedness did not consistently result in worse performance, compared to stories with plausible content and unmarked syntax. However, when looking at the three information sources individually, a significant main effect of syntactic markedness was found on accuracy (β = –1.5, *SD* = 0.36, z = –4.14, p < .001), revealing ca. 18% lower accuracy for target sentences with marked syntactic structures (i.e., double-object). We also found a statistically significant main effect of story-internal congruence on accuracy (*Correct* ~ *Congruence* + ε: β = –3.45, *SD* = 0.56, z = –6.11, p < .001) and RTs (*RTs* ~ *Congruence* + ε: β = 0.29, *SE* = 0.06, t = 4.74, p < .0001): accuracy was higher and RTs faster for stories in which the events described in S1+S2 and S3 were congruent with each other, and irrespective of whether the two cues were both plausible or implausible (*Correct* ~ *Congruence* × *Plausibility* + ε: β = 0.04, *SD* = 0.45, z = 0.09, p = .92) and RTs (*RTs* ~ *Congruence* × *Plausibility* + ε: β = 1.1, *SE* =

0.61, $t = 1.79$, $p = .076$).[1] Moreover, the effect of congruence was independent of the main effect of syntactic markedness observed above (accuracy, *Correct* ~ *Congruence* × *Syntax* + ε: $\beta = -0.04$, $SD = 1.62$, $z = -0.07$, $p = .94$; RTs, *RTs* ~ *Congruence* × *Syntax* + ε: $\beta = 0.15$, $SE = 0.82$, $t = -0.18$, $p = .85$). However, a Bayes Factor analysis did not provide substantial evidence for the null hypothesis in this case, suggesting that additional data is needed (BF = 1.11, Post.Prob. = 0.52).

Error analysis

In order to individuate how the three information sources were weighted during processing, and how they might be driving comprehension errors, we performed an error analysis. For this purpose, we looked at incorrect responses in situations of story-internal incongruence only, since pragmatic and semantic bias can only be fully distinguished in this case. Distribution of errors is presented in Fig. 2. Across conditions, pragmatics-biased errors accounted for 54% of all errors (No-noise = 22% (42 errors), Noise = 32% (97 errors)); semantics-biased errors accounted for 26% (No-noise = 8% (14 errors), Noise = 18% (55 errors)); and syntax-biased errors accounted for 20% (No-Noise = 8% (15 errors), Noise = 12% (36 errors)). Both semantic bias ($\beta = 0.94$, $SE = 0.04$, $t = 2.02$, $p = .043$) and pragmatic bias ($\beta = 0.46$, $SE = 0.04$, $t = 9.9$, $p < .001$) drove significantly more incorrect responses than syntactic bias; syntactic bias was in turn significantly different from zero ($\beta = 0.26$, $SE = 0.034$, $t = 7.79$, $p < .001$, model structure: *Response* ~ *Bias* + ε). We found no significant two-way interactions between the three sources of bias taken individually (i.e., pragmatics, semantics, and syntax) and noise, suggesting that the role of these information sources in eliciting incorrect responses was not affected selectively by the presence of noise. However, Fig. 3 indicates an evident increase in responses due to a semantic bias, when noise was added to the input, although this interaction was not significant: $\beta = 0.16$, $SE = 0.1$, $t = 1.6$, $p = .11$. A Bayes Factor analysis did not provide robust evidence for this null result (*Noise* × *Semantics* + ε: BF = 1.63, Post.Prob. = 0.62), suggesting that further investigation is needed.

Discussion

In this initial study, we investigated how three sources of information commonly acknowledged in the literature on linguistic processing (i.e., pragmatic/contextual expectations, real-world semantics, and syntactic structure) might contribute differently and dynamically to listeners' comprehension of spoken language input in noisy vs. no-noise conditions. Participants were presented with short stories, in which the three information sources under scrutiny either pointed unequivocally to the same interpretation of the narrative or toward conflicting interpretations. This allowed us to assess the relative weight listeners allocated to the different kinds of information in their interpretation of the linguistic input. Half of the participants listened to stories in the presence of Brownian noise. We hypothesized that listeners would change their processing strategy by generally weighting top-down information more in situations of perceptual noise than in noise-free situations. Moreover, we asked whether the relative weight given to the individual information sources would change when noise was added.

The results provided initial support for our hypothesis by showing that listeners relied more on top-down information in noisy contexts, compared to noise-free ones. In general, accuracy was lower for the Noise group, reflecting the fact that the presence of perceptual noise impedes processing. In both Noise and No-noise groups, listeners made incorrect responses that reflected processing biases driven by either the pragmatic, semantic, or syntactic information in the input—

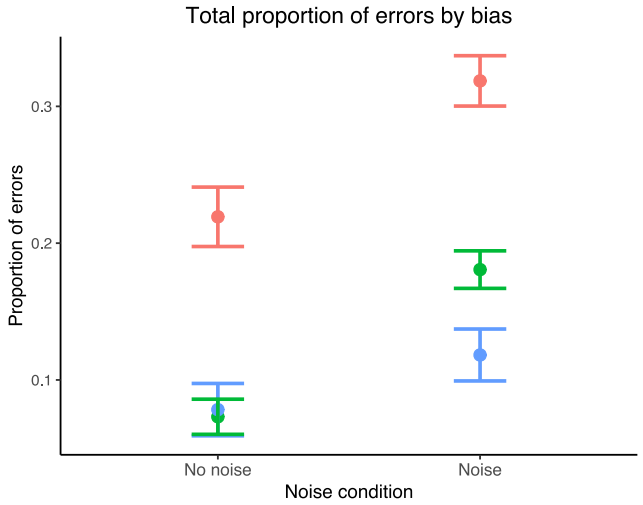


Fig. 2. Distribution of information source biases in incorrect responses (incongruent trials only)

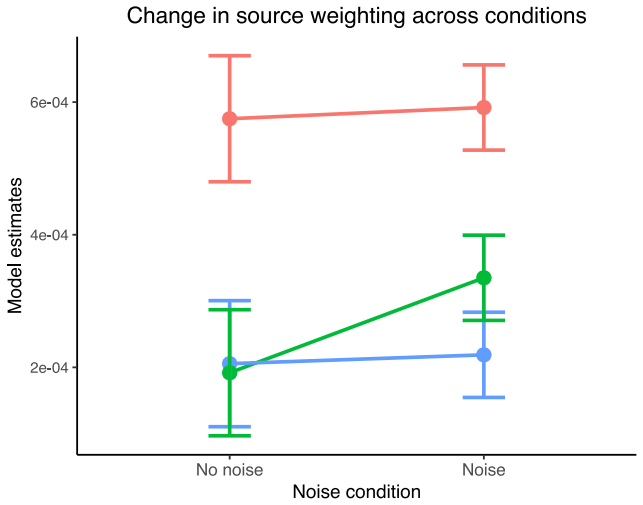


Fig. 3. Predicted values for the model *Response* ~ *Bias* × *Noise* + ε

[1] In the models, plausibility was coded as –1 (S1+S2 and S3 = implausible), 1 (S1+S2 = plausible, S3 = implausible), 2 (S1+S2 = implausible, S3 = plausible), and 3 (S1+S2 and S3 = plausible).

though this happened almost twice as often in the Noise group compared to the No-noise group. Moreover, we found indications that the relative weighting of the different information cues may change when noise is added, with real-world semantics gaining more weight. A number of computational models of language comprehension (e.g., Frank, Koppen, Noordman, & Vonk, 2003, 2008; Venhuizen et al., 2019) have shown that integrating knowledge about the world with lower-level representations of the linguistic input leads to more accurate inferences about the intended meaning of the input. It is possible that the presence of perceptual noise in the signal pressures the processing system and makes it harder for the listener to establish solid representations of the incoming input (e.g., of its syntactic structure and of its pragmatic/contextual information): this may push the listener to rely more on knowledge that is stable over time (i.e., semantic knowledge of the world; see e.g., Kintsch, Patel, & Ericsson, 1999). This mechanism would explain the increase in errors driven by a real-world semantics bias in conditions of noisy signal, but not of those driven by syntax and pragmatics (which are more dependent on establishing representations of the incoming input on the fly). However, this result is only tentative and will need further investigation with more statistical power. Note also that our experimental design only allowed to test comprehension offline (by allowing the participants to make a choice after the end of the story), therefore increasing memory pressure. A more online version of the paradigm (e.g., one that uses mouse tracking/eye tracking) may provide further insights into this issue.

Other interesting results emerged from the study. First, we found a significant main effect of congruence between the pragmatic context of the story and the semantics of the target sentence, with both noisy and non-noisy stimuli. This can be explained in terms of the previously observed mutual influence between story-internal coherence and semantics-based inferences in language comprehension (see e.g., Frank et al., 2003). Second, we found that whenever the pragmatic context of the story and the target-sentence semantics were incongruent (e.g., *the boy walked into the pet store* → *the goldfish bought a boy for its sister*), the pragmatic context "attracted" the listeners' incorrect interpretations to a significantly larger extent than real-world semantics. This evidence is in line with, for instance, previous ERP evidence from Nieuwland and Van Berkum (2006), who showed that listeners' natural tendency to assume animate characters (in our case, human-animate vs. nonhuman-animate) as being agents in stories can be overruled by counterfactual discourse contexts. Third, we found a significant main effect of syntax markedness in the target sentence (S3), in both noisy and noise-free situations, revealing that sentences with a double-object structure are consistently associated with lower accuracy, than sentences with prepositional dative structure. This finding adds to previous psycholinguistic literature documenting the effects of syntactic markedness on language processing (Dabrowska & Street, 2006), and nicely replicates the results from Gibson et al. (2013) and Gibson et al. (2017), in which prepositional dative sentences were shown to lead to literal (although semantically implausible) readings of the sentences more often compared to double-object sentences.

Existing models of language processing under conditions of acoustic challenge (e.g., in hearing-impaired populations) propose that listeners compensate for degraded input by increasing their cognitive effort in terms of memory, attention-based performance monitoring, and allocation of (extralinguistic) neurocognitive resources (e.g., Eckert, Teubner-Rhodes, & Vaden, 2016; Peelle, 2018). However, these compensatory top-down mechanisms have traditionally been thought to only become relevant as a "last resort", when all bottom-up information fails. Instead, our results may suggest that top-down information critically contributes to language processing by default—and more so when the signal itself becomes degraded and therefore less informative. Moreover, our findings hint at a hierarchical weighting of information sources that is flexibly changed in noisy processing situations—at least when the language input is internally incongruent (see e.g., Yurovsky et al., 2017). Reliance on top-down pragmatic context and real-world semantics is largely increased when the language input is degraded by perceptual noise: listeners may rely more heavily on top-down strategies to compensate for the reduced informativeness of the bottom-up cues. Priorities for future studies using the sentence-picture matching design presented here include focusing on languages other than Norwegian, as well as on cross-linguistic differences in the weighting of top-down information. Moreover, it may be important to move away from a binary noise vs. no-noise manipulation and toward a more continuous variation of the amount of noise added to the signal. This may not only lead to stronger patterns of results but also give rise to interesting non-linearities in the data.

Conclusions

Successful language processing depends on the seamless and rapid integration of bottom-up and top-down information. When the bottom-up signal is degraded by noise (as it happens in many everyday situations), listeners become more reliant on top-down information sources. This study presents a novel methodological framework within which to investigate the simultaneous contribution and dynamic weighting of three top-down information sources—pragmatic context, real-world semantics, and sentence syntax—to language processing in the presence of perceptual noise. Our results nicely dovetail with previous findings, while highlighting the need for a unified model of the relative weighting of bottom-up and top-down information in spoken language processing in noisy situations.

Acknowledgments

This research was supported by the Danish Council for Independent Research (FKK) Grant DFF-7013-00074 awarded to Morten H. Christiansen. We are grateful to three anonymous reviewers for useful comments and suggestions for improvement.

References

Bates, E., & MacWhinney, B. (1989). Functionalism and the Competition Model. In B. MacWhinney and E. Bates (Eds.), *The Crosslinguistic Study of Sentence Processing*, 3–73. Cambridge: Cambridge University Press.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1-48.

Bürkner, P.-C. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, 80(1), 1-28.

Christiansen, M.H. & Chater, N. (2016). The Now-or-Never bottleneck: A fundamental constraint on language. *Behavioral & Brain Sciences, 39,* e62.

Dąbrowska, E., & Street, J. (2006). Individual differences in language attainment: Comprehension of passive sentences by native and non-native English speakers. *Language Sciences*, *28*, 604-615.

Eckert, M. A., Teubner-Rhodes, S., & Vaden, K. I. (2016). Is listening in noise worth it? The neurobiology of speech recognition in challenging listening conditions. *Ear & Hearing*, *37*(Suppl. 1), 101S-110S.

Ferreira, F. (2003). The misinterpretation of noncanonical sentences. *Cognitive Psychology*, *47*, 164-203.

Ferreira, F., & Chantavarin, S. (2018). Integration and prediction in language processing: A synthesis of old and new. *Current Directions in Psychological Science*, *27*(6), 443-448.

Fillenbaum, S. (1974). Pragmatic normalization: Further results for some conjunctive and disjunctive sentences. *Journal of Experimental Psychology*, *102*, 574–578.

Frank, S. L., Koppen, M., Noordman, L. G. M., & Vonk, W. (2003). Modeling knowledge-based inferences in story comprehension. *Cognitive Science*, *27*, 875-910.

Frank, S. L., Koppen, M., Noordman, L. G. M., & Vonk, W. (2008). World knowledge in computational models of discourse comprehension. *Discourse Processes*, *45*(6), 429-463.

Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, *110*, 8051–8056.

Gibson, E., Tan, C., Futrell, R., Mahowald, K., Konieczny, L., Hemforth, B., & Fedorenko, E. (2017). Don't underestimate the benefits of being misunderstood. *Psychological Science*, *28*(6), 703-712.

Kintsch, W., Patel, V. L., & Ericsson, K. A. (1999). The role of long-term working memory in text comprehension. *Psychologia*, *42*, 186-198.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(13), 1–26.

Larsen, E. A., & Johansson, C. (2008). Animacy and canonical word order — Evidence from human processing of anaphora. In C. Johansson (Ed.), *Proceedings of the Second Workshop of Anaphora Resolution*, 55-61. Tartu, Estonia: Tartu University Library.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*, 1126–1177.

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review*, *101*, 676–703.

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word-recognition. *Cognition*, *25*(1-2), 71-102.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*, 1-86.

Martin, A. E. (2016). Language processing as cue integration: Grounding the psychology of language in perception and neurophysiology. *Frontiers in Psychology*, *7*(120), 1-17.

Mattys, S. L., Davis, M. H., Bradlow, A. R., & Scott, S. K. (2012). Speech recognition in adverse conditions: A review. *Language and Cognitive Processes*, *27*, 953–978.

Nieuwland, M. S., & Van Berkum, J. J. A. (2006). When peanuts fall in love: N400 evidence for the power of discourse. *Journal of Cognitive Neuroscience*, *18*(7), 1098-1111.

Peelle, J. E. (2018). Listening effort: How the cognitive consequences of acoustic challenge are reflected in brain and behavior. *Ear & Hearing*, *39*, 204-214.

Peirce, J.W., & MacAskill, M.R. (2018). *Building Experiments in PsychoPy*, London: SAGE

R Core Team (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell Systems Technical Journal*, *XXVII*, 379–423.

Snedeker, J., & Trueswell, J. C. (2004). The developing constraints on parsing decisions: The role of lexical-biases and referential scenes in child and adult sentence processing. *Cognitive Psychology*, *49*, 238-299.

Spivey, M. J., Tanenhaus, M. K., Eberhard, K. M., & Sedivy, J. C. (2002). Eye movements and spoken language comprehension: Effects of visual context on syntactic ambiguity resolution. *Cognitive Psychology*, *45*, 447–481.

Szewczyk, J. M., & Schriefers, H. (2001). Is animacy special? ERP correlates of semantic violations and animacy violations in sentence processing. *Brain Research*, *1368*, 108-221.

Venhuizen, N. J., Crocker, M. W., & Brouwer, H. (2019). Expectation-based comprehension: Modeling the interaction of world knowledge and linguistic experience. *Discourse Processes*, *56*(3), 229-255.

Yurovsky, D., Case, S., & Frank, M. (2017). Preschoolers flexibly adapt to linguistic input in a noisy channel. *Psychological Science*, *28*(1), 132-140.

Zwaan, R. A. (2016). Situation models, mental simulations, and abstract concepts in discourse comprehension. *Psychonomic Bulletin & Review*, *23*, 1028-1034.

When is a Visual Perceptual Deficit More Holistic but Less Right-lateralized? The Case of High-school Students with Dyslexia in Chinese

Ricky Van-yip Tso[1,2] (rvytso@eduhk.hk)
Ronald Tsz-chung Chan[1] (ctszchung@eduhk.hk)
1. Department of Psychology, The Education University of Hong Kong, HKSAR
2. Psychological Assessment and Clinical Research Unit, The Education University of Hong Kong, HKSAR

Janet Hui-wen Hsiao[3] (jhsiao@hku.hk)
Department of Psychology, The University of Hong Kong, HKSAR

Abstract

Expert face recognition has been marked by holistic processing and left-side bias/right hemisphere involvement. Hence recognition for Chinese characters, sharing many visual perceptual properties with face perception, was thought to induce stronger holistic processing and left-side bias effect. However, Hsiao & Cottrell (2009) showed that expertise in Chinese character recognition involved reduced holistic processing, while Tso, Au & Hsiao (2014) suggested this effect may be modulated by writing experiences; in contrast, left-side bias was found to be a consistent expertise marker regardless of writing experiences. Here we examine holistic processing and left-side bias effect of Chinese character recognition between adolescents with and without dyslexia. Students with dyslexia were found to recognize Chinese characters with a stronger holistic processing effect than the typical controls. However, compared with the controls, dyslexics showed a more reduced left-side bias in processing mirror-symmetric Chinese characters. The theoretical and educational implications of these results were discussed.

Keywords: Reading, Dyslexia, Left-side bias, Holistic Processing, Perceptual Expertise

Introduction

Holistic Processing

Holistic processing is the tendency to process separable features of an object as a single whole unit. This concept was originally derived from Gestalt psychology, which postulates that the perception of an object as a whole that is a qualitative difference from the sum of its individual parts (Köhler, 1929; see also Wagemans, Elder, et al., 2012; Wagemans, Feldman, et al.,2012). Holistic processing has been a perceptual phenomenon commonly observed in face perception in which all facial parts are integrated and viewed as a whole (Piepers & Robbins, 2012). Holistic processing in face recognition can be demonstrated with the composite paradigm in which it induces the composite face illusion: the two identical top halves of a pair of faces are judged as different when the bottom halves of the two faces are from different faces (see Rossion, 2013). This illusion suggests a failure of selectively attending to facial parts as a result of people obligatorily attending to all facial features as a whole (i.e. holistic processing, see Figure 1; Richler, Wong, & Gauthier, 2011). The holistic processing assessed in the above paradigm demonstrates the second type of configural processing as suggested by Maurer, Le Grand, and Mondloch (2002), which is the inclination to perceive a stimulus as a Gestalt (Pomerantz & Portillo, 2011). Beyond face perception, some studies have posited that expertise-level recognition for subordinate-level objects requires holistic processing (Bukach et al., 2006; though some has suggested limited to face recognition, c.f. Mckone, Kanwisher, & Duchaine, 2007).

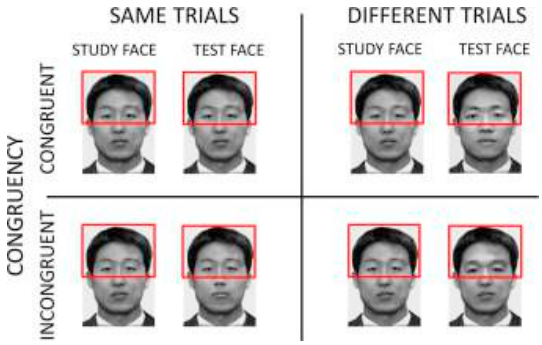


Figure 1. Complete composite paradigm to measure holistic processing for face stimuli. In each trial, participants are cued to attend to top or bottom half of each stimulus pair and judged whether the attended halves are the same or different (attended halves encircled in the figure). Holistic processing is demonstrated by the interference of the irrelevant halves (adapted from Hsiao & Galmar, 2016)

Holistic processing in Chinese character recognition

At first glance, Chinese characters may seem to be a separate class of visual stimuli to that of faces. For example, while the single features (such as eyes and mouth) of each individual face differ but appear in the same positions, the same radicals and strokes can appear in different positions in a character. While faces are always in a symmetrical top-bottom configuration, Chinese characters can appear in more than 10 types of configurations including top-bottom and left-right (Shu, 2003). However, Chinese characters also share many visual properties with faces: They have a homogenous, square configuration—with each character a grapheme mapping onto a morpheme (Shu, 2003). Moreover, strokes are the basic units which combine to form more than 200

basic Chinese character patterns (Hsiao & Shillcock, 2006), which in turn form the Chinese characters. Faces also have homogenous configurations, with facial features combining to form endless different individual faces. A person can differentiate and recognize different faces regardless of their facial expression, similar to a literate typically needing to recognize over 3000 characters regardless of fonts (Hsiao & Cottrell 2009; Wong & Gauthier, 2007). The process of individualizing different faces seems to be comparable with that of naming individual Chinese characters. Hence, theoretically Chinese characters should induce a similar perceptual expertise effect as faces (McCleery *et al*., 2008).

However, using the complete composite paradigm, Hsiao and Cottrell (2009) found that expert Chinese readers had a reduced holistic processing effect (i.e. more analytic) compared with novices. Tso, Au, and Hsiao (2014) showed that the reduced holistic effect of the expert readers in Hsiao and Cottrell's (2009) study may be explained by writing experiences. They showed that compared with novices, expert readers with limited writing performances (Limited-writers) showed increased holistic processing, while expert readers with typical writing abilities (Writers) showed a reduced holistic effect (Tso, Au, & Hsiao, 2014). These findings hint a modulating role of writing abilities on holistic processing: the typical Chinese-reading experts flexibly employ holistic or analytic processing to read and write Chinese characters. It seems that the use of holistic or part-based processing may depend on how readers allocate attention for task relevant information (Chung, Leung, Wong, & Hsiao, 2018).

Holistic processing in the population with special needs

There has been accounts of perceptual differences in processing visual stimuli in populations with a cognitive disability compared with typical controls. For example, reduced holistic processing in has been associated with face-recognition difficulties in patients with prosopagnosia (Avidan, Tanzer, & Behrmann, 2001). Reduced holistic processing also marks a cognitive deficit in people with autism, who were often tested to have poorer abilities in face and facial expression recognition than the general population (Tanaka, Wolf, & Schultz, 2010).

People with dyslexia in the Chinese language is also shown to be characterized by a visuospatial deficit (e.g. visual-orthography processing and visual-spatial attention skills; see Liu et al, 2017), while English dyslexia is generally associated with core deficits in phonological skills. Indeed, developmental dyslexia in an alphabetic script and in the Chinese writing system is characterized by different brain abnormalities (e.g. Siok et al., 2004; Siok et al., 2009): while dyslexia in alphabetic languages is characterized by neurological deficits related to phonological skills (e.g. left temporoparietal regions), dyslexia in Chinese is more associated with abnormalities in regions that are responsible for orthography or visuospatial processing (e.g. middle frontal regions). Chinese-word reading has indeed a strong basis in visual-orthographic processing demonstrated by writing and copying abilities (Tan et al., 2005). Children with reading difficulties are often observed to have a marked discrepancy between reading and writing abilities due to writing in Chinese being a more resource-intensive process than writing in alphabetic languages (Chung & Ho, 2010). As expert reading and writing in Chinese depends on one's ability to analyze local components within a Chinese character (Chung et al, 2018; Hsiao & Cottrell, 2009), people with dyslexia – who generally have backward reading and writing attainments – may fail to employ analytic processing as the components and radicals in a Chinese character may look inseparable to them (Ho, Ng, & Ng, 2003).

Left-side bias

Left-side bias is another visual-perceptual phenomenon commonly reported in face recognition (Burt & Perrett). This effect has also been demonstrated in Chinese character recognition and is suggested to be associated with right-hemisphere involvement (Hsiao & Cottrell, 2009). Left-side bias effect is usually demonstrated using chimeric faces, that is, people often judge faces that composed of two left halves to be more similar to the original face than faces composed of two right halves (Brady, Campbell, & Flaherty, 2005).

Though left-side bias or right-hemisphere lateralization have been thought to correlate with increase in holistic processing in visual object recognition (Gauthier & Tarr, 2002), left-side bias was found to be a consistent behavioral marker of Chinese character recognition regardless of writing experiences, whereas holistic processing could be affected by writing experiences (Tso, Au, & Hsiao, 2014). This effect is consistent with studies that showed right-hemisphere involvement in processing the Chinese orthography (Hsiao, Shillcock, & Lee, 2007; Yang & Cheng, 1999). However, compared with typically developing students, stronger left fusiform and weaker right hemisphere activities have been found in dyslexic children during Chinese character recognition (Siok et al, 2004; Xue et al., 2005). Hence students with dyslexia in this study may display reduced left-side bias compared with the controls.

The present study

This paper hence investigates the role of holistic processing in Chinese recognition by examining how Chinese readers in secondary school with and without a diagnosis of dyslexia process Chinese characters. Reduced holistic processing marks expert Chinese character recognition in Chinese readers with both typical reading and writing abilities. As developmental dyslexia in Chinese in characterized by difficulties in literacy, predominantly in writing performances, students with dyslexia are predicted to processing Chinese characters more holistically than their typical counterpart. Left-side bias of mirror-symmetric Chinese characters was also examined, and this effect was compared between students with and without dyslexia, in relations to holistic processing.

Materials and procedures

Chinese literacy

Dictation performance and Chinese word reading in both timed and untimed context were measured as a reference for their literacy performance. The stimuli were adopted from HKT-P(III). As the purpose of the study was not to yield diagnostic results, we used stimuli from HPT-P(III) for research purposes only to compare the literacy performance between students with and without dyslexia.

i) The untimed Chinese word reading task assessed students' Chinese word reading accuracy. Students read aloud from a set of 150 two-character Chinese words listed in ascending order of difficulty. A participant scored one point for pronouncing both characters of a word correctly.

ii) The Chinese one-minute word reading task assessed students' Chinese word reading fluency. Ninety simple two-character Chinese words were displayed in 9 rows containing 10 words each. Students read aloud as many words as they could in one minute, earning one point every time they read both characters of a word correctly, and the total number of points gave the score.

iii) The Chinese dictation task assessed children's Chinese word writing ability. Students wrote out 45 two-character Chinese words, read out by the examiner in ascending order of difficulty. A student scored one point for writing each character correctly.

Non-verbal Intelligence

To control for the effect of IQ on reading, nonverbal intelligence was assessed using the 9-item subset of Raven's standard progressing matrices (Raven, Court, & Raven, 1996; see Bilker et al., 2012, for its psychometric properties).

Holistic processing

One hundred and sixty pairs of medium to high frequency Chinese characters in Ming font were used as the character stimuli—half of the pairs in top-bottom configuration while the other half in left-right configuration (See Figure 2). 40 pairs were presented in each of the four conditions – same-congruent trials, different-congruent trials, same-incongruent trials and different-incongruent trials. In the congruent trials, the attended halves and the irrelevant halves always led to the same response (i.e. both the attended part and the irrelevant part were the same or different). In the incongruent trials, the attended halves and the irrelevant halves led to different responses – in same incongruent trials, the attended halves were the same while the irrelevant halves were different; whereas in different incongruent trials, the attended halves were different while the irrelevant halves were the same (Figure 3a).

Figure 2. Examples of Chinese characters with a left-right configuration (left) and a top-bottom configuration (right).

The participants' performance in each condition (congruent vs incongruent) is measured by discrimination sensitivity A' as:

$$A' = 0.5 + \left[sign(H-F)\frac{(H-F)^2 + |H-F|}{4\max(H,F) - 4HF} \right]$$

(H and F are the hit and false alarm rate respective) A' is used to measure sensitivity due to its bias-free nonparametric property (Stanislaw & Todorov, 1999). Hence the degree of HP is measured as the A' difference between the congruent trials and the incongruent trials—the larger the discrepancy, the larger the holistic effect. The discrepancy in response time between congruent and incongruent trials was also measured to demonstrate holistic processing. In addition, a misaligned condition was included to tease out the possibility of composite effects due to inhibition abilities, such that if the holistic-processing effect in students with dyslexia is indeed due to interference from the irrelevant halves, misalignment should reduce this effect. See Figure 3.

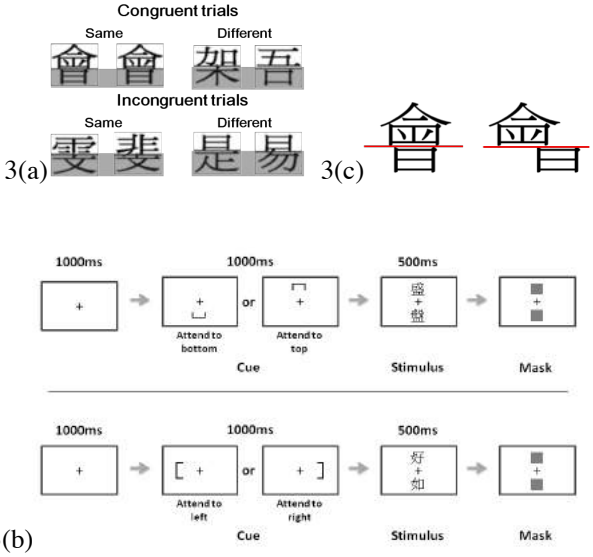


Figure 3. (a) Illustration of stimulus pairs in the complete composite paradigm (b) Trial sequences. (c) character in aligned (left) and misaligned conditions (right).

Left-side bias.

To test for left-side-bias effect, procedures from Tso, Au, & Hsiao (2014) were adopted. Eighty high-frequency mirror-symmetric Chinese characters were selected. Each character was presented once in Ming font. For each character, half of the trials displayed the originals were used on half of the trials, whereas in the other half of the trials displayed chimeric characters constructed from half of the original character and its mirror image, and this was counter-balanced across participants.

For each character stimuli, two left halves constructed the left chimeric character while two right halves formed the right chimeric character (Figure 4a). Each character spanned a visual angle of about 6.7° from a 55 cm viewing distance. After 500 ms of a central fixation cross in each trial, the

original character was displayed either on the left or right side of the screen randomly, at about 7.2° of visual angle away from the center. Each trial displayed the left and right chimeric characters such that one was above and one below an arrow at the screen center which pointed to the original character image. Each chimeric character image subdued about 3°of visual angle away from the center. All image stimuli were displayed on the screen until participants responded to judge which of the two chimeric characters looked more similar to the original one by pressing one of two buttons on the response box. Left-side bias was measured as the percentage of trials in which participants selected chimeric characters composed of two left halves (Figure 4b).

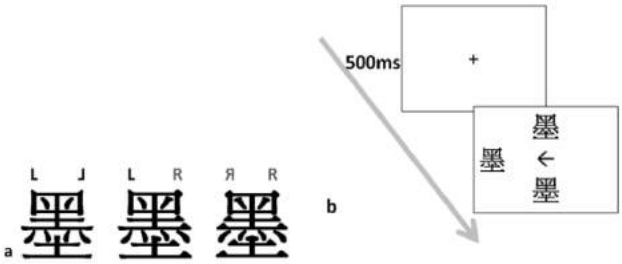


Figure 4. (a) Examples of the stimuli, and (b) the test sequence in the LSB experiment (note that the chimeric characters are still legal Chinese characters).

Results

Literacy abilities and non-verbal intelligence

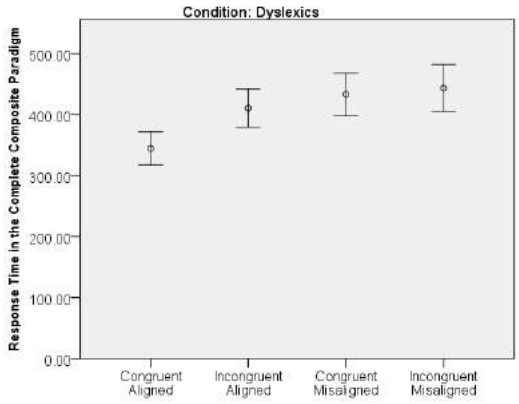
Separate one-way analyses of variance (ANOVAs) were carried out to examine the effect of group (Dyslexics vs Control) on each literacy test. We found that participants in the control condition had a marginally better performance than participants with dyslexia in Chinese word-reading $F(1, 39) = 3.076$, $p = .087$, but their performance in the one-minute word-reading task did not differ, $F(1, 39) = 1.551$, $p = .219$, suggesting that both groups having similar performance in word recognition and fluency in naming over-learned Chinese characters. However, participants in the control condition had significantly better performance in the Chinese word dictation task, $F(1, 39) = 7.229$, $p = .01$, suggesting the students with dyslexia had persistent difficulties in writing Chinese characters even when in high-school grades. The scores are summarized in Table 1.

Table 1. Summary of the scores of Chinese word-reading, Chinese one-minute word reading, Chinese dictation and non-verbal IQ (9-item Raven's) in high-school students with and without dyslexia.

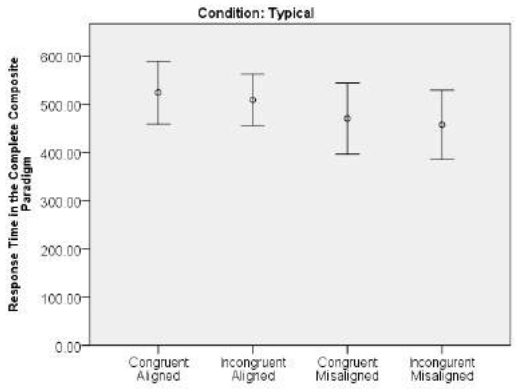
	Control *Mean (SE)*	Dyslexics *Mean (SE)*
Chinese Word-reading	102 (1.82)	99.96 (3.41)
One-minute word reading	93.43 (4.62)	85.91 (3.95)
Chinese dictation	59.73 (13.14)	47.52 (17.35)
Non-verbal IQ	3.91 (1.74)	3.96 (1.34)

Holistic processing

We next examined the ability to holistically process Chinese characters in participants with and without dyslexia. We first conducted a 2 (congruency: congruent vs. incongruent) × 2 (group: dyslexics vs. control) repeated measures ANOVA on A', which showed a main effect of congruency, $F(1, 38) = 27.35$, $p = .000006$, $\eta p2 = .419$, but no interaction between congruency and group, $F(1, 38) = 1.354$, $p = .252$, or main effect of group, $F(1, 38) = 1.342$, $p = .254$, was found. We then conducted a 2 (congruency: congruent vs. incongruent) × 2 (group: dyslexics vs. control) repeated measures ANOVA on response time. We found a significant interaction between congruency and group, $F(1, 38) = 5.854$, $p = .02$, $\eta p2 = .133$, and a main effect of group, $F(1, 38) = 5.306$, $p = .027$, $\eta p2 = .123$, but no main effect of congruency, $F(1, 38) = 2.254$, $p = .150$. Post-hoc ANOVA showed that students with dyslexia responded more slowly in incongruent than in congruent trials, $F(19) = 34.3$, $p < .000012$, $\eta p2 = .644$, whereas response times in congruent and incongruent trials were similar in typically developing students, $F(19) = 0.254$, $p = .620$. See Figure 5.



5a)



5b)

Figure 5. The composite effect in character perception was significant among dyslexics (a), but not the typically developing readers (b). Misalignments significantly reduced the effect in the dyslexics.

The results also showed that misalignment significantly reduced the congruency effect demonstrated by response time among students with dyslexia: a significant interaction between congruency and misalignment (aligned vs. misaligned), $F(1, 19) = 5.662$, $p = .029$, $\eta p2 = .239$; there was no misalignment effect among typically developing students.

Together, these results suggest that participants with dyslexia perceived Chinese characters more holistically than controls.

Left-side Bias

Finally, the results on left-side bias suggests that typical readers have a stronger left-side bias than students with dyslexia, $F(1,38) = 6.439$, $p = .015$, $\eta p2 = .145$. It seems that although the participants with dyslexia were more holistic than typical readers in Chinese character recognition, they revealed weaker left-side bias (Figure 6).

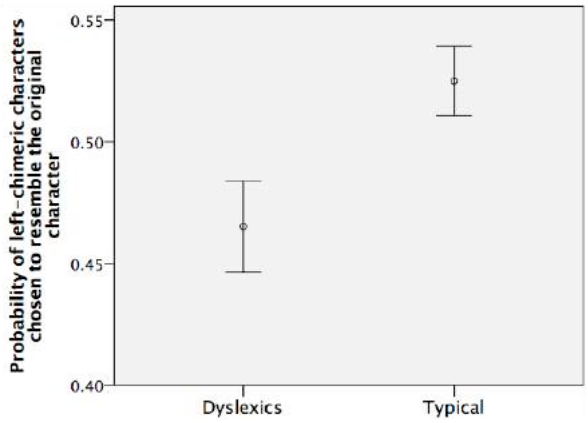


Figure 6. Preference for left chimeric characters in participants with and without dyslexia.

Discussions

This study investigated how high-school students with and without dyslexia differed in how they processed Chinese characters by examining two perceptual expertise phenomena: holistic processing and left-side bias. Our results show that high-school participants with dyslexia demonstrated a stronger holistic processing effect in Chinese character recognition compared with the typically developing controls. This study is consistent with Hsiao and Cottrell's (2009) study in which they showed that reduced HP is associated with expert Chinese character recognition compared with novices, though our study compared between typically developing expert readers and dyslexics in which the dyslexic participants were not completely novices but had relatively weaker Chinese literacy abilities. Our result is also comparable with Tso et al. (2014)'s study that suggested a modulating role of writing abilities on holistic processing: While the students with dyslexia in this study had marginally comparable reading performance to that of typical controls, they recalled and wrote fewer words. Unlike everyday face recognition in which one is not required to recall and draw faces, a typical Chinese reader is fluent in both Chinese character recognition and writing. Indeed, Zhou, et al. (2012) demonstrated reduced HP in artists with face-drawing experience compared with ordinary face-observers. Stronger holistic processing in students with dyslexia than in the typical controls, then, may indicate a perceptual difference between poor and proficient writers.

According to Maurer et al. (2002), holistic processing is a second-order configural processing in which both featural and spatial-distal information within an object are integrated and processed. Hence, the stronger holistic processing effect of the dyslexic students in the present study may also suggest that they recognized characters with an over-dependence on their visuo-spatial information of components, which may hinder developing literacy expertise, particularly in writing. It seems that students with dyslexia demonstrated persistent perceptual abnormalities even when they are in secondary school, which hinders them to selectively attend to individual character components. This in turn hindered Chinese character recognition as it is an ability facilitated by sensitivity to the specific positions of components radicals and structures within a character (Ho, Ng, & Ng, 2003). This speculation warrants future follow-up studies.

This study also echoed with Hsiao and Cottrell's (2009) and Tso, Au and Hsiao's (2014) findings demonstrating that left-side bias was a consistent expertise marker of Chinese character recognition: The dyslexic readers showed reduced left-side bias of Chinese characters than typically developing readers. Our result is also consistent with previous studies that suggested a stronger left-hemisphere but weaker right-hemisphere involvement for Chinese character recognition in readers with dyslexia (Siok et al, 2004; Xue et al., 2005). These effects suggest that dyslexics employ a strategy to process Chinese characters which may be both perceptually and neurologically different from typical readers. Similar to face perceptual processes which involves RH/LSB, our results are consistent with that in prosopagnosic patients who had a reduced left-side bias in facial perception—suggesting a reduced RH involvement in face recognition (Malaspina, Albonic, & Daini, 2016).

However, while HP was previously thought to associate with RH activation as demonstrated in face and subordinate visual-object recognition, the results of this study echoed Hsiao and Cottrell's (2009) study, demonstrating that increased LSB but reduced HP as expertise markers of Chinese character recognition. Holistic processing effect brought about by the composite-face illusion is due to obligatory attention directed to all facial parts, resulting in failure to selectively attending to parts (Hole, 1994; Richler, Tanaka, Brown, & Gauthier, 2008; Richler, Wong, & Gauthier, 2011). Therefore, one reason why Chinese character recognition is different from that of face perception may be because the spacing information between typical Chinese character components may be unimportant to typical Chinese readers (Hsiao & Cottrell, 2009), while spatial information is important in typical face recognition processes (i.e. small changes in spacing between features typically change the face identity; see Farah, et al., 1998). Hence, the relationship between holistic processing and right hemisphere lateralization may be modulated by whether spatial information is used during recognition of visual stimuli. To test above speculations, Hsiao and Galmar (2016) demonstrated through a computational simulation a positive relationship between holistic processing and RH

lateralization when a face recognition task relied purely on spatial information (i.e., all faces stimuli differ only spacing among the same features). On the other hand, when the task recognized faces based purely on features (i.e., all faces differed in features but the same spacing between them), holistic processing correlated negatively with RH lateralization (see also Chung et al., 2018). Therefore, whether the RH engages holistic processing in a recognition task may depend on the type of information used for its processing. Indeed, Chinese character recognition is facilitated by sensitivity to components radicals at specific positions within a character (Ho, Ng, & Ng, 2003), not the spatial distances between components. Hence left-side bias in Chinese character recognition is perhaps related to sensitivity to first-order relations in configural processing, i.e. the relative spatial locations of individual components within a character (Maurer et al., 2002).

To conclude, this study is the first to report the perceptual difference between typically developing and dyslexic students in high school by investigating holistic processing and left-side bias of Chinese character recognition. It has demonstrated preliminary evidence for the link between inability to reduce holistic processing and difficulties in Chinese literacy: dyslexic Chinese are less readily to engage in analytic processing to attend to character components. Finally, the reduced left-side bias of Chinese characters in the dyslexics may be related to deficits in forming first order relationship between components. This study suggested that high-school students with dyslexia in Chinese may still encounter difficulties in reading and writing due to persistent deficits in their literacy-related cognitive abilities, and they may require further supports in their learning to enhance attention to Chinese character components or radicals.

References

Bukach, C. M., Gauthier, I., & Tarr, J. M. (2006). Beyond faces and modularity: The power of an expertise framework. *Trends in Cognitive Sciences, 10,* 156–166

Chung, H. K. S., Leung, J. C. Y., Wong, V. M. Y., & Hsiao, J. H. (2018). When is the right hemisphere holistic and when is it not? The case of Chinese character recognition. *Cognition*, *178*, 50-56.

Goswami, U., Power, A. J., Lallier, M., & Facoetti, A. (2014). Oscillatory "temporal sampling" and developmental dyslexia: Toward an over-arching theoretical framework. *Frontiers in Human Neuroscience*, *8,* 904.

Hsiao, J. H., & Cottrell, G. W. (2009). Not all visual expertise is holistic, but it may be leftist: The case of Chinese character recognition. *Psychological Science*, *20*, 455–463.

Hsiao, J. H., & Shillcock, R. (2006). Analysis of a Chinese phonetic compound database: Implications for orthographic processing. *J Psycholinguist Res, 35,* 405–426

Köhler, W. (1929). *Gestalt psychology*. New York, NY: Liveright.

Malaspina M, Albonico A, & Daini R. (2016). Right perceptual bias and self-face recognition in individuals with congenital prosopagnosia. *Laterality, 118,* 42.

Maurer, D., Le Grand, R., & Mondloch, C. J. (2002). The many faces of configural processing. *T.I.C.S, 6,* 255–260.

McBride-Chang, C., Chow, B. W., Zhong, Y., Burgess, S., & Hayward, W. G. (2005). Chinese character acquisition and visual skills in two Chinese scripts. *Read Writ, 18(*2), 99–128.

McKone, E., Kanwisher, N., & Duchaine, B. C. (2007). Can generic expertise explain special processing for faces? *Trends in Cognitive Sciences, 11*, 8–15.

Piepers, D. W., & Robbins, R. A. (2012). A review and clarification of the terms "holistic," "configural," and "relational" in the face perception literature. *Frontiers in Psychology, 3,* Article 559.

Pomerantz, J. R., & Portillo, M. C. (2011). Grouping and emergent features in vision: Toward a theory of basic Gestalts. *Journal of Experimental Psychology: Human Perception and Performance, 37,* 1331–1349.

Richler, J. J., Wong, Y. K., & Gauthier, I. (2011). Perceptual expertise as a shift from strategic interference to automatic holistic processing. *Cur Dir Psychol Sci, 20,* 129–134.

Rossion, B. (2013). The composite face illusion: A whole window into our understanding of holistic face perception. *Visual Cognition, 21,* 139–253.

Shu, H. (2003). Chinese writing system and learning to read. *International Journal of Psychology, 38,* 274–285.

Siok, W. T., Spinks, J. A., Jin, Z., & Tan, L. H. (2009). Developmental dyslexia is characterized by the co-existence of visuospatial and phonological disorders in Chinese children. *Current Biology, 19*(19), R890–R892.

Tan, L. H., Spinks, J. A., Eden, G. F., Perfetti, C. A., & Siok, W. T. (2005). *Reading depends on writing, in Chinese*. *PNAS, USA, 102*, 8781–8785.

Vellutino, F. R., Fletcher, J. M., Snowling, M. J., & Scanlon, D. M. (2004). Specific reading disability (dyslexia): What have we learned in the past four decades? *Journal of Child Psychology and Psychiatry, 45*(1), 2–40.

Vidyasagar, T. R., & Pammer, K. (2010). Dyslexia: A deficit in visuo-spatial attention, not in phonological processing. *Trends in Cognitive Sciences, 14*(2), 57–63.

Wagemans, J., Elder, J. H., Kubovy, M., Palmer, S. E., Peterson, M. A., Singh, M., & von der Heydt, R. (2012). A century of Gestalt psychology in visual perception: I. Perceptual grouping and figure-ground organization. *Psychological Bulletin, 138*, 1172–1217.

Wagemans, J., Feldman, J., Gepshtein, S., Kimchi, R., Pomerantz, J. R., van der Helm, P. A., & van Leeuwen, C. (2012). *A century of Gestalt psychology in visual perception: II. Conceptual and theoretical foundations*. *Psychological Bulletin, 138*, 1218–1252.

Wong, A. C.-N., & Gauthier, I. (2007). An analysis of letter expertise in a levels-of-categorization framework. *Visual Cognition, 15*, 854–879.

Do Bilingual Infants Possess Enhanced Cognitive Skills?

Angeline Sin Mei Tsui
[astsui@stanford.edu]
Department of Psychology
Stanford University

Christopher T. Fennell
[fennell@uottawa.ca]
School of Psychology
University of Ottawa

Abstract

Prior studies have reported that bilingualism enhances cognitive ability due to the regular conflict management of two language systems (Bialystok, 2015). Here, we explore whether infant bilingualism improves cognitive ability at 9.5 months. Twenty-four monolingual English and 23 bilingual French-English infants were first trained to predict a reward on the right based on a set of tone-shape rule structure (AAB pattern). Infants were later trained to predict a different reward on the left based on another set of new rule structure (ABB pattern). Correct anticipation of reward locations indicates successful learning. If bilingualism improves infants' cognitive skills, bilingual infants would be better at learning a new pattern-reward association. However, we did not find evidence that bilinguals looked at the correct location more than monolinguals or learned the new pattern-reward association faster. Thus, our results suggest bilingualism may not enhance cognitive ability at 9.5 months, as least using the current paradigm.

Keywords: infant bilingualism; cognitive ability; bilingual advantage

Introduction

Bilingual infants face key language learning challenges. For example, as their time listening to language input is divided between the two languages, they therefore hear less input from each language. In addition, bilingual infants often learn their languages in a more variable environment than monolinguals. It is not rare for bilingual infants to listen to parents mixing languages in one conversation (Byers-Heinlein, 2013) or to hear words form speakers with non-native accents (Bosch & Ramon-Casas, 2011). Yet, bilingual infants reach basic language milestones at similar age as monolinguals, such as sensitivity to native sounds (Ferjan Ramìrez et al., 2017) and basic word learning (for a review, see Fennell, Tsui, & Hudon, 2016). This raises the possibility that bilingual infants may have better cognitive control processes because they are remarkably efficient in managing two different language systems.

A number of studies have suggested that bilingual experiences may enhance individuals' executive functioning (see Bialystok, Craik, & Luk, 2012 for a review). Executive functioning is an umbrella term describing a set of cognitive abilities, including inhibition of dominant responses, shifting between tasks or mental sets, as well as updating and monitoring working memory (Miyake et al., 2000). As bilinguals activate both languages even they only speak in one language at a time (e.g., Thierry & Wu, 2007), they must selectively attend to the correct language while inhibiting the other competing language for effective communication in daily life. Earlier theories, such as Green (1998), have proposed that bilinguals' routine of inhibiting the irrelevant language during production can improve their inhibitory ability in non-verbal domains. For example, a number of studies have shown that bilinguals outperform monolinguals in a number of non-verbal conflict tasks, including the flanker task (Costa, Hernández, Sebastián-Gallés, 2008), the Simon task (Bialystok et al., 2004), and the Spatial-Stroop task (Bialystok, 2006).

Subsequent studies have extended research to other aspects of executive control components aside from inhibition. A major reason for this move was that researchers (e.g., Bialystok, 2006; Costa et al., 2008) not only found a bilingual advantage in the incongruent trials of the above conflict tasks, which require inhibitory control, but also in the congruent trials that do not require inhibition. Hence, researchers suggested bilinguals' cognitive advantage is not limited to inhibition ability, but is instead related to an enhancement in executive attention (Bialystok, 2017). Executive attention is the ability to control ones' attention, including ones' ability to maintain attention to the relevant part of the task and avoid directing attention to distractors. Bialystok (2017) has suggested that bilingual experience improves learners' attention systems as bilinguals regularly need to control their attention to accommodate two different language systems. For example, they need to differentiate between the two languages, switch attention between the two, and allocate their attention to the relevant language. Importantly, Bialystok (2015) has highlighted that bilingual experience not only enhances learners' executive function processing when they are selectively producing one of their languages, but also when they are selectively processing the two languages during comprehension. As such, infant bilinguals, who possess richer language comprehension than production, may also enjoy a cognitive advantage before the onset of a large productive vocabulary.

Indeed, some infant studies demonstrate that infant bilingualism may improve executive functioning. In a pioneering study, Kovacs and Mehler (2009a) examined whether Italian monolingual and Italian-other bilingual 7-month-olds differ in cognitive control. Infants were

conditioned to associations between a set of rules and visual rewards. These sets of rules could be conveyed auditorily or visually. For example, two rule structures were composed of three syllables, one with an AAB pattern (e.g. la-la-ga) and one with an ABA pattern (e.g. la-ga-la). In the experiment, infants were first trained to look at a toy (e.g., puppy A) at a particular visual location (e.g., right-side) after hearing one rule structure (e.g., AAB) and later were required to learn a new association (e.g., look at the puppy B on the left when hearing ABA). This task thus required infants to inhibit an earlier learned response and flexibly switch to a new reward response. Across three experiments that used either syllable or visual geometric shape rule patterns, the authors consistently found only bilingual infants successfully learned the new associations. In a similar study, Kovacs and Mehler (2009b) tested whether monolinguals and bilinguals of 12 months were able to simultaneously learn two different tri-syllabic structure rules: AAB and ABA. They again associated each rule to a visual toy at a particular position on the screen (e.g. AAB-object A-left side; ABA-object B-right side). At test, the researchers presented new tri-syllabic auditory stimuli that conveyed either an AAB or ABA structure. They then measured whether infants looked to the correct position after hearing the new auditory stimuli (e.g. AAB structure provokes looks to the left side). If infants looked at the correct corresponding position, it means they learned the corresponding structure rule. The researchers found that bilinguals were able to learn both rules, but monolinguals learned only one of the rule patterns (AAB but not ABA). To summarize, the two studies have suggested that bilingual infants may be better able to control interference and switch between two rule structures.

Over the past decade, however, only a few studies have attempted to replicate the studies in Kovacs and Mehler (2009a). For example, Ibánez-Lillo, Pons, Costa, & Sebastián-Gallés (2010) discovered that both monolingual and bilingual infants of 8 months could inhibit the previously learned cue-reward pairing and successfully learn a new cue-reward pairing, suggesting no early bilingual advantage in cognition. In contrast, Pourllyaei and Byers-Heinlein (2018) found that 7-month-old bilinguals, but not monolinguals, were able to inhibit a previously learned cues' position and anticipate a newly learned cues' position. They specifically presented infants with a visual-auditory cue (e.g., a colorful butterfly paired with a whistle sound) for the first 9 trials on one side of the screen (e.g., left) and then switched the position (e.g., right) of the visual-auditory cue for the next 9 trials. Further, Comishen, Bialystok and Adler (2019) also found that 6-month-old bilingual infants outperformed monolingual infants in a visual expectation cueing paradigm in which infants needed to change their anticipatory looks in response to the varying positions of the rewards. Together, the current evidence of bilingual infants' cognitive advantage is somewhat mixed and the literature is quite limited.

The current paper aims to address this research gap by attempting to replicate Kovacs and Mehler (2009a) in a different population of bilingual infants. Studying whether bilingual infants may have enhanced cognitive skills is a key research question, as it can inform researchers how a variety of language inputs in the early language environment (i.e., processing two language systems) may alter learners' cognitive and/or attention systems. However, we were not able to answer this question based on the current literature as there were only a few studies (n = 3, two of which are unpublished) that have tried to replicate Kovacs and Mehler (2009a) and the limited findings are somewhat mixed. Given that a number of studies have failed to replicate bilingual cognitive advantages in adult participants (e.g., Kousaie & Philips, 2012; Paap & Greenberg, 2013), it is possible that the infant bilingual cognitive advantage is similarly not robust. Replication is a central focus in the current scientific community, with recent specific concerns about replicability in psychological science (Open Science Collaboration, 2015). Infant research is particularly vulnerable to the replication crisis as recruiting and testing infants is difficult, and researchers therefore normally report findings and draw conclusions from small samples (Frank et al., 2017). As such, the current paper contributes to the literature by testing whether bilingual infants would outperform monolingual infants in an experimental paradigm modified from Kovacs and Mehler (2009a) Experiment 3. The current experiment used geometric shapes to convey the abstract rule patterns. This would be a key test of the bilingual cognitive advantage, as researchers have argued that bilinguals' enhanced attention system should transfer to non-verbal cognitive tasks that require attention control (Bialystok, 2017).

There were several methodological differences between our experiment and those reported in Kovacs and Mehler (2009a). First, we tested 9.5-month-old infants, who were slightly older than those reported in Kovacs and Mehler (2009a). As mentioned above, Kovacs and Mehler (2009a) tested infants with abstract rule patterns (i.e., AAB/ABB structures) and paired these structures with different visual rewards at different locations of the screen. Some researchers have raised concerns of testing infants with abstract rules patterns (i.e., extracting patterns from stimuli) as this may quite challenging to participants this young (Comishen et al., 2019). Given that 7-month-old infants can learn some abstract rule patterns (Marcus et al., 1999) and Ibánez-Lillo et al (2010) reported that both monolingual and bilingual infants at 8 months can switch their responses when the cue-reward pairings had changed, we decided to test 9.5-month-old infants, who are a bit older than 8-months, to ensure that the infants were sufficiently mature to handle the task demands. Furthermore, these slightly older infants would have even more bilingual experience, perhaps enhancing any effect of dual language exposure. Second, our abstract rule patterns were conveyed using both visual and auditory (non-linguistic) cues, as presenting

information bimodally should facilitate infant abstract rule learning (Frank et al., 2009). Lastly, we reduced the number of trials during pre-switch and post-switch phases from nine to six. Our decision was based on pilot data (not included in this paper) that revealed 9.5-month-old infants lost interest in the screen after six to seven trials presenting the same abstract rule patterns during the pre-switch phase. Moreover, Kovacs and Mehler (2009a) have shown that 7-month-old bilingual infants began switching their responses during post-switch phase from the fourth trial in Experiment 3. As we were testing older infants, we expected that older infants could switch their responses earlier in the post-switch phase. Thus, we set the number of trials in pre-switch and post-switch phases to six.

Methods

This experiment involved an anticipatory eye movement paradigm, modified from Experiment 3 in Kovács and Mehler (2009a). Infants were trained to predict the locations of visual rewards based on the structures of tone-shape sequences. The key manipulation was that the structure of the tone-shape sequence would change in the middle of the experiment. Successful learners must inhibit the previously learned tone-shape sequence and then learn the new one. Similar to others, we argue that this task measures infants' general cognitive skills, including working memory, attention and inhibitory control. First, infants need to use their working memory to process and track the information in the tone-shape sequences. Next, infants need to pay attention to the common structures across tone-shape sequences and the association between those structures and the locations of visual rewards. Lastly, as mentioned above, infants must rapidly inhibit the previously learned tone-shape sequence structure in order to learn new tone-shape sequence structure during the post-switch phase.

Participants

Twenty-four monolingual English infants and 23 French-English bilingual infants of 9.5 months were tested (Mean = 9.49 months; S.D. = 0.64 months), 22 female). All participants were living in a French-English city in an officially bilingual (French-English) country.

We used the Language Exposure Questionnaire (Bosch & Sebástian-Gallés, 1997) to measure infants' language exposure to English and French. Infants were categorized as monolinguals if they had 90% or greater exposure to English. Infants were categorized as bilinguals if they had a minimum of 20% exposure to one language and a maximum of 80% exposure to the other language (Mean English exposure = 54.96%, SD = 15.64%; Mean French exposure = 44.71%, SD = 15.77%). An additional four infants (2 monolinguals, 2 bilinguals) were tested but not included in the final analysis because of crying or fussiness.

Stimuli

Figure 1 illustrates sample stimuli and the procedure of the study. All stimuli were organized into 3 tone-shape sequences of two structures: AAB or ABB. For AAB sequences, the first two tones and geometric shapes in the sequence were identical and the final pairing differed (e.g., circle-tone C, circle-tone C, star-tone F). By contrast, for ABB sequence, the last two shape-tone pairings were identical and the first pairing differed (e.g., star-tone F, circle-tone C, circle-tone C).

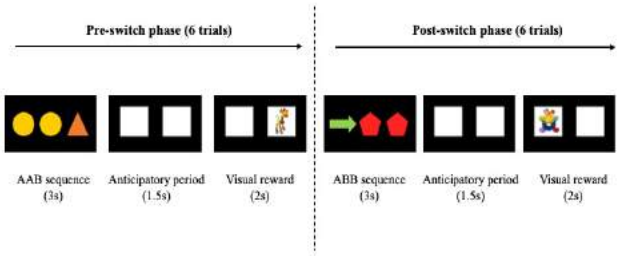


Figure 1. Sample visual stimuli and the procedure of the anticipatory eye movement paradigm.

The visual stimuli comprised six AAB shape sequences and six ABB sequences. Following the methodology in Kovács and Mehler (2009a), specific geometric shapes were used to generate the sequences. Three geometric shapes (arrow, circle, pentagon) were used for position A of the AAB and ABB sequences, whereas three other geometric shapes (moon, 5-pointed star, triangle) were used for position B. Each shape was presented in a different colour. For symmetrical geometric shapes (i.e., circle, pentagon, 5-pointed star and triangle), the size of each shape was 32cm X 32cm. For asymmetric geometric shapes (i.e., arrow and moon), the size of each shape was 34cm X 32cm. The shape sequences appeared on screen in the following manner. The first shape of the sequences was presented on the left side of the screen alone. Next, the second shape was added in the middle of the screen while the first remained onscreen. Finally, the third shape was added on the right side of the screen so that all three shapes appeared simultaneously on the screen for 3 sec. All shape sequences were displayed against a black background.

The audio stimuli were sequences of three musical tones. Two tone structures (i.e., AAB and ABB) were constructed to pair with the corresponding visual shape sequences. Three musical tones (A, D, E) could be paired with objects in position A, whereas three other musical tones (C, F, G) could be paired with objects in position B.

The tone-shape sequence would be followed by a visual presentation of two white squares and the visual reward (see procedure for more details). The two white squares were presented side by side on the screen (each was 52.5cm to 53.5cm in size) for 1.5 sec. For the visual reward, one of two puppets (i.e., a giraffe or hippopotamus toy) appeared inside one of the two white squares on the left or right side of the screen. The puppet loomed from 20cm X 30cm to

28cm X 49cm in size for 2 sec. The presentation of puppets was accompanied by a chime sound.

Apparatus and procedure

Infants were seated on a parent's lap during the experiment. The parent wore headphones playing music with vocals to mask the sounds. The parent was instructed not to turn his/her head to the left-hand side or right-hand side. Instead, parent could either look at the infant or look at the center of the screen in order to minimize their influence of infants' attention to a particular side of the screen. At the beginning of each trial, we presented infants with attention-getting stimuli (e.g., an image of a baby and audio of a baby giggle). Once infants oriented their attention to the screen, the experimenter pressed a key to present the orientation stimuli to the infants. For the first trial, infants saw a video where a rotating ball changed its position from the left side to the right side of the screen. The ball first appeared on the left side and remained onscreen for 3.3 sec. The ball then reappeared on the right side of the screen for another 3.3 sec. The trial served to accustom infants with the experiment procedure where they would be trained to look at the left and right side of the screen to predict different visual rewards based on the tone-shape sequences.

After this orientation trial, there were two phases in the experiment: pre-switch and post-switch (see Figure 1). Each phase consisted of 6 trials. In the pre-switch phase, infants were trained that one tone-shape sequence structure (either AAB or ABB) predicted a visual reward in a particular location (i.e., on the right or left of the screen). On each trial, infants were first presented with a tone-shape sequence (e.g., AAB) for 3 sec. The tone-shape sequence would then be replaced by two white squares on the screen for 1.5 sec. During this window of time, infants could make an anticipatory eye movement by directing their eye gaze to the square where the object would appear (anticipatory window period). Finally, a looming puppet (e.g., giraffe) would appear on one side of the screen (e.g., right side) for 2 sec. After presenting infants with 6 pre-switch trials, infants then entered a post-switch phase where a new structure of tone-shape sequences (e.g., ABB structure) predicted a different reward (e.g., hippopotamus) on a different location (e.g., left side). The procedure and length were identical to the pre-switch phase, aside from the differences above (i.e., sequence structure, visual reward type and location).

Coding

Following Kovács and Mehler (2009a), we coded infants' eye gaze during the anticipatory window period (1.5 sec) for the dependent variable (DV). We coded infants' eye gazes to the right and left positions. We only counted eye gazes to the appropriate reward location as correct responses. Eye gazes to the opposite side of the reward location were all counted as incorrect. Videos were coded frame by frame (30 frames per second). Two trained undergraduate coders coded all trials independently and the reliability between their coding was high (r = 0.90, p < 0.0001). To obtain the proportion of correct anticipatory looks, we divided the number of frames looking at the correct location by the sum of total frames that infants looked at the correct and incorrect positions during anticipatory window period. For example, on a test trial, an infant looked at the correct position for a total of 20 frames and looked at the incorrect position for a total of 10 frames. The proportion of correct anticipatory looks would be 0.67 for this infant on this particular trial.

Data analysis

The DV was infants' proportion of correct anticipatory looks in each trial. Further, we expected that infants would increase the proportion of correct anticipatory looks over time, thus we tested whether the relationship between the DV and the number of trials (i.e., DV-trial slope) was positive or not. We also tested whether infants' language background would influence the DV-trial slope. For example, bilingual infants may have faster rate of learning the association between tone-shape sequence and the corresponding visual rewards. This would be reflected by a steeper DV-trial slope in the bilingual infant group. To model the variations of the DV and the DV-trial slopes in the experiment, we employed hierarchical linear modeling in our analyses. Because we have different hypotheses about infants' performance in the pre-switch and the post-switch phases. We ran two separate hierarchical linear models, one examined infants' performance in the pre-switch phase (pre-switch model) and the other examined infants' performance in the post-switch phase (post-switch model). We predicted that infants' inhibitory control ability would only be reflected in their performance during the post-switch phase. As such, we expected that monolingual and bilingual infants would have similar proportion of correct anticipatory looks and DV-trial slopes during the pre-switch phase. By contrast, monolingual and bilingual infants would differ in terms of their proportion of correct anticipatory looks and DV-trial slopes during the post-switch phase. We used the lme4 package in R (Bates et al., 2015) to perform the hierarchical linear regression models. The regression models were fit by the restricted maximum likelihood approach and the p values in the models were estimated by Satterthwaite approximations in the lmerTest package (Kuznetsova, Brockhoff, & Christensen, 2016). The model was specified as follow: DV ~ trial*language background + (1+trial|Subject) in the lme4 package.

Results

See Figure 2 for infants' performance across trials during the pre-switch and post-switch phases. In the pre-switch model, infants' language background [β = 0.047, *S.E.* = 0.066 p = 0.47] was not a significant predictor. We also

found that the interaction between infants' language background and DV-trial slopes was not significant [β = 0.020, *S.E.* = 0.011, p = 0.997], suggesting that the learning rate between monolingual and bilingual infants were similar in the pre-switch phase. The average DV-trial slope was significantly higher than zero across all participants [β = 0.020, *S.E.* = 0.076, p = 0.012], suggesting that infants made more correct anticipatory looks over time. Finally, a one-tailed t test revealed that the average proportion of correct anticipatory looks in the pre-switch phase was significantly greater than the chance level [M = 0.59, t(276) = 4.909, p < 0.001].

In the post-switch model, we also did not find a significant effect of infants' language background [β = -0.015, *S.E.* = 0.063, p = 0.82]. The interaction between infants' language background and DV-trial slopes was not significant [β = -0.004, *S.E.* = 0.023, p = 0.878]. This again indicates that the learning rate between monolingual and bilingual infants were similar in the post-switch phase. The average DV-trial slope was significantly higher than zero across all participants [β = 0.031, *S.E.* = 0.012, p = 0.009], suggesting that infants also showed a trend of improving their proportion of correct anticipatory looks over time during the post-switch phase. Finally, a one-tailed t test revealed that the average proportion of correct anticipatory looks in the post-switch phase was not significantly greater than the chance level [M = 0.45, t(268) = -2.63, p > 0.99].

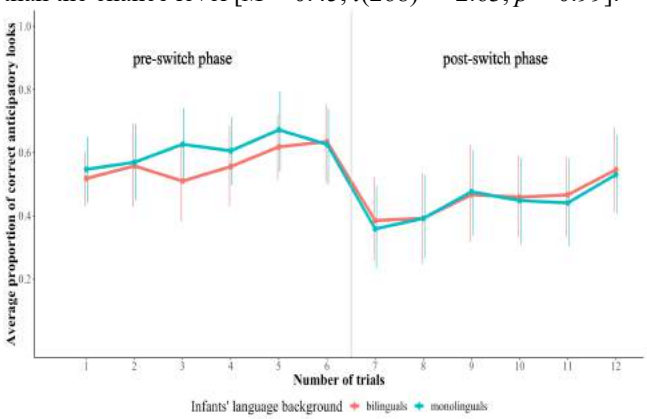


Figure 2. Infants' average proportion of correct anticipatory looks across trials during pre-switch and post-switch phases. Error bars indicate the 95% confidence intervals.

We did some follow up hierarchical linear regression models to further explore whether the degree of bilingual experience influences infants' performance. One aspect of bilingual experience is whether the bilingual infant has balanced exposure to the two languages. The degree of bilingualism was determined by the difference of percentage exposure between dominant language and non-dominant language. For example, an infant with 60% exposure to English and 40% exposure to French would get a score of 20, while an infant with 50% exposure to English and French would get a score of 0. Here, a smaller value in the degree of bilingualism reflected that the infant received a more balanced language exposure to French and English. The degree of bilingualism can be treated as a proxy for the variation of bilingual experience among our bilingual infants. In the following analyses, we only focused on bilingual infants and measured whether the degree of bilingualism contributes to the difference in infants' average proportion of correct anticipatory look during pre-switch and post-switch phases. We specified two separate models for the pre-switch and post-switch phases. Each model was specified as DV~ trial*infants' degree of bilingualism + (1+trial|Subject) in the lme4 package. Across two models, we did not find any significant main effects and interaction between trials and bilingual infants' degree of bilingualism (ps > 0.30). This suggested that bilingual experience (whether they hear more or less balanced exposure in daily environments) did not predict infants' performance in our cognitive task.

Finally, we also computed one-tailed t tests to examine whether infants' average proportion of correct anticipatory look in the last two trials during post-switch phase was significantly greater than chance level. The t test results were both non-significant in trial 11 [M = 0.45, t(40) = -0.95, p = 0.827] and in trial 12 [M = 0.54, t(44) = 0.79, p = 0.217], suggesting that infants' performance across the last two trials did not significantly differ from the chance level.

General Discussion

Our goal was to replicate Kovacs and Mehler (2009a) study and examine the potential positive effects of early bilingualism to infants' cognitive ability. We found that both monolingual and bilingual infants could extract the abstract rule patterns and associate the patterns to particular reward locations. Their performance generally improved as they made more correct anticipatory looks based on the abstract rule patterns over-time. However, we did not find evidence supporting that early bilingualism improves infants' inhibitory control, at least for the experiment tested here. In our study, we tested slightly older infants, who would have even more bilingual experience, and attempted to maximize the learning effects by presenting infants with bimodal visual-audio stimuli. Despite using these manipulations that may enhance any effect of bilingualism, we still found a null effect. Thus, our findings are consistent to recent studies (e.g., Kousaie & Philips, 2012; Paap & Greenberg, 2013) that bilingualism may not have a robust effect on learners' cognition.

Notwithstanding our efforts to maximize the learning effects in the current paradigm, it is important to note that infants' average proportion of correct anticipatory looks during the post-switch phase was not significantly above chance. This implies that infants generally found learning the associations between the tone-shape structure and the visual rewards in the post-switch phase more difficult than

those in the pre-switch phase. Although 7-month-olds were previously reported to succeed in learning using this paradigm (Kovacs & Mehler, 2009a), our results suggest that it may still be cognitively challenging for 9.5-month-olds to learn the new association within the six trials of the post-switch phase. Future work is perhaps needed to explore a range of age-appropriate and simplified cognitive tasks to fully address the question of a correlation between cognitive ability and bilingualism in infancy.

Another possibility is that our null findings may be related to the language contexts in the bilingual infants' home/community. Byers-Heinlein, Morin-Lessard and Lew-Williams (2017) have suggested that bilingual infants' enhanced cognitive ability may be driven by exposure to language mixing contexts in their environments. Language mixing is prevalent in early bilingual environments where one/both parent(s) switch between two languages when speaking to their infants. Byers-Heinlein et al., (2017) have discovered that bilingual French-English infants demonstrated a switching processing cost when hearing sentences that alternated between two languages. The switching costs suggest that bilingual infants need to monitor and control their two languages when listening to speech that mixes languages. Thus, bilingual infants' enhanced cognitive skills may be a result of listening to mixed speech on a daily basis. Although the local area is quite bilingual, perhaps the specific bilingual infants in our study were not living in a home language environment where parents often mix their languages, thus minimizing their need of monitoring and switching between two language systems regularly. In the current paper, we did not collect relevant data to examine this possibility. Future work should explore how the degree of language mixing in early bilingual environment affects infants' cognitive ability.

In conclusion, we did not find support for bilingual cognitive advantages at 9.5 months, suggesting that advantages may not be robustly seen across different bilingual populations or different ages. However, we made note of other possible accounts for the replication failure, including how the bilingual environments and task demands of the current experiment matter. It is our hope that future work can address the existing research gap to further understanding of the effects of early bilingualism on infants' cognitive ability.

References

Bialystok, E. (2006). Effect of bilingualism and computer video game experience on the Simon task. *Canadian Journal of Experimental Psychology, 60(1),* 68-79.

Bialystok, E. (2015). Bilingualism and the development of executive function: The role of attention. *Child Development Perspectives, 9(2),* 117-121.

Bialystok, E. (2017). The Bilingual Adaptation: How Minds Accommodate Experience. *Psychological Bulletin, 143,* 233-262.

Bialystok, E., Craik, F. I. M., Klein, R., & Viswanathan, M. (2004). Bilingualism, aging, and cognitive control: Evidence from the Simon task. *Psychology and Aging, 19,* 290-303.

Bialystok, E., Craik, F. I. M., & Luk, G. (2012). Bilingualism: Consequences for mind and brain. *Trends in Cognitive Sciences, 16,* 240-250.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67,* 1.

Bosch, L., & Ramon-Casas, M. (2011). Variability in vowel production by bilingual speakers: Can input properties hinder the early stabilization of contrastive categories?. *Journal of Phonetics, 39(4),* 514-526.

Byers-Heinlein, K. (2013). Parental language mixing: Its measurement and the relation of mixed input to young bilingual children's vocabulary size. *Bilingualism: Language and Cognition, 16(1),* 32-48.

Byers-Heinlein, K., Morin-Lessard, E., & Lew-Williams, C. (2017). Bilingual infants control their languages as they listen. *Proceedings of the National Academy of Sciences, 114,* 9032-9037.

Comishen, K. J., Bialystok, E., & Adler, S. A. (2019). The Impact of Bilingual Environments on Selective Attention in Infancy. Developmental science, e12797.

Costa, A., Hernández, M., & Sebastián-Gallés, N. (2008). Bilingualism aids conflict resolution: Evidence from the ANT task. *Cognition, 106,* 59–86.

Fennell, C. T., Tsui, A. S. M., & Hudon, T. M. (2016). Speech perception in simultaneously bilingual infants. In S. Montanari & E. Nicoladis (Eds.) *Bilingualism across the lifespan: Factors moderating language proficiency.* Washington, USA: American Psychological Association.

Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., ... & Lew-Williams, C. (2017). A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. Infancy, 22(4), 421-435.

Frank, M. C., Slemmer, J. A., Marcus, G. F., & Johnson, S. P. (2009). Information from multiple modalities helps 5-month-olds learn abstract rules. *Developmental Science, 12*(4), 504–509.

Ferjan Ramìrez, N., Ramìrez, R. R., Clarke, M., Taulu, S., & Kuhl, P. K. (2017). Speech Discrimination in 11-Month-Old Bilingual and Monolingual Infants: A Magnetoencephalography Study. *Developmental Science, 20(1),* e12427.

Green, D. W. (1998). Mental control of the bilingual lexico-semantic system. *Bilingualism: Language and Cognition, 1*, 67–81.

Ibánez-Lillo, A., Pons, F., Costa, A., & Sebastián-Gallés, N. (2010). *Inhibitory control in 8-month-old monolingual and bilingual infants: Evidence from an anticipatory eye movement task.* Poster presented at the 22nd Biennial International Conference on Infant Studies, Baltimore, MD.

Kovács, Á. M., & Mehler, J. (2009a). Cognitive gains in 7-month-old bilingual infants. *Proceedings of the National Academy of Sciences, 106(16),* 6556-6560.

Kovács, Á. M., & Mehler, J. (2009b). Flexible learning of multiple speech structures in bilingual infants. *Science, 325(5940),* 611-612.

Kousaie, S., & Phillips, N. A. (2012). Conflict monitoring and resolution: Are two languages better than one? Evidence from reaction time and event-related brain potentials. *Brain Research, 1446,* 71-90.

Kuznetsova, A., Brockhoff, P.B., Christensen, R. H. B. (2016). *lmerTest: Tests in Linear Mixed Effects Models.* R package version 2.0-29. Retrieved from: http://CRAN.R-project.org/package=lmerTest.

Marcus, G. F., Vijayan, S., Rao, S. B., & Vishton, P. M. (1999). Rule learning by seven-month-old infants. Science, 283(5398), 77-80

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex ‘‘frontal lobe’’ tasks: A latent variable analysis. Cognitive Psychology, 41, 49–100.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349.*

Paap, K. R., & Greenberg, Z. I. (2013). There is no coherent evidence for a bilingual advantage in executive processing. *Cognitive Psychology, 66(2),* 232-258.

Pour Ilyaei, S., & Byers-Heinlein, K. (2018, July). *Cognitive capacity in infancy: How is it linked to bilingualism?* Poster presented at the International Congress on Infant Studies, Philadelphia, PA.

Thierry, G., & Wu, Y. J. (2007). Brain potentials reveal unconscious translation during foreign language comprehension. *Proceedings of the National Academy of Sciences of the United States of America, 104,* 12530–12535.

Draping an Elephant: Uncovering Children's Reasoning About Cloth-Covered Objects

Tomer D. Ullman (tomeru@mit.edu)
Psychology, Harvard University

Eliza Kosoy (eko@mit.edu)
Psychology, UC Berkeley

Ilker Yildrim (ilkery@mit.edu)
Brain and Cognitive Sciences, MIT

Amir A. Soltani (arsalans@mit.edu)
Brain and Cognitive Sciences, MIT

Max Siegel (maxs@mit.edu)
Brain and Cognitive Sciences, MIT

Joshua B. Tenenbaum (jbt@mit.edu)
Brain and Cognitive Sciences, MIT

Elizabeth S. Spelke (spelke@wjh.harvard.edu)
Psychology, Harvard University

Abstract

Humans have an intuitive understanding of physics. They can predict how a physical scene will unfold, and reason about how it came to be. Adults may rely on such a physical representation for visual reasoning and recognition, going beyond visual features and capturing objects in terms of their physical properties. Recently, the use of draped objects in recognition was used to examine adult object representations in the absence of many common visual features. In this paper we examine young children's reasoning about draped objects in order to examine the develop of physical object representation. In addition, we argue that a better understanding of the development of the concept of cloth as a physical entity is worthwhile in and of itself, as it may form a basic ontological category in intuitive physical reasoning akin to liquids and solids. We use two experiments to investigate young children's (ages 3–5) reasoning about cloth-covered objects, and find that they perform significantly above chance (though far from perfectly) indicating a representation of physical objects that can interact dynamically with the world. Children's success and failure pattern is similar across the two experiments, and we compare it to adult behavior. We find a small effect, which suggests the specific features that make reasoning about certain objects more difficult may carry into adulthood.

Keywords: intuitive physics, cloth, cognitive development, object recognition, analysis-by-synthesis

Introduction

Imagine draping an elephant. What shape do you see? Probably not an exact silhouette, but a rough outline with a coarse bottom (Figure 1). This mental image also likely changes as you imagine draping an elephant placed on its side, or turned upside down. This simple feat of the imagination is quite remarkable. Imagining an elephant on its own may involve reactivating a learned representation or a visual memory of an elephant, but imagining an elephant draped by a cloth means 'seeing' a new object (the reader with extensive experience of draped elephants is free to imagine some other animal here). How do we come to this new image? One possible account is that we run a mental simulation and examine the outcome under noisy dynamic laws. But such a simulation requires object representations that go beyond representing image patches. Under this account, objects are represented as three-dimensional bodies, and the mental simulation is able to imagine the transformation and variation of the object under different processes. By examining people's ability to reason about the outcome of draping or to perceive draped object, we examine people's ability to reason visually without most of the traditional visual features that are assumed to play a part in recognition (Yildrim et al., 2016).

Recently, Yildrim, Siegel, and Tenenbaum (Yildrim et al., 2016) have investigated adult reasoning about cloth-covered objects as part of a larger examination of people's object representation as physical entities with the properties necessary for physical interaction. These studies showed that adults can reliably reason about the identity of covered objects in a match-to-sample task, even when the distractor object is within the same category type as the target object. Adult responses were best captured by a model based on a physics and graphics engine, which formalize the proposal that adults base their recognition and reasoning in part on a physical model of objects and a causal dynamical model of their interactions with the world.

More broadly, the Mental Physics Engine proposal suggests that the representations underlying much commonsense physical and visual reasoning are similar to those of modern game engines, software that is useful for quickly rendering an approximate simulation of a physical environment (see e.g. Battaglia et al., 2013; Gerstenberg et al., 2012; Smith & Vul, 2013; Hamrick et al., 2016; Ullman et al., 2017). Such game engines have also been proposed as an essential part of machine intelligence for commonsense reasoning (see e.g. Wu et al., 2015; Lake et al., 2017; Chang et al., 2017). While such a physics engine proposal predicts adult recognition and perception better than neural-network models based on visual image features, it is possible that adults come to this sophisticated understanding of objects and physics over time. Much less is known about children's representation of objects as physical objects for recognition. Here, we propose to examine young children's reasoning about draped objects as a way of examining the development of understanding objects as physical bodies, and of the causal processes that determine the behavior of objects.

Beyond this, we suggest that examining the development of intuitions about cloth is of interest in and of itself. This is because *cloth* (in the sense of a mesh or sheet of connected point masses, which can capture entities such as blankets, towels, and clothes) may be a basic ontological category in intuitive physical reasoning, akin to *rigid body* or *fluid*. At a high level, game engines separate physical entities into several broad classes based on their expected behavior, and the computational resources necessary to simulate them. This

high-level division is limited to only a few classes, and one of the common classes in modern engines is *cloth*, required specialized modular simulation Gregory (2009), and suggesting this may form a basic mental category as well. So, while it may initially seem that there are a large number of intuitive physical categories that can be investigated, of which cloth forms only a small subset, the success of the game engine approach to mental reasoning motivates us to focus on the small number of broad categories that have proven useful for engineers.

While cloth exists as a separate category in modern game engines, it is not obvious that an understanding of cloth has its origin in childhood. On one hand, by their first year many children have extensive experience with clothes, blankets, towels, tissues, and so on. A general mental physics engine with the right computational primitives may use this experience to generate the cloth category. On the other hand, our core knowledge physical reasoning is shared with many other animals and is believed to have a long evolutionary past (Spelke & Kinzler, 2007). Cloth, unlike liquids and rigid bodies, is a relatively recent category, and early human ancestors would not have needed to reason about it on a daily basis. Thus the mental physics engine may lack the right primitives to quickly construct this category.

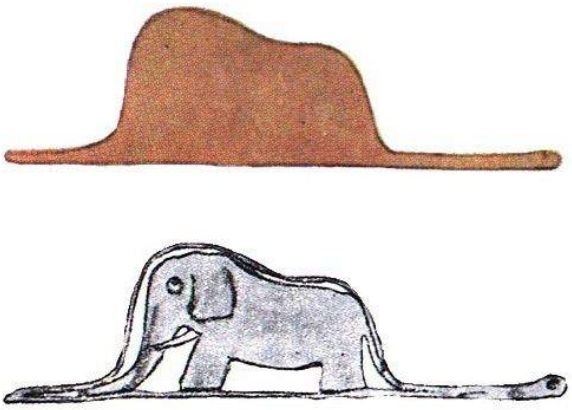


Figure 1: "My drawing was not a picture of a hat. It was a picture of a boa constrictor digesting an elephant. Then, I drew the inside of the boa constrictor, so that the grown-ups could see it clearly. They always need to have things explained." *The Little Prince*, by Antoine de Saint Exupéry.

In this paper, we probe young children's ability to reason about cloth using two basic tasks: reasoning from an uncovered object to a covered image (Experiment 1), and reasoning from a covered object to an uncovered image (Experiment 2). These tasks do not span the full space of the possible behavior of cloth, but they are meant to establish the existence (or lack) of basic competency. We consider an age range of 3–5 years, when children have for the most part not started a formal education, yet possess a sufficiently large vocabulary to understand the language used in the task. We find that children perform above chance in both tasks, and use an adult comparison to examine their patterns of success and failures. In the General Discussion, we consider the implication for generative vs. feature-based models, and the extension of cloth studies to infants.

Experiment 1: Uncovered → Covered

Participants

Sixteen individuals ($N = 16$, 5 female, median age 3.9 years, range 3.2-4.8) were recruited at the [City] Children's Museum. The size of the sample was pre-specified, based on a pilot study which indicated medium-to-large effect sizes can be expected.

Materials and methods

Participants were tested in a designated area in the [City] Children's Museum. Parents gave their informed consent, and advised not to encourage responses from their child.

Participants were presented with a touch-screen device (iPad), and told that they were going to play a game. Participants first played two warm-up rounds, in which they were shown a test-object on top of the screen (e.g. a bird), and asked to match it with one of two possible objects below (e.g. a bird and a horse). The warm-up round was meant to familiarize the participants with making a forced choice between two items based on a target item. By the second warm-up all participants correctly selected the matching object.

During test, participants saw 6 trials in succession, in random order (see Figure 2, top). Each trial contained a pair of objects, for example a mug and a bench. One object in the pair was randomly selected as the test object. The test object was shown at the top of the screen, uncovered. Below the test object were the pair of objects, covered in cloth. Participants were told to imagine that the test object had been covered by a blanket, and asked to indicate what the resulting image would be. Participant choices were automatically stored. Participants were given general encouragement, but no indication of whether their choice was correct.

All the stimuli pairs used in the experiments are shown in Figure 3. Uncovered and covered stimuli images were created in Blender (Blender, 2015). Covered objects were created by draping the uncovered objects using a physical cloth simulation. Objects were chosen from a collection of available objects previously used in experiments with adults (Yildrim et al., 2016). The size of the objects was scaled such that they took up approximately the same amount of visual space when covered.

Results and analysis

Participants' responses were summed across the object pairs, and are shown in Figure 4 (left). The summation resulted in a labeling score going from 0 (no objects correctly identified) to 6 (all objects correctly identified), with chance performance at 3. On average, participants correctly labeled 4.14 objects (95% CI 3.54–4.68, bootstrapped with 10,000 samples). The

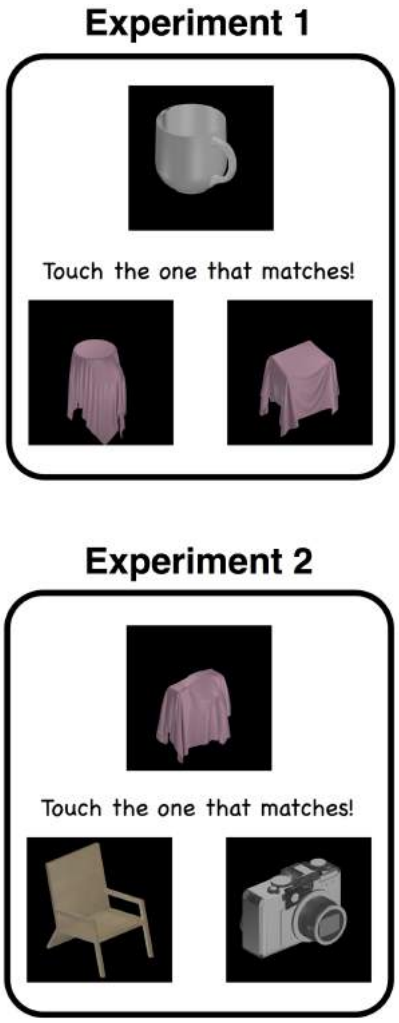


Figure 2: Schematic of example test trials in Experiments 1 and 2. At the top of a touchscreen is the target object (Uncovered in Experiment 1, covered in Experiment 2). Participants were asked to match the target object to one of the pair of objects at the bottom of the screen (Covered in Experiment 1, uncovered in Experiment 2).

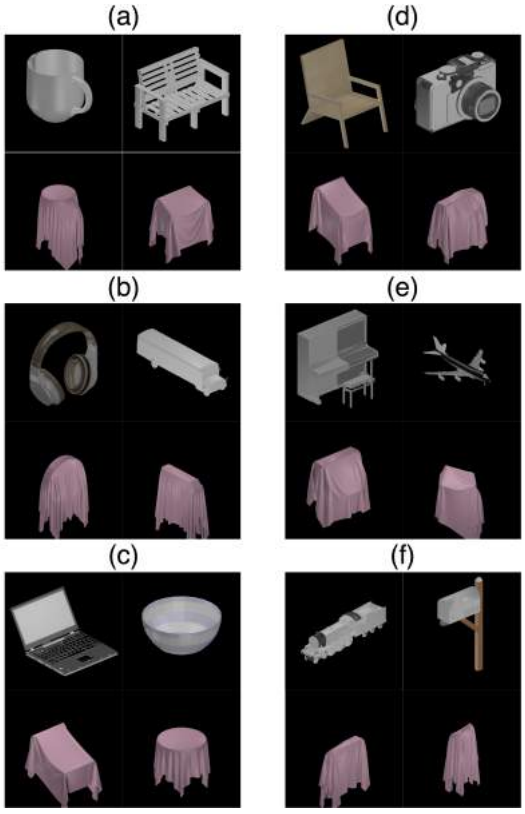


Figure 3: All stimuli pairs used in Experiments 1 and 2, uncovered and covered. In Experiment 1 participants saw one of the objects in the top row as the target, and matched it to the two items in the bottom row. In Experiment 2, participants saw one of the objects in the bottom row as the target, and matched it to one of the items in the top row.

confidence intervals are clearly above chance performance, and a standard two-sided T-test also indicates this result is statistically significant ($t(15) = 3.09, p < 0.01$).

We did not predict nor find a significant effect of age on participant performance. A logistic regression of labeling score on age was not significant, and neither was a median split comparison. Given the small sample size, however, we do not take this to strongly indicate the non-existence of an age effect, but simply the lack evidence for it.

Considering the stimuli by pair, we found that the identity of the objects in a given pair had an effect on participants' labeling. That is, some pairs were harder to discern than others. Specifically, using a standard two-sided binomial test at the $p < 0.05$ level, participants correctly distinguished mug/bench, headphones/bus, and laptop/bowl (Figure 3 a, b, c). Participants were unable to distinguish mailbox/train, piano/airplane, and chair/camera (Figure 3 d, e, f). The exact number of participants correctly labeling the objects by pair is shown in Figure 5.

Experiment 2: Covered → Uncovered

We took the results of Experiment 1 to indicate pre-school children have a general ability to match objects to their cloth-covered representations, though they may have been using one of several different strategies to do so. We next examine whether children were able to go in the inverse direction, inferring the identity of an object hidden under cloth.

Participants

Seventeen individuals ($N = 16$, 5 female, median age 4.0 years, range 3.0-5.0) were recruited at the [City] Children's Museum. The size of the sample was pre-specified at 16 to match Experiment 1.

Materials and methods

Participants were tested in a designated area in the [City] Children's Museum. Parents gave their informed consent, and

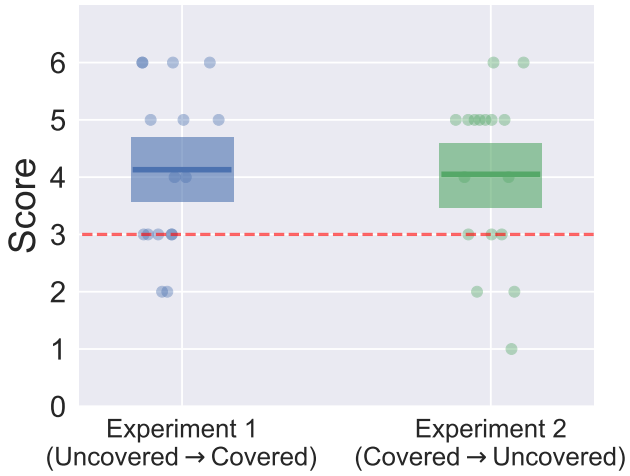


Figure 4: **Left**: Results of Experiment 1, seeing uncovered object and matching to cloth-covered image. **Right:** Results of Experiment 2, seeing cloth-covered image and matching to uncovered object. Score ranges from 0 (no trials correct) to 6 (all trials correct). Bold lines indicate mean score, and shaded colored area indicates 95% CI. Dashed red line indicates chance performance. Each dot indicates the response of one participant, jittered for visibility.

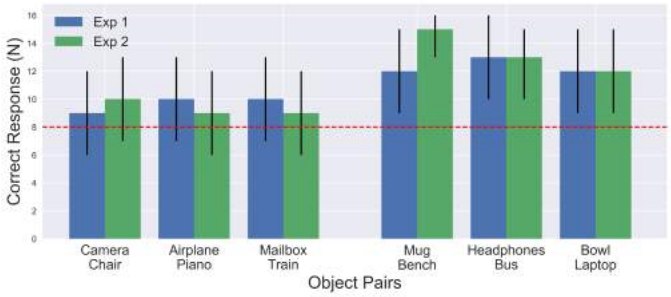


Figure 5: Results of Experiments 1 and 2 by object pair. The number of participants who correctly identified the target object is shown against specific object pairs. Black lines indicate 95% CI, dashed red line indicates chance performance. Children performed at chance or above chance levels for the same object pairs in both experiments.

advised not to encourage responses from their child.

Prior to the touch-screen part of the experiment, participants were shown 6 images of covered items in succession (printed on paper), and asked what they thought was under the cloth covering. That is, participants provided a free-form verbal response. The experimenter did not provide feedback on whether the response was correct or incorrect.

The touch-screen part of the experiment was similar to Experiment 1. Participants were shown an iPad, and told that they were going to play a game. As in Experiment 1, participants first engaged in two warm-up trials, and by the second trial all participants correctly labeled the matching object.

The test phase was also similar to Experiment 1: participants saw 6 trials in succession, in random order. Each trial contained a pair of objects, using the same pairs as in Experiment 1. However, in this experiment, the test object in a pair was covered by cloth, and the two objects below it were uncovered (see Figure 2). Participants were asked to indicate which of the two objects was under the cloth. Participant choices were automatically stored. As before, participants were given general encouragement, but were not told whether their choice was correct.

Results and analysis

The verbal response of participants to the first part of the task (freeform response when prompted to guess what is under a cloth) is summarized in Table 1. We did not predict that participants would correctly guess what was under a cloth, rather we used this task to examine the range of possible guesses. Note that many of the participants responded 'table' as this was a salient object mentioned by the experimenter.

Participants' responses to the forced-choice part of the task were summed across the object pairs, and are shown in Figure 4 (right). The summation resulted in a score going from 0 (no objects correctly identified) to 6 (all objects correctly identified), with chance performance at 3. On average, participants correctly labeled 4.26 objects (95% CI 3.66–4.74, bootstrapped with 10,000 samples). The confidence intervals are above chance performance, and a standard two-sided T-test also indicates this result is statistically significant ($t(15) = 3.87, p < 0.01$).

As in Experiment 1, a logistic regression of labeling score on age was not significant, and neither was a comparison which split participants by median age. We again stress that while we did not expect an age effect, we also do not believe these results necessarily indicate a 'true null' (the nonexistence of an age effect), merely a lack evidence for it.

The identity of the objects in a given pair again had an effect on participant labeling. Interestingly, the exact same pattern emerged when using a standard two-sided binomial test at the $p < 0.05$ level. That is, in Experiment 2 participants correctly distinguished mug/bench, headphones/bus, and laptop/bowl, but did not distinguish mailbox/train, piano/airplane, and chair/camera. Figure 5 shows the performance of participants by pair.

We considered two hypotheses regarding the observation of the same pattern of successes and failures in both experiments:

- H1: Children's performance on both tasks is unrelated
- H2: Children's cloth-related reasoning is affected by object properties due to underlying object features.

We captured hypothesis H1 by assuming children's response is the result of informed inference (a biased coin with weight $\theta = 0.8$) or a random guess ($\theta = 0.5$), and that there are 3 weighted coins and 3 random coins per each experiment, but they are unrelated across experiments. The value of the weighted coin reflects an average of participant performance across the two experiments. We captured hypoth-

Covered object	Verbal description
Chair	table (5), box (3), chair (2), monster (1)
Camera	table (3), box (1), present (1)
Bench	table (3), square box (2), tall present (1)
Mug	table (3), chair (2), box (1), mountain (1), couch (1), squiggle strips (1), circle (1)
Laptop	table (3), square table (1), box(1), dot (1), rectangle (1), square (1), bridge (1)
Bowl	table (3), round table (1), circle (1), chair (1)
Mailbox	table (2), box (1), ghost (1), cat (1), blaster (1), ice-cube (1), gate (1), boat (1)
Train	box (1), chair (1), fence (1), stepstool (1),
Airplane	table (2), present (1), cowboy (1), vacuum cleaner (1), surfboard (1)
Piano	house (3), box (2), ladder (1), table (1), chair or table (1)
Headphones	rainbow machine (1), dog (1), ball (1), mountain (1), band-aid (1), diamond (1), front of crib (1), chair (1), jelly-fish (1), table (1)
Bus	box (1), square (1), fountain (1)

Table 1: Verbal responses of participants in Experiment 2. Numbers in brackets indicate the number of participants giving the preceding response. Numbers do not add up to the total number of participants as not all participants replied in all trials.

esis H2 by assuming the same set-up as hypothesis H1, but with the additional assumption that the weighted coins are matched with the same object pairs in both experiments. Assuming an uninformed uniform prior over both hypotheses, we can assess K, the Bayes factor of the two hypotheses, by estimating the ratio of the likelihood of the data under each hypothesis: $K = \frac{P(H2|D)}{P(H1|D)}$. The data under consideration is passing 3 binomial tests for each experiment, for the same object pairs. Using a bootstrap analysis in which 16 simulated participants have their behavior sampled from the coins described for H1 and H2, using 10,000 samples, we find a Bayes factor of $K = 21$, indicating strong evidence in favor of H2. Put briefly, the 'suspicious coincidence' that children are able to distinguish the same 3 pairs in both experiments is indicative of underlying features of the objects interacting with cloth-based reasoning.

Experiment 3: Adult comparison

While pre-school children were able to overall correctly reason about cloth-covered objects, they also made characteristic mistakes, indicating an underlying difficulty in reasoning about how particular objects will interact with cloth. Such difficulties may be due to simple lower-level feature interaction (for example, covering the mailbox and train both result in elongated rectangular shapes), or due to the end-result of a coarse draping simulation resulting in similar images, or a different reason altogether. Whatever the source of the difficulty, we wanted to examine whether it carried into adulthood. In the next experiment we examined the targeted prediction that adults would overall do worse on the trials that children failed.

Participants

One-hundred and twenty (N=120) participants were recruited online via Amazon Mechanical Turk. Eleven participants were discarded after failing to answer a catch question, and the remaining participants (N=109) are considered in the analysis below ($Median_{age}$ = 33 years, age range 20–70, 48 self-identified as female). We anticipated the task would be easy for adults, and based the number of participants on the expectation of small effect sizes.

Materials and methods

Participants were shown 6 trials, similar to Experiment 1. For each trial, participants were shown a target object and asked to imagine it covered with cloth. On a following page, participants were asked to select which of two covered objects matched the target object. The object pairs were the same as those used in Experiment 1. The order of presentation, the right/left location of the covered objects, and the identity of the target object were all randomized. Participants were asked to respond as quickly as possible. At the end of the 6 trials participants were asked to describe their task was in the study, and irrelevant answers (e.g. 'opinion', 'work', '0') led to discarding their data prior to analysis. Participants were also asked to provide information regarding their age and gender.

Results and analysis

Participants responded within about a second of presentation, with a median response of 1.1 seconds (95% CI 1.04–1.16) per trial. Participants also found the task relatively simple, with an average success rate of 98% (95% CI 96%–99%) per trial.

We considered the average correct response rate for the objects children found easier ('Children pass') and harder ('Children fail'). The average correct response rate by adults for the 'Children pass' trials was 99% (95% CI 98%–100%), whereas the correct response rate for the 'Children fail' trails was 96% (95% CI 94%–98%). The bootstrapped distribution over these variables and the response rate per object pair is shown in Figure 6.

The average correct response rate of adults per trial appears higher for the pairs that children found easier in Experiments 1 and 2, but this effect is very small as adults are nearly at ceiling.

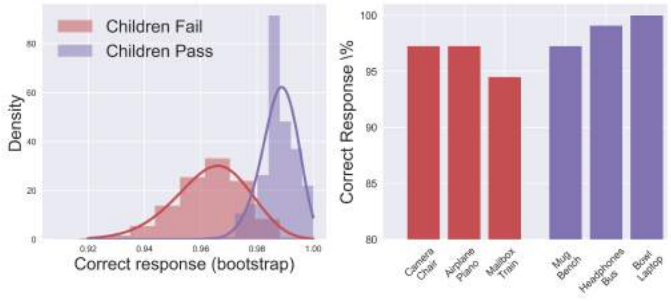


Figure 6: Results of Experiment 3. **Left**: bootstrap posterior distribution over the correct response rate aggregated by trial type ('Children pass' and 'Children fail'), 10,000 samples. **Right**: Correct response rate for each object pair, sorted by trial type.

General Discussion

People can reason intuitively about how things drape, wrap, envelop, sag, and droop. Recent experiments (Yildrim et al., 2016) have shown that adults perform well in a task that requires matching a covered and uncovered object, and that this ability can be captured by a physics and graphics engine which approximately simulates the draping of an object. Motivated by this work, as well as by the general category of 'cloth' in current game engines, we examined whether pre-schoolers can also reason about the interaction of cloth and rigid objects, and found their performance to be above chance in two such tasks. Children's pattern of failure and success was similar across the tasks, and a comparative task with adults found a small effect, suggesting that they too find the same object pairs hard or easy.

The current studies warrant tentative conclusions regarding object representation and the use of dynamic mental simulation in children. Previous studies with adults (Yildrim et al., 2016) rotated the objects, in a way that prevented simple feature-matching and meant in part to examine whether the adults were relying on a generative model reconstructions of the object. We did not use such a rotation in our studies, and we see them as a first step to examine whether children have *any* competence with cloth-based reasoning. It is possible that children's abilities rely on relatively simple feature matching, while adult reasoning is based more on reconstructing a mental representation of the 3D object shape It is also unclear which of several proposals for a generative model of 3D objects (whether for children or adults) is the right one (and see for example Soltani et al. 2017, for a comparison of several such methods for recovered objects from silhouettes). Further studies will need to use object rotations and a wider array of object pairs to examine this question.

The dynamics of cloth go beyond draping objects. For example, cloth sags when objects are placed on top of it, to a degree dependent on internal parameters related to its stiff and stretch. Can children reason about the likely sag of a piece of fabric, based on seeing its motion and knowing an object's felt weight? Are children sensitive to the weight of cloth, or will they reason about it as a weightless 2 dimensional manifold that only interacts geometrically with objects?

Even if both adults and young children rely on similar representations for reasoning about cloth, it is possible that these representations develop late compared to the basic expectations that infants have about rigid bodies (which innate or extremely early developing) and about liquids (which develop over the first year of life). Looking time experiments with infants could test this possibility by familiarizing infants to either cloth or a rigid body of similar proportions and texture, followed by an interaction in which the cloth and rigid body collide with or drape rigid objects.

To wrap up, while many issues remain hanging, this work begins to uncover the origin of cloth-based reasoning, which may form a separate ontological category within intuitive physical reasoning. It opens the door to future research probing the richness and origins of children's reasoning about a human invention that is ubiquitous in human cultures, and that occupies an interesting middle ground between rigid objects and amorphous stuff.

Acknowledgments

We wish to thank the parents and children who participated in the research carried out in [location]. This material is based upon work supported by [center], funded by [funding].

References

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, *110*, 18327–32.

Blender. (2015). Blender - a 3d modelling and rendering package [Computer software manual]. Blender Institute, Amsterdam. Retrieved from `http://www.blender.org`

Chang, M. B., Ullman, T., Torralba, A., & Tenenbaum, J. B. (2017). A compositional object-based approach to learn-

ing physical dynamics. In *Proceedings of the 5th annual international conference on learning representations.*

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2012). Noisy Newtons: Unifying process and dependency accounts of causal attribution. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th annual conference of the cognitive science society.*

Gregory, J. (2009). *Game engine architecture*. CRC Press.

Hamrick, J. B., Battaglia, P. W., Griffiths, T. L., & Tenenbaum, J. B. (2016). Inferring mass in complex scenes by mental simulation. *Cognition*, *157*, 61–76.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*.

Smith, K. A., & Vul, E. (2013). Sources of Uncertainty in Intuitive Physics. *Topics in Cognitive Science*, *5*, 185–199.

Soltani, A. A., Huang, H., Wu, J., Kulkarni, T. D., & Tenenbaum, J. B. (2017). Synthesizing 3d shapes via modeling multi-view depth maps and silhouettes with deep generative networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1511–1519).

Spelke, E. S., & Kinzler, K. D. (2007). Core knowledge. *Developmental science*, *10*(1), 89–96.

Ullman, T. D., Spelke, E. S., Battaglia, P., & Tenenbaum, J. B. (2017). Mind games: Game engines as an architecture for intuitive physics. *Trends in cognitive sciences*, *21*(9), 649–665.

Wu, J., Yildirim, I., Lim, J. J., Freeman, B., & Tenenbaum, J. (2015). Galileo: Perceiving Physical Object Properties by Integrating a Physics Engine with Deep Learning. In *Advances in neural information processing systems* (pp. 127–135).

Yildrim, I., Siegel, M. H., & Tenenbaum, J. B. (2016). Perceiving Fully Occluded Objects via Physical Simulation. In *Proceedings of the 38th annual conference of the cognitive science society.*

Complexity and learnability in the explanation of semantic universals of quantifiers

Iris van de Pol, Shane Steinert-Threlkeld, Jakub Szymanik
{I.P.A.vandePol, S.N.M.Steinert-Threlkeld, J.K.Szymanik}@uva.nl
Institute for Logic, Language and Computation, University of Amsterdam

Abstract

Despite wide variation among natural languages, there are linguistic properties universal to all (or nearly all) languages. An important challenge is to explain why these linguistic universals hold. One explanation employs a learnability argument: semantic universals hold because expressions that satisfy them are easier to learn than those that do not. In an exploratory study we investigate the relation between learnability and complexity and whether the presence of semantic universals for quantifiers can also be explained by differences in *complexity*. We develop a novel application of (approximate) Kolmogorov complexity to measure fine-grained distinctions in complexity between different quantifiers. Our results indicate that the monotonicity universal can be explained by complexity while the conservativity universal cannot. For quantity we did not find a robust result. We also found that learnability and complexity pattern together in the monotonicity and conservativity cases that we consider, while that pattern is less robust in the quantity cases.

Keywords: semantic universals; generalized quantifiers; Kolmogorov complexity; learnability

Introduction

Even though there is huge variability between natural languages, they still share many common features. Such universal linguistic properties have been found at many levels of analysis: phonology (Hyman, 2008), syntax (Chomsky, 1965; Newmeyer, 2008), and semantics (Barwise & Cooper, 1981). Confronted with attested linguistic universals, the question naturally arises: why these properties? What explains the presence of the particular observed universals across languages?

In search of an explanation in terms of the interaction between linguistics and the specifics of human cognition, several theories have presented some form of learnability as an explanation of the presence of semantic universals (see, e.g., Barwise & Cooper, 1981; Keenan & Stavi, 1986; Szabolcsi, 2010). Recently, Steinert-Threlkeld and Szymanik (in press, henceforth ST&S) provided evidence for a version of this learnability hypothesis by using recurrent neural networks as a model for learning and applying this to several different semantic universals.

In this paper, we ask whether these semantic universals could also be explained by some measure of *complexity*, and whether this provides similar results as using a measure of learnability. It is a common expectation that there will be a connection between learnability and complexity and many theories of learning are built around such a connection (Tiede, 1999; Hsu, Chater, & Vitányi, 2013). At the same time, there are few examples that provide evidence for this expectation in concrete cognitive tasks. In particular, it remains open whether a connection between learnability and complexity exists for independently motivated measures of each of these factors in specific domains. In the present work, we study the meaning of generalized quantifiers and compare their complexity (in a sense to be made precise) with the learnability results of ST&S.

The complexity of generalized quantifiers has been intensively studied using methods from logic, automata theory, and computational complexity.[1] However, as we will explain in more detail in a later section, none of these theories have developed a notion of complexity that applies to all quantifiers and can capture the difference between those that are attested and non-attested in natural language. To overcome these limitations, in this paper we propose to evaluate the complexity of quantifiers from an information-theoretic perspective. This perspective has already proven fruitful as an explanatory device in linguistics (Gibson et al., 2019). More specifically, we suggest to adopt (approximate) Kolmogorov complexity (Li & Vitányi, 2008) as a measure for the complexity of quantifiers. Kolmogorov complexity roughly measures how much regularity exists in a string, which enables it to be described by a shorter program that generates it. It is not implausible that universals will have the function of creating "patterns" that enable such compression.

The paper is structured as follows. In the next section, we present generalized quantifier theory and the semantic universals that we will discuss. We also discuss a recent explanation of semantic universals in terms of learnability and previous approaches to measuring the complexity of quantifiers and their limitations with respect to the current study. Following that, we introduce Kolmogorov complexity and a tractable approximation to it, and we explain how we apply this measure to binary encodings of quantifiers. In the section after that, we apply this complexity measure to the same pairs of quantifiers as in the recent learnability study to see (i) whether—in addition to learnability—some of the attested semantic universals can be explained by differences in complexity and (ii) whether complexity and learnability pattern together. We conclude by discussing the results and outlining future work.

[1] See Szymanik (2016) for an overview.

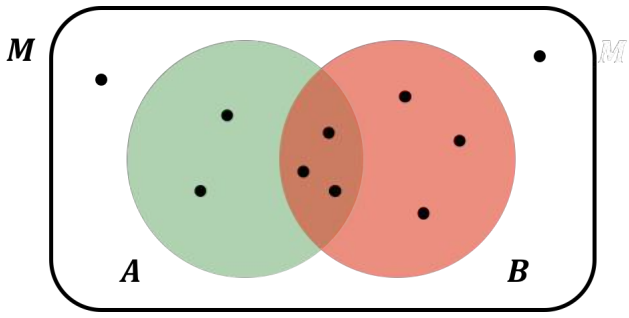


Figure 1: An example of a quantifier model $\mathcal{M} = \langle M, A, B \rangle$ with 10 objects, shown as a vendiagram. This model verifies quantifiers `some` and `most`, but does not varify `all`

.

Quantifiers and their universal properties

Quantifiers are the semantic objects that are expressed by determiners, such as `some`, `most`, or `all`. Determiners are expressions that can combine with common nouns and a verb phrase in simple sentences of the form *Det N VP*, like "some houses are blue". We assume a distinction of the determiners into the grammatically simple (e.g. `some`, `few`, `many`) and the grammatically complex (e.g. `at least 6 or at most 2`, `an even number of`).

We use the framework of generalized quantifiers to represent the meaning of quantifiers as sets of sets. In particular, determiners denote type $\langle 1,1 \rangle$ generalized quantifiers, which are sets of models of the form $\mathcal{M} = \langle M, A, B \rangle$, where M is the domain of the model, and A, B are two unary predicates (that is: $A, B \subseteq M$).[2] See Figure 1 for an illustration. This is an extensional representation of meaning, in which a quantifier is defined as the class of all models satisfying a given property (corresponding to the situations in which a simple sentence with that quantifier would be true). For a given model, $\mathcal{M}$, and quantifier Q we write $\mathsf{Q} \in \mathcal{M}$ if and only if: $\mathcal{M} \models \mathsf{Q}(A,B)$. For example, the meaning of the quantifiers `some`, `most`, and `every` can then be represented as follows:

$$[\![\texttt{some}]\!] = \{\langle M,A,B\rangle : |A \cap B| \neq \emptyset\},$$
$$[\![\texttt{most}]\!] = \{\langle M,A,B\rangle : |A \cap B| > |A \setminus B|\},$$
$$[\![\texttt{every}]\!] = \{\langle M,A,B\rangle : A \subseteq B\}.$$

The semantic universals that we consider build on specific properties of generalized quantifiers, namely *monotonicity*, *quantity*, and *conservativity*. Let Q be a generalized quantifier. Then we call Q *monotone* if it is either upward or downward monotone, which is defined as follows. Q is *upward monotone* := if $\langle M,A,B\rangle \in \mathsf{Q}$ and $B \subseteq B'$, then $\langle M,A,B'\rangle \in \mathsf{Q}$. Q is *downward monotone* := if $\langle M,A,B\rangle \in \mathsf{Q}$ and $B \supseteq B'$, then $\langle M,A,B'\rangle \in \mathsf{Q}$. Barwise and Cooper (1981) formulate and defend the following semantic universal:

MONOTONICITY UNIVERSAL: All simple determiners are monotone.

[2]For a textbook treatment of generalized quantifiers see Peters and Westerståhl (2006).

The property of *quantity* intuitively expresses that the meaning of a determiner only depends on the sizes; i.e., the quantity, of the relevant sets and not on the way those sets are presented or on the particular identity of the objects in those sets. Q is *quantitative* := if $\langle M,A,B\rangle \in \mathsf{Q}$, and $A \cap B$, $A \setminus B$, $B \setminus A$, and $M \setminus (A \cup B)$ have the same cardinality (size) as their primed-counterparts, then $\langle M',A',B'\rangle \in \mathsf{Q}$. Keenan and Stavi (1986) formulate and defend the following semantic universal[3] :

QUANTITY UNIVERSAL: All simple determiners are quantitative.

The property of *conservativity* intuitively expresses that a noun phrase of the form *Det N VP* is genuinely about the *N* and not about the *VP*. That is, to verify a quantifier in a quantifier model only the *A*'s that are *B*'s are relevant, not the *B*'s that are not *A*'s. Q is *conservative* := $\langle M,A,B\rangle \in \mathsf{Q}$ if and only if $\langle M,A,A\cap B\rangle \in \mathsf{Q}$. Barwise and Cooper (1981) formulate and defend the following semantic universal:

CONSERVATIVITY UNIVERSAL: All simple determiners are conservative.

Explaining semantic universals via learnability

The question naturally arises: can a unified explanation be given for these universals? ST&S develop the following *learnability hypothesis*: expressions satisfying semantic universals are easier to learn than those that do not.[4] To anthropomorphize: as languages are developing, they choose to attach lexical items to easy-to-learn meanings, and rely on complex grammatical constructions and compositional interpretation thereof to express hard-to-learn meanings.

The hypothesis immediately raises a challenge: to provide a model of learning on which it's true. ST&S train recurrent neural networks to learn minimal pairs of quantifiers, one satisfying the universal and one that does not.

Figure 2 shows an example learnability result from ST&S: an upward montone quantifier (in blue: `at least 4`) was robustly easier to learn for a neural network than a non-monotone quantifier (in red: `at least 6 or at most 2`). Similar patterns were observed for downward monotone and quantitative quantifiers, while conservative ones were found to be no easier to learn than non-conservative ones (but were argued to arise from a different source than learnability).

These computational results provide strong support for the learnability hypothesis. The approach has also worked well in explaining universals in disparate linguistic domains: color terms (Steinert-Threlkeld & Szymanik, 2019) and responsive predicates (Steinert-Threlkeld, in press).

Previous approaches to the complexity of quantifiers

In the literature on generalized quantifiers one can find several approaches to measuring complexity. Although these measures can capture some of the cognitive difficulty of quantifier

[3]See also Peters and Westerståhl (2006), van Benthem (1984), and ST&S.

[4]Hints of this hypothesis may be found in (van Benthem, 1987; Peters & Westerståhl, 2006; Magri, 2015).

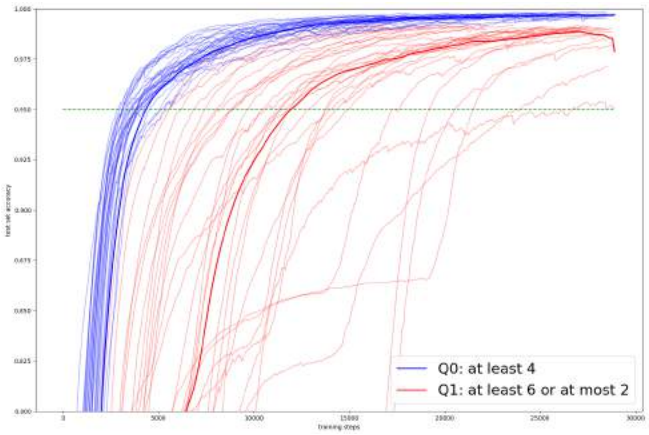


Figure 2: Learning curves on a neural network for the monotone *at least 4* (blue) versus *at least 6 or at most 2* (red). This was Figure 4 in ST&S.

processing,[5] we will see that they are not fine-grained enough to explain the presence of semantic universals.

The earliest approach uses logic, analyzing which quantifiers are definable in progressively more expressive logics. Many natural language quantifiers can be expressed in elementary (i.e. first-order) logic, e.g. `some` or `at least 4`. The seminal result here is that proportional quantifiers cannot be defined in elementary logic: one needs a stronger logical system, like second-order logic, to uniformly express the meaning of, e.g., `most`.[6] This definability criterion cannot, however, distinguish between the complexity of the quantifiers satisfying and not satisfying the universals we study. For example, `all` and `only` can be defined with elementary formulas of exactly the same form (and therefore the same complexity):

$$\texttt{All(A, B)} := \forall x(A(x) \implies B(x))$$
$$\texttt{Only(A, B)} := \forall x(B(x) \implies A(x))$$

Also, both monotone and non-monotone quantifiers can be defined by formulas of the same complexity.

Johan van Benthem (1984) has proposed to study minimal computational devices (automata) corresponding to generalized quantifiers. Under this approach, some quantifiers can be associated with canonical minimal finite automata. One can then use the size of such an automaton (i.e., the number of states) as a measure of quantifier complexity. For example, the automaton for `all` has two states while the automaton for `at least 3` has four states. Other quantifiers—for example, proportional quantifiers—must be associated with more complex computational devices, like push-down automata. This measure of complexity can explain some variance in the cognitive difficulty of quantified sentence verification against pictures (Szymanik & Zajenkowski, 2010). It is, however, not suitable for our purposes. One can easily construct a minimal quantifier pair that cannot be distinguished by this complexity measure. For instance, both `all` and `only` have minimal automata with two states. One can also easily construct a family of quantifiers with the same automaton complexity containing both quantifiers satisfying and not satisfying quantitity (`at least 4`, `first 3`) and monotone and non-monotone quantifiers (`at least 4`, `at least 3 or at most 2`). An extra problem for this approach is that for push-down automata corresponding to proportional quantifiers, there is no accepted complexity measure because they do not have a definition of a minimal automaton. So the measure does not apply to all quantifiers, including ones expressed in natural language.[7]

Another well-studied approach to identify the complexity of generalized quantifiers uses the toolbox of computational complexity theory (Szymanik, 2016). It measures quantifier complexity in terms of the asymptotic growth of the computational resources needed to recognize their meaning. The problem is that computational complexity distinctions are even more crude than the previously described alternatives. Even though computational complexity distinctions have been used to theoretically delimit the borders of natural language expressivity (Ristad, 1993; Kontinen & Szymanik, 2008), these borders include both quantifiers satisfying and not satisfying the semantic universals that we are interested in.

Kolmogorov complexity of quantifiers

To investigate whether the aforementioned semantic universals can be explained by differences in complexity, we need a measure of complexity that is suited for that task. As discussed in the previous subsection, setting up the right framework for this is a non-trivial challenge, as many well-know complexity measures are limited in their ability to distinguish between quantifiers with and without the universal properties under consideration.

Therefore, in this study, we use (approximate) Kolmogorov complexity—a finer-grained measure that has not yet been explored in this domain, and we investigate its potential to explain semantic universals. Because Kolmogorov complexity is more fine-grained than the previously discussed complexity measures, it has greater promise in capturing differences in complexity between the quantifiers that we consider. It makes intuitive sense that humans would be sensitive to Kolmogorov complexity, because it is a mathematical operationalization of the notion of compressibility and various aspects of cognition can plausibly be understood in terms of data compression: storing data compactly in a way that it can be (partially) recovered. Kolomogorov complexity has been shown useful in modelling a cognitive bias towards simplicity in a variety of cognitive domains (see Chater & Vitányi, 2003; Feldman, 2016).

Kolmogorov complexity (K) measures how much an individual sequence of symbols can be compressed. When a sequence contains regularities, these regularities can be exploited to produce a shorter description of that sequence. $K(x)$

[5] See Szymanik (2016) for an overview.

[6] See Peters and Westerståhl (2006) for an overview.

[7] This approach has also inspired learnability models (Gierasimczuk, 2005; Clark, 2010).

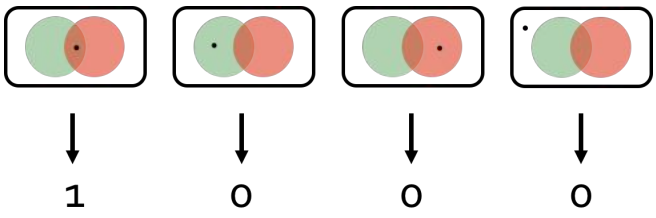


Figure 3: Encoding of some over models of size 1.

of a sequence x is defined as the length of the shortest program p that outputs x (see Li & Vitányi, 2008).[8]

The drawback is that K has been formally proven to be uncomputable. This means that there exists no algorithm that outputs $K(x)$, given x (Li & Vitányi, 2008). For this reason, we use a well-established and tractable approximation to K, that is based on the Lempel-Ziv algorithm for lossless data compression (Lempel & Ziv, 1976). The Lempel-Ziv algorithm parses a sequence x from left to right, and cuts up the sequence into subsequences. At each point it chooses the longest possible subsequence that is identical to an earlier part of the sequence, thereby identifying the number of unique subpatterns in x. The Lempel-Ziv complexity $LZ(x)$ is the number of these unique subpatterns of x. For approximate Kolmogorov complexity $\tilde{K}$, we use $C_{LZ}(x)$, which is defined as $\log_2(\mathsf{len}(x)) \cdot LZ(x)$.[9] Ziv and Lempel (1978) show that $C_{LZ}(x)$ approximates $K(x)$ in the limit; i.e, when $\mathsf{len}(x)$ approaches infinity.Vitányi (2013) shows that, in practice, lossless compression methods give adequate results also for finite sequences. Furthermore, C_{LZ} is considered particularly adequate as a measure for $\tilde{K}$ for shorter strings (Lesne, Blanc, & Pezard, 2009).[10]

To determine the approximate Kolmogorov complexity $\tilde{K}$ of a quantifier we need to represent it as a sequence of symbols. We encode a quantifier as a binary sequence, representing the quantifier as a distribution of truth values over all models (up to a certain size). First, we enumerate all possible models. Then, given such an enumeration, we represent a quantifier by placing a 1 in the sequence for every model that verifies the quantifier and placing a 0 for every model that does not verify the quantifier. See Figure 3 for an example. Given a sequence of models, this gives a unique binary representation for every possible quantifier. Then, for a given sequence of models up to a certain maximum model size, we can determine the complexity of a quantifier Q by computing $\tilde{K}(x_{\mathsf{Q}})$ over the binary representation x_{Q} of Q.

This framework allows us to compare the complexity of different quantifiers and investigate whether semantic universals might be explained by differences in complexity. In doing this, we are not interested in the absolute complexity values of the quantifiers but in the difference in complexity between a quantifier that satisfies a universal and its minimally differing counterpart that does not satisfy that universal. To make any such comparison across quantifiers, we need to fix an enumeration over quantifier models and use that as the base for our quantifier representations.

One way of doing that would be to take a random enumeration over quantifier models. Unfortunately, for our purpose, this is not a suitable method. For a random sequence, the complexity of a quantifier is mainly determined by the uniformity of that quantifier (defined by the ratio of 1's versus 0's in the quantifier representation).[11] When the uniformity of a quantifier is the main determiner for its complexity, differences between the complexity of two quantifiers might not reflect differences due to the presence or absence of a particular universal property.

For our purpose, choosing a structured sequence over models is more suitable than taking a random sequence. The intuition behind this can be understood as follows. If a quantifier that satisfies a universal has lower $\tilde{K}$ complexity than its minimally differing counterpart, then this will be because the universal property causes a regularity in the distribution of truth values across quantifier models. This difference in regularity between quantifiers could disappear when evaluating quantifiers over a random sequence of models, but it might be visible when evaluating those quantifiers over a structured and well-behaved sequence. For this reason, we evaluate our quantifiers over the lexicographic sequence of models, which is standardly used in the literature on generalized quantifiers. For robustness, we look at all 12 uniquely different possible lexicographical orderings, arising from the different ways of ordering the symbols for the four sets $A \cap B$, $A \setminus B$, $B \setminus A$, and $M \setminus (A \cup B)$.[12]

Results

With this framework in place we can now turn to our main question. To test whether approximate Kolmogorov complexity can explain the three proposed semantic universals, we

[8] Formally, K is defined given a particular universal Turing machine (UTM), but, by the Invariance Theorem, K given UTM V or given UTM W will not differ more than some constant c.

[9] In particular, we use the same version of C_{LZ} as used by Dingle, Camargo, and Louis (2018), which uses the average between $LZ(x)$ and $LZ(\mathsf{reverse}(x))$ to obtain an even more fine-grained complexity measure.

[10] There are also other popular lossless compression methods that can be used as approximations to K, such as gzip (based on LZ compression), and bzip2 (a block-sorting compressor). Graphs comparing the LZ and gzip2 complexity of the quantifiers that we considered can be found at https://tinyurl.com/quantifierLZ.

[11] This can be understood from the fact that among all different strings of a given uniformity there are only few strings of low complexity. This is because when a string x of length n has a low complexity, this means that x can be compressed to a shorter string x' of length $n' < n$, and there are only few strings of length n' compared to the amount of strings of length n. Therefore, when taking two quantifier representations with the same uniformity, over a random sequence of models, they are likely to both have complexity values that are close to the maximum complexity for that uniformity (which are thus similar).

[12] In fact, there are in total 24 different lexicographic enumerations over the quantifier models that we use, but only 12 of them are unique, and the other 12 are the reverse of one of those 12 unique sequences. As mentioned earlier, we use a measure that takes the average between the complexity over a sequence and the complexity of the reverse of that sequence. So this leaves 12 lexicographical sequences over which we can compute this measure.

look at minimally differing pairs of quantifiers in which one satisfies the universal and the other does not. To compare our complexity results with the learnability results of ST&S, we test the same pairs of quantifiers. Let $x_{i,\mathsf{Q}}$ be the binary representation of quantifier Q, based on a sequence of all models up to size i. For each quantifier Q, and for each model size i from 1 to 10, we computed $C_{LZ}(x_{i,\mathsf{Q}})$. We repeated this for all 12 lexicographical model sequences. For each pair we plotted the mean complexity against the maximum model size (with confidence intervals), and we compared the differences in complexity between the two quantifiers at each maximum model size and model sequence. The code that we used for generating these data and the data themselves can be found at https://tinyurl.com/quantifierLZ.

Monotonicity

To test the MONOTONICITY UNIVERSAL, we looked at two quantifier pairs, one with a downward- and one with an upward-monotone quantifier. First, we compared the downward-monotone quantifier `at most 3`, meaning $|A \cap B| \leq 3$, with the non-monotone quantifier `at least 6 or at most 2`, meaning $|A \cap B| \geq 6$ or $|A \cap B| \leq 2$. The mean complexity values over all 12 lexicographical model sequences and a 95% confidence interval are plotted in Figure 4. The descriptive statistics show that for all model sizes larger than 2, monotone `at least 4` has a lower complexity than non-monotone `at least 6 or at most 2` (for model size 1 and 2 the differences are 0). This holds for each of the 12 different model sequences. The 12 individual plots for this pair and all the other quantifier pairs can be found at https://tinyurl.com/quantifierLZ.

Second, we compared the upward-monotone quantifier `at least 4`, meaning $|A \cap B| \geq 4$, with the non-monotone quantifier `at least 6 or at most 2`, meaning $|A \cap B| \geq 6$ or $|A \cap B| \leq 2$. The mean complexity values over all 12 model sequences and a 95% confidence interval are plotted in Figure 4. Exactly like for the downward-monotone quantifiers, the descriptive statistics show that that for all model sizes larger than 2, monotone `at most 3` has a lower complexity than non-monotone `at least 6 or at most 2` (for model size 1 and 2 the differences are 0). Again, this holds for each of the 12 different model sequences.

These complexity results show the same patterns as the learnability results of ST&S. This supports the hypothesis that, in addition to learnability, the MONOTONICITY UNIVERSAL might be explained by differences in complexity, with monotone quantifiers being less complex than non-monotone quantifiers.

Quantity

To test the QUANTITY UNIVERSAL, we looked at two quantifier pairs with a quantitative and a non-quantitative quantifier. First, we compared the quantitative quantifier `at least 3`, with the non-quantitative quantifier `first 3`. The mean complexity values over all 12 model sequences and a 95% confidence interval are plotted in Figure 5. For model size 1, 2, and 3, the differences are 0, and for model sizes 4 to 10 the descriptive statistics show that `at least 3` is less complex in 59.5% of the cases, and more complex in 33.3% of the cases.

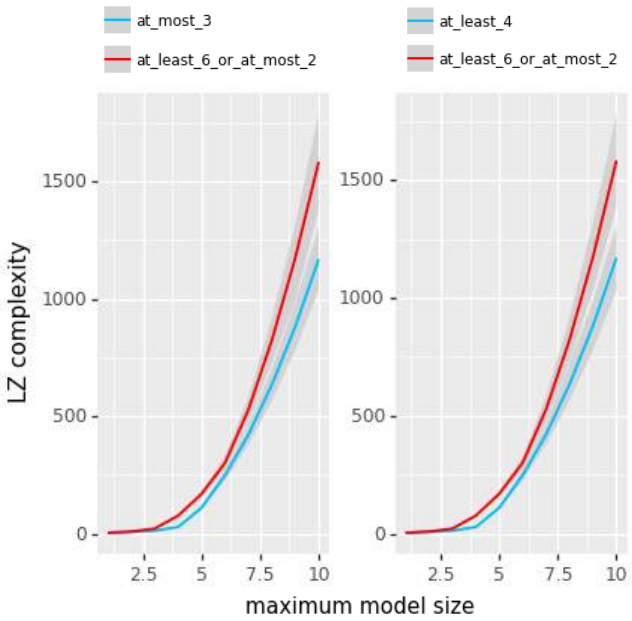


Figure 4: Complexity values for `at most 3` and `at least 6 or at most 2`, and for `at least 4` and `at least 6 or at most 2`. Mean values with 95% confidence interval over all 12 lexicographic model sequences

Second, we compared the quantitative quantifier `at least 3`, with the non-quantitative quantifier `last 3`. The main complexity values over all 12 model sequences and a 95% confidence interval are plotted in Figure 5. Again, for model size 1, 2, and 3, the differences are 0, while for model sizes 4 to 10 the descriptive statistics show that `at least 3` is less complex in 52.4% of the cases and more complex in 42.9% of the cases.

These complexity results do not show a robust pattern. However, they do show a tendency towards the quantitative quantifiers being less complex than the non-quantitative quantifiers. In the learnability results of ST&S, the quantitative quantifiers were significantly easier to learn than the non-quantitative ones. These findings neither confirm nor disconfirm the hypothesis that, in addition to learnability, the QUANTITY UNIVERSAL could be explained by differences in complexity.

Conservativity

To test the CONSERVATIVITY UNIVERSAL, we looked at two quantifier pairs with a conservative and a non-conservative quantifier. First, we compared the conservative quantifier `most`, meaning $|A \cap B| > |A \setminus B|$, with the non-conservative quantifier `M`, meaning $|A| > |B|$. The mean complexity values over all 12 model sequences and a 95% confidence interval are plotted in Figure 6. The descriptive statistics show that that for all model sizes and for all model sequences, conservative `most` has exactly the same complexity as non-conservative `M`.

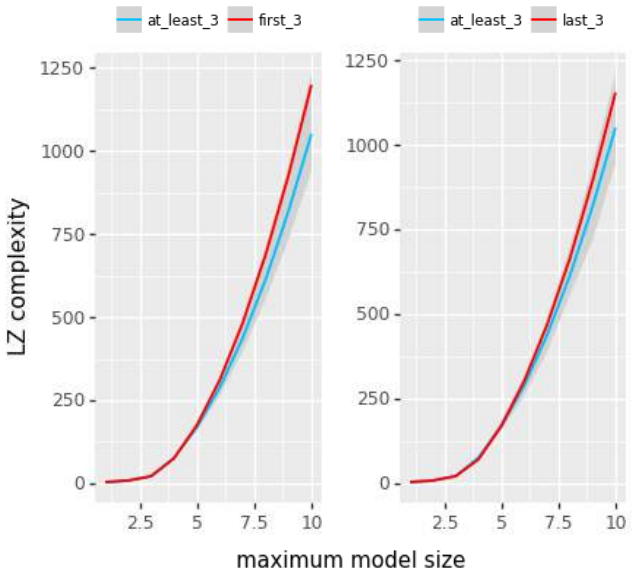


Figure 5: Complexity values for `at least 3` and `first 3`, and for `at least 3` and `last 3`.

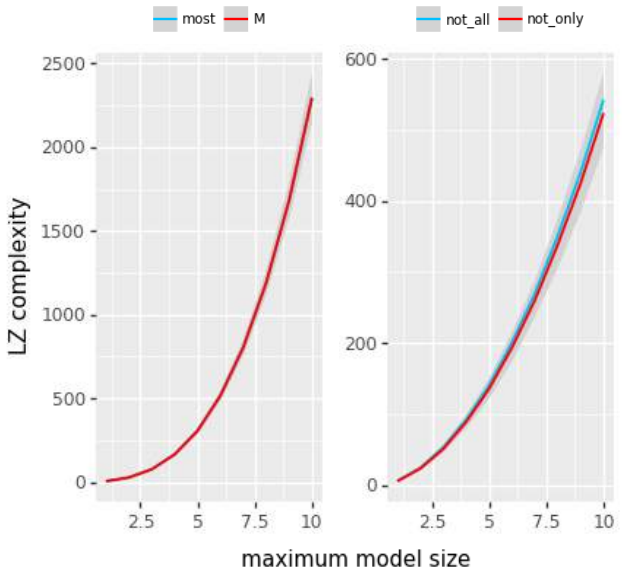


Figure 6: Complexity values for `most` and `M`, and for `not all` and `not only`.

Second, we compared the conservative quantifier `not all`, meaning $A \not\subseteq B$, with the non-conservative quantifier `not only`, meaning $B \not\subseteq A$. Again, the mean complexity values over all 12 model sequences and a 95% confidence interval are plotted in Figure 6. For model size 1 to 10 descriptive statistics show that `not all` is more complex in 55.9% of the cases and less complex in 40.8% of the cases.

These results do not support the hypothesis that the CONSERVATIVITY UNIVERSAL can be explained by differences in complexity. However, these complexity results do show the same patterns as the learnability results of ST&S, as in their results the conservative quantifiers were of similar learnability as the non-conservative ones. This, however, does not constitute a counterexample to the explanation of the universals via learnability. As explained by ST&S one should not expect the difference between conservative and non-conservative quantifiers under their framework. This universal should rather be explained in terms of the syntax-semantics interface.[13]

Discussion

Let us take stock. We have applied tools from algorithmic information theory—in particular, approximate Kolmogorov complexity—to measure the complexity of quantifiers expressed in natural language. We did this in order to see whether the complexity of a quantifier can explain the presence of semantic universals for quantifiers, and whether these complexity results show the same patterns as existing learnability results. We found that monotone quantifiers are robustly less complex than non-monotone quantifiers, and that conservative and non-conservative quantifiers have equal or similar complexity. For quantitative quantifiers we found a slight tendency towards being less complex, but this pattern was not robust. The results for monotonicity and conservativity agree with an existing explanation in terms of learnability due to ST&S, while the results on quantitativity hint in the same direction, but not robustly so.

The results of the exploratory study that we undertook are not decisive. Nevertheless, the results show substantial similarity between the complexity and learnability of quantifiers in the explanation of semantic universals. Our results for monotonicity show that approximate Kolmogorov complexity can indeed capture differences in complexity between quantifiers that could not be captured with the complexity measures from the previous approaches that we discussed. That neither complexity nor learnability distinguishes conservative from non-conservative quantifiers provides further evidence that conservativity has a different source than the other two universals, as suggested by ST&S.

Much work remains to be done. To corroborate our results, one would like to scale up beyond maximum model size of $n = 10$; how to make this computationally efficient is not a simple task. One would also like to expand the experiments beyond the minimal pair methodology employed here. In order to compare with existing results, it would be good to measure the complexity of many quantifiers and see which semantic properties best explain the complexities. The methods here could also be applied to semantic universals in other domains, to test the connection between complexity and learnability in a more general setting. Finally, one can also look at other measures of complexity: for instance, minimal derivation length in a Lanuage of Thought for generating expressions for quantifier meanings (Piantadosi, Tenenbaum, & Goodman, 2012; Goodman, Tenenbaum, & Gerstenberg, 2015).

[13] See Romoli (2015). Hunter and Lidz (2013) observe a difference in children learning conservative vs. non-conservative quantifiers. This result, if replicated, could be due to a bias acquired by the children in earlier exposure to only conservative determiners.

Acknowledgements

We thank the members of the CoCoLab and the CoLaLa at Stanford and Berkeley, respectively, for helpful discussion and Ronald de Haan for technical support. IvP was supported by Gravitation Grant 024.001.006 of the Language in Interaction Consortium from the Netherlands Organization for Scientific Research (NWO). Shane S-T and JS have received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007–2013)/ERC Grant Agreement n. STG 716230 CoSaQ.

References

Barwise, J., & Cooper, R. (1981). Generalized Quantifiers and Natural Language. *Linguistics and Philosophy*, *4*(2), 159–219.

Chater, N., & Vitányi, P. (2003). Simplicity: A unifying principle in cognitive science? *Trends in cognitive sciences*, *7*(1), 19–22.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. MIT Press.

Clark, R. (2010). On the learnability of quantifiers. In J. J. van Benthem & A. ter Meulen (Eds.), *Handbook of logic and language* (pp. 909–922). Elsevier.

Dingle, K., Camargo, C. Q., & Louis, A. A. (2018). Input–output maps are strongly biased towards simple outputs. *Nature Communications*, *9*(1), 761.

Feldman, J. (2016). The simplicity principle in perception and cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *7*(5), 330–340. doi: 10.1002/wcs.1406

Gibson, E., Futrell, R., Piantadosi, S. T., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How Efficiency Shapes Human Language. *Trends in Cognitive Sciences*.

Gierasimczuk, N. (2005). The problem of learning the semantics of quantifiers. In *International tbilisi symposium on logic, language, and computation* (Vol. 4363, pp. 117–126).

Goodman, N. D., Tenenbaum, J. B., & Gerstenberg, T. (2015). Concepts in a probabilistic language of thought. In Morgolis & Lawrence (Eds.), *The Conceptual Mind: New Directions in the Study of Concepts.* MIT Press.

Hsu, A. S., Chater, N., & Vitányi, P. (2013). Language learning from positive evidence, reconsidered: A simplicity-based approach. *Topics in Cognitive Science*, *5*(1), 35–55.

Hunter, T., & Lidz, J. (2013). Conservativity and learnability of determiners. *Journal of Semantics*, *30*(3), 315–334.

Hyman, L. M. (2008). Universals in phonology. *The Linguistic Review*, *25*(1-2), 83–137.

Keenan, E. L., & Stavi, J. (1986). A Semantic Characterization of Natural Language Determiners. *Linguistics and Philosophy*, *9*(3), 253–326.

Kontinen, J., & Szymanik, J. (2008). A remark on collective quantification. *Journal of Logic, Language and Information*, *17*(2), 131–140.

Lempel, A., & Ziv, J. (1976). On the complexity of finite sequences. *IEEE Transactions on Information Theory*, *22*(1), 75–81.

Lesne, A., Blanc, J.-L., & Pezard, L. (2009). Entropy estimation of very short symbolic sequences. *Physical Review E*, *79*(4), 046208:1–10.

Li, M., & Vitányi, P. (2008). *An Introduction to Kolmogorov Complexity and Its Applications*. Springer.

Magri, G. (2015). Universals on natural language determiners from a pac-learnability perspective. In *Proceedings of cogsci 2015.*

Newmeyer, F. J. (2008). Universals in syntax. *The Linguistic Review*, *25*(1-2), 35–82.

Peters, S., & Westerståhl, D. (2006). *Quantifiers in Language and Logic*. Oxford: Clarendon Press.

Piantadosi, S. T., Tenenbaum, J. B., & Goodman, N. D. (2012). *Modeling the acquisition of quantifier semantics: A case study in function word learnability.*

Ristad, E. S. (1993). *The language complexity game*. MIT Press.

Romoli, J. (2015). A Structural Account of Conservativity. *Semantics-Syntax Interface*, *2*(1), 28–57.

Steinert-Threlkeld, S. (in press). An explanation of the veridical uniformity universal. *Journal of Semantics*.

Steinert-Threlkeld, S., & Szymanik, J. (2019). *Ease of Learning Explains Semantic Universals.*

Steinert-Threlkeld, S., & Szymanik, J. (in press). Learnability and Semantic Universals. *Semantics & Pragmatics*. (`https://semanticsarchive.net/Archive/mQ2Y2Y2Z/LearnabilitySemanticUniversals.pdf`)

Szabolcsi, A. (2010). *Quantification*. Cambridge: Cambridge University Press.

Szymanik, J. (2016). *Quantifiers and Cognition: Logical and Computational Perspectives* (Vol. 96). Springer.

Szymanik, J., & Zajenkowski, M. (2010). Comprehension of simple quantifiers. Empirical evaluation of a computational model. *Cognitive Science: A Multidisciplinary Journal*, *34*(3), 521-532.

Tiede, H.-J. (1999). Identifiability in the Limit of Context-Free Generalized Quantifiers. *Journal of Language and Computation*, *1*(1), 93–102.

van Benthem, J. (1984). Questions About Quantifiers. *The Journal of Symbolic Logic*, *49*(2), 443–466.

van Benthem, J. (1987). Toward a Computational Semantics. In P. Gardenfors (Ed.), *Generalized quantifiers: Linguistic and logical approaches* (pp. 31–71). Kluwer.

Vitányi, P. M. (2013). Similarity and denoising. *Philosophical Transactions of the Royal Society A*, *371*(1984).

Ziv, J., & Lempel, A. (1978). Compression of individual sequences via variable-rate coding. *IEEE Transactions on Information Theory*, *24*(5), 530–536.

Preschoolers' Evaluations of Ignorant Agents are Situation-Specific

Alyssa R. Varhol (arv57@cornell.edu)
Department of Human Development, Beebe Hall
Ithaca, NY 14850 USA

Tamar Kushnir (tk397@cornell.edu)
Department of Human Development, Beebe Hall
Ithaca, NY 14850 USA

Melissa A. Koenig (mkoenig@umn.edu)
Institute of Child Development, 51 East River Road
Minneapolis, MN 55455 USA

Abstract

Preschool children's preference for knowledgeable agents over ignorant and inaccurate agents (Sabbagh & Baldwin, 2001; Koenig & Harris, 2005; Rakoczy et al., 2015), is generally interpreted as epistemic vigilance. However, Kushnir and Koenig (2017) recently found that without a contrasting accurate agent, preschoolers will learn new information from an agent who professed ignorance, but not from one who was inaccurate. Employing a two-speaker design contrasting an agent who professed ignorance about familiar object labels with a speaker whose knowledge state was not revealed, we found that preschoolers (N = 41; 3.50-4.89 years, M = 4.08 years) avoided requesting and endorsing novel information from the ignorant agent in the same domain as her previous ignorance (i.e., labels). In different domains, however, (i.e. novel function learning, resource sharing, etc.) they were at chance in choosing the ignorant agent. This suggests that preschoolers' view of ignorance is situational, rather than uniformly negative.

Keywords: learning; testimony; social cognition; credibility; cognitive development; epistemic trust; accuracy; epistemic vigilance

Background

Numerous studies show an overwhelming preference in early childhood for a competent, confident, accurate, or knowledgeable agent over an agent who was inaccurate, ignorant, or uncertain (Birch, Vauthier, & Bloom, 2008; Brosseau-Liard & Birch, 2010, Brosseau-Liard, Cassels, & Birch, 2014; Fusaro, Corriveau, & Harris, 2011; Harris & Corriveau, 2011; Koenig & Harris, 2005; Koenig & Woodward, 2010; Pasquini et al., 2007; Rakoczy et al., 2015; Sabbagh & Shafman, 2009; Scofield et al., 2013; Tenney et al., 2011; Tummeltshammer et al., 2014; *For review, see* Harris et al., 2018). There may be many reasons for this preference--including assessments based on vigilance or trust—but in any case, there seems to be a general negative assessment of all uninformative agents by preschool age.

Recent findings suggest that children do not treat all uninformative agents as equally untrustworthy. Kushnir & Koenig (2017) measured preschoolers' evaluations of either an agent who professed ignorance about familiar object labels or one who was inaccurate. In one condition, 3- and 4-year-old children viewed an agent who professed ignorance about the names of familiar objects. In another, children viewed an agent who was inaccurate in naming the same objects. Kushnir & Koenig found that children were willing to learn new things from the previously ignorant agent, but not from the inaccurate one. This study suggests that children's evaluations of uninformative agents are not uniformly negative or vigilant. Specifically, that they don't see ignorance about some things as a sign to mistrust or avoid learning other things.

We can infer from Kushnir and Koenig (2017) that children respond more negatively to inaccurate agents than ignorant agents, but it remains unclear what these results imply about their evaluations of professed ignorance. It could be that by the presence of a preferred accurate agent overrode information from an ignorant agent in previous studies (e.g., Koenig & Harris, 2005), and that this single-speaker design revealed children's true ignorance evaluations. However, it could be that children were simply agnostic toward the previously ignorant agents, and were willing to learn from them when no alternatives were available.

What is the nature of children's stance on professed ignorance? We suggest that there are at least three possible answers. One is that children view ignorance as situation-specific. Broadly, this means children could discount past ignorance when learning new things (as in the above example) or they might treat an agent's claims of ignorance as specific to one domain of expertise and not another (e.g. Lutz & Keil, 2002; Kushnir, Vredenburgh & Schneider, 2013). The second possibility is that children look favorably on ignorant agents when they make new claims because they will admit what they don't know (i.e. they are "well calibrated" or even "virtuous" e.g. Kominsky, Langthorne, & Keil, 2016; Tenney et al., 2011). This suggests that children could show a preference for those who admit ignorance regardless of domain or situation because. A third possibility is that children only prefer previously ignorant agents when no other agents are available to provide information. This suggests that if any other reasonable (i.e. not inaccurate) source of information was present, children would avoid learning from an ignorant agent. Of course, these need not necessarily be mutually exclusive and could represent contributing factors in a nuanced assessment. We investigate the roles of these three possible interpretations in the current study.

We used a modified version the two-speaker design from Koenig & Harris (2005) which contrasts an ignorant with an accurate speaker to examine these three possibilities. The modification was to contrast an agent who admitted to not knowing the names of familiar objects with a neutral agent whose knowledge state has not been disclosed. To explore the specificity of children's ignorance evaluations, we measured children's willingness to learn from the ignorant agent about novel objects in two domains: labels and functions. To explore the depth of children's evaluations, we measured children's choices of the ignorant agent for requesting information and for endorsing new claims within both domains.

If children's evaluations of ignorance are situation-specific, we expect children to differ in their willingness to learn new information about object labels versus functions from a source who was ignorant about labels. If they instead view ignorance as a virtue or signal of calibration, we expect children to show willingness to learn from the ignorant agent in all cases. Finally, if children show overall vigilance, we expect them to avoid learning new information from the ignorant agent in all cases.

In addition to the learning tasks, we included three different measures of children's ignorant speaker evaluations in non-learning situations. To capture whether they had a preference or general positive regard for the ignorant agent, we measured how often children shared more stickers with her than with the neutral agent across three resource-sharing trials (*see* Chernyak & Sobel, 2015; Kanngiesser & Warneken, 2013; Moore, 2009). Toward testing for general dislike or mistrust, we controlled for agent knowledge state by measuring children's endorsements of claims about the location of a hidden object that both agents could see. Further, to determine whether evaluations permeated children's explicit understanding of agent knowledge, we asked children which of the two agents knows more. Together, these measures can provide evidence about the extent of overall positive or negative evaluations of the ignorant agent.

Method

Participants

We tested 41 preschool age children (16 girls) between 3.50 and 4.89 years old (M = 4.08 yrs., SD = 0.42 yrs.) from a large midwestern city. In addition, one child was excluded for experimenter error, and one child was excluded for ending the study early. Participants were predominately from white, upper-middle class families.

History Phase

Children were shown an image of the two agents and were told they were going to watch some videos of these two friends and then play a game. They then watched alternating videos of the ignorant (*I*) agent (3) and neutral (*N*) agent (3). For each video, the agent sat at a table with a confederate, who initiated a brief exchange with the agent. In order to control for features outside of demonstrated knowledge state, both agents responded to the confederate in a conventional way (e.g., returning a greeting or responding to a question) and were on screen for approximately equal periods of time. Agent who spoke first (*I* vs. *N*; speaker order was constant across all trials within subjects) and actor who was the ignorant agent (*blue shirt* vs. *red shirt*), were counterbalanced between subjects.

Ignorant Agent Videos The confederate handed a familiar object (ball, cup, shoe) to the ignorant agent, asking "Look what I have! Can you tell me what that is called?" (see Kushnir & Koenig, 2017). Each time, agent *I* held the item with both hands, shook her head, and responded "I don't know what that is called". All professions of ignorance concerned labels for familiar objects.

Neutral Agent Videos The confederate and the neutral agent both sat at the table using their cell phones with none of the familiar objects from the Ignorant condition present. The confederate briefly looked up and initiated a common, familiar interaction with agent *N*. ("Hi," "Good morning," and "How are you?") before looking back at her phone. The neutral agent then looked up briefly and responded with an appropriate but non-informative answer ("Hi," "Good morning," and "Fine").

Test Phase

The test phase consisted of 9 trials. The first 6 were two blocks of 3 trials: one novel label and one novel function trial (counterbalanced) followed by a resource sharing trial. The last three trials were (in this order): locations trial, final resource sharing trial, and knowledge attribution. Of the four novel objects, two were always used for label trials and two were always used for function trials. Each trial type is described below:

Novel Label Requests For each novel label trial, the experimenter (*E*) first displayed an image of the novel object on the screen and prompted the child by saying "Look at that thing! I've never seen one of those before! I wonder what it's called. I bet one of our friends can tell us!" *E* then showed the paused opening scene of the novel object video, in which the confederate is standing between the two agents and holding the object, and asked the child, "Who do you want to ask what that is called?" If the child did not reply, *E* prompted once more with "Which friend do you want to ask?" The child's first choice was recorded, and *E* responded with "Ok. Let's see!" regardless of the response.

Novel Label Endorsements *E* then played a video in which the confederate said "Look what I have!" and turned to each agent (order counterbalanced between subjects) and asked "Can you tell me what this is called?" Each agent gave a different label (e.g., danu or koba, counterbalanced). After each video, the child was shown a still image of the two agents with the item between them. *E* pointed to each agent in the order in which they spoke, saying, "So she said it's a danu, and she said it's a koba. What do you think it's called?"

Children's first response was recorded. If the child said "I don't know," *E* followed up with "do you think they could both be right or both be wrong?" Otherwise, no feedback was given.

Novel Function Requests The procedure for novel function requests was identical to that of the novel label requests, except that *E* said "I wonder what it's used for!" and "Who do you want to ask what it's used for?" instead of "I wonder what it's called…Who do you want to ask what it's called?" The objects used for function trials each had features that made both functional claims feasible.

Novel Function Endorsements Endorsement measures for novel functions were also the same as the label endorsement trials, except the confederate asked the agents what the object was for, and they named and demonstrated different functions (e.g., in *Figure 1,* for looking or for stacking).

Resource Sharing At each sharing trial, the experimenter placed two cups, each with a picture of one of the two agents taped to it, in front of the child. The child was then given five identical stickers and was told that for each sticker, they could share with whichever friend they want by putting the sticker in that agent's cup.

Location Endorsement Children watched a video in which the agents had two boxes (equal in size, varying in color) between them. In the video, the confederate showed a small toy, held up a barrier blocking the boxes from the child's view, and then made a motion of placing the toy somewhere behind the barrier while both agents followed the motion with their gaze to indicate they were watching. The confederate then asked where the toy was, and each agent made a different claim about which box it was in (counterbalanced). Children were then asked to endorse one of the locations.

Knowledge Attribution After all the test videos, children were shown the still image of the two agents one more time and were asked, "Who do you think knows more?" First response was recorded, and children were asked "why do you think she knows more?" as a follow-up.

Figure 1: Examples of novel function *(left*) and novel label (*right)* stimuli.

Coding

We coded four categories of responses to our request and endorsement questions. The majority of responses (77.32%) were selections of a single agent (ignorant agent or the neutral agent). The second most frequent response (15.12%) was expressing uncertainty about the choice (e.g. "I don't know"). A small percentage of children (2.44%) picked both agents. On endorse trials, a small percentage of children (5.12%) made up their own label or function (see below).

Requests For each request question (2 label, 2 function), children were given 1 point for each time they asked the ignorant agent (singly or by responding "both") and 0 points for each time they did not (by picking the neutral agent or saying "I don't know").

Endorsements For each endorsement (2 label, 2 function, 1 location) Similar to coding for requests, we gave children 1 point for each time they endorsed the ignorant agent and 0 points for each time they did not. In cases where children used an alternative name or function, we coded their response as a 0. In cases where children responded with uncertainty, we followed up with "Do you think they could both be right or both be wrong?" and assigned 1 point if they selected "both right" and 0 points if they selected "both wrong". (See *Table 2* for responses before follow up question).

Resource Sharing For each of the three sticker sharing trials, we coded two measures. Children were given a score of 1 for each time they gave more stickers to the ignorant agent and a score of 0 each time they gave fewer stickers to the ignorant agent, and we added these scores across the three trials for a possible score of 0-3. We also recorded the number of stickers (0-5) shared with the ignorant agent on each trial and calculated each child's average number of stickers shared with Agent I across all three trials.

Knowledge Attribution Children were given a score of 1 if they indicated that the ignorant agent knew more and 0 if they did not.

Results

McNemar's tests indicated that there were no significant differences in the proportion of Ignorant agent choices in Trial 1 and Trial 2 for label requests (p = 1.00), label endorsements (p = 0.344), function requests (p = 0.146), or function endorsements (p = 0.238). Therefore, we summed across both trials of each of the four questions types, creating four variables with possible scores of 0-2. Pearson's correlations indicated that age in months was not significantly related to ignorant agent choice in any question type or domain (see *Table 1)*, so we did not include age as a covariate in further analyses.

Table 1: Mean choices of ignorant agent and age correlations by task.

Task	*M*	SD	*r* with Age
Novel Labels			
Requests	0.63	0.799	0.23
Endorsements	0.634	0.799	-0.14
Novel Functions			
Requests	0.927	0.848	-0.14
Endorsements	1.024	0.758	0.24

Note. For all tasks, $N = 41$. Range (0-2). For all correlations, $p > .05$

Main Effect of Domain and Question Type

A 2 x 2 repeated-measure ANOVA of domain (label vs. function) by question type (request vs. endorse) revealed a main effect of the domain of the novel information on children's choices of the ignorant agent. Specifically, children were significantly less willing to choose to learn from the ignorant agent in the label domain than in the function domain; F(1) = 7.895, p <.01, 95% CI[-.587, -.096] (see *Table 1* for *M* and *SD*). There was no main effect of question type (F = 0.196) and no domain by question type interaction (F = 0.170).

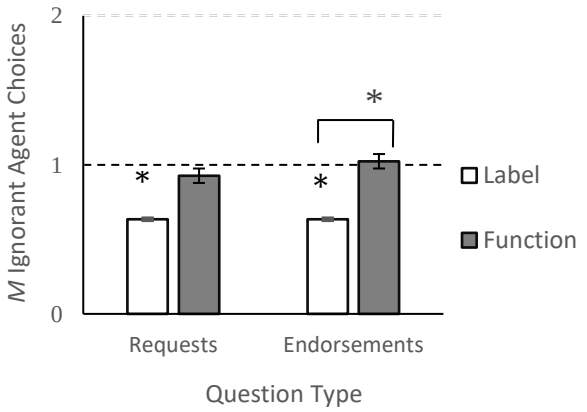


Figure 2: Mean selections of the ignorant agent across domains & question types. Dashed line refers to chance responding.

Domain & Question Type Differences

To further explain this domain effect, we tested choice of ignorant agent against chance for each task and the difference in ignorant agent choices between domains for each question type (e.g. label requests vs. function requests). See *Figure 2* for a visualization of these results.

Children's selections of the ignorant agent were significantly below chance for both requests and endorsement questions in the label domain; $t(40) = -2.933$, $p < .01$, 95% CI [-0.62, -0.11]. In the function domain, children were at chance for choices of the ignorant agent for both function requests (t(40) = -.552, p = .58, 95% CI [-0.34, 0.19]) and function endorsements; t(40) = .206, p = .84, 95% CI [-0.21, 0.26].

Follow up paired-samples t-tests revealed that the domain effect was stronger for endorsements than requests: there was no significant difference between domains on children's requests alone ($t(40) = -1.524$; $p = .153$), but children were significantly less likely to endorse the ignorant agent for novel labels than for novel functions ($t(40) = -2.72$, p = .01, 95% CI[-0.68, -0.10].

To further examine which alternative responses children made when they did not endorse the ignorant agent, we looked descriptively at the counts and percentages of all response categories for each task (see *Figure 3)*. While the percentage of Ignorant Agent choices were noticeably higher in the novel function domain than in the novel label domain, the percentage of choices of the Neutral agent remained similar across all tasks except the novel label requests. The distribution of responses shows that when children were not endorsing the Ignorant agent's label, they were expressing uncertainty or making up their own alternative label as often as they were endorsing the neutral agent's label.

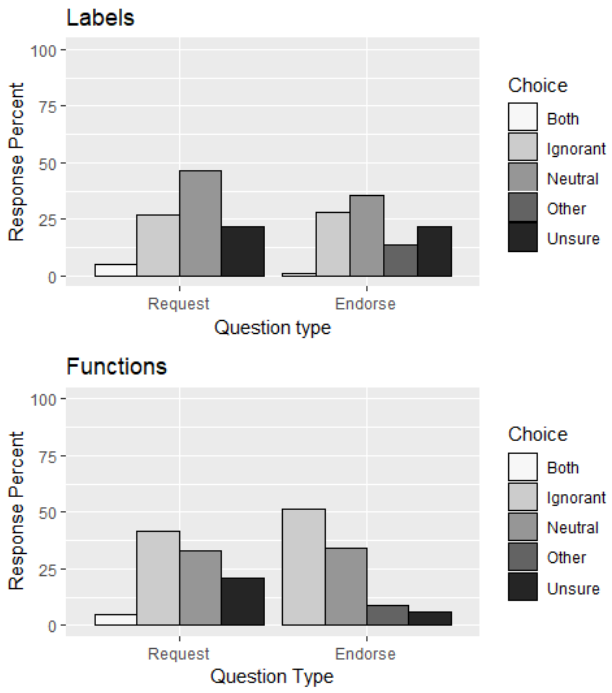


Figure 3: Percentage of raw response types by domain & question type.

Location, Knowledge Attribution, & Sharing

Binomial tests revealed that children were at chance for endorsement of the ignorant agent's hidden object location claim (54% ignorant agent endorsement, p = 0.76) and for attribution of knowledge to the ignorant agent (56%, p = 0.53).

Table 2: Pearson correlations for Ignorant Agent choices in tasks without novel object

	Attribute Knowledge	Share Freq.	Share Avg.	Label Total	Function Total
Endorse Location	.262	.378*	.162	.127	-.236
Attribute Knowledge	1	.564**	.364*	.117	.165
Share Freq.	-	1	.755**	.145	.211
Share Avg.	-	-	1	-.006	.106

Note. "Share Freq." is frequency of sharing more with agent *I* (0-3). "Share Avg" is the avg. amount shared with agent *I* (0-5). "Label total" and "Function Total" are the sum of *agent I* choices across all novel label trials and all novel function trials, respectively*; *p < .05, **p < .01*

Children were also at chance (M = 1.49, SD = .952) for the number of times (0- 3) they shared more stickers with the ignorant agent; t (40) = -.082, p = .935, 95% CI[-0.31, 0.39], and for the average number of stickers (0-5) they shared with the ignorant agent across trials (M = 2.54, SD = 0.774); t (40) = 0.303, p = .764, 95% CI[-0.21, 0.28].

We conducted 2-tailed Pearson correlations to explore the relation of these measures to all the other outcome variables. Agent choices on location endorsement, knowledge attribution, and resource sharing were not related to agent choices on any of the novel label or function questions.

However, ignorant agent choices were strongly correlated between several of these three non-novel object tasks (*Table 2).* Notably, the number of trials in which children shared more stickers with the ignorant agent than with the neutral agent and the average number of stickers they shared with the ignorant agent were positively related to their attribution of more knowledge to the ignorant agent.

Discussion

When preschool children monitor agents' informativeness as evidence about their reliability, they often show an overwhelming social and learning preference for an agent who demonstrates knowledge, certainty, and accuracy over one who is lacking in any of these criteria. By contrasting an ignorant agent with a neutral agent, we tested three possible stances from which children could be considering professed ignorance. We found that children's responses to a previously ignorant agent are more nuanced than a uniform negative or positive judgment.

If children view professed ignorance as specific to the situation or domain in which they have seen evidence of her ignorance—in this case, object labels, we would expect them to respond to her further claims about object labels differently than her claims in another domain. In support of this explanation, we found that children avoided both requesting and endorsing novel labels from the ignorant agent but did not demonstrate this vigilance against her when learning novel object functions. This result suggests that there is a situational constraint of preschooler's pessimism about ignorant agents.

If children look favorably on agents who profess ignorance, perhaps seeing it as evidence of virtue, we would expect them to show a preference for the ignorant agent in their overall learning, perhaps in their resource sharing, and possibly even in their explicit judgments of agent knowledge. Our novel label and novel function data suggest that they avoided learning labels from, or were agnostic toward learning functions from the ignorant agent rather than preferring her over the neutral agent. On sticker sharing trials, which are often used to measure judgment of virtue or general liking of an agent (Chernyak & Sobel, 2015; Kanngiesser & Warneken, 2013; Moore, 2009), our results show no relation between children's learning from and willingness to share with an agent. Therefore, we did not find that children had a general positive regard toward the ignorant agent based on the sharing data, and having positive regard for the ignorant agent did not predict learning from her. However, children who explicitly stated that the ignorant agent knows more also shared more with her. Together, these results suggest that preschoolers do not think of professed ignorance as virtuous, but they may think of knowledge as a virtue.

If children avoid learning from ignorant agents unless no reasonable alternative is available, we would expect preschoolers not to request or endorse new information from the ignorant agent on any novel label or function trials. If children's avoidance of the ignorant agent expanded beyond epistemic vigilance and into mistrust, we would also expect children to share fewer resources with the ignorant agent and reject her claim about the location of a hidden object. Because children were at chance in responding to the ignorant agent across all trials outside of the label domain, we did not find evidence that children are generally pessimistic toward people profess ignorance.

Overall, we propose an explanation that combines elements of two of our three possibilities. Children's stance on professed ignorance is situation-specific—they show epistemic vigilance against new information from an agent only in situations similar to those in which she was ignorant before (e.g., the label domain). However, the extent of their vigilance is influenced by whether there is another reasonable option from whom to learn. When children saw only an familiar-label-ignorant agent in Kushnir & Koenig (2017), they were above chance in endorsing her later novel label and function claims, but when we presented a neutral agent as a contrast, children's domain-specific vigilance emerged, and their willingness to learn from the ignorant in a new domain was reduced to chance. This situational specificity is not apparent in studies contrasting an ignorant and accurate agent (e.g., Koenig & Harris, 2005), which suggests that preschool children are agnostic in their evaluations of ignorant agents outside of the specific situation in which they professed ignorance, treating them similarly to an agent whose knowledge state is unknown. Future studies should include professed ignorance in other, non-linguistic domains in order

to determine whether these situation-specific evaluations are actually specific to ignorance about labels.

The situational nuances in preschoolers' evaluations of agents who profess ignorance aligns with the extant literature on the development of children's understanding of knowledge and expertise. Our findings highlight preschoolers' stance on professed ignorance as part of a greater developmental trajectory for epistemic trust and social learning (as in Kushnir & Koenig, 2017). By 4 years old, children have begun to distinguish ignorant agents from both accurate and inaccurate agents, distinguish agents by their demonstrated domains of expertise (e.g., labels or causal functions), and use these distinctions to inform learning from those agents (Kushnir et al., 2013, see also Brosseau-Liard & Birch, 2011 and Sobel & Corriveau, 2010). This corresponds with our finding that children also evaluate an agent's ignorance—which could be considered the opposite of expertise—based on the domain in which it is demonstrated. However, in alignment with other studies showing that children do not successfully use an agent's calibration of certainty as a sign of epistemic virtue until the end of middle childhood (Tenney et al., 2011; Kominsky et al., 2016; Brosseau-Liard et al., 2014), we found that preschoolers did not show a significant preference of or deference to the ignorant agent on any trials.

Because we only considered one, specific kind of ignorance—familiar object labels—it would be useful to test children's responses to an ignorant versus neutral agent when the ignorance is professed in different domains, such as familiar object functions and causal knowledge (e.g., Bridgers et al., 2016) or with information that is unfamiliar to the child. Further, we are limited in our knowledge of how children evaluate the neutral agent and what exactly makes an agent "neutral" as a source of information, so future studies should explore different presentations of an agents whose knowledge states are not revealed.

In order to draw more detailed, concrete conclusions about children's understanding of knowledge and the development of their epistemic trust, future work should continue to unpack the different ways children respond after evidence of ignorance. We focused on children's willingness to choose the ignorant agent in different situations, but the variety of responses from children who did not choose her suggest that professing ignorance may be influencing children's behavior outside of signaling someone's reliability as a source of information. Because our study showed the ignorant agent later assigning names to unfamiliar objects, the combination of these factors could have given children license to find an answer on their own. In that case, the number of responses where children made up their own answer rather than endorsing either agent could be related to children's increased exploration in the absence of pedagogical cues (Bonawitz et al., 2011). Future studies should consider individual differences in children's responses and in other sensitive developmental areas for preschool-aged children, such as social cognition (e.g., Sabbagh & Baldwin, 2003).

Acknowledgments

The authors would like to thank Huashoua Thao and Lizeth Rodriguez for their contributions to data collection and Sara Wilkerson for her help in overseeing data collection.

References

Birch, S. A., Vauthier, S. A., & Bloom, P. (2008). Three-and four-year-olds spontaneously use others' past performance to guide their learning. *Cognition*, *107*(3), 1018-1034.

Brosseau-Liard, P. E., & Birch, S. A. (2010). 'I bet you know more and are nicer too!': what children infer from others' accuracy. *Developmental Science*, *13*(5), 772-778.

Brosseau-Liard P, Birch S. (2011). Epistemic states and traits: Preschoolers appreciate the differential informativeness of situation-specific and person-specific cues to knowledge. *Child Development*, 82(6):1788–1796. [PubMed: 22004452]

Brosseau-Liard, P., Cassels, T., & Birch, S. (2014). You seem certain but you were wrong before: Developmental change in preschoolers' relative trust in accurate versus confident speakers. *PloS one*, *9*(9), e108308.

Bonawitz, E., Shafto, P., Gweon, H., Goodman, N. D., Spelke, E., & Schulz, L. (2011). The double-edged sword of pedagogy: Instruction limits spontaneous exploration and discovery. *Cognition*, *120*(3), 322-330.

Bridgers, S., Buchsbaum, D., Seiver, E., Griffiths, T. L., & Gopnik, A. (2016). Children's causal inferences from conflicting testimony and observations. *Developmental Psychology*, *52*(1), 9.

Chernyak, N., & Sobel, D. M. (2016). Equal but not always fair: Value-laden sharing in preschool-aged children. *Social Development*, *25*(2), 340-351.

Fusaro, M., Corriveau, K. H., & Harris, P. L. (2011). The good, the strong, and the accurate: Preschoolers' evaluations of informant attributes. *Journal of Experimental Child Psychology*, *110*(4), 561-574

Harris, P. L., & Corriveau, K. H. (2011). Young children's selective trust in informants. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *366*(1567), 1179-1187.

Harris, P. L., Koenig, M. A., Corriveau, K. H., & Jaswal, V. K. (2018). Cognitive foundations of learning from testimony. *Annual Review of Psychology*, *69*.

Kanngiesser, P., & Warneken, F. (2012). Young children consider merit when sharing resources with others. *PloS one*, *7*(8), e43979.

Koenig, M. A., & Harris, P. L. (2005). Preschoolers mistrust ignorant and inaccurate speakers. *Child development*, *76*(6), 1261-1277.

Koenig, M. A., & Woodward, A. L. (2010). Sensitivity of 24-month-olds to the prior inaccuracy of the source: Possible mechanisms. *Developmental psychology*, *46*(4), 815.

Kominsky, J. F., Langthorne, P., & Keil, F. C. (2016). The better part of not knowing: Virtuous ignorance. *Developmental psychology*, *52*(1), 31.

Kushnir, T & Koenig, M. A. (2017). What I don't know won't hurt you: The relation between professed ignorance and later knowledge claims. *Developmental Psychology, 53*(5), 826-835.

Kushnir, T., Vredenburgh, C., & Schneider, L. A. (2013). "Who can help me fix this toy?" The distinction between causal knowledge and word knowledge guides preschoolers' selective requests for information. *Developmental psychology*, *49*(3), 446.

Moore, C. (2009). Fairness in children's resource allocation depends on the recipient. *Psychological Science*, *20*(8), 944-948.

Pasquini, E. S., Corriveau, K. H., Koenig, M., & Harris, P. L. (2007). Preschoolers monitor the relative accuracy of informants. *Developmental psychology*, *43*(5), 1216.

Rakoczy, H., Ehrling, C., Harris, P. L., & Schultze, T. (2015). Young children heed advice selectively. *Journal of Experimental Child Psychology, 138*, 71-87.

Sabbagh, M. A. & Baldwin, D. A. (2001) Learning words from knowledgeable versus ignorant speakers: Links between preschoolers' theory of mind and semantic development. *Child Development*, *72*(4), 1054-1070.

Sabbagh, M. A., & Shafman, D. (2009). How children block learning from ignorant speakers. *Cognition*, *112*(3), 415-422.

Sobel, D. M., & Corriveau, K. H. (2010). Children monitor individuals' expertise for word learning. *Child development*, *81*(2), 669-679.

Scofield, J., Gilpin, A. T., Pierucci, J., & Morgan, R. (2013). Matters of accuracy and conventionality: Prior accuracy guides children's evaluations of others' actions. *Developmental psychology*, *49*(3), 43.

Tenney, E. R., Small, J. E., Kondrad, R. L., Jaswal, V. K., & Spellman, B. A. (2011). Accuracy, confidence, and calibration: How young children and adults assess credibility. *Developmental Psychology, 47*(4), 1065-1077.

Tummeltshammer, K. S., Wu, R., Sobel, D. M., & Kirkham, N. Z. (2014). Infants track the reliability of potential informants. *Psychological Science*, *25*(9), 1730-1738.

Both thematic role and next-mention biases affect pronoun use in Dutch

Jorrig Vogels (j.vogels@rug.nl)
University of Groningen, Center for Language and Cognition (CLCG)
Oude Kijk in 't Jatstraat 26, 9712 EK Groningen, the Netherlands

Abstract

An important question is whether speakers consider listeners' expectations when choosing whether to use a pronoun. It has been suggested that certain thematic roles are more expected to be mentioned again, and are therefore more likely to be pronominalized. In the present study, we aim to disentangle predictability effects on pronoun use from thematic-role effects. To this end, we conducted two web-based continuation experiments in Dutch, in which the next-mention biases associated with Source-Goal and Agent-Patient verbs were manipulated to create a shift in the bias. Experiment 1 confirmed that the manipulations changed the biases. Experiment 2 showed that while thematic role mainly influenced demonstrative and full pronoun use for non-subjects, next-mention biases played a role in the choice between reduced and full pronouns and between pronouns and full NPs, irrespective of thematic role or grammatical function. Thus, thematic role and predictability seem to affect pronoun use in different ways.

Keywords: Dutch; next-mention biases; predictability; pronouns; referring expressions; thematic role

Introduction

Listeners' expectations about whom or what the speaker will mention next influence their interpretation of pronouns. For example, in the sentence *Anna hurt Meryl so she...* the pronoun *she* most likely refers to Meryl (e.g. *so she got mad at her*), while in *Anna recognized Meryl so she...* the pronoun is more likely to refer to Anna (e.g. *so she walked up to her to greet her*). An important question is whether speakers take such expectations into account when choosing to use a pronoun or not. A logical hypothesis would be that speakers use a pronoun when its interpretation is in line with the semantic bias, i.e. when it refers to the person that the listener expects to be mentioned next. If the speaker instead wants to continue with the person that is not expected to be mentioned next, she will signal this by repeating the name.

This is exactly what certain accounts of reference production predict (e.g. Arnold, 2001, 2008; Givón, 1983): Speakers use pronouns for referents that they believe the listener is already expecting to be mentioned, and they use more elaborate referring expressions when the referent is thought to be not very predictable. However, several researchers have found that the choice for a pronoun is not influenced by how predictable the referent is (Fukumura & Van Gompel, 2010; Rohde & Kehler, 2014; Stevenson, Crawley, & Kleinman, 1994). Recently, it has been suggested that whether predictability plays a role in pronoun use may depend on the verb in the preceding clause (Rosa & Arnold, 2017). For example, whereas Fukumura and Van Gompel (2010) did not find a predictability effect on pronoun use in implicit causality contexts with Stimulus-Experiencer verbs, Arnold (2001) and Rosa and Arnold (2017) found in transfer-of-possession verbs that Goal referents were more often pronominalized than Source referents, with the assumption that Goal referents are more predictable than Source referents. However, they did not test the effect of predictability directly (cf. Pickering & Majid, 2007; Kehler & Rohde, 2013). While it may be true that Goals are more likely to be pronominalized because they are more predictable, it may also be the case that this thematic role is more salient for other reasons, for instance because it is often an obligatory argument of the verb (cf. Fukumura & Van Gompel, 2010).

The first aim of this study is therefore to disentangle predictability effects on pronoun use from thematic-role effects. The second aim is to explore whether predictability and thematic role also play a role in the choice of referring expression in Dutch. So far, almost all psycholinguistic studies on this topic have been done on English (but see Bott, Solstad, & Pryslopska, 2018 for a study on German). Dutch is an interesting language to investigate, because it offers more referential options than English. First, Dutch, like German, has a set of demonstrative pronouns that can refer anaphorically to humans as well as inanimates. Second, most personal pronouns in Dutch have two variants: a full form (e.g. *zij* 'she$_{full}$') and a reduced form (e.g. *ze* 'she$_{reduced}$'; see e.g. Kaiser, 2011).[1] It is an open question how the different factors that are argued to play a role in referring expression selection affect speakers' choices between these multiple possible referential forms.

We conducted two web-based written continuation experiments in which participants were presented with a context sentence for which they needed to type a suitable continuation, starting with the connective *vervolgens* 'subsequently'. To be able to generalize across different thematic roles, the context sentences contained either a Source-Goal verb, such as *geven* 'give', or an Agent-Patient verb, such as *bellen* 'call'. All verbs had a default next-mention bias to the second NP (NP2) when combined with either a forward temporal or a consequence coherence relation, as established by previous research (Commandeur, 2010; Koornneef & Sanders, 2012)[2]. That is, when the continuation of a sentence fragment expresses a consecutive event or a consequence of the event expressed by the verb, people tend to interpret a subject pronoun in the continuation as referring to the NP2.

[1] The masculine 3rd person reduced pronoun *ie* 'he' is different from the feminine reduced pronoun in that it mostly occurs in spoken language. It also behaves differently syntactically in that it cannot appear sentence-initially.

[2] We could not find data on next-mention biases for Dutch Source-Goal verbs, so we took these from translations of the English verbs in Rosa and Arnold (2017). Also, some Agent-Patient verbs were translations from English verbs used in Cheng (2016).

We then manipulated these next-mention biases, such that they would shift to the first NP (NP1), which is normally less likely to be mentioned next. We did this in two ways: For some sentences we varied the social status of the referents. When combining social roles with a high or low status with verbs such as 'criticize' or 'mock', the person with low status is expected to be more likely to feature in the continuation of the event (cf. Garvey, Caramazza, & Yates, 1974). Table 1A presents examples for this manipulation, with the expected effect on the next-mention bias. For other sentences we included either a neutral adverb such as *meteen* 'right away', an adverb expressing unintentionality such as *per ongeluk* 'by accident', or the adverb *eerst* 'first'. In combination with the connective *vervolgens* 'subsequently', the latter is expected to create a strong expectation for a subject continuation, because it induces a parallel coherence relation (cf. Kehler, Kertz, Rohde, & Elman, 2008). For the unintentionality adverbs, we expected a tendency to shift the next-mention bias more to the Source/Agent character (cf. Cheng, 2016). Table 1B presents sample sentences for the adverb manipulation.

In Experiment 1, participants were free to continue the context sentences in any way they wanted, as long as they started with the connective *vervolgens* 'subsequently'. The goal of this experiment was to test whether the manipulations indeed affected the next-mention bias. That is, we predicted that participants would be more likely to continue with the NP1 when it refers to a low-status character, or is accompanied by one of the critical adverbs (see Table 1). In Experiment 2, either NP1 or NP2 was underlined, and participants had to refer to this NP as the subject of their continuation. The goal of this experiment was to test whether participants' choice of referring expression would depend on whether they had to refer to a referent that was consistent or inconsistent with the next-mention bias. Here, we predicted that participants choose a more reduced type of referring expression for referents that are more likely to be mentioned next, and a more elaborate expression for less-expected referents. If the thematic-role effect found in previous research is a predictability effect, thematic role should not play a role in referring expression choice. Alternatively, if thematic role has a separate effect, it should affect referring expression choice irrespective of next-mention bias.

Experiment 1

Methods

Participants. Seventy-four Dutch-speaking participants were recruited via social media and email. We discarded the data from participants who did not complete the experiment, leaving 48 participants. Of these, 33 were women, 13 were men, and 2 did not make a choice. Mean age was 27.7 years (range 18-60). Participants were not paid.

Materials. We created 30 Dutch context sentences containing verbs identified as having an NP2 next-mention bias. Of these, 15 were Source-Goal verbs, and 15 were Agent-Patient verbs. For all items, the bias was manipulated either by varying the social status of the characters in the sentence

Table 1. Sample sentences for the social-status (**A**) and adverb (**B**) manipulations, by verb type. The rightmost column shows the expected next-mention bias for each condition. *Unintent.* = Unintentionality adverb.

A. Social-status manipulation			**Expected bias**
Source-Goal	*High-Low*	*De moeder gaf een uitbrander aan haar dochter. Vervolgens ...* 'The mother gave a scolding to her daughter. Next ...'	NP2 (default)
	Low-High	*De dochter gaf een uitbrander aan haar moeder. Vervolgens ...* 'The daughter gave a scolding to her mother. Next ...'	NP1
Agent-Patient	*High-Low*	*De bazin bekritiseerde de assistente. Vervolgens ...* 'The $\text{boss}_{\text{female}}$ criticized the $\text{assistant}_{\text{female}}$. Next …'	NP2 (default)
	Low-High	*De assistente bekritiseerde de bazin. Vervolgens ...* 'The $\text{assistant}_{\text{female}}$ criticized the $\text{boss}_{\text{female}}$. Next …'	NP1
B. Adverb manipulation			
Source-Goal	*Neutral*	*De gravin gaf op het feest de halsketting aan de meid. Vervolgens ...* 'The countess gave the necklace to the maid at the party. Next ...'	NP2 (default)
	Unintent.	*De gravin gaf per ongeluk de halsketting aan de meid. Vervolgens ...* 'The countess gave the necklace to the maid by accident. Next ...'	NP1~NP2
	First	*De gravin gaf eerst de halsketting aan de meid. Vervolgens ...* 'The countess first gave the necklace to the maid. Next ...'	NP1
Agent-Patient	*Neutral*	*De boerin belde meteen de vroedvrouw. Vervolgens ...* 'The farmer's wife called the midwife right away. Next …'	NP2 (default)
	Unintent.	*De boerin belde per ongeluk de vroedvrouw. Vervolgens ...* 'The farmer's wife called the midwife by accident. Next …'	NP1~NP2
	First	*De boerin belde eerst de vroedvrouw. Vervolgens ...* 'The farmer's wife first called the midwife. Next …'	NP1

(e.g. *bazin–assistente* 'boss$_{female}$'–'assistant$_{female}$'; 12 items: 9 Agent-Patient and 3 Source-Goal verbs) or by varying the adverb in the sentence (18 items: 6 Agent-Patient and 12 Source-Goal verbs). The adverb was either neutral (e.g. *meteen* 'right away'), an unintentionality adverb (e.g. *per ongeluk* 'by accident'), or the adverb *eerst* 'first'. To discourage participants to only use pronouns, character pairs were always same-gender (although sometimes they were gender-ambiguous, such as 'officer'–'soldier'), and they did not include proper names. To control for grammatical function, we also created Goal-Source and passive Patient-Agent variants of each item (e.g. *De dochter kreeg een uitbrander van haar moeder* 'The daughter got a scolding from her mother'; *De vroedvrouw werd eerst door de boerin gebeld* 'The midwife was first called by the farmer's wife'). The first word of the participant's continuation was given, and was always the connective *vervolgens* 'subsequently'.

In addition, we created 36 filler items using a variety of syntactic structures, and including proper names, animals and NP conjunctions. The connective was also varied. The items were presented in a pseudo-random order, interspersed with the filler items, such that no two experimental items followed each other directly.

Procedure. The experiment was distributed via the online survey software Qualtrics. Upon clicking on the link, participants received an instruction screen, saying that they would see a series of sentences, for which they had to type a continuation (starting with a pre-given connective) in the text-entry bar, using their first intuition. There was no time limit. After about every 10th trial, a cute animal picture appeared on the screen, and participants were allowed to take a short break. The experiment took about 30 minutes to complete.

Design and analysis. Varying thematic role order (Goal/Patient=NP2, Goal/Patient=NP1) and either social status of the Goal/Patient (low, high) or adverb (neutral, unintentional, first) as within-items factors resulted in a 2x2 or 2x3 design, depending on the manipulation. Given this design, the items were distributed over 6 lists, such that each item occurred only once on a list. Since social status had only two levels, lists 5 and 6 repeated conditions from lists 1 and 2 for this variable. Verb type (Source-Goal, Agent-Patient) was varied between items.

We analyzed the proportion of Goal/Patient references out of all references, in separate analyses for the social-status and the adverb manipulation. The binary predictors were centered. The predictor adverb was contrast coded with neutral adverb as the reference level. Logit mixed-effects analyses including all main effects and second-order interactions with either social status or adverb were run. We aimed for a maximal random-effects structure, but removed random slopes step-by-step in case of non-convergence (see Bates, Kliegl, Vasishth, & Baayen, 2015). We furthermore tested for the inclusion of random slopes and the fixed effects of the control variables verb type and thematic role order using Likelihood Ratio tests.

Results

We excluded trials in which participants did not refer to NP1 or NP2 as the subject of their continuation (298 cases), used a plural expression (47 cases), selected the wrong gender (15 cases), did not produce a completion (9 cases) or did not use verb second word order (4 cases), as well as trials in which three annotators could not reach agreement on the referent (68 cases). This resulted in the removal of 30.6% of the data, leaving 999 cases for analysis.

We found clear effects of both the social-status and the adverb manipulation on the choice of referent. In the social-status analysis, there were significant main effects of social status (β = -1.49, SE = 0.26, p < .001) and thematic role order (β = 0.89, SE = 0.24, p < .001): When the Goal/Patient had a higher social status than the Source/Agent, participants were less likely to continue with the Goal/Patient, and in the canonical (Source-Goal, Agent-Patient) orders even showed a Source/Agent-bias (see Figure 1). In the non-canonical (Goal-Source, Patient-Agent) orders, there was an overall stronger Goal/Patient-bias, suggesting an additional subject-bias. The main effect of verb type was not significant (p = .10), and there were no interactions (ps > .1).

For the adverb manipulation, there was a significant difference between the adverb 'first' and neutral adverbs (β = -1.73, SE = 0.47, p < .001), a significant main effect of the-

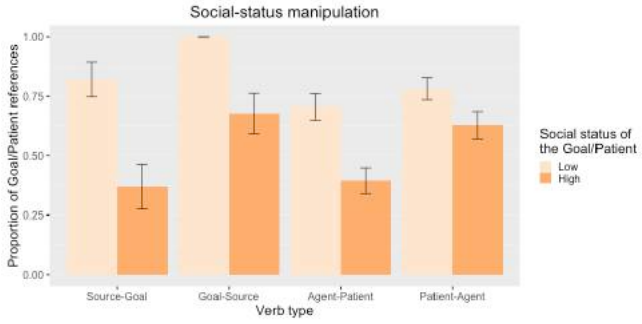


Figure 1: The proportion of Goal/Patient references after Source-Goal and Agent-Patient verbs, including their reversed orders, by the social status of the Goal/Patient.

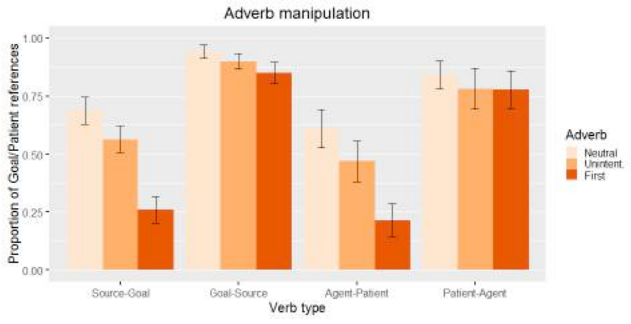


Figure 2: The proportion of Goal/Patient references after Source-Goal and Agent-Patient verbs, including their reversed orders, by type of adverb.

matic role order (β = 2.26, *SE* = 0.25, $p < .001$), and an interaction between the two predictors (β = 1.55, *SE* = 0.67, $p < .05$): In the canonical (Source-Goal, Agent-Patient) orders, participants were less likely to continue with the Goal/Patient when the context sentence contained the adverb 'first', and even showed a Source/Agent-bias (see Figure 2). In the non-canonical (Goal-Source, Patient-Agent) orders, there was again an overall stronger Goal/Patient-bias, and a weaker effect of adverb.[3] The difference between unintentionality and neutral adverbs was not significant (p = .55), and neither was the main effect of verb type (p = .13).

Discussion

The results of Experiment 1 confirm that the next-mention bias of Source-Goal and Agent-Patient verbs can be influenced by manipulating the social status of the referents and by adding certain adverbs. For the canonical thematic role orders, the original Goal/Patient bias even shifted to a Source/Agent bias. In the non-canonical orders, the effect was smaller, probably due to an added bias to refer to the first-mentioned NP (e.g. Gernsbacher & Hargreaves, 1988).

In Experiment 2, we subsequently tested whether manipulating the next-mention bias also affects the choice of referring expression, predicting more pronouns for referents that are consistent with the bias.[4] We used the same method as in Experiment 1, except that one referent in the context sentence was underlined, and participants were asked to start their continuation with this referent.

Experiment 2

Methods

Participants. Ninety-eight Dutch-speaking participants were recruited via social media and email. None had participated in Experiment 1. We removed 44 participants who did not complete the experiment, and 2 participants who were not native Dutch speakers, leaving 52 participants. Of these, 40 were women and 12 were men. Mean age was 37.0 years (range 16-75). Participants were not paid.

Materials. In order to shorten the experiment duration, we selected 16 items from Experiment 1 that showed the largest effect of the next-mention-bias manipulations: 8 from the social-status manipulation (5 Agent-Patient and 3 Source-Goal verbs) and 8 from the adverb manipulation (2 Agent-Patient and 6 Source-Goal verbs). Because the unintentionality adverb condition was not significantly different from the neutral adverb condition, we dropped the former.

We manipulated which referent had to be referred to in the continuations (either NP1 or NP2) by underlining this referent in the context sentences. The referent had either a Source/Agent or a Goal/Patient role. Furthermore, in the social status manipulation the referent had either low or high social status. In the adverb manipulation, it was either combined with a neutral adverb or with *eerst* 'first'. The items were distributed over 6 lists, and interspersed with 24 fillers, in the same way as in Experiment 1.

Procedure. The procedure was identical to Experiment 1, except for the fact that participants were now instructed to start their continuation with the referent that was underlined. The experiment took about 20 minutes to complete.

Design and analysis. We performed separate analyses testing the effect of our next-mention-bias manipulations on three dependent variables: the proportion of pronouns *including* demonstratives out of all references, the proportion of pronouns *excluding* demonstratives out of all references, and the proportion of reduced pronouns out of all pronouns. In all analyses, we included the next-mention-bias manipulation (high/low social status; neutral adverb/first), as well as the referent's grammatical function (Subject (NP1), Non-Subject (NP2)) and thematic role (Source/Agent, Goal/Patient) as predictors, resulting in a 2x2x2 within-items design. Since the effect of verb type was not significant in Experiment 1, we collapsed over Source and Agent, and over Goal and Patient. All predictors were centered. Logit mixed-effects analyses including all main effects and second-order interactions with the bias manipulation were run in the same way as in Experiment 1.

Next, we also tested whether Goal referents were more likely to be pronominalized than Source referents. For this, we ran separate logit mixed-effects analysis on the two verb types (Source-Goal, Agent-Patient), with the referent's thematic role (Goal, Source; Agent, Patient) and grammatical function (Subject (NP1), Non-Subject (NP2)) as predictors, and the proportion of pronouns (including demonstratives) as the dependent variable.

Results

We excluded 61 cases where participants did not refer to the correct referent, 18 cases in which reference was unclear, and 1 case of self-correction, leading to the removal of 9.6% of the data and leaving 752 cases for analysis.

In the social-status analysis, we found a significant main effect of social status on the proportion of reduced pronouns out of all pronouns (β = -1.40, *SE* = 0.65, $p < .05$), with more reduced pronouns when the referent had lower social status (see Figure 3)[5], as well as a main effect of gramma-

[3] Contrary to expectation, in the non-canonical orders the preference to refer to the subject was weaker after 'first' than after a neutral adverb, suggesting a bias towards the Source/Agent rather than the subject. We henceforth consider Source/Agent referents in a sentence with 'first' as consistent with the manipulated bias.

[4] We designed a new experiment to test this question rather than coding the results of Experiment 1 for choice of referring expression because the biases in that experiment would yield a very unbalanced design, i.e. there would be many more references to expected than to unexpected referents.

[5] The data included only one occurrence of the masculine 3rd person reduced pronoun *ie* 'he'. The effect reported here is therefore entirely driven by the feminine reduced pronoun *ze* 'she'.

tical function (β = -2.05, *SE* = 0.80, p < .05), with more reduced pronouns for subjects than for non-subjects. No significant effect of social status was found on the proportion of pronouns out of all references, both including (p = .25) and excluding (p = .26) demonstrative pronouns.

In the adverb analysis, we found a significant interaction between adverb and thematic role on the proportion of pronouns (including demonstratives) out of all references (β = -1.73, *SE* = 0.80, p < .05), with more pronouns for Source/Agent and fewer pronouns for Goal/Patient referents in sentences with 'first' than with a neutral adverb (see Figure 4). This interaction effect was stronger when excluding demonstratives (β = -2.29, *SE* = 0.83, p < .01), suggesting that it is primarily driven by the use of personal rather than demonstrative pronouns. In both analyses, the main effect of grammatical function was also significant (β = -3.47, *SE* = 0.57, p < .001 and β = -5.11, *SE* = 0.78, p < .001, respectively), with more pronouns for subjects than for non-subjects. Although Figure 4 suggests an interaction between adverb and thematic role on the proportion of reduced pronouns out of all pronouns, this was not significant (p = .18).

Finally, Figure 5 shows that, irrespective of next-mention bias, pronouns were more frequent for Goal than for Source non-subjects, although the interaction did not reach significance (p = .06). In addition, the difference seems to be largely due to an increase in demonstrative and full pronouns.[6] For Agent-Patient verbs, pronouns seem to be more frequent for Agent than for Patient non-subjects, but the interaction was not significant (p = .95).

Discussion

The results of Experiment 2 showed effects of the next-mention-bias manipulations on pronoun use: more reduced pronouns for low-status than for high-status referents, and more personal pronouns for Source/Agent referents (as well as fewer pronouns for Goal/Patient referents) in contexts including the adverb *eerst* 'first'.

In addition, thematic role seemed to have an effect on the choice of referring expression beyond these next-mention-bias manipulations. Consistent with Rosa and Arnold (2017), Goal non-subjects were more likely to be pronominalized than Source non-subjects, although not reliably. Moreover, this difference seemed to be due to a larger number of full and demonstrative pronouns for Goal referents. Since full pronouns in Dutch are canonically used for contrastive referents, and demonstrative pronouns for less salient (non-topical) referents (e.g. Kaiser, 2011), this might suggest that Goals are not as salient as subject referents, but salient enough to not be referred to with a full definite NP.

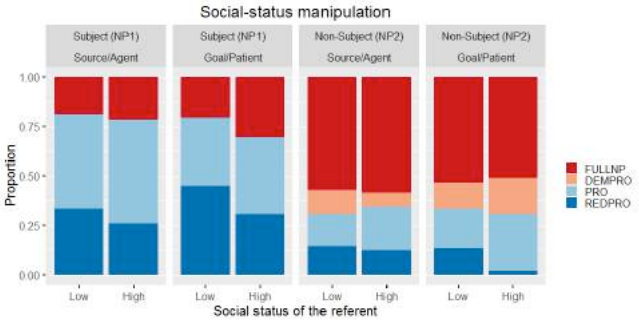


Figure 3. Choice of referring expression in the social status manipulation of Experiment 2, by the referent's social status, grammatical function, and thematic role.

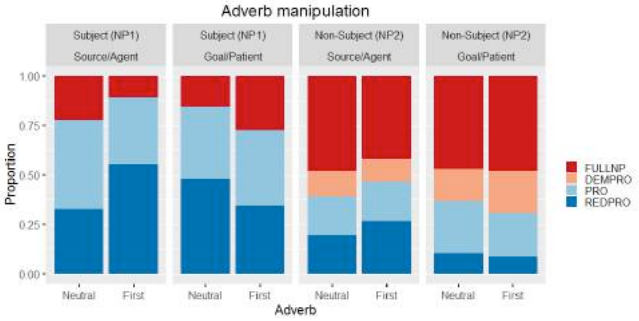


Figure 4. Choice of referring expression in the adverb manipulation of Experiment 2, by adverb and the referent's grammatical function and thematic role.

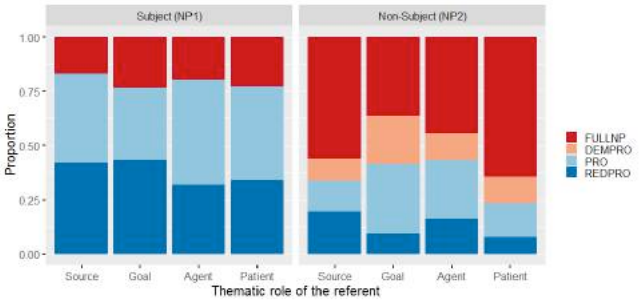


Figure 5. Choice of referring expression in Experiment 2, by the referent's thematic role and grammatical function.

Thus, assuming that next-mention bias is an accurate measure of predictability, the results of Experiment 2 suggest that thematic role and predictability affect referential choices in different ways: Goals stand out among the thematic roles because they attract more demonstrative and full pronoun references, while predictable referents are more likely to be referred to with reduced forms (reduced vs. full pronouns or personal pronouns vs. full NPs).

[6] Post-hoc analyses supported this: When excluding demonstratives, the trend for an interaction between thematic role and grammatical function disappeared (p = .69); An analysis on the proportion of reduced pronouns out of all pronouns showed a significant interaction between thematic role and grammatical function (β = -2.05, *SE* = 0.91, p < .05). Paired comparisons showed a significant increase in reduced pronouns for Goal vs. Source non-subjects (p < .01), but not for subjects (p = .20).

General discussion

In this paper, we investigated whether expectations about what will be mentioned next influence the choice of referring expression. The first aim was to disentangle predictability effects on pronoun use from effects of thematic role. We did this by manipulating the next-mention bias in sentences with Source-Goal and Agent-Patient verbs. For some sentences, we varied the social status of the characters, hypothesizing that the lower-status character would be more likely to be mentioned next. For other sentences, we varied the type of adverb, hypothesizing a stronger Source/Agent bias with unintentionality adverbs and a stronger subject bias with the adverb *eerst* 'first'. The second aim was to explore predictability effects on reference production in Dutch, which has a rich spectrum of anaphoric expressions.

The results of Experiment 1 confirmed that the manipulations affected next-mention biases, especially in sentences in which the Goal or Patient was the second NP, where the bias shifted from the NP2 to the NP1. Sentences including 'first' showed increased references to the Source/Agent rather than to the subject, suggesting that the induced parallel coherence relation was semantic rather than syntactic. The effect of the unintentionality adverb was not as strong (cf. Cheng, 2016), and we therefore removed this condition from Experiment 2.

Experiment 2 showed that the shifts in next-mention bias also affected the choice of referring expression: When the referent had a relatively low social status, participants were more likely to mention it in their continuations (Experiment 1), and they also produced more reduced pronouns as compared to full pronouns, irrespective of grammatical function or thematic role (Experiment 2). This finding is consistent with information-theoretic accounts of language production, which propose that more predictable linguistic material is reduced (e.g. Levy & Jaeger, 2007). It is also in line with a contrastive interpretation of full pronouns (Kaiser, 2011), in which use of the full form pragmatically implicates that it refers to something else than the predictable referent.

When the context sentence anticipated a parallel coherence relation (in the form of 'first…next…'), participants were more likely to mention the Source/Agent referent in their continuations (Experiment 1), and they were also more likely to use a personal pronoun compared to a full NP to refer to these referents. Conversely, they were less likely to pronominalize the Goal/Patient character (Experiment 2). This suggests that next-mention biases may also affect the choice between a pronoun and a full NP in Dutch. Whether there is a fundamental difference between referential biases stemming from the social-status and the adverb manipulations is unclear. The current experiment may simply have lacked the power to detect all the effects.

Irrespective of next-mention bias, thematic role also seemed to have an effect on the choice of referring expression: Goals tended to be more likely to be pronominalized than Sources, at least for non-subjects, in line with Rosa and Arnold (2017). However, this preference was largely driven by the use of demonstrative and full pronouns as opposed to reduced forms. Demonstrative pronouns in Dutch are considered to be used mainly for non-topical referents (e.g. Kaiser, 2011). Indeed, in our study these forms exclusively occurred with non-subjects (see Figures 3-5). The choice of a full over a reduced pronoun is often driven by some form of contrast (Kaiser, 2011). The use of these 'stronger' pronominal forms to refer to Goal non-subjects might suggest that such referents are intermediately salient: They are more salient than other non-subjects, warranting the use of pronouns over full NPs, but not as salient as the average subject to allow for the use of a reduced pronoun.

Taken together, the results of this study have implications for current theories of reference. One line of research argues that what drives referential choices is how likely the referent is to be mentioned next (e.g. Arnold, 2008; Tily & Piantadosi, 2009). If a referent is highly predictable, a pronoun will be used; if it is unexpected, the speaker will signal this by using a full NP. The main evidence for this claim comes from the finding that Goal referents are more likely to be pronominalized than Source referents (Arnold, 2001; Rosa & Arnold, 2017). However, other researchers have argued that what makes a referent predictable is not necessarily its thematic role, but the specific event structure and coherence relation that links two references (Kehler & Rohde, 2013; Pickering & Majid, 2007). The present results point to the possibility that both thematic role and predictability based on event structure affect the choice of referring expression, but in different ways. Although it may still be the case that Goal referents are more likely to be mentioned next, the increase in pronoun use for Goals may also have a different origin. It has been noted, for example, that Goal non-subjects are often an obligatory argument of the verb (indirect object), whereas Source non-subjects are mostly optional (Fukumura & Van Gompel, 2010). This may make Goals more salient. Our results are therefore consistent with a form-specific multiple-constraints approach to reference, in which different referential forms are sensitive to different aspects of the referent (Kaiser & Trueswell, 2008).

A second line of research argues that there is an asymmetry between production and interpretation of referring expressions (e.g. Fukumura & Van Gompel, 2010; Rohde & Kehler, 2014): While reference resolution may be influenced by next-mention biases, reference production is driven only by grammatical or information structural factors. The present results suggest that next-mention biases may in fact influence reference production, at least in Dutch. So far, most studies on this topic have been on English, and investigating referential choices in a language with a richer set of referring expression types, such as Dutch, may reveal patterns that otherwise remain hidden.

Finally, effects of predictability may become more manifest in more engaging communicative settings that involve an actual addressee (cf. Rosa & Arnold, 2017). Since the effects we are seeking are probably small and tend to be overridden by stronger factors, the logical next step is to replicate the current findings in a larger-scale study in a more naturalistic, but still controlled, context.

Acknowledgments

This research was supported by the Netherlands Organisation for Scientific Research (NWO), under Grant 275-89-0360. Many thanks to Daniëlle Fluks, Anouck Braggaar, and Boudewijn Blank for their help with the creation of the materials and the design, conduction and analysis of Experiment 1.

References

Arnold, J. E. (2001). The effect of thematic roles on pronoun use and frequency of reference continuation. *Discourse Processes*, *31*(2), 137–162.

Arnold, J. E. (2008). Reference production: Production-internal and addressee-oriented processes. *Language and cognitive processes*, *23*(4), 495–527.

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *ArXiv Preprint ArXiv:1506.04967*.

Bott, O., Solstad, T., & Pryslopska, A. (2018). Implicit causality affects the choice of anaphoric form. Poster presented at *AMLaP 2018*, September 6-8, 2018, Berlin, Germany.

Cheng, W. (2016). *Implicit Causality And Consequentiality In Native And Non-Native Coreference Processing*. Doctoral dissertation, University of South Carolina.

Commandeur, E. (2010). *Implicit causality and implicit consequentiality in language comprehension*. Doctoral dissertation, Tilburg University.

Fukumura, K., & Van Gompel, R. P. (2010). Choosing anaphoric expressions: Do people take into account likelihood of reference? *Journal of Memory and Language*, *62*(1), 52–66.

Garvey, C., Caramazza, A., & Yates, J. (1974). Factors influencing assignment of pronoun antecedents. *Cognition, 3*(3), 227–243.

Gernsbacher, M. A., & Hargreaves, D. J. (1988). Accessing sentence participants: The advantage of first mention. *Journal of Memory and Language*, 27(6), 699.

Givón, T. (1983). *Topic continuity in discourse: A quantitative cross-language study (Vol. 3)*. John Benjamins Publishing.

Kaiser, E. (2011). Salience and contrast effects in reference resolution: The interpretation of Dutch pronouns and demonstratives. *Language and Cognitive Processes, 26*(10), 1587–1624.

Kaiser, E., & Trueswell, J. C. (2008). Interpreting pronouns and demonstratives in Finnish: Evidence for a form-specific approach to reference resolution. *Language and Cognitive Processes*, *23*(5), 709–748.

Kehler, A., Kertz, L., Rohde, H., & Elman, J. L. (2008). Coherence and Coreference Revisited. *Journal of Semantics*, *25*(1), 1–44.

Kehler, A., & Rohde, H. (2013). A probabilistic reconciliation of coherence-driven and centering-driven theories of pronoun interpretation. *Theoretical Linguistics*, *39*(1–2), 1–37.

Koornneef, A. W., & Sanders, T. J. (2013). Establishing coherence relations in discourse: The influence of implicit causality and connectives on pronoun resolution. *Language and Cognitive Processes*, *28*(8), 1169–1206.

Levy, R. P., & Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. In *Advances in neural information processing systems* (pp. 849–856).

Pickering, M. J., & Majid, A. (2007). What are implicit causality and consequentiality? *Language and Cognitive Processes*, *22*(5), 780–788.

Rohde, H., & Kehler, A. (2014). Grammatical and information-structural influences on pronoun production. *Language, Cognition and Neuroscience*, *29*(8), 912–927.

Rosa, E. C., & Arnold, J. E. (2017). Predictability affects production: Thematic roles can affect reference form selection. *Journal of Memory and Language*, *94*, 43–60.

Stevenson, R. J., Crawley, R. A., & Kleinman, D. (1994). Thematic roles, focus and the representation of events. *Language and Cognitive Processes, 9*(4), 519–548.

Tily, H., & Piantadosi, S. (2009). Refer efficiently: Use less informative expressions for more predictable meanings. In *Proceedings of the workshop on the production of referring expressions: Bridging the gap between computational and empirical approaches to reference*.

Cognitive Abilities to Explain Individual Variation in the Interpretation of Complex Sentences by Older Adults

Margreet Vogelzang (margreet.vogelzang@uni-oldenburg.de)
Institute of Dutch Studies and Cluster of Excellence "Hearing4all"
University of Oldenburg, Germany

Christiane M. Thiel (christiane.thiel@uni-oldenburg.de)
Department of Psychology and Cluster of Excellence "Hearing4all"
University of Oldenburg, Germany

Stephanie Rosemann (stephanie.rosemann@uni-oldenburg.de)
Department of Psychology and Cluster of Excellence "Hearing4all"
University of Oldenburg, Germany

Jochem W. Rieger (jochem.rieger@uni-oldenburg.de)
Department of Psychology and Cluster of Excellence "Hearing4all"
University of Oldenburg, Germany

Esther Ruigendijk (esther.ruigendijk@uni-oldenburg.de)
Institute of Dutch Studies and Cluster of Excellence "Hearing4all"
University of Oldenburg, Germany

Abstract

This paper investigates which cognitive abilities predict the interpretation of complex sentences by older adults. Participants performed a picture-selection task after hearing complex and simpler sentences, as well as a broad test battery of cognitive tests. The results show that different cognitive factors serve as predictors for the interpretation of complex sentences compared to simpler sentences. For complex sentences, verbal intelligence, cognitive flexibility, and working memory capacity are strong predictors. Our study thus shows that older adults' interpretation of sentences of varying complexity is influenced by different cognitive abilities, and stresses the need to take such individual differences into account when studying language processing.

Keywords: language processing; cognitive factors, complex sentences; syntactic structure; age; individual variation

Introduction

It is well-known in cognitive-linguistic research that syntactically complex sentences can be difficult to process (a.o. Bahlmann, Rodriguez-Fornells, Rotte, & Münte, 2007 (object-first sentences); Tun, Benichov, & Wingfield, 2010 (object relative clauses), Bader & Meng, 1999 (embedded clauses)). Especially older adults show difficulties with the processing and interpretation of complex sentences (e.g., Emery, 1985). These difficulties could partially be caused by cognitive abilities, as language processing has long been suggested to be influenced by (working) memory capacity (e.g., King & Just, 1991). In reading research, it has been found that working memory capacity and reading experience (but not vocabulary) can mediate older adults' reading times on temporarily ambiguous sentences (Payne et al., 2014). Contrary, in sentence processing in adverse listening conditions, vocabulary was found to influence older adults' performance, as was cognitive flexibility (also described as mental flexibility; McAuliffe, Gibson, Kerr, Anderson, & LaShell, 2013; Rosemann et al., 2017).

So, several cognitive factors have been suggested to influence older adults' language processing performance. Nevertheless, no consensus has been reached about which cognitive factors exactly influence complex sentence processing in older adults, and little is known about influence of cognitive abilities on the processing and interpretation of complex in comparison to simpler sentences. We therefore ran a broad test battery to examine **which cognitive abilities predict the interpretation of complex sentences by older adults**.

Object-first sentences are a common example of complex sentences. In German, canonical word order in a main clause is subject-verb-object (Zwart, 1997). However, the language allows for structurally more complex object-before-subject sentences, for example:

(1) *Den*$_{ACC}$ *Jungen* *wäscht der*$_{NOM}$ *Vater*
The$_{ACC}$ boy washes the$_{NOM}$ Father
'The father washes the boy'

In (1), case on the determiners indicates which noun phrase is the object (den Jungen) and which is the subject (der Vater). Although unambiguous, such object-first sentences have been found to elicit longer reading times (Hemforth, 1993) and more interpretation errors (Carroll, Uslar, Brand, & Ruigendijk, 2016) compared to subject-first sentences in German. Alternatively, an adjunct can be added at the beginning of the sentence to create a structure in which all information about the protagonists follows the verb, such as in (2).

Table 1: Examples of the four experimental conditions. Note that although the conditions use different word orders, their meaning remains the same.

Subject-object order	Adjunct position	Condition	Example sentence
subject-before-object	3	SVAO	Der Igel berührt am Montag den Hasen
object-before-subject	3	OVAS	Den Hasen berührt am Montag der Igel
subject-before-object	1	AVSO	Am Montag berührt der Igel den Hasen
object-before-subject	1	AVOS	Am Montag berührt den Hasen der Igel
			On Monday the$_{NOM}$ hedgehog touches the$_{ACC}$ hare

(2) *Am Montag wäscht den$_{ACC}$ Jungen der$_{NOM}$ Vater*
On Monday washes the$_{ACC}$ boy the$_{NOM}$ Father
'On Monday the father washes the boy'

In this paper, we describe an auditory sentence-processing paradigm followed by a picture-selection task. Two types of syntactic manipulations are used, namely subject-object order and adjunct position. We measured performance of our older participants on several cognitive factors that have been argued to be related to sentence comprehension: age, years of education, working memory capacity, subjective memory complaints, vocabulary, cognitive flexibility, and a composite measure of cognitive performance, which is widely used as screening for cognitive impairment.

Overall, we expect structurally more complex object-before-subject sentences to be more difficult to interpret than subject-before-object sentences (in line with Carroll et al., 2016). We additionally expect adjunct-first sentences to be more difficult to interpret than adjunct-third sentences, as adjunct-first sentences also violate canonical word order (i.e. verb-subject-object rather than subject-verb-object). Moreover, we expect considerable variation in both the interpretation of complex sentences (cf. Vos, Gunter, Schriefers, & Friederici, 2001) and the performance on the cognitive tasks. We expect the performance on several cognitive factors to influence the interpretation of complex sentences, such as age (Rosemann et al., 2017), working memory (Payne et al., 2014; Vos et al., 2001), and vocabulary and cognitive flexibility (McAuliffe et al., 2013; Rosemann et al., 2017). It will then be investigated which of these cognitive tasks best accounts for the interpretation of complex sentences by older adults.

Methods

Participants

20 older adults (age 51-70, mean age 60; 15 females) participated in the study. All participants had age-normal hearing as tested before the experiment and normal or corrected-to-normal vision. The participants were all monolingual native speakers of German and reported no language impairments and no psychiatric or neurological issues. The ethics committee of the University of Oldenburg approved of the study (reference number Drs. 28/2017) and written informed consent was obtained from all participants.

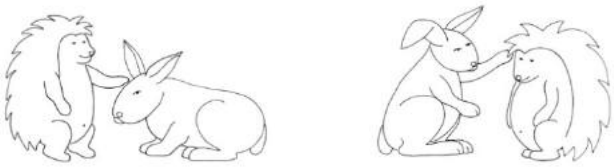


Figure 1: Example pictures corresponding to the sentences in Table 1.

Main Linguistic Task

The linguistic task used auditorily presented German sentences based on the OLACS corpus (Uslar et al., 2013), each followed by two pictures for a picture-selection task. Each sentence consisted of a Subject (S), a transitive Verb (V), an Adjunct (A), and an Object (O). Four different sentence conditions were used (see Table 1): SVAO sentences with canonical word order, OVAS sentences in which the object is placed sentence-initially, adjunct-initial AVSO sentences, in which the verb is placed before its arguments, and adjunct-initial AVOS sentences, in which the subject-object order is additionally manipulated. The task was performed in an fMRI scanner, which inherently creates noise. Therefore, a pre-task was used to control for the loudness of presentation of the stimuli[1].

After each sentence, two pictures (modified from Wendt, Brand, & Kollmeier, 2014) were displayed. These presented both characters mentioned in the sentence performing the mentioned action (the adjunct was not displayed in the pictures, see example pictures corresponding to the sentences in Table 1 in Figure 1). Participants could indicate the picture that best fit the sentence with a response box: the left button for the left picture and the right button for the right picture. The location of the target picture on the screen (left or right) was counterbalanced across trials.

[1] The loudness was adjusted for each participant individually to 80% intelligibility with the Oldenburg Matrix Sentence Test (OLSA; Wagener, Kühnel, & Kollmeier, 1999). The average adjusted loudness of stimuli presentation was 71.5dB (SD = 6.8).

The experiment used 24 sentences per condition, so 96 trials in total. The trials were distributed over two sessions. Two lists with pseudo-randomized presentation orders were created.

Cognitive Tasks

In addition, several cognitive tests were applied: A standard backwards *Digit Span* task (Tewes, 1991) as a measure of simple working memory capacity, the Comprehensive *Trail Making* test (Reynolds, 2002) as a measure of cognitive flexibility, a German *Vocabulary* test called 'Wortschatztest' (Schmidt & Metzler, 1992) as an index of verbal intelligence, the Montreal Cognitive Assessment (*MoCA*; Nasreddine et al., 2005)[2] as a concise screening tool for mild cognitive impairment, and a German version of the self-reported age-related *Memory Assessment* Clinics Questionnaire (Crook, Feher, & Larrabee, 1992) as an index of subjective memory complaints. Finally, participants' *Age* and years of formal *Education* (from primary school up to high school/university/PhD) were assessed through a questionnaire. Of the Trail Making task, following Rosemann et al. (2017), only trail 1 (connecting numbers in order: 1-2-3...) and trail 5 (connecting numbers and letters alternatingly in order: 1-A-2-B-3-C...) were used and participants' score was calculated as the difference in completion time between trail 5 and trail 1.

Procedure

Participants were tested individually at the University of Oldenburg. First, participants were asked to fill out a questionnaire asking for age and years of education, as well as the self-reported Memory Assessment questionnaire. Second, pure-tone audiogram measurements were taken in a soundproof booth. Then, the Trail Making, MoCA, and Digit Span tasks were conducted. After two practice rounds of 6 sentences with the same conditions as in the main experiment, the main linguistic experiment started. The pre-task controlling for the loudness of the stimuli and the main linguistic experiment took place in an MRI scanner; the fMRI results will be published in a separate paper. Participants used headphones during all tasks in the MRI scanner. After the first session of the main experiment, participants came out of the scanner and performed the Vocabulary task, before going back into the scanner for the second session of the experiment and some structural scans. The total testing time was about 3 hours.

Analyses and Results

One participant's Trail Making task was not performed in line with the experiment protocol and therefore excluded from the analysis (the participant took off their glasses halfway through the task). All other participants completed all the tasks.

[2] Approval for the use of this test was obtained from the MoCA Clinic & Institute.

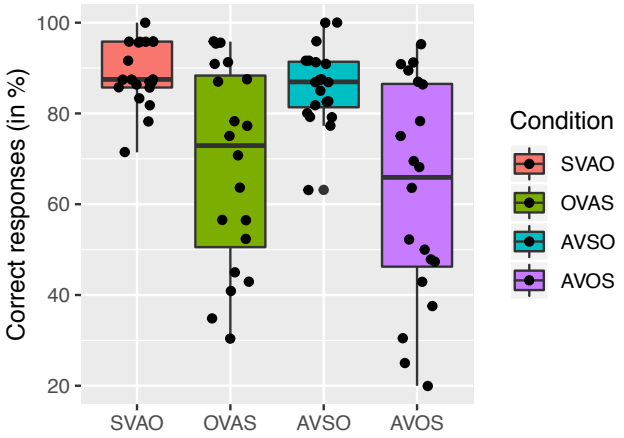


Figure 2: Percentages of correct responses and distribution of these responses on the linguistic task per condition. Each dot indicates the mean score of a participant on a condition. Overall means per condition are SVAO: 89%, OVAS: 69%, AVSO: 86%, and AVOS: 63%.

Main Linguistic Task

We first analyzed the correct responses per condition on the picture-selection task (Figure 2) with generalized linear mixed-effects models (lme4, Bates, Maechler, Bolker, & Walker, 2014). Based on the experimental design, the fixed effects of subject-object order and adjunct position as well as their interaction were included in the model. Based on model comparisons, random intercepts for subjects and items, as well as random slopes for subject-object order per subject and for subject-object order and adjunct position per item were included as random factors in the model. Subject-before-object and adjunct-third were used as the baseline.

The model results (Table 2) show lower performance on object-before-subject sentences (OVAS and AVOS) than on subject-before-object sentences (SVAO and AVSO). This confirms our expectation that object-before-subject sentences are more complex and more difficult to interpret than subject-before-object sentences. No significant effect of adjunct position or interaction between subject-object order and adjunct position was found.

Table 2: Statistical comparison of response accuracies in the linguistic task (corrected for multiple comparisons).

Factor	***ß***	***z*-score**	***p*-value**
Subject-object order	-1.38	-4.32	< 0.001
Adjunct position	-0.30	-2.09	0.07
Subject-object order* Adjunct position	-0.04	-0.16	0.87

As can be seen in Figure 2, on both of the more complex object-before-subject conditions, OVAS and AVOS, participants show very large individual variation; on the subject-before-object conditions participants show less variation. We will now look at whether this individual variation can be explained by the participants' cognitive abilities.

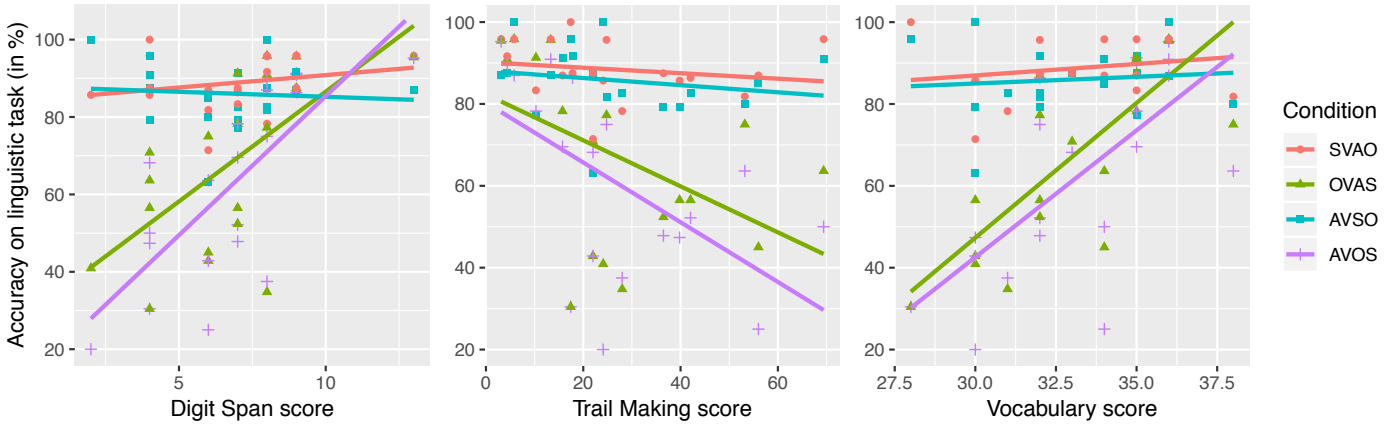


Figure 3: The relation between the performance on the linguistic task and on the Digit Span task (left), indicating working memory capacity, on the Trail Making task (middle), indicating cognitive flexibility, and on the Vocabulary task (right), indicating verbal intelligence.

Cognitive Tasks

We will perform two sets of analyses investigating the relation between the tested cognitive factors and performance on the linguistic task. First, we will examine the influence of each single cognitive factor by running generalized linear mixed-effect models with each factor separately. This analysis will show which cognitive factors influence the processing of simple and complex sentences. Additionally, we will examine the influence of the cognitive factors in combination with each other on the linguistic task by means of an inference tree. This will show which combination of cognitive factors has the strongest predictive power when it comes to interpreting sentences of different complexities and which are thus most useful to take into account when investigating (complex) sentence processing.

Linear mixed-effects models (lme4, Bates, Maechler, Bolker, & Walker, 2014) were developed for each cognitive task separately. The same model was used as for the analysis of the linguistic task, including the fixed effects of subject-object order and adjunct position as well as their interaction, but this time adding the cognitive tasks as co-variates. All significant effects are reported in the text, but only the most interesting effects will be elaborated upon. The results show, after correcting for multiple comparisons, no effects of *MoCA*, *Age*, and *Education* on the responses on the linguistic task (all p's > 0.05). A main effect of *Memory Assessment* was found (β = 0.08; z = 2.89; p < 0.01), indicating that people with more memory complaints actually performed better on the linguistic task.

For *Digit Span*, a significant main effect (β = 0.22; z = 3.38; p < 0.001) as well as an interaction with subject-object order (β = 0.38; z = 4.08; p < 0.001) were found. In Figure 3 (left panel), the relation between participants' scores on the Digit Span task and on the linguistic task per condition is plotted. The figure shows that participants with higher scores on the Digit Span task (indicating better working memory capacity) perform much better on the object-before-subject conditions than participants with lower scores. Conversely, no clear effect of working memory capacity is observed in the subject-before-object order, suggesting that processing the object-before-subject sentences requires additional working memory capacity compared to subject-before-object word order.

For *Trail Making*, also an interaction with subject-object order (β = -0.04; z = -2.86; p < 0.001) was found. Figure 3 (middle panel) shows the relation between participants' scores on the Trail Making task and on the linguistic task per condition. A higher trail making score reflects worse performance on the Trail Making test. Hence, subjects with worse scores on the Trail Making task perform worse on the linguistic task with object-before-subject sentences.

Finally, for *Vocabulary,* a significant main effect (β = 0.21; z = 3.59; p < 0.001) as well as an interaction with subject-object order (β = 0.33; z = 4.04; p < 0.001) were found. In Figure 3 (right panel) the relation between participants' scores on the Vocabulary task and on the linguistic task per condition is shown, making it clear that participants with higher verbal intelligence performed better on the object-before-subject conditions than participants with lower verbal intelligence.

All three interactions occur with object-before-subject sentences, indicating that interpretation of complex object-before-subject but not simpler subject-before-object sentences is influenced by these factors. Thus, it appears that processing more complex sentences draws on additional cognitive resources, whereas processing the simpler subject-before-object sentences requires less resources, presumably because they do not require additional analysis.

Best Predictors

One could argue that performance on Digit Span, Trail Making, and Vocabulary could be intercorrelated, and therefore these tasks may all explain the same effects in the data. Therefore, we will now investigate which cognitive factors *in combination* form the best predictors for the interpretation of sentences with different complexities and thus which are most useful to take into account in future investigations. Because investigating all factors within one mixed-effects model creates difficulties due to the large amount of variables, we favor conditional

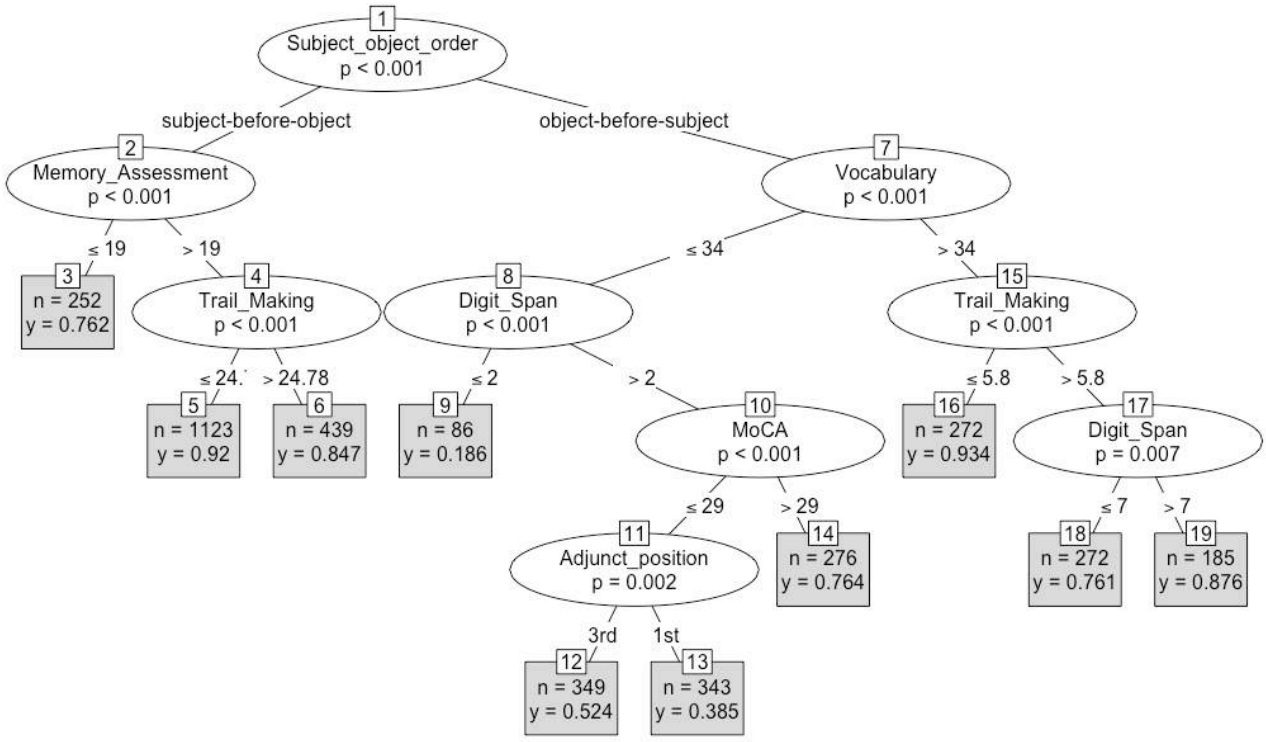


Figure 4: A conditional inference tree showing which of the tested predictors (*Subject-object order, Adjunct position, Digit Span, Trail Making, Vocabulary, MoCA, Memory Assessment, Age,* and *Education*) are the best predictors of response accuracy. The gray boxes indicate the number of trials (n) and proportion of correct responses (y) per branch.

inference trees (ctree from the party package, Hothorn, Hornik, & Zeileis, 2006), which are a type of decision tree. These provide non-parametric tree-based regression models and can handle large numbers of variables.

This method uses a significance test procedure in order to select the variables (cognitive factors) that best predict the response accuracy on the linguistic task. Notably, our two linguistic conditions, subject-object order and adjunct position, are entered as possible predictor variables as well. Using this method, we can investigate which variables most strongly predict the accuracy on our picture-selection task; stronger predictors are higher up in the tree. Variables that are not significant predictors do not occur in the tree at all. Besides showing which variables are predictors of response accuracy, the tree also shows how well individual performance can be predicted given these variables. The results of the analysis are shown in Figure 4 (including *p*-values per variable). Each branch of the tree represents the trials in certain conditions for participants with certain cognitive scores.

The results show that subject-object order is the strongest predictor for the performance on the picture-selection task, as it is highest up in the tree. For subject-before-object sentences, subjective *Memory Assessment* is the strongest predictor, followed by *Trail Making* score lower in the tree. In the gray boxes, the number of trials (n) and the mean proportion of correct responses (y), are displayed for each branch. For example, the 252 trials in the leftmost branch, in the simpler subject-before-object conditions responded to by people with a Memory Assessment score equal to or smaller than 19, were answered correctly 76.2% of the time.

For more complex object-before-subject sentences, *Vocabulary* is the strongest predictor of performance on the linguistic task, followed by *Digit Span* and *MoCA* for participants with a lower Vocabulary score, and *Trail Making* and *Digit Span* for participants with a higher Vocabulary score. Notably, MoCA is a significant predictor only for people with a lower Vocabulary score and a higher digit span score; this explains why there was no main effect of MoCA in the linear mixed-effects models. Interestingly, for participants with a lower Vocabulary score, a higher Digit Span score and a lower MoCA score, adjunct position is a significant predictor of performance on the linguistic task, whereas for other participants no influence of adjunct position is found.

Importantly, Vocabulary, Digit Span, Trail Making (and MoCA) all appear in the decision tree, indicating that they explain different parts of the data, i.e. they are complementary, and therefore that taking all these tests into account is useful when examining complex sentence interpretation.

Discussion

In this paper, we aimed to identify cognitive abilities that can predict the interpretation of complex sentences by older adults. We will focus on the most clear and convincing results here. As predicted, complex object-before-subject sentences were more difficult for older adults to interpret than subject-before-object sentences. Contrary to our predictions, adjunct-first sentences were only more difficult to interpret than adjunct-third sentences for part of the participants. Regarding cognitive abilities, we found that working memory capacity (*Digit*

Span), cognitive flexibility (*Trail Making*), and verbal intelligence (*Vocabulary*) are not only correlated with complex sentence processing as single factors, but also when they are combined. The analysis of all factors combined showed an additional effect of general cognitive performance (*MoCA*) for participants with lower verbal intelligence. Interestingly, age and years of education did not influence participants' performance (compare Stine-Morrow, Ryan, & Leonard, 2000).

The strong predictive power of verbal intelligence is striking. This does not reflect familiarity with the words in the linguistic task, since all conditions, simple and complex, used the same words. Rather, it could indicate that people with a broader and deeper vocabulary are able to access and process words more easily, thereby freeing up capacity for higher-level processing (in line with evidence from speech recognition, McAuliffe et al., 2013, but contra Ramscar, Hendrix, Shaoul, Milin, & Baayen, 2014). Moreover, verbal intelligence could be related to general intelligence or language (e.g., reading) experience (cf. Payne et al., 2014), which may increase familiarity with and aid the processing of complex structures.

Overall, our linguistic picture-selection task was quite challenging, which is reflected in the large amount of individual variation. Some participants showed around chance performance on object-before-subject sentences, suggesting that (1) their working memory capacity was insufficient to keep and manipulate all information in memory (cf. Just & Carpenter, 1992), (2) their verbal intelligence was insufficient to access their lexicon efficiently (cf. McAuliffe et al., 2013), (3) their cognitive flexibility was insufficient to process sentences with non-typical word order, and/or (4) their general cognitive performance was insufficient to process complex sentences. Conversely, simpler subject-before-object sentences do not seem to require high working memory capacity or verbal intelligence. This dissociation is in line with the idea that processing syntactically simpler constructions loads general cognitive resources less than processing syntactically more complex constructions.

One could argue that all cognitive factors that were found to affect complex sentence processing actually all belong to one latent factor. Ramscar et al. (2014), for example, suggest that a larger and more experienced lexicon has more complex representations, which require more demanding searches to be accessed, causing delays and decreased performance on linguistic and other psychometric tests. Our analyses show, however, that the different cognitive factors have a *complementary* effect; they explain different parts of the data, suggesting that the tasks tap into different underlying mechanisms.

Conclusion

Our study identified several cognitive factors that can serve as predictors of the interpretation of complex sentences by older adults, which differ from the factors that predict the interpretation of simpler sentences. The investigation thus highlights the complementary influence of different cognitive abilities on language processing, and emphasizes the need to consider not only working memory capacity, but also factors such as verbal intelligence and cognitive flexibility when investigating complex sentence processing.

Acknowledgements

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2177/1 - Project ID 390895286. We would like to thank Jan Michalsky for his help with recording the stimuli and Rebecca Carroll for her help with recording and preparing the stimuli.

References

Bader, M., & Meng, M. (1999). Subject-object ambiguities in German embedded clauses: An across-the-board comparison. *Journal of Psycholinguistic Research*, *28*(2), 121–143. http://doi.org/10.1023/A:1023206208142

Bahlmann, J., Rodriguez-Fornells, A., Rotte, M., & Münte, T. F. (2007). An fMRI study of canonical and noncanonical word order in German. *Human Brain Mapping*, *28*(10), 940–949. http://doi.org/10.1002/hbm.20318

Bates, D. M., Maechler, M., Bolker, B., & Walker, S. (2014). Package 'lme4.'

Carroll, R., Uslar, V., Brand, T., & Ruigendijk, E. (2016). Processing Mechanisms in Hearing-Impaired Listeners: Evidence from reaction times and Sentence Interpretation. *Ear and Hearing*, *37*(6), e391–e401. http://doi.org/10.1097/AUD.0000000000000339

Crook, T. H., Feher, E. P., & Larrabee, G. J. (1992). Assessment of Memory Complaint in Age-Associated Memory Impairment: The MAC-Q. *International Psychogeriatrics*, *4*(2), 165–176. http://doi.org/10.1017/S1041610292000991

Emery, O. B. (1985). Language and aging. *Experimental Aging Research*, *11*(1), 3–60. http://doi.org/10.1080/03610738508259280

Hemforth, B. (1993). *Kognitives Parsing: Repräsentation und Verarbeitung sprachlichen Wissens*. Sankt Augustin, Germany: Infix Verlag.

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics*, *15*(3), 651–674. http://doi.org/10.1198/106186006X133933

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, *99*(1), 122–149. http://doi.org/10.1037/0033-295X.99.1.122

King, J., & Just, M. A. (1991). Individual differences in syntactic processing: The role of working memory. *Journal of Memory and Language*, *30*(5), 580–602. http://doi.org/10.1016/0749-596X(91)90027-H

McAuliffe, M. J., Gibson, E. M. R., Kerr, S. E., Anderson, T., & LaShell, P. J. (2013). Vocabulary influences older and younger listeners' processing of dysarthric speech. *The Journal of the Acoustical Society of America*, *134*(2), 1358–1368. http://doi.org/10.1121/1.4812764

Nasreddine, Z. S., Phillips, N. A., Bédirian, V.,

Charbonneau, S., Whitehead, V., Collin, I., … Chertkow, H. (2005). The Montreal Cognitive Assessment, MoCA: A brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society*, *53*(4), 695–699. http://doi.org/10.1111/j.1532-5415.2005.53221.x

Payne, B. R., Grison, S., Gao, X., Christianson, K., Morrow, D. G., & Stine-Morrow, E. A. L. (2014). Aging and individual differences in binding during sentence understanding: evidence from temporary and global syntactic attachment ambiguities. *Cognition*, *130*(2), 157–73. http://doi.org/10.1016/j.cognition.2013.10.005

Ramscar, M., Hendrix, P., Shaoul, C., Milin, P., & Baayen, H. (2014). The Myth of Cognitive Decline: Non-Linear Dynamics of Lifelong Learning. *Topics in Cognitive Science*, *6*(1), 5–42. http://doi.org/10.1111/tops.12078

Reynolds, C. R. (2002). *Comprehensive TrailMaking Test*. Austin, TX: Pro-Ed.

Rosemann, S., Gießing, C., Özyurt, J., Carroll, R., Puschmann, S., & Thiel, C. M. (2017). The Contribution of Cognitive Factors to Individual Differences in Understanding Noise-Vocoded Speech in Young and Older Adults. *Frontiers in Human Neuroscience*, *11*(294), 1–13. http://doi.org/10.3389/fnhum.2017.00294

Schmidt, K.-H., & Metzler, P. (1992). *Wortschatztest [MultipleChoice Word Test]*. Weinheim: Beltz Test GmbH.

Stine-Morrow, E. A. L., Ryan, S., & Leonard, J. S. (2000). Age Differences in On-Line Syntactic Processing. *Experimental Aging Research*, *26*(4), 315–322. http://doi.org/10.1080/036107300750015714

Tewes, U. (1991). *Hamburg-Wechsler-Intelligenztest für Erwachsene — Revision 1991 (HAWIE-R)*. Bern: Huber.

Tun, P. A., Benichov, J., & Wingfield, A. (2010). Response latencies in auditory sentence comprehension: Effects of linguistic versus perceptual challenge. *Psychology and Aging*, *25*(3), 730–735. http://doi.org/10.1037/a0019300

Uslar, V. N., Carroll, R., Hanke, M., Hamann, C., Ruigendijk, E., Brand, T., & Kollmeier, B. (2013). Development and evaluation of a linguistically and audiologically controlled sentence intelligibility test. *The Journal of the Acoustical Society of America*, *134*(4), 3039–3056. http://doi.org/10.1121/1.4818760

Vos, S. H., Gunter, T. C., Schriefers, H., & Friederici, A. D. (2001). Syntactic parsing and working memory: The effects of syntactic complexity, reading span, and concurrent load. *Language and Cognitive Processes*, *16*(1), 65–103. http://doi.org/10.1080/01690960042000085

Wagener, K. C., Kühnel, V., & Kollmeier, B. (1999). Entwicklung und evaluation eines satztests in deutscher sprache I: design des oldenburger satztests [Development and evaluation of a German sentence test I: design of the oldenburg oldenburg sentence test]. *Zeitschrift Für Audiologie*, *38*, 4–15.

Wendt, D., Brand, T., & Kollmeier, B. (2014). An eye-tracking paradigm for analyzing the processing time of sentences with different linguistic complexities. *PLoS ONE*, *9*(6). http://doi.org/10.1371/journal.pone.0100186

Zwart, C. J. W. (1997). The Germanic SOV Languages and the Universal Base Hypothesis. In L. Haegeman (Ed.), *The New Comparative Syntax*. London/New York: Longman.

Thinking Locally or Globally? – Trying to Overcome the Tragedy of Personnel Evaluation with Stories or Selective Information Presentation

Momme von Sydow[1,2] (momme.von-sydow@urz.uni-münchen.de)
Niels Braus[2] (n.braus@stud.uni-heidelberg.de)
Ulrike Hahn[1,3] (u.hahn@bbk.ac.uk)
[1]University of Munich (LMU), MCMP, Ludwigstr. 31, D-80539 München, Germany
[2]University of Heidelberg, Department of Psychology, Hauptstr. 47-51, D-69117 Heidelberg, Germany
[3]Birkbeck College London, Department of Psychological Science, Malet Street, London WC1E 7HX, U.K.

Social dilemmas conceptually suggest distinguishing direct individual and group-level effects (also involving indirect effects on others). Furthermore, the success of organizations appears to rely on identifying not only individual excellence but positive impact on others as well. In 'Two-Level Personnel Evaluation Tasks' (T-PETs) participants as human resource managers evaluate employees when individual and group contributions are dissociated. Von Sydow, Braus, & Hahn (2018) have suggested a potential 'Tragedy of Personnel Evaluation': A group-serving employee with the smallest individual contribution but by far the greatest positive effect on the group's overall earnings was often rated the most negatively. Here we investigate, in two experiments with conflicting information, whether emphasizing the group can avert the 'tragic' outcome. Our results suggest that the tragedy is not as complete as suggested, and that contextual information can mitigate the tragedy. Nonetheless, the results also corroborate the stability of underestimating the impact of team players.

Keywords: Co-variation Detection; Inner-Individual Dilemma; Co-operation; Multi-Level Approach Simpsons Paradox

Individual versus Group Utility

Co-operation between individuals – over and above direct *individual* benefit – is essential for the common good of organisations, companies, and society on the whole. Successful co-operation often involves setting personal interests aside and devoting oneself, at least partly, to the success of an organization or team. Thus when evaluating behaviour one must distinguish direct individual utility from group utility.

In evolutionary biology, multi-level selection-approaches stress the differences between behaviour benefiting either the individual or the group; and these models allow for the rise of altruism (Sober & Wilson, 1999; Nowak & Sigmund, 2005; Wilson & Wilson, 2007).

Likewise, social-dilemma situations can be interpreted as conflicts of individual and group interests; and it has been argued that purely self-interested, economically 'rational' behaviour may inevitably lead to the depletion of public and natural resources. This has become associated with the label 'Tragedy of the Commons' (Hardin, 1968). Research in game-theory, behavioural economics and psychology (e.g., on ultimatum games, dictator games and public-good games) has revealed strategies for solving social-dilemma situations, and that many people do *not* act in a purely self-interested manner, but rather demonstrate at least some preference for distributing goods justly or behaving in a group-serving manner (Engel, 2011; Hendrich et al., 2005, Fehr & Gächter, 2002, Gollwitzer, Rothmund, Pfeiffer, & Ensenbach, 2009; Melis, Hare, & Tomasello, 2016).

In Organisational and Social Psychology, the importance of the team-level has been increasingly acknowledged, emphasizing that teams may be greater than the sum of their parts (Haslam, Steffens, Peters, Boyce, Mallett, & Fransen, 2017; Mathieu, Maynard, Rapp, & Gilson, 2008; Memmert, Plessner, Hüttermann, Froese, Peterhänsel, & Unkelbach, 2015). Likewise, the role of pro-social or altruistic role or extra-role behaviours in teams has been identified as central to the success of companies and organizations (Brief & Motowidlo, 1986; Li, Kirkman, & Porter, 2014; Nielsen, Hrivnak, & Shaw, 2009; Organ, 1997; Podsakoff, Whiting, Podsakoff, & Mishra, 2010).

A crucial question, however, is the extent to which people recognize those who clearly serve the overall good at the team level. This question should be particularly pressing for human-resource managers who must evaluate or select employees and must often only base their judgment on abstract performance data (sales numbers, etc.).

Tragedy of Personnel Evaluation

For human-resource management, it seems crucial to address the potential dissociation of employees' individual and collective impact on team performance. While underlining this, recent research has also provided some first evidence that such behaviour is sometimes rewarded – particularly when managers have direct acquaintance with the processes and persons involved (Organ, 1997; Scotter, Cross, & Motowidlo, 2000; cf. Grant & Patil, 2012, 562). However, personnel managers must increasingly evaluate without first-hand experience, often based on abstract performance numbers (Brandl, 2002).

We have begun to study evaluation situations from an experimental perspective as well, using well specified Two-level Personnel-Evaluation Tasks (T-PETs; von Sydow & Braus, 2016, 2017; von Sydow, Braus, & Hahn, 2017, 2018). In these T-PETs, participants obtained information about employees' earnings on individual as well as overall group levels. The T-PET used involves strongly conflicting information at both levels. The group-serving person *A* is characterized by lowest individual earnings yet has a consistent, strongly positive impact on the overall team earnings by substantially increasing the earnings of the other employees. This group-serving person is here called 'altruist'. Although (behavioural) altruism in biology and economics seem to be associated with this kind of indi-

vidual and group impact, it should be noted that such patterns do not imply *motivational* altruism (only 'prosocial' behaviour). For simplicity, however, we call the group-serving person, team-player or positive interactor simply 'altruist' (*A*).

In previous work on T-PETs, participants evaluated the 'altruist' who was best for the team and company as worst, and they tended to ostracize the altruist in selection tasks. These results led to the suggestion of a potential 'tragedy of personnel selection': Personnel managers may neglect or underestimate group impact with substantial damage to personnel and ultimately companies and organisations (von Sydow & Braus, 2016; von Sydow et al., 2018). This occurred despite the strong correlations between group membership and team performance (von Sydow et al., 2018). Further studies showed that, although negative group-interaction or 'egoist' detection (egoism here again defined behaviourally only) differed slightly from 'altruist' detection, they both demonstrated a broadly similar tragedy of ignorance regarding overall group-level effects (von Sydow & Braus, 2017). Another study already has investigated the role of group size in T-PETs. Holding the effects of *A* on *single* other individualy constant, the small group demonstrated no considerable advantage (von Sydow, Braus, & Hahn, 2017).

The current personnel-evaluation experiments again involve similar scenarios with conflicting information at the individual and group levels (T-PETs). However, they investigate whether a shift in the known importance or salience of the group level, by either varying cover-stories (Exp. 1) or selective information presented (Exp. 2), yields improvement. Experiment 2 additionally examines ratings distinguishing explicitly direct impact, impact on others, and overall impact on a team.

Figure 1: Example of shown earnings at the individual and group levels on a particular day.

Experiment 1: Story-Induced Focus

We used a straightforward manipulation, providing participants with texts stressing either the role of the individual or the team as central for personnel management. The experiment had four conditions, with stories focusing on different levels (C1 individual; C2 global; C3 individual & global; C4 control, no focus and no additional text). Additionally, the order of the dependent variables in all four learning phases was counterbalanced (evaluation → selection vs. selection → evaluation).

Method

Participants The experiment was conducted via MTURK with participants from the US. 121 participants passed the two selection-criteria (time spent on the first page, and the correct choice of a rephrasing of the instructions) and finished the experiment. The participants obtained a compensation of $1.50. 49% were male; the mean age was 36; 55% had a Bachelor's or Master's degree and 30% a high school degree as their highest level of education.

Procedure and material The computer experiment resembled previous T-PETs (von Sydow et al., 2018) and was implemented using SociSurvey.

Participants first obtained general instructions that their role as personnel manager was to evaluate the employees of a particular shop. Daily they would obtain information about individual and total earnings of the team working that day. Overall there were five staff members working in the shop, in day-shifts of four people.

On the next slide, participants read that the retired founder of the company had delivered a talk, mentioning the essential role of the personnel management to a company's success. As space precludes exhaustive citation here, we present only the first and fourth of five paragraphs of C1 and C2:

C1: "What is a company? A company is composed of **individual employees working on their tasks**, and it rises and falls with their performance. Thus a company needs to incentivize the performance of **each individual employee**. […] This alone will do justice to those individuals **who do a good job over those who do a bad job**. In particular, you need to detect those **employees** who **individually perform best and worst."**

C2: **"**What is a company? A company is **more than just the sum of its employees**; it is a **whole**, a finely attuned **organism.** It is made up of **teams** in which employees need to **interact** in a positive way. Thus, a company needs to incentivize **team performance**. […]. This alone will do justice to those **teams with positive interaction** and to **good over bad team players**. In particular, you need to detect the **members of the group** who **support and those who exploit the group"** (bold print added).

C3 analogously emphasized the importance of monitoring both individual AND group effects of employees.

In the main part, participants sequentially obtained for each day transparent overview information about the individual earnings of each of the four employees (presented by a picture) working on the shift that day, as well as information on overall earnings (see Figure 1). The structure of the earnings is shown in Table 1 (we added some noise to each value; a normal distribution with SD = 600 €). On the level of individual earnings there were relatively small mean differences (400 €) between the four non-interacting normal workers N_x and the altruist *A*: *N1 > N2 > N3 > N4 > A*.

Apart from the lowest individual earnings of the 'altruist', his/her presence had by far the most positive impact on the group earnings – when *A* is present, the earnings of all co-workers increase by 1000 €, leading to an overall average increase of 2500 € (normally exceeding the salient mark of 10,000 €).

Table 1: Mean earnings of normal workers (*Nx*) and altruist; and mean overall earnings.

	With Altruist	Without Altruist
N1	3,600 €	2,600 €
N2	3,200 €	2,200 €
N3	2,800 €	1,800 €
N4	2,400 €	1,400 €
Altruist	1,500 €	-
Total	10,500 €	8,000 €

The presence of this worker correlated with $r = .99$ with the overall outcome – a correlation easily detectible by machine-learning algorithms.

Overall, 80 panels (Table 1) were sequentially shown and participants could view the overview panels for each day as long as required, with a minimum of four seconds. The role of the altruist (team player) was randomly assigned to one of the five persons of which four are working in a particular shift. He or she appeared randomly 50% of the days (shifts) in the overview panels (Table 1); the four normal workers appeared randomly. We further counterbalanced the presentation-order of the four employees working in each shift.

In four test phases, after the 20, 40, 60 and 80 rounds, the 'personnel managers' evaluated all employees in an evaluation task and a selection task. The order of these tasks was counterbalanced. In the evaluation task, participants rated the contribution of all employees to the overall earnings of the company, on a scale of one to ten. In the personnel selection task they had to answer which four of the five workers they chose to work another day "to optimise the overall profit for the company". At the end of the fourth test phase, we added a total utility task: "Which person is of the greatest/lowest total utility for your business?" and assessed participants' preference for narrow self-interest and pro-social behaviour, using the social-value orientation scale (Murphy, Ackermann, & Handgraaf, 2011). Finally, they provided comments and demographic data.

Results

Figure 2 shows the average evaluation ratings by person in the four test rounds for the four conditions. First, Figure 2 reveals that the average ratings, in all conditions, mainly reflects the average *individual* earnings of the employee, with the ratings of the altruist, *A*, always being lower than all (or at least most) normal workers, *N*. A repeated-measures ANOVA with factors Person (*N1* to *N4*, A) and Phase (R1 to R4) as within-subject factors, and the condition Story (C1 to C4) as between-subject factors, revealed a significant main effect of Person only, $F(2,257) = 216,20$, $p < .001$, $\eta^2 = .65$, but no significant main effects of Condition or Phase.

Second, the ratings reveal that the order of the mean ratings of the normal workers is always (even in Phase 1) in line with the actual (small) differences observed, resulting overall in order (*N1 > N2 > N3 > N4 > A*), with significant post-hoc tests for all four comparisons (each $p < .001$).

Figure 3 shows the percentage of group-based answers in the selection task (selections of *A*), the rating task (all $N \geq A$) through time, and the final comments (coding some insight into the difference between individual and group performance). Selection, comments, and in the beginning rating as well, seem to reveal a similar pattern: a slight advantage for the conditions with global stories (C2 and C3). Comparing these two conditions with those without global stories (C1 and C4), this predicted effect in Phase 4 reaches significance for selection: $\chi^2(\text{df}=1, N=121) = 4.30$, $p < .05$. For ratings there seems only an effect for the global condition, not the individual and global condition. This difference between selection and rating perhaps becomes understandable if one bears in mind that correct rating here involved a stricter test criterion than selection (not only judging *A* higher than *one* other worker, but *all* other workers).

Discussion Experiment 1

First, the results of Experiment 1 overall corroborate the postulated stability of the Tragedy of Personnel Evaluation with no strong impact of the story. The average ratings of the altruist, who actually very consistently causes strong improvements in group performance, are lowest; and at least 50% of the participants seem not to detect the effects of the presence of the 'altruist'. However, the results also suggest

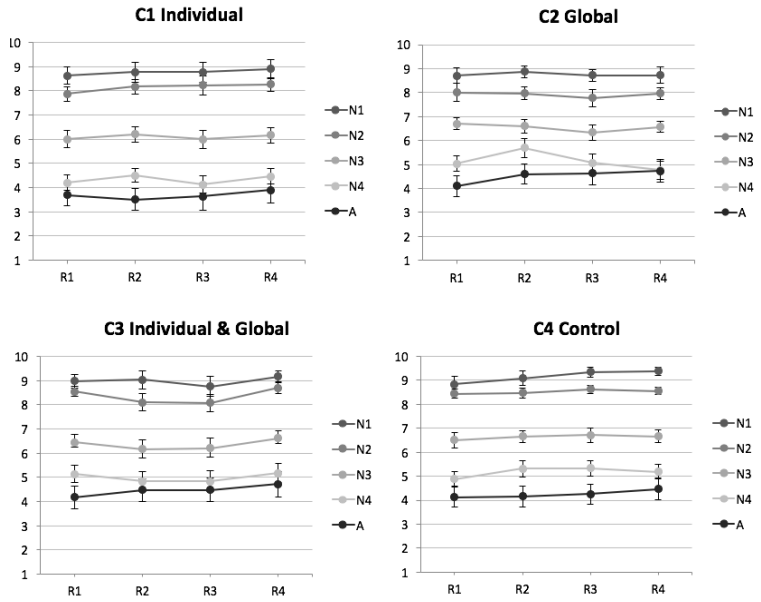


Figure 2: Mean ratings of the employees over test rounds in the four story conditions.

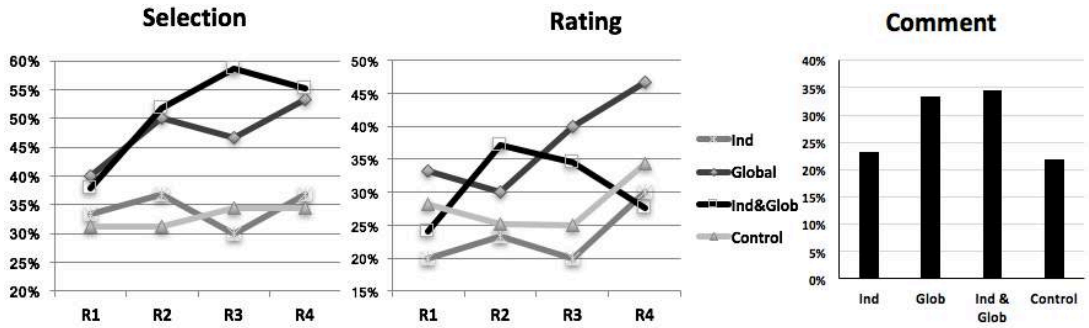


Figure 3: Percentage of group-based selections, evaluations (rating *A* > all *N*), and comments by conditions in Experiment 1.

some small effects of the story, and that thinking about the group level can reduce the Tragedy of Personnel Evaluation at least with regard to some participants.

Experiment 2 – Selective Information Presentation

Experiment 2 varies what information (group versus individual versus both) is presented and at what time. It also increases the number of test phases, to explore the temporal dynamics in more detail and what people could learn even after ten trials. Finally, it tests the ability of participants to detect group-level (and individual-level) effects, also using ratings explicitly differentiating between individual effects, effects on others and overall team effects.

With regard to selective information presentation, there is a condition focusing participants on group-level information alone (*G*) in order to investigate whether and how quickly the 'altruist' could be detected (now using eight test phases). Presenting only individual-level information (*I*), provides the other extreme base-line condition. Always presenting both – individual and a team's overall earnings (*B*) – replicates Experiment 1 but can now be compared to both benchmark conditions. Moreover, we added several further 'mixed conditions, where the three information formats (*G*, *I*, *B*) changed over time (e.g., GIGIGIGI).

The increased number of test phases may allow participants more easily to realize the tension between individual-level success and overall group-level effects. The contrast to both extreme base-line conditions (only *I* or only *G*), should serve as controls for the level of performance in the intermediate conditions, that 'only *B*' or the mixed conditions. In a number of mixed conditions the shown information is varied over time (e.g., GIGIGIGI). We explore whether they may be adventitious in contrasting the global and individual level.

Method

Participants (recruited by Prolific Academic) came from English-speaking countries (i.e. the UK 52%, the US 32%, Ireland, Australia etc.). 172 participants passed all selection criteria and finished the experiment (cf. Exp. 1). Each participant obtained a compensation of £1.80. The mean age was 32 (59% male, 41% female). Regarding education, all participants had at least high school degree or *A* levels.

Procedure and material The materials and procedure apart from some differences strongly resembled the T-PET of Experiment 1. We used a similar, neutral introduction, but without stories. In the main part, participants again obtained, sequentially for each day, overview information (Figure 1) based on the same average earnings as before (Table 1). Thus, on the *individual* level, the rank-order of earnings for the normal (non-interacting) (N_i) workers and the altruist (*A*) worker was *N1 > N2 > N3 > N4 > A*, with small differences (400 €). The altruist impact on the other workers in a shift was larger, increasing each of their average earnings by 1000 € and the average earnings of the team by 2500 €, resulting in a reversed rank order: *A > N1 > N2 > N3 > N4*.

Participants in the role of personnel managers were again shown information on four employees' earning (out of five employees overall) for 80 shifts (days). We again counterbalanced presentation-order of persons shown (see Exp. 1). Participants could study the overview panels as long as required, but with a minimum of four seconds.

Table 2: Information in phases (P_x) and conditions (C_x).

	P1	P2	P3	P4	P5	P6	P7	P8
C1	*G*	*G*	*G*	*G*	*G*	*G*	*G*	*G*
C2	*B*	*B*	*B*	*B*	*B*	*B*	*B*	*B*
C3	*G*	*G*	*G*	*G*	*B*	*B*	*B*	*B*
C4	*G*	*B*	*G*	*B*	*G*	*B*	*G*	*B*
C5	*G*	*G*	*B*	*B*	*G*	*G*	*B*	*B*
C6	*G*	*I*	*G*	*I*	*G*	*I*	*G*	*I*
C7	*I*	*I*	*I*	*I*	*I*	*I*	*I*	*I*

Note: *B* = both individual & group; *G* = group only; *I* = individual only; C_x = condition; R_x = round

In contrast to Experiment 1, Experiment 2 uses eight (instead of four) test rounds and eight preceding learning phases, each composed of ten days. Moreover, we varied the information formats. In a learning phase 10 information panels (10 days, one for each shift) are shown, with information either only on the individual earnings (*I*; see Figure 1 without last row), on the group earnings (*G*; Figure 1 without middle row) or, finally, on both individual and group earnings (*B*; cf. Figure 1). Table 2 presents the information formats in different conditions and phases.

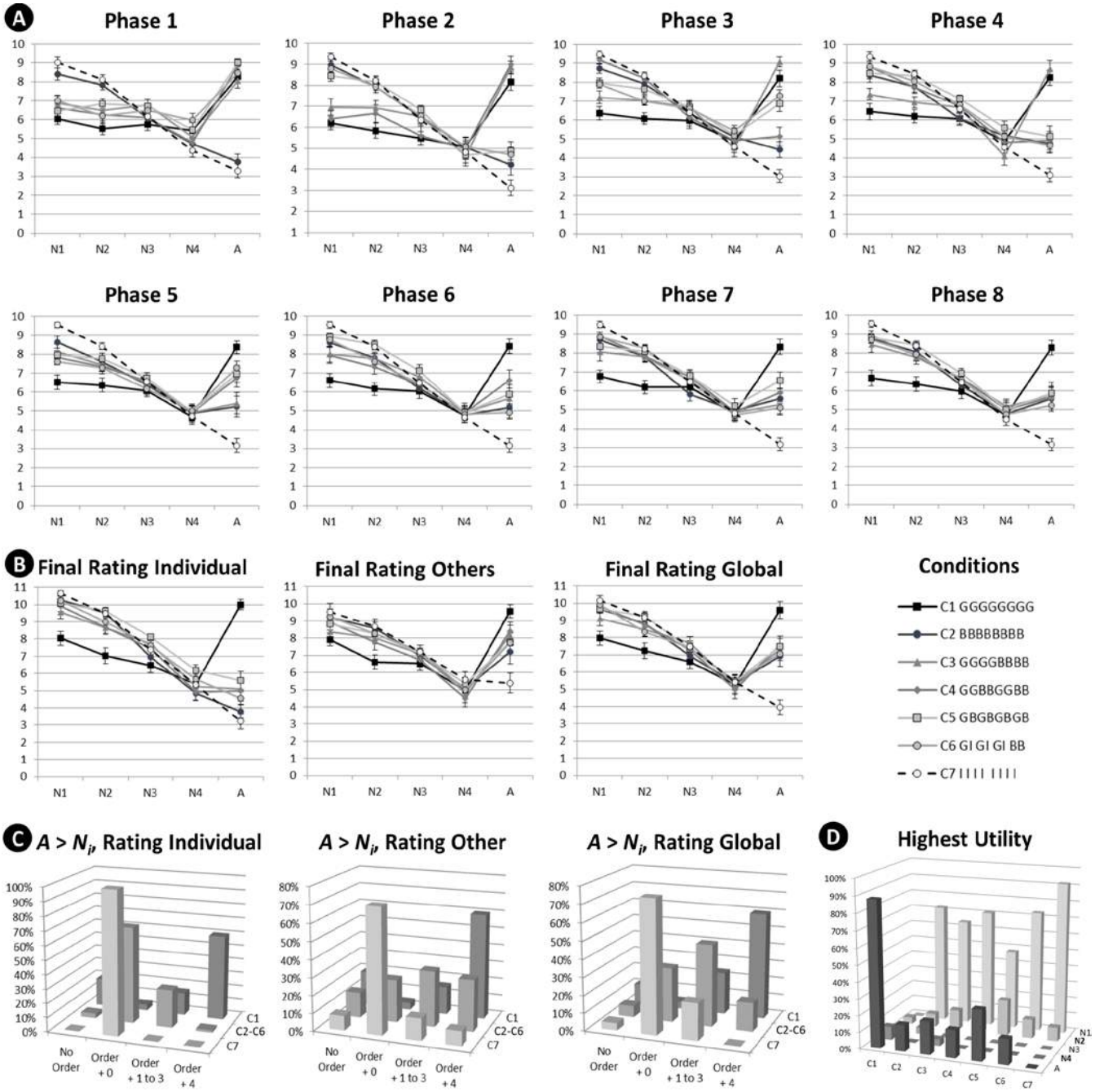


Figure 4: *Panel A* shows participants' mean ratings for the 'altruist' worker (*A*) and the four non-interacting normal workers (*N*) for the seven conditions and eight learning phases. *Panel B* presents the mean of the final ratings, differentiating between workers' individual earnings, effect on others and resulting overall effect (global). For these ratings, *Panel C* shows the percentage of participants who had no sense of individual differences between normal workers ('No Order'; N1+N2 > N3+N4). For those who detected this basic individual order ('Order'), we show the percentage, rating the altruist higher then none of the *N*s ('+ 0'), then one to three Ns ('+1 to 3'), and then all four *N*s ('+ 4'). *Panel D* shows the percentage of participants attributing the overall highest utility for the company to the 'altrust' or a particular normal worker.

In each test round, we administered rating scales and then a personnel selection task. At the end, participants should again provide ratings for all five workers but now explicitly using three rating scales differentiating between individual impact, impact on others and overall contribution. Finally, participants had to choose the employee with greatest or lowest utility and comment on their own behavior.

Results

Figure 4 presents the main results of Experiment 2. *Panel A* shows the participants' mean ratings of the five workers'

contributions in the eight phases. These results show that (a) even after ten presentations (Phase 1) participants' mean judgments suggest a high altruist-detection tendency (at the expense of less clear individual differences between the normal workers); (b) between the phases there was a strong variation of the mean ratings of the mixed middle conditions, suggesting that at least in the beginning participants had difficulties in integrating the results and were strongly influenced by the recent phase; and (c) at the end (Phase 8), however, all middle conditions (C2 to C6) show a middle result between the benchmarks (C1 and C7), and there is no large difference between the middle conditions. Despite still underestimating the altruist, participants clearly do not focus only on individual-level information.

Panel B shows the results of the final ratings differentiating between workers' individual earnings, their impact on others, and the overall impact. (a) Participants in the middle conditions (C2 to C6) clearly differentiate between these ratings. (In the two extreme conditions this was unlikely, but there is even a small effect in C7.) In the middle conditions the individual ratings show $A < N_s$, but the 'other-ratings' show that *A* is, on average, rated higher than N_4, and similarly high as N_3 and N_2. The global rating may be a mixture that seems even more strongly to resemble the 'other-ratings' (but note the incommensurability of the other-rating that used a bipolar scale). In any case, the global rating in the middle conditions clearly does not reach the C1 benchmark, suggesting a remaining tragedy. (b) Comparing the final ratings with the ones in Phase 8 suggests that participants do interpret the latter largely as global ratings (with a small individual influence). (c) Given that the impact-on-others rating would correctly be answered following $A>N_1=N_2=N_3=N_4$, there is not only a correct $A>Ns$, but also an incorrect impact of the individual order $N_1>N_2>N_3>N_4$. This suggests the heuristic '(s)he who is good individually also helps others'.

Panel C investigates the individual differences of the final ratings. (a) Only few participants demonstrate no sense of individual differences ('No Order'; not $N1+N2 > N3+N4$). Though the individual condition C7 does have advantages here, also all other conditions fare reasonably well.

(b) Looking at the others who detected the basic order between the *N*s, the individual and the group condition (C7 and C1) show highly similar results for the three ratings, suggesting a transfer in both directions. In contrast, the middle conditions (C2 to C6) differentiate between the conditions: In the individual ratings, most participants here detect correctly that $A < N$s (Order + 0); and in the 'others-ratings' they rate *A* higher than one, two, three (Order + 1 to 3) or even all four *N*s (Order + 4). The 'others-ratings' and global-ratings make clear that the results of the middle conditions lie between both extreme conditions. Thus it is clearly wrong to claim that all participants completely ignore the group effect; but it is also apparent that only a few rate the altruist as high as would be appropriate based on A's overall (direct and indirect) impact on overall earnings. *Panel D*, finally, shows who the participants judge to be of highest overall utility for the company. (a) C1 shows that, with a focus on group-information, all participants learned the foremost utility of *A*, whereas none learned it in the individual C7. (b) Despite the cited positive effects in the middle conditions, the dependent variable shows that the altruist is still underrated. Nonetheless, a considerable minority also in 'C2 to C6', and more than in C7 ($p < .001$), rates altruist as high as would be appropriate based on A's (direct and indirect) impact on overall earnings ($A >$ *all N*s).

General Discussion

Experiment 1 documents the stability of the Tragedy of Personnel Evaluation. The altruist's or team-player's outstandingly positive *overall* effects on a team were ignored or inadequately acknowledged by most participants; providing a context emphasizing the group had no large effect. However, Experiment 1 suggests that the postulated tragedy could – at least for some participants – be mitigated by contextual cues enhancing focus on the group-level.

Experiment 2 varied the information presented (individual information, group information, or both), used several test-phases, and at the end used rating-scales differentiating between individual earnings, effect on others, and overall impact on group earnings. The results show that participants in principle can learn the overall advantage of *A* as early as in Phase 1 (even after 10 'days') if forced to focus on the global information alone (C1). Second, toggling what information is presented over time did not provide the strong boost we hoped for. Third, and this seems important, the base-line conditions reveal that the tragedy is not at all a total one (at least here after 8 test phases). That is, participants' judgements in all middle conditions (including the only *B* condition) did differ also from an individual-focus condition. Nonetheless, the average ratings, the percentage of correct ratings, and the highest utility judgments show that most participants still substantially underestimate the overall impact of the 'altruist' team player.

Overall, the results are two-fold. They show that the postulated tragedy is neither completely immune to improvement (Exp. 1) nor as radical (Exp. 2) as perhaps originally suggested. However, despite using strong group-level effects, the results obviously do not allow for acquittal but rather corroborate a remaining (but reduced) tragedy. Even though this provides some first evidence for a means to mitigate the tragic, it also continues to underline the danger of potentially similar tragedies in the real world.

Future avenues of research should explore theoretical implications, mediating mechanisms, applications, and boundary conditions of these findings also in real-life settings. Although we here used highly educated participants and strict selection criteria, it would for instance be important to explore the stability of our findings with real personnel managers as well, with or without a number-based task (cf. von Sydow et al., 2018).

Moreover, this research may well be connected to several lines of more theoretically inspired research. For instance, our social-cognition two-level personnel evaluation tasks

may be understood more generally as studies of Simpson's paradox (Waldmann & Hagmayer, 2001; Fiedler et al., 2003; cf. von Sydow, Hagmayer, & Meder, 2016).

Second, we have suggested *some* rational basis for the apparently irrational reluctance to check for large correlations with a high overall outcome (the sum effect of many causal effects). We have suggested that this may be due to a concern with local causal relations rather than ephemeral overall outcomes (von Sydow et al., 2018; cf. Lagnado, Waldmann, Hagmayer, & Sloman, 2006; Sloman & Hagmayer, 2006; Hagmayer & Meder, 2013). Thus other interaction patterns may be more easily detected. For instance, it is known that people are well able to see some logical or causal interaction-patterns if focusing on two or three variables only (e.g., von Sydow, 2016).

In any case, the phenomenon seems of high importance, and the current research warns us that people, at least in the setting of number-based evaluations and perhaps beyond, may well tend to ignore or underestimate the strong overall group effects of team players in contrast to their individual effects.

Acknowledgments

This project was funded by the by the Humboldt Foundation's 'Anneliese Meier Research Award' to Ulrike Hahn and earlier, by a grant Sy111/2-1 from the Deutsche Forschungsgemeinschaft (DFG) to Momme von Sydow.

References

Brief, A. P. & Motowidlo, St. J. (1986), Prosocial Organizational Behavior. *Academy of Management Review, 11*(4), 710–725.

Brandl, J. (2002): Die Problematik der Kennzahlen in Personal-informationssystemen. *Personalführung, 9*, 42-47.

Fiedler, K., Walther, E., Freytag, P., & Nickel, S. (2003). Inductive reasoning and judgment interference: Experiments on Simpson's paradox. *Personality and Social Psychology Bulletin, 29*, 14–27.

Engel, C. (2011). Dictator Games: A Meta Study. *Experimental Economics, 14*, 583–610.

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature, 415*, 137-140.

Gollwitzer, M., Rothmund, T., Pfeiffer, A., & Ensenbach, C. (2009). Why and when justice sensitivity leads to pro- and antisocial behavior. *Journal of Research in Personality, 43(6)*, 999–1005.

Grant, A. M. & Patil, S. V. (2012). Challenging the norm of self-interest. Minority influence and transitions to helping norms in work units. *Academy of Management Review, 37(4)*, 547–588.

Hagmayer, Y., & Meder, B. (2013). Repeated causal decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*, 33-50. doi:10.1037/a0028643.

Hardin, G. (1968). The Tragedy of the Commons. *Science, 162* (3859), 1243–1248.

Haslam, S. A., Steffens, N. K, Peters, K, Boyce, R. A., Mallett C. J., & Fransen, K. (2017). A social identity approach to leadership levelopment. *Journal of Personnel Psychology, 16*(3), 113–124. DOI: 10.1027/1866-5888/a000176

Lagnado, D. A., Waldmann, M. R., Hagmayer, Y., & Sloman, S. A. (2006). Beyond covariation: Cues to causal structure. In A. Gopnik & L. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation*. Oxford: Oxford University Press.

Li, N., Kirkman, B. L., & Porter, C. O. L. H. (2014). Toward a Model of Work Team Altruism. *Academy of Management Review, 39(4)*,541–565. http://dx.doi.org/10.5465/amr.2011.0160

Mathieu, J., Maynard, M. T., Rapp, T., Gilson, L. (2008). Team effectiveness 1997-2007: A Review of Recent Advancements and a Glimpse Into the Future. *Journal of Management, 34*(3), 410-476. DOI: 10.1177/0149206308316061

Melis, A. P., Hare, B., Tomasello, M. (2006). Chimpanzees Recruit the Best Collaborators. *Science, 311*, 1297–1300.

Memmert, D., Plessner, H., Hüttermann, S., Froese, G., Peterhänsel, C., & Unkelbach, C. (2015). Collective fit increases team performances: Extending regulatory fit from individuals to dyadic teams. *Journal of Applied Social Psychology, 45*, 274–281. doi: 10.1111/jasp.12294

Murphy, R. O., Ackermann, K. A. & Handgraaf, M. J. J. (2011). Measuring Social Value Orientation. *Judgment and Decision Making, 6*(8), 771-781.

Nielsen, T. M., Hrivnak, G. A., & Shaw, M. (2009). Organizational Citizenship Behaviour and Performance. A Meta-Analysis of Group-Level Research. *Small Group Research, 40*(5), 555-577.10.1177/1046496409339630

Nowak, M. A., & Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature, 437/27*, 1291–1296.

Organ, D. W. (1997). Organizational Citizenship Behaviour: It's Construct Clean-Up Time. *Human Performance, 10*(2), 85–97.

Podsakoff, N. P., Whiting, S. W., Podsakoff, P. M., & Mishra, P. (2010). Effects of organizational citizenship behaviors on selection decisions in employment interviews. *Journal of Applied Psychology, 96 (2)*, 310–326.

Sloman, S. A., & Hagmayer, Y. (2006). The causal psycho-logic of choice. *Trends in Cognitive Sciences, 10*, 407– 412. doi:10.1016/j.tics.2006.07.00

Sober, E., & Wilson, D. (1999). *Unto Others: The Evolution of Unselfish Behavior*. Harvard University Press.

von Sydow, M. (2016). Towards a Pattern-Based Logic of Probability Judgements and Logical Inclusion "Fallacies". *Thinking & Reasoning, 22(3)*, 297-335. doi:10.1080/13546783.

von Sydow, M., & Braus, N. (2016). On the Tragedy of Personnel Evaluation. In A. Papafragou, et a. (Eds.), *Proceedings of the Thirty-Eighth Annual Conference of the Cognitive Science Society* (pp. 105-110). Austin, TX: Cognitive Science Society.

von Sydow, M., Braus, N. (2017). Altruist vs. Egoist Detection and Individual vs. Group Selection in Personnel Management. *Proceedings of the Thirty-Ninth Annual Conference of the Cognitive Science Society* (3466-3471). Austin, TX: Cognitive Science Society.

von Sydow, M., Braus, N., Hahn, U. (2017). Overcoming the Tragedy of Personnel Selection? *Proceedings of the Thirty-Ninth Annual Conference of the Cognitive Science Society* (pp. 3460-3465). Austin, TX: Cognitive Science Society.

von Sydow, M., Braus, N., & Hahn, U. (2018). On the Ignorance of Group-Level Effects – The Tragedy of Personnel Selection. *Journal of Experimental Psychology: Applied*. Advance online publication. doi: 10.1037/xap0000173

von Sydow, M., Hagmayer, Y., & Meder, B. (2016). Transitive reasoning distorts induction in causal chains. *Memory & Cognition, 44*(3), 469–487. doi:10.3758/s13421-015-0568-5

Waldmann, M. R. & Hagmayer, Y. (2001). Estimating causal strength: the role of structural knowledge and processing effort. *Cognition, 82*, 27-58.

Wilson, D. S., & Wilson, E. O. (2007). Rethinking the theoretical foundation of sociobiology. *Quarterly Review of Biology, 82*(4), 2007, 327–348. doi: 10.1086/52280

Acquiring Agglutinating and Fusional Languages Can Be Similarly Difficult: Evidence from an Adaptive Tracking Study

Svenja Wagner (s1581727@sms.ed.ac.uk),
Centre for Language Evolution, The University of Edinburgh,
3 Charles Street, Edinburgh, EH8 9AD, UK
Kenny Smith (Kenny.Smith@ed.ac.uk),
Centre for Language Evolution, The University of Edinburgh,
3 Charles Street, Edinburgh, EH8 9AD, UK
Jennifer Culbertson (Jennifer.Culbertson@ed.ac.uk)
Centre for Language Evolution, The University of Edinburgh,
3 Charles Street, Edinburgh, EH8 9AD, UK

Abstract

Research on the acquisition of morphology commonly predicts that agglutinating systems should be easier to learn than fusional systems. This is argued to be due to compositional transparency: the mapping between morphemes and meanings is one-to-one in agglutinating systems, but not in fusional systems. This is supported by findings in first and second language learning (Goldschneider & DeKeyser 2001, Slobin 1973), typology (Dressler 2003, Haspelmath & Michaelis 2017), and language evolution (Brighton 2002). We present findings from a series of artificial language learning experiments which complicate this picture. First, we show that when only two features (e.g., NOUN CLASS and NUMBER) are morphologically encoded, the learnability of fusional and agglutinating systems does not differ significantly. This finding holds when learners are given an additional cue to morpheme segmentation–which in principle should make the agglutinating system easier. However, the error patterns of the two groups provide some evidence that learners might have a bias for transparent structures. Our results suggest that the advantages of agglutinating over fusional systems may be overstated, particularly when a small number of features are encoded. Since agglutinating systems likely bear additional costs (e.g., segmentation, longer word length, and the online cost of mapping between morphemes and meanings), such systems do not guarantee learning ease under all circumstances.

Keywords: language acquisition; morphology; agglutinating; fusional; artificial language learning; transparency

Introduction

Classification of languages into morphological types is a commonly used parameter in language typology. Morphological type structures vary within and between languages, and they change over time. One key distinction is between fusional and agglutinating types. The distinction between these two is based on the ratio of morphemes to meaning, where a morpheme is defined as "the smallest meaning-bearing unit of language" (Kortmann, 2005). In fusional languages, morphemes typically express more than one meaning. For example, the German verb *spielst* ('you play') has the suffix –s*t*, which together expresses present tense, second person, and singular number. In comparison, morphemes in agglutinating languages typically only carry a single meaning. For example, the Turkish verb *konuşuyorsunuz* ('you speak') has three suffixes, *-yor*, *-sun*, and *-uz* individually expressing the same pieces of information (present tense, second person, plural number).

While both morphological types are well attested among the languages of the world, it has been proposed that fusional and agglutinating systems may differ in terms of learnability. In particular, it has been claimed that the more meanings a single morpheme carries, the less transparent it is, and therefore the more difficult it is to learn (e.g., Goldschneider & DeKeyser, 2001; Don, 2017; Haspelmath & Michaelis, 2017). For the purpose of this study, we use transparency to mean one-to-one correspondence between a form and its meaning (Don, 2017). Because agglutinating morphology is by definition more transparent, agglutinating systems should be easier to acquire, while fusional systems where a single morpheme encodes multiple meanings should be more difficult.

Support for this idea comes from research on both first and second language acquisition. In first language acquisition, a number of classic studies on morpheme order of acquisition in children have implicated transparency as part of the explanation for why certain English morphemes are acquired earlier than others (Brown, 1973; de Villiers & de Villiers, 1973; Dulay & Bert, 1974). For example, *–s* as in *plays* (which expresses 3rd person, singular, and present tense) is learned later than *–ing* (progressive). These studies build on more general claims relating transparency to ease of acquisition in children (e.g., Slobin, 1973)

More recent work has extended these findings to a number of other languages. For example, Sultana, Stokes, Klee, and Fletcher (2016) argue that the level of transparency of morphological forms predicts the order of acquisition in Bengali. Hengeveld and Leufkens (2018) point out that Turkish children generally master the agglutinating morphology of their language by the age of 3, whereas Dutch children have not yet acquired the fusional verbal system of their language at that age. This is in line with Dressler (2003), who reports earlier acquisition of morphology by children in Turkish than in English.

Second language acquisition research has echoed the role of transparency in morphological learning. In a meta-analysis

of 14 studies on L2 acquisition of English, Goldschneider and DeKeyser (2001) show that transparency correlates with earlier acquisition. For example, L2 learners, like children, acquire the English morpheme *–s* relatively late.

Finally, there is a clear relationship between agglutinating systems and the more general feature of compositionality. In compositional systems, complex signals are formed by combining meaning-bearing parts, with the meaning of the whole being a function of the meaning of the parts; this can be contrasted with holistic systems in which such re-combinable subparts do not exist, the relationship being between whole meanings and unanalyzable signals. A large body of research on the evolution of compositionality in language connects it to learnability (Brighton, 2002; Kirby, Cornish & Smith, 2008; Kirby, Tamariz, Cornish & Smith 2015, a.o.): compositional systems are simpler in that they have a shorter encoding length and are more compressible, making them simpler in a cognitively-relevant sense and therefore easier to learn; compositional systems also permit generalization to unseen meanings and signals. These same characteristics hold for agglutinating systems, suggesting that they too should be easier to learn.

To summarize, various lines of evidence suggest that agglutinating languages should be easier to learn than fusional languages. The inherent transparency and regularity of agglutinating forms, the higher frequency of a given morpheme in the system, and the possibility to generalize all point to a learnability advantage of these systems. However, in many cases, it is difficult to disentangle transparency from other features of the system. Most obviously, agglutinating systems often use *more* morphology overall, which could in principle also serve to obscure this advantage. However, Dressler (2003) argues that the systematic use of morphology in agglutinating languages relative to fusional ones may in fact serve to clue learners into its importance, triggering earlier learning. In this paper, we report a series of artificial language learning experiments which allows us to test the above claims by directly comparing agglutinating to fusional systems, while controlling for systematicity of morpheme use, and number of morphemes across conditions.

Experiment 1

We tested whether learners are faster at acquiring agglutinating systems compared to fusional systems by exposing participants to nouns encoding two binary features, one for NUMBER (singular/plural) and one for CLASS (animate/inanimate). Crucially, we held the number of morphemes to be learned constant across both conditions.

Methods

Participants. 80 participants were recruited on Amazon Mechanical Turk, all self-reported as English native speakers. They were paid $4 for their time. Participants were randomly assigned to one of the conditions described below (38 in the fusional and 42 in the agglutinating condition).

Materials. The language consisted of 96 nouns, referring to objects, and four suffixes, encoding NOUN CLASS (animate and inanimate) and NUMBER (singular and plural). Animate entities were always animals and inanimate entities were everyday objects such as household items and pieces of clothing. All stems were monosyllabic and adhered to English phonotactics. Morphemes used for both languages were identical: *-mu*, *-ka*, *-pi*, *-lo*. In the fusional condition, each of the four morphemes expressed one value for both features: animate+singular, animate+plural, inanimate+singular, inanimate+plural. For example, in Figure 1, *spur* is the noun stem, and *-ka* indicates animate+singular. In the agglutinating condition, the four morphemes each expressed a single value of NUMBER *or* CLASS. For example, in Figure 2, *foog* is the stem, *–ka* indicates inanimate, and *–mu* indicates plural. Note that the stem was directly followed by the CLASS morpheme, which was followed by the NUMBER morpheme. Mappings between morphemes and meanings were randomized across participants. Note that because we use the same set of morphemes in both languages, the words are longer in the agglutinating condition (by one syllable). This is a general characteristic of agglutinating languages, where words tend to be longer than in fusional languages.

Procedure. Participants were instructed that they would be learning part of a new language. On each trial (Figures 1, 2), participants saw an image and were given a choice of four words that could describe it. The four choices always represented the same stem with four possible grammatical combinations of affixes. Participants were instructed to click on the word that they thought correctly described the picture. Immediate feedback was given in every trial: the correct answer was highlighted with color, and the audio of the correct word was played aloud. The study consisted of 96 trials each of which displayed a unique picture. Therefore, no image or stem was ever repeated (and participants were not required to learn the mappings between images and stems). Each combination of grammatical meanings occurred as the correct choice 24 times in total. At the end, participants completed a short questionnaire.

Results. The design of our experiment aimed to compare performance across conditions over time. Since participants were necessarily guessing at the beginning, we expect performance in both conditions to be similar early on, but to potentially diverge over trials as they learned. Figure 3 shows mean accuracy across conditions by trial. As expected, participants generally improved over trials. However, performance appears to improve at a similar rate across conditions. Mean accuracy across all trials for the agglutinating condition was 0.65 (SD=0.24), for the fusional condition 0.60 (SD=0.24). To test whether the rate of improvement differed between conditions we fit a logistic mixed-effects regression model, predicting correct answer by

condition, trial (coded 0-95), and their interaction.[1] Condition was dummy-coded, with agglutinating as the reference level. The by-item intercept was removed because the model failed to converge. The model revealed a significant effect of trial number (b=0.04, SE=0.01, p<0.001), indicating that participants improved their accuracy over the course of the experiment, but no effect of condition (b=0.14, SE=0.18, p=0.42) and most importantly, no condition by trial number interaction (b=-0.01, SE=0.01, p=0.20). The latter would have indicated a difference in the rate of learning in one or the other condition, indicating a learnability advantage.

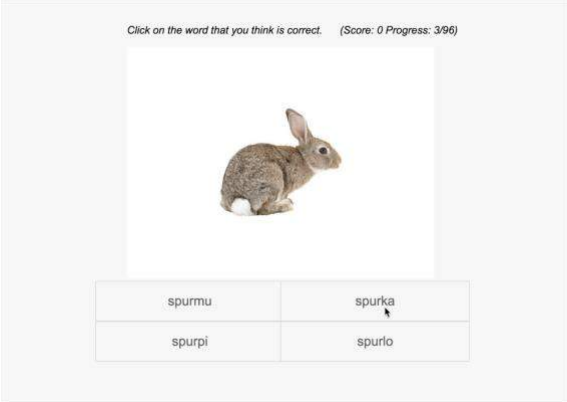


Figure 1: Example trial in Experiment 1, fusional condition. This trial shows an animate, singular object.

Figure 2: Example trial in Experiment 1, agglutinating condition. This trial shows an inanimate, plural object.

We conducted an exploratory analysis of participants' errors to investigate whether the highly similar overall performance masked a difference in error type between the two conditions. As described above, each of the four choices given in every trial constituted a different combination of grammatical meanings. Incorrect responses could either reflect the participant selecting a marker which was appropriate to the CLASS of the noun (e.g. selecting a morpheme marking animacy for an animate referent) but the wrong NUMBER (e.g. selecting a plural morpheme for a singular noun), selecting the wrong CLASS but the correct NUMBER, or selecting a morpheme which was incorrect for both CLASS and NUMBER. The rates for these three classes of error (correct CLASS only, correct NUMBER only, neither correct) are shown in Figure 4. The proportion of errors reflecting correct CLASS appears to be greater in the fusional condition. This impression is confirmed by a logistic mixed-effects regression model testing whether the proportion of CLASS correct only responses was significantly different between conditions. We ran the model predicting correct CLASS in the subset of the data with incorrect answers, including fixed effects of condition, trial number and their interaction. The by-item intercept was removed due to convergence errors. The model revealed a significant effect of trial (b=0.02, SE=0.01, p=0.003) no significant effect of condition (b=-0.02, SE=0.23, p=0.92), and a significant interaction between condition and trial (b=0.02, SE=0.01, p=0.04).

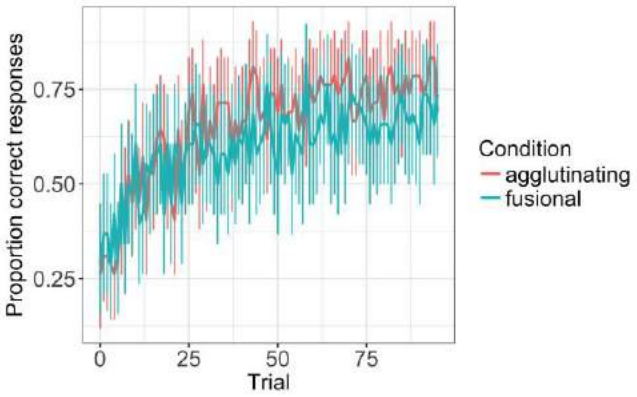


Figure 3: Mean accuracy by trial by condition in Experiment 1.

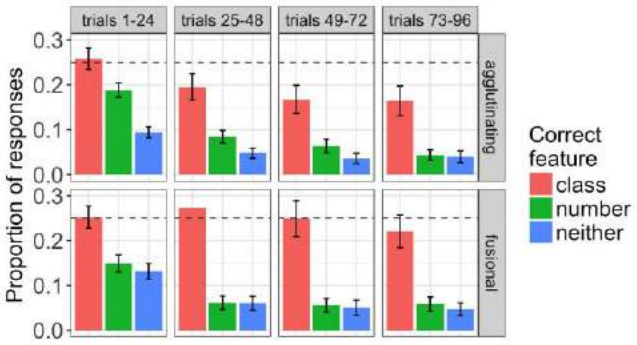


Figure 4: Classification of incorrect choices by correct feature (CLASS correct NUMBER wrong, NUMBER correct CLASS wrong, or neither correct) by trial block by condition

[1] All models were run using the package lme4 in R (Bates 2010). Unless otherwise noted, models included random by-participant and by-item (picture) intercepts, and by-participants slopes for the effect of trial.

in Experiment 1. Trial number is binned for readability. Note that the y-axis range does not display the full range.

This suggests that participants in the fusional condition and agglutinating condition diverged over time in their tendency to choose an answer in which only the CLASS morpheme was correct–while such errors decline in the agglutinating condition, they remain at a fairly constant level in the fusional condition. One possibility is that this reflects a bias for transparency: participants in the fusional condition may have been searching for a single feature with four values, rather than two binary features. While NUMBER is unambiguously binary (one vs. two), the stimuli could in principle encode more fine-grained distinctions of CLASS. In the post-test questionnaire, some participants indeed reported such a strategy, for instance, distinguishing land vs. sea animals and household items vs. clothing. They subsequently tried to map each of these four CLASS values onto one morpheme, ignoring NUMBER altogether. However, this analysis was performed post-hoc since the given distribution of answer types was unexpected. We therefore replicated the experiment.

Experiment 2

Methods

Participants. 100 participants were recruited on Amazon Mechanical Turk, all self-reported as English native speakers. They were paid $4 for their time. Participants were randomly assigned to one of the two conditions (48 in the fusional and 52 in the agglutinating condition).

Materials. Stimuli were identical to those of the previous experiment.

Procedure. The procedure was identical to Experiment 1.

Results. Figure 5 shows mean accuracy by trial across conditions. As in Experiment 1, participants generally improved from the start to the end as expected, and overall performance appears to be similar across conditions (agglutinating M=0.64, SD=0.26; fusional M=0.68, SD=0.25). We ran a model predicting correct answer by condition, trial (coded 0-95), and their interaction. Condition was dummy-coded, with agglutinating as the reference level. The model revealed a significant effect of trial number (b=0.04, SE=0.01, p<0.001), indicating that participants improved their accuracy over the course of the experiment, but no effect of condition (b=-0.10, SE=0.17, p=0.57) and no condition by trial number interaction (b=0.01, SE=0.01, p=0.25). Again, the latter would have indicated a more rapid improvement in one or the other condition, and thus a learnability advantage.

We repeated our analysis of error types across conditions (Figure 6). In this case, the model revealed a significant effect of trial (b=0.02, SE=0.01, p<0.001), no significant effect of condition (b=-0.02, SE=0.20, p=0.93), and no significant interaction between condition and trial (b=-0.004, SE=0.01, p=0.63). This suggests that the apparent difference in CLASS-based errors across conditions seen in Experiment 1 may have been spurious.

The strong expectation from previous research was that, all things equal, an agglutinating system should be easier to learn than a fusional system. This advantage was not borne out in Experiments 1 and 2. However, we are exploring the early stages of learning these systems, and thus one possibility is that participants in the agglutinating condition were not segmenting the morphemes–i.e., they may have been treating the string of two morphemes as a single morpheme, encoding both NUMBER and CLASS. If so, then we would not expect any difference between conditions. Indeed, the post-test questionnaire reveals that at least some participants failed to segment the stems and morphemes. In Experiment 3 we test whether an advantage for the agglutinating system is revealed if we provide a visual cue to aid segmentation.

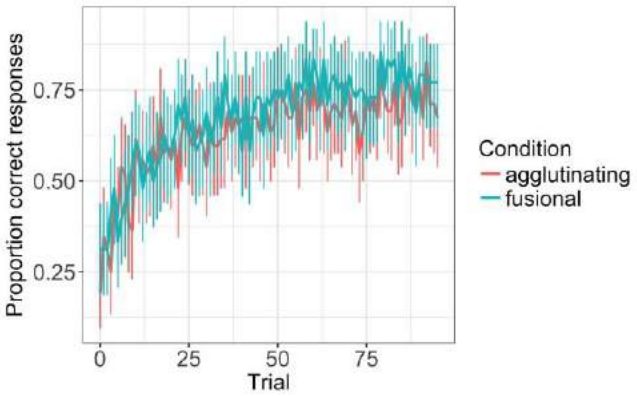


Figure 5: Mean accuracy by trial by condition in Experiment 2.

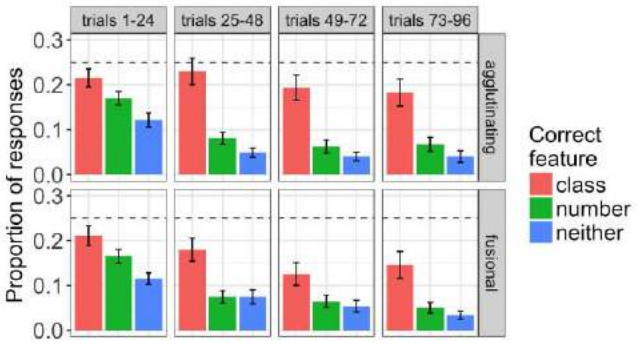


Figure 6: Classification of incorrect choices by correct feature (CLASS correct NUMBER wrong, NUMBER correct CLASS wrong, or neither correct) by trial block by condition in Experiment 2. Trial number is binned for readability. Note that the y-axis range does not display the full range.

Experiment 3

Methods

Participants. 100 participants were recruited on Amazon Mechanical Turk, all self-reported as English native

speakers. They were paid $4 for their time. Participants were randomly assigned to one of the conditions (51 in the fusional and 49 in the agglutinating condition).

Materials. The language was identical to Experiments 1 and 2, however, a visual cue to the segmentation of words and morphemes was provided. In each trial, morphemes were highlighted with color (Figure 7). In the agglutinating condition, the CLASS morpheme was highlighted in one color and the NUMBER morpheme in another. Participants were randomly assigned to see either CLASS in orange and NUMBER in blue, or vice versa. In the fusional condition, all four morphemes were randomly assigned a single color so that a participant would either see all morphemes across all 96 trials in orange or all morphemes in blue.

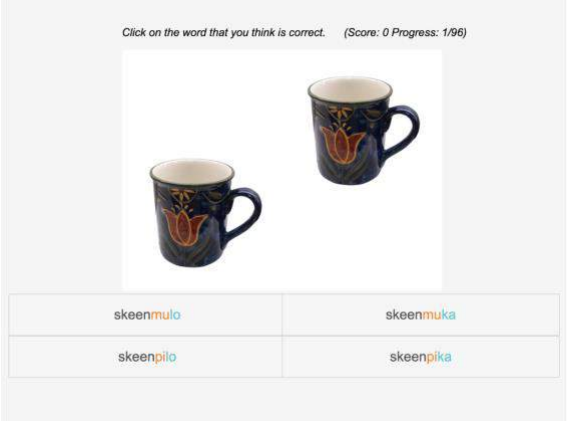


Figure 7: Example trial in Experiment 3, agglutinating condition. This trial shows an inanimate, plural object.

Procedure. The procedure was identical to Experiments 1 and 2.

Results. Figure 8 shows mean accuracy in correct answers across conditions. As in Experiments 1 and 2, participants generally improved from the start to the end as expected, and overall performance appears to be similar across conditions (agglutinating M=0.59, SD=0.25; fusional M=0.59, SD=0.23). We ran a model predicting correct answer by condition, trial (coded 0-95), and their interaction. Condition was dummy-coded, with agglutinating as the reference level. The model revealed a significant effect of trial number (b=0.03, SE=0.005, p<0.001), indicating that participants improved their accuracy over the course of the experiment, but no effect of condition (b=0.23, SE=0.18, p=0.21) and no condition by trial number interaction (b=-0.003, SE=0.01, p=0.64). The latter interaction would have indicated a more rapid improvement in one or the other condition indicating a learnability advantage for one type.

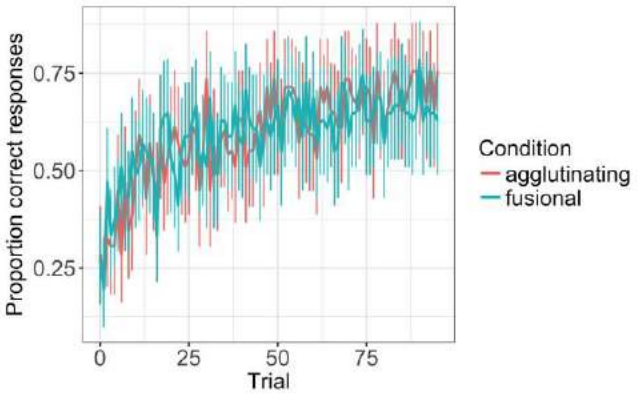


Figure 8: Mean accuracy by trial by condition in Experiment 3.

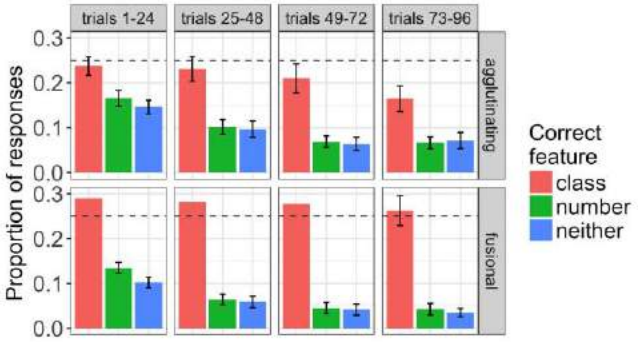


Figure 9: Classification of incorrect choices by correct feature (CLASS correct NUMBER wrong, NUMBER correct CLASS wrong, or neither correct) by trial block by condition in Experiment 3. Trial number is binned for readability. Note that the y-axis range does not display the full range.

We repeated our analysis of error types across conditions (Figure 9). In this case, the model revealed a significant effect of trial (b=0.01, SE=0.004, p=0.03), no significant effect of condition (b=0.29, SE=0.20, p=0.14), and a significant interaction between condition and trial (b=0.02, SE=0.01, p=0.01). Thus, as in Experiment 1, but not 2, participants in the fusional condition were significantly more likely to choose an answer in which only CLASS was correct.

Discussion

It has been claimed that agglutinating systems should be easier to learn because of their inherent transparency: there is a one-to-one mapping between morphemes and meanings in these systems. Here, we directly contrasted a fusional with an agglutinating system, holding the number of morphemes to be learnt constant. We found no clear learnability advantage for agglutinating systems across three experiments. In two of our three experiments, we found a difference in the error patterns between conditions: participants in the fusional condition were more likely to make errors involving NUMBER than CLASS. This error pattern could reflect a bias for

transparency. Participants may have been inferring a single four-way distinction, which was possible for CLASS but not NUMBER. However, this effect was not strong enough to result in an overall advantage for the agglutinating system, and it failed to replicate in one experiment. Our results are surprising, given the general and wide-ranging claims in the literature concerning relative ease of learning of agglutinating systems. It may be that an apparent advantage for agglutinating systems reported in the literature is due to confounding differences between the systems in question. However, below we discuss alternative explanations for our failure to uncover the advantage here.

One possibility is that the paradigms we are testing are too small to result in a discernable difference in learnability. It has been noted that compositionality (and therefore transparency) is increasingly beneficial the larger the paradigm is. This was explored computationally by Brighton (2002), who shows that a compositional system for expressing a few features is hardly more learnable than a holistic system covering the same semantic space; the learnability advantage of compositional systems is maximized when each meaning is composed of many binary features. However, the paradigms we used were intentionally small, consisting of just two features, allowing us to test for a learning advantage arising purely from transparency in the absence of benefits associated with increased generalizability. It is therefore possible that the advantages of agglutinating systems derive purely from the fact that they facilitate more rapid generalization, in which case we would not expect to see that advantage in our paradigm.

Another possibility is that agglutinating systems bear additional costs which have not been much discussed in the literature. One such cost is clearly segmentation. Learners can only profit from compositionality if they are able to segment a word into morphemes, but this process might be costly. To eliminate this issue, we used color highlighting in Experiment 3. However, the null effect of condition on overall accuracy was replicated, suggesting that segmentation alone will not suffice to explain why learners did not have an easier time acquiring the agglutinating system.

Compositional structure typically means more material to process for each word: for example, as is typical cross-linguistically, words were longer in our agglutinating condition than in our fusional condition. It is therefore possible that word length (perhaps combined with segmentation cost) has a detrimental effect on the learnability of the agglutinating system.

Finally, seeing a word and its referent (here, an image) in an agglutinating system does not illustrate the meaning of each individual morpheme. Learners of compositional systems need a set of examples to compare and pin down which morpheme expresses which meaning; learning an agglutinating system therefore potentially poses a cross-situational learning problem (similar to that explored by e.g. Yu & Smith, 2007, where multiple words are simultaneously mapped to multiple referents and the precise word-to-referent mapping can only be disambiguated across trials) that is less pronounced for fusional systems. It is possible that this cost, which is often overlooked, together with the length of words and the small size of the paradigm, did not provide a condition under which an agglutinating system becomes easier to learn than a fusional system.

Conclusion

In this paper, we investigated the frequently-made claim that agglutinating systems are easier to learn than fusional systems due to their inherent transparency. Results from three artificial learning experiments did not show the predicted effect. This held even when a visual cue to segmentation was added to help participants discover morpheme boundaries in the agglutinating condition. While some weak evidence for a possible bias for transparent structures was found in participants' error patterns, this did not lead to an overall difference in learning. We argue that this may be due to the small size of the paradigms, which narrow the extent of the benefit for transparency. Some natural language paradigms are of course larger, and these might provide conditions under which the costs of agglutinating systems outweigh those of fusional systems. Setting aside paradigm size, we also argue that agglutinating systems may present additional costs in processing which have not yet been fully explored.

References

Bates, D. M. (2010). lme4: Mixed-effects modeling with R.

Brighton, H. (2002). Compositional syntax from cultural transmission. *Artificial life*, *8*(1), 25-54.

Brown, R. (1973). *A first language: The early stages*. Harvard U. Press.

De Villiers, J. G., & De Villiers, P. A. (1973). A cross-sectional study of the acquisition of grammatical morphemes in child speech. *Journal of Psycholinguistic Research*, *2*(3), 267-278.

Don, J. (2017). What causes languages to be transparent? *Language Sciences*, *60*, 133–143.

Dressler, W. (2003). Morphological typology and first language acquisition: Some mutual challenges. In *Mediterranean Morphology Meetings* (Vol. 4, pp. 7-20).

Dulay, H. C., & Burt, M. K. (1974). NATURAL SEQUENCES IN CHILD SECOND LANGUAGE ACQUISITION 1. *Language Learning*, *24*(1), 37-53.

Goldschneider, J. M., & DeKeyser, R. M. (2001). Explaining the "Natural Order of L2 Morpheme Acquisition" in English: A Meta-analysis of Multiple Determinants. *Language Learning*, *51*(1), 50.

Haspelmath, M., & Michaelis, S. M. (2017). Analytic and synthetic: Typological change in varieties of European languages. In I. Buchstaller & B. Siebenhaar (Eds.), *Studies in Language Variation* (Vol. 19, pp. 3–22). Amsterdam: John Benjamins Publishing Company.

Hengeveld, K., & Leufkens, S. (2018). Transparent and non-

transparent languages. *Folia Linguistica*, *52*(1), 139–175.
Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, *105*(31), 10681–10686.
Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, *141*, 87–102.
Kortmann, B (2005). *English Linguistics: Essentials*. Berlin: Cornelsen.
Slobin, D. (1973). Cognitive prerequisites for the development. *Charles Ferguson and Dan Slobin, Studies in Child Language Development*, 175-208.
Sultana, A., Stokes, S., Klee, T., & Fletcher, P. (2016). Morphosyntactic development of Bangla-speaking preschool children. *First Language*, *36*(6), 637–657.
Yu, C., & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, *18*, 414–420.

Achievement Goals and Mental Arithmetic: The Role of Distributed Cognition

Anna-Stiina Wallinheimo
School of Psychology, University of Surrey
Guildford, Surrey GU2 7XH, United Kingdom

Adrian Banks
School of Psychology, University of Surrey
Guildford, Surrey GU2 7XH, United Kingdom

Harriet Tenenbaum
School of Psychology, University of Surrey
Guildford, Surrey GU2 7XH, United Kingdom

Abstract

The purpose of these studies was to investigate the role of distributed cognition in defusing the impact of evaluative pressure caused by performance-approach goals on mental arithmetic performance. Performance-approach goals can generate worrying thoughts that can deplete working memory resources. However, some of these working memory limitations can be compensated by off-loading the internal cognitive process to the external environment. We tested this prediction in two experiments. Participants carried out modular arithmetic tasks in a performance-approach goal or mastery-approach goal condition crossed with interactivity or no interactivity. Performance-approach goal manipulation hampered cognitive performance (accuracies), (Experiment 1). However, these negative effects were defused with the help of interactivity (Experiment 2). Interestingly, the mastery-focused individuals had a performance drop in the interactive condition (Experiment 1 and Experiment 2). Finally, experiment 2 reported higher maths anxiety levels for the performance-focused individuals. Reasons for the findings and future implications will be discussed.

Keywords: achievement goals; working memory, mental arithmetic; distributed cognition; interactivity

Introduction

Achievement goals are said to reflect the aim of an individual's achievement pursuits. They are frameworks that can help to understand how individuals react to various achievement situations (Poortvliet & Darnon, 2010). There is a wealth of research on achievement goals and their effects on academic performance. But much less is known about the cognitive processes of these goals and particularly the effects on the working memory and whether distributed cognition could be used to reduce the negative effects of performance goals on academic performance.

Achievement goals

Individuals pursuing performance-approach goals are good at knowing the material that is essential for the task in hand (Elliott, Shell, Henry, & Maier, 2005). They listen to the cues about the future assignments and adjust learning based on these cues. Students perform better when they focus on topics that the teacher deems important and that are tested (Broekkamp, Hout-Wolters, & Van Hout-Wolters, 2007). Performance-focused students concentrate on memorizing rather than elaboration and knowledge construction (Entwistle, 1988). This can lead to surface learning and rote learning (Harackiewicz & Linnenbrink, 2005). Mastery-focused students are freer to pursue their own agenda guided by their own personal interests and curiosity of the current topic. Hence, mastery-approach goals predict the use of adaptive cognitive strategies that lead to deeper processing. This kind of approach might benefit the students in the long run as it promotes deeper learning but might not help in gaining the highest grades as it is based on personal interests rather than the areas that might be tested. When people pursue performance-approach goals, their focus is on the outcome of the task and therefore the individuals might not be fully engaged with the process. On the contrary, mastery-focused individuals focus on the process rather than the activity of outperforming others. Mastery-focused individuals focus on learning and their personal improvement, and therefore have a focus on the task that allows them to explore both intrinsic and utility value (Hulleman, Durik, Schweigert, & Harackiewicz, 2008).

The Effects of Performance-approach Goals on Working Memory

The pressure of outperforming others can generate concerns that deplete available working memory resources. (Crouzevialle & Butera, 2013). When high working memory load tasks were utilized, there was a performance drop in the high evaluative pressure condition (Beilock, Holt, Kulp, & Carr, 2004). Additionally, Avery and Smillie (2013) examined the influence of achievement goal pursuits on working memory capacity when varying levels of executive load were used. Under the high executive load, there was poorer working memory processing during the performance-approach goal than when mastery-approach goal or no-goal control were used (Avery & Smillie, 2013).

Distributed Cognition

Some of the possible working memory limitations can be compensated by off-loading the cognitive process to the

external environment (e.g., by using pen and paper), (Neth & Payne, 2011). According to Kirsh (2010), cognitive processes go to wherever it is easier to perform them. It might be easier to understand a particular sentence by drawing a picture of it rather than just thinking internally. Therefore, with the help of drawing the overall cognitive cost of sense making can be reduced (Kirsh, 2010). Kirsh (1995) conducted a simple coin counting experiment where he observed that complementary strategies could enhance performance (Kirsh, 1995). Neth and Payne (2011) asked participants to add coins on a computer screen in move versus look conditions. Accuracy increased with interactivity but not the speed. Both accuracy and speed were increased with the help of using hands (in the pointing condition) when counting arrays of items (simple arithmetic task), (Carlson, Avraamides, Cary, & Strasberg, 2007). Interactivity enhanced performance, and in particular, accuracy and efficiency for longer sums involving 11 single-digit numbers (Vallée-Tourangeau, 2013). Additionally, interactivity allows the agent to extend their working memory resources when there is a need for it. Dyslexic children (aged between 9 – 11 years) benefited the most form rearranging the letter tiles (interactive condition) in a word production task. By reshaping the physical presentation of the letters, their less efficient working memory capabilities could be compensated. The control group (typically developing children) did not benefit from externalizing the process. In fact, their performance was poorer (with easy set of letters) when they manipulated the letter tiles to produce words (Webb & Vallée-Tourangeau, 2009).

Maths anxiety

Maths anxiety is a multidimensional construct, and a full list of the causes is still undetermined. Maths anxiety can be defined as a feeling of apprehension and tension in a mathematical setting which can also affect overall mathematics performance. The highly maths-anxious individuals avoid mathematics as a topic and choose fewer elective mathematics courses in secondary school and university (Ashcraft, 2002). The maths-anxious individual is pre-occupied with the maths fears and the overall capacity of working memory gets affected. This pre-occupation functions as a secondary task that is heavily working memory resource demanding (Ashcraft & Krause, 2007). Maths anxiety causes a transitory disruption of working memory. The lower working memory capacity of high maths-anxious individuals is partially responsible for the maths performance decrements. This reduced working memory capacity is on-line effect that disrupts information processing in maths tasks (Ashcraft & Kirk, 2001). Finally, maths anxiety is higher among women than men (Ashcraft & Faust, 1994; Luttenberger, Wimmer, & Paechter, 2018). To increase the chances of selecting maths-anxious individuals, we included women only in the sample.

Experiment 1

The aim of the current study was to understand how mastery-approach goal and performance-approach goal engage working memory resources and whether interactivity could be used to reduce any of the negative effects of performance-approach goals on maths performance. If the working memory is loaded due to outcome related worry then there is additional taxation on the working memory (Crouzevialle & Butera, 2013). And together with the horizontally presented maths problems (modular arithmetic tasks) there can be maths performance decrements when in the performance-approach goal condition (Beilock, 2008). We reasoned that, if worries of outperforming others lead to poor maths performance, then giving students the opportunity to externalize the internal cognitive process would enhance this performance.

Method

Participants

Forty-one female undergraduate psychology students (M = 21.88 SD = 3.90) participated in this study for exchange of credits. After consenting to participate in the study, subjects were randomly assigned to one of the experimental conditions (performance-approach goal or mastery-approach goal crossed with interactivity or no interactivity). The participants were tested individually (15 minutes) in a psychology lab.

Material and Measures

Arithmetic task There were two blocks of 24 modular arithmetic tasks that relied heavily on working memory resources, adapted from Beilock and Carr (2005). The purpose of the tasks is to judge the validity of maths problems like $61 \equiv 18 \pmod 4$. The middle number is subtracted from the first number (i.e. 61-18) and then the difference is divided by 4. If the answer is a whole number the maths problem is true (Beilock & Carr, 2005). Modular arithmetic tasks as laboratory tasks are advantageous as most students have not seen them before and therefore previous task experience is controlled.

High-demand problems (e.g. $42 \equiv 27 \pmod 3$) requiring a double-digit subtraction operation were used as they required borrowing, resulting in using more working memory resources. Half of the maths problems required a true response by the participant. The order of the questions was randomized and each question was asked only once. The original questions used by Beilock and Carr (2005) were a mixture of high demand problems (two-digit numbers requiring borrowing) and low-load questions (single-digit numbers, without borrowing). The current study used the high-demand problems only because of limited benefits of using interactivity with low-demand tasks.

The modular arithmetic tasks were presented in a horizontal format as opposed to a vertical format (also called column subtraction). The horizontal presentation of the maths problems is more reliant on phonological resources (the verbal resources) because individuals maintain the required problem steps in their memory verbally (DeStefano & LeFevre, 2004). The possible worries of performing better than others places much heavier demands on working memory (phonological loop, in particular).

Experimental manipulations Participants were informed after completing the baseline block of modular arithmetic tasks (24) that they required to complete a second block (24) of modular arithmetic tasks, and this time their performance would be recorded. The participants in the performance-approach goal condition read the following instructions before starting the task that were aimed at activating performance-approach goals (Darnon, Harackiewicz, Butera, Mugny, & Quiamzade, 2007):

"During the recorded part of the task, the experimenters will assess your performance. It is important for you to be proficient, to perform well and obtain a high score, in order to demonstrate your competence. You should know that a lot of students will do this task. You are asked to keep in mind that you should try to distinguish yourself positively, that is, to perform better than majority of students. In other words, what we ask you here is to show your competencies, your abilities."

The participants in the mastery-approach goal condition read instructions that were designed to activate mastery-approach goals. There is no social comparison being made and the instructions are aimed to create task interest, use for everyday life, and there is no mention about scores or task performance (Crouzevialle, Smeding, & Butera, 2015).

"In previous research, we have observed that practice of the arithmetic task you are solving right now benefits to cognitive functioning and leads to a progressive improvement of mental processes. Hence, this task solving can proved to be beneficial on the long-term. It is however necessary that you focus your attention on calculation mastery, so as to quickly and accurately solve each problem, in order to experience these benefits. Try to master this task as much as you can; keep in mind its practice can be beneficial to you."

Interactivity The participants in the interactive condition were allowed to use pen and paper. The participants in the non-interactive condition were not allowed to use any external artefacts to complete the task.

Procedure

After consenting to participate in the study, the participants were randomly assigned to one of the experimental conditions. There was a short training session before starting the first block. The first block of questions (24) functioned as a base-line. The participants were told that it was a training block, and that their performance was not recorded to avoid any achievement goal activation. The second block of questions was done under the experimental conditions. The participants were told that their performance was recorded this time.

Results

Accuracy

Before the actual statistical analysis was conducted, it was concluded that there were no group differences between the participants in the mastery-approach goal condition and performance-approach goal condition on the baseline modular arithmetic performance (block 1), $F(1, 37) = .08$, $p = .78$, $\eta_p^2 = .002$, confirming that the groups did not differ in their ability to complete the modular arithmetic tasks. Our main performance measure was accuracy of the high working memory load tasks. Accuracy difference score was calculated by subtracting the modular arithmetic performance of block 1 from block 2. Furthermore, a difference score in latencies was used as a covariate in order to avoid any speed-accuracy trade-off of the participants. A 2 (instruction: performance-approach goal or mastery-approach goal) x 2 (level of interactivity: interactivity or control) between-groups analysis of covariance (ANCOVA) was conducted. The covariate, difference score in latencies, was significantly related to the modular arithmetic accuracy, $F(1, 36) = 5.76$, $p = .02$, $\eta_p^2 = .14$. There was a significant two-way interaction of interactivity (interactivity or control) and instruction (performance-approach goal or mastery-approach goal), $F(1, 36) = 4.39$, $p = .043$, $\eta_p^2 = .11$. As expected, performance-focused participants had lower maths performance in the non-interactive condition ($M = -7.43$, $SE = 2.50$) than the mastery-approach goal individuals ($M = 5.44$, $SE = 2.38$), (Figure 1). The post hoc tests confirmed this finding, $F(1, 18) = 11.1$, $p = .004$, $\eta_p^2 = .38$. However, mastery-focused individuals had a performance drop in the interactive condition ($M = -2.37$, $SE = 2.50$) compared with their performance in the non-interactive condition ($M = 5.44$, $SE = 2.38$), (Figure 1). Post hoc tests confirmed this finding, $F(1, 18) = 4.90$, $p = .04$, $\eta_p^2 = .21$. Additionally, there was a main effect of instruction (mastery-approach goal or performance-approach goal), $F(1, 36) = 9.72$, $p = .004$, $\eta_p^2 = .21$. The modular arithmetic performance of the mastery-approach goal participants was enhanced from block 1 to block 2 ($M = 1.53$, $SE = 1.72$). As predicted, there was reduced modular arithmetic performance of the performance-approach goal participants ($M = -6.16$, $SE = 1.76$). Finally, interactivity did not improve modular arithmetic performance, $F(1, 36) = 1.13$, $p = .30$, $\eta_p^2 = .03$.

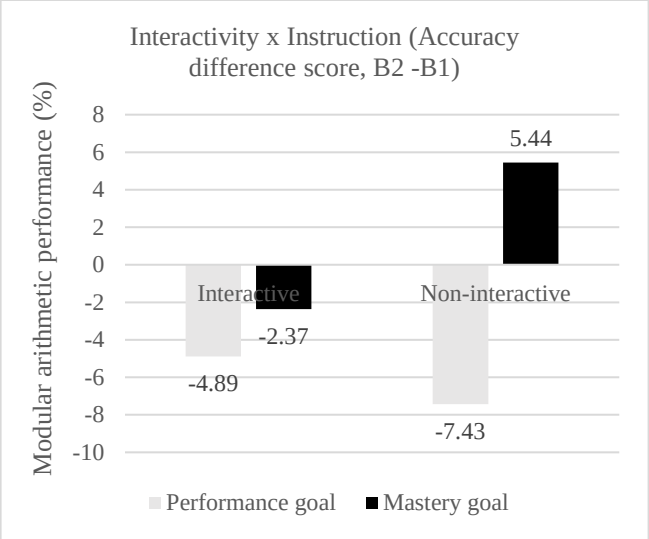


Figure 1: Mean difference in modular arithmetic performance (%) as a function of experimental condition (Experiment 1).

Discussion

We found that when the performance-approach goal was made salient, there was a drop in the mental arithmetic performance compared with the mastery-approach goal participants. Additionally, an interesting finding was made in relation to mastery-focused individuals, their modular arithmetic performance was reduced when the participants were allowed to interact with external resources (with the use of pen and paper).

Experiment 2

It was clear from Experiment 1 that the mental arithmetic performance of the performance-focused participants was depleted compared to the mastery-goal individuals. We therefore argued that it would be the performance-focused individuals that would show higher levels of maths anxiety due to the worrying thoughts of outperforming others, in a mathematical domain. Experiment 2 therefore measures maths anxiety of the participants both before the experiment (trait maths anxiety) and after (state maths anxiety). If maths anxiety is elevated when performance-approach goal is made salient then there should be more benefits of externalizing the internal cognitive process to the outside world (interactivity) for the performance-focused individual.

Method

Participants

Seventy-eight female undergraduate psychology students (*M* = 19.12, *SD* = 1.60) participated in this study for exchange of credits. This study only included females due to their higher levels of maths anxiety. After consenting to participate in the study, the participants were randomly assigned to one of the experimental conditions. The participants were tested individually in a psychology lab (40 minutes).

Material and Measures

Mathematics anxiety (trait) Maths anxiety was measured with the 23-item Mathematics Anxiety Scale (MAS-UK) by Hunt, Clark-Carter, and Sheffield (2011). The test comprises statements that relate to everyday situations that have a mathematics component (e.g., adding up a pile of change). The participants are expected to respond by confirming the level of anxiety that they feel on a 5-point Likert-type scale.
Basic arithmetic skills Basic arithmetic skill (BAS) was measured with the help of 45 simple expressions in a 60-second period (e.g. 10-5).
Computation span (Working memory) Working memory capacity was measured with the help of a computation-based span test. The participants were asked to read a simple arithmetic expression (e.g. 5 + **2** = ?, 9 – **6** = ?) and announce their answer aloud to the researcher (7, 3). Additionally, the participants were asked to remember the second number of each equation to be recalled later (**2**, **6**). The sequences of the simple arithmetic tasks varied from 1 to 7 tasks. The computation span task requires both on-line processing for the problem solution which is simultaneous with storage and maintenance of information in working memory for serial recall. People with maths anxiety have smaller working memory spans. This smaller span can lead to increased reaction times and errors when mental mathematics is completed at the same time as a memory load task (Ashcraft & Kirk, 2001)**.**
Arithmetic task The mental arithmetic task consisted of modular arithmetic tasks (two blocks of 24 questions) that relied heavily on working memory resources, adapted from Beilock and Carr (2005). The arithmetic task was identical to Experiment 1.
Mathematics anxiety (state) Maths anxiety was measured with the 23-item Mathematics Anxiety Scale (MAS-UK) by Hunt, Clark-Carter, and Sheffield (2011). This test was the same as the trait measurement used earlier during the experiment but this time referring to present time (now).
Experimental manipulations Participants were informed that after completing the baseline block of modular arithmetic tasks (24) that they required to complete a second block (24) of modular arithmetic tasks, and this time their performance would be recorded. The actual priming instructions were identical with the experiment 1.
Interactivity As before the participants in the interactive condition were allowed to use pen and paper to come to the solution. The participants in the non-interactive condition were not allowed to use any external artefacts.

Procedure

After consenting to participate in the current study, the participants started with the trait maths anxiety questionnaire. This was then followed by the timed basic arithmetic skills

test. Before commencing with the modular arithmetic tasks in primed conditions, computation span (working memory capacity) was assessed. There was a short training session (2 questions) before starting the first block of the modular arithmetic problems (24). Only high-demand problems requiring a double-digit subtraction operation (e.g. 42 ≡ 27 (mod 3)) were used as they required more of the working memory resources compared to low-demand problems (single-digit operation, and no carrying required) (Ashcraft & Kirk, 2001). After the baseline the participants were primed to either performance-approach goal condition or mastery-approach goal condition. If in the interactive condition, the use of pen and paper was allowed. After completing the second block of arithmetic tasks in primed conditions, the participants were asked to complete the state maths anxiety questionnaire.

Results

Accuracy

There were no group differences between the participants in the two achievement goal groups on the baseline modular arithmetic performance (block 1), $F(1, 74) = 1.77$, $p = .19$, $\eta_p^2 = .02$, confirming that the groups did not differ in their ability to complete the modular arithmetic tasks. Additionally, there were no group differences in working memory capacity, $F(1, 74) = 1.17$, $p = .28$, $\eta_p^2 = .02$, confirming the fact that the two achievement goal groups did not differ in their level of working memory capacity as a baseline measure. To test the hypotheses, accuracy difference in percentage score (block 2 - block 1) of the modular arithmetic tasks was examined. A 2 (level of interactivity: interactivity or control) x 2 (instruction: performance-approach goal or mastery-approach goal) between-groups analysis of covariance (ANCOVA) was conducted. There was a significant two-way interaction of interactivity (interactivity or control) and instruction (mastery-approach goal or performance-approach goal) after controlling for a difference score in latencies, $F(1, 73) = 10.04$, $p = .002$, $\eta_p^2 = .12$. The performance-focused participants benefited from the use of interactivity ($M = 3.70$, $SE = 1.80$) unlike the mastery-focused individuals whose performance was depleted with interactivity ($M = -3.30$, $SE = 1.90$), (Figure 2). The post-hoc test confirmed this finding, $F(1, 36) = 10.67$, $p = .002$, $\eta_p^2 = .23$. The accuracy of the mastery-approach goal participants was reduced in the interactive condition ($M = -3.30$, $SE = 1.90$) compared with the non-interactive condition ($M = 5.40$, $SE = 1.80$) (Figure 2). This finding was confirmed with a post-hoc test, $F(1, 36) = 6.82$, $p = .01$, $\eta_p^2 = .16$. The two main effects (interactivity or instruction) did not reach statistical significance. There was no significant difference in accuracy between the participants in the interactive condition and the participants in the non-interactive condition, $F(1, 73) = 2.35$, $p = .13$, $\eta_p^2 = .03$. Additionally, the main effect of instruction (mastery-approach goal or performance-approach goal) did not reach statistical significance ($F < 1$).

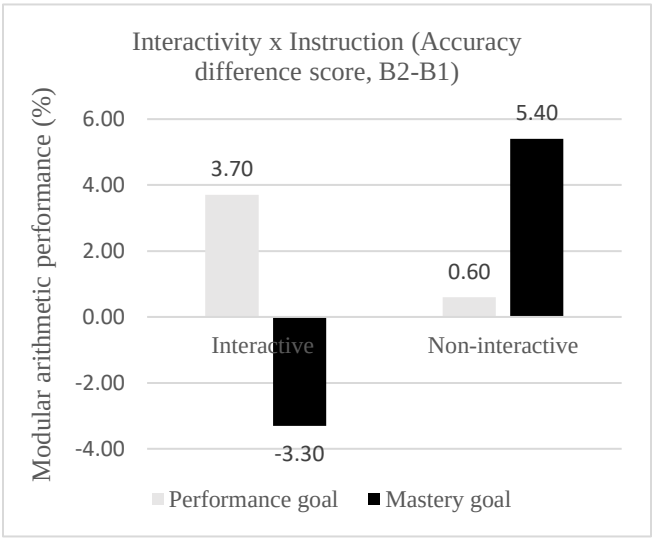


Figure 2: Mean difference in modular arithmetic performance (%) as a function of experimental condition (Experiment 2).

Maths anxiety (state)

A two-way between groups analysis of covariance (ANCOVA) was conducted to compare the effects of interactivity on two levels of instructions that were given to the participants (mastery-approach goals or performance-approach goals) when completing the modular arithmetic tasks. Participants' scores on maths anxiety (trait) were used as the covariate in this analysis. After adjusting for pre-existing maths anxiety levels (trait maths anxiety), there was a significant main effect of instruction (mastery-approach goal or performance-approach goal) on state maths anxiety, $F(1, 73) = 6.07$, $p = .02$, $\eta_p^2 = .08$. The performance-focused individuals showed higher levels of maths anxiety after completing the experiment in primed conditions ($M = 56.0$, $SE = 1.98$) than the mastery goal participants ($M = 49.1$, $SE = 1.98$) confirming the hypothesis set in the beginning. The main effect of interactivity did not reach statistical significance, $F < 1$, as did not the two-way interaction of instruction and interactivity either.

General discussion

The purpose of this study was to see whether the adverse effects of performance-approach goals on mental arithmetic performance could be alleviated with the use of distributed cognition. This investigation reported a performance drop in mental arithmetic performance for the performance-focused individuals in the non-interactive condition compared with the mastery-approach goal individuals (Experiment 1). However, interactivity mitigated the negative effects of performance-approach goal instructions on maths performance (Experiment 2). Additionally, we found that performance-focused participants felt higher levels of state maths anxiety, after completing the maths tasks (Experiment

2). Clearly, the priming instructions of performance-approach goals had strong carry-on effects on maths anxiety as they were still felt after completing the mental arithmetic tasks. However, it was evident that there were no carry-on effects of interactivity at the end of the experiment. An interesting finding was made as there was reduced maths performance for the mastery-focused individual in the interactive condition (Experiment 1 and Experiment 2). It was clear that distributed cognition hindered maths performance for the mastery-focused individual who was less maths anxious after the experiment but allowed the more maths-anxious individual (the participants in the performance-approach goal) to improve mental arithmetic performance. Similar findings have been made by Webb and Vallée-Tourangeau (2009) who concluded that when the agent had the required cognitive resources to complete the word production task, interactivity hampered the performance. If working memory resources are not compromised from increased maths anxiety levels (like in the mastery-approach goal environment), then there are little benefits of externalising the internal cognitive process to the outside world.

Conclusion

To allow for a successful distributed cognition outcome it is of importance to understand how individuals are affected by the different achievement goals. Clearly, the effective manipulation of the physical problem space is relative to the level of the task difficulty (e.g., modular arithmetic tasks) as well as the cognitive abilities (working memory resources in particular) of the individual. Finally, it is important to consider the implications of these studies on a practical level. Future mathematics education should take into consideration the findings of these two experiments in a way to make the learning experience more interactive for the more maths-anxious individuals (performance-focused individuals). The maths-anxious individual should be given the opportunity to reshape the presentation of the mathematical problems to extend their cognitive systems. By doing this, the working memory capacity can be augmented and as a consequence, the maths performance enhanced.

Ashcraft, M. H. (2002). Math anxiety: Personal, educational, and cognitive consequences. *Current Directions in Psychological Science*, *11*(5), 181–185. https://doi.org/10.1111/1467-8721.00196

Ashcraft, M. H., & Faust, M. W. (1994). Mathematics anxiety and mental arithmetic performance: An exploratory investigation. *Cognition & Emotion*, *8*(2), 97–125. https://doi.org/10.1080/02699939408408931

Ashcraft, M. H., & Kirk, E. P. (2001). The Relationships among Working memory, Math anxiety, and Performance. *Journal of Experimental Psychology. General*, *130*(2), 224–237. https://doi.org/10.1037/0096-3445.130.2.224

Ashcraft, M. H., & Krause, J. A. (2007). Working memory, math performance, and math anxiety. *Psychonomic Bulletin & Review*, *14*(2), 243–248. https://doi.org/10.3758/BF03194059

Avery, R. E., & Smillie, L. D. (2013). The impact of achievement goal states on working memory. *Motivation and Emotion*, *37*(1), 39–49. https://doi.org/10.1007/s11031-012-9287-4

Beilock, S. L. (2008). Math performance in stressful situations. *Current Directions in Psychological Science*. https://doi.org/10.1111/j.1467-8721.2008.00602.x

Beilock, S. L., & Carr, T. H. (2005). When high-powered people fail: Working memory and "choking under pressure" in math. *Source: Psychological Science*, *16*(2), 101–105. Retrieved from http://www.jstor.org/stable/40064185

Beilock, S. L., Holt, L. E., Kulp, C. A., & Carr, T. H. (2004). More on the fragility of performance: Choking under pressure in mathematical problem solving. *Journal of Experimental Psychology: General*, *133*(4), 584–600. https://doi.org/10.1037/0096-3445.133.4.584

Broekkamp, H., Hout-Wolters, B. van, & Van Hout-Wolters, B. (2007). The gap between educational research and practice: A literature review, symposium, and questionnaire. *Educational Research and Evaluation*, *13*(3), 203–220. https://doi.org/10.1080/13803610701626127

Carlson, R. A., Avraamides, M. N., Cary, M., & Strasberg, S. (2007). What do the hands externalize in simple arithmetic? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(4), 747–756. https://doi.org/10.1037/0278-7393.33.4.747

Crouzevialle, M., & Butera, F. (2013). Performance-approach goals deplete working memory and impair cognitive performance. *Journal of Experimental Psychology: General*, *142*(3), 666–678. https://doi.org/10.1037/a0029632

Crouzevialle, M., Smeding, A., & Butera, F. (2015). Striving for excellence sometimes hinders high achievers: Performance-approach goals deplete arithmetical performance in students with high working memory capacity. *PLoS ONE*, *10*(9). https://doi.org/10.1371/journal.pone.0137629

Darnon, C., Harackiewicz, J. M., Butera, F., Mugny, G., & Quiamzade, A. (2007). Performance-approach and performance-avoidance goals: When uncertainty makes a difference. *Personality and Social Psychology Bulletin*, *33*(6), 813–827. https://doi.org/10.1177/0146167207301022

DeStefano, D., & LeFevre, J. (2004). The Role of Working memory in Mental arithmetic. *European Journal of Cognitive Psychology*, *16*(3), 353–386. https://doi.org/10.1080/09541440244000328

Elliott, A. J., Shell, M. M., Henry, K. B., & Maier, M. A. (2005). Achievement goals, performance contingencies, and performance attainment: An experimental test. *Journal of Educational Psychology*.

https://doi.org/10.1037/0022-0663.97.4.630
Entwistle, N. (1988). Motivational factors in students' approaches to learning. In *Learning strategies and learning styles: Perspectives on individual differences* (pp. 21–51). https://doi.org/10.1007/978-1-4899-2118-5
Harackiewicz, J. M., & Linnenbrink, E. A. (2005). Multiple Achievement Goals and Multiple Pathways for Learning: The Agenda and Impact of Paul R. Pintrich. *Educational Psychologist*. https://doi.org/10.1207/s15326985ep4002_2
Hulleman, C. S., Durik, A. M., Schweigert, S. A., & Harackiewicz, J. M. (2008). Task Values, Achievement Goals, and Interest: An Integrative Analysis. *Journal of Educational Psychology*, *100*(2), 398–416. https://doi.org/10.1037/0022-0663.100.2.398
Hunt, T. E., Clark-Carter, D., & Sheffield, D. (2011). The Development and Part Validation of a U.K. Scale for Mathematics Anxiety. *Journal of Psychoeducational Assessment*, *29*(5), 455–466. https://doi.org/10.1177/0734282910392892
Kirsh, D. (1995). Complementary Strategies : Why we use our hands when we think. *Seventeenth Annual Conference of the Cognitive Science Society*, (Lave 88), 212–217.
Kirsh, D. (2010). Thinking with External Representations. *AI and Society*, *25*(4), 441–454. https://doi.org/10.1007/s00146-010-0272-8
Luttenberger, S., Wimmer, S., & Paechter, M. (2018). Spotlight on math anxiety. *Psychology Research and Behavior Management*, *Volume 11*, 311–322. https://doi.org/10.2147/PRBM.S141421
Neth, H., & Payne, S. (2011). Interactive Coin Addition: How Hands Can Help Us Think. *Proceedings of the 33rd CogSci*, 279–284. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?rep=rep1&type=pdf&doi=10.1.1.208.4071
Poortvliet, P. M., & Darnon, C. (2010). Toward a More Social Understanding of Achievement Goals. *Current Directions in Psychological Science*, *19*(5), 324–328. https://doi.org/10.1177/0963721410383246
Vallée-Tourangeau, F. (2013). Interactivity, Efficiency, and Individual Differences in Mental Arithmetic. *Experimental Psychology*, *60*(4), 302–311. https://doi.org/10.1027/1618-3169/a000200
Webb, S., & Vallée-Tourangeau, F. (2009). Interactive Word Production in Dyslexic Children. In *The 31st Annual Conference of the Cognitive Science Society* (pp. 1436–1441).

Active information seeking using the Approximate Number System

Jinjing (Jenny) Wang (jinjing.jenny.wang@gmail.com)

Elizabeth Bonawitz (lbaraff@gmail.com)

Department of Psychology, Rutgers University, Newark, NJ 07102

Abstract

Human adults share the ability to approximate large quantities without counting with newborn infants and non-human species. This ability is supported by the Approximate Number System (ANS) - a primitive and domain-specific cognitive system that supports noisy numerical decisions. How does the ANS support active exploratory decisions? Using a numerical comparison task, we found that the amount of active information seeking does not simply increase as the decision becomes more difficult. Instead, there seems to be an inverted U-shaped relationship between trial difficulty and how much one chooses to seek information. Additionally, this effect is not modulated by participants' performance, suggesting that participants' exploratory decisions based on ANS representations are driven by the utility of information seeking actions.

Keywords: Information Seeking; Active Learning; Approximate Number System; Decision Making

Introduction

How the mind processes sensory data and interprets the physical world is the hallmark question in cognitive science. However, the data we receive from the physical world is not readily interpretable. Rather than passively absorb all the information that is available, humans and animals actively explore and selectively attend to aspects of the world (Gottlieb, Oudeyer, Lopes, & Baranes, 2013). This kind of active exploration and selective attention is essential to effective learning and proper cognitive functioning. What determines when we want to explore and to what we choose to attend?

It has been widely demonstrated that observers, humans and animals alike, are drawn to novel and surprising events, which is often explained by a motivation to decrease errors in prediction (Loewenstein, 1994; Schultz & Dickinson, 2000). According to Loewenstein, an observer's desire to learn about a specific topic is driven by a discrepancy between the observer's existing knowledge and what they would like to know. Consistent with this account, infants as young as 10 months old can form expectations about object behavior, and explore more when these expectations are violated (Stahl & Feigenson, 2015). Relatedly, school-age and preschool children prefer to play with toys whose functionality are ambiguous or unexpected (Bonawitz, van Schijndel, Friel, & Schulz, 2012; Schulz & Bonawitz, 2007). This kind of prediction errors cannot only be mathematically defined, but has also been decoded from neuronal activities (Bromberg-Martin & Hikosaka, 2011).

In addition to novelty and surprise, humans and animals are also drawn to more complex stimuli or more difficult situations (Berlyne, 1966). For example, when confined in a minimally-stimulated space, adults prefer to produce light patterns that are the most diverse and unpredictable (Jones, Wilkinson, Braden, 1961). In another experiment, when probed about their curiosity about facts related to different animal species, adult participants were more curious about facts that they knew less about (Berlyne, 1954). These phenomena, that exploration is driven by novelty, surprise, and complexity, are consistent with the information processing account that defines information gain by uncertainty (Berlyne, 1960).

However, this tendency to be drawn to situations with maximum uncertainty (and to reduce it through learning actions) seems counterproductive in many cases. In particular, when the gap between one's current epistemic state and the information provided by the environment is too big, actions of learning and exploration can yield little benefit. For example, no matter how much effort a reader puts into staring at some foreign words without knowing the language or having access to a dictionary, the reader would still have no clue what the words mean.

Instead of linearly increasing exploratory actions as uncertainty increases, numerous studies have demonstrated a trade-off between the cost and benefit of information seeking actions (Coenen, Nelson, & Gureckis, 2018). When reading and rating contentful questions, such as "what instrument was invented to sound like human singing," adult participants' rated level of curiosity was the highest for questions that they had intermediate levels of confidence, and their level of curiosity was the lowest for questions in which they either had extremely low confidence or extremely high confidence (Kang et al., 2009). In a different exploratory situation, where each task option was initially hidden from participants, participants' exploratory decisions also followed a similar U-shaped pattern - they explored the most when the task was moderately difficulty, and explored less when the task was either too easy or too hard (Baranes, Oudeyer, & Gottlieb, 2014).

Consistent with these results, the field of developmental robotics suggests that exploration is based on dynamic changes in the rate of learning (Gottlieb, Oudeyer, Lopes, & Baranes, 2013; Oudeyer, Kaplan, & Hafner, 2007). Robots with this rate-based learning system can efficiently learn

skills in high dimensions without being distracted by activities that are either well learnt or unlearnable (Baranes & Oudeyer, 2013; Pape et al., 2012). Exploration increases as the rate of information increases. In cases when there is very low certainty (or high uncertainty), any particular action may produce new information, but if the problem space is complex enough, then additional information may not produce significant shifts in belief weights -- thus highly complex environments may not produce information that supports learning rates. Instead, learning rate may be highest in the Goldilock's spot (Kidd, Piantadosi, & Aslin, 2012), in which any particular action produces information to support a steeper learning rate. This predicts that, rather than a direct linear relationship, exploration should be lowest at both extremely low and extremely high levels of uncertainty, and exploration should be the highest at an intermediate levels, where information has the highest rate of return.

Results from infants' preference for object complexity are consistent with this account. Seven- and 8-month-old infants' probability of looking at an event was the lowest when looking at either highly predictable or highly surprising content (Kidd, Piantadosi, & Aslin, 2012; 2014; Piantadosi, Kidd, & Aslin, 2014; see also Pelz & Kidd, *in prep*). These results suggest that infants are able to direct their attention to maintain an intermediate rate of information absorption. It is possible that this kind of attentional mechanism is in place to prevent infants, who arguably have the most to learn and the least resources, from wasting cognitive resources on either overly predictable or overly unpredictable information.

One open question is whether adults reveal such trade-offs in active exploratory decision making situations. It is possible that this kind of balance between cognitive resource and exploration is unique to childhood. Another open question is whether such trade-offs are unique to novel learning environments or tasks that require higher-level conceptual reasoning, such as deciding what questions to ask or which route to take in a novel environment. When performing familiar activities using acquired skills, one may not need to adjust exploration based on uncertainty. Alternatively, the expected information gain from exploratory actions may explain information seeking behavior beyond these contexts.

To address these questions, the current study uses adults' exploratory decisions using a primitive cognitive system as a case study to test the relationship between problem difficulty and adults' exploratory decisions. Upon seeing 20 dots and 10 dots, without counting, we can immediately tell which array has more dots. This ability to automatically and effortlessly discriminate large numerosities is supported by the Approximate Number System (ANS; Dehaene, 1997), which produces noisy and ratio-dependent representations in human adults (Halberda, Ly, Wilmer, Naiman, & Germine, 2012), newborn infants (Izard, Sann, Spelke, & Streri, 2009), as well as non-human species (Cantlon, Platt, & Brannon, 2009; Dehaene, Dehaene-Lambertz, & Cohen, 1998). With ANS representations, discriminating 20 dots from 10 is just as easy as discriminating 40 from 20 (a ratio of 2), but both are easier than discriminating 15 from 10 (a ratio of 1.5). The discriminability of numerosities is determined by the numerical ratio, instead of set size, non-numerical dimensions (such as size of individual dots). In other words, the Approximate Number System strictly obeys Weber's Law (Dehaene, 2003). This well-established law allows us clean control over the difficulty and uncertainty of the trials - the less discriminable the trials (the closer the ratio), the more uncertainty. Additionally, infants and adults are able to maintain multiple numerical representations at once (Feigenson, 2008; Zosh, Halberda, & Feigenson, 2011).

This intuitive and automatic cognitive system provides a case study for testing the scope of the expected information gain account - whether adults' decision making using the intuitive and automatic numerical representations also demonstrate a cost and benefit trade-off of information seeking actions. It has been recently suggested that adults and children are sensitive to their internal confidence in numerical decisions (Baer, Gill, & Odic, 2018; Halberda & Odic, 2015), and numerical precision can be influenced by the order of trial difficulty (Odic, Hock, & Halberda, 2014; Wang, Libertus, & Feigenson, 2018; Wang, Odic, Halberda, & Feigenson, 2016). It is possible that this sensitivity to internal confidence or uncertainty drives adults' exploratory decisions in a way that balances the cost and benefit of information seeking actions. On the other hand, neuroimaging studies revealed that the encoding of ANS signals are extremely rapid - as fast as 180ms in the bilateral occipital-parietal sites (Hyde & Spelke, 2009; Park, DeWind, Wordoff, & Brannon, 2015). It is possible that this automatic encoding of numerical information leaves little room for improvement from exploratory actions, and hence adults may show no cost-benefit tradeoff in their exploratory decisions in a numerical task.

To test this, we designed a nonverbal numerical comparison task with four alternative forced choices. This design ensures that the numerical representations can be maintained in adults' working memory (i.e., about four items; Epelboim & Suppes, 2001; Luck & Vogel, 1997). On the other hand, a four-alternative-forced-choice paradigm lowers the chance level to 25%, which increases the performance gap between random guessing and effortful performance by 25% compared to two-alternative-forced choice tasks (which has a 50% chance level), potentially providing more utility for information seeking actions.

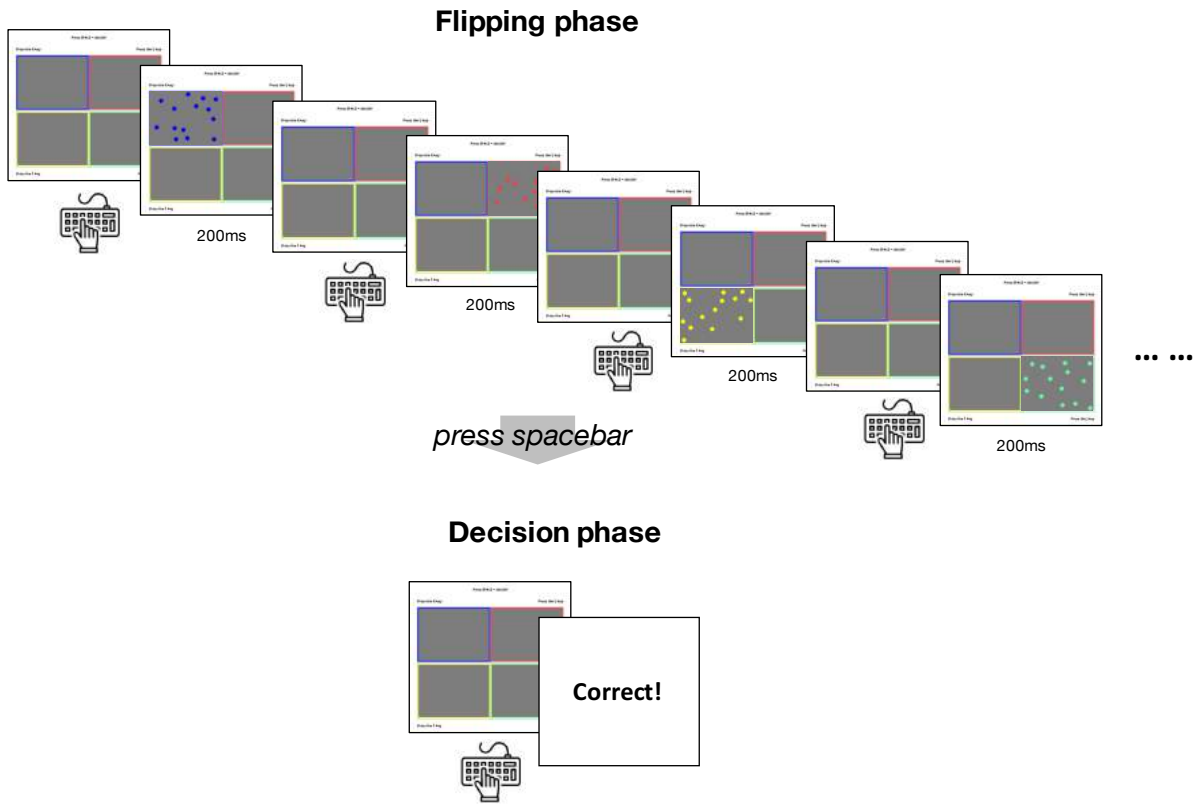


Figure 1: Schematic of the experimental procedure.

To further reduce random guessing, we also offered participants a small reward bonus depending on their performance. Participants can choose to see any one of four large arrays of dots, each for only 200ms which is too brief a window to count the dots. The key difference from traditional numerical comparison tasks is that participants are given the option to re-explore each array as many times as they would like before deciding the largest array. Participants then decide when they are ready to choose the array with the largest numerosity. Numerical comparison tasks allow us to systematically quantify and vary uncertainty and the difficulty of the task by changing the ratio between the numerosities.

If exploration is driven by the utility of information seeking actions, we would expect to see an inverted U-shaped relation between trial difficulty and the amount of exploratory actions. Alternatively, if exploration is driven by performance or error rate, then we should expect participants to explore most in trials in which the difficulty of the trials is the highest. Such an account would reveal information seeking to be linearly related to trial difficulty. Finally, if adults are not sensitive to the uncertainty of the trials, then exploration should not vary with the complexity of the trials.

Method

Participants Forty-two adults were recruited online through Amazon Mechanical Turk.

Stimuli Stimuli consisted of series of arrays containing collections of blue, red, yellow, and cyan dots on a grey background. During all the trials, three of the four arrays always contained the same number of dots (in different layout and configuration), and the fourth array differed from the remaining three with variable ratios. Difficulty was manipulated by changing the ratio between the largest number and the remaining number. Ratios varied between 1 (i.e., all four arrays were the same; the correct answer was pre-determined and randomly generated) and 2 (i.e., the larger number was twofold the smaller number), with at least 6 trials of each ratio. There were a total of 128 trials a participant could possibly complete.

Procedure After reading the instructions, each participant received an untimed practice session with eight practice trials. Participants then completed a timed test session where they had five minutes to complete as many trials correctly as possible. Participants were compensated based on how many trials they answered correctly during the test trials. Each trial started with four empty boxes outlined with distinct colors and paired with a reminding message about which key to press to "flip" the box and reveal the dots.

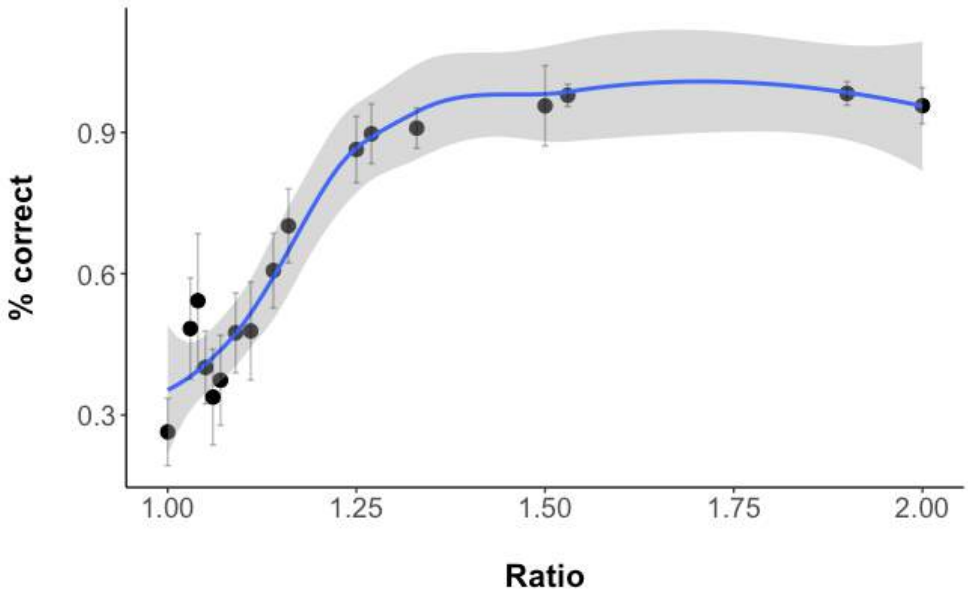


Figure 2: Average accuracy as a function of ratio (larger/ smaller). Error bars represent confidence interval of the mean.

After each keypress, dots appear for 200ms in the chosen box. For example, pressing the "R" key showed blue dots in the blue box, and pressing "U" showed red dots in the red box (Figure 1). During the flipping phase, the participant could press the spacebar to indicate that they were ready to move onto the decision phase at any point. Once the participant had moved to the decision phase, they were prompted to press a key to indicate which box contained the most dots. Feedback was provided after each trial. Participants on average completed 37.67 test trials (SD = 21.80).

Results

We first examined participants' accuracy in the decision phase. On average, participants performed correctly 62% of the time, well above chance (25%; binomial exact test $p < .001$).

We then averaged each participants' performance for each ratio to analyze the effect of ratio on accuracy. If participants used the ANS to solve the task, their performance should show the ratio-dependent signature of the ANS. Alternatively, it is possible that participants were able to count or maintain more precise representation of the numerical arrays after seeing them multiple times. As shown in Figure 2, participants' accuracy increases significantly as the ratio becomes easier. A log-linear regression model predicting accuracy using ratio explains over 72% of the variance, *beta* = .86, t = 6.54, $p < .001$, suggesting that participants primarily relied on ANS representations in the current numerical comparison task, even when they could receive additional information about the numerical stimuli. Consistent with previous research on adults' ANS precision, participants' accuracy on the task plateaued at about 1.5 ratio (Halberda & Feigenson, 2008).

The central question of the current study is how the difficulty of numerical decisions impacts people's information seeking. To test this, we examined the relationship between ratio and the number of boxes participants flipped before the decision phase.

.

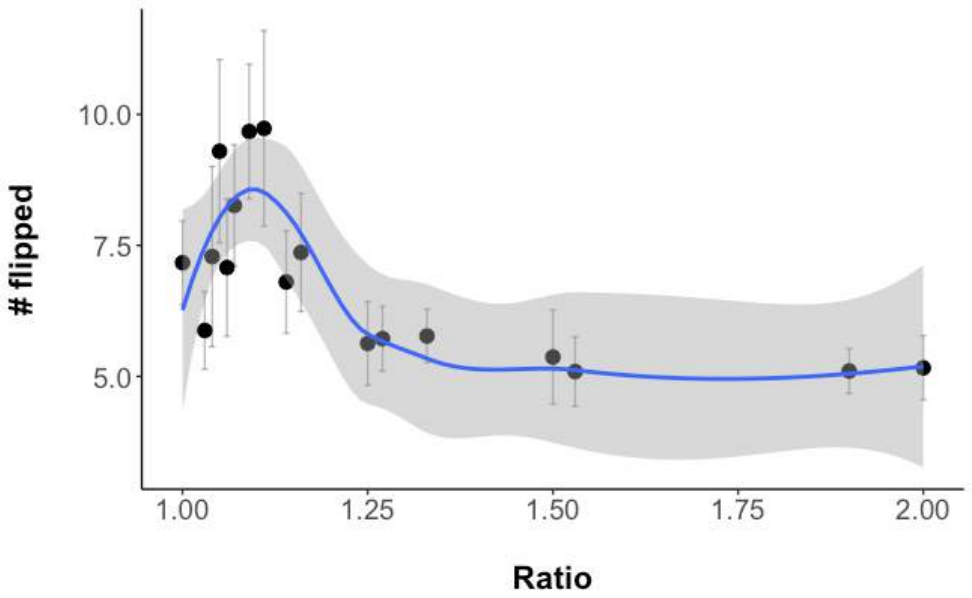


Figure 3: Average number of boxes flipped before decision as a function of ratio (larger/ smaller). Error bars represent confidence interval of the mean.

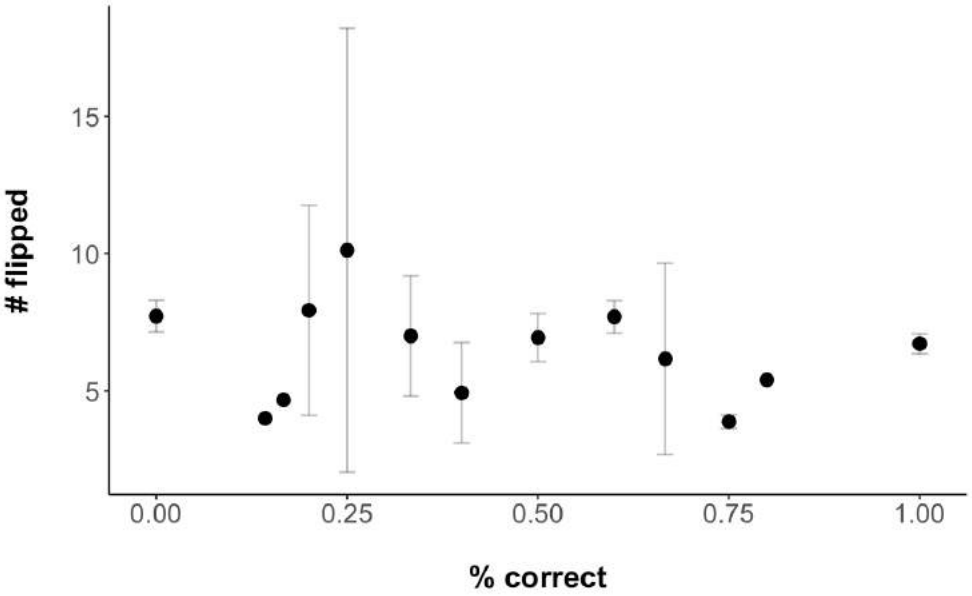


Figure 4: Average number of boxes flipped before decision as a function of accuracy. Error bars represent confidence interval of the mean.

As shown in Figure 3, overall, participants sought more additional information when the trials were more difficult. However, instead of a simple linear increase in number of flips as the ratio decreases, there is an inverted U-shaped relationship between ratio and flips when the ratio was between 1 and 1.25. Indeed, it is precisely in this range that participants steeply shift from near chance performance to near ceiling performance. This supports the claim that exploration is driven by expected information gain.

To test for a quadratic trend of ratio on number of flips, or an inverted U-shaped relationship between exploration and numerical ratio, we ran a model using both linear and quadratic ratio terms. This revealed a significant effect for both ratio (*beta* = -1.08, t = -3.49, p < .001) and ratio-squared (*beta* = .91, t = 2.91, p = .004). However, we found no relationship between average accuracy and the number of boxes flipped before decision (Figure 4). This

suggests that, rather than perceived performance or general motivation, participants' expected information gain drives their information seeking.

Conclusions

The current study investigated the relationship between the difficulty of numerical decisions and exploratory decisions. We found that adults' active search for additional information was the highest for trials with intermediate difficulty, and the lowest when the trials were either too easy or impossibly hard. Moreover, exploration has no clear relationship with numerical performance. These results suggest that numerical difficulty drives adults' exploratory decisions, showing a trade-off between cost of exploratory action and expected benefit from exploration.

Previous research has shown that infants seem to prefer an intermediate flow of information when exploring the environment (Kidd et al., 2012), and adults in novel or complex exploratory tasks explore the most when the task is at intermediate level of uncertainty (Kang et al., 2009; Baranes et al., 2014). These results have been taken to suggest that in learning and exploration, the observer have a tendency to optimize the cost of action and the gained information (Coenen et al., 2018). The current results extends this literature by suggesting that adults remain motivated to show such trade-off in their exploratory decisions even when using the primitive Approximate Number representations that have been active since infancy.

These results are consistent with both the idea that adults can balance the cost and benefit of exploratory actions, and that the rate of information gain drives exploratory behavior adults' numerical decision making. One possibility is that adults were making immediate decisions about whether to explore more solely based on the difficulty of each trial. Alternatively, it is possible that adults were dynamically adapting their exploratory decisions based on observed performance change, or their observed rate of learning, from previous explorations. Future research exploring the benefit of the exploratory actions, such as performance change with and without exploration, will help clarify the mechanisms by which adults make their exploratory decisions.

Where does this ability to dynamically adapt exploration to our own uncertainty come from? The similar U-shaped pattern in infants' attention suggests that infants are able to respond to probabilistic uncertainty in the environment (e.g. Kidd et al., 2012). However, it is possible that the ability to monitor the uncertainty in one's cognitive representations, such as numerical precision, may require more advanced metacognitive skills. Alternatively, infants may already come equipped with implicit representations of their uncertainty in numerical decisions. Recent work suggests that infants as young as 6 months old perform differently in a numerical change detection task as as the order of trial difficulty changes (Wang, Libertus, & Feigenson, 2018). It remains to be tested whether infants can adapt their exploratory behavior when using Approximate Number representations, and whether their exploration has the same kind of relationship with trial difficulty.

Another important question raised by the current study is whether active information seeking boosts numerical precision. In general, we found no relationship between overall accuracy and information seeking. It is possible that seeing the dot arrays more does not actually significantly impact people's accuracy at making numerical decisions. On the other hand, it remains possible that more complex interactions exist between information seeking and numerical precision. Future work examining the difference between people's numerical accuracy with and without information seeking will help test these possibilities.

References

Baer, C., Gill, I. K., & Odic, D. (2018). A domain-general sense of confidence in children. *Open Mind*, *2*(2), 86-96.

Baranes, A. F., Oudeyer, P. Y., & Gottlieb, J. (2014). The effects of task difficulty, novelty and the size of the search space on intrinsically motivated exploration. Frontiers in neuroscience, 8, 317.

Berlyne, D. E. (1954). A theory of human curiosity. *British Journal of Psychology. General Section*, *45*(3), 180-191.

Berlyne, D. E. (1960). Conflict, arousal, and curiosity.

Berlyne, D. E. (1966). Curiosity and exploration. Science, 153(3731), 25-33.

Bonawitz, E. B., van Schijndel, T. J. P., Friel, D., & Schulz, L. (2012). Children balance theories and evidence in exploration, explanation, and learning. *Cognitive Psychology*, *64*(4), 215–234. https://doi.org/10.1016/j.cogpsych.2011.12.002

Bromberg-Martin, E. S., & Hikosaka, O. (2011). Lateral habenula neurons signal errors in the prediction of reward information. *Nature Neuroscience*, *14*(9), 1209–1216. https://doi.org/10.1038/nn.2902

Cantlon, J. F., Platt, M. L., & Brannon, E. M. (2009). Beyond the number domain. *Trends in cognitive sciences*, *13*(2), 83-91.

Coenen, A., Nelson, J. D., and Gureckis, T. M. (2018). Asking the right questions about the psychology of human inquiry: Nine open challenges. Psychonomic bulletin & review, pages 1–41.

Dehaene, S. (1997). *The number sense: How the mind creates mathematics*. Oxford University Press, USA.

Dehaene, S. (2003). The neural basis of the Weber–Fechner law: a logarithmic mental number line. *Trends in cognitive sciences*, *7*(4), 145-147.

Epelboim, J., & Suppes, P. (2001). A model of eye movements and visual working memory during problem solving in geometry. *Vision Research*, *41*(12), 1561–1574. https://doi.org/10.1016/S0042-6989(00)00256-X

Feigenson, L. (2008). Parallel non-verbal enumeration is constrained by a set-based limit. *Cognition*, *107*(1), 1–18. https://doi.org/10.1016/j.cognition.2007.07.006

Gottlieb, J., Oudeyer, P.-Y., Lopes, M., & Baranes, A. (2013). Information-seeking, curiosity, and attention: computational and neural mechanisms. *Trends in Cognitive Sciences*, *17*(11), 585–593. https://doi.org/10.1016/j.tics.2013.09.001

Halberda, J., Ly, R., Wilmer, J. B., Naiman, D. Q., & Germine, L. (2012). Number sense across the lifespan as revealed by a massive Internet-based sample. *Proceedings of the National Academy of Sciences*, *109*(28), 11116-11120.

Hyde, D. C., & Spelke, E. S. (2009). All numbers are not equal: an electrophysiological investigation of small and large number representations. *Journal of cognitive neuroscience*, *21*(6), 1039-1053.

Izard, V., Sann, C., Spelke, E. S., & Streri, A. (2009). Newborn infants perceive abstract numbers. *Proceedings of the National Academy of Sciences*, *106*(25), 10382-10385.

Jones, A., Wilkinson, H. J., & Braden, I. (1961). Information deprivation as a motivational variable. *Journal of Experimental Psychology*, *62*(2), 126.

Kang, M. J., Hsu, M., Krajbich, I. M., Loewenstein, G., McClure, S. M., Wang, J. T. Y., & Camerer, C. F. (2009). The wick in the candle of learning: Epistemic curiosity activates reward circuitry and enhances memory. Psychological Science, 20(8), 963-973.

Kidd, C., Piantadosi, S. T., & Aslin, R. N. (2012). The Goldilocks Effect: Human Infants Allocate Attention to Visual Sequences That Are Neither Too Simple Nor Too Complex. *PLOS ONE*, *7*(5), e36399. https://doi.org/10.1371/journal.pone.0036399

Kidd, C., Piantadosi, S.T., & Aslin, R.N. (2014.) The Goldilocks effect in infant auditory cognition. *Child Development, 85*(5):1795-804.

Loewenstein, G. (1994). The psychology of curiosity: A review and reinterpretation. *Psychological Bulletin*, *116*(1), 75–98.https://doi.org/10.1037/0033-2909.116.1.75

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*(6657), 279–281. https://doi.org/10.1038/36846

Odic, D., Hock, H., & Halberda, J. (2014). Hysteresis affects approximate number discrimination in young children. *Journal of Experimental Psychology: General*, *143*(1), 255.

Oudeyer, P., Kaplan, F., & Hafner, V. V. (2007). Intrinsic Motivation Systems for Autonomous Mental Development. *IEEE Transactions on Evolutionary Computation*, *11*(2), 265–286. https://doi.org/10.1109/TEVC.2006.890271

Park, J., DeWind, N. K., Woldorff, M. G., & Brannon, E. M. (2015). Rapid and direct encoding of numerosity in the visual stream. *Cerebral cortex*, *26*(2), 748-763.

Pelz, M. & Kidd, C. (In prep.) The dynamics of attentional switching in a complex environment.

Piantadosi, S.T., Kidd, C., & Aslin, R.N. (2014) Rich Analysis and Rational Models: Inferring individual behavior from infant looking data. *Developmental Science, 17* (3): 321–337.

Schultz, W., & Dickinson, A. (2000). Neuronal Coding of Prediction Errors. *Annual Review of Neuroscience*, *23*(1), 473–500. https://doi.org/10.1146/annurev.neuro.23.1.473

Schulz, L. E., & Bonawitz, E. B. (2007). Serious fun: Preschoolers engage in more exploratory play when evidence is confounded. *Developmental Psychology*, *43*(4), 1045–1050. https://doi.org/10.1037/0012-1649.43.4.1045

Stahl, A. E., & Feigenson, L. (2015). Observing the unexpected enhances infants' learning and exploration. *Science*, *348*(6230), 91–94. https://doi.org/10.1126/science.aaa3799

Wang, J., Libertus, M. E., & Feigenson, L. (2018). Hysteresis-induced changes in preverbal infants' approximate number precision. *Cognitive Development*, *47*, 107–116. https://doi.org/10.1016/j.cogdev.2018.05.002

Wang, J. J., Odic, D., Halberda, J., & Feigenson, L. (2016). Changing the precision of preschoolers' approximate number system representations changes their symbolic math performance. *Journal of Experimental Child Psychology*, *147*, 82-99.

Zosh, J. M., Halberda, J., & Feigenson, L. (2011). Memory for multiple visual ensembles in infancy. *Journal of Experimental Psychology-General*, *140*(2), 141.

Identifying the Evolutionary Progression of Color from Crosslinguistic Data

Julia Watson
Department of Computer Science
University of Toronto
(jwatson@cs.toronto.edu)

Barend Beekhuizen
Department of Language Studies
University of Toronto
(barend.beekhuizen@utoronto.ca)

Suzanne Stevenson
Department of Computer Science
University of Toronto
(suzanne@cs.toronto.edu)

Abstract

We present a novel statistical analysis of color categorization using a standard method from semantic typology. Our approach shows that crosslinguistic color naming data exhibits latent dimensions whose order of relative importance matches the evolutionary ordering of emergence of those distinctions. Moreover, we show that the importance ordering of these dimensions holds even when controlling for frequency of the distinctions by looking at languages within each stage of evolution. Additionally, we find that the extreme points of the latent color dimensions correspond well to a small set of "universal" focal colors. Thus we show that a simple mathematical method simultaneously derives a consistent match both to the evolutionary stages and to the universal foci.

Keywords: semantic universals; color naming; color evolution.

Introduction

Much work in cognitive science seeks to uncover the basis of human categorization of the world. Semantic typology in particular aims to discover crosslinguistic constraints and tendencies in the ways that lexical semantic systems parcel concepts into named categories. Research across a number of diverse domains – from color to spatial relations to cutting and breaking events (e.g., Berlin & Kay, 1969; Levinson et al., 2003; Majid et al., 2008) – have revealed seemingly universal dimensions that underlie the organization of such lexical categories. For example, there is substantial evidence that color lexicons are organized around a universal set of basic color categories, whose best exemplars – *focal colors*, or *foci* – are clustered within small areas of the perceptual color space (e.g., Berlin & Kay, 1969; Regier et al., 2005; though see, e.g., Roberson et al., 2000, for an alternative view).

The domain of color has been particularly fruitful in revealing such crosslinguistic commonalities. Indeed, research on color is unusual (if not unique) in semantic typology in having revealed another kind of universal as well – that of evolutionary stages of a domain-specific lexicon. Berlin & Kay (1969) proposed that, as the number of basic color terms increase in a language, the named color distinctions emerge in one of a small set of constrained orders; for example, separate terms for yellow and red appear in a language before green is split off from blue. This line of work has been extended to cover a broad range of data from many languages, and the specific proposal refined and adapted. While some counterexamples have been identified, and it has been recognized that some languages do not exhaustively partition the color spectrum, most extant languages largely follow a successive partitioning of the color space according to universal principles (see Kay et al., 2009, for a review).

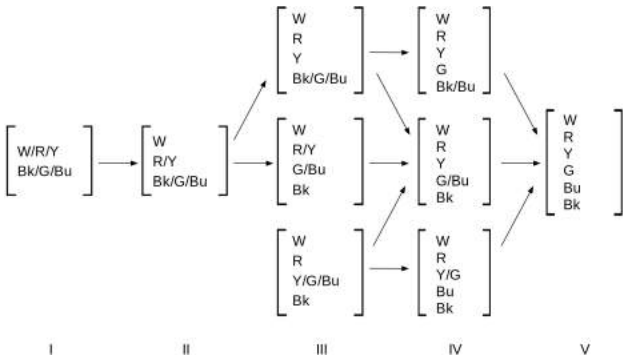


Figure 1: Evolutionary chart from Kay et al. (2009). W=white, R=red, Y=yellow, Bk=black, G=green, Bu=Blue.

Most evolutionary proposals focus on a core set of basic color categories, corresponding to the English terms *white*, *red*, *yellow*, *black*, *green*, and *blue*, because these 'primary colors' follow a consistent evolutionary path (Kay et al., 2009). Fig. 1 illustrates an influential proposal regarding the evolutionary sequence of languages, which we follow here. This diagram says that languages with only two color terms (Stage I) allocate those to the distinction between warm colors (White/Red/Yellow) and cool colors (Black/Green/Blue), while languages with three colors (Stage II) further distinguish White from Red/Yellow. Languages at Stages III through V can follow multiple pathways, depending on the further splits of Red, Yellow, Black, Green, and Blue.

Although modeling of crosslinguistic color data has revealed evidence of the various stages in Fig. 1 (e.g., Regier et al., 2007; Lindsey & Brown, 2009; Jäger, 2012), to our knowledge such work has not (yet) shown how the *ordering* of such stages could be derived from synchronic color naming data alone. Moreover, work on the evolutionary stages has typically focused on the partitioning of the space, and has not linked those stages to the nature and role of focal colors. That is, we know of no single model or method of analysis over the crosslinguistic color data that has derived both a consistent match to the evolutionary stages and to the universal focal areas, showing if or how these two concepts are linked.

Here we show that a standard analysis method from semantic typology, which has been used in work on color as well as

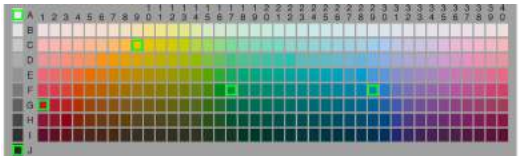


Figure 2: Munsell chart; universal foci (Regier et al., 2005)

in various other semantic domains, can simultaneously derive both the ordering of the evolutionary stages and the universal focal colors, revealing the latter as the drivers of the evolutionary distinctions. To preview our results, we use a simple statistical method to identify the important dimensions of the semantic space of color as represented by the crosslinguistic data. We find that the extremes of the identified dimensions correspond to focal areas of color, and, crucially, that the *importance ordering* of the dimensions corresponds to the *evolutionary ordering* of color distinctions. In addition, we perform the first analysis of subsets of languages at different evolutionary stages, and show that this importance ordering is not simply a by-product of the frequency of color distinctions in a mixed set of languages. We thus provide quantitative confirmation that synchronic color naming patterns reveal an underlying semantic space whose dimensional salience mirrors the evolutionary stages, with "anchors" in focal colors.

Color Data, Foci, and Evolutionary Stages

The World Color Survey (WCS; Kay et al., 2009, http://www.icsi.berkeley.edu/wcs/data.html) is a rich data set, along with a comprehensive qualitative analysis, that underlies much crosslinguistic analysis of color categories. Speakers of 110 diverse languages were surveyed, and languages were manually assigned to an evolutionary stage or transition between stages (cf. Fig. 1). The WCS contains two types of data. First, *naming data* was gathered by asking speakers to provide a single color term for each of the 330 color chips in the Munsell chart (see Fig. 2). Second, *focal data* was collected by asking speakers to select the most representative example, or focal color, of each term.

Both the focal and naming data of the WCS have played a prominent role in semantic typological analyses of color, which seek to derive semantic universals from crosslinguistic usage data. In particular, a body of work has attempted to go beyond qualitative analyses to provide precise mathematical underpinnings for such universals. Our work is in this vein, and we review related research below. Other quantitative analyses have attempted to link the semantic universals apparent from the WCS data to perceptual and/or communicative aspects of cognition. This is not the goal of our work here, but we refer to such research where relevant.

It is a striking finding that the distributions of the focal colors across all of the WCS languages cluster in small areas of the color space, corresponding to the six basic English focal colors, *white*, *red*, *yellow*, *black*, *green*, and *blue* (MacLaury, 1997; Regier et al., 2005; Lindsey & Brown, 2006, see Fig. 2). Some claim that these "universal" focal colors are cognitively privileged areas of the mental representation of color (Heider, 1972; Regier et al., 2005), which play a crucial role in the evolution of color systems (Berlin & Kay, 1969); others propose that they are only epiphenomena of the desired placement of color category boundaries (Roberson et al., 2000).

In their perceptual account, Jameson & D'Andrade (1997) suggest a middle ground in which the focal colors arise due to the nature of categorization in an irregular perceptual space. Abbott et al. (2012) operationalize this approach using Bayesian inference over perceptual color categories whose extents are determined by the WCS naming data. They find a good match between the representative members of these named color categories and the WCS focal data. This suggests that foci may be derivative from color categories whose optimal boundaries are driven by universal properties of the perceptual space (e.g., Jameson & D'Andrade, 1997; Regier et al., 2007). However, the relation of such foci to the evolutionary distinctions among colors is not clear.

The WCS data has also been explored as a source of insight into the evolutionary stages of color term systems, as exemplified in Fig. 1. Lindsey & Brown (2006, 2009) apply clustering techniques over naming patterns to reveal universal constraints over color categories, as well as color naming "motifs" (ways of partitioning the color space), some of which correspond to stages in the evolutionary hierarchy. Jäger (2012) takes a complex, multi-step approach to applying PCA to the WCS naming data, after transforming it in various ways. He derives partitionings of the six basic colors, many (but not all) of which match those of the evolutionary stages. While these approaches use quantitative analyses of the WCS to derive aspects of the evolutionary partitions, none of them derives an *ordering* over the partitions. (Indeed, Lindsey & Brown (2006) explicitly note that their work should not be interpreted as evidence of evolutionary sequencing from synchronic data.)

By contrast, Zaslavsky et al. (2018) combine WCS naming data with a perceptual semantics to derive an order over the emergence of color categories. They assume that color categories are created to optimally balance lexical accuracy with the size of the lexicon. As more color categories are added, their emergence roughly reflects the ordering of categories in color evolution. However, the reliance on perceptual salience leads to some mismatches with the evolutionary stages (overestimating the prominence of yellow), and the method does not address the role of focal colors in the ordering.

The wealth of research analyzing the WCS motivates our exploration of whether this rich synchronic color naming data can directly reveal patterns of evolutionary development, and shed light on the role of focal colors in those stages. We aim for a mathematical method of analysis that is simple and straightforward, with the intention that such an approach would be readily applicable to other semantic domains.

Our Approach

The approach we take in this work complements and seeks to fill in some of the gaps noted in the above body of research. Our goal is to derive the evolutionary sequence from WCS

data using a simple mathematical method – Principal Component Analysis (PCA) – that (along with other dimensionality reduction techniques) has been widely deployed in semantic typology, in diverse semantic domains including color (e.g., Majid et al., 2008; Jäger, 2012; Beekhuizen et al., 2014; Beekhuizen & Stevenson, 2018).

The novelty of our approach is two-fold. First, we extract latent dimensions of the WCS data *in order of importance*, yielding the first quantitative evidence of the evolutionary progression of color naming from the synchronic data. Second, we propose a natural interpretation of the "extremes" of the extracted dimensions as focal areas of color, which indeed show a strong match to empirical foci. Thus, we achieve a simultaneous match of the evolutionary ordering and the focal colors, which has not been shown before. Moreover, we do so with a very simple and straightforward use of PCA, in contrast to other methods that require much more involved mathematical transformations of the data (as in Jäger, 2012).

Our motivation is as follows. If most languages have followed a consistent and small set of orderings in the diachronic emergence of colors, those orderings must be determined by the relative importance of various perceptual/behavioral/cultural/communicative influences (e.g., Jameson & D'Andrade, 1997; Kay et al., 2009; Gibson et al., 2017; Holmes & Regier, 2017; Zaslavsky et al., 2018). Regardless of the source of these influences, if they play a role in the evolution of color systems, they may impact synchronic use of color terminology, and with the same relative importance. Note that this is not necessarily the case; for example, just because a language at Stage V has gone through Stages I through IV does not mean that the current color naming patterns of that language will reflect that a distinction made in an earlier stage (e.g., of Red vs. Yellow) is more important than a final distinction made in Stage V (e.g., of Green vs. Blue). That is, it is an open question whether the factors that exert evolutionary pressure to *create* new terms play a role in how terms are *deployed* in synchronic naming.

Experimental studies suggest that it may indeed be the case that evolutionary factors play a role in cognitive processing of colors by individuals. For example, Holmes & Regier (2017) found that English speakers show a categorical perception effect for the warm–cool distinction of Stage I, even though "warm" and "cool" are not basic color terms in English. Moreover, when English speakers group colors into K categories, the divisions they make roughly follow the evolutionary splits – i.e., with $K = 2$, they select a warm–cool separation, as in Stage I of evolution, with $K = 3$ they add a further distinction of white as in Stage II, etc. (Boster, 1986; Xu et al., 2013). Thus, English speakers are apparently sensitive to the evolutionary factors – and their relative ordering of importance – in color category processing.

Our goal here is to see whether actual color naming behavior, across the many diverse languages of the WCS, show this synchronic realization of the evolutionary influences. Specifically, given a suitable representation of the semantic space of synchronic color naming patterns, we use PCA over this data to extract dimensions of the data in order of importance, and examine whether those dimensions and their relative importance match the evolutionary stages proposed in the literature.

As a suitable representation of the color naming data, we follow a straightforward and standard practice in semantic typology. Specifically, we create a color chip by color term matrix using the color naming data from the WCS (Beekhuizen & Stevenson, 2018). Intuitively, such a matrix forms a semantic space over color, where each row can be viewed as a vector representation of the meaning of a color chip, as determined by the aggregate naming patterns in the data.

Applying PCA to such a matrix finds the latent dimensions characterizing the semantics of color across languages. Moreover, we take advantage of the interpretability of PCA dimensions, which means that points with a minimum or maximum value for a dimension are the most "extreme" example of the property that that dimension captures. Such points represent the "corners" of the data in the space (cf. Fig. 3), which are an indication of the key distinction each dimension is enforcing. We can thus examine these extremes to see if they correspond to the focal colors that have been proposed to "anchor" color categories (Regier et al., 2005).

Methods

Data matrices and PCA. We first create a color chip by term matrix over the naming data. Each cell records the (normalized) number of speakers in a language that used that term for that chip. This matrix compiles the naming data from all or selected subsets of languages in the WCS (as noted below). Thus we create matrices with 330 rows (one per chip in the Munsell chart) and up to 2223 columns (the number of color terms across all WCS languages).

We apply PCA to the resulting matrices to extract the most important dimensions. PCA identifies dimensions in the order along which the data shows the most variance, so the amount of variance accounted for represents the importance of that dimension. As we are looking for important dimensions that could relate to evolutionary development, we only consider dimensions that account for at least 5% of the variance in the data. (In almost all cases, this corresponded to a natural dropping off point in the accounted-for variance.)

For the first three such dimensions, we can plot the data for visualization purposes; i.e., we can plot the 330 color chips as represented by the first dimension of the PCA, by the first two, or by the first three. As shown in Fig. 3, such plots reveal the structure in the data that the PCA finds.

Determining the extreme points. To better understand the dimensions extracted by the PCA, we want to identify the *extreme points* of each. Conceptually, these are the maximally distinguishable points in the data on that dimension; in our visualizations in Fig. 3, these correspond to the endpoints or "corners" of the plotted data. In Fig. 3, the extreme points in 1D correspond to the minimum and maximum values on the x axis; the extreme points in 2D correspond to the corners of a triangle; the extreme points in 3D are the top of the pyramid

and the corners of its triangular base. As we will see in the results, such points generally correspond to focal colors.

We first collect the minimum and maximum points per dimension. However, because points can be tightly clustered at a "corner" of the space, we can end up with multiple extreme points when there is really only one "corner". For example, in the 2D plot in Fig. 3 (middle panel), the bottom right corner of the triangle is near the maximum for the x dimension *and* the minimum for the y dimension, so we might find two distinct points in that same small area. To avoid this, we consider all pairwise combinations of extreme points and merge those that are likely referring to the same "corners" of the space. Extreme points are considered to refer to the same "corner" if there is overlap in their $n = 15$ nearest neighbors, based on Euclidean distance. (We tried other values of n, which made little difference in the pattern of results.)

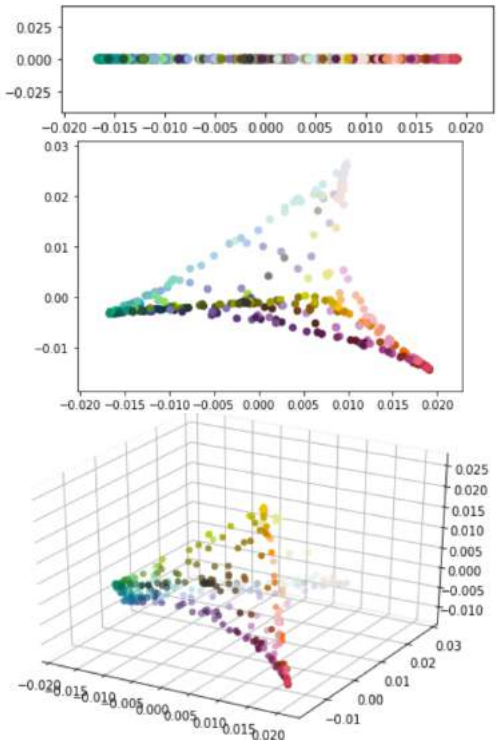


Figure 3: Plots of the 330 color chips in the 1D, 2D, and 3D PCA subspaces of the full WCS.

Visualizations of results. The results of our PCA analysis are a sequence of semantic subspaces defined by the extracted dimensions – the first (1D), the first+the second (2D), the first+the second+the third (3D), etc. – and the points that indicate the extreme "corners" of each subspace. To visually show the results, we plot the extreme points of each subspace (as triangles; or as diamonds for the merged points) in a Munsell chart, along with their closest neighbors (as circles whose size reflects their distance from the extreme point). These extreme point areas show the important color prototypes for each of the components of the PCA.

In addition, we visualize the extreme points as partitioning the PCA subspace, such that every color chip is allocated to the region of the space of its nearest extreme point. This yields a partitioned Munsell chart, with the number of partitions equal to the number of (merged) extreme points, or "corners", in the space. We label these partitions by the focal colors (green squares in the charts) occurring within them, whether they are extreme points or not. Thus each of the six focal colors is allocated to the region of its nearest extreme point, and we label a region by the focal colors it includes.

Fig. 4 shows examples of this visualization. For example, the White, Red, and Green extreme points shown in Fig. 4b for 2D correspond to the white, red, and green "corners" of the PCA plot for 2D shown in Fig. 3. The labels W, R/Y, and Bk/G/Bu correspond to the focal colors within each region as partitioned by the extreme points.

Analysis Over All WCS Languages

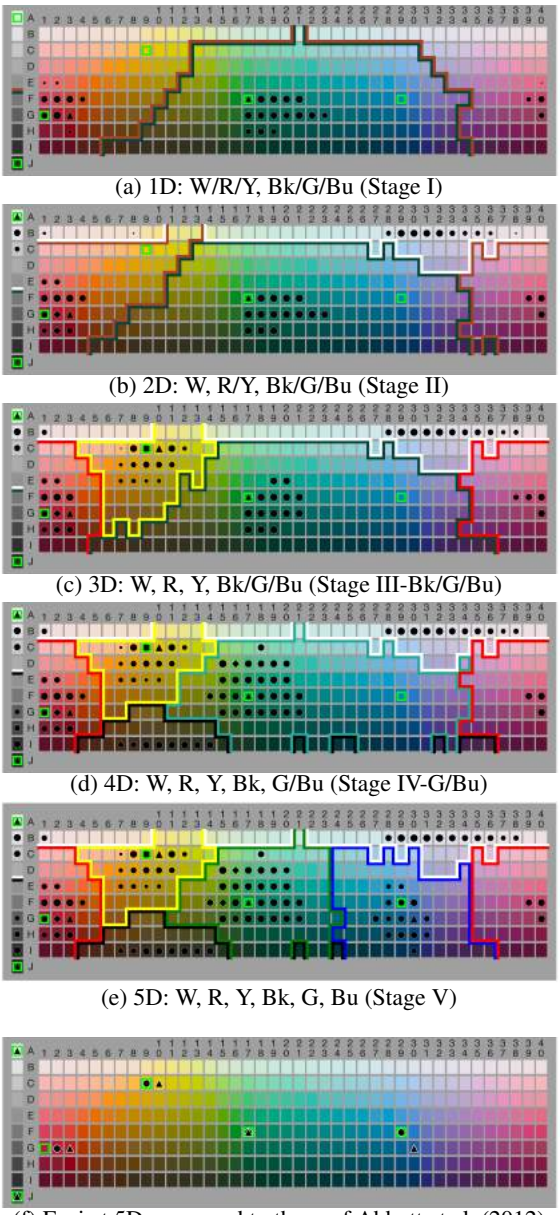
Using the methods above, our goal is to see whether a simple and straightforward application of PCA over the WCS naming data can simultaneously derive both the ordering of the evolutionary stages, as in Fig. 1, and the location of the universal foci, as in Fig. 2.

Results. We first apply our method over the full WCS color naming data. The first five extracted dimensions each account for more than 5% of the variance in the data; the results on the corresponding 1D–5D subspaces are in Fig. 4(a–e).

First, note that all of the primary extreme points for all dimensions of the PCA occur very close to the universal focal points of the 6 basic colors; see the triangles in Fig. 4. The extreme points corresponding to White, Black, and Green occur at the focus, Yellow and Blue adjacent to the focus, and Red two away from the focus. Fig. 4(f) shows the close match between our predicted focal colors and those of Abbott et al. (2012), who draw on both the color naming data and a representation of the named categories in perceptual space. It is important to emphasize that (as in Abbott et al., 2012) our results do not make use of the WCS focal color data, but only naming data. Given evidence for the universality of the focal colors as "anchors" for language-dependent naming of color regions, our results suggest that the extreme points of our PCA space are indeed meaningful in reflecting the important dimensions of color term systems in the WCS.

Second, and crucially, the importance ranking of dimensions in this all-languages data set shows a very strong match to the ordering of the evolutionary stages of Fig. 1, as indicated in the caption below each chart in Fig. 4(a–e). This is the first mathematical demonstration that synchronic color naming patterns reflect the relative importance of latent color dimensions that also underlie their evolutionary emergence.

Discussion. We have shown that a straightforward application of PCA over the WCS yields dimensions of the color naming data whose maximal/minimal values correspond to focal regions of color. That is, the universal foci appear to organize the dimensions along which the data shows the most variance. Moreover, these dimensions are extracted in the order of importance of the evolutionary distinctions among color terms, confirming that naming patterns in the WCS col-



(a) 1D: W/R/Y, Bk/G/Bu (Stage I)

(b) 2D: W, R/Y, Bk/G/Bu (Stage II)

(c) 3D: W, R, Y, Bk/G/Bu (Stage III-Bk/G/Bu)

(d) 4D: W, R, Y, Bk, G/Bu (Stage IV-G/Bu)

(e) 5D: W, R, Y, Bk, G, Bu (Stage V)

(f) Foci at 5D compared to those of Abbott et al. (2012)

Figure 4: Focal colors and associated regions from a PCA over all WCS data. (a–e) Results over each subspace for the first n dimensions; triangles = extreme points, circles = nearest neighbors, diamonds = merged extreme points, green borders = universal foci. Color labels below each chart correspond to the focal colors in each region, with the matching evolutionary stage indicated. (f) Foci in 5D, with our extreme points shown as triangles and the predicted foci of Abbott et al. (2012) as large circles.

lectively exhibit the synchronic influence of the evolutionary factors that shape the progression of color systems.

A legitimate question is whether our finding is simply the expected result of doing PCA over language data that includes languages at all the stages. Specifically, is it a simple frequency effect? That is, if languages at different stages are simply successively partitioning the data (rather than reorganizing colors in a way that changes earlier boundaries), then all languages have some boundary between warm and cool colors, all but Stage I languages have an additional boundary between white and the other warm colors, all but Stage I and II languages have another boundary between two more colors, etc. Thus, the PCA may be finding boundaries based on their frequency across languages at the different stages, rather than based on a true importance ordering.

A crucial observation that argues against this view is that the WCS contains no Stage I languages, and yet the warm–cool split of Stage I emerges as the first dimension in importance. That is: Although all languages in the WCS make *both* the warm–cool distinction of Stage I and the White–Red/Yellow distinction of Stage II, the warm–cool distinction emerges first in the PCA. This suggests that there is a detectable signal in the naming patterns that reveals the relative importance of an evolutionarily-earlier boundary over a later boundary, independently of their frequency in the data. To test this more directly, and across more stages, we next look at subsets of the languages of the WCS grouped by stage.

Analysis Over WCS Languages By Stage

We hypothesize that languages at each stage will show the same importance ranking of dimensions as found in the evolutionary progression, up to and including that stage. The set up here ensures that if we find that languages show evolutionarily-earlier distinctions as more important than later ones, this cannot be explained away as the data including languages at those earlier stages, thus skewing the frequencies toward the earlier distinctions.

Set up. In this analysis, we separately consider subsets of languages of the WCS that are in a single one of the identified evolutionary stages (Kay et al., 2009), yielding 7, 7, 41, and 14 languages at Stages II, III, IV, and V, respectively. (There are few documented Stage I languages, and none in the WCS. Also, we omit languages transitioning between stages, since they can show blends of behavior.) We perform the same PCA analysis as above, once over each of the four naming matrices limited to each stage, with the goal of seeing whether there is a match between the successive dimensions of each PCA analysis and the evolutionary stages.

Results. Tab. 1 presents the sequences of evolutionary stages revealed in the analysis of the subsets of languages by stage (omitting Stage V for space reasons). The table summarizes the color partitions in each subspace using the focal colors in each (we omit Munsell charts due to space reasons), and shows the best-matching stage from Fig. 1. (All and only dimensions accounting for $> 5\%$ of variance shown.)

Overall, the results confirm our hypothesis above: we find a very good match between the PCA analysis and the evolutionary diagram from all sets of languages, except those in Stage V. It is also the case that the majority of extreme points found in all the relevant dimensions of the four PCA analyses are at or very near focal colors. To summarize:

- All of Stages II, III, and IV show a very strong match to

Table 1: Sequences of focal colors in the Stages data.

Data set	1D subspace	2D subspace	3D subspace	4D subspace
Stage II languages	W/R/Y Bk/G/Bu ➔ **Stage I**	W R/Y Bk/G/Bu ➔ **Stage II**		
Stage III languages	W/R/Y Bk/G/Bu ➔ **Stage I**	W R/Y Bk/G/Bu ➔ **Stage II**	W R Y/warm–cool boundary G/Bu Bk ➔ **Stage III**[a,b]	W R Y G/Bu Bk warm–cool boundary ➔ **Stage III**[a,b]
Stage IV languages	W/R/Y/Bk G/Bu ➔ **Stage I**[c]	W R/Y/Bk G/Bu ➔ **Stage II**[c]	W R Y/Bk G/Bu ➔ **Stage III**[c]	W R Y G/Bu Bk ➔ **Stage IV**

[a] Stage III languages show a mix of the top and middle partitions of Stage III in 3D and 4D.

[b] Stage III languages show an additional warm/cool boundary color in 3D and 4D.

[c] Stage IV languages connect Bk with Y through the Brown region in 1D, 2D, and 3D.

the evolutionary stages.

- Stage III has, in addition to predicted extreme colors, a rainbow-like extreme area along the warm–cool boundary.
- Stage IV mostly matches the evolutionary stages, but with Yellow and Black connected through Brown in earlier dimensions. At 4D, there is a precise match to Stage IV.
- Stage V does not yield an ordering of dimensions that match the evolutionary stages. At 6D, all basic focal colors plus Purple have emerged as extreme regions.

Stage III data include languages from different sub-stages (column III in Fig. 1). Follow-up experiments with various subsets of Stage III languages reveal that the observed boundary color appears due to varying ways the different sub-stages divide up the G/Bu/Y region of color, in combination with the fact that one of the languages has an unusual basic color term for this warm–cool boundary region (Kay et al., 2009).

Stage V is a heterogeneous group, with languages having the 6 basic colors plus some number of other derived colors (13 of 14 have at least one derived color). We hypothesized that the variety in naming patterns for the non-basic colors may be swamping the signal from the basic colors. This may indicate a limitation of our method in dealing with a larger number of dimensions of color distinction. To test this, we performed the same PCA analysis over the 8 languages denoted as *approaching* Stage V (which have fewer derived colors). Here, the dimensions of the data emerged in order of the evolutionary stages, including the final stage at which Blue is distinguished from Green.

Discussion. To our knowledge, we are the first to apply a quantitative typological analysis to the languages of the WCS at the various evolutionary stages, as manually analyzed in Kay et al. (2009). Our findings provide strong support for the hypothesis that data from later stage languages can have structure that matches the evolutionary order of earlier stages. By separately analyzing languages at each specific evolutionary stage, we control for the potential frequency explanation of our results on the full WCS data set.

Further work will be required to determine the underlying causes of the cases of mismatches to the stages. Others have found, using more complex procedures, that the WCS data yield color groupings that largely, but not always, correspond to the manually derived partitionings of the color space (Lindsey & Brown, 2009; Jäger, 2012). Our method may be picking up on idiosyncratic patterns of naming, especially on smaller data sets. Regarding the Stage V data in particular, a possible shortcoming of our method is that it may not be sensitive enough to capture regularities beyond the six basic focal colors, which would be necessary to analyze this heterogeneous set of languages.

Conclusions

We present the first statistical analysis of color naming data that both shows a match between the evolutionary ordering of color systems and the importance ordering of informative dimensions of the data, and derives the focal colors from the extremes of those component dimensions. These results arise from a simple and straightforward application of PCA, a standard method from semantic typology for extracting salient dimensions from crosslinguistic naming patterns.

First, our approach reveals a quantitative importance ordering of latent dimensions of color semantics that strongly matches qualitative analyses of the evolutionary stages of color lexicons (e.g., Berlin & Kay, 1969; Kay et al., 2009). Specifically, we find that the color distinctions captured by each successive extracted dimension of the data largely correspond to the distinctions made in successive stages of color term evolution. Moreover, we show that the importance ordering of these dimensions holds even when considering languages at individual evolutionary stages, thus controlling for frequency of earlier vs. later distinctions in the data. Our work thus lends further evidence that speakers are sensitive to evolutionarily-important color distinctions that are not expressed directly by basic terms in their own language (cf. Boster, 1986; Xu et al., 2013; Gibson et al., 2017; Holmes & Regier, 2017).

Second, we find that the extreme points of the identified color dimensions correspond to a small set of focal color regions shown to occur across languages (e.g., MacLaury, 1997). Our work thus reinforces a growing body of research showing that focal colors are important dimensions of color space that serve as "anchors" for color categories (e.g., Regier et al., 2005). It has been proposed that focal colors arise at points of an irregularly-shaped perceptual space that maximize the distance between them (e.g., Jameson & D'Andrade, 1997; Regier et al., 2007). Although our method is agnostic as to the source of the latent dimensions (whether perceptual, and/or salience, as in Gibson et al., 2017, and/or communicative pressures, as in Zaslavsky et al., 2018), our results, like those of Abbott et al. (2012), show that the naming patterns of languages reflect the universal foci. Our approach further sheds light on the focal colors as extremes in the evolutionarily-important dimensions of color semantics.

Acknowledgments

JW and SS are supported by an NSERC Discovery Grant RGPIN-2017-06506 to SS. We thank the anonymous reviewers for their constructive comments.

References

Abbott, J. T., Regier, T., & Griffiths, T. L. (2012). Predicting focal colors with a rational model of representativeness. In *Proceedings of the 34th Annual Meeting of the Cognitive Science Society*.

Beekhuizen, B., Fazly, A., & Stevenson, S. (2014). Learning Meaning without Primitives: Typology Predicts Developmental Patterns. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society.*

Beekhuizen, B., & Stevenson, S. (2018). More than the eye can see: A computational model of color term acquisition and color discrimination. *Cognitive Science*, *42*(8), 2699–2734.

Berlin, B., & Kay, P. (1969). *Basic color terms: Their universality and evolution*. Berkeley, CA: UC Press.

Boster, J. (1986). Can individuals recapitulate the evolutionary development of color lexicons? *Ethnology*, *25*(1), 61–74.

Gibson, E., Futrell, R., Jara-Ettinger, J., Mahowald, K., Bergen, L., Ratnasingam, S., ... Conway, B. R. (2017). Color naming across languages reflects color use. *PNAS*, *114*(40), 10785–10790.

Heider, E. R. (1972). Universals in color naming and memory. *Journal of Experimental Psychology*, *93*(1), 10–20.

Holmes, K., & Regier, T. (2017). Categorical perception beyond the basic level: The case of warm and cool colors. *Cognitive Science*, *41*, 1135–1147.

Jäger, G. (2012). Using statistics for cross-linguistic semantics: A quantitative investigation of the typology of colour naming systems. *Journal of Semantics*, *29*(4), 521–544.

Jameson, K. A., & D'Andrade, R. G. (1997). It's not really red, green, yellow, blue: an inquiry into perceptual color space. In C. L. Hardin & L. Maffi (Eds.), *Color categories in thought and language* (pp. 295–319). CUP.

Kay, P., Berlin, B., Maffi, L., Merrifield, W. R., & Cook, R. (2009). *World color survey*. Stanford: CSLI Publications.

Levinson, S. C., Meira, S., & The Language and Cognition Group. (2003). 'Natural concepts' in the spatial topological domain – Adpositional meanings in crosslinguistic perspective: An exercise in semantic typology. *Language*, *79*(3), 485–516.

Lindsey, D. T., & Brown, A. M. (2006). Universality of color names. *PNAS*, *103*(44), 16608–16613.

Lindsey, D. T., & Brown, A. M. (2009). World color survey color naming reveals universal motifs and their within-language diversity. *PNAS*, *106*(47), 19785–19790.

MacLaury, R. E. (1997). Ethnographic evidence of unique hues and elemental colors. *BBS*, *20*(2), 202-203.

Majid, A., Boster, J. S., & Bowerman, M. (2008). The cross-linguistic categorization of everyday events: A study of cutting and breaking. *Cognition*, *109*(2), 235–250.

Regier, T., Kay, P., & Cook, R. S. (2005). Focal colors are universal after all. *PNAS*, *102*(23), 8386–8391.

Regier, T., Kay, P., & Khetarpal, N. (2007). Color naming reflects optimal partitions of color space. *PNAS*, *104*(4), 1436–1441.

Roberson, D., Davies, I., & Davidoff, J. (2000). Color categories are not universal: Replications and new evidence from a stone-age culture. *Journal of Experimental Psychology: General*, *129*, 369–398.

Xu, J., Dowman, M., & Griffiths, T. L. (2013). Cultural transmission results in convergence towards colour term universals. *Proceedings of the Royal Society B: Biological Sciences*, *280*(1758), 20123073.

Zaslavsky, N., Kemp, C., Regier, T., & Tishby, N. (2018). Efficient compression in color naming and its evolution. *PNAS*, *115*(31), 7937–7942.

Word-Learning Biases Contribute Differently to Late-Talker and Typically Developing Vocabulary Trajectories

Jennifer M. Weber (jennifer.m.ellis@colorado.edu)
Department of Psychology and Neuroscience, 345 UCB
Boulder, CO 80309 USA

Eliana Colunga (eliana.colunga@colorado.edu)
Department of Psychology and Neuroscience, 345 UCB
Boulder, CO 80309 USA

Abstract

This study explores how the vocabulary growth trajectories of typically developing and late-talker children change in relation to their word learning biases. Forty late talkers and 44 typically developing toddlers visited the lab once a month for one year starting at about 18 months of age. Word-learning trajectories were tracked using a parent-reported vocabulary measure, and shape and material bias measures were collected using the novel noun generalization task each month. A two-level hierarchical linear model was utilized for the longitudinal analyses. Results indicate that, at the first visit, a stronger shape bias was significantly associated with a *larger* vocabulary in typically developing talkers. In late talkers, however, a stronger initial shape bias was associated with a *smaller* vocabulary. Over the course of the study, for every additional visit, stronger shape biases were associated with *larger* vocabularies in late talkers, but not in typically developing toddlers. Results for the material bias mirrored the shape bias results. These findings suggest different possible underlying mechanisms for the two groups of children, as well as avenues for the design of language interventions that might support young late talkers.

Keywords: vocabulary acquisition; word-learning bias; late talker; word learning

Vocabulary Acquisition

There is enormous variability in the vocabularies of young children just beginning to speak. By two years of age, an otherwise typically developing toddler may know as few as ten words or well over 300 (Fenson, 1993). These early differences in vocabulary size may lead to long term differences in learning and language skills (Rescorla, 2000). Understanding the mechanisms behind language development that give rise to these different trajectories is vital for informing further research and developing identification and interventions for those children who show delayed word learning.

As children learn words, they also learn important features of the objects represented by these words and how these features relate to word use in general. Children must learn the regularities in their world, such as all balls are round, and all toothpaste is, well, thick and pasty. Children's noun learning progresses from slow and laborious to fast and seemingly effortless. This may be due in part to understanding and taking advantage of the way languages organize categories in the world. For example, by their third year of life, children seem to know to generalize names for solid objects by shape, but names for nonsolid substances by material (Landau, Smith & Jones, 1988). These word learning biases are typically assessed using the novel noun generalization (NNG) task; the child is taught a novel name for a novel item and then asked which other items, matching the exemplar on one or more features, have the same name (Landau et al., 1988). These biases develop in tandem with vocabulary growth, such that when children accrue between 50 and 150 nouns, the tendency to attend to shape for solid objects emerges and becomes robust (Gershkoff-Stowe & Smith, 2004).

Late talkers are children who lag in their vocabulary size compared to their same-aged peers in the absence of any known developmental disorders. Although the label of "late talker" is not a clinical diagnosis in of itself, this group is often defined by being in the lower 25th percentile on productive vocabulary, which is typically measured by the MacArthur-Bates Communicative Inventory (CDI) (Fenson, 1993). However, different researchers use different cut-off points when classifying children as late talkers, ranging from the 10th to 30th percentile.

Evidence suggests that late talkers and typically developing children differ not only in their vocabulary size, but also in the way they learn new words. Thirty-month-old late talkers, when defined as falling at or below the 30th percentile on the CDI, show no or even opposite word learning biases compared to typically developing 30-month-olds do (Jones, 2003). Further, 30-month-old late talkers under the 10th percentile struggle learning new words through fast mapping (Weismer, Venker, Evans & Moyle, 2013). Even before they turn two, children in the top 25th percentile on productive vocabulary show different word learning biases than children in the bottom 25th percentile (Colunga & Sims, 2017). Specifically, these early talkers showed as strong a shape bias for solids as a material bias for nonsolids, whereas late talkers showed a robust shape bias for solids that might be overgeneralized to nonsolids. The fact that late talkers differ from typically developing children in their word-learning biases in the lab may mean that these children acquire language through different mechanisms.

The differences between late talkers and their typically developing peers have long-term impacts, with some late talkers showing persistent deficits in measures such as reading, writing, and oral language skills throughout elementary and middle school (Rescorla, 2000). Although

many of the children labeled late talkers as toddlers do "catch up" to their typically-developing peers, there is a clear need to better understand how late-talker trajectories develop over time, as well as the factors that may influence this development (Heilmann, Weismer, Evans & Hollar, 2005). However, it is unknown exactly how word-learning biases relate to vocabulary growth throughout development, as previous work has investigated these relationships cross-sectionally. For example, the finding that 18-month-old late talkers have a shape bias (Colunga & Sims, 2017) but 30-month-old late talkers do not (Jones, 2003), could be a result of the different task demands of the different novel noun generalization tasks used with these different age groups, or it may suggest something interesting about the different developmental trajectories of children who catch up versus children who remain late talkers between 18 and 30 months of age.

To understand word learning in typically developing and late talkers, word learning biases and vocabulary size need to be examined longitudinally. For example, are late talkers who show stronger word learning biases at 18 months more likely to make greater gains over time? Is the positive relationship between word learning biases and vocabulary size in typically developing children suggested by cross-sectional work also present longitudinally? Further, is that relationship similar among late talkers? By investigating the relationship between vocabulary size as well as both the shape and material biases during the course of development, we can begin to understand the mechanisms that may give rise to the different developmental trajectories. The present study will investigate vocabulary growth trajectories of children over a 12-month period and their relations to word learning biases.

Current Study

The overarching goal of this project is to understand how the vocabulary growth trajectories of typically developing and late-talker children develop vis-à-vis their word learning biases. To accomplish this, we track both the vocabulary and the word learning biases of toddlers over a period in which rapid vocabulary growth is typically observed, 16 to 30 months of age, on a monthly basis for a year. Late talkers were oversampled to account for their expected increased variability. To the extent that there exists a feedback loop between word learning biases and words learned, we would expect a positive relationship between these two measures. This relationship might change throughout development, such that the shape bias for solids is stronger and more strongly related to vocabulary growth early on, and the material bias for non-solids shows a different pattern of development in relation to vocabulary size. Furthermore, if typically developing and late-talker children differ in their learning mechanisms, not just on their vocabulary size, these relationships between word learning biases and words known may differ between the two groups of children.

This study is the first attempt to track, longitudinally, the relationship between word learning biases and vocabulary size in late talkers and typically developing children. Though previous work has documented the relationship between the shape bias and vocabulary composition cross-sectionally (Perry & Kucker, 2019) and longitudinally (Gershkoff-Stowe & Smith, 2004) in typically developing children, and other work had looked at vocabulary growth longitudinally in late talkers (Heilmann et al., 2005) and at the relationship between shape bias and vocabulary cross-sectionally in late talkers (Jones, 2003), this is the first attempt to document the development of both the shape and material biases, longitudinally, in both late talkers and typically developing children.

Method

Participants

One hundred and twelve children were recruited for this study; children were 16-18 months of age at the first visit (M = 17.69, SD = 0.93). Twenty-eight children growing up in bilingual households were excluded for the present analyses, as previous research suggests early differences in the developmental trajectories of vocabulary growth of monolingual vs. bilingual children (e.g., Thordardottir, 2011). Toddlers visited the lab once a month for 12 consecutive months. Seventy-nine of the 84 children attended at least 10 of the expected 12 visits. Forty monolingual children scored below the 25th percentile on the CDI at their first visit, and for the present analyses, these children will constitute the late talker group (CDI percentile M = 11.65, SD = 12.26). The typically developing group consisted of the remaining 44 children (CDI percentile M = 59.14, SD = 21.44). Participating children were screened for known sensory or cognitive developmental disabilities or disorders. Late talkers and their typically developing peers did not differ in their ages at visit 1 or on average throughout the study; $t(82) = 0.71$, $p = 0.48$, $t(82) = 0.73$, $p = 0.47$, respectively.

Materials

Children participated in the novel noun generalization task to assess both the shape and material biases at each visit. The stimuli consisted of a warm-up set made out of common objects, a novel solid test set, and a novel nonsolid test set. The warm-up set had an exemplar, a red plastic ball, two other balls (a tennis ball and a green and blue rubber ball), a plastic spoon, a toy carrot, and a toy cat.

Each solid set consisted of an exemplar and five novel choices; two that matched the exemplar in shape but differed in color and material, one that matched in color, one that matched in material, and another that matched in both color and material. The nonsolid set was analogous, consisting of an exemplar and five choices; two items matching the exemplar's material but differing in shape and color, a color match, a shape match, and a color and shape match.

There were three sets structured in the way described above. The three sets rotated through the study, visit 1 – set A, visit 2 – set B, visit 3 – set C, visit 4 – set A, such that each set was used every 3 months and a total of 4 times over

the 12 visits that encompassed the study.

The Macarthur-Bates Communicative Development Inventory Words and Sentences (CDI) was completed by parents on each visit. The CDI consists of a 680-word checklist asking parents to indicate which words their child *says*. Although the CDI is a parent report measure it has been shown to be reliable and related to performance on child-based vocabulary measures (Fenson, 1993).

Procedure

Children visited the lab once a month for 12 months. At each visit parents filled out a CDI form measuring their child's productive vocabulary. Upon consent, children participated in one rotation of the NNG task measuring their shape bias for solid objects and their material bias for nonsolids. The procedure was modeled after Gershkoff-Stowe and Smith (2004). In the warm-up phase, the experimenter presented all six toys to the child and allowed him or her to look at them and handle and touch them for 30s before removing them outside of the child's reach. The child was then shown the exemplar ball and told, "look at this ball." Then, each child was asked to "get a ball" or get "another ball." If the child failed to retrieve a ball, the child was asked one more time, and finally was told "here's another ball," handed the ball, and was instructed to get it one more time. If the child got one of the nonball distracter items, he or she was told, "that's not a ball, that's a ____", then the distracter was replaced on the tray, and the child was asked again, "is there another ball?" The goal of the warm-up phase was to familiarize toddlers with the procedure and the idea that the display might have multiple things that were or were not in the category.

The procedure during the test phase with the solid and nonsolid novel sets was the same, except without feedback. Children were shown the exemplar and told, "look at this dax" and then asked to "get a dax" or "get another dax" for the solid set or "get more dax" or "get some dax" in the nonsolid set. Children were asked to get another (or more) until they indicated that there were no more, allowing children to accept or reject as few or as many items as they desired. The solid set was presented before the nonsolid set, and a five-minute break and change in testing rooms took place in between the two tests to minimize carry-over effects.

Bias scores were coded by noting the order in which children chose items as members of the queried category. The first choice got three points, second choice two points, and so on. For the solid set, the weighted scores for the items not matching in shape were subtracted from the weighted scores for the two shape-matching objects, yielding a score from -5 to 5. Similarly, the material bias score for the nonsolid set was calculated by subtracting the weighted scores for the two items matching the nonsolid exemplar in shape from the scores for the items matching the exemplar in material.

Data Analysis

We employed a two-level hierarchical linear model to investigate our longitudinal data. We are able to quantify longitudinal growth trends and explore the variation in these trends across individuals. The "level 1" analysis estimates parameters within child, which in turn become the dependent variable for the "level 2" analysis assessing between-child variables. Number of words known, taken from the CDI, was the main outcome of interest across the analyses. Level 1 consisted of each visit within child, whereas level 2 quantified individual characteristics across children (e.g. talker type). We first graphed the trajectories of all children in the study to help visualize the data (*Figure 1*). we elected to use a linear growth description. The graph also indicates great variability in both the initial and ending vocabulary sizes of the children, as well as their trajectories throughout the study. Because of this, we will investigate not only fixed effects but the variance of the modeled growth curves.

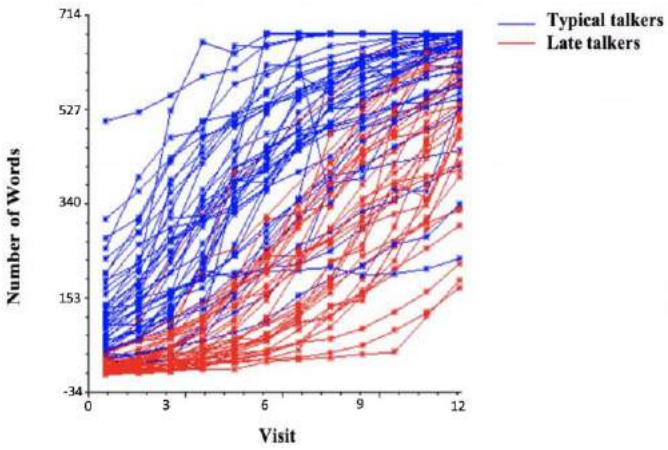


Figure 1: The raw data vocabulary trajectories for our full sample, color-coded by talker type. The x-axis is visit number, and the y-axis is the number of words the children produce from the CDI.

Model 1 Our first analysis seeks to describe the linear trajectories for all children in our sample, in order to establish a baseline for comparison. We centered Visit around visit one and allowed coefficients to vary at level 2. Figure 2 represents model 1 using the hierarchical linear modeling framework.

Level-1 Model: repeated observations of children

$$Vocab_{ti} = \pi_{0i} + \pi_{1i}(Visit_{ti}) + e_{ti}$$
$$where\ e_{ti} \sim iid,\ N(0,\ \sigma^2)$$

Level-2 Model: child characteristics

$$\pi_{0i} = \beta_{00} + r_{0i} \qquad r_{0i} \sim iid,\ N(0,\ \tau_{00})$$
$$\pi_{1i} = \beta_{10} + r_{1i} \qquad r_{1i} \sim iid,\ N(0,\ \tau_{11})$$

Figure 2: a representation of model 1 using the hierarchical linear modeling framework. This same structure will be used for analyses 2 and 3 as well, with added variables at level 1 and 2. For simplicity, we only present the general structure here.

Model 2 Our second analysis further investigates differences in growth trajectories between those children who were initially classified as late talkers (CDI < 25%) and those who were initially classified as typical talkers (CDI >25%). In our analysis, we examined the interaction between this talker type variable and both initial vocabulary size and linear growth. To do this, talker type status was placed at level 2 of the analysis as a between subjects variable, in order to predict the coefficients in our level 1 equation. Level 1 remained the same as in model 1. We centered our talker type variable to test the significance of both late and typical talker slopes separately.

Model 3a and b Our third analyses examine how the two word-learning biases of interest, shape and material, impact the number of words known by children both at visit one and over time. Here we investigate each bias separately. Further, we examine how these biases differentially impact the word learning trajectory for late talkers as compared to typical talkers. To do so, we add the bias score and the bias score by visit interaction to level 1 of our model. Level 2 remains similar to model 2, with talker type status predicting the intercept and visit coefficients. In addition, talker type is also placed in the level 2 equations for bias and the bias by visit interaction. Model 3a investigates the relationship between the shape bias for solid objects and vocabulary size at visit one and over the course of the study, whereas Model 3b does the same for the material bias toward nonsolid substances. A model including both shape and material bias, along with their interactions with each other and over visit, did not account for any more within-child variability in growth than models with either bias alone. Therefore, results will not be reported for such an analysis.

Results

Fixed affects for each model are presented in *Table 1* and variance components in *Table 2*.

Model 1

Results from model 1 indicate that, on average, children in our study had 63.27 words in their vocabularies at the first visit, or when 18 months old. For every month of the study, children, on average, accrued 47.05 new words, t(83)=28.51, p<.001. As expected, all children learned new vocabulary words as they aged. There was significant variability in both initial vocabulary size and visit slope; χ^2 (82, N=84) = 1554.80, p<.001, χ^2 (82, N=84) = 712.73, p<.001 respectively.

Model 2

Analysis 2 investigates differences in the word learning trajectories between late talkers and their typically developing peers (*Figure 3*). Typical talkers are predicted to have, significantly more words in their vocabularies at the first visit than late talkers; t(82)=8.99, p<.001. Both groups made significant vocabulary gains over time. Typical talkers were expected to add 48.19 words each visit, whereas late talkers were predicted to gain 45.53 words each month; t(82)=22.79, p<.001, t(82)=17.67, p<.001 respectively. In fact, vocabulary growth was not significantly different between the two talker types; t(82)=0.80, p=.427.

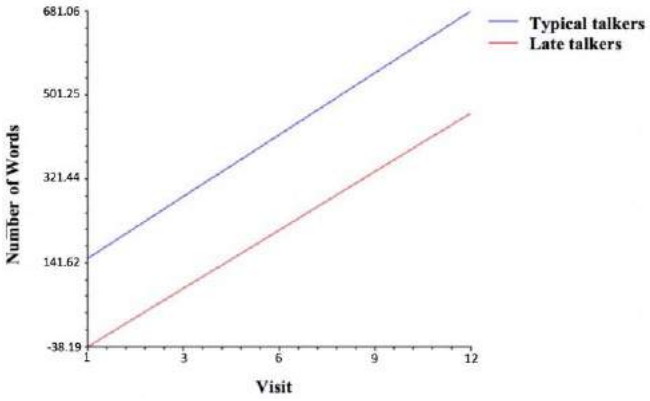


Figure 3: Predicted average vocabulary trajectory for typical and late talkers from model 2. Word-learning trajectories for both talker groups run parallel to each other

There is still significant variability in initial vocabulary size and linear growth; χ^2 (81, N=84) = 793.62, p<.001, χ^2 (81, N=84) = 710.86, p<.001 respectively. Knowing which talker type group a child belonged to did account for 51.85% of the variance between children's initial vocabulary sizes when compared to the first analysis. However, there was no appreciable difference in the variance of vocabulary growth from model 1 to model 2.

Model 3a – Shape Bias

Model 3a investigates how shape bias predicts both vocabulary size at visit one and over the course of the study (*Figure 4*). At the first visit, late talkers did not differ from their typically developing peers in shape bias scores; t(82)=1.61, p=.11. Controlling for shape bias score and its change over time, both late and typical talkers still, as in model 2, know significantly more words at each new visit. Typical talkers learn an average of 51.77 words a month and late talkers learn 41.84 words monthly on average; t(82)=21.18, p<.001, t(82)=16.75, p<.001. However, late talkers make significantly smaller gains in vocabulary size than their typical counterparts once shape bias and its changes over time were accounted for; t(82)=2.82, p<.01.

At the first visit, for every one-point increase in shape bias score, typical talkers were expected to know 6.05 more words, indicating that a stronger shape bias is significantly associated with a larger vocabulary in typically-learning talkers at the beginning of the study; t(82)=3.368, p<.001. In contrast, for every one-point increase in shape bias score, late talkers were predicted to initially know 7.44 fewer words -- for late talkers, a stronger initial shape bias was associated with a smaller vocabulary. This difference between shape

bias and vocabulary size was significantly different for the two groups at their first visit; t(82)=6.069, p<.001.

For every additional visit, a one-point increase in shape bias score predicted a significant 1.39 word decrease in vocabulary size for typical talkers; t(82)=-4.885, p<.001. However, by the 10th visit, 43% of typical talkers already knew at least 90% of the words on the CDI, indicating possible ceiling effects in the typical talker group. Late talkers, on the other hand, showed a significantly more positive relationship between shape bias scores and vocabulary size over the 12-month study period; t(82)=7.216, p<.001. For each visit and one-point increase in shape bias score, late talkers were expected to know 1.61 more words; t(82)=5.31, p<.001.

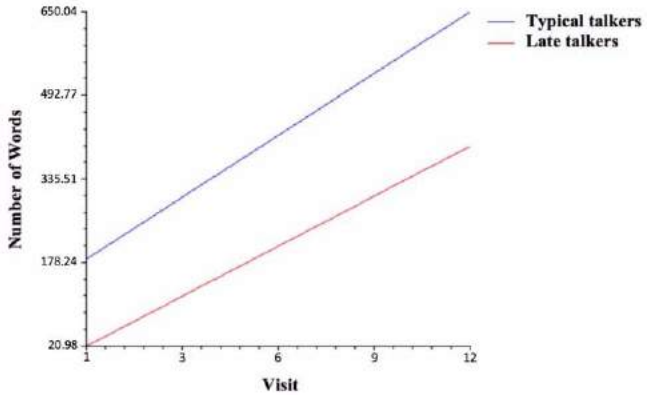


Figure 4: Predicted average vocabulary trajectory for typical and late talkers from model 3a. The words learning trajectories differ between late and typical talkers.

The inclusion of shape bias, its change over time, and how these variables differ between late and typical talkers in the model accounted for 8.7% of the within-child variability in vocabulary. However, there is significant variability in initial vocabulary size and growth; χ^2 (81, N=84) =455.60, p<.001, χ^2 (81, N=84)=475.46, p<.001 respectively.

Model 3b – Material Bias

Model 3b examines how the strength of a child's material bias predicts their vocabulary size at visit one and over the 12-month study period (*Figure 5*). At the first visit, late talkers did not differ from their typically developing peers in material bias scores; t(82)=1.02, p=.31. For a child with an average material bias strength at visit one and over time, both typical and late talkers are predicted to know significantly more words every month of the study, gaining 48.55 and 45.03 words, respectively; t(82)=23.07, p<.001, t(82)=17.50, p<.001. Further, these gains are not significantly different between the two groups; t(82)=1.06, p=0.291. This differs from when shape bias was controlled for, where late talkers did make significantly less gains in word knowledge than typical talkers.

At the first visit, for every one-point increase in material bias score, typical talkers are expected to know 4.32 more words, indicating a significant positive relationship between material bias and vocabulary size; t(82)=2.22, p<.05. The relationship between words known and material bias is significantly more negative for late talkers however; t(82)=3.54, p<.001. Late talkers are expected to know significantly less words (5.3) for every one-point increase in material bias; t(82)=-2.8, p<.01. This directly mirrors the results for shape bias.

The material bias by visit interaction also follows the same pattern as that for the shape bias. For every month aged, a one-point increase in material bias predicts a significant reduction in vocabulary size, by 0.95 words; t(82)=-3.06, p<.01. To note, this reduction in vocabulary size is not as large as the one for the shape bias, at a decrease of 1.39 words. Further, this relationship is significantly more positive for late talkers, as it was for the shape bias; t(82)=4.07, p<.001). For each month aged and a one-point increase in material bias score, late talkers are expected to know 0.92 more words; t(82)=2.72, p<.01. As the year goes by, a stronger material bias predicts more vocabulary gains for late talkers, but fewer gains for typical talkers.

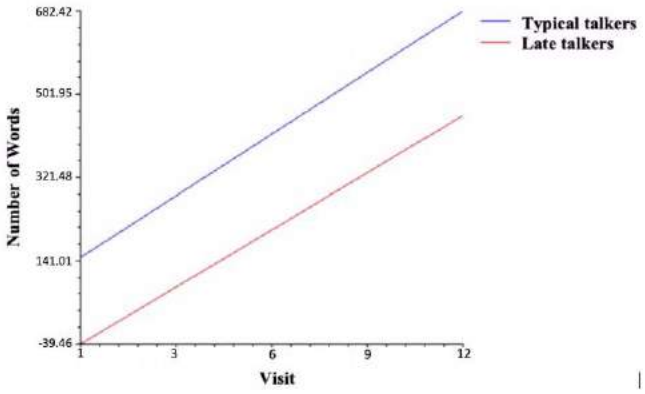


Figure 5: Predicted average vocabulary trajectory for typical and late talkers from model 2. The word-learning trajectories between late and typical talkers do not differ.

Including the material bias and its relationship with visit in the model accounts for about 5.9% of the variance at level one (as compared to 8.7% when shape bias was used). Variability in initial vocabulary size and linear growth are both significant χ^2 (81, N=84) =713.21, p<.001, χ^2 (81, N=84)=588.33, p<.001 respectively.

Table 1: Final estimation of fixed effects for the three analyses. Each intercept is the estimate for typical talkers. We indicate the coefficient for each variable followed by the significance (indicated *** p<.001, **p<.01, *p<.05). In parentheses are standard errors.

Fixed Effects	*Model 1*	*Model 2*	*Model 3a*	*Model 3b*
For Initial Vocabulary Size, π_{0i}				
Initial Vocabulary Size, β_{00}	63.27(14.6)***	150.93(18.5)***	132.24(18.3)***	148.26(18.2)***
Late talker, β_{01}		-189.11(19.6)***	-153.02(19.2)***	-187.70(19.6)***
For Visit slope, π_{1i}				
Visit slope, β_{10}	47.05(1.7)***	48.19(2.1)***	51.77(2.4)***	48.56(2.1)***
Late talker, β_{11}		-2.66(3.3)	-9.93(3.5)**	-3.53(3.3)
For Bias slope, π_{2i}				
Bias, β_{20}			6.05(1.8)***	4.32(2.0)*
Late talker, β_{21}			-13.49(2.2)***	-9.62(2.7)***
*For Visit*Bias slope,* π_{3i}				
*Visit*Bias,* β_{30}			-1.39(0.29)***	-0.95(0.31)**
Late talker, β_{31}			3.01(0.4)***	1.87(0.89***

Table 2: Final estimation of variance components for the three analyses. We indicate the variance component for each variable followed by the significance (indicated *** p<.001, **p<.01, *p<.05). In parentheses are standard deviations.

Variance Components	*Model 1*	*Model 2*	*Model 3a*	*Model 3b*
Initial Vocabulary Size, r_{0i}	17128(130.9)***	8246.53(90.8)***	7360.89(85.8)***	8139.87(90.2)***
Visit, r_{1i}	199.09(14.1)***	199.35(14.1)***	208.06(14.4)***	194.99(14.0)***
Bias, r_{2i}			14.24(3.8)	27.98(5.3)*
*Visit*Bias,* r_{3i}			0.63(0.8)*	0.79(0.9)*
Level-1, e_{ti}	3165.89(56.3)	3166.61(56.3)	2887.73(53.7)	2979.37(54.6)

Discussion

The work presented here looks at the differential contributions of word learning biases to the developmental trajectories of typically developing children and late talkers, and in doing so provides important novel insights. First, word learning biases may not be equally advantageous to all children. At the beginning of the study, typically developing talkers show the expected positive relationship between shape bias score and vocabulary size, suggesting that a shape bias facilitates word learning, in line with decades of work by Linda Smith and colleagues (e.g., Smith, 2000). However, the relationship between word learning biases and vocabulary size among late talkers presents a different pattern.

It is important to note that late talkers and their typically developing peers do not differ in their initial shape bias scores. Although this may seem to contradict Jones' (2003) finding that 30-month-old late talkers do not show a consistent shape bias, that is not the case. Rather, these results complement Jones' by documenting a different point in the developmental timeline; participants in our study were at least a year younger at the beginning of our study. In addition, we used an age-appropriate novel noun generalization task different from that in Jones (2003).

In contrast to the documented positive relationship between the shape bias and vocabulary size in typically developing children, among late talkers there is a negative relationship between the strength of their shape bias and their vocabulary size at the beginning of the study. This intriguing finding suggests our measure of the shape bias might not be distinguishing different underlying mechanisms that these two groups of children might be using. For example, it is possible, given the specific novel noun generalization task we used, that late talkers are exhibiting a generalized shape bias that is not linked to learning new words, but instead simply to attending to shape more generally. The fact that this same pattern of results held also for the material bias, however, suggests that this is not the case, and that at the very least they can shift their attention depending on the physical characteristics of the objects in front of them. Another possible reason our specific shape bias measure might not be detecting different underlying mechanisms is that we do not test retention. It is possible that in the short term, late and typical talkers generalize novel nouns in the same way (though there are documented differences in their fast mapping abilities; Weismer et al., 2013), but typically developing children have an easier time remembering the

word-shape association over time, which would result in different rates of word learning in the real world.

Second, the relationship between word learning biases and vocabulary growth over time in the two groups of children might offer further clues. As the year goes by, we observe that among late talkers, increases in shape bias score is related to vocabulary gains. That is, the 18-month-olds who started as late talkers and grew up to have a robust shape bias by 30 months of age were likely not late talkers at all by that point. Here it is important to note that by the end of the study only seven out of the 39 children who started as late talkers, or about one in five, remained under the 25th percentile; about half of them were above the 50th percentile at the last visit. This could just be a function of the regular course of development, as it is well known that one of the difficulties of dealing with late talkers is that many of them will catch up without the help of any intervention as others will continue to struggle into their school years and beyond. In fact, Heilmann et al. (2005) suggest that the CDI can help identify children with low language skills up to the 11th percentile from children with normal language skills above the 49th percentile. Given that the majority of the late talkers in or sample (27/39) started the study under the 11th percentile mark, our rate of late talker recovery seems higher than expected. Is it possible that participating in this study helped late talkers acquire an effective shape bias? Whether that is the case or not, these findings suggest possible avenues for the design of language interventions that might support young late talkers.

On the other hand, for typically developing children, as the year goes by, increases in shape bias score are related to smaller vocabulary sizes. This unexpected finding is likely an artifact of typically developing children reaching ceiling performance in both the CDI and their word learning bias scores before the end of the study. Because the CDI is a finite set of about 700 words, the vocabulary curves of typical talkers artificially asymptote towards the end of the study, when in fact their vocabularies continue to grow as they acquire words beyond those listed in the CDI. One way to deal with this is to use open-ended diaries rather than vocabulary inventories to measure vocabulary. In addition, this would allow us to capture idiosyncratic differences in vocabulary composition in late talkers as well.

The present study, with the use of hierarchical analysis, sheds light on the differences in language acquisition between those who lag behind in vocabulary size, late talkers, and those that are developing typically. Although the present analyses are just a first step in understanding these trajectories, they suggest interesting targets for future work. With this knowledge, earlier identification of children at risk for delayed vocabulary acquisition, as well as the development of more targeted interventions for such children, might be possible.

Acknowledgments

We are grateful to the families of Boulder, CO, who participated and to the research assistants in the DACS Lab at the University of Colorado Boulder who collected the data. This research was supported by NICHD grant R01 HD067315 to Eliana Colunga.

References

Colunga, E., & Sims, C. E. (2017). Not only size matters: Early-talker and late-talker vocabularies support different word-learning biases in babies and networks. *Cognitive science*, *41*, 73-95.

Fenson, L., Dale, P. S., Reznick, J. S, Thal, D., Bates, E., Hartung, J. P., Pethick, S., & Reilly, J. S. (1993). *The MacArthur Communicative Development Inventories: User's guide and technical manual.* San Diego, CA: Singular Publishing Group.

Fenson, L., Marchman, V. A., Thal, D. J., Dale, P. S., & Reznick, J. S. (2007). *MacArthur-Bates communicative development inventories: User's guide and technical manual*. Baltimore, MD: Brookes.

Ganger, J., & Brent, M. R. (2004). Reexamining the vocabulary spurt. *Developmental psychology*, *40*(4), 621.

Gershkoff-Stowe, L., & Smith, L. B. (2004). Shape and the first hundred nouns. *Child development*, *75*(4), 1098-1114.

Heilmann, J., Weismer, S. E., Evans, J., & Hollar, C. (2005). Utility of the MacArthur—Bates Communicative Development Inventory in identifying language abilities of late-talking and typically developing toddlers. *American Journal of Speech-Language Pathology*, *14*(1), 40-51.

Jones, S. S. (2003). Late talkers show no shape bias in a novel name extension task. *Developmental Science*, *6*(5), 477-483.

Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive development*, *3*(3), 299-321.

Perry, L. K., & Kucker, S. C. (2019). The heterogeneity of word learning biases in late-talking children. *Journal of Speech, Language, and Hearing Research*, *62*(3), 554-563.

Rescorla, L. (2000). Do late-talking toddlers turn out to have reading difficulties a decade later? *Annals of dyslexia*, *50*(1), 85-102.

Smith, L. B. (2000). Learning how to learn words: An associative crane. In *Becoming a word learner*. New York: Oxford University Press.

Thordardottir, E. (2011). The relationship between bilingual exposure and vocabulary development. *International Journal of Bilingualism*, *15*(4), 426-445.

Weismer, S. E., Venker, C. E., Evans, J. L., & Moyle, M. J. (2013). Fast mapping in late-talking toddlers. *Applied Psycholinguistics*, *34*(1), 69-89.

Bayesian Pragmatics Provides the Best Quantitative Model of Context Effects on Word Meaning in EEG and Cloze Data

Markus Werning[1*§], Matthias Unterhuber[1*§], and Gregor Wiedemann[2]
[*]Corresponding authors: {matthias.unterhuber, markus.werning}@rub.de
[1] Department of Philosophy, Ruhr University Bochum
[2] Department of Informatics, University of Hamburg

Abstract

We contrast three views of how words contribute to a listener's understanding of a sentence and compare corresponding quantitative models of how the listener's probabilistic prediction on sentence completion is affected in online comprehension. The Semantic Similarity Model presupposes that the predictor of a word given a preceding discourse is their semantic similarity. The Relevance Model maintains that utterances are chosen to maximize relevance. The Bayesian Pragmatic Model assumes a relevance-guided modulation of a word's lexical meaning that can be regarded as a Bayesian update of statistical regularities stored in memory. In addition to a Cloze test, we perform an EEG study, recording the event-related potential on the predicted word and take the N400 component to be inversely correlated with the word's predictive probability. In a multiple regression analysis, we compare the three models with regard to Cloze values and N400 amplitudes. The Bayesian Pragmatic Model best explains the data.

Keywords: Bayesian Pragmatics, EEG, N400, Cloze Test, Semantic Similarity, Relevance, Generative Lexicon, Probabilistic Prediction, Online Comprehension, Modulation, Predictive Completion Task

Introduction

A preceding discourse can influence the way words contribute to a listener's understanding of a sentence (Recanati, 2012). This contextual influence also affects a listener's implicit probabilistic predictions on the completion of a discourse in the process of online comprehension. The listener's implicit task of predicting the next word w_{n+1} following a discourse that consists of the word sequence $w_1 \dots w_n$ can be described by a predictive probability of the following form:

(1) $P(w_{n+1}|w_1 \dots w_n)$.

In an EEG study, Nieuwland and Van Berkum (2006) showed that discourse contexts can interact with the lexical animacy feature of concrete nouns. A preceding context with a peanut being fictitiously described as dancing and singing can, e.g., invert comparative predictive probabilities such that the predicate *(was) in love* now has a higher probability for the listener than the otherwise more likely predicate *(was) salted*. The inversion of comparative predictive probabilities was revealed by a crossing-over of the N400 components between the two conditions measured on the critical predicates. The N400 component is defined as a negatively-going deflection of the event related potential over centro-parietal electrodes occurring around 400ms after stimulus onset (Kutas & Federmeier, 2011). As reviewed by Kuperberg and Jaeger (2016), the N400 component measured on a word is typically inversely correlated with its conditional probability given the preceding context.[1] There are two dominant interpretations of the underlying neuro-cognitive functions reflected in the N400 component: the semantic integration (e.g., Hagoort, Baggio, Willems, 2009) and the lexical retrieval view (e.g., Brouwer, Crocker, Venhuizen, & Hoeks, 2017). On these views, a negative increase of the N400 amplitude reflects higher processing demands associated with either (a) the integration of the target word's meaning into the compositional meaning of a sentence, or, respectively, (b) the retrieval of the target word's lexical meaning from memory.

Neural network models of the N400's underlying neural mechanisms have been proposed (e.g., Brouwer et al., 2017), some of which take the correlation of the N400 component with its predictive probability as a key explanandum (Rabovsky, Hansen, & McClelland, 2018; Fitz and Chang, to appear).

What has remained unclear in the peanut study is which factors of the context are responsible for the crossing over of the N400 components. The Semantic Similarity View maintains that the contextual influence is due to the degree of the semantic similarity between parts of the discourse context and the words in the target sentence (Otten & Van Berkum, 2008). The semantic similarity between two expressions can be determined by statistical regularities on co-occurrences in large corpora, as described in Distributional Semantics. In the above example the expression *in love*, e.g., has a greater semantic similarity to words in the context story (e.g., *dancing*, *singing*) than *salted*.[2]

The Relevance View, in contrast, holds that the crossing-over and the associated changes in the listeners' predictions

§ MW and MU contributed equally.

[1] A problem for this interpretation of the N400 are so-called semantic illusions, where zero-Cloze cases do not yield an increase in the N400 amplitude (cf. Kuperberg, 2007; Bornkessel-Schlesewsky & Schlesewsky, 2008; Brouwer, Fitz, & Hoeks, 2012).

[2] Existing accounts in Distributional Semantics base semantic similarity values merely on co-occurrences of syntactically unstructured lexical primitives and have problems coping, especially, with certain logical contexts such as negation or negative quantifiers. It would thus not be surprising if semantic similarity values attained by existing Distributional Semantics accounts had problems predicting the N400 in those contexts (e.g., Urbach & Kutas, 2010; Nieuwland, 2016). However, ongoing research in Distributional Semantics attempts to also capture complex logical contexts.

	+ALex	−ALex
AStdCtx	Maria richtet ein Kuchenbuffet her, das ihre Freunde beeindrucken soll, und bereitet alles Notwendige dafür vor. *Maria prepares a cake buffet to impress her friends and makes ready everything necessary for it.* Sie ist schon dabei **Sahne** zu **schlagen**. *She is already about **cream** to **whip**.*	Maria richtet ein Kuchenbuffet her, das ihre Freunde beeindrucken soll, und bereitet alles Notwendige dafür vor. *Maria prepares a cake buffet to impress her friends and makes ready everything necessary for it.* Sie ist schon dabei **Sahne** zu **zeichnen**. *She is already about **cream** to **draw**.*
ANewCtx	Maria übt für ein Bild von einem Kuchenbuffet und benutzt ihr Notizbuch für ihre Vorstudie. *Maria practices for a picture of a cake buffet and uses her notebook for her preliminary study.* Sie ist schon dabei **Sahne** zu **schlagen**.] *She is already about **cream** to **whip**.*	Maria übt für ein Bild von einem Kuchenbuffet und benutzt ihr Notizbuch für ihre Vorstudie. *Maria practices for a picture of a cake buffet and uses her notebook for her preliminary study.* Sie ist schon dabei **Sahne** zu **zeichnen**. *She is already about **cream** to **draw**.*

Table 1. Example of stimuli with English translation (Experiments 1 & 2). In the 2 × 2 design, Agentive (+Alex) and Non-Agentive (– Alex) verbs are combined with a standard (AStdCtx) or a new context (ANewCtx). The word order of the target sentence in the English translation is adjusted to the German original.

are only due to relevance considerations, regarding the relation between the context and the target expression. In the process of comprehension, the listener may assume that the speaker has chosen a particular combination of words in the discourse to maximize relevance (Sperber & Wilson, 1996). In the above example, the preceding fictitious story appears to be more relevant if the noun *peanut* is interpreted as animate. Consequently, a completion with the predicate *in love* should be more probable than one with *salted*. The Relevance View might also be associated with the view that there are no identifiable word meanings in the mental lexicon, at all (Elman, 2004). Consequently, the notion of semantic similarities between lexical meanings would be void.

Both, the Semantic Similarity and the Relevance View contrast with the Bayesian Pragmatic View. The latter accounts for the rational cooperation between speaker and listener by Bayes's Theorem. The predictive probability is identified with the posterior probability of a word, which results from updating its prior probability with its likelihood. The prior is simply a function of the semantic similarity (reflecting overall statistical co-occurrences) between the target word and the words in the preceding context. By being able to update the prior, listeners can incorporate pragmatic considerations on speakers' intentions, such as the thrive for relevance, into to their interpretation and, consequently, adjust their predictions about speakers' continuation of a sentence. Bayesian pragmatics has been successfully used to explain results in behavioral experiments on simple referential games (Frank & Goodman, 2012), scalar implicatures (Degen, Tessler, & Goodman, 2015; Goodman & Stuhlmüller, 2013), gradable adjectives (Lassiter & Goodman, 2013; Qing & Franke, 2014), modal expressions (Lassiter & Goodman, 2015), and figurative meaning (Kao, Wu, Bergen, & Goodman, 2014). It has so far only been validated in EEG by Werning & Cosentino (2017).

In this paper, we test and compare the empirical adequacy of the three different views. For each of the views, a quantitative model of the listener's predictive probability of a word given a preceding discourse is developed. To feed the model with data, values for semantic similarity and relevance are attained. As a measure of semantic similarity, we use GloVe values (see below), a computer linguistic measure in the framework of Distributional Semantics. Values for relevance are collected via relevance ratings in an online questionnaire. To determine listeners' predictive probabilities, we perform two experiments with the same stimulus material: a forced-choice Cloze study employing an online questionnaire and an N400 study in EEG. In a multiple linear regression analysis, finally, the proportions of variance explained by each of the three models are compared with regard to both experiments.

Models

To design the experiment, we build on Pustejovsky's (1995) Generative Lexicon Theory, according to which the lexical entry of a concrete noun (e.g. *cake*) contains a "Qualia Structure", which, among others, specifies an Agentive component (e.g. *bake*). The Agentive component represents the typical way of bringing about the denoted object – it contrasts with the Telic component that relates to a typical purpose or function, e.g., *cake-eat* (Cosentino, Baggio, Kontinen, & Werning, 2017). Triggered by verbs like *begin* and *finish*, the Agentive component of a noun co-composes with the noun in sentence meaning composition (Werning, 2004, 2005). This explains why sentences such as (a) and (c) are typically understood as having the meaning of (b) and, respectively, (d):

(a) Granny finished the cake.
(b) Granny finished baking the cake.
(c) The artist began the statue.
(d) The artist began sculpturing the statue.

Since nouns often co-occur with verbs expressing their Agentive component – so-called Agentive verbs – the semantic similarity of a noun and the respective Agentive verb is usually high. For our stimuli, the high semantic similarity was explicitly confirmed by GloVe values (see below).

Each quadruple of our 2×2 experimental design $\{+\text{ALex}, -\text{ALex}\} \times \{\text{AStdCtx}, \text{ANewCtx}\}$ (see Table 1) was built around a fixed concrete noun n (*cream*). In condition +ALex as [opposed to −ALex] the critical word was chosen as a verb (*whip* [*draw*]) that expressed [did not express] the

Agentive component in the lexical entry of the preceding noun n and had a high [low] semantic similarity to it. In condition AStdCtx, a standard context preceded the target sentence, whereas in condition ANewCtx the preceding discourse sentence suggested a new way of bringing about the object. The semantic similarity between the verbs and each of the context sentences was held invariant over all four combinations of conditions.

The Semantic Similarity Model presupposes that the only predictor for the verb given its preceding discourse is the semantic similarity of the former to the latter (word frequency held constant). The predictive probability $P_n(v|c)$ of the verb v following the noun n given the preceding context sentence c is a monotonously increasing function f_n^+ of the semantic similarity $S(v,n)$ between the verb and the noun and the (however invariant) semantic similarity $S(v,c)$ between the verb and the context sentence c. Accordingly, the listener's predictive probability $P_n(v|c)$ hence comes to:

$$(2)\quad P_n(v|c) = f_n^+\,(S(v,n)).$$

The Relevance Model, in contrast, maintains that listeners assume that speakers aim at maximizing relevance by choosing their utterances. The sole predictor is the relevance of the situation expressed by the context sentence c for the action (expressed by the verb v) to be performed on the object (denoted by the noun n):

$$(3)\quad P_n(v|c) = g_n^+\,(Rel_n(c,v)).$$

The Bayesian Pragmatic Model allows listeners to update their priors regarding the verb following the noun with pragmatic considerations on speakers' intentions, thus, arriving at probabilistic predictions of the verb. The prior, i.e., $P_n(v)$, strictly increases with the semantic similarity between the verb v and the noun n. The update as described by the likelihood, i.e., $P_n(c|v)$, is modelled as the conditional probability of speakers' choice of a context c given their communicative intentions, namely, their intentions to attribute, to a protagonist, an action (denoted by the verb v) to be performed on a given object (denoted by the noun n). The speaker, in other words, has to choose a preceding context sentence (c) to let this action appear relevant such that the choice of c given v strictly increases with the relevance of c for v. This leads to the following identifications: $P(c|v) = g_n^+\,(Rel_n(c,v))$, $P_n(v) = f_n^+\,\big(S(v,n)\big)$, and by Bayes Theorem we get:

$$(4)\quad P_n(v|c) = K \cdot P_n(c|v)\,P_n(v) = K \cdot g_n^+\,(Rel_n(c,v)) \cdot f_n^+\,\big(S(v,n)\big).$$

The Bayesian update of the semantic similarity can be regarded as reflecting a relevance-guided modulation of the Agentive component in the lexical entry of the noun (Recanati, 2012).

Model Predictions. Both, Cloze values and the amplitude of the N400 component, can be assumed to correlate with the predictive probability $P_n(v|c)$. To model the two variables, we assume monotonous functional relations between $P_n(v|c)$ and the values of the Cloze test and, respectively, the N400 amplitude measured on v.

Account	Model
Bayesian Pragmatic View	$a\,Rel_n(c,v) + b\,S(v,n) + k$
Relevance View	$a\,Rel_n(c,v) + k$
Semantic Similarity View	$b\,S(v,n) + k$

Table 2. Linear parametric model predictions for Cloze values and the amplitude of the N400 component measured on the critical verb.

Logarithmization and subsequent linear approximation of the model predictions (2), (3) and (4) lead to the linear parametric model predictions described in Table 2. The negative logarithm of the predictive probability of a word has been interpreted as word surprisal and is directly correlated with the amplitude of the N400, measured on the word (Frank, Otten, Galli, & Vigliocco, 2015).

Experiment 1: Cloze Test

Method

Participants: Cloze Test. Forty German native speakers were recruited via Prolific. Three participants, who failed to have a high-school degree, and two further ones, who answered the test in below four minutes (estimated time: ten minutes), were excluded. Of the remaining thirty-five participants, 57.1% were male, 42.9% female. The average age was 30.20 years (SD=10.86).

Participants: Relevance Rating. Forty participants were recruited via Prolific. Of those, thirty-eight people were included which satisfied the requirement of being German native speakers and having attained a high school diploma.

General Material. For the forced-choice Cloze test and the relevance ratings, forty quadruples of the 2×2 design were generated in German (Table 1). Regarding the Cloze test, we used a forced-choice format to ensure that only Agentive verbs were available for comparison. The critical verbs did neither repeat, nor occur in context sentences.

$\langle +\text{ALex}, -\text{ALex}\rangle$-*verb pairs*. For each noun (e.g., *cream*), an +ALex verb (*whip*) and an −ALex verb (*draw*) were chosen so that the semantic similarity of the +ALex verb to the noun was .15 higher than that of the −ALex verb. $\langle +\text{ALex}, -\text{ALex}\rangle$-pairs did not differ significantly in frequency class (https://wortschatz.uni-leipzig.de/), nor in character length.

Context sentences. For each quadruple, a standard context was chosen (AStdCtx) that was highly relevant for the $\langle \text{n}, +\text{ALex}\rangle$ combination and less relevant for $\langle \text{n}, -\text{ALex}\rangle$. The new contexts (ANewCtx) were designed to reverse this order of relevance. The semantic similarity of $\langle \text{Context}, \text{Verb}\rangle$-pairs was not allowed to differ significantly over all four conditions. The character length of contexts AStdCtx and ANewCtx did not differ significantly either.

Fillers. Twenty filler discourses with congruent and twenty with incongruent $\langle \text{Noun}, \text{Verb}\rangle$-pairs were generated and had the same structure as the test material. The frequency

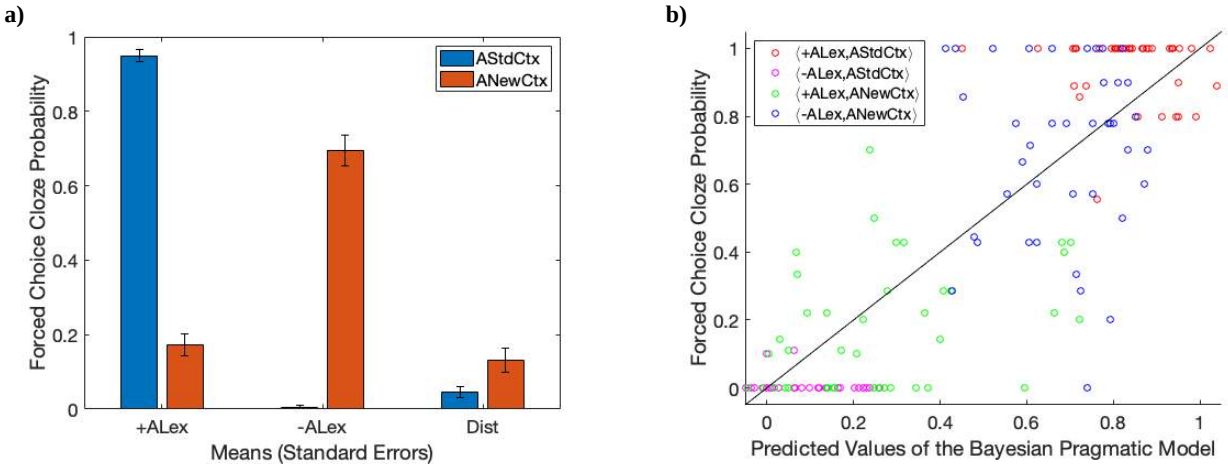


Figure 1. (a) Forced-choice Cloze probabilities. (b) Scatterplot for predictions of the Bayesian Pragmatic Model.

class and the length of the verbs did not differ significantly from the critical verbs of the test stimuli. Nor did the character length of the respective contexts differ significantly. *Semantic Similarity* was measured by an implementation of GloVe (Global Vectors, Pennington, Socher, & Manning, 2014) based on all articles from German Wikipedia and ca. three million news articles from Leipzig Corpora Collection (Goldhahn, Eckart, & Quasthoff, 2012).

Material and Design (Cloze Test). For the forced-choice Cloze test, the stimulus material was randomly distributed over four questionnaires in a counterbalanced way to avoid repetition. Context sentence and target sentence were presented as described in Table 1, with a blank instead of the critical verb. Subjects were instructed to choose the best fit among six alternative verbs. The alternatives always included the +ALex and −ALex verbs, as well as four alternative, agentive verbs drawn from the stimulus material not used in the questionnaire.

Material and Design (Relevance Ratings). The stimulus material including fillers was randomly distributed over four questionnaires in a counterbalanced way to avoid repetition. For each of the vignettes the context sentence surrounded by a box and marked as "A" was shown above the target sentence marked as "B", with the critical verb underlined. Participants were instructed to rate relevance on a 7-point Likert scale by answering the question: How plausible is the situation described in box A, given the action in sentence B.

Procedure. The Cloze test and relevance ratings were done using online questionnaires, via Qualtrics. The relevance ratings were preceded by two practice items.

Results and Discussion

Mean Cloze probabilities are summarized in Figure 1. The following results ensued: (1) $\langle +\text{ALex}, \text{AStdCtx}\rangle$ and $\langle -\text{ALex}, \text{AStdCtx}\rangle$ ($p < .0001$, $d = 9.01$), (2) $\langle -\text{ALex}, \text{ANewCtx}\rangle$ and $\langle -\text{ALex}, \text{AStdCtx}\rangle$ ($p < .0001$, $d = 2.65$), (3) $\langle +\text{ALex}, \text{AStdCtx}\rangle$ and $\langle +\text{ALex}, \text{ANewCtx}\rangle$ ($p < .0001$, $d = 3.88$), and (4) $\langle -\text{ALex}, \text{ANewCtx}\rangle$ and $\langle +\text{ALex}, \text{ANewCtx}\rangle$ ($p < .0001$, $d = 1.31$). To compare the Semantic Similarity, the Relevance and the Bayesian Pragmatic Models, we performed multiple regression analyses of Cloze data (see Table 3). The correlation between relevance ratings and corresponding similarity values was very small ($r = .10$).

The Bayesian Pragmatic Model was the clear winner according to BIC and AIC values, which take the unequal number of predictors into account. Within this model, the relevance values ($\beta = 1.01, p < .001$) and semantic similarity values ($\beta = .31, p < .01$) were significant predictors (for the scatterplot see Figure 1).

Experiment 2: EEG Study

Method

Participants. Twenty-eight participants were recruited. Two participants were excluded due to noisy EEG data and pain medication prior to the experiment. Of the resulting twenty-six participants, 30.8% were male, 69.2% female. The average age was 24.23 years (SD=3.40).

Design and Material. The stimulus material and fillers described in Experiment 1 were used. One of eight three-word phrases succeeded each target sentence (e.g., *und ist fröhlich* [and is happy]) to avoid the critical word being sentence final. Half of those appended phrases expressed positive and, respectively, negative emotional states, randomly chosen for each quadruple in a counterbalanced way.

The stimulus material was randomly split in two parts so that every context sentence and every critical verb appeared only once in each part. The two parts were administered to participants in two separate sessions, at least two weeks apart, to avoid carry-over and contrast effects. All participants had been presented with the entire stimulus material plus fillers. The order was counterbalanced. The complete set of fillers described in Experiment 1 was used in each session, resulting in eighty vignette (twenty per condition) and forty fillers (twenty congruent and twenty incongruent). The order was randomized.

Procedure. Each trial started with a fixation cross (1300ms) followed by the presentation of the context sentences. The context sentences were split up in roughly equal-sized chunks (character lengths: M = 17.91, SD =

Account	Model	N	df	RMSE	r	r_{adj}	BIC	ΔBIC	AIC	AICc
Predicted Variable: Cloze Probabilities (Experiment 1)										
Bayesian Pragmatic Model	Y~A+B+1	160	157	.22	.850*	.848*	-16.40	–	-25.64	-25.47
Relevance Model	Y~A+1	160	158	.23	.839*	.838*	-11.12	+5.30	-17.27	-17.19
Semantic Similarity Model	Y~B+1	160	158	.40	.223*	.209*	175.54	+191.93	169.38	169.46
Predicted Variable: EEG Amplitude (370–500ms, Experiment 2)										
Bayesian Pragmatic Model	Y~A+B+1	160	157	1.90	.489*	.479*	671.06	–	661.84	661.99
Relevance Model	Y~A+1	160	158	1.96	.426*	.419*	677.71	+6.64	671.56	671.63
Semantic Similarity Model	Y~B+1	160	158	2.08	.283*	.273*	696.26	+25.20	690.11	690.19

Table 3. Model comparisons for the regression analysis of Cloze probabilities (Experiment 1) and EEG amplitudes (370–500ms, Experiment 2). The EEG Amplitude was averaged across electrodes CP1, CPz, CP2, P1, Pz and P2. AICc = Aikaike Information Criterion corrected for sample size. *p<.001.

3.41). Each chunk was presented for 1300ms with a random interval of 200–400ms.

The target sentence, including the three-word phrase, was then presented word by word. In order to allow for a sufficiently long time interval for measuring ERP components, the target word was always presented for 450ms followed by an inter-stimulus interval of 450ms. All other words in the target sentence were presented for either 400 or 450ms, with a random inter-stimulus interval of 250–450ms. 700ms after the vignette's last word, participants had to judge, on a keypad, whether the complete vignette – consisting of context and target sentence (including the three-word phrase) – was plausible. The left-right orientation was randomized. The plausibility judgment should ensure that participants read the stimulus material carefully. Three examples with clearly congruent and incongruent ⟨Noun, Verb⟩-pairs served as practice material, prior to the testing phase.

Electroencephalogram recording and data processing. A BrainAmp Acticap system was used to record the electroencephalogram (EEG) from 66 active electrodes including four electro-oculogram electrodes for monitoring horizontal and vertical eye movements. The Brain Vision Analyzer 2.0 was employed to filter the data, correct for eye movements via independent component analysis (ICA). We used automatic artifact rejection to remove an episode per electrode if the change in currency exceeded 150μV in a 150ms interval. In case more than 10 electrodes were affected, or the trial had been interrupted before the critical word occurred, the whole episode was removed.

ERP data on the critical word (for each participant and trial) was exported to Matlab and individual trials were sorted and averaged across participants for each of the 160 vignettes, based on the Fieldtrip format for EEG data.

Results and Discussion

Mean amplitudes in the interval 370–500ms for posterior-central electrodes (CP1, CPz, CP2, P1, Pz, P2) across the four conditions are described in Figure 2. Bonferroni-corrected t-tests revealed significant differences between (1) $\langle +\text{ALex}, \text{AStdCtx}\rangle$ and $\langle -\text{ALex}, \text{AStdCtx}\rangle$ ($p < .01$, $d = 1.80$), (2) $\langle -\text{ALex}, \text{ANewCtx}\rangle$ and $\langle -\text{ALex}, \text{AStdCtx}\rangle$ ($p < .01$, $d = 1.13$), and (3) $\langle +\text{ALex}, \text{AStdCtx}\rangle$ and $\langle +\text{ALex}, \text{ANewCtx}\rangle$ ($p < .05$, $d = .61$). However, no significant difference was found for (4) $\langle -\text{ALex}, \text{ANewCtx}\rangle$ and $\langle +\text{ALex}, \text{ANewCtx}\rangle$ ($d = .17$). For the respective ERP results as measured on the CPz see Figure 3.

To compare the Semantic Similarity, the Relevance and the Bayesian Pragmatic Models, we performed multiple regression analyses of the mean, baseline-corrected ERP in

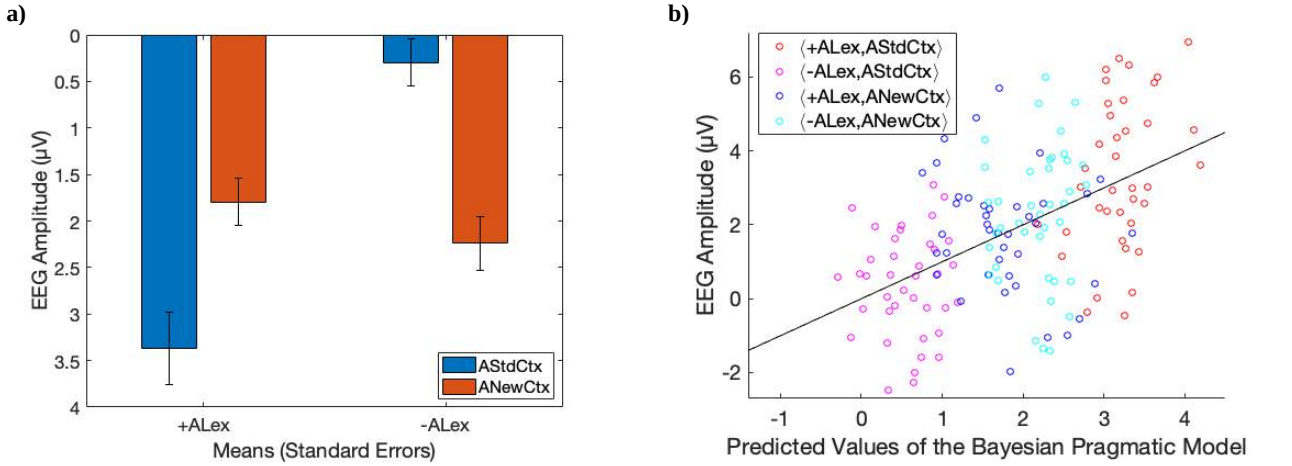


Figure 2. (a) EEG amplitudes (370–500ms, μV) for the pooled central-posterior electrodes (CP1, CPz, CP2, P1, Pz, and P2) after baseline correction. (b) Scatterplot for predictions of the Bayesian Pragmatic Model.

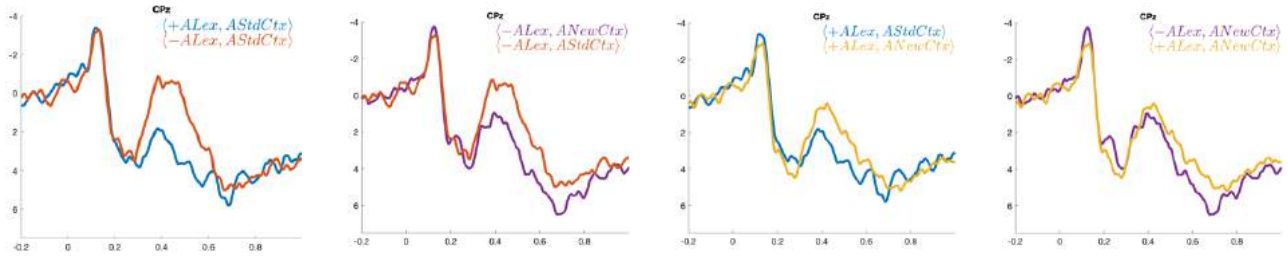


Figure 3. ERPs as measured on the CPz (µV).

the interval 370–500ms of the pooled posterior-central electrodes (see Table 3). Again, the Bayesian Pragmatic Model turned out as the clear winner according to BIC and AIC. Within this model, the relevance values ($\beta = 2.54, p < .0001$) as well as the semantic similarity values ($\beta = 2.81, p < .001$) were significant predictors. See Figure 2 for the scatterplot of the averaged N400 amplitudes and the values predicted by the Bayesian Pragmatic Model.

General Discussion

We contrasted three views of how words contribute to a listener's understanding of a sentence and compared three corresponding quantitative models of how the listener's implicit probabilistic predictions on the completion of a discourse is affected in online comprehension. The Semantic Similarity Model presupposes that the only predictor for a word given a preceding discourse is the semantic similarity between the two. The Relevance Model maintains that listeners assume that speakers aim at maximizing relevance by choosing their utterances. The Bayesian Pragmatic Model assumes a relevance-guided modulation of a word's lexical meaning that can be regarded as a Bayesian update of learnt statistical regularities stored in semantic memory.

To compare the explanatory power of the three models with regard to our Cloze and EEG data, we used the BIC and AIC criteria. Unlike r^2, BIC and AIC penalize for the unequal number of predictors. The clear winner in the comparisons, regarding both, the Cloze and EEG data, was the Bayesian Pragmatic Model. The Bayesian Pragmatic Model is not merely a combination of relevance and semantic similarity, but relates the two as the relevance-guided Bayesian update of a similarity-based prior probability. Interestingly, the two factors, relevance and similarity, did not contribute equally to the success of the Bayesian Pragmatic Model. Relevance outperformed semantic similarity, as indicated by the relative success of the Relevance over the Semantic Similarity Model. With regard to the EEG data, relevance explains 2.27 times as much variance as semantic similarity does. With regard to the Cloze data, however, this ratio is dramatically higher and equals 14.16. This pattern is also evidenced by comparing the EEG and the Cloze data with respect to the relative differences in BIC and AIC values of the three models. At first sight, this suggests that relevance is the dominant factor for the predictive probability of a word. A closer look reveals that semantic similarity still plays a larger role in truly incremental online comprehension, as observed in EEG, than in the Cloze test, which allows for backward-looking and untimed deliberation.

The lack of a significant difference of the ERPs in the time window of the N400 (370–500ms) for the comparison of the conditions ⟨+ALex, ANewCtx⟩ and ⟨−ALex, ANewCtx⟩ – with an effect size of only $d = .17$ – indicates that a greater semantic similarity value can compensate for a lower relevance value. This qualitatively illustrates the still prevalent importance of semantic similarity and thus the superiority of the Bayesian Pragmatic Model over the Relevance Model.

Our results also have clear implications for both the retrieval and the integration account of the N400. In light of our results, both approaches have to be modified to explicitly address the relevance-guided modulation of lexical meaning, described as updating by the Bayesian Pragmatic Model. Following the retrieval account, it will be the *modulated* lexical meaning of a preceding word that facilitates or impedes the retrieval of a subsequent word's meaning. Within the integration account, it is the *modulated* lexical meaning that determines the ease of integrating a subsequent word's meaning into the compositional meaning of the sentence.

Acknowledgments

We are grateful to the German Research Foundation for financial support through the grant WE 4984/4-1 as part of the Priority Program XPrag.de (SPP1727).

References

Bornkessel-Schlesewsky, I., & Schlesewsky, M. (2008). An alternative perspective on "semantic P600" effects in language comprehension. *Brain Research Reviews, 59*, 55–73.

Brouwer, H., Fitz, H., & Hoeks, J. (2012). Getting real about semantic illusions: Rethinking the functional role of the P600 in language comprehension. *Brain Research 1446*, 127–143.

Brouwer, H., Crocker, M. W, Venhuizen, N. J., Hoeks, J. (2017). A Neurocomputational model of the N400 and the P600 in language processing. *Cognitive Science, 41*(Suppl. 6), 1318–1352.

Cosentino, E., Baggio, G., Kontinen, J., & Werning, M. (2017). The time-course of sentence meaning composition. N400 effects of the interaction between context-induced and lexically stored affordances. *Frontiers in Psychology, 8*, 813.

Degen, J., Tessler, M. H., & Goodman, N. D. (2015). Wonky

worlds: Listeners revise world knowledge when utterances are odd. *Proceedings of the 37th Annual Conference of the Cognitive Science Society*, 548–553.

Elman, J. L. (2004). An alternative view of the mental lexicon. *Trends in Cognitive Sciences*, *8*(7), 301–306.

Fitz, H. & Chang, F. (to appear). Language ERPs reflect learning through prediction error propagation. *Cognitive Psychology*.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998.

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, *140*, 1–11.

Goldhahn, D., Eckart, T., & Quasthoff, U. (2012). Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*, 759–765.

Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science*, *5*(1), 173–184.

Hagoort, P., Baggio, G., and Willems, R. M. (2009). Semantic unification. In M. S. Gazzaniga (ed.), *The Cognitive Neurosciences* (pp. 819–836). Boston, MA: MIT Press.

Kao, J. T., Wu, J. Y., Bergen, L., & Goodman, N. D. (2014). Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences of the United States of America*, 12002–12007.

Kuperberg, G. R. (2007). Neural mechanisms of language comprehension: challenges to syntax. *Brain Research*, *1146*, 23–49.

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language Cognition & Neuroscience*, *31*(1), 32–59.

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology*, *62*, 621–647.

Lassiter, D., & Goodman, N. D. (2013). Context, scale structure, and statistics in the interpretation of positiveform adjectives. *Proceedings of SALT*, *23*, 587–610.

Lassiter, D., & Goodman, N. D. (2015). How many kinds of reasoning? Inference, probability, and natural language semantics. *Cognition*, *136*, 123–134.

Nieuwland, M. S. (2016). Quantification, prediction, and the online impact of sentence truth-value: Evidence from event-related potential. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(2), 316–334.

Nieuwland, M. S., & Van Berkum, J. J. A. (2006). When peanuts fall in love: N400 evidence for the power of discourse. *Journal of Cognitive Neuroscience*, *18*(7), 1098–1111.

Otten, M., & Van Berkum, J. J. (2008). Discourse-based Word Anticipation During Language Processing: Prediction or Priming? *Discourse Processes*, *45*(6), 464–496.

Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (pp. 1532–1543). Doha, Qatar: Association for Computational Linguistics.

Pustejovsky, J. (1995). *The Generative Lexicon*. Cambridge, MA: MIT Press.

Qing, C., & Franke, M. (2014). Gradable adjectives, vagueness, and optimal language use. *Proceedings of SALT*, *24*, 23–41.

Rabovsky, M., Hansen, S. S., & McClelland, J. L. (2018). Modelling the N400 brain potential as change in probabilistic representation of meaning. *Nature Human Behavior*, *2*, 693–705.

Recanati, F. (2012). Compositionality, Semantic Flexibility, and Context-Dependence. In M. Werning, W. Hinzen, & E. Machery (eds.), *Oxford Handbook of Compositionality* (pp. 175–191). Oxford: Oxford University Press.

Sperber, D., & Wilson, D. (1996). Precis of "relevance: communication and cognition." In H. Geirsson & M. Losonsky (eds.), *Readings in Language and Mind* (pp. 460–486). Oxford: Blackwell.

Urbach, T. P., & Kutas (2010). Quantifiers more or less quantify online: ERP evidence for partial incremental interpretation. *Journal of Memory and Language*, *63*(2), 158–179.

Werning, M. (2004). Compositionaltity, context, categories and the indeterminacy of translation. *Erkenntnis*, *60*(2), 145–178.

Werning, M. (2005). Right and wrong reasons for compositionality. In M. Werning, E. Machery, & G. Schurz (eds.), *The Compositionality of Meaning and Content* (Vol. I, pp. 285–309). Frankfurt: Ontos Verlag.

Werning, M., & Cosentino, E. (2017). The Interaction of Bayesian Pragmatics and Lexical Semantics in Linguistic Interpretation: Using Event-related Potentials to Investigate Hearers' Probabilistic Predictions. In *Proceedings of the 39th Annual Conference of the Cognitive Science Society*, 3504–3509.

A Trade-Off in Learning Across Levels of Abstraction in Adults and Children

Erika Wharton-Shukster (e.wharton.shukster@mail.utoronto.ca)
Department of Psychology, 100 St George St
Toronto, ON M5S 3G3 Canada

Amy Sue Finn (amy.finn@utoronto.ca)
Department of Psychology, 100 St George St
Toronto, ON M5S 3G3 Canada

Abstract

Learning about novel objects not only involves noticing information that makes the object unique, but also what makes objects the same. Yet, these two levels of learning involve different pieces of information, meaning that learning one well could come at the cost of the other. Moreover, children may categorize in a fundamentally different way, resulting in these levels of learning interacting differently. To investigate this, we had adults and children perform a categorization task followed by an item recognition test. We found that adults showed a trade-off, such that the ability to categorize items came at the cost of memory for those items. Using a subset of more unique lures, children's memory trended towards a trade-off with category learning. However, this was only observed among the older children. This suggests that adults' efficient learning comes at a cost, and this trade-off may start to appear in the elementary school years.

Keywords: cognitive development; category learning; abstraction; generalization; memory; selective attention

Background

Learning about objects goes beyond simply identifying unique features. We also learn abstract information, picking up on consistencies across objects. These different levels of abstraction are each informative in different ways, as the former provides details specific to an individual object while the latter facilitates categorization and generalization across objects. Given the discrepancy in the information learned, it seems likely that learning information really well at one level of abstraction might impede learning at the other, resulting in a trade-off in learning.

In fact, we know that categorical knowledge can impact memory in important ways. For instance, the Deese-Roediger-McDermott (DRM) paradigm has been found to show memory distortions in adults for words on a list when the presented words are categorically related. In this case, the category-level information that connects the words on the list causes memory distortions, a loss of memory of the specific words themselves (item-level information) and false memory for words that did not occur. This memory distortion is not seen when the word list is not categorical, suggesting that it is the abstraction of the category that is obscuring the details and producing distortions (see Brainerd, Reyna, & Ceci, 2008).

In addition, Sloutsky and Fisher (2004) found a drop in adults' memory for specific animals after participants sorted them categorically. When left uncategorized, memory for the individual animals was good, however following an induction task that required sorting the animals into species, memory for the individuals dropped to chance levels.

One possible explanation for this pattern of behaviour is the longstanding fuzzy-trace theory, in which abstraction necessarily involves a distillation of information to a vague, detail-free, gist representation (Brainerd & Reyna, 1990). In other words, abstraction is facilitated by a lack of detailed, item-level information. Similarly, the schema literature would suggest a comparable process in the organization of our knowledge, as schemas are abstract representations of something (be it a place, animal, social interaction) accumulated through experience that create expectations for the future (Mandler, 1984). Details from specific experiences are filtered out and consistencies are used to create a generic representation. Again, it is the loss of detailed information that makes a schema so generalizable.

Given this previous work and theory, there could be a trade-off in learning item-level and category-level information. Yet, work exploring this relationship has to date only included pre-existing categories and hasn't tackled whether a trade-off might occur *during* or in the service of category *learning*. It remains unclear how novel category learning would affect the interaction of item- and category-level information.

An eye-tracking study that assessed attention during a novel category learning task may provide some insight into this question (Rehder & Hoffman, 2005). In this study, as participants were learning to categorize the stimuli, they were found to fixate on all features of a stimulus. However, once they had successfully learned to categorize, they were found to fixate only on the diagnostic feature. This narrowing of focus would likely result in better categorization behavior, likely at the expense of learning about non-category relevant features of the objects, thereby producing a trade-off in object and category learning.

In contrast, when it comes to category learning, an abstract representation could still be formed not by ignoring irrelevant features of objects, but by learning *all* of the features of objects—both relevant and irrelevant—well. After all, abstracting to learn a category only necessitates learning the

relevant information for that category and not necessarily ignoring what is irrelevant. This approach to category learning would of course result in a different relationship with item memory: there would not be a trade-off.

There is some research suggesting that children may not demonstrate a trade-off in item-level and category-level information. For instance, in the DRM paradigm discussed above, adults consistently fall prey to false memories when the word lists are thematic. Interestingly, children do not succumb to the same memory distortions. In fact, this paradigm finds that young children have few false memories, with that number steadily increasing across the elementary school years and peaking in adulthood (Brainerd et al., 2008). In this case, children seem to have no thematic intrusions and remember the item-level information despite their category membership, resulting in no trade-off.

A similar result was found by Sloutsky and Fisher (2004), discussed above. Whereas the adults' memory for individual animals dropped after categorization, children's memory remained consistent. Here again, children maintained memory for the item information despite categorization. Given these findings, it is likely that children will similarly not show a trade-off during category *learning*, although the answer to this question is as yet unknown.

Along these lines, it has been suggested that instead of utilizing only the diagnostic dimensions, children categorize by including all item-level information (Sloutsky, 2010). Indeed, support for a holistic approach to categorization was found in children but not adults; adults were found to use only diagnostic features to categorize, while children were found to use the entirety of the item, basing their categorization on overall similarity (Smith & Kemler, 1977). This difference may reflect children's developing ability to selectively attend, as they are less successful at suppressing irrelevant information (Rueda, Posner, & Rothbart, 2005).

Furthermore, these differences in attention are likely to impact memory. For instance, during change detection and search tasks, children have also been found to have superior memory for task-irrelevant information compared to adults, suggesting that children's distributed attention facilitates memory for task-irrelevant information (Plebanek & Sloutsky, 2017). Together, these findings raise a final question. If children are attending to all available information when learning to categorize, will they retain item-level information despite successfully learning to categorize? In other words, might children be immune to the trade-off in item-level and category-level learning that we expect to see in adults?

To answer these questions, two experiments were performed, one with adults and one with children. In each, participants performed an A/B categorization task to measure category learning and a recognition memory test to measure item memory. To assess how specifically the items were remembered, half of the recognition foils were similar to the categorization stimuli along an orthogonal (not categorically-diagnostic) dimension, and half were dissimilar.

Experiment One

Methods

Participants Participants included 60 undergraduate students from the University of Toronto participating for course credit (M = 19.73 years, 76% female).

Materials and Procedure The category learning task consisted of 60 trial-unique trials of a feedback driven A/B sort task. Participants were instructed to sort "amoebas" into one of two categories based on the feedback given. They were not told what features defined category membership. Each stimulus was presented for 1.5 seconds or until a response was given, and stimulus presentation order was randomized between participants. The task was conducted on an Apple desktop computer using PsychoPy (Peirce, 2008).

The stimuli were designed to vary categorically along one dimension and orthogonally along two dimensions. Category membership was defined by distortions of two prototypical dot patterns shown below (Fried & Holyoak, 1984; Seger et al., 2000; Figure 1a). The 84 exemplars were generated by allowing dots a 7% chance of differing from the original. No exemplars were repeated across stimuli.

The orthogonal dimensions included colour and shape. Unique colours were randomly assigned, and shapes were created by making two extremely different shapes—generated from images of paint splatter—and morphing them together to varying degrees to create a series of related shapes (Figure 1b). All three dimensions were combined by placing the dot pattern in black on the coloured shape to create a total of eighty-four unique items (Figure 1d).

The item memory task consisted of a surprise item recognition test that always took place after the categorization task. In this recognition task, participants were asked if stimuli were present in the categorization task (old) or were new. Of the 48 stimuli presented at test, 24 were old, 12 were novel-shape lures, and 12 were same-shape lures. *Same-shape lures* were generated from the morphing procedure that was used to generate the categorization stimuli, but all twelve were unique and had not occurred during the categorization phase. Novel shape lures were created outside of the shape space used to generate the categorization stimuli, but were likewise generated from paint splatter images (see Figure 1c for examples). Order of presentation was randomized, and there was no time limit for response.

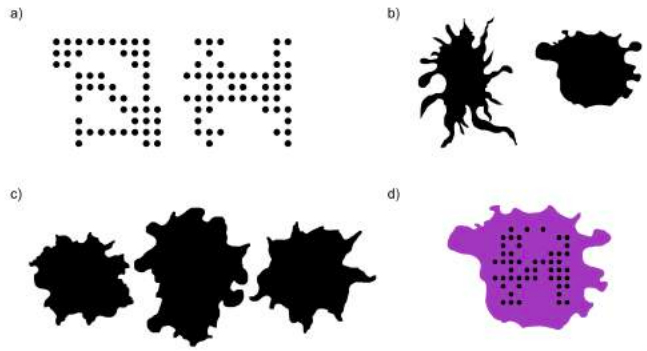


Figure 1: a) Diagnostic features defining category membership, b) Two shapes were morphed together to create a shape space, c) Novel-shape lures: Shapes were created outside of the shape space, d) An example of a complete stimulus: unique colours, shapes, and dot patterns were combined to create a set of completely unique stimuli.

Statistical Analysis We conducted all statistical analyses in R (R Core Team, 2017). Categorization accuracy was operationalized by calculating percentage correct in the category learning task. Item memory was calculated using d' (Z(hit rate) – Z(false alarm rate)), and scores were compared with chance using an independent-samples t-test. The general linear model was applied to all basic correlations, and general linear mixed-effects models were applied for analyses involving trial number using the lme4 package in R (Bates, Mächler, Bolker, & Walker, 2015). The fixed effects were categorization trial number and accuracy. Models contained random intercepts and slopes grouped by stimulus.

Results

Participants demonstrated learning in the categorization task, with accuracy increasing across trial number (β = 0.025, z = 9.793, p <.001) and an average overall accuracy of 76% (SD = 16.827). Similarly, participants demonstrated memory for the items at test, successfully distinguishing old items from new at a rate significantly different from 0 (M = 0.268, SD = 0.374; t(59) = 5.558, p <.001). As predicted, a tradeoff was also observed such that participants' categorization scores were negatively related to their recognition scores (F(1,58) = 14.31, p <.001, Figure 2). Thus, individuals who performed the categorization task better, had worse memory for exemplars.

To determine if memory was different for items that were categorized accurately from those that were not, we performed a t-test comparing the memory (d') for correctly and incorrectly categorized items. Memory was equivalent across correctly and incorrectly categorized items (t(495.18) = 0.463, p = 0.644). However, this relationship shifted over time, such that there was an interaction between memory for correctly and incorrectly categorized items and trial number (β = -0.016, z = -2.062, p = 0.039) such that memory for incorrectly categorized items moderately increased with an increasing number of trials (β = 0.001, z = 1.873, p = 0.061), and memory for correct trials moderately decreased with an increasing number of trials (β = -0.006, z = -1.648, p = 0.099).

To determine how specifically items were remembered, item memory was analyzed separately for novel-shape and same-shape lures by calculating d' using each as a unique false alarm score. A paired sample t-test determined that the two sets of scores were significantly different (t(59) = -7.643, p < .001). Using novel-shape lures, memory was significantly different from 0 (M = 0.7272, SD = 0.627; t(59) = 8.986, p <.001). Using same-shape lures, however, memory did not differ from 0 (M = -0.103, SD = 0.538; t(59) = -1.488, p = 0.142). Relating this to categorization performance across individuals, d' calculated using novel-shape lures was negatively correlated with categorization scores (F(1,58) = 25.7, p <.001), while d' calculated using same-shape lures was not (F(1,58) = 0.57, p = 0.453). Thus, the trade-off is only observed when using the novel-shape lures, for which there is evidence of memory.

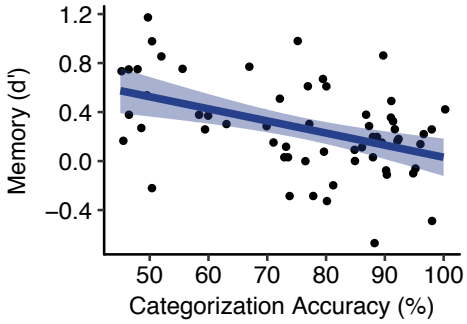


Figure 2: Item memory (d') by category accuracy (%). Each dot is an individual, the line signifies the slope and shading indicates standard error of the mean.

Discussion

As a group, participants successfully learned to categorize and remembered the items at post-test. Individual difference scores showed a trade-off between levels of learning such that those who performed well at the categorization task, performed more poorly at the item recognition task.

Participants failed to exhibit memory in comparison to the same-shape lures, but they did demonstrate memory when the lures were more distinct. Taken together, these data show that for adults, learning to categorize well impedes memory for items.

Experiment Two

Methods

Participants Participants included 61 children between the ages of 5- and 8- years old (M = 6.42 years, 48% female) recruited at a science museum. Exclusion criteria included lack of English skills to understand instructions, with one child meeting exclusion criteria.

Materials and Procedure The same two tasks were completed as in Experiment One, with a different, age-appropriate cover task. For the category learning task,

participants were told to sort two alien families onto their correct spaceship. Stimuli were presented for 3 seconds or until the participant responded. For the item memory test, 32 stimuli were presented randomly. Again, half were previously seen and half were new. Of the new, half were novel-shape lures and half were same-shape lures. Children were asked to verbally confirm their response after each button press. Tasks were completed on an Apple laptop using PsychoPy (Peirce, 2008).

Statistical Analysis We conducted the same analyses as Experiment One, as well as independent samples t-tests comparing adult and child scores.

Results

Children demonstrated learning in the categorization task, with accuracy increasing across trial number (β = 0.015, z = 6.619, p <.001). However, accuracy was significantly lower than adults (M = 65.889, SD = 17.664; β = -0.073, t = -3.399, p <.001, Figure 3). Children's item memory was poor: d' did not differ from 0 (M = 0.065, SD = 0.515; t(59) = 0.984, p = .329). Their item memory (d') was also significantly worse than adults (t(110) = 2.743, p = .007, Figure 4).

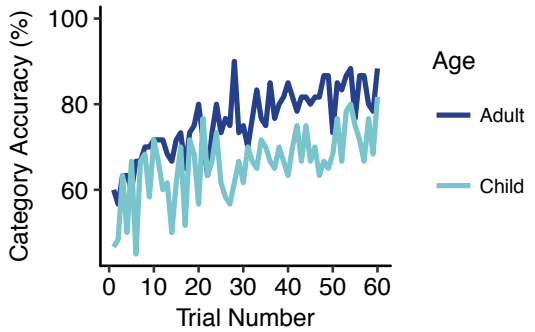


Figure 3: Category accuracy (%) by trial number and age group. The lines signify average group accuracy by trial. The dark blue line signifies the adult group, and the light blue line signifies the child group.

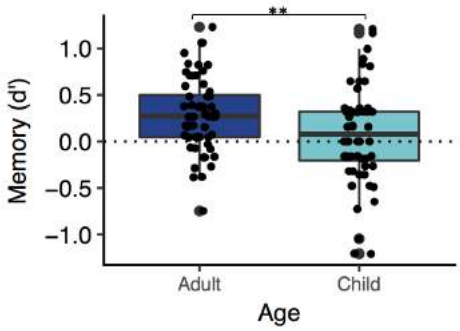


Figure 4: Item memory (d') by age group. d' is plotted separately for adults (dark blue) and children (light blue). The boxes signify the interquartile range and the whiskers signify the first quartile and below and the third quartile and above, respectively. The dotted line signifies chance, or no evidence of memory, and individual dots represent participants.

No trade-off was found between children's categorization accuracy and item memory (d') (F(1,58) = 1.766, p = .1891, Figure 5). Moreover, children's age was not found to interact with this relationship (F(3,56) = 0.647, p = 0.5885).

To determine if memory was different for items that were categorized accurately from those that were not, we performed a t-test comparing memory (d') for correctly and incorrectly categorized items. While not significant, we found a marginal difference in memory for correctly and incorrectly categorized items (t(690.2) = 1.950, p = .052), such that incorrectly categorized items were remembered moderately better. In addition, this relationship showed a trend toward shifting over time, with a trending interaction between time and accuracy (β = 0.013, z = 1.660, p = .097). Incorrectly categorized items were moderately better remembered at the beginning of the task, and correctly categorized items at the end of the task. While not significant, it is important to note that this relationship is the opposite pattern of that observed in adults.

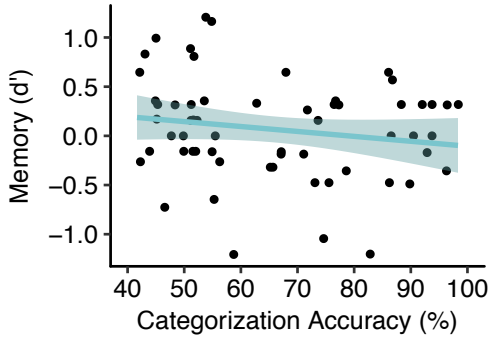


Figure 5: Child item memory (d') by category accuracy (%). Each dot is an individual, the line signifies the slope and shading indicates standard error of the mean.

To determine how specifically items were remembered, d' was calculated twice, once with each type of lure. Using a paired sample t-test, the two sets of scores were found to be significantly different (t(59) = -3.302, p = 0.002). When calculated with only novel-shape lures, memory was significantly different from 0 (M = 0.2701, SD = 0.822; t(59) = 2.547, p = 0.013), but memory did not differ from 0 when calculated with only the same-shape lures (M = -0.092, SD = 0.526; t(59) = -1.361, p = .179). While this is a similar pattern to that found in the adults, when compared to adults, the former was significantly lower (t(110.22) = 3.43, p <.001, Figure 6). Thus, children did show memory, but needed more distinct lures to demonstrate it, and it was poorer than that observed in the adults.

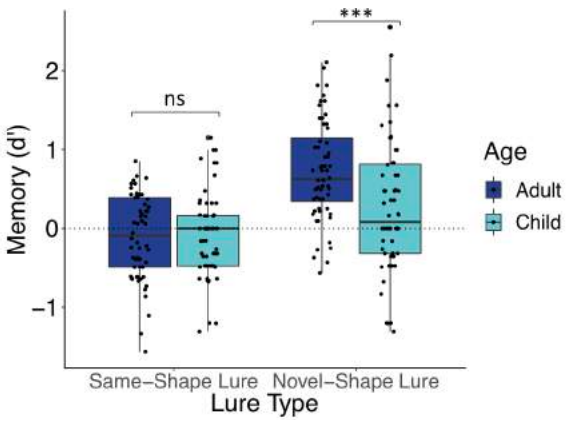


Figure 6: Item memory (d') calculated with novel- and same-shape lures across age group. The boxes signify the interquartile range and the whiskers signify the first quartile and below and the third quartile and above, respectively. The dotted line signifies chance, or no evidence of memory, and individual dots represent participants. The two left-hand boxes signify scores calculated with same-shape lures and those on the right signify scores calculated with novel-shape lures.

Looking at individual differences, d' calculated using novel-shape lures was not significant but had a moderate effect trending towards a trade-off with categorization accuracy, as more successful category learners had worse memory ($F(1,58) = 3.504$, $p = .0663$). No relationship was found between d' calculated with same-shape lures and categorization accuracy ($F(1,58) = 0.07289$, $p = .788$). The trade-off did not interact with children's age for d' calculated with the novel-shape lures ($F(3,56) = 1.618$, $p = 0.196$) or same-shape lures ($F(3,56) = 0.591$, $p = 0.623$). Nonetheless, to further assess the impact of children's age on the trade-off, we broke the children into two age groups: 35 5- and 6-year old children (young) and 25 7- and 8-year old children (old). Upon analyzing the trade-off in each group, the moderate effect found in the novel-shape lures continued in the old children ($F(1,23) = 3.303$, $p = 0.082$), but disappeared in the younger children ($F(1,33) = 0.790$, $p = 0.380$, Figure 7). No relationship between categorization accuracy and d' calculated with the same-shape lures was found in the old ($F(1,23) = 0.094$, $p = 0.762$) or young children ($F(1,33) = 0.182$, $p = 0.672$).

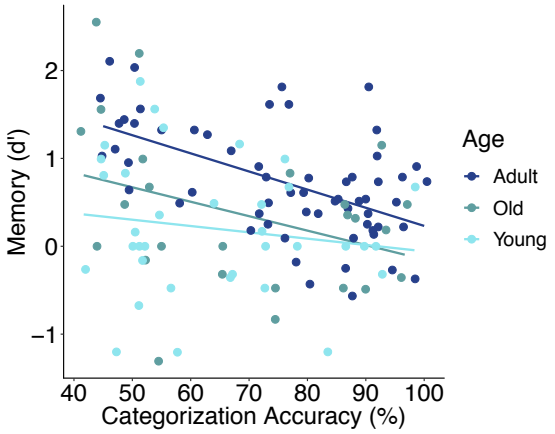


Figure 7: Item memory (d') by category accuracy (%) across adults, 7-8 year olds ("Old"), and 5-6 year olds ("Young"). Each dot is an individual, and the lines signify the slope. Adults are represented in dark blue, older children in teal, and young children in light blue.

Discussion

Although children learned to categorize, their memory for items was very poor overall. Like adults, children had no memory when compared to the same-shape lures but demonstrated memory when compared to the novel-shape lures, suggesting that they could only distinguish new and old items when the lures were distinctive.

Item memory did not significantly predict category learning, which is not surprising given how low children's memory was. However, a moderate effect was observed when d' was calculated with only the novel-shape lures, such that category learning scores were negatively correlated with memory (d'). Finally, this trend was only present in the older, 7- and 8-year old children but was not found in the younger, 5- and 6-year old children.

General Discussion

Building off of prior research showing a trade-off in pre-existing category knowledge and item memory, we found that a trade-off also occurs during the process of learning new categorical structures. For the adults, there was a cost to category learning, as those who performed well at the category learning task demonstrated worse memory for the items at post-test. This effect was driven by the more distinct lures, as memory for the similar lures was overall quite poor. In comparison, the children did not show a trade-off in category learning and item learning. However, when d' was calculated using only the distinct lures, a moderate effect was observed. This effect was found to be driven by the older, 7- and 8-year old children. When divided into two age groups, the moderate trade-off was observed in the older children, but there was no trade-off found in the younger, 5- and 6-year olds. These data may point to adults and young children approaching categorization in different ways.

While the mechanism at play remains unclear, a selective attention account provides an explanation for the pattern of data observed in adults. Selective attention is a top-down

process that not only directs attention to *relevant* information, but also suppresses *irrelevant* information (Pashler, Johnston, & Ruthruff, 2001). In a categorization task, the relevant information is the diagnostic feature, while the remaining information is irrelevant. Indeed, as discussed above, upon learning what defines a category, learners have been shown to fixate their gaze on the diagnostic feature (Rehder & Hoffman, 2005). In the current task, successful category learners likely fixated on the dot pattern while suppressing the irrelevant, item-level information. Given that unattended information is not remembered well (Simons, 2000), this could explain the successful learners' poor memory performance. Conversely, the poor category learners may have failed to learn which feature was diagnostic, thereby never selectively attending to it and, thereby continuing to attend to item-level information.

In comparison, it is less clear what mechanism best accounts for the children's pattern of behaviour, as they displayed only a moderate trade-off and, more specifically, only in the older children when calculating memory using the most distinctive lures. One possibility is that the younger children are utilizing a holistic, similarity-based categorization style, leading to a lack of trade-off. In comparison, the older children may be beginning to shift from this categorization style towards a more adult-like style focused on a diagnostic feature. A developmental shift in this age group would account for the moderate trade-off observed.

Prior research suggests that young children may categorize based on overall item similarity (Smith & Kemler, 1977), and if this were the case, children would attend to all features equally instead of selectively attending to a single feature. As such, their item memory would not drop upon categorization. Interestingly, Smith and Kemler (1977) found that this approach to categorization was consistently used among 5-year olds, but results were more ambiguous among 8-year olds. Perhaps the ambiguity reflects the beginning of adult-like categorization, and explains the moderate trade-off that we see here.

Indeed, a shift away from holistic processing would reflect the developmental course of selective attention, as the ability to filter irrelevant information has been found to improve across the elementary school years (Enns & Akhtar, 1989). An increase in selective attention would facilitate a more adult-like approach and result in a trade-off between category learning and item memory. Future research assessing the role of selective attention and its developmental course on the trade-off would help clarify the mechanism behind the patterns of behaviour observed in each age group.

Interestingly, children's memory was quite poor overall, which is not aligned with a holistic processing approach. Prior studies found children to have superior memory to adults for item-level information since they processed more information overall (e.g., Sloutsky & Fisher, 2004; Plebanek & Sloutsky, 2017). However, it is also well established that children have poor memory compared to adults (Ghetti, Angelini, & Annunzio, 2008; Rubin, 2000), and the adult group's item memory was not particularly strong either. It may be the case that children learned to categorize in a different way than the adult group but did not have the memory capacity to demonstrate it.

Alternatively, children's poor memory may not reflect poor memory overall, but may be symptomatic of poor pattern separation. Work by Ngo, Newcombe, & Olson (2018) found that 4-year olds were significantly worse than 6-year olds and adults at distinguishing old items from very similar items, irrespective of overall memory scores. Due to the similarity across items in the current study, it is possible that the younger children were disproportionately unable to discriminate the items. While unclear at this time, boosting children's memory in the future by increasing discriminability between items would help us to better understand how category learning and item memory interact across development.

The different patterns of learning across categorization trials in the adult and child groups suggest that the groups could be using different learning strategies. First, adults remembered incorrectly categorized items better towards the end of the task, while children remembered them better towards the beginning and moderately better overall than correctly categorized items. It may be the case that with increased learning across trials, errors became rare and surprising to adults and were, thus, remembered better. In comparison, children's heightened memory for errors throughout may reflect their tendency to respond more reactively than adults (Chatham, Frank, & Munakata, 2009), the surprise of which could have a memory boosting effect throughout.

Second, adults remembered correctly categorized items better towards the beginning of the task while children remembered them better towards the end. Given our assertion of increased selective attention with category learning in the adult group, it follows that correctly categorized items would be remembered more poorly towards the end of the task after learning had occurred and irrelevant information became unattended. Since children showed the reverse pattern in memory, this may provide further support for a more holistic approach to categorization than one of selective attention. This pattern would suggest that children maintain distributed attention throughout the task, as their memory for task-irrelevant information does not drop. The boost in memory observed towards the end may be a simple recency effect. Whatever the explanation, these divergent patterns of learning show that adults' and children's online categorization performance impacts memory.

Across the lifespan, our approach to learning changes as our needs change. Children are still figuring out what information is important, so it makes sense that they would attend to much of the information available. On the other hand, adults have a good sense of what information to prioritize and so attend to only what they deem informative. Inevitably, this means that a certain amount of information is always going to be missed. These findings make clear that we are always only seeing a piece of the picture, but perhaps we did not all start out that way.

Acknowledgments

We would like to thank Meg Schlichting, Katherine Duncan, Tess Forest, Michael Dubois, Alexandra Decker, and the rest of the Finn lab for all your help and support. We would also like to thank the Ontario Science Centre, Child Study Centre, NSERC and SSHRC, as well as the reviewers for their helpful commentary and feedback. Finally, a special thank you to all the parents and children who made this research possible.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using **lme4**. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Brainerd, C. J., & Reyna, V. F. (1990). *Gist Is the Grist: Fuzzy-Trace Theory and the New lntuitionism. DEVELOPMENTAL REVIEW* (Vol. 10). Retrieved from https://ac-els-cdn-com.myaccess.library.utoronto.ca/027322979090003M/1-s2.0-027322979090003M-main.pdf?_tid=5ee0e9d6-0629-4cf8-845e-a1078571a776&acdnat=1535355885_93d0c33e7a0ae9b5767f328477ae24f1

Brainerd, C. J., Reyna, V. F., & Ceci, S. J. (2008). Developmental Reversals in False Memory: A Review of Data and Theory. https://doi.org/10.1037/0033-2909.134.3.343

Chatham, C. H., Frank, M. J., & Munakata, Y. (2009). Pupillometric and behavioral markers of a developmental shift in the temporal dynamics of cognitive control. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(14), 5529–33. https://doi.org/10.1073/pnas.0810002106

Enns, J. T., & Akhtar, N. (1989). A Developmental Study of Filtering in Visual Attention. *Child Development*, *60*(5), 1188–1199. Retrieved from https://www.jstor.org/stable/pdf/1130792.pdf?casa_token=Ocv6L18S3SAAAAAA:xIZmcqlfE9qULHcqa8ACTrYA34m3RD_JdC53X9APaBKZoVZou0T2AIre-14C5ygR4zo_lIZ9FWAG8KbIEIrei5a4sjvR0c6LxYYa_XwEa3EXeT7aTu8

Fried, L. S., & Holyoak, K. J. (1984). Induction of Category Distributions: A Framework for Classification Learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(2), 234–57. https://doi.org/10.1037/0278-7393.10.2.234

Ghetti, S., Angelini, L., & Annunzio, G. D. (2008). The Development of Recollection and Familiarity in Childhood and Adolescence : Evidence From the Dual-Process Signal Detection Model, *79*(2), 339–358.

Mandler, J. M. (1984). Stories, scripts, and scenes: Aspects of schema theory. *Hillsdale: Erlbaum*.

Ngo, C. T., Newcombe, N. S., & Olson, I. R. (2018). The Ontogeny of Relational Memory and Pattern Separation. *Developmental Science*, *21*(2). https://doi.org/10.1111/desc.12556

Pashler, H., Johnston, J. C., & Ruthruff, E. (2001). Attention and Performance. *Annual Review of Psychology*, *52*(1), 629–651. https://doi.org/10.1146/annurev.psych.52.1.629

Peirce, J. W. (2008). Generating stimuli for neuroscience using PsychoPy. *Frontiers in Neuroinformatics*, *2*, 10. https://doi.org/10.3389/neuro.11.010.2008

Plebanek, D. J., & Sloutsky, V. M. (2017). Costs of Selective Attention: When Children Notice What Adults Miss. *Psychological Science*, *28*(6), 723–732. https://doi.org/10.1177/0956797617693005

Rehder, B., & Hoffman, A. B. (2005). Eyetracking and selective attention in category learning. *Cognitive Psychology*, *51*(1), 1–41. https://doi.org/10.1016/J.COGPSYCH.2004.11.001

Rubin, D. C. (2000). The distribution of early childhood memories. *Memory*, *8*(4), 265–269. https://doi.org/10.1080/096582100406810

Rueda, M. R., Posner, M. I., & Rothbart, M. K. (2005). The Development of Executive Attention: Contributions to the Emergence of Self-Regulation. Retrieved from https://www.researchgate.net/profile/Maria_Rueda/publication/7617847_The_Development_of_Executive_Attention_Contributions_to_the_Emergence_of_Self-Regulation/links/09e41505b54a74a51f000000.pdf

Seger, C. a, Poldrack, R. a, Prabhakaran, V., Zhao, M., Glover, G. H., & Gabrieli, J. D. (2000). Hemispheric asymmetries and individual differences in visual concept learning as measured by functional MRI. *Neuropsychologia*, *38*(9), 1316–1324. https://doi.org/10.1016/S0028-3932(00)00014-2

Simons, D. J. (2000). Attentional capture and inattentional blindness. *Trends in Cognitive Sciences*, *4*(4), 147–155. https://doi.org/10.1016/S1364-6613(00)01455-8

Sloutsky, V. M. (2010). From Perceptual Categories to Concepts: What Develops? *Cognitive Science*, *34*(7), 1244–1286. https://doi.org/10.1111/j.1551-6709.2010.01129.x

Sloutsky, V. M., & Fisher, A. V. (2004). When Development and Learning Decrease Memory. *Psychological Science*, *15*(8), 553–558. https://doi.org/10.1111/j.0956-7976.2004.00718.x

Smith, L. B., & Kemler, D. G. (1977). Developmental trends in free classification: Evidence for a new conceptualization of perceptual development. *Journal of Experimental Child Psychology*, *24*(2), 279–298. https://doi.org/10.1016/0022-0965(77)90007-8

The Role of Prior Beliefs in The Rational Speech Act Model of Pragmatics: Exhaustivity as a Case Study

Ethan Wilcox[1] and **Benjamin Spector**[2]

[1]Department of Linguistics, Harvard University, wilcoxeg@g.harvard.edu
[2]Institut Jean Nicod, Département d'études cognitives, ENS, EHESS, PSL University, CNRS , benjamin.spector@ens.fr

Abstract

This paper examines the interaction between prior beliefs and pragmatic inferences, focusing on exhaustivity effects. We present three experiments that tests how prior beliefs influence both interpretation and production of language, and compare the results with the predictions of the Rational Speech Act model, a Bayesian model of linguistic interpretation. We find that prior beliefs about conditional probabilities have no affect on language production, but do affect interpretation, producing *anti-exhaustivity* effects. We find that the RSA model achieves a relatively good fit both for the human production and interpretation data, but only for highly-implausible utterance costs.

Keywords: Pragmatics, Rational Speech Act model, Exhaustivity.

Introduction

The interpretation of linguistic utterances in context depends on the prior beliefs of speakers and hearers. For instance, if someone says "Mary visited a cardiologist today", one will infer that Mary is more likely than a random person to have a heart-related medical condition. It is easy to account for such inferences as a probabilistic inference: the hearer starts with a prior probability distribution over possible world states, and then conditionalizes this distribution with the new information that Mary visited a cardiologist today. Typically, though, pragmatic inferences go well beyond what can be predicted with such a simple model of linguistic interpretation. They also involve, for instance, reasoning about *other sentences* that the speaker could have uttered given some assumptions about their communicative goals (Grice, 1975). For instance, if I'm asked a question such as *Among Peter, Mary and Sue, who attended the show today?*, an answer such as *Mary did* tends to trigger the inference that the others did not, even if there is no expectation that what any of them does depends on what the others do. Such *exhaustivity effects* are typically accounted for in terms of Grice's maxim of quantity: if in fact both Peter and Mary had attended the show, a knowledgeable speaker would say that rather than just talking about Mary.

Now, in some situations these two types of effects (effect of prior beliefs, exhaustivity effects) are pitted against each other. For instance, we might know that Peter and Mary are a couple, who usually go out together, so that, upon learning that Mary attended the show, one would assign a high probability to the possibility that Peter did too, which would go against the exhaustivity effect just mentioned.

The Rational Speech Act Model (RSA) is a model of pragmatic reasoning which integrates both the role of prior beliefs and that of pragmatic reasoning about alternative utterances (Frank & Goodman, 2012). It can in principle make very precise predictions about their interactions. The RSA model views the speaker as being engaged in a trade-off between two goals: maximizing informational content and minimizing the cognitive cost of an utterance. As we will see shortly, in the baseline RSA model, this trade-off is affected in a drastic way by the prior beliefs shared by listeners and speakers, to the extent that, in some situations, an *anti-exhaustivity effect* is predicted: in some cases, the utterance *Mary did*, in the above context, is expected to be the best message to use to convey that both Mary and Peter attended the show, and thus to be interpreted in this way. However, because an RSA model has several free parameters, it is difficult to assess a) whether it is compatible with a given set of data, and b) whether it provides an *explanatory* account of the data. The goal of this paper is to gather data about the effect of priors on exhaustivity effects, both for interpretation and production, and to assess how well the baseline RSA model can account for these data in a principled way.

Degen et al. have already tested the effect of priors on pragmatic interpretation within the RSA framework, focusing on a similar but different type of inference, namely the inference from *some* to *not all* (Degen, Tessler, & Goodman, 2015). We will discuss the relationship between our study and Degen et al. (2015) in the next section.

The Rational Speech Act Framework and exhaustivity effects

In the basic RSA model, we start from a *literal listener* L_0 who has a prior probability distribution over worlds and knows the literal meanings of sentences. When hearing an utterance u, L_0 updates her prior distribution by conditionalizing it with the proposition expressed by the literal meaning of u. Then we define a speaker S_1 who wants to communicate her beliefs to L_0 and knows how L_0 interprets sentences. S_1 is characterized by a utility function U_1 such that the utility of a message u if S_1 believes w is *increasing* with the probability that L_0 assigns to w after updating her distribution with u, and *decreasing* with the *cost* of u. A *rationality parameter* α determines the extent to which S_1 maximizes her utility. Next, we define a more sophisticated listener, L_1, who, when receiving a message u, uses Bayes's rule to update her prior distribution on worlds, under the assumption that the author of u is S_1. A speaker S_2 is then defined exactly like S_1, except that now S_2 assumes that she talks to L_1, not L_0. And so on.[1].

[1]See Bergen, Levy, and Goodman (2016) for the mathematical description of the model.

Now consider a case where world states are individuated by the truth-values of two propositions A and B (for instance *Mary attended* and *Peter attended*), and where the available utterances are A, B, A *and not* B, B *and not* B and A *and* B. Consider a situation where the speaker wants to communicate the world state $\{A\}$ (where A is true and B is false) . She can choose between the two messages A and A *and not* B. While A is less informative than A *and not* B, it has nevertheless a significant probability of use, because it is less costly. Upon hearing A, the first-level pragmatic listener L_1 will reason as follows, if the priors are sufficiently uniform across world states. The message is only compatible with two world states, namely $\{A\}$ And $\{A,B\}$. But the the speaker is more likely to mean $\{A\}$ than to mean $\{A,B\}$. If she wanted to communicate $\{A,B\}$, there were two other possible messages, namely B and A *and* B. B is furthermore no more costly than A, and A *and* B is costly but also more informative. In contrast with this, if she wanted to communicate $\{A\}$, there was only one other possible message, namely A *and not* B, and furthermore this message, while more informative, is very costly (more than A *and* B). As a result, it is likely that the intended meaning was in fact $\{A\}$, and the exhaustivity effect is derived.

However, things can change drastically with non-uniform priors. Imagine now a speaker who wants to communicate $\{A,B\}$. She has a choice between using the messages A, B, A *and* B. While the latter message is the most informative, it is also more costly than the two others. Suppose further that the prior conditional probability of B given A is very high. The literal listener L_0, upon hearing A, will assign a high probability to the world state $\{A,B\}$. In this case, A may turn out to have a higher utility than A *and* B for S_1: it is quite good at communicating the world state $\{A,B\}$ (given the priors), and it is less costly than A *and* B. Furthermore, with such non-uniform priors, a speaker who would want to communicate $\{A\}$ might be very unlikely to use the message A: despite the fact that A is less costly than A *and not* B, it is so poor at conveying the intended world state (due to the priors), that the speaker now has an extra incentive to use the costly sentence A *and not* B. Now, upon hearing A, the pragmatic listener L_1 will reason as follows. The intended world state is either $\{A\}$ or $\{A,B\}$. If the latter, S_1 was in fact quite likely to use A. If the former, the speaker was more likely to use A *and not* B. So the intended world state is probably $\{A,B\}$. This time an *anti-exhaustivity effect* is derived (Roni Katzir, p.c.). However, this prediction is highly sensitive to the values of the free parameters of the model (rationality, costs).

Degen et al. (2015) discuss a related case. The RSA model, under a broad range of values for the free parameters, predicts that when the conditional probability of an *all*-statement given the truth of the corresponding *some*-statement is very high, *some* is going to be used to convey *all* and to be so understood. Degen et al. consider a discourse such as: *Max threw fifteen marbles in the water. Some of the marbles sank*. Because we expect all marbles to sink, this is a case where the prior probability of $\forall$ (the world where all marble sank) is very high, and where the basic RSA model predicts that the sentence will in fact convey that all marbles sank. But the experimental results show that actual listeners typically derive a *some but not all*-reading. In Degen et al.'s model, unlike in the basic RSA model, the pragmatic listener is uncertain about the speaker's beliefs about the listener's priors. Even if $\forall$ has a very high prior probability for the listener, the pragmatic listener L_1 assigns a substantial probability to the possibility that the speaker believes that the literal listener L_0 is in fact entertaining uniform priors over world states. So the pragmatic listener L_1 has a higher-order prior probability distribution over the set of first-order prior distributions (over world-states) that the speaker might attribute to the literal listener L_0. When processing a sentence, this listener updates both her probability distribution over worlds and her higher-order probability distribution over the set of priors that the speaker is considering. The proposed model is such that when hearing *some*, the listener concludes that the speaker probably believes that the listener is using uniform priors, and as a result *some* ends up conveying $\exists\neg\forall$. Simulations show that in order to obtain this result, the pragmatic listener L_1 must view the speaker (S_1) as believing that there is a high probability that the literal listener's prior distribution over world states is uniform. For the range of values that are typically used in RSA models for α (somewhere between 1 and 10), this probability must be substantial (Degen et al. report that it has to be equal to .5 to achieve the best fit with experimental data). Given this, a conceptual limitation of this account is that it models the listener as believing that the speaker views the listener as likely to be unaware that marbles typically sink when thrown into water (despite the fact that the priors over world states that Degen et al. collected show that people do in fact expect that when marbles are thrown into water, they will all sink). But no empirical evidence is provided to support these assumptions, and so it is not clear that much is gained compared to a model that would simply ignore the actual priors and take as input relatively uniform priors.

Now, in the case of exhaustivity effects, the situation is even more extreme. In the *some-all* case, the *all* sentence is no more costly than the *some*-sentence. Because of this, even with extremely biased priors, a fully rational speaker would always choose *all* to convey *all*, since it is still more informative than *some*, and would never use *some* (*some but not all* would be used to convey $\exists\neg\forall$). For this reason, with very high values for α (corresponding to a very rational speaker), a correct result is derived in Degen et al.'s model, even if the probability that the speaker assigns to the possibility that the listener does not expect all marbles to sink is very low (but still positive). In the exhaustivity case, avoiding the anti-exhaustivity effect is harder, because the message A *and* B is more costly than the message A, and so will not necessarily be the message used by a fully rational speaker who believes A *and* B, if the prior conditional probability of B given A is very high (the gain in informativity provided by A *and* B compared to A might be too small to justify the extra cost). Even with

a fully rational speaker, for a broad range of reasonable cost values, there exist contexts where the speaker is predicted to use *A* to mean *A and B*. In this paper, we will compare the predictions of the baseline RSA model with experimental data pertaining to exhaustivity and anti-exhaustivity effects.

Independently of this theoretical goal, our contribution is to provide experimental data pertaining to cases where priors are biased against the exhaustive reading of a sentence *A* in the context of *Which of A and B is true?*.

Human Judgement Experiments

To test the effect of priors on human linguistic judgements, we conducted three online experiments. Each experiment involved a simple scenario in which a character was moving furniture from her apartment onto the street, and questions were asked about what the character was able to move or how she was likely to report the progress of her moving to a friend. Experiments were hosted on IbexFarm. Participants were recruited on Amazon Mechanical Turk.[2]

Experiment 1: Priors

As this work aims to test the effect of priors on human linguistic judgements, our first experiment gathered prior probabilities for two scenarios, which were used in later experiments. In the priors experiment subjects were shown a scenario in which a character is moving her apartment and tests the weight of two furniture items. The character picks up one item, at which point respondents were asked whether they thought she could pick up the second item as well. Input format were forced-choice, yes/no radio buttons. The experiment was divided into two conditions: In the first, High Conditional Probability condition, the character was shown picking up a chair and asked whether she could also pick up a footstool, which was visually about half the size. In the second, Low Conditional Probability condition participants saw the character picking up the footstool and asked if they thought she could also pick up the chair. Participants were asked two simple comprehension questions at the end of the experiment, and only responses from participants who answered both correctly were used. We collected 60 responses, of which 57 (95%) were usable. The proportion of respondents who selected yes in each condition was taken as the population-level prior on conditional probability in each case.

The results can be seen in Fig. 1, on the left-hand panel. Error bars represent binomial 95% confidence intervals using the binconf function in R on default settings (Wilson method). A Fishers Exact Test indicates that participants were significantly less likely to endorse the yes response in the Low Conditional Probability condition (p=0.02225).

Experiment 2: Elicitation

We conducted a second experiment to test the effect of priors on the elicitation of simple conjunctives. If humans subjects incorporate priors in their utterance and endorsement of simple conjunctives, then we expect the relative rate of the conjoined utterance ("A and B") to be lower in high-conditional probability contexts, where P(B‖A) is very high (because in this case the utterance A is quite good at communicating the $A \wedge B$ world state (which we will denote by $\{A,B\}$ henceforth). Furthermore, we also expect that, if they want to communicate the world $A \wedge \neg B$ (which we will now notate$\{A\}$), there will be less likely to use the message A in the high-probability condition, and more likely to use *A and not B*.

In this setup, participants were shown the same 'moving' scenario from the previous experiment, involving a chair, a footstool and a character who tells a friend that she would move 'everything I can' down to the curb. In the subsequent panel participants were shown the character with the furniture she was able to move depending on the condition to which the participant was assigned, which are enumerated in Table 1, along with the condition name and a *tag*, with which we refer to the condition in charts and figures. Participants are asked to endorse an utterance that they think the character would use to describe the situation to a friend, who has prior familiarity with the items, over the telephone. Input were force-choice radio buttons with six possible utterances: 'I moved the chair', 'I moved the footstool', 'I moved the chair but not the footstool', 'I moved the footstool but not the chair', 'I moved the chair and the footstool' and ' I moved the footstool and the chair.'

Experimental Stimuli	Tag	Condition Name
Chair + Footstool	{A,B}	[BOTH, HIGH PROB]
Chair	{A}	[SINGLE, HIGH PROB]
Footstool + Chair	{A,B}	[BOTH, LOW PROB]
Footstool	{A}	[SINGLE, LOW PROB]

Table 1: Elicitation Experimental Conditions

Following the critical question, we asked two simple comprehension questions and whether the participant was a native speaker of English. Only data from those respondents who answered both correctly and identified as a native English speaker were used. The experiment was given to 174 subjects, of which 126 (72.4%) answered the follow-up questions satisfactorily. A further 33 subjects were filtered as repeat subjects from one of our other experiments, bringing the total number of responses to 93.

The results from this experiment can be seen in Figure 1, in the middle panel, with world state on the x-axis and the proportion of "A and B" responses on the y-axis. Red dots represent proportion of "A and B" responses in the high probability condition, blue dots the low probability condition; error bars represent 95% confidence intervals. Endorsements of the "A and B" utterance were near floor in the $\{A\}$ world (m=0.02, m=0.11 in the High Probability and Low Probability conditions, respectively). However, the endorsements were not at ceiling in the $\{A,B\}$ world state (m=0.68, m=0.84 in the High Prob and Low Prob conditions, respectively).

[2]Experiments were pre-registered online at http://aspredicted.org/blind.php?x=7qm9pz

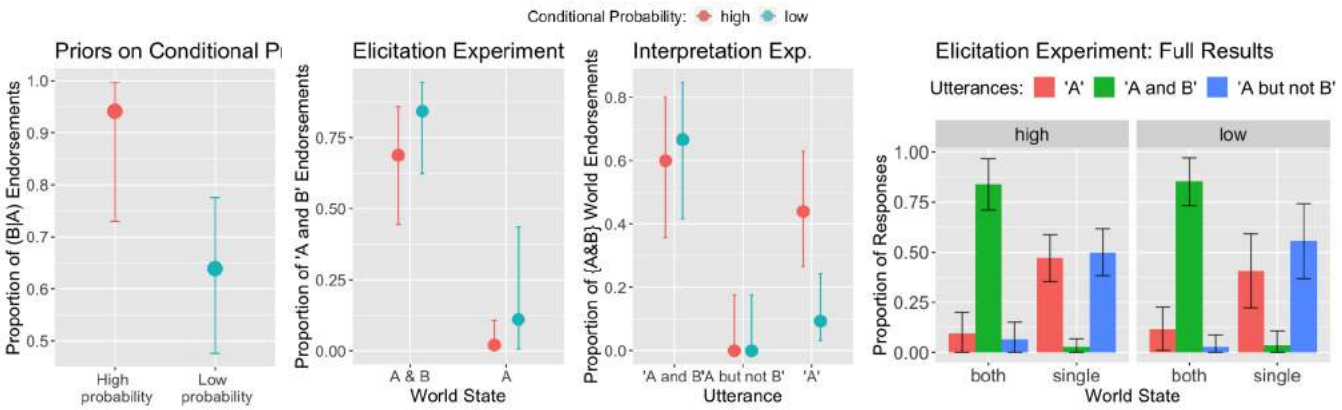


Figure 1: Human Judgements from the Online Study

To test whether priors on conditional probability had an effect on utterance endorsement, we fit a linear model to the data using the proportion of 'A and B' responses as our dependent measure, and experimental conditions as predictors, which were coded using 1/-1. We found a main effect of WORLD STATE, whereby participants were less likely to endorse "A and B" in the $\{A\}$ world ($p<0.001$), as expected. However, we found no interaction between world state and prior conditional probability ($p=0.507$), which is visually evident from the fact that both prior probability conditions fall within each other's confidence intervals. In fact, the relative rate of "A and B" endorsement in the high and low conditional probability conditions, ran counter to our expectations, with respondents marginally less likely to endorse the relevant utterance in the high conditional probability condition. Note that while the subjects in this study were willing to use 'A' to endorse the {A,B} world, their rates of endorsement in both conditions (between 15-32%) were well below their expectation of P({A,B}|{A}) (between 65-95%).

In addition to our pre-registered analyses, we conducted a follow-up analysis to assess whether the priors on conditional probability affected the rate of endorsements for the exhaustified utterances, 'A but not B.', in order to communicate the world $\{A\}$ The results for all utterance endorsements can be seen in Fig. 1, on the far right panel. As the conditional probability increases, we might expect the rate of endorsements for the exhaustified utterance (the blue bar) to increase in the SINGLE condition, given that the bare utterance, 'A' might be quite bad at communicating $A \wedge \neg B$. Our pre-registered analysis, which examines only the rate of endorsement for the 'A and B' utterance, would not capture these dynamics.

In order to assess the impact of conditional probability priors on the rate of the exhaustified utterance we fit a linear regression model using the proportion of exhaustified ('A but not B') utterance endorsements as our dependent variable and utterance types as our predictors. We found a main effect of world state ($p < 0.001$) whereby exhaustified utterances were less likely to be endorsed in the BOTH condition (as fully expected), but no significant interaction between world state and conditional probability ($p = 0.515$).

Experiment 3: Interpretation

Experimental Stimuli	Tag	Condition Name
"The chair and the footstool"	'A and B'	[BOTH, HIGH PROB]
"The footstool and the chair"	'A and B'	[BOTH, LOW PROB]
'The chair but not the footstool"	'A but not B'	[ONLY, HIGH PROB]
"The footstool but not the chair"	'A but not B'	[ONLY, LOW PROB]
"The chair"	'A	[SINGLE, HIGH PROB]
"The footstool	'A'	[SINGLE, LOW PROB]

Table 2: Interpretation Experimental Items

The third experiment aimed to test the effects of prior conditional probability on utterance interpretation. The RSA model predicts that human subjects will be more likely to interpret the utterance 'A' as referring to an {A,B} world in cases where P(B|A) is higher.

For this experiment, participants were shown the same 'moving' scene as in the others. A character commits to moving 'what I can lift' down to the curb, and tells a friend what she is capable of lifting depending on the condition to which the subject was assigned. There were six conditions, corresponding to six possible utterances: 'I can lift the chair and the footstool','I can lift the footstool and the chair', 'I can lift the chair but not the footstool', 'I can lift the footstool but not the chair', 'I can lift the chair', 'I can lift the footstool'(cf. Table 2).

In the subsequent slide, participants see the character by the curb, with a grayed-out area where the furniture would be, are told that the character 'has moved all the items she can lift down the curb', and are asked to select which items they believe have been moved down. They are provided with a visual reference of the furniture items, scaled to size, at the bottom of the screen. The input form was a check box, and

in the instructions to the experiment participants were told that they could check as many or as few of the boxes as they wished.

Following the critical question, participants were asked two comprehension questions and whether or not they were a native speaker of English. The survey was given to 475 participants of which 338 (71%) answered the follow-up questions satisfactorily. Another 77 were filtered out, as they were repeat responders from the previous experiment, leaving the total number of responses analyzed to 261.

The results from this experiment can be seen in 1, on the right-hand panel. The utterance types are on the x-axis, with the proportion of respondents who checked both boxes (thereby endorsing the {A,B} world) on the y-axis. Red dots indicate responses for the high conditional probability condition, blue for the low conditional probability condition. Error bars indicate 95% confidence intervals. The proportion of {A,B} world endorsements is at floor when respondents heard the "A but not B" utterance, as predicted. However, when respondents heard the "A and B" utterance, endorsements of the {A,B} were relatively low (m=0.6, m=0.66 in the high and low probability conditions, respectively). This means that when respondents read "I will move what I can lift down to the curb" followed "I can lift the chair and the footstool down", and are then told that the character moved all the furniture he could lift, they are willing to endorse a world where only one had been moved (the footstool in 72% of the cases). We believe this behavior is partly due to the experimental setup: subjects may expect the character to do as little work as possible without the help of her friend, who they were told would assist in the moving process later on. We had initially thought that the commitment to 'move what I can lift' would ensure that modalized sentence of the form 'I can do X' would be interpreted as implying that the character did X, but this result suggests that this was not always the case.

To test whether the conditional probability had an effect on the rate of {A,B} world endorsements, we fit a linear regression model using experimental conditions as predictors. We found a significant main effect of ONLY utterances and SINGLE utterances ($p<0.001$ for both), whereby subjects were less likely to endorse the {A,B} world for these two conditions. In addition, we found an interaction between the prior probability and the SINGLE utterance types ($p=0.0144$), whereby participants were more likely to endorse the {A,B} world in the high conditional probability after hearing the non exhaustified utterance. Overall these results indicate that gradient prior probabilities gradiently affect utterance interpretation, raising the question why we did not observe a similar gradience in the elicitation experiment.

Model Fit

We fit the vanilla Recursive Speech act Model presented in (Frank & Goodman, 2012) to the human data we collected, with one level of recursion depth (that is, we fit S_1 and L_1).

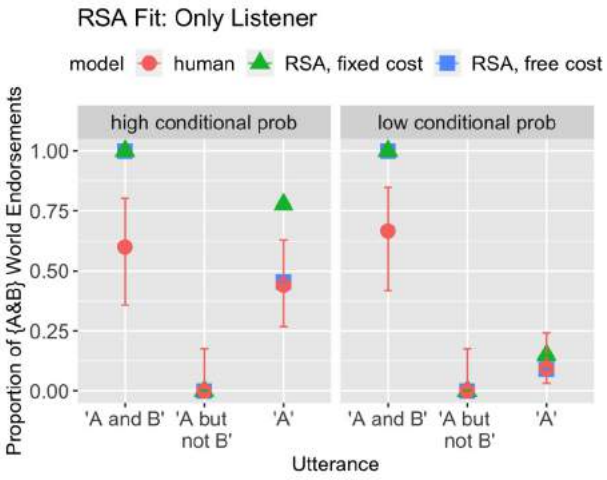


Figure 2: RSA Model fit with fixed cost (green triangles) and free cost ratios (blue squares) to human judgements (red circles).

The model has three possible world states: {A}, {B} and {A,B}. World {A} had a prior of 0.32, world {B} had a prior of 0.14 and world {A,B} had a prior of 0.54, rendering $P(\{A,B\} \mid \{A\}) = 0.8$, close to the human *high conditional probability* prior, and $P(\{A,B\} \mid \{B\}) = 0.63$, close to the human *low conditional probability* prior. The model includes seven messages: 'A', 'B', 'A but not B', 'B but not A', 'A and B' and 'null', which is defined as true in every situation, and was assigned a fixed cost of 100. 'A' and 'B' were assigned a cost of 0.[3] The costs of 'A and B' (c_1) and 'A but not B' (c_2) were free parameters, as was the rationality parameter, α.

In order to assess how well the RSA model captured the human judgements, we conducted four fits, which are summarized in Table 3. Each fit was made by iterating through a wide range of alphas and cost parameters (0-20 for each). This technique guarantees that we found a locally optimal fit within the range of cost and optimally parameters typically seen in the rest of the Recursive Speech Act literature (Scontras, Tessler, & Franke, 2017). In the fixed cost ratio fits, the cost for "A but not B" must be greater than but could not be more than 2 times that of "A and B". This constraint makes sense if we view cost as reflecting, for instance, the number of logical operators in a sentence, or the number of words used. Thus, we wanted to see if an fit existed with cognitively plausible relative costs between these two types of utterances. But we also relaxed this constraint in the 'free cost' fit, where the only constraint that the the cost of "A but not B" is higher than that of "A and B".

The results for the listener-only fit can be seen in Figure 2. Here, the x-axis is the possible utterances, and the x-axis is the proportion of respondents who endorsed the {A,B} world (in the human case) or the posterior distribution on the {A,B} world (in the model case). The left panel represents the high

[3] In the RSA model, it is the *difference* between relative costs that matters: adding a fixed constant to each cost value has no effect.

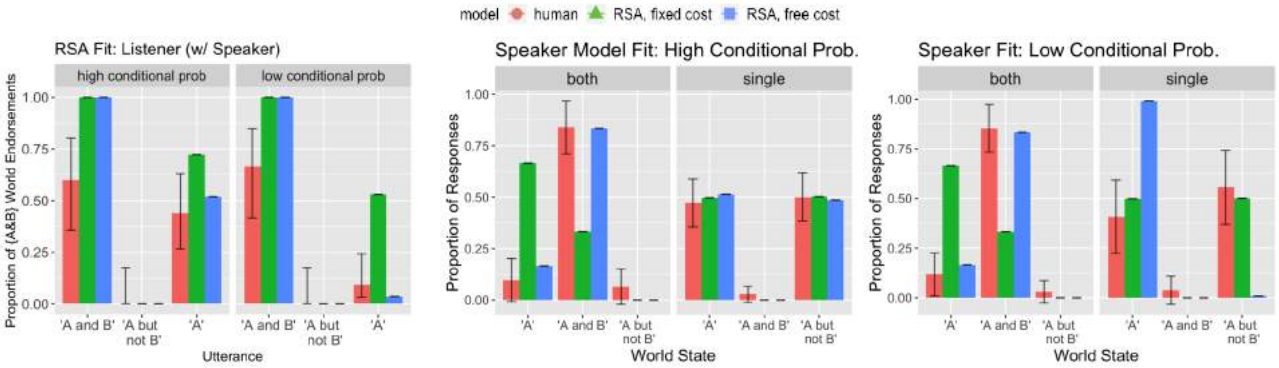


Figure 3: RSA Model fit with fixed-cost ratios (green) and free cost ratios (blue) to human judgements (red).

Name	Layers	Restrictions	α	c_1	c_2	MSE
Fit 1	L_1	$c_2 < 2 * c_1$	9.19	1.57	3.17	0.097
Fit 2	L_1	$c_1 < c_2$	5.52	0.010	8.42	0.068
Fit 3	L_1, S_1	$c_2 < 2 * c_1$	0.01	0.27	0.54	0.092
Fit 4	L_1, S_1	$c_1 < c_2$	7.97	0.01	1.59	0.059

Table 3: Summary of The Four Fits and Optimal Parameters

conditional probability condition and the right panel the low conditional probability condition. The vanilla RSA's meaning function constrains the listener's posterior for utterances 'A and B' and 'A but not B' such that all probability is assigned to the {A,B} world and the {A} world, respectively. Therefore, it is entirely the model's posterior on the 'A' utterance that determines the relative goodness of the fit. For the restricted cost ratio fit (green triangles) the best model is able to match human behavior in the low conditional probability condition, but favors the {A,B} much more greatly than do human respondents in the high conditional probability condition (the green triangle is well above the red error bars), resulting in a mean squared error of 0.097. When the restriction on relative costs is relaxed (blue squares) the model is able to achieve a very precise fit, with a mean squared error of 0.0678. The reason why the free-cost fit is able to perform significantly better than the fixed cost fit is that it can assign much higher relative cost to "A but not B" than to "A and B".

For example, in Fit 2 the utterance "A but not B" is 840 times more costly than the utterance "A and B". This results in strong model performance because high relative cost of the exhaustified utterance counterbalances its informativity at communicating the $\{A\}$ world. This renders the 'A' utterance a good choice to communicate the $\{A\}$ world, despite the strong priors on the $\{A, B\}$. Furthermore, the low cost of "A and B" ensures that it will be chosen often in the $\{A, B\}$ world, even in the high-probability condition.

The results for Fits 3 and 4, which fit both the speaker and listener layers, can be seen in Figure 3, with the listener layer graphed at left and the speaker layer graphed in the center and right images. For the listener layer, the x-axis shows utterances, and the y-axis posterior probability endorsements for the {A,B} world. For the speaker layer, the facets represent the different worlds conditions and the x-axis shows the possible utterances, with the relative proportion assigned to each utterance (for the RSA models) or proportion of endorsements (for the human) on the y-axis.

As to performance of the model: in the restricted cost ratio fit (the green bars) the performs only moderately well. For the ten critical conditions where the posterior distributions are not constrained to either 100% of 0%, the best fit falls outside of the human judgements' 95% confidence intervals 6 times, resulting in a mean squared error of 0.092. For the free cost model (blue bars) the model is able to perform slightly better, falling outside of the human judgements' 95% confidence intervals only twice (both in the {A} world, low probability condition). This fit gives an MSE of 0.059. Two remarks are in order. First, in the free cost model, "A but not B" is 158 times more costly than the utterance "A and B". Second, the best model achieves a good fit for the listener and for the speaker in the high-probability condition, but drastically underestimates the rate of endorsement of "A but not B" in the low probability condition as a way to express the $\{A\}$-world.

Discussion

The results of the interpretation experiment establishes that prior probabilities modulate exhaustivity effects, as is expected under the RSA approach. In our data, they do so for interpretation, but not for production. The RSA model can achieve a good fit with our experimental data for the interpretation experiment only with implausible parameters. With the kind of cost values that are typically assumed (cf. fixed cost fit), it overestimates the effect of prior probabilities. When we relax constraints on costs, an excellent fit is achieved, but the cost of "A but not B" has to be 832 times that of "A and B". When we want to fit both interpretation and production, the best model drastically underestimates the use of sentences such as "A but not B" - precisely because it assigns it an extremely prohibitive cost. Note that we are only evaluating the baseline RSA model. More sophisticated models have been proposed within the RSA framework, and we are not evaluating those. What our results suggest is that a key ingredient of the baseline RSA model, namely the tradeoff between infor-

mativity and cost, which predicts a huge influence of priors on interpretation and production, might make it hard to capture both interpretation and production data. On the interpretation side, the model needs to assign a very high cost to *A but not B*, but then on the production side, the model predicts that *A but not B* is not usable.

That being said, this conclusion is provisional, as caution is in order when interpreting the results we present here. We only tested two different conditions, in one type of scenario, and the data are somewhat noisy (cf. the high rate of rejection of $\{A, B\}$ after hearing "I can do A and B"). The fact that we used modal sentences when we collected priors and in the interpretation task is a limitation of this study.[4] Future work is needed to a) gather additional and less noisy data so as to reach more reliable conclusions, b) construct alternative models, including refined versions of the baseline RSA model, which could then be compared to it.

Acknowledgments

B.S. would like to thank Leon Bergen, Danny Fox and Roni Katzir for relevant discussions, and acknowledges funding from a grant from the Agence Nationale de la Recherche (ANR-17-EURE-0017). E.G.W. would like would like to acknowledge support from the Mind Brain Behavior Interfaculty Initiative Graduate Student Grant.

References

Bergen, L., Levy, R., & Goodman, N. (2016). Pragmatic reasoning through semantic inference. *Semantics and Pragmatics*, *9*.

Degen, J., Tessler, M. H., & Goodman, N. D. (2015). Wonky worlds: Listeners revise world knowledge when utterances are odd. In *Cogsci.*

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998–998.

Grice, H. P. (1975). Logic and conversation. *1975*, 41–58.

Scontras, G., Tessler, M. H., & Franke, M. (2017). *Probabilistic language understanding: An introduction to the rational speech act framework* (Tech. Rep.). Retrieved 2018-1-9 from https://problang.org.

[4] In another experiment that we do not present here, we collected judgments for non-modal sentences in near-identical scenarios, and we got a much higher endorsement (85%) for the $\{A, B\}$ world for the conjunctive sentence *A and B*.

Phonological Cues to Syntactic Structure in a Large-Scale Corpus

Ethan Wilcox
Harvard University, Cambridge, Massachusetts, United States

Abstract

The Prosodic Bootstrapping Theory (PBT) states that prosodic and phonetic cues assist infant language learners to segment the speech stream into words and assemble those words into phrase structures. However, many of the studies demonstrating a link between prosody and syntax were conducted on small data sets and on a narrow range of syntactic structures. This work uses a state-of-the-art parser to syntactically annotate the BU Radio News Corpus of around 16,000 diverse sentences, which are prosodically tagged and annotated. A decision tree classifier was fit, using six prosodic features and achieving 87% accuracy at differentiating words internal to major syntactic phrases vs. words that mark phrase boundaries. However, the models tested are unable to differentiate between phrasal categories based on prosodic information alone. These results provide new evidence in support of the Prosodic Bootstrapping Theory, suggesting it is possible to identify phrasal boundaries based on prosodic information alone.

The Accuracy of Causal Learning over 24 Days

Ciara L. Willett (clw137@pitt.edu)
Benjamin M. Rottman (rottman@pitt.edu)
Department of Psychology, University of Pittsburgh,
3939 O'Hara Street, Pittsburgh, PA 15260 USA

Abstract

Humans often rely on past experiences stored in long-term memory to predict the outcome of an event. In traditional lab-based experiments (e.g., causal learning, probability learning, etc.), these observations are compressed into a successive series of learning trials. The rapid nature of this paradigm means that completing the task relies on working memory. In contrast, real-world events are typically spread out over longer periods of time, and therefore long-term memory must be used. We conducted a 24 day smartphone study to assess how well people can learn causal relationships in extended timeframes. Surprisingly, we found few differences in causal learning when subjects observed events in a traditional rapid series of 24 trials as opposed to one trial per day for 24 days. Specifically, subjects were able to detect causality for generative and preventive datasets and also exhibited illusory correlations in both the short-term and long-term designs. We discuss theoretical implications of this work.

Keywords: causal learning; probability learning; illusory correlation; long-term memory; smartphone

Introduction

Every day we use our experiences to make inferences. For example, is your new medication improving an ailment or causing a negative side-effect? Does meditating have a positive impact on your mental health? If we can accurately predict the outcomes of our experiences and actions, we can use this information to behave adaptively in the world.

Trial-by-trial learning paradigms, in which cue-outcome pairs are presented to subjects sequentially, are used extensively to study learning across many different fields including causal learning, probability learning, fear learning, stereotype formation, associative learning with non-human animals, and others. The trial-by-trial paradigm is supposed to simulate an important aspect of the world: most of our experiences occur sequentially over time, rather than in a summarized form. Typically the 'inter-trial-interval', the time between trials, is a couple seconds. However, we contend that there are few real-world learning situations that involve experiencing repeated cue-outcome pairs separated by seconds, perhaps with a few exceptions (e.g., flipping through records rather than first-hand experiences).

The goal of the current study is to compare trial-by-trial learning in the normal rapid format vs. trial-by-trial learning in which the trials are spaced out once per day. Day-by-day learning simulates many natural processes (e.g., does a medicine have an influence on a health outcome, does exercising have an influence on sleep, etc.). Importantly, whereas working memory is believed to support learning in short timeframes, long-term memory must take over when learning occurs over many days. In the current study we investigated how effectively people are able to learn cue-outcome relations across multiple days.

Trial-by-Trial Causal Learning

Prior research has evaluated how people detect causation from data shown over a successive series of trials. In a typical experiment, participants observe data in which the putative cause and the outcome are either present or absent. This information can be organized into a 2x2 table where each cell *A-D* represents the number of times that the cause/outcome combination occurs for a particular dataset (see Figure 1). Most often, participants are shown the data rapidly, for example two or three seconds per trial. After observing the entire dataset, subjects judge the degree to which the cause influences the outcome.

		Outcome: Present	Outcome: Absent
Cause	Present	*A*	*B*
	Absent	*C*	*D*

Figure 1: A 2x2 table depicting the four possible types of data in a traditional binary design.

One normative model of causation is the Δ*P* rule, a measure of contingency that suggests an optimal way to infer causation is by comparing the probability of the outcome in the presence of the cause and the probability of the outcome in the absence of the cause: $\Delta P = A/(A+B) - C/(C+D)$. When Δ*P* is positive, the causal relationship is generative. When Δ*P* is negative, the causal relationship is preventive.

Although prior research suggests that people are able to discriminate generative vs. preventive causation (Shaklee & Mims, 1982), individuals sometimes exhibit biases in causal reasoning. One such bias, "illusory correlation" or "illusory causation", occurs when people inaccurately infer causation when no causal relationship exists.

An "A-cell bias" is when individuals believe that a causal relation exists merely because of a high number of A-cell trials (e.g., Kao & Wasserman, 1993). In the A-cell bias condition in Table 1, even though there is zero relation between the cue and outcome (the outcome occurs with a chance of .625 regardless of whether the cue is present or absent, so $\Delta P = 0$), people tend to infer that they are positively correlated. An "outcome density bias" is when people incorrectly assign causation to a dataset in which the overall probability of the outcome is high (Table 1), even though the

probability of the outcome is the same (.75) whether the cause is present or absent, so $\Delta P = 0$ (e.g., Jenkins & Ward, 1965).

Table 1: Cell Frequencies for the 4 Datasets

Dataset	*A*	*B*	*C*	*D*	ΔP
Generative	9	3	3	9	0.5
Preventive	3	9	9	3	-0.5
Outcome-Density	9	3	9	3	0
A-cell	10	6	5	3	0

Causal Learning and Memory

Many causal learning experiments use rapidly successive trial-by-trial paradigms. In the real world, however, you would not experience each data point in rapid succession. This raises a number of challenges for long term memory. For example, imagine learning whether going to yoga improves your mood; some days you do yoga and other days you do not. After a few weeks, would you be able to remember the days you did or did not do yoga? Could you remember your mood on those days? How might your memories for these events impact your ability to detect causation? Would you be more susceptible to biases such as illusory correlations? Currently, there is no research on how well people can learn causal relations over long timespans.

One basis for making hypotheses about causal learning in long timeframes is research on short timeframes that has increased working memory (WM) demands. Studies have found stronger illusory correlations in a rapid trial-by-trial paradigm (higher WM demands) than in a "summary" paradigm (lower WM demands) in which all the trials are presented simultaneously (Kao & Wasserman, 1993). Adding a distractor task on top of the trial-by-trial paradigm leads to less accurate judgments (Shaklee & Mims, 1982), and older adults with lower WM have less accurate causal learning (Mutter & Pliske, 1996). If causal learning is worse when WM is taxed, we expected learning to get even worse when long-term memory must be used to assess causation. Still, people are often able to navigate the world successfully, suggesting a reasonable causal-learning ability when relying on long-term memories to make inferences. This raises the question: how well can we learn causal relations across many days?

Summary of Current Study

In the current study, we investigated the implications of learning a cause-effect relationship quickly from a rapid sequence of trials vs. learning the same relationship over an extended period of time – one trial per day for 24 days. We investigated how subjects learned about four causal relations using different datasets: generative, preventative, 'outcome-density', and 'A-cell' (Table 1).

The motivation for studying the generative and preventive datasets was to determine whether or not participants were capable of detecting a causal relationship or if learning is hampered when the experiences occur spread out in time. Because memories might be noisier in the long-term condition, we predicted that participants' judgments might be closer to zero, implying a weaker causal relationship.

For the A-cell and outcome density datasets, we wanted to assess the effect of long-term memory on illusory correlations. Prior research mainly found exaggerated illusory correlations with increased WM demand, so one hypothesis was that illusory correlations would be exaggerated in the long-term condition. Another hypothesis was that, if memories of the experiences are weaker in the long timeframe condition, then the judgments might actually be closer to zero – more accurate.

Methods

Participants

There were 476 participants. The main requirements were owning a smartphone and intending to complete the entire study; however, we mainly targeted college students to have a similar sample to most other causal learning studies and since they frequently use smartphones. Participants were paid $30 if they successfully completed the entire study.

Our goal was to have around 400 participants, 100 for each of the 4 datasets in the long timeframe condition. The large number was used because the four datasets need to be analyzed separately, and to have power to detect small effects. The final data analyses included 409 participants after dropping 13 participants who admitted to writing down data during the study, 1 who was not trying during the task, 39 due to a programming error, and 14 who skipped too many days of the long timeframe task.

Datasets

Participants learned about five datasets: four short-timeframe (generative, preventative, A-cell, and outcome density) and one long-timeframe (one of the four from the short-timeframe

Table 2: Example Datasets for a Subject

Task Order	Day	Length	Dataset	Context	Valence	Authenticity
1	1	Short	A-Cell*	Restaurant	Positive*	Real*
2	1	Short	Preventive	House	Negative	Vitamin
3	1-24	Long	A-Cell*	Library	Positive*	Real*
4	25	Short	Generative	Street	Positive	Vitamin
5	25	Short	Outcome Density	Park	Negative	Real

Note. *Matched short and long timeframe conditions.

condition). This design allowed for a within-subjects comparison of one of the four datasets across the long vs. short conditions (see Table 2 for an example). By having subjects learn all four datasets in the short timeframe condition, it also reduces the likelihood that subjects were aware that one of the short timeframe datasets was the same as the long timeframe dataset. Each dataset consisted of 24 trials ordered randomly. The two illusory correlation datasets were previously used by Kao and Wasserman (1993).

Procedure

Participants completed the entire study on their own smartphones by logging into our website created with our PsychCloud.org framework. The procedure for the short-term and long-term tasks were identical, except that subjects observed one trial per day in the long timeframe condition, and they did trials back-to-back in the short timeframe condition. On Day 1 of the study, participants completed two short-term tasks and began Day 1 of the long-term task.

On Days 2 – 24, participants received automated text-message reminders at 10am, 3pm, and 8pm to complete their daily trial for the long-term task and stopped receiving reminders if they had already participated that day. They returned to the lab on Day 25 to complete the remaining short-term tasks and receive payment. The order of the short-term tasks was randomized so that participants completed the short version of the long-term task either on Day 1 or on Day 25 - before or after the long-term task.

Within a Trial Each task consisted of 24 trials in which participants were told whether or not the putative cause was present or absent. A number of procedures were taken to facilitate encoding, including asking subjects to verify the state of the cause and effect (rather than just observe them), and to spend extra time to look each image. Each trial proceeded as described in the following example, which uses the 'Facebook' cover story – other cover stories are explained below. In the Facebook cover story, subjects were asked to judge whether using Facebook during their lunch break improves or worsens or has no influence on their mood, based on the hypothetical dataset.

At the beginning of each trial, subjects were shown a contextual image. These images allowed us to ask a number of episodic memory questions that are not analyzed in this report. In the Facebook cover story, they saw an image from the inside of a restaurant and were told "This is the scene from your lunch break." After three seconds, an icon and text were superimposed over the contextual image to show the presence or absence of the cause (e.g., whether they used or did not use Facebook during their lunch break). They pressed a radio button to confirm the state of the cause and could not move on until selecting the correct button (e.g., Facebook vs. No Facebook). Next, they pressed a radio button to predict the effect as present or absent (e.g., Very Sad Mood vs. Normal Mood). They received text feedback for whether their prediction was correct or incorrect and an icon representing the effect was superimposed on the image. After clicking the correct radio button to verify the state of the effect, subjects were instructed to "Take a couple of seconds to imagine this scene", which was displayed for an additional four seconds.

At the end of a trial in the short timeframe condition, subjects were permitted to move on to the next trial. In the long timeframe condition, subjects were told that their task was over and to come back to the website the following day. Once a trial was over, the website did not allow subjects to see the data for that trial or prior trials, not even by clicking the back button on their web browser.

Figure 2. Screenshot of the end of a trial.

After Trial 24 (either immediately afterwards for the short timeframe condition, or on Day 25 in the long timeframe condition), participants judged the strength of the causal relationship. First, they answered whether the cause (Facebook) "improves or worsens or has no influence" on the effect (mood). If participants said the cause had no influence, they were assigned a causal judgment of 0. If they responded "improve" or "worsen", they answered "How strongly does [the cause] [improve/worsen] [the effect]?" on a scale of 1 (*very weak*) to 10 (*very strong*), which produced a scale from -10 to +10. In this report we only discuss the judgments after Trial 24, though subjects also made similar judgments before Trial 9 and before Trial 17.

In addition to the causal strength judgments, participants also made a number of other judgments, for example, memories of the number of experienced cells of types A, B, C, and D (before Trials 5, 13, 21, and after Trial 24), as well as a number of judgments about the memories for the contextual images after Trial 24. These measures will not be analyzed in the current report due to space.

Cover Stories Since subjects learned about five cause-effect relations, we created the following five authentic '*contexts*', randomly assigned to the five tasks so that each was viewed as a separate learning task: the relation between using Facebook during lunch in a *restaurant* and mood, eating a

healthy dinner in a friend's *house* and having an upset stomach, using notecards to study in a *library* and grades on a daily quiz, biking to work on *city streets* and productivity at work, and bringing your dog on a walk in a *park* and stress. The five stories were chosen so that it would be plausible for the cause to either improve or worsen the outcome; the influence of prior beliefs will be analyzed in other reports.

Because this study is the first to use a long timeframe paradigm, is unlikely to be replicated, and is focused on external validity, we conducted two manipulations of the cover stories. Specifically, we manipulated the "authenticity" and "valence" of the cover stories. If subjects in the long timeframe condition exhibited very poor learning, we wanted to rule out some potential explanations and to know how to best design future studies. Although we will explain the manipulations here, they are not of primary importance and will not be analyzed in this report.

First, though it is typical in causal learning studies to use entirely novel and abstract cover stories to minimize the influence of prior beliefs, we worried that abstract stimuli could be hard to remember in a long timeframe condition.[1] For this reason, we manipulated the 'authenticity' of the cover stories. The 'authentic' cover stories were the five stories mentioned previously. In the 'novel' cover stories, we used the same effects but replaced the causes with a hypothetical vitamin that a subject took on some days but not others (e.g., does the vitamin have an influence on mood, upset stomach, etc.). The matched short-term and long-term datasets were assigned to different contexts but were matched on authenticity. Of the four short timeframe conditions, two were assigned to 'novel' vitamin cover stories and two were assigned to authentic cover stories (Table 2).

Second, we manipulated the 'valence' of the effect; whether the presence of the effect is good or bad.[2] The absence of the effect was always described as normal (e.g., normal mood, normal grade on a quiz, etc.). The presence of the effect was described as either very good or very bad (e.g., very happy or very sad; very good grade or very bad grade, etc.). For participants in the negative valence condition, we reverse coded their causal strength judgments, so positive causal strength means "improved" for the positive valence condition and "worsened" for the negative valence condition. The matched short-term and long-term datasets were assigned the same valence. Of the four short-term conditions, two had positive and two had negative valence (see Table 2). Authenticity and valence are not analyzed due to space.

Participation Before starting the experiment, participants were told that if they missed more than three days in the long timeframe task, the study would be terminated and that they would not be paid. 462 (97%) participants successfully completed the study. On any given day, 83% of subjects participated before the 3pm reminder, 96% before the 8pm reminder, and 99% by midnight. If a subject missed one, two, or three days, the subsequent days were automatically pushed back the appropriate number of days.

The causal strength judgments and other measures for the long timeframe task occurred during the second in-lab testing session. We worked hard to have subjects come back to the lab for the second in-lab testing session on Day 25, one day after the last trial in the long timeframe condition. Of the 409 subjects in the final analyses, 83% returned to the lab on Day 25. If they skipped one day of the long timeframe task, sometimes this session occurred on the same day as their 24th trial (13%). If the session had to be moved, sometimes it occurred two (3%) or three (1%) days after the last trial. Overall, the protocol was followed with high fidelity.

Results

Causal Strength Judgments

In this paper, we only analyzed data from the matched short-term and long-term conditions. We analyzed the generative ($N = 98$), preventive ($N = 102$), A-cell ($N = 105$), and outcome density ($N = 104$) conditions separately.

Average causal strength judgments are presented in Figure 3. Significance values above each column indicate whether the value was significantly different from zero. The significance value above the horizontal lines indicates whether the judgments in the short and long-term conditions were significantly different from each other. We calculated Bayes Factors (*BF*) for each t-test, where a $BF > 1$ is support for the alternative hypothesis and a $BF < 1$ is support for the null. Often *BF*s > 10 (or < 1/10) are considered "strong" evidence for the alternative (or null), *BF*s > 30 or < 1/30 are considered "very strong" and *BF*s >100 or < 1/100 are considered "extreme" (e.g., Lee & Wagenmakers, 2013).

Generative and Preventive Conditions First, we wanted to assess whether participants were capable of detecting causation in the generative and preventive conditions. For the generative dataset, causal judgments were significantly different from zero in both the short-term condition, $t(97) =$

[1] For example, we suspect that in short learning tasks using novel stimuli, subjects might use other cues such as the position of stimuli on the screen rather than the semantic meanings of the cues. Such alternative methods of learning might be less salient in the long timeframe condition. Instead, we thought that semantically meaningful cause-effect relations might be easier to remember and also have higher external validity.

[2] Most studies on causal learning use cues that are either present or absent. Presence/absence of the cause and the effect is theoretically important in some theories of causal learning (e.g., Cheng, 1997). Further, the definition of the cells as A-D only makes sense with cues that are present/absent (not "high"/"low" or "2"/"1", etc.; see Figure 1). In order to stick close to prior studies and to be able to study the A-cell bias, we used present/absent cues. However, one consequence of using presence/absence is that most outcomes have an implicit valence of being good or bad. For example, many prior studies have used outcomes like the presence/absence of a headache (bad) or of a flower blooming (good). We did not want to arbitrarily use outcomes of one particular valence, or to confound valence with cover story. Furthermore, valence can influence the strength of illusory correlations (Mullen & Johnson, 1990). For all these reasons, we counterbalanced the valence of the cover story.

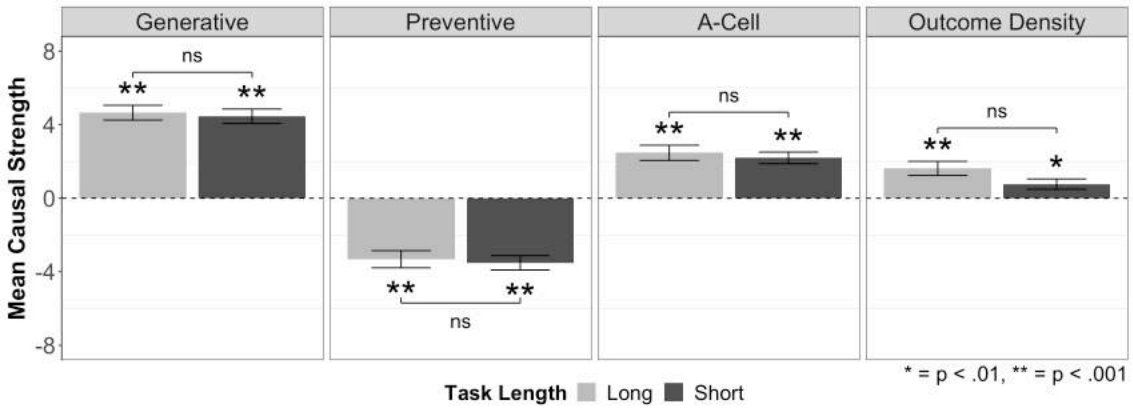


Figure 3: Average causal strength judgments after reviewing 24 trials. Error bars indicate standard error.

11.27, $p < .001$, $d = 1.14$, $BF = 2.13 * 10^{16}$, and the long-term condition, $t(97) = 11.53$, $p < .001$, $d = 1.17$, $BF = 7.37 * 10^{16}$. For the preventive dataset, judgments were less than zero in both the short-term, $t(101) = -9.03$, $p < .001$, $d = -0.89$, $BF = 5.72 * 10^{11}$, and long-term condition, $t(101) = -7.13$, $p < .001$, $d = -0.71$, $BF = 6.05 * 10^{7}$.

We predicted that for both the generative and preventive datasets, causal judgments would be closer to zero in the long-term condition because participants' memories would be noisier. However, paired t-tests revealed no significant differences between judgments in the short-term and long-term conditions for either the generative, $t(97) = -0.37$, $p = .707$, $d = -0.04$, $BF = 0.12$, or preventive datasets, $t(101) = -0.33$, $p = .741$, $d = 0.03$, $BF = 0.12$. Thus, participants were just as capable of detecting causation in the short and long timeframe conditions.

Illusory Correlation Conditions In the outcome-density and A-cell datasets, an optimal causal judgment would be zero. In line with our predictions, we found significant illusory correlations for both datasets**.** For the A-cell dataset, causal judgments were significantly greater than zero in both the short-term, $t(104) = 7.13$, $p < .001$, $d = 0.70$, $BF = 6.75 * 10^{7}$, and long-term, $t(104) = 6.11$, $p < .001$, $d = 0.60$, $BF = 6.36 * 10^{5}$, conditions. We found similar results for the outcome-density dataset; judgments were also positive and significantly different from zero in the short-term, $t(103) = 2.73$, $p = .008$, $d = 0.27$, $BF = 3.60$, and long-term, $t(103) = 4.23$, $p < .001$, $d = 0.41$, $BF = 341.33$, conditions.

We hypothesized that the illusory correlations could be either exacerbated or diminished in the long timeframe condition. However, there were no differences between causal judgments in the short and long-term conditions for the A-cell dataset, $t(105) = -0.67$, $p = .500$, $d = 0.07$, $BF = 0.13$. Illusory correlations appeared slightly stronger in the long-term condition for the outcome-density bias dataset, but this trend only approached significance, $t(104) = -1.87$, $p = .065$, $BF = 0.45$, with a small effect size of $d = 0.18$. These results suggest that illusory correlations in traditional trial-by-trial experiments are similar to what we observe in a long timeframe task.

Predictive Strength

Another way to measure learning, aside from causal strength judgments, is through subjects' predictions of whether the outcome was present or absent each day. To ensure that participants had observed enough experiences to make predictions, we analyzed the predictions from Trials 13 – 24.

We transformed participants' predictions into a measure of causal strength by subtracting the probability that they predicted that the outcome would be present given the absence of the cause from the probability that the outcome would be present given the presence of the cause. This measure of "predictive strength" is conceptually similar to ΔP. These results are displayed in Figure 4.

Generative and Preventive We found very similar results using subjects' predictions to assess learning as from their causal strength judgments. In the generative condition, predictive strength was significantly greater than zero for both the short-term, $t(97) = 11.58$, $p < .001$, $d = 1.17$, $BF = 9.01 * 10^{16}$, and long-term conditions, $t(97) = 12.47$, $p < .001$, $d = 1.26$, $BF = 6.32 * 10^{18}$. In the preventive condition, predictive strength was significantly less than zero in both the short-term, $t(101) = -11.87$, $p < .001$, $d = -1.18$, $BF = 6.77 * 10^{17}$, and long-term conditions, $t(101) = -9.38$, $p < .001$, $d = -0.93$, $BF = 3.12 * 10^{12}$. We found no difference in predictive strength between the short-term and long-term conditions for either the generative, $t(97) = -0.36$, $p = .718$, $d = -0.04$, $BF = 0.12$, or preventive, $t(101) = -0.49$, $p = .623$, $d = -0.05$, $BF = 0.12$, datasets. In sum, participants learned to accurately predict the effect, to the same extent, in both conditions.

Illusory Correlation Conditions In the A-cell bias condition, we found a similar pattern of results to the strength judgments. Subjects did infer an illusory correlation; they were more likely to predict the effect as present when the cause was present in both the short-term, $t(104) = 3.66$, $p < .001$, $d = 0.36$, $BF = 51.08$, and long-term condition, $t(104) = 3.66$, $p < .001$, $d = 0.36$, $BF = 50.64$. Furthermore, we found no difference between predictions in the short-term vs. long-term conditions, $t(104) = -0.66$, $p = .512$, $d = 0.06$, $BF = 0.13$.

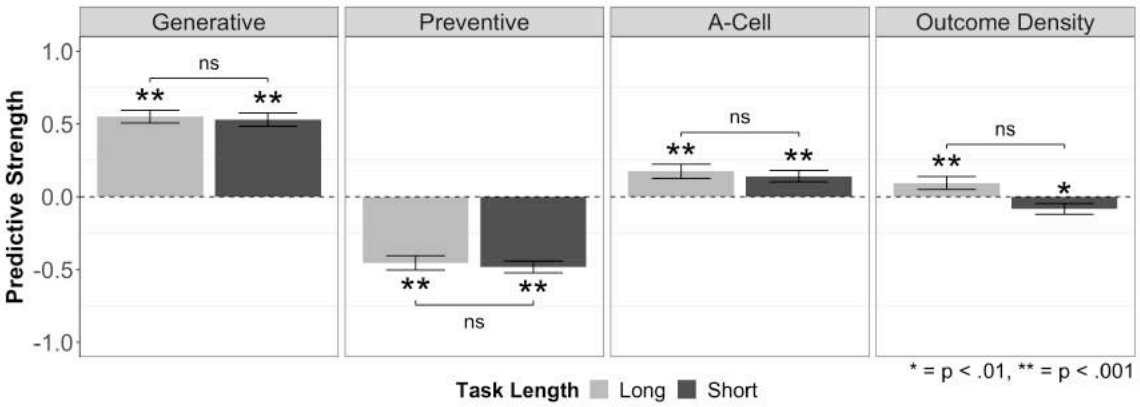


Figure 4: Δ P calculated from predictions about the effect given the cause. Error bars indicate standard error

In the outcome-density condition, the predictions were significantly negative in the short-term condition, $t(103) = -2.24$, $p = .028$, $d = -0.22$, $BF = 1.17$. However, they were significantly positive in the long-term, $t(103) = 2.13$, $p = .036$, $d = 0.21$, $BF = 0.94$, and the difference was statistically significant, $t(103) = -3.60$, $p < .001$, $d = -0.35$, $BF = 42.20$.

This difference was only marginally significant for the causal strength analyses, and the causal strength judgments for the short timeframe were significantly positive, not negative. Because this is the only difference between the two conditions, and it was only found for predictive strength (not the causal strength judgments) in the outcome density condition (not the other illusory correlation condition), we do not want to over-interpret it.

Discussion

We sought to evaluate the external validity of traditional trial-by-trial causal learning experiments by comparing trial-by-trial learning when presented rapidly vs. one trial per day for 24 days. Presumably the former relies on working memory, whereas the latter requires long term memory. Our findings suggest that people are capable of learning generative and preventive causal relationships and also exhibit illusory correlations when learning causal relations over 24 days. Critically, we found few differences between the short-term and long-term tasks, and in fact most of the Bayes factors were roughly 8 to 1 in favor of the null.

From a practical perspective, this research provides an optimistic perspective on the validity of the trial-by-trial paradigm as a simulation of causal learning that occurs in the real world across longer periods of time. Assessing the external validity of this paradigm is important given that it has been used in hundreds of published studies on causal learning, and many thousands of studies when including studies of probability learning and other related topics.

From a theoretical perspective, we find it striking that there are so few differences in learning across the short and long timeframe condition. We intentionally used large samples to have the power to detect small effects. The robust learning in the long timeframe condition is surprising considering that participants completed the long-term trials outside of the lab and likely participated with many distractors and interruptions, comparable to everyday causal learning. Still, we hypothesized that the learning in the long timeframe condition would be plagued by considerably worse learning due to noisy memories. The fact that we found few differences raises a number of questions.

One question has to do with how learning occurs (e.g., Bornstein et al., 2017). Are subjects recording individual episodic memories and using them for causal learning? Or are they merely encoding them as generic events of the four cell types? Or are they using a process more similar to reinforcement learning in which an estimate of the strength of the relation between the cause and outcome gets updated as new evidence is experienced? Some of these questions can be addressed with our contextual image memory questions.

Another question is how well long-term memory can support other types of learning. It is possible that a single cause-effect relation is simple enough for long-term memory to robustly support learning, but that long-term memory might not be able to support more complex cause-effect relations (e.g., with multiple causes or long delays). We are actively studying such questions.

This research also has potential implications for whether learning and memory processes are fundamentally the same for shorter vs. longer timeframes. In associative learning, there is a debate about "timescale independence or invariance" (Gallistel & Gibbon, 2000), in which learning phenomena tend to replicate if the sequence is stretched or compressed. In memory, there are debates about the similarities and differences in short vs. long-term memory (e.g., Cowan, 2008) and whether memories across short and long timespans can be modeled with the same forgetting curves (e.g., Wixted & Ebbesen, 1991). Perhaps researchers invested in these debates may be able to use these results.

More generally, we believe that the current research provides an important step towards generalizing current learning paradigms to more real-world settings. The current findings are optimistic in terms of how well the paradigm generalizes; however, future research may also reveal areas in which standard learning paradigms generalize poorly.

Acknowledgments

This work was supported by NSF 1651330. We would like to thank all of the research assistants who helped with data collection, including Aleks Brown, Beatrice Langer, Caitlin Viele Haggerty, Chris Shon, Elise Faut, Gabriela Cuadro, Joanna Ye, Julia Gillow, Melinda Rosen, Michael Datz, Minbae Le, Priya Chandrasekaran, Rachel Hopkins.

References

Bornstein, A. M., Khaw, M. W., Shohamy, D., & Daw, N. D. (2017). Reminders of past choices bias decisions for reward in humans. *Nature Communications*, *8*, 15958.

Cheng, P. W. (1997). From covariation to causation: a causal power theory. *Psychological review*, *104*(2), 367-405.

Cowan, N. (2008). What are the differences between long-term, short-term, and working memory? *Processes in brain research, 169,* 323-338.

Gallistel, C. R., & Gibbon, J. (2000). Time, rate, and conditioning. *Psychological Review, 107*(2), 289-344.

Jenkins, H. M., & Ward, W. C. (1965). Judgment of contingency between responses and outcomes. *Psychological Monographs: General and Applied, 79*(1), 1-17.

Kao, S. F., & Wasserman, E. A. (1993). Assessment of an information integration account of contingency judgment with examination of subjective cell importance and method of information presentation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*(6), 1363-1386.

Lee, M. D., & Wagenmakers, E-J. (2013). Bayesian cognitive modeling: A practical course. Cambridge University Press.

Mullen, B. and Johnson, C. (1990), Distinctiveness-based illusory correlations and stereotyping: A meta-analytic integration. *British Journal of Social Psychology, 29*, 11–28. doi: 10.1111/j.2044-8309.1990.tb00883.x

Mutter, S. A., & Pliske, R. M. (1996). Judging event covariation: Effects of age and memory demand. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *51*(2), 70-80.

Shaklee, H., & Mims, M. (1982). Sources of error in judging event covariations: Effects of memory demands. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *8*(3), 208-224.

Wixted, J. T., & Ebbesen, E. B. (1991). On the form of forgetting. *Psychological science*, *2*(6), 409-415.

Modeling Expertise with Neurally-Guided Bayesian Program Induction

Catherine Wong
Massachusetts Institute of Technology, Cambridge, Massachusetts, United States

Kevin Elllis
Massachusetts Institute of Technology, Cambridge, Massachusetts, United States

Mathias Sabl-Meyer
PSL/Collge de France, Paris, France

Josh Tenenbaum
Massachusetts Institute of Technology, Cambridge, Massachusetts, United States

Abstract

Studies of human expertise suggest that experts and novices "see" problems differently. Experts not only acquire a body of domain-specific strategies and knowledge, but also learn to quickly identify when those concepts apply to problems within the domain. We propose modeling these elements as an iterative process of domain-specific language (DSL) learning, while jointly training a neural network to recognize when learned concepts apply to new problems. We show that the algorithm solves problems more accurately and quickly than either a neural network alone, or a model that simply acquires new concepts without learning when to use them. We also examine the implicit problem representations learned by the neural network recognition model, and find that they increasingly come to reflect abstract relationships between problems, rather than surface features, as the model acquires domain expertise. A full paper and additional details are available at: https://sites.google.com/view/neurally-guided-expertise-mit

Semantic and Visual Interference in Solving Pictorial Analogies

Emily F. Wong (emilyfwong@ucla.edu)
Department of Psychology, University of California, Los Angeles

Guido F. Schauer (guido.f.schauer@ucla.edu)
Department of Psychology, University of California, Los Angeles
Los Angeles, CA, USA

Peter C. Gordon (pcg@email.unc.edu)
Department of Psychology, University of North Carolina at Chapel Hill

Keith J. Holyoak (holyoak@lifesci.ucla.edu)
Department of Psychology, University of California, Los Angeles

Abstract

Neuropsychological investigations with frontal patients have revealed selective deficits in selecting the relational answer to pictorial analogy problems when the correct option is embedded among foils that exhibit high semantic or visual similarity. In contrast, normal age-matched controls solve the same problems with near-perfect accuracy regardless of whether high-similarity foils are present (in the absence of speed pressure). Using more sensitive measures, the present study sought to determine whether or not normal young adults are subject to such interference. Experiment 1 used eye-tracking while participants answered multiple-choice 4-term pictorial analogies. Total looking time was longer for semantically similar foils relative to an irrelevant foil. Experiment 2 presented the same problems in a true/false format with emphasis on rapid responding and found that reaction time to correctly reject false analogies was greater (and errors rates higher) for those based on semantically or visually similar foils. These findings demonstrate that healthy young adults are sensitive to both semantic and visual similarity when solving pictorial analogy problems. Results are interpreted in relation to neurocomputational models of relational processing.

Keywords: Analogy, semantics, perception, interference, eye-tracking, reaction time

Introduction

Relational reasoning—inferential processes constrained by the relational roles that entities play rather than the specific features of those entities—is a hallmark of human cognition. The basic components of relational processing have been investigated using a wide variety of analogy tasks. The simplest format for analogies involves four terms, expressed as either words or pictures, in the form *A:B::C:?*, where the task is to complete the analogy by selecting the best *D* term from a small set of options. By varying the alternative options, it is possible to assess the degree to which analogical reasoning is influenced by foils that pit semantic and/or visual similarity of individual

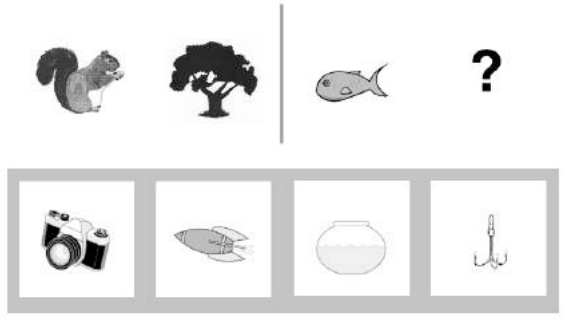


Figure 1. Example of a 4-term pictorial analogy with four alternatives, used in the present experiments (from Krawzcyk et al., 2008).

concepts or objects against relational similarity between pairs of concepts or objects. In the example shown in Figure 1, the task is to select the analogical option (*fish bowl*, based on the relation "lives in" that matches the *A:B* relation) from among a semantic distractor similar to the *C* term (*fish hook*), a visual distractor (*rocket*) and an unrelated option (*camera*).

Krawczyk et al. (2008) administered a set of picture analogies (from which the example shown in Figure 1 is drawn) to neuropsychological patients suffering from frontotemporal lobar degeneration (FTLD) and age-matched neurotypical controls (mean age approximately 60 years). Some of these problems were adapted from an earlier set created by Goranson (2002), and hence are dubbed the Goranson Analogy Test (GAT). In the study by Krawczyk et al., problems were administered one at a time, without speed pressure. In one problem set the options included distractors as in Figure 1; in an alternative set the semantic and perceptual distractors were replaced by two additional unrelated options.

For the set with similar distractors, patients with frontal-variant FTLD were correct on only 49% of the problems, rising to 84% correct for the set without similar distractors. An additional group of patients with temporal-variant FTLD showed a similar level of impairment regardless of whether similar foils were present, suggesting a general semantic deficit (see also Morrison et al., 2004). When similar distractors were present the patients with frontal damage selected similar distractors (mainly semantic, but also visual) more often than control participants. Indeed, the control group achieved near-perfect accuracy (98% correct). Thus, frontal damage appeared to selectively impair the ability to inhibit responding to pictorial analogy problems on the basis of superficial object similarity.

The near-perfect performance of the control participants in solving pictorial analogies even in the presence of similar distractors raises the question of whether and how cognitively unimpaired adults screen out object similarity (both semantic and visual) so as to focus on similarity of relations. Adults sometimes respond on the basis of object similarity when comparing more complex visual scenes (Markman & Gentner, 1993; Walz et al., 2000); however, the simple format of four-term pictorial analogies may allow non-relational information to be filtered out at a very early processing stage, so that choice of the analogical solution is not influenced by the presence of similar but non-relational foils. Alternatively, more sensitive measures may reveal evidence of response competition based on different varieties of similarity.

In two experiments, we investigated this question by administering versions of the GAT analogies used by Krawczyk et al. (2008) to healthy young adults. Eye-tracking methods provide one avenue for investigating online processing that occurs during analogical reasoning prior to making an overt decision (Gordon & Mozer, 2006; Glady, French, & Thibaut, 2016; Hayes, Petrov, & Sederberg, 2011; Vendetti et al., 2017). Accordingly, in Experiment 1 we collected data on gaze durations for the various response options while solving the GAT problems.

Another potentially more sensitive measure is reaction time (RT) to solve analogies under speed pressure. In Experiment 2 we changed the format of the GAT problems from four-alternative forced choice to true/false. For each of the original problems, each of the three foils was used to create a false picture analogy in the form *A:B::C:D*. In addition, participants were instructed to respond as quickly as possible. If semantic and/or visual similarity is screened out easily, then the various types of false analogies should take about the same length of time to reject. However, if college students are unable to avoid processing more superficial types of similarity, then decisions about false analogies in which the *D* term is similar (semantically or visually) to the *C* term may be relatively slow and error-prone.

Experiment 1

If superficial similarity intrudes into analogical reasoning for healthy adults, then they may spend more time looking at semantic and/or visual distractors than at an unrelated option.

Method

Participants. Participants were 32 undergraduates (24 female), mean age 20.4 years (range: 17–34) from the University of California, Los Angeles (UCLA), with normal or corrected to normal vision. They received course credit for participating.

Materials. Picture analogies were based on the 18 GAT problems used by Krawczyk et al. (2008). Two of these served as practice items, and 16 as experimental items. As in the Krawczyk et al. study, two sets of the 16 problems were created, one of which included similar foils and one of which replaced the semantic and visual foils with unrelated options.

Procedure. Pictorial analogies were presented on a computer screen one at a time. The size of each individual image (framed by a gray box) was 128 x 128 pixels (one-tenth of the screen width). A fixation cross was presented for 2 s, followed by the problem. The problem remained on until the participant pressed one of four response keys (corresponding to letters F, G, H, and J) to indicate which of the four alternatives was the correct analogical solution. When a response was made, the screen showed the reverse grayscale image for .25 s, after which the next trial began. Instructions did not emphasize speed of responding. During the experiment eye-tracking data were recorded using an Eyelink II gaze tracker (SR Research Ltd., Mississauga, Ontario, Canada), running under Eyelink Toolbox, PsychToolbox, and MATLAB on dual PCs. No feedback was provided.

For each participant, eight problems were included in the set with similar distractors (Distractor condition), and the other eight in the set with only unrelated foils (No-Distractor condition). Assignment of problems to set was counterbalanced across participants, as was the order of the four response options for each problem. Presentation order of the problems was randomized for each participant.

Results

Data were missing for one participant, who was excluded from analyses. Accuracy overall was 92% correct and did not vary reliably across the Distractor and No-Distractor conditions.

To guide analyses of eye movements, an invisible square of size 192 x 192 pixels around each individual image was defined as the location of that image. Figure 2 presents an example of a pattern of eye movements for an individual analogy problem in the Distractor condition.

To provide evidence of a possible pre-decisional influence of superficial similarity, we focused on dwell time (i.e., total looking times summed across all fixations) for each response option. Figure 3 plots the mean dwell time for each option in both the Distractor and No-Distractor conditions.

Participants' mean total time looking for each of the three foil images, in descending order, was: semantic foil (522 ms, *SE* = 38.2), visual foil (518 ms, *SE* = 54.9), and unrelated foil (404 ms, *SE* = 31.1). Overall, there was significant variation in dwell times depending on the foil condition, $F(2,60) = 3.93$, $p = 0.025$, $\eta^2 = .12$. Individual comparisons between conditions are reported with Bonferroni-corrected *p*-values. Semantic foils had longer dwell times relative to unrelated foils, $t(30) = 3.67$, $p < .001$, $\eta_P^2 = .31$.

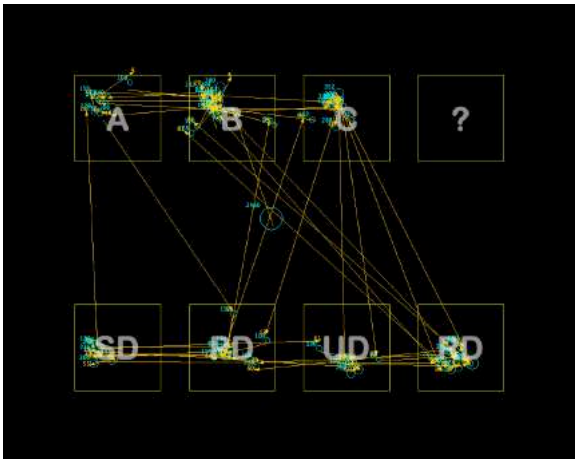


Figure 2. Example of pattern of eye movements during solution of a picture analogy. The above boxes (not visible to participants) indicate regions around the four images in the problem (*A*, *B*, *C*, *?*) and the four response options: semantically similar (S), visually similar (P), unrelated (U), and relational (R, the correct response). The D on each option label indicates this trial is from the Distractor condition.

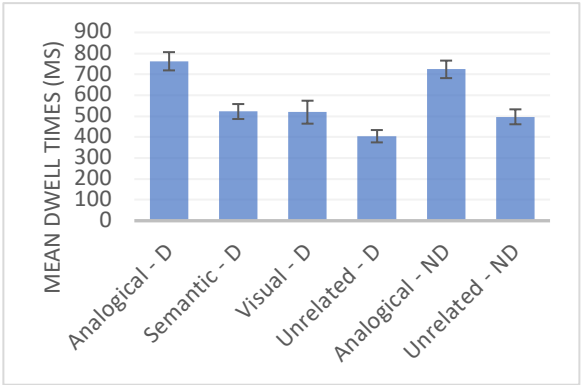


Figure 3. Total dwell time for each type of response image. The dwell time for Unrelated–ND is the mean across the three unrelated options provided in the No-Distractor condition. Error bars indicate +/- 1 standard error of the mean.

Visual foils also tended to have longer dwell times relative to unrelated foils. However, due to greater error variance in the visual foil condition, the latter difference was not reliable after the Bonferroni correction, $t(30) = 2.24$, $p = .098$, $\eta_P^2 = .14$. Dwell times for the two types of similar foils did not differ, n.s.

The eye-tracking data from Experiment 1 provide clear evidence that healthy adults are influenced by the presence of semantic and possibly visual distractors. Although response accuracy was high even in the presence of distractors, participants looked longer at semantically similar foils than at an unrelated option, suggesting that participants were sensitive to superficial similarity prior to making a decision.

Experiment 2

Experiment 2 used the same basic GAT analogies as in Experiment 1 but changed the format from 4-alternative forced choice to true/false. Instead of eye-tracking, the main dependent measure was RT to evaluate the problems under speed pressure.

Method

Participants. A total of 60 UCLA undergraduates (83% female) participated in the experiment. Their mean age was 20.8 years (range: 18–28), with normal or corrected to normal vision. They received course credit for participating.

Materials and Procedure. The experiment was conducted using a computer to display problems and record responses. The materials were based on the GAT problems used by Krawczyk et al. (2008). Each original problem was used to generate four true/false

problems, each showing four pictures. As shown in Figure 4, in each problem the *A:B* pair appeared at the top of the display and the *C:D* pair on the bottom. The *D* picture was either the correct analogical response (true), the semantic foil (false), the visual foil (false), or the unrelated option (false). Thus 25% of the problems were true analogies and 75% were false.

A set of four practice problems was created, using two of the GAT problems plus two additional problems taken from other sources. For the actual test trials, 16 analogy sets were created, one from each of the remaining 16 GAT problems. Figure 4 shows one of these sets. This procedure resulted in a total of 64 analogy problems. Each participant solved all 64 problems (i.e., a within-subjects design). To control for order effects the items were counterbalanced in the following way. The 16 sets were randomly assigned in equal numbers to Group A, B, C, or D. Thus, there were a total of 4 sets in each of the groups. Then, four test combinations were formed (I, II, III, IV). Combination I included only the items in Group A that had the analogical option, items in Group B that had the semantic foil option, items in Group C that had the visual foil option, and items in Group D that included the unrelated choice. Combinations II–IV were formed in the same basic manner, completing the counterbalancing of the four problems in each set. The presentation order of combinations I–IV was then counterbalanced across participants. Finally, the order of items within each combination was randomized for each participant.

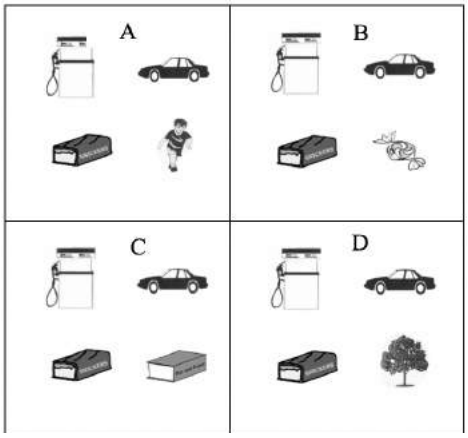


Figure 4. Example set of true/false picture analogies used in Experiment 2, created from the four alternatives of a single GAT problem. In each problem the *A:B* pair appears on top and the *C:D* pair on the bottom. The *D* term varies across problems. Panel A: Analogical (true); Panel B: semantic (false); Panel C: visual (false); Panel D: unrelated (false).

Participants were instructed to respond as quickly as possible while avoiding errors. They were told to press the *v* key to indicate "true" and the *n* key to indicate "false". Before the actual test trials, participants completed the four practice items (illustrating each of the four basic problem types) and were given feedback after each one. The correct answer was presented for 3,000 ms. No feedback was provided after test trials. On each test trial, a fixation cross appeared in the center of the screen for 1,000 ms before presentation of the analogy problem. The analogy problem remained visible until a response was made. The screen then went blank for 1,000 ms, after which the next fixation cross was presented.

Results

Both error rates and RTs for correct trials were analyzed. In Experiment 1, where the task was a four-alternative forced choice without speed pressure, error rates were low. In Experiment 2, by contrast, the speeded true/false task led to a substantial error rate. The mean error rate was 25% for analogical (true) problems, 48% for the problems with a semantic foil (false), 16% for the problems with a visual foil (false), and 7% for the problems with an unrelated foil (false). For the three types of false problems, a one-way within-subjects ANOVA was highly significant, $F(2, 118) = 150.31$, $p < .001$, $\eta^2 = .72$. Error rates were higher for the semantic foils than the unrelated foils, $t(59)=13.63$, $p < .001$, $\eta_P^2 = .76$. Error rates were also higher for the visual foils than the unrelated foils, $t(59)=5.42$, $p < .001$, $\eta_P^2 = .33$. Finally, semantic foils produced more errors than visual foils, $t(59)=12.66$, $p < .001$, $\eta_P^2 = .73$.

Figure 5 presents the mean correct RTs for each problem type. On average, participants took 3,047 ms to correctly verify problems with the analogical completion, 3,396 ms to correctly reject problems with the semantic foil, 2917 ms to reject those with the visual foil, and 2,518 ms to reject those with the unrelated foil. A within-subjects one-way ANOVA provided strong evidence for variation in RTs among the three types of false analogies, $F(2, 58) = 34.98$, $p < .001$, $\eta^2 = .37$. A Bonferroni correction was again applied to pairwise comparisons between foil conditions. False problems with semantic foils took longer to reject than those with unrelated foils, $t(59) = 6.62$, $p < .001$, $\eta_P^2 = 0.43$. Those with visual foils also yielded longer RTs compared with unrelated foils, $t(59) = 6.27$, $p < .001$, $\eta_P^2 = .40$. Finally, problems with semantic foils produced longer RTs than those with visual foils, $t(59) = 4.46$, $p < .001$, $\eta_P^2 = .77$.

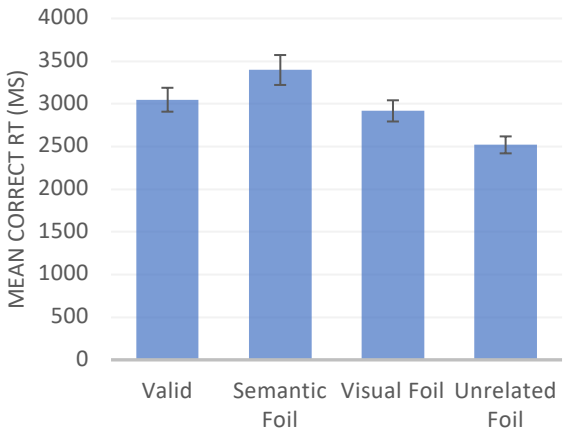


Figure 5. Mean correct RT for each type of picture analogy problem (Experiment 2). Error bars indicate +/- 1 standard error of the mean.

Discussion

The goal of the present study was to assess whether or not healthy young adults are influenced by semantic and/or visual similarity of distractors when solving four-term picture analogy problems. Previous work had indicated that in the absence of speed pressure, healthy older adults show little if any tendency to actually choose similar distractors over the correct, analogical option (Krawczyk et al., 2008). One possibility is that for reasoners with a fully functional frontal cortex, any tendency to select similar distractors is successfully inhibited (Morrison et al., 2004; Knowlton et al., 2012). But an alternative possibility is that healthy adults are able to reason by analogy without evoking a more superficial strategy based on comparing the similarity of *C* and *D* terms, so that superficial similarity simply does not enter into the analogical decision process.

Employing two different methods, the present study found evidence that college students are in fact influenced by both semantic and visual similarity when solving picture analogies. Using a four-alternative forced choice paradigm, in Experiment 1 we tracked eye movements while college students solved picture analogies in the absence of speed pressure. We found that dwell time (total looking time) was elevated for semantic (and possibly visual) foils during the period prior to selection of a response, even though the presence of similar distractors had little impact on the final choice. This finding suggests that similar distractors tended to draw extra attention, even though they were almost always rejected in favor of the analogical solution.

Experiment 2 examined solutions of the same basic picture analogies after they were recast in a true/false format and administered with instructions that emphasized speed of responding. In this situation, the similar distractors (especially the semantic foil) strongly influenced performance by college students. False analogies containing a semantic distractor as the *D* term were often erroneously judged to be true and took longer to correctly reject than any other condition. False analogies based on visual distractors also yielded higher error rates and higher correct RTs than did false analogies based on unrelated *D* terms.

The much more salient impact of similar distractors in Experiment 2 may be related to two ways in which its design differed from that used in both Experiment 1 and in the previous neuropsychological study by Krawczyk et al. (2008). First, speed pressure may be critical. When pressed to respond quickly, as in Experiment 2, there may not be time for inhibitory processes to effectively suppress a tendency to base decisions on superficial similarity.

Second, the true/false format used in Experiment 2 may also have played a role. In the four-alternative forced choice set-up, all options are simultaneously available for comparison, and a common criterion can be applied on an individual trial to determine the "best" alternative (e.g., Lu, Wu, & Holyoak, 2019; Lu, Liu, Ichien, Yuille, & Holyoak, 2019). In the true/false set-up, by contrast, each option has to be evaluated in isolation, and a criterion must be set on each trial to decide whether the analogy is "good enough" to respond "true". Since feedback was never given in our experiments, participants may have been uncertain about the appropriate criterion (especially since the ratio of true and false analogies was unbalanced). Given that the analogies were best solved on the basis of semantic relations, problems including a semantic lure (i.e., those in which the *C* term is semantically related to the *D* term, but not in the same way that *A* is related to *B*) may have often passed the subjective decision criterion, resulting in errors.

Taken together, the present findings seem to rule out the hypothesis that superficial similarity plays no role in analogical reasoning for healthy adults. Depending on test conditions, semantic and visual lures may have relatively subtle effects (a tendency to attract visual attention) or extremely salient effects (generating either errors or slow correct responses).

It would seem, therefore, that our results favor the standard view that analogical reasoning is susceptible to interference from a non-analogical strategy of simply evaluating the similarity of the *C* and *D* terms, without reference to the *A:B* relation. However, another alternative deserves consideration. The analogy "game" bases the correct answer on the most specific possible relation(s) in common across *A:B* and *C:D* (e.g., for the analogy shown in Figure 1, the specific relation "lives in" links squirrel to tree and

also fish to fishbowl). But suppose relations emerge in a gradual fashion during the reasoning process, rather than simply being retrieved in an all-or-none fashion. Then the *A:B* and *C:D* relations may at first be vague or incomplete, and only over time reach full specificity. Early in this process of relation encoding, the active relation between *A:B* may be something very general (e.g., a squirrel is somehow related, either semantically or visually, to a tree). At this point, one or both of the foils may match the crude *A:B* relation about equally well as the analogical answer (e.g., a fish is associated with a fishbowl, and similar visually to the pictured rocket). Under this view, speed pressure may force the reasoner to choose the "best" answer before the relations are fully encoded, at a point in time when the analogical answer and the similar foils may be comparable in their degree of match to the partially-encoded *A:B* relation.

This alternative account of interference implies its source may not be a rival non-analogical strategy (e.g., simply comparing *C* and *D* while ignoring *A:B*). Rather, interference may emanate from the analogy process itself, if a fast decision is required when relations are as yet poorly encoded. Future research should attempt to test these alterative accounts of how superficial similarity can infiltrate a process that aims to focus on relations.

Acknowledgements

We thank Robin Gruber and Matthew Weiden for assistance in running Experiment 1, and Brandon Valenica, Justin Shin and Kiran Cherian for assistance with Experiment 2. A preliminary report of Experiment 1 was presented at the Forty-ninth Annual Meeting of the Psychonomic Society (Chicago, November 2008). Preparation of this paper was supported by NSF Grant BCS-1827374 to KJH.

References

Glady, Y., French, R. M., & Thibaut, J. P. (2016). Comparing competing views of analogy making using eye-tracking technology. In *Proceedings of the 38th Annual Meeting of the Cognitive* Science Society (pp. 1349-1354). Austin, TX: Cognitive Science Society.

Goranson, T. E. (2002). On diagnosing Alzheimer's disease: Assessing abstract thinking and reasoning. *Dissertation Abstracts International: Section B: Sciences and Engineering*, *62*, 4785.

Gordon, P. C., & Moser, S. (2007). Insight into analogies: Evidence from eye movements. *Visual Cognition, 15*(1), 20–35.

Hayes, T. R., Petrov, A. A., & Sederberg, P. B. (2011). A novel method for analyzing sequential eye movements reveals strategic influence on Raven's Advanced Progressive Matrices. *Journal of Vision, 11*, 1–11.

Knowlton, B. J., Morrison, R. G., Hummel, J. E., & Holyoak, K. J. (2012). A neurocomputational system for relational reasoning. *Trends in Cognitive Sciences*, *16*, 373-381.

Krawczyk, D. C., Morrison, R. G., Viskontas, I., Holyoak, K. J., Chow, T. W., Mendez, M. F., Miller, B. L., & Knowlton, B. J. (2008). Distraction during relational reasoning: The role of prefrontal cortex in interference control. *Neuropsychologia*, *46*(7), 2020–2032.

Lu, H., Liu, Q., Ichien, N., Yuille, A. L., & Holyoak, K. J. (2019). Seeing the meaning: Vision meets semantics in solving visual analogy problems. In A. Goel, C. Seifert, & C. Freksa (Eds.), *Proceedings of the 41st Annual Meeting of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Lu, H., Wu, Y. N., & Holyoak, K. (2019). Emergence of analogy from relation learning. *Proceedings of the National Academy of Sciences, USA*, *116*, 4176-4181.

Markman, A. B., & Gentner, D. (1993). Structural alignment during similarity comparisons. *Cognitive Psychology, 25,* 431-467.

Morrison, R. G., Krawczyk, D. C., Holyoak, K. J., Hummel, J. E., Chow, T. W., Miller, B. L., & Knowlton, B. J. (2004). A neurocomputational model of analogical reasoning and its breakdown in frontotemporal lobar degeneration. *Journal of Cognitive Neuroscience*, *16*(2), 260–271.

Vendetti, M. S., Starr, A., Johnson, E. L., Modavi, K., & Bunge, S. A. (2017). Eye movements reveal optimal strategies for analogical reasoning. *Frontiers in Psychology*, *8*, 932. doi:10.3389/fpsyg.2017.00932

Waltz, J. A., Lau, A., Grewal, S. K., & Holyoak, K. J. (2000). The role of working memory in analogical mapping. *Memory & Cognition*, *28*, 1205-1212.

An Examination of Perseveration Terms in Reinforcement Learning Models

Darrell Worthy
Texas A&M University, College Station, Texas, United States

Astin Cornwall
Texas A&M University, College Station, Texas, United States

Hilary Don
Texas A&M University, College Station, Texas, United States

Abstract

Perseveration, or stickiness parameters have been added to reinforcement-learning (RL) models to capture autocorrelation in choices. Here, we systematically examined whether perseveration terms simply improve a models ability to fit noise in the data, thereby making them overly flexible. We simulated data with basic versions of a Delta and Prediction-Error Decay model with no perseveration terms added, and for half of the simulated data sets we added random noise to expected RL values on each trial. We then performed cross-fitting analyses where the simulated data sets were fit by the basic data-generating models as well as extended models with perseveration terms added. The addition of perseveration terms improved model fit, particularly when noise was added to the simulation process. Parameter recovery was generally poorer for the extended models. These results suggest simpler models may be more useful for prediction and generalization to novel environments, as well as for theory development.

Generalization as diffusion: human function learning on graphs

Charley M. Wu[1] **(cwu@mpib-berlin.mpg.de), Eric Schulz**[2]**, Samuel J. Gershman**[2]
[1]Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin, Germany
[2]Department of Psychology, Harvard University, Cambridge, MA, USA

Abstract

From social networks to public transportation, graph structures are a ubiquitous feature of life. How do humans learn functions on graphs, where relationships are defined by the connectivity structure? We adapt a Bayesian framework for function learning to graph structures, and propose that people perform generalization by assuming that the observed function values diffuse across the graph. We evaluate this model by asking participants to make predictions about passenger volume in a virtual subway network. The model captures both generalization and confidence judgments, and provides a quantitatively superior account relative to several heuristic models. Our work suggests that people exploit graph structure to make generalizations about functions in complex discrete spaces.
Keywords: Function Learning, Graph structures, Gaussian Process, Generalization, Successor Representation

Introduction

Most of function learning research has focused on how people learn a relationship between two continuous variables (Mcdaniel & Busemeyer, 2005; Lucas, Griffiths, Williams, & Kalish, 2015; DeLosh, Busemeyer, & McDaniel, 1997). How much hot sauce should I add to enhance my meal? How hard should I push a child on a swing? While function learning on continuous spaces is ubiquitous, many other relationships in the world are defined by functions on discrete spaces. For example, navigating a subway network and constructing a bookshelf both require representation of functions mapping discrete inputs (subway stops and configurations of components) to continuous outputs (passenger volume and probability of success). Likewise, language, commerce, and social networks are all defined partly by discrete relationships. How do people learn functions on discrete graph structures?

We propose that a diffusion kernel provides a suitable similarity metric based on the transition structure of a graph. When combined with the Gaussian Process (GP) regression framework, we arrive at a model of how humans learn functions and perform inference on graph structures. Using a virtual subway network prediction task, we pit this model against heuristic alternatives, which perform inference with lower computation demands, but are unable to capture human inference and confidence judgments. We also show that the diffusion kernel can be related to prominent models in continuous function learning and models of structure learning. This opens up a rich set of theoretical connections across theories of human learning and generalization.

Computational Models of Function Learning

Based on a limited set of observations, how can you interpolate or extrapolate to predict unobserved data? This question has been the focus of human function learning research, which has traditionally studied predictions in continuous spaces (e.g., the relationship between two variables; Busemeyer, Byun, DeLosh, & McDaniel, 1997). Function learning research has revealed how inductive biases guide learning (Kwantes & Neal, 2006; Kalish, Griffiths, & Lewandowsky, 2007; Schulz, Tenenbaum, Duvenaud, Speekenbrink, & Gershman, 2017) and which types of functions are easier or harder to learn (Schulz, Tenenbaum, Reshef, Speekenbrink, & Gershman, 2015).

Several theories have been proposed to account for how humans learn functions. Earlier approaches used rule-based models that assumed a specific parametric family of functions (e.g., linear or exponential; Brehmer, 1974; Carroll, 1963; Koh & Meyer, 1991). However, the rigidity of rule-based learning struggled to account for order-of-difficulty effects in interpolation tasks (Mcdaniel & Busemeyer, 2005), and could not capture the biases displayed in extrapolation tasks (DeLosh et al., 1997).

An alternative approach relied on similarity-based learning, using connectionist networks to associate observed inputs and outputs (DeLosh et al., 1997; Kalish, Lewandowsky, & Kruschke, 2004; Mcdaniel & Busemeyer, 2005). The similarity-based approach is able to capture how people interpolate, but fails to account for some of the inductive biases displayed in extrapolation and in the partitioning of the input space. In some cases, hybrid architectures were developed to incorporate rule-based functions in a associative framework (e.g., Kalish et al., 2004; Mcdaniel & Busemeyer, 2005) in an attempt to gain the best of both worlds.

More recently, a theory of function learning based on GP regression was proposed to unite both accounts (Lucas et al., 2015), because of its inherent duality as both a rule-based and a similarity-based model. GP regression is a non-parametric method for performing Bayesian function learning (Schulz, Speekenbrink, & Krause, 2018), has successfully described human behavior across a range of traditional function learning paradigms (Lucas et al., 2015), and can account for compositional inductive biases (e.g., combining periodic and long range trends; Schulz et al., 2017).

While the majority of function learning research has studied continuous spaces, many real-world problems are discrete (Kemp & Tenenbaum, 2008). In a completely unstructured discrete space, the task of function learning is basically hopeless, because there is no basis for generalization across inputs. Fortunately, most real-world problems have structure

(Tenenbaum, Kemp, Griffiths, & Goodman, 2011), which we can often represent as a connectivity graph that encodes how inputs (nodes) relate to each other (see Kemp & Tenenbaum, 2008, for a similar argument). By assuming that functions vary smoothly across the graph (a notion we formalize below), functions can be generalized to unobserved inputs. Although this idea has been studied extensively in machine learning, it has not yet fully permeated into studies of human function learning.

Goals and Scope

We describe a model of learning graph-structured functions using a diffusion kernel. The diffusion kernel specifies the covariance between function values at different nodes of a graph based on its connectivity structure. When combined with the GP framework, it allows us to make Bayesian predictions about unobserved nodes. Even though previous work has investigated how people learn the relational structure of a graph (Kemp & Tenenbaum, 2008; Kemp, Tenenbaum, Griffiths, Yamada, & Ueda, 2006; Tomov, Yagati, Kumar, Yang, & Gershman, 2018), or infer properties of unobserved inputs (Kemp & Tenenbaum, 2009; Kemp, Shafto, & Tenenbaum, 2012), less is known about how people learn functions in discrete spaces with real-valued outputs.

We present an experiment where participants are shown a series of randomly generated subway maps and asked to predict the number of passengers at unobserved stations to test our model of function learning on graphs. In addition, we collected confidence judgments from participants. We compared the GP diffusion kernel model to heuristic models based on nearest-neighbor interpolation.

Function Learning on Graphs

We can specify a graph $G = (\mathcal{N}, \mathcal{E})$ with nodes $n_i \in \mathcal{N}$ and edges $e_i \in \mathcal{E}$ to represent a structured state space (Fig. 1a). Nodes represent states and edges represent connections. For now, we assume that all edges are undirected (i.e., if $x \to y$ then $y \to x$).

The diffusion kernel (Kondor & Lafferty, 2002) defines a similarity metric $k(s,s')$ between any two nodes on a graph based on the matrix exponentiation of the graph Laplacian:

$$k(s,s') = e^{\alpha L}, \tag{1}$$

where L is the graph Laplacian:

$$L = D - A, \tag{2}$$

with the adjacency matrix A and the degree matrix D. Each element a_{ij} is 1 when nodes i and j are connected, and 0 otherwise, while the diagonals of D describe the number of connections of each node. The graph Laplacian can also describe graphs with weighted edges, where D becomes the weighted degree matrix and A becomes the weighted adjacency matrix.

Intuitively, the diffusion kernel assumes that function values diffuse along the edges similar to a heat diffusion process (i.e., the continuous limit of a random walk). The free parameter α governs the rate of diffusion, where $\alpha \to 0$ assumes complete independence between nodes, while $\alpha \to \infty$ assumes all nodes are perfectly correlated.

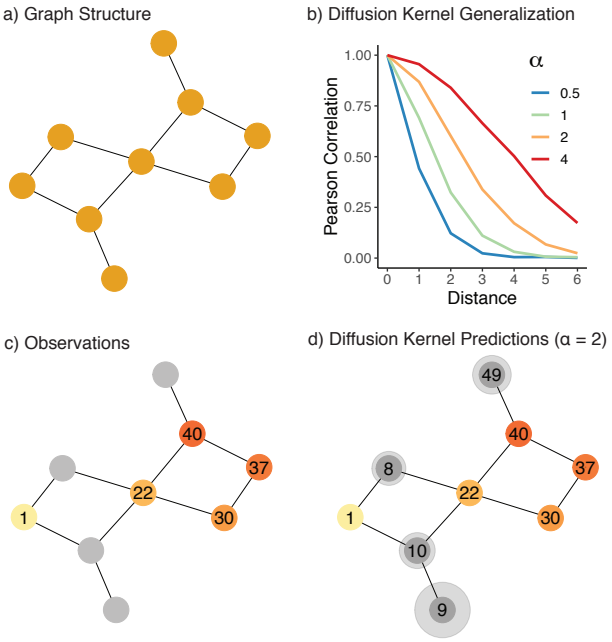


Figure 1: Graph-structured function learning. **a**) An example of a graph structure, where nodes represent states and edges indicate the transition structure. **b**) A diffusion kernel is a similarity metric between nodes on a graph, allowing us to generalize to unobserved nodes based on the assumption that the correlation between function values decays as an exponential function of the distance between two nodes. The diffusion parameter (α) governs the rate of decay. **c**) Given some observations on the graph (colored nodes), we can use the diffusion kernel combined with the Gaussian Process framework to make predictions (**d**) about expected rewards (numbers in grey nodes) and the underlying uncertainty (size of halo) for each unobserved node.

From the similarity metric defined by the diffusion kernel, we can use the GP regression framework (Rasmussen & Williams, 2006) to perform Bayesian inference over graph-structured functions. A GP defines a distribution over functions $f : \mathcal{S} \to \mathbb{R}^n$ that map the input space $\mathcal{S}$ to real-valued scalar outputs:

$$f \sim \mathcal{GP}(m,k), \tag{3}$$

where $m(s)$ is a mean function specifying the expected output of s, and $k(s,s')$ is the covariance function (kernel) that encodes prior assumptions about the smoothness of underlying function. Any finite set of function values drawn from a GP is multivariate Gaussian distributed.

We use the diffusion kernel (Eq. 1) to represent the covariance $k(s,s')$ based on the connectivity structure of the graph, and follow the convention of setting the mean function to zero, such that the GP prior is fully defined by the

kernel.

Given some observations $\mathcal{D}_t = \{\mathbf{y}_t, \mathbf{s}_t\}$ of observed outputs $\mathbf{y}_t$ at states $\mathbf{s}_t$, we can compute the posterior distribution $p(f(s_*)|\mathcal{D}_t)$ for any target state s_*. The posterior is a normal distribution with mean and variance defined as:

$$m(s_*|\mathcal{D}_t) = \mathbf{k}_*^\top (\mathbf{K} + \sigma_\varepsilon^2 \mathbf{I})^{-1} \mathbf{y}_t \quad (4)$$

$$v(s_*|\mathcal{D}_t) = k(s_*, s_*) - \mathbf{k}_\star^\top (\mathbf{K} + \sigma_\varepsilon^2 \mathbf{I})^{-1} \mathbf{k}_* \quad (5)$$

where $\mathbf{K}$ is the $t \times t$ covariance matrix evaluated at each pair of observed inputs, and $\mathbf{k}_* = [k(s_1, s_*), \ldots, k(s_t, s_*)]$ is the covariance between each observed input and the target input s_*, and σ_ε^2 is the noise variance. Thus, for any node in the graph, we can make Bayesian predictions (Fig. 1e) about the expected function value $m(s_*|\mathcal{D}_t)$ and also the level of uncertainty $v(s_*|\mathcal{D}_t)$.

The posterior mean function of a GP can be rewritten as:

$$m(s) = \sum_{i=1}^{t} w_i k(s_i, s) \quad (6)$$

where each s_i is a previously observed state and the weights are collected in the vector $\mathbf{w} = \left[k(\mathbf{S}_t, \mathbf{S}_t) + \sigma_\varepsilon^2 \mathbf{I}\right]^{-1} \mathbf{y}_t$. Intuitively, this means that GP regression is equivalent to a linearly-weighted sum using basis functions $k(s_i, s)$ to project observed states onto a feature space (Schulz, Speekenbrink, & Krause, 2018). To generate new predictions for an unobserved state s, each output y_t is weighted by the similarity between observed states s_t and the target state s.

Connections to Function Learning On Continuous Domains

The GP framework allows us to relate similarity-based function learning on graphs to theories of function learning in continuous domains. Consider the case of an infinitely fine lattice graph (i.e., a grid-like graph with equal connections for every node and with the number of nodes and connections approaching continuity). Following Kondor and Lafferty (2002) and using the diffusion kernel defined by Eq. 1, this limit can be expressed as

$$k(s, s') = \frac{1}{\sqrt{(4\pi\alpha)}} \exp\left(\frac{-|s - s'|}{4\alpha}\right), \quad (7)$$

which is equivalent to a Radial Basis Function (RBF) kernel. Models similar to the RBF kernel are prominent in the literature on function learning in continuous domains (Busemeyer et al., 1997; Lucas et al., 2015). The RBF kernel has also been used to model how humans generalize about unobserved rewards in exploration tasks (Wu, Schulz, Speekenbrink, Nelson, & Meder, 2018). Thus, the RBF kernel can be understood as a special case of the diffusion kernel, when the underlying structure is symmetric and infinitely fine.

More broadly, both the RBF and diffusion kernel can be understood as instantiations of Shepard's (1987) "universal law of generalization" in a function learning domain, by expressing generalization as an exponentially decaying function of the distance between two stimuli. Shepard famously proposed that the law of generalization should be the first law of psychology, while recent work has further entrenched it in fundamental properties of efficient coding (Sims, 2018) and measurement invariance (Frank, 2018).

Heuristic Models

We compare the GP model to two heuristic strategies for function learning on graphs, which make predictions about the rewards of a target state s_* based on a simple nearest neighbors averaging rule. The *k-Nearest Neighbors* (kNN) strategy averages the function values of the k closest states (including all states with same shortest path distance as the k-th closest), while the *d-Nearest Neighbors* (dNN) strategy averages the function values of all states within path distance d. Both kNN and dNN default to a prediction of 25 when the set of neighbors are empty (i.e., the median value in the experiment).

Both the dNN and kNN heuristics approximate the local structure of a correlated graph structure with the intuition that nearby states have similar function values. While they sometimes make the same predictions as the GP model and have lower computational demands, they fail to capture the connectivity structure of the graph and are unable to learn directional trends. Additionally, they only provide point-estimate predictions, and thus do not capture the underlying uncertainty of a prediction (which we use to model confidence judgments).

Experiment: Subway Prediction Task

We used a "Subway Prediction Task" to study how people perform function learning in graph-structured state spaces. Participants were shown a series of graphs described as subway maps, where nodes corresponded to stations and edges indicated connections (Fig. 2). Participants were asked to predict the number of passengers (in a randomly selected train car) at a target station, based on observations from other stations.

Methods and procedure

We recruited 100 participants ($M_{age} = 32.7$; $SD = 8.4$; 28 female) on Amazon MTurk to perform 30 rounds of a graph prediction task. On each graph, numerical information was provided about the number of passengers at 3, 5, or 7 other stations (along with a color aid), from which participants were asked to predict the number of passengers at a target station and provide a confidence judgment (Likert scale from 1-11). The subway passenger cover story was used to provide intuitions about graph correlated functions. Additionally, participants observed 10 fully revealed graphs to familiarize themselves with the task and completed a comprehension check before starting the task. Participants were paid a base fee of $2.00 USD for participation with an additional performance contingent bonus of up to $3.00 USD. The

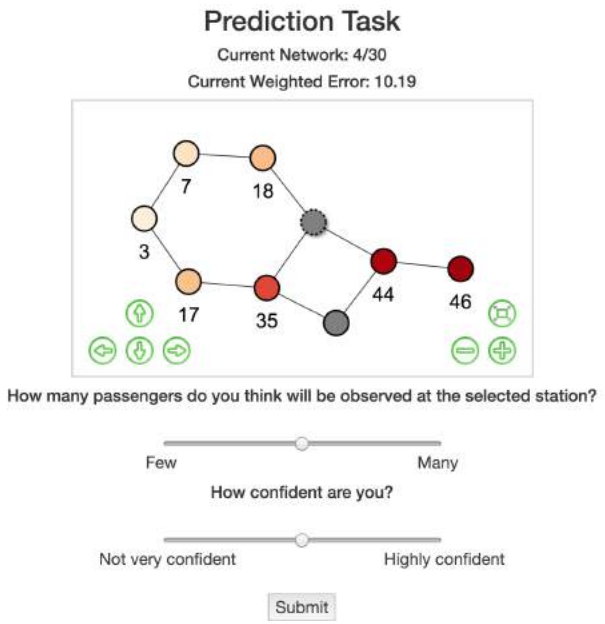


Figure 2: Screenshot from the Subway Prediction Experiment. Observed nodes (3, 5, or 7 randomly sampled nodes depending on the information condition) are shown with a numerical value and a corresponding color aid (darker indicates larger values). The target node is indicated by the dashed line, and dynamically changes color and displays a numerical value when participants move the top slider. Confidence judgments were used to compute a weighted error (i.e., more confident answers having a larger contribution), which was used to determine the performance contingent bonus.

bonus payment was based on the mean absolute judgment error weighted by confidence judgments: $R_{bonus} = \$3.00 \times (25 - \sum_i \tilde{c}_i \varepsilon_i)/25$ where $\tilde{c}_i$ is the normalized confidence judgment $\tilde{c}_i = \frac{c_i}{\sum c_j}$ and ε_i is the absolute error for judgment i. On average, participants completed the task in 8.09 minutes ($SD = 3.7$) and earned \$3.87 USD ($SD = \$0.33$).

In each of the 30 rounds, a different graph was sampled without replacement. We used three different information conditions (observations $\in [3,5,7]$; each used in 10 rounds in randomly shuffled order) as a within-subject manipulation determining the number of randomly sampled nodes with revealed information. In each round, participants were asked to predict the value of a target node, which was randomly sampled from the remaining unobserved nodes.

All participants observed the same set of 40 graphs that were sampled without replacement for the 10 fully revealed examples in the familiarization phase and for the 30 graphs in the prediction task. We generated the set of 40 graphs by iteratively building 3×3 lattice graphs (also known as mesh or grid graphs), and then randomly pruning 2 out of the 12 edges. In order to generate the functions (i.e., number of passengers), we fit a diffusion kernel to the graph and then sampled a single function from a GP prior, where the diffusion parameter was set to $\alpha = 2$.

Results

Figure 3 shows the behavioral and model-based results of the experiment. We applied linear mixed-effects regression to estimate the effect of the number of observed nodes on participant prediction errors, with participants as a random effect. Participants made systematically lower error predictions as the number of observable nodes increased ($\beta = -.11$, $t(99) = -6.28$, $p < .001$, $BF > 100$[1]; Fig. 3a). Repeating the same analysis but using participant confidence judgments as the dependent variable, we found that confidence increased with the number of observable nodes ($\beta = .16$, $t(99) = 11.3$, $p < .001$, $BF > 100$; Fig. 3b). Finally, participants were also able to calibrate confidence judgments to the accuracy of their predictions, with higher confidence predictions having consistently lower error ($\beta = -.19$, $t(99) = -9.0$, $p < .001$, $BF > 100$; Fig. 3c). There were no substantial effects of learning over rounds ($\beta = .01$, $t(99) = 0.47$, $p = .642$, $BF = 0.2$), suggesting the familiarization phase and cover story were sufficient for providing intuitions about graph correlated structures.

Model comparison

We compare the predictive performance of the GP with the dNN and kNN heuristic models. Using participant-wise leave-one-out cross-validation, we estimate model parameters for all but one judgment, and then make out-of-sample predictions for the left-out judgment. We repeat this procedure for all trials and compare predictive performance using Root Mean Squared Error (RMSE) over all left-out trials.

Figure 3d shows that the GP made better predictions than both the dNN ($t(99) = -4.06$, $p < .001$, $d = 0.41$, $BF > 100$) and kNN models ($t(99) = -7.19$, $p < .001$, $d = 0.72$, $BF > 100$). Overall, 58 out of 100 participants were best predicted by the GP, 31 by the dNN, and 11 by the kNN. Figure 3e shows individual parameter estimates of each model. The estimated diffusion parameter α was not substantially different from the ground truth of $\alpha = 2$ ($t(99) = -0.66$, $p = .51$, $d = 0.07$, $BF = 0.14$), although the distribution appeared to be bimodal, with participants often underestimating or overestimating the correlational structure. Estimates for d and k were highly clustered around the lower limit of 1, suggesting that averaging over larger portions of the graph were not consistent with participant predictions.

Finally, an advantage of the GP is that it produces Bayesian uncertainty estimates for each prediction. While the dNN and kNN models make no predictions about confidence, the GP uncertainty estimates correspond to participant confidence judgments ($\beta = -.10$, $t(99) = -3.39$, $p < .001$, $BF > 100$; linear mixed-effects model with participant as a random effect).

[1] β is the standardized effect size $\in [-1,1]$ and we approximate the Bayes Factor using bridge sampling (Gronau, Singmann, & Wagenmakers, 2017) to compare our model to an alternative intercept only null model, where both models were hierarchical regressions but only the alternative model contained the variable of interest as a regressor.

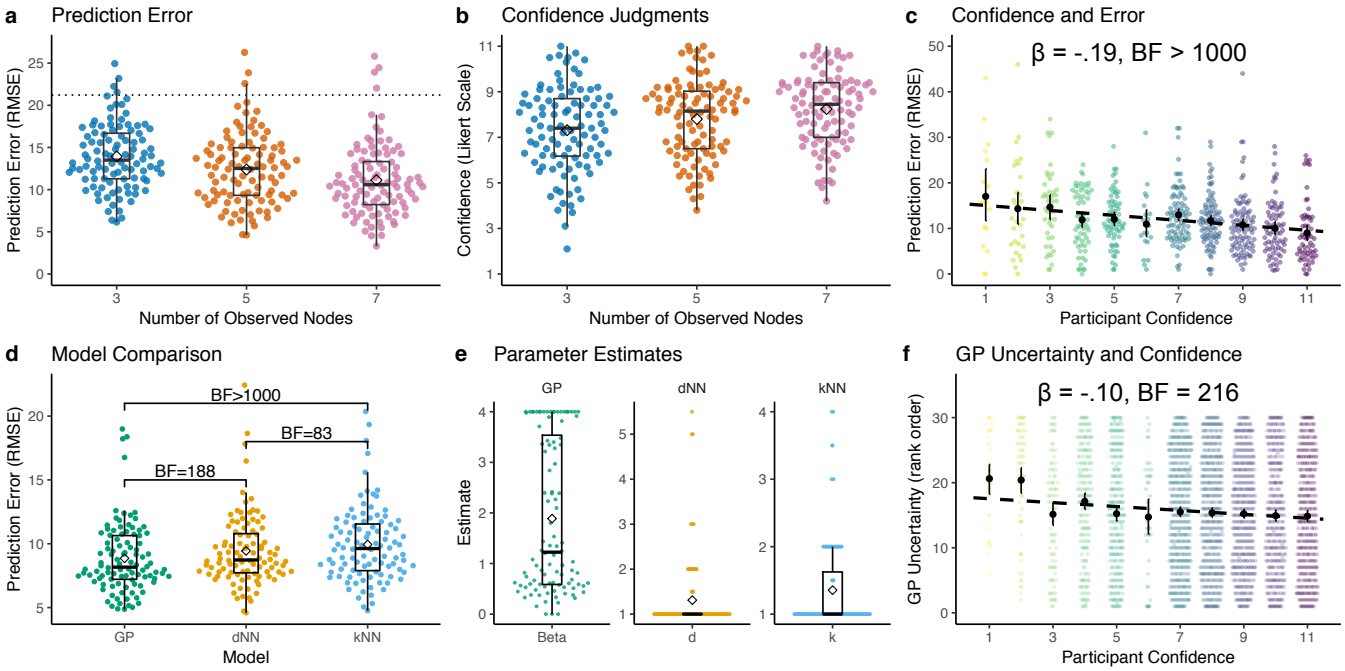


Figure 3: Results. **a-b**) Participant judgment errors and confidence estimates. Each dot is a single participant (averaged over each number of observed nodes), with Tukey boxplots and diamonds indicating group means. The dotted line in **a**) is a random baseline. **c**) Judgment error and confidence. Each colored dot is a participant (averaged over each confidence level), dashed line is a linear regression, with black dots and error bars indicating group means and 95% CI. We report the mixed-effects regression coefficient and Bayes Factor above. **d**) Cross-validated model comparison between the Gaussian Process with diffusion kernel (GP), d-nearest neighbors (dNN), and k-nearest neighbors (kNN). Each point is a single participant with a Tukey boxplot overlaid and diamonds indicating group means. Comparisons are for a Bayesian one-sample t-test, where the null hypothesis posits no difference between models and assumes a Cauchy prior with the scale set to $\sqrt{2}/2$. **e**) Parameter estimates, where each dot is the mean cross-validated estimate for each participant, with Tukey boxplots and diamonds indicating group means. **f**) GP uncertainty estimates (rank ordered within participant) and participant confidence judgments (Likert scale). Dotted line is a linear regression, with black dots and error bars indicating mean and 95% CI. We report the mixed-effects regression coefficient and Bayes factor (see text for details).

Discussion

How do people learn about functions on structured discrete spaces like graphs? We show how a GP with a diffusion kernel can be used as a model of function learning that produces Bayesian predictions about unobserved nodes. Our model integrates existing theories of human function learning in continuous spaces, where the RBF kernel (commonly used in continuous domains) can be seen as a special limiting case of the diffusion kernel. Using a virtual subway task, we show that the GP was able to capture how people make judgments about unobserved nodes and is also able to generate uncertainty estimates that correspond to participant confidence ratings.

Related work

Previous work has also investigated how people perform inference over graphs (Kemp & Tenenbaum, 2009, 2008; Shafto, Kemp, Baraff, Coley, & Tenenbaum, 2005; Tomov et al., 2018). Whereas these studies were geared towards probing how people inferred underlying structure (Kemp & Tenenbaum, 2008) and how (implicit or explicit) representations of structure influenced causal property judgments (Kemp & Tenenbaum, 2009; Shafto et al., 2005), the goal of our Subway Prediction Task was to study how people perform functional inference given explicit knowledge of a relational structure. Thus, our study can be seen as a real-valued extension of the experiments presented in Kemp and Tenenbaum (2009) and Shafto et al. (2005), where we explicitly present the underlying structure and modeled both participants predictions and their confidence judgments simultaneously.

Our approach also has formal similarities to Kemp and Tenenbaum (2008, 2009), who used a kernel defined as $k(s,s') = (L + \frac{I}{\sigma^2})^{-1}$ to generate feature vectors over structured representations, in order to approximate a prior over properties distributed across the graph. This kernel is a reformulation of the regularized Laplacian kernel[2] (Zhu, Lafferty, & Ghahramani, 2003), which belongs to the same broad framework of regularization operators (Smola & Kondor, 2003) as the diffusion kernel (Eq. 1), with both providing similar inductive biases of smoothness over the graph struc-

[2] $k(s,s') = (I + \sigma^2 L)^{-1}$

ture.

Future Work and Limitations

Currently, we have only studied how people learn functions on spatial representations of graph structures, where all nodes and edges are visible simultaneously. However, people can perform inferences over discrete structures that are more conceptual such as social hierarchies (Lau, Pouncy, Gershman, & Cikara, 2018) or causal connections (Rothe, Deverett, Mayrhofer, & Kemp, 2018). Given that the GP framework can be used to compare how people learn functions over different (i.e., spatial and conceptual) domains (Wu, Schulz, Garvert, Meder, & Schuck, 2018), comparing functional inference over conceptual and spatial graphs seems like promising extension for future studies.

Additionally, one could also assess the suitability of the diffusion kernel as a model for more complex problems, such as multi-armed bandit tasks with structured rewards (e.g., Schulz, Franklin, & Gershman, 2018) and in planning problems, where exploration plays a fundamental role. One advantage of the GP diffusion kernel model is that it makes prediction with estimates of the underlying uncertainty. Whereas many models of generalization only make point-estimates about the value of a state, the GP framework offers opportunities for using uncertainty-guided exploration strategies (e.g., Auer, 2002).

One limitation of the diffusion kernel is that it assumes *a priori* knowledge of the graph structure. While this may be a reasonable assumption in problems such as navigating a subway network where one can simply look at a map, this is not always the case. In contrast, the SR can learn the graph structure through experience (using prediction-error updating). Thus, the connection between the SR and the diffusion kernel presents a promising avenue for incorporating a plausible process model of structure learning.

Conclusion

We show that GP regression, together with a diffusion kernel, captures how participants learn functions and make confidence ratings on graph structures in a virtual subway prediction task. Our model opens up a rich set of theoretical connections to theories of function learning on continuous domains and models of structure learning and property induction.

Acknowledgements

CMW is supported by the International Max Planck Research School on Adapting Behavior in a Fundamentally Uncertain World; ES is supported by the Harvard Data Science Initiative

References

Auer, P. (2002). Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, *3*, 397–422.

Brehmer, B. (1974). Hypotheses about relations between scaled variables in the learning of probabilistic inference tasks. *Organizational Behavior and Human Performance*, *11*(1), 1–27.

Busemeyer, J. R., Byun, E., DeLosh, E. L., & McDaniel, M. A. (1997). Learning functional relations based on experience with input-output pairs by humans and artificial neural networks. In K. Lamberts & D. Shanks (Eds.), *Concepts and Categories* (p. 405-437). Cambridge: MIT Press.

Carroll, J. D. (1963). Functional learning: The learning of continuous functional mappings relating stimulus and response continua. *ETS Research Bulletin Series*, *1963*(2), i–144.

DeLosh, E. L., Busemeyer, J. R., & McDaniel, M. A. (1997). Extrapolation: The sine qua non for abstraction in function learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(4), 968.

Frank, S. A. (2018). Measurement invariance explains the universal law of generalization for psychological perception. *Proceedings of the National Academy of Sciences*, *115*(39), 9803–9806. doi: 10.1073/pnas.1809787115

Gronau, Q. F., Singmann, H., & Wagenmakers, E.-J. (2017). Bridgesampling: An R package for estimating normalizing constants. *arXiv preprint arXiv:1710.08162*.

Kalish, M. L., Griffiths, T. L., & Lewandowsky, S. (2007). Iterated learning: Intergenerational knowledge transmission reveals inductive biases. *Psychonomic Bulletin & Review*, *14*(2), 288–294.

Kalish, M. L., Lewandowsky, S., & Kruschke, J. K. (2004). Population of linear experts: knowledge partitioning and function learning. *Psychological Review*, *111*(4), 1072.

Kemp, C., Shafto, P., & Tenenbaum, J. B. (2012). An integrated account of generalization across objects and features. *Cognitive Psychology*, *64*(1-2), 35–73.

Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences*, *105*(31), 10687–10692.

Kemp, C., & Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychological Review*, *116*(1), 20–58.

Kemp, C., Tenenbaum, J. B., Griffiths, T. L., Yamada, T., & Ueda, N. (2006). Learning systems of concepts with an infinite relational model. In *AAAI* (Vol. 3, p. 5).

Koh, K., & Meyer, D. E. (1991). Function learning: Induction of continuous stimulus-response relations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(5), 811–836.

Kondor, R. I., & Lafferty, J. (2002). Diffusion kernels on graphs and other discrete structures. In *Proceedings of the 19th International Conference on Machine Learning* (Vol. 2002, pp. 315–322).

Kwantes, P. J., & Neal, A. (2006). Why people underestimate y when extrapolating in linear functions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(5), 1019–1030.

Lau, T., Pouncy, H. T., Gershman, S. J., & Cikara, M. (2018). Discovering social groups via latent structure learning. *Journal of Experimental Psychology: General*, *147*(12), 1881–1891.

Lucas, C. G., Griffiths, T. L., Williams, J. J., & Kalish, M. L. (2015). A rational model of function learning. *Psychonomic Bulletin & Review*, *22*, 1193–1215.

Mcdaniel, M. A., & Busemeyer, J. R. (2005). The conceptual basis of function learning and extrapolation: Comparison of rule-based and associative-based models. *Psychonomic Bulletin & Review*, *12*(1), 24–42.

Rasmussen, C., & Williams, C. (2006). *Gaussian Processes for Machine Learning*. MIT Press.

Rothe, A., Deverett, B., Mayrhofer, R., & Kemp, C. (2018). Successful structure learning from observational data. *Cognition*, *179*, 266–297.

Schulz, E., Franklin, N. T., & Gershman, S. J. (2018). Finding structure in multi-armed bandits. *bioRxiv*, 432534.

Schulz, E., Speekenbrink, M., & Krause, A. (2018). A tutorial on gaussian process regression: Modelling, exploring, and exploiting functions. *Journal of Mathematical Psychology*, *85*, 1–16.

Schulz, E., Tenenbaum, J. B., Duvenaud, D., Speekenbrink, M., & Gershman, S. J. (2017). Compositional inductive biases in function learning. *Cognitive Psychology*, *99*, 44–79.

Schulz, E., Tenenbaum, J. B., Reshef, D. N., Speekenbrink, M., & Gershman, S. (2015). Assessing the perceived predictability of functions. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society* (p. 2116-2121). Cognitive Science Society.

Shafto, P., Kemp, C., Baraff, E., Coley, J., & Tenenbaum, J. (2005). Context-sensitive induction. In *Proceedings of the 27th Annual*

Conference of the Cognitive Science Society (pp. 2003–2008). Austin, TX: Cognitive Science Society.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323.

Sims, C. R. (2018). Efficient coding explains the universal law of generalization in human perception. *Science*, *360*(6389), 652–656.

Smola, A. J., & Kondor, R. (2003). Kernels and regularization on graphs. In *Learning theory and kernel machines* (pp. 144–158). Springer.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*(6022), 1279–1285.

Tomov, M., Yagati, S., Kumar, A., Yang, W., & Gershman, S. (2018). Discovery of hierarchical representations for efficient planning. *bioRxiv*, 499418.

Wu, C. M., Schulz, E., Garvert, M. M., Meder, B., & Schuck, N. W. (2018). Connecting conceptual and spatial search via a model of generalization. In T. T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.), *Proceedings of the 40th annual conference of the cognitive science society* (pp. 1183–1188). Austin, TX: Cognitive Science Society.

Wu, C. M., Schulz, E., Speekenbrink, M., Nelson, J. D., & Meder, B. (2018). Generalization guides human exploration in vast decision spaces. *Nature Human Behaviour*, *2*, 915-924. doi: 10.1038/s41562-018-0467-4

Zhu, X., Lafferty, J. D., & Ghahramani, Z. (2003). *Semi-supervised learning: From gaussian fields to gaussian processes* (Tech. Rep. No. CMU-CS-03-175). Carnegie Mellon University.

Detecting presupposition failure and accommodation with EEG

Alice Xia (alicexia@cmail.carleton.ca)
Institute of Cognitive Science, Carleton University, 1125 Colonel By Dr
Ottawa, ON K1S 5B6 Canada

Roxana M. Barbu (roxanamariabarbu@cmail.carleton.ca)
Institute of Cognitive Science, Carleton University, 1125 Colonel By Dr
Ottawa, ON K1S 5B6 Canada

Kathleen Van Benthem (kathy.vanbenthem@carleton.ca)
Institute of Cognitive Science, Carleton University, 1125 Colonel By Dr
Ottawa, ON K1S 5B6 Canada

Daniel A. Di Giovanni (daniel.digiovanni@mail.mcgill.ca)
Montréal Neurological Institute and Hospital, 3801 Rue Université
Montréal, QC H3A 2B4 Canada

Ida Toivonen (ida.toivonen@carleton.ca)
Institute of Cognitive Science, Carleton University, 1125 Colonel By Dr
Ottawa, ON K1S 5B6 Canada

Raj Singh (raj.singh@carleton.ca)
Institute of Cognitive Science, Carleton University, 1125 Colonel By Dr
Ottawa, ON K1S 5B6 Canada

Abstract

Sentence comprehension in part involves introducing, storing, and retrieving information about individuals. Natural languages provide various means for performing this computational work. One popular idea is that indefinite noun phrases provide instructions for updating the discourse model by adding a new discourse referent, while definite noun phrases presuppose the existence of a discourse referent available in memory, as well as instructions for retrieving it. When no antecedent is available, the definite's presupposition fails to be satisfied, resulting in the so-called 'presupposition failure' and pragmatic infelicity. However, under certain conditions, definite noun phrases *can* felicitously be used even when no antecedent is available in memory. In such cases, a conversational repair strategy called 'presupposition accommodation' can rescue the discourse by adding the required referent. It is natural to expect greater processing costs for adding a discourse referent with a definite than with an indefinite: although both involve the process of adding a referent, definites go through a stage of presupposition failure and a subsequent decision to accommodate. The experimental challenge has been to apply a method sensitive enough to detect expected costs in discourse, even when the participant is unaware of the presupposition failure and repairs it rapidly. The present study addresses this challenge by using EEG to capture temporally fine-grained processing differences between definite and indefinite noun phrases when both introduce new discourse referents in plausible and implausible contexts. Our main finding is that definite noun phrases elicit the Left Anterior Negativity (LAN) effect, compared to indefinite noun phrases, both in implausible contexts where there is a sense of oddness and in perfectly coherent contexts. We take this as evidence of a specific cognitive stage at which presupposition failure is detected and when an accommodation decision occurs. This also supports the idea that, when encountering a definite, the LAN is tightly linked to working memory processes involving the search for discourse elements that are presupposed to exist in memory. When none are found, definites are subsequently accommodated and bridged to other entities in the discourse.

Keywords: discourse; presuppositions; context; accommodation; EEG

Introduction

Presuppositions in natural language are commonly viewed as pieces of information that impose constraints on the contexts in which they are triggered. Just as pronouns like "she" require that the context furnish a (uniquely) salient female, presuppositions require that the context entail them. For example, consider a command like (1):

(1) # After you read this paper, go call the waiter.

The sentence in (1) is strange when uttered out of the blue. It is strange for the same reason that *go call her* is strange when there is no salient female in the context. The sentence has been uttered in a context that is missing something that the sentence needs - in this case, a uniquely salient waiter. We refer to such cases as 'presupposition failure': the failure is technical (i.e., the context does not entail the presupposition), and this technical failure leads to a discourse failure. Note that there is nothing inherently odd about the sentence in (1); it is odd when the context in which it is uttered fails to provide a waiter as antecedent for the definite phrase *the waiter*. If we introduce a waiter into the prior context, meaning that there

is no longer any technical presupposition failure, the oddness disappears:

(2) A waiter and a cook came by and left a flyer. After you read this paper, go call the waiter.

Note also that the oddness disappears when we change "the" to "a"; since the latter has no presuppositions, there is no threat of presupposition failure, and hence none of the oddness that is experienced in (1).

(3) After you read this paper, go call a waiter.

You might not know why the speaker is telling you to call a waiter, but nothing has gone wrong as far as language itself is concerned.

This connection between presupposition and the prior context is the centrepiece of the so-called "satisfaction theory" of presupposition (e.g., Karttunen, 1974; Stalnaker, 1974; Heim, 1983). Its basic assumption is that a sentence S with presupposition p, S_p, may be used in context c only if c 'satisfies' p, i.e., only if $c \subseteq p$. The satisfaction theory adds additional auxiliary assumptions to deduce predictions about presupposition projection, that is, about the presuppositions of complex sentences. For example, it predicts that $\neg S_p$ presupposes p but that *if T then* S_p presupposes $A \rightarrow p$ (e.g., Heim, 1983). We will limit our attention to atomic sentences here.

A *prima facie* challenge for the satisfaction theory comes from the observation that it is often possible to felicitously use S_p even when c does *not* satisfy p. In other words, there appear to be instances of technical presupposition failure without any sense of a higher discourse failure. For example, consider the following text (modified from Singh, Fedorenko, Mahowald, & Gibson, 2016):

(4) I went to a restaurant last night. The waiter yelled at me.

In (4), the context in which *the waiter* is uttered does not entail that there was a waiter. It is plausible, of course, that there should be waiters in the restaurant, but this information is not strictly entailed by the context. Nevertheless, there is no hint of oddness here.[1]

The satisfaction theory explains the contrast between (4) and (1) by appealing to what is called 'presupposition accommodation' (Lewis, 1979). When addressees hear or read a definite description like *the waiter* in a context that does not furnish an antecedent, they face a choice: they can either accommodate the required presupposition; that is, they can 'quietly and without fuss' (Von Fintel, 2008) adjust the context by adding the missing presupposition, or they can let the discourse come to a crashing halt. If the context is one that makes it reasonable to accommodate, say if the presupposition is unsurprising or uncontroversial, then cooperative speakers will recognize that they should keep the discourse running and will therefore simply accommodate.[2] Viewed in this light, accommodation is a repair mechanism that can fix a context so that technical presupposition failure – the failure to initially find the required antecedent – does not become pragmatic presupposition failure. By 'pragmatic presupposition failure' we mean that the context does not get amended and the discourse is interrupted because the definite noun phrase is unable to do its job. It is considered bad conversational practice to rely on accommodation when the presupposition is somehow noteworthy. If the addressee faces the choice of having to either let the discourse crash because of pragmatic presupposition failure, or accommodate a presupposition that is surprising, controversial, or otherwise hard to incorporate into the context, then the addressee would rightly feel that the speaker is asking too much of them.

The appeal to accommodation has been controversial (e.g., Gazdar, 1979; Van der Sandt, 1992; Gauker, 1998; Abbott, 2006). The satisfaction theory predicts that the addressee has passed through a stage of processing at which technical presupposition failure was detected but was then overcome by the accommodation repair. However, there is no trace of this failure detection in our conscious awareness, and it would be desirable to find a way to measure whether accommodation is real and whether it is indeed triggered by a stage of technical presupposition failure.

Previous psycholinguistic studies have found that S_p is easier to process in contexts that satisfy p than in contexts that do not (e.g., Haviland & Clark, 1974; Crain & Steedman, 1985; Burkhardt, 2006; Schwarz, 2007). This might be thought to lend support to the satisfaction theory. However, this processing difference may not be about presupposition accommodation itself; instead, it could have arisen from the fact that in contexts in which p is not satisfied, there is an extra step of adding p to the context. This additional step may have been responsible for the extra costs whether or not there is any purported technical presupposition failure from which the addressee may choose to recover using accommodation (see Singh et al., 2016 for discussion).

To control for this, we would need a minimal pair that would also involve adding p to the context, but through assertion rather than presupposition accommodation. Indefinite articles provide the required contrast:

(5) I went to a restaurant last night. A waiter yelled at me.

[1] A reviewer raises the question about the relative appropriateness of "I went to a restaurant last night. The waiter yelled at me" and "I went to a wedding. The bride talked to me." Weddings typically have one and only one bride, while there may be no waiters or many waiters at a restaurant, and the reviewer suggests that these considerations might lead to differences in appropriateness judgments. We are not aware of work on this, and we hope to return to it in future work. We intentionally designed contexts that would (i) allow multiple referents, such as multiple waiters – this is that indefinites could also be used in these contexts), with one being uniquely salient (e.g., the waiter who serves you).

[2] Heim (1982) argued that with definites, there must be a prior discourse referent that the definite can 'bridge' to (in the sense of Clark, 1975). For example, in (4) the introduced waiter can 'bridge' to the restaurant mentioned in the first sentence, such that *the waiter* is roughly understood as 'the waiter at the restaurant', and typically with further identifying features (e.g., the waiter who served you at the restaurant – see Note 1).

In both (4) and (5), the existence of a waiter is added to the discourse context (e.g., Heim, 1982). For example, suppose with Heim (1982) that discourse referents can be thought of as file cards that can be introduced, referred to, or taken out of the 'file' that collects the discourse information as it accumulates. An indefinite noun phrase such as *a waiter* simply adds a new file card with 'is a waiter' on it. A definite like *the waiter* scans the file to find a file card with 'is a waiter' on it. If one exists, it refers to it; if it doesn't, then either communication fails, or the missing file card is accommodated. This is what appears to happen in (4). Thus, the processing of both (5) and (4) involves adding a file card corresponding to a waiter, but only in (4) do you also go through a stage of recognizing that something is wrong with the context (there is no antecedent file card). The addition of a (file card corresponding to a) waiter is then an accommodation response to this recognition; we will sometimes use 'referents' and 'antecedents' when we mean file cards, as they are probably more familiar, but it is worth noting that file cards are the underlying technical object that is being manipulated (see Heim, 1982). What we would like, then, is to test whether there is indeed a stage at which the processing mechanism seeks but fails to find an antecedent for *the waiter*, and then repairs for this by accommodating one. The satisfaction theory predicts that there should be such stages; presumably the performance system executes these computations demanded by the competence system, and if so, we might expect to find reflexes of them during language processing.

Singh et al. (2016) performed an online incremental stops-making-sense (SMS) task to examine participants' appropriateness judgments about indefinite and definite noun phrases in plausible contexts like (4) and (5) as well as in implausible contexts like the following:

(6) I went to a jail last night. {A/the} waiter yelled at me.

They found a main effect of plausibility, such that implausible conditions had more and earlier SMS judgments. This was unsurprising, given that implausible information is generally harder to process than plausible information (e.g., Trueswell, Tanenhaus, & Garnsey, 1994; Gibson & Perlmutter, 1998). More interestingly, they found an interaction, such that implausible definites had earlier and more SMS judgments than implausible indefinites. This provides support for the claim that accommodation is subject to stricter requirements than assertion. In particular, it is inappropriate to force your addressee to accommodate implausible information as a presupposition; such information is better expressed as an assertion so that your addressee is at least given the opportunity to challenge it (e.g., Soames, 1989; Heim, 1992; Beaver, 2001; Von Fintel, 2008).

However, together with appropriate auxiliary assumptions about how the competence theory is realized in performance (see above), it is plausible that the satisfaction theory would expect accommodation difficulty relative to indefinite controls not only in implausible contexts but also in plausible contexts (though cf. Stalnaker, 2002). There is technical presupposition failure in both plausible and implausible contexts. Pragmatic presupposition failure of course is more easily averted in plausible contexts than implausible ones. Thus, the enhanced difficulty of implausible definites makes sense. However, the predicted stage of technical failure in plausible contexts did not reveal itself. Perhaps the method was inappropriate for detecting such a stage, if there is one. We have seen that accommodation is not sensed as odd or costly when the presupposition is sufficiently supported in the context. Thus, it is perhaps not surprising that participants' SMS judgments did not differentiate between plausible definites and plausible indefinites. It remains an open question, then, whether an empirical cognitive account can be found for when the absence of an expected antecedent is noticed and when we decide to accommodate in response.

We explored this question by means of an electroencephalography (EEG) study using materials from Singh et al. (2016). We give a comprehensive account of our materials and methods momentarily, but briefly our goal was to compare definites and indefinites in contexts in which both would have the effect of introducing a new discourse referent into the context. The relevant difference between the two is that definites presuppose the existence of an object in memory and aim to retrieve it while indefinites introduce a new object into memory. Ideally, we want these objects to be the same, or as close to that as possible. That way, any detected difference between the two could be plausibly attributed to the assumption that definites introduce the desired object only when the search for it fails. We wanted to see if we could find an EEG signature of this hypothesized failure and repair. Previous EEG studies comparing definites and indefinites did not isolate this difference between definites and indefinites. Experiment 2 of Anderson and Holcomb (2005) had a definite and indefinite condition but the definite in these cases *had* an antecedent and the indefinite (as it does) introduced a new discourse referent. Schumacher (2009), building on Burkhardt (2006) (which investigated definites alone), included a definite given condition and an indefinite given condition (in which an indefinite NP in the second sentence has a matching indefinite NP in the first sentence); but the latter texts are odd (for reasons we discuss shortly), and hence the definite and indefinite are not properly matched, and in any event, this 'given' condition breaks our desired symmetry under which definites and indefinites both introduce a new discourse referent in all conditions.

Methods

Participants

Thirty-four participants were recruited from Carleton University. As compensation for participating in the experiment, students received 3% class credit towards a first-year cognitive science course. All students were English speakers between the ages of 18 and 24.

Materials

We used shortened versions of all 128 sentence pairs from Singh et al. (2016) as our experimental stimuli. The sentence pairs were divided evenly into four blocks using a Latin Square design as illustrated in Table (1):

Table 1: Sample stimuli

Indefinite Plausible:
Philip went to *a pool* on Tuesday evening.
A swim instructor insulted him there.
Definite Plausible:
Philip went to *a pool* on Tuesday evening.
The swim instructor insulted him there.
Indefinite Implausible:
Philip went to *a laboratory* on Tuesday evening.
A swim instructor insulted him there.
Definite Implausible:
Philip went to *a laboratory* on Tuesday evening.
The swim instructor insulted him there.

Procedure

Participants sat in a Faraday cage in front of a computer monitor and were instructed to read all sentence stimuli for comprehension. Using PsychoPy, all stimuli were presented visually in the center of the monitor in white letters against a grey background. A practice session consisting of four trials was completed before beginning each session. The first sentence in the sentence pair appeared in full for 3000 ms, followed by 100 ms of a blank screen. Our critical noun phrase (e.g., "the lion") in the second sentence of the pair then appeared on screen for 600 ms, followed by another 100 ms of a blank screen. The remaining non-critical segments of the second sentence, which had an average length of three words, appeared for 400 ms. All participants saw all items in all conditions, counterbalancing block orders.

EEG Recording

A 128-channel HydroCel Geodesic Net was used to record continuous EEG signals against Cz as reference, at a sampling rate of 250 Hz, with Net Station 4.3.1. Electrode impedance was kept below 5 kOhms.

Data Analysis

Data from two participants were excluded due to excessive noise during EEG recording. Four channels (E68, E73, E88, E94) were removed prior to preprocessing as is common for high-density electrode nets (to allow the plug-in of other external biometric devices). EEG recordings were re-referenced offline to the average and digitally filtered with a low-pass of 0.5 Hz and a high-pass of 30 Hz. Filtered data were then epoched from 500 ms before to 1000 ms after the critical noun phrase.

Subject data were preprocessed using a combination of EEGLAB 14.1.2 (Delorme & Makeig, 2004) and custom-written MATLAB scripts. Independent component analysis (ICA) in EEGLAB was used to first remove eye-blinks and other physiological noise. The CleanLine toolbox (Mullen, 2012) was used to reduce drift. Channels that were three standard deviations away from the mean, based on a power spectrum threshold, were removed. Lastly, an automatic component rejection was performed using the MARA toolbox (Winkler, Haufe, & Tangermann, 2011).

Following previous literature, two time windows were selected for analysis: 300-500 ms and 500-700 ms after onset of the critical noun phrase. This allowed us to examine the N400/P600 complex (Burkhardt, 2006), as well as the LAN effect (Kutas & Federmeier, 2007). Using the EEGLAB Darbeliai extension, event-related potentials were computed for the 1000 ms after stimulus onset relative to a 100 ms pre-stimulus baseline for each participant, for each condition, from electrodes clustered in each of the following four regions: left anterior (F3/F7/FC3/FT7), right anterior (F4/F8/FC4/FT8), left posterior (P3/T5/CP5/T5), and right posterior (P4/T6/CP6/T6).

For statistical analyses, mean amplitude data were submitted to linear mixed-effects models using the lme4 package (Bates, Mächler, Bolker, & Walker, 2014) in R (R Core Team, 2013). Significance testing was done using the lmerTest package (Kuznetsova, Brockhoff, & Christensen, 2017). In both time windows of interest, our models evaluated mean amplitude as a function of a three-way interaction among Plausibility (plausible, implausible), Definiteness (definite, indefinite), and Electrode Region (left anterior, right anterior, left posterior, right posterior). Participant was included as a random factor. We performed planned comparisons between our four conditions (Definite Plausible, Indefinite Plausible, Definite Implausible, Indefinite Implausible) if significant interactions were found between region and definiteness or plausibility. Pairwise contrasts were investigated using the emmeans package in R (Lenth, 2018) and p-values were adjusted using the Bonferroni correction.

Results

300-500 ms time window: We found a significant interaction between Region and Definiteness ($F(3, 1946) = 5.64$, $p < .001$). In particular, the Definite - Indefinite condition contrast in the group of left anterior electrodes was significant (beta = -.37, t = -5.24, $p < .001$) (Figure 1). This general determiner effect was further corroborated by a scalp map of the same left anterior electrodes in the 300-500 ms time window, averaged across all participants for each condition (Figure 2). We further found a significant difference in the Definite Implausible - Indefinite Implausible

contrast in the same region (beta = -.35, t = -3.59, p = .002), which was reflected in greater negativity elicited by definite noun phrases (e.g., "the lion") compared to indefinite noun phrases (e.g., "a lion") (Figure 3). Similarly, in the plausible context, the Definite Plausible - Indefinite Plausible contrast revealed a significantly more negative deflection for definite noun phrases (beta = -.37, t = -3.81, $p < .001$) (Figure 4).

500-700 ms time window: In the left anterior, we again found a significant interaction between Region and Definiteness ($F(1, 1946) = 22.17$, $p < .001$). Contrast analyses were significant for Definite - Indefinite (beta = -.22, t = -4.05, $p <$ .001) (Figure 1), as well as for Definite Implausible - Indefinite Implausible (beta = -.30, t = -3.86, $p < .001$) (Figure 4).

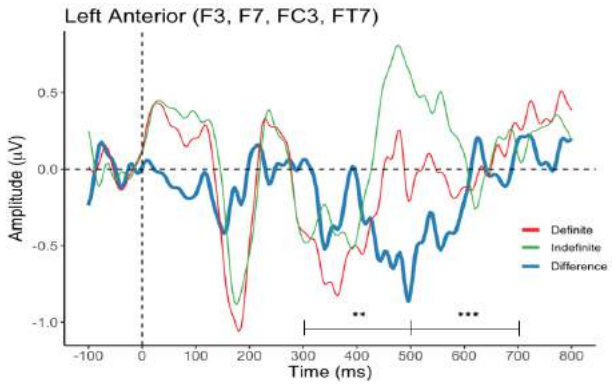


Figure 1: Significant negative-going ERP elicited by the Definite condition relative to Indefinite condition in the 300-500 ms time window in left anterior electrodes, reminiscent of the Late Anterior Negativity (LAN). Significant divergence between the two conditions continues into the 500-700 ms time window.

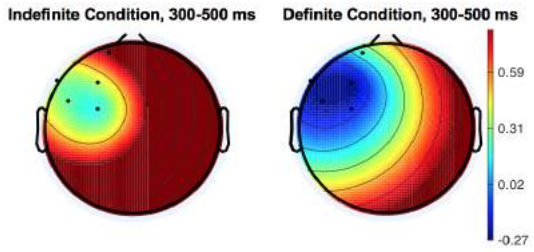


Figure 2: Scalp map of left anterior electrodes for Indefinite (left) and Definite (right) conditions in the 300-500 ms time window.

Discussion

Our study used EEG to explore the consequences of processing definite and indefinite noun phrases in plausible and implausible sentence contexts. Our goal was to test both types of determiners together, to isolate the crucial stages of presupposition failure and accommodation of a new discourse

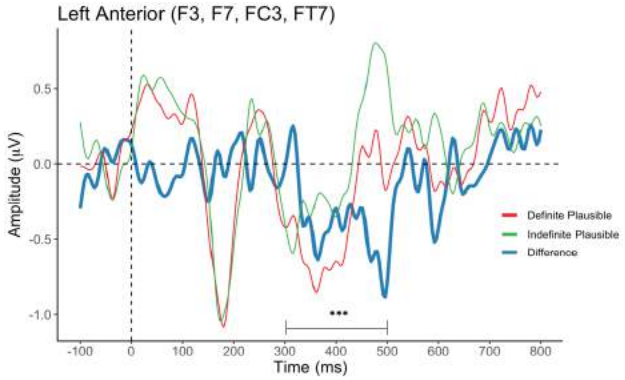


Figure 3: Greater negativity elicited by the Definite Plausible condition relative to the Indefinite Plausible condition in the 300-500 ms time window.

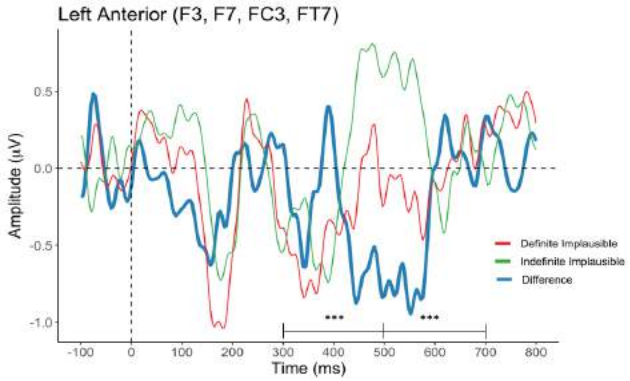


Figure 4: Greater negativity elicited by the Definite Implausible condition relative to Indefinite Implausibl in the 300-500 ms time window. Significant divergence between the two conditions continues into the 500-700 ms time window.

referent (see our earlier discussion of Burkhardt, 2006 and Schumacher, 2009; see also Hirotani & Schumacher, 2011). While Schumacher (2009) also investigated definite and indefinite phrases together, the pragmatic appropriateness of the phrase pairs in that task was not stable across contexts. For example, consider a given definite text like *Peter has recently visited a speaker in Munich. He said that the speaker had been very nice*. The text is coherent and there is no sense of oddness. This cannot be said of the indefinite given counterpart: *Peter has recently visited a speaker in Munich. He said that a speaker had been very nice.* This text is decidedly odd, confirmed by offline data reported in Schumacher (2009). Presumably, the oddness is due to so-called *Maximize Presupposition!* effects (e.g., Heim, 1991; Singh, 2011), which essentially demand that the speaker use a presuppositional alternative (like a definite) when its presupposition is satisfied instead of a non-presuppositional minimal variant (like an indefinite). *Maximize Presupposition!* has been proposed to explain the oddness of sentences like *A sun is*

shining; the definite variant is preferred because of *Maximize Presupposition!* (see also Note 1). Similarly, in the case under current consideration, the indefinite *a speaker* would be ruled out in favour of *the speaker* (if the same speaker is intended), or in favour of *another speaker* (if a different speaker is intended).

In our study, the definite/indefinite pairs are both either appropriate or both inappropriate in their given contexts (cf. stimuli norming results in Singh et al., 2016). There is no influence of *Maximize Presupposition!* because there is no 'given' context. All conditions required the addition of a new discourse referent, and hence were expected to not differ with respect to the P600 (given the findings in Burkhardt, 2006; Schumacher, 2009). Somewhat to our surprise, our results showed a greater positive deflection for indefinites than definites in the implausible context during the 500-700ms window. This might suggest a P600, but we are not confident that it is since the P600 is typically found over parietal lobes (e.g., Osterhout & Holcomb, 1992). Whatever this effect's proper classification, the difference does not replicate the finding in Schumacher (2009) that definites and indefinites both generate a late positivity (a P600 in her studies) that indexes the addition of a discourse referent. The difference here might be teaching us that by this stage the accommodation for the definite has already occurred, or that the introduction of referents via assertion is different than via accommodation, and the late positivity we found for indefinites indexes only assertive updates.

Here we tentatively suggest that by that late stage the accommodation for the definite has already taken place. According to the satisfaction theory, assertions are updates to the context, and the presuppositions in a context get updated by the assertion. Thus, there is an implied temporality: presuppositional matters are resolved prior to assertive updates (hence the *pre-*). Thus, it is conceivable that the accommodation step occurs early, right after the detection of the technical presupposition failure. Perhaps this bundle of computations is what our left-lateralized frontal negativity for definite noun phrases in the 300-500ms window was indexing. This Left Anterior Negativity (LAN) has been found in previous studies of (in-)definiteness (e.g., Anderson & Holcomb, 2005; Schumacher, 2009), but for reasons discussed earlier it is hard to interpret such findings because the contrasts between definite and indefinite conditions were not quite minimal. More generally, the LAN has been linked to processes of working memory resources that involve 'reactivating' previous entities or forming dependencies between new and old entities (see e.g., King & Kutas, 1995; Kirsten et al., 2014, among others). We tentatively propose here that the detection of presupposition failure and accommodation are among the computations the LAN indexes. Note that definites involve the search for entities in memory ('reactivation'), and that accommodation when no antecedent is found typically involves 'bridging' the new entity to a previous entity (e.g., linking 'the waiter' to the previously mentioned restaurant – see also Note 2).

The design of our study was based on a previous stops-making-sense task that investigated temporal decisions during the silent reading of definite and indefinite phrases in contexts that varied in plausibility (Singh et al., 2016). Based on the results of that study, we initially expected sentences with implausible contexts to result in a semantic violation that would be captured by the N400, relative to sentences with plausible contexts. Our results did not support this expectation. There may be several reasons for this. First, our stimuli did not include traditional semantic violation phrases that are used elsewhere in the N400 literature (e.g., *He spread the warm bread with socks*, Kutas & Hillyard, 1984). The process of reading an otherwise well-formed phrase in an implausible context (e.g., "the lion" in the context of a restaurant) may not map directly onto the process of reading semantic violation phrases, which typically are incoherent. Our implausible texts are not incoherent like the traditionally studied ones; they are merely implausible, and this distinction might be relevant to the N400 component. Second, unlike our instructions, the stops-making-sense task used in Singh et al. (2016) explicitly required participants to make judgments about nonsense. This may have led participants to pay greater attention to coherence and sensibility than our instructions. Third, it is possible that there is a lag between the time at which the brain first detects implausibility/incoherence and the time at which our minds become consciously aware of this, and the N400 may be sensitive to the first and not necessarily the second.

Our results thus sharpen Schumacher's finding that the LAN appears to be associated with failure to find an appropriate antecedent when triggered by a uniqueness presupposition, i.e., *the*. By removing 'given' conditions, and the need to compare context updates, we have isolated the cognitive cost associated with technical presupposition failure and accommodation: the brain registers it as a LAN effect. If this is correct, we would expect the LAN to show up in other environments that require accommodation but which are not odd. For example, *The psychology department is facing a crisis. Both of their neuroscientists left* should elicit a LAN relative to *The psychology department is facing a crisis. All of their neuroscientists left*.

References

Abbott, B. (2006). *Unaccommodating presuppositions: A neogricean view.* (Manuscript, Michigan State University)

Anderson, J., & Holcomb, P. (2005). An electrophysiological investigation of the effects of coreference on word repetition and synonymy. *Brain and Language*, *94*(2), 200-216.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

Beaver, D. (2001). *Presupposition and assertion in dynamic semantics*. Stanford, CA: CSLI.

Burkhardt, P. (2006). Inferential bridging relations reveal distinct neural mechanisms: Evidence from event-related

brain potentials. *Brain and Language*, *98*(2), 159–168.

Clark, H. H. (1975). Bridging. In R. Schank & B. Nash-Webber (Eds.), *Theoretical issues in natural language processing* (p. 169-174). New York, NY: Association for Computing Machinery.

Crain, S., & Steedman, M. (1985). On not being led up the garden path: the use of context by the psychological syntax processor. In D. R. Dowty, L. Karttunen, & A. Zwicky (Eds.), *Natural language parsing: Psychological, computational, and theoretical perspectives* (p. 320-358). Cambridge: Cambridge University Press.

Delorme, A., & Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*(1), 9–21.

Gauker, C. (1998). What is a context of utterance? *Philosophical Studies*, *91*, 149-172.

Gazdar, G. (1979). *Pragmatics*. New York, NY: Academic Press.

Gibson, E., & Perlmutter, N. (1998). Constraints on sentence comprehension. *Trends in Cognitive Science*, *2*, 262-268.

Haviland, S., & Clark, H. H. (1974). What's new? acquiring new information as a process in comprehension. *Journal of Verbal Learning and Verbal Behavior*, *13*, 512-521.

Heim, I. (1982). *On the semantics of definite and indefinite noun phrases.* (Doctoral Dissertation, University of Massachusetts at Amherst)

Heim, I. (1983). On the projection problem for presuppositions. In *WCCFL 2* (p. 114-125).

Heim, I. (1991). Artikel und definitheit. In *Semantics: An international handbook of contemporary research.* Berlin: de Gruyter.

Heim, I. (1992). Presupposition projection and the semantics of attitude verbs. *Journal of Semantics*, *9*, 183-221.

Hirotani, M., & Schumacher, P. B. (2011). Context and topic marking affect distinct processes during discourse comprehension in japanese. *Journal of Neurolinguistics*, *24*(3), 276–292.

Karttunen, L. (1974). Presupposition and linguistic context. *Theoretical Linguistics*, *1*, 181-193.

King, J. W., & Kutas, M. (1995). Who did what and when? using word-and clause-level erps to monitor working memory usage in reading. *Journal of cognitive neuroscience*, *7*(3), 376-395.

Kirsten, M., Tiemann, S., Seibold, V. C., Hertrich, I., Beck, S., & Rolke, B. (2014). When the polar bear encounters many polar bears: event-related potential context effects evoked by uniqueness failure. *Language, Cognition and Neuroscience*, *29*(9), 1147-1162.

Kutas, M., & Federmeier, K. (2007). Event-related brain potential (erp) studies of sentence processing. In G. Gaskell (Ed.), *Oxford handbook of psycholinguistics.* Oxford: Oxford University Press.

Kutas, M., & Hillyard, S. A. (1984). Event-related brain potentials (erps) elicited by novel stimuli during sentence processinga. *Annals of the New York Academy of Sciences*, *425*(1), 236–241.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmertest package: tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13).

Lenth, R. (2018). Emmeans: Estimated marginal means, aka least-squares means. *R Package Version*, *1*(2).

Lewis, D. (1979). Scorekeeping in a language game. *Journal of Philosophical Logic*, *1*, 339-359.

Mullen, T. (2012). *NITRC: Cleanline: Tool/resource info.*

Osterhout, L., & Holcomb, P. J. (1992). Event-related brain potentials elicited by syntactic anomaly. *Journal of memory and language*, *31*(6), 785-806.

R Core Team. (2013). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria.

Schumacher, P. B. (2009). Definiteness marking shows late effects during discourse processing: evidence from erps. In *Discourse anaphora and anaphor resolution colloquium* (pp. 91–106).

Schwarz, F. (2007). Processing presupposed content. *Journal of Semantics*, *24*, 373-416.

Singh, R. (2011). *Maximize Presupposition!* and local contexts. *Natural Language Semantics*, *19*, 149-168.

Singh, R., Fedorenko, E., Mahowald, K., & Gibson, E. (2016). Accommodating presuppositions is inappropriate in implausible contexts. *Cognitive Science*, *40*, 607-634. (DOI: 10.1111/cogs.12260)

Soames, S. (1989). Presupposition. In D. Gabbay & F. Guenther (Eds.), *Handbook of philosophical logic, vol. iv.* Dordrecht: Reidel.

Stalnaker, R. (1974). Pragmatic presuppositions. In M. Munitz & P. Unger (Eds.), *Semantics and philosophy.* New York, NY: NYU Press.

Stalnaker, R. (2002). Common ground. *Linguistics and Philosophy*, *25*, 701-721.

Trueswell, J. C., Tanenhaus, M. K., & Garnsey, S. M. (1994). Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *Journal of Memory and Language*, *33*, 285-318.

Van der Sandt, R. (1992). Presupposition projection as anaphora resolution. *Journal of Semantics*, *9*, 333-377.

Von Fintel, K. (2008). What is presupposition accommodation, again? *Philosophical Perspectives*, *22*, 137-170.

Winkler, I., Haufe, S., & Tangermann, M. (2011). Automatic classification of artifactual ICA-components for artifact removal in EEG signals. *Behavioral and Brain Functions*, *7*(1), 30.

How should we incentivize learning? An optimal feedback mechanism for educational games and online courses

Lin Xu
Rationality Enhancement Group, MPI for Intelligent Systems, Tübingen, Germany
Maria Wirzberger
Rationality Enhancement Group, MPI for Intelligent Systems, Tübingen, Germany
Falk Lieder
Rationality Enhancement Group, MPI for Intelligent Systems, Tübingen, Germany
Bernstein Center for Computational Neuroscience, Tübingen, Germany

Abstract

There are plenty of opportunities for life-long learning but people rarely seize them. Game elements are an increasingly popular tool to keep students engaged in learning. But gamification only works when it is done properly. Here, we introduce the first principled approach to gamifying learning environments. Our feedback mechanism rewards students' efforts and study choices according to how beneficial they are in the long run. The rewards are conveyed by game elements that we call "optimal brain points". In our experiment, these optimal brain points significantly increased the proportion of participants who attempted to learn a difficult skill, persisted through failure, and succeeded to master it. Our method provides a principled approach to designing incentive structures and feedback mechanisms for both educational games and online courses. We are optimistic that this can help people overcome the motivational obstacles to self-directed life-long learning.

Keywords: gamification; artificial intelligence in education; persistence; educational games; incentive structures

Introduction

As the technological development accelerates, self-directed life-long learning is becoming critically important. Massive Open Online Courses (MOOCs) and other digital resources provide unprecedented opportunities for life-long learning. However, only about 15% of the students who enroll in a MOOC actually finish it (Jordan, 2019). One of the reasons might be that learning something new often requires confronting one's own incompetence and persisting through several failed attempts to understand a new concept or do something new. Many people tend to irrationally avoid such hardships (Urdan & Midgley, 2001; Baker et al., 2008) even though they are often necessary to master new skills (Ericsson, Krampe, & Tesch-Römer, 1993). People who have become experts in using an outdated tool by doing the same work in the same way for many years may be especially resistant to learning how to use a new tool because in the short-run it is much more comfortable for them to exploit their outdated expertise than to become a novice again.

When students are given choices in online courses or educational software they sometimes procrastinate on learning something new by repeatedly practicing skills they already know (Baker, Corbett, & Koedinger, 2004; Mostow et al., 2002).

To help student's overcome such motivational obstacles, educational software increasingly relies on game elements, such as points, levels, and badges, to encourage continued engagement with the learning materials (Kapp, 2012; Dicheva, Dichev, Agre, & Angelova, 2015; Huang & Soman, 2013). The trend of gamification has outpaced the development of an adequate theoretical foundation, and it has been noted that gamification is often ineffective and sometimes even harmful (Toda, Valle, & Isotani, 2018). This raises the question how the incentive structures of digital learning environments such as educational games and online courses should be designed to optimally incentivize good study choices and effective learning strategies.

The points students receive in educational games usually convey performance feedback. But making performance feedback more gameful does not address the fundamental problem that – in the short run – performance feedback might discourage trying to learn something new. Rather, by making student's failures more salient to them, gamified performance feedback can have a negative effect on their study choices – thereby making things worse rather than better (Shute, 2008). O'Rourke, Haimovitz, Ballweber, Dweck, and Popović (2014) argue that to address this problem, gamification should give students "brain points" that reward effort and persistence rather than performance. In support of this view, they found that incentivizing students' effort and learning strategies in an educational game significantly increased their persistence and the total amount of time they spent in the game. However, the hand-crafted incentive system was imperfect and could be exploited by discovering easy ways to earn brain points without doing the hard work of learning a new skill (O'Rourke, Peach, Dweck, & Popovic, 2016). The high prevalence of students "gaming the system" across many intelligent tutoring systems (Baker et al., 2008) underlines that designing good incentives by hand is hard and fallible. This illustrates the deeper issue that we lack a principled theory for designing reward structures in learning systems that incentivize learning properly.

Recent work has begun to establish such principles in the domain of decision-support (Lieder & Griffiths, 2016; Lieder, Chen, Krueger, & Griffiths, 2019). There, the basic idea of this approach is to align the immediate reward of each decision with its long-term value. This addresses the problem that people's decisions are usually overly swayed by the anticipated immediate outcomes (e.g., the unpleasantness of strug-

gling with a difficult math problem vs. the fun of watching a YouTube video) rather than their long-term consequences (e.g., the benefits of a good education). This, so called, *present bias* (O'Donoghue & Rabin, 1999) manifests in a wide range of sub-optimal, short-sighted decisions and problems such as impulsivity and procrastiation that have been explained in terms of hyperbolic discounting and temporal motivation theory (Steel & König, 2006; Steel, 2007).

Considering people's present bias, the real world is far from being an optimal learning environment because the immediate reward for practicing a new skill is usually failure and negative feedback – when it should be something much more positive, namely the value of learning. Conversely, neglecting skill development in favor of exploiting existing skills is usually rewarded because it leads to higher immediate productivity. This suggests that the present bias could be one of the major reasons why students often quit studying too soon or procrastinate on learning a difficult skill – especially when this requires persisting through a series of failed attempts. This suggests that the optimal gamification approach developed by Lieder et al. (2019) might also be applicable to support students' study choices in MOOCs and educational games.

Here, we leverage the framework of optimal gamification (Lieder & Griffiths, 2016; Lieder et al., 2019) to develop a formal mathematical theory of optimal incentives for self-directed learning and an automatic method for computing such incentives from basic assumptions about the skills to be learned and the process of skill acquisition. To achieve this, we develop a mathematical model of the value of practice and apply optimal gamification to it. Our method can be used to automatically compute *optimal* brain points that encourage learning behaviors that are consistent with the growth mindset that the intervention by O'Rourke et al. (2014) was meant to encourage. We postulate that optimal brain points can not only increase the amount of time students invest into learning, as has been demonstrated for hand-designed brain points (O'Rourke et al., 2014), but also their learning outcomes. We test this prediction in a behavioral experiment that simulates a scenario where people have to choose between exploiting their old skill (Skill 1) or learning a new skill (Skill 2) that would allow them to solve a recurring task more efficiently.

We found that participants incentivized with optimal brain points were less likely to give up on trying to learn a new skill, became more likely to master it, and consequently performed better at their tasks. This suggests that our method for computing optimal brain points can help us overcome the pitfalls of incentivizing students study choices manually.

These findings suggest that our principled approach to incentivizing skill acquisition can help people overcome the motivational challenges of self-directed learning and could be used to make educational games and online courses more effective and to avoid the pitfalls of previous attempts to gamify education. Optimal brain points are a principled way to incentivize good study choice and might be able to help students develop a growth mindset (Dweck, 2008).

The plan for this paper is as follows: We first derive the long-term value of practicing a new skill using a simple model of skill acquisition. Next, we translate the value of practice into an optimal gamification method for encouraging skill acquisition. We then evaluate the efficacy of this method in a behavioral experiment mimicking the motivational obstacles to life-long learning. We conclude with the implications of our findings for designing educational games and directions for future work.

Quantifying the value of practice

When should you complete a task using the skills you already have and when should you try to learn a better way to accomplish it? If you would like to invest into learning a new skill, which one should you pick? And if trying to learn this skill is proving difficult, then how long should you keep trying before you give up and do it in the old, familiar way? To help people make these difficult choices, we derive the value of practicing an unfamiliar skill.

The first step of our derivation postulates a simplistic but general and tractable model of skill acquisition through trial and error. If a task has k potential solutions – only one of which is correct – then the probability of discovering the skill in the first attempt is $\frac{1}{k}$. Conversely, the probability that the first attempt will fail is $\frac{k-1}{k}$. After a failure the probability of success increases to $\frac{1}{k-1}$.

Based on this probabilistic model, we can describe skill acquisition as a Markov Decision Process (Sutton & Barto, 1998)

$$M_{\text{skill}} = \{\mathcal{S} \times \mathcal{D}, \mathcal{A}, \gamma, T, r\} \tag{1}$$

where $\mathcal{A}$ includes one action for each skill, $\mathcal{S}$ is the set of all possible skill levels the learner could attain through practice and $\mathcal{D} \subset \mathbb{N}_0$ denotes how much more work is required to complete the current task. The learner's skill level $\mathbf{s_t} \in \mathcal{S}$ reflects how close they are to having mastered each of n different skills at time t and how likely they are to succeed at the task by using each of those skills in their next attempt. We formalize it by the tuple $(k_1, k_2, \cdots, k_n)$ where k_i is the number of potential ways in which the i^{th} skill might work given what the learner knows so far. The transition matrix T encodes that unsuccessfully attempting skill i decreases k_i by 1 and that discovering how it works sets k_i equal to 1. It also encodes how the successful application of each skill would reduce the amount of work required to complete the task and that unsuccessful attempts do not decrease it. The reward function $r((\mathbf{s_t}, d_t), a_t, (\mathbf{s_{t+1}}, d_{t+1}))$ encodes the immediate effort of using or attempting to learn a skill and the value of completing the current task. For simplicity, we assume that the cost of each action is -1 and add the value of completing the current task when d changes to 0. Finally, $1-\gamma \in [0,1]$ is the probability that the current type of task will become obsolete in the next time step.

Abstracting away the details of how specific skills are acquired makes this model very general and broadly applicable. It can therefore be used to incentivize student effort in

any learning context. Our model can either be applied out of the box or tailored to specific learning contexts by measuring how specific learning activities increase the probability that the student will successfully learn a particular skill and plugging the measured probabilities into the model's transition matrix T .

Having modelled the process of skill acquisition as an MDP allows us to leverage standard dynamic programming methods (Sutton & Barto, 1998) to compute the value of practice. For instance, we can apply the value iteration algorithm to compute $V^{\star}((\mathbf{s},d))$ – which is the value of having the skill set s when the current task has difficulty d – and $Q^{\star}((\mathbf{s},d),a)$ which is the value of choosing action a (e.g., trying out a new tool versus reusing an old one). To work out under which conditions it is worthwhile to invest in extending one's skill set, we can then translate these value functions into the value of practice which we define as

$$\mathrm{VOP}\left((\mathbf{s},d),a\right) = Q^{\star}\left((\mathbf{s},d),a\right) - V^{\pi_{\text{stop learning}}}\left((\mathbf{s},d)\right), \quad (2)$$

where $V^{\pi_{\text{stop learning}}}$ is the expected return of the strategy that always exploits existing skills without making any investment into learning new skills.

The simplicity of our model allows us to derive the value of practice analytically for the dilemma of choosing between exploiting a mastered skill and attempting to learn a new skill that would make you more effective. When the value of completing the task is g, the mastered skill achieves it in d time steps, and the to be learned skill could achieve it in 1 time step, then the value of practicing the second skill is

$$\begin{aligned}\mathrm{VOP}\left((k_2,d),a_2\right) = {} & \frac{1}{k_2}\cdot\left[(g-1)+\gamma\cdot V^{\star}(((1,1),d))\right] \\ & + \left(1-\frac{1}{k_2}\right)\cdot\left[\gamma\cdot V^{\star}((1,k_2-1),d)-1\right] \\ & - V^{\pi_{\text{stop learning}}}\left(((1,k_2),d)\right),\end{aligned} \quad (3)$$

and the value of ceasing to learn and exploiting Skill 1 is

$$V^{\pi_{\text{stop learning}}}\left((\mathbf{s},d)\right) = g\cdot\frac{\gamma^{d-1}}{1-\gamma^{d}} - \frac{1}{1-\gamma}, \quad (4)$$

where $\frac{\gamma^{d-1}}{1-\gamma^{d}}$ is the expected number of times one can complete the task using only Skill 1, and $-\frac{1}{1-\gamma}$ is the expected cumulative cost of using the skill. This allows us to characterize under which conditions it is valuable to invest in learning a new skill and under which conditions it is better to exploit the skills one already has.

As shown in Figure 1, we found that the value of practice decreases with the relative effectiveness of the skill one has already mastered ($\frac{d_2}{d_1}=\frac{1}{d}$ if d_1 and d_2 are the number of time steps it takes to complete the task with Skill 1 vs. Skill 2 respectively in this example), but increases with the expected number of times one will have to perform the task in the future (i.e., $\frac{1}{1-\gamma}$). This means that learning a new skill becomes more valuable the more often it might be useful in the future and becomes less worthwhile the more effective the skill is that one has already mastered. By quantifying these effects, the value of practice reveals under which conditions it is worthwhile to learn something new (green-blue) and under which conditions it is better to stick with what one already knows (orange-red). Since the recommendations of our mathematical framework appear to be intuitively correct, we proceed to apply our model of the value of practice to automatically incentivize people's study choices. Furthermore, future work might leverage Equations 2–4 to assist people with decisions about their personal or professional development.

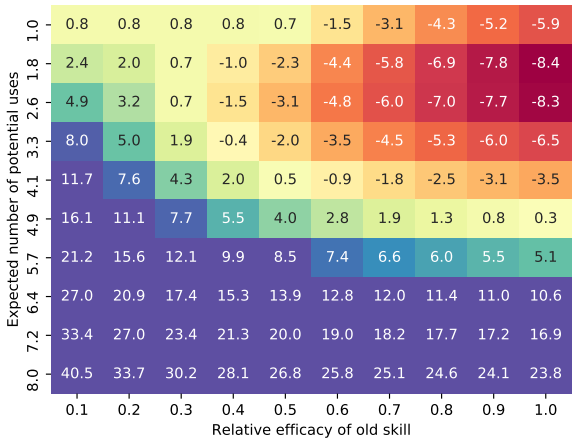


Figure 1: The value of practice. Each square of this heat map shows the difference between the value of attempting to learn a new skill versus exploiting an old skill (colors and numbers) depending on relative efficacy of the old skill (x-axis) and the expected number of occasions on which either skill could be used (y-axis; increasing from top to bottom).

An optimal gamification method for incentivizing skill acquisition

Optimal brain points. Having quantified the value of practice with the skill acquisition MDP defined above, we can now use it to incentivize learning behaviors according to their expected contributions to the learner's competency. Formally, the expected increase in the value of the learner's skill set s achieved by action a is

$$\Delta V(\mathbf{s_t},a) = \gamma\cdot\mathbb{E}[V^{*}(\mathbf{S_{t+1}})|\mathbf{s_t},a] - V^{*}(\mathbf{s_t}),$$

where the random variable $\mathbf{S_{t+1}}$ denotes the learner's skill set after performing action a and $V^{\star}$ is the optimal value function of the skill acquisition MDP defined above. The discount factor γ accounts for the possibility that the practiced skill might become obsolete.

The value of learning by doing is twofold: it increases the value of the learner's skill set (ΔV) and it produces potentially

valuable outcomes ($r(\mathbf{s_t}, a)$). Our optimal brain points capture both sources of value, that is

$$\text{BrainPoints}(\mathbf{s}, a) = \Delta V(\mathbf{s}, a) + r(\mathbf{s}, a). \quad (5)$$

The way in which optimal brain points are constructed is a direct application of the optimal gamification method developed by Lieder et al. (2019). It satisfies the necessary and sufficient conditions of the *shaping theorem* (Ng, Harada, & Russell, 1999) which thereby guarantees that the resulting incentives do not encourage sub-optimal learning strategies. Rather, by using the value of the learner's skill set ($V^\star$) as the basis for constructing the brain points, they are making optimal study choices immediately rewarding. We predict that they should therefore help learners overcome the present bias and invest more in acquiring difficult skills that will benefit them in the future. In the next section, we test this hypothesis with a simple behavioral experiment.

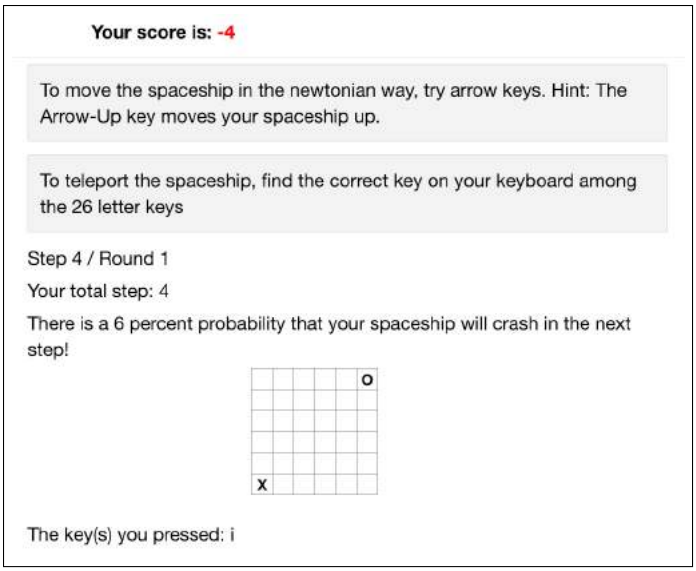


Figure 2: Screenshot of the spaceship game.

Optimal brain points improve learning and performance

To evaluate the potential of our approach to help people overcome the motivational obstacles to learning new skills, we conducted an online experiment where people repeatedly solve a task and can choose to either solve it using a skill that they already possess (Skill 1) or try to learn a new skill that, once mastered, would allow them to solve the task more efficiently (Skill 2). The experimental group received optimal brain points for their choices between exploiting Skill 1 versus attempting to learn Skill 2 whereas the control condition received no brain points. We predicted that a) most participants in the control condition would neglect investing the time and effort necessary to acquire the new skill – even if their investment in learning would pay off in the long run, and b) that optimal brain points can help them overcome this irrational bias.

Methods

We ran our experiment using psiTurk (Gureckis et al., 2016). We recruited a total of 450 participants from Amazon Mechanical Turk between 15:30 EST and 18:30 EST on January 19, 2019, and we restricted the worker region to the United States of America. Participants received \$0.75 for about 6 ± 2 minutes of work and could earn a bonus of up to \$1 (average bonus \$0.10, standard deviation \$0.10) for their performance in the task. Of our 450 participants, 226 were assigned to the control condition and 224 were assigned to the experimental condition with optimal brain points according to psiTurk's counterbalancing method.

Experimental paradigm. We created the *Spaceship Adventure* game shown in Figure 2 and used it to evaluate the efficacy of optimal brain points. The game world is a board with 6×6 cells. The task for the participants is to control the spaceship so as to move from its initial position (0, 0) to its destination (5, 5). Participants play the game for several rounds. After each time they arrive at the destination, the game board is reset and the spaceship is returned to its initial position.

The instructions inform participants that they will be playing the game for multiple rounds. Participants are also informed that there are two modes of moving the spaceship: The spaceship can be moved one step at a time whereas an unknown letter key could be used to teleport the spaceship directly to its final destination. Each step (using the arrow keys or trying out a new letter key) incurs a cost of -1, whereas reaching the destination earns a reward of $+20$. Following each round there was a 6% chance that the game would end and a 94% chance that it would continue (i.e., $\gamma = 0.94$) and participants were informed about that.. The two skills involved in the game are using the arrow key to move the spaceship forward one square at a time (Skill 1) and teleporting the spaceship directly to the destination using one of the 26 letter keys (Skill 2). For each participant the letter key that would teleport their spaceship was independently selected at random before they started their first round and remained the same until the end of their last round.

In the control condition, the only points being shown were the cost of controlling the spaceship and the reward for reaching the goal. In the experimental condition, participants additionally received the optimal brain points described above. Brain points were given for each of the participant's choices between exploiting Skill 1 versus attempting to learn Skill 2. As illustrated in Figure 3, brain points were conveyed using a color-coded score that was accompanied by the image of a brain. The first time, the participant received brain points, those were explained as conveying the value of learning a new skill. To make the brain points more rewarding, a pleasant crystal sound, which is often used to convey a sense

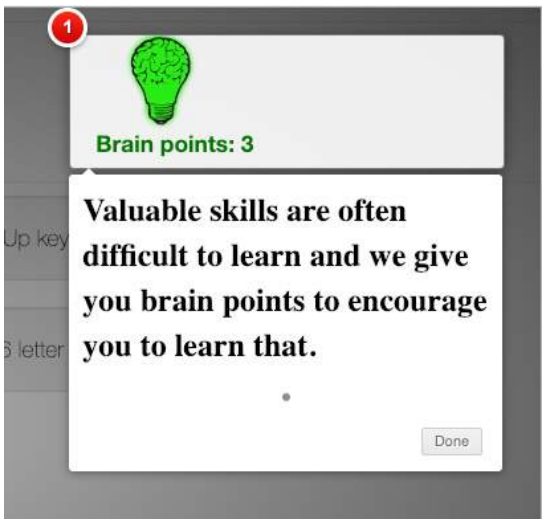


Figure 3: Screenshot illustrating the brain points a participant would receive for trying Skill 2 in the first step.

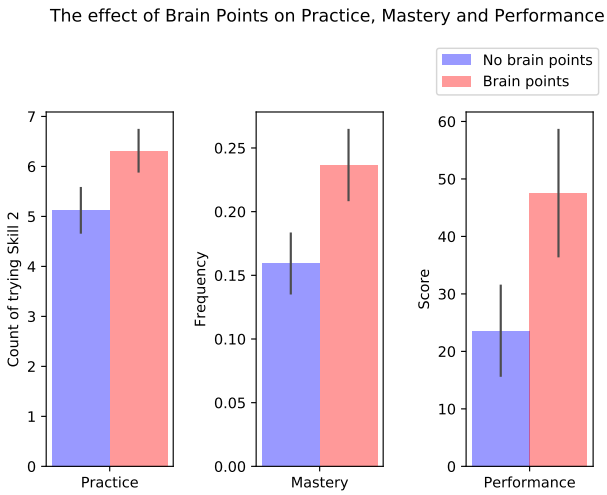


Figure 4: Effect of brain points on practice, learning, and performance. The left bar chart shows the average number of times that people who tried to teleport (skill 2) at least once attempted to figure out how it works until they discovered it or gave up. The middle bar chart shows the proportion of participants in either condition who succeeded to learn skill 2 by discovering which letter key would teleport their spaceship to its target location ("Mastery"). The right bar chart shows the average total score by condition. The error bars represent the standard error of the mean for the bar charts on Practice and Performance and the standard error of the proportion for the Mastery bar chart.

of enlightenment when the players find something valuable in a video game, was played when the number of brain points increased, whereas an unpleasant sound, which could be intuitively perceived as something shrinking, was played when it decreased. Additionally, in both conditions a cheerful sound is played when the spaceship reaches its destination. The brain points score was cumulative as is customary in computer games.

Our code, the experiment, and the data are available on the Open Science Framework at `https://osf.io/k6wjp/`.

Results

As predicted, we found that, when left to their own devices, 42% of the participants never even tried to learn Skill 2 and relied exclusively on Skill 1, although learning Skill 2 could have allowed them to reap higher rewards; that is always attempting Skill 2 would have yielded 154 points on average whereas always exploiting Skill 1 yielded only 8 points on average. This highlights that while there are some situations where people adequately invest into exploring new things (Wilson, Geana, White, Ludvig, & Cohen, 2014), the choice between solving a recurring task with a skill one has already mastered versus using trial-and-error to learn a new skill to be able to handle future occurrences of the task more efficiently might not be one of them for many people.

Encouragingly, we found that optimal brain points significantly increased the proportion of people who attempted to learn the difficult skill (i.e., teleportation, henceforth "Skill 2") from 32% to 46% of participants who had not already tried it in the first step ($\chi^2(1) = 5.74, p = .0165$)[1].

As illustrated in Figure 4, optimal brain points also increased the amount of effort people invested into acquiring Skill 2 from 2.8 to 3.9 attempts on average ($t(448) = 2.52$, $p = .006$; the median number of attempts were 1 and 2 respectively, $Z = 2.59$, $p = .0048$). Furthermore, our optimal brain points also made the people who tried learning Skill 2 at least once more persistent, doubling their median number of additional attempts at learning Skill 2 from 2 to 4 ($Z = 2.49$, $p = .0064$; 4.1 vs. 5.3 on average, $t(448) = 1.86$, $p = .0323$). As a consequence, the proportion of participants who mastered Skill 2 increased from 15% to 24% ($\chi^2(1) = 3.77, p = .0523$), and their average total score doubled from 24 points to 48 points ($t(448) = 1.74, p = .0414$).

These findings suggest that optimal brain points successfully motivated our participants to learn the more difficult skill and thereby improved their learning outcomes and performance.

Conclusion

We derived the expected value of attempting to learn a new skill and translated it into an optimal feedback mechanism for encouraging students to persist in learning valuable skills. Our results suggest that optimal brain points could be useful for helping people overcome the motivational obstacles towards life-long learning. Its basic idea is to reward people's efforts to learn a new skill according to the long-term value of having mastered it and the expected progress towards mas-

[1] We excluded the first action from this analysis because the conditions are identical up until the first feedback is displayed after the participant's first action.

tery.

Our principled computational method for incentivizing learning might become part of the theoretical foundation for the gamification of digital learning environments such as MOOCs or educational games. We hope that the approach illustrated in this article will eventually help people overcome the motivational obstacles that stand in the way of life-long self-directed learning.

Our admittedly simplistic experiment was merely the first step towards evaluating the potential of optimal brain points for increasing student effort. Follow-up experiments should use more naturalistic skill acquisition paradigms and evaluate the proposed feedback mechanism against simpler, heuristic approaches to the gamification of learning environments (Huang & Soman, 2013; Dicheva et al., 2015; Kapp, 2012; O'Rourke et al., 2014). Before we can make any practical recommendations randomized field experiments will have to evaluate our intervention with real students learning real skills.

Future work will evaluate the practical utility of our optimal feedback mechanisms for increasing the student retention rates of MOOCs, encouraging students to use educational games and intelligent tutoring systems more effectively, and building apps that facilitate deliberate practice. These applications may use our method as it is or refine its model of skill acquisition with domain-specific learner models.

While there is a lot of value in being able to motivate students to practice a specific skill inside a digital learning environment, it would be even more valuable if we could help them internalize the value of learning new skills. Future work will therefore investigate whether giving people optimal brain points for their efforts to learn a new skill in one environment can also improve their motivation to learn other skills in different environments and help them develop a growth mindset (Dweck, 2008).

References

Baker, R. S., Corbett, A. T., & Koedinger, K. R. (2004). Detecting student misuse of intelligent tutoring systems. In *International conference on intelligent tutoring systems* (pp. 531–540).

Baker, R. S., Walonoski, J., Heffernan, N., Roll, I., Corbett, A., & Koedinger, K. (2008). Why students engage in gaming the system behavior in interactive learning environments. *Journal of Interactive Learning Research*, *19*(2), 185–224.

Dicheva, D., Dichev, C., Agre, G., & Angelova, G. (2015). Gamification in education: A systematic mapping study. *Journal of Educational Technology & Society*, *18*(3).

Dweck, C. S. (2008). *Mindset: The new psychology of success*. Random House Digital, Inc.

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, *100*(3), 363.

Gureckis, T. M., Martin, J., McDonnell, J., Rich, A. S., Markant, D., Coenen, A., ... Chan, P. (2016). psiTurk: An open-source framework for conducting replicable behavioral experiments online. *Behavior Research Methods*, *48*(3), 829–842.

Huang, W. H.-Y., & Soman, D. (2013). Gamification of education. *Research Report Series: Behavioural Economics in Action, Rotman School of Management, University of Toronto*.

Jordan, K. (2019). *MOOC Completion Rates: The Data.* Retrieved from `www.katyjordan.com/MOOCproject.html`

Kapp, K. M. (2012). *The gamification of learning and instruction: game-based methods and strategies for training and education*. John Wiley & Sons.

Lieder, F., Chen, O. X., Krueger, P. M., & Griffiths, T. L. (2019). Cognitive prostheses for goal achievement. *Nature Human Behavior*.

Lieder, F., & Griffiths, T. L. (2016). Helping people make better decisions using optimal gamification. In A. Papafragou, D. Grodner, D. Mirman, & J. Trueswell (Eds.), *Proceedings of the 38th Annual Meeting of the Cognitive Science Society* (pp. 2075–2080). Austin, TX: Cognitive Science Society.

Mostow, J., Beck, J., Chalasani, R., Cuneo, A., Jia, P., Kadaru, K., et al. (2002). A la recherche du temps perdu, or as time goes by: Where does the time go in a reading tutor that listens? In *International conference on intelligent tutoring systems* (pp. 320–329).

Ng, A. Y., Harada, D., & Russell, S. (1999). Policy invariance under reward transformations: Theory and application to reward shaping. In I. Bratko & S. Dzeroski (Eds.), *Proceedings of the 16th Annual International Conference on Machine Learning* (pp. 278–287). San Francisco, CA: Morgan Kaufmann.

O'Donoghue, T., & Rabin, M. (1999). Doing it now or later. *American Economic Review*, *89*(1), 103–124.

O'Rourke, E., Haimovitz, K., Ballweber, C., Dweck, C., & Popović, Z. (2014). Brain points: a growth mindset incentive structure boosts persistence in an educational game. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 3339–3348). ACM.

O'Rourke, E., Peach, E., Dweck, C. S., & Popovic, Z. (2016). Brain points: A deeper look at a growth mindset incentive structure for an educational game. In *Proceedings of the Third (2016) ACM Conference on Learning@ Scale* (pp. 41–50).

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, *78*(1), 153–189.

Steel, P. (2007). The nature of procrastination: a meta-analytic and theoretical review of quintessential self-regulatory failure. *Psychological Bulletin*, *133*(1), 65–94.

Steel, P., & König, C. J. (2006). Integrating theories of motivation. *Academy of Management Review*, *31*(4), 889–913.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.

Toda, A., Valle, P. H., & Isotani, S. (2018). The dark side of gamification: An overview of negative effects of gamification in education. In A. I. Cristea, I. I. Bittencourt, & F. Lima (Eds.), *Higher Education for All. From Challenges to Novel Technology-Enhanced Solutions.* Cham, Switzerland: Springer Nature.

Urdan, T., & Midgley, C. (2001). Academic self-handicapping: What we know, what more there is to learn. *Educational Psychology Review*, *13*(2), 115–138.

Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A., & Cohen, J. D. (2014). Humans use directed and random exploration to solve the explore–exploit dilemma. *Journal of Experimental Psychology: General*, *143*(6), 2074.

Evaluating Levels of Emotional Contagion with an Embodied Conversational Agent

Özge Nilay Yalçın (oyalcin@sfu.ca)
Interactive Arts and Technology, 250 -13450 102 Avenue
Surrey, BC, V3T 0A3 CANADA

Steve DiPaola (sdipaola@sfu.ca)
Interactive Arts and Technology, 250 -13450 102 Avenue
Surrey, BC, V3T 0A3 CANADA

Abstract

This paper presents an embodied conversational agent framework as a controlled environment to test components of empathy. We implement levels of emotional contagion which includes mimicry and affective matching along with necessary communicational capabilities. We further demonstrate an examination of these foundational behaviors in isolation, to better understand the effect of each level on the perception of empathy in a social conversational scenario with a human actor. We report three studies where the agent shows levels of emotional contagion behavior during (1) the listening act in comparison with baseline backchanneling behavior (2) additional verbal response matching simple emotional storyline (3) the verbal response to the human actor performing complex emotional behaviors. Results revealed that both mimicry and affective matching behaviors were perceived as more empathic than the baseline listening behavior, where the difference between these behaviors was only significant when the agent verbally responded to complex emotional behaviors.

Empathy; Emotional Contagion; Mirroring; Affect Matching; Affective Computing; Social Interaction; Embodied Conversational Agents

Introduction

Empathy, as the capability to understand and react to the emotions of another (Iacoboni, 2011; Coplan & Goldie, 2011), is a complex behavior that arises from the interaction of these basic affective mechanisms with higher-level cognitive functions (de Waal & Preston, 2017). Emotional contagion is said to be the foundation of empathic capacity, as it includes innate and automatic synchronization of the motor and affective responses during an interaction (Hatfield, Cacioppo, & Rapson, 1994). Behaviors such as mimicry and affective matching are levels of the emotional contagion that results from the innate capability of resonating with the other during social interaction.

The literature suggests the sustained act of mimicry results in a feeling of the mimicked emotion and affective matching through muscular feedback (Hatfield et al., 1994; Hatfield, Bensman, Thornton, & Rapson, 2014), while categorizing both behaviors as emotional contagion. Others use affect matching as a highly connected but distinct phenomenon to the mimicry, pointing out the differences between the subjective quality of experience in the emotional contagion and the automatic matching of expressions in mimicry (Hess & Fischer, 2014). However, both ideas converge on the foundational role of mimicry and affect matching in empathic behavior. This notion is consistent with the Perception-Action-Model (PAM) (Preston & De Waal, 2002) and the Russian Doll model of empathy (de Waal, 2007), which integrates the neuroscience studies on mirror neurons as a baseline for the hierarchical levels of empathy mechanisms. However, it is difficult to study the levels of emotional contagion in isolation.

Research efforts often rely on behavioral experiments, neuroscientific techniques (EEG, fMRI) and pathology studies conducted to understand the effects of emotional contagion during social interactions (Hess & Fischer, 2014; Hatfield et al., 2014). As an alternative, computational empathy studies have recently gained attention in a way to simulating the empathy mechanism within the agent and examining empathic responses of the users towards the agent (Paiva, Leite, Boukricha, & Wachsmuth, 2017; Yalçın & DiPaola, 2018). The perception of empathy in artificial agents is shown to increase the length of the interaction (Leite, Castellano, Pereira, Martinho, & Paiva, 2014), user performance (Partala & Surakka, 2004), user satisfaction (Prendinger, Mori, & Ishizuka, 2005), and lead to more trust (Brave, Nass, & Hutchinson, 2005). These findings suggest that equipping interactive systems with empathic capacity would not only improve our understanding of the interaction between cognitive and affective processes in the human mind but may also help us enhance our interaction between artificial systems.

In this work, we use the simulation approach to study empathic behavior in virtual agents and try to understand the differences between the levels of emotional contagion behavior and the perception of empathy during a conversation. We examine the basic emotional contagion capabilities in an embodied conversational agent (ECA) in order to evaluate the perception of empathy during mimicry and affect matching behaviors. We present an agent framework and implementation with necessary communicational capabilities as a baseline. In the following section, we will present our implementation for an ECA that incorporates different levels of emotional contagion as a foundation for empathic capacity. Next, we will demonstrate three experiments that examine the effect of these levels on the perception of empathy during a social interaction scenario with a human actor. Our approach and results show the potential of computational empathy studies as a reliable alternative to test mechanisms for empathic behavior in isolation.

Agent Behavior

Our empathy framework is implemented in an embodied conversational agent that is capable of responding to an emotional conversation with the user using verbal and non-verbal behaviors. Our socially situated 3D virtual character system can perform a set of behavioral acts and context-specific dialogue in response to the speech and video input received from the user (see (Yalçın, in press) for a detailed explanation of the framework). Inputs are gathered using a standard webcam and a microphone. We use the Smartbody behavior realizer (Thiebaux, Marsella, Marshall, & Kallmann, 2008), that can provide face and body gestures, gaze, and speech output for virtual characters. We use the standard Behavior Markup Language (BML) (Kopp et al., 2006) as the basis for the two way connection between the framework and the behavor realizer.

The implementation includes mimicry and affect matching behaviors as the foundational capabilities of empathy in combination with basic conversational capabilities such as backchanneling. In order to achieve this, our system incorporates a perceptual module, a behavior controller and a behavior generation module. The visual and verbal input from the user is processed through the perceptual module, reasoned within the behavior controller according to the selected empathy mechanism and prepared for a behavioral output in the behavior manager before being displayed in the ECA.

Low-level empathic behaviors, such as mimicry and affective matching require a fast response to the emotional stimuli presented by the interaction partner. The fundamental components of this first level of empathic behavior include the perception of emotion, representation of emotion and expressing emotion. This cycle is realized with Perceptual Module and Controller and Behavior Generation modules of our system.

Perceptual Module

The perceptual module is responsible for handling the input received from the user and creating internal representations of these inputs to be used by the controller. Currently, our system is capable of handling audio, video and textual inputs to be used in recognition systems. The audio input includes verbal signals from the user to be recognized as speech and pauses. The initiation, pauses and termination in the speech signal are used to provide information about the dialogue state as well as backchannel timing.

Emotion recognition is a sub-module within the perceptual module that is specialized for emotion recognition and fusion processes. Here, three types of modalities can be used for further processing using the first level of recognition from the perceptual module: facial emotion recognition, tone analysis and speech emotion recognition. During listening, emotion recognition is based on the facial gestures and tone analysis, which is derived from the video and speech inputs for immediate listening feedback. After the speech signal from the user ended, the complete utterance is also being processed in speech emotion recognizer for emotion detection based on the textual output of the speech recognizer. Outputs from this sub-module are used by the behavior controller depending on the dialogue state as well as the selected empathy mechanisms.

Behavior Controller

The behavior controller module is a central unit in the framework which provides a link between inputs and the outputs. It decides which input channel or information to be used depending on the state of the conversation, required empathy mechanisms and the behavioral capabilities of the agent. It is also responsible for providing the information necessary to the behavior manager module to prepare verbal and non-verbal behavior. The Controller acts as a decision-making component, which determines behavioral choices concerning the percepts of the agent and its internal state. Currently, the behavior controller provides a link between the perception-action mechanisms as a key component in computational empathy (de Waal, 2007). During a conversation, the agent should decide which behavioral state it is in depending on the user input: listening, thinking, speaking or waiting. According to the state of the interaction (listening, speaking, thinking and waiting) and the current emotional value (arousal, valence and emotion category), the controller assigns the proper behavior categories to the behavior generation component.

If the user is speaking, the agent should be in the listening mode. Here, the agent is expected to provide proper backchanneling to the user as well as the emotional feedback depending on the empathy mechanisms. After the speech of the user is over, the speech signal should be sent to the dialogue manager component through the controller with the assigned emotional value. The agent will be in thinking mode during the processing of this input by the dialogue manager component. The prepared output sentence will then be sent back to the controller to be sent to the behavior manager which will prepare the output behaviors including face gestures, body gestures and the verbal response to be presented in the speaking mode. After the speech behavior of the agent is done, the waiting or idle mode will be activated until the user speaks again. This cycle can be interrupted via the controller at any stage.

Behavior Generation

The behavior generation module is responsible for preparing the output for the virtual character depending on the emotion, dialogue state and speech information received from the behavior controller. During listening behavior, this module is relatively passive in preparing behaviors. It uses the backchanneling signal to select an appropriate head nod for the agent and a facial expression. When these behaviors are sent and consumed by the behavior realizer, the behavior generation module receives a signal back that indicates the behavior was successfully generated by Smartbody system.

Method

In this paper, we used the simulation approach to study low-level empathic behavior in virtual agents to show the differences between the levels of emotional contagion behavior in the perception of empathy. We examined the effect of mimicry and affect matching behavior on perceived empathy during conversational interaction using three studies.

Participants

Participants for all three experiments were recruited using Amazon's Mechanical Turk platform and were paid for their participation to the study. Because we were focusing on the emotional expressions during verbal communication, we only included participants who had English as their first language. Additionally, users that participate with mobile devices and tablets were excluded to ensure a consistency in the display quality.

A total of 84 subjects participated in the studies. 36 participants with ages ranging from 20 to 60 (M=37.6, SD=10.7) completed the first study. 19 of the participants were male and 16 of them female, while 1 participant defined themselves as 'other'. 24 subjects participated in the second study with ages ranging between 21 and 64 (M=36.17, SD=10.82). 10 of the participants were female and 13 of them male, while 1 participant defined themselves as 'non-binary'. The last study included 24 participants with ages ranging between 23 and 59 (M=37.82, SD=10.64), 12 Male and 12 Female.

Procedure

Studies followed the same procedure, where the participants are asked to evaluate the recorded interaction between the agent and a human (see Figure 1). The interaction scenario consists of a student/participant expressing an emotional story to the agent. We have chosen three stories inspired by the work of Omdahl (Omdahl, 2014), that includes three basic emotion categories: anger, joy and sadness. Other basic emotions such as fear, surprise and disgust were not considered for this study due to the involvement of facial action units that controlled mouth movements during the expression of these emotions. Furthermore, we selected the emotions that would be consistent with the facial emotions, that would not provide an advantage to the affective matching over mimicry.

All of the experiments were deployed in Amazon's Mechanical Turk environment using scripts written in Python 3.6 with psiturk and jspych libraries. Each of the studies takes about 10 minutes to complete. Participants were first shown a test video and were asked to answer two questions about the visual and verbal content of the video, to make sure they can hear and see the videos that are displayed. This was required for the workers to participate in the study.

Each participant is then displayed a short video clip of an interaction, where the agent and a student are shown in a video-conferencing scenario in different conditions (see Figure 1). During the interaction, the student in the video talks about an emotional story in one of three basic emotions: joy, sadness and anger. After displaying each video, the participant is asked to report what the story in the video was about, and also the main emotion of the user and the virtual agent. This is done to make sure the participants are paying attention to the video clips. The participants then evaluated the perceived empathy of the agent towards the student. The perceived empathy of the agent is evaluated by using a modified version of the Toronto empathy questionnaire (Spreng, McKinnon, Mar, & Levine, 2009) which is a 16-item survey that originally is used as a self-report measure. Each item on the questionnaire are scored in a 5-item likert scale (Never = 0; Rarely = 1; Sometimes = 2; Often = 3; Always = 4), where half of the items are worded negatively. Scores are summed to derive total for the perceived empathy and can be varied between -32 to +32. Similar evaluations were suggested by Paiva and colleagues (Paiva et al., 2017), as a modification of Davis's Interpersonal Reactivity Index (Davis et al., 1980).

Figure 1: An image from the video chat between the student and the avatar. Here, the student (left) converses with the avatar.

We used repeated measures design, where each participant is shown all levels of agent behavior in emotional contagion. The type of the interaction study and the order of the conditions are counterbalanced accordingly.

Experiment Conditions

Experiment conditions include three distinct agent behaviors that signifies levels of emotional contagion mechanisms in the empathy framework.

The baseline behavior of the agent is the backchanneling behavior, which is activated depending on the pauses during the speech signal from the audio input component in the perceptual module. In the following subsections, we will provide a detailed examination of three different listening behaviors depending on the level of empathic behavior of the agent: backchannel only, mimicry with backchannel, affective matching with backchannel.

Backchanneling as baseline behavior Listener behavior in humans include backchannels such as head nods, fa-

cial feedback, short vocalizations or a combination of them (Yngve, 1970). These behaviors might show information about listener agreement, acknowledgment, turn-taking or attitude (Schroder et al., 2012; Cassell, Bickmore, Campbell, & Vilhjálmsson, 2000). Backchannel feedback can occur due to change in pitch, disfluency or loudness of the speech signal, as well as shifts in speaker's posture, gaze and head movements (Maatman, Gratch, & Marsella, 2005). In our current implementation we included backchanneling based on the pauses during speech, which is a form of disfluency in the speech signal (Maatman et al., 2005). Information about pauses are extracted from the perceptual module and sent to the controller, which in turn is used to trigger backchanneling as head nods. More advanced methods of adding backchannel that are compatible with the valence of the interaction partner or adding specific facial expressions such as smile, would have interfere with the empathy mechanisms that we would like to test. Therefore, we omitted these behaviors from the baseline behavior.

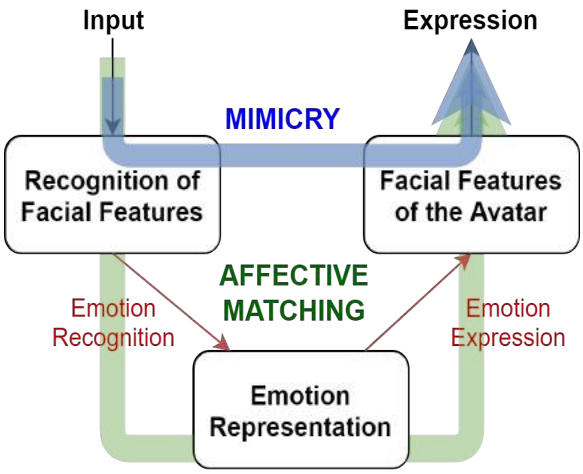


Figure 2: Two paths for emotional contagion. Basic emotional communication competence that results in low-level empathic capabilities of mimicry and affective matching by following distinct routes during the interaction process.

Mimicry Mechanism Mimicry is the lowest level of empathy behavior in our empathy model. It is achieved by a direct mapping between the gestures of the user to the gestures of the agent without being assigned to any type of emotional category. Facial mimicry behavior during listening is a result of mapping the perceived facial action units (AUs) extracted from the perceptual module, to the AUs of the embodied agent in the behavior generation module. The amount, duration and speed of these AUs match the perceived values of the interaction partner without any regulations. In order to avoid mimicking of the lip movements during the speaking of the user, we removed direct mapping of AU18 (lip puckerer), AU26 (jaw drop) and AU24 (lip pressor). As a side-effect of this modification, certain emotions that requires these AUs (fear, surprise and disgust) were not properly expressed. In order to avoid bias, interactions that include these emotions were not used during the evaluation of the system for this study. However, this drawback should be noted for future studies.

After the listening cycle, the agent will sustain the mimicry behavior until it retrieves a response from the dialogue manager. The dialogue manager will then retrieve an emotionally neutral response, due to the lack of emotional representation that is needed to be acquired during the interaction.

Affective Matching Another type of low-level or affective empathy behavior is affective matching (de Waal & Preston, 2017). It is achieved by the emotion recognition and the emotion expression cycle that is connected through emotion representation. As it can be seen in Figure 2, the facial features are mapped to the representation of the basic emotion categories which in turn triggers the facial expressions of the agent that represents those emotions. The amount, duration and speed of these expressions depend directly on the values from the perceived emotions. In contrast to the mimicry behavior, this allows the agent to present and regulate emotions that are better perceived by the users. Moreover, excluded emotion categories in mimicry can be used without the disturbance of the AUs that control mouth muscles as explained in the previous section.

After the listening cycle, the agent will give an emotional feedback that reflects the overall emotion of the interaction partner until it retrieves a response from the dialogue manager. In the affective matching condition, the dialogue manager is able to use the representation of the interaction partner's emotions to pick an emotional response. Without the effect of the higher level emotion regulation capabilities, the agent will pick a response that reflects the emotion of the interaction partner.

Study 1

In order to evaluate the perception of empathic behaviors we compared the listening behavior of the agent in backchannel, mimicry and affective matching conditions. For our study, we used within subjects design where three conditions of agent behavior are shown to the same subject for the evaluation. The conditions are baseline backchanneling behavior, mimicry with backchanneling and affective matching with backchanneling during only the listening act. We used three emotional stories told by the same person, which displays three different emotions as the main theme: joy, sadness and anger. Each video starts with a neutral remark, that is followed by the emotional story.

The experiment counterbalanced on the order of the type of interaction (backchannel, mimicry, affect matching), and the order of type of emotional story (angry, sad, happy). 36 (6x6) different conditions presented to subjects.

Evaluation In the evaluation of the first study, Mauchly's Test of Sphericity indicated that sphericity had not been vio-

lated, X2(2) = 1.748, p = .417. A one-way repeated measures ANOVA was conducted to compare the effect of (IV) level of emotional contagion behavior on (DV) the perception of empathy in backchanneling, mimicry, and affective matching conditions. The results showed that perceived empathy is significantly effected by the type of listening feedback F(2, 70)= 16.721 $p < .0001$, 95%CI (see Figure 3). Pairwise comparisons showed backchannel feedback only (M=-5.47, SD=12.45) is perceived to have significantly lower empathy than both mimicry ($p < .001$) and affective matching ($p < .0001$). However, listening behavior with mimicry (M=5.16, SD=10.64) and affective matching (M=8.22, SD=13.72) did not have any significant difference (p = .18).

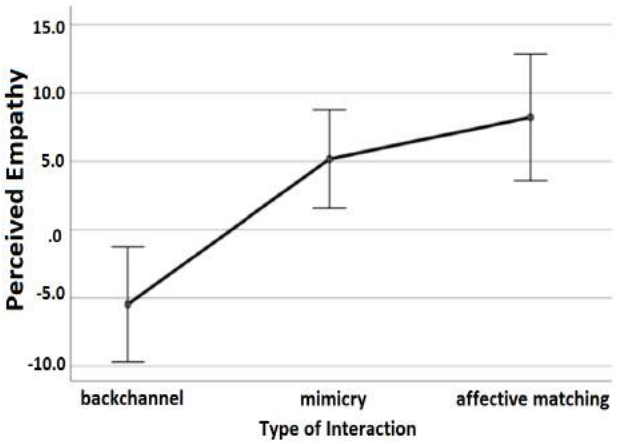


Figure 3: Results of our study showed significant differences in the perceived empathy levels between backchannel, mimicry and affective matching behavior (95%CI).

Study 2 and Study 3

Following up the first study, we further examined the effect of the verbal feedback produced by the dialogue manager in both mimicry and affect matching conditions. Our hypothesis is, due to the effect of emotional representation during the affect matching mechanism, the verbal response behavior of the avatar will be perceived as more empathic. However, this result might show difference when the interaction partner shows more complex emotions, where the context and information about the overall emotion representation is required to understand the semantics of the behavior. Therefore, we conduct two additional studies where one is focused on simple emotions and the other examines the effect of complex emotional behavior. For the following experiments, the participants were asked to evaluate the interaction stories, where the agent listens to different types of emotional stories told by the interaction partner and verbally reacts to it. As the first study showed significant differences over the baseline backchanneling behavior, the following studies did not compare the baseline behavior to emotional contagion.

In both conditions the listening behaviors of the agent will be the same as the first study, which showed no significant difference. The behavior of the agent between will differ from the first study in terms of verbal feedback during the conversational cycle. In the mimicry condition, the agent will produce an emotionally neutral feedback such as "I understand" or "I know what you mean" while sustaining the reflective facial expression of the interaction partner. In the affect matching condition, due to the additional information the dialogue manager will receive from the emotional representation of the interaction partner, the agent will produce an emotionally charged sentence. The emotional category of this sentence will be the same as the emotions of the interaction partner. For example, a happy story will trigger a happy remark such as "That sounds wonderful", an angry story will trigger a response such as "That is really frustrating", and a sad story will trigger a sad response such as "I am sorry to hear that".

The third experiment focused on more complex emotional stories, where the human actor will talk about two scenarios mentioning a dog and a plant. In the dog scenario, the actor will go through excitement, disgust, worry and happiness emotions while mentioning a story about their new pet dog. In the plant scenario, the actor will go through neutral, surprise, worry and happiness emotions while mentioning a story about their friend's plant. The listening behavior of the agent will be matching the emotions both in mimicry and affective matching conditions. Similar to the second study, mimicry condition will result in a generic verbal response from the agent while affective matching condition will give an emotionally charged feedback due to emotional representation.

The second experiment counterbalanced on the order of the type of interaction (mimicry, affect matching), and the order of the type of emotional story (angry, sad, happy). 12 (2x6) different conditions presented to subjects. The third experiment is also counterbalanced on the order of the type of interaction (mimicry, affect matching), and the order of the type of emotional story (dog, plant). 4 (2x2) different conditions presented to the subjects. Both experiments followed the same procedure as the first study.

Evaluations In the second study, one-way repeated measures ANOVA was conducted to compare the effect of (IV) level of emotional contagion behavior on (DV) the perception of empathy in mimicry, and affective matching conditions. The results showed that perceived empathy is not significantly different between mimicry (M=7.62, SD=11.66) and affect matching (M=9.5, SD=8.03) conditions F(1, 23) = 1.030, p = .321.

Following up these results, in the third study, one-way repeated measures ANOVA was conducted to compare the effect of (IV) level of emotional contagion behavior on (DV) the perception of empathy in mimicry, and affective matching conditions during the interaction with complex emotional behavior. The results showed that perceived empathy is significantly different between mimicry (M=0.75, SD=10.45) and affect matching (M=7.21, SD=9.98) conditions F(1, 23) = 7.731, p = .011 (see Figure 4).

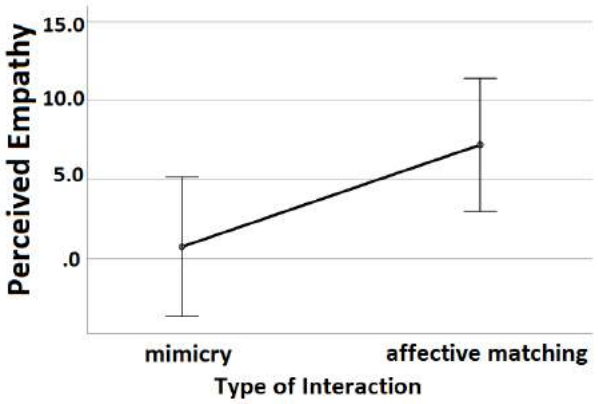


Figure 4: Results of the third study showed significant differences in the perceived empathy levels between mimicry and affective matching behavior in complex emotional interaction (95%CI).

Discussion

The results of the studies showed a significant difference in the perception of empathy between the baseline backchanneling behavior and the emotional contagion behavior during the listening act. As expected, the perceived empathy was significantly higher in the emotional contagion behavior with respect to the baseline behavior in the first study. However, there was no significant difference between the different levels of empathic behavior (mimicry and affective matching) in this experiment. This result is a direct consequence of the similarity in the expressions of these two conditions during listening.

Even though mimicry and affect matching behaviors have important differences in terms of processing of input information, the real-time expressions of these behaviors during listening behavior show dramatic similarities. During the listening act, mimicry captures the facial expressions of the interaction partner and reflects them using the same facial muscles. In affective matching behavior, instead of copying the facial muscles, the system copies the emotions perceived from these facial expressions. As the emotions are expressed as a result of the facial muscles, these two behaviors are expected to show very similar expressions.

One advantage of affective matching that it allows the expression of emotions that are more suitable to the virtual agent, while any emotion will be expressed in terms of the virtual agent's repertoire instead of the expressions of the conversation partner. Moreover, affective matching allows processing of other input channels to conclude the emotion of the interaction partner, such as voice stress, the context of the speech and body expressions. However, the first and second studies only included simple emotions, and therefore such an effect was not present.

Another distinction between mimicry and affect matching conditions is present during the verbal response after the listening act is completed. This response is created by examining the overall emotion of the story told by the interaction partner. As mimicry behavior does not provide the representation of the emotions of the interaction partner, the virtual agent cannot generate a response that is aligned to that emotion. In contrast, the verbal response for the affective matching behavior can be generated from the emotion representation (see Figure 2 for a comparison of these two strategies). Study 2 and 3 are designed to show this distinction.

Interestingly, Study 2 did not show a significant difference between the mimicry and affect matching behaviors for simple emotional stories, where we see a significant difference in Study 3. In these studies, there are two main differences between mimicry and affective matching conditions: the content of the verbal response, and the facial emotions shown during the verbal response. In mimicry condition, the verbal response is generic where the affect matching condition generates an emotionally appropriate response. The facial expressions in mimicry response are sustained regardless of the overall emotions, where the affect matching condition generates facial expression based on the overall emotional representation for the whole story. The difference between the two studies was the emotional complexity of the overall story told by the interaction partner.

We argue that the mimicry response for the simple emotional stories in Study 2, did not show a significant difference on the perception of empathy due to the match between the overall emotion of the story and the sustained facial expression. Where in Study 3 the sustained emotion of the mimicry response was contrasting the overall emotions of the story, due to the complexity of the emotions presented by the interaction partner. We further examined the comments provided by the participants on how they perceived the behavior of the agent in response to the story told by the interaction partner. The comments of the participants in Study 2 showed that the mimicry condition is seen as "understanding" and "sympathy", where the affective matching behavior is seen as "concerned" and "empathy". In contrast, in Study 3, participant comments on the behavior of the virtual agent included descriptions such as "confused" and "indifferent", where the affect matching response was seen more as "attentive", "understanding" and "empathy". However, this distinction should be examined more systematically before reaching to a conclusion.

Overall, these results show that low-level emotional contagion behaviors of the agent during conversational interaction lead to an increased perception of empathy. Additionally, the results show that higher levels of emotional contagion behavior are perceived as more empathic behavior when the interaction includes more complex emotional behaviors. The proposed framework shows promise in providing a foundation to examine the perception of higher levels of empathic behavior during an interaction.

Conclusion and Future Work

Artificial systems provide means to test the empathy theories while allowing the manipulation of parameters in a controlled and isolated way. In this work, we proposed an embodied conversational agent framework to test empathy components and demonstrated three studies that evaluate the foundational empathy mechanisms along with basic communication behaviors. We found that during listening, mimicry and affective matching behaviors are perceived significantly more empathetic compared to backchannel behavior. We also found that the difference between the two levels of affective contagion only significant while the interaction involves complex emotional behaviors, where the context of the interaction is crucial for producing matching behavior. Our framework and the results of our initial study shows promising results that allows for easy integration and testing of higher level components of empathy. The suggested framework, study and evaluation methods shows the potential as a reliable alternative to test mechanisms for empathic behavior in isolation.

Our contributions were to provide a framework, implement the baseline behavior for real-time interaction with a highly realistic conversational avatar, and provide the first study for testing the theoretical assumptions. We hope this baseline for is useful the emerging community of researchers that study empathy in artificial agents and that it can be expanded through this framework and evaluation methods.

References

Brave, S., Nass, C., & Hutchinson, K. (2005). Computers that care: investigating the effects of orientation of emotion exhibited by an embodied computer agent. *International journal of human-computer studies*, *62*(2), 161–178.

Cassell, J., Bickmore, T., Campbell, L., & Vilhjálmsson, H. (2000). Designing embodied conversational agents. *Embodied conversational agents*, 29–63.

Coplan, A., & Goldie, P. (2011). *Empathy: Philosophical and psychological perspectives*. Oxford University Press.

Davis, M. H., et al. (1980). A multidimensional approach to individual differences in empathy.

de Waal, F. B. (2007). The 'russian doll'model of empathy and imitation. *On being moved: From mirror neurons to empathy*, 35–48.

de Waal, F. B., & Preston, S. D. (2017). Mammalian empathy: behavioural manifestations and neural basis. *Nature Reviews Neuroscience*, *18*(8), 498.

Hatfield, E., Bensman, L., Thornton, P. D., & Rapson, R. L. (2014). New perspectives on emotional contagion: A review of classic and recent research on facial mimicry and contagion. *Interpersona: An International Journal on Personal Relationships*, *8*(2), 159–179.

Hatfield, E., Cacioppo, J. T., & Rapson, R. L. (1994). Emotional contagion. *Current directions in psychological science*, *2*(3), 96–100.

Hess, U., & Fischer, A. (2014). Emotional mimicry: Why and when we mimic emotions. *Social and Personality Psychology Compass*, *8*(2), 45–57.

Iacoboni, M. (2011). Within each other. *Empathy: Philosophical and Psychological Perspectives*, 45.

Kopp, S., Krenn, B., Marsella, S., Marshall, A. N., Pelachaud, C., Pirker, H., ... Vilhjálmsson, H. (2006). Towards a common framework for multimodal generation: The behavior markup language. In *International workshop on intelligent virtual agents* (pp. 205–217).

Leite, I., Castellano, G., Pereira, A., Martinho, C., & Paiva, A. (2014). Empathic robots for long-term interaction. *International Journal of Social Robotics*, *6*(3), 329–341.

Maatman, R., Gratch, J., & Marsella, S. (2005). Natural behavior of a listening agent. In *International workshop on intelligent virtual agents* (pp. 25–36).

Omdahl, B. L. (2014). *Cognitive appraisal, emotion, and empathy*. Psychology Press.

Paiva, A., Leite, I., Boukricha, H., & Wachsmuth, I. (2017). Empathy in virtual agents and robots: a survey. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *7*(3), 11.

Partala, T., & Surakka, V. (2004). The effects of affective interventions in human–computer interaction. *Interacting with computers*, *16*(2), 295–309.

Prendinger, H., Mori, J., & Ishizuka, M. (2005). Using human physiology to evaluate subtle expressivity of a virtual quizmaster in a mathematical game. *International journal of human-computer studies*, *62*(2), 231–245.

Preston, S. D., & De Waal, F. B. (2002). Empathy: Its ultimate and proximate bases. *Behavioral and brain sciences*, *25*(1), 1–20.

Schroder, M., Bevacqua, E., Cowie, R., Eyben, F., Gunes, H., Heylen, D., ... others (2012). Building autonomous sensitive artificial listeners. *IEEE Transactions on Affective Computing*, *3*(2), 165–183.

Spreng, R. N., McKinnon, M. C., Mar, R. A., & Levine, B. (2009). The toronto empathy questionnaire: Scale development and initial validation of a factor-analytic solution to multiple empathy measures. *Journal of personality assessment*, *91*(1), 62–71.

Thiebaux, M., Marsella, S., Marshall, A. N., & Kallmann, M. (2008). Smartbody: Behavior realization for embodied conversational agents. In *Proceedings of the 7th international joint conference on autonomous agents and multiagent systems-volume 1* (pp. 151–158).

Yalçın, Ö. N. (in press). Empathy framework for embodied conversational agents. *Cognitive Systems Research*.

Yalçın, Ö. N., & DiPaola, S. (2018). A computational model of empathy for interactive agents. *Biologically Inspired Cognitive Architectures*, *26*, 20 - 25. doi: https://doi.org/10.1016/j.bica.2018.07.010

Yngve, V. H. (1970). On getting a word in edgewise. In *Chicago linguistics society, 6th meeting, 1970* (pp. 567–578).

Mouse Tracking Measures Reveal Cognitive Conflicts Better than Response Time and Accuracy Measures

Takashi Yamauchi (takashi-yamauchi@tamu.edu), Anton Leontyev, Moein Razavi
Department of Psychological and Brain Sciences
College Station, TX 77843 USA

Abstract

Mouse-tracking is said to provide a real-time record of decision making in a conflict situation (Stillman, Shen, & Ferguson, 2018); yet precise benefit of this method is unknown. Using two versions of the attention network task (ANT-R) (Fan et al., 2009), we investigated the extent to which mouse movement measures capture cognitive conflicts created in flanker and Simon tasks. The movement measures collected in the augmented ANT-R (mouse movement condition) were responsive to both flanker and Simon incongruency but response time and accuracy measures in the regular ANT-R (key-press condition) were responsive primarily to flanker incongruency only. The mouse movement measures were also sensitive to interaction effects involving incongruency and gender, trial order and congruency sequence, while response time and accuracy in the regular ANT-R (key-press condition) were mostly insensitive to these interactions. These results suggest that mouse movement measures are more perceptive to cognitive conflicts.

Keywords: mouse-cursor movement; cognitive conflict; cognitive control; flanker and Simon effect

Introduction

One of the major goals of cognitive science is to elucidate the mental mechanism of cognitive operations (i.e., reverse engineering, Marr, 1981), and developing analytic tools that aide this endeavor has been a main preoccupation in cognitive science. Nearly all theoretical debates in the field involve the assessment and interpretation of behavioral data that these tools provide. Bayesian cognitive models, linear mixed effect models, model-based and model-free experimental designs and tasks are geared to help inference of perceptual, cognitive, and affective mechanisms that enable complex human behavior (Barr, Levy, Scheepers, & Tily, 2013; Daw, Gershman, Seymour, Dayan, & Dolan, 2011; Lee & Wagenmakers, 2014).

Ironically, these sophisticated theories and models are based on the age-old dependent measures—how fast and accurately the subject presses a computer key. Yet, it is unclear how reliable these measures are as analytic tools. The problem is that, until recently, cognitive science has had few other viable measures of human behavior.

Using two versions of the attention network task (ANT-R, Fan et al., 2009)—one that primarily measures response time and accuracy through a key press and the other that uses mouse tracking, we compared the extent to which these measures capture cognitive conflicts created in flanker and Simon tasks.

Detecting cognitive conflicts in motor behavior: Mouse-cursor tracking

The theoretical foundation of the mouse-cursor motion research originated from Michael Spivey's conceptualization of human cognitive processing (Spivey, 2007). Traditional theories suggest that cognitive functions such as reasoning, decision making, and problem solving result from symbol manipulations, and computational algorithms for perception, decision, and action are explained by procedures transforming one representational state to another (Marr, 1981). Spivey conceptualizes cognitive functions as a fluid process where probabilistically weighted perceptual-cognitive processing units interact continuously.

Instrumental in Spivey's continuous cognition theory is a series of experiments that measure goal-directed action and decision making (i.e., choice reaching). In a typical choice reaching task, two competing options are pitted against each other (2AFC) and participants are instructed to select one of the choices by clicking on a button by the computer mouse. Unlike a traditional 2AFC task where response time and accuracy are key dependent measures, a choice reaching task has the subject navigate the computer cursor to select a button. By analyzing the navigational path of the cursor from the initial starting position to the end position, researchers found that trajectory features such as AUC (area under the curve) and MAD (maximum absolute deviation) (the degree of deviations from the straight line connecting the starting position to the end position) reveal the subject's perceptual, cognitive, and social conflicts in the decision process (Maldonado, Dunbar, & Chemla, 2019).

The findings in support of this principle come from a broad range, including numerical judgment (Xiao & Yamauchi, 2015), categorization (Dale, Kehoe, & Spivey, 2007), inductive reasoning (Yamauchi, Kohn, & Yu, 2007), linguistic judgment (Spivey, Grosjean, & Knoblich, 2005), racial and gender judgment of morphed face pictures (Freeman & Ambady, 2009; Freeman, Pauker, Apfelbaum, & Ambady, 2009), attitudinal ambivalence toward certain topics (e.g., abortion) (Schneider et al., 2015; Wojnowicz, Ferguson, Dale, & Spivey, 2009), uncertainty in economic choices (Calluso, Committeri, Pezzulo, Lepora, & Tosoni, 2015), and among others (see for review, Freeman, 2018; Stillman et al., 2018; Yamauchi, Leontyev, & Wolfe, 2017). Studies have shown that mouse movement measures can capture semantic incongruency that is processed subliminally

(Xiao & Yamauchi, 2014, 2015, 2017); they even allow automated recognition of emotion, gender and feelings of computer users (Yamauchi & Xiao, 2018; Yamauchi & Bowman, 2014).

One critical question is exactly how well these "continuous" motor measures capture cognitive conflicts as compared to traditional response time and accuracy measures. Is there any advantage of assessing motor measures to study executive control? To address this question, we employed two versions of the attention network task (Fan et al., 2009) and compared the extent to which different dependent measures—traditional response time and accuracy measures and mouse-cursor movement measures—capture flanker and Simon effects.

Cognitive Conflicts in ANT-R

Because attention plays a pivotal role in a wide range of perceptual, cognitive and affective behavior (Posner & Rothbart, 2007), the attention network task provides an ideal testbed to investigate how well cognitive conflicts are reflected in different dependent measures.

The attention network theory (Petersen & Posner, 2012) posits that there are three separate but interactive functions of attention—alerting (being vigilant), orienting (selecting stimuli), and executive control (resolving conflict). A revised version of the attention network task (ANT-R, Fan et al., 2009) has been used widely to probe the interaction and integration of these attention functions, especially cognitive conflicts. The task combines the flanker task (Eriksen & Eriksen, 1974) and the Simon task (Simon & Berbaum, 1990) and creates different types of cognitive conflict (Figure 1). In a flanker task, conflicts are generated by surrounding arrows pointing opposite to the center (target) arrow. In a Simon task, conflicts are created by the stimulus location presented opposite to the center (target) arrow (Figure 1). In both cases, the task of the participant is to indicate the direction of the target (center) arrow.

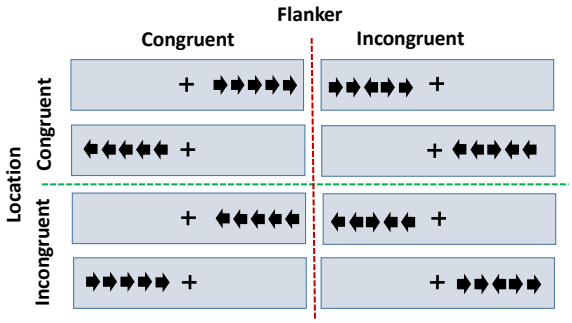


Figure 1: Illustration of flanker and Simon (location) tasks. Flanker congruent and flanker incongruent stimuli are shown in the two columns. Location congruent and location incongruent stimuli are shown in the four rows. The task is to identify the left-right direction of the target (center) arrow.

We devised two versions of the attention network task—traditional and augmented—and contrasted how well traditional response time and accuracy measures and mouse-cursor movement measures can capture the flanker and Simon effects. The traditional attention network task collects only response time and accuracy. Here the subject is to indicate their responses by pressing a designated computer key. The augmented version of the attention network task is identical to the traditional version, except that subjects indicate their response by clicking a button presented on the screen. For this, the subject has to navigate the mouse from the bottom of the screen and press the button. In the augmented version, the x-y coordinate location of the cursor is recorded every 15ms.

(a) (b)

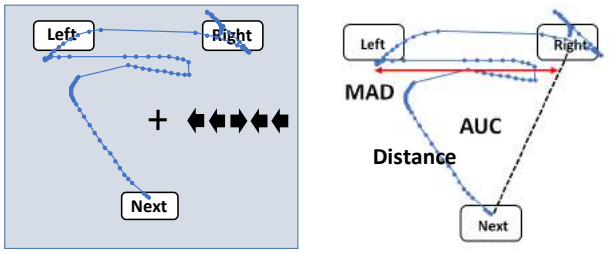


Figure 2: (a) An illustration of an augmented ANT-R trial in the mouse movement condition. To indicate the left/right direction of the target (center) arrow, the participant moves the cursor from the center of the Next button to the final posisiton. The trajectory of the cursor is shown for illustrative purpose and were invisible to participants. (b) AUC (area under curve) is the area enclosed by the trajectory and the straight line connecting the starting position and the end position. MAD (maximum absolute deviation) is the signed maximum absoluite deviation from the direct path. Distance is the sum of Euclidean displacements of the cursor at each sampling point (dots).

The critical question addressed here is how well these depend measures collected from the traditional and augmented ANT-R tasks can capture cognitive conflicts (Figure 1). Although researchers claim the advantage of mouse-cursor measures over traditional measures in extracting cognitive conflicts, this idea has never been explicitly tested. By contrasting the two types of the attention network task, the experiment described below investigate this question directly.

Experiment

The flanker and Simon effects are known to produce robust conflict effects (Eriksen & Eriksen, 1974; Stillman et al., 2018). Although the traditional ANT-R is well suited for the assessment of a flanker-type conflict, the task fails to capture a Simon effect (Fan et al., 2009). Indeed, the Simon effect is particularly difficult to replicate unless the stimulus allows explicit spatial coding (Hommel, 2011). With its emphasis on spatial coding (Figure 1), we predict that the augmented ANT-R are suitable for the assessment of both flanker and Simon effects.

What is unknown is the nature of the effects. Both flanker and Simon effects are subject to contextual factors, such as gender and sequential modulation. The flanker and Simon effects are generally larger in women than men (Stoet, 2017); they are also subject to the trial order. For example, flanker and Simon effects are smaller when two incongruent stimuli are shown in sequence (Egner, 2017). The question addressed here is how well these contextual impacts are reflected in the four dependent measures. If mouse-cursor movement measures are more sensitive than traditional response time and accuracy measures, these interaction effects should be well captured by the mouse-cursor movement measures as compared to the response time and accuracy measures collected in the key-press condition.

Method

Participants Participants (*N*=261) were undergraduate students who enrolled in an introductory psychology course. Participants participated in the experiment for course credit. These participants were randomly assigned to one of two between-subjects conditions—the key-press or mouse movement conditions (key-press = 135, female = 105 male = 30; mouse movement = 126, female = 92, male = 34).

Procedure We employed a revised version of the attention network task (ANT-R, Fan et al., 2009). The ANT-R task is a combination of an arrow flanker (Eriksen & Eriksen, 1974)(Eriksen & Eriksen 1974) and a Simon task (Simon & Berbaum, 1990). A stimulus consisted of five arrows—one center arrow sandwiched by four arrows (two arrows placed both sides). The task of the participant was to indicate the left-right direction of the center arrow (i.e., target arrow). Stimuli (five arrows) were shown either the left or right side of the monitor and the direction of the target arrow was either congruent or incongruent to side arrows (Figure 1).

The key-press and the mouse movement conditions were identical except for one critical point. In the key-press condition, participants indicated the left-right direction of the center arrow by pressing the left or right arrow keys on the keyboard. In the mouse movement conditions, participants used the mouse to indicate the left-right direction of the center arrow. In this condition, two buttons were placed on top left or top right corner of the screen and participants had to navigate the cursor to press the button. (Figure 2a).

ANT-R also incorporates different attention cues (rectangular boxes), which were shown before the presentation of the stimulus at (Figure 3). No cue, double cue, invalid cues, and valid cues were randomly assigned. Because no impacts of attention cues were observed in the present study, the procedure and results involving attention cues are not discussed further.

Altogether each participant received 144 trials, which were divided into eight possible combinations of flanker congruency (congruent, incongruent) and location congruency (congruent, incongruent) and target direction (left, right) (18 trials for each condition and see Figure 1). Eight stimuli in each combination were shown 18 times (8 x 18 = 144), comprising of 144 trials. The order of presenting individual stimuli was determined randomly.

The schedule of stimulus presentation is illustrated in Figure 3. A blank screen with a square is shown; 500ms after the subject clicks the Next button, a fixation sign appears and remains on the screen between 2000ms to 12000ms. The duration between the offset of the target and the onset of the next trial (the cue is shown) varied (approximating an exponential distribution, 2000 to 12,000ms, mean 4000ms). A cue is shown for 100ms. Another fixation is shown for 0, 400, or 800ms (uniform random). A target figure is shown for 500ms. At the onset of the target frame, the cursor is placed at the center of the next button in the mouse movement condition.

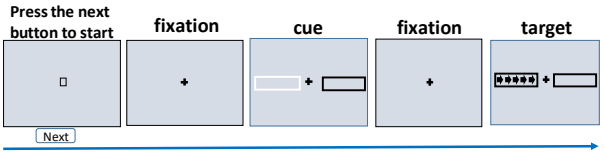


Figure 3: A trial sequence of an ANT-R trial. As the subject press the Next button, a fixation sign appears, followed by a cue, and another fixation sign. Soon after a target frame flashed for 500ms.

Prior to the experiment, all participants received a minimum of 24 practice trials. In the practice trial, corrective feedback was provided after each trial. Practice trials ended when the accuracy was 90% or above in the last 24 trials or a maximum of 48 trials. In 24 practice trials, all possible combinations of flanker congruency, location congruency, and target directions. No cue, double cue, invalid cues, and valid cues were randomly assigned.

Design The experiment had a 2(flanker; congruent, incongruent) x 2(location; congruent, incongruent) x 2 (block order; early, late) x 2 (gender; male, female) design. The key-press and the mouse movement conditions were analyzed separately. Dependent measures in the key-press condition were response time and accuracy (error rate). Dependent measures in the mouse movement condition were AUC, MAD and distance. To analyze the impact of trial sequence, we introduced another factor, congruency sequence (cog_seq; congruent, incongruent), which indicate a congruent or incongruent condition of the stimulus given right before the current stimulus.

To compare the efficacy of the dependent measures, we applied linier mixed-effects models (LMEMs), which are particularly suited to detect population-level systematic effects of manipulations while controlling random variations stemming from individual participants and stimuli. Following the suggestion by Barr et al. (2013), we applied a maximal random-effects structure that was allowed by the experimental design with four fixed factors with two levels; flanker (congruent, incongruent), location (congruent, incongruent), trial order (early, late), and gender (female, male) and two-way interactions among the factors combined with subject-specific random intercepts and item-specific

random intercepts. The first three factors, flanker, location, and trial order are within-subjects variables and gender is a between-subjects variable.

Trials that took longer than and equal to 5000 milliseconds and trials shorter than and equal to 100ms were removed from our data analysis. Outliers were removed using the median-based procedure suggested by Wilcox (p. 77, Wilcox, 2003) (9% of the trials were removed in the key-press condition and 7% of the trials were removed in the mouse movement condition). To ensure that each dependent variable was approximately normally distributed in a similar degree, we transformed each dependent variable with ordered quantile transformation using R package bestNormalize. For all LMEM analyses, we used R packages lme4 and afex, and all dependent variables were rescaled to -1 to 1 (mean = 0). All trajectories were time-normalized using linear interpolation method (101 constant time steps, and see Spivey et al., 2005). We used R package mousetrap (Kieslich & Henninger, 2017) for time normalization and feature extraction (AUC, MAD, and distance).

Result

We first report the results from LMEM analysis followed by a direct comparison of effect sizes. Summaries of these results are shown in Table 1 and Figures 4-6. Following this analysis, we report the impact of congruency sequence.

Table 1: *p*-values from LMEM ANOVA

	RT	Accuracy	AUC	MAD	Dist.
flanker	****	****	****	****	****
location	(**)		****	****	****
flanker x location	(****)	+		+	****
flanker x blkOrder	*		****	***	**
location x blkOrder			+	**	****
flanker x gender			+	*	
location x gender			*	*	

Note. $^{+}p < .10$. $*p < .05$. $**p < .01$. $***p < .001$. $****p < .0001$. Dist. = Distance. (*) opposite direction (congruent > incongruent)

Response time. The response time measure in the key-press condition was quite robust in capturing the flanker effect; $F(1, 172.9) = 662.4$, $p<0.0001$. However, this measure was ineffective for the Simon (location) effect. Although we found a significant main effect of location, the direction of the effect was opposite—participants took longer for location-congruent stimuli than location-incongruent stimuli; $F(1, 172.5) = 7.6$, $p<0.01$. A similar significant "opposite" Simon effect was reported in the Fan et al. (2009) study. The flanker-location interaction effect was significant; $F(1, 132.41) = 17.7$, $p<0.001$.

In general, response time was not very effective in capturing interaction effects. Except for the flanker by block order interaction ($F(1, 17474.4)= 4.2$, $p<0.05$), no other

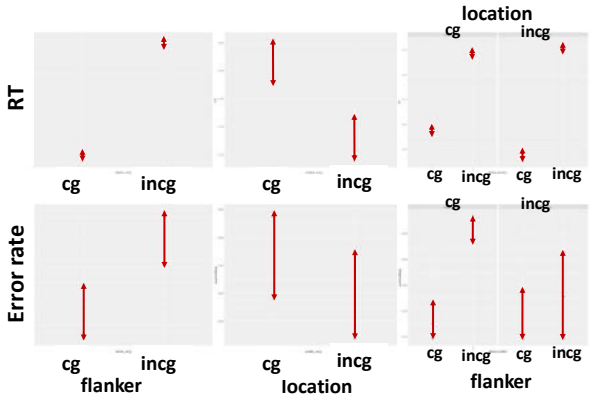


Figure 4: Mean response time (RT) and error rate (accuracy) with flanker (cg=congruent, incg=incongruent, left), location (center), and flanker by location (right) interactions. The arrows represent 95% CI.

interaction effects were significant; location x block order, $F(1, 17473.6)=1.5$, $p=0.22$; flanker x gender, $F<1.0$; location x gender, $F<1.0$ (Figure 4).

Accuracy (error rate) Accuracy (error-rate) was effective in capturing the flanker effect, but not the location (Simon) effect; flanker, $F(1, 133) = 43.0$, $p<0.0001$; location, $F(1, 133) = 1.4$, p=0.24; flanker x location, $F(1, 133) = 2.9$, $p=0.09$. No other interaction effects were observed in accuracy; flanker x block order, $F(1, 133) = 1.0$, $p=0.32$; location x block order, $F(1, 133) = 2.0$, $p=0.16$; flanker x gender, $F<1.0$; location x gender, $F<1.0$ (Figure 4).

AUC (Area under curve). AUC was effective in capturing both the flanker and location (Simon) effects very well. This measure was also sensitive to interaction effects involving gender and block order; flanker, $F(1, 198.9) = 371.9$, $p<0.0001$; location, $F(1, 199.2) = 898.8$, $p<0.0001$; flanker x location, $F<1.0$; flanker x block order, $F(1, 16714.4) = 16.0$, $p<0.0001$; location x block order, $F(1, 16709.2) = 2.9$, $p=0.09$, flanker x gender, $F(1, 16567.9)=3.2$, $p=0.07$; location x gender, $F(1, 16570.9) = 6.3$, $p<0.05$.

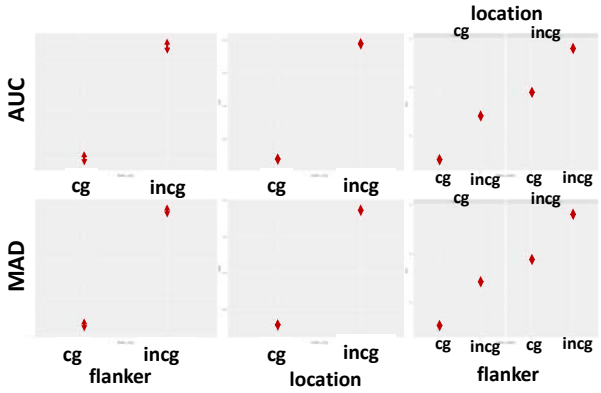


Figure 5: Mean AUC and MAD with flanker (cg=congruent, incg=incongruent, left), location (center), and flanker by location (right) interactions. The arrows represent 95% CIs.

MAD (Maximum Absolute Deviation). MAD was sensitive to flanker and location (Simon) effects, as well as interactions between these terms and block orders; flanker, $F(1, 209.9) = 518.9$, $p<0.0001$; location, $F(1, 210.0)=885.5$, $p<0.0001$; flanker x location, $F(1, 209.8) = 3.3$, $p=0.07$; flanker x block order, $F(1, 17226.3)=12.1$, $p<0.0005$; location x block order, $F(1, 17218.3) = 6.8$, $p<0.01$; flanker x gender, $F(1, 17074.7) = 4.0$, $p=0.05$; location x gender, $F(1, 17075.8) = 4.3$, $p<0.05$.

Distance. Distance responded well to flanker and location effects; flanker, $F(1, 215.3)=214.2$, $p<0.0001$; location, $F(1, 215.3)=109.1$, $p<0.0001$; flanker x location, $F(1, 214.5) = 17.6$, $p<0.0001$. This measure was also sensitive to interactions between these terms and block order; flanker x block order, $F(1, 15822.2) = 9.5$, $p<0.002$; location x block order, $F(1, 15818.9) = 22.7$, $p<0.0001$, but not gender; flanker x gender, $F(1, 15698.3) = 1.2$, $p=0.28$; location x gender, $F<1.0$.

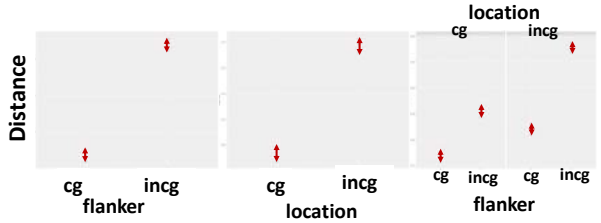


Figure 6: Distance with flanker (cg=congruent, incg=incongruent, left), location (center), and flanker by location (right) interactions. The arrows represent 95% CI.

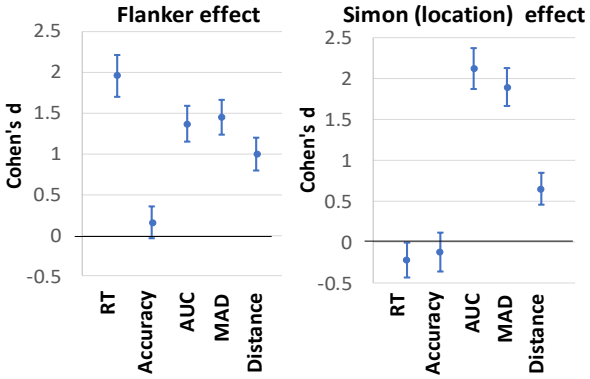


Figure 7: Effect sizes (95% CIs) of flanker (left) and Simon (location) effects. Following Cumming (p. 290, 2012), effect sizes and their CIs for congruent and incongruent conditions were calculated as independent groups.

Effect sizes We compared effect sizes of the flanker and location (Simon) effects captured by the five dependent measures (Figure 7). We observed a large effect size of the flanker effect in the response time measure, as compared to AUC, MAD, and distance; for all comparisons Z's> 2.9, p's<0.001. However, both response time and accuracy measures were ineffective for the location (Simon) effect. In contrast, the effect sizes obtained in the mouse-cursor movement measures were considerably above chance level (Figure 7).

Congruency sequence effects Another important characteristic of cognitive conflict is congruency sequence effects. Flanker and Simon effects are generally smaller when two incongruent stimuli are shown in sequence (Egner, 2017). We examined sequence effects with another factor, congruency sequence (cog_seq; congruent, incongruent), which informs whether preceding stimuli were congruent or incongruent (e.g., flanker (cog, incog) x seq(cog, incog)). This analysis shows that congruency sequence effects were well captured by AUC, MAD, and distance, but not response time and accuracy (Table 2); flanker x seq, RT and accuracy, F's<1.0; AUC, MAD, distance, F's>37.0, p's<0.0001; location x seq, RT and accuracy, F's<1.0; AUC, MAD, F's<1.0; distance, $F(1, 162.4)=4.2$, $p<0.05$ (Table 2).

Table 2: *p*-values for congruency sequence effects

	RT	Acc.	AUC	MAD	Dist.
flanker x seq			****	****	****
location x seq					*

Note. $^{+}p < .10$. $^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$. $^{****}p < .0001$. Dist.=Distance, seq=congruency sequence, RT=response time, Acc.=accuracy (error rate)

Discussion

The cursor movement measures, AUC, MAD, and distance, collected in the mouse-movement condition were responsive to incongruency created in flanker and Simon (location) tasks but response time and accuracy measures in the key-press condition were primarily responsive to flanker incongruency but not location (Simon) incongruency. The mouse movement measures were also sensitive to interaction effects involving incongruency and gender, trial order and congruency sequence, while response time and accuracy in the key-press condition were mostly insensitive to these interactions. These results suggest that the mouse movement measures, as compared to traditional response time and accuracy measures, are more perceptive to flanker and Simon (location) effects.

Researchers have advocated that mouse tracking measures are advantageous for the examination of cognitive conflicts (Freeman, 2018; Stillman et al., 2018). Our results provide empirical support for this idea: the mouse movement measures are statistically more sensitive to various aspects of cognitive conflicts than traditional response time and accuracy measures.

Our results are also consistent with recent findings that performance-based behavior tests for cognitive control (e.g., go/No-go task and stop signal task) can be improved with augmentation of mouse movement measures. Although go/No-go and stop signal tests have been applied widely for the assessment of mental disorders (e.g., ADHD), these tests are ineffective in assessing sub-clinical populations (Toplak, West, & Stanovich, 2013). By augmenting regular go/No-go or stop signal tasks with mouse movement measures,

Leontyev et al. (Leontyev, Sun, Wolfe, & Yamauchi, 2018) demonstrated that these cognitive tests become more reliable in separating individuals with weak and strong symptoms of ADHD-related impulsivity.

Given that the mouse motion measures allow more nuanced examination of cognitive conflict, mouse-tracking measure helps further our theoretical understanding of cognitive control. For example, determinants, boundary conditions, and neural correlates of the congruency sequence effect have been developed, revised and evaluated primarily on the basis of how well the theory accounts for response time and accuracy performance (Egner, 2007). Our results show that different dependent measures can produce different outcomes. In this vein, the validity of these theories (e.g., bottom-up associative theory and top-down control-based theory) can be reexamined with mouse-tracking measures.

Conclusion

For decades, scientific analysis of human behavior has been made mainly on the basis of how fast and accurately an individual responds to a task. Response time and accuracy has served as the primal dependent measures and formidable theories have been developed from these two measurements. The results from this study show that these traditional measures can be supplemented with motor measures, and the mouse-cursor motion analysis provides a viable analytic tool to probe cognitive conflict.

References

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Calluso, C., Committeri, G., Pezzulo, G., Lepora, N., & Tosoni, A. (2015). Analysis of hand kinematics reveals inter-individual differences in intertemporal decision dynamics. *Experimental Brain Research*, *233*(12 LB-Calluso2015), 3597–3611. https://doi.org/10.1007/s00221-015-4427-1

Dale, R., Kehoe, C., & Spivey, M. (2007). Graded motor responses in the time course of categorizing atypical exemplars. *Memory & Cognition*, *35*(1), 15–28.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*(6), 1204–1215.

Egner, T. (2007). Congruency sequence effects. *Cognitive, Affective & Behavioral Neuroscience*, *7*(4), 380–390. https://doi.org/10.3758/CABN.7.4.380

Egner, T. (2017). Past, Present, and Future of the Congruency Sequence Effect as an Index of Cognitive Control. *The Wiley Handbook of Cognitive Control*, 64.

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics*, *16*(1), 143–149.

Fan, J., Gu, X., Guise, K. G., Liu, X., Fossella, J., Wang, H., & Posner, M. I. (2009). Testing the behavioral interaction and integration of attentional networks. *Brain and Cognition*, *70*(2), 209–220. https://doi.org/10.1016/j.bandc.2009.02.002

Freeman, J B, & Ambady, N. (2009). Motions of the hand expose the partial and parallel activation of stereotypes. *Psychological Science*, *20*(10 LB-mouse movement), 1183–1188.

Freeman, J B, Pauker, K., Apfelbaum, E. P., & Ambady, N. L. B. move. (2009). Continuous dynamics in the real-time perception of race. *Journal of Experimental Social Psychology*, *46*, 179–185.

Freeman, Jonathan B. (2018). Doing psychological science by hand. *Current Directions in Psychological Science*.

Hommel, B. (2011). The Simon effect as tool and heuristic. *Acta Psychologica*, *136*(2), 189–202. https://doi.org/10.1016/j.actpsy.2010.04.011

Kieslich, P. J., & Henninger, F. (2017). Mousetrap: An integrated, open-source mouse-tracking package. *Behavior Research Methods*, *49*(5), 1652–1667.

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge university press.

Leontyev, A., Sun, S., Wolfe, M., & Yamauchi, T. (2018). Augmented Go/No-Go Task: Mouse Cursor Motion Measures Improve ADHD Symptom Assessment in Healthy College Students. *Frontiers in Psychology*, *9*, 496.

Maldonado, M., Dunbar, E., & Chemla, E. (2019). Mouse tracking as a window into decision making. *Behavior Research Methods*. https://doi.org/10.3758/s13428-018-01194-x

Marr, D. (1981). *Vision*. New York: Freeman.

Petersen, S. E., & Posner, M. I. (2012). The Attention System of the Human Brain: 20 Years After. *Annual Review of Neuroscience*, *35*(1), 73–89. https://doi.org/10.1146/annurev-neuro-062111-150525

Posner, M. I., & Rothbart, M. K. (2007). Research on Attention Networks as a Model for the Integration of Psychological Science. *Annual Review of Psychology*, *58*(1), 1–23. https://doi.org/10.1146/annurev.psych.58.110405.085516

Schneider, I. K., van Harreveld, F., Rotteveel, M., Topolinski, S., van der Pligt, J., Schwarz, N., & Koole, S. L. (2015). The path of ambivalence: tracing the pull of opposing evaluations using mouse trajectories. *Frontiers in Psychology*, *6*, 1–12.

Simon, J. R., & Berbaum, K. (1990). Effect of conflicting cues on information processing: the 'Stroop effect'vs. the 'Simon effect.' *Acta Psychologica*, *73*(2), 159–170.

Spivey, M. J. (2007). *The Continuity of Mind*. Oxford: Oxford University Press.

Spivey, M. J., Grosjean, M., & Knoblich, G. (2005). Continuous attraction toward phonological competitors. *Proceedings of National Academy of Sciences of the United States of America*, *102*(29), 10393–10398.

Stillman, P. E., Shen, X., & Ferguson, M. J. (2018). How

mouse-tracking can advance social cognitive theory. *Trends in Cognitive Sciences*.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2013). Practitioner review: Do performance-based measures and ratings of executive function assess the same construct? *Journal of Child Psychology and Psychiatry*, *54*(2), 131–143.

Wilcox, R. R. (2003). *Applying contemporary statistical techniques*. Elsevier.

Wojnowicz, M. T., Ferguson, M. J., Dale, R., & Spivey, M. J. (2009). The self-organization of explicit attitudes. *Psychological Science*, *20*(11), 1428–1435.

Xiao, K., & Yamauchi, T. (2014). Semantic priming revealed by mouse movement trajectories. *Consciousness and Cognition*, *27*(1). https://doi.org/10.1016/j.concog.2014.04.004

Xiao, K., & Yamauchi, T. (2015). Subliminal semantic priming in near absence of attention: A cursor motion study. *Consciousness and Cognition*, *38*. https://doi.org/10.1016/j.concog.2015.09.013

Xiao, K., & Yamauchi, T. (2017). The role of attention in subliminal semantic processing: A mouse tracking study. *PLoS ONE*, *12*(6). https://doi.org/10.1371/journal.pone.0178740

Yamauchi, T., Kohn, N., & Yu, N.-Y. (2007). Tracking mouse movement in feature inference: Category labels are different from feature labels. *Memory and Cognition*, *35*(5). https://doi.org/10.3758/BF03193460

Yamauchi, T., & Xiao, K. (2018). Reading Emotion From Mouse Cursor Motions: Affective Computing Approach. *Cognitive Science*, *42*(3). https://doi.org/10.1111/cogs.12557

Yamauchi, T, & Bowman, C. (2014). Mining cursor motions to find the gender, experience and feelings of computer users. *IEEE International Conference on Data Mining (ICDM 2014): Workshop on Domain Driven Data Mining*, 221–230. https://doi.org/10.1109/ICDMW.2014.131

Yamauchi, Takashi, Leontyev, A., & Wolfe, M. (2017). Choice reaching trajectory analysis as essential behavioral measures for psychological science. *Insights in Psychology*, *1*(4), 1.

A perspective-change based account of creativity evaluation: An investigation in simile assessments

Shiyu Yang (shiyuy3@illinois.edu)

Jeffrey Loewenstein (jloew@illinois.edu)
Gies College of Business, University of Illinois at Urbana-Champaign
1206 S. 6th St, Champaign, IL 61820

Abstract

Why do people experience something as creative? We propose a perspective-change based account of creativity evaluation. Drawing upon structure mapping theory (Gentner, 1983), we show that people evaluate a simile to be creative when they spontaneously (Study 1) or are induced (Study 2) to experience a change in perspective. This account further predicts that people are unlikely to find a simile creative if they are unable to form a working perspective, as is in the case of anomalies. In addition, a simile is unlikely to be evaluated as creative when people's initial perspectives are sufficient to interpret the simile, as in the case of literal statements. We further show that repeated use of the same perspective suppresses the experience of perspective change and thus reduces creativity perception (Study 3).

Keywords: creativity evaluation; analogy; simile; perspective-change; structure mapping theory

Introduction

Author J. K. Rowling apparently received many rejection letters for her first Harry Potter novel. Stories abound in academia about seminal, award-winning papers that were initially rejected by journals. Innovation requires more than just generating creative ideas—it also requires being able to evaluate ideas. Consequently, creativity evaluation is a critical and challenging step in the innovation process (Mueller, 2017). However, our knowledge of how lay people form creativity judgments is still limited (Zhou, Wang, Bavato, Tasselli, & Wu, 2019). This paper contributes by proposing and providing initial tests of a perspective-change based account of creativity evaluation.

We focus on perspective change following a proposal that evaluating creative ideas is somewhat like generating creative ideas (Cronin & Loewenstein, 2018). There is a long tradition in creativity research emphasizing the key role of changing one's perspective, discussed variably as, for example, the reorganization of cognitive structures (Mumford & Gustafson, 1988), breaking set (Boring, 1950), restructuration (Duncker & Lees, 1945), deviation from habitual use of knowledge (Luchins, 1960), and transformation (Boden, 2004). These ideas are related to work beyond the creativity literature on conceptual change (Chi, 2009) and re-representation (Gentner & Wolff, 2000). Following terminology from Page (2008) and Cronin and Loewenstein (2018), we describe it as perspective change. Briefly, as any mental representation is a partial rather than a complete account, it necessarily only provides a perspective on whatever is being represented. It follows then that adopting a particular mental representation of a situation leaves open the possibility of changing to an alternative mental representation that is both appropriate to the situation and incompatible with the first mental representation. The possibility pursued here, building on the argument by Cronin and Loewenstein (2018), is that if the process of generating creative ideas involves a change in perspective, then it might also be the case that when the process of forming an interpretation of an item leads us to change our perspective, we are likely to perceive the item to be creative. Thus, the proposal is that creativity evaluations rest at least in part on the process of forming interpretations, and that process echoing the process of generating creative ideas.

To explore this perspective change account of creativity evaluation, we asked participants to evaluate similes: A is like B. Prior work has established that such statements can convey fresh analogies, anomalies, or mundane literal similarities or class inclusions (Bowdle & Gentner, 2005; Gentner, 1989). The perspective change account of creativity evaluation makes predictions about each case. Specifically, a simile is unlikely to be judged creative if people cannot form a coherent interpretation of it—that is, if it is an anomaly. A simile is unlikely to be judged creative if people's initial, default interpretation is apt—that is, if it is a mundane literal similarity comparison or class inclusion. In contrast, a simile is likely to be judged creative if people's initial, default interpretation is not apt but they are able to find an alternative interpretation that is appropriate—that is, it is experienced as a fresh analogy.

To further examine the role of perspective change, we draw upon the habituation paradigm (Rankin et al., 2009) to show that repeated exposure to a perspective can lead to diminished perspective change and thus reduces creativity evaluation.

Study 1: Spontaneous Perspective Change

Method

Participants This study involved 147 students from a mid-west university who participated in the study for course credit (49% male, $Mean_{age}$ = 20.65, SD_{age} = 2.11, 48.30% white, 5.44% black, 41.50% Asian, 4.68% other). No participant was excluded from the analysis.

Materials and Design We generated five groups of similes, with each group using the same target and three different bases. The five targets were: diamond, crib, snowflake, pencil, and closet. Drawing upon the structure mapping framework for analogy (Gentner, 1983), we composed three types of

similes for each of the five targets: 1) anomaly, e.g., a crib is like seaweed; 2) literal similarity, e.g., a crib is like a bed; 3) analogy, e.g., a crib is like a cocoon.

Participants saw the 15 similes twice. First, they rated them for creativity on a 7-point Likert scale (1="highly uncreative", 7="highly creative"). Next, they categorized each statement as: "nonsensical", "a literal comparison", "a metaphor". Participants gave ratings of perspective change by being asked the extent to which the statement made them think differently about the target in the statement on a 5-point Likert scale (1 = "Not at all differently", 5 = "Extremely differently"). Participants also wrote down their interpretations of the statements.

Results and Discussion

Table 1 provides descriptive data. Consistent with the materials design, most of the anomalies, literal similarity statements, and analogies were categorized by participants as such ("Nonsensical," "Literal" and "Metaphor"; bold numbers in Table 1).

Creativity evaluations and perspective-change scores are presented in the rows of Table 1. As expected, aggregating across items, we found that the analogy type ($Mean_{All3} = 4.84$) was evaluated to be more creative than the anomaly type ($Mean_{All1} = 3.74$), $F(1, 1468) = 171.64$, $p < .00$ as well as the literal type ($Mean_{All2} = 2.62$), $F(1, 1468) = 852.80$, $p < .00$. The same pattern held for perspective-change scores. The analogy type ($Mean_{All3} = 2.56$) was rated as more perspective-changing than the anomaly type ($Mean_{All1} = 1.77$), $F(1, 1468) = 164.88$, $p < .00$ as well as the literal type ($Mean_{All2} = 1.33$), $F(1, 1468) = 554.7$, $p < .00$.

We also analyzed creativity evaluations and perspective-change scores for the similes by how participants categorized, and so presumably how they experienced, them (Table 1). Aggregating across items, planned contrasts showed that items categorized by participants as metaphors ($M = 5.18$, $SD = 1.38$) were evaluated to be more creative than items categorized as nonsensical ($M = 3.36$, $SD = 1.62$), $F(1, 1320) = 400.07$, $p < .00$, $\eta^2 = 0.23$. They were also evaluated to be more creative than items categorized as literal ($M = 2.72$, $SD = 1.51$), $F(1, 1320) = 842.3$, $p < .00$, $\eta^2 = 0.39$. A similar pattern held for the perspective change ratings. Similes that participants categorized as metaphors ($M = 2.84$, $SD = 1.13$) were rated as more perspective-changing than those categorized as nonsensical ($M = 1.30$, $SD = 0.93$), $F(1, 1320) = 733.21$, $p < .00$, $\eta^2 = 0.36$ They were also rated as more perspective-changing than similes categorized as literal ($M = 1.46$, $SD = 0.82$), $F(1, 1320) = 627.07$, $p < .00$, $\eta^2 = 0.32$.

The consistency in the patterns between creativity evaluations and perspective-change scores held not just in the aggregate but also at the level of individual items. The correlation between creativity evaluations and perspective change scores was high, $r = 0.50$, $p < .00$.

Taken together, these findings are consistent with the possibility that people evaluate similes to be creative to the extent that they formed interpretations that differed from how they usually interpreted the target.

Table 1: Categorizations, creativity evaluations, and perspective change scores in Study 1

List	Category (%)			Creativity			P-change		
	Ns	Lit	Met	Ns	Lit	Met	Ns	Lit	Met
All1	**71**	6	23	3.41	3.5	4.81	1.32	2.5	2.94
All2	12	**85**	3	3.08	2.52	3.57	1.46	1.29	2.24
All3	8	12	**80**	3.68	4.05	5.07	1.35	2.44	2.70

Note: Ns-Nonsensical, Lit-Literal, Met-Metaphor; 1, 2, and 3 denotes nonsensical, literal, and metaphor, respectively. Note that Ns, Lit, and Met denote participants' categorizations, whereas 1, 2, and 3 denote the intended type of statement in the design of the materials.

This was seen in the highest creativity evaluations being given to those similes intended as analogies as well as in the high correlation between creativity evaluations and perspective-change scores. But perhaps the most compelling aspect of the data is that it was not the similes themselves that mattered so much as participants' own categorizations of the similes. The same simile could be and were categorized differently by different participants. What was perceived to be an anomaly by some was perceived to be a metaphor by others, and the perspective-change scores and creativity evaluations followed from those subjective interpretations.

Taken together, the results of Study 1 found that a simile is likely to be evaluated as creative to the extent that people experience a change in perspective as they form an interpretation of the similarities between the target and the base. This is initial evidence consistent with a perspective-change based account of creativity evaluation—that what drives creativity judgments is experiencing a perspective change in the course of forming an interpretation of the item one is evaluating.

Study 2: Induced Perspective Change

Study 2 builds on Study 1 by randomly assigning individuals to conditions that should encourage or discourage them from experiencing a change in perspective (cf., Day & Asmuth, 2017). Specifically, before participants evaluated an anomalous simile, they first read a short paragraph. That paragraph contained information that was either relevant or irrelevant to comprehending the simile as an analogy. Thus, we sought to enable participants to form a coherent change in perspective or limit them from doing so, and as a result encourage or hinder them from perceiving the simile to be creative.

Method

Participants We recruited 237 participants from Mturk (45% male, $Mean_{age} = 37.59$, $SD_{age} = 12.22$, 78.90% white, 11.39% black, 7.17% Asian, 2.53% other). Participants qualified for the study if they were located in the United States and had an approval rate above 95% in previous ''Human Intelligence Tasks" (HITs) on MTurk. None of the participants was excluded from the analysis.

Materials and Design Participants saw one of two similes: (1) Pigeons are like snowflakes, or (2) Seaweed is like a crib. Before reading the simile itself, they first read a paragraph that was either relevant or irrelevant to interpreting the target of the simile in a way that supports interpreting the simile as an analogy.

Specifically, for participants assigned to evaluate the simile about pigeons, they read one of the two paragraphs:

Pigeons often move together and descend to the ground, blanketing it and changing its color. At some times of year, out in the countryside it seems that the wind brings pigeons and covers the roofs with the feathered creatures. Or go to a town square at certain times of day or the year and soon you may find that pigeons gradually cover the entire square (relevant information condition*).*

Pigeon is a French word that derives from the Latin pipio, for "peeping", based on the sounds the birds make. Pigeons are a common species. They are stout-bodied birds with short necks and slender bills. Pigeons primarily feed on seeds, fruits, and plants. Most pigeons lay one or two eggs at a time, and both the male and female pigeons care for the young (irrelevant information condition).

For participants assigned to evaluate the simile about seaweed, they read one of the two paragraphs:

Fish sometimes benefit from the protection provided by seaweed. The ribbons of seaweed extending upwards from the sea floor form a safe space in which fish can place their eggs. Seaweed provides a shelter for the baby fish. They can rest in the protected space that the seaweed provides. When baby fish outgrow their seaweed home, they can explore the open waters until they are ready to lay eggs of their own (relevant information condition).

Seaweed is a popular snack. All seaweed food is low in calories and fat. Dried seaweed comes in various flavors and is sold in sheets, flakes, or handy snack packs. Fresh seaweed, on the other hand, is commonly sold as an ingredient in prepared foods like sushi or seaweed salad. Canned seaweed snacks are also becoming trendy now; you can easily find them in the refrigerated section of the supermarkets (irrelevant information condition).

After reading these passages and rating them for how informative they were, participants then saw the target simile. Specifically, they rated how creative the simile was on a 5-point Likert scale (1= "Not at all creative", 5= "Highly creative"). They also categorized each simile ("nonsensical" or "metaphor"), provided perspective change ratings by indicating the extent to which the simile made them think differently about the target on a 5-point Likert scale (1 = "Not at all different", 5 = "Extremely differently"), and wrote interpretations.

Results and Discussion

As predicted, the initial passages that participants read influenced their interpretations of the similes and their judgments of creativity. Specifically, for the simile about pigeons, participants assigned to the relevant information condition rated it as more creative than those in the irrelevant information condition ($Mean_{relevant}$ = 3.08, $SD_{relevant}$ = 1.21, $Mean_{irrelevant}$ = 2.34, $SD_{irrelevant}$ = 1.22, t = 3.33, p < .00, Cohen's d = 0.61). In addition, a higher proportion of participants categorized the simile as a metaphor in the relevant information condition than in the irrelevant information condition ($Metaphor_{relevant}$ = 68%, $Metaphor_{irrelevant}$ = 30%, χ^2 = 17.61, p < .00, φ = 0.38). Lastly, participants assigned to the relevant information condition also rated the simile as more perspective-changing than those in the irrelevant information condition ($Mean_{relevant}$ = 2.59, $SD_{relevant}$ = 1.12, $Mean_{irrelevant}$ = 1.97, $SD_{irrelevant}$ = 1.08, t = 3.12, p < .00, Cohen's d = 0.57).

The same pattern was found for the simile about seaweed. Specifically, compared to participants assigned to the irrelevant information condition, those assigned to the relevant information condition rated the simile as more creative ($Mean_{relevant}$ = 3.26, $SD_{relevant}$ = 1.17, $Mean_{irrelevant}$ = 1.84, $SD_{irrelevant}$ = 1.20, t = 6.51, p < .00, Cohen's d = 1.20), were more likely to categorize it as a metaphor rather than a nonsensical statement ($Metaphor_{relevant}$ = 75%, $Metaphor_{irrelevant}$ = 12%, $\chi 2$ = 46.47, p < .00, φ = 0.63), and rated it as more perspective-changing ($Mean_{relevant}$ = 2.72, $SD_{relevant}$ = 1.06, $Mean_{irrelevant}$ = 1.67, $SD_{irrelevant}$ = 1.16, t = 5.22, p < .00, Cohen's d = 0.97).

Study 2 found that providing participants with particular information about the target in a simile could encourage them to see a coherent, novel metaphor in what otherwise would likely have been an anomalous statement. This was likely to be experienced as a change in perspective and likely to have led to considering the simile to be creative. Thus, in inducing a perspective-change and prompting evaluations of creativity, this study offers further support for perspective change playing a role in the process of forming creativity evaluations.

Study 3: Suppressed Perspective Chang

Study 3 tests whether minimizing a perspective change will lead to lower evaluations of creativity. If experiencing a change in perspective contributes to judging something to be creative, then continuing with an existing perspective could dampen judgments of creativity. For example, examining several items in a row could provide an opportunity to compare judgments of the same item when it is either distinct from what has come before and so a change in perspective, or in line with what has come before and so consistent with the existing perspective.

A variety of research examines sequences of judgments of potentially similar and potentially different items, ranging from research using a habituation paradigm (Rankin et al., 2009), to research on deviant items (Sakamoto & Love, 2004), to work on expectation violations (Loewenstein, 2019). We used work on repetition-break structures (Loewenstein & Heath, 2009; Loewenstein, Raghunathan & Heath, 2011) to design sequences of items. The repetition-break structure allows us to identify items that are likely to be experienced as a perspective change: the first item and the break item. It also allows us to identify items that are likely to be experienced as consistent with the existing perspective: the second and any subsequent items that are highly similar to the first one. Thus, we can use sequencing to lead to either a diminished perspective change and therefore diminished

creativity judgments, or a perspective change and therefore expected creativity judgments.

The study uses sequences of similes, some of which are arranged using the repetition-break structure. For a given target in a simile, we generated two sets of bases incorporating two distinctive perspectives. We used three similar bases to establish an initial repetition pattern, and then broke this pattern by contrasting this set of three similes with a fourth simile that whose base drew from a different perspective. In this way, for each target we are able to create two distinctive interpretations of it. Thus, we expected an evaluation experience that can be described by the following process: initial exposure to perspective A → habituation to perspective A with the second and third exposures → initial exposure to a different perspective B. It is at the last stage of this evaluation process that we expect to observe the switch from perspective A to a second different perspective B, and so be experienced as a perspective change.

If the experience of a perspective change underlies creativity evaluation as we proposed, we should expect to see the following patterns: First, the gradual habituation to a perspective will result in reduced ratings of creativity. Through repeated exposure to similar bases incorporating the same perspective, the initial perspective will be gradually assimilated into participants' knowledge structure and thus should lose its freshness. Since it can no longer induce any departure from or change to participants already existing interpretations regarding the target, similes incorporating this perspective will be perceived as less creative. We therefore predict that there will be a decrease in creativity and perspective-change ratings over the course of encounters with the first, second, and third simile.

Second, in the repetition break condition, given that the last simile conveys a second perspective of the target that is different from the first one, we should observe that comprehending the fourth simile will induce a change in perspective—switching from one way of interpreting the target to another way. As a result of this experienced perspective-change, we predict that there will be a jump in creativity and perspective-change ratings for the last simile.

Method

Participants We recruited 428 participants from Mturk. There were 14 participants who failed the attention check. They were excluded, leaving 414 in the final sample (44% male, $\text{Mean}_{\text{age}} = 36.91$, $SD_{\text{age}} = 12.31$, 78.83% white, 7.79% black, 6.81% Asian, 6.38% other). Unless otherwise noted, inclusion of the 14 participants did not change any of the results reported below substantially. Participants qualified for the study if they were located in the United States and had an approval rate above 95% in previous ''Human Intelligence Tasks" (HITs) on MTurk. This sample size was determined using G*Power (Faul, Erdfelder, Lang, & Buchner, 2007), based on a priori power analysis by setting 80% statistical power with an effect size of 0.50, which was obtained through pilot studies.

Materials and Design To generate similes that can be organized in the repetition break structure, we conducted several rounds of pretests. Our final materials included similes with two targets: (1) poverty and (2) marriage. For each target, we generated two sets of bases, with four bases in each set. Each set of bases are synonyms or phrases with similar meanings, such that they are centered around one specific interpretation of the target (i.e., a perspective). The two perspectives for interpreting poverty were that it is a destructive, spreading influence (i.e., poverty is like an infection/an illness/a disease/a virus) and that it is a barrier (i.e., poverty is like a fence/a barricade/a moat/a wall). The two perspectives for interpreting marriage were that it is a thrilling activity (i.e., marriage is like skydiving/bungee jumping/parachuting/hang gliding) and that it is a nurturing activity (i.e., marriage is like growing flowers/caring for your lawn/farming/gardening).

We used a 3 (three structure conditions) by 2 (two targets) by 2 (two sets of bases) design. The first factor is structure, which has three conditions: repetition break, consistent, and baseline. Using "poverty is like a virus" as an example, this simile conveys the interpretation that poverty is destructive and contagious. In the repetition break condition, three similes conveying the other interpretation of poverty were presented in a sequence prior to it (i.e., poverty is like a fence/a barricade/a moat/a wall). This way, when appeared at the fourth position, the focal simile "poverty is like a virus" constituted a break from the initial perspective for poverty.

In the consistent condition, the preceding three similes conveyed the same perspective (i.e., poverty is like an infection/an illness/a disease) as the ending simile, such that the focal simile "poverty is like a virus" is a continuance of rather than a break from the initial perspective. In this condition the focal simile in the fourth position is consistent with the preceding perspective. Lastly, in the baseline condition, the focal simile "poverty is like a virus" was presented on its own without any preceding similes about poverty.

Across the three conditions, our focus is on comparing the creativity and perspective-change ratings for the focal simile— "poverty is like a virus". We expect that: (1) the consistent condition will generate lower creativity and perspective-change ratings than the baseline condition; (2) the consistent condition will generate lower creativity and perspective-change ratings than the repetition break condition; and (3) given that the standing-alone condition provides us with a baseline level of how creative and perspective-changing the focal simile is, we expect to see that the repetition break condition will lead to a jump in perceived creativity and perspective-change of the focal simile such that the ratings are restored to a level comparable to that in the baseline condition.

The second factor in our design was the target (i.e., poverty vs. marriage). We chose to use two rather than just one target in order to show that the pattern of creativity evaluation we predicted did not hinge on the idiosyncrasies of one specific simile target, and that the effect of perspective-change can be generalized to similes with various targets.

In a similar spirit, for each target, we fully counterbalanced the position of the two perspectives, such that each perspective was placed as the opening one (i.e., the repetition stage) in one condition and the ending one (i.e., the break

stage) in another condition. Our intention is to show that the pattern of creativity evaluation we predicted did not rely on one specific perspective being in a specific position. Regardless of the specific content of a perspective, it is the cognitive experience of viewing it as either a continuance of or a break from the already existing perspective that predicts how creative people perceive it to be.

In the current study we have four focal similes: poverty is like a virus/a wall, and marriage is like gardening/hang gliding. Each of the four focal simile is presented in three different structures: repetition break, consistent, and baseline. We thereby generated a total of 12 conditions. In each condition, participants were asked to rate one (i.e., the baseline condition) or four similes (i.e., consistent condition and repetition break condition) for creativity (1 = "not at all creative", 5 = "extremely creative") and perspective-change (1 = "not at all perspective-changing", 5 = "extremely perspective-changing").

In the repetition break and the consistent conditions, the focal simile always appeared at the fourth position. Given that the focus of the current study is on examining the experience of switching from one coherent interpretation to another one, being able to understand a simile and form a perspective in the first place is therefore an important prerequisite. Thus, we gave participants the option of marking a simile as non-sensical if they failed to form an interpretation of it ("this simile doesn't make sense to me"), in which case they were excluded from the analysis.

Results and Discussion

The creativity ratings showed that the consistent condition generated the lowest level of creativity (M = 2.07, *SD* = 0.95), lower than both the baseline condition (M = 2.70, *SD* = 1.07, $t = 7.13$, $p = 0.000$, Cohen's $d = 0.62$) and the repetition break condition (M = 2.89, *SD* = 1.09, $t = 9.06$, $p = 0.000$, Cohen's $d = 0.80$). Although we didn't predict to see a significant difference across the repetition break condition and the baseline condition, results showed that while the difference was small in absolute magnitude (i.e., 0.19), it reached statistical significance level ($t = 2.02$, $p = 0.04$, Cohen's $d = 0.18$). These are overall effects of the structure condition; there were no reliable effects or interactions due to the particular targets or bases so we collapsed across them.

Analysis of perspective-change ratings showed a similar pattern, such that the consistent condition generated the lowest level of perspective-change (M = 1.80, *SD* = 1.12), lower than both the baseline condition (M = 2.45, *SD* = 1.20, $t = 5.43$, $p = 0.000$, Cohen's $d = 0.47$) and the repetition break condition (M = 2.50, *SD* = 1.33, $t = 6.82$, $p = 0.000$, Cohen's $d = 0.60$). There was no difference across the repetition break condition and the baseline condition ($t = 1.52$, $p = 0.13$, Cohen's $d = 0.13$)[1].

Figure 1 shows the ratings for similes in each of the four positions, and once again there were no effects or interactions due to the particular targets or bases. As we predicted, creativity and perspective-change ratings declined from position 1 to position 4 in the consistent condition. On the contrary, in the repetition break condition we observed a decline over the first three similes but a jump in the last one. Switching from the initial perspective to a different one restored the perceptions of creativity and perspective-change to a level comparable to that in the baseline condition.

Taken together, Study 3 provided evidence largely in support of our predictions that repeated exposure to the same perspective will result in reduced creativity perception, whereas breaking from one perspective to another one (i.e., a perspective-change) will lead to an increase in creativity perception to the baseline level. This is evidence in support of the general proposition that the experience of a perspective-change underlies creativity evaluation.

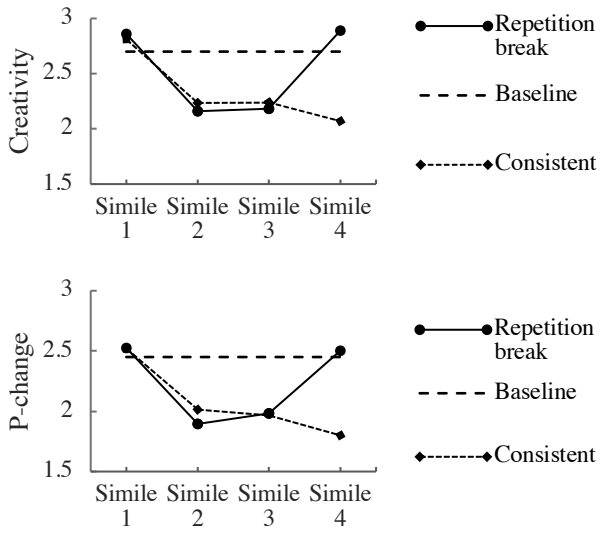


Figure 1: Results in Study 3

General Discussion

The proposal that some kind of change in perspective is involved in generating creativity has long drawn the attention of theorists, but there has been much less said about the process of evaluating creativity. We outlined and tested the beginnings of a perspective change account of creativity evaluation. Three studies found strong relationships between a simile prompting changes in perspective and evaluating the simile to be creative.

Critical to the account was specifying that an item, such as a simile, is likely to be experienced as creative if in the process of comprehending it people experienced a change in perspective. We followed existing research on changes in representation within the cognitive science literature, particularly work on analogy and comparison, to provide plausible specifications of what a change in perspective might involve and what might make a change more and less compelling. Other approaches could also be useful. Our intent was not to delve into accounts of mental representation

[1] Inclusion of the 14 participants who failed the attention check yielded marginally significant difference in perspective-change across the repetition break condition (M = 2.68, *SD* = 1.19) and the baseline condition (M = 2.51, *SD* = 1.20, $t = 1.15$, $p = 0.09$, Cohen's $d = 0.14$).

but rather to emphasize why doing so could advance research on creativity evaluations.

The current results offer initial support for the value in thinking about perspective change as a driver of creativity evaluations. Using both spontaneous (Study 1) and induced (Study 2) changes in perspective, we found consistent evidence that a simile is likely to be judged creative if it is experienced as a new way to interpret the target. Further, we also found (Study 3) that repeated use of the same perspective suppresses the experience of perspective change and thus reduces creativity perception. Taken together, these studies indicate that the process of forming an interpretation of an item, and the kind of interpretation we form, influences our evaluation of its creativity.

Focusing on the cognitive process of forming and changing perspectives opens a new area for research on creativity evaluation. Most work has focused on whether an item is novel and useful for a community or domain (e.g., Amabile, 1983, 1988, 1996; Cropley & Cropley, 2010; Oldham & Baer, 2012; Runco & Jaeger, 2012; Shalley, Zhou, & Oldham, 2004; Sternberg & Lubart, 1999; Woodman, Sawyer, & Griffin, 1993). The current work shifts the focus of creativity judgments from the product to the process—from the characteristics (i.e., novelty and usefulness) of the item to the work of making sense of the item and the perspective that results.

In emphasizing the forming and changing of perspectives, this work opens up new ways to think about the role of expertise and culture in shaping creativity evaluations. There might be expertise needed to appreciate an item as creative, as absent that expertise on might not change perspective or perceive the change to have much potential. There might be cultural assumptions that resonate or impede the change in perspective. These are cognitive issues, and they also lead to considerations around attitudes and values. Creativity evaluations are, after all, judgments about worth.

Developing an account of creativity evaluations resting on perspective change provides an opening. The considerable amount of research on knowledge and knowledge change in cognitive science is not always or even usually at the center of discussions about creativity. Yet it may hold significant potential to help advance such discussions. As researchers whose central task is innovation, we are aware of the imperfect evaluations generated by grant review panels, journals, and promotion review committees. It is clear that deepening our understanding of the creativity evaluation process is consequential. It is likely to be similarly consequential for all the other domains of innovation in our societies.

References

Amabile, T. M. (1983). The social psychology of creativity: A componential conceptualization. *Journal of personality and social psychology*, *45*(2), 357.

Amabile, T. M. (1988). A model of creativity and innovation in organizations. *Research in organizational behavior*, *10*(1), 123-167.

Amabile, T. M. 1996. Creativity in context. Boulder, CO: Westview

Boden, M. A. (2004). *The creative mind: Myths and mechanisms*. Routledge.

Boring, E. G. (1950). Great men and scientific progress. *Proceedings of the American Philosophical Society*, *94*(4), 339-351.

Bowdle, B. F., & Gentner, D. (2005). The career of metaphor. *Psychological Review*, *112*(1), 193-216.

Chi, M. T. (2009). Three types of conceptual change: Belief revision, mental model transformation, and categorical shift. In *International handbook of research on conceptual change* (pp. 89-110). Routledge.

Cronin, M. A., & Loewenstein, J. (2018). *The Craft of Creativity*. Stanford University Press.

Cropley, D. H., & Cropley, A. J. (2010), Functional creativity: Products and the generation of effective novelty. In J. C. Kaufman & R. J. Sternberg (Eds.), Cambridge Handbook of Creativity (pp. 301-320). New York: Cambridge University Press.

Day, S. B., & Asmuth, J. (2017). Re-representation in comparison and similarity. *Proceedings of the Cognitive Science Society*.

Duncker, K., & Lees, L. S. (1945). On problem-solving. *Psychological monographs*, *58*(5), i.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods, 39*(2), 175-191.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, *7*, 155-170.

Gentner, D. (1989). The mechanisms of analogical learning. In S. Vosniadou & A. Ortony (Eds.), Similarity and analogical reasoning (pp. 199-241). London: Cambridge University Press.

Gentner, D., & Wolff, P. (2000). Metaphor and knowledge change. *Cognitive dynamics: Conceptual change in humans and machines*, 295-342.

Loewenstein, J. (2019). Surprise, recipes for surprise, and social influence. *Topics in cognitive science*, *11*(1), 178-193.

Loewenstein, J., & Heath, C. (2009). The Repetition-Break plot structure: A cognitive influence on selection in the marketplace of ideas. *Cognitive science, 33*(1), 1-19.

Loewenstein, J., Raghunathan, R., & Heath, C. (2011). The repetition break plot structure makes effective television advertisements. *Journal of Marketing*, *75*(5), 105-119.

Luchins, A. S. (1960). On some aspects of the creativity problem in thinking. *Annals of the New York Academy of Sciences*, *91*(1), 128–140. https://doi.org/10.1111/j.1749-6632.1961.tb50921.x

Mumford, M. D., & Gustafson, S. B. (1988). Creativity syndrome: Integration, application, and innovation. *Psychological bulletin, 103*(1), 27.

Oldham, G. R., & Baer, M. (2012). Creativity and the work context. *Handbook of organizational creativity*, 387-420.

Page, S. (2008). The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies. Princeton: Princeton University Press, 2007. 448p. *Perspectives on Politics*, *6*(4), 828-829.

Rankin, C. H., Abrams, T., Barry, R. J., Bhatnagar, S., Clayton, D. F., Colombo, J., ... & McSweeney, F. K. (2009).

Habituation revisited: an updated and revised description of the behavioral characteristics of habituation. *Neurobiology of learning and memory*, *92*(2), 135-138.

Runco, M. A., & Jaeger, G. J. (2012). The standard definition of creativity. *Creativity Research Journal*, *24*(1), 92-96.

Sakamoto, Y., & Love, B. C. (2004). Schematic influences on category learning and recognition memory. *Journal of Experimental Psychology: General*, *133*(4), 534.

Shalley, C. E., Zhou, J., & Oldham, G. R. (2004). The effects of personal and contextual characteristics on creativity: Where should we go from here?. *Journal of management*, *30*(6), 933-958.

Sternberg, R. J., & Lubart, T. I. (1999). The concept of creativity: Prospects and paradigms. *Handbook of creativity*, *1*, 3-15.

Woodman, R. W., Sawyer, J. E., & Griffin, R. W. (1993). Toward a theory of organizational creativity. *Academy of management review*, *18*(2), 293-321.

Zhou, J., Wang, X.M., Bavato, D., Tasselli, S., & Wu, J.F. (2019). Understanding the receiving side of creativity: a multidisciplinary review and implications for management, *Journal of Management Annual Review Issue* (forthcoming)

Race and gender are automatically encoded in visual working memory

Xin Yang
Yale University, New Haven, Connecticut, United States

Joshua Langfus
John Hopkins University, Baltimore, Maryland, United States

Justiin Halberda
John Hopkins University, Baltimore, Maryland, United States

Yarrow Dunham
Yale University, New Haven, Connecticut, United States

Abstract

Research has suggested that perceivers automatically categorize faces based on gender and race but gaps remain regarding whether effects emerge at encoding or recall and the extent to which they are reducible to perceptual similarities (since faces from the same category are generally more similar to each other). We address these limitations using change detection paradigms adapted from visual working memory literature where one face from an array of faces changes to a face from the same or a different gender or racial category. We show that individuals are considerably faster and more accurate to identify changes that cross a category boundary, even when controlling for a range of perceptual differences and subjective features of faces. Our results suggest that social category information is automatically encoded in visual working memory in a format that is not reducible to lower-level perceptual features.

The Effect for Category Learning on Recognition Memory: A Signal Detection Theory Analysis

Siyuan Yin[1,2,*], Kevin O'Neill[3,*], Timothy F. Brady[4], Felipe De Brigard[1,2,3,5]

[1]Duke Institute for Brain Sciences, Duke University, Durham, NC 27708, USA.
[2]Department of Philosophy, Duke University, Durham, NC 27708, USA.
[3]Center for Cognitive Neuroscience, Duke University, Durham, NC 27708, USA.
[4]Department of Psychology, University of California, San Diego, CA 92093, USA.
[5]Department of Psychology and Neuroscience, Duke University, Durham, NC 27708, USA.

Abstract

Previous studies have shown that category learning affects subsequent recognition memory. However, questions remain as to how category learning affects discriminability during recognition. In this three-stage study, we employed sets of simulated flowers with category- and non-category-inclusion features appearing with equal probabilities. In the learning stage, participants were asked to categorize flowers by identifying the category-inclusion feature. Next, in the studying stage, participants memorized a new set of flowers, a third of which belonged to the learned category. Finally, in the testing stage, participants received a recognition test with old and new flowers, some from the learned category, some from a not-learned category, some from both categories, and some from neither category. We applied hierarchical Bayesian signal detection theory models to recognition performance and found that prior category learning affected both discriminability as well as criterion bias. That is, people that learned the category well, exhibited improved discriminability and a shifted bias toward flowers from the learned relative to the not learned category.

Keywords: category learning; recognition memory; signal detection theory; Bayesian modeling

Introduction

Memory research has shown that prior learning experience affects recognition memory. It is often thought that prior learning is encoded into knowledge structures or *schemas* (Bartlett, 1932). In turn, schemas increase recognition of schema-inconsistent information compared to schema-consistent information, while also increasing false alarms to schema-consistent lures compared to schema-inconsistent lures. Because schema acquisition takes time and learning experiences vary among people, most recognition memory tasks have employed either within-subject designs for pre-acquired schemas (Graesser & Nakamura, 1982) or between-subject designs for individuals with different expertise (Castel et al, 2007). As such, traditional experimental designs do not easily allow manipulation of schema acquisition in a way that enables us to assess their effect on recognition memory performance.

A number of recent studies have unveiled strong connections between schematic and categorical knowledge, leading many researchers to postulate profound similarities in the cognitive processes underlying schematic and categorical learning (Sakamoto & Love, 2004). To contribute to the integration of schematic and categorical learning, and to further explore the effects of prior learning on recognition memory, De Brigard et al. (2017) recently employed a set of computer-generated stimuli (flowers) to explore how learning a novel category affects participants' recognition memory for items from the learned category relative to items from a category they did not learn. However, the studies reported by De Brigard et al. (2017) left several unanswered questions. In particular, the findings could not differentiate between discriminability changes for items from the learned category and a change in response bias because their experiments did not include foils of both learned and not-learned categories, and thus could not provide measures of discriminability and bias for all options. In addition, De Brigard et al.'s (2017) findings did not discriminate between those who learned best and those who learned least during the category-learning phase, potentially obscuring effects on discriminability in recognition memory.

To explore these issues, in the present study we used a modified version of De Brigard et al.'s (2017) paradigm in which flowers from learned and non-learned categories appeared in the learning and study phases with equal probability. Additionally, the current study included lures from both learned and not-learned categories during the recognition test. As such, we were able to implement full hierarchical Bayesian signal detection theory (SDT) models to data from all participants, as well as separate people by the strength of their learning. This modified experimental paradigm, and the SDT models with which the results are analyzed, enables us to further understand the effect of category learning on recognition memory.

Category Learning and Recognition Experiment

Participants

113 individuals participated via Amazon Mechanical Turk (https://www.mturk.com) for monetary compensation. All participants were from the United States and had at least 100 approved hits and overall hit rate $\geq$ 95%. Three participants were excluded because of failure to follow instructions or terminated the experiment in the middle, so data were analyzed with the remaining 110 individuals. All participants

were provided informed consent under a protocol approved by the Duke University IRB.

Materials

Stimulus consisted of MATLAB (2018b)-generated flowers from De Brigard et al. (2017). Flowers varied across five dimensions, with each dimension taking one of three possible values: number of petals (4, 6, or 8), color of petals (blue, green, or yellow), shape of center (circle, triangle, and square), color of center (orange, purple, or turquoise), and number of sepals (1, 2, or 3). Figure 1 illustrates three examples of flowers with different combinations of the features (see further details in De Brigard et al., 2017).

Procedure

We closely followed the procedure from the fourth experiment in De Brigard et al. (2017), with some modifications (see below). The experiment had three phases: learning, study, and test. At the beginning of each phase participants read the instructions for 90s.

Figure 1. Examples of MATALB-generated flowers. From left to right: 4 blue petals- orange circle center -1 sepal; 6 green petals- purple triangle center-2 sepals; and 8 yellow petals- blue square center-3 sepals. See more in De Brigard et al., 2017.

In the learning phase, participants were told they would see a flower on the screen and will have to determine whether or not it belonged to the species *avlonia*. Participants were told that avlonias differed from other flowers in one simple way (e.g., only avlonias have four petals), and their task was to find out what the simple way was. At the beginning of the learning phase, participants were informed of all five possible dimensions—number of petals, color of petals, etc.—across which flowers may vary and saw two example flowers for illustration. They then made binary choices "yes" or "no" on each trial to categorize each flower by pressing "y" or "n", respectively, and there were 54 trials in total. Immediately after their responses, feedback with the word "Correct" or "Incorrect" was displayed. Participants were ensured that they could guess at the beginning but eventually they would find out the simple way that made a flower an avlonia. Each participant was assigned to a category-inclusion feature consisting of one possible value from one of the five dimensions. Additionally, participants were also assigned a "Not-learned" category, defined by a value of a different dimension, of which participants were never informed or given feedback. Both of these assignments were counterbalanced across participants. In all phases of the experiment, all values of all stimulus features did not differ in their statistical properties, such that flowers having the learned feature (i.e., that were avlonias) appeared on one-third of the trials, while the other two-thirds of the trials included flowers displaying the other two values of the Learned category-inclusion feature. Likewise, one-third of the trials presented flowers in the Not-learned category, while the other two-thirds of the trials included flowers with the other two values of the Not-learned category-inclusion feature. Importantly, the category-inclusion features for the Learned and Not-learned categories were independent, such that one-ninth of all flowers were in both the Learned and the Not-learned categories (Both condition), two-ninths of all flowers were in the Learned category but not the Not-Learned category (Learned condition), two-ninths of all flowers were in the Not-learned category but not the Learned category (Not-learned condition), and four-ninths of all flowers were in neither the Learned nor the Not-learned category (Neither condition). Table 1 summarizes the distribution of values for the Learned and Not-Learned category-inclusion features.

In the study phase, participants were asked to memorize 18 flowers. Each flower was shown alone for 5s followed by a 1s blank. Of the 18 flowers, four were in the Learned category but not the Not-Learned category (Learned), four were in the Not-learned category but not the Learned category (Not-Learned), two were in both categories (Both), and eight were members of neither category (Neither). To incentivize memorization, participants were told that they would receive an extra bonus for remembering above 85% of the stimuli. None of these 18 flowers were presented during the learning phase (Table 1).

Finally, in the testing phase, participants were told that they would see 54 flowers, one on each trial, and that their task was to remember whether or not stimuli were shown before in the study phase by pressing "yes" or "no". Of the 54 flowers, 18 were *old*—i.e. were presented in the studying phase—while the remaining 36 were *new*. Of these new flowers, four were from the Learned category only, four were from the Not-learned category only, two were from Both, and eight were from Neither. Of note, these new flowers were not shown during the study phase. All flowers were presented randomly and each trial was self-paced.

In sum, there were four types of trials in these three phases. Table 1 illustrates some possible combinations of Learned and Not-learned features. For each subject, one-third of trials included the learned category inclusion feature, which was chosen randomly from the three possible values from one of the five dimensions. Orthogonally, one-third of the trials included a not-learned category inclusion feature, i.e., the value of a dimension that could define a category of which participants were not aware of. This not-learned category inclusion feature was chosen randomly from the values belonging to the remaining four dimensions different from the dimension with the learned category inclusion feature. Membership in the Learned and Not-learned categories was independent of one another.

Table 1: Examples of possible combinations of Learned and Not-learned feature. Each row indicates one possible combination for one participant. A_1, A_2 and A_3 indicate three

possible values (denoted by 1, 2, and 3) of one randomly selected dimension out of five dimensions (denoted by A, B, C, D, and E; here we use only A and B for illustration purpose) -- number of petals, color of petals, shape of center, color of center, and number of sepals. B_1, B_2 and B_3 indicate three possible values of another randomly selected dimension out of the remaining four dimensions. Both condition has learned category inclusion feature and not-learned category inclusion feature features, and Neither condition does not have learned category inclusion feature or not-learned category inclusion feature features. The number of trials shown in the table is for learning and testing phases only. The number of trials for each feature during the study phase is 2 (not shown).

Learned feature	Not-learned feature	Number of trials	Probabilities
A_1	B_1	6	1/9
A_1	B_2	6	1/9
A_1	B_3	6	1/9
A_2	B_1	6	1/9
A_2	B_2	6	1/9
A_2	B_3	6	1/9
A_3	B_1	6	1/9
A_3	B_2	6	1/9
A_3	B_3	6	1/9

Results

Learning. We measured the learning performance by calculating the percentage of correct responses in the learning phase (Figure 2). We found participants were, in general, able to detect the single feature that categorized avlonias. The overall accuracy rates for both stimuli during the last twenty trials were 82.3%. Note that because we do not inform participants of the feature in advance, they necessarily begin at 50% accuracy at the beginning of the learning phase.

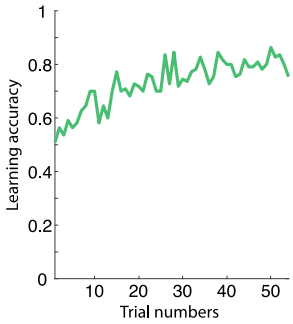


Figure 2. Learning performance during learning phase.

Memory Accuracy. We analyzed hit and false alarm (FA) rates separately for flowers of each type (Figure 3). To examine the learning effects for the four conditions (i.e., flowers that belong to the Learned category, Not-learned category, Both categories, Neither category), we implemented a two-way Bayesian repeated measures ANOVA. People exhibited increased hit rates for stimuli containing learned features included in Learned (M_{Hit} = 0.65, SD_{Hit} = 0.28) and Both (M_{Hit} = 0.70, SD_{Hit} = 0.35) conditions during the testing phase, but not toward stimuli not including those features in Not-learned (M_{Hit} = 0.56, SD_{Hit} = 0.28) and Neither (M_{Hit} = 0.58, SD_{Hit} = 0.22) conditions (Figure 2 and Table 2). We followed up with Bayesian paired samples t-tests which showed evidence supporting that hit rates in the Learned condition were higher than those in the Not-learned (BF_{10} = 2.15) and Neither conditions (BF_{10} = 1.70), but not in the Both condition (BF_{10} = 0.25) (See the scale of evidence in Jeffreys, 1998). Similarly, there was evidence indicating that hit rates for the Both condition were higher than those in the Not-learned (BF_{10} = 135.38) and Neither condition (BF_{10} = 30.56). Hit rates in the Not-learned condition were not different from the Neither condition (BF_{10} = 0.14). Also, there was weak evidence for FA rates in the Learned condition (M_{FA} = 0.59, SD_{FA} = 0.24) being higher than for the Not-learnsed (M_{FA} = 0.53, SD_{FA} = 0.23; BF_{10} = 0.43) and Neither condition (M_{FA} = 0.53, SD_{FA} = 0.19; BF_{10} = 0.61). We found no evidence for differences in other pairs of conditions (Both condition: M_{FA} = 0.57, SD_{FA} = 0.30).

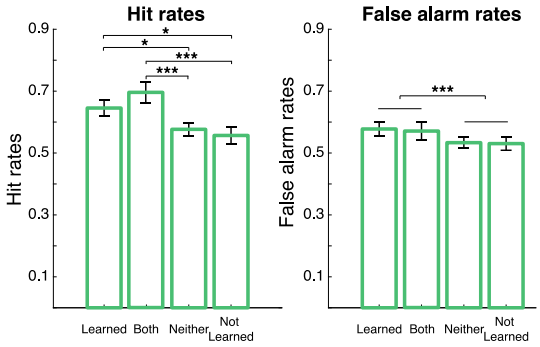


Figure 3. Hit and false alarm rates during testing phase. Left panel: hit rates. Right panel: false alarm rates. Stimulus with the category-inclusion value appeared in the Learned and Both conditions, and not in the Neither and Not-Learned conditions. * BF_{10} > 1, *** BF_{10} > 10.

Table 2: Bayesian repeated measures ANOVA

Rates	Best models	BF_{Model}	BF_{10}
Hit	Learned	11.02	6.93×10^3
False Alarm	Learned	13.93	32.35

BF: Bayes Factor.

We followed up with Bayesian paired samples t-tests which showed evidence supporting that hit rates in the Learned condition were higher than those in the Not-learned (BF_{10} = 2.15) and Neither conditions (BF_{10} = 1.70), but not in the Both condition (BF_{10} = 0.25) (See the scale of evidence in Jeffreys, 1998). Similarly, there was evidence indicating that hit rates for the Both condition were higher than those in the Not-learned (BF_{10} = 135.38) and Neither condition (BF_{10} = 30.56). Hit rates in the Not-learned condition were not different from the Neither condition (BF_{10} = 0.14). Also, there was weak evidence for FA rates in the Learned condition (M_{FA} = 0.59, SD_{FA} = 0.24) being higher than for the

Not-learned (M_{FA} = 0.53, SD_{FA} = 0.23; BF_{10} = 0.43) and Neither condition (M_{FA} = 0.53, SD_{FA} = 0.19; BF_{10} = 0.61). We found no evidence for differences in other pairs of conditions (Both condition: M_{FA} = 0.57, SD_{FA} = 0.30).

To explore the effect of category learning separately on response bias and discriminability (e.g., d'), we conducted a hierarchical Bayesian parameter estimation analysis within a SDT framework. To that end, we fit the accuracy data from three groups, i.e., (1) *all* participants (n = 110), (2) *experts*, i.e., participants whose accuracy of the last twenty learning trials was greater than or equal to 80% (n = 66), and (3) *non-experts*, i.e., participants whose accuracy of the last twenty learning trials was less than 80% (n = 44), to a SDT model in which the parameters were estimated using a hierarchical Bayesian approach (Lee, 2008). As such, two parameters of discriminability were estimated: (1) the sensitivity, d', that is measured by the distance between the signal and noise distributions indicating the discriminability of the signal trials from the noise trials; and (2) the criterion or bias, c, that is measured by the distance between the actual criterion used for responding and the unbiased criterion (i.e., $d'/2$).

The hierarchical model of SDT is able to partially pool individual parameters by taking into account group-level distributions, thus yielding more reliable estimates than non-hierarchical, full individual difference models. In this model, individual parameters are drawn from group-level (normal) distributions with estimated means and standard deviations. The model assumes that the estimated means quantify discriminability and criterion-bias for each of the four conditions, and precision quantifies the similarity among individual behavior.

In this implementation, our SDT model has four parameters per condition, reflecting properties of the average subject and how the subjects vary: mean discriminability μ_d, precision of discriminability τ_d, mean criterion μ_c, and precision of criterion τ_c. The prior on the mean discriminability was set to be very wide so as to be uninformative over the range of reasonable d' values (i.e., 0-4), with only a slight pull toward 0, consistent with previous research. Specifically, individual d_i was drawn from a normal distribution with mean and precision $\mu^d \sim N(0, 0.001)$ and $\tau^d \sim Gamma(0.001, 0.001)$, respectively. Individual c_i was then drawn from the normal distribution with two group-level parameters $\mu^c \sim N(0, 0.001)$ and $\tau^c \sim Gamma(0.001, 0.001)$. We implemented the hierarchical SDT model in JAGS, a sampler that utilizes a version of the BUGS programming language (Version 3.3.0) called from MATLAB (The MathWorks, Inc., Natick, Massachusetts, United States). Posterior distributions were approximated by 3 Monte Carlo Markov Chain methods with 5000 samples from each chain, after a burn-in of 1000 samples. Convergence of chains was evaluated with the $\hat{R}$ statistic.

We first estimated the mean sensitivity and mean criterion-bias for each condition by calculating the posterior distributions of hit and FA rates for all participants--group (1). We found that in the Learned condition, this was skewed toward 1 for both hit and false alarm rates, significantly above the other three conditions (Figure 4A), indicating the people had both more hits and more false alarms in this condition. Furthermore, for participants from group (2, expert-learners), hit and FA rates in both Learned and Both conditions were skewed toward 1, significantly above than those under Not-learned and Neither conditions (Figure 4B), whereas for participants from group (3, non-expert-learners) there were no differences (Figure 4C), suggesting the main effect was driven by the expert-learners.

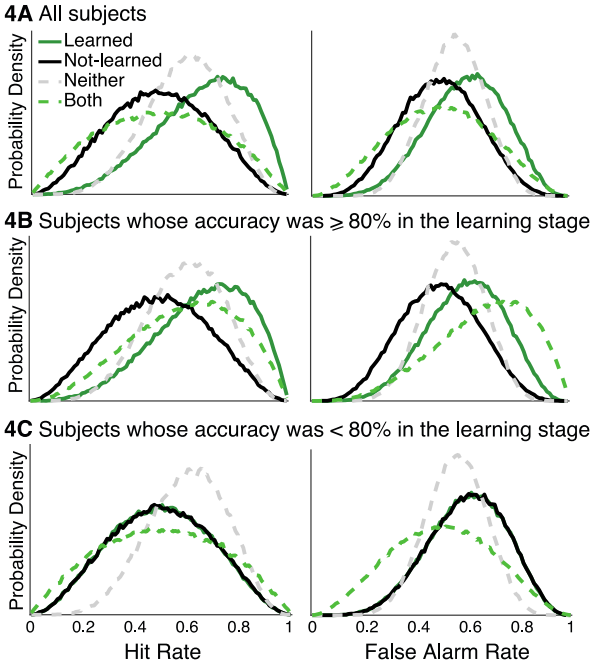


Figure 4. Posterior distribution of hit and FA rates for each of the four conditions.

To further investigate differences in discriminability/bias we performed a two-way Bayesian repeated measures ANOVA on estimated individual sensitivity and criterion-bias measures from each group. For individual sensitivity/d' of all subjects (1), we found main effects for Learned and Not-learned categories as well as their interaction, while for criterion-biases, we only found a main effect for the Learned category. For group (2, expert-learners), we found a significant main effect for the Learned category and a significant interaction between the Learned and Not-learned categories for both sensitivity and criterion-bias measures. For individual sensitivity of group (3, non-expert-learners), we found main effects for Learned and Not-learned categories as well as their interaction, while for criterion-biases we did not observe main effects of categories or their interaction (Table 3). These results indicate that participants who clearly excelled at learning the category during the learning stage—which we here operationalize as those participants whose accuracy for the last twenty trials was above 80%—were more sensitive to (i.e., increased discriminability/d') other flowers in this category and also and tended to say 'old' more often for these in general (i.e.,

Learned and Both conditions) compared to flowers not in the category (i.e., Not-learned and Neither conditions).

Follow-up Bayesian paired sample t-tests on sensitivity and criterion-bias for participants from group (1)—i.e. all participants—showed decisive evidence supporting that the sensitivity d_i for the Learned condition ($d_{Learned} = 0.18 \pm 0.05$) was higher than for the other three conditions ($d_{Not-learned} = 0.07 \pm 0.11$, $d_{Neither} = 0.12 \pm 0.03$, $d_{Both} = 0.38 \pm 0.18$), while the sensitivity d_i for Both was higher than the Not-learned and Neither conditions. As for the criterion-bias c_i, the evidence was also decisive supporting that the bias c_i for the Learned condition ($c_{Learned} = -0.32 \pm 0.38$) was lower than for the other three conditions ($c_{Not-learned} = -0.12 \pm 0.23$, $c_{Neither} = -0.15 \pm 0.34$, $c_{Both} = -0.39 \pm 0.28$), while the criterion-bias c_i for Both was lower than the Not-learned and Neither conditions. No strong evidence supported any differences between Not-learned and Neither conditions for both sensitivity d_i and bias c_i. (Table 4A).

For participants from group 2, this analysis revealed strong evidence that support differences in sensitivity d_i ($d_{Learned} = 0.290 \pm 0.048$, $d_{Not-learned} = 0.119 \pm 0.107$, $d_{Neither} = 0.126 \pm 0.084$, $d_{Both} = 0.339 \pm 0.083$) in almost all pairwise contrasts except Not-learned versus Neither. The same trend was also found in bias c_i ($c_{Learned} = -0.360 \pm 0.359$, $c_{Not-learned} = -0.073 \pm 0.262$, $c_{Neither} = -0.161 \pm 0.365$, and $c_{Both} = -0.543 \pm 0.191$). These results suggest that participants who mastered the learned features well in the learning stage were overall more sensitive to flowers with those features.

For participants from group 3, the sensitivity d_i of Learned condition ($d_{Learned} = 0.069 \pm 0.077$) was higher than those of Not-learned condition ($d_{Unlearned} = 0.003 \pm 0.080$) and lower than those of Both condition ($d_{Both} = 0.391 \pm 0.413$), but not different from those of Neither condition ($d_{Neither} = 0.096 \pm 0.019$). The sensitivity d_i of Both condition were higher than those of the other conditions, and the d_i of Not-learned condition were lower than those of Neither condition. As for biases c_i, the Bayesian paired t test did not show strong evidence supporting any differences between pairs of conditions ($c_{Learned} = -0.264 \pm 0.415$, $c_{Unlearned} = -0.183 \pm 0.155$, $c_{Neither} = -0.135 \pm 0.294$, and $c_{Both} = -0.132 \pm 0.176$), except moderate evidence suggesting differences in c_i between Learned and Neither conditions as well as Learned and Both conditions.

Figure 5 illustrates the joint posterior distributions of discriminability and bias for each condition. The main panel shows 15000 samples form the joint posterior of the mean μ^d and μ^c. The side panels show the marginal distribution for each of the group-level means. For all subjects, the group-level sensitivity d_i differed the most between Both and Not-learned conditions, and the group-level biases c_i were negative in Both and Learned conditions. That is, participants exhibited better sensitivity toward flowers with learned features and a tendency to overrespond "yes" in the recognition memory tasks (Figure 5A).

Table 3: Bayesian repeated measures ANOVA

Datasets	SDT parameters	Best models	BF_{Model}	BF_{10}
All subjects	d_i	Learned+Not-learned+Learned×Not-learned	5.39×10^{26}	1.43×10^{73}
	c_i	Learned	12.04	4.17×10^{13}
Experts	d_i	Learned+Not-learned+Learned× Not-learned	25.97	1.27×10^{51}
	c_i	Learned+Not-learned+Learned× Not-learned	884.46	1.27×10^{19}
Non-Experts	d_i	Learned+Not-learned+Learned× Not-learned	1.40×10^{8}	1.19×10^{13}
	c_i	Null model	4.26	1.00

d_i and c_i are individual sensitivity and biases estimated by the hierarchical Bayesian parameter estimation.

Table 4A: Bayesian paired sample t test for sensitivity and bias with all subjects. Numbers shown in the table indicate Bayes Factors.

Category comparison	Sensitivity d_i	Bias c_i
Learned vs. Not-learned	4.19×10^{12}	6.04×10^{4}
Learned vs. Neither	1.73×10^{23}	9.03×10^{2}
Learned vs. Both	8.77×10^{16}	0.773
Not-learned vs. Neither	3.24×10^{2}	0.183
Not-learned vs. Both	1.53×10^{27}	3.87×10^{10}
Neither vs. Both	8.72×10^{25}	2.03×10^{6}

Table 4B: Bayesian paired sample t test for sensitivity and bias with only subjects whose accuracy of the last twenty learning trials was above or equal to 80% (i.e., experts). Numbers shown in the table indicate Bayes Factors.

Category comparison	Sensitivity d_i	Bias c_i
Learned vs. Not-learned	2.37×10^{14}	1.75×10^{5}
Learned vs. Neither	7.29×10^{22}	1.48×10^{2}
Learned vs. Both	3.06×10^{2}	1.04×10^{3}
Not-learned vs. Neither	0.154	0.810
Not-learned vs. Both	8.56×10^{19}	3.25×10^{18}
Neither vs. Both	3.48×10^{19}	3.85×10^{8}

Table 4C: Bayesian paired sample t test for sensitivity and bias with only subjects whose accuracy of the last twenty learning trials was less than 80% (i.e., non-experts). Numbers shown in the table indicate Bayes Factors.

Category comparison	Sensitivity d_i	Bias c_i
Learned vs. Not-learned	1.15×10^{2}	0.411
Learned vs. Neither	1.35	1.63
Learned vs. Both	2.64×10^{3}	1.24
Not-learned vs. Neither	4.26×10^{6}	0.30
Not-learned vs. Both	4.92×10^{4}	0.41
Neither vs. Both	9.94×10^{2}	0.16

For subjects whose accuracy in the last twenty learning trials was greater than or equal to 80%, the difference in the group-level sensitivity d_i between Learned and Both conditions was less but the difference between Learned and Not-learned or Neither were greater. The group-level biases c_i in Learned and Both conditions were more negative than those in Not-learned and Neither conditions (Figure 5B). The results suggested that participants who learned category-relevant features well had better discriminability and stronger biases toward flowers with learned features. For subjects whose accuracy in the last twenty learning trials was less than 80%, the group-level sensitivity d_i differed the most between Both and Not-learned conditions, whereas the group-level biases c_i became closer to each other across conditions (Figure 5C). The results indicated that participants who did not learn the category-relevant features well had worse discriminability and little biases toward flowers with learned features.

Discussion

In this study we measured the extent to which learning novel categories influences recognition memory, and we focused on sensitivity and biases estimated in Bayesian SDT modeling. First, we corroborated previous findings that people exhibited biases toward stimuli within a learned category compared to stimuli not in the category, even when the relevant features are equally sampled during learning and study (De Brigard et al., 2017). That is, hit rates of stimuli with learned features (i.e., Learned and Both trials) were higher than stimuli with other values for that feature (i.e., Not-learned and Neither trials) (Figure 2). False alarm rates showed the same pattern. Going beyond this, we first fit full Bayesian SDT models and compared two measures of discriminability—sensitivity and criterion-bias—in four conditions. We observed that experts exhibited greater sensitivity and more negative criterion-bias than non-experts. We found greater discriminability for Learned and Both conditions than Not-learned and Neither conditions, which suggested people formed better memories of studied flowers with learned features. It is also clear that there was a response bias for Learned and Both conditions (Figure 4), indicating a tendency to overrespond "yes" (i.e., the flower was shown in the study stage) for these, in addition to the actual improved memory sensitivity. These results suggest that category learning affected recognition memory, improving discriminability as well as affecting response bias.

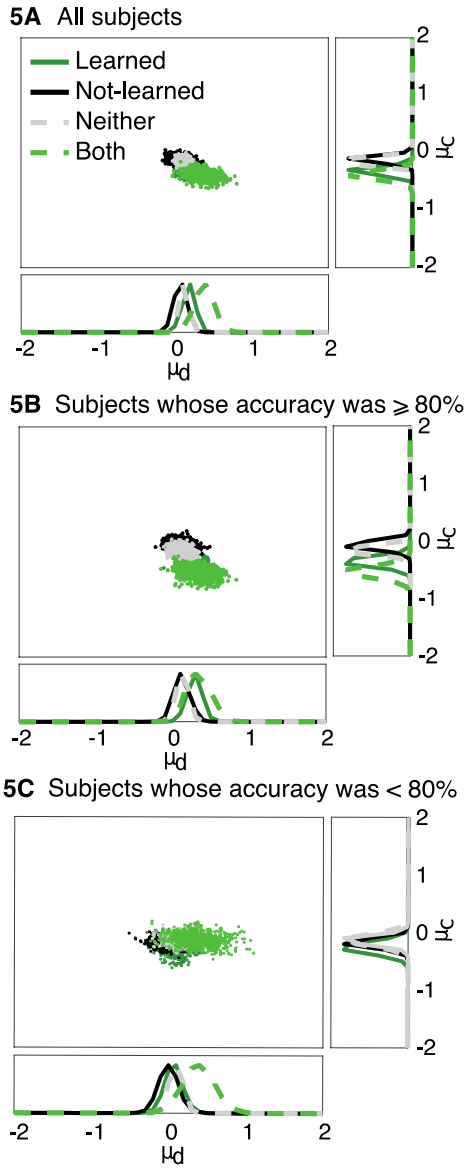


Figure 5: the joint distribution of mean discriminability d and mean bias c. The side panels show the corresponding marginal distribution. μ_d and μ_c are the group-level means of discriminability and criterion.

In the current study we employed a yes/no learning strategy to create new categories for novel stimuli and ask how they influence subsequent recognition memory. Our findings of the influence of category information on recognition memory are consistent with findings that show the influence of existing categories (Bae et al., 2015; Persaud & Hemmer, 2016) as well as newly learned episodic information about a category (Brady et al., 2018) on continuous recall measures. This suggests more insight into the influence of newly learned categories on memory looking at the effect of novel category learning on recognition memory employing

continuous measures. In addition, future studies may investigate whether different learning strategies may elicit the same biases. For example, supervised (i.e., with explicit guidance on category-inclusion criteria) and unsupervised (i.e., without explicit guidance), or active (i.e., trying to learn category-inclusion criteria with instant feedback) and passive (i.e., merely observing stimuli and their corresponding categories) learning processes may largely change biases toward stimuli with learned features.

It is worth noting that, in the current study, we used a somewhat arbitrary threshold to classify expert and non-expert learners. Future studies may apply Bayesian analyses to explore individual differences in learning and compare estimates of individual learning rates to individual recognition memory effects. This could provide a better characterization of the data rather than a binary division.

Previous studies have mainly focused on category learning and memory during the course of an experiment, but how these categories are acquired is also critical in this processing. In this study, we used a set of well controlled stimuli – computer simulated flowers – so that we can manipulate the degree of exposure of different features and reveal how the learning process affects recognition memory. In future work, it would be useful to adopt more naturalistic stimuli to examine the mechanisms of category learning in real world settings and how this varies as a function of context and with different age groups.

We applied SDT models to measure the effect of category learning on recognition memory. This effect may also be related to different learning procedures: for example, explicit reasoning and the nature and timing of feedback, which may or may not be directly associated with the learned feature only (Ashby & Maddox, 2005). Other categorization models such as the generalized context model (GCM, Nosofsky, 1986), the general recognition theory (GRT, Ashby & Townsend, 1986), or the deterministic exemplar model (DEM, Ashby & Maddox, 1993) will be worth exploring to make more refined quantitative accounts of the influence of category learning on recognition memory.

Finally, it is important to note that in the current study, participants were given binary choices in the testing stage (old/new). While this allowed us to apply signal detection models to probe the effect of category learning on recognition memory, to do so we needed to assume an equal variance signal detection model. Adopting a confidence scale and ROC analysis based on confidence rating data would provide a refined gauge of discriminability in the recognition memory task and allow us to measure the memory signal accurately, even in the case of unequal variance (as is common in recognition memory experiments). This would allow us to be more certain we had separately measured response bias and discriminability and address the nature of the memory signal more clearly (e.g., address whether unequal variance signal detection model, or a hybrid threshold and signal detection model is more applicable; Wixted, 2007). Broadly, however, our results show that participants discriminate more toward stimuli with learned features than those with not-learned features. These results contribute to our understanding of how prior category learning influences recognition memory.

Acknowledgments

This study is supported by a grant from the Office of Naval Research (N00014-17-1-2603) to FDB.

References

Ashby, F. G., & Townsend, J. T. (1986). Varieties of Perceptual Independence. *Psych. Rev., 93(2),* 154-179.

Ashby, F. G., & Maddox, W. T. (1993). Relations between Prototype, Examplar, and Decision Bound Models of Categorization. *J. Math. Psych., 37*, 372-400.

Ashby, F. G., & Maddox, W. T. (2005). Human Category Learning. *Ann. Rev. Psych., 56*, 149-178.

Bae, G.-Y., Olkkonen, M., Allred, S. R., & Flombaum, J. I. (2015). Why some colors appear more memorable than others: A model combining categories and particulars in color working memory. Journal of Experimental Psychology: General, 144(4), 744-763.

Brady, T. F., Schacter, D.L., and Alvarez, G.A. (2018). The adaptive nature of false memories is revealed by gist-based distortion of true memories. PsyArXiv Preprint. https://doi.org/10.31234/osf.io/zeg95

Castel, A. D., McCabe, D. P., Roediger, H. L., III, & Heitman, J. L. (2007). The dark side of expertise: Domain specific memory errors. *Psych. Sci.*, 18, 3–5.

De Brigard, F., Brady, T. F., Ruzic, L., & Schacter, D. L. (2017). Tracking the emergence of memories: A category-learning paradigm to explore schema-driven recognition. *Mem Cogn*, *45*, 105-120.

Graesser, A. C., & Nakamura, G. V. (1982). The impact of a schema on comprehension and memory. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 16, pp. 59–109). New York: Academic Press.

Jeffreys, H. (1998). The Theory of Probability (3rd ed.). Oxford, England.

Lee, M. D. (2008). Three case studies in the Bayesian analysis of cognitive models. *Psych. Bull. Rev., 15(1),* 1-15.

Nosofsky, R. M. (1986). Attention, Similarity, and the Identification-Categorization Relationship. *J. Exp. Psych.: General, 115(1)*, 39-57.

Persaud, K., & Hemmer, P. (2016). The dynamics of fidelity over the time course of long-term memory. *Cognitive Psychology, 88*, 1-21.

Sakamoto, Y., & Love, B. C. (2010). Learning and retention through predictive inference and classification. *J. Exp. Psych.: Applied*, 16, 361–377.

Wixted, J.T.. (2007) Dual-Process Theory and Signal-Detection Theory of Recognition Memory. *Psych. Rev., 114 (1)*, 152-76.

Yonelinas, A., Dobbins, I., Szymanski, M. D., Dhaliwal, H. S. & King L. (1996) Signal-Detection, Threshold, and Dual-Process Models of Recognition Memory: ROCs and Conscious Recollection. *Cons. Cog., 5*, 418-441.

The process of art-making: An analysis of artist's modification of conditions in the art-making process

Sawako Yokochi (yokochi-sawako@tokyomirai.jp)
School of Child Psychology, Tokyo Future University,
34-12, Senju Akebono-tyo, Adachi-ku, Tokyo, Japan 120-0023

Takeshi Okada (okadatak@p.u-tokyo.ac.jp)
Graduate School of Education, The University of Tokyo,
7-3-1, Hongo, Bunkyo-ku, Tokyo, Japan 113-0033

Abstract

The present study investigated how younger and expert artists create artwork, paying special attention to the modification of conditions in the art-making process. Here, "process modification" is the means by which artists generate new artistic ideas/concepts by modifying elements of one's own previous artwork. To examine whether younger artists use such modifications in the same manner as experts, we interviewed 28 contemporary artists (including 14 experts). Results revealed that most of the younger artists modified their work unsystematically. Younger artists drastically changed the subject/motif, method, and concept for their new artwork. Experts, in contrast, actively used process modification to create a new technique and generated a new concept based on their creative vision.

Keywords: artistic creation; creative process; art-making process; process modification

Introduction

How do people create artwork and become experts in this domain? Are there differences in the cognitive processes underlying artistic creation between younger artists and experts? The present study investigated potential experienced-related differences in artistic creation by analyzing retrospective interviews with artists during the early and progressive stage of an artistic creation.

Artistic Creation as Problem-Solving

Artistic creation can be regarded as a creative problem-solving process (Simon, 1973). In such a process, people search for goals, tools, and ways to make art in ill-defined problem space(s). To do so, they need to explore new frames in a problem space or generate a new problem space itself (i.e., problem finding). In such cases, it is difficult to use algorism or well-known heuristics, because goals and methods are not known in advance. Therefore, exploration becomes an essential process in such an endeavor. This study tries to describe the detailed process of such explorations adopted by artists during their long-term artistic activities.

Cognitive Process of Art-Making

Several studies have been conducted on human creativity, especially within the artistic realm (e.g., Getzels & Csikszentmihalyi, 1976; Mace & Ward, 2002; Okada, Yokochi, Ishibashi, & Ueda, 2009; Patrick, 1937; Stokes, 2014). For instance, Getzels and Csikszentmihalyi (1976) and Mace and Ward (2002) emphasized the importance of the "problem-finding process" within artistic creativity. In particular, Getzels and Csikszentmihalyi (1976) studied art majors' constructions of still life drawings. Their results indicated that students whose drawings were evaluated as highly creative were more exploratory in their work both before and during their drawings (i.e., arranging still life objects or changing tools more frequently). The authors also observed that after graduating from art school, the students who participated more in problem-finding activities were more successful in their careers. Thus, problem-finding could be a key activity for quality artistic creation.

Mace and Ward (2002) conducted interviews with artists to identify how they generated ideas during creation process. The authors developed a process model of art-making with the following four steps: "Artwork conception," "Idea development," "Making artwork and idea development," and "Finishing the artwork." While the process model is helpful, Mace and Ward could only describe these four steps; they did not assess potential underlying mechanism for progressing through these steps.

Through the using of art historical biographies, Stokes (2014) analyzed the process of artistic creation as problem-solving activity, revealing that paired constraints play an important role. Constraints have been regarded as having the function of both promotion and inhibition (e.g., Simon, 1973). Applying this idea to the artistic creation process, Stokes argued that precluding a constraint on the creative process (e.g., realism) helps promote another aspect of the paired constraint (e.g., abstraction). Using this framework to describe the creation process of famous artists such as Mondrian, Klee, Monet, and Chuck Close, Stokes suggested that the creative process proceeds as a cascading cycle until a new artwork is created.

Recently, Okada and colleagues have conducted research on the medium-term or long-term creative process of art-making, including the process of making a series of artwork based on certain artistic styles or themes/concepts (Okada, Yokochi, Ishibashi, & Ueda, 2009; Takagi, Okada, & Yokochi, 2013, Takagi, Kawase, Yokochi, & Okada, 2015; Yokochi & Okada, 2006, 2007). Of specific focus is how an artistic theme/concept is formed in an artist's mind or how

sub-themes/sub-concepts are derived from the main theme/concept from a cognitive psychological perspective (Okada, Yokochi, Ishibashi, & Ueda, 2009; Yokochi & Okada, 2006, 2007).

Yokochi & Okada (2007) revealed that artists develop expertise through several phases over the years. For instance, artists often construct a main theme, "creative vision," after about 12 years of practice post-art school. Creative vision is a somewhat abstract theme/concept, such as "life and death," "viewing/seeing," and "relationship with others," and is formed through long-term practice. The authors claimed that creative vision guides the construction of an artwork series in a certain direction, giving the artwork the consistency as a common base. Based on this creative vision, an artist finds suitable motifs/subjects and generates new artistic methods and creative ideas.

Okada et al. (2009) investigated the creation process of art concepts focusing on "analogical modification," which refers to cognitive processes tasked with generating new artistic ideas/concepts by analogically modifying elements of the artists' previous artwork. The authors claimed that 1) patterns of art concept formation gradually change as artists accumulate experiences; 2) artists use their creative vision for analogical modification of their art-making process; and 3) analogical modification enables artists to generate various artwork series, which are mutually connected with each other under the same creative vision. Takagi et al. (2013) also discovered, through ten months of qualitative and quantitative analyses of interviews with an artist, that the artist generated a new art concept for a new series. Here, the artist modified his creative process in multiple ways, including the modification of perception and action. We refer to these modifications in art-making process, including analogical modification and modification of perception and action, as "process modification" throughout this paper.

For the present study, we investigated how artists form their goals, art concepts, and creative vision, as well as how they develop methods for creating artwork series, paying special attention to "process modification." In terms of the development of artistic creative expertise, a creative vision, which is formed through many years of creative activity and consists of long-term intentions or goals for creation, serves as a framework to guide the process of creation (Yokochi & Okada, 2007). Because of such a creative vision, experts' creative process would be substantially different from young ones. Therefore, we also examined similarities and differences between younger and expert artists in terms of this concept formation process.

Methods

Participants: We interviewed 28 Japanese contemporary artists, comprising 14 younger artists, "YNG" (including 7 art major graduate students; 7 women, age range = 20-30 years, mean age = 28.3 years, mean work experience = 8.64 years, *SD* = 4.19), and 14 expert artists, "EXP" (4 women, age range = 40-60 years, mean age = 44.9years, mean work experience = 23.14 years, *SD* = 7.84). These artists have created various art forms, including paintings, sculptures, installations, photographs, and so on. All artists have participated in solo or group exhibitions every year, especially the expert artists, who have exhibited their work worldwide (including the USA and Europe). Those who participated were recommended by their peers, and in the case of the graduate students, they were nominated by their advisers.

Procedure (Portfolio-interview): The present study was conducted from 2005 to 2018. Because each artist's whole body of work was large in size, we interviewed each artist individually several times, using a portfolio of his/her entire work, which we referred to as a "portfolio-interview." The average interview time was 8 hours for YNG and 10 hours for EXP. This difference in interview time was because experts had a longer career and created more artwork than did younger artists. The portfolio-interview was conducted in a quiet room, which was either an art studio, home, or our university office. All conversations were recorded with IC recorder and a video camera.

The portfolio-interview was conducted as follows; First, we asked artists to explain each of their artwork pieces (e.g., "when and how was the artworks made?" "What kind of materials was used?" and "What was the idea/concept for the artwork?"). Second, we asked artists to identify what aspects of their work were kept and which were changed from prior work (e.g., "What (element) was changed from previous work?" and "What was a new or additional idea of this artwork?"). Finally, after explaining all of their work by reflecting on their entire career, the artists were asked whether they had their main art concept/theme (i.e., creative vision); if so, they were asked to report when they had realized this vision (e.g., "What is your main art concept/theme?" "When did you recognize the theme?" and "When were the turning points in your own art career?"). Additionally, we conducted semi-structured interviews as follows to gather information on: 1) originality in making and evaluating artwork (e.g., "What do you think about originality in your artwork?" and "Do you think it is important to represent originality in your work?"), 2) general process of making art (e.g., "How long do you usually spend on making/thinking about your work each day?" and "When and how do new ideas come up?"), and 3) educational background and biography of the artists.

Analysis procedure (Analysis of the process modification type and developmental trajectory of creation):

We analyzed the words used by the artists during the portfolio-interview and the features of their artwork. The coding framework was both theory and data driven. The categories for process modification included the categories for analogical modification (Okada, et al., 2009), and were guided by related theories regarding creativity and education/expertise, such as exploration (Boden, 2004), and reflection (Schön, 1983; Zimmerman, 2006). Further, the categories were inductively derived from the transcripts of the portfolio interview data, using the KJ method, which consists of a set of systematic procedures that seek to derive

a common (affinity) feature of data and ideas (Kawakita, 1967).

First, we specified the "main art concept and related sub art concept," "method and related methodology," and "motif (subject)" of each of artwork, and identified how each was changed from previous artwork. Second, we refined the categories for analogical modification (Okada et al., 2009) reflected in the interview data and features of works. Finally, the categories for process modification included and defined eight codes reflected in the interview data (see Table 1).

The interview data and all photographs of artwork were organized and stored using the computer package, MAXQDA, which is designed to organize unstructured data in qualitative and quantitative analyses. We developed the categories of process modification, and coded the portfolio interview data with the help of MAXQDA.

Results and Discussion

Following analyses based on the process modification categories, we examined distinctions between YNG and EXP artists, particularly comparing *before* finding a creative vision, "EXP_before," and *after* finding a creative vision, "EXP_after." Besides, we assessed how the artists generated new art concepts and series after realizing their creative vision.

Group comparisons in process modification types

Table 2 shows the number of artists who used each type of process modification, and the mean number of times YNG and EXP used each type (before and after realizing their creative vision), and artists who had their creative vision, "AwCV" (before and after realizing their creative vision), respectively.

Table 2 shows that there is little difference between the number of types between YNG and EXP_before groups in terms of "Subject modification" (YNG 79% vs. EXP_before 93%), "Structure modification" (29% vs. 36%), and "Concept modification" (14% vs. 0%). Although there is a subtle difference in "Unsystematic change" (57% vs. 64%), which refers to changing the art subject/motif, methods, and concepts from prior work, YNG tended to use "Searching for suitable subjects and methods" (86% vs. 50%) and "Subject modification with reconsideration of artistic methods" (73% vs. 43%) more often than the EXP_before

Table 1. Types of process modifications and definitions

Reference Frame for Modification	Modification Type	Definition
None	Type 0 No modification	Reproducing a previous work
Idea	Type 1_1 Unsystematic change	Changing both a previous motif, method, and concept without any specific goal (or sub-goal) e.g., changing all based on a temporal (casual) idea
	Type 1_2 Searching for suitable subjects and methods based on prior artistic ideas	Changing both motifs/subjects and methods to make artwork more suitable for the prior idea e.g., searching motifs and methods based on the idea for prior work
Methodology	Type 2_1 Quantitative modification	Changing size or material of previous work without changing subjects and concepts (becoming bigger/smaller size than previous work) e.g., changing the size of Mobiles
	Type 2_2 Subject modification	Changing motifs/subjects to make a new artwork by using the same methodology as for prior artwork e.g., applying Mobiles to various motifs
	Type 2_3 Subject modification with reconsideration of methods	Reconsidering the methodology while making new artwork by changing subjects and realizing availability/possibility of the methodology e.g., reconsidering availability of Mobiles methodology
Sub or Main art concept	Type 3 Structure modification	Generating a new methodology, in line with a sub art concept or a main art concept of artwork series e.g., generating "Mobiles" as a new methodology of sculpture
Creative vision	Type 4 Concept modification	Forming a main art concept and generating sub-concept (artwork series) according to a creative vision e.g., generating "Constellations" series based on Calder's main theme "Universe"

artists. Moreover, artists using "Subject modification" ended up in a stalemate/dead-end (21% vs. 43%).

Comparing the YNG and EXP_after conditions, although both used "Subject modification" (YNG 79% vs. EXP_after 100%), YNG tended to use more "Unsystematic change" (57% vs. 27%), "Searching for suitable subjects and methods" (86% vs. 8%). In contrast, EXP_after tended to use more "Structure modification" (29% vs. 73%) and "Concept modification" (14% vs. 82%). Furthermore, the number of EXP_after artists who experienced dead-end was reduced (21% vs. 9%).

A two-way factorial analysis of variance (mixed plan, factor 1: artists (3 levels, YNG, EXP_before, and EXP_after) × factor 2: types of process modification (8 levels)) was conducted on the number of times each artist group used the various process modification types. First, Mauchly's test of sphericity revealed a sphericity violation ($p < .01$); hence, a Greenhouse-Geisser correction was used to adjust the p-values and degrees of freedom for interaction and main effects; p-values for simple main effects and multiple comparisons were determined based on Benjamini and Hochberg (1995).

The results revealed a significant interaction, F (6.48, 116.61) = 2.868, p = .0372, η^2 = .121, and a significant simple main effect of factor 1 at "Searching for suitable subjects and methods" and "Concept modification" (F (2, 36) = 6.384, p = .0169, η^2 = .262; F (2, 36) = 4.733, p = .0449, η^2 = .208, respectively). YNG used more "Searching for suitable subjects and methods" than EXP_after (p = .0032); in contrast EXP_after used more "Concept modification" than YNG and EXP_before (p = .0022, p = .0218, respectively).

The results indicate that artists in their early careers changed their artwork unsystematically and searched for suitable subjects and methods based on their previous ideas/concepts. Unsystematic refers to taking "a big jump" in creation, whereby it is difficult to identify commonality between new and previous artwork. Seeking suitable subjects and methods, however, is a means by which artists make more suitable artwork while keeping a prior art idea/concept. In fact, after enacting unsystematic changes, 36% (YGN 18%) of the artists searched for suitable subjects and methods. This suggests that the artists generated sub goals within their art-making process to find appropriate methods and motifs after taking "a big leap" in their creative activity.

Comparison in process modification types within AwCV group

To examine differences among usage types before and after

Table 2. - The number of artists using each type of process modification and the mean number of times each process was used

	YNG (n=14)		EXP _before vision (n=14)		EXP _after vision (n=11)		AwCV_before vision (n=14)		AwCV_after vision (n=14)	
Process Modification Type	No. of artists	Mean no. of times (*SD*)	No. of artists	Mean no. of times (*SD*)	No. of artists	Mean no. of times (*SD*)	No. of artists	Mean no. of times (*SD*)	No. of artists	Mean no. of times (*SD*)
Type 0 No modification	1	0.1 (0.27)	0	0.0 (0.00)	0	0.0 (0.00)	0	0.0 (0.00)	0	0.0 (0.00)
Type 1_1 Unsystematic change	8	1.1 (1.23)	9	1.9 (2.27)	3	0.5 (0.93)	8	1.6 (1.39)	3	0.4 (0.84)
Type 1_2 Searching for suitable subjects and methods	12	1.3 (0.83)	7	0.7 (0.83)	1	0.2 (0.60)	6	0.9 (0.95)	0	0.1 (0.53)
Type 2_1 Quantitative modification	1	0.2 (0.80)	2	1.4 (4.29)	2	0.5 (1.51)	1	1.1 (4.28)	2	0.4 (1.34)
Type 2_2 Subject modification	11	2.6 (3.13)	13	5.4 (3.98)	11	5.9 (3.24)	11	4.8 (3.96)	12	5.1 (3.37)
Type 2_3 Subject modification with reconsideration of methods	11	1.7 (1.82)	6	3.0 (5.82)	7	1.8 (2.23)	4	0.9 (2.37)	9	1.7 (2.02)
Type 3 Structure modification	4	0.5 (0.85)	5	0.6 (1.01)	8	3.5 (6.71)	2	0.4 (0.76)	6	2.9 (6.05)
Type 4 Concept modification	2	0.4 (1.34)	0	0.0 (0.00)	9	4.3 (6.90)	0	0.0 (0.00)	10	3.8 (6.22)
Dead end	3	0.2 (0.43)	6	0.5 (0.65)	1	0.1 (0.30)	2	0.4 (0.65)	0	0.1 (0.27)

finding a creative vision, we focused on AwCV (artists with creative vision) and compared the number of artists using each type of process modification from before to after realizing this vision (see Table 3). McNemar's test was used for the matrix. The results indicate that the number of artists using "Concept modification" increased significantly after finding a creative vision (Holm's adjusted $p = .008$).

Next, we summed the number of times artists used each type of process modification before and after finding their creative vision and then conducted a two-way factorial analysis of variance (within-subjects, factor 1: types of process modification (8 levels) × factor 2: before and after finding creative vision (2 levels)). Mauchly's test of sphericity revealed a sphericity violation ($p < .01$). Therefore, the Greenhouse-Geisser correction and Benjamini-Hochberg adjustment were employed.

The results indicated a significant interaction, F (2.06, 26.77) = 3.902, $p = .0315$, $\eta^2 = .231$, and a significant simple main effect of types of process modification before and after finding a creative vision (respectively, F (2.58, 33.61)= 7.036, $p = .0014$, $\eta^2 = .351$; F (1.67, 21.73)= 5.048, $p = .0201$, $\eta^2 = .280$). AwCV before finding a creative vision used more "Unsystematic change" and "Searching for suitable subjects and methods" than after finding a vision. Conversely, AwCV after finding a vision used more "Concept modification" than before (respectively, F (1, 13)= 6.421, $p = .0249$, $\eta^2 = .331$; F (1, 13)= 4.924, $p = .0449$, $\eta^2 = .275$; F (1, 13)= 5.192, $p = .0402$, $\eta^2 = .285$). Additionally, "Subject modification" was used more frequently than all of other types of process modification before realizing a creative vision ($p < .05$), and more frequently than "Quantitative modification," "Unsystematic change," "Searching for suitable subjects and methods," and "Subject modification with reconsideration of artistic methods" after realizing a creative vision ($p < .05$).

These results suggest that artists who have not yet found their creative vision tended to change their artwork unsystematic or search for suitable subjects and methods to produce satisfactory work. After finding a creative vision, the artists typically generate new ideas/concepts and are productive based on this vision. For example, Figure 1 shows the developmental trajectory of EXP_SG, who is one of our expert artists. He realized his creative vision on "How to See" eight years after beginning his career as a contemporary artist. During his first artwork series, called "Inside Outside," the size of artwork became increasingly large; thus the series reached a deadlock. Because of his creative vision, he was able to generate new art concept, called "Institute of Intimate Museums (IIM)", which aims to encourage viewers/visitors of his work to create their own private museums in spaghetti boxes. This "IIM" concept has helped him develop many series, referred to as "museums in …" (e.g., windowed envelopes, garments, and toy boxes). Additionally, he generated new related ideas, including "Director in museum," "Viewer in museum," and so on. Other artists showed a similar pattern of development. We calculated z-scores on the mean number of artwork series before and after finding a creative vision (before: 14.69 vs. after: 23.39). This result suggests that the number of series increased after artists found their creative vision.

General Discussion

Several features of younger artists and experts (or artists before and after finding a creative vision) can be reviewed in terms of art-making process, specifically in terms of how artists engaged in process modification.

Overall, the results suggest that younger artists and artists *before* finding a creative vision create successful work through the following processes:

1) Using the same process modification, such as "Subject modification," as experts.

2) Using different types of process modification from experts, including "Unsystematic change" and "Searching

Table 3. Number of artists (AwCV) using each type of process modification before and after finding a creative vision

Type 0 No modification		After vision: Absence	After vision: Presence
Before vision	Absence	14	0
	Presence	0	0

Type 1_1 Unsystematic		Absence	Presence
Before vision	Absence	2	2
	Presence	9	1

Type 1_2 Search		Absence	Presence
Before vision	Absence	5	1
	Presence	8	0

Type 2_1 Quantitative		Absence	Presence
Before vision	Absence	12	1
	Presence	0	1

Type 2_2 Subject		Absence	Presence
Before vision	Absence	1	1
	Presence	0	12

Type 2_3 Reconsider		After vision: Absence	After vision: Presence
Before vision	Absence	5	4
	Presence	0	5

Type 3 Structure		Absence	Presence
Before vision	Absence	5	5
	Presence	0	4

Type 4 Concept **		Absence	Presence
Before vision	Absence	3	11
	Presence	0	0

Dead end		Absence	Presence
Before vision	Absence	8	5
	Presence	1	0

* $p < .05$, ** $p < .01$
$n = 14$, including 3 YNGs

for suitable subjects and methods."

Conversely, expert artists and artists *after* finding a creative vision create their work by:

3) Using "Concept modification" based on their creative vision.

4) Generating new art concepts and producing more artwork series than before finding a creative vision.

Half of the younger artists and artists in the early career stage tend to use Unsystematic change, which refers to changing the art subjects, methods, and concepts while creating artwork. As these artists are yet to clearly realize their superordinate concepts (or main theme/creative vision), they are unable to use Structure and Concept modification effectively. These younger artists, however, make new artwork while seeking suitable subjects and methods, which are based on concepts from prior work. This helps the younger artists form (or recognize) their own art-making theme.

After realizing a creative vision, artists create their work by implementing Structure and Concept modification. A creative vision, which is formed through many years of activity and consists of long-term intentions or goals for creation, plays a vital role in guiding the use of process modification. Thus, artists who have found their creative vision are able to work more productively and creatively.

Our results are consistent with the claim that artistic creation does not derive from "irrational and random thoughts/ideas" in creative writing (e.g., Oatley & Djikic, 2017), while several researchers claim that creativity depends on blind variation and random retention (e.g., Campbell, 1960). Creative writing studies indicate that writers continue to explore the same theme (or related themes) in their literary work (e.g., Patrick, 1937; Oatley & Djikic, 2017). Although previous studies described the exploration in the creation of poetry, literature, and fine art (e.g., Boden, 2010), they have not revealed how the exploration occurs or what kind of exploration contributes to longitudinal creative work.

Regarding these questions, using in-depth analysis of dancers' practice, Shimizu and Okada (2018) revealed that expert breakdancers engaged in "exploratory practice" to generate new and original skills. They claimed that "The dancers practiced with multiple goals, that is, not only to improve the quality of the skills but also to develop original and flexible skills that fit well into a performance by varying aspects of domain skills and by combining those domain skills with other domain skills" (Shimizu & Okada, 2018, p. 2392).

Artistic creation is also a process of exploring for a theme, concept, method, and motif to achieve one's goal as an artist. The present study reveals extensive explorations in artistic creation via process modification. A creative vision guides artists' creation and enables them to give consistency to their work. The formation of a creative vision seems to correlate with reconsiderations of the methods, subjects, and ideas for artwork series while reflecting on art-making processes and experiences. Thus, the process modification framework is useful when analyzing the details of the development of artwork series and creative expertise.

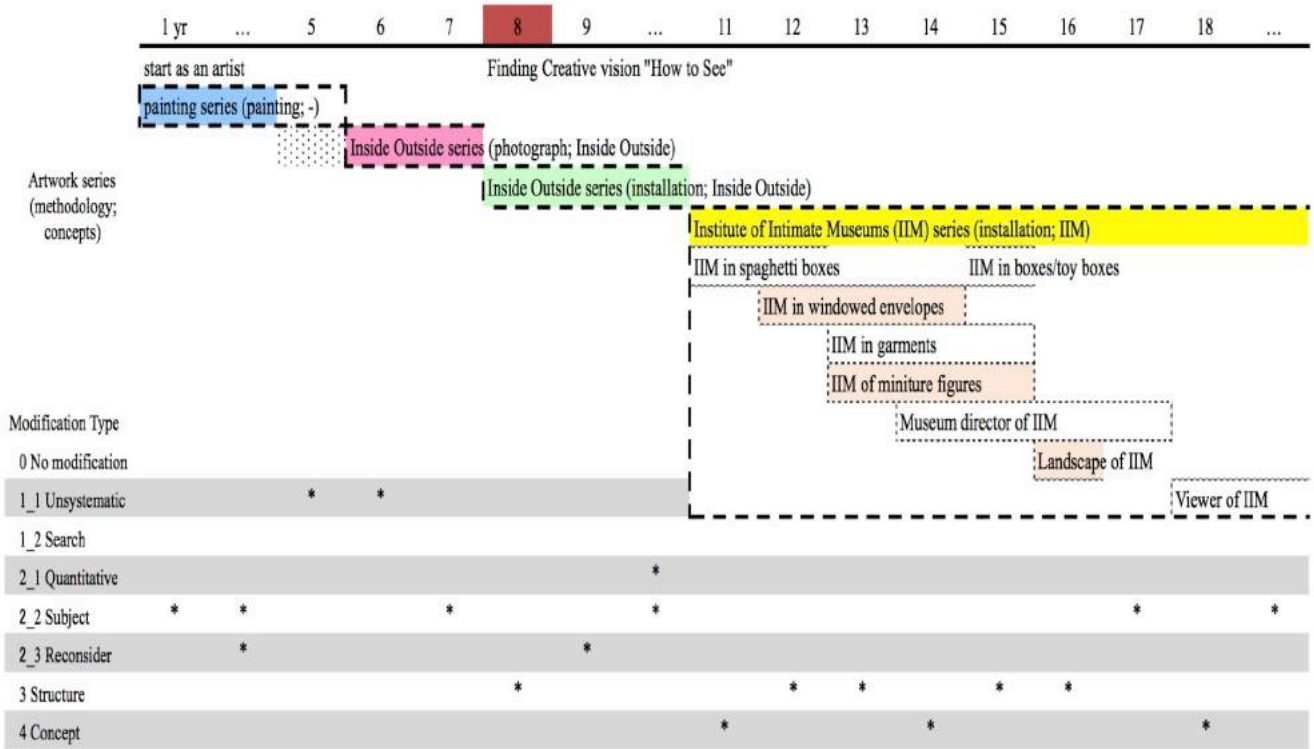


Figure 1. The developmental trajectory of artwork series created by EXP_SG

Acknowledgments

This work was supported by JSPS KAKENHI Grant number JP26780364 and 20243032. We thank the artists for their cooperation.

References

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*, 289-300.

Boden, A. M. (2004). Creative minds: Myths and mechanisms (second edition). Routledge.

Boden, A. M. (2010). Creativity and art: Three roads to surprise. Oxford: Oxford University Press.

Campbell, D. T. (1960). Blind variation and selective retention in creative thought as in other knowledge processes. *Psychological Review*, *67*, 380-400.

Getzels, J. W., & Csikszentmihalyi, M. (1976). *The Creative vision: A longitudinal study of problem finding in art*. Wiley.

Kawakita, J. (1967). *Abduction* (Hasso-ho). Tokyo: Chukoshinsho. (in Japanese).

Mace, M., & Ward, T. (2002). Modeling the creative process: A grounded theory analysis of creativity in the domain of art making. *Creativity Research Journal*, *14*, 179-192.

Oatley, K., & Djikic, M. (2017). The creativity of literary writing. In J. M. Kaufman, V. P. Glaveanu, & J. Baer. (Eds.), *The Cambridge handbook of creativity across domains*. (pp. 63-79). Cambridge University Press.

Okada, T., Yokochi, S., Ishibashi, K., & Ueda, K. (2009). Analogical modification in the creation of contemporary art. *Cognitive Systems Research*, *10*, 189-203.

Schön, D. A. (1983). *The reflective practitioner*. New York: Basic Books.

Shimizu, D. & Okada, T. (2018). How do creative experts practice new skills? Exploratory practice in breakdancers. *Cognitive Science*, *42*, 2364-2396.

Simon, H.A. (1973). The structure of ill-structured problems. *Artificial Intelligence*, *4*, 181–201.

Stokes, P. D. (2014). Thinking inside the tool box: Creativity, constraints, and the colossal portraits of chuck close. *Journal of Creative Behavior*, *48*, 276–289.

Takagi, K., Kawase, A., Yokochi, S., & Okada, T. (2015). Formation of an art concept: A case study using quantitative analysis of a contemporary artist's interview data. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the 37th Annual Meeting of the Cognitive Science Society* (p. 2332-2337). Austin, TX: Cognitive Science Society.

Takagi, K., Okada, T., & Yokochi, S. (2013). Formation of an art concept: How is visual information from photography utilized by the artist in concept formation. *Cognitive Studies: Bulletin of the Japanese Cognitive Science Society*, *20*, 59-78. (in Japanese)

Patrick, C. (1937). Creative thought in artists. *Journal of Psychology, 4*, 35-73.

Yokochi, S., & Okada, T. (2006). Artists' long-term process of making art. *Proceedings of 28th Annual Meeting of the Cognitive Science Society*, 2635.

Yokochi, S., & Okada, T. (2007). Creative expertise of contemporary artists. *Cognitive Studies: Bulletin of the Japanese Cognitive Science Society*, *14*, 437-454. (in Japanese)

Zimmerman, B. J. (2006). Development and adaptation of expertise: The role of self-regulatory processes and beliefs. In K. A. Ericsson, N. Charness, P. J. Feltovich, & R. R. Hoffman (Eds.), *The Cambridge handbook of expertise and expert performance*, (pp. 705–722). Cambridge: Cambridge University Press.

Preschool children's understanding of polite requests

Erica J. Yoon and **Michael C. Frank**
{ejyoon, mcfrank} @stanford.edu
Department of Psychology, Stanford University

Abstract

As adults, we use polite speech on a daily basis. What do children understand about polite speech? Looking at children's polite speech comprehension can help examine children's pragmatic understanding more generally, and can be informative for caregivers who want to teach children what it means to be polite. Even though children start to produce polite speech from early on, there is little known about whether they understand intentions behind polite language. Here we show that by 3 years, English-speaking preschool children understand that it is more polite and nicer (and less rude and mean) to use politeness markers such as "please" when making requests, and by 4 years, they understand that the use of these politeness markers indicates that the speaker is more socially likeable and is more likely to gain compliance from their conversational partners. This work can help lay the foundation for future work on children's understanding of polite speech and pragmatic development more generally.

Keywords: Politeness, pragmatic development, online experiment

Introduction

We use and hear polite speech on a daily basis: polite utterances range from simple words of apology ("sorry") or gratitude ("thanks") to compliments ("I love your dress!") and requests ("Can you please open the window?"). Yet polite utterances are seemingly inefficient and even misinformative: speakers say "Can you please ..." when it should suffice to say, "Open the window." These facts are a mystery for frameworks which describe communication in terms of efficient information transfer (e.g., Buhler, 1934; Goodman & Stuhlmuller, 2013; Shannon, 1948): If language is a tool for transferring information, speakers should be as efficient as possible in their communication to prioritize informativity. Nonetheless, everyday politeness is ubiquitous in everyday language use, and adults tend to use strategies to be polite even while arguing (Holtgraves, 1997).

So why do people speak politely? Linguistic theories assume that people's utterance choices are motivated by social concerns, framed as either maxims (Leech, 1983), social norms (Ide, 1989), or listener's and/or speaker's public identity (*face*; Brown & Levinson, 1987). For example, Brown & Levinson (1987)'s theory predicts that if a speaker's intended meaning contains a threat to the listener's face or self-image, the speaker's utterance will be less direct and less informative. For example, if a speaker considered that saying "Open the window" will give the impression that she is in a position to give orders to the listener, she could instead say "Can you please open the window?", using a more indirect form of request to give the other person a sense of autonomy or freedom from imposition (Clark & Schunk, 1980). Thus, while it may hinder the goal of efficient information transfer, using polite speech can help the speaker save the listener's face while simultaneously communicating her own positive social goals (Yoon, Tessler, Goodman, & Frank, 2017).

Do children speak politely, and if so, what do they understand about polite speech? Previous research shows that children begin producing polite speech early on; They produce "please" at 2.5 years (Read & Cherry, 1978), and request forms increase in their variety and frequency with age (Bates, 1976; Bates & Silvern, 1977; Bock & Hornsby, 1981; Ervin-Tripp, 1982; Nippold, Leonard, & Anastopoulos, 1982). Young children learn to produce different forms of requests depending on context: For example, by three years children are able to vary their utterances based on whether they are instructed to "tell" versus "ask" an addressee to given them a puzzle piece (Bock & Hornsby, 1981). And even at two years, children are able to modify their requests to make them more polite ("ask in the nicest way possible"; Bates & Silvern, 1977). Hence, children's production of polite speech seems to parallel adult speakers' desires to produce utterances with appropriate levels of face-saving.

While children appear to produce polite speech from an early age, less is known about whether they *understand* polite speech. Examining children's comprehension of polite speech is important for a number of reasons. First, children's polite speech understanding can reveal their inferential abilities underlying more general pragmatic understanding: going beyond what was literally said to infer what was intended. For example, children need to understand that, in saying "can you open the window?" the speaker does not literally question the listener's ability to open the window but rather wants to make a polite request. Thus, understanding what children comprehend about polite speech can help see how children are able to infer speaker's intentions behind utterances.

Second, understanding polite speech can have practical implications for education, as caregivers often care about teaching their children to be more polite. Indeed, from very early on, parents teach children to follow normative rituals to say "please", "thank you", "hello" and "good-bye" (Gleason, Perlmann, & Greif, 1984). It can be enlightening to know

whether and when children understand positive implications of following these norms.

Third, examining children's *comprehension* of polite speech as compared to their *production* is meaningful, in that children's comprehension can reveal more abstract representations and inferences about language than their productivity (e.g., Fisher, 2002): Children's ability to say "please" early on does not necessarily indicate that they understand saying "please" is more polite, nicer and socially apt, as children may simply obey or imitate what their caregivers tell them to say without understanding its meaning.

Research on children's comprehension of polite speech has received less focus than research on their production of polite speech. Moreover, the few studies that did examine children's understanding of polite speech have been largely inconclusive. Though there was some initial evidence to suggest that producing a request with "please" is judged to be polite by three years of age (Bates, 1976; Bates & Silvern, 1977), in a later study, the judgment of "please" as being polite was only replicated starting at five years of age, but not younger (Nippold et al., 1982). These initial studies also lacked statistical tests to assess each age group's performance, and did not systematically manipulate cues other than linguistic markers (e.g., prosody or facial expressions).

In addition to children's recognition of politeness markers, there are also many open questions about their abilities to recognize the intentions underlying polite speech. For example, do children know that the word "polite" should be associated with politeness rules people abide by (e.g., saying "please")? Relatedly, do children recognize polite speech as being positively valenced, such that they think it is better and nicer to say polite things? Do children understand the social implications of speaking politely? For example, polite people may be more likely to get their wishes granted ("I will pour him more water because he was nice") and may be better social play partners compared to those who are impolite. Finally, what cues to politeness do children recognize? Do they recognize linguistic politeness markers such as "please," or "can you," or both? Or do they rely on prosodic cues that make utterances sound more respectful, or on facial expressions that make a person look kind?

In this current work, we sought to answer these questions, and test what 2- to 4-year-old children understand about requests using politeness markers. Across three experiments, we presented stories about speakers who decided to speak politely (e.g., "Please pour me more water") or impolitely ("Pour me more water") and asked child participants to make judgments between the two speakers. We examined in each experiment whether: (1) children are able to reason about speakers using polite speech as being relatively more "polite" and "nice" and less "rude" or "mean" than speakers not using polite speech; (2) they can reason about social implications of using polite speech (e.g., politeness as a sign of a nice play partner, or greater likelihood of compliance from the addressee); and (3) they show improvement with age for these lines of reasoning. We also examined whether children need additional cues to politeness such as facial expressions (Expt 1) or prosodic cues (Expt 2), or they can make use of linguistic politeness markers alone (Expt 3) to make appropriate inferences about speakers.

Experiment 1

In Experiment 1, we tested whether 3- to 4-year-old children were able to understand the implications of using simple politeness markers, based on linguistic cues of interest (whether the speaker says "please," "can you") and other cues (facial expressions and prosodic cues) that make polite speech more salient and naturalistic.

Methods

Participants 3-year-old ($n = 20$; 12 F, M_{age} = 3.61 years, SD_{age} = 0.22) and 4-year-old children ($n = 18$; 6 F, M_{age} = 4.38 years, SD_{age} = 0.25) were recruited from a local preschool. An additional 3 children were tested but excluded due to failure on the practice questions ($n = 2$) or completion of fewer than half of the test trials ($n = 1$).

Stimuli and design We designed a picture book (see Figure 1) with twelve stories in which a protagonist is approached by two speakers, one of whom makes a request by producing an utterance with a politeness marker (e.g., "Please pour me more water"), and the other produces an utterance without ("Pour me more water"). There were three types of politeness marker that could be used: "please" (as in "Please pour me more water"), "can you" ("Can you pour me more water"), and "can you please" ("Can you please pour me more water").

We designed six question types to ask participants following the presentation of the stories: four *speaker attribute* questions (*polite*: "Which one was more polite?"; *rude*: "Which one was more rude?"; *nice*: "Which one was nicer?"; *mean*: "Which one was meaner?") and two *social implication* questions (*play partner*: "Which one would you rather play with?"; *compliance*: "Which one will [get what they want]?"). Each participant would be asked one of the four speaker attribute questions, followed by one of the two social implication questions.

In Experiment 1, all utterances were produced live by the experimeter, with appropriate proodic cues and facial expressions for each request: Utterances with politeness markers were produced by kind voice and facial expression, whereas utterances lacking politeness marker were produced with angry voice and facial cues.

Procedure The experimenter presented to the child a storybook with a total of thirteen stories about different characters. In the *practice* phase, the child heard a story with one clearly mean character (*Drew kicked Carol*) and one clearly nice character (*Graham gave Carol a gift*). After a reminder of what each character did, the experimenter asked the participant: *Which one was being meaner?* and *Which one was being nicer?* If the child answered the question wrong the first

Figure 1: Example story.

time, the experimenter read the story one more time, saying, "Let's think about the story one more time." Only children who correctly answered both questions in the first or second attempt were included in the analyses.

In the *test* phase, the child heard twelve stories, in each of which they saw one speaker who decided to speak politely (*Jean wanted more water in her cup. Jean said to Fred, "Please pour me more water"*) and another speaker who spoke impolitely (*Suzy also wanted more water in her cup. Suzy said to Fred, "Pour me more water."*). After a reminder about what each speaker said, the child was asked a total of two questions. For the first question, the experimenter asked one out of four possible questions for speaker attribute: "Which one was being more polite [more rude/nicer/meaner]?" For the second, social implication question, the experimenter either asked about play partner (*Which one would you rather play with?*) or likelihood of compliance (e.g., *Which one will Fred give water to?*). The order of story types and question types was counterbalanced.

Results and Discussion

We looked at the proportion of correct responses to various questions comparing speakers who used a politeness marker and spoke kindly, and speakers who did not use a politeness marker and spoke meanly (Figure 2, top row). A mixed-effects logistic regression predicting accuracy based on age, question type and politeness marker type[1] showed there was an improvement with age ($\beta = 0.2$, $p = .026$). The regression model also revealed that children seemed to find some question types easier than others: Responses to *nice* and *mean* questions were more accurate than to *polite* and *rude* questions ($\beta = 0.8$, $p = .002$), whereas social implication questions (*play partner* and *compliance*) were overall more difficult compared to speaker attribute questions (*polite*, *rude*, *nice*, and *mean*; $\beta = -0.33$, $p = .006$).

[1]for Experiments 1 and 2, we use this model structure with a maximal random effect structure that converges: `accuracy ~ age x question type x politeness marker type + (1 | item)`, where age is continuous, centered and scaled. All categorical variables were deviation coded, with specified contrasts of interest for the question type. Significance was calculated using the standard normal approximation to the t distribution (Barr, Levy, Scheepers, & Tily, 2013).

Looking more closely at responses for each of the question types, children from both age groups tended to accurately answer the *polite*, *nice*, *mean*, *rude*, and *play partner* questions overall (3-year-olds' mean accuracy range: 0.58 - 0.88; 4-year-olds' mean accuracy range: 0.68 - 0.9), indicating correctly that the speaker who used a politeness marker was more polite and nicer, and less mean and rude, and was likely a better play partner. For the *compliance* question, 4-year-olds overall answered correctly that the speaker who used politeness marker will likely get what they want from the listener ($M_{4y} = 0.75$, $p < .01$), but 3-year-olds did not perform above chance ($M_{3y} = 0.58$). As for the different politeness marker types, both age groups overall tended to give correct answers based on all three markers, but especially "can you please" (3-year-olds: $M_{please} = 0.66$, $M_{canyou} = 0.72$, $M_{canyouplease} = 0.74$; 4-year-olds: $M_{please} = 0.73$, $M_{canyou} = 0.77$, $M_{canyouplease} = 0.84$).

In sum, in this first experiment, we saw preliminary evidence that children pay attention to some cues to politeness and are able to use these cues to infer whether speakers are relatively polite, rude, nice or mean, and whether speakers are good play partners and are likely to get what they wanted from their addressees. 4-year-olds answered questions accurately more often compared to 3-year-olds, especially for the question about addressee's compliance with the speaker's request. In general, however, both age groups tended to be accurate when all the possible cues were used to signal that one speaker was polite (used "can you please", spoke with a kind tone and face) and the other speaker wasn't (did not use a politeness marker, spoke with an angry tone and face).

There were a number of remaining issues from Experiment 1. Children may not have used the linguistic politeness markers (e.g., "please") per se, and rather prosodic and facial cues that accompany these markers. That is, children may have relied on the speaker's kind voice and face rather than their use of "please" to evaluate their niceness or likeability as a play partner. Similarly, greater accuracy for some questions over others (e.g., *nice* > *polite*) may have been due to greater association between some of the words and prosodic and facial cues (e.g., a kind face may be seen to signal niceness more than politeness), not due to greater understanding for those words or concepts. Another concern is that the ex-

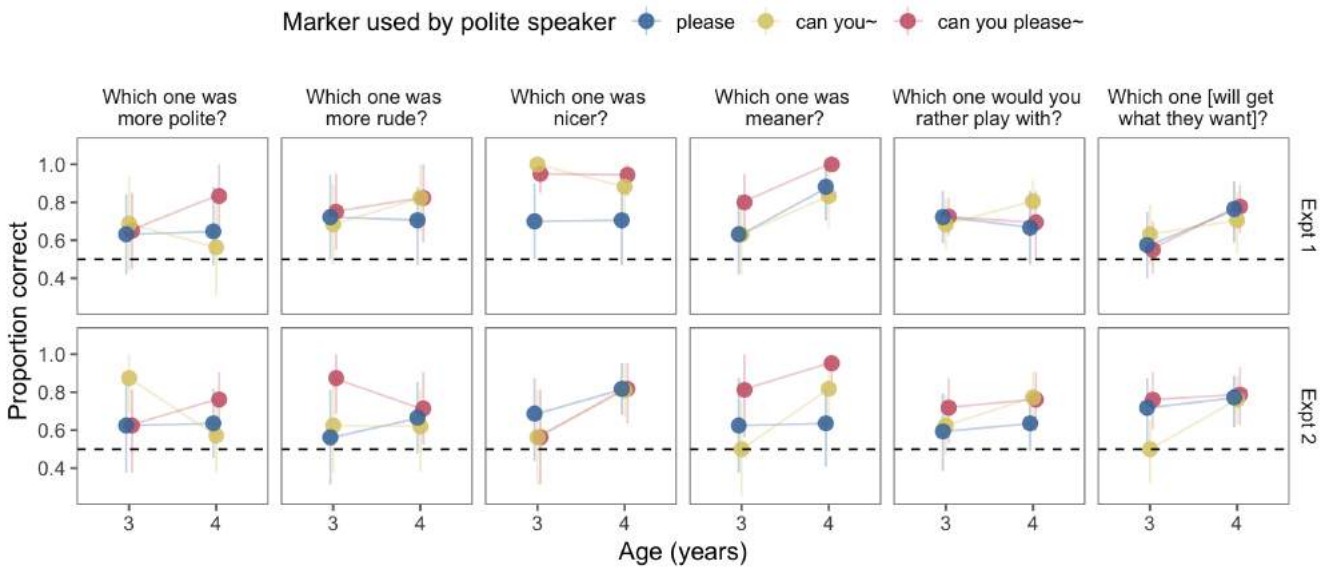


Figure 2: Experiment 1 and 2 results. Proportion of correct responses to questions comparing between a speaker who used a politeness marker (where blue indicates "please", yellow "can you", and red "can you please") versus a speaker who did not. Data are binned into one-year age groups. Each row represents data from a different Experiment. Columns represent different questions asked. Dashed line represents chance level at 50% (i.e., if participant were guessing at random).

perimenter was aware of the manipulations (i.e., they knew which speaker was supposed to be "polite") and thus could have affected the presentation of these speakers in ways that are not consistent across all participants. In our next two experiments, we sought to remove these potential confounds.

Experiment 2

In Experiment 1, we saw initial evidence that children can use some combinations of linguistic, prosodic, and facial cues to politeness. In Experiment 2, we examined whether children can make similar judgments using linguistic and prosodic cues only, without facial expressions. For this, we conducted a preregistered experiment where we used prerecorded voiceovers to present speaker utterances, so that (1) we could look at children's judgments based on linguistic markers and prosodic cues only, and (2) we could remove the role of the experimenter in presentation of these utterances.

Methods

Participants 3-year-old (n = 16; 8 F, M_{age} = 3.56 years, SD_{age} = 0.29) and 4-year-old children (n = 22; 13 F, M_{age} = 4.5 years, SD_{age} = 0.32) were recruited from a local preschool. An additional 5 children were tested but excluded due to failure on the practice questions.

Stimuli and design The design was identical to Experiment 1. Stimuli were the same as Experiment 1 except two changes: (1) Instead of a picture book, we presented the stories on a tablet; (2) the speakers' utterances were now presented as recorded voiceovers. The voiceovers were recorded by native English speakers, and contained prosodic cues that matched the presence/absence of a politeness marker (e.g., the speakers were instructed to record "Please pour me more water" with a "kind voice"" and "pour me more water" with an "angry voice").

Procedure The procedure was identical to Experiment 1, except for the following change: The participants now had to tap on a speaker on tablet in order either to hear them speak, or to choose an answer to the questions asked.

Results and Discussion

Overall we saw similar patterns of results in Experiment 2 (Figure 2, bottom row) compared to Exp. 1. A mixed-effects logistic regression predicting accuracy based on age, question type and politeness marker type showed that accuracy improved with age (β = 0.25, p = .002), and children made accurate judgments more often when the politeness marker was "can you please" than when the marker was "please" or "can you" (β = 0.33, p = .019). There was no main effect of question type, but there was an interaction between age and question type such that performance for *nice* and *mean* questions saw greater improvement with age than for *polite* and *rude* questions (β = 0.57, p = .011).

For children's responses to different question types, 3-year-olds' accuracy did not differ from chance level for *nice*, *mean*, and *play partner* questions, but their means numerically exceeded 50% for all question types, and 4-year-olds accurately answered questions of all types (3-year-olds' mean accuracy range: 0.6 - 0.88; 4-year-olds' mean accuracy range: 0.66 - 0.9). For politeness marker types, 3-year-olds' performance did not differ from chance for "please" and "can you", but

both age groups tended to answer questions about different politeness markers accurately overall (3-year-olds: M_{please} = 0.63, M_{canyou} = 0.61, $M_{canyouplease}$ = 0.72; 4-year-olds: M_{please} = 0.7, M_{canyou} = 0.72, $M_{canyouplease}$ = 0.8).

In sum, across Experiments 1 and 2, we saw that children tend to make accurate judgments about speakers given their use of politeness markers, especially "can you please," together with prosodic cues, and children get better with age in their use of politeness cues to respond to questions about speaker attributes and social implications.

Experiment 3

We conducted a third, preregistered experiment to see whether children are able to evaluate speakers based on linguistic markers only, without any other supporting cues such as prosodic cues or facial expressions.

Methods

Participants We recruited two samples of participants: one from the same local nursery school as Experiments 1 and 2, and the other from Lookit (`https://lookit.mit.edu/`), an online platform for child research participation, in which parents and their children can participate together. The nursery school sample consisted of 3-year-old ($n = 24$; 11 F, M_{age} = 3.65 years, SD_{age} = 0.26) and 4-year-old children ($n = 25$; 13 F, M_{age} = 4.48 years, SD_{age} = 0.28). An additional 3 children were tested but excluded due to failure on the practice questions. The online sample consisted of 2-year-old ($n = 23$; 12 F, M_{age} = 2.48 years, SD_{age} = 0.29), 3-year-old ($n = 31$; 15 F, M_{age} = 3.59 years, SD_{age} = 0.27) and 4-year-old children ($n = 27$; 12 F, M_{age} = 4.46 years, SD_{age} = 0.29). An additional 28 children were tested but excluded due to failure on the practice questions ($n = 19$) or completion of fewer than half of the test trials ($n = 9$).

Stimuli For the nursery school sample, stimuli were identical to Experiment 2 except that the voiceovers for all utterances had the same prosody: All utterances ended with a rising intonation. For the online sample, stimuli were identical to what the nursery school participants saw except that the story narrations (other than speaker utterances) were also pre-recorded such that parents did not need to read the stories aloud to their children.

Procedure For the nursery school sample, the procedure was identical to Experiment 2. For the online sample, the procedure was similar except that parents and children participated together at home and there was no experimenter present. Parents accessed the webpage for the study and gave their consent for participation, and then read instructions to proceed through the different stories, which specifically asked the parents to not tell their children correct answers for the questions.

Results and Discussion

Experiment 3 For Experiment 3, we were able to look at how children answered the *polite* and *rude* questions given the same three politeness marker types as in Experiments 1 and 2, with three age groups including 2-year-olds. (Fig. 3).

A mixed-effects logistic regression controlling for the effect of sample[2] showed improvement with age ($\beta = 0.19$, $p = .033$) as well as better performance for "can you please" than "please" and "can you" together ($\beta = 0.42$, $p = .002$), consistent with Experiment 2 results. Performance for "please" was also better than for "can you please" and "can you" together ($\beta = .3$, $p = .027$), which may be surprising given that we previously did not see the same effect in Experiments 1 and 2. One possible explanation is that controlling for prosodic cues in Experiment 3 actually made it *easier* to use "please" as a politeness cue. Because we had stripped all the other variations, it may have made the contrast between the presence and absence of the marker "please" *more* salient.

Additionally, children were better with the *polite* questions than *rude* overall (β = -0.19, $p = .04$), but especially given "please" (β = .42, $p = .002$). Finally, children showed a greater improvement with age for "can you please" compared to "please" and "can you" together ($\beta = 0.38$, $p = .004$).

All experiments Did children perform better given facial and/or prosodic cues, or were linguistic politeness markers sufficient? To see any potential effect of experiment on children's performance, we conducted an exploratory mixed-effects logistic regression on all three experiments together[3]. The regression model showed no significant main effect of experiment, suggesting that children did not perform more poorly when facial and prosodic cues were removed, and they were able to make accurate judgments based on linguistic cues alone. The model also showed that children improved with increasing age ($\beta = 0.33$, $p < .001$) and that children were more accurate with "can you please" than "please" and "can you" ($\beta = 0.25$, $p = .011$), confirming results from each individual experiment. Additionally, the model showed that children became better at judging the politeness marker "can you please" with age ($\beta = 0.73$, $p = .005$), and that children answered *polite* questions better than *rude* questions about the marker "please" ($\beta = 0.26$, $p = .006$)

General Discussion

What do young children understand about polite speech? In three experiments, we looked at how 2- to 4-year-old children reason about making requests with or without simple politeness markers such as "please", "can you" and "can you please." By 3 years, children pay attention to the use of politeness markers to accurately judge whether that speaker is relatively more polite, rude, nicer or meaner compared to another speaker. By 4 years, children reliably infer that a speaker who uses a politeness marker is a better play partner and more likely to get what they want. Across all three experiments, we saw a clear developmental trend such that children improved

[2]Model structure: `accuracy ~ sample + age x question type x politeness marker type + (1 | item)`

[3]Model structure: `accuracy ~ sample + experiment + age x question type x politeness marker type + (1 | item)`

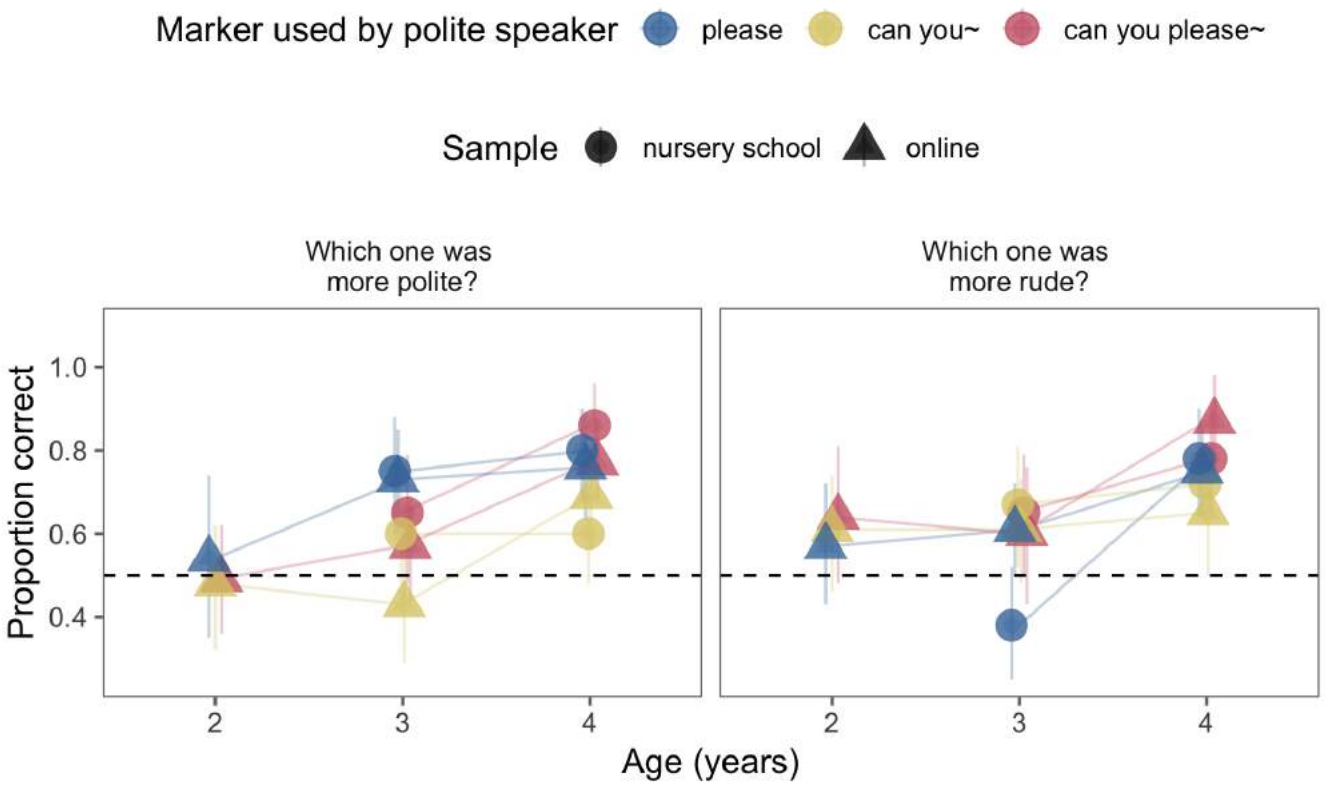


Figure 3: Experiment 3 results. Conventions are identical to Figure 2.

in their reasoning about polite speech with increasing age. We observed no large experiment effects as we eliminated facial and prosodic cues; instead, all these inferences appeared to be supported by linguistic markers alone.

Even though children have been shown to produce polite speech such as "please," evidence has been sparse and inconclusive for whether young children below 5 years comprehend speaker attributes and intentions based on polite speech. Here, we found that children are sensitive to the use of politeness markers in speech, and are able to use these markers to infer the speaker's attributes (e.g., niceness) by 3 years, and consequent social implications by 4 years. These ages are closer to the age of first reliable production of polite speech than have been suggested by earlier work.

Children in the US are often explicitly taught and prompted to use politeness markers such as "please" in their requests from early on (e.g., "What's the magic word?"; Gleason et al., 1984), thus they may quickly learn to use these markers as a rule in order to get what they want. They also might hear other remarks that pair politeness markers with positive words (e.g., "You should be *nice* and say *please*"), which may help them learn the association between polite speech and positive attributes. Gradually, children may recognize more subtle social processes that are related to polite speech production: Adults may praise and reward children who spoke politely, and children themselves may like peers who ask for permission to play with their toys rather than take the toys away without asking. Future work with corpus data analysis looking at these interactions between children and others may reveal important conversational patterns that help children acquire social meanings of polite speech.

There are limitations to the current work that present other opportunities for future research. Because this work looked only at the behaviors of English-speaking children with a relatively high socioeconomic status in the US, it is an open question how children with different language and cultural background may develop understanding of polite speech. Cross-cultural investigation of what markers are present in other languages, cultures and backgrounds, as well as how those markers are acquired, will be informative.

Also, we did not manipulate the social status of speakers or addressees. Though not explicitly stated, the visual depiction and narration used for the current work suggested that speakers were communicating with their peers only. However, one key prediction from politeness theory is that speakers will adjust their utterances based on the status of the addressees (Brown & Levinson, 1987). Indeed children adjust own their speech based on the listener status and age: Even at two years, children use a polite form of request ("Can I have…") to an adult but an imperative form ("Give me…") to a peer (Shatz & Gelman, 1973). Thus, future work should examine how children use cues to politeness to judge speaker

intentions in different contexts, including varied status differences between speakers and listeners.

In sum, the current work showed that young children understand implications of using simple politeness markers in requests. A broader understanding of the emergence of politeness may offer insights into how children become proficient users of language across the wide range of social situations that they encounter.

All experiments, data, and analysis code are available in the public repository for the project: (link will be available upon acceptance)

References

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Bates, E. (1976). Acquisition of polite forms: Experimental evidence. *Language and Context: The Acquisition of Pragmatics*, 295–326.

Bates, E., & Silvern, L. (1977). Social adjustment and politeness in preschoolers. *Journal of Communication*, *27*(2), 104–111.

Bock, J. K., & Hornsby, M. E. (1981). The development of directives: How children ask and tell. *Journal of Child Language*, *8*(01), 151–163.

Brown, P., & Levinson, S. C. (1987). *Politeness: Some universals in language usage* (Vol. 4). Cambridge university press.

Buhler, K. (1934). *Sprachtheorie*. Oxford, England: Fischer.

Clark, H. H., & Schunk, D. H. (1980). Polite responses to polite requests. *Cognition*, *8*(2), 111–143.

Ervin-Tripp, S. M. (1982). Ask and it shall be given unto you: Children's requests. *Georgetown University Roundtable on Languages and Linguistics. Contemporary Perceptions of Language: Interdisciplinary Dimensions*, 235–245.

Fisher, C. (2002). The role of abstract syntactic knowledge in language acquisition: A reply to tomasello (2000). *Cognition*, *82*(3), 259–278.

Gleason, J. B., Perlmann, R. Y., & Greif, E. B. (1984). What's the magic word: Learning language through politeness routines? *Discourse Processes*, *7*(4), 493–502.

Goodman, N. D., & Stuhlmuller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science*, *5*(1), 173–184.

Holtgraves, T. (1997). YES, but... positive politeness in conversation arguments. *Journal of Language and Social Psychology*, *16*(2), 222–239.

Ide, S. (1989). Formal forms and discernment: Two neglected aspects of universals of linguistic politeness. *Multilingua-Journal of Cross-Cultural and Interlanguage Communication*, *8*(2-3), 223–248.

Leech, G. (1983). *Principles of pragmatics*. London, New York: Longman Group Ltd.

Nippold, M. A., Leonard, L. B., & Anastopoulos, A. (1982). Development in the use and understanding of polite forms in children. *Journal of Speech, Language, and Hearing Research*, *25*(2), 193–202.

Read, B. K., & Cherry, L. J. (1978). Preschool children's production of directive forms. *Discourse Processes*, *1*(3), 233–245.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.*, *27*, 623–656.

Shatz, M., & Gelman, R. (1973). The development of communication skills: Modifications in the speech of young children as a function of listener. *Monographs of the Society for Research in Child Development*, 1–38.

Yoon, E. J., Tessler, M. H., Goodman, N. D., & Frank, M. C. (2017). "I won't lie, it wasn't amazing": Modeling polite indirect speech. In *Proceedings of the thirty-ninth annual conference of the Cognitive Science Society*.

Modeling Number Sense Acquisition in A Number Board Game by Coordinating Verbal, Visual, and Grounded Action Components

Arianna Yuan (xfyuan@stanford.edu) **James L. McClelland (jlmcc@stanford.edu)**
Department of Psychology, Center for Mind, Brain, Computation and Technology, Stanford University, Stanford, CA

Abstract

Previous studies including Ramani and Siegler (2008) have shown that playing a number board game improved students performance on several numerical tasks, including numeral identification, magnitude comparison, counting and number line estimation. However, the computational mechanism underlying such number sense acquisition remains unclear. Here, we aim to fill this gap by building a model that simulates play of the game as well as the basic numerical tasks. We hypothesize that cognitive components that are used in the basic tasks are recruited to work together when children play the game, so that the learning induced by playing the game also manifests itself in those tasks. We reproduced the empirical findings with a neural network model implementing our hypothesis. This computational approach demonstrates how a single model that coordinates components of number processing in different modalities (visual, language and spatially-guided action) can explain the number sense acquisition in number board game playing.

Keywords: Numerical Cognition; Mathematical Education; Neural Networks; Board Games

Introduction

Mathematical concepts are notoriously hard to learn, perhaps because they often involve a range of distinct properties. Even the seemingly simple mathematical concepts, such as the natural numbers, may have diverse properties and multiple representations. For instance, the concept of "five" can be grounded as the *cardinality of a set*; as a *position on a line*; as a *distance in space*; or as a *number of events in a temporal sequence*. In fact, researchers have summarized these observations and proposed that humans use several grounding metaphors to understand numbers (Lakoff & Núñez, 2000), including *arithmetic as object collection*, *arithmetic as the use of a measuring stick* and *arithmetic as motion along a path*.

Given these diverse groundings, the perceptual variance of natural numbers may be much larger than that of many ordinary concepts. For example, "five dogs", "five houses", and the position in a row between 4 and 6 are perceptually very different, yet they can all instantiate the number "five". How do children learn to link different representations of numbers, particularly non-symbolic ones, such as cardinality and distance in space, to symbols such as verbal number words and written Arabic numerals? This problem is called the symbol grounding problem (Harnad, 1990), thought to be equivalent to the problem of determining how we assign a meaning to a symbol (in our case, a number word).

Many researchers have attempted to provide an answer to this problem. One popular account, the *approximate number system (ANS) mapping account*, assumes that a symbol acquires its numerical meaning by being mapped on a non-verbal and innate ANS. Evidence supporting this hypothesis includes longitudinal studies in developmental psychology. For instance, there is a large body of literature showing a correlation between non-symbolic number processing and symbolic math (Halberda, Mazzocco, & Feigenson, 2008; Libertus, Feigenson, & Halberda, 2011) and arguing for a causal link between the two (Park & Brannon, 2014).

Whatever the origin of non-symbolic number may be, the question remains, what is the process whereby the many diverse aspects of non-symbolic number and symbolic numbers are acquired, to support skills such as number-space mapping and numerical magnitude comparison? In the current paper, we begin to address this question. Particularly, we will focus on a number board game that has been used in several studies to provide a rich learning environment that grounds various aspects of numerical processing and links them to the printed and spoken symbols for numbers. In the seminal paper by Ramani and Siegler (2008), the authors showed that playing this number board game for roughly an hour spaced over several sessions increased low-income preschoolers' proficiency on four diverse numerical tasks: numerical magnitude comparison, number line estimation, counting (defined as reciting the count list from 1 to 10) and numeral identification. Below we describe the details of their intervention study.

In the board game, the board includes 10 horizontally arranged squares of the same size, with the word "Start" at the left end and "End" at the right end. There are two conditions in the study. In the experimental condition, the board the squares are numbered consecutively from 1 to 10 in order from left to right. In the control condition, the squares only have alternating colors. In addition, in the experimental condition, the game has an associated spinner with a "1" half and a "2" half, whereas in the color board version it has a spinner with colors that correspond to the colors of the squares on the board. Before playing the game, the participants were tested on 4 numerical tasks: numeral identification, magnitude comparison between two numbers, counting and number line estimation.

In the game, children chose an animal token and used it to mark their progress on the board. Children were instructed to take turns spinning the spinner and were told that the one who reaches the end first wins the game. Children in the experimental condition were told that on each turn, they would move their token the number of spaces indicated on the spinner. Also, they were required to say the number that they spun and the numbers on each of the squares they reached as they moved. For instance, if they were on 5 and they spun a "2", they would first say "two" then say "six", "seven" as they moved. In the control condition, children were told that

they would move their token to the next square with the color that matched the one they spun. Similar to the experimental group, they need to say the color they spun and the colors on the squares they reached as they moved.

After the participants played the game several times in each of four short sessions over the course of several weeks, they were tested a second time on the 4 numerical tasks mentioned previously. The experimental group demonstrated significant improvement, whereas the control group did not. The gains remained 9 weeks later in a follow-up session, in which the same tasks were tested a third time.

Here we propose a mechanistic account for the enhancement of numerical processing skills that was induced by playing the game. We hypothesize that multiple cognitive components (visual, language and spatially-guided action) are recruited and coordinated in the game environment. Particularly, the process of moving the token incrementally along the number board (motor) is coordinated with saying the next number word through the count list (verbal) as well as naming the printed numeral on each square tile on the board (visual). We suggest that the various components engaged in this process are also engaged in the basic numerical tasks, allowing learning occurring in the game to transfer to the basic tasks.

Another motivation for the current paper is that we want to address one of the common shortcomings of neural network models, which is their lack of flexibility in multi-task learning scenarios. In the current paper, we would like to show that as long as the model is composed of meaningful cognitive components that are also used in other tasks, it will be possible to demonstrate that training on one task could result in improvement on other tasks. The idea of constructing neural networks with multiple components that are responsible for different sub-tasks aligns with some recent advance in machine learning, e.g., neural networks composed of distinct modules have been used to solve language grounding problems (Andreas, Rohrbach, Darrell, & Klein, 2016; Johnson et al., 2017).

Below we first describe the architecture of our model, followed by the experiments we ran to simulate the learning in the number board game and the resulting learning effect. Finally, we present the implications of our results, some limitations of the current work and some future directions.

Model Architecture

There are three neural network modules in our model: the Visual component, the Language component (Figure 1) and the Action component (Figure 2).

Visual Component

The Visual component is composed of a pre-trained neural network named the ResNet (He, Zhang, Ren, & Sun, 2016) and two fully-connected readout layers. The ResNet is a deep neural network trained to recognize objects in ImageNet, a large image database of natural images with hand-annotated labels (Deng et al., 2009). The ResNet consists of stacked

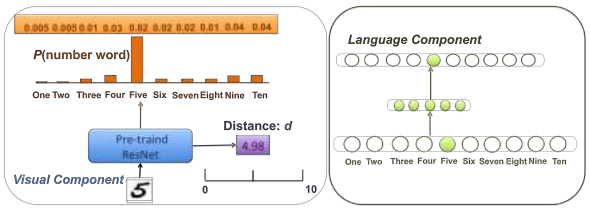


Figure 1: Illustration of the Visual component (left) and the Language component (right).

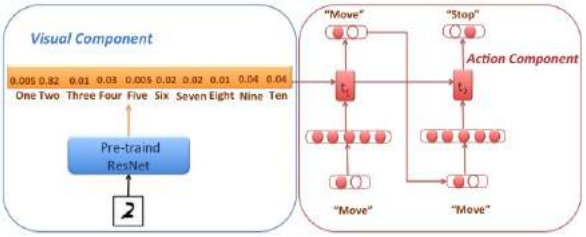


Figure 2: Illustration of the link between the number word output of the Visual component (left) and the Action component (right).

convolutional layers with non-linear activation functions and identity shortcut connections that skip one or more layers (see He et al. for details). We use the ResNet to process each image in our dataset and use the hidden activation of the last hidden layer as the feature vector of the image, i.e., as the image embedding. We then apply two fully-connected readout layers to the image embeddings, one *number word* readout layer and one *magnitude* readout layer. All of the images in our dataset are images of Arabic numerals ranging from 1 to 10 (see the Experiment Section for details). The *number word* readout layer is used to decode the numbers in the images. It outputs a probability distribution over the ten possible number words (one to ten). The *magnitude* layer is used to decode the magnitude of the number represented as a scalar, thought to be provided as a target for learning by the perceived distance of the number's location on the number line from zero. When simulating the number board game, we make the assumption that children attend to the board in two frames of reference, a global one and a local one: in the local frame of reference, they attend to and recognize the digit printed on the current square. In the global frame of reference, they keep track of how far their token has traveled from the "Start" point. These two processes are implemented by the *number word* readout layer and the *magnitude* readout layer, respectively. In the simulation of the number board game playing, these two layers were trained simultaneously. During the training we did not update the weights of the ResNet and only the connection weights of the two readout layers were updated. The equations of the Visual component can be written as follows:

$$
\begin{aligned}
e_i &= \text{ResNet}(I) \\
P(\text{number word}) &= \phi(W_{nw} e_i + b_{nw}) \\
m &= \text{ReLU}(W_m e_i + b_m)
\end{aligned} \tag{1}
$$

where $I \in \mathbb{R}^{28\times28\times3}$ is the raw pixels of the image, $e_i \in \mathbb{R}^{512}$ is the image embedding, $W_{nw} \in \mathbb{R}^{512\times10}$ ($W_m \in \mathbb{R}^{512\times1}$) are the connection weights from the embedding layer to the *number word* (*magnitude*) readout layer, $b_{nw} \in \mathbb{R}^{10}$ ($b_m \in \mathbb{R}^1$) are the biases of the *number word* (*magnitude*) readout layer. The softmax function $\phi(\boldsymbol{x})_i = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}}$ is used to normalize the logits of the *number word* readout layer in order to get the probabilities over number classes. The prediction is thus the one with the highest probability. The ReLU activation function is used to restrict the magnitude output to be non-negative, i.e., $\text{ReLU}(x) = \max(0, x)$.

Language Component

This component learns the successor function in counting, i.e., it predicts the next number in a count list given the current number (Figure 1, right). Particularly, during training, we use (current-number, next-number) pairs as data, e.g., ("five", "six"), and the model takes the current-number as input and predict the next-number. During testing, we iterate from zero, and the score is coded as correct up to the point of the first error. The number words are represented as one-hot vectors and they are fed to a fully-connected layer with rectifier activation function (ReLU), which is followed by a fully connected output layer with the softmax activation function. The mathematical formula for this module can be written as:

$$
\begin{aligned}
h_l &= \text{ReLU}(W_l e_w + b_l) \\
P(\text{next word}) &= \phi(W_n h_l + b_n)
\end{aligned} \tag{2}
$$

where $e_w \in \mathbb{R}^{10}$ are the word embeddings for the stimuli (one-hot vector), $W_l \in \mathbb{R}^{10\times10}$ are the connection weights from the input embedding layer to the hidden layer. We use the softmax function $\phi(x)$ to normalize the logits to get the probabilities over the possible number words.

Action Component

The Action component is a recurrent network that learns to output a sequence of "MOVE" actions before it outputs a "STOP" action (Figure 2, right). We use the Visual component described above to read the spinner, which shows either "1" or "2", to generate a probability distribution over all possible number words. This serves as the initial hidden state of the Action recurrent network. The Action network then takes the initial hidden state h_0 and the initial action a_0 to predict the next action, which is either "MOVE" or "STOP"[1], and the new hidden state and the predicted action are used in the next time step t. The mathematical formulae can be written as:

$$
\begin{aligned}
h_t &= h_{t-1} + ReLU(W_a e_a) \\
O_t &= \phi(W_m h_t + b_m)
\end{aligned} \tag{3}
$$

[1]In this board game the first action is always "MOVE", so there's no need to make a prediction for the first action.

Figure 3: Digits stimuli used in the current task. 1 - 9 are randomly selected from the MNIST dataset and the 10s are generated by randomly selecting 1s and 0s from the dataset and combine them (LeCun et al., 2010).

where $e_a \in \mathbb{R}^5$ are the embedding of the action, $W_a \in \mathbb{R}^{5\times10}$ are the connection weights from the embedding layer to the hidden layer. We use the softmax function to normalize the logits of the output layer to get the probabilities over the two possible actions ("MOVE" or "STOP").

Experiments

We use digit images from the MNIST database (LeCun, Cortes, & Burges, 2010), a dataset of handwritten digits with labels. We use a randomly selected subset of the original training data to construct our training dataset, which contains 10,000 samples (Figure 3), and our test dataset contains 10,000 samples that do not overlap with our training data.

Our simulations have two conditions that correspond to the experimental condition and the control condition in Ramani and Siegler's empirical study. In the experimental condition, there are three phases: the pre-test training, the number board game training and the post-test training. For simplicity, and because no numbers were used in Ramani and Siegler's control condition, the simulation control condition included only the pre-test training and the post-test training. In the pre-test training phase, the *number word* readout layer of the Visual component is trained on the numeral identification task for 34 batches, the *magnitude* readout layer is trained on the number line estimation task for 46 batches and the Language component is trained on the counting task for 16 batches. The numbers of batches for each task are chosen to produce the pre-test accuracies that approximate the ones in Ramani and Siegler's original paper.

Each batch contained 100 trials. The numerals were equally distributed in trials of each task. In counting, each trial involved a single transition from a "current-number" to the "next-number". We used backpropogation to train the network. In the board game training phase the models were trained on board game trials corresponding to individual turns in the game for 50 batches. We next describe in detail how all the cognitive components work together in a single trial. Assuming that the agent's token is at square "3" and the spinner yields "2", the modules will perform the following sequence of computations:

1. The Visual component reads the spinner ("2") and feeds the computed P(number word) to the Action component as its initial hidden state.

2a. The agent moves one step forward. The Visual component takes the image of the square where the agent's token is currently located ("4") and recognizes the number

printed on it, i.e., "four".
2b. The *magnitude* readout layer of the Visual component predicts the distance between the token's current location to zero, i.e., 4.0, as the supervision signal is available in the game.
3. The Action component takes the initial hidden state and the embedding of the initial action "MOVE" to predict the next action "MOVE".
4a. The agent moves one step forward. The Visual component takes the image of the square where the agent's token is currently located ("5") and recognizes the number, i.e., "five".
4b. The *magnitude* readout layer of the Visual component predicts the distance between the token's current location and the "Start" point, i.e., 5.0.
4c. The Language module takes the number that the Visual module recognized in the last time step ("four") to predict the current number, i.e., "five".
5. The Action component takes the last hidden state and the last action "MOVE" to output the next action "STOP".

To summarize, in this single trial, the Visual component needs to map the image of "4" to ("four", 4.0) and "5" to ("five", 5.0); the Language component needs to map "four" to "five" and the Action component needs to map the image of "2" (the spinner's output) to the action sequence "MOVE (given), MOVE, STOP". On trials where the spinner's output is 1, then only the steps 1, 2, 5 will be performed.

Finally, in the post-test training phase, in both conditions the neural modules are trained on one batch to simulate the learning experience gained during the 9 weeks between the immediate post-test and the follow-up test. This is termed as the "1-batch post-test training" simulation. This amount of training yields changes in performance that are comparable to the change from the post-test to the follow-up test in the control condition of Ramani and Siegler's experiment. To get a comprehensive understanding of the advantage of the experimental group, we also simulate another scenario in which we train the neural modules with the same number of batches as in the pre-test training phase ("*n*-batch post-test training" simulation). This gives us a sense of how the model will continue evolving if we train it on more than one batch, although it is unlikely that participants actually got that much training during the 9-week period of time in Ramani and Siegler's experiment.

When measuring the model performance on each numerical processing task, the Visual component is used to perform the numeral identification task and the number line estimation task using randomly selected digit images not used in training, and the Language component is used to perform the counting task, starting from 0, until an error is made or the whole count list is completed. To measure the performance of the number line estimation, we report the linearity of the estimation (measured by the square of the coefficient of correlation R^2 between the models prediction and the target) and the slope of number line estimation. These are the same dependent variables as measured in (Ramani & Siegler, 2008). A perfect number line estimate should yield a R^2 equal to 1.0 and a slope equal to 1.0. For the magnitude comparison task, we randomly select images of two different numbers from 1 to 10 and feed them to the Visual component separately. We then use the output of its *magnitude* readout layer to determine a response, i.e., the one with larger magnitude output is determined to be the "bigger number". The training and the testing were simulated 20 times with different random initialization of the network parameters and random sampling of the data.

Results

Accuracies of the Cognitive Components across the Training

We find consistent patterns in results across all the tasks that were tested (numeral identification, number line estimation, counting and magnitude comparison). As expected, up to the point when number board game was introduced, accuracy did not differ between the experimental condition and the control condition were. However, after the number board game was introduced, across all tasks the experimental condition demonstrated better performance than the control condition. Figure 4 shows the learning curves of different tasks in the "*n*-batch post-test training" simulation.

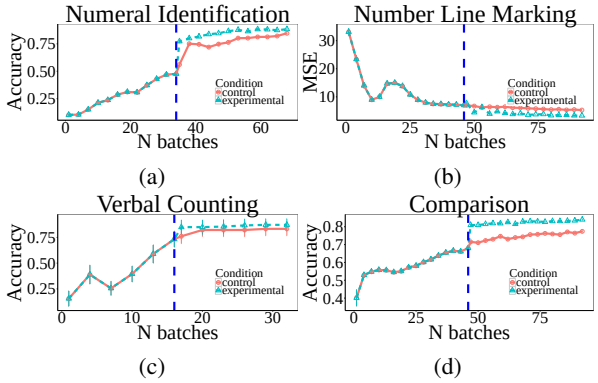


Figure 4: (a) Accuracy of the Visual component on the numeral identification task. (b) Mean Square Error of the Visual component on the number line estimation task. (c) Accuracy of the Language component on the counting task. (d) Accuracy of the magnitude comparison task. The vertical blue lines indicate the time points when number board game is introduced in the experimental condition.

Pre-test, Post-test and Follow-up Test

To compare our results with the results in Ramani and Siegler (2008), particularly their Figure 2, we also plot the pre-test scores, the post-test scores and the follow-up scores in the "1-batch post-test training" simulation. We run several paired *t*-tests to compare the pre-test scores and the post-test scores, as well as the pre-test scores and the follow-up test scores. In the control condition, as expected because

there is no training between the pre- and post-tests, the post-test scores are not significantly different than the pre-test scores. However, in the experimental condition, across all tasks the post-test scores are significantly higher than the pre-test scores (Table 1). The same results hold true for the comparison between the follow-up measures and the pre-test.

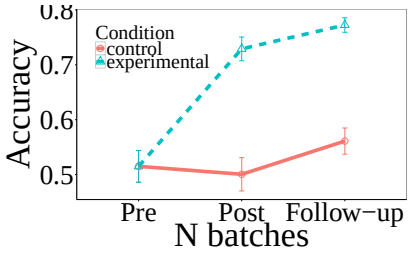


Figure 5: The model performance at the pre-test, the post-test, and the follow-up measures on the numeral identification task.

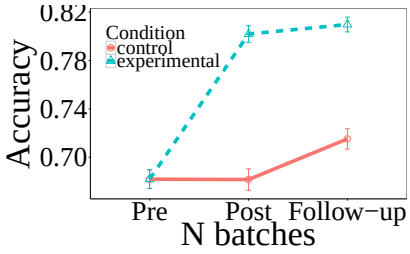


Figure 6: The model performance at the pre-test, the post-test, and the follow-up measures on the magnitude comparison task.

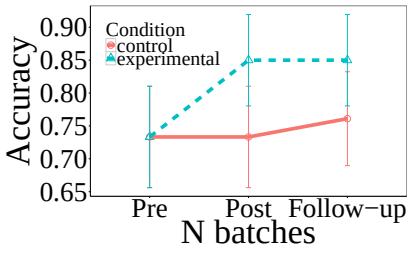


Figure 7: The performance of the Language component at the pre-test, the post-test, and the follow-up measures on the counting task.

As shown in Table 1, in the experimental condition, after the model is trained on the number board game, the numeral identification accuracy increases from 52% to 73%, the magnitude comparison accuracy increases from 68% to 80%, the counting performance increases from 73% to 85%, the square of the coefficient of correlation between the model's prediction and the target in number line estimation improves from 0.29 to 0.62, and the slope increases from 0.12 to 0.47. No improvement is observed in the control condition.

	Pre	Post	FU	Post - Pre	t_1	FU - Pre	t_2
Experimental Condition							
Numeral Identification	0.52	0.73	0.77	0.21***	5.91	0.26***	7.65
Magnitude Comparison	0.68	0.80	0.81	0.12***	12.66	0.13***	11.46
Counting	0.73	0.85	0.85	0.12*	2.25	0.12*	2.25
Number Line Linearity	0.29	0.62	0.64	0.33***	4.81	0.36***	5.99
Number Line Slope	0.12	0.47	0.46	0.35***	11.31	0.34***	13.97
Control Condition							
Numeral Identification	0.52	0.50	0.56	-0.01	-0.94	0.05*	2.41
Magnitude Comparison	0.68	0.68	0.72	0	-0.03	0.03***	3.56
Counting	0.73	0.73	0.76	0		0.03	1.75
Number Line Linearity	0.29	0.30	0.41	0.01	0.20	0.12	1.95
Number Line Slope	0.12	0.10	0.18	-0.01	-0.61	0.07*	2.99

*$p < .05$, **$p < .01$, ***$p < .001$.

Table 1: Mean scores of the different tasks at the pre-test (Pre, 1st column), the post-test (Post, 2nd column) and the follow-up test (FU, 3rd column); Differences of the scores between the post/follow-up test and the pre-test(Post-Pre/FU-Pre, 4th/6th column) and the corresponding t statistics (t_1/t_2, 5th/7th column) in the paired t-test (df=20).

Viewed from a different perspective of the data, we also show that although the performances in the two conditions do not differ at the pre-test, there is a significant difference between the two conditions at the post-test (Table 2) across all tasks. Such gap remains significant in all the follow-up measures except for the counting task.

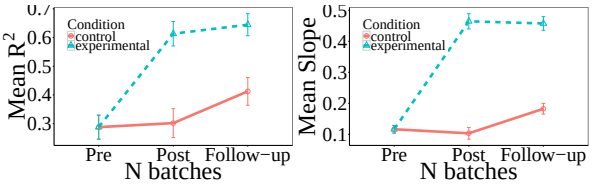
	Post	t_3	FU	t_4
Experimental - Control				
Numeral Identification	0.23***	6.06	0.21***	7.05
Magnitude Comparison	0.12***	10.09	0.09***	8.30
Counting	0.12*	2.25	0.09	1.90
Number Line Linearity	0.31***	4.34	0.23***	4.24
Number Line Slope	0.36***	11.56	0.28***	8.88

*$p < .05$, **$p < .01$, ***$p < .001$.

Table 2: Differences of scores between the experimental condition and the control condition at the post-test (Post, 1st column) and the follow-up test (FU, 3rd column) with their corresponding t statistics (the 2nd and the 4th column).

Linearity of the Number Line Estimation

One of the major motivations of Ramani and Siegler's work was to test whether playing the number board game could improve the linearity of children's number line estimation. As can be seen in Table 1 and Figure 8, in the experimental condition, both the linearity (measured by the square of the coefficient of correlation between the models prediction and the target) and the slope of number line estimation significantly increase after playing the number board game, and these learning outcomes are still significant at the follow-up test (Table 1).



(a) (b)

Figure 8: The linearity (a) and the slope (b) of the model performance in the number line estimation task evaluated at different stages.

Discussion

In the current study, we created a neural network model to provide a mechanistic account for the enhancement of numerical processing skills that was induced by playing a number board game (Ramani & Siegler, 2008; Ramani, Siegler, & Hitti, 2012). We hypothesized that various parts of the number board game actually train different components of the player that could later be used to perform other numerical tasks, such as numeral identification, magnitude comparison, counting and number line estimation. We reproduced the empirical learning effect in our computational model that implemented our hypothesis.

In our model, different cognitive components need to coordinate with each other to perform the task. For instance, as the Action component outputs the right actions, the agent constantly gets good teaching signals from the external environment which can be used by the Visual component and the Language component. As the Visual component recognizes the identities of the digits on the board, it further provides the language component with the proper inputs needed to predict the next number word. At the same time, as the agent's token moves along the board, the Visual component learns to associate the symbolic numeral with distance along the number line and the Visual and Language components jointly determine the number word to be uttered.

One limitation of the current paper is that we did not explicitly train the model's attention, e.g., we feed the Visual component with images of the current digits treating the "MOVE" action as simultaneously moving the player's token and shifting the focus of attention. We assume that the participants could allocate their attention to different parts of the environment and deploy their cognitive components in a synergistic way. Coordination of these processes to actually play the game requires executive control and working memory (Barnes et al., 2016). In future work we will explore how the selective attention and the executive control ability of the model can also be learned through training. This is a promising direction since recent advances in language grounding research in the artificial intelligence field have shown that neural network models can learn to attend different parts of an image in order to answer questions about the image (Yang, He, Gao, Deng, & Smola, 2016).

Another limitation is that we did not fully model the color board game control condition in Ramani and Siegler's work, i.e., we only modeled the aspect that participants did not receive any number-related training during the color board game playing, but we did not simulate the color-related training. In future work, we will fully model this color board game control condition. In addition to this control condition, in later studies researchers also compared the count-on procedure (reciting the number words for each new tile reached) used in Ramani and Siegler's study with the standard count-from-1 procedure (reciting the number words corresponding to the number of onward steps), and found that playing the same game using the standard procedure led to considerably less transfer to the other tasks (Laski & Siegler, 2014). This might occur because participants' attention was not directed either to the token position on the board or to the numerals on each square, as the task can be performed without this information. Also, the count-from-1 procedure provides no practice counting beyond the number two. It would be interesting to see whether modeling those control conditions within a network that learns to deploy its attention could provide a mechanistic account for why one of the interventions worked while the others did not.

In summary, the current work is a first step towards building a comprehensive computational model for numerical cognition that coordinates different modalities and integrates various training stimuli and paradigms. Our approach allows us to simulate the acquisition of number concepts as a process through which a set of component skills are assembled in different configurations in diverse task settings, promoting transfer across tasks that share components.

References

Andreas, J., Rohrbach, M., Darrell, T., & Klein, D. (2016). Neural module networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 39–48).

Barnes, M. A., Klein, A., Swank, P., Starkey, P., McCandliss, B., Flynn, K., ... Roberts, G. (2016). Effects of tutorial interventions in mathematics and attention for low-performing preschool children. *Journal of Research on Educational Effectiveness*, *9*(4), 577–606.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 248–255).

Halberda, J., Mazzocco, M. M., & Feigenson, L. (2008). Individual differences in non-verbal number acuity correlate with maths achievement. *Nature*, *455*(7213), 665.

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, *42*(1-3), 335–346.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770–778).

Johnson, J., Hariharan, B., van der Maaten, L., Hoffman, J., Fei-Fei, L., Zitnick, C. L., & Girshick, R. (2017). Inferring and executing programs for visual reasoning. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 3008–3017).

Lakoff, G., & Núñez, R. (2000). *Where mathematics comes from: How the embodied mind brings mathematics into being*. Basic Books.

Laski, E. V., & Siegler, R. S. (2014). Learning from number board games: You learn what you encode. *Developmental Psychology*, *50*(3), 853.

LeCun, Y., Cortes, C., & Burges, C. (2010). MNIST handwritten digit database. *AT&T Labs [Online]. Available: http://yann. lecun. com/exdb/mnist*, *2*.

Libertus, M. E., Feigenson, L., & Halberda, J. (2011). Preschool acuity of the approximate number system correlates with school math ability. *Developmental Science*, *14*(6), 1292–1300.

Park, J., & Brannon, E. M. (2014). Improving arithmetic performance with number sense training: An investigation of underlying mechanism. *Cognition*, *133*(1), 188–200.

Ramani, G. B., & Siegler, R. S. (2008). Promoting broad and stable improvements in low-income childrens numerical knowledge through playing number board games. *Child Development*, *79*(2), 375–394.

Ramani, G. B., Siegler, R. S., & Hitti, A. (2012). Taking it to the classroom: Number board games as a small group learning activity. *Journal of Educational Psychology*, *104*(3), 661.

Yang, Z., He, X., Gao, J., Deng, L., & Smola, A. (2016). Stacked attention networks for image question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 21–29).

Crossmodal Spatial Mappings as a Function of Online Relational Analyses?

Yordanka Zafirova (jovasileva@abv.bg)
Department of Cognitive Science and Psychology, New Bulgarian University,
21 Montevideo Street, 1618 Sofia, Bulgaria

Yolina Petrova (yolina.petrovaa@gmail.com)
Central and East European Center of Cognitive Science,
Department of Cognitive Science and Psychology, New Bulgarian University,
21 Montevideo Street, Sofia 1618, Bulgaria

Georgi Petkov (gpetkov@nbu.cogs.bg)
Central and East European Center of Cognitive Science,
Department of Cognitive Science and Psychology, New Bulgarian University,
21 Montevideo Street, Sofia 1618, Bulgaria

Abstract

Crossmodal correspondences are innate, language-based and statistically derived. They occur across all sensory systems and in different cultures. Despite their multiformity, they are exhibited analogously, mainly through robust congruency effects. One plausible explanation is that they rely on a common underlying mechanism, reflecting the fundamental ability to transfer relational patterns across different domains. We investigated the pitch-height correspondence in a bimodal sound-discrimination task, where the context of one relative sound pitch was changed online. The intermediate sound frequency was presented in successive blocks with lower *or* higher equidistant sounds and two squares at fixed up and down vertical positions. Congruency effects were transferred across sound contexts with ease. The results supported the assumption about the relational basis of the crossmodal associations. In addition, vertical congruency depended critically on the horizontal spatial representations of sound.

Keywords: crossmodal associations; relational mapping; pitch-height correspondence; SMARC effect

Introduction

Multidimensional information is integrated not only within the neural frame of one sensory system (e.g. Garner, 1974), but also across different modalities. Thus, certain features extracted from one perceptual realm interact with other, modality-specific attributes, and create coherent multimodal percepts, or intersensory Gestalts (for a review, see Spence, 2015). During the process of integration, particular aspects of the polysensory flow may modulate one another (like in McGurk & MacDonald, 1976, where visual stimuli modified auditory content, creating perceptual illusion), or bind together in bistable crossmodal entities with corresponding features. Examples of such corresponding features for pitch are shape (Melara & O'Brien, 1987; Walker et al., 2010); brightness (Marks, 1974; Martino &Marks, 1999; Melara, 1989); hue (Simpson, Quinn, Ausubel, 1956); smell (Belkin, Martin, Kemp, & Gilbert, 1997); size (Evans & Treisman, 2009; Mondloch & Maurer, 2004; Parise & Spence, 2012); height (Ben-Artzi & Marks, 1995; Chiou & Rich, 2012; Mudd, 1963; Patching & Quinlan, 2002; Rusconi, Kwan, Giordano, Umilta, & Butterworth, 2006) etc.

Apparently, invariable crossmodal associations might occur across all sensory systems. Examples of such associations are observed in different cultures (e.g. Bremner et al., 2013; Levitan et al., 2014; Parkinson, Kohler, Sievers, Wheatley, 2012; Wan et al., 2014) and are more or less implicit (Chen, Tanaka, Namatame, & Watanabe, 2016; Evans & Treisman, 2009; Parise & Spence, 2012). Given the broad scale and the diversity of the correspondence effect, at least three assumptions come to mind. First, these mappings should be nonrandom and in some way meaningful for perception. Second, they might comprise different processes and originate in different perceptual and cognitive networks. Third, they might be related to a common underlying mechanism, reflecting more general, inherent adaptive framework.

It was demonstrated that 4-month-olds are already sensitive to the associations between pitch and height, and pitch and sharpness (Walker et al., 2010). Lewkowicz and Turkewitz (1980) found evidence for mappings between brightness and loudness in infants 21 to 31 days of age. Along with that, crossmodal couplings are reported in nonhuman animals (see Ratcliffe, Taylor, & Reby, 2016, for a recent review). These findings are critical for the validity of the notion that multisensory associations are semantically mediated. It seems that they emerge on a lower, perceptual level and congruency effects are dependent on available attentional resources. More specifically, "attention is likely to play an important role in cross-modal perceptual organization" (Spence, 2015, p. 12). At the same time, it might be argued that even newborns have enough experience with environment, given the fact that they are

extremely sensitive to certain statistical frequencies. In fact, 6-month-olds were better than adults in extracting implicit crossmodal information from the context (Rohlf, Habets, von Frieling, & Röder, 2017). Additionally, it was suggested that human perceptual system might foster the development of language by auxiliary crossmodal mappings of speech sounds to other percepts, for example shapes (Ozturk, Krehm, & Vouloumanos, 2013). The opposite is also true – linguistic experience systematically promotes progressive coupling of nonlinguistic dimensions (Martino & Marks, 1999). In short, the importance of language in crossmodal perception is undeniable. However, not all data can be explained with semantics (consider the animal studies). By all appearances, perceptual constraints, environment and language contribute synchronously to the establishment of stable crossmodal links. According to Spence (2011) certain structural correspondences might be innate, while statistical and semantic ones are evidently learned, and all are a function of environment.

A basic tool for exploring the nature and the strength of crossmodal associations is the speeded discrimination task, developed by Garner (1974), where participants respond to one task-relevant dimension while ignoring another, task-irrelevant dimension. Corresponding dimensions are integrated and trigger congruency effects that influence selective attention and performance. Apart from that, particular dimensions might interact on different levels. There is evidence that certain crossmodal couplings might actually prompt perceptual change (e.g. Evans & Treisman, 2009; McGurk & MacDonald, 1976). Others communicate on decisional level (e.g. Melara, 1989; Rusconi et al., 2006) or result from semantic inconsistencies (e.g. Martino & Marks, 1999). Notably, they can be correlated directly – for instance, psychophysical dimensions, like pitch and height usually covary in magnitude. On the other hand, it is generally accepted that these mappings are relative – one level of the first dimension can be mapped on different levels of the second dimension, depending on the context (for a review, see Spence, 2011). However, the mechanisms behind these relative mappings remain unclear.

In short, crossmodal correspondences engage all sensory systems, they are universal, innate, automatic, can be learned and assist learning, and interact on lower, bottom-up and higher, top-down levels. Considering the relative nature of the mappings, it is plausible to assume that they recruit the mechanisms of another, major cognitive process, reflecting the ability to build and compare relations. What is more, "the ability to pick out patterns, to identify recurrences of these patterns despite variation in the elements that compose them, to form concepts that abstract and reify these patterns, and to express these concepts in language" is considered a fundamental core of cognition (Holyoak, Gentner, & Kokinov, 2001, p. 2). Besides, relational analyses can be performed online and automatically, with or without utilizing attentional resources. Evidence is piling up that relations can be retrieved unconsciously and transferred across corresponding sets of data (Hristova, 2017; Li, Li, Zhang, Shi, & He, 2018). That being said, the capacity to construct and compare associations across modalities is one possible explanation for the pervasiveness of crossmodal correspondences. Thus, perceptual dimensions are represented not in their absolute values but as correlated dyads. To check this hypothesis, we investigated the congruency effects between one sound frequency and two vertical spatial positions – higher and lower. In other words, we measured the interaction between pitch and height in a speeded sound-discrimination task. Crucially, the context of the sound was changed during the task – it was presented with either higher or lower pitch, so that it was perceived as relatively lower or higher than the other sound. Previous studies demonstrated that pitch and height were positively correlated – higher frequencies were consistently associated with higher vertical positions (e.g. Ben-Artzi & Marks, 1995; Evans & Treisman, 2009; Rusconi et al., 2006 etc.). If crossmodal correspondences are indeed represented as relations, are we should expect comparable congruency effects in both contexts – i.e., one and the same sound should be mapped to different vertical positions in accordance with its relative frequency. Importantly, this shift should be effortless and almost instantaneous.

Experiment: Sound-Discrimination Task

Method

Participants 24 students (7 males) with mean age 23.8 years (standard deviation SD=6.3) from the Psychology Department were recruited for the task, after approval from the Cognitive Science and Psychology Ethics Committee. All signed the informed consent form and reported no problems performing the task.

Stimuli and design The stimuli were three sinusoidal sound waves with different frequencies, and a black square presented at two vertical positions. The sounds were generated on *Audacity* at 600 Hz, 900Hz and 1200Hz; 16 bits, mono, on an amplitude level of -2.5 dBFS (decibels relative to full scale) and duration 1000 ms. The square was solid black, 100x100 pixels (px) JPEG image. Participants had to perform a sound-discrimination task with bimodal presentation of the stimuli – i.e. the values of the task-relevant (sound pitch) and task-irrelevant (square height) dimensions were coupled randomly for each trial and presented simultaneously in both visual and auditory modalities.

Procedure The experiment was conducted in the presence of the experimenter in one of the booths of the Experimental Psychology Laboratory. Presentation and timing were controlled by the E-prime software (Schneider, Eschman, & Zuccolotto, 2012) and the multifunctional USB-based stimuli and response device *Chronos* which recorded accuracy and response times (RTs) with 1 ms resolution

(*Chronos operator manual*, 2015). The sounds in the experiment were presented via *Chronos* accessory headset.

The experiment started with onscreen instruction about the task requirements. Participants were asked to keep their eyes on the screen (as cued by the fixation cross) and to respond by pressing one button for the *thick* sound and another button for the *squeaky* sound, with the index finger of their dominant hand. The buttons of the response box (the first and the third, with the finger resting on the middle one between trials) were counterbalanced across participants. The words *low* and *high* were avoided in the instruction because of possible semantic priming. The sounds were presented in pairs – 600/900 Hz or 900/1200 Hz, in two continuous blocks with a pair change in between. Crucially, participants were informed that at one point the sounds would be changed but they should continue performing the same task. The experimenter urged the participants to look at the screen and monitored the execution of the task.

Each trial started with 500 ms fixation cross (*Consolas*, 22 pt, black, on a white background). Then participants heard one of two sounds – lower or higher, presented pseudorandomly, no more than four of the same pitch successively (to avoid motor fluency). The sounds were randomly coupled with a black square, presented on a white background below or above fixation – at 20% or 80% along the vertical midline of a 24-inch monitor with 60 Hz refresh rate and resolution 1920x1080 px (at 7.5 cm below or above the center of the display). Each pitch was accompanied by an equal number of high and low squares. As the viewing distance was approximately 60 cm, the side of the square corresponded to 2.5°-3° horizontal visual angle. The sound and the square were presented simultaneously (bimodally) for 1000 ms or until response, and were followed by 1500 ms intertrial interval (white screen). The sequence of one trial is represented in Figure 1.

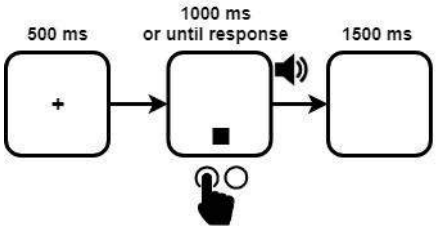


Figure 1: Schematic representation of one trial, not to scale.

For all participants the experiment started with one sound pair in the first block and ended with another sound pair in the second block, in counterbalanced order. There were 216 experimental trials. During the first 104 trials some participants responded to 600/900 Hz, and other – to 900/1200 Hz, and vice versa for the rest of the trials. The first 8 trials after the sound change were treated as practice and were analyzed separately. That way, the two blocks consisted of 104 trials each. Additionally, there were 8 practice trials before the experimental part, always with the sound pair of the following block. These trials were excluded from the analyses. Thus, there were 224 trials overall – 8 practice trials separated with a break from the experimental part, and 216 experimental trials (104 before the sound change, 8 practice trials after the sound change and 104 trials with the second sound pair) with no break. The experiment lasted about ten minutes.

Results

First, the overall accuracy was assessed (.94, range .83-1). No participant was excluded on that account. As 900 Hz was the pitch of interest, *only* responses to that pitch were considered in the subsequent analyses. Then, accuracy and RT were aggregated by trial in order to examine the nature of transition between the two contexts. The progress of the values over time is visualized on Figure 2.

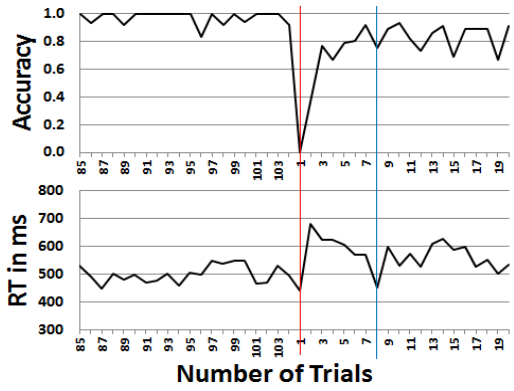


Figure 2: Accuracy and RT around context change, responses to relative pitch only (900 Hz), by trial. The red line indicates the point of change. The blue line marks the end of the online practice trials.

As participants had eight trials of practice before the experimental part, the first eight trials after the context change were also considered as practice and disregarded. Accuracy was analyzed as a function of experimental block and pitch-height congruency. There was no main effect of congruency – no difference in accuracy for congruent and incongruent responses ($F(1,23)=2.32$, $p=.142$; $\eta^2_p=.092$). However, there was main effect of experimental block ($F(1,23)=8.12$, $p=.009$; $\eta^2_p=.261$), i.e. more accurate responses in the first block; and an interaction between experimental block and congruency ($F(1,23)=5.24$, $p=.032$; $\eta^2_p=.185$). Newman-Keuls post-hoc revealed difference between the incongruent trials of the second block and the congruent and incongruent trials of the first block ($p<.001$), between the congruent trials of the first block and the second block ($p=.045$) and between the congruent and incongruent trials of the second block ($p=.008$). That is, participants made more mistakes in the second part of the experiment, especially in the incongruent trials.

There were a total of 5.6% errors in the experimental trials (.08% with no RT) which were also excluded. The data were then trimmed with 2.5 SDs from the subject means per condition (another 2.8%). Only trimmed data from the experimental blocks, concerning the correct responses to the relative pitch (900 Hz) were examined for differences in response times. The analyses were again performed with experimental block and pitch-height congruency as within factors. Means and standard deviations per condition are presented in Table 1.

Table 1: Descriptive statistics: mean RT (SD) per condition

First block		Second block	
Congruent	Incongruent	Congruent	Incongruent
512 (110)	529 (119)	497 (96)	522 (83)

There was no main effect of experimental block ($F(1,23)=.761$, $p=.392$; $\eta^2_p=.032$). At the same time, in accordance with the presumed crossmodal interaction between pitch and height, there was a substantial congruency effect ($F(1,23)=5.65$, $p=.026$; $\eta^2_p=.197$). Crucially for our hypothesis, there was *no* interaction between congruency and experimental block ($F(1,23)=.533$, $p=.473$; $\eta^2_p=.023$) – in other words, congruent responses were faster regardless of the relative sound context. Mind that only the context of the 900 Hz sound was changed between the blocks. And yet, in the presence of the higher pitch (1200 Hz) it was mapped to the lower vertical position, and subsequently to the higher vertical position in the presence of the lower pitch (600 Hz). Figure 3 illustrates the remapping of the sound across the two experimental blocks.

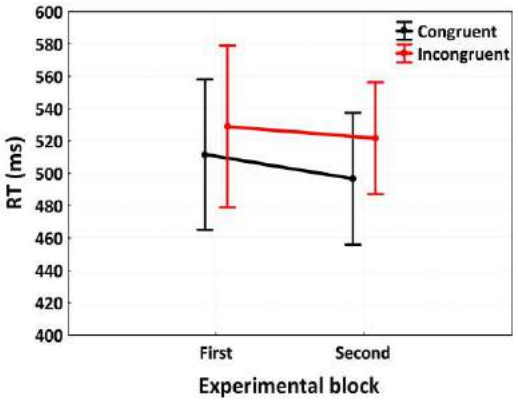


Figure 3: Transfer of vertical congruency effects across different sound contexts in the experimental blocks. Vertical bars denote .95 confidence intervals.

Due to the nature of the experiment, additional analyses were performed to account for alternative explanations. In the bimodal sound-discrimination task, sound pitch is coupled with visual stimuli presented along the vertical dimension. However, crossmodal correspondences were already demonstrated between pitch and the horizontal space (e.g. Rusconi et al., 2006). More specifically, lower pitch was mapped more readily to the left, and higher pitch was mapped more readily to the right. Our participants responded by pressing a left or a right button for the lower or higher pitch, in counterbalanced order. That is, for one half of the participants the sounds were horizontally congruent (they always pressed the left button for the lower pitch), while the other responded in horizontally incongruent manner.

To estimate the possible interaction between horizontal and vertical congruency, the mapping of the response was added as a categorical predictor in the above analyses. There was an interaction between horizontal and vertical congruency for accuracy ($F(1,22)=5.75$, $p=.025$; $\eta^2_p=.207$) and for RT ($F(1,22)=25.13$, $p<.001$; $\eta^2_p=.533$) (Figure 4).

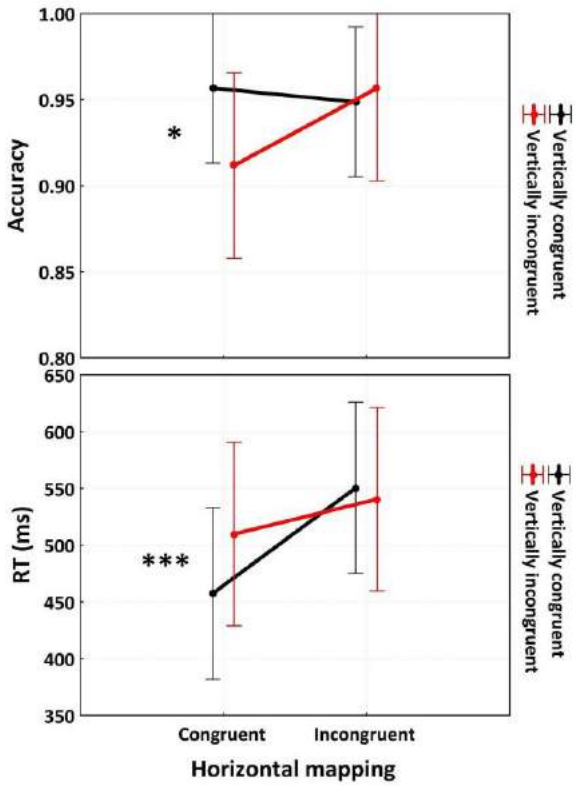


Figure 4: Interaction between horizontal and vertical congruency for accuracy (up) and RT (down). Vertical bars denote .95 confidence intervals.
* $p<.05$; *** $p<.001$

Newman-Keuls post-hoc revealed difference between the vertically congruent and incongruent trials ($p=.041$ for accuracy; and $p<.001$ for RT), but *only* when the responses were horizontally congruent. Remarkably, there was no difference when responses were mapped incongruently.

Discussion

On the whole, we did find the typical pitch-height congruency effects (Ben-Artzi & Marks, 1995; Chiou & Rich, 2012; Evans & Treisman, 2009; Patching & Quinlan, 2002; Rusconi et al., 2006 etc.). That way, we supported experimentally one of the major claims in the field – that the mappings between two dimensions are relative. For example, Gallace and Spence (2006) reported similar effects in a bimodal size-discrimination task with task-irrelevant sound frequencies, relatively mapped to the values of the visual stimuli. In another study, Smith, Grabowecky and Suzuki (2007) presented participants with the same visual stimuli (androgynous faces) accompanied by a sound in the male or female frequency speaking range, and found differences in perception of gender. Here, we demonstrated online remapping of one intermediate pitch to lower and higher vertical spatial positions in a speeded sound-discrimination task. Note that the intermediate pitch was presented throughout the whole experiment, but in a different context – with equidistant lower and subsequently higher pitch for one half of the participants, and the other way around for the other half of the participants. As follows, participants had to change their response as well – those who responded to the lower pitch with the *thick* button had to remap the same sound to the *squeaky* button, and vice versa. Moreover, in accordance with the expected congruency effects, responses to the same pitch were faster and more accurate when it was coupled with a square in the lower or higher visual field and perceived as *thick* or *squeaky*, respectively.

Crucially, the crossmodal correspondence between pitch and vertical space depended on the correspondence between pitch and horizontal space. Rusconi and colleagues (2006) provided conclusive evidence for explicit and implicit vertical and horizontal spatial mappings of pitch (but see Pitteri, Marchetti, Priftis, & Grassi, 2017). In their experiments responses were gathered vertically and horizontally, and participants performed the tasks with crossed and uncrossed hands. Higher pitch was mapped to upper and right buttons, and lower pitch was mapped to lower and left buttons (the so-called SMARC effect), even when responses did not require explicit processing of pitch, as in a wind vs. percussion sounds discrimination task. This implies that pitch-height correspondences are not solely semantically modulated, as horizontal mapping of sound is not linguistically promoted. Crossmodal correspondences depend mostly on failures in selective attention, especially within the speeded discrimination task (Spence 2011, 2015). When the task-irrelevant dimension is visuospatial, we might expect interaction between generated spatial and response codes (see Lu & Proctor, 1995, for a review of Simon and spatial Stroop effects). Interaction was reported also for mental representations and responses in horizontal space (Dehaene, Bossini, & Giraux, 1993). In our task, responses were gathered horizontally, while visual stimuli were presented along the vertical axis. And yet, we found substantial interaction between horizontal mental representation of sound and response side (unlike Pitteri et al., 2017, who reported the same effect for pitch and brightness, but only for musicians). It can be speculated that sound is represented both horizontally and vertically. That way, mentally generated horizontal and vertical spatial codes interact with stimuli-generated vertical spatial codes and modulate the crossmodal congruency effect. As a result, horizontal congruency emerged as an essential prerequisite for vertical congruency effects.

In addition, it seems that crossmodal mappings happen automatically and effortlessly. Our results are in line with previous findings, demonstrating that relations can be retrieved unconsciously and transferred across domains. (Hristova, 2017; Li, Li, Zhang, Shi, & He, 2018) Thus, as Holyoak, Gentner and Kokinov (2001) pointed out, the ability to manipulate relations might be basic for cognition.

That being said, the experiment is a beginning of a larger experimental work within the field of crossmodal correspondences. The hypothesis about the online relational analyses should be explored further. Additional experimental settings should investigate whether similar associations exist among isolated features, or are integrated in larger cognitive frameworks. Another major challenge would be to outline the dissimilarities between given crossmodal mappings and relating them to other forms of associations.

Acknowledgements

This research was supported financially by the European Office for Aerospace Research and Development under grant FA9550-15-1-0510 (Anticipating Future by Analogy-Making).

References

Belkin, K., Martin, R., Kemp, S. E., & Gilbert, A. N. (1997). Auditory pitch as a perceptual analogue to odor quality. *Psychological Science*, *8*(4), 340-342.

Ben-Artzi, E., & Marks, L. E. (1995). Visual-auditory interaction in speeded classification: Role of stimulus difference. *Perception & Psychophysics*, *57*(8), 1151-1162.

Bremner, A. J., Caparos, S., Davidoff, J., de Fockert, J., Linnell, K. J., & Spence, C. (2013). "Bouba" and "Kiki" in Namibia? A remote culture make similar shape–sound matches, but different shape–taste matches to Westerners. *Cognition*, *126*(2), 165-172.

Chen, N., Tanaka, K., Namatame, M., & Watanabe, K. (2016). Color-Shape Associations in Deaf and Hearing People. *Frontiers in psychology*, *7*. Retrieved from https://doi.org/10.3389/fpsyg.2016.00355

Chiou, R., & Rich, A. N. (2012). Cross-modality correspondence between pitch and spatial location modulates attentional orienting. *Perception*, *41*(3), 339-353.

Chronos operator manual. (2015). Pittsburgh, PA: Psychology Software Tools, Inc.

Dehaene, S., Bossini, S., & Giraux, P. (1993). The mental representation of parity and number magnitude. *Journal of Experimental Psychology: General*, *122*(3), 371-396.

Evans, K. K., & Treisman, A. (2009). Natural cross-modal mappings between visual and auditory features. *Journal of Vision*, *10*(1), 6:1-12.

Fastl, H. (2004). Audio-visual interactions in loudness evaluation. *Proceedings of the 18th International Congress on Acoustics, Kyoto, Japan* (pp. 1161–1166).

Gallace, A., & Spence, C. (2006). Multisensory synesthetic interactions in the speeded classification of visual size. *Perception & Psychophysics*, *68*(7), 1191-1203.

Garner, W. R. (1974). *The processing of information and structure*. Hillsdale, NJ: Erlbaum.

Holyoak, K. J., Gentner, D., & Kokinov, B. N. (2001). Introduction: The place of analogy in cognition. In D. Gentner, K. J. Holyoak & B. N. Kokinov (Eds.), *The analogical mind: Perspectives from cognitive science*, (1-19). Cambridge, MA: MIT press

Hristova, P. (2017). Can we think unconsciously via analogy?. In Z. Radman (Ed.), *Before consciousness: In search of the fundamentals of mind*, (270-295). Exeter, UK: Imprint Academic.

Levitan, C. A., Ren, J., Woods, A. T., Boesveldt, S., Chan, J. S., McKenzie, K. J., ... & van den Bosch, J. J. (2014). Cross-cultural color-odor associations. *PloS One*, *9*(7), e101651.

Lewkowicz, D. J., & Turkewitz, G. (1980). Cross-modal equivalence in early infancy: Auditory–visual intensity matching. *Developmental Psychology*, *16*(6), 597-607.

Li, J., Li, X., Zhang, X., Shi, K., & He, Y. (2018). Can unconscious thought detect relational similarities?. *International Journal of Psychology*, 1-7. doi:10.1002/ijop.12550

Lu, C. H., & Proctor, R. W. (1995). The influence of irrelevant location information on performance: A review of the Simon and spatial Stroop effects. *Psychonomic Bulletin & Review*, *2*(2), 174-207.

Marks, L. E. (1974). On associations of light and sound: The mediation of brightness, pitch, and loudness. *The American Journal of Psychology*, 173-188.

Martino, G., & Marks, L. E. (1999). Perceptual and linguistic interactions in speeded classification: Tests of the semantic coding hypothesis. *Perception*, *28*(7), 903-923.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746-748.

Melara, R. D. (1989). Dimensional interaction between color and pitch. *Journal of Experimental Psychology: Human Perception and Performance*, *15*(1), 69-79.

Melara, R. D., & O'Brien, T. P. (1987). Interaction between synesthetically corresponding dimensions. *Journal of Experimental Psychology: General*, *116*(4), 323-336.

Mondloch, C. J., & Maurer, D. (2004). Do small white balls squeak? Pitch-object correspondences in young children. *Cognitive, Affective, & Behavioral Neuroscience*, *4*(2), 133-136.

Mudd, S. A. (1963). Spatial stereotypes of four dimensions of pure tone. *Journal of Experimental Psychology*, *66*(4), 347-352.

Ozturk, O., Krehm, M., & Vouloumanos, A. (2013). Sound symbolism in infancy: evidence for sound–shape cross-modal correspondences in 4-month-olds. *Journal of Experimental Child Psychology*, *114*(2), 173-186.

Parise, C. V., & Spence, C. (2012). Audiovisual crossmodal correspondences and sound symbolism: a study using the implicit association test. *Experimental Brain Research*, *220*(3-4), 319-333.

Parkinson, C., Kohler, P. J., Sievers, B., & Wheatley, T. (2012). Associations between auditory pitch and visual elevation do not depend on language: Evidence from a remote population. *Perception*, *41*(7), 854-861.

Patching, G. R., & Quinlan, P. T. (2002). Garner and congruence effects in the speeded classification of bimodal signals. *Journal of Experimental Psychology: Human Perception and Performance*, *28*(4), 755-775.

Pitteri, M., Marchetti, M., Priftis, K., & Grassi, M. (2017). Naturally together: pitch-height and brightness as coupled factors for eliciting the SMARC effect in non-musicians. *Psychological Research*, *81*(1), 243-254.

Ratcliffe, V. F., Taylor, A. M., & Reby, D. (2016). Cross-modal correspondences in non-human mammal communication. *Multisensory Research*, *29*(1-3), 49-91.

Rohlf, S., Habets, B., von Frieling, M., & Röder, B. (2017). Infants are superior in implicit crossmodal learning and use other learning mechanisms than adults. *eLife*, *6*, e28166.

Rusconi, E., Kwan, B., Giordano, B. L., Umilta, C., & Butterworth, B. (2006). Spatial representation of pitch height: the SMARC effect. *Cognition*, *99*(2), 113-129.

Schneider, W., Eschman, A., & Zuccolotto, A. (2012). *E-Prime: User's guide.* Pittsburgh, PA: Psychology Software Tools, Inc.

Smith, E. L., Grabowecky, M., & Suzuki, S. (2007). Auditory-visual crossmodal integration in perception of face gender. *Current Biology*, *17*(19), 1680-1685.

Spence, C. (2011). Crossmodal correspondences: A tutorial review. *Attention, Perception, & Psychophysics*, *73*(4), 971-995.

Spence, C. (2015). Cross-modal perceptual organization. In J. Wagemans (Ed.), *The Oxford handbook of perceptual organization* (649-664). Oxford, UK: Oxford University Press.

Simpson, R. H., Quinn, M., & Ausubel, D. P. (1956). Synesthesia in children: Association of colors with pure tone frequencies. *The Journal of Genetic Psychology*, *89*(1), 95-103.

Walker, P., Bremner, J. G., Mason, U., Spring, J., Mattock, K., Slater, A., & Johnson, S. P. (2010). Preverbal infants' sensitivity to synaesthetic cross-modality correspondences. *Psychological Science*, *21*(1), 21-25.

She Helped Even Though She Wanted to Play: Children Consider Psychological Cost in Social Evaluations

Xin Zhao & Tamar Kushnir
Cornell University

Abstract

Sometimes we incur a high psychological cost (for example, forgo something we really like) in order to fulfill social or moral obligations. How would the information of incurring psychological costs influence children's social evaluations? Prior work suggests that children do not recognize the virtue of resolving inner conflicts until age 8. In two studies, we de-confounded costs from inner conflicts and found that when the difficulty was not explicitly stated as having conflicting desires (a self-interested desire and a moral desire) at once, most 8- to 9-year-olds and some 6 to 7-year-olds gave adult-like favorable evaluations of the character who overcame psychological or physical difficulty to act morally. Moreover, neither adults nor children inferred conflicting moral and personal desires spontaneously. These together suggest that children's evaluation of moral virtue depends on understanding of cost rather than conflict: Physical cost is incorporated early in development, and psychological cost later.

Keywords: cognitive development, social cognition, moral development, moral cognition, costs

Introduction

Suppose that you ask two of your friends to help you with a paper you have to finish tonight; at the same time there is a really good show on tv. One of your friends really likes this show. The other friend does not have any interest in the show at all. If each one of these friends offered to help you with your paper, would you evaluate their actions towards you differently? Even if both friends ended up helping you, the one who gave up watching her favorite show incurred a higher psychological cost to do so, and intuitively this might lead us to evaluate her as nicer, kinder, perhaps a better friend. The costliness of her choice to help seems to weigh heavily in our evaluation. We investigate children and adults' intuitions about psychological cost as it relates to moral status in the current studies.

The ability to make social evaluations about others develops early in childhood (Hamlin, Wynn, & Bloom, 2007; Hamlin, Wynn, Bloom, & Mahajan， 2011; Burns & Sommerville, 2014; Geraci & Surian, 2011; Sloane, Baillargeon, & Premack, 2012; Olson & Spelke, 2008). Even infants and young children prefer someone who helps another person fulfill a goal (e.g., climbing a mountain or opening a box) over someone who hinders another person from goal completion (e.g., Hamlin et al., 2007) and prefer someone who shares equally with others over someone who does not share equally (e.g., Olson & Spelke, 2008). This research has mainly focused on comparing actions that bring about different outcomes (usually a positive outcome vs. a negative outcome). By preschool age, children consistently consider the intention behind an action even when it is inconsistent with its outcome (e.g., attempted or innocent harm; see Baird & Astington, 2004; Cushman, Sheketoff, Wharton, & Carey, 2013; Killen, Mulvey, Richardson, Jampol, & Woodward, 2011). Prior work suggests a link between the development of intent-based social evaluation and theory of mind (Killen et al., 2011; Smetana et al., 2012).

Previous research has examined young children's consideration of costs in their inferences of individual's goals and preferences. For example, infants consider the cost that someone expends to achieve a goal when making inferences on how much the agent values the goal. After seeing someone achieve two goals one at a larger cost than the other (e.g. has to jump over a higher barrier), infants expect her to value the goal that incurs a larger cost more than the other goal (Liu, Ullman, Tenenbaum, & Spelke, 2017). Similarly, toddlers are more likely to exonerate a non-helper for whom helping would have been hard than someone for whom helping would have been easy (Jara-Ettinger, Tenenbaum, & Schulz, 2015). Preschoolers even consider the cost they themselves incur to share with others in interpreting if their own actions are prosocial (Chernyak & Kushnir, 2013, 2018).

To date, studies of young children's evaluation of agents' psychological or moral status based on cost have focused on tangible goods - physical obstacles such as distance or barriers or valuable resources such as toys or stickers. Our initial example of the friend who gives up her favorite tv show is both like and unlike these cases. It is like resource sharing because the tv show can be thought of as having value, like stickers or toys. However, it is unlike resource sharing in that the value is intangible rather than tangible, a mental state rather than an object. Less is known about how children's understanding of this, more psychological, type of cost plays a role in their social evaluations.

Several pieces of evidence suggest that understanding psychological cost may be challenging for young children. First, one recent study (Starmans & Bloom, 2016) looked at children's evaluation of inner moral conflicts. In this study, children of 3 to 8 years old and adults were asked to compare two characters who both ultimately acted morally, but one acted morally without experiencing inner conflict, while the other resolved an inner conflict between a self-interested desire and a moral desire in order to act morally. Starmans & Bloom (2016) found that although adults evaluated the conflicted character more favorably than the unconflicted character, children of 3 to 8 years old showed the opposite evaluation. This result shows that children do not recognize the moral virtue of resolving inner conflicts until after age 8. However, it leaves open the question of whether the conflict

itself was difficult for children to understand (having both a moral and selfish desire at once), or the psychological cost was difficult to understand (forgoing something one likes in order to act morally).

Second, much recent evidence has shown that during early and middle childhood children increasingly recognize the possibility and positivity of overcoming immediate self-interested desires. For example, between 4 and 7, children increasingly believe that one can choose to act contrary to personal desires (e.g., Kushnir et al., 2015). Children also increasingly predict that individuals will act against personal desires (e.g. play) to comply with moral rules (e.g. help brother) and would feel good about it (Lagattuta, 2005; Lagattuta, Nucci, & Bosacki, 2010). Similarly, they also increasingly predict that an individual will act towards higher-order goals (e.g. doing homework) rather than succumbing to immediate desires (e.g. watching cartoons) (Yang & Fyre, 2018) Therefore, it is likely that, during early and middle childhood, as children view forgoing immediate self-interested desires to be possible and positive, they may increasingly favorably evaluate someone who endures high psychological cost to do the right thing.

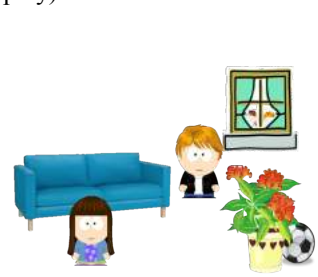
In two studies, we investigate how information about psychological costs affects children's social evaluations. Our first research question was, at what age can children evaluate someone who incurs higher psychological costs to fulfill social or moral obligations as more virtuous? In Study 1, we asked children and adults to compare two characters who ultimately did the right thing, but one incurred a larger psychological cost (i.e., forewent something she really likes) in order to do the right thing, while the other incurred a smaller psychological cost (i.e., forwent something she does not like). We closely followed the procedure of Starmans & Bloom (2016) but, importantly, we removed expressions of inner conflict from the procedure by mentioning moral actions without stating moral desires. We focused on children of 4 to 9 years old. Our second research question was how children make inferences on the agents' moral desires based on the information on psychological costs incurred to perform the moral action. Thus, after asking children to make evaluations, we also asked children to make inferences about the unstated moral desires of each character. Our final question was whether children's social evaluations may differ by the types of costs. Thus, in Study 2, we tested how children's evaluation of incurring psychological cost compare to their understanding of incurring physical cost.

Study 1

Method

Seventy-six 4- to 9-year-olds (4.02- 7.98, M = 5.80, SD = 1.06, 41 boys) from Ithaca, New York were recruited for this study. Mirroring the procedure in Starmans & Bloom (2016), we divided the children into three age groups: 4- to 5-year-olds, 6- to 7-year-olds, 8- to 9-year-olds. Specifically, 39 4- to 5-year-olds (4.02- 5.85, M = 4.99, SD = .52, 21 boys), 37 6- to 7-year-olds (6.00 - 7.98, M = 6.94, SD = .64, 17 boys) and 24 8- to 9-year-olds (8.03 – 9.65, M = 8.84, SD = .55, 11 boys) were included in the analyses. In addition, 92 adults were recruited from Amazon Mechanical Turk.

Each child was read four pairs of stories and shown accompanying pictures adapted from Starmans & Bloom (2016). See Figure 1 for an example of the stories. Each pair of stories described two characters who both performed a good action (e.g. helping her brother). One character (i.e., the "high psychological cost" character) incurred a higher psychological cost and forewent something she really liked in order to perform the good action. The other character (i.e., the "low psychological cost" character) incurred a lower psychological cost and forewent something she did not really like. Two story items (one Helping Story about helping siblings, one Honesty Story about telling truth to mom) were adapted from Starmans & Bloom (2017) and concerned moral obligations. We added two other pairs of stories about following rules (one Dishes Story about cleaning up dishes as mom asks, one Toys Story about playing the toy mom asks to play).



"High Psychological Cost" Character:
This is Ana, and this is Ana's brother. Ana's brother lost his favorite ball. He has been looking for it really hard, but he still hasn't found it. So he asked Ana to help him look for it. Helping people is the right thing to do. Ana's friends are playing right outside. Ana likes playing outside. She wants to go play outside right now.
But Ana helped her brother look for his ball. It was really hard for Ana to help her brother because she wanted to play outside. But Ana helped her brother anyway.

"Low Psychological Cost" Character:
This is Tiffany, and this is Tiffany's brother. Tiffany's brother lost his favorite ball. He has been looking really hard, but he still hasn't found it. So he asked Tiffany to help him look for it.
Helping people is the right thing to do. Tiffany's friends are playing right outside. But Tiffany doesn't like playing outside. She doesn't want to go play outside right now.
So Tiffany helped her brother look for his ball. It was really easy for Tiffany to help her brother because she didn't want to play outside. So Tiffany helped her brother.

Figure 1 Example of the images and scripts in Study 1.

All the characters were the same gender as the participants. The order of presenting the four stories was counterbalanced across participants. The order of presenting the high psychological cost character and the low psychological cost character was counterbalanced across story items for each participant. After hearing each pair of stories, the child was asked two remember check questions: "Who found it easy to do something good?" and "Who found it hard to do something good?" Children answered 95% of the trials correctly. We only included those trials where both remember check questions were answered correctly. Including those trials where the remember check questions were answered incorrectly did not change the pattern or significance of results reported here.

Following each story, we asked children two *social evaluation* questions. The first was (i.e., Prize question) "Which of the two characters would you give a prize to?" This was followed by a second question (i.e., Nicer question), "which one do you think is nicer?"

We then asked children a *moral desire rating* question for each character in each pair of stories: "How much do you think she (the "high cost" character) wants to do the right thing?" and "How much do you think she (the "low cost" character) wants to do the right thing?" For each question,

children were asked to use a 3-point rating scale ("a lot", "a little bit", "not at all") to infer the degree of moral desire.

The adults received identical stimuli and questions, but read through these materials themselves online, and the characters were not matched to adult participants' gender.

Results

Social Evaluation First, we examined our first research question that at what age can children evaluate someone who incurs higher psychological costs to fulfill social or moral obligations as more virtuous. See Figure 3.2 for results on children and adults' responses to the social evaluation questions. We conducted a binary logistic regression, with their responses ("low cost" character = 1, "high cost" character = 0) as the dependent variable and age group (4- to 5-year-olds, 6- to 7-year-olds, 8- to 9-year-olds, adults) as a between-subjects factor, and story item (helping, honesty, toys, dishes) and question (prize, nicer) as within-subjects factors. We found a significant main effect of age group (Wald $\chi^2(3, N = 192) = 71.08$, $p < .001$). Specifically, adults were more likely to choose the "high cost" character than either 6- to 7-year-olds Wald $\chi^2(1, N = 129) = 17.83$, $p < .001$, or the 4- to 5-year-olds, Wald $\chi^2(1, N = 131) = 62.65$, $p < .001$. The 8- to 9-year-olds were not significantly different from the adults, $p = .84$, and were more likely to choose "high cost" character than were either the 6- to 7-year-olds Wald $\chi^2(1, N = 61) = 8.79$, $p = .003$, or the 4- to 5-year-olds, Wald $\chi^2(1, N = 63) = 27.14$, $p < .001$. The 6- to 7-year-olds were also more likely to choose the character who incurred a higher psychological cost than the 4- to 5-year-olds, Wald $\chi^2(1, N = 76) = 6.94$, $p = .008$. No effects of questions or story item were found (p's > .06).

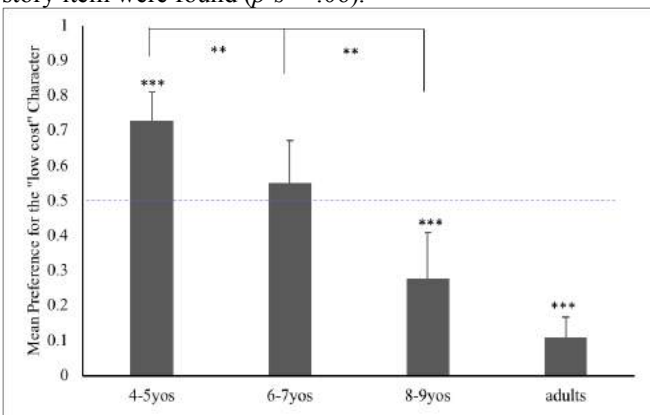


Figure 2. Children's and adults' mean preference for the "low cost" character in Study 1. Error bars represent 95% confidence intervals. Asterisks indicate significance of two-tailed t-tests. (**) $p < .01$, (***) $p < .001$.

Since no significant effects of question or story item were found, we averaged participants' responses in two dependent measure questions across four story items and ran two-tailed one-sample t-tests to compare to chance (0.5) for each age group. Adults significantly favored the "high cost" character ($M = .25$), $t(91) = -6.87$, $p < .001$, 95% $CI = [-.32, -.17]$. In contrast, the 4-to 5-year-olds significantly favored the "low cost" character ($M = .73$), $t(38) = 5.43$, $d =$, 95% $CI = [.14, .31]$. Responses of the 6- to 7-year-olds did not differ from chance ($M = .55$), $t(36) = .83$, $p = .41$, 95% $CI = [-.07, .18]$. The 8- to 9-year-olds significantly favored the "high cost" character ($M = .28$), $t(23) = -3.06$, $p = .006$, 95% $CI = [-.37, -.07]$.

Moral Desire Ratings We then examined participants' ratings of the characters' moral desires (see Figure 3). We ran an ordinal GEE with age group (4-to 5-year-olds, 6-to 7-year-olds, 8- to 9-year-olds, adults) as a between-subject factor, character ("low cost" character, "high cost" character) and story item as within-subject factors. We found a significant main effect of character (Wald $\chi2(1, N = 192) = 221.45$, $p < .001$) that participants' ratings of moral desire were higher for the "low cost" character than the "high cost" character. We also found a significant main effect of story item (Wald $\chi2(3, N = 192) = 47.18$, $p < .001$). Specifically, participants' ratings of moral desire were lower for the Dishes story than the three other stories (p's < .004). No significant differences were found among other stories. We found no significant main effect of age group ($p = .08$) but found a significant interaction between age group and character (Wald $\chi2(3, N = 192) = 24.57$, $p < .001$). To further investigate the interaction, for each age group, we ran an ordinal GEE with character ("low cost" character, "high cost" character) and story item as within-subject factors. We found that although participants in all age groups rated higher moral desire for the "low cost" character than the "high cost" character (4- to 5-year-olds: Wald $\chi2(1, N = 39) = 31.93$, $p < .001$; 6- to 7-year-olds: Wald $\chi2(1, N = 37) = 73.61$, $p < .001$; 8- to 9-year-olds: Wald $\chi2(1, N = 24) = 44.70$, $p < .001$; adults: Wald $\chi2(1, N = 92) = 58.03$, $p < .001$), the difference were strongest among the 6- to 7-year-olds.

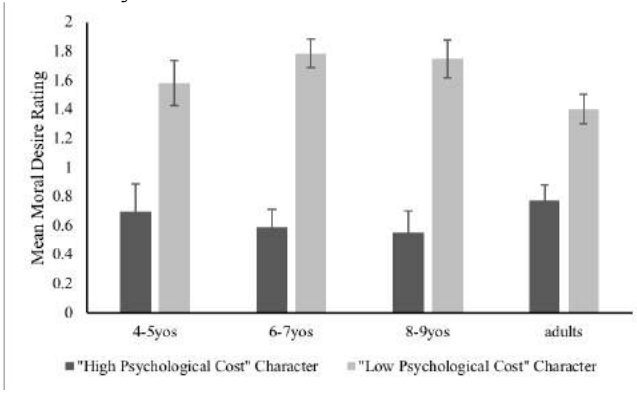


Figure 3. Mean moral desire ratings split by character and age group in Study 1. Error bars represent 95% confidence intervals.

Discussion

In Study 1, adults considered a person who incurred a higher psychological cost to do the right thing (e.g., help brother) more favorably than a person who incurred a lower psychological cost to do the same thing. We found a

developmental change in this evaluation among children. Four- to five-year-olds showed completely opposite evaluation from adults. With age, children increasingly showed a preference for the character who incurred a higher psychological cost to help.

Our results clarify age differences found in Starmans and Bloom (2016) in a few ways. First, in contrast to this prior study, when the difficulty was not explicitly stated as having conflicting desires (a self-interested desire and a moral desire) at once, 8- to 9-year old children gave adult-like favorable evaluations of the character who overcame the difficulty to act morally. Moreover, 6- and 7-year-olds were at chance, rather than favoring the easy action. The reversal from the adult pattern only appeared in the youngest group.

Both children and adults inferred that the person who incurred a lower psychological cost had stronger desire to do the right thing than the person who incurs a higher psychological cost. This suggests that neither children nor adults intuitively inferred coexistence of two conflicting desires (e.g., a self-interested desire and a moral desire).

Although ideally a direct replication and comparison to Starmans & Bloom (2016) would be more informative, we speculate that our results so far may together rule out moral conflict as the central understanding driving children's and adults' social evaluations. Instead, our findings suggest the importance of developing understanding the virtue of incurring costs to do the right thing in children's evaluations. To further investigate this developmental change, in Study 2, we look at how children's consideration of psychological costs may compare to their consideration of physical costs in social evaluations. We focused on the youngest children from study 1, 4- to 7-year-olds, since we found that their evaluations were significantly different from adults. We tested a group of adults as a reference group.

Study 2

Method

Data collection is still ongoing. We set our sample size as 36 children per age group (4- to 5-year-olds and 6- to 7-year-olds). So far, sixty-one 4- to 7-year-olds (4.00- 7.99, M = 5.32, SD = 1.17, 28 boys) from Ithaca, NY were recruited for this study. We divided the children into a younger group (4- to 5-year-olds) and an older group (6- to 7-year-olds). Specifically, 37 4- to 5-year-olds (4.00- 5.95, M = 4.89, SD = .58, 21 boys), 24 6- to 7-year-olds (M = 7.07, SD = .56, 6.03 - 7.99, 7 boys) were included in the preliminary analyses. addition, 101 adults took part in this study and were included in the analyses.

Participants were told four pairs of stories with accompanying pictures, each contrasting a "high cost" character (who incurred a high physical or psychological cost to do the right thing) with a "low cost" character (who incurred a low cost to do the right thing). Two pairs of the stories featured psychological costs and were the same as the Helping Story and the Dishes story in Study 1. The other two pairs of stories featured physical cost (see Figure 4). For example, in the Helping Story, the "high cost" character climbed up the stairs to picked up the ball for her brother, while the "low cost" character walked behind the sofa next to her and picked up the ball.

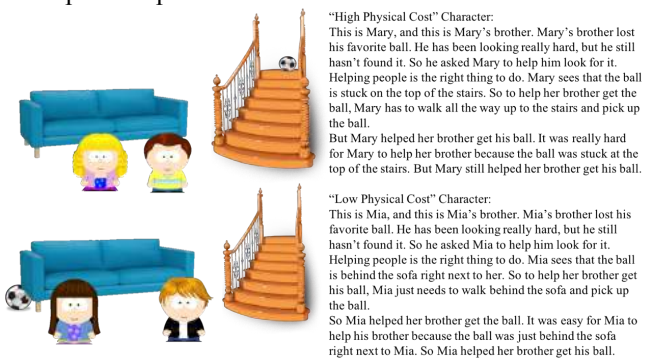


Figure 4 Example of the images and scripts featuring physical cost presented in Study 2.

All the characters were the same gender as the participants. The order of presenting the stories about psychological cost and stories about physical cost were counterbalanced across participants. The order of the high psychological cost character and the low psychological cost character were counterbalanced across stories for each participant. After hearing each pair of stories, the child was asked two remember check questions: "Who found it easy to do something good?" and "Who found it hard to do something good?" Children answered 93% of the trials correctly. We only included those trials where both remember check questions were answered correctly. Including those trials where the remember check questions were answered incorrectly did not change the pattern or significance of results reported here.

Following each story, children were asked the same two *social evaluation* questions (order counterbalanced) as in Study 1. One was (i.e., Prize question) "Which of the two characters would you give a prize to?" The other question (i.e., Nicer question) was "which one do you think is nicer?" We then asked children one *moral desire rating* question for each character using the same measures as Study 1.

The adults received identical stimuli and questions, but read through these materials themselves online, and the characters were not matched to adult participants' gender.

Results

Social Evaluations First, we examined participants' evaluation of the two characters (See Figure 5). We ran a binary logistic regression, with their responses ("low cost" character = 1, "high cost" character = 0) as the dependent variable and age group (4- to 5-year-olds, 6- to 7-year-olds, adults) as a between-subjects factor, and cost type (psychological vs. physical), story item (helping, dishes) and questions (prize, nicer) as within-subjects factors. We found a significant main effect of age group (Wald $\chi^2(2, N = 162) = 93.69$, $p < .001$). Specifically, adults were more likely to choose the "high cost" character than either the 6- to 7-year-

olds (Wald $\chi^2(1, N = 137) = 11.62$, $p = .001$), or the 4- to 5-year-olds (Wald $\chi^2(1, N = 125) = 83.69$, $p < .001$). The 6- to 7-year-olds were also more likely to choose the "high cost" character than the 4- to 5-year-olds (Wald $\chi^2(1, N = 61) = 25.06$, $p < .001$). We also found a significant main effect of cost type (Wald $\chi^2(1, N = 61) = 8.44$, $p = .004$). Specifically, participants were more likely to choose the "high cost" character in the physical stories than in the psychological stories. Interestingly, we also found a significant interaction between age group and cost type, Wald $\chi^2(2, N = 61) = 8.99$, $p = .011$. To further investigate the interaction, for each age group, we ran a binary logistic regression with responses ("low cost" character = 1, "high cost" character = 0) as the dependent variable and cost type (psychological vs. physical), story item (helping, dishes) and question (prize, nicer) as within-subjects factors. We found a marginal effect of cost type for 4- to 5-year-olds (Wald $\chi^2(1, N = 37) = 3.32$, $p = .068$), a significant main effect of cost type for 6- to 7-year-olds (Wald $\chi^2(1, N = 24) = 8.25$, $p = .004$), and no main effect of cost type for adults ($p = .66$). No significant effects question type or story item were found (p's > .25).

Since no significant effects of question type or story item were found, we averaged participants' responses in two dependent measure questions across two story items for each type of cost. We then ran one-sample t-tests to compare participants' responses in each type of story to chance (0.5) for each age group. Adults significantly favored the "high cost" character both for psychological stories ($M = .20$, $t(93) = -8.75$, $p < .001$, 95% $CI = [-.36, -.23]$) and physical stories ($M = .13$, $t(95) = -13.42$, $p < .001$, 95% $CI = [-.43, -.32]$). In contrast, the 4-to 5-year-olds significantly favored the "low cost" character both for physical costs ($M = .64$, $t(35) = 2.28$, $p = .029$, 95% $CI = [.02, .26]$) and psychological costs ($M = .75$, $t(36) = 5.16$, $p < .001$, 95% $CI = [.15, .35]$). The 6- to 7-year-olds significantly favored the "high cost" character for the physical stories ($M = .23$, $t(22) = -4.81$, $p < .001$, 95% $CI = [-.39, -.15]$) but their responses did not differ from chance for the psychological stories ($M = .40$, $t(22) = -1.18$, $p = .25$, 95% $CI = [-.27, .07]$).

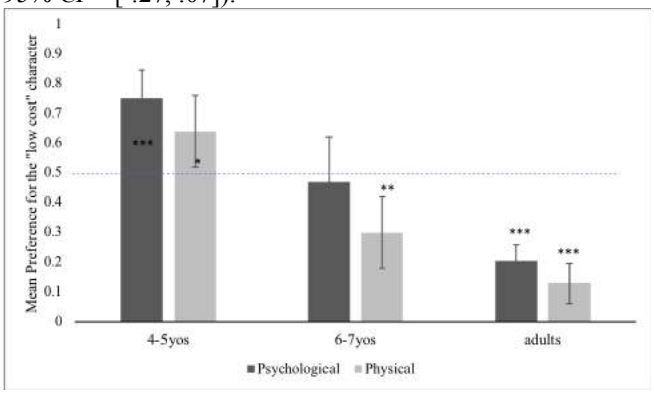


Figure 5. Children's and adults' mean preference for the "low cost" character in Study 2. Error bars represent 95% confidence intervals. Asterisks indicate significance of two-tailed t-tests. (**) $p < .01$, (***) $p < .001$.

Moral Desire Ratings We then examined participants' moral desire ratings for the characters (see Figure 6). We ran an ordinal GEE with age group (4-to 5-year-olds, 6-to 7-year-olds, adults) as a between-subject factor and character ("low cost" character, "high cost" character), cost type (psychological, physical) and story item (Helping, Dishes) as within-subject factors. We found a significant main effect of character (Wald $\chi^2(1, N = 158) = 88.03$, $p < .001$) that participants' moral desire ratings are higher for the "low cost" character than the "high cost" character. We also found a significant main effect of cost type (Wald $\chi^2(1, N = 158) = 4.98$, $p = .026$), that participants' moral desire ratings for the characters are higher in the psychological stories than the physical stories. We also found a significant main effect of story item (Wald $\chi^2(1, N = 158) = 32.74$, $p < .001$), that the moral desire ratings for the characters are higher in the Helping stories than the Dishes stories. Interestingly, we also found a significant interaction between character and cost type (Wald $\chi^2(1, N = 158) = 28.08$, $p < .001$). Follow-up analyses showed that participants rated stronger moral desire for the "low cost" character than the "high cost" character for both psychological cost (Wald $\chi^2(1, N = 158) = 78.21$, $p < .001$) and physical cost (Wald $\chi^2(1, N = 158) = 23.71$, $p < .001$), but the difference is stronger for psychological cost than for physical cost. We also found a significant interaction between age group and character (Wald $\chi^2(1, N = 158) = 28.08$, $p < .001$). Follow-up analyses showed that participants in all age groups rated stronger moral desire for the "low cost" character than the "high cost" character (4- to 5-year-olds: Wald $\chi^2(1, N = 37) = 33.25$, $p < .001$; 6- to 7-year-olds: Wald $\chi^2(1, N = 23) = 36.58$, $p < .001$; Wald $\chi^2(1, N = 98) = 15.67$, $p < .001$), but the difference is stronger among children than adults. Also, we found no significant main effect of age group ($p = .53$).

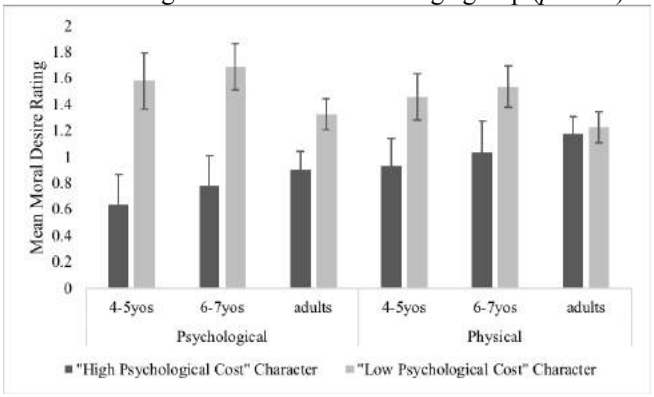


Figure 6. Mean moral desire ratings split by character, cost type and age group in Study 2. Error bars represent 95% confidence intervals.

Discussion

In Study 2, we looked at how children's considerations of psychological costs compare to their considerations of physical costs in social evaluations. Adults consistently demonstrated a favorable evaluation for someone who

incurred a high psychological or physical cost to do the right thing. Preschool-age children (4- to 5-year-olds) demonstrated an opposite evaluation from adults, favoring the person who incurred a lower psychological cost or physical cost. Most six- and seven-year-olds recognize the virtue of acting at a physical cost. Some of them also recognize the virtue of acting at a psychological cost. These results further support the idea that children's evaluation of moral virtue depends on their understanding of cost rather than conflict: Physical cost is incorporated early in development, and psychological cost later.

General Discussion

In this paper, we investigated children's consideration of costs in their social and moral evaluations. Prior studies have mostly focused on children's understanding of physical costs including physical obstacles or valuable resources. Across two studies, we show that young children may start out with an intuitive preference for individuals who find it easy to do something good, and that they gradually transition to an adult-like understanding that incurring costs to do something good is positive, praiseworthy and morally virtuous. Importantly, neither adults nor children inferred conflicting moral and personal desires spontaneously. This helps clear the findings in our study and findings in Starmans &Bloom (2016). It seems that children recognize the virtue of incurring costs before recognizing the virtue of resolving conflicting desires. Moreover, children's recognition of the positivity of incurring costs to do the right thing seems to develop in two stages: They first recognize the positivity of overcoming *physical* obstacles at around 6 to 7 years old, and then understanding the positivity of overcoming *psychological* obstacles at around 8 to 9 years old.

The difference we found between children's consideration of the psychological costs and physical costs add to prior work on children's understanding about costs. Understanding psychological costs is similar to understanding physical costs in that they both involve recognizing the possibility and positivity of making efforts and overcoming some kind of difficulty. However, they are also different in that understanding psychological costs relies on understanding that people may have different desires and that they need to make mental efforts to overcome the psychological obstacles, which may be part of higher-order theory-of-mind understanding (Lagattuta et al. 2015). Exploring interactions of understanding of costs and children's mental state understanding is an important direction for future work.

What underlies the development between ages 4 and 9? There are at least three possible explanations for this developmental change. First, it is possible that, as children age, they increasingly experience situations where they need to incur physical or psychological costs (for example, giving up something they really like) in order to achieve certain social or moral goals. Through such experience of they may gradually recognize the effort one needs to put in this process, and thus understand the virtue of incurring costs to do the right thing. Second, it is also possible that as children get older, they may be increasingly praised and encouraged for making efforts to overcome some physical or psychological difficulties to achieve certain goals by caregivers or teachers. The final possibility is that younger children may have a bias that someone who incurs a lower cost simply has higher competence, while only later they gradually understand that easiness is not necessarily the indicator for competence. This possibility is consistent with prior work in children's reasoning about ability showing that 4-year-olds judge someone who finds a task easy to be smarter than one who find the same task hard (Heyman, Gee, & Giles, 2003). These possibilities are certainly not mutually exclusive. It might be that children's first-person experience, the linguistic input they receive, and their increasingly mature understanding of competence together guide their development of an understanding of the virtue of incurring costs to do the right thing.

Acknowledgments

We gratefully thank the members of the Early Childhood Cognition Lab, and especially Jason Lin and Andrew Lee for assistance with recruitment and coding. We thank the staff of the Sciencenter and preschools, and the children and parents who participated in this research.

References

Baird, J. A., & Astington, J. W. (2004). The role of mental state understanding in the development of moral cognition and moral action. *New directions for child and adolescent development*, *2004*(103), 37-49.

Burns, M. P., & Sommerville, J. (2014). "I pick you": the impact of fairness and race on infants' selection of social partners. *Frontiers in Psychology*, *5*, 93.

Cushman, F., Sheketoff, R., Wharton, S., & Carey, S. (2013). The development of intent-based moral judgment. *Cognition*, *127*(1), 6-21.

Chernyak, N., & Kushnir, T. (2013). Giving preschoolers choice increases sharing behavior. *Psychological Science*, *24*(10), 1971-1979.

Chernyak, N., & Kushnir, T. (2018). The influence of understanding and having choice on children's prosocial behavior. *Current opinion in psychology*, *20*, 107-110.

Geraci, A., & Surian, L. (2011). The developmental roots of fairness: Infants' reactions to equal and unequal distributions of resources. *Developmental science*, *14*(5), 1012-1020.

Hamlin, J. K., Wynn, K., & Bloom, P. (2007). Social evaluation by preverbal infants. *Nature*, *450*(7169), 557.

Hamlin, J. K., Wynn, K., Bloom, P., & Mahajan, N. (2011). How infants and toddlers react to antisocial others. *Proceedings of the national academy of sciences*, *108*(50), 19931-19936.

Jara-Ettinger, J., Tenenbaum, J. B., & Schulz, L. E. (2015). Not so innocent: Toddlers' inferences about costs and culpability. *Psychological science*, *26*(5), 633-640.

Killen, M., Mulvey, K. L., Richardson, C., Jampol, N., & Woodward, A. (2011). The accidental transgressor:

Morally-relevant theory of mind. *Cognition*, *119*(2), 197-215.

Lagattuta, K. H. (2005). When you shouldn't do what you want to do: Young children's understanding of desires, rules, and emotions. *Child Development*, *76*(3), 713-733.

Lagattuta, K. H., Nucci, L., & Bosacki, S. L. (2010). Bridging theory of mind and the personal domain: Children's reasoning about resistance to parental control. *Child Development*, *81*(2), 616-635.

Liu, S., Ullman, T. D., Tenenbaum, J. B., & Spelke, E. S. (2017). Ten-month-old infants infer the value of goals from the costs of actions. *Science*, *358*(6366), 1038-1041.

Olson, K. R., & Spelke, E. S. (2008). Foundations of cooperation in young children. *Cognition*, *108*(1), 222-231.

Sloane, S., Baillargeon, R., & Premack, D. (2012). Do infants have a sense of fairness?. *Psychological science*, *23*(2), 196-204.

Smetana, J. G., Jambon, M., Conry-Murray, C., & Sturge-Apple, M. L. (2012). Reciprocal associations between young children's developing moral judgments and theory of mind. *Developmental Psychology*, *48*(4), 1144.

Starmans, C., & Bloom, P. (2016). When the spirit is willing, but the flesh is weak: Developmental differences in judgments about inner moral conflict. *Psychological science*, *27*(11), 1498-1506.

Big, Little, or Both? Exploring the Impact of Granularity on Learning for Students with Different Incoming Competence

Guojing Zhou, Xi Yang, and Min Chi
⟨*gzhou*3, *yxi*2, *mchi*@*ncsu.edu*⟩
Department of Computer Science, College of Engineering
North Carolina State University, Raleigh, NC, USA

Abstract

We explored the impact of three types of decision granularity, problem level (Prob), step level (Step), and both problem and step levels (Both), on student learning. We first conducted an empirical study to directly compare the three conditions and then three subsequent studies to evaluate one or two of the three conditions. Overall our empirical results showed there was no significant difference among the three conditions. We further split students into different groups based on their performances on the single-principle and the multiple-principle problems in the pre-test. Solving the single-principle problems only involves one step while solving the multiple-principle ones involves generating multiple steps in a logic order. We define High students as those who were correct on *all* single-principle problems and *at least one* multiple-principle ones in the pre-test, Low students as those who were correct on *some or all* single-principle problems *but no* multiple-principle ones, and the rest are in the Medium group. Our empirical results showed that for Low students, Both can be better than Step. For the Medium and High students, no clear conclusions could be drawn because of small sample sizes. As a result, in a post-hoc analysis all students were combined by their assigned conditions. Overall, while no significant difference was found among the three conditions, we found that the impact of three types of granularity, Prob, Step, and Both differs significantly for High vs. Low students: $Both, Step > Prob$ for the High students and $Both, Prob > Step$ for the Low students. No clear conclusions could be drawn for the Medium group due to its small sample sizes. In short, while Prob could be effective for Low students but ineffective for High ones and Step could be effective for High students but ineffective for Low ones, Both seemed to be effective for both High and Low students.

Keywords: granularity, worked example, problem solving, student competence

Introduction

In STEM domains like math, probability and science, solving a *problem* often requires producing an argument, proof or derivation consisting of one or more inference steps, and each *step* is the result of applying a domain principle, operator or rule. For instance, an algebraic equation 2x+5=21 can be solved via two steps: 1) subtract the same term 5 from both sides of the equation; and 2) divide both sides by the non-zero term 2. As a result, tutoring in such domains is often structured as a two-loop procedure. An outer loop selects the *problem* or task the student should work on next, while the inner loop governs *step* level decisions such as whether or not to give a hint (Vanlehn, 2006).

In this paper, we directly explored the impact of three types of decision granularity on student learning by comparing three conditions: problem level (Prob), step level (Step), and both problem and step levels (Both). In the Prob condition, the tutor randomly decides whether the next problem is worked example (WE) or problem solving (PS). In WE, students observe how the tutor solves a problem, while in PS the students solve the problem themselves. In the Step condition, a random decision is made on whether the next *step* should be WE or PS. To differentiate it from the problem level PS and WE, we refer to such step level interleaving as *Faded Worked Example (FWE)*. Finally, the Both condition involves both levels of decisions: at the problem level, it randomly decides whether the next problem should be WE, PS or FWE; if FWE is selected, step level decisions will be randomly made.

A series of studies were conducted to evaluate the three types of decision granularity in the domain of probability using an Intelligent Tutoring System (ITS) named Pyrenees from 2014-2017. Pyrenees allowed us to rigorously control the content and vary only the types of decision granularity. In Fall 2014 (Fall'14), all three conditions were *empirically* compared; for the subsequent studies, only one or two conditions were examined.[1] In a post-hoc comparison, students from all studies were combined by their conditions because all conditions across different years went through the same standard 4-phase procedure: textbook, pre-test, training on ITS, and post-test, and all materials in each of the four phases were kept to be *identical* across different years. Overall, our results showed that there was no significant difference among the three conditions either in Fall'14 (Zhou, Price, Lynch, Barnes, & Chi, 2015) or in the post-hoc analysis.

On the other hand, the *aptitude-treatment interaction (ATI)* effect states that some instructional interventions can be more or less effective for particular students depending upon their specific abilities or knowledge (Cronbach & Snow, 1977; Snow, 1991). Here we argue that WE, PS, and FWE involve different *learning mechanisms*. More specifically, in WEs, students learn by *observing* how the tutor solves a problem; in PSs, students learn by *doing* – solving the problem with the tutor's assistance; in FWEs, students learn by *collaboratively constructing* the solution with the tutor. As a result, we argue that in the Prob condition students switched between learning by observing (WE) and learning by doing (PS); in the

[1] Please note that another purpose of the subsequent studies was to compare reinforcement learning induced policies with random policies. Due to participant limit, we were not able to compare the three conditions again.

Step condition, students learn by *collaboratively constructing* answers for (FWEs) with the tutor; in the Both condition, students experienced all three types of learning mechanisms. Therefore, we expect that Prob, Step, and Both can be more or less effective for different students.

To investigate whether there is indeed an ATI effect, we split students into High, Medium and Low groups based on their incoming competence measured by their performances on six single-principle and four multiple-principle problems in the pre-test. Solving the single-principle problems only involves one step while solving the multiple-principle ones involves generating multiple steps in a logic order. We define High students as those who were correct on *all six* single-principle problems and *at least one* multiple-principle ones in the pre-test, Low students as those who were correct on *some or all* single-principle problems *but no* multiple-principle ones, and the rest are in the Medium group. Our results from Fall'14 showed that for the Low students, both levels of the granularity (Both) is significantly more effective than the step level decisions (Step); for the Medium and High students, no clear conclusions could be drawn because of small sample sizes. In the post-hoc analysis, no clear conclusions could be drawn for the Medium group due to its small sample sizes and for the other two groups, we have: $Both, Step > Prob$ for the High students and $Both, Prob > Step$ for the Low students. In short, our post-hoc analysis suggested that the problem level decisions (Prob) could be effective for Low students but ineffective for High ones; on the other hand, the step level decisions (Step) could be effective for High students but ineffective for Low ones; finally, the both level decisions (Both) seemed to be effective for both High and Low students.

Background and Related Work

The Impact of Granularity Involving WE, PS, FWE

Much of prior research has investigated the effectiveness of WE, PS, FWE, and their various combinations (Sweller & Cooper, 1985; McLaren, Lim, & Koedinger, 2008; McLaren & Isotani, 2011; McLaren, van Gog, Ganoe, Yaron, & Karabinos, 2014; Van Gog, Kester, & Paas, 2011; Renkl, Atkinson, Maier, & Staley, 2002; Schwonke et al., 2009; Najar, Mitrovic, & McLaren, 2014; Salden, Aleven, Schwonke, & Renkl, 2010; Zhou et al., 2015; Zhou, Lynch, Price, Barnes, & Chi, 2016; Zhou & Chi, 2017; Zhou, Wang, Lynch, & Chi, 2017; Zhou, Azizsoltani, Ausin, Barnes, & Chi, 2019). Here we only include those that involved any of the three types of granularity. At the problem level granularity, for example, McLaren et al. (2008) found no significant difference in learning performance between Prob (WE-PS pairs) and PS-only, but the former spent significantly less time than the latter. In a subsequent study, McLaren and Isotani (2011) compared three conditions: WE-only, PS-only, and Prob (WE-PS pairs). Similarly, no significant differences were found among them in terms of learning gains, but the WE condition spent significantly less time than the other two; and no significant time on task difference was found between the PS and the Prob (WE-PS pairs) condition.

A series of studies compared the Step level and the Both level granularity with PS only (Schwonke et al., 2009; Salden et al., 2010). Results showed that the former two can be more effective than the latter. For example, Salden et al. compared three conditions: Both (WE-FWE-PS), Step (FWE), and PS-only (Salden et al., 2010). Their results showed that Step outperformed Both, which in turn outperformed PS-only, and no significant time on task difference was found among the three conditions. Note that in this study, the order of WE, FWE, and PS was fixed in Both; while in Step, the tutor used an adaptive pedagogical policy, expert rules combined with data-driven student models, to determine whether the next step should be WE or PS. Therefore, it is was not clear whether it was the adaption or the granularity that made the Step condition more effective than the other two conditions. In our studies, we factored out the impact of adaption by employing random policies.

While the studies described above mainly used PS-only as baselines, several studies directly compared different types of granularity. Overall, results suggested that the Both level granularity could be more effective than the Prob level (Renkl et al., 2002; Najar et al., 2014). For example, Renkl et al. (2002) compared Both (WE-FWE-PS) with Prob (WE-PS pairs) and the former significantly outperformed the latter on student learning performance while no significant difference was found between them on time on task. Similarly, Najar et al. (2014) compared Both (adaptive WE/FWE/PS) with Prob (WE-PS pairs). They found that the former significantly outperformed the latter in terms of learning outcomes and the former also spent significantly less time on task. Here, an adaptive pedagogical policy was also employed to make both the problem and step level decisions. Thus, it is quite possible that the superiority of Both over Prob stemmed from the adaption rather than from the granularity. In sum, while different decision granularities were involved in prior studies, the WEs and PSs were provided following some fixed or adaptive pedagogical policies. In this work, we factor out the impact of pedagogical policies by employing a random policy for all three types of granularity.

The ATI Effect of WE, PS, FWE

Some prior studies have also investigated the ATI effect of WE, PS, FWE, and their combinations (Kalyuga, Chandler, Tuovinen, & Sweller, 2001; Najar & Mitrovic, 2013; Najar, Mitrovic, & McLaren, 2016). For example, Najar and Mitrovic (2013) compared three conditions: 1) WE-only, 2) PS-only and 3) Prob (WE-PS pairs) in the domain of Structured Query Language and students were split into High vs. Low groups based on their pre-test scores. The results showed that for the High students: Prob, PS-only $>$ WE-only; while for their Low peers: Prob $>$ PS-only, WE-only. In a subsequent study, Najar et al. (2016) compared Both (adaptive WE/FWE/PS) with Prob (WE-PS pairs) and students were divided into High and Low groups by a median split on pre-test scores. Results showed that for the High students, Both

Table 1: Single-principle Problem vs. Multiple-principle Problem

Type	Single-principle Problem	Multiple-principle Problem
Question	If $p(A \cap B) = 0.2$ and $p(B) = 0.5$, find $P(A\|B)$.	If $p(B) = 0.06$, $p(\sim A \cap \sim B) = 0.87$ and $p(A \cap B) = 0.03$, find $p(A)$.
Answer	Apply the definition of conditional probability: $p(A\|B) = p(A \cap B)/p(B) = 0.2/0.5 = 0.4$	1) Apply the complement theorem: $p(\sim B \cap \sim A) + p(\sim (\sim B \cap \sim A)) = 1$ 2) Apply the de morgan's law: $p(A \cup B) = p(\sim (\sim B \cap \sim A)) = 1 - 0.87 = 0.13$ 3) Apply the addition theorem: $p(A \cup B) = p(A) + p(B) - p(A \cap B)$, $p(A) = 0.13 + 0.03 - 0.06 = 0.1$.

is more effective than Prob; while for the Low students, no significant difference was found.

In short, prior research investigating the ATI effect of WE, PS, FWE, and their combinations showed that for Low students, Prob could be more effective than doing WE and PS only; but for High students, Both can be more effective than Prob. While much of prior ATI research involved one or two types of granularity, to the best of our knowledge, no prior study has investigated the ATI effect when comparing the three types of granularity directly.

High, Medium, vs. Low Students

To investigate the ATI effect, we need to first distinguish students based on some specific abilities or knowledge. Learning in STEM domains such as math and science often involves acquiring two types of knowledge: declarative and procedural (Anderson, 1993). Declarative knowledge includes facts that we know and that can be described to others, for example, "the probability of TRUE is always 1". Procedural knowledge specifies how to retrieve and use declarative knowledge to solve problems. It is a type of knowledge that display with behaviors and often times cannot be explicitly described. Procedural knowledge often requires the interplay of many cognitive factors including but not limited to the following five ones in order of occurrence: 1) acquisition of declarative knowledge, 2) identification and retrieval of the proper declarative knowledge, 3) application of declarative knowledge, 4) organization and production of solution plans; 5) execution of solution plans and evaluation of answers.

Similar to previous research, we used pre-test to measure students' incoming competence. Our pre-test contains single-principle problems which involve applying one domain principle once and multiple-principle problems which involve applying multiple domain principles and for some principles more than once. Table 1 shows an example for each of them. The second column shows the question and answer for a single-principle problem. As we can see, the problem can be solved by directly applying a single-principle. The third column shows a multiple-principle problem. Solving the problem needs to not only apply three algebraic principles but also organize them in a logical order.

Based on the five cognitive factors described above, we argue that solving single-principle problems mainly involves factors 1-3, while solving multiple-principle ones involves all five of them. Thus, students must be able to solve single-principle problems before they can solve multiple-principle problems. Our data supported this point, showing that students who could solve multiple-principle problems always had the perfect score on all single-principle problems in the pre-test. Therefore, in the following we refer to students who could solve at least one multiple-principle problem correctly as High students, those who could only solve some or all of the six single-principle problems correctly as Low students, and the rest as the Medium students.

Methods

Participants

Four studies were conducted in each of the Fall semesters from 2014-2017 to evaluate the three conditions: Prob, Step, and Both using an ITS named Pyrenees in the undergraduate-level Discrete Mathematics course at North Carolina State University. They were assigned to students as one of their regular homework assignments and the completion of the tutor was required for full credit. Students were told that the assignment will be graded based on their demonstrated effort rather than performance. In different studies, different conditions were evaluated and in each study, students were randomly assigned to each condition. In Fall'14, all three conditions were *empirically* compared while for the subsequent three studies, only one or two conditions were examined and in the post-hoc analysis, students from all studies across the four years were combined by their conditions.

Table 2 shows an overview of participants in the four studies and the post-hoc analysis: the first two columns show the semester of the study and its corresponding conditions; columns 3 and 4 list the number of students initially assigned and finally completed in each condition. Overall, Pearsons Chi-squared test showed that there was no significant difference among the three conditions on their completion of study: $\chi^2(2) = 1.13, p = 0.57$ for Fall'14 and $\chi^2(2) = 0.65, p = 0.72$ for the post-hoc analysis. Here we only focus on Fall'14 and the post-hoc analysis because all three conditions are present.

Finally, students with perfect pre-test scores were excluded because we could not measure the improvement they made through training. The last column in Table 2 shows the number of students included in the following analysis.

Probability Tutor

Pyrenees is a web-based tutor that teaches students a general problem solving strategy and 10 major probability principles, such as the Complement Theorem and Bayes' Rule. It provides students with step-by-step instruction, immediate feedback, and on-demand help. Specifically, the help is provided via a sequence of increasingly specific hints. The last hint in

Table 2: Participants for E ach Study and Condition

Study	Cond	Distributed	Completed	Included
Fall'14	Prob	58	38	37
	Step	59	39	37
	Both	59	34	34
Fall'15	Prob	47	38	38
	Step	47	35	34
Fall'16	Prob	40	32	31
	Step	41	35	35
Fall'17	Both	70	57	56
Post-hoc	Prob	145	108	106
	Step	147	109	106
	Both	129	91	90

the sequence, i.e., the bottom-out hint, tells student exactly what to do. The ITS has three basic modes. In the WE mode, all the steps in a problem were solved by the tutor while in the PS mode, they were solved by the student. In the FWE mode, each step has a 50% chance to be solved by the tutor and 50% chance by the student. Except for the decision granularity, the remaining components of the tutor, including the GUI interface, the training problems and the tutorial support were identical for all students.

Procedure

All four studies include the four identical phases: 1) textbook, 2) pre-test, 3) training, and 4) post-test. The only difference among the three conditions was the decision granularity level, problem level for Prob; step level for Step; and both the problem and the step level for Both.

During textbook, all students studied the domain principles through a probability textbook. They read a general description of each principle, reviewed some examples of it, and solved some single- and multiple-principle problems. After solving each problem, the student's answer was marked in green if it was correct and red if incorrect. They were also shown an expert solution at the same time. If the students failed to solve a single-principle problem, then they were asked to solve an isomorphic one. This process was repeated until they either failed three times or succeeded once. The students had only one chance to solve each multiple-principle problem and were not asked to solve an isomorphic problem if their answer was incorrect.

The students then took a pre-test which contained 10 problems. They were not given feedback on their answers, nor were allowed to go back to earlier questions (this was also true for the post-test).

During training, students in all three conditions received the same 12 problems in the same order. Each main domain principle was applied at least twice. The minimal number of steps needed to solve each training problem ranged from 20 to 50. Such steps included variable definitions, principle applications, and equation solving. The number of domain principles required to solve each problem ranged from 3 to 11. The problems were given as WE, PS, or FWE, based upon the students' experimental condition. All students could access the textbook.

Finally, all students took the post-test which contained 16 problems in total. 10 of the problems were isomorphic to the pre-test problems given in phase 2. The remainder were non-isomorphic multiple-principle problems.

Grading criteria

The pre- and post-test problems required students to derive an answer by writing and solving one or more equations. We used three scoring rubrics: binary, partial credit, and one-point-per-principle. Under the binary rubric, a solution was worth 1 point if it was completely correct or 0 if not. Under the partial credit rubric, each problem score was defined by the proportion of correct principle applications evident in the solution. A student who correctly applied 4 of 5 possible principles would get a score of 0.8. The One-point-per-principle rubric in turn gave a point for each correct principle application. All of the tests were graded in a double-blind manner by a single experienced grader. The results presented below were based upon the partial-credit rubric but the same results hold for the other two. For comparison purposes, all test scores were normalized to the range of $[0,1]$.

Results

The three conditions were compared on test scores. For the Fall'14 study, a One-way ANOVA analysis on the pre-test score showed no significant difference among the three condition: $F(2,105) = 1.12$, $p = 0.33$, $\eta = 0.021$. A One-way ANCOVA analyses on the post-test score using the pre-test score as a covariate also showed no significant difference: $F(2,104) = 1.70$, $p = 0.19$, $\eta = 0.021$. Similar insignificant results were found in the host-hoc analysis: $F(2,299) = 0.68$, $p = 0.51$, $\eta = 0.005$ for the pre-test and $F(2,298) = 0.98$, $p = 0.38$, $\eta = 0.004$ for the post-test. In terms of time on task, contrast analysis revealed that Prob spent significantly less time than Step in both Fall'14: $t(105) = -2.62$, $p = 0.010$, $d = 0.61$ and post-hoc: $t(299) = -3.00$, $p = 0.003$, $d = 0.40$.

To evaluate the ATI effect, we split students based on their pre-test scores. Our pre-test included six single-principle and four multiple-principle problems. Following our splitting criteria discussed above, we refer to students who could solve at least one multiple-principle problem correctly (pre ≥ 0.7) as High students [2]; those who could only solve some or all of the six single-principle problems correctly (pre ≤ 0.6) as Low students, and the rest as Medium ones. As expected, in the pre-test the High group scored significantly higher than the Medium group: $t(105) = 6.94$, $p < 0.0001$, $d = 3.16$ in Fall'14 and $t(299) = 9.71$, $p < 0.0001$, $d = 2.37$ in post-hoc; the Medium group significantly outperformed the Low group: $t(105) = 8.41$, $p < 0.0001$, $d = 2.14$ in Fall'14 and $t(299) = 11.82$, $p < 0.0001$, $d = 2.08$ in post-hoc.

Incoming competence combined with three conditions partitioned the students into nine groups for both Fall'14 and

[2]Note that in our grading rubrics, all problems were weighted equally in both pre- and post-tests.

Table 3: Students Performance and Time (minutes) on Fall'14 Empirical Study and Post-hoc Analysis

Cond	Fall'14 Empirical Study					Post-hoc Analysis				
	N	Pre	Iso	Post	Time	N	Pre	Iso	Post	Time
Prob_H	12	.857(.065)	.817(.125)	.700(.169)	85.5(19.9)	56	.822(.086)	.844(.138)	.740(.178)	111.1(42.5)
Step_H	8	.800(.080)	.868(.156)	.769(.154)	125.2(40.0)	47	.827(.077)	.908(.089)	.819(.126)	128.2(29.4)
Both_H	13	.826(.064)	.863(.136)	.767(.160)	113.2(30.1)	46	.818(.073)	.902(.110)	.821(.156)	106.6(29.3)
Prob_M	5	.636(.017)	.728(.134)	.598(.168)	96.8(17.4)	10	.652(.021)	.813(.148)	.688(.173)	101.1(24.8)
Step_M	6	.647(.017)	.687(.187)	.559(.124)	124.1(22.2)	12	.649(.022)	.818(.190)	.715(.191)	125.7(28.5)
Both_M	6	.653(.028)	.740(.163)	.618(.189)	111.5(25.8)	11	.652(.025)	.736(.216)	.616(.216)	113.1(26.3)
Prob_L	20	.453(.117)	.657(.190)	.528(.205)	95.8(29.3)	40	.455(.117)	.703(.234)	.596(.244)	98.7(35.7)
Step_L	23	.441(.103)	.592(.192)	.458(.153)	104.2(38.1)	47	.439(.110)	.628(.219)	.500(.190)	109.8(34.6)
Both_L	15	.414(.110)	.703(.170)	.550(.154)	105.1(37.4)	33	.415(.119)	.707(.208)	.565(.185)	110.3(36.3)

post-hoc. The number of students in each group is listed in the "N" column for Fall'14 in Table 3 (Left) and for post-hoc in Table 3 (Right). Fortunately, random assignment balanced the three conditions for ability, and this balance persisted even after the groups were subdivided into High, Medium, and Low. No significant difference was found on pre-test among the three High groups, the three Medium groups, or the three Low groups in both Fall'14 and post-hoc.

Empirical Fall'14 Study

In Table 3, the first column shows the condition-competence group and then followed by a section presenting the learning performance and time on task (in minutes) for Fall'14. Here it shows the number of students (N) and the mean and SD of pre-test score (Pre), isomorphic post-test score (Iso), overall post-test score (Post) and time on task (Time). A Chi-square test showed that there was no significant relation between condition and incoming competence $\chi^2(4) = 2.94, p = 0.57$.

To measure student learning improvement, we compared their isomorphic post-test scores with their pre-test scores. A repeated measures analysis using test type (pre-test vs. isomorphic post-test) as a factor and test score as the dependent measure showed that there is a main effect for test type: $F(1,107) = 50.82$, $p < 0.0001$, $\eta = 0.322$ in that they scored significantly higher on the isomorphic post-test problems than pre-test. Thus, our tutor is indeed effective on improving student learning. More specifically, all three conditions scored significantly higher in the isomorphic post-test than in the pre-test: $F(1,36) = 13.56$, $p = 0.0008$, $\eta = 0.274$ for Prob, $F(1,36) = 16.26$, $p = 0.0003$, $\eta = 0.311$ for Step, and $F(1,33) = 20.92$, $p < 0.0001$, $\eta = 0.388$ for Both respectively. This showed that the basic practices and problems, domain exposure, and interactivity of our ITS might be effective to help students acquire knowledge.

Finally, to obtain a comprehensive evaluation of students' final performance, analyses were performed on the overall post-test which contains six additional multiple-principles. A two-way ANCOVA analysis on the factors of granularity and incoming competence using the pre-test score as a covariate showed no significant interaction or main effect. A subsequent pairwise contrast analysis revealed that for Low students, the Both_L group scored significantly higher than the Step_L group: $t(98) = -2.01$, $p = 0.047$. The results suggested that the Both levels of decisions can be more effective than the step level decisions for the Low students.

In terms of time on task, a two-way ANOVA analysis on granularity and incoming competence showed a main effect on granularity: $F(2,99) = 3.97$, $p = 0.02$, $\eta = 0.071$ in that the Prob condition spent significantly less time than the Step condition $t(105) = -2.62$, $p = 0.01$, $d = 0.61$ and the Both condition $t(105) = -2.22$, $p = 0.029$, $d = 0.58$. Subsequent contrast analyses showed that such difference mainly came from the High students in that: Prob_H spent significantly less time than Step_H and Both_H: $t(99) = -2.72$, $p = 0.008$, $d = 1.35$ and $t(99) = -2.17$, $p = 0.03$, $d = 1.08$ respectively; no significant difference was found among the three Low groups.

Overall, Fall'14 results showed that on learning performance, Both was better than Step for the Low students; while on time on task, Prob spent less time than the other two for the High students. Note that since some of the groups are in small size, the absence of significant differences might be due to insufficient statistical power.

Post-hoc Analysis

The right section of Table 3 presents the post-hoc analysis results. Numbers in the "N" column revealed that the three High and the three Low groups are in reasonable size while the three Medium groups remain small. A Chi-square test showed no significant relation between condition and incoming competence: $\chi^2(4) = 2.11, p = 0.72$.

A repeated measures analysis using test type (pre-test vs. isomorphic post-test) as a factor and test score as the dependent measure showed that there was a main effect for test type $F(1,301) = 177.38$, $p < 0.0001$, $\eta = 0.371$ in that students scored significant higher in the isomorphic post-test than in the pre-test. Similarly, for each of the three conditions, students scored significantly higher in the isomorphic post-test than in the pre-test: $F(1,105) = 42.79, p < 0.0001, \eta = 0.290$ for Prob; $F(1,105) = 72.27$, $p < 0.0001$, $\eta = 0.408$ for Step and $F(1,89) = 67.46$, $p < 0.0001$, $\eta = 0.431$ for Both. The

results confirmed that our tutor is effective over the years.

For the overall post-test scores, a two-way ANCOVA analysis on the factors of granularity and incoming competence using the pre-test score as a covariate showed a significant interaction effect: $F(4,292) = 3.66$, $p = 0.006$, $\eta = 0.029$. Subsequent contrast analyses showed that for High students, the Step_H group and the Both_H group scored significantly higher than the Prob_H group: $t(292) = 2.25$, $p = 0.03$ and $t(292) = -2.50$, $p = 0.01$ respectively. For Low students, the Prob_L group and the Both_L group scored significantly higher than the Step_L group: $t(292) = 2.29$, $p = 0.02$ and $t(292) = 2.19$, $p = 0.03$ respectively. The results suggest that for High students, the Step level decisions and the Both level decisions are more effective than the Prob level while for Low students, the Prob level decisions and the Both level decisions are more effective than the Step level.

For time on task, a two-way ANOVA analysis on granularity and incoming competence showed a significant main effect on granularity: $F(2,293) = 4.98$, $p = 0.007$, $\eta = 0.032$ in that the Step condition spent more time than the Prob condition: $t(299) = 3.00$, $p = 0.003$, $d = 0.40$ and the Both condition: $t(299) = 2.22$, $p = 0.027$, $d = 0.34$. Subsequent contrast analysis revealed that for High students: the Step_H group spent longer time than the Prob_H group: $t(293) = 2.51$, $p = 0.01$, $d = 0.46$ and the Both_H group: $t(293) = 3.03$, $p = 0.003$, $d = 0.74$. No such significant difference was found among the three Low groups.

Overall, the results suggest that on learning performance, the problem level decisions can be effective for Low students but ineffective for High students, the step level decisions could be effective for High students but ineffective for Low students, while Both level decisions seem to be effective for both High and Low students. For time on task, the High students, the Step_H group can spend more time than the Prob_H and the Both_H groups while no significant difference was found among the three Low groups.

Conclusion & Discussion

In this paper, we explored the impact of three types of decision granularity on student learning by comparing three conditions: Prob involving WE and PS, Step involving FWE only, and Both involving all WE, PS and FWE. Overall, while no significant difference was found among the three conditions on learning performance, a significant difference was found among them on time on task in that Prob spent significantly less time than Step for both Fall'14 and the post-hoc.

We hypothesized that different learning mechanisms are involved in WE, PS and FWE and thus there may exist an ATI effect. Students were then split into High, Medium and Low groups based on their pre-test performance. Results from Fall'14 show that on learning performance, for Low students Both is more effective than Step; on time on task, for High students Prob would spend less time than Step. Overall because of small sample sizes, more general conclusions cannot be drawn here. Furthermore, our post-hoc results suggest that on learning performance, Prob can be effective for Low students but ineffective for High ones on the other hand, Step could be effective for High students but ineffective for Low ones; finally, Both seemed to be effective for both High and Low students; as for time on task, while no significant difference was found among the three Low groups either in Fall'14 or post-hoc, significant difference was found among the three High groups in that Prob_H spent significantly less time than Step_H in both Fall'14 and post-hoc.

Our results showed a difference between the Prob and Step granularity. In terms of time on task, students spent less time when learning with Prob than with Step. For learning performance, each of them can be effective for some students but ineffective for some other students, depending on students' knowledge level. This suggests that the granularity can have an impact on student learning. Additionally, results for the Both granularity suggest that mixing this two types of granularity together has the potential to get a more robust instructional intervention. The Prob granularity can be ineffective for the High students and the Step granularity can be ineffective for Low students, but our results suggest that Both can be effective for both High and Low students.

One possible explanation for our results is that different cognitive load were involved in the three conditions. At the problem level, students pay attention to either the tutor's solution in WE or their own solution in PS; while at the step level, they need to pay attention to both the tutor's solution and their own solution and integrate them. Compared with PSs, in FWEs the tutor may solve certain steps for students but on the other hand, students need to devote extra effort to understand and to integrate their answers with the tutor's answers. Thus, we hypothesized that in terms of cognitive load, WE <PS <FWE. This explains why the Step condition spent more time than the Prob condition (in both Fall'14 and post-hoc) despite that students in these two condition completed the same amount of work (as measured by the number of PS steps in our subsequent log analysis). Assuming that FWEs are more challenging than WEs or PSs, the results that Step benefits the High students more than Prob while Prob benefits the Low ones more than Step can be explained by the conjecture that High students have more prior knowledge and learning capacity than the Low ones. However, this is only our hypothesis and much more research is needed to fully understand it. More importantly, more research is needed to explain why the Both levels of granularity benefits both High and Low students.

Lots of prior research has shown that studying WEs help students learn. However, questions about how and when WEs should be presented remain open. Our findings inform researchers that the granularity can have an impact on student learning and the impact of granularity can differ for students at distinct knowledge levels. Thus, it urges researchers to consider the impact of granularity when designing instructions and adapt the instruction based on students' knowledge level.

Acknowledgements

This research was supported by the NSF Grants #1432156: "Educational Data Mining for Individualized Instruction in STEM Learning Environments", #1651909: "CAREER: Improving Adaptive Decision Making in Interactive Learning Environments", #1726550: "Integrated Data-driven Technologies for Individualized Instruction in STEM Learning Environments", and #1916417: "MetaDash: A Teacher Dashboard Informed by Real-Time Multichannel Self-Regulated Learning Data". We would also like to thank the anonymous reviewers for their valuable feedback.

References

Anderson, J. R. (1993). Problem solving and learning. *American Psychologist*, *48*(1), 35.

Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods: A handbook for research on interactions.* Irvington.

Kalyuga, S., Chandler, P., Tuovinen, J., & Sweller, J. (2001). When problem solving is superior to studying worked examples. *Journal of educational psychology*, *93*(3), 579.

McLaren, B. M., & Isotani, S. (2011). When is it best to learn with all worked examples? In *International conference on artificial intelligence in education* (pp. 222–229).

McLaren, B. M., Lim, S.-J., & Koedinger, K. R. (2008). When and how often should worked examples be given to students? new results and a summary of the current state of research. In *Proceedings of the 30th annual conference of the cognitive science society* (pp. 2176–2181).

McLaren, B. M., van Gog, T., Ganoe, C., Yaron, D., & Karabinos, M. (2014). Exploring the assistance dilemma: Comparing instructional support in examples and problems. In *Intelligent tutoring systems* (pp. 354–361).

Najar, A. S., & Mitrovic, A. (2013). Do novices and advanced students benefit differently from worked examples and its. In *Proceedings of international conference icce* (pp. 20–29).

Najar, A. S., Mitrovic, A., & McLaren, B. M. (2014). Adaptive support versus alternating worked examples and tutored problems: Which leads to better learning? In *Umap* (pp. 171–182). Springer.

Najar, A. S., Mitrovic, A., & McLaren, B. M. (2016). Learning with intelligent tutors and worked examples: selecting learning activities adaptively leads to better learning outcomes than a fixed curriculum. *UMUAI*, *26*(5), 459–491.

Renkl, A., Atkinson, R. K., Maier, U. H., & Staley, R. (2002). From example study to problem solving: Smooth transitions help learning. *The Journal of Experimental Education*, *70*(4), 293–315.

Salden, R. J., Aleven, V., Schwonke, R., & Renkl, A. (2010). The expertise reversal effect and worked examples in tutored problem solving. *Instructional Science*, *38*(3), 289–307.

Schwonke, R., Renkl, A., Krieg, C., Wittwer, J., Aleven, V., & Salden, R. (2009). The worked-example effect: Not an artefact of lousy control conditions. *Computers in Human Behavior*, *25*(2), 258–266.

Snow, R. E. (1991). Aptitude-treatment interaction as a framework for research on individual differences in psychotherapy. *Journal of consulting and clinical psychology*, *59*(2), 205.

Sweller, J., & Cooper, G. A. (1985). The use of worked examples as a substitute for problem solving in learning algebra. *Cognition and Instruction*, *2*(1), 59–89.

Van Gog, T., Kester, L., & Paas, F. (2011). Effects of worked examples, example-problem, and problem-example pairs on novices learning. *Contemporary Educational Psychology*, *36*(3), 212–218.

Vanlehn, K. (2006). The behavior of tutoring systems. *International journal of artificial intelligence in education*, *16*(3), 227–265.

Zhou, G., Azizsoltani, H., Ausin, M. S., Barnes, T., & Chi, M. (2019). Hierarchical reinforcement learning for pedagogical policy induction. In *International conference on artificial intelligence in education.*

Zhou, G., & Chi, M. (2017). The impact of decision agency & granularity on aptitude treatment interaction in tutoring. In *Proceedings of the 39th annual conference of the cognitive science society* (pp. 3652–3657).

Zhou, G., Lynch, C., Price, T. W., Barnes, T., & Chi, M. (2016). The impact of granularity on the effectiveness of students' pedagogical decision. In *Proceedings of the 38th annual conference of the cognitive science society* (pp. 2801–2806).

Zhou, G., Price, T. W., Lynch, C., Barnes, T., & Chi, M. (2015). The impact of granularity on worked examples and problem solving. In *Proceedings of the 37th annual conference of the cognitive science society* (pp. 2817–2822).

Zhou, G., Wang, J., Lynch, C., & Chi, M. (2017). Towards closing the loop: Bridging machine-induced pedagogical policies to learning theories. In *Edm* (pp. 112–119).

Robustness of Object Recognition under Extreme Occlusion in Humans and Computational Models

Hongru Zhu[1] **Peng Tang**[2] **Jeongho Park**[3] **Soojin Park**[4] **Alan Yuille**[1,2]
[1]Department of Cognitive Science, The Johns Hopkins University, Baltimore, MD 21218 USA
[2]Department of Computer Science, The Johns Hopkins University, Baltimore, MD 21218 USA
[3]Department of Psychology, Harvard University, Cambridge, MA 02138 USA
[4]Department of Psychology, Yonsei University, Seoul, Republic of Korea
hzhu38@jhu.edu {tangpeng723,p9j8h7,sjpark31,alan.l.yuille}@gmail.com

Abstract

Most objects in the visual world are partially occluded, but humans can recognize them without difficulty. However, it remains unknown whether object recognition models like convolutional neural networks (CNNs) can handle real-world occlusion. It is also a question whether efforts to make these models robust to constant mask occlusion are effective for real-world occlusion. We test both humans and the above-mentioned computational models in a challenging task of object recognition under extreme occlusion, where target objects are heavily occluded by irrelevant real objects in real backgrounds. Our results show that human vision is very robust to extreme occlusion while CNNs are not, even with modifications to handle constant mask occlusion. This implies that the ability to handle constant mask occlusion does not entail robustness to real-world occlusion. As a comparison, we propose another computational model that utilizes object parts/subparts in a compositional manner to build robustness to occlusion. This performs significantly better than CNN-based models on our task with error patterns similar to humans. These findings suggest that testing under extreme occlusion can better reveal the robustness of visual recognition, and that the principle of composition can encourage such robustness.

Keywords: visual recognition; occlusion; computational model; neural network; psychophysics

Introduction

Objects in the visual world are occluded much more than objects in typical visual science experiments. The ability to handle occlusion is essential for survival and everyday activities. For instance, in order to safely drive on the road, one must be able to swiftly detect other vehicles and pedestrians in advance even when they are only partially visible. However, both humans and object recognition models are rarely tested on the level of occlusion we encounter in the real world. Several studies have addressed this important issue by investigating real-world object recognition under occlusion (Tang et al., 2018; Rajaei, Mohsenzadeh, Ebrahimpour, & Khalligh-Razavi, 2018). These authors successfully developed models that could handle constant mask occlusion as shown in Figure 1 (a), and produced results consistent with human performance. In this paper, we take a step further and propose a more challenging task of object recognition under extreme occlusion to test humans and object recognition models.

We designed our task using a public occlusion image dataset from the computer vision community (Wang et al., 2017). In the proposed task, target objects are heavily occluded by several superimposed irrelevant real-world objects (occluders) with an average target occlusion ratio above 0.6, see Figure 1 (b). The biggest difference from previous studies is that both targets and occluders are real objects in real backgrounds. This task provides new insights in two ways. First, compared to testing in an occlusion-free domain, it challenges both humans and object recognition models and better tests the robustness of visual recognition. Second, it provides a way to check whether the ability to handle constant mask occlusion can entail robustness to real-world occlusion.

To begin with, we experimentally measured human performance on our task. Figure 1 (c) provides an example stimulus used in the behavioral experiments. Humans were very good at recognizing such extremely occluded objects and showed great robustness to occlusion. The results also suggest that our task is feasible and that an ideal object recognition model should be able to accomplish it.

Therefore, we subsequently tested several recent object recognition models on our task. The first model tested is the hierarchically feed-forward model, represented by convolutional neural networks (CNNs) (LeCun, Bengio, & Hinton, 2015). These models mimic the feed-forward process in biological vision. They achieve impressive performance and can explain some human data in several non-challenging visual tasks (DiCarlo & Cox, 2007; Yamins et al., 2014). However, there is still considerable variability in human neural and behavioral data at the individual image level that CNNs cannot explain (Rajalingham et al., 2018; Schrimpf et al., 2018). Our experiments show that CNNs perform very well without occlusion but their performance is not very good under extreme occlusion, suggesting that they lack robustness to occlusion.

The second model is a hybrid model that combines CNNs with models of recurrent computations. In biological vision, recurrent computations are essential for recognition under occlusion (Lamme & Roelfsema, 2000; Lamme, Zipser, & Spekreijse, 2002; Tang et al., 2018; Rajaei et al., 2018; Wyatte, Curran, & O'Reilly, 2012). Tang et al. (2018) modelled recurrent computations as lateral connections realized by a Hopfield network, which acted as a content addressable memory (Hopfield, 1982). They used Hopfield networks to store CNN activations of occlusion-free objects and later recover activations of occluded objects. This model improved CNN performance for recognition under constant mask occlusion, but we did not observe improvements under extreme occlusion, implying that the ability to handle constant mask occlusion does not entail robustness to real-world occlusion.

(a) (b) (c)

Figure 1: Examples of different types of occlusion (on car images). (a) An image of constant mask occlusion used in (Tang et al., 2018). (b) An image of extreme occlusion used to test computational models in our paper. (c) An image of extreme occlusion used in our behavioral experiments.

As a comparison to the models above, we propose a third model designed with the principle of composition. Throughout this paper, we use the term "composition" in its traditional sense meaning the process where smaller parts are composed together to form larger parts. This principle is not inherently addressed by CNNs (Stone et al., 2017), but it is supported by biological evidence showing that populations of neurons in macaque V4 and IT areas represent complete shapes with aggregates of shape fragments (Brincat & Connor, 2006; Pasupathy & Connor, 2002; Yamane, Carlson, Bowman, Wang, & Connor, 2008). Our model uses two stages to recognize objects from parts/subparts in a compositional manner. In the first stage, we consider spatial relations among subparts and detect parts. In the second stage, we consider spatial relations among detected parts and compose them into objects. Both stages are designed to be robust to occlusion. This two-stage model can handle missing parts under occlusion as long as the visible parts conform to reasonable spatial constraints. Our model performed better than the other two models under extreme occlusion with similar error patterns to humans, demonstrating a way to build robustness to occlusion by exploiting object compositional structures.

Task and Dataset

Recognition Under Extreme Occlusion Task

Object recognition under occlusion involves recognizing targets occluded by other entities (called occluders). Depending on task specifics, targets could be as simple as letters, digits, and symbols, or as complex as objects in real scenes. Occluders can also vary substantially. There are simple occluders like constant masks and also complex ones like real objects.

Simple occlusion by constant masks and complex occlusion by real objects are actually treated differently during recognition. For instance, in CNNs, neurons tuned to fur textures may fire on the presence of cats as occluders and distort the CNN activation of the target. However, fewer misguiding neurons are likely to fire on constant masks without textures. Thus, real objects as occluders are more likely to distort target object recognition by providing irrelevant context.

To test the robustness of visual recognition to real-world occlusion, we propose a difficult task of object recognition under extreme occlusion. Specifically, it involves recognizing vehicles occluded by other irrelevant real occluders, including animals, furniture and other objects, in real backgrounds (see Figure 1 (b)). This task is challenging because occluders are irrelevant real objects and the occlusion ratio is high, which discourages the use of context during recognition.

Training and Testing Dataset

For the purpose of training computational models to perform our task, we propose that only occlusion-free images should be used. A single object can be occluded in an exponential number of ways and it is unlikely for a limited training set to cover all occluder appearances, positions and so on. Therefore, we used 4049 occlusion-free training images covering five types of vehicles, including *aeroplane*, *bicycle*, *bus*, *car* and *motorbike* from VehicleSemanticPart dataset (Wang et al., 2017). 113 different types of object parts, such as car wheels, bicycle pedals and jet engines, are annotated with part identities and bounding box positions.

For testing purposes, we built an occlusion testing set using another 500 images from VehicleOcclusion dataset (Wang et al., 2017). To our knowledge, this is the only public occlusion dataset with accurate occlusion annotations of parts and objects. In each image, 2-4 randomly-positioned real occluders are placed onto the single target object (see Figure 1 (b)). The target occlusion ratio is constrained. 77% of the images have an occlusion ratio of 0.6-0.8; 18.4% of the images have an occlusion ratio of 0.4-0.6; 4.6% of the images have an occlusion ratio of 0.2-0.4. Furthermore, we also created an occlusion-free testing set by collecting the corresponding 500 clean images before occluders were placed. Neither these clean images nor superimposed occluders are met in the training set.

For evaluation metrics, we evaluated human and model performance both quantitatively by recognition accuracy and qualitatively by confusion matrices and representational dissimilarity matrices (Kriegeskorte, Mur, & Bandettini, 2008).

Behavior Experiments

Participants

We designed a survey on Amazon Mechanical Turk to collect human responses for the task of object recognition under extreme occlusion. 25 human subjects completed our survey.

Procedure

We had 1,250 human intelligence tasks (HITs) with 20 stimuli in each, so that there were 50 repetitions of the occlusion testing set (500 stimuli). In each HIT, subjects were given unlimited time to observe and respond to 20 stimuli one at a time. The stimuli had red bounding boxes around the targets (see Figure 1 (c)). Subjects were asked to type the names of the objects in the bounding boxes without knowing that they should belong to the aforementioned five categories.

Data Processing and Exclusions

We collected 25,000 typed strings as subject responses. There were 785 different strings from the responses and we manually assigned them to the five vehicle categories. We first

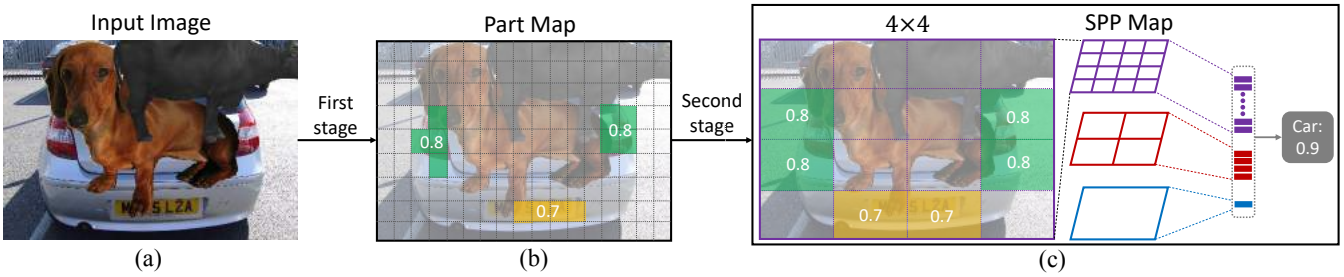


Figure 2: The overall framework of our two-stage model. The input image (a) contains a car whose back trunk is occluded. In the first stage, we detect several parts, including tail lights (green) and the license plate (yellow), and output part maps (b). In the second stage, we apply spatial pyramid pooling (SPP) to part maps and obtain SPP maps (c) (4×4 scale shown here). We consider spatial constraints over parts and aggregate part confidence scores at different scales to determine the object identity.

excluded 300 responses whose corresponding images were oversized for computational models. We further excluded 5,359 responses assigned to either none or more than one category. The rest of 19,341 responses with valid reported and ground-truth category labels were used for data analysis.

Computational Models

CNNs: AlexNet, ResNet and VGG16

Recently, CNNs achieved impressive performance in object recognition. They use stacked convolutional layers and pooling layers to extract image features for classification. We used AlexNet (Krizhevsky, Sutskever, & Hinton, 2012), ResNet (He, Zhang, Ren, & Sun, 2016) and VGG16 (Simonyan & Zisserman, 2014) as three representative CNN-based object recognition models and modified them to get the best possible performance. In AlexNet, we substituted fully-connected layers fc6 and fc7 with two equivalent convolutional layers with kernel size 6×6 and 1×1, followed by a global average pooling layer. In VGG16, we also made similar modifications to the fully-connected layers. In ResNet, we used ResNet-18 and substituted the last average pooling layer with a global average pooling layer. These modifications enable the networks to handle inputs of variable sizes. The classification layers were also changed to fit our five object categories.

Training and Testing The training and testing images were resized so that the short object edge had 224 pixels. During the training phase, we froze weights in earlier layers and fine-tuned ImageNet pre-trained models. In AlexNet, we froze weights in the first two convolutional layers. In ResNet, we froze weights in the first two residual stages. In VGG16, we froze weights in the first two convolutional stages. The training inputs were randomly cropped image patches of size 224×224 containing at least part of the target objects. During the testing phase, inputs were full-sized images.

Hybrid Model: AlexNet+Hopfield Network

Tang et al. (2018) proposed a hybrid model of CNNs with Hopfield networks that improved object recognition under constant mask occlusion and produced results consistent with human performance. Concretely, they adopted a fully-connected Hopfield network with binary threshold nodes. It can store patterns as local minima and later recover incomplete patterns by iteratively processing them until convergence. The capacity of a Hopfield network with N nodes is only about 0.15N memories. When the number of patterns to store increases, local minima are more likely to be spurious minima. We followed the experimental settings from Tang et al. (2018) and used 4096-dimensional features from the fc7 layer in ImageNet-trained AlexNet as patterns to store and recover in the Hopfield network.

Training and Testing During the training phase, we fed a random subset of 500 resized images (224×224) to ImageNet-trained AlexNet. Following Tang et al. (2018), we extracted fc7 features, binarized them with a threshold of 0 and used them to train a Hopfield network with 4096 nodes (implemented in MATLAB's newhop function) and a linear multiclass Support Vector Machine (SVM). We used a small training set due to the limited capacity of the Hopfield network. During the testing phase, we fed images to AlexNet, binarized fc7 features and used them to initialize the Hopfield network. Each node in the network receives weighted inputs from connected nodes and updates its binary state accordingly and synchronously. Converged outputs (timestep=256) are classified using the SVM.

Ours: Two-stage Voting Model

Motivation Object appearance and context can change drastically under occlusion yet spatial constraints over objects and parts are largely preserved. With a few parts missing, an occluded object can be recognized as long as the positions of visible parts make sense, that is, if they conform to reasonable spatial constraints. This motivates us to exploit object compositional structures for object recognition under occlusion. We developed a two-stage object recognition model that could utilize these spatial constraints (Figure 2). Specifically, in the first stage, we detected different object parts in the images. In the second stage, we considered spatial constraints over objects and parts and used those detected parts to determine object identities. In both stages, we utilized deep networks

to capture these spatial constraints via spatial voting mechanisms. The core idea of spatial voting is to detect larger parts by considering spatial relations of smaller parts and gathering their votes, which are confidence scores for their presence. We now elaborate our two stages in more details.

Stage 1: Part Detection In the first stage, we want to robustly detect object parts. Zhang, Xie, Wang, Xie, and Yuille (2018) developed a robust voting model to detect semantic parts under occlusion. We adopted their method and produced part maps as shown in Figure 2 (b), which contained confidence scores for the presence of different object parts at different locations. Following Zhang et al. (2018), we first obtained subparts by clustering corresponding CNN intermediate activations at the *pool-4* layer from a VGG network (Simonyan & Zisserman, 2014). We subsequently implemented the spatial voting mechanism as a convolutional layer with kernel size 15×15 on top of the subparts to capture spatial constraints over subparts and parts in a larger spatial region. The intuition for this voting stage is that a partially occluded object part could be robustly detected as long as it gathered enough votes from a set of subparts which conformed to certain spatial constraints in a spatial region.

Stage 2: Object Recognition In this stage, we used part maps as inputs and designed another voting method to aggregate parts to form objects. There are two challenges for our voting method. First, it needs to capture spatial constraints over parts and produce fixed-length vectors for classification. Second, it needs to tolerate within-category variation so that the learned spatial constraints are generalized for most objects in a category. To this end, we applied spatial pyramid pooling (SPP) (Lazebnik, Schmid, & Ponce, 2006) at three different scales (4×4, 2×2, 1×1) to the normalized part maps and obtained three SPP maps as shown in Figure 2 (c). This method maintained some spatial information and was different from the sliding window pooling of deep networks. Concretely, we evenly divided part maps into $n \times n$ local spatial bins ($n = 4, 2, 1$) and applied max pooling to each bin. SPP maps contained maximum confidence scores for the presence of parts in each spatial bin. Later, we concatenated SPP maps and appended a dropout layer (dropout ratio 0.1) and a fully-connected layer. The dropout layer randomly dropped a subset of part votes during training and improved the robustness of our model under occlusion. The fully-connected layer aggregated part votes and learned spatial constraints over parts at different scales to determine the object identities. The learned weights of this layer are referred to as object-part spatial heatmaps (visualized in Figure 3). Both stages in our model were robust to occlusion due to the use of spatial voting and dropout mechanisms, which allowed larger parts to be detected using only a subset of smaller parts under certain spatial constraints.

Training and Testing Training and testing images were resized so that the short object edge had 224 pixels. Two stages were trained separately. When training the first stage, inputs were randomly cropped image patches of size 224×224 containing at least part of target objects. More details are available in (Zhang et al., 2018). When training the second stage, inputs were parts maps obtained by feeding training images to the first stage. During testing, inputs were full-sized images.

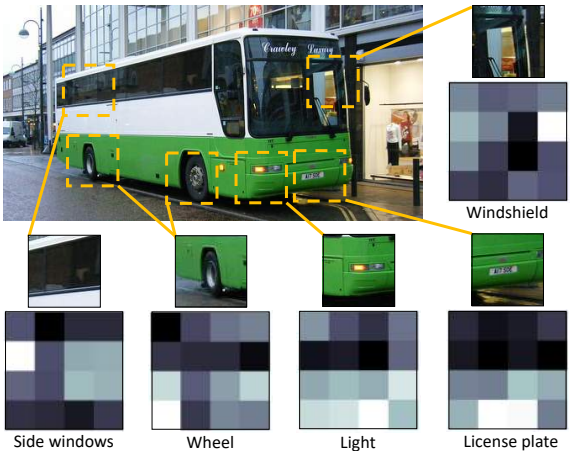


Figure 3: Visualization of object-part spatial heatmaps at the 4×4 scale for the bus category. Each heatmap shows the learned spatial constraints between the bus and the part. Brighter regions indicate higher voting weights. For instance, a license plate in the lower region of a image often casts a highly weighted vote in favor of the presence of a bus.

Results and Discussions

Testing without Occlusion

First, we test three computational models on the task of recognizing occlusion-free vehicles. Human subjects are not tested for this task because it is very easy for humans when they are given unlimited time to recognize vehicles without occlusion.

Both CNNs and our model perform reasonably well on this task (Table 1). Our model has a comparable accuracy (92.9%), showing that spatial constraints over parts/subparts are useful information for the recognition of occlusion-free objects. However, these results tell little about the robustness of different models given their similar performances.

The hybrid model of AlexNet and Hopfield networks gets a relatively lower accuracy without occlusion. When we use SVM to directly classify binarized fc7 features without using the Hopfield network, the accuracy increases from 77.7% to 85.4%. This implies that the relatively lower accuracy may be caused by the Hopfield network. In order to check whether the Hopfield network setting was implemented correctly, we followed Tang et al. (2018) and tested the hybrid model on mask occlusion images from five categories with an average occlusion ratio above 0.7; see Figure 1 (a) for an example testing image. The use of the Hopfield network increased the accuracy from 40.9% to 46.8%, which was qualitatively similar to the improvement reported in Tang et al. (2018). This shows that the Hopfield network was implemented correctly and improved object recognition performance under constant

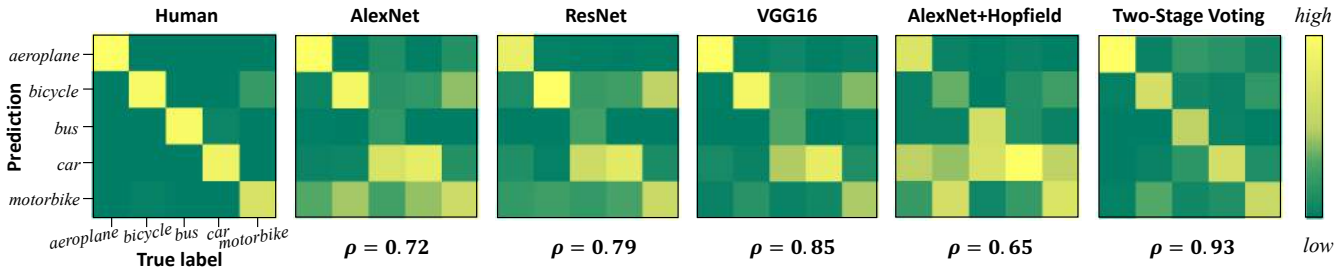


Figure 4: Category-level confusion matrices under extreme occlusion. The Pearson correlation coefficients between the human confusion matrix and each model confusion matrix are listed below the matrices.

Table 1: Testing Accuracy under No/Extreme Occlusion.

Humans/Models	w/o occlusion	w/ occlusion
Humans	-	**93.3%**
AlexNet	89.8%	50.0%
ResNet	90.1%	54.0%
VGG16	**94.7%**	62.6%
AlexNet+Hopfield	77.7%	46.0%
Two-stage Voting (Ours)	92.9%	**67.0%**
Ablation 1	91.2%	47.5%
Ablation 2	89.9%	58.9%

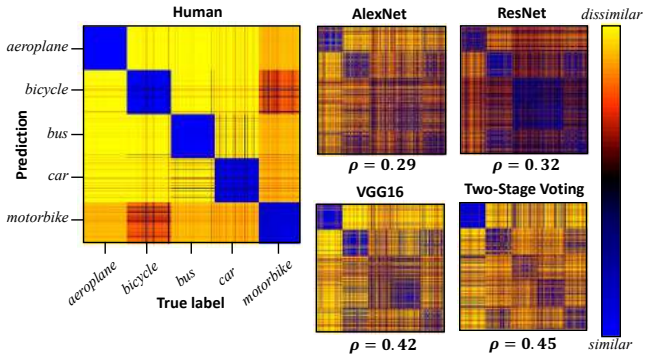


Figure 5: Image-level representational dissimilarity matrices (RDMs) under extreme occlusion. Each testing image is characterized by a 5-dimensional categorical distribution obtained from either human response frequencies for five vehicle categories on that image or the Softmax output in the final layer of each model. The dissimilarity between two images is measured as the Euclidean distance between two vectors representing the associated categorical distributions. The Pearson correlation coefficients between the human RDM and each model RDM are listed below the matrices.

mask occlusion. We will further discuss the possible causes of the relatively low accuracy of the hybrid model on our task later with testing results under extreme occlusion.

Testing under Extreme Occlusion

We further test humans and these models on recognizing objects under extreme occlusion with our occlusion testing set.

Table 1 shows that humans have very high accuracy at recognizing occluded vehicles and are robust to extreme occlusion. It also confirms that our task of object recognition under extreme occlusion is feasible and the information in these occlusion images is sufficient to determine object identities.

For CNNs, the accuracy is relatively low (Table 1). Despite their good performance without occlusion, CNNs do not manifest robustness under extreme occlusion as humans do. Our results support previous findings that CNN activation is not inherently compositional and cannot explicitly address contextual and non-contextual information (Stone et al., 2017).

For the hybrid model, we do not observe performance gains compared to CNNs under extreme occlusion. If we use SVM to classify binarized fc7 features without using Hopfield networks, the accuracy rises from 46.0% to 48.6%. This result together with previous ones shows that the Hopfield network did not improve performance on our task, either without occlusion or under extreme occlusion, although it improved performance under constant mask occlusion. There are two possible reasons. First, it may require more representative training features because our dataset is more complex than the one from Tang et al. (2018). However, given the limited capacity of Hopfield networks, too many training features may result in more spurious minima. Second, mask occlusion may only suppress some neurons in CNNs while real occluders can activate additional misguiding neurons, making it hard for pattern recovery. Thus, the ability to handle constant mask occlusion does not entail robustness to real-world occlusion.

Our model outperforms other models under extreme occlusion in terms of accuracy. We further compare the performance of humans and different models by analyzing their category-level confusion matrices (Figure 4) and image-level representational dissimilarity matrices (RDMs) (Figure 5). These provide a better way to qualitatively compare the robustness of human vision and these computational models. Although the accuracy of our model is still lower than humans, it shows greater robustness than other models and produces the most similar results to humans. Figure 6 also shows some representative improvements and errors from our model. The performance of our model suggests that spatial constraints over parts are important cues for object recogni-

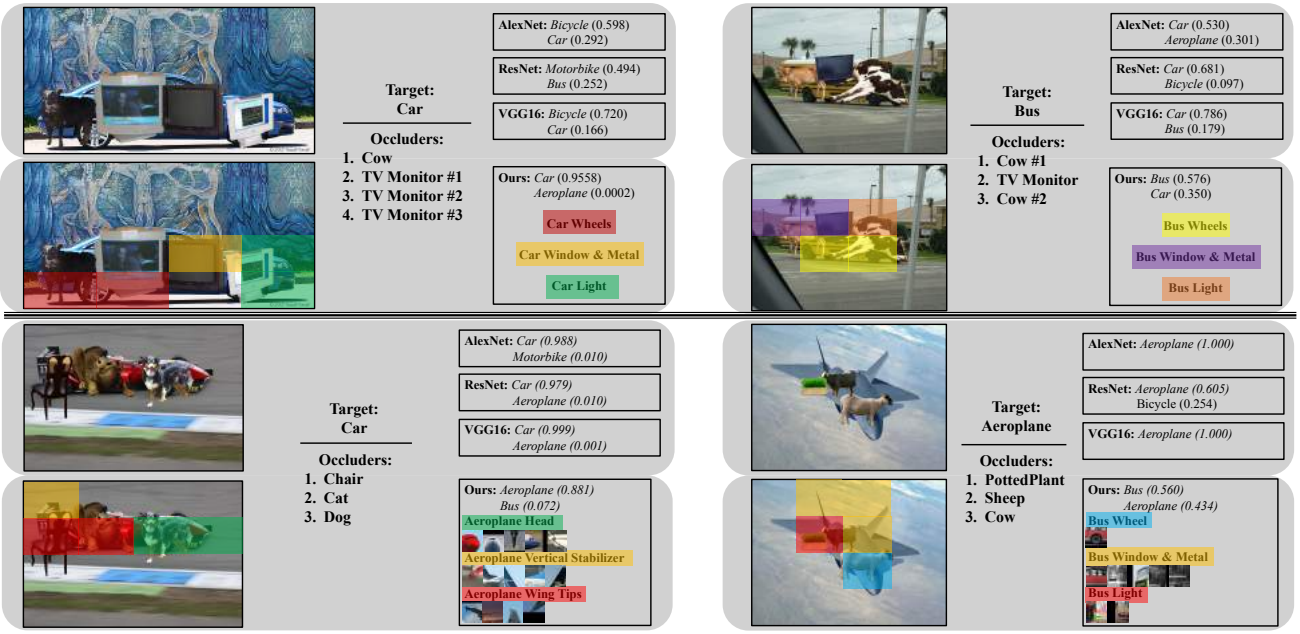


Figure 6: Representative improvements (top two) and errors (bottom two) from our model. In the top two cases, our model produced correct object labels by exploiting spatial relations among detected object parts. In the bottom left case, our model misclassified the race car as an aeroplane partly because they are similar in the aerodynamic design. In the bottom right case, our model incorrectly detected bus parts and was subsequently misled by these false part detection results.

tion under occlusion, and that the principle of composition is a promising solution for bridging the gap between the robustness of human vision and these computational models.

Ablation Experiments

Finally, we show the effectiveness of each voting stage in our model. We substituted each stage respectively with alternative models to do the same task with the same supervision.

In the first experiment (Ablation 1 in Table 1), we substituted the first voting stage with Faster-RCNN (Ren, He, Girshick, & Sun, 2015), a state-of-the-art object detection model. We trained Faster-RCNN to detect parts and obtained SPP maps. Later, we trained our second stage and tested the whole model. The accuracy changed little without occlusion but dropped from 67.0% to 47.5% under extreme occlusion. As Zhang et al. (2018) pointed out, it is difficult for proposal-based detection methods including Faster-RCNN to extract good proposals under occlusion and even with correct proposals, the classifier may still go wrong due to the presence of occluders. This result further confirmed the robustness of the first stage model on detecting parts under occlusion.

In the second experiment (Ablation 2 in Table 1), we substituted the second stage with a bag-of-words module where all spatial relations were discarded by a global max pooling layer. We concatenated the highest confidence score in each part map into a vector and appended a dropout layer and a fully-connected layer. We trained the bag-of-words module and tested the whole model. The accuracy changed little without occlusion but dropped from 67.0% to 58.9% under extreme occlusion, implying the effectiveness of spatial voting under occlusion. Figure 3 also shows that the learned spatial constraints are meaningful. Wheels and windshields often appear in lower and higher regions of bus images respectively.

Together, the results suggest that the higher accuracy of our model is not purely a result of additional part-level supervision but also due to the use of object compositional structures.

Conclusion

Occlusion is often present in everyday visual tasks yet humans and models are rarely tested under real-world occlusion. We proposed a task of object recognition under extreme occlusion and tested humans and models, including CNNs, a hybrid model of CNNs with Hopfield networks and our two-stage voting model. Our findings lead us to three conclusions.

First, testing under extreme occlusion can better reveal the robustness of visual recognition than testing without occlusion. Object recognition models that can compete with humans in the occlusion-free domain may not show the same robustness under extreme occlusion as humans do.

Second, the ability to handle constant mask occlusion does not entail robustness to real-world occlusion. Different types of occlusion may alter context differently yet object inherent structures could still be exploited for recognition purposes.

Third, the performance of our model is better and more correlated with human results under occlusion, suggesting that the principle of composition is a possible solution for building robustness to occlusion as demonstrated by human vision.

Acknowledgments

We thank Tal Linzen, Dan Kersten, Tom McCoy and the JHU CCVL group for helpful comments. This work was supported by ONR with grant N00014-19-S-B001.

References

Brincat, S. L., & Connor, C. E. (2006). Dynamic shape synthesis in posterior inferotemporal cortex. *Neuron*, *49*(1), 17–24.

DiCarlo, J. J., & Cox, D. D. (2007). Untangling invariant object recognition. *Trends in Cognitive Sciences*, *11*(8), 333–341.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of CVPR*.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, *79*(8), 2554–2558.

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*, 4.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Proceedings of NIPS*.

Lamme, V. A., & Roelfsema, P. R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends in Neurosciences*, *23*(11), 571–579.

Lamme, V. A., Zipser, K., & Spekreijse, H. (2002). Masking interrupts figure-ground signals in v1. *Journal of Cognitive Neuroscience*, *14*(7), 1044–1053.

Lazebnik, S., Schmid, C., & Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. *Proceedings of CVPR*.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436.

Pasupathy, A., & Connor, C. E. (2002). Population coding of shape in area v4. *Nature Neuroscience*, *5*(12), 1332.

Rajaei, K., Mohsenzadeh, Y., Ebrahimpour, R., & Khaligh-Razavi, S.-M. (2018). Beyond core object recognition: Recurrent processes account for object recognition under occlusion. *bioRxiv*, 302034.

Rajalingham, R., Issa, E. B., Bashivan, P., Kar, K., Schmidt, K., & DiCarlo, J. J. (2018). Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *Journal of Neuroscience*, *38*(33), 7255–7269.

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Proceedings of NIPS*.

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., ... others (2018). Brain-score: Which artificial neural network for object recognition is most brain-like? *bioRxiv*, 407007.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Stone, A., Wang, H., Stark, M., Liu, Y., Scott Phoenix, D., & George, D. (2017). Teaching compositionality to cnns. *Proceedings of CVPR*.

Tang, H., Schrimpf, M., Lotter, W., Moerman, C., Paredes, A., Caro, J. O., ... Kreiman, G. (2018). Recurrent computations for visual pattern completion. *Proceedings of the National Academy of Sciences*, *115*(35), 8835–8840.

Wang, J., Zhang, Z., Xie, C., Zhu, J., Xie, L., & Yuille, A. L. (2017). Detecting semantic parts on partially occluded objects. *Proceedings of BMVC*.

Wyatte, D., Curran, T., & O'Reilly, R. (2012). The limits of feedforward vision: recurrent processing promotes robust object recognition when objects are degraded. *Journal of Cognitive Neuroscience*, *24*(11), 2248–2261.

Yamane, Y., Carlson, E. T., Bowman, K. C., Wang, Z., & Connor, C. E. (2008). A neural code for three-dimensional object shape in macaque inferotemporal cortex. *Nature Neuroscience*, *11*(11), 1352.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624.

Zhang, Z., Xie, C., Wang, J., Xie, L., & Yuille, A. L. (2018). Deepvoting: A robust and explainable deep network for semantic part detection under partial occlusion. *Proceedings of CVPR*.

Why Decisions Bias Perception: An Amortised Sequential Sampling Account

Jian-Qiao Zhu (J.Zhu@warwick.ac.uk)
Warwick Business School, University of Warwick
Adam N. Sanborn (A.N.Sanborn@warwick.ac.uk)
Department of Psychology, University of Warwick
Nick Chater (Nick.Chater@wbs.ac.uk)
Warwick Business School, University of Warwick
Coventry, UK CV4 7AL

Abstract

The judgments that people make are not independent – initial decisions can bias later perception. This has been shown in tasks in which participants first decide whether the direction of moving dots is to one side or the other of a reference line: their subsequent estimates are biased away from this reference line. This interesting bias has been explained in past work as either a consequence of weighting sensory neurons, or as a consequence of participants adjusting their estimate to match their decision. We propose a new explanation: that people sequentially sample evidence to make their decision, and reuse these samples to make their estimate (i.e., amortised inference). Because optimal stopping leads to samples that strongly favor one or another decision alternative, the subsequent estimates are also biased away from the reference line. We introduce a sequential sampling model for posterior samples that does not assume constant thresholds, and provide evidence for our explanation in a new experiment that generalizes the perceptual bias to a new domain.

Keywords: decision biases, adaptive sampling, amortised inference.

Introduction

Experiments in motion perception show that making a perceptual decision biases subsequent perception. As illustrated in Figure 1A, participants in these random-dot-motion experiments are first asked whether the motion was clockwise (CW) or counter-clockwise (CCW) of a decision boundary. After making this decision, participants are then asked to estimate the direction of motion. While participants' estimates are unsurprisingly consistent with their decision, these estimates also show a surprising perceptual bias: estimates are biased *away from* the decision boundary (Jazayeri & Movshon, 2007; Luu & Stocker, 2018; Zamboni, Ledgeway, McGraw, & Schluppeck, 2016).

Two main theories have been proposed to explain this perceptual bias: (a) the optimal weighting of outputs of orientation-tuned neurons used in the decision task is also used in the estimation task (Jazayeri & Movshon, 2007), or (b) people employ self-consistent reasoning by only considering hypotheses consistent with their initial decision when making an estimate (Luu & Stocker, 2018). Existing comparisons favor the self-consistency account (Luu & Stocker, 2018; Zamboni et al., 2016).

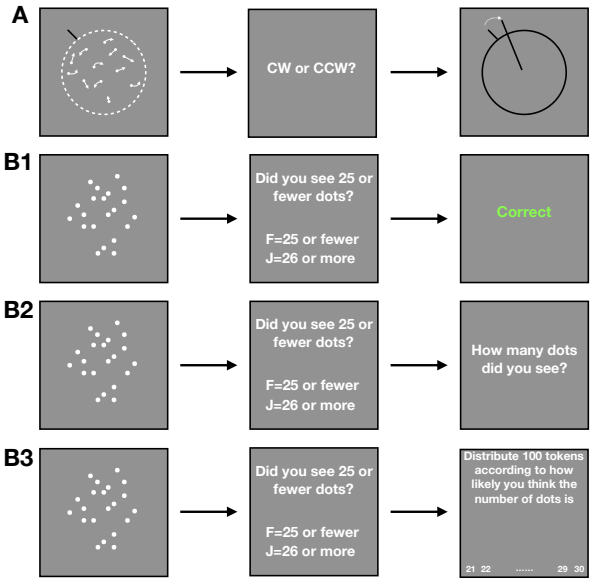


Figure 1: Illustration of experimental tasks. **(A)** Decision-estimation (D-E) task for random-dot-motion. **(B)** Numerosity experiment. **(B1)** Decision with feedback (D-F) task. **(B2)** D-E task. **(B3)** Decision-histogram (D-H) task.

However, it may be that self-consistency is unnecessary to explain the perceptual bias. We take a sequential sampling approach to modelling this task, following a long history of models in human decision making that sequentially draw perceptual or posterior samples and optimally accumulate evidence (Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006; Vul, Goodman, Griffiths, & Tenenbaum, 2014). We introduce a sequential sampling model that optimally accumulates posterior samples, and then demonstrate that the perceptual bias in estimation is produced by simply averaging samples that were optimally accumulated for the initial decision. The intuition for why the bias is produced is simple: because sequential sampling models stop when the samples favor one of the alternatives, the estimate (i.e., the average) also favors one of the alternative. Interestingly, similar Bayesian analysis have arisen in a different experimental domain: the studies of probability estimation from sequential samples, where

optimal stopping has shown to predict a distorting effects on subsequent judgments (Coenen & Gureckis, 2016). However, when the sampling process is external, such distortion did not receive empirical supports (Coenen & Gureckis, 2016).

To discriminate between these accounts, we first show that weighted decoding, self-consistency, and sequential sampling make qualitatively different predictions about the perceptual beliefs that people will have about individual stimuli. We then test these predictions in a new experiment which generalizes the perceptual bias from a simple random-dot-motion task to a perceptually more complex numerosity task (see Figure 1B), finding that the perceptual bias is best explained by sequential sampling.

Computational Models

In this section, we introduce and compare computational models of the perceptual bias.

Weighted Decoding

The Weighted Decoding (WD) model argues that post-decision bias is a result of optimally tuning the sensory representation to boost responses for the initial decision (Jazayeri & Movshon, 2007; Zamboni et al., 2016). For example, to discriminate whether a random-dot-motion stimulus is coherently moving clock-wise or counter-clockwise of a reference line, the neurons that respond maximally to motion directions that are slightly different from the reference line are the most informative. This leads to a optimal weighting profile that is bimodal: emphasizing directions that are slightly away from the reference line.

WD assumes this weighting profile is also used in the estimation task, and that the mode of the weighted sensory distribution is taken as the estimate. As a result, a post-decision bias naturally emerges (Figure 2A).

Self-Consistency

To predict the post-decision bias, WD must assume that the selective read-out of sensory information in the decision task is carried over to the subsequent estimation task. However, more recent work has demonstrated that the perceptual bias is actually a late decision-related bias, rather than a sensory bias (Luu & Stocker, 2018; Zamboni et al., 2016).

The Self-Consistency (SC) model is a Bayesian model that makes the initial decision according to which option has highest posterior probability, given noisy sensory evidence. However, because the estimate is made after the decision, this model assumes that the quality of the sensory evidence has decayed by the time participants are asked to make an estimate. Instead of relying on the low-quality sensory evidence alone, SC assumes that the participant treats their initial decision (which was made with high-quality sensory evidence) as information as well (cf. Fleming & Daw, 2017), only considering hypotheses that are consistent with the initial decision. SC's estimate is then the mean of the posterior distribution over hypotheses consistent with the initial choice. As shown in Figure $2B_1$, SC also produces estimates that are biased away from the decision boundary.

Our implementation of SC also predicts a bias toward the decision boundary for true stimuli that are far away from the boundary. This is for an uninteresting reason: in the task we will describe below the response range was restricted, and so we also truncated the posterior at the edges of the allowable response range – this leads presentations of extreme values to be biased toward the center of the range.

Simple Amortised Sampling

Because perfectly storing and representing probability distributions can easily become computationally daunting, sample-based approximations have been proposed as a way for the brain to approximate Bayesian inference (Sanborn & Chater, 2016; Zhu, Sanborn, & Chater, 2018). On each trial, the Simple Amortised Sampling (SAS) model generates a set of N samples from the posterior distribution:

$$x_i \overset{\text{i.i.d}}{\sim} P(X|S), \quad i = 1, 2, ..., N \tag{1}$$

where $P(X|S)$ is the posterior sensory representation of number of dots given the stimulus. For the decision task, SAS chooses the alternative that attracts the larger number of samples, which introduces a natural stochasticity into the decision.

In the later estimation task, it makes little sense to draw a new set of samples, as an average of the samples drawn to make the decision can serve as the estimate. This effort-saving strategy is a form of *amortised inference* (Gershman & Goodman, 2014). Reusing samples in this way ensures a high degree of consistency between the decision and the estimate SAS make. However, SAS will not produce a perceptual bias away from the decision boundary – the average of a fixed number of samples is unbiased (Figure $2C_3$), and it only shows a bias toward the center for extreme stimuli. Thus, this model is not actually a candidate for explaining the perceptual bias, but instead serves to illustrate why the following model does.

Bayesian Amortised Sequential Sampling

If samples are drawn sequentially and require effort to generate, it often makes no sense to continue sampling until a fixed number are obtained. Instead, it is more efficient to stop sampling when it is no longer worthwhile.

Many different kinds of sequential sampling models have been proposed, including those that accumulate sensory information (Bogacz et al., 2006), and those that accumulate the kind of posterior samples similar to SAS (Vul et al., 2014). We take as a starting point the sequential model introduced in Vul et al. (2014), which accumulates samples until there are a threshold T more in favor of one alternative than the other. This scheme has the advantages of producing a fixed probability of choosing the better alternative regardless of the number of samples, and it is possible to find the optimal threshold for maximizing utility.

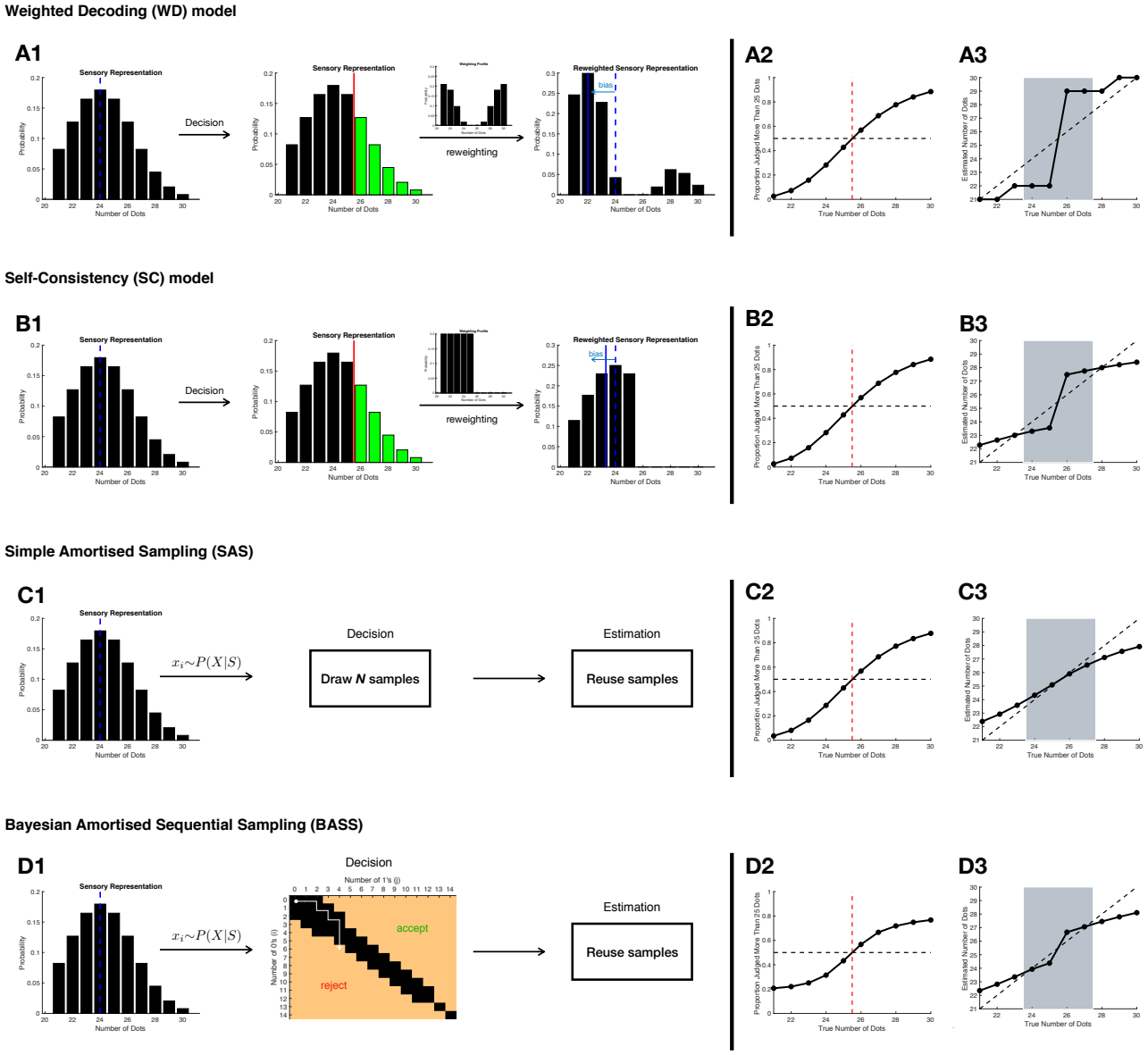


Figure 2: A comparison of model mechanisms and predictions for the numerosity task. For illustrative purposes, we assume a Gaussian likelihood of dot numbers. Then, the posterior distribution combines this likelihood with the prior (i.e., a uniform distribution from 21 to 30). A posterior distribution with its mean at 24 dots (dashed blue line) is shown for each model. **(A)** Schematic illustration of the WD model (Jazayeri & Movshon, 2007). The sensory representation is reweighted according to the optimal bimodal weighting profile. The initial decision is based on the relative probability above and below the decision boundary. The estimate is then the mode of this same reweighted sensory representation. **(B)** Schematic illustration of the SC model (Luu & Stocker, 2018). The initial decision is based on whether there is more probability above or below the decision boundary. The estimate is the mean of the portion of the posterior distribution that is consistent with the initial decision. **(C)** Schematic illustration of the SAS model. A fixed number of posterior samples are drawn and the alternative that attracts the larger number of samples is chosen in the initial decision. The average of these samples is the estimate. **(D)** Schematic illustration of the BASS model. Samples are sequentially drawn from the posterior distribution until it is no longer worthwhile to continue. The alternative that attracts the larger number of samples is chosen in the initial decision. The average of these samples is the estimate.

However, it is possible to sequentially draw samples from a posterior distribution even more efficiently by allowing a non-constant threshold and determining before every sample whether it is better to continue or stop sampling. We term this

scheme Bayesian Amortised Sequential Sampling (BASS). The problem of finding for the optimal changing threshold was solved by Wald (1950) for deciding when to stop drawing binomial samples from an external source, and a similar approach to external samples was investigated empirically by Coenen and Gureckis (2016). We simply adopt Walds approach to optimally drawing internal posterior samples.

The posterior probability p that one decision alternative is true is assumed to be unknown, but we assume that binomial samples can be sequentially drawn with probability p. We perform Bayesian inference using the obtained samples, by first placing a prior distribution over p and assume a fixed cost c of drawing a sample, reflecting the time and effort of doing so. After drawing j samples in favor of a decision alternative and i against, we denote p_{ij} as the posterior probability of a decision alternative given those samples, with p_{00} being the prior probability. The binary decision task essentially becomes a sequential test on whether $p_{ij} < 1/2$ and the optimal stopping rule for this test can then be derived from the following using dynamic programming:

$$F(i,j) = \min \begin{cases} F_0(i,j), \\ c + p_{ij}F(i,j+1) + (1 \pm p_{ij})F(i+1,j). \end{cases} \tag{2}$$

where $F(i,j)$ and $F_0(i,j)$ are respectively the expected cost of sampling and expected cost of termination after i samples against and j in favor have been observed. The sampling process should terminate whenever $F(i,j) \geq F_0(i,j)$.

Because $F_0(i,j)$ represents the expected cost of stopping the sampling process when the posterior probability of an alternative is p_{ij}, if the punishment for an incorrect decision is one unit of utility and thus,

$$F_0(i,j) = \min \begin{cases} i/(i+j), \\ j/(i+j). \end{cases} \tag{3}$$

This is the expected cost of incorrectly choosing an alternative when the posterior probability of that alternative is $p_{ij} = \text{Beta}(i,j)$.

The expected cost of drawing another sample is the sum of (a) the cost of generating one sample c, (b) the expected cost if the new sample turns out to be in favor $p_{ij}F(i,j+1)$, and (c) the expected cost if the new sample turns out to be against $(1 \pm p_{ij})F(i+1,j)$.

While the exact solution of the Bayesian optimal stopping problem is difficult to obtain, once computed it can also be reused across different cognitive tasks. For illustrative purpose, we set cost of collecting one sample $c = 0.006$ and prior probability to $\text{Beta}(1,1)$. This leads to the termination conditions for sequential sampling in Figure $2D_1$, which shows a collapsing threshold.

We assume that BASS is performing amortised inference, and because the decision and estimation tasks are so similar, the samples drawn for decision are simply averaged to produce the estimate. Like SAS, BASS produces a high degree of consistency between decision and estimate and a biased toward the decision boundary for extreme stimuli. However, unlike SAS, BASS produces estimates that are biased away from the decision boundary for central stimuli (Figure $2D_3$). The reason for the model's behavior can be seen in the termination conditions shown in Figure $2D_1$. The sampling process is very unlikely to stop when there are an equal number of samples in favor of the two alternatives, instead waiting until there are more samples in favor of one of the alternatives. Then, after averaging the resulting samples to produce an estimate, these estimates are unlikely to be close to the decision boundary.

Table 1: Summary of model predictions on empirical effects

Effects	WD	SC
Decision bias	yes	yes
Self consistency	low	high
Belief distribution	bimodal	one-sided
	SAS	BASS
Decision bias	no	yes
Self consistency	high	high
Belief distribution	undistorted	favors one side

Note. WD=Weighted Decoding, SC=Self-Consistency, SAS=Simple Amortised Sampling, BASS=Bayesian Amortised Sequential Sampling.

Comparing the Models

As seen across Figure 2, qualitatively similar patterns of decision and estimation bias are predicted by the WD, SC, and BASS models. What distinguishes the models are the beliefs about the probability of each possible response in the estimation task. WD predicts that the optimal weighting will result in a bimodal belief distribution. SC predicts a one-sided belief distribution: that only estimates consistent with the decision will be considered. In contrast, BASS predicts that people will believe that several estimates are possible, including a low probability of those that are not consistent with the decision (see Table 1 for a summary).

We now test these predictions in a new experiment that includes a new type of trial that is used to elicit participants' belief distributions over possible estimates. We use a numerosity task in this experiment both to generalize the results, and because the discrete responses required in a numerosity experiment make it easier to elicit a belief distribution.

Experiment

Participants

Twenty-four participants (12 Males, ages between 18 and 35) were recruited through SONA system, University of Warwick. They received £4 for completing the experiment.

Materials

Participants were shown a briefly appearing number of dots (0.5 sec) on computer screen in a series of trials. The true number of dots was uniformly distributed between 21 and 30, and participants were explicitly told this at the beginning of the experiment. To generate a stimulus, dots were randomly positioned within a circular field subject to a minimum spacing between any two dots of four times the dot size. To encourage reliance on numerosity, rather than low-level visual features, the dot sizes varied uniformly between 3 and 9 pixels, and the radius of the dot field also varies uniformly between 150 and 450 pixels.

There were three different trial types (Figure 1B). For all trial types, participants made an initial decision as to whether there were 25 or fewer or 26 or more dots, so that the trials were identical until after this point. Following the decision, participants either immediately received feedback (D-F trials), were immediately asked to estimate the number of dots (D-E trials), or were immediately asked to state their beliefs about the number of dots using a histogram (D-H trials). When given a histogram, participants were asked to distribute 100 tokens among all of the possible numbers of dots according to how likely they believed these numbers were on that particular trial.

Procedure

Before the main experiment, participants received one practice example for each of the D-F, D-E, and D-H trials (Figure 1B). Participants additionally received feedback during practice, to introduce them to the point system used in the experiment that was used to encourage them to engage with the more demanding histogram trials. Correct decisions and estimates were both worth one point, while the number of points assigned to a histograms was $R = 100 \times [(1 \pm T_i/100)^2]$, where T_i was the number of tokens placed on the correct response. This formula is based on the Brier score, which incentivizes accurately reporting a belief distribution. Participants were also told, "If you had placed all the tokens on the correct number of dots, you would have scored 100 points. But if you had placed no tokens on the correct number of dots, you would have scored 0 points." Points were tallied throughout the experiment, but were only displayed to participants at the end of the experiment.

Results and Discussion

Decisions As shown in Figure 3A, participants were more likely than not to pick the correct answer for each true number of dots, but were never perfect.

Estimates Figure 3B shows the results of the estimates from the D-E trials for each true number of dots. True numbers that were close to the edge of the range showed an average bias towards the decision boundary, as participants tended not to respond outside the allowable range, and these responses outside the allowable range were excluded from further analysis (7.32%).

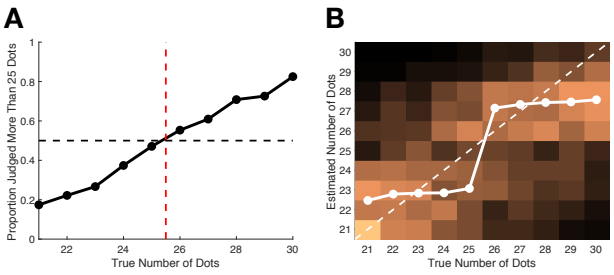


Figure 3: Behavioural data from the decision-estimation trials. **(A)** The psychometric curve describing the relationship between the true number of dots displayed and the average proportions of times participants judged the number of dots to be 26 or more. **(B)** The estimated number of dots (solid white line) systematically deviates from the true number of dots (dashed white line), constituting a perceptual bias.

For true numbers of dots near the decision boundary, particularly for numbers 25 and 26, participants were biased away from the decision boundary, in line with past work on the perceptual bias. When true number of dots was 25, estimates were on average smaller than 25, $t(23) = \pm 14.97, p < .001$. When true number of dots was 26, estimates were on average larger than 26, $t(23) = 6.37, p < .001$. This estimation bias is, as expected, further qualitative evidence against the SAS model.

Histograms Figure 4A shows the average belief histogram following a decision of 25 or fewer, and the average belief histogram following a decision of 26 or more. These average histograms show greater mass on the side of the boundary consistent with the decision, indicating that overall participants engaged with these trials. They also show no evidence of bimodality as WD predicts, nor is the mass completely on one side of the boundary as SC predicts. Instead qualitatively these average histograms are most consistent with BASS.

To quantitatively test for whether there were the kind of bimodal histograms that WD predicts, we computed the average proportions, for each participant, that the mean token mass on the boundary numbers (i.e., 25 and 26) would be lower than either the mean token mass on 21-24 or the mean token mass on 27-30 (Figure 4B). If responses were random, we expect a $1/3$ chance that the mean token mass on 25 and 26 would be smallest. However, there were very few histograms that were bimodal, fewer than would be predicted by random responding, $t(23) = \pm 7.68, p < .001$.

We then quantitatively tested whether all of the belief mass was on one side of the boundary, as SC predicts. We first estimated how much true responding there was, by calculating the proportion of trials on which a participant's histogram was consistent with their decision. Per participant, we calculated the proportion of trials on which the majority of token mass matched the decision. Next, we calculated,

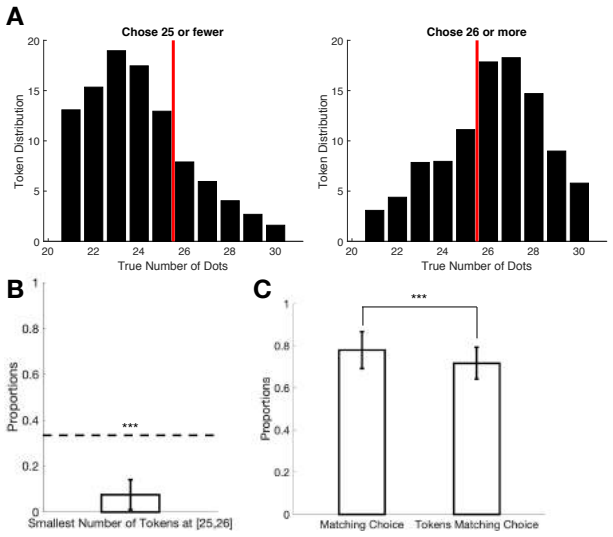


Figure 4: Results from the belief distributions reported in decision-histogram trials. **(A)** Average histograms when people chose "25 or fewer" (*left*) and when people chose "26 or more" (*right*). **(B)** The proportion of times when the token mass on central boundary numbers (25 and 26) is smallest. Error bar indicates 95% confidence interval across participants. **(C)** The proportion of times when preponderance of tokens matches the choice (*left*) and the percentages of token mass consistent with the choice (*right*). Error bars indicate 95% confidence interval across participants.

per participant, the proportion of tokens on the same side of boundary as the decision, as a measure of whether inconsistent estimates were considered. These two quantities, shown in Figure 4C, were different, $t(23) = 3.46, p = .002$, showing that tokens were certainly placed on numbers that disagreed with the decision, a number that exceeded what was expected from our estimate of noisy responding.

The results of the histogram trials are in line with the predictions of BASS, in which samples from both sides of the boundary are expected to be reused for the estimate. According to BASS, the amount of tokens placed on the opposite side of choice is a consequence of stochastic samples from the posterior distribution. Due to the termination conditions, there are always more samples that match the decision than those that mismatch the decision, but there are often samples on both sides of the boundary.

Conclusions

We proposed a new explanation for the decision bias in perception. To make a decision, we assume that participants sequentially sample hypotheses about the true nature of the environment, and stop when they have strong enough evidence in favor of one alternative over the other. As an application of amortised inference, we assumed that participants then reuse the samples to save cognitive effort, averaging them to produce their estimate. The bias in the estimate occurs because the samples that were sequentially obtained are never balanced between the options, and so estimates tend to be biased away from the decision boundary.

We generalized the bias from orientation tasks to a numerosity task, showing that it also occurs when participants give discrete responses to these perceptually complex stimuli. Using a novel type of trial in which we elicited participants' beliefs about which numbers were likely on a single trial, we found evidence for sequential sampling over other explanations of the decision bias in perception.

The sequential sampling model we evaluated here, BASS, is a novel application of the work of Wald (1950) for optimally deciding when to stop sampling from a binomial distribution with unknown probability. Of course it is almost certain that other sequential sampling approaches, such as those by Vul et al. (2014) and Bogacz et al. (2006), would predict the same qualitative results. Discriminating between these sequential sampling approaches will likely require quantitative comparisons, which is an interesting avenue for future work.

Additionally, a soft version of the self-consistency model, which relaxes the assumption that self-consistent estimates are made by every participant on every trial, could reproduce the qualitative results here. To distinguish the BASS model from a soft self-consistency account, we could test whether the belief distributions are a mixture of self-consistent sensory representations and unmodified sensory representations. However, to properly answer this question, we will need a further experiment that characterizes unmodified sensory representations for comparison.

Another avenue for future work is to explore the extent to which amortised inference and sequential sampling can explain other psychological biases. One interesting possibility is the anchoring bias (Tversky & Kahneman, 1974). In anchoring, participants are first asked to make a decision about whether a number, such as the percentage African countries in the UN, is smaller or larger than an often transparently irrelevant number, such as a number that results from the spin of wheel of fortune. Then participants are asked to make their estimate. While the task participants engage in is almost identical to the one that we used here, the effect that is found is the opposite: participants estimates are biased *toward* the anchor (Tversky & Kahneman, 1974). A unified explanation of these similar perceptual and cognitive biases will need to account for both the push and pull that decisions can exert on subsequent estimates.

Acknowledgements

J.Q.Z, A.N.S, and N.C. were supported by a grant from the National Institute of Economic and Social Research from their program Rebuilding Macroeconomics. A.N.S was supported by a European Research Council consolidator grant (817492-SAMPLING). N.C. was supported by the Economic and Social Research Council Network for Integrated

Behavioural Science (ES/P008976/1) and the Leverhulme Trust (RP2012-V-022).

References

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review*, *113*(4), 700.

Coenen, A., & Gureckis, T. M. (2016). The distorting effect of deciding to stop sampling. In *Proceedings of the annual meeting of the cognitive science society.*

Fleming, S. M., & Daw, N. D. (2017). Self-evaluation of decision-making: A general bayesian framework for metacognitive computation. *Psychological Review*, *124*(1), 91.

Gershman, S., & Goodman, N. (2014). Amortized inference in probabilistic reasoning. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 36).

Jazayeri, M., & Movshon, J. A. (2007). A new perceptual illusion reveals mechanisms of sensory decoding. *Nature*, *446*(7138), 912.

Luu, L., & Stocker, A. A. (2018). Post-decision biases reveal a self-consistency principle in perceptual inference. *eLife*, *7*, e33334.

Sanborn, A. N., & Chater, N. (2016). Bayesian brains without probabilities. *Trends in Cognitive Sciences*, *20*(12), 883–893.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131.

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? optimal decisions from very few samples. *Cognitive Science*, *38*(4), 599–637.

Wald, A. (1950). Statistical decision functions.

Zamboni, E., Ledgeway, T., McGraw, P. V., & Schluppeck, D. (2016). Do perceptual biases emerge early or late in visual processing? decision-biases in motion perception. *Proceedings of the Royal Society B*, *283*(1833), 20160263.

Zhu, J.-Q., Sanborn, A., & Chater, N. (2018). Mental sampling in multimodal representations. In *Advances in Neural Information Processing Systems* (pp. 5753–5764).

Modeling Judgment Errors in Naturalistic Numerical Estimation

Wanling Zou (wanlingz@sas.upenn.edu)
Department of Psychology

Sudeep Bhatia (bhatiasu@sas.upenn.edu)
Department of Psychology, Wharton Marketing
University of Pennsylvania
Philadelphia, PA 19104, USA

Abstract

We quantitatively modeled and compared two types of errors in numerical estimation for naturalistic judgment targets: *mapping errors* and *knowledge errors*. *Mapping errors* occur when people make mistakes reporting their beliefs about a particular numerical quantity (e.g. by inflating small numbers), whereas *knowledge errors* occur when people make mistakes using their knowledge about the judgment target to form their beliefs (e.g. by overweighting or underweighting cues). In two studies, involving estimates of the calories of common food items and estimates of infant mortality rates in various countries, we found that knowledge error models predicted participant estimates with very high out-of-sample accuracy rates, significantly outperforming the predictions of mapping error models. The knowledge error models were also able to identify the objects and concepts most associated with incorrect estimates, shedding light on the psychological underpinnings of numerical judgment.

Keywords: judgment errors; numerical estimation; word embeddings; word vectors; knowledge representation; cognitive model

Introduction

Decades of research on judgment and decision making has established that people make systematic errors when estimating numerical quantities, such as the frequencies of lethal events, proportions of demographic groups, or the calories of food items (Chernev & Chandon, 2011; Landy, Guay, & Marghetis, 2017; Lichtenstein et al., 1978). Researchers studying these judgment errors have identified a number of factors responsible for numerical mis-estimation, such as the use of non-linear weighting functions (e.g. Gonzalez & Wu, 1999; Hollands & Dyre, 2000; Landy et al., 2017; Tversky & Kahneman, 1992) or the use of heuristic cue-aggregation rules (Brown & Siegler, 1993; von Helversen & Rieskamp, 2008).

These factors can be understood in terms of two types of errors – *mapping errors* and *knowledge errors*. *Mapping errors* occur when people make mistakes reporting their beliefs about a particular numerical quantity. In other words, people may have the correct belief about the numerical quantity but incorrectly map this belief into a response. For example, prior literature on numerical estimation has found an inverse-S-shape pattern when plotting participant estimations against objective statistics, with overestimation of small values and underestimation of large values (e.g. Erlick, 1964; Hollands & Dyre, 2000; Landy et al., 2017; Varey, Mellers, & Birnbaum, 1990). Such patterns have typically been modeled using non-linear functions, e.g. power functions and their variants (Curtis, Attmeave, & Harrington, 1968; Hollands & Dyre, 2000), log-odds transformations (Shepard, 1981; Zhang & Maloney, 2012), and probability weighting functions (Fennell & Baddeley, 2012; Tversky & Kahneman, 1992). These models all assume that a systematic distortion takes place when transforming correct internal beliefs into an explicit numerical response.

In contrast, *knowledge errors* occur when people make mistakes using their knowledge about the judgment target to form their beliefs. These can lead to the formation of incorrect beliefs (e.g. through the biased use of memory cues), though people may still be able to accurately report these beliefs. For example, Chernev and Chandon (2011) have documented halo biases in food calorie estimation, according to which health-related cues are given an incorrectly high weight, which can then lead to the underestimation of food calories. Media coverage or word frequency has also been shown to be used as a cue in probability estimation (Dougherty, Franco-Watkins, & Thomas, 2008; Tversky & Kahneman, 1974) and frequency estimation (Hertwig, Pachur, & Kurzenhäuser, 2005; Lichtenstein et al., 1978), which can lead to the overestimation of the size of minority groups (Gallagher, 2003; Herda, 2013). More generally, many researchers in cognitive psychology have proposed that people use heuristics to weigh and aggregate judgment cues, which can, at times, lead to systematic errors in numerical estimation. These heuristics simplify the decision process by ignoring certain cues (and thus assigning them incorrectly low weights), or by using equal weights for all cues (and thus over-weighting irrelevant cues and under-weighting relevant cues) (see Hertwig, Hoffrage, & Martignon, 1999; Juslin, Olsson, & Olsson, 2003; von Helversen & Rieskamp, 2008). Our division of numerical judgment errors into mapping and knowledge errors has precedent. For example, Lichtenstein et al. (1978) suggested that there are two types of biases in frequency estimation – a primary bias (i.e. overestimation of small numbers and underestimation of large numbers) and a secondary bias (which may due to media bias, disproportionate exposure, imaginability, etc.). Likewise, Brown and Siegler (1993) argued that there are two types of knowledge that come into play in quantitative estimation – metric knowledge (e.g. statistical induction) and mapping knowledge (e.g.

heuristics). Von Helversen and Rieskamp (2008) extended this work by showing that people are likely to sample objects that are similar to the judgment target (where *knowledge errors* may occur) and make estimation based on some transformation of the sampled objects' values (where *mapping errors* may occur). Finally, Landy et al. (2017) showed the presence of two features that lead to errors in demographic estimation – domain-general bias (i.e. a log-odds mental representation of proportion) and domain-specific bias (e.g. media bias and xenophobia).

Although this work has greatly expanded our understanding of numerical estimation, most of it pertains to estimates of simple frequencies, rather than more general (and complex) numerical quantities. Additionally, experiments that do examine such complex numerical estimates, typically use artificial experimental stimuli – such as toxicity of fictional bugs (Juslin et al., 2003), percentage of dots in a mixture of black and white dots (Varey et al., 1990) and proportion of letters (e.g. "A") in a random letter string (Erlick, 1964) – and/or experimenter-generated cues that provide only an abstract representations of the rich knowledge present in the human mind (Brown, 2002; Juslin et al., 2003; von Helversen & Rieskamp, 2008). Although artificial stimuli and simplified knowledge representations help establish theoretical foundations, it is also necessary to model how people make quantitative estimates in the real world where they usually possess rich and complex knowledge and apply it at their discretion.

Our goal in this paper is to address these issues by formalizing, fitting, and comparing mapping and knowledge error models of numerical estimation with policy-relevant consequences. We consider two domains: estimates of food calories and estimates of infant mortality rates in countries. For the mapping error model, we drew insights from prior work and fit various non-linear functions to predict participant estimates from correct answers. For the knowledge error model, we used word embeddings – models trained on large text corpora that preserve semantic knowledge of words and phrases in high-dimensional vectors – to obtain rich quantitative representations for food items and countries, and then attempted to predict participant estimates from these representations by implicitly learning cue weights on semantic knowledge. Word embeddings have been shown to mimic representations at play in human cognition, such as associative judgment (Bhatia, 2017; Caliskan, Bryson, & Narayanan, 2017), free recall in memory (Manning & Kahana, 2012; Steyvers, Shiffrin, & Nelson, 2004), priming (Günther, Dudschig, & Kaup, 2016), and stereotypes (Garg, Schiebinger, Jurafsky, & Zou, 2018). Thus these representations are likely to capture common knowledge about the judgment targets that may hinder or facilitate numerical estimation. More importantly, these representations will offer insights into the psychological qualities and cues that most contribute to over- and underestimation.

Experimental Methods

Participants

We recruited a total of 101 participants – 50 participants (mean age = 30 years, 52% were female) in Study 1 and 51 participants (mean age = 31.4 years, 60.78% were female) in Study 2 from Prolific Academic, an online experiment platform. All participants were from the U.S. and had an approval rate of 80% or above. They were paid at a rate of approximately $6.50 per hour.

Stimuli

For Study 1, we obtained 200 food items and their true calorie amounts from a United States Department of Agriculture (USDA) database. Sample items include lamb, butter, mint, etc. For Study 2, we obtained the infant mortality rates of 200 countries from the Central Intelligence Agency (CIA)[1]. These countries include Denmark, Nepal, Estonia, etc.

Procedure

In Study 1, participants were asked to estimate how many calories (in kcal) there are in 100 grams of a particular food item; in Study 2, they were asked to estimate the infant (child under 1 year old) mortality rate in number of deaths per 1,000 live births in a particular country. Each participant estimated all 200 stimuli and saw only one item on each screen. The order of the 200 stimuli was randomized and there was a 30-second break after every 50 stimuli. After completing all questions, participants were asked for their age and gender.

Predicting Estimates

Computational Methods

For each target i (e.g. peanuts), we obtained both the average participant estimate y_i (e.g. estimated calories in peanuts) and the correct answer z_i (e.g. actual calories in peanuts). To quantitatively study mapping explanations for these errors, we fit three different mapping models that transformed correct answers into participant responses. Formally, our mapping error models predicted y_i as some function (linear or nonlinear) of z_i. The first function we used was a simple linear function (Eq.1); the second function was a third-degree polynomial (Eq.2); and the third function was a power function with a constant term (Eq.3)[2]. Parameters were estimated by minimizing the residual sum of squares.

$$y_i = \beta_0 + \beta_1 z_i \tag{1}$$

$$y_i = a z_i^3 + b z_i^2 + c z_i + d \tag{2}$$

[1] There are 224 countries and regions in CIA database. We excluded the ones that do not have a vector representation in our word embedding model (see the computational methods in the next section) and those whose public data are limited (e.g. no electricity usage data, no literacy rates, etc.)

[2] Although linear pattern was rarely found in previous literature, we included the linear model here as a baseline. A third-degree polynomial served to model any potential S-shape or inverse-S-shape pattern. We incorporated a power function due to its prevalence in prior work.

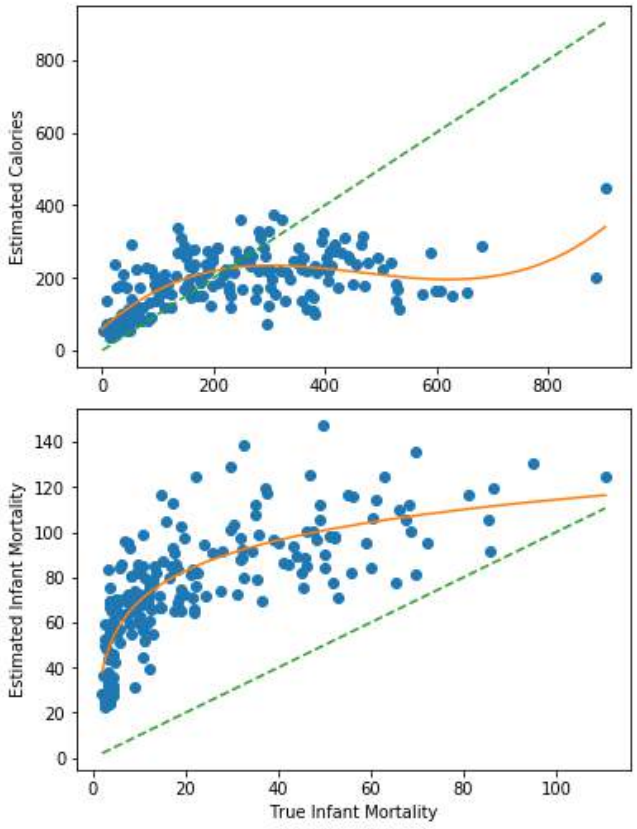


Figure 1: Scatterplots of average participant estimates vs. correct answers for Study 1 (top) and Study 2 (bottom). The green dotted lines indicate perfect calibration where participant estimates are equal to the correct answers. The red curves represent the best fitting mapping error models – third-degree polynomial (Eq.2) for Study 1 and power function (Eq.3) for Study 2.

$$y_i = \lambda z_i^{\delta} + \gamma \tag{3}$$

To examine knowledge errors, we used pretrained Word2Vec word embeddings (Mikolov et al., 2013) to obtain rich quantitative representations for food items and countries. These gave us a 300-dimensional vector x_i for each target i. Our knowledge error model involved fitting a (regularized) linear function with a 300-dimensional weight vector w (w_1 for Study 1 and w_2 for Study 2), to predict y_i using $w * x_i$[3]. The weight vectors (300-dimensional) transform semantic knowledge represented in a 300-dimensional space to an one-dimensional numerical estimation line. Intuitively, each dimension of x_i can be seen as a semantic cue that participants might rely on to facilitate estimation and these weight vectors can be seen as capturing cue weights on semantic knowledge. We compared mapping and knowledge error models through leave-one-out cross-validation (LOOCV)[4], on both the aggregate and the individual level.

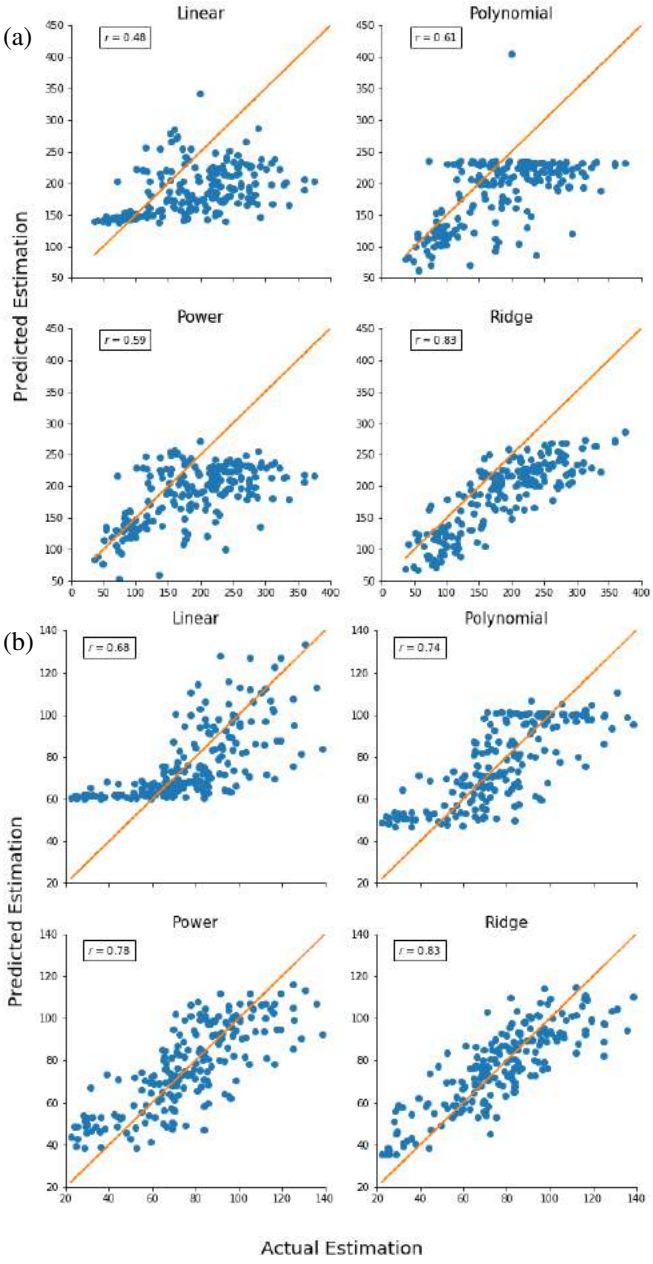


Figure 2: Scatterplots of predicted estimates using leave-one-out cross-validation (LOOCV) vs. actual participant estimates for (a) Study 1 and (b) Study 2, along with Pearson correlations.

Results

Study 1 and 2 elicited a total of 40,400 participant estimates for 400 distinct naturalistic entities in two domains calories

[3]Specifically, we implemented a ridge regression in the Scikit-Learn Python machine learning library (Pedregosa et al., 2011). There was a set of hyperparameters in this library. To avoid manipulating the hyperparameters to improve model performance, we took the default values of all these hyperparameters. We focused on ridge regression because previous results (e.g. Bhatia, 2019; Richie, Zou, & Bhatia, 2018, Dec) suggested that compared to other models such as lasso and support vector regression, ridge regression often works best in predicting human judgments from word embeddings.

[4]For each domain, we trained our models on 199 judgment targets and then used the trained model to predict participant estimates of the left-out target. This procedure was repeated for each judgment target to get LOOCV predictions

Table 1: Aggregate level predictive accuracy of the three mapping error models – linear (Eq.1), polynomial (Eq.2), power (Eq.3), and knowledge error model (ridge) using leave-one-out cross-validation (LOOCV) for Study 1 and Study 2.

	Study 1			Study 2		
	Correlation	R^2	RMSE	Correlation	R^2	RMSE
Linear	0.48	0.23	4609.04	0.68	0.46	348.72
Polynomial	0.61	0.37	3772.07	0.74	0.55	291.82
Power	0.59	0.35	3902.78	0.78	0.60	258.64
Ridge	0.83	0.68	1924.11	0.83	0.68	208.10

of 200 common food items and infant mortality rates of 200 countries. Consistent with prior work, we found that participants made substantial errors. The average absolute differences between the average participant estimate, y_i, and the correct answer, z_i, for food calories and infant mortality rates were -45.28kcal per 100g ($se = 10.8$) and 53.44 deaths per 1,000 live births ($se = 3.78$) respectively, indicating an overall underestimation of food calories and overestimation of infant mortality rates. Figure 1 reflects some overestimation of low calories, significant underestimation of high calories, and overall overestimation of infant mortality rates.

Table 1 summarizes the aggregate level performance of the three mapping error models and one knowledge error model. We evaluated model performance using the Pearson correlation between observed y_i and predicted y_i, R^2, and root mean square error (RMSE), in the out-of-sample tests. Figure 2 shows scatterplots of predicted estimates using LOOCV and average participant estimates, along with Pearson correlations. We found that the knowledge error model was able to predict average participant estimates fairly accurately, with out-of-sample correlation rates of .83 for both domains on the aggregate level. In contrast, the best mapping error models were only able to achieve aggregate-level out-of-sample correlation rates of .61 for foods and .78 for countries (all $p < 10^{-22}$). We obtained similar results on the individual level. The best mapping error model achieved average individual-level out-of-sample correlation rates of .37 for foods and .43 for countries, while the knowledge error model achieved .51 for food and 0.47 for countries. Our results showed statistically significant improvements in predictive accuracy when using the knowledge error model compared to the mapping error models on both the aggregate and the individual level.

Traces of Judgment Errors

Computational Methods

In the previous section, we showed that the word-embedding-based vector representations could be used to predict estimates of food calories and infant mortality rates by multiplying x_i (the vector representations for the foods and countries) with different weight vectors w_1 (Study 1) and w_2 (Study 2). As mentioned in computational methods of last section, these weight vectors can be seen as capturing cue weights on semantic knowledge. In this section, we hope to better understand the psychological substrates of the judgment errors that these weights generate. What are the features that lead to the overestimation or underestimation of food calories and infant mortality rates?

To address this, we took the 5,000 most frequent words from the corpus of contemporary American English (`http://corpus.byu.edu/coca/`) that were not judgment targets and for each word j, we also obtained a 300-dimensional vector, s_j, from the Word2Vec model. Intuitively, the weight vector w in the previous section could be seen as a function that projects the semantic knowledge represented by x_i onto a numerical estimation line y_i. By multiplying s_j by the weight vector w, we got a vector representation e_j for these 5,000 words in the numerical estimation line. Similarly, we also trained a weight vector w' to predict the correct answer, z_i, using $w' * x_i$. Multiplying s_j by this new weight vector w' would give us a vector t_j that pinpoints the location of the 5000 words in a line of correct answers. The difference between e_j and t_j then informs us of what words and concepts might be overweighted (or underweighted) in the estimation process. In other words, this difference would offer a quantitative measure of how much any given word contributes to overestimation (or underestimation).

Results

Figure 3 has word clouds of 50 words[5] that greatly contribute to over- and under- estimation for both domains. These words reveal potential conceptual underpinnings of judgment biases that match our intuition. For example, words related to dining out (e.g. restaurant, menu, chef, wine) bias toward overestimation of calories; words appearing to be small in portion (e.g. flour, candy, powder, dust) bias toward underestimation of calories; developing-country-related words (e.g. Iraqi, Cuban, Palestinian, Arab) contribute to overestimation of infant mortality rates; and European-country-related words (e.g. Dutch, German, French, European) contribute to underestimation of infant mortality rates.

[5]We included 50 words because that was the maximum number of legible words that could be fit into the graphs.

Figure 3: Word clouds of words that greatly contribute to overestimation and underestimation of (a) food calories and (b) infant mortality rates. Font reflects the magnitude of over- and under-estimation. Color has no meanings.

Discussion

We built computational models to compare two types of errors in numerical estimation: *mapping errors* and *knowledge errors*. We applied these models to study naturalistic numerical estimates in two studies involving judgments of calories of food items and judgments of infant mortality rates of countries. Consistent with previous findings, the best fitting mapping error models in both studies were not linear and drastically outperformed simple linear regression (baseline). We found the common inverse-S-shape pattern in food calorie estimation but not in the infant mortality domain. Although almost all countries' infant mortality rates were overestimated, the magnitude of overestimation appeared to be larger for countries with low infant mortality than for countries with high infant mortality.

Although our mapping error models were able to provide a good account of our data, we obtained even higher predictive accuracy rates from our knowledge error models. This indicates that we can predict participant estimates better if we assume some flexible (potentially incorrect) use of memory cues instead of some flexible (potentially incorrect) use of correct beliefs, and that judgment errors appear to stem primarily from the incorrect use of knowledge, rather than the incorrect mapping of the true quantities. Here it is also useful to note that our knowledge error models, unlike our mapping error models, did not know the correct responses. Rather they were able to predict participant estimates merely by proxying the rich semantic knowledge that participants used in their own judgments. This showcases both the descriptive power of the models as well as their domain-general applications: We can use these models to predict participant estimates even when we (the researchers) do not know the correct answers.

So far, we've only studied *mapping errors* and *knowledge errors* separately. It is likely that these two types of errors take place simultaneously as suggested by prior work (e.g. Brown & Siegler, 1993; von Helversen & Rieskamp, 2008). Future work should investigate the interaction between the two types of errors and also the interplay between error type and judgment domain – how do people balance between these two types of errors to minimize overall errors and in which domains are people more likely to make one type of errors than the other?

For our knowledge error model, we used semantic knowledge which was captured by word embeddings to predict participant estimates. Building upon recent successful applications of such models in naturalistic judgments (e.g. Bhatia, 2019; Richie et al., 2018, Dec), we showed how a single knowledge representation derived from natural language data was able to predict participant numerical estimation with high out-of-sample accuracy. We acknowledge that due to its high dimensionality, word-embedding-based representations are likely to contain more knowledge about judgment targets than what people actually use to estimate numerical quantities, and due to its generality, they also don't account for individual differences, such as personal experience and level of expertise. One way to address this may involve training different word embeddings for different populations, which is a potential topic for future work.

Although our mapping error models and knowledge error models lack some interpretability in terms of cognitive process underlying numerical estimation, the different non-linear mapping error models for food calories and infant mortality rates suggest different transformations from internal beliefs to external responses. Likewise, the words generated from the knowledge error models reveal intuitive conceptual bases of judgment errors, e.g. a misuse of food size associations in food judgment and a misuse of poverty associations in infant mortality judgment. These results show that our knowledge

error model has explanatory value, and can shed light on the types of associations that contribute to judgment errors across different domains.

Finally, we would like to emphasize the naturalism of the two domains examined in this paper. Our approach is unique in that it can be applied to numerical estimates for arbitrary natural entities, such as food items and countries. This opens up new avenues for applying cognitive science theory to policy-relevant applications, such as those pertaining to health-related and humanitarian issues. We look forward to future work that extends our approach to model the types of errors at play in the many important judgments that people make on a day-to-day basis.

References

Bhatia, S. (2017). Associative judgment and vector space semantics. *Psychological review*, *124*(1), 1–20.

Bhatia, S. (2019). Predicting risk perception: new insights from data science. *Management Science*.

Brown, N. R. (2002). Real-world estimation: Estimation modes and seeding effects. *Psychology of learning and motivation*, *41*, 321–359.

Brown, N. R., & Siegler, R. S. (1993). Metrics and mappings: A framework for understanding real-world quantitative estimation. *Psychological review*, *100*(3), 511–534.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186.

Chernev, A., & Chandon, P. (2011). Calorie estimation biases in consumer choice. In R. Batra, P. Keller, & V. Strecher (Eds.), *Leveraging consumer psychology for effective health communications: The obesity challenge* (pp. 104–121). New York, NY: M.E. Sharpe.

Curtis, D. W., Attmeave, F., & Harrington, T. L. (1968). A test of a two-stage model of magnitude judgment. *Perception & Psychophysics*, *3*(1), 25–31.

Dougherty, M. R., Franco-Watkins, A. M., & Thomas, R. (2008). Psychological plausibility of the theory of probabilistic mental models and the fast and frugal heuristics. *Psychological Review*, *115*(1), 199–213.

Erlick, D. E. (1964). Absolute judgments of discrete quantities randomly distributed over time. *Journal of Experimental Psychology*, *67*(5), 475–482.

Fennell, J., & Baddeley, R. (2012). Uncertainty plus prior equals rational bias: An intuitive bayesian probability weighting function. *Psychological Review*, *119*(4), 878–887.

Gallagher, C. A. (2003). Miscounting race: Explaining whites' misperceptions of racial group size. *Sociological Perspectives*, *46*(3), 381–396.

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, *115*(16), E3635–E3644.

Gonzalez, R., & Wu, G. (1999). On the shape of the probability weighting function. *Cognitive psychology*, *38*(1), 129–166.

Günther, F., Dudschig, C., & Kaup, B. (2016). Latent semantic analysis cosines as a cognitive similarity measure: Evidence from priming studies. *The Quarterly Journal of Experimental Psychology*, *69*(4), 626–653.

Herda, D. (2013). Too many immigrants? examining alternative forms of immigrant population innumeracy. *Sociological Perspectives*, *56*(2), 213–240.

Hertwig, R., Hoffrage, U., & Martignon, L. (1999). Quick estimation: Letting the environment do the work. In G. Gigerenzer, P. Todd, & the ABC Research Group (Eds.), *Simple heuristics that make us smart* (pp. 209–234). New York: Oxford University Press.

Hertwig, R., Pachur, T., & Kurzenhäuser, S. (2005). Judgments of risk frequencies: tests of possible cognitive mechanisms. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(4), 621–642.

Hollands, J., & Dyre, B. P. (2000). Bias in proportion judgments: the cyclical power model. *Psychological review*, *107*(3), 500–524.

Juslin, P., Olsson, H., & Olsson, A.-C. (2003). Exemplar effects in categorization and multiple-cue judgment. *Journal of Experimental Psychology: General*, *132*(1), 133–156.

Landy, D., Guay, B., & Marghetis, T. (2017). Bias and ignorance in demographic perception. *Psychonomic bulletin & review*, 1–13.

Lichtenstein, S., Slovic, P., Fischhoff, B., Layman, M., & Combs, B. (1978). Judged frequency of lethal events. *Journal of experimental psychology: Human learning and memory*, *4*(6), 551–578.

Manning, J. R., & Kahana, M. J. (2012). Interpreting semantic clustering effects in free recall. *Memory*, *20*(5), 511–517.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Vanderplas, J. (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research*, *12*(Oct), 2825–2830.

Richie, R., Zou, W., & Bhatia, S. (2018, Dec). Semantic representations extracted from large language corpora predict high-level human judgment in seven diverse behavioral domains. Retrieved from `https://psyarxiv.com/g9j83` doi: 10.31234/osf.io/g9j83

Shepard, R. N. (1981). Psychological relations and psychophysical scales: On the status of direct psychophysical measurement. *Journal of Mathematical Psychology*, *24*(1), 21–57.

Steyvers, M., Shiffrin, R. M., & Nelson, D. L. (2004). Word association spaces for predicting semantic similarity effects in episodic memory. *Experimental cognitive psychology*

and its applications: Festschrift in honor of Lyle Bourne, Walter Kintsch, and Thomas Landauer, 237–249.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *science*, *185*(4157), 1124–1131.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and uncertainty*, *5*(4), 297–323.

Varey, C. A., Mellers, B. A., & Birnbaum, M. H. (1990). Judgments of proportions. *Journal of Experimental Psychology: Human Perception and Performance*, *16*(3), 613–625.

von Helversen, B., & Rieskamp, J. (2008). The mapping model: A cognitive theory of quantitative estimation. *Journal of Experimental Psychology: General*, *137*(1), 73–96.

Zhang, H., & Maloney, L. T. (2012). Ubiquitous log odds: a common representation of probability and frequency distortion in perception, action, and cognition. *Frontiers in Neuroscience*, *6*, 1–14.

Semantic coordination of speech and gesture in young children

Olga Abramov
Bielefeld University, Bielefeld, Germany

Stefan Kopp
Bielefeld University, Bielefeld, Germany

Katharina Rohlfing
Paderborn University, Paderborn, Germany

Friederike Kern
Bielefeld University, Bielefeld, Germany

Ulrich Mertens
Paderborn University, Paderborn, Germany

Anne Nmeth
Bielefeld University, Bielefeld, Germany

Abstract

People use speech and gesture together when describing an event or action, where both modalities have different expressive opportunities (Kendon, 2004). One question is how the two modalities are semantically coordinated, i.e. how meaning is distributed across speech and accompanying gestures. While this has been studied only for adult speakers so far, here, we present a study on how young children (4 years of age) semantically coordinate speech and gesture, and how this relates to their cognitive and (indirectly) their verbal skills. Results indicate significant positive correlations between cognitive skills of the children and gesture-speech coordination. In addition, high cognitive skills correlate with the number of semantically relevant child descriptions revealing a link between verbal and cognitive skills.

Visuo-Motor Control Using Body Representation of a Robotic Arm with Gated Auto-Encoders

Julien Abrossimoff
ETIS, CNRS UMR8051, ENSEA, Universit de Cergy-Pontoise, Cergy-Pontoise, France

Alexandre Pitti
ETIS, CNRS UMR8051, ENSEA, Universit de Cergy-Pontoise, Cergy-Pontoise, France

Philippe Gaussier
ETIS, CNRS UMR8051, ENSEA, Universit de Cergy-Pontoise, Cergy-Pontoise, France

Abstract

We present an auto-encoder version of gated networks for learning visuomotor transformations for reaching targets and representating the location of the robot arm. Gated networks use multiplicative neurons to bind correlated images from each others and to learn their relative changes. Using the encoder network, motor neurons categorize the induced visual displacements of the robot arm when applying their corresponding motor commands.Using the decoder network, it is possible to infer back the visual motion and location of the robot arm from the activity of the motor units, aka body image.Using both networks at the same time, near targets can simulate a fictious visual displacement of the robot arm and induce the activation of the most probable motor command for tracking it. Results show the effectiveness of our approach for 2 degree of freedom and 3 degree of freedom robot arms. We discuss then about the network and its use for reaching task and body representation, future works and its relevance for modeling the so-called gain-field neurons in the parieto-motor cortices for learning visuomotor transformation.

Culture as ground for cross modality unidimensional timelines

Roberto Aguirre
Universidad de la Repblica, Montevideo, Uruguay

Alejandro Fojo
Universidad de la Repblica, Montevideo, Uruguay

Mauricio Castillo
Universidad de la Repblica, Montevideo, Uruguay

Mara Macedo
Universidad de la Repblica, Montevideo, Uruguay

Adriana de Len
Universidad de la Repblica, Montevideo, Uruguay

Maximiliano Meliande
Universidad de la Repblica, Montevideo, Uruguay

Germn Tourn
Universidad de la Repblica, Montevideo, Uruguay

Yliana Rodrguez
Universidad de la Repblica, Montevideo, Uruguay

Abstract

Current evidence supports the idea that time is mentally represented by unidimensional spaces. One main question is whether the language modality grounds differences on using these spaces when signers and speakers share the cultural framing of time (e.g., by clocks, calendars, etc.). We tested whether past and future events are represented along a Left-Past Right-Future and a Behind-Past Ahead-Future mental timeline in two language modalities. In Experiments 1 and 2 deaf signers of Uruguayan Sign Language (LSU) categorized the temporal reference of LSU sentences by pressing a directional key. The congruency effect was registered for the Left-Past Right-Future trials and for hand setting counterbalanced Behind-Past Ahead-Future trials. Experiments 3 and 4 replicated the congruency effect for Spanish speakers. The findings answered the research question in line with the suggestion that when signers and speakers share the cultural framing of time the tested space-time mappings activates on the same fashion.

Information Theory Meets Expected Utility: The Entropic Roots of Probability Weighting Functions

Mikaela Akrenius
Indiana University Bloomington, Bloomington, Indiana, United States

Abstract

This paper proposes that the shape and parameter fits of existing probability weighting functions can be explained with sensitivity to uncertainty (as measured by information entropy) and the utility carried by reductions in uncertainty. Building on applications of information theoretic principles to models of perceptual and inferential processes, I suggest that probabilities are evaluated relative to the distribution of maximum entropy (the uniform distribution) and that the perceived distance between a probability and uniformity is influenced by the shape (relative entropy) of the distribution that the probability is embedded in. These intuitions are formalized in a novel probability weighting function, VWD(p), which is simpler and has less free parameters than existing probability weighting functions. VWD(p) captures characteristic features of existing probability weighting functions, introduces novel predictions, and provides a parsimonious account of findings in probability and frequency estimation related tasks.

The Effect of Chronic Regulatory Focus on Sampling Behavior and Risky Decisions

Lujain Al Alamy
Columbia University, New York, New York, United States

James E. Corter
Columbia University, New York, New York, United States

Abstract

Prior research on a possible role of regulatory focus orientation (Higgins, 1998) in financial decision-making has focused on description-based tasks in which people receive explicit information about the characteristics of a decision problem a priori. However, relatively few real-world decisions resemble this type of laboratory task. Here, we examine how regulatory focus orientation influences peoples decision behavior in an experience-based sampling paradigm (Hertwig et al., 2004), where people learn about the characteristics of a decision problem only through experience. We investigated if individuals chronic regulatory focus orientation (promotion-focus or prevention-focus) predicts process (sampling) and outcomes (risky versus sure-thing choices) in a sampling paradigm task. Regulatory focus did not predict sampling behavior, nor the number of risky choices in the gain domain, but promotion focus orientation was correlated with the prevalence of risky choices in the loss domain. Also, the big-5 personality trait of Openness was found to be related to number of sampled outcomes for losses and to risky choices for gains.

Showing without telling: Indirect identification of psychosocial risks during and after pregnancy

Kristen Allen
Carnegie Mellon University, Pittsburgh, Pennsylvania, United States

Alex Davis
Carnegie Mellon University, Pittsburgh, Pennsylvania, United States

Tamar Krishnamurti
University of Pittsburgh, PITTSBURGH, Pennsylvania, United States

Abstract

During the perinatal period, psychosocial health risks, including depression and intimate partner violence, are associated with serious adverse health outcomes for both parent and child. To appropriately intervene, healthcare professionals must first identify those at risk, yet stigma often prevents people from disclosing the information needed to prompt an assessment. We use techniques from natural language processing to indirectly identify psychosocial risks in the perinatal period. We apply latent Dirichlet allocation (LDA) and latent semantic indexing (LSI) to categorize themes from brief diary entries by pregnant and postpartum women and apply sentiment analysis to characterize affect, then perform regularized regression to predict diagnostic measures of depression and emotional intimate partner violence. Journal text entries quantified through sentiment analysis and topic models show promise for improved identification of depression and intimate partner violence, both stigmatized risks. Such methods may serve as an initial or complementary screening approach.

Modeling Gaze Distribution in Cross-situational Learning

Andrei Amatuni
Indiana University, Bloomington, Indiana, United States

Chen Yu
Indiana University, Bloomington, Bloomington, Indiana, United States

Abstract

Here we investigate the performance of two models in predicting human gaze behavior in cross situational word learning. Previous work has developed two diverging accounts of potential mechanisms that might serve this learning ability. The first, associative learning, relies on the integration of contextual statistics across time. The second, hypothesis testing of the "propose-but-verify" sort, suggests that learners do not track co-occurrence statistics, instead only tracking a single label-object mapping at a time. To adjudicate between these two mechanisms, we examine real time selective attention behavior as a window into learning processes. We demonstrate systematic biasing in gaze allocation as a function of the associative evidence accumulated for a label-object pairing over time, favoring the associative learning account. Moreover, we predict learning outcomes with model parameters controlling sensitivity and noise in memory encoding. This is novel evidence supporting associative learning and highlights the unique role of memory in cross-situational learning.

Learning by doing: Supporting experimentation in inquiry-based modeling

Sungeun An
Georgia Institute of Technology, Atlanta, Georgia, United States

Robert Bates
Georgia Institute of Technology, Atlanta, Georgia, United States

Jennifer Hammock
Smithsonian Institution, Washington, District of Columbia, United States

Spencer Rugaber
Georgia Institute of Technology, Atlanta, Georgia, United States

Emily Weigel
Georgia Institute of Technology, Atlanta, Georgia, United States

Ashok Goel
Georgia Institute of Technology, Atlanta, Georgia, United States

Abstract

Inquiry-based modeling plays an important role in science; Science makes progress through formulating and evaluating questions, hypothesis, and arguments. The inquiry-based modeling approach also enhances learning about science. However, engaging in modeling requires domain knowledge as well as quantitative skills. The Virtual Ecological Research Assistant (VERA) is an interactive learning environment that supports inquiry-based modeling for citizen and student scientists. VERA provides a visual language for conceptual modeling in the domain of ecology and an AI compiler for automatic generation of agent-based simulations in the process of constructing, evaluating, and revising the models. We conducted a pilot study with college-level biology students (N=15) using VERA for modeling ecological phenomena. The 2-hour pre- and post-test study demonstrates gains in the learning of ecological content knowledge. We also found that the use of the Encyclopedia of Life as a source of domain knowledge helped students create more complex models.

Composing Indeterminate Event Information In Context: Evidence from an Eye-Tracking Memory Paradigm

Caitlyn Antal
Yale University, New Haven, Connecticut, United States

Roberto de Almeida
Concordia University, Montreal, Quebec, Canada

Abstract

A sentence such as ”We finished the paper” is indeterminate regarding what we finished doing with the paper. These sentences constitute a test case for two major issues regarding the nature of language comprehension: (1) whether or not semantic composition is simple (classical) or enriched with intended or implicit constituents; and (2) the nature of the linguistic and cognitive resources that help us interpret the event the sentence conveys. We conducted an eye-tracking study to investigate whether indeterminate sentences embedded within biasing contexts would trigger event interpretations, using a long-term memory paradigm. In each trial, participants were presented with one of three recognition probe types for reading while having their eyes monitored. Recognition probes were presented 0 seconds (s) after having read the indeterminate sentence, or following an additional 25s of neutral discourse. Results suggest that abductive processes, relying on the propositional content of supporting context, drive indeterminate sentence interpretation.

Linguistic Distributional Information and Sensorimotor Similarity Both Contribute to Semantic Category Production

Briony Banks
Lancaster University, Lancaster, United Kingdom

Cai Wingfield
University of Lancaster, Lancaster, United Kingdom

Louise Connell
University of Lancaster, Lancaster, United Kingdom

Abstract

We investigated the contribution of sensorimotor and linguistic distributional information in a semantic category production task, hypothesizing that the task would rely on both but particularly on linguistic distributional information, which may provide a shortcut for conceptual processing. In a pre-registered study, we asked participants to name members of semantic categories and tested whether responses were predicted by a novel measure of sensorimotor proximity (based on an 11-dimension representation of sensorimotor experience) and linguistic proximity (based on word co-occurrence derived from a large subtitle corpus). Both proximity measures predicted the order and frequency of responses and, critically, linguistic proximity had an effect above and beyond sensorimotor proximity. Our findings support linguistic-sensorimotor accounts of the conceptual system and suggest that category production is based on both the similarity of sensorimotor experience between the category and member concepts, and on the linguistic distributional relationship between the category and member labels.

Listeners use descriptive contrast to disambiguate novel referents

Claire Bergey
The University of Chicago, Chicago, Illinois, United States

Dan Yurovsky
University of Chicago, Chicago, Illinois, United States

Abstract

People often face referential ambiguity; one cue to resolve it is adjectival description. Beyond narrowing potential referents to those that match a descriptor, listeners may infer that a described object is one that contrasts with other present objects of the same type (tall cup contrasts with another, shorter cup). This contrastive inference guides the visual identification of a familiar referent as an utterance progresses (Sedivy et al., 1999). We extend this work, asking whether listeners use this type of inference to guide explicit referent choice when reference is ambiguous, and whether this varies with adjective type. We find that participants consistently use size adjectives contrastively, but not color adjectives (Experiment 1)even when color is described with more relative language (Experiment 2) or emphasized with prosodic stress (Experiment 3). Listeners can use adjective contrast to disambiguate a novel words referent, but do not treat all adjective types as equally contrastive.

Emulating Human Developmental Stages with Bayesian Neural Networks

Marcel Binz
Philipps-Universitt Marburg, Marburg, Germany

Dominik Endres
Philipps-Universitaet, Marburg, Hesse, Germany

Abstract

In this work we compare the acquisition of knowledge in humans and machines. Research from the area of developmental psychology indicates, that human-employed hypothesis are initially guided by simple rules, before evolving into more complex theories. This observation is shared across many tasks and domains. We investigate whether the stages of development in artificial learning systems are based on similar characteristics. We operationalize developmental stages as the size of the data-set on which the artificial system is trained. For our analysis we look at the developmental progress of Bayesian Neural Networks on three different data-sets, including occlusion, support and quantity comparison tasks. We compare the results with prior research from the developmental psychology literature and find agreement between the family of optimized models and pattern of development observed in infants and children on all three tasks, indicating common principles for the acquisition of knowledge.

An asymmetry between distance estimates made to and from a target

David Bosch
New York University, New York, New York, United States

Yaacov Trope
New York University, New York, New York, United States

Abstract

In three experiments, we demonstrated that the self can act as a cognitive reference point, producing an egocentric asymmetry effect on distance judgments such that targets are judged as closer to the viewer than the viewer is to the target. Egocentric asymmetry was observed even when there was a fixed reference object that people could use to anchor distance estimates across trials (Experiment 2). Further, egocentric asymmetry was greater to a non-human artifact than to a human avatar (Experiment 3). In addition, distances from a mailbox to a human avatar were estimated as shorter than distances from an avatar to a mailbox, suggesting that the special status of the self may extend to other people when compared to non-human objects even in allocentric distance judgments (Experiment 2).

Neither the time nor the place: Omissive causes yield temporal inferences

Gordon Briggs
U.S. Naval Research Laboratory, Washington, District of Columbia, United States

Hillary Harner
US Naval Research Laboratory, Washington, District of Columbia, United States

Christina Wasylyshyn
U.S. Naval Research Laboratory, Washington, District of Columbia, United States

Paul Bello
U.S. Naval Research Laboratory, Washington, District of Columbia, United States

Sangeet Khemlani
Naval Research Laboratory, WASHINGTON, District of Columbia, United States

Abstract

Is it reasonable to draw temporal conclusions from omissive causal assertions? For example, if you learn that not charging your phone caused it to die, is it sensible to infer that your failure to charge your phone occurred before it died? The conclusion seems intuitive, but no theory of causal reasoning explains how reasoners make the inference other than a recent proposal by Khemlani and colleagues (2018a). We present that theory and describe its consequences. If people conceive of omissions as non-events, i.e., events unmoored in space and time, they might refrain from drawing conclusions when asked whether an omissive cause precedes its effect. Two experiments speak against these predictions of the non-event view and in favor of a view that omissive causation imposes temporal constraints on events and their effects. We conclude by considering whether drawing a temporal conclusion from an omissive cause constitutes a reasoning error.

Modeling Long-Distance Cue Integration Strategies in Phonetic Categorization

Wednesday Bushong
University of Rochester, Rochester, New York, United States

T. Florian Jaeger
University of Rochester, Rochester, New York, United States

Abstract

Language temporally unfolds, with relevant cues arriving at different moments in time. For comprehension to be optimal, listeners must maintain gradient representations of cues in order to integrate with later-arriving cues. Several studies have established during speech perception listeners integrate cues that occur far apart in time. There are several proposals about how restricted this is, but there's little rigorous work establishing and testing models of long-distance cue integration strategies. We take a first step at addressing this gap by formalizing four different models of how listeners use cue information during real-time processing, testing them on two perception experiments. In one experiment, we find support for optimal integration of cues. In another, more attention-taxing experiment, we find evidence in favor of a strategy that avoids maintaining detailed representations of cues in memory. These results represent a first step toward understanding how listeners change their cue integration strategies across contexts.

Simplicity preferences in young childrens decision-making

Rebecca Canale
University of Rochester, Rochester, New York, United States

George Loewenstein
Carnegie Mellon University, Pittsburgh, Pennsylvania, United States

Celeste Kidd
University of California, Berkeley, Berkeley, California, United States

Abstract

Classic theories of multi-attribute choice typically assume that preferences are an additive function of attribute values. However recent work (Evers et al.) demonstrates a preference for simplicity that can violate the most basic assumptions and predictions of conventional models. For example, a set of 7 colored pencils that are all unique colors are preferred over a set of 8 colored pencils with one redundant color. This preferential choice, however, cannot be explained by the utility of consumption itself. Does this preference emerge as a result of adults substantial experience with such sets in the world (e.g., through shopping or organizing ones possessions), or is this preference present much earlier? Does the preference for simplicity, in fact, facilitate cognitive encoding? We investigate these questions through a series of experiments conducted with children in an effort to understand the emergence of this simplicity bias, and its connection to the development of working memory.

Exploring the Role of Social Priming in Alcohol Attentional Bias

Stephen Cantarutti
City, University of London, London, United Kingdom

Emmanuel Pothos
City, University of London, London, United Kingdom

Abstract

Recent studies have linked the Stroop Effect with social priming, suggesting that social concept priming tends to trigger automatic behaviour aligned with the primed concept (Augustinova & Ferrand, 2014; Goldfarb, Aisenberg, & Henik, 2011). This study attempts to test the efficacy of social priming on alcohol attentional bias, integrating a social priming interference task into an alcohol-Stroop test to measure Stroop interference before and after participants have been socially primed. Results show no significant interaction between stimulus category (alcohol and neutral), experiment block, and social priming condition (alcohol addiction, alcohol preoccupation and control) to indicate that social priming had triggered expedited, automatic behaviour. Our results do show a significant interaction between experiment block and social priming condition ($F(6, 426) = 2.166$, $p = .045$), suggesting the alcohol social priming tasks may have induced a greater general interference for participants in those conditions, than for participants receiving the neutral interference task.

Visual Spatial Attention Skills and Holistic Processing in High School Students With and Without Dyslexia

Ronald Chan
The Education University of Hong Kong, Hong Kong, Hong Kong

Chin-wai Kwok
The Education University of Hong Kong, Hong Kong, Hong Kong

Duo Liu
The Education University of Hong Kong, Hong Kong, Hong Kong

Ricky Van-yip Tso
The Education University of Hong Kong, Taipo, N.T., Hong Kong

Abstract

Visual-spatial attention has been shown to influence literacy development, yet studies investigating its influence on reading in non-alphabetic scripts such as Chinese are scarce, despite recent studies demonstrating orthographic and visuo-spatial skills to be key deficits in people with dyslexia in Chinese. Here, we investigate visual-spatial processing skills in Chinese adolescents by measuring their 1) exogenous and endogenous attentional orienting, and 2) holistic processinga phenomenon typically demonstrated in face perceptionin Chinese character recognition. Compared with typically developing students, Chinese high-school students with dyslexia showed deficits in both endogenous and exogenous visual-spatial attention. Dyslexics also perceived characters more holistically than the controls, suggesting that they selectively attended to individual components within Chinese characters less readily. These results demonstrated irregularities in visual-spatial processing skills in students in dyslexia. This study provides implications for reading intervention programs in order to facilitate selective attention to character components to enhance literacy.

Elucidating the Cognitive Anatomy of Representation Systems

Peter Cheng
University of Sussex, Brighton, United Kingdom

Grecia Garcia Garcia
University of Sussex, Brighton, United Kingdom

Holly Sutherland
University of Sussex, Falmer, United Kingdom

Daniel Raggi
University of Cambridge, Cambridge, United Kingdom

Aaron Stockdill
University of Cambridge, Cambridge, United Kingdom

Mateja Jamnik
University of Cambridge, Cambridge, United Kingdom

Abstract

We present a framework to assess the relative cognitive cost of alternative representational systems for problem solving. The framework consists of 19 cognitive properties of representational systems, which are distributed across 4 dimensions (registration, semantic encoding, inference, and solution) and three scales of granularity (symbol, expression, and sub-representations). It examines components and processes spanning the internal mental representation and external physical display, and also addresses heterogenous representations of problems. We provide functions to evaluate the cost of each cognitive property by examining, for example, types of matches between display symbols and concepts, the arity of relations, or the depth of solution trees. The cognitive costs for each property are combined to estimate the overall cognitive cost, and hence the relative effectiveness, of a representation. The frameworks development is motivated by our goal of engineering an automated system that will select representations suited to specific classes of problems for individual users.

Why Some Events Are More (or Less) Random: The Role of Alternation Rate and Number of Occurrence

Karen H. H. Chu
University of Macau, Macau SAR, China

Sophia Deng
University of Macau, Macau SAR, Macao

Abstract

How do people tell the difference between random and nonrandom events? What affects peoples understanding of randomness? In two experiments, we investigated the role of two characteristics of a sequencealternation rate and number of occurrencein peoples perception of randomness. We presented participants with a pair of binary sequences of length 6 (e.g., OXOXXO vs. XOXXXX) and asked them to evaluate which of the two was more likely to occur. In Experiment 1, we examined how participants randomness perception changed as the difference in alternation rate and the difference in the number of occurrence changed. In Experiment 2, we further examined whether participants exhibited differential reliance on alternation rate and number of outcomes. Results suggest that people exhibit differential reliance on alternation rate and number of occurrence. When the two characteristics are in conflict, people tend to rely more on the alternation rate in their randomness judgement.

Integrating Methods to Improve Model-based Performance Prediction

Michael Collins
Air Force Research Laboratory, Dayton, Ohio, United States

Kevin Gluck
Air Force Research Laboratory, Wright-Patterson AFB, Ohio, United States

Abstract

The initial performance of individuals is often difficult for models of learning and retention to predict. One such model is the predictive performance equation (PPE) a mathematical model of learning and retention that uses regularities seen in human learning to predict future performance.To generate predictions, PPEs free parameters must be calibrated to a minimum amount of historical performance data, preventing valid predictions for initial learning events.Prior research (Collins, Gluck, Walsh, Krusmark & Gunzelmann, 2016; Collins, Gluck, & Walsh, 2017), has shown that the generalization of best fitting parameters from prior data can improve initial performance predictions.Here we build on that research, using Bayesian hierarchical modeling to estimate free parameters from various sources of prior data. Bayesian hierarchal modeling allows an opportunity to improve and add structure to the parameters used by PPE, improving its application to cognitive technology in education and training.

Compositional subgoal representations

Carlos Correa
Princeton University, Princeton, New Jersey, United States

Frederick Callaway
Princeton University, Princeton, New Jersey, United States

Mark Ho
UC Berkeley, Berkeley, California, United States

Tom Griffiths
University of California, Berkeley, Berkeley, California, United States

Abstract

When faced with a complex problem, people naturally break it up into several simpler problems. This hierarchical decomposition of an ultimate goal into sub-goals facilitates planning by reducing the number of factors that must be considered at one time. However, it can also lead to suboptimal decision-making, obscuring opportunities to make progress towards multiple subgoals with a single action. Is it possible to take advantage of the hierarchical structure of problems without sacrificing opportunities to kill two birds with one stone? We propose that people are able to do this by representing and pursuing multiple subgoals at once. We present a formal model of planning with compositional goals, and show that it explains human behavior better than the standard "one-at-a-time" subgoal model as well as non-hierarchical limited-depth search models. Our results suggest that people are capable of representing and pursuing multiple subgoals at once; however, there are limitations on how many subgoals one can pursue concurrently. We find that these limitations vary by individual.

Rule-following, Lexical Competence and Categorization Processes

Marco Cruciani
University of Trento, Trento, Italy

Francesco Gagliardi
ORCID:0000-0002-4270-1636, Naples, Italy

Abstract

The article addresses the issues of extending a category and updating a lexical concept, and determining its reference. We try to answer the questions: how can an object seen for the first time extend a category or update a concept? How is it possible to determine the reference of a concept that represents a behaviour? Firstly, we discuss the learning of inferential linguistic competence used to update a concept through an approach based on prototype theory. Secondly, we discuss the learning of referential linguistic competence used to determine the reference of a concept through an approach based on embodied cognition. Finally, on the basis of the dual dimension of the praxis of rule-following, we show how it is possible to combine the two approaches into a single model that deals with both the extension of a category and the updating of a concept, and the determination of the reference.

Magnitude Comparisons of Improper Fractions

Lucy Cui
UCLA, Los Angeles, California, United States

Zili Liu
UCLA, Los Angeles, California, United States

Abstract

Previous studies examining the mental representations of fractions have focused on fractions with magnitudes less than one (e.g., 2/3). In the current study, we examine the mental representations of fractions with magnitudes greater than one, specifically those of improper fractions. Participants were asked to make magnitude comparisons of these improper fractions to a reference that was in an improper fraction, a mixed fraction, or a decimal format. Results show that magnitudes of improper fractions were more accurately accessed when they were compared to mixed fractions and decimals. This suggests that the reinterpretation of these improper fractions benefited magnitude processing. Distance effects on error rate and response time were observed for all three reference formats and more consistently took the form of a Welford function, which predicts worse performance above rather than below the reference. Possible explanations of these results are discussed.

Magnitude Comparisons of Discounted Prices: Are They Similar to Fractions?

Lucy Cui
UCLA, Los Angeles, California, United States

Zili Liu
UCLA, Los Angeles, California, United States

Abstract

The present study examines whether peoples mental representation of discounted prices, which have a part-whole relationship of the current price to the original price, is similar to that of fractions. Participants performed a fraction comparison task and a deal comparison task on the same set of fractional magnitudes. In two experiments, we observed worse performance (error rate, RT of correct trials) on the deal comparison task. The distance effect, where magnitude comparisons are made more slowly and less accurately the closer two magnitudes are, observed in the two tasks was best modeled using logarithmic distance between the fractional magnitudes as a predictor of performance.

Magnitude Processing of Improper Fractions When Comparing Bundle Deals

Lucy Cui
UCLA, Los Angeles, California, United States

Zili Liu
UCLA, Los Angeles, California, United States

Abstract

People encounter improper fractions in real life contexts on a regular basis. One such example is with bundling at the grocery store (2/$4 or two for $4). The present study seeks to understand how people process these bundle prices compared to improper fractions. Participants completed a magnitude comparison task with different bundling formats (2/$4 vs. $4/2) and their fractional equivalents. We found a reliable difference between the bundle format (2/$4) seen in grocery stores and the most visually similar fraction (2/4). This difference shows that participants are not using a heuristic (larger fraction means cheaper per item) when comparing these bundle deals and instead do need to process them like improper fractions. Overall, we found that participants were better at comparing fractional magnitudes in a math context than in a financial context and that this effect of context also depended on format (2/4 vs. 4/2).

Category-Specific Verb-Semantic Naming Deficit in Alzheimers Disease: Evidence from a Dynamic Action Naming Task

Roberto de Almeida
Concordia University, Montreal, Quebec, Canada

Forouzan Mobayyen
Concordia University, Montreal, Quebec, Canada

Eva Kehayia
McGill University, Montreal, Quebec, Canada

Caitlyn Antal
Yale University, New Haven, Connecticut, United States

Vasavan Nair
Douglas Mental Health University Institute, Montreal, Quebec, Canada

George Schwartz
Douglas Mental Health University Institute, Montreal, Quebec, Canada

Abstract

Numerous studies have found category-specific semantic deficits in Alzheimers disease (AD). Thus far, however, only a small number of studies have investigated how semantic categories lexicalized by verbs are represented, and how these categories might be impaired in AD. We investigated the representation and breakdown of verb knowledge employing different syntactic and semantic classes of verbs in a group of probable AD patients (N=10) and matched controls. In our main task, we employed movies of events and states depicting verbs belonging to three different classes: causatives, perception/psychological, and movement verbs. These verbs differ with regards to their argument structure, the thematic roles they assign, and their hypothetical semantic templates. Patients had more difficult employing verbs of the perception/psychological class. We suggest that thematic roles play the most important role in verb semantic representations. We further suggest that verbs are not represented by decompositional semantic templates.

A Reservoir Model for Intra-Sentential Code Switching Comprehension in French and English

Pauline Detraz
Inria, Bordeaux, France

Xavier Hinaut
Inria, Bordeaux, France

Abstract

Some people can mix two languages within the same sentence: this is known as intra-sentential code-switching. The majority of computational models on language comprehension are dedicated to one language. Some bilingual models have also been developed, but very few have explored the code-switching case. We collected data from human subjects that were required to mix pairs of given sentences in French and English. Truly bilingual subjects produced more switches within the same sentence. The corpus obtained have some very complex mixed sentences: there can be until eleven language switches within the same sentence. Then, we trained ResPars, a Reservoir-based sentence Parsing model, with the collected corpus. This Recurrent Neural Network model processes sentences incrementally, word by word, and outputs the sentence meaning (i.e. thematic roles). Surprisingly the model is able to learn and generalize on the mixed corpus with performances nearly as good as the unmixed French-English corpus.

Assessment of Cognitive Load in the Context of Neurosurgery

Daniel Di Giovanni
McGill University, Montreal, Quebec, Canada

Simon Drouin
Montreal Neurological Institute (McGill University), Montreal, Quebec, Canada

Marta Kersten-Oertel
Concordia University, Montreal, Quebec, Canada

Louis Collins
McGill University, Montreal, Quebec, Canada

Abstract

The work presented in this paper explores the amount of effort, defined by cognitive load, needed to understand depth visualization while navigating a virtual space in the context of planning for image guided surgery. In this context, cognitive load is evaluated by measuring brain activity through event-related electroencephalography (EEG). We found a significant difference between dynamic depth cue renders versus statically rendered cues. The work presented here demonstrates the usefulness of EEG as an acceptable and efficient method to inspect brain activity for future user studies in the operating room, and that cognitive load can serve as an objective measure of visualization effectiveness.

Skill Acquisition in a Dynamic Collaborative Task

Cvetomir Dimov
Carnegie Mellon University, Pittsburgh, Pennsylvania, United States

John R. Anderson
Carnegie Mellon University, Pittsburgh, Pennsylvania, United States

Shawn Betts
Carnegie Mellon University, Pittsburgh, Pennsylvania, United States

Dan Bothell
Carnegie Mellon University, Pittsburgh, Pennsylvania, United States

Abstract

Skill acquisition studies have generally focused on individual tasks, such as language learning, learning how to use a text editor or how to play video games. Here we present a study that investigates how subjects learn to work in a team in a dynamic collaborative task. The task - Coop Space Fortress - is a modification of a computer game used extensively in research, in which subjects fly space ships in a frictionless environment and coordinate to destroy a space fortress. When learning to play this computer game, subjects not only master the game controls, but also typically settle on team roles to more efficiently achieve their goal, despite not being allowed to communicate.

Liars Intent: A Multidimensional Recurrence Quantification Analysis Approach to Deception Detection

Hannah Douglas
Macquarie University, Macquarie Park, NSW, Australia

Adriana Rossi
Macquarie University , Sydney, NSW, Australia

Rachel W. Kallen
Macquarie University, Macquarie Park, NSW, Australia

Michael J Richardson
Macquarie University, Sydney, NSW, Australia

Abstract

The current study utilizes dynamical systems and embodiment theory to better understand how movement dynamics impact deception detection. While research has consistently revealed humans are often no better than chance at discriminating a truth from a lie, individuals may reveal more than they know through the dynamic movement of the face and the body beyond discrete gestures traditionally examined in deception detection research (e.g., rise of a brow). As expected, the present study found that the dynamic stabilities of facial and body movements were significantly influenced by deceptive intent such that untruthful statements elicited less stability in both the face and upper body. Moreover, despite detection levels no greater than chance, the accuracy of observers to detect deceptive intent covaried with these dynamic stabilities. The research presented provides another piece to the illusive puzzle of deception detection, affording researchers and practitioners a possible tool to tap into deceptive intent.

Human-level but not human-like: Deep Reinforcement Learning in the dark

Rachit Dubey
Princeton University, PRINCETON, New Jersey, United States

Pulkit Agrawal
UC Berkeley, Berkeley, California, United States

Deepak Pathak
UC Berkeley, Berkeley, California, United States

Alyosha Efros
UC-Berkeley, Berkeley, California, United States

Tom Griffiths
University of California, Berkeley, Berkeley, California, United States

Abstract

Deep reinforcement learning (RL) algorithms have recently achieved impressive results on a range of video games, learning to play them at or beyond a human level just from raw pixel inputs. However, do they leverage visual information in the same manner as humans do? Our investigations suggest that they do not: given a static game, we find that a state-of-the-art deep RL algorithm solves that game faster without visual input (only the agent location was provided to the algorithm). We posit that this is because deep RL attacks each problem tabula rasa, i.e. without any prior knowledge, as also suggested by other recent work. We further propose that in certain settings, an agent is better off having no visual input compared to having no visual priors. To demonstrate this, we conduct an experiment with human participants and find that people solve a game that hid all visual input (except agent location) much faster than a game that prevented the use of various visual priors. These results highlight the importance of prior knowledge and provide a compelling demonstration of how the lack of prior knowledge leads to deep RL algorithms approaching a problem very differently from humans.

Exergame Training of Executive Function in Preschool Children: Generalizability and Long-term Effects

Cassondra Eng
Carnegie Mellon University, Pittsburgh, Pennsylvania, United States

Melissa Pocsai
Carnegie Mellon University, Pittsburgh, Pennsylvania, United States

Dominic Calkosz
Carnegie Mellon University, Pittsburgh, Pennsylvania, United States

Nathan Williams
Carnegie Mellon University, Pittsburgh, Pennsylvania, United States

Erik Thiessen
Carnegie Mellon University, Pittsburgh, Pennsylvania, United States

Anna Fisher
Carnegie Mellon University, Pittsburgh, Pennsylvania, United States

Abstract

Studies with older children and adults have found that physically engaging video games (i.e., Exergames) that promote both cognitive control and physical activity improve executive function (EF) skills; yet, children below school age remain understudied with regard to the impact of Exergames on EF. Additionally, research on the extent of the impact of Exergames resulting in prolonged changes, and whether training generalizes to EF-related behaviors in a real-world context remains scarce. This study examined the short- and long-term changes in EF of 4- to 5-year-olds after participation in two 20-minute Exergame sessions. Results indicate that Exergame training improved performance on EF tasks and resulted in higher teacher ratings of EF in the classroom compared to a sex-/classroom-/age-matched control group. The improvements in EF persisted over a one-month period. This study provides novel insights into the short-term and long-term effects of Exergame training on executive function in preschool-aged children.

Using Known Words to Learn More Words: A Distributional Analysis of Child Vocabulary Development

Andrew Flores
University of Illinois Urbana-Champaign, Champaign, Illinois, United States

Jessica Montag
University of Illinois, Champaign-Urbana, Champaign, Illinois, United States

Jon Willits
University of Illinois at Urbana-Champaign, Champaign, Illinois, United States

Abstract

Why do children learn some words before others? Understanding individual variability across children and also variability across words, may be informative of the learning processes that underlie language learning. We investigated item-based variability in vocabulary development using lexical properties of distributional statistics derived from a large corpus of child-directed speech. Unlike previous analyses, we predicted word trajectories cross-sectionally, shedding light on trends in vocabulary development that may not have been evident at a single time point. We also show that whether one looks at a single age group or across ages as a whole, the best distributional predictor is whether a child knows a word is the number of other known words with which that that word tends to co-occur.

Agent framing moderates concerns about moral contagion

Stephen Flusberg
Purchase College, SUNY, Purchase, New York, United States

Carly LaPlace
Purchase College, SUNY, Purchase, New York, United States

Abstract

Concerns about moral contamination shape peoples attitudes towards the objects they encounter in daily life. For example, money seems less desirable when it comes from a robbery (Tasimi & Gelman, 2017). Drawing on the theory of dyadic morality, we hypothesized that increasing an individuals sense of agency would reduce the salience of moral contagion and make people feel less vulnerable to moral contamination. Across two experiments, we adapted the study design of Tasimi and Gelman (2017), asking participants how much they desired a $1 (Experiment 1) or $100 (Experiment 2) bill associated with different negative moral histories. We modified the stimulus language so that participants were framed as either the moral agent or patient for all scenarios. As predicted, participants in the agent language condition expressed nearly the same level of desire regardless of the bills moral history, highlighting the role that feelings of agency play in moral decision-making.

The Impact of Speech Complexity on Preschooler Attention, Speaker Preference, and Learning

Ruthe Foushee
University of California, Berkeley, Berkeley, California, United States

Mahesh Srinivasan
UC Berkeley, Berkeley, California, United States

Fei Xu
UC Berkeley, Berkeley, California, United States

Abstract

How do children decide what speech to tune into and learn from? We extend the idea that learners preferentially attend to stimuli at an intermediate level of complexity to the domain of spoken language. Preschoolers (2.5-6.5 years in Exp.1 and 3.5-5.5 years in Exp. 2) watched two speakers alternate narrating pages of a textless picture book, before selecting who they wanted to hear finish the story. We manipulated the complexity of the narrators speech, such that the SIMPLE speaker used earlier-acquired words than the COMPLEX speaker. In Experiment 1, both speakers introduced rare target words that children were later tested on. While children learned an impressive number of them, the inclusion of these rare words may have made both speech streams too complex for children to show a systematic preference for one over the other. In Experiment 2, we narrowed our age range, and amplified the contrast in complexity between the two speech streams. Preliminary results suggest that children discriminated between the two levels of complexity, systematically selecting the simpler speaker to finish the story. These results suggest that preschoolers can track the relative complexity of different linguistic inputs, opening the possibility that they may actively direct their attention toward linguistic input that is more appropriate for them.

Experimental Investigation on Top-down and Bottom-up Processing in Comprehension of Graphs to Justify Decisions

Misa Fukuoka
Nagoya University, Nagoya Aichi, Japan

Kazuhisa Miwa
Nagoya University, Nagoya-shi, Aichi-ken, Japan

Abstract

Authors (2017) examined decision-making processes together with graph comprehension and in particular the influence of bottom-up and top-down processing on them. Using an altered procedure, this study examined bottom-up and top-down processing relative to graph comprehension where a decision is made first, followed by graph comprehension. We compared the results of the two studies. Some of the results observed in the previous study were not observed in this study, suggesting that the influence of impressions provisionally formed on graph comprehension was mitigated to justify the declared decision in advance. Attitude s that individuals have in a daily life were observed to have an influence in the decision in both the previous and current studies, showing that it strongly influences decision making regardless of the degree to which the graph is comprehended.

A New Class Of Proximity Data Obtained From Dictionary Networks

Camilo Garrido
Universidad de Chile, Santiago, Chile

Claudio Gutierrez
Universidad de Chile, Santiago, Chile

Guillermo Soto
Universidad de Chile, Santiago, Chile

Abstract

Background. Proximity data is a notion that indicates the degree of psychological closeness of concepts. It includes, among others, judgments of similarity, relatedness and cause-effect. Obtaining proximity data is challenging because it involves experts, corpora and people. On the other hand, dictionaries are fair representations made by experts (and thus, good proxies) of the lexicon and linguistic heritage of people.

Methods. We present a method to automatically obtain proximity data from dictionaries. We construct a network representation of a dictionary; exploit classical techniques on networks to build a similarity matrix; extract parameterized clouds of lexical proximity; test them with native speakers.

Results. Preliminary evaluations show that the method captures word associations significant to humans. Although the research was done in Spanish, the methods are easily reproducible in other languages.

Conclusions. Dictionaries are good sources of proximity data. We conjecture that dictionary networks are good proxies to human mind semantic associations.

Human Visual Object Similarity Judgments are Viewpoint-Invariant and Part-Based as Revealed via Metric Learning

Joseph German
University of Rochester, Rochester, New York, United States

Robert Jacobs
University of Rochester, Rochester, New York, United States

Abstract

We describe and analyze the performance of metric learning systems, including deep neural networks (DNNs), on a new dataset of human similarity judgments of Fribbles, naturalistic, part-based objects. Metrics trained using pixel-based or DNN-based representations fail to explain our experimental data, but a metric trained with a viewpoint-invariant, part-based representation produces a good fit. We also find that although neural networks can learn to extract the part-based representation—and therefore should be capable of learning to model our data—networks trained with a triplet loss function based on similarity judgments do not perform well. We analyze this failure, providing a mathematical description of the relationship between the metric learning objective function and the triplet loss function. The comparatively poor performance of neural networks appears to be due to the nonconvexity of the optimization problem in network weight space. We discuss the implications for neural network research as a whole.

Reinstatement of Old Memories and Integration with New Memories

Pierre Gianferrara
Carnegie Mellon University, Pittsburgh, Pennsylvania, United States

Marlieke van Kesteren
Vrije Universiteit, Amsterdam, Netherlands

Martijn Meeter
Vrije Universiteit, Amsterdam, Netherlands

Abstract

The acquisition of new knowledge relies on our ability to connect old information to new information using semantic networks. This process can be referred to as memory integration. In this study, we investigated how such integration may aid memory reactivation, defined as the retrieval of previously encoded information. In addition, we were interested in whether congruency (or semantic similarity) between two separately learned associations (AB-AC) enhances memory integration. University students learned congruent and incongruent AB-AC associations in an fMRI scanner and reported subjective reactivation. In addition to a behavioral score, we measured the degree of neural activity in the PPA to test for potential effects of reinstatement (neural reactivation) using the multivoxel pattern analysis (MVPA) technique. Our analyses revealed a robust effect of memory reactivation (behaviorally) and reinstatement (neurally). An effect of congruency was also found behaviorally, but was not evident in the PPA.

Why Are Some Online Educational Programs Successful?: A Cognitive Science Perspective

Marissa Gonzales
Georgia Institute of Technology, Atlanta, Georgia, United States

Ashok Goel
Georgia Institute of Technology, Atlanta, Georgia, United States

Abstract

Massive Open Online Courses (MOOCs) once offered the promise of accessibility and affordability. However, MOOCs typically lack expert feedback and social interaction, and have low student engagement and retention. Thus, alternative programs for online education have emerged including an online graduate program in computer science at a major public university in USA. This program is considered a success with over 9000 students now enrolled in the program. We adopt the perspective of cognitive science to answer the question why do only some online educational courses succeed? We measure learner motivation and self-regulation in one course in the program, specifically a course on artificial intelligence (AI). Surveys of students indicate that students self-reported assessments of self-efficacy, cognitive strategy use, and intrinsic value of the course are not only fairly high, but also generally increase over the course of learning. This data suggests that the online AI course might be a success because the students have high self-efficacy and the class fosters self-regulated learning.

A Convolutional Self-organizing Map for Visual Category Learning

Chris Gorman
University of Otago, Dunedin, New Zealand

Lech Szymanski
University of Otago, Dunedin, New Zealand

Anthony Robins
University of Otago, Dunedin, New Zealand

Alistair Knott
University of Otago, Dunedin, New Zealand

Abstract

In this paper we present a novel neural network architecture that aims to combine the highly popular and successful convolutional neural network architecture with the learning mechanism of an unsupervised self-organizing map. The convolutional self-organizing map (ConvSOM) is a hierarchical network consisting of several independent self-organizing maps. It incorporates features associated with convolutional networks, such as weight sharing, spatial pooling, and hierarchical abstraction, with the unsupervised, topographically organized self-organizing map. We will show that the resulting architecture performs poorly on the MNIST data set, but offers interesting avenues for further research.

Boundaries of Creativity: Thick or Thin Organization?

Jean-Christophe Goulet-Pelletier
University of Ottawa, Ottawa, Ontario, Canada

Denis Cousineau
University of Ottawa, Ottawa, Ontario, Canada

Abstract

Semantic organization of knowledge has a long history in theories of creativity. Flexibility of thinking and distant connections are indispensable elements of a creative network. Simultaneously, convergence of thoughts and evaluation of ideas are essential at many stages of the creative process. The current study evaluates these complementary aspects through the lens of an exploratory concept known as mental boundaries. Correlation analyses are used to compare flexible and rigid tendencies of organizing the world, the concepts of intellect, schizotypy, perfectionism, divergent thinking and self-perceived creativity. Results (n = 316) reveal an interesting contrasting pattern where divergent thinking is significantly related to flexible internal and external organizations, whereas self-perceived creativity is significantly related to rigid external and non-significantly related to rigid internal organizations. The present findings have implications for the measurement of creativity and the identification of factors that facilitate the creative process.

Failing to see what you are a part of: Wisdom among crowd members

Ulrike Hahn
Birkbeck, University of London, London, London, United Kingdom

Toby Pilditch
University College London, London, United Kingdom

Nicole Cruz
Birkbeck, University of London, London, United Kingdom

Abstract

One of the key features that make human cognition so successful is its social basis. The fact that we can exchange information with others is integral to the knowledge humans have collectively built up over centuries. One place where this can readily be seen is in the aggregation of judgments. As is well documented, aggregates of individual judgments are often considerably more accurate than the individual judgments themselves, giving rise to so-called wisdom of the crowd effects. A key determinant of the benefits of aggregation is the degree of dependency between judgments. Here, we probed experimentally lay peoples understanding both of the value of aggregation and informational dependency, using a numerical prediction task. We found only an equivocal trend in people's understanding of the value of aggregation, and no clear evidence of people's understanding of the accuracy benefit of diversity.

Demonstrating the Impact of Prior Knowledge in Risky Choice

Mathew Hardy
Princeton University, Princeton, New Jersey, United States

Tom Griffiths
University of California, Berkeley, Berkeley, California, United States

Abstract

Bayesian models that optimally integrate prior probabilities with observations have successfully explained many aspects of human cognition. Research on decision-making under risk, however, is usually done through laboratory tasks that attempt to remove the effect of prior knowledge on choice. To test the effects of manipulating prior probabilities on participants' choices, we ran a large online experiment in which risky options paid out according to the distribution of Democratic and Republican voters in unknown congressional districts in known US states. This setup allows us to directly manipulate prior probabilities while holding observations constant and to compare people's choices with the options' true posterior values. We find that people's choices are appropriately influenced by prior probabilities, and discuss how the study of risky choice can be integrated into the Bayesian approach to studying cognition.

The role of AMPA receptor exchange in systems memory reconsolidation: A computational model

Peter Helfer
Dept. of Psychology, Montreal, Quebec, Canada

Thomas Shultz
McGill University, Montreal, Quebec, Canada

Abstract

In the mammalian brain, a newly acquired memory depends on the hippocampus for maintenance and recall, but over time the neocortex takes over these functions, rendering the memory hippocampus-independent. The process responsible for this transformation is called systems memory consolidation. Interestingly, retrieval of a well-consolidated memory can trigger a temporary return to a hippocampus-dependent state, a phenomenon known as systems memory reconsolidation. The neural mechanisms underlying systems memory consolidation and reconsolidation are not well understood. Here, we propose a neural model based on well-documented mechanisms of synaptic plasticity and stability and describe a computational implementation that demonstrates the models ability to account for a range of findings from the systems consolidation and reconsolidation literature. Based on the computational model, we derive a number of predictions and suggest experiments that may put them to the test.

Statistical Learning Ability as a Measure of Cognitive Function

Steffen Herff
Agency for Science, Technology and Research (A*STAR), Singapore, Singapore

Nur Amirah Abdul Rashid
Institute of Mental Health, Singapore, Singapore

Jimmy Lee
Institute of Mental Health, Singapore, Singapore

Tih Shih Lee
Duke-NUS Medical School, Singapore, Singapore

Kat Agres
Institute of High Performance Computing, A*STAR, Singapore, – Select State/Province –, Singapore

Abstract

Statistical Learning (SL), the ability to extract probabilistic information from the environment, is a subject of much debate. It appears intuitive that such a profound mechanism of learning should carry predictive power towards general cognitive ability. Yet, previous attempts have struggled to link SL ability to measures of general cognitive function, suffering from low correlations and mediocre test-retest reliability. Here, we deploy a new continuous auditory SL task that achieves high test-retest reliability (r = .8) and shows that SL ability does significantly correlate with general cognitive function (up to r =. 56). Results are discussed in light of i) the theoretical implications of the high test-retest reliability of our novel SL task, ii) SL ability as a marker of general cognitive function, and iii) future methodological considerations.

Prepare to Swear: Considering Phonological Preparation of Taboo Words

Kathryn Hodges
Muhlenberg College, Allentown, Pennsylvania, United States

Alyce Huot
Muhlenberg College, Allentown, Pennsylvania, United States

Alexandra Frazer
Muhlenberg College, Allentown, Pennsylvania, United States

Hazem Abdelaal
Muhlenberg College, Allentown, Pennsylvania, United States

Jessica Oxer
Muhlenberg College, Allentown, Pennsylvania, United States

Abstract

The current studies investigated whether speakers can prepare to swear the same way they prepare non-taboo words. Swearing, when produced reflexively, has greater right hemisphere activation than normal production suggesting that swearing is a different linguistic process. We used a form preparation paradigm to consider phonological preparation for non-reflexive swearing. Participants were given two types of lists; homogeneous - all words shared phonological onset (e.g. /f/ - feet, fork, film, fuck), and heterogeneous nothing shared (e.g. film, shit, dock, poll). Results indicated the taboo words did not contravene preparation for homogeneous sets, and taboo words were facilitated similarly to non-taboo words. Next, we tested variable homogeneous sets (taboo item was inconsistent with majority onset, e.g. shit, film, fork, feet) to understand whether increased attention to taboo items would disable preparation. Results showed reduced preparation for items sharing the majority onset in variable sets, but preparation was still significant.

The Phenomenological Mind: Foregrounding Experience Through Cognitive Anti-realism and Quantum Cognition

Pamela Hoyte
Queensland University of Technology, Brisbane, Qld, Australia

Peter Bruza
Queensland University of Technology, Brisbane, QLD, Australia

Greg Thompson
Queensland University of Technology , Brisbane, Qld, Australia

Abstract

Two perspectives on human cognition are contrasted: the computational mind and the phenomenological mind. The computational mind derives from the cognitivist hypothesis and is based on representation, computation and realism. While useful for cognitive modelling, it is limited as it cannot cater for a cognitive agents experience. The phenomenological mind foregrounds experience by drawing on the concept of the enactive mind. The phenomenological mind refers to a view of cognition that is not predicated on the pre-existing mental representation of an objective world, and so is cognitively anti-realist and non-representational. Quantum cognition offers the prospect for cognitive modelers to step out of the computational mind but still have tools to rigorously and formally explore the anti-realism inherent to the phenomenological mind. The concept of contextuality from quantum cognition is proposed as a signature of experience in the phenomenological mind.

Understanding Individual Differences in Eye Movement Pattern During Scene Perception through Co-Clustering of Hidden Markov Models

Janet Hsiao
University of Hong Kong, Hong Kong, Hong Kong

Kin Yan Chan
University of Hong Kong, Hong Kong, Hong Kong

Yuefeng Du
City University of Hong Kong, Hong Kong, Hong Kong

Antoni Chan
City University of Hong Kong, Hong Kong, China

Abstract

Here we combined the Eye Movement analysis with Hidden Markov Models (EMHMM) method with the data mining technique co-clustering to discover participant groups with consistent eye movement patterns across stimuli during scene perception. We discovered explorative (switching between foreground and background information) and focused (mainly on foreground) eye movement strategy groups among Asian participants. In contrast to previous research suggesting a cultural difference where Asians adopted explorative and Caucasians used focused eye movement strategies, we found that explorative patterns were associated with better foreground object recognition performance whereas focused patterns were associated with better feature integration in the flanker task and higher preference rating of the scenes. In addition, images with a salient foreground object relative to the background induced larger individual differences in eye movements. Thus, eye movements in scene perception not only contribute to scene recognition performance, but also reflects individual differences in cognitive ability and scene preference.

The Effect of Semantic Diversity on Serial Recall for Words

Yaling Hsiao
University of Oxford, Oxford, Oxfordshire, United Kingdom

Matthew H.C. Mak
University of Oxford, Oxford, United Kingdom

Kate Nation
University of Oxford, Oxford, United Kingdom

Abstract

We investigated whether semantic diversity (SemD) influences immediate serial recall for words. SemD was calculated using LSA to quantify semantic similarity across contexts in large corpus. We examined the effects of SemD and imageability, a classic semantic variable. Participants saw and recalled the 6-word list by typing out the words in correct serial order. Experiment 1 was conducted in the laboratory (N=40). There was no main effect of SemD or imageability but exploratory analyses showed that SemD was modulated by list position and imageability. Among high-imageability words, low-SemD words were better recalled in latter positions (4 & 5) of the list. Experiment 2 conducted online (N=44) replicated the results, showing better recall of low-SemD words in the high-imageability condition at Position 5. These findings suggest that the availability of more semantic connections induces more competition between items, which impacts on performance later on in serial recall.

Examining the association between elementary students lexcio-syntactic writing features and cognitive-motivational profiles using Natural Language Processing

Melissa Hunte
University of Toronto (OISE), Toronto, Ontario, Canada

Christine Barron
University of Toronto, Toronto, Ontario, Canada

Jeanne Sinclair
University of Toronto (OISE), Toronto, Ontario, Canada

Hyunah Kim
OISE, University of Toronto, Toronto, Ontario, Canada

Samantha McCormick McCormick
University of Toronto (OISE), Toronto, Ontario, Canada

Megan Vincett Vincett
University of Toronto (OISE), Toronto, Ontario, Canada

Eunhee Eunice Jang
University of Toronto, Toronto, Ontario, Canada

Abstract

Natural language processing (NLP) provides an innovative avenue to understand and explore human language content, yet minimal research has utilized it to classify students literacy, cognition, or motivation. This study investigated the association between grade 4-6 students (n = 143) writing and their cognitive-motivational profiles (CMPs) based on their self-regulated learning, locus of control, writing self-efficacy, and goal-orientation. LPA (Mplus 7.4) results indicated a two-class CMP solution with predominantly positive or negative CMPs. Using NLP, 404 lexico-syntactic writing features were extracted from students writing. Random forest with 10-fold cross-validation was implemented in Weka 3.8 (with SMOTE to equate class instances) to accurately (93%) classify students CMPs (class 1 True Positive Rate (TPR) = .942; class 2 TPR = .925) based on the NLP-processed lexico-syntactic writing features. These results highlight the potential for machine learning to analyze students writing and accurately classify learner profiles to provide formative feedback and customized interventions.

How does art appreciation promote artistic inspiration?

Chiaki Ishiguro
Kanazawa Institute of Technology, Kanazawa, Japan

Takeshi Okada
The University of Tokyo, Tokyo, Japan

Abstract

Through art appreciation, viewers are sometimes inspired to express or implement creative ideas. Such an experience is thought to be important for art learning. In this study, we conduct a questionnaire to examine how art appreciation promotes creative inspiration in non-experts. We hypothesize that: (a) individual experience of art-related activities and self-evaluation of artistic expression affect creative inspiration, mediated by the method of appreciation of artworks; and (b) the type of artworks affects creative inspiration, mediated by the method of appreciation of artworks. The participants were 373 adults, who were not art professionals (179 women, age: M = 45.02, SD = 13.45, range: 20-69 years). The data are analyzed using structured equation modeling for the two hypotheses. The two hypotheses are mostly supported, suggesting that self-evaluation of artistic expression and the type of artworks (especially classic works of art) influence creative inspiration, mediated by the method of appreciation of artworks. However, experience of art-related activities has no significant direct effect on inspiration for artistic creation.

Learning to control the others body facilitates the embodied perspective taking

Ryota Ishikawa
University of Tsukuba, Tsukuba, Ibaraki, Japan

Kyohei Sasaki
University of Tsukuba, Tsukuba, Japan

Saho Ayabe-Kanamura
University of Tsukuba, Tsukuba, Ibaraki, Japan

Jun Izawa
University of Tsukua, Tsukuba, Japan

Abstract

Perspective taking, a cognitive process of understanding information from the others viewpoint, is essential for forming communication skills. Whereas this process is considered to involve detachment of the reference frame from the own eye and attachment of it to the others eye, we instead hypothesized here that it is mediated by representing the others intrinsic (i.e., proprioceptive) coordinate frame, since our cognitive abilities often rely on the physical presence. To examine this possibility, we asked the participants to learn to control avatars motion in the virtual reality space from the third-person perspective and sought interaction between the ability to represent avatars intrinsic coordinate systems via motor adaptation and the ability to take avatars spatial perspective. We found significant facilitation of perspective taking ability by the motor adaptation experience, which supports our hypothesis that the perspective taking encompasses a process of representing the others intrinsic coordinate frame. We suggest that the perspective taking is an embodied cognitive process which underpins theory of mind and empathy.

Spatial Updating Based on Visually Signaled Self-motion in Virtual Reality

Georg Jahn
Chemnitz University of Technology, Chemnitz, Germany

Manuel Dudczig
Institute for Machine Tools and Production Processes (IWP), Chemnitz, Saxony, Germany

Philipp Klimant
Institute for Machine Tools and Production Processes, Chemnitz, Saxony, Germany

Abstract

Spatial updating during self-motion can be effortless, however, in virtual reality if there are inconsistent cues about self-motion, spatial updating of egocentric representations of object locations usually relies on perceived scene motion or imagery of a spatial situation model. Strong presence and illusory self-motion with a quick onset are presumed necessary for effortless spatial updating if self-motion is signaled visually only. In the reported experiment, participants performed spatial updating compensating for visually signaled forward self-motion in a virtual scene presented in a head-mounted display. Higher visual detail in the scene improved performance only slightly. Overall, the result pattern suggests that participants did not experience illusory self-motion that could support effortless updating despite more favorable conditions than in a previous study. Several modifications to the experiment are discussed as further tests of conditions fostering effortless updating in virtual reality.

Emergence: A Proposal for a Foundational Revolution in Cognitive Science

Jay Jennings
Institute of Cognitive Science, Ottawa, Ontario, Canada

Abstract

Emergence has been a fundamental part of physics, chemistry, and biology since the turn of the century. The sub-disciplines of cognitive science have all adopted emergentist approaches in many areas within their field, yet cognitive science as a whole lacks an overarching theory between the sub-disciplines. Therefore, I propose that emergence is a valuable conceptual tool for unifying the sub-disciplines of cognitive science, as it will facilitate communication via a shared emergentist framework. Although there are several definitions of emergence, cognitive science can benefit from an overarching view that regardless of discipline, reductionistic approaches are unable to describe cognition from the macro to the micro without invoking emergent stages of explanation. The reluctance to adopt an emergent paradigm surrounds the issue that emergent phenomena cannot be predicted from their component parts, which challenges the way experiments in cognitive science are designed and conducted, and how cognition is modeled computationally.

Do Deep Neural Networks Model Nonlinear Compositionality in the Neural Representation of Human-Object Interactions?

Aditi Jha
IIT Delhi, New Delhi, India

Sumeet Agarwal
IIT Delhi, New Delhi, India

Abstract

Visual scene understanding often requires the processing of human-object interactions. Here we seek to explore if and how well Deep Neural Network (DNN) models capture features similar to the brain's representation of humans, objects, and their interactions. We investigate brain regions which process human-, object-, or interaction-specific information, and establish correspondences between them and DNN features. Our results suggest that we can infer the selectivity of these regions to particular visual stimuli using DNN representations. We also map features from the DNN to the regions, thus linking the DNN representations to those found in specific parts of the visual cortex. In particular, our results suggest that a typical DNN representation contains encoding of compositional information for human-object interactions which goes beyond a linear combination of the encodings for the two components, thus suggesting that DNNs may be able to model this important property of biological vision.

Single Template vs. Multiple Templates: Examining the Effects of Problem Format on Performance

Yang Jiang
Educational Testing Service, Princeton, New Jersey, United States

Ma. Victoria Almeda
TERC, Cambridge, Massachusetts, United States

Shimin Kai
Teachers College Columbia University, New York, New York, United States

Ryan Baker
University of Pennsylvania, Philadelphia, Pennsylvania, United States

Korinn Ostrow
Worcester Polytechnic Institute, Worcester, Massachusetts, United States

Paul Salvador Inventado
California State University Fullerton, Fullerton, California, United States

Peter Scupelli
Carnegie Mellon University, Pittsburgh, Pennsylvania, United States

Abstract

Classroom and lab-based research have shown the advantages of exposing students to a variety of problems with format differences between them, compared to giving students problem sets with a single problem format. The rapid development of technologies such as intelligent tutoring systems (ITS) in education affords the opportunity to automatically generate and adapt problem content for practice and assessment purposes. In this paper, we investigate whether this approach can be effectively deployed to an ITS, conducting a randomized controlled trial to compare students who practiced problems based on a single template and those who were exposed to problems based on multiple templates, both in the same ITS. Results show no statistically significant difference in the two conditions on students post-test performance and hint request behavior. However, students who saw multiple templates spent more time to answer the practice items compared to students who solved problems of a single structure.

Assessing Integrative Complexity as a Measure of Morphological Learning

Tamar Johnson
University of Edinburgh, Edinburgh, United Kingdom

Jennifer Culbertson
University of Edinburgh, Edinburgh, United Kingdom

Hugh Rabagliati
University of Edinburgh, Edinburgh, United Kingdom

Kenny Smith
University of Edinburgh, Edinburgh, United Kingdom

Abstract

Morphological paradigms differ widely across languages in their size and number of contrasts they mark. Recent work on morphological complexity has argued that certain features of even very large paradigms make them easy to learn and use. Specifically, Ackerman & Malouf, 2013 propose an information-theoretic measure, i-complexity, which captures the extent to which forms in the paradigm predict each other, and show that languages which differ widely in surface complexity exhibit similar i-complexity; in other words, paradigms with many contrasts reduce the learnability challenge for learners by having predictive relationships between inflections. We present three artificial language learning experiments testing whether i-complexity in fact predicts learnability of nominal paradigms where nouns inflect for class and number. Our results reveal only weak evidence that paradigms with low i-complexity are easier to learn than paradigms with high i-complexity. We suggest that alternative aspects of complexity may have a larger impact on learning.

Elicitation and Assessment of Emotion in Computational Rationality

Jussi Jokinen
Aalto University, Helsinki, Finland

Viet Ba Hirvola
Aalto University, Espoo, Finland

Abstract

Computational modelling of human emotion has a promising outlook within the approach of computational rationality, which formalises adaptive behaviour as a bounded optimisation problem. However, testing different hypothetical emotion models under this approach is hindered by lack of structured data, that have been collected in experimentation coherent with the underlying formal assumptions. Here, we design an interactive task that is used to elicit and assess emotion, and aligns with the problem solving formalism of a partially observable Markov decision problem. From the literature on emotion modelling, we derive hypotheses about what affects emotional responses, and use the collected data to test the hypotheses. We demonstrate how emotion can be assessed in a semi-continuous manner throughout the trials of the experiment, and in a way that can be used to test computational rationality models of emotion.

How the Organization of Autobiographical Memories Changes Over Time

Yoed Kenett
University og Pennsylvania, Philadelphia, Pennsylvania, United States

Alexa Tompary
University of Pennsylvania, Philadelphia, Pennsylvania, United States

Sharon Thompson-Schill
University of Pennsylvania, Philadelphia, Pennsylvania, United States

Abstract

Cognitive scientists have discovered much about the acquisition, consolidation, and retrieval of episodic memories; however, much less is known about how memories of our daily experiences are organized, nor how this organization may change as memories become consolidated. Here, we apply computational network science methodologies to quantify the organization of recent (within the past year) and remote (5 10 years ago) autobiographical memories and quantitatively examine how these networks change over time. We found that remote memories exhibited higher global connectivity relative to recent memories, and that this increased connectivity is coupled with lower subjective ratings of vividness. Our results demonstrate how such cognitive features of episodic memory can be quantitatively examined and shed novel light on the organization and reconfiguration of episodic memories over time.

Learning to Recognize Uncertainty: Effects of Disconfirming Evidence on Confidence Scale Use in Preschoolers

Isabella Killeen
University of California- San Diego, La Jolla, California, United States

Caren Walker
University of California San Diego, La Jolla, California, United States

Abstract

Although young learners often express overconfidence, research has demonstrated that 3- to 4-year-old children may be able to use a confidence scale to discriminate between their correct and incorrect responses. The current research introduces a novel paradigm to facilitate childrens reflection and reporting of confidence, based on the presentation of disconfirming evidence. This paradigm presents 3-, 4- and 5-year-olds with windows of varying occlusion (none, partial, and full). Children used a 3-point scale to assess their confidence that a target shape was behind each window. In four conditions, we varied when and whether children received disconfirming evidence. Results suggest that violating childrens expectations about the presence of the target shape on the first trial results in improves confidence calibration on future trials. Results also suggest that baseline confidence scale use improves with age. We discuss theoretical implications for development of uncertainty monitoring and potential applications of this novel paradigm.

Measuring Selective Sustained Attention in Children with TrackIt and Eyetracking

Jaeah Kim
Carnegie Mellon University, Pittsburgh, Pennsylvania, United States

Shashank Singh
Carnegie Mellon University, Pittsburgh, Pennsylvania, United States

Emily Keebler
Carnegie Mellon University, Pittsburgh, Pennsylvania, United States

Erik Thiessen
Carnegie Mellon University, Pittsburgh, Pennsylvania, United States

Anna Fisher
Carnegie Mellon University, Pittsburgh, Pennsylvania, United States

Abstract

Measuring selective sustained attention (SSA) development in preschool-aged children has been difficult due to challenges in designing age-appropriate measurement paradigms. The TrackIt task, together with eye-tracking and a recently proposed Bayesian-model based eye-tracking analysis method, creates opportunity for fine-grained measurement of SSA in young children. The current study 1) provides the first rigorous validation of this method by comparing model judgments with human video-coding of the data, and 2) further explores potential uses of this method for providing nuanced measures of SSA. More specifically, we use the analysis method to explore different ways of characterizing SSA based on eye-gaze data obtained during TrackIt with 3- to 6-year old children. We look at patterns of in-trial eye-gazing across age and across time.

Information Distribution Depends on Language-Specific Features

Josef Klafka
University of Chicago, Chicago, Illinois, United States

Dan Yurovsky
University of Chicago, Chicago, Illinois, United States

Abstract

Language can be thought of as a code: A system for packaging a speakers thoughts into a signal that a listener must decode to recover some intended meaning. If language is a near-optimal code, then speakers should structure information in their utterances to minimizes the impact of errors in production or comprehension. To examine the distribution of information within utterances, we apply information-theoretic methods to a diverse set of languages in various spoken and written corpora. We find reliably non-uniform and cross-linguistically variable information distributions across languages. These distributions are consistent across contexts, and are predictable from typological features, most notably canonical word order. However, when we include even a small amount of predictive context (bigrams or trigrams), the language-specific shapes disappear, and all languages are characterized by uniform information distribution. Despite cross-linguistic variability in communicative codes, speakers structure their utterances to preserve uniform information distribution and support successful communication.

Exploring Monaural Auditory Displays that Convey Positional Information to Users

Takanori Komatsu
Meiji University, Tokyo, Japan

Masahiro Yamada
Meiji University, Tokyo, Japan

Seiji Yamada
National Institute of Informatics, Tokyo, Japan

Abstract

The purpose of this study is to confirm whether monaural auditory displays that indicate leftward and rightward directions to users can be used together with speech sounds in order to convey positional information to users. We conducted two experiments; experiment 1 was for investigating how a speech sound followed by auditory displays can convey three positions, right, center, and left, to participants, and experiment 2 was for exploring the effects of the durations of these auditory displays on how users interpreted these pieces of positional information. As a result of experiment 1, a speech sound followed by monaural auditory displays with durations of 0.25, 0.50, and 0.75 sec succeeded in conveying the three pieces of positional information to users. As a result of experiment 2, the speech sound followed by monaural auditory displays with durations of 0.25, 0.50, 0.75 or 1.00 sec was interpreted by users correctly.

How to find axioms for finite domains: A computational exploration of mathematical discovery

Gordon Krieger
McGill University, Montral, Quebec, Canada

Dirk Schlimm
McGill University, Montral, Quebec, Canada

Abstract

Axioms are pervasive in mathematics and formulating the axioms for a particular discipline has often been an important step in the development of mathematics. One way mathematicians arrive at axioms is by characterizing a given domain that consists of objects (e.g., numbers or points and lines) and relations between them. We present a software system that, given a set of objects and relations as input, determines, first, a set of first-order formulas that are satisfied in that domain, and, second, a set of axioms from which all of these formulas can be derived. Several domains are used to illustrate our program. By comparing the axioms for different domains, analogies between these domains can be expressed, such as structural and invariance properties. From the complexities of the implementation and the discussion of various examples, conclusions are drawn about the process of axiomatization in mathematical practice.

Choosing the unimaginable: Social psychological factors in seeking transformative experiences

Marta Kryven
MIT, Cambridge, Massachusetts, United States

Laura Niemi
University of Toronto, Toronto, Ontario, Canada

Laurie Paul
Yale University, New Haven, Connecticut, United States

Josh Tenenbaum
MIT, Cambridge, Massachusetts, United States

Abstract

How do people make transformative decisions (the outcomes of which are hard to imagine, and which might change one's self in lasting ways)? We investigate social psychological factors that contribute to making transformative decisions in contrast to ordinary decisions (with easily imaginable outcomes). We show that transformative decisions are uniquely predicted by a desire for self-improvement and forming new social bonds. However, contrary to our expectations, epistemic curiosity did not play a role in making transformative decisions. In contrast, ordinary decisions are uniquely predicted by the preferences of the community, and younger age. We identify important differences that point to separate cognitive mechanisms used to evaluate transformative decisions.

Various sources of distraction in analogical reasoning

Hanna Kucwaj
Jagiellonian University, Krakow, Poland

Jan Jastrzebski
Jagiellonian University, Krakw, Poland

Micha Ociepka
Jagiellonian University, Krakw, Poland

Adam Chuderski
Jagiellonian University, Krakow, Poland

Abstract

Two leading analogical reasoning paradigms: A:B::C:D task and scene analogies, to date studied in isolation, were applied to the same 61 participants. The former task included 3 types of distracting response options (relational, semantic, and perceptual); the latter task imposed cross-mapping (response options that suggested a wrong structure to be mapped). First, relational and semantic, but not perceptual, distractors were similarly frequently selected, but their choices were weakly correlated. These choices were unrelated to cross-mapping in the other task, either. So, various sources of distraction can play a role in the analogical reasoning process.

Temporal Structure in Reaction Time Data is sensitive to exercised control

Devpriya Kumar
Indian Institute of Technology, Kanpur, Uttar Pradesh, India

Narayanan Srinivasan
University of Allahabad, Allahabad, India

Akanksha Malik
Indian Institute of Technology, Kanpur, India

Abstract

Hierarchical control theories of perception-action conceptualize action as control of input, occurring simultaneously at multiple levels. These levels differ in terms spatio-temporal proximity of the perception controlled. However, it is not clear how this interaction between different levels in a control hierarchy can be measured from the behavior of the organism. We propose that Long Range Temporal Correlations (LRTC) in RT data can be used as a measure of coupling between different control levels within such complex system. Participants perform the task of controlling a hierarchical stimulus either at global level or at local level in a noisy presentation, while the level of control and noise are manipulated. The results suggest that LRTC in control task is higher for global level of control compared to local level of control in the no noise condition. We discuss implications of the results for understanding of perception-action interactions as a complex dynamic system.

Rudimentary modeling of acceptability judgement from a large scale, unbiased data

Kow Kuroda
Kyorin University, Mitaka, Tokyo, Japan

Hikaru Yokono
Fujitsu Laboratories, Ltd., Kawasaki, Japan

Keiga Abe
Gifu Shotoku Colledge, Gifu, Japan

Tomoyuki Tsuchida
Kyushu University, Hakata, Japan

Yoshihiko Asao
National Institute of Communications Technology, Kyoto, Japan

Yuichiro Kobayashi
Nihon University, Tokyo, Japan

Toshiyuki Kanamaru
Kyoto University, Kyoto, Japan

Takumi Tagawa
University of Tsukuba, Tsukuba, Japan

Abstract

Acceptability Rating Data for Japanese (ARDJ) is a project that explores the true nature of acceptability judgement based on a large-scale survey using theoretically unbiased stimuli. Its main survey was carried out in 2018 in two phases with carefully constructed 300 stimulus sentences: Phrase 1 was a smaller scale experiment with roughly 300 college students; Phase 2 was a large scale web survey with over 1,600 participants.

This paper reports on phase 2 and provides two results: Analysis 1 brought us a good typology of 300 sentences; Analysis 2 implements explicit modeling of acceptability judgement using Semi-supervised local Fisher discriminant analysis.

The results, if combined, suggest that i) acceptability is not a simple dichotomous partitioning of stimuli; ii) acceptability is a complex property that emerges through an interplay among the three factors: 1) degree or strength of deviance, 2) syntactic and/or semantic complexity of stimulus, and 3) localizability of deviance.

How the Brain Learns Language: an Exploration of The Brain Areas Involved in Statistical Language Learning

Imme Lammertink
University of Amsterdam, Amsterdam Center for Language and Communication (ACLC), Amsterdam, Netherlands

Gillian Clark
Deakin University, School of Psychology, Cognitive Neuroscience Unit, Melbourne, Australia

Judith Rispens
University of Amsterdam, Amsterdam Center for Language and Communication (ACLC), Amsterdam, Netherlands

Jarrad Lum
Deakin University, School of Psychology, Cognitive Neuroscience Unit, Melbourne, Australia

Abstract

It has been suggested that the detection of statistical regularities in language a skill fundamental to language acquisition is supported by brain areas that are also involved in implicit motor skill learning. The present study is one of the first to explore this claim in an artificial language learning experiment. We used continuous theta-burst transcranial magnetic stimulation (cTBS) to temporarily inhibit functioning of the left dorsolateral prefrontal cortex (DLPFC) or the primary motor cortex (M1) in healthy adults. We hypothesized that the left DLPFC plays a role in adults detection of nonadjacent dependencies (NADs) and therefore that learning should be disrupted in the group of adults receiving cTBS to this area. Our results provide no evidence for (or against) this claim, however. An interesting exploratory result is that learning of NADs may be enhanced in adults who received cTBS to the M1 as compared to participants who received sham cTBS.

Expertise and Anchoring Bias in Medical Decision Making

Aron Liaw
UC San Francisco, San Francisco, California, United States

Matthew Welsh
University of Adelaide, Adelaide, South Australia, Australia

Hillary Copp
UC San Francisco, San Francisco, California, United States

Benjamin Breyer
UC San Francisco, San Francisco, California, United States

Abstract

Anchoring bias describes the tendency to base an estimate around a previously given value, the anchor. Herein, a cohort of 124 medical providers and trainees, from medical students to practicing physicians, were shown to display anchoring bias when faced with medical scenarios including an anchoring value in the form of a prior assessment. Anchoring bias did not vary significantly with participants level of training although tolerance to risk did. However, they showed increased reliance on the anchor when its source had greater expertise. Analyses showed no correlation between anchoring susceptibility and participants preference for Rationality or Intuition as measured by the Decision Styles Scale. The results suggest that medical decisions can be vulnerable to anchoring effects, particularly when the anchor is sourced from an authoritative source. Given that authoritative sources should be more knowledgeable, this is reasonable, but will hold true regardless of the accuracy of the anchoring value.

Selecting and evaluating evidence: The garden of forking information paths

Alice Liefgreen
University College London, London, United Kingdom

Toby Pilditch
University College London, London, United Kingdom

David Lagnado
University College London, London, United Kingdom

Abstract

In order to make accurate inferences and judgments, one needs to not only be able to aptly evaluate and integrate information, but be able to seek and acquire the right information in the first place. The present work explored human information acquisition and evaluation in a novel probability context and utilising a more naturalistic criminal investigation scenario. Focus was placed on exploring the relationship between searching for information, evaluating it and integrating it within ones belief model in order to make a causal judgement. Results indicated that although participants search choices approximated those of informed Bayesian OED models, belief updating accuracy systematically decreased throughout the task. Findings suggested a dichotomy between information evaluation and belief integration, questioning the descriptive abilities of OED principles to account for these processes. The implications of these finding in relation to the psychological literature of human inquiry are discussed.

Different Frames of Players and their Empathy as Motive of Prosocial Behavior in Digital Games

Ji Soo Lim
Dokkyo University, Soka, Saitama, Japan

Abstract

Advanced technologies used in games allow players to behave freely in the game world. Like in the real world, there may be complex motives for a behavior. Although how a player behaves in a game is afforded by the games rules, motives may differ depending on the type of player. For example, a player who regards the game as mere rule-based play may behave differently as compared to a player who perceives the game as another reality with its own rules and sociality. This study focuses on understanding players prosocial behavior in games and empathy as their motive. A survey was conducted to look at the relationships between prosocial behavior, empathy, and different types of players (depending on their interpretation of gameplay). The results showed that the type of player did not affect their levels of empathy, but it moderated the effect of empathy on prosocial behavior toward other characters.

Comparison of Chinese and Western Categorization: Based on Bayesian Model

Junyao Liu
Department of Psychology, Beijing Forestry University, Beijing, China

Yifei Wang
Department of Psychology, Beijing Forestry University, Beijing, China

Yingying Yin
Department of Psychology, Beijing Forestry University, Beijing, China

Wenxuan Hao
Department of Psychology, Beijing Forestry University, Beijing, China

Mingyi Wang
Department of Psychology, Beijing Forestry University, Beijing, Beijing, China

Fei Xu
UC Berkeley, Berkeley, California, United States

Abstract

Xu and Tenenbaum (2007a, 2007b) applied the Bayesian model to explain the impact of differences in exemplification on words learning, and they achieved milestones. It remains unexplored if there are differences when native language and culture are changed. Taking the same method as the original research, we added test after a long time interval, and use between-subject design to eliminate the practice effect. The results of Chinese adults and children show that: (1) The Bayesian model has stability over time and culture. (2) When the objects in the same category differ greatly from each other, the Bayesian model's predictive power on children's results is significantly reduced. (3) Since the low-level words in Chinese vocabulary are often composed of high-level words and adjectives, Chinese easier to generalize. (4) Results of Chinese subjects reflect more instinct rather than logical reasoning stylewhich is differ from westerners.

Gestures for Self Help Learning by Creating Models

Yang Liu
Teachers College, Columbia University, New York, New York, United States

Melissa Zrada
Teachers College, Columbia University, New York, New York, United States

Barbara Tversky
Columbia Teachers College/Stanford, New York, New York, United States

Abstract

People spontaneously gesture when studying spatial descriptions. Doing so improves comprehension and learning. Their gestures create spatial models of the described environments. Here, we address two questions in two experiments: will people gesture to study descriptions that are not inherently spatial, and will people gesture when information is presented visually rather than text. The answers to both questions are yes. Together, the results suggest that gestures facilitate comprehension and learning by creating spatial-motor representations that directly reflect meaning.

Inferring the social meaning of objects with intuitive physics and Theory of Mind

Michael Lopez-Brau
Yale University, New Haven, Connecticut, United States

Julian Jara-Ettinger
Yale University, New Haven, Connecticut, United States

Abstract

Humans primarily communicate through words and gestures. In some cases, however, humans also communicate indirectly through objects, such as trafc cones or stanchion ropes. How does the human mind generate and interpret the social meaning of objects? Here we show that a computational model that uses commonsense physics and Theory of Mind spontaneously gives rise to the ability to communicate through objects. As predicted by our model, we show that people can infer the communicative meaning of novel objects by reasoning about the costs they impose, even in the absence of a pre-existing convention. Moreover, we show that people store the meaning of an object after a single exposure and recognize it in subsequent encounters. Our model sheds light on how humans bootstrap cognitive capacities that we share with other animals to give rise to uniquely-human cognition.

Integration of gaze information during online language comprehension and learning

Kyle MacDonald
University of California Los Angeles, Los Angeles, California, United States

Elizabeth Swanson
Stanford University, Stanford, California, United States

Michael Frank
Stanford University, Stanford, California, United States

Abstract

Face-to-face communication provides access to visual information that can support language processing. But do listeners automatically seek social information without regard to the language processing task? Here, we present two eye-tracking studies that ask whether listeners' knowledge of word-object links changes how they actively gather a social cue to reference (eye gaze) during real-time language processing. First, when processing familiar words, children and adults did not delay their gaze shifts to seek a disambiguating gaze cue. When processing novel words, however, children and adults fixated longer on a speaker who provided a gaze cue, which led to an increase in looking to the named object and less looking to the other objects in the scene. These results suggest that listeners use their knowledge of object labels when deciding how to allocate visual attention to social partners, which in turn changes the visual input to language processing mechanisms.

Comparing cognitive models in dynamic agent-based models: A methodological case study

Jens Madsen
University of Oxford, Oxford, United Kingdom

Richard Bailey
University of Oxford, Oxford, United Kingdom

Ernesto Carrella
University of Oxford, Oxford, United Kingdom

Nicolas Payette
University of Oxford, Oxford, United Kingdom

Abstract

Dynamic models, such as agent-based models (ABMs), are becoming an increasingly common modelling tool in cognitive sciences. They enable cognitive scientists to explore how computational, analytic models scale up when placed in complex, interactive, and dynamic environments where agents can sequentially interact over time and in space. Frequently, ABMs are built to yield a particular behaviour (riots, echo chamber emergence, etc.). As such, some models may bake in the desired behaviour. However, many models may yield this behaviour, making it difficult to discriminate between the adequacies of each computational model. The paper directly addresses this methodological challenge. We explore a case study (fisheries). Agents make decisions in this dynamic and complex environment. Given a rich data set against which to calibrate and validate model predictions, we compare and contrast statistical, adaptive, and perfect agents. We show that adaptive computational agents equal statistical agents in calibration and outperform them for validation. In addition, we show that perfect and random agents fare poorly. This provides a method for using dynamic, agent-based models to choose between computational models

Spatial Representations of Symbolic Fractions and Nonsymbolic Ratios: SNARC Effect and Number Line Estimation

Rui Meng
University of Wisconsin Madison, Madison, Wisconsin, United States

Percival Matthews
University of Wisconsin - Madison, Madison, Wisconsin, United States

Abstract

Recent research on numerical cognition has begun to systematically detail the ability to perceive the magnitudes of symbolic fractions and non-symbolic ratios. The current study extended this line of research by investigating spatial representations of symbolic fractions and nonsymbolic ratios with two behavioral measures: the Spatial-Numerical Association of Response Codes (SNARC) effect and number line estimation. The two research questions were: 1) what are the similarities and differences of spatial representations between symbolic fractions and nonsymbolic ratios? 2) do mechanisms driving the SNARC effect and performance on number line estimation rely on a shared cognitive mechanism? Participants completed four tasks: magnitude comparison with symbolic fractions, magnitude comparison with nonsymbolic ratios, number line estimation with symbolic fractions, and number line estimation with nonsymbolic ratios. Results suggested the existence of both shared and specific spatial representations of symbolic fractions and nonsymbolic ratios. Moreover, individual participants SNARC effects and number line estimation performances were not correlated with each other. Findings further elucidate the relations between different spatial representations for symbolic fractions and nonsymbolic ratios and cast doubt on the prospect of their sharing common cognitive mechanisms.

An experiment in the neuroscience of learning interactions: The effect of agency on emotional processing in dyads learning physics with a serious computer game

Julien Mercier
UQAM, Montreal, Quebec, Canada

Ariane Paradis Ph.D. Student
Universit du Qubec Montral , Montral, Quebec, Canada

Kathleen Whissell-Turner
UQAM, Montreal, Quebec, Canada

Ivan Avaca
UQAM, Montreal, Quebec, Canada

Abstract

Many educational approaches assume that making the learner active leads to better learning although this improvement in learning has not be firmly quantified experimentally. The goal of this paper is to articulate a model of agency in cooperative learning based on a predictive cognitive architecture and to explore methodological strategies as well as theoretical and applied implications of agency in the study of cooperative learning, in this case with data on emotional processing. Results from 27 dyads (1 player and 1 watcher) who played a serious game for learning physics for 120 minutes show that agency has no effect on the overall quantity of emotional processing, but that the emotional processing of a watcher and player is synchronized. A watchers emotional processing may precede or be delayed from the players. The cornerstone of this framework is the notion of predictions, which unites top-down and bottom-up influences as modulated by the possibility for action (agency). The model presented is the foundation for process-oriented studies of cooperative learning from an educational neuroscience perspective.

Interlocutors preserve complexity in language

Madeline Meyers
University of Chicago, Chicago, Illinois, United States

Dan Yurovsky
University of Chicago, Chicago, Illinois, United States

Abstract

Why do languages change? One possibility is they evolve in response to two competing pressures: (1) to be easily learned, and (2) to be effective for communication. In a number of domains, variation in the worlds natural languages appears to be accounted for by different but near-optimal tradeoffs between these pressures. Models of these evolutionary processes have used transmission chain paradigms in which errors of learning by one agent become the input for the subsequent generation. However, a critical feature of human language is that children do not learn in isolation. Rather, they learn in communicative interactions with caregivers who draw inferences from their errorful productions to their intended interests. In a set of iterated reproduction experiments, we show that this supportive context can have a powerful stabilizing role in the development of artificial patterned systems, allowing them to achieve higher levels of complexity than they would by vertical transmission alone while retaining equivalent transmission accuracies.

The Role of Sketch Quality and Visuo-Spatial Working Memory in Science Accuracy

Dana Miller-Cotto
University of Pittsburgh, Pittsburgh, Pennsylvania, United States

Nicole Hallinen
Temple University, Philadelphia, Pennsylvania, United States

Julie Booth Ph.D.
Temple University, Philadelphia, Pennsylvania, United States

Abstract

Sketching is often a helpful strategy for solving science problems. We examined the role of visuo-spatial working memory and sketching in predicting science problem solving accuracy. Sketches were coded for quality based on whether they included elements and relationships in the sketches. Regression analyses were done regressing working memory on to science problem solving. A mediation analysis was also conducted to determine whether sketch quality mediated the relationship between working memory and science accuracy. Findings are discussed in terms off implications for education and classroom instruction.

The Cognitive Process of Reinterpreting Non-art Objects in an Art Context

Koto Minami
The University of Tokyo, Graduate School of Interdisciplinary Information Studies, Tokyo, Japan

Daichi Shimizu
Graduate shool of education, Bunkyo-ku, Tokyo, Japan

Takeshi Okada
The University of Tokyo, Tokyo, Japan

Abstract

In this study, we investigated the reinterpretation process of a non-art object. It is often said that a unique perspective different from daily life arises in the cognitive process of an art activity. We assumed that such a unique viewpoint can also be applied to non-art objects and people will discover new aspects of objects and/or their own viewpoints through the reinterpretation of non-art objects. We conducted a between-subjects experiment to investigate the process in detail. We expected the artistic context of the participant to influence the interpretation. We conducted two types of interventions to manipulate participants artistic context. The result of the experiment suggests that the artistic context influenced the interpretation process of non-art objects.

L1 Influence on Content Word errors in Learner English Corpora: Insights from Distributed Representation of Words.

Kanishka Misra
Purdue University, West Lafayette, Indiana, United States

Hemanth Devarapalli
Purdue University, West Lafayette, Indiana, United States

Julia Rayz
Purdue University, West Lafayette, Indiana, United States

Abstract

The first language of a person has been shown to influence the processing of words when they learn a new language. This has been previously researched in behavioral studies, as well as by using lexical distributions or co-occurrence counts between word combinations to detect errors. In this paper we follow the findings of two recent studies and test their hypotheses within the framework of two different word embedding models, based on their representation of the erroneous usage of concent words. Using an error-annotated corpus of essays written by learnings bellowing to 16 different first languages, we compare incorrect words and their correct replacements as vectors in English and the learners first language. The results are consistent with previous findings that the first language has an influence on errors in the second language. The relationships between typologically similar languages differed between the models of embedding, suggesting an avenue for future explorations.

Planning failures induced by budgetary overruns cause intertemporal impulsivity

Arjun Mitra
University of Allahabad, Allahabad, Uttar Pradesh, India

Narayanan Srinivasan
University of Allahabad, Allahabad, India

Nisheeth Srivastava
Indian Institute of Technology, Kanpur, UP, India

Abstract

Recent research has identified intertemporal impulsivity as a critical cognitive variable for explaining the autocatalytic nature of socioeconomic status (SES). But how exactly this relationship transpires has not been clearly identified. We present results from a novel experimental study, demonstrating that decision-makers' time preference becomes more present-focused when they experience budgetary overruns in a sequential decision-making task. On the basis of these results, we hypothesize that steep intertemporal discounting in low SES individuals may arise as a rational metacognitive adaptation to persistently experiencing planning and control failures in long-term plans. Consilient evidence in support of this hypothesis and downstream policy implications are briefly discussed.

Evaluation of Methods for Tracking Strategies in Complex Tasks

Jarrod Moss
Mississippi State University, Mississippi State, Mississippi, United States

Aaron Wong
Mississippi State University, Mississippi State, Mississippi, United States

Kevin Barnes
Mississippi State University, Mississippi State, Mississippi, United States

Jaymes Durriseau
Mississippi State University, Mississippi State, Mississippi, United States

Gary Bradshaw
Mississippi State University, Mississippi State, Mississippi, United States

Abstract

In complex tasks, high performers often have better strategies than low performers even with similar practice. Relatively little research has examined how people form and modify strategies in tasks that permit a large set of possible strategies. One challenge with such research is determining strategies based on behavior. Three algorithms were developed to track the task features people employ in their strategies while performing a complex task. An ACT-R model that performs the task was created to collect simulated data with a range of known strategies. The performance of the three algorithms is compared, and a decision tree classification algorithm yielded the best performance across the test cases. Summary data from applying the algorithms to human data on the tasks is also presented and highlights potential challenges for future work. However, this approach to tracking strategy exploration may enable further development of theories about how people search for good strategies.

”Give me a break”: Can brief bouts of physical activity reduce elementary children’s attentional failures and improve learning?

Grace Murray
Kent State University, Kent, Ohio, United States

Karrie Godwin
Kent State University, Kent, Ohio, United States

Abstract

In classroom settings, young children are frequently off-task, which may be due in part to childrens still-maturing attentional system. Lapses in attention may impede academic success by reducing the amount of time spent engaged in instructional activities. One popular strategy to increase on-task behavior is to provide brief physical activity (PA) breaks in between instructional tasks. Though PA breaks are hypothesized to increase on-task behavior, much is unknown regarding the effectiveness of breaks and their underlying mechanism(s). The present study systematically investigated the effectiveness of PA breaks, using direct measures of attention and learning. Break type (PA vs. Sedentary control) was manipulated within-subjects. Preliminary results indicate PA breaks benefit learning compared to the sedentary control (p=.03, Cohens d=.389). A marginally significant increase in on-task behavior was also found following the PA break. These results provide tentative support for the benefit of PA breaks for childrens attention and learning.

Gradations in task engagement emerge from metacognitive priority control

Dominic Mussack
University of Luxembourg, Belval, Luxembourg

Paul Schrater
University of Minnesota, Twin Cities, Minneapolis, Minnesota, United States

Abstract

Engagement is a critical motivational factor that has broad effects on learning, productivity, performance, and even satisfaction and happiness. However, it can also be impacted by a myriad of factors which have made it difficult to model and design interventions. Here we address this problem by developing an integrated metacognitive framework for understanding task engagement. We treat engagement as resulting from a unified metacognitive decision process where the gradient of engagement results from a common priority calculation. Priority signals are computed relative to a set of available tasks and updated across time and environmental changes. We propose a metacognitive controller makes decisions about both task switching (when to quit, next task) and cognitive resourcing (working memory, attention, etc) using the graded priority signals. By simultaneously choosing the task and allocating resources using the same graded signals, we capture the complex dependencies of engagement with task errors, performance, and time allocation.

The impact of sequences on the learning of contingencies at UK traffic lights

William G. Nicholson
University of Exeter, Exeter, United Kingdom

Ciro Civile
University of Exeter, Exeter, United Kingdom

IPL McLaren
University of Exeter, Exeter, Devon, United Kingdom

Abstract

Previous work has found that the contingencies experienced at UK traffic lights can affect drivers behavior potentially leading to risky driving. However, these studies did not account for the sequences experienced at traffic lights. This experiment seeks to rectify this. As with previous research we used an incidental go/no-go task in which colored shapes were stochastically predictive of whether a response was required. The stimuli encoded the contingencies of traffic lights and their appropriate response, for example, stimuli G was a go cue, mimicking the response to a green light. Crucially, cues were displayed in the sequences experienced at traffic lights. Supporting earlier work, the 50/50 cue that mimicked amber traffic lights was experienced as a go cue, and the stop cue that represented red lights was responded to as a neutral cue. The sequences seemed to enhance this pattern of learning with much larger effect sizes than previously found.

Investigating the Role of Future-orientated Feedback in Self-Monitoring Devices

Milena Nikolic
Queen Mary University London, London, United Kingdom

Magda Osman
Queen Mary University, London, United Kingdom

Abstract

Standard self-monitoring devices provide real-time daily feedback. This may not help users learn the long-term future cumulative effects of their behaviour because it orientates attention on the now. We test the hypothesis that future oriented feedback is more effective than real-time feedback in increasing users propensity to exercise. We asked 54 female treadmill users in a gym to report the feedback they got from the machine (calories burnt, time spent running and distance covered) upon finishing their workout and were then provided with additional feedback which varied in format across three between-subject conditions: day only feedback (no additional feedback), monthly feedback (additional projection of the future cumulative effect of the activity repeated daily after one month), and all times feedback (additional projection of the future cumulative effect of the activity repeated daily after one month and after one year). All participants were then asked about the extent to which they felt their own running workout affected their weight loss, as well the extent to which running leads to weight loss in general. They also all answered two questions aimed at measuring their time perspective after being exposed to the various feedbacks. In comparison to participants who had been exposed to the standard real time feedback, participants who had been exposed to the future oriented feedbacks perceived the causal connection between their own running workout and their weight loss as significantly higher, and reported a significantly more future oriented time perspective. The results highlight the need to consider time orientation as an important dimension to aid decisions through technologies.

On Language and Thought: How Bilingualism Affects Conceptual Associations

Siqi Ning
Northwestern University, Evanston, Illinois, United States

James Bartolotti
Northwestern University, Evanston, Illinois, United States

Viorica Marian
Northwestern University, Evanston, Illinois, United States

Abstract

Language experience influences cognition. Using behavioral and ERP measures, the present study examines whether experience with multiple languages can change how we form associations between concepts. Four experiments comparing bilingual and monolingual groups on semantic relatedness judgments indicate that highly proficient bilinguals perceive concepts as more related to one another than monolinguals. Results suggest that bilinguals denser lexical and phonological connections across their two languages may shorten semantic distances between concepts. This finding is consistent with connectionist models of language and suggests that the structure of the lexical and phonological systems may influence conceptual level associations. We conclude that bilingualism has consequences for the structure of the language system at the level of lexical-semantic connections.

Bringing Order to the Cognitive Fallacy Zoo

Ardavan S. Nobandegani
McGill University, Montreal, Quebec, Canada

William Campoli
McGill University, Montreal, Quebec, Canada

Thomas Shultz
McGill University, Montreal, Quebec, Canada

Abstract

Investigations into human decision-making have led to the discovery of numerous cognitive biases and fallacies, with new ones continually emerging, leading to a state of affairs which can fairly be characterized as the cognitive fallacy zoo! In this work, we formally present a principled way to bring order to this zoo. We introduce the idea of establishing implication relationships (IRs) between cognitive fallacies, formally characterizing how one fallacy implies another. IR is analogous to, and partly inspired by, the concept of reduction in computational complexity theory. We present several examples of IRs involving experimentally well-documented fallacies: base-rate neglect, availability bias, conjunction fallacy, decoy effect, framing effect, and Allais paradox. We conclude by discussing how our work: (i) allows for identifying those pivotal cognitive fallacies whose investigation would be the most rewarding research agenda, and (ii) permits a systematized, guided research program on cognitive fallacies, motivating influential theoretical as well as experimental avenues of future research.

On Robustness: An Undervalued Dimension of Human Rationality

Ardavan S. Nobandegani
McGill University, Montreal, Quebec, Canada

Kevin da Silva-Castanheira
McGill University, Montreal, Quebec, Canada

Timothy ODonnell
McGill University, Montreal, Quebec, Canada

Thomas Shultz
McGill University, Montreal, Quebec, Canada

Abstract

Human rationality is predominantly evaluated by the extent to which the mind respects the tenets of normative formalisms like logic and probability theory, and is often invoked by appealing to the notion of optimality. Drawing on bounded rationality, there has been a surge in the understanding of human rationality with respect to the mind's limited computational and cognitive resources. In this work, we focus on a fairly underappreciated, yet crucial, facet of rationality, robustness: insensitivity of a model's performance to miscalculations of its parameters. We argue that an integrative pursuit of three facets (optimality, efficient use of limited resources, and robustness) would be a fruitful approach to understanding human rationality. We present several novel formalizations of robustness and discuss a recently proposed metacognitively-rational model of risky choice (Nobandegani et al., 2018) which is surprisingly robust to under- and over-estimation of its focal parameter, nicely accounting for well-known framing effects in human decision-making under risk.

Decoy Effect and Violation of Betweenness in Risky Decision Making: A Resource-Rational Mechanistic Account

Ardavan S. Nobandegani
McGill University, Montreal, Quebec, Canada

Kevin da Silva-Castanheira
McGill University, Montreal, Quebec, Canada

Thomas Shultz
McGill University, Montreal, Quebec, Canada

A. Ross Otto
McGill University, Montral, Quebec, Canada

Abstract

A wealth of experimental evidence shows that, contrary to normative models of choice, people's preferences are markedly swayed by the context in which options are presented. In this work, we present the first resource-rational, mechanistic account of the decoy effect—a major contextual effect in risky decision making. Our model additionally explains a related, well-known behavioral departure from expected utility theory: violation of betweenness. We demonstrate that, contrary to widely held views, these effects can be accounted for by a variant of normative expected-utility maximization—sample-based expected utility model (SbEU; Nobandegani et al., 2018)—which acknowledges cognitive limitations. Our work is consistent with two empirically well-supported hypotheses: (i) In probabilistic reasoning and judgment, a cognitive system accumulates information through sampling, and (ii) People engage in pairwise comparisons when choosing between multiple alternatives.

1.9 Million Hits and Counting: An Investigation of the Cognitive Alignment of Hundred Boards for Subtraction Thinking

Julie Nurnberger-Haag
Kent State University, Kent, Ohio, United States

Karrie Godwin
Kent State University, Kent, Ohio, United States

Rachael Todaro
Kent State University, Kent, Ohio, United States

Abstract

The primary numerical activities in kindergarten through third grade are aimed at developing an understanding of the structure of base-ten numbers and learning to add and subtract with increasingly larger numbers. Many students in the U.S. continue to find this difficult. Thus, the most common instructional tools intended to support childrens learning of these ideas should be analyzed for their cognitive alignment and, if needed, redesigned for optimal learning. This study reports the findings from a study examining the cognitive alignment of a standard hundred board for the more difficult subtraction operation. Additionally, we investigate whether redesigning the hundred board such that addition goes up and subtraction goes down is more optimal for subtraction.

Verb arguments in Japanese picture books

Naho Orita
Tokyo University of Science, Tokyo, Japan

Asumi Suzuki
Tohoku University, Sendai, Japan

Yuichiro Matsubayashi
Tohoku University, Sendai, Japan

Abstract

Previous experiments have demonstrated that Japanese children can use the number of arguments and the case markers to learn novel verbs. However, these cues are mostly omitted in child-directed speech. We revisit this gap between the ability of children to use syntactic cues and the deficiency of such input by examining a different mode of input in the form of picture books. We built a Japanese picture book predicate-argument structure corpus containing annotations of predicate-argument structure and non-linguistic information. The analyses show that Japanese picture books contain more overt arguments and accusative case markers, and that these cues have significant influence on the prediction of verb transitivity. In addition, this study demonstrates that non-linguistic information (animacy and the numbers of potential referents) could help predict transitivity if learners are able to use these cues to infer the presence of null arguments.

How Different Metaphor Styles Impact on Creativity of the Poetry Receivers?

Magorzata Osowiecka
SWPS University of Social Sciences and Humanities, Sopot, Poland

Alina Kolaczyk
SWPS University of Social Sciences and Humanities, Sopot, Poland

Abstract

Poetry is one of the most creative uses of language. Yet the influence of poetry on creativity has received little attention. The present research aimed to determine how the reception of different types of poetry affect creativity levels. In two experimental studies, participants were assigned to two conditions: poetry reading and non-poetic text reading. Participants read poems (Study 1 = narrative/open metaphors; Study 2 = descriptive/conventional metaphors) or control pieces of non-poetic text. Before and after the reading manipulation, participants were given a test to determine levels of divergent thinking. In Study 1 (N = 107), participants showed increased fluency and flexibility after reading a narrative poem. In Study 2 (N = 131) reception of conventional, closed metaphorization significantly lowered fluency and flexibility (compared to reading non-poetic text). The most critical finding was that poetry exposure could either increase or decrease creativity level depending on the type of poetic metaphors.

Does Expressive Writing About Negative Emotions Influence Divergent Thinking?

Magorzata Osowiecka
SWPS University of Social Sciences and Humanities, Sopot, Poland

Radosaw Sterczyski
SWPS University of Social Sciences and Humanities, Sopot, Poland

Abstract

Many researchers claim that negative emotions inhibit creativity. However, describing emotions in a safe environment has beneficial effects: it allows for the restructuring of difficult experiences, as a result, we discover the world again, which can influence creativity. The classic method of writing about emotions is long-term one. The hereby study was an attempt to verify, if one session of expressive writing improves creative thinking. This hypothesis was tested in an experimental study by exposing participants (N = 60) to emotionally laden content. Participants viewed unpleasant images. The first group wrote about their emotions in connection with the images. The second described their typical day. At the end all participants solved creativity measure (Alternative Uses Task). After each stage, emotions of respondents were measured. The conducted analyses had shown that, performance was better in the unpleasant emotions describing condition. At the same time, negative emotions persistence has been observed.

Testing Accuracy, Additivity, and Sufficiency of Human Use of Probability Density Information in a Visuo-Cognitive Task

Keiji Ota
New York University, New York, New York, United States

Jakob Phillips
New York University, New York, New York, United States

Laurence Maloney
New York University, New York, New York, United States

Abstract

We tested three fundamental properties of Bayesian Decision Theoryaccuracy, additivity, and sufficiency. In Experiment 1, observers were shown a sample of dots from a bivariate Gaussian and estimate the probability that an additional sample would fall into specified regions. There were three types of regions: symmetric around the mean (S), the upper and lower halves of the symmetric region (SU and SL). In Experiment 2, the same observers were asked to maximize the expected rewards based on jointly sufficient statistics for given the sample (herein, mean and covariance of a Gaussian). In Experiment 1, We found that the observers estimates of symmetric region P[S] were close to accurate. However, they showed a highly patterned super-additivity: the sum of P[SU] + P[SL] ¿ P[S]. In Experiment 2, the observers violated sufficiency by assigning too much weight to a feature of the sample rather than jointly sufficient statistics.

Domestic Dogs Sensitivity to the Accuracy of Human Informants

Madeline Pelgrim
University of Toronto, Toronto, Ontario, Canada

Emma Tecwyn
Birmingham City University, Birmingham, United Kingdom

Julia Espinosa
University of Toronto, Toronto, Ontario, Canada

Angie Johnston
Yale University, New Haven, Connecticut, United States

Sarah MacKay Marton
University of Toronto, Toronto, Ontario, Canada

Daphna Buchsbaum
University of Toronto, Toronto, Ontario, Canada

Abstract

Domestic dogs excel at understanding human social-communicative gestures. The present study explores whether dogs can use peoples past accuracy when deciding who to trust. In Experiment 1, dogs watched an informant hide a treat under one of two containers and then point at one of them. Dogs were more likely to follow an accurate (pointed correctly) vs. the inaccurate (pointed incorrectly) informants point. In Experiment 2, dogs interacted separately with an accurate and inaccurate informant and again were more likely to follow an accurate point. In test trials, dogs did not witness hiding of the treat and saw the same two informants simultaneously point at different locations. Here, they chose between the locations at chance-level. Dogs inability to selectively follow a previously accurate informant when presented with conflicting information suggests that, unlike children, they may not be able to use past informant accuracy when choosing whose information to use.

The inverse operation modulates confidence

Gabriel Penagos
Pontificia Universidad Javeriana, Bogota, Colombia

Santiago Alonso Diaz
Universidad Javeriana, Bogota, Colombia

Abstract

Inversion is an essential operation, for instance in math (negatives) and action (to move in an opposite direction). Even though humans can invert is unclear how is implemented. There are two alternative hypotheses. The first possibility (H1) is that only positives are represented and negatives (inverses) are implemented as either a response (e.g. left to right) or task demand flip (e.g. ¿ to ¡). The second possibility (H2) is that both positives and negatives (inverses) are encoded. To disambiguate them, we ran two experiments where participants had to apply the inverse while implicitly reporting confidence. If inverting modifies encoding of otherwise identical stimulation then confidence should differ. We found that confidence was lower in inverse trials than direct/positive trials. This suggests that the inverse is not a simple response strategy or modification of task demands (H1), rather inverting modulates how cognitive information is encoded and used in the brain (H2).

Phonological and semantic processing in short-term memory

Theresa Pham
The University of Western Ontario, London, Ontario, Canada

Lisa Archibald
The University of Western Ontario, London, Ontario, Canada

Abstract

Much research has focused on phonological representation in verbal short-term memory (STM), with less attention paid to semantic representations despite evidence of linguistic long-term memory (LTM) effects. We investigate when phonological and semantic representations are activated in verbal STM: does it occur during retrieval (redintegration account) or there is direct access to language knowledge stored in LTM (language-based account). A probe recognition paradigm was used to test phonological and semantic encoding in verbal STM. Participants studied a list of words and then judged whether a probe word presented after the list rhymed or was synonymous to any item in the word list. Probe recognition was better was semantically processed words than the phonological task, suggesting that semantic encoding was evident at first exposure during encoding rather than a redintegration effect. It appears that semantic knowledge, in addition to and separate from phonological knowledge, is actively maintained in verbal STM.

Linguist Alignment in Collaborative and Conversational Contexts

Ramon Pieternella
Tilburg University, Tilburg, Noord-Brabant, Netherlands

Travis Wiltshire
Tilburg University, Tilburg, North Brabant, Netherlands

Abstract

Effective communication is a crucial factor contributing to successful collaborative problem solving (CPS) teams. Research in cognitive science has long shown evidence of linguistic alignment, or convergence in ways of speaking, but its functional role, if any, during CPS is unknown. Based on recent theorizing, we expected that both goal-oriented dialogue and non-goal-oriented dialogue should exhibit alignment. However, if linguistic alignment contributes to effective CPS, then conversations in this context should exhibit higher levels of alignment. In this study, we compared levels of syntactic and lexical alignment between a corpus of CPS dialogue and a corpus of spontaneous dialogue. Contrary to our prediction, we observed that the mean lexical alignment level was lower in the CPS corpus than in the Switchboard corpus. Implications for future research into linguistic alignment in CPS are discussed.

A round Bouba is easier to remember than a curved Kiki: Sound-symbolism can support associative memory

Marie Poirier
City, University of London, London, United Kingdom

Ren-Pierre Sonier
Universit de Moncton, Moncton, New Brunswick, Canada

Dominic Guitard
Universit de Moncton, Moncton, New Brunswick, Canada

Jean Saint-Aubin
Universit de Moncton, Moncton, New Brunswick, Canada

Abstract

Past research has shown that prior knowledge can support our episodic memory for recently encountered associations (Chalfonte & St-Giles, 1996; Naveh-Benjamin, 2000). Badham, Estes and Maylor (2012) for example, showed that integrative relationships between words help associative memory, even if the relationships are highly unfamiliar. A pair of words is integrative if the words make sense when considered together (e.g. monkey-foot). We extend this phenomenon to sound-symbolism associations; here, the latter refer to relationships between phonemes and object characteristics– relationships that participants readily find natural, even without prior knowledge of the items. For instance, the non-word maluma is much more readily associated with a random shape with rounded contours than with a shape that has sharp angles (Khler, 1929, 1947). In our study, 70 participants completed paired-associate memory tests after studying lists of three shape / non-word pairs. The sound-shape pairs that relied on known sound-symbolism links facilitated associative memory.

How much harder are hard garden-path sentences than easy ones?

Grusha Prasad
Johns Hopkins University, Baltimore, Maryland, United States

Tal Linzen
Johns Hopkins University, Baltimore, Maryland, United States

Abstract

The advent of broad-coverage computational models of human sentence processing has made it possible to derive quantitative predictions for empirical phenomena of longstanding interest in psycholinguistics; one such case is the disambiguation difficulty in temporarily ambiguous sentences (garden-path sentences). Adequate evaluation of the accuracy of such quantitative predictions requires going beyond the classic binary distinction between ”hard” and ”easy” garden path sentences and obtaining precise quantitative measurements of processing difficulty. We report on a self-paced reading study designed to estimate the magnitude of the disambiguation difficulty in two temporarily ambiguous sentence types (NP/Z and NP/S ambiguities). Disambiguation was more than twice as hard in NP/Z sentences as in NP/S sentences. This contrasts with the predictions of surprisal estimates derived from current broad-coverage language models, which lead us to expect a smaller difference between the two.

Proposing a Cognitive System for Universal Mental Spatial Transformations

Kai Preuss
Technical University Berlin, Berlin, Germany

Nele Russwinkel
Technische Universitt Berlin, Berlin, Germany

Abstract

Mental spatial transformation processes are often modeled by assuming imaginal processes, highly task-specific assumptions, or both. We propose the existence of a dedicated, unified cognitive system for the simulation of spatial processes, and show ways to model this system, including an ACT-R implementation that is currently in development. Results of spatial cognition and brain-imaging research support this proposal. Operations of this system are proposed to be influenced by their complexity, which we assume to be a product of the extent and amount of necessary transformation steps. This complexity is further assumed to be limited in its extent, possibly explaining decision time effects between task difficulties in a mental folding task as being caused by cognitive re-encoding processes. A model for the mental folding task lacking such a spatial system is presented, serving as a baseline to demonstrate the need of a system dedicated to mental transformations.

SpotLight on Dynamics of Individual Learning

Roussel Rahman
Rensselaer Polytechnic Institute, Troy, New York, United States

Wayne Gray
Rensselaer Polytechnic Institute, Troy, New York, United States

Abstract

How do individuals learn a complex task? Averaging performance over a group of individuals implicitly assumes that only one set of methods exists for accomplishing the task and that all learners acquire those methods in the same sequence. Rather than profiling a mythical "average subject", we focus on individuals using SpotLight – a tool for analyzing changes in individual performance as a complex task is learned. Specifically, we investigate 9 individuals who spent 31 hours learning the task of Space Fortress (SF). The SpotLight enables us to uncover the evolution of strategies and the iterative efforts of individuals to explore and devise new ways to improve performance. To our surprise, these players seem to have followed a common 'design for the weakest link' rule, in which after the current weakest link of performance is strengthened, an individual's attention turns to the next weakest link.

Cue Validity, Feature Salience, and the Development of Inductive Inference

Robert Ralston
The Ohio State University, Columbus, Ohio, United States

Vladimir Sloutsky
The Ohio State University, Columbus, Ohio, United States

Abstract

Young children can generalize properties to novel stimuli, but the mechanism underlying these early inductions is still debated. Some researchers argue that from an early age induction relies on category information and undergoes little development, while others believe that early induction is similarity-based, and the use of categories emerges over time. This present study brings new evidence to the debate by exploring the kinds of features 4-year-old children and adults (N = 123) rely on in their induction. Our results indicate that induction undergoes dramatic development: young children tend to rely on salient features when performing induction, whereas adults rely primarily on category information. We argue that the reported findings present evidence challenging category-based accounts of early induction, while supporting similarity-based accounts.

I Never Even Considered That!: Investigating explanations for adults failures to learn conjunctive causal rules

Alexandra Rett
UC San Diego, La Jolla, California, United States

Elizabeth Bonawitz
Rutgers University - Newark, Newark, New Jersey, United States

Koeun Choi
Virginia Tech, Blacksburg, Virginia, United States

Caren Walker
University of California San Diego, La Jolla, California, United States

Abstract

Despite having sophisticated causal reasoning skills, there are a variety of cases in which adult learners consistently ignore the available evidence and make an incorrect inference. Here, we focus on a specific case in which adults fail to infer and apply a conjunctive causal rule (Lucas et al., 2014), and examine two explanations for this failure. In Experiment 1, we manipulate information about the probabilistic nature of the events to test whether adults failure results from an endorsement of noisy relations. In Experiment 2, we manipulate the physical design of the causal system to test an alternative account: that this phenomenon is due to a failure to consider the correct, conjunctive hypothesis. Taken together, our results suggest that failures to learn the conjunctive rule may not be entirely due to a noisy prior that affords discounting of the evidence, but instead results from a failure to generate the relevant hypothesis.

Distinguishing the Phenomenal from the Cognitive: An Empirical Investigation into the Folk Concepts of Emotions

Kevin Reuter
University of Bern, Bern, Switzerland

Rodrigo Daz
University of Bern, Bern, Switzerland

Abstract

The two dominant theories on the nature of emotions are feeling theories (e.g., Prinz 2004) and cognitive theories (e.g., Lazarus 1991). The former take feelings to be the essential core aspect of emotions. The latter argue that emotions are based on judgements or some other conceptual states in order to account for the datum that emotions always seem to be directed towards events or objects. In this paper we argue that the controversy between feeling theories and cognitive theories rests on the false assumption that people do not distinguish emotional feelings from emotional judgements, i.e., that expressions of the form I feel x and I am x are largely intersubstitutable (Bennett & Hacker 2003). We present new empirical evidence from both corpus studies and a vignette study showing that feeling happy (sad/angry) and being happy (sad/angry) are two separate states that people are able to conceptually and linguistically distinguish.

You must know something I dont: risky behavior implies privileged information

Emory Richardson
Yale University, New Haven, Connecticut, United States

Julian Jara-Ettinger
Yale University, New Haven, Connecticut, United States

Abstract

People make sense of each others behavior by assuming that beliefs and desires vary across agents. We propose that people are more conservative when it comes to risk: when an agent takes an extreme risk, we assume they have privileged information rather than high risk tolerance. Participants watched an agent choose either to obtain three guaranteed tokens, or guess which box from a set had four tokens. After watching the agents choice, participants played the game themselves. In Study 1, participants were quicker to imitate an agent who immediately made extremely risky bets than one who started out making low-risk bets that became progressively riskier, suggesting that they inferred that risk-seeking agents were knowledgeable. In Study 2, participants ceased taking risky bets when the anonymous agent did, suggesting that participants choices depend on mental state inferences rather than contagious but mind-blind risk-seeking behavior.

Preparing not to Forget: Actions Take to Plan for Memory Error

Lorena Rosales
California Lutheran University, Thousand Oaks, California, United States

AndreaJ. Sell
California Lutheran University, Thousand Oaks, California, United States

Abstract

The present study was designed to examine actions people take in everyday life to prevent potential memory errors. Many past studies focus on the nature of forgetting, and additional studies have assessed cognitive interventions for those with varying degrees of impairment from aging or injury. However, there are a limited number of studies examining everyday remembering for healthy, functioning adults. In this study, across two experiments (n1=136; n2=85), participants completed a self-reported questionnaire regarding various types of daily prospective memory actions. We hypothesized that people would report using external memory aids (ex. technology) rather than internal aids (ex. mnemonics) and participants would report lower forget scores when using external aids. Results showed that participants overwhelmingly used external memory aids to prevent future memory errors for all tasks analyzed. Results also showed that levels of self-reported forgetting were not associated with particular types of preventative actions. Thus, the results imply that people tend to use what they perceive to work.

(A)symmetry (Non)monotonicity: Towards a Deeper Understanding of Key Cognitive Di/Trichotomies and the Common Model of Cognition

Paul Rosenbloom
University of Southern California, Los Angeles, California, United States

Abstract

Many dichotomies from across the cognitive sciences can be reduced to one of two fundamental distinctions (a)symmetry and (non)monotonicity of processing simplifying greatly the space of dichotomies needed to structure this broad interdisciplinary discipline. Taking the cross-product of these two dichotomies then yields a 2x2 structure of cells that in its turn yields a deeper understanding of two key trichotomies based on control and content hierarchies with each mapping to three out of the four cells. This cross-product and its four cells further provide a deeper understanding of the structure of the Common Model of Cognition an attempt to develop a community consensus concerning the processes and structures implicated in human-like minds as well as cognitive architectures that map onto it, such as ACT-R, Sigma and Soar and even AlphaZero with results that bear on the structure of integrative architectures, models and systems; and on their commonalities, differences and gaps.

Learning a novel rule-based conceptual system

Joshua Rule
MIT, Cambridge, Massachusetts, United States

Josh Tenenbaum
MIT, Cambridge, Massachusetts, United States

Steven Piantadosi
University of California, Berkeley, Berkeley, California, United States

Abstract

Humans have developed complex rule-based systems to explain and exploit the world around them. When a learner has already mastered a system's core dynamicsidentifying its primitives and their interrelationsfurther learning can be effectively modeled as discovering useful compositions of these primitives. It nevertheless remains unclear how the dynamics themselves might initially be acquired. Composing primitives is no longer a viable strategy, as the primitives themselves are what must be explained. To explore this problem, we introduce and assess a novel concept learning paradigm in which participants use a two-alternative forced-choice task to learn an unfamiliar rule-based conceptual system: the MUI system (Hofstadter, 1980). We show that participants reliably learn this system given a few dozen examples of the systems rules, leaving open the mechanism by which novel conceptual systems are acquired but providing a useful paradigm for further study.

Modeling Axonal Plasticity in Artificial Neural Networks

james ryland
University of Texas at Dallas, Richardson, Texas, United States

Abstract

Axonal growth and pruning is the brains primary method of controlling the structured sparsity of its neural circuits, as without long distance axon branches connecting distal neurons no direct communication is possible. Further, artificial neural networks have almost entirely ignored axonal growth and pruning instead relying on implicit assumptions that prioritize dendritic/synaptic learning above all other concerns. This project proposes a new model called the Axon Game, which allows the incorporation of biologically inspired axonal plasticity dynamics into most artificial neural network models with computational efficiency. We will explore the qualities of receptive windows grown under this methodology and discuss how they can integrate with neural network simulations.

How Productivity and Compositionality May Emerge from a Neural Dynamics of Perceptual Grounding

Daniel Sabinasz
Ruhr-Universitt Bochum, Bochum, Germany

Mathis Richter
Ruhr-Universitt Bochum, Bochum, Germany

Jonas Lins
Ruhr-Universitt Bochum, Bochum, North Rhine-Westphalia, Germany

Gregor Schner
Ruhr-Universitt Bochum, Bochum, Germany

Abstract

The productivity and compositionality of language and thought have often been taken as evidence that higher cognition is a form of information processing on systems of symbols with combinatorial syntax and semantics. We present a non-symbolic neural dynamic architecture that can ground combinatorial concepts in perception, i.e., establish a link between a combinatorial concept and an object in the perceptual array. The components of a combinatorial concept tree are sequentially grounded from the leaves to the root, while the output of each grounding step is passed on to the next grounding step by means of a mental map. This way, compositionality is an emergent property of the neural dynamics and does not require any form of symbolic information processing. We discuss how this process account contrasts with other neural accounts of compositionality and conclude with implications for the modeling of higher cognition.

Assessing the role of matching bias in reasoning with disjunctions

Mathias Sabl-Meyer
NeuroSpin center, CEA DRF/I2BM, INSERM, Universit Paris-Sud, Universit Paris-Saclay, 91191 Gif-Sur-Yvette, France

Salvador Mascarenhas
Ecole Normale Suprieure, Paris, France

Abstract

On mental models theories, reasoners create mental representations of information which they manipulate in order to derive new conclusions. These theories have been uniquely successful at explaining a class of attractive fallacies involving disjunctions. The original theories have appealed to low-level matching mechanisms (Walsh & Johnson-Laird, 2004; Koralus & Mascarenhas, 2013) to compare the models of the premises and the models of the conclusion and predict an answer. In three experiments, we show that the check for overlap in content involved in these accounts must take place at a high level of cognition in order to incorporate complex world knowledge. We introduce variants of illusory inferences from disjunction whose acceptance is accurately predicted by independant measures of confidence in causal connections. We conclude that the Revised Mental Model Theory of Khemlani et al. (2018) holds promise, but cannot account for our data out of the box.

On the purpose of ambiguous utterances

Gregory Scontras
UC Irvine, Irvine, California, United States

Asya Achimova
University of Tbingen, Tbingen, Germany

Christian Stegemann
University of Tbingen, Tbingen, Germany

Martin Butz
University of Tuebingen, Tuebingen, Germany

Abstract

Traditionally, linguists have treated ambiguity as a bug in the communication system, something to be avoided or explained away. More recent research has taken notice of the efficiency ambiguity affords us. The current work identifies an additional benefit of using ambiguous language: the extra information we gain from observing how our listeners resolve ambiguity. We propose that language users learn about each others private knowledge by observing how they resolve ambiguity. If language does not do the job of specifying the information necessary for full interpretation, then listeners are left to draw on their private knowledgeopinions, beliefs, and preferencesto fill in the gaps; by observing how listeners fill those gaps in, speakers learn about the private knowledge of their listeners. We implement this hypothesis as a computational model within the Rational Speech Act framework. We then test our hypothesis by using the model to predict behavioral data from naive participants.

A Smile Goes a Long Way: Exploring the Effect of Culture, Weather, and Connectedness on Smile Diffusion with an Agent-based Modell

Victoria Scotney
University of British Columbia, Kelowna, British Columbia, Canada

Fabian Cid Yanez
University of British Columbia, Kelowna, British Columbia, Canada

Joshua Cooper
University of British Columbia, Kelowna, British Columbia, Canada

Liane Gabora
University of British Columbia, Kelowna, British Columbia, Canada

Abstract

This paper first synthesizes research showing that (a) people reciprocate smiles, (b) smiling and being smiled at elevates mood, and (c) elevated mood is associated with proclivity to smile. Collectively, these findings suggest that smiling is contagious, i.e., smiles diffuse through a social network. The paper then presents experiments carried out to investigate how various factors affect the contagiousness of smiling using an agent-based model in which smiling affects a mood variable, which in turn affected proclivity to smile. The society consistently stabilized on a proportion of smilers, the magnitude of which was a function of social connectivity. Using previous data on the effect of weather and cultural differences on smile reciprocity, we simulated how these factors affect smile diffusion. Smile diffusion was greater in the sunny condition than the cloudy condition, and in the American condition than the Japanese condition, and both effects were magnified by increased social connectivity.

Learning and Production in the Explanation of Regularization Behaviour: a Computational Model

Chiara Semenzin
University of Edinburgh, Edinburgh, United Kingdom

Vanessa Ferdinand
University of Melbourne, Melbourne, VIC, Australia

Simon Kirby
The University of Edinburgh, Edinburgh, United Kingdom

Abstract

We propose a computational model to account for the regularization behaviour that characterizes language learning and that has emerged from experimental studies, specifically from concurrent multiple frequency learning tasks (Ferdinand, 2015). These experiments show that learners regularize the input frequencies they observe, suggesting that domain-general factors might underlie regularization behaviour. Standard models have failed to capture this pattern, so we explore the consequences of adding a production bias that follows the learning stage in a probabilistic model of frequency learning. We simulate and fit to experimental data a beta-binomial Bayesian sampler model, which allows an explicit quantification of both the learning and the production bias. Our results reveal that adding a production component to the model leads to a better fit to data. Given our results, we hypothesize that linguistic regularization may result from general-domain constraints on learning combined to biases in production, which need not to be considered innate.

An Associative Theory of Semantic Representation

Kevin Shabahang
University of Melbourne, Melbourne, VIC, Australia

Hyungwook Yim
The University of Melbourne, Melbourne, Australia

Simon Dennis
The University of Melbourne, Melbourne, VIC, Australia

Abstract

We present a new version of the Syntagmatic-Paradigmatic model (SP; Dennis, 2005) as a representational substrate for encoding meaning from textual input. We depart from the earlier SP model in three ways. Instead of two multi-trace memory stores, we adopt an auto-associative network. Instead of treating a sentence as the unit of representation, we go down a scale to the level of words. Finally, we specify all stages of processing within a single architecture. We show how the model is capable of forming representations of words that are independent of the surface-form through some question-answering examples. We end with a discussion of how the current model can provide a mechanistic account of elaborative and inferential processes during comprehension.

Associations versus Propositions in Memory for Sentences

Kevin Shabahang
University of Melbourne, Melbourne, VIC, Australia

Hyungwook Yim
The University of Melbourne, Melbourne, Australia

Simon Dennis
The University of Melbourne, Melbourne, VIC, Australia

Abstract

Propositional accounts of organization in memory have dominated theory in compositional semantics, but it is an open question whether their adoption has been necessitated by the data. We present data from a narrative comprehension experiment, designed to distinguish between a propositional account of semantic representation and an associative account based on the Syntagmatic-Paradigmatic (Dennis, 2005; SP) model. We manipulated expected propositional-interference by including distractor sentences that shared a verb with a target sentence. We manipulated paradigmatic-interference by including two distractor sentences, one of which contained a name from a target sentence. That is, we increased the second-order co-occurrence between a name in a target sentence and a distractor. Contrary to the propositional assumption, our results show that subjects are sensitive to second-order co-occurrence, hence favouring the associative account.

One-Object Decision-Making model: Fast and Frugal Heuristic for Human Activity Classification

karan sharma
Keysight Technologies, Atlanta, Georgia, United States

Suchendra Bhandarkar
The University of Georgia, Athens, Georgia, United States

Abstract

Consider an uncertain situation where an artificial intelligence (AI) system is called upon to determine a human action or activity in an image or scene. The AI system has not been previously trained to recognize any human action or activity, and has no prior information on pose, parts, spatial layout of the object in an image. In such a situation, what is the AI system supposed to do? Its options are limited, and it must determine the action or activity with the aid of the most probable inanimate object (other than the human actor) that it can detect in the image. The AI system needs to formulate two hypotheses to infer the action or activity in a zero-shot manner; first, that the most probable inanimate object detected in the image is one that is involved in the action or activity, and second, that the most likely action or activity associated with this object in the real world is the one actually occurring in the image. To what extent are these hypotheses valid? We propose that correct detection of the highly probable object and use of natural language word embeddings obtained via training on a general text corpus such as Wikipedia could enable the AI system to determine the underlying human action or activity in an image with reasonable classification accuracy. We conducted studies on the HICO dataset, which is a challenging dataset containing many rare human action/activity categories. Our experimental results show that if the AI system can reliably detect the most probable inanimate object in the image and then infer the corresponding verb in a zero-shot manner using language models trained on general text corpora, then it has a reasonable chance of correctly guessing the underlying action/activity in an image.

A CTA-DCD Model to Determine Design Requirements for Technology to Support People with Mild Cognitive Impairment / Dementia at Work

Karan Shastri
University of Waterloo, Waterloo, Ontario, Canada

Jennifer Boger
University of Waterloo, Waterloo, Ontario, Canada

Parminder Flora
Ontario Shores Centre for Mental Health Sciences, Whitby, Ontario, Canada

Arlene Astell
Ontario Shores Centre for Mental Health Sciences, Whitby, Ontario, Canada

Ann-Charlotte Nedlund
Linkoping University, Linkpoing, Sweden

Katja Karjalainen
University of Eastern Finland, Joensuu, Finland

Anna Mki-Petj-Leinonen
University of Eastern Finland, Joensuu, Finland

Louise Nygrd
Karolinska institutet, Huddinge, Sweden

Abstract

Work is an integral and meaningful part of many peoples lives. Research has shown that the consequences of MCI and dementia (MCI/dem) before the age of sixty-five can profoundly affect a persons vocational situation. Technology plays a significant role in supporting different abilities for people with MCI/dem at communities and home; however, there is little research to investigate the role of technology and address the technological requirements of people with MCI/dem at work who are employed. We propose a new systematic human factors model to study peoples tasks, activities, and requirements derived from in-depth interviews with six people living with MCI/dem and one caregiver. By characterizing the barriers or problems faced by people with MCI/dem in the context of cognitive work, we organized individual barriers of the participants in terms of macrocognitive activities and cognitive support requirements.

An Empirical investigation of JointSeparate Effect on Preference of Causal Explanation

Asaya SHIMOJO
Nagoya University, Nagoya city, Japan

Kazuhisa Miwa
Nagoya University, Nagoya-shi, Aichi-ken, Japan

Hitoshi Terai
Kindai University, Iizuka-shi, Fukuoka, Japan

Abstract

What makes an explanation better than another explanation? Previous studies have suspected that explanatory virtues, such as Simplicity and Scope, affect individuals' evaluation of the explanatory goodness. Although almost all of these studies have focused on the situation that some explanations are presented simultaneously, we do not always obtain some explanations in daily life. In this research, we conducted an experiment to investigate the preference change in causal explanation between Joint and Separate Evaluations. The results showed that Latent scope has a large effect as a criterion for evaluating explanatory goodness regardless of Joint and Separate Evaluation. Furthermore, Simplicity affects the evaluation of explanatory goodness differently between these situations of evaluation; however, the effect of comparison was observed only by online reflection in which evaluation is performed for two explanations simultaneously and not by offline reflection in which evaluation is re-performed after ending all evaluations.

Recombinant building: the ability to generate and recombine navigation structures is difficult to acquire through just reinforcement learning

Ganesh Shinde
Centre for Modeling and Simulation,Savitribai Phule Pune University , Pune, Maharashtra, India

Harshit Agrawal
Homi Bhabha Centre for Science Education, TIFR, Mumbai, Maharashtra, India

Sanjay Chandrasekharan
Homi Bhabha Centre for Science Education, TIFR, Mumbai, India

Abstract

Humans build novel tools, external knowledge structures (markers, maps etc.), and internal structures (analogies, mental models etc.) to facilitate cognition. Humans also recombine these building strategies to suit any task. Other organisms generate such structures as well, but they use them to optimize single tasks. This suggests that the human species' cognitive advantage stems from the capability to recombine built structures, and the resulting extended mind. Chandrasekharan & Stewart (2007) hypothesized that this capacity could emerge from reinforcement learning. We tested this proposal, by studying three foraging models, which examined whether novel recombinations of building (external and internal navigation structures) emerged in reactive agents, from just reinforcement learning. Results showed that recombination does not emerge with just reinforcement. This was because the building of external structures provided a very high reward profile, including free riding, thus acting as an attractor, blocking the recombination strategy. We discuss the implications of these results.

Can a forward posture enhance willingness to change ones own attitude in decision making? Nudging with embodied cognition approach

Masaru Shirasuna
The University of Tokyo, Tokyo, Japan

Hidehito Honda
Yasuda Womens University, Hiroshima, Japan

Kazuhiro Ueda
The University of Tokyo, Tokyo, Japan

Abstract

Recently, nudging approaches wherein peoples decisions are altered in a predictable direction have attracted attention. Conversely, many embodied cognition approaches that relate peoples mind with their body have been studied in cognitive science. Based on these approaches, we investigated whether a forward posture (defined by leaning forward in a chair) generated by the environment can enhance a particular decision. We also evaluated the types of decisions that are likely to be enhanced by the forward posture. Behavioral experiments via a forward or normal chair where the seat allows little or no lean revealed that a forward posture can affect the decision making, particularly participants willingness to change their own attitude. We discuss the possible applications of leading predictable decisions from the environment and setting the decision environment in the real world.

Real-time inference of physical properties in dynamic scenes

Kevin Smith
Massachusetts Institute of Technology, Cambridge, Massachusetts, United States

Mario Belledonne
MIT, Cambridge, Massachusetts, United States

Ilker Yildirim
Massachusetts Institute of Technology, Cambridge, Massachusetts, United States

Jiajun Wu
MIT, Cambridge, Massachusetts, United States

Josh Tenenbaum
MIT, Cambridge, Massachusetts, United States

Abstract

Human scene understanding involves not just localizing objects, but also inferring the latent causal properties that give rise to the scene for instance, how heavy those objects are. These properties can be guessed based on visual features (e.g., material texture), but we can also infer them from how they impact the dynamics of the scene. Furthermore, these inferences are performed rapidly in response to dynamic, ongoing information. Here we propose a computational framework for understanding these inferences, and three models that instantiate this framework. We compare these models to the evolution of human beliefs about object masses. We find that while peoples judgments are generally consistent with Bayesian inference over these latent parameters, the models that best explain human judgments are approximations to this inference that hold and dynamically update beliefs. An earlier version of this work was published in the proceedings of CCN 2018 at https://ccneuro.org/2018/proceedings/1091.pdf.

Cognitively-Inspired Salience Computation for Intelligent Agents

Sterling Somers
Carnegie Mellon University, Pittsburgh, Pennsylvania, United States

Konstantinos Mitsopoulos
Carnegie Mellon University, Pittsburgh, Pennsylvania, United States

Christian Lebiere
Carnegie Mellong University, Pittsburgh, Pennsylvania, United States

Robert Thomson
United States Military Academy , West Point, New York, United States

Abstract

We describe a method for determining feature salience of action decisions in intelligent agents based on cognitively-inspired salience. Salience is defined as the degree of influence that a factor has on a given decision. This is generated by having a cognitive model using instance-based learning theory to mirror the actions of an intelligent agent, and then determining which features most uniquely contributed to the actions of the agent. We present three examples of this salience techniques, including reinforcement learning agents based in the StarCraft II and autonomous drone domains, as well as part of a risk assessment model. A benefit of our method is that it does not rely on a specific implementation of an agent, it only requires the underlying decision feature-space. It is also capable of utilizing features at different levels of abstraction

An Attractor Neural-network Simulation of Decision Making

Ashley Stendel
McGill University, Montreal, Quebec, Canada

Thomas Shultz
McGill University, Montreal, Quebec, Canada

Abstract

We apply an attractor neural-network model to experiments on monkeys who decided which direction tokens are moving, while firing rates of large numbers of neurons in premotor cortex are being recorded. Using pools of artificial excitatory and inhibitory neurons, our network model accurately simulates the neural activity and decision behavior of the monkeys. Among the simulated phenomena are decision time and accuracy, commitment, patterns of neural activity in trials of varying difficulty, and an urgency signal that builds over time and resets at the moment of decision. Predictive simulations of decision change are also presented, suggesting gradual passing through an uncertain region on the way to a new decision. The model shows that committed decisions need not involve any explicit threshold detection mechanism. Instead, competition, suppression, decision, and commitment naturally emerge from the dynamics of the system.

A Cognitive Modeling Approach for Predicting Behavioral and Physiological Workload Indicators

Christopher Stevens
Air Force Research Laboratory , Dayton, Ohio, United States

Megan Morris
Ball Aerospace, Fairborn, Ohio, United States

Christopher Fisher
Air Force Research Laboratory, Dayton, Ohio, United States

Christopher Myers
Air Force Research Laboratory, WPAFB, Ohio, United States

Abstract

Measuring cognitive workload is a persistent challenge in cog-nitive science. Cognitive architectures may offer a principledway to measure, define, and understand workload and its be-havioral and physiological consequences in terms of under-lying cognitive dynamics. Previous research has shown thatmodel-based workload relates to subjective workload judg-ments in simple tasks. Our goal was to further validate model-based workload measurement with known physiological work-load indicators in a complex task characterized by varying de-grees of workload levels. Participants completed an unmannedvehicle management task while their physiology was recorded.Correlations between model-based workload and physiologi-cal metrics generally trended in the predicted direction, andthe engagement index showed the strongest and most consis-tent relationship to model workload. The results provide pre-liminary validation for model-based workload measurement.

A Geometric Interpretation of Feedback Alignment

Andreas Stckel
University of Waterloo, Waterloo, Ontario, Canada

Terrence Stewart
University of Waterloo, Waterloo, Ontario, Canada

Chris Eliasmith
University of Waterloo, Waterloo, Ontario, Canada

Abstract

Feedback alignment has been proposed as a biologically plausible alternative to error backpropagation in multi-layer perceptrons. However, feedback alignment currently has not been demonstrated to scale beyond relatively shallow network topologies, or to solve cognitively interesting tasks such as high-resolution image classification. In this paper, we provide an overview of feedback alignment and review suggested mappings of feedback alignment onto biological neural networks. We then discuss a novel geometric interpretation of the feedback alignment algorithm that can be used to analyze its limitations. Finally, we discuss a series of experiments in which we compare the performance of backpropagation and feedback alignment. We hope that these insights can be used to systematically improve feedback alignment under biological constraints, which may allow us to build better models of learning in cognitive systems.

A case study of formation of an art concept by a contemporary artist: Analysis of the utilization of drawing in the early phase

Kikuko Takagi
The University of Tokyo, Tokyo, Japan

Takeshi Okada
The University of Tokyo, Tokyo, Japan

Sawako Yokochi
Tokyo Future University, Tokyo, Japan

Abstract

When producing a new series of artworks, an artist may engage in a variety of activities in the formation of an art concept. In a specific instance, a contemporary artist was demonstrated first to draw his ideas on paper, as an initial phase of developing his art concept. This paper utilizes data from a previous study to analyze the drawings and interviews conducted during this drawing phase. The results show that the artist used various types of modification of his art-making process. By changing his own creative activity, the artist often reflected upon his creative process, asking himself what he really wanted to do, and explored new images in response to unexpected findings and the feeling of confusion at his own drawings.

What Factors of Background Music Disrupt Task Performance? Influence of Types of Sound, Tasks, and Working Memory Capacity on Irrelevant Sound/Speech Effect

Maiko Takahashi
University of Tokyo, Tokyo, Japan

Mika Ishikawa
Nagoya University, Nagoya, Japan

Sachiko Kiyokawa
Nagoya University, Nagoya, Japan

Abstract

Task-irrelevant background speech or sounds are known to have detrimental effects on task performance which are called irrelevant-speech/sound effects (ISEs). In this study, we have investigated the contributing factor responsible for magnitude of ISE focusing on the meaningfulness of the background noise and working memory capacity (WMC). Participants were asked to perform reading comprehension task (Exp. 1), serial recall task (Exp 2), and match-to-sample task (Exp.3) with or without task-irrelevant instrumental music and lyrics, and their WMC was measured with the Reading Span Test. The results revealed that the irrelevant sounds with lyrics, but not instrumental music disrupted the performance of the participants in both the reading comprehension and serial recall tasks , while that in match-to-sample task was not interfered by either sound types. The moderating effect of WMC was not observed in any experiments. The results implied that ISEs were observed when phonological loop was used to conduct these tasks. Based on these results, the function of a learners WMC in the ISE is discussed.

What strategies do adults use to solve fraction arithmetic problems?

Shawn Tan
Carleton University, Ottawa, Ontario, Canada

Jo-Anne LeFevre
Carleton University, Ottawa, Ontario, Canada

Abstract

When children perform fraction arithmetic, they generate a variety of solutions. In this study, we extended this research to adults. We report that adults performance is best for addition and subtraction, worse for division, and is susceptible to the same kinds of strategy errors observed in 6th grade children. Specifically, solvers common strategy errors involved maintaining the values of fractions with common denominators even when that strategy was not appropriate. We also present two other findings that were not observed in children. First, adults applied an incorrect division algorithm; they incorrectly inverted the first, rather than the second operand in fraction division problems. Second, adults applied reduction procedures for fraction multiplication and division in order to simplify numerator-denominator pairs during fraction arithmetic. Our results suggest that strategy selection was cued by identifying common fraction components within problems.

Cognitive Complexity of Logical Reasoning in Games: Automated Theorem Proving Perspective

Katrine Thoft
Technical University of Denmark, Lyngby, Denmark

Nina Gierasimczuk
Technical University of Denmark, Lyngby, Denmark

Abstract

We use formal proof techniques from artificial intelligence and mathematical logic to analyse human reasoning in problem solving. We focus on the Deductive Mastermind game, as implemented in the Dutch massive online learning system for children, Math Garden. The game is formalised in propositional logic and the game-playing procedure is given a form of a logical proof. We use Resolution and Natural Deduction proof methods (implemented in JAVA). The difficulty of a particular logical reasoning step is associated with the computationally obtained parameters of the proofs, which are compared with each other, and against the empirical difficulty of the game. We show, among others, that the complexity parameters derived from Natural Deduction agree with the Analytical Tableaux parameters, and with the empirical difficulty as experienced by human subjects.

(Nina Gierasimczuk is supported by NCN OPUS Grant 2015/19/B/HS1/03292.)

Estimating Average Body Size of Sets of Bodies

Michelle To
Lancaster University, Lancaster, Lancashire, United Kingdom

James Brand
University of Canterbury, Christchurch, New Zealand

Georgia Hampton
Lancaster University, Lancaster, United Kingdom

Martin Tovee
Lincoln University, Lincoln, Lincolnshire, United Kingdom

Abstract

In two behavioral experiments, we demonstrated that human observers can extract average body size from a group of individuals. In Experiment 1, we asked 38 participants to estimate the average body size from a group of 5, 10 or 15 bodies that were presented in various angles of view (Profile, Three-Quarter, Frontal, and Mixed). Participants were able to extract the average body size, but they systematically overestimated thinner body groups, and underestimated larger body groups. Biases were generally reduced for smaller sets sizes and when bodies were shown in profile view, but the trend was reversed for sets with larger bodies. In Experiment 2, we tested 37 participants and showed that the accuracy of their estimates was modulated by presentation time: Accuracy was poorest when groups were presented for 1s, but significantly improved for 3s and 5s presentations. Implications of these finding are discussed.

Be timely: when gaps are more than symptoms

John Tomlinson Jr
Leibniz Centre for General Linguistics, Berlin, Germany

Abstract

Recently, turn-taking gaps, or unfilled pauses, have been viewed as a symptom or by-product of predictive planning mechanisms in speech production (Levinson & Torreria, 2015). Other works has shown that gaps can take signaling functions and that this is governed by politeness (Bgels, Kendrick, & Levinson, 2015). Two mouse-tracking experiments examined when gaps are interpreted as a symptom of processing or as a signal. This was tested by examining how gaps are interpreted in tandem to scalar implicatures (Bonneferon, Dahl, & Holtgraves, 2015). Experiment 1 found that longer gaps slightly reduce implicature rates at longer gaps and these longer gaps do not lead to faster implicature responses. Experiment 2 found that filled and unfilled pauses (gaps) both signal hesitation, though filled pauses signaled hesitation at short gaps. Overall, these experiments show that gaps lengths can have signaling functions beyond politeness and question bias.

Sub-morphemic form-meaning systematicity: the impact of onset phones on word concreteness

Sean Trott
UC San Diego, San Diego, California, United States

Arturs Semenuks
University of California, San Diego, San Diego, California, United States

Benjamin Bergen
UC San Diego, La Jolla, California, United States

Abstract

Do individual sounds carry meaning? The relationship between sound and meaning in human languages is typically assumed to be arbitrary, though recent research provides evidence for the existence of both iconicity and systematicity between word forms and their meaning. However, this research has not asked whether individual sounds in a language covary in systematic ways with aspects of meaning. In two analyses, we find evidence for more systematicity between the initial phones of words and those words concreteness ratings than one would expect in a truly arbitrary lexicon. This suggests that initial phones may act as cues to aspects of word meaning, and raises questions about whether language learners detect and exploit these cues.

The Scaffolding of Inferential Reasoning: Intuitive Analysis of Variance

David Trumpower
University of Ottawa, Ottawa, Ontario, Canada

Nicolas Robinson
University of Ottawa, Ottawa, Ontario, Canada

Abstract

In the present study, we explored the effect of a scaffolding exercise designed to make salient the importance of within-group variance on participants informal reasoning during a subsequent intuitive analysis of variance task. Participants were first presented with several datasets that varied with respect to within-group differences and were asked to provide examples of extraneous factors that could be the source of the variance. Afterwards, participants were given additional datasets that differed with respect to both within and/or between-group variability, and were asked to rate the strength of evidence provided by the dataset in support of a hypothetical theory. Consistent with prior research, the majority of participants tended to place a strong emphasis on between-group variability while minimizing the importance of within-group variation. However, the results indicate that for a subset of participants (n=6), the scaffolding exercise was effective in highlighting the significance of within-group variation. We found that all participants who reasoned normatively on the scaffolding exercise were able to successfully complete the analysis of variance task in a normative manner.

Group Discussion Clarifies the Difference between Maximin and Equality Principles in Social Distribution for Others

Atsushi Ueshima
The University of Tokyo, Bunkyo, Tokyo, Japan

Tatsuya Kameda
The University of Tokyo, Bunkyo, Tokyo, Japan

Abstract

The allocation of scarce resources is a ubiquitous process in human societies, yet it is challenging to aggregate peoples diverse distributive viewpoints into group consensus. We investigate whether such heterogeneity in preferences may be reduced when people participate in group discussion in a distribution task. In two interactive experiments, we found that after group discussion, participants became less inequity-averse and preferred the maximin allocation. Analyses of participants conversations and information-search behaviors showed that such shifts toward the maximin allocation were facilitated by a strong concern for the worst-off recipient during group discussion. These results suggest that a maximin concern exhibited in discussion helped participants to understand the difference between the inequity-aversion principle and the maximin principle, which are often confounded in individual judgments. These results provide empirical insight into how social interaction can help to aggregate peoples diverse distributive preference into a social consensus.

The Role of Sensorimotor and Linguistic Information in the Basic-Level advantage

Rens van Hoef
Lancaster University, Lancaster, United Kingdom

Louise Connell
University of Lancaster, Lancaster, United Kingdom

Dermot Lynott
Lancaster University, Lancaster, United Kingdom

Abstract

The basic-level advantage is one of the best-known effects in human categorisation. Traditional accounts argue that basic-level categories present a maximally informative or entry-level into a taxonomic organisation of concepts in semantic memory. However, these explanations are not fully compatible with most recent views on the structure of the conceptual system, which emphasise the role of sensorimotor (i.e., perception-action experience of the world) and linguistic information (i.e., statistical distribution of words in language) in conceptual processing. In a pre-registered wordpicture categorisation study, we hypothesised that our novel measures of sensorimotor and linguistic distance would contribute to categorical decision making, and would outperform traditional taxonomic levels (i.e., subordinate, basic, superordinate) in predicting the basic-level advantage. Results showed that, overall, our measures predicted the basic-level advantage at least as well as taxonomic level. Sensorimotor information best explained processing speed, whereas taxonomic level best explained participants choices.

Analyzing Performance Differences in Artists and Engineers- An RPM Study

sravya vatsavayi
International Institute of Information and Technology, Hyderabad, Telengana, India

Priyanka Srivastava
International Institute of Information Technology, Hyd, Hyderabad, Telangana, India

Kavita Vemuri
International Institute of Information Technology - Hyderabad, Hyderabad, Telangana, India

Abstract

Analytic reasoning differences, as gauged from intelligence metrics, in students engaged in streams requiring a predominantly divergent (arts) or convergent thinking (science and engineering) is a topic of interest. In this paper we have examined this difference by a modified sequence of two sections (D & E) of the Standard Ravens Progressive matrices (RPM). The scan path gaze behavior was analyzed with an eye tracker. The 30 engineering students (half of them are also trained in fine arts) scored higher than the 15 fine arts students. In the former cohort, the artistic and the non-artistic set show no difference in performance but the scan path, fixation count and time taken indicate possible differences in visual strategies for pattern identification. From the detailed analysis, we argue that intelligence as measured by RPM is enhanced by training in reasoning and logic as in engineering streams and might not reflect an innate ability.

Understanding the design neurocognition of industrial designers when designing and problem-solving.

Sonia Vieira
Faculty of Engineering University of Porto, Porto, Porto, Portugal

John Gero
UNCC, Charlotte, North Carolina, United States

Jessica Delmoral
University of Porto, Porto, Portugal

Valentin Gattol
AIT Austrian Institute of Technology GmbH, Vienna, Austria

Carlos Fernandes
University of Porto, Porto, Portugal

Marco Parente
University of Porto, Porto, Portugal

Antnio Fernandes
University of Porto, Porto, Portugal

Abstract

This paper presents results from an experiment to determine brain activation differences between problem-solving and designing of industrial designers. The study adopted and extended the tasks described in a previous fMRI study of design cognition and measured brain activation using EEG. The experiment consists of 4 tasks: problem-solving, basic design and open design tasks using a tangible interface and sketching. By taking advantage of EEG's temporal resolution we focus on time-related neural responses during problem-solving compared to design tasks. Statistical analyses indicate increased activation when designing compared to problem-solving. Results of time-related neural responses connected to Brodmann areas cognitive functions, contribute to a better understanding of industrial designers' cognition. The study is part of a research project whose goal is to correlate design cognition with brain behavior across design domains. Bringing neuroscience methods to design research is contributing to a better understanding of the emergent field of design neurocognition.

Social Learning and Decisional Constraints in Uncertain Environments

Marius Vollberg
Harvard University, Cambridge, Massachusetts, United States

Matthias Hofer
MIT, Cambridge, Massachusetts, United States

Mina Cikara
Harvard University, Cambridge, Massachusetts, United States

Abstract

The ability to learn from others is central to our species. At the same time, we are more than able to independently learn from our own experience. Investigating how these pathways function in concert, past research has looked at how we integrate what can be learned from others with our own observations. To do so, social information is typically operationalized as observed behavior. However, social information often comes in the form of normative advice. Humans have been shown to value decisional freedom and reject constraints to it. Some forms of social information, such as normative advice, plausibly comprise potential for both social learning and perceived constraint. Past research on decisional constraints posed by social information has been of limited granularity. We present an experimental framework to study behavior in the face of normative social information and explore data from two experiments using computational modeling.

The Temporal Dynamics of Belief-based Updating of Epistemic Trust: Light at the End of the Tunnel?

Momme von Sydow
LMU Munich, Munich, Bavaria, Germany

Christoph Merdes
Universitt Erlangen, Erlangen, Bavaria, Germany

Ulrike Hahn
Birkbeck, University of London, London, London, United Kingdom

Abstract

We start with the distinction of outcome- and belief-based Bayesian models of the sequential update of agents beliefs and subjective reliability of sources (trust). We then focus on discussing the influential Bayesian model of belief-based trust update by Eric Olsson, which models dichotomic events and explicitly represents anti-reliability. After sketching some disastrous recent results for this perhaps most promising model of belief update, we show new simulation results for the temporal dynamics of learning belief with and without trust update and with and without communication. The results seem to shed at least a somewhat more positive light on the communicating-and-trust-updating agents. This may be a light at the end of the tunnel of belief-based models of trust updating, but the interpretation of the clear findings is much less clear.

Foundations of search behavior, beyond the exploration-exploitation trade-off

oana vuculescu
Aarhus University, Aarhus, Denmark

Carsten Bergenholtz
Aarhus University, Aarhus, Denmark

Ali Amidi
Aarhus University, Aarhus, Denmark

Abstract

We investigate the cognitive micro-foundations of individual search. The aim of this study is to identify important cognitive antecedents of the heterogeneity of individual level search behavior. We introduce a problem-solving task that not only requires a binary trade-off between either exploration or exploitation, but solicits the individual to understand the underlying problem structure in order to be able to optimize the search. Combining data collected from individuals solving this experimental task (N = 407) with a quantitative survey of cognitive styles as well as a neuropsychological test of cognitive ability (g-factor) we explain how different cognitive micro-foundations translate into substantial variation in search behaviors.

Successes of the Intuitive Psychologist: Observers make reasonable judgments in the role conferred advantage paradigm

Drew Walker
UCSD, San Diego, California, United States

Nicholas Christenfeld
University of California, San Diego, La Jolla, California, United States

Ed Vul
University of California, San Diego, La Jolla, California, United States

Abstract

In a now classic experiment Ross, Amabile & Steinmetz (1977) showed that observers think that a participant who is randomly assigned to invent questions has more general knowledge than a participant assigned to answer these questions. This is taken to be an error arising from a reasoning process in which observers ignore social roles, and instead rely on surface behavior to make social judgments. Here we test two potential explanations for this observation: (1) observers are using a flawed reasoning process in which they do not consider the advantages and disadvantages that different social roles may confer, or (2) observers are using an unbiased reasoning process in which they do consider the influence of social role, but they are simply operating with an imperfect estimate of the advantage afforded the questioner. In a series of five studies, we show that not only is reasoning in this task consistent with an unbiased inference account, but, that observers are also surprisingly well calibrated to the influence of the social roles used in this paradigm.

Evidence for constructive influences from simple evaluations

Lee White
City, University of London, London, London, United Kingdom

Emmanuel Pothos
City, University of London, London, United Kingdom

Michael Jarrett
INSEAD, Singapore, Singapore

Abstract

There have been several demonstrations of constructive influences from choice paradigms, for example, when a decision maker has to commit to one of the available options and abandon the rest. In such cases, an expectation of constructive influences, whereby the preference for the chosen option increases, while the preference for the abandoned ones decreases, is perhaps reasonable (e.g., as a way to reduce cognitive dissonance). However, this reasoning is harder to translate to situations such that there is a simple evaluation. We employ an organizational questionnaire to show that a simple evaluation of an earlier statement can lead to systematic influences on a later one. Our results generalize our understanding for when constructive influences may occur. We outline a technical framework for predicting this bias (which we label evaluation bias), based on quantum theory. Quantum theory is an appropriate framework for modelling constructive influences, because the theory involves a fundamental process of state change when a measurement (evaluation) is made.

Testing Gender Markedness of Nouns with Self a Paced Reading Study

Ethan Wilcox
Harvard University, Cambridge, Massachusetts, United States

Abstract

Some English nouns occur in gender-marked pairs, which fall into two classes: In the Superordinate class, the unmarked (masculine) form is available to refer to female referents ("Allison Janey is a good actor"), whereas in the specific class it is not (*"Diana is a good prince"). Two theories account for this alternation: The Featural Theory proposes that the unmarked are unspecified for gender features. The second, Frequency Theory proposes relative frequency of the marked vs. unmarked forms are responsible (Haspelmath, 2006). This work provides evidence against the frequency theory by employing a self-paced reading study that tests relative processing times of anaphoric pronouns referring to gendered nouns. If noun pairs are split along Specific/Superordinate class lines, a processing slowdown is found for processing processing pronoun gender mismatches, except for nouns like 'actor', as expected. However, when the noun pairs are split by relative frequency the effect disappears.

Do typically and atypically developing children learn and generalize novel names similarly: the role of conceptual distance during learning and at test

Arnaud Witt
University of Bourgogne Franche-Comt, Dijon, France

Annick Comblain
University of Liege, Liege, Belgium

Jean-Pierre Thibaut
University of Bourgogne Franche-Comt, Dijon, Bourgogne, France

Abstract

There is a large body of evidence showing that comparison leads to better conceptualization and generalization of novel names than no-comparison settings in typically developing (TD) children (e.g., Gentner, 2010). So far, comparison situations have not been studied with children with intellectual disabilities (ID) (Chapman & Kay-Raining Bird, 2012). In the present research children with ID and TD children matched on mental age with the Ravens coloured progressive matrices RCPM (Raven, 1965) were tested in several comparison conditions. We manipulated the conceptual distance between stimuli in the learning phase and between the learning phase stimuli and the generalization phase stimuli for object and relational nouns. Results showed that overall both populations had rather similar performance profile when matched on their cognitive skills (low vs. high functioning). Unexpectedly, ID childrens performance was equivalent or better than their TD peers. We discuss our results in terms of the role of conceptual distance on participants conceptual generalization as a function of their intellectual abilities and cognitive functioning.

Surprisingly unsurprising! Infants looks to probable vs. improbable events is modulated by others expressions of surprise

Yang Wu
Stanford University, Stanford, California, United States

Hyowon Gweon
Stanford University, Stanford, California, United States

Abstract

Research in diverse disciplines suggests that agents own prediction errors enhance their learning. Yet, human learners also possess powerful capacities to learn from others. Here we ask whether infants can use others expressions of surprise as vicarious prediction error signals to infer hidden states of the world. First, we conceptually replicated Xu & Garcia (2008), showing that infants (12.0-17.9 months) looked longer at improbable than probable sampling outcomes (Experiment 1). Then we added emotional cues to the design (Experiment 2). Before revealing an outcome to an infant, the experimenter looked at the outcome and expressed either happiness or surprise. While infants still looked longer at the improbable than the probable outcome following the experimenters happy expression, this trend was reversed when the experimenter had expressed surprise at the outcome. Such early-emerging ability to use others surprise as vicarious prediction error may guide infants own learning about the world. Preprint:https://psyarxiv.com/8whuv

Revealing Long-term Language Change with Subword-incorporated Word Embedding Models

Yang Xu
San Diego State University, San Diego, California, United States

Jiasheng Zhang
The Pennsylvania State University, State College, Pennsylvania, United States

David Reitter
Penn State, University Park, Pennsylvania, United States

Abstract

We propose an augmented word embedding model that better incorporates subword information with additional parameters that characterize the semantic weights of characters in composing words. Our model can reveal some interesting patterns of long-term change in Chinese language, which provides novel evidence and methodology that enriches existing theories in evolutionary linguistics. The resulting word vectors also has decent performance in NLP-related tasks.

Demonstrative This and Hand Pointing Can Promote Socio-Centric Interpretations About Invisible Objects

Tetsuya Yasuda
Tokyo Denki University, Saitama prefecture, Japan

Kei Kashiwadate
Graduate school of TokyoDenki University, Saitama prefecture, Japan

Harumi Kobayashi
Tokyo Denki University, Saitama prefecture, Japan

Abstract

Conveying referential intention is essentially important to cooperate with others. It is reported that even adults sometimes take ego-centric perspective (i.e., perspective that is based on one's own perspective ignoring other's perspective) in comprehending others utterances. In the present study we used a modified version of Keysars paradigm of 4x4 grid, and examined whether the interpretation of the instruction by the addressee was affected by the directors use of two social-pragmatics aspects; demonstratives and gestures. Results showed if the director did not use a demonstrative and hand pointing, the addressees interpreted the object from ego-centric perspective. In contrast, if the director used a demonstrative and hand pointing, the addressees correctly interpreted the referred object showing their use of the directors perspective. The result suggested that demonstratives and hand pointing may promote the addressees interpretation based on the directors perspectives.

Can Paradigmatic Relations be Learned Implicitly?

Hyungwook Yim
The University of Melbourne, Melbourne, Australia

Olivera Savic
The Ohio State University, Columbus, Ohio, United States

Layla Unger
Ohio State University, Columbus, Ohio, United States

Vladimir Sloutsky
The Ohio State University, Columbus, Ohio, United States

Simon Dennis
The University of Melbourne, Melbourne, VIC, Australia

Abstract

A wealth of statistical learning research has provided evidence that regularities in which items co-occur (referred to here as syntagmatic) can be learned implicitly. However, it is not known whether higher-order relations can also be learned implicitly. Here we present two experiments that investigate whether regularities, where items do not co-occur but instead share co-occurrence with each other (referred to here as paradigmatic), can be learned implicitly. In Experiment 1, we used a traditional auditory statistical learning paradigm where participants passively listened to an auditory stream containing syntagmatic and paradigmatic regularities and found evidence only of syntagmatic learning. In Experiment 2, we instructed participants to attend to items during the training session and found evidence of learning paradigmatic relations in participants who demonstrated high-level of syntagmatic learning. The results are discussed in terms of the limits of implicit learning and the role of attentional mechanisms in learning higher-order statistical regularities.

Understanding Human Memory for Where Using Experience Sampling Data

Hyungwook Yim
The University of Melbourne, Melbourne, Australia

Bree Wan Rong Ong
The University of Melbourne, Melbourne, VIC, Australia

Benjamin Stone
The University of Melbourne, Melbourne, VIC, Australia

Simon Dennis
The University of Melbourne, Melbourne, VIC, Australia

Abstract

We examined how people remember 'where' a certain event happened given the time and date of the event (i.e., memory for where). We especially focused on the kinds of information people use when trying to retrieve their memory for where. In order to increase ecological validity, we used experience sampling technology. In the task, participants watched a video that depicted a 3rd person's life for a month period, which was generated by using the 3rd person's experience sampling data. Then, participants were cued with a certain time and were asked where the person was at that time as well as how confident they were with their response. Using a conditional logit model, we found that, temporal and spatial distances were the main predictors of participants' choice. We also found that generic knowledge about one's life and repeating events (or locations) also affect participants retrieval of memory for where.

Exploring How People Use Star Rating Distributions

Jingqi Yu
Indiana University Bloomington, Bloomington, Indiana, United States

David Landy
Indiana University, Bloomington, Bloomington, Indiana, United States

Abstract

When purchasing products online, often two products may have similar mean ratings and numbers of reviews, but such apparent similarities may hide important differences. Sometimes, the distribution of star ratings is also available to decision makers in addition to these two attributes. Will the decision still be as undifferentiated as before or will the distributions of stars engender a preference towards one of the products? To answer this question, the current study manipulated the displayed variability of ratings for choices with the same average rating. The behavioral studies showed that participants exhibited distinctive choice patterns when the distribution of ratings was provided even when the average rating and total number of reviews were the same between two compared products. A utility-based cognitive model was therefore developed to identify the underlying mechanism as to why people chose the way they did.

Neural Network Modeling of Learning to Actively Learn

Lie Yu
McGill University, Montreal, Quebec, Canada

Ardavan S. Nobandegani
McGill University, Montreal, Quebec, Canada

Thomas Shultz
McGill University, Montreal, Quebec, Canada

Abstract

Humans are not mere observers, passively receiving the information provided by their environment; they deliberately engage with their environment, actively participating in the information acquisition stage to improve their learning performance. Despite being a hallmark of human cognition, the computational underpinnings of this active (or self-directed) mode of learning have remained largely unexplored. Drawing on recent advances in machine learning, we present a neural-network model simulating the process of learning how to actively learn. To our knowledge, our work is the first neural-network model of learning to actively learn. Extensive simulations demonstrate the efficacy of our model, particularly in handling high dimensional domains. Notably, our work serves as the first computational account of the recent experimental finding by MacDonald and Frank (2016) showing that prior passive learning improves subsequent active learning. Our work exemplifies how a synergistic interaction between machine learning and cognitive science helps develop effective, human-like artificial intelligence.

Lexical diversity and language development

Yawen Yu
University of California, Los Angeles, LOS ANGELES, California, United States

Dan Yurovsky
University of Chicago, Chicago, Illinois, United States

Abstract

Previous research has demonstrated a relationship between quantity of language input and childrens rate of language development: Children who hear more words learn faster. This work takes on two mutually-constraining questions: (1) How should we define quality, and (2) what is the relationship between input quality and language development? We analyzed a longitudinal corpus of interactions between 50 children and their parents using four measures of lexical diversity: Type Token Ratio (TTR), Moving Average TTR, and two more recent measuresvocd-D and MTLD. We found that only MTLD gave a prima-facie correct characterization of childrens development, and parents MTLD was correlated with childrens over development. Results of simulations showed that MTLD was distinct from the other measures in its sensitivity to both lexical diversity and word order, suggesting that quality should be defined not just by diversity of words, but also by the variability of sentence structures in which they occur.

Chinese Children Learning Higher-Order Generalizations through Free Play: The Influence of Parenting Style

Li Zhao
Department of Psychology, Beijing Forestry University, Beijing, Beijing, China

Zi L. Sim
Department of Psychology, University of California, Berkeley, Berkeley, California, United States

Mingyi Wang
Department of Psychology, Beijing Forestry University, Beijing, Beijing, China

Fei Xu
Department of Psychology, University of California, Berkeley, Berkeley, California, United States

Abstract

Rational constructivism believes children are active learners, they are able to learn causal rules through free play. Empirical evidence has demonstrated that 2- and 3-year-old children successfully identified causality and acquired higher-order generalizations using self-generated evidence during free play, and their performances were same as in didactic learning(Sim & Xu, 2017). However, if this conclusion is true across cultures? In the current study, we used the same methods and found that 2.5- to 4-year-old Chinese children could also acquire higher-order generalizations under two different learning conditions, but their performances were better in the didactic condition than that in the free play condition. One of the reasons affected childrens learning is parenting styles, but only in the free play condition: children with authoritative parents performed significantly better than children with authoritarian parents.

Key words: free play, active learning, higher-order generalization, parents cultural belief systems, parenting style

The Role of Causal Information and Perceived Knowledge in Decision-Making

Min Zheng
Stevens Institute of Technology, Hoboken, New Jersey, United States

Jessecae Marsh
Lehigh University, Bethlehem, Pennsylvania, United States

Samantha Kleinberg
Stevens Institute of Technology, Hoboken, New Jersey, United States

Abstract

Causal knowledge is key to making effective decisions, yet little is known about how we combine new causal information with what we already know. This scenario, with a mix of prior beliefs and new information is common to many settings, and is pervasive in health decisions. We specifically examine how decision-making with causal models differs in abstract decisions versus those more reminiscent of daily life, and how new information interacts with people's perceived knowledge about the decision-making domains. We found that while people can successfully use causal models to answer abstract questions, causal models can lead to worse choices in everyday decisions, especially when people believe they know a lot about the domain (Experiment 1). We then used an IOED task to determine if showing people how little they actually understand about a domain may improve the use of causal models in decision-making (Experiment 2).

Change and social distribution of figurative language

on Uruguayan female population

Roberto Aguirre. Center of Basic Research in Psychology. Uruguay

Manuel García-Ruiz. University Institute of de Lisbon. CIES-IUL. Portugal

Yliana Rodríguez. Foreign Languages Center. Uruguay

Mauricio Castillo. Center of Basic Research in Psychology. Uruguay

María Noel Macedo. Center of Basic Research in Psychology. Uruguay

Metaphors change through time in different cultures, languages and across generations. This research aimed to test the change and social distribution of some metaphors in Uruguayan Spanish. This study tested figurative expressions for the metaphors BEING IN THE OVEN IS DIFFICULTIES / HAZARDNESS, BANKING SOMETHING OR SOMEBODY IS BEARING IT and TO BE FLYING IS DOING SOMETHING WELL. On a multiple choice online questionnaire 267 Uruguayan female chose the meaning and the frequency that they believe they use previous metaphors. By using Multiple Correspondence Analysis (MCA) as a visual exploratory statistical tool, the study suggested Cultural Immersion and Metaphorical Proficiency as dimensions for explaining the social distribution of the aforementioned metaphors. But even though MCA seems to be a useful tool for understanding the metaphors' vitality, the short percentage of the variance explained by the dimensions suggests introducing additional categories for obtaining an adequate proportion of this variance.

Modulation of mood on eye movement pattern and performance in face recognition

Jeehye An
University of Hong Kong, Hong Kong, Hong Kong

Janet Hsiao
University of Hong Kong, Hong Kong, Hong Kong

Abstract

Research has suggested negative mood facilitates local attention while positive mood facilitates global attention. In face recognition, looking at the eyes has been associated with engagement of local attention as well as better recognition performance. Accordingly, negative mood changes may lead to more eyes-focused eye movements and consequently enhance recognition performance. We tested this hypothesis using mood induction. Through Eye Movement analysis with Hidden Markov Models (EMHMM), we discovered eyes-focused and nose-focused strategies. Although negative mood changes predicted increased eye movement pattern similarity to the eyes-focused strategy, it did not predict changes in recognition performance. Furthermore, most participants did not switch between eyes-focused and nose-focused strategies despite changes in mood. We conclude that mood changes lead to eye movement pattern changes that are not sufficient to modulate recognition performance as individuals may have preferred eye movement strategies impervious to transitory mood changes.

Surprise-Based Learning with Non-Solid Substances

Erin Anderson
Northwestern University, Evanston, Illinois, United States

Natasha Zeigler
Northwestern University, Evanston, Illinois, United States

Susan Hespos
Northwestern University, Evanston, Illinois, United States

Lance Rips
Northwestern University, Evanston, Illinois, United States

Abstract

Violating infants expectations about solid objects (e.g., a ball passing through a wall) leads to increased exploration and learning about the objects properties (Stahl & Feigenson, 2015). How limited is this type of learning? Infants can anticipate how non-solid substances behave and interact (Hespos et al., 2009; 2016), but the non-cohesive nature of substances means that they have less predictable shapes and boundaries. Across four trials, we presented 12- to 14-month-olds with items that looked solid or liquid. For half the trials, the items behavior was consistent with its appearance, so, for example, it looked solid and remained cohesive. For the other half, the behavior was inconsistent. Infants spent significantly more time exploring the inconsistent items, whether solid or non-solid, $F(1, 57) = 24.00$, $p = .001$, $p = .29$. These results suggest that infants preference for learning from violations might be a general mechanism responsible for new knowledge.

Explicit cues lead to reward-related enhancements in motor skill performance

Sean Anderson
University of Michigan, Ann Arbor, Michigan, United States

Taraz Lee
University of Michigan, Ann Arbor, Michigan, United States

Abstract

A large body of evidence suggests that motor sequencing skills can be trained either implicitly or explicitly. That is, participants can learn implicitly outside of conscious awareness or they can be explicitly told and/or cued to existence of repeating sequences. Although explicit learning often coincides with faster skill acquisition, the role of conscious awareness in skill learning is still debated. Some recent work has suggested that the benefits seen from explicit learning are not due to added conscious knowledge per se, but rather an increase in intrinsic motivation. Here we show that although performance-contingent monetary incentives lead to improved performance in all subjects, this effect is larger for explicitly trained subjects. This suggests that intrinsic motivation alone cannot explain the superior performance in explicitly trained tasks and that explicit knowledge can confer an additional benefit in that it can allow individuals to better contextually modulate their behavior.

Childrens Unscientific Conceptions Before and After Instruction in Space Science

Florencia Anggoro
College of the Holy Cross, Worcester, Massachusetts, United States

Benjamin Jee
Worcester State University, Worcester, Massachusetts, United States

Amanda McCarthy
College of the Holy Cross, Worcester, Massachusetts, United States

Victoria Jackson
College of the Holy Cross, Worcester, Massachusetts, United States

Demitria Tsitsopoulos
College of the Holy Cross, Worcester, Massachusetts, United States

Ioli Karageorgiou
College of the Holy Cross, Worcester, Massachusetts, United States

Abstract

Research has documented childrens difficulty reconciling observations of the sky (Earth-based perspective) with scientific models of the solar system (space-based perspective) (e.g., Vosniadou & Brewer, 1994). We developed a coding rubric to capture childrens explanations before and after instruction that emphasized relational learningmapping the spatial, temporal, and causal relations inherent in the day-night cycle. We focused on several key dimensions including the perspective of the child and their causal attributions, focusing primarily on their mental model (e.g., Sun goes up/down). We coded pre- and post-test videos from 3rd graders from two experiments (N=205) using the rubric. Results suggest that (a) consistent with prior findings, children who received the instruction demonstrated fewer unscientific conceptions about Sun motion at posttest, and (b) these conceptions were more pronounced in modeling than in verbal responses. We conclude that topics that require integration between Earth- and space-based perspectives are particularly challenging for young children.

Using Eye Tracking to Examine Morphological Features and Working Memory Capacity in Agreement Processing

Erik Arnold
Brigham Young University, Provo, Utah, United States

Deryle Lonsdale
Brigham Young University, Provo, Utah, United States

Abstract

Morphosyntactic agreement refers to a head-dependent relation where similar features are shared between syntactic constituents. Several grammatical features are expressed in agreement relations through different manifestations of exponence (e.g. separative and cumulative). Whereas prior research has largely examined features in separative exponence (e.g. gender and number), this study investigates differences in the on-line processing of features in cumulative exponence. Using eye tracking, we investigated differences between second language (L2) learner processing of person, number, and tense features in Spanish verbal agreement. We also examined the effect of working memory capacity (WMC) on learners on-line processing of these same features. The results of our linear mixed effects model indicated learners had greater perturbation in processing person and tense agreement violations compared to number agreement violations. The results also revealed that learners with higher WMC demonstrated less perturbation to agreement violations of each feature type than learners with lower WMC.

A computational cognitive modeling approach to understand test-takers strategy use in drag-and-drop math questions

Burcu Arslan
Educational Testing Service, Princeton, New Jersey, United States

Yang Jiang
Educational Testing Service, Princeton, New Jersey, United States

Tao Gong
Educational Testing Service, Princeton, New Jersey, United States

Madeleine Keehner
ETS, Princeton, New Jersey, United States

Irvin Katz
Educational Testing Service, Princeton, New Jersey, United States

Abstract

Computer-based educational assessments often include questions with a drag-and-drop response. Logged data obtained from drag-and-drop responses allow us to go beyond scores, investigating the response strategies test-takers use to reach an answer. There is no previously published research on strategies used by test-takers in answering drag-and-drop questions. We tested 476 MTurk participants under five conditions where key design features of mathematics questions were manipulated. Regardless of the design manipulations, participants mostly used one of the two possible systematic response strategies. Using PRIMs cognitive architecture (Taatgen, 2013), we constructed computational cognitive models to simulate the differences between these two strategies. The models were able to capture participants reaction time patterns. Our conclusion based on the models is that most participants apply a cognitively less demanding strategy by offloading cognition on action, which is in line with the idea of strategy selection as rational metareasoning (Falk & Griffiths, 2017).

Co-thought gestures during abstract relational reasoning

Misha Ash
University of Chicago, Chicago, Illinois, United States

Kensy Cooperrider
University of Chicago, Chicago, Illinois, United States

Dedre Gentner
Northwestern University, Evanston, Illinois, United States

Susan Goldin-Meadow
University of Chicago, Chicago, Illinois, United States

Abstract

When talking about abstract relations like better and worse, people often use gestures arrayed in space to get their point across. But are these analogical gestures solely communicative props that make abstract content more accessible for listeners, or do they also reflect an integral part of reasoning? To address this question, we investigated whether people would produce analogical gestures outside of a communicative context. In a linear syllogism task, participants spontaneously gestured on 52.4% of trials on average; most participants (87.5%) gestured on at least one trial. Trials involving spatial relational terms prompted more gestures per trial than those with non-spatial terms (spatial: M = 2.87; non-spatial: M = 2.29; $F(1, 23) = 7.62$, $p = .011$). Analogical gestures thus do occur outside of communicative contexts, suggesting that they serve to aid the reasoning process itself. An in-progress follow-up study replicates and extends these findings.

Role of Variety in Cognitive Improvement From Action Video Games

Katie Bainbridge
University of California, Santa Barbara, Santa Barbara, California, United States

Richard Mayer
University of California, Santa Barbara, Goleta, California, United States

Abstract

Participants were divided into three groups. One group played Call of Duty: Black Ops Multiplayer in a variety of maps for 9 hours over 2 weeks, another played in the just one map for 9 hours over 2 weeks, and the last did not play any video games for the duration of the study. All groups took three measures of visual attention skill at the start and close of the study: Useful Field of View (UFOV), Multiple Object Tracking (MOT) and Attentional Blink (AB). Results indicate that those who played Call of Duty did not improve more than those who did not from pretest to posttest, regardless of group.

Embodied Measurements of Ideological Positioning

Brandon Batzloff
University of California Merced, Merced, California, United States

Michael Spivey
UC Merced, Merced, California, United States

Abstract

Prior studies have shown tests for scales used to describe an individuals ideological position are not replicable. We examined ideological positioning of individuals through two mouse tracking tasks. First, participants were asked to select from six ideologies, mixed with distractors, they believed described them. They were then shown ten defined traits of these ideologies. Next, participants were asked to choose between pairs of compared traits and assign them to a displayed ideology. The first task was to determine which ideologies participants were most closely associated with, while the second was used to determine how each individual defined ideologies. In this way, we were able to gain insight into how people define themselves when completing discrete tasks, such as answering political questionnaires. Results show differences in individual ideological definitions. We have begun grouping statistically similar responses. It is our hope that this data will help develop realistic scales of ideological positioning.

A multi-study neuroeducational perspective on vocabulary learning

Peta Baxter
Radboud University Nijmegen: Donders Institute for Brain, Cognition and Behavior, Nijmegen, Netherlands

Randi Goertz
Radboud University Nijmegen: Donders Institute for Brain, Cognition and Behavior, Nijmegen, Netherlands

Lukas Ansteeg
Radboud University Nijmegen: Donders Institute for Brain, Cognition and Behavior, Nijmegen, Netherlands

Josh Ring
Radboud University Nijmegen: Donders Institute for Brain, Cognition and Behavior, Nijmegen, Netherlands

Marianne van den Hurk
Radboud University Nijmegen: Behavioral Science Institute, Nijmegen, Netherlands

Mienke Droop
Radboud University Nijmegen: Behavioral Science Institute, Nijmegen, Netherlands

Ton Dijkstra
Radboud University Nijmegen: Donders Institute for Brain, Cognition and Behavior, Nijmegen, Netherlands

Harold Bekkering
Radboud University Nijmegen: Donders Institute for Brain, Cognition and Behavior, Nijmegen, Netherlands

Frank Leone
Radboud University: Donders Institute for Brain, Cognition and Behavior, Nijmegen, Netherlands

Abstract

We aim to apply cognitive neuroscience insights to vocabulary learning practice. Towards this end, we review current educational methods in relation to important characteristics of the mental lexicon, such as similarity-coding. This shows that methods relate poorly to the mental lexicon, and that especially contrasting - explicitly distinguishing similarities - receives little attention. To remedy this, we run experiments to put these findings into practice. First, we ask participants to learn artificial vocabulary using retrieval practice multiple-choice, manipulating the orthographic and semantic similarity of distractors. The prediction is that learning will be harder but more effective depending on similarity and translation direction. Second, we test whether participants show indications of gradient descent learning when guessing in recall retrieval practice. Thirdly, we use cognitive neuroscience and large scale word learning data to model the mental lexicon. Combined, these studies potentially offer relevant scientific and societal insights, applicable to school settings.

Inferior frontal gyrus involvement during search and solution in verbal creative problem solving: A parametric fMRI study

Maxi Becker
University Medical Center Hamburg-Eppendorf, Hamburg, Hamburg, Germany

Tobias Sommer
University Medical Centre Hamburg-Eppendorf, Hamburg, Germany

Simone Khn
University Medical Center Hamburg-Eppendorf, Hamburg, Germany

Abstract

In verbal creative problems like compound remote associates (CRAs), the solution is semantically distant and there is no predefined path to the solution. Therefore, people first search through the space of possible solutions before retrieving the correct semantic content by extending their search space. We assume that search and solution are both part of a semantic control process which involves the inferior frontal gyrus (IFG). Furthermore, the degree of the IFG involvement depends on how much the search space needs to be extended, i.e. how semantically distant the solution is. To demonstrate this, we created a modified CRA paradigm which systematically modulates the semantic distance from the first target word to the solution via priming. We show that brain areas (left inferior frontal gyrus and middle temporal gyrus) associated with semantic control are already recruited during search. In addition, we found a linear correlation between the BOLD activation of the IFG (pars orbitalis and triangularis) and the search space extension. However, this linear relationship could only be observed during and shortly before the correct solution but not during search. We discuss the role of the IFG in accessing semantically distant information during verbal creative problem solving.

Systematic ambiguity: the effect of creativity and fractal dimension on pareidolia

Antoine Bellemare
Concordia University, Montreal, Quebec, Canada

Yann Harel
Universit de Montral, Montreal, Quebec, Canada

Julien Besle
American University of Beirut, Beirut, Lebanon

Arne Dietrich
American University of Beirut, Beirut, Lebanon

Karim Jerbi
Universit de Montral, Montreal, Quebec, Canada

Abstract

Pareidolia refers to the perception of recognizable forms in noisy or ambiguous stimuli. It has mostly been studied in the context of pathologies such as schizophrenia and dementia. However, pareidolic perception occurs in general population without associated psychotic symptoms. This phenomenon is conceived as a compensatory perceptual mechanism that enables the brain to deal with ambiguous information. It has been hypothesized that pareidolia would be related to the emergence of creative ideation. In this study, we investigated the effect of fractal dimension on pareidolic perception by asking participants to perceive as many recognizable forms as possible in a set of Fractional Brownian Motion images with varying fractal dimensions. In addition, we further investigated, using questionnaires, whether creativity, openness personality trait and schizotypy are linked to pareidolic perception. Results show that creativity facilitates pareidolic perceptions and that this effect interacts significantly with the state of flow.

HOT: Higher Order Tetris, Experts' Subgoals and Activities

Jacquelyn Berry
Rensselaer Polytechnic Institute, Troy, New York, United States

Wayne Gray
Rensselaer Polytechnic Institute, Troy, New York, United States

Abstract

For Tetris, clearing 4 lines at once (a "Tetris") results in 7.5 times as many points as clearing one line four times. Getting a Tetris requires a solid block of filled cells, 9 columns wide and 4 rows high. That block leaves vacant one column. If an I-beam appears, all 4 rows can be cleared. Finalists at the Classic Tetris World Championships have an explicit subgoal structure not seen in lesser players. Among the 32 competitors, the 4 finalists are those who are most adept at maintaining or preparing the board for a Tetris by executing one of these subgoals, as needed. We present a video-based analysis which compares the proportion of time spent on each activity between those eliminated on the first tournament round and those who survive to the final round.

Masterminding in Education: Bringing cognition, emotion and motivation together in a unified mathematical framework

Lara Bertram
University of Surrey, Guildford, United Kingdom

Eric Schulz
Harvard, Boston, Massachusetts, United States

Elif zel
Ludwigsburg University of Education, Ludwigsburg, Germany

Matthias Hofer
Massachusetts Institute of Technology, Cambridge, Massachusetts, United States

Laura Martignon
Ludwigsburg University of Education, Ludwigsburg, Baden Wrttemberg, Germany

Jonathan D. Nelson,
University of Surrey, Guildford, United Kingdom

Abstract

In this research project, a novel app-based version of the code breaking game Mastermind, Entropy Mastermind, was introduced and evaluated as a learning medium in undergraduate cognitive psychology and in primary mathematics education. In a quasi-experimental pre- and posttest design we investigated a) the role of individual differences in game play and learning, b) the effectiveness of Entropy Mastermind for giving students of different age groups experientially grounded access to the fundamental concepts of proportions and mathematical entropy, and c) effects of game play on students academic emotions, motivation and attitudes. Data analyses revealed significant associations between cognitive variables, emotional-motivational factors and game play parameters. We present computational modeling results of students search strategies and entropy intuitions within a unified framework of entropy measures, the Sharma-Mittal space. Potential applications in digitalized learning environments at the interface between mathematics and computer science will be discussed.

Movements and Visuospatial Working Memory: Examining the Role of Movement and Attention to Movement

divya bhatia
Sapienza University of Rome, Rome, Italy

Pietro Spataro
Universitas Mercatorum, Rome, Italy

Clelia Rossi-Arnaud
Sapienza University of Rome, Rome, Italy

Abstract

Previous studies have shown that, under specific conditions, pointed-to arrays can be recognized better than arrays that are only visually observed. In the present study we investigated whether this memory advantage is due to movement per se or to attention to the movement. In two experiments we modulated the amount of attention devoted to the execution of pointing movements by comparing the effects of passive and active pointing in a visuo-spatial working memory (VSWM) task. In Experiment 1, participants were instructed that their hands would be moved by the experimenter (passive pointing); in Experiment 2, participants performed active and passive pointing movements in random alternation. Results showed that passive movements benefitted VSWM only when they were alternated with active movements. This finding suggests that the key factor underlying the positive effect of pointing on VSWM is the increased attention devoted to them in the mixed pointing conditions of Experiment 2.

The Effect of Graphics on Mind Wandering in Online Video Lectures

Laura Bianchi
University of Waterloo, Waterloo, Ontario, Canada

Kristin Wilson
University of Waterloo, Waterloo, Ontario, Canada

Evan Risko
University of Waterloo, Waterloo, Ontario, Canada

Abstract

There is a rising interest in determining the most effective (i.e., the most conducive for learning) way to present online lecture information. The cognitive load model of multimedia learning suggests that learners are capacity limited. Lecture graphics that are interesting but extraneous to the content (e.g., a celebrity), have been shown to impair comprehension of the material (i.e., the seductive detail effect). The seductive detail effect likely results from a lack of cognitive resources available to maintain attention. Across 2 experiments, the use of graphics was manipulated in a psychology online video lecture. We demonstrate no differences across conditions (i.e., no images, relevant images, and seductive images) in overall comprehension and limited differences mind wandering behaviour.

Its About Time: Temporal Problem Solving With Static Drawings in Animation Design

Janet Blatter
Independent Research, Montreal, Quebec, Canada

Abstract

Drawings and diagrams have long been researched as supporting design thinking in many domains. However, real-world design that deals with, in, and about time as part of the process and outcome is less studied. How do designers in authentic practices use static drawings to think about time in different frames of reference? With a view of situated, mediated cognition as in Activity Theory, this presentation is a case study of an expert animator at the National Film Board of Canada. It focuses on the use of static drawings in finding temporal problems in the key frames of references used in creating narrative animation. The study suggests that the icons forming the basis of his drawings are used strategically, as indices to his design process, the fictive motion, and the sequence and duration of actions that must be seen at 24 frames per second.

Improving Fraction Knowledge to Open the Door to Algebra

Julie Booth Ph.D.
Temple University, Philadelphia, Pennsylvania, United States

Kristie J. Newton
Temple University, Philadelphia, Pennsylvania, United States

Christina Barbieri
University of Delaware, Newark, Delaware, United States

Laura K. Young
Temple University, Philadelphia, Pennsylvania, United States

Nicole Hallinen
Temple University, Philadelphia, Pennsylvania, United States

Abstract

Recent studies have established that students knowledge about fractions is predictive of their readiness, performance, and learning in Algebra (Booth & Newton, 2012; Booth, Newton, & Twiss-Garrity, 2013). However, it is yet unknown whether the relationship between fractions and algebra is causal; that is, would improving students' knowledge of fractions cause improvements in their ability to perform in and learn Algebra? The present study examines the impact of improving fraction computation and fraction magnitude knowledge in real world classrooms on middle school students' learning of key concepts and problem-solving techniques in Algebra. Individual differences in the impact of improved fraction knowledge will also be investigated and discussed.

Stability of Core Language Skill from Infancy to Adolescence in Typical and Atypical Development

Marc Bornstein
NICHD & IFS, Bethesda, Maryland, United States

Abstract

Individual differences are a central characteristic of child language, and a conceptual issue in language and developmental science is stability. Language was evaluated at 6 months and annually through 15 years in 5167 (50.2% girls) white, monolingual singletons: 4111 typically developing children; 435 moderate-late and 51 very preterm children; 322 children with dyslexia; 89 children with autism; and 221 children who had mild and/or moderate hearing impairment. Structural equation modelling showed both typical and atypically developing childrens language skills had medium to large average stabilities between successive waves over the span of 15 years, even accounting for child nonverbal intelligence and sociability and maternal age and education. The strong stability of child language skill from early in development across typical and at-risk groups points to a highly conserved and robust individual-differences characteristic and underscores the importance of identifying lagging language skills and promoting childrens language environment well before formal schooling.

The Effect of Multiple Repetitions on Scanning in Long-Term Memory

Ian Bright
Boston University, Boston, Massachusetts, United States

Rebecca DiDomenica
Boston University, Boston, Massachusetts, United States

Rui Cao
Boston University, Boston, Massachusetts, United States

Marc Howard
Boston University, Boston, Massachusetts, United States

Abstract

Cognitive psychologists have hypothesized that episodic recall is caused by the recovery of a gradually-changing state of spatiotemporal context. Little is known about the processes that cause successful recovery of this temporal context. Recent behavioral evidence suggests that in continuous recognition tasks, the retrieval time necessary to recover a previous context depends on the recency of the memory. Previous work has found that the non-decision time to retrieve a memory goes up with the logarithm of its recency. This suggests retrieval of temporal context proceeds via scanning along a compressed timeline but also contradicts earlier work suggesting that recency affects the drift rate of retrieval more than the non-decision time. Here we explore the effect of multiple repetitions on this counterintuitive result in continuous recognition. Our results find that while repeating items speeds up the time to access a memory, the recency effect persists out to at least five repetitions.

A Formalization of Cognitive Continuity/Discontinuity, to Settle the Darwin's-Mistake Debate

Selmer Bringsjord
RPI, Troy, New York, United States

Naveen Sundar Govindarajulu
Rensselaer Polytechnic Institute, Troy, New York, United States

Atriya Sen
RPI, Troy, New York, United States

Christina Elmore
RPI, Troy, New York, United States

Abstract

Darwins /Origin/ doesn't discuss the evolution of the human mind. He saved treatment of this topic for the subsequent /Descent of Man/, in which he advanced two claims: (C1) If the cognitive powers of nonhuman animals are discontinuous with those possessed by humans, then the human mind isnt the product of evolution by mutation and natural selection. (C2) The cognitive powers of nonhuman animals, including specifically reasoning powers, are continuous with those enjoyed by humans; continuity is established. Penn, Holyoak, and Povinelli (2008) have in /BBS/ written "Darwin's Mistake," in which they purport to refute C2 by establishing discontinuity (they don't affirm C1). Many vehemently disagree with PHP, and the debate remains intense, and unresolved. Yet, (1) the hitherto informal concept of continuity can be formalized, and (2) that formalization, applied to the debate, settles it. We provide the formalization, and with it settle the debate (in favor of PHP).

Using Graph Theory to Understand the Structure of Event Knowledge in Memory

Kevin Brown
Oregon State University, Corvallis, Oregon, United States

Nickolas Christidis
University of Western Ontario, London, Ontario, Canada

Jeffrey Elman
University of California, San Diego, California, United States

Ken McRae
University of Western Ontario, London, Ontario, Canada

Abstract

There are several competing theories regarding how event knowledge is represented in the mind, ranging from a strictly temporally ordered list of activities to sets of connected scenes which may themselves consist of ordered activities. We employed a network science approach to provide data-driven insight into event structure. We converted sets of human generated activity sequences, in which roughly 25 participants list up to 12 activities for 81 different events (making a sandwich, cleaning the house, taking money out of an ATM, etc.), into directed, weighted networks. Analyses of the event networks revealed a complex and varied temporal structure to events. In addition, we were able to identify scenes within events, and use graph theory to understand activity centrality, popularity, and influence, as well as the coupling between these activity characteristics. In the aggregate, we find that network science makes multiple data-driven, empirically testable predictions about event structure.

Who are you talking to like that? Exploring adults' ability to discriminate child- and adult-directed speech across languages

John Bunce
University of Manitoba , Winnipeg, Manitoba, Canada

Melanie Soderstrom
University of Manitoba, Winnipeg, Manitoba, Canada

Md Momin Al Aziz
University of Manitoba, Winnipeg, Manitoba, Canada

Marisa Casillas
Max Planck Institute for Psycholinguistics, Nijmegen, Netherlands

Abstract

Child-directed speech (CDS) shows similar characteristics across many languages, but is known to vary across cultural and demographic groups (Lieven, 1994). Is CDS consistently discriminable from adult-directed speech (ADS) despite these differences? Perhaps: adults listening to scripted female CDS can discriminate ADS-vs-CDS in a language they dont speak (Bryant et al., 2012). We build on this finding by asking North American English speakers to classify utterances from the natural language input of 10 Tseltal Mayan children as ADS or CDS (n = 1836 utterances). Binomial mixed-effects regressions of accuracy show that listeners are more accurate on utterances from females (mFemale = .81, mMale = .67) and adults (mAdult = .82, mChild = .72), with a larger gender effect for child speakers (m: Girl-Boy = 0.31, Woman-Man = 0.09). This suggests that (a) ADS-CDS discrimination of natural speech in an unrelated, non-familiar language is reliable (mAll = 0.78) and also (b) modulated by speaker type.

The Effects of Video Interviews on Perceptions of Applicant Quality

Devin Burns
Missouri University of Science & Technology, Rolla, Missouri, United States

Denise Baker
Missouri University of Science & Technology, Rolla, Missouri, United States

Clair Kueny
Missouri University of Science & Technology, Rolla, Missouri, United States

Matthew Jordan
Missouri University of Science & Technology, Rolla, Missouri, United States

Abstract

Previous research has shown that job candidates are rated significantly higher if evaluators are allowed to listen to their pitches rather than just reading the transcript (Schroeder & Epley, 2015). That research did not find any additional benefit from seeing the candidate on video, but did not examine whether watching a video interview was different from watching an interview in-person. Our experiment had 50 participants watch a mock interview in-person while 50 other participants watched the same interviews ostensibly through a live video feed in another room. Those who watched through video rated the job applicant significantly lower on all measured dimensions including agency, hireability, and intellect. These findings indicate that job applicants who are interviewed through a video-conference service or whose interviews are recorded and watched later are at a significant disadvantage to those who can be observed live. Potential causes and ameliorations of these effects are discussed.

Task Characteristics and Individual Differences in Judgments of Relative Direction

Heather Burte
University of Texas at Arlington, Arlington, Texas, United States

Abstract

Judgments of relative direction (JRD) have been frequently used to understand peoples mental representation of outdoor and indoor spaces. In JRD experiments, experimenters need to identify a signal within the trial-by-trial and participant-by-participant variability. However, it is not well understood how characteristics of the task and differences between individuals contributes to performance variability. In this paper, I investigated task characteristics (i.e., reference frames used in instructions, orienting and target headings, and distances between headings) and individual differences (i.e., gender, sense-of-direction, familiarity, and strategy use) to provide insights into the factors that influence JRD accuracy and variability. Using the findings of this study, I make recommendations for best-practices in JRD methods and analyses.

The Role of Task Characteristics and Individual Differences in Pointing to Unseen Locations

Heather Burte
University of Texas at Arlington, Arlington, Texas, United States

Abstract

Pointing tasks have been used for decades to investigate peoples understanding of environmental-scale spaces. Most of this research has used the variability of pointing estimates to provide insights into peoples cognitive maps. In pointing experiments, experimenters need to identify a signal within the trial-by-trial and participant-by-participant variability. However, it is not well understood how characteristics of the task and differences between individuals contribute to pointing variability. In this paper, I investigated characteristics of pointing tasks and individual differences (i.e., gender, sense-of-direction, familiarity, and strategy use) to provide insights into the factors that influence pointing accuracy and its variability. Using the findings of this study, I make recommendations for best-practices in pointing task methods and analyses.

Motivated Reasoning in Causally Ambiguous Explore-Exploit Situations

Zachary Caddick
University of Pittsburgh, Pittsburgh, Pennsylvania, United States

Benjamin Rottman
University of Pittsburgh, Pittsburgh, Pennsylvania, United States

Abstract

Two studies investigated how political attitudes affect causal learning. Participants were tasked with testing economic policies to maximize the economic output of an imaginary country. Based on their political attitudes, participants were either strongly in favor or strongly against the policies (Study 1), or could also have neutral attitudes (Study 2). Some policies had fairly clear positive or negative effects. But some were more ambiguous; they initially had positive effects but eventually had negative effects on the economy, or vice versa. After testing the policies, participants falsely believed that the policies that fit with their political attitudes were more effective, and this bias was exacerbated for the policies that had different short vs. long-term effects. This research shows the power of motivated reasoning and provides a well-controlled method to study the effects of motivated reasoning on causal learning in explore-exploit situations.

A Dynamic Neural Field Model of the McGurk Effect and Incongruous Audiovisual Speech Stimuli

Ryan Cannistraci
University of Tennessee, Knoxville, Knoxville, Tennessee, United States

Jessica Hay
University of Tennessee, Knoxville, Knoxville, Tennessee, United States

Aaron Buss
University of Tennessee, Knoxville, Knoxville, Tennessee, United States

Abstract

Our Dynamic Neural Field (DNF) model aims to simulate audiovisual integration in speech perception, including the well-known McGurk effect (McGurk & MacDonald, 1976). The classic McGurk effect is characterized by a fusion effect, whereby incongruent audio and visual stimuli are fused into a single percept, however other interesting audiovisual effects are present in the extant literature. Our DNF model uses the same architecture and parameters across stimulus combinations to simulate a host of audiovisual illusory effects as well as audiovisually congruent, auditory-only, and visual-only controls. Our simulation results replicate rates of visual-dominant percepts, audiovisual fusion percepts, auditory-dominant percepts, and auditory dichotic fusion found in the extant literature, and illustrate how a complex pattern of responses across different stimuli configurations can arise from common neural dynamics involved in binding information across sensory modalities. We are currently exploring how hemodynamic response predictions generated through our neural simulations relate to real-time behavior.

Origins of cross-domain asymmetries

Daniel Casasanto
Cornell University, Ithaca, New York, United States

Yamur Deniz Ksa
University of Chicago, Chicago, Illinois, United States

Abstract

Why do people use space to talk about time, and to think about time, more than vice versa? On one proposal, this space-time asymmetry arises from the greater perceptual availability of space. Alternatively, a space-time asymmetry in language could give rise to the space-time asymmetry in thought during early language acquisition. If this language-first view is correct, then parents should use space-time words (e.g., long) more often in their spatial senses than in their temporal senses, imparting to children the primacy of the spatial senses. More generally, childrens space-time word use should reflect the statistics of parental input. Results of a corpus analysis contradict both predictions: English speaking adults used polysemous words more often in their temporal senses than in their spatial senses, whereas young children showed the opposite pattern, in the same conversations. Asymmetries between space and time appear to precede and guide the acquisition of spatio-temporal language.

Eye-tracking as a Measure of Table Tennis Expert-Novice Differences in Theory of Mind

Ting-Hsuan Chang
National Cheng Kung University, Tainan, Taiwan

Fu-Zen Shaw
National Cheng Kung University, Tainan, Taiwan

Sheng-Fu Liang
National Cheng Kung University, Tainan, Taiwan

Hung-Ta Chiu
National Cheng Kung University, Tainan, Taiwan

Jon-Fan Hu
National Cheng Kung University, Tainan City, Taiwan

Wei-En Chang
National Cheng Kung University, Tainan, Taiwan

Abstract

Theory of Mind (ToM) refers to the ability of individuals to understand beliefs, desires, and emotions of others. Our study is based on the expert-novice paradigm and aims to investigate the operations of ToM of table tennis novices and experts by the patterns of eye movement. Stimuli integrated cognitive and affective ToM dimensions analogical to the table tennis situations and recorded response by eye-tracking technique. Reaction time, accuracy and eye movement data were analysis indexes. Study results revealed that experts could predict the shot actions and emotional states of opponents more quickly and accurately than novices, also there were differences in eye trajectory traces. The findings clearly show that eye-tracking technique can be used to illustrate table tennis expert-novice differences in ToM and provide suggestions for the development of table tennis training programs in use of eye tracker facilities.

The effect of word-by-word presentation on reading of Chinese texts by native Chinese readers and learners of Chinese as a second language

Jenn-Yeu Chen
National Taiwan Normal University, Taipei, Taiwan

Yalin Chuang
Nanya Institute of Technology, Taoyuan, Taiwan

Abstract

There are no spaces between words in Chinese texts and this can present a challenge in reading for learners of Chinese as a second/foreign language (CSL) and native Chinese alike. We designed a self-paced reading computer platform on which individual words were shown or highlighted successively as participants pressed the spacebar to read a text without word spaces. CSL learners could read faster in this way than the traditional way where the entirety of the unspaced text appeared as a whole. Native Chinese readers did not show such a beneficiary effect. The results support the Processing Cost Hypothesis which states that word segmentation when reading unspaced texts consumes processing resources and therefore saving the resources by providing segmentation cues could benefit readers only when processing resources are overtaxed under certain circumstances, e.g., reading difficult texts, under time pressure, for beginner readers, and for foreign learners.

Providing Stroke Sequence of Chinese Characters Facilitates Handwriting Learning in Children with Developmental Coordination Disorder

Rong-Ju Cherng
National Cheng Kung University, Tainan, Taiwan

Yi-Wen Liao
National Cheng Kung University, Tainan, Taiwan

Jenn-Yeu Chen
National Taiwan Normal University, Taipei, Taiwan

Abstract

The study investigated whether providing instruction on the stroke sequence would facilitate the learning of writing Chinese characters in children with developmental coordination disorder (DCD) and typically developing (TD) children. The children wrote six characters, three with stroke sequence instruction and three without. Each character was repeated 40 times. Trajectory, speed, on-paper time, in-air time, and number of changes in velocity direction per stroke (NCV) were measured with Wacom Intuos 5 digitizing writing tablet. The results showed a significant group effect, time (practice) effect and instruction effect but no interaction effects. Both groups of children showed a similar trend of improvement over practice with decreasing trajectory, increasing speed, decreasing on-paper time and in-air time. With stroke sequence instruction, both groups of children learned at a similar rate on most of the writing parameters. Instruction on stroke sequences helped the character writing of both the DCD children and the TD children.

Exploring Aha! moments during science learning

Christine Chesebrough
Drexel University, Philadelphia, Pennsylvania, United States

Jennifer Wiley
University of Illinois Chicago, Chicago, Illinois, United States

Abstract

The Aha! experience has mainly been studied in the context of insightful problem solving, but less work has investigated Aha! experiences that can occur during learning. In these studies, participants were asked to self-report Aha! moments when learning about principles in Biology, such as symbiosis or mimicry, from sets of three divergent examples. In the problem-oriented condition, participants saw the examples and were asked to generate their common principle. In the direct instruction condition, participants were told the principle directly. Participants were significantly more likely to report Aha! moments in the problem-oriented condition. Although having an Aha! experience did not always lead to better learning, the likelihood of having an Aha! moment was positively correlated with several student characteristics, particularly in the problem-oriented condition. These studies offer another perspective on the potential benefits of learning from invention activities.

Modeling the Costly Rejection of Wrongdoers by Children using a Bayesian Approach

Theodore Cheung
University of Toronto, Toronto, Ontario, Canada

Rachel Eng
University of Toronto, Toronto, Ontario, Canada

Daphna Buchsbaum
University of Toronto, Toronto, Ontario, Canada

Abstract

In previous work, young children avoided associating with a wrongdoer, despite incurring a personal cost. Such aversion to wrongdoers, arguably a reflection of moral development, weakens when the cost becomes very large (Tasimi & Wynn, 2016). We model this moral decision-making process using the nave utility calculus (Jara-Ettinger et al., 2016), assuming utility maximization amidst uncertainty using Bayesian framework. The cost is defined as the number of stickers forgone by choosing a nice persons smaller offer over a mean persons larger one, following the ratios of 1:2, 1:4, 1:8, and 1:16. Our model aims to explain previous findings, and test predictions for new ratios. Compared to a baseline condition where no background information is available, children are predicted to choose the nice person when the cost is low, but reverse their preference when the cost becomes increasingly high, which would suggest a utility account for moral decision making.

Math ability varies independently of number estimation in the Tsiman

Samuel Cheyette
UC Berkeley, Berkeley, California, United States

Benjamin Pitt
UC Berkeley, Berkeley, California, United States

Steven Piantadosi
UC Berkeley, Berkeley, California, United States

Edward Gibson
Massachusetts Institute of Technology, Cambridge, Massachusetts, United States

Abstract

How do people reason about mathematical concepts like addition and subtraction? According to one proposal, mathematical thinking is supported in part by the approximate number system (ANS), a primitive cognitive system for estimating the numerosity of a set, without counting. Here we tested this proposal in the Tsiman, a culture of farmer-foragers in the Bolivian Amazon. Compared to industrialized societies like the US, the Tsiman have high variability in their level of education and number knowledge. In a large sample of Tsiman adults, math ability was positively correlated with ANS performance, consistent with previous findings. However, this correlation disappeared when controlling for participants education, and when controlling for their ability to sustain attention. These findings challenge the claim that the ANS supports math ability. Rather, performance on ANS tasks and math tasks may both be shaped by non-numerical abilities practiced (or selected for) in educational settings.

L2 learners' phonemic sensitivity: MMN & L2 proficiency

Jeongwha Cho
Seoul National University, Seoul, Korea, Republic of

Sun-Young Lee
Cyber Hankuk University of Foreign Studies, Seoul, Korea, Republic of

Mijung Sung
Seoul National University, Seoul, Korea, Republic of

Ki-Chun Nam
Korea University, Seoul, Korea, Republic of

Hyeon-Ae Jeon
Daegu Kyungbook Institue of Science and Technology (DGIST), Daegu, Korea, Republic of

Youngjoo Kim
Kyung Hee University, Youngin, Korea, Republic of

Abstract

This study examined the acquisition of Korean stop sounds /t/(), /t/() and /th/() by Chinese learners of Korean using ERP focusing on the role of L2 proficiency. A total of 28 learners (16 advanced and 11 intermediate) and 18 native controls participated in the experiment with four conditions: (i) standard /ta/ vs. deviant /tha/, (ii) standard /ta/ vs. deviant /t/, (iii) standard /tha/ vs. deviant /ta/, and (iv) standard /ta/ vs. deviant /ta/. The results of the AX discrimination task found no significant differences between groups showing high accuracy rates from 73% to 84%. However, their brain responses were different: P3 was found only for the intermediate group in condition (iii) although MMN was elicited in both groups in the other three conditions. The results indicate that learners sensitivity to the differences of stop sounds develops as their general proficiency improves. Still, their sensitivity is weaker than native speakers.

Comparing the social judgements between American and Taiwanese cultures

Yun Chuang
National Cheng Kung University, Tainan, Taiwan

Jon-Fan Hu
National Cheng Kung University, Tainan, Taiwan

Abstract

While observing others in the society, people make explanations and judgements about others' behaviors. However, there should be cultural differences in affecting behavior judgments. The aim of the present study is to examine whether there are cognitive or emotional differences between Eastern and Western cultures while judging other peoples behaviors. Vignettes stimuli and the questions developed by Knutson et al. (2010) were used to measure how Taiwanese participants think and react while making behavior judgements. Factor analysis is conducted to compare the results with the original study completed in the US. The results revealed that for the Taiwanese participants, emotional aversion was more related to the norm violation, while for the American participants, according to the original study, aversion was more related to the social affect. The results of this comparison have demonstrated cultural differences between Taiwan and the US in how aversion could be evoked by observing others behaviors.

Go big and go grounded: Categorical structure emerges spontaneously from the latent structure of sensorimotor experience

Louise Connell
University of Lancaster, Lancaster, United Kingdom

James Brand
University of Canterbury, Christchurch, New Zealand

James Carney
Brunel University, London, United Kingdom

Marc Brysbaert
Ghent University, Ghent, Belgium

Dermot Lynott
Lancaster University, Lancaster, United Kingdom

Abstract

Many theories of semantic memory assume that categories spontaneously emerge from commonalities in the way we perceive and interact with the world around us. However, efforts to test this assumption computationally have been hampered by use of abstracted features without clear sensorimotor grounding and over-reliance on small samples of concepts from a limited number of categories. Taking a radically different approach, we examined whether categorical structure emerges spontaneously from the latent structure of sensorimotor experience by creating a fully-grounded multidimensional sensorimotor space at the scale of a full-size human conceptual system (i.e., 11 sensorimotor dimensions x 40,000 concepts). We found evidence for (a) a high-level separation of abstract and concrete categories (which was not enhanced by the inclusion of affective information); (b) a hierarchical structure of concrete concepts that separated categories commonly impaired in double dissociations, such as fruit/vegetables, animals, tools, and musical instruments; and (c) a flatter hierarchy of abstract concepts that separated categories such as negative emotions, units of time, social relationships, and political systems. These findings demonstrate that grounded sensorimotor information is fundamental to the representation of all conceptual knowledge.

Metacognitive Modeling; using cognitive modeling to clarify philosophical metacognitive concepts

Brendan Conway-Smith
Carleton University, Ottawa, Ontario, Canada

Robert West
Institute of Cognitive Science, Carleton University, Ottawa, Ontario, Canada

Abstract

Metacognitive research is integral to understanding cognition, but a problem persists metacognition remains poorly defined and its basic terminology contested. To address this problem, we propose a new philosophical method for understanding metacognition in a bottom up, computational way. We follow John Andersons principle that complex problems become systematic when analyzed within a cognitive model. Researchers agree that metacognition is cognition acting upon itself. Accepting this, we first define the fundamental units of cognition and then define how these units act upon themselves. We ground this within human cognition by using the Standard Model of Cognition (Laird et at. 2017, also known as the Common Model). This model defines the mechanisms common to all computational architectures modeling human cognition. Our model is then compared to metacognitive theories within psychology, philosophy, and neuroscience. This method clarifies metacognition by grounding it both within a computational cognitive architecture and present research literature.

Audio-Visual Integration: Point Light Gestures Influence Listeners Behavior

Susan Cook
University of Iowa, Iowa City, Iowa, United States

Abstract

Listeners are influenced by speakers hand gestures. However, it is not clear what processes support gesture processing. We investigated listeners behavior after observing speech with videotaped gestures or with point light gesture trajectories in the Tower of Hanoi task. Listeners were influenced by the synchrony of the visual and auditory information but not the nature of the information both videotaped and point light gestures reliably influenced behavior. Thus, visual information that is not perceived as produced by the speaker nonetheless reliably influences listeners behavior, so long as information is synchronized across modalities. Thus, observers do not appear to rely on functional or biological links between speech and hand gesture but rather on more general processes of multimodal integration. The principles underlying integration of auditory language with visual information from hand gestures appear to different from those underlying integration of auditory language and visual speech.

Scrape, rub, and roll: causal inference in the perception of sustained contact sounds

Maddie Cusimano
Massachusetts Institute of Technology, Cambridge, Massachusetts, United States

Josh McDermott
Massachusetts Institute of Technology, Cambridge, Massachusetts, United States

Abstract

We experience our soundscape in terms of physical events; for instance, a friend sweeping up after a plate crashed on the floor. The underlying perceptual inferences are typically ill-posed: without constraints, there are infinite possible causes of the observed sound. Thus, a core task for cognitive science is specifying the variables we perceive along with the constraints that allow them to be estimated. We identified sustained contact sounds (e.g., hands rubbing together, scraping a pan) as a rich domain with which to explore perceptual constraints. We developed a simple physics-based sound-synthesis model that can generate a diverse set of realistic scraping sounds. We find that listeners perceive the generative physical variables from scraping sounds, including velocity, motion trajectory, and surface roughness. Further experiments and acoustic analyses will address whether perception is constrained by a holistic generative model of sound or by invariant features that specify each perceived variable.

From wugged to wug: Reverse generalisation of stems from novel past tense verbs

Christine Cuskley
Newcastle University, Newcastle, United Kingdom

Stella Frank
University of Edinburgh, Edinburgh, United Kingdom

Kenny Smith
University of Edinburgh, Edinburgh, United Kingdom

Abstract

When native and non-native English speakers inflect novel verb forms for the past tense, non-natives are more likely to produce irregular (non -ed) forms than natives (Cuskley et al., 2015). We test whether participants can reverse-engineer the correct present tense stem from regular and irregular past tense forms of novel verbs. All participants are better able to identify the stem of regularly inflected forms than irregular forms, but we find no difference between native and non-native speakers. Phonological similarity to existing irregulars interferes with recognition of regularly inflected non-verbs (e.g., proximity of sleened to sling/slung makes it more difficult than drocked). While non-natives are more likely to produce irregular past tense forms, they are not better than native speakers at interpreting them. Non-native over-production of irregulars may reflect statistical patterns in their more limited input, but these factors do not seem to affect the process of inferring stems.

Pupillometry as a Measure of Effort Exertion in Cognitive Control Tasks

Kevin da Silva-Castanheira
McGill University, Montreal, Quebec, Canada

Myles LoParco
McGill University, Montreal, Quebec, Canada

A. Ross Otto
McGill University, Montreal, Quebec, Canada

Abstract

Despite recent interest in pupillometry as a psychophysiological measure, it remains unclear what construct the physiological measure is assessing in cognitive control tasks: task load or mental exertion. This debate is of particular interest as cognitive effort remains an elusive construct partly due to the difficulty in empirically quantifying mental exertion. The current research aims to differentiate these disparate accounts by leveraging rewards as motivation for effort exertion. Using an individual differences approach, a sample of 80 undergraduate students performed a cognitive control taskTask switching. Critically, monetary incentives were used to motivate participants to exercise cognitive control, and found to improve overall performance. Pupillary responses were found to increase in response to trials requiring more cognitive control, and relate to performance improvements in the rewarded conditions. The present findings provide some support for the effort account, and suggest that pupillometry may be a viable index of cognitive effort.

Contextual Effects in Value-Based Decision Making: A Resource-Rational Mechanistic Account

Kevin da Silva-Castanheira
McGill University, Montreal, Quebec, Canada

Ardavan S. Nobandegani
McGill University, Montreal, Quebec, Canada

Thomas Shultz
McGill University, Montreal, Quebec, Canada

A. Ross Otto
McGill University, Montreal, Quebec, Canada

Abstract

A wealth of experimental evidence shows that, contrary to normative models of choice, peoples preferences are markedly swayed by the context in which options are presented. Particularly, there exist a well-known triad of effects, dubbed the contextual effects, which consistently show that preferences change depending on the availability of other options: the attraction effect, the similarity effect, and the compromise effect. In this work, we present the first resource-rational, process-level account of these three contextual effects by extending Nobandegani et al.'s (2018) sample-based expected utility model to the realm of multi-attribute value-based decision-making. Importantly, our work is consisted with two empirically well-supported findings: (1) People tend to draw only a few samples in their probabilistic judgment and decision-making, and (2) People tend to overestimate the probability of extreme events in their judgment.

Towards building AI Life-coach agent for honing creativity

Amarnath Dasaka
IIIT hyderabad, Hyderabad, Telangana, India

Preeti S
Georgia Tech, Atlanta, Georgia, United States

Bapiraju Surampudi
IIIT hyderabad, Hyderabad, Telangana, India

Abstract

World Economic Forum report predicts that 35% of the skills needed to navigate the world of work will have changed by 2020. By 2020, creativity will be the third most sought-after skill, behind complex problem solving and critical thinking. Creative skills are future-proof, in that they cannot be Automated. Art and creativity are essentially what makes us human and this is being backed up by research. (Elaine Rumbol) How do you hone creativity? This seems to be an open question. The present study aims to build an architecture for AI agent(life-coach) that incorporates the latest research on creativity and guides the user based on the users personality traits, context, emotions, mood and cognitive load. The agent will detect the users emotional valance & Motivational Intensity which in turn will influence the attention focus (Broaden the mind (for free floating ideas) or result in narrow focus (linear, step by step goal attainment)). Toward this aim, we plan to run a series of tests for gathering user feedback. Design of the tests are underway.

The Jig-saw of Part-task Training in Dynamic Task Environments

Ropafadzo Denga
Rensselaer Polytechnic Institute, Troy, New York, United States

Wayne Gray
Rensselaer Polytechnic Institute, Troy, New York, United States

Abstract

Part-task training is a technique which involves separating the target task into parts and presenting them during training. This approach has been used to train users to perform optimally in dynamic task environments. The present study investigated the effects of fractionation, a part-task training approach, versus whole-task training to improve performance in the video game Tetris by focusing on an important sub-task element of the game. Seventy-eight young adults were trained on Tetris with one of three training regimens: 1) Part-task training with feedback, 2) Part-task training with no feedback, and 3) Whole-task training in which participants practiced the whole game to obtain the highest overall score. Results show that baseline performance influences training gains and feedback may not be helpful for learning. Training gains from the different training regimens show that tasks with highly interdependent components may benefit most from whole-task training.

Linguistic descriptions of action influence object perception: The role of action readiness

Victoria DiRubba
SUNY at Purchase, Purchase, New York, United States

Tommy Anderson
SUNY at Purchase, Purchase, New York, United States

Alexia Toskos Dils
SUNY at Purchase, Purchase, New York, United States

Abstract

Does hearing a story about performing an action activate corresponding motor representations? If so, can linguistically-activated motor representations affect our visual experience of the world? The present study tested whether hearing a story about performing power or precision grasps would cause people to perceive an ambiguous object in a grasp-congruent manner. Participants listened to a story in which they tossed water balloons either (1) without touching their knots (power grasp condition) or (2) by only touching their knots (precision grasp condition). Afterward, participants interpreted an object that could either be seen as an apple (power grasp) or cherry (precision grasp). To further manipulate participants availability for subsequent action in the story, participants either (1) had just grasped, (2) prepared to grasp, or (3) had repeatedly grasped the water balloons before the ambiguous image appeared. People perceived the object in a grasp-congruent manner only when their hands were available for action.

Modeling Causal Learning with the Linear Ballistic Accumulator

Yuhui Du
Ohio State University, Columbus, Ohio, United States

Nitisha Desai
Ohio State University, Columbus, Ohio, United States

Renlai Zhou
Nanjing University, Nanjing, Jiangsu, China

Abstract

Learning causal relationships is critical in our daily lives. To learn these causal relationships, one strategy we may use is the positive testing strategy (PTS), in which we attempt to confirm a hypothesis about the causal relationship. Also, we may use the expected information gain (EIG) strategy to distinguish between multiple hypotheses. Here we use an experimental paradigm in which subjects decide which of two causal patterns underlies a four-node causal system (Coenen, Rehder, & Gureckis, 2015) and fit the Linear Ballistic Accumulator (LBA) model to our data to investigate the precise mechanisms of different age groups using these strategies. We find that children and the elderly use PTS more than other groups. Yet, comparing drift rate and relative threshold parameters, we find no evidence for biases in strategy selection across age groups, but find that the elderly are more cautious when choosing a strategy.

Lying in public: Revealing the microstructure of real-time false responding through action dynamics

Nicholas D. Duran
Arizona State University, Glendale, Arizona, United States

Denis O'Hora
National University of Ireland Galway, Galway, Ireland

Sam Redfern
National University of Ireland Galway, Galway, Ireland

Arkady Zgonnikov
University of Aizu, Aizuwakamatsu, Fukushima, Japan

Abstract

It is commonly agreed that, in most scenarios, deception involves cognitive demands. Prime amongst these demands is competition between a default true response and an alternative false response. What is less understood are issues surrounding the mechanistic underpinnings of how and when this competition enacts its influence during responding. In previous work (Duran, Dale, & McNamara, 2011), we have used an action dynamics paradigm to capture millisecond-timing information in how people use their mouse movements to respond yes or no to autobiographical information. In the current study, we employed a similar paradigm to collect response data from hundreds of anonymous participants, who freely used an interactive touchscreen exhibit at a public science museum exhibit, aiming to replicate and extend our previous findings. As expected, during false responding, the truth appears to be initially activated and dissipates continuously over the course of the response.

The dark side of conceptual metaphor

Frank Durgin
Swarthmore college , Swarthmore, Pennsylvania, United States

Jessica Lewis
Swarthmore College, Swarthmore, Pennsylvania, United States

Abstract

Zhu (2017) used the implicit association test (IAT) to assess metaphorical alignment between concepts such as black and white and good and evil. Here we asked whether self-identified Black people have similar metaphoric alignments as those who identify as White. In an initial experiment, we tested pairwise metaphoric associations between black and white, dirty and clean, and good and evil. Measured strength of the 3 alignment pairings for these 3 sets of concepts was statistically the same among Black participants as that measured by Zhu for white participants. In a follow-up experiment, we compared self-identified Black and White participants IAT-scores for race (i.e., faces) and for color (i.e., chess pieces) IATs. For White participants, mean strength of white-positive alignment was identical for race and color; Black participants showed only slight white-positive bias for race IATs, and an intermediate level of white-positive bias for color IATs.

The role of affect in sentence perception

Veena Dwivedi
Brock University, ST CATHARINES, Ontario, Canada

Abstract

The role of affect and sentence processing is an understudied topic. In an event-related potential (ERP) language experiment, we investigated modulation of the P300 ERP component by dispositional affect. Using our previous ERP paradigm, we employed a 3x2 design where 32 participants read sentences presented in 1- and 2-word chunks (Berent et al., 2005; Patson & Warren, 2010). Sentences started with subject nouns that were either universally quantified or not, and continued with a direct object which was either indefinite, definite singular, or plural e.g., (i) Every kid climbed a tree/the tree/the trees vs. (ii) The kid climbed a tree/the tree/the trees. Number judgments were required at tree(s), which was always presented alone (and never final). Reduced P300 amplitudes were observed for the plural condition indicating interference; furthermore, low positive affect individuals showed responses sensitive to local high probability features associated with the control singular condition.

Do Humans Look Where Deep Convolutional Neural Networks "Attend"?

Anonymous CogSci submission

Abstract

Convolutional Neural Networks (CNNs) have recently begun to exhibit human level performance on some visual perception tasks. Performance remains relatively poor on vision tasks like object detection. We hypothesized that this gap is largely due to the fact that humans exhibit selective attention, while most object detection CNNs have no corresponding mechanism. We investigated some well-known attention mechanisms in the deep learning literature, identifying their weaknesses and leading us to propose a novel CNN approach to object detection: the Densely Connected Attention Model. We then measured human spatial attention, in the form of eye tracking data, during the performance of an analogous object detection task. By comparing the learned representations produced by various CNNs with that exhibited by human viewers, we identified some relative strengths and weaknesses of the examined attention mechanisms. The resulting comparisons provide insights into the relationship between CNN object detection systems and the human visual system.

Keywords: Visual Spatial Attention; Computer Vision; Convolutional Neural Networks; Densely Connected Attention Maps; Class Activation Maps; Sensitivity Analysis

Investigating bidirectionality of associations in young infants as an approach to the symbolic system

Milad Ekramnia
Neurospin, Gif sur Yvette, il de france, France

Ghislaine Dehaene
neurospin, Gif Sur Yvette, France

Abstract

Symbolic associations in human children and adults are based on forming equivalence classes which include three main relations between the tokens. 1) A = A (Reflexivity), if 2) A –¿ B and B –¿ C then A –¿ C (Transitivity) and 3) if A –¿ B then B –¿ A or Symmetry (1). Extensive studies on non-human primates have demonstrated success in Reflexivity and Transitivity in several species but a consistent failure in Symmetry in any given association. Comprehension of symmetry of an association can be a key contribution to linking abstract words to their corresponding tokens and later on in coupling writing forms of words to their spoken form (2). However to our knowledge it hasnt been investigated whether infants are capable of spontaneously reversing the direction of an association to any extent. In two EEG studies we investigated if 4.5-month-old infants are capable of applying symmetry in the context of word-learning.

In the first study we trained 2 groups of 25 infants, to two pairs of word-categories (bird or vehicle). At each trial infants were presented with a word and an image. The critical consideration was to introduce a 1 s of SOA between the two stimuli. In one group infants were trained on words always preceding the images (Word-Image group) and in the other group infants were trained on the opposite direction (Image-Word group). In the test blocks 70% of trials were as in the training and the other 30% were either with the incongruent trials in the original direction or the congruent and incongruent trials in the reversed direction. We observed significant cluster of electrodes, mainly in the right temporal, in both the trained and reversed directions while contrasting the congruent and incongruent conditions, with the word-image group showing a stronger effect.

In a 2nd experiment, designed as a comparative study between infants, adult humans and adult macaques, we sought to train each participant on 4 pairs of word-images, 2 pairs following a word-image direction and the other 2 an image-word direction, with a 1s SOA between the two stimuli similar to experiment 1. In this experiment the infants attended the training phase at home prior to the experiment through three YouTube videos on three consecutive days and on the test day, they were being tested either on the trained or the reversed direction of each single pair in a similar ERP design as in study 1. The results in a group of 54 4.5-month-old infants follow the pattern of results in study 1 that infants show an early as well as a late surprise effect relative to the onset of the second stimulus of the trial, while contrasting the incongruent versus congruent trials in both directions. Furthermore we utilized frequency tagging in both studies as an extra measure to compare the conditions of interest. The overall results suggest that contrary to the consistent failure of non-human animals, infants can readily learn an association in a bi-directional manner, which can be suggestive of an early access to their symbolic system.

1. Sidman and Tailby 1982 2. T. Medam, et al, Anim Cogn, 2016,

Visual exploration of emotional scenes in aging during a free visualization task depending on arousal level of scenes

PONCET Elie
Univ. Grenoble Alpes, CNRS, LPNC UMR 5105, 38000 Grenoble, France

NICOLAS Galle
Univ. Grenoble Alpes, CNRS, LPNC UMR 5105, 38000 Grenoble, France

Nathalie Guyader
Grenoble Alpes University, Grenoble, France

MORO Elena
Movement Disorders Unit, Department of Psychiatry Neurology and Neurological Rehabilitation CHU Grenoble, 38000 Grenoble, France

Aurlie CAMPAGNE
University Grenoble Alpes, Laboratory in Psychology and Neurocognition, CNRS, 38000 Grenoble, France

Abstract

Research on emotion suggests that the attentional preference observed toward the negative stimuli in young adults tends to disappear in normal aging and, sometimes, to shifts towards a preference for positive stimuli. However, this age-related effect called the positivity effect may be modulated by several factors, such as the arousal level of stimuli. The present study investigated visual exploration of natural scenes of different emotional valence in three age groups (young, middle-aged and older adults) depending on arousal level of scenes using an eye-tracking paradigm. Participants visualized pairs of emotional scenes either in low or high arousal condition. In contrast with the literature, the preliminary results revealed a reduction in prevalence of negative stimuli relative to other ones in older adults regardless of the arousal conditions. No difference between young adults and middle aged adults was observed.

Domestic dog understanding of containment and occlusion events

Julia Espinosa
University of Toronto, Toronto, Ontario, Canada

Daphna Buchsbaum
University of Toronto, Toronto, Ontario, Canada

Abstract

Intuitive physical concepts help humans navigate the world. One such concept, object containment, has been studied extensively in infants and nonhuman primates. Evidence indicates objects hidden inside of containers are more difficult to find than covered or occluded objects, possibly due to the prerequisite understanding that containers are hollow. Dogs encounter containers in daily life, and canine studies commonly require subjects to locate hidden treats. The present research provides the first test of the hypothesis that dogs, like primates, find it harder to make inferences about containment compared to other hiding events. To address this hypothesis, across 24 trials dogs (N=90) searched between 2 possible locations, one of which concealed a treat. They watched 3 different methods of hiding: i) inside containers, ii) behind containers, and iii) under containers. As predicted, dogs were less likely to locate treats inside containers. Results will be discussed in a comparative context.

Beyond divergent thinking: Measuring creative process and achievement in young children

Natalie Evans
Temple University, Philadelphia, Pennsylvania, United States

Abstract

Creativity is an elusive construct that is difficult to measure in children, and divergent thinking tasks have been overused and may be unreliable as measures of creativity (Baer, 2011). This study examines creative process and achievement in children using a problem-solving task (Daehler & Chen, 1993). Children (N=98) ages 4 to 6 tried removing a ball from a jar using common objects. Success with retrieving the ball was a measure of creative achievement. Creative process was assessed by coding creative behaviors such as object exploration, combinations, manipulation, and ball retrieval attempts. Results suggest differences in creative behaviors between successful and unsuccessful children. Successful participants created more unique object combinations ($p=0.02$), spent more time manipulating ($p=0.05$), and spent less time attempting to retrieve the ball ($p=0.02$) than unsuccessful children. Results suggest that this task moves beyond divergent thinking assessments by measuring both creative process and achievement in children.

Learned social values modulate representations of faces in the Fusiform Face Area

Ariana Familiar
University of Pennsylvania, Philadelphia, Pennsylvania, United States

Alice Xia
University of Pennsylvania, Philadelphia, Pennsylvania, United States

Sharon Thompson-Schill
University of Pennsylvania, Philadelphia, Pennsylvania, United States

Abstract

Social value processing has been shown to recruit specific neural systems, yet how they are associated with person-specific information, such as facial identity, processed in separate regions remains to be established. The present study examined changes in neural representations in face-selective visual areas due to social value learning. Over four days, participants learned combinations of social (generosity) and reward (point) values orthogonally assigned to naturalistic face images. We found that after learning, activity similarity (measured with fMRI) in the fusiform face area evoked by viewing the faces was related to social value as well as a measure of future social preferences, but was not related to reward value. This shows how learned social values can influence representations in face-selective brain regions thought to primarily encode visual information, and provides a potential neural mechanism for the association of social and visual information relevant to propensities in future social behavior.

Experimental conditions affect how social cues guide the regularisation of unpredictable variation

Olga Feher
University of Warwick, Coventry, Warwickshire, United Kingdom

Simon Kirby
University of Edinburgh, Edinburgh, United Kingdom

Kenny Smith
University of Edinburgh, Edinburgh, United Kingdom

Abstract

Unpredictable variation is widely used to investigate how cognitive and communicative biases impact on language evolution and change. Learning, interactive and cultural biases all contribute to universal linguistic patterns. We explored the effects of social cues using a miniature artificial language exhibiting unpredictable lexical variation distributed either within or between multiple speakers. We compared the effects of testing modality (spoken vs. forced-choice), experimental population (students vs. online workers) and setting (laboratory vs. online). Learners were sensitive to social cues, but reliable differences only emerged in the laboratory. In an online setting, students were much more likely to regularise across conditions. In addition, task difficulty increased rates of regularisation but only online. Online workers showed high levels of regularisation throughout. Our experiments suggest that the conditions in which learning and recall take place have a large impact on the biases which shape language and our ability to measure them.

Improv exercises promote uncertainty tolerance and improve creativity outcomes

Peter Felsman
University of Michigan, Social Work and Psychology, Ann Arbor, Michigan, United States

Sanuri Gunawardena
University of Michigan, Ann Arbor, Michigan, United States

Colleen Seifert
U Michigan, Ann Arbor, Michigan, United States

Abstract

Improvisational theater is defined broadly as a theatrical setting in which, process and product co-occur (Sowden, Clements, Redlich, & Lewis, 2015). Therefore, practicing improvisational theater involves embracing uncertainty (Napier, 2004). In this context, individuals may learn to tolerate uncertainty with greater comfort, a common treatment outcome across many psychological disorders (e.g. Boswell et al., 2013). The current study employs a lab-based paradigm linking brief improvisational theater experience to increased divergent thinking outcomes (Lewis & Lovatt, 2013). We set out to replicate and extend this finding by including an explicit measure of uncertainty tolerance. Across two studies, our results show increased uncertainty tolerance for people who improvised, significantly more than people who participated in a social interaction control with limited uncertainty. Additionally, the improvising condition predicted relative improvement on a subset of divergent thinking measures, offering partial support for the Lewis and Lovatt (2013) finding that improvisational theater exercises can improve creativity.

Space Matters: Investigating the influence of spatial information on subjective time perception

Can Fenerci
McGill University, Montreal, Quebec, Canada

Myles LoParco
McGill University, Montreal, Quebec, Canada

Kevin da Silva-Castanheira
McGill University, Montreal, Quebec, Canada

Signy Sheldon
McGill University, Montreal, Quebec, Canada

Abstract

Although understood that time perception is subjective, the underlying cognitive mechanisms are not well described. Event segmentation theories propose that spatial information serves to segment experienced information in discrete units which then can be used to estimate time. Based on this theory, we explored whether subjective time perception is influenced by the amount of perceived spatial information. A group of young participants viewed short videos of episodes that included a spatial change (e.g., moving through doorways) or no spatial change. In one experiment, participants were asked to estimate a given time duration while viewing the video and in a second experiment, participants estimated the time of the video after viewing. Across experiments, videos with spatial change were associated with more accurate time perception estimates than those without spatial changes. These results highlight the important role of spatial processing in directing the experience of time.

No Morphological Markers, No Problem: ERP Study Reveals Semantic Factors Differentiating Neural Mechanisms of Noun and Verb Processing

Jun Feng[1], Tao Gong[2], Lan Shuai[2], Yicheng Wu[3]

[1] Hangzhou Normal University

[2] Educational Testing Service

[3] Zhejiang University

Abstract

Neural mechanisms behind noun and verb processing are ubiquitously separate, yet it remains controversial which factor, syntax or semantics, is behind such separation. We conducted an ERP study using Chinese sentences with a specific construction, noun phrase + mei ("not/no") + noun/verb/noun-verb-ambiguous-word, and excluding other grammatical or syntactic factors that could hint at the target words' part-of-speech. Results showed significantly distinct P200, N400 and P600 between noun and verb processing in native speakers, indicating that semantic factors are essential for the differentiated neural mechanisms behind noun and verb processing. Similar results were also found between noun-verb-ambiguous-word and noun processing, but not between noun-verb-ambiguous-word and verb processing, suggesting that lacking clues on part-of-speech makes the dynamic properties of the ambiguous words more salient than the static ones, thus causing interpretation of such words more likely as verbs. This further elaborates the crucial role of semantic factors in noun and verb processing.

The impact of frequency on the evolution of category systems

Vanessa Ferdinand
University of Melbourne, Melbourne, VIC, Australia

Charles Kemp
University of Melbourne, Melbourne, VIC, Australia

Amy Perfors
University of Melbourne, Melbourne, VIC, Australia

Abstract

How do category systems reflect the information content of their environments? One basic kind of information in a linguistic environment is the frequency of objects or meanings: some things are just spoken about more often than others. A great deal is known about frequency effects on the evolution of lexical items (e.g. Lieberman et al, 2007); however analogous effects on category systems are not understood. Two theories point in opposite directions: the generalized context model (Nosofsky, 2011) predicts that categories containing high-frequency items will expand over time, while information theory (Cover & Thomas, 2012) predicts tighter boundaries around high-frequency items. We explore the impact of frequency on the evolution of category systems over time in an iterated category learning experiment that manipulates object frequency. How does this manipulation affect category boundaries? Does the result change if transmission is between different individuals or within the same person over time?

Cover, T. M., & Thomas, J. A. (2012). Elements of information theory. John Wiley & Sons.

Lieberman, E., Michel, J. B., Jackson, J., Tang, T., & Nowak, M. A. (2007). Quantifying the evolutionary dynamics of language. Nature, 449(7163), 713.

Nosofsky, R. M. (2011). The generalized context model: An exemplar model of classification. Formal approaches in categorization, 18-39.

How victim framing shapes attitudes towards sexual assault

Stephen Flusberg
Purchase College, SUNY, Purchase, New York, United States

Sarah Husney
Colorado College, Colorado Springs, Colorado, United States

Casey Pollard
Colorado College, Colorado Springs, Colorado, United States

Kevin Holmes
Colorado College, Colorado Springs, Colorado, United States

Abstract

Crimes typically involve a perpetrator and a victim, but alleged perpetrators are often cast as the true victim, as happened recently in the case of U.S. Supreme Court nominee Brett Kavanaugh. Across two experiments, we investigated the efficacy of this type of victim framing. Participants read a brief report about an alleged college campus sexual assault and expressed their support for the male and female protagonists. The report either framed the woman as the victim (of sexual assault), the man as the victim (of false accusations), or was relatively neutral about victimhood (baseline control). Relative to baseline, the framing manipulation was effective at eliciting more support for the character described as a victim, regardless of participants gender or political affiliation. These findings suggest that the language of victimhood, or its co-opting to cast alleged perpetrators in a more favorable light, can shape public opinion about a politically polarized issue.

Language stability and change in age-dependent networks

Stella Frank
University of Edinburgh, Edinburgh, United Kingdom

Kenny Smith
University of Edinburgh, Edinburgh, United Kingdom

Abstract

People's social and linguistic environment changes over the course of their life: infants learn language from a small set of caregivers; children and adolescents practice language skills with their peers; adults speak to other adults and also pass on their language to the next generation (Kerswill, 1996, Sankoff 2018). Population models of language change have explored network effects but neglected changing networks as a function of agent age. We model a population of Bayesian agents that go through life phases of initial learning, subsequent peer interactions, and transmission to the next generation. We find these age-dependent networks to be more stable than other network architectures. This stability counters previous Bayesian modelling results in which languages reliably and rapidly change, converging to the learners prior, suggesting that languages spoken in populations in which interactions are organised assortatively by age may only weakly reflect human priors on language learning.

Investigating the factorial structure of widespread false beliefs

Vincent Frigo
UW-Madison, Madison, Wisconsin, United States

Timothy Rogers
University of Wisconsin- Madison, Madison, Wisconsin, United States

Abstract

Cognitive science often views human learning as rational. Why then do false beliefs arise, and why are they resistant to change? False beliefs might arise when people (1) lack knowledge in some domain, (2) adopt beliefs aligning with implicit causal theories, or (3) encounter, through media or social networks, sets of beliefs that strongly covary. To test these hypotheses we composed a survey assessing beliefs about matters of fact across a wide range of knowledge domains and collected responses from 500 MTurkers. We then conducted a factor analysis to determine which false beliefs co-vary together, clustered respondents to find groups that adopt comparable false belief sets, and used regression to identify sociodemographic and media-consumption features that predict susceptibility to different kinds of false beliefs. The results suggest that some kinds of false belief may arise and persist merely from covariance in the opinions learners encounter in social life.

Inflated inflation and superseded supersession: testing counterfactual sampling accounts of causal strength judgments

Maureen Gill
Princeton University , Princeton, New Jersey, United States

Jonathan Kominsky
Harvard University, Cambridge, Massachusetts, United States

Joshua Knobe
Yale University, New Haven, Connecticut, United States

Thomas Icard
Stanford University, Stanford, California, United States

Abstract

Norm violations have been shown to influence causal judgments. Icard, Kominsky, and Knobe (2017) explained the influence of norms by appeal to a model of norm-weighted sampling of counterfactual possibilities. This model explains two well-known effects (among others): When two agents must act to bring about an outcome (i.e. both actions are necessary), if an agent S violates a norm, they are judged more causal than when they do not violate a norm (abnormal inflation), and the other agent B is judged to be less causal than when S does not violate a norm (causal supersession). In the present study (N = 1008), we find empirical support for two untested further predictions of this sampling model of causal strength judgments: Abnormal inflation of S is greater when B violates a norm (inflation increase), and causal supersession of B is smaller when S violates a norm (supersession decrease).

Spatial-Numeric Associations Distort Estimates of Causal Strength

Kelly Goedert
Seton Hall University, South Orange, New Jersey, United States

Daniel W. Czarnowski
Seton Hall University, South Orange, New Jersey, United States

Abstract

When individuals provide magnitude estimates using numeric scales, they may be influenced by spatio-numeric biases. In Western, English-speaking cultures smaller magnitudes are associated with the left side of space and larger with the right. We demonstrated the impact of spatial-numeric associations on judgments of causal strength in two trial-by-trial causal learning experiments. Causes appeared on either the left or right side of a computer screen. In Experiment 1, participants made casual judgments using a number line either increasing in magnitude from left to right or decreasing in magnitude from left to right. In Experiment 2, participants made judgments using a non-linear circular target with the depth of hue saturation representing causal strength. In Experiment 1, participants gave higher causal ratings to causes appearing in the space associated with larger numbers on the number line. These influences disappeared when the linearity of spatial-numeric associations was removed in Experiment 2.

Can children develop novel tools to solve problems via analogical generalization? Kind of!

Micah Goldwater
University of Sydney, Sydney, Australia

Abstract

Recent research has examined whether children can modify tools to solve novel problems. For example, when children are given a pipe cleaner with the goal to retrieve a little bucket at the bottom of a tube, will they realize that bending the pipe cleaner into a hook will solve the problem? Children younger than 7 almost all fail at this task, and children under 10 are far from ceiling. Because problem solving is often helped via generalization from analogous problems, the current study examined whether children in this task could take advantage of being read a story (with pictures) about fishing, emphasising the importance of hooks. Interesting we found an interaction wherein preschool children were helped by the analogy, while school-aged children were not, who also solved the task at much higher rates overall (but still below ceiling).

Detecting Students Problem Solving Strategies Using Sankey Diagrams

Tao Gong
Educational Testing Service, Princenton, New Jersey, United States

Christopher Agard
Educational Testing Service, Princeton, New Jersey, United States

Gary Feng
Educational Testing Service, Princeton, New Jersey, United States

Gabrielle Cayton-Hodges
Educational Testing Service, Princeton, New Jersey, United States

Luis Saldivia
Educational Testing Service, Princeton, New Jersey, United States

Abstract

Process data (e.g., logs of actions, keystrokes, times, or eye tracks) recording students interactions with digital assessments are available in many digital educational assessments. They have become the primary focus of cognitive scientists to detect and analyze students strategies during problem solving. This study developed a Sankey diagram-based method to visualize process data of multiple-choice items. Such diagram has been widely adopted in industry and ecology to trace flow of information, energy, or resource. Using released items from the 2017 National Assessment of Educational Progress Mathematics Tests, we illustrated how to use such a diagram to elucidate frequent answer formulation patterns of students, their common mistakes, and estimated probabilities of reaching correct/wrong answers at various answering stages. These help reveal the problem solving strategies adopted by students and their underlying cognitive processes. Assessment developers, teachers, and students could use such insights to improve assessments and learning outcomes for confusing concepts.

Evaluating systematicity in neural networks with natural language inference

Emily Goodwin
McGill University, Montreal, Quebec, Canada

Koustuv Sinha
McGill University, Montreal, Quebec, Canada

Timothy ODonnell
McGill University, Montreal, Quebec, Canada

Abstract

Compositionality makes linguistic creativity possible. By combining words, we can express uncountably many thoughts; by learning new words, we can extend the system and express a vast number of new thoughts. Recently, a number of studies have questioned the ability of neural networks to generalize compositionally (Dasgupta, Guo, Gershman & Goodman, 2018). We extend this line of work by systematically investigating the way in which these systems generalize novel words.

In the setting of a simple system for natural language inference, natural logic (McCartney & Manning, 2007), we systematically explore the generalization capabilities of various neural network architectures. We identify several key properties of a compositional system, and develop metrics to test them. We show that these architectures do not generalize in human-like ways, lacking inductive leaps characteristic of human learning.

Experimental Study on the Decision Making process in a Centipede Game

Dhriti Goyal
International Institute of Infromation Technology, Hyderabad, India, Hyderabad, India

Dhiraj Jagadale
International institute of information technology, Hyderabad, Hyderabad, Telangana, India

Kavita Vemuri
International Institute of Information Technology - Hyderabad, Hyderabad, Telangana, India

Abstract

The studys objective was to measure the somatic state response (skin conductance and heart rate) and understand the decision making processes in a two-player Centipede game, an extensive form game, with a modified payoff. The experiment included fixed and random termination for analyzing the effect of players mutual trust on risk-taking behavior. The behavioral results reveal that trust controls the game rounds (that is, the number of pass decisions) in known or random termination game conditions, though the exit points were higher in the former compared to the latter condition. Higher skin conductance and heart rate during the game-play is noticed as compared to the baseline data showing anxiety during the gameplay and interestingly opponents action induced higher skin conductance amplitude than during self-play for the same decision. The data provides strong preliminary evidence of trust influencing cooperative gameplay.

Optimal categorisation: the nature of nominal classification systems

Alexandra Grandison
University of Surrey, Guildford, United Kingdom

Michael Franjieh
University of Surrey, Guildford, United Kingdom

Greville Corbett
University of Surrey, Guildford, United Kingdom

Abstract

Effective categorisation should be simple, to minimise cognitive load, and informative, to maximise communicative efficiency. Nominal classification systems (gender, classifiers) are a functional means of categorisation that vary enormously across languages, revealing a trade-off between simplicity and informativeness. Closely related Oceanic languages of Melanesia show staggering variation in their number and type of classifiers. How does the Iaai language carve up nouns into 23 semantic groups whilst the Merei language uses only two; and what implications do these vastly different systems have for the cognitive representations of their related concepts? We combined typological enquiry and psycholinguistic experimentation (free listing, card sorting, video vignettes, possessive labelling, eye tracking, storyboards, category training) comparing nominal classification systems in six Oceanic languages of Vanuatu and New Caledonia. We discuss how these experiments uncover the nature of nominal classification systems, comparing objective data across languages and experimental contexts to reveal a model for optimal categorisation.

Do You Need More than Two Subjects: Using Cognitive Modeling to Make Accurate Predictions for Individual Subjects

Emily Greve
Carleton University, Ottawa, Ontario, Canada

Elisabeth Reid
Carleton University, Ottawa, Ontario, Canada

Robert West
Carleton University, Ottawa, Ontario, Canada

Abstract

In experimental research, large numbers of participants are used to average out individual differences in the data. However, differences in task performance may be largely due to two factors; lack of task training, and different micro-strategies. We implement a methodology that removes the effect of these factors, requires only 23 participants, and still produces large amounts of data. Other studies have been published using a similar methodology (Cousineau & Shiffrin, 2004; Gray & Boehm-Davis, 2000). Our study is a revision of previous research using a mobile game (West et al., 2018). Participants are trained extensively on the game to ensure they are experts. The study includes a predictive cognitive model and the game-design is based on an apparent micro-strategy. We hypothesize that the same micro-strategies under identical conditions, should produce identical results across participants and the model. Suggesting the model may exist in the mind of human experts.

Language facilitates 2.5-year-olds reasoning by the disjunctive syllogism

Myrto Grigoroglou
University of Toronto, Toronto, Ontario, Canada

Sharon Chan
University of Toronto, Toronto, Ontario, Canada

Patricia Ganea
University of Toronto, Toronto, Ontario, Canada

Abstract

Children and animals successfully reason by elimination: if a reward is hidden in A or B, and they see A empty, they search in B (Call, 2004; Hill et al., 2012). Twenty-seven-month-olds also solve similar tasks when emptiness is conveyed verbally, through negation (The toy is not in the box, Feiman et al., 2017). However, it is unclear whether participants solved these tasks with the disjunctive syllogism (A OR B, NOT A, THEREFORE B); in a 4-cup paradigm requiring disjunctive reasoning only 3-5-year-olds but not 2.5-year-olds succeeded (Mody & Carey, 2016). We used a linguistic version of the 4-cup task to examine childrens ability to reason disjunctively using verbal negation. We found that 3- and 2.5-year-olds performed significantly above chance (58.1%, 54.2%, respectively, ps¡.05). Thus, presenting the negative premise verbally facilitated 2.5-year-olds deductions. We conclude that older 2-year-olds have a robust understanding of negation, which they apply in abstract reasoning.

Exploring cognitive states through real-time classification and sonification of brain data

Yann Harel
Universit de Montral, Montral, Quebec, Canada

Antoine Bellemare
Concordia University, Montreal, Quebec, Canada

Arthur Dehgan
Universit de Montral, Montral, Quebec, Canada

Anne-Lise Saive
Universit de Montral, Montral, Quebec, Canada

Karim Jerbi
Universit de Montral, Montral, Quebec, Canada

Abstract

With the recent advances in EEG technology and the popularization of low-cost mobile EEG devices, brain-computer interface (BCI) systems and neurofeedback tools have become more accessible. Real-time EEG signal processing is increasingly popular in the context of digital arts projects powered by a neuroaesthetic approach. CoCo Brain Channel is one such project : designed to use real-time processing of EEG signal in order to generate a musical environment, it provides the user with a means to hear and control his own brain activity. This is achieved by hooking-up a commercial mobile EEG device to a music generation algorithm built in PureData. The generative algorithm uses features from EEG signals to modulate harmonic and rhythmic structures of multiple oscillators. The result is a continuous musical soundscape reflecting the evolution of EEG signals. Improvements and possible applications for basic research will be discussed.

When circumstances change, update your pronouns

Joshua K. Hartshorne
Boston College, Chestnut Hill, Massachusetts, United States

Mariela V Jennings
Boston College, Chestnut Hill, Massachusetts, United States

Tobias Gerstenberg
Stanford University, Stanford, California, United States

Josh Tenenbaum
MIT, Cambridge, Massachusetts, United States

Abstract

Language is frequently ambiguous, with the same sentence having several possibleinterpretations. One prevalent example is third-person pronouns. Hartshorne, Gerstenberg, & Tenenbaum (2014) HGT2014 model pronoun interpretation as an inference over a generative model of the speaker. An advantage of the generative intuitive theory approach is that it incorporates a flexible, quantitative model of world knowledge rather than a list of facts and heuristics. The authors formalized this world knowledge as inference over a generative model of the world. We directly test this flexibility by changing the rules of the world (e.g., through scenarios that reverse the normal relationship between strength and probability of winning tug-of-war), which according to HGT2014 should directly affect pronoun interpretation. We find that model predictions and participant judgments align well in such scenarios, supporting HGT2014 and challenging other theories of pronoun resolution. We discuss this work in the context of recent work on intuitive theories.

Strategy shifting in navigation: Insights from trial-level effects in a virtual navigation task

Chuanxiuyue He
University of California, Santa Barbara, Santa Barbara, California, United States

Alexander Boone
University of California, Santa Barbara, Santa Barbara, California, United States

Mary Hegarty
University of California, Santa Barbara, Santa Barbara, California, United States

Abstract

In the dual-solution paradigm (DSP), people learn a route through a virtual environment. After learning, people are asked to navigate to locations in the environment. Individuals vary in the degree to which they rely on the learned route (response strategy) versus a shortcut (place strategy). The present study characterizes trial-level features such as relative target locations, Euclidean distance and number of turns or intersections between locations, and uses a Rasch Model to investigate how spatial attributes of these trials influence participants strategy-choice. Additionally, a post-task questionnaire shows a partial disassociation between navigation behaviors in the virtual environment and navigation in daily life. It is proposed that this dissociation can be explained by differences in environment features. This study has unique potential to advance understanding of factors that affect navigation strategy choice, and to inform ecological validity of the Dual Solution Paradigm and other navigation paradigms.

Explaining without Information: The Role of Label Entrenchment

Babak Hemmatian
Brown University, Providence, Rhode Island, United States

Steven Sloman
Brown University, Providence, Rhode Island, United States

Abstract

In categorical explanation a category label is used to explain an associated property. We show that label entrenchment, whether a label is commonly used by ones community, affects the judged quality of a categorical explanation whether the explanation offers substantive information or not. In Experiments 1 and 2, explanations using unentrenched labels are rated as less comprehensive and less natural independent of causal or featural information, even when the label is merely a name for the explanandum. Experiments 3 and 4 replicate the effect with unentrenched labels coined by groups of expert discoverers and rule out explanations like familiarity and communicative principles. Most participants in Experiments 3 and 4 could not report the impact of entrenchment on their judgments. We argue that reliance on entrenchment arose because the community often has useful information. Common use of labels as conduits for this knowledge induces reliance on community cues even when uninformative.

Consequential Consensus: A Decade of Online Discourse about Same-sex Marriage

Babak Hemmatian
Brown University, Providence, Rhode Island, United States

Sabina Sloman
Carnegie Mellon University, Pittsburgh, Pennsylvania, United States

Uriel CohenPriva
Brown University, Providence, Rhode Island, United States

Steven Sloman
Brown University, Providence, Rhode Island, United States

Abstract

Framing issues as matters of non-negotiable values can increase the perceived intractability of debates. Focusing on the concrete consequences of policies instead can facilitate conflict resolution. Using a topic model of Reddit comments from January 2006 to September 2017, we show that the contribution of certain topics concerned with protected values to the debate increased prior to the emergence of a public consensus in support of same-sex marriage and declined afterwards. These topics related to religious arguments and freedom of opinion. In contrast, discussion of certain concrete consequences (the impact of politicians stances and policy implications) showed the opposite pattern, their increased prominence coinciding with improved public support for same-sex marriage after 2012. Our results reinforce the meaningfulness of protected values and consequentialism as relevant dimensions for describing public discourse and highlight the usefulness of unsupervised machine learning methods in tackling questions about social attitude change.

Untangling indices of emotion in music using neural networks

Dorien Herremans
Singapore University of Technology and Design, Singapore, Singapore

Kin Wai Cheuk
Singapore University of Technology and Design, Singapore, Singapore

Yin-Jyun Luo
Singapore University of Technology and Design, Singapore, Singapore

Kat Agres
Institute of High Performance Computing, A*STAR, Singapore, – Select State/Province –, Singapore

Abstract

Emotion and music are intrinsically connected, and researchers have had limited success in employing computational models to predict perceived emotion in music. Here, we use computational dimension reduction techniques to discover meaningful representations of music. For static emotion prediction, i.e., predicting one valence/arousal value for each 45s musical excerpt, we explore the use of triplet neural networks for discovering a representation that differentiates emotions more effectively. This reduced representation is then used in a classification model, which outperforms the original model trained on raw audio. For dynamic emotion prediction, i.e., predicting one valence/arousal value every 500ms, we examine how meaningful representations can be learned through a variational autoencoder (a state-of-the-art architecture effective in untangling information-rich structures in noisy signals). Although vastly reduced in dimensionality, our model achieves state-of-the-art performance for emotion prediction accuracy. This approach enables us to identify which features underlie emotion content in music.

Emotion attributions echo the structure of people's intuitive theory of psychology

Sean Houlihan
MIT, Cambridge, Massachusetts, United States

Max Kleiman-Weiner
Harvard, Cambridge, Massachusetts, United States

Josh Tenenbaum
MIT, Cambridge, Massachusetts, United States

Rebecca Saxe
MIT, Cambridge, Massachusetts, United States

Abstract

We present a generative model of how observers think about the emotions experienced by players in a socially-charged game: a public, high-stakes, one-shot Prisoner's Dilemma. The model extends inverse planning frameworks to capture observers' judgments about players' reactions to hypothetical events. Observers attribute different beliefs and values to players based on what decisions the players make. We model how observers' noisy inferences of players' mental contents bias emotion predictions. Incorporation of non-monetary features into forward planning enables us to model emotions that reflect complex social concerns (e.g. Embarrassment depends on how much players think others will infer that they tried to take advantage of their opponents). In addition to matching the intensities of twenty attributed emotions, the model reflects how observers' emotion judgments covary within single stimuli, indicating that the model captures important aspects of the generative process underlying humans' emotion attributions in this game.

The Intervention of Affective and Cognitive Theory of Mind on Impacting Social Norm Violation Judgements

Nai Ching Hsiao
National Cheng Kung University, Tainan, Taiwan

Jon-Fan Hu
National Cheng Kung University, Tainan City, Taiwan

Abstract

Individual's judgment on the appropriateness of social norm includes perceiving others mental states (theory of mind), but it might differ with the intervention aspects in real social contexts. Therefore, in this study we mainly focus on evaluating whether affective and cognitive theory of mind would affect social norm violation judgments and investigate whether the timing of mentalization involves the judgments. As a result, preconceived intention intervention (both affective and cognitive theory of mind) significantly affected the judgments of the appropriateness. However, only cognitive theory of mind in attributing violation intentions after encountering the social norm statement was found to affect in the judgments of the appropriateness of norm violations. In summary, theory of mind plays an important role on the judgment of appropriateness for social norm violation, but the timing of intervention matters significantly.

A tool to analyze verb phrase and noun phrase relationship in sentences

Te-En Huang
National Cheng Kung University, Tainan, Taiwan

Tao-Hsing Chang
National Kaohsiung University of Science and Technology, Kaohsiung City, Taiwan

ADAT Technology Co.
ADAT Technology Co., Tainan, Taiwan

Jon-Fan Hu
National Cheng Kung University, Tainan City, Taiwan

Abstract

SPACY is a well-known package for NLP analysis for delineating the Verb phrases and Direct Objects in English by applying the default structures to define noun phrase. However, SPACY lacks a function to include the status of adjectives and vast amount of noun phrase structures for identifying the relationship between Verbs and Nouns efficiently. The present study develops a SPACY-based program to customize practical noun phrase structures written in industrial SOPs for machine operations. It performs better at merging overlapping structures, for example, a sentence An important thing of NLP is hard to define can be processed to be An important thing, NLP, thing of NLP; and then automatically merged into one noun phrase An important thing of NLP. The capacity of the program can abstract the core concepts of sentences and recognize the co-occurrences of noun phrases and their associated verbs from the corpus for research and application purposes.

Examining Prefrontal Cortex Contributions to Creative Problem Solving With Noninvasive Electric Brain Stimulation

Kent Hubert
Drexel University, Philadelphia, Pennsylvania, United States

Evangelia G. Chrysikou
Drexel University, Philadelphia, Pennsylvania, United States

Abstract

Cognitive neuroscience studies of creativity typically employ divergent thinking tasks that prioritize bottom-up processes to generate novel responses. However, real-world creative problem solving is guided by top-down thinking that puts an emphasis on the goal to be achieved. Here, we introduce the Alternative Objects Task (AOT)a novel task that incorporates both bottom-up and down-down thought during problem solving. Guided by functional neuroimaging findings, we employed transcranial direct current stimulation (tDCS) over frontopolar cortex to investigate causally the impact of transient changes in activity in this region for problem solving performance on the AOT. Participants were presented with a series of goals and generated either a common or an uncommon object that could satisfy each, while undergoing either excitatory (anodal) or sham tDCS. Analyses of accuracy, reaction times, and semantic distance highlight the importance of goal-orientation during creative problem solving and its reliance on prefrontal cortex.

A Two-Process Model of Semantic Development

Philip Huebner
University of Illinois, Urbana-Champaign, Urbana, Illinois, United States

Jon Willits
University of Illinois at Urbana-Champaign, Champaign, Illinois, United States

Abstract

How do children acquire semantic knowledge? In this work, we explore an old answer to this question: Semantic development is a hybrid of two distinct processes. The first process involves unsupervised learning of relations between objects, providing a representation of objects that is useful for a wide range of possible goals. The second process involves explicitly learning to put objects and their relations into categories. Critically, this second process uses the representations of the first process as its starting point. Here, we demonstrate this using a two-process model, where the first process is a distributional semantic model (e.g. HAL, Word2Vec, RNN), and the second process is a transformation of representations learned during process 1 into a task-specific target space. This approach improves performance on multiple semantic tasks, compared to using the representations learned by process 1 directly. We believe this model demonstrates that a task- or goal-oriented perspective of semantic cognition has promise for furthering our understanding of semantic development.

The Relationship between Inhibitory control and Creativity

tal ivancovsky
Bar Ilan University, Ramat Gan, Israel

Moshe Bar
Bar Ilan University, Ramat Gan, Israel

Abstract

There is a debate in the literature as to whether inhibitory control improves or hinders creativity. Alternatively, we propose that flexible alterations between these two states would actually benefit creativity best. Therefore, the purpose of the current study was to resolve the debate by inducing inhibited/disinhibited/flexible states of mind and subsequently examine the influence on creative performance. To do so, the Stop-Signal task (SST) was deployed through the use of differential task instructions. Afterwards, participants completed two creativity tasks: a free association task (FAT) and the alternate uses task (AUT). Results indicated that while the inhibited group scored higher in the FAT, the flexible group scored higher in the AUT. Based on the results, we propose that there is an inverted U-shaped relationship between inhibitory control and creativity: while some cognitive control is needed to generate original ideas; excessive control might hinder creativity as it may lead to premature closure of ideas that could otherwise be further developed.

Does Motor Engagement Influence Memory for STEM Abstract Concepts?

Constanza Jacial
Drexel University, Philadelphia, Pennsylvania, United States

Evangelia G. Chrysikou
Drexel University, Philadelphia, Pennsylvania, United States

Abstract

Theories of embodied cognition have suggested that motor activity may influence the consolidation of conceptual knowledge. In line with this prediction, behavioral studies have shown retrieval interference effects of a manual motor task for manipulable object concepts. On the other hand, research investigating such effects for abstract concepts is limited. Here, we examined in a behavioral experiment potential effects of the recruitment of the motor system for the consolidation of different kinds of abstract concepts. Participants were presented auditorily and asked to memorize abstract concepts with movement referents (e.g., fluidity), abstract concepts without movement referents (e.g., theory), and concrete concepts (e.g., microscope) while engaging in a full-body motor task. All concepts were specific to Science Technology Engineering and Mathematics (STEM) disciplines. Analysis of free recall and recognition performance suggests influence of motor engagement for certain types of STEM concepts during memory encoding and subsequent retrieval.

Symbol grounding boosts transfer in addition learning

Clint Jensen
University of Wisconsin - Madison, Madison, Wisconsin, United States

April D. Murphy
University of Wisconsin - Madison, Madison, Wisconsin, United States

Andrew Young
Occidental College, Los Angeles, California, United States

Martha Alibali
University of Wisconsin-Madison, Madison, Wisconsin, United States

Timothy Rogers
University of Wisconsin- Madison, Madison, Wisconsin, United States

Chuck Kalish
University of Wisconsin-Madison, Madison, Wisconsin, United States

Abstract

Early math instruction often prioritizes rapid retrieval of mathematical facts, (e.g. 4 + 6 = __; 10), an approach that promotes quick recall of sums but with limited transfer to unstudied problems. We consider how this pattern changes when the learning scenario highlights the quantities that underlie symbols. Adult participants learned a novel base 8 addition task using alphabetic symbols to indicate quantities (e.g. AG + AF = __). They practiced with symbols only or with symbols grounded in quantitative representations. When tested in the same format as participants were trained, studied problems were learned equally well but symbol-only learners transferred only to identical-elements problems (e.g. AG + AF transferred to AF + AG). Grounded learners showed better transfer to problems involving novel quantities. The results suggest, in contradiction to some other recent findings, that arithmetic transfer is boosted when the learning scenario highlights quantitative meaning denoted by number symbols.

Boundedness in event and object cognition

Yue Ji
University of Delaware, Newark, Delaware, United States

Anna Papafragou
Unversity of Delaware, Newark, Delaware, United States

Abstract

The semantic property of boundedness characterizes the presence of well-defined spatio-temporal boundaries for events or objects in language (Bach, 1986; Frawly, 1992; Jackendoff, 1991). Little research has tested whether this property actually characterizes event and object cognition (but see Wellwood, Hespos, & Rips, 2018). We showed participants videos of bounded events where a salient change in state of the affected object(s) occurred (e.g., dressing a teddy bear) and unbounded events that lacked a salient change (e.g., waving a handkerchief). Participants decided whether a video matched with a picture of a single novel object or a picture of a novel substance (object/substance pictures were adopted from Li, Dunham, & Carey, (2009)). Participants tended to pair a bounded event with an object and an unbounded event with a substance, and were in fact better at establishing the former connection. We conclude that boundedness underlies the cognitive representation of both events and objects.

Pupillometry measures of cognitive load in meta-T dynamic task environment

Chris Joanis
Rensselaer Polytechnic Institute, Troy, New York, United States

Evan Pierce
Rensselaer Polytechnic Institute, Troy, New York, United States

Wayne Gray
Rensselaer Polytechnic Institute, Troy, New York, United States

Abstract

Pupillometry uses pupil diameter as a physiological measure of cognitive effort and load. In static tasks, pupillometry has revealed that cognitive effort varies with expertise, and, combined with gaze analysis, shows that experts can exert effort to focus on non-salient visual input. Much real-life expertise is practiced in dynamic tasks, and expert effort in dynamic tasks remains unstudied. Using tetris as a dynamic task environment, we collected pupil and gameplay data from individuals of varying expertise levels. We then use collected data and examine cognitive workload differences across levels of expertise. Consistent with studies of image saliency and gaze, our results indicate that experts and novices engage differently with the task and do not experience the same cognitive workload. Further inspection will likely reveal strategy-level sources of these differences.

Equanimity moderates approach/avoidance motor-responses and evaluative conditioning

Catherine Juneau
LAPSCO, Clermont-Ferrand, France

Laurent Waroquier
LAPSCO, Clermont-Ferrand, France

Michael Dambrun
LAPSCO, Clermont-Ferrand, France

Abstract

A growing body of research investigates equanimity as an outcome of mediation practices. Equanimity has been defined as a stable and impartial mental state or trait, regardless the affective valence of stimuli or situations (Desbordes et al., 2015). Few experimental studies focused on its understanding. After created and validated an equanimity questionnaire (EQUA-S, N = 265), we conducted a laboratory study (N = 38) to examine the effect of equanimity on both approach-avoidance motor-behavior with positive and negative stimuli (Rougier et al., 2018) and evaluative conditioning. While classical approach/avoidance and evaluative conditioning effects were significantly reproduced with evidence in favor of H1 among the participants with a low level of equanimity (N = 17), evidence in favor of H0 was found among those with a high level of equanimity. Thus, equanimity seems to moderate automatic cognitive responses toward valenced stimuli.

Do children extend pragmatic principles to non-linguistic communication?

Alyssa Kampa
University of Delaware, Newark, Delaware, United States

Catherine Richards
University of Delaware, Newark, Delaware, United States

Anna Papafragou
Unversity of Delaware, Newark, Delaware, United States

Abstract

In conversation, speakers are expected to offer as much information as required by the purposes of the exchange. (Grice, 1975). Classic theories of communication assume that the principle of informativeness extends beyond linguistic interactions (Grice, 1989; Sperber & Wilson, 1986), but relevant evidence so far is limited. We replicated the paradigm of a referent selection study in which preschool-aged children successfully apply the principle of informativeness to linguistic exchanges (Stiller et al., 2015) and added a matched non-linguistic condition in which the referent choice was communicated through pictures instead of verbal descriptions. Children between the ages of 3.5 to 5 performed significantly better in both the linguistic and non-linguistic conditions compared to a control condition, and there were no significant differences between linguistic and non-linguistic conditions for 3-year-olds, 4-year-olds, or 5-year-olds. We conclude that preschool-aged children apply pragmatic principles to pictures as well as words.

When do iconic gestures facilitate word learning? The case of L2 lessons for preschoolers led by a robot or human tutor

Junko Kanero
Sabanci University, Tuzla/Istanbul, Istanbul, Turkey

Cansu Oran
Ko University, Istanbul, Turkey

Smeyye Kokulu
Koc University, stanbul, Turkey

Tilbe Gksun
Ko University, Istanbul, Turkey

Aylin C Kuntay
Ko University, Istanbul, Turkey

Abstract

Gestures help us understand language (e.g., Hostetter, 2011). However, less is known about how good gestures must be to facilitate word learning. Turkish-speaking preschoolers learned five English verbs with corresponding iconic gestures, varying in the verb-gesture match (i.e., how well the gesture represented the verb), in a one-on-one lesson led by either a human adult or the humanoid robot NAO. Our preliminary results (N = 43) suggest that the verb-gesture match predicts word learning, and this match might even be more important when the robot was the tutor (though the interaction was not statistically significant). In addition, while both tutors were effective in teaching verbs, preschoolers learned better with the robot than with the human. This study not only makes a theoretical contribution by demonstrating the effects of the match between words and iconic gestures, but also provides practical implications for designing of robot- and human-led L2 lessons.

Confirmation Bias Trumps Performance Optimization in Overt Active Learning

Yul Kang
University of Cambridge, Cambridge, United Kingdom

Daniel Wolpert
Columbia University, New York, New York, United States

Mate Lengyel
University of Cambridge, Cambridge, United Kingdom

Abstract

When gathering information, different sources typically have distinct levels of informativeness. Therefore, it is optimal to actively select the source of information to learn from (i.e., perform active learning). It has been debated whether humans optimize task performance in active learning or use a simple heuristic of seeking information that confirms their beliefs. Critically, depending on ones subjective beliefs, confirmation bias can in fact be optimal. Thus, without measuring subjective beliefs, previous approaches were unable to distinguish between these alternatives. Using a perceptual decision-making task, we measured participants subjective beliefs before and after a new piece of information was presented. We then characterized confirmation-based and performance optimizing strategies with respect to these subjective beliefs. We found that participants strategy was dominated by confirmation bias, modulated only weakly by the performance optimization. We discuss potential reasons that may limit performance optimization in active learning.

High-Dimensional Vector Spaces as the Architecture of Cognition

Matthew Kelly
The Pennsylvania State University, University Park, Pennsylvania, United States

Nipun Arora
Carleton Univesity, Ottawa, Ontario, Canada

Robert West
Carleton University, Ottawa, Ontario, Canada

David Reitter
Penn State, University Park, Pennsylvania, United States

Abstract

We demonstrate that the key components of cognitive architectures - declarative and procedural memory - and their key capabilities - learning, memory retrieval, judgement, and decision-making - can be implemented as algebraic operations on vectors in a high-dimensional space. Modern machine learning techniques have an impressive ability to process data to find patterns, but typically do not model high-level cognition. Traditional, symbolic cognitive architectures can capture the complexities of high-level cognition, but have limited ability to detect patterns or learn. Vector-symbolic architectures, where symbols are represented as vectors, bridge the gap between these two approaches. Our vector-space model accounts for primacy and recency effects in free recall, the fan effect in recognition, human probability judgements, and human performance on an iterated decision task. Our model provides a flexible, scalable alternative to symbolic cognitive architectures at a level of description that bridges symbolic, quantum, and neural models of cognition.

Offloading memory: serial position effects

Megan Kelly
University of Waterloo, Waterloo, Ontario, Canada

Evan Risko
University of Waterloo, Waterloo, Ontario, Canada

Abstract

Despite the long history and pervasiveness of cognitive offloading as a memory strategy, the memorial fate of offloaded information is not well understood. Recent work has suggested that offloading information may engage similar mechanisms as instructions to forget (directed forgetting). Presently, we test this prediction by examining the serial position effect for offloaded information. Previous research has demonstrated that forget instructions can eliminate the primacy effect while leaving an intact recency effect. Across two experiments, participants completed multiple free recall trials using an external aid and then a final recall trial without the external aid. We compared a group that was expecting to use the aid for the final trial (offloading) with a group that was not (no offloading). We found a memory impairment for offloaded items that was characterized by a reduced primacy effect but intact recency effect, similar to what has been reported in research on directed forgetting.

The reassurance of the Complex Trial Protocol against ecologically validated countermeasures

Hyemin Kim
Korea University, Seoul, Korea, Republic of

Abstract

The P300-based Complex Trial Protocol (CTP), developed by Rosenfeld et al. (2008), is known to compensate for accuracy degradation and countermeasure issues of the Concealed Information Test. Although a myriad of CTP studies using electroencephalogram has been investigated, the lack of crime-related details and the complexity of the previously used countermeasures have revealed the necessity of in-depth experiment. In the present study, fifty participants were divided into three groups: guilty, innocent, and guilty-countermeasure. Participants engaged in a mock-crime scenario and only the guilty-countermeasure group performed ecologically validated countermeasures during the CTP. Participants reaction time and the amplitude of P300 components of event-related potential were analyzed and there was a significant difference ($p_i0.05$). Moreover, using the bootstrapping method, participants were correctly classified as guilty or innocent, regardless of the use of countermeasure, with accuracy above 80%. The results support the possibility of the on-site usage of the CTP.

Making Young Childrens Design Cognition Visible

Mi Song Kim
University of Western Ontario, London, Ontario, Canada

Abstract

There are emerging innovative educational interventions through automated computational analytics so-called learning analytics (LA) to utilize a large amount of student participation. However, LA is a relatively unexplored area in Early Childhood Education (ECE). To respond to this gap, LA is defined as a tool for co-designing pedagogical documentation practices with ECE teachers to visualize student design cognition. Drawing upon a Multiliteracies pedagogy framework, this qualitative study investigates how two kindergarten teachers co-designed pedagogical documentation practices using a digital portfolio app (Seesaw) to leverage 25 young childrens design cognition in multiple modes and technologies. Using the constant comparison method, two themes were emerged from multiple data sources (e.g., digital portfolios on Seesaw, teacher assessment, fieldnotes, interviews): teachers-as-(Co)Designers of LA Interventions; and Portfolio of Student Learning Progression, not Portfolio of Student Work. Our findings suggest the need for effective pedagogical supports for young childrens design cognition and their teachers LA interventions.

Downloading Culture.zip: Social learning by program induction with execution traces

Max Kleiman-Weiner
Harvard University, Cambridge, Massachusetts, United States

Felix Sosa
Harvard University, Cambridge, Massachusetts, United States

Samuel Gershman
Harvard University, Cambridge, Massachusetts, United States

Fiery Cushman
Harvard University, Cambridge, Massachusetts, United States

Abstract

Cumulative culture ultimately depends on the fidelity of learning between successive generations. When humans learn from others in addition to observing inputs and outputs we often observe the process which led to that output. For instance, when preparing a meal we don't just observe a pile of vegetables and then a ratatouille. Instead, we observe a causal process by which those ingredients are transformed. Here we use programs to represent a cultural process and show that the observation of an execution trace speeds up program induction even when learning from only a single example. This mechanism could account for (1) the high fidelity of social learning which leads to cumulative culture in humans (2) unify the role of emulation and imitation in social learning and (3) account for aspects of moral learning such as ritualization.

Curiouser and Curiouser: Childrens intrinsic exploration of mazes and its effects on reaching a goal.

Eliza Kosoy
UC Berkeley , Berkeley, California, United States

Deepak Pathak
UC Berkeley, Berkeley, California, United States

Pulkit Agrawal
UC Berkeley, Berkeley, California, United States

Alison Gopnik
University of California at Berkeley, Berkeley, California, United States

Abstract

Children are naturally curious, and now even reinforcement learning models within machine learning are channeling this child-like curiosity. Pathak et-al (2017) created the ICM (Intrinsic Curiosity Model) in which curiosity serves as an intrinsic reward signal to enable the agent to explore its environment and learn skills, in this case a maze game called Doom. We study this inherent ability in children by having them explore mazes, with and without goals built using DeepMind software. In our pilot data we found that kids are adept at exploring the maze, readily and without prompt. We suggest a relationship between exploration and performance on a maze task, such that performance in the curiosity driven maze exploration task, is correlated with finding a goal in a second separate maze, even when the initial path to the goal is blocked. We also show side-by-side comparisons of the ICM vs. children exploring on our mazes.

Emotional Speech Processing With the Help of F2 Syntactic Parser

Artemy Kotov
Kurchatov Institute, Moscow, Russian Federation

Nikita Arinkin
National Research Center Kurchatov Institute, Moscow, Russian Federation

Liudmila Zaidelman
National Research Center Kurchatov Institute, Moscow, Russian Federation

Anna Zinina
National Research Center Kurchatov Institute, Moscow, Russian Federation

Abstract

F2 syntactic parser is a part of F2 emotional robot, designed to support natural emotional communication with the help of gestures, facial expressions and speech. The parser constructs syntactic and semantic representations (frame networks) of an input text, saves them to memory (database) and selects a communicative reaction for the robot in BML (behavior markup language) format. The model of reactions and inferences is based on scripts if-then operators, competing for the processing of semantics. In particular, scripts detect emotionally relevant meanings: when it is declared, that somebody threatens the robot, does not care about it, behaves inadequately 13 negative scripts, and also when the robot is superior, attracts attention, etc 21 positive scripts. Parser may run in a standalone mode, daily processing sentences from news and blogs. Balancing of scripts allows us to tune the understanding and reproduce different emotional profiles for the robot. (Research is supported by RSF, project No 17-78-30029).

Visual, auditory, and temporal sensorimotor discrimination abilities and their relationships with complex cognition

Bartomiej Kroczek
Jagiellonian University, Cracow, Poland

Jan Jastrzebski
Jagiellonian University, Krakw, Poland

Micha Ociepka
Jagiellonian University, Krakw, Poland

Hanna Kucwaj
Jagiellonian University, Krakow, Poland

Adam Chuderski
Jagiellonian University, Krakow, Poland

Abstract

At dawn of cognitive science, it was hypothesized that performance on diverse sensorimotor tasks is rooted in unitary sensory discrimination ability that shares the same neural resource with complex cognition. A century of research yielded inconclusive evidence. We modelled the factor structure for 33 diverse visual sensorimotor, memory, and reasoning tasks, completed by 234 young adults. Covariance structure models indicated two considerably correlated, yet statistically separate, sensorimotor abilities reflecting temporal vs. non-temporal processing. However, initially moderate relationships of each simple ability with reasoning disappeared when mediated by working memory, suggesting that sensory discrimination plays no explanatory role for complex cognition. These results were replicated in another study of 255 young adults, who additionally attempted auditory sensorimotor tasks. The latter appeared to be separate from temporal and visual abilities. Overall, sensory discrimination does not constitute unitary ability. Moreover, individual differences in complex cognition cannot be reduced to sensory discrimination.

Sizing Up Relations: Dimensions on Which Stimuli Vary Affect Likelihood of Adults' Relational Processing

Ivan Kroupin
Harvard University, Cambridge, Massachusetts, United States

Abstract

Relational reasoning is central to much of human-unique cognition including artistic metaphor, scientific analogy. While much research has addressed the process of relational reasoning, the conditions under which relational reasoning is engaged in at all remains under-explored.

This work examines the relationship between dimensions on which stimuli vary and the likelihood that these stimuli will be processed relationally by adults. We use a modified relational-match-to-sample paradigm: One of the two choices contains a relational match with the target, the other contains a partial object match. Changing dimensions on which the stimuli vary dramatically effects the likelihood that adults process them relationally (i.e. make relational matches) - from 56% when stimuli vary on shape and color to 98% when stimuli vary on size alone. This is despite the relational content of the task remaining identical throughout.

We discuss implications of these results for designing stimuli, and for theories of relational reasoning generally.

Look out, its going to fall!: Does physical instability capture attention and lead to distraction?

Marta Kryven
Massachusetts Institute of Technology, Cambridge, Massachusetts, United States

Sholei Croom
MIT, Cambridge, Massachusetts, United States

Brian Scholl
Yale University, New Haven, Connecticut, United States

Josh Tenenbaum
MIT, Cambridge, Massachusetts, United States

Abstract

Physical scene understanding requires not only detecting the current state of the world, but also predicting how the future will unfold. The need for such prediction is especially salient in the context of physical instability as when an object is teetering, about to fall off a surface. Here we asked whether such scenes automatically capture attention, such that the mere presence of instability will impair performance on a central attention-demanding task. Observers viewed scenes in which an object (e.g. an open laptop) was either sitting stably, or was about to fall off a table. Observers simply completed a central Multiple Object Tracking (MOT) task (e.g which could appear on the screen of the depicted laptop). MOT Performance was indeed worse in the presence of physical instability, despite its task irrelevance, and even when observers failed to notice the physical stability vs. instability in the first place.

Verbal Insight Revisited: fMRI evidence for subliminal processing in bilateral insulae for solutions with AHA! experience shortly after trial onset

Simone Khn
University Medical Center Hamburg-Eppendorf, Hamburg, Germany

Tobias Sommer
University Medical Centre Hamburg-Eppendorf, Hamburg, Germany

Maxi Becker
University Medical Center Hamburg-Eppendorf, Hamburg, Hamburg, Germany

Abstract

In insight problem solving solutions with AHA! experience have been assumed to be the consequence of restructuring of a problem which usually takes place shortly before the solution. However, evidence from priming studies suggests that solutions with AHA! are not spontaneously generated during the solution process but already relate to prior subliminal processing. We test this hypothesis by conducting an fMRI study using a modified compound remote associates paradigm which incorporates semantic priming. We observe stronger brain activity in bilateral anterior insulae already shortly after trial onset in problems that were later solved with than without AHA!. This early activity was independent of semantic priming but may be related to other lexical properties of attended words helping to reduce the amount of solutions to look for. In contrast, there was more brain activity in bilateral anterior insulae during solutions that were solved without than with AHA!. This timing (after trial start / during solution) x solution experience (with / without AHA!) interaction was significant. The results suggest that a) solutions accompanied with AHA! relate to early solution-relevant processing and b) both solution experiences differ in timing when solution-relevant processing takes place. In this context, we discuss the potential role of the anterior insula as part of the salience network involved in problem-solving by allocating attentional resources.

An Investigation on the Relationships Among Social Cognition Processes by Eye-Tracking Techniques

Pei-Ling Kuo
College of Social Science, Tainan, Taiwan

Ting Yun Chen
National Cheng Kung University, Tainan, Taiwan

Ting-Hsuan Chang
National Cheng Kung University, Tainan, Taiwan

Shiau-Wen Chen
National Cheng Kung University, Tainan, Taiwan

Mingzhe Liu
National Cheng Kung University, Tainan, Taiwan

Jon-Fan Hu
National Cheng Kung University, Tainan City, Taiwan

Abstract

The present study integrates four primary social cognition processes Joint Attention(JA), Intention Detection(ID), Perspective Taking(PT), and Social Reference(SR) into lively comic scenarios in order to disentangle their relationships and possible one-to-one connections. By using eye-tracking technique, gaze patterns in terms of Total Fixation Duration were considered as indexes to examine the hypotheses. It is found that PT is positively correlated with JA, ID, and SR whereas JA is positively correlated with ID and PT. As a criteria-related validation, the scores of Geneva Social Cognition Scale(GeSoCS) were used to delineate the gaze performance. Participants with higher score in GeSoCS showed different eye-movement patterns to those with lower score, indicating the pattern of eye movements could be a reliable indicator of social cognition status. Moreover, the correlations revealed in the present study suggest that close connections exist between social cognition processes and eye gaze scanning toward pictorial scenarios.

Automated cognitive modeling with Bayesian active model selection

Vishal Lall
UC Berkeley, Berkeley, California, United States

Jordan Suchow
Stevens Institute of Technology, Hoboken, New Jersey, United States

Gustavo Malkomes
Washington University in St. Louis, St. Louis, Missouri, United States

Tom Griffiths
Princeton University, Princeton, New Jersey, United States

Abstract

Behavioral experiments are often feed-forward: they begin with designing the experiment, and proceed by collecting the data, analyzing it, and drawing inferences from the results. Active learning is an alternative approach where partial experimental data is used to iteratively design subsequent data collection. Here, we study experimental application of Bayesian Active Model Selection (BAMS), which designs trials to discriminate between a set of candidate models. We consider a model set defined by a generative grammar of Gaussian Process kernels that can model both simple functions and complex compositions of them. To validate the method experimentally, we use BAMS to discover how factors such as contrast and number affect numerosity judgements. We compare the rate of convergence of the active-learning method to a baseline passive-learning strategy that selects trials at random. Active learning over a structured model space may increase the efficiency and robustness of behavioral data acquisition and modeling.

Using interpersonal movement coordination to investigate gender differences in adults with autism

Nida Latif
McGill University, Montreal, Quebec, Canada

Cynthia Di Francesco
McGill University, Montreal, Quebec, Canada

Aparna Nadig
McGill University, Montreal, Quebec, Canada

Abstract

When individuals engage in social interactions, they coordinate their nonverbal movements. Atypical movement coordination may contribute to social difficulties in autism. Further, distinct gender differences have been found in autism: males show reduced socio-communicative behaviours relative to females. Here, we explored whether interpersonal movement coordination differs between males and females with autism, compared to neurotypical (NT) adults. Thirteen adults with autism participated. Twenty-six NT controls are currently being tested. Participants complete a semi-structured interview while being video-recorded. Coordination between participant and examiner is measured using a video-based movement analysis. Females with autism demonstrated significantly greater movement coordination with their conversational partner, within a smaller range, than males. Given past findings, we expect that coordination differences between autistic and NT males will be greater than between autistic and NT females. These preliminary results suggest that investigating movement coordination during interaction may provide a tool for better understanding gender differences in ASD.

Novel labels modify visual attention in 2-year-old children

Alexander LaTourrette
Northwestern University, Evanston, Illinois, United States

Miriam A. Novack
Northwestern University, Evanston, Illinois, United States

Sandra Waxman
Northwestern University, Evanston, Illinois, United States

Abstract

Labeling objects enhances fundamental cognitive capacities like categorization, individuation, and memory in young children. However, the mechanism by which labels support these cognitive processes remains unknown. One possibility is that providing a label for an object changes childrens online visual processing of that object. To address this, we considered several indices of visual attention, asking whether 2-year-old children attend to an object differently if it is labeled (Look at the dax) than if it is paired with a non-labeling phrase (Look at that). We find that 2-year-old childrens visual fixations are longer when objects are paired with a labeling phrase, rather than a non-labeling phrase. Indeed, after hearing a label, children showed a sustained increase in fixation duration. However, the number of fixations children made did not change as a function of labeling. This illustrates an attentional mechanism by which language might enhance learning in 2-year-old children.

Modal concepts: developing thoughts of the possible and the impossible

Brian Leahy
Harvard, Cambridge, Massachusetts, United States

Susan Carey
Harvard University, Cambridge, Massachusetts, United States

Abstract

What is it to represent a single world as having alternative, mutually inconsistent possible futures? A large literature explores this question from philosophical and linguistic perspectives, along with a growing literature in developmental psychology. Recent findings suggest that 36 month olds (Redshaw and Suddendorf 2016) or even 14 month olds (Cesana-Arlotti et al. 2018) prepare for multiple alternative possibile futures. These experiments did not require participants to contrast the possible with the impossible. We replicated Redshaw and Suddendorf (2016), and added conditions that required participants to contrast the possible with the impossible. 36 month olds now failed, as did many 48 month olds, suggesting that their representations do not capture the structure of possibilities. 48 month olds tended to pass our test, but their understanding of possibilities was still fragile. These data converge with other results suggesting that concepts of possibility and impossibility are constructed in the late preschool years.

Drawing conclusions from spatial coincidences: a cumulative clustering account

Jennifer Lee
NYU, New York, New York, United States

Wei Ji Ma
New York University, New York, New York, United States

Abstract

Spatial coincidences allow us to infer the presence of latent causes in the world. For instance, an unusually large cluster of ants allows us to infer the presence of a food source. The leading cognitive model for such inferences is Bayesian, but the Bayesian algorithm is computationally taxing. Humans likely employ a more efficient, approximative algorithm. To characterize the cognitive algorithms used, we had subjects judge whether a set of dots was drawn from a uniform distribution or from a mixture of a uniform and a gaussian source (tending to produce clusters). Responses systematically deviate from Bayesian optimality: as the number of dots increase, subjects more often report a latent cause where none exists. The bias is accounted for by a Bayesian clustering algorithm that cumulatively considers the next-nearest dot to a putative source. This finding helps characterize our tendency to perceive causal patterns where none exist.

Brain responses to verbal mismatches and case marking mismatches: adolescents vs. adults

Sun-Young Lee
Cyber Hankuk University of Foreign Studies, Seoul, Korea, Republic of

Dr. Jinhee Jeong
Hankuk University of Foreign Studies, Seoul, Korea, Republic of

Eun Kyoung Lee
University of Maryland, College Park, Maryland, United States

Ha-A-Yan Jang
Sogang University, Seoul, Korea, Republic of

Dr. Sook Whan Cho
Sogang University, Seoul, Korea, Republic of

Abstract

This study investigated Korean adolescents behavioral and neural responses to the semantic and syntactic anomalies in Korean compared with adults, focusing on the case marking mismatches. EEG data were collected from 16 Korean adolescents (12 males, aged 12-14 years) using a picture sentence verification task regarding (A) verbal mismatch [AGENT-NOM + Verb/*Verb] (e.g., - /*; Brother-ka catches/*bites) and (B) case marker mismatch [AGENT-NOM/*ACC + Verb] (e.g., -/*- ; Brother-ka/*-lul catches). The behavioral results showed 95% accuracy of their judgment regardless of conditions.The ERP data revealed differences between the conditions: N400 was elicited for verbal mismatches as well as for case marker mismatches. The results are different from data collected from Korean adults, where the syntactic anomalies elicited early negativity at the case marker in addition to the N400 at the verb. The different ERP responses between adults and adolescents to the syntactic anomalies provide evidence for the continuous development of human brains.

Evidence for a 30-million-word gap across language environments of children with cochlear implants

Matthew Lehet
Michigan State University, East Lansing, Michigan, United States

Meisam K. Arjmandi
Michigan State University, East Lansing, Michigan, United States

Laura Dilley
Michigan State University, East Lansing, Michigan, United States

Abstract

Hart and Risley (1995) found evidence of a 30-million-word gap by the age of three between children experiencing the most and the least spoken input. In the present study, we investigated the magnitude of differences in amount of linguistic input in environments of a clinical population: children with cochlear implants. We identified a 30 million word gap over three years between children who received the most and the least spoken language input in their home environments. Further, we identified a 22 million word gap in numbers of infant-directed spoken words experienced by children hearing the most and the least input. Together, the results suggest that some children with cochlear implants may be doubly disadvantaged in acquiring spoken language, due to the degradation of the speech signal associated with electronic hearing, and due to the dearth of quality linguistic input in sufficient quantity in their language environments.

Approximate Inference through Sequential Measurements of Likelihoods Accounts for Hicks Law

Xiang Li
New York University, New York, New York, United States

Luigi Acerbi
University of Geneva, Geneve, Switzerland

Wei Ji Ma
New York University, New York, New York, United States

Abstract

In Bayesian categorization, exactly computing likelihoods and posteriors might be hard for humans. We propose an approximate inference framework inspired by Bayesian quadrature and Thompson sampling. An agent can pay a fixed cost to make a noisy measurement of the likelihood of one category. By sequentially making measurements, the agent refines their beliefs over the likelihoods. When the agent stops measuring and chooses a category, they get rewarded for being correct; the agent chooses the category that maximizes probability correct. To decide whether to make another measurement, the agent simulates one measurement for each category. If any of the gains in expected reward exceeds the cost, they make a real measurement corresponding to the simulation with the largest gain. We find that the average number of measurements grows approximately logarithmically with the number of categories, reminiscent of Hicks law. Furthermore, our model makes predictions for decision confidence among multiple alternatives.

Do children really have a trust bias? Preschoolers reject labels from previously inaccurate robots but not inaccurate humans

Xiaoqian Li
Singapore University of Technology and Design, Singapore, Singapore

Wei Quin Yow
Singapore University of Technology and Design, Singapore, Others, Singapore

Abstract

Past research suggests that young children have a bias to believe what they are told so that they often trust an informant regardless of the informants previous accuracy. With the ubiquity of new technology, children regularly come in contact with non-human agents such as robots, yet little is known how children are trusting and thus willing to learn from these artificial beings. In our study, 3.5- to 5.5-year-old children (N=120) watched a single informant (either a robot NAO or a human adult) name familiar objects either accurately or inaccurately. The same informant subsequently tested children on their willingness to accept novel labels for novel objects provided. While children trusted the accurate robot and the accurate human to the same extent, they were less likely to accept information from the inaccurate robot than the inaccurate human. This suggests that preschoolers may not readily extend their trust bias to robots as informants.

Predicting human decisions in a sequential planning puzzle with a large state space

Yichen Li
New York University, New York, New York, United States

Zahy Bnaya
New York University, New York, New York, United States

Wei Ji Ma
New York University, New York, New York, United States

Abstract

We study human sequential decision-making in large state spaces using a puzzle game called Rush Hour. A puzzle consists of a dense configuration of rectangular cars on a 6x6 grid. Each car moves only horizontally or vertically. The goal is to move a target car to an exit. In a given state (board position), a subject (n=86) could move a car, restart the puzzle, or surrender. A move is correct if it reduces the distance (number of moves) to the goal. Using mixed-effects logistic regression modeling, we find that the probabilities of an error, a restart, and a surrender are higher with a longer distance to goal, higher mobility, and when the previous move was an error. The effects of distance to goal and mobility are consistent with tree search. As a next step, we plan to investigate the heuristics that people might use for such tree search.

Scientific knowledge organized through question network

Zhiwei Li
NYU, new york, New York, United States

Kai Ren
National University of Singapore, Singapore, Singapore

Abstract

Research in science is usually built uponcomplex background knowledge and assumptions, making it difficult to organize and overview. We propose using question network to dynamically maintain scientific knowledge, with each nodes being either a question or an answer, linked with relations such as specification, contrast and so on. Publications can then be fitted into nodes of the network. By constructing example networks around cognitive concepts, we observed a big question (e.g. What is curiosity?) being answered with theoretical speculation initially, then specified into the operationalized definition (How to measure curiosity as a personality?) and computational algorithms. Similar patterns are repeated in different branches of the network. We also compare research topics starting with similar questions yet develop differently.

Causal Structure and Probability Information Modulate the Preference for Simple Explanations

Emily Liquin
Princeton Univerisity, Princeton, New Jersey, United States

Tania Lombrozo
Princeton University, Princeton, New Jersey, United States

Abstract

Are simple explanations better? Research has shown that people favor simple explanations (defined as number of unexplained causes; Lombrozo, 2007; Pacer & Lombrozo, 2017), but new findings suggest that under some conditions, complexity is preferred (Johnson et al., in press; Zemla et al., 2017). We explore three features that could affect preferences: causal structure, baserates, and likelihoods. Adults (N=544) read one simple and one complex explanation following one of three causal structures. Simplicity preferences were strongest for one vs. two causes explaining two independent effects, modest for one vs. two jointly sufficient causes explaining one effect, and reversed (to favor complexity) for one vs. two independently sufficient causes explaining one effect. When baserates and likelihoods were specified and matched, simplicity preferences were attenuated, while complexity preferences were sometimes reversed. These findings suggest that simplicity preferences are moderated by several factors and point to a more unified account of explanatory reasoning.

The Development of Children's Understanding of Arguments by Analogy

Nicole Lobo
Arizona State University, Glendale, Arizona, United States

Zachary Horne
Arizona State University, Phoenix, Arizona, United States

Abstract

Analogical reasoning allows humans to make inferences about novel experiences and transfer learning across contexts. There is substantial literature on how analogical reasoning develops, but less is known about how children understand a common use of analogyargument by analogy. Considering the importance argument by analogy plays in politics and the law, we examined the developmental trajectory of the ability to understand arguments by analogy. We measured childrens (N = 128, ages 3-12 years old) performance on a commonly used analogical reasoning task (i.e., a picture-mapping task; see Richland et al., 2006) and their understanding of arguments by analogy. We found that at age 4, children have as much difficulty understanding arguments by analogy as they do performing a picture-mapping task. However, by age five, childrens performance improves more rapidly in an argument by analogy task compared to a picture-mapping task.

Modeling practice-related reaction time speedup using hierarchical Bayesian methods: Evidence for a process-shift account

Jarrett Lovelett
University of California San Diego, San Diego, California, United States

Ed Vul
University of California, San Diego, La Jolla, California, United States

Tim Rickard
University of California San Diego, La Jolla, California, United States

Abstract

In skill-learning tasks, reaction times (RTs) typically decrease with practice. For example, in alphabet arithmetic tasks (e.g. J + 7 = ?), learners respond correctly (e.g. Q) faster on later than on earlier trials. A number of mathematical models have been proposed to account for the functional form of practice-related RT speedup. We aim to evaluate which of two candidates better fits observed speedup data for individual learners across several tasks. In particular, we compare a process-shift account in which learners initially execute an algorithm in constant time, but as trials accumulate, exhibit power-law speedup as they directly retrieve a memorized solution to a delayed exponential model in which RTs decrease exponentially after learners eventually achieve insight into a task-appropriate strategy. Using hierarchical Bayesian models of each account (which can flexibly model learning in individual subjects), we show that the process-shift model better predicts out-of-sample data than the delayed-exponential model.

The Effects of Contextual Cues on the Learning of Prepositions

Michelle Luna
University of California, Los Angeles, Los Angeles, California, United States

Catherine Sandhofer
University of California, Los Angeles, Los Angeles, California, United States

Abstract

Language has the power to shape the way people organize their thoughts and concepts. Some concepts, like spatial words, are categorized differently cross-linguistically. Conflicting language-to-concept mappings, such as the Spanish en translating to both in and on, may pose difficulty to Spanish speakers learning English. This study investigated how contextual cues can help children learn prepositions. Three-year-olds were read preposition books that were arranged in one of two conditions: separation or control. The separation condition had each instance of in appear in one visual context (e.g., Bear put the apple in the box, blue page) and each instance of on appear in a separate context (e.g., Penguin put the ball on the grass, green page). The control condition eliminated the contextual cues by presenting instances of in and on in both contexts. This study informs our understanding of strategies to improve the learning of spatial words in everyday adult-child interactions.

How does temperature affect behaviour? A meta-analysis of effects in experimental studies

Dermot Lynott
Lancaster University, Lancaster, United Kingdom

Katherine Corker
Grand Valley State University, Allendale, Michigan, United States

Louise Connell
University of Lancaster, Lancaster, United Kingdom

Kerry O'Brien
Monash University, Melbourne, VIC, Australia

Abstract

The surrounding environment has a profound impact on human behaviour. Historically, studies have shown that higher temperatures are associated with increases in antisocial behaviours (aggression, violence). More recently, studies have linked higher temperature experiences to increases in prosocial behaviours (altruism, co-operation). Such contrasting patterns leave the status of temperature-behaviour links unclear. Here we conduct a series of meta-analyses of laboratory-based empirical studies that measure either prosocial (monetary reward, gift giving, helping) or antisocial (retaliation, horn honking, sabotage) outcomes, with temperature as an independent variable. Overall, we found that there was no reliable effect of temperature on the behavioural outcomes measured. In follow-up analyses, there was no reliable effect of temperature on prosocial or antisocial outcomes when analysed separately. We consider why the evidence to support temperature-behaviour links from laboratory-based studies is weak, assess potential moderators, and examine how future studies can attempt to reconcile seemingly contradictory patterns in the literature.

Measuring Creativity in the Classroom: Linking Group Patterns with Individual Outcomes

Leanne Ma
OISE/University of Toronto, Toronto, Ontario, Canada

Abstract

Although creativity has traditionally been measured as an individual trait (Runco & Jaeger, 2012), contemporary research on workplace innovation (Kelley & Littman, 2001; Nonaka, 2008) suggests that creativity is a collaborative process of working with ideas (Amabile & Pratt, 2016). Furthermore, organizational creativity can be measured using social network analysis (Gloor, 2006) the more emergent leaders, the more creative the outcome (Gloor et al., 2016). Gloor's creativity measure was adapted in a grade 1 class (n=22) to explore whether leaders would emerge when students engaged in creative problem-solving through online discussions in Knowledge Forum (Scardamalia, 2017). Social network analysis reveals that 13 students emerged as leaders, and content analysis of the discussion indicates that leaders proposed new ideas that helped deepen the progression of ideas. Additional analyses are underway to explore correlations between leadership and creativity scores. Educational implications for developing the creative potential of young students are discussed.

Deconvolving a Complex, Real-Life Task: Do standard lab tasks predict CPR learning and retention?

Sarah Maa
University of Groningen, Groningen, Netherlands

Florian Sense
University of Groningen, Groningen, Netherlands

Michael Krusmark
Wright-Patterson Air Force Base, Dayton, Ohio, United States

Kevin Gluck
Air Force Research Laboratory, Wright-Patterson AFB, Ohio, United States

Hedderik van Rijn
University of Groningen, Groningen, Netherlands

Abstract

Cardiopulmonary resuscitation (CPR), a basic life-saving skill, requires a combination of procedural and declarative knowledge. CPR proficiency was assessed and re-trained to criterion across four sessions (spaced weeks to months apart). In addition, three laboratory tasks were administered: continuation tapping, paired-associate learning, and Raven matrices. These served as proxies for procedural learning, declarative learning, and general cognitive ability, respectively. Even though a computational model (Predictive Performance Equation, Walsh et al., 2018) predicted long-term CPR performance, none of the lab tasks correlated with any aspect of CPR performance (initial performance, (re-)learning, or retention of CPR; see https://osf.io/m8bxe/ for details). These results highlight the challenges faced when translating lab results into real-world domains and can serve as a benchmark for applying computational models to real-life learning and forgetting.

Controlling Automobiles During Unconsciousness of the Driver using Brainwaves

Nilakshi Mahanta
North Eastern Hill University, Guwahati, Assam, India

Abstract

Introduction: Controlling Automobiles during unconsciousness of the driver using Brainwaves. Brainwave based accident avoidance system is an effective way to prevent accident caused due to drowsy driving. Every year number of road mishaps are caused by drowsy driving. The proposed idea brainwave based accident avoidance system is to avoid this kind of accident using Electroencephalography (EEG) of human brain and speed control in automobiles. Human brain consists of millions of interconnected neurons. The patterns of interaction between these neurons are represented as thoughts and emotional states. According to the human thoughts, this pattern will be changing which in turn produce different electrical waves. A muscle contraction will also generate a unique electrical signal. All these electrical waves will be sensed by the brain wave sensor and it will convert the data into packets and transmit through Bluetooth medium. Level analyzer unit (LAU) will receive the brainwave raw data and it will extract and process the signal using MATLAB platform. Then the control commands will be transmitted to the motor to process. With this entire system, we can control / stop the vehicle according to human thoughts. Electroencephalography (EEG) is the fundamental idea utilized as a part of this framework. Neurosky mind wave sensor is utilized as primitive segment to examine the Brainwave signals. In this way by controlling vehicles it can spare numerous mishaps and can spare numerous lives. Among these bands, theta and alpha are the signals which represent drowsiness to relaxed sleep. Methods: In a brain controlled vehicle, controller is based on Brain Computer Interface (BCI). BCIs are systems that can bypass conventional channels of communication to provide direct communication and control between the human brain and physical devices by translating different patterns of brain activity into commands in real time. With these commands a vehicle can be controlled. The intention of this work is to design and develop a system that can assist the person during their unhealthy condition to avoid the accident on the road. Results: Brainwave based accident avoidance system for unhealthy condition of the drivers which predict the signals and system in engaging with processing of signals to alert the drivers unconscious situation. The biggest challenge about the system is that to determine the signal from the headset. Proper identification is needed for the signals so that wrong signal does not trigger the routine even when driver is not unconscious. Every person is different and every person has different thoughts and emotions so they might have slightly different brainwave signals. So before adapting this system, the interface should be configured according to the brain activity of the driver. Discussion: The research and development of brainwave controlled vehicle during unconsciousness of the driver has received a great deal of attention because they can help to avoid the accident on the road. Improving the BCI system performance to make brainwave controlled vehicles usable in real-world situations. Keywords: Brain Computer Interface (BCI), Brain Wave Sensor, EEG, Bluetooth

Cultural difference of the effect of analytical / intuitive thinking style on reasoning, JDM, and belief tasks.

Yoshimasa Majima
Hokusei Gakuen University, Sapporo, Japan

Abstract

Research within the dual-process framework have repeatedly suggested that individuals thinking style can predict their performance on reasoning, judgment, decision making, and acceptance of religious and paranormal statements. However, some studies also suggested that the link between analytical thinking and epistemically unwarranted beliefs was peculiar to so-called WEIRD societies. The present study aimed to explore the possible cultural (Western and Eastern) difference on the relationship between performance and style of our thinking. Participants were presented with various tasks including belief bias, denominator neglect bias, numeracy, temporal discounting, risk preference, and paranormal belief. They were also presented with tasks measuring their thinking styles (CRT and Rational-Experiential Inventory). Results showed that the effects of thinking style on heuristics-bias and decision-making tasks were almost similar between two cultures, however we find a significant style-culture interaction in paranormal beliefs. This may suggest a cultural difference of the role of analytical thinking on belief-based response.

Testing human use of probability in a visuo-motor conjunction task

Laurence Maloney
New York University, New York, New York, United States

Jinsoo Kim
New York University, New York, New York, United States

Keiji Ota
New York University, New York, New York, United States

Abstract

People overestimate the conjunctive probability of independent events (Bar Hillel, 1973). We examined conjunctive performance in a task involving motor uncertainty and binomial sampling. Human probabilistic judgment is typically near-optimal with either of these sources of uncertainty alone. Four subjects attempted to earn rewards by reaching to circular targets. They chose between a single smaller target and one of N larger targets. Hitting the single target always earned a reward but only one on the N larger targets was rewarded: they chose between P[Smaller] and the conjunctive probability (1/N)*P[Larger] as we varied N and the sizes of the targets. The ideal observer should be indifferent when P[Smaller] = (1/N)*P[Larger]. We also asked observers to estimate the probability of hitting targets of different sizes to verify that they could do so accurately. Remarkably, three out of four observers ignored numerosity N in their preferences.

The Influence of Implicit Normative Commitments in Decision-Making

Alexia Cristina Martinez
Princeton University, Princeton, New Jersey, United States

Abstract

We approach some decisions (e.g., choosing an investment plan) by deliberating about our options, and others (e.g., choosing dessert) by relying on intuition. In a study with 259 participants evaluating hypothetical decisions, we investigate factors that predict whether deliberation and/or intuition is judged appropriate. We find that participants are more inclined to endorse deliberation, and less inclined to endorse intuition, when they believe the means and ends involved in a decision can be objectively evaluated (consistent with Inbar, Cone, & Gilovich, 2010). We also find that violations of coherence (i.e., endorsing contradictory beliefs about a decision) predict higher ratings for intuition, as does belief that a given decision reflects ones identity. These findings hold after adjusting for perceived effort, importance, and stakes. We suggest that deliberation is judged appropriate when people believe that norms governing rational action apply, and we consider the implications for real-world decision-making.

Forming Action-Effect Contingencies through Observation of a Dot-Control Task

Jasmine Mason
Illinois State University, Normal, Illinois, United States

J. Scott Jordan
Illinois State University, Normal, Illinois, United States

Abstract

Previous research suggests the possibility that observers have access to action plans of others (Jordan & Hommel, 2008). To examine this we design three experiments. The first examines action-plan coding in participants performing the task (controllers) using a Hommel-like 'compatibility' test measuring reaction times (Hommel, 1996). We manipulated the inclusion of task irrelevant auditory tones during the dot-control game. The second experiment utilized the same design to examine observer's action-plans after watching the experimenter play the dot control game. Experiment 3 allows us to examine the additional effects of the controller's skill level and observer's level of access to the task. So far the results support the hypothesis that participants can learn action plans by observing the distal effects of another's actions. Further research will help unearth the factors mediating observer's action plan coding and the differences between how controllers and observer's encode actions and their different effects.

Analysis on learning a latent structure in a probabilistic reversal learning task

Akira Masumi
National Institute of Technology, Okinawa College, Nago, Okinawa, Japan

Takashi Sato
National Institute of Technology, Okinawa College, Nago, Okinawa, Japan

Abstract

We need to be flexible to adapt to dynamically changing circumstances. A probabilistic reversal learning task is one of the experimental paradigms to characterize flexibility of a subject. In recent studies, it is hypothesized that a subject may utilize not only a reward history but also a cognitive map representing a latent structure of the task. In this study, we conducted an experiment using the task toward understanding a process of learning a latent structure of the task. We found subjects choose a rewarding option with relatively high frequency in a later phase of the task. Analyzing the subjects decision making, it is suggested that they make decision based on their own estimation about the latent structure. A statistical model selection suggested that a reinforcement learning model with state representations fit behavioral data in the later phase. These results suggest the subjects learn the latent structure during the task.

The role of environment and body in divergent thinking tasks

Heath Matheson
University of Northern British Columbia, Prince George, British Columbia, Canada

Yoed Kenett
University og Pennsylvania, Philadelphia, Pennsylvania, United States

Alexander LePage LePage
University of Northern British Columbia, Prince George, British Columbia, Canada

Mathew Sargent
University of Northern British Columbia, Prince George, British Columbia, Canada

Abstract

Humans are creative tool users. We investigated whether body posture and environmental context influence creative output in the divergent thinking task. Participants adopted either flexion or extension body postures and were shown images of kitchen utensils or work tools. Each image was primed with an image of either a congruent environment or an incongruent environment. Results show that body posture, specifically extension, results in faster generation of responses, especially when the object is primed by a congruent environment, and that extension increases sensitivity to environmental primes, increasing fluency overall. Our results shed light on the cognitive mechanisms of generating creative object uses.

Spatial Alignment Enhances Comparison of Complex Educational Visuals

Bryan Matlen
WestEd, San Francisco, California, United States

Benjamin Jee
Worcester State University, Worcester, Massachusetts, United States

Nina Simms
Northwestern University, Evanston, Illinois, United States

Dedre Gentner
Northwestern University, Evanston, Illinois, United States

Abstract

Grasping relational concepts is facilitated by comparing their representations. Previously, Matlen et al (2014; under review) found that for simple visual figures, the comparison process was optimized when the visuals were placed in direct spatial alignment, such that the main axes of the visuals run perpendicular to their placement (e.g., horizontal figures placed vertically), relative to impeded spatial alignment, when the axes run parallel to their placement. In the present work, we tested this spatial alignment effect using complex naturalistic stimuli, consisting of skeletal structures. Participants identified anomalous bones by comparing a correct skeleton with a skeleton that had an incorrect bone. Participants were more accurate when skeletal structures were placed in direct (M=.90) relative to impeded (M=.84) alignment (p¡.01). Given the relevance of these findings to education, we are formally coding visuals in middle-school science textbooks based on their spatial alignment and will present these results at the conference.

Quality of STEM Learning from Childrens Books

Hilary Miller
Emory University, Atlanta, Georgia, United States

Lucy Cronin-Golomb
Emory University, Atlanta, Georgia, United States

Patricia J. Bauer
Emory University, Atlanta, Georgia, United States

Abstract

Promoting STEM knowledge early in development helps prepare children for school success. Exposing children to STEM books may be a simple and effective means for promoting early STEM knowledge. However, whether preschool-aged childrens STEM books are optimally designed is unknown. Children and adults learn new information more effectively when there is support for encoding and demand for active processing. We have conducted a textual analysis of 50 STEM books designed for preschool-aged children. The books are coded for (a) support for encoding (narratively cohesive and topic maintaining), and (b) demand of active processing (posing questions and including interactive prompts). Preliminary data shows that on average the books include limited support for encoding and demand for active processing. This suggests that these books are not fulling their potential of promoting early STEM knowledge. Next steps in this research involve identifying means for enhancing STEM childrens books efficacy.

The Development of Reasoning About Abductive, Inductive and Deductive Conditionals

Patricia Mirabile
Sorbonne Universit, Paris, France

Zachary Horne
Arizona State University, Phoenix, Arizona, United States

Abstract

Conditionals are statements of the form "If P, Then Q". Reasoning about conditionals is a core component of human cognition. However, studies of how adults and children interpret and use conditionals have highlighted discrepancies between human reasoning and logic inference rules. Recently, Douven and Verbrugge (2010) have found that a classification of conditionals based on the type of inferential connection between the antecedent and the consequent (e.g., deductive, inductive and abductive conditionals) allowed for a finer analysis of adult conditional reasoning. Do these findings extend to child conditional reasoning? We report a study (N=200, ages 4 to 11) that examines how performance in modus ponens and modus tollens tasks depends on the type of conditional embedded in the argument. These results will shed light on how the development of conditional reasoning in children is sensitive to the nature of the inferential relationship of conditionals.

Looks delicious? Cerebral blood flow in young adults with eating disorder tendencies on exposure to food pictures

Kozue Miyashiro
Utsunomiya University, Utsunomiya-City, Tochigi Prefecture, Japan

Reiko Ohmori
Utsunomiya University, Utsunomiya-City, Tochigi Prefecture, Japan

Satoko Shiraishi
Utsunomiya University, Utsunomiya-City, Tochigi Prefecture, Japan

Yumiko Ishikawa
Utsunomiya University, Utsunomiya-City, Tochigi Prefecture, Japan

Abstract

We examined the physiological changes brought on by the sight of foods in people with high eating disorder tendencies relative to normal controls. Graduate students were assessed for eating disorder tendencies using a questionnaire. Functional near-infrared spectroscopy was used to observe participants when five pictures were presented, in five categories: popular food (fried chicken), non-popular food (Japanese simmered dishes), inedible object (screw), comfortable animal (rabbit), and uncomfortable animal (cockroach). Most participants oxyhemoglobin density was found to be different in response to two pictures (fried chicken and cockroach). This indicates that this level of cerebral blood flow corresponds to unpleasant feelings. However, students with higher eating disorder tendencies showed high-level oxyhemoglobin density in the same channel, indicating discomfort, in response to popular food, neutral objects, and the uncomfortable animal. Our study implies the attitudes toward foods totally differ at cognition in people with high eating disorder tendencies compared with healthy people.

Interactive Cognitive Modeling: Understanding and Supporting Individual Human Cognition

Junya Morita
Shizuoka University, Hamamatsu, Shizuoka, Japan

Abstract

Cognitive modeling, approximation of human cognitive functions in a computational system, is a traditional methodology in the field of cognitive science. Usually this methodology has been used as a tool for scientific understanding of human mind, and evaluated by fitting to human data. In this presentation, the author proposes a framework of interactive cognitive modeling as an application of the above methodology for understanding and supporting individual human cognition. The framework consists of cognitive architecture, visualization of the model behavior, knowledge database of personal user and sensing devices to include the users reaction. This presentation shows two systems of interactive cognitive modeling in the field of web browsing and photo browsing.

Lexical iconicity facilitates word learning in situated and displaced learning contexts

Yasamin Motamedi
University College London, London, United Kingdom

Elizabeth Wonnacott
University College London, London, London, United Kingdom

Chloe Marshall
University College London, London, London, United Kingdom

Pamela Perniss
University of Cologne, Cologne, Germany

Gabriella Vigliocco
University College London, London, United Kingdom

Abstract

We present an experimental study that examines how lexical iconicity (i.e. onomatopoeia) affects early word learning, across learning contexts. Children aged 24-36 months (N=37) were first trained on labels that are either iconic or neutral with respect to the referent event, and then tested using a forced-choice task to select the correct referent given a label. We assessed learning across two contexts: situated, where label and referent co-occur, and displaced, where children learn the label following the referent event. We predicted that iconicity would aid word learning, and would have a more facilitatory effect in the displaced condition, helping the child to associate label and referent. Our findings demonstrate that children learn iconic labels in the experiment better than they do neutral labels. However, we find no difference across learning contextsiconicity facilitates word learning in both situated and displaced learning scenarios.

The Effect of Alternative Outcomes on Perceived Counterfactual Closeness

Matthew Myers
Northwestern University, Evanston, Illinois, United States

Lance Rips
Northwestern University, Evanston, Illinois, United States

Abstract

Assessing the likelihood that a counterfactual event would have happened involves contrasting a factual outcome with the counterfactual alternative. In many situations, the number of alternatives will influence the perceived closeness of a particular alternative. For example, losers of a game in which participants guess which door conceals a prize will likely believe they were closer to winning when there were three doors compared to six. This reflects accurate probabilistic reasoning because more doors will be associated with a lower probability of winning. However, we test whether the number of alternatives has a unique influence on beliefs about counterfactual closeness. Experiments 1 and 2 show that, even when probability is held fixed, people believe counterfactual closeness decreases when there are more alternatives.

On falsification and Optimal Experimental Design approaches to the value of information

Jonathan D. Nelson
University of Surrey, Guildford, United Kingdom

Vincenzo Crupi
University of Turin, Torino, Italy

Flavia Filimon
University of Surrey, Guildford, Surrey, United Kingdom

Garrison Cottrell
UCSD, La Jolla, California, United States

Abstract

There is a great deal of discussion about whether people intuitively seek to falsify their working hypothesis. But there has been little consideration of the relationships between falsificationist and probabilistic Optimal Experimental Design (OED) approaches to evaluating the usefulness of possible experiments. Recent work has shown that a variety of important OED and heuristic models can be derived as special cases of the generalized Sharma-Mittal framework of information gain measures. We show how falsification-like behavior can also derive from a quasi-information gain model, based on high-degree Tsallis entropies. Our analysis shows that falsificationist and probabilistic approaches are not as far apart as the east and the west. Rather, they can be built out of virtually the same set of ingredients, within a probabilistic framework. We report simulation studies showing how important falsificationist, OED, and hybrid models could be differentiated as possible descriptive accounts of information-seeking behavior.

Effects of implicit processes on conversion from a sub-optimal to an optimal solution

YUKI NINOMIYA
Nagoya University, Nagoya-shi, Aichi-ken, Japan

Hitoshi Terai
Kindai University, Iizuka-shi, Fukuoka, Japan

Kazuhisa Miwa
Nagoya University, Nagoya-shi, Aichi-ken, Japan

Abstract

Conversion from an initial representation for gaining insight has mainly been studied in experimental settings where solution through that initial representation is impossible.Many studies of insight problem solving have shown that an implicit process engage in conversion from an inadequate initial representation. However, few studies exist about such conversion in a situation in which solution by the initial representation is possible. A typical situation is conversion from a sub-optimal to an optimal solution. In such a situation, solution by the initial representation is inefficient, but possible. Therefore, participants received no negative feedback that the solution is impossible.In this study, by measuring eye movement, we investigated the hypothesis that the implicit process also emerges in such a situation. We found that the implicit process related to relaxation of fixedness on the sub-optional solution was observed prior to conscious finding of the optimal solution.

Bayesian Item Response Model with Condition-specific Parameters for Evaluating the Differential Effects of Perspective-taking on Emotional Sharing

Keishi Nomura
The University of Tokyo, Meguro-ku, Tokyo, Japan

Aiko Murata
Nippon Telegraph and Telephone Corporation, Atsugi, Kanagawa, Japan

Yuko Yotsumoto
The University of Tokyo, Meguro-ku, Tokyo, Japan

Shiro Kumano
Nippon Telegraph and Telephone Corporation, Atsugi, Kanagawa, Japan

Abstract

It is known that perspective-taking helps humans recognize anothers emotional state on an individual basis. Here, we investigated how perspectives influence emotional sharing, namely the act of understanding mood, or a relationship between other people in a multiparty conversation. In order to capture the effects of perspectives on sensitivity and bias in responses, we introduced condition-specific parameters in a Bayesian item response model. The model revealed that interlocutors are more sensitive and biased to emotional incongruency when they give ratings for a pair including themselves than that excluding them. This relationship holds for observers who did not participate in the conversation and took the respective perspectives. The findings support the assimilating effects of perspective-taking through which people can perceive mood as the target does.

The influence of mental fatigue on delay discounting

Samuel Nordli
Indiana University, Bloomington, Indiana, United States

Peter Todd
Indiana University, Bloomington, Bloomington, Indiana, United States

Abstract

The capacity to continually exert self control appears to become temporarily depleted over time, leading to mental fatigue and self-control failures. Some researchers have proposed that self control requires limited resources which must be periodically replenished, but no direct evidence supports this theory. An alternative explanation is that mental fatigue is an evolutionarily-adaptive feature for managing motivations, serving to temporarily disincentivize the present course (or type) of action, thereby redirecting behavior towards other goals that may better serve an individuals evolutionary fitness. Since self control is typically associated with delayed gratification and self-control failures with immediate gratification, mental fatigue may generally encourage immediately-gratifying behavior by temporarily increasing the extent to which individuals devalue all future rewards (delay discounting). To test this hypothesis, the present study examines whether delay discounting increases for participants who have recently completed a fatiguing task.

Learning Preferences as an Index of Individual Differences in Cognitive Flexibility

Hayley O'Donnell
Drexel University, Philadelphia, Pennsylvania, United States

Evangelia G. Chrysikou
Drexel University, Philadelphia, Pennsylvania, United States

Abstract

Recent findings suggest that when solving problems involving cognitive flexibility (CF), individuals who approach a learning task using reinforcement learning (RL), outperform those who approach the task using supervised learning (SL). Based on these data, we hypothesized that CF is a function of individual differences in learning preference and task demands. Healthy native English speakers were administered three CF tasks that incorporated (i) shifting, (ii) divergent thinking, or (iii) both shifting and divergent thinking elements. Participants response selection history on a reward-based learning task, which could be approached either through SL or RL, was used to determine each participants learning style and predict CF performance. Results showed that different CF task components (i.e., whether the task involved divergent thinking) interacted with participants learning preferences as measured by the independent learning task. We discuss how learning preferences might capture individual differences in CF.

An Engineered Approach: Examining the Role of Child-directed Speech With Automatic Speech Recognition and Network Science

Erick Oduniyi
University of Kansas, Lawrence, Kansas, United States

Rebekah Manweiler
University of Kansas, Lawrence, Kansas, United States

Jonathan Brumberg
University of Kansas, Lawrence, Kansas, United States

Abstract

Language acquisition is a significant developmental process children undertake automatically but is only partially understood. Though researchers have long debated the influence of internal knowledge and external stimuli in language acquisition, both features are required for this process. External stimuli are dominated by child-directed speech for the first few years of life. Accordingly, the role of child-directed speech (CDS) in early language acquisition continues to attract cognitive and developmental researchers. Here, we use statistical and computational tools from Automatic Speech Recognition (ASR) and Network Science to explore the statistical nature of CDS. In particular, we examine CDS using two complementary computational approaches: a bottom-up approach using ASR as a representation of auditory processing, and a top-down approach using networks to represent semantic and syntactic knowledge. Exploring CDS with both methods offers the unique opportunity to model the role of CDS in language acquisition from a more holistic perspective.

The Influence of Emotional Cues on Toddler Word Learning

Marissa Ogren
University of California, Los Angeles, Los Angeles, California, United States

Catherine Sandhofer
University of California, Los Angeles, Los Angeles, California, United States

Abstract

Prior research indicates that the physical context in which a word is spoken can influence how well young children learn the word. Yet, it is unclear how variability in social contexts (e.g. emotion) may impact word learning. To assess this, the present study used a novel noun generalization task with 2-year-old children. Participants were randomly assigned to one of four emotional labeling conditions: consistently angry, consistently happy, consistently sad, or variable (one label in each emotional tone per trial). We investigated whether the number of correct responses out of eight trials varied by emotional condition. Preliminary data from 28 (14 female) participants suggests that the percentage of correct responses in the sad (59.4%) and happy (64.3%) conditions may be lower than in the angry (70.8%) or variable (69.6%) conditions. These results hold implications for how emotional contexts may influence childrens ability to learn new words.

Modeling Intuitive Teaching as Sequential Decision Making Under Uncertainty

Pamela Osborn Popp
New York University, New York, New York, United States

Todd Gureckis
New York University, New York, New York, United States

Abstract

Informal teaching is a ubiquitous social behavior with a rich evolutionary history. We model teaching as the decision making problem of planning a sequence of actions to convey information to a naive learner. We compare humans intuitive teaching actions in a simple collaborative game to the optimal solution of a Partially Observable Markov Decision Process (POMDP). In a teaching POMDP, the current state is the latent, unobservable knowledge of the student and pedagogical actions may yield changes in that knowledge or provide partial information about the students state. In our experiment, human teachers balance assessment and instruction while incorporating prior information about student knowledge. Viewing teaching as a POMDP suggests specific predictions for when different teaching actions (e.g., testing versus instruction) should be preferred under different conditions. Improving our understanding of the decision making strategies that underlie intuitive teaching has a range of implications from education to clinical rehabilitation.

Congruency Effects and Individual Differences in Bilingual Experience Influence Simon Task Performance

Pauline Palma
McGill University, Montreal, Quebec, Canada

Jason Gullifer
McGill University, Montral, Quebec, Canada

Naomi Vingron
McGill University, Montral, Quebec, Canada

Veronica Whitford
University of Texas at El Paso, El Paso, Texas, United States

Deanna Friesen
The University of Western Ontario, London, Ontario, Canada

Debra Jared
The University of Western Ontario, London, Ontario, Canada

Debra Titone
McGill University, Montral, Quebec, Canada

Abstract

Prior work examining executive control during the Simon task has focused on global congruency alone and/or has primarily contrasted bilinguals with monolinguals. This is problematic for two reasons: (1) prior trial experience on current trial performance is unaccounted for (Grundy et al., 2017) and (2) bilinguals are not a homogeneous group. Here, we examined the interaction between prior and current trial congruency in the Simon Task for 65 bilingual young adults who varied continuously in bilingual experience. Generally, current trial congruency effects were larger when the prior trial was congruent vs. incongruent. However, as non-L1 experience increased, this interaction diminished; the overall prior trial effect was reduced independently of age of acquisition. Crucially, neither non-L1 experience nor age of acquisition influenced current trial congruency alone. Although preliminary, these results suggest that both congruency effects and bilingual experience influence performance on a non-linguistic executive control task.

Is Font Type and General Recommendation Really Playing Role in Dyslexic Comfortable Reading?

Tereza Pailov
Masaryk University, Brno, Czech Republic, Czech Republic

Bruno Mik
Masaryk University, Brno, Czech Republic

Abstract

Different visualizations of texts have been studies within dyslexia and significant effect of font attributes have been proved. However, the newest studies show that dyslexia is not only a matter of visual or phonological deficit and could be connected to blue cone area spots. We present a study that was designed on the basis of previous published articles and recommendations. Participants were splitted into two groups of dyslexic and nondyslexic readers. We measured reading time, comprehension and personal preferences of font types. The results show that the fastest reading time does not correspond with highest preference. Moreover, we have an interesting observation concerning preferences and reading time of participants with computer science background. This article brings new insights which could serve for further research and new design of effect of font type studies and can support blue cone theory and critical role that different languages play in dyslexia.

Semi-supervised Learning with 2D Categories

John Patterson
Binghamton University, Binghamton, New York, United States

Kenneth Kurtz
Binghamton University, Binghamton, New York, United States

Abstract

Research has shown that 1D category representations acquired through supervision change after unsupervised exposures that suggest a different boundary. However, it is unclear whether this effect generalizes to categories in which multiple dimensions are relevant. To address this question, we trained participants on a 2D information integration structure (a diagonal boundary) under supervision. Participants then classified unsupervised items that implied either a steeper or flatter boundary than that established by supervision creating a conflict region where items should switch membership. Participants classified a grid of the stimulus space both immediately before (pretest) and after (posttest) unsupervised learning to assess for differences. We found that conflict-region items were more likely to be classified as members of the opposite class on the posttest, relative to pretest in a manner consistent with the unsupervised learning condition. Implications of these findings for semi-supervised learning research and theories of category learning are discussed.

Five aspects of compositionality and a universal principle

Steven Phillips
National Institute of Advanced Industrial Science and Technology, Tsukuba, Ibaraki, Japan

Abstract

Compositionality supposedly explains structure-sensitive features of cognition, such as productivity and systematicity. However, the nature of compositionality is still controversial: e.g., symbolic versus subsymbolic. Category theory—a formal theory of structure—provides an explanation for systematicity in terms of universal morphisms: the optimal factorization of cognitive components (Phillips & Wilson, 2010). We survey five aspects of compositionality as they relate to formal properties of universal morphisms. The emerging view is a unified (universal) principle for compositionality. This category theoretical view affords a novel perspective on the emergence of symbol systems, i.e. as the construction of universal morphisms, which is illustrated in regard to some empirical data.

Reference: Phillips & Wilson (2010). Categorial compositionality: A category theory explanation for the systematicity of human cognition. PLoS Computational Biology, 6(7), e1000858. doi:10.1371/journal.pcbi.1000858

Scheduling an Information Search: Heuristics and Meaningful Metrics

Toby Pilditch
University College London, London, United Kingdom

Alice Liefgreen
University College London, London, United Kingdom

Abstract

Many domains involve gathering evidence, from forensic investigations and medical diagnosis, to everyday life. How should one order this collection, given the costs involved (e.g. time, financial, information)? Scheduling theory offers optimal solutions, but requires clear metrics. Evidence can have many influences on it, which affect prioritization, e.g. degradation, contamination, etc. However, to date there has been no clear way to bring this into a unified metric, and thus optimal scheduling has remained out of reach. We propose a new information-based measure, KL, as a way of encapsulating these information costs, and present maximum KL preservation as a clear rule & metric for scheduling. We go on to test several heuristic rules for scheduling evidence collection, based on optimally derived algorithms, providing novel formal backing for a dominant heuristic strategy for scheduling information gathering.

Mental simulation: A cognitive linguistic approach to language teaching

Laura Pissani
Concordia University, Montreal, Quebec, Canada

Abstract

This paper illustrates the neural mechanisms underlying language processing. Based on evidence from neuroscience, the Neural Theory of Language supports the idea that, to fully understand an utterance, one should be able to imagine the scene evoked by that utterance. To achieve that, brain regions responsible for the action associated with that utterance are activated in order to mentally simulate the action that is being described. In this report, I propose four activities that implement these findings to language teaching in order to boost the learning process and provide meaningful content, not only about language itself but also about the processes behind.

Ordinality trumps cardinality: What we spatialize when we spatialize numbers

Benjamin Pitt
UC Berkeley, Berkeley, California, United States

Daniel Casasanto
Cornell University, Ithaca, New York, United States

Abstract

People implicitly map numbers onto space, but what aspect of numbers do people spatialize? When cardinality (i.e. magnitude; 5 objects) is pitted against ordinality (i.e., sequential position; the 5th object), people show an implicit ordinality mapping, at least in lateral space. We hypothesized that if people spatialize numerical magnitude at all, they should do so on the vertical axis, according to the way they talk about numbers (i.e. low, high). Participants memorized sequences of randomized numbers (e.g. 85913) and then classified them (as small or large) using two response keys, oriented either laterally or vertically. Participants showed reliable ordinality mappings on both axes; they were faster to press the left/upper key for numbers earlier in the memorized sequence and the right/bottom key for later numbers, regardless of numbers magnitudes. People map exact numbers onto both lateral and vertical space according to their ordinality.

The Diagram Disconnect: An Examination of Note-Taking Behaviors In College Students

Blaire Porter
Emory University, Atlanta, Georgia, United States

Julia Wilson
Emory University, Atlanta, Georgia, United States

Hilary Miller
Emory University, Atlanta, Georgia, United States

Patricia J. Bauer
Emory University, Atlanta, Georgia, United States

Abstract

Note-taking in college courses is prevalent yet often ineffective. One potential reason is a disconnect between the information in lectures and that recorded in notes. Whereas science-based lectures frequently include diagrams, students notes often fail to include them. This disconnect likely inhibits learning and may be exacerbated by digital note-taking. We investigated students note-taking during two mini neuroscience lectures and its relation to recall. Students were assigned to diagram presence (diagram embedded in notes for first or second lecture) and note-taking method (typed or handwritten) conditions. Students recalled more in the diagram first condition. There was no recall difference based on note-taking method. Including diagrams in notes for the first lecture likely primed participants to attend to diagrams in the subsequent lecture, helping them realize the importance of the diagram. The lack of a note-taking method effect is inconsistent with past research, but may reflect increasing use of digital note-taking.

Parent comparison and contrast speech is affected by variation of present visual display and child language comprehension

Gwendolyn Price
University of California, Los Angeles, Los Angeles, California, United States

Catherine Sandhofer
University of California, Los Angeles, Los Angeles, California, United States

Abstract

Sometimes parents use comparison in speech to children and sometimes they do not. Comparison has been shown to have multiple benefits for learning. This study investigates what types of situations afford and engender parent comparison talk to 12 children 20 to 24 months of age in a free form picture book context. Each page contained three pictures that varied on color and/or object. Parent speech was analyzed for color, object, question/statement use, and comparison/contrast use. Childrens color comprehension and MCDI score were also measured. The results indicated a quadratic relationship where parents used comparison and contrast more often when their children knew few or many color words. Parents also used comparison more when the page had one dimension held constant across pictures. The results of this study inform existing understanding of comparison and demonstrate how this speech correlates with childrens understanding of language, and specifically color words.

(Mis)interpretations of implausible passive sentences pattern with N400 amplitudes

Milena Rabovsky
Freie Universitt Berlin, Berlin, Germany

Kazunaga Matsuki
BEworks (Behavioral Economics Works), Toronto, Ontario, Canada

Ken McRae
University of Western Ontario, London, Ontario, Canada

Abstract

Representations formed during language comprehension do not always accurately reflect the linguistic input, but are sometimes just good enough (Ferreira et al., 2003). Here, we examined the electrophysiological correlates of such heuristic processing. Participants were presented with passive sentences where the plausibility of the fillers of the agent and patient thematic roles was manipulated. As expected, they made more errors in the interpretation of implausible sentences (e.g., The doctor was treated by the patient). Intriguingly, N400 amplitudes patterned with (mis)interpretation, with increased amplitudes to the second noun in correctly processed implausible sentences, and equally small amplitudes in plausible sentences and in incorrectly interpreted implausible sentences. These results are in line with the view that N400 amplitudes reflect the change in an initial heuristic representation of sentence meaning (Rabovsky et al., 2018), but seem difficult to explain by accounts suggesting that the N400 reflects lexical retrieval (Brouwer et al., 2017).

Working memory, strategy, and distraction on gF tasks

Megan Raden
Mississippi State University, Starkville, Mississippi, United States

Andrew Jarosz
Mississippi State University, Mississippi State, Mississippi, United States

Abstract

Recent work suggests that strategy differences may play an important role on gF tasks and are related to WMC. The present study utilized eye tracking to assess the consistency of strategy use across tasks, focusing on constructive matching (CM) and response elimination (RE) strategies. Across two gF tasks (the Raven Matrices and a figural analogies task), participants were highly consistent in their strategy use, regardless of WMC. However, high-WMC individuals were more likely to utilize the CM strategy, though this was influenced by task order. Those who utilized RE were more likely to have their attention captured by salient, incorrect responses in the response bank and time on those responses was negatively related to accuracy. However, on select items where the response bank was necessary to make a response, these relationships disappeared. Results are discussed in terms of the implications of strategy differences on our understanding of WMC and gF.

Modeling students' fraction arithmetic strategies using inverse planning

Anna Rafferty
Carleton College, Northfield, Minnesota, United States

Rachel Jansen
University of California, Berkeley, Berkeley, California, United States

Tom Griffiths
University of California, Berkeley, Berkeley, California, United States

Abstract

Fraction arithmetic is a challenging topic for students. Past work has found that many errors can be accounted for by a limited number of malrules, reflecting both execution errors and incorrect strategies (Braithwaite, Pyke, and Siegler 2017). We develop an inverse planning model for fraction arithmetic that computes students' affinity for particular malrules based on their problem solutions. Inverse planning models people's choices when solving problems, and has been used to model data from solving algebraic equations and playing educational games. The output of the fraction arithmetic inverse planning model gives a more detailed assessment of a student's knowledge than the number of problems she answers correctly, and does not require human interpretation of students' solutions. Applying the model to the two datasets in Braithwaite et al. (2017) and inferring tendencies to use two specific malrules shows that its output is consistent with manual annotations of students' strategies.

Individual spatial reasoning skills support different kinds of physics tasks

Ilyse Resnick
University of Canberra, Bruce, ACT, Australia

Daniel Jackson
Pennsylvania State University Lehigh Valley, Center Valley, Pennsylvania, United States

Abstract

The majority of undergraduate students fail to achieve a basic understanding of fundamental concepts in science, technology, engineering, and mathematics (Bao et al., 2009). A major barrier may be spatial reasoning (Wai, Lubinski, & Benbow, 2009). Spatial reasoning is the ability to mentally manipulate the 2D and 3D relations within and between objects. The current study examines the casual relation between spatial reasoning and performance in an undergraduate introductory physics course. All students enrolled in the course took tests of mental rotation, hidden figures, form board, and perspective-taking at the beginning of the semester and again at the end of the semester. Post-test scores were significantly higher compared to pre-test scores, ts(38) ¡ 10.82, ps¡.02. Growth in spatial reasoning is predictive of exam performance, with performance on individual spatial reasoning tests being correlated with specific kinds of exam items. This suggests individual spatial reasoning skills differentially support different physics understanding.

Geometric Significance of Topological Neighborhood in Standard and Oscillating SOM Models

Spyridon Revithis
UNSW, Sydney, Australia

Abstract

The role of Topological Neighborhood (TN) in SOM cognitive modeling has biological and computational implications. The modeling significance of the TN width function (epoch) is associated with the initial TN width parameter 0. Furthermore, 0 is decisive in determining the geometric area under the TN-width function curve through the epochs of SOM training; measures training "opportunity". From this perspective, what is considered narrow (or wide) TN during SOM formation is a function of the TN width area covered.

In computer simulations of standard-TN SOM and of our previously proposed oscillating-TN SOM models, we calculated the area using the Riemann integral of the corresponding (epoch) function (standard, oscillating) and epoch-interval. The results show: a) for the same 0 and epoch-interval, the value remains unchanged irrespective of the (epoch) function used; b) when reducing 0, it reduces and directly affects the SOM representation of the input space.

Neuromodulation of electrophysiological correlates of reinforcement learning in humans

Patrick Rice
University of Washington, Seattle, Washington, United States

Mathi Manavalan
University of Washington, Seattle, Washington, United States

Andrea Stocco
University of Washington, Seattle, Washington, United States

Abstract

The feedback-related negativity (FRN) is an event-related potential that differentiates between positive and negative feedback, occurring most prominently at frontocentral electrodes 200-300ms after delivery of feedback. The FRN seems to be reflective of a reward prediction error, as the magnitude of the ERP component has been related to the magnitude of prediction error estimated through reinforcement learning (RL) models. We aim to further understanding of the FRN and its relationship to behavior by replicating the study of Reinhart & Woodman (2014), replacing tDCS with focal, targeted transcranial magnetic stimulation (TMS) over the frontocentral region. Preliminary data shows that our participants reliably generate a FRN when presented with incorrect feedback, and that single-trial estimates of theta power are significantly correlated with RL-derived single-trial estimates of prediction error for correct trials. We will examine the effect of stimulation both on participant behavior as well as on RL parameter estimates.

Do Verbal Labels Enhance Detection of Visual Targets?

Catherine Richards
University of Delaware, Newark, Delaware, United States

James Hoffman
University of Delaware, Newark, Delaware, United States

Timothy Vickery
University of Delaware, Newark, Delaware, United States

Anna Papafragou
Unversity of Delaware, Newark, Delaware, United States

Abstract

Cognitive penetrability describes cognition and perception as interconnected, with cognition impacting the process of perception rather than just the interpretation. The current study addresses this claim in the domain of language, asking if language helps people detect nearly-invisible stimuli. Two experiments were adapted from Lupyan and Spivey (2010), where auditory cues were found to be more beneficial than visual cues in recognizing letters. Participants reported the presence of a target letter that was either preceded by an auditory or visual cue (e.g., cues were either hearing emm or seeing M, followed by a visual M as a target). Detection sensitivity was calculated and compared within cue presentation type. Neither visual nor auditory cues helped participants recognize target letters more than the no-cue condition. These results differ from previous work demonstrating linguistic facilitation and indicated that neither linguistic nor visual information aid in perceiving a matching item.

Categorical rhythms shared between songbirds and humans

Tina Roeske
Max Planck Institute for Empirical Aesthetics, Frankfurt, Hessen, Germany

Ofer Tchernichovski
Hunter College, New York, New York, United States

David Poeppel
Max Planck Institute for Empirical Aesthetics, Frankfurt, Deutschland, Germany

Nori Jacoby
Max Planck Institute for Empirical Aesthetics, Frankfurt, Deutschland, Germany

Abstract

Rhythm the organization of sounds in time is a universal feature of human music. Of the infinite ways of organizing events in time, human rhythms are distributed categorically. We compared rhythms of classical piano playing and finger tapping to rhythms of thrush nightingale songs. Across species, we found similar common rhythms, as relative durations of intervals formed three categories: isochronous 1:1 rhythms, small integer ratio rhythms, and high ratio ornaments. In both species, those categories were invariant within extended ranges of tempi, indicating natural classes. In all cases, the number of rhythm categories decreased with higher tempi. Finally, in birdsong, high ratios (ornaments) were derived from very fast rhythms containing inflexible (probably uncontrollable) interval ratios. These converging results indicate that birds and humans similarly create simple rhythm categories from a continuous temporal space. Such natural categories can promote cultural transmission of rhythmic sounds a feature that songbirds and humans share.

Socio-economic related differences in the use of variation sets in naturalistic child directed speech. A study with Argentinian population

Celia Rosemberg Rosemberg
Consejo Nacional de Investigaciones Cientficas y Tcnicas-CONICET- (National Council of Scientific Research, Argentina), Ciudad de Bs. As., Ciudad de Buenos Aires, Argentina

Florencia Alam Alam
National Council of Scientific and Technical Research, Buenos Aires, Capital federal, Argentina

Leandro Garber
CONICET, Ciudad de Bs. As., Argentina

Alejandra Stein
CONICET, Buenos Aires, Argentina

Maia Julieta Migdalek
Consejo Nacional de Investigaciones Cientficas y Tcnicas (CONICET, Ciudad de Buenos Aires, Ciudad de Buenos Aires, Argentina

Abstract

Child-directed speech (CDS), compared to speech between adults, shows a higher amount of repetitiveness, particularly of sequences of utterances with self-repetitions. This phenomena, known as variation sets, has been found to be beneficial for learning. Although previous findings indicated socio-economic status (SES) effects on the quantity of variation sets, they were based on data from child-parent dyadic interactions in play situations. Given that SES comprises interrelated factors affecting childrens quotidianity, here we examine SES effects on the use of variation sets in long recordings of the family naturalistic environment of 30 low and middle SES Argentinian children (8 to 20 months). Variation sets were automatically extracted from CDS provided by all the participants. Results demonstrated the effects of two factors related to SES-differences: while parents education showed a positive relation to the quantity and extension of variation sets, the number of people living in the household influenced it negatively.

Modelling eye tracking dynamics with quantum theory

Agnes Rosner
University of Zurich, Zurich, Switzerland

Irina Basieva
City Universtiy London, London, United Kingdom

Albert Barque-Duran
City University London, London, United Kingdom

Andreas Gloeckner
University of Cologne, Cologne, Germany

Bettina von Helversen
University of Zurich, Zurich, Switzerland

Andrei Khrennikov
Linnaeus University, Kalmar, Sweden

Emmanuel Pothos
City, University of London, London, United Kingdom

Abstract

Eye movements during decision making show systematic patterns such as increased fixations to the chosen option (i.e. gaze cascades) and multiple gaze transitions between fixated options. Existing formalisms, such as multivariate decision field theory, only provide limited scope for describing multiple reversals in the attentional focus and it is therefore unclear how they can be applied to the underlying attentional dynamics. Here, we present an open systems dynamical model from quantum theory to describe gaze transitions between choice options and the gaze cascade effect. Our model was tested on a decision task, in which participants repeatedly decided among two complex options (i.e. that lacked easily quantifiable, matched characteristics). The model can describe the gaze patterns on the individual trial level. It reveals structure in the gaze dynamics that is predictive for choice behavior. The explanatory value of this account for studying attentional dynamics during decision making will be discussed.

Priming Effects on the Interpretation of Ambiguous Discourse Relations

Eyal Sagi
University of St. Francis, Joliet, Illinois, United States

Abstract

Many theories of discourse structure rely on the idea that the segments comprising the discourse are linked through inferred relations such as causality and temporal contiguity. These theories often suggest that the information needed to determine the relation can be found when the discourse is interpreted through the application of world knowledge. However, Sanders (1997) found that the interpretation of ambiguous relations can be affected by the discourses genre. Similarly, Sagi (2006) reported that participants were faster to interpret discourse relations when they were preceded by the same discourse relation. The present study demonstrates that exposure to discourse relations such as result (e.g., John passed Mark in a marathon. He won.) or explanation (e.g., John ... He was in great shape.) can affect the interpretation of subsequent ambiguous relations encountered in an unrelated context. This result suggests that discourse relations are represented independently of the context in which they appear.

Animal Vocalization Generative Network (AVGN): A method for visualizing, understanding, and sampling from animal communicative repertoires

Tim Sainburg
University of California, San Diego, La Jolla, California, United States

Marvin Thielk
University of California, San Diego, La Jolla, California, United States

Timothy Gentner
University of California, San Diego, La Jolla, California, United States

Abstract

We propose here a set of machine-learning algorithms to produce a generative low-dimensional and visually-understandable space of the communicative repertoire of vocal species such as songbirds. As opposed to human speech, where individual elements are well defined and grounded in principled ways, the methods for defining units of animal communication systems are often more varied and rely on human-centric heuristics. Using our method, we can automatically discover latent structure in the vocal repertoire of individuals and use these to define-well principled categorical boundaries between vocal elements in communicating species. Further, we can sample from latent representations to generate novel vocal units that can be used to probe perceptual and physiological representations of communication. We demonstrate two use cases: (1) automated labeling of songbird vocal repertoires showing novel structure in vocal communication, and (2) a perceptual task demonstrating that behavioral and physiological representational spaces can be biased by contextual information. GitHub.com/timsainb/AVGN

Reducing Smartphone Overuse through Behavioural Nudges

Dasha Sandra
McGill University, Montreal, Quebec, Canada

Jay A. Olson
McGill University, Montreal, Quebec, Canada

Denis Chmoulevitch
McGill University, Montreal, Quebec, Canada

Signy Sheldon
McGill University, Montreal, Quebec, Canada

Amir Raz
Chapman University, Orange, California, United States

Samuel Veissire
McGill University, Montreal, Quebec, Canada

Abstract

We identified smartphone usage patterns predicting overuse and developed an intervention to reduce these effects. In Study 1, 54 undergraduate students reported their daily screen time and the reasons for their smartphone use. A cluster analysis revealed two usage patterns: as a tool (e.g., for directions), and to socialize or pass time. Only the latter pattern correlated with daily phone use (r=.35). In Study 2, 28 pilot participants underwent a two-week-long behavioural intervention involving disabling non-essential notifications and keeping their phone out of reach when not in use. All participants complied with these guidelines, leading to a 1.2 hours/day reduction in usage (4h to 2.8h), a decrease in smartphone addiction scores to normal levels, and a 30% decrease of scores on the Beck Depression Inventory-II (10.1 to 7). We explore potential cognitive benefits of the intervention on memory and attention (measured by Operational Span and Sustained Attention to Response tasks).

The Price of Good Intentions

Arunima Sarin
Harvard University, Cambridge, Massachusetts, United States

Fiery Cushman
Harvard University, Cambridge, Massachusetts, United States

Abstract

Prior work has shown that positively intentioned agents are held more responsible, causal, and blameworthy for subsequent bad outcomes than negatively intentioned agents are held for good outcomes. Across a series of studies, we investigate the underlying expectations that produce this asymmetry. We find that, in in the absence of explicit information about the action performed, actions of positively intentioned agents who produce bad outcomes are inferred to be worse than actions of negatively intentioned agents who produce good outcomes (Study 1). While both agents are judged to be incompetent (Study 2), positively intentioned agents are attributed more control over subsequent negative outcomes (Study 3) and are also considered more pivotal in bringing them about (Study 4). Together these results suggest that well-intentioned agents are seen as having more control, perhaps because, we believe they are in a better position to modify their future behavior to bring about positive outcomes.

The posterior probability of a null hypothesis given a statistically significant result

Daniel Schad
University of Potsdam, Potsdam, Germany

Shravan Vasishth
University of Potsdam, Potsdam, Germany

Abstract

When researchers carry out a null hypothesis significance test, it is tempting to assume that a statistically significant result lowers Prob(H0), the probability of the null hypothesis being true. Technically, such a statement is meaningless for various reasons: e.g., the null hypothesis does not have a probability associated with it. However, it is possible to relax certain assumptions to compute the posterior probability Prob(H0) under repeated sampling. We show that the intuitively appealing belief, that Prob(H0) falls when significant results have been obtained under repeated sampling, is in general incorrect and depends greatly on: (a) the prior probability of the null being true; (b) Type I error, and (c) Type II error. Through simulation we quantify uncertainty and find that uncertainty about the null hypothesis often remains high despite a significant result. To help the reader develop intuitions about this common misconception, we provide a Shiny app (https://danielschad.shinyapps.io/probnull/).

Temporal dynamics of preschoolers novel word learning and categorization

Christina Schonberg
UW-Madison, Madison, Wisconsin, United States

Haley Vlach
University of Wisconsin-Madison, Madison, Wisconsin, United States

Abstract

Word learning paradigms often teach children the name of a novel object and then immediately ask them to generalize the label to another object. This study uses a new paradigm that affords the ability to determine how childrens generalization changes over time. Participants (N=22, Mage=3.8 years) saw a novel object labeled by the experimenter (e.g., wug) and then were shown five novel objects that each had an additional feature changed from the exemplar (i.e., the fifth object had five changed features), either immediately after the exemplar or after a five minute delay. Category membership endorsement of the five test objects was higher at immediate test than delayed test, suggesting that children represent novel categories broadly at first but more narrowly over time. We propose that childrens forgetting of exemplars across time leads to shifts in childrens generalization; as children forget exemplar features, category membership becomes more specific.

Spatial Preferences in Everyday Activities

Holger Schultheis
University of Bremen, Bremen, Germany

Abstract

Many everyday activities pose only weak constraints on the order, in which certain actions have to be performed. When setting the table, for example, any order of putting the required items on the table will be fine as long as all necessary items are on the table eventually. Despite the commonality of weakly constrained sequences in everyday activities, little is known about how humans deal with such sequences. In this contribution, we argue that humans do not order weakly constrained actions arbitrarily, but exhibit systematic patterns of orderings, which we term ordering preferences. Moreover, we argue that the task environment's spatial layout and its mental representation are key factors in determining such preferences. An initial empirical study on table setting corroborates this reasoning by revealing ordering preferences that seem to be based on a regionalization of space and the distances between the regions.

Using eye-tracking to examine the role of fluency in the number line placement task

Samantha Schwarz
Susquehanna University, Selinsgrove, Pennsylvania, United States

Jennifer Asmuth
Susquehanna University, Selinsgrove, Pennsylvania, United States

Abstract

The number line placement task, in which individuals are presented with a target number and mark where it would be located along a number line, has played an important role in the investigation of numerical cognition. However, recent work suggests that different factors may influence performance on the task, making it a poor proxy for mental representation of number. In this study, adults completed a computer-based number line placement task with either standard or non-standard endpoints. Consistent with previous research, responses in the standard condition were best fit by a linear model, while responses in the non-standard condition were best fit by a logarithmic model. In addition, eye-tracking data revealed different looking patterns between conditions, including greater fixations on and more frequent alternation between endpoints in the non-standard condition and a leftward bias in the standard condition. This behavior may reflect differences in number familiarity and strategy use.

The Visual Representation of Abstract Verbs: Merging Verb Classification with Iconicity in Sign Language

Simone Scicluna
University of Trento, Trento, Italy

Carlo Strapparava
FBK-Irst, Trento, Italy

Abstract

Theories like the picture superiority effect prove that visual information is vital in the acquisition of knowledge, such as in language learning. Words can be graphically represented to illustrate the meaning of a message and facilitate its understanding, but this rarely applies to abstract words. The current research turns to sign languages to explore the common semantic elements that link abstract words to each other, pointing towards the possibility of creating clusters of iconic meanings. By using sign language insight and VerbNets organisation of verb predicates, this study presents a novel organisation of 500 English abstract verbs classified by visual shape. Graphic animation was used to visually represent 20 classes of abstract verbs (see on www.vroav.online). An online survey was created to achieve judgements on the graphic visuals representativeness. Significant agreement between participants suggests a positive way forward for further research and applications within multimodal communication and computer assisted learning.

Mathematical Creativity: Incubation, Serial Order Effect, and Relation to Divergent Thinking

Stacy Shaw
Univeristy of California, Los Angeles, Los Angeles, California, United States

Gerardo Ramirez
Ball State University, Muncie, Indiana, United States

Abstract

The current study explored whether creative processes specifically incubation and the serial order effect extend to creativity in mathematics, and if there is a relation to divergent thinking. A total of 155 postsecondary students completed an unusual use task and a multiple-strategy math task. Participants were given 8 minutes to generate as many strategies as they could for the math task, and then after a brief break, were given another problem with the same underlying structure for 4 minutes. We find evidence for a serial order effect in math, whereas across trials it became more difficult for participants to generate a new strategy, but the strategies were rated as more creative. The brief break also provided some evidence of incubation, as there was a boost in the number of overall strategies and creativity. We also found that divergent thinking and mathematical creativity were significantly related.

When Experts Err: Using Tetris Models to Detect True Errors From Deliberate Sub-Optimal Choices

Catherine Sibert
Rensselaer Polytechnic Institute, Troy, New York, United States

Wayne Gray
Rensselaer Polytechnic Institute, Troy, New York, United States

Abstract

Error detection and correction is a vital part of skill acquisition, but when training a complex, real time, dynamic task, it can be difficult to isolate a true mistake in a sequence of decisions without clear correct choices. We use previously developed high-performing, human-like models of the video game Tetris (Sibert et al., 2017) to analyze individual piece placement decisions for players of high and low skill. In cases where the model's choice differed from the human's choice, we examine the eye fixations made during the placement decision to determine if the disagreement is caused due to the player performing at lower level than the model (i.e. not being aware of a better placement), the player performing at a higher level than the model (i.e. deliberately making a suboptimal move in service of a long term strategy), or the player making a true error.

Instructions to Incorporate Music Themes into a Haiku Increases Perceived Creativity of the Haiku

Cynthia Sifonis Sifonis
Oakland University, Rochester, Michigan, United States

Paul Sullivan
Oakland University, Rochester, Michigan, United States

Abstract

The current research examines the degree to which thematic/referential music affects performance in Amabiles American Haiku task. Thematic music conveys meaning to the listener by activating concepts associated with the music in semantic memory. Ward (1994) demonstrated that generating novel exemplars is influenced by activated concepts in memory. Consequently, participants listening to thematic music before writing a haiku should be more likely to incorporate thematic elements into the haiku which increases the perceived creativity of the haiku. Participants specifically instructed to incorporate thematic elements into the haiku should include more thematic elements and write more creatively than participants not instructed to include thematic elements and participants who wrote their haiku without having listened to thematic music beforehand. 206 undergraduates listened to a 90 second sample of unfamiliar lullaby- or war-themed music. Participants were instructed to write a haiku inspired by the music (Inspire), write a haiku after listening to the music (Neutral) or write a haiku before listening to the music (Control). We found a significant main effect of the Inspire instruction on incorporation of thematic elements into the haiku. Participants in the Inspire condition included significantly more thematic elements of the music into their haiku than participants in the Neutral condition or Control conditions. Participants in the Inspired condition wrote haikus that were marginally more likely to be rated as more negatively valenced and were more creative than the haikus written in the Neutral and Control conditions. Results suggest ways of increasing creativity through use of thematic music.

Flexible Strategy Use in ACT-R's Tic-Tac-Toe

Julian Skirzyski
McGill University, Montreal, Quebec, Canada

Dr Piotr Wasilewski
University of Warsaw, Warsaw, Poland

Abstract

Modeling cognitive processes is one of the major tasks of cognitive science. This work presents a computer model of a study described in "Flexible Strategy Use in Young Children's Tic-Tac-Toe" (Crowley & Siegler, 1993) in which authors made an attempt to characterize decision-making in a conflict-of-interests-like environment. In the experiments, kindergarten/primary school children and an algorithm-based opponent played a series of games in Tic-Tac-Toe. The outcomes seemed to indicate existence of a hierarchy of rules that is constructed with experience. Although already tested algorithmically, the simulation detailed in the paper was applicable to a narrow class of problems only. The model shown in this work was built using a cognitive architecture, i.e. computer-based structure mimicking general functioning of the human mind. We used a rule-based system ACT-R that operates in mental rules paradigm and successfully replicated results of the mentioned study.

Adult Prediction Error Processing is Associated with Vocabulary Size

Katherine Snelling
Queen's University, Kingston, Ontario, Canada

Stanka Fitneva
Queen's University, Kingston, Ontario, Canada

Abstract

How do individuals learn language when there are so many possible potential referents for each word? Prediction-based theories of language learning propose that predictions enable individuals to learn from implicit negative evidence by comparing the predictions to outcomes. However, the role of prediction errors for learning has yet to be established. Traditionally, prediction errors have been believed to hinder learning. Recently though, prediction errors have been associated with improved novel word acquisition in cross situational learning. This present study used a cross-situational word learning task to examine the relation between prediction error-based processes during word learning and vocabulary size. The results showed that learners who switched their gaze more quickly from the non-target to the target image when they had to detect and correct prediction errors had higher productive vocabularies. This research supports the theory that productive vocabulary is strongly tied to predictive processes.

Introducing Recursive Linear Classification (RELIC) for Machine Learning

Sean Snoddy
Binghamton University, Binghamton, New York, United States

Kenneth Kurtz
Binghamton University, Binghamton, New York, United States

Abstract

Numerous classifiers for machine learning are powerful and effectivean important path forward is decreasing the complexity and increasing the transparency of the solutions achieved. RELIC (REcursive LInear Classifer) consists of recursively applying a classifier to the training items not successfully accounted for in the previous iteration to find subsets within the training data that yield simpler classification schemes. Chooser models are iteratively added and trained on item-to-subset assignments to learn a mapping between input space and the classifier ensemble. Test examples are passed through the set of choosers to select the appropriate subset-classifier pairing to generate a classification. While applicable to any classifier, we begin by evaluating RELIC using logistic regression and linear SVM to determine whether they perform better under the recursive approach and become competitive with non-linear classifiers. Application of this approach to non-linear classifiers and potential implications for the broader science of learning are also addressed.

Compositionality in emerging multi-agent languages: Marrying Language Evolution and Natural Language Processing

Kees Sommer
Leiden University, Leiden, Netherlands

Jae Perris
Leiden University, Leiden, Netherlands

Arianna Bisazza
Leiden University, Leiden, Netherlands

Tessa Verhoef
Leiden Institute of Advanced Computer Science, Leiden, Netherlands

Abstract

The mainstream approach in NLP is to train systems on large amounts of data. Such passive learning contrasts with the way language is learnt by humans. Human language is acquired within communities, it is culturally transmitted and changes dynamically. These evolutionary mechanisms have been extensively studied in the field of Language Evolution. Despite limited prior interaction between fields, such mechanisms are now increasingly incorporated into NLP systems. Such models have the potential to both study the evolution of language in multi-agent simulations with state-of-the-art (deep) learning systems in more naturalistic settings and improve NLP systems by having language emerge organically. We examine how findings from a model by Havrylov & Titov (2017) compare to those from traditional Language Evolution models and quantify the emerging compositionality using an existing Language Evolution method (Tamariz, 2011). This approach reveals novel insights into the generated data, the applied methodology and the nature of compositionality.

Using Occam's razor and Bayesian modelling to compare discrete and continuous representations in numerostiy judgements

Jake Spicer
University of Warwick, Coventry, United Kingdom

Adam Sanborn
University of Warwick, Coventry, United Kingdom

Ulrik Beierholm
Durham University, Durham, United Kingdom

Abstract

Previous research has suggested that numerosity judgements are based not just on perceptual data but also past experience, and so may be influenced by the form of this stored information. The representation of such experience is unclear, however: numerical data can be represented by either continuous or discrete systems, each predicting different generalisation effects. This study therefore contrasts discrete and continuous prior formats within numerical estimation using both direct comparisons of computational models using these representations and empirical contrasts exploiting different predicted reactions of these formats to uncertainty via Occam's razor. Both computational and empirical results indicate that numeroisty judgements rely on a continuous prior format, mirroring the analogue approximate number system, or number sense. This implies a preference for the use of continuous numerical representations even where both stimuli and responses are discrete, with learners seemingly relying on innate number systems rather than symbolic forms acquired in later life.

Creativity and Machine Learning: Divergent Thinking EEG Analysis and Classification

Carl Stevens
University of Arkansas, Fayetteville, Arkansas, United States

Darya Zabelina
University of Arkansas , Fayetteville, Arkansas, United States

Abstract

Prior research has shown that greater EEG alpha power (8-13 Hz) is characteristic of greater creativity. This study investigates the potential for machine learning to classify more and less creative brain states. Participants completed an alternate use task, in which they thought of normal or uncommon (more demanding) uses for everyday objects (e.g., brick). We hypothesized that alpha power and reaction time would be greater for uncommon uses, and that a trained machine learning model would be able to reliably classify data from the two conditions. Participants responded much faster in the normal condition, compared to uncommon; alpha was significantly greater for the uncommon condition; and 73.3% classification accuracy was attained when a trained model was applied to new data. Future research will attempt to implement neurofeedback training to maintain optimally creative states.

Effects of Instructor Presence in Video Lectures: Rapport, Attention, and Learning

Andrew Stull
University of California, Santa Barbara, Santa Barbara, California, United States

Logan Fiorella
University of Georgia, Athens, Georgia, United States

Rebecca Similuk
University of California, Santa Barbara, Santa Barbara, California, United States

Stevi Ibonie
University of California, Santa Barbara, Santa Barbara, California, United States

Richard Mayer
University of California, Santa Barbara, Goleta, California, United States

Abstract

Do students learn better from video lectures when an on-screen instructor is socially presentthat is, when students can see the instructor's face and eye gaze during the lecture? The present study explores how access to the instructors face and eye gaze affects students feelings of social rapport, attention to the lesson, and learning outcomes. The study compares a video lecture about the human kidney where students either have access to the instructors face and eye gaze during the lecture or do not (i.e., the instructor does not face the camera). Students reported higher levels of engagement, directed more eye fixations to the lecture material rather than the instructor (based on eye-tracking metrics), and performed better on both retention and transfer posttests after viewing a video lecture with a socially present, on-screen instructor. Results suggest that social cues play a role in guiding academic learning from instructional video.

Aha! Under Pressure: Is the Aha! Experience Constrained by Cognitive Load?

Hans Stuyck
Catholic University of Leuven, Leuven, Belgium

Axel Cleeremans
Universit libre de Bruxelles, Brussels, Belgium

Eva Van den Bussche
KU Leuven, Leuven, Belgium

Abstract

Suddenly comprehending the solution to a vexing problem is often accompanied by an Aha! experience. The driving mechanisms behind this experience are unclear. One way to address this, is to study Aha! under cognitive load. If Aha! is the result of the same explicit process that we use to solve everyday problems, it should be influenced by cognitive load in a similar way. However, if it constitutes a different, more implicit process, cognitive load might not affect it at all. Using a dual-task paradigm where participants solved word puzzles under different memory loads, we found that word puzzles solved with Aha! were more accurate and led to higher solution confidence. When memory load increased, only puzzles without Aha! were solved more slowly. The fact that solution retrieval with Aha! was unaffected by memory load, implies that Aha! experiences rely on a process that does not compete for limited cognitive resources.

Eye Movement Assessment in High and Low Social Anxiety Individuals: An Eye-Tracker Study

Wei-Ling Su
National Cheng Kung University, Tainan, Taiwan

Min-Hsien Wu
National Cheng Kung University, Tainan, Taiwan

Po-Yi Chi
National Cheng Kung University , Taipei, Taiwan

Hua Feng
Graduate Institute of Rehabilitation Counseling, Changhua, Taiwan

TSE-MING CHEN
Graduate Institute of Rehabilitation Counseling, Changhua, Taiwan

Chia-Hua Chang
National Changhua University of Education, Changhua, Taiwan

Ting-Hsuan Chang
National Cheng Kung University, Tainan, Taiwan

Te-En Huang
National Cheng Kung University, Tainan, Taiwan

Jon-Fan Hu
National Cheng Kung University, Tainan City, Taiwan

Abstract

Previous studies have suggested that, socially anxious individuals tend to avoid eye contact while looking toward faces. The study designed an emotional faces task consisted of human and comic face stimuli with 6 different emotions (happy, angry, sad, scared, stunned, confused), and recorded the eye movements to examine the hypotheses above. The results revealed that high social anxiety (HSA), medium social anxiety (MSA) and low social anxiety (LSA) individuals have no significant difference on total fixation duration of the eyes, nose, and mouth among 6 different emotions. However, while focusing on the angry expression, LSA have significantly higher total fixation duration, visit count and area normalized score on the nose. It shows that LSA tend to focus on the nose intentionally when a person shows an angry face. Furthermore, HSA show lower proportion of eyes to eyes, nose and mouth fixation duration than MSA in happy, sad and stun faces.

Keywords: eye tracking, social anxiety, emotional faces

Learning to calibrate age estimates

Jordan Suchow
Stevens Institute of Technology, Hoboken, New Jersey, United States

Abstract

Age is a primary social category and, with little effort, we can quickly approximate it from photographs. Here, we analyze 1.5 million age judgments derived from a popular online website where participants estimate the age of a person depicted in a photograph, with feedback. We find that median age judgments across participants are linear in the actual age, with little bias. However, the slope is considerably less than one, such that the aggregate overestimates the age of younger people and underestimates the age of older people. Age estimates are found to be unbiased at 37.5 years, which coincides with the median age across all the depicted persons. These results are consistent with an account in which, over time, participants learn to calibrate an analogue magnitude to the learned distribution of encountered ages, combining photographic evidence with distributional information to arrive at an estimate that balances the two.

The development of compound word processing in young children

Takayo Sugimoto
Aichi University, Toyohashi, Aichi, Japan

Abstract

Hirose & Mazuka (2015 & 2017) demonstrate that Japanese speaking adults and first graders both show anticipatory compound processing, using the language-specific compound accent rule (=CAR). That is, six- to seven-year-old children can exploit compound prosody to disambiguate the structure and meaning of a given compound. However, we do not know exactly when and how children start exploiting the CAR to properly comprehend compounds. Thus, we investigated Japanese-speaking childrens acquisition of the CAR and their development of compound processing. We conducted longitudinal experiments using compound comprehension tasks on 65 Japanese-speaking children aging from two- to four-years. We found that childrens compound processing strategies changed after their acquiring the CAR. Before acquiring it, children could not identify the compound head; instead they showed a language-general parsing preference for the left-most part of a compound. Our results suggest that childrens acquisition of the language-specific CAR enables their compound processing.

Shame on you! A computational linguistic analysis of shame expressions

Anonymous CogSci submission

Abstract

The current study explored the unique linguistic characteristics of the self-conscious emotion shame. The data used for the analyses were part of two larger studies in which semi-structured interview techniques were used that had learners describe shameful or frustrating experiences in the context of psychology and engineering courses. Results revealed when describing an experience of shame, learners use significantly more positive emotional words, significantly more words associated with anxiety, and significantly fewer words associated with anger. Additionally, learners use simpler syntax, more abstract words, and have less cohesive speech. Educational implications are discussed.

Keywords: emotions; shame; learning centered emotions; cognition; computational linguistic analysis

Objective

A gap currently exists in the literature regarding a quantitative exploration of the self-conscious emotion of shame. Adding to the body of literature on negative emotions, this study explored the unique linguistic characteristics of shame and frustration with the hope that we can better understand students' experiences of these emotions.

Theoretical Framework

Language is a powerful cognitive communicative process that has been the focus of research for centuries. Speaking and writing are expressive through the specific words chosen by individuals, as well as the frequency of specific words, and become one's "style". One's linguistic style in speech and writing has been suggested to be indicative of individual differences and personality (Groom & Pennebaker, 2002). We explored differences with respect to descriptions of the emotions of shame and frustration to better understand cognitive aspects of these emotions through speech-analysis.

Linguistic Analyses

From the study of dead languages to the biological nature of language within the brain, researchers have sought to understand how humans possess complex language abilities, the impact of language on humans, and countless other aspects of human language-use. Human language is undeniably expressive in content and dialect, however, this does not account for the full expressive power of language. The style of which we speak and write is also critically expressive but is frequently unnoticed. Speaking and writing is expressive through the particular words chosen by individuals and the frequency of specific words; these linguistic styles in speech and writing have been suggested to be indicative of individual differences and personality (Groom & Pennebaker, 2002; Pennebaker & King, 1999).

The study of linguistic style and content has numerous applications, but, until recently, conducting these analysis has been a difficult task that consisted of counting and organizing words with the use of individual judges (Pennebaker, Mehl, & Niederhoffer, 2002). However, an objective analysis of language patterns through word counting software has led to an increase in our understanding of what particular parts of speech contain a deeper level that is not naturally perceived (Pennebaker & Graybeal, 2001). We believe that using this type of analyses, we can gain insight into students' experiences of emotions.

Shame

Although there are many ways to define shame, for the purposes of this study, shame is an acutely painful affective state that is brought on by a failure to meet internally set rules, ideals, goals, or standards (Turner, Husman, & Schallert, 2002). A gap currently exists in the literature regarding a quantitative exploration of shame. Of the research that has been conducted, much has been qualitative in nature and not focused on "academic" shame (i.e., shame affiliated with learning and education). One possible reason for the underdeveloped exploration of this construct is due to the difficulty in studying it. More specifically, research has shown that individuals may deny their feelings of shame, they tend to self-isolate when they feel shame, and they may be unwilling or unable to express themselves when they feel shame (citation needed). In fact, one's difficulty in communicating a shameful experience may be a distinctive characteristic of shame (Turner, 2014; Babcock & Sabini, 1990, Lunde, 1958).

Although research has suggested the *difficulties* in studying shame, the difficulty does not detract from the *importance* of studying shame. Tangney and Dearing (2002) suggested that, "Guilt, and especially shame ... are powerful, ubiquitous emotions that come into play across most important areas of life." (p. 8). Contemporary research has shown that experiences of shame can have a "negative impact on interpersonal behavior and functioning" (Tangney &

Dearing, 2002, p. 5). Within the context of education, a number of educational psychologists have asserted that feeling shame can interfere with motivation, and negatively impact students' academic goals and achievement (Pekrun, Frenzel, Goetz, & Perry, 2007; Weiner, 1986). Indeed, once students experience shame, their ability to become cognitively engaged may be hindered, they may lose motivation for studying, and, they may feel reluctant to attend class (Turner, Husman, & Schallert, 2002).

Given the importance of gaining a better understanding of this self-conscious emotion, the current study sought to compare the unique linguistic characteristics of shame with that of frustration. Our intent was to better understand the underlying composition of shame expressions.

Data Sources, Evidence, Objects, or Materials

Linguistic Inquiry and Word Count (LIWC)

The present study used a program called Linguistic Inquiry and Word Count (LIWC) to analyze speech. LIWC allows researchers to efficiently enter text files into the program in order to obtain outputs that cover a number of language indices. For example, if we were to convert *Of Mice and Men* by John Steinbeck into a text file and enter it into LIWC we would obtain the exact word count, words per sentence, and a description of approximately 90 indices. These indices are extremely insightful in objectively understanding what a text consists of and the mental state of the author or speaker (Groom & Pennebaker, 2002). For the current study, we focused only on indices that were theoretically relevant: 1) Affective processes (e.g., happy, cried) 2) Positive emotion (e.g., love, nice, sweet) 3) Negative emotion (e.g., hurt, ugly, nasty) 4) Anxiety (e.g., worried, fearful) 5) Anger (e.g., hate, kill, annoyed) and 6) Sadness (e.g., crying, grief, sad).

Coh-Metrix

Coh-Metrix, is a system for computing computational cohesion and coherence for written and spoken texts. For the purpose of the current study, we explored five specific indices within a Coh-Metrix: Narrativity, Syntactic Simplicity, Word Concreteness, Referential Cohesion, and Deep Cohesion. Narrative text tells a story, with characters, events, places, and things that are familiar to the reader. Syntactic simplicity reflects the degree to which the sentences in the text contain fewer words and use simple, familiar syntactic structures, which are less challenging to process by the reader. Word concreteness refers to texts that contain content words that are concrete, meaningful, and evoke mental images. Texts high in referential cohesion contain words and ideas that overlap across sentences and the entire text. Deep cohesion reflects the degree to which the text contains causal, intentional, and temporal connectives (McNamara, Graesser, Cai, & Kulikowich, 2011). The theoretical purpose of focusing solely on these five indices is because previous research has found that dozens of measures funnel into these five major factors (Graesser, McNamara, Cai, Conley, Li, & Pennebaker, 2014).

Methods

The data used for analysis are subsets from two larger studies. As part of one study, participants were recruited from an upper-division psychology course at a midwestern R1 university. Five-weeks into the semester, after obtaining in-class feedback on their midterm exam, students completed a survey (Experiential Shame Scale, Turner, 2014, Cronbach's alpha = .86) to determine the extent to which they perceived their grade was a failure and if they were experiencing the emotion of shame. Eight students, who indicated they experienced shame after their midterm exam, agreed to participate in semi-structured interviews two weeks before the final exam. All interviews were recorded and transcribed verbatim.

We compared the shame interviews with that obtained in a second study, one that used an interpretative phenomenological analysis (IPA) of students' experiences of frustration in the context of college-level science and engineering courses. The semi-structured interviews were conducted by an undergraduate student who had been extensively trained to conduct phenomenological interviewing. Select portions of these interviews comprised our frustration corpus (n = 5) (Huff & Clements, 2018).

The interviews from both studies were approved by the IRB offices of each investigator for the respective studies. Additionally, the procedures of the present investigation were approved by lead author's institutional IRB.

Results

Results from our LIWC analysis indicated that students describing a shameful experience tended to use more positive emotional words than students describing a frustrating experience, t (11) = 1.629, p = .06 (one-tailed), d = .92. Shame-describing students also used significantly more words associated with anxiety than students who described their frustration, t (11) = 2.644, p = .023, d = 1.50. Lastly, results showed that when describing a frustrating experience, students tended to use significantly more words associated with anger, t (4.409) = 2.623, p = .05, d = 1.49. See Figure 1.

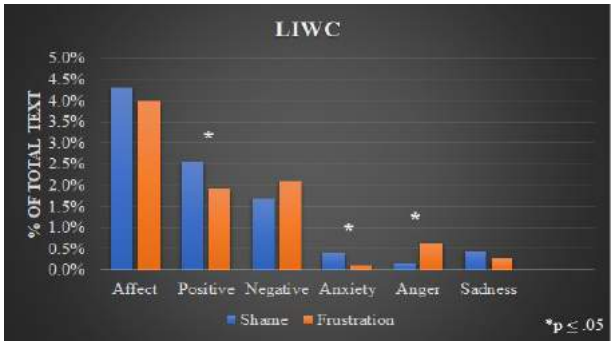


Figure 1: Linguistic Inquiry and Word Count (LIWC) results.

The results from the Coh-Metrix revealed that, when discussing an experience of shame, students tended to use significantly simpler syntax, t (11) = 6.616, p = .000, d = 3.77. They also used more abstract words, t (4.326) = -2.909, p = .04, d = 1.66, and had less referential cohesion, t (3.062) = .01, d = 1.75. See Figure 2.

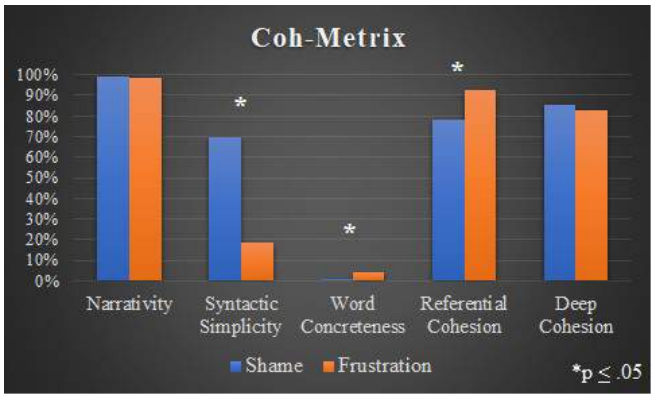


Figure 2: Coh-Metrix results.

Scholarly Significance

The results from the current study revealed that shame does in fact have a unique linguistic profile when compared to frustration. Surprisingly, learners who described an experience of shame tended to use significantly more positive emotional words (e.g., love, nice, sweet), along with more words associated with anxiety (e.g., worried, fearful). Additionally, students who described a shameful experience used significantly fewer words associated with anger compared to learners describing frustration. The Coh-Metrix results revealed that, when discussing a moment of shame, learners tended to use significantly simpler syntax, more abstract words, and demonstrated less referential cohesion.

Our results supported the notion that, when individuals talk about shame shame-experiences, the use of language is difficult. Students spoke abstractly about their shame experiences, while they were more able to articulate their frustration experiences. Shame-experiencing students also used less linguistic complexity and their narrative had less cohesion than students describing frustration. A teacher could learn to pick up on these linguistic elements and use this information to help students bounce back from the debilitating effects of experiencing academic shame.

Imagine a student who, after having failed an exam is staying after class to talk to the instructor about his/her performance. What if it could be determined, based on speech alone, whether these individuals are experiencing shame? What if a teacher was able to figure out which subset of students were actually experiencing shame and were able to be *proactive* to the potential negative consequences? Mitigating shame-consequences by understanding linguistic components of the *what*- and *how*-indicators of shame experiences, could facilitate teachers' ability to provide motivational interventions. Recognizing linguistic components of shame may be especially important given that individuals may deny their feelings, and may be unwilling or unable to express themselves, particularly if they self-isolate. In other words, as of now, we have no reliable way (other than perhaps self-report measures) to determine who is experiencing shame. Thus, intervention is near impossible without perceiving reliable indicators.

We do note that this study is limited in making comparative claims between the two sets of interview transcript-data. While the two sets of transcripts used in this analysis did, indeed, focus on different constructs (i.e., frustration and shame), they also differed according to other characteristics, such as the overall study-design, the methodology driving the investigations, and the institutions in which the data collection occurred. Thus, we make our claims with sensitivity to the multiple ways in which the two interview datasets can be compared. Yet, even with these limitations considered, we maintain that the linguistic profile that accompanies students' experiences of discussing shame provides compelling implications for educators.

References

Huff, J. L., & Clements, H. R. (2018). The hidden person within the frustrated student: An interpretative phenomenological analysis of a student's experience in a programming course. *Proceedings of the 2018 American Society for Engineering Education Conference*, Columbus, OH.

Groom, C., & Pennebaker, J. (2002). Brief report: Words. *Journal of Research in Personality, 36,* 615–621.

Pennebaker J.W., & Graybeal A. 2001. Patterns of natural language use: disclosure, personality, and social integration. *Curr. Dir. Psychol. Sci. 10*, 90–93

Turner, J. E., Husman, J., & Schallert, D. L. (2002). The importance of students' goals in their emotional experience of academic failure: Investigating the precursors and consequences of shame. *Educational Psychologist, 37*, 79 – 89.

Turner, J. E. (2014). Researching state shame with the experiential shame scale. *The Journal of Psychology, 148(*5), 577–601.

Pekrun, R., Frenzel, A. C., Goetz, T., & Perry, R. P. (2007). The control-value theory of achievement emotions: An integrative approach to emotions in education. In P. A. Schutz & R. Pekrun (Eds.), *Emotions in education.* San Diego: Academic Press.

McNamara, D.S., Graesser, A.C., Cai, Z., & Kulikowich, J.M. (2011, April). *Coh-Metrix easability components: Aligning text difficulty with theories of text comprehension.* Paper presented at the annual meeting of the AERA, New Orleans, LA.

Graesser, A. C., McNamara, D. S., Cai, Z., Conley, M., Li, H., & Pennebaker, J. (2014). Coh-Metrix measures text characteristics at multiple levels of language and discourse. *Elementary School Journal, 115* (2).

Event Perception Differs Across Cultures

Khena Swallow
Cornell University, Ithaca, New York, United States

Qi Wang
Cornell University, Ithaca, New York, United States

Abstract

Event segmentation divides continuous experience into meaningful events and guides attention, memory, and learning. Culture could impact event segmentation by emphasizing the importance of different aspects of experiences (attentional focus), and by providing different exemplars of everyday activities (familiarity). In this study, Indian and US viewers identified large (coarse) and small (fine) events in videos of everyday activities recorded in Indian and US settings. Analyses revealed that US viewers segmented the activities at a higher rate than Indian viewers. In addition, while the boundaries identified by US viewers were more strongly associated with visual change, boundaries identified by Indian viewers were more strongly associated with changes in actions and goals. However, there was no evidence that familiarity with an activity, as indicated by the match between a viewers culture and the activity setting, impacted segmentation. Culture appears to affect how people define events during perception, independent of familiarity.

A re-examination of the interrelationships between attention, eye behavior, and creative thought

Shadab Tabatabaeian
University of California Merced, Merced, California, United States

Colin Holbrook
University of California, Merced, Merced, California, United States

Carolyn Jennings
University of California, Merced, Merced, California, United States

Abstract

Internally focused attention, characterized by reduced sensory input, is often correlated with memory retrieval and the ability to combine memories to generate new ideas. Accordingly, the attenuation of external distractors (e.g., via reduced visual input) may be expected to enhance idea generation. We conducted a study requiring participants to perform an alternative uses task, in either a well-lit or totally dark environment. We also measured eye movements, as they have been linked with idea generation and attention. Departing from prior studies, our participants were not presented with visual stimuli, but received auditory task instructions. Preliminary analyses replicated the eye behavior attributed to internal attention in previous research, including more and shorter fixations and greater saccade amplitude in the dark. While these results suggest a positive relationship between darkness and internal attention, task performance was not significantly influenced by darkness manipulation. The findings and suggestions for future studies will be discussed.

Incorporating Semantic Constraints into Algorithms for Unsupervised Learning of Morphology

Abi Tenenbaum
Commonwealth School, Boston, Massachusetts, United States

Roger Levy
Massachusetts Institute of Technology, Cambridge, Massachusetts, United States

Abstract

A key challenge in language acquisition is learning morphological transforms relating word roots to derived forms. Unsupervised learning algorithms can perform morphological segmentation by finding patterns in word strings (e.g. Goldsmith, 2001), but struggle to distinguish valid segmentations from spurious ones because they look only at sequences of characters (or phonemes) and ignore meaning. For example, a system that correctly discovers ¡add -s¿ as a valid suffix from seeing dog, dogs, cat, cats, etc, might incorrectly infer that ¡add -et¿ is also a valid suffix from seeing bull, bullet, mall, mallet, etc. We propose that learners could avoid these errors with a simple semantic assumption: morphological transforms should approximately preserve meaning. We extend an algorithm from Chan (2008) by integrating proximity in vector-space word embeddings as a criterion for valid transforms. On the Brown CHILDES corpus, we achieve both higher accuracy and broader coverage than the purely syntactic approach.

Individual Differences in Second Language Age of Acquisition and Language Entropy Predict Non-Verbal Reinforcement Learning Among Bilingual Adults

Mehrgol Tiv
McGill University, Montreal, Quebec, Canada

Jason Gullifer
McGill University, Montreal, Quebec, Canada

A. Ross Otto
McGill University, Montral, Quebec, Canada

Debra Titone
McGill University, Montreal, Quebec, Canada

Abstract

We investigated whether bilingualism affects non-verbal model-free vs. model-based reinforcement learning (RL). This dual-systems theory posits independent valuation systems in controlling choices and may overlap with systems of bilingual executive control. Forty-five bilingual adults completed a two-stage decision making task with transition and probability of reward dynamically varying. First, we calculated a model-based index to measure how much participants integrate environmental structure with reward when planning choices. Consistent with monolingual results, we found that bilinguals display model-free and model-based RL to differing degrees. Next, we assessed whether individual differences in second language (L2) age of acquisition (AoA) and language entropy interact with these RL systems. Bilinguals with earlier L2 AoA and greater language entropy demonstrated model-free RL, whereas bilinguals with later L2 AoA and lower language entropy demonstrated greater sensitivity to model-based reward frequencies. This suggests an interesting link between bilingual experience and how reward shapes decision-making strategies.

Emergent Compositionality in Signaling Games

Nicholas Tomlin
Brown University, Providence, Rhode Island, United States

Ellie Pavlick
Brown University, Providence, Rhode Island, United States

Abstract

Understanding the origins of linguistic compositionality is a fundamental challenge in evolutionary linguistics. Prior work has explored this topic through dynamical computational modeling and experiments in iterated learning. We explore these questions using RL agents tasked with developing cooperative communication strategies in a signaling game. We analyze how various mechanisms (such as Bayesian pragmatic reasoning) and constraints (such as limited memory) may affect compositionality and generalizability in the invented communication protocols. In particular, our preliminary results suggest that incremental pragmatic reasoning induces a bias towards lexical compositionality. To evaluate the extensibility of our model, we compare the behavior of the RL agents to the behavior of humans on the same task. That is, we ask humans to coordinate in a reference game task by repeatedly composing non-linguistic symbols. We discuss ways in which the resulting protocol mirrors and differs from that produced by the RL agents.

Agent-based modeling of how national identity affects party preferences in voting

Taiji Ueno
Takachiho University, Tokyo, Japan

Ryu Hakche
Neo Career, Co. Ltd., Tokyo, Japan

Nobuko Asai
Kyoto-Bunkyo University, Uji, Japan

Minoru Karasawa
Nagoya University, Nagoya, Aichi, Japan

Abstract

Attitudes concerning national identity (e.g., nationalism, patriotism) are known to correlate with various social behaviors such as party preferences in voting. For instance, survey data indicates that Japanese citizens who are proud of being Japanese (i.e., patriots) and those who are high in a right-wing tendency are more willing to vote for the conservative party (Liberal Democratic Party). In this study, we employed an agent-based modeling approach to understand how national identity affects individual voting intentions. The individual agents and the political party agents interacted with each other by spreading their political attitudes (e.g., VAT should be increased to maintain the pension insurance system), and the recipients of the messages changed their attitudes (e.g., persuasion). The simulation revealed that the effects of persuasive messages were moderated by the strength of its own national identity attitudes, and the resultant individual agents voting preferences simulated the human participants data more precisely.

Exploring the role of visuospatial processes in surgical skill acquisition: A longitudinal study

Tina Vajsbaher
University of Bremen, Bremen, Germany

Holger Schultheis
University of Bremen, Bremen, Germany

Verena Uslar
University of Oldenburg, Oldenburg, Germany

Dirk Weyhe
Pius-Hospital Oldenburg, Oldenburg, Germany

Hseyin Bektas
Klinikum Bremen-Mitte, Bremen, Germany

Nader Francis
Yeovil District Hospital, Yeovil, United Kingdom

Abstract

Surgical error is the most frequent and costly type of medical error, posing a direct threat to patient safety. Surgical errors have been described as a 'cognitive phenomenon', as it is largely the shortcomings of the surgeons cognitive processing that leads to error. In laparoscopic surgery, visuospatial processes are known to be crucial for skill acquisition, although it remains unclear as to which exact processes are important, how these develop over time and intraoperatively, and how they influence competency development. We will report interim spatial cognitive baseline results of 35 surgeons, 17 residents and 18 specialists, taking part in an on-going longitudinal study at two major hospitals in Germany. Our results offer new insight into the role of visuospatial cognition in domain-specific expertise, and shed new light on the malleability of visuospatial processes in the skill acquisition process.

RunTheLine: An infinite runner serious game to train comprehension of societally relevant large numbers

Thijs van Den Hout
Radboud University Nijmegen: Donders Institute for Brain, Cognition and Behavior, Nijmegen, Netherlands

Hanna Schraffenberger
Radboud University, Nijmegen, Netherlands

Florian Krauze
Radboud University Nijmegen: Donders Institute for Brain, Cognition and Behavior, Nijmegen, Netherlands

Tibor Bosse
Radboud University, NIjmegen, Netherlands

Frank Leone
Radboud University: Donders Institute for Brain, Cognition and Behavior, Nijmegen, Netherlands

Abstract

Large numbers play a significant role in personal and political financial choices and the understanding of exponential growth. Large numbers are also often misjudged, showing a logarithmic number understanding. Small numbers are however represented in a linear fashion, due to direct experience on for example number lines. Earlier, it was shown that large number comprehension can be trained, influencing societally relevant choices. We trained large number comprehension using a serious game (RunTheLine): an infinite runner game where an avatar runs on a number line ranging till one billion. Due to the game mechanics, the players walk the number line at both small and large numbers in small steps, making them aware of the continuity of the number line. Pre-post test differences show a change in economic judgments compared to a control group. This offers a scientific manipulation of behavioral and cortical number line representations and potential educational applications.

Automatic Model Generation with Symbolic Deep Learning

Vladislav Veksler
DCS Corp, U.S. Army Research Laboratory, Aberdeen, Maryland, United States

Norbou Buchler
U.S. Army Research Laboratory, Aberdeen Proving Ground, Maryland, United States

Abstract

Automatic model generation based on user-task interactions is of great use for behavioral predictions and understanding of cognition. Mapping which environment features cause which actions seems like a classification problem suited for Deep Learning (DL). Unfortunately, DL does not create an observable model, and is more suitable to making predictions from billions of examples than from limited observations. There are, however, many tasks that lend themselves to symbolic input, allowing an alternative approach - Symolic Deep Learning (SDL). Symbolic hierarchical representations have a long history in Psychological literature, though some of these were integraged as models of memory without action-selection (e.g. EPAM/CHREST), and some have run into computational limitations (e.g. configural-cue). SDL stands to benefit from better model integration and modern growth in computational power and algorithmic efficiency, and promises to be the right paradigm for automatic model generation from limited user observations.

The Importance of Explanations in Guided Science Activities

Vaunam Venkadasalam
University of Toronto, Toronto, Ontario, Canada

Nicole Larsen
University of Toronto, Toronto, Ontario, Canada

Patricia Ganea
University of Toronto, Toronto, Ontario, Canada

Abstract

This study examined whether embedding explanations in guided activities promotes conceptual change about a physical science concept. One common misconception that children have is that heavy objects fall at a faster rate than light ones. We used a pre-, post-, and delay test design to address this misconception. Forty 5-year-old children were assigned to one of two conditions: a guided play activity that included an explanation about gravity, or the same guided play activity but with no explanation provided. Childrens explanations improved immediately at post-test (p =.001, 95% CI [0.58, 2.33]) and after a one-week delay test (p ¡.001, 95% CI [1.23, 2.95]) when the explanation about gravity was embedded in the activity. There was no improvement at post-test (p =.36) or delay-test (p =.93) when children completed the activity only. This study shows that conceptually rich explanations are an effective pedagogical tool for promoting belief revision in children.

Exploring the linguistic landscape: How individual differences among bilingual adults modulate eye movements when viewing multilingual artificial signs

Naomi Vingron
McGill University, Montral, Quebec, Canada

Jason Gullifer
McGill University, Montreal, Quebec, Canada

Debra Titone
McGill University, Montreal, Quebec, Canada

Abstract

Eye movement research reveals how people allocate visual attention when reading, scanning the environment around them (Rayner, 2012). These cognitive processes come together when people view what sociolinguists refer to as, the linguistic landscape, consisting of signage in the public space. Linguistic landscapes around the world are jointly determined by top-down socio-legal provisions, and bottom-up capacities and attitudes of individual people (Leimgruber, Vingron, & Titone, 2019). In a preliminary study, we found that bilinguals differed in how they viewed naturally occurring linguistic landscape images (Vingron et al., 2018). We are currently analyzing data from a follow-up study that investigated whether individual differences in language experience among bilinguals modulate their eye movements to artificial linguistic landscape images that systematically manipulate text language, position, and size, while controlling for linguistic content.

Integrating stereotypes and individuating information based on informativeness under cognitive load

Thalia Vrantsidis
University of Toronto, Toronto, Ontario, Canada

William Cunningham
University of Toronto, Toronto, Ontario, Canada

Abstract

When making inferences about another person (the target), perceivers often have to integrate multiple sources of information. This can include stereotypes about the targets groups (e.g., age, race, occupation) as well as other information about the target (individuating information). In simple situations, perceivers approximate ideal Bayesian information integration, relying more heavily on information that is more informative for the judgement. However under cognitive load with cognitive resources taken up by other demands people may instead rely on simplifying heuristics. We investigate several possible heuristics that people may use under load, including relying primarily on stereotypes rather than individuating information, as suggested by previous research, and we test if and how these heuristics depend on how informative each source of information is. By clarifying how stereotypes are used in less-than-ideal cognitive conditions, this work has implications for when stereotypes will tend to be overused in real-world situations.

Children with immature intuitive theories seek domain-relevant information

Jinjing (Jenny) Wang
Rutgers University-Newark, Newark, New Jersey, United States

Yang Yang
Rutgers University-Newark, Newark, New Jersey, United States

Carla Macias
Rutgers University-Newark, Newark, New Jersey, United States

Elizabeth Bonawitz
Rutgers University - Newark, Newark, New Jersey, United States

Abstract

A growing body of research suggests that infants and children are sensitive to signals of information gain. However, the value of a piece of information may also change as the learner knows more. How do changes that occur naturally in childrens intuitive theories contribute to their subsequent learning? Here we tested whether children who are at different stages of understanding an intuitive theory also differ in their interest in acquiring more information in the same domain. We tested childrens performance in three distinct domains, including intuitive biology, psychology, and beliefs about psychosomatic events. We found that children at earlier stages of their intuitive theories were more likely to seek information in the related domain than children with mature knowledge. These results are the first to show the relationship between natural changes in childrens existing knowledge and childrens future learning preferences.

Visual Statistical Learning Contributes to Word Segmentation during Reading of Unspaced Chinese Sentences

Tsanyu Wang
National Taiwan Normal University, Taipei, Taiwan

Jenn-Yeu Chen
National Taiwan Normal University, Taipei, Taiwan

Abstract

We investigated whether Chinese readers learn to segment words automatically while reading unspaced sentences through statistical learning. Experiment 1 replicated Saffran et al.s (1997) study using Chinese monosyllables presented auditorily to foreign learners of Chinese. The learning outcome was .57 on a two-alternative forced-choice test, statistically better than guessing (.5). Experiment 2 repeated Experiment 1 but presented the Chinese monosyllable string visually as a character string. Experiment 3 repeated Experiment 2 but doubled the exposure. Experiment 4 repeated Experiment 2 with characters of fewer numbers of strokes. The learning outcomes were .53, .52, and .52., not significant when tested individually, but was significant when the data were combined. At least 60% of the participants in each experiment showed the effect. We conclude that visual statistical learning does contribute to automatic word segmentation in Chinese reading.

A tradeoff between generalization and perceptual capacity in recurrent neural networks

Taylor Webb
Princeton University, Princeton, New Jersey, United States

Steven Frankland
Princeton University, Princeton, New Jersey, United States

Simon Segert
Princeton University, Princeton, New Jersey, United States

Alexander Petrov
Ohio State University, Columbus, Ohio, United States

Randall O'Reilly
University of Colorado Boulder, Boulder, Colorado, United States

Jonathan Cohen
Princeton University, Princeton, New Jersey, United States

Abstract

In a classic paper, Miller (1956) summarized findings showing that people can only identify a limited number of distinct stimuli at a time. One puzzling aspect of this capacity limitation is that it is approximately invariant to range. That is, the number of accurately identifiable stimuli is approximately the same regardless of how far apart the stimuli are spaced. Models of this phenomenon have suggested that people operate in a context-coding mode when performing these tasks, effectively carrying out a form of contextual normalization, but why such normalization might take place is unclear. Here, we propose an explanation by appealing to a tradeoff with generalization. Specifically, we implement contextual normalization in a recurrent neural network and show that this normalization enables stronger generalization in a relational reasoning task, but also results in a perceptual capacity limitation which captures many of these classic phenomena.

Wriggly, Squiffy, Lummox, and Boobs: What Makes Some Words Funny?

Chris Westbury
University of Alberta, Edmonton, Alberta, Canada

Geoff Hollis
University of Alberta, Edmonton, Alberta, Canada

Abstract

Theories of humor suffer from insufficient operationalization. We build on the Engelthaler & Hills (2017) humor rating norms, by analyzing the semantic and word form factors that play a role in the judgments. Our model can predict the original humor rating norms and ratings for previously unrated words with greater reliability than the split half reliability in the original norms. The model is consistent with several theories of humor, while suggesting that those theories are too narrow. In particular, it is consistent with incongruity theory, which suggests that experienced humor is proportional to the degree to which expectations are violated. Words are judged funnier if they are less common and have an improbable orthographic or phonological structure. We also describe and quantify the semantic attributes of funny words that are judged funny and show that they are partly compatible with the superiority theory of humor, which focuses on humor as scorn.

Language in Math Problem Solving

Renee Whittaker
Carleton University , Ottawa, Ontario, Canada

Chang Xu
Carleton University, Ottawa, Ontario, Canada

Jo-Anne LeFevre
Carleton University, Ottawa, Ontario, Canada

Helena P. Osana
Concordia University, Montreal, Quebec, Canada

Jill Turner
Carleton University, Ottawa, Ontario, Canada

Heather Douglas
Carleton University, Ottawa, Ontario, Canada

Anne Lafay
Concordia University, Montreal, Quebec, Canada

Sheri-Lynn Skwarchuk
University of Winnipeg, Winnipeg, Manitoba, Canada

Abstract

Children enrolled in language-immersion programmes may be required to learn math in the immersion language. Following the framework of the Pathways Model (LeFevre et al., 2010; Sowinski et al., 2014), the goal of the present study was to understand how instructional language supports math learning by comparing patterns of performance of immersion and non-immersion students. Participants included 182 grade 2 students (Mean age= 7.8 years): 108 students were enrolled in French immersion programs and were learning math in French (their second language) and 74 were enrolled in non-immersion programs and were learning math in English (their home language). Participants were tested on a number of general cognitive measures as well as math specific outcome measures. Results show that overall, across both immersion and non-immersion students, linguistic, quantitative and working memory components contributed to math problem solving. However, within the linguistic component there were differences between the direct and indirect pathways.

Using low-level sensory mechanism to bootstrap high order thinking in EFL reading

HingYi Wong
The Education University of Hong Kong, Hong Kong, China

Duo Liu
The Education University of Hong Kong, Hong Kong, Hong Kong

Zi Yan
The Education University of Hong Kong, Hong Kong, China

Abstract

The goal of the study was to compare potential changes in architecture when different set sizes were manipulated as a function of age difference and reading group difference in the Visual Search Task in Coglab. Based on the RT performance of Chinese EFLs aged 11 15 years old in feature and conjunction search when target was absent/present across three different set sizes (display size 4, 16 & 64), we conducted tests for architecture, stopping rule and dependency in visual search between typical and poor readers. What we are interested in are as follows: First, how a parallel/serial mental architecture in visual search might be predicted by both item features and person characteristics; and second, if stopping rule in target absent search is self-terminating/ exhaustive in nature. The aim of the study was to find cognitive behaviour that would accommodate developmental deficiency in EFL reading.

Semantic structure of infant first-person scenes changes with development

Ziyu Xiang
Indiana University, Bloomington, Indiana, United States

Linda Smith
Indiana University, Bloomington, Indiana, United States

David Crandall
Indiana University, Bloomington, Indiana, United States

Abstract

The co-occurrence of objects in visual scenes reflects the semantic structure of the world: cups are more likely to appear in scenes with tables than airplanes, for example. Both human and machine vision use these co-occurrences to support recognition of individual objects. A reasonable assumption is that these co-occurrences are ubiquitous and present for all perceivers. However, the scenes observed by infants are highly dependent on their body postures and locations, both of which change dramatically over the first year of post-natal life. To measure these changing co-occurrences in infant-perspective scenes, we collected images from infants wearing head cameras in everyday home environments comparing three age groups: 1-3, 6-8 and 11-12 months. Using graph theoretical analysis, we conclude that the semantic structure of scenes in 6-8 months differs from whats in younger and older infants.

Abstract Syntactic Knowledge or Limited-Scope Formulae: A Computational Study of Childrens Early Utterances

Qihui Xu
Graduate Center, City University of New York, New York, New York, United States

Martin Chodorow
Graduate Center, City University of New York, New York, New York, United States

Virginia Valian
Graduate Center, City University of New York, New York, New York, United States

Xiaomeng Ma
Graduate Center, City University of New York, New York, New York, United States

Abstract

Do childrens early utterances reflect abstract syntactic knowledge or slot-filler formulae developed through word imitation? This study compares development of part-of-speech (POS) sequences with word sequences using language models (LMs) trained on mothers utterances (N=1,272,139) from CHILDES English corpora, in which POS tags are automatically assigned by MOR and POST programs (MacWhinney, 2000). Word-based and POS-based LM probabilities for childrens multi-word utterances in the Providence corpus (Brschinger et al., 2013, 15-36 months, Nchildren=6, Nutterances=50,717) were calculated as a function of age. Word-based LM probability of childrens multi-word utterances first increases with age and then levels off after 23 months. By contrast, POS-based probability remains high and stable across all ages. This suggests children have adult-like syntactic knowledge even at a very early age when their word sequences are still not adult-like. The pattern of results supports the abstract syntax view. Additional studies will use more accurate POS-taggers and larger datasets.

The effects of object motion observations on physical prediction

Moyuru Yamada
Fujitsu Laboratories Ltd., Kanagawa, Japan

Kevin Smith
Massachusetts Institute of Technology, Cambridge, Massachusetts, United States

Josh Tenenbaum
MIT, Cambridge, Massachusetts, United States

Abstract

People use knowledge about physical objects to predict and plan their actions, but this knowledge about objects can be directly perceived or simply inferred. In this experiment, participants chose the direction to shoot computerized cannons to hit targets. These cannons differed in how fast they shot the cannonball, but participants could learn this information either from observing the full trajectory of a prior shot, or just observing the outcome. While the cannonballs initial speed can be determined from the end state alone, additional information in the full trajectory might improve these estimates. We find that performance is only worse in the end-state trials if these trials were tried first; if participants judged the full trajectory trials first, their performance did not decline on the end-state trials. We explore this order effect using a model of noisy physical inference that assumes learning from prior trial blocks.

Commonality search shares processes with alternative categorization

Mayu Yamakawa
Nagoya University, Nagoya, Japan

Sachiko Kiyokawa
Nagoya University, Nagoya, Japan

Abstract

We investigated how people find commonalities between unrelated objects as a basis of generating creative ideas by examining the relationship between performances on commonality search and alternative categorization tasks. We predicted a positive correlation between performances on the two tasks because one needs to focus on some obscure features of objects to do both tasks well. Thirty-one undergraduates were asked to engage in both commonality search and alternative categorization tasks. They were asked to list as many as commonalities between nine unrelated object pairs for 90 seconds for each pair. They were then asked to list as many categories as possible that each of five objects belong to for 60 seconds per object. The results showed a significant positive correlation between the performances on these tasks. We concluded that commonality search and alternative categorization both focus on obscure features of objects.

Minimal but meaningful: Probing the limits of randomly assigned social identities

Xin Yang
Yale University, New Haven, Connecticut, United States

Yarrow Dunham
Yale University, New Haven, Connecticut, United States

Abstract

The present studies (total n = 151) experimentally manipulated meaningfulness in novel social groups and measured any resulting ingroup biases. Study 1 showed that even when groups were arbitrary and presumptively meaningless, 5- to 8-year-olds developed equally strong ingroup biases as did children in more meaningful groups. Study 2 explored the lengths required to effectively reduce ingroup biases by stressing the arbitrariness of the grouping dimension. Even in this case ingroup bias persisted in resource allocation behavior, though it was attenuated on preference and similarity measures. These results suggested that one has to go to great lengths to counteract childrens tendency to imbue newly encountered social groups with rich affiliative meaning.

Corpus-based topic modeling for the cognitive study of the 21st century sociocultural challenges

Vera Zabotkina
Russian State University for the Humanities, Moscow, Russian Federation

Boris M. Velichkovsky
Kurchatov Institute, Moscow, Russian Federation

Artemy Kotov
Kurchatov Institute, Moscow, Russian Federation

Dmitry Orlov
Russian State Univercity for the Humanities, Moscow, Russian Federation

Alexander Piperski
Russian State University for the Humanities, Moscow, Russian Federation

ELENA POZDNYAKOVA
Russian State Univercity for the Humanities, Moscow, Russian Federation

Abstract

The results were obtained in the course of a two-stage study. At the first stage (2018) linguists analyzed the conceptual domain sociocultural challenges on the basis of purposely elaborated Russian language THREAT-corpus (10.4 m words) and built a frame of the domain. At the second stage (2018-2019) the research was carried out with methods of automated topic modeling for two Russian language corpora: THREAT-corpus and alternative corpus collected using WebBootCaT tool in the SketchEngine corpus management system. Methods of topic modeling (PLSA, LDA, BigARTM et al.) allowed eliciting thematic profiles for texts of both corpora. Comparison of two datasets was carried out by applying set theory, graph theory, and probabilistic analysis. Combining topic modeling with linguistic frame analysis resulted in more precise configurations of cognitive models in the conceptual domain sociocultural challenges. Word frequency for lexemes manifesting sociocultural challenges proved to be an important factor of conceptual structures representation.

Communicative need and color naming

Noga Zaslavsky
The Hebrew University, Jerusalem, Israel

Charles Kemp
University of Melbourne, Melbourne, VIC, Australia

Naftali Tishby
Hebrew University of Jerusalem, Jerusalem, Israel

Terry Regier
UC Berkeley, Berkeley, California, United States

Abstract

Color naming across languages has traditionally been held to reflect the structure of color perception. At the same time, it has often, and increasingly, been suggested that color naming may be shaped by patterns of communicative need. However, much remains unknown about the factors that drive communicative need, how need interacts with perception, and how this interaction may shape color naming systems across languages. We engage these open questions by building on general information-theoretic principles, and on a recent account of color naming that integrates the roles of need and perception. On this basis, we present a systematic evaluation of several factors that may influence need, and that have been proposed in the literature: capacity constraints, linguistic usage, and the visual environment. Our findings suggest that communicative need and resulting patterns of color naming are shaped more by linguistic usage than they are by the visual environment alone.

Constructing a category prototype from statistical regularities under uncertainty

Haiyun Zeng
University of Pennsylvania, Philadelphia, Pennsylvania, United States

John Trueswell
University of Pennsylvania, Philadelphia, Pennsylvania, United States

Sharon Thompson-Schill
University of Pennsylvania, Philadelphia, Pennsylvania, United States

Abstract

Learning the meaning of a word requires forming a semantic representation that characterizes the referential exemplars encountered with that word. However, each learning instance is ambiguous in that the word may plausibly refer to multiple entities. To the extent that learners consider multiple referents under conditions of referential uncertainty, how do these alternatives enter into learning word meaning? We employed a cross-situational word-learning paradigm with novel creatures to investigate whether co-occurring exemplars that were considered but not selected as the words referent would influence the category prototype. We contrasted a condition where all exemplars were labeled with a word and a condition where only some of the exemplars of a category were labeled with the word later in the learning phase. Preliminary results are consistent with the prediction that referents that are considered but not selected contribute less to the semantic representation of the word than do the selected referents.

Interpretation of Generic Language is Dependent on Listener's Background Knowledge

Xiuyuan Zhang
University of Chicago, Chicago, Illinois, United States

Dan Yurovsky
University of Chicago, Chicago, Illinois, United States

Abstract

Generic statements, like "birds lay eggs" or "dogs bark" are simple and ubiquitous in naturally produced speech. However, the inherent vagueness of generics makes their interpretation highly context-dependent. Building on work by Tessler & Goodman (in press) showing that generics can be thought of as inherently relative (i.e. more birds lay eggs than you would expect), we explore the consequences of different implied comparison categories on the interpretation of novel generics. In Experiments 1 and 2, we manipulated the set of categories salient to a listener by directly providing them the comparison sets. In Experiments 3 and 4, we collected participants demographic information and used these naturally occurring differences as a basis for differences in the participants' comparison sets. Our results confirmed the hypothesis that prevalence judgments of features in novel categories are sensitive to differences in their corresponding comparison categories. These results suggest a possible source for well-intentioned miscommunications.

Deep Learning of Chinese Characters

Xiaowei Zhao
Emmanuel College, Boston, Massachusetts, United States

Abstract

In this study, the printing forms (different fonts) of about 3000 common Chinese characters were sent into a Deep Neural Network (DNN), along with their sounds. The network can successfully learn the association between the form and the sound of these characters. It also develops certain generalizability when facing new characters. In addition, the internal representations on different layers of the network show the emergence of basic writing structures of Chinese characters (i.e. strokes, radicals, left-right, top-down structures). The learning pattern of the network is further compared with that of the elementary school students.

Bayesian Inference Causes Incoherence in Human Probability Judgments

Jianqiao Zhu
University of Warwick, Coventry, United Kingdom

Adam Sanborn
University of Warwick, Coventry, United Kingdom

Nicholas Chater
University of Warwick, Coventry, United Kingdom

Abstract

Human probability judgements appear systematically biased, in apparent tension with Bayesian models of cognition. But perhaps the brain does not represent probabilities explicitly, but approximates probabilistic calculations through a process of sampling, as used in computational probabilistic models in statistics. The Bayesian sampling viewpoint provides a simple rational model of probability judgements, which generates biases such as conservatism. The Bayesian sampler provides a single framework for explaining phenomena associated with diverse biases and heuristics, including availability and representativeness. The approach turns out to provide a rational reinterpretation of noise in an important recent model of probability judgement, the probability theory plus noise model (Costello & Watts, 2014; 2016; 2017; Costello, Watts, & Fisher, 2018), and captures the empirical data supporting this model.

A resource-rational model of physical abstraction for efficient mental simulation

Tina Zhu
DeepMind, London, United Kingdom

Jessica Hamrick
DeepMind, London, United Kingdom

Kevin McKee
DeepMind, London, United Kingdom

Raphael Koster
DeepMind, London, United Kingdom

Jan Balaguer
DeepMind, London, United Kingdom

Peter Battaglia
DeepMind, London, United Kingdom

Matthew Botvinick
DeepMind, London, United Kingdom

Abstract

Physical simulation enables people to make intuitive predictions about physical scenes and interact flexibly with the objects around them, from a stack of books balanced on a ledge to the turrets and moats of a sandcastle. We hypothesize that when the number of possible objects makes simulation intractable, people use chunked abstractions that reduce the number of objects they need to simulate while also minimizing simulation error. We tracked participants gaze while they viewed complex towers of blocks and predicted whether the towers would remain stable under gravity. We developed a resource-rational model of how people might optimally partition towers into chunks of blocks. Subsequently, we compared this model to participants fixations over the scene. We explore how efficient, resource-rational chunkings of physical scenes might underlie peoples ability to make rapid and robust inferences in this domain.

Modeling of Complex Communicative Behavior for F-2 Companion Robot

Anna Zinina
National Research Center Kurchatov Institute, Moscow, Russian Federation

Nikita Arinkin
National Research Center Kurchatov Institute, Moscow, Russian Federation

Liudmila Zaidelman
National Research Center Kurchatov Institute, Moscow, Russian Federation

Artemy Kotov
Kurchatov Institute, Moscow, Russian Federation

Abstract

We design F-2 companion robot, supporting natural multimodal communication. The robot is operated by a set of scripts, triggered by input speech and generating behavioral patterns in BML format. To make robots behavior as close as possible to humans, we extract natural communication patterns from the Russian Emotional Corpus REC (over 400.000 annotations), reproduce key patterns in Blender 3D editor and export them to MySQL database (n = 220). For each generated BML the software retrieves the corresponding movement from the database, joins compatible patterns and performs them on the robot. Robot can also receive the coordinates of surrounding human faces and simulate direct gazes towards the eyes of the addressee. It can also perform oriented (pointing) gestures: switch between directions or between several interlocutors. This allows us to model complex robot behavior, as shown in our experiment, increasing human satisfaction from robot-to-human interaction (Research is supported by the Russian Science Foundation, project No 19-18-00547).

A Visual Remote Associates Test and its Initial Validation

Faheem Zunjani
Freie Universitt Berlin, Berlin, Germany

Ana-Maria Olteteanu
Freie Universitat Berlin, Berlin, Germany

Abstract

The Remote Associates Test (RAT) is a test used for measuring creativity as relying on the power of making associations, and it normally takes a linguistic form (i.e., given three words, a fourth word associated with all three is asked for). While other visual creativity tests exist, no creativity test to date can be given in both a visual and linguistic form. Such a test would allow the study of differences between various modalities, in the creativity domain. In this paper, a visual version of the well known Remote Associates Test is constructed. This visual RAT is validated in relation to its linguistic counterpart in a study with 42 participants. A significant correlation of 0.431 (p ¡ 0.01) between visual RAT scores and comRAT-G scores was observed.

Author Index

List of Reviewers

Aakre, Anthony
ABDELFATTAH, Ahmed M. H.
Abney, Drew H
Abrahamson, Dor
Acarturk, Cengiz
Admoni, Henny
Aguado-Orea, Javier
Aguilar, Wendy
Ahmad, Sheeraz
Ahmed, Faez

Ainooson, James
AKNINE, Samir
Alacam, Ozge
Aldosari, Bushra
Alhama, Raquel G.
Alikaniotis, Dimitrios
Alishahi, Afra
Allbritton, David
Allen, Michael G
Alnajjar, Khalid
Altmann, Erik
Alves, Ana Oliveira
Alviar, Camila
Ambridge, Ben
Anchan, Mona
Andonova, Elena
Andre, Elisabeth
Ansuini, Caterina
Aparicio, Helena
Arfini, Selene
Arjmandi, Meisam K.
Armstrong, Blair
Arner, Tracy
Arnold, Jennifer
Arnon, Inbal
ARSLAN, BURCU
Asaba, Mika
Atagi, Natsuki
Atari, Mohammad
Atit, Kinnari
Atkinson, Emily
Attar, Nada
Aulet, Lauren
Austerweil, Joseph Larry
Awad, Edmond
Baart, Martijn
Backer, Kristina
Bae, Gi-Yeul
Bakdash, Jonathan
Balasubramani, PragathiPriyadharsini
Ball, Tonio
Baltaretu, Adriana
Bandara, H.M. Ravindu T.
Banks, Adrian
Banks, Briony
Bannerjee, Bonny
Baranski, Michael
Barbu, Roxana-Maria
Barley, Mike
Bartels, Daniel
Bartlett, Madeleine
Bascandziev, Igor
Battaglia, Peter
Bauer, Malcolm
Bauters, Merja
Beaty, Roger
Becker, Maxi
Beekhuizen, Barend
Beierholm, Ulrik
beim Graben, Peter
Belardinelli, Anna
Bello, Paul
Belpaeme, Tony
Bender, Andrea
Benedek, Mathias
Benitez, Viridiana L
Bennett, Erin
Bent, Tessa
Benuzzi, Francesca
Bergelson, Elika
Bergen, Leon
Bertel, Sven
Bertero, Dario
Bertolotti, Tommaso W
Besold, Tarek R.
Bhat, Ajaz Ahmad
Bhatia, Sudeep
Bianchini, Francesco
Bilalic, Merim
Biro, Tamas

Bittner, Jennifer
Bixter, Michael
Blass, Joe
Blomsma, Pieter A.
Blumenthal, Anna
Boduroglu, Aysecan
Boghrati, Reihane
Boland, Julie E
Boncoddo, Rebecca
Boone, Alexander
Borghi, Anna
Bortfeld, Heather
Bosch, David A.
Boswijk, Vincent
Botezatu Ph.D., Mona Roxana
Botvinick, Matthew
Boucart, Muriel
Bourgonje, Peter
Boyer, Ty W.
Brainerd, Charles
Bramley, Neil R
Brand, Daniel
Brand, James
Brandone, Amanda C
Bratitsis, Tharrenos
Breaux, Brooke O.
Bredeweg, Bert
Brennan, Jonathan R.
Bridgers, Sophie
Briggs, Gordon

Brisson, Janie
Brooks, Connor
Brooks, Thomas R
Brouwer, Harm
Brown, Gordon
Brunstein, Angela
Brunye, Tad T
Bryan-Kinns, Nicholas
Bryant, David J.
Buchsbaum, Daphna
Buechner, Simon J.
Bunce, John P
Bunger, Ann
Burling, Joseph
Burns, Bruce
Burns, Devin M.
Buschmeier, Hendrik
Busemeyer, Jerome
Bushong, Wednesday
Buss, Aaron
Butler, Lucas
Butz, Martin V.
Byers, Patrick
Cakir, Murat Perit
Callaway, Charles
Callaway, Frederick
Canale, Rebecca
Capobianco, Antonio
Carcassi, Fausto
Cardoso, Amilcar
Carlisle, Nancy
Carlson, Richard
Carmeci, Floriana
Caro, Manuel Fernando
Caro, Marta
Carr, Jon W
Carruthers, Peter
Carstensen, Alexandra
Carvalho, Paulo
Casillas, Marisa
Caterina, Gianluca
Catrambone, Richard
Caulfield, Meghan
Cayton-Hodges, Gabrielle
Ceja, Cristina R
Cevolani, Gustavo
Chan, Jenny Yun-Chen
Chan, Margaret
Chandler, Jesse
Chandrasekharan, Sanjay
Chang, Maria D
Changizi, Mark
Charfi, Selem
Chaudhary, Aashish
Chemla, Emmanuel
Chen, Lang
Chen, Peiyao
Cheng, Jiuqing
Cheng, Peter
Chesney, Dana
Chetail, Fabienne
Cheung, Pierina
Cheyette, Samuel J.
Chiang, Cindy
Childers, Jane B.
Chin-Parker, Seth
Chin, Jessie
Chodroff, Eleanor
Choi, Soonja
Chrysikou, Evangelia G.
Chu, Mingyuan
Cifuentes-Férez, Paula
Cimiano, Philipp
Clapper, John
Clausner, Tim
Clegg, Jennifer M
Cleland, Alexandra
Clifton Jr., Charles
Cochrane, Aaron
Coco, Moreno
Coco, Moreno I
Coello, Yann
Cohen, Andrew
Collazos, César
Collins, Michael Gordon
Colombo, Matteo
Colunga, Eliana
Committee, Test
Connor Desai, Saoirse
Cooper Borkenhagen, Matthew
Cooperrider, Kensy
Corps, Ruth Elizabeth
Corriveau, Kathleen
Corter, James E.
Costello, Fintan
Cottrell, Garrison W
Couch, Nathan
Courtland, Maury
Coutanche, Marc N
Coutrot, Antoine
Cox, Christopher R
Cox, Gregory Edward
Craig, Scotty
Crick, Christopher
Croijmans, Ilja
Crupi, Vincenzo
Culbertson, Jennifer
Cullen, Clare
Cummins, Chris
Cunnings, Ian
Cuskley, Christine
Dale, Rick
Dam, Gregory
Damen, Debby
Dames, Hannah
Danks, David
Dasgupta, Ishita
Dautriche, Isabelle
Davies, Jim
Davis, Isaac
Davis, Nicholas
Dayton, Andrew
De Deyne, Simon
De Freitas, Julian
de Kleer, Johan
de Leeuw, Josh
De Palma, Paul
de Sa, Virginia
de Visser, Ewart

Deak, Gedeon
Dean, Roger
DeBrigard, Felipe
Declercq, Christelle
Degen, Judith
Dehghani, Morteza
DeJesus, Jasmine
deKamps, Marc
DeLiema, David
Dellantonio, Sara
Deng, Sophia W
Denison, Stephanie
Dennis, Tam
Dessalles, Jean-Louis
Destruel-Johnson, Emilie
Dey, Sanorita
Di Caro, Luigi
Diard, Julien
Díaz Agudo, M Belén
Dilley, Laura
Dillon, Brian
Dingemanse Dingemanse, Mark
Distefano, Lizanne
Do, Tuan
Doebel, Sabine
Dorfman, Hayley
Dow, Steven
Dragovich, Colleen
Dubé, Adam Kenneth
Dubey, Rachit
Duff, John
Dufour, Sophie

Dunn, Kirsty
Duran, Nicholas D.
Dye, Melody
Dymarska, Agata
Dziura, Sarah
Eagleman, David
Echols, Catharine H.
Eckstein, Maria K
Edelsbrunner, Peter
Edmiston, Pierce
Edmunds, Charlotte
Endres, Dominik M
Engelmann, Neele
Epstein, Susan L.
Eranki, Kiran L.N
Erdener, Doğu
Erdogan, Goker
Erdogmus, Deniz
Erickson, Brian
Ervas, Francesca
Falomir, Zoe
Fan, Judith E.
Fang, Wen-Chieh
Farina, Mirko
Farkas, Igor
Fausey, Caitlin
Favela, Luis H.
Fedyk, Mark
Feher, Olga
Fehringer, Benedict C. O. F.
Feiman, Roman
Feinstein, Jonathan
Felsman, Peter
Feng, Shuo
Ferguson, Brad
FernandezDuque, Diego
Ferry, Alissa
Files, Benjamin T
Finlayson, Mark A
Finley, Sara
Fiorella, Logan
Firestone, Chaz
Fisher, Anna

Fitneva, Stanka A.
Flor, Nick
Floyd, Sammy
Flusberg, Stephen
Folstein, Jonathan
Foltz, Anouschka
Forbus, Ken
Foster-Hanson, Emily
Fourtassi, Abdellah
Fox, Amy
Frame, Mary
Frank, Stefan
Frank, Stella
Franke, Michael
Frankenstein, Julia
Franklin, Nicholas
Frassinelli, Diego
Frazer, Alexandra K
Freed, Michael
Freksa, Christian
Friemann, Paulina
Frixione, Marcello
Frost, Rebecca
Fu, Kate
Fu, Wai-Tat
Fuhl, Wolfgang
Furman, Reyhan
Fürnkranz, Johannes
Futrell, Richard
Gabora, Liane
Gabriel, Rami
Gagliardi, Francesco
Galati, Alexia
Gallagher, Natalie
Gallardo, Judith Charlene
Gambi, Chiara
Garcia, Raul Royden
Gardony, Aaron L
Garner, Kelly Grace
Garsoffky, Bärbel
Garten, Justin
Gavish, Nirit
Gazzo Castaneda, Lupita Estefania
Gelman, Susan
Genc, Yegin
Gentner, Dedre
Gero, John
Gershman, Samuel
Gerstenberg, Tobias
Gervás, Pablo
Gessell, Bryce
Getz, Heidi
Ghafurian, Moojan
Ghiran, Ana-Maria

Giardino, Valeria
Gierasimczuk, Nina
Giersch, Anne
Gil, Sandrine
GIuntini, Roberto
Glavan, Joseph
Glebkin, Vladimir
Gliozzi, Valentina
Gluck, Kevin
Godwin, Karrie
Goebel, Peter Michael
Goel, Ashok
Göksun, Tilbe
Goldberg, Adele
Goldwater, Micah
Golinkoff, Roberta M
Gong, Tao
Gonnerman, Laura
Gonzalez-Marquez, Monica
Gonzalez, Cleotilde
Gosavi, Radhika
Goudbeek, Martijn
Govindarajulu, Naveen Sundar
Graham, Erin N
Gramann, Klaus
Grant, Erin
Gray, H.M.
Greenauer, Nathan
Grieben, Raul
Gries, StefanTh.
Griffiths, Tom
Grigoroglou, Myrto
Grimm, Lisa
Grinberg, Maurice
Gruenenfelder, Thomas M
Grzyb, Beata
Gu, Yan
Guerrero, Ivan
Guha, Amal
Gunderson, Elizabeth
Gunzelmann, Glenn
Guo, Cai
Guo, Cheng
Gureckis, Todd M
Gussow, Arella
Gutzwiller, Robert S
Haertl, Holden
Hafri, Alon
Hall, Jessica
Hallam, Glyn
Halpern, David J.
Halverson, Tim
Hamker, Fred
Hampton, James
Hamrick, Jessica
Haring, Kerstin S
Harmon, Stephen
Harner, Hillary
Haroz, Steve
Harris, Jesse
Harrison, Anthony
Hawkins, Robert
Hayashi, Yugo
Hayes, Brett
Hayhoe, Mary
He, Angela Xiaoxue
Hegarty, Mary
Heidari Kapourchali, Masoumeh
Helie, Sebastien
Heller, Daphna
Hemmatian, Babak
Hendrickson, Andrew
Heng, Li
Henne, Paul
Hernández-Orallo, José
Heyman, Gail D.
Hiatt, Laura
Hidaka, Shohei
Hill, Katherine
Hinaut, Xavier
Hinrichs, Thomas
Hirshberg, Matt
Hitczenko, Kasia
Ho, Mark K
Hochman, Guy
Hoek, Jet
Hoetjes, Marieke
Hoffman, Paul
Hohenberger, Annette
Holbrook, Colin
Holmes, Kevin J.
Holyoak, Keith
Hoover, Joe
Hopman, Elise W. M.
Horne, Zachary
Horton, William S
Hough, Alexander
Houlihan, Sean Dae
Howard, Marc
Howes, Christine
Hristova, Evgenia
Hristova, Penka
Hsiao, Janet
Hsu, Nina
Hubbard, Edward M.
Hubscher, Roland
Hudson Kam, Carla
Huette, Stephanie
Hunter, Lindsay
Hupp, Julie
Husband, Edward Matthew
Hussey, Erika
Icard, Thomas
Imai, Mutsumi
Indurkhya, Bipin
Isaac, Alistair
Ishikawa, Toru
Ito, Aine
Izquierdo, Eduardo
Jachmann, Torsten Kai
Jacob, Mikhail
Jacobs, Cassandra
Jacobs, Robert
Jäger, Gerhard
Jain, Ajit
Jain, Saransh
Jäkel, Frank
James, Ariel
Jansen, Rachel

Jennifer, Hofmann
Jern, Alan
Jimenez, Guillermo
Johnson, Samuel
Johnston, Angie
Jokinen, Jussi
Jones, Gary
Jones, Lara L.
Joyner, David
Juvina, Ion
Kabasele, Philothe
Kachergis, George
Kadihasanoglu, Didem
Kahl, Sebastian
Kajić, Ivana
Kan, Irene
Kandaswamy, Subu
Kang, Wang-Cheng
Kannan, Amar Viswanathan
Kapnoula, Efthymia C
Kapucu, Aycan
Karaminis, Themis
Karanam, Saraschandra
Karimi, Hossein
Kaygusuz, Yasin
Keane, Mark T
Kearns, John T
Kelley, Troy Dale Kelley
Kello, Christopher
Kelly, Laura J
Kelty-Stephen, Damian
Kenett, Yoed
Kennedy, Brendan
Kennedy, William
Kershaw, Trina
Kersten, Alan W.
Kersten, Luke
Khan, Azam
Khemlani, Sangeet
Kholodova, Aline
Khooshabeh PhD, Peter
Kidd, Celeste
Kietzmann, Tim
Kim, Mi Song
Kim, Woojae
Kirfel, Lara
Kiss, Szabolcs
Klatzky, Roberta L.
Kleiman-Weiner, Max
Kleinschmidt, Dave
Klenk, Matthew
Kliesch, Christian
Knoeferle, Pia
Koedinger, Ken
Koehne, Judith
Kogon, Drew
Komatsu, Takanori
Kominsky, Jonathan F.
Konderak, Piotr
Konopka, Agnieszka E
Kontogiorgos, Dimosthenis
Kopp, Stefan
Korpan, Raj
KORTE, JESSICA L
Kosie, Jessica E.
Kotelova, Rossitza
Kothiyal, Aditi
Krafft, Peter
Krahmer, Emiel
Krajcsi, Attila
Kranjec, Alexander
Krems, Josef
Krishnaswamy, Nikhil
Krogh-Jespersen, Sheila
Kronmuller, Edmundo
Kroupin, Ivan
Krukar, Jakub
Krusienski, Dean
Krypotos, Angelos
Kryven, Marta
Krzyzanowska, Karolina
Kuchinsky, Stefanie E
Kucker, Sarah
Kuhn, Gustav
Kukona, Anuenue
Kumar, Kiran N
Kumova, Bora
Kunda, Maithilee
Kurdi, Benedek
Kurtz, Kenneth
Kurumada, Chigusa
Kushnir, Tamar
Kvam, Peter
Lake, Brenden
Lakeland, Corrin
Lampinen, Andrew
Landau, Barbara
Landy, David
Lange, Nicholas
Lapesa, Gabriella
Larue, Othalia
Lauer, Jillian
Law, Edith
Lawson, Chris
Lazaroff, Emma
Le Corre, Mathieu
Le Guen, Olivier
Lea, Brooke
Leake, David
Lebani, Gianluca
Ledda, Antonio
Lee, Jessica
Lee, Michael
Lee, Saebyul
Legare, Cristine
Lehet, Matthew I.
Lehman, Blair
Lenci, Alessandro
León, Carlos
Leonard, Julia
Leonardi, Giuseppe
Leshinskaya, Anna
Lester, Nicholas
Levy, Roger
Lewis, Ashley G
Lewis, Rick
Li, Peggy
Li, Rui
Li, Sara Tze Kwan

Licato, John
Lichtenstein, Patricia
Lieberman, Amy
Lieder, Falk
Lieto, Antonio
Lignos, Constantine
Ligorio, Tiziana
Lin, Hause
Linder, Rhema
Lindner, Felix
Lindsay, Shane
Lins, Jonas
Linzen, Tal
Liquin, Emily G
Little, Daniel
Little, Hannah
LIU, NIAN
Liu, Shari
Liu, Tong
Logan, John
Lohmann, Johannes
Lombrozo, Tania
Long, Bria
Long, Duri
Long, Lyle N.
Lonsdale, Deryle
Lorenz, Tamara
Lotte, Fabien
Lovett, Andrew
Loy, Jia
Lu, Hongjing
Lucas, Chris
Lucero, Che
Luchkina, Elena
Luhmann, Christian
Luo, Yiwei
Lupyan, Gary
Lynch Ph.D., Michael F.
Lynott, Dermot
Lyons, Joseph
Ma, Tengyu
MacDonald, Kyle
MacDonald, Maryellen
Macedo, Luís
Mack, Michael
Madan, Piyush
Maehigashi, Akihiro
Magerko, Brian
Magnani, Lorenzo
Maier, Emar
Maihot, Fred
Majid, Asifa
Maldonado, Mora
Mallot, Hanspeter
Maloney, Laurence T
Malt, Barbara C.
Man, Kingson
Mandelbaum, Eric
Maravilla, Francisco
Marelli, Marco
Marghetis, Tyler
Markant, Doug
Markman, Art
Marno, Hanna
Marsh, Jessecae
Marti, Louis
Martin, Alexander
Martinez, David
Masnick, Amy
Mason, Robert
Masumi, Akira
Mather, Emily
Matlen, Bryan
Matsumuro, Miki
Matthews, Percival
Matusevych, Yevgen
May, Kaitlyn
Mayrhofer, Ralf
Mayseless, Naama
McClelland, Jay
McClimens, Brian
McDonald, David D.
McLean, Janet F
McRae, Ken
McShane, Marjorie
Meadows, Ben
Mech, Emily N
Meder, Bjoern
Mehrotra, Siddharth
Mekik, Can
Melcher, David
Mello, Catherine
Melnick, Robin
Mendonca, David
Meng, Yuan
Menon, Mythili
Merritt, Haily
Metcalf, Katherine
Metcalfe, Janet
Meylan, Stephan C.
Michaelides, Christos
Michaelis, Laura A.
Milburn, Evelyn
Miller, Craig
Miller, Hilary
Miller, Josh Aaron
Miller, Lee
Mills, Gregory
Minelli, Alessandro
Mistry, Percy
Miton, Helena
Mizuguchi, Takashi
Modoni, Gianfranco
Modrek, Anahid S.
Mollica, Francis
Monaghan, Padraic
Monroy, Claire
Moon, Jung Aa
Morett, Laura M.
Morgan, Emily
Morisseau, Tiffany
Morris, Bradley
Morton, Neal W
Moshkina, Lilia
Moss, Jarrod
Moss, Larry
Mostafazadeh Davani, Aida
Motamedi, Yasamin
Mousas, Christos

Mousavi, Mahta
Mueller, Shane
Munnich, Edward
Munro, Paul
Muntanyola-Saura, Dafne
Münzer, Stefan
Murdock IV, J William
Murphy, Gregory
Murray, Grace
Musslick, Sebastian
Myers, Christopher
Myers, Matthew Ross
Nahum, Mor
Naik, Shweta
Nakamura, Kuninori
Narasimhan, Bhuvana
Nayyar, Mollik
Negen, James
Nelson„ Jonathan D.
Newcombe, Nora
Ngoon, Tricia J.
Nielsen, Alan K.S.
Nirenburg, Sergei
Nobandegani, Ardavan S.
Noelle, David C.
Noelle, Jonas
Noles, Nicholaus
Noll-Hussong, Michael
Nordmeyer, Ann E.
Nozari, Nazbanou
Nyhout, Angela
O'Grady, Shaun
O'Mara Shimek, Michael P
O'Neill, Kevin J.
O'Shaughnessy, David
Oates, Tim
Obaidellah, Unaizah
Odic, Darko
Oduniyi, Erick
Ojha, Amitash
Ojha, Suman
Okano, Kana
Olkkonen, Maria
Olney, Andrew
Olteteanu, Ana-Maria
Ong, Desmond
Onnis, Luca
Ontanon, Santiago
Opfer, John
Orbán, Gergő
Orenes, Isabel
Osherson, Daniel
Ostrow, Korinn
Oury, Jacob David
Ouyang, Iris Chuoying
Over, David
Overmann, Karenleigh Anne
Oxenham, Andrew
Padilla, Lace
Pajak, Bozena
Pala, Kiran
Palatnik, Alik
Pande, Prajakt
Pani, John R
Papafragou, Anna
Parker, Dan
Parker, Jeffrey R.
Parnamets, Philip
Parpart, Paula
Pastore, Luigi
Patalano, Andrea
Patel, Purav
Patel, Purav J
Patrick, Sturt
Patson, Nikole
Patterson, John
Pautler, David
Pazzaglia, Francesca
Pecher, Diane
Peebles, David
Peloquin, Benjamin
Pennisi, Antonio Antonio
Pereira, Alfredo F.
Perelman, Brandon Scott
Pérez y Pérez, Rafael
Perfors, Amy
Perikos, Isidoros
Perlman, Marcus
Perniss, Pamela
Perry, Bob
Persaud, Kimele
Peters, Megan
Peterson, Joshua
Petkov, Georgi
Petrovych, Veronika
Pezzelle, Sandro
Philalithis, Eugene
Phillips , Jonathan S
Phillips, Steven
Piantadosi, Steven
Piñango, Maria M
Pipergias Analytis, Pantelis
Pitt, Benjamin
Pitti, Alex
Plate, Rista C
Ploran, Elisabeth
Poeppel, Jan
Politzer-Ahles, Stephen
Popov, Vencislav
Pothos, Emmanuel
Pouncy, Thomas
Powell, Derek
Pozniak, Celine
Prasada, Sandeep
Pyke, Aryn
Pynadath, David V.
Qian, Zhiying
Queen, Jennifer
Quinn, Molly S
Rabagliati, Hugh
Rabkina, Irina
Rabovsky, Milena
Rafferty, Anna
Ragni, Marco
Ralph, Jason
Ramirez-Aristizabal, Adolfo G.
Ransom, Keith James
Räsänen, Okko
Rashedi, Roxanne

Raviv, Limor
Rayz, Julia
Rebedea, Traian
Recio García, Juan A.
Reddy, Jayasankara
Redeker, Gisela
Reed, Stephen
Rehder, Bob
Reichherzer, Thomas
Reinholtz, Nicholas
Rekabdar, Banafsheh
Relaford-Doyle, Josephine
Remington, Roger
Reschke PhD, Peter
Rescorla, Michael
Resnick, Ilyse
Reviewer, Test
Rhodes, Darren
Ricciardi, Emiliano
Rice, Caitlin
Rice, Patrick
Richey, J. Elizabeth
Richland, Lindsey Engle
Richter, Kai-Florian
Richter, Mathis
Riesterer, Nicolas Oliver
Riley, Sean
Rissman, Lilia
Risto, Malte
Robbins, Philip
Robert, Serge
Robinette, Paul
Roby, Erin
Roche, Jennifer
Roehrbein, Florian
Roembke, Tanja
Roettger, Timo B
Rohlfing, Katharina
Roose, Kaitlyn M
Roscoe, Rod D.
Rosenbloom, Paul S.
Rosner, Agnes
Rossano, Federico
Rotaru, Armand
Rothe, Anselm
Rothganger, Fred
Rottman, Benjamin
Ruggeri, Azzurra
Rule, Joshua S
Russwinkel, Nele
Ryali, Chaitanya K
Ryskin, Rachel
Sadeghi, Sepideh
Sagi, Eyal
Sakas, William
Sala, Giovanni
Salapska, Joanna
Saldana, Carmen
Saleh, Mai Sabry
Salvi, Carola
Salvucci, Dario
Sanches de Oliveira, Guilherme
Sandamirskaya, Yulia
Sanz, Ricardo
Sapienza, Alessandro
Sarin, Arunima
Sato, Yuri
Satyadas, Antony
Sauerwald, Kai
Saunders, Robert
Savoye, Yann
Saxe, Andrew
Scassellati, Brian
Schachner, Adena
Schlegel, Daniel R
Schloesser, Daniel S.
Schlotterbeck, Fabian
Schmid, Ute
Schneegans, Sebastian
Schommer, Christoph R
Schöner, Gregor
Schorlemmer, Marco
Schotter, Liz
Schroeder, Noah
Schuler, Kathryn
Schultheis, Holger
Schulz, Eric
Schulz, Laura
Schunn, Chris
Schwartz, Nora Alejandrina
Schwering, Steven
Scontras, Gregory
Scott, Daniel
Seegelke, Christian
Seifert, Colleen
Sell, AndreaJ.
Semenuks, Arturs
Sense, Florian
Sera, Maria
Serrano PhD, J.Ignacio
Serre, Thomas
Setzler, Matthew
Seymour, Travis
Shafto, Meredith
Shafto, Michael G
Shafto, Patrick
Shah, Priti
Shanidze, Natela
Shantz, Kailen
Shaw, Jason A.
Sheppard, Chris
Shevlin, Henry
Shi, Wenlei
Shinohara, Kazuko
Shtulman, Andrew
Shubeck, Keith
Shukla, Mo
Shvartsman, Michael
Siegel, Max
Siegelman, Noam
Silliman, Daniel
Silvey, Catriona
Simko, Juraj
Simms, Nina
Sims, Chris R
Sinha, Priyanka
Sinha, Tanmay
Sinnett, Scott
Sloman, Sabina

Sloman, Steven
Smaldino, Paul
Smith, Carl
Smith, Kenny
Smith, Kevin A
Snoddy, Sean
Snow, Erica
Sobel, David M.
Solari, Fabio
Somers, Sterling
Sosa, Felix Anthony
Soylu, Dr. Firat
Spaulding, Shannon
Speed, Laura
Speekenbrink, Maarten
Spenader, Jennifer
Spencer, John
Spike, Matthew
Spiller, Stephen A
Spino, Joseph M
Spranger, Michael
Spurrett, David
Srinivasan, Mahesh
Srivastava, Nisheeth
Srivastava, Priyanka
Stafford, Tom
Stefanini, Fabio
Stenning, Keith
Stephan, Simon
Steven, Steven Smith M
Stewart, Terrence C
Stocco, Andrea
Stojanov Stojanov, Georgi Kiril
Suanda, Sumarga H.
Suchow, Jordan
Sugden, Nicole
Sulik, Justin
Sullins, Jeremiah
Sullivan, Jess
SUMER, BEYZA
Sumner, Emily
Sun, Chen
Sun, Ron
Sussman, Abigail
Swanson, Link
Syrett, Kristen
Tabor, Whitney
Tan, Shawn
Tang, Shuai
Tang, Yun
Tappe, Heike
Tas, Caglar
Taxitari, Loukia
Tekülve, Jan
Temperley, David
Tenbrink, Thora
Tenenbaum, Josh
tentori, katya
Tessler, Michael Henry
Test, Test
Teubner-Rhodes, Susan
Thai Ph.D., Khanh-Phuong
Thibodeau, Paul
Thill, Serge
Thomas, Bobby
Thomas, Rick P
Thompson, Bill
Thompson, Clarissa A
Thompson, Jim
Thomson, Robert
Thorstad, Robert
Thrash, Tyler
TIJUS, Charles Albert
Tillas, Alex
Tillman, Katharine A
Timpf, Sabine
Todaro, Rachael D
Todorov, Alex
Toivonen, Hannu
Toivonen, Ida
Tomkins-Flanagan, Eilene
Tomlinson Jr, John Michael
Tomov, Momchil
Toskos Dils, Alexia
Trafton, Greg
Travers, Eoin
Trent, Scott
Trickett, Susan Bell
Trninic, Dragan
Tso, Ricky Van-yip
Tsutsui, Satoshi
Tubridy, Shannon
Tummolini, Luca
Tupper, Paul
Tutunjian, Damon
Tversky, Barbara
Twomey, Katherine E
Tylen, Kristian
Ullman, Tomer
Ünal, Ercenur
Ungemach, Christoph
Unger, Layla
University, Jidong State
Uno, Ryoko
Urminsky, Oleg
Utsumi, Akira
Vales, Catarina
Valle-Lisboa, Juan C
Vallee-Tourangeau, Frederic
Van de Cruys, Tim
van de Pol, Iris
van den Berg, Ronald
van der Wal, Natalie
van Gompel, Roger
van Hoef, Rens
van Schijndel, Marten
van Vugt, Marieke Karlijn
vanderWel, Robrecht
Vankov, Ivan
Varma, Sashank
Vasilyeva, Nadya
Veale Veale, Tony
Veale, Richard
Veerabadran, Vijay
Ventura, Dan
Vergilova, Yoana
Vertolli, Michael Olias
Vigliocco, Gabriella
Villareal, Manuel

Vitale, Jonathan
Vogel, Carl
Vollmeyer, Regina
von der Malsburg, Titus
von Sydow, Momme
Vong, Wai Keen
Vosoughi, Soroush
Votsis, Ioannis
Vul, Ed
Vuong, Loan
Wagner, Alan
Walker, Drew
Wang, Jane
Wang, Panqu
Wang, Tianlin
Wang, Tony
Ward, Nigel
Warick, Walter
Washburn, Auriel
Weidemann, Christoph T
Weisberg, Steven
Weitz, Katharina
Welsh, Matthew Brian
Wen, Nicole
Wertheim, Julia
Westbury, Chris
Whalen, Andrew
White, Aaron Steven
White, James
White, Katherine
Wiese, Eliane Stampfer
Wiggins, Geraint A.
Wilkinson, Meredith
Willems, Roel M.
Williams, Tom
Willits, Jon
Wilson, Colin
Wilson, Jason R.
Wilson, Nicholas
Wiltshire, Travis J.
Wirzberger, Maria
Wnuk, Ewelina
Wong, HingYi Orieta
Wood, Sharon
Worthy, Darrell
Wu, Charley Mingshuo
Wu, Jiajun
Wu, JU-CHUAN
Xie, Xin
Xu, Fei
Xu, Linger
Xu, Mingze
Xu, Xiaotong (Tone)
Xu, Yang
Xu, Yuhang
Yan, Xiaogang
Yang, Charles
Yang, Lee-Xieng
Yang, Scott Cheng-Hsin
Yang, Shan
Yeo, Amelia
Yildirim, Ilker
Yim, Hyungwook
Yin, Siyuan
Yip, Michael C. W.
Yoon, Erica
Yoon, Si On
Yoon, SiOn
Yoshida, Hanako
Young, Andrew G
Yovel, Galit
Yuan, Lei
Yun, Jiwon
Yurovsky, Dan
Zabelina, Darya L.
Zaitchik, Deborah
Zanatto, Debora
Zaslavsky, Noga
Zedelius, Claire Marie
Zemla, Jeffrey
Zepeda, Cristina D.
Zettersten, Martin
Zhang, Muye
Zhang, Ningyu
Zhang, Qiong
Zhao, Fangyun
Zhou, Peiyun
Zhu, Jerry
Zhu, Rebecca
Zhu, Yixin
Zipitria, Iraide
Zish, Kevin
Zourou, Filio